

Get confidence to solve real live industry projects in Big Data & Data Science.

Get 50% off by June 27.

View Project List



100+ Data Science Interview Questions and Answers for 2021

Last Updated: 23 Jun 2021

Get 1250+ Data Science code snippets

GET NOW



100 Common Data Science Interview Questions & Answers
Hone yourself to be the ideal candidate at your next data science interview with these frequently asked data science interview questions. Data Scientist interview questions asked at a job interview can be categorized into the following categories -



- Technical Data Scientist Interview Questions based on programming languages like Python, R, etc.
- Technical Data Scientist Interview Questions based on probability, math, machine learning, etc.
- Practical experience or Role-based data scientist interview questions based on the projects you have worked on, and how you solved them.

Apart from interview questions, we have also put together 100+ ready-to-use Data Science solved code examples. Each example solves a specific use-case for your project. They can help in answering interview questions and also a handy guide when working on data science projects.

In collaboration with data scientists, industry experts, and top counselors, we have put together a list of general data science interview questions and answers to help you with your preparation in applying for data science jobs. This first part of a series of data science interview questions and answers article focuses only on common topics like questions around data, probability, statistics, and other data science concepts. This also includes a list of open-ended questions that interviewers ask to get a feel of how often and how quickly you can think on your feet. There are some data analyst interview questions in this blog that can also be asked in a data science interview. These kinds of analytics interview questions are asked to measure if you were

Relevant Projects

[Machine Learning Projects](#)

[Data Science Projects](#)

[Python Projects for Data Science](#)

[Data Science Projects in R](#)

[Machine Learning Projects for Beginners](#)

[Deep Learning Projects](#)

[Neural Network Projects](#)

[Tensorflow Projects](#)

[NLP Projects](#)

[Kaggle Projects](#)

[IoT Projects](#)

[Big Data Projects](#)

[Hadoop Real-Time Projects Examples](#)

[Spark Projects](#)

[Data Analytics Projects for Students](#)

You might also like

[Data Scientist Salary](#)

[How to Become a Data Scientist](#)

[Data Analyst vs Data Scientist](#)

[Data Scientist Resume](#)

successful in applying data science techniques to real-life problems.



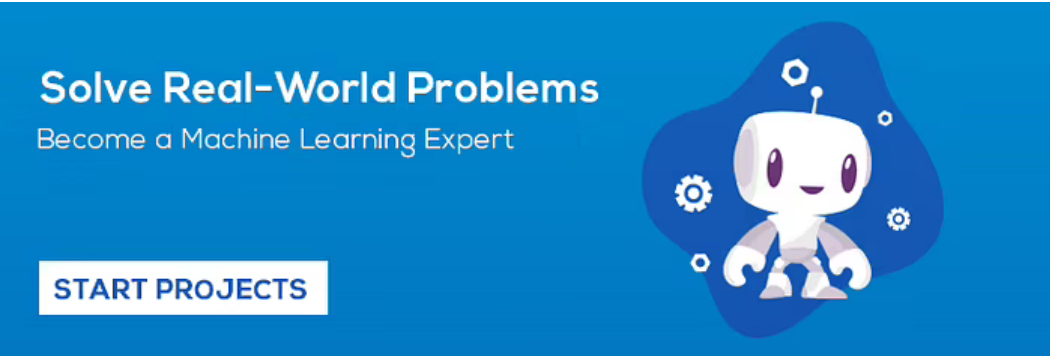
Table of Contents

- [Data Science Interview Questions and Answers](#)
- [Top 100 Common Data Scientist Interview Questions and Answers](#)
- [Data Science Puzzles-Brain Storming/ Puzzle based Data Science Interview Questions asked in Data Scientist Job Interviews](#)
- [Probability Interview Questions for Data Science](#)
- [Statistics Interview Questions for Data Science](#)
- [Python Data Science Interview Questions](#)
- [Frequently Asked Open-Ended Machine Learning Interview Questions for Data Scientists](#)
- [Suggested Answers by Data Scientists for Open-Ended Data Science Interview Questions](#)
- [3 Secrets to becoming a Great Enterprise Data Scientist](#)

Data Science Interview Questions and Answers

Data Science is not an easy field to get into. This is something all data scientists will agree on. Apart from having a degree in mathematics/statistics or engineering, a data scientist also needs to go through intense training to [develop all the skills required for this field](#). Apart from the degree/diploma and the training, it is important to prepare the right resume for a data science job and to be well versed with the data science interview questions and answers.

Consider our top 100 Data Science Interview Questions and Answers as a starting point for your data scientist interview preparation. Even if you are not looking for a data scientist position now, as you are still working your way through hands-on projects and [learning programming languages like Python and R](#) – you can start practicing these Data Scientist Interview questions and answers. These Data Scientist job interview questions will set the foundation for data science interviews to impress potential employers by knowing about your subject and being able to show the practical implications of data science.



Top 100 Common Data Scientist Interview

[Data Science Projects for Beginners](#)

[Machine Learning Engineer](#)

[Machine Learning Projects for Beginners](#)

[Datasets](#)

[Pandas Dataframe](#)

[Machine Learning Algorithms](#)

[Regression Analysis](#)

[MNIST Dataset](#)

[Data Science Interview Questions](#)

[Python Data Science Interview Questions](#)

[Spark Interview Questions](#)

[Hadoop Interview Questions](#)

[Data Analyst Interview Questions](#)

[Machine Learning Interview Questions](#)

[AWS vs Azure](#)

[Hadoop Architecture](#)

[Spark Architecture](#)

Blog Categories

[Apache Flume Projects](#)

[Big Data](#)

[CRM](#)

[Data Science](#)

Questions and Answers

1)Differentiate between Data Science, Machine Learning, and AI.

[Access 100+ ready-to-use, sample Python and R codes for data science](#) to prepare for your Data Science Interview

Data Science vs Machine Learning

Criteria	Data Science	Machine Learning	Artificial Intelligence
Definition	Data Science is not exactly a subset of machine learning but it uses machine learning to analyze and make future predictions.	A subset of AI that focuses on a narrow range of activities.	A wide term that focuses on applications ranging from Robotics to Text Analysis.
Role	It can take on a business role.	It is a purely technical role.	It is a combination of both business and technical aspects.
Scope	Data Science is a broad term for diverse disciplines and is not merely about developing and training models.	Machine learning fits within the data science spectrum.	AI is a sub-field of computer science.
AI	Loosely integrated	Machine learning is a subfield of AI and is tightly integrated.	A sub-field of computer science consisting of various tasks like planning, moving around in the world, recognizing objects and sounds, speaking, translating, performing social business transactions, creative work.

2) Python or R – Which one would you prefer for text analytics?

The best possible answer for this would be Python because it has a Pandas library that provides easy-to-use data structures and high-performance data analysis tools.



[Data Science Projects in Python](#)

[Data Science Projects in R](#)

[Deep Learning Projects](#)

[IoT Projects](#)

[Live Courses](#)

[Machine Learning Projects in Python](#)

[Mobile App Development](#)

[NLP Projects](#)

[NoSQL Database](#)

[Web Development](#)

Tutorials

[Hadoop Online Tutorial – Hadoop HDFS Commands Guide](#)

[MapReduce Tutorial–Learn to implement Hadoop WordCount Example](#)

[Hadoop Hive Tutorial-Usage of Hive Commands in HQL](#)

[Hive Tutorial-Getting Started with Hive Installation on Ubuntu](#)

[Learn Java for Hadoop Tutorial: Inheritance and Interfaces](#)

[Learn Java for Hadoop Tutorial: Classes and Objects](#)

[Learn Java for Hadoop Tutorial: Arrays](#)

[Apache Spark Tutorial - Run your First Spark Program](#)

[PySpark Tutorial-Learn to use Apache Spark with Python](#)

[R Tutorial- Learn Data Visualization with R using GGVIS](#)

[Neural Network Training Tutorial](#)

[Python List Tutorial](#)

[Matplotlib Tutorial](#)

[Decision Tree Tutorial](#)



Ooops something went wror

Return to home page to continue browsing.

Create your own LeadQuizzes

3) Which technique is used to predict categorical responses?

Classification technique is used widely in mining for classifying data sets.

4) What is logistic regression? Or State an example when you have used logistic regression recently. (Access a sample use-case on Logistic Regression)

Logistic Regression often referred to as the logit model is a technique to predict the binary outcome from a linear combination of predictor variables. For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

5) What are Recommender Systems?

A subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

6) Why data cleaning plays a vital role in the analysis? (Access popular Python and R Codes for data cleaning.)

Cleaning data from multiple sources to transform it into a format that data analysts or data scientists can work with is a cumbersome process because - as the number of data sources increases, the time take to clean the data increases exponentially due to the number of sources and the volume of data generated in these sources. It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

7) Differentiate between univariate, bivariate, and multivariate analysis.

These are descriptive statistical analysis techniques that can be differentiated based on the number of variables involved at a given point in time. For example, the pie charts of sales based on territory involve only one variable and can be referred to as univariate analysis.

If the analysis attempts to understand the difference between 2 variables at the time as in a scatterplot, then it is referred to as bivariate analysis. For example, analyzing the volume of sales and spending can be considered as an example of bivariate analysis.

Analysis that deals with the study of more than two variables to understand the effect of variables on the responses is referred to as multivariate analysis.

[Neural Network Tutorial](#)

[Performance Metrics for Machine Learning Algorithms](#)

[R Tutorial: Data.Table](#)

[SciPy Tutorial](#)

[Step-by-Step Apache Spark Installation Tutorial](#)

[Introduction to Apache Spark Tutorial](#)

[R Tutorial: Importing Data from Web](#)

[R Tutorial: Importing Data from Relational Database](#)

[R Tutorial: Importing Data from Excel](#)

[Introduction to Machine Learning Tutorial](#)

[Machine Learning Tutorial: Linear Regression](#)

[Machine Learning Tutorial: Logistic Regression](#)

[Support Vector Machine Tutorial \(SVM\).](#)

[K-Means Clustering Tutorial](#)

[dplyr Manipulation Verbs](#)

[Introduction to dplyr package](#)

[Importing Data from Flat Files in R](#)

[Principal Component Analysis Tutorial](#)

[Pandas Tutorial Part-3](#)

[Pandas Tutorial Part-2](#)

[Pandas Tutorial Part-1](#)

[Tutorial- Hadoop Multinode Cluster Setup on Ubuntu](#)

[Data Visualizations Tools in R](#)

[R Statistical and Language tutorial](#)

[Introduction to Data Science with R](#)

[Apache Pig Tutorial: User Defined Function Example](#)

8) What do you understand by the term Normal Distribution? ([Learn how to plot normal distribution using Seaborn](#))

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve. The random variables are distributed in the form of a symmetrical bell shaped curve.

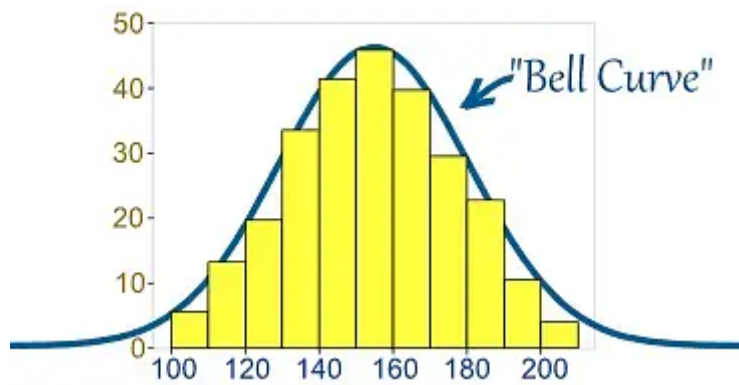


Image Credit: mathisfun.com

9)What is Linear Regression?

Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y as the criterion variable.

10) What are Interpolation and Extrapolation?

Estimating a value from 2 known values from a list of values is Interpolation. Extrapolation is approximating a value by extending a known set of values or facts.

11) What is power analysis?

An experimental design technique for determining the effect of a given sample size.

12) What is K-means? How can you select K for K-means?

Get Closer To Your Dream of Becoming a Data Scientist with 70+ Solved [End-to-End ML Projects](#)

13) What is Collaborative filtering?

The process of filtering used by most of the recommender systems to find patterns or information by collaborating viewpoints, various data sources, and multiple agents.

14) What is the difference between Cluster and Systematic Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. A cluster sample is a probability sample where each sampling unit is a collection or cluster of elements. Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example for systematic sampling is the equal probability method.

15) Are expected value and mean value different?

They are not different but the terms are used in different contexts. Mean is

[Apache Pig Tutorial Example: Web Log Server Analytics](#)

[Impala Case Study: Web Traffic](#)

[Impala Case Study: Flight Data Analysis](#)

[Hadoop Impala Tutorial](#)

[Apache Hive Tutorial: Tables](#)

[Flume Hadoop Tutorial: Twitter Data Extraction](#)

[Flume Hadoop Tutorial: Website Log Aggregation](#)

[Hadoop Sqoop Tutorial: Example Data Export](#)

[Hadoop Sqoop Tutorial: Example of Data Aggregation](#)

[Apache Zookeeper Tutorial: Example of Watch Notification](#)

[Apache Zookeeper Tutorial: Centralized Configuration Management](#)

[Hadoop Zookeeper Tutorial](#)

[Hadoop Sqoop Tutorial](#)

[Hadoop PIG Tutorial](#)

[Hadoop Oozie Tutorial](#)

[Hadoop NoSQL Database Tutorial](#)

[Hadoop Hive Tutorial](#)

[Hadoop HDFS Tutorial](#)

[Hadoop hBase Tutorial](#)

[Hadoop Flume Tutorial](#)

[Hadoop 2.0 YARN Tutorial](#)

[Hadoop MapReduce Tutorial](#)

[Big Data Hadoop Tutorial for Beginners- Hadoop Installation](#)

generally referred to when talking about a probability distribution or sample population whereas expected value is generally referred in a random variable context.

For Sampling Data

The mean value is the only value that comes from the sampling data.

Expected Value is the mean of all the means i.e. the value that is built from multiple samples. The expected value is the population mean.

For Distributions

Mean value and Expected value are the same irrespective of the distribution, under the condition that the distribution is in the same population.

16) What does P-value signify about the statistical data?

P-value is used to determine the significance of results after a hypothesis test in statistics. P-value helps the readers to draw conclusions and is always between 0 and 1.

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



[Access 50+ solved end-to-end Data Science and Machine Learning Projects](#) to build a job-winning data science portfolio

17) Do gradient descent methods always converge to the same point?

No, they do not because in some cases it reaches a local minima or a local optima point. You don't reach the global optima point. It depends on the data and starting conditions

18) What is the benefit of shuffling a training dataset when using a batch gradient descent algorithm for optimizing a neural network?

As we all know, in Batch gradient descent, the entire training data is considered for taking a single step of gradient descent. Mini batch gradient descent is a compromise between Batch gradient descent and Stochastic gradient descent where small batches of data are considered to take each step of gradient descent. In both Batch gradient descent and Mini Batch gradient descent, shuffling of data after each epoch is crucial.

Shuffling the data after each epoch will help us create batches that are more general representations of the overall dataset. To put it in other words, it helps us create batches that include all the training examples in different possible combinations. This helps us obtain the gradient estimate much more efficiently. This in turn reduces the variance of the model so that a more generalized model is obtained. Apart from this, another obvious scenario to shuffle the data is before splitting the data which is arranged by the order of its output class.

19) A test has a true positive rate of 100% and false-positive rate of 5%. There is a population with a 1/1000 rate of having the condition the test identifies. Considering a positive test, what is the probability of having that condition?

Let's suppose you are being tested for a disease, if you have the illness the test will end up saying you have the illness. However, if you don't have the illness- 5% of the time the test will end up saying you have the illness, and 95% of the time the test will give an accurate result that you don't have the illness. Thus there is a 5% error in case you do not have the illness.

Out of 1000 people, 1 person who has the disease will get a true positive result.

Out of the remaining 999 people, 5% will also get true positive results.

Close to 50 people will get a true positive result for the disease.

This means that out of 1000 people, 51 people will be tested positive for the disease even though only one person has the illness. There is only a 2% probability of you having the disease even if your reports say that you have the disease.

20) How you can make data normal using Box-Cox transformation?

21) What is the difference between Supervised Learning and Unsupervised Learning?

If an algorithm learns something from the training data so that the knowledge can be applied to the test data, then it is referred to as Supervised Learning. Classification is an example of Supervised Learning. If the algorithm does not learn anything beforehand because there is no response variable or any training data, then it is referred to as unsupervised learning. Clustering is an example of unsupervised learning.

22) Why it is not advisable to use a softmax output activation function in a multi-label classification problem for a one-hot encoded target?

The **Sigmoid function**, which is also known as the logistic function is used when the problem is a binary classification problem. The formula of sigmoid function goes as given below.



Generally, if the output of the sigmoid function is greater than 0.5, then it corresponds to class 1 and class 0 otherwise.

The **Softmax function** is a generalized version of the sigmoid function to multiple dimensions or classes. The softmax function assumes that the outputs are **mutually exclusive**. If the outputs are one-hot encoded, then they might not be mutually exclusive which is why we do not prefer Softmax

activation functions in such cases.

Another reason is that, when we say that the labels are one-hot encoded, then it means that the output will contain either a 0 or a 1 which is a more comfortable scenario for the Sigmoid function rather than the softmax function.

23) Why is vectorization considered a powerful method for optimizing numerical code?

All computer CPUs support **SIMD (Single Instruction Multiple Data)** where a single instruction can be applied to multiple data points simultaneously. Vectorization can be defined as a process of transforming the system from operating on a single data point at a time to multiple data points simultaneously. Hence when we say that we have vectorized the code, it means that we are applying a single instruction to multiple data points simultaneously. With a conventional for loop (or while loop or any other looping techniques for that matter), we apply the instructions on only one data point at each iteration but when we use a vectorized approach, the instruction can be applied to n (say $n = 3$) data points at each iteration. If we have N such data points, and if the instruction takes 1 second to run on each data point, the conventional for loop might take $1 * N = N$ seconds to run. But in a vectorized approach, the time taken will be N/n seconds. i.e., the time taken is reduced n fold.

24) What is the goal of A/B Testing?

It is a statistical hypothesis testing for a randomized experiment with two variables A and B. The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. An example of this could be identifying the click-through rate for a banner ad.

25) What is an Eigenvalue and Eigenvector?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing, or stretching. 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.

Get hands-on experience on real-time [Data Science and Machine Learning Projects](#) to showcase in your Interview

26) What is Gradient Descent?

Gradient descent is an iterative procedure that aims at minimizing the cost function parametrized by model parameters. It is an optimization method based on convex function and trims the parameters iteratively to help the given function attain its local minimum. Gradient measures the change in parameter with respect to the change in error. Imagine a blindfolded person on top of a hill and wanting to reach the lower altitude. The simple technique he can use is to feel the ground in every direction and take a step in the direction where the ground is descending faster. Here we need the help of the learning rate which says the size of the step we take to reach the minimum. The learning rate should be chosen in such a way that it should not be too high or too low. When the learning rate chosen is too high it tends to bounce back and forth between the convex function of the gradient descent



and when it is too low we will reach the minimum very slowly.

27) Differentiate between a multi-label classification problem and a multi-class classification problem.

A multi-class classification problem is a classification problem with more than two output classes. For example, if we have a collection of fruit images, we can classify them into an apple, mango, banana, guava, etc. In a multi-class classification problem, each sample can belong to only one class. For example, fruit can either be mango or banana. It can not be both. This is why we say that in a multiclass classification problem, the classes are mutually exclusive.

A multi-label is a classification problem in which more than one class can be assigned to a sample input. For example, if we have a collection of news articles, each text sample can be associated with more than one topic such as religion, political, social, sports, etc. Hence the classes are not mutually exclusive and they can be related to each other.

28) What is the difference between gradient descent optimization algorithms Adam and Momentum?

MOMENTUM ALGORITHM

Vanilla gradient descent with momentum is a method of **accelerating the gradient descent** to move faster towards the global minimum.

Mathematically, a decay rate is multiplied to the previous sum of gradients and added with the present gradient to get a new sum of gradients. When the decay rate is assigned zero, it denotes a normal gradient descent. When the decay rate is set to 1, it oscillates like a ball in a frictionless bowl without any end. Hence decay rate is typically chosen around **0.8 to 0.9** to arrive at an end. The momentum algorithm gives us the advantage of escaping the local minima and getting into global minima.

ADAM ALGORITHM

Adaptive Moment Estimation, shortly called ADAM, is a combination of **Momentum and RMSProp**. In the AdaGrad algorithm, the sum of gradients is squared which only grows and it is incredibly slow. RMSProp is nothing but root mean square propagation which fixes the issue by considering a decay factor. In the Adam algorithm, when mathematically explained, two decay rates are used namely beta1 and beta2 where beta1 denotes the first momentum in which the sum of the gradient is considered and beta2 denotes the second momentum in which the sum of gradient squared is considered. Since the Momentum algorithm gives us a faster way and RMSProp provides the ability to gradient to restyle in different directions, the combination of the two works well, and thus Adam algorithm is considered as the go-to choice of deep learning algorithms.

29) What are the various steps involved in an analytics project?

- Understand the business problem
- Explore the data and become familiar with it.
- Prepare the data for modeling by detecting outliers, treating missing values, transforming variables, etc.
- After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step till the best possible outcome is

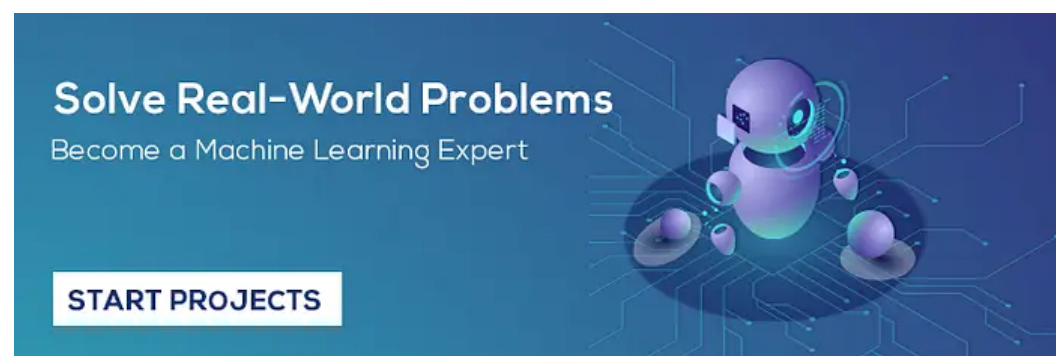


achieved.

- Validate the model using a new data set.
- Start implementing the model and track the result to analyse the performance of the model over the period of time.

30) How can you iterate over a list and also retrieve element indices at the same time?

This can be done using the enumerate function which takes every element in a sequence just like in a list and adds its location just before it.



31) During analysis, how do you treat missing values? ([get sample code here](#))

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights. If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored. There are various factors to be considered when answering this question-

- Understand the problem statement, understand the data and then give the answer. Assigning a default value which can be the mean, minimum, or maximum value. Getting into the data is important.
- If it is a categorical variable, the default value is assigned. The missing value is assigned a default value.
- If you have a distribution of data coming, for normal distribution give the mean value.
- Should we even treat missing values is another important point to consider? If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

32) Explain the box cox transformation in regression models.

For some reason or the other, the response variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow the skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality. Applying a box cox transformation means that you can run a broader number of tests.

33) Can you use machine learning for time series analysis?

Yes, it can be used but it depends on the applications.

34) Write a function that takes in two sorted lists and outputs a sorted

list that is their union.

The first solution which will come to your mind is to merge two lists and sort them afterwards .

Python code-

```
def return_union(list_a, list_b):
    return sorted(list_a + list_b)
```

R code-

```
return_union <- function(list_a, list_b)
{
list_c<-list(c(unlist(list_a),unlist(list_b)))
return(list(list_c[[1]][order(list_c[[1]])]))
}
```

Generally, the tricky part of the question is not to use any sorting or ordering function. In that case, you will have to write your own logic to answer the question and impress your interviewer.

Python code-

```
def return_union(list_a, list_b):
    len1 = len(list_a)
    len2 = len(list_b)
    final_sorted_list = []
    j = 0
    k = 0

    for i in range(len1+len2):
        if k == len1:
            final_sorted_list.extend(list_b[j:])
            break
        elif j == len2:
            final_sorted_list.extend(list_a[k:])
            break
        elif list_a[k] < list_b[j]:
            final_sorted_list.append(list_a[k])
            k += 1
        else:
            final_sorted_list.append(list_b[j])
            j += 1
    return final_sorted_list
```

Similar function can be returned in R as well by following similar steps.

```
return_union <- function(list_a,list_b)
{
#Initializing length variables
len_a <- length(list_a)
len_b <- length(list_b)
len <- len_a + len_b

#initializing counter variables

j=1
k=1
```

```
#Creating an empty list which has length equal to sum of both the lists
```

```

list_c <- list(rep(NA,len))

#Here goes our for loop

for(i in 1:len)
{
  if(j>len_a)
  {
    list_c[i:len] <- list_b[k:len_b]
    break
  }
  else if(k>len_b)
  {
    list_c[i:len] <- list_a[j:len_a]
    break
  }
  else if(list_a[[j]] <= list_b[[k]])
  {
    list_c[[i]] <- list_a[[j]]
    j <- j+1
  }
  else if(list_a[[j]] > list_b[[k]])
  {
    list_c[[i]] <- list_b[[k]]
    k <- k+1
  }
}
return(list(unlist(list_c)))

}

```

[CLICK HERE](#)

**to get free access to 120 Data Science Interview
Questions and Answers PDF**

35) What is the difference between Bayesian Estimate and Maximum Likelihood Estimation (MLE)?

In bayesian estimate, we have some knowledge about the data/problem (prior). There may be several values of the parameters which explain data and hence we can look for multiple parameters like 5 gammas and 5 lambdas that do this. As a result of Bayesian Estimate, we get multiple models for making multiple predictions i.e. one for each pair of parameters but with the same prior. So, if a new example needs to be predicted then computing the weighted sum of these predictions serves the purpose.

Maximum likelihood does not take prior into consideration (ignores the prior) so it is like being a Bayesian while using some kind of a flat prior.

36) What is Regularization and what kind of problems does regularization solve?

Regularization is basically a technique that is used to push or encourage the coefficients of the machine learning model **towards zero** in order to reduce the **over-fitting** problem. The general idea of regularization is to penalize complicated models by adding an **additional penalty to the loss function** in order to generate a larger loss. In this way, we can discourage the model



from learning too many details and the model is much **more general**.

There are two ways of assigning the additional penalty term to the loss function giving rise to two types of regularization techniques. They are

1. L2 Regularization
2. L1 Regularization

In **L2 Regularization**, the penalty term is the **sum of squares** of the magnitude of the model coefficients while in L1 Regularization, it is the **sum of absolute values** of the model coefficients

37) How will you tackle an exploding gradient problem?

38) How will you tackle a vanishing gradient problem?

39) How do you decide whether your linear regression model fits the data?

A good fitting regression model results in predicted values closer to the observed values. We can use any of the metrics below to check the performance of a linear regression model on our data.

1. R-squared: It is based on Sum of Squares Total (SST) and Sum of Squares Error (SSE). SST measures how far the data are from the mean of the data and SSE measures how far the data are from the model's predicted values. Dividing the difference of SST and SSE with SST will give us the R-squared value. This proportion indicates how well the model is fit. R-squared ranges from zero to one, where zero indicates that the model makes poor predictions and one indicates perfect predictions. An increase in R square is proportional to improvement in the regression model.

2. F-test: f-test assesses with a null hypothesis that all coefficients in the regression model are zero and an alternate hypothesis that at least one is not zero. We accept the null hypothesis when R-squared equals to zero.

3. RMSE: It is the square root of the variance of the residuals. Lower the value of RMSE, the better the model is. The formula for RMSE is -

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

R-squared is considered as a relative measure of fit whereas RMSE is an absolute measure of fit.

40) What is the difference between squared error and absolute error?

Mean Absolute Error

Mean absolute error is the **average absolute difference** between the predicted and the actual values across the validation set. It gives us the average residual of the validation data. The formula for mean absolute error is

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

MAE = mean absolute error
 y_i = prediction
 x_i = true value
 n = total number of data points

Mean square error

Mean square error is the **average of the squared difference** between the predicted and the actual values across the validation set. This gives us the variance of the residuals in the validation data. Unlike MAE, **MSE punishes large errors more** since it is a squared metric. The formula for mean squared error is

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
 n = number of data points
 Y_i = observed values
 \hat{Y}_i = predicted values

41) What is Machine Learning?

The simplest way to answer this question is – we give the data and equation to the machine. Ask the machine to look at the data and identify the coefficient values in an equation.

For example for the linear regression $y=mx+c$, we give the data for the variable x , y and the machine learns about the values of m and c from the data.

Recommended Reading: [Deep Learning Interview Questions and Answers](#)

42) How are confidence intervals constructed and how will you interpret them?

A confidence interval provides a range of values that is likely to contain the population parameter of interest. In most statistical case studies, we tend to estimate the population mean. We can calculate the confidence interval of the average of a population if the standard deviation of the population is known using the formula below -

$\bar{x} \pm z * \frac{\sigma}{\sqrt{n}}$, where \bar{x} is the sample mean, σ is the population standard

Here, z stands for the z value from the normal distribution. The z value changes according to the desired confidence level.

z*-values for Various Confidence Levels	
Confidence Level	z*-value
80%	1.28
90%	1.645 (by convention)
95%	1.96
98%	2.33
99%	2.58

While interpreting a confidence interval, It is always necessary to remember that when we are estimating a confidence interval we are estimating a population parameter using the data from a sample.

The correct way to interpret a 95% confidence interval can be "we are 95% confident that the population parameter is between X (lower limit) and X (upper limit)."

43) If the training loss of your model is high and almost equal to the validation loss, what does it mean? What should you do?

If the training loss is much higher than the validation loss or if the training loss is almost equal to the validation loss, it means that your model is underfitting. Or in other words, your model is not complex enough to properly understand the underlying relationships between the dependent and the independent variables. We can overcome the problem of underfitting by trying one or more of the following techniques

- Training the model longer, i.e., for more epochs.
- Choosing a more complex learning model than the current one.
- Creating new features which are nonlinear combinations of the previous ones using feature engineering techniques to increase the data features.
- Performing data augmentation if possible.

44) How can you overcome Overfitting?

We can overcome overfitting using one or more of the following techniques

1. Simplifying the model: We can reduce the overfitting of the model by reducing the complexity of the model. We can either remove layers or reduce the number of neurons in the case of a deep learning model, or prefer a lesser order polynomial model in case of regression.

2. Use Regularization: Regularization is the common technique used to remove the complexity of the model by adding a penalty to the loss function. There are two regularization techniques namely L1 and L2. L1 penalizes the sum of absolute values of weight whereas L2 penalizes the sum of square values of weight. When data is too complex to be modeled, the L2 technique is preferred and L1 is better if the data to be modeled is quite simple. However, L2 is more commonly preferred.

3.Data Augmentation: Data augmentation is nothing but creating more data samples using the existing set of data. For example, in the case of a convolutional neural network, producing new images by flipping, rotation, scaling, changing brightness of the existing set of images helps in increasing the dataset size and in turn reducing overfitting.

4.Early Stopping: Early stopping is a regularization technique that identifies the point from where the training data leads to generalization error and begins to overfit. The algorithm stops training the model at that point.

5. Feature reduction: If we have a small number of data samples with a large number of features, we can prevent overfitting by selecting only the most important features. We can use various techniques for this such as F-test, Forward elimination, and Backward elimination.

6. Dropouts: In the case of neural networks, we can also randomly deactivate a proportion of neurons in each layer. This technique is called dropout and it is a form of regularization. However, when we use the dropout technique, we have



to train the data for more epochs.

45) Differentiate between wide and long data formats?

Structured data can be arranged in different formats, mainly the wide and tall format. Wide data format arranges the data in horizontal manner and tall data format arranges the data in vertical format. In wide data format, a single row represents the data with multiple column variables. While in a long data format, each row represents the data of every column variable.

46) Is Naïve Bayes bad? If yes, under what aspects.

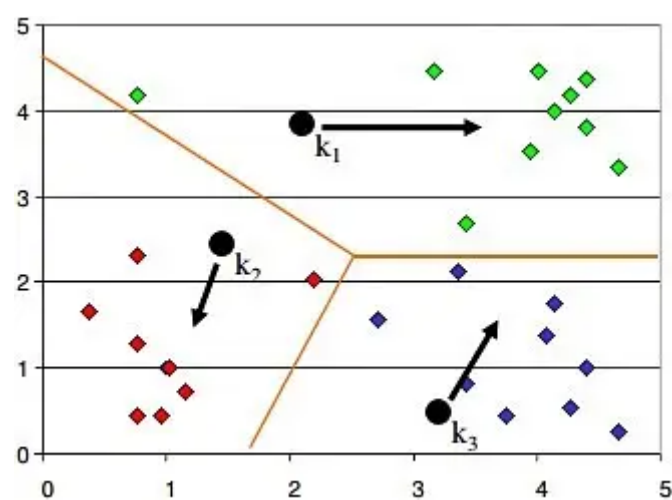
Naïve Bayes is a machine learning algorithm based on the Bayes Theorem. This is used for solving classification problems. It is based on two assumptions, first, each feature/attribute present in the dataset is independent of another, and second, each feature carries equal importance. But this assumption of Naïve Bayes turns out to be disadvantageous. As it assumes that the features are independent of each other, but in real life scenario, this assumption cannot be true as there is always some dependence present in the given set of features. There is another disadvantage of this algorithm: the 'zero-frequency problem' where the model assigns value zero for those features in the test dataset that were not present in the training dataset.

47) How would you develop a model to identify plagiarism?

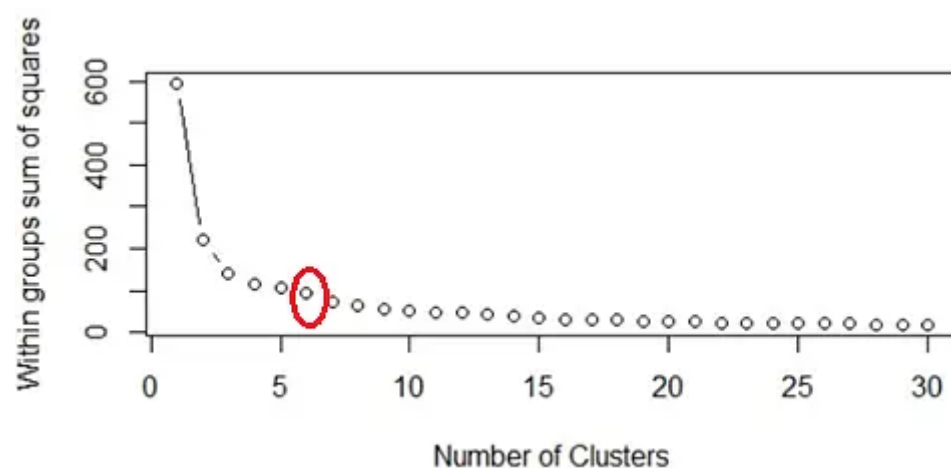
48) How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question will mostly be asked in reference to K-Means clustering where "K" defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other but the groups are different from each other.

For example, the following image shows three different groups.



Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below. The Graph is generally known as Elbow Curve.



Red circled point in above graph i.e. Number of Cluster =6 is the point after

which you don't see any decrement in WSS. This point is known as bending point and taken as K in K – Means.

This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendograms and identify the distinct groups from there.

49) How important it is to introduce non-linearities in a neural network and why?

Get FREE Access to [Machine Learning Example Codes](#) for Data Cleaning, Data Munging, and Data Visualization

50) Is it possible to perform logistic regression with Microsoft Excel?

It is possible to perform logistic regression with Microsoft Excel. There are two ways to do it using Excel.

- a) One is to use Add-ins provided by many websites which we can use.
- b) Second is to use fundamentals of logistic regression and use Excel's computational power to build a logistic regression

But when this question is being asked in an interview, interviewer is not looking for a name of Add-ins rather a method using the base excel functionalities.

Let's use a sample data to learn about logistic regression using Excel. (Example assumes that you are familiar with basic concepts of logistic regression)

	A	B	C
6			
7	X1	X2	Y
8	39	4	0
9	36.5	4	0
10	36.5	2.5	0
11	35.5	3.5	0
12	34	2.5	0
13	29.5	2	0
14	28.5	3.5	0
15	24.5	2.5	0
16	17.5	2	0
17	13.5	3.5	0
18	29.5	1.5	1
19	28.5	2	1
20	22	2.5	1
21	19	2.5	1
22	18	2	1
23	18	1	1
24	11	3	1
25	11	2.5	1
26	7.5	2	1
27	5	3	1

Data shown above consists of three variables where X1 and X2 are independent variables and Y is a class variable. We have kept only 2 categories for our purpose of binary logistic regression classifier.

Next we have to create a logit function using independent variables, i.e.

Logit = $L = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$

	A	B	C	D	E	F
1			Decision Variables			
2				B0	0.1	
3				B1	0.1	
4				B2	0.1	
5						

6					
7	X1	X2	Y	Logit	
8	39	4	0	=SE\$2+\$E\$3*A8+\$E\$4*B8	
9	36.5	4	0		
10	36.5	2.5	0		
11	35.5	3.5	0		
12	34	2.5	0		
13	29.5	2	0		
14	28.5	3.5	0		
15	24.5	2.5	0		
16	17.5	2	0		
17	13.5	3.5	0		
18	29.5	1.5	1		
19	28.5	2	1		
20	22	2.5	1		
21	19	2.5	1		
22	18	2	1		
23	18	1	1		
24	11	3	1		

We have kept the initial values of beta 1, beta 2 as 0.1 for now and we will use Excel Solve to optimize the beta values in order to maximize our log likelihood estimate.


Assuming that you are aware of logistic regression basics, we calculate probability values from Logit using following formula:

$$\text{Probability} = \frac{e^{\text{Logit}}}{1 + e^{\text{Logit}}}$$

e is base of natural logarithm i.e. $e = 2.71828163$

Let's put it into excel formula to calculate probability values for each of the observation.

	A	B	C	D	E	F
1			Decision Variables			
2				B0	0.1	
3				B1	0.1	
4				B2	0.1	
5						
6						
7	X1	X2	Y	Logit	Probability	
8	39	4	0	4.4	=EXP(D8)/(1+EXP(D8))	
9	36.5	4	0	4.15		
10	36.5	2.5	0	4		
11	35.5	3.5	0	4		
12	34	2.5	0	3.75		
13	29.5	2	0	3.25		
14	28.5	3.5	0	3.3		
15	24.5	2.5	0	2.8		
16	17.5	2	0	2.05		
17	13.5	3.5	0	1.8		
18	29.5	1.5	1	3.2		
19	28.5	2	1	3.15		
20	22	2.5	1	2.55		
21	19	2.5	1	2.25		
22	18	2	1	2.1		
23	18	1	1	2		
24	11	3	1	1.5		

The conditional probability  is the probability of Predicted Y, given set of independent variables X.

And this p can be calculated as-

$$P[(X)]^{Y_{\text{actual}}} [1 - P[(X)]^{(1 - Y_{\text{actual}})}]$$

Then we have to take natural log of the above function-

$$\ln \left[[P[(X)]^{Y_{\text{actual}}} [1 - P[(X)]^{(1 - Y_{\text{actual}})}] \right]$$

Which turns out to be –

$$Y_{\text{actual}} * \ln [P[(X)]^{Y_{\text{actual}}} [1 - P[(X)]^{(1 - Y_{\text{actual}})}] + (1 - Y_{\text{actual}}) * \ln [1 - P[(X)]^{Y_{\text{actual}}} [1 - P[(X)]^{(1 - Y_{\text{actual}})}]]$$

$$P(Y=y|X) = \frac{e^{B_0 + B_1X_1 + B_2X_2}}{1 + e^{B_0 + B_1X_1 + B_2X_2}}$$

Log likelihood function LL is the sum of above equation for all the observations

	A	B	C	D	E	F	G
1			Decision Variables				
2				B0	0.1		
3				B1	0.1		
4				B2	0.1		
5							
6							
7	X1	X2	Y	Logit	Probability	P(Y=y X)	
8	39	4	0	4.4	0.987871565	=C8*LN(E8)+(1-C8)*LN(1-E8)	
9	36.5	4	0	4.15	0.984480243		
10	36.5	2.5	0	4	0.98201379		
11	35.5	3.5	0	4	0.98201379		
12	34	2.5	0	3.75	0.97702263		
13	29.5	2	0	3.25	0.962673113		
14	28.5	3.5	0	3.3	0.964428811		
15	24.5	2.5	0	2.8	0.942675824		
16	17.5	2	0	2.05	0.885947619		
17	13.5	3.5	0	1.8	0.858148935		
18	29.5	1.5	1	3.2	0.960834277		
19	28.5	2	1	3.15	0.958908722		
20	22	2.5	1	2.55	0.927573515		
21	19	2.5	1	2.25	0.904650535		
22	18	2	1	2.1	0.890903179		
23	18	1	1	2	0.880797078		
24	11	3	1	1.5	0.817574476		

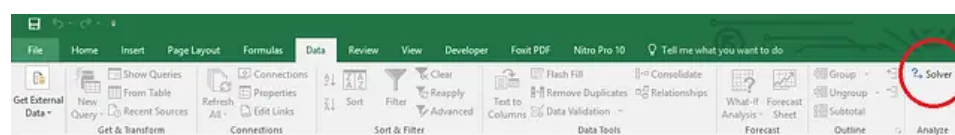
Log likelihood LL will be sum of column G, which we just calculated

	A	B	C	D	E	F	G	H	I
1			Decision Variables				Log Likelihood		
2				B0	0.1			=SUM(F8:F27)	
3				B1	0.1				
4				B2	0.1				
5									
6									
7	X1	X2	Y	Logit	Probability	P(Y=y X)			
8	39	4	0	4.4	0.987871565	-4.41220258			
9	36.5	4	0	4.15	0.984480243	-4.16564145			
10	36.5	2.5	0	4	0.98201379	-4.01814993			
11	35.5	3.5	0	4	0.98201379	-4.01814993			
12	34	2.5	0	3.75	0.97702263	-3.77324546			
13	29.5	2	0	3.25	0.962673113	-3.28804137			
14	28.5	3.5	0	3.3	0.964428811	-3.33621926			
15	24.5	2.5	0	2.8	0.942675824	-2.85903283			
16	17.5	2	0	2.05	0.885947619	-2.17109745			
17	13.5	3.5	0	1.8	0.858148935	-1.95297761			
18	29.5	1.5	1	3.2	0.960834277	-0.03995333			
19	28.5	2	1	3.15	0.958908722	-0.04195939			
20	22	2.5	1	2.55	0.927573515	-0.07518323			
21	19	2.5	1	2.25	0.904650535	-0.10020656			
22	18	2	1	2.1	0.890903179	-0.11551952			
23	18	1	1	2	0.880797078	-0.12692801			
24	11	3	1	1.5	0.817574476	-0.20141328			

The objective is to maximize the Log Likelihood i.e. cell H2 in this example. We have to maximize H2 by optimizing B₀, B₁, and B₂.

We'll use Excel's solver add-in to achieve the same.

Excel comes with this Add-in pre-installed and you must see it under Data Tab in Excel as shown below



If you don't see it there then make sure if you have loaded it. To load an add-in in Excel,

Go to *File >> Options >> Add-Ins* and see if checkbox in front of required add-in is checked or not? Make sure to check it to load an add-in into Excel.

If you don't see Solver Add-in there, go to the bottom of the screen (Manage Add-Ins) and click on OK. Next you will see a popup window which should have your Solver add-in present. Check the checkbox in front of the add-in name. If you don't see it there as well click on browse and direct it to the required folder which contains Solver Add-In.

Once you have your Solver loaded, click on Solver icon under Data tab and You will see a new window popped up like –

Put $H2$ in set objective, select max and fill cells $E2$ to $E4$ in next form field.

By doing this we have told Solver to Maximize $H2$ by changing values in cells $E2$ to $E4$.

Now click on Solve button at the bottom –

You will see a popup like below -

This shows that Solver has found a local maxima solution but we are in need of Global Maxima Output. Keep clicking on Continue until it shows the below popup

It shows that Solver was able to find and converge the solution. In case it is not able to converge it will throw an error. Select "Keep Solver Solution" and Click on OK to accept the solution provided by Solver.

Now, you can see that value of Beta coefficients from B_0 , B_1 , B_2 have changed and our Log-Likelihood function has been maximized.

	A	B	C	D	E	F	G	H
1			<i>Decision Variables</i>				<i>Log Likelihood</i>	
2				B0	12.48309171			-6.65456
3				B1	-0.23406877			
4				B2	-2.93832567			
5								
6								
7	X1	X2	Y	Logit	Probability	P(Y=y X)		
8	39	4	0	-8.3988928	0.000225066	-0.00022509		
9	36.5	4	0	-7.8137209	0.000403988	-0.00040407		
10	36.5	2.5	0	-3.4062324	0.032101254	-0.0326278		
11	35.5	3.5	0	-6.1104893	0.002214549	-0.00221701		
12	34	2.5	0	-2.8210605	0.056196661	-0.05783746		
13	29.5	2	0	-0.2985882	0.425902646	-0.55495629		
14	28.5	3.5	0	-4.4720079	0.011295312	-0.01135959		
15	24.5	2.5	0	-0.5974072	0.354937108	-0.43840746		
16	17.5	2	0	2.510237	0.924856362	-2.58835382		
17	13.5	3.5	0	-0.9609765	0.276682734	-0.32390733		
18	29.5	1.5	1	1.1705746	0.763248868	-0.27017113		
19	28.5	2	1	-0.0645194	0.483875735	-0.72592715		
20	22	2.5	1	-0.0122353	0.496941214	-0.69928354		
21	19	2.5	1	0.689971	0.665960476	-0.40652496		

Using these values of Betas you can calculate the probability and hence response variable by deciding the probability cut-off.

51) What do you understand by Fuzzy merging? Which language will you use to handle it?

52) What is the difference between skewed and uniform distribution?

When the observations in a dataset are spread equally across the range of distribution, then it is referred to as uniform distribution. There are no clear perks in a uniform distribution. Distributions that have more observations on one side of the graph than the other are referred to as skewed distribution. Distributions with fewer observations on the left (towards lower values) are said to be skewed left and distributions with fewer observations on the right (towards higher values) are said to be skewed right.

[The #1 question in your interview will be "What experience do you have?"](#)

[Get hands-on experience with free access to 100+ code examples solved by industry experts.](#)

[Click here \(you can rapidly get some project experience before your interviews\)](#)

53) You created a predictive model of a quantitative outcome variable using multiple regressions. What are the steps you would follow to validate the model?

Since the question asked, is about post model building exercise, we will assume that you have already tested for null hypothesis, multi collinearity and Standard error of coefficients.

Once you have built the model, you should check for following –

- Global F-test to see the significance of group of independent variables on dependent variable

- R^2

- Adjusted R^2
- RMSE, MAPE

In addition to above mentioned quantitative metrics you should also check for-

- Residual plot
- Assumptions of linear regression

54) What do you understand by Hypothesis in the content of Machine Learning?

[CLICK HERE](#)

to get free access to 120 Python Data Science Interview Questions and Answers PDF

55) What do you understand by Recall and Precision?

Recall measures "Of all the actual true samples how many did we classify as true?"

Precision measures "Of all the samples we classified as true how many are actually true?"

We will explain this with a simple example for better understanding -

Imagine that your wife gave you surprises every year on your anniversary in the last 12 years. One day all of a sudden your wife asks - "Darling, do you remember all anniversary surprises from me?".

This simple question puts your life in danger. To save your life, you need to Recall all 12-anniversary surprises from your memory. Thus, Recall(R) is the ratio of the number of events you can correctly recall to the number of all correct events. If you can recall all the 12 surprises correctly then the recall ratio is 1 (100%) but if you can recall only 10 surprises correctly of the 12 then the recall ratio is 0.83 (83.3%).

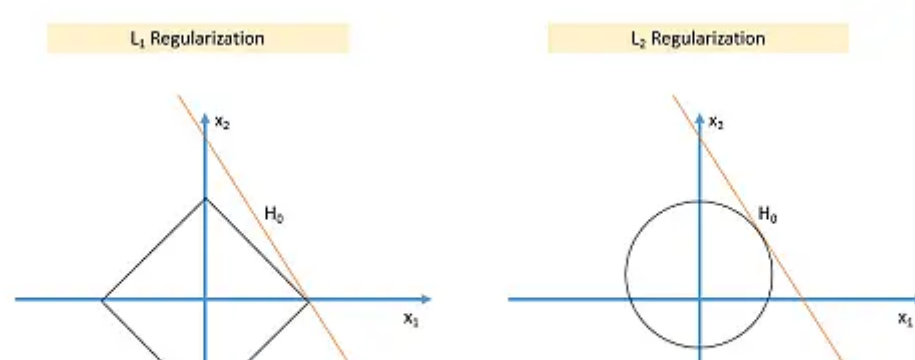
However, you might be wrong in some cases. For instance, you answer 15 times, 10 times the surprises you guess are correct and 5 wrong. This implies that your recall ratio is 100% but the precision is 66.67%.

Precision is the ratio of the number of events you can correctly recall to the number of all events you recall (combination of wrong and correct recalls).

56) How will you find the right K for K-means?

57) Why L1 regularizations cause parameter sparsity whereas L2 regularization does not?

Regularizations in statistics or in the field of machine learning is used to include some extra information in order to solve a problem in a better way. L1 & L2 regularizations are generally used to add constraints to optimization problems.





In the example shown above H_0 is a hypothesis. If you observe, in L_1 there is a high likelihood to hit the corners as solutions while in L_2 , it doesn't. So in L_1 variables are penalized more as compared to L_2 which results into sparsity.

In other words, errors are squared in L_2 , so model sees higher error and tries to minimize that squared error.

Get More Practice, More [Data Science and Machine Learning Projects](#), and More guidance. Fast-Track Your Career Transition with ProjectPro

58) How can you deal with different types of seasonality in time series modelling? ([get 100+ solved code examples here](#))

Seasonality in time series occurs when time series shows a repeated pattern over time. E.g., stationary sales decreases during holiday season, air conditioner sales increases during the summers etc. are few examples of seasonality in a time series.

Seasonality makes your time series non-stationary because average value of the variables at different time periods. Differentiating a time series is generally known as the best method of removing seasonality from a time series.

Seasonal differencing can be defined as a numerical difference between a particular value and a value with a periodic lag (i.e. 12, if monthly seasonality is present)

59) Explain evaluation protocols for testing your models? Compare hold-out vs k-fold cross validation vs iterated k-fold cross-validation methods of testing.

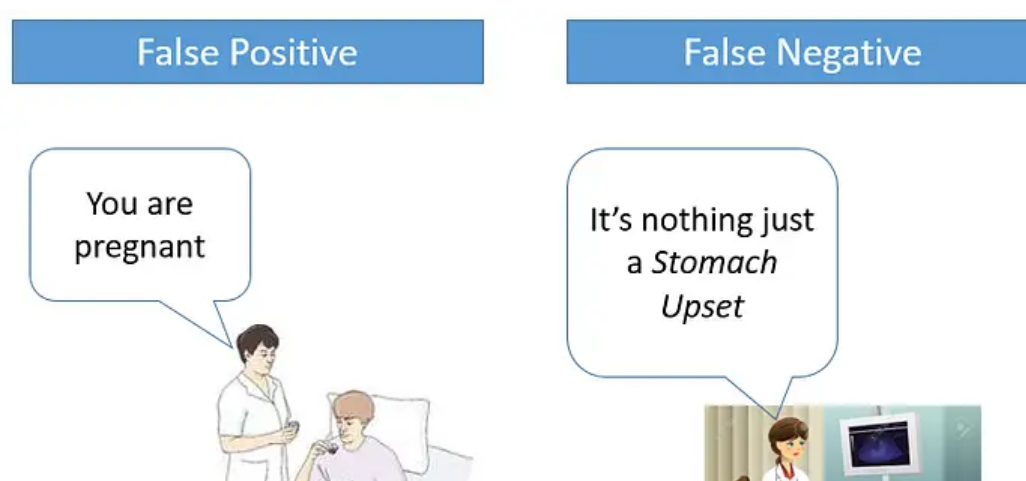
60) What do you understand by conjugate-prior with respect to Naïve Bayes?

61) Can you cite some examples where a false positive is important than a false negative?

Before we start, let us understand what are false positives and what are false negatives.

False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.

And, False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.





In medical field, assume you have to give chemo therapy to patients. Your lab tests patients for certain vital information and based on those results they decide to give radiation therapy to a patient.

Assume a patient comes to that hospital and he is tested positive for cancer (But he doesn't have cancer) based on lab prediction. What will happen to him? (Assuming Sensitivity is 1)

One more example might come from marketing. Let's say an ecommerce company decided to give \$1000 Gift voucher to the customers whom they assume to purchase at least \$5000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above 5K.

Now what if they have sent it to false positive cases?

62) Can you cite some examples where a false negative important than a false positive? ([get 100+ solved code examples here](#))

Assume there is an airport 'A' which has received high security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to shortage of staff they decided to scan passenger being predicted as risk positives by their predictive model.

What will happen if a true threat customer is being flagged as non-threat by airport model?

Another example can be judicial system. What if Jury or judge decide to make a criminal go free?

What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after few years and realize that you had a false negative?

63) Can you cite some examples where both false positive and false negatives are equally important?

In the banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don't want to lose good customers and at the same point of time they don't want to acquire bad customers. In this scenario both the false positives and false negatives **become very important to measure**.

These days we hear many cases of players using steroids during sport competitions Every player has to go through a steroid test before the game starts. A false positive can ruin the career of a Great sportsman and a false negative can make the game unfair.

[Get hands-on experience for your interviews with free access to solved code examples found here \(these are ready-to-use for your projects\)](#)

64) Can you explain the difference between a Test Set and a Validation

Set?

Set:

Validation set can be considered as a part of the training set as it is used for parameter selection and to avoid Overfitting of the model being built. On the other hand, test set is used for testing or evaluating the performance of a trained machine learning model.

In simple terms, the differences can be summarized as-

- Training Set is to fit the parameters i.e. weights.
- Test Set is to assess the performance of the model i.e. evaluating the predictive power and generalization.
- Validation set is to tune the parameters.

65) What makes a dataset gold standard?

66) What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, RF etc.). Sensitivity is nothing but "Predicted TRUE events/ Total events". True events here are the events which were true and model also predicted them as true.

Calculation of sensitivity is pretty straight forward-

Sensitivity = True Positives /Positives in Actual Dependent Variable

Where, True positives are Positive events which are correctly classified as Positives.

67) What is the importance of having a selection bias? ([get 100+ solved code examples here](#))

Selection Bias occurs when there is no appropriate randomization achieved while selecting individuals, groups or data to be analyzed. Selection bias implies that the obtained sample does not exactly represent the population that was actually intended to be analyzed. Selection bias consists of Sampling Bias, Data, Attribute, and Time Interval.

68) Given that you let the models run long enough, will all gradient descent algorithms lead to the same model when working with Logistic or Linear regression problems?

69) Differentiate between Batch Gradient Descent, Mini-Batch Gradient Descent, and Stochastic Gradient Descent.

Gradient descent is one of the most popular machine learning and deep learning optimization algorithm used for updation of parameters of a learning model. There are 3 variants of the gradient descent. Batch Gradient Descent: In batch gradient descent, computation is carried on the entire dataset Stochastic Gradient Descent: In stochastic gradient descent, computation is carried over only one single training sample. Mini Batch Gradient Descent: In mini batch gradient descent, a small number/batch of training samples is used for computation. For example, if a dataset has 1000 datapoints, then batch GD, will train on all the 1000 datapoints, Stochastic GD, will train on only a single sample and the mini batch GD will consider a batch size of say 100 data points and update the parameters.



70) How do data management procedures like missing data handling make selection bias worse?

Missing value treatment is one of the primary tasks which a data scientist is supposed to do before starting data analysis. There are multiple methods for missing value treatment. If not done properly, it could potentially result in selection bias. Let see few missing value treatment examples and their impact on selection-

Complete Case Treatment: Complete case treatment is when you remove an entire row in data even if one value is missing. You could achieve a selection bias if your values are not missing at random and they have some pattern. Assume you are conducting a survey and few people didn't specify their gender. Would you remove all those people? Can't it tell a different story?

Available case analysis: Let say you are trying to calculate a correlation matrix for data so you might remove the missing values from variables that are needed for that particular correlation coefficient. In this case, your values will not be fully correct as they are coming from population sets.

Mean Substitution: In this method, missing values are replaced with the mean of other available values. This might make your distribution biased e.g., standard deviation, correlation and regression are mostly dependent on the mean value of variables.

Hence, various data management procedures might include selection bias in your data if not chosen correctly.

71) What are the advantages and disadvantages of using regularization methods like Ridge Regression?

When we say that a model is overfitting, essentially, we get a low bias and high variance model. So, in order to minimize the overfitting, the technique called regularization is used. The Lasso regularization is termed as L1 regularization and ridge regularization is termed as L2 regularization. Ridge Regression is a further extension of linear regression where a penalty is added to the RSS (residual sum of squares) which is equal to the square of coefficients term. This penalty term is nothing but alpha multiplied by slope and its square. This helps getting rid of overfitting. Ridge regression = $\min(\text{Sum of squared errors} + \alpha * \text{slope}^2)$

Advantages: Ridge regression helps avoid overfitting of the model. Ridge regression works great with the data having high multicollinearity.

Disadvantages: Ridge regression trades variance for bias which means the result is not unbiased. All the predictors are included in the final model. The coefficient term shrinks towards Zero.

72) What do you understand by long and wide data formats?

73) What do you understand by outliers and inliers? What would you do if you find them in your dataset?

In your dataset when, when the data points are several standards away from the mean, we call those data points outliers. In a dataset when the points are in the interior of the distribution where most of the data occur, these points are known as inliers. There are various ways to deal with the outliers and inliers, the most common being completely removing them from your dataset. But this



dealing of outliers differs according to the data under consideration. There are mainly 3 ways one can deal with them, either keep them in the data, remove them or change them to another variable.

74) Write a program in Python that takes input as the diameter of a coin and weight of the coin and produces output as the money value of the coin.

75) What are the basic assumptions to be made for linear regression? ([get sample code here](#))

Normality of error distribution, statistical independence of errors, linearity, and additivity.

76) Can you write the formula to calculate R-square?

R-Square can be calculated using the below formula -

$1 - (\text{Residual Sum of Squares} / \text{Total Sum of Squares})$

77) What is the advantage of performing dimensionality reduction before fitting an SVM?

Support Vector Machine Learning Algorithm performs better in the reduced space. It is beneficial to perform dimensionality reduction before fitting an SVM if the number of features is large when compared to the number of observations.

78) How will you assess the statistical significance of insight whether it is a real insight or just by chance?

The statistical importance of insight can be accessed using Hypothesis Testing.

79) How would you create a taxonomy to identify key customer trends in unstructured data?

[Tweet: Data Science Interview questions #1 - How would you create a taxonomy to identify key customer trends in unstructured data? - http://ctt.ec/sdqZ0+](#)

The best way to approach this question is to mention that it is good to check with the business owner and understand their objectives before categorizing the data. Having done this, it is always good to follow an iterative approach by pulling new data samples and improving the model accordingly by validating it for accuracy by soliciting feedback from the stakeholders of the business. This helps ensure that your model is producing actionable results and improving over the time.

80) How will you find the correlation between a categorical variable and a continuous variable ?

You can use the analysis of covariance technique to find the correlation between a categorical variable and a continuous variable.

81) Is it better to have too many false negatives or too many false positives?

When one receives a positive result of a test but should have received a negative result is known as a false positive. Similarly when a positive result is expected out of a test but a negative result is received is known as a false negative. The answer to the question is entirely dependent on the application.



For example, a cancer screening test is negative for many patients but the doctor expects it to be positive for the maximum number of patients. Here there are many false negatives which is not a good thing, as if the patient really has cancer and would not get treated immediately may suffer a lot and eventually succumb.

82) In experimental design, is it necessary to do randomization? If yes, why?

83) What are the benefits of using a convolutional neural network over a fully connected network when working with image classification problems?

84) What are the benefits of using a recurrent neural network over a fully connected network when working with text data?

85) What do you understand by feature vectors?

86) How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for a large number of outliers, the values can be substituted with either the 99th or the 1st percentile values. All extreme values are not outlier values. The most common ways to treat outlier values –

1) To change the value and bring in within a range

2) To just remove the value.

87) How can you assess a good logistic model?

There are various methods to assess the results of logistic regression analysis-

- Using Classification Matrix to look at the true negatives and false positives.
- Concordance that helps identify the ability of the logistic model to differentiate between the event happening and not happening.
- Lift helps assess the logistic model by comparing it with random selection.

88) What are categorical variables?

A variable that takes one of a limited set of values, usually fixed, is known as a categorical variable. A categorical variable can only take discrete values, unlike the continuous variable which can take an unlimited number of values. As suggested by the name, categorical variables have limited categories or levels. For example, a variable representing the blood type of a human can only take A, AB, B, O values, which is a categorical variable. Ideally, the height of humans can take any positive value which can be termed as a continuous variable. A special type of categorical variable that can only take two values is known as a binary variable.

89) What is the benefit of weight initialization in neural networks?

90) How does the use of dropout work as a regularizer for deep neural networks?

91) How beneficial is dropout regularization in deep learning models? Does it speed up or slow down the training process and why?



92) How will you explain logistic regression to an economist, physician-scientist, and biologist?

93) What is the benefit of batch normalization?

94) Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm and vice-versa.

95) What is multicollinearity and how you can overcome it?

SVM and Random Forest are both used in classification problems.

a) If you are sure that your data is outlier free and clean then go for SVM. It is the opposite - if your data might contain outliers then Random forest would be the best choice

b) Generally, SVM consumes more computational power than Random Forest, so if you are constrained with memory go for the Random Forest [machine learning algorithm](#).

c) Random Forest gives you a very good idea of variable importance in your data, so if you want to have variable importance then choose the Random Forest machine learning algorithm.

d) Random Forest machine learning algorithms are preferred for multiclass problems.

e) SVM is preferred in multi-dimensional problem set - like text classification

but as a good data scientist, you should experiment with both of them and test for accuracy, or rather you can use an ensemble of many Machine Learning techniques.

96) What is the curse of dimensionality?

97) Do you need to do feature engineering and feature extraction when applying deep learning models?

98) How will you calculate the accuracy of a model using a confusion matrix?

99) According to the universal approximation theorem, any function can be approximated as closely as required using single collinearity. Then why people use more?

100) Explain the use of Combinatorics in data science.

101) You are given a dataset with 1500 observations and 15 features. How many observations you will select in each decision tree in a random forest?

Each decision tree has a subset of features but includes all the observations from the dataset. In this case, the answer will be 1500 as the tree will include all the observations from the dataset.

102) How will you evaluate the performance of a logistic regression model?

It's very much obvious that you would mention accuracy as the answer to this question but since logistic regression is not the same as linear regression it will mislead. You should mention how you will use the confusion matrix to evaluate



the performance and the various statistics related to it like Precision, Specificity, Sensitivity, or Recall. You get bonus points for mentioning Concordance, Discordance, and AUC.

[Access Data Science and Machine Learning Project Code Examples](#)

Data Science Puzzles-Brain Storming/ Puzzle based Data Science Interview Questions asked in Data Scientist Job Interviews

1) How many Piano Tuners are there in Chicago?

To solve this kind of problem, we need to know –

Can you tell if the equation given below is linear or not?

$$\text{Emp_sal} = 2000 + 2.5(\text{emp_age})^2$$

Yes it is a linear equation as the coefficients are linear.

What will be the output of the following R programming code ?

```
var2 <- c("I", "Love", "ProjectPro")
```

```
var2
```

It will give an error.

How many Pianos are there in Chicago?

How often would a Piano require tuning?

How much time does it take for each tuning?

We need to build these estimates to solve this kind of a problem. Suppose, let's assume Chicago has close to 10 million people and on an average there are 2 people in a house. For every 20 households there is 1 Piano. Now the question how many pianos are there can be answered. 1 in 20 households has a piano, so approximately 250,000 pianos are there in Chicago.

Now the next question is-“How often would a Piano require tuning? There is no exact answer to this question. It could be once a year or twice a year. You need to approach this question as the interviewer is trying to test your knowledge on whether you take this into consideration or not. Let's suppose each piano requires tuning once a year so on the whole 250,000 piano tunings are required.

Let's suppose that a piano tuner works for 50 weeks in a year considering a 5 day week. Thus a piano tuner works for 250 days in a year. Let's suppose tuning a piano takes 2 hours then in an 8 hour workday the piano tuner would be able to tune only 4 pianos. Considering this rate, a piano tuner can tune 1000 pianos a year.

Thus, 250 piano tuners are required in Chicago considering the above estimates.

2) There is a race track with five lanes. There are 25 horses of which you want to find out the three fastest horses. What is the minimal number of races needed to identify the 3 fastest horses of those 25?

Divide the 25 horses into 5 groups where each group contains 5 horses. Race between all the 5 groups (5 races) will determine the winners of each group. A race between all the winners will determine the winner of the winners and must be the fastest horse. A final race between the 2nd and 3rd place from the

be the fastest horse. A final race between the 2nd and 3rd place from the winners group along with the 1st and 2nd place of the second place group along with the third place horse will determine the second and third fastest horse from the group of 25.

- 3) Estimate the number of french fries sold by McDonald's everyday.
- 4) How many times in a day does a clock's hand overlap?
- 5) You have two beakers. The first beaker contains 4 litre of water and the second one contains 5 litres of water.How can you our exactly 7 litres of water into a bucket?
- 6) A coin is flipped 1000 times and 560 times heads show up. Do you think the coin is biased?
- 7) Estimate the number of tennis balls that can fit into a plane.
- 8) How many haircuts do you think happen in US every year?
- 9) In a city where residents prefer only boys, every family in the city continues to give birth to children until a boy is born. If a girl is born, they plan for another child. If a boy is born, they stop. Find out the proportion of boys to girls in the city.

Probability Interview Questions for Data

Science

1. There are two companies manufacturing electronic chip. Company A is manufactures defective chips with a probability of 20% and good quality chips with a probability of 80%. Company B manufactures defective chips with a probability of 80% and good chips with a probability of 20%.If you get just one electronic chip, what is the probability that it is a good chip?
2. Suppose that you now get a pack of 2 electronic chips coming from the same company either A or B. When you test the first electronic chip it appears to be good. What is the probability that the second electronic chip you received is also good?
3. A dating site allows users to select 6 out of 25 adjectives to describe their likes and preferences. A match is said to be found between two users on the website if the match on atleast 5 adjectives. If Steve and On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match is declared between two users if they match on at least 4 adjectives. If Brad and Angelina randomly pick adjectives, what is the probability that they will form a match?
4. A coin is tossed 10 times and the results are 2 tails and 8 heads. How will you analyze whether the coin is fair or not? What is the p-value for the same?
5. Continuation to the above question, if each coin is tossed 10 times (100 tosses are made in total). Will you modify your approach to test the fairness of the coin or continue with the same?
6. An ant is placed on an infinitely long twig. The ant can move one step backward or one step forward with the same probability during discrete time steps. Find out the probability with which the ant will return to the starting point.

Statistics Interview Questions for Data

Science

1. Explain the central limit theorem.
2. What is the relevance of the central limit theorem to a class of freshmen in the social sciences who hardly have any knowledge about statistics?
3. Given a dataset, show me how Euclidean Distance works in three dimensions.
4. How will you prevent overfitting when creating a statistical model?

Python Data Science Interview Questions

- 1) Explain about the range function.
- 2) How can you freeze an already built machine learning model for later use ?
What is the command you would use?
- 3) Differentiate between func anf func()
- 4) Write the command to import a decision tree classification algorithm using sklearn library.
- 5) What do you understand by pickling in Python?

Frequently Asked Open-Ended Machine

Learning Interview Questions for Data

Scientists

1. Which is your favorite machine learning algorithm and why?
2. In which libraries for Data Science in Python and R, does your strength lie?
3. What kind of data is important for specific business requirements and how, as a data scientist will you go about collecting that data?
4. Tell us about the biggest data set you have processed to date and for what kind of analysis.
5. Which data scientists you admire the most and why?
6. Suppose you are given a data set, what will you do with it to find out if it suits the business needs of your project or not.
7. What were the business outcomes or decisions for the projects you worked on?
8. What unique skills you think can you add to our data science team?
9. Which are your favorite data science startups?
10. Why do you want to pursue a career in data science?
11. What have you done to upgrade your skills in analytics?
12. What has been the most useful business insight or development you have found?
13. How will you explain an A/B test to an engineer who does not know statistics?
14. When does parallelism helps your algorithms run faster and when does it make them run slower?
15. How can you ensure that you don't analyze something that ends up producing meaningless results?
16. How would you explain to the senior management in your organization why a particular data set is important?



17. Is more data always better?
18. What are your favorite imputation techniques to handle missing data?
19. What are your favorite data visualization tools?
20. Explain the life cycle of a data science project.
21. What according to you are the limitations of deep learning?

Suggested Answers by Data Scientists for Open-Ended Data Science Interview Questions

How can you ensure that you don't analyze something that ends up producing meaningless results?

- Understanding whether the model chosen is correct or not. Start understanding from the point where you did Univariate or Bivariate analysis, analyzed the distribution of data and correlation of variables and built the linear model. Linear regression has an inherent requirement that the data and the errors in the data should be normally distributed. If they are not then we cannot use linear regression. This is an inductive approach to find out if the analysis using linear regression will yield meaningless results or not.
- Another way is to train and test data sets by sampling them multiple times. Predict on all those datasets to find out whether or not the resultant models are similar and are performing well.
- By looking at the p-value, by looking at r square values, by looking at the fit of the function and analyzing as to how the treatment of missing value could have affected- data scientists can analyze if something will produce meaningless results or not.

- Gaganpreet Singh, Data Scientist

So, there you have over 120 data science interview questions and answers for most of them too. These are some of the more common interview questions for data scientists around data, statistics, and data science that can be asked in the interviews. We will come up with more questions – specific to language, Python/ [R](#), in the subsequent articles, and fulfill our goal of providing 120 data science interview questions PDF with answers to our readers.

3 Secrets to becoming a Great Enterprise Data Scientist

- Keep on adding technical skills to your data scientist's toolbox.
- Improve your scientific axiom
- Learn the language of business as the insights from a data scientist help in reshaping the entire organization.

The important tip, to nail a data science interview is to be confident with the answers without bluffing. If you are well-versed with a particular technology whether it is Python, R, [Hadoop](#), Spark or any other big data technology ensure that you can back this up but if you are not strong in a particular area do not mention it unless asked about it. The above list of data scientist job interview questions is not an exhaustive one. Every company has a different approach to interviewing data scientists. However, we do hope that the above data science technical interview questions elucidate the data science interview process and provide an understanding of the type of data scientist job interview questions asked when companies are hiring data people.

We request industry experts and data scientists to chime in their suggestions in

we request industry experts and data scientists to chime in their suggestions in comments for open-ended data science interview questions to help students understand the best way to approach the interviewer and help them nail the interview.

Related Posts

[Python Data Science Interview Questions](#)

[Data Science Interview Questions for R](#)

[Data Scientist Interview Questions asked at Top Tech Companies](#)

[Data Analyst Interview Questions](#)

[PREVIOUS](#)

[NEXT](#)



Trending Project Categories

- Machine Learning Projects
- Data Science Projects
- Deep Learning Projects
- Big Data Projects
- Apache Hadoop Projects
- Apache Spark Projects
- Show more

Trending Projects

- Walmart Sales Forecasting Data Science Project
- BigMart Sales Prediction ML Project
- Music Recommender System Project
- Credit Card Fraud Detection Using Machine Learning
- Resume Parser Python Project for Data Science
- Time Series Forecasting Projects
- Show more

Trending Blogs

- Machine Learning Projects for Beginners with Source Code
- Data Science Projects for Beginners with Source Code
- Big Data Projects for Beginners with Source Code
- IoT Projects for Beginners with Source Code
- Data Analyst vs Data Scientist
- Data Science Interview Questions and Answers
- Show more

Trending Recipes

- Search for a Value in Pandas DataFrame
- Pandas Create New Column based on Multiple Condition
- LSTM vs GRU
- Plot ROC Curve in Python
- Python Upload File to Google Drive
- Optimize Logistic Regression Hyper Parameters
- Show more

Trending Tutorials

- PCA in Machine Learning Tutorial
- PySpark Tutorial
- Hive Commands Tutorial
- MapReduce in Hadoop Tutorial
- Apache Hive Tutorial -Tables
- Linear Regression Tutorial
- Show more

[Contact us](#)

[Privacy policy](#)

[User policy](#)

