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. In the field of Data Analytics, Data Wrangling is the process cleaning up a dataset to make it easier to analyze. This process includes removing errors and combining complex datasets so that they become more accessible.
2. **What are the differences between data analysis and data analytics?**
   a. Data analytics is a broader term encompassing data analysis, and it refers to the general field of using data and various tools to make business decisions. Data analysis is a subset of data analytics that refers to the actual processes employed to manipulate, organize, and visualize the data under consideration.
3. **What are the differences between machine learning and data science?**
   a. Data science is a field that focuses on studying data and extracting meaning from it. Machine learning is a branch of artificial intelligence that works to give machines the human-like ability to learn and solve problems, through statistical models and algorithms. The two fields meet by data science producing the data that is used to train the machine learning algorithms.
4. **What are the various steps involved in any analytics project?**
   a. In any analytics project there are several common steps that must be taken. You must first obtain your dataset. Next you need to check for missing data. If there are missing data values, depending on the datatype you should visualize that variable to help determine the best way to fill the missing values. After filling missing values, you need to check the datatypes of the variables because in order to use certain data it needs to be a certain datatype, so if it is not the correct datatype you must change the datatype of the variable. Once all of these steps are complete the data should be ready for analysis. You can then perform various functions and create many different visualizations to obtain valuable information from the data.
5. **What are the common problems that data analysts encounter during analysis?**
   a. One of the major problems data analysts encounter during analysis is missing and/or inconsistent data. This problem is often solved by effect data cleaning, but depending on the type of data, if too many values are missing or there is inconsistent data within a particular variable, that variable may not be usable at

all, and may then be dropped from the dataset. Another problem that can be encountered, is just trying to figure out what analysis would be effective for producing useful information from the data. There are many different types of analysis that can be performed on data, but not all of it will produce useful results for your purposes. To help solve this problem, it is helpful to have goals in mind of what you want to learn and what questions you want answered. This will help direct what analysis you choose to do.

6. **Which technical tools have you used for analysis and presentation purposes?**
    a. Throughout my data analysis experience, I have used several tools to aid in my analysis and presentation of data. At the broadest level I have been able to use various programs in this process, such as Python, R, and Tableau. Withing this programs I have made use of several premade programs including but not limited to matplotlib, pyplot, pandas, and seaborn to perform various aspects of data analysis and visualization.
7. **What is the significance of Exploratory Data Analysis (EDA)?**
    a. Exploratory data analysis is significant because of its usefulness in discovering previously unknown relationships within a dataset. This type of analysis is extremely helpful when you do not have a clear question or hypothesis regarding data you need to analyze. In other words, it help provide direction, when there is no direction at the start.
8. **What are the different methods of data collection?**
    a. There are a plethora of methods that can be used to collect data. The reliability of data received from these different methods varies a lot, but each serve there purpose. Some common methods are: surveys, interviews, focus groups, general observation, experiments, and secondary data analysis.
9. **Explain descriptive, predictive, and prescriptive analytics.**
    a. Descriptive analytics is the first kind of data analysis that is performed on a dataset. Generally, it is applied to large volumes of data, and it almost always includes certain techniques. These core techniques are frequency distributions (like a histogram), the common measures of centrality (mean, median, mode), and evaluation of the dispersion of the distribution (range, interquartile range, variance, standard deviation). This type of analysis helps you get an idea of the spread, average, middle, and what value occurs the most of a given variable. It is also most applicable to numerical data, the functions of this type that can be performed on categorical data are much more limited.
    b. The purpose of predictive analytics is to take given data from the past and use it to predict what is likely to happen in the future. It involves the use of a variety of functions on past data to then extrapolate that to predict behavior in the future. This type of analysis is extremely important for all business to help increase sales and reduce costs.
    c. Prescriptive analytics is an area of business analytics that analyzes data to solve a problem. Functions are used on the given data to help determine the best course of action in a particular situation or problem. It seeks to find connections between

various parameters and variables to address a specific question often still in a predictive way.

10. **How can you handle missing values in a dataset?**
    a. For missing categorical variables your options are limited. Typically, one might fill it with what occurs most for that variable. For numerical data, if there is a normal distribution then you might use the mean to fill missing values, but if the data is skewed then it is better to use the median to fill missing values.
11. **Explain the term Normal Distribution.**
    a. In its simplest definition, a normal distribution is one whose data falls in line with a normal bell curve distribution.
12. **How do you treat outliers in a dataset?**
    a. If the outliers are so few and so abnormal that they are likely erroneous data entries then they can be eliminated from the dataset. If you think the outliers that are present are important to consider they should remain, but keep in mind how they are affecting the distribution of your data.
13. **What are the different types of Hypothesis testing?**
    a. There are a few different types of hypothesis testing. An alternative hypothesis seeks to explain and define the relationship between two different variables, showing that they have a statistical bond of some sort. Null hypothesis states that there is no relation at all between statistical variables. A non-directional hypothesis is one that indicates the true value does not equal the predicted value. The directional hypothesis states that there is a direct relationship between two variables. Lastly, a statistical hypothesis is one which helps to understand the nature and character of a population. This method is great method for determining if the values and data satisfy the given hypothesis.
14. **Explain the Type I and Type II errors in Statistics?**
    a. A type I error is occurs when you get a false positive result, and a type II error occurs when you receive a false negative result.
15. **Explain univariate, bivariate, and multivariate analysis.**
    a. Univariate analysis is a statistical analysis that only summarizes one variable at a time. As the name suggests, a bivariate analysis is a statistical analysis that summarizes two variables, and multivariate is the same for more than two variables.
16. **Explain Data Visualization and its importance in data analytics?**
    a. Data visualization is the process of creating charts and graphs from a dataset. Data visualization is extremely important in data analytics because it is one of the most effective ways to communicate a lot of information that is learned from a dataset. It can also help show trends and correlation more easily than the numerical representation of the same.
17. **Explain Scatterplots.**
    a. A scatter plot is a graph in which the values of two variables are plotted along two axes. The patter or lack thereof that is revealed by the plotted points reveals if there is any correlation present between the variables.

18. **Explain histograms and bar graphs.**
    a. Histograms are the most commonly used graphs of frequency distributions. They show how many values fall into equally spaced “buckets” of data values. A bar graph visualizes the quantity of occurrences of categorical data.
19. **How is a density plot different from histograms?**
    a. At the basic level, the main difference between these two visualizations is that a histogram shows the counts of values in each range, where as a density plot shows the proportion of values in each range.
20. **What is Machine Learning?**
    a. Machine learning is a branch of artificial intelligence that works to give machines the human-like ability to learn and solve problems, through statistical models and algorithms.
21. **Explain which central tendency measures to be used on a particular data set?**
    a. On a normally distributed dataset we use the mean as the primary measure of central tendency, but with a skewed dataset we use the median, because a skew throws off the mean.
22. **What is the five-number summary in statistics?**
    a. The five-number summary gives the basic layout of numerical data. It helps to understand the spread of the data. It gives the minimum, maximum, first quartile, third quartile, and the median.
23. **What is the difference between population and sample?**
    a. A population is the entire group that you are seeking to learn about. A sample is a smaller group that is supposed to reflect the population that you actually collect data from.
24. **Explain the Interquartile range?**
    a. The interquartile range is the difference between the third and first quartiles. It tells that 50% of the data falls between those two points.
25. **What is linear regression?**
    a. Linear regression is the most basic and commonly used form of predictive analytics. It is used to predict the value of a variable based on the value of another variable. So, if two variables are strongly correlated then we should be able to use one to accurately predict the value of the other.
26. **What is correlation?**
    a. Correlation shows the interdependence of variable quantities. Or in other words, it shows how strongly the value of one variable is determined by the value of another.
27. **Distinguish between positive and negative correlations.**
    a. Positive correlation is when the two variables being considered are correlated and move in the same direction. As the value of one increases so does the value of the other. Negative correlation is the opposite. As the value of one increases the value of the other decreases proportionately.
28. **What is Range?**

    a. The range is the difference between the highest and the lowest values of a given variable. This is the broadest spread of the dataset.

29. **What is the normal distribution, and explain its characteristics?**
    a. A normal distribution, is a probability distribution that is symmetrical about the mean of the data meaning that data closer to the mean occur more frequently and vice versa.
30. **What are the differences between the regression and classification algorithms?**
    a. The main difference between the regression and classification algorithms is that regression algorithms are used to determine continuous values and classification algorithms are used to forecast or classify the distinct values (i.e. real or false, male or female, etc.).
31. **What is logistic regression?**
    a. Logistic regression is a type of regression analysis that is best to use when the dependent variable is binary.
32. **How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?**
    a. The RMSE is the standard deviation of the difference between the prediction and the actual of a dataset. The MSE is the average of the squares of all the differences between the actual points and their predicted points.
33. **What are the advantages of R programming?**
    a. There are several potential advantages of R programming depending on what you are wanting to use it for. It is open source, it has great support for data wrangling, it can create high quality plots and other visualizations, and it has available machine learning operations.
34. **Name a few packages used for data manipulation in R programming?**
    a. Some common packages used for data manipulation in R include dpylr, readr, and tidyr among others.
35. **Name a few packages used for data visualization in R programming?**
    a. Some common packages used for data visualization in R include ggplot, Leaflet, Plotly, and dygraphs among others.