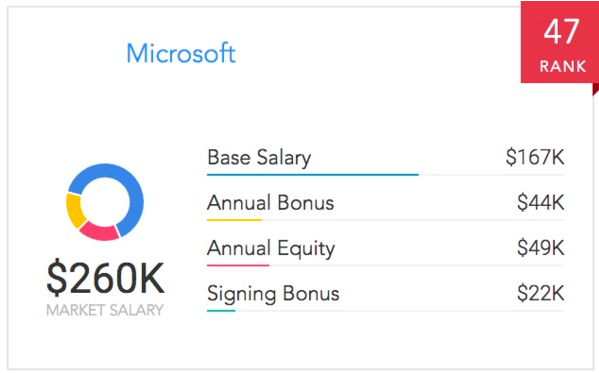
100+ Common Data Science Interview Questions

* Posted by [Amelia Matteson](https://www.datasciencecentral.com/profile/AmeliaMatteson) on July 19, 2017 at 12:00pm
* [View Blog](https://www.datasciencecentral.com/profiles/blog/list?user=1rfgcpmf8pi90)

*This article was written by Michael Rundell.*

**A Curated List of Data Science Interview Questions**

Preparing for an interview is not easy – naturally there is a large amount of uncertainty regarding the data science interview questions you will be asked. No matter how much work experience or technical skill you have, an interviewer can throw you off with a set of questions that you didn’t expect. For a data science interview, an interviewer will ask questions spanning a wide range of topics, requiring strong technical knowledge and communication skills from the part of the interviewee. Your statistics, programming, and data modeling skills will be put to the test through a variety of questions and question styles – intentionally designed to keep you on your feet and force you to demonstrate how you operate under pressure. Preparation is a major key to success when in pursuit of a career in data science career in data science.

[](https://api.ning.com/files/QzmRjEe-2pbvni-FxVw4SCDUlZGi0xJAJJUwJOExOravVpzLqM*dXuwJRPEXBkYPyEh0bmcQncrYbxsVLnUfplFFNQNl*Myx/Capture.PNG)

*Source for picture:*[*click here*](https://www.paysa.com/blog/2017/03/29/5-data-science-job-interview-questions/)

This guide contains all of the data science interview questions an interviewee should expect when interviewing for a position as a data scientist.

We set off to curate, create and edit different data science interview questions and provided answers for some. 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 data science interview questions into six different categories: statistics, programming, modeling, behavior, culture, and problem-solving.

**Table of Contents**

1. Statistics  
2. Programming

* General
* Big Data
* Python
* R
* SQL

3. Modeling  
4. Behavioral  
5. Culture Fit   
6. Problem-Solving

**1. Statistics**

Statistical computing is the process through which data scientists take raw data and create predictions and models backed by the data. Without an advanced knowledge of statistics it is difficult to succeed as a data scientist – 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?
2. What is sampling? How many sampling methods do you know?
3. What is the difference between Type I vs Type II error?
4. What is linear regression? What do the terms P-value, coefficient, R-Squared value mean? What is the significance of each of these components?
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?
7. What is selection bias?
8. What is an example of a dataset with a non-Gaussian distribution?
9. What is the Binomial Probability Formula?

**2. Programming**

To test your programming skills, employers will ask two things during their 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.

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 dataset in (insert language here)?
7. Tell me about the coding you did during your last project?

2.2 *Big Data*

1. What are the two main components of the Hadoop Framework?
2. Explain how MapReduce works as simply as possible.
3. How would you sort a large list of numbers?
4. Here is a big dataset. What is 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. What are the supported data types in Python?
3. What is the difference between a tuple and a list in Python?

2.4 *R*

1. What are the different types of sorting algorithms available in R language? -- *There are insertion, bubble, and selection sorting algorithms.*
2. What are the different data objects in R?
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?
5. What is the command used to store R objects in a file?
6. What is the best way to use Hadoop and R together for analysis?
7. How do you split a continuous variable into different groups/ranks in R?
8. Write a function in R language to replace the missing value in a vector with the mean of that vector.

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 be asked to extract relevant data, filter and order the data as you see fit, and report your findings.

1. What is the purpose of the group functions in SQL? Give some examples of group functions.
2. Group functions are necessary to get summary statistics of a dataset. COUNT, MAX, MIN, AVG, SUM, and DISTINCT are all group functions
3. Tell me the difference between an inner join, left join/right join, and union.
4. What does UNION do? What is the difference between UNION and UNION ALL?
5. What is the difference between SQL and MySQL or SQL Server?
6. 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?

**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 the model you created 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 kNN different from k-means clustering? -- *kNN, or k-nearest neighbors is a classification algorithm, where the k is an integer describing the the number of neighboring data points that influence the classification of a given observation. K-means is a clustering algorithm, where the k is an integer describing the number of clusters to be created from the given data. Both accomplish different tasks.*
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.
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.*
9. Explain the difference between L1 and L2 regularization methods.
10. What is root cause analysis?
11. What are hash table collisions?
12. What is an exact test?
13. In your opinion, which is more important when designing a machine learning model: Model performance? Or model accuracy?
14. What is one way that you would handle an imbalanced dataset 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 5 days developing a 90% accurate solution, or 10 days for 100% accuracy?
19. What are some situations where a general linear model fails?
20. Do you think 50 small decision trees are better than a large one? Why?
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 negatives?

*To read the full original article click*[*here*](https://www.springboard.com/blog/data-science-interview-questions/)*. For more related articles on data science interview questions on DSC click*[*here*](https://www.datasciencecentral.com/page/search?q=data/science/interview/questions)*.*

**DSC Resources**

* Services: [Hire a Data Scientist](http://careers.analytictalent.com/jobs/products) | [Search DSC](https://www.datasciencecentral.com/page/search?q=Python) | [Classifieds](http://classifieds.datasciencecentral.com/) | [Find a Job](http://www.analytictalent.com/)
* Contributors: [Post a Blog](https://www.datasciencecentral.com/profiles/blog/new) | [Ask a Question](https://www.datasciencecentral.com/forum/topic/new)
* Follow us: [@DataScienceCtrl](https://www.twitter.com/datasciencectrl) | [@AnalyticBridge](https://www.twitter.com/analyticbridge)

Popular Articles

* [Difference between Machine Learning, Data Science, AI, Deep Learnin...](https://www.datasciencecentral.com/profiles/blogs/difference-between-machine-learning-data-science-ai-deep-learning)
* [What is Data Science? 24 Fundamental Articles Answering This Question](https://www.datasciencecentral.com/profiles/blogs/20-articles-about-core-data-science)
* [Hitchhiker's Guide to Data Science, Machine Learning, R, Python](https://www.datasciencecentral.com/profiles/blogs/hitchhiker-s-guide-to-data-science-machine-learning-r-python)
* [Advanced Machine Learning with Basic Excel](https://www.datasciencecentral.com/profiles/blogs/advanced-machine-learning-with-basic-excel)