Data Visualization and Pre-processing Perform Below Tasks to complete the assignment:-

Tasks:-

1. Download the dataset: Dataset

2. Load the dataset.

3. Perform Below Visualizations.

● Univariate Analysis

● Bi - Variate Analysis

● Multi - Variate Analysis

4. Perform descriptive statistics on the dataset

5. Handle the Missing values.

6. Find the outliers and replace the outliers

7. Check for Categorical columns and perform encoding.

8. Split the data into dependent and independent variables.

9. Scale the independent variables

10. Split the data into training and testing

Data Visualization and Pre-processing

Perform Below Tasks to complete the assignment:-

Tasks:-

1.Download the dataset: Dataset

2.load the data

If the DataSet already contains rows, the incoming data from the data source is merged with the existing rows. The Load method can be used in several common scenarios, all centered around **getting data from a specified data source and adding it to the current data container** (in this case, a DataSet ).

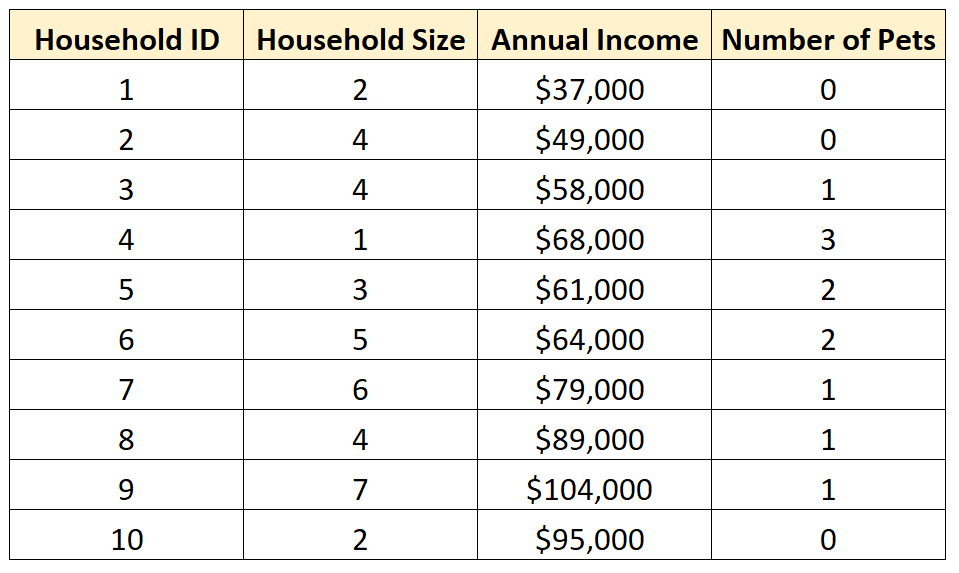
3. Visualizations.

● Univariate Analysis

The term **univariate analysis**refers to the analysis of one variable. You can remember this because the prefix “uni” means “one.”

The purpose of univariate analysis is to understand the distribution of values for a single variable.

For example

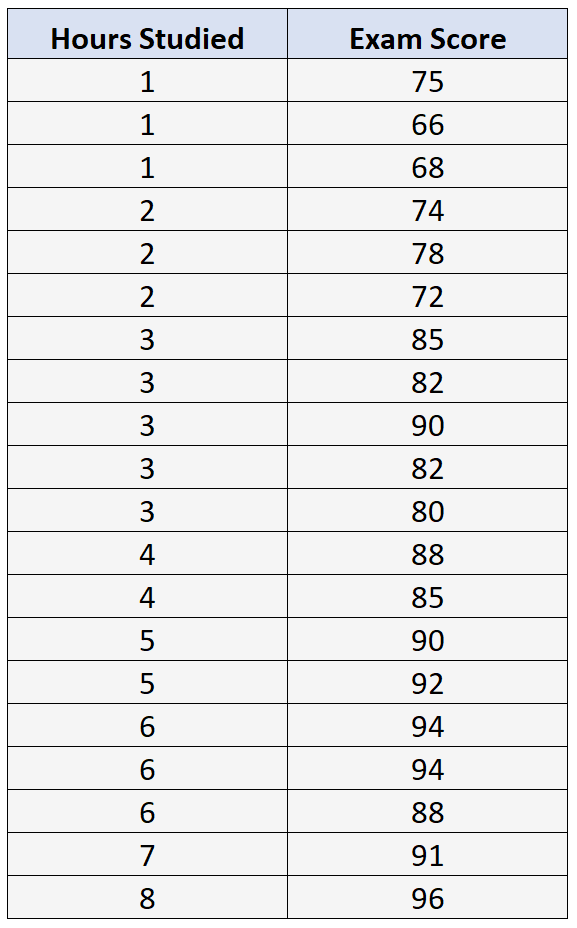


● Bi - Variate Analysis

The term **bivariate analysis**refers to the analysis of two variables. You can remember this because the prefix “bi” means “two.”

The purpose of bivariate analysis is to understand the relationship between two variables.

**(1)** Hours spent studying and **(2)** Exam score received by 20 different students:



● Multi - Variate Analysis

Multivariate analysis (MVA) is **a Statistical procedure for analysis of data involving more than one type of measurement or observation**. It may also mean solving problems where more than one dependent variable is analyzed simultaneously with other variables.

4. descriptive statistics on the dataset.

Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables

5. Handling the Missing values.

The problem of missing value is quite common in many real-life datasets. Missing value can bias the results of the machine learning models and/or reduce the accuracy of the model. This article describes what is missing data, how it is represented, and the different reasons for the missing data. Along with the different categories of missing data, it also details out different ways of handling missing values with examples.

6. Find the outliers and replace the outliers

outliers are values in data that differ extremely from a major sample of the data, the presence of outliers can significantly reduce the performance and accuracy of a predictable model.

The measure of how good a machine learning model depends on how clean the data is, and the presence of outliers may be as a result of errors during the collection of data, but some of this extreme values may be valid and legitimate

7. Check for Categorical columns and perform encoding.

**Hash Encoding** represents the categorical data into numerical value by the hashing function. Hashing is often used in data encryption or data comparison, but the main part is still similar — transform one feature to another using hashing function

8. Split the data into dependent and independent variables.

The independent variable is the cause. Its value is independent of other variables in your study. The dependent variable is the effect. Its value depends on changes in the independent variable.

9. Scale the independent variables

Commonly, we scale all the features to the same range (e.g. 0 - 1). In addition, remember that all the values you use to scale your training data must be used to scale the test data. As for the dependent variable y **you do not need to scale it**.

10. Split the data into training and testing

The simplest way to split the modelling dataset into training and testing sets is to **assign 2/3 data points to the former and the remaining one-third to the latter**. Therefore, we train the model using the training set and then apply the model to the test set.