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 process of analyzing, transforming, and validating data from its raw and messy form to a more structured and high-quality form. It is used to make data more useful and consumable for analytical purposes, machine learning, and model predictions. The steps of data wrangling include thinking about questions you want to answer using the data, how you want to restructure the data, examining the data, structing data in an effective manner, cleaning the data, enriching the data, and publishing the data.
2. What are the differences between data analysis and data analytics?
    - Data analytics is broader than data analysis which requires the use of data analysis. Data analytics is a broad field which uses data and tools to make decisions. We are typically using models to make predictions about the future. The main goal is to find trends, uncover opportunities, make predictions, and make actions.
    - Data analysis is a subset of data analytics which deals with more hands-on data processing, evaluation, and exploration. We are typically looking at what happened in order to explain further. This is through cleaning, manipulating, and modeling data to find useful information. The main goal is give others access to the data, provide data visualization, and suggest what actions must be taken.
3. What are the differences between machine learning and data science? Though machine learning is utilized within data science the two concepts are different. Data science works with processing, managing, analyzing, and interpreting big data while utilizing data analysis and data analytics for decision making. Data science utilizes data wrangling. Machine learning utilizes software algorithms to specifically analyze the data in order to learn from it and form predictions for the future.
4. What are the various steps involved in any analytics project? First is defining the problem or question you wish to answer. Then exploring the data, preparing the data through transforming and cleaning for machine learning. Then the data modeling process to run the specific model and evaluate the results. Next is validating the model to test the accuracy of the model. Finally, selecting a specific model and implementing it to track the results and performance.

5. What are the common problems that data analysts encounter during analysis? Some common problems include missing or incomplete data, inconsistent data, inaccurate or false data, data security concerns, scaling data, duplicates of data, lack of good quality data, etc.
6. Which technical tools have you used for analysis and presentation purposes? Some technical tools I have used for analysis and presentation purposes include Python, R, SQL, Tableau, Microsoft Excel, Jupyter Notebook, and Microsoft PowerPoint.
7. What is the significance of Exploratory Data Analysis (EDA)? The EDA is crucial for analyzing rhe data before making assumptions about the data. Through an EDA, you can detect errors, outliers, and anomalies. It helps you have a better understanding of the data along with its patterns. An EDA helps in identifying possible relationships and correlations between variables which in turn can make the modeling process easier. Additionally, an EDA can help identify specific problems which can be expanded on to provide useful insight on the specific data.
8. What are the different methods of data collection? Surveys, interviews, focus groups, online tracking, transactional tracking, social media measurements, forms, secondary data collection, experiments, direct observations, etc.
9. Explain descriptive, predictive, and prescriptive analytics. Descriptive analytics tells you what already occurred. It is used to find patterns and trends based on what happened in the past. Predictive analytics tells you what could occur. It uses probability for its machine learning to make assessments of what could happen. Prescriptive analytics tells you what should occur in the future. It uses statistics and mathematical methods for algorithms to see what should happen in the future.
10. How can you handle missing values in a dataset? When using python, you can isnull().sum() to check whether the data has any missing values. From there you can use the fillna() function to fill in the missing values. If the data is numerical, you can fill in the missing values using mean, median, mode, or a constant. If the data is categorical, you can fill in missing values using the mode or a constant.
11. Explain the term Normal Distribution. Normal distribution refers to when the data is symmetrically distributed with no skew. This forms a bell-shaped curve with the mean at the center. This implies that most of the data is at or near the mean.
12. How do you treat outliers in a dataset? You can drop outliers to avoid skewing the data, you can reduce the weight of outliers, you can change the value of outliers, you can use the interquartile method, etc.
13. What are the different types of Hypothesis testing?
    - There are Two-Sided Hypothesis testing (two-tailed) and One-Sided hypothesis testing (right tail or left tail).
    - The Right Tailed hypothesis test refers to when the p-value is greater than the statistic. In other words, the alternative hypothesis claims that the value of the parameter specified by the null hypothesis is greater than what the null hypothesis claims. The testing area is on the right side of the normal distribution.
    - The Left Tailed hypothesis test refers to when the p-value is less than the statistic. In other words, the alternative hypothesis claims that the value of the parameter is

  less than what the null hypothesis claims. The testing area is on the left side of the normal distribution.
- The Two Tailed hypothesis test refers to when the testing area is on both sides of the normal distribution.

14. Explain the Type I and Type II errors in Statistics? A type I error is when you reject a null hypothesis that is true. This is known as a false positive. A type II error is when you fail to reject a null hypothesis that is false. This is known as a false negative.
15. Explain univariate, bivariate, and multivariate analysis. Univariate involves a single variable. Bivariate involves two variables. Multivariate involves two or more variables.
16. Explain Data Visualization and its importance in data analytics? Data visualization is the process of organizing and displaying data is a useful manner through graphs and charts to aid in understanding of the dataset. Data visualization is important because visualizations are an easy way to analyze and absorb information, visuals can help to easily understand complex problems. Visuals can help in identifying patterns, relationships, and outliers in data. Additionally, they can help in understanding business problems better while building a compelling story.
17. Explain Scatterplots. Scatterplots are used to visualize the relationship between two numeric variables. This can help to identify trend patterns and correlations between two variables. It can also help identify outliers. Scatterplots are often used in Machine learning concepts like regression, where x and y are continuous variables. It is also used in clustering scatters or outlier detection.
18. Explain histograms and bar graphs. Histograms are often used when you need to know the count of a variable. It shows the frequency of data using rectangles. The data is often grouped in equal sized bins. Bar graphs show the distribution of data over several groups using rectangles. It helps compare multiple numeric values. Histograms represent quantitative data and bar graphs represent categorical data.
19. How does a density plot differ from a histogram? Histograms are made up of rectangular bars while the density plot is made up of a smooth curve. The histogram shows the count of data per range while the density plot shows the proportion of data per range.
20. What is Machine Learning? Machine learning is a subfield of artificial intelligence and computer science. It is heavily utilized in machine learning, and it is a way of analyzing and studying statistical and software algorithms needed to make predictions and perform tasks without specific human instruction. It is often utilized for deep learning and data mining for big data.
21. Explain which central tendency measures are to be used on a particular data set?
    - Mean and median are both measures of centrality.
    - When the values of a dataset have the same units, you should use the arithmetic mean. When the values of a dataset have different units, you should use the geometric mean. When the values in the dataset are rates, use the harmonic mean.
    - When the distribution of data is symmetric, use the mean since it is a non-resistant measure. When the distribution of the data is skewed, use the median since it is a resistant measure.

22. What is the five-number summary in statistics? The 5 number summary includes the values for the minimum value of a dataset, the maximum value of a dataset, the upper quartile, the lower quartile, and the median.
23. What is the difference between population and sample? The population encompasses the entire dataset while the sample represents a small group/number within the population.
24. Explain the Interquartile range? It is the upper quartile minus the lower quartile. The interquartile range (IQR) shows the middle 50% of the data and shows how spread out the middle half of the dataset is. It is also a useful way to find outliers. The IQR can be visualized in a boxplot.
25. What is linear regression? It is a way of describing a relationship between a response variable (dependent variable) and one or more explanatory variables (independent variables). It is a way of predicting the value of an unknown data given other known data. It is used to see the differences between the actual value and predicted value (the residuals). When there is only one explanatory variable and one response variable, it is a simple linear regression. When there are multiple explanatory variables and one response variable, it is a multiple linear regression. Linear regression is used to minimize the prediction error for data points.
26. What is correlation? Correlation refers to the relationship or association between two variables. Statistically, it is a way to measure how linearly related two variables are.
27. Distinguish between positive and negative correlations.
    - A positive correlation refers to when the independent variable (x) increases, and the dependent variable (y) also increases. Or when the independent variable (x) decreases, and the dependent variable (y) also decreases. Both variables go in the same direction.
    - A negative correlation refers to when the independent variable (x) increases, and the dependent variable (y) decreases. Or when the independent variable (x) decreases, and the dependent variable (y) increases. Each variable goes in a different direction.
28. What is Range? The range refers to the maximum value minus the minimum value within a dataset.
29. What is the normal distribution, and explain its characteristics? Normal distribution forms a bell-shaped curve and it is centered at the mean. The distribution is symmetric and the mean, median, and mode are the exact same.
30. What are the differences between the regression and classification algorithms? Regression algorithms are used for numerical values while classification algorithms are used for categorical algorithms. Regression algorithms help to determine continuous values while classification algorithms help classify and forecast distinct values.
31. What is logistic regression? Logistic regression refers to the process of modeling the probability of a discrete (discontinuous) outcome given an input variable. It is often used to predict a dependent categorical variable. Logistic regression models work with binary outcomes or something that can take two variables. This includes true or false, 0 or 1, and yes or no. Unlike a linear regression, a logistic regression does not require a linear

relationship between the input and output variables and its range is bounded between 0 and 1.

32. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)? The Mean Square Error (MSE) is found by squaring the value of subtracting the observed value of y by the predicted value of y. Then dividing that by the number of datapoints. THE MSE tells you how close the regression line is to the datapoints and measures the variances of residuals. The Root Mean Square Error (RMSE) is found by taking the square root of the mean square error (MSE).
33. What are the advantages of R programming? Some advantages of R programming include the fact that it does not require a compiler and is open source. It is also known for its statistics. Thus, making it a better choice when dealing with regression models. R can also be used on multiple platforms including UNIX and LINUX. Alongside this, R is very compatible with other programming languages such as Python, Java, C++, etc. It also has high quality graphs and visuals needed for data analysis. Additionally, it is capable of graphing and plotting data. R also has numerous built-in packages. Lastly, R is known for its easy data wrangling.
34. Name a few packages used for data manipulation in R programming?
    - Tidyverse
    - Matrix
    - tidyr
    - list
    - dplyr
    - stringr
35. Name a few packages used for data visualization in R programming?
    - ggplot
    - ggcor
    - corrpot
    - PerformanceAnalytics
    - heatmap
    - plotly
    - shiny