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 variety of processes that are carried out to convert raw data into more readily formats. This is also known as data cleaning.

2. **What are the differences between data analysis and data analytics?**
   Data analytics is a type of analytics used in enterprises to make data-driven decisions. Data analysis is a specialized type of analytics used in businesses to evaluate data and gain insights. It has one or more users and generally consists of data collection and inspection.

3. **What are the differences between machine learning and data science?**
   Data science is a field that studies data and how to extract meaning from it, whereas machine learning is a field devoted to understanding and building methods that utilize data to improve performance or inform predictions.

4. **What are the various steps involved in any analytics project?**
   In an analytics project, the first step is to define the question (forming a hypothesis and figuring out how to test it). Next, data preparation consists of collecting data and storing the information in an accessible format. Then, data cleaning consists of removing errors/outliers and filling in major gaps. After, data analysis takes place to find relationships, patterns, and trends within the dataset. Finally, sharing results and actively making use of the data and analysis results.

5. **What are the common problems that data analysts encounter during analysis?**
   Besides the possibility of messy data due to the high volume, they also face other challenges such as collecting meaningful data, selecting the right analytics tool, data visualization, multiple-source data, low-quality data, lack of skills, scaling challenges, data security, budget limitations, lack of a data culture, and inaccessibility.

6. **Which technical tools have you used for analysis and presentation purposes?**
   Microsoft Excel, Python (NumPy, Matplotlib, Seaborn, Pandas), R, Tableau, SQL.

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, and find interesting relations among the variables.

8. **What are the different methods of data collection?**
   Surveys, interviews, observations, focus groups, experiments, and secondary data analysis.

9. **Explain descriptive, predictive, and prescriptive analytics.**
   Descriptive analytics tells us what has already happened; predictive analytics shows us what could happen, and finally, prescriptive analytics informs us what should happen in the future.

10. **How can you handle missing values in a dataset?**
    One way of handling missing values is the deletion of the rows or columns having null values. Another way is by imputing missing values with mean or median values.

11. **Explain the term Normal Distribution.**
    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 infinity. The middle of the range is also known as the mean of the distribution. The standard normal distribution is often described as a bell-shaped curve. However, the normal distribution can also be right/left skewed, depending on outliers or extreme values.

12. **How do you treat outliers in a dataset?**
    You could treat outliers by deleting the outlier if they are due to data entry errors caused due to human error, or data processing errors. Alternatively, you can replace the outlier with the mean value, or median value.

13. **What are the different types of Hypothesis testing?**
    Alternative hypothesis explains and defines the relationship between two variables; null hypothesis states that there is no relation between statistical variables; non-directional hypothesis is a tow-tailed hypothesis that indicates the true value does not equal the predicted value; directional hypothesis is when there is a direct relationship between two variables; statistical hypothesis testing helps in understanding the nature and character of the population.

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

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

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

16. **Explain Data Visualization and its importance in data analytics?**
    Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The importance of data visualization is to make it easier to identify patterns, trends and outliers in large data sets.

17. **Explain Scatterplots.**
    A scatterplot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

18. **Explain histograms and bar graphs.**
    Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data.

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

20. **What is Machine Learning?**
    Machine learning is a subfield of artificial intelligence. It is the capability of a machine to imitate intelligent human behavior.

21. **Explain which central tendency measures to be used on a particular data set?**
    Central tendency is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or center of its distribution. There are three main measures of central tendency: mean, median, and mode.

22. **What is the five-number summary in statistics?**
    Minimum value, lower quartile (Q_1), median value (Q_2), upper quartile (Q_3), maximum value.

23. **What is the difference between population and sample?**

A population is the entire group that you want to draw conclusions about. A sample is the specific group that you will collect data from.

24. Explain the Interquartile range?

The interquartile range tells you the spread of the middle half of your distribution

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 a statistical measure that expresses the extent to which two variables are linearly related.

27. Distinguish between positive and negative correlations.

A positive correlation means that when one variable increases, the other variable increases too. A negative correlation means that when one variable increases, the other variable decreases.

28. What is Range?

The range is the difference between the lowest and highest values.

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

Normal distributions are symmetric, unimodal, and asymptotic, and the mean, median, and mode are all equal. A normal distribution is perfectly symmetrical around its center. That is, the right side of the center is a mirror image of the left side. There is also only one mode, or peak, in a normal distribution.

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

Regression algorithms are used to determine continuous values, such as price and age. Whereas, classification algorithms are used to determine distinct values, such as male/female or true/false.

31. What is logistic regression?

Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other.

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. Repeat that for all observations. Then, sum all the squared values and divide by the number of observations. To find the RMSE, you square root the MSE.

33. What are the advantages of R programming?

R is an open-source programming language, R provides exemplary support for data wrangling, R has a vast array of packages, and R facilitates quality plotting and graphing.

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

ggplot2, tidyverse, dplyr, tidyr.

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

ggplot2, Lattice, and Plotly.