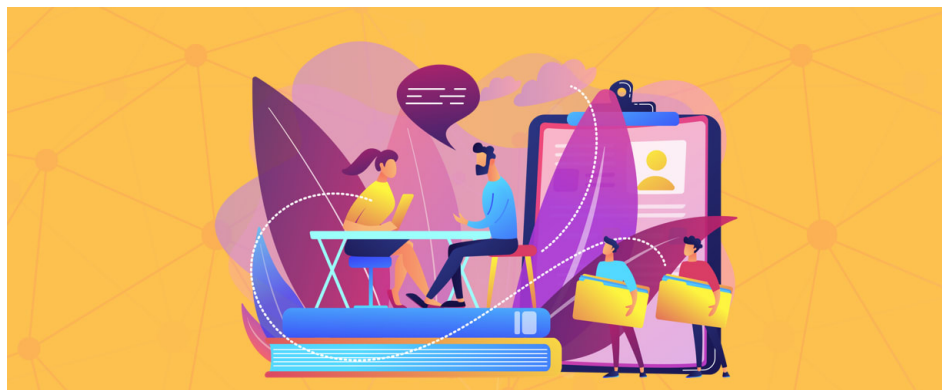Springboard **Blog**

≡

November 29, 2018

109 Data Science Interview Questions and Answers

Michael Rundell

A Curated List of Data Science Interview Questions and Answers

Preparing for an interview is not easy–there is significant uncertainty regarding the data science interview questions you will be asked. No matter how much work experience or what data science certificate you have, an interviewer can throw you off with a set of questions that you didn't expect.

During a data science interview, the interviewer will ask questions spanning a wide range of topics, requiring both strong technical

220 shares

through a variety of questions and question styles that are intentionally designed to keep you on your feet and force you to demonstrate how you operate under pressure.

Preparation is the key to success when pursuing a career in data science, and that includes the interview process.

This guide contains all of the data science interview questions you should expect when interviewing for a position as a data scientist. At Springboard, we teach data science through our self-guided, mentor-supported data science workshops. They're a great way to learn data science and get expert guidance on how to get a data science job.

We previously created a free data science interview guide, yet we still felt we had more to explore. So we curated this list of real questions asked to data science interview candidates. From this list of **data science interview questions**, an interviewee should be able to prepare for the tough questions, learn what answers will positively resonate with an employer, and develop the confidence to ace the interview.

We've broken the interview questions for data scientists into six different categories: statistics, programming, modeling, behavior, culture, and problem-solving.

1. **Statistics**
2. **Programming**
    1. *General*
    2. *Big Data*
    3. *Python*
    4. *R*
    5. *SQL*
3. **Modeling**
4. **Behavioral**
5. **Culture Fit**
6. **Problem-Solving**

1. Statistics Interview Questions

Statistical computing is the process through which data scientists take

accordingly, it is likely a good interviewer will try to probe your understanding of the subject matter with statistics-oriented data science interview questions. Be prepared to answer some fundamental statistics questions as part of your data science interview.

Here are examples of rudimentary statistics questions we've found:

1. What is the Central Limit Theorem and why is it important?
   - "Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible. While we can't obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample. The Central Limit Theorem addresses this question exactly." *Read more* here.
2. What is sampling? How many sampling methods do you know?
   - "Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined." *Read the full answer* here.
3. What is the difference between type I vs type II error?
   - "A type I error occurs when the null hypothesis is true, but is rejected. A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected." *Read the full answer* here.
4. What is linear regression? What do the terms p-value, coefficient, and r-squared value mean? What is the significance of each of these components?
   - A linear regression is a good tool for quick predictive analysis: for example, the price of a house depends on a myriad of factors, such as its size or its location. In order to see the relationship between these variables, we need to build a linear regression, which predicts the line of best fit between them and can help conclude whether or not these two factors have a positive or negative relationship. *Read more* here *and* here.
5. What are the assumptions required for linear regression?

- There are four major assumptions: 1. There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data, 2. The errors or residuals of the data are normally distributed and independent from each other, 3. There is minimal multicollinearity between explanatory variables, and 4. Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.

6. What is a statistical interaction?
   - "Basically, an interaction is when the effect of one factor (input variable) on the dependent variable (output variable) differs among levels of another factor." *Read more here*.

7. What is selection bias?
   - "Selection (or 'sampling') bias occurs in an 'active,' sense when the sample data that is gathered and prepared for modeling has characteristics that are not representative of the true, future population of cases the model will see. That is, active selection bias occurs when a subset of the data are systematically (i.e., non-randomly) excluded from analysis." *Read more here*.

8. What is an example of a data set with a non-Gaussian distribution?
   - "The Gaussian distribution is part of the Exponential family of distributions, but there are a lot more of them, with the same sort of ease of use, in many cases, and if the person doing the machine learning has a solid grounding in statistics, they can be utilized where appropriate." *Read more here*.

9. What is the Binomial Probability Formula?
   - "The binomial distribution consists of the probabilities of each of the possible numbers of successes on N trials for independent events that each have a probability of π (the Greek letter pi) of occurring." *Read more here*.

Examples of similar data science interview questions found on Glassdoor:

2. Programming

To test your programming skills, employers will typically include two
specific data science interview questions: they'll ask how you would
solve programming problems in theory without writing out the code, and
then they will also offer whiteboarding exercises for you to code on the
spot. For the latter types of questions, we will provide a few examples
below, but if you're looking for in-depth practice solving coding
challenges, visit HackerRank. With a "learn by doing" philosophy, there
are challenges organized around core concepts commonly tested during
interviews.

2.1 *General*

1. With which programming languages and environments are you
   most comfortable working?
2. What are some pros and cons about your favorite statistical
   software?
3. Tell me about an original algorithm you've created.
4. Describe a data science project in which you worked with a
   substantial programming component. What did you learn from
   that experience?
5. Do you contribute to any open-source projects?
6. How would you clean a data set in (insert language here)?
7. Tell me about the coding you did during your last project?

2.2 *Big Data*

- The Hadoop Distributed File System (HDFS), MapReduce, and YARN. *Read more here.*

2. Explain how MapReduce works as simply as possible.
   - "MapReduce is a programming model that enables distributed processing of large data sets on compute clusters of commodity hardware. Hadoop MapReduce first performs mapping which involves splitting a large file into pieces to make another set of data." *Read more here.*

3. How would you sort a large list of numbers?

4. Say you're given a large data set. What would be your plan for dealing with outliers? How about missing values? How about transformations?

2.3 *Python*

1. What modules/libraries are you most familiar with? What do you like or dislike about them?

2. In Python, how is memory managed?
   - In Python, memory is managed in a private heap space. This means that all the objects and data structures will be located in a private heap. However, the programmer won't be allowed to access this heap. Instead, the Python interpreter will handle it. At the same time, the core API will enable access to some Python tools for the programmer to start coding. The memory manager will allocate the heap space for the Python objects while the inbuilt garbage collector will recycle all the memory that's not being used to boost available heap space. *Read more here.*

3. What are the supported data types in Python?
   - "Python's built-in (or standard) data types can be grouped into several classes. Sticking to the hierarchy scheme used in the official Python documentation these are numeric types, sequences, sets and mappings." *Read more here.*

4. What is the difference between a tuple and a list in Python?
   - "Apart from tuples being immutable there is also a semantic distinction that should guide their usage." *Read more here.*

Related: 20 Python Interview Questions with Answers

1. What are the different types of sorting algorithms available in R language?
   - There are insertion, bubble, and selection sorting algorithms. *Read more here.*
2. What are the different data objects in R?
   - "R objects can store values as different core data types (referred to as modes in R jargon); these include numeric (both integer and double), character and logical." *Read more here.*
3. What packages are you most familiar with? What do you like or dislike about them?
4. How do you access the element in the 2nd column and 4th row of a matrix named M?
   - "We can access elements of a matrix using the square bracket [indexing method. Elements can be accessed as var[row, column]." *Read more here.*
5. What is the command used to store R objects in a file?
   - save (x, file="x.Rdata")
6. What is the best way to use Hadoop and R together for analysis?
   - "Hadoop and R complement each other quite well in terms of visualization and analytics of big data. There are four different ways of using Hadoop and R together." *Read more here.*
7. How do you split a continuous variable into different groups/ranks in R?
   - *Read about this here.*
8. Write a function in R language to replace the missing value in a vector with the mean of that vector.
   - *Read about this here.*

Related: Interview Questions on R and Text Mining in R: A Tutorial will help with data mining interview questions.

2.5 *SQL*

Often, SQL questions are case-based, meaning that an employer will task you with solving an SQL problem in order to test your skills from a practical standpoint. For example, you could be given a table and asked

finally report your findings. If you do not feel ready to do this in an interview setting, Mode Analytics has a delightful introduction to using SQL that will teach you these commands through an interactive SQL environment.

1. What is the purpose of the group functions in SQL? Give some examples of group functions.
   - Group functions are necessary to get summary statistics of a data set. COUNT, MAX, MIN, AVG, SUM, and DISTINCT are all group functions.
2. Tell me the difference between an inner join, left join/right join, and union.
   - "In a Venn diagram the inner join is when both tables have a match, a left join is when there is a match in the left table and the right table is null, a right join is the opposite of a left join, and a full join is all of the data combined." *Read more here.*
3. What does UNION do? What is the difference between UNION and UNION ALL?
   - "UNION removes duplicate records (where all columns in the results are the same), UNION ALL does not." *Read more here.*
4. What is the difference between SQL and MySQL or SQL Server?
   - "SQL stands for Structured Query Language. It's a standard language for accessing and manipulating databases. MySQL is a database management system, like SQL Server, Oracle, Informix, Postgres, etc." *Read more here.*
5. If a table contains duplicate rows, does a query result display the duplicate values by default? How can you eliminate duplicate rows from a query result?
   - Yes. One way you can eliminate duplicate rows with the DISTINCT clause. *Read more here.*

For additional SQL questions that focus on looking at specific snippets of code, check out this useful resource created by Toptal.

Examples of similar data science interview questions found on Glassdoor:

3. Modeling

Data modeling is where a data scientist provides value for a company. Turning data into predictive and actionable information is difficult, talking about it to a potential employer even more so. Practice describing your past experiences building models—what were the techniques used, challenges overcome, and successes achieved in the process? The group of questions below are designed to uncover that information, as well as your formal education of different modeling techniques. If you can't describe the theory and assumptions associated with a model you've used, it won't leave a good impression.

Take a look at the questions below to practice. Not all of the questions will be relevant to your interview—you're not expected to be a master of all techniques. The best use of these questions is to re-familiarize yourself with the modeling techniques you've learned in the past.

1. Tell me about how you designed a model for a past employer or client.
2. What are your favorite data visualization techniques?
3. How would you effectively represent data with 5 dimensions?
4. How is k-NN different from k-means clustering?
   - k-NN, or k-nearest neighbors is a classification algorithm, where the k is an integer describing the number of neighboring data points that influence the classification of a

220 shares

the k is an integer describing the number of clusters to be created from the given data.

5. How would you create a logistic regression model?

6. Have you used a time series model? Do you understand cross-correlations with time lags?

7. Explain the 80/20 rule, and tell me about its importance in model validation.

   ○ "People usually tend to start with a 80-20% split (80% training set – 20% test set) and split the training set once more into a 80-20% ratio to create the validation set." *Read more here.*

8. Explain what precision and recall are. How do they relate to the ROC curve?

   ○ Recall describes what percentage of true positives are described as positive by the model. Precision describes what percent of positive predictions were correct. The ROC curve shows the relationship between model recall and specificity–specificity being a measure of the percent of true negatives being described as negative by the model. Recall, precision, and the ROC are measures used to identify how useful a given classification model is. *Read more here.*

9. Explain the difference between L1 and L2 regularization methods.

   ○ "A regression model that uses L1 regularization technique is called Lasso Regression and model which uses L2 is called Ridge Regression. The key difference between these two is the penalty term." *Read more here.*

10. What is root cause analysis?

    ○ "All of us dread that meeting where the boss asks 'why is revenue down?' The only thing worse than that question is not having any answers! There are many changes happening in your business every day, and often you will want to understand exactly what is driving a given change — especially if it is unexpected. Understanding the underlying causes of change is known as root cause analysis." *Read more here.*

11. What are hash table collisions?

    ○ "If the range of key values is larger than the size of our hash

two different keys can hash to the same table index. There are a few different ways to resolve this issue. In hash table vernacular, this solution implemented is referred to as collision resolution." *Read more* *here*.

12. What is an exact test?

    ○ "In statistics, an exact (significance) test is a test where all assumptions, upon which the derivation of the distribution of the test statistic is based, are met as opposed to an approximate test (in which the approximation may be made as close as desired by making the sample size big enough). This will result in a significance test that will have a false rejection rate always equal to the significance level of the test. For example an exact test at significance level 5% will in the long run reject true null hypotheses exactly 5% of the time." *Read more* *here*.

13. In your opinion, which is more important when designing a machine learning model: model performance or model accuracy?

    ○ *Here's* *one approach to this question*.

14. What is one way that you would handle an imbalanced data set that's being used for prediction (i.e., vastly more negative classes than positive classes)?

15. How would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regression?

16. I have two models of comparable accuracy and computational performance. Which one should I choose for production and why?

17. How do you deal with sparsity?

18. Is it better to spend five days developing a 90-percent accurate solution or 10 days for 100-percent accuracy?

19. What are some situations where a general linear model fails?

    ○ *Read about this* *here*.

20. Do you think 50 small decision trees are better than a large one? Why?

    ○ *Read about this* *here*.

21. When modifying an algorithm, how do you know that your changes are an improvement over not doing anything?

22. Is it better to have too many false positives or too many false

Examples of similar data science interview questions found on Glassdoor:

Data Scientist at Apple was asked...   Nov 6, 2013

How do you take millions of users with 100's of transactions each, amongst 10k's of products and group the users together in a meaningful segments?

2 Answers ⌄

Senior Data Scientist at SparkCognition was asked...   Feb 26, 2017

academic-like questions where you could just look up the answer. just memorize a thing or two about the basic algos like neural nets, random forests, svm, linear regression. that satisfies their data scientists.

2 Answers ⌄

Data Scientist at Instacart was asked...   Jun 21, 2016

How would you tune a random forest?

1 Answer ⌄

Data Scientist at Groupon was asked...   Jul 17, 2014

I signed an NDA, so I can't give much details, but one interviewer asked me a very open ended question that involved how I would create/design/implement a certain algorithm from start to end.

1 Answer ⌄

4. Past Behavior

Employers love behavioral questions. They reveal information about the work experience of the interviewee and about their demeanor and how that could affect the rest of the team. From these questions, an interviewer wants to see how a candidate has reacted to situations in the past, how well they can articulate what their role was, and what they learned from their experience.

There are several categories of behavioral questions you'll be asked:

1. Teamwork
2. Leadership
3. Conflict management
4. Problem-solving
5. Failure

Before the interview, write down examples of work experiences related to these topics to refresh your memory—you will need to recall specific examples to answer the questions well. When asked about a prior experience, make sure you tell a story. Being able to concisely and

example: "I was asked X, I did A, B, and C, and decided that the answer was Y."

Of course, if you can highlight experiences having to do with data science, these questions present a great opportunity to showcase a unique accomplishment as a data scientist that you may not have discussed previously.

Here are examples of these sorts of questions/prompts:

1. Tell me about a time when you took initiative.
2. Tell me about a time when you had to overcome a dilemma.
3. Tell me about a time when you resolved a conflict.
4. Tell me about a time you failed and what you have learned from it.
5. Tell me about (a job on your resume). Why did you choose to do it and what do you like most about it?
6. Tell me about a challenge you have overcome while working on a group project.
7. When you encountered a tedious, boring task, how would you deal with it and motivate yourself to complete it?
8. What have you done in the past to make a client satisfied/happy?
9. What have you done in your previous job that you are really proud of?
10. What do you do when your personal life is running over into your work life?

Examples of similar data science interview questions found on Glassdoor:

Database Analyst/Data Scientist at Constellation Brands was asked...   Sep 6, 2016

Look at this email this jerks sent me, what would you do about this? (Guy actually shows me his Outlook In-box and lets me read a very demanding email from another department in the company.)

2 Answers ∨

Data Scientist at Beachbody was asked...   Apr 6, 2015

How would you like to use data to change the world.

1 Answer ∨

5. Culture Fit

trying to gauge where your interest in data science and in the hiring company come from. Take a look at these examples and think about what your best answer would be, but keep in mind that it's important to be honest with these answers. There's no reason to not be yourself. There are no right answers to these questions, but the best answers are communicated with confidence.

1. Which data scientists do you admire most? Which startups?
   - There are plenty of amazing data scientists to choose from —take a look at this article on top data science influencers for interesting information about some of the top data scientists in the world.
2. What do you think makes a good data scientist?
3. How did you become interested in data science?
4. Give a few examples of "best practices" in data science.
5. What is the latest data science book / article you read? What is the latest data mining conference / webinar / class / workshop / training you attended?
   - If you haven't read a good data science book recently, Springboard compiled a list of the best data science books to read! And check out these data science podcasts.
6. What's a project you would want to work on at our company?
7. What unique skills do you think you'd bring to the team?
8. What data would you love to acquire if there were no limitations?
9. Have you ever thought about creating your own startup? Around which idea / concept?
10. What can your hobbies tell me that your resume can't?
11. What are your top 5 predictions for the next 20 years?
12. What did you do today? Or what did you do this week / last week?
13. If you won a million dollars in the lottery, what would you do with the money?
14. What is one thing you believe that most people do not?
15. What personality traits do you butt heads with?
16. What (outside of data science) are you passionate about?

Examples of similar data science interview questions found on Glassdoor:

6. Problem-Solving

Interviewers will, at some point during the interview process, want to test your problem-solving ability through data science interview questions. Often these tests will be presented as an open-ended question: How would you do X? In general, that X will be a task or problem specific to the company you are applying with. For example, an interviewer at Yelp may ask a candidate how they would create a system to detect fake Yelp reviews.

Some quick tips: Don't be afraid to ask questions. Employers want to test your critical thinking skills—and asking questions that clarify points of uncertainty is a trait that any data scientist should have. Also, if the problem offers an opportunity to show off your white-board coding skills or to create schematic diagrams—use that to your advantage. It shows technical skill, and helps to communicate your thought process through a different mode of communication. Always share your thought process—process is often more important than the results themselves for the interviewer.

1. How would you come up with a solution to identify plagiarism?
2. How many "useful" votes will a Yelp review receive?
3. How do you detect individual paid accounts shared by multiple users?
4. You are about to send a million emails. How do you optimize delivery? How do you optimize response?
5. You have a data set containing 100,000 rows and 100 columns, with one of those columns being our dependent variable for a problem we'd like to solve. How can we quickly identify which columns will be helpful in predicting the dependent variable. Identify two techniques and explain them to me as though I were 5

6. How would you detect bogus reviews, or bogus Facebook accounts used for bad purposes?
   - This is an opportunity to showcase your knowledge of machine learning algorithms; specifically, sentiment analysis and text analysis algorithms. Showcase your knowledge of fraudulent behavior—what are the abnormal behaviors that can typically be seen from fraudulent accounts?
7. How would you perform clustering on a million unique keywords, assuming you have 10 million data points—each one consisting of two keywords, and a metric measuring how similar these two keywords are? How would you create this 10 million data points table in the first place?
8. How would you optimize a web crawler to run much faster, extract better information, and better summarize data to produce cleaner databases?

Examples of similar data science interview questions found on Glassdoor:

Conclusion

There is no single "best" way to prepare for data science interview, but hopefully by reviewing these common interview questions for data scientists you will be able to walk into your interviews well-practiced and confident. If you have any suggestions for questions, let us know! Good luck.

220 shares

Our guide to data science interviews.

A look at 40 artificial intelligence interview questions.

What we learned analyzing hundreds of data science interviews. This also includes a selection of data science interview questions.

Sources

Glassdoor – Data Scientist Interview Questions

KDnuggets

DeZyre

Udacity

Data Science Central – 66 Interview Questions for Data Scientists

AnalyticsVidhya – 40 Interview Questions asked at Startups in Machine Learning/Data Science

Workable – Data Scientist Coding Interview Questions

Codementor – 15 Essential Python Interview Questions

MaxNoy – Coding Interviews

DeZyre – 100 Hadoop Interview Questions and Answers

Tutorials Point – Python Interview Questions

Tutorials Point – SQL Interview Questions

(This post was originally published October 26, 2016. It was last updated November 29, 2018.)

Knowing the interview questions to prepare for is just one part of the interview process. Learn step-by-step everything you need to know to not only land an interview, but ace the data science interview with Springboard's Ultimate Guide to Data Science Interviews.

E-book

Ultimate Guide to Data Science Interviews

Learn what it takes to get more interviews and land a job offer.

Download now

Michael Rundell

Data scientist in training, avid football fan, day-dreamer, UC Davis Aggie, and opponent of the pineapple topping on pizza.

You might also be interested in...

19 Free Public Data Sets for Your Data Science Project

Completing your first project is a major milestone on the road to becoming a data scientist and helps to both reinforce your skills and

READ MORE

Data Mining in Python: A Guide

Data mining and algorithms Data mining is the process of discovering predictive information from the analysis of large

READ MORE

DATA SCIENCE

Data Science Career Paths: Different Roles

Data Science Career Paths: Introduction We've just come out with the first data science bootcamp with a job guarantee to help you

READ MORE

UI/UX Design Bootcamp

UX Bootcamp

Machine Learning Bootcamp

Data Analytics Bootcamp

RESOURCES

Free Learning Paths

E-books and Guides

View All Resources

Jobs

Become a Mentor

Hire Our Students

Corporate Training

Affiliates

Partners

SCHOLARSHIPS

Career Tracks

Veterans

Women In Tech

Like Us on Facebook

Follow Us on Twitter

Read Our Stories on Medium

Springboard

220 shares