Ryan Fox
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. Data wrangling is the process of cleaning or transforming raw data into a usable format.
2. What are the differences between data analysis and data analytics?
    a. Data analysis is the process of cleaning, transforming, and using mathematical/statistical tests to better understand the data.
    b. Data analytics is a much broader field that includes data analysis. This is used for large data sets including those used in business settings.
3. What are the differences between machine learning and data science?
    a. Data science involves using statistical and mathematical methods to gain an insight on the dataset. This includes data cleaning, data analysis, and data visualization.
    b. Machine Learning is a part of data science that involves using statistical models to learn from data and making predictions without having to execute a line of specific code.
4. What are the various steps involved in any analytics project?
    a. The steps involved in analytical projects are obtaining your data, exploring and cleaning your data, filling in missing or null values, using statistical processes to gain information from the data, use models to visualize the data, and make conclusions.
5. What are the common problems that data analysts encounter during analysis?
    a. Some common problems experienced by data analysts are missing/incomplete data, data quality issues, bias, or outliers.
6. Which technical tools have you used for analysis and presentation purposes?
    a. I have used Tableau, python, SQL, and R studio.
7. What is the significance of Exploratory Data Analysis (EDA)?
    a. Exploratory Data Analysis is very important in current times where data is so abundant. It helps data scientists learn more about trends in today’s society.
8. What are the different methods of data collection?

a. The different methods are: surveys, interviews, focus groups, observations, experiments, case studies, and secondary data sources.

9. Explain descriptive, predictive, and prescriptive analytics.
    a. Descriptive analysis – The first kind of data analysis prepared, usually applied to large volumes of data. The techniques involved are frequency distributions, measuring centrality, and dispersion of distributions.
    b. Predictive analysis – Predictive analysis is about the understanding of what will happen in the future by using data from the past. It uses things like trendlines to predict what will happen in the future.
    c. Prescriptive analysis – used after an event occurs. This is an area of business analytics that finds the best course of action using statistical measurements.
10. How can you handle missing values in a dataset?
    a. You can handle missing values in datasets by omitting that subject or filling in values using either the mean or median of the dataset based on bias and skew.
11. Explain the term Normal Distribution.
    a. A normal distribution is what ideal statistical data would look like. In a normal distribution, the mean and median of the dataset would be the exact middle. A normal distribution is in the shape of a bell curve.
12. How do you treat outliers in a dataset?
    a. It is best to find the reason for outliers instead of omitting them. If possible, you can transform the data to make them less of an impact on the mean.
13. What are the different types of Hypothesis testing?
    a. The different types of hypothesis testing are the T-test, Z-test, and Chi-Square. Sometimes other regression models are used.
14. Explain the Type I and Type II errors in Statistics?
    a. A type I error occurs when the null hypothesis is rejected when it is actually true. This is also known as a false positive.
    b. A type II error is when the null hypothesis is not rejected when it is actually false. This is also known as a false negative.
15. Explain univariate, bivariate, and multivariate analysis.
    a. Univariate analysis is the statistical analysis of a single variable. It describes things such as the mean, median, mode, range, and standard deviation.
    b. Bivariate analysis is the statistical analysis of two variables. This analysis compares the relationship of the two variables using things like, correlation, regression, and contingency. This is a cause-and-effect analysis.
    c. Multivariate analysis is the statistical analysis of more than two variables. This analysis compares the relationships between the multiple variables. This can be seen when using a multiple regression test.
16. Explain Data Visualization and its importance in data analytics?
    a. Data visualization is using different kinds of graphs or demonstrations to show import findings within a dataset. This is important to data analytics because it allows data scientists to show demonstrate their conclusions as well as makes it

easier to explain and communicate their findings to the general public or other data scientists.

17. Explain Scatterplots.
    a. A scatterplot is a graph that shows the comparison of two quantitative variables by plotting data points on (x,y) coordinates. Trendlines can often be used to help explain scatterplots.
18. Explain histograms and bar graphs.
    a. A histogram is a graph that represents the distribution of a dataset. This graph can either be normal, skewed left, or skewed right.
    b. A bar graph is a graph that demonstrates values by the height of a bar. The higher the bar the more data points in that grouping. This allows for easier comparison when using things like demographics.
19. How is a density plot different from histograms?
    a. Density plots show the distribution of a continuous variable. They are different from histograms because it includes a curve over histogram data and predicts the probability density function of the variable.
20. What is Machine Learning?
    a. Machine learning involves algorithms and models that allow computer systems to learn from data and perform a specific task without being explicitly programmed.
21. Explain which central tendency measures to be used on a particular data set?
    a. If a dataset is normally distributed then the measure of centrality to use is the mean.
    b. If a dataset is skewed in any direction, it is better to use the median because it represents a better center of the data.
22. What is the five-number summary in statistics?
    a. Min, Max, Q1, Median, Q3
23. What is the difference between population and sample?
    a. A sample is just a small portion of the population. If you were to sample everyone that a certain test applied to, then you would have a population. A sample can be done with a much smaller amount of a data.
24. Explain the Interquartile range?
    a. The interquartile range is the range between the first quartile and the third quartile. This means it is the 50% of data between the 25% data value and 75% data value.
25. What is linear regression?
    a. Linear regression is a statistical model that shows the relationship between a dependent variable and one or more independent variables. This comparison shows the correlation between the variables, or how much one variable effects the other. The closer the value is to 1 or -1, the stronger the correlation.
26. What is correlation?
    a. Correlation shows the strength and direction of a relationship between two variables. You can have either positive or negative correlation.
27. Distinguish between positive and negative correlations.

    a. In a negative correlation, as one variable increases, the other decreases.

28. What is Range?
    a. Range is the number value between the lowest data point and the highest data point.
29. What is the normal distribution, and explain its characteristics?
    a. A normal distribution is a dataset where the sample mean is exact or close to the true mean. There is no skew in this distribution and the data points are distributed about 68% within the first standard deviation, 95% within two standard deviations, and 99% within three standard deviations.
30. What are the differences between the regression and classification algorithms?
    a. Regression algorithms predict quantitative values, classification algorithms predict categorical values.
31. What is logistic regression?
    a. Logistic regression is like linear regression except it uses categorical values instead of quantitative values.
32. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?
    a. You find MSE by taking the sum of the squares difference between the actual and predicted values. RMSE is found by taking the square root of MSE.
33. What are the advantages of R programming?
    a. The advantages of R studio are that it is a free data software that has many statistical functions that are accessible for everyone. It also provides ease of use and many visualization tools.
34. Name a few packages used for data manipulation in R programming?
    a. Tidyverse, tidyr, reshape2
35. Name a few packages used for data visualization in R programming?
    a. Ggplot, ggmap, plotly