[29-30]

**Data Scientist Interview Preparation**

COURSES

* 29 - Data Analyst Interview Preparation
  + 30 - [Data Analyst Interview Preparation | Codecademy](https://www.codecademy.com/learn/paths/intprep-data-analyst-interview-prep)
* 30 Data Scientist Interview Preparation
  + <https://www.codecademy.com/enrolled/paths/intprep-ds-ds-interview-prep>

# 29 - Data Scientist Interview Preparation

# **Welcome to the Data Analyst Interview Preparation Skill Path**

**You have the data analysis skills, now it’s time to ace the data analyst interview!**

### **Goals of this Skill Path**

The Data Analyst Interview Preparation Skill Path is designed to give you the opportunity to practice answering the types of questions you will see in a data analyst interview. In this Skill Path, you’ll learn how to decode job postings, tailor your materials to a specific job, and answer almost any type of interview question. By the end of this Path, you will have familiarity with data analyst interview question types and the key elements that make a good response, so that you’ll be confident and prepared for even the toughest interviews.

After this Path, you will be able to:

* Specify what kind of role you want to apply for and tailor your job search materials and interview preparation to that specific job description.
* Anticipate the stages of a Data Analyst Interview process, from building your resume to the on-site interview.
* Respond to general data management questions—from cleaning to exploratory analysis.
* Explain your data visualization and communication process.
* Describe your approach to experimental design and project management.
* Ace the SQL technical challenge.
* Effectively communicate your solution to coding challenges.
* Successfully complete the data analyst interview take-home challenge.
* Present yourself professionally as a Data Analyst.

Along the way, you’ll be able to put your knowledge of data analytics into practice with Portfolio Projects where you can showcase your skills and interests.

### **Navigating this Skill Path**

This Skill Path covers the most common topics across a variety of roles within data analytics. Not everything will apply to every situation. Feel free to pick and choose what content is relevant to you and the specific role you are applying for.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned

# **Welcome to Getting Started with Data Analyst Interview Preparation**

**Get started with an overview of data analyst job and interview types.**

### **Goals of this Unit**

The goal of this unit is to introduce you to the different types of data analyst roles and the corresponding job descriptions.

After this unit, you will be able to:

* Read a job description and better understand what skills the company is looking for.
* Tailor your job search materials to meet the requirements stated in a job description.
* Determine what type of role(s) you would be most interested in.
* Prepare for the stages of a data analyst job interview.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

**Overview of the Data Analyst Interview Process**

## **Introduction**

<1 min

Approaching the data analyst interview process for the first time can be daunting. Phone screenings, take-home assessments, whiteboard interviews – does it ever end?

In this lesson, we will demystify the data analyst interview process so that you know exactly what to expect going in. You will learn about:

* Polishing your resume
* Building a digital presence
* Creating a portfolio
* The phone screening
* The take-home assessment
* Technical and whiteboard interviews
* Basic interview etiquette

After this lesson, the rest of this skill path will be focused mainly on portfolio-building, the technical interviews, and take-home assessments.

Let’s get started!

**Instructions**

In this lesson, you will get an introduction to the Data Analyst Interview Process.  
When you’re ready, move on to the next exercise.

**Overview of the Data Analyst Interview Process**

### **Getting Your Foot in the Door**

1 min

The first step of the data analyst interview process is getting your foot in the door.

To set yourself up for success, you will need to polish your resume and build a digital presence.

##### **Resume**

You might have a deep understanding of data analysis, but that alone will not help you get your foot in the door. To stand out from the crowd and secure an interview, you will need a strong resume that showcases your skills and experience.

**First impressions matter**Make sure your resume is free of any spelling or grammar mistakes. Also, make sure it is consistent in style and content. These simple things are often ignored but can go a long way.

**Tailor to the role**It is important to recognize that there is no one-size-fits-all resume in the field of data analytics or data science at large. Carefully read the description for the role you are interested in and make sure your resume reflects the relevant technical skills required. You should do this each time you apply for a new role.

**Be concise**Keep your resume concise by eliminating any skills or points that aren’t relevant to the role. Every word you choose to put on your resume matters, so choose carefully.

**Tell a story**A good resume doesn’t just list out everything you’ve ever done. It should be a story about growth in a particular direction and have a common theme that is relevant to the role.

Try to read your resume in the voice of the hiring manager. Does it catch your attention? Are you the perfect candidate for the position?

What your resume should look like depends largely on the kind of role you are applying for. For example, [here](https://www.indeed.com/career-advice/resume-samples/information-technology-resumes/business-analyst) is a helpful example and guide on how to craft a strong resume for a business analyst role and [here](https://www.indeed.com/career-advice/resume-samples/information-technology-resumes/data-scientist) is one for more of a data scientist role.

##### **LinkedIn**

In this digital age, recruiters usually check your [LinkedIn](https://www.linkedin.com/) profile before inviting you for an interview.

Make sure to upload a professional-looking photo of yourself and summarize your role, skills, and experience in the “About” section.

When listing your previous roles in the “Experience” section, keep your descriptions short and to the point but give a clear breakdown of your role responsibilities. This will help recruiters better understand your skillset.

The “Skills and Endorsements” section is another great way to showcase your skills. Here are a few data analytics skills you might include:

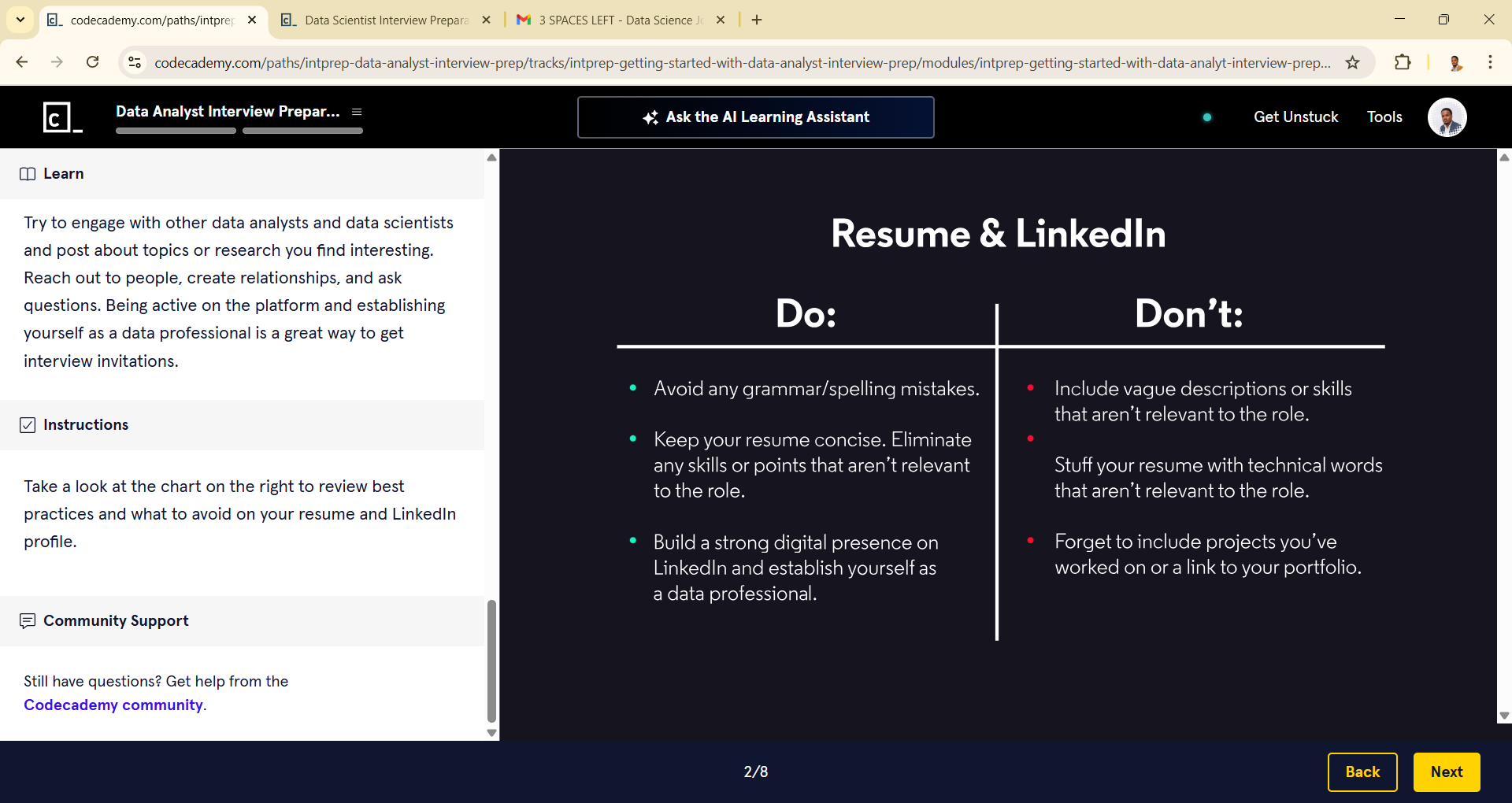
* Python/R
* SQL
* Tableau/Power BI/Looker
* Data visualization

LinkedIn can also be a great way to establish yourself as a leader in your field.

Try to engage with other data analysts and data scientists and post about topics or research you find interesting. Reach out to people, create relationships, and ask questions. Being active on the platform and establishing yourself as a data professional is a great way to get interview invitations.

**Instructions**

Take a look at the chart on the right to review best practices and what to avoid on your resume and LinkedIn profile.



### **Showing Off What You've Built**

11 min

An important part of the data analyst interview process is showcasing the projects that you’ve worked on.

This skill path provides many opportunities to build a strong portfolio that shows off skills companies want to see.

##### **GitHub**

When it comes to showcasing your portfolio,

[GitHub](https://www.codecademy.com/resources/docs/general/github)

Preview: Docs Loading link description

allows you to share your projects publicly, including collaborations with other data analysts and data scientists.

These days, recruiters and hiring managers may turn to GitHub to find talent. GitHub allows them to:

* *See how active you are in the data analytics and data science community at large.*
* *Dig into the source code itself to see your skills and how you approach a problem.*

Make sure that every project you include in your portfolio has a reason for being there and follows best practices:

* *Is the code clean and well documented?*
* *Did you include a README file that explains the basic functionality and how to run the code?*
* *Do your portfolio projects show growth over time?*

GitHub can also be a great way to establish yourself as an expert in a specific domain. For example, if your goal is to work as a data analyst at a finance company, then you should showcase projects using financial data.

You can also find open-source projects from your industry or a specific company and contribute to the project. When companies see you are an active contributor, they will often reach out to you instead of the other way around!

**Instructions**

* Checkpoint 1 Passed
* **1.**
* You’re getting ready to apply to your dream job but you realize that one of your projects doesn’t follow best practices. Take some time to review the code on the right.  
  In this code, we analyze housing data and visualize the relationship between SalePrice and variables such as GarageArea, YearBuilt, and 1stFlrSF.  
  When you’re ready, continue to the next step.
* This code explores housing data found in the housing\_data.csv file.  
  First, we load in the housing data. Then we perform simple analysis such as:
  + Descriptive statistics
  + Histogram
  + Log transformation
  + Visualizations
* Checkpoint 2 Passed
* **2.**
* Currently, you don’t have any comments in your Python file.  
  Write short but helpful comments for each code block that has # add comment here above it.  
  If you need help, we’ve included a housing\_commented.py file that is well-commented.
* Including comments that are relevant and useful can make it easier for others to understand your code and collaborate with you.  
  Check the housing\_commented.py file to see what a well-commented file looks like.
* Checkpoint 3 Passed
* **3.**
* Your project’s README file is blank!  
  Write a helpful README file in README.md that gives a basic overview of the project.
* Note: this is a relatively simple project so the README doesn’t need to be too extensive. Generally, the more complex the project the more extensive your README should be.  
  Here is a great [article](https://towardsdatascience.com/how-to-write-an-awesome-readme-68bf4be91f8b) on what to include in a README file. (We also recommend including some information concerning where your dataset comes from.)
* Checkpoint 4 Passed
* **4.**
* Nice job! Your project is already much more presentable. Try to find any other ways you can polish your project:
  + Do all of the variable names make sense?
  + Are there more efficient ways to write certain parts of the code?

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# load in housing data

housing\_data = pd.read\_csv('housing\_data.csv')

# check the columns

housing\_data.columns

# descriptive statistics summary

housing\_data['SalePrice'].describe()

# histogram

sns.displot(housing\_data['SalePrice'])

plt.show()

plt.clf()

# perform log transformation on sale price data

log\_data = np.log(housing\_data['SalePrice'])

print(log\_data.skew())

sns.displot(log\_data)

plt.show()

plt.clf()

# visualize relationship between 1st floor sq feet and sale price

data = pd.concat([housing\_data['SalePrice'], housing\_data['1stFlrSF']], axis=1)

data.plot.scatter(x='1stFlrSF', y='SalePrice', ylim=(0,800000))

plt.show()

plt.clf()

# visualize relationship between garage area and sale price

data = pd.concat([housing\_data['SalePrice'], housing\_data['GarageArea']], axis=1)

data.plot.scatter(x='GarageArea', y='SalePrice', ylim=(0,800000))

plt.show()

plt.clf()

# visualize relationship between year built and sale price

data = pd.concat([housing\_data['SalePrice'], housing\_data['YearBuilt']], axis=1)

f, ax = plt.subplots(figsize=(16, 8))

fig = sns.boxplot(x='YearBuilt', y="SalePrice", data=data)

fig.axis(ymin=0, ymax=800000)

plt.xticks(rotation=90)

plt.show()

plt.clf()

### **Passing the Screening**

3 min

You perfected your resume and sent in your application. The company sees a potential fit and decides to invite you to a phone screening. Nice job!

##### **What to expect**

The phone screening is an opportunity for the company to get to know you better and also for you to learn more about the company. The interviewer is trying to get a better understanding of:

* Your professional and technical background
* Projects you’ve worked on
* Why you’re applying to the company

The call usually lasts between 15 minutes and 1 hour. Depending on the company, the screening may be with a recruiter, hiring manager, or both. If it is with a recruiter, you can expect behavioral questions such as:

* “Tell me about yourself”
* “What is a challenge you recently faced on a project?”
* “Why do you want to work for us?”

If the call is with a hiring manager, you can expect to get a mix of behavioral and light technical questions. Always make sure to look up the interviewer on LinkedIn beforehand to get more context about who you are speaking with.

In addition, checking the company’s common interview questions on Glassdoor is a great way to get a sense of the types of questions you might expect at this early stage in the process.

##### **Nailing the screening**

The phone screening is meant to gauge your depth of experience, ability to communicate, and interest in the role.

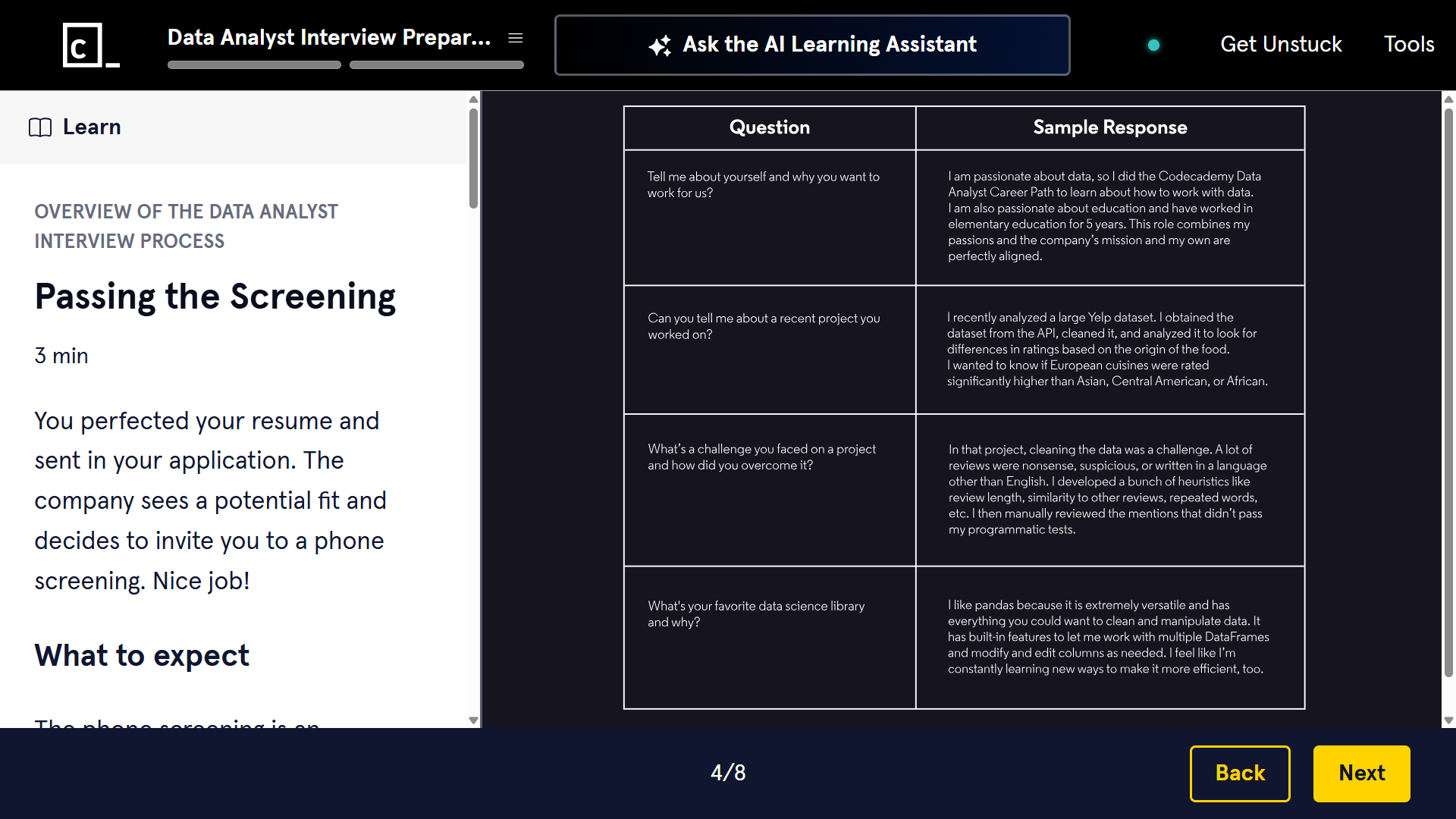
Be professional and personable. When going through your experience, don’t just read from your resume – tell a story and connect it back to why you’re applying to this new role.

In addition, *show* the interviewer that you are genuinely interested by asking thoughtful questions and being an attentive listener. Do research prior to the interview and prepare original questions related to the role or company.

If you are applying to a mission-driven organization, be sure to speak to the mission at this stage.

**Instructions**

Take a look at the chart to review some common questions you can expect during the phone screening.  
If you’d like to view the image in a separate tab, you can view it [here](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/intro/DataAnalyticsInterview_Excercise4.svg).



### **The Take-Home Assessment**

2 min

Your phone screening went well and the recruiter just sent you a take-home assessment. Take a moment to breathe. You’re well on your way to getting your dream data analytics job.

##### **What is the Take-Home?**

The take-home is a technical assessment that helps interviewers get a sense of your skills before moving you to an onsite interview.

Take-homes can vary widely. At a minimum, you can expect:

* a dataset
* instructions for some task(s) or outcome(s)

In this path, you’ll have opportunities to practice take-home assessments.

A benefit of the take-home is that you can show off your technical skills without the pressure of someone watching over your shoulder. The company will usually indicate how much time to spend on the assessment. Though the timeframe should give a sense of what they are expecting you to produce, it is common to spend more time than quoted.

##### **Best Practices**

The take-home allows companies to see how you approach a problem and also whether you follow good coding practices. Here are some best practices to follow:

* Make sure your solution is readable, technically accurate, and well-documented
* Ask yourself whether there is a more efficient or effective solution
* Include interpretable variable names and comments when appropriate

Before you jump into a solution or coding, make sure you have a clear understanding of the data you’re working with and the specifications of the project. If you have questions about the prompt, be sure to reach out to your recruiter. It is better to ask than to answer the wrong question.

The best way to improve at take-home assessments is to practice doing them. Throughout this Skill Path, you will get plenty of practice. You can expect to code in Python (or R) and SQL.

In addition, a great way to prepare is to participate in data analytics and visualization competitions like [Tableau’s Iron Viz](https://www.tableau.com/community/iron-viz), which offer take-home style challenges. There you can also see how different data analysts approach a problem and the practices they follow.

**Instructions**

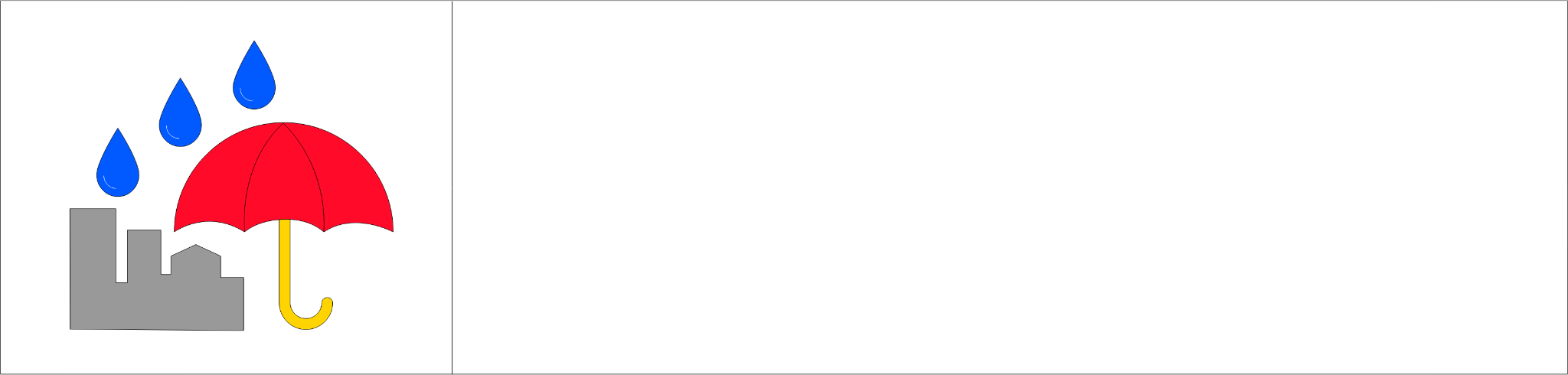
Your take-home assessment might involve interacting with data that the company uses.  
For example, Airbnb might ask you to solve a business problem for them using public housing [data](http://insideairbnb.com/get-the-data.html) in the Seattle area.  
Try to think about the company you’re applying to and how they might interact with data. Look at the chart to see some examples of how companies use data.



### **Technical and Whiteboard Interviews**

1. 4 min
2. You’ve made it through the screening and passed the take-home assessment. At long last, the company invites you to the final part of the interview process: the onsite interview.
3. The onsite consists of a mix of behavioral and technical questions, and can include whiteboarding or other technical interviewing techniques. These may include:
4. SQL questions or Python/R coding challenges
5. working with a dataset live on a computer (e.g., Here’s a dataset -> Conduct some EDA -> What insights are you able to find?)
6. walking through a project you built and explaining the technical implementation
7. general   
   [statistics](https://www.codecademy.com/resources/docs/general/statistics)
8. Preview: Docs Loading link description
9. questions
10. business problem-solving questions
11. behavioral questions
12. This Skill Path will help you prepare for the types of technical questions you might encounter.
13. You can also check Glassdoor to see common behavioral and technical questions for the company you’re applying to and the format of their interview process.
14. **Instructions**
15. On the right, you’ll see a popular data science interview question asked by tech companies. Take a moment to read and think about the question.  
    Try to come up with a solution on your own. Then, take a look below for the complete solution.  
    **Click to see the solution**First, it’s important to understand what the question is asking you to solve.  
    You need to figure out the probability that it’s raining in Seattle, given that all three friends tell you that it is raining.  
    Abstractly, we can write this as:
16. P(Raining∣Yes,Yes,Yes)
17. *P*(*Raining*∣*Yes*,*Yes*,*Yes*)
18. Next, let’s use Bayes’ Theorem to rewrite the above. Here is a refresher on Bayes’ Theorem:
19. P(A∣B)=P(B∣A)⋅P(A)P(B)
20. *P*(*A*∣*B*)=
21. *P*(*B*)
22. *P*(*B*∣*A*)⋅*P*(*A*)
23. ​
24. Using this formula, we get:
25. P(Raining∣Yes,Yes,Yes)=P(Yes,Yes,Yes∣Raining)⋅P(Raining)P(Yes,Yes,Yes)
26. *P*(*Raining*∣*Yes*,*Yes*,*Yes*)=
27. *P*(*Yes*,*Yes*,*Yes*)
28. *P*(*Yes*,*Yes*,*Yes*∣*Raining*)⋅*P*(*Raining*)
29. ​
30. Let’s break this down even further.  
    We can easily solve for P(Yes, Yes, Yes|Raining) since the problem states that each friend has a 2/3 chance of telling the truth and all three said it was actually raining. So if all three friends tell you it’s raining and all three friends are telling the truth:
31. P(Yes,Yes,Yes∣Raining)=2∗2∗23∗3∗3=827
32. *P*(*Yes*,*Yes*,*Yes*∣*Raining*)=
33. 3∗3∗3
34. 2∗2∗2
35. ​
36. =
37. 27
38. 8
39. ​
40. What about P(Raining)? This is the probability that it rains on any given day (which has nothing to do with what your friends say). We’re missing this crucial piece of information!  
    This is a good time to ask your interviewer a clarifying question. Let’s say that your interviewer tells you to assume that P(raining) is 25%.  
    Great! You now have the knowledge you need to solve the problem.  
    Let’s solve for P(Yes, Yes, Yes), which can be rewritten as:
41. P(Yes,Yes,Yes)=P(Raining)∗P(Yes,Yes,Yes∣Raining)+P(Not−raining)∗P(Yes,Yes,Yes∣Not−raining)
42. *P*(*Yes*,*Yes*,*Yes*)=*P*(*Raining*)∗*P*(*Yes*,*Yes*,*Yes*∣*Raining*)+*P*(*Not*−*raining*)∗*P*(*Yes*,*Yes*,*Yes*∣*Not*−*raining*)
43. Now we have all of the necessary information to solve for P(Yes, Yes, Yes):
44. P(Raining)=0.25
45. *P*(*Raining*)=0.25
46. P(Yes,Yes,Yes∣Raining)=827
47. *P*(*Yes*,*Yes*,*Yes*∣*Raining*)=
48. 27
49. 8
50. ​
51. P(NotRaining)=0.75
52. *P*(*NotRaining*)=0.75
53. P(Yes,Yes,Yes∣NotRaining)=1∗1∗13∗3∗3=127
54. *P*(*Yes*,*Yes*,*Yes*∣*NotRaining*)=
55. 3∗3∗3
56. 1∗1∗1
57. ​
58. =
59. 27
60. 1
61. ​
62. Therefore, P(Yes, Yes, Yes) can be simplified to:
63. P(Yes,Yes,Yes)=0.25∗827+0.75∗127
64. *P*(*Yes*,*Yes*,*Yes*)=0.25∗
65. 27
66. 8
67. ​
68. +0.75∗
69. 27
70. 1
71. ​
72. Let’s go back to what we are trying to solve (the probability that it is raining given that all three friends said it was):
73. P(Raining∣Yes,Yes,Yes)=P(Yes,Yes,Yes∣Raining)⋅P(Raining)P(Yes,Yes,Yes)
74. *P*(*Raining*∣*Yes*,*Yes*,*Yes*)=
75. *P*(*Yes*,*Yes*,*Yes*)
76. *P*(*Yes*,*Yes*,*Yes*∣*Raining*)⋅*P*(*Raining*)
77. ​
78. Putting everything together, we get:
79. P(Raining∣Yes,Yes,Yes)=(827⋅0.25)0.25∗827+0.75∗127
80. *P*(*Raining*∣*Yes*,*Yes*,*Yes*)=
81. 0.25∗
82. 27
83. 8
84. ​
85. +0.75∗
86. 27
87. 1
88. ​
89. (
90. 27
91. 8
92. ​
93. ⋅0.25)
94. ​
95. We can simplify further by canceling out the 27’s and multiplying the numerator and denominator by 4.
96. P(Raining∣Yes,Yes,Yes)=811
97. *P*(*Raining*∣*Yes*,*Yes*,*Yes*)=
98. 11
99. 8
100. ​

Therefore, the probability that it is raining in Seattle, given that all three of your friends said it is raining, is 8/11.

****

### **Interview Etiquette**

1 min

During the interview process, it’s important to follow some basic etiquette rules.

If you are invited to interview, make sure to do thorough research about the company and role beforehand.

* What are the history, mission, and values of the company?
* What is the company’s technology stack?
* Why is this role important to the company mission?

##### **Look into the company and role**

Doing your homework will allow you to have more engaging conversations with the interviewer. Show them that you care about the company mission and express your curiosity by asking specific questions.

##### **Learn about the interviewers**

Look up your interviewer(s) on LinkedIn beforehand so that you can better understand who you’re speaking with. The more context you have, the easier it will be to develop rapport and have a smooth conversation. It’s also helpful to know how technical your interviewer is.

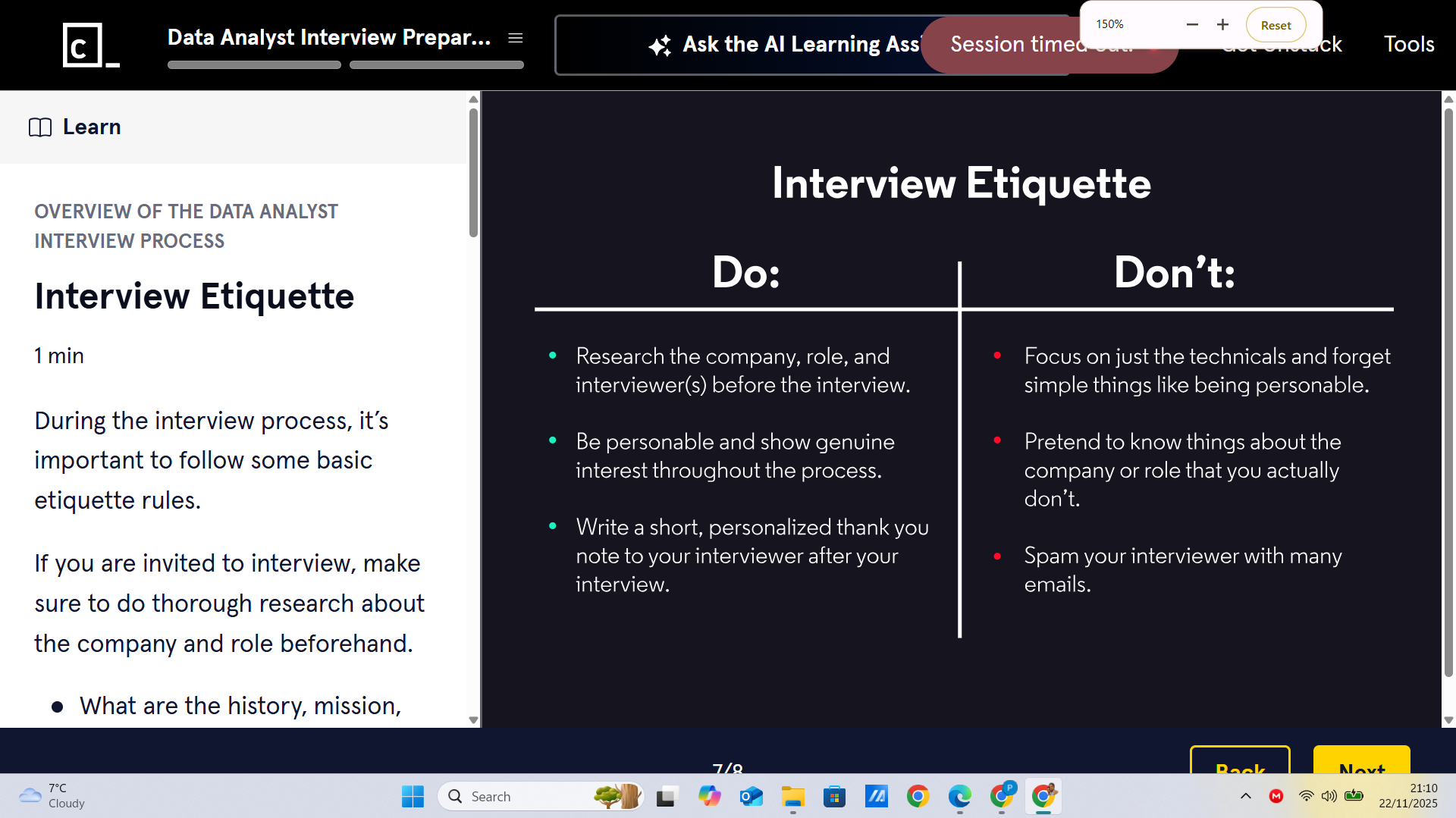
##### **Follow up**

After each interview, send a thank you note to your interviewer (preferably within 24 hours after the interview). Keep it short and simple, but also make it unique. Try to mention something that you discussed during the interview if possible.

While it is true that companies are trying to hire the most talented data analysts, they are also trying to gauge whether you are passionate about the company mission and a good fit for the company culture.

**Instructions**

Take a look at the chart to review basic interview etiquette.



### **Review**

<1 min

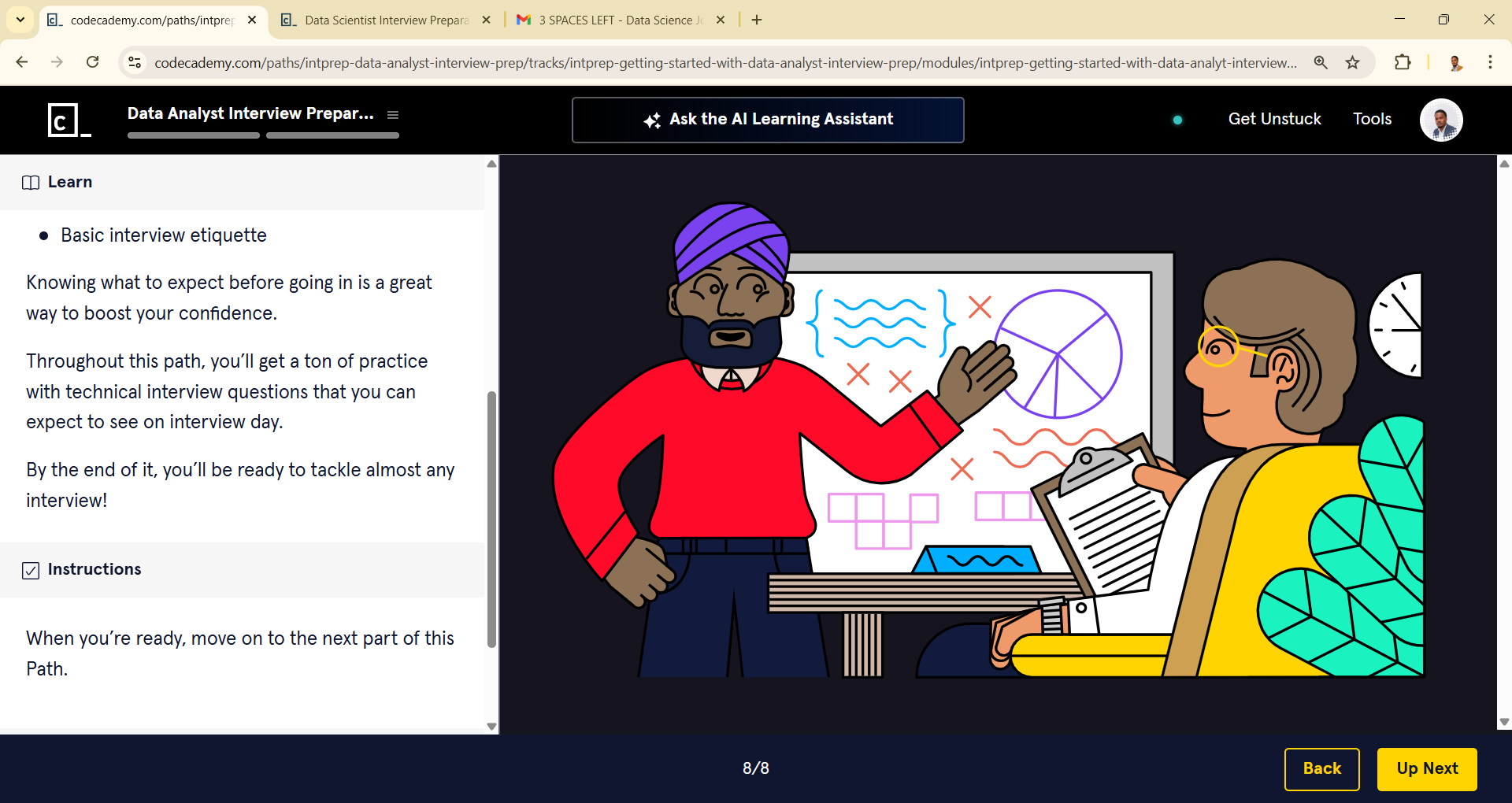
Congratulations! You now know about the basic components that make up the Data Analyst Interview Process:

* Polishing your resume
* Building a digital presence
* Creating a portfolio
* The phone screening
* The take-home assessment
* Technical and whiteboard interviews
* Basic interview etiquette

Knowing what to expect before going in is a great way to boost your confidence.

Throughout this path, you’ll get a ton of practice with technical interview questions that you can expect to see on interview day.

By the end of it, you’ll be ready to tackle almost any interview!



# **Common Data Analytics Roles**

**An introduction to navigating data analyst and data scientist roles across industries**

### **Introduction**

Searching for jobs in data analytics or data science can be a tricky process to navigate. For example, you might find different job titles such as *Data Analyst*, *Decision Scientist*, *Quantitative Analyst*, or *Business Analyst*.

Which one do you choose? How do you prepare for each?

Learning what to expect in your job search will help you approach the process with confidence. In this article, you will learn:

* Common data analyst and data scientist roles
* How to recognize roles in job postings
* How to prepare for each type of role

Let’s get started!

Free response

Describe your ideal role. What do you want to spend the majority of your day working on?

**Submit Response**

### **Common Data Analyst Roles**

Data analytics is a broad, interdisciplinary field that combines other disciplines such as probability, statistics, visualization, and databases. The responsibilities could range from writing reports to building linear regression models. This means you are likely to find a wide variety of role titles when searching for jobs in the field of data analytics.



When searching for jobs in data analytics, common titles you might see include:

* Data Analyst
* Quantitative Analyst
* Business Intelligence Engineer/Analyst/Developer
* Business Analyst
* Operations Analyst
* Data Scientist - Analytics
* Data Scientist - Inference

In many cases, the job title might simply be “Data Analyst” but other times it will be something more specific. This has a lot to do with the company and industry you are looking at. For example, if you apply for a data analyst position at an investment firm, you are likely to see something like “Quantitative Analyst” or “Financial Analyst.”

It is also important to remind yourself that data analyst job titles don’t always reflect the actual responsibilities. Make sure to dig in the job description and responsibilities to make sure the role is what you expect.

### **Common Data Scientist Roles**

Data science is a broader field that encompasses data analysis, but typically includes more of an emphasis on programming, as well as predictive modeling and machine learning. Common data scientist titles you might see include:

* Data Scientist
* Data Scientist - Modeling
* Decision Scientist
* Applied Research Scientist

Note that the job title you see does not always give a complete picture of what is expected on the job. How one company defines the role of a data analyst or scientist can be very different from how another company chooses to do so. As we’ll discuss shortly, it is much more helpful to look at the job description and requirements.

### **Recognizing Roles in Job Postings**

Job titles can be deceiving – they don’t always reflect the actual responsibilities of the job. This is especially true for data science and data analytics, which are relatively new fields that aren’t fully defined yet. Many data science jobs are actually data analyst jobs in disguise and vice versa.

That’s why it is much more helpful to look at the job description and responsibilities. Sometimes there will be a “Skills/Tools” section that lists out all of the tools, programming languages, and libraries that are required for the job. This can be a helpful way to see if the job matches your expectations.

From our perspective, these are the typical responsibilities of each role (acknowledging that this can vary widely):

**Data Analyst:**

* Access and query different data sources (e.g., SQL)
* Process, manipulate, and clean data
* Summarize data
* Employ statistics and mathematical techniques
* Data visualizations and reports

**Data Scientist:** Requires knowledge of all of the above, plus:

* Much more emphasis on programming (Python, R)
* Statistical modeling
* Machine learning knowledge

In general, data analysts look at the past to find trends, summarize data, and communicate findings, while data scientists use the past to make a prediction about the future. If you see phrases like “Predictive analysis,” “Statistical modeling,” “Machine learning,” or “Extensive Python/R experience,” that means you are probably looking at a data science role. If the role lists skills such as Excel, SQL, Tableau, and minor programming experience, then you are most likely looking at a data analyst role.

Context is also really important when going through job postings. Make sure to pay close attention to the company and industry that you are applying to. How a marketing company chooses to call its data analysts (“Marketing analyst”) might be very different from how a financial data company chooses to do so (“Quantitative analyst”).

Also, each company has a unique workflow and slightly different expectations for its data scientists and analysts. Some will even require that you have industry-specific knowledge or experience working with a particular type of data.

### **Preparation**

Your interview preparation will vary based on the role that you decide on and the company and industry you are interested in. Generally speaking, if you apply to a data science role at a large company, you should make sure to have a strong understanding of general data questions, Python, machine learning, with less focus on SQL and data visualization/communication.

For a data analyst or data scientist role at a smaller company, you are less likely to see machine learning or Python questions. You should focus your attention on general data questions, SQL, and a heavy focus on data visualization and communication.

For example, if you are applying to the Alexa team at Amazon, in addition to machine learning knowledge, you will need to brush up on your natural language processing knowledge. If you are set on working as a quantitative scientist at a hedge fund, you will need to practice analyzing financial data with Python.

The reality is that there is no definitive set of questions to study for a data science or analytics interview. But! That is actually a strength of the field because the roles are so diverse that there is something to fit every kind of data professional. So, be sure to tailor your job search based on your interests and the company you are applying to.

# **Review of Getting Started with Data Analyst Interview Preparation**

**See what you learned about different data analyst roles and job interviews.**

### **Review**

Congratulations! The goal of this unit was to introduce you to the different types of data analyst roles and the corresponding job descriptions.

Having completed this unit, you are now able to:

* Read a job description and better understand what skills the company is looking for.
* Tailor your job search materials to meet the requirements stated in a job description.
* Determine what type of role(s) you would be most interested in.
* Prepare for the stages of a data analyst job interview.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# **Welcome to Fundamental Data Interview Questions**

**Learn how to communicate about general data principles and practices.**

### **Goals of this Unit**

The goal of this unit is to give you practice answering general data questions. These questions may or may not appear as stand-alone questions during your interview, but may be embedded in others. Data analysts are expected to be fluent with all of these topics.

After this unit, you will be able to:

* Talk through how you work with data and manage different data types.
* Define your data cleaning process.
* Implement basic summary statistics and explain the significance of the results.
* Speak to when you use different sampling methods.
* Walk through solutions to probability interview questions.
* Approach estimations problems logically and systematically.
* Apply hypothesis testing to practical business problems.
* Effectively communicate your thought process and problem-solving skills.
* Clean and summarize a real dataset.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# **Data Types Interview Questions**

**Practice answering common interview questions about general data concepts and data types.**

### **Introduction**

Hi! My name is Codey Cademy. I’m so excited because today I’m interviewing candidates for the data analyst role on my team, which means that I get to hire someone to help me analyze all this data! I got my budget approved, went through all of the applications, did my phone screens, and now I’m down to just a few candidates.

Whoever I hire needs to be data-fluent. I’ve heard stories from my friends about hiring people who know all about the latest algorithms but just didn’t know the basics. That can really cost a lot of time and money, and I don’t want that to happen to me. So, I’m going to make sure my next hire is skilled in all stages of data analysis.

There are three types of tasks I want to be sure they can do:

* Exploratory analysis
* Diagnostic analyses
* Data organization

I’m going to ask them a mix of conceptual and applied questions to be sure they know the concepts and that they’d be able to apply them to the real world.

I might integrate some of these basic concepts into the other questions I ask them, too, but the basics are so important for this role that I am going to ask directly.

I really hope they talk through their answers. It’s so much easier to understand how someone thinks about data when they vocalize their thought process and assumptions. If I could give them one piece of advice for our day together, it would be:

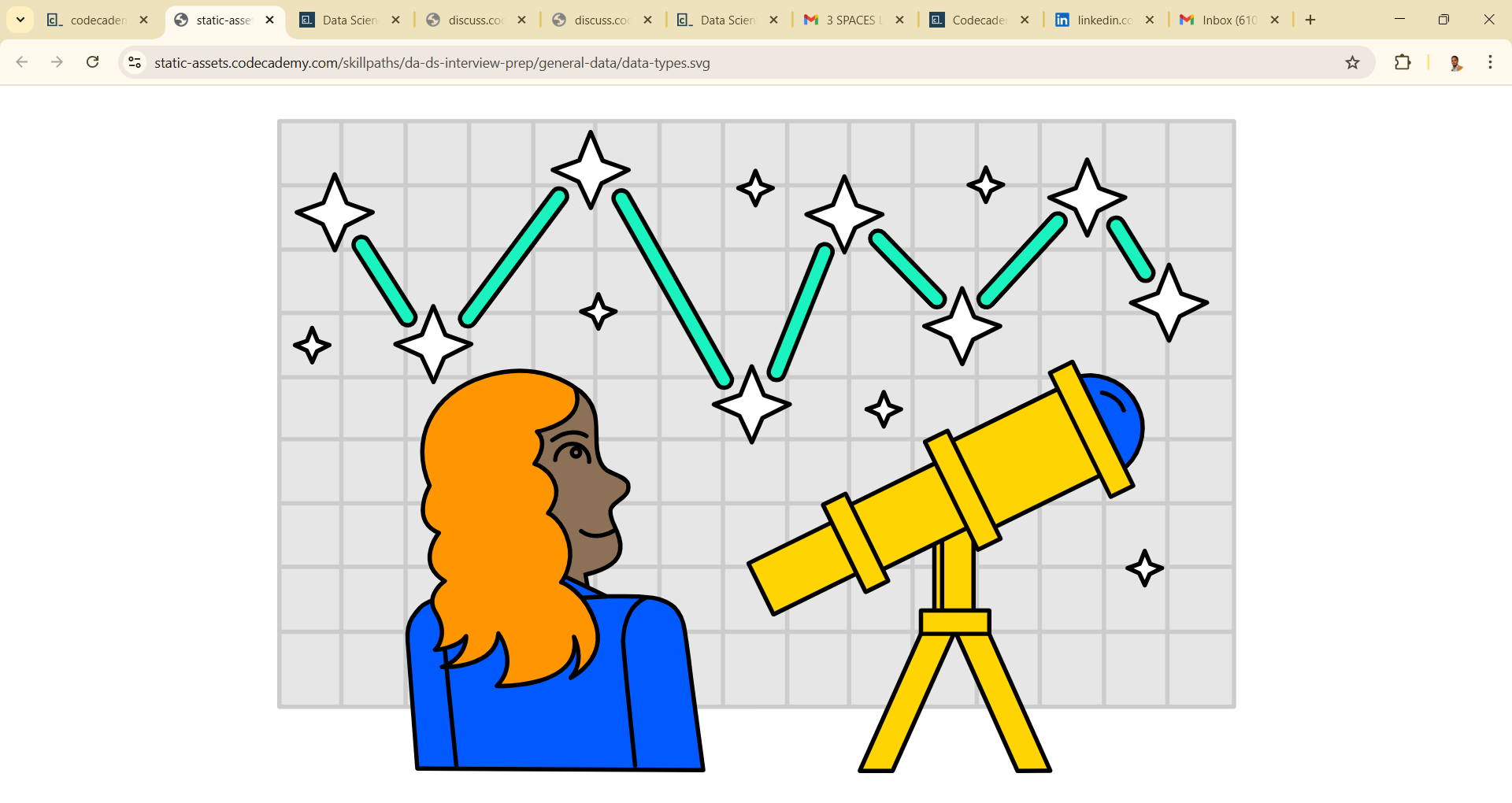
When in doubt, talk it out!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Theoretical Questions**

Hi! Welcome to your basic data types interview! We’re going to start off with a few conceptual questions about working with data.



Free response

How do you typically start an exploratory data analysis?

I start by get sense of the data through their types (numericals (continuous and discrete), categorical (nominal and ordinal)), missing values, or incorrect entries, format and size. From this undesratnd I start to apply most appropriate tecnique. Also, depending on the goal of my task I can start technique to organize data into insight for report, otherwise after clenaing Ipass on to featureing new data

**Feedback**

Partially correct

**What you did well**

* You’ve correctly identified the importance of understanding the types of data (numerical and categorical) you’re working with.
* You’ve also recognized the need to handle missing or incorrect entries, which is a crucial part of data cleaning.
* Your mention of tailoring your approach based on the goal of your task shows a good understanding of the context-dependent nature of data analysis.

**What could be improved**

* Try to elaborate more on the specific techniques you would use for data cleaning and exploration. For example, how would you handle outliers or skewed data?
* It would be beneficial to mention the use of summary statistics and visualizations to understand the data better.
* Your response could be improved by discussing how you would approach categorical variables differently from numerical ones.
* It would be helpful to mention the iterative nature of exploratory data analysis, where you might need to cycle through different steps to gain a deeper understanding of the data.

You’re on the right track! Keep refining your process and don’t forget to delve deeper into each step. Remember, data analysis is as much an art as it is a science. Keep practicing and you’ll continue to improve!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

The first thing I always do is clean my data and by cleaning it, I also get a sense of what I’m working with. In the cleaning process, I usually end up touching all of the variables.

After cleaning, I first run summary statistics on all of the variables and look at the mean, median, interquartile range, and check for outliers. If anything stands out, I dig deeper to see if there is a story I can start to find. I make sure to note the range of values in this dataset.

I then look at the distribution of each variable. I like to make a pair plot to visualize the distribution and the correlations between variables. If anything stands out, I dig deeper with more visualizations or by looking at the value of the correlations.

If there are outliers, I inspect the observations or rows with the outliers. Sometimes the outliers give a good entry point to constructing a data story.

For categorical variables, I like to see what the relative distribution is, so I usually make a bar chart for each variable with categorical values.

From there I start building a data story. I often cycle through these steps to figure out what to spend more time with to better understand the data.

WHY THIS RESPONSE?

The key elements of this answer are:

* looking for trends and relationships
* iterating through this process and
* exploring different variable types.

You might have a slightly different process to exploring your data, but being able to articulate that process demonstrates that you know how to get acquainted with data and how to get started with a dataset.

Free response

Tell me about a data analytics project\* you worked on and the major steps you took to complete it.

(\*Note that this could be a project you complete as part of this Skill Path.)

Your response

I have worked on the customer reviews where it was composed by numeric, categorical and text data. The goal was to build Machine learning model which was implemented using neural netowrk ML. First step that I’ve taken was to separate data from text and other then combining laker on. Cleaning data was the first step afterward which was conducted summary statistics look into range, mean, middiam, quartile and outliers. Visualization was performance to better understand the data distribution using pairplot to its power in showing both distribution and correction. With categorical dat barplot were used to better understand their distrbuition. Surely, it was an iterative process including similar process done with text data - bag-of-word, and word embbeding through CountVectorizer. Naturaly, I ve creature new feature from relation undesratnding of features in data set , which ended spliting data, creating model, testing, tuning applying gridsearch dozens parameters and few algorithms

Our answer

I recently completed a project working with YouTube comments. The goal was to better understand what makes something go viral. My hypothesis was that viral videos have a lower like-to-dislike ratio than non-viral videos. The idea is that something goes viral because it is highly polarizing rather than because it is well-liked.

I decided that viral videos were anything that was an outlier in terms of shares since that’s how we think about things that are viral–they are very different from the norm.

First, I downloaded a collection of about 500,000 videos and their metadata from Kaggle. This dataset was divided by country, so I combined and then cleaned the dataset, checking for missing data, outliers, etc. Then I created a new column for the like-to-dislike ratio. I then grouped by the video id and made a new dataset that only included the first day each video was live. I then separated the “viral” videos from the rest.

To find the outliers, I visualized the distribution with a histogram, then calculated the interquartile range, and used both to determine a cutoff point in the data.

To test my hypothesis, I used a t-test because, within each sample, the data was mostly normally distributed. I found that the difference in the mean like-to-dislike ratio of viral videos versus non-viral videos was statistically significant at 95% confidence level. I interpreted that to mean that there is something special about the like-to-dislike ratio of viral videos. I then isolated the videos with the like-to-dislike ratio nearest to one (meaning that there were almost as many likes as dislikes). I found that this also represented the videos with the lowest like-to-dislike ratio as there were very few videos with more dislikes than likes.

The next step would be to watch them and do a qualitative analysis to determine if they are actually more polarizing.

WHY THIS RESPONSE?

This answer addresses the steps of working with a real dataset. The purpose of this kind of question is to be sure that you have worked on a project and give you the opportunity to talk about something that you should be very familiar with. This answer works because it contains all the following elements:

* A short summary of what the project was about, include the data source and the question.
* A description of any assumptions that you made and why.
* A brief summary of how you cleaned the data, explored it, and reshaped it to meet your needs.
* A description of how you analyzed it. Did you do hypothesis testing? A/B Testing?
* A brief summary of the results and how you interpreted them, including future questions or limitations.

Great! You sound like you know your way around a dataset! Let’s move on to applying these concepts to real-life problems.

### **Situational Questions**

There are a lot of companies and organizations out there and we all face different challenges. The best interviewees I’ve met with really knew my company. They did their research and gave some thought to the types of problems we face and the data we work with. Let’s go ahead and apply the general concepts to some business problems.

Free response

An e-commerce site: We have been collecting data on user behavior including clicks, time to checkout, unique visits, etc. Yesterday we saw a 50% drop in checkouts. We didn’t receive any customer complaints though. Using just the site data, how would you start to diagnose the problem?

…

**Feedback**

Not quite

**What you did well**

Unfortunately, your response was left blank, so I couldn’t identify any strengths in your answer.

**What could be improved**

It seems like you didn’t provide an answer. Next time, try to share your thoughts on the question. Even if you’re unsure, it’s better to attempt an answer as it can demonstrate your current understanding and allow for more specific feedback.

Don’t worry, everyone starts somewhere. Keep going, and don’t hesitate to express your thoughts. You’re here to learn, and every attempt is a step forward!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

I would start by doing some Exploratory Data Analysis, looking at the distribution of each variable, their relationship to each other, and if any are directly correlated. I would investigate the outliers and look into any strange behavior by going deep into the rest of the record for each outlier instance.

I would analyze this data with respect to the question by finding or creating a 30 day benchmark with a 68% confidence interval for clicks on each page, unique visits, and time on page for each page. I would systematically compare each of those benchmarks with yesterday’s metrics and flag any metrics outside of that 68% confidence interval. I would then look into any pages that were flagged in that process to see if there were any changes to those pages.

WHY THIS RESPONSE?

For Data Analysis roles, questions about diagnostics can be approached with benchmark analysis. The key is to specify how you will create your benchmark and if you will use a point estimate or a confidence interval. We kept the confidence interval narrow in order to identify anything even slightly out of the ordinary. This will make it look like more things are out of the “norm” than actually are, but that is preferred when diagnosing a problem.

Free response

A rideshare startup: We have demographic data, car ownership, and insurance premium data from 2019-2021 for residents in the Detroit area. We want to expand our services there in 2022. With this in mind, what questions would you develop for this dataset and how would you begin to explore this data?

..

**Feedback**

Not quite

**What you did well**

It seems like you didn’t provide an answer to the question. It’s okay, everyone has moments where they’re unsure of how to proceed.

**What could be improved**

Try to read the question carefully and understand what it’s asking. Then, think about the information you have and how you can use it to answer the question. Don’t hesitate to take your time and think it through.

Don’t worry, you’re still learning and it’s okay to make mistakes. Keep trying and you’ll get there!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

I would start by clearly and concretely defining the goal. So can I assume that our goal is to identify which neighborhoods we should expand our services into first? Can I also assume that we are going to make this decision by leveraging all of the geographically situated demographic and economic data we have?

[Codey] That sounds like a good summary.

Ok, then once everyone has agreed on the goal, I would start by exploring all of the variables available to determine exactly what demographic information is available. I assume it would be income and profession. I will also assume that it includes some location information. I would then segment the population based on income to think about each of these as different potential markets.

I would first look at people who are cost-burdened by owning a car. These would be people in the lowest income quartile who also own a car. What percentage of their income goes towards insurance? What neighborhoods do they live in? What jobs do they have? Are those jobs typically 9-5 jobs? What are their age ranges? What is their driving history overall? This would help develop a profile for this group that could be a starting point for understanding their unmet needs and how a new rideshare might meet those needs.

I would then look at the same variables for middle-income and upper-income car owners.

I would also look at non-car owners divided by the same income breakdowns. What trips do they take? How often do they take them? What alternatives do they have? Did they typically own cars at some other point in time? If so, what is the average length of time that they have been without a car and what are they doing now for transportation?

I would use the results of this analysis to identify potential markets, and consider how those markets might change throughout a day, week, or year.

WHY THIS RESPONSE? This response addressed the specific problem that was presented by segmenting users into different potential markets and exploring as many variables as possible to understand those markets. This answer works because it focuses both on addressing the need and articulates a clear method for identifying and segmenting the data to make sense of it.

Thanks! It sounds like you’ve really done your homework on how we think about data! Let’s see how you do with some of the specialist data types we work with.

### **Temporal Data**

We just got access to new data sources, and one of them is time-bound! It looks to me like there’s some seasonality to it, but I’m not sure. I know a little bit about working with time, but not a lot. Looks like you also list familiarity with time series on your resume. Wonderful! Let’s talk about some simple time series analysis.

Free response

What is time series analysis, and what aspects would you keep in mind when working with time series?

Give the best answer

**Feedback**

Not quite

**What you did well**

* You’ve shown a willingness to participate and engage with the question.

**What could be improved**

* Try to provide a detailed explanation of what time series analysis is and the aspects to consider when working with time series data.
* Remember to include key concepts such as autocorrelation and seasonality, and how to handle them in your analysis.

Don’t worry, everyone starts somewhere. Keep learning and practicing, and you’ll get the hang of it in no time!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

Time series is any data that has a time element to it. Typically it’s some variable or variables measured at evenly spaced times. The hard thing about time series data is that we expect variables that are independent to be independent, but if they are correlated over time, they aren’t really independent.

The biggest thing I take into consideration is the potential for autocorrelation among the variables. In addition to plotting the variables, when running correlations between them, I plot the residuals. If there appears to be a pattern in the residuals, I use an autocorrelation function to account for it.

In the case of seasonality, I would take data from previous seasons to create a seasonal index and then scale the seasonal effects to a normalized standard before making any generalizations or analyzing overall trends.

Finally, because this is an additional process that is often invisible to stakeholders, I make explicit mention of it in my analysis and interpretation.

WHY THIS RESPONSE?

The important part of this question is explicitly mentioning that autocorrelation and/or seasonality are often present in time series data. Time series are unique because they often have an internal structure. An awareness of this fact and some strategies to deal with it will be essential if you are working with time series data.

Coding question

## **Questions**

Change a string into a timestamp in Python.

### **Code**

1

2

3

4

5

6

7

Output:

**Run**

Run your code to check your answer

Free response

Toy company: We launched a new toy campaign this year and want to know if our marketing was successful. How would you account for seasonal variation to determine if our advertisements affected sales?

..

**Feedback**

Not quite

**What you did well**

Unfortunately, your response was not provided, so I can’t identify any strengths in your answer.

**What could be improved**

It seems like you didn’t provide an answer. Next time, try to engage with the question and provide your thoughts. Remember, there’s no wrong answer in learning. It’s all about trying and improving.

Don’t be afraid to make mistakes. That’s how we learn. Keep going, you’re doing great!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

First I would look at the previous 3-5 years’ toy sales over the entire year for toys in the same category to calculate a seasonal adjustment factor. I would then look at the overall trend of sales for that category before the new campaign launched and the relative performance of other toys after the campaign. I would focus on the time period immediately preceding the launch as well as the relative performance of other products at the same time. Using these three pieces of information, I would adjust the data to try to isolate the effect of the campaign.

WHY THIS RESPONSE?

This response addresses the fact that there is some seasonality to non-perishable product sales and taking that into account is essential for measuring impact. Secondly, it acknowledges that there are overall trends in consumption in a market, so normalizing by both dimensions is essential.

Wow! You really know your stuff. Turns out that we also have some unstructured text. Looks to me like Tweets and reviews. I know it’s asking a lot from one person to be able to work with all of these different data types, but let’s go through some textual data questions.

### **Text Data**

I only took a brief look at the new data we are pulling into our data lake, but I’m hoping you can help make some sense of it. I’m going to ask a few questions about how you’ve worked with text.

Free response

What steps would you take to normalize text?

..

**Feedback**

Generating...

Feedback can take up to 5-10 seconds to generate.

Something went wrong and we could not generate any feedback.

**Try again**

### **Our Answer**

Normalization is always context-dependent, so I would start by inspecting the corpus to understand what type of data we are looking at and what the goals are. I will assume that we are working with relatively long-form data and we are trying to cluster our documents together.

I would start with making everything lowercase and removing punctuation. Then I would tokenize the words. Assuming it is a relatively small corpus, I would lemmatize the words. If it is large, I would stem them. This would help reconcile the different word forms into their basic form.

I might remove stop words depending on what we are doing. If we are working with short-form text, I might not do this, but again it’s always situation-dependent. Finally, I would deduplicate if relevant. Again, if we are working with Tweets or something like that, deduplicating will be essential, if we’re working with something more formal, it might not be necessary.

There are a few ways of deduplicating, if some overlap is ok, I might just use a simple match. If I’m working with something like Tweets which can possibly be very similar or include typos, I might also use something like Levenshtein distance.

In the end, I always make sure that my normalization process is appropriate for the task at hand.

WHY THIS RESPONSE?

This answer is good because it emphasizes that the process is context-dependent but also that there are a few key steps that are frequently used. It also addresses some thinking about why each technique is or is not appropriate for the data and goals of the project.

Free response

Large manufacturing company: We have a collection of employee generated incident reports. We would like to use these records to identify trends across departments. What would you do to process this data?

..

**Feedback**

Not quite

**What you did well**

Unfortunately, your response was not provided, so I couldn’t identify any strengths in your answer.

**What could be improved**

It seems like you didn’t provide an answer. Please make sure to provide a response to the question. Remember, it’s okay to make mistakes - that’s how we learn!

Don’t hesitate to try again! You’re here to learn and every attempt brings you one step closer to mastering the subject. Keep going!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

I am going to assume that the reports do not have a rating system or other quantifiable information attached and that each report is just text maybe with a timestamp and some identification of the author.

I would start by collecting all of the reports into one corpus. I would do some light normalization on it. Since it is short-form text, I wouldn’t normalize very heavily. After that, I would conduct topic modeling and probably start with LDA.

If my topic modeling doesn’t turn up any coherent topics, I would first inspect the LDA-vis and look at the reports in the outlier clusters.

I would then re-run the topic model with more or fewer topics. I would then read through a random sample of documents in each cluster to be sure that the topics are coherent, write a summary for each cluster, and collapse the clusters with significant overlap. If we expect to get additional reports, I would save the weights from the topic model to classify new documents into the existing clusters.

WHY THIS RESPONSE?

This response details an interactive and creative solution to working with a novel dataset. The interviewee stated their assumptions and did not try to leverage other data that may or may not be present, but focused on the text data that the interviewer was asking about. The solution focuses on solving the problem at hand, and the steps are all in service to this particular dataset.

### **Review**

Great job! I was really impressed with your answers and think you will be a great fit!

We covered:

* Basic data management,
* applied data management concepts to business problems, and
* touched on working with both time-series and text data.

You showed a thoughtful and organized approach, and I’m starting to get a sense of how you approach problems because you were able to articulate your thought process.

We’d love to move you along to the next stage in the hiring journey. Let’s take a look at some more questions.

# **Data Cleaning Interview Questions**

**Practice data cleaning and data wrangling interview questions.**

### **Introduction**

Welcome back!

In this portion of the interview, are going to talk about your strategies for cleaning data. Data cleaning is a huge part of my job (sometimes it feels like my only job). Some people estimate that up to 80% of data scientists’ analysts’ jobs is data cleaning!

I’m going to ask you two types of questions:

* Conceptual questions asking about techniques
* Situational questions asking how you would approach a dataset

You’ll have a chance to talk through how you think about cleaning data and the steps you take. This is not the last time you will have to clean data, though. It will surely be a part of your take-home challenges. Since cleaning data is so central to our jobs, I need to be sure our next data scientist is up for the challenge!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Conceptual Questions**

Free response

What steps do you take when you first clean a dataset?

…

**Feedback**

Not quite

**What you did well**

It seems like you didn’t provide an answer to the question. It’s okay, everyone has moments of uncertainty.

**What could be improved**

Next time, try to provide an answer that reflects your understanding of the topic. In this case, you could discuss the steps involved in cleaning a dataset, such as inspecting, repairing, validating, and documenting the data.

Don’t worry, learning is a journey and it’s okay to have bumps along the way. Keep going, you’re doing great!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

I start by giving myself a checklist:

* Inspect
* Repair
* Validate
* Document

For each new dataset, I follow this sequence for each aspect of the data:

I start with column names and metadata.

* I **inspect** to make sure there are no stray characters, all the codes have been made human-readable (unless they won’t ever be read by a human), and that they describe the data that follows.
* I **repair** and reconcile anything that is not in a clean format.
* I **validate** that the changes took place and everything is in order, and
* I **document** anything that I changed.

One I’m happy with the column formats and meta data, I move on to the data types. Again,

* I **inspect** to make sure that what looks like a string is a string, and what looks like a number is a number, etc. If any of these don’t look right, I go back to how the data was collected and what that variable is supposed to represent.
* I then **repair** try to convert it to the appropriate type. If that works, great!
* I then **validate** that everything looks good and
* **document** that I changed the data type. If it doesn’t work, I look at the values and try to find the value–or values–causing the incorrect type.

I might manually **repair** it by converting to the correct type, then **validate** and **document** again.

I then go looking for missing data. If there is missing data:

* I **inspect** it by looking for patterns. If any variable has a lot of data missing, I look into it more closely to try to figure out why.
* If there is a systemic reason why the data is missing, I see if I can recover it and **repair** the missing data. I avoid imputing data because it can introduce as many problems as it solves.
* I then **validate** that everything was repaired and there is no more missing data that I haven’t accounted for.
* Finally, I **document** by making sure that the missing data points are appropriately labeled so that I know what was missing.

I then get the summary statistics for each variable, and **inspect** for outliers. I look at extreme they are, and for each record that contains an outlier, what the other values in that record are. I also check to be sure they are real and not mistakes.

WHY THIS RESPONSE?

The key to answering this question is that you have an organized data cleaning process and are aware of the biggest sources of error—including data types, missing data, and outliers.

This answer also touches on the fact that data cleaning is context-dependent, and any modification to the original dataset needs to be documented.

Free response

What data validation methods have you used?

…

**Feedback**

Not quite

**What you did well**

Unfortunately, your response was left blank and does not provide any information related to the question.

**What could be improved**

It’s important to provide a response to the question. In this case, you could discuss the different data validation methods you have used, such as form-level validation, regular expressions, checking for duplicated data, and so on.

Don’t worry, everyone has moments of uncertainty. Take a moment to review the material and try again. You’re capable of understanding this!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

I think about data validation in two phases:

* data collection
* data analysis

Whenever possible, I try to do data validation at the time of collection, and always do form-level validation if users are inputting data. For example, I use multiple choice or drop-down questions whenever possible, and regular expressions to be sure that user input matches the format I expect it to take, such as checking for the at-sign in an email. Finally, when collecting data I set ranges for everything that I can. For example, if I’m collecting age data, I make sure that it is within the bounds of what is humanly possible.

If I’m working with a dataset that has already been collected, I typically use regular expressions to be sure that all of the values are in the expected format. I also check for duplicated data and partially duplicated data using fuzzy matching techniques. It often comes up in data cleaning, but I also check for missing data at the data validation stage.

For very large datasets, I may take a sample because data validation can be very time-intensive. Plus, if there is one error of a particular type, there are likely to be multiple of the same type. I can then focus in on that type of error in a more efficient way.

WHY THIS RESPONSE?

This answer addresses the fact that data validation can occur at multiple stages in the data analysis pipeline. It also draws on the fact that errors are typically repeated, so you can work through more efficient mechanisms. Though there are a few open-source tools to help with data validation, you should already be very familiar with data validation using regular expressions from the [Data Analyst](https://www.codecademy.com/learn/paths/data-analyst/) or [Data Scientist](https://www.codecademy.com/learn/paths/data-science/) Career Paths.

A graph with two lines where there are many round holes of varying sizes cut out, in a fashion that resembles the holes of Emmental cheese.

Free response

Do you ever impute data? Why or why not?

**Get feedback**

Powered by OpenAI

Sometimes data is missing and you have to fabricate data, but it must be done cautiously.

Fantastic! It sounds like you have a good theoretical foundation for how to clean data - let’s try it out in some code.

### **Coding Questions**

Coding question

## **Questions**

Find the rows with missing values in this dataset.

* If a record has more than two missing values, remove it from the dataset.
* If a record is only missing one value, replace the missing value with the median of that variable.
* Create a new column where you indicate that a value was imputed (you do not need to indicate which value was imputed). Call this column Imputed.

The first 5 lines of this dataset are:

Name Salary Age Children Company

Vanna Holland 5738.45 40 5

Gretchen Kemp 9765.50 53 2 Feugiat Lorem Ipsum Company

Brittany Porter 8557.05 37 Magna Ut Industries

Kaseem Crawford 1424.80 20 4

Lamar Payne 2442.63 28 5 Amet Incorporated

### **Code**

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

Output:

Salary Age Children

count 91.000000 98.000000 95.000000

mean 5555.502088 44.112245 3.073684

std 2469.829777 15.158192 2.179411

min 1082.800000 18.000000 0.000000

25% 3754.985000 31.000000 1.000000

50% 5719.620000 44.500000 3.000000

75% 7507.900000 57.000000 5.000000

max 9765.500000 67.000000 6.000000

Salary Age Children

count 91.000000 98.000000 95.000000

mean 5555.502088 44.112245 3.073684

std 2469.829777 15.158192 2.179411

min 1082.800000 18.000000 0.000000

25% 3754.985000 31.000000 1.000000

50% 5719.620000 44.500000 3.000000

75% 7507.900000 57.000000 5.000000

max 9765.500000 67.000000 6.000000

**Run**

Loading Complete

**Check answer**

Answer Check Complete

Please define your new column Imputed.

### **Situational Questions**

Free response

Healthcare Company: We have millions of records of anonymized data, and a lot of appointment notes are missing. We want to determine if there is a relationship between the amount of time nurses spend with patients and their complaint category, but we cannot do that without the notes.

How would you deal with the missing data?

**Get feedback**

Powered by OpenAI

Missing data often tells its own story, either about the event itself or the collection process.

Free response

Non-profit organization: We are conducting a survey about attitudes towards policing. How would you determine if survey responses were filled in at random by certain individuals?

Your response

..

Our answer

I am assuming that there are multiple questions in this survey and it is a multiple choice. If I am involved in the design of the survey, I would include similar questions worded differently to use as a control. Likewise, I might include logically incompatible questions like “I eat beef” and “I observe a vegan diet”.

Once the data is already collected, I think about it as a flagging system where suspicious activities get flagged, and no one suspicious activity is itself a flag, but when there are multiple flags, that can indicate that the respondent filled in questions at random.

For example: First, I would look at how long it took each participant to fill in each question, and focus on the participants who filled in the questions very quickly as one suspicious group.

Second, I would look at any participant who answered all the questions with the same ranking (i.e., the highest, lowest, first, or last option). I would also consider this group to be a suspicious group. Finally, I would cluster respondents together, assuming that there are trends in the responses. Any respondent that does not fit into a cluster deserves closer investigation.

If I were involved in the design of the survey, I might include the same question worded differently to be sure that the answers match, and then use this as another flag.

When all of the flags were assigned, I would then look at the responses to determine if there are any patterns that I am missing or any other identifying information. I would then indicate that these are suspicious responses. I would then make a decision about whether to include them or not.

WHY THIS RESPONSE?

This answer demonstrates an awareness that there is no simple solution to this problem, and the best solutions will draw on multiple methods. It highlights multiple ways to think about how real people answer surveys, and what data relics are left when they do.

Also, this answer communicates a conservative approach to data management by including a step of manually evaluating the flagged records and checking your work.

### **Review**

Great job! It’s clear to me that you have really worked with cleaning data before!

In this interview, we covered:

* Dealing with missing data
* Imputing data
* Data validation
* General data cleaning

You gave some really thoughtful responses to all of these questions. Let’s move on to the next sectionl

**Summary Statistics Interview Questions**

### **Introduction**

<1 min

Welcome back!

In this part of the interview, we will cover some basic statistical operations in Python. This means that you will get a break from answering questions and have a chance to write some code.

Since this is an interview, you will still need to explain your thought process. In fact, we hope that after every function, you will create a short explanation about what your results mean.

We want to be sure that you know some basic data analysis functions and can implement them in Python.

Many companies ask these questions during the onsite interview; other companies may embed these questions into the take-home exercise or ask them in another way.

Get ready to show off your summary

[statistics](https://www.codecademy.com/resources/docs/general/statistics)

Preview: Docs Loading link description

skills!

**Instructions**

You’ll be working with a dataset of avocado prices, ripe for you to show off your summary statistics skills!

### **Measures of Central Tendency**

10 min

In this exercise, we are going to ask you some questions about price. You can ignore the total number sold—these questions are asking about ‘daily price’ rather than ‘price’.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Start by importing pandas with the usual pd convention, and any other packages you think you will need.  
  Then import the avocado.csv file as a DataFrame. Call this DataFrame df.  
  Print the first 5 rows to the console.
* Checkpoint 2 Passed
* **2.**
* Let’s get started by finding the average (simple mean) daily price of avocados for the entire dataset.  
  Save this variable as avg\_price and print it to the console.
* Check the pandas documentation or Stackoverflow if you get stuck.
* Checkpoint 3 Passed
* **3.**
* Now let’s get the median average daily price.  
  Save this variable as med\_price and print it to the console.
* Check the pandas documentation or Stackoverflow if you get stuck.
* Checkpoint 4 Passed
* **4.**
* In a few sentences, answer each of the following questions:  
  What does avg\_price mean?  
  Why are the mean and the median different?  
  When would you report the median instead of the mean?  
  Record your answers as multi-line strings.

import pandas as pd

df = pd.read\_csv("avocado.csv")

print(df.head())

avg\_price = df.AveragePrice.mean()

print(avg\_price)

med\_price = df.AveragePrice.median()

print(med\_price)

"""The average daily price of avocadoes was about $1.41. This is not the average price of an avocado, but the average price paid by day."""

"""The mean and the median are different because they represent different values. The mean is just the simple average whereas the median is the midpoint. These numbers suggest that there were some days when the average price was very high. There are probably some outliers making the mean price higher than the median."""

"""I would report the median when I want a number that is representative of avocado prices on an average day."""

### **Measures of Spread**

14 min

In this exercise, we are going to ask you some more questions about price. This time we will ask about the range of the prices. The price questions are asking about ‘daily price’ rather than ‘price’, so there is no need to take the units sold into account.

We are also going to ask you some questions about Total Volume sold. This value is an estimation across multiple cities (which is why you will see partial avocados sold).

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Start by importing pandas as pd and any other packages you might need.  
  Then import the avocado.csv as a DataFrame. Call this DataFrame df.  
  Print the top 5 rows of the dataset to the console. Notice that not all of the variables are displayed.
* Checkpoint 2 Passed
* **2.**
* First, let’s get a list of the column names.  
  Save this list as my\_columns and print it to the console.
* If you’re stuck, use your preferred search engine to check how to turn a pandas DataFrame’s column names into a list.
* Checkpoint 3 Passed
* **3.**
* This time, let’s take a look at the Total Volume column.  
  Find the maximum and minimum daily volume of avocados sold. Save these as max\_vol and min\_vol and print them to the console.
* This column presents a challenge because it has a space in the column name. If you have been referencing your columns with dot notation (i.e., df.AveragePrice), you’ll need to use bracket notation (i.e., df['AveragePrice']) when working with Total Volume.  
  If you’re stuck, use your preferred search engine to check how to get the maximum and minimum values from a DataFrame column.
* Checkpoint 4 Passed
* **4.**
* Wow! That is quite a range!  
  Calculate the range, save it as range\_vol, and print it to the console.
* If you’re stuck, use your preferred search engine to check how to get the range within a given DataFrame column.
* Checkpoint 5 Passed
* **5.**
* As in the previous exercise, the most important thing we can do is interpret these values.  
  What do min\_vol and max\_vol represent? What does range\_vol represent?  
  Save your responses as multi-line strings in the Output Terminal.
* Only 85 avocados were sold on the day with the fewest avocado sales. On the day with the most avocado sales, 62,505,647 avocados were sold.  
  This means that there was a range of 62,505,562 avocados!
* Checkpoint 6 Passed
* **6.**
* Let’s also find the following for the average daily price:
  + the minimum
  + the maximum
  + the range
* Save these values to min\_price, max\_price, and range\_price, and print them to the screen.
* Checkpoint 7 Passed
* **7.**
* Again, particularly in an interview setting, interpreting your result gives it meaning. So go ahead and explain what min\_price, max\_price, and range\_price represent.  
  Save your responses as multi-line strings in the Output Terminal.
* On the cheapest day for avocados, they sold for just 44 cents. On the most expensive day, they were $3.25. That’s a $2.81 range in avocado prices.

import pandas as pd

df = pd.read\_csv("avocado.csv")

print(df.head())

my\_columns = list(df)

print(my\_columns)

max\_vol = df['Total Volume'].max()

min\_vol = df['Total Volume'].min()

print(max\_vol)

print(min\_vol)

range\_vol = max\_vol - min\_vol

print(range\_vol)

"""Only 85 avocados were sold on the day with the fewest avocado sales. On the day with the most avocado sales, 62,505,647 avocados were sold."""

"""This means that there was a range of 62,505,562 avocados!"""

max\_price = df['AveragePrice'].max()

min\_price = df['AveragePrice'].min()

print(max\_price)

print(min\_price)

range\_price = max\_price - min\_price

print(range\_price)

"""On the cheapest day for avocados, they sold for just 44 cents. On the most expensive day, they were $3.25. That's a $2.81 range in avocado prices."""

### **Histogram**

13 min

Oftentimes, it is more useful to visualize the distribution of a variable before trying to conduct any analysis on it. This allows you to determine if it is skewed or normally distributed.

Since you’ve already calculated the range of these

[variables](https://www.codecademy.com/resources/docs/general/julia/variables)

Preview: Docs Loading link description

, we can dive right into visualizing the distribution of both Total Volume and AveragePrice.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Import pandas as pd and matplotlib.pyplot as plt.  
  Then load the avocado.csv dataset into a DataFrame called df.  
  Note: You’ll notice the codecademylib3 library has been imported for you. codecademylib3 is a package that Codecademy requires to plot your histogram. Don’t worry about this library — you won’t see it outside the Codecademy platform.
* Checkpoint 2 Passed
* **2.**
* Start by reminding yourself of the minimum and maximum values for Total Volume and AveragePrice from the previous exercise.  
  Call these max\_vol, min\_vol, max\_price, and min\_price.
* Review the previous exercise.
* Checkpoint 3 Passed
* **3.**
* Create a histogram of the AveragePrice column using the plt.hist() function.  
  Set the range from 0 to 3.5 and the bins to 14.  
  Remember to include plt.show() for all of your plots.
* Checkpoint 4 Passed
* **4.**
* Now let’s display a histogram of Total Volume.  
  This is a little more complicated since there are outliers. There are only a few days in our dataset when avocados enjoyed sales in the millions. We don’t need to display them on our histogram.  
  For this histogram, pick a range that makes sense to you.  
  Set the bins to 10.
* Checkpoint 5 Passed
* **5.**
* Finally, write 1-2 sentences describing what these histograms tell us about the distribution of our variables. Be sure to include whether they are normally distributed, skewed left, or skewed right.  
  Write your response as a multi-line string in the code editor.
* The daily price of avocados is normally distributed, and the daily total volume sold is right-skewed. This means that there are many more days when there were just a few avocado sales versus days with lots of sales.

import codecademylib3

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv('avocado.csv')

print(df.head())

max\_vol = df["Total Volume"].max()

min\_vol = df["Total Volume"].min()

max\_price = df.AveragePrice.max()

min\_price = df.AveragePrice.min()

plt.hist(df.AveragePrice, range=(0, 3.5), bins=14)

plt.show()

plt.hist(df["Total Volume"], range=(0, 7500000), bins=10)

plt.show()

### **Interquartile Range**

10 min

While range is a very common statistic to report, so is interquartile range. In this exercise, we will investigate the interquartile range of the two

[variables](https://www.codecademy.com/resources/docs/general/julia/variables)

Preview: Docs Stores data that can be manipulated and referenced throughout the code.

we have been working with:

* AveragePrice
* Total Volume

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Import any libraries you think you’ll need to calculate interquartile range.  
  Go ahead and start by loading the avocado.csv dataset into a DataFrame called df.
* We recommend importing the pandas and numpy libraries and scipy.stats.  
  However, you can perform the following calculations using the technique of your choice.
* Checkpoint 2 Passed
* **2.**
* Calculate the interquartile range of AveragePrice and save it with the variable name, iqr\_price. Print iqr\_price to the screen.
* We recommend using the SciPy function dedicated to computing an interquartile range (from scipy.stats import iqr)
* Checkpoint 3 Passed
* **3.**
* Now calculate the interquartile range for Total Volume and save it with the variable name, iqr\_vol.  
  Print iqr\_vol to the screen.
* Checkpoint 4 Passed
* **4.**
* Wow! The range of Total Volume was over 62,000,000! Why is the interquartile range so much smaller?  
  Explain what iqr\_price and iqr\_vol mean for our data. Write your answer in the code editor as a multi-line string.
* The interquartile range is the middle 50% of observations. So, 50% of all daily avocado prices fall within $0.56 of each other.  
  For Total Volume, it means that 50% of all daily sales are within 422,123 of each other.  
  Those spans are much narrower than the total range, which suggests that there are outliers above and below the median.

import numpy as np

import pandas as pd

from scipy.stats import iqr

df = pd.read\_csv("avocado.csv")

print(df.head())

iqr\_price = iqr(df['AveragePrice'])

print(iqr\_price)

iqr\_vol = iqr(df['Total Volume'])

print(iqr\_vol)

"""The interquartile range for the price is just 56 cents. This means that the majority of avocados, or the middle 50% of avocadoes were sold within 56 cents of each other"""

"""The middle 50% of total volume of daily sales was within 422,123. The fact that this is so much smaller than 62,000,000 demonstrates that there is a huge range in the top quartile, incidating some outlier days for the sale of avocados"""

### **Association between Variables**

11 min

In this final exercise, we’d like you to determine if there is an association between AveragePrice and Total Volume.

If we lower the price of avocados, do people also purchase fewer of them? You do not need to determine if there is a causal relationship between the

[variables](https://www.codecademy.com/resources/docs/general/julia/variables)

Preview: Docs Loading link description

, but you should be able to determine if there is an association.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Import matplotlib and pandas, as well as any additional packages you’ll need.  
  We recommend importing the SciPy function that will help you calculate a correlation coefficient.  
  Load the avocado.csv dataset into the workspace.  
  Call this DataFrame df.  
  Note: You’ll notice the codecademylib3 library has been imported for you. codecademylib3 is a package that Codecademy requires to plot your histogram. Don’t worry about this library—you won’t see it outside the Codecademy platform.
* We recommend importing the pearsonr function from scipy.stats to calculate this.
* Checkpoint 2 Passed
* **2.**
* We want to determine if there is an association between AveragePrice and Total Volume.  
  Start your investigation by creating a scatter plot of the AveragePrice and Total Volume variables. Place AveragePrice on the x-axis.
* Checkpoint 3 Passed
* **3.**
* Even though we can visually determine that there is no association between these variables, we still need a quantitative measure to report.  
  We’d like you to calculate the Pearson Correlation coefficient to determine definitively if there is a linear relationship between the daily price of avocados and the number of avocados sold.  
  Save your coefficient value to corr\_price\_vol and print it to the console.
* Checkpoint 4 Passed
* **4.**
* Great, you have a number to describe this association (or lack of association).  
  How would you report these findings to relevant stakeholders? Write your answer as a multi-line string.
* A Pearson correlation coefficient near to 0 indicates that there is no linear association between the variables.

import codecademylib3

import pandas as pd

import matplotlib.pyplot as plt

from scipy.stats import pearsonr

df = pd.read\_csv("avocado.csv")

# create your scatter plot here:

plt.scatter(df['AveragePrice'], df['Total Volume'])

plt.xlabel('Average Daily Price')

plt.ylabel('Total Daily Sales')

plt.show()

# calculate corr\_sqfeet\_beds and print it out:

corr\_price\_vol, p = pearsonr(df['AveragePrice'], df['Total Volume'])

print(corr\_price\_vol)

"""The Pearson correlation coefficient is approximately -.19. This is very low, and indicates no linear relationship between these variables. The scatterplot of Price versus sales shows two distinct clusters. It may be worth investigating these clusters independently in another round of evaluation."""

### **Review**

<1 min

Congratulations! You are ready to answer summary

[statistics](https://www.codecademy.com/resources/docs/general/statistics)

Preview: Docs Statistics is the science that is concerned with methods for collecting, organizing, analyzing, and interpreting data.

questions that might come your way in your data analyst or data scientist interview.

In this lesson, we covered:

* Calculating the mean and median of a variable and how to interpret them.
* Calculating and reporting the range of a variable
* Creating a histogram and evaluating it for skewedness.
* Calculating the interquartile range and comparing it to the standard range.
* Evaluating the relationship between quantitative   
  [variables](https://www.codecademy.com/resources/docs/general/julia/variables)
* Preview: Docs Loading link description
* to determine if there is a linear association.

Good luck and see you in the next interview!

Note that the dataset used in this lesson was downloaded from [kaggle](https://www.kaggle.com/neuromusic/avocado-prices) and originally compiled from the [Hass Avocado Board](http://www.hassavocadoboard.com/retail/volume-and-price-data).

# **Sampling Interview Questions**

**Practice answering common sampling interview questions**

### **Introduction**

Welcome back! This is the sampling portion of your interview. Sampling is a large part of what we do on our team.

It would be great if we could collect data from our entire population, but that just isn’t realistic. So smart sampling is absolutely essential. I hope you are up for the challenge and well versed in all the different ways to sample from a population. On our team, we use a variety of tools, but the most important tool of them all is knowing when to use each one!

You’ve done a great job talking through your thought process and explaining why you implement different methods so far, and I look forward to your answers in this section, too!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Conceptual Questions**

Let’s start with some general questions about sampling.

Free response

How do you decide which sampling technique to use and what do you take into account when deciding?

**Get feedback**

Powered by OpenAI

Organize your thoughts by making a list of research goals and a list of potential sampling techniques.

Four individuals standing in front of an image of the Earth. The person on the left has medium-dark skin and a beard, and is wearing an orange checked shirt and purple turban. The second person has dark skin, longish black hair, and a pink and purple striped shirt. The third person has light skin, a bright red headscarf, and purple and pink dotted shirt. The fourth person has yellow hair, medium-dark skin, and a blue buttoned shirt

Free response

How do you typically create a representative sample?

**Get feedback**

Powered by OpenAI

Consider your sampling frame first. Then consider which characteristics you need to control for, how you pick those characteristics, and how you represent them in your sample.

Free response

What is selection bias and how do you account for it in your data?

**Get feedback**

Powered by OpenAI

Free response

How do you determine sample size?

**Get feedback**

Powered by OpenAI

Great job! It’s clear to me that you have a good sense of the theory behind sampling. Let’s move on to some applied questions.

### **Applied Questions**

Free response

A publishing house: We have a user base of about 1 million customers. About 80% only access Freemium content, and the other 20% have some mixture of paid subscriptions. We want to learn more about customers’ interests so we can maximize our revenue. We have a budget for a survey. How would you approach this project? Random sampling or Stratified sampling - why? Focus on all customers or only the 20% paid?

**Get feedback**

Powered by OpenAI

A Random Sample would probably over-represent the freemium customers. How could you be sure to capture enough of the paid subscribers but still use random sampling?

Free response

A grocery store:

200 customers enter our store every day. We want to know what path they take through the store and what sections they visit first. How will you determine which customers to track?

**Get feedback**

Powered by OpenAI

Consider the fact that there is a time aspect of when people visit grocery stores.

Free response

A healthcare company:

We want to know how many people in a community have a respiratory illness. We have established a testing center where people can come to be tested for free. What steps would you take to account for the bias in the sample?

Your response

Give the best

Our answer

Even when something is offered for “free”, in the sense that it is given without payment, people have to have the time and transportation to access it. I tend to think of this as the intangible ways that people have to pay.

This sample will be biased towards those who are able to reach the location and who have the flexibility in their schedule to make the trip. They can pay with non-monetary resources. We can assume that more people who live closer will be represented, but that is an assumption. Depending on the demographic makeup of the community and specific transportation routes, that could be an inaccurate assumption.

The best way to get a measure of the community incidence would be to take a survey of test takers and align that to the demographic distribution of the community. However, it wouldn’t be a direct demographic comparison. Since there are other socio-economic factors that will contribute to who is able to take the test, these have to be considered.

For example, if taking a test meant missing work, people who cannot afford to miss work might avoid going to take a test or might only take one if they are very very ill. This has the potential to both make it seem like a higher proportion of that community is actually ill, but also may limit the ability to make an accurate estimate.

It would also be beneficial to ask some survey questions alongside demographic information to address these issues. However, this has to be balanced with concerns around privacy and confidentiality. Because those concerns are outside of my area of expertise, I would likely work with other stakeholders who can advise on what questions would be appropriate.

WHY THIS RESPONSE?

This is a good response because it addresses real world practicalities that are often overlooked when providing “free” services or items. The interviewee further considers the limitations of strictly normalizing on demographics and introduces through the use of a clear example some socio-economic factors that have to be considered. Finally, the interviewee talks about external issues outside of their control and the relevant stakeholders that they would work with in order to resolve them. This is another very realistic possibility that the interviewee takes into consideration which shows that they are also able to think outside of abstract concepts.

### **Review**

Congratulations! Sampling can be tricky, but you really know how to find the sample from a population!

In this interview, we covered:

* How to decide what sampling techniques to use,
* How to create a representative sample,
* What selection bias is and how to deal with it, and
* The best way to determine sample size.

You also made use of some general strategies to answer sampling problems such as:

* The ability to think out loud and consider a wide range of constraints when identifying the right sampling technique.
* The use of practical examples to illustrate the abstract concepts such as creating a representative sample.
* Determining and estimating the inputs to answer complex questions like determining sample size.
* Why picking a different technique could cause us to lose out on valuable information.

You also applied your skills by explaining how:

* A publishing house could use Stratified Random Sampling to answer their questions,
* A healthcare company could estimate the presence of an illness in a community, and
* A grocery store could measure how customers move through their store.

You are ready to showcase your sampling skills in a real-life interview, good luck!

# **Probability Interview Questions**

**Practice answering common interview questions about probability.**

### **Introduction**

Hi again! It’s me, Codey. Have you ever wondered what the odds are of running into probability problems as a data professional? You guessed it - pretty high!

That is why in this section I am excited to test your problem-solving skills through some fairly common, yet somewhat tricky, probability problems. While the problems we solve in our work are of a much broader scope, successful candidates have an intuitive understanding of counting and probability concepts, and are able to quickly take the simpler and more efficient route towards the solution. I certainly hope you are able to as well!

Do you have your pen and paper ready? You’ll have to work your way through some probability problems fairly quickly, but please don’t leave me in the dark! Whenever you can, I would really like to hear about your thought process to understand how you approach and solve problems. Let’s get started!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Conceptual Questions**

**Image of person waving from inside a train car. The person is on the leftmost window. There are two windows and 2 doors in the center. The car is atop 2 lines representing the tracks.**

Free response

On any given day, there is a 10% chance you will see a friend on the subway. Assuming you ride the subway 5 days a week, what is the probability that you will see a friend at least once?

**Get feedback**

Powered by OpenAI

It may be easier to calculate the probability of never seeing a friend rather than actual probability of seeing a friend.

Free response

A couple has 2 children. At least 1 is a girl. What is the likelihood that both are girls?

**Get feedback**

Powered by OpenAI

Try writing out all the possible combinations.

Free response

A fair six-sided die is rolled twice. What is the probability of getting 1 on the first roll and not getting 6 on the second roll?

**Get feedback**

Powered by OpenAI

They are independent events.

Great job with the conceptual questions! It’s clear to me that you have a solid grasp of the theory! Let’s move on to implementing these principles in code.

### **Coding Challenges**

Coding question

## **Questions**

How would you simulate the behavior of a fair coin?

### **Code**

1

2

Output:

**Run**

Loading Complete

**Check answer**

Answer Check Complete

👏

You got it!

Coding question

## **Questions**

How would you simulate the behavior of an unfair coin?

### **Code**

1

Output:

**Run**

Run your code to check your answer

Great job! You clearly have the ability to simulate probability problems computationally, which is a really important skill to have when working with data! Let’s move on to applying these concepts to real-world scenarios.

### **Applied Questions**

Free response

Facebook: This question was famously asked at Facebook and has since become representative of questions about probability. We have two options for serving ads within Newsfeed:

* Out of every 25 stories, one will be an ad.
* Every story has a 4% chance of being an ad.

For each option, what is the expected number of ads shown in 100 news stories? If we go with option 2, what is the chance a user will be shown only a single ad in 100 stories? What about no ads at all?

**Get feedback**

Powered by OpenAI

Free response

You call 2 UberX’s and 3 Lyfts at the same time. The time to reach you can be considered independent and identically distributed. What is the probability that all the Lyfts arrive first? What is the probability that all the UberX’s arrive first?

**Get feedback**

Powered by OpenAI

Free response

If three friends in London told you it’s raining, and there is a 1/3 probability that each person is lying, what is the probability that it’s raining in London?

**Get feedback**

Powered by OpenAI

### **Review**

Thanks so much for taking the time to work through some fairly tricky probability problems! In this interview, you encountered a range of problems and demonstrated your understanding of key concepts through the application of problem-solving strategies and clear communication of your ideas.

Here are some concepts that you used in the problems that you encountered:

* The idea of complementary, conditional, and independent events.
* Creating a script to solve a relatively simple simulation problem.
* The distinction that the chance of an event happening is not the same as an event that is guaranteed to occur x times every n trials.
* Conditionally independent events, and how they can make traditional methods fail (such as complementary counting).

You really showcased your strong and intuitive grasp of probability concepts today! Now you are ready to do the same in a real-life probability interview as well. Good luck!

# **Estimation Interview Questions**

**Have fun estimating answers for estimation interview questions**

### **Introduction**

Hello, again! It’s me, Codey.

Have you ever wondered how much space it might take to gather the entire population of the world? But wait a minute—nobody can magically command everybody to gather at the same place at the same time. So who knows, right?

Indeed, it would be impossible to conduct a real-world experiment for these kinds of goofy problems. And that is where some clever estimation skills can come in handy. Don’t get thrown off by the seemingly bizarre nature of these problems because the same estimation skills used to answer them can be extended to a real-world data analysis project as well. This is what you’ll be tested on here.

I love it when interviewees talk through their solution by first laying out the parameters:

* What information do you need to know to calculate this estimate?
* What pieces of information are dependent on other pieces of information?

It’s also really helpful when candidates break their answers down into steps.

Do your best and try to make a reasonable guess!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Estimation Questions**

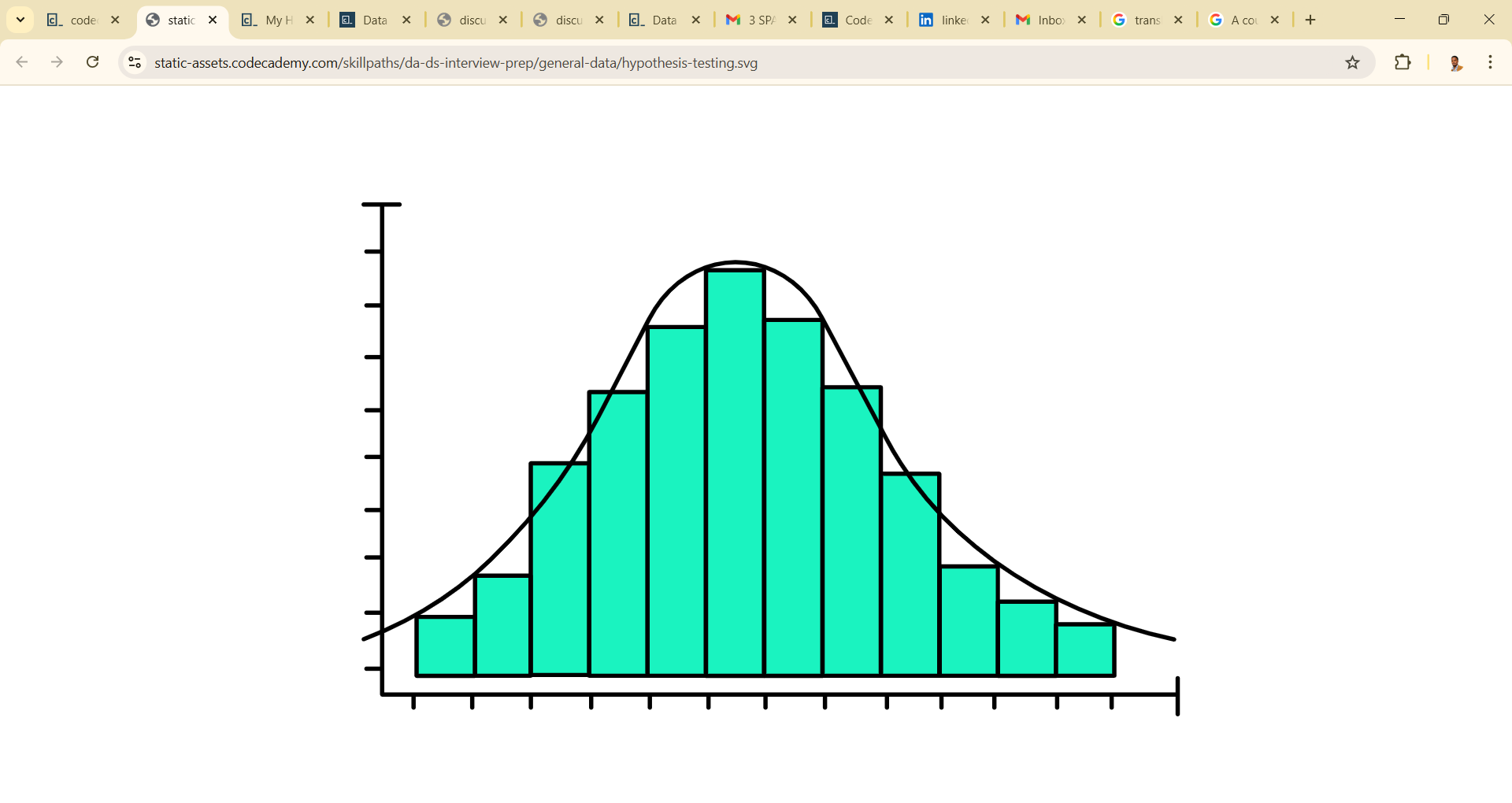
Free response

How many people can fit on a subway train in Tokyo?

**Get feedback**

Powered by OpenAI

What do you need to know to figure out how many people fit into one train car?



Free response

How many boxes of strawberries are for sale in Paris today?

**Get feedback**

Powered by OpenAI

How many households are there in Paris and how many strawberries do you think they consume?

Free response

How many bicycles would you need to set up a rideshare program in Boston?

**Get feedback**

Powered by OpenAI

How many people would use the bikes at any given moment?

Free response

How would you estimate the value of pi?

**Get feedback**

Powered by OpenAI

Leverage the ratio of a circle to a square and conduct a simple experiment.

Free response

How many tourists will see a billboard in Times Square in a typical year?

**Get feedback**

Powered by OpenAI

Start with the total population of the world and take small steps to reduce the population to something closer to the one in question.

### **Review**

Here are some key ideas you used in answering these estimation problems:

* Laying out the parameters—what information do you need to know and which pieces of information are dependent on others?
* The number of units that can be fit into some space is equal to the volume of the space divided by the volume of the total units (e.g., a metro).
* The proportion of people doing an activity right now is roughly equal to the proportion of the entire time an average person spends doing the activity (e.g., a ride-share program).
* Taking the geometric mean of the minimum and maximum range allows you to preserve the order of magnitude across the entire range and narrow in on a somewhat precise estimate (e.g., NY tourists who see a billboard).
* Solving estimation problems bottom-up (e.g., boxes of strawberries for sale) as well as top-down (e.g., NY tourists who see a billboard)

Great job estimating these somewhat tricky numbers! If you apply these same concepts and strategies to a real-life interview, we estimate *your estimates* will be quite precise.

In the worst case, even if you are off on a problem or two, you will leave the interview demonstrating your ability to:

* consider different parameters
* break down open-ended problems into steps

And both of these skills are commendable traits in a successful candidate. Good luck!

# **Hypothesis Testing Interview Questions**

**Practice answering common interview questions about hypothesis testing**

### **Introduction**

Hi, welcome back!

As you know, my team is in need of someone who can help us answer our data questions. We have a lot of data, I’d say maybe too much for one person to handle. I’m looking for a candidate who can work with samples and draw conclusions with a certain level of confidence.

As with any candidate, I am also looking for a problem solver who can clearly communicate their thinking and knowledge, someone who is curious about our data and has ideas on how to solve our specific business problems. I bet that person can be you!

Let’s dive back in with a few specific questions to give you a chance to show off your knowledge of hypothesis testing.

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Conceptual Questions**

Free response

What is the difference between type I and type II errors? Which one is worse? Why? Can you give me an example of each?

**Get feedback**

Powered by OpenAI

* Type I error is a false positive.
* Type II error is a false negative.

Free response

What is a p-value and what does it tell you?

**Get feedback**

Powered by OpenAI

* P-value tells us if we can accept the null hypothesis or reject it.

An unlabeled histogram with 12 bars depicts an almost-normal distribution, slightly positively skewed.

Free response

What is the Central Limit Theorem and why does it matter?

**Get feedback**

Powered by OpenAI

* The larger the sample size, our sample distribution will approach a normal distribution

Great job! It really sounds like you understand hypothesis testing. Let’s move on to applying these concepts to real-life scenarios.

### **Applied questions**

Free response

Finance company:

We assume that the average rate of return of a mutual fund is 9%. Talk through the steps of how you would validate if this is true. State any assumptions you would need to make and the data that you would need to validate your results.

**Get feedback**

Powered by OpenAI

This is a typical hypothesis testing scenario, so start by stating the null hypothesis.

Free response

Mortgage Lender:

We think that first-time homebuyers take out smaller loans than other borrowers. What information do you need to assess this and what would you do?

**Get feedback**

Powered by OpenAI

* You’ll want to perform a hypothesis test, what are the steps you need to take?
* Are there any assumptions you need to make about your sample data?

Free response

A marketing company:

We think that at least 60% of American households have an iPad. Our research team surveyed 200 households and found that 123 of them had iPads. Can we be sure that 60% is accurate? Why or why not?

**Get feedback**

Powered by OpenAI

* You’ll need to find your sample mean and build a confidence level around to determine the range of means we believe the population falls in.

Congratulations! You have done a great job with these hypothesis testing questions. It is clear to me that you know how to organize an experiment and evaluate the results.

### **Review**

In this portion of the interview, we covered:

* Step-by-step process for conducting hypothesis tests.
* The importance of representative samples when performing hypothesis tests.
* Code that can be used to calculate the confidence interval of a sample with 95% confidence that it is reflective of the true population.
* The theory behind the central limit theorem and a clear example.
* How vital the p-value is to any hypothesis test and what information you can gain from that value.

With all this preparation, you are going to do great in your next interview! You are ready to talk about hypothesis testing in a real interview setting!

# **Data Cleaning & Summarizing Take-Home**

**Practice your data cleaning and summarizing skills with a sample take-home assessment.**

### **Overview**

Welcome to the take-home assessment!

This exercise helps interviewers get a sense of your technical skills before moving you to the next stage of the interview process. The style and format of these projects can vary widely depending on the company and role you are applying for. In this project, we will give you a prompt and you are free to answer it however you wish. You can use any coding language and any tools you are familiar with. This is an opportunity to show off your skills.

The projects that we have prompted here are typical for data analyst roles but are not definitive. However, there are some widespread practices and conventions that you should be aware of before starting the take-home.

* There may be one or more very simple question(s). These are not tricks – they are meant to easily exclude applicants who cannot perform basic tasks.
* There will be some discussion, analysis, or reflection. Expect to have to write a summary of your findings. A large part of both Data Analyst and Data Scientist roles is communication. How well you communicate your findings is just as important as what you find and how you find it.
* You may be asked to submit your files in a specific way. Pay attention to these instructions and submit your file in the requested format. This is another area that may initially seem unimportant but can have an outsized influence on your candidacy.
* You may *not* be asked to submit your files in any particular way. If this is the case, Python files should be .py or .ipynb, R should be in .R or .Rmd format. Similarly, your written analysis should be in .pdf format or .md if submitted online.
* If anything is unclear or you are not sure what they are asking you to do, ask! It is much better to get clarification at the beginning of a project than submit something that doesn’t answer the question. This assignment is an opportunity to showcase your problem-solving skills. If you are stuck on a particular technical skill, try to find the answer on your own. Google and StackOverflow will be your friends during this process! If you still can’t figure it out, and the interviewer has allowed for questions, reach out and explain what you’ve already tried. This will show you have the ability to problem-solve on your own before reaching out for help.
* Your interviewers may give you an estimate of how long they think it should take. It is very common to go far beyond the time estimate.
* The take-home assessment is also a good opportunity for you to decide if this is the kind of work you are interested in doing. If you enjoy it, you are probably in the right place. If it is something you do not enjoy, it may be worth clarifying what the expectations of the job are.

Let’s get started!

### **Context**

You are applying to be a Data Analyst at a digital marketing agency. A client in the UK wants to start advertising on YouTube and needs your help to perform some analysis on video data.

**Datasets:**

* [GBvideos.csv](https://www.kaggle.com/datasnaek/youtube-new?select=GBvideos.csv)
* [USvideos.csv](https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv)

Note: These datasets were both compiled by Mitchell J and are available on [Kaggle](https://www.kaggle.com/datasnaek/youtube-new).

### **Instructions:**

* Load your datasets and perform some simple exploratory data analysis.
* Create a new column to represent the like-dislike ratio (likes divided by dislikes) for the videos in each dataset.
* What is the average like-dislike ratio for all of the videos? (average number of likes per video divided by the average number of dislikes per video).
* Perform data manipulation to find the average number of likes per video for the US and GB in the year 2018.
* Business report: Your client hypothesizes that the most polarizing videos get shared the most. Analyze the data to determine whether your client is correct, and what other metrics should be investigated. Be sure to define any terms or cutoffs you make in the data. Present your findings in a short report for the client to review.

### **Checklist:**

Here is a checklist of things to consider as you work on this project:

* Make sure to perform EDA (descriptive statistics, data visualizations, etc).
* Make sure your code is clean and well-documented.
* Assessment questions such as Q2 and Q3 only have one correct answer. Make sure to double-check your math and write 1-2 sentences interpreting the values.
* Don’t forget to deal with special edge cases throughout, such as a division by 0 error.
* Make sure to clearly address the client’s issue in the business report.

**Note**: a take-home won’t typically include a checklist like this one, but we’ve included one here for you as you become more familiar with the take-home.

# **Introduction to Visualization, Communication, and Experimental Design Interviews**

**See what you’ll cover in Visualization, Communication, and Experimental Design**

### **Goals of this Unit**

The goal of this unit is to get practice answering questions about data visualization, communication, and experimental design.

After this unit, you will be able to:

* Discuss how you pick a data visualization tool.
* Use data visualization to answer a practical business question.
* Describe your approach to experimental design.
* Talk to your interviewer about how you conduct A/B tests.
* Evaluate data from a real A/B test to make a recommendation about website layout.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# **Review of Fundamental Data Interview Questions**

**See what you’ve learned about answering fundamental data interview questions**

### **Review**

Congratulations! The goal of this unit was to give you practice answering general data questions.

Having completed this unit, you are now able to:

* Talk through how you work with data generally, including how you manage different types of data.
* Define your data cleaning process and prepare to answer questions about how you approach raw datasets.
* Implement basic summary statistics methods in Python and explain the significance of the results.
* Speak to how you use different sampling methods for different problems.
* Walk through solutions to probability interview questions.
* Approach estimations problems logically and systematically.
* Apply hypothesis testing to practical business problems.
* Effectively communicate your thought process and problem solving skills.
* Clean and summarize a real dataset.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# **Visualization Tools and Techniques**

**Practice answering data visualization interview questions**

### **Introduction**

Hi, it’s me, Codey, again!

In this part of the interview, we will talk about how you approach the data visualization process and which tools and techniques you use.

Data visualization is an important skill to have as a data professional because it allows you to tell meaningful stories and give your data visual context. It is especially important for a business intelligence role or a data analyst role in a product or marketing department.

There is no better place to show off your data visualization skills than your project portfolio. But I also like to ask candidates like you some visualization questions to make sure you really know your stuff.

I try to evaluate a few things:

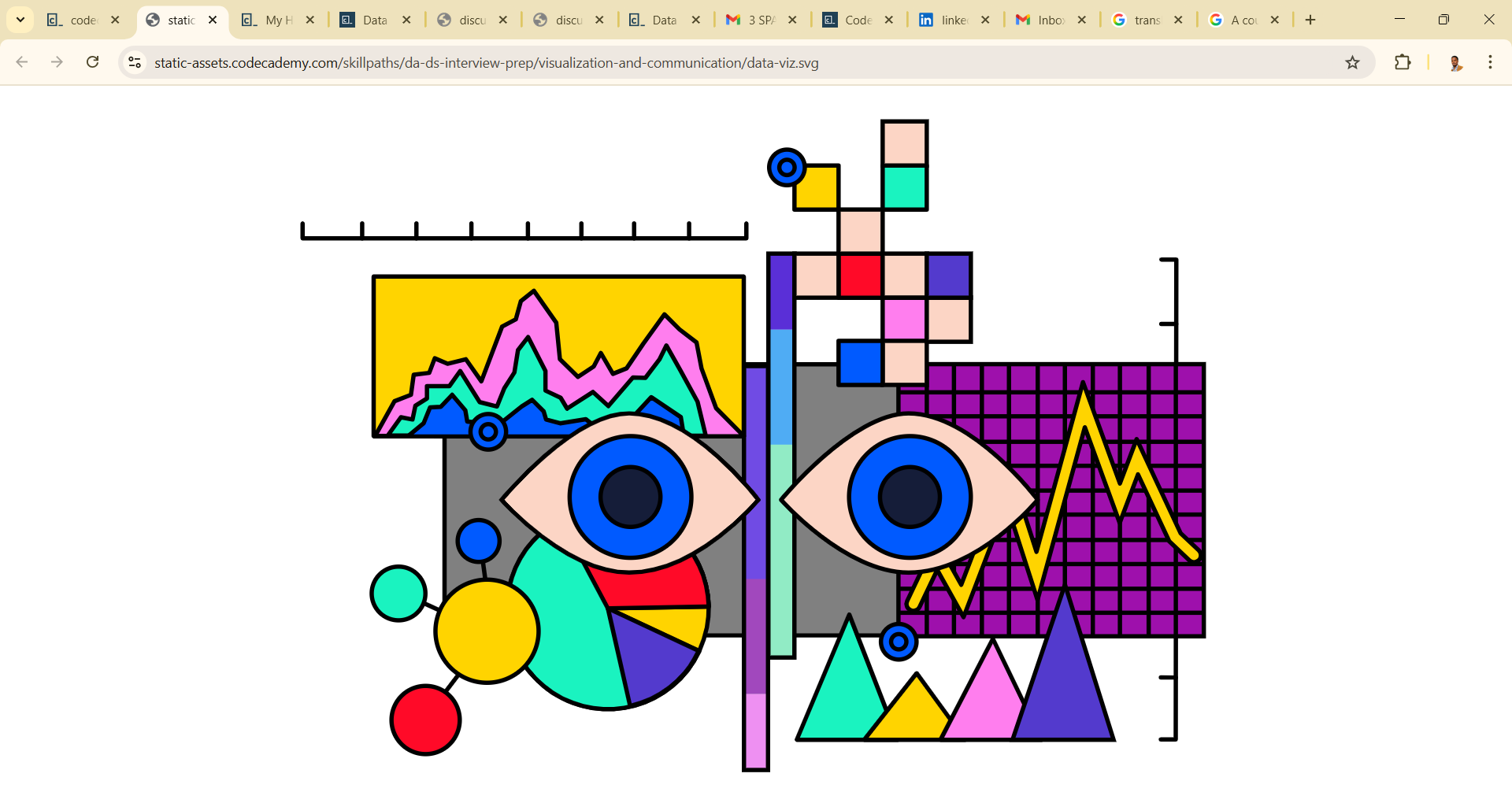
* Do you understand the data visualization process?
* How familiar are you with common data visualization charts, tools, and techniques?
* Do you know which tools to use for EDA (Exploratory Data Analysis) and Communication?

I hope you’re prepared to talk about your experience with visualizing data because I’m excited to hear about it!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Fundamental Data Visualization Questions**

****

Free response

What makes a good visualization?

**Get feedback**

Powered by OpenAI

To answer this question, think about the main components that make up the data visualization process:

* Preparing/Cleaning
* Visualizing
* Styling

Free response

What types of charts do you gravitate towards and why?

**Get feedback**

Powered by OpenAI

Think about the types of charts you often turn to when visualizing data. Do you have a favorite? Feel free to talk about why you like it so much.

Also consider the charts you use in different situations, such as when you want to represent one of the following:

* Relationship
* Comparison
* Distribution
* Composition

Free response

How would you represent six variables of a dataset?

**Submit Response**

To answer this question, think about how you would approach multivariate analysis differently than univariate and bivariate analysis.

Remember that the more variables you work with, the more difficult it becomes to visualize data in a chart. What are some ways to deal with this?

Free response

Which tools do you use for EDA and communication?

**Get feedback**

Powered by OpenAI

This is not a trick question! The interviewer just wants to see whether you are familiar with tools used for various purposes.

### **Review**

Great job! It is clear that you have the data visualization skills and knowledge required for the job. Just to recap, we’ve talked about:

* What makes a good data visualization
* Common visualization charts, tools, and techniques
* Visualization for EDA and communication

You are now ready to move to the next stage of the interview process - see you there!

# **What is a Portfolio Project?**

This project is a little different from other Codecademy projects you’ve encountered. In this project, you will bring together what you have learned in previous lessons to build a project in your own development environment and publish it to the web!

Create your own files outside of the Codecademy platform

Write your own code with less guided instructions

Use common project management processes to track your progress

# **How Do Portfolio Projects Work?**

We’ll provide you with high-level tasks to guide your project to completion, but you will be responsible for deciding how to implement them in your code.

There are many possible ways to correctly fulfill all of these requirements, and you should expect to use the internet, Codecademy, and other resources when you encounter a problem that you cannot easily solve.

Note that there are hints that can assist you, but they will only provide one potential implementation. Do not worry if your program looks different from ours!

# **Project Prompt**

## **Visualization Portfolio Project**

In this project, you will use your data visualization skills to help a company solve a business problem.

Click Begin to get started!

**Project Overview**

Companies like to see that you are able to create meaningful visualizations and solve real-world problems.

In this portfolio project, you will get an opportunity to show off your data visualization skills while solving a business problem for a technology company using real-world data. The following datasets are good candidates to showcase your data visualization skills:

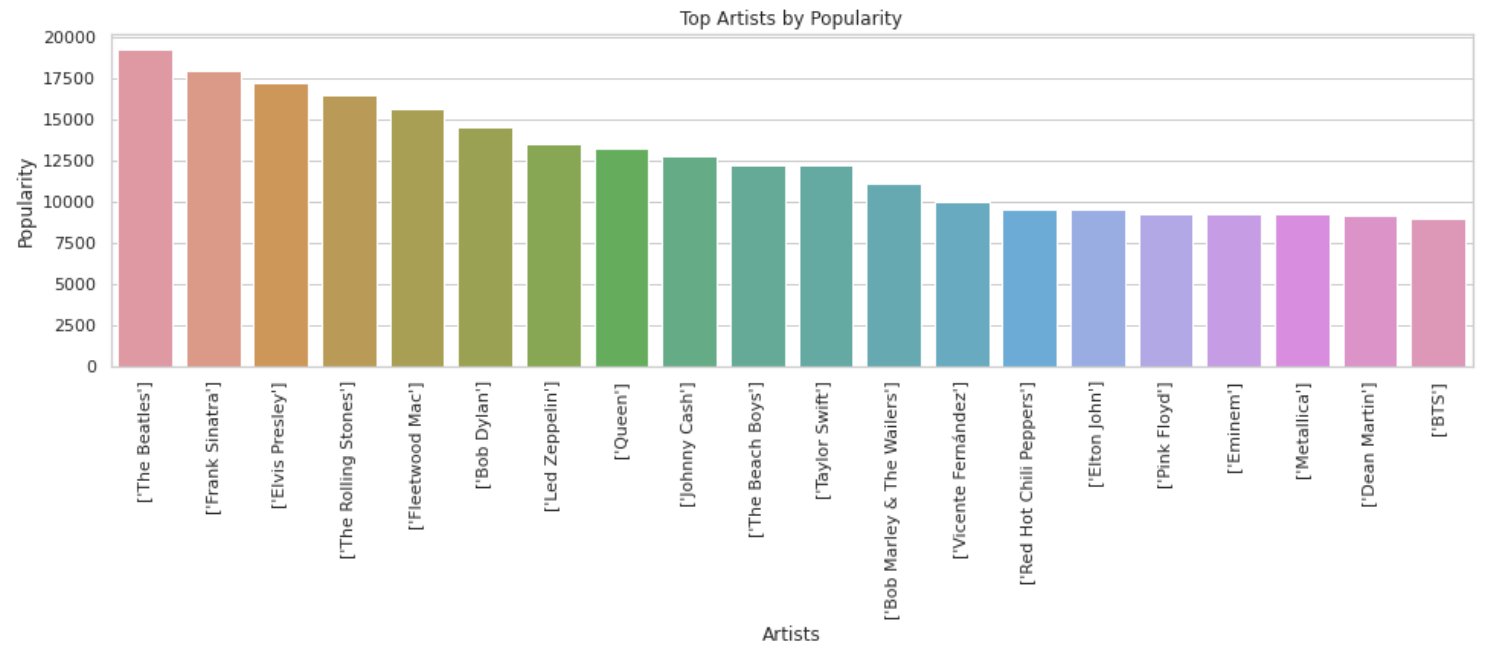
* Spotify [dataset](https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks) featuring 600,000 songs released between 1922 and 2021.
* Airbnb [dataset](http://insideairbnb.com/get-the-data.html) featuring listings in a city of your choice.
* Google [dataset](https://www.kaggle.com/lava18/google-play-store-apps) featuring 10,000 Google Play Store apps.

If you’d like, feel free to explore other datasets. Here are some places you can find datasets:

* [Kaggle](https://www.kaggle.com/datasets)
* [FiveThirtyEight](https://data.fivethirtyeight.com/)
* [UCI](https://archive.ics.uci.edu/ml/datasets.php)
* [awesomedata](https://github.com/awesomedata/awesome-public-datasets)

Think about the company you are interested in applying to and the type of data they interact with on a daily basis. This project is a great opportunity for you to show off your visualization skills using that data.

**Example Project**

****

**Project Objectives**

* Complete a project to add to your portfolio
* Choose a dataset and come up with a business problem to solve
* Create data visualizations to help solve a problem or answer a question
* Write documentation describing the dataset, the problem, and key takeaways

**Prerequisites**

* Data Wrangling and Tidying
* Summarizing Quantitative Data
* Data Visualization
* Communicating Data Science Findings

## **Project Tasks**

Keep track of your progress by dragging each task from "To Do" to "In Progress" to "Done" as you work on them. You can also click on a task to see more information about it.

To Do

**Setting up your Git Repository**

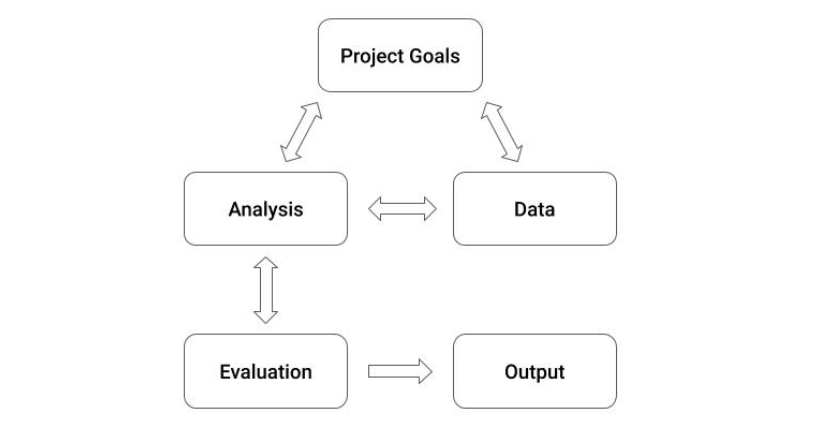
Create a new Git repository for this project.

**Project Scoping**

Properly scoping your project will greatly benefit you; scoping creates structure while requiring you to think through your entire project before you begin.

1. Start by stating the goals for your project.
2. Gather the data.
3. Consider the analytical steps required.

A proper project scope can be a great road map for your project, but keep in mind that some down-stream tasks may become dead ends which will require adjustment to the scope.



**Choose a Dataset**

Select one of the following datasets to work with or choose another dataset you’d like to explore:

* Spotify music data
* Airbnb listing data
* Google apps data

If you’re applying to a company that interacts with a particular type of data, this is a great opportunity to show off your data visualization skills using that data.

**Pose and Solve a Problem**

Look at your dataset and get curious!

Define a question you want to answer or problem you can explore. Go through each variable, what do you want to know more about?

Companies love to see that you are able to create meaningful data visualizations that answer real-world questions and address practical problems.

Make sure that the problem is worth exploring and can be solved in a reasonable amount of time.

**Load and Check Data**

Load your data into Python, R, or another environment. Check for missing data, consistency, and outliers. If necessary, clean it to make sure it is ready for data exploration.

**Explore and Explain Data**

Once you have your data, it’s a good idea to get acquainted with it. Gather some summary statistics and visually examine your data. Don’t forget to keep track of the questions you think of while exploring. It is also a good idea to start writing out some of the insights you have gain in this phase.

**Visualize Data**

Create meaningful data visualizations that help answer the business question(s) you set out to explore. Feel free to use any of the following libraries and tools for your visualizations:

**Python:**

* Matplotlib
* seaborn
* Plotly

**Business Intelligence:**

* Tableau (BI)
* Looker
* Power BI

**R:**

* ggplot2

**Web:**

* [D3.js](http://d3.js)

**Communicate Findings**

Once you’ve performed your analysis and visualizations, you’re ready to communicate your findings. This can be a written document, slide deck, or other format.

Make sure to clearly communicate:

* The data you worked with
* The question or problem you set out to solve
* Key findings and takeaways

# **Experimental Design Interview Questions**

**Practice answering experimental design interview questions to solve real-world business problems**

### **Introduction**

Welcome back! It’s me, Codey, again.

In this part of the interview, I’m going to ask you a few questions about setting up experiments and designing data collection to meet business needs. This is my favorite part of the interview because I love talking to candidates about real problems and hearing all of their great ideas. We have tons of data and I’m so excited to get a fresh perspective on it!

The best candidates are going to show me that they can think through problems, organize a research plan, and get to a data-driven solution quickly. They are also going to be able to draw connections between data and real business problems.

I’ve interviewed a lot of people for this position and heard a lot of answers. The best answers have a few things in common:

* The metrics they propose are related to the problem. For example, if I want to increase traffic to a website, I want the metrics to directly relate to page visits. If I want to reduce unnecessary medical expenses, the metrics should directly relate to cost and usage.
* They suggest a few metrics and focus on how they will leverage those, paying attention to how they will use the metrics rather than just giving a list of things to measure.
* They state their assumptions at the beginning and ask clarifying questions as needed.

In the very best interviews, I walk away with a new way of thinking about my data and how to design new experiments.

Let’s go ahead and get started!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Conceptual Questions**

Free response

What is lifetime value and how can you measure it?

**Get feedback**

Powered by OpenAI

Free response

We’ve identified reducing churn by 10% as a goal this year. What would you do to create a model that can identify who is susceptible to churn?

**Get feedback**

Powered by OpenAI

Excellent! You seem like you have a good idea about designing experiments, let’s move on to applying these concepts.

### **Applied Questions**

Free response

e-Commerce website: We have a search feature on our site that we have had for about 5 years. We are trying to decide if we need to redesign it. How would you measure the effectiveness of this search feature?

**Get feedback**

Powered by OpenAI

A pair of yellow headphones is plugged into a computer application window. Inside the application window, there is a box with what is likely text. Next to the box there is a flow chart represented by shapes stacked vertically, connected by lines. The first level in the flow chart is a purple rectangle, the second is a green rhombus, and the third is two yellow squares.

Free response

How would you measure the performance of a song recommendation model for a music streaming platform?

**Get feedback**

Powered by OpenAI

What user behaviors are measurable? What do those behaviors tell you about the performance of the recommendation engine?

Free response

A video streaming service: We recently launched a new feature that allows viewers to create playlists of movies, shows, and other content. This will allow them to pre-plan binge sessions. We have already rolled out this feature in our latest release, so it is not possible to conduct an A/B Test.

How would you evaluate the success of this feature? What data would you need and how would you analyze it?

**Get feedback**

Powered by OpenAI

The key here will be segmenting users after the fact. Since you cannot divide users into control and experiment groups, how could you analyze the data that they have already generated and data that they are likely to generate in the future?

### **Review**

Excellent job!

It’s clear that you have a solid sense of what data we need to measure and how to set up experiments.

You’ve shown your expertise in:

* measuring lifetime value
* predicting churn rates
* normalizing data to isolate effects
* measuring the impact of features after launch, and
* setting benchmarks to measure success.

Be sure to keep getting familiar with the company’s product for the next round of interviews. See you there!

# **A/B Testing**

**Get practice answering A/B testing interview questions**

### **Introduction**

Hi again! It’s me, Codey. I am sure you are just as excited as I am to talk about A/B Testing. Since so much of our job is evaluating options, a deep understanding of A/B Tests is absolutely crucial.

I love when interviewees:

* Start by stating their assumptions and
* Talk through why they collect certain features or run certain tests.

It gives me a much better understanding of how they are going to collaborate with me to think through experiments going forward.

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

There are two images of hamsters. On the left, the hamster has aqua-colored liquid in a water bottle. This hamster is running on its wheel. On the right, the hamster has yellow-colored liquid in its water bottle. This hamster appears to be asleep and maybe dizzy.

Free response

What is the goal of A/B Testing, and why is it important?

**Get feedback**

Powered by OpenAI

Free response

What considerations do you need to make when determining the number of participants needed for an A/B Test?

**Get feedback**

Powered by OpenAI

Free response

If you had to stop an A/B Test early, what considerations would you make?

**Submit Response**

Fantastic! It really sounds like you understand what goes into A/B Testing, now let’s go over some applied questions to talk about how this plays out in real life.

### **Applied Questions**

Free response

Social networking app: We are considering adding a saved comments feature on our platform. When users are messaging each other, they can flag comments to be saved to a special folder. How would you determine if we should implement this new feature?

**Get feedback**

Powered by OpenAI

Free response

A news site: We have new versions of four features that we want to evaluate:

* The header,
* Signup page,
* Newsfeed layout, and
* Home button.

How would you prioritize these features for your A/B Test?

**Get feedback**

Powered by OpenAI

Free response

So we decided to run four variations of one feature at the same time, and we found that one variation won with a p-value of less than .05! Should we make the change?

**Get feedback**

Powered by OpenAI

Free response

Food delivery service: How do you manage spillover between your control and treatment groups? For example, we are launching a new promotion for food delivery and want to know which promotion is most effective at getting people to share their promo codes.

**Get feedback**

Powered by OpenAI

### **Review**

Thank you so much! It’s clear to me that you have a great sense of how to apply A/B Testing to real-world scenarios.

Just to recap, we’ve talked about:

* The goals of A/B Testing.
* How to determine how long an A/B Test should be.
* What to think about if you have to stop an experiment early.
* Working with multiple feature changes and multiple variations
* Evaluating the potential success of a feature.
* Preventing and managing spillover.

It was great talking to you and I look forward to chatting with you in the rest of the interviews! See you there.

# **A/B Testing Take Home**

**Practice your A/B Testing skills with a Take-Home Assessment**

### **Overview**

Welcome to the take-home assessment! Take-home assessments help interviewers get a sense of your technical skills before moving you to the next stage of the interview process. The style and format of these projects can vary widely depending on the company and role you are applying for. In this project, we are going to give you a prompt and you are free to answer it however you wish. You can use any coding language and any tools you are familiar with. This is an opportunity to show off your skills.

The projects that we have prompted here are typical for data analyst roles but are not definitive. However, there are some widespread practices and conventions that you should be aware of before starting the project.

* There may be one or more very easy questions. These are not tricks, they are meant to easily exclude applicants who cannot perform basic functions.
* There will be some discussion, analysis, or reflection. Expect to have to write a summary of your findings. A large part of both Data Analyst and Data Scientist roles is communication. How well you communicate your findings is just as important as what you find and how you find it.
* You may be asked to submit your files in a specific way. Pay attention to these instructions and submit your file in the requested format. This is another area that may initially seem unimportant but can have an outsized influence on your candidacy.
* You may *not* be asked to submit your files in any particular way. If this is the case, Python files should be .py or .ipynb, R should be in .R or .Rmd format. Similarly, your written analysis should be in .pdf format or .md if submitted online.
* If anything is unclear or you are not sure what they are asking you to do, ask! It is much better to get clarification at the beginning of a project than submit something that doesn’t answer the question. This assignment is an opportunity to showcase your problem-solving skills. If you are stuck on a particular technical skill, try to find the answer on your own. Google and StackOverflow will be your friends during this process! If you still can’t figure it out, and the interviewer has allowed for questions, reach out and explain what you’ve already tried. This will show you have the ability to problem-solve on your own before reaching out for help.
* Your interviewers may give you an estimate of how long they think it should take. It is very common to go far beyond the time estimate.
* The take-home assessment is also a good opportunity for you to decide if this is the kind of work you are interested in doing. If you enjoy it, you are probably in the right place. If it is something you do not enjoy, it may be worth clarifying what the expectations of the job are.

### **Context:**

You are applying for a role that might be situated in marketing, web development, or a business analytics department of a large company. The dataset provided is for a library homepage, but the principles are the same across institutions and industries.

### **Take Home Challenge:**

Congratulations on being invited to complete this analysis project for our Analytics team!

This assignment consists of two parts.

* In Part 1 you will read, interpret, and explain the dataset provided.
* In Part 2, you will demonstrate your data analysis skills using the dataset.

You can use the tool(s) of your choice. The key to this assignment is to clearly document your steps and state any assumptions you’ve made along the way. If you choose to use a programming language, you’ll need to provide your code. If you choose to use a spreadsheet, you’ll need to clearly document your steps. This way, our team will be able to easily follow along with your logic and fully understand your approach.

When you turn in your work we expect to receive:

* Documentation of your steps (either as a code file or process notes).
* A written summary of your findings as a .pdf or Microsoft Office file.
* Any supplemental visualizations or materials that are necessary to illustrate your findings.
* All of your files combined as one zip file.

Best of luck, we look forward to reading your results.

**Necessary Files:** The five site variations were given sequential names:

| **Experiment Name** | **Category Title** | **Url** |
| --- | --- | --- |
| Control | Interact | /index.php |
| Variation 1 | Connect | /index2.php |
| Variation 2 | Learn | /index3.php |
| Variation 3 | Help | /index4.php |
| Variation 4 | Services | /index5.php |

* Google Analytics Data for this site is linked [here](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/general-data/GoogleAnalytics-2014-07-11.zip).
* Page Views and clickthrough data collected by CrazyEgg is linked [here](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/general-data/CrazyEgg-2014-07-11.zip).
* Details about this dataset are available [here](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/general-data/AB-testing-web-analytics.pdf).

You may need to use one or both data files. This data was obtained from [Young, Scott W.H. (2014)](http://dx.doi.org/10.3998/weave.12535642.0001.101) A/B Testing Web Analytics Data. Montana State University Scholar Works. [dataset](https://scholarworks.montana.edu/xmlui/handle/1/3507)

For more on how CrazyEgg works, [see here](https://www.crazyegg.com/overview)

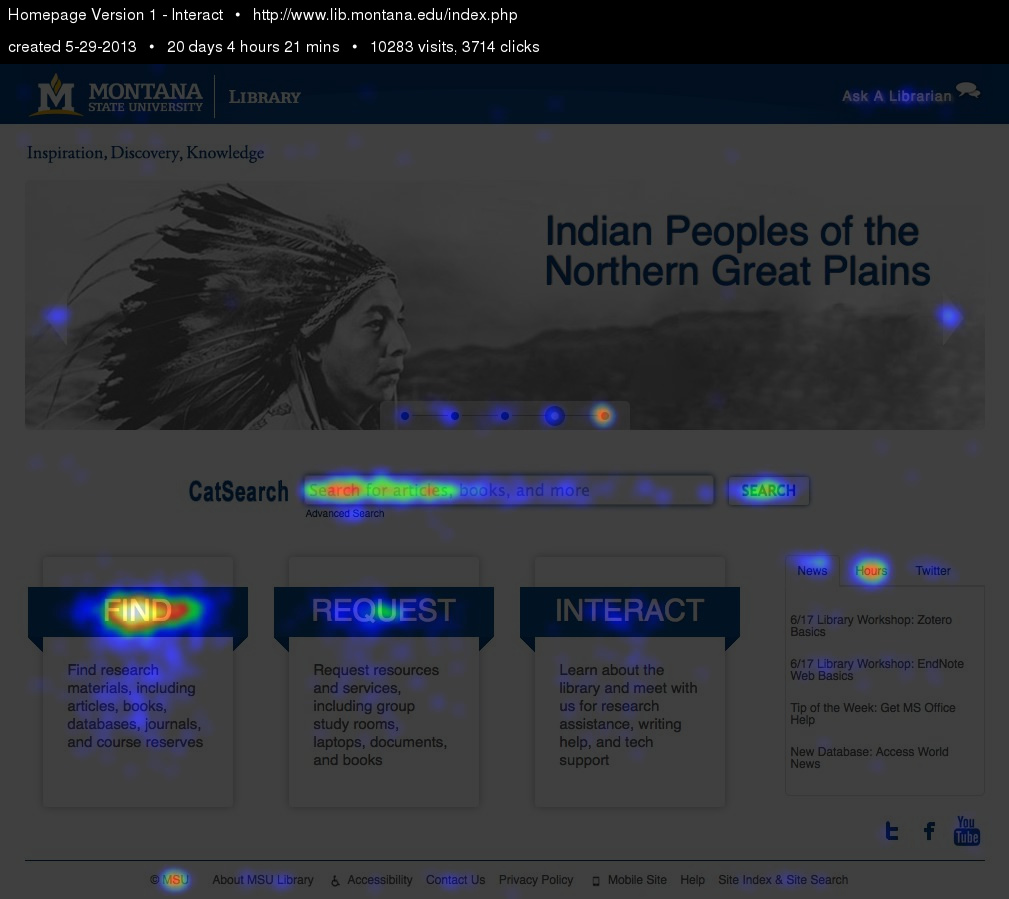
### **Instructions**

#### **Part 1: Exploratory Data Analysis**

* What is the median average time on a page?
* What is the average total daily page views?
* Are there any outliers?

#### **Part 2: Analysis of Testing Results**

We’ve provided a sample, open-source, dataset from a library website that is similar to the data we typically work with. Our goal is to have users follow through further into our website. In this example, we noticed that our current “Interact” page is often neglected by visitors, as you can see from this heatmap.



We want to know if changing the design of the webpage will engage them better. Our question is: “Which title will generate the most engagement: Help, Services, Learn, Connect, or Interact”

**Your task:**

* Explore the dataset.
* Create a visual representation of the click-through rate, bounce rate, and exit rate for all five variations of the A/B test.
* Write a short paragraph explaining why you chose the visualizations you did, how you arrived at your metrics, and be sure to include what variation you would recommend to our team based on your findings.

### **Checklist:**

Here is a checklist of things to consider as you work on this project:

* Make sure your code is clean, well-documented, and that your steps are clearly laid out.
* Assessment questions in Part 1 only have one correct answer. Make sure to double-check your math and write 1-2 sentences interpreting the values.
* There is an academic paper associated with this dataset asking the same questions in Part 2. Your results may be similar. The process you take to get your answer and the way you communicate your results are most important in a take-home challenge.
* Make sure to provide a clear, data-driven recommendation.
* In your presentation, you can include the heatmaps and screenshots as supporting evidence. If it is in the data files, you can use it in your analysis.

**Note**: a take-home won’t typically include a checklist like this one, but we’ve included one here for you as you become more familiar with the take-home.

Good luck and have fun!

# **Review of Visualization, Communication, and Experimental Design Interviews**

**See all you’ve covered in Visualization, Communication, and Experimental Design**

### **Review**

Congratulations! The goal of this unit was to get practice answering questions about data visualization, communication, and experimental design.

Having completed this unit, you are now able to:

* Discuss how you pick a data visualization tool.
* Use data visualization to answer a practical business question.
* Describe your approach to experimental design.
* Talk your interviewer through how you conduct A/B tests.
* Evaluate data from a real A/B test to make a recommendation about website layout.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# **Introduction to Data Analyst Interview Coding Questions**

**See what coding questions you’ll cover.**

### **Goals of this Unit**

The goal of this unit is to prepare for coding questions in your interview. Since most data analyst roles require knowledge of SQL and familiarity with a scripting language (such as Python or R), this unit focuses on only what you will need for the interview. The focus is on SQL and general whiteboarding. Whiteboarding is a practice that is applicable to a wide variety of coding challenges.

After this unit, you will be able to:

* Approach whiteboarding and live coding challenges in an organized and systematic way.
* Communicate your coding decisions to your interviewers.
* Implement SQL to answer common data analyst interview questions.
* Describe your thought process when performing database queries.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# **Technical Data Science and Data Analyst Interviews: Whiteboarding**

# **Introduction**

1 min

* Aspiring data analysts and data scientists must pass a technical interview to be hired as a professional. These interviews evaluate the candidate’s ability to write code, articulate their thought process, and problem solve in a timed environment.
* Technical interviews take many forms: writing code alongside an employee of the company, testing online with a third-party service, take-home challenges spanning several days, and more!
* A *whiteboarding* technical interview requires the interviewee to plan and code the solution by hand. The candidate relies on their ability to communicate without the assistance of a text editor or web browser.
* This lesson covers strategies to improve performance in a whiteboarding interview. Each exercise features a step in the interview with a video demonstration of the technique.
* The steps are:
* **Clarify** the Problem
* Create **Inputs**
* **Outline** the Solution
* **Code** the Solution
* **Test** the Solution
* **Analyze** the Solution
* **Instructions**
* Whiteboarding does not use a text editor, so this lesson will ask you to follow along with a pen and paper.  
  Grab a notebook and something to write with before moving forward!  
  **Note:** The instructor in this lesson’s videos will be using Python, but these whiteboarding tips and tricks can be applied to any other general-purpose programming language.

### **Clarifying the Problem**

3 min

Whiteboarding interviews begin with a problem from the interviewer. The interviewee must be confident they understand the dimensions of the problem!

Working with data is full of ambiguity. Programming requires concrete deliverables, but company needs can be murky. Even when the need is clear, a feature could have dozens of possible implementations. The ability to clearly define a problem is an important skill to demonstrate.

When the interviewer presents their technical question, repeat the question back to the interviewer **in your own words**. This gives you a moment to think and will resolve any glaring misunderstandings.

Once you’ve repeated the question, **ask every clarifying question that comes to mind.**

Assumptions must be communicated to the interviewer so there is agreement on the scope of the problem.

For example, if asked:

*Write a function that returns duplicate characters in string.*

Here are some questions which may come to mind:

* What is the desired return value?
  + True|False, a list of characters, or …?
* Do punctuation and spaces count as “characters”?
* Should case be considered?
  + are "a" and "A" duplicates?
* Should we be checking for [Unicode](https://en.wikipedia.org/wiki/Unicode) characters?
* Can we assume it’s a 26 character alphabet?

**Instructions**

1. We’ll apply these steps to a single problem through the rest of the lesson.
2. Given a list of numbers, return whether the list contains Pythagorean Triplets.
3. Copy to Clipboard
4. Rephrase this problem in your own words and write that down.
5. Then, write down every question or assumption you have.
6. Watch the video to see how we clarified this question.

### **Producing Inputs and Finding Edge Cases**

3 min

When the question is clear, we then produce concrete inputs and outputs. These inputs guide a solution for the remainder of the interview so write them on the board!

You may still be unclear how to solve the problem in code, but it’s certain that given an input, X, your function will produce an output, Y.

Make one input the *happy path*: input that reflects a common scenario.

For example, you’re asked to write a function which capitalizes the first letter of an input

[string](https://www.codecademy.com/resources/docs/general/data-types/string)

Preview: Docs Stores a sequence of indexed characters that can be of any length and is contained within a pair of single or double quotes.

.

A good input could be "apple", which returns "Apple" because this demonstrates the function’s purpose.

If the input were "Apple", it would return "Apple". That’s correct but less informative.

Also think about *edge cases*, or inputs which **do not** reflect a common scenario and may cause problems.

For the capitalization function, what should you return if given None as input? Or a number?

**Instructions**

Write inputs and outputs for the Pythagorean Triplet problem.  
Try to give yourself a few different cases.  
Watch the video to see how we made inputs/outputs.

### **Writing the Outline**

3 min

It’s time to start breaking down the problem by category.

Given a question which requires the use of a stack, what do you know about stacks? Have you encountered other problems that use stacks and how were they solved?

Is this a searching question? Can you sort the input and will that help? Does this problem sound like it can be modeled as a graph, with vertices and connected edges?

Understanding the applications of different

[data structures](https://www.codecademy.com/resources/docs/general/data-structures)

Preview: Docs Loading link description

is very useful! The more questions you practice, the more you will be able to see patterns between problems.

This step varies the most because it requires details of the specific problem, but regardless of the question make certain you are communicating with the interviewer as a potential co-worker. **Show them your thought process!**

During this step, the interviewer may make suggestions on how to proceed. Acknowledge the interviewer and incorporate their suggestions into your approach.

**Don’t disregard their input! It gives the impression you would be difficult to work with on the job.**

When you and the interviewer are satisfied with a workable solution, write the steps next to your input. Follow these steps as you write code on the board.

**Instructions**

Come up with an outline of how to solve the Pythagorean Triplet problem.  
Don’t worry about efficiency, just aim for high-level steps which will produce the correct output.  
Watch the video when you’re ready to move on.

### **Coding the Solution**

3 min

We’ve finally come to the portion where you’ll write some code! Previous steps improve your odds for success because you have an outline and clear inputs for testing.

Writing code on the board is a collaborative process. Refer to your outline and explain the step you’re implementing.

The goal is to be facing the interviewer when talking through the implementation and facing the board when you’re writing the code.

Try to avoid writing code in silence or narrating at a low level like “for… i… in… range… length of the list… colon” when writing for i in range(len(input\_list)):.

When you’re finished with the implementation, look it over for any mistaken syntax or logical errors.

**Instructions**

Code the solution to the Pythagorean Triplets problem by writing it out on paper. Use your outline from the previous exercise as a guide.  
Before writing each step on paper, **say what you will do out loud.** Better to practice by yourself than in a real interview!

### **Testing with Inputs**

4 min

With a working implementation on the whiteboard, use the test inputs to walk through the evaluation of your code.

Write out any temporary

[variables](https://www.codecademy.com/resources/docs/general/julia/variables)

Preview: Docs Loading link description

on the board and update them when they change during execution.

This is another opportunity to showcase your communication skills, and will help you catch any logical errors which you may not have noticed in earlier steps.

If you catch an error, don’t panic! Mistakes happen. Explain the issue and talk through what you can do to fix the bug.

Your interviewer may want to see you write the correction or they may be satisfied with the explanation.

**Instructions**

Use your test inputs and implementation to talk through your solution.  
Again, **say each line aloud** so you become more comfortable speaking about the code you’ve written.  
Track the variables used and update them as you go.

### **Analyzing Time and Space Complexity**

3 min

You’re satisfied with your implementation and you’ve demonstrated how it works, but you’re not quite done.

Analyze the time and space complexity of the solution. With this step you are demonstrating that you care about the efficiency of your code. Even though it is not essential for typical Data Science and Data Analyst roles, considering how long your code will take to run will make your code better.

If you can optimize to a more efficient

[runtime](https://www.codecademy.com/resources/docs/general/runtime)

Preview: Docs Loading link description

, explain how that would work. If you can’t optimize, explain why it’s not possible.

**Instructions**

Talk through the time and space complexity of your solution.  
Explain how an optimization would be possible or not.

### **Review**

<1 min

Technical interviews judge the candidate’s ability to perform technical aspects of the job. Whiteboarding is a type of technical interview which requires the candidate to code a solution without the assistance of a text editor.

Coding an efficient solution is important, but the ultimate goal as an interviewee is to convince the interviewer you would be a desirable co-worker.

A desirable co-worker is able to articulate their thought process and collaborate without friction.

Our whiteboarding strategies revolve around communication at every step.

The steps are:

* **Clarify** the Problem
* Create **Inputs**
* **Outline** the Solution
* **Code** the Solution
* **Test** the Solution
* **Analyze** the Solution

**Instructions**

Practice makes perfect!  
Run through these steps with a new practice problem:  
Write a function which prints every number from 0 up to the given input. If divisible by 3, print “Fizz” instead of the number. If divisible by 5, print “Buzz”. If input is divisible by 3 AND 5, print “FizzBuzz”.

# **Common SQL Interview Questions**

**Practice responding to common SQL interview questions.**

Welcome back!

In this part of the interview, we are going to cover some common questions that come up around databases–specifically SQL. Though some positions don’t require SQL knowledge and others require extensive knowledge, we expect our team members to have a solid understanding of SQL even though they don’t have to be experts.

These questions are a bit more direct than the other sections because we assume that if you know these basic concepts, you have the SQL skills we are looking for.

So, without further ado, let’s jump in to talking about SQL and all the powerful things it can do!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Questions**

Free response

What is SQL?

**Get feedback**

Powered by OpenAI

Free response

What is a database? And what is a relational database?

**Get feedback**

Powered by OpenAI

Free response

What is a RDBMS?

**Get feedback**

Powered by OpenAI

Free response

What is a table? And what are a row and column within a table?

**Get feedback**

Powered by OpenAI

Free response

What is a data type?

**Get feedback**

Powered by OpenAI

Free response

What is a primary key and a foreign key?

**Get feedback**

Powered by OpenAI

Free response

What is the difference between ALTER and UPDATE?

**Get feedback**

Powered by OpenAI

Free response

What is a query and what is a subquery?

**Get feedback**

Powered by OpenAI

Free response

What are constraints?

**Get feedback**

Powered by OpenAI

Free response

What is a statement?

**Get feedback**

Powered by OpenAI

Free response

How do you check if a field has a value or not?

**Get feedback**

Powered by OpenAI

Free response

What is the difference between DISTINCT and UNIQUE?

**Get feedback**

Powered by OpenAI

Free response

What is a join?

What is the difference between an INNER JOIN and LEFT JOIN?

**Get feedback**

Powered by OpenAI

Free response

What is the purpose of windows functions?

**Get feedback**

Powered by OpenAI

Free response

What are indexes and why are they needed?

**Get feedback**

Powered by OpenAI

Great job! It is clear that you know your way around a database and are ready to talk about them in an interview.

### **Review**

In this interview, we covered basic SQL concepts and practices, including:

* The basic structure of a database.
* How data are represented in a database.
* Some simple functions for extracting data from a database.
* Best practices for keeping a database well structured.

Looking forward to seeing you again in the next interview!

### **Code Challenge 14**

2 min

The purchases table has the following columns:

* purchase\_id - the id of the purchase
* purchase\_date - the date of the purchase

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Using string formatting and substitutions, get the month and day for each purchase in the form ‘mm-dd’.  
  Give this new column a name of ‘reformatted’.
* SELECT STRFTIME('%m-%d', purchase\_date) AS 'reformatted'
* FROM purchases;

SELECT STRFTIME('%m-%d', purchase\_date) AS 'reformatted'

FROM purchases;

# **Review of Data Analyst Interview Coding Questions**

**See what you’ve covered in Coding Questions.**

### **Review**

Congratulations! The goal of this unit was to prepare you for coding questions in your interview, with a focus on SQL.

Having completed this unit, you are now able to:

* Approach whiteboarding and live coding challenges in an organized and systematic way.
* Communicate your coding decisions to your interviewers.
* Implement SQL to answer common data analyst interview questions.
* Describe your thought process when performing database queries.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# **What is a Portfolio Project?**

This project is a little different from other Codecademy projects you’ve encountered. In this project, you will bring together what you have learned in previous lessons to build a project in your own development environment and publish it to the web!

Create your own files outside of the Codecademy platform

Write your own code with less guided instructions

Use common project management processes to track your progress

# **How Do Portfolio Projects Work?**

We’ll provide you with high-level tasks to guide your project to completion, but you will be responsible for deciding how to implement them in your code.

There are many possible ways to correctly fulfill all of these requirements, and you should expect to use the internet, Codecademy, and other resources when you encounter a problem that you cannot easily solve.

Note that there are hints that can assist you, but they will only provide one potential implementation. Do not worry if your program looks different from ours!

# **Project Prompt**

## **Data Analyst Interview Preparation Portfolio-Building Project**

Your portfolio is the key to getting a data analyst interview because it lets you show off all of your awesome data analysis skills. This project will give you a chance to bring together everything you have worked on and fill in any gaps with new projects.

Click Begin to get started!

### **Project Overview**

This project is not one project, but many different sub-projects. Depending on how much of the Data Analyst Career Path you have completed, or how many other projects you have done, this project may be primarily about organization and refinement. If you don’t have any completed projects, this project may be the beginning of your journey to developing a portfolio.

Regardless of where you are starting, by the end of this portfolio-building project, you will have a job search-ready portfolio that shows off your skills.

We will not be providing step-by-step instructions, but we will provide a checklist of elements your portfolio should have and suggestions on how to develop those elements.

### **Project Objectives:**

* Collect and organize completed projects
* Create new projects to fill in any skills gaps
* Align your portfolio to your career goals

### **Prerequisites:**

* Data Analyst Career Path

## **Project Tasks**

Keep track of your progress by dragging each task from "To Do" to "In Progress" to "Done" as you work on them. You can also click on a task to see more information about it.

**Establish an Online Presence**

Create a website or GitHub repository with the following characteristics:

* Your name, initials, or professional handle
* A short 1-4 sentence professionally-oriented biography
* Links to your projects
* Title each project with a descriptive name (i.e., the title of the project)

Specific for GitHub:

* Each project has its own repository
* Pin repositories for the projects that you want to showcase
* Use the first sentence of each project’s **README.md** for a short description of your project

Specific for a Website:

* An organization system for projects (i.e., a menu or gallery)

**Tidy Your Code**

Good code is well documented and organized. Most people have to clean up their code to make it intelligible to others. While this step is not essential, it does demonstrate that you can work in an organized way.

* Write comments for any functions you created.
* Write comments for complicated or clever steps.
* Break functions into their own files and call those files in your main script.
* Separate phases of a project into separate files. For example, clean in one file, analyze in another.

**Regression Project**

This project should involve determining if there is a linear association between two variables. The variables should both be numeric, and the predictor and dependent variables should be clearly indicated.

If you completed the Data Analyst Career Path, Reggie’s Linear Regression is a great project to demonstrate this skill.

Be sure to include an analysis of your findings!

**Data Aquisition Project**

Data acquisition is not necessary for all jobs, but can showcase your versatility and adaptability. You could choose to do web scraping with Beautiful Soup or work with a popular API such as the Dataverse API. This project does not need to stand on its own, getting data for another project will demonstrate this skill.

The Chocolate Scraping with Beautiful Soup Project from the Data Analyst Career Path is a good example of a web scraping project, though to include it in your portfolio you would have to recreate it off-platform and use a different website. Remember if you are webscraping, be a good internet citizen and do not bombard someone’s site with more requests than necessary.

For working with an API, the Dataverse API is a great repository of datasets that can be accessed via an API.

**Data Cleaning Project**

At least one project should work with a dataset from an original source. Original sources are typically either data that you have curated yourself or data that is distributed by someone before it is cleaned.

Be sure to document your cleaning steps and why you performed each step. This can be comments in your code, notes in a Jupyter Notebook, or an affiliated document.

You can find original sources in a variety of repositories. Some popular ones are:

* FiveThirtyEight
* Buzzfeed
* UCI
* Data is Plural
* Google Data Repository

**Hypothesis Testing**

At least one project should involve testing a hypothesis. Regardless if you can or cannot reject the null hypothesis, the purpose is to demonstrate that you know how to set up a hypothesis test and evaluate the results.

Your hypothesis test can take the form of an A/B Test or it can be a more traditionally academic type. The A/B Testing Take-Home Project from this Skill Path can fulfill this requirement or you can recreate any of the Hypothesis Testing Challenge Projects in the Data Analyst Career Path.

Be sure to summarize your findings in a written document. This can be a text file, presentation, or other format, but the summary and interpretation is as important as the analysis.

**Data Visualization Project**

Many of your projects will likely involve some sort of visualization. If any project leverages a lot of visualizations, it can fulfill this task. The Data Visualization portfolio project from this Skill Path or the portfolio project in the Data Analyst Career Path (Life Expectancy and GDP) is also a good option.

**Role-specific Data Project**

Consider the industry and specific role you are interested in. If you are applying to a marketing agency, a healthcare company, or conducting financial analysis, you will want to have something in your portfolio that showcases your ability to work with that kind of data in particular.

In Progress

# **Review of Data Analyst Interview Preparation**

**You have the data analysis skills, and now you have the interview skills to match!**

### **Review**

Congratulations!

The goal of the Data Analyst Interview Preparation Skill Path was to give you the opportunity to practice answering the types of questions you will see in a data analyst interview. In this Skill Path, you learned how to decode job postings, tailor your materials to a specific job, and answer almost any interview question. After completing this Path, you have familiarity with data analyst interview question types and the key elements that make a good response.

Having completed this unit, you are now able to:

* Specify what kind of role you want to apply for and tailor your job search materials and interview preparation to that specific job description.
* Anticipate the stages of a Data Analyst Interview from the initial phone screen to the on-site interview.
* Explain your data visualization and communication process.
* Respond to general data management questions—from cleaning to exploratory analysis.
* Describe your approach to experimental design and project management.
* Ace the Python and SQL technical challenges.
* Effectively communicate your solution to coding challenges.
* Successfully complete the data analyst interview take-home challenge.
* Present yourself professionally as a Data Analyst.

You can walk into your next data analyst job interview with confidence!

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

Syllabus

**1. Welcome to the Data Scientist Interview Preparation Skill Path**

Welcome to the Data Scientist Interview Preparation Skill Path

1 Informational

**2.Getting Started with Data Science Interview Prep**

Welcome to Getting Started with Data Science Interview Preparation

1 Informational

Getting Started with Data Science Interview Preparation

1 Lesson, 1 Article

Review of Getting Started with Data Scientist Interview Preparation

1 Informational

**3. General Data Questions for Data Scientists**

Welcome to General Data Questions

1 Informational

50% progress

50%

Data Cleaning and Types for Data Science Interviews

2 Articles

Classic Statistics for Data Science Interviews

Data Cleaning & Summarizing Data Science Project

1 Article

Review of General Data Questions

1 Informational

**4. Visualization, Communication, and Experimental Design**

Welcome to Visualization, Communication, and Experimental Design

Visualization and Communication

Experimental Design

Review of Visualization, Communication, and Experimental Design

5. Coding Challenges for Data Scientist Interview Preparation

Welcome to Coding Questions

Python Technical Interview: Whiteboarding

SQL for Data Science Interview Preparation

Review of Coding Questions

6. Machine Learning and Algorithms for Data Science Interviews

Welcome to Machine Learning and Algorithms

1 Informational

General Machine Learning Interview Questions

1 Article

Supervised Learning Interview Questions

4 Articles, 2 Lessons

Unsupervised Learning Interview Question

2 Articles, 1 Lesson

Model Tuning Interview Questions

3 Articles, 1 Lesson

Review of Machine Learning and Algorithms Interview Questions

1 Informationa

7. Data Scientist Interview Preparation Portfolio Project

Data Scientist Interview Preparation Portfolio Project

1 Kanban project

8. Review of the Data Scientist Interview Preparation Skill Path

Review of the Data Scientist Interview Preparation Skill Path

Review all that you’ve learned in the Data Scientist Interview Prep Skill Path.

Informational

Review of the Data Scientist Interview Preparation Skill Path

Company

About

Careers

Affiliates

Partnerships

Resources

Articles

Blog

Cheatsheets

Code challenges

Docs

Projects

Videos

Workspaces

Support

Help Center

Plans

For individuals

# 

# 30 - Data Scientist Interview Preparation

# **Welcome to the Data Scientist Interview Preparation Skill Path**

**You have the data science skills, now it’s time to ace the data scientist interview!**

### **Goals of this Skill Path**

The Data Scientist Interview Preparation Skill Path is designed to give you the opportunity to practice answering the types of questions you will see in a data scientist interview. In this Skill Path, you’ll learn how to decode job postings, tailor your materials to a specific job, and answer almost any type of interview question. By the end of this Path, you will have familiarity with data scientist interview question types and the key elements that make a good response, so that you’ll be confident and prepared for even the toughest interviews.

After this Path, you will be able to:

* Specify what kind of role you want to apply for and tailor your job search materials and interview preparation to that specific job description.
* Anticipate the stages of a Data Scientist Interview process, from building your resume to the on-site interview.
* Respond to general data management questions—from cleaning to exploratory analysis.
* Explain your data visualization and communication process.
* Describe your approach to experimental design and project management.
* Ace the SQL technical challenge.
* Effectively communicate your solution to coding challenges.
* Successfully complete data scientist interview take-home challenges.
* Give detailed answers to machine learning interview questions.
* Implement and explain your code for building machine learning models.
* Present yourself professionally as a Data Scientist.

Along the way, you’ll be able to put your knowledge of data analytics into practice with Portfolio Projects where you can showcase your skills and interests.

### **Navigating this Skill Path**

This Skill Path covers the most common topics across a variety of roles within data science. Not everything will apply to every situation. Feel free to pick and choose what content is relevant to you and the specific role you are applying for. Note that this path provides a general introduction, and some roles will require more preparation either in Python or SQL, and depending on the job you are applying for, you may want to supplement with a “Pass the [Language] Technical Interview” Skill Path. But you don’t have to figure that out alone, this path will help guide you to make that decision.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned

# **Welcome to Getting Started with Data Scientist Interview Preparation**

**Get started with an overview of data scientist and data analyst job and interview types.**

### **Goals of this Unit**

The goal of this unit is to introduce you to the different types of data scientist roles and the corresponding job descriptions.

After this unit, you will be able to:

* Read a job description and better understand what skills the company is looking for.
* Tailor your job search materials to meet the requirements stated in a job description.
* Determine what type of role(s) you would be most interested in.
* Prepare for the stages of a data scientist job interview.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# Overview of the Data Science Interview Process

# 

# Introduction

<1 min

Approaching the data science interview process for the first time can be daunting. Phone screenings, take-home assessments, whiteboard interviews – does it ever end?

In this lesson, we will demystify the data science interview process so that you know exactly what to expect going in. You will learn about:

* Polishing your resume
* Building a digital presence
* Creating a portfolio
* The phone screening
* The take-home assessment
* Technical and whiteboard interviews
* Basic interview etiquette

After this lesson, the rest of this skill path will be focused mainly on portfolio-building, the technical interviews, and take-home assessments.

Let’s get started!

**Instructions**

In this lesson, you will get an introduction to the Data Science Interview Process.  
When you’re ready, move on to the next exercise.

### **Getting Your Foot in the Door**

1 min

The first step of the data science interview process is getting your foot in the door.

To set yourself up for success, you will need to polish your resume and build a digital presence.

##### **Resume**

You might have a deep understanding of data science knowledge, but that alone will not help you get your foot in the door. To stand out from the crowd and secure an interview, you will need a strong resume that showcases your skills and experience.

**First impressions matter**

Make sure your resume is free of any spelling or grammar mistakes. Also, make sure it is consistent in style and content. These simple things make a big difference.

**Tailor to the role**

Of course there is no one-size-fits-all resume in the field of data science. Carefully read the description for the role you are interested in and make sure your resume reflects the required technical skills. Do this each time you apply for a new role.

**Be concise**

Keep your resume concise by eliminating any skills or points that aren’t relevant to the role. Every word you choose to put on your resume matters, so choose carefully.

**Tell a story**

A good resume doesn’t just list everything you’ve ever done. It should be a story about growth in a particular direction and culminate with the contributions you will make to the new role.

Try to read your resume in the voice of the hiring manager. How will you help their organization? Why you are the right person to fill their needs?

[Here](https://www.indeed.com/career-advice/resume-samples/information-technology-resumes/data-scientist) is a helpful guide on how to craft a strong data science resume.

##### **LinkedIn**

Recruiters often check your [LinkedIn](https://www.linkedin.com/) profile before inviting you for an interview.

Make sure to upload a professional-looking photo of yourself and summarize your role, skills, and experience in the “About” section.

When listing out your previous roles in the “Experience” section, keep your descriptions short and to the point but give a clear breakdown of your role responsibilities and concrete skills and activities.

The “Skills and Endorsements” section is another great way to showcase your skills. Typical data scientist skills include:

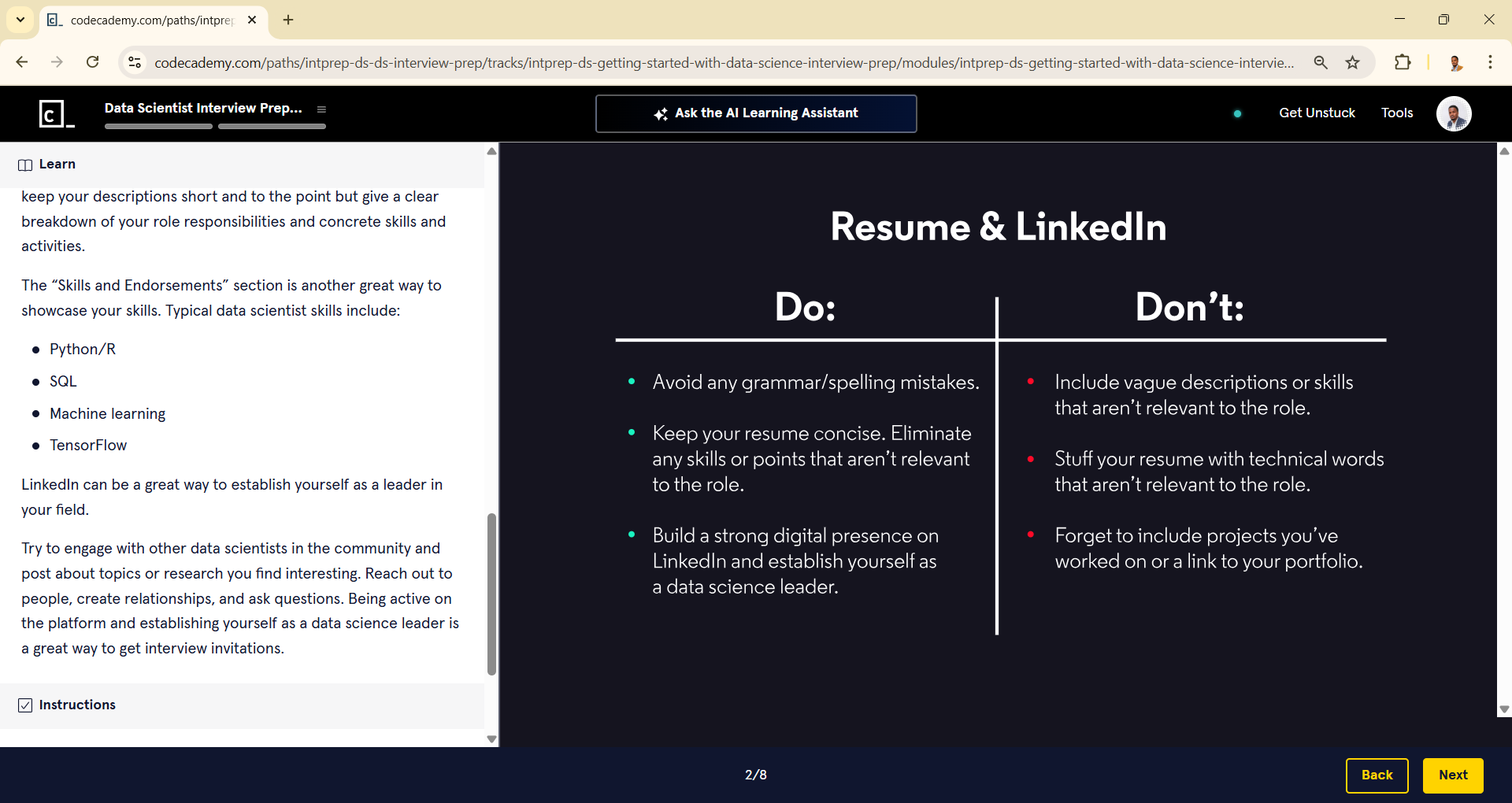
* Python/R
* SQL
* Machine learning
* TensorFlow

LinkedIn can be a great way to establish yourself as a leader in your field.

Try to engage with other data scientists in the community and post about topics or research you find interesting. Reach out to people, create relationships, and ask questions. Being active on the platform and establishing yourself as a data science leader is a great way to get interview invitations.

**Instructions**

Take a look at the chart on the right to review best practices and what to avoid on your resume and LinkedIn profile.



### **Showing Off What You've Built**

12 min

An important part of the data science interview process is showcasing the projects that you’ve worked on.

This skill path provides many opportunities to build a strong portfolio that shows off skills companies want to see.

##### **GitHub**

When it comes to showcasing your portfolio, GitHub allows you to share your projects publicly, including collaborations with other data scientists.

Recruiters and hiring managers may turn to GitHub to find talent. GitHub allows them to:

* See how active you are in the data science community.
* Dig into the source code itself to see your skills and how you approach a problem.

Make sure that every project you include in your portfolio has a reason for being there and follows best practices:

* Is the code clean and well documented?
* Did you include a README file that explains the basic functionality and how to run the code?
* Do your portfolio projects show growth over time?

GitHub can also be a great way to establish yourself as an expert in a specific domain. For example, if your goal is to work as a data scientist at a finance company, then you should showcase projects using financial data.

You can also find open-source projects from your industry or a specific company and contribute to the project. When companies see you are an active contributor, they will often reach out to you instead of the other way around!

**Instructions**

* Checkpoint 1 Passed
* **1.**
* You’re getting ready to apply to your dream job but you realize that one of your projects doesn’t follow best practices. Take some time to review the code on the right.  
  In this code, we analyze housing data and visualize the relationship between SalePrice and variables such as GarageArea, YearBuilt, and 1stFlrSF.  
  When you’re ready, continue to the next step.
* This code explores housing data found in the housing\_data.csv file.  
  First, we load in the housing data. Then we perform simple analysis such as:
  + Descriptive statistics
  + Histogram
  + Log transformation
  + Visualizations
* Checkpoint 2 Passed
* **2.**
* Currently, you don’t have any comments in your Python file.  
  Write short but helpful comments for each code block that has # add comment here above it.  
  If you need help, we’ve included a housing\_commented.py file that is well-commented.
* Including comments that are relevant and useful can make it easier for others to understand your code and collaborate with you.  
  Check the housing\_commented.py file to see what a well-commented file looks like.
* Checkpoint 3 Passed
* **3.**
* Your project’s README file is blank!  
  Write a helpful README file in README.md that gives a basic overview of the project.
* Note: this is a relatively simple project so the README doesn’t need to be too extensive. Generally, the more complex the project the more extensive your README should be.  
  Here is a great [article](https://towardsdatascience.com/how-to-write-an-awesome-readme-68bf4be91f8b) on what to include in a README file. (We also recommend including some information concerning where your dataset comes from.)
* Checkpoint 4 Passed
* **4.**
* Nice job! Your project is already much more presentable. Try to find any other ways you can polish your project:
  + Do all of the variable names make sense?
  + Are there more efficient ways to write certain parts of the code?

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# load in housing data

housing\_data = pd.read\_csv('housing\_data.csv')

# check the columns

housing\_data.columns

# descriptive statistics summary

housing\_data['SalePrice'].describe()

# histogram

sns.displot(housing\_data['SalePrice'])

plt.show()

plt.clf()

# perform log transformation on sale price data

log\_data = np.log(housing\_data['SalePrice'])

print(log\_data.skew())

sns.displot(log\_data)

plt.show()

plt.clf()

# visualize relationship between 1st floor sq feet and sale price

data = pd.concat([housing\_data['SalePrice'], housing\_data['1stFlrSF']], axis=1)

data.plot.scatter(x='1stFlrSF', y='SalePrice', ylim=(0,800000))

plt.show()

plt.clf()

# visualize relationship between garage area and sale price

data = pd.concat([housing\_data['SalePrice'], housing\_data['GarageArea']], axis=1)

data.plot.scatter(x='GarageArea', y='SalePrice', ylim=(0,800000))

plt.show()

plt.clf()

# visualize relationship between year built and sale price

data = pd.concat([housing\_data['SalePrice'], housing\_data['YearBuilt']], axis=1)

f, ax = plt.subplots(figsize=(16, 8))

fig = sns.boxplot(x='YearBuilt', y="SalePrice", data=data)

fig.axis(ymin=0, ymax=800000)

plt.xticks(rotation=90)

plt.show()

plt.clf()

### **Passing the Screening**

3 min

You perfected your resume and sent in your application. The company sees a potential fit and decides to invite you to a phone screening. Nice job!

##### **What to expect**

The phone screening is an opportunity for the company to get to know you better and also for you to learn more about the company. The interviewer is trying to get a better understanding of:

* Your professional and technical background
* Projects you’ve worked on
* Why you’re applying to the company

The call usually lasts between 15 to 30 minutes. Depending on the company, the screening may be with a recruiter, hiring manager, or both. If it is with a recruiter, you can expect general questions such as:

* “Tell me about yourself”
* “What is a challenge you recently faced on a project?”
* “Why do you want to work for us?”

If the call is with a hiring manager, you can expect to get a mix of behavioral and light technical questions. Always make sure to look up the interviewer on LinkedIn beforehand to get more context about who you are speaking with.

In addition, checking the company’s common interview questions on Glassdoor is a great way to get a sense of the types of questions you might expect at this early stage in the process.

##### **Nailing the screening**

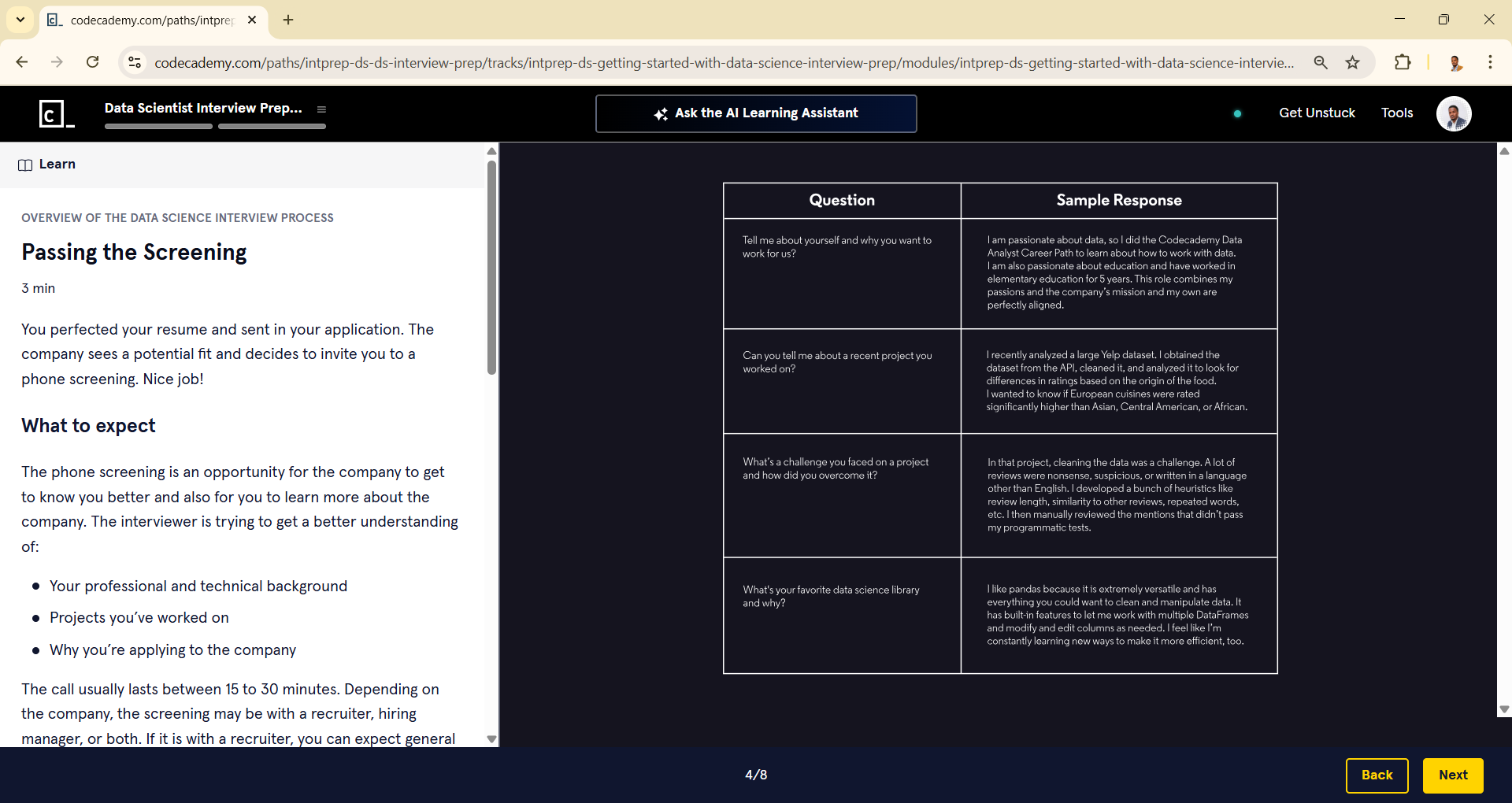
The phone screening is meant to gauge your depth of experience, ability to communicate, and interest in the role.

Be professional and personable. When discussing your experience, don’t just read from your resume – tell a story and connect it back to why you’re applying to this new role.

In addition, *show* the interviewer that you are genuinely interested by asking thoughtful questions and being an attentive listener. Do research prior to the interview and prepare original questions related to the role or company.

**Instructions**

Take a look at the chart to review some common questions you can expect during the phone screening.  
If you’d like to view the image in a separate tab, you can view it [here](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/intro/DataAnalyticsInterview_Excercise4.svg).



### **The Take-Home Assessment**

2 min

Your phone screening went well and the recruiter just sent you a take-home assessment. Take a moment to breathe. You’re well on your way to getting your dream data science job.

##### **What is the Take-Home?**

The take-home is a technical assessment that helps interviewers get a sense of your skills before moving you to the main onsite technical interview.

Here’s what to expect from a take-home:

* a dataset
* instructions for some task(s) or outcome(s)

In this path, you’ll have opportunities to practice take-home assignments.

A benefit of the take-home is that you can show off your technical skills without the pressure of someone watching over your shoulder. The company will usually indicate how much time to spend on the assessment and you should generally stick to those guidelines (plus or minus an hour).

##### **Best Practices**

The take-home allows companies to see how you approach a problem and also whether you follow best coding practices. Here are some best practices to follow:

* Make sure your solution is readable, technically accurate, and well-documented
* Ask yourself whether there is a more efficient or effective solution
* Include proper variable names and comments when appropriate

Before you jump into a solution or coding, make sure you have a clear understanding of the data you’re working with and the specifications of the project. If you ignore this step, things can get sloppy very quickly.

The best way to improve at take-home assessments is to practice doing them. Throughout this skill path, you will get plenty of practice. You can expect to code in Python (or R) and SQL.

In addition, a great way to prepare is to participate in [Kaggle](https://www.kaggle.com/competitions) competitions, which offer take-home style challenges. There you can also see how different data scientists approach a problem and the practices they follow.

**Instructions**

Your take-home assessment might involve interacting with data that the company uses.  
For example, Airbnb might ask you to solve a business problem for them using public housing [data](http://insideairbnb.com/get-the-data.html) in the Seattle area.  
Try to think about the company you’re applying to and how they might interact with data. Look at the chart to see some examples of how companies use data.



### **Technical and Whiteboard Interviews**

6 min

You’ve made it through the screening and passed the take-home assessment. At long last, the company invites you to the final part of the interview process: the onsite interview.

The onsite consists of a mix of behavioral and technical questions, and can include whiteboarding or other technical interviewing techniques. These may include:

* solving technical questions or brain-teasers
* pair programming to solve something
* working with a dataset live on a computer (e.g., Here’s a dataset -> Conduct some EDA -> What model would you use here? -> Build X model -> What insights are you able to find? -> What predictions can you make using this data?)
* walking through a project you built and explaining the technical implementation

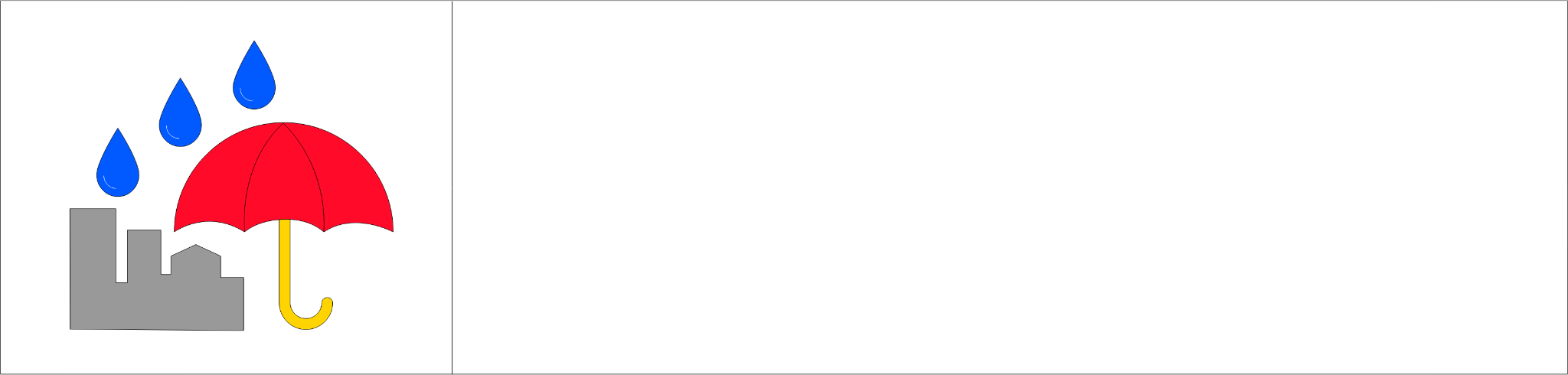
This skill path will help you prepare for the types of technical questions you might encounter.

You can also check Glassdoor to see common behavioral and technical questions for the company you’re applying to and the format of their interview process.

**Instructions**

1. On the right, you’ll see a popular data science interview question asked by tech companies. Take a moment to read and think about the question.  
   Try to come up with a solution on your own. Then, take a look below for the complete solution.  
   **Click to see the solution**First, it’s important to understand what the question is asking you to solve.  
   You need to figure out the probability that it’s raining in Seattle, given that all three friends tell you that it is raining.  
   Abstractly, we can write this as:
2. P(Raining∣Yes,Yes,Yes)
3. *P*(*Raining*∣*Yes*,*Yes*,*Yes*)
4. Next, let’s use Bayes’ Theorem to rewrite the above. Here is a refresher on Bayes’ Theorem:
5. P(A∣B)=P(B∣A)⋅P(A)P(B)
6. *P*(*A*∣*B*)=
7. *P*(*B*)
8. *P*(*B*∣*A*)⋅*P*(*A*)
9. ​
10. Using this formula, we get:
11. P(Raining∣Yes,Yes,Yes)=P(Yes,Yes,Yes∣Raining)⋅P(Raining)P(Yes,Yes,Yes)
12. *P*(*Raining*∣*Yes*,*Yes*,*Yes*)=
13. *P*(*Yes*,*Yes*,*Yes*)
14. *P*(*Yes*,*Yes*,*Yes*∣*Raining*)⋅*P*(*Raining*)
15. ​
16. Let’s break this down even further.  
    We can easily solve for P(Yes, Yes, Yes|Raining) since the problem states that each friend has a 2/3 chance of telling the truth and all three said it was actually raining. So if all three friends tell you it’s raining and all three friends are telling the truth:
17. P(Yes,Yes,Yes∣Raining)=2∗2∗23∗3∗3=827
18. *P*(*Yes*,*Yes*,*Yes*∣*Raining*)=
19. 3∗3∗3
20. 2∗2∗2
21. ​
22. =
23. 27
24. 8
25. ​
26. What about P(Raining)? This is the probability that it rains on any given day (which has nothing to do with what your friends say). We’re missing this crucial piece of information!  
    This is a good time to ask your interviewer a clarifying question. Let’s say that your interviewer tells you to assume that P(raining) is 25%.  
    Great! You now have the knowledge you need to solve the problem.  
    Let’s solve for P(Yes, Yes, Yes), which can be rewritten as:
27. P(Yes,Yes,Yes)=P(Raining)∗P(Yes,Yes,Yes∣Raining)+P(Not−raining)∗P(Yes,Yes,Yes∣Not−raining)
28. *P*(*Yes*,*Yes*,*Yes*)=*P*(*Raining*)∗*P*(*Yes*,*Yes*,*Yes*∣*Raining*)+*P*(*Not*−*raining*)∗*P*(*Yes*,*Yes*,*Yes*∣*Not*−*raining*)
29. Now we have all of the necessary information to solve for P(Yes, Yes, Yes):
30. P(Raining)=0.25
31. *P*(*Raining*)=0.25
32. P(Yes,Yes,Yes∣Raining)=827
33. *P*(*Yes*,*Yes*,*Yes*∣*Raining*)=
34. 27
35. 8
36. ​
37. P(NotRaining)=0.75
38. *P*(*NotRaining*)=0.75
39. P(Yes,Yes,Yes∣NotRaining)=1∗1∗13∗3∗3=127
40. *P*(*Yes*,*Yes*,*Yes*∣*NotRaining*)=
41. 3∗3∗3
42. 1∗1∗1
43. ​
44. =
45. 27
46. 1
47. ​
48. Therefore, P(Yes, Yes, Yes) can be simplified to:
49. P(Yes,Yes,Yes)=0.25∗827+0.75∗127
50. *P*(*Yes*,*Yes*,*Yes*)=0.25∗
51. 27
52. 8
53. ​
54. +0.75∗
55. 27
56. 1
57. ​
58. Let’s go back to what we are trying to solve (the probability that it is raining given that all three friends said it was):
59. P(Raining∣Yes,Yes,Yes)=P(Yes,Yes,Yes∣Raining)⋅P(Raining)P(Yes,Yes,Yes)
60. *P*(*Raining*∣*Yes*,*Yes*,*Yes*)=
61. *P*(*Yes*,*Yes*,*Yes*)
62. *P*(*Yes*,*Yes*,*Yes*∣*Raining*)⋅*P*(*Raining*)
63. ​
64. Putting everything together, we get:
65. P(Raining∣Yes,Yes,Yes)=(827⋅0.25)0.25∗827+0.75∗127
66. *P*(*Raining*∣*Yes*,*Yes*,*Yes*)=
67. 0.25∗
68. 27
69. 8
70. ​
71. +0.75∗
72. 27
73. 1
74. ​
75. (
76. 27
77. 8
78. ​
79. ⋅0.25)
80. ​
81. We can simplify further by canceling out the 27’s and multiplying the numerator and denominator by 4.
82. P(Raining∣Yes,Yes,Yes)=811
83. *P*(*Raining*∣*Yes*,*Yes*,*Yes*)=
84. 11
85. 8
86. ​

Therefore, the probability that it is raining in Seattle, given that all three of your friends said it is raining, is 8/11.



### **Interview Etiquette**

1 min

During the interview process, it’s important to follow some basic etiquette rules.

If you are invited to interview, make sure to do thorough research about the company and role beforehand.

* What are the history, mission, and values of the company?
* What is the company’s technology stack?
* Why is this role important to the company mission?

##### **Look into the company and role**

Doing your homework will allow you to have more engaging conversations with the interviewer. Show them that you care about the company mission and express your curiosity by asking specific questions.

##### **Learn about the interviewers**

Look up your interviewer(s) on LinkedIn beforehand so that you can better understand who you’re speaking with. The more context you have, the easier it will be to develop rapport and have a smooth conversation. It’s also helpful to know how technical your interviewer is.

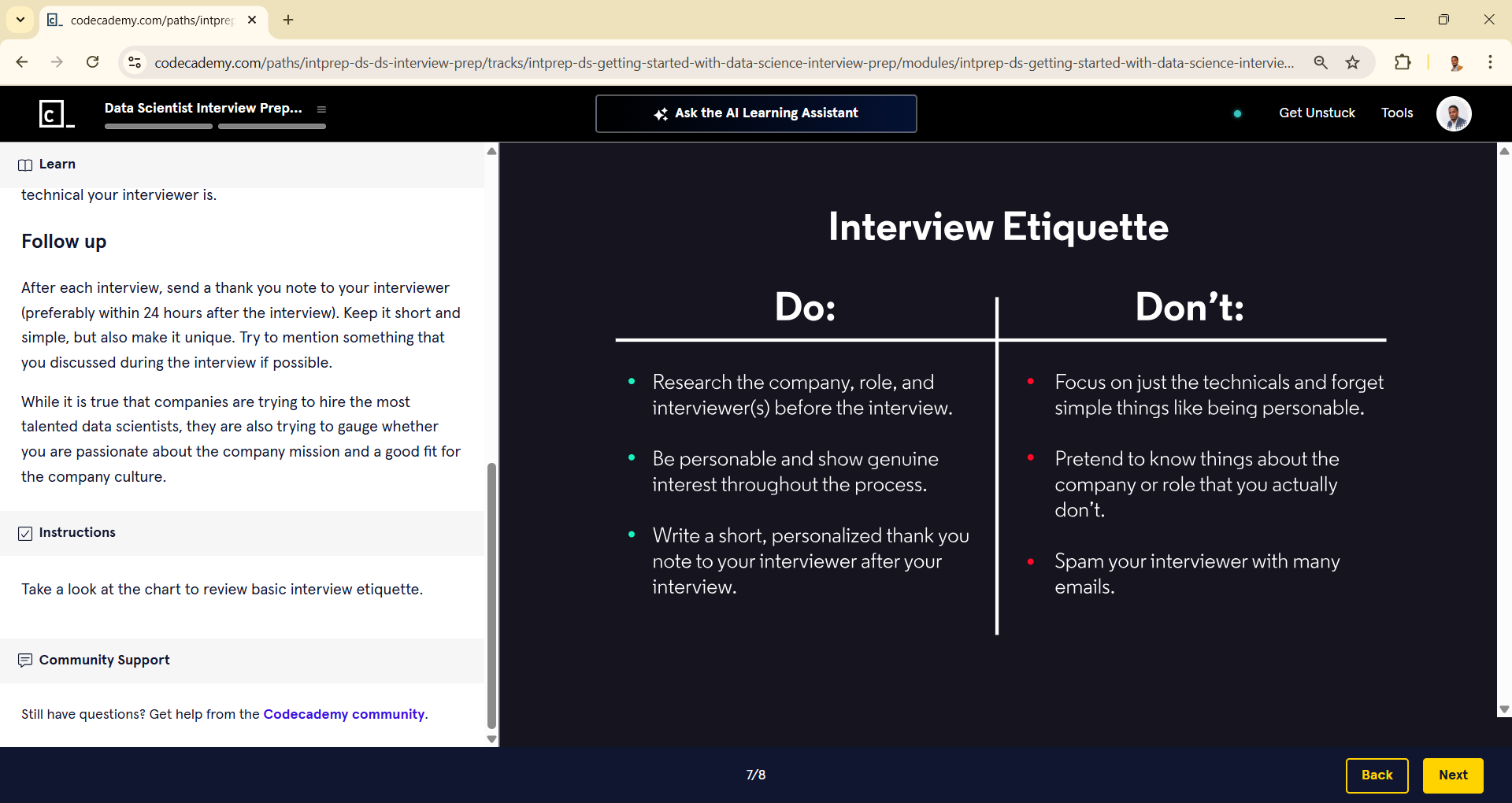
##### **Follow up**

After each interview, send a thank you note to your interviewer (preferably within 24 hours after the interview). Keep it short and simple, but also make it unique. Try to mention something that you discussed during the interview if possible.

While it is true that companies are trying to hire the most talented data scientists, they are also trying to gauge whether you are passionate about the company mission and a good fit for the company culture.

**Instructions**

Take a look at the chart to review basic interview etiquette.



**Overview of the Data Science Interview Process**

### **Review**

<1 min

Congratulations! You now know about the basic components that make up the Data Science Interview Process:

* Polishing your resume
* Building a digital presence
* Creating a portfolio
* The phone screening
* The take-home assessment
* Technical and whiteboard interviews
* Basic interview etiquette

Knowing what to expect before going in is a great way to boost your confidence.

Throughout this path, you’ll get a ton of practice with technical interview questions that you can expect to see on interview day.

By the end of it, you’ll be ready to tackle most interviews.

**Instructions**

When you’re ready, move on to the next part of this Path.

# Common Data Science Roles

**An introduction to navigating data analyst and data scientist roles across industries**

#### **Introduction**

Searching for jobs in data analytics or data science can be a tricky process to navigate. For example, you might find different job titles such as *Data Analyst*, *Decision Scientist*, *Quantitative Analyst*, or *Applied Research Scientist*.

Which one do you choose? How do you prepare for each?

Learning what to expect in your job search will help you approach the process with confidence. In this article, you will learn:

* Common data analyst and data scientist roles
* How to recognize roles in job postings
* How to prepare for each type of role

Let’s get started!

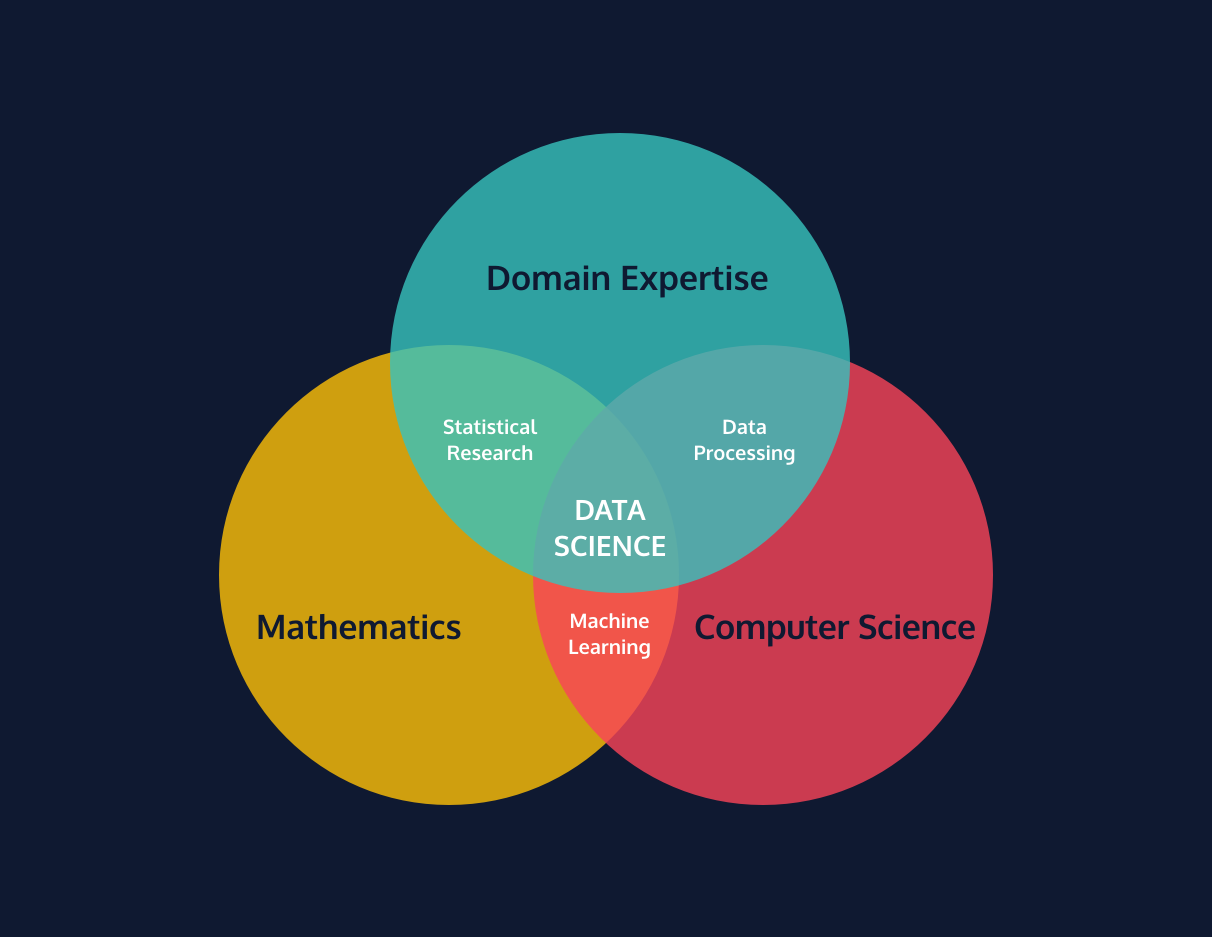
Free response

Describe your ideal role. What do you want to spend the majority of your day working on?

**Submit Response**

#### **Common Data Science Roles**

Data science is a broad, interdisciplinary field that combines other disciplines such as probability, statistics, analytics, and computer science. The work that you might do could range from writing up reports to building machine learning models. This means you are likely to find a wide variety of role titles when searching for jobs in the field of data science.



Common titles you might see include:

* Data Analyst
* BI Analyst
* Business Analyst
* Data Scientist
* Decision Scientist
* Applied Research Scientist
* Predictive Modeler/Scientist
* Data Scientist - Algorithms
* Data Scientist - Machine Learning

Note that the job title you see does not always give a complete picture of what is expected on the job. How one company defines the role of a data analyst can be very different from how another company chooses to do so. As we’ll discuss shortly, it is much more helpful to look at the job description and requirements.

#### **Common Data Analytics Roles**

Data analytics is similar to data science in many ways, but there is much less emphasis on programming and does not include things like statistical/predictive modeling and machine learning.

When searching for jobs in data analytics, common titles you might see include:

* Data Analyst
* Quantitative Analyst
* Business Intelligence Engineer/Analyst/Developer
* Business Analyst
* Operations Analyst
* Data Scientist - Analytics
* Data Scientist - Inference

In many cases the job title might simply be “Data Analyst” but other times it will be something more specific. This has a lot to do with the company and industry you are looking at. For example, if you apply for a data analyst position at an investment firm, you are likely to see something like “Quantitative Analyst” or “Financial Analyst.”

It is also important to remind yourself that data analyst job titles don’t always reflect the actual responsibilities. Make sure to dig in the job description and responsibilities to make sure the role is what you expect.

#### **Specialized Roles**

The field of data science and analytics is still relatively new. As it continues to grow, people are becoming more specialized in specific domains such as machine learning and natural language processing.

Here are some examples of roles that require specialized knowledge:

* **Machine Learning Engineer/Scientist:** Requires knowledge of big data, deep learning, DevOps, and preferably expertise in a particular industry or working with a specific type of data. For example, a self-driving car startup might want to see that you have experience working with computer vision and GPS data.
* **NLP Engineer/Scientist:** Requires knowledge of deep learning, natural language processing, and extensive experience working with text data. As large companies like Google and Amazon race to build the best speech/voice products, they are eager to hire the top engineers in the field of natural language processing.
* **Data Engineer:** Requires knowledge of databases, data structures, how to set up data pipelines, big data/cloud computing, backend engineering experience.

This trend of specialization is likely to continue as more use cases arise in data science and analytics. It is already becoming more common to see job titles such as “Computer Vision Engineer” or “Senior Language Engineer” as companies look for talented people who specialize in these domains.

#### **Recognizing Roles in Job Postings**

Job titles can be deceiving – they don’t always reflect the actual responsibilities of the job. This is especially true for data science and data analytics, which are relatively new fields that aren’t fully defined yet. Many data science jobs are actually data analyst jobs in disguise and vice versa.

That’s why it is much more helpful to look at the job description and responsibilities. Sometimes there will be a “Skills/Tools” section that lists out all of the tools, programming languages, and libraries that are required for the job. This can be a helpful way to see if the job matches your expectations.

A quick way to distinguish between a data analyst and data scientist role is to remind yourself of the responsibilities for each role. Here is a quick review of each role:

**Data Analyst:**

* Access and query different data sources (e.g., SQL)
* Process, manipulate, and clean data
* Summarize data
* Employ statistics and mathematical techniques
* Data visualizations and reports
* **Data Scientist:** Requires knowledge of all of the above, plus:
* Much more emphasis on programming (Python, R)
* Statistical modeling
* Machine learning knowledge

In general, data analysts look at the past to find trends, summarize data, and communicate findings, while data scientists use the past to make a prediction about the future. If you see phrases like “Predictive analysis,” “Statistical modeling,” “Machine learning,” or “Extensive Python/R experience,” that means you are probably looking at a data science role. If the role lists skills such as Excel, SQL, Tableau, and minor programming experience, then you are most likely looking at a data analyst role.

Context is also really important when going through job postings. Make sure to pay close attention to the company and industry that you are applying to. How a marketing company chooses to call its data analysts (“Marketing analyst”) might be very different from how a financial data company chooses to do so (“Quantitative analyst”).

Also, each company has a unique workflow and slightly different expectations for its data scientists and analysts. Some will even require that you have industry-specific knowledge or experience working with a particular type of data.

#### **Preparation**

Your interview preparation will vary based on the role that you decide on and the company and industry you are interested in. Generally speaking, if you apply to a data science role at a large company, you should make sure to have a strong understanding of general data questions, Python, machine learning, with less focus on SQL and data visualization/communication.

For a data analyst or data scientist role at a smaller company, you are less likely to see machine learning or Python questions. You should focus your attention on general data questions, SQL, and a heavy focus on data visualization and communication.

The reality is that there is not one clear answer on what to study for a data science or analytics interview. To fully prepare for your interview:

* Set yourself up with a solid understanding of the topics listed above.
* Tailor your approach based on the company and industry you are applying to.

For example, if you are applying to the Alexa team at Amazon, in addition to machine learning knowledge, you will need to brush up on your natural language processing knowledge. If you are set on working as a quantitative scientist at a hedge fund, you will need to practice analyzing financial data with Python.

# Review of Getting Started with Data Scientist Interview Preparation

**See what you learned about different data scientist and data analyst roles and job interviews.**

### **Review**

Congratulations! The goal of this unit was to introduce you to the different types of data scientist roles and the corresponding job descriptions.

Having completed this unit, you are now able to:

* Read a job description and better understand what skills the company is looking for.
* Tailor your job search materials to meet the requirements stated in a job description.
* Determine what type of role(s) you would be most interested in.
* Prepare for the stages of a data scientist job interview.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# Welcome to Fundamental Data Science Interview Questions

**Learn how to communicate about general data principles and practices.**

### **Goals of this Unit**

The goal of this unit is to give you practice answering general data questions. These questions may or may not appear as stand-alone questions during your interview, but may be embedded in others. Data scientists are expected to be fluent with all of these topics.

After this unit, you will be able to:

* Talk through how you work with data and manage different data types.
* Define your data cleaning process.
* Implement basic summary statistics and explain the significance of the results.
* Speak to when you use different sampling methods.
* Walk through solutions to probability interview questions.
* Approach estimations problems logically and systematically.
* Apply hypothesis testing to practical business problems.
* Effectively communicate your thought process and problem-solving skills.
* Clean and summarize a real dataset.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# Data Types Interview Questions

**Practice answering common interview questions about general data concepts and data types.**

### **Introduction**

Hi! My name is Codey Cademy. I’m so excited because today I’m interviewing candidates for the data scientist role on my team, which means that I get to hire someone to help me analyze all this data! I got my budget approved, went through all of the applications, did my phone screens, and now I’m down to just a few candidates.

Whoever I hire needs to be data-fluent. I’ve heard stories from my friends about hiring people who know all about the latest algorithms but just didn’t know the basics. That can really cost a lot of time and money, and I don’t want that to happen to me. So, I’m going to make sure my next hire is skilled in all stages of data science.

There are three types of tasks I want to be sure they can do:

* Exploratory analysis
* Diagnostic analyses
* Data organization

I’m going to ask them a mix of conceptual and applied questions to be sure they know the concepts and that they’d be able to apply them to the real world.

I might integrate some of these basic concepts into the other questions I ask them, too, but the basics are so important for this role that I am going to ask directly.

I really hope they talk through their answers. It’s so much easier to understand how someone thinks about data when they vocalize their thought process and assumptions. If I could give them one piece of advice for our day together, it would be:

When in doubt, talk it out!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Theoretical Questions**

Hi! Welcome to your basic data types interview! We’re going to start off with a few conceptual questions about working with data.

A person with long orange hair and light brown skin is looking up at a graph with stars as data points. The larger stars are connected by lines, creating a line graph. There is a telescope in front of the person.

Free response

How do you typically start an exploratory data analysis?

**Get feedback**

Powered by OpenAI

Free response

Tell me about a data analytics project\* you worked on and the major steps you took to complete it.

(\*Note that this could be a project you complete as part of this Skill Path.)

**Submit Response**

Great! You sound like you know your way around a dataset! Let’s move on to applying these concepts to real-life problems.

### **Situational Questions**

There are a lot of companies and organizations out there and we all face different challenges. The best interviewees I’ve met with really knew my company. They did their research and gave some thought to the types of problems we face and the data we work with. Let’s go ahead and apply the general concepts to some business problems.

Free response

An e-commerce site: We have been collecting data on user behavior including clicks, time to checkout, unique visits, etc. Yesterday we saw a 50% drop in checkouts. We didn’t receive any customer complaints though. Using just the site data, how would you start to diagnose the problem?

**Get feedback**

Powered by OpenAI

Free response

A rideshare startup: We have demographic data, car ownership, and insurance premium data from 2019-2021 for residents in the Detroit area. We want to expand our services there in 2022. With this in mind, what questions would you develop for this dataset and how would you begin to explore this data?

**Get feedback**

Powered by OpenAI

Thanks! It sounds like you’ve really done your homework on how we think about data! Let’s see how you do with some of the specialist data types we work with.

### **Temporal Data**

We just got access to new data sources, and one of them is time-bound! It looks to me like there’s some seasonality to it, but I’m not sure. I know a little bit about working with time, but not a lot. Looks like you also list familiarity with time series on your resume. Wonderful! Let’s talk about some simple time series analysis.

Free response

What is time series analysis, and what aspects would you keep in mind when working with time series?

**Get feedback**

Powered by OpenAI

Coding question

## **Questions**

Change a string into a timestamp in Python.

### **Code**

Output:

Run your code to check your answer

Free response

Toy company: We launched a new toy campaign this year and want to know if our marketing was successful. How would you account for seasonal variation to determine if our advertisements affected sales?

**Get feedback**

Powered by OpenAI

Wow! You really know your stuff. Turns out that we also have some unstructured text. Looks to me like Tweets and reviews. I know it’s asking a lot from one person to be able to work with all of these different data types, but let’s go through some textual data questions.

### **Text Data**

I only took a brief look at the new data we are pulling into our data lake, but I’m hoping you can help make some sense of it. I’m going to ask a few questions about how you’ve worked with text.

Free response

What steps would you take to normalize text?

**Get feedback**

Powered by OpenAI

Free response

Large manufacturing company: We have a collection of employee generated incident reports. We would like to use these records to identify trends across departments. What would you do to process this data?

**Get feedback**

Powered by OpenAI

### **Review**

Great job! I was really impressed with your answers and think you will be a great fit!

We covered:

* Basic data management.
* Applying data management concepts to business problems.
* Working with both time-series and text data.

You showed a thoughtful and organized approach, and I’m starting to get a sense of how you approach problems because you were able to articulate your thought process.

We’d love to move you along to the next stage in the hiring journey. Let’s take a look at some more questions.

# Data Cleaning Interview Questions

**Practice data cleaning and data wrangling interview questions.**

### **Introduction**

Welcome back!

In this portion of the interview, are going to talk about your strategies for cleaning data. Data cleaning is a huge part of my job (sometimes it feels like my only job). Some people estimate that up to 80% of data scientists’ analysts’ jobs is data cleaning!

I’m going to ask you two types of questions:

* Conceptual questions asking about techniques
* Situational questions asking how you would approach a dataset

You’ll have a chance to talk through how you think about cleaning data and the steps you take. This is not the last time you will have to clean data, though. It will surely be a part of your take-home challenges. Since cleaning data is so central to our jobs, I need to be sure our next data scientist is up for the challenge!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Conceptual Questions**

Free response

What steps do you take when you first clean a dataset?

…

**Feedback**

Not quite

**What you did well**

It seems like you didn’t provide an answer to the question. It’s okay, everyone has moments of uncertainty.

**What could be improved**

Next time, try to provide an answer that reflects your understanding of the topic. In this case, you could discuss the steps involved in cleaning a dataset, such as inspecting, repairing, validating, and documenting the data.

Don’t worry, learning is a journey and it’s okay to have bumps along the way. Keep going, you’re doing great!

Powered by OpenAI

Is this feedback helpful?

Review our answer

Free response

What data validation methods have you used?

…

**Feedback**

Not quite

**What you did well**

Unfortunately, your response was left blank and does not provide any information related to the question.

**What could be improved**

It’s important to provide a response to the question. In this case, you could discuss the different data validation methods you have used, such as form-level validation, regular expressions, checking for duplicated data, and so on.

Don’t worry, everyone has moments of uncertainty. Take a moment to review the material and try again. You’re capable of understanding this!

Powered by OpenAI

Is this feedback helpful?

Review our answer

A graph with two lines where there are many round holes of varying sizes cut out, in a fashion that resembles the holes of Emmental cheese.

Free response

Do you ever impute data? Why or why not?

**Get feedback**

Powered by OpenAI

Fantastic! It sounds like you have a good theoretical foundation for how to clean data - let’s try it out in some code.

### **Coding Questions**

Coding question

## **Questions**

Find the rows with missing values in this dataset.

* If a record has more than two missing values, remove it from the dataset.
* If a record is only missing one value, replace the missing value with the median of that variable.
* Create a new column where you indicate that a value was imputed (you do not need to indicate which value was imputed). Call this column Imputed.
* The first 5 lines of this dataset are:

Name Salary Age Children Company

Vanna Holland 5738.45 40 5

Gretchen Kemp 9765.50 53 2 Feugiat Lorem Ipsum Company

Brittany Porter 8557.05 37 Magna Ut Industries

Kaseem Crawford 1424.80 20 4

Lamar Payne 2442.63 28 5 Amet Incorporated

### **Code**

### **Situational Questions**

Free response

Healthcare Company: We have millions of records of anonymized data, and a lot of appointment notes are missing. We want to determine if there is a relationship between the amount of time nurses spend with patients and their complaint category, but we cannot do that without the notes.

How would you deal with the missing data?

**Get feedback**

Powered by OpenAI

Free response

Non-profit organization: We are conducting a survey about attitudes towards policing. How would you determine if survey responses were filled in at random by certain individuals?

Your response

..

Our answer

I am assuming that there are multiple questions in this survey and it is a multiple choice. If I am involved in the design of the survey, I would include similar questions worded differently to use as a control. Likewise, I might include logically incompatible questions like “I eat beef” and “I observe a vegan diet”.

Once the data is already collected, I think about it as a flagging system where suspicious activities get flagged, and no one suspicious activity is itself a flag, but when there are multiple flags, that can indicate that the respondent filled in questions at random.

For example: First, I would look at how long it took each participant to fill in each question, and focus on the participants who filled in the questions very quickly as one suspicious group.

Second, I would look at any participant who answered all the questions with the same ranking (i.e., the highest, lowest, first, or last option). I would also consider this group to be a suspicious group. Finally, I would cluster respondents together, assuming that there are trends in the responses. Any respondent that does not fit into a cluster deserves closer investigation.

If I were involved in the design of the survey, I might include the same question worded differently to be sure that the answers match, and then use this as another flag.

When all of the flags were assigned, I would then look at the responses to determine if there are any patterns that I am missing or any other identifying information. I would then indicate that these are suspicious responses. I would then make a decision about whether to include them or not.

WHY THIS RESPONSE?

This answer demonstrates an awareness that there is no simple solution to this problem, and the best solutions will draw on multiple methods. It highlights multiple ways to think about how real people answer surveys, and what data relics are left when they do.

Also, this answer communicates a conservative approach to data management by including a step of manually evaluating the flagged records and checking your work.

### **Review**

Great job! It’s clear to me that you have really worked with cleaning data before!

In this interview, we covered:

* Dealing with missing data
* Imputing data
* Data validation
* General data cleaning

You gave some really thoughtful responses to all of these questions. Let’s move on to the next sectionl

### **Review**

<1 min

Congratulations! You are ready to answer summary

[statistics](https://www.codecademy.com/resources/docs/general/statistics)

Preview: Docs Loading link description

questions that might come your way in your data analyst or data scientist interview.

In this lesson, we covered:

* Calculating the mean and median of a variable and how to interpret them.
* Calculating and reporting the range of a variable
* Creating a histogram and evaluating it for skewedness.
* Calculating the interquartile range and comparing it to the standard range.
* Evaluating the relationship between quantitative   
  [variables](https://www.codecademy.com/resources/docs/general/julia/variables)
* Preview: Docs Loading link description
* to determine if there is a linear association.

Good luck and see you in the next interview!

Note that the dataset used in this lesson was downloaded from [kaggle](https://www.kaggle.com/neuromusic/avocado-prices) and originally compiled from the [Hass Avocado Board](http://www.hassavocadoboard.com/retail/volume-and-price-data).

# Sampling Interview Questions

**Practice answering common sampling interview questions**

### **Introduction**

Welcome back! This is the sampling portion of your interview. Sampling is a large part of what we do on our team.

It would be great if we could collect data from our entire population, but that just isn’t realistic. So smart sampling is absolutely essential. I hope you are up for the challenge and well versed in all the different ways to sample from a population. On our team, we use a variety of tools, but the most important tool of them all is knowing when to use each one!

You’ve done a great job talking through your thought process and explaining why you implement different methods so far, and I look forward to your answers in this section, too!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Conceptual Questions**

Let’s start with some general questions about sampling.

Free response

How do you decide which sampling technique to use and what do you take into account when deciding?

**Get feedback**

Powered by OpenAI

Four individuals standing in front of an image of the Earth. The person on the left has medium-dark skin and a beard, and is wearing an orange checked shirt and purple turban. The second person has dark skin, longish black hair, and a pink and purple striped shirt. The third person has light skin, a bright red headscarf, and purple and pink dotted shirt. The fourth person has yellow hair, medium-dark skin, and a blue buttoned shirt

Free response

How do you typically create a representative sample?

**Get feedback**

Powered by OpenAI

Free response

What is selection bias and how do you account for it in your data?

**Get feedback**

Powered by OpenAI

Free response

How do you determine sample size?

**Get feedback**

Powered by OpenAI

Great job! It’s clear to me that you have a good sense of the theory behind sampling. Let’s move on to some applied questions.

### **Applied Questions**

Free response

A publishing house: We have a user base of about 1 million customers. About 80% only access Freemium content, and the other 20% have some mixture of paid subscriptions. We want to learn more about customers’ interests so we can maximize our revenue. We have a budget for a survey. How would you approach this project? Random sampling or Stratified sampling - why? Focus on all customers or only the 20% paid?

**Get feedback**

Powered by OpenAI

Free response

A grocery store:

200 customers enter our store every day. We want to know what path they take through the store and what sections they visit first. How will you determine which customers to track?

**Get feedback**

Powered by OpenAI

Free response

A healthcare company:

We want to know how many people in a community have a respiratory illness. We have established a testing center where people can come to be tested for free. What steps would you take to account for the bias in the sample?

Your response

Give the best

Our answer

Even when something is offered for “free”, in the sense that it is given without payment, people have to have the time and transportation to access it. I tend to think of this as the intangible ways that people have to pay.

This sample will be biased towards those who are able to reach the location and who have the flexibility in their schedule to make the trip. They can pay with non-monetary resources. We can assume that more people who live closer will be represented, but that is an assumption. Depending on the demographic makeup of the community and specific transportation routes, that could be an inaccurate assumption.

The best way to get a measure of the community incidence would be to take a survey of test takers and align that to the demographic distribution of the community. However, it wouldn’t be a direct demographic comparison. Since there are other socio-economic factors that will contribute to who is able to take the test, these have to be considered.

For example, if taking a test meant missing work, people who cannot afford to miss work might avoid going to take a test or might only take one if they are very very ill. This has the potential to both make it seem like a higher proportion of that community is actually ill, but also may limit the ability to make an accurate estimate.

It would also be beneficial to ask some survey questions alongside demographic information to address these issues. However, this has to be balanced with concerns around privacy and confidentiality. Because those concerns are outside of my area of expertise, I would likely work with other stakeholders who can advise on what questions would be appropriate.

WHY THIS RESPONSE?

This is a good response because it addresses real world practicalities that are often overlooked when providing “free” services or items. The interviewee further considers the limitations of strictly normalizing on demographics and introduces through the use of a clear example some socio-economic factors that have to be considered. Finally, the interviewee talks about external issues outside of their control and the relevant stakeholders that they would work with in order to resolve them. This is another very realistic possibility that the interviewee takes into consideration which shows that they are also able to think outside of abstract concepts.

### **Review**

Congratulations! Sampling can be tricky, but you really know how to find the sample from a population!

In this interview, we covered:

* How to decide what sampling techniques to use,
* How to create a representative sample,
* What selection bias is and how to deal with it, and
* The best way to determine sample size.

You also made use of some general strategies to answer sampling problems such as:

* The ability to think out loud and consider a wide range of constraints when identifying the right sampling technique.
* The use of practical examples to illustrate the abstract concepts such as creating a representative sample.
* Determining and estimating the inputs to answer complex questions like determining sample size.
* Why picking a different technique could cause us to lose out on valuable information.

You also applied your skills by explaining how:

* A publishing house could use Stratified Random Sampling to answer their questions,
* A healthcare company could estimate the presence of an illness in a community, and
* A grocery store could measure how customers move through their store.

You are ready to showcase your sampling skills in a real-life interview, good luck!

# Probability Interview Questions

**Practice answering common interview questions about probability.**

### **Introduction**

Hi again! It’s me, Codey. Have you ever wondered what the odds are of running into probability problems as a data professional? You guessed it - pretty high!

That is why in this section I am excited to test your problem-solving skills through some fairly common, yet somewhat tricky, probability problems. While the problems we solve in our work are of a much broader scope, successful candidates have an intuitive understanding of counting and probability concepts, and are able to quickly take the simpler and more efficient route towards the solution. I certainly hope you are able to as well!

Do you have your pen and paper ready? You’ll have to work your way through some probability problems fairly quickly, but please don’t leave me in the dark! Whenever you can, I would really like to hear about your thought process to understand how you approach and solve problems. Let’s get started!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Conceptual Questions**

**Image of person waving from inside a train car. The person is on the leftmost window. There are two windows and 2 doors in the center. The car is atop 2 lines representing the tracks.**

Free response

On any given day, there is a 10% chance you will see a friend on the subway. Assuming you ride the subway 5 days a week, what is the probability that you will see a friend at least once?

**Get feedback**

Powered by OpenAI

Free response

A couple has 2 children. At least 1 is a girl. What is the likelihood that both are girls?

**Get feedback**

Powered by OpenAI

Free response

A fair six-sided die is rolled twice. What is the probability of getting 1 on the first roll and not getting 6 on the second roll?

**Get feedback**

Powered by OpenAI

Great job with the conceptual questions! It’s clear to me that you have a solid grasp of the theory! Let’s move on to implementing these principles in code.

### **Coding Challenges**

Coding question

## **Questions**

How would you simulate the behavior of a fair coin?

### **Code**

Coding question

## **Questions**

How would you simulate the behavior of an unfair coin?

### **Code**

Great job! You clearly have the ability to simulate probability problems computationally, which is a really important skill to have when working with data! Let’s move on to applying these concepts to real-world scenarios.

### **Applied Questions**

Free response

Facebook: This question was famously asked at Facebook and has since become representative of questions about probability. We have two options for serving ads within Newsfeed:

* Out of every 25 stories, one will be an ad.
* Every story has a 4% chance of being an ad.

For each option, what is the expected number of ads shown in 100 news stories? If we go with option 2, what is the chance a user will be shown only a single ad in 100 stories? What about no ads at all?

**Get feedback**

Powered by OpenAI

Free response

You call 2 UberX’s and 3 Lyfts at the same time. The time to reach you can be considered independent and identically distributed. What is the probability that all the Lyfts arrive first? What is the probability that all the UberX’s arrive first?

**Get feedback**

Powered by OpenAI

Free response

If three friends in London told you it’s raining, and there is a 1/3 probability that each person is lying, what is the probability that it’s raining in London?

**Get feedback**

Powered by OpenAI

### **Review**

Thanks so much for taking the time to work through some fairly tricky probability problems! In this interview, you encountered a range of problems and demonstrated your understanding of key concepts through the application of problem-solving strategies and clear communication of your ideas.

Here are some concepts that you used in the problems that you encountered:

* The idea of complementary, conditional, and independent events.
* Creating a script to solve a relatively simple simulation problem.
* The distinction that the chance of an event happening is not the same as an event that is guaranteed to occur x times every n trials.
* Conditionally independent events, and how they can make traditional methods fail (such as complementary counting).

You really showcased your strong and intuitive grasp of probability concepts today! Now you are ready to do the same in a real-life probability interview as well. Good luck!

# Estimation Interview Questions

**Have fun estimating answers for estimation interview questions**

### **Introduction**

Hello, again! It’s me, Codey.

Have you ever wondered how much space it might take to gather the entire population of the world? But wait a minute—nobody can magically command everybody to gather at the same place at the same time. So who knows, right?

Indeed, it would be impossible to conduct a real-world experiment for these kinds of goofy problems. And that is where some clever estimation skills can come in handy. Don’t get thrown off by the seemingly bizarre nature of these problems because the same estimation skills used to answer them can be extended to a real-world data analysis project as well. This is what you’ll be tested on here.

I love it when interviewees talk through their solution by first laying out the parameters:

* What information do you need to know to calculate this estimate?
* What pieces of information are dependent on other pieces of information?

It’s also really helpful when candidates break their answers down into steps.

Do your best and try to make a reasonable guess!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Estimation Questions**

Free response

How many people can fit on a subway train in Tokyo?

**Get feedback**

Powered by OpenAI

two bright red strawberries surrounding the Eiffel Tower

Free response

How many boxes of strawberries are for sale in Paris today?

**Get feedback**

Powered by OpenAI

Free response

How many bicycles would you need to set up a rideshare program in Boston?

**Get feedback**

Powered by OpenAI

Free response

How would you estimate the value of pi?

**Get feedback**

Powered by OpenAI

Free response

How many tourists will see a billboard in Times Square in a typical year?

**Get feedback**

Powered by OpenAI

### **Review**

Here are some key ideas you used in answering these estimation problems:

* Laying out the parameters—what information do you need to know and which pieces of information are dependent on others?
* The number of units that can be fit into some space is equal to the volume of the space divided by the volume of the total units (e.g., a metro).
* The proportion of people doing an activity right now is roughly equal to the proportion of the entire time an average person spends doing the activity (e.g., a ride-share program).
* Taking the geometric mean of the minimum and maximum range allows you to preserve the order of magnitude across the entire range and narrow in on a somewhat precise estimate (e.g., NY tourists who see a billboard).
* Solving estimation problems bottom-up (e.g., boxes of strawberries for sale) as well as top-down (e.g., NY tourists who see a billboard)

Great job estimating these somewhat tricky numbers! If you apply these same concepts and strategies to a real-life interview, we estimate *your estimates* will be quite precise.

In the worst case, even if you are off on a problem or two, you will leave the interview demonstrating your ability to:

* consider different parameters
* break down open-ended problems into steps

And both of these skills are commendable traits in a successful candidate. Good luck!

# Hypothesis Testing Interview Questions

**Practice answering common interview questions about hypothesis testing**

### **Introduction**

Hi, welcome back!

As you know, my team is in need of someone who can help us answer our data questions. We have a lot of data, I’d say maybe too much for one person to handle. I’m looking for a candidate who can work with samples and draw conclusions with a certain level of confidence.

As with any candidate, I am also looking for a problem solver who can clearly communicate their thinking and knowledge, someone who is curious about our data and has ideas on how to solve our specific business problems. I bet that person can be you!

Let’s dive back in with a few specific questions to give you a chance to show off your knowledge of hypothesis testing.

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Conceptual Questions**

Free response

What is the difference between type I and type II errors? Which one is worse? Why? Can you give me an example of each?

**Get feedback**

Powered by OpenAI

Free response

What is a p-value and what does it tell you?

**Get feedback**

Powered by OpenAI

An unlabeled histogram with 12 bars depicts an almost-normal distribution, slightly positively skewed.

Free response

What is the Central Limit Theorem and why does it matter?

**Get feedback**

Powered by OpenAI

Great job! It really sounds like you understand hypothesis testing. Let’s move on to applying these concepts to real-life scenarios.

### **Applied questions**

Free response

Finance company:

We assume that the average rate of return of a mutual fund is 9%. Talk through the steps of how you would validate if this is true. State any assumptions you would need to make and the data that you would need to validate your results.

**Get feedback**

Powered by OpenAI

Free response

Mortgage Lender:

We think that first-time homebuyers take out smaller loans than other borrowers. What information do you need to assess this and what would you do?

**Get feedback**

Powered by OpenAI

Free response

A marketing company:

We think that at least 60% of American households have an iPad. Our research team surveyed 200 households and found that 123 of them had iPads. Can we be sure that 60% is accurate? Why or why not?

**Get feedback**

Powered by OpenAI

Congratulations! You have done a great job with these hypothesis testing questions. It is clear to me that you know how to organize an experiment and evaluate the results.

### **Review**

In this portion of the interview, we covered:

* Step-by-step process for conducting hypothesis tests.
* The importance of representative samples when performing hypothesis tests.
* Code that can be used to calculate the confidence interval of a sample with 95% confidence that it is reflective of the true population.
* The theory behind the central limit theorem and a clear example.
* How vital the p-value is to any hypothesis test and what information you can gain from that value.

With all this preparation, you are going to do great in your next interview! You are ready to talk about hypothesis testing in a real interview setting!

# Data Cleaning & Summarizing Take-Home for Data Science

**Practice your data cleaning and summarizing skills with a sample take-home assessment.**

### **Overview**

Welcome to the take-home assessment!

This exercise helps interviewers get a sense of your technical skills before moving you to the next stage of the interview process. The style and format of these projects can vary widely depending on the company and role you are applying for. In this project, we will give you a prompt and you are free to answer it however you wish. You can use any coding language and any tools you are familiar with. This is an opportunity to show off your skills.

The projects that we have prompted here are typical for data scientist roles but are not definitive. However, there are some widespread practices and conventions that you should be aware of before starting the take-home.

* There may be one or more very simple question(s). These are not tricks – they are meant to easily exclude applicants who cannot perform basic tasks.
* There will be some discussion, analysis, or reflection. Expect to have to write a summary of your findings. A large part of both Data Analyst and Data Scientist roles is communication. How well you communicate your findings is just as important as what you find and how you find it.
* You may be asked to submit your files in a specific way. Pay attention to these instructions and submit your file in the requested format. This is another area that may initially seem unimportant but can have an outsized influence on your candidacy.
* You may *not* be asked to submit your files in any particular way. If this is the case, Python files should be .py or .ipynb, R should be in .R or .Rmd format. Similarly, your written analysis should be in .pdf format or .md if submitted online.
* If anything is unclear or you are not sure what they are asking you to do, ask! It is much better to get clarification at the beginning of a project than submit something that doesn’t answer the question. This assignment is an opportunity to showcase your problem-solving skills. If you are stuck on a particular technical skill, try to find the answer on your own. Google and StackOverflow will be your friends during this process! If you still can’t figure it out, and the interviewer has allowed for questions, reach out and explain what you’ve already tried. This will show you have the ability to problem-solve on your own before reaching out for help.
* Your interviewers may give you an estimate of how long they think it should take. It is very common to go far beyond the time estimate.
* The take-home assessment is also a good opportunity for you to decide if this is the kind of work you are interested in doing. If you enjoy it, you are probably in the right place. If it is something you do not enjoy, it may be worth clarifying what the expectations of the job are.

Let’s get started!

### **Context**

You are applying to be a Data Scientist at a digital marketing agency. A client in the UK wants to start advertising on YouTube and needs your help to perform some analysis on video data.

**Datasets:**

* [GBvideos.csv](https://www.kaggle.com/datasnaek/youtube-new?select=GBvideos.csv)
* [USvideos.csv](https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv)

Note: These datasets were both compiled by Mitchell J and are available on [Kaggle](https://www.kaggle.com/datasnaek/youtube-new).

### **Instructions:**

* Load your datasets and perform some simple exploratory data analysis.
* Create a new column to represent the like-dislike ratio (likes divided by dislikes) for the videos in each dataset.
* What is the average like-dislike ratio for all of the videos? (average number of likes per video divided by the average number of dislikes per video).
* Perform data manipulation to find the average number of likes per video for the US and GB in the year 2018.
* Business report: Your client hypothesizes that the most polarizing videos get shared the most. Analyze the data to determine whether your client is correct, and what other metrics should be investigated. Be sure to define any terms or cutoffs you make in the data. Present your findings in a short report for the client to review.

### **Checklist:**

Here is a checklist of things to consider as you work on this project:

* Make sure to perform EDA (descriptive statistics, data visualizations, etc).
* Make sure your code is clean and well-documented.
* Assessment questions such as Q2 and Q3 only have one correct answer. Make sure to double-check your math and write 1-2 sentences interpreting the values.
* Don’t forget to deal with special edge cases throughout, such as a division by 0 error.
* Make sure to clearly address the client’s issue in the business report.

**Note**: a take-home won’t typically include a checklist like this one, but we’ve included one here for you as you become more familiar with the take-home.

# Review of Fundamental Data Science Interview Questions

### **Review**

Congratulations! The goal of this unit was to give you practice answering general data questions.

Having completed this unit, you are now able to:

* Talk through how you work with data generally, including how you manage different types of data.
* Define your data cleaning process and prepare to answer questions about how you approach raw datasets.
* Implement basic summary statistics methods in Python and explain the significance of the results.
* Speak to how you use different sampling methods for different problems.
* Walk through solutions to probability interview questions.
* Approach estimations problems logically and systematically.
* Apply hypothesis testing to practical business problems.
* Effectively communicate your thought process and problem solving skills.
* Clean and summarize a real dataset.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# Introduction to Visualization, Communication, and Experimental Design Interviews

**See what you’ll cover in Visualization, Communication, and Experimental Design**

### **Goals of this Unit**

The goal of this unit is to get practice answering questions about data visualization, communication, and experimental design.

After this unit, you will be able to:

* Discuss how you pick a data visualization tool.
* Use data visualization to answer a practical business question.
* Describe your approach to experimental design.
* Talk to your interviewer about how you conduct A/B tests.
* Evaluate data from a real A/B test to make a recommendation about website layout.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# Visualization Tools and Techniques

**Practice answering data visualization interview questions**

### **Introduction**

Hi, it’s me, Codey, again!

In this part of the interview, we will talk about how you approach the data visualization process and which tools and techniques you use.

Data visualization is an important skill to have as a data professional because it allows you to tell meaningful stories and give your data visual context. It is especially important for a business intelligence role or a data analyst role in a product or marketing department.

There is no better place to show off your data visualization skills than your project portfolio. But I also like to ask candidates like you some visualization questions to make sure you really know your stuff.

I try to evaluate a few things:

* Do you understand the data visualization process?
* How familiar are you with common data visualization charts, tools, and techniques?
* Do you know which tools to use for EDA (Exploratory Data Analysis) and Communication?

I hope you’re prepared to talk about your experience with visualizing data because I’m excited to hear about it!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Fundamental Data Visualization Questions**

**Image of two eyes overlaid on many different types of visualizations used to represent data.**

Free response

What makes a good visualization?

**Get feedback**

Powered by OpenAI

Free response

What types of charts do you gravitate towards and why?

**Get feedback**

Powered by OpenAI

Free response

How would you represent six variables of a dataset?

**Submit Response**

Free response

Which tools do you use for EDA and communication?

**Get feedback**

Powered by OpenAI

### **Review**

Great job! It is clear that you have the data visualization skills and knowledge required for the job. Just to recap, we’ve talked about:

* What makes a good data visualization
* Common visualization charts, tools, and techniques
* Visualization for EDA and communication

You are now ready to move to the next stage of the interview process - see you there!

### **Project Overview**

Companies like to see that you are able to create meaningful visualizations and solve real-world problems.

In this portfolio project, you will get an opportunity to show off your data visualization skills while solving a business problem for a technology company using real-world data. The following datasets are good candidates to showcase your data visualization skills:

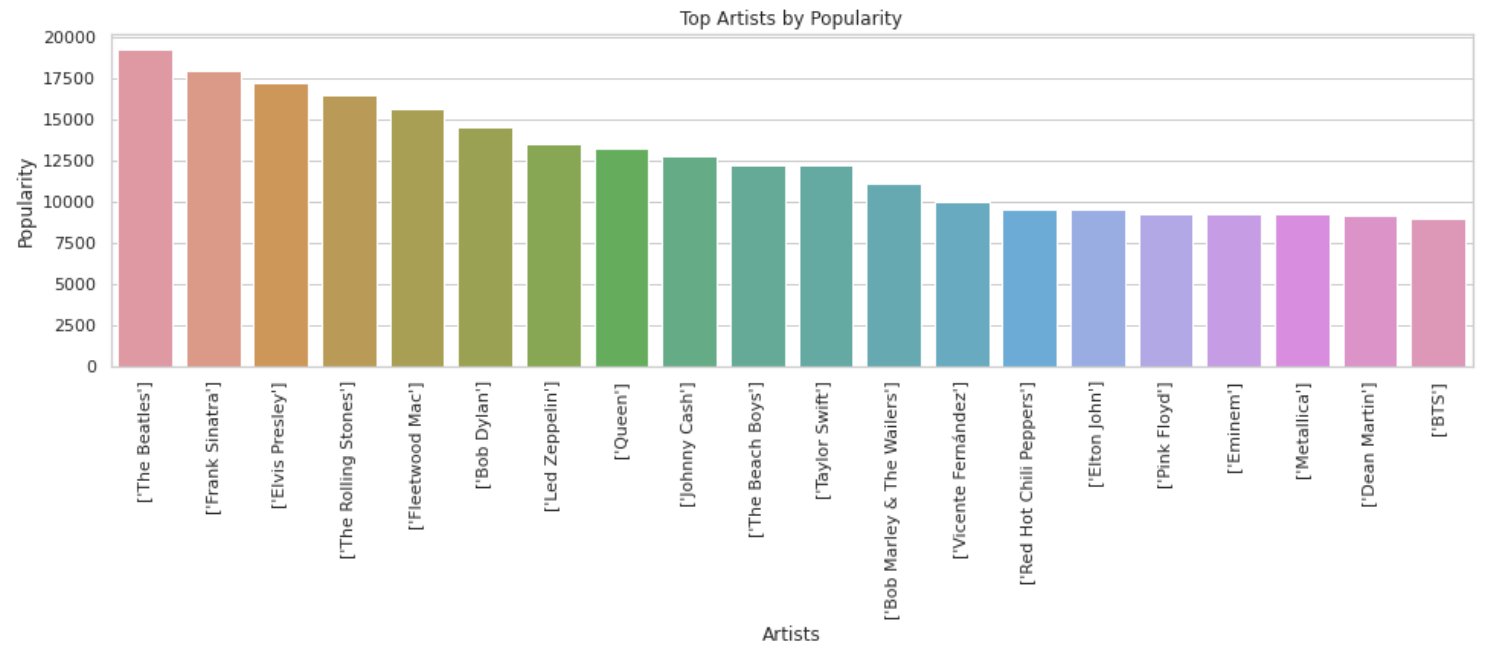
* Spotify [dataset](https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks) featuring 600,000 songs released between 1922 and 2021.
* Airbnb [dataset](http://insideairbnb.com/get-the-data.html) featuring listings in a city of your choice.
* Google [dataset](https://www.kaggle.com/lava18/google-play-store-apps) featuring 10,000 Google Play Store apps.

If you’d like, feel free to explore other datasets. Here are some places you can find datasets:

* [Kaggle](https://www.kaggle.com/datasets)
* [FiveThirtyEight](https://data.fivethirtyeight.com/)
* [UCI](https://archive.ics.uci.edu/ml/datasets.php)
* [awesomedata](https://github.com/awesomedata/awesome-public-datasets)

Think about the company you are interested in applying to and the type of data they interact with on a daily basis. This project is a great opportunity for you to show off your visualization skills using that data.

**Example Project**

****

**Project Objectives**

* Complete a project to add to your portfolio
* Choose a dataset and come up with a business problem to solve
* Create data visualizations to help solve a problem or answer a question
* Write documentation describing the dataset, the problem, and key takeaways

**Prerequisites**

* Data Wrangling and Tidying
* Summarizing Quantitative Data
* Data Visualization
* Communicating Data Science Findings

## **Project Tasks**

Keep track of your progress by dragging each task from "To Do" to "In Progress" to "Done" as you work on them. You can also click on a task to see more information about it.

To Do

In Progress

Done8 / 8 done

**Setting up your Git Repository**

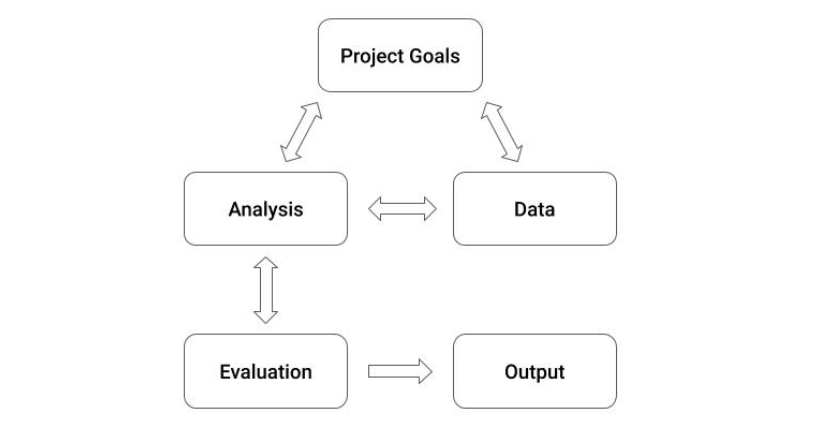
Create a new Git repository for this project.

**Project Scoping**

Properly scoping your project will greatly benefit you; scoping creates structure while requiring you to think through your entire project before you begin.

1. Start by stating the goals for your project.
2. Gather the data.
3. Consider the analytical steps required.

A proper project scope can be a great road map for your project, but keep in mind that some down-stream tasks may become dead ends which will require adjustment to the scope.



**Choose a Dataset**

Select one of the following datasets to work with or choose another dataset you’d like to explore:

* Spotify music data
* Airbnb listing data
* Google apps data

If you’re applying to a company that interacts with a particular type of data, this is a great opportunity to show off your data visualization skills using that data.

**Pose and Solve a Problem**

Look at your dataset and get curious!

Define a question you want to answer or problem you can explore. Go through each variable, what do you want to know more about?

Companies love to see that you are able to create meaningful data visualizations that answer real-world questions and address practical problems.

Make sure that the problem is worth exploring and can be solved in a reasonable amount of time.

**Load and Check Data**

Load your data into Python, R, or another environment. Check for missing data, consistency, and outliers. If necessary, clean it to make sure it is ready for data exploration.

**Visualize Data**

Create meaningful data visualizations that help answer the business question(s) you set out to explore. Feel free to use any of the following libraries and tools for your visualizations:

**Python:**

* Matplotlib
* seaborn
* Plotly

**Business Intelligence:**

* Tableau (BI)
* Looker
* Power BI

**R:**

* ggplot2

**Web:**

* D3.js

**Communicate Findings**

Once you’ve performed your analysis and visualizations, you’re ready to communicate your findings. This can be a written document, slide deck, or other format.

Make sure to clearly communicate:

* The data you worked with
* The question or problem you set out to solve
* Key findings and takeaways

**Explore and Explain Data**

Once you have your data, it’s a good idea to get acquainted with it. Gather some summary statistics and visually examine your data. Don’t forget to keep track of the questions you think of while exploring. It is also a good idea to start writing out some of the insights you have gain in this phase.

# Experimental Design Interview Questions

**Practice answering experimental design interview questions to solve real-world business problems**

### **Introduction**

**Welcome back! It’s me, Codey, again.**

**In this part of the interview, I’m going to ask you a few questions about setting up experiments and designing data collection to meet business needs. This is my favorite part of the interview because I love talking to candidates about real problems and hearing all of their great ideas. We have tons of data and I’m so excited to get a fresh perspective on it!**

**The best candidates are going to show me that they can think through problems, organize a research plan, and get to a data-driven solution quickly. They are also going to be able to draw connections between data and real business problems.**

**I’ve interviewed a lot of people for this position and heard a lot of answers. The best answers have a few things in common:**

* **The metrics they propose are related to the problem. For example, if I want to increase traffic to a website, I want the metrics to directly relate to page visits. If I want to reduce unnecessary medical expenses, the metrics should directly relate to cost and usage.**
* **They suggest a few metrics and focus on how they will leverage those, paying attention to how they will use the metrics rather than just giving a list of things to measure.**
* **They state their assumptions at the beginning and ask clarifying questions as needed.**

**In the very best interviews, I walk away with a new way of thinking about my data and how to design new experiments.**

**Let’s go ahead and get started!**

### **Directions**

**For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.**

### **Conceptual Questions**

**Free response**

**What is lifetime value and how can you measure it?**

**Get feedback**

**Powered by OpenAI**

**Free response**

**We’ve identified reducing churn by 10% as a goal this year. What would you do to create a model that can identify who is susceptible to churn?**

**Get feedback**

**Powered by OpenAI**

**Excellent! You seem like you have a good idea about designing experiments, let’s move on to applying these concepts.**

### **Applied Questions**

**Free response**

**e-Commerce website: We have a search feature on our site that we have had for about 5 years. We are trying to decide if we need to redesign it. How would you measure the effectiveness of this search feature?**

**Get feedback**

**Powered by OpenAI**

**A pair of yellow headphones is plugged into a computer application window. Inside the application window, there is a box with what is likely text. Next to the box there is a flow chart represented by shapes stacked vertically, connected by lines. The first level in the flow chart is a purple rectangle, the second is a green rhombus, and the third is two yellow squares.**

**Free response**

**How would you measure the performance of a song recommendation model for a music streaming platform?**

**Get feedback**

**Powered by OpenAI**

**Free response**

**A video streaming service: We recently launched a new feature that allows viewers to create playlists of movies, shows, and other content. This will allow them to pre-plan binge sessions. We have already rolled out this feature in our latest release, so it is not possible to conduct an A/B Test.**

**How would you evaluate the success of this feature? What data would you need and how would you analyze it?**

**Get feedback**

**Powered by OpenAI**

### **Review**

**Excellent job!**

**It’s clear that you have a solid sense of what data we need to measure and how to set up experiments.**

**You’ve shown your expertise in:**

* **measuring lifetime value**
* **predicting churn rates**
* **normalizing data to isolate effects**
* **measuring the impact of features after launch, and**
* **setting benchmarks to measure success.**

**Be sure to keep getting familiar with the company’s product for the next round of interviews. See you there!**

# A/B Testing

**Get practice answering A/B testing interview questions**

### **Introduction**

Hi again! It’s me, Codey. I am sure you are just as excited as I am to talk about A/B Testing. Since so much of our job is evaluating options, a deep understanding of A/B Tests is absolutely crucial.

I love when interviewees:

* Start by stating their assumptions and
* Talk through why they collect certain features or run certain tests.

It gives me a much better understanding of how they are going to collaborate with me to think through experiments going forward.

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

There are two images of hamsters. On the left, the hamster has aqua-colored liquid in a water bottle. This hamster is running on its wheel. On the right, the hamster has yellow-colored liquid in its water bottle. This hamster appears to be asleep and maybe dizzy.

Free response

What is the goal of A/B Testing, and why is it important?

**Get feedback**

Powered by OpenAI

Free response

What considerations do you need to make when determining the number of participants needed for an A/B Test?

**Get feedback**

Powered by OpenAI

Free response

If you had to stop an A/B Test early, what considerations would you make?

**Submit Response**

Fantastic! It really sounds like you understand what goes into A/B Testing, now let’s go over some applied questions to talk about how this plays out in real life.

### **Applied Questions**

Free response

Social networking app: We are considering adding a saved comments feature on our platform. When users are messaging each other, they can flag comments to be saved to a special folder. How would you determine if we should implement this new feature?

**Get feedback**

Powered by OpenAI

Free response

A news site: We have new versions of four features that we want to evaluate:

* The header,
* Signup page,
* Newsfeed layout, and
* Home button.

How would you prioritize these features for your A/B Test?

**Get feedback**

Powered by OpenAI

Free response

So we decided to run four variations of one feature at the same time, and we found that one variation won with a p-value of less than .05! Should we make the change?

**Get feedback**

Powered by OpenAI

Free response

Food delivery service: How do you manage spillover between your control and treatment groups? For example, we are launching a new promotion for food delivery and want to know which promotion is most effective at getting people to share their promo codes.

**Get feedback**

Powered by OpenAI

### **Review**

Thank you so much! It’s clear to me that you have a great sense of how to apply A/B Testing to real-world scenarios.

Just to recap, we’ve talked about:

* The goals of A/B Testing.
* How to determine how long an A/B Test should be.
* What to think about if you have to stop an experiment early.
* Working with multiple feature changes and multiple variations
* Evaluating the potential success of a feature.
* Preventing and managing spillover.

It was great talking to you and I look forward to chatting with you in the rest of the interviews! See you there.

# A/B Testing Take Home

**Practice your A/B Testing skills with a Take-Home Assessment**

### **Overview**

Welcome to the take-home assessment! Take-home assessments help interviewers get a sense of your technical skills before moving you to the next stage of the interview process. The style and format of these projects can vary widely depending on the company and role you are applying for. In this project, we are going to give you a prompt and you are free to answer it however you wish. You can use any coding language and any tools you are familiar with. This is an opportunity to show off your skills.

The projects that we have prompted here are typical for data analyst roles but are not definitive. However, there are some widespread practices and conventions that you should be aware of before starting the project.

* There may be one or more very easy questions. These are not tricks, they are meant to easily exclude applicants who cannot perform basic functions.
* There will be some discussion, analysis, or reflection. Expect to have to write a summary of your findings. A large part of both Data Analyst and Data Scientist roles is communication. How well you communicate your findings is just as important as what you find and how you find it.
* You may be asked to submit your files in a specific way. Pay attention to these instructions and submit your file in the requested format. This is another area that may initially seem unimportant but can have an outsized influence on your candidacy.
* You may *not* be asked to submit your files in any particular way. If this is the case, Python files should be .py or .ipynb, R should be in .R or .Rmd format. Similarly, your written analysis should be in .pdf format or .md if submitted online.
* If anything is unclear or you are not sure what they are asking you to do, ask! It is much better to get clarification at the beginning of a project than submit something that doesn’t answer the question. This assignment is an opportunity to showcase your problem-solving skills. If you are stuck on a particular technical skill, try to find the answer on your own. Google and StackOverflow will be your friends during this process! If you still can’t figure it out, and the interviewer has allowed for questions, reach out and explain what you’ve already tried. This will show you have the ability to problem-solve on your own before reaching out for help.
* Your interviewers may give you an estimate of how long they think it should take. It is very common to go far beyond the time estimate.
* The take-home assessment is also a good opportunity for you to decide if this is the kind of work you are interested in doing. If you enjoy it, you are probably in the right place. If it is something you do not enjoy, it may be worth clarifying what the expectations of the job are.

### **Context:**

You are applying for a role that might be situated in marketing, web development, or a business analytics department of a large company. The dataset provided is for a library homepage, but the principles are the same across institutions and industries.

### **Take Home Challenge:**

Congratulations on being invited to complete this analysis project for our Analytics team!

This assignment consists of two parts.

* In Part 1 you will read, interpret, and explain the dataset provided.
* In Part 2, you will demonstrate your data analysis skills using the dataset.

You can use the tool(s) of your choice. The key to this assignment is to clearly document your steps and state any assumptions you’ve made along the way. If you choose to use a programming language, you’ll need to provide your code. If you choose to use a spreadsheet, you’ll need to clearly document your steps. This way, our team will be able to easily follow along with your logic and fully understand your approach.

When you turn in your work we expect to receive:

* Documentation of your steps (either as a code file or process notes).
* A written summary of your findings as a .pdf or Microsoft Office file.
* Any supplemental visualizations or materials that are necessary to illustrate your findings.
* All of your files combined as one zip file.

Best of luck, we look forward to reading your results.

**Necessary Files:** The five site variations were given sequential names:

| **Experiment Name** | **Category Title** | **Url** |
| --- | --- | --- |
| Control | Interact | /index.php |
| Variation 1 | Connect | /index2.php |
| Variation 2 | Learn | /index3.php |
| Variation 3 | Help | /index4.php |
| Variation 4 | Services | /index5.php |

* Google Analytics Data for this site is linked [here](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/general-data/GoogleAnalytics-2014-07-11.zip).
* Page Views and clickthrough data collected by CrazyEgg is linked [here](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/general-data/CrazyEgg-2014-07-11.zip).
* Details about this dataset are available [here](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/general-data/AB-testing-web-analytics.pdf).

You may need to use one or both data files. This data was obtained from [Young, Scott W.H. (2014)](http://dx.doi.org/10.3998/weave.12535642.0001.101) A/B Testing Web Analytics Data. Montana State University Scholar Works. [dataset](https://scholarworks.montana.edu/xmlui/handle/1/3507)

For more on how CrazyEgg works, [see here](https://www.crazyegg.com/overview)

### **Instructions**

#### **Part 1: Exploratory Data Analysis**

* What is the median average time on a page?
* What is the average total daily page views?
* Are there any outliers?

#### **Part 2: Analysis of Testing Results**

We’ve provided a sample, open-source, dataset from a library website that is similar to the data we typically work with. Our goal is to have users follow through further into our website. In this example, we noticed that our current “Interact” page is often neglected by visitors, as you can see from this heatmap.



We want to know if changing the design of the webpage will engage them better. Our question is: “Which title will generate the most engagement: Help, Services, Learn, Connect, or Interact”

**Your task:**

* Explore the dataset.
* Create a visual representation of the click-through rate, bounce rate, and exit rate for all five variations of the A/B test.
* Write a short paragraph explaining why you chose the visualizations you did, how you arrived at your metrics, and be sure to include what variation you would recommend to our team based on your findings.

### **Checklist:**

Here is a checklist of things to consider as you work on this project:

* Make sure your code is clean, well-documented, and that your steps are clearly laid out.
* Assessment questions in Part 1 only have one correct answer. Make sure to double-check your math and write 1-2 sentences interpreting the values.
* There is an academic paper associated with this dataset asking the same questions in Part 2. Your results may be similar. The process you take to get your answer and the way you communicate your results are most important in a take-home challenge.
* Make sure to provide a clear, data-driven recommendation.
* In your presentation, you can include the heatmaps and screenshots as supporting evidence. If it is in the data files, you can use it in your analysis.

**Note**: a take-home won’t typically include a checklist like this one, but we’ve included one here for you as you become more familiar with the take-home.

Good luck and have fun!

# Review of Visualization, Communication, and Experimental Design Interviews

**See all you’ve covered in Visualization, Communication, and Experimental Design**

### **Review**

Congratulations! The goal of this unit was to get practice answering questions about data visualization, communication, and experimental design.

Having completed this unit, you are now able to:

* Discuss how you pick a data visualization tool.
* Use data visualization to answer a practical business question.
* Describe your approach to experimental design.
* Talk your interviewer through how you conduct A/B tests.
* Evaluate data from a real A/B test to make a recommendation about website layout.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# Introduction to Data Analyst Interview Coding Questions

**See what coding questions you’ll cover.**

### **Goals of this Unit**

The goal of this unit is to prepare for coding questions in your interview. Since most data analyst roles require knowledge of SQL and familiarity with a scripting language (such as Python or R), this unit focuses on only what you will need for the interview. The focus is on SQL and general whiteboarding. Whiteboarding is a practice that is applicable to a wide variety of coding challenges.

After this unit, you will be able to:

* Approach whiteboarding and live coding challenges in an organized and systematic way.
* Communicate your coding decisions to your interviewers.
* Implement SQL to answer common data analyst interview questions.
* Describe your thought process when performing database queries.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

### **Review**

<1 min

Technical interviews judge the candidate’s ability to perform technical aspects of the job. Whiteboarding is a type of technical interview which requires the candidate to code a solution without the assistance of a text editor.

Coding an efficient solution is important, but the ultimate goal as an interviewee is to convince the interviewer you would be a desirable co-worker.

A desirable co-worker is able to articulate their thought process and collaborate without friction.

Our whiteboarding strategies revolve around communication at every step.

The steps are:

* **Clarify** the Problem
* Create **Inputs**
* **Outline** the Solution
* **Code** the Solution
* **Test** the Solution
* **Analyze** the Solution

**Instructions**

Practice makes perfect!  
Run through these steps with a new practice problem:  
Write a function which prints every number from 0 up to the given input. If divisible by 3, print “Fizz” instead of the number. If divisible by 5, print “Buzz”. If input is divisible by 3 AND 5, print “FizzBuzz”.

# Common SQL Interview Questions

**Practice responding to common SQL interview questions.**

Welcome back!

In this part of the interview, we are going to cover some common questions that come up around databases–specifically SQL. Though some positions don’t require SQL knowledge and others require extensive knowledge, we expect our team members to have a solid understanding of SQL even though they don’t have to be experts.

These questions are a bit more direct than the other sections because we assume that if you know these basic concepts, you have the SQL skills we are looking for.

So, without further ado, let’s jump in to talking about SQL and all the powerful things it can do!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Questions**

Free response

What is SQL?

**Get feedback**

Powered by OpenAI

Free response

What is a database? And what is a relational database?

**Get feedback**

Powered by OpenAI

Free response

What is a RDBMS?

**Get feedback**

Powered by OpenAI

Free response

What is a table? And what are a row and column within a table?

**Get feedback**

Powered by OpenAI

Free response

What is a data type?

**Get feedback**

Powered by OpenAI

Free response

What is a primary key and a foreign key?

**Get feedback**

Powered by OpenAI

Free response

What is the difference between ALTER and UPDATE?

**Get feedback**

Powered by OpenAI

Free response

What is a query and what is a subquery?

**Get feedback**

Powered by OpenAI

Free response

What are constraints?

**Get feedback**

Powered by OpenAI

Free response

What is a statement?

**Get feedback**

Powered by OpenAI

Free response

How do you check if a field has a value or not?

**Get feedback**

Powered by OpenAI

Free response

What is the difference between DISTINCT and UNIQUE?

**Get feedback**

Powered by OpenAI

Free response

What is a join?

What is the difference between an INNER JOIN and LEFT JOIN?

**Get feedback**

Powered by OpenAI

Free response

What is the purpose of windows functions?

**Get feedback**

Powered by OpenAI

Free response

What are indexes and why are they needed?

**Get feedback**

Powered by OpenAI

Great job! It is clear that you know your way around a database and are ready to talk about them in an interview.

### **Review**

In this interview, we covered basic SQL concepts and practices, including:

* The basic structure of a database.
* How data are represented in a database.
* Some simple functions for extracting data from a database.
* Best practices for keeping a database well structured.

Looking forward to seeing you again in the next interview!

**Interview Prep Problems**

### **Code Challenge 14**

2 min

The purchases table has the following columns:

* purchase\_id - the id of the purchase
* purchase\_date - the date of the purchase

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Using string formatting and substitutions, get the month and day for each purchase in the form ‘mm-dd’.  
  Give this new column a name of ‘reformatted’.
* SELECT STRFTIME('%m-%d', purchase\_date) AS 'reformatted'
* FROM purchases;

SELECT STRFTIME('%m-%d', purchase\_date) AS 'reformatted'

FROM purchases;

# Review of Data Scientist Interview Coding Questions

**See what coding questions you’ve prepared for.**

### **Review**

Congratulations! The goal of this unit was to prepare you for coding questions in your interview, with a focus on SQL.

Having completed this unit, you are now able to:

* Approach whiteboarding and live coding challenges in an organized and systematic way.
* Communicate your coding decisions to your interviewers.
* Implement SQL to answer common data analyst interview questions.
* Describe your thought process when performing database queries.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# Introduction to Machine Learning and Algorithms

**Review what you covered in Machine Learning and Algorithms.**

### **Review**

The goal of this unit is to get familiarity with both open-ended machine learning and algorithms questions and put some of your solutions into code. The purpose of this module is to prepare you specifically for the Machine Learning section of the interview. If you are applying to an engineering-heavy role, be sure to also take the [Pass the Technical Interview with Python Skill Path](https://www.codecademy.com/learn/paths/pass-the-technical-interview-with-python). If you don’t know if you’ll need that or not, be sure to go back to the beginning of this Path to revisit the section on how to read a job advertisement.

After completing this unit, you will be able to:

* Answer questions about supervised and unsupervised classification methods.
* Implement supervised and unsupervised algorithms on real data to answer business questions.
* Discuss evaluation of classification and regression tasks.
* Talk your interviewer through how you improve upon machine learning models.
* Evaluate predictions from supervised and unsupervised machine learning models.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# The Relationship Between Data and Machine Learning

**Get ready for interview questions on the relationship between training data and machine learning**

### **Introduction**

Hi again, Welcome back! Today we are going to go over some high-level data science interview questions. For this role, we are hoping for someone who is well versed in a variety of machine learning algorithms. We have so many different data sources and are looking for a data scientist who can help us answer the long list of business questions we have.

As with any candidate, I am looking for someone who is a problem solver and can clearly communicate their thinking and knowledge. Someone who is curious about our data and has ideas on how to solve our business problems.

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

Free response

What is the bias/variance trade-off?

**Get feedback**

Powered by OpenAI

Free response

What are the different types of machine learning, can you give an example of each?

**Get feedback**

Powered by OpenAI

Free response

What is your favorite algorithm and can you explain to me how it works?

**Submit Response**

Free response

What is the difference between neural networks and other types of ML algorithms?

**Get feedback**

Powered by OpenAI

Free response

How do you handle imbalanced training sets when training a model?

**Get feedback**

Powered by OpenAI

image of person looking at the stars with a telescope and drawing a line chart between them.

Free response

Where do you usually get datasets from for your own projects?

**Submit Response**

Free response

What is feature engineering and why is it important?

**Get feedback**

Powered by OpenAI

Free response

What is feature selection and why is it important?

**Get feedback**

Powered by OpenAI

Great job! It is clear to me that you have a really firm understanding of how data relates to models and are ready to talk more specifically about each one!

### **Review**

We have covered questions about:

* The bias/variance trade-off and how this needs to be addressed when building models that will perform well with unseen data.
* A few different types of machine learning models, examples of each, and when to use which model.
* The difference between deep learning and ML, and demonstrated awareness of debates in the field.
* How to handle imbalanced classes and maximize the available data.
* The importance of both feature selection and feature engineering and how they affect our models.
* And where you find datasets.

# Common Regression Questions

**Practice answering common interview questions about regression.**

### **Introduction**

Welcome back! In this part of the interview, we are going to talk about linear regression. For some problems (and therefore at some companies), linear regression is very popular. Whenever you have a bunch of predictor variables and an output, multiple linear regression is likely your best solution.

It is therefore essential that Data Scientists are familiar with it and understand when to apply it, how to improve it, and what assumptions are being made about the data to begin with.

So, without further ado, let’s get started. Just remember that like all the other interviews, explaining your thought process and contextualizing your answer is the key to standing out in your interview.

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Questions**

Free response

What is linear regression?

**Get feedback**

Powered by OpenAI

Free response

What are the assumptions of OLS linear regression?

**Get feedback**

Powered by OpenAI

Free response

What happens when OLS assumptions are violated?

**Get feedback**

Powered by OpenAI

Free response

How do you evaluate a regression model?

**Get feedback**

Powered by OpenAI

scatter plot with lots of different colors and a line of best fit drawn through them

Free response

What is the difference between R^2 and adjusted R^2?

**Get feedback**

Powered by OpenAI

Free response

What are some different types of regression models?

**Get feedback**

Powered by OpenAI

Great job working through the questions about regression! It is clear that you understand linear regression and predictive modeling very well.

### **Review**

In this interview, we covered:

* What linear regression is,
* The assumptions of linear regression and what happens when they are violated,
* How to evaluate linear regression models
* The difference between R^2 and adjusted-R^2 and when to use each,
* Other regression models.

Congratulations! You are ready to show off those skills in your next interview.

### **Introduction - Common Regression Questions**

<1 min

Welcome back!

In this part of the interview, we will cover how to solve a classification problem with linear regression in Python. We are going to ask some questions about linear regression and ask you to write your response as a multi-line string in the console. Typically the coding questions would be asked as part of the take-home challenge or during a technical interview, but we are going to pull them out here to give you a chance to practice implementing code for linear regression and walking through a problem.

There is no code evaluation for these exercises because there are many ways to implement these solutions and it is important that you use your own. However, if you select ‘View Solution’ you can see our sample responses.

Let’s get started!

**Common Regression Questions**

### **Inspect Dataset**

3 min

In this exercise, you will get familiar with a dataset about wine ratings. This dataset includes both red and white wine, though you have only been given qualitative data about it. You do not have the names or grapes. There is a lot you can do with this dataset, but in this part of the interview, we are going to focus on linear regression.

Start by getting acquainted with the **wine\_ratings.csv** dataset.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Start by importing the dataset that we’ve loaded into the workspace, **wine\_ratings.csv**. Also import any packages you think you might need. We’ve imported the Codecademylib3 library to help display tables in a more readable format.
* Be sure to lead your data as a pandas DataFrame.
* Checkpoint 2 Passed
* **2.**
* Inspect the variables, be sure to describe the dataset, and look at how many unique values each variable has.
* Pandas has a helpful number of unique values function. Check it out in the Pandas [documentation](https://pandas.pydata.org/docs/).
* Checkpoint 3 Enabled
* **3.**
* Describe what you will need to do to transform the variables and make sure that they will be usable in your model. Write your plan as a multi-line string in the text editor.

import codecademylib3

import pandas as pd

import numpy as np

df = pd.read\_csv("wine\_ratings.csv")

print(df.head())

print(df.describe())

print(df.columns)

print(df.dtypes)

print(df.nunique())

print(df.corr())

"""This is a relatively clean dataset! It looks like all of the variables are continuous and formatted as floats except for category which is a string. There is a unique identifier column that I will be sure not to include in the model. """

**Common Regression Questions**

### **Single Linear Regression**

8 min

Let’s put this dataset to work trying to create a model to predict the rating of the wine.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Start by creating a linear regression model with one continuous variable. In our example, we will use alcohol content, though you are free to use whatever variable interests you.
* Checkpoint 2 Passed
* **2.**
* Now fit that model to the existing data. Be sure to plot your prediction versus the actual values.
* The [Scikit-learn documentation](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html) and [matplotlib documentation](https://matplotlib.org/stable/contents.html) should help you get started.
* Checkpoint 3 Passed
* **3.**
* Now calculate the R^2 value of the model. Is this a reliable model? Why or why not? Write your answer as a multi-line string in the text editor.
* Remember that the [scikit-learn package](https://scikit-learn.org/stable/modules/model_evaluation.html) has a .score() method that can help you.
* Checkpoint 4 Passed
* **4.**
* Calculate the residuals and plot them to ensure that the residuals are normally distributed and our assumptions have been satisfied.
* We will use a qq plot, though you could also conduct a test if you prefer.
* Checkpoint 5 Passed
* **5.**
* Decide if you need to transform any variables. Why or why not? Does this model fulfill assumptions of heteroscedasticity and normality? Write your answer as a multi-line string in the text editor.

import codecademylib3

import pandas as pd

import numpy as np

import scipy.stats as stats

import statsmodels.api as sm

import matplotlib.pyplot as plt

from sklearn.linear\_model import LinearRegression

df = pd.read\_csv("wine\_ratings.csv")

#reshape the predictor (alcohol) and target (rating) values.

alcohol = df['alcohol'].values.reshape(-1,1)

rating = df['rating'].values.reshape(-1,1)

#create a linear regression model fitted to the alcohol feature.

lr\_alcohol = LinearRegression()

lr\_alcohol.fit(alcohol, rating)

#predict the ratings based on the alcohol model

rating\_predict\_alcohol = lr\_alcohol.predict(alcohol)

df['alcohol\_predict'] = rating\_predict\_alcohol

#plot the alcohol versus the real ratings and the alcohol versus the predicted ratings on the same plot

plt.plot(alcohol, rating, 'o')

plt.plot(alcohol, rating\_predict\_alcohol)

plt.show()

#calculate the R^2 score

alcohol\_r2 = lr\_alcohol.score(alcohol, rating)

print(alcohol\_r2)

"""The R^2 value is very low, suggesting that the model does not fit the data well. Since R^2 is on a scale of 0 to 1 and this model is less than .2, it is not a good fit for the existing data. That is, the difference between the real and predicted values are very high."""

#calculate the alcohol residuals

df['alcohol\_residuals'] = abs(df['rating']) - abs(df['alcohol\_predict'])

#make a qq plot of the residuals to test normalcy

fig = sm.qqplot(df['alcohol\_residuals'] , stats.t, fit=True, line="45")

plt.show()

"""The residuals are close to normally distributed, which fulfills our assumptions about normality. Likewise, the data seems linear even though the model is not a particularly good fit. So I will not transform the alcohol variable."""

### **Multiple Linear Regression**

6 min

Often you will want to use all of our features to create a model, not just one. So, in this part of the interview, you can work with the whole dataset to create a regression model.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Start by scaling the variables. You know that you’re going to want to determine which variable contributes the most to the model, so you will need to scale your variables to be able to determine that.
* Use the [scikit-learn Standard Scaler](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html).
* Checkpoint 2 Passed
* **2.**
* Now create a linear regression model using all of the continuous features in the dataset. You will need to evaluate this model after you’ve built it, so be sure to split the data into training and testing sets.
* [Scikit-learn Linear Regression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html) applies to both single and multiple regression.
* Checkpoint 3 Passed
* **3.**
* Once you’ve established the model, fit it to the training set.
* This is a simple fit function.
* Checkpoint 4 Passed
* **4.**
* Create a scatterplot of the predicted versus actual values.
* This will help show how far from the actual values your predictions are.
* Checkpoint 5 Passed
* **5.**
* Print out the regression coefficients and the R^2 value. What is the biggest predictor variable? Are any variables good predictors of overall rating? How well does this model fit the data? Write your answer as a multi-line string in the text editor.
* Checkpoint 6 Passed
* **6.**
* Could we improve the model by removing highly correlated features or normalizing the data?  
  Conduct any tests you feel are appropriate to search for highly correlated variables to determine if we can improve the model by removing any. Explain your findings. Write your answer as a multi-line string in the text editor.

import codecademylib3

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import r2\_score

df = pd.read\_csv("wine\_ratings.csv")

print(df.head())

#separate data from labels

data = df.drop(['category','uuid', 'rating'], axis=1)

label = df['rating']

#scale data

scaler = StandardScaler()

data\_scaled = scaler.fit\_transform(data)

#split into training and testing

x\_train, x\_test, y\_train, y\_test = train\_test\_split(data\_scaled, label, train\_size=0.8, test\_size=0.2)

# fit a multiple linear regression model

mlr = LinearRegression()

mlr.fit(x\_train, y\_train)

#make predictions based on that model

y\_predict = mlr.predict(x\_test)

#create scatter plot of predictions vs. actual ratings

plt.scatter(y\_test, y\_predict, alpha=0.4)

plt.show()

#print the coefficients and max coefficient

mlr\_coef = mlr.coef\_

print(mlr\_coef)

#print the maximum coefficient

print("The max coefficient is: ",np.max(mlr\_coef))

#find the index of the maximum coefficient

max\_coef\_index = np.argmax(mlr\_coef)

#make a list of the columns in the dataset

column\_name\_list = data.columns.tolist()

#use the index of the max coefficient to find the feature with the maximum coefficient

print("The feature with the max coefficient is: ", column\_name\_list[max\_coef\_index])

print("The R^2 score is: ",r2\_score(y\_test, y\_predict))

"""None of the predictor variables are very good predictors. Alcohol contributes the most to the model, but it is still a very weak predictor. The R^2 value shows that the model does not fit the data very closely as it is very low (less than .3)"""

df\_corr = df.corr()

print(df\_corr)

#that's hard to read, so use the unstack and sort method

df\_corr\_list = df\_corr.unstack()

df\_corr\_rank = df\_corr\_list.sort\_values(ascending = False)

#print starting at 13 to exclude self-correlations

print(df\_corr\_rank[13:30])

"""The sulfur dioxide variables are highly correlated. We can remove one of those to see if the model will become more robust. Density and residual sugar also seem to be highly correlated, but not strongly enough to warrant removing them."""

### **Improved Multiple Linear Regression**

1 min

Identifying ways to improve your model is only part of the challenge. In this part of the interview, we would like you to improve upon your features in order to improve upon your model.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Since you will be rebuilding the model you just built, start by either rewriting or copying your code into the new workspace.
* There’s a ‘Copy’ button in the text editor.
* Checkpoint 2 Passed
* **2.**
* Start by transforming and selecting your variables as necessary. You can use any method you prefer.  
  Explain why you did what you did in a multi-line string in the text editor.
* There is at least one categorical variable that needs to be changed.
* Checkpoint 3 Passed
* **3.**
* Create a new model with the selected features, evaluate it, and determine whether the new model or the old model performs better.  
  Write your response as a multi-line string in the text editor.
* Feature selection doesn’t always improve the model. Why not?

import codecademylib3

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import r2\_score

df = pd.read\_csv("wine\_ratings.csv")

print(df.head())

#change category to 'red'

df['category'].replace(['red','white'],[1,0], inplace=True)

df.rename({'category':'red'}, axis=1, inplace=True)

df\_selected = df.drop(['uuid', 'total sulfur dioxide'], axis=1)

"""I changed 'category' to a binary variable to be able to use it in the model. I also dropped the uuid and total sulfur dioxide because it was highly correlated with free sulfur dioxide"""

data = df\_selected.drop(['rating'], axis=1)

label = df\_selected['rating']

#scale data

scaler = StandardScaler()

data\_scaled = scaler.fit\_transform(data)

x\_train, x\_test, y\_train, y\_test = train\_test\_split(data\_scaled, label, train\_size=0.8, test\_size=0.2)

mlr = LinearRegression()

mlr.fit(x\_train, y\_train)

y\_predict = mlr.predict(x\_test)

plt.scatter(y\_test, y\_predict, alpha=0.4)

plt.show()

print(r2\_score(y\_test, y\_predict))

"""Removing the total sulfur dioxide and adding the wine type variable actually degraded the model. I would like to try to make 2 more versions: one with just the category removed and one with just the total free sulfide removed. It appears that the total free sulfide may have contributed slightly to the model. However, no matter what, the model does not fit the data very well and is not a good predictor of the rating. """

**Common Regression Questions**

### **Review Common Regression Questions**

<1 min

Congratulations! You are ready to answer the most common questions about linear regression in your interview. Ultimately Data Science is an iterative process, so though we only went through this model one time, you will be expected to iterate many more times in order to improve the model. This could include using Regularization techniques, further feature selection, or other techniques. Being able to speak to this process in your interview and showing how you reached your answers in your take-home assessment will demonstrate your skill in fitting Linear Regression models to your data. Whenever possible, document your thinking and the steps you took to reach your result.

In this lesson, we covered:

* Inspecting data,
* Building a single linear regression model,
* Building a multiple linear regression model,
* Improving upon a multiple linear regression model.

# Regression Take-Home Assessment

**Practice your regression and predictive machine learning skills with a sample take-home assessment.**

### **Overview:**

Welcome to the take-home assessment!

Take-home assessments help interviewers get a sense of your technical skills before moving you to the next stage of the interview process. The style and format of these projects can vary widely depending on the company and role you are applying for. In this project, we are going to give you a prompt and you are free to answer it however you wish. You can use any coding language and any tools you are familiar with. This is an opportunity to show off your skills.

The projects that we have prompted here are typical for data scientist roles, but are not definitive. However, there are some widespread practices and conventions that you should be aware of before starting the project.

* There may be one or more very easy question(s). These are not tricks, they are meant to easily exclude applicants who cannot perform basic functions.
* There will be some discussion, analysis, or reflection. Expect to have to write a summary of your findings. A large part of Data Scientist roles is communication. How well you communicate your findings is just as important as what you find and how you find it.
* You may be asked to submit your files in a specific way. Pay attention to these instructions and submit your file in the requested format. This is another area that may initially seem unimportant but can have an outsized influence on your candidacy.
* You may *not* be asked to submit your files in any particular way. If this is the case, Python files should be .py or .ipynb, R should be in .R or .Rmd format. Similarly, your written analysis should be in .pdf format or .md if submitted online.
* If anything is unclear or you are not sure what they are asking you to do, ask! It is much better to get clarification at the beginning of a project than submit something that doesn’t answer the question. This assignment is an opportunity to showcase your problem-solving skills. If you are stuck on a particular technical skill, try to find the answer on your own. Google and StackOverflow will be your friends during this process! If you still can’t figure it out, and the interviewer has allowed for questions, reach out and explain what you’ve already tried. This will show you have the ability to problem-solve on your own before reaching out for help.
* Your interviewers may give you an estimate of how long they think it should take. It is very common to go far beyond the time estimate.
* The take-home assessment is also a good opportunity for you to decide if this is the kind of work you are interested in doing. If you enjoy it, you are probably in the right place. If it is something you do not enjoy, it may be worth clarifying what the expectations of the job are.

### **Context:**

You are applying for a role at an online news organization in a very small data science team.

#### **Take Home Assessment:**

Congratulations on being invited to complete this analysis project for our Data Science team!

You can use the tool(s) of your choice. The key to this assignment is to clearly document your steps and state any assumptions you’ve made along the way. Please submit your code as well as a summary of your findings.

When you turn in your work we expect to receive:

* Documentation of your steps (either as a code file or process notes).
* A written summary of your findings as a .pdf or Microsoft Office file.
* Any supplemental visualizations or materials that are necessary to illustrate your findings.
* All of your files combined as one zip file.

Best of luck, we look forward to reading your results. Dataset: Buza, K. (2014). [Feedback Prediction for Blogs](https://archive.ics.uci.edu/ml/datasets/BlogFeedback#). In Data Analysis, Machine Learning and Knowledge Discovery (pp. 145-152). Springer International Publishing.

[The Data Package](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/general-data/blogData.zip)

In this dataset, there is training and testing data. The testing data is presented daily for March 2012. Each file represents one day. We recommend using each file as a separate test or validation set.

The features refer to “basetime.” We choose a basetime (in the past) and selected the blog posts that were published at most 72 hours before the selected base date/time. The actual time that the basetime is does not matter, and it is not the same for all of the blogs. You can think of the basetime as the time the crawler found the blog. However, all of the times are with respect to the selected basetime, i.e., 24 or 48 hours before/after the basetime.

Each instance in the dataset corresponds to one blog post.

### **Instructions**

Please answer the following questions:

* What is the median ratio of comments in the first 24 hours to the second 24 hours after the blog was posted? (be sure to explain your answer)
* What feature(s) is the strongest predictor of comments?
* Please create a model that will predict the number of comments a blog post will generate.

We would like to see all of your code, including any EDA or feature engineering as well as the end performance of your model.

### **Checklist:**

Here is a checklist of things to consider as you work on this project:

* Make sure your code is clean, well-documented, and that your steps are clearly laid out.
* The assessment question in question 1 only has one correct answer. Make sure to double-check that there aren’t any nulls or errors and the data and that you did the math correctly. If there is a specified format, be sure to submit it in that format.
* There is an academic paper associated with this dataset. Your results may be similar. The process you take to get your answer and the way you communicate your results are most important in a take-home assessment.
* Make sure to provide a clear, data-driven analysis in a written or presentation format.

**Note**: a take-home won’t typically include a checklist like this one, but we’ve included one here for you as you become more familiar with the take-home.

Good luck and have fun!

# Common Classification Questions

**Practice answer common interview questions about classification.**

### **Introduction**

Welcome! In this section of the interview, we will cover Classification. Classification is obviously central to a Data Scientist’s work, and I want to be sure that our next hire is fluent in classification techniques.

Let’s dive into it.

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Questions**

Free response

What is the difference between classification and regression?

**Get feedback**

Powered by OpenAI

Free response

How do you evaluate a classification model?

**Get feedback**

Powered by OpenAI

Free response

What is logistic regression?

**Get feedback**

Powered by OpenAI

Free response

How do you interpret coefficients from a logistic regression model?

**Get feedback**

Powered by OpenAI

Free response

What are some different types of classification models? When would you use certain models?

**Get feedback**

Powered by OpenAI

image of tree with big green branches and a thick brown trunk with the boxes of a decision tree model overlaid

Free response

Describe a decision tree classifier.

**Get feedback**

Powered by OpenAI

#### **Review**

We have covered:

* Classification and Regression
* Evaluating a Classification Model
* Logistic Regression and Coefficients
* Different types of Classification Models
* Decision Tree Classifier

Great job! You are on your way to being a classification expert – you should be ready to tackle most classification problems in a real interview. Good luck!

### **Inspect Dataset**

2 min

In this exercise, we are going to get acquainted with the Telco dataset. At this point, you can just explore the dataset and make a plan about how to work with it. Taking a little extra time to get your bearings and think about the data will pay off when you have to answer questions about the data later on.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Start by importing the dataset that we’ve loaded into the workspace, **telco\_churn.csv**. Also import any packages you think you might need. We’ve imported the Codecademylib3 library to help display tables in a more readable format.
* Be sure to lead your data as a pandas DataFrame.
* Checkpoint 2 Passed
* **2.**
* Inspect the variables, be sure to describe the dataset, and look at how many unique values each variable has.
* Pandas has a helpful number of unique values function. Check it out in the [Pandas documentation](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.nunique.html).
* Checkpoint 3 Passed
* **3.**
* Describe what you will need to do to transform the variables and make sure that they will be usable in your model. Write your plan as a multi-line string in the console.

import codecademylib3

import pandas as pd

import numpy as np

df = pd.read\_csv("telco\_churn.csv")

print(df.describe())

print(df.columns)

print(df.dtypes)

print(df.nunique())

print(df.corr())

"""I will need to transform all the binary variables from yes/no into 1/0. I will remove the 'No phone service' and 'No internet service' values and make them into 'No' so that I can treat those columns as binary variables rather than multinomial. If my model does not perform well, I may revert back to multinomial. Since there are only 2 genders listed, I will replace Male/Female with 1/0 and name the column 'Male'."""

### **Prepare Data for the Model**

1 min

After inspecting the data, you will need to transform it and prepare it for your model. The basic thing to keep in mind when practicing code while preparing for an interview is to document your steps and your thinking. Well-documented code demonstrates that you can communicate about your thought process. So, for each checkpoint, be sure to document why you are performing that step.

Again, we’ve included the codecademylib3 library in case you need to make any visualizations.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Prepare data for model – this could include handling missing values, one-hot-encoding of categorical features, and other transforms/standardization.
* Be sure to replace values where needed and use one-hot encoding.

### **Build a Classification Model**

2 min

Now that your data is prepared, it is time to build and evaluate a classification model. The follow up questions will be about Logistic Regression models, so please implement a Logistic Regression model here. You can write it from scratch or use the scikit-learn library’s function. You will need to copy and paste your code from the last exercise into this exercise to save your transformations.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Separate your data into predictor features and labels. Create at least 3 different variations of predictors and labels. For example, you might want to create a model with all of the predictor features, and another with only demographic features, and another with only service-related features. However you choose to separate the groups, be sure to label what they are.
* Revisit the Data Manipulation with Pandas track in the Data Scientist Career Path for a refresher on segmenting datasets or visit the [pandas documentation](https://pandas.pydata.org/).
* Checkpoint 2 Passed
* **2.**
* Time to build your Logistic Regression! We recommend writing a function that will take predictors and labels as input and return the model and results as output.  
  You can do this from scratch like we did in the Data Scientist Career Path or you can use scikit-learn’s methods.
* Check out the scikit-learn documentation [here](https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression).
* Checkpoint 3 Passed
* **3.**
* Be sure to include evaluation metrics for your model (and create a train/test split in the data). The most common evaluation metrics are Precision, Recall, and F1, but you may also want to include a confusion matrix to better understand what your model is getting right and wrong.
* [Scikit-learn](https://scikit-learn.org/stable/) has functions to create a train/test split automatically as well as evaluate your models.
* Checkpoint 4 Passed
* **4.**
* Run your model on the datasets that you created.
* Checkpoint 5 Passed
* **5.**
* Compare the results of each model. Which dataset is the best predictor of churn? Why? Write your answer as a multi-line string in the console.

### **Interpret Logistic Regression Coefficients**

3 min

Great! Now that you have built a model, it is time to interpret the coefficients. This will help us identify which features are the strongest predictors of churn. This is where Logistic Regression shines in a business sense. Once we have weights for each coefficient, we can make business decisions around those values.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* First, copy the code that you wrote in the last exercise. We need to retrieve the features we’ve created as well as the training and testing split.
* You actually don’t need the model and evaluate function, but you do need the feature engineering and feature segmentation work you’ve already done.
* Checkpoint 2 Passed
* **2.**
* Extract the coefficients from the model. Remember that the values that will be returned from the .coef\_ function are the log odds ratio, so remember to transform the scale to get the odds ratio.  
  We recommend filling a DataFrame with the coefficients for easy comparison.
* Remember to fit the training data you used previously to the model you will use for evaluation.
* Checkpoint 3 Passed
* **3.**
* Interpret the coefficients. What feature is the strongest predictor of churn in the best model? What about other features or other models? What does that tell us?  
  Be sure to explain how you are interpreting these values - given that they are the odds ratio.
* Sorting your DataFrame will help you spot the biggest contributor. If you need a refresher on what the coefficients mean, be sure to return to the Logistic Regression lesson of the Data Scientist Career Path.

import pandas as pd

import numpy as np

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import precision\_recall\_fscore\_support

df = pd.read\_csv("telco\_churn.csv")

# replace no internet/phone service with no

df.replace(['No internet service', 'No phone service'], 'No', inplace=True)

# replace yes and no with 1 and 0

df.replace(['YES', 'Yes', 'NO', 'No'],[1,1,0,0], inplace=True)

# since only 2 genders listed, make 'gender' into a binary column for female

df['gender'].replace(['Male', 'Female'], [1,0], inplace=True)

df.rename({'gender':'female'}, axis=1)

# one hot encoding for categorical variables

# drop\_first=True removes redundant values

temp = pd.get\_dummies(df[['InternetService', 'Contract', 'PaymentMethod']], drop\_first=True)

# rejoin modified datasets

df1 = pd.concat([df, temp.reindex(df.index)], axis=1)

#drop columns that are now redundant

df1.drop(columns = [ 'InternetService', 'Contract', 'PaymentMethod'], inplace=True)

def model\_and\_eval(my\_training, my\_testing):

X\_train, X\_test, y\_train, y\_test = train\_test\_split(my\_training, my\_testing, test\_size=0.2, random\_state=42)

# evaluation

clf = LogisticRegression(random\_state=0).fit(X\_train, y\_train)

y\_pred = clf.predict(X\_test)

print(confusion\_matrix(y\_test, y\_pred))

tn, fp, fn, tp = confusion\_matrix(y\_test, y\_pred).ravel()

print("tn:", tn, "fp:", fp, "fn:", fn, "tp:", tp)

print(precision\_recall\_fscore\_support(y\_test, y\_pred, average='weighted'))

# separate labels

df\_labels = df1['Churn']

# make a dataset of everything

df\_all\_training = df1.drop(['Churn', 'customerID'], axis=1)

# make a dataset of just the user demographics

df\_demographics\_training = df1[['gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure']]

#make a dataset of just services

df\_service\_training = df1[['PhoneService', 'MultipleLines', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaperlessBilling', 'MonthlyCharges', 'InternetService\_DSL', 'InternetService\_Fiber optic']]

model\_and\_eval(df\_all\_training, df\_labels)

model\_and\_eval(df\_demographics\_training, df\_labels)

model\_and\_eval(df\_service\_training, df\_labels)

def interp\_coef(my\_training, my\_testing):

logreg = LogisticRegression()

X\_train, X\_test, y\_train, y\_test = train\_test\_split(my\_training, my\_testing, test\_size=0.2, random\_state=42)

logreg.fit(X\_train, y\_train)

my\_coef = np.exp(logreg.coef\_[0])

coef\_df = pd.DataFrame(my\_coef,

X\_test.columns,

columns=['coefficient']).sort\_values(by='coefficient', ascending=False)

print(coef\_df)

interp\_coef(df\_all\_training, df\_labels)

interp\_coef(df\_demographics\_training, df\_labels)

interp\_coef(df\_service\_training, df\_labels)

"""Since we are looking at the odds ratio, I was careful to interpret these values with respect to each other and not as linearly scaled proportions. In the model that uses all of the available features as predictors, whether someone is on a month-to-month contract has the biggest effect on whether someone will terminate service or not. It also seems that whether they are a fiber optic subscriber plays a large role. I would want to look into this more to determine if there are very few fiber optic subscribers, but it may be that people are unhappy with the service."""

### **Review**

<1 min

Congratulations! You have answered some tough questions about classification. There are, of course, more ways to improve our model (i.e., Regularization, Tuning, and Decision Boundaries) that may come up in more advanced interviews. But, by completing this lesson, you have demonstrated that you are ready to:

* Select features to use in a model
* Transform features to prepare them for use in a Logistic Regression model
* Build a Logistic Regression model
* Evaluate a Logistic Regression model
* Decide which model performs the best
* Interpret the coefficients of a Logistic Regression model.

You are ready to show off your classification skills in your next data science interview!

# Classification Take-Home Assessment

**Practice completing take-home assessments with a focus on classification.**

### **Overview:**

Welcome to the take-home assignment! Take-home assessments help interviewers get a sense of your technical skills before moving you to the next stage of the interview process. The style and format of these projects can vary widely depending on the company and role you are applying for. In this project, we are going to give you a prompt and you are free to answer it however you wish. You can use any coding language and any tools you are familiar with. This is an opportunity to show off your skills.

The projects that we have prompted here are typical for data analyst roles, but are not definitive. However, there are some widespread practices and conventions that you should be aware of before starting the project.

* There may be one or more very easy question(s). These are not tricks, they are meant to easily exclude applicants who cannot perform basic functions.
* There will be some discussion, analysis, or reflection. Expect to have to write a summary of your findings. A large part of both Data Analyst and Data Scientist roles is communication. How well you communicate your findings is just as important as what you find and how you find it.
* You may be asked to submit your files in a specific way. Pay attention to these instructions and submit your file in the requested format. This is another area that may initially seem unimportant but can have an outsized influence on your candidacy.
* You may *not* be asked to submit your files in any particular way. If this is the case, Python files should be .py or .ipynb, R should be in .R or .Rmd format. Similarly, your written analysis should be in .pdf format or .md if submitted online.
* If anything is unclear or you are not sure what they are asking you to do, ask! It is much better to get clarification at the beginning of a project than submit something that doesn’t answer the question. This assignment is an opportunity to showcase your problem-solving skills. If you are stuck on a particular technical skill, try to find the answer on your own. Google and StackOverflow will be your friends during this process! If you still can’t figure it out, and the interviewer has allowed for questions, reach out and explain what you’ve already tried. This will show you have the ability to problem-solve on your own before reaching out for help.
* Your interviewers may give you an estimate of how long they think it should take. It is very common to go far beyond the time estimate.
* The take-home assessment is also a good opportunity for you to decide if this is the kind of work you are interested in doing. If you enjoy it, you are probably in the right place. If it is something you do not enjoy, it may be worth clarifying what the expectations of the job are.

### **Context**

You are applying for a role at a medical center on the data science team.

#### **Take Home Challenge**

Congratulations on being invited to complete this analysis project for our Data Science team!

You can use the tool(s) of your choice. The key to this assignment is to clearly document your steps and state any assumptions you’ve made along the way. Please submit your code as well as a summary of your findings.

When you turn in your work we expect to receive:

* Documentation of your steps (either as a code file or process notes).
* A written summary of your findings as a .pdf or Microsoft Office file.
* Any supplemental visualizations or materials that are necessary to illustrate your findings.
* All of your files combined as one zip file.

Best of luck, we look forward to reading your results.

#### **Dataset**

[Download dataset here](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/general-data/heart_failure_clinical_records_dataset_edited.csv)

The original dataset version was collected by Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza (Government College University, Faisalabad, Pakistan) and made available by them on FigShare under the Attribution 4.0 International (CC BY 4.0: freedom to share and adapt the material) copyright in July 2017. The current version of the dataset was elaborated by Davide Chicco (Krembil Research Institute, Toronto, Canada) and donated to the University of California Irvine Machine Learning Repository under the same Attribution 4.0 International (CC BY 4.0) copyright in January 2020

Attribute Information: Thirteen (13) clinical features:

* age: age of the patient (years)
* anaemia: decrease of red blood cells or hemoglobin (boolean)
* high blood pressure: if the patient has hypertension (boolean)
* creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
* diabetes: if the patient has diabetes (boolean)
* ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
* platelets: platelets in the blood (kiloplatelets/mL)
* sex: male or female (categorical)
* serum creatinine: level of serum creatinine in the blood (mg/dL)
* serum sodium: level of serum sodium in the blood (mEq/L)
* smoking: if the patient smokes or not (boolean)
* time: follow-up period (days)
* [target] death event: if the patient deceased during the follow-up period (boolea
* n)**Instructions**

Please answer the following questions:

* Why would larger times be associated with LOWER death rates?
* is there a difference in the feature values between the sexes in the study?

We would like to know which, if any, factors can predict the death of patients, and which factor(s) make the biggest contribution(s) to the model. Secondly, what metric(s) did you choose to evaluate your model and why?

Please include both the code and an explanation of which features you found to be most predictive and how you found them. We will use this information to determine which tests we should run at every appointment to be sure that we can provide the best possible care for our patients.

We would like to see all of your code, including any EDA or feature engineering as well as the end performance of your model.

### **Checklist**

Here is a checklist of things to consider as you work on this project:

* Make sure your code is clean, well-documented, and that your steps are clearly laid out.
* The assessment question in question 1 only has one correct answer. Make sure to double-check that there aren’t any nulls or errors and the data and that you did the math correctly. If there is a specified format, be sure to submit it in that format.
* There is an academic paper associated with this dataset. Your results may be similar. The process you take to get your answer and the way you communicate your results are most important in a take-home assessment.
* Make sure to provide a clear, data-driven analysis in a written or presentation format.

**Note**: a take-home won’t typically include a checklist like this one, but we’ve included one here for you as you become more familiar with the take-home.

# Common Unsupervised Learning Questions

**Practice answering common unsupervised learning interview questions.**

Welcome back! In this portion of the interview, we are going to dive into some unsupervised learning techniques. I know there is so much information to gain from the data we currently have, and we need a strong candidate who can identify patterns in our data that we cannot see.

I am interested in hiring someone who is familiar with popular techniques and also someone who knows how to evaluate a model when there isn’t a definitive answer. As with any candidate, I am looking for someone who is a problem solver and can clearly communicate their thinking and knowledge. Someone who is curious about our data and has ideas on how to solve our specific business problems.

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

#### **Conceptual Questions**

Free response

How do you choose the number of clusters for k-means?

**Submit Response**

*How do you choose the number of clusters for k-means?*

*Your response*

*..*

*Our answer*

*I have two methods I like to use when I begin to explore the optimal number for k. I first start by testing out a few different clusters. This depends in part on how big my dataset is, but in my experience, a k value between 3 and 15 has been a good place to start.*

*I test this by building a k-means model with these different values and store each model’s inertia in a list. I then create a line graph to visualize the inertial value vs. k number of clusters. With this graph, I would expect to see a high inertia value for when k = 3 and then see it begin to lower as we increase k.*

*I will be looking for the “elbow” within the plot. That is the point where our inertia value starts to level out. The number of clusters that are located at the “elbow” can be a good indicator of an optimal number of clusters.*

*I know that inertia will continue to get lower for each number of clusters I increase, which makes sense because inertia is the mean squared distance between each instance and its closest centroid. The more centroids there are, the closer our instances will be to at least one centroid.*

*Once I see where the “elbow” is, I have a good idea for the number of clusters to create, but I also know that a silhouette score is maybe a more precise approach. I will take the same number of clusters I evaluated for my inertia vs. cluster graph and evaluate each cluster’s silhouette score.*

*I’ll be looking for a silhouette score closer to 1 which tells me that the instance is close to its cluster. I can create another line graph of silhouette scores vs. the number of clusters for a richer interpretation of how my instances are relative to their centroid.*

*Just like any data science model, there is always a trade-off. Sure, the more clusters I create the lower my inertia and higher my silhouette score, but I may be separating my data into too many clusters, and if my number of clusters is too low I may not be able to derive business insights from fewer groups. Determining the right amount of clusters is an important step during our training phase to ensure when the model is put into production it will be able to handle new information.*

*WHY THIS RESPONSE?*

* *This response is great because you have listed out two ways in which to determine the optimal number of clusters based on your data. Codey will be thrilled to see that you have such a strong knowledge of both the inertia and silhouette scores and what information you can gain for both graphs.*

**Three clusters of space objects. Stars on the left, satellites in the center, and planets on the right.**

Free response

How do you evaluate the performance of a clustering?

**Submit Response**

Free response

| *How do you evaluate the performance of a clustering?*  *Your response*  *…*  *Our answer*  *Taking the time to evaluate each model I build is essentials to my process, especially with clustering. If I specify a k that is too high, we may be creating multiple clusters that could be better represented as one. Then again, if k is too low our clusters are too big and may not tell us the underlying patterns we are searching for.*  *One thing I am always aware of when clustering is the distribution of our data. Clustering techniques work best with data that has a non-uniform distribution and we can perform a statistical test to determine if our data is or is not normally distributed.*  *Let’s say our data is non-uniform and we want to see how a model is doing.*  *Ideally, we are looking for small intra-cluster distances, meaning that instances in the same cluster are close to each other, and large inter-cluster distance(s), meaning that the distance from an instance in one cluster is far away from an instance in another cluster. Another way to say this is that they are maximally compact.*  *The graph I like to use to visualize this is a silhouette diagram, where each cluster represents a knife-like shape, I know it is a crazy description but it’s an accurate one. The height of the shape tells me the number of instances in that cluster, and the length of the shape will tell me the sorted silhouette coefficients. What I am really looking for is to have all of the clusters reach beyond the dashed line which represents the silhouette score.*  *If the knife shape does not reach that line, I know that my instances are close to another cluster and have a high intra-cluster distance. This means that I need to revisit k. In practice, this means that I need to experiment with a different number of clusters. Usually, I run my experiments with multiple clusters and create a silhouette diagram for each one. It’s the first method I choose to quickly assess the performance of my model.*  ***WHY THIS RESPONSE?***  *This response is good because it demonstrates that the interviewee has a preferred method for assessing the performance of the model and is able to quickly compare multiple models in order to select the optimum value for k. Though there is no singular quantitative metric for evaluating the performance of the clustering, the interviewee walks through an objective and fairly typical method for evaluating how well a model has clustered the instances.* |
| --- |

What are some clustering algorithms, other than k-means?

**Submit Response**

Free response

| *What are some clustering algorithms, other than k-means?*  *Your response*  *…*  *Our answer*  *There are many different clustering algorithms, one of the most popular alternatives to k-means is a hierarchical cluster algorithm called Agglomerative clustering. This algorithm starts from the bottom and works its way up. By that I mean it starts with an individual instance and finds a similar instance by using a specified distance metric. The algorithm works its way through the dataset connecting pairs of instances and grouping the pairs into clusters. It eventually builds these clusters further and further until it has transversed the entire dataset. As the analyst, we can then review the final visual of every stage of that model (i.e., clusters of just two up to the entire dataset as one giant cluster) and determine at what level to choose the clusters.*  *Another popular model is DBSCAN. This algorithm is good for arbitrary-shaped clusters and can also be used to detect outliers. This model has two important hyper-parameters and let’s say we choose epsilon = 0.5 and minPts = 30. An instance is chosen at random and a radius of 0.5 is drawn around that instance. If within that radius there are at least 29 other instances - because the instance itself counts - then it can be labeled as a core point. Thus beginning the cluster formation. The algorithm will continue this process, and return a cluster index for each instance. A result of -1 would indicate an anomaly or an instance that is further than epsilon away from any instance.*  *I know I’ve already shared a lot about two clustering algorithms, but I also am aware of and have worked with the following: Mean-Shift, BIRCH, and Affinity propagation.*  *WHY THIS RESPONSE? This is a good response because the interviewee both listed alternative algorithms and described how they work. Providing the extra context, demonstrates that the interviewee can make good decisions about when to use each based on the data that they have.* |
| --- |

When would you use clustering versus a supervised learning model?

**Submit Response**

Free response

| *When would you use clustering versus a supervised learning model?*  *Your response*  *..*  *Our answer*  *Since clustering is an unsupervised machine learning algorithm, it means that we don’t have the answer as a feature in our dataset. Clustering techniques help us better understand our data when we don’t know where to start. These techniques can help us gain valuable insight that will drive business decisions. For example, if we have a wide user base but want to start targeting our advertisements based on user behavior, we might use clustering to identify the user groups we want to target.*  *However, we would use a supervised learning model when we have the target value listed as a feature in our dataset. Supervised learning depends on labeled training data, so if we have labels on our data or the labels are a feature within our dataset, supervised learning could be the best option. These techniques can help us leverage what we know to make predictions about new inputs.*  *WHY THIS RESPONSE? This is a strong answer because it highlights what both clustering and supervised algorithms do well and hits the same points for both methods. Drawing simple and clear business applications for both approaches.* |
| --- |

Explain how the k-means algorithm works.

**Submit Response**

| *Explain how the k-means algorithm works.*  *Your response*  *..*  *Our answer*  *At its most basic, there are three steps in a k-means algorithm. Find a spot, assign an instance to a centroid, make a movement.*  *K-means models are all based on how many clusters I want it to organize my data into that is ‘k’. ‘K’ is also called a hyper-parameter, or one of the global parameters that shape how the algorithm proceeds. The algorithm starts by creating k centroids and placing each one in a randomly assigned location of our data. Then, it takes each instance or observation and calculates how far it is from each centroid.*  *The centroid with the shortest distance is assigned to that instance, creating a cluster of data points associated with each centroid. Next, the values of each centroid are calculated by taking the number of instances in that cluster divided by the number of instances in the whole data set, then moving the centroid to a better location.*  *These three steps: place a centroid, assign instances to a centroid, calculate new value - will continue until the centroids find an optimal spot and the movement becomes smaller and start to stabilize.*  ***WHY THIS RESPONSE?*** *This is a good response because it begins with a very basic model, expands into the details of the model, and summarizes with another basic explanation of how the algorithm works. The basic-complex-basic structure showcases the fact that the interviewee both understands the details of the algorithm and can communicate about it to a variety of different audiences.* |
| --- |

Great job! You’ve given clear and understandable answers to some tough unsupervised classification interview problems! You are ready to head into your next interview with confidence that you can show off your knowledge.

### **Review**

In this interview, we covered:

* ways we can find an optimal number of clusters when working with different clustering algorithms.
* methods to evaluate each of our clustering models and ways in which to see how well our clusters are capturing the patterns in our data.
* clustering algorithms beyond just k-means.
* the difference between clustering and supervised learning.
* how exactly a k-means algorithm works.

**Common Clustering Questions**

### **Introduction to Common Clustering Questions**

<1 min

Welcome to common clustering interview questions, we are happy to see you back. We’ll review additional information that may be asked at any stage in the interview process. These are questions you may be asked in an interview with the hiring manager, the recruiter, perhaps during a whiteboard challenge, or even as part of your take-home assessment. This is not meant to be teaching content, but concepts learners should already be familiar with. Think of this as a chance to test your skills with unsupervised learning!

Best of luck, you got this.

**Instructions**

1. We will be working with data from clients of a wholesale distributor. The prices are in Monetary Units (m.u.), and there are 8 features in this dataset:
   * FRESH: annual spending on fresh products;
   * MILK: annual spending on milk products;
   * GROCERY: annual spending on grocery products;
   * FROZEN: annual spending on frozen products
   * DETERGENTS\_PAPER: annual spending on detergents and paper products;
   * DELICATESSEN: annual spending on and delicatessen products;
   * CHANNEL: customes type - Horeca (Hotel/Restaurant/Cafe) or Retail
   * REGION: customer region - Lisnon, Oporto, or Other

Descriptive Statistics: Let’s get started with our dataset and see what business insights we can find!

**Common Clustering Questions**

### **Set up your data frame for clustering**

2 min

We’ve provided you with a dataset that we’d like for you to use a K-Means algorithm to see what insights you can gain for the various clusters you will find. At this point, we hope you can explore the dataset and prepare it for the K-Means algorithm.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Begin by importing the dataset we’ve loaded for you in the workspace, wholesale\_dataset.csv. Be sure to import any packages you think you might need. We’ve imported the Codecademylib3 library to help you display any tables you might want to explore.
* Load your data frame with pandas
* Checkpoint 2 Passed
* **2.**
* Inspect your data frame, and get a good understanding of what it includes.
* Pandas will be helpful to inspect your features.
* Checkpoint 3 Passed
* **3.**
* You may have noticed that not all columns are numerical, what can you do to that column to prepare your dataset for a K-Means algorithm? Be sure to document your code so that anyone in the interview process can follow you steps.
* You’ll want to one-hot encode that feature.
* Checkpoint 4 Passed
* **4.**
* Be sure to have all of your features on the same scale. Write your plan and steps as a multi-line string.
* You can standardize for normalize your data.
* Checkpoint 5 Passed
* **5.** Copy your code, you’ll need to paste it into the next exercise!

import codecademylib3

#import libraries

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

#import dataframe

wholesale = pd.read\_csv('wholesale\_dataset.csv')

#inspect dataframe

print(wholesale.head())

print(wholesale.info())

print(wholesale.describe())

#one-hot encode categorical feature

channel\_onehot = pd.get\_dummies(wholesale['Channel'])

new\_wholesale = pd.concat([wholesale, channel\_onehot], axis = 1)

new\_wholesale.drop('Channel', axis = 1, inplace = True)

print(new\_wholesale.head())

#standardize your data

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

wholesale\_stand = scaler.fit\_transform(new\_wholesale)

### **Choosing K for K-Means**

1 min

Great job, you’ve inspected your data frame, you have a good understanding of what the data is. You’ve prepared it for modeling and you documented your thinking as a multi-line string.

Now it’s time to start clustering your data.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Paste in your code from exercise 2.
* There is a Copy button at the bottom of the text editor.
* Checkpoint 2 Passed
* **2.**
* Run your data frame through a few different values of K and determine which number of clusters is best for this data.
* Decide how you will measure the accuracy of each cluster.
* Checkpoint 3 Passed
* **3.**
* Visualize the different clusters you created, and determine what is the most appropriate number of clusters.
* Plot inertia vs cluster size.
* Checkpoint 4 Passed
* **4.**
* Determine the appropriate value for k based on your results. Be sure to list your reasoning as a multi-line string.
* Answers may vary here, just be sure to fully explain why you chose the number of clusters you did.
* Checkpoint 5 Passed
* **5.**
* Copy your code, you’ll need to paste it into the next exercise.

import codecademylib3

#import libraries

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import silhouette\_score

#import dataframe

wholesale = pd.read\_csv('wholesale\_dataset.csv')

#inspect dataframe

print(wholesale.head())

print(wholesale.info())

print(wholesale.describe())

#one-hot encode categorical feature

channel\_onehot = pd.get\_dummies(wholesale['Channel'])

new\_wholesale = pd.concat([wholesale, channel\_onehot], axis = 1)

new\_wholesale.drop('Channel', axis = 1, inplace = True)

print(new\_wholesale.head())

#standardize your data

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

wholesale\_stand = scaler.fit\_transform(new\_wholesale)

#storing inertia and silhouette score to evaluate each cluster

inertia\_values = []

silhouette\_scores = []

#looking at different values for k

for k in range(3, 16):

kmeans = KMeans(n\_clusters = k)

kmeans.fit(wholesale\_stand)

inertia\_values.append(kmeans.inertia\_)

silhouette\_scores.append(silhouette\_score(wholesale\_stand, kmeans.labels\_))

#visualize inertia

plt.plot(range(3,16), inertia\_values, marker = 'o')

plt.xlabel('Number of Clusters')

plt.ylabel('Inertia')

plt.show();

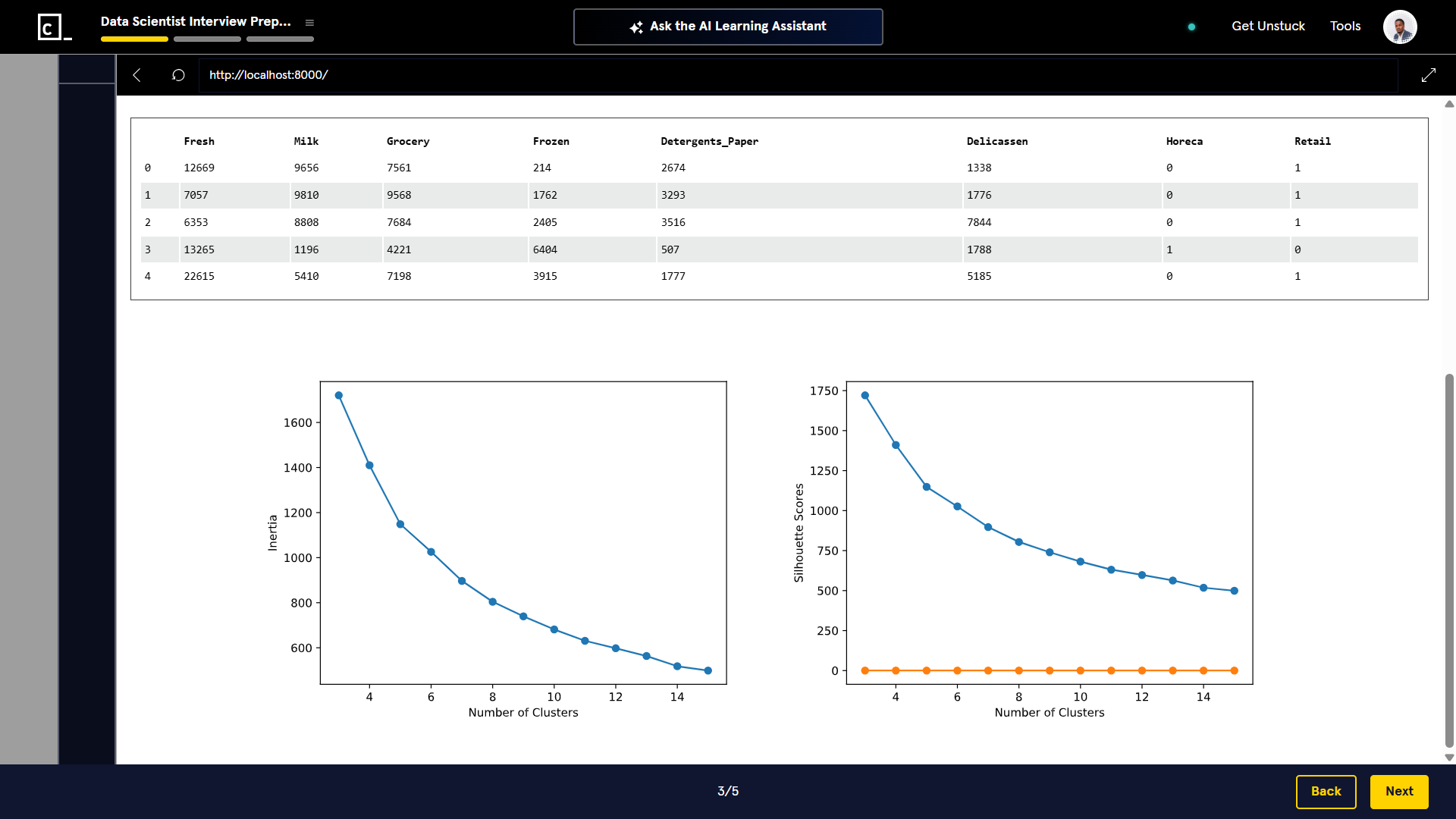
#visualize inertia CHECKPOINT 3

plt.plot(range(3,16), silhouette\_scores, marker = 'o')

plt.xlabel('Number of Clusters')

plt.ylabel('Silhouette Scores')

plt.show();



### **Interpreting Clusters**

1 min

Now that you have chosen your k, and documented your reasoning let’s cluster and see what we can find out about each cluster.

**Instructions**

* **1.** Paste in your code from the previous excerises.
* Checkpoint 2 Passed
* **2.** Run the k-means algorithm on however many number of clustered you decided.
* Checkpoint 3 Passed
* **3.** Create a visualization of the final clusters using a pairwise scatter plot.
* Add the cluster number onto the data frame and graph the pairwise scatter plot for each feature.

import codecademylib3

#Carry over

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

wholesale = pd.read\_csv('wholesale\_dataset.csv')

channel\_onehot = pd.get\_dummies(wholesale['Channel'])

new\_wholesale = pd.concat([wholesale, channel\_onehot], axis = 1)

new\_wholesale.drop('Channel', axis = 1, inplace = True)

scaler = StandardScaler()

wholesale\_stand = scaler.fit\_transform(new\_wholesale)

#run the k-means algorithm

kmeans = KMeans(n\_clusters = 4)

kmeans.fit(wholesale\_stand)

#add cluster column to dataframe

wholesale['cluster'] = kmeans.labels\_

#import seaborn to visualize pairwise scatter plot

import seaborn as sns

df = wholesale[['Fresh', 'Milk', 'Grocery', 'Frozen', 'Detergents\_Paper', 'Delicassen', 'cluster']].copy()

sns.pairplot(df, vars=df.columns[:-1])

plt.show();



### **Clustering Interview Review**

<1 min

Great job, you just provided great insight to your interview panel. During an interview around unsupervised learning there often is no ground-truth value to evaluate the output against. What hiring teams will be looking for is someone who can try a variety of different approaches (here you tried a variety of values for K to find the right number of clusters) and clearly communicate their findings, assumptions, and recommendations.

Many areas of businesses (marketing, loyalty, product recommendations) are increasingly using personalization to drive decisions. Customer segmentation is often at the center of such programs, and within this lesson, you were able to better understand purchasing orders.

We covered:

* How to prepare your data for a clustering algorithm
* How to visualize the different number of clusters
* How to determine an appropriate number of clusters
* How to turn your results into actionable information that the company can use to better understand / market / communicate with their users

# Unsupervised Learning Take-Home Assessment

**Practice your unsupervised learning skills with a sample take-home assessment.**

### **Overview**

Welcome to the take-home assignment! Take-home assessments help interviewers get a sense of your technical skills before moving you to the next stage of the interview process. The style and format of these projects can vary widely depending on the company and role you are applying for. In this project, we are going to give you a prompt and you are free to answer it however you wish. You can use any coding language and any tools you are familiar with. This is an opportunity to show off your skills.

The projects that we have prompted here are typical for data analyst roles, but are not definitive. However, there are some widespread practices and conventions that you should be aware of before starting the project.

* There may be one or more very easy question(s). These are not tricks, they are meant to easily exclude applicants who cannot perform basic functions.
* There will be some discussion, analysis, or reflection. Expect to have to write a summary of your findings. A large part of both Data Analyst and Data Scientist roles is communication. How well you communicate your findings is just as important as what you find and how you find it.
* You may be asked to submit your files in a specific way. Pay attention to these instructions and submit your file in the requested format. This is another area that may initially seem unimportant but can have an outsized influence on your candidacy.
* You may *not* be asked to submit your files in any particular way. If this is the case, Python files should be .py or .ipynb, R should be in .R or .Rmd format. Similarly, your written analysis should be in .pdf format or .md if submitted online.
* If anything is unclear or you are not sure what they are asking you to do, ask! It is much better to get clarification at the beginning of a project than submit something that doesn’t answer the question. This assignment is an opportunity to showcase your problem-solving skills. If you are stuck on a particular technical skill, try to find the answer on your own. Google and StackOverflow will be your friends during this process! If you still can’t figure it out, and the interviewer has allowed for questions, reach out and explain what you’ve already tried. This will show you have the ability to problem-solve on your own before reaching out for help.
* Your interviewers may give you an estimate of how long they think it should take. It is very common to go far beyond the time estimate.
* The take-home assessment is also a good opportunity for you to decide if this is the kind of work you are interested in doing. If you enjoy it, you are probably in the right place. If it is something you do not enjoy, it may be worth clarifying what the expectations of the job are.

### **Context**

You are applying for a role at a small marketing agency with a newly established data science team.

#### **Take Home Challenge**

Congratulations on being invited to complete this analysis project for our Data Science team!

You can use the tool(s) of your choice. The key to this assignment is to clearly document your steps and state any assumptions you’ve made along the way. Please submit your code as well as a summary of your findings.

When you turn in your work we expect to receive:

* Documentation of your steps (either as a code file or process notes).
* A written summary of your findings as a .pdf or Microsoft Office file.
* Any supplemental visualizations or materials that are necessary to illustrate your findings.
* All of your files combined as one zip file.

Best of luck, we look forward to reading your results.

#### **Dataset**

[Download dataset here](https://static-assets.codecademy.com/skillpaths/da-ds-interview-prep/general-data/Groceries_dataset.csv)

Attribute Information: This dataset contains three columns: a customer id, date, and purchases. It was compiled by [Heeral Dedhia](https://www.kaggle.com/heeraldedhia) and is available on [Kaggle](https://www.kaggle.com/heeraldedhia/groceries-dataset/metadata)

### **Instructions**

Please answer the following questions:

* What customer had the most visits?
* What is the median number of distinct items a customer buys per visit?

You can assume that on a given day, a customer makes only one visit.

For this project, would like to identify customer types. We have data about customers who shopped with us in 2015. All we have is a unique id, the date they shopped with us, and what was purchased. We would like to know how to group them so that we can better market to them. We need the number of groups, and a description of each group. There is no correct answer, so please justify why you think these groups are the correct ones.

You can use any additional dataset that you think may support you in this project. For example, if you want to classify the products into different categories, you can use a taxonomy, just be sure to document where it came from and why you used it.

We would like to see all of your code, including any EDA or feature engineering as well as a description of each of these groups, what are the key products that distinguish them from others?

### **Checklist:**

Here is a checklist of things to consider as you work on this project:

* Make sure your code is clean, well-documented, and that your steps are clearly laid out.
* The assessments in questions 1and 2 only have one correct answer. Make sure to double-check that there aren’t any nulls or errors and the data and that you did the math correctly. If there is a specified format, be sure to submit it in that format.
* Make sure to provide a clear, data-driven analysis in a written or presentation format.

**Note**: a take-home won’t typically include a checklist like this one, but we’ve included one here for you as you become more familiar with the take-home.

# Types of Problems in Model Tuning

**Practice answering questions about model tuning.**

### **Introduction**

Hi again! It’s me, Codey! I am sure you are eager to show how well you knew about machine learning model tuning. Most part of our job is highly related to the production of machine learning models, it is absolutely vital for the person to have a good understanding of model tuning, and I care much about how they think and analyze a problem instead of memorizing some concepts or methods.

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Questions**

Let’s get started!

Free response

What are training, validation, and test sets? How would you split data?

*Your response*

*..*

*Our answer*

*In supervised learning, the training set is the set of data that is used to fit the model at first place. The test set is a holdout set that is never used in model tuning. It is used to give a final evaluation of a model.*

*The validation set is used for hyper-parameter tuning, and helps reduce over-fitting, and is used in model selection. The model hasn’t seen the validation set during training, so the performance of models with different combinations of hyper-parameters can be evaluated on the validation set.*

*However, if we don’t expect to tune hyperparameters or if our goal is to get a general sense of model performance rather than optimizing the model, splitting data into training and test set is enough. Likewise, if we don’t have a lot of data and can’t really afford to create a validation set, I have used just a training and testing set.*

*However, in most cases to make sure getting a model good enough to put into production, a validation set is required. Normally I would have a 70-15-15 split - I use 70% of the original data as my training set, and 15% each for validation and test. If I have sufficient data and ensuring that the hyperparameters are well tuned, I might adopt a 60-20-20 split to be extra confident with the performance.*

*If I have limited data, I might use an 80-10-10 split or use cross-validation instead. This is not ideal, but might ultimately be better than skipping the validation set altogether. In the end, deciding whether or not to have a validation set and the breakdown to train/test/validate comes down to the goals of building the model and how much data I have.*

*WHY THIS RESPONSE?*

*This answer clearly describes what training, validation and test sets are and gives some examples of how to decide what to use and how to make the split. The interviewee demonstrates how they decide what to use and for what purposes.*

A blue rectangle with green and yellow dots representing a dataset is in the background. A clipboard with empty lines on it and a yellow pencil are in the foreground

Free response

What is cross-validation and why is it used?

*Your response*

*..*

*Our answer*

*Cross-validation means that we split the data into k parts, and train the model k times. Each time one of the parts is held out as the test set. Afterward, it is folded back into the training data and another is held out.*

*Typically when we’re building a machine learning model, we split the data into training/testing or training/testing/validation sets. This lets us build the model and keep the evaluation set separate. Of course, that is expensive. So, if we don’t have a lot of data or if we need to use all of our data, we can use cross-validation.*

*For example, data annotation can be expensive, and if the goal is to create the most robust model possible, we wouldn’t want to have data sitting around that we couldn’t use for training.*

*With cross-validation, we can still evaluate the performance and tune the hyperparameters, but we also don’t waste any training data. It’s also great because we get a more robust score. That is if we train K models and evaluate the average F1 score. This score is susceptible to variations in the testing set. That test may have slightly different features just by chance, which would affect the evaluation of our model–making it look better or worse than it actually is.*

*In an ideal situation, the performance of all the iterations would be similar. However, if there is an outlier, we know that we should do some investigation. In that case, we would want to go back to the data to be sure it is what we think it is and is nice and clean.*

*WHY THIS RESPONSE?*

*This answer provides a clear and succinct response to the question and gives practical reasons why cross-fold validation is sometimes the best approach. The interviewee goes on to give a second reason why cross-fold validation is useful and how it can diagnose problems in the data as well. Finally, they give a brief explanation of what to do if unexpected results are returned.*

Free response

How do you determine if a model is overfit or underfit?

*Your response*

*..*

*Our answer*

*If a model performs well on a self-test, that is, it performs well on the training data, but performs poorly on data it did not see during training, we can say that it is overfit. I like to think of this as having memorized the core features and the random noise in the training data.*

*Overfitting is easy to spot – if accuracy against the training set is very good, but accuracy against the validation set is not, then it’s an overfit model.*

*If the model does not perform well on its own training data and performs about the same–or even better–on unseen data, then it is not complex enough, and is underfit to the training data. Overfitting is not always the worst case if the problem space is very narrow. That is, if the model is only going to be used against data that is very very similar to the training data, then erring on the side of overfitting may fulfill the goals of the model better than erring on the side of underfitting.*

*Underfitting, on the other hand, is harder to spot. Typically if accuracy and validation accuracy are both poor, then I suspect it’s a situation of underfitting.*

*I sometimes like to think about this in terms of the bias/variance trade off. Just like the bias/variance trade off, there is a balance between a model being overfit and underfit - it’s the perfect fit! Overfitting is a model with high variance, and underfitting is a model with high bias. The goal is to find the balance that best suits our problem.*

*WHY THIS RESPONSE?*

*This is a solid response because the interviewee compared the two problem types, gave methods for identifying both, and contextualized the relationship between them in terms of another problem type: the bias-variance trade off. Furthermore, the interviewee demonstrated their understanding of the relationship between the practical problem and the performance of the model, showing that they consider the practical implications of the theoretical challenges.*

Free response

Describe k-fold vs LOO cross-validation. When is random assignment of data into folds not advisable?

*Your response*

*..*

*Our answer*

*Both k-fold and Leave One Out cross-validation are techniques to maximize training data. They are ways to use every data point for both training and testing. K-fold means that the data is divided into k subsets. The model is then trained k times, leaving on of the subsets out for testing each time. This allows us to both evaluate the model and use all of the data for training.*

*LOO is the most extreme form of k-folds because k is just equal to n. So, as many models are trained as there are data points, each time with one data point held out. Then a prediction is made for that one data point. LOO can be computationally expensive, though there are ways to combat that. More importantly, it can lead to overfitting because the predictions are made on single data points.*

*These methods are best for datasets with balanced classes. If the classes are unbalanced, random assignment of instances into k-folds could lead to very uneven splits. This will affect both what the model learns on and the distribution of the data it is evaluated against. As always, it is possible to get a dataset that doesn’t have any examples of one class or the other, but with very imbalanced datasets, that situation becomes even more probable. One way to combat these problems is to control the distributions with stratified k-fold cross-validation. This means that the data balance in each split is controlled.*

*Though stratified sampling will never solve the problem of imbalanced classes, it does help to control the situation and ensure even distribution across splits.*

*WHY THIS RESPONSE?*

*This is a strong answer because the interviewee both explained the difference between the two versions and followed it up with an explanation about some of the trade-offs. Second, the interviewee answered the question about random selection and also offered additional context for how to address it. Adding the solution demonstrated that they both understand the problem and have thought about ways to solve it.*

Free response

What methods are used to tune hyperparameters?

*Your response*

*..*

*Our answer*

*There are at least four different methods to tune hyperparameters:*

* *Manual tuning*
* *Grid Search*
* *Randomized Search*
* *Bayesian optimization*

*Manual tuning is obviously labor-intensive and error-prone, but it can be the most effective method especially when used at the end, when other methods have been exhausted.*

*Grid searches are thorough but time and energy intensive. Grid search means that we make a list of all of the combinations of hyperparameter candidates and then build a model for each combination and evaluate it. We then pick the model with the best performance. For example to find the optimal k in k-NN, we can build 98 k-NN classifiers by changing k from 3 to 100, and select the model with the best performance.*

*Randomized search is kind of like grid search but trieds to be more efficient. For randomized search, we don’t provide a discrete set of values oe value combinations, instead, we provide a statistical distribution for each hyperparameter from which the values can be randomly sampled. Theoretically, random search outperforms grid search in cases where hyperparameters are not equally important.*

*Finall, Bayesian optimization is the most different. The search-based methods are essentially experiments conducted in isolation across the logically possible options. Bayesian optimization allows improvement of sampling method based on previous result. It starts out similar to a search-type method by conducting an experiment with some selected hyperparameters. This experiment tells us something about the importance of the hyperparameters, and that is called our “prior.” Another experiment is conducted taking what we know to be true (our prior), and the result is used to update the prior. This process is repeated iteratively until convergence. The only problem is that it is possible to get stuck at a local maximum or minimum and not reach convergence.*

*WHY THIS RESPONSE?*

*This answer is good because in addition to listing and describing each method, the interviewee compared them and explained some of the benefits and limitations of each method.*

*Free response*

Explain the difference between parameters and hyperparameters of a model.

*Your response*

*…*

*Our answer*

*The main difference between model parameters and hyperparameters is whether they can be learned from data in the model training process. The model parameters are learned from data, and they are used to making predictions or estimations of new data. For example, the coefficients of a linear regression model are the model parameters for this regressor; another example of model parameters are the support vectors of an SVM (support vector machine).*

*Hyperparameters, on the other hand, cannot be “learned” from data. They are often used to help estimate model parameters. For a k-NN (k-nearest neighbor classifier), the value of k is a hyperparameter; for an SVM, the C and sigma are hyperparameters.*

*Sometimes they are used interchangeably though they shouldn’t be, but I understand when someone says something like “passing the parameters to the model”, they are talking about both the parameters and the hyperparameters.*

*WHY THIS RESPONSE?*

*This is a good response because the interviewee answered the question clearly and also acknowledged why the question would need to be asked to begin with. This demonstrates that they are aware of the discussions and confusions within the field.*

*Good job! It’s clear that you really understand the types of problems associated with model tuning, I’m looking forward to hearing more about hwo you otimize your models in the next interview!*

### **Review**

In this interview, we covered:

* The difference between training, testing, and validation sets;
* When to use cross validation and why;
* Different types of validation;
* How to deal with model under- and over-fitting;
* The difference between parameters and hyperparameters, and how to tune them.

It was great talking to you and I look forward to chatting with you in the rest of the interviews! See you there.

# Common Evaluation Questions

**Practice answering common questions about machine learning evaluation metrics.**

Great! I get to hire someone to join our group. In addition to knowing what algorithms to use and how to solve problems, they also need to know how to evaluate their models. There are some tried and true evaluation metrics that I want to be sure they are fluent in. They need to be able to both conduct an evaluation and interpret the result. I hope my next candidate is up for the challenge!

Welcome back to the next section of your Data Science interview! In this part, we are going to talk about model evaluation. Obviously, this is a really broad topic, but we will focus here on some specific questions. Let’s get started!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Questions**

Free response

How would you evaluate a machine learning model?

Your response

..

Our answer

Are you asking for binary classification models or regression models?

[Interviewer] Let’s talk about binary classification models first.

Got it! Thanks!

The choice of evaluation metric depends on the problem. But in general, before training the model, I split my data into training, validation, and test sets. I like to use a 70-15-15 random split if possible. But if there is not a lot of data, I might use 80-10-10. I could imagine using 60-20-20 if there is sufficient data and I need to be extra sure about the performance.

Sometimes a simple train/test split may be ok if we don’t expect to tune the hyperparameters. For example, if we are just trying to create a model to get a general sense of the data–not to put it into production.

No matter what, though, the training set is the largest proportion of data, and is used to train the model. The test set is used to evaluate the model. The validation set is where it gets interesting. A validation set is used to improve the model, to identify where it got things wrong, and tune those features. The validation set and test set have to be different because otherwise it leads to overfitting.

Overfitting is when the features of the model are too strongly associated with the features in the training or testing set. An overfit model will look like it is performing very well on data it has seen before or the specific set of data in the test set. However, when deployed on new instances, it won’t do as well because it has been overfit to the existing data.

Once I’ve separated my training, testing and evaluation sets, there are several metrics to evaluate the model’s performance.

The most basic way to evaluate a model is classification accuracy. However, it only works well with balanced data and gives very limited insight into what the model performs well on what it fails on.

A better method is to evaluate the model based on the Precision, Recall, and F1 scores. I also rely a lot on confusion matrices. Sometimes I might use a ROC curve or Precision-Recall Curve.

Precision tells me how many of the mentions my model labeled relevant are actually relevant. This is the ratio is True Positives (things labeled Relevant that are actually Relevant) to everything labeled Relevant (also called True and False Positives). Sometimes I need really high precision. This is especially true when misidentification is costly. Then, letting a few relevant instances go may not be bad.

For example, let’s say I am trying to market to customers. Since it’s expensive to market to people who won’t become customers, I want high precision. If I miss a few people, that is better than spending lots of money marketing to people who won’t buy my product.

Recall tells me how many instances of the real relevant class my model labeled as relevant. This is the ratio of True Positives to everything that is actually relevant. Sometimes I need really high recall. This is especially true when something bad would happen if any relevant instance is missed.

Disease detection is a good example of this. If I am screening for a disease, it’s ok if I misclassify some people as potentially having it when they don’t. However, it is not ok if I tell someone who has the disease that they don’t because they won’t get the treatment they need. Sure it will cause the healthy people some unnecessary anxiety, but it will also make sure everyone gets the treatment they need.

F1 is the harmonic mean between Precision and Recall, and gives me a good overall, comparative score to evaluate different models.

However, to really understand where my model is doing well and where it isn’t, I use a confusion matrix. A confusion matrix lists all of the True and False Positives and Negatives in one table so I can focus on what the model is classifying correctly and incorrectly.

Finally, if I’m trying to find the right probability threshold and a balance between true positives and false positives or precision and recall, I use a graph. If my Relevant and Irrelevant classes are balanced, I will use a ROC Curve, if they are not, I will use a Precision-Recall Curve.

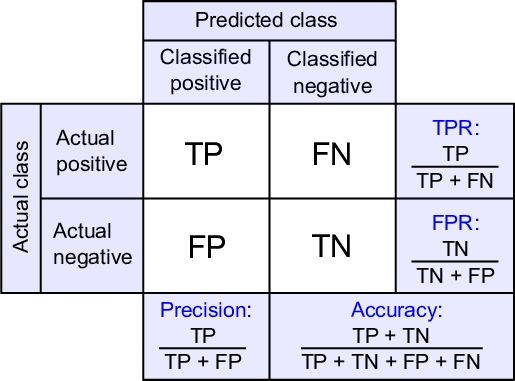
If I am explaining the results to others, I make sure to report it in terms that will be understandable and help contextualize the performance of the model in terms of why I’m making the model in the first place.

WHY THIS RESPONSE?

This answer is good because the interviewee asked for clarification first. The way the question is phrased, there actually is not enough information to answer it. While this may seem like a trick at first, the reality is that Data Scientists have to be confident enough in their knowledge to know when they simply need more information. It is therefore not uncommon to be asked questions that are incomplete in some way. Knowing when to ask for more details and when all the details are there is a skill unto itself. So, if you suspect you don’t have all the details you need, be sure to ask.

The interviewee covered the most important elements of binary classification model evaluation, including:

* splitting data into training, validation and test sets *before* training the model
* highlighting that simple accuracy can obscure the result when dealing with imbalanced data
* leveraging a confusion matrix to pinpoint exactly where the model performs well and where it does not.



Free response

How are precision and recall defined?

Your response

..

Our answer

Can I write it down?

[Codey] Yes of course. You can use the white board there.

Thanks!

Precision=TPTP+FP

*Precision*=

*TP*+*FP*

*TP*

​

Precision is defined as the number of true positives divided by the sum of true positives and false positives. Abstractly, this tells us the proportion of relevant identifications that were actually correct. It’s important to maximize precision when false positives (also known as irrelevant instances that were labeled relevant) are expensive or time-consuming. I would prioritize precision when I’m trying to make sure that everything labeled relevant really is – even if it means some relevant instances are misclassified.

Recall is defined as the number of true positives divided by the sum of true positives and false negatives. Abstractly, this tells me the proportion of actual positives that were identified correctly. It’s important to maximize recall when False Negatives (or relevant instances that are miscategorized as irrelevant) are costly. I would prioritize recall when I need to make sure that every relevant instance is captured.

Recall=TPTP+FN

*Recall*=

*TP*+*FN*

*TP*

​

WHY THIS RESPONSE?

First, the interviewee asked for the tools they need to successfully answer the question. For some, it might be easier to write a formula down and then explain it, and that is an acceptable accommodation to ask for if you need or want it. The interviewee also talked through the formula and contextualized how they would use it. Simply writing it down or reciting it is not enough. Knowing when to apply methods is more important than knowing what they are, so talking through those scenarios is key to this kind of answer.

Free response

How is F1 score computed?

Your response

..

Our answer

Can I use the whiteboard again?

[Codey] Of course!

F1=2precision×recallprecision+recall

*F*1=2

*precision*+*recall*

*precision*×*recall*

​

F1 Score is the harmonic mean between precision and recall. F1 is a good way to evaluate a model when the classes are not balanced. It is often compared to the Accuracy, and while they both give a good overview of model performance, F1 score is much better for quantifying both Type I and Type II errors.

However, I’ve found that F1 score can be difficult to understand because it is really abstract. Ultimately it is an average of averages. So, depending on the audience I sometimes also report the accuracy.

WHY THIS RESPONSE?

This answer first describes how F1 is calculated and gives some context for why F1 is used and compares it to accuracy. It would be possible to simply answer this question with an equation. However, that doesn’t give the interviewer the chance to see that you truly understand how F1 is used and why.

Free response

Is it better to have higher precision or higher recall?

..

**Feedback**

Not quite

**What you did well**

It seems like you didn’t provide an answer to the question. It’s great that you’re participating and attempting the questions!

**What could be improved**

Next time, try to provide an answer to the question. If you’re unsure, it’s okay to make an educated guess or provide your thoughts based on what you’ve learned so far.

Don’t worry, everyone starts somewhere. Keep going and don’t hesitate to try again!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

It really depends on the problem. Sometimes we want higher precision, and other times we care more about achieving a higher recall. In any situation, there is a trade-off between Precision and Recall, so optimizing our model depends on the goals.

For example, if we are developing a model to predict if a patient has a disease or not, we need to decrease the false negatives as much as possible, in other words, we don’t want to predict a person with a disease to be classified as healthy. If a healthy person is told they have a disease, it will be stressful, but not harmful to their health in the same way.

However, if we are building a model to classify if a video is suitable for children or not, we want the false positive rate to be as low as possible, so we want very high precision. In other words, we wouldn’t want videos with adult content to be classified as “kid-friendly”. Meanwhile, we don’t care as much about the false-negative rate because if we incorrectly label some videos as not “kid-friendly”, it won’t hurt anyone.

We could even imagine the same algorithm in different contexts being tuned to favor precision or recall. For example, if we developed a model to classify potential customers, if it is very cheap to market to people, we would want to favor recall and be sure to send it to everyone. If it is expensive to market to people, we would want to favor precision to save our budget. This could be achieved by simply modifying the confidence threshold.

WHY THIS RESPONSE?

This answer highlights that the precision-recall trade-off is a major struggle for Data Scientists. Since there will always be consequences to favoring one or the other, the answer will always depend on the situation. You cannot have both high precision and recall. This response also gives three examples: one for Recall, one for Precision, and another that highlights that even in the same environment, the balance depends on the details.

Free response

How would you improve the precision of a model?

..

**Feedback**

Not quite

**What you did well**

Unfortunately, your response was not provided. Therefore, it’s challenging to identify the strengths in your answer.

**What could be improved**

It seems like your response was not submitted correctly. Please make sure to provide a complete answer to the question. In this case, you were asked to explain how you would improve the precision of a model. You could discuss strategies such as adjusting the confidence threshold, overfitting the training data in certain situations, or performing feature engineering.

Don’t worry, everyone makes mistakes. Let’s try again and I’m sure you’ll do great!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

If possible, I would always add more high-quality training data, but that is not often possible.

So, first I manage expectations that if we improve precision, we are likely to reduce recall unless we change the algorithm. The first thing we can do is adjust the confidence threshold.

So, let’s say that we’ve set the threshold for a relevant or positive class to .55, I would increase it incrementally. This will of course mean that more truly relevant instances are classified as irrelevant, but it will also increase the precision of the model with minimal effort.

Another option is to overfit the training data. While this not a good practice, for some highly constrained situations, it might be appropriate to place greater emphasis on some features in order to be sure that our relevant instances are truly relevant.

If I’m trying to improve the performance of the model overall, I would likely do some feature engineering (assuming it has not been attempted yet). For example, if I’m using a Random Forest Classifier, I might use something like Randomized Search to find the right combination of hyperparameters which would return the parameters I should maximize. This is is better than hand-tuning because it leads to less over-fitting, but any time we are trying to maximize precision, we may be overfitting our model to the test set.

WHY THIS RESPONSE?

This answer is good because it acknowledges that improving the precision will almost always involve reducing the recall, and a big part of the data scientist’s job is making sure that all of the stakeholders understand the implications. So, leading with an acknowledgement of the fact that there are tradeoffs and managing expectations around them is essential for this kind of question.

The remainder of the question explores various strategies to improve precision, ranging from simple and quick to more in-depth and time-consuming. By exploring all of these options, the interviewee demonstrates that they understand the realities of data science within an organization and that there isn’t always time to improve the model if there just needs to be an increase in precision.

Free response

What is a ROC curve?

..

**Feedback**

Not quite

**What you did well**

It seems like you didn’t provide an answer to the question. It’s okay, everyone has moments of uncertainty.

**What could be improved**

Try to research and understand what a ROC curve is. Remember, it’s a graph used to show the performance of a classification model and it describes the relationship between true positive rate and false positive rate at different classification thresholds.

Don’t be discouraged! Learning is a journey and every step, even the challenging ones, brings you closer to your goal. Keep going!

Powered by OpenAI

Is this feedback helpful?

### **Our Answer**

A ROC curve is a graph that is used to show the performance of a classification model. ROC stands for is a receiver operating characteristic, but what it really does is describe the relationship between true positive rate and false positive rate at different classification thresholds.

It is a good way to find the right balance for precision and recall when the classes are imbalanced.

WHY THIS RESPONSE?

This answer both defines what an ROC Curve is and also demonstrates awareness of how to use one.

Great job! It is clear that you have a great sense of how to evaluate models and I’ve really enjoyed talking about it with you!

### **Review**

In this part of the interview, we covered:

* How to evaluate a machine learning model
* Precision, Recall, Accuracy, and F1 scores
* The ROC curve
* Overfitting and how to avoid it

It was great talking to you, and I look forward to talking to you in the next stage in the hiring journey. Let’s take a look at some more questions!

# Common Cross-Validation, Boosting, Bagging, and Loss Questions

**Practice answering interview questions about cross-validation, boosting, bagging, and loss.**

### **Introduction**

Welcome! In this last section, you will be asked about a wide range of topics in data science and machine learning. Since the field is constantly expanding and growing every day, it is important to keep your fundamentals sharp while also staying updated on new topics. Don’t worry if some of these are new concepts to you. You already have a really solid foundation in machine learning, so will be able to pick up the ideas quickly.

Let’s get started!

### **Directions**

For each question below, fill in your response in the free-response box provided. After you submit your answer, you will see our sample response. There are a lot of correct answers. Your answer may look different from ours, so we will touch on a few key points you should cover.

### **Questions**

Free response

What is the difference between bagging and boosting? Which would you recommend to reduce bias and/or variance?

..

***Feedback***

*Not quite*

***What you did well***

* *Unfortunately, your response was blank. Therefore, it’s difficult to identify specific strengths in this instance.*

***What could be improved***

* *It’s important to attempt the question, even if you’re unsure of the answer. Remember, it’s okay to make mistakes - that’s how we learn!*
* *Try to understand the difference between bagging and boosting. Both are ensemble machine learning techniques, but they use different approaches.*
* *Once you understand the concepts, try to explain them in your own words. This will help you solidify your understanding.*

*Don’t be discouraged! Learning is a journey, and it’s okay to not know everything right away. Keep studying and practicing, and you’ll get there. You can do it!*

*Powered by OpenAI*

*Is this feedback helpful?*

### ***Our Answer***

*Both bagging and boosting are ensemble machine learning techniques, which means combining weaker models to produce a more powerful model with a good balance between bias and variance.*

*While bagging uses a more parallel approach of averaging the predictions from many weaker models, boosting uses a sequential approach of incrementally improving upon a previous weaker model to arrive at the final model.*

*Bagging uses random sampling with replacement to create multiple sample datasets through which several models can be generated. The predictions of these independent models are then averaged depending on the nature of the problem to arrive at a more robust prediction. In this way, bagging helps to reduce the overall variance of the model’s predictions, thereby reducing overfitting.*

*Boosting on the other hand is more focused on reducing the bias of a model and is a corrective procedure that improves the model at each step of the learning process. Boosting also uses random sampling but the data can be weighted unlike in bagging. For example, if the weak model made some incorrect predictions during one of the stages, these data points are then given extra weights to make sure the improved model is able to better capture the feature space in an adaptive way.*

*WHY THIS RESPONSE?*

*The interviewee does a good job of explaining the term ‘ensemble’ before describing the difference between the two techniques. The rest of the response follows a clear logical progression where both boosting and bagging are explained independently but in a way that positions them in relationship to each other.*

Free response

What is a 5-fold cross-validation?

..

***Feedback***

*Not quite*

***What you did well***

*Unfortunately, your response was left blank and did not provide any information related to the question.*

***What could be improved***

*Next time, try to provide an answer that explains what 5-fold cross-validation is. Remember, it’s okay if you’re not completely sure. Making an attempt can help you learn and improve!*

*Don’t be discouraged! Learning is a journey and every step, even the challenging ones, brings you closer to your goal. Keep going!*

*Powered by OpenAI*

*Is this feedback helpful?*

### ***Our Answer***

*5-fold cross-validation is a specific case of k-fold cross-validation where the parameter k is equal to five. This means that the given dataset is split into 5 groups or “folds” to evaluate the model’s performance.*

*Just like in k-fold cross-validation, one of the groups is used as the test dataset while the model is trained on the remaining four groups. Then the model is evaluated on the group that was labeled the test dataset. This process is repeated for each of the five groups to get a rough idea of the evaluation scores for the model.*

*In the end, the model can be trained on all five groups and we can still have a robust evaluation metric.*

*WHY THIS RESPONSE?*

*While this is a short response, it is clear and to the point. Since cross-validation has been covered earlier in the interview process, providing a clear and direct answer both demonstrates knowledge of the method and what 5-fold cross-validation is specifically.*

Free response

Name an ensemble technique that uses bagging and one that uses boosting.

..

***Feedback***

*Not quite*

***What you did well***

*It seems like you didn’t provide an answer to the question. It’s okay, everyone has moments of uncertainty.*

***What could be improved***

*Next time, try to provide an answer even if you’re not completely sure. In this case, you could have mentioned ensemble techniques such as Random Forests for bagging and AdaBoost for boosting. Don’t forget to explain how these techniques work in machine learning.*

*Don’t be discouraged! Learning is a journey and every step, even the challenging ones, brings you closer to your goal. Keep going!*

*Powered by OpenAI*

*Is this feedback helpful?*

### ***Our Answer***

*Random Forests are a good example of bagging. Random Forests are typically used to reduce the variance of something like a decision tree. Bagging is means to reduce variance on models with high variance, and since decision trees tend to have very high variance (since they can get tuned a little too closely to the data they were trained on), bagging is a good way to improve them.*

*This would mean making a bunch of decision trees from samples of the data. Train a decision tree on each one and then get the output from each one and take the most frequently predicted result.*

*AdaBoost is a good example of a technique that uses boosting. AdaBooks fits weak learners to weighted training data. If the first learner makes an incorrect prediction then it gives a higher weight to that learner. This has the effect of giving a higher weight to difficult-to-classify instances versus those that the model is already handing well.*

*WHY THIS RESPONSE?*

*This is a good response because it both lists some examples of techniques that use bagging and boosting and explains each one with respect to how it is applied to real machine learning problems.*

Free response

What is a loss function?

..

***Feedback***

*Not quite*

***What you did well***

*It seems like you didn’t provide an answer to the question. It’s okay, everyone has moments of uncertainty.*

***What could be improved***

*Try to read the question carefully and provide an answer that explains what a loss function is, how it works, and why it’s important in machine learning. You can also provide examples of commonly used loss functions.*

*Don’t worry, you’re still learning and it’s okay to make mistakes. Keep going, you’re doing great!*

*Powered by OpenAI*

*Is this feedback helpful?*

### ***Our Answer***

*As the name suggests, a loss function is a mathematical tool to give us an idea of how well our chosen algorithm models the behavior of a given dataset. In other words, it measures how much “loss” has resulted when the model makes predictions on the dataset.*

*The loss function calculates the model error for a neural network type model. The most commonly used loss functions are cross-entropy and the mean squared error. Neural networks are complex, so it’s not really possible to calculate the weights of the predictors. So, to optimize the performance of the algorithm, we try to minimize loss.*

*Ultimately, machine learning is inherently an optimization problem, which means that we are looking for the best way to represent the problem space. Candidates for the best solution are evaluated with an objective function. For neural networks, the objective function is called a loss function. The value of the loss function is the loss.*

*WHY THIS RESPONSE?*

*This response is good because the interviewee demonstrated that they understand both what the loss function is and that it is a special type of another class of functions: objective functions. By defining the dependencies, they also demonstrated that they are aware of the purpose for a loss function and how it affects model selection.*

Free response

What are some common loss functions for classification tasks and regression tasks?

..

***Feedback***

*Not quite*

***What you did well***

*Unfortunately, your response was left blank, so I couldn’t identify any strengths in your answer.*

***What could be improved***

*Next time, try to provide an answer to the question. For this particular question, you could discuss some common loss functions for classification tasks and regression tasks. For example, Binary Cross-Entropy and Hinge Loss are common for classification tasks, while Mean Squared Error, Mean Absolute Error, and R^2 are common for regression tasks. Providing an explanation for why these loss functions are used in their respective tasks would also enhance your answer.*

*Don’t worry, everyone has moments where they’re unsure of an answer. Keep studying and practicing, and you’ll get the hang of it!*

*Powered by OpenAI*

*Is this feedback helpful?*

### ***Our Answer***

*Some common loss functions for classification tasks are Binary Cross-Entropy and Hinge Loss. For regression tasks, the most common ones are Mean Squared Error, Mean Absolute Error, and R^2. There are different loss functions because the algorithms are doing different things and have different goals.*

*For classification tasks with Support Vector Machines, the goal is maximum-margin classification, which means that the model is trying to make sure that the predictions are as confident as possible. I know that is a massive simplification, so I’d be happy to talk more about that if you’d like.*

*Anyhow, with max-margin type classifiers, hinge loss is most appropriate for training because the value is based on how far from the decision boundary a prediction is.*

*For Regression tasks, Mean Squared Error is a very popular way to calculate the loss of a model. Mean Squared Error is, as the name suggests, the average fo the squared errors. The error is the residual error or the difference between the real and predicted values. This is good for regression tasks because it is differentiable, so it is possible to use it to find the point where the model has the least error. This is good because*

*WHY THIS RESPONSE?*

*This response is good because the interviewee both listed functions to calculate loss and explained a few of them. When giving the context, they also offered to go into more detail about points that they oversimplified, indicating that they are aware that they are making an oversimplification, but are ready to provide a more in-depth explanation on any of the points.*

Butterflies being caught in a net against a grid. The grid is skewed where the net is.

Congratulations! You are ready to talk about boosting, bagging, and loss in your next interview!

### **Review**

In this interview, we covered:

* Bagging vs Boosting
* K-fold cross-validation
* Ensemble Techniques
* Loss Functions

Great job! You are grounded on the fundamentals and also well-updated on various techniques on boosting, bagging, and loss. You are going to do great in your next interview!

# Common PCA questions

# Introduction

<1 min

Welcome back to the last section of your interview! In this section, we are going to cover Principle Component Analysis (PCA). We are going to ask some questions about PCA and ask you to write your response as a multi-line string in the console. Typically the coding questions would be asked as part of the take-home challenge, but we are going to pull them out here to give you a chance to practice implementing code for PCA and walking through a problem.

There is no code evaluation because there are many ways to implement these solutions and it is important that you use your own. However, if you select ‘View Solution’ you can see our sample responses.

**Instructions**

We’ve uploaded a dataset about wild blueberry yield that was developed by Saurabh Shahane and made [available on Kaggle](Saurabh Shahane). There are 18 variables, but since we are focusing on PCA, we ultimately won’t use those variables in their current shape. But, you can learn more about this data in the Readme.txt file that is loaded into your learning environment.  
Without further ado, let’s get started with a dataset that is sweet enough to eat (or at least do PCA on)!

### **When to use PCA**

1 min

We will start by asking a few conceptual questions. Write your response as a multi-line string in the text editor. Wait until you’ve finished both questions before getting the solution because they will both be displayed in the solution.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* What is PCA used for?
* There are at least two primary uses for PCA.
* Checkpoint 2 Passed
* **2.**
* What is PCA?
* Please explain what PCA is and what its objectives are.

"""There are two primary applications for PCA. The first is to reduce the dimensions so that a multidimensional space can be visualized. A common application of this is with clustering algorithms. There may be a lot of clusters and we would want to compare how similar they are to each other. In order to visually represent that on a 2 dimensional plane, we could use PCA to get the two dimensions that contribute the most variance to the model.

The second is to reduce the feature space to make training a model faster. Training a model with lots of features takes up a lot of memory and time. By reducing the feature space into principal components, we can speed that training time up and potentially use fewer resources with minimal loss in model performance. """

### **Inspect the blueberries dataset**

1 min

The most important part of any machine learning project is to get familiar with your dataset. Since we know that PCA will be an important part of working with this dataset, be sure to explore it with dimensionality reduction in mind.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Start by getting familiar with the **blueberries.csv** dataset. Since we know we will need to do PCA, be sure to find out how many features there are.
* Checkpoint 2 Passed
* **2.**
* The last column is our yield, which is our target variable. Separate the target variable and the row numbers from the **blueberries.csv** dataset.
* Checkpoint 3 Passed
* **3.**
* Check for how many highly correlated features there are. You can use any correlation value you want, but we will use 0.8.
* There are many ways to find out how many highly correlated features there are. In fact, you have implemented at least one way in the Data Scientist Interview Preparation Skill Path already.
* Checkpoint 4 Passed
* **4.**
* Make a heatmap of the absolute value of the correlations. You can use any color scheme, but we will use Grey.
* A heatmap will help put labels to the highly correlated features to check to be sure it makes sense.

import codecademylib3

import pandas as pd

import numpy as np

import seaborn as sns

from matplotlib import pyplot as plt

data = pd.read\_csv("blueberries.csv")

#determine how many variables there are

print(data.shape)

print(data.dtypes)

#look at the head of the dataset

print(data.head())

# check for missing data

print(data.isnull().sum().sum())

# separate target

target = data.iloc[:,-1]

# drop target and Row# from data

data.drop(data.columns[[0]], axis=1, inplace=True)

data.drop(data.columns[[-1]], axis=1, inplace=True)

# check number of highly correlated features

correlated\_features = set()

correlation\_matrix = data.corr()

for i in range(len(correlation\_matrix .columns)):

for j in range(i):

if abs(correlation\_matrix.iloc[i, j]) > 0.8:

colname = correlation\_matrix.columns[i]

correlated\_features.add(colname)

print(len(correlated\_features))

# map absolute value of correlations

sns.heatmap(data.corr().abs(),cmap="Greys")

plt.show()

### **Standardize the features**

<1 min

I’m sure you noticed that the features in this dataset are not standardized. Before moving on to dimensionality reduction you will need to standardize the features.

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Scale the features using your favorite scaling method.
* We will use the [Standard Scaler from scikit-learn](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html) in our example.

import codecademylib3

import pandas as pd

import numpy as np

import seaborn as sns

from matplotlib import pyplot as plt

data = pd.read\_csv("blueberries.csv")

#determine how many variables there are

print(data.shape)

print(data.dtypes)

#look at the head of the dataset

print(data.head())

# check for missing data

print(data.isnull().sum().sum())

# separate target

target = data.iloc[:,-1]

# drop target and Row# from data

data.drop(data.columns[[0]], axis=1, inplace=True)

data.drop(data.columns[[-1]], axis=1, inplace=True)

# check number of highly correlated features

correlated\_features = set()

correlation\_matrix = data.corr()

for i in range(len(correlation\_matrix .columns)):

for j in range(i):

if abs(correlation\_matrix.iloc[i, j]) > 0.8:

colname = correlation\_matrix.columns[i]

correlated\_features.add(colname)

print(len(correlated\_features))

# map absolute value of correlations

sns.heatmap(data.corr().abs(),cmap="Greys")

plt.show()

### **Select Principal Components**

1 min

It’s finally time to find out what your principal components will look like!

**Instructions**

* Checkpoint 1 Passed
* **1.**
* Copy the code to scale your features from the previous exercise, and start by implementing PCA on the entire dataset.
* In our solution, we are just trying to reduce the feature space (not visualize the data), so we used all of the components before determining how many to keep.
* Checkpoint 2 Passed
* **2.**
* Create a scree plot to visualize how much variation is explained by the principal components.
* Plot the [explained variance ratio](https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html) on the y axis.
* Checkpoint 3 Passed
* **3.**
* Given the shape of the scree plot, how many principal components will you keep? Why? Write your answer as a multi-line string in the text editor.
* There are elbows at 5 and 10 components. It may be easier to find the cumulative sum of the variance.
* Checkpoint 4 Passed
* **4.**
* Recreate the PCA with just the selected number of components.
* We used 6 components.
* Checkpoint 5 Passed
* **5.**
* Inspect the loadings for the dataset to be sure that none of the features contribute strongly to any of the principal components.
* Use the same technique for finding highly correlated values.
* Checkpoint 6 Passed
* **6.**
* What do the loadings mean in terms of your principal components? Write your answer as a multi-line string in the text editor.
* Should you change anything about the principal components?

import codecademylib3

import pandas as pd

import numpy as np

import seaborn as sns

from matplotlib import pyplot as plt

from sklearn.preprocessing import StandardScaler

from sklearn import decomposition

data = pd.read\_csv("blueberries.csv")

# standardize

scaler = StandardScaler()

std\_feats = scaler.fit\_transform(data)

std\_feats = pd.DataFrame(std\_feats, columns=data.columns)

print(std\_feats.head())

# PCA

pca = decomposition.PCA()

PCs = pd.DataFrame(pca.fit\_transform(std\_feats))

# scree plot

PCs = np.arange(pca.n\_components\_) + 1

plt.plot(PCs, pca.explained\_variance\_ratio\_, 'ro-', linewidth=2)

plt.title('Scree Plot')

plt.xlabel('Principal Component')

plt.ylabel('Proportion of Variance Explained')

plt.show()

#calculate proportion of variance explained by principal components

print ("Proportion of Variance Explained : ", pca.explained\_variance\_ratio\_)

#calculate the cumulative proportion of variance explained by principal components

out\_sum = np.cumsum(pca.explained\_variance\_ratio\_)

print ("Cumulative Prop. Variance Explained: ", out\_sum)

"""In the scree plot, it seems like there are 2 'elbows'. The firs is around 5 and the second is around 10. When I look at the cumulative proportion of variance, it seems that if I keep 6 principal components, I've explained almost 95% of the variation in the dataset. So I will keep the first 6 components."""

#perform pca with just 6 components

pca\_selected = decomposition.PCA(n\_components = 6)

PCs\_selected = pd.DataFrame(pca\_selected.fit\_transform(std\_feats))

#calculate the loadings

loadings = pd.DataFrame(pca\_selected.components\_.T, index=data.columns)

print(loadings)

#determine if any features contribute strongly to any of the components

largest\_loadings = set()

for i in range(len(loadings.columns)):

for j in range(i):

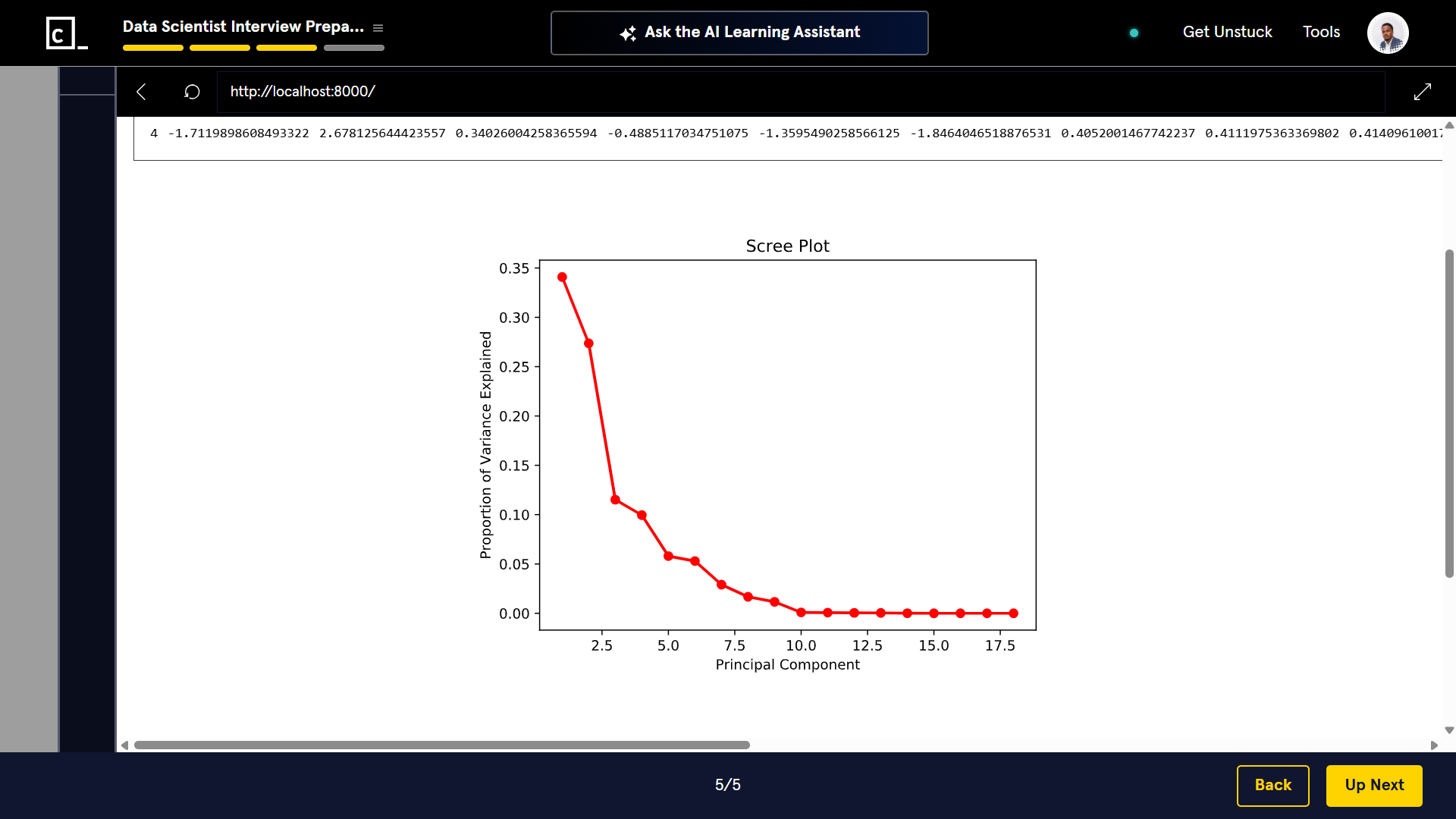
if abs(loadings.iloc[i, j]) > 0.7:

colname = loadings.columns[i]

largest\_loadings.add(colname)

print(len(largest\_loadings))

"""None of the features contribute strongly to the components, so it seems that the principal components are the best most efficient way to describe the variation in our dataset."""



# Review of Algorithms and Machine Learning Interviews

**See what you’ve covered in Algorithms and Machine Learning.**

### **Review**

Congratulations! The goal of this unit was to get practice answering questions about machine learning and algorithms.

Having completed this unit, you are now able to:

* Answer questions about supervised and unsupervised classification methods.
* Implement supervised and unsupervised algorithms on real data to answer business questions.
* Discuss evaluation of classification and regression tasks.
* Talk your interviewer through how you improve upon machine learning models.
* Evaluate predictions from supervised and unsupervised machine learning models.

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

# What is a Portfolio Project?

This project is a little different from other Codecademy projects you’ve encountered. In this project, you will bring together what you have learned in previous lessons to build a project in your own development environment and publish it to the web!

Create your own files outside of the Codecademy platform

Write your own code with less guided instructions

Use common project management processes to track your progress

# How Do Portfolio Projects Work?

We’ll provide you with high-level tasks to guide your project to completion, but you will be responsible for deciding how to implement them in your code.

There are many possible ways to correctly fulfill all of these requirements, and you should expect to use the internet, Codecademy, and other resources when you encounter a problem that you cannot easily solve.

Note that there are hints that can assist you, but they will only provide one potential implementation. Do not worry if your program looks different from ours!

### **Project Overview**

This project is not one project, but many different sub-projects. Depending on how much of the Data Scientist Career Path you have completed, or how many other projects you have done, this project may be primarily about organization and refinement. If you don’t have any completed projects, this project may be the beginning of your journey to developing a portfolio.

Regardless of where you are starting, by the end of this portfolio-building project, you will have a job search-ready portfolio that shows off your skills.

We will not be providing step-by-step instructions, but we will provide a checklist of elements your portfolio should have and suggestions on how to develop those elements.

### **Project Objectives:**

* Collect and organize completed projects
* Create new projects to fill in any skills gaps
* Align your portfolio to your career goals

### **Prerequisites:**

* Data Scientist Career Path

## **Project Tasks**

Keep track of your progress by dragging each task from "To Do" to "In Progress" to "Done" as you work on them. You can also click on a task to see more information about it.

To Do

**Establish and Online Presence**

Create a website or GitHub repository with the following characteristics:

* Your name, initials, or professional handle
* A short 1-4 sentence professionally-oriented biography
* Links to your projects
* Title each project with a descriptive name (i.e., the title of the project)

Specific for GitHub:

* Each project has its own repository
* Pin repositories for the projects that you want to showcase
* Use the first sentence of each project’s **README.md** for a short description of your project

Specific for a Website:

* An organization system for projects (i.e., a menu or gallery)

**Tidy Your Code**

Good code is well documented and organized. Most people have to clean up their code to make it intelligible to others. While this step is not essential, it does demonstrate that you can work in an organized way.

* Write comments for any functions you created.
* Write comments for complicated or clever steps.
* Break functions into their own files and call those files in your main script.
* Separate phases of a project into separate files. For example, clean in one file, analyze in another.

**Supervised Learning Regression Project**

This project should use methods such as Single or Multiple Linear Regression and be sure to document any feature engineering or selection that you do.

If you completed the Data Scientist Career Path, the Yelp Rating Predictor Cumulative Project is a good example.

**Supervised Learning Classification Project**

This project should use methods such as KNN Classification or Logistic Regression and be sure to document any feature engineering or selection that you do. If you choose to use Logistic regression, be sure to document how you are interpreting the coefficients since it’s a common interview question.

If you completed the Data Scientist Career Path, the Twitter Classification Cumulative Project is a good example.

**Unsupervised Learning Project**

Include at least one unsupervised project. This might be a clustering question or incorporating PCA into another project. Be sure to document any feature engineering or selection you implement or–if you choose to fulfill this requirement with PCA–why you are using that method.

If you completed the Data Scientist Career Path, the Masculinity Survey Project is a good example.

**Data Acquisition Project**

Data acquisition is not necessary for all jobs, but can showcase your versatility and adaptability. You could choose to do web scraping with Beautiful Soup or work with a popular API such as the Dataverse API. This project does not need to stand on its own, getting data for another project will demonstrate this skill.

The Chocolate Scraping with Beautiful Soup Project from the Data Scientist Career Path is a good example of a web scraping project, though to include it in your portfolio you would have to recreate it off-platform and use a different website. Remember if you are webscraping, be a good internet citizen and do not bombard someone’s site with more requests than necessary.

For working with an API, the Dataverse API is a great repository of datasets that can be accessed via an API.

**Data Cleaning Project**

At least one project should work with a dataset from an original source. Original sources are typically either data that you have curated yourself or data that is distributed by someone before it is cleaned.

Be sure to document your cleaning steps and why you performed each step. This can be comments in your code, notes in a Jupyter Notebook, or an affiliated document.

You can find original sources in a variety of repositories. Some popular ones are:

* FiveThirtyEight
* Buzzfeed
* UCI
* Data is Plural
* Google Data Repository

**Hypothesis Testing**

While all of your projects should be question-oriented (meaning that they are designed around a specific question), at least one should also test a formal hypothesis.

Regardless if you can or cannot reject the null hypothesis, the purpose is to demonstrate that you know how to set up a hypothesis test and evaluate the results.

Your hypothesis test can take the form of an A/B Test or it can be a more traditionally academic type. The A/B Testing Take-Home Project from this Skill Path can fulfill this requirement or you can recreate any of the Hypothesis Testing Challenge Projects in the Data Scientist Career Path.

Be sure to summarize your findings in a written document. This can be a text file, presentation, or other format, but the summary and interpretation is as important as the analysis.

**Data Visualization Project**

Many of your projects will likely involve some sort of visualization. If any project leverages a lot of visualizations, it can fulfill this task. The Data Visualization portfolio project from this Skill Path or the portfolio project in the Data Scientist Career Path (Life Expectancy and GDP) are also good options.

**Role-specific Data Project**

Consider the industry and specific role you are interested in. If you are applying to a marketing agency, a healthcare company, or conducting financial analysis, you will want to have something in your portfolio that showcases your ability to work with that kind of data in particular.

# Review of the Data Scientist Interview Preparation Skill Path

**You have the data science skills, and now you have the interview skills to prove it!**

### **Review**

Congratulations!

The goal of the Data Analyst Interview Preparation Skill Path was to give you the opportunity to practice answering the types of questions you will see in a data analyst interview. In this Skill Path, you learned how to decode job postings, tailor your materials to a specific job, and answer almost any interview question. After completing this Path, you have familiarity with data analyst interview question types and the key elements that make a good response.

Having completed this unit, you are now able to:

* Specify what kind of role you want to apply for and tailor your job search materials and interview preparation to that specific job description.
* Anticipate the stages of a Data Scientist Interview process, from building your resume to the on-site interview.
* Respond to general data management questions—from cleaning to exploratory analysis.
* Explain your data visualization and communication process.
* Describe your approach to experimental design and project management.
* Ace the SQL technical challenge.
* Effectively communicate your solution to coding challenges.
* Successfully complete data scientist interview take-home challenges.
* Give detailed answers to machine learning interview questions.
* Implement and explain your code for building machine learning models.
* Present yourself professionally as a Data Scientist.

You can walk into your next data analyst job interview with confidence!

Learning is social. Whatever you’re working on, be sure to connect with the Codecademy community in the [forums](https://discuss.codecademy.com/). Remember to check in with the community regularly, including for things like asking for code reviews on your project work and providing code reviews to others in the [projects category](https://discuss.codecademy.com/c/project/1833), which can help to reinforce what you’ve learned.

Want to build your portfolio and help others? Contribute to Codecademy Docs! You can go to our [Contribution Guide](https://www.codecademy.com/resources/docs/contribution-guide) to get started.

|  | |
| --- | --- |
|  | |
| **Subjects and languages** Most progress  + 575 XP  newly added  **All**  Languages  Subjects   * 0 * 2.5K * 5K * 7.5K * 10K * **Data science** * 8,835 XP * **Python** * 5,794 XP * **Machine learning** * 3,995 XP * **Data analytics** * 3,540 XP * **Computer science** * 3,494 XP * **Math** * 1,846 XP * **Data engineering** * 1,491 XP * **Data visualization** * 1,425 XP * **R** * 1,160 XP * **SQL** * 1,015 XP * **Artificial intelligence** * 915 XP * **Web development** * 845 XP * **IT** * 520 XP * **Bash** * 395 XP * **JavaScript** * 236 XP * **DevOps** * 105 XP * **Cybersecurity** * 85 XP * **Web design** * 85 XP * **HTML & CSS** * 50 XP * **Mobile development** * 45 XP * **Cloud computing** * 10 XP * **Game development** * 10 XP   [Learn more about XP](https://help.codecademy.com/hc/en-us/articles/26497354279195-Skill-XP-FAQ)  **More available** | **Subjects and languages** Most progress  + 575 XP  newly added  All  **Languages**  Subjects   * 0 * 1.5K * 3K * 4.5K * 6K * **Python** * 5,794 XP * **R** * 1,160 XP * **SQL** * 1,015 XP * **Bash** * 395 XP * **JavaScript** * 236 XP * **HTML & CSS** * 50 XP   **More available** |
| **Subjects and languages** Most progress  + 575 XP  newly added  All  Languages  **Subjects**   * 0 * 2.5K * 5K * 7.5K * 10K * **Data science** * 8,835 XP * **Machine learning** * 3,995 XP * **Data analytics** * 3,540 XP * **Computer science** * 3,494 XP * **Math** * 1,846 XP * **Data engineering** * 1,491 XP * **Data visualization** * 1,425 XP * **Artificial intelligence** * 915 XP * **Web development** * 845 XP * **IT** * 520 XP * **DevOps** * 105 XP * **Cybersecurity** * 85 XP * **Web design** * 85 XP * **Mobile development** * 45 XP * **Cloud computing** * 10 XP * **Game development** * 10 XP   **More available** |  |
|  |  |
|  |  |
|  |  |
|  |  |