**Braden Heuglin**

**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 is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.**

1. What are the differences between data analysis and data analytics?

**Data analysis is a process involving the collection, manipulation, and examination of data for getting a deep insight. Data analytics is taking the analyzed data and working on it in a meaningful and useful way to make well-versed business decisions.**

1. What are the differences between machine learning and data science?

**Machine learning helps make artificial intelligence — the science of making machines capable of human-like decision-making — possible. Data science is the process of developing systems that gather and analyze disparate information to uncover solutions to various business challenges and solve real-world problems.**

1. What are the various steps involved in any analytics project?

**First, you have to find and access the data you wish to work with. Next, you read that data into whatever program you are using. Then, you check the data for any missing or null values, and remove or fix any of those values. After that you have free reign to manipulate the data in any way you wish, from creating plots to just simply displaying what values the data contains.**

1. What are the common problems that data analysts encounter during analysis?

**The most common problem is missing or null values in a dataset. If you do not check the data for missing values and try to create plots with an incomplete dataset, your results are not going to be accurate or useful.**

1. Which technical tools have you used for analysis and presentation purposes?
2. What is the significance of Exploratory Data Analysis (EDA)?

**The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.**

1. What are the different methods of data collection?

**Some common data collection methods include surveys, interviews, observations, focus groups, experiments, and secondary data analysis.**

1. Explain descriptive, predictive, and prescriptive analytics.
2. How can you handle missing values in a dataset?

**As I mentioned in a previous answer, you must fill or remove any missing values you find in a dataset. If you do not, the methods you use to display your results will be inaccurate and therefore, useless.**

1. Explain the term Normal Distribution.
2. How do you treat outliers in a dataset?

**There are two ways to handle outliers. First, and the best way, is to remove the outliers so the data is more normal and useful, be sure to let people know you had to remove outliers though. Second, you can leave the outliers in but you *must* tell anyone who looks at your work which values are outliers and how they affect your findings.**

1. What are the different types of Hypothesis testing?
2. Explain the Type I and Type II errors in Statistics?
3. Explain univariate, bivariate, and multivariate analysis.
4. Explain Data Visualization and its importance in data analytics?
5. Explain Scatterplots.
6. **Explain histograms and bar graphs.**
7. **How is a density plot different from histograms?**
8. **What is Machine Learning?**
9. **Explain which central tendency measures to be used on a particular data set?**
10. **What is the five-number summary in statistics?**

**It is a series of five numbers which explain the distribution of a dataset to help in understanding the data and making decisions with it. The numbers are min, Q1, median, Q3, and max.**

1. **What is the difference between population and sample?**
2. **Explain the Interquartile range?**
3. **What is linear regression?**
4. **What is correlation?**
5. **Distinguish between positive and negative correlations.**
6. **What is Range?**
7. **What is the normal distribution, and explain its characteristics?**
8. **What are the differences between the regression and classification algorithms?**
9. **What is logistic regression?**
10. **How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?**
11. **What are the advantages of R programming?**
12. **Name a few packages used for data manipulation in R programming?**
13. **Name a few packages used for data visualization in R programming?**