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.
    a. Data wrangling is the process in which raw data is turned into usable data by going though a cleaning process
2. What are the differences between data analysis and data analytics?
    a. Data analysis is looking at the data and coming up with conclusions about it while data analytics is more about the visual side of the data and processing it.
3. What are the differences between machine learning and data science?
    a. Data science is more of the study of the data and how to make conclusions from it where machine learning is more about the background work in making programs that can be used to analyze the data.
4. What are the various steps involved in any analytics project?
    a. Gathering and preparing the data, analyzing it, interpreting the results
5. What are the common problems that data analysts encounter during analysis?
    a. Missing data/null values, bad data/not enough,
6. Which technical tools have you used for analysis and presentation purposes?
    a. Python, R, SQL,
7. What is the significance of Exploratory Data Analysis (EDA)?
    a. It allows you to get a full understanding of the data before making conclusions on it.
8. What are the different methods of data collection?
    a. Surveys, experiments, interviews
9. Explain descriptive, predictive, and prescriptive analytics.
    a. descriptive analytics, which tell us what has already happened; predictive analytics, which show us what could happen, and finally, prescriptive analytics, which inform us what should happen in the future.
10. How can you handle missing values in a dataset?
    a. Find the mean or median of the category you need and then implement it into the data.
11. Explain the term Normal Distribution.
    a. The data is symmetrical with no skew

12. How do you treat outliers in a dataset?
    a. You can delete them or replace it with the median value
13. What are the different types of Hypothesis testing?
    a. Simple and composite testing
14. Explain the Type I and Type II errors in Statistics?
    a. Type I is a false positive – false in the hypothesis and true in real life, and type II is a false negative – true in hypothesis and false in real life
15. Explain univariate, bivariate, and multivariate analysis.
    a. Univariate statistics summarize only one variable at a time. Bivariate statistics compare two variables. Multivariate statistics compare more than two variables.
16. Explain Data Visualization and its importance in data analytics?
    a. The process of turning the data into a visual aid. This allows people to get a good understanding of the data and it can provide some good insights and show trends
17. Explain Scatterplots.
    a. It's a graph made up of a bunch of individual points that are related to the x and y axis
18. Explain histograms and bar graphs.
    a. Histograms and bar graphs are very simile with their sections or bins but histograms x axis contains all numbers where bar graphs are specific numbers
19. How is a density plot different from histograms?
    a. A density plot is a continuous line while a histogram contains many bins
20. What is Machine Learning?
    a. Machine learning is the process of a computer being able to learn from itself and complete task with few instructions
21. Explain which central tendency measures to be used on a particular data set?
    a. The mean or median
22. What is the five-number summary in statistics?
    a. Min, Q1, Q2, Q3, max
23. What is the difference between population and sample?
    a. The population is the whole group of people while a sample is only a small portion of the population
24. Explain the Interquartile range?
    a. There are four equal sections and the IQR is the middle two sections. Its 50% of the data
25. What is linear regression?
    a. The correlation between x and y
26. What is correlation?
    a. How closely related two things are/the strength of their relationship
27. Distinguish between positive and negative correlations.
    a. A positive correlation has both variables moving in the same direction while a negative correlation the two variables work opposite of eachother
28. What is Range?
    a. Range is the difference between the highest and lowest value

29. What is the normal distribution, and explain its characteristics?
    a. Normal distribution is when the distribution is symmetrical about the mean
30. What are the differences between the regression and classification algorithms?
    a. Regression helps predict a continuous quantity while classification predicts discrete class labels
31. What is logistic regression?
    a. Logistic regression is a statistical method used for modeling the relationship between a categorical dependent variable and one or more independent variables
32. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?
    a. rmse = np.sqrt(np.mean((actual - predicted) ** 2))
    b. mse = np.mean((actual - predicted) ** 2)
33. What are the advantages of R programming?
    a. Its an open source program with a bunch on packages and is really good for statistical analysis
34. Name a few packages used for data manipulation in R programming?
    a. Dplyr, stringr, tidyr, data.table
35. Name a few packages used for data visualization in R programming?
    a. Ggplot2, leaflet,