**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 asymmetrical bell shaped curve.

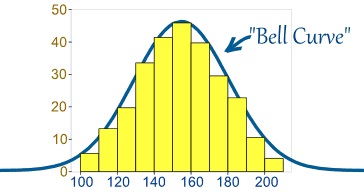


Image Credit : mathisfun.com

**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.

**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?**

**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?**

**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.

**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.

**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.**

**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 data?**

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

Learn  [Data Science in Python](https://www.dezyre.com/data-science-in-python-online-training/36" \o "IBM Certified Data Science in Python Training" \t "_blank) and [R  Programming](https://www.dezyre.com/data-science-in-R-programming-online-training/37" \o "Data Science in R Programming" \t "_blank) to nail data science interviews at top tech companies!

**Frequently Asked Open Ended Interview Questions for Data Scientists**

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)         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.

6)         What were the business outcomes or decisions for the projects you worked on?

7)         What unique skills you think can you add on to our data science team?

8)         Why do you want to pursue a career in data science?

9)         What have you done to upgrade your skills in analytics?

10)       What has been the most useful business insight or development you have found?

11)       How will you explain an A/B test to an engineer who does not know statistics?

12)       When does parallelism helps your algorithms run faster and when does it make them run slower?

13)       How can you ensure that you don’t analyse something that ends up producing meaningless results?

14) How would you explain to the senior management in your organization as to why a particular data set is important?

15) Is more data always better?

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](https://www.dezyre.com/Hadoop-Training-online/19" \o "IBM Certified Big Data Hadoop Training Online" \t "_blank) 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.

If you have any words of wisdom for data science students to ace a data science interview, share with us in comments below!

1. What is the biggest data set that you processed, and how did you process it, what were the results?
2. Tell me two success stories about your analytic or computer science projects? How was lift (or success) measured?
3. What is: lift, KPI, robustness, model fitting, design of experiments, 80/20 rule?
4. What is: collaborative filtering, n-grams, map reduce, cosine distance?
5. How to optimize a web crawler to run much faster, extract better information, and better summarize data to produce cleaner databases?
6. How would you come up with a solution to identify plagiarism?
7. How to detect individual paid accounts shared by multiple users?
8. Should click data be handled in real time? Why? In which contexts?
9. What is better: good data or good models? And how do you define "good"? Is there a universal good model? Are there any models that are definitely not so good?
10. What is probabilistic merging (AKA fuzzy merging)? Is it easier to handle with SQL or other languages? Which languages would you choose for semi-structured text data reconciliation?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is your favorite programming language / vendor? why?
13. Tell me 3 things positive and 3 things negative about your favorite statistical software.
14. Compare SAS, R, Python, Perl
15. What is the curse of big data?
16. Have you been involved in database design and data modeling?
17. Have you been involved in dashboard creation and metric selection? What do you think about Birt?
18. What features of Teradata do you like?
19. You are about to send one million email (marketing campaign). How do you optimze delivery? How do you optimize response? Can you optimize both separately? (answer: not really)
20. Toad or Brio or any other similar clients are quite inefficient to query Oracle databases. Why? How would you do to increase speed by a factor 10, and be able to handle far bigger outputs?
21. How would you turn unstructured data into structured data? Is it really necessary? Is it OK to store data as flat text files rather than in an SQL-powered RDBMS?
22. What are hash table collisions? How is it avoided? How frequently does it happen?
23. How to make sure a mapreduce application has good load balance? What is load balance?
24. Examples where mapreduce does not work? Examples where it works very well? What are the security issues involved with the cloud? What do you think of EMC's solution offering an hybrid approach - both internal and external cloud - to mitigate the risks and offer other advantages (which ones)?
25. Is it better to have 100 small hash tables or one big hash table, in memory, in terms of access speed (assuming both fit within RAM)? What do you think about in-database analytics?
26. Why is naive Bayes so bad? How would you improve a spam detection algorithm that uses naive Bayes?
27. Have you been working with white lists? Positive rules? (In the context of fraud or spam detection)
28. What is star schema? Lookup tables?
29. Can you perform logistic regression with Excel? (yes) How? (use linest on log-transformed data)? Would the result be good? (Excel has numerical issues, but it's very interactive)
30. Have you optimized code or algorithms for speed: in SQL, Perl, C++, Python etc. How, and by how much?
31. Is it better to spend 5 days developing a 90% accurate solution, or 10 days for 100% accuracy? Depends on the context?
32. Define: quality assurance, six sigma, design of experiments. Give examples of good and bad designs of experiments.
33. What are the drawbacks of general linear model? Are you familiar with alternatives (Lasso, ridge regression, boosted trees)?
34. Do you think 50 small decision trees are better than a large one? Why?
35. Is actuarial science not a branch of statistics (survival analysis)? If not, how so?
36. Give examples of data that does not have a Gaussian distribution, nor log-normal. Give examples of data that has a very chaotic distribution?
37. Why is mean square error a bad measure of model performance? What would you suggest instead?
38. How can you prove that one improvement you've brought to an algorithm is really an improvement over not doing anything? Are you familiar with A/B testing?
39. What is sensitivity analysis? Is it better to have low sensitivity (that is, great robustness) and low predictive power, or the other way around? How to perform good cross-validation? What do you think about the idea of injecting noise in your data set to test the sensitivity of your models?
40. Compare logistic regression w. decision trees, neural networks. How have these technologies been vastly improved over the last 15 years?
41. Do you know / used data reduction techniques other than PCA? What do you think of step-wise regression? What kind of step-wise techniques are you familiar with? When is full data better than reduced data or sample?
42. How would you build non parametric confidence intervals, e.g. for scores? (see the [AnalyticBridge theorem](http://www.analyticbridge.com/profiles/blogs/how-to-build-simple-accurate-data-driven-model-free-confidence-in" \t "_blank))
43. Are you familiar either with extreme value theory, monte carlo simulations or mathematical statistics (or anything else) to correctly estimate the chance of a very rare event?
44. What is root cause analysis? How to identify a cause vs. a correlation? Give examples.
45. How would you define and measure the predictive power of a metric?
46. How to detect the best rule set for a fraud detection scoring technology? How do you deal with rule redundancy, rule discovery, and the combinatorial nature of the problem (for finding optimum rule set - the one with best predictive power)? Can an approximate solution to the rule set problem be OK? How would you find an OK approximate solution? How would you decide it is good enough and stop looking for a better one?
47. How to create a keyword taxonomy?
48. What is a Botnet? How can it be detected?
49. Any experience with using API's? Programming API's? Google or Amazon API's? AaaS (Analytics as a service)?
50. When is it better to write your own code than using a data science software package?
51. Which tools do you use for visualization? What do you think of Tableau? R? SAS? (for graphs). How to efficiently represent 5 dimension in a chart (or in a video)?
52. What is POC (proof of concept)?
53. What types of clients have you been working with: internal, external, sales / finance / marketing / IT people? Consulting experience? Dealing with vendors, including vendor selection and testing?
54. Are you familiar with software life cycle? With IT project life cycle - from gathering requests to maintenance?
55. What is a cron job?
56. Are you a lone coder? A production guy (developer)? Or a designer (architect)?
57. Is it better to have too many false positives, or too many false negatives?
58. Are you familiar with pricing optimization, price elasticity, inventory management, competitive intelligence? Give examples.
59. How does Zillow's algorithm work? (to estimate the value of any home in US)
60. How to detect bogus reviews, or bogus Facebook accounts used for bad purposes?
61. How would you create a new anonymous digital currency?
62. Have you ever thought about creating a startup? Around which idea / concept?
63. Do you think that typed login / password will disappear? How could they be replaced?
64. Have you used time series models? Cross-correlations with time lags? Correlograms? Spectral analysis? Signal processing and filtering techniques? In which context?
65. Which data scientists do you admire most? which startups?
66. How did you become interested in data science?
67. What is an efficiency curve? What are its drawbacks, and how can they be overcome?
68. What is a recommendation engine? How does it work?
69. What is an exact test? How and when can simulations help us when we do not use an exact test?
70. What do you think makes a good data scientist?
71. Do you think data science is an art or a science?
72. What is the computational complexity of a good, fast clustering algorithm? What is a good clustering algorithm? How do you determine the number of clusters? How would you perform clustering on one million unique keywords, assuming you have 10 million data points - each one consisting of two keywords, and a metric measuring how similar these two keywords are? How would you create this 10 million data points table in the first place?
73. Give a few examples of "best practices" in data science.
74. What could make a chart misleading, difficult to read or interpret? What features should a useful chart have?
75. Do you know a few "rules of thumb" used in statistical or computer science? Or in business analytics?
76. What are your top 5 predictions for the next 20 years?
77. How do you immediately know when statistics published in an article (e.g. newspaper) are either wrong or presented to support the author's point of view, rather than correct, comprehensive factual information on a specific subject? For instance, what do you think about the official monthly unemployment statistics regularly discussed in the press? What could make them more accurate?
78. Testing your analytic intuition: [look at these three charts](http://www.analyticbridge.com/profiles/blogs/how-to-detect-a-pattern-problem-and-solution" \t "_blank). Two of them exhibit patterns. Which ones? Do you know that these charts are called scatter-plots? Are there other ways to visually represent this type of data?
79. You design a robust non-parametric statistic (metric) to replace correlation or R square, that (1) is independent of sample size, (2) always between -1 and +1, and (3) based on rank statistics. How do you normalize for sample size? Write an algorithm that computes all permutations of n elements. How do you sample permutations (that is, generate tons of *random* permutations) when n is large, to estimate the asymptotic distribution for your newly created metric? You may use this asymptotic distribution for normalizing your metric. Do you think that an exact theoretical distribution might exist, and therefore, we should find it, and use it rather than wasting our time trying to estimate the asymptotic distribution using simulations?
80. More difficult, technical question related to previous one. There is an obvious one-to-one correspondence between permutations of n elements and integers between 1 and n! Design an algorithm that encodes an integer less than n! as a permutation of n elements. What would be the reverse algorithm, used to decode a permutation and transform it back into a number? **Hint**: An intermediate step is to use the [factorial number system](http://en.wikipedia.org/wiki/Factorial_number_system" \t "_blank) representation of an integer. Feel free to check this reference online to answer the question. Even better, feel free to browse the web to find the full answer to the question (this will test the candidate's ability to quickly search online and find a solution to a problem without spending hours reinventing the wheel).
81. How many "useful" votes will a Yelp review receive? **My answer**: Eliminate bogus accounts ([read this article](http://www.analyticbridge.com/forum/topics/how-do-you-estimate-the-proportion-of-bogus-accounts-on-facebook" \t "_blank)), or competitor reviews (how to detect them: use taxonomy to classify users, and location - two Italian restaurants in same Zip code could badmouth each other and write great comments for themselves). Detect [fake likes](http://www.analyticbridge.com/profiles/blogs/invented-by-a-data-scientist-the-first-anti-scam" \t "_blank): some companies (e.g. [FanMeNow.com](http://www.fanmenow.com/buy-twitter-followers/" \t "_blank)) will charge you to produce fake accounts and fake likes. Eliminate prolific users who like everything, those who hate everything. Have a blacklist of keywords to filter fake reviews. See if IP address or IP block of reviewer is in a blacklist such as "Stop Forum Spam". Create honeypot to catch fraudsters.  Also watch out for disgruntled employees badmouthing their former employer. Watch out for 2 or 3 similar comments posted the same day by 3 users regarding a company that receives very few reviews. Is it a brand new company? Add more weight to trusted users (create a category of trusted users).  Flag all reviews that are identical (or nearly identical) and come from same IP address or same user. Create a metric to measure distance between two pieces of text (reviews). [Create a review or reviewer taxonomy](http://www.bigdatanews.com/profiles/blogs/fast-clustering-algorithms-for-massive-datasets" \t "_blank). Use [hidden decision trees](http://www.analyticbridge.com/forum/topics/hidden-decision-trees-vs" \t "_blank) to rate or score review and reviewers.
82. What did you do today? Or what did you do this week / last week?
83. What/when is the latest data mining book / article you read? What/when is the latest data mining conference / webinar / class / workshop / training you attended? What/when is the most recent programming skill that you acquired?
84. What are your favorite data science websites? Who do you admire most in the data science community, and why? Which company do you admire most?
85. What/when/where is the last data science blog post you wrote?
86. In your opinion, what is data science? Machine learning? Data mining?
87. Who are the best people you recruited and where are they today?
88. Can you estimate and forecast sales for any book, based on Amazon public data? Hint: [read this article](http://www.fonerbooks.com/surfing.htm" \t "_blank).
89. [What's wrong with this picture?](http://www.analyticbridge.com/forum/topics/job-interview-question-what-is-wrong-with-this-picture" \t "_blank)
90. Should removing stop words be Step 1 rather than Step 3, [in the search engine algorithm described here](http://www.datasciencecentral.com/profiles/blogs/building-better-search-tools-problems-and-solutions" \t "_blank)? **Answer**: Have you thought about the fact that mine and yours could also be stop words? So in a bad implementation, data mining would become data mine after stemming, then data. In practice, you remove stop words before stemming. So Step 3 should indeed become step 1.
91. [Experimental design and a bit of computer science with Lego's](http://www.analyticbridge.com/forum/topics/analytics-for-kids" \t "_blank)