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.
    a. Data wrangling refers to the process of cleaning, transforming, and organizing raw data into a usable format for analysis.
2. What are the differences between data analysis and data analytics?
    a. Data analysis involves the examination of data to uncover patterns and insights, while data analytics involves using techniques and tools to analyze and interpret data to inform decision-making.
3. What are the differences between machine learning and data science?
    a. Machine learning is a subset of data science that involves using algorithms to enable machines to learn from data, while data science is a broader range of things, including data analysis, machine learning, and other related techniques.
4. What are the various steps involved in any analytics project?
    a. The steps in an analytics project typically include defining the problem, collecting and cleaning data, exploring the data through visualization and analysis, building models, and communicating results.
5. What are the common problems that data analysts encounter during analysis?
    a. Common problems faced by data analysts include dealing with missing or inaccurate data, identifying relevant variables, and selecting appropriate models.
6. Which technical tools have you used for analysis and presentation purposes?
    a. Some technical tools we used for analysis and presentation include Python, R, SQL, and Tableau.
7. What is the significance of Exploratory Data Analysis (EDA)?
    a. Exploratory Data Analysis (EDA) is a big step in data analysis that involves examining and summarizing the main features of a dataset, including identifying patterns, relationships, and outliers.
8. What are the different methods of data collection?
    a. The main methods of collecting data include surveys, experiments, observational studies and data mining.
9. Explain descriptive, predictive, and prescriptive analytics.
    a. Descriptive- involves summarizing and visualizing data

    b. Predictive – involves looking at past data and historical facts to create predictive models and make predictions about the future
    c. Prescriptive- involves suggestions of options on how to take advantage of the information gained in the analytic process

10. How can you handle missing values in a dataset?
    a. Missing values in a dataset can be handled by either imputing them with a value based on similar data points or removing them from the analysis altogether.
11. Explain the term Normal Distribution.
    a. Normal Distribution refers to a probability distribution where the data points are symmetrically distributed around the mean and follow a bell-shaped curve.
12. How do you treat outliers in a dataset?
    a. Outliers in a dataset can be treated by either removing them or transforming them to minimize their impact on the analysis.
13. What are the different types of Hypothesis testing?
    a. Types of Hypothesis testing include t-tests, ANOVA, chi-square tests, and correlation tests.
14. Explain the Type I and Type II errors in Statistics?
    a. Type I error occurs when a null hypothesis is rejected when it is actually true, while Type II error occurs when a null hypothesis is accepted when it is actually false.
15. Explain univariate, bivariate, and multivariate analysis.
    a. Univariate- focuses on examining one variable at a time
    b. Bivariate- examines the relationship between two different variables
    c. Multivariate- examines the relationship between multiple variables
16. Explain Data Visualization and its importance in data analytics?
    a. Data visualization refers to the use of graphical representations to communicate complex data patterns and insights to an audience. It is important in data analytics as it can help identify trends, patterns, and outliers that may not be apparent from raw data.
17. Explain Scatterplots.
    a. Scatterplots are graphs that display the relationship between two quantitative variables as a series of points.
18. Explain histograms and bar graphs.
    a. Histograms are graphs that display the distribution of a single quantitative variable, while bar graphs are graphs that display the relationship between a categorical variable and a quantitative variable.
19. How is a density plot different from histograms?
    a. Density plots display the distribution of a single quantitative variable as a smooth curve, unlike histograms, which use bins.
20. What is Machine Learning?
    a. It is a field of study involving teaching machines to learn from the data without explicitly programming them to do so
21. Explain which central tendency measures to be used on a particular data set?

    a. The central tendency measure to be used on a particular dataset depends on the type of data being analyzed. For example, the mean is typically used for normally distributed data, while the median is used for skewed data.

22. What is the five-number summary in statistics?
    a. The five-number summary consists of the minimum value, first quartile, median, third quartile, and maximum value of a dataset.
23. What is the difference between population and sample?
    a. Population refers to the entire group of individuals or observations being studied, while a sample is a subset of the population used to draw inferences about the population.
24. Explain the Interquartile range?
    a. The IQR contains 50% of the data and is the middle 50%, so everything between the Q1 and Q3.
25. What is linear regression?
    a. Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables, assuming a linear relationship between them.
26. What is correlation?
    a. Correlation is a statistical measure that indicates the strength and direction of the relationship between two variables.
27. Distinguish between positive and negative correlations.
    a. A positive correlation indicates that two variables move in the same direction, while a negative correlation indicates that they move in opposite directions.
28. What is Range?
    a. Range is a measure of variability that represents the difference between the maximum and minimum values in a dataset.
29. What is the normal distribution, and explain its characteristics?
    a. The normal distribution is a continuous probability distribution that is symmetric, bell-shaped, and characterized by its mean and standard deviation. It is also known as the Gaussian distribution or the bell curve.
30. What are the differences between the regression and classification algorithms?
    a. Regression algorithms are used to predict a continuous value, while classification algorithms are used to predict a categorical value or class.
31. What is logistic regression?
    a. Logistic regression is a statistical method used to model the probability of a binary response variable based on one or more predictor variables.
32. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?
    a. Root Mean Square Error (RMSE) and Mean Square Error (MSE) are both measures of the difference between predicted and actual values in a regression model, with RMSE being the square root of MSE.

33. What are the advantages of R programming?
    a. Some advantages of R programming include its extensive package library for statistical analysis and visualization, and its flexibility for data manipulation and modeling
34. Name a few packages used for data manipulation in R programming?
    a. Tidyverse, data.table and caTools
35. Name a few packages used for data visualization in R programming?
    a. Ggplot and plotly