Top 19 Data Science Interview Questions for Beginners

* Posted by [Rahul Dharia](https://www.datasciencecentral.com/profile/RahulDharia) on February 20, 2020 at 2:30am
* [Send Message](https://www.datasciencecentral.com/profiles/blogs/top-17-data-science-interview-questions-for-beginners)   [View Blog](https://www.datasciencecentral.com/profiles/blog/list?user=3c737twvhl12v)

Job interviews make everyone nervous. But this is what they are designed to do. It is the most common medium to assess a candidate’s presence of mind and his/her ability to remain calm and composed in a tense situation.  In order to ace the interview, you need to have in-depth knowledge of the role you are interviewing for and what is expected. Presence of mind and strong subject knowledge assumes added significance when you are preparing for a Data Scientist’s interview as it is definitely going to test your capabilities.

During a Data Science interview, you can expect a variety of questions being asked to test your knowledge of a myriad number of topics ranging from Statistics, Data Analysis, ML, and Deep Learning to Big Data and AI among others.

To help you prepare exceptionally well for the interview, we present here a list of most frequently asked [Data Science interview questions](https://vasitum.com/blog/career-advice/19-data-science-interview-questions-for-professionals/) along with the answers.

So, without further delay, let’s get started!

**#1 What is Data Science?**

Data Science, to put it in the simplest form, is the study of data which is collected from different sources and then stored, organized and analyzed to derive meaningful information from it.

**#2 What knowledge you need to have to extract intended information from raw data?**

Data Scientists and analysts need to have good knowledge of mathematics, statistics computer science, machine learning, data visualization, cluster analysis, and data modelling.

**#3 What is the difference between structured and unstructured data?**

Structured data, as the name suggests, is data that is highly organized and neatly formatted so it's easily searchable in relational databases. Unstructured data on the other hand is data that is not organized or formatted.

**#4 What is the difference between Supervised and Unsupervised ML?**

Supervised ML uses training data set, and the input data is labelled. Unsupervised ML uses the input data set and the input data remains unlabeled. Also, supervised ML is used for prediction, and unsupervised ML is used for analysis.

**#5 What Is Logistic Regression?**

It is a statistical technique used to predict a binary outcome that is either zero or one or a yes or no.

**#6 Why Data Cleansing is of critical importance in Data Analysis?**

Data is accumulated from a variety of sources. It is important to ensure that the data collected is good enough for analysis. Data cleansing makes sure that data is complete and accurate and devoid of redundant or irrelevant components.

**#7 What is the Binomial Probability Formula?**

[](https://storage.ning.com/topology/rest/1.0/file/get/3874519461?profile=original)

**#8 What is a Linear Regression?**

Linear Regression is a statistical tool employed for quick predictive analysis. In linear regression, the score of a variable (say A) is predicted from what is the score of another variable (say B). For example, the price of a house which depends on factors such as its size or location.

**#9 What Are Feature Vectors?**

Feature Vectors are n-dimensional vectors of numerical features that represent some object.

**# 10 What does A/b Testing aim to achieve?**

It is a statistical hypothesis testing used to detect any changes to the web page so that steps can be taken to maximize the possibility of the desired outcome.

**#11 What Is the Law Of Large Numbers?**

It is a theorem that deals with the results produced when the same experiment is performed multiple times.

*We relied on the Law of Large Numbers when performing Monte Carlo Simulation*

*From Wikipedia: In probability theory, the law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed.*

**#12 What is sampling?**

Data sampling is a statistical analysis technique used to select and analyze a representative subset of data to determine patterns in a larger set.

**#13 What is the difference between SQL and MySQL or SQL Server?**

SQL stands for Structured Query Language. It’s a standard language employed to assess and manipulate databases. MySQL is a database management system, such as SQL Server, Oracle, etc.

**# 14 How can you eliminate duplicate rows from a query result?**

One way you can eliminate duplicate rows from a query result is with the DISTINCT clause.

**# 15 What Are The Types Of Biases That Can Occur During Sampling?**

* Selection bias
* Under coverage bias
* Survivorship bias
* What are Artificial Neural Networks?

**# 16 What are Artificial Neural Networks (ANN)?**

Artificial Neural Networks (ANN) are computing systems designed to simulate the human brain.

**# 17 Python or R – Which is better for text analytics?**

Python is a better option for text analytics because of its Pandas library that includes user-friendly data structures and solid data analysis tools. R is a better option for Machine Learning applications.

**#18. How will you define the number of clusters in a clustering algorithm?**

The primary purpose of clustering is to group together similar identities in a way so as the entities within the group remain same but the groups remain dissimilar to one another.

**# 19. What is TF/IDF vectorization?**

tf–idf stands for term frequency–inverse document frequency. It is a numerical statistic that is used to determine the importance of a word in a document in a collection or corpus.