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 can be defined as the process of cleaning and structuring data into a usable format that makes the process of data analysis much more efficient. One important aspect of data wrangling is handling missing values in your data set, which you can deal with by imputing values or removing them entirely.

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

Data analysis is the broad term for statistical inference that is used to draw inference from a dataset. It is a very general definition. However, data analytics involves the use of advanced, modern technologies and tools to process and analyze large datasets. Some activities that are involved in analytics are predictive modeling and machine learning, which are more advanced and extreme that basic data analysis.

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

Data science is a broad field of using various techniques to collect, analyze, and derive insights from data. Machine learning is a specific tool that is used in data science but is more specific and involves using data to create a model that can be used to make predictions and assist in decision-making.

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

The first step in an analytics project is to identify the problem or objective that will be addressed. Then, data needs to be collected that could potentially help provide insight into the problem at hand. The next step is to clean the data and conduct preprocessing that may involve many different things such as handling missing values or designating which variables will be used. Exploratory data analysis will be used to gain insight from the data and then modeling can be done to complete the project.

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

One problem that data analysts encounter during analysis is missing values. This is a problem because it can be hard to gain a trustworthy insight from a dataset if there is missing data that could be important or if the values are not entirely accurate. This can be handled by imputing mean or median values for integer variables. You can decide to use mean or median based on what type of distribution you have. Another common problem that analysts encounter is lack of context for the data. If an analyst is not aware of what their variables are or what the data means, then it can be hard to create a quality model that provides the most insight into the problem that is possible.

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

One technical tool that I have used for analysis is R Studio. This software aids in predictive modeling and statistical analysis because it is very efficient to use its pre-programmed functions and have it run tests on data. One tool I have used for presentation purposes is Tableau, which is extremely efficient and user-friendly because it is very simple and creates highly effective visuals that can be used for presentation purposes.

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

EDA is significant because it helps analysts gain a deeper understanding of the dataset that is being worked with. You can determine characteristics of the data, spread of distribution, and what variables are important for the data. Gaining context on the variables is important with EDA because you can determine what variables will help create the most efficient model. Another reason EDA is significant is because of data visualization. Creating visuals form the dataset is important because you can turn the data into information and use it solve problems and inform people who may not be as versed in the world of statistics.

8. What are the different methods of data collection?

There are many methods of data collection. A few methods are surveys with statistical questioning, observation, interviews, experiments, and social media analysis. These are just a few examples but are all important because they aid in collecting data that will be ultimately used to provide insight into a problem.

9. Explain descriptive, predictive, and prescriptive analytics.

Descriptive analytics is basically a summary of the data with visualization or numerical summary, like the 5-number summary. Predictive analytics involves the use of statistical modeling to determine what the data indicates and using it to predict what may happen in the future and make decisions. Prescriptive analytics is using the data and insight to make decisions on what to do moving forward to enhance business activity or productivity.

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

If there are only a few missing values for a numerical variable in a dataset, you can handle this by either imputing the mean or median of the values. You will decide to use mean or

median based on what type of distribution you have. If there are too many missing values that may cause the data to be unreliable, you can remove that variable entirely. If you are working with categorical data, you can fill in the missing values with “unknown” or something along those lines.

11. Explain the term Normal Distribution.

A normal distribution means that there is a bell-shaped curve consisting of random, continuous variables that are results of a sample. There should not be any outliers and normal distribution should follow the 68-95-99.7 rule.

12. How do you treat outliers in a dataset?

You can treat outliers in a dataset in multiple ways. One way would be by relying on median measurements rather than mean, since outliers to do not impact median as much as they can affect the mean. Another way would be to remove them entirely, because sometimes they can be unreliable and negatively skew the data.

13. What are the different types of Hypothesis testing?

There are many different types of hypothesis testing. The two hypotheses are the null hypotheses, and the alternative hypotheses. Usually, the null hypotheses must be assumed and is tested to see if there is evidence that it is false. Some statistical hypothesis tests include t-tests, anova tests, and chi-square tests.

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

A type I error is when you reject the null hypotheses even though it is exactly true. This is known as a false positive and indicates a difference in populations even though there is actually no difference. A type II error is when a false null hypothesis is not rejected. A type I error is much more dangerous, as it is always better to fail to reject the null hypotheses in cases of uncertainty.

15. Explain univariate, bivariate, and multivariate analysis.

Univariate analysis is when you examine the characteristics of one variable such as spread, center, and visualizations with histograms. Bivariate analysis is when you run statistical tests on two variables, usually one independent and dependent. This is usually a test of the relationship between the two variables. Lastly, multivariate analysis involves the study of three or more variables and the relationship and interaction between them.

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

Data visualization is when you use graphs, charts, and other modes of visualization to display the data or relationships between variables in a dataset. It is important because datasets usually look like a large collection of meaningless numbers, but visualization presents the data in an organized manner that is easy to understand. It can be very useful in presenting information to people in a quick and effective manner.

17. Explain Scatterplots.

Scatterplots are the graphical representation of two continuous, numerical variables and are used to identify patterns and trends. Each data point from the variables is on the graph and both values are represented. They are very good for identifying outliers and determining the relationship between two variables by finding the correlation value and linear regression.

18. Explain histograms and bar graphs.

Histograms and bar graphs are commonly misinterpreted to be the same type of graph, but that is not the case. A histogram displays counts of numeric, continuous data that is split into intervals. The bars on histograms represent the counts of the data that falls into the intervals decided and stated on the x-axis. Bar graphs on the other hand, display the counts of categorical data. The bars represent the value or count associated with a specific category. The main difference is that histograms display numerical data counts by intervals, while bar graphs display categorical data counts by group.

19. How is a density plot different from histograms?

Density plots do not display counts of numerical data in the intervals, instead they display the probability density at different points along the x-axis. Advanced mathematical equations are used to make density plots, so they are much more complex than a histogram. Just like a histogram, it is used to represent where most of the continuous data lies in a distribution.

20. What is Machine Learning?

Machine learning is a field of artificial intelligence that focuses on the development of algorithms and statistical models enabling computers to learn from and make predictions or decisions based on data, without being explicitly programmed.

21. Explain which central tendency measures to be used on a particular data set?

The central tendency measures that are most often used on a dataset are mean and median. The mean is the average of the dataset, while the median is the number that lies directly in the middle of the list of all datapoints. Mean is usually used when the data is normally distributed and is often used to compare different groups when doing different statistical hypothesis tests. Median is used when the data is not completely normally distributed and may be skewed or have some outliers. The reason median is used is because it is not as greatly impacted by outliers and still gives an accurate representation of the center of a dataset.

22. What is the five-number summary in statistics?

The five number summary is the list of numbers that are used to describe the spread and center of a distribution and is also used to make a boxplot. The five number summary consists of: minimum, first quartile, median, third quartile, and maximum.

23. What is the difference between population and sample?

The sample is a small group of subjects that are being tested on or observed. The population is the entire group of all people who have similar characteristics as the sample. It is difficult to test an entire population, so samples are used to draw inference on an entire population from tests on different samples.

24. Explain the Interquartile range?

The interquartile range is the range of the first quartile to the third quartile. The first quartile sits at about 25% of the way through the list of numbers made by the dataset, while the third quartile sits at about 75% of the way through the dataset. By subtracting the first quartile measurement from the third quartile measurement, you get the interquartile range.

25. What is linear regression?

Linear regression is a method for modeling the relationship between a dependent variable, or multiple dependent variables, and an independent variable that is affected by the dependent variables. This is done by formulating an equation that is best fit to describe the relationship between the data. The regression line can be used to show the pattern or relationship that dependent variables on the x-axis have with the independent variable on the y-axis. A regression equation is created from this line and is very important in predictive modeling, as an accurate regression equation will give you more accurate and trustworthy predicted outputs.

26. What is correlation?

Correlation is a measurement that shows how strong the relationship between two input variables is. It is a very great indicator of the relationship between two variables.

27. Distinguish between positive and negative correlations.

The correlation values range from -1 to 1. -1 dictates a very strong, inverse relationship and 1 dictates a very strong, direct relationship. If the number is close to zero, the relationship is very weak.

28. What is Range?

Range is a measure that represents the difference between the highest value in a dataset and the smallest. It is a very simple indication of the spread of a dataset. It is very sensitive to outliers and is not an indicator of the shape of a distribution or mean and median.

29. What is the normal distribution, and explain its characteristics?

A normal distribution means that there is a bell-shaped curve consisting of random, continuous variables that are results of a sample. There should not be any outliers and normal distribution should follow the 68-95-99.7 rule.

30. What are the differences between the regression and classification algorithms?

A regression algorithm predicts a continuous numerical variable, while a classification algorithm predicts a category that an object with qualities of the input variables belongs to. They are both used in predictive analysis, but regression predicts continuous variables while regression predicts discrete values/classes.

31. What is logistic regression?

Logistic regression is used in binary classification and uses classification algorithms to predict the class that the output variable will be, which is usually going to be a 0 or 1. Logistic regression uses a logarithmic function to predict the probability that an object with specific characteristics will belong to a certain class.

32. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?

There are complicated statistical methods that are used to find these equations, but all we actually need to do is call a couple functions in programs like R or Python and plug the regression model in, and it will return the value for these measurements.

33. What are the advantages of R programming?

R is very beneficial for statistical tests and visualization of distributions. This software aids in predictive modeling and statistical analysis because it is very efficient to use its pre-programmed functions and have it run tests on data.

34. Name a few packages used for data manipulation in R programming?

There are packages like Metrics, TidyVerse, and caTools.

35. Name a few packages used for data visualization in R programming?

The package that we mainly used was ggplot. This package has many pre programmed plots and the user can plug values in to the functions and create a visually pleasing and informative graph.