**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.

Data wrangling is the preprocessing of data by cleaning, structuring, and transforming to prepare for analysis. Common tasks of data wrangling are handling missing values, dealing with outliers, converting data types, and merging datasets.

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

Data analysis: inspecting, cleaning transforming, and modeling data to discover useful information and draw conclusions informing decision-making. A smaller part of data analytics.

Data analytics: encompassing a broader scope, data analytics also focuses on deriving insights and making predictions to drive business decisions through the use of statistical and computational methods (not just through data).

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

Machine learning: a subset of data science involving the use of algorithms and statistical models to enable systems to learn and improve from experience without explicit programming.

Data Science: broader than machine learning, includes data cleaning, exploration, and interpretation, as well as using machine learning for predictive modeling and pattern discovery.

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

Problem definition, data collection, data cleaning and wrangling, exploratory data analysis (EDA), feature engineering, model building, model evaluation, deployment, and finally monitoring and maintenance. Then the process repeats either with the same problem or a different problem.

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

Missing data, outliers, data quality issues, inconsistent data, poorly defined goals, results difficult to interpret, overfitting, and underfitting.

1. Which technical tools have you used for analysis and presentation purposes?

Analysis: python (pandas, numply, scipy), r, sql

Presentation: jupyter notebook (matplotlib.pylot, seaborn), tableau, power BI

1. What is the significance of Exploratory Data Analysis (EDA)?

EDA helps in understanding the main characteristics of a dataset and identifies patterns, relationships, and potential outliers to help the process of subsequent analysis and model building to make informed decisions.

1. What are the different methods of data collection?

Surveys, experiments, observational studies, existing databases, web scraping.

1. Explain descriptive, predictive, and prescriptive analytics.

Descriptive: describes the past.

Predictive: predicts future based on past.

Prescriptive: recommends actions to achieve goals.

1. How can you handle missing values in a dataset?

Removing rows or columns, imputation (using mean, median, or mode), interpolation, machine learning-based imputation.

1. Explain the term Normal Distribution.

A symmetric, upside down bell-shaped probability distribution characterized by its mean and standard deviation.

1. How do you treat outliers in a dataset?

Removing outliers, transforming data, capping or flooring values, using robust statistical measures.

1. What are the different types of Hypothesis testing?

Parametric tests: t-tests, ANOVA, non-parametric Tests: Mann-Whitney U, Kruskal-Wallis, Chi-Square Test: for categorical data

1. Explain the Type I and Type II errors in Statistics?

Type 1 error: false positive, type 2 is false negative.

1. Explain univariate, bivariate, and multivariate analysis.

univariate explains a single variable. Bivariate analyzes the relationship between two variables. Multivariate examines three or more variables simultaneously.

1. Explain Data Visualization and its importance in data analytics?

Presenting data in graphically helps in understanding patterns, trends, and insights and enhances communication of complex information.

1. Explain Scatterplots.

A type of graphic representation of the relationship between two continuous variables using points that represent single observations.

1. **Explain histograms and bar graphs.**

**Histograms represent the distribution of a continuous variable graphically using vertical bars.**

**Bar graphs: display the distribution of a categorical variable also using vertical bars.**

1. **How is a density plot different from histograms?**

**Density plots shows the proportion of values in each range while histogram**

1. **What is Machine Learning?**
2. **Explain which central tendency measures to be used on a particular data set?**
3. **What is the five-number summary in statistics?**
4. **What is the difference between population and sample?**
5. **Explain the Interquartile range?**
6. **What is linear regression?**
7. **What is correlation?**
8. **Distinguish between positive and negative correlations.**
9. **What is Range?**
10. **What is the normal distribution, and explain its characteristics?**
11. **What are the differences between the regression and classification algorithms?**
12. **What is logistic regression?**
13. **How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?**
14. **What are the advantages of R programming?**
15. **Name a few packages used for data manipulation in R programming?**
16. **Name a few packages used for data visualization in R programming?**