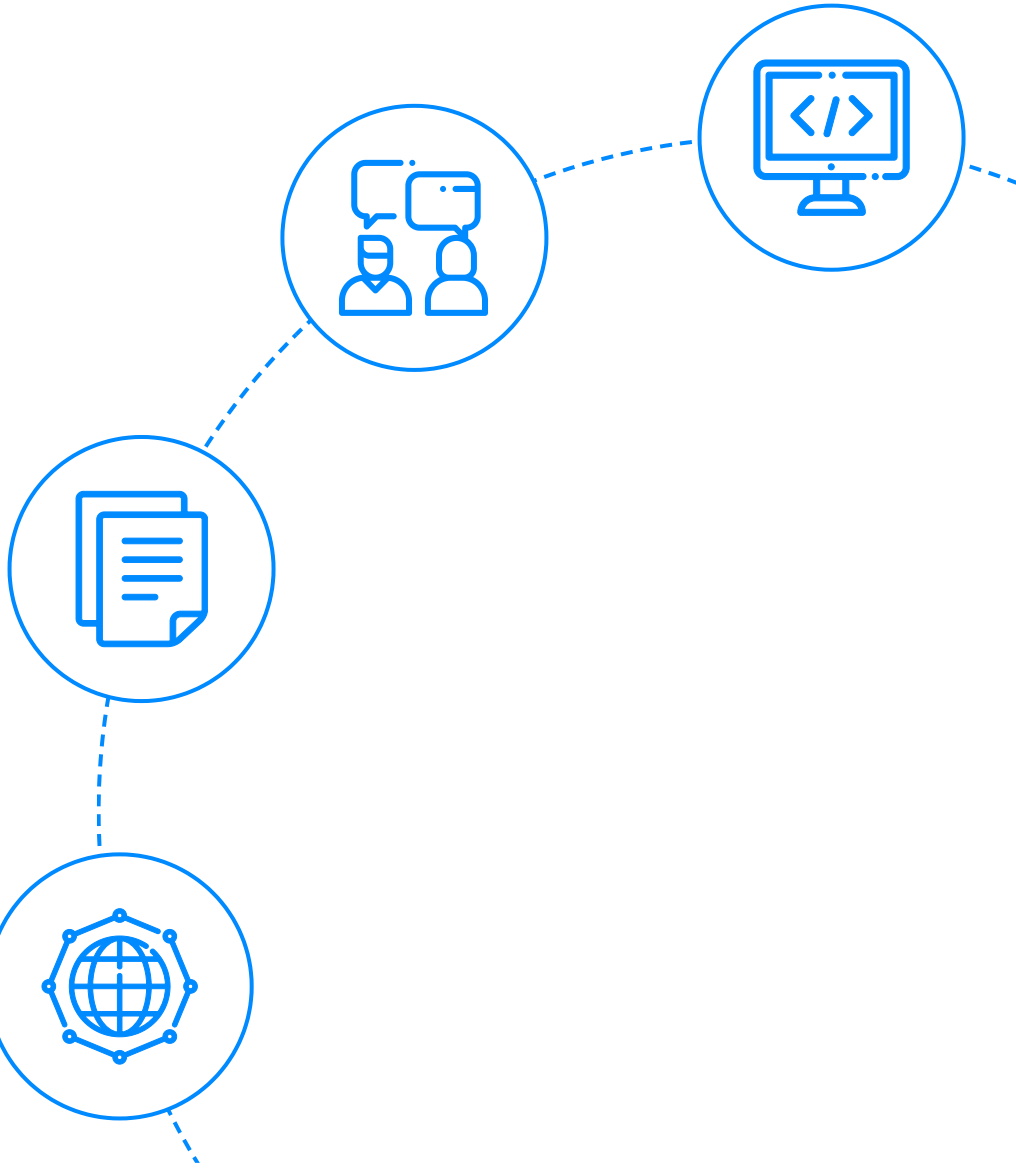
InterviewBit

Data Science Interview Questions



To view the live version of the page, click here.

Contents

Data Science Interview Questions for Freshers

Data Science Interview Questions for Experienced

Data Science Interview Questions for Experienced

(.....Continued)

19. Why do we need selection bias?
20. Why is data cleaning crucial? How do you clean the data?
21. What are the available feature selection methods for selecting the right variables for building efficient predictive models?
22. Will treating categorical variables as continuous variables result in a better predictive model?
23. How will you treat missing values during data analysis?
24. What does the ROC Curve represent and how to create it?
25. What are the differences between univariate, bivariate and multivariate analysis?
26. What is the difference between the Test set and validation set?
27. What do you understand by a kernel trick?
28. Differentiate between box plot and histogram.
29. How will you balance/correct imbalanced data?
30. What is better - random forest or multiple decision trees?
31. Consider a case where you know the probability of finding at least one shooting star in a 15-minute interval is 30%. Evaluate the probability of finding at least one shooting star in a one-hour duration?
32. Toss the selected coin 10 times from a jar of 1000 coins. Out of 1000 coins, 999 coins are fair and 1 coin is double-headed, assume that you see 10 heads. Estimate the probability of getting a head in the next coin toss.
33. What are some examples when false positive has proven important than false negative?
34. Give one example where both false positives and false negatives are important equally?
35. Is it good to do dimensionality reduction before fitting a Support Vector Model?
36. What are various assumptions used in linear regression? What would happen if they are violated?

Data Science Interview Questions for Experienced

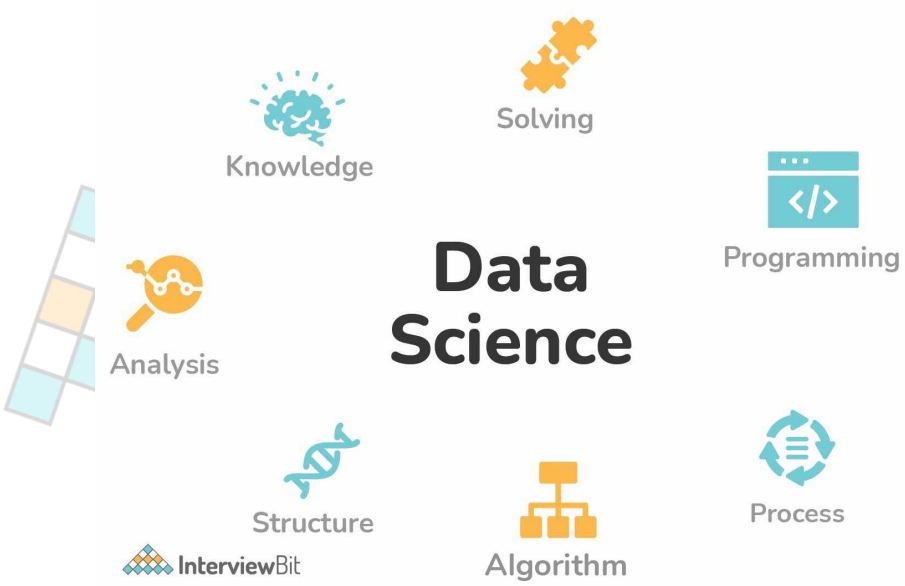
(.....Continued)

39. What is the importance of dimensionality reduction?
40. How is the grid search parameter different from the random search tuning strategy?

Let's get Started

Introduction:

Data science is an interdisciplinary field that mines raw data, analyses it, and comes up with patterns that are used to extract valuable insights from it. Statistics, computer science, machine learning, deep learning, data analysis, data visualization, and various other technologies form the core foundation of data science.



Over the years, data science has gained widespread importance due to the importance of data. Data is considered as the new oil of the future which when analyzed and harnessed properly can prove to be very beneficial to the stakeholders. Not just this, a data scientist gets the exposure to work in diverse domains, solving real-life practical problems all by making use of trendy technologies. The most common real-time application is fast delivery of food in apps such as Uber Eats by aiding the delivery person shows the fastest possible route to reach the destination from the restaurant. Data Science is also used in item recommendation systems in e-commerce sites like Amazon, Flipkart, etc which recommends the user what item they can buy based on their search history. Not just recommendation systems, Data Science is becoming increasingly popular in fraud detection applications to detect any fraud involved in credit-based financial applications. A successful data scientist can interpret data, perform innovation and bring out creativity while solving problems that help drive business and strategic goals. This makes it the most lucrative job of the 21st century. Learn More.

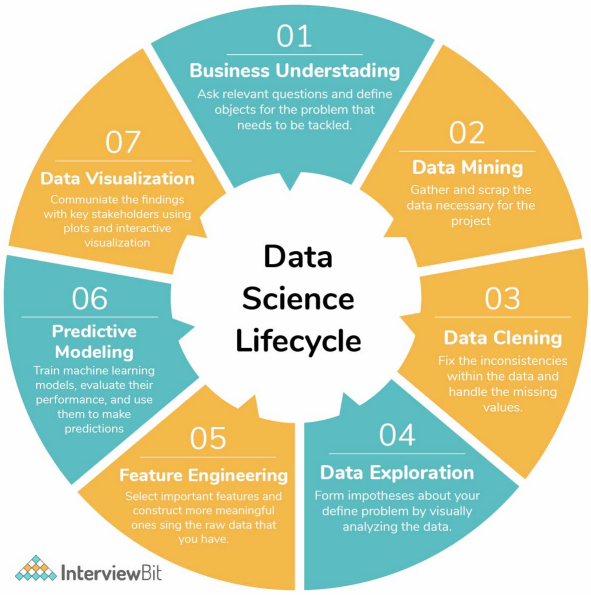
In this article, we will explore what are the most commonly asked interview questions which will help both aspiring and experienced data scientists.

Data Science Interview Questions for Freshers

1. What does one understand by the term Data Science?

An interdisciplinary field that constitutes various scientific processes, algorithms, tools, and machine learning techniques working to help find common patterns and gather sensible insights from the given raw input data using statistical and mathematical analysis is called Data Science.

The following figure represents the life cycle of data science.

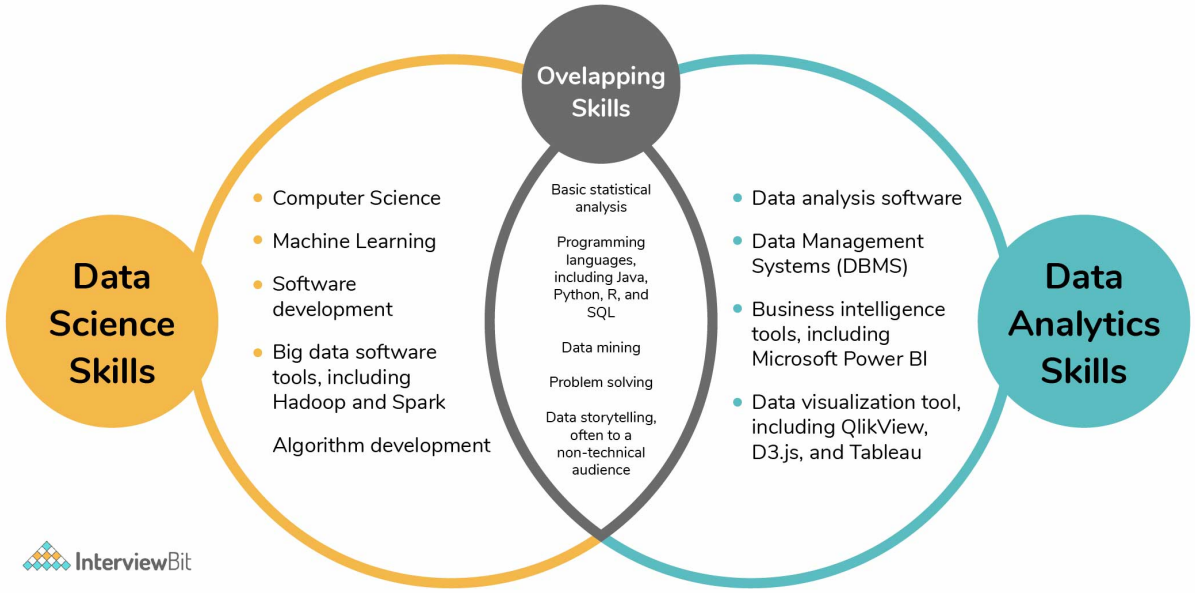


- It starts with gathering the business requirements and relevant data.
- Once the data is acquired, it is maintained by performing data cleaning, data warehousing, data staging, and data architecture.
- Data processing does the task of exploring the data, mining it, analyzing it which can be finally used to generate the summary of the insights extracted from the data.
- Once the exploratory steps are completed, the cleansed data is subjected to various algorithms like predictive analysis, regression, text mining, recognition patterns, etc depending on the requirements.
- In the final stage, the results are communicated to the business in a visually appealing manner. This is where the skill of data visualization, reporting, and different business intelligence tools come into the picture.

2. What is the difference between data analytics and data science?

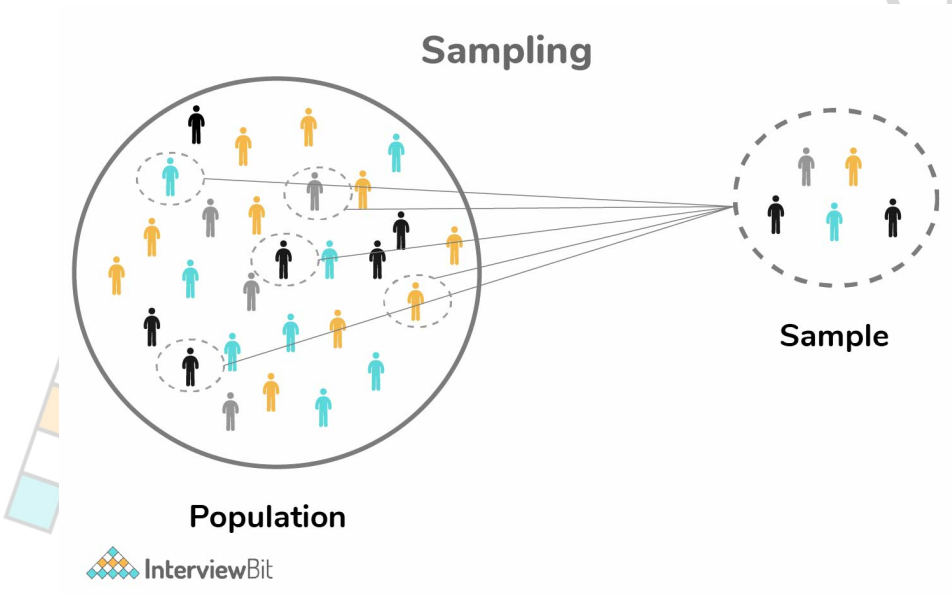
- Data science involves the task of transforming data by using various technical analysis methods to extract meaningful insights using which a data analyst can apply to their business scenarios.
- Data analytics deals with checking the existing hypothesis and information and answers questions for a better and effective business-related decision-making process.
- Data Science drives innovation by answering questions that build connections and answers for futuristic problems. Data analytics focuses on getting present meaning from existing historical context whereas data science focuses on predictive modeling.
- Data Science can be considered as a broad subject that makes use of various mathematical and scientific tools and algorithms for solving complex problems whereas data analytics can be considered as a specific field dealing with specific concentrated problems using fewer tools of statistics and visualization.

The following Venn diagram depicts the difference between data science and analytics clearly:



3. What are some of the techniques used for sampling? What is the main advantage of sampling?

Data analysis can not be done on a whole volume of data at a time especially when it involves larger datasets. It becomes crucial to take some data samples that can be used for representing the whole population and then perform analysis on it. While doing this, it is very much necessary to carefully take sample data out of the huge data that truly represents the entire dataset.

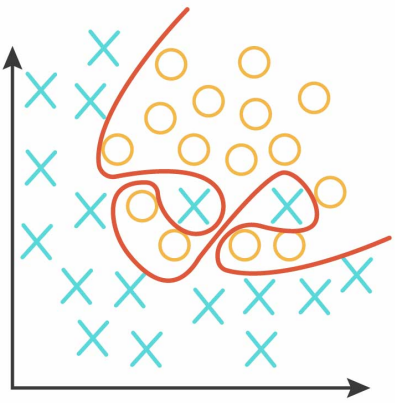


There are majorly two categories of sampling techniques based on the usage of statistics, they are:

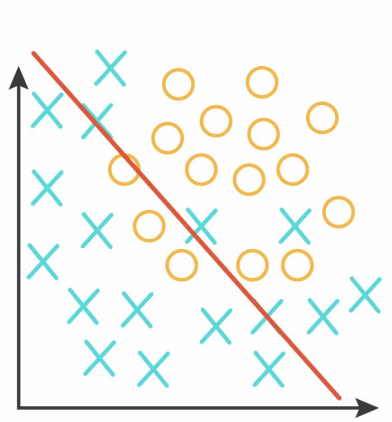
- **Probability Sampling techniques:** Clustered sampling, Simple random sampling, Stratified sampling.
- **Non-Probability Sampling techniques:** Quota sampling, Convenience sampling, snowball sampling, etc.

4. List down the conditions for Overfitting and Underfitting.

Overfitting: The model performs well only for the sample training data. If any new data is given as input to the model, it fails to provide any result. These conditions occur due to low bias and high variance in the model. Decision trees are more prone to overfitting.



Underfitting: Here, the model is so simple that it is not able to identify the correct relationship in the data, and hence it does not perform well even on the test data. This can happen due to high bias and low variance. Linear regression is more prone to Underfitting.



5. Differentiate between the long and wide format data.

Long format Data	Wide-Format Data
Here, each row of the data represents the one-time information of a subject. Each subject would have its data in different/ multiple rows.	Here, the repeated responses of a subject are part of separate columns.
The data can be recognized by considering rows as groups.	The data can be recognized by considering columns as groups.
This data format is most commonly used in R analyses and to write into log files after each trial.	This data format is rarely used in R analyses and most commonly used in stats packages for repeated measures ANOVAs.

The following image depicts the representation of wide format and long format data:

Name	Height	Weight
John	160	67
Christopher	182	78

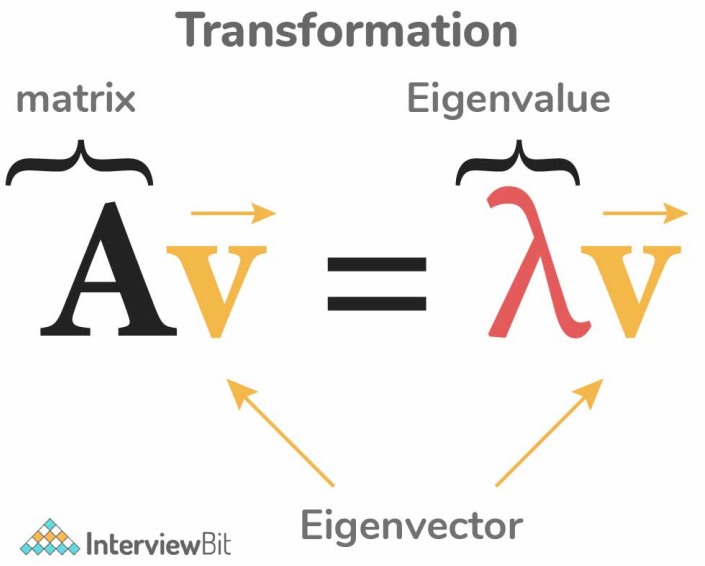
Figure: Wide Format

Name	Attribute	Value
John	Height	160
John	Weight	67
Christopher	Height	182
Christopher	Weight	78

Figure: Long Format

6. What are Eigenvectors and Eigenvalues?

Eigenvectors are column vectors or unit vectors whose length/magnitude is equal to 1. They are also called right vectors. Eigenvalues are coefficients that are applied on eigenvectors which give these vectors different values for length or magnitude.



A matrix can be decomposed into Eigenvectors and Eigenvalues and this process is called Eigen decomposition. These are then eventually used in machine learning methods like PCA (Principal Component Analysis) for gathering valuable insights from the given matrix.

7. What does it mean when the p-values are high and low?

A p-value is the measure of the probability of having results equal to or more than the results achieved under a specific hypothesis assuming that the null hypothesis is correct. This represents the probability that the observed difference occurred randomly by chance.

- Low p-value which means values ≤ 0.05 means that the null hypothesis can be rejected and the data is unlikely with true null.
- High p-value, i.e values ≥ 0.05 indicates the strength in favor of the null hypothesis. It means that the data is like with true null.
- p-value = 0.05 means that the hypothesis can go either way.

8. When is resampling done?

Resampling is a methodology used to sample data for improving accuracy and quantify the uncertainty of population parameters. It is done to ensure the model is good enough by training the model on different patterns of a dataset to ensure variations are handled. It is also done in the cases where models need to be validated using random subsets or when substituting labels on data points while performing tests.

9. What do you understand by Imbalanced Data?

Data is said to be highly imbalanced if it is distributed unequally across different categories. These datasets result in an error in model performance and result in inaccuracy.

10. Are there any differences between the expected value and mean value?

There are not many differences between these two, but it is to be noted that these are used in different contexts. The mean value generally refers to the probability distribution whereas the expected value is referred to in the contexts involving random variables.

11. What do you understand by Survivorship Bias?

This bias refers to the logical error while focusing on aspects that survived some process and overlooking those that did not work due to lack of prominence. This bias can lead to deriving wrong conclusions.

12. Define the terms KPI, lift, model fitting, robustness and DOE.

- **KPI:** KPI stands for Key Performance Indicator that measures how well the business achieves its objectives.
- **Lift:** This is a performance measure of the target model measured against a random choice model. Lift indicates how good the model is at prediction versus if there was no model.
- **Model fitting:** This indicates how well the model under consideration fits given observations.
- **Robustness:** This represents the system's capability to handle differences and variances effectively.
- **DOE:** stands for the design of experiments, which represents the task design aiming to describe and explain information variation under hypothesized conditions to reflect variables.

13. Define confounding variables.

Confounding variables are also known as confounders. These variables are a type of extraneous variables that influence both independent and dependent variables causing spurious association and mathematical relationships between those variables that are associated but are not casually related to each other.

Data Science Interview Questions for Experienced

14. How are the time series problems different from other regression problems?

- Time series data can be thought of as an extension to linear regression which uses terms like autocorrelation, movement of averages for summarizing historical data of y-axis variables for predicting a better future.
- Forecasting and prediction is the main goal of time series problems where accurate predictions can be made but sometimes the underlying reasons might not be known.
- Having Time in the problem does not necessarily mean it becomes a time series problem. There should be a relationship between target and time for a problem to become a time series problem.
- The observations close to one another in time are expected to be similar to the ones far away which provide accountability for seasonality. For instance, today's weather would be similar to tomorrow's weather but not similar to weather from 4 months from today. Hence, weather prediction based on past data becomes a time series problem.

15. Suppose there is a dataset having variables with missing values of more than 30%, how will you deal with such a dataset?

Depending on the size of the dataset, we follow the below ways:

- In case the datasets are small, the missing values are substituted with the mean or average of the remaining data. In pandas, this can be done by using `mean = df.mean()` where df represents the pandas dataframe representing the dataset and mean() calculates the mean of the data. To substitute the missing values with the calculated mean, we can use `df.fillna(mean)` .
- For larger datasets, the rows with missing values can be removed and the remaining data can be used for data prediction.

16. What is Cross-Validation?

Cross-Validation is a Statistical technique used for improving a model's performance. Here, the model will be trained and tested with rotation using different samples of the training dataset to ensure that the model performs well for unknown data. The training data will be split into various groups and the model is run and validated against these groups in rotation.

The most commonly used techniques are:

- K- Fold method
- Leave p-out method
- Leave-one-out method
- Holdout method

17. What are the differences between correlation and covariance?

Although these two terms are used for establishing a relationship and dependency between any two random variables, the following are the differences between them:

- **Correlation:** This technique is used to measure and estimate the quantitative relationship between two variables and is measured in terms of how strong are the variables related.
- **Covariance:** It represents the extent to which the variables change together in a cycle. This explains the systematic relationship between pair of variables where changes in one affect changes in another variable.

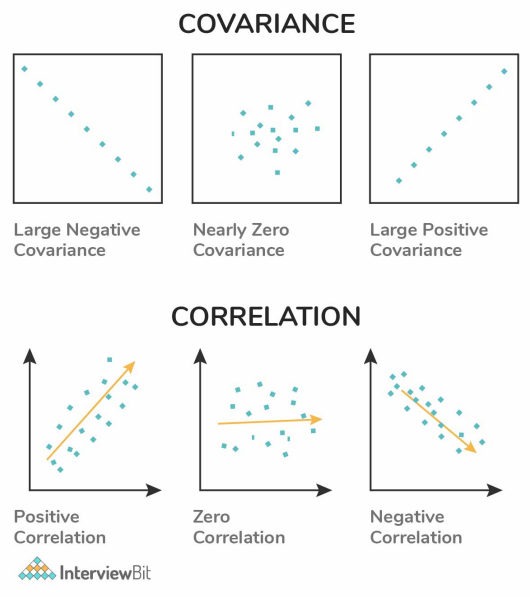
Mathematically, consider 2 random variables, X and Y where the means are represented as μX and μY respectively and standard deviations are represented by σX and σY respectively and E represents the expected value operator, then:

- covarianceXY = E[(X-μX),(Y-μY)]
- correlationXY = E[(X-μX),(Y-μY)]/(σXσY)
  so that

```
correlation(X,Y) = covariance(X,Y)/(covariance(X) covariance(Y))
```

Based on the above formula, we can deduce that the correlation is dimensionless whereas covariance is represented in units that are obtained from the multiplication of units of two variables.

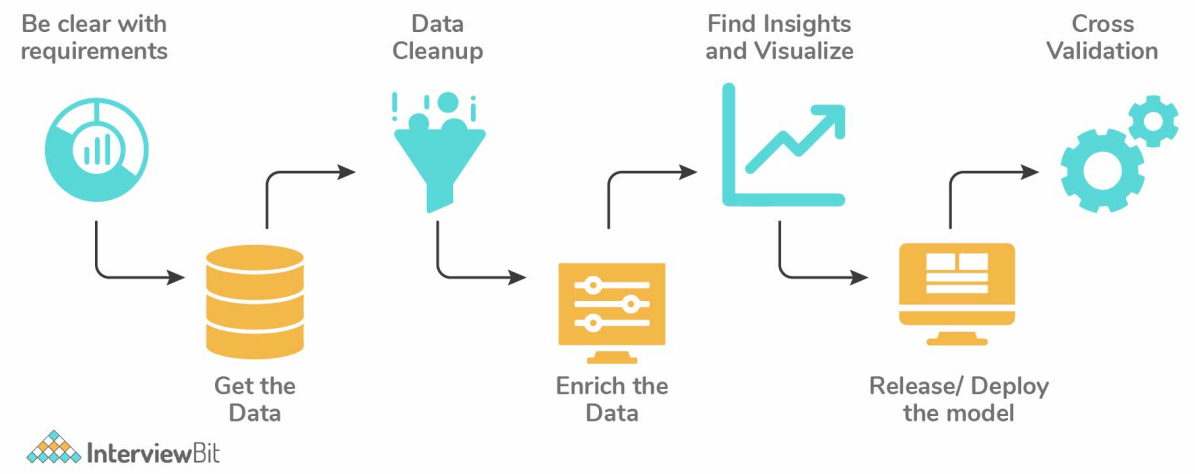
The following image graphically shows the difference between correlation and covariance:



18. How do you approach solving any data analytics based project?

Generally, we follow the below steps:

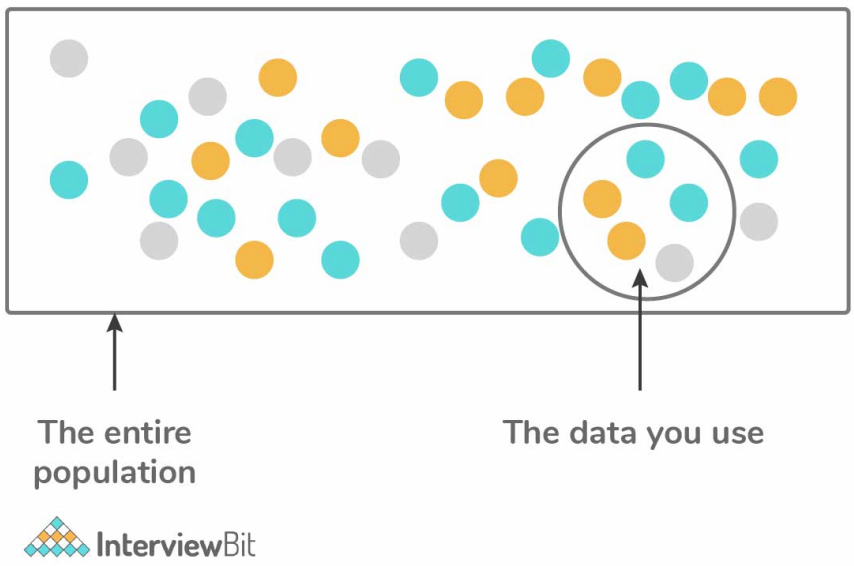
- First step is to thoroughly understand the business requirement/problem
- Next, explore the given data and analyze it carefully. If you find any data missing, get the requirements clarified from the business.
- Data cleanup and preparation step is to be performed next which is then used for modeling. Here, the missing values are found and the variables are transformed.
- Run your model against the data, build meaningful visualization and analyze the results to get meaningful insights.
- Release the model implementation, track the results and performance over a specified period to analyze the usefulness.
- Perform cross-validation of the model.



19. Why do we need selection bias?

Selection Bias happens in cases where there is no randomization specifically achieved while picking a part of the dataset for analysis. This bias tells that the sample analyzed does not represent the whole population meant to be analyzed.

- For example, in the below image, we can see that the sample that we selected does not entirely represent the whole population that we have. This helps us to question whether we have selected the right data for analysis or not.



20. Why is data cleaning crucial? How do you clean the data?

While running an algorithm on any data, to gather proper insights, it is very much necessary to have correct and clean data that contains only relevant information. Dirty data most often results in poor or incorrect insights and predictions which can have damaging effects.

For example, while launching any big campaign to market a product, if our data analysis tells us to target a product that in reality has no demand and if the campaign is launched, it is bound to fail. This results in a loss of the company's revenue. This is where the importance of having proper and clean data comes into the picture.

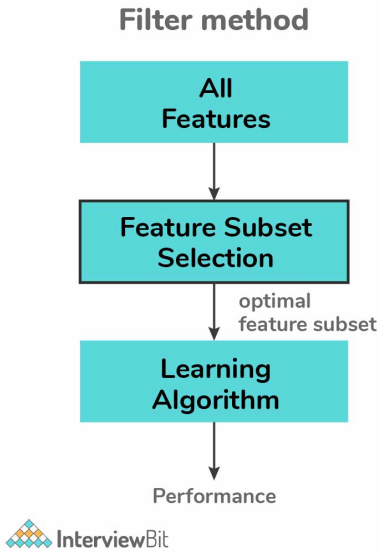
- Data Cleaning of the data coming from different sources helps in data transformation and results in the data where the data scientists can work on.
- Properly cleaned data increases the accuracy of the model and provides very good predictions.
- If the dataset is very large, then it becomes cumbersome to run data on it. The data cleanup step takes a lot of time (around 80% of the time) if the data is huge. It cannot be incorporated with running the model. Hence, cleaning data before running the model, results in increased speed and efficiency of the model.
- Data cleaning helps to identify and fix any structural issues in the data. It also helps in removing any duplicates and helps to maintain the consistency of the data.

The following diagram represents the advantages of data cleaning:

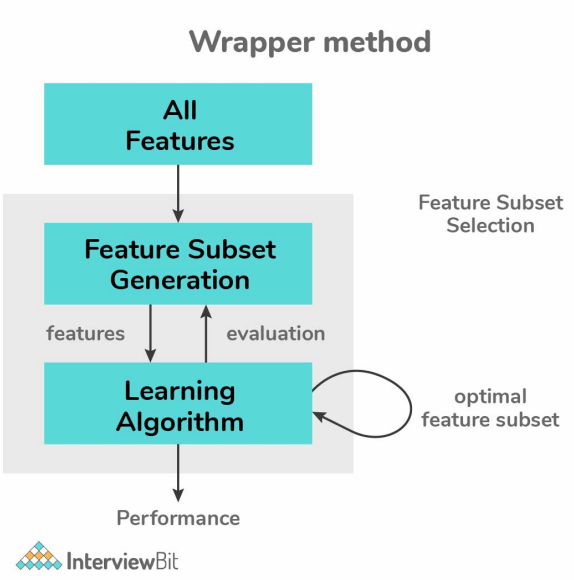
21. What are the available feature selection methods for selecting the right variables for building efficient predictive models?

While using a dataset in data science or machine learning algorithms, it so happens that not all the variables are necessary and useful to build a model. Smarter feature selection methods are required to avoid redundant models to increase the efficiency of our model. Following are the three main methods in feature selection:

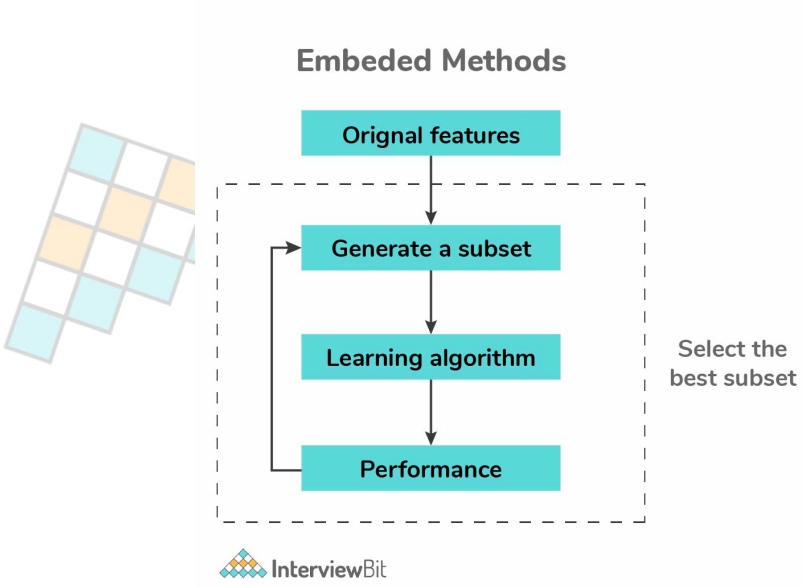
- **Filter Methods:**
  - These methods pick up only the intrinsic properties of features that are measured via univariate statistics and not cross-validated performance. They are straightforward and are generally faster and require less computational resources when compared to wrapper methods.
  - There are various filter methods such as the Chi-Square test, Fisher's Score method, Correlation Coefficient, Variance Threshold, Mean Absolute Difference (MAD) method, Dispersion Ratios, etc.



- **Wrapper Methods:**
  - These methods need some sort of method to search greedily on all possible feature subsets, access their quality by learning and evaluating a classifier with the feature.
  - The selection technique is built upon the machine learning algorithm on which the given dataset needs to fit.
  - There are three types of wrapper methods, they are:
    - **Forward Selection:** Here, one feature is tested at a time and new features are added until a good fit is obtained.
    - **Backward Selection:** Here, all the features are tested and the non-fitting ones are eliminated one by one to see while checking which works better.
    - **Recursive Feature Elimination:** The features are recursively checked and evaluated how well they perform.
  - These methods are generally computationally intensive and require high-end resources for analysis. But these methods usually lead to better predictive models having higher accuracy than filter methods.



- **Embedded Methods:**
  - Embedded methods constitute the advantages of both filter and wrapper methods by including feature interactions while maintaining reasonable computational costs.
  - These methods are iterative as they take each model iteration and carefully extract features contributing to most of the training in that iteration.
  - Examples of embedded methods: LASSO Regularization (L1), Random Forest Importance.



22. Will treating categorical variables as continuous variables result in a better predictive model?

Yes! A categorical variable is a variable that can be assigned to two or more categories with no definite category ordering. Ordinal variables are similar to categorical variables with proper and clear ordering defines. So, if the variable is ordinal, then treating the categorical value as a continuous variable will result in better predictive models.

23. How will you treat missing values during data analysis?

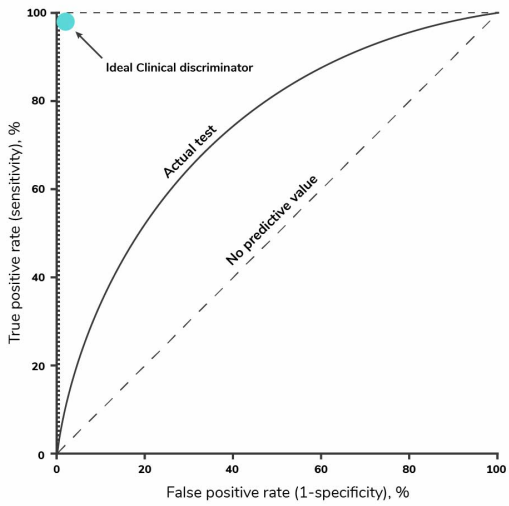
The impact of missing values can be known after identifying what kind of variables have the missing values.

- If the data analyst finds any pattern in these missing values, then there are chances of finding meaningful insights.
- In case of patterns are not found, then these missing values can either be ignored or can be replaced with default values such as mean, minimum, maximum, or median values.
- If the missing values belong to categorical variables, then they are assigned with default values. If the data has a normal distribution, then mean values are assigned to missing values.
- If 80% values are missing, then it depends on the analyst to either replace them with default values or drop the variables.

24. What does the ROC Curve represent and how to create it?

ROC (Receiver Operating Characteristic) curve is a graphical representation of the contrast between false-positive rates and true positive rates at different thresholds. The curve is used as a proxy for a trade-off between sensitivity and specificity.

The ROC curve is created by plotting values of true positive rates (TPR or sensitivity) against false-positive rates (FPR or (1-specificity)) TPR represents the proportion of observations correctly predicted as positive out of overall positive observations. The FPR represents the proportion of observations incorrectly predicted out of overall negative observations. Consider the example of medical testing, the TPR represents the rate at which people are correctly tested positive for a particular disease.



25. What are the differences between univariate, bivariate and multivariate analysis?

Statistical analyses are classified based on the number of variables processed at a given time.

Univariate analysis	Bivariate analysis	Multivariate analysis
This analysis deals with solving only one variable at a time.	This analysis deals with the statistical study of two variables at a given time.	This analysis deals with statistical analysis of more than two variables and studies the responses.
Example: Sales pie charts based on territory.	Example: Scatterplot of Sales and spend volume analysis study.	Example: Study of the relationship between human's social media habits and their self-esteem which depends on multiple factors like age, number of hours spent, employment status, relationship status, etc.

26. What is the difference between the Test set and validation set?

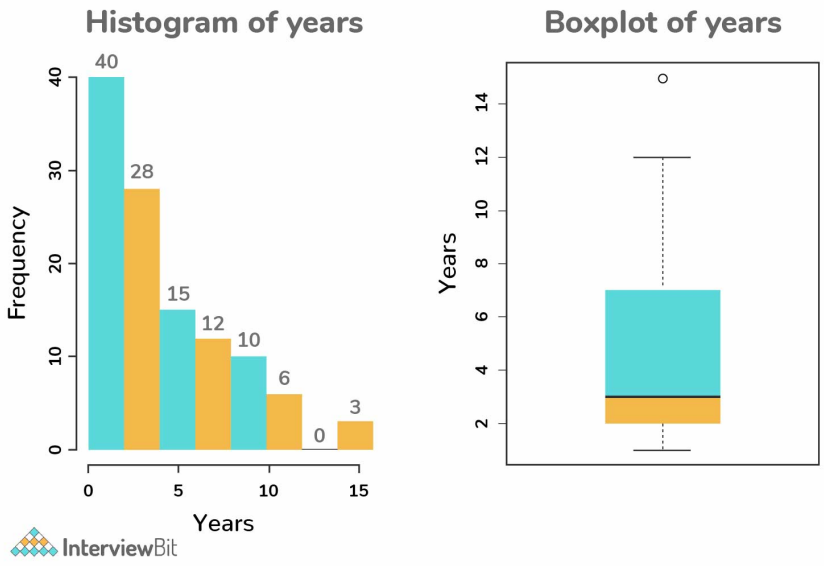
The test set is used to test or evaluate the performance of the trained model. It evaluates the predictive power of the model.
The validation set is part of the training set that is used to select parameters for avoiding model overfitting.

27. What do you understand by a kernel trick?

Kernel functions are generalized dot product functions used for the computing dot product of vectors xx and yy in high dimensional feature space. Kernal trick method is used for solving a non-linear problem by using a linear classifier by transforming linearly inseparable data into separable ones in higher dimensions.

28. Differentiate between box plot and histogram.

Box plots and histograms are both visualizations used for showing data distributions for efficient communication of information.
Histograms are the bar chart representation of information that represents the frequency of numerical variable values that are useful in estimating probability distribution, variations and outliers.
Boxplots are used for communicating different aspects of data distribution where the shape of the distribution is not seen but still the insights can be gathered. These are useful for comparing multiple charts at the same time as they take less space when compared to histograms.



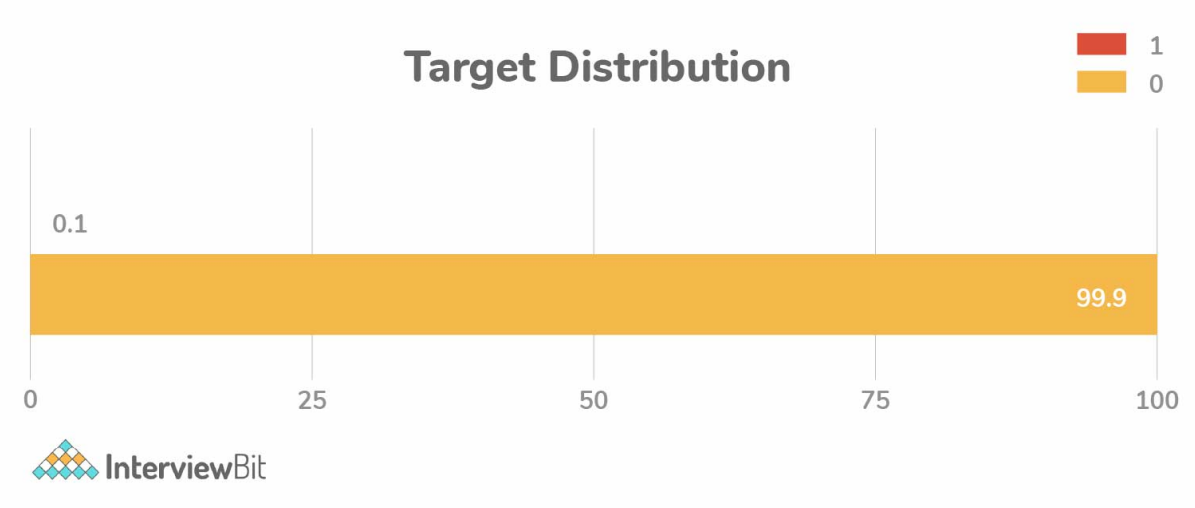
29. How will you balance/correct imbalanced data?

There are different techniques to correct/balance imbalanced data. It can be done by increasing the sample numbers for minority classes. The number of samples can be decreased for those classes with extremely high data points. Following are some approaches followed to balance data:

- **Use the right evaluation metrics:** In cases of imbalanced data, it is very important to use the right evaluation metrics that provide valuable information.
    - **Specificity/Precision:** Indicates the number of selected instances that are relevant.
    - **Sensitivity:** Indicates the number of relevant instances that are selected.
    - **F1 score:** It represents the harmonic mean of precision and sensitivity.
    - **MCC (Matthews correlation coefficient):** It represents the correlation coefficient between observed and predicted binary classifications.
    - **AUC (Area Under the Curve):** This represents a relation between the true positive rates and false-positive rates.

For example, consider the below graph that illustrates training data:

Here, if we measure the accuracy of the model in terms of getting "0"s, then the accuracy of the model would be very high -> 99.9%, but the model does not guarantee any valuable information. In such cases, we can apply different evaluation metrics as stated above.



- **Training Set Resampling:** It is also possible to balance data by working on getting different datasets and this can be achieved by resampling. There are two approaches followed under-sampling that is used based on the use case and the requirements:
    - **Under-sampling** This balances the data by reducing the size of the abundant class and is used when the data quantity is sufficient. By performing this, a new dataset that is balanced can be retrieved and this can be used for further modeling.
    - **Over-sampling** This is used when data quantity is not sufficient. This method balances the dataset by trying to increase the samples size. Instead of getting rid of extra samples, new samples are generated and introduced by employing the methods of repetition, bootstrapping, etc.
- **Perform K-fold cross-validation correctly:** Cross-Validation needs to be applied properly while using over-sampling. The cross-validation should be done before over-sampling because if it is done later, then it would be like overfitting the model to get a specific result. To avoid this, resampling of data is done repeatedly with different ratios.

30. What is better - random forest or multiple decision trees?

Random forest is better than multiple decision trees as random forests are much more robust, accurate, and lesser prone to overfitting as it is an ensemble method that ensures multiple weak decision trees learn strongly.

31. Consider a case where you know the probability of finding at least one shooting star in a 15-minute interval is 30%. Evaluate the probability of finding at least one shooting star in a one-hour duration?

```
We know that,
Probability of finding atleast 1 shooting star in 15 min = P(sighting in 15min) = 30% =
Hence, Probability of not sighting any
shooting star in 15 min = 1-P(sighting in 15min)
                        = 1-0.3
                        = 0.7

Probability of not finding shooting star in 1 hour
                        = 0.7^4
                        = 0.1372
Probability of finding atleast 1
shooting star in 1 hour = 1-0.1372
                        = 0.8628
```

So the probability is 0.8628 = 86.28%

32. Toss the selected coin 10 times from a jar of 1000 coins. Out of 1000 coins, 999 coins are fair and 1 coin is double-headed, assume that you see 10 heads. Estimate the probability of getting a head in the next coin toss.

We know that there are two types of coins - fair and double-headed. Hence, there are two possible ways of choosing a coin. The first is to choose a fair coin and the second is to choose a coin having 2 heads.

P(selecting fair coin) = 999/1000 = 0.999
P(selecting double headed coin) = 1/1000 = 0.001

Using Bayes rule,

```
P(selecting 10 heads in row) = P(selecting fair coin)* Getting 10 heads + P(selecting c
P(selecting 10 heads in row) = P(A)+P(B)

P (A)  =  0.999 * (1/2)^10
      =  0.999 * (1/1024)
      =  0.000976
P (B)  =  0.001 * 1  =  0.001
P( A / (A + B) )  = 0.000976 /  (0.000976 + 0.001)                    = 0.4939
P( B / (A + B))   = 0.001 / 0.001976
                  = 0.5061
P(selecting head in next toss) = P(A/A+B) * 0.5 + P(B/A+B) * 1
                               = 0.4939 * 0.5 + 0.5061
                               = 0.7531
```

So, the answer is 0.7531 or 75.3%.

33. What are some examples when false positive has proven important than false negative?

Before citing instances, let us understand what are false positives and false negatives.

- False Positives are those cases that were wrongly identified as an event even if they were not. They are called Type I errors.
- False Negatives are those cases that were wrongly identified as non-events despite being an event. They are called Type II errors.

Some examples where false positives were important than false negatives are:

- In the medical field: Consider that a lab report has predicted cancer to a patient even if he did not have cancer. This is an example of a false positive error. It is dangerous to start chemotherapy for that patient as he doesn't have cancer as starting chemotherapy would lead to damage of healthy cells and might even actually lead to cancer.
- In the e-commerce field: Suppose a company decides to start a campaign where they give $100 gift vouchers for purchasing $10000 worth of items without any minimum purchase conditions. They assume it would result in at least 20% profit for items sold above $10000. What if the vouchers are given to the customers who haven't purchased anything but have been mistakenly marked as those who purchased $10000 worth of products. This is the case of false-positive error.

34. Give one example where both false positives and false negatives are important equally?

In Banking fields: Lending loans are the main sources of income to the banks. But if the repayment rate isn't good, then there is a risk of huge losses instead of any profits. So giving out loans to customers is a gamble as banks can't risk losing good customers but at the same time, they can't afford to acquire bad customers. This case is a classic example of equal importance in false positive and false negative scenarios.

35. Is it good to do dimensionality reduction before fitting a Support Vector Model?

If the features number is greater than observations then doing dimensionality reduction improves the SVM (Support Vector Model).

36. What are various assumptions used in linear regression? What would happen if they are violated?

Linear regression is done under the following assumptions:

- The sample data used for modeling represents the entire population.
- There exists a linear relationship between the X-axis variable and the mean of the Y variable.
- The residual variance is the same for any X values. This is called homoscedasticity
- The observations are independent of one another.
- Y is distributed normally for any value of X.

Extreme violations of the above assumptions lead to redundant results. Smaller violations of these result in greater variance or bias of the estimates.

37. How is feature selection performed using the regularization method?

The method of regularization entails the addition of penalties to different parameters in the machine learning model for reducing the freedom of the model to avoid the issue of overfitting.
There are various regularization methods available such as linear model regularization, Lasso/L1 regularization, etc. The linear model regularization applies penalty over coefficients that multiplies the predictors. The Lasso/L1 regularization has the feature of shrinking some coefficients to zero, thereby making it eligible to be removed from the model.

38. How do you identify if a coin is biased?

To identify this, we perform a hypothesis test as below:
According to the null hypothesis, the coin is unbiased if the probability of head flipping is 50%. According to the alternative hypothesis, the coin is biased and the probability is not equal to 500. Perform the below steps:

- Flip coin 500 times
- Calculate p-value.
- Compare the p-value against the alpha -> result of two-tailed test (0.05/2 = 0.025). Following two cases might occur:
  - **p-value > alpha:** Then null hypothesis holds good and the coin is unbiased.
  - **p-value < alpha:** Then the null hypothesis is rejected and the coin is biased.

39. What is the importance of dimensionality reduction?

The process of dimensionality reduction constitutes reducing the number of features in a dataset to avoid overfitting and reduce the variance. There are mostly 4 advantages of this process:

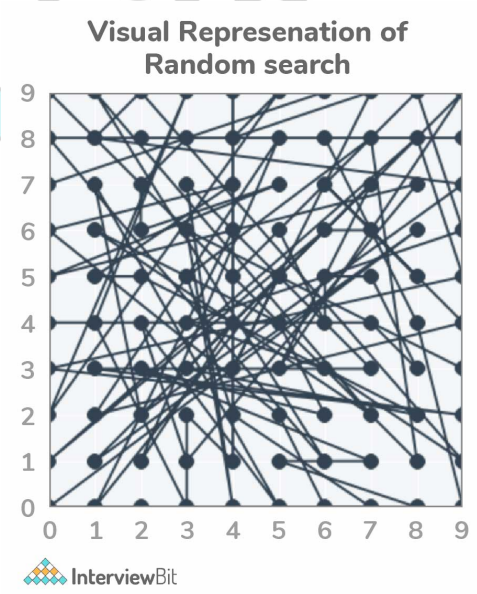
- This reduces the storage space and time for model execution.
- Removes the issue of multi-collinearity thereby improving the parameter interpretation of the ML model.
- Makes it easier for visualizing data when the dimensions are reduced.
- Avoids the curse of increased dimensionality.

40. How is the grid search parameter different from the random search tuning strategy?

Tuning strategies are used to find the right set of hyperparameters. Hyperparameters are those properties that are fixed and model-specific before the model is tested or trained on the dataset. Both the grid search and random search tuning strategies are optimization techniques to find efficient hyperparameters.

- **Grid Search**:
  - Here, every combination of a preset list of hyperparameters is tried out and evaluated.
  - The search pattern is similar to searching in a grid where the values are in a matrix and a search is performed. Each parameter set is tried out and their accuracy is tracked. after every combination is tried out, the model with the highest accuracy is chosen as the best one.
  - The main drawback here is that, if the number of hyperparameters is increased, the technique suffers. The number of evaluations can increase exponentially with each increase in the hyperparameter. This is called the problem of dimensionality in a grid search.

- **Random Search:**
  - In this technique, random combinations of hyperparameters set are tried and evaluated for finding the best solution. For optimizing the search, the function is tested at random configurations in parameter space as shown in the image below.
  - In this method, there are increased chances of finding optimal parameters because the pattern followed is random. There are chances that the model is trained on optimized parameters without the need for aliasing.
  - This search works the best when there is a lower number of dimensions as it takes less time to find the right set.



Conclusion:

Data Science is a very vast field and comprises many topics like Data Mining, Data Analysis, Data Visualization, Machine Learning, Deep Learning, and most importantly it is laid on the foundation of mathematical concepts like Linear Algebra and Statistical analysis. Since there are a lot of pre-requisites for becoming a good professional Data Scientist, the perks and benefits are very big. Data Scientist has become the most sought out job role these days. In this article, we have seen the most commonly asked interview questions on Data Science for both freshers and experienced.

Useful Resources:

Links to More Interview Questions

C Interview Questions	Php Interview Questions	C Sharp Interview Questions
Web Api Interview Questions	Hibernate Interview Questions	Node Js Interview Questions
Cpp Interview Questions	Oops Interview Questions	Devops Interview Questions
Machine Learning Interview Questions	Docker Interview Questions	Mysql Interview Questions
Css Interview Questions	Laravel Interview Questions	Asp Net Interview Questions
Django Interview Questions	Dot Net Interview Questions	Kubernetes Interview Questions
Operating System Interview Questions	React Native Interview Questions	Aws Interview Questions
Git Interview Questions	Java 8 Interview Questions	Mongodb Interview Questions
Dbms Interview Questions	Spring Boot Interview Questions	Power Bi Interview Questions
Pl Sql Interview Questions	Tableau Interview Questions	Linux Interview Questions
Ansible Interview Questions	Java Interview Questions	Jenkins Interview Questions