**Individual Project 9**

**DS160-02**

**Introduction to Data Science**

**Spring 2023**

**Data Science Questions (35 points)**

**Goal:** This project aims to do a basic knowledge check that we covered in this class.

**Instructions:** For this project, create a pdf script titled **IP9\_XXX.pdf**, where **XXX** are your initials. Also create a GitHub repository titled **IP9\_XXX** to which you can **push your pdf file along with the Word file.**

1. Define the term 'Data Wrangling’ in Data Analytics. Data wrangling refers to the process of cleaning and transforming raw data into a usable format for analysis. It involves various tasks such as data cleaning, data integration, data formatting, and data aggregation.
2. What are the differences between data analysis and data analytics? Data analysis involves examining, cleaning, transforming, and modeling data to extract meaningful insights from it. Data analytics, on the other hand, is a broader term that encompasses the entire process of extracting insights from data, including data collection, data cleaning, data analysis, and data visualization.
3. What are the differences between machine learning and data science? Machine learning is a subset of data science that involves training algorithms to learn patterns and make predictions from data. Data science involves a broader range of activities, including data cleaning, data analysis, data visualization, and communication of insights.
4. What are the various steps involved in any analytics project? The various steps involved in any analytics project include defining the problem, collecting, and cleaning the data, exploring and analyzing the data, building and testing the model, communicating insights, and monitoring and updating the model.
5. What are the common problems that data analysts encounter during analysis? Some common problems that data analysts encounter during analysis include missing or incomplete data, outliers, data formatting issues, and data quality issues.
6. Which technical tools have you used for analysis and presentation purposes? For analysis purposed we have used sql, python and R. For presentation purposed we have used some tableu
7. What is the significance of Exploratory Data Analysis (EDA)? Exploratory Data Analysis (EDA) is a crucial step in data analysis as it helps in understanding the data's distribution, patterns, relationships, and outliers. It also helps in identifying data quality issues and selecting appropriate statistical techniques for analysis.
8. What are the different methods of data collection? The different methods of data collection include surveys, interviews, focus groups, observation, and experiments.
9. Explain descriptive, predictive, and prescriptive analytics. Descriptive analytics involves summarizing and describing past events or data. Predictive analytics involves forecasting future outcomes based on historical data and patterns. Prescriptive analytics involves recommending actions based on predictive models and optimization algorithms.
10. How can you handle missing values in a dataset? Some common methods for handling missing values in a dataset include deleting the rows with missing values, imputing the missing values with mean or median, and using machine learning algorithms to predict the missing values.
11. Explain the term Normal Distribution. Normal distribution is a bell-shaped probability distribution that is symmetric around the mean. In a normal distribution, 68% of the data falls within one standard deviation of the mean, and 95% falls within two standard deviations of the mean.
12. How do you treat outliers in a dataset? Some common methods for treating outliers in a dataset include removing the outliers, transforming the data, or using statistical techniques that are less sensitive to outliers.
13. What are the different types of Hypothesis testing? Some different types of hypothesis testing include one-sample t-test, two-sample t-test, paired t-test, ANOVA, chi-square test, and correlation analysis.
14. Explain the Type I and Type II errors in Statistics? Type I error occurs when the null hypothesis is rejected, but it is actually true. Type II error occurs when the null hypothesis is not rejected, but it is actually false.
15. Explain univariate, bivariate, and multivariate analysis. Univariate analysis involves analyzing one variable at a time. Bivariate analysis involves analyzing the relationship between two variables. Multivariate analysis involves analyzing the relationship between multiple variables.
16. What is Data Visualization and why is it important in data analytics? Data visualization is the process of representing data visually, using charts, graphs, and other visual aids to make data more easily understandable and accessible. It is important in data analytics because it allows analysts to identify patterns, trends, and outliers quickly and easily in large data sets, and to communicate their findings to others in a clear and concise way.
17. What are Scatterplots? Scatterplots are graphs that use dots to represent individual data points, with one variable plotted on the x-axis and another variable plotted on the y-axis. They are used to show the relationship between two variables and to identify any patterns or trends in the data.
18. What are histograms and bar graphs? Histograms are graphs that use bars to represent the distribution of a single variable, with the values of the variable plotted on the x-axis and the frequency of those values plotted on the y-axis. Bar graphs are similar, but they are used to compare values across different categories or groups, with each category or group represented by a separate bar.
19. How is a density plot different from histograms? A density plot is a graph that uses a smooth line to represent the distribution of a single variable, rather than bars. It is similar to a histogram in that it shows the shape of the distribution, but it provides a smoother and more continuous representation of the data.
20. What is Machine Learning? Machine Learning is a field of study that focuses on the development of algorithms and models that can learn from data and make predictions or decisions based on that data. It is a subset of Artificial Intelligence (AI) that uses statistical and mathematical techniques to analyze data and identify patterns and relationships.
21. Which central tendency measures should be used on a particular data set? The choice of central tendency measure depends on the type of data being analyzed and the research question being addressed. The mean is generally used for normally distributed data, while the median is more appropriate for skewed or non-normal data. The mode is used to identify the most common value in a data set.
22. What is the five-number summary in statistics? The five-number summary is a descriptive statistical tool that provides a quick summary of the distribution of a data set. It consists of the minimum value, the first quartile, the median, the third quartile, and the maximum value.
23. What is the difference between population and sample? A population is the entire group of individuals, objects, or events that a researcher is interested in studying, while a sample is a subset of the population that is selected for analysis. In statistical analysis, samples are used to make inferences about the population.
24. What is the Interquartile range? The interquartile range is a measure of the spread of a data set, calculated as the difference between the first and third quartiles. It is often used to identify outliers in a data set and to provide a more robust measure of variability than the range.
25. What is Linear Regression? Linear regression is a statistical method for modeling the relationship between two variables, where one variable (the dependent variable) is predicted based on the other variable (the independent variable) using a linear equation. The goal of linear regression is to find the best-fitting line through a set of data points.
26. What is correlation? Correlation is a statistical measure that indicates the strength and direction of the relationship between two variables. It is used to determine how closely related two variables are and whether they move in the same or opposite directions.
27. Distinguish between positive and negative correlations. Positive correlation means that as one variable increases, the other variable also increases, while a negative correlation means that as one variable increases, the other variable decreases. In other words, in a positive correlation, the two variables move in the same direction, while in a negative correlation, they move in opposite directions.
28. What is Range? Range is a measure of the spread of a data set, calculated as the difference between the highest and lowest values in the data set. It is a simple and straightforward measure of variability, but it can be affected by outliers and may not provide a complete picture of the distribution of the data.
29. What is the normal distribution, and explain its characteristics? The normal distribution, also known as the Gaussian distribution, is a bell-shaped probability distribution that is symmetric around the mean. It is characterized by two parameters, the mean and the standard deviation, which determine the shape, location, and spread of the distribution. The normal distribution is important in statistics because many natural phenomena, such as heights and weights, tend to follow this pattern.
30. What are the differences between the regression and classification algorithms? Regression algorithms are used to predict a continuous numerical value, while classification algorithms are used to predict a categorical value or a binary outcome. Regression algorithms include linear regression and polynomial regression, while classification algorithms include logistic regression and decision trees.
31. What is logistic regression? Logistic regression is a statistical method for analyzing a binary outcome, where the dependent variable is either 0 or 1. It is used to model the probability of the outcome based on one or more independent variables. Logistic regression is a type of classification algorithm and is widely used in machine learning and data analysis.
32. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)? RMSE and MSE are measures of the difference between predicted and actual values in a regression analysis. RMSE is calculated as the square root of the average of the squared differences between predicted and actual values, while MSE is calculated as the average of the squared differences. Both RMSE and MSE are used to evaluate the performance of a regression model and to compare different models.
33. What are the advantages of R programming? R programming is a popular language for data analysis and statistical computing, with many advantages for data scientists and researchers. Some of the advantages include a wide range of statistical and graphical capabilities, a large and active community of users and developers, and the ability to handle large and complex data sets.
34. Name a few packages used for data manipulation in R programming? Some of the popular packages used for data manipulation in R programming include dplyr, tidyr, reshape2, and data.table. These packages provide a wide range of functions for manipulating and reshaping data, such as filtering, sorting, merging, and aggregating data sets.
35. Name a few packages used for data visualization in R programming? Some of the popular packages used for data visualization in R programming include ggplot2, lattice, plotly, and ggvis. These packages provide a wide range of functions for creating and customizing a variety of visualizations, such as scatterplots, bar graphs, histograms, and heatmaps.