Ayush Agarwal

SMDM Project Report

Contents

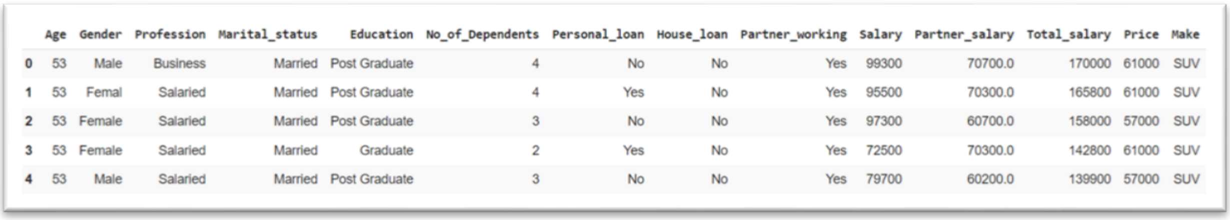
Problem 1 : Austo Motor Company

Analysts are required to explore data and reflect on the insights. Clear writing skill is an integral part of a good report. Note that the explanations must be such that readers with minimum knowledge of analytics is able to grasp the insight.

Austo Motor Company is a leading car manufacturer specializing in SUV, Sedan, and Hatchback models. In its recent board meeting, concerns were raised by the members on the efficiency of the marketing campaign currently being used. The board decides to rope in an analytics professional to improve the existing campaign.

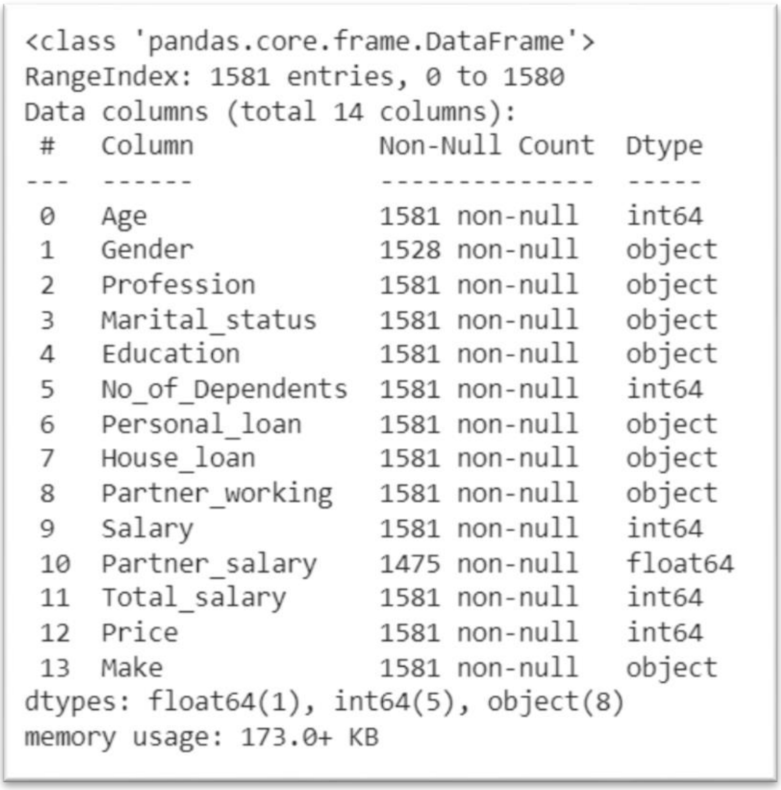
You as an analyst have been tasked with performing a thorough analysis of the data and coming up with insights to improve the marketing campaign.

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



	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	Price	Make
0	53	Male	Business	Married	Post Graduate	4	No	No	Yes	99300	70700.0	170000	61000	SUV
1	53	Femal	Salaried	Married	Post Graduate	4	Yes	No	Yes	95500	70300.0	165800	61000	SUV
2	53	Female	Salaried	Married	Post Graduate	3	No	No	Yes	97300	60700.0	158000	57000	SUV
3	53	Female	Salaried	Married	Graduate	2	Yes	No	Yes	72500	70300.0	142800	61000	SUV
4	53	Male	Salaried	Married	Post Graduate	3	No	No	Yes	79700	60200.0	139900	57000	SUV

Table gives an at a glance look at the data provided.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581 entries, 0 to 1580
Data columns (total 14 columns):
 #   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)
memory usage: 173.0+ KB
```

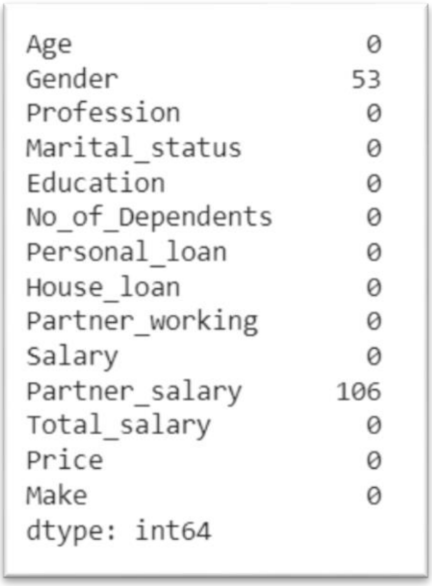
Table gives the information about the details for each column and number of entries (1581).

There are **8 categorical variables** and **6 numerical variables**.

There are null values in two variables:

1. **Gender (1528)**
2. **Partner_salary (1475)**

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?

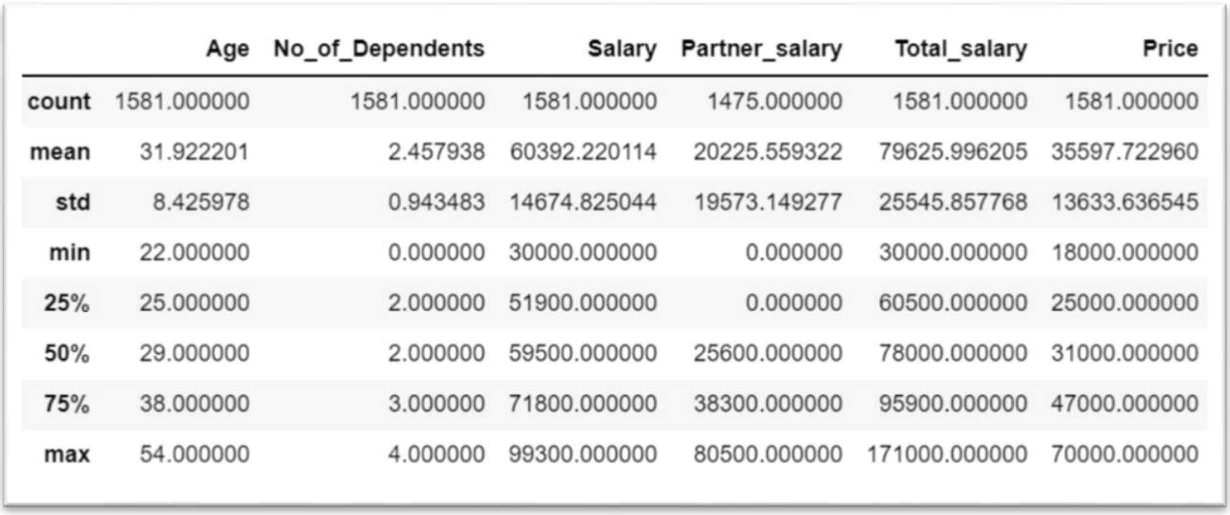


```
Age                  0
Gender              53
Profession           0
Marital_status       0
Education            0
No_of_Dependents     0
Personal_loan        0
House_loan           0
Partner_working      0
Salary               0
Partner_salary     106
Total_salary         0
Price                0
Make                 0
dtype: int64
```

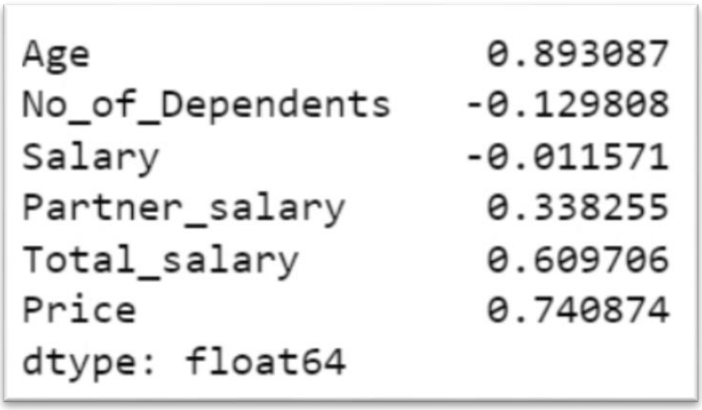
As we can see from the above table **Gender** and **Partner_salary** has **53** and **106** null values respectively.

Handling the null values:

1) Gender: As it is a categorical variable hence the null value is replaced by the mode in the date that is male.
   Before imputing the value: **Male = 1199 ; Female = 329**
   After imputing the value: **Male = 1252 ; Female = 329**
2) Partner_salary: Treating the null values of Partner_salary from the data from the dataframe.
   If partner is not working the value is imputed by **0**
   If partner is working the value is imputed by **(Total_salary-Salary)**



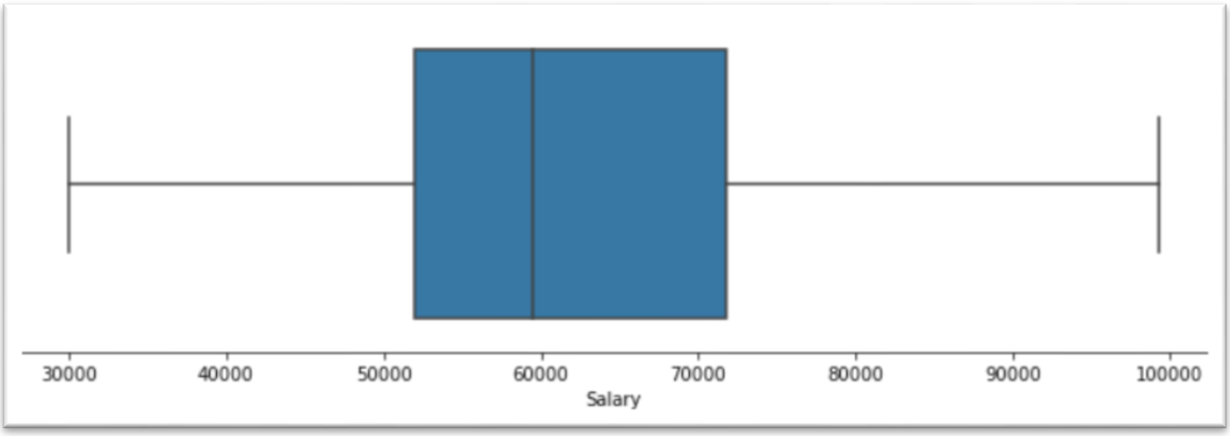
	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
count	1581.000000	1581.000000	1581.000000	1475.000000	1581.000000	1581.000000
mean	31.922201	2.457938	60392.220114	20225.559322	79625.996205	35597.722960
std	8.425978	0.943483	14674.825044	19573.149277	25545.857768	13633.636545
min	22.000000	0.000000	30000.000000	0.000000	30000.000000	18000.000000
25%	25.000000	2.000000	51900.000000	0.000000	60500.000000	25000.000000
50%	29.000000	2.000000	59500.000000	25600.000000	78000.000000	31000.000000
75%	38.000000	3.000000	71800.000000	38300.000000	95900.000000	47000.000000
max	54.000000	4.000000	99300.000000	80500.000000	171000.000000	70000.000000

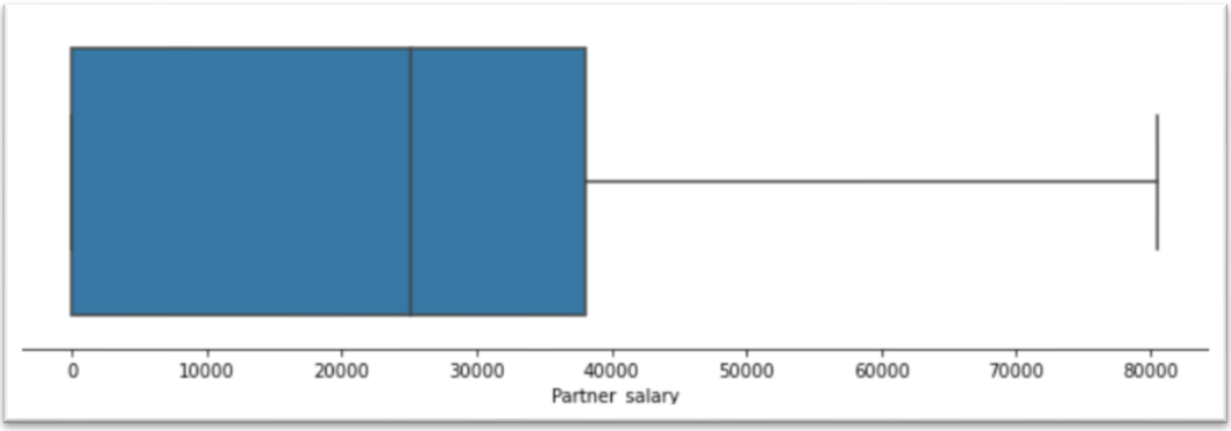


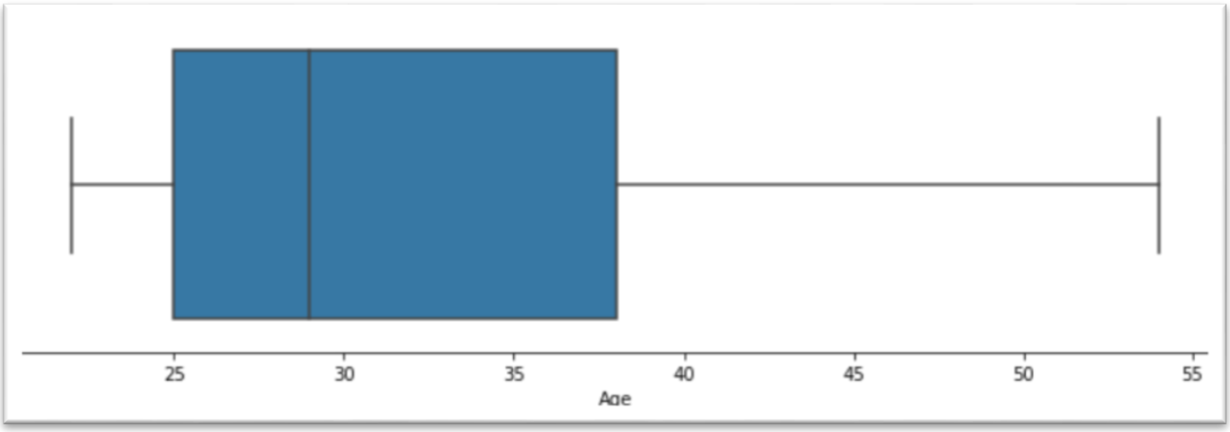
```
Age                 0.893087
No_of_Dependents   -0.129808
Salary             -0.011571
Partner_salary      0.338255
Total_salary        0.609706
Price               0.740874
dtype: float64
```

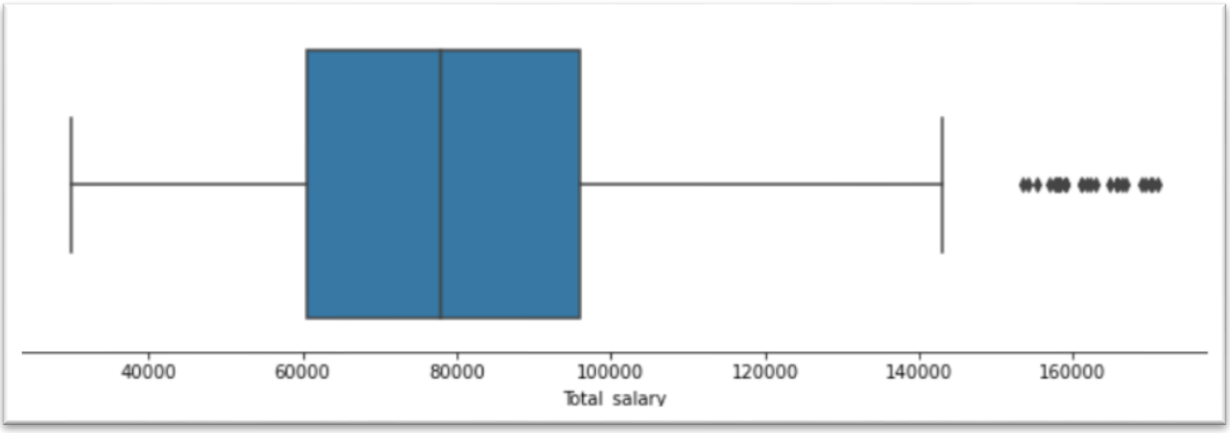
The above data of the dataset tells us the following information :

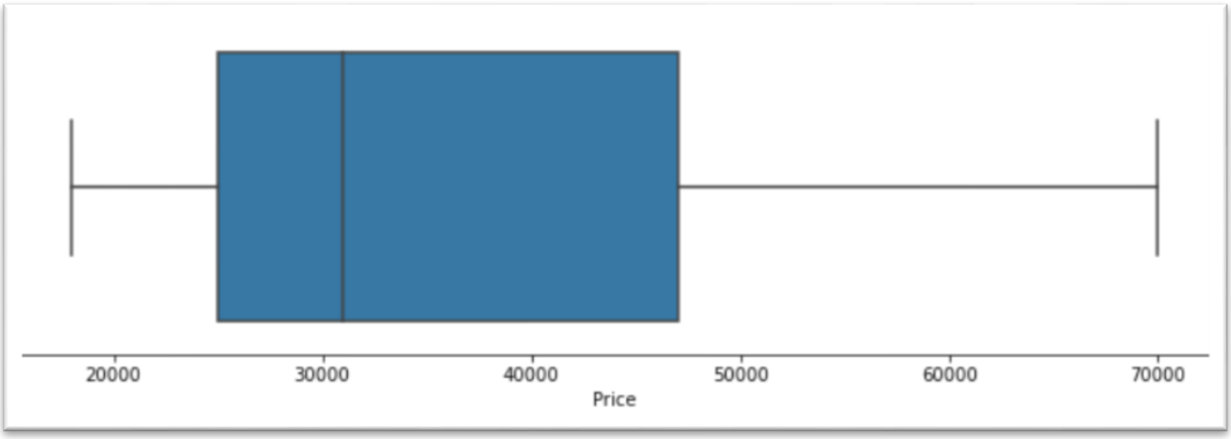
- Customers of age group between **22 to 54 years** old. Average age of the people are **31.92** and Median age is **29 years**. This indicates the age distribution in positively skewed. The value of **skewness is 0.89**.
- The salary of the customers ranges between **30K to 99.3K** and the distribution is symmetric as mean and the median values are very close and skewness is very close to **0.**
- Total_salary ranges between **30K and 171K.**
- Price of the purchased mobile is **minimum 18K and maximum is 70K**. Price is having skewness of **0.74.**


30000
40000
50000
60000
70000
80000
90000
100000
Salary


0
10000
20000
30000
40000
50000
60000
70000
80000
Partner_salary


25
30
35
40
45
50
55
Age



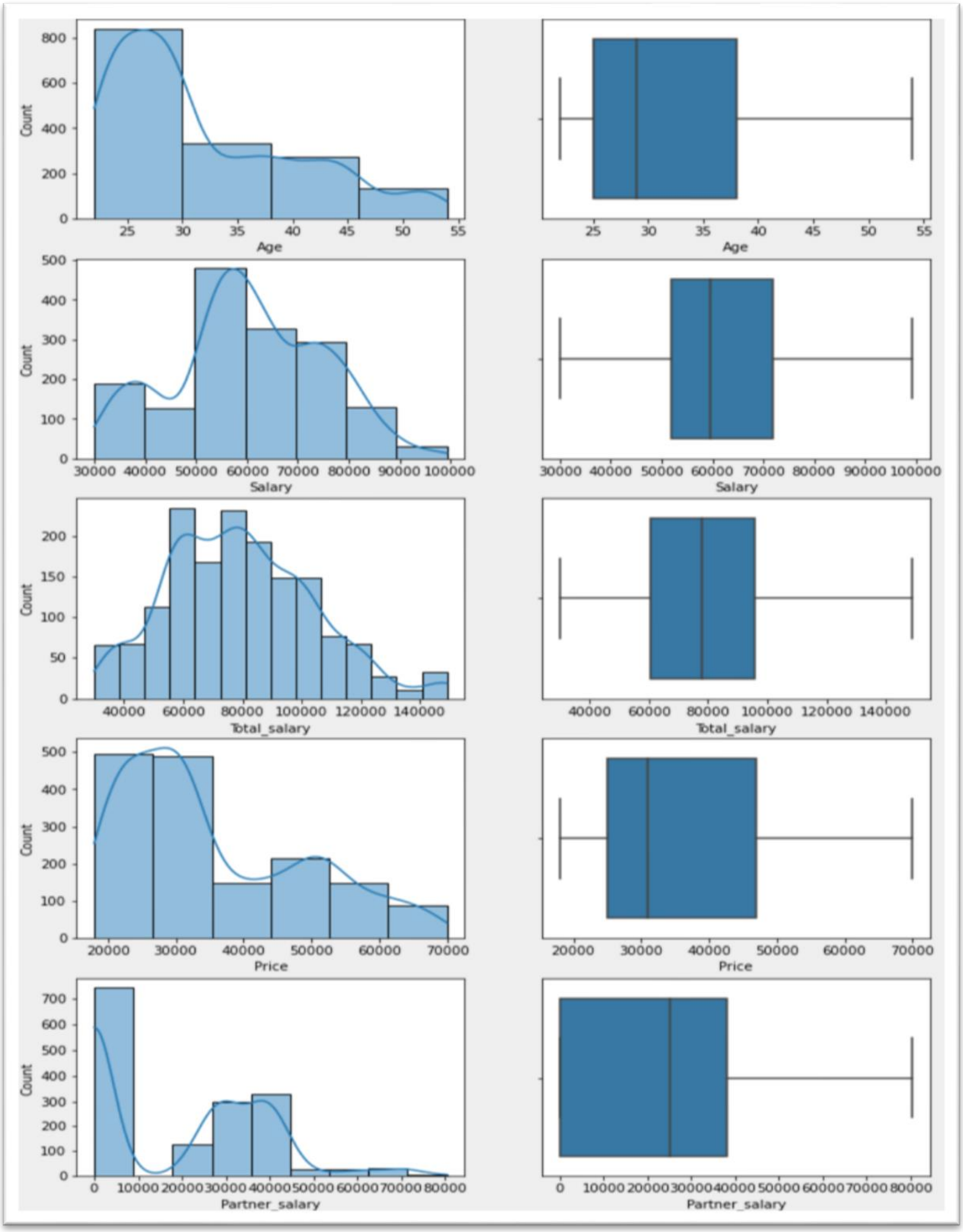


As we can see the Total_salary contains outliers that can be treated by replacing the extreme values by maximum and minimum.

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

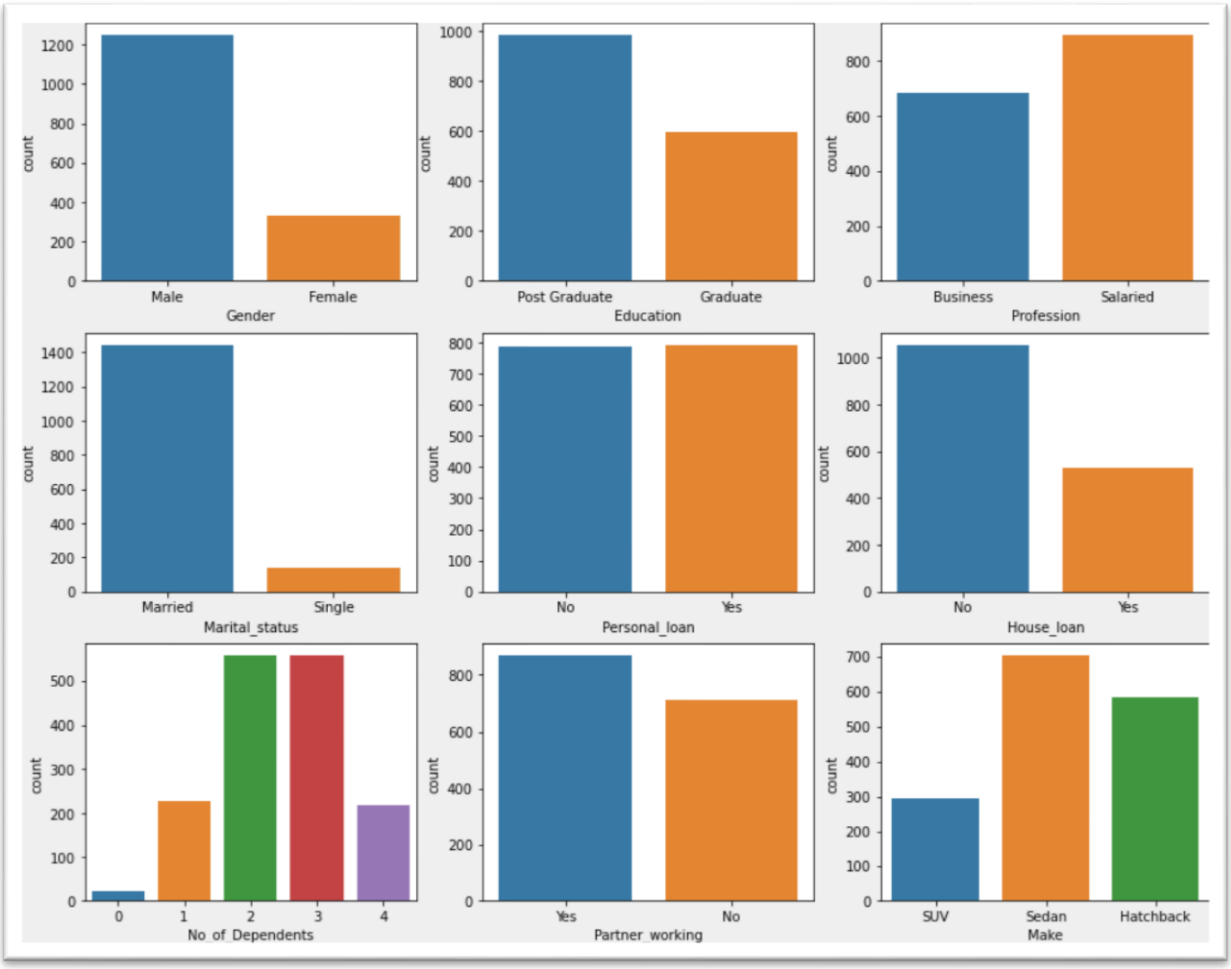
Univariate analysis of Numerical field.

For performing Univariate analysis we will be taking a look at the Boxplots and Histograms to get better understanding of the distribution.



Inferences:

1) **Age** is positively skewed with **multi modal distribution**.
2) **Salary** is also a **multi modal distribution** with most of the data lying between **50k to 70k**.
3) After the treatment of outliers **Total_salary** seems to be **multi modal** with most of the data lying between lying between **60k to 100k.**
4) **Price** seems to have a **bimodal distribution.**



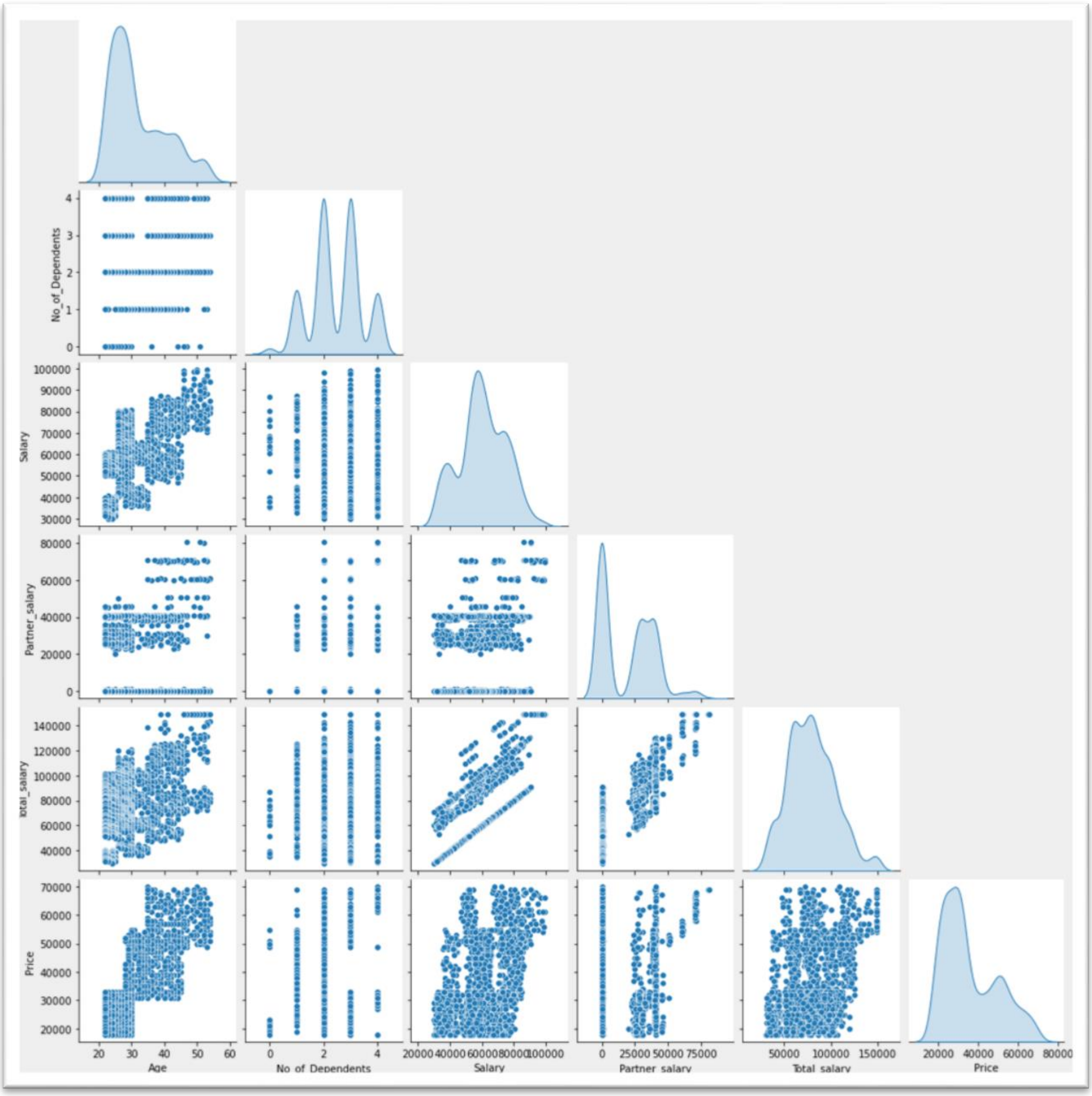
Inferences:

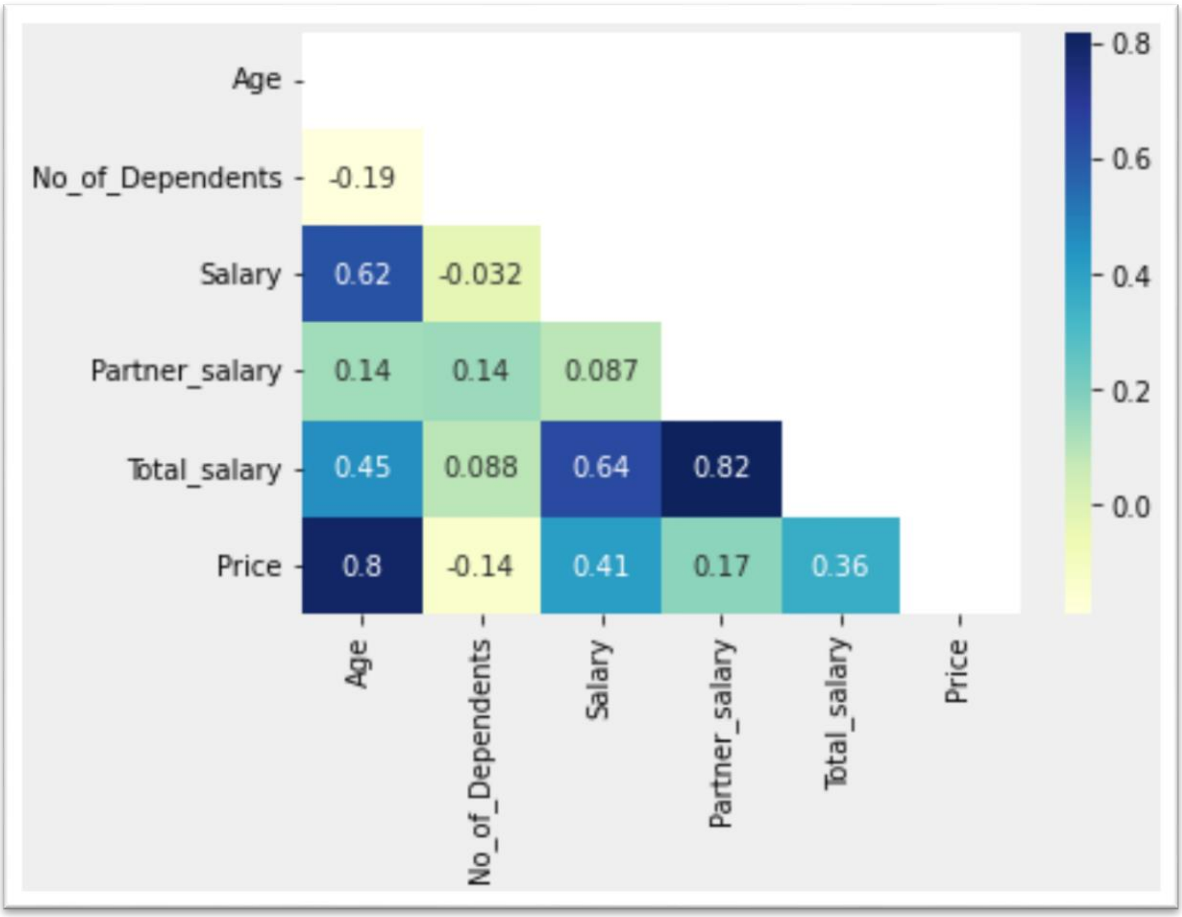
1) Count of **Male is more than female** in Gender category.
2) Count of customers with **post-graduation degree is higher** than customer who are graduate.
3) There are **more salaried customers buying cars** than business owners.
4) **Married customers are very high** as compared to single.
5) When it comes to loan, there are equal number of customer with or without personal loan but for with **house loan customer is nearly half without house loan.**
6) **Most of the customer lies with 2 or 3** dependents followed by 1 and 4.

7) Customers with **working partners are more** than customers with no working partners.
8) In terms of make **sedan is the highest sold car** followed by hatchback followed by SUV.

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.

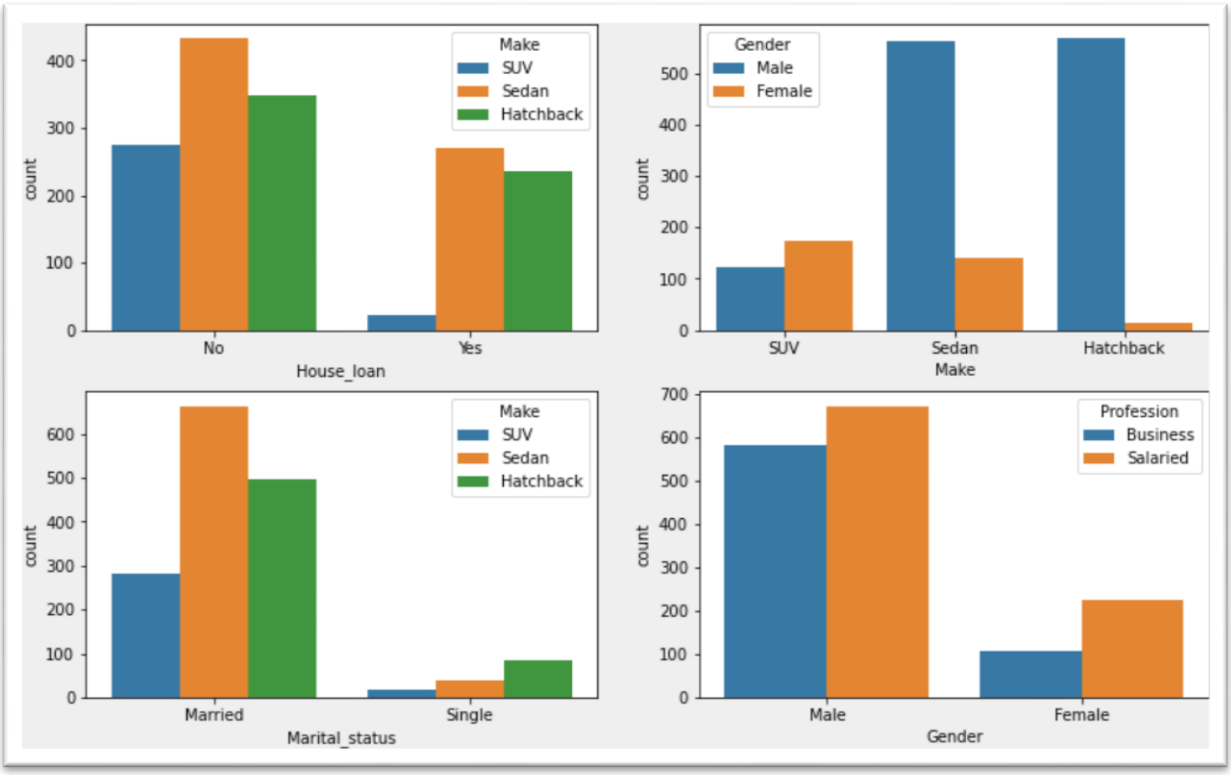
Bivariate analysis for Numerical variable :





Inference:

From the above it is clear that highest and the **positive correlation exist between Age and Price and Total_salary and Partner_salary** that is if one increases other also increases. Whereas the highest **negative correclation exists between Age and No_of_Dependents.**

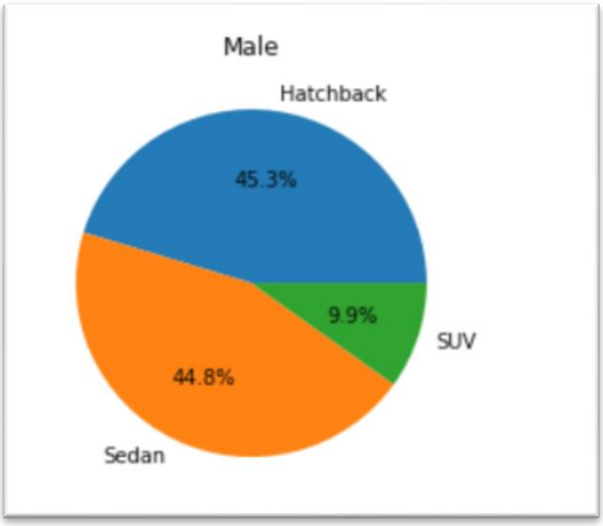


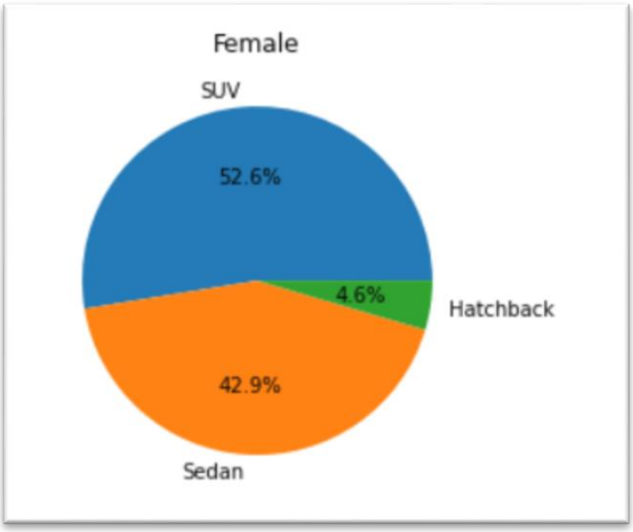
Inferences:

1) Customers with **House loan are more likely to buy sedan** and rarely buys SUV whereas customers **with no house loan mostly buys sedan** followed by hatchback and sedan.
2) When it comes to Gender **male mostly prefer sedan** and hatchback where as **female prefer SUV** more as compare to male.
3) **Married customer prefer mostly sedan** cars **whereas single goes for hatchback**.

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.

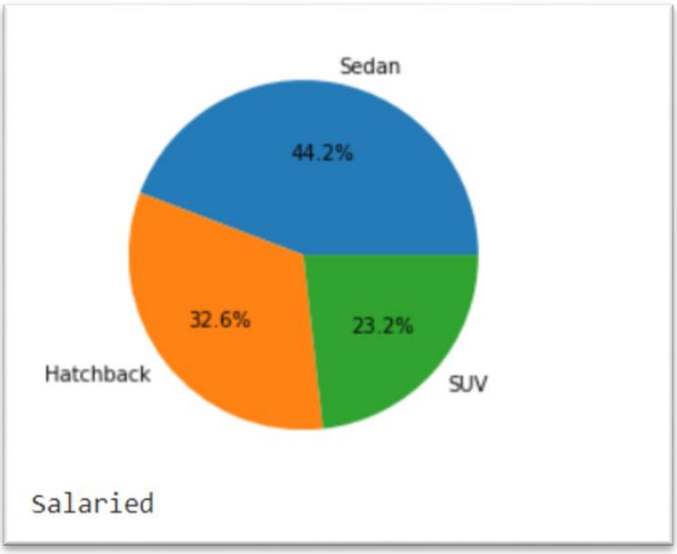
E1) Steve Roger says "Men prefer SUV by a large margin, compared to the women"





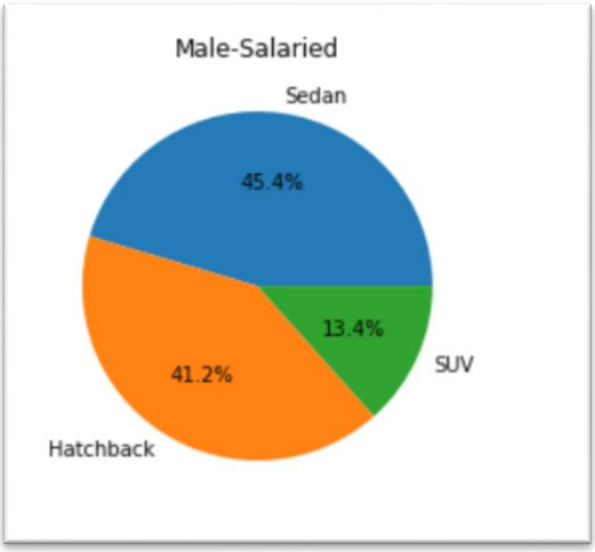
From the above pie chart it is clear that out of the **total female customers 52.6% goes for SUV** whereas out of **the total male customers only 9.9% goes for SUV,** hence **the statement by Steve Roger "Men prefer SUV by a large margin, compared to the women" is absolutely false.**

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



From the above pie chart it is clear that out of **the total salaried customers 44.2% goes for Sedan, hence the statement by that a salaried person is more likely to buy a Sedan is correct.**

E3) 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.

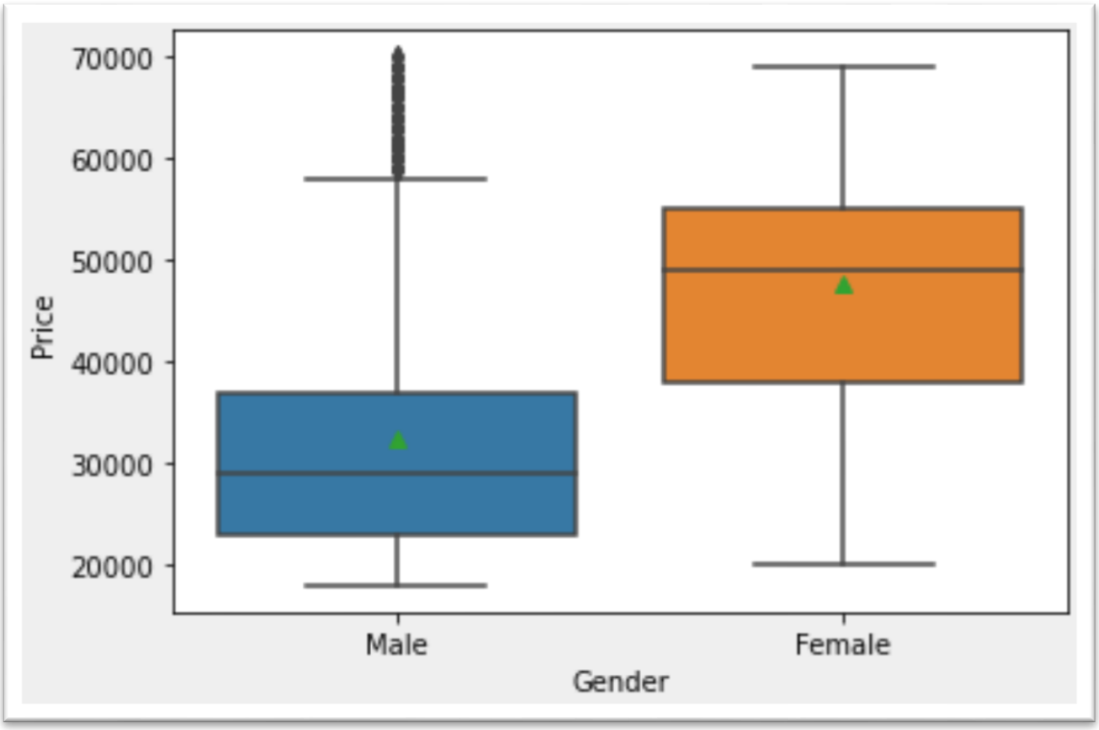


From the above pie chart which is male specific, it is clear that out of the total salaried male **customers 45.4% goes for Sedan and only 13.4% goes for SUV**, **hence the statement by that a salaried male is an easier target for a SUV sale over a Sedan Sale is incorrect.**

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

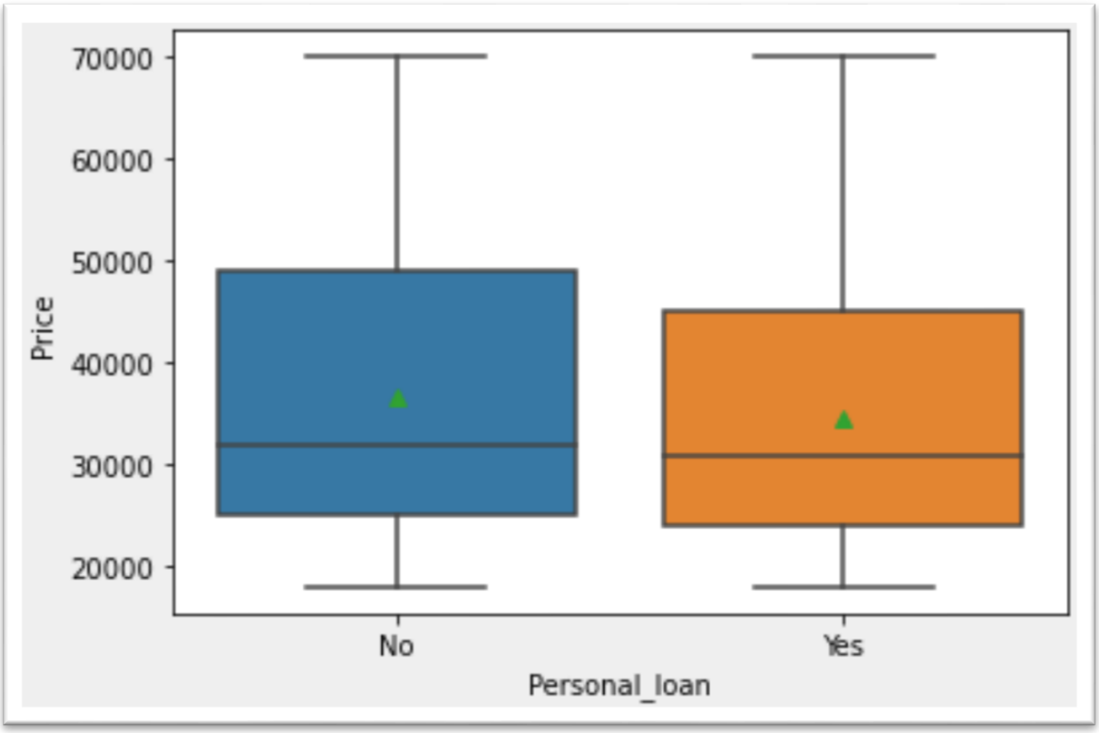


```
Mean Gender
Female    47705.167173
Male      32416.134185
Name: Price, dtype: float64

 Median Gender
Female    49000.0
Male      29000.0
```

It is evident from the above data female spends more money than male while purchasing a car.

F2) Personal_loan

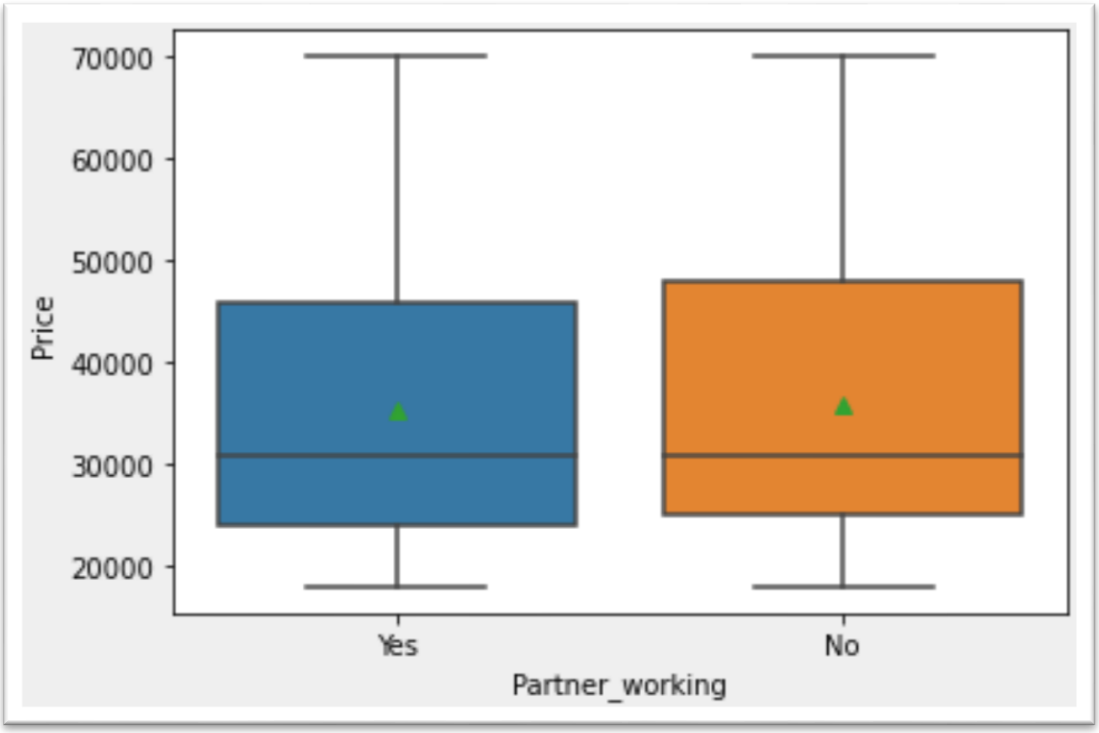


Mean Personal_loan
No   36742.712294
Yes   34457.070707

Median Personal_loan
No   32000.0
Yes   31000.0

There is not much difference between the amount spend by the customer with or without personal loan.

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

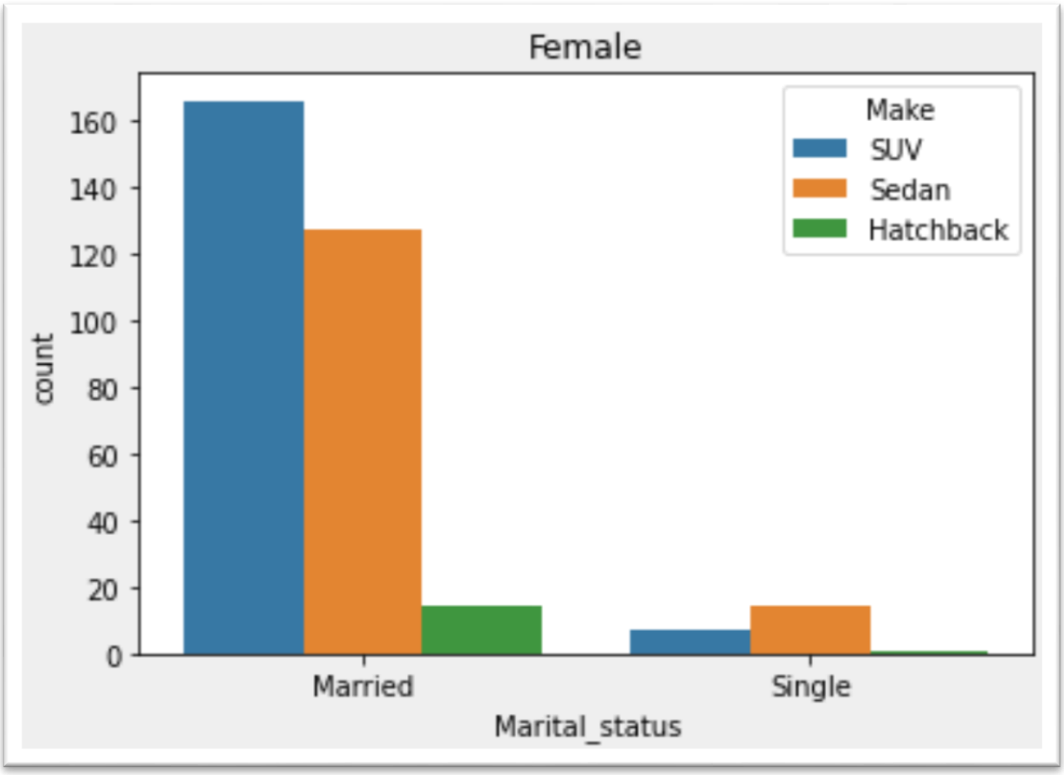


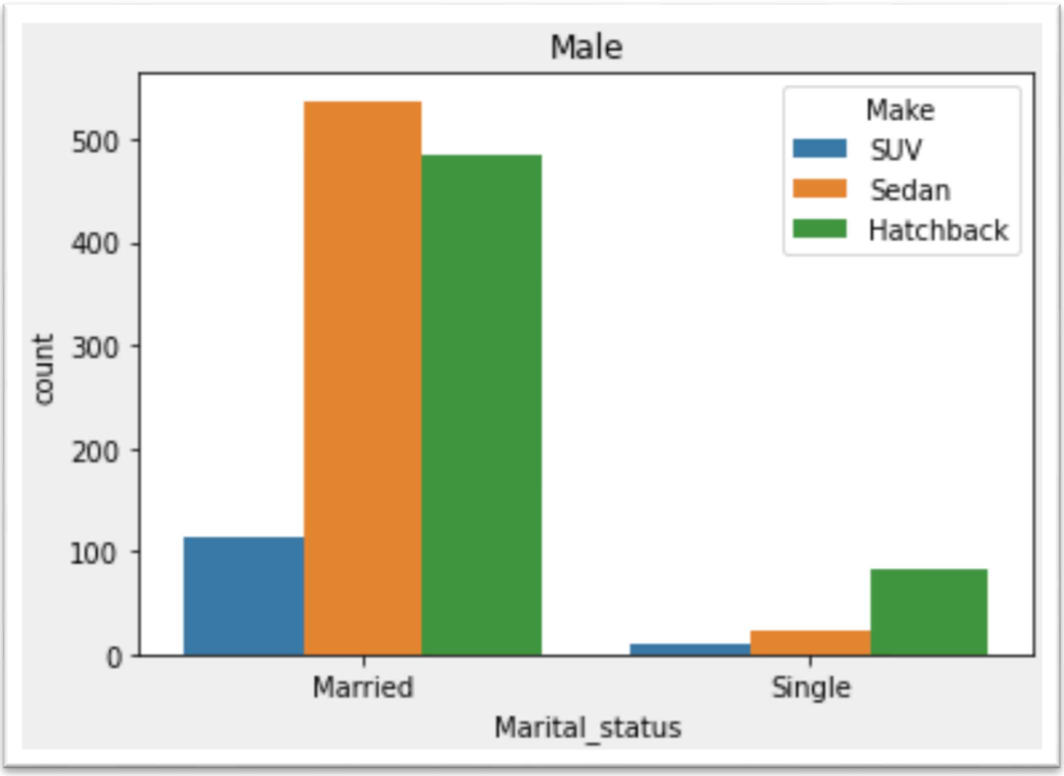
```
Partner_working
No     36000.000000
Yes    35267.281106

Partner_working
No     31000.0
Yes    31000.0
```

There is not much difference between the amount spend by the customer with or without working partners.

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.





From the past trends following inferences can be made based on gender and martial status of a customer to the make of the car he prefers to purchase.

- **Married female prefers SUV**
- **Single female prefers Sedan**
- **Married male prefers Sedan**
- **Single male prefers Hatchback**

Problem 2

A bank can generate revenue in a variety of ways, such as charging interest, transaction fees and financial advice. Interest charged on the capital that the bank lends out to customers has historically been the most significant method of revenue generation. The bank earns profits from the difference between the interest rates it pays on deposits and other sources of funds, and the interest rates it charges on the loans it gives out.

GODIGT Bank is a mid-sized private bank that deals in all kinds of banking products, such as savings accounts, current accounts, investment products, etc. among other offerings. The bank also cross-sells asset products to its existing customers through personal loans, auto loans, business loans, etc., and to do so they use various communication methods including cold calling, e-mails, recommendations on the net banking, mobile banking, etc.

GODIGT Bank also has a set of customers who were given credit cards based on risk policy and customer category class but due to huge competition in the credit card market, the bank is observing high attrition in credit card spending. The bank makes money only if customers spend more on credit cards. Given the attrition, the Bank wants to revisit its credit card policy and make sure that the card given to the customer is the right credit card. The bank will make a profit only through the customers that show higher intent towards a recommended credit card. (Higher intent means consumers would want to use the card and hence not be attrite.)

Below are the Top 5 important variables from the given dataset with justification.

1) Annual_income_at_source : Annual income tells us about the purchasing power of the user hence it plays a very important role in making decisions related to risk profiling, targeted ads, campaigns, offers, loan limits etc.

2) cc_limit : It defines the credit limit given to a customers based on different attributes (such as CIBIL Score, income, etc.) wherein the banks try to minimize the number of defaulters through risk management.

3) cc_active30 : Flag variables such as cc_active30 are used to understand over how frequently does the customer use the credit card, and study the usage of customer over time.

4) T+1_month_activity : Flag variables such as T+1_month_activity can be used to plan out campaigns and promotional offers so as to increase activity in the credit card.

5) avg_spends_l3m : The avg_spends_l3m gives the idea about spending of a customer. It can be used to identify paying capacity of the customer and campaigns can be done accordingly, customized offers and rewards can be given to customers.