SMDM PROJECT REPORT
DSBA

Abhishek Pradhan

Contents

Contents

Problem 1

A. What is the important technical information about the dataset that a database administrator would be interested in?

The Austo Motor Company dataset has 1581 rows and 14 columns consisting of one float,5 integers, and 8 object datatypes, with some missing data with the features Gender and Partner_salary.

```
#   Column            Non-Null Count  Dtype
---  ------           --------------  -----
 0   Age              1581 non-null   int64
 1   Gender           1528 non-null   object
 2   Profession       1581 non-null   object
 3   Marital_status   1581 non-null   object
 4   Education        1581 non-null   object
 5   No_of_Dependents 1581 non-null   int64
 6   Personal_loan    1581 non-null   object
 7   House_loan       1581 non-null   object
 8   Partner_working  1581 non-null   object
 9   Salary           1581 non-null   int64
 10  Partner_salary   1475 non-null   float64
 11  Total_salary     1581 non-null   int64
 12  Price            1581 non-null   int64
 13  Make             1581 non-null   object
dtypes: float64(1), int64(5), object(8)
```

```
Age 0
Gender 53
Profession 0
Marital_status 0
Education 0
No_of_Dependents 0
Personal_loan 0
House_loan 0
Partner_working 0
Salary 0
Partner_salary 106
Total_salary 0
Price 0
Make 0
```

B. 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? If yes, perform preliminary treatment of data.

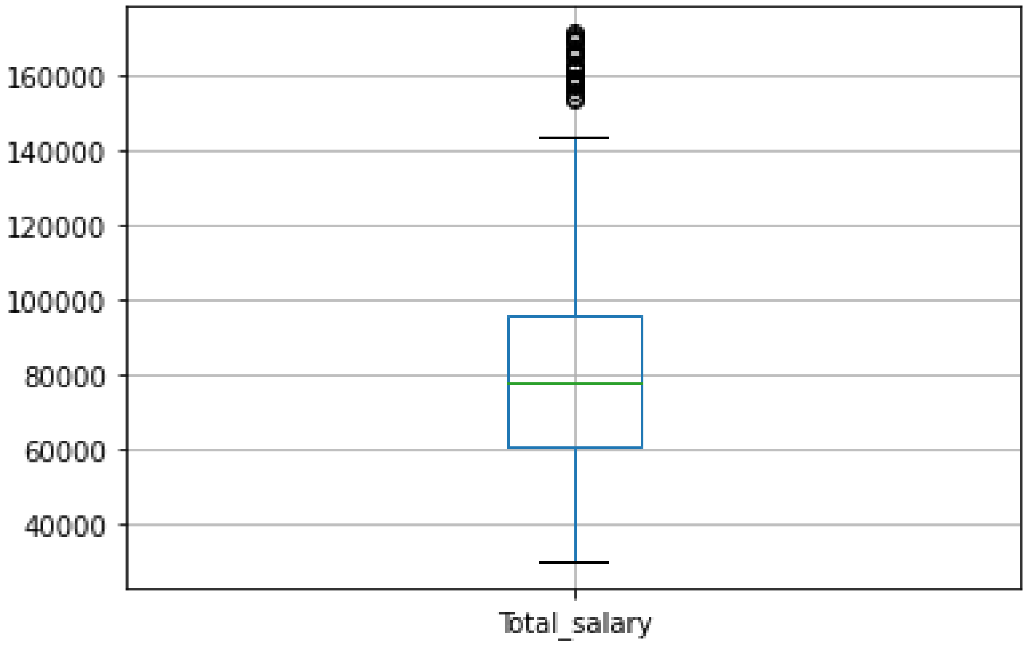
The Gender column has some bad data and 53 missing variables in it whereas in Partner_salary there are 106 missing values that need to be treated.

```
array(['Male', 'Femal', 'Female', nan, 'Femle']
```

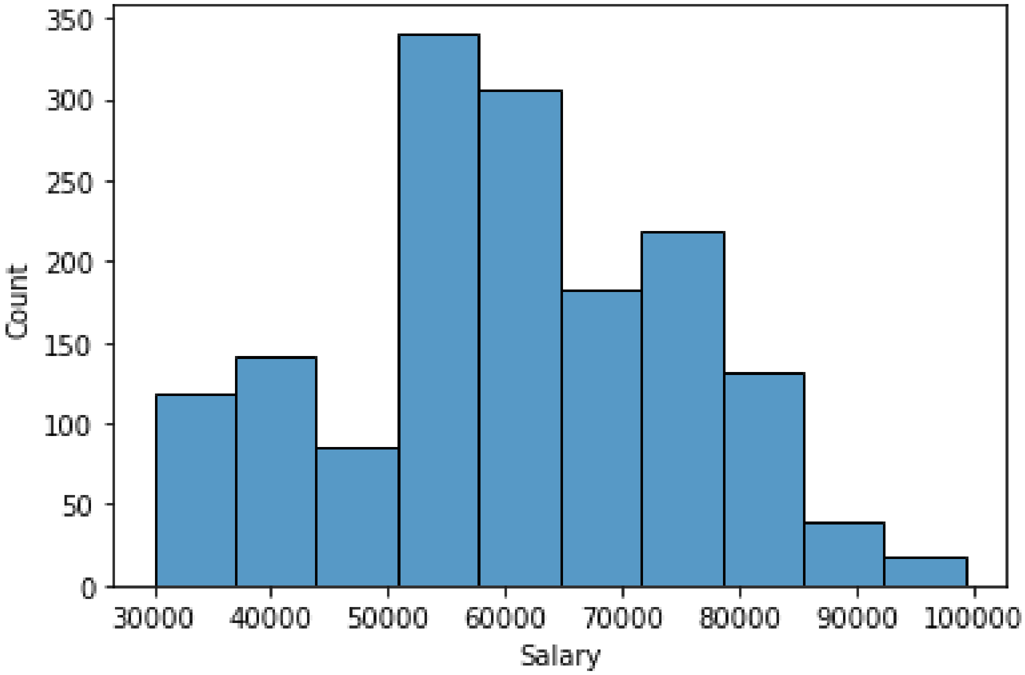
The Female has been mistyped in different rows and was fixed by replacing "Female" with it.
In addition to this as we know that Column Salary+Partner_salary=Total Salary. So filled in those null values using the equation.

And hence the data is treated.

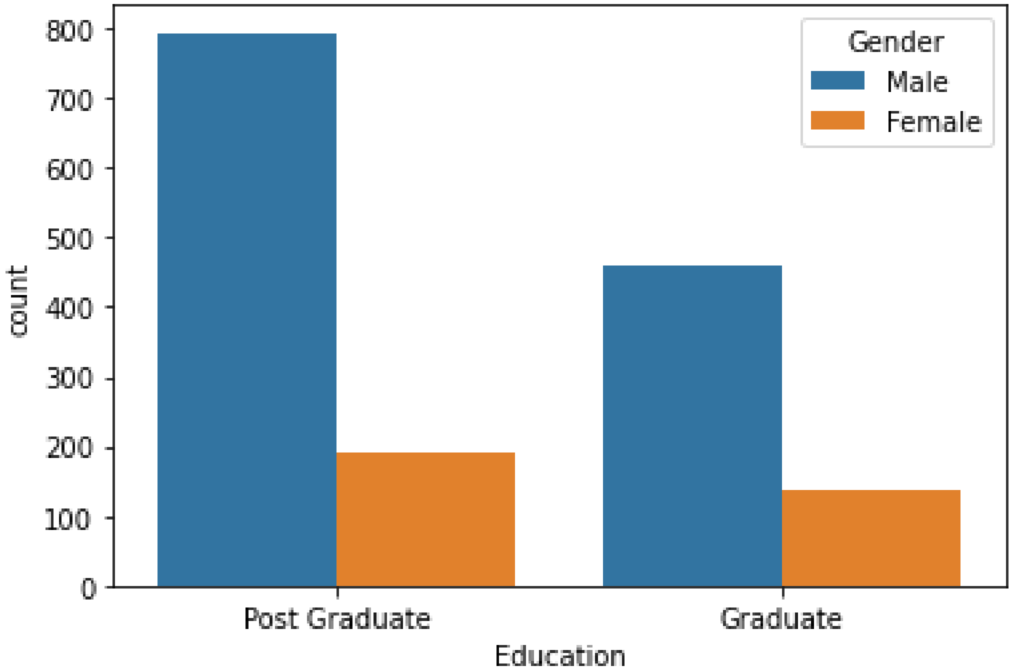
Next there are outliers in the Total Salary which was imputed using the IQR Techniques(Inter Quartile Ranges)



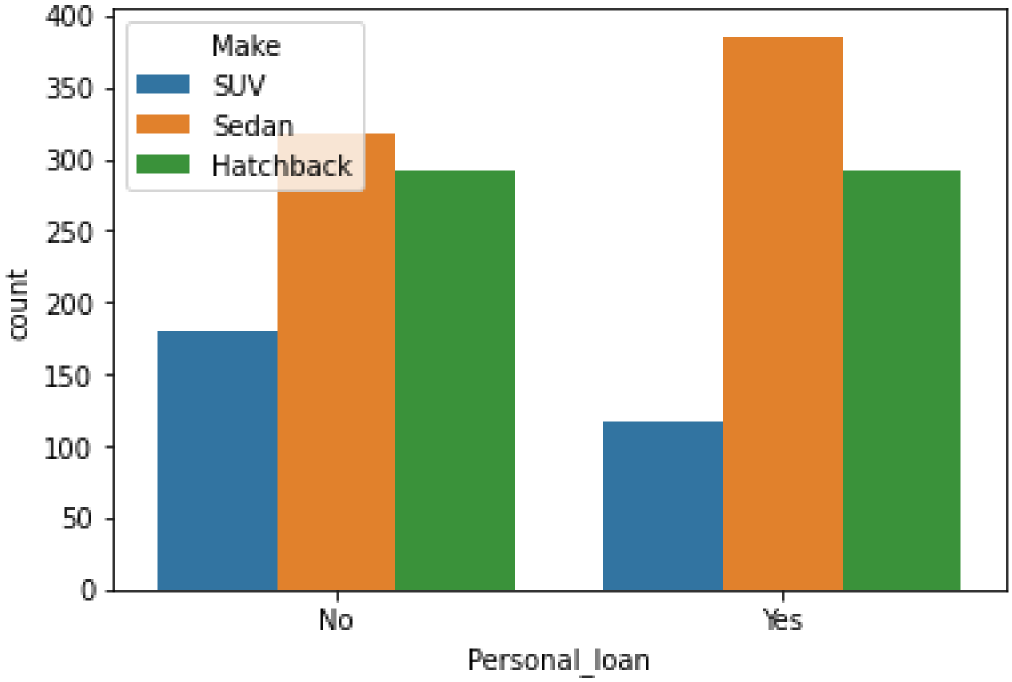
C. C Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.



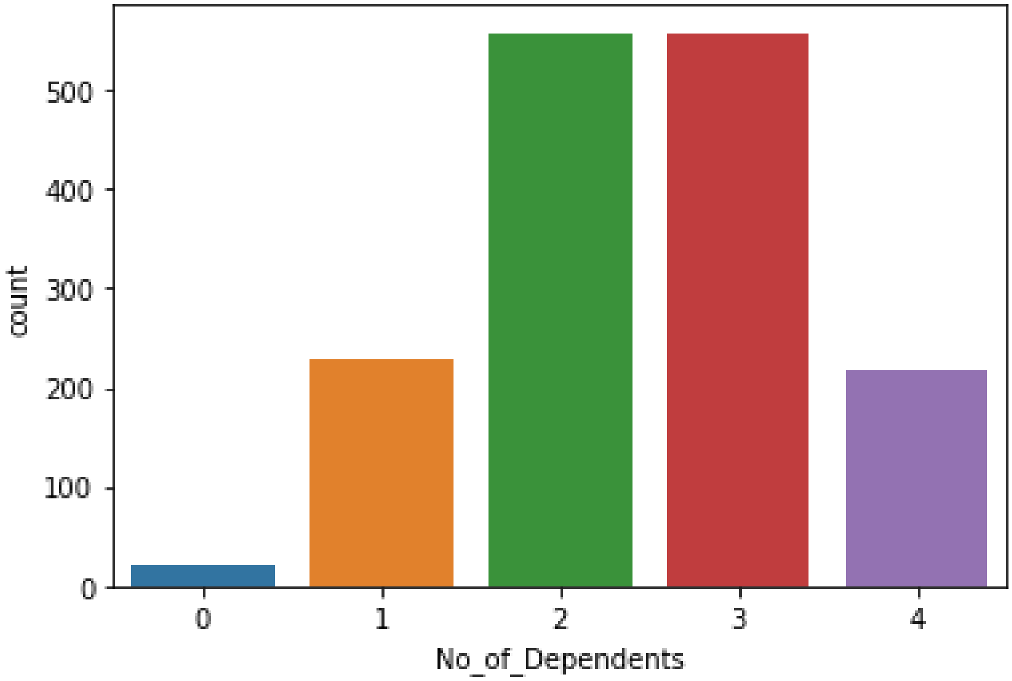
Here we can interpret that between 50000 to 60000 salaried people are the most in numbers.



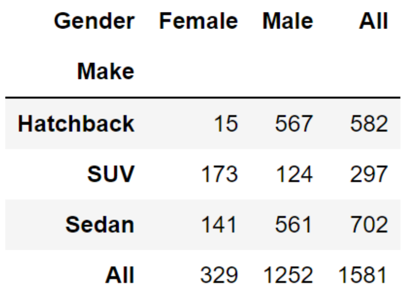
As per the education wise ,Post Graduate male and female are more than the Graduate ones.



Her ,without the personal loan people are buying more SUVs.



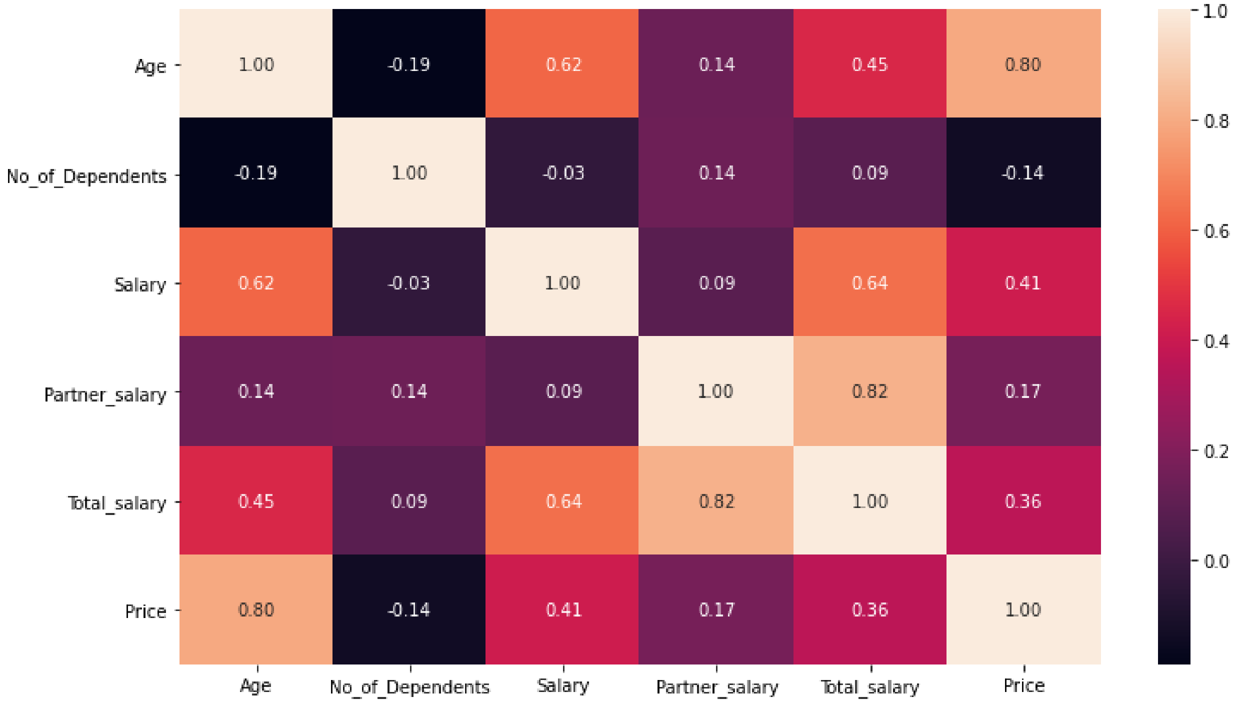
Here we can see number of dependents are more of 2 and 3.



Gender	Female	Male	All
Make			
Hatchback	15	567	582
SUV	173	124	297
Sedan	141	561	702
All	329	1252	1581

Here we can see males buying more than female.

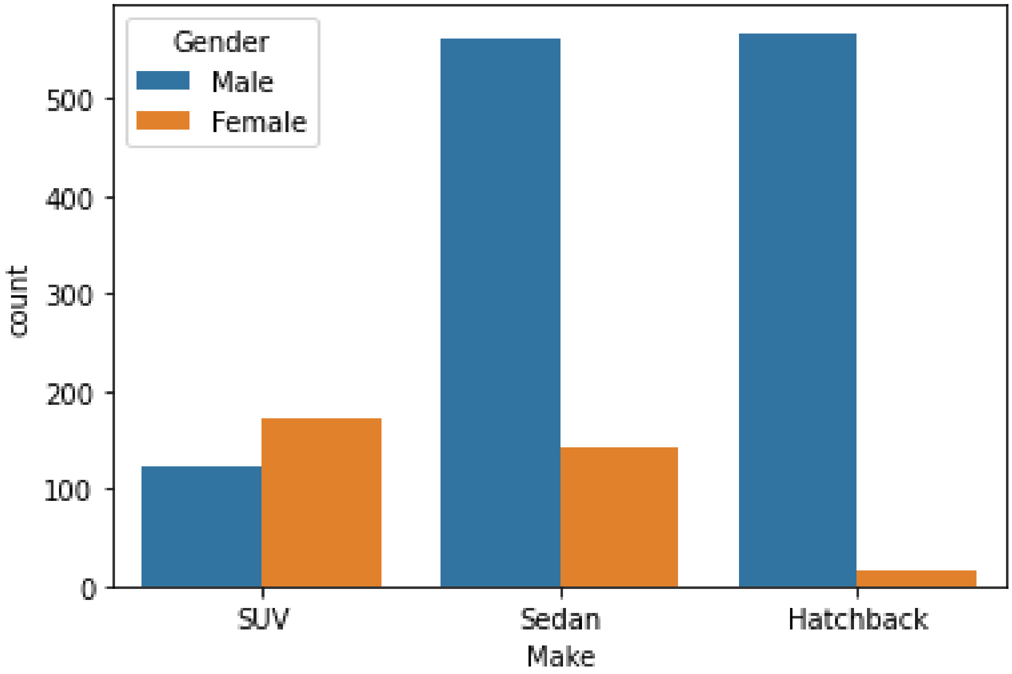
D. D. Understanding the relationships among the variables in the dataset is crucial for every analytical project. Perform analysis on the data fields to gain deeper insights. Comment on your understanding of the data.



To dig deeper we can see here a relation between Age and price highly correlating to each other, followed by Age and Salary, Salary and Total Salary, and Partner salary and Total Salary.

E. E. Employees working on the existing marketing campaign have made the following remarks. Based on the data and your analysis state whether you agree or disagree with their observations. Justify your answer Based on the data available.

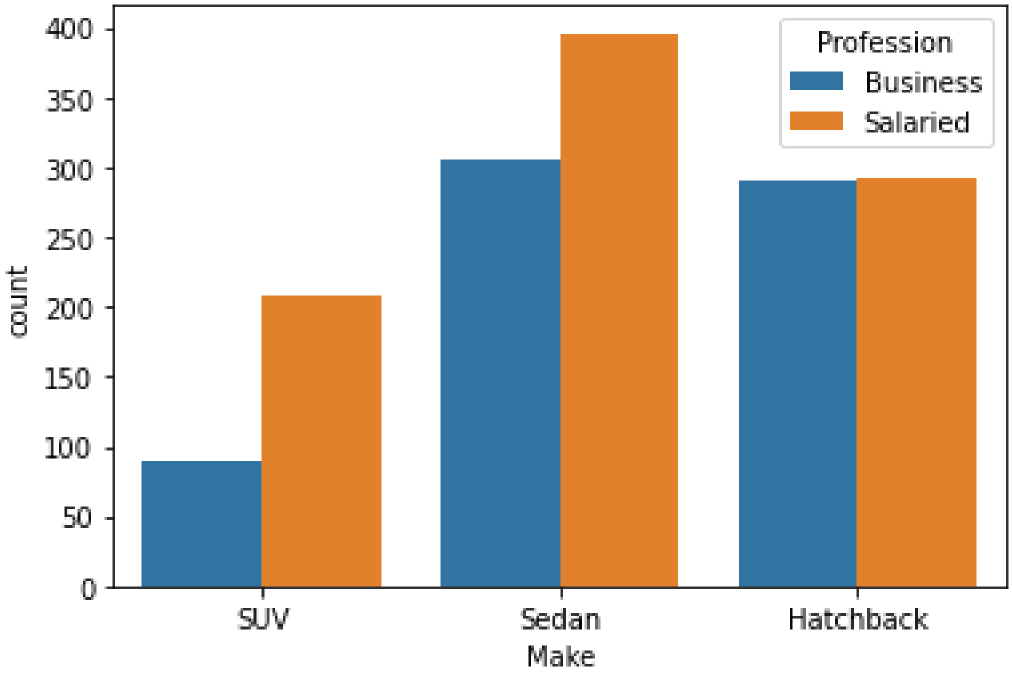
E1) Steve Roger says “Men prefer SUV by a large margin, compared to the women”



I totally disagree with Steve Roger’s statement. As per the findings, women prefer SUVs more than men. Whereas men prefer Sedans and Hatchbacks by a large margin than women.

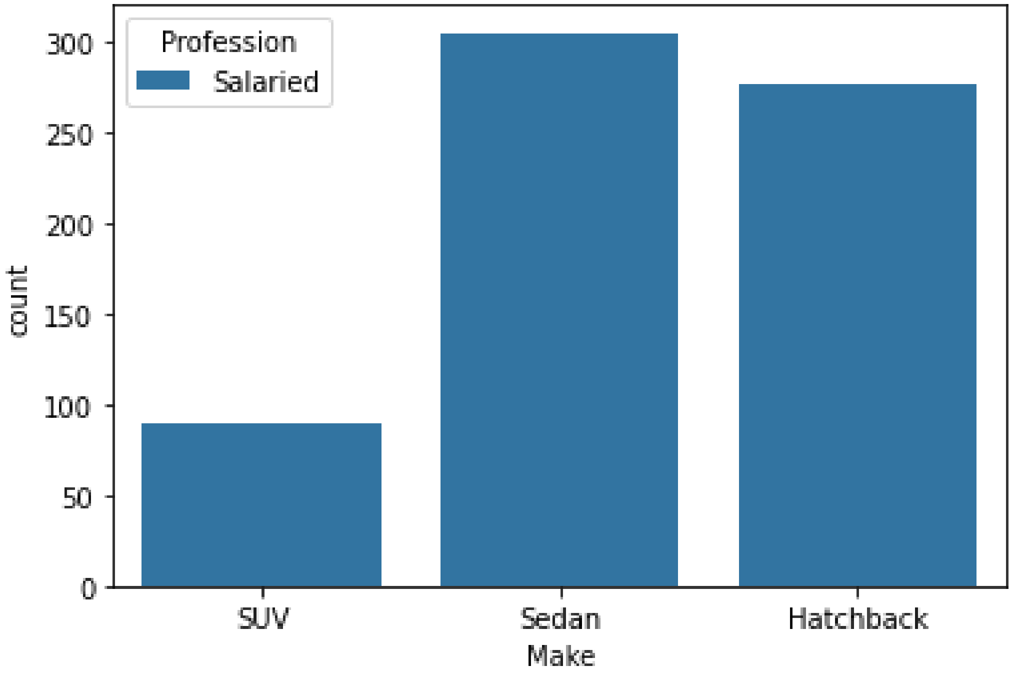
E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.

Yes, a salaried person is more likely to buy a sedan.



E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for an SUV sale over a Sedan Sale.

But as per the findings, it clearly states that a salaried male is an easier target for a Sedan or a Hatchback, not SUV.

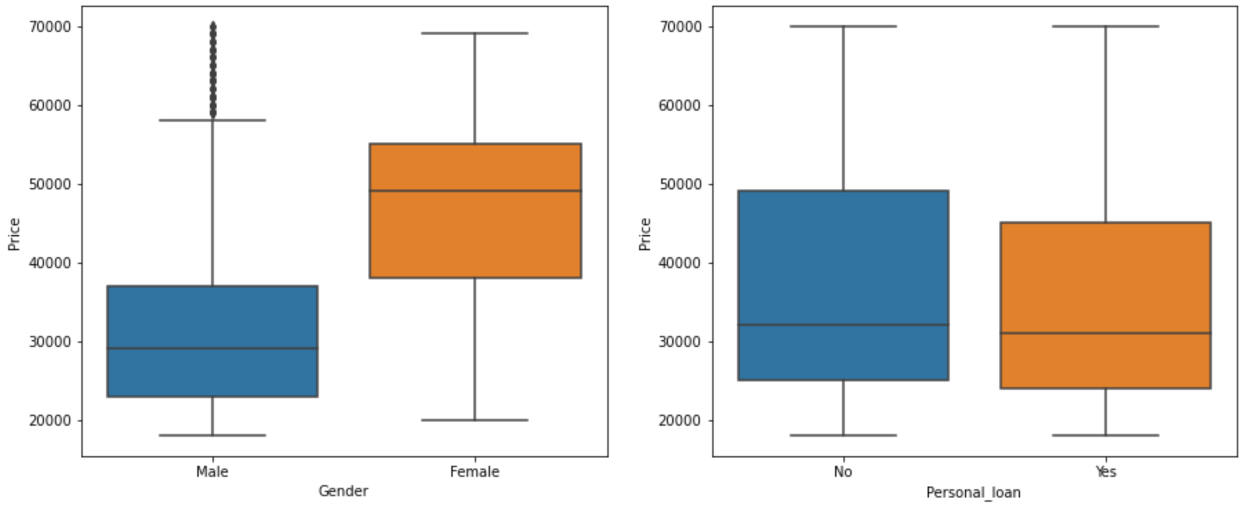


F. F. From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a Business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions.

Give justification along with presenting metrics/charts used for arriving at the conclusions.

F1) Gender

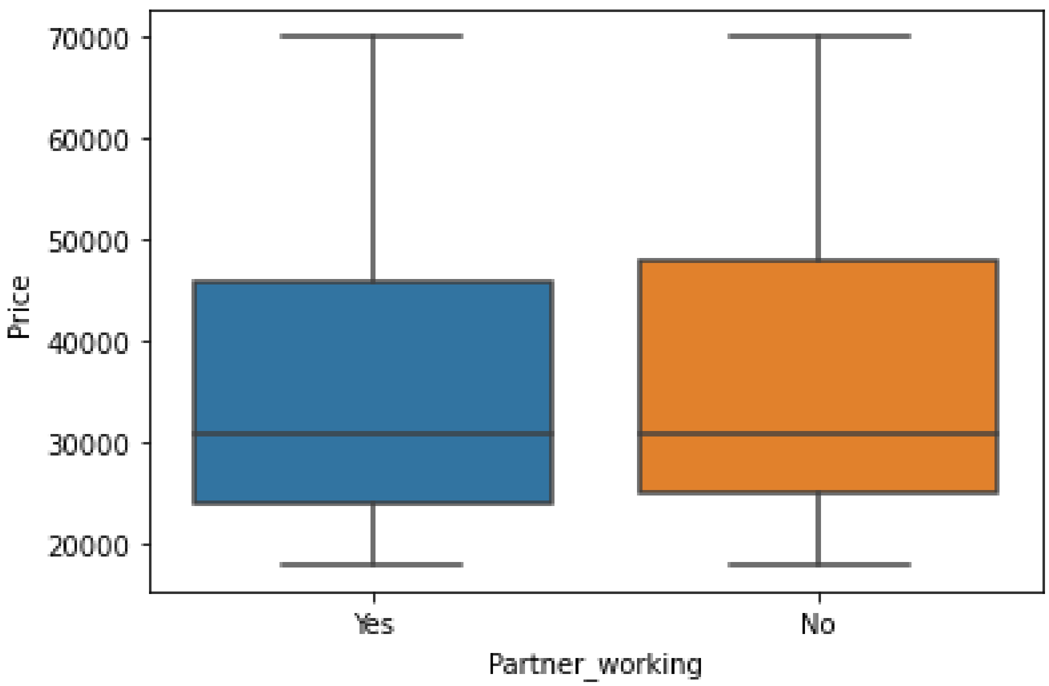
F2) Personal loan



Here we can see from the first graph as females are purchasing more priced cars as compared to males.

And people without personal loans are spending more on cars and people with personal loans are doing little less.

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



As per the dataset, the mean is the same,s o having a working partner leads to the purchase of a higher-priced car might not be true.

H. 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 the Gender and Marital_status - fields to arrive at groups with similar purchase history.

Problem 2

Analyse the dataset and list down the top 5 important variables, along with the business justifications:

Acc_Active30 – This will give the activity on the Credit Card for last 30 days whether the customer has used the card or not.

annual_income_at_source – This metric will give us an insight into the spending power and the income record for the whole year. The more the salary more will be the spending power.

[T+1_month_activity, T+2_month_activity, T+3_month_activity] – This will give information on the customer's propensity to use their credit card during the next one, two, or three months. This will make it easier for the bank to forecast consumer spending patterns and assign the appropriate offers and credit cards.

avg_spends_l3m – This variable will provide information about the customer's typical spending over the last three months and aid in predicting the best card for the consumer based on average spend.

cc_limit – From this variable, I got a pattern i.e. having less credit card limit is inversely proportional to the higher expenditure and vice versa with a higher credit card limit where bank can send more offers inorder to increase the expenditure.