Leighann Robinson

Individual Project 9

Descriptive Questions

DS160 SP 2023

1. Define the term 'Data Wrangling' in Data Analytics.

In data analytics, data wrangling refers to the process of collecting, selecting, and cleaning data to transform the raw data into a different format. This different format can then be used to help answer analytical questions.

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

Data analytics is a broad and general form of analytics used in companies making decisions that are data driven. Data analysis is a specialized form of analytics used in businesses to gain insights and evaluate collected data. Data analysis consists of cleaning, transforming, modeling, and questioning data to find useful information. Analytics usually models the future or predicts a result, whereas analysis is used to describe what happened. Data analysis is a subset of data analytics.

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

Machine learning is a branch of artificial intelligence that consists of algorithms that help improve through supervised, unsupervised, and reinforcement learning. Machine learning is its own field that deals with understanding and building methods that utilize data to improve performance or inform predictions. Data science, however, combines domain knowledge, programming skills, and mathematical and statistical knowledge to extract meaningful findings from data.

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

   The various steps in any analytics project are as follows:

   - Understand the business issues
   - Understand the dataset
   - Prepare the data
   - Perform exploratory analysis and modeling
   - Validate the data
   - Visualize and present findings

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

Common problems that data analysts encounter during analysis include poorly formatted data, numerous missing values, misclassified data, and having inconsistent, incomplete, or duplicate entries within the data. All these issues can be encountered during analysis and must be properly handled to draw useful findings from the data.

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

The technical tools that we have used for analysis and presentation purposes in this course have been Python libraries such as numpy, matplotlib, pandas, and seaborn, as well as RStudio, SQL, and Tableau. Analysis of data has been done primarily in RStudio, SQL, and the Python libraries pandas and numpy. Presentation and data visualization has been done in Tableau and the Python libraries of matplotlib, numpy, seaborn, and pandas.

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

The main purpose of EDA is to help look at data before making assumptions. EDA helps to find errors in the data that can affect the distribution. Also, EDA helps to further analyze and find patterns and outliers within the data. EDA is meant to find the underlying structure of the dataset.

8. What are the different methods of data collection?

Data collection is the process of collecting, measuring and analyzing different types of information from a dataset. Different methods of collection include primary and secondaryh collection. Examples of primary data collection methods are surveys, interviews and statistical methods. Examples of secondary data collection methods are finanical and sales reports, as well as government reports such as the census.

9. Explain descriptive, predictive, and prescriptive analytics.

Descriptive analytics provides insight into historical data. It is commonly the first kind of data analysis performed on a data set. Description and interpretation processes are different steps. Descriptive analytics deals with measuring frequency distributions, measures of centrality, and measures of dispersion of the data. Predictive analytics, much like the name suggests, predicts changes in markets or customer demand. Predictive analytics does to some extent look into the past using purchase behavior, market trends, rates, etc. Prescriptive analysis recommends decisions to be made. It is dedicated to finding the best solution for a given situation.

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

Missing data values in datasets can be imputed, meaning any values that show as null can be changed to match either the mean or the median depending on the data's distribution and skew. Leaving null values can further skew the data found in the database.

11. Explain the term Normal Distribution.

A normal distribution, also called a bell-curve, shows data distributed symmetrically around the center of all data values (median). Normal distributions represent a lack of skew in the data values.

12. How do you treat outliers in a dataset?

Outliers, while causing the data to be skewed either positively or negatively depending on the value, help to showcase any anomalies within the data. Outliers can generally be ignored if they are due to errors in entry. Outliers can be replaced by the mean or median value or dropped entirely from the dataset to avoid any possible bias.

13. What are the different types of Hypothesis testing?

There are five different types of hypothesis testing. The alternative hypothesis explains the relationship between two variables. It simply indicates a positive relationship between the two variables showcasing that they have a statistical bond. The null hypothesis states that there is no relation between statistical variables. If the facts presented at the start of exploration do not match with the outcomes, the testing is null hypothesis testing. The non-directional hypothesis means that there is no direction between the two variables. This type of hypothesis indicates that the true value and the predicted value are not equal. In the Directional hypothesis, there is a direct relationship between two variables, meaning any of the variables have an influence on the others. Statistical hypothesis testing helps in understanding the nature and character of the population. It is a great method to decide whether the values and the data presented satisfies the given hypothesis or not.

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

Type I errors in statistics refer to a false positive. This occurs if an investigator rejects a null hypothesis that is true in the population. A type II error occurs if the investigator fails to reject a null hypothesis false in the population. Type II errors are called false negatives.

15. Explain univariate, bivariate, and multivariate analysis.

Univariate analysis goes about exploring each variable within a dataset separately. Univariate analysis explores the range of values as well as the measurements of central tendency. Bivariate analysis examines how two different variables are related. It can be used to measure how strong of a statistical link there is between two variables. Multivariate analysis is used to study more complex sets of data. It is used to describe analyses of a dataset where there are multiple outcomes and observations made for each individual variable.

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

Data visualization is the representation of information from a dataset through the usage of charts, diagrams, infographics, among other visuals. Data visualization aids in making complex data relationships easy to understand and communicate. It is an important part of analytics due to the fact that visualization helps to display findings in a way that anyone can understand and read.

17. Explain Scatterplots.

Scatterplots help to visualize two numeric values. The position of each dot on the x and y axes indicates values for an individual data point. Scatter plots are useful in detecting outliers within the dataset as well as trends and correlation between the variables using regression.

18. Explain histograms and bar graphs.

A histogram shows the distribution of a variable. It converts numerical data into bins as columns. On the x axis is the range of values and on the y axis, the frequency of those values in the dataset. Bar graphs show the distribution of data over several groups. Bar graphs are commonly confused with a histogram but differ in the fact that bar graphs compare multiple numeric values and do not have bins.

19. How is a density plot different from histograms?

Density plots show a smooth curve that shows the distribution of the data in a continuous way whereas histograms have touching bars to show the distribution.

20. What is Machine Learning?

Machine learning is a branch of artificial intelligence that consists of algorithms that help improve through supervised, unsupervised, and reinforcement learning. Machine learning is its own field that deals with understanding and building methods that utilize data to improve performance or inform predictions.

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

Mean imputation is not always a very good idea. The mean is the average of all of the data values and averages can easily be skewed. The imputation of missing data with mean values can only be done with numerical data. When the data is skewed, it is a good idea to consider using the median value for replacing the missing values. The median is the middle value of all of the data values. Median imputation can only be done with numerical data. Mode imputation is good to consider for replacing the missing values when data is also skewed. Imputing missing data with mode values is mostly done with categorical data as a dataset can have multiple numerical modes.

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

The five-number summary in statistics consists of the most extreme values in the dataset (the minimum and maximum values), the median, and the first and third quartiles. The median is equivalent to the second quartile.

23. What is the difference between population and sample?

A sample is the specific group from which data is collected from. A population is the entire group from which conclusions are going to be drawn from. A sample is always smaller than the population.

24. Explain the Interquartile range?

The interquartile range (IQR) is a number that indicates how spread the middle half of the dataset is. This can help determine outliers. The IQR can be found by subtracting the value of the first quartile (Q1) from the value of the third quartile (Q3).

25. What is linear regression?

Linear regression is a statistical regression method used for predictive analysis. Linear regression shows the linear relationship between the independent variable (x-axis) and the dependent variable (y-axis).

26. What is correlation?

Correlation refers to the explanation of how one or more variables are related to one another. Correlation is a part of diagnostic analysis.

27. Distinguish between positive and negative correlations.

A positive correlation means that the variables move in the same direction. As one variable increases so does the other and if one variable decreases so does the other. A negative correlation means that the variables move in opposite directions. As one variable increases, the other decreases.

28. What is Range?

Range is the difference between the largest and smallest values in a dataset.

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

A normal distribution, also called a bell-curve, shows data distributed symmetrically around the center of all data values (median). Normal distributions represent a lack of skew in the data values.

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

Regression algorithms are used to predict continuous quantities. They are able to showcase changes over periods of time. Classification algorithms are used to forecast or classify different values. Classification algorithms are used in statistics most commonly with data that binary outcomes.

31. What is logistic regression?

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome, something such as true/false or yes/no. It differs from linear regression in that it has a range bounded between 0 and 1.

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

Root Mean Square Error (RMSE) can be found in Jupyter Notebook by the following code: print (f"RMSE: {math.sqrt(mean_squared_error(y_test,y_pred)):.2f}"). This code is in formatted string, printing that RMSE is equal to the square root of the mean squared error (MSE) between the variables in the parenthesis. MSE can be found in Jupyter Notebook by the following code: print (f"MSE: {mean_squared_error(y_test,y_pred):.2f}"). This code is formatted like RMSE but does not take the square root. Both RMSE and MSE can be found in R using the summary of the database.

33. What are the advantages of R programming?

The advantages of R programming include clean, easy to read graphs, it has an easy-to-use library available, and can run locally. The R programming language is used most in statistical analysis.

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

Some packages used for data manipulation in R programming include tidyverse , and caTools which further include pipe operators (%>%), the mutate () function, and the select () function for manipulating the dataset.

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

Packages used for data visualization in R programming include ggplot which helps to plot different graphs and visuals in RStudio.