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 cleaning and preparing raw data for analysis.

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

data analysis is hands-on data exploration and evaluation. Data analytics is a broader term that includes data analysis as a necessary part.

3. What are the differences between machine learning and data science?

Data science involves collecting, processing, cleaning, analyzing, and interpreting large and complex datasets to gain insights into possible business solutions. Machine learning is a subset of data science that involves algorithms that can learn from data on their own and make predictions for the data.

4. What are the various steps involved in any analytics project?

You need to collect the data, refine the data, store it, analyze it, and then deliver it.

5. What are the common problems that data analysts encounter during analysis?

Some common problems are incomplete data (data has missing values), outliers in the data, bias in the data, or inconsistent data (multiple different types of data that can't work together).

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

Some technical tools I have used are pandas – a Python library to organize, preprocess, and view the data, Seaborn – a Python library to graph/chart the data and sklearn to use both ML models and view the data's R squared and mean squared values.

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

It is important because it requires you to identify important relationships with the data. Without a plan and something to look for, the data is useless. It is important to look at the data with all tools available to discover patterns, relationships, and possible errors in the data.

8. What are the different methods of data collection?

You can find data online through websites, you can collect data yourself, you can collect data from other people, and you can scrape data from online.

9. Explain descriptive, predictive, and prescriptive analytics.

Descriptive analysis is the exploration of historical data to identify patterns trends and relationships. Predictive uses statistical and ML models to make predictions on future events. Prescriptive involves using optimization techniques to identify the best action to take in a situation.

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

You can just remove the missing values, and impute them with the common mean/median/mode, or you can predict the missing values with ML models.

11. Explain the term Normal Distribution.

A normal distribution is a distribution that follows a bell-shaped curve and is symmetrically distributed around the mean, with the majority of data within one standard deviation of the mean.

12. How do you treat outliers in a dataset?

We can either remove them from the dataset, impute the outlier values with estimated values, or we can keep the outliers.

13. What are the different types of Hypothesis testing?

There are parametric tests that assume that the data is normally distributed and these include t-tests (mean of 2 groups are significantly different from each other) and z-tests (used when population standard deviation is known). There are also nonparametric tests that do not assume that the data is normally distributed. This is used when the sample size is small as well.

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

A type 1 error occurs when the hypothesis is rejected when it is actually true (false positive). A type 2 error is when the hypothesis is not rejected it is actually false (false negative).

15. Explain univariate, bivariate, and multivariate analysis.

The univariate analysis involves a single variable at a time. It is used to describe the distribution of the variable and to identify patterns in the data. Commonly used to identify mean, median, and mode. Bivariate analysis is done with 2 variables. Used to determine the strength and direction of the relationship and is often done with correlation coefficients. Multivariate analysis is done with 3 or more variables and is used commonly for regression and cluster analysis.

16. Explain Data Visualization and its importance in data analytics?

Data visualization is the graphical representation of data. It usually involves charts, graphs, and maps to show complex data and their relationships. It is important to data analytics because it helps make trends/patterns and variables in data more understandable and concise.

17. Explain Scatterplots.

Scatter plots are used to visualize the relationship between 2 variables and to identify patterns in the data. Useful for identifying outliers and identifying linear/nonlinear relationships. Each point on the plot represents a combination of the 2 variables for a single observation in the data set.

18. Explain histograms and bar graphs.

Histograms are a display of the distribution of a numerical variable. It is made by dividing the range of the variable into different bins and counting the number of occurrences that falls in each bin. They are used to show the shape of the distribution of the variable. Bar graphs are a representation of categorical data. It divides categories of the variable along the x-axis and then shows the count of occurrences for each category as a vertical bar.

19. How is a density plot different from histograms?

They are both used to visualize data but histograms display the frequency of occurrences within each interval while density plots display the probability density of the variable.

20. What is Machine Learning?

Subset of artificial intelligence involves developing algorithms and models that computer systems can automatically learn and improve from experience, without being explicitly told so. The goal is to recognize patterns and relationships in data.

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

Mean is used when the data is normally distributed and does not contain outliers. Median is useful when the data has skewed values or outliers. Mode is useful for nominal or categorical data.

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

The five-number summary provides a summary of the distribution of a dataset. It shows the minimum value, the maximum values, the first quartile, the second quartile (median), and the third quartile.

23. What is the difference between population and sample?

The population is the entire group while the sample is a subset of that group.

24. Explain the Interquartile range?

The interquartile range measures the middle of the dataset and it is defined as the difference between the 3^{rd} and 1^{st} quartile of the data. It is commonly used in box plots and useful when discussing where the majority of the dataset lies.

25. What is linear regression?

It is a statistical method used to model the linear relationship between a dependent variable and one or more independent variables. It is considered a supervised learning technique. The goal is to estimate the values of a and b that minimize the difference between the predicted values of Y and the actual values of Y.

26. What is correlation?

It is a technique used to measure the strength and direction of the relationship between 2 or more variables. In other words, how many variables are related to one another?

27. Distinguish between positive and negative correlations.

A positive correlation means as one variable increases, the other variable also increases. A negative correlation refers to when one variable increases, the other decreases.

28. What is Range?

Range measures the spread of a dataset. It's the difference between the maximum and minimum values in a dataset.

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

The normal distribution is a probability distribution that is symmetrical around its mean. It has a bell-shaped curve with the peak being mean. The mean, median, and mode are all equal as well. Finally, the skew is zero on the graph/data.

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

Regression is used to predict a continuous variable such as the price of a house. Classification is used to predict a discrete variable such as whether or not it will rain. Regression is typically a numerical range of values while classification is categorical.

31. What is logistic regression?

It is a regression ML algorithm used for classification problems. It is used to estimate the probability of a binary response (yes/no) based on multiple variables. Can be used for binary classification or multi-class classification and can handle both continuous and categorical variables.

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

MSE is (1/n) * the sum of (y actual – y predicted)^2. RMSE is the same equation but we take the square root of it. N is the number of data points.

33. What are the advantages of R programming?

It has a wide range of packages, good for the visualization of datasets, is open source, and it can be set up and run very easily. It is a very popular programming language for data manipulation and visualization.

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

Tidyr, plyr, dplyr

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

Ggplot2, lattice, plotly