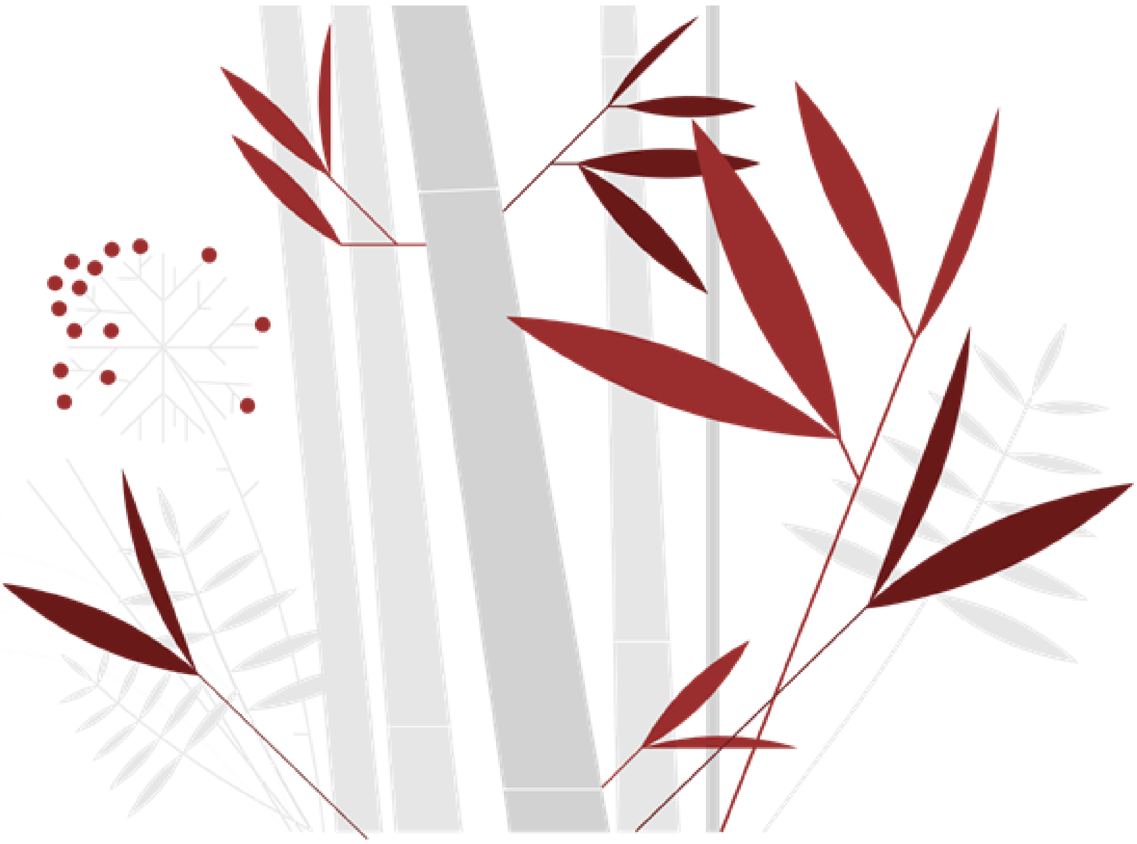
STATISTICAL METHODS FOR DECISION MAKING (SMDM) PROJECT-CODED

BY

Harsh Patel

10th March 2024



Sr No.	Contents	Page No.
1.	Problem-1: Austo Auto Company	6
1.1	Data Structure.	6
1.1.1	Dataset quality check and corrections	9
1.1.2	Observation and Insights	10
1.2	Uni-variate Analysis	11
1.2.1	Graphs for Categorical Variables	11
1.2.2	Observation and Insights	15
1.2.3	Boxplots and Histograms for Numerical Variables	16
1.2.4	Outliers treatment	18
1.2.5	Observation and Insights	19
1.3	Bi-Variate Analysis	19
1.3.1	Numerical vs Numerical	19
1.3.2	Observation and Insights	22
1.3.3	Numerical vs Categorical each with Observation and Insights	22
1.4	Key Questions	28
1.4.1	Do men tend to prefer SUVs more compared to women?	28
1.4.2	What is the likelihood of a salaried person buying a Sedan?	29
1.4.3	What evidence or data supports Sheldon Cooper's claim that a salaried male is an easier target for a SUV sale over a Sedan sale?	30
1.4.4	How does the amount spent on purchasing automobiles vary by gender?	31
1.4.5	How much money was spent on purchasing automobiles by individuals who took a personal loan?	32
1.4.6	How does having a working partner influence the purchase of higher-priced cars?	33
1.5	Conclusion and Business Justification	34
2.	Problem-2: GODIGT Bank	35

2.1	***Key questions***	***36***
2.1.1	***How much does the customer on an average uses the credit cards? Can we find out the correlation between there average spend, income and credit card limit?***	***36***
2.1.2	***What type of card of does the customer uses the most?***	***37***
2.1.3	***What is the income and occupation of the customer?***	***38***
2.1.4	***In last 30 days how much has the customer spend and what is there cc_limit compared to their Occupation?***	***39***
2.2	***The five important variables with business justification***	***40***

Fig no.	Figure and Chart names	Page no.
1.	Gender Bar and Pie chart	11
2.	Profession Bar and Pie chart	12
3.	Marital Status Bar and Pie chart	12
4.	Education Bar and Pie chart	13
5.	Personal loan Bar and Pie chart	13
6.	House loan Bar and Pie chart	14
7.	Partner working Bar and pie chart	14
8.	Make Bar and Pie chart	15
9.	No. of Dependents Bar graph	16
10.	Age boxplot and histogram	16
11.	Price boxplot and histogram	17
12.	Salary boxplot and histogram	17
13.	Partner salary boxplot and histogram	17
14.	Total Salary boxplot and histogram before outliers treatment	18
15.	Total Salary boxplot and histogram after outliers treatment	18
16.	Pair-plot for all numerical variables	20
17.	Heatmap of correlation for all numerical variables	21
18.	Age vs All categorical variables	22
19.	No. of Dependents vs All categorical variables	23
20.	Salary vs All categorical variables	24
21.	Partner salary vs All categorical variables	25
22.	Total salary vs All categorical variables	26
23.	Price vs All categorical variables	27
24.	Gender vs Make Bar graph	28
25.	Profession vs Make Bar graph	29

26.	Profession vs Make Bar graph for male	30
27.	Gender vs Price boxplot	31
28.	Personal loan vs Price for single person boxplot	32
29.	Partner working vs Price Box plot	33
30.	Marital Status vs Price Box plot for Male and Female.	34
31.	Heatmap of correlation between all numerical variables.	36
32.	Count plot for card types	37
33.	Annual income vs credit card limit boxplot	38
34.	Occupation vs Credit card limit boxplot for credit card activity 30 days	39

Table No.	***Table Name***	***Pg no.***

1.	First five Rows	6
2.	Last five Rows	6
3.	Information on variable datatype	7
4.	Statistical summary before any treatment	8
5.	Missing values in Variables	8
6.	Irregularities in Categorical Variable	9
7.	No missing values in variables	10
8.	Statistical summary after treatment	10
9.	Correlation table between numerical variables	21
10.	Likelihood of salaried person to get different car models.	29

Problem-1: Austo Auto Company

The Austo Auto Company is a car manufacturer that produces SUVs, Sedan and Hatchback car models. During their recent board meeting some members raised the question on the efficiency of the marketing campaign done by the marketing team.

We as a data scientist have been asked to help them to make better strategies for their marketing campaign.

1.1 Data structure

First, we will look at the structure of the dataset given to us. First, we import the necessary libraries like matplotlib, pandas, seaborn etc. as per our requirement and load the "austo_automobile.csv" data file. Now, we see first and last 5 rows using head and tail function in the dataset for our visual inspection to get an idea about the type of variables and their values available to us.

	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	P
0	53	Male	Business	Married	Post Graduate	4	No	No	Yes	99300	70700.0	170000	61
1	53	Femal	Salaried	Married	Post Graduate	4	Yes	No	Yes	95500	70300.0	165800	61
2	53	Female	Salaried	Married	Post Graduate	3	No	No	Yes	97300	60700.0	158000	57
3	53	Female	Salaried	Married	Graduate	2	Yes	No	Yes	72500	70300.0	142800	61
4	53	Male	Salaried	Married	Post Graduate	3	No	No	Yes	79700	60200.0	139900	57

Table-1 First five rows

	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary
1576	22	Male	Salaried	Single	Graduate	2	No	Yes	No	33300	0.0	33300
1577	22	Male	Business	Married	Graduate	4	No	No	No	32000	NaN	32000
1578	22	Male	Business	Single	Graduate	2	No	Yes	No	32900	0.0	32900
1579	22	Male	Business	Married	Graduate	3	Yes	Yes	No	32200	NaN	32200
1580	22	Male	Salaried	Married	Graduate	4	No	No	No	31600	0.0	31600

Table-2 Last five rows

From the above table-1 and table-2 we got the idea about different variable datatypes and values present in them. But for better, understanding of dataset we further explore it using shape, info, describe function one by one.

Now we move on to see the shape of the dataset using shape function and using that function on the dataset we get that there are 1581 rows and 14 columns available for analysis in the dataset.

Now, we check for the data type of the variables and see if that the datatype is different form the one, we observed in the first and last rows of the dataset. For this we use the info function and we observed that there are 5 integer data type, 1 float data type, 8 object datatype.

```
<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: 173.1+ KB
```

Table-3 Information on Variable Datatypes

We can see that all the numerical variables are in numerical datatype. From this we can conclude that there are no anomalies/bad values in the numerical variables. We also, conclude that that in Partner Salary and in Gender variables that there are some missing values.

But we do not know if there are any irregularities or anomalies present in the dataset. First, we will find out all the empty values and any anomalies present in the dataset then we will resolve those one by one.

We will get the counts, mean, median (Q2 or 50%), standard deviation, quantiles (Q1/25% and Q3/75%) values, min and max of all the numerical variables before treating any empty values. And from table-3 we know that there are no irregularities present in any numerical variables. As, we are currently only exploring the dataset we will deal with null values and anomalies later steps. We use describe function to get Statistical summary before any corrections done as shown in table-3 below.

	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

Table-4 Statistical Summary before treatment

So, we check for any missing values present in the dataset using isnull function on the dataset and we got that there are 53 and 106 missing values in Gender and Partner working variables, as shown in table-5 below.

```
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
```

Table-5 Missing values in Variables

Now, we will also see if there are any irregularities or corrupted data in the categorical variables in the dataset and for the is we us value count function for each of the categorical variable. We found out that there are two anomalies present in Gender column, as shown in table-6 below.

```
Gender
Male        1199            Personal_loan
Female       327            Yes     792
Femal          1            No      789
Femle          1            Name: count, dtype: int64
Name: count, dtype: int64   House_loan
Profession                  No      1054
Salaried     896            Yes      527
Business     685            Name: count, dtype: int64
Name: count, dtype: int64   Partner_working
Marital_status              Yes     868
Married     1443            No      713
Single       138            Name: count, dtype: int64
Name: count, dtype: int64   Make
Education                   Sedan        702
Post Graduate     985       Hatchback    582
Graduate          596       SUV          297
Name: count, dtype: int64   Name: count, dtype: int64
```

Table-6 Irregularities in Categorical Variable

1.1.1 Data set quality check and corrections:

From table-5 and table-6 we got all the irregularities and empty values present in the dataset. First, we will deal with the 2 anomalies found in the Gender variable. We replace the “Femal”, “Femle” with “Female” using replace function or using loc method after finding out their location in the dataset. We will check using unique function we conclude that there no anomalies present in the Gender variable anymore.

For dealing with the missing values there are two ways to fix it, first we can impute it with its mean, median or mode and second way is to drop those rows entirely. We use second method only when our missing values are less than 1.5%. And, we have almost 3% missing values in the whole dataset. So, we use the first method.

For the empty values in the Gender column, we find out the most recurring value using mode function and we get that “Male” is the most recurring variable. Thus, by using the “Fillna” function we fill those empty values with “Male”. And for Partner salary variable we know that if we subtract the Salary column from the Total salary column, we get the Partner salary column.

By, doing so we have dealt with all the missing values and anomalies. Now, again we use info and describe function after the treatment of the dataset, and we obtain table-7 and table-8 as shown below.

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

Table-7 No missing values in variables

	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1581.0	19233.776091	19670.391171	0.0	0.0	25100.0	38100.0	80500.0
Total_salary	1581.0	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	1581.0	35597.722960	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0

Table-8 Statistical Summary after treatment

1.1.2 Observation and In-sights:

Thus, from the above procedures we learned the shape of dataset is 1581 row and 14 columns.

The variables are of two data types:

1) Categorical variables: Gender, Marital status, Profession, Education, Personal loan, House loan, Partner working, Make.
2) Numerical variables: Age, No. of dependents, Salary, Partner salary, Total Salary, Price.
3) We also got to see the Statistical Summary for all the numerical variables and find out that Q1 and min for partner salary column is same. And got the 5 important values from them for all numeric variables.

1.2 Uni-variate Analysis:

For each variables weather they are categorical or numerical type we will perform univariate analysis on them for better understanding of their distributions. We will see observations and insights at the end for categorical and numerical separately.

1.2.1 Graphs for Categorical Variables:

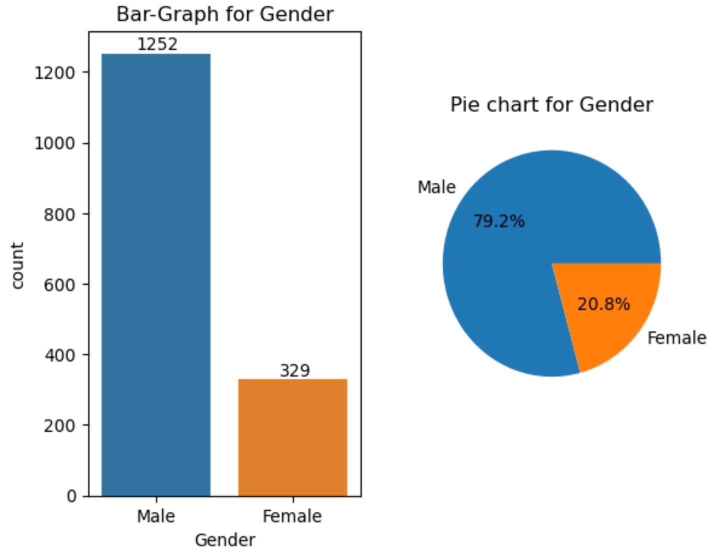


Fig-1 Gender Bar and Pie chart

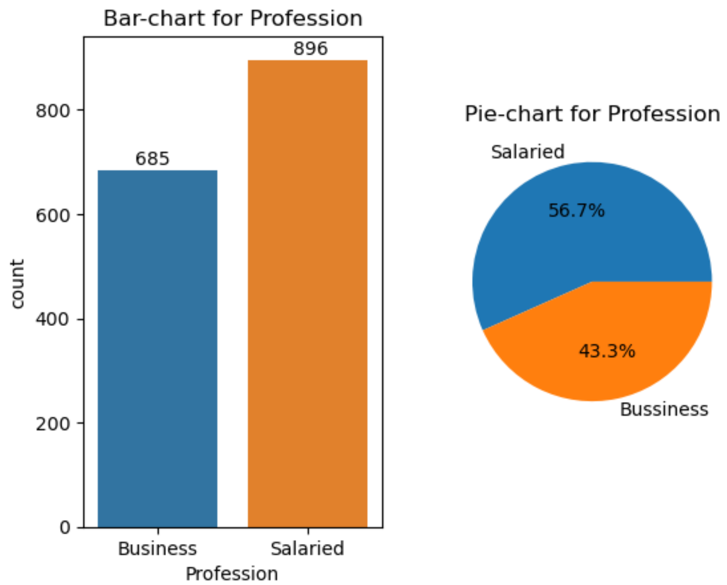


Fig-2 Profession Bar and Pie chart

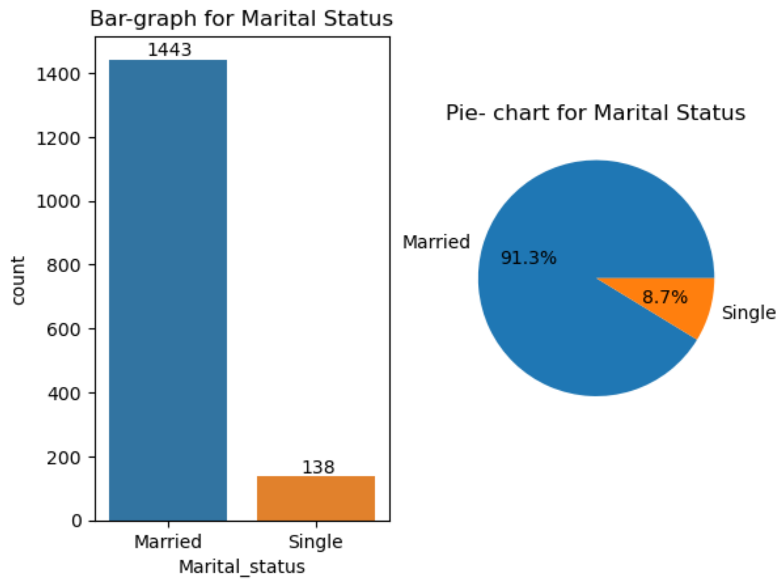


Fig-3 Marital Status Bar and Pie chart

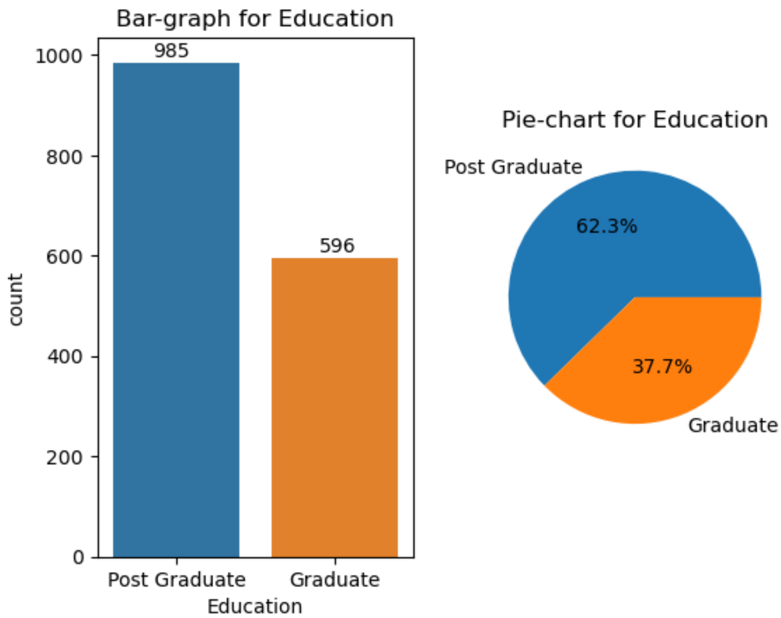


Fig-4 Education Bar and Pie chart

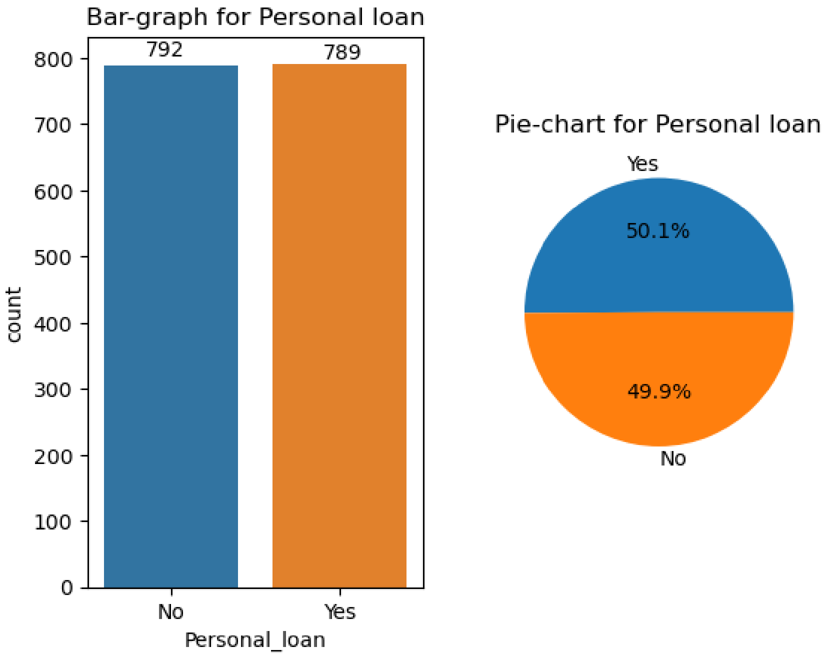


Fig-5 Personal loan Bar and Pie chart

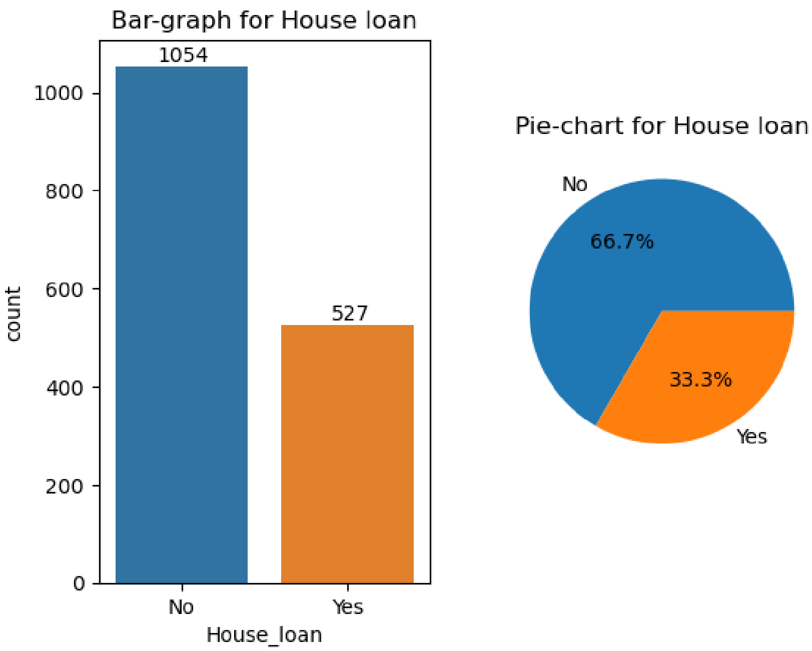


Fig-6 House loan Bar and Pie chart

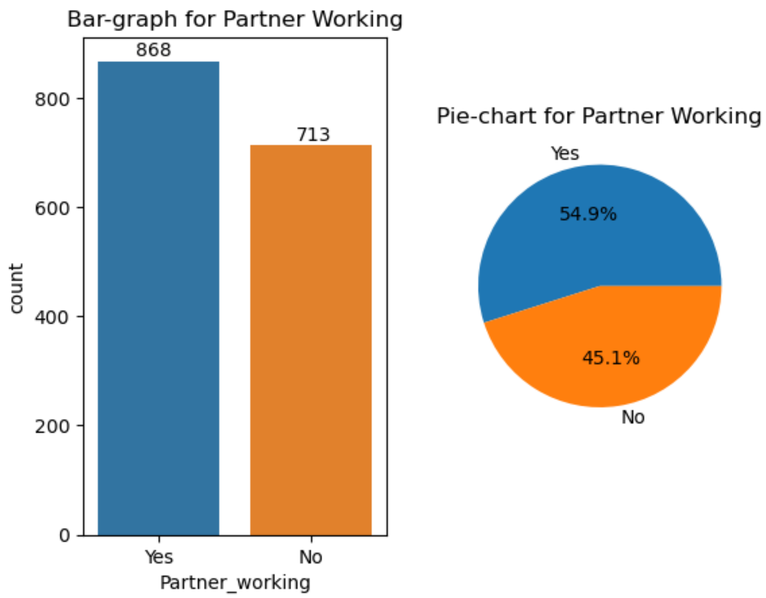


Fig-7 Partner working Bar and Pie chart

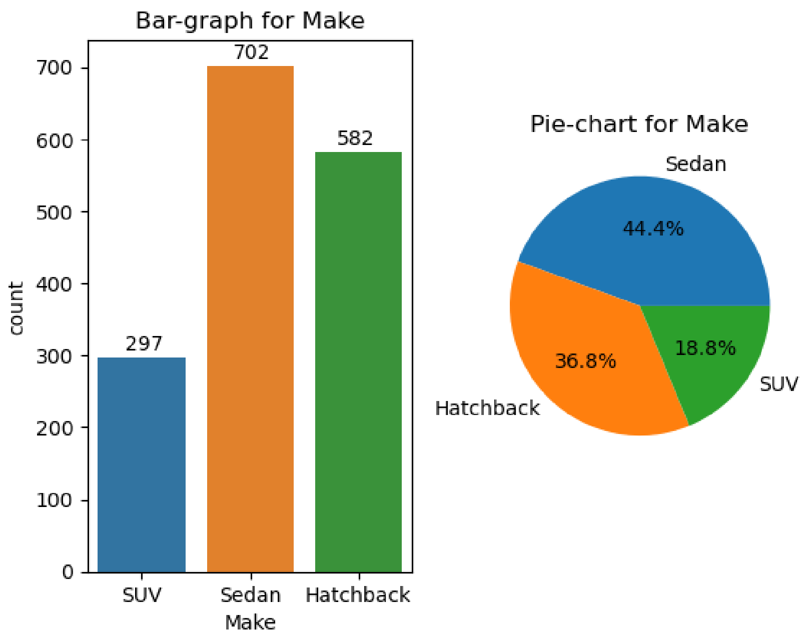


Fig-8 Make Bar and Pie chart

1.2.2 Observations and Insights:

1) We learned that that we have more male values than female in our dataset.
2) Salaried people are more than the Businessperson.
3) Most of them are married people and only approx. 8% are single.
4) Most of the customers are post-graduates.

5) People who took personal loan are almost same as who did not.
6) More people have taken house loan than those who did not.
7) Customers whose partner are working are more than those whose don't.
8) Customers first preference is sedan then hatchback and last SUVs.

1.2.3 Boxplot and Histograms for numerical variables:

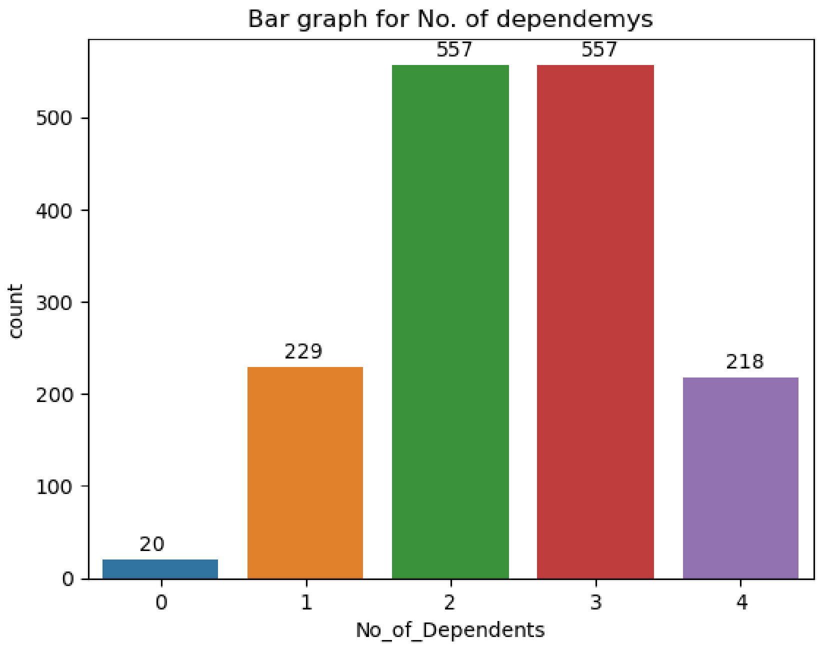


fig-9 No. Of Dependents Bar graph

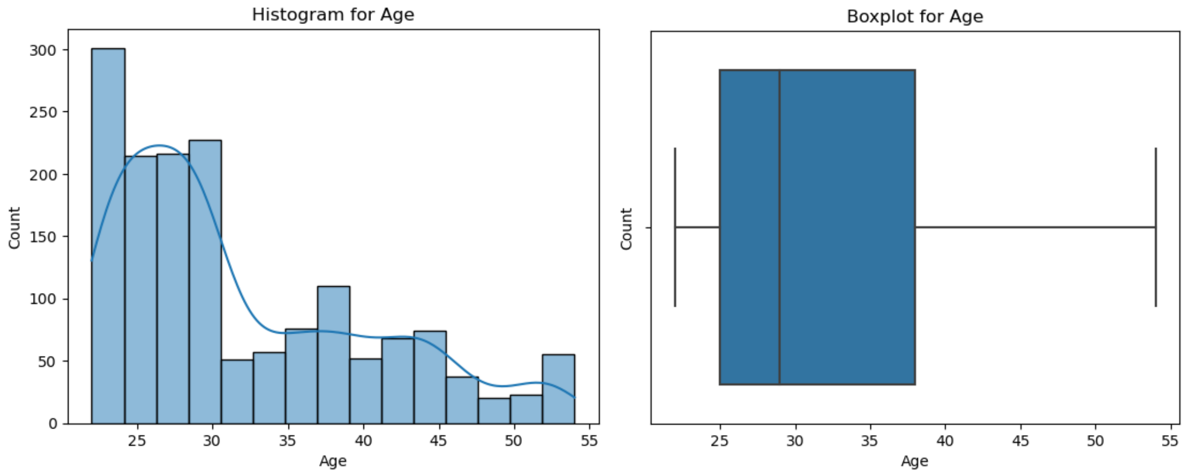


fig-10 Age boxplot and Histogram

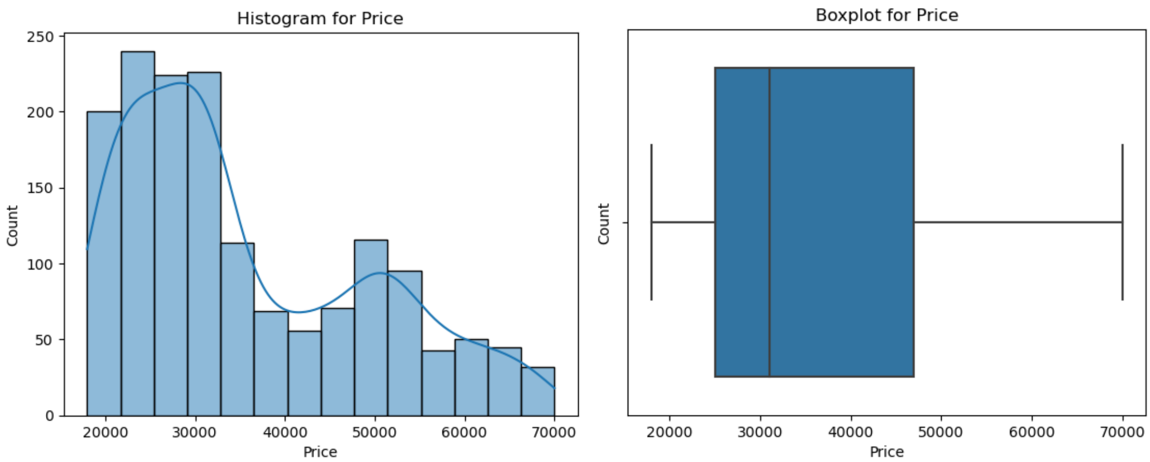


fig-11 Price boxplot and Histogram

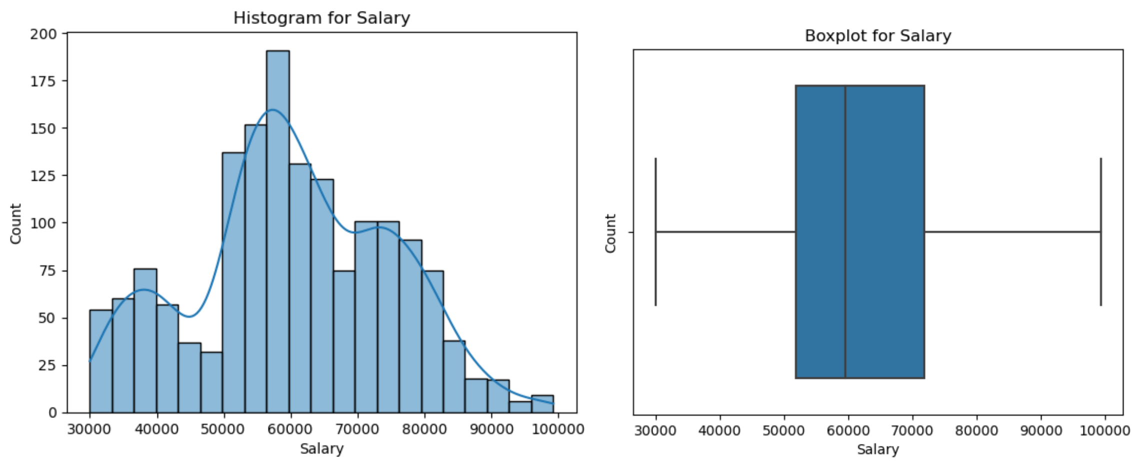


fig-12 Salary boxplot and Histogram

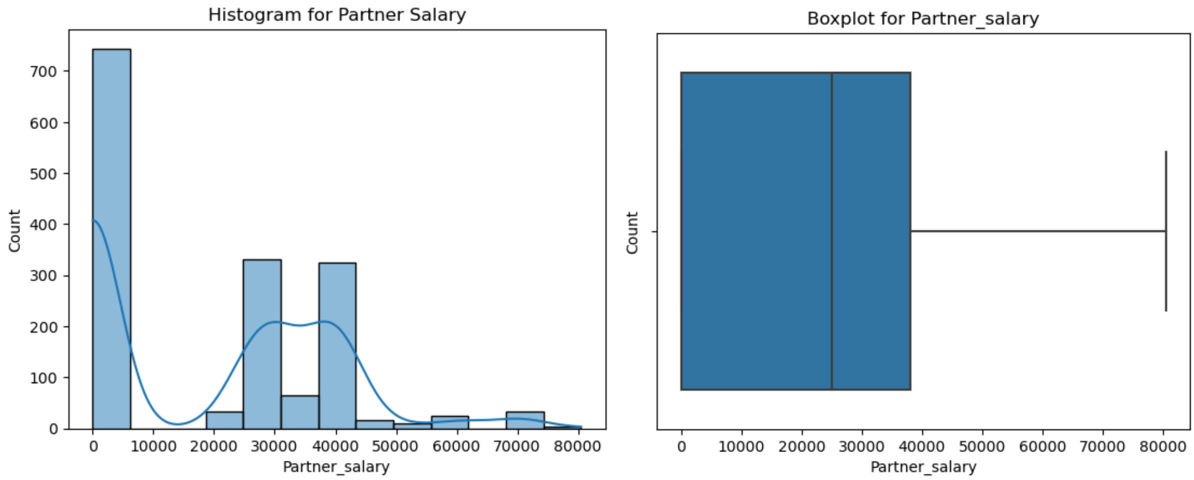


fig-13 Partner Salary boxplot and Histogram

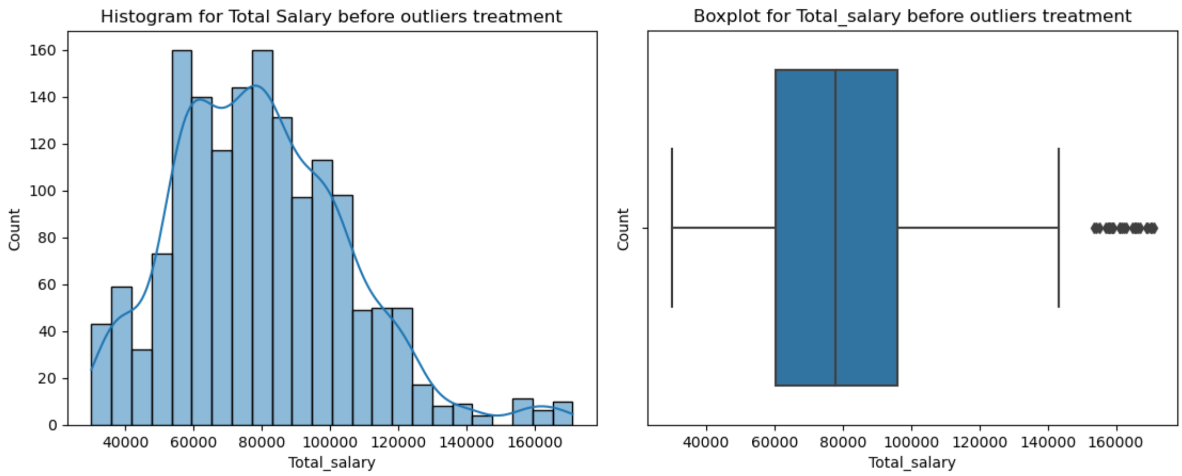


fig-14 Total Salary boxplot and Histogram before outliers treatment

1.2.4 Outliers Treatment:

There are few methods for treating the outliers in our boxplot. And we will perform outliers treatment on Total salary variable using Winsorization method. In this method we will first find out Q1 and Q3 for this variable, we use them to find put IQR value (IQR = Q3-Q1).

After, we get the IQR value we use it to bring larger outliers to upper whiskers using the formula (Q3 + 1.5*IQR) and smaller outliers to lower whiskers using (Q1-1.5*IQR).We plot the graphs again after treatment and we get the below fig-15.

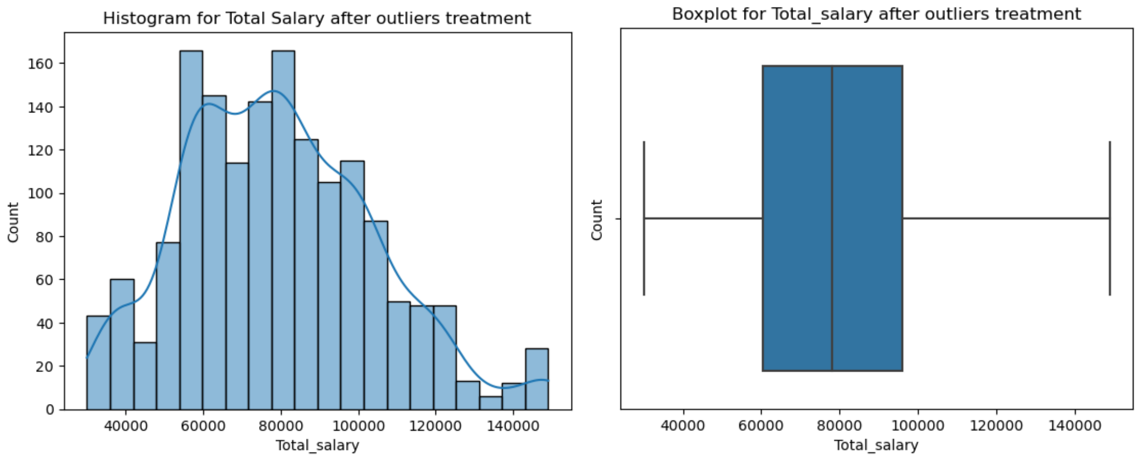


fig-15 Total Salary boxplot and Histogram after outliers treatment

1.2.5 Observation and Insights:

1) Age has a multi-modal distribution and has a positive skewness.
2) Salary variable also has multi-modal distribution within range of 50,000 to 70,000.
3) After the outliers treatment the Total salary variable's skewness has reduced and has shown to have multi-modal distribution.
4) Price has a positive skewness and is a Bi-modal distribution.
5) Except for Total salary that has near normal distribution all other variables have some skewness.

1.3 BI-variate Analysis:

In Bivariate analysis we will explore the relationship between all the numerical variables. Then, see the correlation between all these numerical variables. And, at last we will explore the relationship between categorical vs numerical variables. Also the observations and insights will be written along with the graphs and charts same way as we did in univariate analysis

1.3.1 Numerical Vs Numerical:

First, to explore the relationship between all the numerical variables we will use pair-plot graph from seaborn and find out the patterns and relationships between these variables.

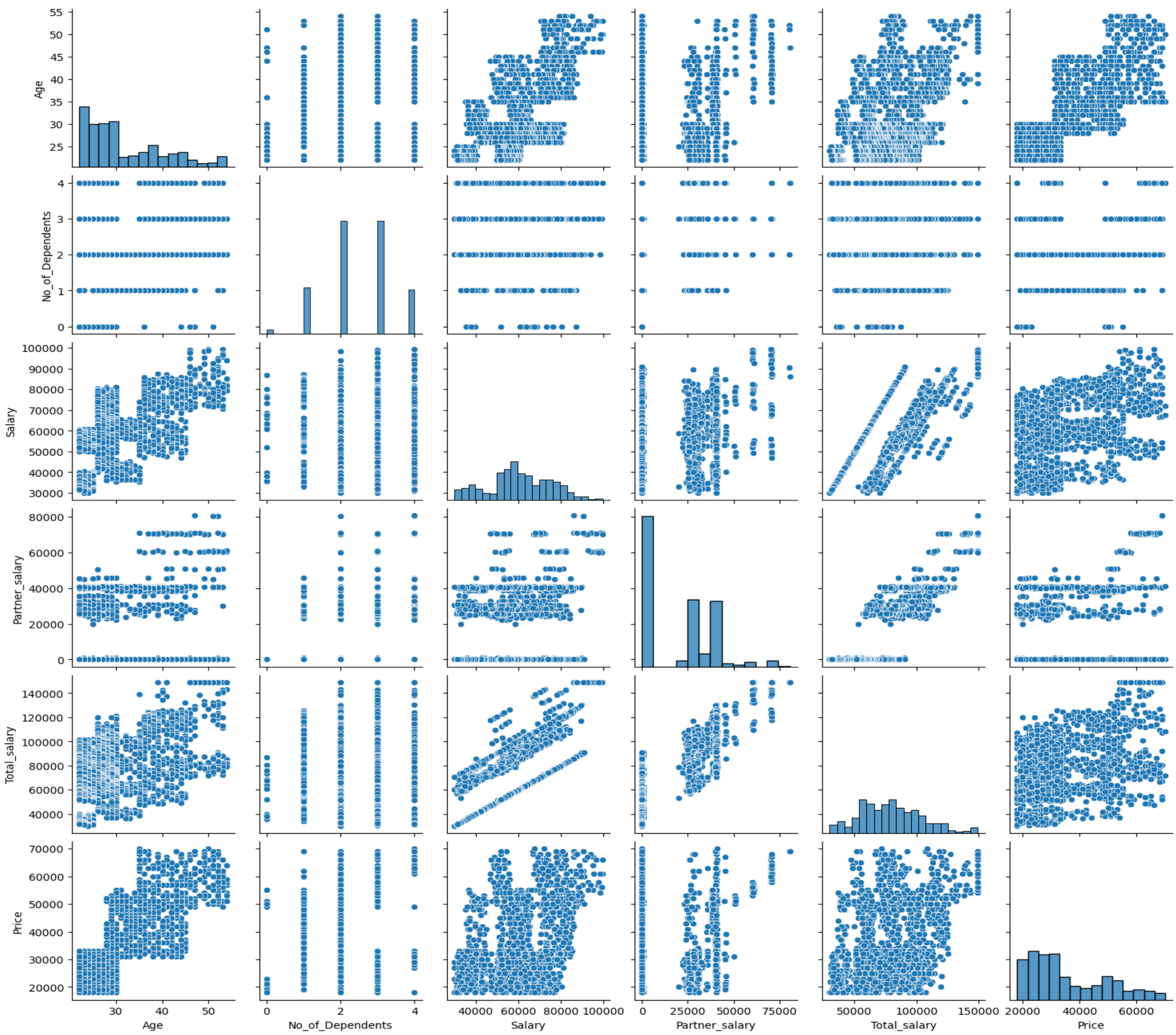


Fig-16 Pair-Plot for all Numerical Variables

Now, we look at correlation between these variables using "corr()" function and we get the below table-9.

	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
Age	1.000000	-0.189614	0.616899	0.135702	0.452844	0.797831
No_of_Dependents	-0.189614	1.000000	-0.031746	0.144320	0.087606	-0.135839
Salary	0.616899	-0.031746	1.000000	0.087155	0.638625	0.409920
Partner_salary	0.135702	0.144320	0.087155	1.000000	0.819248	0.171875
Total_salary	0.452844	0.087606	0.638625	0.819248	1.000000	0.359651
Price	0.797831	-0.135839	0.409920	0.171875	0.359651	1.000000

Table-9 Correlation table of numerical variables

For better understanding we use heatmap for further clarification and get the heatmap graph as shown below.

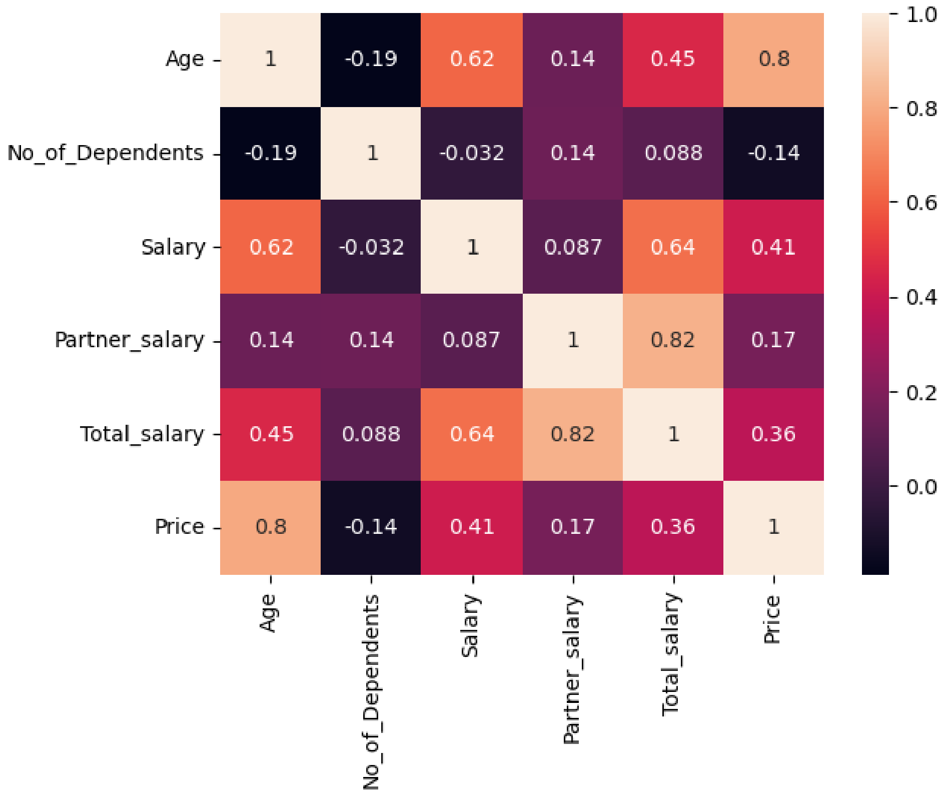


Fig-17 Heatmap of correlation between numerical variables.

1.3.2 Observation and Insights:

1) Age and Price as well as Total salary and Partner salary have the highest correlation of 0.8 and 0.82 respectively.
2) Salary and Total salary seems to have almost a linear relationship as the Salary increase so does the Total salary.
3) Age and Salary seems to have 0.6 correlation which can also be seen in pair-plot.

1.3.3 Numerical vs Categorical each with Observation and Insights :

Now we look at relationship between all numerical vs categorical variables. Each figure shows relationship between a numerical variable vs all categorical variables respectively.

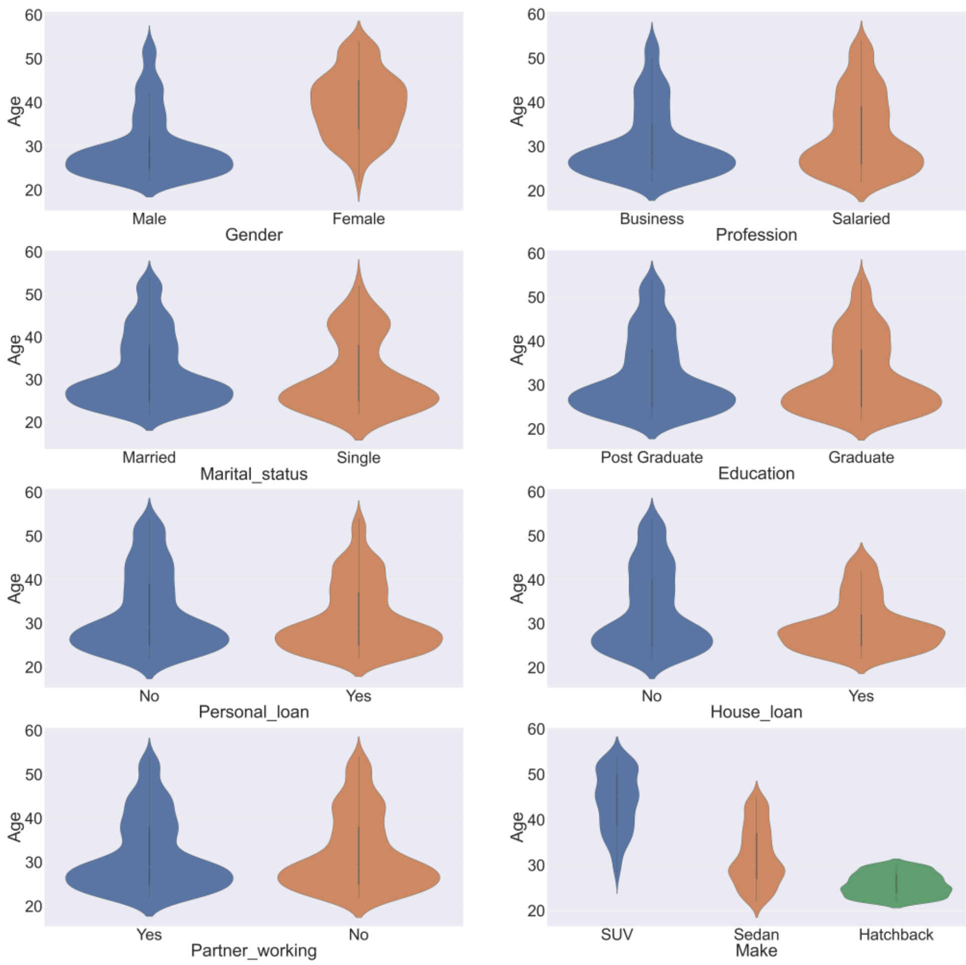


Fig-18 Age vs All categorical variables

Observation and Insight:

- From fig.18 it can be said that most male customers are of age between 20 to 30.
- Younger customer prefers cheaper cars compared to older ones.
- Most of the violin plots are equally distributed.

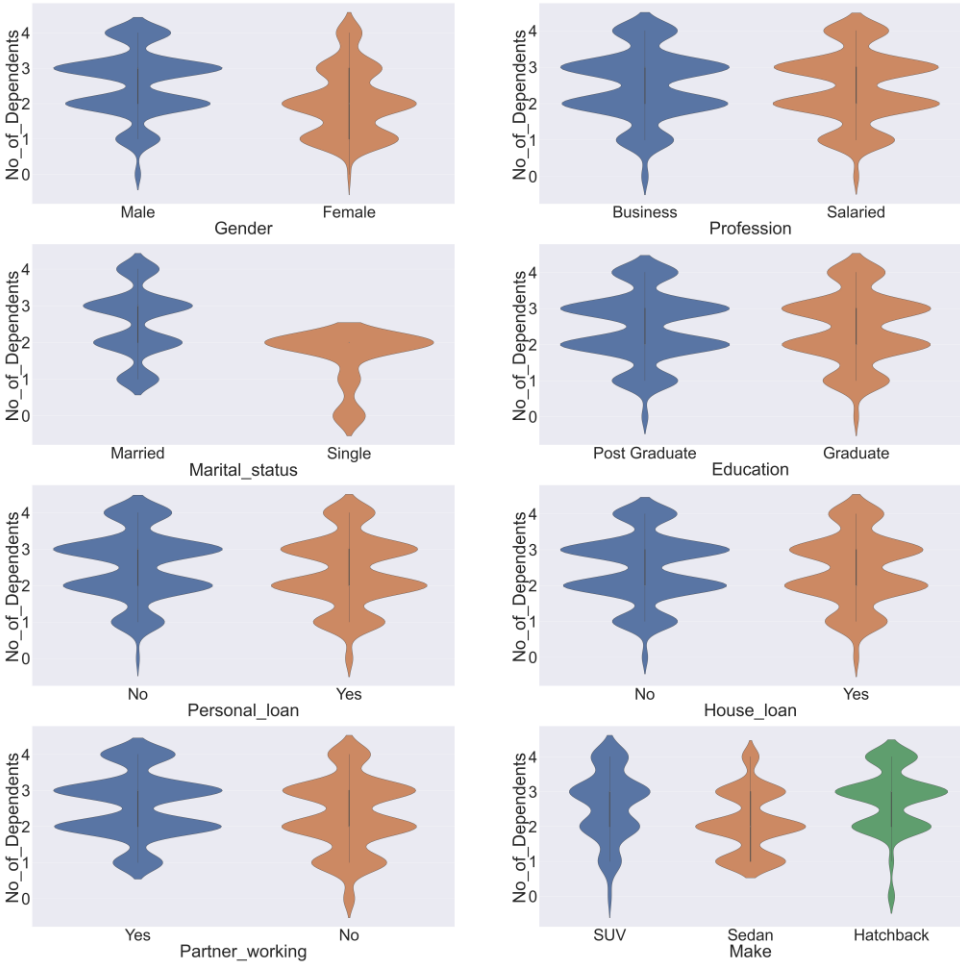


Fig-19 No.of Dependents vs All categorical variables

Observation and Insight:

From fig-19 it can be concluded that the No of Dependents variable has no effect on all the numerical and categorical variables.

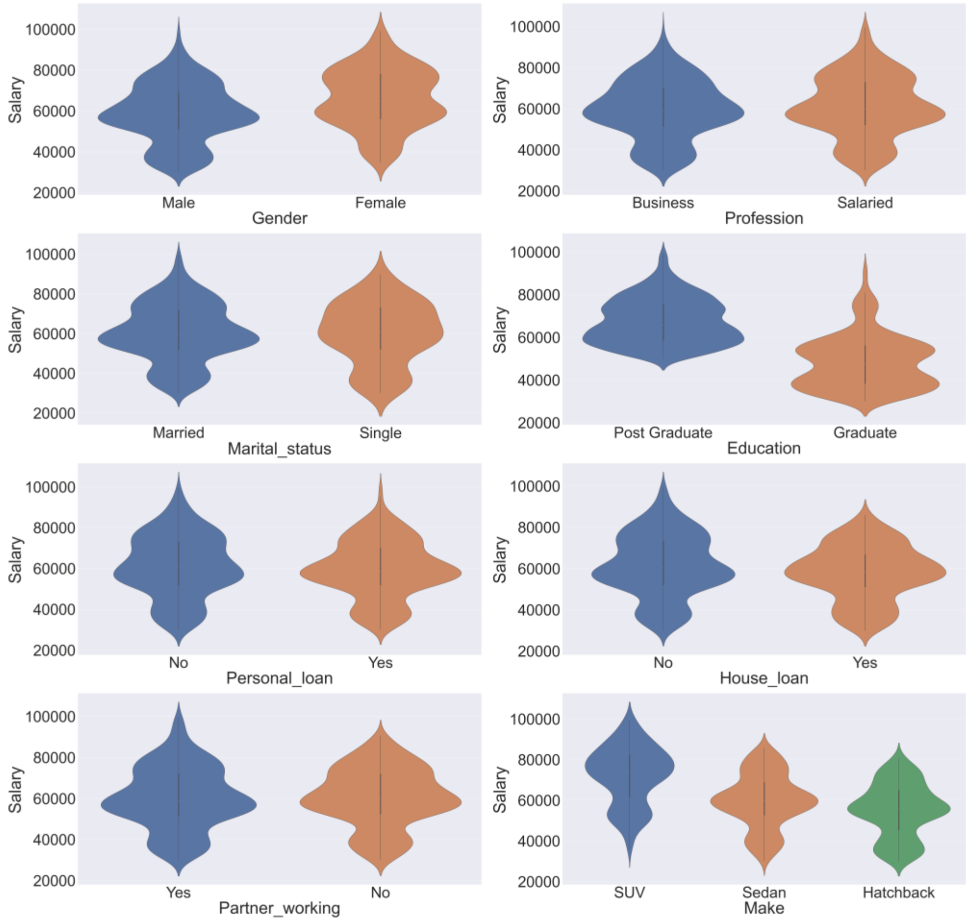


Fig-20 Salary vs All categorical variables

Observation and Insight:

- From fig-20 can be said that salary for post graduate people is higher than the graduate people.
- Salary between 40k to 60k have take a personal loans.

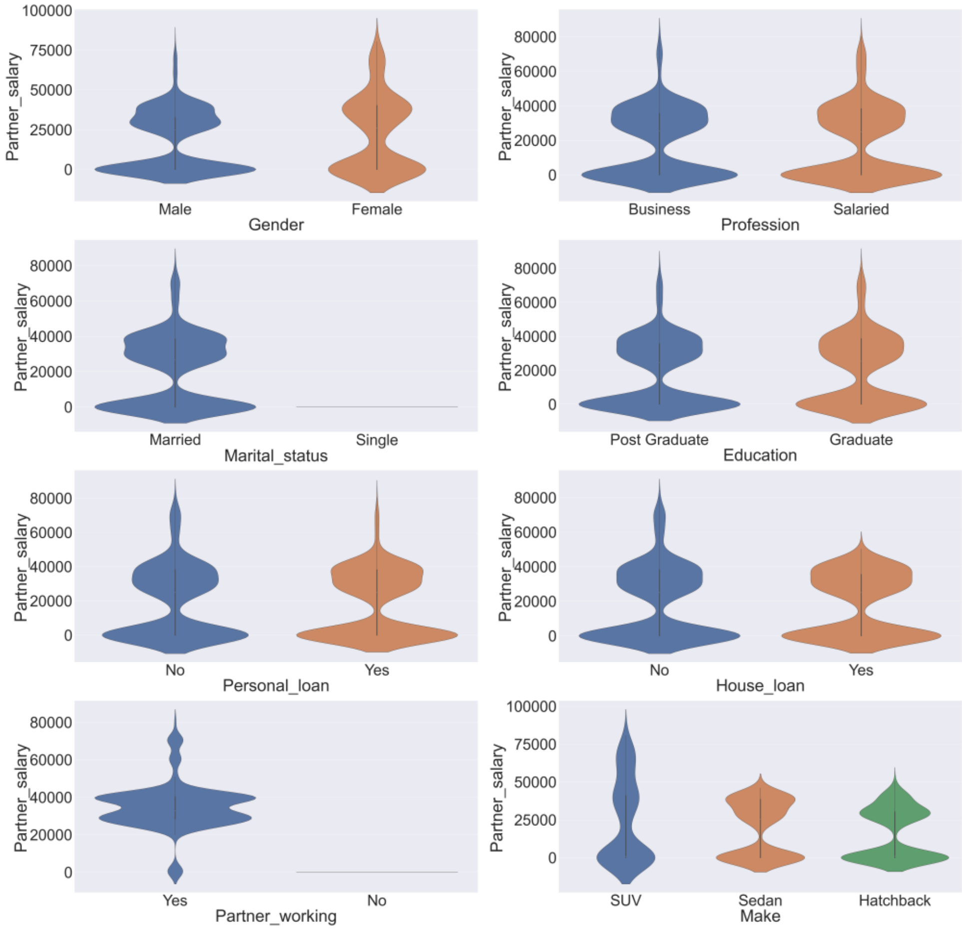


Fig-21 Partner Salary vs All categorical variables

Observation and Insight:

- From fig-21 it is observed that for partner working and marital status for single is zero.
- Most of the partners of Male are at zero compared to females.
- Education plot is almost same for graduate and postgraduate.

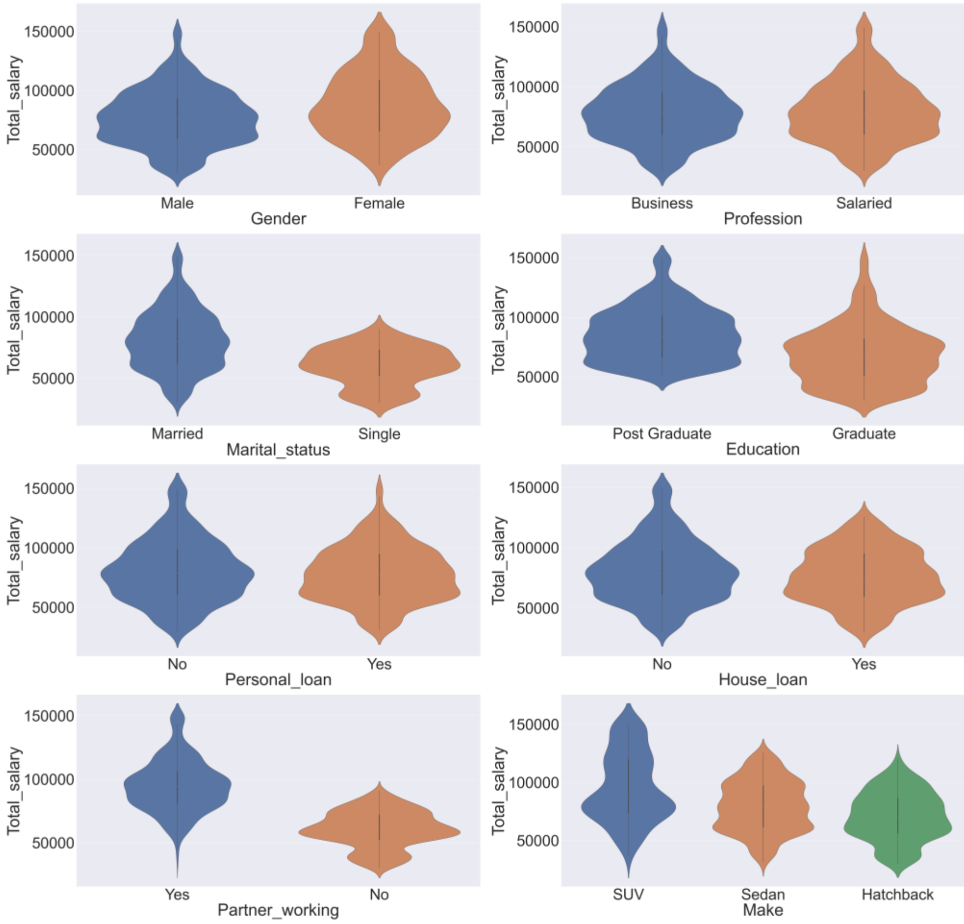


Fig-22 Total Salary vs All categorical variables

Observation and Insight:

- The total salary for married person is higher than that of single person.
- Total salary for whose partner working is higher than for those who do not.

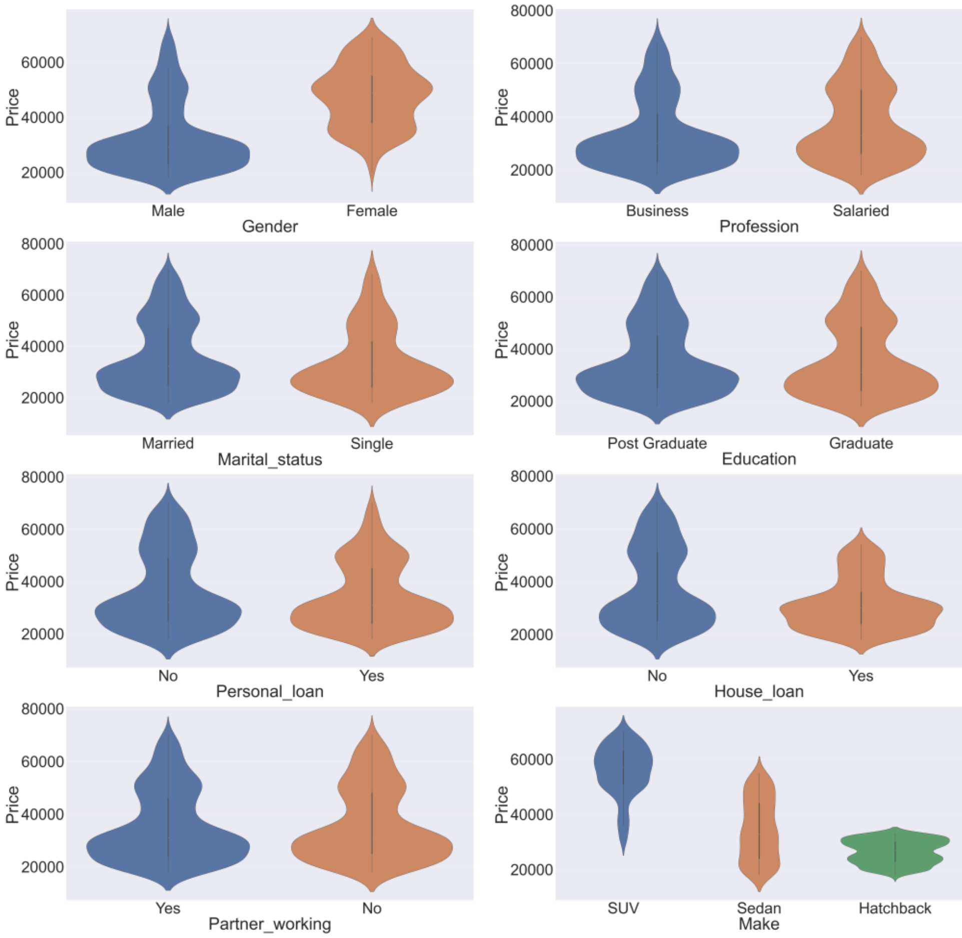


Fig-23 Price vs All categorical variables

Observation and Insights:

1) Male prefer to buy cheaper car then female customers.
2) Partner working variable has no effect on price.

1.4 Key Question:

1.4.1. Do men tend to prefer SUVs more compared to women?

Ans: From the below graph it can be said that, Men prefer SUVs the least. In fact Men prefer sedan and hatchback by a large margin.

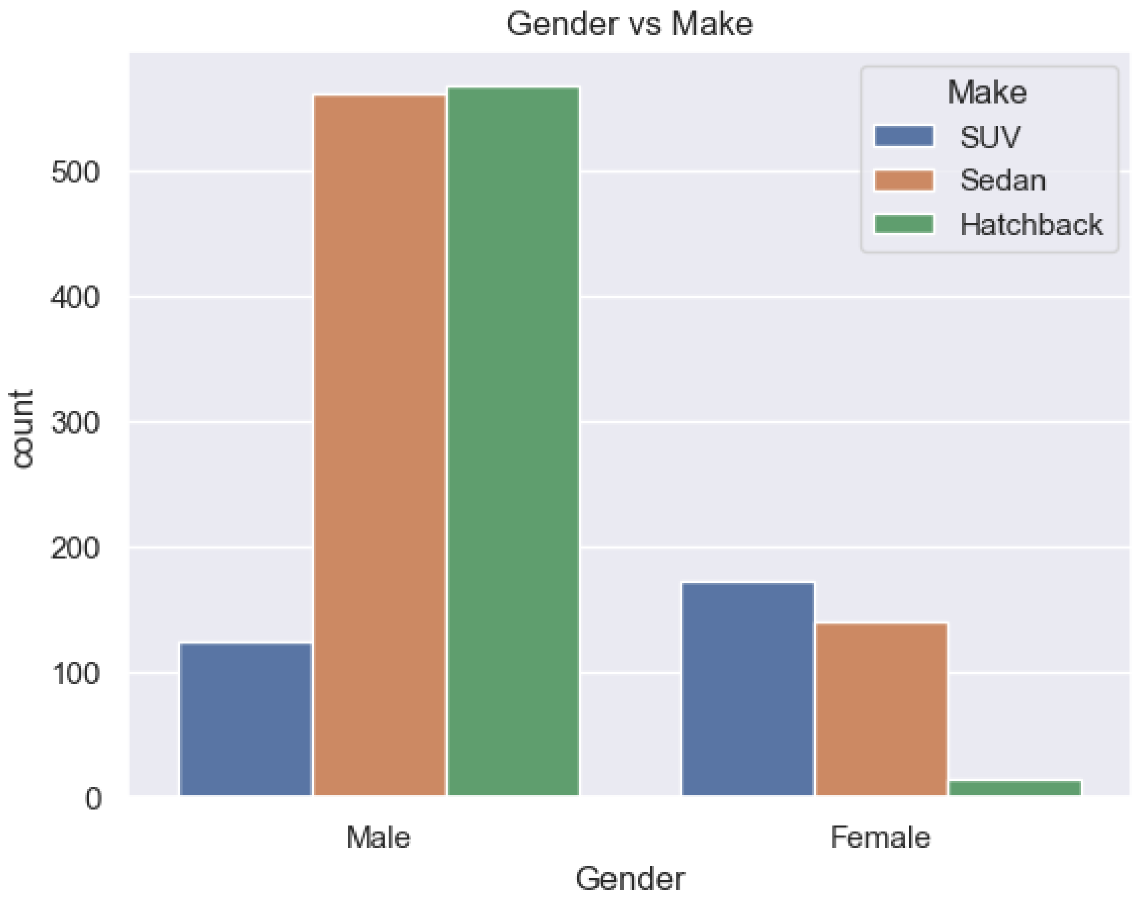


Fig-24 Gender vs Make Bar-graph

1.4.2. What is the likelihood of a salaried person buying a Sedan?

Ans: From the below graphs it is evident that the salaried person's likelihood of buying a sedan is 44.19%.

```
Profession  Make
Business    SUV           89
            Hatchback    290
            Sedan        306
Salaried    SUV          208
            Hatchback    292
            Sedan        396
Name: count, dtype: int64
Profession
Salaried    896
Business    685
Name: count, dtype: int64
Salaried person likelihood for getting Hatchbacks =  0.32589285714285715
Salaried person likelihood for getting SUV =  0.23214285714285715
Salaried person likelihood for getting Sedan =  0.4419642857142857
```

Table-10 Likelihood of Salaried person to get different car model.

Fig-25 Profession vs Make Bar-graph

1.4.3. What evidence or data supports Sheldon Cooper's claim that a salaried male is an easier target for a SUV sale over a Sedan sale?

Ans: For the below graph it can be concluded that Sheldon Cooper's claim is false. In fact, salaried male are easier target for Sedan and hatchback over SUVs

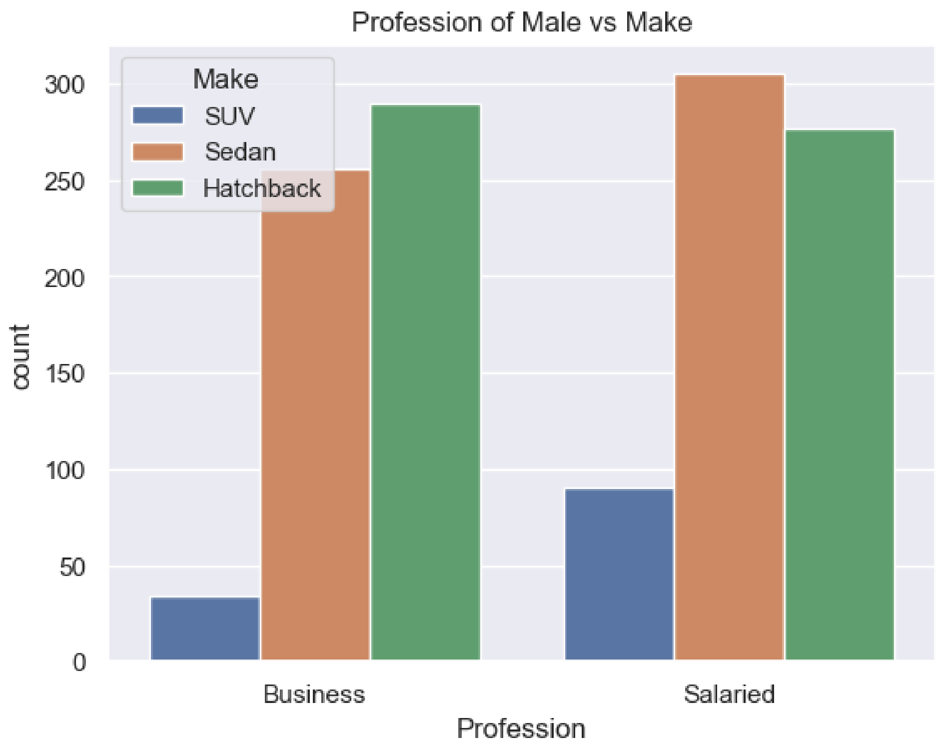


Fig-26 Profession vs Make for men Bar-graph

1.4.4. How does the amount spent on purchasing automobiles vary by gender?

Ans: On an average female customers spend more on car purchase compared to male customers as shown below.

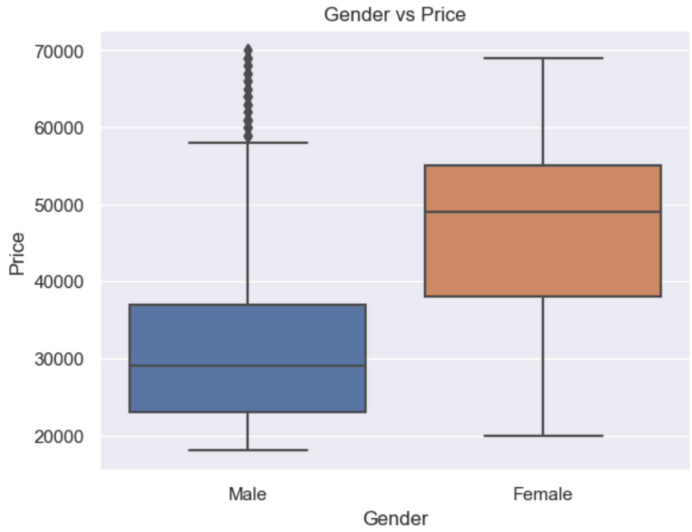


Fig-27 Gender vs Price Box plot

1.4.5. How much money was spent on purchasing automobiles by individuals who took a personal loan?

Ans : As seen from below plot it is evident that the person with personal loan spends slightly more compared to who does not have a personal loan.

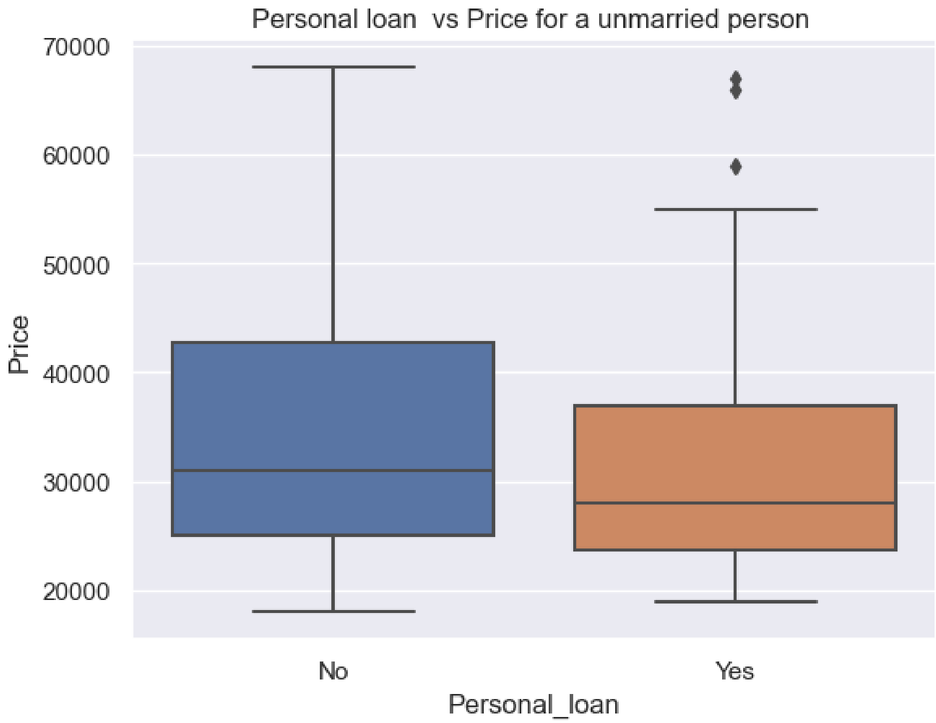


Fig-28 Personal loan vs Price for single person Box plot

1.4.6. How does having a working partner influence the purchase of higher-priced cars?

Ans: It has no influence on purchasing of car as seen below it is almost the same. Thus, it does not matter whether the partner is working or not.

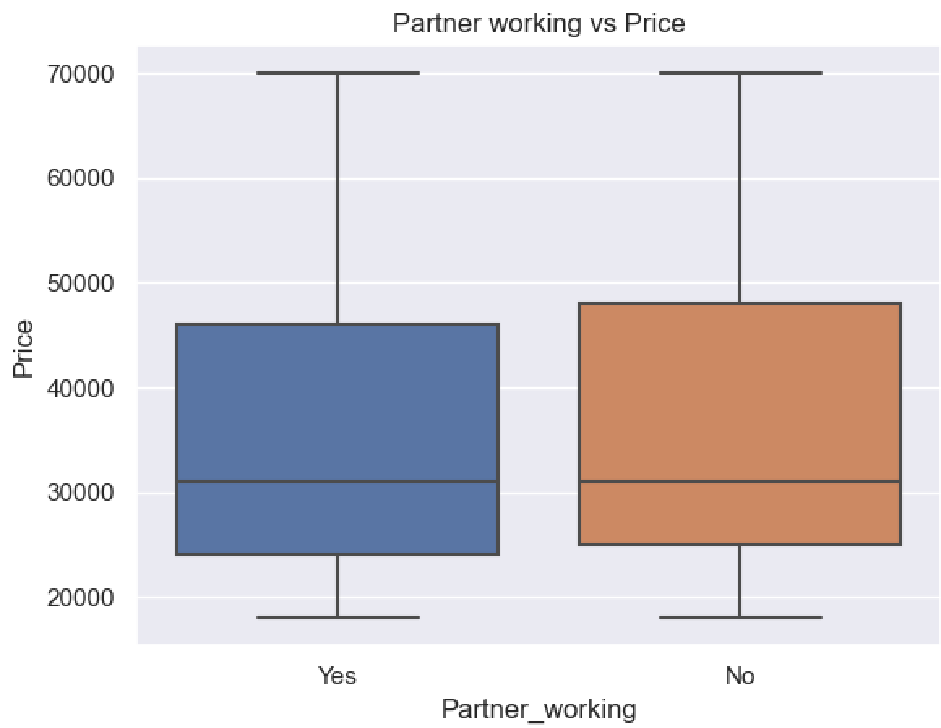


Fig-29 Partner working vs Price Box plot

1.5 Conclusion and business justification:

Thus, we can conclude that from all the analysis and from the below graphs and table that we must focus on Gender, Marital status and Make variables.

From the below graph we can recommend that since, there are more married costumers who come to buy cars. Most married men prefer to buy a sedan and hatchback. While most married women prefer to buy SUVs. The marketing campaign can be focused on married people and their preferences to increase their revenue by making targeted ads, schemes, discounts etc. on their preferred choices.

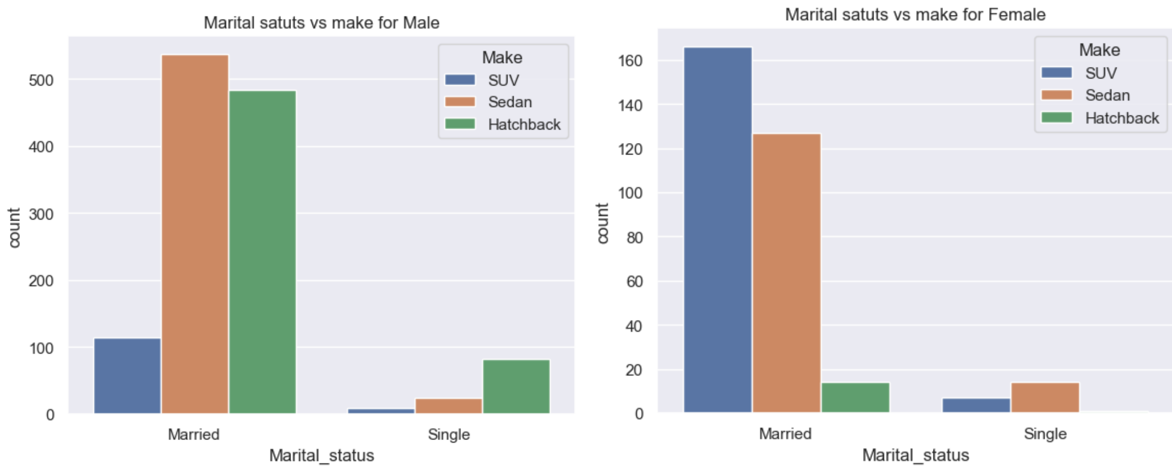


Fig-30 Marital Status vs Price Box plot for Male and Female.

PROBLEM 2: GO DIGT Bank

This bank has been facing high credit card attrition, that lead them to reevaluate there policies related to credit card to ensure customers recieve the right cards for higher spending and intent, which in turn results in profitable relationship between customers and bank.

We as a Data scientist are tasked to analyse the dataset and find out the Five key variables along with business justification.

In this unlike problem-1 where we first analysed whole dataset and then gave our business recommendations without any hypothesis. In this we will first hypothesis some question. And then filter out the variables as per requirements.

We will follow the following steps for seeing the Data structure:

1) we will load the dataset, look at the first and last five rows in the dataset.
2) Then look for duplicated rows using duplicated function. After performing this we found no duplicated rows.
3) Now we will look at the datatypes using info function and we got datetime(1), int64(19), object(8) variables.
4) Then we see for any empty or missing values using isnull function. We found that Transactor_revolver variable has 38 null values. By forming mode on this variable, we found that 'T' data value is the most recurring one, so we replace those empty values with the mode value.
5) Now we performed count values on categorical variables and found 261 anomalies present in Occupation_at_souce variable having value "0". We replace these zero values with the most recurring value that is "Salaried".
6) Now we recheck using isnull to see if we have dealt with the missing values and after performing that yes, we have.
7) To see the statistical summary, we used to describe function on the dataset. And got the min, max, 25%, 50%, 75%, std, mean, count of all numerical variables.

2.1 Key questions:

After the Data Structure is created lets think and ask hypothesis some question based on the requirements and select some variable.

a) How much does the customer on an average uses the credit cards? Can we find out the correlation between there aveage spend, income and credit card limit?
b) What type card of does the customer uses the most?
c) What is the income and occupation of the customer?
d) In last 30 days how much has the customer spend and what is there cc_limit compared to their occupation?

2.1.1. Can we find out any relationship between the customers income and their average spend?

Ans: First we will find out the relationship and correlation between the numerical variables.

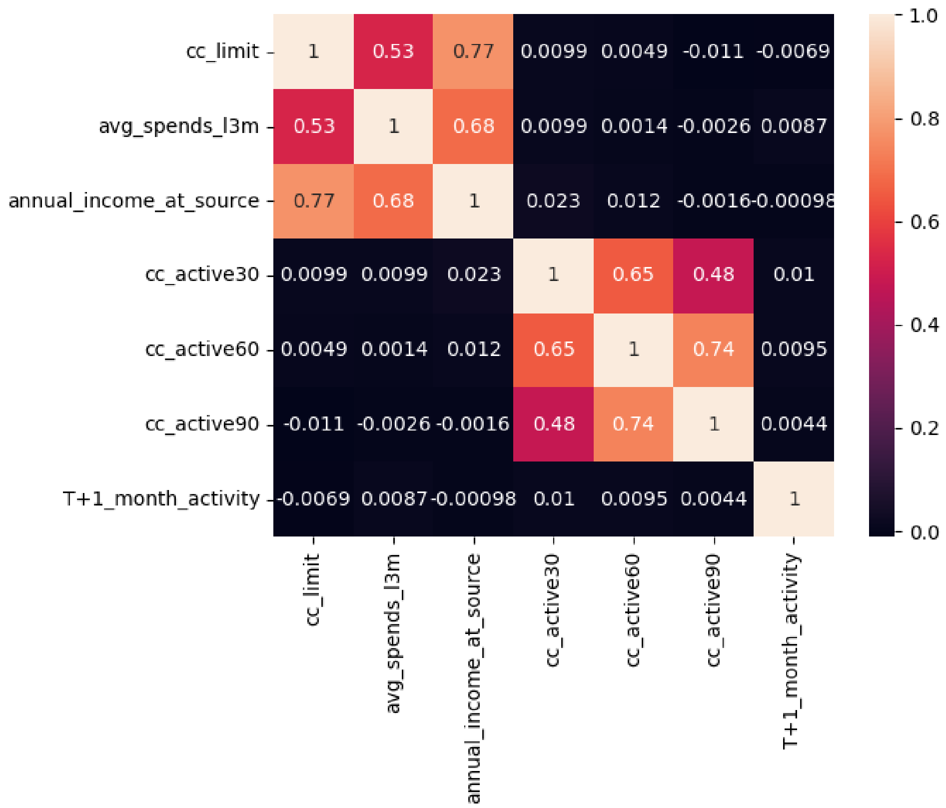


Fig-31 Heatmap of correlation between the numerical values

From the above we found out the three key variables cc_limit, avg_spends_l3m and annual_income_at_source.

There a possiblity of cc_active30 variable of being on the key variable but we must dig much deeper in the dataset.

2.1.2. What type of card of does the customer uses the most?

Ans: From below plot we can say that coustmers uses the reward card type the most and platimum type the least.

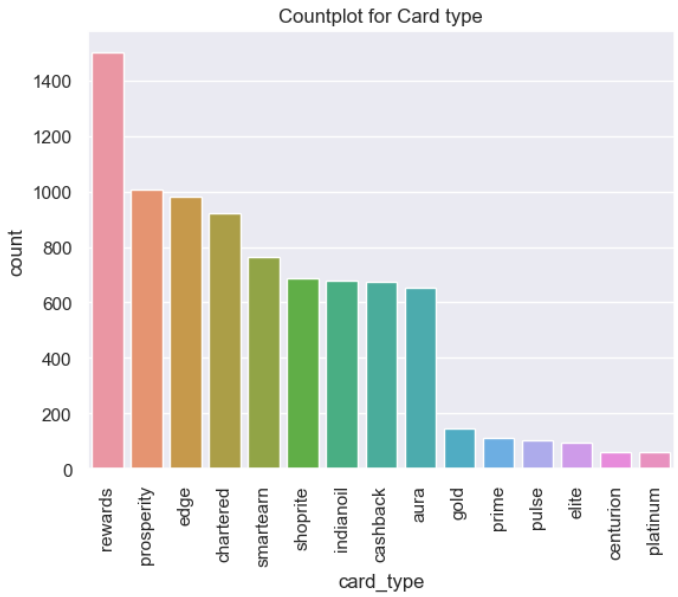


Fig-32 Countplot for Card types.

2.1.3. What is the income and occupation of the customer?

Ans: We can see that income of retired and housewife is less compared to other categories.

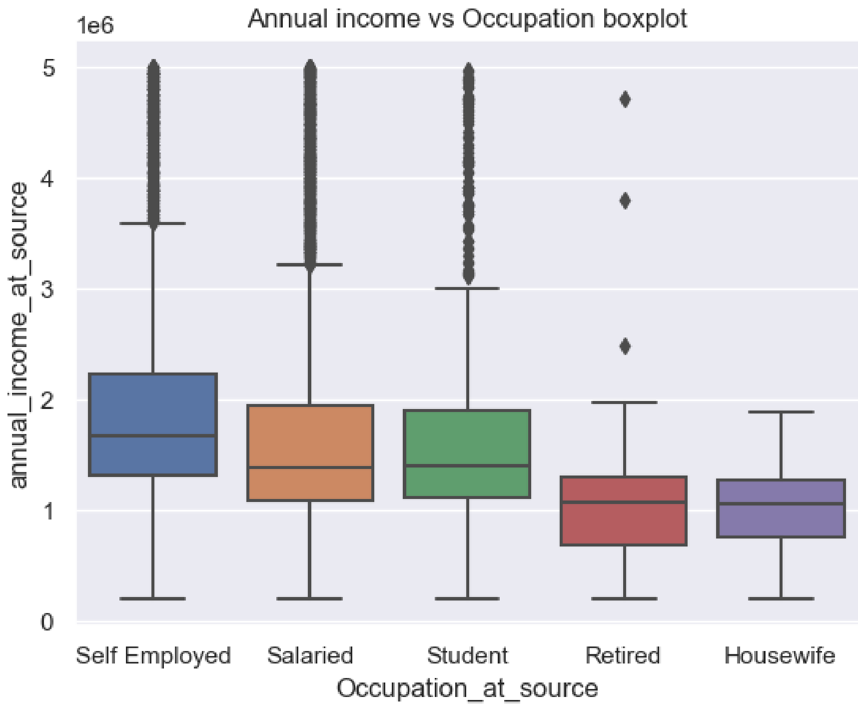


Fig-33 Annual income vs Occupation boxplot

2.1.4. In last 30 days how much has the customer spend and what is there cc_limit compared to their occupation?

Ans: As seen from the below plot it can be said that only the self-employed, salaried and retired are the ones that have actively used their credit cards.

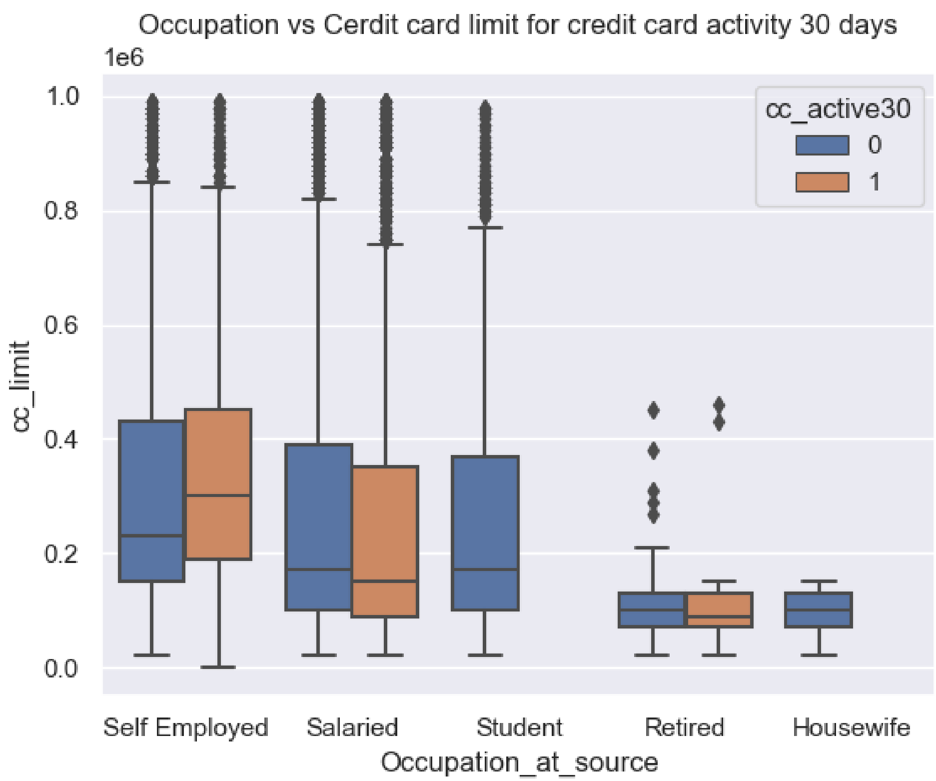


Fig-34 Occupation vs Credit card limit boxplot for credit card activity 30 days.

2.2 The important Five variables with business justifications:

1. avg_spends_l3m:

It represents the average credit card usage in the last three months. This can be as a pattern for customers spending behaviour. It tells me how higher or lower is my customer spending the money. Depending on the preferences a customized campaigns can be organized so, that the bank has customer spend more frequently using credit card.

2. cc_limit:

This variable tells me about the current credit card limit of the customers. Depending on the customers income and occupation as well as credit card activity the limit can be increased or decreased for risk management to lower the number of defaulters.

3. Annaual_income_at_source:

This variable represents the annual income that is recorded during the customers credit card applications. As, this also say about the persons spending power or ability. Knowing the income makes banks decision making better for targeted ads, offers, loans, risk profiling etc.

4. cc_active30:

This variable represents the credit card activity in last 30 days. It tells the bank that how frequently the customer uses the credit card and due this variable it will also help the bank to know that has the use of credit card increased or decreased.

5. Occupation_at_source:

This variable represents the occupation given to the bank at the time of credit card application by the customer. Depending on the occupation target ads, schemes, loans, funds, etc. can be provided to the customer's needs.