Module 1: Introduction to Data analytics and life cycle Weightage: 10 Marks

1. What is an analytic sandbox, and why is it important?
Ans:

- An analytic sandbox is a controlled environment where data analysts and data scientists can access, manipulate, and analyze data without affecting the live production system. It is typically a separate instance of a database or a virtual environment that is isolated from the production system, and it contains a subset of the data that is relevant to the analysis.
- The importance of an analytic sandbox lies in the fact that it provides a safe and secure place for data analysts and scientists to experiment with data without the risk of corrupting or disrupting the live production system. In the sandbox environment, analysts can run queries, develop models, test hypotheses, and validate their findings before deploying them to the production environment.
- Additionally, an analytic sandbox provides greater flexibility and agility to data analysis teams by enabling them to quickly prototype and iterate on their work without disrupting the production system. This means that they can experiment with new data sources, techniques, and tools without worrying about the impact on the live system.

2 Explain the differences between BI and Data Science.
Ans:

Parameters	**Business Intelligence (BI)**	**Data Science (DS)**
Focus	Past and Present	Future
Purpose	Reporting and Analysis	Prediction and Optimization
Data Volume	Historical and Structured	Historical and Unstructured
Data Sources	Internal and External	Mostly Internal
Data Analysis	Descriptive and Diagnostic	Predictive and Prescriptive
Tools	BI Tools, Dashboards	Programming languages (Python, R), Machine Learning Libraries
Skills	SQL, Data Warehousing, Reporting	Programming, Statistics, Machine Learning
Outcome	Insights and Recommendations	Actionable Insights and Recommendations

In essence, BI is focused on providing insights based on historical data to improve business operations. It is mostly used for descriptive and diagnostic analysis, and the data is often structured and comes from internal and external sources. In contrast, DS is focused on predicting future trends and behaviors based on historical and unstructured data. It involves using predictive and prescriptive analysis techniques, and the data is mostly from internal sources.

3. Describe the challenges of the current analytical architecture for data scientists.
Ans:
The current analytical architecture presents several challenges for data scientists, including:

1. Data Silos: Data silos arise when data is stored in multiple disparate systems or formats, making it difficult for data scientists to access and integrate data from different sources. This can lead to incomplete or inconsistent data, which hinders the accuracy and reliability of analytical insights.
2. Data Quality: Data quality is a crucial aspect of analytical work, and poor data quality can lead to inaccurate insights and decisions. Data scientists must spend a significant amount of time cleaning and pre-processing data to ensure its quality, which can be time-consuming and labor-intensive.
3. Scalability: Analytical workloads are often computationally intensive and require scalable architectures to handle large volumes of data. However, scaling up analytical architectures can be

challenging due to infrastructure limitations, data storage costs, and the need for specialized hardware and software.

4. Skill Set: Analytical work requires a diverse set of skills, including statistics, machine learning, programming, and data visualization. However, finding individuals with the right skill set to fill data science roles can be challenging, leading to skill gaps within organizations.
5. Security and Privacy: Analytical work often involves sensitive data, such as personal information or financial data. Therefore, data scientists must ensure that their analytical architectures are secure and comply with privacy regulations, such as GDPR or CCPA.
6. Speed and Agility: In the fast-paced business environment, data scientists need to work quickly and efficiently to deliver insights that drive business value. However, legacy analytical architectures may be slow and inflexible, hindering their ability to deliver insights in a timely manner.

4. What are the key skill sets and behavioral characteristics of a data scientist?

Ans: The key skill sets and behavioral characteristics of a data scientist include:

1. Statistics and Mathematics: Data scientists need a strong foundation in statistics and mathematics to understand and apply statistical models and algorithms to data analysis.
2. Programming Languages: Data scientists must have expertise in programming languages such as Python, R, or SQL to manipulate and analyze data, create predictive models, and build data visualizations.
3. Machine Learning: Data scientists need a deep understanding of machine learning algorithms and techniques to apply them to predictive modeling, clustering, classification, and other applications.
4. Data Visualization: Data scientists must be skilled in data visualization to communicate insights effectively to stakeholders and decision-makers.
5. Business Acumen: Data scientists need a strong understanding of business objectives, operations, and strategies to identify opportunities for data-driven insights and recommendations.
6. Communication Skills: Data scientists must be able to communicate complex technical concepts and analytical results to both technical and non-technical stakeholders.
7. Curiosity and Creativity: Data scientists must have a natural curiosity and creative problem-solving skills to explore and experiment with new approaches to data analysis and modeling.
8. Attention to Detail: Data scientists must have strong attention to detail to ensure the accuracy and reliability of their analyses and models.
9. Teamwork and Collaboration: Data scientists must be able to work effectively in cross-functional teams with other data scientists, data engineers, and business stakeholders to achieve common goals.

In summary, a successful data scientist needs a combination of technical, analytical, communication, and business skills, as well as personal characteristics such as curiosity, creativity, attention to detail, teamwork, and collaboration.

5. What are the key Roles and stakeholders for a successful analytics project?

Ans: The key roles and stakeholders for a successful analytics project include:

1. Project Sponsor: The project sponsor is typically a senior executive who provides support, resources, and guidance to the project team.
2. Project Manager: The project manager is responsible for overseeing the project's planning, execution, and delivery, and for managing the project team.
3. Data Analysts and Data Scientists: Data analysts and data scientists are responsible for data collection, cleaning, preparation, analysis, modeling, and visualization.
4. Data Engineers: Data engineers are responsible for designing, building, and maintaining the data infrastructure and pipelines that support the analytics project.
5. Business Stakeholders: Business stakeholders include department heads, managers, and analysts who provide input on the project's goals, requirements, and objectives, and who use the insights generated by the project to make decisions.

6. IT and Infrastructure Teams: IT and infrastructure teams are responsible for providing the necessary hardware, software, and networking infrastructure to support the analytics project.

6 In which phase would the team expect to invest most of the project time? Why? Where would the team expect to spend the least time?
Ans:

- The team would typically invest most of the project time in the execution phase. This is because the execution phase involves collecting, cleaning, preparing, analyzing, modeling, and visualizing data, which are typically the most time-consuming tasks in an analytics project. During the execution phase, the team will also refine and adjust their approach based on feedback and results, which can require additional time and effort.
- The team would expect to spend the least amount of time in the planning phase. This is because the planning phase typically involves establishing the project's goals, objectives, requirements, and timelines, which can be completed relatively quickly. However, it is important not to overlook this phase, as it sets the foundation for the rest of the project and can help prevent costly delays and mistakes later on.

7. What kinds of tools would be used in the following phases, and for which kinds of use scenarios? a. Phase 2: Data preparation b. Phase 4: Model building
Ans:
a. Phase 2: Data preparation
Data preparation involves collecting, cleaning, transforming, and integrating data from various sources into a format suitable for analysis. Some of the tools used in this phase include:

1. ETL Tools: ETL (Extract, Transform, Load) tools are used to extract data from various sources, transform it into a consistent format, and load it into a target database or data warehouse. Examples of ETL tools include Apache Nifi, Talend, and Pentaho.
2. Data Cleaning and Quality Tools: Data cleaning and quality tools are used to identify and fix data quality issues such as missing values, duplicates, inconsistencies, and errors. Examples of data cleaning and quality tools include OpenRefine, Trifacta, and Talend Data Quality.
3. Data Integration Tools: Data integration tools are used to integrate data from multiple sources into a single dataset. Examples of data integration tools include Apache NiFi, Talend, and Pentaho.
4. Data Profiling Tools: Data profiling tools are used to analyze and understand the characteristics and quality of the data. Examples of data profiling tools include Trifacta, Talend, and Apache Nifi.

b. Phase 4: Model building
Model building involves developing and testing predictive models using statistical and machine learning algorithms. Some of the tools used in this phase include:

1. Statistical Analysis Tools: Statistical analysis tools are used to conduct exploratory data analysis and statistical inference, and to build descriptive and predictive models. Examples of statistical analysis tools include R, SAS, and SPSS.
2. Machine Learning Tools: Machine learning tools are used to build predictive models using algorithms such as regression, classification, clustering, and deep learning. Examples of machine learning tools include Python libraries such as scikit-learn, TensorFlow, and Keras.
3. Model Development and Deployment Tools: Model development and deployment tools are used to develop and deploy predictive models into production environments. Examples of model development and deployment tools include Azure Machine Learning, AWS SageMaker, and Google Cloud AI Platform.
4. Data Visualization Tools: Data visualization tools are used to visualize and communicate the results of the model building process. Examples of data visualization tools include Tableau, Power BI, and ggplot2 in R.

8. What are the activities carried out in each phase in the Life cycle of data analytics.

Ans:

The life cycle of data analytics is a process that includes several phases that are followed to extract meaningful insights and value from data. Here are the activities carried out in each phase:

1. Requirements Gathering: This phase involves understanding the business problem, defining the objectives and goals, and identifying the data sources that are required to achieve those goals. The activities in this phase include interviewing stakeholders, reviewing existing data, and identifying the key performance indicators (KPIs).
2. Data Preparation: This phase involves collecting, cleaning, and transforming the data so that it is ready for analysis. This involves activities like data cleaning, data transformation, and data integration. The goal of this phase is to ensure that the data is accurate, complete, and relevant.
3. Data Analysis: In this phase, the data is analyzed using statistical and machine learning techniques to identify patterns, relationships, and trends. The activities in this phase include exploratory data analysis, hypothesis testing, regression analysis, and clustering.
4. Data Visualization: Once the insights are obtained from the data, they are visualized using various charts, graphs, and other visualizations. The goal of this phase is to communicate the insights effectively to the stakeholders.
5. Deployment: This phase involves implementing the insights obtained from the data analysis into the business process. This involves activities like creating dashboards, reports, and implementing automation.
6. Monitoring: In this phase, the data is continuously monitored to ensure that the insights obtained are still relevant and accurate. The goal of this phase is to ensure that the insights are up-to-date and useful for making business decisions.

These phases in the life cycle of data analytics are iterative and can be repeated multiple times until the desired insights are obtained.

***(skip) 9. Consider an example of a retail store chain that wants to optimize its products' prices to boost its revenue. The store chain has thousands of products over hundreds of outlets, making it a highly complex scenario. Once you identify the store chain's objective, you find the data you need, prepare it, and go through the Data Analytics lifecycle process.There are different types of customers, such as ordinary customers and customers like contractors who buy in bulk. a. How would you apply the data analytics life cycle to this problem?**

Ans:

To apply the data analytics lifecycle to the retail store chain's objective of optimizing its product prices to boost revenue, we would follow the following steps:

1. Business Understanding: The first step is to understand the business problem and define the objectives. In this case, the objective is to optimize product prices to increase revenue. We would also need to consider the different types of customers and their buying behavior, such as ordinary customers and contractors who buy in bulk.
2. Data Understanding: The second step is to gather and explore the relevant data. In this case, we would need to collect data on the prices of products, sales volume, customer behavior, demographics, and other relevant factors. We would also need to assess the quality and completeness of the data and identify any data gaps or inconsistencies.
3. Data Preparation: The third step is to prepare the data for analysis. This would involve cleaning the data, transforming it into a suitable format for analysis, and integrating data from multiple sources if necessary. We may also need to apply statistical techniques to impute missing values or correct data errors.
4. Data Modeling: The fourth step is to develop and test predictive models to identify the optimal prices for different products and customer segments. We would need to select appropriate machine learning algorithms, train the models on the data, and validate their performance using metrics such as accuracy and precision.

5. Evaluation: The fifth step is to evaluate the performance of the models and determine whether they meet the business objectives. We may need to test the models on a holdout dataset or conduct A/B testing to compare the performance of the optimized prices against the current prices.
6. Deployment: The final step is to deploy the models in a production environment and monitor their performance over time. We may need to integrate the models with the store chain's pricing and inventory management systems to ensure that the optimized prices are implemented effectively.

Overall, the data analytics lifecycle process would help the retail store chain to make data-driven decisions and optimize its product prices to increase revenue while considering the different types of customers and their buying behavior.

Module 2: Regression Models Weightage: 20 Marks

1. Define the following terms related to regression analysis:-
a. Overfitting
b. Cross validation
c. R²
d. Residuals
Ans:

a. Overfitting: Overfitting occurs when a regression model is too complex and fits the training data too closely, resulting in poor generalization to new data. In other words, the model is tailored too closely to the training data and may not be able to accurately predict new data.
b. Cross-validation: Cross-validation is a technique used to assess the performance of a regression model. It involves dividing the data into multiple subsets and using each subset in turn as the validation set while the other subsets are used as the training set. This allows the model to be tested on multiple subsets of the data to ensure that it can accurately predict new data.
c. R^2: R^2, also known as the coefficient of determination, is a measure of how well the regression model fits the data. It ranges from 0 to 1, with higher values indicating a better fit. R^2 represents the proportion of the variance in the dependent variable that is explained by the independent variable(s).
d. Residuals: Residuals are the differences between the actual values and the predicted values of the dependent variable in a regression model. In other words, they represent the error of the model. The goal of regression analysis is to minimize the residuals, as smaller residuals indicate a better fit of the model to the data.

2 . Explain Logit/log-odds function in detail?
Ans:

- The logit, or log-odds, function is a mathematical function used in logistic regression analysis to model the probability of a binary outcome. The logit function is defined as the natural logarithm of the odds of the event occurring, where the odds are the ratio of the probability of the event occurring to the probability of the event not occurring.
- The formula for the logit function is: **logit(p) = ln(p / (1-p))** where p is the probability of the event occurring.
- The logit function maps probabilities ranging from 0 to 1 to values ranging from negative infinity to positive infinity. The logit function is useful in logistic regression because it allows the linear regression model to be transformed into a form that can model probabilities between 0 and 1.
- In logistic regression, the logit function is used to model the probability of the dependent variable taking on the value of 1, given the independent variables. The logit function is then transformed back into a probability using the logistic function, which is defined as:
  p = 1 / (1 + exp(-logit(p))) where exp is the exponential function.

- The logistic regression model is used to estimate the coefficients of the independent variables that best predict the probability of the dependent variable taking on the value of 1. The logit function is then used to transform the linear combination of the independent variables into a probability, which can be interpreted as the predicted probability of the event occurring.

3 For the customer service data, the proportion of customers who would recommend the service in the sample of customers is p hat= 0.84, so calculate the proportion of customers who would not recommend the service department?

Ans: If the proportion of customers who would recommend the service department in the sample is p hat= 0.84, then the proportion of customers who would not recommend the service department can be calculated as follows:

Proportion of customers who would not recommend = 1 - p hat Proportion of customers who would not recommend = 1 - 0.84 Proportion of customers who would not recommend = 0.16

Therefore, the proportion of customers who would not recommend the service department in the sample is 0.16 or 16%.

4 What is the difference between Linear Regression and Logistic Regression?

Ans:

Linear Regression	**Logistic Regression**
Dependent variable is continuous	Dependent variable is binary
Independent variables can be continuous or categorical	Independent variables can be continuous or categorical
Predicts a value of the dependent variable	Predicts the probability of the dependent variable taking on the value of 1
Uses a linear equation to model the relationship between independent and dependent variables	Uses the logit function to model the relationship between independent variables and the probability of the dependent variable taking on the value of 1
Goal is to minimize the sum of squared errors	Goal is to find coefficients that maximize the likelihood of observing the actual values of the dependent variable given the independent variables
Example: predicting housing prices based on features like square footage, number of bedrooms, etc.	Example: predicting whether a customer will buy a product based on demographic and behavioral data

5. What is the difference between Linear Regression and Multiple Regression?

Ans:

Linear Regression	**Multiple Regression**
Simplest form of regression analysis, with one independent variable	Includes multiple independent variables
Dependent variable is continuous	Dependent variable is continuous
Independent variable is continuous	Independent variables can be continuous or categorical

Linear Regression	**Multiple Regression**
Predicts a value of the dependent variable	Predicts a value of the dependent variable based on multiple independent variables
Uses a linear equation to model the relationship between independent and dependent variables	Uses a linear equation to model the relationship between multiple independent variables and the dependent variable
Goal is to minimize the sum of squared errors	Goal is to find the coefficients that best predict the dependent variable based on the independent variables
Example: predicting housing prices based on square footage	Example: predicting housing prices based on square footage, number of bedrooms, location, etc.

6. Describe how logistic regression can be used as a classifier
Ans:

- Logistic regression can be used as a classifier by using a threshold probability value to predict the binary outcome of a dependent variable.
- In logistic regression, the dependent variable is the probability of an event occurring. The output of logistic regression is a value between 0 and 1, which represents the probability of the event occurring. To use logistic regression as a classifier, we need to set a threshold probability value (e.g., 0.5), above which we predict a positive outcome (i.e., a value of 1), and below which we predict a negative outcome (i.e., a value of 0).
- For example, suppose we have a dataset of customers who either buy or do not buy a product. We can use logistic regression to model the probability of a customer buying a product based on demographic and behavioral data. Once we have trained the logistic regression model, we can use it as a classifier to predict whether a new customer will buy the product or not. If the predicted probability is greater than the threshold value (e.g., 0.5), we predict that the customer will buy the product. Otherwise, we predict that the customer will not buy the product.

7. Discuss how the ROC curve can be used to determine an appropriate threshold value for a classifier.
Ans:
The Receiver Operating Characteristic (ROC) curve can be used to determine an appropriate threshold value for a classifier by evaluating the trade-off between the true positive rate (TPR) and false positive rate (FPR) at different threshold values.
The ROC curve plots the TPR against the FPR for different threshold values, providing a visual representation of the classifier's performance at different thresholds. An ideal classifier would have a TPR of 1 and an FPR of 0, resulting in a point in the top-left corner of the ROC curve.
The optimal threshold value for the classifier is the point on the ROC curve that is closest to the top-left corner. This point maximizes the difference between TPR and FPR and minimizes the classification error rate. The optimal threshold value can be selected based on the specific requirements of the problem. For example, if minimizing false positives is more important than maximizing true positives, a threshold value with a lower FPR may be preferred.

8 . If the probability of an event occurring is 0.4, then
a. What is the odds ratio?
b. What is the log odds ratio?
Ans:
a. If the probability of an event occurring is 0.4, **then the odds ratio is 0.4/(1-0.4) = 0.67.**

b. The log odds ratio is the natural logarithm (ln) of the odds ratio. Therefore, **the log odds ratio for an event with a probability of 0.4 is ln(0.67) = -0.40.**

9. Find Linear regression equation for the following data
x 2 3 5 8 y 3 6 5 12
Solution: To find the linear regression equation for the given data, we need to follow these steps:

Step 1: Calculate the means of x and y:
$\bar{x} = (2+3+5+8)/4 = 4.5$
$\bar{y} = (3+6+5+12)/4 = 6.5$

Step 2: Calculate the deviations from the means for both x and y:
$x_i - \bar{x}$: (-2.5, -1.5, 0.5, 3.5)
$y_i - \bar{y}$: (-3.5, -0.5, -1.5, 5.5)

Step 3: Calculate the sum of squares of the deviations of x:
$\Sigma(x_i - \bar{x})^2 = (-2.5)^2 + (-1.5)^2 + (0.5)^2 + (3.5)^2 = 25$

Step 4: Calculate the sum of the product of deviations of x and y:
$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = (-2.5)(-3.5) + (-1.5)(-0.5) + (0.5)(-1.5) + (3.5)(5.5) = 31$

Step 5: Calculate the slope (β1):
$\beta 1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2 = 31/25 = 1.24$

Step 6: Calculate the y-intercept (β0):
$\beta 0 = \bar{y} - \beta 1\bar{x} = 6.5 - 1.24(4.5) = 0.1$

Therefore, the linear regression equation for the given data is:
$y = 1.24x + 0.1$

10. Find Linear regression equation for the following data
x 2 4 6 8
y 3 7 5 10
Solution:
To find the linear regression equation for the given data, we need to follow these steps:

Step 1: Calculate the means of x and y:
$\bar{x} = (2+4+6+8)/4 = 5$
$\bar{y} = (3+7+5+10)/4 = 6.25$

Step 2: Calculate the deviations from the means for both x and y:
$x_i - \bar{x}$: (-3, -1, 1, 3)
$y_i - \bar{y}$: (-3.25, 0.75, -1.25, 3.75)

Step 3: Calculate the sum of squares of the deviations of x:
$\Sigma(x_i - \bar{x})^2 = (-3)^2 + (-1)^2 + (1)^2 + (3)^2 = 20$

Step 4: Calculate the sum of the product of deviations of x and y:
$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = (-3)(-3.25) + (-1)(0.75) + (1)(-1.25) + (3)(3.75) = 29.5$

Step 5: Calculate the slope ($\beta1$):
$\beta1 = \Sigma(xi - \bar{x})(yi - \bar{y}) / \Sigma(xi - \bar{x})^2 = 29.5/20 = 1.475$

Step 6: Calculate the y-intercept ($\beta0$):
$\beta0 = \bar{y} - \beta1\bar{x} = 6.25 - 1.475(5) = -0.5$

Therefore, the linear regression equation for the given data is:
$y = 1.475x - 0.5$

11. Find multiple regression equation for the following data
x1 60 62 67 70 71 72 75 78
x2 22 25 24 20 15 14 14 11
y 140 155 159 179 192 200 212 215
Solution:
To find the multiple regression equation for the given data, we need to follow these steps:

Step 1: Create a matrix of the predictor variables (x1 and x2) and the response variable (y):

x1	x2	y
60	22	140
62	25	155
67	24	159
70	20	179
71	15	192
72	14	200
75	14	212
78	11	215

Step 2: Calculate the means of x1, x2, and y:
$\bar{x}1 = (60+62+67+70+71+72+75+78)/8 = 70.5$
$\bar{x}2 = (22+25+24+20+15+14+14+11)/8 = 18.5$
$\bar{y} = (140+155+159+179+192+200+212+215)/8 = 182.375$

Step 3: Calculate the deviations from the means for x1, x2, and y:
$xi - \bar{x}1$: (-10.5, -8.5, -3.5, -0.5, 0.5, 1.5, 4.5, 7.5)
$xi - \bar{x}2$: (3.5, 6.5, 5.5, 1.5, -3.5, -4.5, -4.5, -7.5)
$yi - \bar{y}$: (-42.375, -27.375, -23.375, -3.375, 9.625, 17.625, 29.625, 32.625)

Step 4: Calculate the sum of squares of the deviations for x1, x2, and y:
$\Sigma(xi - \bar{x}1)^2 = (-10.5)^2 + (-8.5)^2 + (-3.5)^2 + (-0.5)^2 + (0.5)^2 + (1.5)^2 + (4.5)^2 + (7.5)^2 = 538$
$\Sigma(xi - \bar{x}2)^2 = (3.5)^2 + (6.5)^2 + (5.5)^2 + (1.5)^2 + (-3.5)^2 + (-4.5)^2 + (-4.5)^2 + (-7.5)^2 = 252$
$\Sigma(yi - \bar{y})^2 = (-42.375)^2 + (-27.375)^2 + (-23.375)^2 + (-3.375)^2 + (9.625)^2 + (17.625)^2 + (29.625)^2 + (32.625)^2 = 32856.03125$

Step 5: Calculate the sum of the product of deviations of x1 and y, and x2 and y:
$\Sigma(xi - \bar{x}1)(yi - \bar{y}) = (-10.5)(-42.375) + (-8.5)(-27.375) + (-3.5)(-23.375) + (-0.5)(-3.375) + (0.5)(9.625) + (1.5)(17.625) + (4.5)(29.625) + (7.5)(32.625)$

Module 3: Time Series Weightage: 10 Marks

1 List some applications that deal with time series data.

Ans:

Applications that deal with time series data include:

- Forecasting demand for products or services
- Predicting stock prices or market trends
- Analyzing website traffic or social media engagement
- Monitoring environmental variables such as temperature or pollution levels
- Tracking healthcare metrics such as patient outcomes or disease incidence
- Predicting crime rates or traffic congestion
- Analyzing sensor data in industrial processes or machinery
- Forecasting energy consumption or production
- Analyzing financial data such as exchange rates or commodity prices.

2 . What are the components of time series data in Box-Jenkins Methodology?

Ans:

The components of time series data in Box-Jenkins Methodology are:

Autoregressive (AR) component: This is the component that captures the influence of past values of the time series on its present value. It is denoted by "p" and represents the number of lagged values of the time series that are used to predict the present value.

Moving average (MA) component: This is the component that captures the influence of past errors or residuals of the time series on its present value. It is denoted by "q" and represents the number of lagged errors that are used to predict the present value.

Differencing (d) component: This is the component that represents the number of times the data needs to be differenced to make it stationary. It is denoted by "d" and can be used to remove any trend or seasonality that may be present in the data.

3 Define the following terms with respect to time series data:

a. Trend

b. Seasonality

c. Cyclic

d. Random

e. Stationarity

f. Differencing

Ans:

Definitions of terms related to time series data:

a. Trend: A long-term, gradual movement of the data in a particular direction over time. Can be upward or downward, and may be linear or non-linear.

b. Seasonality: Patterns that repeat themselves over fixed, regular intervals of time, often influenced by factors such as the time of year or day of the week.

c. Cyclic: Patterns that repeat themselves over longer, non-fixed intervals of time, often influenced by external factors such as the economy or business cycles.

d. Random: Unpredictable, erratic fluctuations in the data that cannot be attributed to any particular trend, seasonality, or cyclic pattern.

e. Stationarity: Statistical properties of the data, such as mean, variance, and autocorrelation, remain constant over time. A stationary time series is easier to model and forecast than a non-stationary time series.

f. Differencing: Transformation of non-stationary time series data into stationary data by taking the difference between consecutive observations, with the goal of removing any trend or seasonality that may be present. The order of differencing (d) represents the number of times the data needs to be differenced to make it stationary.

4 Justify which of the following series are stationary?
(a) Google stock price for 200 consecutive days;
(b) Daily change in the Google stock price for 200 consecutive days;
(c) Annual number of strikes in the US;
(d) Monthly sales of new one-family houses sold in the US;
(e) Annual price of a dozen eggs in the US (constant dollars);
(f) Monthly total of pigs slaughtered in Victoria, Australia;
(g) Annual total of lynx trapped in the McKenzie River district of northwest Canada;
(h) Monthly Australian beer production;
(i) Monthly Australian electricity production.
Ans:
(a) Google stock price for 200 consecutive days - Non-Stationary
(*b) Daily change in the Google stock price for 200 consecutive days - Stationary*
(c) Annual number of strikes in the US - Non-Stationary
(d) Monthly sales of new one-family houses sold in the US - Non-Stationary
(e) Annual price of a dozen eggs in the US (constant dollars) - Stationary
(f) Monthly total of pigs slaughtered in Victoria, Australia - Non-Stationary
(g) Annual total of lynx trapped in the McKenzie River district of northwest Canada - Non-Stationary
(h) Monthly Australian beer production - Non-Stationary
(i) Monthly Australian electricity production - Non-Stationary

A stationary time series is one whose statistical properties such as mean, variance, and covariance remain constant over time. For a series to be stationary, it should not have any trend or seasonality. Series (b) and (e) have no trend or seasonality, so they can be considered stationary. The remaining series have either a trend, seasonality or both, which make them non-stationary.

5. What is the significance of differencing in time series data analysis?
Ans:

- Differencing is a technique used in time series data analysis to make the data stationary. Stationarity is an important property of time series data, and it is required to model the data accurately using various time series models. Stationarity refers to the statistical properties of the data that remain constant over time, such as the mean, variance, and covariance.
- Differencing involves computing the difference between consecutive observations of a time series. This can be done once or multiple times, depending on the nature of the data. Differencing is useful because it can help remove trends, seasonality, and other forms of non-stationarity from the data.
- When a time series exhibits a trend or a seasonal pattern, it can be difficult to model it accurately. Differencing can help remove these patterns and make the data stationary, which allows for the use of various time series models such as ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average).
- In summary, differencing is an essential technique used in time series data analysis to make the data stationary, which is a crucial prerequisite for accurate modeling and forecasting.

6. Why is time series required to be stationary.?
Ans:
Stationarity is a fundamental requirement for time series data analysis because it helps to simplify and generalize the statistical properties of the data. A stationary time series is one whose statistical properties remain constant over time, such as the mean, variance, and covariance.
There are several reasons why stationarity is important in time series analysis:
1<u>) Accurate modeling:</u> Stationary time series are easier to model and forecast accurately than non-stationary time series. Time series models such as ARIMA and SARIMA assume that the data is stationary, and the models are designed to capture the statistical properties of the stationary data. If the data is non-stationary, the models may not work as intended, and the forecasts may be unreliable.

2) Generalization: Stationarity allows us to generalize the statistical properties of the data over time. If the statistical properties of a time series change over time, it can be challenging to generalize the data to new observations or periods. By ensuring stationarity, we can make better predictions and more reliable inferences about the future behavior of the data.
3) Hypothesis testing: Stationarity is a requirement for many statistical tests and hypothesis testing. These tests assume that the statistical properties of the data remain constant over time, and if the data is non-stationary, the tests may produce unreliable results.
In summary, stationarity is crucial for time series data analysis because it simplifies and generalizes the statistical properties of the data, allows for accurate modeling and forecasting, and is a requirement for many statistical tests and hypothesis testing.

7 Define the following components of time series
a. Trend b. Seasonality
Ans: a. Trend: A trend in a time series refers to the long-term, gradual movement of the data in a particular direction over time. This movement can be either upward or downward, and may be linear or non-linear. Trends are often used to identify changes or shifts in the underlying data, and can be helpful in making predictions or forecasting future values. A positive trend indicates an increase in the data over time, while a negative trend indicates a decrease in the data over time.

b. Seasonality: Seasonality in a time series refers to patterns that repeat themselves over fixed, regular intervals of time. These patterns can be influenced by factors such as the time of year, the day of the week, or even the time of day. Seasonality can be seen in many different types of data, such as sales data for a particular product or website traffic data. By identifying and understanding seasonality in a time series, analysts can make more accurate predictions about future values and plan accordingly.

8 Define the following terms related to time series
a. Stationary b. Differencing
Ans:
a. Stationary: A stationary time series is one whose statistical properties, such as mean, variance, and autocorrelation, remain constant over time. In other words, the data does not exhibit any trend or seasonality and is said to be stable over time. Stationarity is an important concept in time series analysis, as it simplifies the modeling process and enables more accurate predictions of future values.

b. Differencing: Differencing is a technique used to transform a non-stationary time series into a stationary one. It involves taking the difference between consecutive observations in the data, with the goal of removing any trend or seasonality that may be present. By performing differencing, the mean and variance of the data can be made constant over time, making it easier to model and forecast future values. Differencing can be performed at different orders, with first-order differencing involving the difference between consecutive observations, second-order differencing involving the difference between the first differences, and so on.

9 . Write the steps to develop ARIMA model using Box Jenkins Methodology.
Ans: The Box-Jenkins methodology is a widely used approach for developing ARIMA models to forecast time series data. The steps to develop an ARIMA model using the Box-Jenkins methodology are as follows:
1) Identification: In this step, the time series data is analyzed to identify any trends or patterns that may be present. This involves plotting the data, looking for trends, seasonality, and other patterns. If the data is non-stationary, differencing is performed to make it stationary.
2) Estimation: Once the data has been identified, the next step is to estimate the parameters of the ARIMA model. This involves selecting the appropriate order of the model (p, d, q) based on the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.

3) Diagnostic checking: After the model has been estimated, it is important to perform diagnostic checks to ensure that the model is a good fit for the data. This involves examining the residuals of the model to ensure that they are white noise and do not exhibit any patterns or trends.
4) Forecasting: Once the model has been validated, it can be used to forecast future values of the time series data. This involves using the estimated parameters of the model to generate forecasts for future time periods.
5) Model refinement: Finally, the model can be refined by revisiting the previous steps and adjusting the model as necessary to improve its accuracy and performance.
Overall, the Box-Jenkins methodology provides a systematic approach to developing ARIMA models that can be used to forecast time series data in a variety of contexts.

10. What are auto regressive models?
Ans:

- Autoregressive models are statistical models that use past observations of a time series variable to predict future values. These models assume that the current value of the variable is a linear function of its past values, with the coefficients of the past values called "lags".
- For example, an AR(1) model assumes that the current value of the variable is a linear combination of its previous value, with a coefficient that reflects the strength of the relationship between the two values. Similarly, an AR(p) model uses the previous p values to predict the current value. By analyzing the patterns in the data from the past, autoregressive models can provide insights into future trends and behaviors. These models are commonly used in finance, economics, and other fields to forecast time series data.

11. What is the moving average model in time series?
Ans:

- The moving average model is a statistical model that uses past errors (the difference between the predicted and actual values) to predict future values of a time series variable. Unlike the autoregressive model, which uses past values of the variable, the moving average model uses past errors to capture any patterns or trends in the data that were not captured by the previous predictions. The moving average model is denoted as MA(q), where q is the number of lagged errors used in the model.
- For example, an MA(1) model uses the previous error to predict the current value, while an MA(q) model uses the previous q errors. The moving average model is often used in conjunction with the autoregressive model to form the ARMA model, which combines both approaches to provide more accurate predictions of time series data. The moving average model is widely used in finance, economics, and other fields to forecast time series data and identify trends in the data.

Module 4: Text Analytics Weightage: 20 Marks

1 . Define the following terms:

a. Term Frequency

b. Inverse Document Frequency

c. Bag of words

d. Corpus

Ans:

a. Term Frequency: Term frequency is a numerical representation of how often a term (word) appears in a document or a collection of documents. It is usually calculated as the number of times a term appears in a document divided by the total number of terms in the same document. The resulting value represents the importance or relevance of the term in the document.

b. Inverse Document Frequency: Inverse document frequency (IDF) is a numerical representation of how important or rare a term is in a corpus of documents. It is calculated as the logarithm of the ratio of the total number of documents in the corpus to the number of documents that contain the term. The resulting value is used to weigh the importance of a term in a document or query, with less common terms being given a higher weight.

c. Bag of words: Bag of words is a simple and commonly used technique in natural language processing (NLP) that represents a text document as a collection of words, ignoring the order and structure of the sentences. It involves breaking down a document into individual words and then counting the frequency of each word in the document. The resulting collection of words and their frequency is called a bag of words.

d. Corpus: A corpus is a large collection of text documents that are used for linguistic analysis or natural language processing (NLP) tasks. The documents in a corpus may be from a variety of sources, such as books, articles, emails, or social media posts, and they may be in different formats, such as text, HTML, or PDF. A corpus is used to train or evaluate machine learning models for various NLP tasks such as text classification, sentiment analysis, or machine translation.

2 . What are the steps in text analytics?

Ans: Text analytics is the process of extracting valuable insights and information from unstructured text data. The following are the typical steps involved in text analytics:

1. Data collection: Collect the data from various sources, such as social media, news articles, blogs, customer feedback, etc.
2. Data preprocessing: Clean and preprocess the data by removing unwanted characters, stop words, and perform stemming/lemmatization.
3. Text tokenization: Break down the preprocessed text into individual words or phrases, which are called tokens.
4. Feature extraction: Convert the tokens into numerical features, such as term frequency or TF-IDF, to represent the text data in a numerical format that can be used for further analysis.
5. Data exploration: Explore the data by visualizing the word frequencies, co-occurrences, and relationships between words using techniques such as word clouds, topic modeling, and sentiment analysis.
6. Model building: Build machine learning or deep learning models to classify the text data into categories or predict outcomes based on the text features.
7. Model evaluation: Evaluate the performance of the models by measuring metrics such as accuracy, precision, recall, and F1-score.
8. Model deployment: Deploy the models in production to extract valuable insights from the text data and make data-driven decisions.

3 . What are the methods of representing text in text analytics? What are the challenges associated with them?

Ans:

There are several methods of representing text in text analytics, including:

1. Bag-of-words (BoW): In this method, each word in a document is treated as a separate entity, and the frequency of each word is used as a feature. This method is simple and easy to implement, but it doesn't take into account the context of the words and may result in a high-dimensional and sparse representation.
2. Term Frequency-Inverse Document Frequency (TF-IDF): This method measures the importance of a word in a document by taking into account its frequency in the document and inversely proportional to its frequency across all documents in the corpus. This method reduces the impact of common words and helps to identify important words, but it still suffers from the high-dimensional and sparse representation problem.
3. Word embeddings: This method represents words as dense vectors in a high-dimensional space, where the position of the word vector is learned based on its contextual relationship with other words. This method captures the semantic relationships between words and provides a dense representation, but it requires large amounts of data and computational power to train the models.

- The challenges associated with text representation include the high-dimensional and sparse nature of the data, the presence of noisy and irrelevant information, and the difficulty of capturing the semantic meaning of the text. Preprocessing techniques such as tokenization, stemming, and stop-word removal can help to reduce noise and improve the quality of the representation. Additionally, techniques such as dimensionality reduction and clustering can be used to reduce the dimensionality of the data and identify meaningful patterns in the text.

4 .What are the different transformation techniques used to represent raw data?

Ans:

Transformation techniques are used to modify the original data to make it more suitable for analysis. Some of the commonly used transformation techniques are:
1. Scaling: Scaling is used to standardize the data. This is done to ensure that all the variables are on the same scale, which helps in comparison and analysis. Scaling can be done using methods such as z-score normalization and min-max normalization.
2. Log Transformation: Log transformation is used to reduce the effect of outliers and to make the data more symmetrical. It is commonly used when the data is skewed.
3. Box-Cox Transformation: Box-Cox transformation is used to transform non-normal data into normal data. It is a power transformation that works by raising the data to a power.
4. PCA (Principal Component Analysis): PCA is used to reduce the number of variables in a dataset. It is used to identify patterns and relationships between variables.
5. Binning: Binning is the process of grouping data into bins or categories. It is commonly used to reduce the noise in the data.

5. The term frequency matrix given in the table below shows the frequency of terms per document. Calculate the TF-IDF value of Terms T1, T2, T3, T4, T5, T6 in Document D1

Document / Term	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
D1	**5**	**9**	**4**	**0**	**5**	**6**
D2	**0**	**8**	**5**	**3**	**10**	**8**

D3	**3**	**5**	**6**	**6**	**5**	**0**
D4	**4**	**6**	**7**	**8**	**4**	**4**

Solution:

To calculate the TF-IDF value of terms in Document D1, we need to follow these steps:

1. Calculate the Term Frequency (TF) of each term in Document D1 by dividing the frequency of the term in Document D1 by the total number of terms in Document D1.

TF(T1, D1) = 0/24 = 0

TF(T2, D1) = 5/24 = 0.2083

TF(T3, D1) = 9/24 = 0.375

TF(T4, D1) = 4/24 = 0.1667

TF(T5, D1) = 0/24 = 0

TF(T6, D1) = 6/24 = 0.25

2. Calculate the Inverse Document Frequency (IDF) of each term by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that ratio.

IDF(T1) = log(4/1) = 1.3863

IDF(T2) = log(4/1) = 1.3863

IDF(T3) = log(4/1) = 1.3863

IDF(T4) = log(4/1) = 1.3863

IDF(T5) = log(4/2) = 0.6931

IDF(T6) = log(4/2) = 0.6931

3. Calculate the TF-IDF value of each term in Document D1 by multiplying the TF and IDF values.

TF-IDF(T1, D1) = 0 * 1.3863 = 0

TF-IDF(T2, D1) = 0.2083 * 1.3863 = 0.2886

TF-IDF(T3, D1) = 0.375 * 1.3863 = 0.5192

TF-IDF(T4, D1) = 0.1667 * 1.3863 = 0.2311

TF-IDF(T5, D1) = 0 * 0.6931 = 0

TF-IDF(T6, D1) = 0.25 * 0.6931 = 0.1733

Therefore, the TF-IDF values of Terms T1, T2, T3, T4, T5, T6 in Document D1 are 0, 0.2886, 0.5192, 0.2311, 0, and 0.1733, respectively.

6. Give three benefits of using the TF IDF.

Ans:

The three benefits of using the TF IDF (Term Frequency-Inverse Document Frequency) are:

1. Better document similarity and relevance ranking: TF IDF provides a more accurate way to measure the importance of a term in a document, which leads to better ranking of documents based on relevance.
2. Reduced impact of common words: Common words like "the," "and," and "a" appear frequently in documents but do not provide much information about the topic. TF IDF reduces the impact of such common words by giving them a lower score.
3. Better feature selection: TF IDF can be used to select the most important features (terms) from a large set of data. This helps in reducing the dimensionality of the data and improving the efficiency of the analysis.

7 . What methods can be used for sentiment analysis?

Ans: Sentiment analysis is the process of identifying and extracting subjective information from text, such as opinions and emotions. There are several methods that can be used for sentiment analysis, including:

1. Lexicon-based methods: In this method, sentiment scores are assigned to words based on their dictionary meanings. The sentiment scores are then aggregated to produce an overall sentiment score for a piece of text.
2. Machine learning-based methods: In this method, a machine learning algorithm is trained on a labeled dataset to predict the sentiment of a given piece of text. The algorithm can then be used to classify new text as positive, negative, or neutral.
3. Hybrid methods: These methods combine lexicon-based and machine learning-based methods to improve the accuracy of sentiment analysis.
4. Rule-based methods: In this method, a set of rules are defined to identify sentiment in a piece of text. The rules may be based on linguistic rules or domain-specific rules.

The choice of method depends on the specific requirements of the analysis, the quality and quantity of data, and the available resources.

8 . What is the definition of topic in topic models?

Ans: In topic modeling, a topic is a group of words that are commonly used together in a specific context or theme. It is a set of related words that tend to co-occur in a text document. The goal of topic modeling is to identify these latent topics in a collection of documents, and to assign each document a distribution over these topics, indicating the extent to which each topic is present in that document. Topics can be interpreted as themes, subjects, or concepts that are discussed in the collection of documents. Topic models can be used to extract meaningful insights from large volumes of text data and to understand the underlying structure and patterns in the data

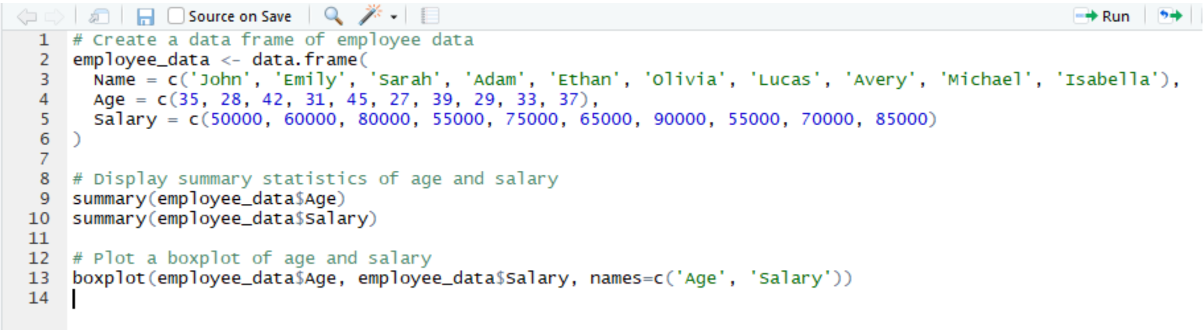
9. Explain precision and recall.

Ans: In data analytics, precision and recall are two metrics used to evaluate the performance of a classification or information retrieval model.

1. Precision refers to the ratio of correctly predicted positive cases to the total predicted positive cases. In other words, it measures how many of the predicted positive cases are actually true positive cases. A high precision indicates that the model is very accurate in predicting positive cases.
2. Recall, on the other hand, refers to the ratio of correctly predicted positive cases to the total actual positive cases. In other words, it measures how many of the actual positive cases are correctly identified by the model. A high recall indicates that the model is able to identify most of the positive cases in the data.

In summary, precision and recall are two important measures to evaluate the performance of a model, and they are often used together to give a more comprehensive picture of the model's effectiveness in identifying positive cases.

Q1. Create a data frame of 10 employee names, age and salary. Display the summary of salary and age. Plot a boxplot for age and salary in R.

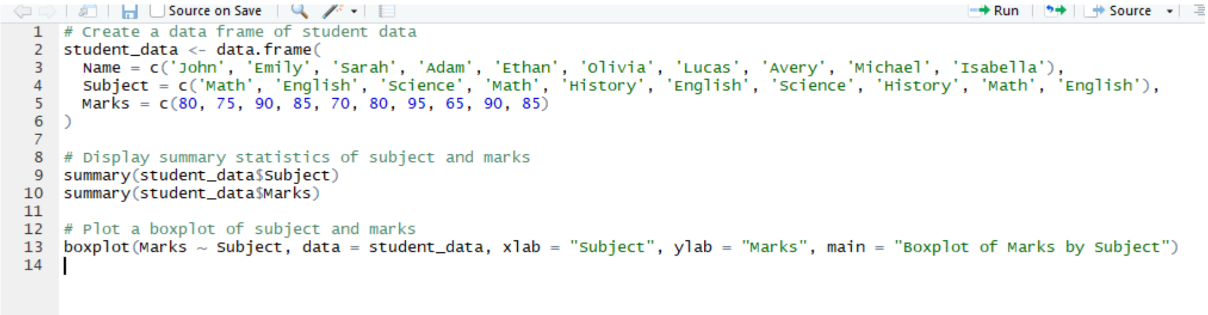


```r
# Create a data frame of employee data
employee_data <- data.frame(
  Name = c('John', 'Emily', 'Sarah', 'Adam', 'Ethan', 'Olivia', 'Lucas', 'Avery', 'Michael', 'Isabella'),
  Age = c(35, 28, 42, 31, 45, 27, 39, 29, 33, 37),
  Salary = c(50000, 60000, 80000, 55000, 75000, 65000, 90000, 55000, 70000, 85000)
)

# Display summary statistics of age and salary
summary(employee_data$Age)
summary(employee_data$Salary)

# Plot a boxplot of age and salary
boxplot(employee_data$Age, employee_data$Salary, names=c('Age', 'Salary'))
```

Q2. Create a data frame of 10 students' names, subject and marks. Display the summary of subject and marks. Plot a boxplot for subject and marks in R.

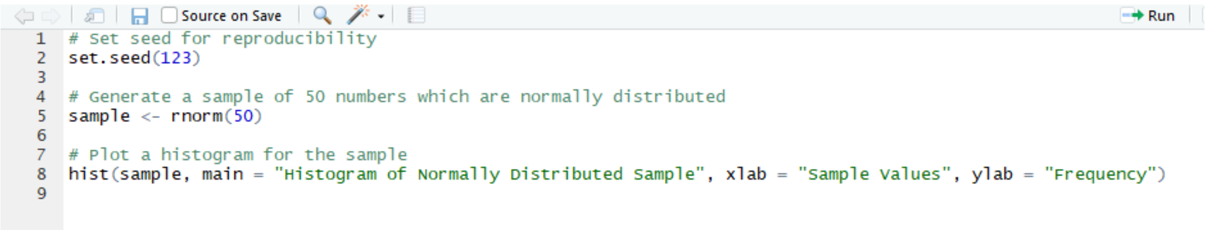


```r
# Create a data frame of student data
student_data <- data.frame(
  Name = c('John', 'Emily', 'Sarah', 'Adam', 'Ethan', 'Olivia', 'Lucas', 'Avery', 'Michael', 'Isabella'),
  Subject = c('Math', 'English', 'Science', 'Math', 'History', 'English', 'Science', 'History', 'Math', 'English'),
  Marks = c(80, 75, 90, 85, 70, 80, 95, 65, 90, 85)
)

# Display summary statistics of subject and marks
summary(student_data$Subject)
summary(student_data$Marks)

# Plot a boxplot of subject and marks
boxplot(Marks ~ Subject, data = student_data, xlab = "Subject", ylab = "Marks", main = "Boxplot of Marks by Subject")
```

Q3. Create a sample of 50 numbers which are normally distributed. Plot the histogram for this sample in R.

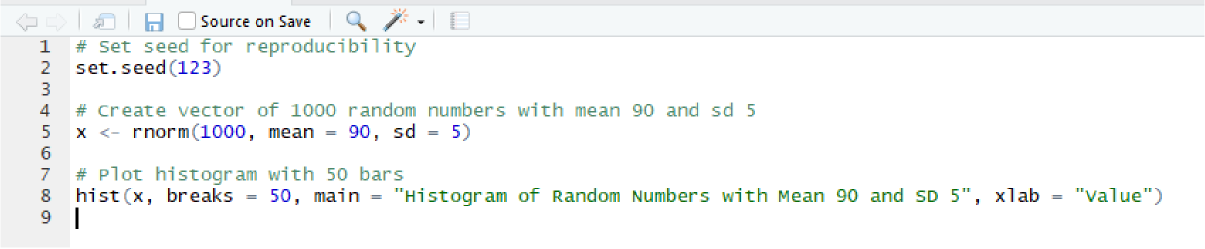


```r
# Set seed for reproducibility
set.seed(123)

# Generate a sample of 50 numbers which are normally distributed
sample <- rnorm(50)

# Plot a histogram for the sample
hist(sample, main = "Histogram of Normally Distributed Sample", xlab = "Sample Values", ylab = "Frequency")
```

Q4. Create a vector of 1000 random numbers with mean = 90 and sd=5. Create the histogram with 50 bars in R.



```r
# Set seed for reproducibility
set.seed(123)

# Create vector of 1000 random numbers with mean 90 and sd 5
x <- rnorm(1000, mean = 90, sd = 5)

# Plot histogram with 50 bars
hist(x, breaks = 50, main = "Histogram of Random Numbers with Mean 90 and SD 5", xlab = "Value")
```

Q5. If a graph of data is skewed and all the data is positive, what mathematical technique may be used in R to help detect structures that might otherwise be overlooked?

1. One mathematical technique that can be used in R to help detect structures in skewed data is a log transformation.
2. The log transformation is a mathematical process that converts the data into a logarithmic scale, making the differences between the smaller values more visible.
3. By taking the logarithm of the data, the range of values can be compressed, making it easier to identify patterns and structures.
4. This technique is especially useful for skewed data with positive values, as it can help identify patterns that may have been overlooked when looking at the original data.
5. After transforming the data, it can be plotted and analyzed to identify trends and patterns that were not easily visible in the original data.

Q6. What function can be used to fit a nonlinear line to the data in R?

- In R, the function that can be used to fit a nonlinear line to the data is nls() which stands for Nonlinear Least Squares.
- Nonlinear lines can describe complex relationships between variables that cannot be represented by simple linear models.
- The 'nls()' function fits a nonlinear model to the data by minimizing the sum of squared residuals between the observed and predicted values.
- The function requires specifying a mathematical formula that describes the relationship between the variables in the data.
- The 'nls()' function then estimates the parameters of the formula that best fit the data, using an iterative optimization algorithm.
- This technique is particularly useful when trying to model complex relationships between variables, such as exponential growth, logistic curves, or other non-linear patterns.

Q7. Suppose everyone who visits a retail website gets one promotional offer or no promotion at all. We want to see if making a promotional offer makes a difference. What statistical method would you recommend for this analysis?

A statistical method called hypothesis testing can be used for this analysis. The steps for this analysis would be:

- Formulating null and alternative hypotheses: The null hypothesis would state that the promotional offer makes no difference, while the alternative hypothesis would state that the promotional offer does make a difference.

- Collecting data: Random samples of visitors to the website should be taken, and each visitor should be randomly assigned to either the control group (no promotion) or the treatment group (promotion).

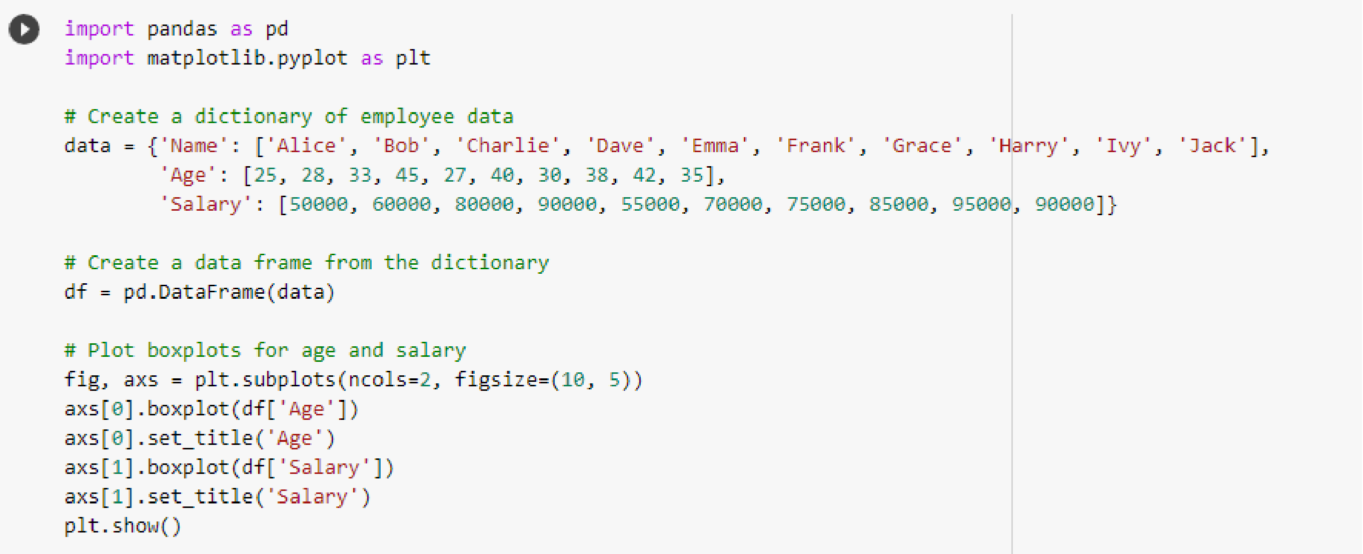
- Analysing the data: The data collected from both groups should be compared to determine if there is a significant difference between them. This can be done using statistical tests such as a t-test or chi-square test.

- Drawing conclusions: Based on the results of the statistical analysis, a decision can be made regarding whether or not the promotional offer makes a significant difference. If the null hypothesis is rejected and the alternative hypothesis is accepted, it can be concluded that the promotional offer does make a difference. If the null hypothesis is not rejected, it can be concluded that there is no evidence to suggest that the promotional offer makes a difference.

Q8. Explain exploratory data analysis in R.

- EDA is the process of examining and understanding the data before performing any statistical analysis.
- The goal of EDA is to discover patterns, relationships, and anomalies in the data that can be used to generate hypotheses and guide further analysis.
- EDA can be done through visualizations such as histograms, box plots, scatter plots, and correlation matrices, as well as summary statistics such as mean, median, and standard deviation.
- Through EDA, one can determine the distribution of the data, the presence of outliers, the relationship between variables, and the presence of missing data.
- EDA also helps in identifying potential problems such as collinearity, data entry errors, and measurement errors.
- R provides various tools for performing EDA, such as ggplot2, dplyr, and tidyr packages, which enable data manipulation, visualization, and summary statistics.
- Overall, EDA is a crucial step in the data analysis process as it helps in understanding the data better and ensuring that the analysis is based on sound assumptions.

Q1. Create a data frame of 10 employee names, age and salary. Plot a boxplot for age and salary in python.



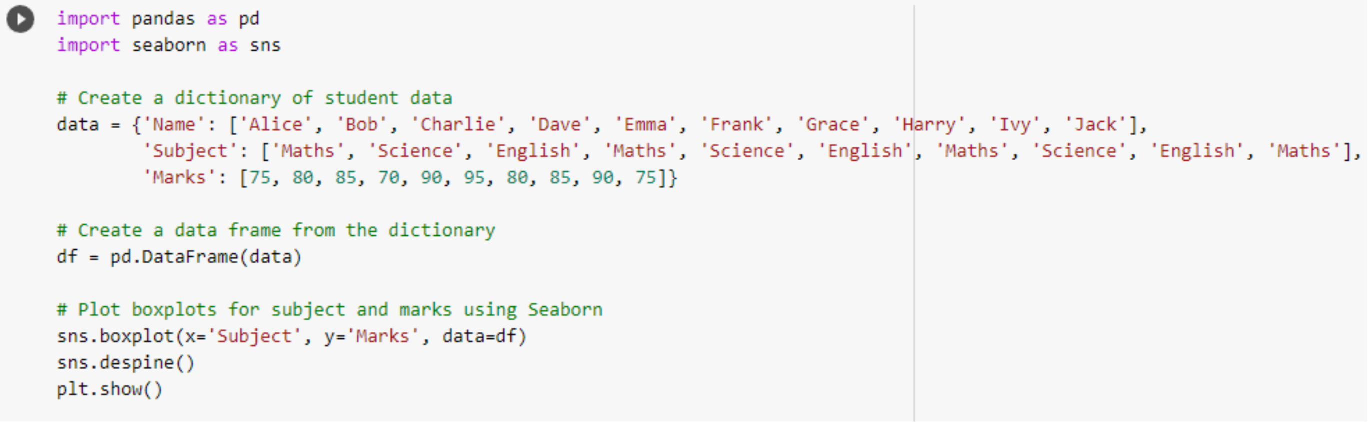
```
import pandas as pd
import matplotlib.pyplot as plt

# Create a dictionary of employee data
data = {'Name': ['Alice', 'Bob', 'Charlie', 'Dave', 'Emma', 'Frank', 'Grace', 'Harry', 'Ivy', 'Jack'],
        'Age': [25, 28, 33, 45, 27, 40, 30, 38, 42, 35],
        'Salary': [50000, 60000, 80000, 90000, 55000, 70000, 75000, 85000, 95000, 90000]}

# Create a data frame from the dictionary
df = pd.DataFrame(data)

# Plot boxplots for age and salary
fig, axs = plt.subplots(ncols=2, figsize=(10, 5))
axs[0].boxplot(df['Age'])
axs[0].set_title('Age')
axs[1].boxplot(df['Salary'])
axs[1].set_title('Salary')
plt.show()
```

Q2. Create a data frame of 10 students' names, subject and marks. Plot a boxplot for subject and marks in python.



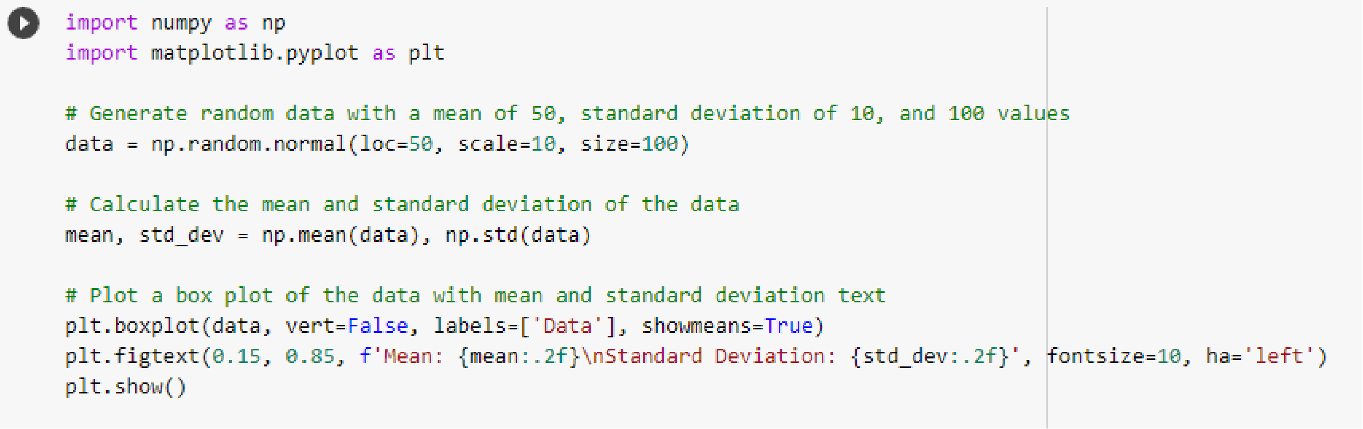
```
import pandas as pd
import seaborn as sns

# Create a dictionary of student data
data = {'Name': ['Alice', 'Bob', 'Charlie', 'Dave', 'Emma', 'Frank', 'Grace', 'Harry', 'Ivy', 'Jack'],
        'Subject': ['Maths', 'Science', 'English', 'Maths', 'Science', 'English', 'Maths', 'Science', 'English', 'Maths'],
        'Marks': [75, 80, 85, 70, 90, 95, 80, 85, 90, 75]}

# Create a data frame from the dictionary
df = pd.DataFrame(data)

# Plot boxplots for subject and marks using Seaborn
sns.boxplot(x='Subject', y='Marks', data=df)
sns.despine()
plt.show()
```

Q3. Create the box plot by using some random data, give mean, standard deviation, and the desired number of values.



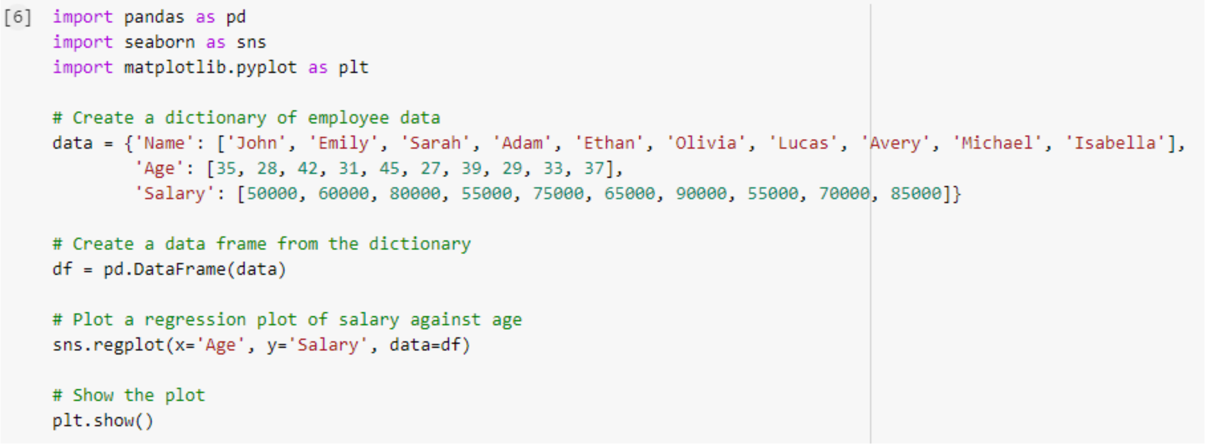
```
import numpy as np
import matplotlib.pyplot as plt

# Generate random data with a mean of 50, standard deviation of 10, and 100 values
data = np.random.normal(loc=50, scale=10, size=100)

# Calculate the mean and standard deviation of the data
mean, std_dev = np.mean(data), np.std(data)

# Plot a box plot of the data with mean and standard deviation text
plt.boxplot(data, vert=False, labels=['Data'], showmeans=True)
plt.figtext(0.15, 0.85, f'Mean: {mean:.2f}\nStandard Deviation: {std_dev:.2f}', fontsize=10, ha='left')
plt.show()
```

Q4. Create a data frame of 10 employee names, age and salary. Plot a regression plot for age and salary in python.



```
[6]  import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt

     # Create a dictionary of employee data
     data = {'Name': ['John', 'Emily', 'Sarah', 'Adam', 'Ethan', 'Olivia', 'Lucas', 'Avery', 'Michael', 'Isabella'],
             'Age': [35, 28, 42, 31, 45, 27, 39, 29, 33, 37],
             'Salary': [50000, 60000, 80000, 55000, 75000, 65000, 90000, 55000, 70000, 85000]}

     # Create a data frame from the dictionary
     df = pd.DataFrame(data)

     # Plot a regression plot of salary against age
     sns.regplot(x='Age', y='Salary', data=df)

     # Show the plot
     plt.show()
```