Name: Sanket Patil Assignment No: 10
Roll no: COTB031

Title: Download the Iris flower dataset, compute and summarize statistics for features of dataset Perform data visualization and plot histogram as well as boxplot for the same.

problem statement: Data Visualization III
Download the Iris flower dataset or any other dataset into a dataframe.

1. List down the features and their types (e.g numeric, nominal) available in the dataset.
2. create a histogram for each features in the dataset to illustrate the features the distribution
3. create a boxplot for each feature in the dataset.

objectives: learn classification techniques how to plot histogram and boxplot for given dataset.

pre-requisites: R- programming Rstudio

Theory: Let's design a basic data analysis program in R using Rstudio by utilizing the features of R studio by ~~utilizing~~ to create some visual representation of their data

How to Install Rstudio?
we have to follow three basic steps in same order to run R and R studio on your system.

- Install R (Download the binary setup file for R from the following link)
- Install Rstudio
- Install R Packages.

following steps will be performed to achieve your goal.

- Downloading/importing data in R
- Transforming Data / Running queries on data.
- Basic data analysis using statistical averages.
- Plotting data distribution

Typical data analysis process

Data analysis deals with collecting, inspecting, cleansing, transforming and modelling data to glean valuable, insights and support better decision in an-organization.

Data Exploration:- Having Indentified business problem, a data analyst has to go through the data provided by client to analyse the root cause of the problem.

Data preparation:- This is the most crucial step of the data analysis process wherein any data anomalies. with the data have to be modelled in the right direction.

1 Data Modelling:-

The modelling step begins once the data has been prepared modelling is an iterative process wherein the model is run repeatedly for improvement.

2. Validation-

In this step, the model provided by the client and the model developed by the data analyst are validated against each other to find out if the developed model meets the business requirement.

3. Implementation of the Model and Tracking.

This is final step of the data analysis process wherein the model is implemented in production and is tested for accuracy and efficients.

1. Importing Data in R studio:-

for this tutorial we will use the sample census data set Acs. There are two ways to import this data in R. one way is to import the data programaticaly by executing the following command in the data console window of Rstudi

The second way to import the data set into R studio is to first download it is onto you local computer and use the import dataset featurer of R studio.

1. click on the import dataset button in the top-right section under the environment tab. select the file you want to import and then click open.
2. After setting up the preferences of seperator, name and other parameters click on the import button.

2. Transforming Data.

once you are done with importing the data in R studio. you can use various transformation features of R to manipulate the data.

- To Access a particular column, Ex. age husband in our case.
- To run some queries on data, you can use the subset function of R. Let's say I want those rows from the dataset in which the age husband is greater than age-wife.

The first parameters to the subset function is the dataframe you want to apply that function to and the second parameter is the boolean condition that needs to be checked for each row to be included or not.

Getting statistical Averager from data:

Following functions can be used to calculate the averagas of the dataset.

- For mean of any column, run: mean(acs$age_husband)
- Median, run: median(acs$age_husband)
- Quntite, run : quantile(acs$age-wife)
- variance, run: var(acs $age wife)

4. Plotting Data

A very liked featurer of R studio is its built in data visualizer for R. Any data set imported in R can visualized using the plot and several other functions of R.

Where s is the subset of the original dataset and type='p' set the plot type as point. you can also choose thae and other change variables to Letc.

To draw a Histogram of dataset, you can run the command nist(acs$number_children)

Similarly for bar plots, run the following set of command.

Conclusion:

Thus we have learned how to use features of R studio when learning data visualization. How to Create histogram for each features of dataset, boxplot etc