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.

-is the process of converting raw data into a usable form

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

-data analytics converts raw data

-data analysis is a specialized type of analytics used in business to evaluate data and gain insights

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

-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?
- Defining the question, collecting the data, cleaning the data, analyze the data, sharing results and summary
5. What are the common problems that data analysts encounter during analysis?

-collecting meaningful data, selecting the right analytics tool, data visualization, multiple source data, low-quality data

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

-python, R, matplotlib, seaborn, pandas

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

-to help look at data before making any assumptions. It can help identify any obvious errors as well as better understand patterns within the data

8. What are the different methods of data collection?

-questionnaires, surveys, documents, and records

9. Explain descriptive, predictive, and prescriptive analytics.

-descriptive analytics tell us what has already happened

-predictive shows what could happen

Prescriptive informs what should happen in the future

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

-using the function isnull( ).sum( )

11. Explain the term 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

12. How do you treat outliers in a dataset?

-you can replace the outlier with the mean or median value or in some cases dropping the suspected outlier to avoid any bias

13. What are the different types of Hypothesis testing?

-simple, complex, alternative, composite, directional, non-directional, logical, empirical, statistical, associative, exact, and inexact

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

-type 1 error is false-positive which occurs if an investigator rejects a null hypothesis that is true in the population

-type 2 is a false-negative which occurs if the investigator fails to reject a null hypothesis that is false in the population

15. Explain univariate, bivariate, and multivariate analysis.

-univariate summarizes one variable at a time

-bivariate compares two variables

-multivariate compares more than two variables

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 the data easier to visual

17. Explain Scatterplots.

-uses dots to present values for two different numeric variables

18. Explain histograms and bar graphs.

-a bar graph is a graphical representation of categorical data using rectangular bars where the length of each bar is proportional to the values they represent

-a histogram is the where the data is grouped into continuous number ranges and each number range corresponds to a vertical bar

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. A density plot is a smooth curve that shows the distribution of the data in a more continuous way.

20. What is Machine Learning?

-is devoted to understanding and methods that let the machine "learn" which helps to improve accuracy

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

-it measures mean and median

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

-the most extreme values in the data set, max and min values, the lower and upper quartiles, and the median. These values are presented together and ordered from lowest to highest

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?
- Contains the second and third quartiles, or the middle half of the data set. The range is spread throughout the whole data set

25. What is linear regression?

-basic and commonly used predicative analysis, its used to determine the character and strength of the association between a dependent variable and a series of other independent variables

26. What is correlation?

-relationship between two or more objects

27. Distinguish between positive and negative correlations.
- When two variables operate in unison so that when one variable rises or falls the other does the same. A negative correlation is when two variables move opposite one another si that when one variable rises the other falls

28. What is Range?

-the difference between the smallest and highest numbers in a set or list

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

- symmetric, unimodal, asymptotic and the mean, median and mode are equal. Normal distribution is perfectly symmetrical around its center

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

-regression algorithms are used to predict the continuous values such as price, salary and age while classification algorithms are used to predict the discrete values such as male/female, true/false

31. What is logistic regression?

-statistical model that uses the logistic function between x and y

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

-the sum of the absolute difference between actual and predicted values

33. What are the advantages of R programming?

-its open source, platform independent, has a lot of packages, well suited for machine learning

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

-dplyr, ggplot, readr, data.table

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

- Ggplot, plotly