Braden Heuglin
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.
   Data wrangling is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

2. What are the differences between data analysis and data analytics?
   Data analysis is a process involving the collection, manipulation, and examination of data for getting a deep insight. Data analytics is taking the analyzed data and working on it in a meaningful and useful way to make well-versed business decisions.

3. What are the differences between machine learning and data science?
   Machine learning helps make artificial intelligence — the science of making machines capable of human-like decision-making — possible. Data science is the process of developing systems that gather and analyze disparate information to uncover solutions to various business challenges and solve real-world problems.

4. What are the various steps involved in any analytics project?
   First, you have to find and access the data you wish to work with. Next, you read that data into whatever program you are using. Then, you check the data for any missing or null values, and remove or fix any of those values. After that you have free reign to manipulate the data in any way you wish, from creating plots to just simply displaying what values the data contains.

5. What are the common problems that data analysts encounter during analysis?
   The most common problem is missing or null values in a dataset. If you do not check the data for missing values and try to create plots with an incomplete dataset, your results are not going to be accurate or useful.

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

We've used a lot of tools to analyze data and present it in graphs and plots. From matplotlib and seaborn in python, to ggplot and tidyr in r, almost every program we've used has had the ability to analyze and present data in various ways.

7. What is the significance of Exploratory Data Analysis (EDA)?
   The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

8. What are the different methods of data collection?
   Some common data collection methods include surveys, interviews, observations, focus groups, experiments, and secondary data analysis.

9. Explain descriptive, predictive, and prescriptive analytics.
   Descriptive analytics is the analysis of historical data using trio key methods - data aggregation and data mining - which are used to uncover trends and patterns. it is concerned with representing what has happened in the past.
   Predictive analytics is a more advanced method of data analysis that uses probabilities to make assessments of what could happen in the future. Like descriptive analytics, prescriptive analytics uses data mining – however it also uses statistical modeling and machine learning techniques to identify the likelihood of future outcomes based on historical data.
   While predictive analytics shows companies the raw results of their potential actions, prescriptive analytics shows companies which option is the best.

10. How can you handle missing values in a dataset?
    As I mentioned in a previous answer, you must fill or remove any missing values you find in a dataset. If you do not, the methods you use to display your results will be inaccurate and therefore, useless.

11. Explain the term Normal Distribution.
    t\A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme.

12. How do you treat outliers in a dataset?
    There are two ways to handle outliers. First, and the best way, is to remove the outliers so the data is more normal and useful, be sure to let people know you had to remove outliers though. Second, you can leave the outliers in but you *must* tell anyone who looks at your work which values are outliers and how they affect your findings.

13. What are the different types of Hypothesis testing?
    Alternative, null, non-directional, directional, and statistical.

14. Explain the Type I and Type II errors in Statistics?
    Type I is a false positive, you say the null is true when it is actually false. Type II is a false negative, you say the null is false when it is actually true.

15. Explain univariate, bivariate, and multivariate analysis.
    Univariate analysis looks at one variable, bivariate looks at two variables, and multivariate looks at more than two variables and their relationship.

16. Explain Data Visualization and its importance in data analytics?
    Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

17. Explain Scatterplots.
    A scatter plot compares two variables and maps out how related they are. You can use scatter plots to find the correlation between two variables as well as see the spread of the points laid out.

18. Explain histograms and bar graphs.
    A bar graph is a graphical representation of categorical data, whereas a histogram is a graphical representation of quantitative data.

19. How is a density plot different from histograms?
    A histogram shows the counts of quantitative values, while a density plot shows the proportion of values.

20. What is Machine Learning?
    Machine learning focuses on the use of data and algorithms to imitate the way that humans learn, *learning* to improve its accuracy.

21. Explain which central tendency measures to be used on a particular data set?
    There are two ways to measure the center of a dataset, mean and median. The mean is the average of all values in the dataset, whereas the median is the middle value. In a case where there are many outliers or skew, we use the median because the mean is affected by the outliers and skew.

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

It is a series of five numbers which explain the distribution of a dataset to help in understanding the data and making decisions with it. The numbers are min, Q1, median, Q3, and max.

23. What is the difference between population and sample?
The population is the entire group you want to learn something about, whereas a sample is a small group in that population.

24. Explain the Interquartile range?
The IQR is the difference between Q3 and Q1 of a dataset. Q1 is the point halfway between the minimum value and median, and Q3 is the point halfway between the maximum value and median.

25. What is linear regression?
Linear regression analysis is used to predict the value of a variable based on the value of another variable.

26. What is correlation?
Correlation is how closely related two variables are to each other. If the correlation is high, or close to 1, that means the two variables are very closely related and as one changes, it either positively or negatively changes the other.

27. Distinguish between positive and negative correlations.
A positive correlation is when two variables follow each other, so when one rises, so does the other and vice versa. A negative correlation is the opposite, when one variable rises, the other falls and vice versa.

28. What is Range?
Range is the difference between the minimum and maximum values in a dataset. It shows how large the dataset can potentially be.

29. What is the normal distribution, and explain its characteristics?
A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. It is also called a bell curve because that is the shape the plot takes in a normal distribution.

30. What are the differences between the regression and classification algorithms?
The key distinction between regression and classification algorithms is regression algorithms are used to determine continuous values, whereas classification algorithms are used to classify distinct values.

31. What is logistic regression?

Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors.

32. How do you find Root Mean Square Error (RMSE) and Mean Square Error (MSE)?
    To find the MSE, take the observed value, subtract the predicted value, and square that difference. Do that for all values, then sum all of those squared values and divide by the number of values in the dataset.
    To find the RMSE, simply take the square root of the MSE.

33. What are the advantages of R programming?
    R programming is open source, meaning you don't need to pay or get a license to use it. It's great for clearing up data by using dplyr and readr, as well as visualizing that data with ggplot and tidyr. It also has machine learning operations and statistical analysis operations.

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

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