100 Data Science Interview Questions and Answers (General) for 2017

01 Dec 2015

Latest Update made on June 17,2017.

In collaboration with data scientists, industry experts and top counsellors, 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, focusses only on the general 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

interview. These kind of analytics interview questions also measure if you were successful in applying data science techniques to real life problems.

If you would like more information about Online Data Science course, please click the orange "Request Info" button on top of this page.

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.

DeZyre

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 Data Scientist Interview Questions and Answers

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

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.

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

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

3) Which technique is used to predict categorical responses?

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

- HDFS Interview Questions and Answers
- Real-Time Hadoop Interview Questions and Answers
- Hadoop Admin Interview Questions and Answers
- Basic Hadoop Interview Questions and Answers
- Apache Spark Interview Questions and Answers
- Data Analyst Interview Questions and Answers
- 100 Data Science Interview Questions and Answers (General)

logistic regression recently.

Logistic Regression often referred as 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 analysis?

Cleaning data from multiple sources to transform it into a format that data analysts or data scientists can work with is

- 100 Data Science in Python Interview Questions and Answers
- Recap of Data Science News for June 2017
- Recap of Apache Spark News for June 2017
- Recap of Hadoop News for June 2017
- Top Machine Learning Interview Questions and Answers for 2017
- Hadoop Cluster Overview: What it is and how to setup one?
- Spark SQL for Relational Big

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 analysis task.

7) Differentiate between univariate, bivariate and multivariate analysis.

These are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of 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 time as in a scatterplot, then it is referred to as bivariate analysis. For example, analysing the volume of sale and a 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.

3.0 -Features and Enhancements

- Recap of Data Science News for May 2017
- Recap of Apache Spark News for May 2017
- Recap of Hadoop News for May 2017

Blog Categories

- Big Data
- CRM
- Data Science
- Mobile App Development
- NoSQL Database
- Web Development

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 an symmetrical bell shaped curve.

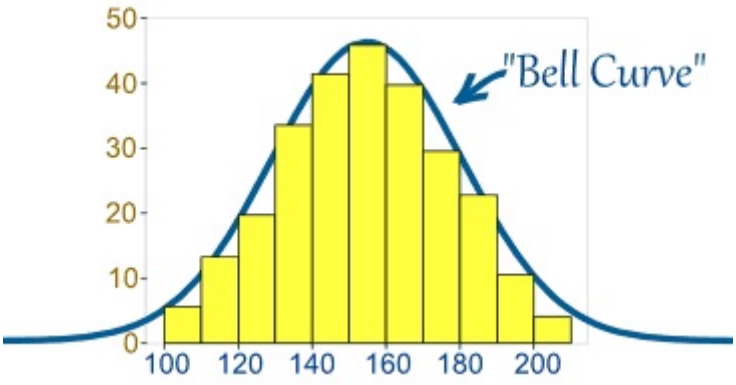


Image Credit : mathisfun.com

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 is Interpolation and Extrapolation?

- 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:

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?

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. 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

method.

15) Are expected value and mean value different?

They are not different but the terms are used in different contexts. Mean is generally referred when talking about a probability distribution or sample population whereas expected value is generally referred in a random variable context.

For Sampling Data

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. Expected value is the population mean.

For Distributions

Mean value and Expected value are 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?

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

17) Do gradient descent methods always converge to 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 are categorical variables?

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,

- 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

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 times the test will end up saying you have the illness and 95% of the times the test will give 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 true positive result.

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

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 an Unsupervised Learning?

an example for 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 for unsupervised learning.

22) Explain the use of Combinatorics in data science.

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

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

It is a statistical hypothesis testing for 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 an interest. An example for 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

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

26) What is Gradient Descent?

27) 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 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.

28) How can you assess a good logistic model?

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

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.

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

- Understand the business problem
- Explore the data and become familiar with it.
- Prepare the data for modelling by detecting outliers, treating missing values, transforming variables, etc.
- After data preparation, start running the model, analyse 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.

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?

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

- 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 about the box cox transformation in regression models.

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.

First solution which will come to your mind is to merge two lists and short them afterwards

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])
```

```
  return final_sorted_list
```

Similar function can be returned in R as well by following the 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)
```

```
  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)))

}
```

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

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

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

38) What is the curse of dimensionality?

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

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.

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

43) How will you explain logistic regression to an economist, physican scientist and biologist?

44) How can you overcome Overfitting?

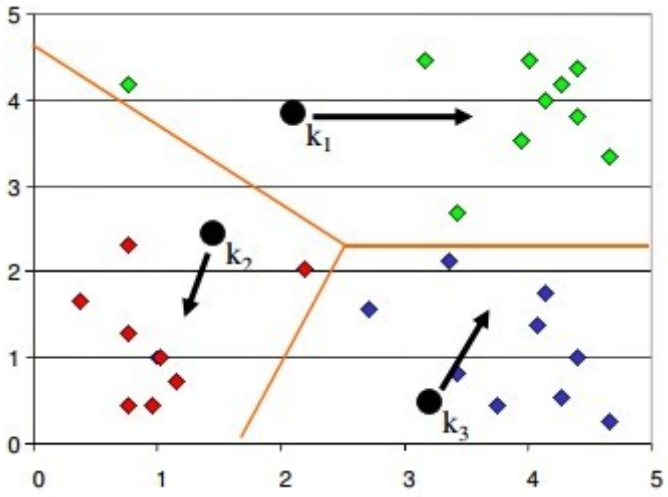
45) Differentiate between wide and tall data formats?

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

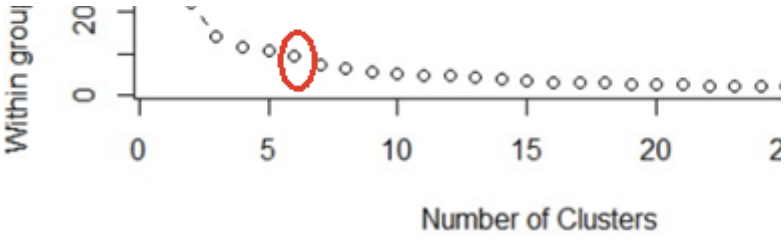
47) How would you develop a model to identify plagiarism?

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

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

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

using independent variables, i.e.

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

	A	B	C	D	E
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	=$E$2+$E$3*A8+$E$4*B	
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		

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?

53) You created a predictive model of a quantitative outcome variable using multiple regressions. What are the

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

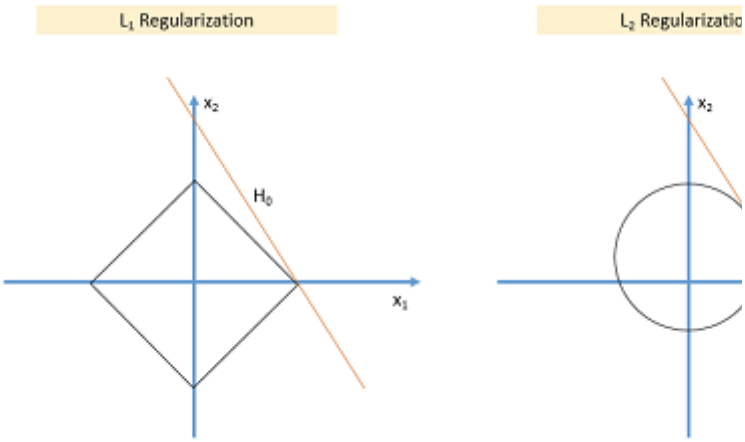
· Assumptions of linear regression

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

55) What do you understand by Recall and Precision?

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 L2 which results into sparsity.

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

modelling?

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

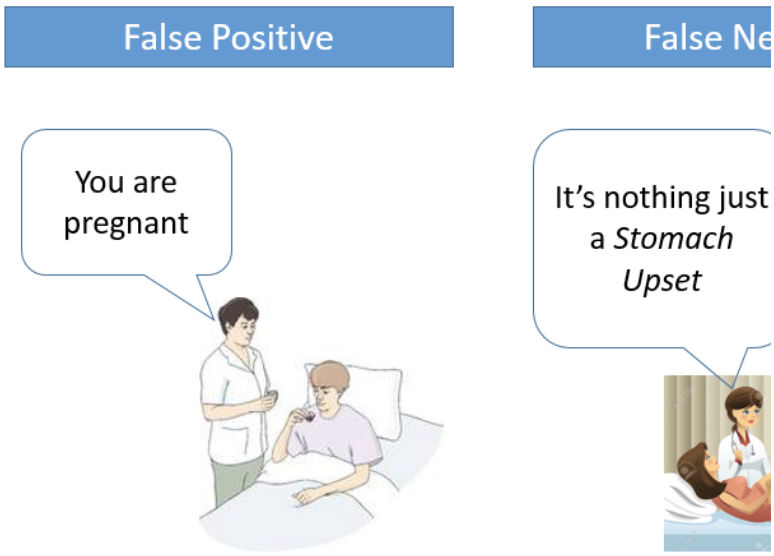
60) What do you understand by conjugate-prior with respect to Naïve Bayes?

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

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

marketing. Lets 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?

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?

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 be**come 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.

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 leaning 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".

Calculation of seasonality is pretty straight forward-

Seasonality = 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?

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

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 Random Forest machine learning algorithm.

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 ensemble of many Machine Learning techniques.

69) What do you understand by feature vectors?

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 into selection bias. Let see few missing value treatment examples and their impact on selection-

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 correlation matrix for data so you might remove the missing values from variables which 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 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?

and inliers? What would you do if you find them in your dataset?

74) Write a program in Python which 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?

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

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

R-Square can be calculated using the below formular -

1 - (Residual Sum of Squares/ 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

significance of an insight whether it is a real insight or just by chance?

Statistical importance of an insight can be accessed using Hypothesis Testing.

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 a problem, we need to know –

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

Emp_sal= 2000+2.5(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 ?

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.

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 2^{nd} and 3^{rd} place from the winners group along with the 1^{st} and 2^{nd} place of thee second place group along with the third place horse will determine the second and third fastest horse from the group of 25.

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

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

each coin is tossed 10 times (100 tosses are made in total). Will you modify your approach to the 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 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 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 ?

1. Which is your favourite 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 till 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 on to our data science team?
9. Which are your favorite data science startups?

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 analyse something that ends up producing meaningless results?
16. How would you explain to the senior management in your organization as to why a particular data set is important?
17. Is more data always better?
18. What are your favourite imputation techniques to handle missing data?
19. What are your favorite data visualization tools?
20. Explain the life cycle of a data science project.

How can you ensure that you don't analyse 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, analysed 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 analysing as to how the treatment of missing value could

- Gaganpreet Singh,Data Scientist

These are some of the more general questions 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 fulfil our goal of providing a set of 100 data science interview questions and answers.

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 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 unless asked about it. The above list of data

above data science technical interview questions elucidate the data science interview process and provide an understanding on 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 comments for open ended data science interview questions to help students understand the best way to apporach the interviewer and help them nail the interview.If you have any words of wisdom for data science students to ace a data science interview, share with us in comments below!

Related Posts

Data Science Interview Questions for Python

Data Science interview Questions for R

Data Scientist Interview Questions asked at Top Tech Companies

Data Analyst Interview Questions

PREVIOUS NEXT

Utkarsh Kasana • 9 months ago

Answer to Q12 is missing. Please look into that!
Edit: Most of the questions are unanswered!

Reply • Share ›

Ashok Matta • 9 months ago

Few questions i have encountered are :
1) What are false positives ?
2) What are the methods for anomaly detection in time series data?
3) How can you fill data gaps ?
4) My data science coding exercise
Task Description

Your task is to rapidly load, understand, and build a machine learning model for house price data.

Your deliverable should include:
* A machine learning model trained on the SalePrice (hint: train the model on log(SalePrice))
* An evaluation of the predictions, score them using RMSE of the log(prediction) vs. the log(SalePrice)

see more

Reply • Share ›

data Jing • a year ago

Normal distribution is symmetric. There is a typo in the post.

Reply • Share ›

Khushbu Shah Mod ➚ data Jing • a year ago

Thanks Data Jing for letting us know on the typo mistake. It has been corrected.

Reply • Share ›

poor understanding.

Reply • Share ›

Khushbu Shah Mod
greyswan • 2 years ago

Hey Greyswan,

Glad that you liked the starter list of questions. Well yes, it is true that the answers are not detailed and sufficient enough to crack the interviews. But the idea here is to give an idea of what kind of questions to expect in these job interviews. We are still working on getting quality answers from industry experts and will keep updating the content further.
Also this is just part 1 of the data science interview questions and answers.We will upload more questions on Python and R that can be asked in data science job interviews.

Reply • Share ›

Suraj Rastogi
Khushbu Shah
• 2 years ago

still khusbhu instead off researching every single question do we have pdf or something wherein we can find all these answers at one place

Reply • Share ›

Khushbu Shah
Mod Suraj Rastogi
• 2 years ago

Hey Suraj,
We understand your

the blog and have quality answers for all the questions from the experts. We

Big Data and Hadoop Training Courses in Popular Cities

- Hadoop Training in Texas
- Hadoop Training in California
- Hadoop Training in Dallas
- Hadoop Training in Chicago
- Hadoop Training in Charlotte
- Hadoop Training in Dubai
- Hadoop Training in Edison
- Hadoop Training in Fremont
- Hadoop Training in San Jose
- Hadoop Training in Washington

- Hadoop Training in New Jersey
- Hadoop Training in New York
- Hadoop Training in Atlanta