1. What is meant by selection bias?
Answer: Selection bias is a type of error that arises when the researcher decides on whom he is going to conduct the study. It happens when the selection of participants takes place not randomly. Selection bias is also sometimes referred to as a selection effect. It works more effectively and sometimes if the selection bias is not taken into account, the conclusions of the study may go wrong.
2. What is a Boltzmann machine?
Answer: Boltzmann developed with simple learning algorithms that allow them to find the important information that was presented in the complex regularities in the data. These machines are generally used to optimize the quantity and the weights of the given problem. The learning program works very slow in networks due to many layers of feature detectors. When we consider Restricted Boltzmann Machines, this has a single algorithm feature detectors that make it faster compared to others.
3. What is the difference between Cluster and Systematic Sampling?
Answer: 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 of systematic sampling is equal probability method.
4. What is the Law of Large Numbers?
Answer: It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of frequency-style thinking. It says that the sample means, the sample variance and the sample standard deviation converge to what they are trying to estimate.
5. What are Eigenvectors and Eigenvalues?
Answer: 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.

6. Can you cite some examples where both false positive and false negatives are equally important?
Answer: 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. (E learning portal)
Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

7. What is logistic regression? State an example when you have used logistic regression recently.
Answer: Logistic Regression often referred to as the logit model is a technique to predict the binary outcome from a linear combination of predictor variables.
For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent on election campaigning of a particular candidate, the amount of time spent in campaigning, etc. (tableau training online)
8. What is the role of the Activation Function?
Answer: The Activation function is used to introduce non-linearity into the neural network helping it to learn more complex function. Without which the neural network would be only able to learn linear function which is a linear combination of its input data. An activation function is a function in an artificial neuron that delivers an output based on inputs.
9. What do you mean by cluster sampling and systematic sampling?
Answer: When studying the target population spread throughout a wide area becomes difficult and applying simple random sampling becomes ineffective, the technique of cluster sampling is used. A cluster sample is a probability sample, in which each of the sampling units is a collection or cluster of elements.
Following the technique of systematic sampling, elements are chosen from an ordered sampling

frame. The list is advanced in a circular fashion. This is done in such a way so that once the end of the list is reached, the same is progressed from the start, or top, again. (data science training)

10. Please explain Gradient Descent?

Answer: The degree of change in the output of a function relating to the changes made to the inputs is known as a gradient. It measures the change in all weights with respect to the change in error. A gradient can also be comprehended as the slope of a function.

Gradient Descent refers to escalating down to the bottom of a valley. Simply, consider this something as opposed to climbing up a hill. It is a minimization algorithm meant for minimizing a given activation function.

11. What do you know about Autoencoders?

Answer: Autoencoders are simplistic learning networks used for transforming inputs into outputs with minimum possible error. It means that the outputs resulted are very close to the inputs.

A couple of layers are added between the input and the output with the size of each layer smaller than the size pertaining to the input layer. An autoencoder receives unlabeled input that is encoded for reconstructing the output. (oracle apex training online)

12. How and by what methods data visualizations can be effectively used?

Answer: In addition to giving insights in a very effective and efficient manner, data visualization can also be used in such a way that it is not only restricted to bar, line or some stereotypic graphs. Data can be represented in a much more visually pleasing manner.

One thing has to be taken care of is to convey the intended insight or finding correctly to the audience. Once the baseline is set. Innovative and creative part can help you come up with better looking and functional dashboards. There is a fine line between the simple insightful dashboard and awesome looking 0 fruitful insight dashboards.

13. What is the common perception of visualization?

Answer: People think visualization as just charts and summary information. But they are beyond that and drive business with a lot of underlying principles. Learning design principles can help anyone build effective and efficient visualizations and this Tableau prep tool can drastically increase our time on focusing more important part. The only issue with Tableau is, it is paid and companies need to pay for leveraging that awesome tool.

14. Where to seek help in case of discrepancies in Tableau?

Answer: When you face any issue regarding Tableau, try searching in the Tableau community forum. It is one of the best places to get your queries answered. You can always write your question and get the query answered with an hour or a day. You can always post on LinkedIn and follow people.

15. Why is data cleaning essential in Data Science?

Answer: Data cleaning is more important in Data Science because the end results or the outcomes of the data analysis come from the existing data where useless or unimportant need to be cleaned periodically as of when not required. This ensures the data reliability & accuracy and also memory is freed up.

Data cleaning reduces the data redundancy and gives good results in data analysis where some large customer information exists and that should be cleaned periodically. In businesses like e-commerce, retail, government organizations contain large customer transaction information which is outdated and needs to be cleaned.

Depending on the amount or size of data, suitable tools or methods should be used to clean the data from the database or big data environment. There are different types of data existing in a data source such as dirty data, clean data, mixed clean and dirty data and sample clean data.

Modern data science applications rely on machine learning model where the learner learns from the existing data. So, the existing data should always be cleanly and well maintained to get sophisticated and good outcomes during the optimization of the system. (devops training videos)

16. What is A/B testing in Data Science?

Answers: A/B testing is also called Bucket Testing or Split Testing. This is the method of comparing and testing two versions of systems or applications against each other to determine which version of application performs better. This is important in the cases where multiple versions are shown to the customers or end-users in order to achieve the goals.

In the area of Data Science, this A/B testing is used to know which variable out of the existing two variables in order to optimize or increase the outcome of the goal. A/B testing is also called Design of

Experiment. This testing helps in establishing a cause and effect relationship between the independent and dependent variables.
This testing is also simply a combination of design experimentation or statistical inference. Significance, Randomization and Multiple Comparisons are the key elements of the A/B testing. The significance is the term for the significance of statistical tests conducted. Randomization is the core component of the experimental design where the variables will be balanced. Multiple comparisons are the way of comparing more variables in the case of customer interests that causes more false positives resulting in the requirement of correction in the confidence level of a seller in the area of e-commerce.

17. How Machine Learning Is Deployed In Real World Scenarios?
Answer: Here are some of the scenarios in which machine learning finds applications in the real world:
Ecommerce: Understanding customer churn, deploying targeted advertising, remarketing.
Search engine: Ranking pages depending on the personal preferences of the searcher
Finance: Evaluating investment opportunities & risks, detecting fraudulent transactions
Medicare: Designing drugs depending on the patient's history and needs
Robotics: Machine learning for handling situations that are out of the ordinary
Social media: Understanding relationships and recommending connections
Extraction of information: framing questions for getting answers from databases over the web.
18. What Is Power Analysis?
Answer: power analysis is a vital part of the experimental design. It is involved with the process of determining the sample size needed for detecting an effect of a given size from a cause with a certain degree of assurance. It lets you deploy a specific probability in a sample size constraint.
The various techniques of statistical power analysis and sample size estimation are widely deployed for making the statistical judgment that is accurate and evaluates the size needed for experimental effects in practice.

Power analysis lets you understand the sample size estimate so that they are neither high nor low. A low sample size there will be no authentication to provide reliable answers and if it is large there will be wastage of resources.

19. What Is K-means? How Can You Select K For K-means?
Answer: K-means clustering can be termed as the basic unsupervised learning algorithm. It is the method of classifying data using a certain set of clusters called K clusters. It is deployed for grouping data in order to find similarity in the data.
It includes defining the K centers, one each in a cluster. The clusters are defined into K groups with K being predefined. The K points are selected at random as cluster centers. The objects are assigned to their nearest cluster center. The objects within a cluster are as closely related to one another as possible and differ as much as possible to the objects in other clusters. K-means clustering works very well for large sets of data.

20. Why is resampling done?
Answer: Resampling is done in any of these cases:
Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points
Substituting labels on data points when performing significance tests
Validating models by using random subsets (bootstrapping, cross-validation)
21. What tools or devices help you succeed in your role as a data scientist?
Answer: This question's purpose is to learn the programming languages and applications the candidate knows and has experience using. The answer will show the candidate's need for additional training of basic programming languages and platforms or any transferable skills. This is vital to understand as it can cost more time and money to train if the candidate is not knowledgeable in all of the languages and applications required for the position. (python training)
22. Why do you want to work at this company as a data scientist?
Answer: The purpose of this question is to determine the motivation behind the candidate's choice of applying and interviewing for the position. Their answer should reveal their inspiration for working for the company and their drive for being a data scientist. It should show the candidate is pursuing the

position because they are passionate about data and believe in the company, two elements that can determine the candidate's performance. Answers to look for include:
Interest in data mining
Respect for the company's innovative practices
Desire to apply analytical skills to solve real-world issues with data
"I have a passion for working for data-driven, innovative companies. Your firm uses advanced technology to address everyday problems for consumers and businesses alike, which I admire. I also enjoy solving issues using an analytical approach and am passionate about incorporating technology into my work. I believe that my skills and passion match the company's drive and capabilities."

23. What are the differences between overfitting and underfitting?
Answer: In statistics and machine learning, one of the most common tasks is to fit a model to a set of training data, so as to be able to make reliable predictions on general untrained data.
In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfitting has poor predictive performance, as it overreacts to minor fluctuations in the training data.
Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

24. What is Machine Learning?
Answer: Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Closely related to computational statistics. Used to devise complex models and algorithms that lend themselves to a prediction which in commercial use is known as predictive analytics. (hadoop training)
25. Can you enumerate the various differences between Supervised and Unsupervised Learning?
Answer: Supervised learning is a type of machine learning where a function is inferred from labeled training data. The training data contains a set of training examples.
Unsupervised learning, on the other hand, is a type of machine learning where inferences are drawn from datasets containing input data without labeled responses. Following are the various other differences between the two types of machine learning:

Algorithms Used – Supervised learning makes use of Decision Trees, K-nearest Neighbor algorithm, Neural Networks, Regression, and Support Vector Machines. Unsupervised learning uses Anomaly Detection, Clustering, Latent Variable Models, and Neural Networks.
Enables – Supervised learning enables classification and regression, whereas unsupervised learning enables classification, dimension reduction, and density estimation
Use – While supervised learning is used for prediction, unsupervised learning finds use in analysis

26. What is underfitting?
Answer: Any prediction rate which has provides low prediction in the training error and the test error leads to a high business problem, if the error rate in training set is high and the error rate in the test set is also high, then we can conclude it as overfitting model.
27. How to understand the problems faced during data analysis?
Answer: Most of the problem faced during hands-on analysis or data science is because of poor understanding of the problem in hand and concentrating more on tools, end results and other aspects of the project.
Breaking the problem down to a granular level and understanding takes a lot of time and practice to master. Coming back to square one in data science projects can be seen in a lot of companies and even in your own project or kaggle problems.
28. What does SAS stand out to be the best over other data analytics tools?
Answer: Ease to understand: The provisions included in SAS are remarkably easy to learn. Further, it offers the most suitable option for those who already are aware of the SQL. On the other hand, R comes with a steep training cover which is supposed to be a low-level programming style.
Data Handling Capacities: it is at par the most leading tool which also includes the R& Python.

If it advances before handling the huge data, it is the best platform to engage Graphical Capacities: it comes with functional graphical capacities and has a limited knowledge field.

It is useful to customize the plots Better tool management: It benefits in a release the updates with regards to the controlled conditions.

This is the main reason why it is well tested. Whereas if you considered R&Python, it has open contribution also the risk of errors in the current development is also high.

29. What is the best Programming Language to use in Data Science?
Answer: Data Science can be handled by using programming languages like Python or R programming language. These two are the two most popular languages being used by the Data Scientists or Data Analysts. R and Python are open source and are free to use and came into existence during the 1990s.
Python and R have different advantages depending on the applications and required a business goal. Python is better to be used in the cases of repeated tasks or jobs and for data manipulations whereas R programming can be used for querying or retrieving datasets and customized data analysis.
Mostly Python is preferred for all types of data science applications where some time R programming is preferred in the cases of high or complex data applications. Python is easier to learn and has less learning curve whereas R has a deep learning curve.

Python is mostly preferred in all the cases which is a general-purpose programming language and can be found in many applications other than Data Science too. R is mostly seen in Data Science area only where it is used for data analysis in standalone servers or computing separately.

30. What is a Linear Regression in Data Science?
Answer: This is the frequently asked Data Science Interview Questions in an interview. Linear Regression is a technique used in supervised machine learning the algorithmic process in the area of Data Science. This method is used for predictive analysis.
Predictive analytics is an area within Statistical Sciences where the existing information will be extracted and processed to predict the trends and outcomes pattern. The core of the subject lies in the analysis of existing context to predict an unknown event.
The process of Linear Regression method is to predict a variable called target variable by making the best relationship between the dependent variable and an independent variable. Here the dependent variable is the outcome variable and also response variable whereas the independent variable is the predictor variable or explanatory variable.
For example in real life, depending on the expenses occurred in this financial year or monthly expenses, the predictions happen by calculating the approximate upcoming months or financial years expenses.
In this method, the implementation can be done by using Python programming technique where this is the most important method used in Machine Learning technique under the area of Data Science.
Linear regression is also called Regression analysis that comes under the area of Statistical Sciences which is integrated together with Data Science.
31. What Is A Recommender System?
Answer: A recommender system is a today widely deployed in multiple fields like movie recommendations, music preferences, social tags, research articles, search queries and so on. The recommender systems work as per collaborative and content-based filtering or by deploying a personality-based approach. This type of system works based on a person's past behavior in order to build a model for the future. This will predict future product buying, movie viewing or book reading by people. It also creates a filtering approach using the discrete characteristics of items while recommending additional items.
32. How Do Data Scientists Use Statistics?
Answer: Statistics help Data Scientists to look into the data for patterns, hidden insights and convert Big Data into Big insights. It helps to get a better idea of what the customers are expecting. Data Scientists can learn about consumer behavior, interest, engagement, retention and finally conversion all through the power of insightful statistics. It helps them to build powerful data models in order to

validate certain inferences and predictions. All this can be converted into a powerful business proposition by giving users what they want at precisely when they want it.

33. What Do You Understand By The Term Normal Distribution?

Answer: It is a set of a continuous variable spread across a normal curve or in the shape of a bell curve. It can be considered as a continuous probability distribution and is useful in statistics. It is the most common distribution curve and it becomes very useful to analyze the variables and their relationships when we have the normal distribution curve.

The normal distribution curve is symmetrical. The non-normal distribution approaches the normal distribution as the size of the samples increases. It is also very easy to deploy the Central Limit Theorem. This method helps to make sense of data that is random by creating an order and interpreting the results using a bell-shaped graph.

34. What is collaborative filtering?

Answer: Filtering is a process used by recommender systems to find patterns and information from numerous data sources, several agents, and collaborating perspectives. In other words, the collaborative method is a process of making automatic predictions from human preferences or interests.

35. Explain the difference between overfitting and underfitting?

Answer: In machine learning as well as in statistics, the common task to undergo is to fit a model to a set of training data. It helps us in making reliable predictions using general untrained data.

In overfitting, a statistical model will help us in letting know the random noise or errors instead of the underlying relationship. Overfitting comes into light when the data is associated with too much complexity, which means it is associated with so many parameters relative to the number of observations. A model that is overfitted is always performed poor in predictive performance and acts overly to the minor fluctuations in the training data.

Unnderfittinng happens when a machine learning algorithm or statistical model is unable to focus on the underlying insights of the data. The case when you are trying to fix a linear model to a nonlinear one. This kind of model would result in poor predictive performance.

36. What is systematic sampling?

Answer: Systematic sampling is a technique, and the name resembles that it follows some systematic way, and the samples are selected from an ordered sampling frame. In systematic sampling, the list is actually in a circular manner and the selection starts from one end and reaches the final, and the cycle goes on. Equal probability method would be the best example for the systematic sampling.

37. What are recommender systems?

Answer: Recommender systems are also treated as information filtering systems that work to predict or likeness of a user for a product. These recommender systems are widely used in areas like news, movies, social tags, music, products, etc.

We can see the movie recommenders in Netflix, IMDB, & bookMyShow, and product recommender e-commerce sites like eBay, Amazon, Flipcart, Youtube video recommendations, and game recommendations.

38. What are Artificial Neural Networks?

Answer: Artificial neural networks are the main elements which have made the machine learning popular. These neural networks are developed based on the functionality of a human brain. The Artificial neural networks are trained to learn from the examples and experiences without being programmed explicitly. Artificial neural networks work based on nodes called artificial neurons that are connected to one another. Each connection acts similar to synapses in the human brain that helps in transmitting the signals between the artificial neurons.

39. Explain the role of Activation function?

Answer: The activation function helps in introducing the nonlinearity into the neural network that enables the neural network to learn the complex functions. Without this, it is challenging for the linear function to analyze complex data. An activation function is a function is an artificial neuron which delivers the output based on the input given.

40. What is the difference between Supervised Learning an Unsupervised Learning?

Answer: 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 of unsupervised learning.

41. What is the Central Limit Theorem and why is it important?
Answer: "Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible. While we can't obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample.
42. What are the feature vectors?
Answer: A feature vector is an n-dimensional vector of numerical features that represent some object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics, called features, of an object in a mathematical, easily analyzable way.
43. What is Cluster Sampling?
Answer: 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.
For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.
44. What are the various steps involved in an analytics project?
Answer: The following are the various steps involved in an analytics project:
Understand the Business problem
Explore the data and become familiar with it.
Prepare the data for modeling by detecting outliers, treating missing values, transforming variables, etc.
After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step until the best possible outcome is achieved.
Validate the model using a new data set.
Start implementing the model and track the result to analyze the performance of the model over the period of time.
45. Please explain Eigenvectors and Eigenvalues?
Answer: Eigenvectors help in understanding linear transformations. They are calculated typically for a correlation or covariance matrix in data analysis.
In other words, eigenvectors are those directions along which some particular linear transformation acts by compressing, flipping, or stretching.
Eigenvalues can be understood either as the strengths of the transformation in the direction of the eigenvectors or the factors by which the compressions happens.
46. What are outlier values and how do you treat them?
Answer: Outlier values, or simply outliers, are data points in statistics that don't belong to a certain population. An outlier value is an abnormal observation that is very much different from other values belonging to the set.
Identification of outlier values can be done by using univariate or some other graphical analysis method. Few outlier values can be assessed individually but assessing a large set of outlier values require the substitution of the same with either the 99th or the 1st percentile values.

There are two popular ways of treating outlier values:
To change the value so that it can be brought within a range
To simply remove the value
Note: Not all extreme values are outlier values.
47. How to choose the right chart in case of creating a viz?
Answer: Using the right chart to represent data is one of the key aspects of data visualization and design principle. You will always have options to choose from when deciding on a chart. But fixing to the right chart comes only by experience, practice and deep understanding of end-user needs. That dictates everything in the dashboard.
48. What is the basic responsibility of a Data Scientist?
Answer: As a data scientist, we have the responsibility to make complex things simple enough that anyone without context should understand, what we are trying to convey.
The moment, we start explaining even the simple things the mission of making the complex simple goes away. This happens a lot when we are doing data visualization.

Less is more. Rather than pushing too much information on to readers brain, we need to figure out how easily we can help them consume a dashboard or a chart.

The process is simple to say but difficult to implement. You must bring the complex business value out of a self-explanatory chart. It's a skill every data scientist should strive towards and good to have in their arsenal.

49. What is the difference between Machine learning Vs Data Mining?
Answer: Data mining is about working on unlimited data and then extract it to a level anywhere the unusual and unknown patterns are identified.
Machine learning is any method about a study whether it closely relates to design, development concerning the algorithms that provide an ability to certain computers to capacity to learn.
50. What are the types of biases that can occur during sampling?
Answer: Some simple models of selection bias are described below. Undercoverage occurs when some members of the population live badly represented inside the sample. … The survey relied on a service unit, drawn of telephone directories and car registration lists.

- Selection bias
- Under coverage bias
- Survivorship bias

51. Why data cleaning plays a vital role in the analysis?
Answer: Cleaning data from multiple sources to transform it into a format that data analysts or data scientists can work with is a cumbersome process because – as the number of data sources increases, the time take to clean the data increases exponentially due to the number of sources and the volume of data generated in these sources. It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.
52. What are an Eigenvalue and Eigenvector?
Answer: 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.
53. Define some key performance indicators for the product
Answer: After playing around with the product, think about this: what are some of the key metrics that the product might want to optimize? Part of a data scientist's role in certain companies involves working closely with the product teams to help define, measure, and report on these metrics. This is an exercise you can go through by yourself at home, and can really help during your interview process
54. Why is data cleaning important for analysis?
Answer: This is a knowledge-based question with a relatively simple answer. So much of a data scientist's time goes into cleaning data – and as the data gets bigger, so does the time it takes to clean. Cleaning it right is the foundation of analysis, and the time it takes to clean data, alone, makes it important.
55. Do you prefer Python or R for text analytics?
Answer: Here, you're being asked to insert your own opinion. However, most data scientists agree that the right opinion is Python. This is because Python has Pandas library which has strong data analysis tools and an easy-to-use structure. What's more, Python is typically faster for text analytics.
56. Explain Star Schema?
Answer: It is a traditional database schema with a central table. Satellite tables map ID's to physical name or description and can be connected to the central fact table using the ID fields; these tables are known as lookup tables, and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.
57. What do you mean by word Data Science?
Answer: Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured, which is a continuation of the field data mining and predictive analytics, It is also known as knowledge discovery and data mining.
58. What do you understand by term hash table collisions?
Answer: Hash table (hash map) is a kind of data structure used to implement an associative array, a structure that can map keys to values. Ideally, the hash function will assign each key to a unique

bucket, but sometimes it is possible that two keys will generate an identical hash causing both keys to point to the same bucket. It is known as hash collisions.

59. How can you assess a good logistic model?

Answer: There are various methods to assess the results of logistic regression analysis-
Using Classification Matrix to look at the true negatives and false positives.

Concordance that helps identify the ability of the logistic model to differentiate between the event happening and not happening.

Lift helps assess the logistic model by comparing it with random selection.

60. Why do you want to work as a data scientist?

Answer: This question plays off of your definition of data science. However, now recruiters are looking to understand what you'll contribute and what you'll gain from this field. Focus on what makes your path to becoming a data scientist unique – whether it be a mentor or a preferred method of data extraction.

61. How have you overcome a barrier to finding a solution?

Answer: This question directly asks you to draw upon your experiences and your ability to problem-solve. Data scientists are, after all, numbers-based problem-solvers, so, it's important to determine an example of a problem you've solved ahead of time. Whether it's through re-cleaning data or using a different program, you should be able to explain your process to the recruiter.

62. How To Work Towards A Random Forest?

Answer: The underlying principle of this technique is that several weak learners combined provide a strong learner. The steps involved are
Build several decision trees on bootstrapped training samples of data
On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates, out of all pp predictors
Rule of thumb: at each split m=p√m=p
Predictions: at the majority rule.

63. Explain Cross-validation?

Answer: It is a model validation technique for evaluating how the outcomes of a statistical analysis will generalize to an independent data set. Mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice.
The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting and get an insight on how the model will generalize to an independent data set.

64. What is a Linear Regression?

Answer: 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.

65. Can you explain the difference between a Test Set and a Validation Set?

Answer: 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, the test set is used for testing or evaluating the performance of a trained machine learning model.
In simple terms, the differences can be summarized as-

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

66. How do you define data science?

Answer: This question allows you to show your interviewer who you are. For example, what's your favorite part of the process, or what's the most impactful project you've worked on? Focus first on what data science is to everyone – a means of extracting insights from numbers – then explain what makes it personal.

67. What devices or tools help you most as a data scientist?
Answer: By asking this question, recruiters are seeking to learn more about your qualifications. Explain how you utilize every coding language you know, from R to SQL, and how each language helps complete certain tasks. This is also an opportunity to explain more about how your education or methods go above and beyond.
68. How often should an algorithm be updated?
Answer: This quasi-trick question has no specific time-based answer. This is because an algorithm should be updated whenever the underlying data is changing or when you want the model to evolve over time. Understanding the outcomes of dynamic algorithms is key to answering this question with confidence.
69. Python or R – Which one would you prefer for text analytics?
Answer: 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.
70. What is an Auto-Encoder?
Answer: The Auto-Encoders are learning networks that work for transforming the inputs into outputs with no errors or minimized error. It means the output must be very close to the input. We add a few layers between the input and output and the sizes of these layers would be smaller than the input layer. Actually, the Auto-encoder is provided with the unlabelled input then it would be transmitted into reconstructing the input.
71. What is back Propagation?
Answer: Backpropagation is an algorithm used in Deep Learning to train the multilayer neural network. Using this method, we can move an error form an end of a network to the inside of it, and that brings the efficient computation of gradient.
It consists of the below-mentioned steps:
Forward data propagation of data that is being used for training
Derivatives are computed with the help of output and target.
Backpropagation for computing the derivative error.
You can use the output that was previously calculated for output.
Update the weights.
72. How can the outlier values be treated?
Answer: We can identify the outlier values by using graphical analysis method or by using Univariate method. It becomes easier and can be assessed individually when the outlier values are few but when the outlier values are more in number then these values required to be substituted either with the 1st or with the 99th percentile values.
Below are the common ways to treat outlier values.

To bring down and change the value
To remove the value

73. Explain the difference between Univariate, Bivariate and Multivariate analysis?
Answer: Univariate analysis is a descriptive analysis and can be used to differentiate the number of variables involved at a given point of time. For instance, the sales of a particular territory include only one variable, and then the same is treated as a Univariate analysis.
Bivariate analysis is used to understand the difference between two variables at a given time on the scatter pilot. The best example for bivariate analysis of the difference between the sale and expenses happens for a particular product.
Multivariate analysis is used to understand the more than two variables responses for the variables.
74. What makes the difference between "Long" and "Wide" Format data?
Answer: In a wide format method, when we take a subject, the repeated responses are recorded in a single row, and each recorded response is in a separate column. When it comes to Long format data, each row acts as a one-time point per subject. In wide format, the columns are generally divided into groups whereas in a long-form the rows are divided into groups.
75. Do we have different Selection Biases, if yes, what are they?
Answer: Sampling Bias: This bias arises when you select only particular people or when non-random selection of samples happened. In general terms, it is nothing but a selection of the majority of the people belong to one group.
Time Interval: sometimes a trial may be terminated earlier than actual time (probably due to some ethical reasons) but the extreme value finally taken into consideration is the most significant value even though all other variables have similar Mean.

Data: We can name it as a Data bias when a separate set of data is taken to support a conclusion or eliminates terrible data based on the arbitrary grounds, instead of generally relying on generally stated criteria.
Attrition bias: Attrition bias is defined as an error that occurs due to Unequal loss of participants from a randomized controlled trial (RCT).

76. What is meant by supervised and unsupervised learning in data?

Answer: Supervised Learning: Supervised learning is a process of training machines with the labeled or right kind of data. In supervised learning, the machine uses the labeled data as a base to give the next answer.
Unsupervised learning: It is another form of training machines using information which is unlabeled or unstructured one. Unlike Supervised learning, there is no special teacher or predefined data for the machine to quickly learn from.

77. What is Data Science?

Answer: Data science is defined as a multidisciplinary subject used to extract meaningful insights out of different types of data by employing various scientific methods such as scientific processes and algorithms. Data science helps in solving the analytically complex problems in a simplified way. It acts as a stream where you can utilize raw data to generate business value.

78. What is Cross-validation?

Answer: It is a model validation technique used to evaluate how the statistical analysis would generalize to an independent dataset. This could be helpful in the areas of backgrounds where the objective is exactly forecasted, and the people want to estimate how accurately the model would work in real-time.
The main ambition of cross-validation is to test a model that is to test a model which is in the training phase and limit the problems like overfitting and to get insights on how to generalize the to an independent data set.

79. How can the outlier values be treated?

Answer: We can identify the outlier values by using graphical analysis method or by using Univariate method. It becomes easier and can be assessed individually when the outlier values are few but when the outlier values are more in number then these values required to be substituted either with the 1st or with the 99th percentile values.
Below are the common ways to treat outlier values.

To bring down and change the value
To remove the value

80. List the variants of backpropagation?

Answer: Below mentioned are the three different variants of backpropagation
Stochastic Gradient Descent: In this module, we take the help of the single training as an example for updating the parameters and for calculation of gradient.
Batch Gradient Descent: in this backpropagation method, we consider whole data to calculating the gradient and executes the update at each iteration.
Mini-batch Gradient Descent: It is considered as a popular optimization algorithm in deep learning. In this Mini-batch gradient Descent instead of single training example, mini-batch of samples is used.

81. What is a Boltzmann machine?

Answer: Boltzmann developed with simple learning algorithms that allow them to find the important information that was presented in the complex regularities in the data. These machines are generally used to optimize the quantity and the weights of the given problem. The learning program works very slow in networks due to many layers of feature detectors. When we consider Restricted Boltzmann Machines, this has a single algorithm feature detectors that make it faster compared to others.

82. Do gradient descent methods at all times converge to a similar point?

Answer: No, they do not because in some cases they reach a local minimum or a local optima point. You would not reach the global optima point. This is governed by the data and the starting conditions.

83. What are Eigenvalue and Eigenvector?

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

84. What is Selection Bias?
Answer: Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn't random. It is sometimes referred to as the selection effect. It is the distortion of statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.
The types of selection bias include:
Sampling bias: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
Time interval: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.
Data: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
Attrition: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.
85. How does data cleaning plays a vital role in the analysis?
Answer: Data cleaning can help in the analysis because:
Cleaning data from multiple sources helps to transform it into a format that data analysts or data scientists can work with.
Data Cleaning helps to increase the accuracy of the model in machine learning.
It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.
It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.
86. Can you explain the difference between a Validation Set and a Test Set?
Answer: A 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, a Test Set is used for testing or evaluating the performance of a trained machine learning model.
In simple terms, the differences can be summarized as; training set is to fit the parameters i.e. weights and test set is to assess the performance of the model i.e. evaluating the predictive power and generalization.
87. What do you mean by Deep Learning and Why has it become popular now?
Answer: Deep Learning is nothing but a paradigm of machine learning which has shown incredible promise in recent years. This is because of the fact that Deep Learning shows a great analogy with the functioning of the human brain.
Now although Deep Learning has been around for many years, the major breakthroughs from these techniques came just in recent years.

This is because of two main reasons:
The increase in the amount of data generated through various sources
The growth in hardware resources required to run these models
GPUs are multiple times faster and they help us build bigger and deeper deep learning models in comparatively less time than we required previously.

88. What are the variants of Back Propagation?
Answer: Stochastic Gradient Descent: We use only a single training example for calculation of gradient and update parameters.
Batch Gradient Descent: We calculate the gradient for the whole dataset and perform the update at each iteration.
Mini-batch Gradient Descent: It's one of the most popular optimization algorithms. It's a variant of Stochastic Gradient Descent and here instead of single training example, mini-batch of samples is used.
89. Please explain the role of data cleaning in data analysis.
Answer: Data cleaning can be a daunting task due to the fact that with the increase in the number of data sources, the time required for cleaning the data increases at an exponential rate.
This is due to the vast volume of data generated by additional sources. Also, data cleaning can solely take up to 80% of the total time required for carrying out a data analysis task.

Nevertheless, there are several reasons for using data cleaning in data analysis.

Two of the most important ones are:
Cleaning data from different sources helps in transforming the data into a format that is easy to work with
Data cleaning increases the accuracy of a machine learning model

90. What do you understand by linear regression and logistic regression?
Answer: Linear regression is a form of statistical technique in which the score of some variable Y is predicted on the basis of the score of a second variable X, referred to as the predictor variable. The Y variable is known as the criterion variable.
Also known as the logit model, logistic regression is a statistical technique for predicting the binary outcome from a linear combination of predictor variables.

91. What do you understand by Deep Learning?
Answer: Deep Learning is a paradigm of machine learning that displays a great degree of analogy with the functioning of the human brain. It is a neural network method based on convolutional neural networks (CNN).
Deep learning has a wide array of uses, ranging from social network filtering to medical image analysis and speech recognition. Although Deep Learning has been present for a long time, it's only recently that it has gained worldwide acclaim. This is mainly due to:
An increase in the amount of data generation via various sources
The growth in hardware resources required for running Deep Learning models
Caffe, Chainer, Keras, Microsoft Cognitive Toolkit, Pytorch, and TensorFlow are some of the most popular Deep Learning frameworks as of today.

92. What is overfitting?
Answer: Any prediction rate which has a high inconsistency between the training error and the test error leads ta a high business problem, if the error rate in training set is low and the error rate ithe n test set is high, then we can conclude it as overfitting model.
93. Advantages of Tableau Prep?
Answer: Tableau Prep will reduce a lot of time like how its parent software (Tableau) does when creating impressive visualizations. The tool has a lot of potentials in taking professionals from data cleaning, merging step to creating final usable data that can be linked to the Tableau desktop for getting visualization and business insights. A lot of manual tasks will be reduced and the time can be used to make better findings and insights.
94. How make you 3D plots/visualizations using NumPy/SciPy?
Answer: Like 2D plotting, 3D graphics is beyond the scope of NumPy and SciPy, but just as in this 2D example, packages exist that integrate with NumPy. Matplotlib provides primary 3D plotting in the mplot3d subpackage, whereas Mayavi produces a wide range of high-quality 3D visualization features, utilizing the powerful VTK engine.
95. Compare Sas, R, And Python Programming?
Answer:
SAS: it is one of the most widely used analytics tools used by some of the biggest companies on earth. It has some of the best statistical functions, graphical user interface, but can come with a price tag and hence it cannot be readily adopted by smaller enterprises
R: The best part about R is that it is an Open Source tool and hence used generously by academia and the research community. It is a robust tool for statistical computation, graphical representation, and reporting. Due to its open source nature, it is always being updated with the latest features and then readily available to everybody.
Python: Python is a powerful open source programming language that is easy to learn, works well with most other tools and technologies. The best part about Python is that it has innumerable libraries and community created modules making it very robust. It has functions for statistical operation,
96. Describe Univariate, Bivariate And Multivariate Analysis?
Answer: As the name suggests these are analysis methodologies having a single, double or multiple variables.

So a univariate analysis will have one variable and due to this, there are no relationships, causes. The major aspect of the univariate analysis is to summarize the data and find the patterns within it to make actionable decisions.

A Bivariate analysis deals with the relationship between two sets of data. These sets of paired data come from related sources, or samples. There are various tools to analyze such data including the chi-squared tests and t-tests when the data are having a correlation.

If the data can be quantified then it can be analyzed using a graph plot or a scatterplot. The strength of the correlation between the two data sets will be tested in a Bivariate analysis.

97. What Are Interpolation And Extrapolation?
Answer: The terms of interpolation and extrapolation are extremely important in any statistical analysis. Extrapolation is the determination or estimation using a known set of values or facts by extending it and taking it to an area or region that is unknown. It is the technique of inferring something using data that is available.
Interpolation, on the other hand, is the method of determining a certain value which falls between a certain set of values or the sequence of values.
This is especially useful when you have data at the two extremities of a certain region but you don't have enough data points at a specific point. This is when you deploy interpolation to determine the value that you need.

98. How Is Data Modeling Different From Database Design?
Answer: Data Modeling: It can be considered as the first step towards the design of a database. Data modeling creates a conceptual model based on the relationship between various data models. The process involves moving from the conceptual stage to the logical model to the physical schema. It involves the systematic method of applying data modeling techniques.
Database Design: This is the process of designing the database. The database design creates an output which is a detailed data model of the database. Strictly speaking, database design includes the detailed logical model of a database but it can also include physical design choices and storage parameters.
99. Differentiate between Data modeling and Database design?
Answer: Data Modeling – Data modeling (or modeling) in software engineering is the process of creating a data model for an information system by applying formal data modeling techniques.
Database Design– Database design is the system of producing a detailed data model of a database. The term database design can be used to describe many different parts of the design of an overall database system. (hadoop training)
100. What is selection bias and why does it matter?
Answer: Selection bias is a product of inadequately or improperly randomized data leading to data sets that are not representative of the whole. In an interview, you should express the importance of this in terms of its effect on your solution. If your data is not representative, your solutions likely are not either.
101. Differentiate between univariate, bivariate and multivariate analysis?
Answer: Univariate analyses 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 the analysis can be referred to as univariate analysis.
The bivariate analysis attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sale and spending can be considered as an example of bivariate analysis.

The multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.

102. Can you cite some examples where a false negative important than a false positive?
Answer: 1: 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 a shortage of staff, they decide to scan passengers 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?
2: What if the Jury or judge decides to make a criminal go free?
3: What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after a few years and realize that you had a false negative?

103. Describe the structure of Artificial Neural Networks?

Answer: Artificial Neural Networks works on the same principle as a biological Neural Network. It consists of inputs which get processed with weighted sums and Bias, with the help of Activation Functions. (oracle apex training online)

104. What do you understand by the Selection Bias? What are its various types?

Answer: Selection bias is typically associated with research that doesn't have a random selection of participants. It is a type of error that occurs when a researcher decides who is going to be studied. On some occasions, selection bias is also referred to as the selection effect.
In other words, selection bias is a distortion of statistical analysis that results from the sample collecting method. When selection bias is not taken into account, some conclusions made by a research study might not be accurate. Following are the various types of selection bias:

Sampling Bias: A systematic error resulting due to a non-random sample of a populace causing certain members of the same to be less likely included than others that results in a biased sample.
Time Interval – A trial might be ended at an extreme value, usually due to ethical reasons, but the extreme value is most likely to be reached by the variable with the most variance, even though all variables have a similar mean.
Data – Results when specific data subsets are selected for supporting a conclusion or rejection of bad data arbitrarily.
Attrition – Caused due to attrition, i.e. loss of participants, discounting trial subjects or tests that didn't run to completion.

105. Please explain Recommender Systems along with an application?

Answer: Recommender Systems is a subclass of information filtering systems, meant for predicting the preferences or ratings awarded by a user to some product.
An application of a recommender system is the product recommendations section on Amazon. This section contains items based on the user's search history and past orders.

106. Could you explain how to define the number of clusters in a clustering algorithm?

Answer: The primary objective of clustering is to group together similar identities in such a way that while entities within a group are similar to each other, the groups remain different from one another. Generally, Within Sum of Squares is used for explaining the homogeneity within a cluster. For defining the number of clusters in a clustering algorithm, WSS is plotted for a range pertaining to a number of clusters. The resultant graph is known as the Elbow Curve.
The Elbow Curve graph contains a point that represents the point post which there aren't any decrements in the WSS. This is known as the bending point and represents K in K–Means.
Although the aforementioned is the widely-used approach, another important approach is the Hierarchical clustering. In this approach, dendrograms are created first and then distinct groups are identified from there.

106. What are the types of machine learning?

Answer:

- Supervised learning
- Unsupervised learning
- Reinforcement Learning

107. What is a Random Forest?

Answer: Random forest is a versatile method in machine learning that performs both classification and regression tasks. It also helps in areas like treats missing values, dimensionality reduction, and outlier values. It is like gathering the various weak modules comes together to form a robust model

108. What is Reinforcement learning?

Answer: Reinforcement learning maps the situations to what to do and how to map actions. The end result of this Reinforcement learning is to maximize the numerical reward signal. The learner is not defined with what action to do next but instead must discover which actions will give the maximum reward. Reinforcement learning is developed from the learning process of human beings. It works based on the reward/penalty mechanism.

109. What does P-value signify about the statistical data?
Answer: 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.

Get hands-on experience for your interviews with free access to the solved code example.

110. What is an example of a data set with a non-Gaussian distribution?
Answer: The Gaussian distribution is part of the Exponential family of distributions, but there are a lot more of them, with the same sort of ease of use, in many cases, and if the person doing the machine learning has a solid grounding in statistics, they can be utilized where appropriate
111. How regularly must an algorithm be updated?
Answer:
You will want to update an algorithm when:
You want the model to evolve as data streams through infrastructure
The underlying data source is changing
There is a case of non-stationarity
Planning for Data Science Certification in R – Programming? Here're 100 Data Science Foundations Questions. Take this free practice test to know where you stand.
112. How has your prior experience prepared you for a role in data science?
Answer: This question helps determine the candidate's experience from a holistic perspective and reveals experience in demonstrating interpersonal, communication and technical skills. It is important to understand this because data scientists must be able to communicate their findings, work in a team environment and have the skills to perform the task.
Here are some possible answers to look for:
Project management skills
Examples of working in a team environment
Ability to identify errors
A substantial response may include the following: "My experience in my previous positions has prepared me for this job by giving me the skills I need to work in a group setting, manage projects and quickly identify errors.

113. What is Unsupervised learning?
Answer: Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.
Algorithms: Clustering, Anomaly Detection, Neural Networks, and Latent Variable Models
Data Science Mock interviews for you
Interviews by Industry ExpertsPersonalized detailed interview feedback access to exclusive and curated content
E.g. In the same example, a fruit clustering will categorize as "fruits with soft skin and lots of dimples", "fruits with shiny hard skin" and "elongated yellow fruits".

114. Could you draw a comparison between overfitting and underfitting?
Answer: In order to make reliable predictions on general untrained data in machine learning and statistics, it is required to fit a model to a set of training data. Overfitting and underfitting are two of the most common modeling errors that occur while doing so.
Following are the various differences between overfitting and underfitting:
Definition – A statistical model suffering from overfitting describes some random error or noise in place of the underlying relationship. When underfitting occurs, a statistical model or machine learning algorithm fails in capturing the underlying trend of the data.
Occurrence – When a statistical model or machine learning algorithm is excessively complex, it can result in overfitting. Example of a complex model is one having too many parameters when compared to the total number of observations. Underfitting occurs when trying to fit a linear model to non-linear data.

Poor Predictive Performance – Although both overfitting and underfitting yield poor predictive performance, the way in which each one of them does so is different. While the overfitted model overreacts to minor fluctuations in the training data, the underfit model under-reacts to even bigger fluctuations.

115. Can you compare the validation set with the test set?

Answer: A validation set is part of the training set used for parameter selection as well as for avoiding overfitting of the machine learning model being developed. On the contrary, a test set is meant for evaluating or testing the performance of a trained machine learning model.

116. Please explain the concept of a Boltzmann Machine.

Answer: A Boltzmann Machine features a simple learning algorithm that enables the same to discover fascinating features representing complex regularities present in the training data. It is basically used for optimizing the quantity and weight for some given problem.

The simple learning algorithm involved in a Boltzmann Machine is very slow in networks that have many layers of feature detectors.

117. What are the time series algorithms?

Answer: Time series algorithms like ARIMA, ARIMAX, SARIMA, Holts winters are very interesting to learn and use as well to solve a lot of complex problems for businesses. Data preparation for time series analysis plays a vital role. The stationarity, seasonality, cycles, and noises need time and attention. Take as much time as you would like to make the data right. Then you can run any model on top of it.

118. Now companies are heavily investing their money and time to make the dashboards. Why?

Answer: To make stakeholders more aware of the business through data. Working on visualization projects helps you develop one of the key skills every data scientist should possess i.e. Thinking from the shoes of the end-user.

If you're learning any visualization tool, download a dataset from kaggle. Building charts and graphs for the dashboard should be the last step. Research more about the domain and think about the KPIs you would like to see in the dashboard if you're going to be the end-user. Then start building the dashboard piece by piece.

119. Explain The Various Benefits Of R Language?

Answer: The R programming language includes a set of a software suite that is used for graphical representation, statistical computing, data manipulation, and calculation.

Some of the highlights of the R programming environment include the following:

An extensive collection of tools for data analysis
Operators for performing calculations on matrix and array
Data analysis technique for graphical representation
A highly developed yet simple and effective programming language
It extensively supports machine learning applications
It acts as a connecting link between various software, tools, and datasets
Create high-quality reproducible analysis that is flexible and powerful
Provides a robust package ecosystem for diverse needs
It is useful when you have to solve a data-oriented problem

120. Why Data Cleansing Is Important In Data Analysis?

Answer: With data coming in from multiple sources it is important to ensure that data is good enough for analysis. This is where data cleansing becomes extremely vital. Data cleansing extensively deals with the process of detecting and correcting data records, ensuring that data is complete and accurate and the components of data that are irrelevant are deleted or modified as per the needs. This process can be deployed in concurrence with data wrangling or batch processing.

Once the data is cleaned it confirms with the rules of the data sets in the system. Data cleansing is an essential part of the data science because the data can be prone to error due to human negligence, corruption during transmission or storage among other things. Data cleansing takes a huge chunk of time and effort of a Data Scientist because of the multiple sources from which data emanates and the speed at which it comes.