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 – process of transforming raw data into usable, contextualized dataset.
2. What are the differences between data analysis and data analytics?
    a. Data analysis – specific process of treatment of collected data to draw conclusions from it, concerned with mathematics, for statistical analysis
    b. Data analytics – broader version of data analysis where expects collect data to synthesize, form of business intelligence, for predictive analysis
3. What are the differences between machine learning and data science?
    a. Data science – study of large amounts of data and understanding where it came from and its use within a company, experts convert raw data into business analytics
    b. Machine learning – field of study that gives computers the ability to learn without having an explicit coding line, a way of making the process less human
4.
5. What are the various steps involved in any analytics project?
    a. Obtain the data that is necessary for the task, explore data, clean data or fill missing values, conduct programming to gather information from the data, use visualizations, make conclusive statements based on evidence
6. What are the common problems that data analysts encounter during analysis?
    a. Missing data, duplicate data or columns, outliers
7. Which technical tools have you used for analysis and presentation purposes?
    a. R studio, Python, Tableau, SQL
8. What is the significance of Exploratory Data Analysis (EDA)?
    a. Helps summarize the data at an overall broader level. It identifies the main characteristics of the dataset and is usually done with data visualization models.
9. What are the different methods of data collection?
    a. Experiment, survey, observations, interviews
10. Explain descriptive, predictive, and prescriptive analytics.
    a. Descriptive – shows what happened in a business and describes why it happened.
    b. Predictive – uses data to predict what might happen based on past performance.

    c. Prescriptive – uses data to assess where the company is and what it needs to change to meet new goals.
11. How can you handle missing values in a dataset?
    a. You must first check whether the variable is categorical or numerical. If categorical, you can use the most common entry (mode) as a value to fill in the missing entries. If the column is numerical, you must first plot the column to assess the distribution. If it is normal, you can use the mean as the missing value filler, if not, use median.
12. Explain the term Normal Distribution.
    a. A normal distribution is one that follows a bell curve. This means that most of the entries are in the middle of the range of values and that if not in the middle, they are within two standard deviations.
13. How do you treat outliers in a dataset?
    a. Because outliers are still a true part of the dataset, it is best not to exclude them. Instead take note of them and their impact on the mean and median of the dataset.
14. What are the different types of Hypothesis testing?
    a. Z-test – used to determine if a relationship between variables are statistically significant, can only be used when population standard deviation is known
    b. T-test – most frequently used to determine statistical significance
    c. Chi-square – test of whether the model is a good fit
15. Explain the Type I and Type II errors in Statistics?
    a. Type I – false positive, the analysis predicted a positive but the actual result was a negative
    b. Type II – false negative, the analysis predicted a negative but the actual result was a positive.
16. Explain univariate, bivariate, and multivariate analysis.
    a. Univariate, bivariate, and multivariate analysis only differ by the amount of variables involved with univariate being one, bivariate being two, and multivariate being more than two.
17. Explain Data Visualization and its importance in data analytics?
    a. Data visualization is a way of formatting large amounts of data into digestible models such as scatterplots, bar charts, or line graphs. It is important to data analytics due to the benefit of allowing assumptions and conclusions to be made about the data at a broader level.
18. Explain Scatterplots.
    a. Scatterplots are a type of data visualization that plots two data columns on different axes to see their pattern of the relationship and determine its linear or non linear relationship.
19. Explain histograms and bar graphs.
    a. Histograms and bar graphs are both graphs that show the distribution of a data set. Usually, the x axis is the independent variable and the y axis is the frequency of that variable. Bar graphs tend to be discrete variables while histograms tend to be continuous.

20. How is a density plot different from histograms?
    a. Histograms use bars to measure frequency while the density plot compiles the data into a curve to see the distribution more smoothly.
21. What is Machine Learning?
    a. Machine learning – field of study that gives computers the ability to learn without having an explicit coding line, a way of making the process less human
22. Explain which central tendency measures to be used on a particular data set?
    a. The three measures of center are mean, median, and mode. If the distribution of the data is normal, the mean is the best measure. If the distribution is skewed, the median is the best measure. The mode is another measure for variables where the other two are not suitable.
23. What is the five-number summary in statistics?
    a. The five number summary is the range description of the data set with the minimum value (Q1), lower quartile (Q2), median (Q3), upper quartile (Q4), and the maximum (Q5).
24. What is the difference between population and sample?
    a. Population – the main type of subject that you want to study (ex: women 18-25 years old)
    b. Sample – a small selection of the population in an attempt to predict the results of the population (ex: 30 interviews from women 18-25 years old)
25. Explain the Interquartile range?
    a. The interquartile range is the difference of the lower quartile and the upper quartile and it represents the middle 50% of the data.
26. What is linear regression?
    a. Linear regression is a model done to data to find an equation or a line that best fits the data's trend and relationship.
27. What is correlation?
    a. Correlation is the description of the relationship between variables and the determination if the variables moved in a way where one effects the other. Correlation is a number that is stronger the closer it is to 1, and weaker the closer it is to 0.
28. Distinguish between positive and negative correlations.
    a. A negative correlation indicates that as one variable increases, the other decreases and vice versa. A positive correlation indicates that both variables move in the same direction.
29. What is Range?
    a. Range is the difference between the highest and the lowest value in a data set. It is a measure of dispersion and the spread of a distribution.
30. What is the normal distribution, and explain its characteristics?
    a. A normal distribution is one that follows a bell curve. This means that most of the entries are in the middle of the range of values and that if not in the middle, they are within two standard deviations.
31. What are the differences between the regression and classification algorithms?

    a. Regression analysis is a model used for prediction of numerical values (trying to predict a price or a quantity) while classification is used to predict a categorical value (having a quality or trait based on data)
32. What is logistic regression?
    a. Logistic regression is the process of making a prediction model for a categorical variable. The model will come out with four outcomes: true positive, true negative, false positive, and false negative.
33. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?
    a. MSE is the sum of the squares of the difference between the predicted and actual y value of a prediction model. The RMSE is the square root of the MSE.
34. What are the advantages of R programming?
    a. Free for anyone, plenty of data visualization models (including density plot), multiple packages available with statistics already installed.
35. Name a few packages used for data manipulation in R programming?
    a. Mutate() – adds a column that can be used to manipulate the data into new calculations.
    b. Tidyverse – cleans up the data and allows it to be more manageable
36. Name a few packages used for data visualization in R programming?
    a. Ggplot – used for data visualization and plots a variety of graphs with different kinds of manipulation (scatter, density)