**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. The process of transforming raw data into usable forms
2. What are the differences between data analysis and data analytics?
    a. Data Analysis: Specific type of analytics that uses data to draw useful conclusions and information – often numerical/statistical in nature
    b. Data Analytics: Broad term to cover collection, cleaning, inspection to make decisions
3. What are the differences between machine learning and data science?
    a. Machine Learning: focused on tools, libraries and methods of building models that learn by themselves using data
    b. Data Science: focused on learning about data and how to gain and understand information from it through analysis
4. What are the various steps involved in any analytics project?
    a. Obtain/load the data, prepare the data for analysis by filling in values, explore the data with basic visualizations, identify relationships, form and test hypotheses, make conclusive statements regarding the relationships and hypotheses
5. What are the common problems that data analysts encounter during analysis?
    a. Incomplete data, outliers, repetitive information/data
6. Which technical tools have you used for analysis and presentation purposes?
    a. Python, Tableau, R-Studio, SQL
7. What is the significance of Exploratory Data Analysis (EDA)?
    a. EDA is the initial analysis used to discover relationships and trends. It is significant because this step is completed without bias and allows a deeper understanding of the data and to identify relationships not previously thought of.
8. What are the different methods of data collection?
    a. Observation, surveys, interviews, experiments
9. Explain descriptive, predictive, and prescriptive analytics.
    a. Descriptive Analysis: Describe and summarize the data – identify trends and relationships
    b. Predictive Analysis: Using the data to predict different things and relationships
    c. Prescriptive Analysis: Using data to find best course of action or solution

10. How can you handle missing values in a dataset?
    a. First, plot the values to visualize the distribution. Then, use the best fitting statistical measure, mean, median, etc., to fill in the missing values. The best fitting statistical measure is based upon the distribution – for an example, median works best for skewed data
11. Explain the term Normal Distribution.
    a. Normal distribution means the dataset is symmetric regarding the mean, meaning that points near the mean occur more frequently than those far from the mean – it also follows a bell curve
12. How do you treat outliers in a dataset?
    a. Outliers can skew the mean and median of the dataset so they are often dropped or disregarded in analysis – they do however matter and should still be considered
13. What are the different types of Hypothesis testing?
    a. Z Test: Tests if the relationship is statistically significant by comparing the null hypothesis – can only be used if the standard deviation of the population is known
    b. T Test: The same as a Z Test when population standard deviation is unknown
    c. Chi-Square Analysis: Used with categorical variables – compares predicted values from assuming null is true to what actually happened to gage the relationship
14. Explain the Type I and Type II errors in Statistics?
    a. Type I: False positive – saying something happened when it doesn't
    b. Type II: False negative – saying something does not happen when it actually does
15. Explain univariate, bivariate, and multivariate analysis.
    a. Univariate Analysis: Explore each variable of the dataset independent/separately
    b. Bivariate Analysis: Explore and compare two variables to see how they are related
    c. Multivariate Analysis: Explore multiple variables to discover relationships between them
16. Explain Data Visualization and its importance in data analytics?
    a. Data visualization is the use of graphs and charts to depict the data – these images help to communicate important information about relationships and distribution
17. Explain Scatterplots.
    a. Scatterplots depict the relationship between two variables with dots representing each point
18. Explain histograms and bar graphs.
    a. Histograms and bar graphs both depict the distribution of a variable. Histograms visualize continuous data while bar graphs visualize categorical data.
19. How is a density plot different from histograms?
    a. Histograms are bars stacked together to visualize distribution while density plots visualize distribution in a continuous way – a continuous line
20. What is Machine Learning?
    a. Machine learning is a branch of data science concerned with tools, libraries and methods of building models that learn by themselves using data

21. Explain which central tendency measures to be used on a particular data set?
    a. Mean, median, and mode – mean is used under normal distribution, median with skewed distribution, and mode when mean and median are not suitable/accurate
22. What is the five-number summary in statistics?
    a. Five numbers that describe the data and its distribution. It consists of the min, max, median, first quartile, and third quartile
23. What is the difference between population and sample?
    a. Population: the entire group you are drawing conclusions about
    b. Sample: small portion of the entire group you are collecting data from
24. Explain the Interquartile range?
    a. The difference between Quarter 1 and Quarter 3 – the range of the middle of the dataset
25. What is linear regression?
    a. Model applied to a dataset to predict values based off an equation
26. What is correlation?
    a. Correlation describes the relationship between two variables meaning they affect each other
27. Distinguish between positive and negative correlations.
    a. Positive Correlations: The two variables move in the same direction meaning they increase or decrease together
    b. Negative Correlations: The two variables move in opposite directions meaning one increase while the other decreases and vice versa
28. What is Range?
    a. Range is the distance between the minimum and maximum value of a dataset
29. What is the normal distribution, and explain its characteristics?
    a. Normal distribution means the data is not skewed and values near the mean occur more frequently than those not near the mean. It creates a bell curve where the center of the hump is equal distance to both ends of the curve
30. What are the differences between the regression and classification algorithms?
    a. Regression algorithms are used with continuous variables for prediction while classification algorithms work to predict categorical variables based on related qualities
31. What is logistic regression?
    a. Model that estimates the probability of the predicted event or value occurring
32. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?
    a. To find MSE find the difference between the actual and predicted values, square these difference, average them, and then sum all of the averaged values. To find RMSE, find the square root of the MSE.
33. What are the advantages of R programming?
    a. Open source programming with uses for data wrangling and machine learning. It is also compatible and independent so it can be used on any computer.
34. Name a few packages used for data manipulation in R programming?

    a. Tidyverse: consists of things like filter, mutate, and select that allow users to filter out and change data to prepare it for visualization
    b. CaTools: prepares data for plotting by splitting data for regression

35. Name a few packages used for data visualization in R programming?
    a. ggplot: plotting function of RStudio – includes things like density plots, scatter plots, and bar charts