Question Bank

1. A city government is collecting sensor data from traffic cameras, weather stations, and public transportation systems. Explain the challenges of managing and analyzing such big data, and discuss the relevant components of the big data ecosystem (storage, processing, analytics).
2. A social media platform wants to analyse user engagement with specific hashtags. Use Python libraries (Pandas, Matplotlib) to Read data containing hashtags, user IDs, and engagement metric. Filter for specific hashtags. Impute missing engagement values or filter out rows with missing data. Group by hashtag and calculate average engagement metric (likes/post, comments/post). Sort by average engagement. Create bar charts showing average engagement for top 5 hashtags.
3. Analyze how the big data ecosystem has impacted the advancement of data science techniques. Examine how developments in data processing, storage, and analytical tools have influenced the scope and complexity of data science tasks.
4. A financial analyst wants to analyze historical stock prices. Use NumPy and Matplotlib to Load daily closing prices for multiple stocks using NumPy arrays. Calculate daily mean, standard deviation, and correlation coefficients for each stock. Create time series plots using Matplotlib to visualize how stock prices fluctuate over time.
5. You are tasked with building a model to predict housing prices. Explain different methods for retrieving data relevant to this task (public datasets, web scraping, API access). Describe the challenges of cleaning, integrating, and transforming this data for analysis.
6. A researcher wants to study the distribution of body mass index (BMI) in different age groups. Calculate the mean, median, and standard deviation of BMI for different age segments. Analyse the distribution using boxplots and investigate any potential skewness or outliers. Explore if there are differences in BMI distributions across age groups
7. A school wants to analyse student performance on various exams. Calculate the mean, median, and standard deviation of exam scores for different subjects and classes. Construct histograms and boxplots to visualize the distribution of scores. Identify any outliers and consider their impact on overall performance.
8. Implement the data science method into an actual scenario of your choice. Describe the sequential steps involved in data retrieval, data purification, data integration and transformation, data analysis, model building, and results presentation for a chosen industry or topic. Give instances of the methods and resources employed at each phase of the procedure and describe how they help to resolve the data issue in the selected field.
9. Suppose you have a dataset containing historical weather data for a city over several years. Your task is to use pandas to read and manipulate the data for analysis for following

- Load and preprocess the weather dataset.
- Visualize the average temperature trends over different seasons.
- Identify and visualize any patterns or anomalies in precipitation.
- Plot temperature distributions and highlight extreme weather events.
- Explore correlations between temperature, humidity, and other weather parameters.

10. Consider a situation in which a data scientist is employed on a project that requires them to analyze, manipulate, and visualize data. For their code, they choose to use an integrated development environment (IDE) and the core Python libraries for data scientists, which include NumPy, SciPy, Scikit-learn, Pandas, and Matplotlib. Illustrate on libraries with appropriate examples.
11. You work as a data analyst for an online retailer. You have access to a dataset that the company sent you that includes details about product orders, user interactions on the platform, and client orders. Your job is to gather information and support decision-making by performing an exploratory data analysis, or EDA.|
12. A financial analyst wants to investigate if there is a significant correlation between the daily closing prices of two stocks (Stock A and Stock B) over a period of 200 days. They collect daily closing price data for both stocks. Calculate the Pearson correlation coefficient between the two stock prices. Interpret the magnitude and direction of the correlation. Conduct a hypothesis test to determine if there is a statistically significant correlation (at a significance level of 0.05) between the two stocks. Explain your decision based on the test results. How can the analyst use this information to inform their investment decisions?
13. You are an educational researcher analysing the performance of students in a standardized test. Discuss how you would use the Frequentist Approach to measure variability in test scores and conduct hypothesis testing Explain the concept of point estimates and their relevance in educational assessment. Describe the process of calculating confidence intervals for test scores. Discuss the interpretation of a 90% confidence interval for test scores. Explain how you would use confidence intervals to compare the performance of different student groups. Discuss the application of p-values in hypothesis testing to identify significant differences in test scores.
14. A medical research team wants to develop a system to assist doctors in diagnosing diseases based on medical images like X-rays or MRIs. Data: You have a large dataset of medical images labelled with specific diseases and healthy controls. Explain why kNN might not be the best choice for this task. Discuss how alternative approaches like deep learning techniques can be utilized for image classification in medical diagnosis. Briefly describe the concept of regularization and its importance in preventing overfitting when training models with complex data like medical images.

15. You are analyzing public sentiment towards a new product launch on social media. You have collected thousands of tweets mentioning the product. Describe how you would use descriptive statistics for the following
    - What is the overall sentiment towards the product? Is it positive, negative, or neutral?
    - Are there any differences in sentiment across different demographics, such as age or location?
    - What are the most common keywords and phrases used in positive and negative tweets?
16. A researcher is conducting a hypothesis test to determine if a new drug has a significant effect on blood pressure. The null hypothesis states that the drug has no effect, and the alternative hypothesis states that the drug does have an effect. The researcher collects a sample of 30 patients, administers the drug, and records their blood pressure changes. After conducting the test, they calculate a t-statistic of -2.5 and obtain a p-value of 0.015.
    - What does the p-value of 0.015 indicate in the context of this hypothesis test?
    - If the researcher had chosen a significance level (α) of 0.05, would they reject the null hypothesis? Explain why or why not.
17. A medical researcher claims that a new drug reduces blood pressure in patients with hypertension by more than 10 mmHg on average. To test this claim, a random sample of 25 patients with hypertension is selected. After administering the drug for a period, their blood pressures are measured, and the average reduction is found to be 9.2 mmHg with a standard deviation of 3.6 mmHg. Conduct a hypothesis test at a 1% level of significance to determine if the new drug is effective in reducing blood pressure.
18. A healthcare organization is conducting a study to predict the likelihood of patient readmission within 30 days of discharge. They aim to use a machine learning model to identify patients at high risk of readmission, allowing for proactive intervention and improved patient care. The organization decides to employ the CART (Classification and Regression Tree) algorithm to build a predictive model based on patient characteristics and medical history.
19. A company wants to develop a spam email filter to automatically classify incoming emails as either "spam" or "non-spam." They decide to utilize the kNN classifier algorithm for this task. The company has a labeled dataset of 10,000 emails, where each email is represented by a set of features such as the presence of certain keywords, the number of links, and the length of the email. The team plans to use the kNN classifier to classify new emails as they arrive.
20. A news website wants to predict daily website traffic based on historical data and external factors. You have a dataset containing information on daily website traffic, along with

external factors like weather conditions, trending topics, and competing news event. Discuss the potential challenges of using linear regression for predicting website traffic, which can be influenced by various complex factors. Explain how decision trees (CART) can be helpful in this scenario. Describe their ability to capture non-linear relationships and complex decision

21. A digital marketing team is keen on understanding the engagement patterns of visitors on their website. The team tracks the time spent on the website (in minutes) for a sample of 20 visitors. This data provides valuable insights into user behavior and helps the team optimize the website's content and features for better user engagement.

    Time Spent (minutes): [8, 12, 5, 15, 10, 20, 7, 18, 25, 8, 12, 22, 15, 10, 30, 7, 18, 15, 20, 12]

    - Calculate the median of the time spent on the website. The median provides a measure of central tendency and represents the middle value in the dataset.
    - Calculate the interquartile range (IQR) to understand the spread of the data. IQR measures the range of values within which the central 50% of the data falls.
    - Create a box plot to visually represent the distribution of time spent on the website. The box plot displays the median, quartiles, and potential outliers.

22. You are analyzing public sentiment towards a new product launch on social media. You have collected thousands of tweets mentioning the product. Describe how you would use descriptive statistics for the following
    - What is the overall sentiment towards the product? Is it positive, negative, or neutral?
    - Are there any differences in sentiment across different demographics, such as age or location?
    - What are the most common keywords and phrases used in positive and negative tweets?

23. Consider a study investigating the average time spent by individuals on a specific website. A random sample of 50 users is taken, and the sample mean time spent is found to be 25 minutes, with a sample standard deviation of 4 minutes.
    - Construct a 95% confidence interval for the true average time spent on the website.
    - If the margin of error is desired to be within 1.5 minutes, what sample size would be required to achieve this at a 99% confidence level?

    Perform the necessary calculations for both parts and present the results with appropriate interpretations.

24. Consider a scenario where a manufacturing company claims that the average lifespan of their product is at least 50 hours. To test this claim, a sample of 30 products is selected, and the average lifespan is found to be 48 hours with a standard deviation of 5 hours.

Formulate the null hypothesis (H0) and the alternative hypothesis (H1) for this situation. Given a significance level (α) of 0.05, perform a hypothesis test to determine whether there is enough evidence to reject the company's claim. Perform the necessary calculations and state the conclusion based on the p-value.

25. Propose a systematic approach to identify the optimal value for K, taking into account potential challenges. Your objective is to employ the K-means clustering algorithm to categorize customers according to their purchasing patterns in an online retail dataset. Outline an analytical methodology for this task, addressing the complexities of choosing the most suitable number of clusters (K).
26. Consider a Multi-Layer Perceptron (MLP) model designed for a specific task. The network architecture consists of one hidden layer with 10 neurons and one output layer with 3 neurons. The input to the MLP is a 5-dimensional vector. Each neuron is fully connected to every neuron in the previous layer, and a bias term is included for each neuron. The activation function used is the sigmoid function.

    Calculate the total number of parameters in the hidden layer, considering weights and biases. Explain the formula used for the calculation.

    b) Similarly, calculate the total number of parameters in the output layer, considering weights and biases. Provide a detailed explanation of the computation.

    c) Discuss the significance of the activation function (sigmoid) in this context and how it influences the model's learning capacity.

27. Imagine you are analyzing sales data for a retail company that wants to identify patterns in customer purchasing behavior. You have a dataset with information on customers' average spending per visit and the frequency of their visits. The company is interested in using K-means clustering to group customers based on their spending habits. The dataset has two variables: Average Spending per Visit (in dollars) and Visits per Month. Customer A: Average Spending = $50, Visits per Month = 4

    - Customer B: Average Spending = $30, Visits per Month = 8
    - Customer C: Average Spending = $40, Visits per Month = 6
    - Customer D: Average Spending = $60, Visits per Month = 3

    Apply K-means clustering to group these customers into two clusters. Define the centroids and assign each customer to the nearest cluster

28. The distribution of body mass index (BMI) among age groups is of interest to a researcher. Determine the BMI's mean, median, and standard deviation for each age group. Use boxplots to analyze the distribution and look for any possible skewness or outliers. Examine whether the distributions of BMI vary throughout age groups.

29. A marketing manager wants to investigate if there is a significant correlation between the daily website traffic and the number of online purchases made over a period of one month. They collect daily data for both website traffic and online purchases. Calculate the Pearson correlation coefficient between the two variables. Interpret the magnitude and direction of the correlation. Conduct a hypothesis test to determine if there is a statistically significant correlation (at a significance level of 0.05) between website traffic and online purchases. Explain your decision based on the test results. How can the manager use this information to optimize marketing strategies?

30. A retail supply chain is looking to enhance its customer experience and boost sales through personalized marketing strategies. How can data analytics and machine learning techniques be leveraged to achieve these objectives? Discuss the key data sources, and data analysis techniques that can be utilized in this context