Be the Outlier

How to Ace Data Science Interviews

Shrilata Murthy

BE THE OUTLIER

HOW TO ACE DATA SCIENCE INTERVIEWS

SHRILATA MURTHY

New Degree Press

Be the Outlier
How to Ace Data Science Interviews

ISBN 978-1-64137-985-4 *Paperback*

978-1-64137-877-2 *Kindle Ebook*

978-1-64137-878-9 *Ebook*

To Ganapati and Samyra, for always being there.

CONTENTS

INTRODUCTION

If you work in the field of data science, you probably know how daunting the interview process can be. As a candidate, you are being tested on your skills in statistics, technology, business, and countless other subsections of these broad categories. This can make the process overwhelming, and you may feel unprepared no matter how much you've practiced.

As a data scientist myself, I know how nerve-racking preparing for interviews can be, and they're even worse when you aren't sure what to expect. But if you weed through and peel back the layers, you will uncover a pattern—a pattern that is intuitive and speaks to how the different elements in a data scientist's job role come together.

That's exactly why I wrote this book, which covers all aspects of a data science interview in today's industry and highlights how you can differentiate yourself from your peers and Be the Outlier.

My data science journey began back in 2013 when I joined the Masters in Analytics program at Northwestern University. Ours was the second graduating cohort at Northwestern from the program at the time.

From then to now, the data science field has evolved a lot over the past years. Indeed, one of the top job sites reported that data science postings have rocketed 256 percent—more than tripling since December 2013.[1] The industry, overall, has seen staggering growth in the use of data science and machine learning in its day-to-day business, and this has translated into an increase in data science jobs over the years.

Over the last decade, not only have jobs available for data science seen incredible growth, but the nature of the job role itself has evolved. Closer to graduation, as I was preparing for interviews, I realized how data science roles differed across companies depending on why, what, and how data science is

used in a company. Needless to say, the strategy to prepare for interviews for each company changed accordingly.

I made extensive notes about the key concepts I prepared for, programming questions I was typically asked, and probability questions I was quizzed on. Every now and then, when prospective data scientists reached out to me to get pointers on how to crack these interviews, I found myself referring to my notes from my graduation days.

Although the notes I made were current to when I was interviewing, I saw gaps in what I had collated versus what the industry is asking for now. While I was up to date with how the interview process and candidate expectations have changed in the industry that I work in, I was curious to know how things have progressed industry wide.

I reached out to my network in the field, including those who worked at consulting companies, product-based and technology firms, industry-specific companies, startups, and everything in between. I learned a lot in the process and did my best to relay this knowledge to the folks who reached out to me with questions of this nature.

When talking to people on both sides of the table—job seekers and experienced data scientists—I noticed that while good material to prepare for software-programmer, product-manager, and data-analyst types of roles was out there, there was a lack of comprehensive material for a data science interview. I found many question banks online with pointers to various types of questions asked, but not a one-stop-shop resource to prepare for data science interviews.

With this book, I will walk you through the different types of data science roles and the specific skill sets you need to focus on when preparing for the pertinent interview. This book will also provide you with sample answers that have been vetted by experienced data scientists with tips and tricks to help you stand out as a stellar candidate.

Whether you are an aspiring data scientist looking for a break in the field, a data scientist looking for a job change, or just someone who wants to learn more about data science, this book is for you.

I wish you the very best as you prepare for your next interview. A strong foundation and practice will get you where you aim to be!

---

1 “Data Scientist: A Hot Job That Pays Well,” Indeed Hiring Lab, US, accessed May 5, 2020.

PART 1

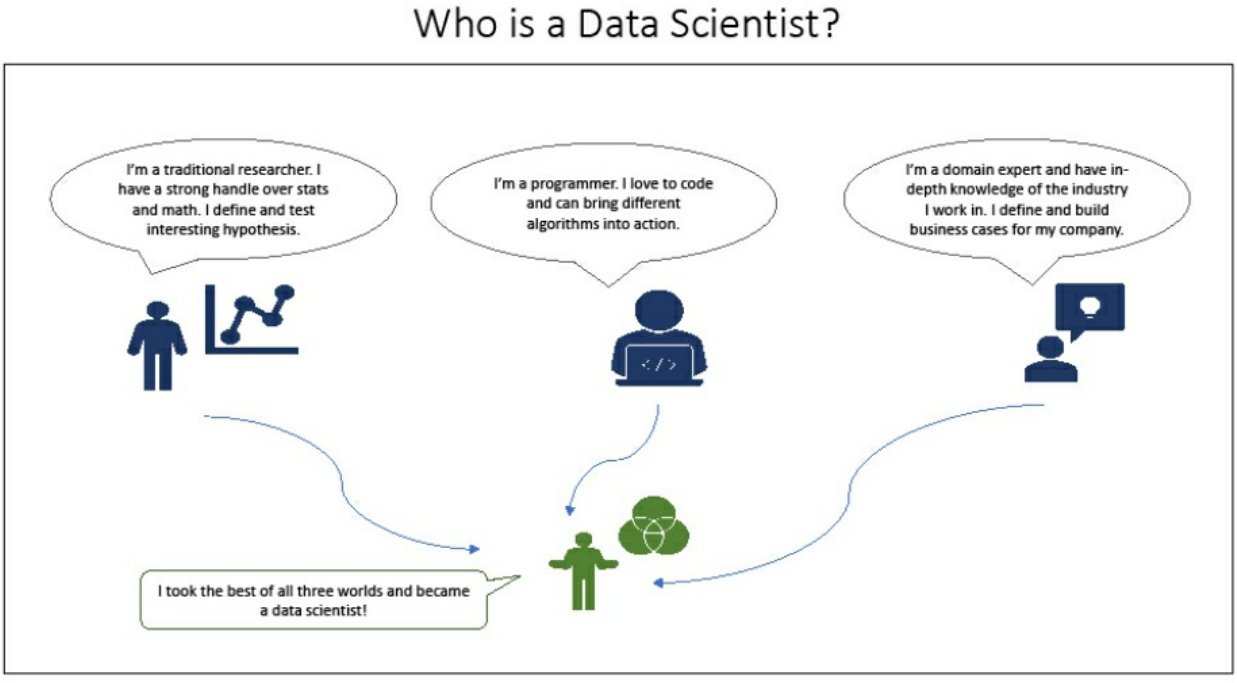
GETTING STARTED

CHAPTER 1

WHAT IS DATA SCIENCE?

Before we jump into the nitty-gritty of a data science-focused interview, let's take a step back and understand what data science is and how it all started.

At a grassroots level, I like to think of the data science role as a combination of a traditional researcher, a computer programmer, and a business analyst. The graphic below helps paint this picture of the three key skill sets coming together and forming what we now call data science.



Who is a data scientist?

Essentially, I like to think of a data scientist as someone who can extract meaning from complex data and assist in decision-making. A data scientist makes use of methods from statistics and machine learning for their analysis and has the technical skills to codify their work.

The term "data science" was coined in 2008 by DJ Patil and Jeff Hammerbacher, who were then the respective leads of data and analytics efforts at LinkedIn and Facebook. In an interview with *Observer,* Patil, the former chief data scientist for the Obama administration, said, "I was at LinkedIn building the data team, and Jeff Hammerbacher [co-founder of Cloudera] was bustling at Facebook's data team, and we would collaborate and compare notes sometimes. One of the things we realized was that we didn't know what to call ourselves."[2]

When asked about the definition of the data scientist role, Patil said, "I am generally opposed to trying to define it too rigorously. The important thing is how you use data to interact with the world, study it, and try to come up with new things. Some of those things are new products, like a self-driving car or your weather app. Others are data analyses used to help people make assessment for everything from loans to health care decisions. There are all kinds of data scientists. Maybe the title survives and maybe it turns into something else. But I think the most powerful thing here is that we are using data in novel ways to build things."[3]

More than a decade later, the title not only survives—it thrives and continues to evolve.

The impact of data science on the job market is huge. One can find multiple reports online that speak to this in depth. According to LinkedIn's third annual US Emerging Jobs Report, the data scientist role has seen a 37 percent annual growth per this report and is ranked third among the top fifteen emerging jobs in the US.[4]

Data science skills proliferate across industries and aren't just predominant in any one industry alone. When I spoke to people working in consulting, who worked across industries to serve their clients in the analytics space, they mentioned an increasing use of data science applications in the financial, consumer products, healthcare, and insurance industries.

As different industries continue to invest in data science resulting in more innovation in the field, the need for a workforce to serve in this rapidly emerging job market is growing.

2 Sissi Cao, “What on Earth Is a Data Scientist? The Buzzword’s Inventor DJ Patil Spills All,” Observer, November 9, 2019.
3 Sissi Cao, “What on Earth Is a Data Scientist? The Buzzword’s Inventor DJ Patil Spills All,” Observer, November 9, 2019.
4 “LinkedIn: 2020 Emerging Jobs Report,” LinkedIn, accessed May 16, 2020.

CHAPTER 2

WHAT IS THE MARKET LOOKING FOR?

Now that we understand what data science entails, we must also understand what skills the market is looking for in a data scientist.

When I spoke to Laura Siahaan, business IT data scientist and team lead at NASA's Jet Propulsion Laboratory, she explained that a key aspect of her job involves communicating analytical findings to a non-technical audience. Hence, when interviewing data scientists, one of the main skills she looks for is the ability to explain complex data science concepts in a simple and precise manner. She highlighted the ability to tell a story using data as a differentiator among candidates.

Translating complex findings into actionable insights can be a daunting task at times, but having this skill set in your back pocket will make your work rewarding.

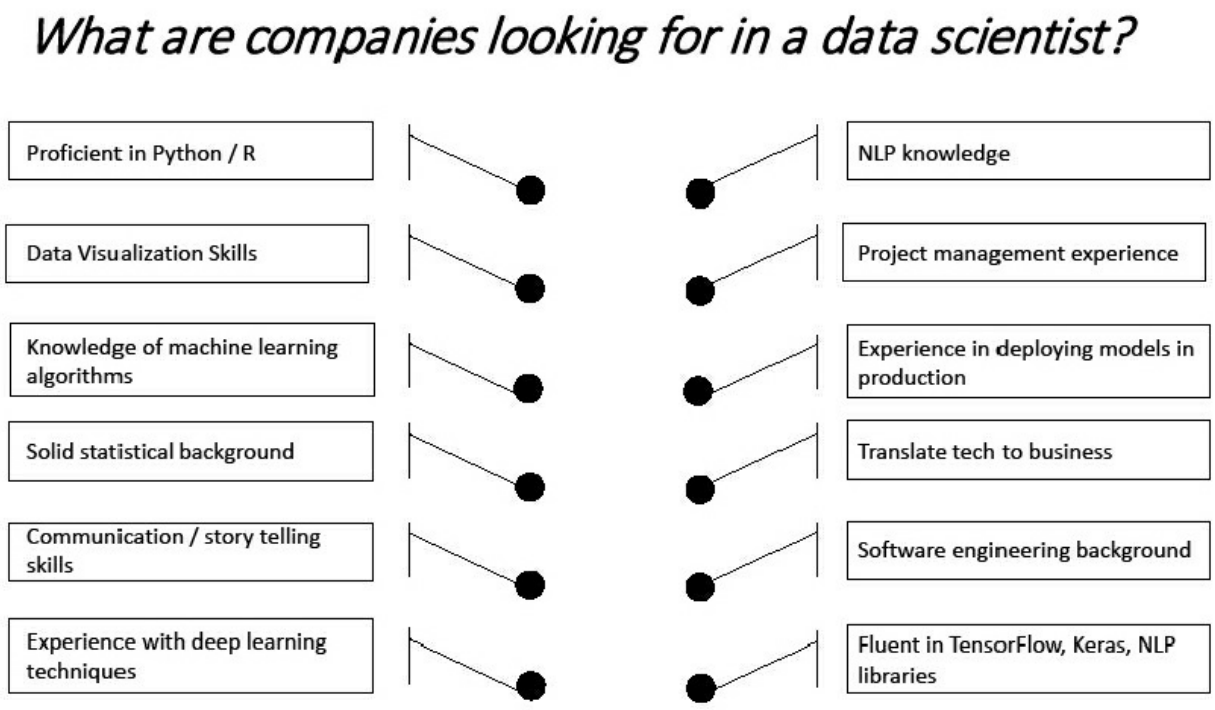
A consultant in the analytics space from the Big Four said, "When you apply for a role in a consulting company, you are a consultant first then a data scientist. It is crucial to be a strategic thinker and understand holistically the business problem that you are trying to solve. When you absorb all the information at a macro level to begin with, it becomes a lot easier to deep dive and solve the problem analytically. With the business requirements and expectations nailed down, it should then come naturally for a data scientist to formulate a data request and recommend analytical techniques to be used to solve a problem."

It was then that I realized how important having a bird's eye view of the

problem at all times is. As a data scientist, getting lost in the weeds is easy. Having this as a constant reminder will help course correct if needed!

While both of these inputs touch more on the non-tangible skill sets of problem solving and effective communication, roles exist where the knowledge of a particular tool or programming language is a must. A product manager with Google told me their team almost always uses Python as the language of choice. Given that their team heavily relies on deep learning techniques to conduct their analysis, knowledge of Pytorch and TensorFlow is required.

After having conversations with people in the data science field, I noticed the wide range of skill sets in demand in the marketplace today. The visual below captures the different answers I received from data scientists across industries.



What are companies looking for in a data scientist?

With such a wide range of answers, I sought to understand how these skill sets vary across different data science roles. If you search for “data scientist” job openings on LinkedIn or any other job posting site, you will notice that the results are listed under varying job titles. Some of the popular ones are

Artificial Intelligence Specialist, Data Scientist, Machine Learning Engineer, Data Analyst, Data Visualization Specialist, Business Analyst, Data Engineer, and Data Architect.

I grouped these for brevity and to allow for a more strategic approach for anyone preparing for an interview in these fields.

Machine Learning Engineer: Per LinkedIn's Emerging Jobs Report, this role encompasses skills in the areas of machine learning, deep learning, TensorFlow, Python, and natural language processing.[5] Here are a few highlights of job requirements posted by Apple and Spotify for the machine learning engineer role.

- Apple:

– Expertise and experience in various facets of machine learning and natura language processing, such as classification, feature engineering information extraction, structured prediction, clustering, semi-supervise learning, topic modeling, and ranking.
– Programming experience in one or more of the following: Java, C+- Python or equivalent.
– Contributions to research communities, e.g., ACL, NIPS, ICML, an CVPR.

- Spotify:

– You have hands-on experience implementing production machin learning systems at scale in Java, Scala, Python, or similar language Experience with XGBoost, TensorFlow is also a plus.
– You care about agile software processes, data-driven developmen reliability, and disciplined experimentation.

As you may notice in both job descriptions, they emphasize being proficient at an at-scale programming language such as Python or Java. They also have a distinguished requirement of being well-versed in the latest in machine learning techniques. A machine learning engineer typically will have a chance to work on the state-of-the-art technological environments and have a chance

to use the cutting-edge techniques to differentiate the company's product in the competitive market.

Data Scientist: Touted as "The Sexiest Job of the Twenty-First Century," the data scientist role at a broader level encompasses knowledge of descriptive, predictive, and prescriptive analytics.[6] A few nuggets of the job requirements from QuantumBlack, a McKinsey Company, showcase that this role popularly requires knowledge of machine learning algorithms, Python, R, and SQL.

- Experience in statistical modelling and machine learning techniques.
- Programming experience in at least two of the following language Python, Scala, SQL.
- Experience in applying data science methods to business problems.
- Good presentation and communication skills, with the ability to ex complex analytical concepts to people from other fields.

Business Analytics Specialist: This role is different from the machine learning engineer and data scientist role in that it has more emphasis on use case development for analytics and its requirement of the specific industry/domain knowledge in which it is applicable. This role also requires excellent communication and presentation skills, considering you need to be able to get buy-ins from key stakeholders in the organization, who are often an executive audience of C-suite members. Additionally, the role may require good interpersonal skills to enable adoption of analytical techniques through the organization. Here's a highlight from Unilever's job posting for a Business Analytics Manager:

- Serve as the trusted data and analytics consultant for leaders withi organization that provide products-solutions that rapidly solve for bus issues.
- Delivery of the total joint business plan priorities between different team
- Ensure the career development of the in-market data scientist team as a t pipeline into the cluster hubs.

Additionally, note the below job requirement snippets for the Business Analytics role, as they occurred in multiple other companies as well.

- Expert knowledge in statistics (Regression, Clustering, Random Fo Decision Trees, Optimization, Time Series, Probability, and other re advanced methodologies).
- Working knowledge of a visualization tool such as Microsoft Pow QlikView, or Tableau.

- Working knowledge and basic skills to code in R, R Shiny, and Micı Azure Machine Learning.

As it may have come to your attention as well, the job requirements vary across the different roles that we have seen so far. Thus, how you prepare for an interview varies widely depending on the job role. In the next chapter, we'll explore how you can efficiently identify what to prep for!

---

5 "LinkedIn: 2020 Emerging Jobs Report," LinkedIn, accessed May 16, 2020.

6 Thomas H. Davenport and D.J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," Harvard Business Review, October 2012.

CHAPTER 3

WHAT DO YOU WANT TO DO?

With a strong handle on the different roles available in the market, let's move on to *you*! This book will help you prepare for interviews for one of the three roles mentioned in the previous chapter. We will discuss the popular questions observed in data science-related interviews and provide sample answers to each of the questions. Additionally, I will help you formulate a framework you can use to answer questions not seen in this book to differentiate yourself from other candidates.

The book spans many different areas of data science, covering machine learning, deep learning, statistical knowledge, and technology/programming questions. We will also strategize how to tackle online coding tests and take-home exams and dive into some case studies to ensure you are well prepped for any case interviews rounds you may have.

To be clear, this book will not tackle interview strategies for data engineer and data architect roles as those focus more on developing and maintaining databases and large-scale processing systems, which is out of scope for this book.

To help you prepare for the different areas in data science, I constructed the visual below to serve as a guideline.

To the left, I have listed the key areas to prepare for and then drilled down into the specific skill sets in each. To the right, I have laid out the three key job role groupings we discussed and a heat map indicating how the importance of different skill sets varies across roles.

		Machine Learning Engineer	Data Scientist	Business Analytics Specialist
Stats / Machine Learning	Fundamental statistical knowledge such as probability, t-tests, ANOVA, research design			
	Fluent in Machine Learning techniques Regression, Trees, Clustering			
	Fluent in TensorFlow, Keras, NLP libraries			
	Knowledge of current modeling tools such as XGBoost or Vowpal Wabbit (or equivalent)			
Technology / Programming	Extract and manipulate large datasets with fluency in SQL			
	Experience in one of R, Python, Spark / Scala			
	Strong technical background working in a Big Data ecosystem			
	Familiarity with software development tools and practices (Git, code review, etc.)			
Communication / Presentation	Communicate complex concepts and the results of the analyses in a clear and effective manner			
	Strong verbal, written communication, and presentation skills			
	Experience creating compelling dashboards in a visualization tool (Tableau, Power BI, QlikView, IBM Cognos, etc.)			
Analytics use case development	Problem-solving mindset with the ability to understand business challenges			
	Industry specific knowledge			

Data science skill sets by roles

You can use this as a starting point to evaluate what skills you have right now versus what skills you need to fit the bill of the role you want to be in. This gap analysis will help to identify the key areas you need to extensively prepare for and what areas may need just a refresher.

Doing the exercise above will help you immensely if you're looking to move from one job role to the other; for example, if you want to move from a business analytics type of role to a machine learning engineer role. Additionally, you can utilize the information in the grid above to narrow down the specific areas you need to focus on before your interview.

Each of the following chapters in this book will discuss the skills sets listed above and will guide you on how you can ace the different rounds in a data science interview.

PART 2

CRACKING THE TECHNICAL ROUNDS

CHAPTER 4

MODELING AND MACHINE LEARNING QUESTIONS

If you see any of the following bullet points as part of a job description, then the insights in this chapter are going to be key for your interview:

- Develop and validate data-driven experiments, forecasting algorithms, machine learning models.
- Engage with statistical methods such as forecasting, time series, hypot testing, regression, classification, and clustering.
- Employ extensive knowledge in data mining, predictive modeling, ma learning, and text analytics.

An interview for a job role that involves the use of machine learning algorithms and predictive modeling techniques will test you on your knowledge in these areas. Expect the interview to get into the weeds of the algorithms you have worked on in the past.

In this chapter, we will walk through a few questions within the realm of modeling and machine learning. Please note this list isn't exhaustive, but it does offer a flavor of the kinds of interview questions one can expect.

I have added quite a bit of detail to the answers in this chapter so you can use this book to recap some of the key concepts before an interview. Note that you may not be required to go into that level of detail unless probed by your interviewer, but being prepared is always good!

Practice Question #1—Overfitting in predictive models

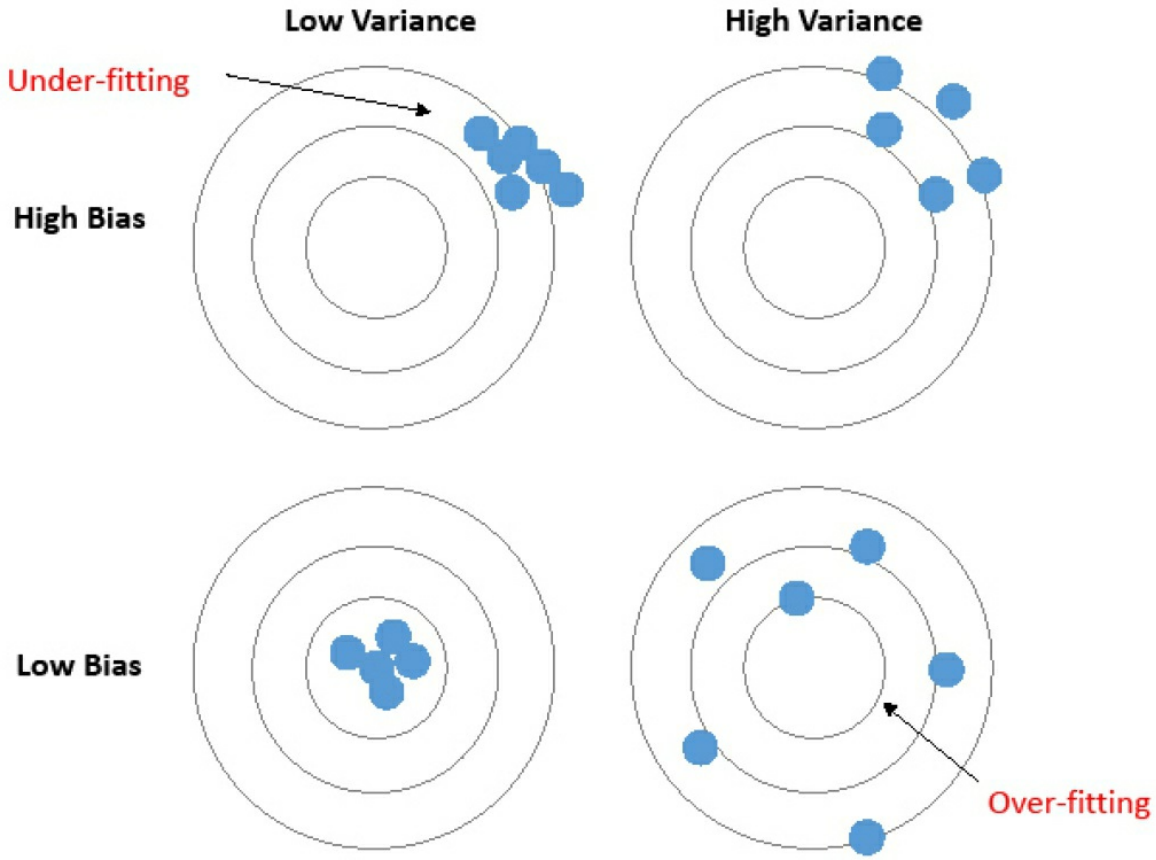
Interviewer: How can you avoid overfitting in predictive models?

Before we dive into the answer, let us look at why this is important.

Why is this important? When we build a predictive model, one of the key things to look at is the prediction error we obtain from the model built. Prediction error in a regression model is how well your model can predict the response variable. For classification models, it is a measure of how well the model classifies to the correct category.

Prediction error can be explained by bias and variance errors. Bias essentially is the difference between the forecast and the actual that we are trying to predict. Variance is the variability of the forecasted value and gives an estimate of the spread of the model data.

To understand bias and variance better, let us visualize them in the form of targets on a dart board. For example, high-variance, low-bias would be hitting darts on a dartboard at varied spots but far away from the center, whereas low-variance, high-bias would be hitting the darts in close proximity but far away from the center target.



Bias-variance trade-off

Underfitting in supervised machine learning algorithms happens when a model has high-bias, low-variance. Hence, these models are not able to capture the underlying trend in the data. This could happen due to insufficient data being available to build the model or too few features being included in the model.

Overfitting in supervised machine learning algorithms happens when a model has low-bias, high-variance. These models fit the train data too well but aren't able to predict with a high accuracy on the test data. Overfitting occurs when the model fits the noise rather than the actual trend in the data, making it incapable of predicting with good accuracy on data it hasn't seen before. These models tend to react significantly to minor fluctuations in data leading to poor accuracy.

To build a good model, we need a balance between the bias-variance trade-off.

Now that we see why these concepts are important, let's attempt to answer the original question.

Candidate: A model is said to overfit when it describes the noise in the data rather than the underlying relationship. Based on my experience, overfitting can be avoided in the following ways:

- **Reduce model complexity**: Overfitting can be caused by having too r features in the model, which makes it more complex than required. Cu down on the number of features by including only those that truly influ the model's prediction can prevent overfitting. A simple approach to do is to run a forward stepwise regression. In this method, a predictor w added to the model which has the smallest p-value, and this will be across multiple iterations until the stopping criterion is met. Albeit an solution, this approach could miss the "best" model.

- **Regularization**: When making use of regularization techniques shrink/regularize the coefficient estimates toward zero. An added bene this technique is that it avoids modeling of any noise in the data whicl prevent overfitting. Ridge and Lasso techniques can be used to reduc number of predictors used in the model when you have a scenar

described in the question.

Interviewer: Yes, those two approaches make sense. Can you explain a little more about how Ridge and Lasso regression work?

Candidate: Sure, I have used regularized regression in a few of my projects in the past. At a high level, in ridge regression, the loss function or the residual sum of square errors is minimized by adding a shrinkage quantity. Ridge regression makes use of lamda, which acts as a tuning parameter for the model. As the value of lamda increases, the coefficient estimates tend toward zero.

Lasso is another regularization method and has the capability of "selecting" variables by penalizing the high value coefficients. Lasso in comparison to ridge will shrink the coefficient values to a zero-value, allowing the model to select a small number of variables as the final predictors in the model, whereas ridge regression will shrink them close to zero but not necessarily make them zero. In other words, the ridge regression model will include almost all predictors whereas lasso will perform feature selection.

Interviewer: That's great. Are there any other ways you can think of using which you can prevent overfitting in models?

Candidate: Yes, cross-validation is another preventive measure against overfitting and an important one at that. Using cross-validation, you can generate multiple smaller train-test splits. For example, in k-fold cross-validation, you will use k-1 folds to train the model while the remaining fold (or holdout fold) will be used as a test set. When training on k-1 folds, cross-validation can be used to tune the parameters.

The rationale behind using cross-validation is that we are varying what we train on and what we test on. By generalizing well, we can avoid potential overfitting.

Interviewer: That sounds good. Thank you very much.

***Differentiator*:** While this sufficiently answers the interviewer's question, a good way to distinguish yourself from the other candidates is to layer in an example of when you have dealt with such a scenario in the past.

The key when giving an example is to keep it concise to demonstrate real-life experience. If the interviewer wishes to go into details, they may ask you further questions on the topic.

Extra questions for practice on this topic:

- What is the difference between forward stepwise and backward step regression?
- Why does Lasso tend to shrink estimates to zero whereas Ridge shrinks close to zero but not zero?

Practice Question #2—Determine number of clusters in k-means

Interviewer: How will you define the number of clusters in a clustering algorithm?

Why is this important? Clustering is a technique that's used to group together objects with similar characteristics.

In their simplest form, *clusters* are sets of data points that share similar attributes, and *clustering algorithms* are the methods that group these data points into different clusters based on their similarities. You'll see clustering algorithms used for disease classification in medical science, but you'll also see them used for customer classification in marketing research and for environmental health risk assessment in environmental engineering.[7]

The purpose of clustering and classification algorithms is to make sense of and extract value from large sets of structured and unstructured data. If you're working with huge volumes of unstructured data, it only makes sense to try to partition the data into some sort of logical groupings before attempting to analyze it.[8]

Before beginning to answer this question, let's understand how clustering algorithms work.

How does this work? One of the popular algorithms used for clustering is the k-means algorithm. K-means uses distance as a measure of similarity.

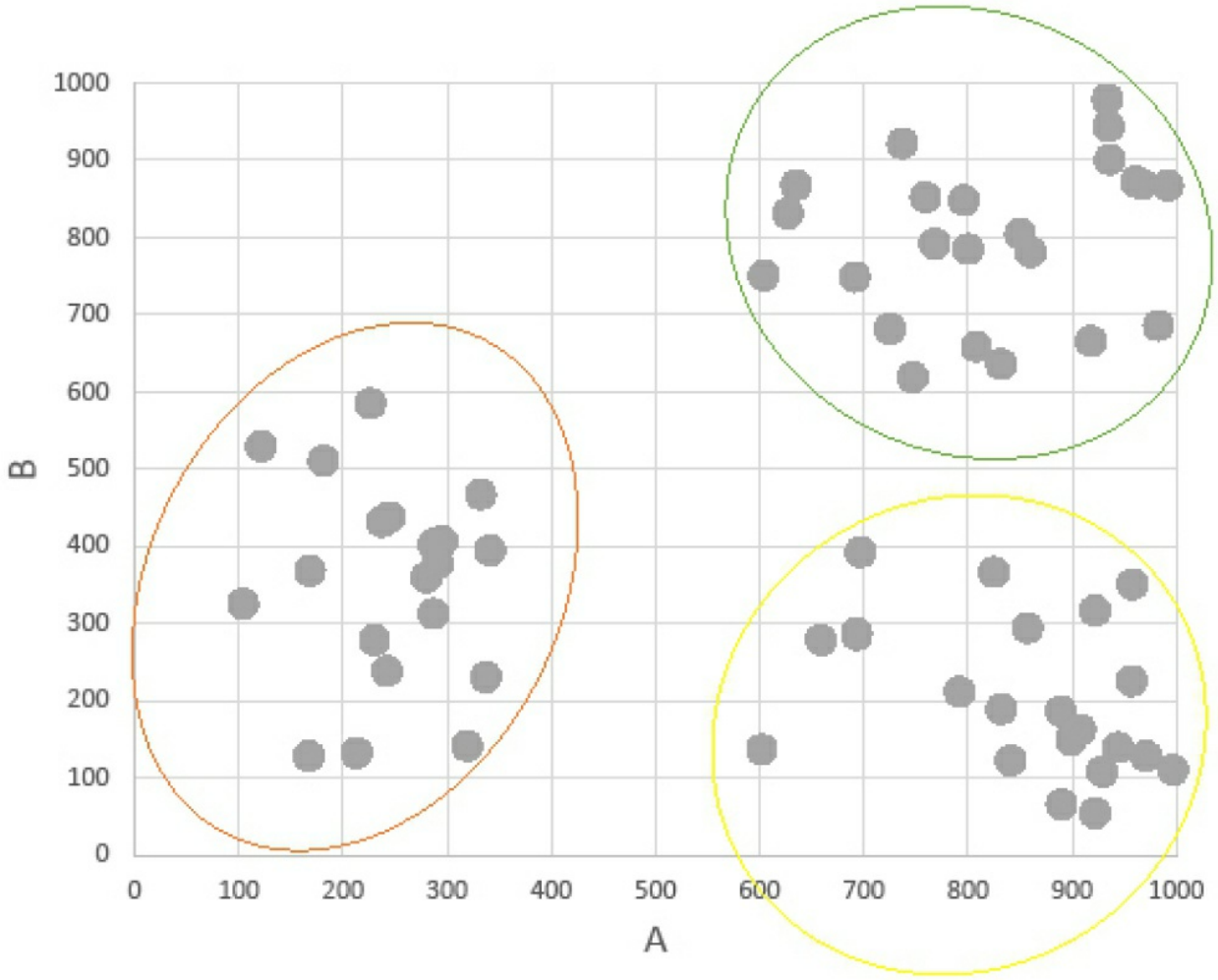
Let's say you have a data set with two columns, column A and column B.

Now, you would like to cluster this data set based on similarities observed in the two columns. And you would like to use k-means algorithm to create the clusters. K-means needs you to specify the number of clusters you want at the onset. Suppose for this example, you think three would be a good start. Once you have set this hyperparameter, here's how the algorithm would then work:

- The algorithm will randomly assign any three points as the cluster center
- Then, it will compute the distance between each observation and the cl centers.
- After this, it will assign each observation to a cluster that is closest to value.

- It will then calculate the mean of the new clusters and keep doing this there is no change in the clusters.

The clusters, for our example, can be visualized as below:



Scatterplot of Variable A, B

Let's now dive into the specifics of the answer for the question asked by the interviewer.

Candidate: Algorithms like k-means that are used popularly for clustering need the user to input the number of clusters to be built. Given this requirement, you must determine the optimal number of clusters to be generated. A commonly used method to determine this is using the elbow curve.

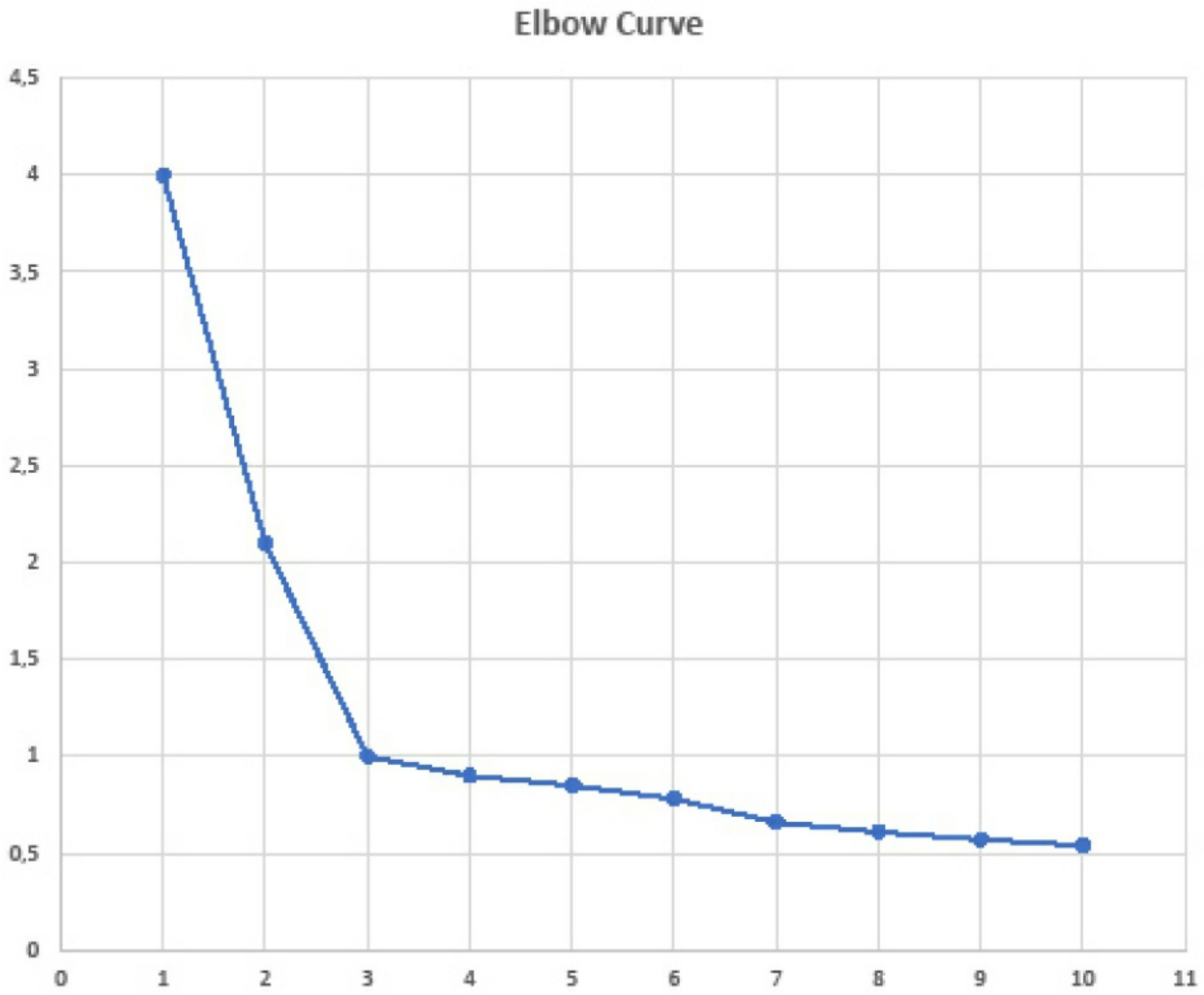
Interviewer: How do you recommend using the elbow curve to determine the number of clusters?

Candidate: To answer this, let me recap a bit on how k-means work. K-means work toward reducing the intra-cluster variation; in other words, the within-cluster sum of square (WSS) is minimized. This WSS statistic is plotted

against the number of clusters, forming the elbow curve. Typically, one can see a huge drop in the WSS after X number of clusters. The point where the drop is observed is to be taken as the optimal number.

For reference: The graph below shows a sample elbow curve that plots the WSS* metric against the number of clusters. For this particular example, you can see that the graph changes drastically at cluster size = 3. The curve becomes linear in shape, and that's the optimal point in the curve that determines the number of clusters to be used for that data set.

*WSS metric: average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used.



Elbow curve

Differentiator: Similar to the first question in this chapter, I would

recommend layering in a short example of when you have done this in the past. For this question, you can talk about a clustering project you worked on and how many segments you ended up with.

While looking at the elbow curve to determine the number of clusters is a good way to go about it, a validation check toward the end is to see whether the clusters formed are intuitive or not. You can tie them back to the problem you're trying to solve and see if the clusters formed are sufficient to answer the business problem. Depending on the objective for clustering, you may find yourself iterating to find the optimal number of clusters that answers the business question sufficiently and is statistically sound.

Extra questions for practice on this topic:

- What are some other clustering techniques? E.g., hierarchical cluste density-based clustering.
- Which clustering algorithm is typically sensitive to outliers?
- Can clustering be used to improve the accuracy of a linear regression m If so, how?
- Can categorical variables be used in k-means clustering?

- If there is mix of categorical and continuous variables, which clust technique would you use?

Practice Question #3—Favorite algorithm

*Interviewer***:** What is your favorite algorithm and why?

Why is this important? In a conversation with Andy Fox, a senior director at Opex Analytics, I asked if he had a go-to question when he conducts interviews for data scientists. And this is the question Andy shared with me.

This is an interesting one because it's open-ended in nature and allows the candidate to lead the interviewer to an answer they have a good deal of knowledge about, essentially allowing them to show off their modeling skills! On the other hand, it allows the interviewer to test the candidate's depth of knowledge in a particular area.

Pro tip: For a broad question such as this, you should frame your answer so it

covers a few key aspects. Here is an outline as a reference point:

- My favorite algorithm is…
- This model can be used for…[*List use cases you know for this model, classification or regression*]
- This algorithm works as follows…[*Give a brief overview of how this r works behind the scenes*]
- This is my favorite algorithm because…[*List a couple of reasons why you this model is good*]

- I used this algorithm to…[*Describe a project where and how you used it*]

For practice, let's walk through an example of how this question can be answered. For simplicity and ease of understanding, let's use a popular model.

Candidate: My favorite algorithm is Random Forest. It can be used for classification and regression use cases.

Random Forest consists of a large number of decision trees and works as an ensemble technique. Random Forests works on the principle of bagging, wherein each decision tree is built on a sample of the training data set with replacement. The results from these multiple decision trees are then aggregated to come up with the final forecast. For the purposes of classification, the mode of all the predictions is used, and for regression, the mean of all predictions is deemed as the final output.

This is my favorite algorithm for a few reasons:

- Random Forests work well with both categorical and numerical data scaling or transformation of variables is usually necessary.
- Random Forests implicitly perform feature selection and gen uncorrelated decision trees. It does this by choosing a random set of fea to build each decision tree. This also makes it a great model when you ha work with a high number of features in the data.
- Random Forests are not influenced by outliers to a fair degree. It does th binning the variables.
- Random Forests can handle linear and non-linear relationships well.

- Random Forests generally provide a high accuracy and balance the variance trade off well. Since the model's principle is to average the r across the multiple decision trees it builds, it averages the variance as wel

Differentiator: This answer covers all main aspects of the algorithm that you like and leaves room for more conversation on this topic. A good differentiator for such a question would be to follow up with two more points:

- Describe two to three drawbacks of this algorithm and describe a situ where you wouldn't recommend using it.

- Give a quick overview on an algorithm that performs better than the on chose to describe, allowing the interviewer to know that you are resear and exploring further.

For our example, you can follow up and conclude your answer in the following way:

- While Random Forests are great in a number of applications, there are ce places where they may not be an ideal choice. Here are a couple reasons this could be so:

  – Random Forests aren't easily interpretable. Although they provide featu importance, they do not provide complete visibility into the coefficien like a linear regression does.
  – Random Forests can be computationally intensive for large datasets.

- Another algorithm that I have been working on recently that I quite l XGBoost. XGBoost improves upon the capabilities that Random Fores by making use of the gradient descent framework. It also has the abil build trees in parallel and optimizes hardware as it does so. XGBoost ha in-built capability to penalize complex models by using regulariz techniques. It also comes with in-built cross validation that can be us determine the number of boosting iterations required in a run.

Additionally, choice of the algorithm largely depends on its use case and data availability. While the question we discussed above assumes you can answer

for any use case without data constraints, feel free to ask your interviewer if they are looking for a specific use case for you to address.

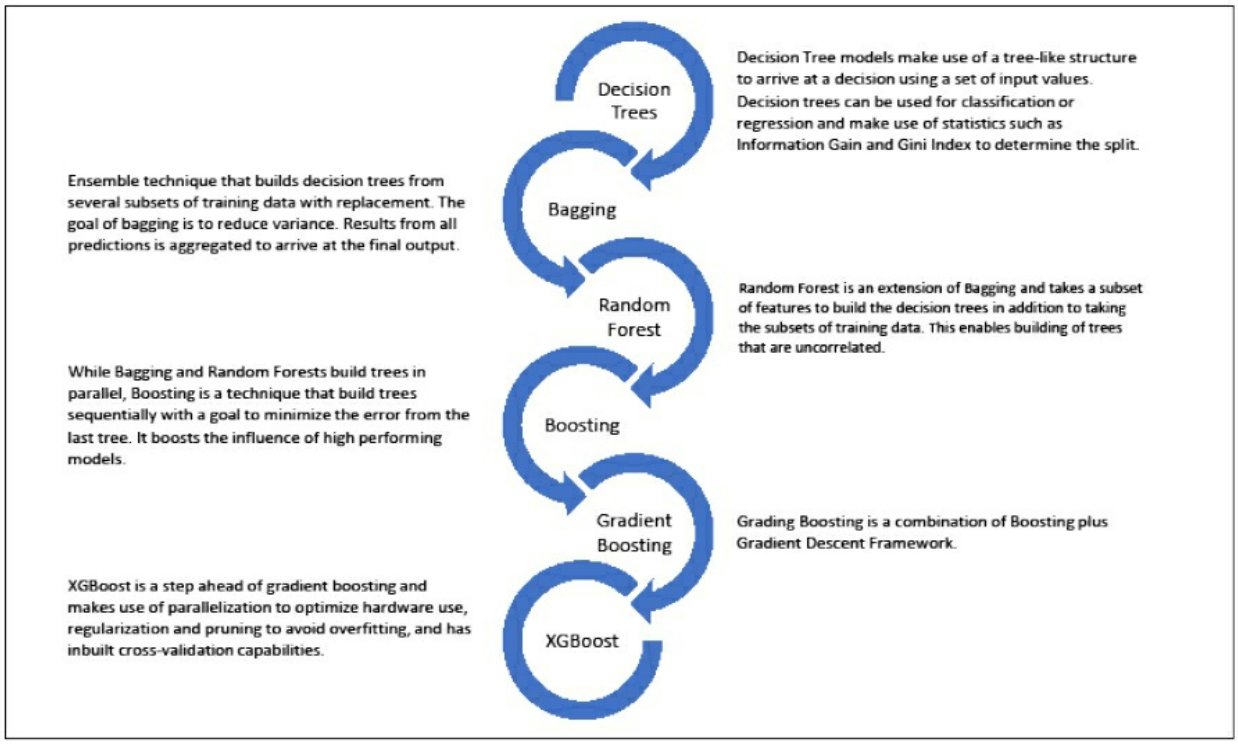
In the sample answer above, we focused on a supervised learning algorithm, feel free to use an unsupervised learning algorithm if your expertise lies in that area.

Extra questions for practice on this topic:

- How do models like Random Forest determine feature importance?
- What are the key hyper parameters to specify when building a Rar Forest?
- How can overfitting be avoided in Random Forests?

- What is your favorite deep learning algorithm?

If you have worked on tree-based models in the past, expect the interviewer to go into detail about how they work, to suggest advantages and disadvantages of using them, and to discuss use cases of where you have used them in the past. For a handy reference, here's a cheat sheet I like to refer to when revising my knowledge of tree-based models![9]



Practice Question #4—Evaluate performance of a predictive model

Interviewer: How do you evaluate performance of a predictive model?

Why is this important? Once you have built a predictive model, you need a way to evaluate whether it is performing at the level you expect. The success criteria of a predictive model are driven by what the model is going to be used for. For example, if the goal of the business is to forecast demand for a certain product, then metrics relating to accuracy can be used to evaluate performance.

However, at the same time, certain metrics can be deceiving. For example, let's say you have a model that detects fraud, and fraud is expected to happen only 1 percent of the time. Such a model may have a high accuracy, say up to 99 percent, but the model may fail to detect any fraud cases at all. In this scenario, looking at the overall accuracy metric in an isolated fashion will not be the correct choice. For this particular case, it may be useful to look at the true positive rate and false positive rate produced by the model.

Choosing the correct set of metrics to evaluate the model against is key.

For this question, let's look at the metrics available for the supervised learning models.

Quick Recap: Supervised learning is where the data is "labeled" when training the model. Examples of supervised learning are classification and regression models. Unsupervised learning is where the data is not "labeled" or the output values are not known. Examples of unsupervised learning are clustering and principal component analysis.

Candidate: To evaluate the success of a predictive model, we look at a few key metrics, and this differs by the type of the model, e.g., classification or regression. Let's start with the metrics used for classification.

Classification:

- **Confusion Matrix**: A confusion matrix is popularly used to eva performance of classification models. I have used overall accuracy, sensit and specificity metrics when I have built classification models in the past

For reference: Here's a quick recap on the formulae used in the confusion matrix.

- **Overall accuracy**: Proportion of number of predictions that were cor classified.
- **Positive predictive value or precision**: Proportion of positive cases that correctly classified.
- **Negative predictive value**: Proportion of negative cases that were cor classified.
- **Sensitivity or recall**: Proportion of actual positive cases that were cor classified.

- **Specificity**: Proportion of actual negative cases that were correctly classif

Interviewer: In what instances do you recommend using one metric over the other for classification models?

Candidate: So, depending on the use case of the classification problem, one of these metrics will be more suited than the other. For example, a pharmaceutical company will be concerned with minimal incorrect positive diagnosis. Hence, the metric to evaluate for this use case will be specificity (aiming for a high specificity value). On the other hand, for a model that predicts loan default rates, the model should capture the slightest chance of a default, hence we need the model to be highly sensitive.

		Target		
		Positive	**Negative**	
Actuals	**Positive**	True Positive (TP)	False Negative (FN) ***Type II Error***	Sensitivity / Recall TP / (TP + FN)
	Negative	False Positive (FP) ***Type I Error***	True Negative (TN)	Specificty TN / (TN + FP)
		Precision TP / (TP + FP)	Negative Predicted Value TN / (TN + FN)	Overall Accuracy (TP + TN) / (TP + TN + FP + FN)

Confusion matrix cheat sheet

Cheat sheet: Use this diagram to visually remember the different metrics.

Interviewer: Thank you for that example. Are there any other metrics you

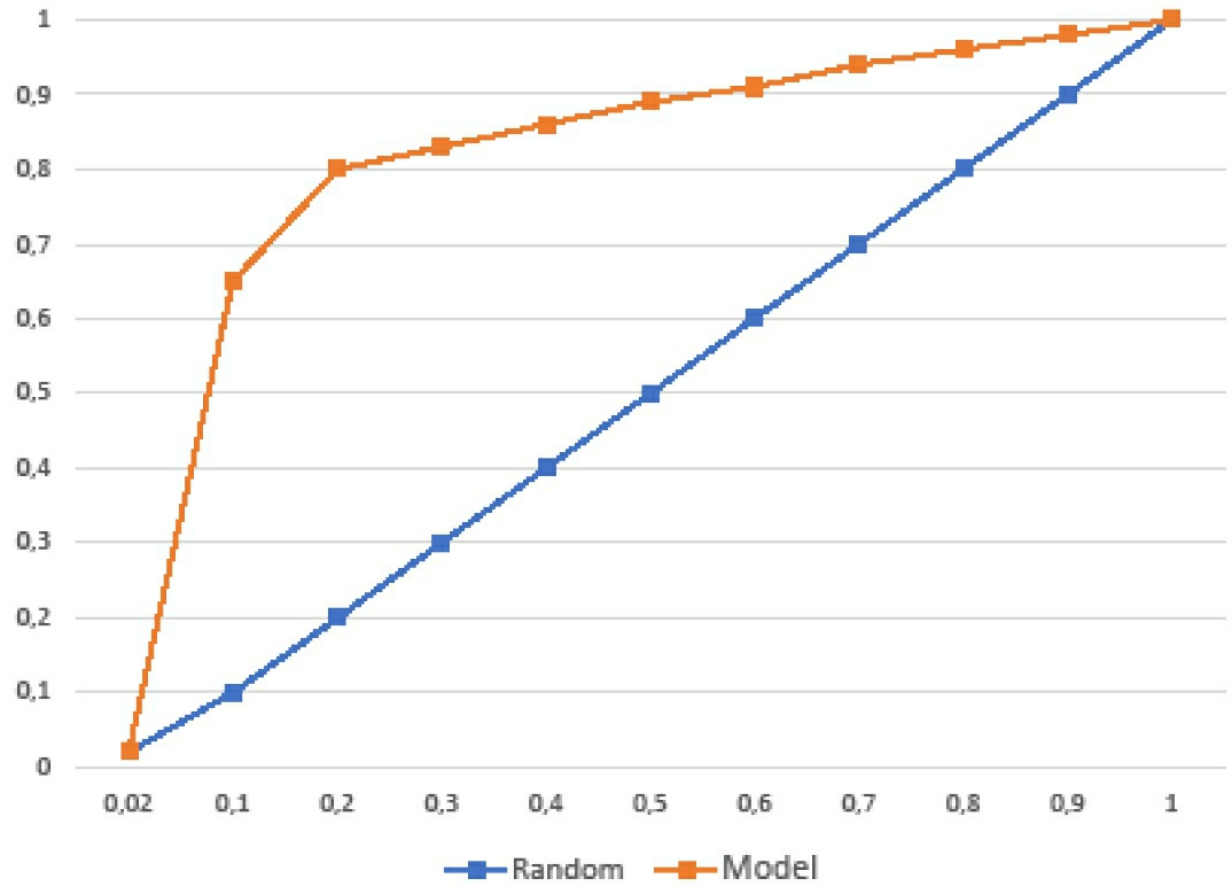
look at for classification models?

Candidate: Yes, I have used F1 score and area under the ROC curve as well in the past.

- **F1 score:** For a use case that requires a balance between precision and r the F1 score is used. F1 score is the harmonic mean of precision and values. It punishes extreme values more to keep the balance.

$$F_1 = \frac{recall^{-1} + precision^{-1}}{2}$$

- **Area under the ROC curve:** The ROC chart shows 1-Specificity (or Positive Rate) on the x-axis and Sensitivity (or True Positive Rate) on t axis. Area under the curve metric is often used to measure the quality c classification model and visually shows the trade-off between sensitivity specificity. If the model is good, the chart quickly climbs up in value on t axis. Typically, a value between 0.5 and 1 for this metric is considered go



Example of an ROC curve

Candidate: That being said, multiple metrics can be used to evaluate a classification model. The choice should be made based on the end use case of the model.

Interviewer: That sounds good. Can you now walk me through the metrics that can be used to evaluate the performance of a regression model?

Candidate: Often for regression models, I have used root mean squared error, mean absolute error, and R-square.

Interviewer: Can you explain the difference between R-square and Adj R-square metrics?

Candidate: R-square may keep increasing when a greater number of features are added in the model even though the model isn't improving, whereas Adj R-square will likely remain the same. Adj R-square penalizes for adding more variables in the model, and Adj R-square will always be less than or equal to R-square.

For reference: If asked to go into details, here's a recap of these metrics.

- **Root mean squared error (RMSE):** RMSE is a measure of how spread ou residuals are, where residuals are the difference between actuals predicted. RMSE penalizes the higher prediction errors more than the M

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

- **Mean absolute error (MAE):** MAE is the average of the absolute diffe between the predicted values and observed value.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_{0j}|$$

- **R-square/Adjusted R-square:** R-squared is a goodness of fit measure use linear regression models. R-square and Adjusted R-square give insight how the independent variables explain the variability in the depe variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

$${R_{adj}}^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

Note: n is the total number of observations and k is the number of predictors.

Differentiator: While this answer expertly covers the different metrics that can be used to evaluate the model, a few other factors can be taken into consideration for an added differentiation.

- For a business scenario that values transparency in the model, you must *if the model is directionally correct*. For example, when building a l regression model, check whether the sign of the coefficients is directio intuitive.

- Check if the model is robust and ensure model is not overfitting.

Extra questions for practice:

- How will you deal with an imbalanced data set when building a pred model?
- What metrics will you evaluate for a classification model built on imbala data?
- What metric will you look at to evaluate whether a model is overfitting?

- What are the limitations of using R-square to evaluate model performanc

Practice Question #5—Determine parameter values in ARIMA

Interviewer: How can you determine the parameter values in a time series ARIMA model?

Why is this important? Real-world applications of time series analysis are endless. Accurate forecasting enables businesses to plan and execute better. For example, time series analysis can be used for stock market analysis, inventory projections, and economic forecasting, to name a few.

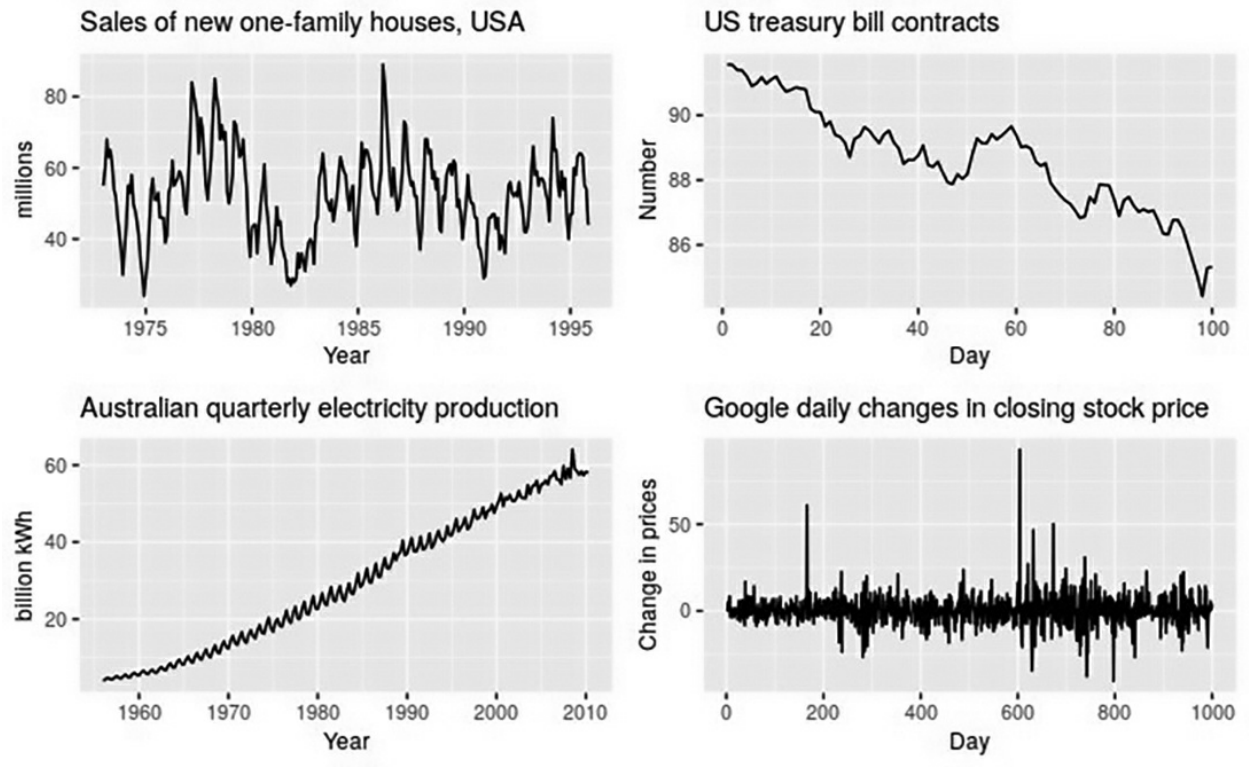
Before we jump into the details of time series modeling, let's understand the basics of time series data. Time series data is essentially a set of observations arranged in a chronological order. This data can be primarily broken down into four main components:

- **Trend**: The trend component of the time series data shows the ge tendency of the data to increase or decrease with time. For example, pri commodities such as gold have a tendency to increase over time.
- **Seasonality**: Any systematic and calendar-driven effects observed in the series data make up the seasonal component. For example, turkey sale expected to shoot up around Thanksgiving.
- **Cyclicity**: Patterns that depict an upward or downward movement arou given trend is the cyclical component. For example, the stock markets h tendency to cycle between highs and lows, but the duration of time bet these fluctuations isn't known like it is in seasonality.

- **Randomness**: This component of a time series is unpredictable and ran

The random component of a time series is unlikely to be repeated.

Let's look at an example that shows different combinations of the above components.



Four examples of time series showing different patterns

1. The monthly housing sales (top left) show strong seasonality within year, as well as some strong cyclic behavior with a period of about six t years. We can see no apparent trend in the data over this period.
2. The US treasury bill contracts (top right) show results from the Ch market for one hundred consecutive trading days in 1981. Here we se seasonality, but an obvious downward trend. Possibly, if we had a l longer series, we would see that this downward trend is actually part of a cycle, but when viewed over only one hundred days it appears to be a trer
3. The Australian quarterly electricity production (bottom left) shows a st increasing trend with strong seasonality. We see no evidence of any c behavior here.
4. The daily change in the Google closing stock price (bottom right) ha trend, seasonality, or cyclic behavior. We can see random fluctuations v do not appear to be very predictable and no strong patterns that would

with developing a forecasting model.[10]

With a brief overview of what a time series data set consists of, we can jump into understanding ARIMA models. ARIMA stands for autoregressive integrated moving average and is used for time series analysis. ARIMA models have three key parameters we need to tune when building the model:

- **P**: Parameter P refers to the number of periods to lag for and correspon the autoregressive (AR) part of the ARIMA model. A purely autoregre model is essentially a linear regression where the dependent variables lagged variables.
- **D**: Parameter D stands for differencing. Data fed to an ARIMA mo required to be stationary, i.e., a series without any trend or seasonalit remove the trend/seasonality, the variables are differenced. D refers t number of differencing that needs to be applied to the data to ma stationary.

- **Q**: Parameter Q denotes the lag of the error component. The component is the one that cannot be explained and is random at best. corresponds to the moving average (MA) part of the ARIMA model a used to smooth out any random jumps observed in the data.

The value of these parameters is decided depending on what is observed in the data.

Candidate: Typically, an ARIMA model consists of three key parameters (P, D, Q). The order of the autoregressive (AR) and moving average (MA) components of an ARIMA model, or P and Q respectively, can be identified by ACF and PACF plots.

Interviewer: Can you tell me a little more about the ACF and PACF plots and how they can be used for determining the P and Q values?

Candidate: Sure, ACF is an auto-correlation plot that gives us the correlation between the predictor variable and its lagged values.

PACF is a partial auto-correlation plot and finds correlation with the residuals. The partial autocorrelation at lag Q is the correlation that results

after removing the effect of any correlations due to the terms at shorter lag.[11]

Here's the idea behind using these plots: When the correlation is plotted against the lagged values (ACF plot), one would expect the plot to tail off as the effect weakens. When tailing off is observed in an ACF plot, we can say that an AR term is needed in the ARIMA model and the value for the term AR term (P) can be determined at the point where we observe it to tail off.

Similarly, if we observe a tailing off in the PACF plot that plots the correlation of residuals against the lags, we can say that a moving average component is required in the model and the value of the parameter can be at the lag that it tails off. Alternatively, you would see that the ACF plot cuts off after a certain lag, indicating a moving average process the cuts off after a lag Q.

Interviewer: Got it, and how would you determine the value of parameter D?

Candidate: Parameter D is used to indicate the number of differencing to be used to make the time series data stationary. Since the data is said to be stationary when it does not have any trend or seasonality component (i.e., the mean and the variance are constant), plots of the differenced variables can be inspected to find out at what value stationarity is achieved.

Interviewer: That makes sense, thank you.

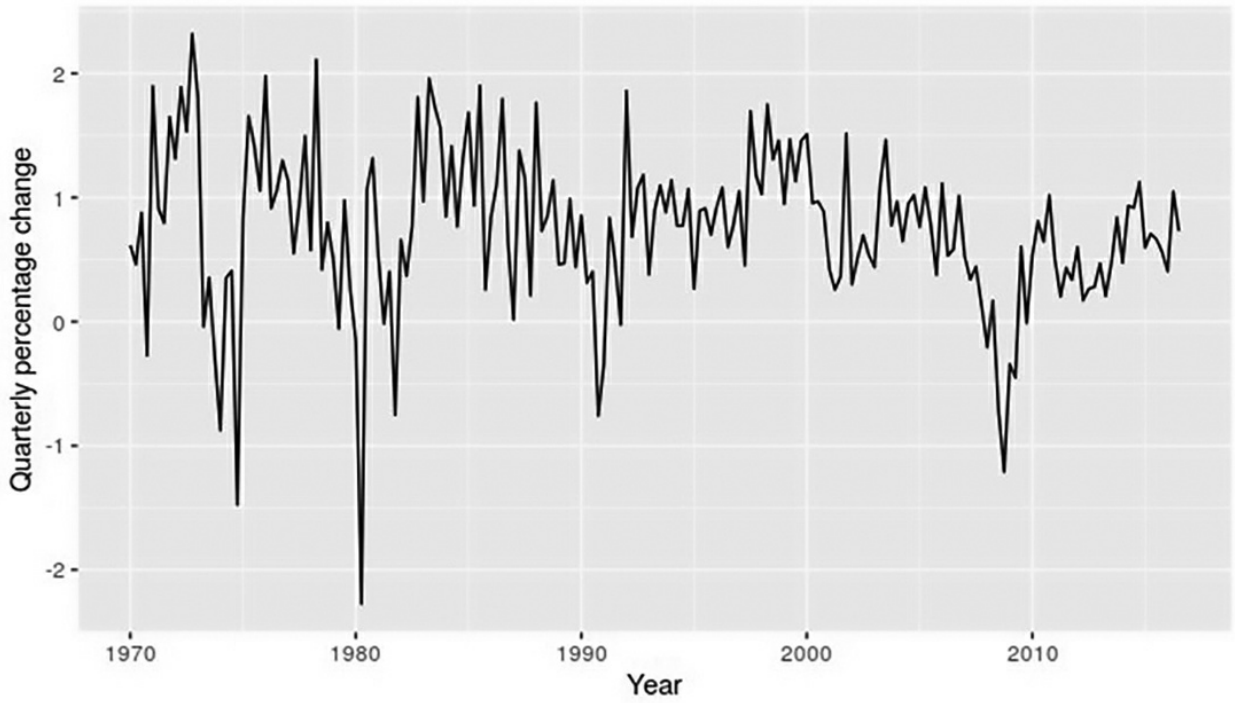
Time series models are widely used. To help you prepare well for interviews, here are some more details about time series models.

Cheat sheet: Here's a cheat sheet for how to use ACF and PACF values to determine P, Q parameters.

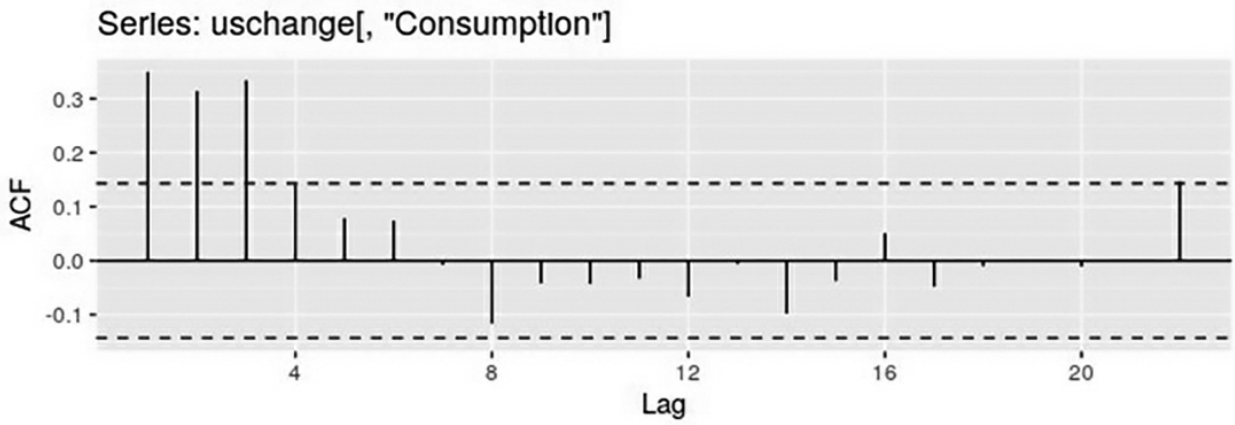
	AR (p)	MA (q)	ARMA (p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

Determine P and Q using ACF and PACF plots

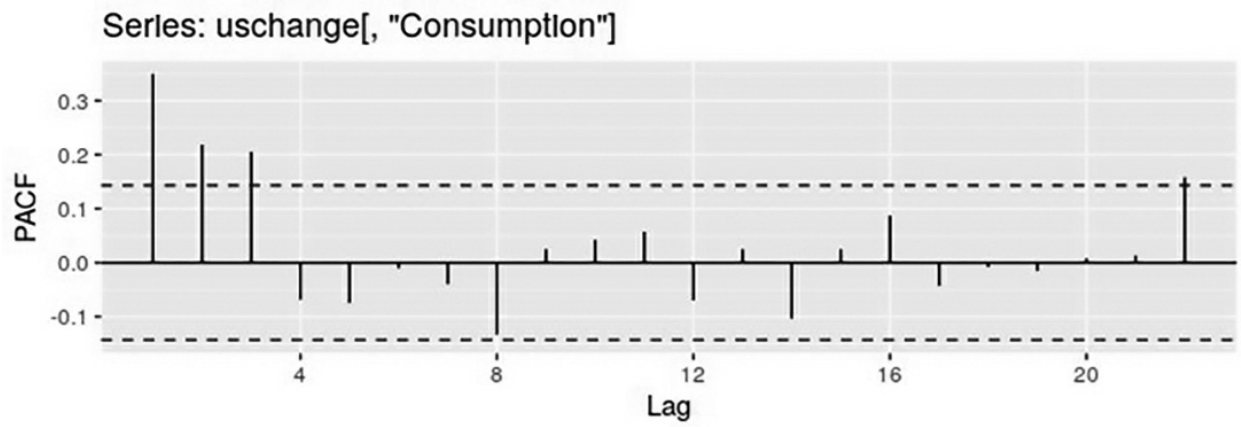
Quick Recap: Here are some examples of ACF and PACF plots for quarterly percentage change in US consumption expenditure data:



Quarterly percentage change in US consumption expenditure



ACF of quarterly percentage change in US consumption



PACF of quarterly percentage change in US consumption

In the plots above, we see three spikes in the ACF followed by an almost significant spike at lag 4. In the PACF, we see three significant spikes and then no significant spikes thereafter (apart from one just outside the bounds at lag twenty-two). We can ignore one significant spike in each plot if it is just outside the limits, and not in the first few lags. After all, the probability of a spike being significant by chance is about one in twenty, and we are plotting twenty-two spikes in each plot. The pattern in the first three spikes is what we would expect from an ARIMA (3,0,0), as the PACF tends to decrease. So, in this case, the ACF and PACF lead us to think an ARIMA (3,0,0) model might be appropriate.[12]

Extra questions for practice on this topic:

- Why should a time series be transformed into a stationary series b building an ARIMA model?
- How do you test for stationarity of time series data?

- Can you explain how ARIMA and ARIMAX models are different?

You can expect to encounter the types of questions we covered in this chapter in a data science technical screen. Typically, in these screens, the company wants to evaluate you on fundamental data science skills before moving you on to the next rounds. To adequately prepare for such questions, go over the basic machine learning techniques and ensure your conceptual understanding of them is sound. Once you have done that, prepare extensively on the techniques you have used in past projects so you are well

prepped for situations in which the interviewer may want to do a deep dive.

---

7 “The Importance of Clustering and Classification in Data Science," Dummies.com, accessed May 17, 2020.

8 Ibid.

9 Rob J. Hyndman and George Athanasopoulos, *Forecasting: principles and practice*, 2nd edition (Melbourne: OTexts, 2018).

10 Inspired from: Vishal Morde, “XGBoost Algorithm: Long May She Reign!,” Medium.com, accessed May 20, 2020.

11 Andrew V. Metcalfe and Paul S.P. Cowpertwait, *Introductory Time Series with R* (New York: Springer, 2009), 81.

12 Rob J. Hyndman and George Athanasopoulos, *Forecasting: principles and practice*, 2nd edition (Melbourne: OTexts, 2018).

CHAPTER 5

PROBABILITY, STATISTICS, AND EXPERIMENTAL DESIGN

Probability and statistics form the foundation of any data science work. By and large, any analysis that you do will have these two topics at its base and for this reason, the probability that an interviewer may test your understanding of these subjects is high (pun intended). Think of these topics as prerequisites for data science. To showcase your understanding of the fundamentals, you must cover the basics before an interview.

In this chapter, I will walk you through a few key concepts and demonstrate how an interviewer can probe into your comprehension of these subjects. This chapter is divided into three sections:

1. Probability
2. Statistics
3. Experimental Design

Probability

Our goal for this section is to recap key concepts in probability and work through some practice problems so you have in-depth clarity on the fundamentals. While the subject of probability in itself is very vast and worthy of its own book, in this chapter we will look at a few concepts that will act as a refresher as you prepare for your interviews.

Let's dive right in.

What is Probability? Probability is the likelihood of the occurrence of an event. An event can be anything, such as drawing a card from a deck or

tossing a coin.

There are varying degrees of probability of whether an event can occur. If an event cannot occur at all, its probability is zero. If an event is certain to occur, its probability is one. Hence, the possibility of an event occurrence is always between zero and one.

The probability of an event is the number of ways the event can occur divided by the total number of possible outcomes.

$$P(A) = \frac{No.of\ ways\ A\ can\ occur}{Total\ number\ of\ possible\ outcomes}$$

E.g., Probability of getting a head on the toss of a coin = 1/2

Compound Events: When we use the term compound event in math, we are referring to the probability of two or more events happening at the same time.[13]

Compound probability is a mathematical term relating to the likeliness of two independent events occurring. Compound probability is equal to the probability of the first event multiplied by the probability of the second event.[14]

$$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$$

P(A or B) is the probability of the occurrence of at least one of the events.

P(A and B) is the probability of the occurrence of both A and B at the same time.

Mutually Exclusive Events: When two events cannot occur at the same time, they are called mutually exclusive events.

For mutually exclusive events A and B:

$$P(A\ and\ B) = 0$$

Independent Events: An event is said to be independent when the outcome of it is not affected by the outcome of any other event. For example, if you toss a fair coin, the outcome of the first toss does not affect the outcome of the second toss.

Practice Questions

1. What is the probability of getting a 1 or a 6 on the roll of a die?

```
P(1 or 6) = P(1) + P(6) - P(1 and 6)
 = 1/6 + 1/6 - 0
 = 1/3.
```

2. What is the probability of getting a black card or a king from a deck of ca

```
P(Black Card or King) = P(Black Card) + P(King) - P(Black Card and
King)

 P(Black Card) = 26/52
 P(King) = 4/52
 P(Black Card and King) = 2/52

P(Black Card or King) = (26/52) + (4/52) - (2/52)
 = 28/52
 = 7/13
```

3. A box contains 3 black, 2 blue, and 5 red balls. What is the probabili picking 2 black and 1 red ball with replacement?

Hint: Since the balls are picked with replacement, the event of picking a single ball is an independent event.

```
P(2 Black and 1 Red) = (3/10) *( 3/10) * (5/10) = 9/200
```

Let's now look at conditional probability.

Conditional Probability: Conditional probability is used when you have to calculate the probability of an event given another event has occurred.[15]

Conditional probability is often denoted as P(A|B), read as P(A given B). The formula for conditional probability is: [16]

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Practice Questions

4. A math teacher gave her class two tests. 25 percent of the class passed tests and 42 percent of the class passed the first test. What is the proba that of those who passed the first test also passed the second test?[17]

```
P(Second | First) = P(First and Second) / P(First)
 = 0.25/0.42
 = 0.6
```

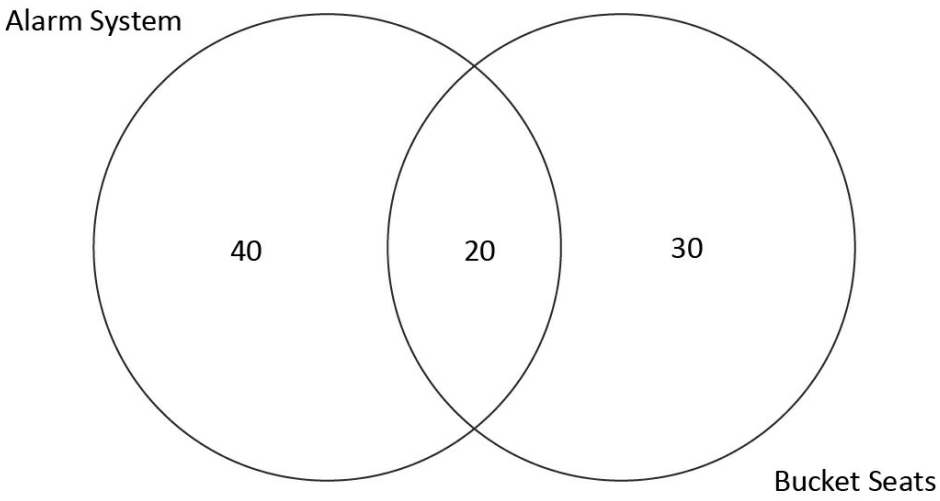
5. In a group of 100 sports car buyers, 40 bought alarm systems, 30 purcl bucket seats, and 20 purchased an alarm system and bucket seats. If buyer chosen at random bought an alarm system, what is the probability also bought bucket seats? 18

Consider B as the event of buying bucket seats.

Consider A as the event of buying alarm systems.

```
P(B|A) = P(A∩B) / P(A) = 0.2 / 0.4 = 0.5.
```

Pro tip: For questions like these, make use of a Venn diagram to quickly visualize the numbers. This approach is very useful when you're asked a question that involves the intersection of three events. Venn diagrams help simplify the question and lead to an answer faster.



Venn diagram for alarm system and bucket seats

Let us now look at Bayes' theorem, which is pivotal in probability theory and statistics.

Bayes: Bayes' theorem is central to many machine learning techniques today. For an interview where you are tested on probability, it is highly likely you may be asked a question related to Bayes.

Let's say we are trying to find out the probability of being asked an interview question based on Bayes. For this, you have additional evidence that the interview process will consist of a dedicated round on the topic of probability.

Before we jump into the actual calculation for our example, let's look at how Bayes' rule is typically depicted.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

A: the event we want the probability of

B: the evidence that A is related in some way

P(A|B) is called the **posterior.** In our example, this would be the "probability of being asked a Bayes question given that the interview process contains a dedicated round on the topic of probability."

P(B|A) is called the **likelihood.** In our example, this would be the "probability of having a dedicated interview round on the topic of probability given that there was an interview question on Bayes."

P(A) is called the **prior**. In our example, this would be the "probability of being asked a Bayes interview question."

P(B) is called the **marginal likelihood**. In our example, this would be the "probability of having a dedicated interview round on the topic of probability."

Let us now plug in some numbers so we can use the equation to come to our answer.

```
P(being asked a Bayes interview question) = 0.1
P(having a dedicated interview round on the topic of probability) =
0.4
P(having a dedicated interview round on the topic of probability
given that there was an interview question on Bayes) = 0.6
```

Therefore,

```
P(being asked a Bayes question given that the interview process
contains a dedicated round on the topic of probability)
 = (0.6*0.1)/(0.4)
 = 3/20
```

Using Bayes theorem, we are able to use more information to update our probabilities to get a better estimate.

Solving problems using probability trees and tables: Probability trees are an

easy way to visualize a probability problem and an excellent tool to use during your interview. They allow you to show your thought process to your interviewer on paper and brings them along on the journey when you're solving the problem they've asked.

MIT's Introduction to Probability and Statistics class has a very good example that showcases the use of trees and tables. Let's look at it here.

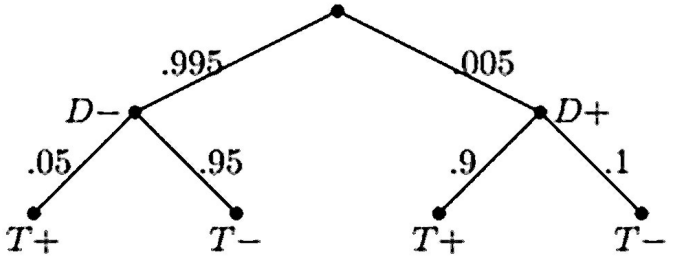
PRACTICE QUESTION

6. Consider a routine screening test for a disease. Suppose the frequency c disease in the population (base rate) is 0.5 percent. The test is highly acc with a 5 percent false positive rate and a 10 percent false negative rate. take the test and it comes back positive. What is the probability that you the disease?

Answer: We will do the computation three times: using trees, tables, and symbols. We'll use the following notation for the relevant events:
D+ = you have the disease
D− = you do not have the disease
T+ = you tested positive
T− = you tested negative

We are given P(D+) = .005 and therefore, P(D−) = .995. The false positive and false negative rates are (by definition) conditional probabilities. P(false positive) = P(T + |D−) = .05 and P(false negative) = P(T − |D+) = .1. The complementary probabilities are known as the true negative and true positive rates: P(T − |D−) = 1 − P(T + |D−) = .95 and P(T + |D+) = 1 − P(T − |D+) = .9.

Trees: All of these probabilities can be displayed quite nicely in a tree.



Probability tree

The question asks for the probability that you have the disease given that you tested positive; i.e., what is the value of P(D+|T+)? We aren't given this value, but we do know P(T +|D+), so we can use Bayes' theorem.

$$P(D+|T+) = \frac{P(T+|D+) \cdot P(D+)}{P(T+)}.$$

The two probabilities in the numerator are given. We compute the denominator P(T+) using the law of total probability. Using the tree, we sum the probabilities for each of the nodes marked T+.

$$P(T+) = .995 \times .05 + .005 \times .9 = .05425$$

Thus,

$$P(D+|T+) = \frac{.9 \times .005}{.05425} = 0.082949 \approx 8.3\%.$$

Tables: Another trick that is useful for computing probabilities is to make a table.

Let's redo the previous example using a table built with 10,000 total people divided according to the probabilities in this example. We construct the table as follows: Pick a number, say 10,000 people, and place it as the grand total in the lower right. Using P(D+) = .005, we compute that 50 out of the 10,000 people are sick (D+). Likewise, 9,950 people are healthy (D−). At this point the table looks like:

	$D+$	$D-$	total
$T+$			
$T-$			
total	50	9950	10000

Using P(T +|D+) = .9, we can compute that the number of sick people who tested positive as 90 percent of 50 or 45. The other entries are similar. At this point the table looks like the table below on the left. Finally, we sum the T+ and T− rows to get the completed table on the right.

	$D+$	$D-$	total
$T+$	45	498	
$T-$	5	9452	
total	50	9950	10000

	$D+$	$D-$	total
$T+$	45	498	543
$T-$	5	9452	9457
total	50	9950	10000

Using the complete table, we can compute,

$$P(D+|T+) = \frac{|D+ \cap T+|}{|T+|} = \frac{45}{543} = 8.3\%.$$

Symbols: For completeness, we show how the solution looks when written out directly in symbols.[19]

$$\begin{aligned} P(D+|T+) &= \frac{P(T+|D+) \cdot P(D+)}{P(T+)} \\ &= \frac{P(T+|D+) \cdot P(D+)}{P(T+|D+) \cdot P(D+) + P(T+|D-) \cdot P(D-)} \\ &= \frac{.9 \times .005}{.9 \times .005 + .05 \times .995} \\ &= 8.3\% \end{aligned}$$

Before we move on to the next section of statistics, let's practice another question that is a bit conversational but has a flavor of probability. Interviewers often tend to embed technical questions within a case-type problem to get a sense of your thought process and overall problem-solving approach.

Practice Question

7. *Interviewer*: Let's say you walk into a fair and see a game with the follo instructions: There are six closed boxes and at any given time only tv them have a toy inside it. You can ask to open any box by paying $5 for turn. If the box you open has the toy in it, you win $10. Each time you the box, the game will be reset. Will you play this game?

Candidate: Hmm, can I ask a clarifying question before I continue?

Interviewer: Sure.

Candidate: Is there a limit on how many times I can ask to open a box?

Interviewer: No, as long as you're paying the $5 fee for each turn, you can play an unlimited number of times.

Candidate: Okay, so if I think through this, at any time there are two boxes out of six that have a toy inside it. Hence, the probability of getting the toy is 2 by 6 or 33 percent. And the return on my investment of $5 if I win is $10. Essentially, my return is two times that of my investment. On the other hand, the probability that I win is one to three. My risk is greater than the return on my investment, so this does not look favorable to me. Given this reason, I will not play.

Interviewer: Okay, that's fair. What amount if offered as the prize money will you be willing to pay?

Candidate: If my return is greater than or the same as my risk, I will be willing to play. So, in this scenario, if my investment is 33 percent of the prize money offered, I will play. Given that the cost to open a box is $5, if I am offered $15 as the prize money, I will be game to play as my risk and return will break even at those values.

Pro tip: While the interviewer is looking for the right numeric answer to their question, their main intention of asking probability-type questions is to test your knowledge of the fundamentals. So always make sure to explain your thought process as you solve the problem as well as any considerations or assumptions you have taken into account as you derive the final answer. You might be tempted to solve the question quickly and showcase your quick arithmetic abilities, but always remember to balance it with explaining the rationale behind it.

The "how" and "why" is always as important as the "what."

Statistics

In my experience of going through the data science interview process, I have often encountered a statistics-related question in some shape or form during the interview process. The goal of these questions is to test the candidate's fundamentals and their ability to explain the concepts with ease.

Let us walk through a couple example questions that you may encounter in an interview and sample answers for them.

8. *Interviewer*: Can you explain p-value?

Background: Before we dive into the answer, let's go over the basics of hypothesis testing.

Hypothesis testing is used to test the validity of a claim via a null and an alternative hypothesis. This testing is done on a sample rather than on the entire population.

Let's walk through an example so you can easily follow along. Let's say you have a friend that claims that the average GMAT score for the students at his university is 650. Now, you want to test his claim but don't have access to the data of all students. However, you are friends with quite a few people on Facebook and can send them a survey link to fill out their score so you can see what the average is like for that sample.

In this example, your hypotheses can be stated as follows:

Null hypothesis: The average GMAT score at your friend's university is 650 or less.

Alternative hypothesis: The average GMAT score at your friend's university is greater than 650.

The idea here is that we test which hypothesis is better supported by the sample data you obtain via the survey.

Let us now look at how our hypothetical candidate answers this question.

Candidate: P-value is used to determine the statistical significance of hypothesis testing. Largely, p-value is the probability of obtaining the observed results of a test, assuming the null hypothesis is correct.

Interviewer: Hmm, that makes sense. Can you explain how the significance of p-values is used for testing the null hypothesis?

Candidate: When using p-value to test the null hypothesis, you…

- Can reject null hypothesis and say that we have evidence for the alterr hypothesis when the *p-value is significant.*

- Cannot reject null hypothesis but can state that we do not have evidenc the alternative hypothesis when the *p-value is not significant.*

Interviewer: Okay, that sounds great so far. Can you walk me through an example of where you used p-value to determine significance?

Candidate: I conducted a small experiment to test a claim my friend made about the average GMAT score at his university being 650. I designed the hypothesis as follows:

Null hypothesis: The average GMAT score at my friend's university is 650 or less.

Alternative hypothesis: The average GMAT score at my friend's university is greater than 650.

After this, I collected data via a survey I sent to a subset of students at his university. I calculated the p-value based on the sample data and found that the average GMAT score was 700 with a p-value of 0.02. With a significance level of 0.05, I found the p-value to be significant. This meant I could reject the null hypothesis and say I had evidence for the alternative hypothesis.

P-values don't exactly prove something. In my experiment, it does not prove that my friend is incorrect but rather enables me to challenge his belief given that my p-value was significant.

Pro tip: Notice how the interviewer went into some of the details to check if the candidate thoroughly understood the concept or if their knowledge of the subject was superficial. Often, the best way to demonstrate expertise is to explain the concept using an example, preferably from a project you have worked on in the past. This highlights your practical knowledge of the subject and confirms your solid understanding of the fundamentals.

9. *Interviewer***:** Can you explain confidence interval to a nontechnical audie

*Candidate***:** Hmm, let me take a stab at it.

We use confidence intervals when we do not know what the true values are but want to know the range within which the true value can fall with a degree of certainty. Let me explain this with an example. Let's say we want

to know the average height of toddlers in the US. The exhaustive way of calculating this would be to measure each toddler's height and then take an average. However, this may not be a feasible approach, so we draw out a random sample of toddlers and measure their height and take an average of that. Now, to get a sense of the average height of all toddlers in the US, we calculate the confidence interval of the sample. For example, we can say that, based on a 95 percent confidence interval, the true average height of all toddlers is likely to be between thirty-three and thirty-eight inches.

Interviewer: Thank you for that explanation. Now statistically speaking, what are the key elements you need to calculate the confidence interval? You don't have to state the formula—just help me understand the different elements needed.

Candidate: Okay, so for calculating the confidence interval, you would need to know the number of observations in the sample, the mean, and the standard deviation of the sample. Depending on the confidence interval, for example 95 percent or 99 percent or any other, you would need the corresponding z-score. Using these elements, you can calculate the confidence interval.

***Pro tip**:* Typically, you won't be asked to state the formula for any statistic, but understanding the key elements behind it is helpful. Also, notice how the interviewer checks for a technical and nontechnical explanation in the same question. A data scientist is often required to communicate with both sets of audiences and fluency in explaining the concepts in both manners can be crucial for roles requiring both.

Extra questions for practice:

- What is the difference between one-tailed and two-tailed tests?
- What is the difference between Type I and Type II errors?
- What is a normal distribution and how are z-scores calculated?
- What is the central limit theorem? Why is it important?

- What potential biases can you inflict when sampling?

Experimental Design

In industries where design and modification of products and processes are routine (e.g., research and development, product development), an approach called *design of experiments* (DOE) is used to manage product and process improvements. DOE is a stepwise method based on statistics that efficiently guides the identification and selection of changes that collectively will optimize performance. Typically, this involves iterative testing of different factors, settings, and configurations, using the results of successive tests to further refine the product or process.[20]

When properly done, a DOE approach produces more precise results while using many fewer experimental runs than other methods (e.g., one factor at a time or trial and error). The outcome is a robust design that better meets customer specifications and production and delivery constraints.[21]

Experimental design has different applications depending upon the industry. Within product companies, it is popularly known as A/B testing and is carried out to test potential improvements to a website or a mobile application. Within the health care or insurance industries, you may see applications involving experiments carried out to test process improvement ideas, such a decreasing patient/customer dissatisfaction.

If the job role that you are applying to has experimental design or A/B testing as a required/preferred component, practice a few questions that you may anticipate in this area. Even if you don't have direct experience in the area, being familiar with the basics and the process will be beneficial and will allow you to have a conversation with the interviewer on this subject.

Let's walk through a broad question on this topic so we can familiarize ourselves with the overall process and also along the way dive into a few technical aspects.

Practice Question

10. *Interviewer*: Have you used experimental design on any of your projects?

    Candidate: Yes, I worked on a clinical intervention project as part of the practicum during my masters. Our practicum project was with HealthFirst insurance company, and they wanted us to design and analyze an experiment as part of it.

Interviewer: Okay, great. Can you walk me through the details of this project? Specifically, how you designed the experiment and then analyzed the results coming of it?

Candidate: Sure. To begin with, the business goal of the broader project was to reduce the claims cost for diabetes patients. HealthFirst wanted to launch a program for its members with diagnosed diabetes condition. This program would monitor the members' health on a periodic basis and encourage a suitable and healthy lifestyle to them to avoid any health-related complications arising out of having diabetes. With regular monitoring and maintaining a healthy lifestyle, HealthFirst was hoping to reduce the number of claims from diabetes patients, thereby reducing the cost to the company.

Before the program was launched for all patients, it wanted to test whether launching such a program would be effective in reducing claims. This is where our team that designed and analyzed the experiment came in.

We took a sample of one thousand patients with diabetes and split them into a test and control group. The test and control groups were selected based on similarity on certain parameters such as demographics, history of hospitalization and ER visits, and type of medicines prescribed.

Once our test and control groups were identified, HealthFirst launched the program on the test group, and we monitored key metrics affecting the result of the test over the next six months. The metrics we analyzed at the end of the six-month experiment were primarily around the number and number of claims of the test and control groups. Pertaining to this, we also looked at what the claims were for, e.g., visit to primary care provider versus specialist, regular visit versus ER visit, or in-network versus out-of-network provider used.

Based on our analysis, we found a significant drop in number of claims for the test group versus the control group proving the program to be effective.

Interviewer: That's great. Thank you for walking me through the details. Can you tell me how you made sure the test and control groups you chose were comparable?

Candidate: Sure, we took last twelve months of data as observation period. From this observation period, we drew out samples for test and control and compared them against the parameters I mentioned earlier: key demographic elements, number of visits, regular versus ER visits, rate of hospitalization, etc. We ran a t-test to compare the means of the two samples across these different parameters.

Interviewer: Got it. Was there a different test that you used for the categorical data that you had? Let's say you looked at gender as part of the demographics parameters you used.

Candidate: Yes, absolutely. For the categorical variables, I used the chi-squared test to check if the samples were similar.

Interviewer: That sounds good. Was there a reason you used the t-test and not the z-test for the numerical variables to compare the means?

Candidate: Yes, the z-test is used when the population parameters, such as mean and standard deviation, are known whereas the t-test is used when the population parameters are unknown. Given we drew out a sample from the total population of diabetes patients, I used the t-test.

***Comments**:* Experimental design is crucial because a cost is often associated with implementing something. In the example the candidate gave above, a cost is associated with the launch of the program. With the use of the testing that HealthFirst did, it was able to evaluate whether there is merit in launching the program at scale. Connecting the dots to the value your project brings in will help you differentiate yourself from your peers in the interview.

Extra questions for practice:

- How will you decide on the sample size to use for conducting the A/B tes
- How long should the experiment run?
- How often will you run an experiment?

- What should your test versus control split be?

In interview rounds such as this one, the primary aspect the interviewer is testing you on is your fundamentals. Make sure to revise the foundational

concepts used in data science, and you will breeze through this round!

---

13 "Probability of Compound Events: Definition & Examples," Study.com, accessed June 5, 2020.
14 "Compound Probability", Investopedia.com, accessed June 5, 2020.
15 Probability | Theory, solved examples and practice questions," MBA Crystal Ball, accessed May 17, 2020.
16 Probability | Theory, solved examples and practice questions," MBA Crystal Ball, accessed May 17, 2020.
17 "Conditional Probability", Math Goodies, accessed May 17, 2020.
18 "Conditional Probability: Definition & Examples", Statistics How To, accessed May 17, 2020.
19 Jeremy Orloff and Jonathan Bloom, "Conditional Probability, Independence and Bayes' Theorem," *18.05 Introduction to Probability and Statistics* (Spring 2014). Massachusetts Institute of Technology: MIT OpenCourseWare.
20 Patton McGinley, "Tests of Change: Simulated Design of Experiments in Healthcare Delivery," Patient Safety & Quality Healthcare, July 14, 2009.
21 Ibid.

CHAPTER 6

PROGRAMMING QUESTIONS

PROGRAMMING IN SQL

As any data science work starts with extracting and wrangling data, Structured Query Language (SQL) has become a must-have skill set for a data scientist. You will also notice that SQL by far is one of the most requested skills in a data science job description. So, if your interview process has a programming component to it, your SQL skills will very likely be tested.

First, let me briefly recap relational databases. Relational database management systems (RDBMS) is the basis for SQL and uses a structured format to store its data. The data in a relational database is organized in the form of tables and stored in a rows/columns format. Typically, in an RDBMS, you will find multiple tables that are related to each other.

If you would like to brush up on your knowledge of RDBMS a bit more or if this is relatively new to you, check out the references at the end of this book.

You can use SQL in many different database systems, such as MySQL, Microsoft SQL Server, MS Access, Oracle, Sybase, Informix, and Postgres. These database systems will have minor differences in syntax from each other, but having knowledge of at least one will suffice for your interview. For this chapter, we will be using the SQL syntax for Microsoft SQL Server.

The easiest and often best way to practice SQL for interviews is by writing and executing queries in a SQL environment so that you are confident with the syntax and can view the results being generated from your queries. If you would like to follow along in a SQL environment, you can download the Microsoft SQL Server trial version here: https://www.microsoft.com/en-us/sql-server/sql-server-downloads.

To go through key SQL concepts and to practice questions asked in a SQL interview, let us start by creating the following three tables in SQL Server:

Customers

```
CREATE TABLE Customers(
 CustomerId INT,
 FirstName VARCHAR(100),
 LastName VARCHAR (100),
 Email VARCHAR(100),
 City VARCHAR(100),
 State VARCHAR(100)
);

INSERT INTO Customers (CustomerId , FirstName, LastName, Email,
City, State) VALUES
(1, 'John', 'Collins', 'john@example.com', 'Chicago', 'IL'),
(2, 'Jennifer','McCarthy','jennifer@example.com', 'New York', 'NY'),
(3, 'Anna','Doe','anna@example.com', 'Stamford', 'CT'),
(4, 'Maria', 'Clark', 'maria@example.com', 'Hartford', 'CT'),
(5, 'William', 'Yang', 'william@example.com', 'White Plains', 'NY');
```

Orders

```
CREATE TABLE Orders(
 OrderId INT,
 OrderQty INT,
 CustomerId INT,
 ProductId INT,
 CreatedAt Datetime
);

INSERT INTO Orders(OrderId , OrderQty , CustomerId , ProductId,
CreatedAt) VALUES
 (1, 6, 2, 1, '20190618 10:34:00 AM'),
 (2, 5, 1, 3, '20190721 11:21:00 AM'),
 (3, 6, 3, 4, '20191115 09:12:00 PM'),
 (4, 2, 4, 5, '20190911 12:04:00 AM'),
 (5, 3, 3, 2, '20190325 01:06:00 PM'),
 (6, 4, 5, 2, '20190204 05:22:00 AM');
```

Products

```
CREATE TABLE Products(
 ProductId INT,
 ProductName VARCHAR(100),
 Price NUMERIC
);

INSERT INTO Products(ProductId, ProductName, Price) VALUES
 (1, 'Shampoo', 11.99),
 (2, 'Conditioner', 15.99),
 (3, 'Face Wash', 5.99),
 (4, 'Moisturizer', 8.99),
 (5, 'Tooth Paste', 2.99);
```

The basics for SQL are pretty straightforward, but given how often you may use some of these operations in your job, interviewers tend to test you on the basics first.

Let's go through a few practice questions based on the tables we defined above so you have the SQL fundamentals down!

Here are some basic SQL Concepts we will be looking at in the following section:

- Filter
- Sort
- Aggregate
- Joins

Practice Questions—SQL

Practice Question #1:

Write a query to show customers who live in the state of NY and order them by their last name in ascending order.

Solution:

```
SELECT *
FROM Customers
WHERE state = 'NY'
ORDER BY LastName;
```

Result set:

CustomerId	FirstName	LastName	Email	City	State
2	Jennifer	McCarthy	jennifer@ example.com	New York	NY
5	William	Yang	william@ example.com	White Plains	NY

Practice Question #2:

Write a query to show the total number of orders placed by each customer and sort by highest orders. Showing TotalOrders by CustomerId will suffice.

Hint: You will need to use joins and aggregate operations for this query.

Solution:

```
SELECT CustomerId, COUNT(OrderId) as TotalOrders
FROM Orders
GROUP BY CustomerId
ORDER BY TotalOrders DESC;
```

Result set:

CustomerId	TotalOrders
3	2
4	1
5	1
1	1
2	1

Pro tip: The aggregate functions popularly used are sum, min, max, count, and avg.

Practice Question #3:

Write a query to show states that have more than one customer record.

Hint: The WHERE keyword cannot be used with aggregate functions, hence HAVING is used.

Solution:

```
SELECT State, COUNT(CustomerId) as TotalCustomers
FROM Customers
GROUP BY State
HAVING COUNT(CustomerId) > 1;
```

Result set:

State	TotalCustomers
CT	2
NY	2

With the basics covered, let's now dive into the following intermediate SQL concepts:

- Union, Intersect
- Sub-queries

- Case statements

For the next question, suppose you have another table in addition to the three above:

Suppliers

```
CREATE TABLE Suppliers(
 SupplierId INT,
 Name VARCHAR(100),
 Email VARCHAR(100),
 City VARCHAR(100),
 State VARCHAR(100)
);

INSERT INTO Suppliers(SupplierId, Name, Email, City, State) VALUES
(1, 'ABC Group', 'abc@example.com', 'Chicago', 'IL'),
(2, 'XYZ Group','xyz@example.com', 'New York', 'NY');
```

Practice Question #4:

Write a query to show the list of cities where the customers and suppliers are from.

Solution:

```
SELECT City
FROM Customers
UNION
SELECT City
FROM Suppliers;
```

Hint: You do not have specify DISTINCT, as UNION will display each city only once.

Result set:

City
Chicago
Hartford
New York
Stamford
White Plains

Practice Question #5:

Write a query to show OrderIds that have OrderQty greater than the average across all orders.

Solution:

```
SELECT OrderId
FROM Orders
WHERE OrderQty >
(SELECT avg(OrderQty)
FROM Orders);
```

Result set:

OrderId
1
2
3

Practice Question #6:

Write a query to show the first order of each customer. Showing CustomerId, OrderId will suffice.

Solution:

```
SELECT o.CustomerId, o.OrderId, o.CreatedAt
FROM Orders AS o
INNER JOIN (
 SELECT CustomerId, MIN(CreatedAt) AS MinCreatedAt
 FROM Orders
 GROUP BY CustomerId
) AS r
 ON o.CustomerId= r.CustomerId
 AND o.CreatedAt= r.MinCreatedAt
 ORDER BY o.CreatedAt
```

Result set:

CustomerId	OrderId	CreatedAt
5	6	2019-02-04 05:22:00 000
3	5	2019-03-25 13:06:00 000
2	1	2019-06-18 10:34:00 000
1	2	2019-07-21 11:21:00 000
4	4	2019-09-11 00:04:00 000

Practice Question #7:

Write a SQL query to display the full name of a state. For example, if the state is IL, show Illinois. For this question, show the full names for the three states present in the Customers table.

Solution:

```
SELECT DISTINCT State,
CASE
 WHEN State = 'IL' THEN 'Illinois'
 WHEN State = 'NY' THEN 'New York'
 WHEN State = 'CT' THEN 'Connecticut'
 ELSE 'State mapping is not available'
END AS State_FullName
FROM Customers;
```

Result set:

State	State_FullName
IL	Illinois
NY	New York
CT	Connecticut
CT	Connecticut
NY	New York

While the majority of problems can be solved with filters, joins, and aggregates, you may encounter situations where using CTE or window functions may be beneficial. For example, window functions are great when you need to aggregate without losing the individual line item information. Window functions do not output aggregates as a single row for the group but rather as a column to the original line items that you have in the table. CTEs are great to make your queries computationally less expensive.

For example, with a query involving a WHERE and JOIN clause, SQL will process the JOIN first and then the WHERE clause. In cases where the table size is quite large, JOIN on all rows will be an expensive computation, especially if you are planning to remove those rows anyway with the WHERE clause. In such cases, using a CTE is computationally more effective where you can apply the WHERE clause first and then do the JOIN in the following statement.

Let's look at a couple of advanced practice questions so we are familiar with how to use CTEs and window functions. We will review the following concepts:

- Common table expressions (CTEs)
- Window functions

Practice Question #8:

Write a query to show the total revenue generated by each city. Include state as a column in the final result set. Revenue for each product can be calculated as a multiplication of price and order amount for that respective product.

Solution:

—First, find the amount for each order

```
with amt (OrderId, CustomerId, OrderQty, Price, Amount) as (
SELECT o.OrderId, o.CustomerId, o.OrderQty, p.Price, (p.Price *
o.OrderQty) as Amount
FROM
Orders as o
LEFT JOIN
Products as p
ON o.ProductId = p.ProductId),
```

—Next, find the total amount for each customer:

```
totalAmt (CustomerId, TotalAmount) as (
SELECT CustomerId, sum(Amount) as TotalAmount
FROM amt
GROUP BY CustomerId)
```

—Finally, find the total revenue by city, state:

```
SELECT c.City, c.State, sum(t.TotalAmount) as Revenue
FROM Customers as c
JOIN totalAmt as t
ON c.CustomerId = t.CustomerId
GROUP BY City, State;
```

Result set:

City	State	Revenue
Hartford	CT	6
Stamford	CT	102
Chicago	IL	30
New York	NY	72
White Plains	NY	64

Practice Question #9:

Write a query to show what percentage of a state's total revenue comes from each city in that state. Assume for this you have the result set named RS with columns City, State, Revenue (similar to the result set obtained as part of Q8).

Solution:

```
with StateTotal as (
 SELECT City, State, Revenue,
 SUM(Revenue) OVER (PARTITION BY State) as StateRevenue
 FROM RS)

SELECT City, State,
 Revenue/StateRevenue as Percent_of_StateRevenue
 FROM StateTotal ;
```

Result set:

City	State	Percent_of_StateRevenue
Hartford	CT	0.05556
Stamford	CT	0.94444
Chicago	IL	1
New York	NY	0.52941
White Plains	NY	0.47059

Here are some additional questions for practice:

- Write a query to show total number of orders placed by a customer. orders have been placed by a customer, display 0. (**Hint**: Use left join.)
- Write a query to show customers that are from the same city. (**Hint**: Use join.)

- Write a query to show the list of cities that the customers are from. (**Hint** DISTINCT keyword.)

Programming in Python, R

Today's data science recruiting process almost always has a technical screening round. The company interviewing you is likely to test how adept your programming skills are. These assessments are more streamlined these days by way of an online test on a platform such as CodeSignal, HackerRank, or TestDome. Some companies may even test you outside of these online platforms by giving you an at-home exam or by conducting an in-person whiteboarding interview.

Irrespective of the modes of the tests, they are primarily geared toward testing whether a candidate is familiar with the coding fundamentals to extract, analyze data, and suggest conclusions.

Currently, Python and R are the two most popular languages. Your employer will often let you choose between them, but if they are an exclusive Python or R shop, you will have to test on the language of their choice.

In the following section, we will look at a few practice questions so the most frequently tested concepts are at the front of your mind before an interview. If either language is completely new to you, at the end of the book you will see a few resources that are good places to start.

While you can use vanilla Python or R to work with data, using libraries makes this process faster and more efficient. Both languages have many useful libraries, but I'm listing a few below that are popularly used. Please note that these aren't ranked, as they are all useful in their own way.

Packages in Python:

- numpy—used for basic and advanced array operations
- pandas—used for working with data frames
- scikit-learn—used for common machine learning algorithms
- py-torch, tensorflow—used for deep learning
- matplot-lib—used for data visualization

Packages in R:

- dpylr—used for data manipulation
- data.table
- tidyverse
- ggplot2—used for data visualization
- caret—used for classification and regression problems
- stringr—used for string operations

The data science programming universe covers a wide range of items. I have broken them down into three logical sections so you can focus your preparation in a sequential approach, covering the most tested areas.

Let's look at these three sections:

1. **Programming basics**: This section covers basic programming construct use of different data types, arithmetic/logical/string operations, funct looping constructs, and decision constructs.
2. **Data wrangling**: This part covers the essential skills required during processing of data such as filter, subset, aggregate, and merge datasets. T functionalities are crucial for any data analysis work.
3. **Data science/machine learning functionality**: This section is specific t use of libraries relevant to data science/machine learning. This includes, l not limited to, building of classification and regression models, data m solutions, and deep learning techniques.

Now that we have a handle on the key areas we need to focus on, let us dive into a few practice questions so we can flex our muscles for any programming questions that come up during an interview.

Practice Questions—Python and R

Practice Question #10:

Given a list of strings containing the name of the subject and its associated test score, find the subject with the highest test score. You can assume that this data is only for a single student. The string is in the format of "stats,099" where the first four characters represent the subject and the last three characters represent the test score. The test score is always between 0 and 100 and will be padded with leading zeroes if required. For example, if the score is 85, it will be displayed as '085' in the string.

Write a function to display the subject with the highest test score.

```
Python Input Data: data = ['math,067', 'stat,085', 'econ,054',
'hist,092']
R Input Data: data = list('math,067', 'stat,085', 'econ,054',
'hist,092')
```

Python Solution:

```
def getMaxScore(data):
 max_score = 0
 max_subject = ''
 for data_point in data:
 subject, score = data_point.split(',')
 if int(score) > max_score:
 max_score = int(score)
 max_subject = subject
 return max_subject
```

```
data = ['math,067', 'stat,085', 'econ,054', 'hist,092']
result = getMaxScore(data)
print("Subject with the highest test score is: " +result)
```

```
Subject with the highest test score is: hist
```

R Solution:

```
getMaxScore = function (data) {
 max_score = "000"
 max_subject = ""
 for (i in data) {
 x = strsplit(i, ",")[[1]]
 if (x[2] > max_score) {
 max_score = x[2]
 max_subject = x[1]
 }
 }
 return(max_subject)
}

data = list('math,067', 'stat,085', 'econ,054', 'hist,092')
result = getMaxScore(data)print(paste0
 ("Subject with the highest test score is: ",result))
```

```
[1] "Subject with the highest test score is: hist"
```

Practice Question #11:

To practice our data wrangling skills, let's focus on writing code to the same result set as we did in the last SQL question. Write code to show what percentage of a state's total revenue comes from each city in that state. For this question, use the three tables as a starting point (Customers, Orders, Products).

Here are the tables again:

Customers

CustomerId	FirstName	LastName	Email	City	State
1	John	Collins	john@example.com	Chicago	IL
2	Jennifer	McCarthy	jennifer@example.com	New York	NY
3	Anna	Doe	anna@example.com	Stamford	CT
4	Maria	Clark	maria@example.com	Hartford	CT
5	William	Yang	william@example.com	White Plains	NY

Orders

OrderId	OrderQty	CustomerId	ProductId	CreatedAt
1	6	2	1	6/18/2019 10:34
2	5	1	3	7/21/2019 11:21
3	6	3	4	11/15/2019 21:12
4	2	4	5	9/11/2019 00:04
5	3	3	2	3/25/2019 13:05
6	4	5	2	2/4/2019 05:22

Products

ProductId	ProductName	Price
1	Shampoo	12
2	Conditioner	16
3	Face Wash	6
4	Moisturizer	9
5	Tooth Paste	3

Python Solution: For this solution, I read in the three data files using a CSV in Python and then proceeded with the required calculation.

```
import pandas as pd
#### Read in data files
customers = pd.read_csv("customers.csv")
orders = pd.read_csv("orders.csv")
products = pd.read_csv("products.csv")

##### Merge orders with product to get Price column
ord_prod = pd.merge(orders, products, on='ProductId', how='left')

#### Calculate $ Amount per Order
ord_prod['Amount'] = ord_prod['OrderQty'] * ord_prod['Price']

#### Merge in customers table to get City, State columns
ord_prod = pd.merge(ord_prod, customers, on='CustomerId',
how='left')

#### Aggregate to get revenue by city and state
rev_by_city = ord_prod.groupby(['City', 'State']).agg({'Amount' :
['sum']})
rev_by_city.columns = ['Rev_City']
rev_by_state = ord_prod.groupby(['State']).agg({'Amount' : ['sum']})
rev_by_state.columns = ['Rev_State']
rev_by_city = pd.merge(rev_by_city, rev_by_state, on = 'State',
how='left')

#### Calculate percent Revenue by City
rev_by_city['Perc_Rev'] = rev_by_city['Rev_City'] /
rev_by_city['Rev_State']
```

```
print(rev_by_city)
```

State	Rev_City	Rev_State	Perc_Rev
IL	30	30	1.000000
CT	6	108	0.055556
NY	72	136	0.529412
CT	102	108	0.944444
NY	64	136	0.470588

R Solution: Similar to the Python solution, I read in the CSV files in R before proceeding with the required calculation.

```
#### Load libraries needed
library(data.table)

#### Read in data files
customers = read.csv("customers.csv")
orders = read.csv("orders.csv")
products = read.csv("products.csv")

##### Merge orders with product to get Price column
ord_prod = merge(x=orders, y=products[,c('ProductId', 'Price')],
by=c('ProductId'), all.x = T)

#### Convert to data.table
ord_prod = as.data.table(ord_prod)

#### Calculate $ Amount per Order
ord_prod[, Amount:= OrderQty*Price]

#### Merge in customers table to get City, State columns
ord_prod = merge(x=ord_prod, y=customers[,c('CustomerId', 'City',
'State')],
by=c('CustomerId'), all.x = T)

#### Aggregate to get revenue by city and state
rev_by_city = ord_prod[,.(Rev_City=sum(Amount)), by=c('City',
'State')]
rev_by_state = ord_prod[,.(Rev_State=sum(Amount)), by=c('State')]
rev_by_city = merge(x=rev_by_city, y=rev_by_state,
by=c('State'), all.x = T)

#### Calculate percent Revenue by City
rev_by_city[, Perc_Rev:= Rev_City/Rev_State]
rev_by_city
```

```
> rev_by_city
   State         City Rev_City Rev_State   Perc_Rev
1:    CT     Stamford      102       108 0.94444444
2:    CT     Hartford        6       108 0.05555556
3:    IL      Chicago       30        30 1.00000000
4:    NY     New York       72       136 0.52941176
5:    NY White Plains       64       136 0.47058824
```

Practice Question #12:

For this question, you are given an input data set with the following columns: Sales Volume, Price. The end goal is to develop a linear regression model to predict sales volume based on the price of the product. To measure model performance, you will calculate the mean average percentage error (MAPE) for the test data.

Here are the specific functions you need to develop to get to the end result:

1. Write a function to impute missing price values by average.
2. Write a function to calculate correlation of the dependent and indepen variable.
3. Write a function to train and score the model on train and test respectively. For this, you can split the given input data set into train and set by using an 80:20 split.
4. Write a function to calculate and display the MAPE for the test data. For you can assume that Sales_Volume will never be 0.

Python Solution:

```
def impute_data(data):
 data.fillna(data.mean(), inplace=True)
 return data

def calc_corr(data):
 correlation = data['Sales_Volume'].corr(data['Price'])
 return round(correlation,2)

def run_linear_reg(x_train, y_train):
 linear_mod = LinearRegression()
 linear_mod.fit(x_train, y_train)
 return linear_mod

def score_linear_reg(linear_mod, x_test, y_test):
 y_pred = linear_mod.predict(x_test)
 predictions = pd.DataFrame({'Actuals': y_test, 'Predicted':
 y_pred})
 return predictions

def calc_mape(predictions):
 mape = metrics.mean_absolute_error(predictions['Actuals'],
 predictions['Predicted'])
 return round(mape,2)

# Load libraries
import numpy as np
import pandas as pd
```

```
from numpy import cov
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics

#### Read in data filesdata = pd.read_csv("SalesData.csv")

#### Impute missing price values
data = impute_data(data)

#### Calculate correlation between the dependent and independent
variable
correlation = calc_corr(data)
print("Correlation between the dependent and independent variable
is: ", correlation)

#### Split into train test
x_train, x_test, y_train, y_test = train_test_split(data[['Price']],
data['Sales_Volume'], test_size=0.2, random_state=1)

#### Develop linear regression model on train data
linearMod = run_linear_reg(x_train, y_train)

#### Score linear regression model on test data
predictions = score_linear_reg(linearMod, x_test, y_test)
print(predictions)

#### Calculate MAPE for the test periodMAPE = calc_mape(predictions)
print("MAPE for the linear regression model is: ", MAPE, "percent")
```

```
Correlation between the dependent and independent variable is:  -0.85
   Actuals   Predicted
3      256  238.111111
7      523  530.680556
6      621  647.708333
MAPE for the linear regression model is:  17.43 %
```

R Solution:

```
imputeData = function (data) {
 data$Price[is.na(data$Price)] = mean(data$Price, na.rm=TRUE)
 return(data)
}

calcCorr = function (data) {
  corr = cor(data$Sales_Volume, data$Price)
  return(round(corr,digits=2))
}

runLinearReg = function (data) {
  linearMod = lm(Sales_Volume ~ Price, data=data)
  return(linearMod)
}

scoreLinearReg = function (linearMod, data) {
 pred = predict(linearMod, data)
 predictions = as.data.frame(cbind(pred, data$Sales_Volume))
```

```
 names(predictions) = c('Predicted', 'Actuals')
 return(predictions)

}

calcMAPE = function (predictions) {
 MAPE = mean(abs(predictions$Actuals-
 predictions$Predicted)/(predictions$Actuals))
 return(round((MAPE100),digits=2))
}

#### Read in data files
data =  read.csv("SalesData.csv")

#### Impute missing price values
data = imputeData(data)

#### Calculate correlation between the dependent and independent
variable
corr = calcCorr(data)
print(paste0("Correlation between the dependent and independent
variable is: ", corr))

#### Split into train test
sampleSize <- floor(0.80  nrow(data))
# Set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(data)), size = sampleSize)
trainData <- data[train_ind, ]
testData <- data[-train_ind, ]

#### Develop linear regression model on train data
linearMod = runLinearReg(trainData)
summary(linearMod)

#### Score linear regression model on test data
predictions = scoreLinearReg(linearMod, testData)

#### Calculate MAPE for the test period
MAPE = calcMAPE(predictions)
print(paste0("MAPE for the linear regression model is: ", MAPE,
"percent"))
```

```
[1] "Correlation between the dependent and independent variable is: -0.85"
```

```
> summary(linearMod)

Call:
lm(formula = Sales_Volume ~ Price, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-212.614  -54.857    1.669   64.355  182.130

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1315.21     194.55   6.760 4.98e-05 ***
Price         -56.26      12.45  -4.519  0.00111 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 110.6 on 10 degrees of freedom
Multiple R-squared:  0.6713,    Adjusted R-squared:  0.6384
F-statistic: 20.42 on 1 and 10 DF,  p-value: 0.00111
```

```
> print(paste0("MAPE for the linear regression model is: ", MAPE, "%"))
[1] "MAPE for the linear regression model is: 2.27%"
```

```
> print(predictions)
   Predicted Actuals
8   527.6135     523
9   471.3569     455
13  696.3833     713
```

Comments: I have kept the code simple to make it easy to follow. However, this code can be generalized further if needed. For example, the correlation calculation function references the columns directly by their name (Sales_Volume, Price). Instead of hard coding the column names in the functions, you can pass them by reference so the correlations can be calculated for any two columns.

When answering these types of questions in the coding test, follow a simple approach first, and then you can work on generalizing and making the code more efficient if time permits. The primary goal should be to solve the problem at hand in a simple manner and pass all test cases.

Additionally, here a few resources you can use to practice your coding skills:

- LeetCode
- HackerRank
- CodeSignal
- TestDome

- CoderByte

CHAPTER 7

CASE QUESTIONS

Interviews involving case studies are quite common for many job roles. For data science roles, these have evolved to mimic the work the data scientist will be doing if hired. They're a great way for companies to simulate a data science project and assess how a candidate will perform in the real world.

Before we dive into any practice questions, let's understand what interviewers are looking for in a case interview and how to best tackle them.

What is an interviewer testing you on in a case interview?

- **Ability to synthesize information**—Interviewers are looking to see if yo piece together information on multiple fronts and come up with coheren concise ideas. Most case interviews begin with a broader prompt givi overarching view of the problem statement. The interviewer is looki whether you can break down this top-level information and work tow solution by putting together the jigsaw pieces along the way.
- **Analytical thinking**—Given the nature of the job role for a data scientis interviewer will be judging you on your ability to think from an anal lens and highlight how data can be best used to solve the business pro posed to you.
- **Skills**—Interviewers are looking for how well you communicate approach and the findings and how you summarize the overall problem. will also be looking at your ability to communicate with both a technica nontechnical audience.

- **Product Sense**—If you're interviewing for a product company, they w evaluating you on your "product sense"—what makes a product grea what doesn't. You can practice developing your product sense by looki

everyday things critically. You may ask yourself questions like: Why i iPhone so easy to use? What is so great about the Instagram story fea What makes it easy to search for bed and breakfasts on Airbnb? If you w like to take it a bit further and wear the product manager hat for a while, start to think about what can make this product better. While you may n asked product enhancement questions in a data science interview, pract along these lines will allow you to think outside the box and will push creative mindset.

What is a good approach to tackle case questions?

- **Ask and understand the "*What*"**—When you are given the case prompt a minute or two to carefully understand the question and the objective be it, and confirm your understanding of what is being asked. You may know the answer to the question off the bat, and that's okay. Rather, t how it should be. You will get to the finer aspects later on, but as the first understand what the problem statement is and what is it you are task solve for.
- **Dive into the "*Why & How*"**—Once you have a handle on the goals, n into the "*Why & How*." For example, if it's a profitability case, identif drivers of profitability and then move to understanding how changes in drivers' impact profitability. Write down your structure as you do this s interviewer can give you early feedback if you are steering away from w they want you to go.
- **Hypothesize and brainstorm options**—Based on the overall goal, hypotheses and talk through the various options as you see fit for the This is also a great time in the interview to ask if additional data point available to dig in further and to validate your assumptions with interviewer along the way.
- **Actively communicate thought process**—Having given multiple interviews as part of the recruiting process after his MBA, Gan Raghunathan offered the following insight: "It is key to keep communic your thought process with the interviewer. The interviewer will usually the candidate back on track if they go on a tangent."

- **Summarize and recommend**—Once you have evaluated different op

and essentially walked the interviewer through your "decision
summarize your recommendation/conclusion, list out any key assump
you have made while arriving at the conclusion, and state any risks tha
recommendation may bring in.

Other things to watch out for:

- If in the process you ever stuck and things seem daunting, think of yours
the interviewer's *colleague.* Think of the question as if you were leadin
project. This will put you right back in the driver's seat and will give yo
confidence to ask the right questions and formulate an approach.
- If you find yourself rambling and giving long-winded answers, pause. T
step back and ask for a couple minutes to work through your thou
Structure your response and write down key points and then start to ar
again. Remember, there is never a wrong time to ask for a couple minut
structure your response.
- Do not be fixated on only one area, especially if the interviewer is not kee
going in that direction. Read the cues from your interviewer as they ofte
nudge you in the correct direction.

- Do not get swayed into building a perfect data-driven solution only. Va
it with your business intuition and assess whether it fulfills the br
business objective before making a recommendation.

Let us now look at a few practice cases.

Practice Case #1—Churn Prediction

The first practice case that we will look at starts off as a generic case question and then gets into the specifics of how a data science approach can be used to solve the problem.

Case Prompt: BeautyBox is an online monthly subscription service that sends its subscribers a box of seven selected samples of makeup and beauty-related products. The products include skincare items, perfumes, and other cosmetics. Over the past year or so, they have noticed a decline in profits. What factors may be contributing to this and what can they do to alleviate the situation?

Candidate: Can I take a couple minutes to frame my thoughts?

Interviewer: Absolutely! Take your time.

***Pro tip**:* Taking a couple minutes at the start of the case interview to organize your thoughts is extremely beneficial. This will allow you time to consider the different factors involved in solving the business challenge. You can also take this time to understand the case question better and ask any clarifying questions to the interviewer.

Candidate: I had a few questions before I dive into the answer. Can you help me understand the business model of BeautyBox? Do its customers subscribe for a given period of time and then renew after that? And are they allowed to cancel at any time? Also, does it have a different product mix that its customers can choose from?

Interviewer: Great questions! So, its customers can choose from a three-, six-, or twelve-month subscription option. And they are allowed to cancel at any time. Currently, it only has one subscription box to offer to their customer base. Each month it sends a beauty box containing seven different cosmetic items for its customers to try.

Candidate: Makes sense. Thank you for that information. As I am thinking through the question you initially asked me, I think there could be a few different factors that could affect their profits.

Here's a breakdown of the factors I'm considering:

- Revenue

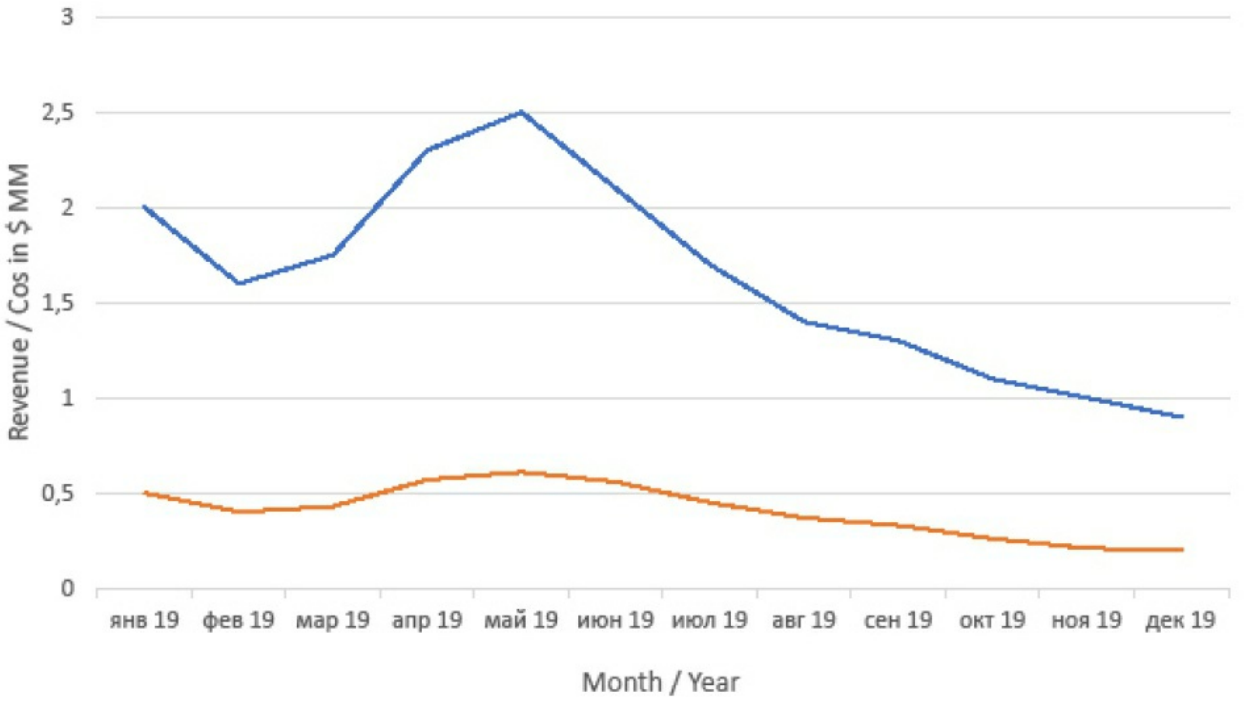
  - Number of subscriptions sold
  - Price per subscription

- Cost

  - Cost of goods produced
    - Product cost
    - Fulfillment cost
  - Cost per customer

- Customer acquisition cost
  - ► Discounts for first-timers, marketing cost, etc.
- Customer retention cost
  - ► Customer support, renewal management, etc.

Interviewer: That makes sense. Here's what we know about the volume of subscriptions and the total costs to the company in 2019.



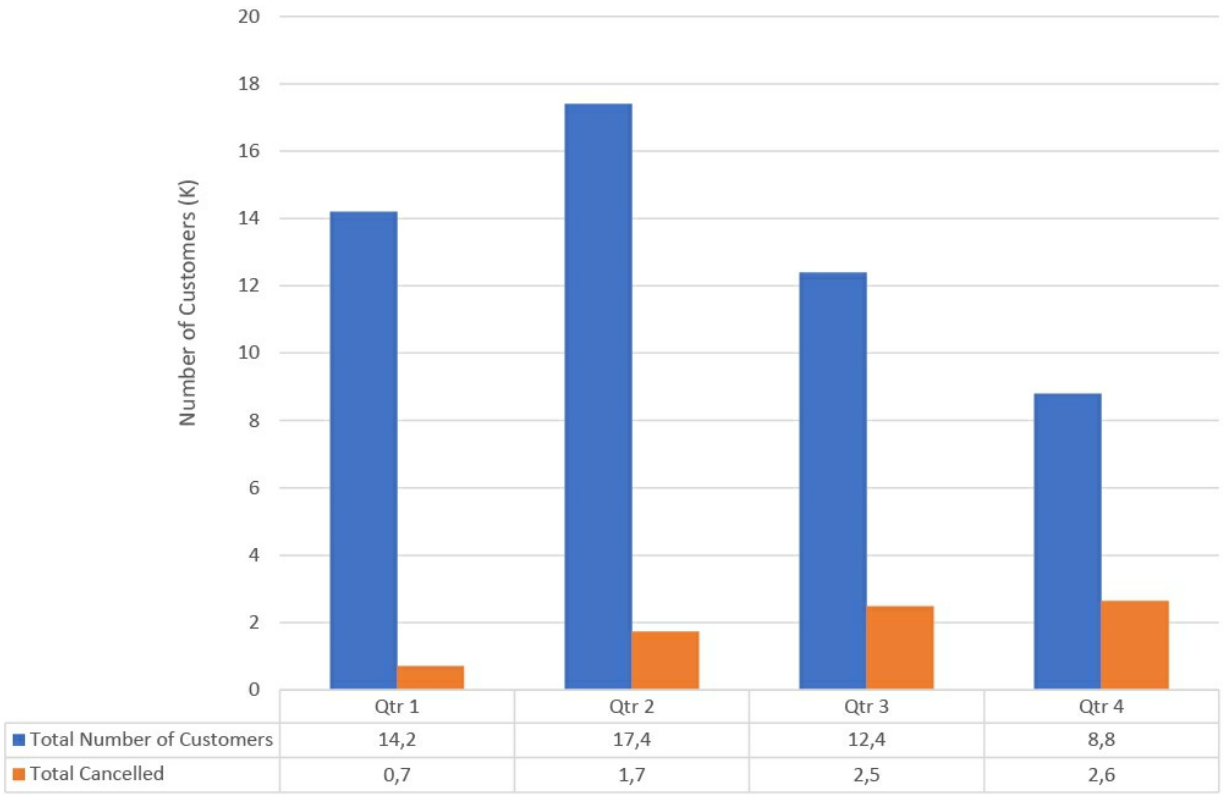
Breakdown of revenue and cost for BeautyBox in 2019

Candidate: Hmm, given that a single product is available to the customers, I assume the price for all is the same. Let me know if my assumption is incorrect. And I can see the costs have remained relatively steady over the year from the graph you shared with me. Looking at this data, I suspect the number of subscriptions is driving the profit down. Do we know anything more on how the subscription volume has changed over the months in 2019?

Interviewer: Your assumption about the price is correct. BeautyBox has seen a steady decline in the number of subscriptions over the last three quarters. It has seen an increasing number of cancellations. Looking at this graph, can you tell me what is the average quarterly churn rate?

***Comment**:* Notice how the candidate has stated their assumptions along the way. Openly stating the assumptions you make is a good strategy so that, in

case they are not valid, the interviewer can jump in and course correct if needed.



	Qtr 1	Qtr 2	Qtr 3	Qtr 4
Total Number of Customers	14,2	17,4	12,4	8,8
Total Cancelled	0,7	1,7	2,5	2,6

Breakdown of total number of customers vs. total cancelled in 2019

Candidate: Hmm, looking at the graph, I can see that the average churn rate is 16 percent. I calculated the churn rate for each quarter as the number of subscribers cancelling divided by the total number of subscribers in that quarter.

Comments*:* In case you do not recall a particular formula in the interview, you could ask the interviewer. Although, before you directly ask for the formula, state your understanding of what the calculation should be and confirm the understanding. This lets the interviewer know that even if you don't have the solution at your immediate disposal, you are able to think of an approach.

Interviewer: That's correct. The team at BeautyBox wants to know if it can identify the customers who are likely to churn before they actually cancel. What model can you build for this?

Candidate: Absolutely. We can identify the customers who are likely to cancel with the help of machine learning techniques. In fact, if the company's profits are declining due to an increasing number of customers canceling the subscription, implementing churn modeling techniques to retain the customers is definitely worth exploring. From what I understand, usually the cost to acquire a customer significantly outweighs the cost to retain a customer. So, if BeautyBox is able to identify the customers who are likely to leave, it can take any necessary action to try to retain them.

***Comments**:* Notice how the candidate not only mentions that it is possible to predict churn but also shares what BeautyBox can do once they are able to predict it. This shows that the candidate has technical understanding of the subject and is also able to explain the advantages of the method in business terms.

Candidate: Coming to the modeling aspect for this, I believe you can use regression or tree-based models to predict the likelihood of customers who would cancel. In the case of regression, one of the ways to do it is to build a logistic regression model that predicts the probability of a customer churning. Similarly, for tree-based models, random forest or boosted trees can be used to make the same prediction. While these are some of the techniques you can use to build the models, you can assess which technique gives you the best results with the help of cross validation and comparing out of sample accuracy across different models.

Interviewer: That sounds great. Can you tell me what data you would use to build such a model?

Candidate: Multiple data points can affect the churn rate for BeautyBox. Broadly speaking, they can look into the following categories:

- **Demographic information**: Demographic data of the customers such as gender, income level, and employment status can be used as an input int model.
- **Interactions with support team**: Data such as number of times a cust has interacted with the support team, total number of complaints made, taken to address each complaint, and time between complaints are informers of how the customer interacts with the company's support t

Customer satisfaction can be inferred from these data points to an exten in all, they are great inputs to estimate churn.

- **Customer account information**: It is also useful to look at wher customer first subscribed, what brought them in (promotional offer, fri referral, etc.), how long they have been a customer, subscription type use, and if they have canceled and resumed the subscription at any point.

- **Social media information**: If possible, layering in the social media would be an excellent addition. This can be done from the perspective of the customers are talking about the product on social media to get an ic the customer sentiment.

Interviewer: Excellent! Those are some good data points to include. Now, can you tell me what metrics you will use to measure model accuracy?

Pro tip: We have covered the details on these accuracy metrics in Chapter 4. Feel free to browse the exact definitions in there to recap!

Candidate: To measure the accuracy of a churn prediction model, I'll start by developing the 2x2 confusion matrix with the predicted classes represented in the columns and the actual cases represented in the rows. The first metric to look at is the total accuracy of the model measured as number of cases correctly predicted/total number of cases.

Next, with the help of the confusion matrix, I will look at the false positives and the false negatives. False positive will falsely identify someone who was not going to churn as someone who churned, whereas false negative will measure the number of cases in which you failed to identify someone who was likely to churn.

These two metrics are important because if false positives are too high, you end up investing in customers who would have stayed anyway. If false negatives are too high, then you miss out on identifying people who are going to churn and lose the opportunity to persuade them to stay.

A cost is associated with both of these scenarios. Hence, we need a balance between both metrics. The F-score metric can be looked at to identify the required balance.

Interviewer: That sounds great. What would you recommend the business to do once they have identified the customers as likely to churn?

Candidate: Once the business knows which customers are likely to cancel the subscription, they can take action to prevent this from happening. From the models built to predict churn, they will be able to isolate the key indicators of churn. Those indicators will be the call to action. For example, an indicator can be multiple phone calls to the customer care team with the issue status still resolved. When these indicators are identified, BeautyBox can design a churn mitigation plan and take action before the customer decides to cancel. The mitigation plan can be formulated at a customer segment level to make it more targeted and to concentrate efforts toward its high-valued customers. The segmentation, for example, can group its customers into three categories: high-valued, medium-valued, and low-valued customers. In terms of priority, the high- and medium-valued customers should be action-ed first.

Comments*:* Overall, the candidate did a good job in answering the case questions. They were able to talk about the data science techniques applicable to this problem and to give an overall recommendation as to how the insights from that can be leveraged for the business. They stated their assumptions along the way and confirmed their understanding before stating any conclusions.

Practice Case #2—A/B Testing

When interviewing for a product company, a case question determines whether a candidate can methodically tackle a broad question and make a recommendation for the team to act on.

The nature of work for a data scientist can differ based on the organization structure of the data science team.

Data scientists can be embedded in the product team working alongside the product managers, engineers, UI/UX designers, etc. Data scientists working within the product team bring in an analytical lens when working on problems. They are expected to explore and analyze data sets to uncover insights for the team. Within the product realm, when a decision about whether or not to launch a feature comes up, the data scientist helps by

designing experiments and tracking metrics to measure experiment performance.

Typically, data scientists within a product team aren't tasked with data engineering activities. While there is an element of merging and aggregating data sets that is involved with any analytics work, building ETL pipelines isn't part of a data scientist's role description. The same goes with deploying machine learning models to productions. Data scientists build models and conduct the analysis as needed, but within a product team, the engineers are often the ones who deploy the models to production.

Data scientists can also be a part of a centralized/core function. Data scientists as part of a central team consult and work on different projects across the organization. The core team can be made up of data scientists, machine learning engineers, and data engineers. The work of the core team can potentially be of research nature unlike the data scientist on the product team.

Let us now dive into our practice case #2. This question throws light on the type of cases you can expect when interviewing for a product company.

Interviewer: Facebook is considering showing more ads in place of the "People you may know" feature. Do you think the team should go ahead with this implementation?

Candidate: I would like to start by talking about the overall goals for Facebook and what I think the goals for the features you described are and then go from there.

Interviewer: Sounds good.

Candidate: Facebook enables people to stay connected with their family and friends, and its goal is to give people a platform to build a community with whom they can share their experiences. From what I know, the "People you may know" feature directly correlates with their goal and enables users to build their existing networks. Furthermore, Facebook's primary source of revenue is ad sales, some of which come from the ads directly displayed on the news feed, which is the feature we are talking about I assume.

Interviewer: Yes, you are correct about the news feed assumption and the goals you mentioned are aligned correctly as well.

Candidate: Great, now I would like to think through the pros and cons of each feature.

Pros and cons by feature

Feature	Pros	Cons
People you may know	Increases user engagement and retention	For a mature user, this feature may not be as relevant
Ads	Increase revenue for the company	May make the user experience a bit cluttered and impact brand perception

While there are trade-offs for choosing either feature, we should align these decisions to any business objective that Facebook has. Is there a specific short- or long-term goal the team is working toward?

Interviewer: No, not specifically for the scope of this question.

Candidate: Okay. In that case, I propose running an A/B test to assess the impact of showing more ads to a randomized group of users and compare it against a control group who will see the news feed as is.

Interviewer: Sounds good. What metrics will you track when you run the A/B test?

Candidate: Let's see. From the point of view of the two options, I will look at the following:

- Ads (metrics related to monetization)

  – Average revenue per user
  – Average revenue per active user

- People who you may know (metrics related to engagement and retention

  – Likes

- Comments
- Shares
- Clicks
- Total number of users
- Total number of active users

Interviewer: Thank you for listing those out. I see that you have listed out average revenue per *active* user specifically. Is there a reason you have in mind to track that?

Candidate: Yes, absolutely. From what I understand, Facebook users can be grouped into two key segments:

1. Active Users
2. Inactive Users

Active users engage with the product often and are the first to react to any product changes. Hence, we must track any changes in the product usage of the test group.

Interviewer: Okay, let's say we got the results back from the A/B test run over a two-week period. Based on these results, what is your recommendation?

A/B test results

Metric	**Lift**
Likes	-4.10 percent
Comments	2.80 percent
Shares	-3.20 percent
Avg. revenue per user	1.70 percent
Avg. revenue per active user	2.30 percent

Candidate: On one hand, while I see average revenue going up by 1 percent overall and 2 percent for active users, the number of likes and shares have gone down, indicating that the user engagement has taken a hit. While the comments metric has gone up, we must evaluate if these are positive or negative sentiments. People could be negatively commenting about the increase in ads, so looking at this increase in isolation may be deceptive.

In summary, given that Facebook's primary goal is to build a platform to connect users, I think a significant impact in the engagement metrics is a sign that something is not right. Facebook needs to find that optimal point where engagement is not impacted but the revenue generated increases. For now, I would recommend to not show ads in place of the "People you may know" feature.

Comments: The candidate offered a very structured and thoughtful discussion around the pros and cons of using one feature more over the other. The candidate was able to use the data provided well and to convince the interviewer that they have a reasonable plan in place to approach such a problem.

Questions for practice relating to case #2:

- How do you measure the success of Instagram stories?
- How will you forecast user sign up for a new product?

- An online platform suddenly saw an increase in sign-ups. What could b reasons for this?

Practice Case #3—Business Analytics

Let us look at another case that centers around the use of business analytics to solve the problem at hand.

Interviewer: A major hotel chain has been experiencing slow growth and declining profits in the last few years. The CEO has hired you to help them understand the drivers behind this and ultimately help increase profitability.

The hotel chain has thirty-nine hotels spread across the United States and, despite major back-end cost savings, it is experiencing low profits.

Candidate: I have a few questions before I jump into the case.

Interviewer: Sure, go ahead.

Candidate: I want to understand the hotel's business model better. Is it centrally owned or does each location operate as its own franchisee and has an owner?

Interviewer: The hotel chain is centrally owned by our client, and each location has a hotel manager and supporting staff to run the operations.

Candidate: Okay, I would also like to understand the hotel profitability better. Are all stores equally profitable or do we see variations?

Interviewer: Some are more profitable than the others, and we see variations throughout. But that being said, irrespective of the location, all hotels have seen a decline in profits over the last two years.

Candidate: Thank you. You also mentioned in the beginning that the hotel chain has cut down on back-end costs. Can you tell me how the cost savings were achieved?

Interviewer: Sure, the hotel chain cut down on the costs by optimizing labor scheduling and staff training. With this approach, the staff was not underutilized, and cross-trained employees could lend a hand in another role when needed.

Candidate: Got it. Additionally, I would like to understand if the number of bookings has changed over the past two years.

Interviewer: We have the following data from the CEO on how the costs and bookings have changed over 2018 and 2019.

Costs and bookings for 2018, 2019.

Year-Quarter	Number of Bookings (in thousands)	Operating Costs (in thousands)
2018-Q1	783	1,279
2018-Q2	715	1,492
2018-Q3	689	1,394
2018-Q4	612	891
2019-Q1	499	873
2019-Q2	396	657
2019-Q3	468	761
2019-Q4	256	612

Candidate: The information provided here suggests the decrease in operating costs is correlated with the decline in number of bookings. We see a significant dip in the operating costs post-Q4 in 2018, which is when we see

the decline in number of bookings start.

If cost saving was primarily driven by cutting down on labor costs, this to me suggests a few things:

- It looks like in controlling costs the customer service got impacted. rationale behind my thinking is that in optimizing labor scheduling cross-training the staff, the level of service delivered wasn't at the same And if this is true, within the hospitality industry, bad service will customers.

- We see this especially in Q4 in 2018, where seasonality otherwise would suggested a peak in bookings, as that is when most folks travels d holidays. But in this case, we see a gradual drop in number of boo starting to show in Q4 2018 itself. We can validate this by looki competitor bookings and see how the client's numbers compare to them.

If the hotel chain cut down on costs in other areas, it would have an impact in the number of bookings, such as:

- The hotel chain cut down on marketing/advertising, leading to l bookings

- It reduced the number of promotions it runs

The hotel chain could also get impacted by external factors, such as:

- Decrease in competitor prices leading to customers switching to other chains
- General decrease in bookings due to any external factors such as the r COVID-19 virus

- Launch of a new hotel chain

Interviewer: So, we know that in general the competitors haven't had any major change in their pricing strategy and overall bookings in the industry have remained in the range as one would expect.

Your hypothesis on the cost control measures impacting customer

satisfaction is correct. The hotel chain has seen a decrease in customer satisfaction as the cost control measures were put in place. What would you suggest the hotel chain do in this case then?

Candidate: We need to better understand what aspects of the service customers are dissatisfied with. Once we have identified these factors, a strategy to improve the existing service while balancing the costs need to be put in place.

Interviewer: Okay, can you tell me how the hotel chain can measure customer dissatisfaction?

Candidate: Sure, they can do this in a few ways:

- By looking at the complaints data at the hotel chain service desk
- By looking at the customer service logs to see the volume and typ complaints logged
- By doing sentiment analysis to see what the hotel guests are saying on s media

Essentially, the hotel chain can leverage different data avenues available to them to see how the key metrics relating to customer satisfaction have changed over the last two years.

While doing so, my recommendation is to identify the root causes of customer satisfaction as well. With the insights into the drivers of low customer satisfaction, the hotel chain can develop a plan toward enhancing customer experience, thereby increasing its bookings and profits.

Interviewer: That sounds good. The CEO is keen on using data and analytics to enhance their customer satisfaction and experience. How do you recommend they do this?

Candidate: I can think of a few for this.

- **App for essentials**: Technology use for hotels is evolving a lot these days providing the guest with the latest in technology can be a game changer example, the hotel chain can have a mobile app that allows them to re

for services by the click of a button, allow them to check in online, and k track of their reward points.

- **Personalization**: The hotel chain can look at how it can personalize customer's experience while staying at its facility. For example, it can lo its different customer segments and understand what feature/service w valued by each. For families coming in to stay at its hotels, it can give gifts to the children to keep them entertained and ensure extra towels a the room before the family checks in. Small gestures such as this have sl to enhance customer experience, and such customers are likel recommend the hotel to their friends.
- **Improved recommendations**: The hotels can establish a connection w guest as soon as they have made a booking. The communication can in details on what they can do during the stay and the facilities they can u the hotel. These recommendations can be tailored to the personal tastes guest. For example, custom itineraries can be curated based on the g personal preferences and tastes. The hotel chain can look at a wa integrating their CRM system with the social media data. With a view o social media activity of its guests, it can easily unlock insights into the g habits and interests and provide tailored recommendations for their trav
- **Ask and deliver**: The hotel chain can get real-time feedback from its g during the stay and improve their service during the same trip. This can shift from gaining retrospective feedback to having a continuous feed loop with the guest and addressing their needs in real time. It can leverag mobile app to get feedback from the guests during their stay and rout check in to see if they have everything they need.

- **Provide exceptional service**: The hotel chain needs to look at what the complaints so far have been and come up with a tactical plan to resolve t If the service levels have dropped due to labor optimization, ther scheduling needs to be revised to ensure the staff isn't overwo Overworked staff could lead the service levels to be poor. The hotel needs to re-establish its service standards and train its employees and e them to give its guests a fabulous experience when staying at the hotel.

The hotel chain has ample opportunity to use data and analytics to improve its customer experience. From a tactical standpoint and immediate next steps,

my recommendation for the hotel chain is to address the primary reasons for customer dissatisfaction. And in parallel, it should look at further enhancing the customer experience and increase retention with the help of data analytics.

Comments: The candidate overall did a good job with the case and asked good clarifying questions at the outset. The candidate, however, could have done a better job at structuring their response when answering the question around the drivers of low profitability. The candidate thought through the reasons sequentially, but putting together a broader framework and then diving into the details one by one would have been a better approach in this case.

They also immediately jumped to conclusions when data around cost and bookings was given. An ideal approach would involve thinking through the different drivers of low profitability (internal and external), stating the observations based on the data given, and then stating the findings connected to the large business problem.

In conclusion, at some point in the interview process, you can expect to encounter a case-type question in some shape or form. To be well prepared for one, consider doing mock cases with a friend. The more cases you practice, the more comfortable you will be with the case interview process.

And be sure to practice cases in the relevant focus areas of the company you're interviewing for. Aniket Deshpande, senior data scientist at Aetna, shared this excellent advice with me: "Most companies ask questions related to their domain. The cue here is to research some practice questions related to their domain before interviewing for any company."

With that being said, I will leave you with a few tips to help you crack the case interview.

Pro tip: When given a case:

- Take a step back and think of a) any clarifying questions you want t before you dive in and b) possible approaches to solve the problem.
- When presented with data, take the time to think through the observa and see how it translates back to the larger problem asked.

- Always state any assumptions you are making.
- Talk through any mathematical calculations so the interviewer can f
  along with you and course correct if needed.
- When concluding the case, connect your recommendations with the l
  problem and lead with that. Suggest any additional analysis you recom
  doing to arrive at an answer that takes into consideration all aspects req
  to provide a recommendation. Remember, you may not dive into
  possible aspect during the interview, but broadly stating them help
  interviewer understand that you're able to look at the bigger picture an
  look into the details when needed.
- Additionally, state any risks you see with the recommendation you're ma

- At each step, remember to take the time to pause and think, and you
  great!

PART 3

SHOWCASING THE RIGHT EXPERIENCE

CHAPTER 8

TELL ME ABOUT A PROJECT YOU WORKED ON

Over the course of writing this book, I spoke to interviewers and interviewees both. Almost every person I spoke to mentioned asking or being asked the question, "Tell me about a project you worked on," in an interview. I probed my interviewers a little further, and some of them said this was in fact one of their favorite questions to ask! The reason was that the easiest way for an interviewer to find out how good a candidate is was by asking them about work they did previously.

Discussion on past projects isn't unique only to data science interviews but can be seen in almost any interview. Nailing this question down will help you come across as thorough and showcase your potential value to the interviewer.

What is an interviewer looking for?

An interviewer is largely looking for the breadth and depth of your knowledge when asking you this question. Let's dive into the specifics of what an interviewer is looking for.

Breadth:

- **Ability to explain a project you worked on from start to end**: interviewer will be looking at how you communicate what the projec about, what was done overall, what your specific contribution was, and value was generated. An effective response here shows the interviewe have an understanding of the entire pipeline of work.
- **Ability to articulate the business problem and how the project conne**

the company's broader strategic objectives: The project you were wo on had a specific problem statement, of course, but its implementation w have helped fulfill an overarching strategic objective of the company. W you answer what the project was about, connecting it to the comp objectives, you show that you have an understanding of the bigger pi and the larger impact your specific project has on the company's goals.

- **Interaction among teams**: Even an individual contributor role requir some extent, collaboration with people across teams. Your work may involved working with the business to understand the problem stateme detail, with the data engineering team to get the data, or with the end us test the intuitiveness of the results. Demonstrating your ability to work people from different teams displays interpersonal skills, a key in mak successful data scientist.

- **Ability to communicate impact of analytics solution**: To achieve b from the appropriate stakeholders, it is imperative to be able to quantif impact of the analytical solution and communicate it across the board. becomes key when driving adoption of the solution you're working on.

Depth:

- **Approach to solution building**: A great candidate is able to talk throug different options thought of and evaluated before one is selected. W explaining the approach, you should share details on any requiremen constraints taken into consideration.
- **Technical details of the solution**: This question can be a great segue fo interviewer to dive into some of the technical details of the project example, if you built a predictive model as part of the project, be prepar answer questions about the model: How does it work? What parameter you tune? How did you evaluate model fit? Is the model scalable?
- **Data used**: For any analytical role, you must have a detailed understandi the data involved in solving the problem. The interviewer could go details and ask questions along the lines of, what data was used to cor the analysis? Why that data specifically? How did you do the data collec Did you face any challenges when collecting the data?

- **Ability to synthesize technical information for a business audience**: I able to explain the technical nuances of the project in simple layman's t is important. While the interviewer will look for you to explain the workings of the models, the ability to explain things to a nontech audience is equally important.

Framework to answer the question

Now that we have a fair understanding of what the interviewer is looking for, let's take a stab at how best to answer this question. I have put together a rudimentary framework you can use for this. Feel free to use this framework as a baseline and customize to your specific projects.

- Brief overview of the project in two to three lines

– What was the problem statement?
– What was the solution?
– What was the impact?

- Deep dive into the problem statement

- Approach to problem-solving

– What were the various options available to solve the problem?
– What option was chosen and why?

- Deep dive into the technical details of the solution

– What were some of the challenges faced during solution implementation?

- Impact created and ROI calculation

– State the impact and ROI from the project
– Walk through how this was measured and calculated

Practice Question

Interviewer: Can you tell me about a project you worked on?

Candidate: Sure. In the last year, I built a dashboard as part of the larger workforce analytics project to track the key metrics around incoming job applications. The goal of the project was to increase efficiency and effectiveness of talent acquisition. With the insights generated using the dashboard, the recruiting managers were able to track against the critical metrics and identify opportunities for improvements. The talent team was able to reduce the time to hire by 15 percent and the cost to hire by 22 percent over the course of the next quarter.

I was primarily involved with building of the dashboard in PowerBI as part of this project. I worked closely with our talent teams to understand the use case requirements, and in the process, I gained an in-depth understanding of the data available to build to dashboard.

Interviewer: That sounds excellent. Can you walk me through what data you used to build the dashboard and the key metrics you tracked?

Candidate: Absolutely. Once I had an understanding of the dashboard requirements, I worked with our IT team to find the relevant sources of the data. Let me first walk you through the metrics I tracked and then dive into the data sources if that works.

Interviewer: Sure, no problem.

Candidate: As the overall goal of the project was to increase efficiency and effectiveness of the talent acquisition process, I broke down the hiring outcomes in the following way:

Efficiency

- Average cost to hire
- Average time to hire
- Applicants to Interview Invites ratio

- Interview to offer ratio

Effectiveness

- Average retention rate over six and twelve months from hire

- Average individual performance rating over six and twelve months from

The efficiency metrics were geared more toward how the talent acquisition process works, and tracking them over time will provide insight into how the process can be improved. The effectiveness metrics looked at how successful the hiring process was after the event.

Furthermore, I provided a drill down into each of these metrics by recruitment source. For example, we found that the cost to hire via referrals was on average 15 percent lower than other sources such as LinkedIn or external recruiting agencies. The retention and individual performance rating for employees coming in via referrals was also higher compared to the other sources.

Another drill down that I added was to show these metrics by different business units. Using this drill down, we found that the number of referrals was very low for one business unit compared to the others. This alluded to the employee satisfaction being low for that business unit and was an actionable insight for our management.

Now, coming to the data sources, I primarily used the data within our Human Resources Management System. We had tables that stored employee information, hiring source details, pipeline and budget allocations, and interview process details. I also merged in data from tables that had performance ratings for all employees.

I organized meetings between our talent teams and the IT team to ensure I was mapping the right data source to the appropriate metric. These open conversations allowed us to be on the same page.

Interviewer: Great. I think you mentioned earlier you used PowerBI to create the dashboards. Any specific reason for the choice of that tool versus others?

Candidate: That's a great question. So, during the initial discussions, we looked at what the different options were from a visualization perspective with Tableau, QlikView, and PowerBI being the primary ones given their ease of implementation and ability to create interactive visualizations. The primary reason we were in favor of PowerBI was its low cost and company-wide access, as it is part of the Microsoft 365 Enterprise suite. Although tools

like Tableau are excellent in handling large amounts of data, for our use case we were typically dealing with only about one hundred thousand rows, so processing large amounts of data was not a concern. PowerBI, like other tools, provided the capability to directly link to the MS SQL Server tables to access the data for the visualizations. Essentially, we were able to make use of all the visualization functionality for our use case at a lower cost, and that drove the decision in favor of PowerBI.

Interviewer: That sounds good. Thank you for walking me through this project that you worked on. Sounds like this was well received by the people in your company and was a success.

Candidate: Absolutely. We were able to launch the dashboard for the talent team to access within two months, and the dashboard was very well received by the business overall. In fact, because of this project, our analytics team got exposure to the HR data present in the system, and I recommended the use of machine learning techniques to predict employee churn based on factors such as performance rating, appraisals, employee satisfaction surveys, and leaves taken. This idea was liked by the business and HR teams, and my team and I are currently working on its implementation.

Interviewer: That's very well thought out and will add value to the organization. Thank you for your time today.

Comments: Choosing a project to talk about is a mix of what you feel confident about and the type of role you're applying for. Note in this case, the candidate chose to speak about a data analysis and data visualization project. This would be a good choice for a role that requires these skill sets and also requires you to liaise between different teams. Remember that not every project you talk about needs to be machine learning driven or predictive modeling type of work. You can use examples where you have done extensive data analysis, gathered requirements to define analytics use cases, or built data visualizations to give strategic recommendations. And these are just a few examples—there's definitely a lot out there in terms of the different data science related projects that you can talk about.

If you're right out of college or if this is the first job you're applying for, you can talk about a project you did as part of your coursework or something you

self-initiated and did outside of school.

Differentiator: Notice how the candidate gave a quick view into the recommendation they made to the talent team to leverage data analytics for their work further. Time permitting in an interview, you can always add what you think could bring further value to the team/company based on your work. This will showcase to your interviewer that you continually think of the next steps, are able to think of the bigger picture, and give recommendations based on your analysis.

Expect to be asked this question in an interview in some shape or form. Prepping at least two to three projects you can talk about in depth during your interview will set you up for success.

CHAPTER 9

PRESENTATION INTERVIEW

The ultimate goal for data science is to help make better business decisions. To enable this, the data scientist must convert data into actionable insights and communicate the results of their technical analysis to an executive audience in a concise manner.

When I spoke to interviewers about how they were testing a candidate's ability to communicate, the topic of presentation interviews came about. Based on my conversations with interviewers across companies, I found that presentation interviews are currently popular among product and consulting companies.

They typically work in two ways:

1. Interviewers allow the candidate to choose a project of their choice OR
2. Interviewers give a specific problem statement they want the candida work on and present the results for.

In the first scenario, you are primarily expected to present a past project from start to finish, including the approach you took to solve the problem and the findings that came out of it. In the latter case where a problem statement or case prompt is given to you, interviewers will often share with you a supporting data set. You will use this data set to answer the questions stated and present your key findings.

Presentation interviews usually happen toward the end of the interview process. This means you can expect it after technical screening rounds are done and before behavioral interviews occur. These interviews are forty-five to sixty minutes in duration, with the first fifteen to twenty minutes reserved

for the candidate to make the presentation. The next fifteen to twenty minutes are slotted for the interviewers to ask you questions on your presentation, and the remainder time is for candidate's questions to the interviewer.

What is an interviewer looking for?

You will notice that acing a presentation interview is in some ways similar to answering the question, "Tell me about a project you worked on," which we discussed previously. The similarities are in what the interviewer expects to see from the candidate, and this is applicable for both scenarios in which you either choose your own topic for presentation or you're given a specific question to answer.

Let's quickly recap the themes we looked at in the last chapter. These speak to the interviewer expectations for the presentation interview and also for the "Tell me about a project you worked on" questions:

- Ability to explain a project you worked on from start to end
- Ability to articulate the business problem
- Approach to problem-solving
- Depth of technical knowledge
- Ability to synthesize technical information for a business audience
- Ability to communicate impact of analytics solution

How do you prepare for a presentation interview?

With a handle on what the interviewer expectations are, let us now move on to a few pointers that will be helpful as you prepare for a presentation interview.

- **Tell it like a story**: Describing a project you have worked on is best when you look at it like narrating a story. Irving Wladawsky-Berger in a *Street Journal* blog emphasizes the importance of storytelling in bus "Good storytelling is particularly important when introducing a comple potentially disruptive offering in the marketplace whose value is not understood. It's a natural way of explaining what the new innovation about."[22]

What Irving says is especially true for data science projects because the inner workings of machine learning techniques can be quite complex, but they can also be a differentiator in bringing value to the business. To encourage adoption of these techniques, the impact from such projects needs to be communicated effectively. And the skill to articulate the complex nature of data science is exactly the skill any employer is looking for when hiring a data scientist.

- **Tailor it to your audience:** What you narrate largely depends or audience you are speaking to. For the presentation interview, under from the recruiter who the audience is going to be.

In most interviews, you can expect a panel that consists of a technical and a nontechnical audience. Your nontechnical audience primarily wants to understand what the business problem was and what value you created or impact you made by implementing your solution. Your technical audience will go into the details of the approach you used and the chosen solution.

Collectively, they want to see if you are adept at communicating what you worked on from start to end. This is especially important because as a data scientist, you will talk to people from different teams on a regular basis. These people could be anyone from the leadership, data engineering team, fellow data scientists, developers, end users, etc.

- **Make it visual:** Edward Tufte explained the value of good visuals per when he said, "Graphical excellence is that which gives to the viewe greatest number of ideas in the shortest time with the least ink in the sm space."[23]

Here's the thing, though: Putting together good visuals is easier said than done. But investing time in them is generally worth it! Visualizations tend to catch your audience's attention quickly and also allow you to creatively share the key messages you want to highlight in a presentation.

From a data science perspective, you can make use of visualizations to show data/insights with the use of charts, histograms, heat maps, plots, word clouds, etc. These allow you to organize your data better and also show more information in a smaller space.

- **Anticipate the questions:** Once you have laid out your story board, fli hat you are wearing and think from the interviewer's perspective. If you the interviewer, what questions would you ask to understand the details project? While this is not an exhaustive list, you can expect questions the lines of:

- Why did you choose this methodology specifically?
- Did you consider any alternative approaches? Why or why not?
- If you were to do this project again, what would you do differently?
- What were some of the key challenges you faced when working on th project?
- What was the response you got when you presented the key findings fro this project?
- Can you explain the analytical technique used to an executive audience?

If you have some extra time on your hands, putting together a couple slides that speak to the anticipated questions could be worthwhile. You can add these slides in the appendix and bring them up in the conversation if needed. This shows that you have thought through multiple facets of the project and will set you apart from other candidates.

- **Do a practice run:** Once you have the presentation ready and you've w on the anticipated questions, it's time to do a mock run! The best way to is to present to a friend live and have them ask you the list of question have identified. See how you do in the practice run and improvise from Get feedback from your friend and identify any gaps in the presentatio can address. A useful tactic here could be recording this conversatio playing to see for yourself how you did.

If your friend happens to be in the field of data science, have them roleplay and wear the hat of a technical and a nontechnical audience. If your friend isn't from a data science background, have them probe your understanding of the business problem and the value you created with your solution.

Doing this even once will set you up for success in the presentation interview.

How do you structure and story board the presentation?

Now, let's come to the meat of this chapter! How would you structure and story board your presentation? Below, I have provided an elementary guideline you can use when you put together the presentation.

- Executive summary

– What was the problem statement?
– Why was this important to the business?
– Broadly, how was the problem solved?
– What value did the solution bring?

- Deep dive into the company and problem statement

– What does the company do and what is its business model?
– What is the exact business problem the company is facing?
– What are the different parameters involved in the problem?

- Solution overview

– Was there a need for an analytical solution and why?
– What are the different options available to solve this problem?
– What type of analytical solution was used and why?
– What are some of the technical details behind the solutio implementation?
  - Highlight any special considerations

- Value created

– What was the value created?
– How did you measure the impact?
– How did this translate to hard dollars for the company (if applicable)?

In situations where the impact wasn't measured, state the approach in which it could have been done. This can be applicable for a presentation where the

company gives you the problem statement to solve for. It's always good to show your thought process on how you would do it if you had all the information needed. Feel free to make and state any assumptions while doing so.

The guidelines in this chapter will put you in a great spot, but the key is to practice so you can sound like the pro you are in your interview!

---

22 Irving Wladawsky-Berger, "The Growing Importance of Storytelling in the Business World," *The Wall Street Journal*, March 17, 2017.

23 Edward Tufte, *The Visual Display of Quantitative Information,* (Cheshire: Graphics Press, 2001).

CHAPTER 10

TAKE-HOME EXAM

Let's say you just got off the phone with the recruiter about an open data science role. It looks like a great fit based on your discussion, and now you're awaiting the next steps when you receive the following email:

"We are excited to move further in the interview process with you. The next step in the process is a data science assessment that you will take remotely. Please find the instructions enclosed."

So how do you now prepare for it?

In this chapter, I will walk you through what to expect in these take-home exams and give you tips on how to crack them.

Take-home exams are a popular screening technique and sometimes are used even before you have a live interview with someone from the company. They are often part of the technical screening process of the interview and are designed to mostly mimic the type of work you may be expected to perform. The company will likely evaluate you against the key skill sets they need in their data scientist.

The format of these varies from company to company. Based on my research, I have found the following two formats popularly used:

1. **Online assessment**: Depending on the company, you may be aske complete a coding challenge on interview platforms like HackerI CoderByte, and Code Signal. These tests vary in duration but typically between one to three hours, and you have the flexibility to choos programming language you want to code in. Often, the platform tests against the completeness of the code, correctness of the result, and the s

at which you developed the solution.

2. **Custom take-home assessment**: Some companies may ask you to comp take-home data science project. As part of this, they usually give y broader business statement to tackle and an accompanying data set to u solve the problem. These projects can be completed from anywhere bet six to ten hours. Again, most companies give you the flexibility in cho the programming language you want to work in.

What is an interviewer looking for?

I spoke to a couple of senior data scientists and when it comes to evaluating performance, this is the rubric they mentioned the most:

- **Problem-solving skills**: This is one of the key elements of the rubric you be tested against. The evaluator wants to see your ability to comprehen problem statement, break it down into smaller tasks, and develop a sol and recommendation.
- **Code-writing skills**: These exams are designed to test whether you hav basic code-writing skills expected of a data scientist.
- **Speed**: In an online exam where tracking of completion time is possible may also be evaluated on the speed with which you can code.

- **Completeness**: In the scenario that you aren't able to fully complete the you may be evaluated against what portion of the test you were ab complete.

***Pro tip*:** In cases where getting through the whole challenge appears difficult to you, address the key questions first and tackle secondary questions later.

What should you expect in a take-home exam?

If you're asked to complete such an exam, you can typically expect a few tasks types:

- **Query language**: You may be asked to perform data operations like jo two tables, aggregating key columns to generate metrics, and other operations such as filtering, sorting, case when, and if/else statements. S tests may look at your know-how of advanced SQL techniques like u window functions in SQL.

- **Predictive modeling**: You may be asked to build a predictive model, e classification or regression in most cases, and be asked to score a test data
- **Clustering**: You can be given an unsupervised learning problem and be a to develop clusters that are statistically sound and that have the de business application.
- **Probability and statistics basics**: In some cases, you can expect to see n choice questions testing you on probability and statistics basics. probability these could be around conditional probability, Bayes' theo etc. For statistics, these could be around p-values or z-scores. I have prov sample questions and answers on these topics in Chapter 5: Probal Statistics, Experimental Design.

- **Data interpretation**: I have seen product and consulting companies ask interpretation questions, especially if the role has a heavier "data anal component to it. You may be given a raw or summarized data set ar asked to answer specific questions or to generate insights from the dat Most of these questions are posed in the business context and may re you to make a formal recommendation to the business on the problem y asked to solve.

How should you prep for take-home exams?

Here are a few things you can do to prepare before you are even asked to take an assessment at home:

- **Create an often-used code library**: You can build out basic functions programming language of your choice so you have them handy when have to work on a take-home project. If you're assigned an online assess and aren't allowed access to any outside material during the exam, bui out these functions can be great practice and will help you code faster d the test. Here are a few functions you can consider building out:

  - Impute missing values
  - Plots to do exploratory data analysis
  - Building out summary statistics on a data set
  - Train and score a classification and regression model

- **Practice coding exercises**: You can practice coding challenges on o platforms like CoderByte, HackerRank, LeetCode, and CodeSignal. Mo these online platforms offer free exercises for you to practice. Doing this ahead of time will help you evaluate where you are when it comes to tac these tests and address any gaps you see beforehand. It is also very help familiarize yourself with at least a couple of these platforms so that you k how you can write, run, test, and debug the code on these websites. You lose time learning to navigate the test environment if you see it for the time on the day of your test.

Pro tip: Before you are given the assessment, check with the recruiter on what format you can expect the areas you will be tested on and the rubric you will be scored with. Your recruiter may have already given you this information, but in case they haven't, you can always ask this question so you can adequately prepare for it.

To ace these tests, you should practice beforehand and make sure your fundamentals in programming and data science concepts are solid. Although they may seem daunting initially, with sufficient practice, you will be able to ace them.

Extra questions for practice:

- You have been given a data set containing five columns titled A, B, C, with numeric data that contains test scores corresponding to a subject scores range between zero to one hundred, and each row corresponds student. With this data, build clusters of students that show si performance on tests. The clusters created will be used by the teache understand the strengths and weaknesses of the group and build custon teaching plans for each.
- You are given stock prices for three stocks over the last two years. Predi stock prices for the next month. The data is at a daily level with the price at close of business.

- You have a table named "Squares" given to you that has four columns. column contains a number that corresponds to the length of a side. another column to this table called "Flag," which is a bo

(TRUE/FALSE). The column “Flag” will be true if it is a square. Note t all four sides are equal then flag as true.

CHAPTER 11

BEHAVIORAL INTERVIEW

A behavioral interview can be a free playground to test the candidate on the skills that are core to the company. Put yourself in the shoes of the interviewer and you will understand their thought process. The interviewer is probably thinking:

- Do I want this person on my team?
- Can I rely on and trust this person?
- Is this person a good team player?
- Can this person start contributing to the project straight away?
- Can this person handle stressful situations and navigate conflict?

Now, let's discuss how you can best prepare for this kind of interview.

What is an interviewer looking for?

Behavioral interviews are quite common for a lot of job roles, but your interviewer may be looking for a few specific things that are key to the data science field. Here's a breakdown of the rubric you may be tested on:

Reactive versus proactive: When given a task, do you just do what you're told, or do you bring in your point of view to the job and enhance the solution? To illustrate this point further, let's look at the following conversation.

Joon Lim, a data science leader at LinkedIn, spoke to the students of Northwestern University as part of Eunhee Ko's class. Here's a snippet of that transcript that lays this out beautifully.

Question asked by a student: *I am curious to hear more about the difference*

between junior and senior data scientists. What's expected of a junior data scientist (DS), and what makes senior data scientists senior?

Joon's response: *You may be expecting answers like a more in-depth understanding of ML algorithms such as deep learning, more effective partnership management, and more engaging presentation. These technical skills and soft skills are all important, but I believe the real differentiator is the ability to make proactive contributions. Junior DS tend to be reactive to business partners. The partners make prescriptive requests to a junior DS who just focuses on delivering the result.*

Business partner: *"Hey Joon, could you help check how many LinkedIn members are seeing this promotion for a week?"*

Junior DS: *"Sure, here is the number that you are looking for: 25M."*

Business Partner: *"Thank you."*

Joon's comment: *Junior DS don't bother to understand the motivation behind the question and tend to quickly jump into the conclusion. Some argue this is effective in building trust and relationships with partners. I don't disagree, but this type of contribution is limited as you assume the prescriptive direction that your partner outlined is the most effective in solving the business problem at hand. Instead, senior DS would seek to understand the motivation and think together on the best way to achieve the fundamental objective.*

Business partner: *"Hey Joon, could you help check how many LinkedIn members are seeing this promotion for a week?"*

Senior DS: *"Sure, I can help you get this number, but first, could you help me understand how you plan to use this data for?"*

Business partner: *"Oh, I think the promotion targeting criteria is too stringent."*

Senior DS: *"Okay, then I'd suggest looking into the underlying eligibility distribution instead, and we can check whether it makes sense to move the threshold. But if we loosen up the criteria, we may face potential side effects such as increasing member complaints, unsubscriptions, turning off the promo setting, etc."*

Business partner: *"Got it, that makes sense to me. Could you help find the sweet spot?"*

Senior DS: *"Of course, I should be able to get back to you by this Friday."*

Business Partner: *"Awesome, thank you."*[24]

When framing your answers, focus on being proactive and demonstrate how you added value on top of what was expected.

Culture fit: Culture fit is a two-way street. The organization you want to work for should be a fit for you as much as you should be a fit for them. Culture fit questions can revolve around the following:

- Why do you want to work for our company?
- How do you see yourself fitting into our company?
- How do you handle stress or tight deadlines?
- How do you handle failures?
- Would you rather work alone or with a team?
- How do you like to lead and delegate?
- How do you like to be managed?

These are a few questions you can expect in the culture fit assessment. But the idea here is to see if both parties involved are compatible with each other for a successful stint. It's just as important for *you* to be happy with the company as it is for the company to be happy with you.

Conflict resolution and stressful situations: The interviewer may ask you to describe situations where you had to resolve a conflict or work through any times that were stressful on the project. While there may not be an exact right or wrong answer to these types of questions, the interviewer is looking to understand your working style and if it aligns with the company's overall culture.

Stories to prepare

When preparing for behavioral interviews, you will set yourself up for success if you prepare key examples/instances, or "*stories*," as I prefer to call them. Depending on the question, you can pick the *story* to highlight during your

interview.

While I am categorizing them under wide umbrellas here, stories under these headlines can be customized to many different questions asked in a behavioral interview. The key is to have this list handy so you have a strong base to go in with for your interview.

Describe an instance when you…

- led a team
- handled and resolved a conflict
- went out of your way to make a project successful
- went outside of your comfort zone
- faced a challenging situation on one of your projects
- failed and how you adapted to the situation

Approach: alternative what-how technique

To answer behavioral questions, I suggest using the alternative what-how technique. Let's walk through the details below:

What was it?

Briefly describe the project, problem statement being solved, and parties involved. Think of this part as setting the stage for your answer.

How was it done?

In this part, describe the details of the task and the work done to complete it. State any other alternatives that were taken into consideration, if applicable. This is a great way to set yourself apart by showcasing that you evaluated other options before finalizing the approach you took.

What did you do?

Once you have set the stage and described the task at hand, walk the interviewer through your specific contributions. This is where the "I vs. We" differentiator comes into picture. While the full team's contributions are key to make any project successful, your specific part in it is what you need to highlight.

How did you bring value?

Round up the answer by summarizing the impact the project made and how your work brought value to the team and the organization.

Let us now look at a practice question.

Practice Questions

Practice Question #1

Interviewer: Can you tell me about a time when you were asked to complete some work in an accelerated time frame than what was decided earlier?

Candidate: Sure. In my current role, I am part of a centralized data science team, and I work with different functions within the organization on data-driven projects. The instance that I want to share with you is about a project I was working on with the marketing team.

We were working on an analysis leading to an optimization solution that allowed the marketers to optimize digital marketing campaign spend to target the right audience. This in turn would optimize the organization's marketing budget. The full project, as such, was scoped and the time frame based on the scope decided was two months.

However, a week into the project, our marketing team reached out to us and told us to deliver this project within a week from then. This was quite accelerated given the scope of work involved. Although my immediate reaction was that what they are asking for is just not feasible, I took a step back, tried to understand where they were coming from, and then set up a conversation with them to understand the need behind the request for accelerated delivery.

From that conversation, I understood that the business was pushing to close campaigns on a particular channel, and it wanted to use our solution to gather the data/insights it needed to counter this decision if necessary. Now, that made sense to me. I explained to the business that we could pull the data and help it with the analysis it needed to determine the impact of pulling the campaigns from that channel within a week. And I proposed that we keep this as the short-term goal for the project and retain the original scope and timing

for the larger optimization project.

The business agreed to this, and I was able to work with the marketing team to understand their detailed requirements and then with the data engineering team to track the exact data sources we needed to hit to pull the data for this analysis. We collaborated across teams in an agile fashion and delivered the analysis to the marketing team. We then put forth a recommendation to the business of not pulling that campaign channel down as the impact on revenue would be considerable, and the ROI from the channel wasn't as low as the business suspected.

***Comments**:* Overall, the candidate did a good job in demonstrating how they delivered value in an accelerated time frame and how they navigated the ask from the marketing team. The candidate was able to break down the task into what was an immediate priority versus long term and promptly delivered what was imperative to the business. The candidate was also able to show their ability to collaborate and communicate effectively to get things done.

Practice Question #2

Interviewer: Can you tell me a situation where you handled conflict at work?

Candidate: Sure. On a recent project, I led a team of three junior data scientists. The project had tight deadlines and very specific deliverables tied to each milestone. Each of us was focusing on a deliverable and tracking against the timeline in place. However, for one of the milestones, a deliverable that my teammate was working on wasn't ready, and there was a potential that we would be late.

I checked in with my teammate and asked why the deliverable was running late. They did not receive that question well and got very defensive. I quickly realized my question could have been worded better to make them feel more comfortable to talk about the issues at hand. We decided to take a small break from that conversation and talk again in a few minutes.

I took the time to collect my thoughts and then met with my teammate again. When we spoke at length, they told me they had other competing priorities with similar tight deadlines and that is why they were running late on the project we were working on together. I sat down with them and made a list of

all the items they were working on and met with the team lead of their other project. Together, we were able to shuffle a few items and reorganize their priorities without impacting any deadlines.

I took up a few tasks from their list and had my other team members lean in too so we could collectively cover ground.

Interviewer: Sounds like you handled the situation well. Were there any learnings from this that you used moving forward?

Candidate: Yes, in fact I updated our planning process so that our full team has visibility into all of the projects we are working on and how much bandwidth each team member has by week. This allowed me to prioritize task items better and reallocate as necessary.

Comments*:* If you are interviewing for a managerial role, the interviewer will want to know how you lead your junior team members and how you deal with any challenging situations that crop up as part of being a lead on the project. Notice in this question how the candidate effectively managed to continually improve as a leader while being receptive to the input from their juniors.

In conclusion, when you are asked a question in a behavioral interview, take a few minutes at the outset to identify an example that would be the most fitting for the question asked. With that example in mind, lay out the structure in which you want to talk through the story. Taking some time in the beginning will help you navigate the question well and will lead to meaningful conversation with the interviewer.

Extra questions for practice:

- Tell me about an instance where you solved a challenging problem.
- Have you worked under pressure or a stressful situation in the past? Hov you handle it?
- Tell me about an instance where you failed at something, and how dic navigate that situation?

- Has there been a time when you pitched an idea to senior leadership what was the outcome?

24 Eunhee (Emily) Ko, "A Conversation With Joon Lim About Data Science At LinkedIn," LinkedIn.com, accessed June 07, 2020.

PART 4

PUTTING YOUR BEST FOOT FORWARD

CHAPTER 12

CRAFTING A DATA SCIENCE RESUME

Your resume is a snapshot of who you are, what you have done so far, and what potential you have to offer. When you're applying for jobs, you don't want your resume to get lost in the mythical black box that never gets opened after the first read. You may be an excellent data scientist with the best git repository of analytics solutions, but if your resume isn't up to par, then you may not get a call back from the company you're applying to.

Cracking an interview gets you the job, but an excellent resume is what gets you in the door!

In their book *Cracking the PM Interview*, Gayle Laakmann McDowell and Jackie Bavaro speak about the 15 Second Rule. They say, "A resume isn't read; it's skimmed. A resume screener will glance at your resume for about 15 seconds (or maybe less) to make a decision about whether or not to interview you."[25] This is exactly why a resume is crucial and the first step you need to get right when applying for jobs.

Pro tip: My biggest advice when creating a resume is to objectively take some time out to build it. By investing this time early on, you can, in one shot:

- Create a solid and favorable resume
- Prepare for a "resume-based" interview

How do you do these two things at once? You can do this using what I call "The Hundred-Word Story."

The Hundred-Word Story

The Hundred-Word Story is a framework I put together that will help you naturally build out a resume and prepare for interview questions based on it.

At some point in the interview process, you can expect questions on the projects that you have mentioned on your resume. Instead of spending additional time prepping for those, with The Hundred-Word Story framework, you can do both at once.

You can use the framework by following the steps below:

Step 1: List out each project that you have done in the past that you think is a good candidate to talk about in any interview or to go on your resume.

Step 2: Write an approximately hundred-word story on each project, answering the questions below:

- What was the project's overarching objective?
- What was your role and contribution on the project?
- What impact was made by the work you did?

Step 3: Write another hundred-word story answering the questions below:

- Why did you choose the approach that you used to solve the problem?
- What were the major challenges you faced during this project?
- Is there anything you would do differently now that the project is done?

If you do these above three steps correctly, you will not only prepare solid content for your resume, but you'll also have material handy for any resume-based questions that pop up in an interview.

Let's walk through an example of this so we can see how it plays out.

Hundred-Word Story #1

The marketing team within our company wanted to build a positioning strategy for one of the cosmetic products that our company sells. They wanted to do this to drive revenue growth for that line of product. I worked

with the marketing and IT teams to develop the customer segments. Following this, I identified the most profitable target segment for that product. The marketing team then used the insights from the segmentation to position the product in the market. The new and improved positioning of the product enabled a potential market gain of $1.2 million over the quarter after it was launched.

Hundred-Word Story #2

I chose to segment the customers using the k-means clustering technique and used the elbow curve to find the optimal number of segments. I also tried the agglomerative clustering technique, but I found more meaningful results using the k-means technique. I compared the mean and standard deviations of each cluster across the two techniques to assess their performance. We used survey data to create the clusters, and designing meaningful features from those was a bit challenging. In retrospect, I would involve the business stakeholders from early on, in this case the marketing team, so their input into feature engineering and result validation could be built into the process and not just taken at the end.

Now that we have our two stories in place, let's look at how we use them.

The first story is what you will use to craft a bullet on your resume. Let's look at a very basic way to do this and then at an improved example of the same thing.

Basic Example:

- Developed customer segments using k-means techniques to help marketing team position a product. Segmentation was done in Python.

New and Improved Example:

- Enabled a potential market gain of 1.2 million by improving the positio strategy of a cosmetic product. The new product positioning was drive identifying the most profitable segment. Customer segmentation devel was based on survey data and involved use of k-means clust implementation in Python.

Comments: Notice how in the basic example, the candidate leaves out the impact made by their project. When you lead with the impact or at least mention it as part of your bullet, you demonstrate that you have an understanding of how your data science work drives business decisions.

Let us now look at a sample resume to get a full picture of what a data science resume can look like.

Sample Resume

Jane Doe

Education

North Carolina State University
M.S., Analytics (2010–2012)

San Francisco State University (2005–2008)
B.S., Computer Science

Work Experience

Cosmoline (2012–Present)
Data Scientist

- Enabled a potential market gain of 1.2 million by improving the positic strategy of a cosmetic product. The new product positioning was drive identifying the most profitable segment. Customer segmentation devel was based on survey data and involved use of k-means clust implementation in Python.
- Decreased customer churn by 3 percent by building a model to isolat factors causing churn so immediate action could be taken to inc retention. Churn model was built in R based on logistic regression techı and PCA for feature reduction.

- Cut down call duration for the customer care team by 27 percer identifying the most commonly asked questions with the use of text mini

Ziofy (2008–2010)
Product Analyst

- Coordinated with the data engineering and product teams to design and a metric tracking dashboard in Tableau for the launch of a new pro

feature. Streamlined process to track metrics for other future feature laur saving 2-week analyst work effort each time to do this process.

- Analyzed external data and provided recommendations on compe trends. Insights generated were leveraged to reduce loss of custome competitor offerings.

Awards and Activities

- Outstanding Performance Award during the Capstone Case Study Proje Corporate Development Program, Ziofy.

- President of quarterly data science hackathon at NCSU that inv participation from 100+ students from computer science, analytics, indu engineering, and other departments. Event raised $5000 in prize money sponsorship from tech giants including Facebook, Google, and Microsoft

Languages and Tools

- Programming Languages: Python, R, SAS, Java, SQL
- Database: Experience working with Microsoft SQL Server, MySQL, O Sybase

- Visualization tools: Tableau, QlikView, PowerBI

Additional Information

- Kaggle rank 87 as of Dec 2019

- Active blog contributor for "Towards Data Science" on medium.com

Your resume is a gateway to any job you want, and this makes it one of the most important things you will work on in the entire job application and interview process. Hence, take the time to build a solid resume and make sure to use The Hundred-Word Story technique we discussed in this chapter. This way, not only will you have a strong resume, but you'll also be sufficiently prepared for any interview questions based off of your resume. And as always, get it reviewed by a friend so you have a second set of eyes to give you 360-degree feedback!

25 Gayle Laakmann McDowell and Jackie Bavaro, *Cracking the PM Interview: How To Land A Product Manager Job In Technology* (Palo Alto: CareerCup, 2013).

CHAPTER 13

DATA SCIENCE PORTFOLIO

If you talk to any artist or creative professional, you will realize that the majority will showcase their work with the help of a portfolio. A portfolio is nothing but a collection of work samples that documents their professional accomplishments and offers demonstrable proof of their work. It's a great way to showcase conceptual, creative, technical, and presentation abilities.

Why build a portfolio for data science?

"Now, why consider building a portfolio for data science?" you may ask, and rightly so. If you set aside the technical and mathematical aspects of data science, then it is as much art as it is science. Weeding through rows and rows of data to find something substantial needs scientific prowess, no doubt, but this skill also requires a certain creative ability to ask the right questions and tease out the insights.

Working on a canvas and seeing it evolve day by day is analogous in many ways to the work of a data scientist. Let's say, for example, you are building a machine learning model and start by exploring your playing field, i.e., the data. You work through the variables to see what makes sense to include in the model and iterate on different features and modeling methodologies before you paint the final picture. The whole process is very discovery-based, and a portfolio in some ways does justice to demonstrate in a tangible fashion the thought process and the amount of work going behind building the final product.

Another strategic reason to build a portfolio is because it allows you to play to your strengths. Data science as a field in itself is so varied and has wide-ranging skill sets in demand in the market. You may be someone who writes efficient, flawless code designed to be scalable from the get-go or you may be

someone who can narrate complex analytical findings in a simple visual fashion. With the help of a portfolio you can showcase where in the large scheme of things your specialty lies and in what areas you can contribute from day one.

Who can/should build one?

Anyone in the industry can build and use a portfolio to showcase their skills. It's an excellent way to differentiate yourself from your competitors when applying for a job.

If you fall into any of these three categories or a combination of them, you can benefit from building a data science portfolio:

- **You are looking for your first job**: Consider building a portfolio if y right out of college looking for your first job or if you have limited experience. A portfolio is an excellent way to put forth tangible products that speak to your credibility as a data scientist.
- **You are new to data science**: If you have a few years of work exper under your belt but it's not directly classified as data science currently can build a portfolio to showcase your potential in the field. Additionally will be able to strengthen the case for things you are already good at current job. For example, you are a software developer looking to move i machine learning engineer role. Having a GitHub account with ma learning projects to show can help bridge that gap and land you the de job.

- **You are an experienced data scientist**: If you are an experienced scientist looking for your next big opportunity, you can tactically curat different projects you've worked on in the past to share with pot employers. This will allow you to play to your strengths and get you i door of a company of your choosing.

What can you include?

When building a portfolio, show what you can do well—and what you can do extremely well—to make yourself stand apart. You don't have to tick every box when building a portfolio, but show the things you're really good at because what may appeal to one company may not appeal to the other. By

showcasing your strengths and being recruited for those, not only does a company find a good fit in you, you find a good fit in the company.

That being said, depending on where your strengths lie, you can showcase your skills using any of the following. Please note this is not an exhaustive list but rather a sample of different avenues that can be explored when building a data science portfolio.

- **Projects on GitHub**: Sharing your GitHub repository with concrete pr examples lends you certain credibility and shows that you have confidence to share and speak about the projects you have worked on detailed level. Your projects on GitHub can span a wide range of to including but not limited to exploratory data analysis, super or unsuper learning models, descriptive data analysis, and data cleaning.
- **Kaggle**: Kaggle hosts data science competitions and is often a great w practice solving data science problems. You can also see how you com against folks in the industry. You can leverage Kaggle projects to show rankings in the competition or as a way to share your approach to anal problem solving.
- **Dashboard/Visualizations**: Explaining complex analytical concepts in si visualizations that are intuitive to an executive audience is often challen If you're able to depict your data and insights in a visual fashion, show this ability in your portfolio. It is a crucial skill to have and is a quick w the eyes of a potential employer.
- **Blogs**: Another great way to get the word out about your skills and intere to blog. You can present your point of view to the world using your blog demonstrate thought leadership in the field. Or you can use your bl show your technical and nontechnical communication skills.
- **YouTube videos**: Often, teaching is the fastest way to learn something. I are passionate about a certain data science area, you can c training/informational videos on that topic and share. You don't have this for the sake of building a portfolio but if it is something you enjoy d then take a stab at it, and your portfolio will build itself.

- **Presentation**: Do you like making slides and telling a story? Put toget presentation of a project you worked on. Being able to narrate a story thr

slides shows your ability in written and visual communication.

In summary, a data science portfolio isn't mandatory for every candidate to have, but if you have the time and passion to build one, it can go a long way. It's an excellent technique to practice and showcase your skills. In the process of creating a portfolio, you might even build a network of people who have similar interests as you.

PART 5

LAST BUT NOT THE LEAST

CHAPTER 14

WRAPPING UP

Now that you have reached the end of this book, I hope you understand the different aspects of a data science interview and have a good handle on how to crack each round.

Before we wrap up, though, I want to share a few pointers about optimizing your job application process.

What to look for when applying to jobs

You should know a few things about the company you are applying to and the role you are applying for. This can be done by digging into three key areas.

1. **Company:** As you prepare for your interview, here are a few things should learn about the company.

   a. The company's culture, mission, and values.
   b. Its primary business and how its business model works.
   c. The key players in the organization.
   d. Latest news articles about the company.

2. **People:** The people perspective is more of a spin on how the teams withi company are structured. Typically, you might observe one of these organizational structures in place at most companies:

   a. A centralized data science team.
   b. Dedicated data science teams embedded within different functions.

Simply understanding how the teams are structured will give you a sense of

what to expect in the interview. For example, let's say the data scientists are embedded within the specific product teams. You can expect interview questions around that product, and you could potentially hypothesize the kind of cases they ask you and then study in depth the relevant analytical techniques generally used for working on those problems.

3. **Technology:** While the company aspect answers *who they are* and the p piece answers *who does what*, the technology piece will help answer *ho they do it*. Researching this bit will help you understand things like…

   a. Is it a Python shop, an R shop, or something else?
   b. Does it emphasize on certain skill sets like data visualization, storytellin presentation skills?
   c. How is its infrastructure set up? What technology should you be familia to work with its infrastructure?

Once you narrow down the company and role you want to apply to, try to get an informational interview with a current employee to learn more about the company culture and the work they do. You can do this by reaching out to someone you may know directly or by asking a mutual connection.

Knowing these things will not only help you be prepared for your interviews but will also allow you to evaluate if the company/role is a fit for *you*. At the same time, remember you don't need to have answers to all the questions I have mentioned above. If you can't find answers to some of these questions, ask them in your interview to learn more about the company and role. Based on my experience, some of these led to excellent conversations with my interviewers. It also gave me an insight into how favorable this job would be for me if I were to get an offer and accept it.

I say this to anyone I talk to when it comes to a job hunt—it's a two-way street. You should like them as much as they like you!

In closing

When it comes to cracking a job interview, no substitute exists for adequate practice.

Reach out to people within your network and ask them to conduct mock

interviews for you. Be it case interviews, coding challenges, or behavioral interviews, you will feel very confident in an actual interview when you have practiced a few times. Practicing with the use of mock interviews will help you tackle these situations with confidence.

In the context of identifying what areas to practice on, Eric Weber, GM of Experimentation & Data Science Leader at Yelp, gave me the following advice: "Candidates should push for details about what the interview process entails. Although this could be tough to do in a candidate's position, doing so gives you a structure to base your preparation on."

Leverage your recruiter to understand the role in depth and what to expect as part of the interview process. They are an excellent resource and want to see you succeed.

Lastly, remember to brush up on your fundamentals before any interview and know in depth about the projects you've worked on. You can capitalize on the sample answers given in this book and use them as a frame of reference when tackling a question in an interview. Even if you encounter a completely new question, you will be able to think on your feet using the frameworks and structure mentioned in this book.

I am confident that if you are passionate about data science and diligent in your preparation as recommended in my book, you will certainly be an outlier in the recruiting process and land the job that you want!

All the best. Stay in touch.

https://www.linkedin.com/in/shrilata-murthy/

ACKNOWLEDGMENTS

I started writing this book when I was pregnant with my daughter, Samyra. So, it is safe to say that I have never been alone when writing this book. And this has been true throughout the journey.

I have had all-around support from my family, friends, publishing team, and my beta readers during the course of writing *Be the Outlier*.

Writing a book is hard work. For every sentence you write, you wonder if it could have been structured differently. For every idea you discuss, you ponder if another one would have been more appropriate. For every chapter you include, you probe whether you said everything you wanted to say.

It is an intense and grueling process and one that is not possible without a solid support system. I deeply appreciate everyone who has supported me in this process.

To my best friend and husband, Ganapati. You are the pillar that always stood strong when I needed you. Thank you for always being there for me.

To my beautiful daughter, Samyra. One day you will be old enough to read this book and know how you were always next to Mumma every step of the way.

To my family, Shrikant Murthy, Vijaya Murthy, Mythili Payyalur, Shrideep Murthy, Rutuja Deshpande, Lakshmi Kollengode, and Anand Kuppaswamy. Thank you for always encouraging me to do my best.

To my data science support system, Alice Zhao, Andy Fox, Aniket Deshpande, Anirban Bhattacharya, Anthony Tockar, Arindam Paul, Biswanath Banik, Chris Pease, Diego Klabjan, Ehsan Rehman, Eric Weber, Gaurav Dhingra, Kyle Hundman, Laura Siahaan, Macario Lullo, Peter Schmidt, Samar Kaushal, Shantanu Raghav, and Shel Singh. Thank you for

being so willing to ideate with me and for providing excellent input. With your collective help, I was able to metamorphose all ideas into this very book.

To my publishing team from New Degree Press, Eric Koester, Brian Bies, Emily Price, and Jennifer Psujek. With your help, this book has come into being.

To my beta readers, I could not have done this without you. Your support early on in the process made publishing this book possible.

Abhishek Kedia

Ahsan Rehman

Ameer Khan

Ameeta Japtiwale

Amit Rao

Amol Khor

Amruta Pathak

Anand Kuppaswamy

Andrew Fox

Aniket Deshpande

Anirban Bhattacharya

Anisha Gurumurthy

Ankit Koradia

Ankita Nagori

Anupama Mangalvedhe

Aprit Arora

Apurva Limaye

Ashwini Senan

Biswanath Banik

Cinjal Shah

Corey N Paulish

Daniel Medina

Darryl Malcolm DCosta

Devon Weiss

Dr. Halasya Ramanathan

Eric Koester

Erika Enomoto

Faza Chugtai

G. Ragunathan

Ganapati Raghunathan

Garima Singh

Hema Chandra Chittuluri

Jake Whitesides

Jayanth Srinivasan

Kalyani Patil

Kapil Gadhire

Karthik Sastry

Kedar Guttikar

Ketan Patil

Ketki Gandhe

Kirti Khopkar

Krutika Vyas

Kushal Mahajan

Lakshmi Kollengode

Laura Siahaan

Macario Lullo

Manasi Mangalvedhe

Meghana Mangalvedhe

Miranda Zhu

Namesh Kher

Naveen Shankar

Neil Jones

Nikhil Khekade

Nikhil Nambiar

Nitin Mangalvedhe

Nisha Shetty

Nistarani Ramanathan

Nithya Hariharan

Northwestern University, Master of Science in Analytics

P.S. Mythili

Parul Deswal

Peter J. Schmidt

Pooja Sunder

Pratap Kaul

Preeti Ramaraj

Prithvi Mali

Rahul Goswami

Rajalakshmi Balan

Ravindar Kumar

Richa Kumari

Riddhi Bhatt

Rohit Bhangale

Ronak Parpani

Rutuja Deshpande

Salil Sabade

Sandesh Doddameti

Sanjeevni Wanchoo

Saurabh Jha

Saurabh Vishwas Joshi

Shantanu Raghav

Shel Singh

Shilpa Sreekrishnan

Shirish Gupta

Shivee Singh

Shraddha Hegde

Shreya Thacker

Shreyans Mulkutkar

Shreyas Tendulkar

Shrideep Murthy

Shrikant Murthy

Shruti Jalali

Sneha paul

Sneha Tiku

Soumyajit Mallick

Tanmay Shah

Umananthini Abhishek

Varsha Jagdale

Varshini Ramaraj

Venky Munivenkata

Vignesh Venkatachalam

Vijaya Murthy

Vikesh Chauhan

Vishal Pareek

Vivek Ajmera

William Chiu

Xiaowei Li

Yahya Cheema

Yugandhar Garde

To *you*, my reader. Thank you for choosing to read this book and joining me in the data science journey.

ADDITIONAL RESOURCES

Database Systems

- Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer D. Widom, Dat Systems: The Complete Book, 2nd Edition, Prentice Hall, 2009.

- Ramakrishnan, Raghu and Johannes Gehrke. *"Database manage systems."* 3rd Edition. New York: McGraw-Hill, 2002

SQL

- LeetCode. Category – Database. https://leetcode.com/problemset/databa
- W3School. SQL Tutorial. https://www.w3schools.com/sql/
- Select Star SQL. https://selectstarsql.com/
- HackerRank. Practice SQL. https://www.hackerrank.com/domains/sql

- CodeAcademy. Learn SQL. https://www.codecademy.com/learn/learn-sc

Python

- Python.org. The Python Tut https://docs.python.org/3/tutorial/index.html
- CodeAcademy. Learn Python. https://www.codecademy.com/learn/l python-3
- Udacity. Programming Foundations With Py https://www.udacity.com/course/introduction-to-python--ud1110
- Google. Google's Python Class. https://developers.google.com/edu/pytho

- PracticePython. Beginner Python Exercises. http://www.practicepython.

R

- Grolemund, Garett and Hadley Wickham. *R for Data Science*. Sebast O'Reilly, 2017
- Tidyverse. R packages for data science. https://www.tidyverse.org/

- Ismay Chester and Albert Y. Kim "*Statistical Inference via Data Scien ModernDive into R and the tidyverse*." Boca Raton: CRC Press, 2020.

APPENDIX

Introduction

Indeed Hiring Lab, US. "Data Scientist: A Hot Job That Pays Well." accessed May 5, 2020

Chapter 1

Cao, Sissi. "What on Earth Is a Data Scientist? The Buzzword's Inventor DJ Patil Spills All." *Observer*, November 9, 2019. https://observer.com/2019/11/data-scientist-inventor-dj-patil-interview-linkedin-job-market-trend/

LinkedIn. "LinkedIn: 2020 Emerging Jobs Report." accessed May 16, 2020. https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_U.S._FINAL.pdf

Chapter 2

Davenport, Thomas H. and D.J. Patil. "Data Scientist: The Sexiest Job of the 21st Century." Harvard Business Review, October 2012 Issue. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

LinkedIn. "LinkedIn: 2020 Emerging Jobs Report." accessed May 16, 2020. https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_U.S._FINAL.pdf

Chapter 4

Dummies. "The Importance of Clustering and Classification in Data Science." Dummies. Accessed May 17, 2020. https://www.dummies.com/programming/big-data/data-science/the-importance-of-clustering-and-classification-in-data-science/

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed May 17, 2020.

Metcalfe, Andrew V. and Paul S.P. Cowpertwait. *Introductory Time Series with R.* (New York: Springer, 2009), 81

Morde, Vishal. "XGBoost Algorithm: Long May She Reign!" Medium.com. Accessed May 20, 2020. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Chapter 6

Introduction to Probability and Statistics. *Conditional Probability, Independence and Bayes' Theorem Class 3, 18.05*. Spring 2014. Massachusetts Institute of Technology: MIT OpenCouseWare,

https://ocw.mit.edu/. License: Creative Commons BY-NC-SA.

Investopedia.com. "Compound Probability." Accessed June 5, 2020. https://www.investopedia.com/terms/c/compound-probability.asp

Math Goodies. "Conditional Probability". Accessed May 17, 2020. https://www.mathgoodies.com/lessons/vol6/conditional

MBA Crystal Ball. "Probability | Theory, solved examples and practice questions". Accessed May 17, 2020. https://www.mbacrystalball.com/blog/2015/07/03/probability/

McGinely, Patton. "Tests of Change: Simulated Design of Experiments in Healthcare Delivery". *Patient Safety & Quality Healthcare*, July 14, 2009. https://www.psqh.com/analysis/tests-of-change/

Statistics How To. "Conditional Probability: Definition & Examples". Accessed May 17, 2020. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/conditional-probability-definition-examples/

Study.com. "Probability of Compound Events: Definition & Examples." Accessed June 5, 2020. https://study.com/academy/lesson/probability-of-compound-events-definition-examples-quiz.html

Chapter 9

Tufte, Edward. *The Visual Display of Quantitative Information*. Cheshire: Graphics Press, 2001.

Wladawsky-Berger, Irving. The Growing Importance of Storytelling in the Business World." *The Wall Street Journal*, March 17, 2017. https://blogs.wsj.com/cio/2017/03/17/the-growing-importance-of-storytelling-in-the-business-world/

Chapter 11

Ko, Eunhee (Emily). "A Conversation With Joon Lim About Data Science At LinkedIn" LinkedIn.com. Accessed June 07, 2020. https://www.linkedin.com/pulse/conversation-joon-lim-data-science-linkedin-eunhee-emily-ko/

Chapter 12

Laakmann McDowell, Gaye, and Jackie Bavaroman. *Cracking the PM Interview: How To Land A Product Manager Job In Technology*. Palo Alto: CareerCup, 2013.

ABOUT THE AUTHOR

Shrilata Murthy is a practicing data scientist with experience in statistical modeling, data mining, and data visualization. She holds a Master's degree in Analytics from Northwestern University and a B.E. in Information Technology from Mumbai University. As a programmer-turned-data scientist, Murthy now enjoys consulting to help improve data-driven decision making.

When she's not building machine-learning models, Murthy loves spending time with her daughter and going for outdoor runs. Her debut book *Be the Outlier: How to Ace Data Science Interviews* is a practical guide for any novice or experienced data scientist looking to improve their interview technique.