Report

Abstract:

We have a data of automobile in which I analyze the features, getting the insights of data using different techniques and also handle the business problems and it helps us to make future decisions. We also have data of go digit. Go Digit is the private bank that have many products and we analyze the data using python libraries and also take insights from data.

1-What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables)

Firstly we explain the automobile dataset there are 1581 rows in the dataset and there are 14 columns in this dataset and total size of dataset is 22134

Age: This column tells us the age of all customers

Gender: This column tells us the gender of all the customers

Profession: This column tells us the profession of customer that is going to purchase the car

Marital Status: This column tells us is the customer is married or still he is single

Education: This column tells us the level of education of customers

No of dependents: This column tell us total the number of dependents

Personal Loan: This column tells us the total personal loan of customers who is going to buy a car

House loan: This column tells us the total house loan of customers who is going to buy a car

Partner Working: This column tells us the partner or spouse of customer is working or not

Salary: Salary of customers

Partner salary: This column tells us the salary of customer spouse

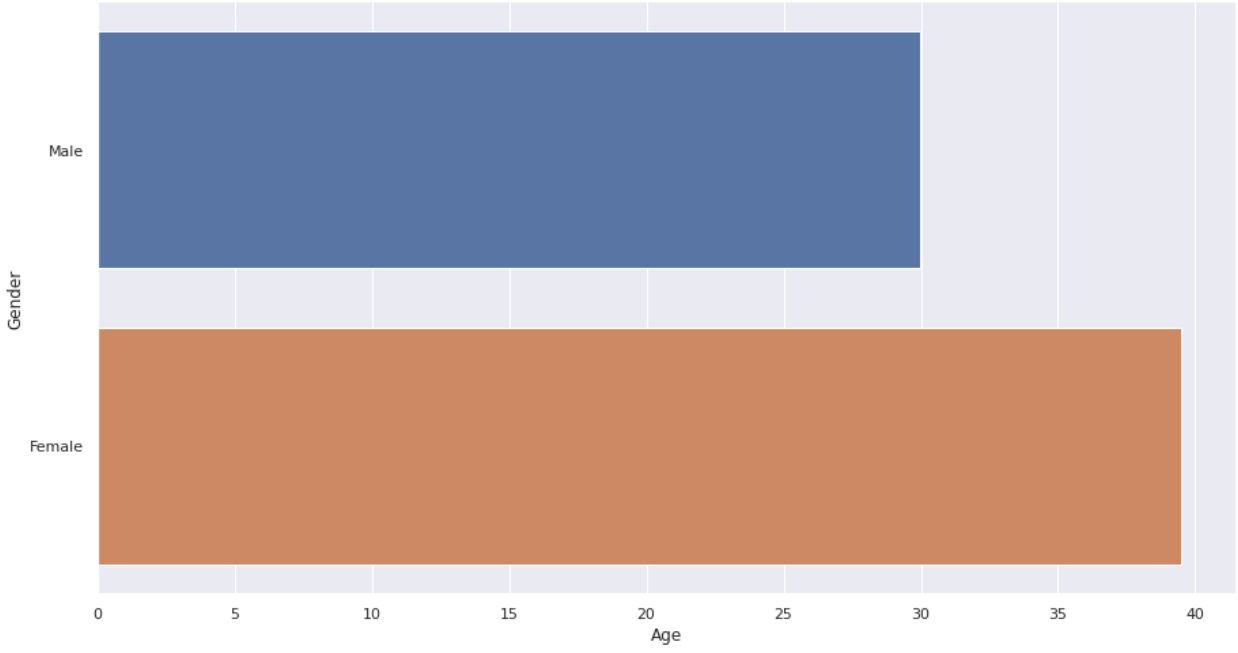
Total Salary: Total salary of customers and his/her spouse

Price: Price of Car that customer trying to buy

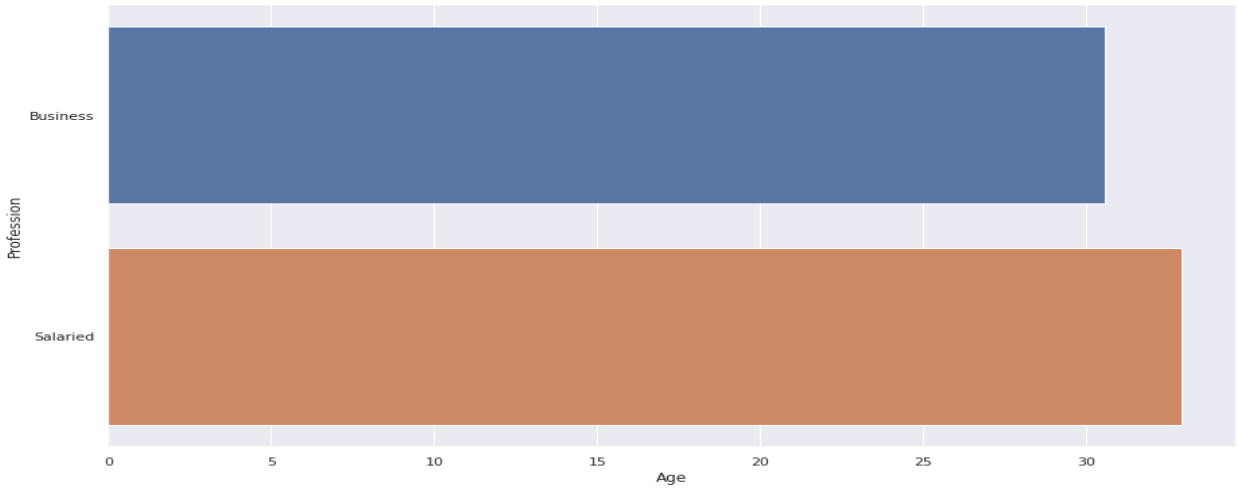
Make: The category of car

2-Take a critical look at the data and do a preliminary analysis of the variables. Do a quality check of the data so that the variables are consistent. Are there any discrepancies present in the data?

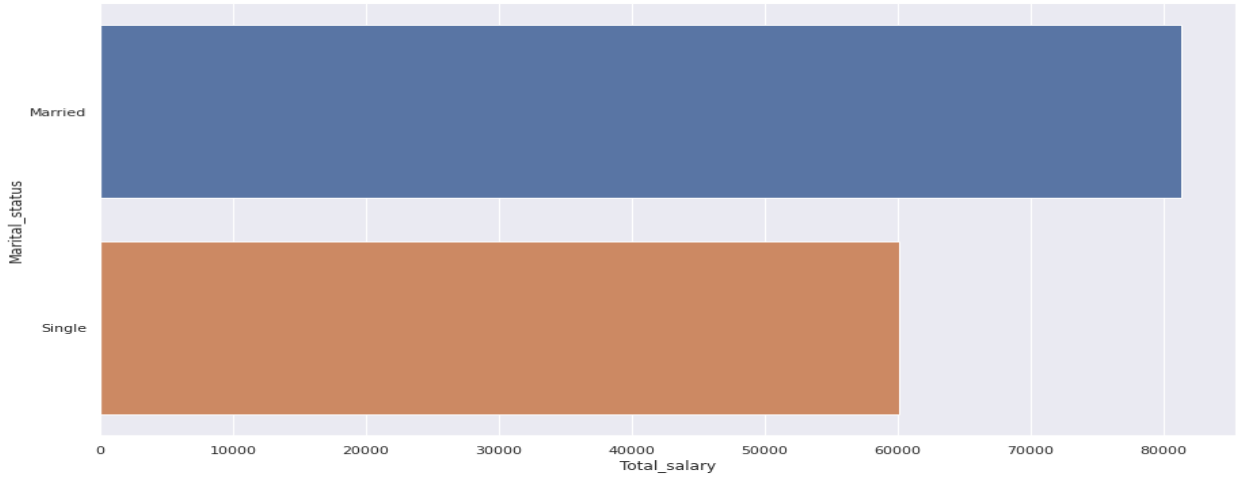
So firstly of all we analyze the age of Male and female. We see that the age of female is larger then age of male. Normally this is unusual behaviour of any dataset because this is automobile dataset and normally male customer age is greater the female but in this dataset the age of female customer is upto 40years and age of male is upto 30 years so there is clear difference between them.



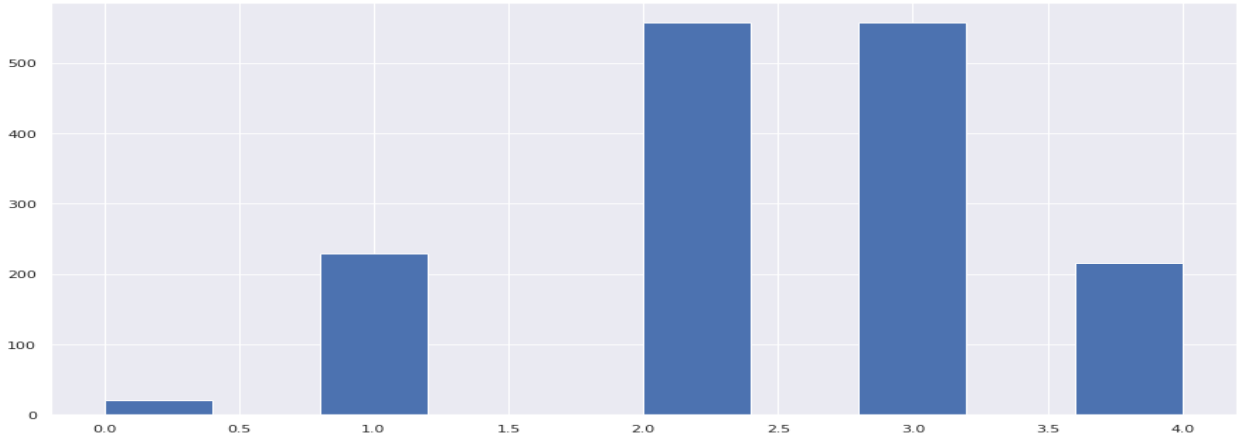
Similarly when we see the age of customer profession its clearly shown that average salaried customer buy car at the age of 40 and business man buy car at the age of 30 so business man earn more than salaried person and this is obvious thing



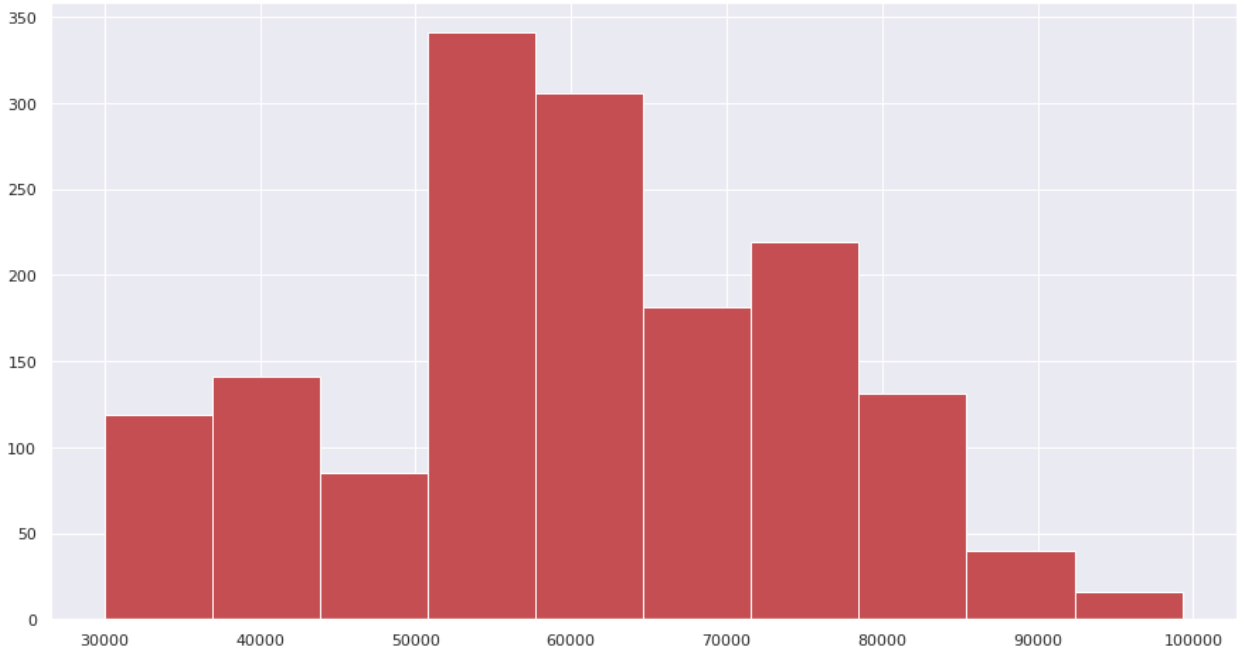
When we see the person who is married earn more salary as compared to customer who is still single. Because sometime married customer spouse is also doing some job so this thing is explain in the feature that name is partner working. So thatswhy married people spent a good life and also buy a car on the start of the carrier.



This is the distribution of number of dependents so the is almost normal and average number of dependents of each customer is 2 or 3 . So if the customer have more dependent then he or she have more responsibility and not able to save payment so thatswhy he or she not buy a car in the early age of his carrier.

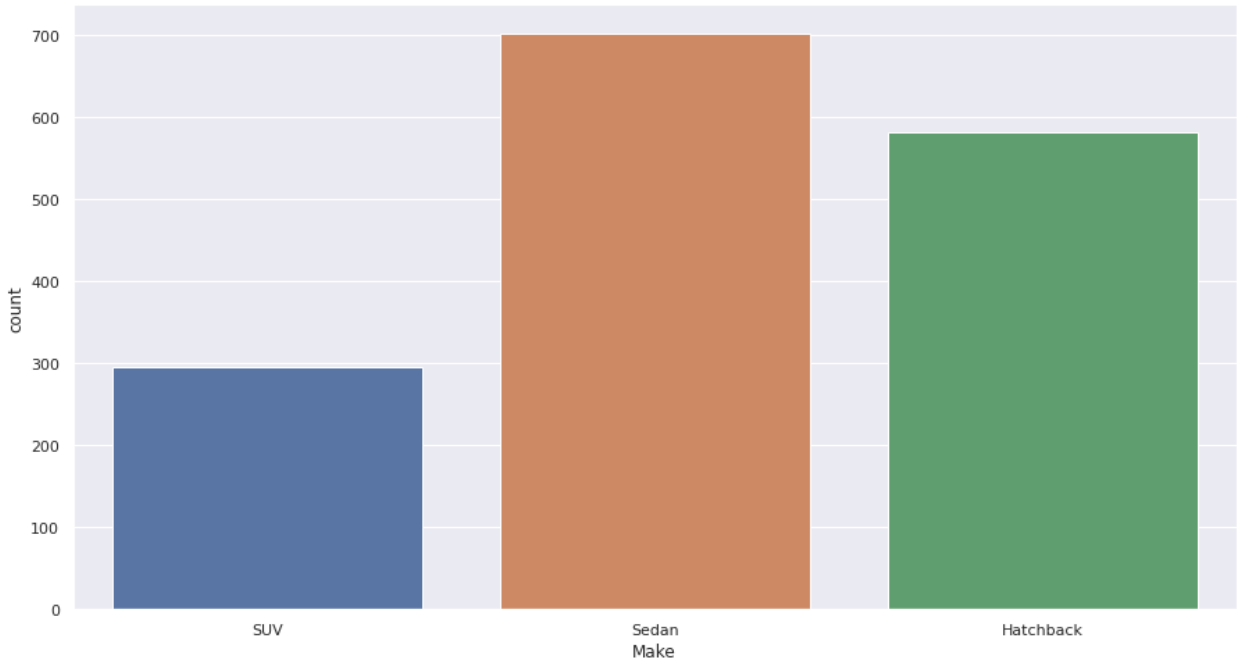


This is the distribution of salary so the is binormal because there is two peaks in the distribution and average salary of each customer is 5000 to 6000. So if the customer have good salary almost greater than 70000 then it helps to buy a car

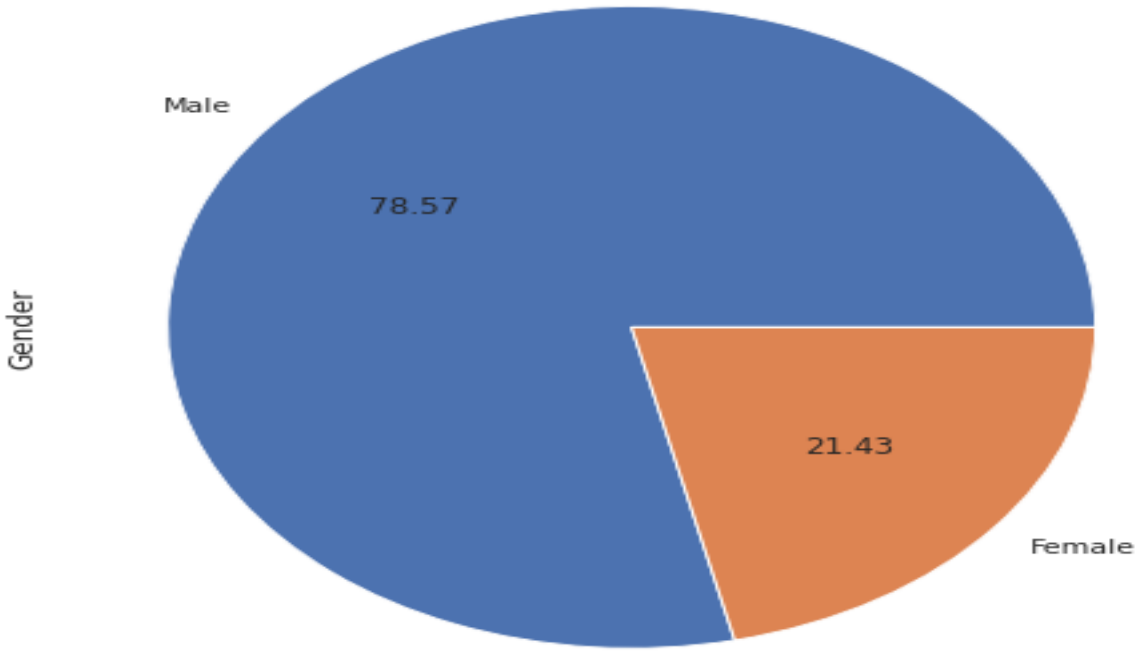


3-Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business

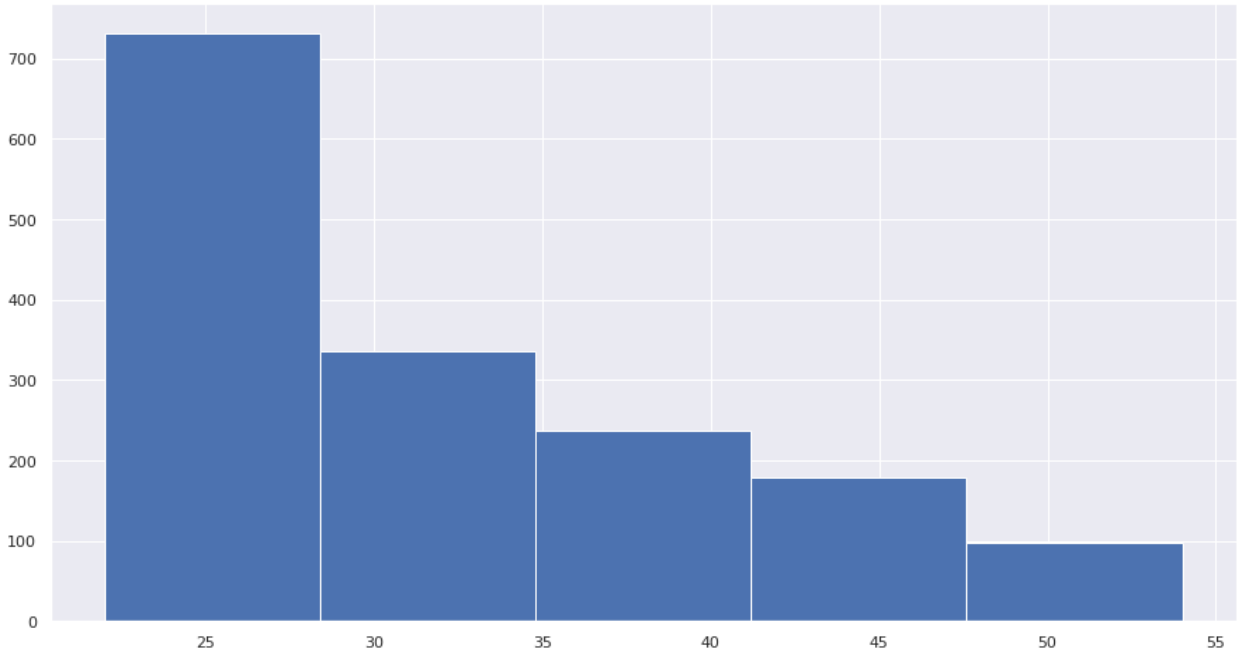
In below plot we do univariate analysis of Make column in which I plot the distribution of Make column and 700 customers buy a sedan and 300 customers SUV 590 customer buy hatchback cars.



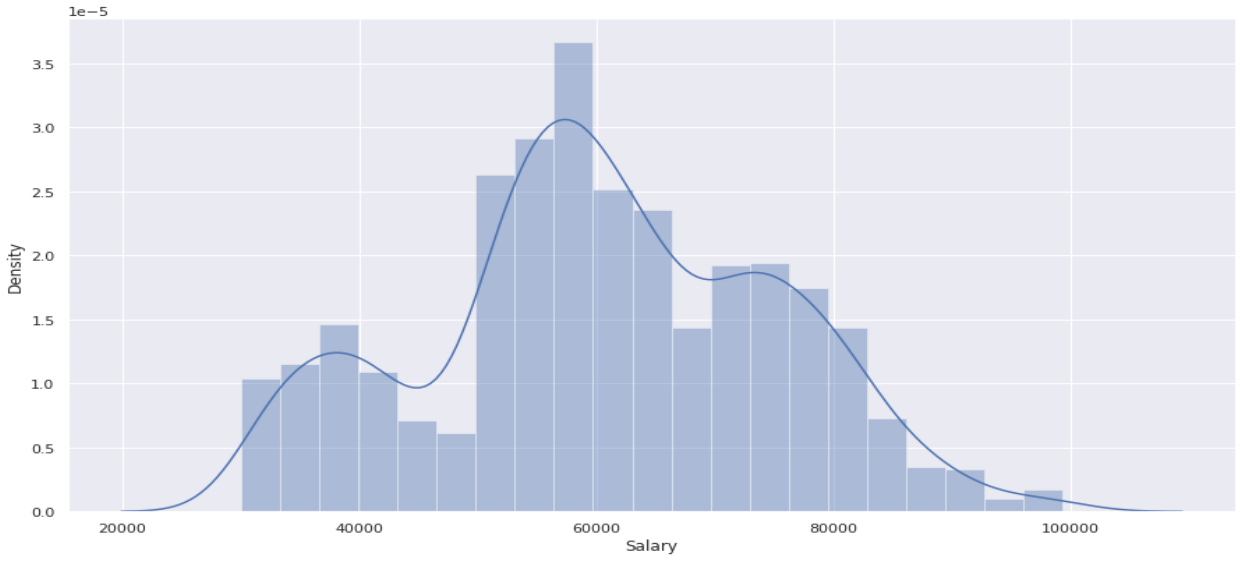
In below chart we see the distribution of gender column and we see than 77% customer is male and 23% is female so its mean most of this data contains male and there is only few females



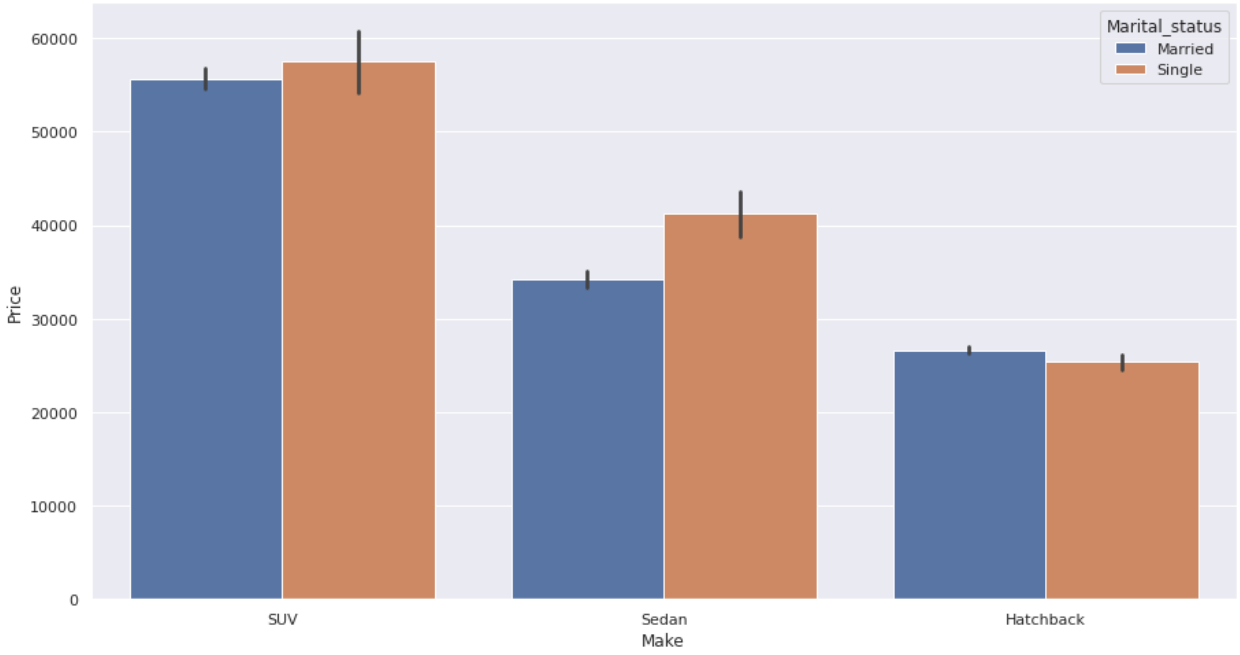
In below we see the distribution of age column the distribution of age column is right skewed its mean data is not normal so the most of the person is around 25 so most of the customer is young and very interested to buy a car as compared the customer who is aged



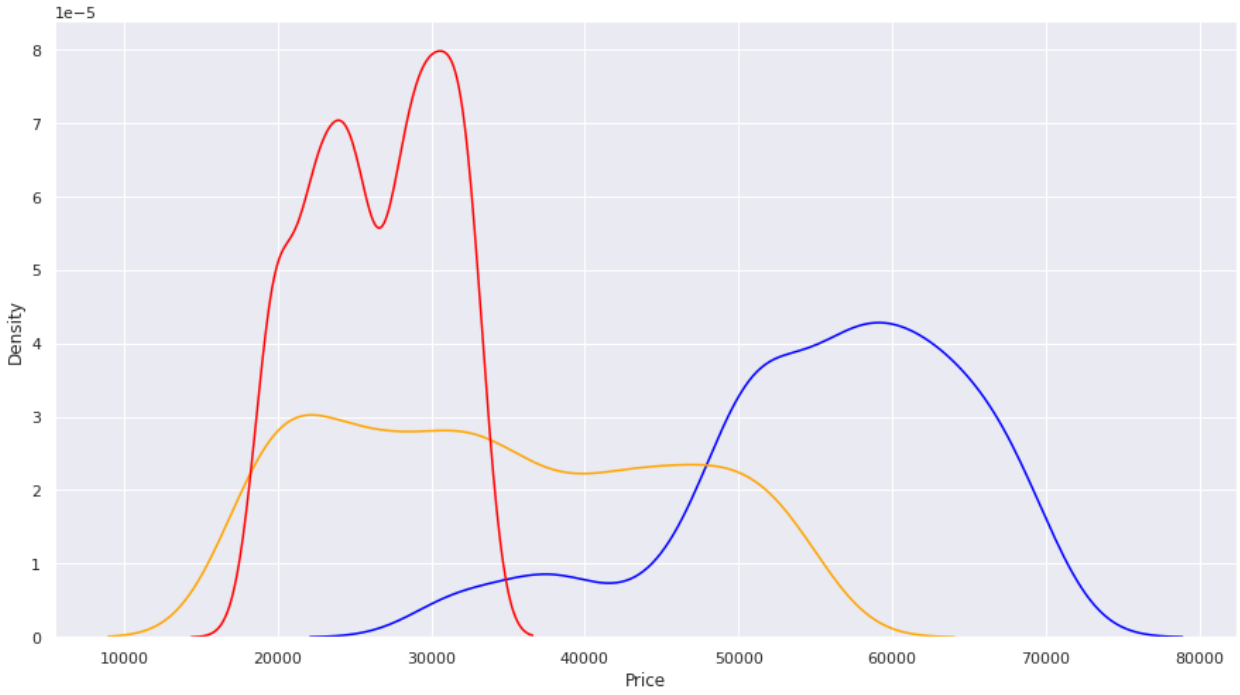
This is the distribution of salary so the is binormal because there is two peaks in the distribution and average salary of each customer is 5000 to 6000. So if the customer have good salary almost greater than 70000 then it helps to buy a



After that we see the price of all category of cars, this is a bivariate analysis. We clearly see that. Married people buy a low price car and single customers who do have spouse who save money and buy a car that price is high. So we see this trend in all categorize of car and most of the customers want to buy a SUV car .



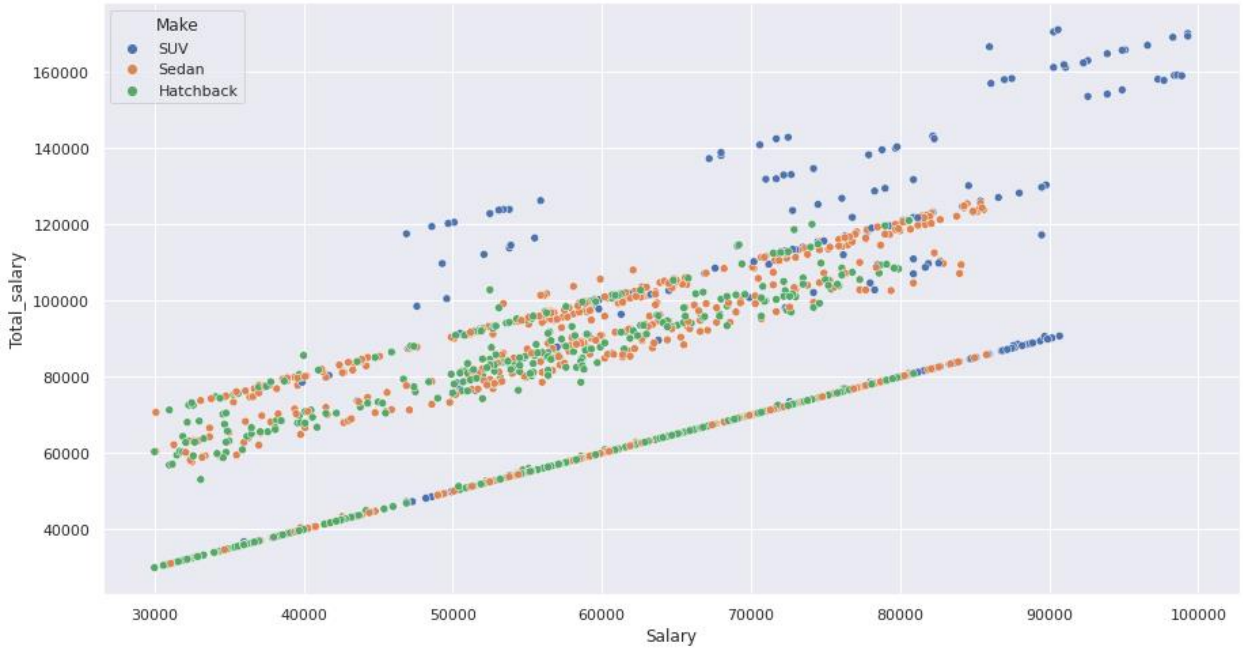
In below graph we plot the distribution of price of all cars. Blue line is the distribution of SUV car and orange line is the distribution sedan car and red line is the distribution of hatchback car. There is major different in the price of these cars. The price of SUV is much more deviate as compared to other cars and price of hatchback is very limited normally this car is buy by customers in the range of 20000 to 30000



4-Understanding the relationships among the variables in the dataset is crucial for every analytical project. Perform analysis on the data fields to gain deeper insights. Comment on your understanding of the data.

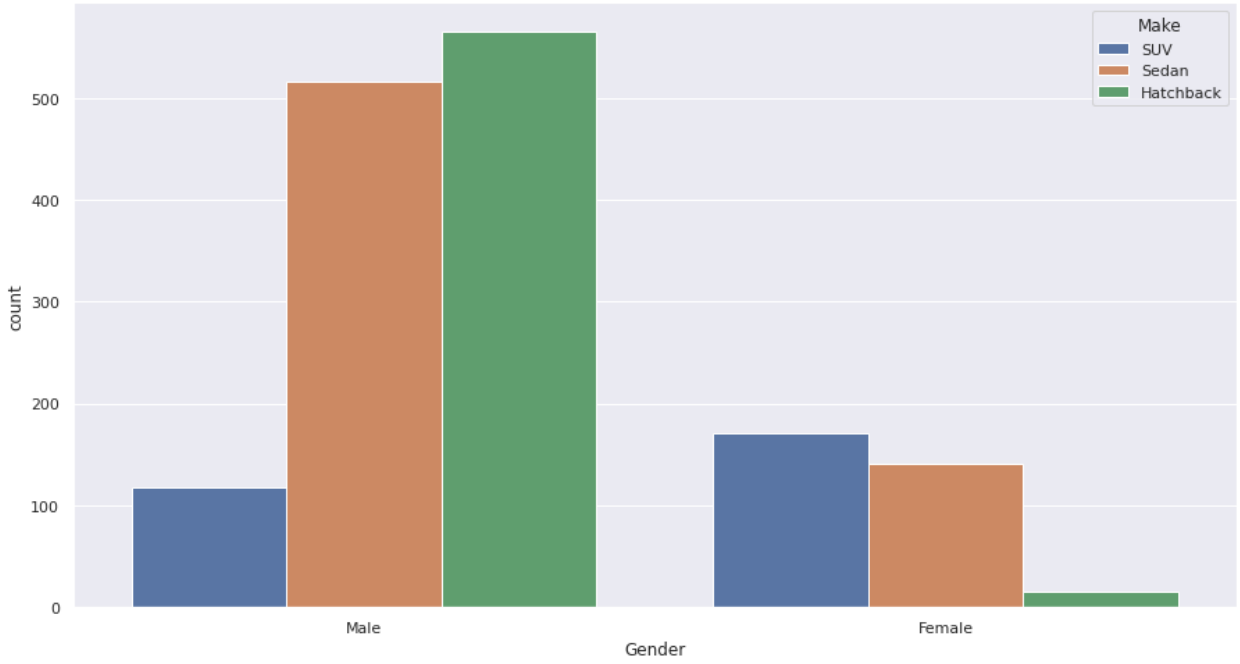
In this question we see the correlation of two features the one is partner salary and other is price. And there is no correlation between these two features. There is zero correlation between these features. Correlation normally tells us how much correlated two features are. The value correlation is between 0 to 1.

In this question we see the correlation of two features the one is salary and total salary. And there is correlation between these two features. There is some correlation between these features. Correlation normally tells us how much correlated two features are. The value correlation is between 0 to 1.



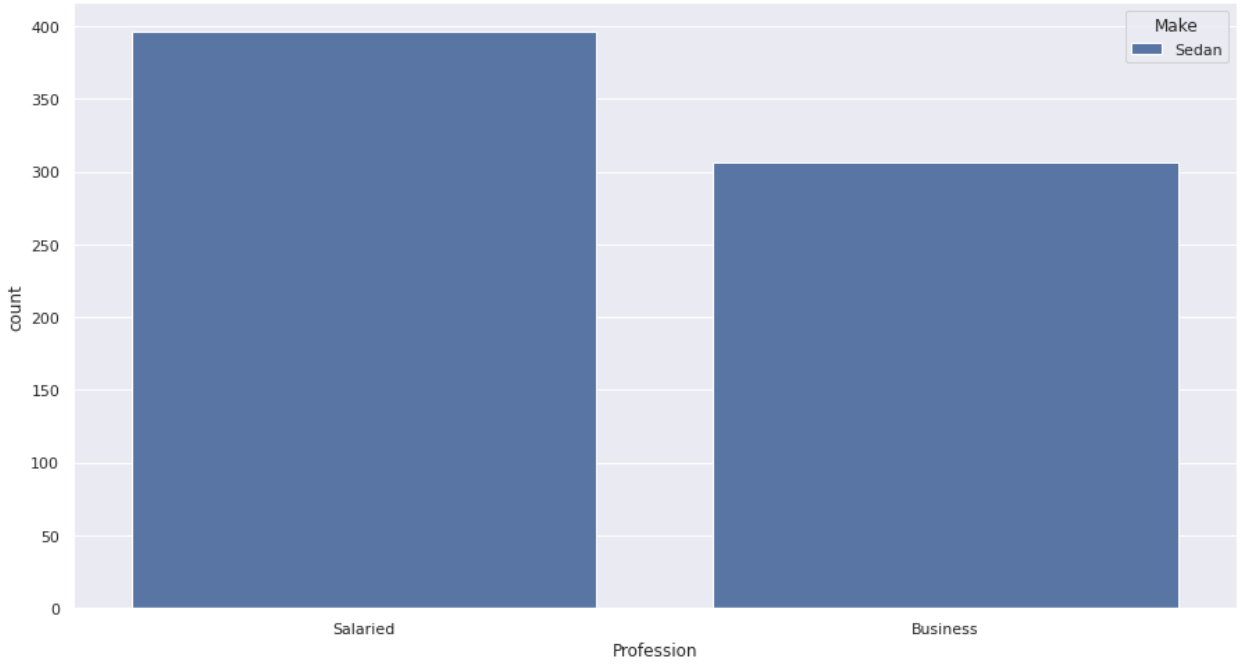
Steve Roger says "Men prefer SUV by a large margin, compared to the women"

Yes we see the result of this business questions by count plot and we conclude that Men prefer SUV by a large margin, compared to the women



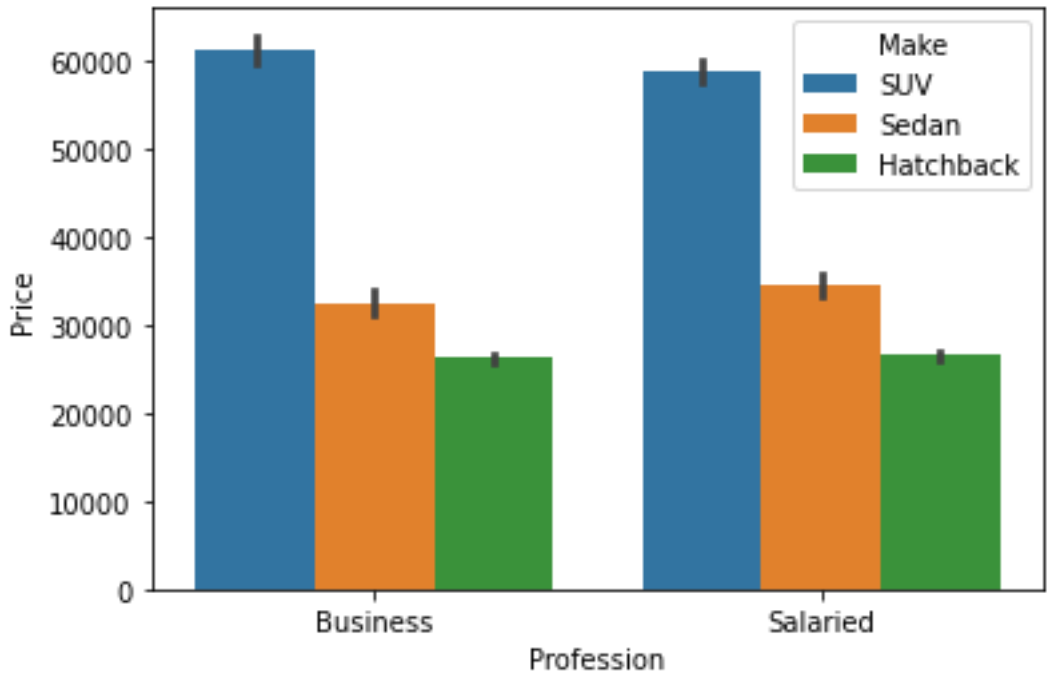
5-1-Ned Stark believes that a salaried person is more likely to buy a Sedan.

Yes we again plot the count plot of both salaried customers and business customers. I conclude Ned Stark believes that a salaried person is more likely to buy a Sedan



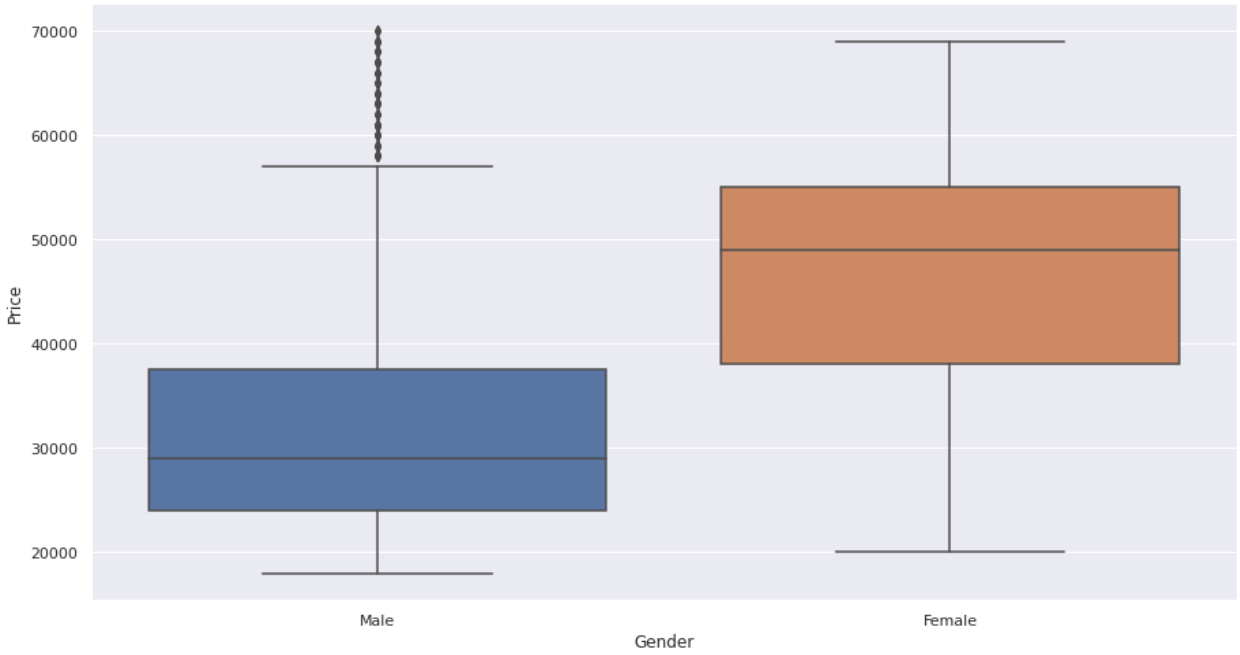
5-2-Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale.

No business male also interested in purchasing the SUV car over sedan and ratio of businessman customer who interested to buy a SUV car is more the salaried customers

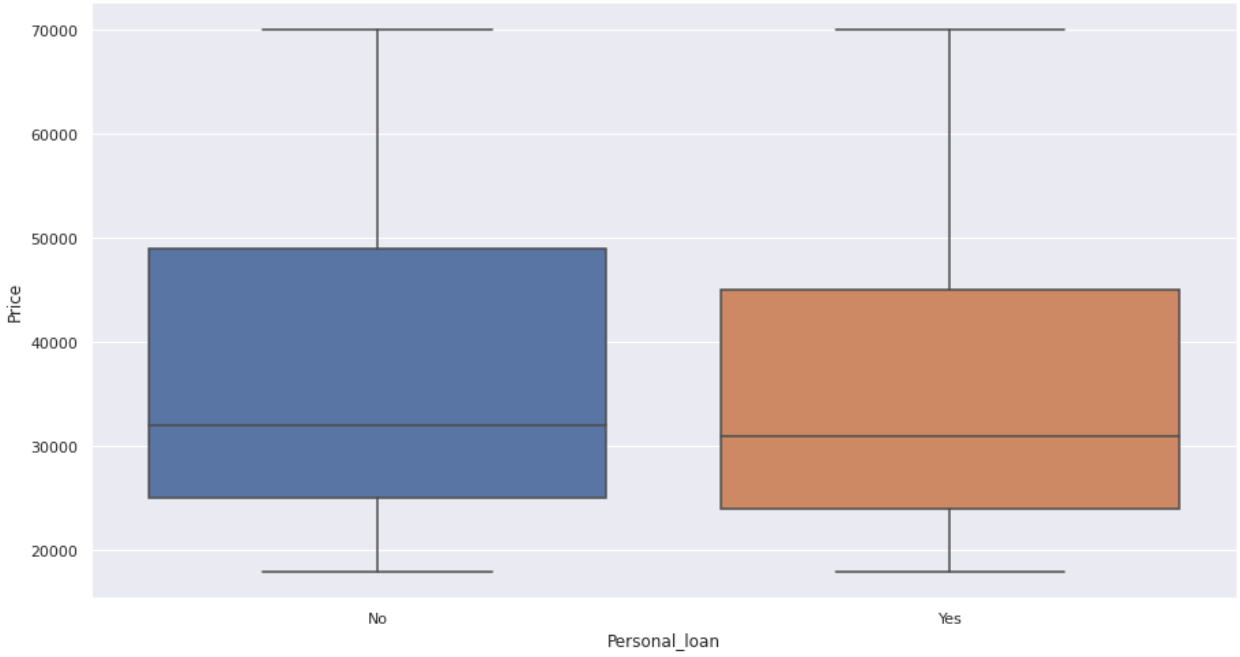


6-From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a Business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions. Give justification along with presenting metrics/charts used for arriving at the conclusions. ***F1) Gender ***F2) Personal_loan

When we discussed the amount spent on purchasing automobile then we get to know that price that is use by male and female to purchase a car. So we see that there are many outliers in case of male because some male spend so much amount and most of the male people spend amount in the range 27000 to 37000. The average spending of female is greater than male customers

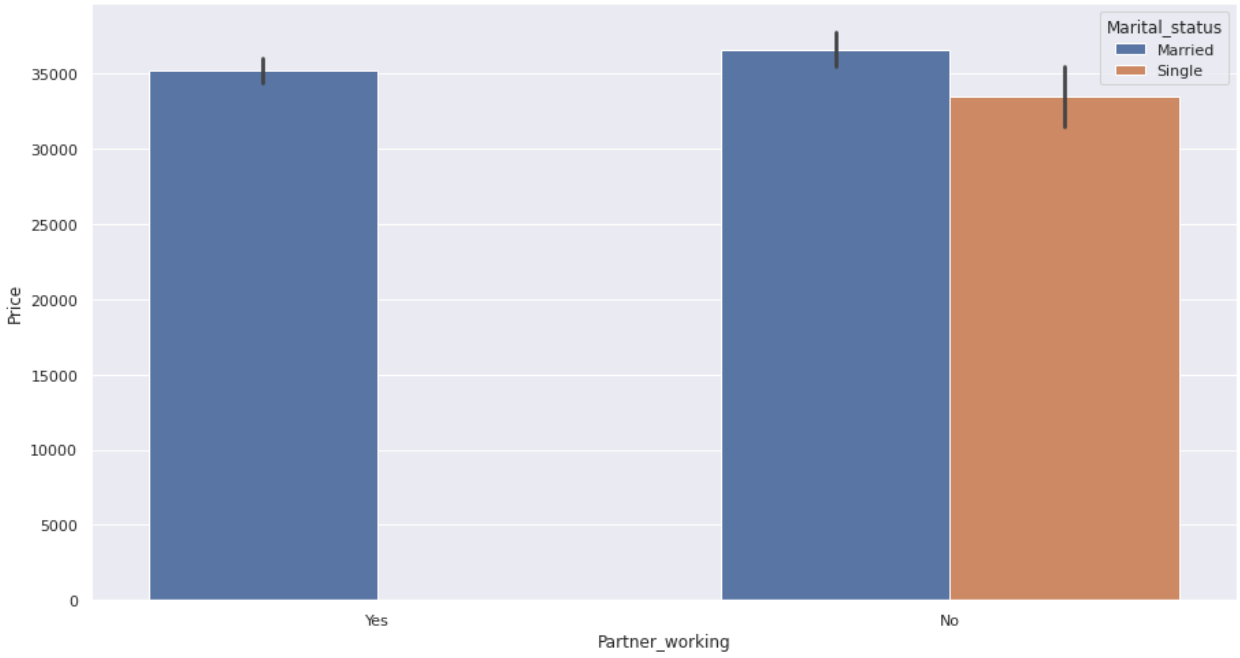


Those customers that don’t need loan spend more money as compared to those who don’t need personal loan. The customer who don’t need loan its mean they already rich and have a lot of money to purchase a new car so customer who don’t have personal loan purchase a high price car.



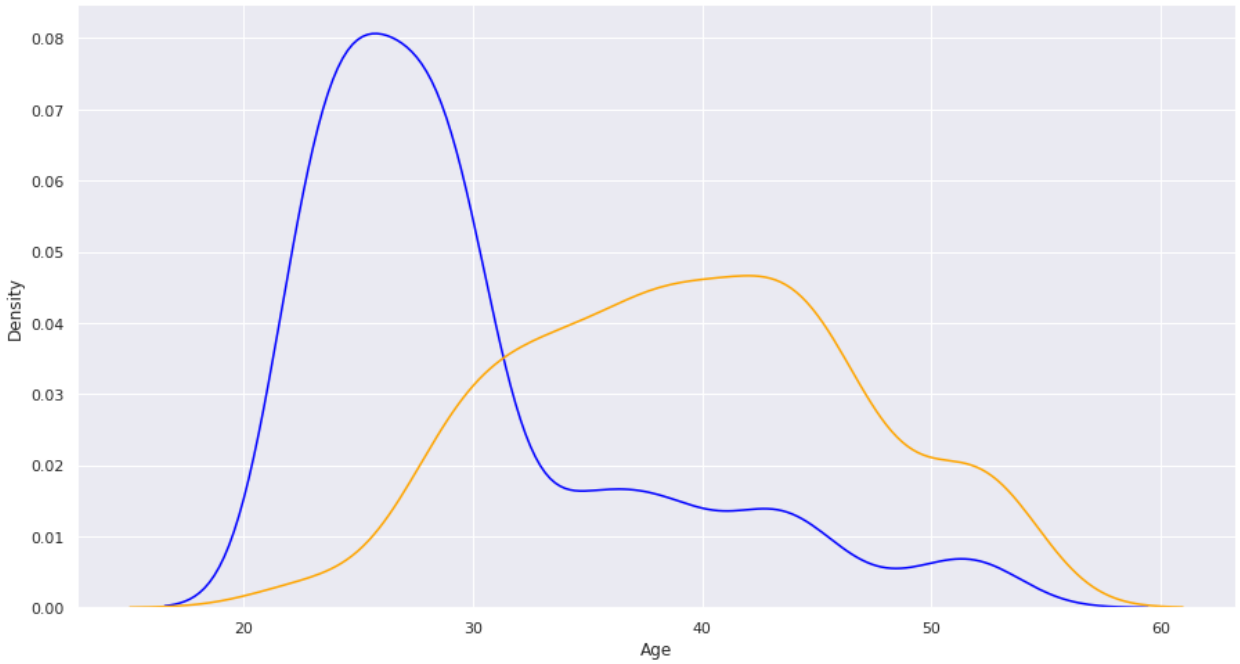
7-From the current data set comment if having a working partner leads to purchase of a higher priced car

No working partners don't leads to purchase a high priced car you clearly seen in below chart the blue bar on right is the married customer who's partner is not doing any job but still purchase higher priced car

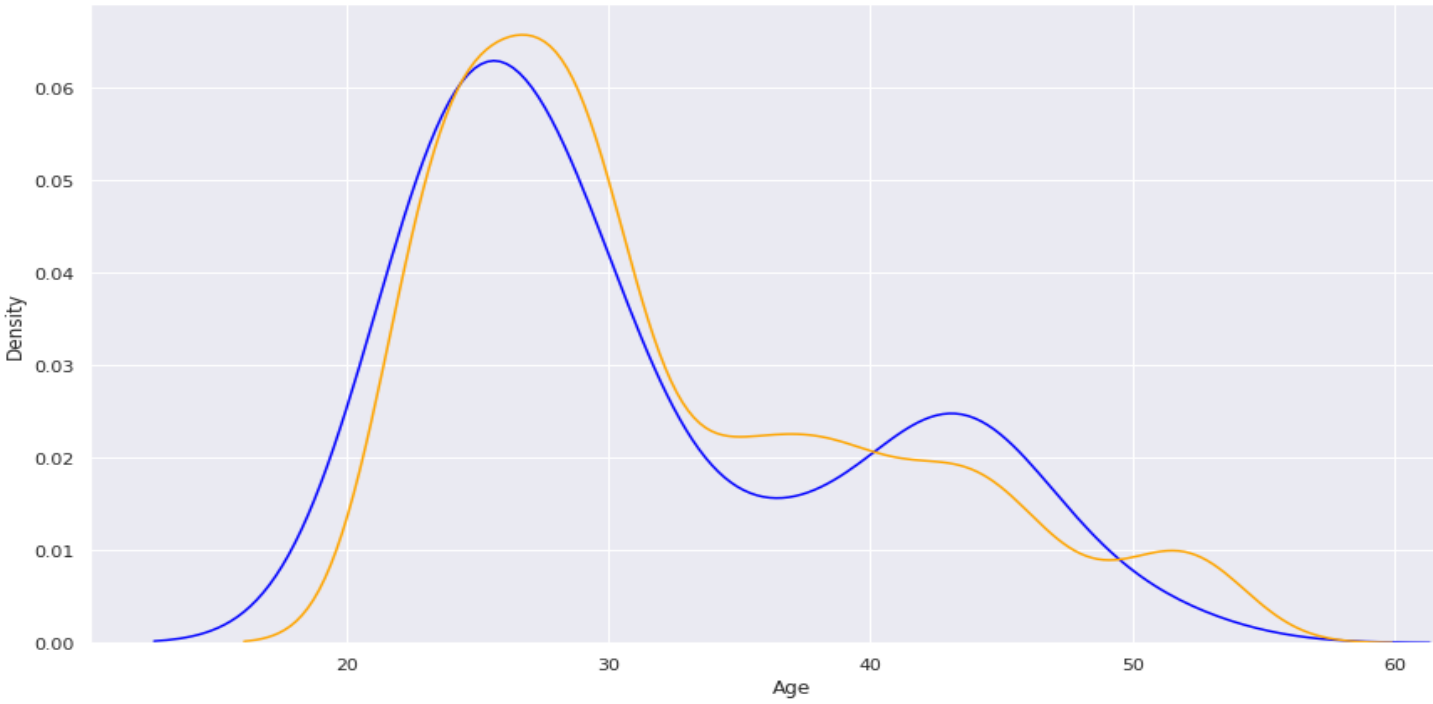


8-The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use Gender and Marital_status - fields to arrive at groups with similar purchase history.

In this section we improve the marketing strategy to send targeted information to different groups. So blue line is the line of age of male customers and orange line is the age of female so we clearly see that female customer is more aged then male customers and the distribution of male customers is also different as compared to female customers the distribution of age of male customers is right skewed and age of female customers is almost normal.



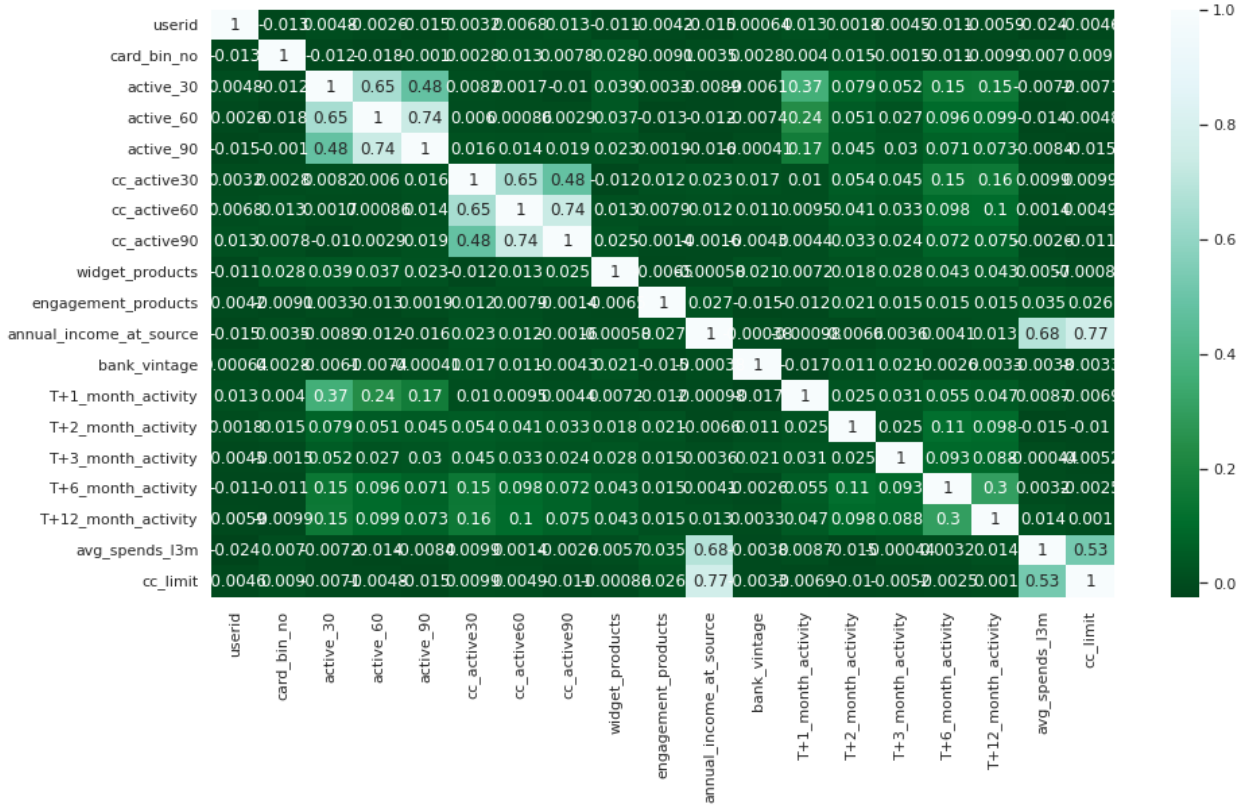
Similarly, So blue line is the line of age of married customers and orange line is the age of single customer so we clearly see that single customer is almost same then male customers and the distribution of male customers is also same as compared to female customers .



Problem-2:

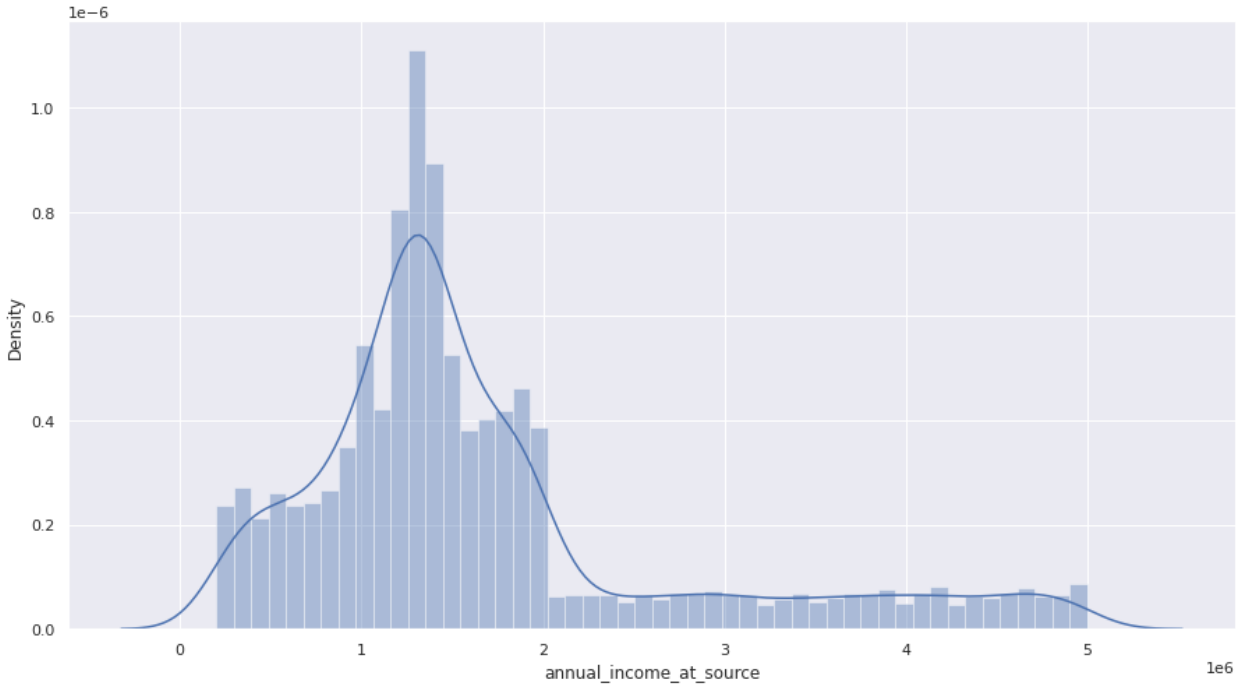
Framing An Analytics Problem Analyze the dataset and list down the top 5 important variables, along with the business justifications.

In this section we see the correlation between features using heatmap. The value of correlation between 0 to 1. If the value close to mean the feature is highly positive correlated if the value close to zero its means there is no correlation between these features and if the value is close to -1 then feature is highly negative correlated so we extract the correlated feature by this techniques. This technique is called feature selection using correlation**.**

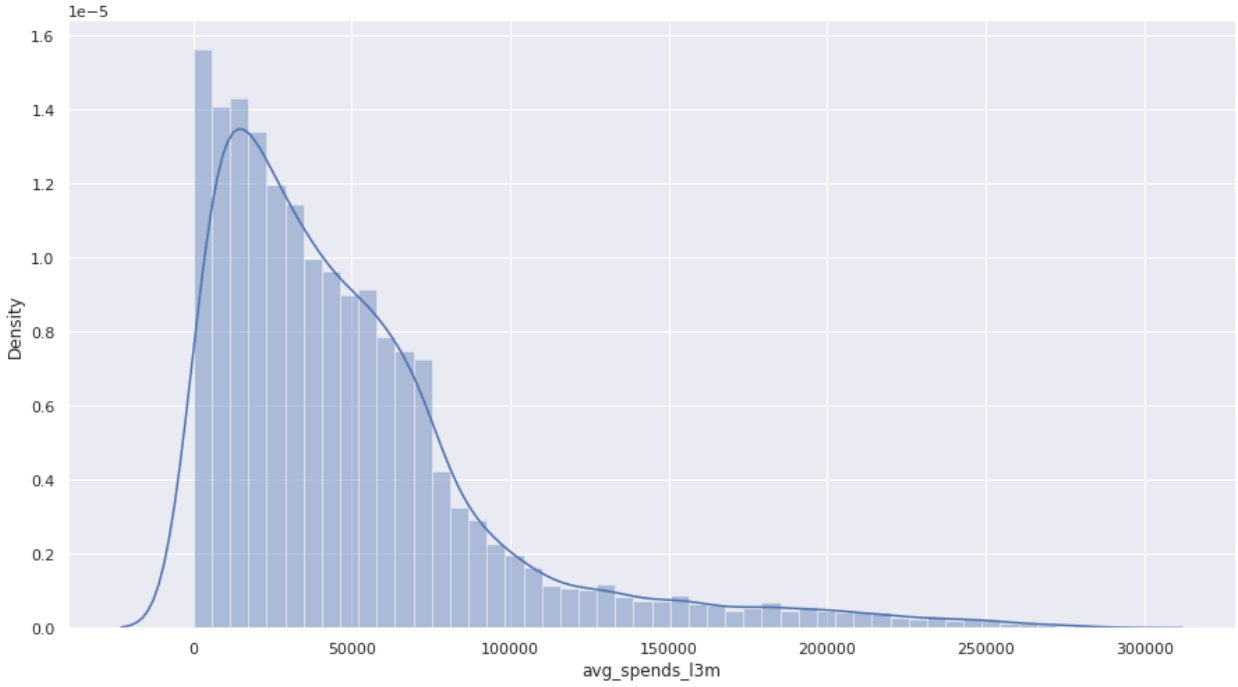


Then we select some useful features that is highly correlated. The name of features are active_30,annual income at source,avg speed_l3m , active 60 and cc_active_60. We plot the distribution of these features one by one.

So firstly we discussed the distribution of annual income at source and we see that this is right skewed the distribution is not normal



After that we see the distribution of average speed l3m this distribution is also right skewed the distribution is not normal and most of the values of this features is around 0- 7000



After that we see the correlation of annual income at source and cc-limit we see that the correlation is not linear but there is very unusual behavior of correlation sometime cc_limit increase sharply and sometime annual income show a unusual behavior so there is no positive or negative relationship between features