1. Normal vs skewed distribution
    a. <u>Normal distribution</u>: Normal distribution, also known as Gaussian distribution, is a continuous probability distribution that is commonly used in statistics to describe real-world phenomena that tend to cluster around a central value.
    b. The normal distribution is characterized by its bell-shaped curve, which is symmetrical around the mean (the central value). The mean, median, and mode of a normal distribution are all equal, and the curve is completely determined by its mean and standard deviation.
    c. <u>Skewed distribution</u>: Skewed distribution is a type of probability distribution in which the data is not evenly distributed around the mean. In a skewed distribution, one tail of the distribution has more extreme values than the other tail.
    d. There are two types of skewed distributions:
        i. Positive skew: In a positive skew, the tail on the right side of the distribution is longer than the tail on the left side. This means that the mean is greater than the median and the mode.
        ii. Negative skew: In a negative skew, the tail on the left side of the distribution is longer than the tail on the right side. This means that the mean is less than the median and the mode.
2. What are mean median and mode?
    a. <u>Mean</u>: The mean, also known as the average, is the sum of all the values in a dataset divided by the total number of values. It represents the central value of the dataset and is influenced by extreme values, or outliers. The formula for calculating the mean is:
       mean = (sum of values) / (number of values)
    b. <u>Median</u>: The median is the middle value in a dataset when the values are arranged in order from smallest to largest (or largest to smallest). It represents the value that is exactly in the middle of the dataset, with half the values above and half below it. The median is not affected by extreme values. To calculate the median, you need to arrange the values in order and find the value that is exactly in the middle. If there are an even number of values, the median is the average of the two middle values.
    c. <u>Mode</u>: The mode is the value that occurs most frequently in a dataset. It represents the most common value in the dataset. A dataset can have one or more modes, or no mode at all. The mode can be used for categorical data as well as numerical data.
3. What is bias?
    a. Bias in statistics refers to a systematic error or distortion in the data or analysis that leads to incorrect results or conclusions. Bias can occur in different stages of the data analysis process, such as data collection, data pre-processing, sampling, or analysis.
    b. There are different types of bias, including:
        i. <u>Selection bias</u>: This occurs when the sample of data is not representative of the population of interest, and some groups or characteristics are overrepresented or underrepresented in the sample. This can lead to incorrect inferences about the population.
        ii. <u>Measurement bias</u>: This occurs when the measurement or observation method used in the data collection is flawed, and it leads to inaccurate or incomplete data. For example, a survey that asks leading questions may produce biased responses.
        iii. <u>Confounding bias</u>: This occurs when there is a third variable that affects both the exposure and the outcome of interest, and it leads to a false association between them. For example, a study that finds a positive correlation between ice cream consumption and crime rates may be confounded by temperature, which affects both ice cream consumption and crime rates.
4. What is Variance?
    a. Variance is a statistical measure that describes the variability or spread of a dataset. It is a measure of how much the individual data points deviate from the mean, or central value, of the dataset. In other words, variance tells us how much the data points are spread out around the average value.
    b. The formula for calculating the variance of a dataset is:
        i. variance = (sum of squared deviations from the mean) / (number of data points - 1)
       where the deviation is the difference between each data point and the mean, squared and added together.
    c. A high variance indicates that the data points are widely spread out from the mean, while a low variance indicates that the data points are clustered closely around the mean.
5. How to handle missing values?
    a. Missing values are a common problem in data analysis and can occur for various reasons, such as incomplete data collection, data entry errors, or data loss. Handling missing values is important because they can affect the accuracy and reliability of statistical analysis and machine learning models. Here are some common techniques for handling missing values:
        i. <u>Deletion</u>: This involves removing the rows or columns that contain missing values from the dataset. There are two types of deletion: listwise deletion, which removes the entire row with a missing value, and pairwise deletion, which removes only the observations with missing values for the specific variables being analysed. Deletion is a simple approach but can lead to loss of valuable information and reduce the sample size.

   ii. Imputation: This involves filling in the missing values with estimated or imputed values. There are several methods for imputation, including mean imputation, median imputation, mode imputation, regression imputation, and multiple imputation. The choice of imputation method depends on the nature of the data and the missing data pattern.
   iii. Prediction: This involves using machine learning algorithms to predict the missing values based on the available data. This approach can be more accurate than simple imputation methods but requires more computational resources and can be affected by the quality of the prediction model.
   iv. Expert knowledge: In some cases, missing values can be filled in using expert knowledge or domain expertise. For example, a medical expert may be able to provide a plausible value for a missing medical record.

6. How to detect outliers?
   a. Outliers are data points that are significantly different from other data points in the dataset and can have a large impact on statistical analysis or machine learning models. Detecting outliers is an important step in data pre-processing and can be done using various techniques, including:
      i. Visual inspection: This involves creating graphical representations of the data, such as scatter plots or box plots, and looking for data points that fall outside the expected range or pattern. Outliers can often be visually identified as data points that are far away from the bulk of the data.
      ii. Statistical methods: This involves using statistical tests to identify data points that are significantly different from the rest of the data. Common statistical methods include z-score, modified z-score, and interquartile range (IQR). These methods calculate the distance between the data point and the mean or median of the dataset and compare it to a threshold value to identify outliers.
      iii. Machine learning methods: This involves using machine learning algorithms, such as clustering or anomaly detection, to identify data points that are significantly different from the rest of the data. These methods can be more accurate than statistical methods but may require more computational resources and domain expertise.

7. Explain Box Plot?
   a. A box plot, also known as a box-and-whisker plot, is a graphical representation of a dataset that summarizes the distribution of the data. It shows the median, quartiles, and outliers of the data in a simple and easy-to-interpret way.
   b. A box plot is made up of several components:
      i. Median: The median is the middle value of the dataset, where half of the data points are below it and half are above it. It is represented by a horizontal line inside the box.
      ii. Quartiles: The quartiles divide the dataset into four equal parts, each containing 25% of the data. The first quartile (Q1) is the value below which 25% of the data fall, while the third quartile (Q3) is the value below which 75% of the data fall. They are represented by the bottom and top edges of the box, respectively.
      iii. Interquartile range (IQR): The IQR is the range between the first and third quartiles and contains 50% of the data. It is represented by the height of the box.
      iv. Whiskers: The whiskers are lines that extend from the top and bottom edges of the box to the highest and lowest non-outlier data points within a certain distance from the quartiles. The distance can be set to a fixed value or calculated as a multiple of the IQR. Data points outside the whiskers are considered outliers and are represented by individual points.
      v. Outliers: Outliers are data points that fall outside the whiskers and are represented by individual points.

8. What is Z-score?
   a. Z-score, also known as the standard score, is a statistical measure that represents the number of standard deviations a data point is away from the mean of the dataset. It is calculated as:

      $$Z = (X - \mu) / \sigma$$

      where X is the data point, μ is the mean of the dataset, and σ is the standard deviation of the dataset.
   b. The Z-score indicates whether a data point is above or below the mean, and by how many standard deviations. A positive Z-score means that the data point is above the mean, while a negative Z-score means that it is below the mean. A Z-score of 0 means that the data point is exactly at the mean.
   c. Z-scores are useful for identifying outliers and extreme values in a dataset, as well as for comparing values from different datasets that may have different scales and units. A common threshold for identifying outliers is a Z-score greater than or equal to 3 or less than or equal to -3, which represents data points that are more than three standard deviations away from the mean.
   d. Z-scores are widely used in various fields, such as finance, quality control, and sports, to standardize and compare data values.

9. What Is IQR?

   a. IQR stands for Interquartile Range. It is a measure of variability in a dataset and is used to describe the spread of the middle 50% of the data. The IQR is calculated as the difference between the first quartile (Q1) and the third quartile (Q3) of the dataset:

      IQR = Q3 - Q1

      The first quartile (Q1) is the value that separates the lowest 25% of the data from the rest, while the third quartile (Q3) separates the highest 25% of the data from the rest. The IQR therefore contains the middle 50% of the data.
   b. The IQR is useful for detecting outliers in a dataset, as data points that fall outside the IQR are often considered to be potential outliers. Specifically, data points below Q1 - 1.5IQR or above Q3 + 1.5IQR are commonly considered outliers.

10. What is Over fitting?
    a. Overfitting is a common problem in machine learning and occurs when a model is trained too well on the training data and as a result, it becomes overly specialized to the training data and performs poorly on new, unseen data.
    b. In other words, overfitting occurs when a model is too complex and captures the noise in the training data as well as the underlying patterns. This can result in a model that is too tightly tuned to the training data and fails to generalize well to new data.
    c. Signs of overfitting include a high accuracy on the training data but a low accuracy on the test data or new data. This is because the model has learned to recognize the specific patterns in the training data but is unable to generalize to new data that may have different patterns.
11. What is Under fitting?
    a. Underfitting is the opposite of overfitting and occurs when a machine learning model is too simple and unable to capture the underlying patterns in the data, resulting in poor performance on both the training and test data.
    b. An underfit model is characterized by high bias and low variance, which means that the model is not flexible enough to fit the training data well and also unable to generalize to new data. Signs of underfitting include low accuracy on both the training and test data, which indicates that the model is not capturing the underlying patterns in the data.
    c. Underfitting can occur when a model is too simple or when the dataset is too small or noisy. To overcome underfitting, it is important to use more complex models, such as deep neural networks, or to add more features to the dataset. In addition, increasing the complexity of the model can help reduce bias and improve performance. However, care must be taken not to overfit the model in the process. Cross-validation can also be used to assess whether the model is underfitting or overfitting.
12. How to avoid overfitting?

To avoid overfitting, there are several techniques that can be used:

a. Regularization: Regularization is a technique used to reduce the complexity of a model and prevent overfitting. It involves adding a penalty term to the loss function of the model, which discourages large weights and biases. There are two commonly used regularization techniques: L1 regularization and L2 regularization.
b. Cross-validation: Cross-validation is a technique used to evaluate the performance of a model on new data. It involves dividing the dataset into training and validation sets and training the model on the training set, then evaluating its performance on the validation set. This process is repeated multiple times with different splits of the data, and the average performance is used to assess the model's performance.
c. Early stopping: Early stopping is a technique used to prevent overfitting by stopping the training of a model before it overfits. It involves monitoring the model's performance on a validation set during training and stopping the training process when the performance on the validation set starts to degrade.
d. Dropout: Dropout is a regularization technique used in neural networks. It involves randomly dropping out neurons during training, which prevents the network from relying too heavily on any one feature.
e. Data augmentation: Data augmentation involves generating new data from the existing dataset by applying transformations such as rotation, scaling, and cropping. This can help prevent overfitting by increasing the size and diversity of the dataset.
f. Simplify the model architecture: If the model is too complex, it may overfit the data. One way to avoid this is to simplify the model architecture by reducing the number of layers or nodes. This can help the model generalize better to new data.
g. Increase the size of the dataset: Increasing the size of the dataset can help prevent overfitting by providing more data for the model to learn from. This can help the model generalize better to new data.

13. How to avoid underfitting?

To avoid underfitting, there are several techniques that can be used:

a. Increase model complexity: If the model is too simple, it may underfit the data. One way to avoid this is to increase the complexity of the model by adding more layers or neurons. This can help the model capture more complex patterns in the data.

b. Add more features: If the dataset is too simple, it may underfit the model. One way to avoid this is to add more features to the dataset. This can help the model capture more complex patterns in the data.

c. Reduce regularization: Regularization can help prevent overfitting, but too much regularization can also lead to underfitting. If the model is underfitting, reducing the amount of regularization can help.

d. Increase training time: If the model is not learning enough from the data, increasing the training time can help. This can allow the model to learn more complex patterns in the data.

e. Change the model architecture: If the model is not able to capture the patterns in the data, changing the model architecture can help. This can involve trying different types of models, such as neural networks or decision trees, to find the best fit for the data.

f. Tune hyperparameters: Hyperparameters are settings that determine the behaviour of the model. If the model is underfitting, tuning the hyperparameters can help. This can involve adjusting the learning rate, batch size, or other settings to find the best fit for the data.

g. Increase the size of the dataset: If the dataset is too small, it may not provide enough data for the model to learn from. Increasing the size of the dataset can help the model generalize better to new data and prevent underfitting.

14. What is supervised and unsupervised learning?

a. Supervised learning: In supervised learning, the algorithm is trained on a labelled dataset, where each data point is associated with a label or output. The goal of the algorithm is to learn a function that maps inputs to outputs based on the training data, so that it can predict the correct output for new, unseen data. Examples of supervised learning algorithms include linear regression, logistic regression, decision trees, and neural networks.

b. Unsupervised learning: In unsupervised learning, the algorithm is trained on an unlabelled dataset, where there are no output labels. The goal of the algorithm is to identify patterns and structure in the data, such as clusters or groups of similar data points. Examples of unsupervised learning algorithms include k-means clustering, principal component analysis (PCA), and autoencoders.

The main difference between supervised and unsupervised learning is the availability of labelled data. Supervised learning algorithms require labelled data to learn from, while unsupervised learning algorithms can learn from unlabelled data. Supervised learning is often used in tasks such as classification and regression, where the goal is to predict a specific output. Unsupervised learning is often used in tasks such as clustering and dimensionality reduction, where the goal is to identify patterns and structure in the data.

15. Regression V/S Classification

a. Regression: Regression is a type of supervised learning used when the output variable is continuous or numerical. The goal of regression is to predict the value of the output variable based on the input variables. Examples of regression problems include predicting the price of a house based on its features, predicting the sales of a product based on marketing spend, and predicting the amount of rainfall based on temperature and humidity.

b. Classification: Classification is a type of supervised learning used when the output variable is categorical or discrete. The goal of classification is to predict the class or category of the output variable based on the input variables. Examples of classification problems include predicting whether an email is spam or not, predicting whether a customer will churn or not, and predicting whether a patient has a certain disease or not.

The main difference between regression and classification is the type of output variable. Regression is used for numerical output variables, while classification is used for categorical output variables. The algorithms used for regression and classification are also different. Regression algorithms include linear regression, polynomial regression, and decision trees, while classification algorithms include logistic regression, decision trees, and support vector machines.

In summary, regression is used to predict a numerical value, while classification is used to predict a categorical value.

16. Types of regression and classification

a. Types of Regression:

i. Linear Regression: A technique where the relationship between the input and output variable is assumed to be linear. It is used to predict a continuous output variable.

ii. Polynomial Regression: A technique where the relationship between the input and output variable is modelled as an nth degree polynomial equation. It is also used to predict a continuous output variable.

iii. Logistic Regression: A technique where the output variable is a binary value (0 or 1). It is used to predict the probability of an event occurring.

iv. Ridge Regression: A technique used to prevent overfitting in linear regression models. It introduces a penalty term to the loss function, which helps in reducing the magnitude of the coefficients.

v. Lasso Regression: A technique used to perform variable selection and reduce the number of features in the model. It introduces a penalty term that shrinks the coefficients of irrelevant features to zero.

b. Types of Classification:

i. Binary Classification: A technique where the output variable has two possible values (0 or 1). Examples include spam detection, fraud detection, and disease diagnosis.
ii. Multi-class Classification: A technique where the output variable can have more than two possible values. Examples include image classification, handwriting recognition, and language translation.
iii. Decision Tree Classification: A technique that uses a tree-like model of decisions and their possible consequences. It is used to classify inputs into different categories or classes.
iv. Random Forest Classification: A technique that uses an ensemble of decision trees to improve the accuracy and reduce overfitting. It is used for classification problems with many features.
v. Support Vector Machines (SVM) Classification: A technique that finds the best separating boundary between the different classes. It is used for both binary and multi-class classification problems.

17. What is confusion matrix?
  a. A confusion matrix is a table that is used to evaluate the performance of a classification model by comparing its predicted output with the actual output. It is a matrix that summarizes the results of a binary or multi-class classification problem.
  b. A confusion matrix has four main components:
    i. True Positives (TP): The number of instances where the actual output and predicted output are both positive.
    ii. False Positives (FP): The number of instances where the actual output is negative but the predicted output is positive.
    iii. True Negatives (TN): The number of instances where the actual output and predicted output are both negative.
    iv. False Negatives (FN): The number of instances where the actual output is positive but the predicted output is negative.
  c. A confusion matrix is typically represented as a 2x2 or 3x3 table, depending on the number of classes in the classification problem. It can be used to calculate various evaluation metrics such as accuracy, precision, recall, F1 score, and others.
    i. Accuracy = (TP + TN) / (TP + TN + FP + FN)
    ii. Precision = TP / (TP + FP)
    iii. Recall = TP / (TP + FN)
    iv. F1 Score = 2 * ((Precision * Recall) / (Precision + Recall))
  d. Overall, a confusion matrix provides a detailed picture of how well a classification model is performing and can help in identifying areas of improvement.

18. Define Accuracy, Precision, Recall and F1 Score.
  a. Accuracy: Accuracy is the most basic evaluation metric used to measure the performance of a classification model. It measures the percentage of correctly classified instances among all the instances in the dataset.

     Accuracy = (TP + TN) / (TP + TN + FP + FN)

     where, TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives

  b. Precision: Precision is a metric that measures the percentage of true positive instances among all the instances predicted as positive by the model. It measures how accurate the model is in predicting positive instances.

     Precision = TP / (TP + FP)

  c. Recall: Recall is a metric that measures the percentage of true positive instances among all the instances that are positive. It measures how well the model can identify positive instances.

     Recall = TP / (TP + FN)

  d. F1 Score: F1 Score is a metric that combines both precision and recall into a single value. It is the harmonic mean of precision and recall and provides a balance between the two metrics.

     F1 Score = 2 * ((Precision * Recall) / (Precision + Recall))

19. What P-values?
  a. P-values are a statistical measure used to determine the significance of an observed effect or result. They are used to test a null hypothesis and determine the probability of obtaining a result as extreme as the observed result, assuming the null hypothesis is true.
  b. The null hypothesis is a statement or assumption that there is no significant difference or relationship between two variables or groups in a population. The alternative hypothesis is the opposite of the null hypothesis, and it states that there is a significant difference or relationship between the variables or groups in the population.
  c. The p-value is a probability value between 0 and 1. If the p-value is less than the level of significance (usually set to 0.05 or 0.01), the null hypothesis is rejected, and the alternative hypothesis is accepted. This means that the observed effect or result is statistically significant and not due to chance. If the p-value is greater than the

level of significance, the null hypothesis is accepted, and the observed effect or result is not considered statistically significant.

   d. In simple terms, a p-value tells us how likely it is to obtain a result as extreme as the observed result, assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence against the null hypothesis.

20. What is residual?
    a. A residual is the difference between the observed value of a dependent variable and the predicted value of that variable. It is the vertical distance between the observed data point and the regression line or curve.
    b. Residuals are an important concept in regression analysis, where the goal is to find a model that best fits the data. A regression model makes predictions for the dependent variable based on the values of one or more independent variables. The difference between the predicted value and the actual value of the dependent variable is the residual.
    c. A residual can be positive or negative, depending on whether the observed value is greater or less than the predicted value. The sum of the squared residuals (SSR) is used as a measure of how well the regression model fits the data. A smaller SSR indicates a better fit, while a larger SSR indicates a poorer fit.
    d. Residual plots are often used to visually inspect the residuals and check for patterns or trends that indicate the regression model may not be appropriate for the data. Ideally, residuals should be randomly scattered around zero, with no discernible pattern. If there is a pattern in the residuals, it suggests that the model may be missing some important factors that are affecting the dependent variable.
21. Dimensionality reduction
    a. Dimensionality reduction is a technique used in machine learning and data science to reduce the number of features or variables in a dataset while retaining as much relevant information as possible. The goal of dimensionality reduction is to simplify the data without losing critical information or insights.
    b. The need for dimensionality reduction arises when a dataset contains many variables or features, which can cause several problems such as:
        i. Increased complexity: Many variables can make a dataset more complex and difficult to analyse.
        ii. Computational inefficiency: large datasets can require more computational resources and time to process.
        iii. Overfitting: A model trained on a dataset with too many features can overfit the training data and perform poorly on new data.
    c. Dimensionality reduction can be achieved in two main ways:
        i. Feature selection: Feature selection is the process of selecting a subset of the most relevant features from a dataset based on certain criteria, such as correlation with the target variable, variance, or importance. The selected features are used to build a model or perform further analysis.
        ii. Feature extraction: Feature extraction is the process of transforming the original features into a new set of features, which captures the essential information in the data. This can be achieved through techniques such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and Non-negative Matrix Factorization (NMF).
    d. Principal Component Analysis (PCA) is a popular technique for dimensionality reduction, where the original variables are transformed into a new set of variables called principal components. These components are orthogonal to each other and capture the maximum variance in the data. The components with the highest variance are retained, while the ones with low variance are discarded.
22. Steps involved in data science?
    a. Problem definition: Define the problem or question to be answered by the data analysis. This step involves understanding the business problem or research question, identifying the data sources, and determining the data required.
    b. Data collection: Collect the data required for analysis. This may involve obtaining data from internal or external sources, such as databases, APIs, or web scraping.
    c. Data cleaning: Clean the data by removing missing values, correcting errors, dealing with outliers, and transforming the data into a suitable format for analysis.
    d. Data exploration: Explore the data to understand the characteristics of the data, identify patterns, and relationships among variables.
    e. Feature engineering: Create new features or transform existing ones to better represent the underlying patterns in the data.
    f. Modelling: Develop a predictive model based on the data and the problem definition. This involves selecting an appropriate algorithm, training the model on a subset of the data, and validating the model's performance on a separate test set.
    g. Evaluation: Evaluate the model's performance and refine it if necessary. This may involve adjusting the model parameters, changing the feature set, or trying different algorithms.
    h. Deployment: Deploy the model in a production environment, if applicable. This may involve integrating the model into a software system or creating a user interface.
    i. Monitoring: Monitor the performance of the model over time and update it as necessary. This may involve retraining the model on new data or adjusting the model parameters to account for changing conditions.

j. Communication: Communicate the results of the analysis to stakeholders, including visualizations, reports, and presentations. This step involves translating technical results into business insights that can be used to inform decision-making.

23. Explain deployment.

Deployment in the context of data science refers to the process of taking a predictive model developed during the modelling phase of a project and integrating it into a production environment where it can be used to make predictions on new data. This is an important step in the data science workflow, as the value of a model is only realized when it is put into use.

Deploying a model involves several steps, including:

a. Saving the model: The first step is to save the trained model as a file or package that can be loaded into a production environment.
b. Pre-processing the data: The input data to the model may need to be pre-processed to ensure it is in the correct format and contains all the required features.
c. Integrating the model: The model must be integrated into the software infrastructure of the production environment. This may involve developing an API or library to interface with the model.
d. Testing the model: The deployed model must be tested to ensure that it is producing accurate and reliable predictions. This testing should include edge cases, or scenarios that may not be present in the training data.
e. Monitoring the model: Once the model is deployed, it must be monitored to ensure that it continues to produce accurate and reliable predictions. This may involve logging predictions and monitoring their accuracy over time.
f. Updating the model: Over time, the model may need to be updated as new data becomes available or the underlying patterns in the data change. This may involve retraining the model or adjusting its parameters.

Deployment is a critical step in the data science workflow, as it allows organizations to leverage the insights gained from the analysis to make better decisions and improve their operations.

24. Explain decision tree and random forest.
    a. Decision trees are a type of supervised learning algorithm that is used for classification and regression. The decision tree algorithm builds a tree-like model of decisions and their possible consequences, where each node represents a decision based on a feature, and each branch represents a possible outcome or decision based on the feature value. The decision tree algorithm splits the data based on the feature that provides the best separation between the classes or the highest reduction in variance for regression tasks. The decision tree continues to split the data recursively until a stopping criterion is met, such as the maximum depth of the tree, minimum number of samples required to split a node, or minimum improvement in the criterion.
    b. Random forests are a type of ensemble learning algorithm that combines multiple decision trees to improve the accuracy and reduce the overfitting of the model. The random forest algorithm creates multiple decision trees on different subsets of the training data and feature subsets. During the training process, each tree is trained on a random subset of the features and data samples, and the final prediction is made by taking the majority vote of all the trees. This technique is called bagging, which helps to reduce the variance and improve the generalization of the model.
    c. Random forests also use a technique called feature bagging or random subspace method, where each tree is trained on a random subset of the features, which helps to reduce the correlation between the trees and improve the diversity of the ensemble.
25. Explain pruning.
    a. Pruning is a technique used in decision tree algorithms to prevent overfitting and improve the accuracy and generalization of the model. Overfitting occurs when the decision tree is too complex and captures noise or irrelevant features in the training data, which results in poor performance on new data.
    b. Pruning involves removing some of the branches or nodes from the decision tree to simplify the model and reduce its complexity. There are two types of pruning techniques: pre-pruning and post-pruning.
        i. Pre-pruning is a technique that stops the tree building process early, before the tree becomes too complex. The decision tree algorithm stops splitting a node if a certain condition is met, such as the maximum depth of the tree, the minimum number of samples required to split a node, or the minimum improvement in the criterion. Pre-pruning is a simple and efficient technique, but it may not result in the optimal tree, and the chosen stopping criterion may not be the best for the specific problem.
        ii. Post-pruning is a technique that prunes the fully grown tree after the tree building process is complete. Post-pruning involves removing some of the branches or nodes from the tree using a validation set or cross-validation. The algorithm evaluates the effect of removing each node on the accuracy of the model on the validation set and chooses the node that results in the highest increase in accuracy. The process continues until removing any additional nodes does not result in an improvement in accuracy. Post-pruning is a more computationally expensive technique but may result in a more optimal and generalized model.
26. Explain RMSE.

a. RMSE stands for Root Mean Squared Error, which is a commonly used evaluation metric for regression problems. It measures the average deviation of the predictions made by a regression model from the actual target values in the dataset.
b. RMSE is calculated by taking the square root of the average of the squared differences between the predicted and actual values. It is calculated as follows:

RMSE = sqrt(mean((y_pred - y_true)^2))

Where y_pred is the predicted values from the regression model, y_true is the actual target values in the dataset, and mean() is the average function.

c. The lower the RMSE value, the better the performance of the regression model. However, it should be noted that RMSE should be used in combination with other evaluation metrics and domain knowledge to fully understand the performance of the model and make informed decisions.

27. What is elbow method
    a. The elbow method is a graphical technique used in unsupervised machine learning, particularly in clustering algorithms, to determine the optimal number of clusters for a given dataset.
    b. The elbow method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters, where WCSS measures the sum of the squared distances between each point and its assigned cluster centre. The plot resembles an elbow, and the optimal number of clusters is usually located at the "elbow point," where the addition of another cluster does not result in a significant reduction in WCSS.
28. Explain K-means clustering.
    a. K-means clustering is a popular unsupervised machine learning algorithm used to group a dataset into a specified number of clusters. The algorithm works by iteratively partitioning the data into K clusters, where K is a user-defined number of clusters.
    b. The K-means algorithm works as follows:
       i. Initialize K cluster centres randomly in the feature space.
       ii. Assign each data point to the nearest cluster centre based on the Euclidean distance.
       iii. Recalculate the cluster centres as the mean of all the data points assigned to each cluster.
       iv. Repeat steps 2-3 until the cluster centres converge or a maximum number of iterations is reached.
    c. The goal of K-means clustering is to minimize the sum of squared distances between each data point and its assigned cluster centre, which is known as the Within-Cluster Sum of Squares (WCSS).
29. Explain naïve bayes.
    a. The Naive Bayes algorithm works as follows:
       i. Given a dataset with labelled instances, calculate the prior probability of each class.
       ii. For each feature in the dataset, calculate the likelihood of that feature given each class.
       iii. For a new instance with unknown class, calculate the posterior probability of each class using Bayes' theorem and the calculated priors and likelihoods.
       iv. Assign the new instance to the class with the highest posterior probability.
    b. The Naive Bayes algorithm assumes that all the features in the dataset are independent of each other, which is often not true in real-world applications. Despite this oversimplified assumption, the Naive Bayes algorithm is still effective and efficient in many applications, especially in text classification and spam filtering, where the assumption of conditional independence holds reasonably well.
30. Disadvantages of a linear model.
    a. Limited flexibility: Linear models assume a linear relationship between the independent and dependent variables, which can be limiting in certain situations where the relationship is not linear.
    b. Overfitting: If the model is too complex, it can overfit the data, meaning that it will perform well on the training data but poorly on new data.
    c. Underfitting: On the other hand, if the model is too simple, it may underfit the data and not capture the true relationship between the variables.
    d. Assumptions: Linear models have certain assumptions about the data, such as linearity, independence, homoscedasticity, and normality of residuals, which may not always hold true in practice.
    e. Outliers: Linear models are sensitive to outliers, which can distort the line of best fit and affect the model's performance.
    f. Non-linear relationships: Linear models cannot capture non-linear relationships between variables, and may not be suitable for data with complex relationships.
    g. Interpretability: While linear models are generally easy to interpret, they may not provide a complete understanding of the data and may miss important interactions between variables.
31. Explain normalization and standardization.
    a. Normalization is a process of scaling numeric data to a range of 0 to 1, where the minimum value in the dataset is scaled to 0, and the maximum value is scaled to 1. This technique is useful when the scale of the data varies widely, as it can help to avoid certain variables having a disproportionate impact on the analysis. Normalization can be performed using the formula:

X_normalized = (X - X_min) / (X_max - X_min)

where X is the original value, X_min is the minimum value in the dataset, and X_max is the maximum value in the dataset.

   b. Standardization, on the other hand, is a process of scaling data to have zero mean and unit variance. This technique is useful when the distribution of the data is not Gaussian, and when the scale of the data varies widely. Standardization can be performed using the formula:

      X_standardized = (X - X_mean) / X_std

      where X is the original value, X_mean is the mean value of the dataset, and X_std is the standard deviation of the dataset.

32. Steps involved in hypothesis testing.
    a. State the null hypothesis (H0) and the alternative hypothesis (Ha): The null hypothesis is a statement about the population parameter that is assumed to be true until there is sufficient evidence to reject it. The alternative hypothesis is a statement that contradicts the null hypothesis and represents the alternative explanation for the observed data.
    b. Choose the level of significance (α): The level of significance is the probability of rejecting the null hypothesis when it is actually true. It is usually set at 0.05 or 0.01.
    c. Select the appropriate test statistic: The choice of test statistic will depend on the type of hypothesis being tested and the nature of the data.
    d. Calculate the p-value: The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the one observed in the sample data, assuming that the null hypothesis is true.
    e. Compare the p-value to the level of significance: If the p-value is less than or equal to the level of significance, then the null hypothesis is rejected in favour of the alternative hypothesis. If the p-value is greater than the level of significance, then the null hypothesis cannot be rejected.
    f. Draw a conclusion and interpret the results: If the null hypothesis is rejected, it means that there is sufficient evidence to support the alternative hypothesis. If the null hypothesis is not rejected, it means that there is not enough evidence to support the alternative hypothesis.
    g. Check the assumptions: It is important to check the assumptions of the statistical test to ensure that the results are valid and reliable.
33. training test vs validation set vs test set
    a. Training set: This is the subset of the data used to train the machine learning model. The model is trained on the training set, and the parameters are adjusted to minimize the error between the predicted values and the actual values.
    b. Validation set: This is the subset of the data used to evaluate the performance of the model during training. The validation set is used to tune the hyperparameters of the model, such as the learning rate, regularization, and number of layers. The model is evaluated on the validation set to check if it is overfitting or underfitting the training data.
    c. Test set: This is the subset of the data used to evaluate the final performance of the model after it has been trained and validated. The test set is used to check how well the model generalizes to new, unseen data. The model should not have been exposed to the test set during training or validation to avoid bias.
34. Differential statistical methods used in data science.
    a. T-test: A t-test is used to compare the means of two groups and determine if they are statistically different. The independent t-test is used when the two groups are independent of each other, and the paired t-test is used when the two groups are dependent on each other.
    b. ANOVA: Analysis of variance (ANOVA) is used to compare the means of more than two groups and determine if there are significant differences between them. ANOVA can be one-way or two-way, depending on the number of factors being compared.
    c. Chi-squared test: The chi-squared test is used to determine if there is a significant association between two categorical variables. It is commonly used to test for independence between two variables.
    d. Z-test: A Z-test is a statistical hypothesis test used to determine whether two population means are different when the standard deviation is known and the sample size is large. The test is based on the normal distribution and the standard normal distribution, which has a mean of 0 and a standard deviation of 1.
    e. Mann-Whitney U test: The Mann-Whitney U test is used to compare the medians of two groups when the data is not normally distributed or the sample size is small.
    f. Kruskal-Wallis test: The Kruskal-Wallis test is used to compare the medians of more than two groups when the data is not normally distributed or the sample size is small.
    g. Wilcoxon signed-rank test: The Wilcoxon signed-rank test is used to compare the medians of two dependent samples when the data is not normally distributed.
    h. Friedman test: The Friedman test is used to compare the medians of more than two dependent samples when the data is not normally distributed.
35. SVM

a. SVM (Support Vector Machines) is a supervised machine learning algorithm that can be used for classification or regression tasks. The goal of SVM is to find the optimal hyperplane that separates two classes of data points with the largest margin. In other words, SVM tries to find a decision boundary that maximizes the distance between the two classes of data points.

b. The key concept behind SVM is the use of support vectors, which are the data points closest to the decision boundary. These support vectors are used to define the optimal hyperplane and classify new data points based on their location relative to the decision boundary.

c. SVM can be used with different kernel functions to handle non-linearly separable data by mapping the data to a higher-dimensional feature space where the data becomes linearly separable. Some commonly used kernel functions are linear, polynomial, radial basis function (RBF), and sigmoid.

d. SVM has many advantages, such as being effective in high-dimensional spaces, handling non-linearly separable data, and having a unique solution. However, SVM can be sensitive to the choice of kernel function and the parameter tuning process. Also, SVM can be computationally expensive when dealing with large datasets.

e. Overall, SVM is a powerful and versatile machine learning algorithm that can be used for a variety of classification and regression tasks.

36. What is a kernel?

a. In machine learning, a kernel is a function that takes two inputs, computes the similarity between them, and returns a scalar value. Kernels are commonly used in support vector machines (SVM) and other machine learning algorithms that involve comparing pairs of data points.

b. The kernel function is used to map the input data points to a higher-dimensional feature space, where it may be easier to find a linear decision boundary that separates the data points. This is known as the kernel trick, and it allows the algorithm to perform well even when the data is not linearly separable in its original feature space.

c. Some commonly used kernel functions include:

i. Linear kernel: This kernel computes the dot product between two vectors, which is equivalent to a linear transformation of the data

ii. Polynomial kernel: This kernel computes the dot product between two vectors raised to a power, allowing for non-linear transformations of the data.

iii. Radial basis function (RBF) kernel: This kernel computes the distance between two vectors using the Gaussian function, which allows for a smooth, non-linear transformation of the data

iv. Sigmoid kernel: This kernel is similar to the RBF kernel, but uses the sigmoid function instead of the Gaussian function.

d. The choice of kernel function depends on the nature of the data and the specific machine learning algorithm being used. The kernel function is often tuned through a process called hyperparameter tuning, where different kernel functions and their parameters are evaluated to find the best combination for a given problem.

37. Explain correlation and covariance.

a. Covariance measures the degree to which two variables vary together. Specifically, covariance measures the expected value of the product of the deviations of two random variables from their respective means. If the covariance between two variables is positive, it means that they tend to increase or decrease together, while a negative covariance means that they tend to vary in opposite directions. A covariance of zero means that the variables are uncorrelated, which does not necessarily imply independence.

b. Correlation, on the other hand, is a standardized measure of the linear relationship between two variables. Correlation is calculated as the covariance between two variables divided by the product of their standard deviations. Correlation ranges from -1 to 1, where a correlation of -1 indicates a perfect negative linear relationship, a correlation of 0 indicates no linear relationship, and a correlation of 1 indicates a perfect positive linear relationship.

38. What is cross validation?

a. Cross-validation is a technique used in machine learning and data science to evaluate the performance of a model on a limited data sample. The main idea behind cross-validation is to divide the available data into two or more subsets, where one subset is used for training the model and the other subset is used for testing its performance. This approach helps to estimate how well the model can generalize to new data that was not used during training.

b. Cross-validation can help to prevent overfitting of the model, where the model is too closely tailored to the training data and performs poorly on new data. It can also help to select the best model or tuning parameters by comparing their performance across different cross-validation folds.

39. What is sampling and techniques used in sampling?

a. Sampling is a technique used in statistics and data science to select a representative subset of data from a larger population. The goal of sampling is to gather enough information from the subset to make inferences about the population, while reducing the cost and time required to collect data.

b. There are two main types of sampling: probability sampling and non-probability sampling.

i. Probability sampling is a method where every member of the population has an equal chance of being selected for the sample. Some common techniques of probability sampling include:

1. Simple random sampling: where every member of the population has an equal chance of being selected.
2. Systematic sampling: where members of the population are selected at regular intervals from a random starting point.
3. Stratified sampling: where the population is divided into subgroups (or strata) based on some characteristic, and a sample is taken from each subgroup in proportion to its size.
4. Cluster sampling: where the population is divided into clusters, and a sample of clusters is selected at random. All members within the selected clusters are included in the sample.

ii. On the other hand, non-probability sampling methods do not involve random selection and therefore do not ensure that every member of the population has an equal chance of being selected. Some common techniques of non-probability sampling include:

1. Convenience sampling: where the sample is selected based on convenience or availability.
2. Quota sampling: where a specific number of individuals with certain characteristics are selected for the sample.
3. Snowball sampling: where participants are asked to recruit others to participate in the study.

c. While non-probability sampling methods are easier and less expensive to conduct, they may introduce bias into the sample and may not be representative of the population. Probability sampling methods, on the other hand, ensure that the sample is representative and can provide more reliable estimates of population parameters.

40. What is ROC curve?

a. A ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classification model that predicts the probability of a positive outcome. It is used to evaluate the accuracy of the model's predictions by comparing the true positive rate (TPR) and false positive rate (FPR) at different classification thresholds.

b. The ROC curve plots the TPR (also known as sensitivity or recall) on the y-axis and the FPR (1-specificity) on the x-axis. Each point on the curve represents a different threshold for classifying a positive or negative outcome. The area under the ROC curve (AUC) is a commonly used metric to evaluate the performance of a classification model, where a value of 1 indicates a perfect classifier and a value of 0.5 indicates a classifier that performs no better than chance.

c. A good classifier should have an ROC curve that is close to the top left corner of the plot, indicating a high TPR and low FPR across all classification thresholds. The ROC curve can also be used to compare the performance of different classification models or to choose the optimal threshold for a specific application.

41. Explain univariate bivariate and multivariate.

a. Univariate refers to the analysis of a single variable or feature in isolation, without considering the relationship with other variables. This type of analysis typically involves descriptive statistics and data visualization techniques such as histograms and box plots to summarize the distribution of the variable.

b. Bivariate analysis, on the other hand, examines the relationship between two variables. This type of analysis can help to identify patterns, trends, and correlations between variables, and can be used to build predictive models. Common techniques for bivariate analysis include scatter plots and correlation analysis.

c. Multivariate analysis involves the analysis of more than two variables, often with the aim of identifying complex patterns and relationships. Multivariate techniques include linear regression, principal component analysis, factor analysis, and cluster analysis, among others. Multivariate analysis can help to identify the underlying structure of data, reveal hidden patterns and relationships, and provide insights for decision making.

42. What are the libraries used in Data science.

a. NumPy: for numerical computing, including arrays, matrices, and mathematical functions
b. Pandas: for data manipulation and analysis, including data cleaning, merging, and grouping
c. Matplotlib: for data visualization, including creating charts and graphs
d. Scikit-learn: for machine learning, including classification, regression, and clustering algorithms
e. TensorFlow and Keras: for deep learning and neural networks

43. What is deep learning?

i. Deep learning is a subfield of machine learning that uses artificial neural networks to model and solve complex problems. These neural networks are composed of multiple layers of interconnected nodes, which allow the network to learn and extract increasingly complex features from the input data. Deep learning algorithms are particularly well-suited for tasks such as image and speech recognition, natural language processing, and playing games. They have achieved state-of-the-art results in many areas, and have enabled significant advancements in fields such as computer vision, robotics, and autonomous vehicles.