AUSTO MOTOR COMPANY DATA ANALYSIS REPORT

Project - 1

JULY 21, 2024

SUBMITTED BY:
Justin John

List of Contents

Objective:

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. In this report we will be generating the relevant insights to understand the purchase pattern of customers belonging to diverse field and category based on various attributes.

```
(1581, 14)
```

The dataset given to us is Automobile data which contains 14 columns and 1581 rows.

1. DATA OVERVIEW

1.1 Data dictionary:

```
<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 1.1

The data has 1581 observations with 14 entries.

5 integer type as numerical variables,

1 float type as numerical variables and

8 object type as categorical variables.

1.2. Missing Values

Checking for the missing values in the dataset:

```
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 1.2

From the above table 1.2, there are 53 null values in Gender and 106 null values in Partner_Salary dataset.

1.2.1. Treating NAN values:

When we check the Gender and Partner_salary we can find some discrepancies and data is not consistent.

Also, when we check unique values of Gender column, we see another extra attributes like **Femle**, **Femal**, nan as mentioned so, may be the backend agent would have mistyped it to feed the data.

```
Gender
Male      1199
Female     327
Femal        1
Femle        1
Name: count, dtype: int64
```

Table 1.2.1

1.2.2. Treating Missing values:

So after fixing the discrepancies, such as fixing the null values and adding the count value of **Femle** and **Femal** to **Female** attribute. Then the data looks like:

```
Gender
Male      1199
Female     329
Name: count, dtype: int64
```

Table 1.2.2.a

Now we can see only 2 attributes such as **Male** and **Female**.

And there are 53 missing values in Gender column, which has been added to the Male attribute by taking the Mode value of Gender column we see the output shows as 0 Male (as in Table 1.2.2.b) hence we filled all the missing values in gender column as Male.

```
0    Male
Name: Gender, dtype: object
```

Table 1.2.2.b

Since Partner_Salary attribute had 106 missing values so we imputed all missing values using **Mean**.

```
Age                1581
Gender             1581
Profession         1581
Marital_status     1581
Education          1581
No_of_Dependents   1581
Personal_loan      1581
House_loan         1581
Partner_working    1581
Salary             1581
Partner_salary     1581
Total_salary       1581
Price              1581
Make               1581
dtype: int64
```

Table 1.2.2.c

From the above plot, now we can see missing values in Gender and Partner_salary has been already treated.

1.3 Checking for duplicates

Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	Price	Ma

Table 1.3

As we can see in the above Table there are no duplicates records found.

1.4 Summary of the dataset

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	1581.0	NaN	NaN	NaN	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
Gender	1581	2	Male	1252	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Profession	1581	2	Salaried	896	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Marital_status	1581	2	Married	1443	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Education	1581	2	Post Graduate	985	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_Dependents	1581.0	NaN	NaN	NaN	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Personal_loan	1581	2	Yes	792	NaN	NaN	NaN	NaN	NaN	NaN	NaN
House_loan	1581	2	No	1054	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Partner_working	1581	2	Yes	868	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	1581.0	NaN	NaN	NaN	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1581.0	NaN	NaN	NaN	20230.65588	18909.850652	0.0	0.0	24900.0	38000.0	80500.0
Total_salary	1581.0	NaN	NaN	NaN	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	1581.0	NaN	NaN	NaN	35597.72296	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0
Make	1581	3	Sedan	702	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 1.4.1

	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	20230.655880	18909.850652	0.0	0.0	24900.0	38000.0	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 1.4.2

The above table 1.4.2 depicts the six-point summary of the continuous attributes.

Analyzing at the age column, we can see that the distribution of the adult population is between the minimum of age 22 years and with maximum age of 54 years. And 25% & 50% of the people having age of 25 and 29. And they have number of

dependents as value 2 with a Total salary of 60500 and 78000. And 75% of the people having age of 38 years has number of dependents as value 3 with a total salary of 95900. And the number of dependents shows as 0 for minimum age of 22 years and value 4 being the highest number of dependents for people having age with 54 years.

1.4.1 Let's plot the histogram to see the distribution of the continuous features continuously.

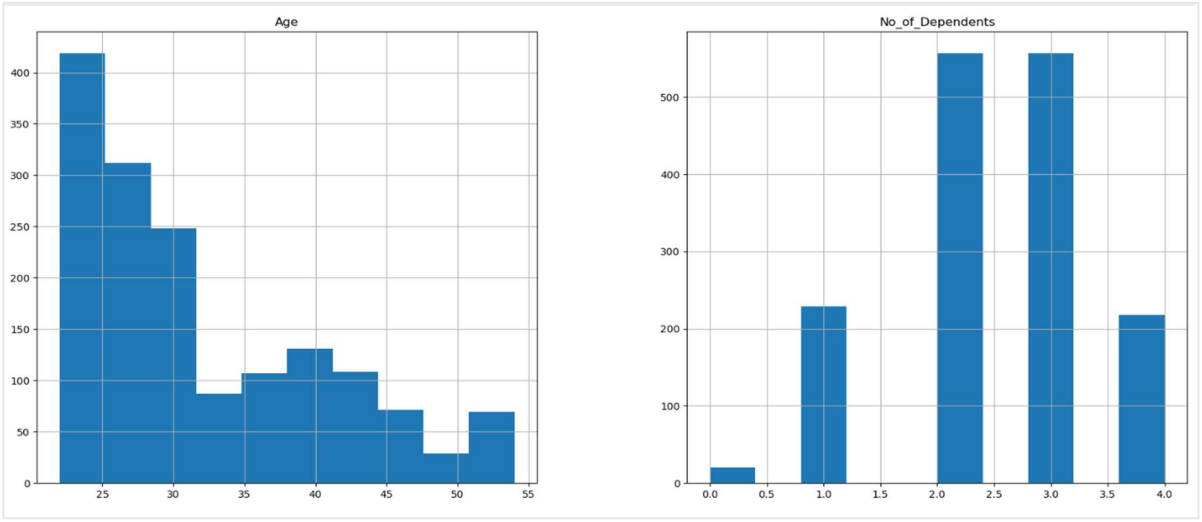

Fig 1.4.1

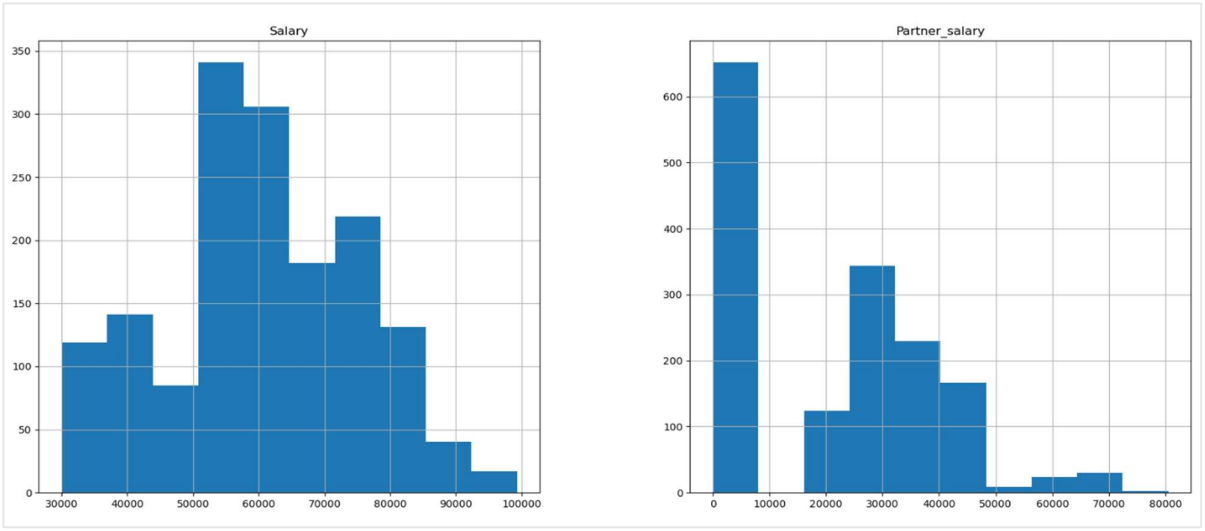

Fig 1.4.2

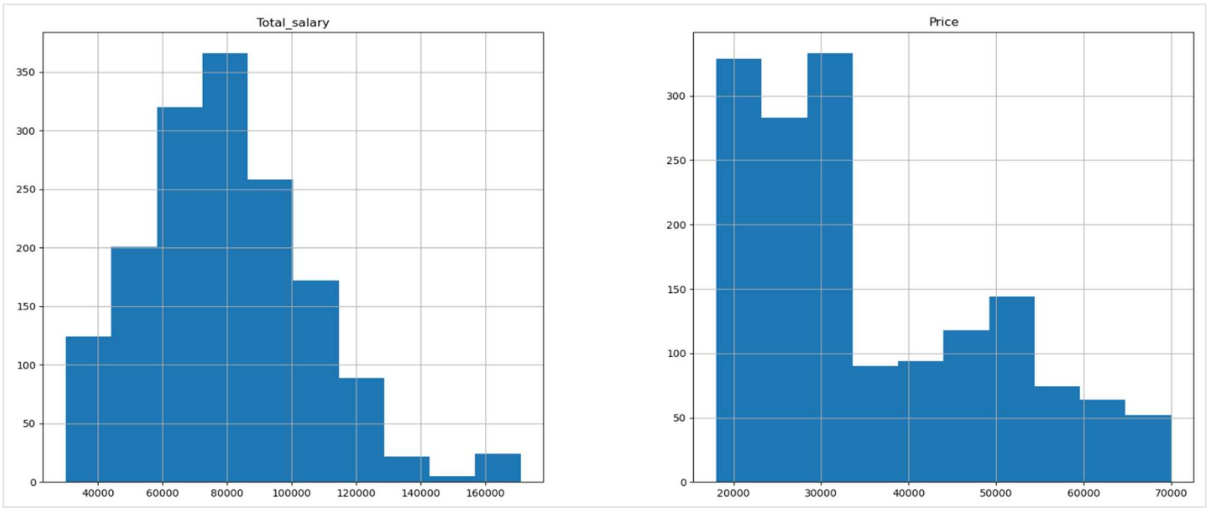

Fig 1.4.3

Looking at the above histograms (Fig 1.4.1,1.4.2, 1.4.3), we can see that age (left-skewed) no_of_dependents is not uniformly distributed (right-skewed), salary is uniformly distributed, partner_salary is not uniformly distributed (left-skewed), total_salary (left-skewed) and price is left-skewed.

1.4.2 Now we shall look at how the variables are distributed with the help of countplot.

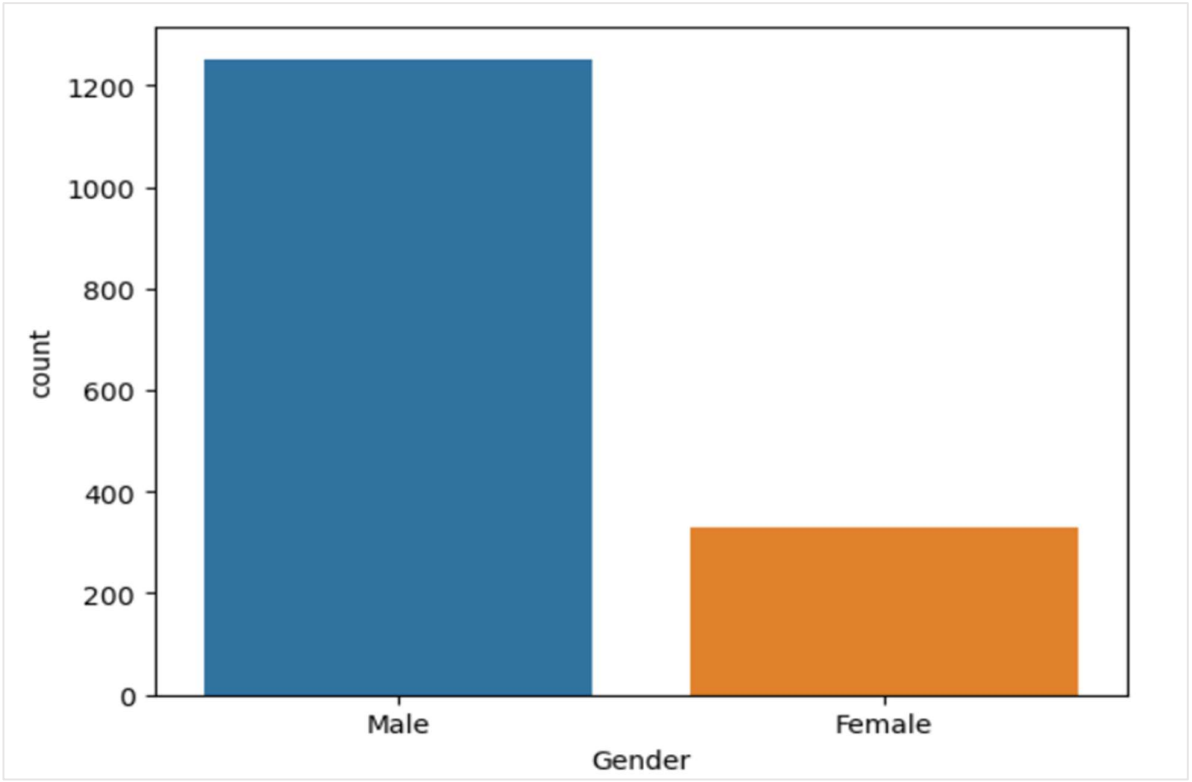

Fig 1.4.2.a

We can see that, the Gender count of Male is much higher than Female.

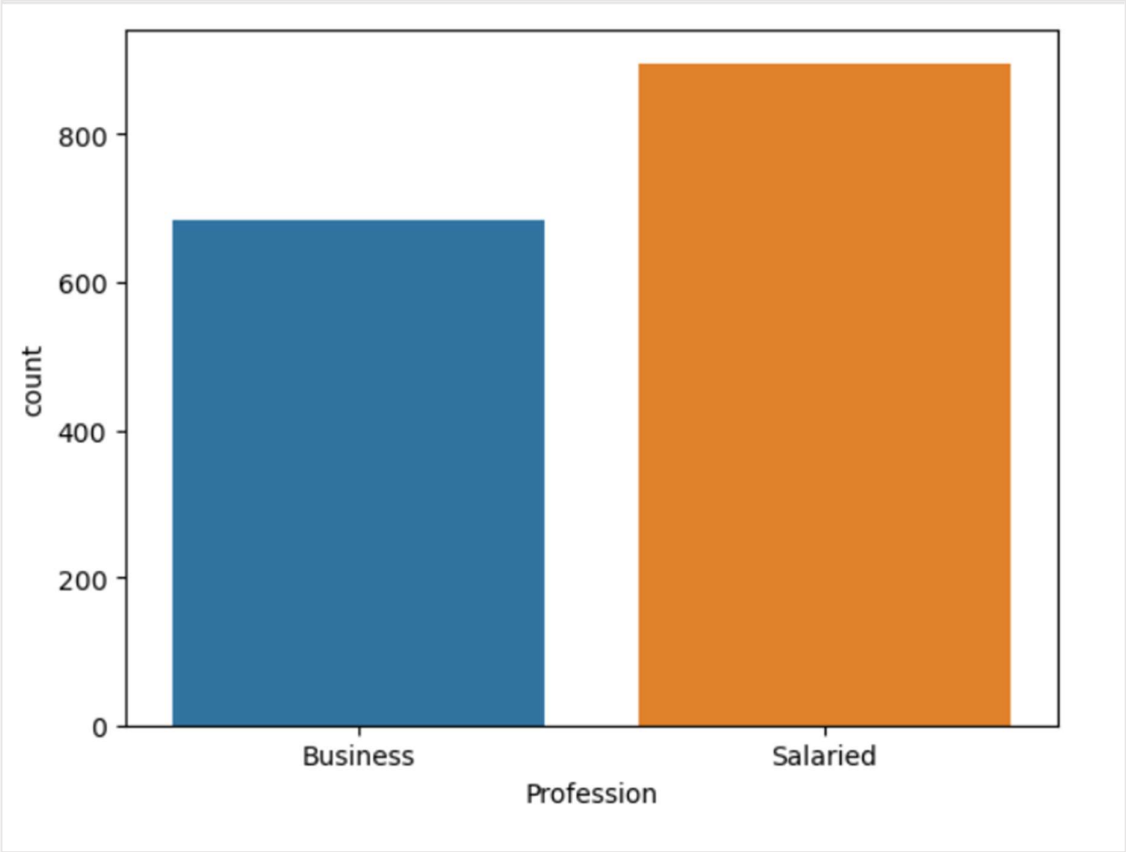

Fig 1.4.2.b

We can see the Salaried Profession is more than the Business Profession.

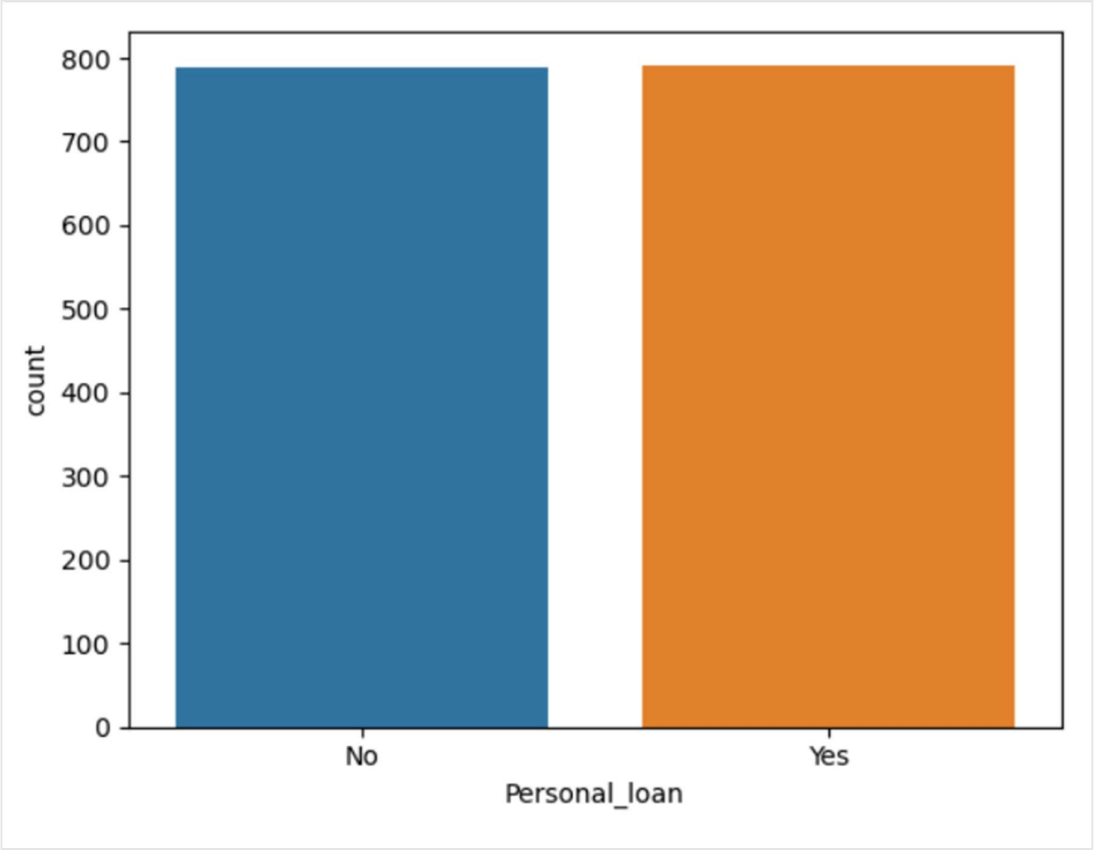

Fig 1.4.2.c

As depicted, we can see personal_loan status shows as same for both the values : Yes and No.

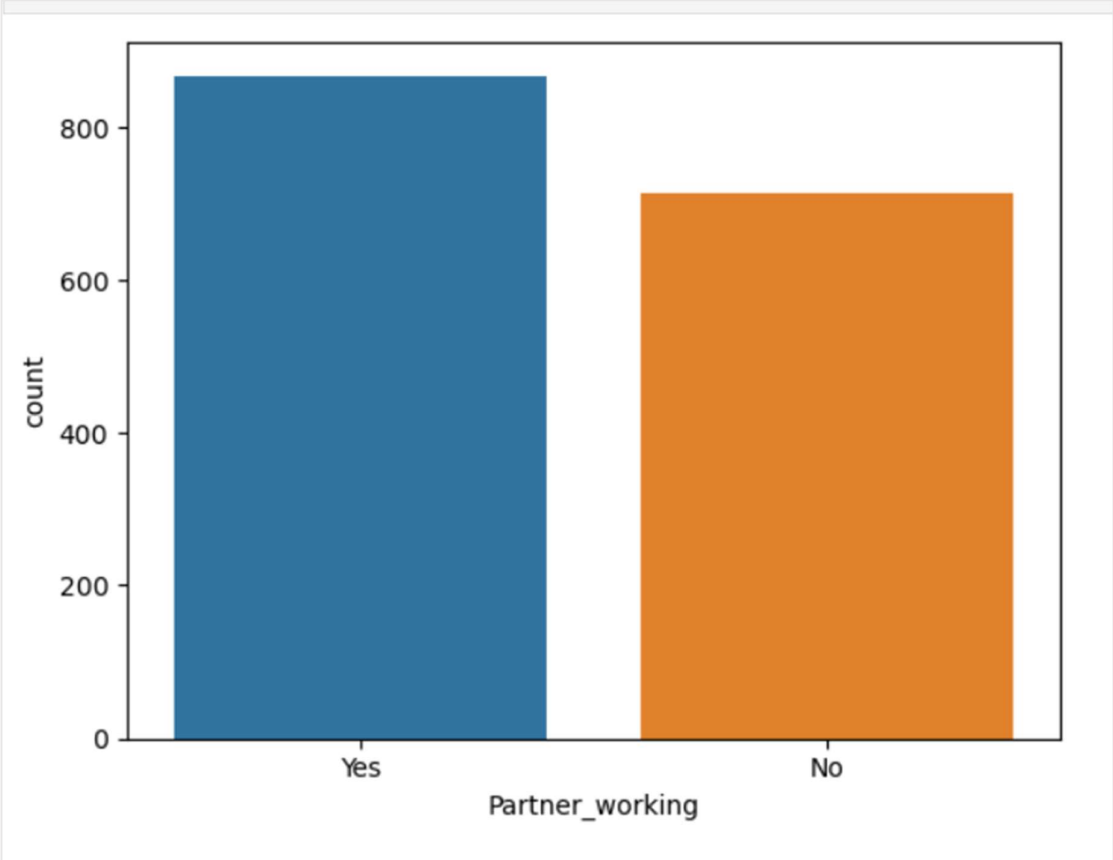

Fig 1.4.2.d

The most of the Partners have a job.

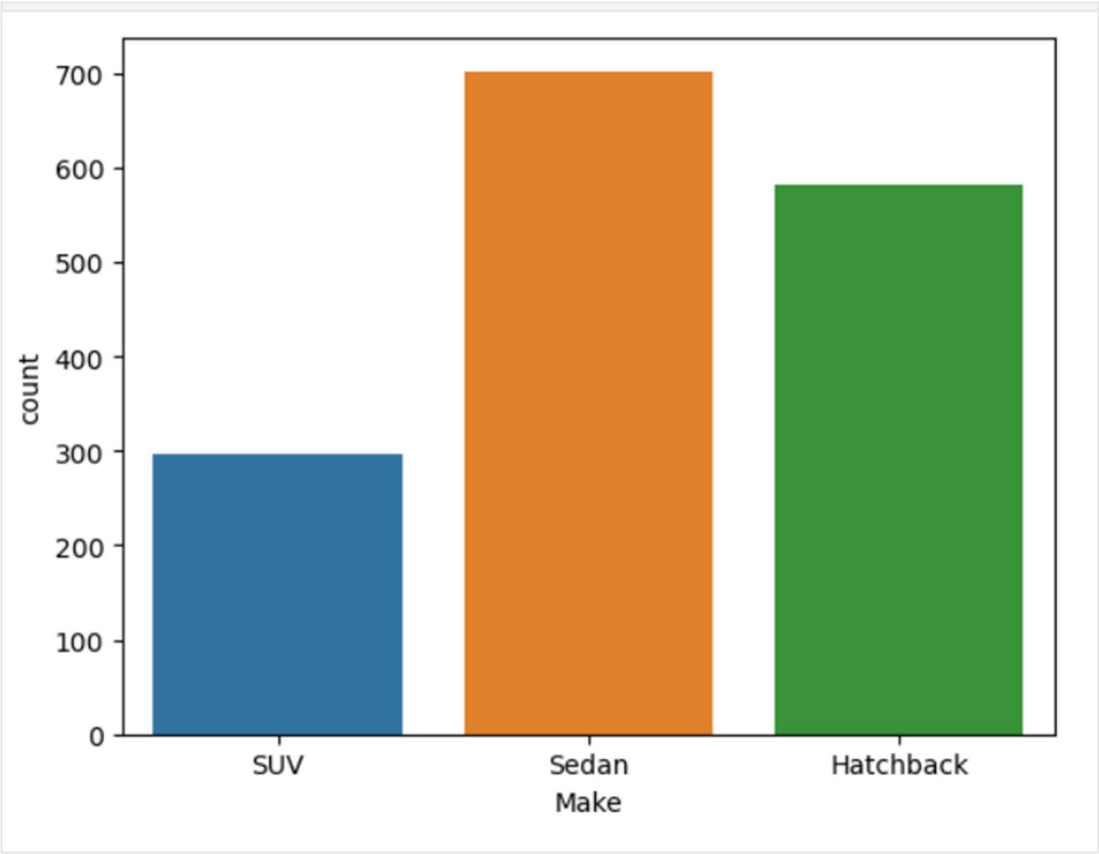

Fig 1.4.2.e

The above plot depicts that Brand ‘Sedan’ is the most purchased followed by ‘Hatchback’ and the least is ‘SUV’.

Gender	Female	Male
Make		
Hatchback	15	567
SUV	173	124
Sedan	141	561

Table 1.4.2.a

So, here we can see that most of the Female Gender uses SUV automobile and Gender Male uses more Hatchback automobile.

Marital_status	Married	Single
Make		
Hatchback	498	84
SUV	281	16
Sedan	664	38

Table 1.4.2.b

The table depicts that the married person buys more cars than the un-married person. And most of them have purchased Sedan Brand models.

1.4.3 Variables are distributed with the help of Histplot.

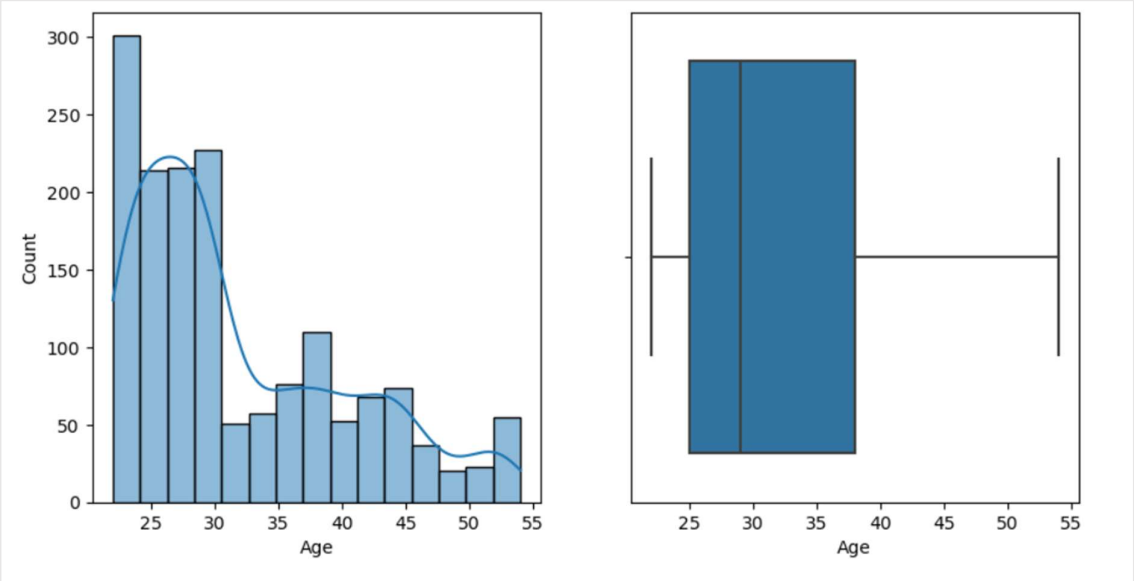

Fig 1.4.3.a

Here using histplot, we can see the Age is rightly-skewed distribution and with the help of boxplot, we can see no outliers found in Age variable.

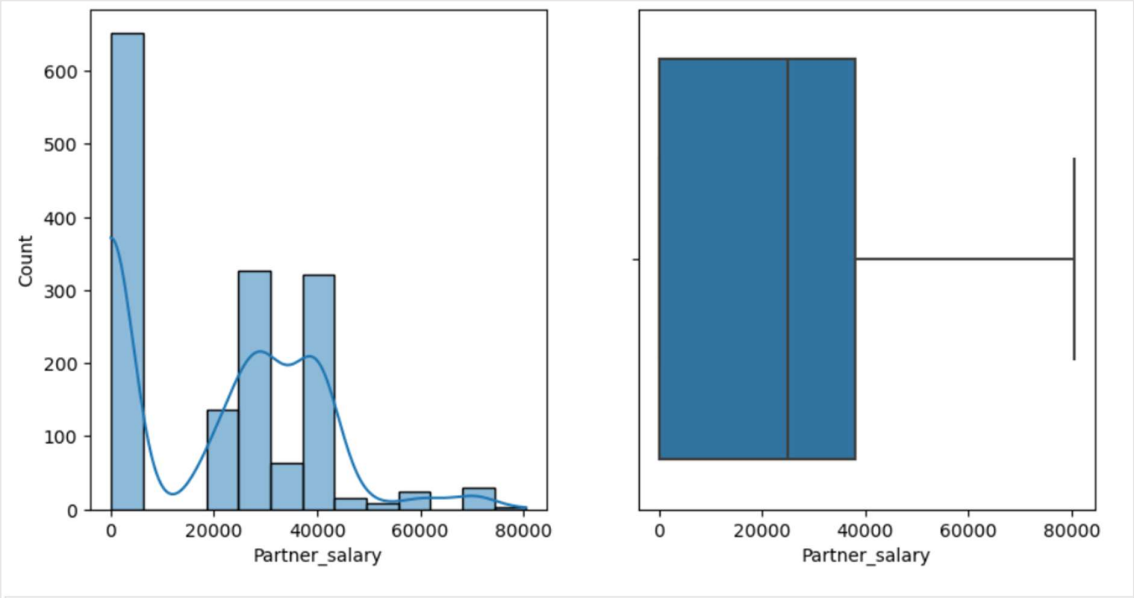

Fig 1.4.3.b

In the same manner, we can analyze that even Partner_Salary has no outliers which we tried to check using boxplot and through the help of histplot, we see the distribution of Partner_Salary is again right-skewed distribution.

1.4.4 Bi-variate distribution of every possible attribute pair

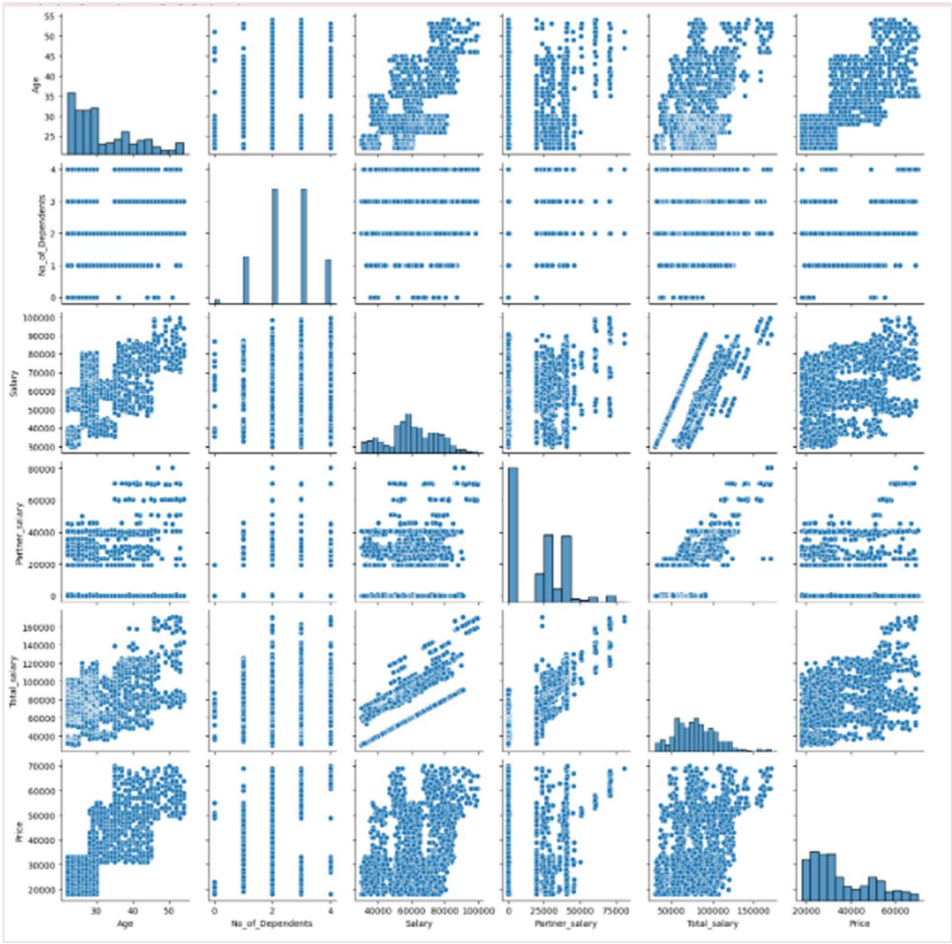

Fig 1.4.3

By using the pairplot we can see the bivariate distribution. As age gets increased even the salary is getting increased and looks like people with more age are spending more price on the purchases.

1.4.5 Correlation :

Now let's have a look at correlation with the help of heatmap.

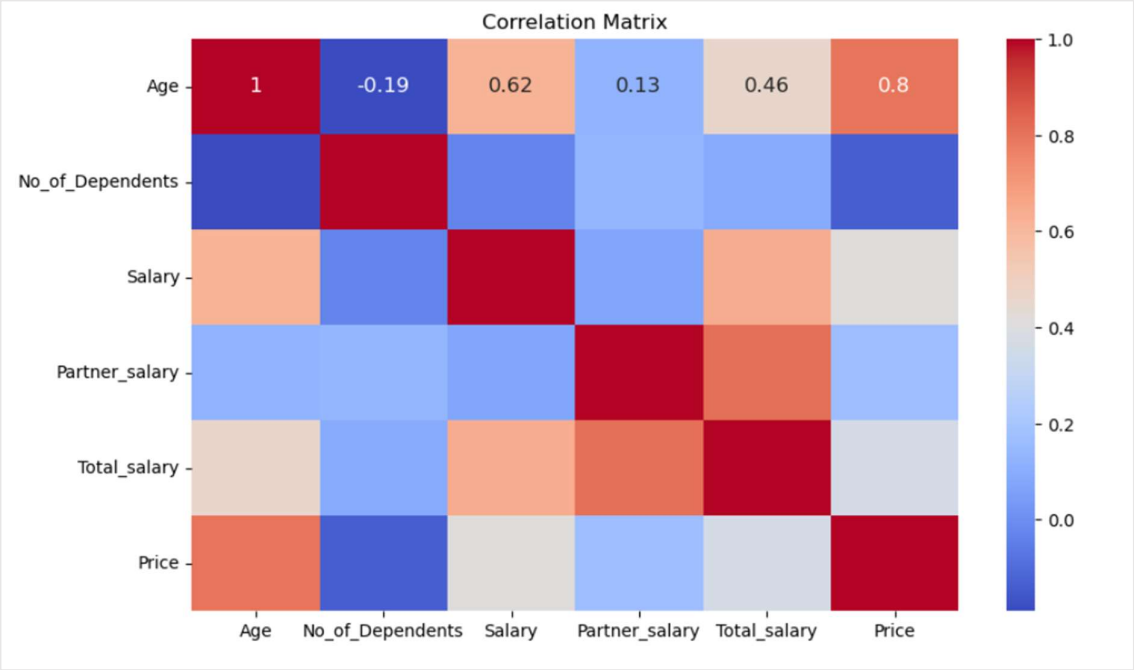

Fig 1.5

2. Implementing Univariate Analysis

2.1 Variables used : Age, No_of_Dependents (Numeric with numeric)

	Age	No_of_Dependents
count	1581.000000	1581.000000
mean	31.922201	2.457938
std	8.425978	0.943483
min	22.000000	0.000000
25%	25.000000	2.000000
50%	29.000000	2.000000
75%	38.000000	3.000000
max	54.000000	4.000000

Table 2.1

As in table 2.1, the Age attribute has a minimum of 22 years and maximum of 54 years. From the above data, we can see that 50% of the people has age 29. Minimum no_of_Dependents is 0 and maximum is 4. Also, we can see 25% and 50% of the age group ranging between 25 to 29 years has no-of-Dependents as 2 and 3 for 75% of the people having age 38

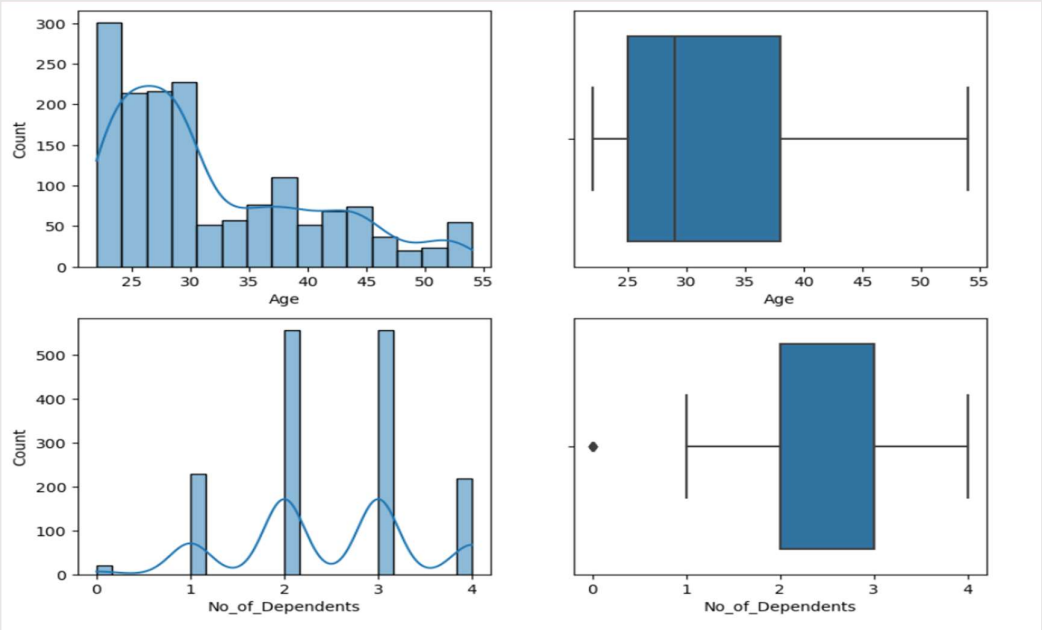

Fig 2.1

We can observe from the above 2 plots, there is no outliers for Age but we can see **outliers present in No_of_Dependents**.

2.2 Now we shall take another 2 numeric variables to draw some more insights.

Variables used : Total_salary and Partner_Salary.

	Total_salary	Partner_salary
count	1581.000000	1581.000000
mean	79625.996205	20230.655880
std	25545.857768	18909.850652
min	30000.000000	0.000000
25%	60500.000000	0.000000
50%	78000.000000	24900.000000
75%	95900.000000	38000.000000
max	171000.000000	80500.000000

Table 2.2

We can conclude that the minimum salary of Total_salary is 30,000 with maximum salary as 1.71lakhs. In Partner_salary we can see minimum salary as 0 and 25% of the partner who are working also do not contribute anything. So, the maximum salary from Partner_salary shows as 80500.

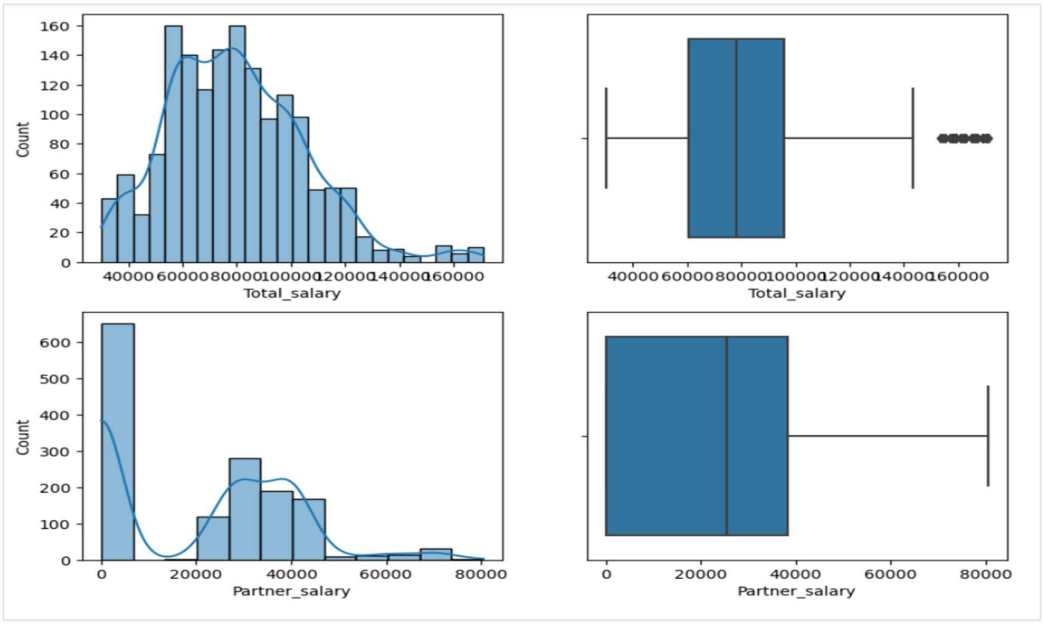

Fig 2.2

We can see more **outliers found in Total_Salary** which we will treat it later. But no outliers found in Partner_salary.

2.3 Variables used : Salary and Price.

	Salary	Price
count	1581.000000	1581.000000
mean	60392.220114	35597.722960
std	14674.825044	13633.636545
min	30000.000000	18000.000000
25%	51900.000000	25000.000000
50%	59500.000000	31000.000000
75%	71800.000000	47000.000000
max	99300.000000	70000.000000

Table 2.3

As in Table 2.3, The minimum salary is 30,000 with maximum salary as 99300. In Price we can see minimum price as 18000 and a maximum price of the car as 70000.

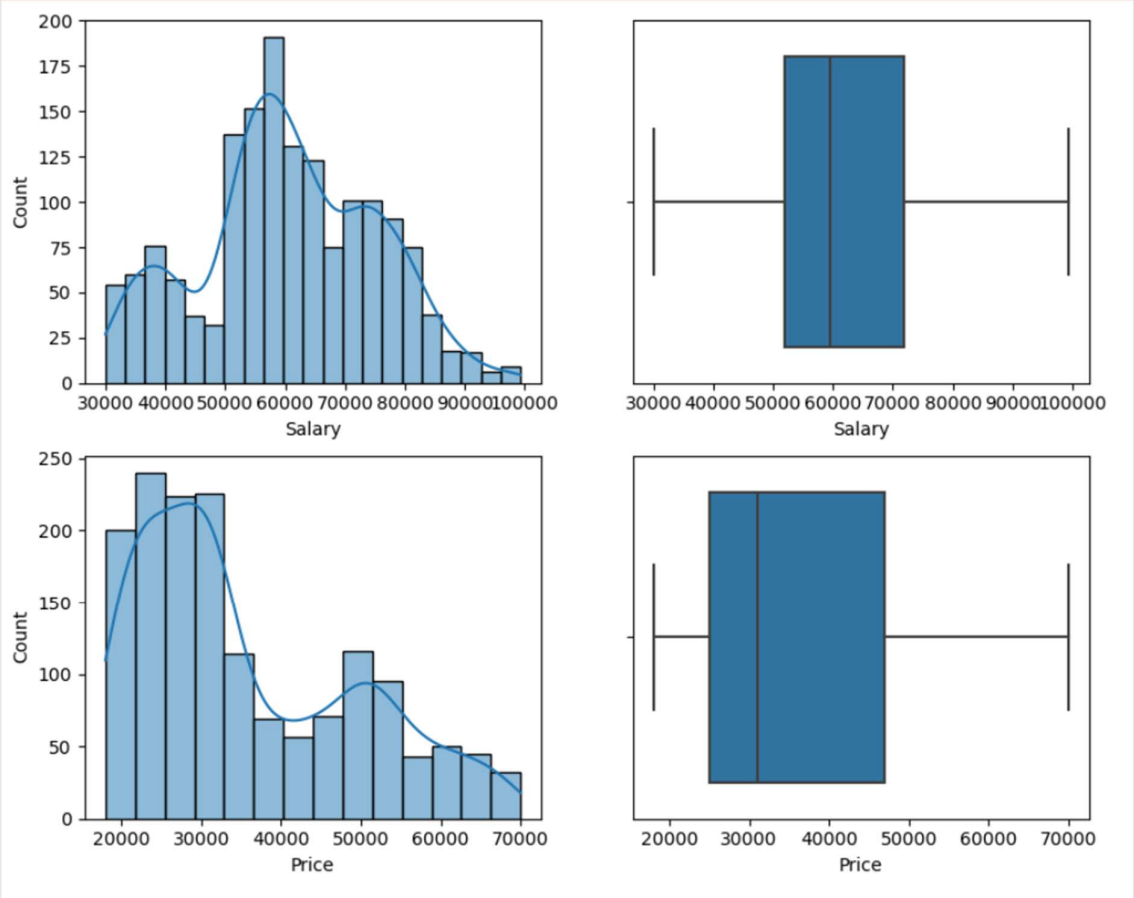

Fig 2.3

We couldn't see any outliers in both Salary and Price.

2.4 Categorical variable
Variable Used : Gender

```
Gender
Male      0.791904
Female    0.208096
Name: proportion, dtype: float64
```

Table 2.4

So, the Gender categorical variable we have displayed the results in percentage form which contributes 0.80% as Male and %0.20 as Female.

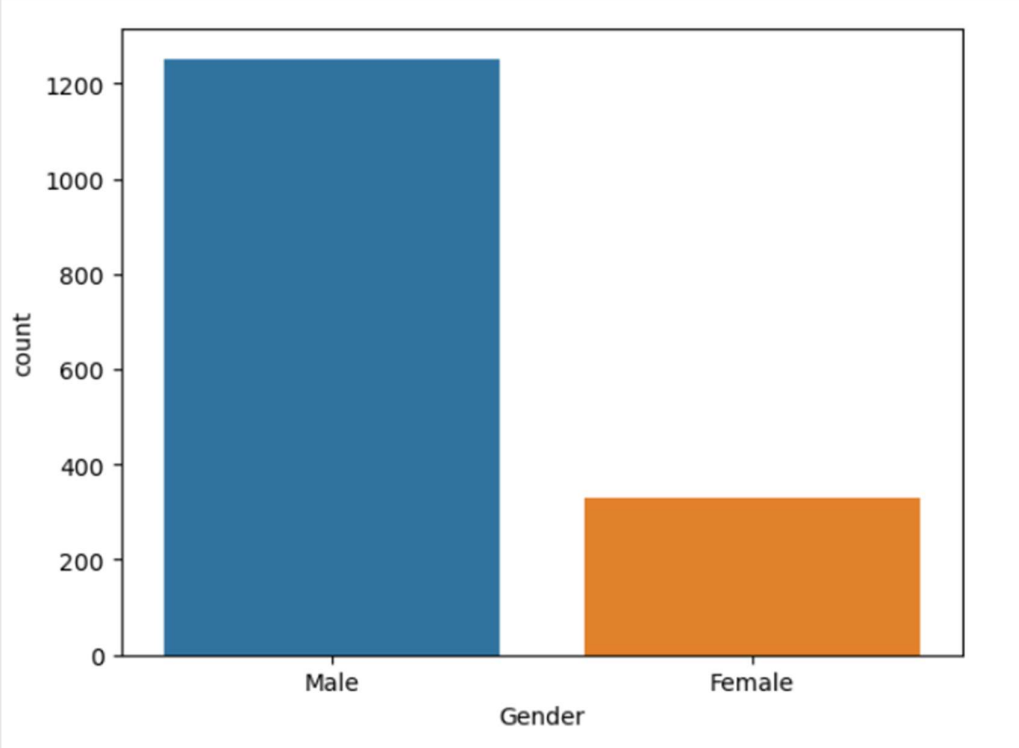

Fig 2.4

From the above plot, we can depict that count of Male gender is more when compared to Female gender.

2.5 Let’s draw some more insights with another categorical variable. Variable used : Education.

```
Education
Post Graduate    0.623023
Graduate         0.376977
Name: proportion, dtype: float64
```

Table 2.5

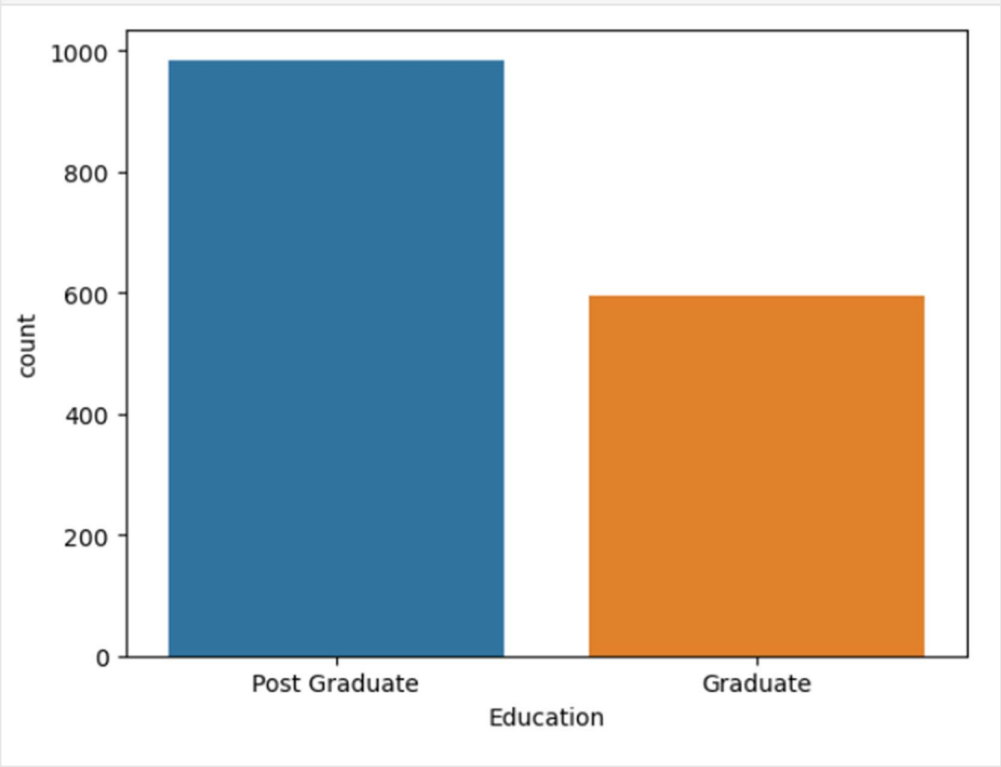

Fig 2.5

So, we can conclude that 0.63% people are Post graduates and remaining 0.37% are Graduates.

3. Let's do Bivariate Analysis

3.1 So here we take 2 numeric variables.

Variables used are : Age and No of Dependents.

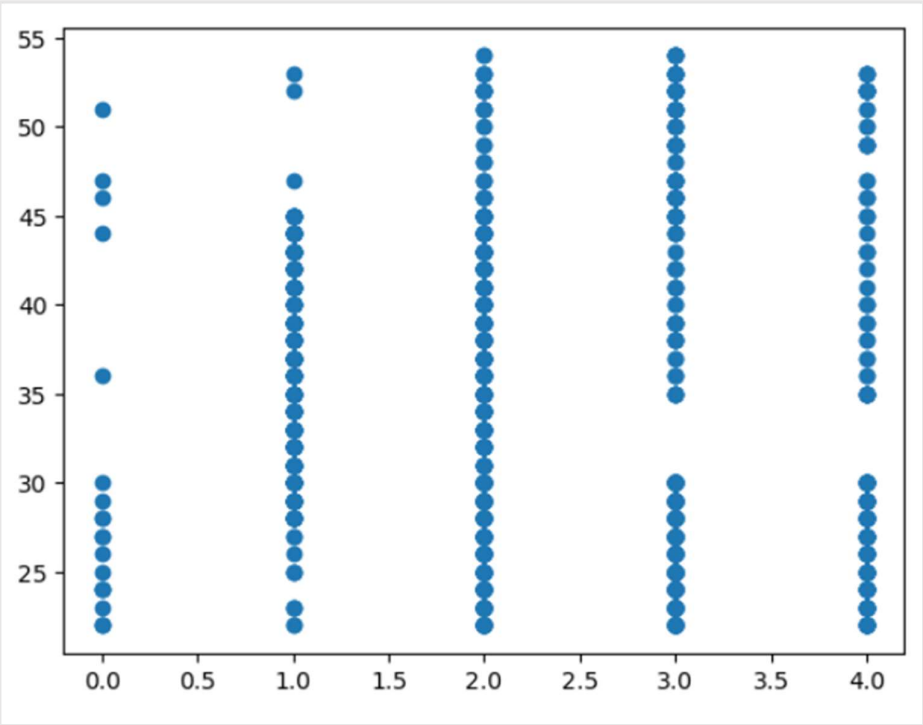

Fig 3.1

From the above plot, there seems to be a higher concentration of individuals with 2 and 3 dependents compared to other values.

3.2 So let's do some more bivariate analysis on 2 categorical variables. Here variables used are : Marital_Status and Partner_working

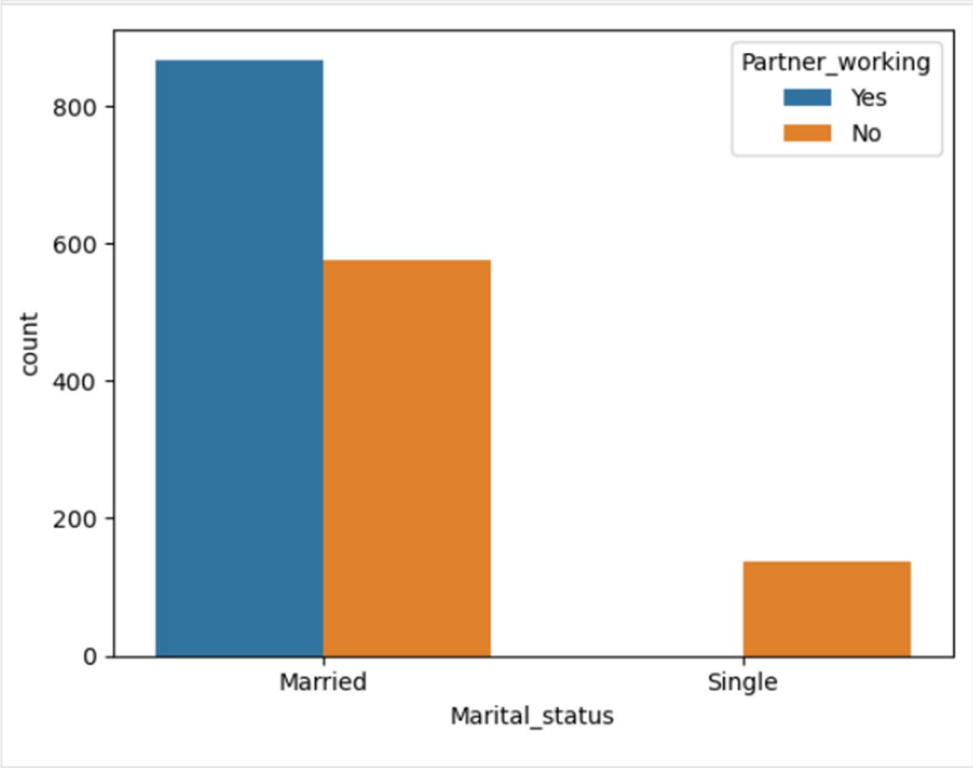

Fig 3.2

Partner_working	No	Yes	All
Marital_status			
Married	0.363694	0.54902	0.912713
Single	0.087287	0.00000	0.087287
All	0.450980	0.54902	1.000000

Table 3.2

So from both the data and plot above, we can assume that 0.087% are single and 0.91% constitutes to married people which includes partner_working.

3.3 Now let's see bivariate analysis on categorical variable with numeric variable.

Variables used : Gender , Personal_Loan.

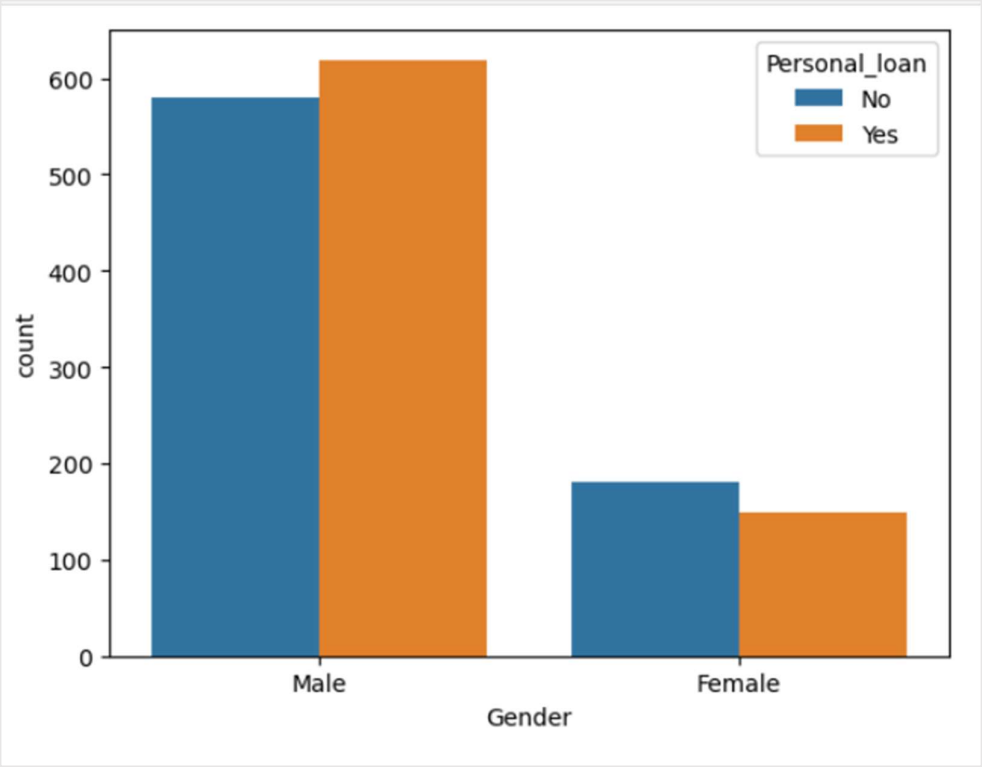

Fig 3.3

So when we see the above figure, the personal loan taken by Male is more when compared to Female.

Personal_loan	No	Yes	All
Gender			
Female	0.113852	0.094244	0.208096
Male	0.385199	0.406705	0.791904
All	0.499051	0.500949	1.000000

Table 3.3

From the table it states that only 0.09% of Females had taken Personal Loan, where 0.40% of men have personal loans.

3.4 Let’s analyze with some more variables.
Variables used : Salary , Gender

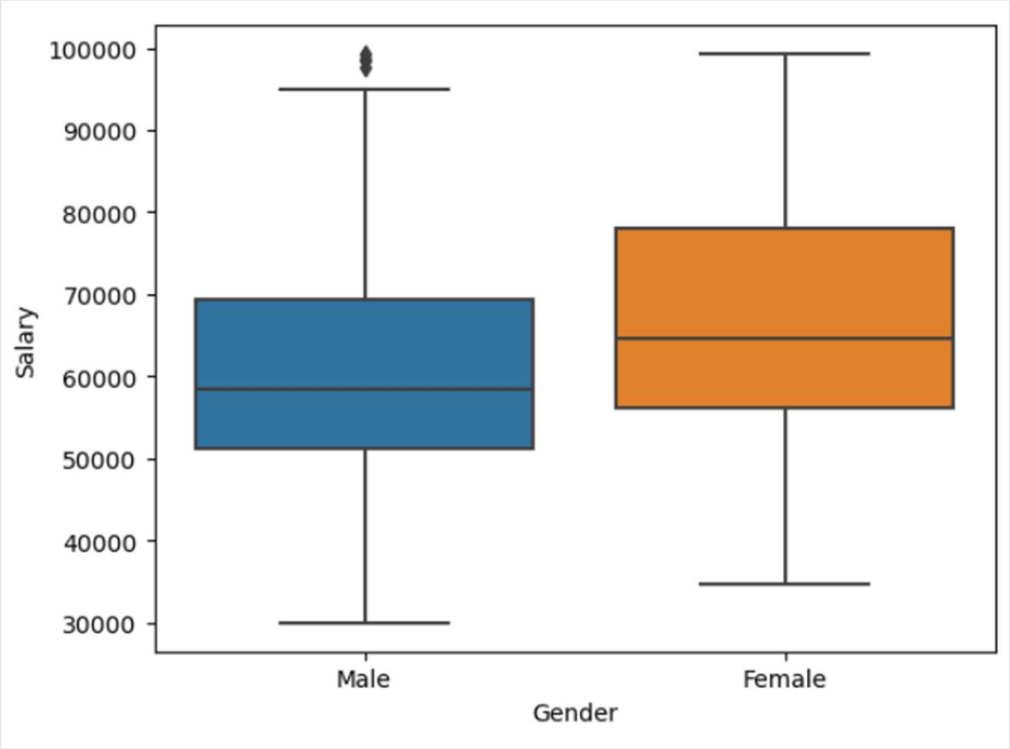

Fig 3.4

We can observe here that the median salary for females is higher than that of males.

3.5 Variables used : Salary , Make

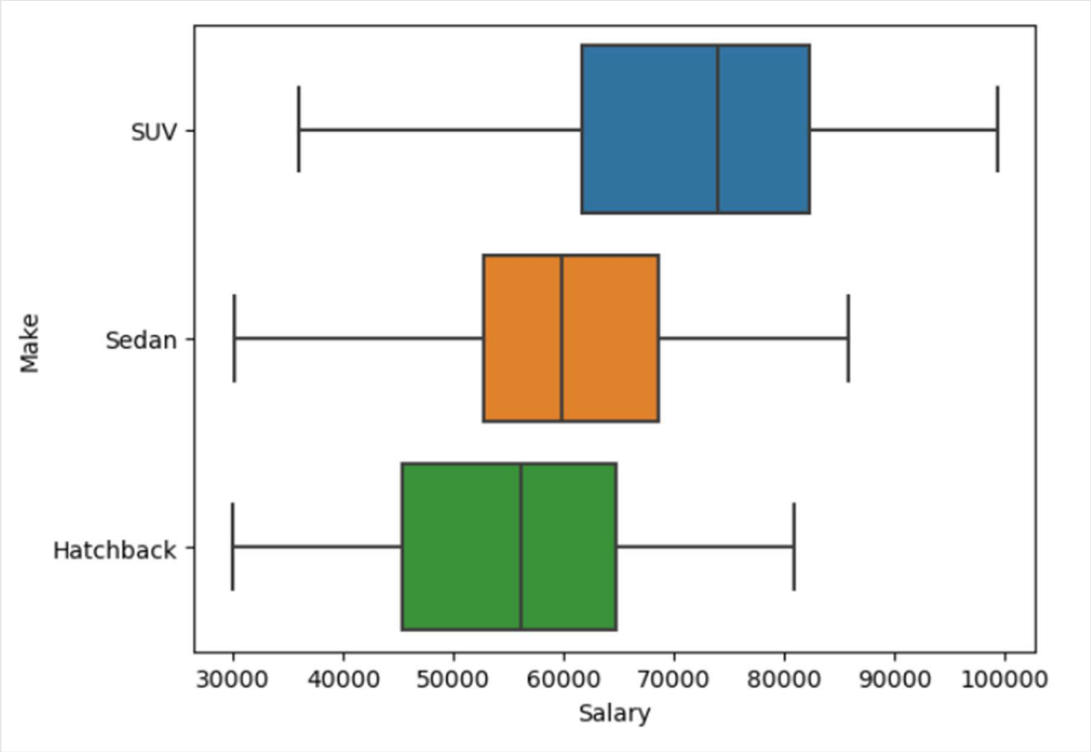

Fig 3.5

SUVs tend to have the highest median salary, followed by Sedans and Hatchbacks. A larger variation in earnings for people owning SUVs compared to Sedans or Hatchbacks.

4. Checking for Outliers

We shall make use of boxplot to check for outliers.

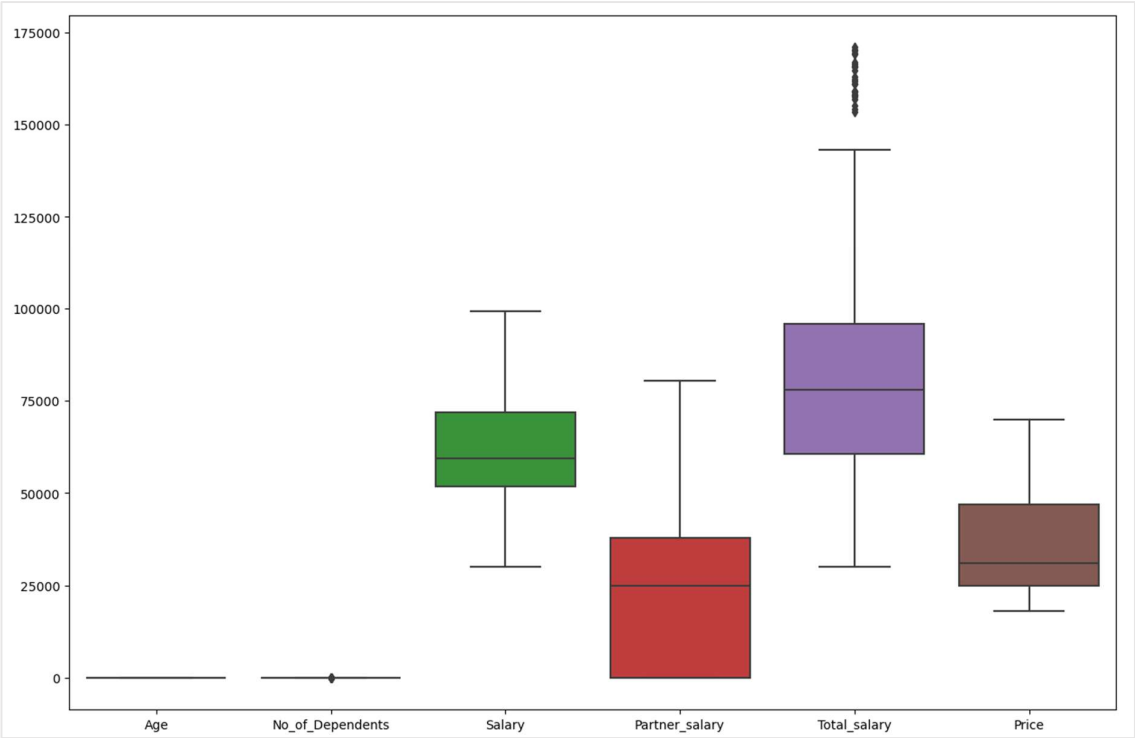

Fig 4.1

We can see lot of extreme outliers present in **Total_salary** as it is highly skewed and only 1 outlier in **no-of_dependents** but no outliers present in other variables.

Removing the outliers :

As we had seen outliers in **Total_salary** and **No_of_dependents.** Let's remove it by replacing the outlier value using IQR.

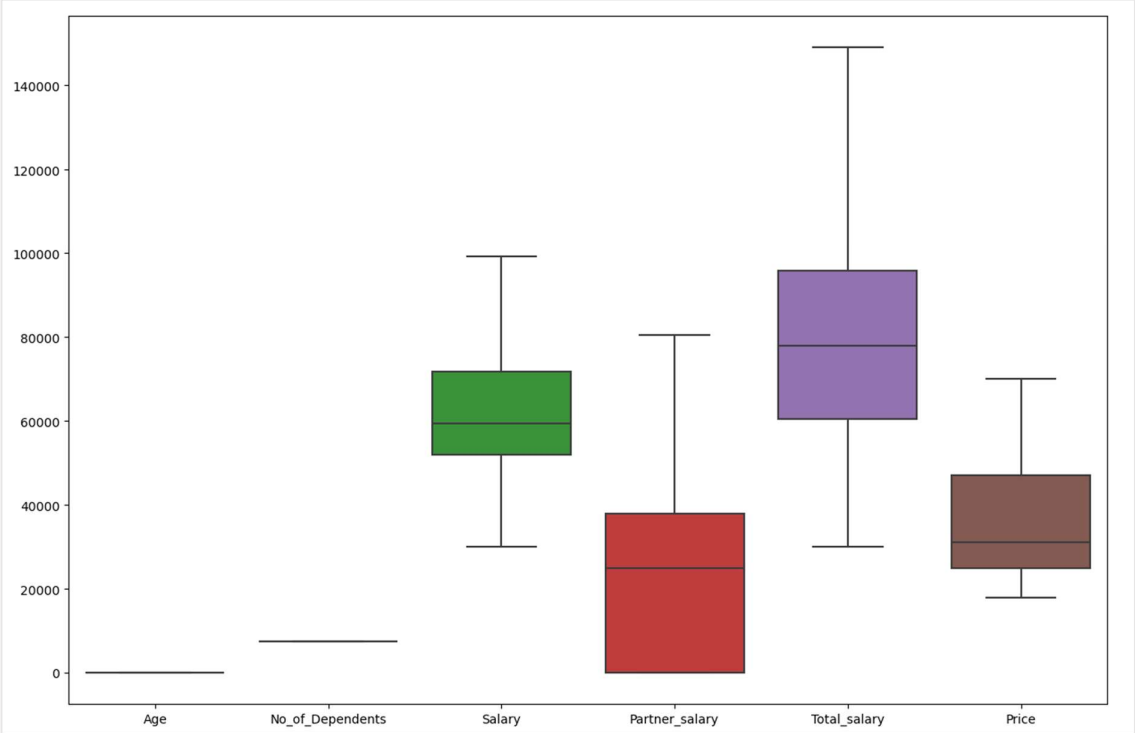

Fig 4.2

So once outliers removed then we can see the following attributes does not contain outliers now.

5. Answers to Key Questions

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

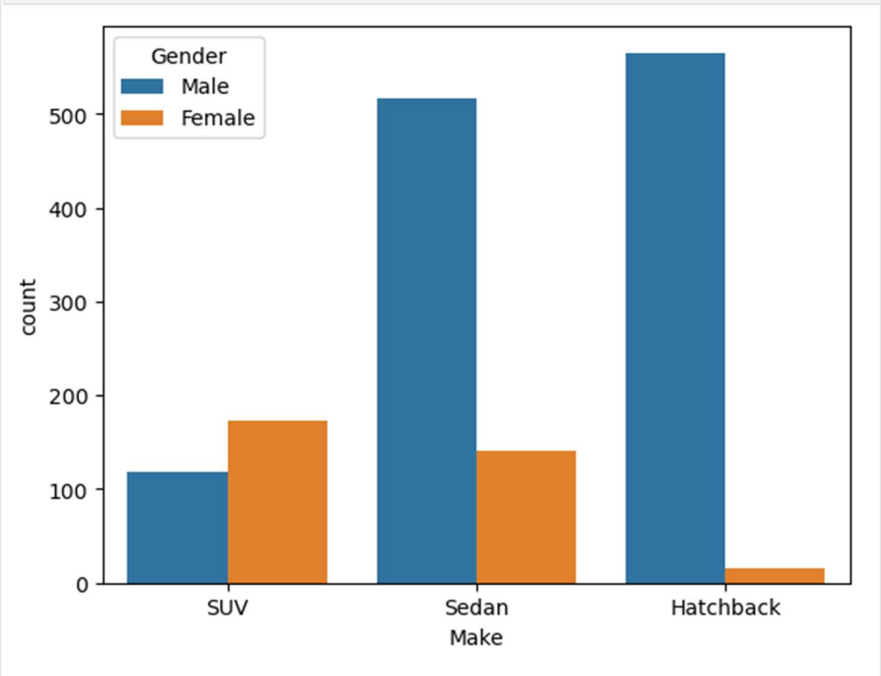

Fig 5.1

Gender	Make	
Female	SUV	173
	Sedan	141
	Hatchback	15
Male	Hatchback	565
	Sedan	516
	SUV	118

Table 5.1

From the above analysis, we can see count of Female gender who prefer SUV is 173 but Male Gender who prefers SUV count is 118 hence we can say that Female prefer SUV by a medium margin when compared to Men. **So "Female prefer SUV by a large margin, compared to the Male" that is Men doesn't trend to prefer SUV's while comparing to Women.**

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

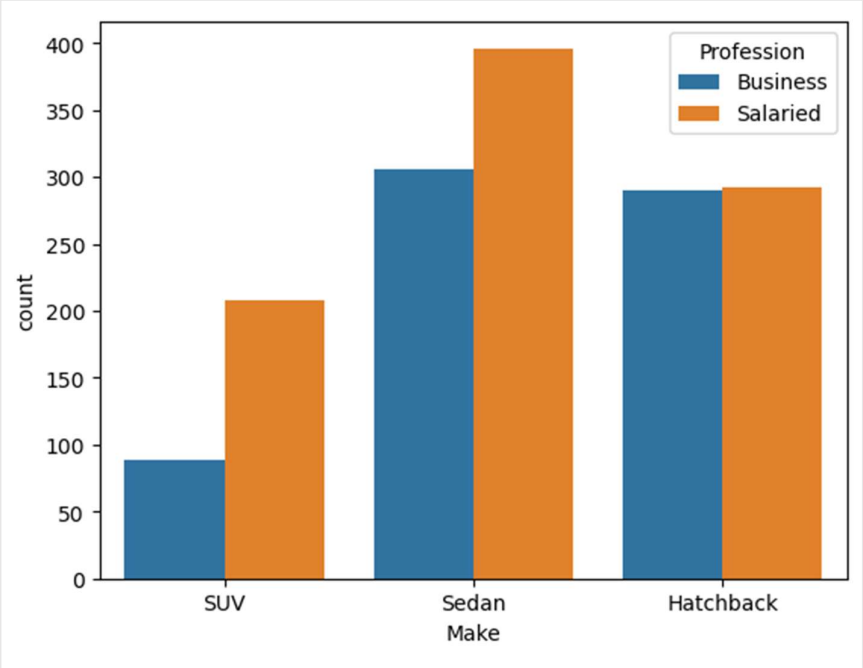

Fig 5.2

Profession	Make	
Business	Sedan	306
	Hatchback	290
	SUV	89
Salaried	Sedan	396
	Hatchback	292
	SUV	208

Table 5.2

We can see the count of Salaried people buying Sedan is 396 whereas the Business profession people buying Sedan count is 306 which highlights that "**Salaried person is more likely to buy a Sedan** to buy a Sedan than Business person".

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?

Profession	Make	Gender	
Business	Hatchback	Male	289
	SUV	Female	55
		Male	33
	Sedan	Female	50
		Male	237
Salaried	Hatchback	Female	15
		Male	276
	SUV	Female	118
		Male	85
	Sedan	Female	91
		Male	279

Table 5.3

As per the plot, we can see that the count of salaried Male people buys Sedan more than SUV when compared to Female salaried people. Hence the Sheldon Cooper statement is **False stating that a salaried male is an easier target for a SUV sale over a Sedan sale.**

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

Gender	Make	
Female	SUV	173
	Sedan	141
	Hatchback	15
Male	Hatchback	565
	Sedan	516
	SUV	118

Name: count, dtype: int64

Table 5.4.1

Gender	
Female	329
Male	1199

Name: Make, dtype: int64

Table 5.4.2

Total number of cars brought by Female is 329 and Male is 1199.

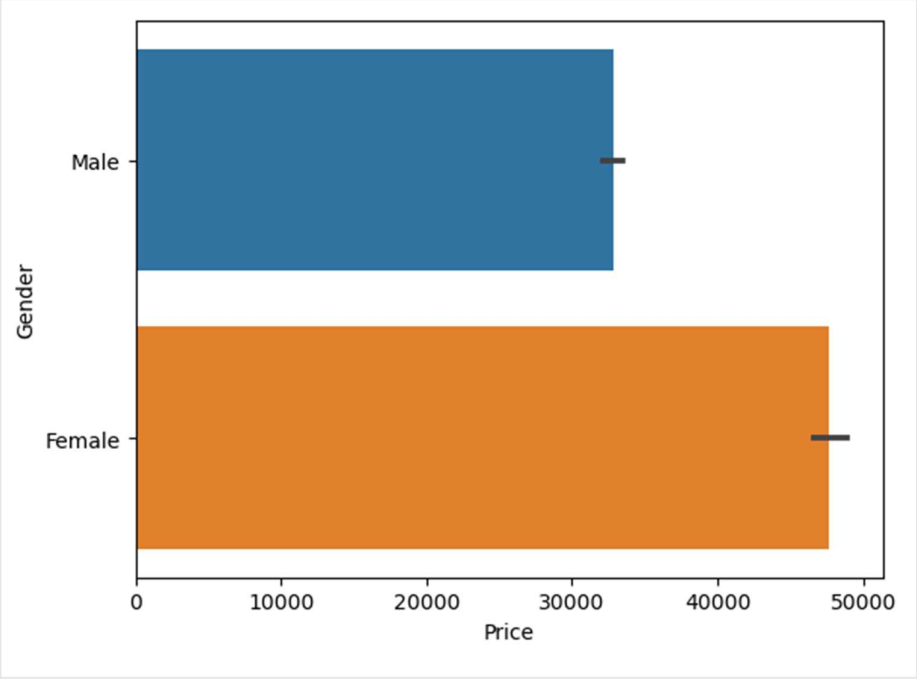

Fig 5.4.1

If we see the plot above even though the cars brought by Men are more, we can see that the Females are spending more than Male.

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

```
Gender  Personal_loan
Female  No              180
        Yes             149
Male    Yes             619
        No              580
Name: count, dtype: int64
```

Table 5.5.1

In total 619 Men took personal loan where only 149 Women took loan. Which convey that men took more loan compared to women. And in total 758 people purchases Vehicle using personal loan.

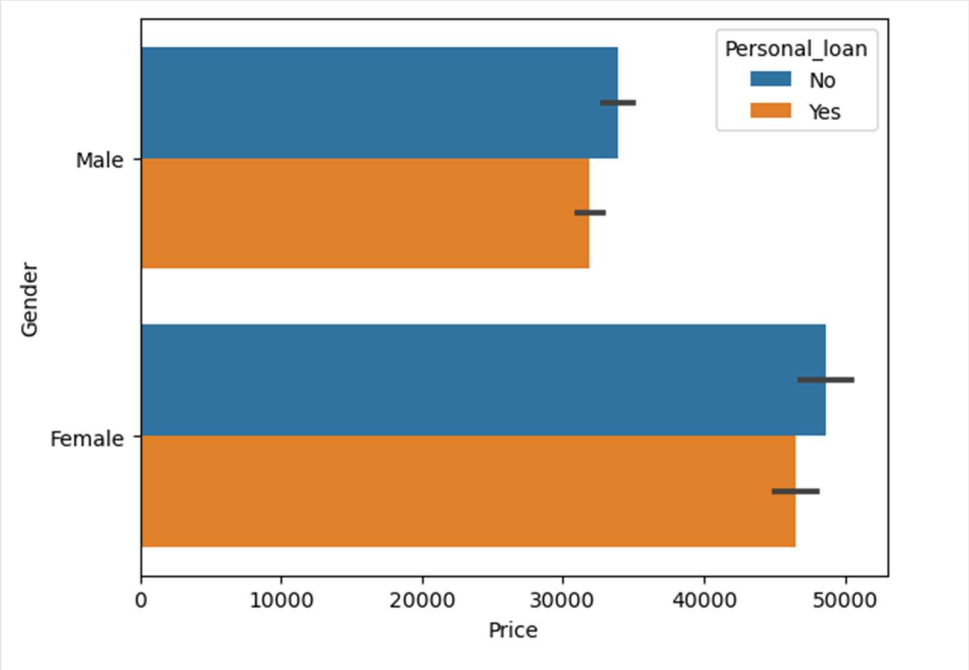

Fig 5.5.1

With regards to personal_loan analysis, we can conclude that Males are taking more loan than females but we can see that Without having any personal_loan Female are spending more on purchasing automobiles when compared to Male and also Female spending money by personal_loan is more than that of Male.

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

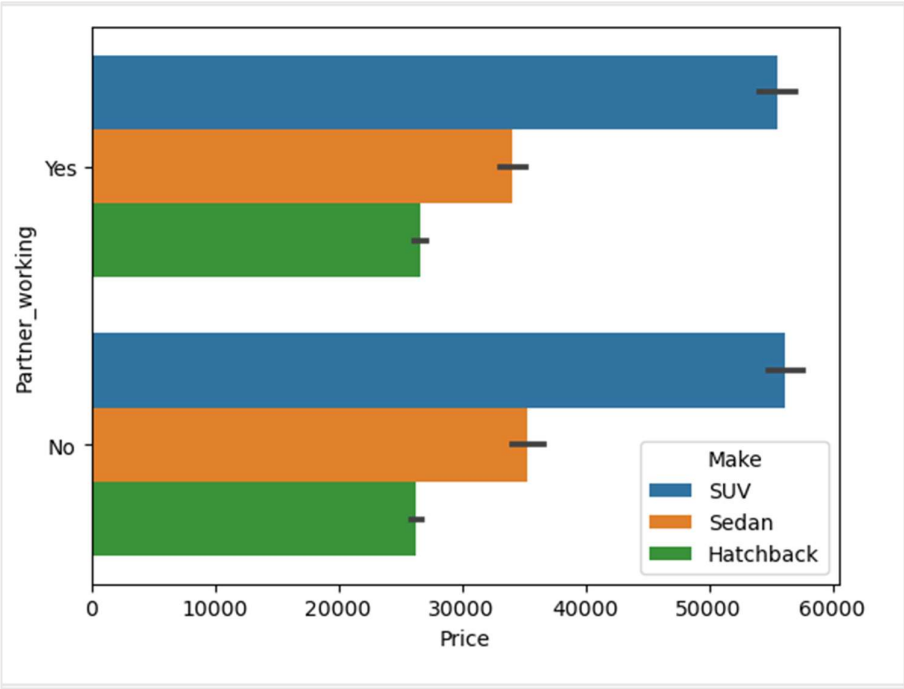

Fig 5.6

We can see that partner who are not working seems to be slightly more edged than partner who are working with regards to purchasing in high priced car. But partner who are working seems to be very close near to that of partner who are not working. **So, conclusion is, if having a working partner leads to the purchase of a higher-priced car – it's not agreed, it's no and statement is proved to be wrong with the analysis.**

6. Summary & recommendations:

1. Sedan contributes 44% of the entire sales and 43% of the total revenue.
2. 20-30 age group customers are purchasing 60% cars.
3. In terms of marital status as married 34% Males are purchasing Sedans, 31% males are purchasing Hatchbacks. 10% females are purchasing SUVs and also females are mostly interested in purchasing SUVs.
4. Customers who have taken personal loans are going after Sedans and in terms of money and quantity sedan is the highest. It means that most of the customers are purchasing sedans on personal loans.
5. Working partners purchase more cars because they contribute more in terms of revenue, also they are purchasing more sedans. But for SUVs partner working doesn't show much impact.
6. Profession weather it is business or salaried doesn't show much impact on the purchase pattern.
7. 57% employee total salary is between the 50000-100000 and 16% employee total salary falls between 70000-80000.
8. Most of the male are taking Personal Loans for the purchase of a car.
9. Also, most of the working Female are preferring to buy SUV model than other 2 models. Hence, We should check the other 2 models features compared to SUV model and price analysis so that the other 2 models can be bought into competition and the Female can purchase it based on the customization of the features , price , etc which is found in SUV model.