SEMESTER PROJECT

Submitted to:
Sir Irfan

Submitted from:
Aleezay Amir 2020-SE-27

Course Name:
Introduction to Data Science

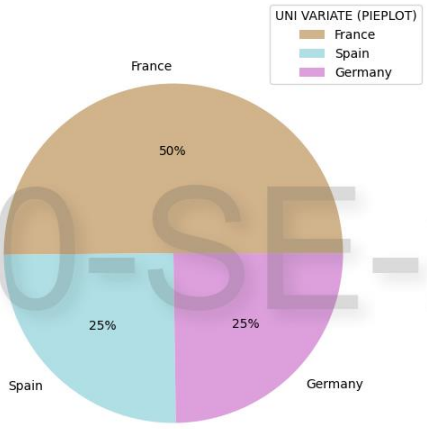
EXPLORATORY DATA ANALYSIS
(DATA_SET_1)

Department of Computer Science

UNIVERSITY OF ENGINEERING AND TECHNOLOGY, NEW CAMPUS, LAHORE

FIGURE # 01

UNIVARIATE (PIE PLOT)



EXPLANATAION:

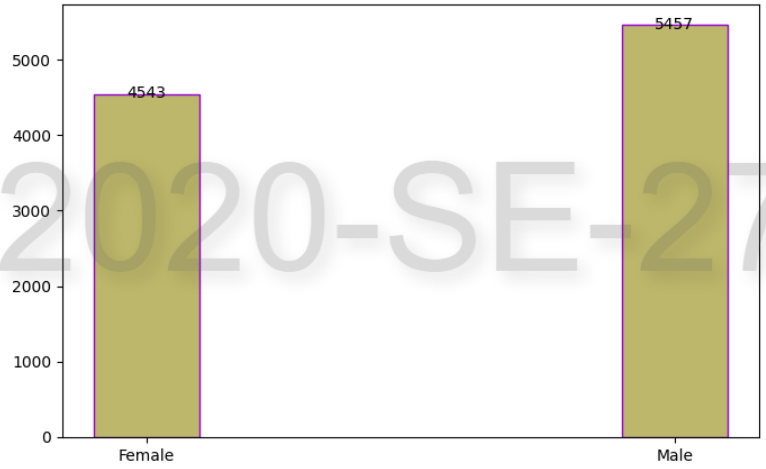
This graph Pie Plot is plotted on the dataset of those customers who wanted to withdraw their accounts from the bank due to the loss that has occurred due to some reasons. The pie plot consists of the categories based on the countries or nationalities of the customers from all over the world, but only three kinds of countries were using the services of this bank including Spain, France and Germany. The colors represent different countries in the pie plot.

When we study the graph, it is not hard to see that the half of the bank account users were from France i.e. 50 percent. Similarly, if we have a closer look at the data analysis we can infer that the other half of the bank accounts of the customer are from both Spain and Germany, divided in equal halves i.e. 25 percent each.

Therefore, the visual representation of the division of the number of accounts is made easy by using the pie plot for the following dataset.

FIGURE # 02

UNIVARIATE (BAR PLOT)



EXPLANATAION:

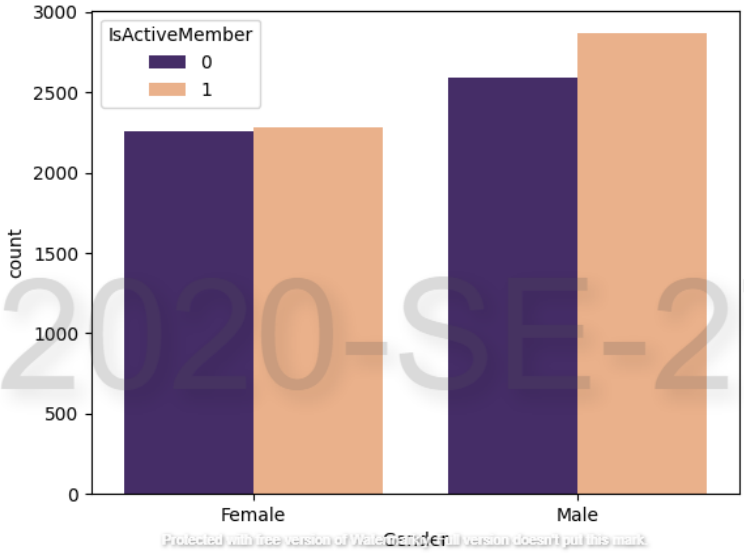
While talking about the Gender of the Customers who were having their bank accounts in the bank, this bar plot or bar chart illustrates the total number of Male and Female account holders (customers). As this is a univariate graph so the gender itself is taken as an independent variable and the number of the gender specified customers are plotted.

The bar plot clearly explains the difference between the number of bank account holders on the basis of their gender. The bar plot for the male customers goes high till around 5000 plus customers whereas the female customers are at the number of 4543.

Conclusion we get from the graph is that the Male customers were more in number than the female customers who wanted to withdraw from their bank accounts. So, if your data is categorical but you want to get a count against that the data, this bar plot can be proved beneficial in case of Univariate data analysis.

FIGURE # 03

BIVARIATE (COUNT PLOT)



EXPLANATAION:

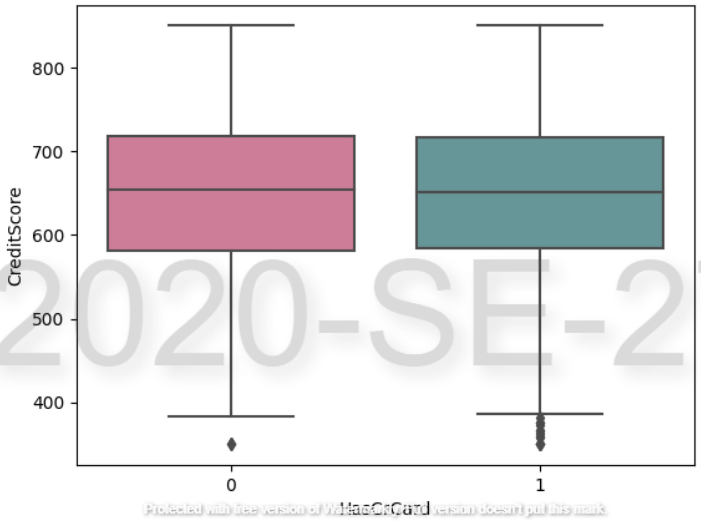
Talking about the Bivariate data analysis on the given dataset, this count plot explains the relation between the Gender of the customers and on the other hand, the customers who were active member of the bank accounts and who weren't. Both the data variables are of categorical nature.

Instead of plotting the Dependent variable of the y-axis we take the advantage of using **hue** in the count plot because it provides a count of the customers on the y-axis as obvious by the name of the graph. Both the genders male and female are taken and the hue of two different colors shows that which gender was more likely to be an active member in the bank.

The results clearly show that the more male customers were active member in comparison to the female customers. And side-by-side if we talk about the customers who weren't active members, still the plot of male customers is higher that the female customers. But still there is a difference between being an active member and for not being an active member in both the genders by some ratio.

FIGURE # 04

BIVARIATE (BOX PLOT)



EXPLANATAION:

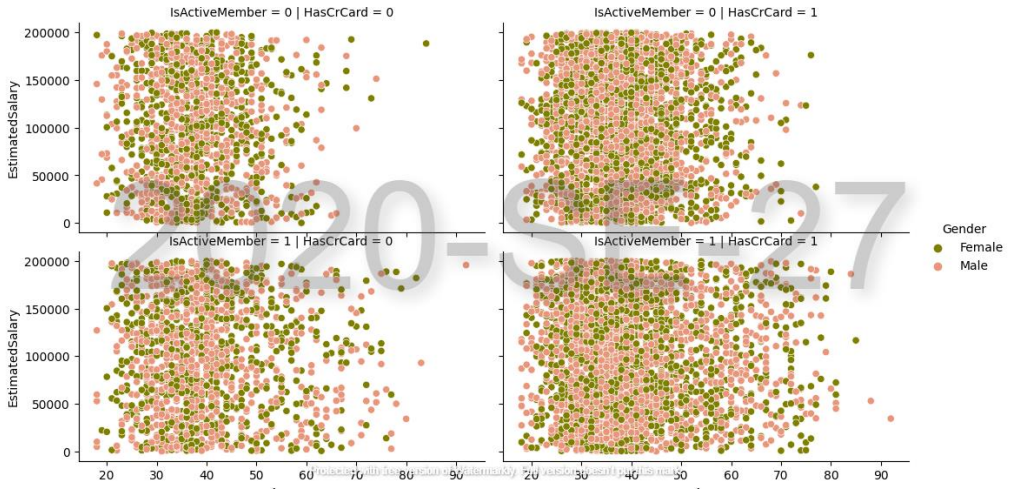
This Box plot explains the relation between two categorical variables. On x-axis we take the data of those customers who have credit card and on the y-axis we take the total credit points a customer has.

From the plot it is obvious that the customer's most of the data lies between the range of 600 to 700 both for the customers who have credit cards and for those who don't have credit cards. For box plot, the plot is divided into 4 quartiles. For this case, the 25 percent lies below 600 and the two upper boxes contains the 50 percent of the data i.e. 25 percent each. Similarly, the other 25 percent is splitted into 20 percent which is above the two boxes and the remaining 5 percent is represented in the form of outliers.

For the customers who do not have a credit card the outlier is the only single entity in this box plot. But for the customers who have a credit card the outliers are more than one which lies under the range of box plot.

FIGURE # 05

MULTIVARIATE (SCATTER PLOT)



EXPLANATAION:

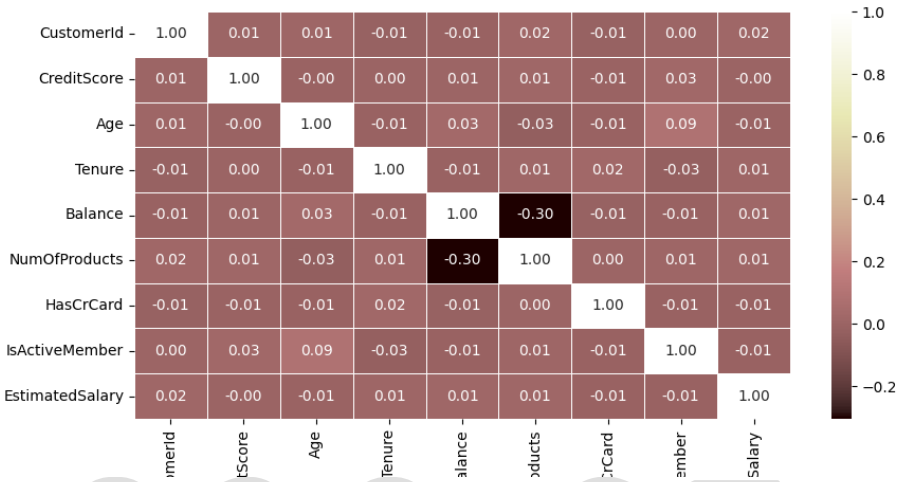
This Scatter plot is divided into 4 parts for the further clarification of the results plotted on the given dataset. On the y-axis the Estimated Salary is taken while on the x-axis, two categorical variables are taken i.e. The customers who are active members or not and the Customers who have a credit card or not. As this is a Multivariate graph, the variable Gender is taken side-by-side to explain the functioning of the scatter plot.

As obvious, there are two genders and they are represented with the help of different colors. With the green color the female customers, and with pink color for the male customers, overall customers are depicted who do not have a card and simultaneously are not an active member on the upper-left corner plot.

For the customers who have a credit card but were not an active member are illustrated in the upper-right corner. While talking about the customers who are active members but do not have accredit card are shown the graph at the lower-left corner and lastly, the customers who were active members as well as who had a credit card are shown in the lower-right corner.

FIGURE # 06

MULTIVARIATE (HEAT MAP)



EXPLANATAION:

This heat map explains the correlation between all the columns of the datasets taken as variables in this heat map. The x-axis and the y-axis are plotted as the variables each compared with another variable to create a relation between them and define it likewise.

We can also take a limited number of variables on the axis to make it less complicated by making a list of variables. A color heat map guide is given at the right side of the heat map to explain the color of the map. The lightest color is the greatest value and the darkest color represents the lowest value.

The exceptional black boxes in the heat map represents the values which fall beyond the temperature guide value i.e. -0.3. The interesting point to know is the diagonal values which are 1 and are represented as white boxes. These values are 1 because of the similar variable column and rows merge at the given point.

2020-SE-27