DATA SCIENCE

Data Science Interview Questions and Answers for 2021

by Rashmi Karan No Comments

May 10 2021

Data has now become the raw material for business and the vast amounts of structured and unstructured information are used to create a new form of economic value for businesses of every size. In this sense, Data Science is precisely the field of study where collective processes, theories, and technologies are combined that allow the review, analysis, and extraction of valuable knowledge and information from hard data. Now with the market growing leaps and bounds, there is a significant dearth of skilled data scientists, who can help businesses sift through an overabundance of data and come up with meaningful insights.

So if you are planning to move on the path to becoming a data scientist, you need to prepare well and create a fabulous impression on your prospective employers with your knowledge. This write-up brings you some important data science interview questions and answers to help you crack your data science interview.

To learn more about data science, read our blog on – What is data science?

The article is segmented on different data science topics–

Basic Data Science Interview Questions

Q1. What is the difference between data science and big data?

Ans. The common differences between data science and big data are –

Big Data	Data Science
Large collection of data sets that cannot be stored in a traditional system	An interdisciplinary field that includes analytical aspects, statistics, data mining, machine learning, etc.
Popular in the field of communication, purchase and sale of goods, financial services, and educational sector	Common application are digital advertising, web research, recommendation systems (Netflix, Amazon, Facebook), speech and handwriting recognition applications
Big Data solves problems related to data management and handling, and analyze insights resulting in informed decision making	Data Science uses machine learning algorithms and statistical methods to obtain accurate predictions from raw data
Popular tools are Hadoop, Spark, Flink, NoSQL, Hive, etc.	Popular tools are Python, R, SAS, SQL, etc.

You may also be interested in exploring:

Popular Data Science Basics Online Courses & Certifications	Popular Machine Learning Online Courses & Certifications
Popular Statistics for Data Science Online Courses & Certifications	Popular Python for data science Online Courses & Certifications

Q2. How do you check for data quality?

Ans. Some of the definitions used to check for data quality are:

- Completeness
- Consistency
- Uniqueness
- Integrity
- Conformity
- Accuracy

Q3. Suppose you are given survey data, and it has some missing data, how would you deal with missing values from that survey?

Ans. There are two main techniques for dealing with missing values –

- Debugging Techniques – It is a Data Cleaning process consisting of evaluating the quality of the information collected, increasing its quality, in order to avoid lax analysis. The most popular debugging techniques are –

Searching the list of values: It is about searching the data matrix for values that are outside the response range. These values can be considered as missing, or the correct value can be estimated from other variables

Filtering questions: It is about comparing the number of responses of a filter category and another filtered category. If any anomaly is observed that cannot be solved, it will be considered as a lost value.

Checking for Logical Consistencies: The answers that may be considered contradictory to each other are checked.

Counting the Level of representativeness: A count is made of the number of responses obtained in each variable. If the number of unanswered questions is very high, it is possible to assume equality between the answers and the non-answers or to make an imputation of the non-answer.

- Imputation Technique

This technique consists of replacing the missing values with valid values or answers by estimating them. There are three types of imputation:

- Random imputation
- Hot Deck imputation
- Imputation of the mean of subclasses

Q4. How would you deal with missing random values from a data set?

Ans. There are two forms of randomly missing values:

MCAR or Missing completely at random. Such errors happen when the missing values are randomly distributed across all observations.

We can confirm this error by partitioning the data into two parts –

1. One set with the missing values
2. Another set with the non-missing values.

After we have partitioned the data, we conduct a t-test of mean difference to check if there is any difference in the sample between the two data sets.

In case the data are MCAR, we may choose a pair-wise or a list-wise deletion of missing value cases.

MAR or Missing at random. It is a common occurrence. Here, the missing values are not randomly distributed across observations but are distributed within one or more sub-samples. We cannot predict the probability from the variables in the model. Data imputation is mainly performed to replace them.

Q5. What is Hadoop, and why should I care?

Ans. Hadoop is an open-source processing framework that manages data processing and storage for big data applications running on pooled systems.

Apache Hadoop is a collection of open-source utility software that makes it easy to use a network of multiple computers to solve problems involving large amounts of data and computation. It provides a software framework for distributed storage and big data processing using the MapReduce programming model.

Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packets of code to nodes to process the data in parallel. This allows the data set to be processed faster and more efficiently than if conventional supercomputing architecture were used.

Q6. What is ‘fsck’?

Ans. ‘fsck ‘ abbreviation for ‘ file system check.’ It is a type of command that searches for possible errors in the file. fsck generates a summary report, which lists the file system’s overall health and sends it to the Hadoop distributed file system.

Q7. Which is better – good data or good models?

Ans. This might be one of the frequently asked data science interview questions.

The answer to this question is very subjective and depends on the specific case. Big companies prefer good data; it is the foundation of any successful business. On the other hand, good models couldn't be created without good data.

Based on your personal preference, you will probably choose no right or wrong answer (unless the company requires one specifically).

Q8. What are Recommender Systems?

Ans. Recommender systems are a subclass of information filtering systems, used to predict how users would rate or score particular objects (movies, music, merchandise, etc.). Recommender systems filter large volumes of information based on the data provided by a user and other factors, and they take care of the user's preference and interest.

Recommender systems utilize algorithms that optimize the analysis of the data to build the recommendations. They ensure a high level of efficiency as they can associate elements of our consumption profiles such as purchase history, content selection and even our hours of activity, to make accurate recommendations.

To know more about the job profile and responsibilities of a Data Scientist, refer to this article on What is Data Scientist?

Q9. What are the different types of Recommender Systems?

Ans. There are three main types of Recommender systems.

Collaborative filtering – Collaborative filtering is a method of making automatic predictions by using the recommendations of other people. There are two types of collaborative filtering techniques –

- User-User collaborative filtering
- Item-Item collaborative filtering

Content-Based Filtering– Content-based filtering is based on the description of an item and a user's choices. As the name suggests, it uses content (keywords) to describe the items, and the user profile is built to state the type of item this user likes.

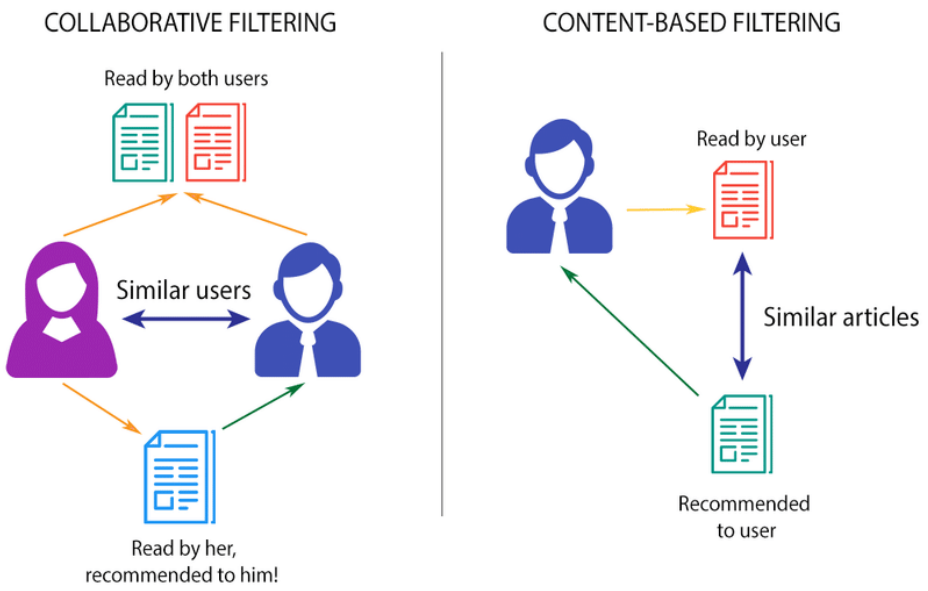


Image – Collaborative filtering & Content-based filtering (Source)

Hybrid Recommendation Systems – Hybrid Recommendation engines are a combination of diverse rating and sorting algorithms. A hybrid recommendation engine can recommend a wide range of products to consumers as per their history and preferences with precision.

Q10. Differentiate between wide and long data formats.

Ans. In a wide format, categorical data are always grouped.

The long data format is in which there are a number of instances with many variables and subject variables.

Q11. What are Interpolation and Extrapolation?

Ans. Interpolation – This is the method to guess data points between data sets. It is a prediction between the given data points.

Extrapolation – This is the method to guess data point beyond data sets. It is a prediction beyond given data points.

Also Read>>Skills That Employers Look For In a Data Scientist

Q12. How much data is enough to get a valid outcome?

Ans. All the businesses are different and measured in different ways. Thus, you never have enough data and there will be no right answer. The amount of data required depends on the methods you use to have an excellent chance of obtaining vital results.

Q13. What is the difference between 'expected value' and 'average value'?

Ans. When it comes to functionality, there is no difference between the two. However, they are used in different situations.

An expected value usually reflects random variables, while the average value reflects the population sample.

Q14. What happens if two users access the same HDFS file at the same time?

Ans. This is a bit of a tricky question. The answer itself is not complicated, but it is easy to confuse by the similarity of programs' reactions.

When the first user is accessing the file, the second user's inputs will be rejected because HDFS NameNode supports exclusive write.

Q15. What is power analysis?

Ans. Power analysis allows the determination of the sample size required to detect an effect of a given size with a given degree of confidence.

Q16. Is it better to have too many false negatives or too many false positives?

Ans. This question will depend on how you show your viewpoint. Give examples

These are some of the popular data science interview questions. Always be prepared to answer all types of data science interview questions— technical skills, interpersonal, leadership, or methodologies. If you are someone who has recently started your career in Data Science, you can always get certified to improve your skills and boost your career opportunities.

Statistics Interview Questions

Q17. What is the importance of statistics in data science?

Ans. Statistics help data scientists to get a better idea of a customer's expectations. Using statistical methods, data Scientists can acquire knowledge about consumer interest, behavior, engagement, retention, etc. It also helps to build robust data models to validate certain inferences and predictions.

Q18. What are the different statistical techniques used in data science?

Ans. There are many statistical techniques used in data science, including –

The arithmetic mean – It is a measure of the average of a set of data

Graphic display – Includes charts and graphs to visually display, analyze, clarify, and interpret numerical data through histograms, pie charts, bars, etc.

Correlation – Establishes and measures relationships between different variables

Regression – Allows identifying if the evolution of one variable affects others

Time series – It predicts future values by analyzing sequences of past values

Data mining and other Big Data techniques to process large volumes of data

Sentiment analysis – It determines the attitude of specific agents or people towards an issue, often using data from social networks

Semantic analysis – It helps to extract knowledge from large amounts of texts

A / B testing – To determine which of two variables works best with randomized experiments

Machine learning using automatic learning algorithms to ensure excellent performance in the presence of big data

Check Out Our Data Science Courses

Q19. What is an RDBMS? Name some examples for RDBMS?

Ans. This is among the most frequently asked data science interview questions.

A relational database management system (RDBMS) is a database management system that is based on a relational model.

Some examples of RDBMS are MS SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access.

Q20. What is a Z test, Chi-Square test, F test, and T-test?

Ans. Z test is applied for large samples. Z test = (Estimated Mean – Real Mean)/ (square root real variance / n).

Chi-Square test is a statistical method assessing the goodness of fit between a set of observed values and those expected theoretically.

F-test is used to compare 2 populations' variances. F = explained variance/unexplained variance.

T-test is applied for small samples. T-test = (Estimated Mean – Real Mean)/ (square root Estimated variance / n).

Q21. What does P-value signify about the statistical data?

Ans. The p-value is the probability for a given statistical model that, when the null hypothesis is true, the statistical summary would be the same as or more extreme than the actual observed results.

When,

P-value>0.05, it denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.

P-value <= 0.05 denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.

P-value=0.05is the marginal value indicating it is possible to go either way

Q22. Differentiate between univariate, bivariate, and multivariate analysis.

Ans. Univariate analysis is the simplest form of statistical analysis where only one variable is involved.

Bivariate analysis is where two variables are analyzed and in multivariate analysis, multiple variables are examined.

Q23. What is association analysis? Where is it used?

Ans. Association analysis is the task of uncovering relationships among data. It is used to understand how the data items are associated with each other.

Also Read – Top 6 Industries Hiring Data Scientists in 2021

Q24. What is the difference between squared error and absolute error?

Ans. Squared error measures the average of the squares of the errors or deviations—that is, the difference between the estimator and what is estimated.

Absolute error is the difference between the measured or inferred value of a quantity and its actual value.

Q25. What is an API? What are APIs used for?

Ans. API stands for Application Program Interface and is a set of routines, protocols, and tools for building software applications.

With API, it is easier to develop software applications.

Q26. What is Collaborative filtering?

Ans. Collaborative filtering is a method of making automatic predictions by using the recommendations of other people.

Q27. Why do data scientists use combinatorics or discrete probability?

Ans. It is used because it is useful in studying any predictive model.

Also Read>>How are Data Scientist and Data Analyst different?

Q28. What do you understand by Recall and Precision?

Ans. Precision is the fraction of retrieved instances that are relevant, while Recall is the fraction of relevant instances that are retrieved.

Become Machine Learning Expert Now>>

Q29. What is market basket analysis?

Ans. Market Basket Analysis is a modeling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items.

Q30. What is the central limit theorem?

Ans. The central limit theorem states that the distribution of an average will tend to be Normal as the sample size increases, regardless of the distribution from which the average is taken except when the moments of the parent distribution do not exist.

Q31. Explain the difference between type I and type II error.

Ans. Type I error is the rejection of a true null hypothesis or false-positive finding, while Type II error is the non-rejection of a false null hypothesis or false-negative finding.

Q32. What is Linear Regression?

Ans. Linear regression is the most popular type of predictive analysis. It is used to model the relationship between a scalar response and explanatory variables.

Q33. What are the limitations of a Linear Model/Regression?

Ans.
• Linear models are limited to linear relationships, such as dependent and independent variables
• Linear regression looks at a relationship between the mean of the dependent variable and the independent variables, and not the extremes of the dependent variable
• Linear regression is sensitive to univariate or multivariate outliers
• Linear regression tend to assume that the data are independent

Q34. What is the goal of A/B Testing?

Ans. A/B testing is a comparative study, where two or more variants of a page are presented before random users and their feedback is statistically analyzed to check which variation performs better.

Q35. What is the main difference between overfitting and underfitting?

Ans. Overfitting – In overfitting, a statistical model describes any random error or noise, and occurs when a model is super complex. An overfit model has poor predictive performance as it overreacts to minor fluctuations in training data.
Underfitting – In underfitting, a statistical model is unable to capture the underlying data trend. This type of model also shows poor predictive performance.

Q36. What is a Gaussian distribution and how it is used in data science?

Ans. Gaussian distribution or commonly known as bell curve is a common probability distribution curve. Mention the way it can be used in data science in a detailed manner.

Q37. Explain the purpose of group functions in SQL. Cite certain examples of group functions.

Ans. Group functions provide summary statistics of a data set. Some examples of group functions are –
a) COUNT
b) MAX
c) MIN
d) AVG
e) SUM
f) DISTINCT

Q38. What is Root Cause Analysis?

Ans. Root Cause is defined as a fundamental failure of a process. To analyze such issues, a systematic approach has been devised that is known as Root Cause Analysis (RCA). This method addresses a problem or an accident and gets to its “root cause”.

Q39. What is the difference between a Validation Set and a Test Set?

Ans. The validation set is used to minimize overfitting. This is used in parameter selection, which means that it helps to verify any accuracy improvement over the training data set.
Test Set is used to test and evaluate the performance of a trained Machine Learning model.

Q40. What is the Confusion Matrix?

Ans. Confusion Matrix describes the performance of any classification model. It is presented in the form of a table with 4 different combinations of predicted and actual values.

Q41. What is the p-value?

Ans. A p-value helps to determine the strength of results in a hypothesis test. It is a number between 0 and 1 and Its value determines the strength of the results.

Q42. What is the difference between Causation and Correlation?

Ans. Causation denotes any causal relationship between two events and represents its cause and effects.
Correlation determines the relationship between two or more variables.
Causation necessarily denotes the presence of correlation, but correlation doesn’t necessarily denote causation.

Q43. What is cross-validation?

Ans. Cross-validation is a technique to assess the performance of a model on a new independent dataset. One example of cross-validation could be – splitting the data into two groups –

training and testing data, where you use the testing data to test the model and training data to build the model.

Q44. What do you mean by logistic regression?

Ans. Also known as the logit model, logistic regression is a technique to predict the binary result from a linear amalgamation of predictor variables.

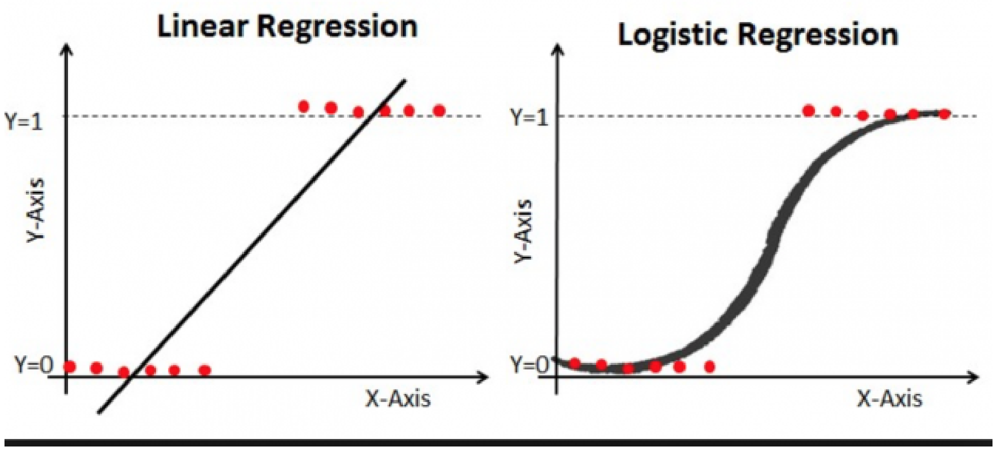


Fig – Linear Regression vs. Logistic Regression

Source

Q45. What is ‘cluster sampling’?

Ans. Cluster sampling is a probability sampling technique where the researcher divides the population into separate groups, called clusters. Then a simple cluster sample is selected from the population. The researcher conducts his analysis of data from the sample pools.

Q46. What happens if two users access the same HDFS file at the same time?

Ans. This is a bit of a tricky question. The answer itself is not complicated, but it is easy to confuse by the similarity of programs’ reactions.

When the first user is accessing the file, the second user’s inputs will be rejected because HDFS NameNode supports exclusive write.

Q47. What are the Resampling methods?

Ans. Resampling methods are used to estimate the precision of the sample statistics, exchanging labels on data points, and validating models.

Q48. What is selection bias, and how can you avoid it?

Ans. Selection bias is an experimental error that occurs when the participant pool, or the subsequent data, is not representative of the target population.

Selection biases cannot be overcome with statistical analysis of existing data alone, though Heckman correction may be used in special cases.

Q49. What is the binomial distribution?

Ans. A binomial distribution is a discrete probability distribution that describes the number of successes when conducting independent experiments on a random variable.

$$P(k;n,p) = \binom{n}{k} p^{k} (1-p)^{n-k}$$

Formula –

Where:

n = Number of experiments

x = Number of successes

p = Probability of success

q = Probability of failure (1-p)

Q50. What is covariance in statistics?

Ans. Covariance is a measure of the joint variability of two random variables. The covariance between two variables x and y can be calculated as follows:

$$\text{Cov}\ (X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

Where:

- Xi – the values of the X-variable
- Yj – the values of the Y-variable
- $\bar{X}$ – the mean (average) of the X-variable
- $\bar{Y}$ – the mean (average) of the Y-variable
- n – the number of data points

Q51. What is Root Cause Analysis?

Ans. Root Cause Analysis (RCA) is the process of uncovering the root causes of problems to identify appropriate solutions. The RCA assumes that it is much more useful to systematically prevent and resolve underlying issues than just treating symptoms ad hoc and putting out fires.

Q52. What is Correlation Analysis?

Ans. Correlation Analysis is a statistical method to evaluate the strength of the relationship between two quantitative variables. It consists of autocorrelation coefficients, estimated and calculated to make a different spatial relationship. It is used to correlate data based on distance.

Q53. What is imputation? List the different types of imputation techniques.

Ans. Imputation is the process that allows you to replace missing data with other values. Types of imputation techniques include –

Single Imputation: Single imputation denotes that the missing value is replaced by a value.

Hot-deck: The missing value is imputed from a similar register, which is chosen at random, based on a punched card.

Cold deck Imputation: Select donor data from other sets.

Mean Imputation: Substitute the stored value for the mean of that variable in other cases.

Mean Imputation: Its purpose is to replace the missing value with predicted values of a variable that is based on others.

Stochastic Regression: equal to the regression, but adds the mean regression variance to the regression imputation.

Multiple Imputation: It is a general approach to the problem of missing data, available in commonly used statistical packages. Unlike single imputation, Multiple Imputation estimates the values multiple times.

Q54. What is the difference between a bar graph and a histogram?

Ans. Bar charts and histograms can be used to compare the sizes of the different groups. A bar chart is made up of bars plotted on a chart. A histogram is a graph that represents a frequency distribution; the heights of the bars represent observed frequencies.

In other words, a histogram is a graphical display of data using bars of different heights. Generally, there is no space between adjacent bars.

Bar Charts

- The columns are placed on a label that represents a categorical variable.
- The height of the column indicates the size of the group defined by the categories

Histogram

- The columns are placed on a label that represents a quantitative variable.
- The column label can be a single value or a range of values.

In bar charts, each column represents a group defined by a categorical variable; and with histograms, each column represents a group defined by a quantitative variable.

Q55. Name some of the prominent resampling methods in data science.

Ans. The Bootstrap, Permutation Tests, Cross-validation, and Jackknife.

Q56. What is an Eigenvalue and Eigenvector?

Ans. Eigenvectors are used for understanding linear transformations.

Eigenvalue can be referred to as the strength of the transformation in the direction of the eigenvector or the factor by which the compression occurs.

Q57. Which technique is used to predict categorical responses?

Ans. Classification techniques are used to predict categorical responses.

Programming Language Interview Questions

Q58. Which would you prefer – R or Python?

Ans. One of the most important data science interview questions.

Both R and Python have their own pros and cons. R is mainly used when the data analysis task requires standalone computing or analysis on individual servers. Python, when your data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database.

Read More – What is Python?

Q59. Which package is used to do data import in R and Python? How do you do data import in SAS?

Ans. In R, RODBC is used for RDBMS data and data.table for fast-import.

In SAS, data and sas7bdat are used to import data.

In Python, Pandas package and the commands read_csv, read_sql are used for reading data.

Q60. What packages are used for data mining in Python and R?

Ans. There are various packages in Python and R:

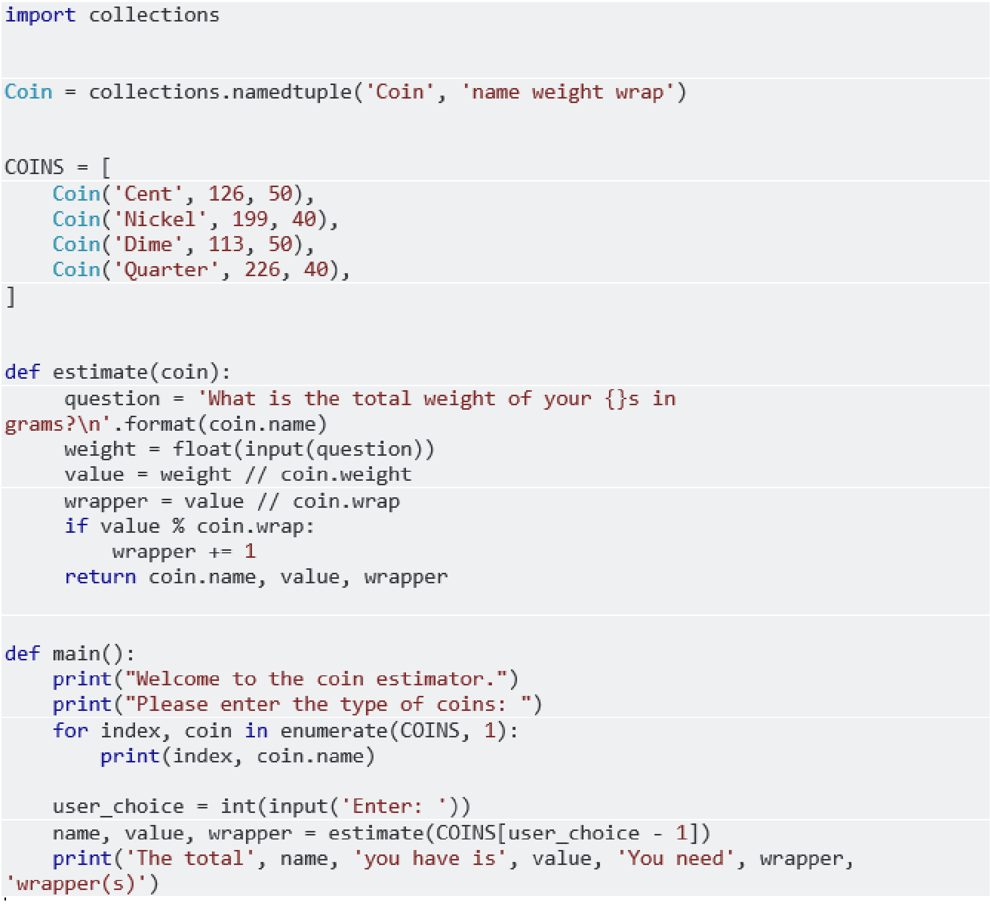
Python – Orange, Pandas, NLTK, Matplotlib, and Scikit-learn are some of them.

R – Arules, tm, Forecast and GGPlot are some of the packages.

Explore – Python Online Courses & Certifications

Q61. Write a program in Python that takes input as the weight of the coins and produces output as the money value of the coins.

Ans. Here is an example of the code. You can change the values.



```
import collections

Coin = collections.namedtuple('Coin', 'name weight wrap')

COINS = [
    Coin('Cent', 126, 50),
    Coin('Nickel', 199, 40),
    Coin('Dime', 113, 50),
    Coin('Quarter', 226, 40),
]

def estimate(coin):
     question = 'What is the total weight of your {}s in
grams?\n'.format(coin.name)
     weight = float(input(question))
     value = weight // coin.weight
     wrapper = value // coin.wrap
     if value % coin.wrap:
         wrapper += 1
     return coin.name, value, wrapper

def main():
    print("Welcome to the coin estimator.")
    print("Please enter the type of coins: ")
    for index, coin in enumerate(COINS, 1):
        print(index, coin.name)

    user_choice = int(input('Enter: '))
    name, value, wrapper = estimate(COINS[user_choice - 1])
    print('The total', name, 'you have is', value, 'You need', wrapper,
'wrapper(s)')
```

Q62. Why does Python score high over other programming languages?

Ans. Python has a wealth of data science libraries; it is incredibly fast and easy to read and learn. The Python suite specializing in deep learning and other machine learning libraries includes popular tools such as sci-kit-learn, Keras, and TensorFlow, which allow data scientists to develop sophisticated data models directly integrated into a production system.

To discover data revelations, you will need to use Pandas, the data analysis library for Python. It can handle large amounts of data without the lag of Excel. You can do numerical modeling analysis with Numpy, do scientific computation and calculation with SciPy, and access many powerful machine learning algorithms with the Sci-Kit-learn code library. With the Python API and the iPython Notebook that comes with Anaconda, you will have robust options to visualize your data.

Q63. What are the data types used in Python?

Ans. Python has the following built-in data types:

- Number (float, integer)
- String
- Tuple
- List
- Set
- Dictionary

Numbers, strings, and tuples are immutable data types, which means that they cannot be modified at run time. Lists, sets, and dictionaries are mutable, which means they can be modified at run time.

Q64. What is a Python dictionary?

Ans. A dictionary is one of the built-in data types in Python. Defines a messy mapping of unique keys to values. Dictionaries are indexed by keys, and the values can be any valid Python data type (even a user-defined class). It should be noted that dictionaries are mutable, which means that they can be modified. A dictionary is created with braces and is indexed using bracket notation.

Q65. What libraries do data scientists use to plot data in Python?

Ans. Matplotlib is the main library used to plot data in Python. However, graphics created with this library need a lot of tweaking to make them look bright and professional. For that reason, many data scientists prefer Seaborn, which allows you to create attractive and meaningful charts with just one line of code.

Q66. Explain the difference between lists and tuples.

Ans. Both lists and tuples are made up of elements, which are values of any Python data type. However, these data types have a number of differences:

Lists are mutable, while tuples are immutable.

Lists are created in brackets (for example, my_list = [a, b, c]), while tuples are in parentheses (for example, my_tuple = (a, b, c)).

Lists are slower than tuples.

Q67. What are lambda functions?

Ans. Lambda functions are anonymous functions in Python. They are very useful when you need to define a function that is very short and consists of a single expression. So instead of formally defining the little function with a specific name, body, and return statement, you can write everything in a short line of code using a lambda function.

Q68. What is PyTorch?

Ans. PyTorch is a Python-based scientific computing package designed to perform numerical calculations using the programming of tensors. It also allows its execution on GPU to speed up calculations. PyTorch is used to replace NumPy and process calculations on GPUs and for research and development in the field of machine learning, mainly focused on the development of neural networks.

PyTorch is designed to seamlessly integrate with Python and its popular libraries like NumPy and is easier to learn than other Deep Learning frameworks. PyTorch has a simple Python interface, provides a simple but powerful API, and provides the ability to run models in a production environment, making it a popular deep learning framework.

Q69. What are the alternatives to PyTorch?

Ans. Some of the best-known alternatives to PyTorch are –

Tensorflow – Google Brain Team developed Tensorflow, which is a free software designed for numerical computation using graphs.

Caffe – Caffe is a machine learning framework designed with the aim of being used in computer vision or image classification. Caffe is popular for its library of training models that do not require any extra implementation.

Microsoft CNTK – Microsoft CNTK is the free software framework developed by Microsoft. It is very popular in the area of speech recognition although it can also be used for other fields such as text and images.

Theano – Theano is another python library. It helps to define, optimize and evaluate mathematical expressions that involve calculations with multidimensional arrays.

Keras – Keras is a high-level API for developing neural networks written in Python. It uses other libraries internally such as Tensorflow, CNTK, and Theano. It was developed to facilitate and speed up the development and experimentation with neural networks.

Machine Learning Interview Questions

Must Read – What is Machine Learning?

Q70. What is the main difference between supervised and unsupervised machine learning?

Ans. Supervised learning includes training labeled data for a range of tasks such as data classification, while unsupervised learning does not require explicitly labeling data.

Q71. What is Deep Learning?

Ans. It is among the most frequently asked data science interview questions. Deep Learning is an artificial intelligence function used in decision making. Deep Learning imitates the human brain's functioning to process the data and create the patterns used in decision-making. Deep learning is a key technology behind automated driving, automated machine translation, automated game playing, object classification in photographs, and automated handwriting generation, among others.

Read More – What is Deep Learning?

Q72. Name different Deep Learning Frameworks.

Ans.
a) Caffe
b) Chainer
c) Pytorch
d) TensorFlow
e) Microsoft Cognitive Toolkit
f) Keras

Also Explore – Machine Learning Online Courses & Certifications

Q73. What are the various types of classification algorithms?

Ans. There are 7 types of classification algorithms, including –
a) Linear Classifiers: Logistic Regression, Naive Bayes Classifier
b) Nearest Neighbor
c) Support Vector Machines
d) Decision Trees
e) Boosted Trees
f) Random Forest
g) Neural Networks

Q74. What is Gradient Descent?

Ans. Gradient Descent is a popular algorithm used for training Machine Learning models and find the values of parameters of a function (f), which helps to minimize a cost function.

Q75. What is Regularization and what kind of problems does regularization solve?

Ans. Regularization is a technique used in an attempt to solve the overfitting problem in statistical models.

It helps to solve the overfitting problem in machine learning.

Q76. What is a Boltzmann Machine?

Ans. Boltzmann Machines have a simple learning algorithm that helps to discover interesting features in training data. These machines represent complex regularities and are used to optimize the weights and the quantity for the problems.

Q77. What is hypothesis testing?

Ans. Hypothesis testing is an important aspect of any testing procedure in Machine Learning or Data Science to analyze various factors that may have any impact on the outcome of the experiment.

Q78. What is Pattern Recognition?

Ans. Pattern recognition is the process of data classification that includes pattern recognition and identification of data regularities. This methodology involves the extensive use of machine learning algorithms.

Q79. Where can you use Pattern Recognition?

Ans. Pattern Recognition has multiple usabilities, across-

- Bio-Informatics
- Computer Vision
- Data Mining
- Informal Retrieval
- Statistics
- Speech Recognition

Q80. What is an Autoencoder?

Ans. These are feedforward learning networks where the input is the same as the output. Autoencoders reduce the number of dimensions in the data to encode it while ensuring minimal error and then reconstruct the output from this representation.

Also Explore – Deep Learning Online Courses & Certifications

Q81. What is the bias-variance trade-off?

Ans. Bias – Bias is the difference between the average prediction of a model and the correct value we are trying to predict.

Variance – Variance is the variability of model prediction for a given data point or a value that tells us the spread of our data.

Models with high variance focus on training data and such models perform very well on training data. On the other hand, a model with high bias doesn't focus on training data and oversimplifies the model, leading to increased training and test data error.

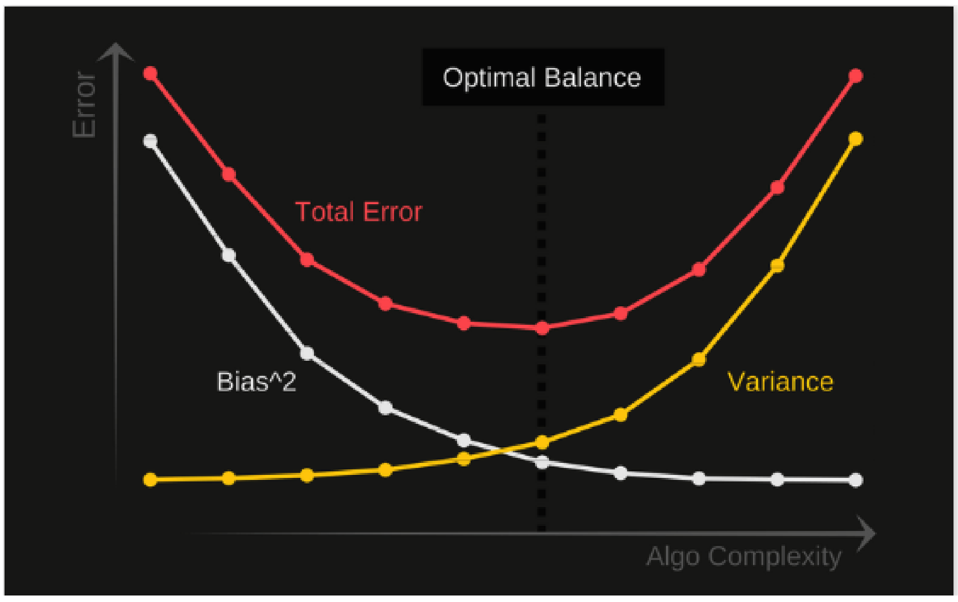


Fig – Optimal balance – Bias vs. Variance (Source – towardsdatascience.com)

Q82. When do you need to update the algorithm in Data science?

Ans. You need to update an algorithm in the following situation:

- You want your data model to evolve as data streams using infrastructure
- The underlying data source is changing
- If it is non-stationarity

Q83. Why should you perform dimensionality reduction before fitting an SVM?

Ans. These SVMs tend to perform better in reduced space. If the number of features is large as compared to the number of observations, then we should perform dimensionality reduction before fitting an SVM.

Q84. Name the different kernels of SVM.

Ans. There are nine types of kernels in SVM.

- Polynomial kernel
- Gaussian kernel
- Gaussian radial basis function (RBF)
- Laplace RBF kernel
- Hyperbolic tangent kernel
- Sigmoid kernel
- Bessel function of the first kind Kernel
- ANOVA radial basis kernel
- Linear splines kernel in one-dimension

Q85. What is the Hierarchical Clustering Algorithm?

Ans. Hierarchical grouping algorithm combines and divides the groups that already exist, in this way they create a hierarchical structure that presents the order in which the groups are split or merged.

Q86. What is 'Power Analysis'?

Ans. Power Analysis is a type of analysis used to determine what kind of effect a unit will have based simply on its size. Power Analysis can be used to estimate the minimum sample size required for an experiment and is directly related to hypothesis testing. The primary purpose underlying power analysis is to help the investigator determine the smallest sample size that is adequate to detect the effect of a certain test at the desired level of significance.

Q87. Have you contributed to any open source project?

Ans. This question seeks a continuous learning mindset. It also tells the interviewer that a candidate is curious and how well they work as a team. Good data scientists are collaborative people, sharing new ideas, knowledge, and information with each other to keep up with rapidly changing data science.

You must say specifically which projects you have worked on and what was their objective. A good answer would also include what you have learned from participating in open source projects.

Q88. How to deal with unbalanced data?

Ans. Machine learning algorithms don't work well with imbalanced data. We can handle this data in a number of ways –

- Using appropriate evaluation metrics for model generated using imbalanced data
- Resampling the training set through undersampling and oversampling
- Properly applying cross-validation while using the over-sampling method to address imbalance problems
- Using more data, primarily by ensembling different resampled datasets
- Resampling with different ratios, where the best ratio majorly depends on data and models used
- Clustering the abundant class
- Designing your own models and be creative in using different techniques and approaches to get the best outcome

Q89. How will you recover the information from a given data set? What are the most common issues in the process of information retrieval?

Ans. The recovery process is carried out through queries to the database where the structured information is stored, using a suitable interrogation language. It is necessary to take into account the key elements that allow the search to be carried out, determining a greater degree of relevance and precision, such as indexes, keywords, thesauri, and the phenomena that can occur in the process such as noise and documentary silence.

One of the most common problems that arise when searching for information is whether what we retrieve is "a lot or a little", that is, depending on the type of search, a multitude of documents or simply a very small number can be retrieved. This phenomenon is called Silence or Documentary Noise.

Documentary silence – These are the documents stored in the database but are unrecovered, because the search strategy has been too specific or because the keywords used are not adequate to define the search.

Documentary noise – These are the document recovered by the system but are irrelevant. This usually happens when the search strategy has been defined as too generic.

Q90. What is Big Data?

Ans. Big Data is a set of massive data, a collection of huge in size and exponentially growing data, that cannot be managed, stored, and processed by traditional data management tools.

To learn more about Big Data, read our blog – What is Big Data?

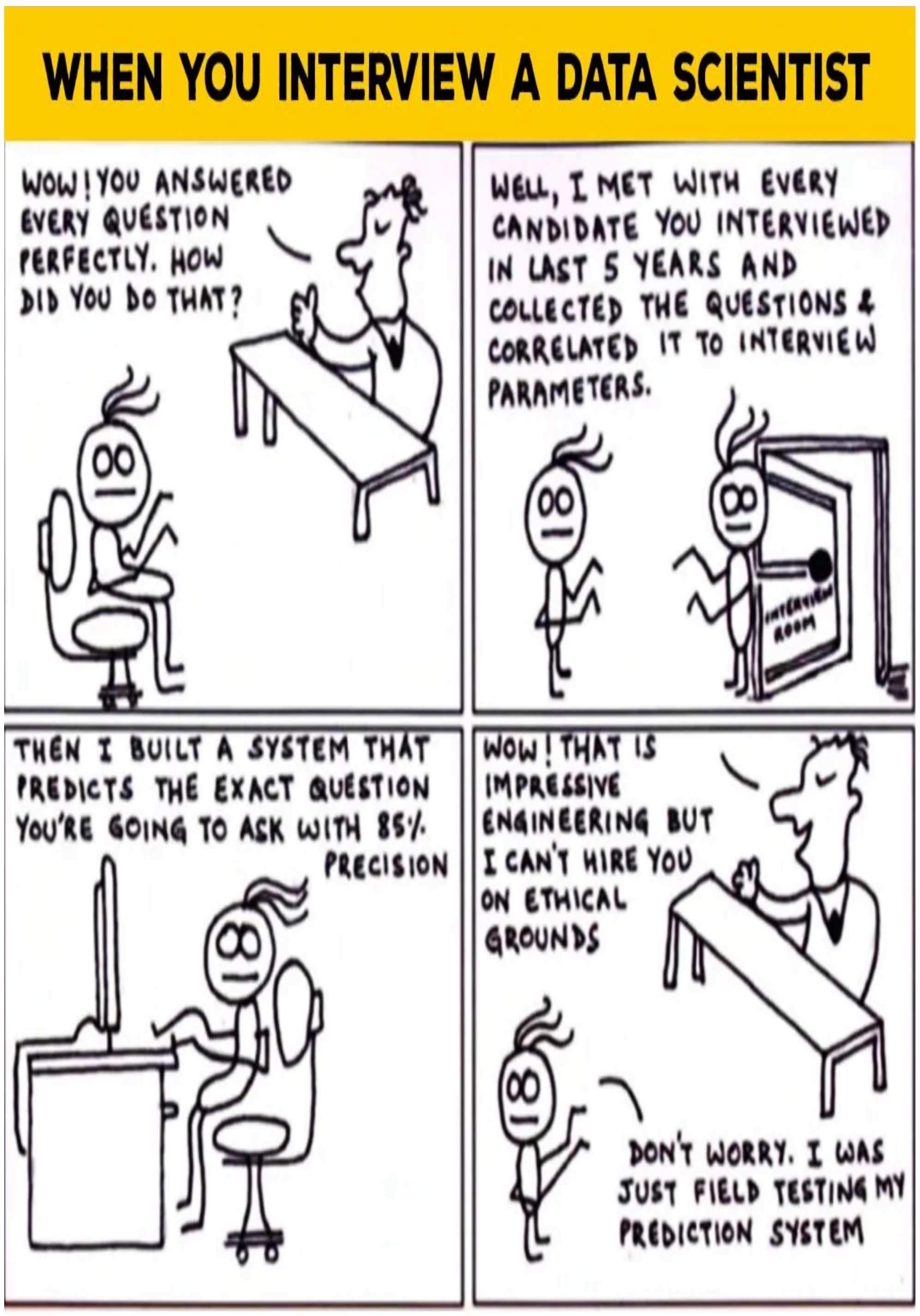
Q91. What are some of the important tools used in Big Data analytics?

Ans. The important Big Data analytics tools are –

- NodeXL
- KNIME
- Tableau
- Solver
- OpenRefine
- Rattle GUI
- Qlikview

We hope these data science interview questions would help you crack your next interview. Always go well-prepared and be ready to share your experience of working on different projects. All the best!

On a lighter note –


WHEN YOU INTERVIEW A DATA SCIENTIST
WOW! YOU ANSWERED EVERY QUESTION PERFECTLY. HOW DID YOU DO THAT?
WELL, I MET WITH EVERY CANDIDATE YOU INTERVIEWED IN LAST 5 YEARS AND COLLECTED THE QUESTIONS & CORRELATED IT TO INTERVIEW PARAMETERS.
INTERVIEW ROOM
THEN I BUILT A SYSTEM THAT PREDICTS THE EXACT QUESTION YOU'RE GOING TO ASK WITH 85% PRECISION
WOW! THAT IS IMPRESSIVE ENGINEERING BUT I CAN'T HIRE YOU ON ETHICAL GROUNDS
DON'T WORRY. I WAS JUST FIELD TESTING MY PREDICTION SYSTEM

Source:Internet

——

If you have recently completed a professional course/certification, click here to submit a review and get FREE certification highlighter worth Rs. 500.

★★★★★

5.00 avg. rating (99% score) - 8 votes

DATA SCIENCE, FEATURED, INTERVIEW QUESTIONS, TRENDING ARTICLES

Browse Courses by Categories

Machine Learning Courses Deep Learning Courses

Data Science Basics Courses

Related Posts

Oct 10 2019	Feb 10 2021	Apr 21 2021
DATA SCIENCE	DATA SCIENCE	DATA SCIENCE

13 Examples of Machine Learning Applications in Real World

Top Machine Learning and Data Science Tools To Make You Job-ready in 2021

Best Rated Data Science Courses on Coursera

Popular Courses

Artificial intelligence and Machine Learning Courses
Data Science Courses
Management Courses
Digital Marketing Courses
Product Management Courses
Business Analytics Courses
Python Courses
Sales Management Courses
Healthcare Courses
Accounting and Finance Courses

Popular Platforms

Coursera Courses
Udemy Courses
Swayam & NPTEL Courses
Edx Courses
IIM Courses
Udacity Courses
Ivy league Courses
IIT Courses
Google Courses

Courses & Certifications

Free online courses
Top courses starting soon
Programming for managers
Top 10 Free AI Courses
Executive programs
Trending courses
Top Interview Questions & Answers

Information

About us
Contact us
Sitemap
Customer Support
Feedback
Privacy Policy
Terms & Conditions
Grievances
Trust & Safety
Locate Branch
Register Here
Blog

Contact Us

Toll Free - 1800-102-5557
(9.00 AM to 8.00PM IST)

Email - nl.feedback@naukri.com

Connect with us

Terms & Conditions

While Naukri Learning services have helped many customers over the years, we do not guarantee any interview calls or assure any job offers with any of our services. The services associated with Naukri Learning are only provided through the website Naukri.com. You are advised to be cautious of calls/emails asking for payment from other web sites that claim to offer similar services under the name of Naukri.com. We have no associates/agents other than the partner sites that have been specifically named on the homepage of the website Naukri.com. We also recommend that you read the Security Guidelines and Terms and Conditions

Partner Sites

Naukri Naukrigulf Shiksha 99acres Firstnaukri Jeevansathi Ambitionbox Policybazaar Brijj Zomato Meritnation