**Reese Bottorff**

**Individual Project 5**

**DS160**

**Introduction to Data Science**

**Fall 2023**

**Data Science Questions (70 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 **IP5\_XXX.pdf**, where **XXX** are your initials. Also create a GitHub repository titled **IP5\_XXX** to which you can **push your pdf file along with the Word file.** Show your best work and keep the document for your future journey.

1. Define the term 'Data Wrangling in Data Analytics.
   1. The process of transforming data from one raw form into a desired format to improve the quality of data
2. What are the differences between data analysis and data analytics?
   1. Data analysis focuses on the past and current state of the dataset where data analytics focuses on the future improvements or changes
3. What are the differences between machine learning and data science?
   1. Data Science studies data and how to read the data, machine learning is to understand and build methods to improve performance
4. What are the various steps involved in any analytics project?
   1. Identify business questions, collect data, clean and prepare data, analyze data, visualize and communicate data
5. What are the common problems that data analysts encounter during analysis?
   1. A common problem is the lack of quality data. It makes it hard in the decision making process
6. Which technical tools have you used for analysis and presentation purposes?
   1. Python, tableau, R, excel
7. What is the significance of Exploratory Data Analysis (EDA)?
   1. It identifies general patterns of data. It is very important in the first step of data analysis
8. What are the different methods of data collection?
   1. Surveys, interviews, observations, experiments
9. Explain descriptive, predictive, and prescriptive analytics.
   1. Descriptive – tells us what has already happened. Predictive – shows us what could happen. Perspective – informs us what should happen in the future
10. How can you handle missing values in a dataset?
    1. You can delete rows or columns that have null values. Columns can also be dropped if more than half the rows have null values
11. Explain the term Normal Distribution.
    1. A probability distribution that is symmetric about the mean, showing the data near the mean is more frequent in occurrence
12. How do you treat outliers in a dataset?
    1. Deleting the outliers can help the dataset if they cause data errors and data processing errors
13. What are the different types of Hypothesis testing?
    1. Simple, complex, null, alternative, directional, logical, statistical
14. Explain the Type I and Type II errors in Statistics?
    1. I – occurs if an investigator rejects a null hypothesis that is true in the population. II – occurs if the investigator fails to reject a null hypothesis that is actually false
15. Explain univariate, bivariate, and multivariate analysis.
    1. Univariate – summarize only one variable at a time. Bivariate – compare two variables. Multivariate – compares two or more variables
16. Explain Data Visualization and its importance in data analytics?
    1. The practice of translating information into a visual context to make data easier for the brain to understand
17. Explain Scatterplots.
    1. Shoes how two continuous variables are related by putting one variable on the x and a second variable on the y
18. **Explain histograms and bar graphs.**
    1. **Bar – the graphical representation of categorical data. Histogram – the graphical representation of quantitative data**
19. **How is a density plot different from histograms?**
    1. **Density – shows the proportion of values in each range. Histogram – shows the counts of values in each range**
20. **What is Machine Learning?**
    1. **A branch of artificial intelligence and computer science which focuses on the use of data**
21. **Explain which central tendency measures to be used on a particular data set?**
    1. **Mean, median, and mode**
22. **What is the five-number summary in statistics?**
    1. **Median, lower extreme, upper extreme, lower quartile, upper quartile**
23. **What is the difference between population and sample?**
    1. **Population – refers to the entire group of individuals. Sample – refers to the group of people from which you will be collecting data**
24. **Explain the Interquartile range?**
    1. **A measure of where the middle is in a dataset**
25. **What is linear regression?**
    1. **A process of drawing a line through data in a scatter plot**
26. **What is correlation?**
    1. **Used to estimate the relationship between two characteristics**
27. **Distinguish between positive and negative correlations.**
    1. **Positive – the variables move in the same direction. Negative – means that the variables move in opposite directions**
28. **What is Range?**
    1. **The spread of your data from the lowest to the highest value**
29. **What is the normal distribution, and explain its characteristics?**
    1. **Normal distribution is perfectly symmetrical around its center. Mean median and mode are all equal**
30. **What are the differences between the regression and classification algorithms?**
    1. **Regression – helps predict a continuous quantity. Classification – predicts discrete class labels**
31. **What is logistic regression?**
    1. **Data analysis technique that uses mathematics to find the relationships between two data factors**
32. **How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?**
    1. **MSE is measured in units that are the square of the target variable. RMSE is measure in the same units as the target variable**
33. **What are the advantages of R programming?**
    1. **Platform independent, great for statistics, well suited for machine learning, open source**
34. **Name a few packages used for data manipulation in R programming?**
    1. **Dplyr, matrix, tidyr, tibble, stringr**
35. **Name a few packages used for data visualization in R programming?**
    1. **Ggvis, esquisse, plotly, shiny**