DreambigCareer

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.

4) What is logistic regression? Or State an example when you have used 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 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 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.

8) What do you understand by the term Normal Distribution?

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 an symmetrical bell shaped curve.

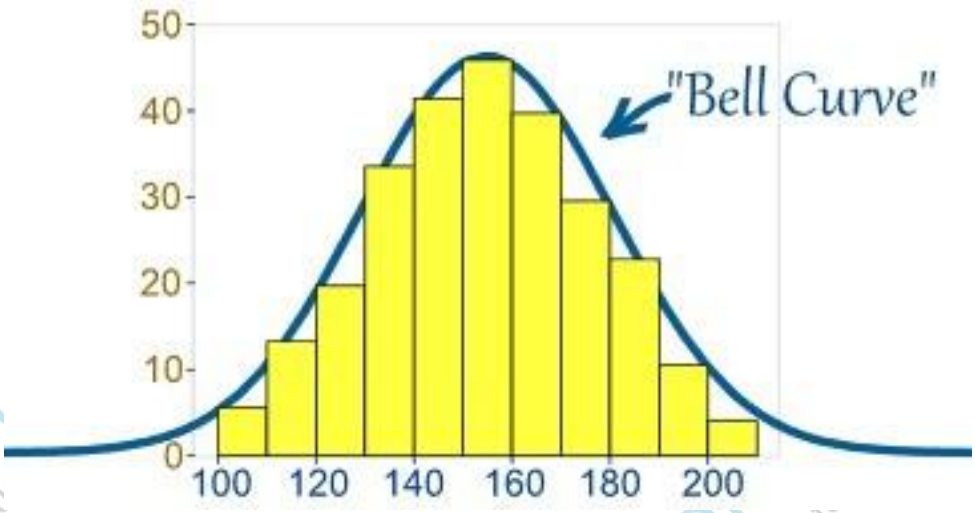

Image Credit : mathisfun.com

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?

Estimating a value from 2 unknown 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?

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 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 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 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?

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=0.05is 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, 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 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?

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 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 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 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-

- 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.

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.

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?

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 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 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

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 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)
   {
    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)))
}
```

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 the curse of dimensionality?

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

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

40) 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.

38) How are confidence intervals constructed and how will you interpret them?
39) How will you explain logistic regression to an economist, physican scientist and biologist?
40) How can you overcome Overfitting?
41) Differentiate between wide and tall data formats?
42) Is Naïve Bayes bad? If yes, under what aspects.
43) How would you develop a model to identify plagiarism?
44) How will you define the number of clusters in a clustering algorithm?
45) Is it better to have too many false negatives or too many false positives?
46) Is it possible to perform logistic regression with Microsoft Excel?
47) What do you understand by Fuzzy merging ? Which language will you use to handle it?
48) What is the difference between skewed and uniform distribution?
49) You created a predictive model of a quantitative outcome variable using multiple regressions. What are the steps you would follow to validate the model?
50) What do you understand by Hypothesis in the content of Machine Learning?
51) What do you understand by Recall and Precision?
52) How will you find the right K for K-means?
53) Why L1 regularizations causes parameter sparsity whereas L2 regularization does not?
54) How can you deal with different types of seasonality in time series modelling?
55) In experimental design, is it necessary to do randomization? If yes, why?
56) What do you understand by conjugate-prior with respect to Naïve Bayes?
57) Can you cite some examples where a false positive is important than a false negative?

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

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

60) **Can you explain the difference between a Test Set and a Validation 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 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.

61) What makes a dataset gold standard?

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

63) What is the importance of having a selection bias?

64) 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

65) What do you understand by feature vectors?

66) How do data management procedures like missing data handling make selection bias worse?
67) What are the advantages and disadvantages of using regularization methods like Ridge Regression?
68) What do you understand by long and wide data formats?
69) What do you understand by outliers and inliers? What would you do if you find them in your dataset?
70) 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.

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

1) How many Piano Tuners are there in Chicago?

To solve this kind of a problem, we need to know –
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 winners group along with the 1st and 2nd 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.

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 and Statistics Interview Questions for Data Science

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?
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?

A coin is tossed 10 times and the results are 2 tails and 8 heads. How will you analyse whether the coin is fair or not? What is the p-value for the same?
Continuation to the above question, if 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?

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.

Frequently Asked Open Ended Interview Questions for Data Scientists

Which is your favourite machine learning algorithm and why?
In which libraries for Data Science in Python and R, does your strength lie?
What kind of data is important for specific business requirements and how, as a data scientist will you go about collecting that data?
Tell us about the biggest data set you have processed till date and for what kind of analysis.
Which data scientists you admire the most and why?

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.
What were the business outcomes or decisions for the projects you worked on?
What unique skills you think can you add on to our data science team?
Which are your favorite data science startups?
Why do you want to pursue a career in data science?
What have you done to upgrade your skills in analytics?
What has been the most useful business insight or development you have found?
How will you explain an A/B test to an engineer who does not know statistics?
When does parallelism helps your algorithms run faster and when does it make them run slower?
How can you ensure that you don't analyse something that ends up producing meaningless results?
How would you explain to the senior management in your organization as to why a particular data set is important?
Is more data always better?
What are your favourite imputation techniques to handle missing data?
What are your favorite data visualization tools?
Explain the life cycle of a data science project.

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

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 have affected- data scientists can analyse if something will produce meaningless results or not.

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 scientist job interview questions is not an exhaustive one. Every company has a different approach for 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 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!