Business Report SMDM

Problem 1

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? (Hint: Information about the size of the dataset and the nature of the variables)

From a given dataset Austo Motor Company, we import numpy, pandas, matplotlib, seaborn in jupyter notebook.

	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

	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

We can see from the above table the first 5 and last 5 records in the dataset.

There are total 1581 rows and 14 columns are present in the dataset.

Below table shows the basic statistical information of the numerical variables. Do not consider ‘Age’ as numerical variable.

	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	1475.0	20225.559322	19573.149277	0.0	0.0	25600.0	38300.0	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

Below table shows the business information of the dataset.

```
<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
```

A quick look at all the dataset information tells us that there are 6 are numerical and 8 categorical variables. There are few null records present in two variables (Gender and Partner_salary). Which will be analysed in detail in the next section.

There are no duplicates that we have found in the dataset.

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.**

Checking the null values in variables.
"Gender" has 53 Null values present
"Partner_salary" has 106 Null values present

Missing values can be imputed. But if there is larger amount of values are missing then we drop the column.

```
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
```

	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	1475.0	20225.559322	19573.149277	0.0	0.0	25600.0	38300.0	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

From the dataset we described mean, median, std, max, min value here we get the whole summary of the data

- The Gender age between 22 to 54 are belong to working people with median age is 29

- The overall data of salary given people ranges from 30000 to 99300
- The total salary is ranges from 30000 to 171000
- The minimum purchase of the car is 18000, where maximum car purchased 70000

Now we check each categorical value in the data, help us to check the issues.

"Gender"

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

"Profession"

```
Salaried    896
Business    685
Name: Profession, dtype: int64
```

"Marital_status"

```
Married    1443
Single      138
Name: Marital_status, dtype: int64
```

"Education"

```
Post Graduate    985
Graduate         596
Name: Education, dtype: int64
```

"Personal_loan"

```
Yes    792
No     789
Name: Personal_loan, dtype: int64
```

"House_loan"

```
No     1054
Yes     527
Name: House_loan, dtype: int64
```

"Partner_working"

```
Yes    868
No     713
Name: Partner_working, dtype: int64
```

"Make"

```
Sedan        702
Hatchback    582
SUV          297
Name: Make, dtype: int64
```

From the value counts we can see the in “Gender” variable “Female” misspelled to “Feml” and “Femal” but the rest of the variables are free from the issues.

We can impute the data from “Feml” and “Femal” to “Female” in the data.

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

After imputing “Female” Count will be changed to 329.

Still there are some missing values in the “Gender” so we take “Male” as a maximum frequency and put it into the missing values place. After that we get.

```
Male      1252
Female     329
Name: Gender, dtype: int64
```

Now the “Gender” variable is perfect and we have our “Male” is 1252 and “Female” 329.

```
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      106
Total_salary          0
Price                 0
Make                  0
dtype: int64
```

As we see from the above table still there are missing values in “Partner_salary”.

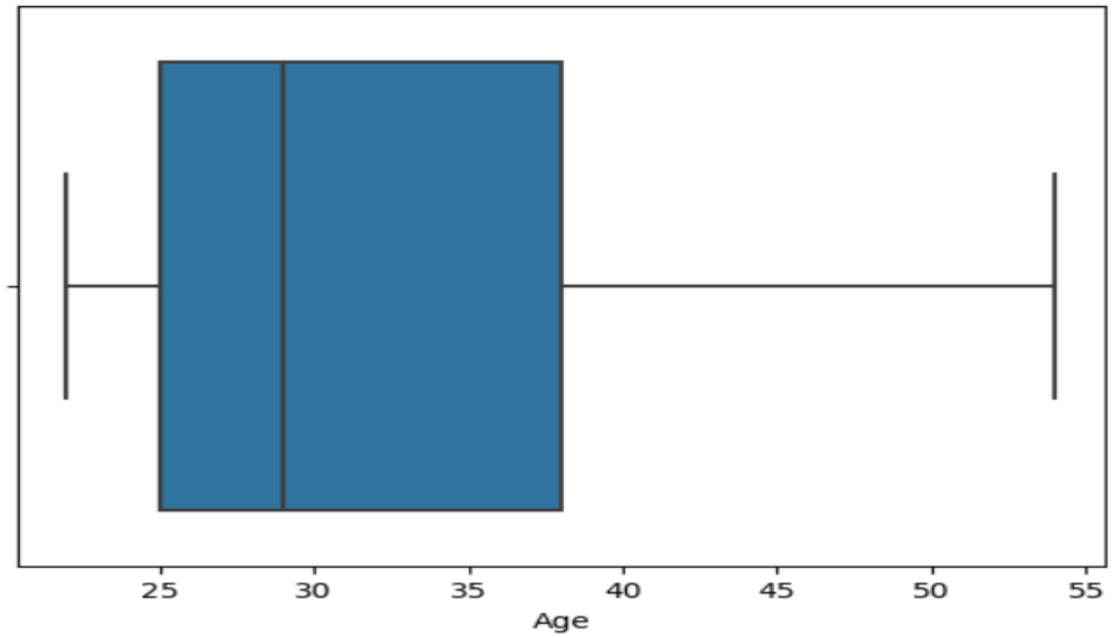
As we know, Total_salary = Salary + Partner_salary

By using above formula we fill the all null values in “Partner_salary” and as you see there no any-other null values present in the dataset.

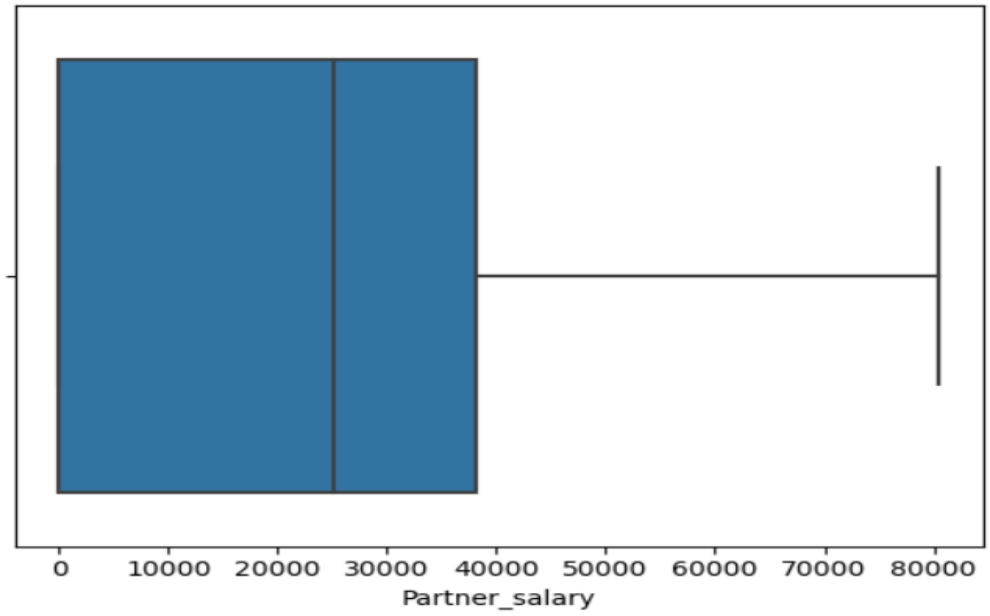
```
Age                0
Gender             0
Profession         0
Marital_status     0
Education          0
No_of_Dependents   0
Personal_loan      0
House_loan         0
Partner_working    0
Salary             0
Partner_salary     0
Total_salary       0
Price              0
Make               0
dtype: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

Now we will see the data, if there is any outliers present or not by using boxplot.

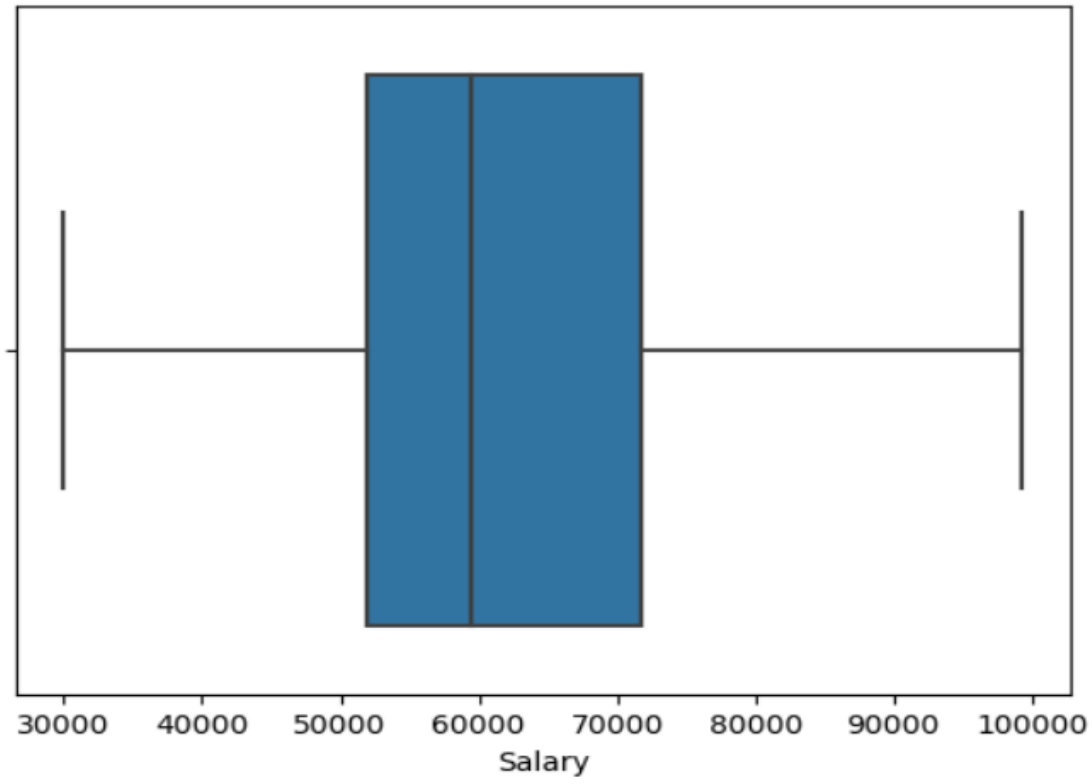
```
<Axes: xlabel='Age'>
```



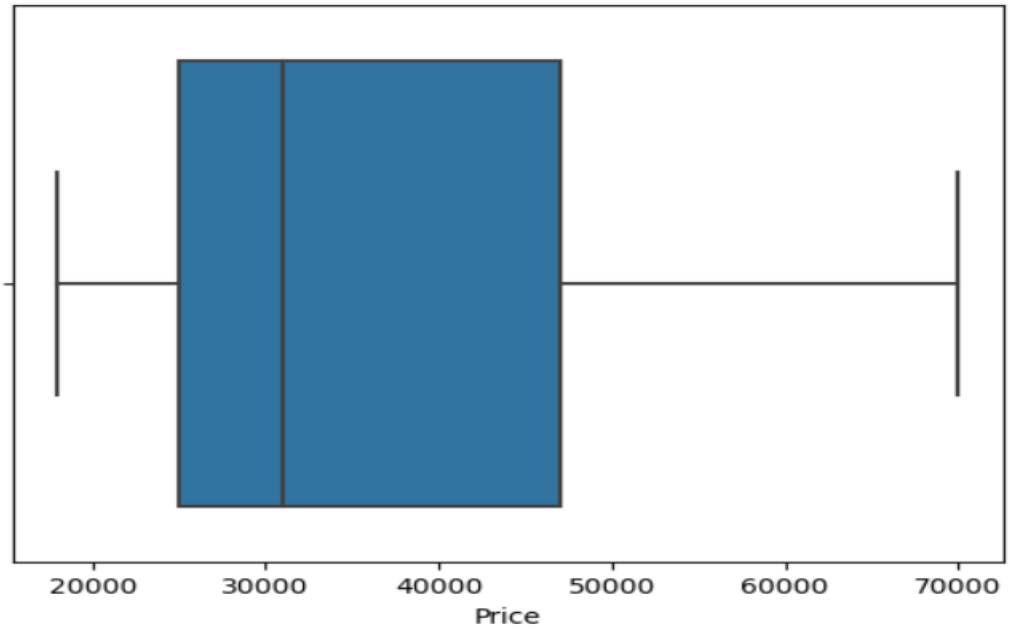
```
<Axes: xlabel='Partner_salary'>
```



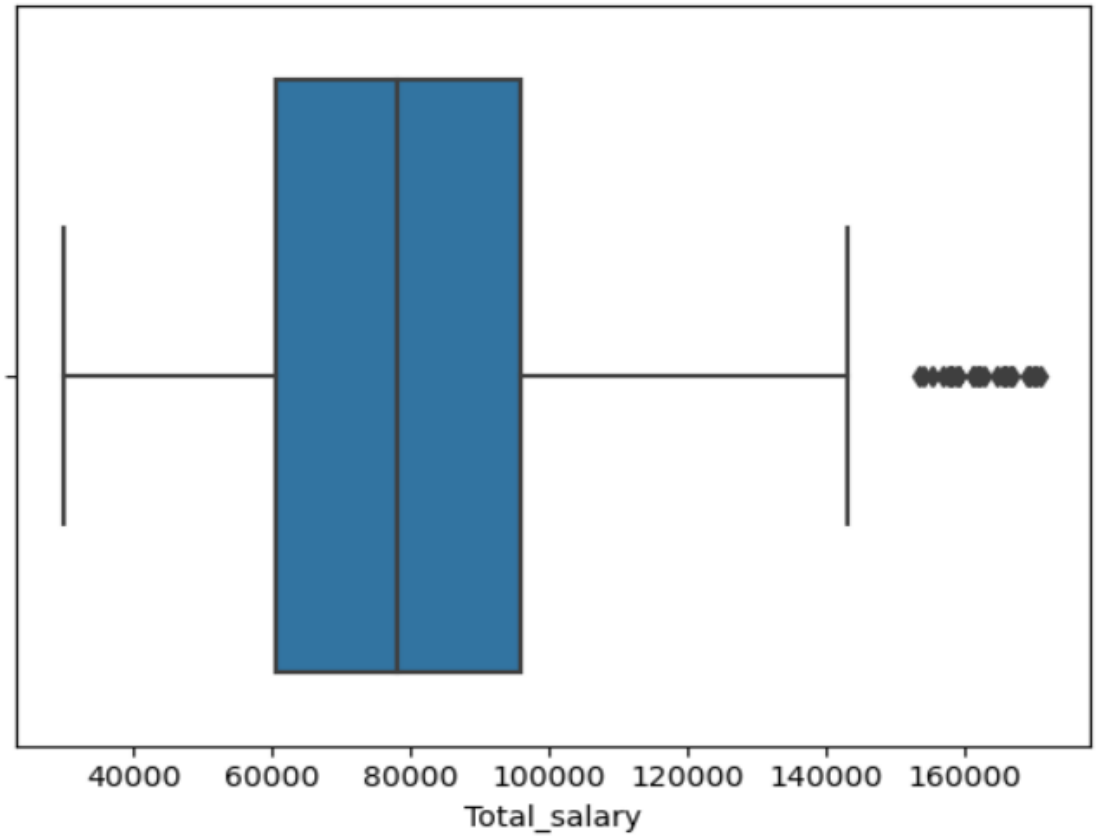

```
<Axes: xlabel='Salary'>
```



```
<Axes: xlabel='Price'>
```



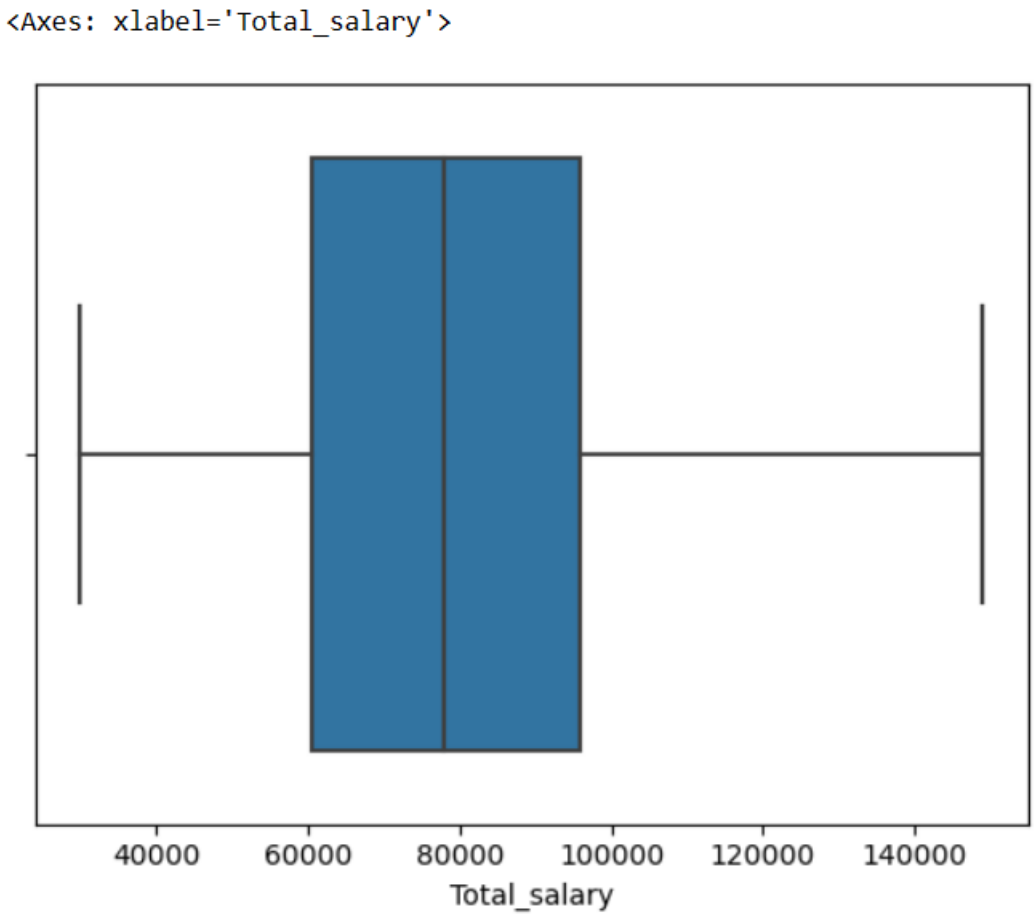

```
<Axes: xlabel='Total_salary'>
```



We don’t see any outliers in Age, Partner_salary, Salary and Price.

But we can see there are outliers in “Total_salary”

The Boxplot for Total_salary contains outliers, we need to treat according to IQR Equation to get more insights in the analysis.



We have treated the outliers by using IQR rule.

(Q1-1.5*IQR) is used to treat the lower value.

(Q3+1.5*IQR) is used to treat the higher value.

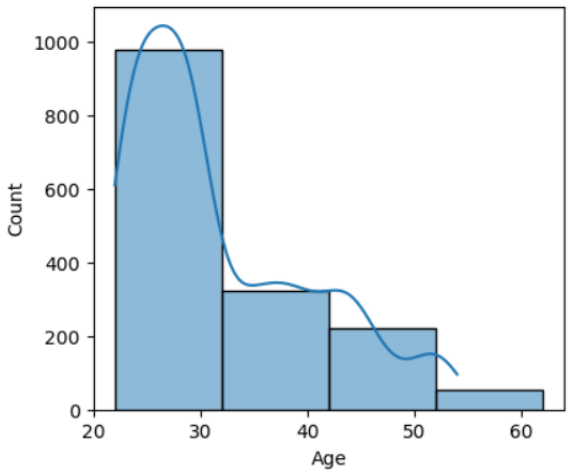
Using the formula, The has been imputed.

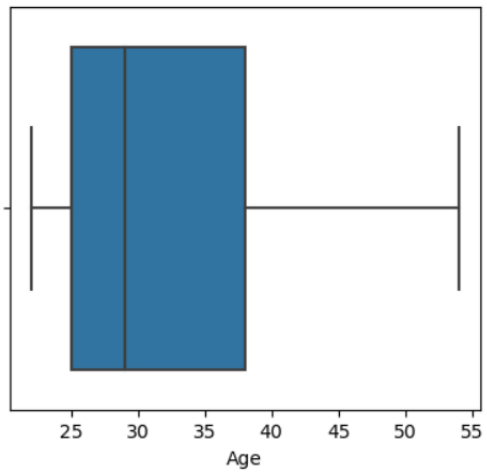
C. **Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.**

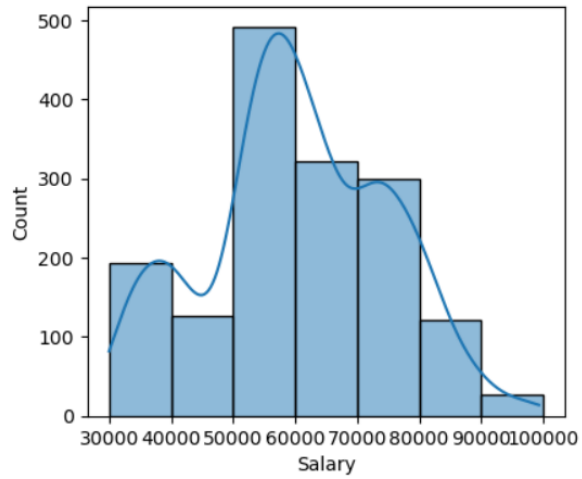
We have separated categorical value and numerical value to see the insights of the business.

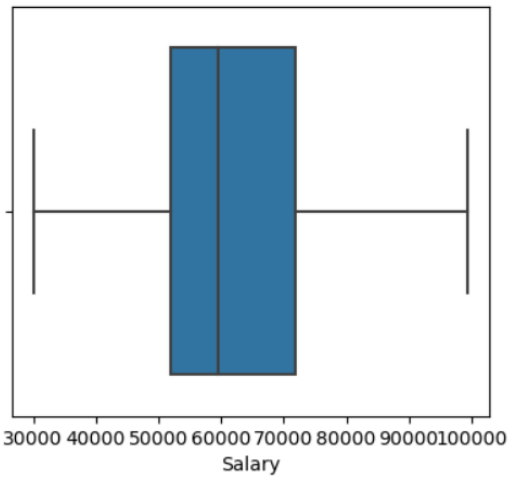
For better visualisation we are going to use Histogram and Boxplot for better understanding .

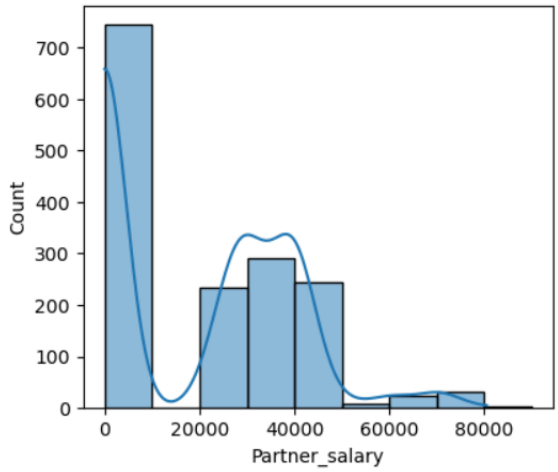
The Graph shows the output of the preloaded data.

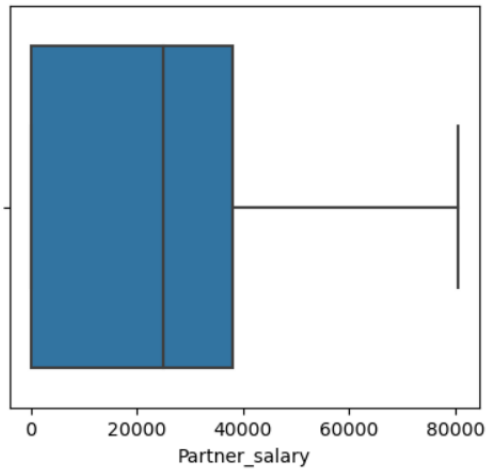


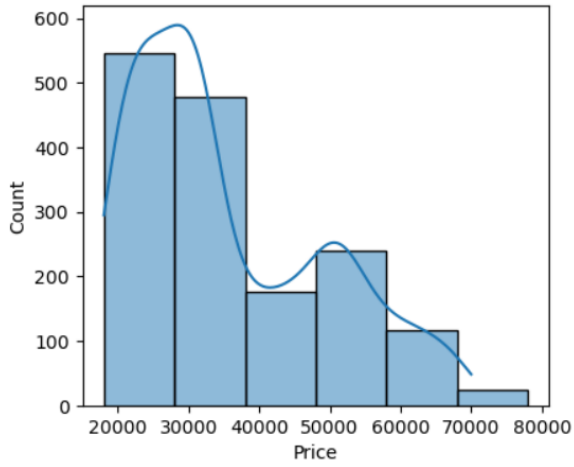


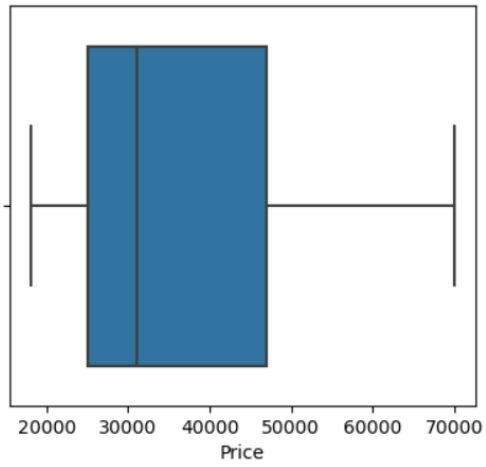


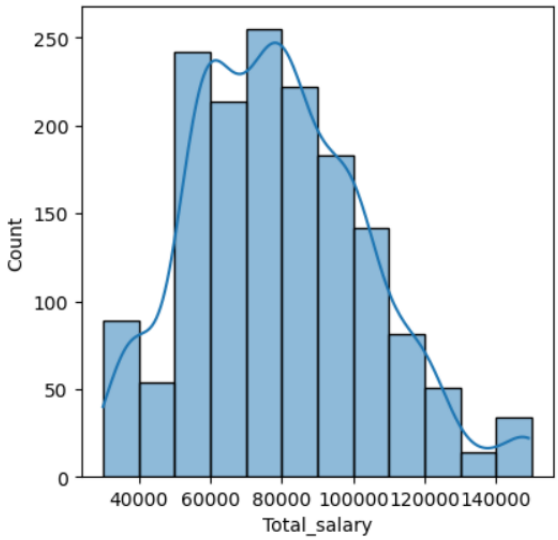


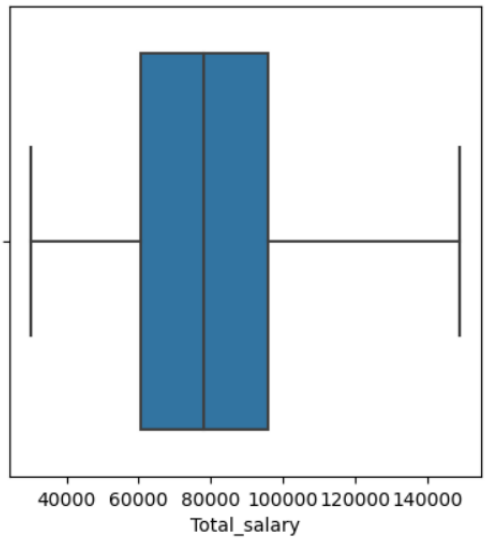












Age is multi distribution from the 50000 to 80000

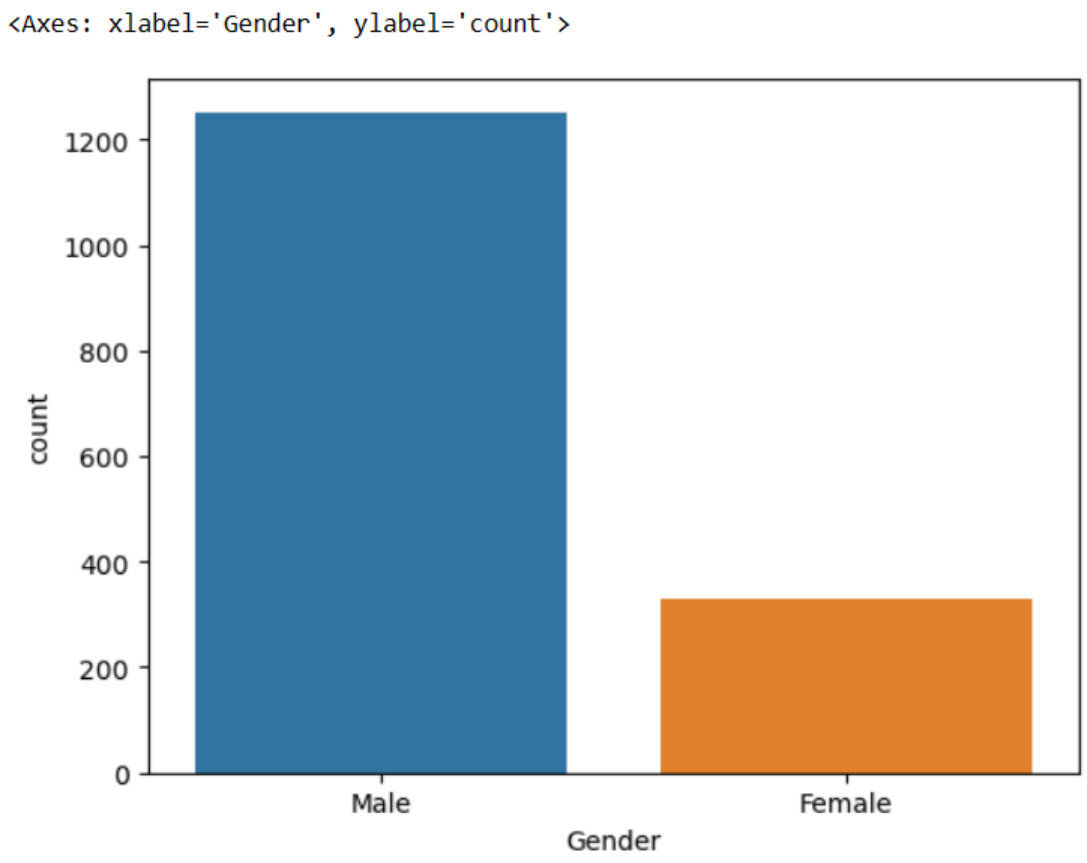
Salary has bulk distribution from the 50000 to 70000

Total_salary contains outliers but after treating we can see a perfect Histogram and Boxplot for Total_salary.

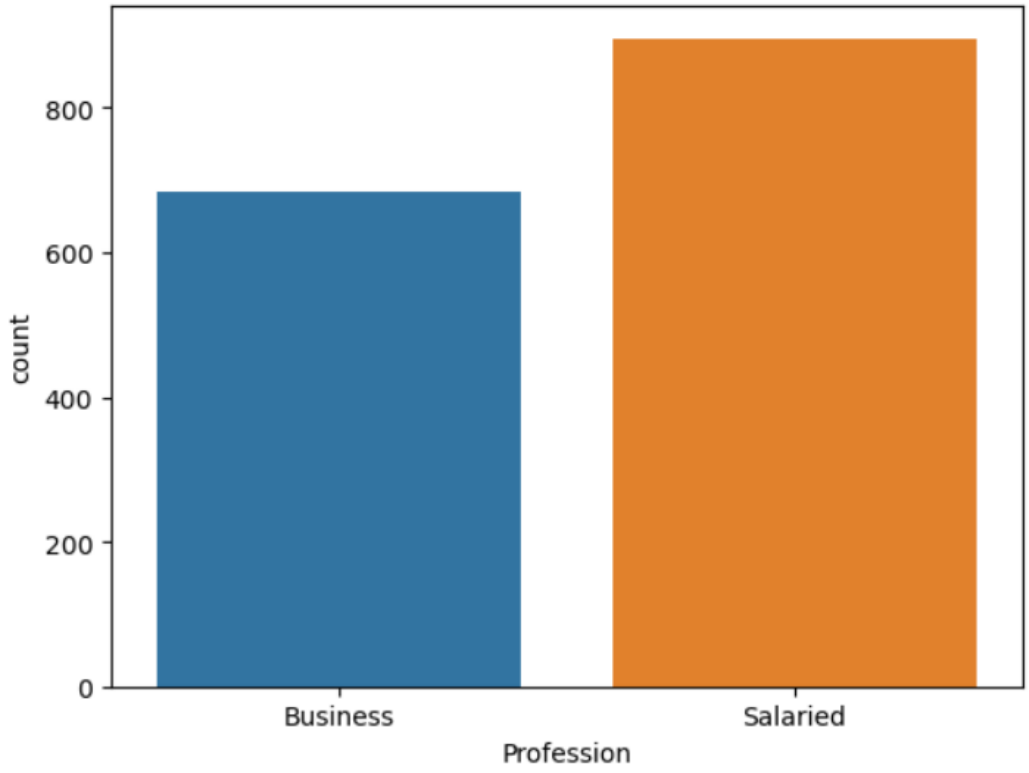
Price seems to be positive skewed of 0.74.

From the above items there is not any normal distribution in this data.

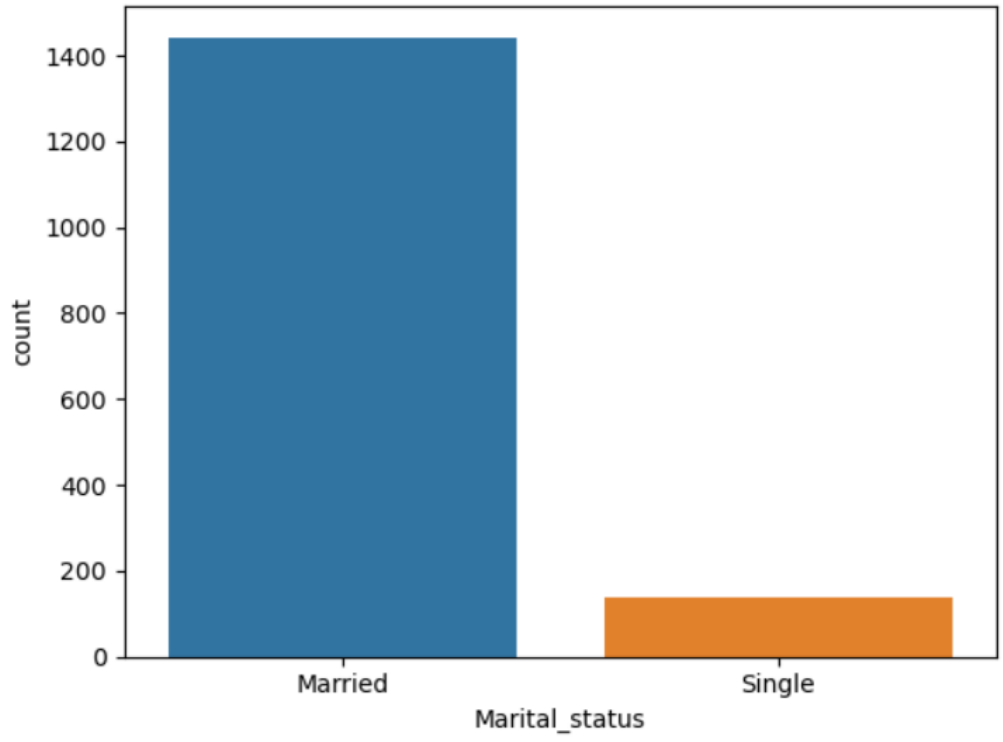
Univariate Analysis for Categorical variable.



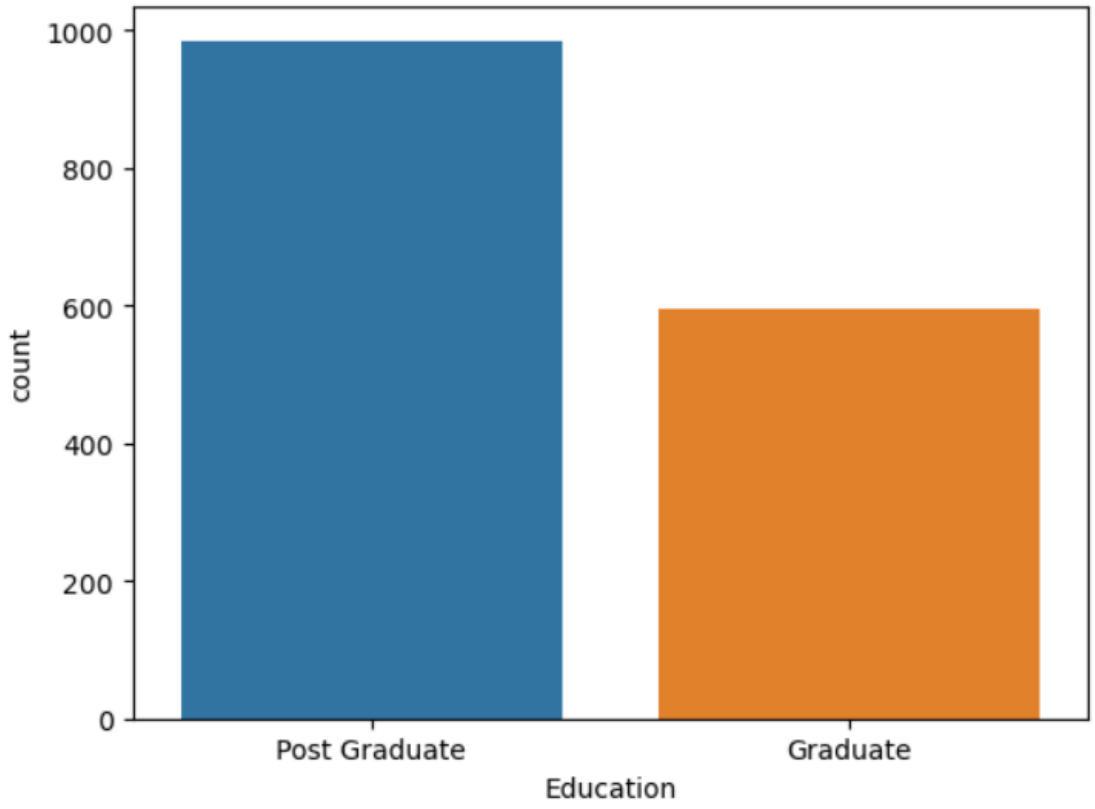
```
<Axes: xlabel='Profession', ylabel='count'>
```



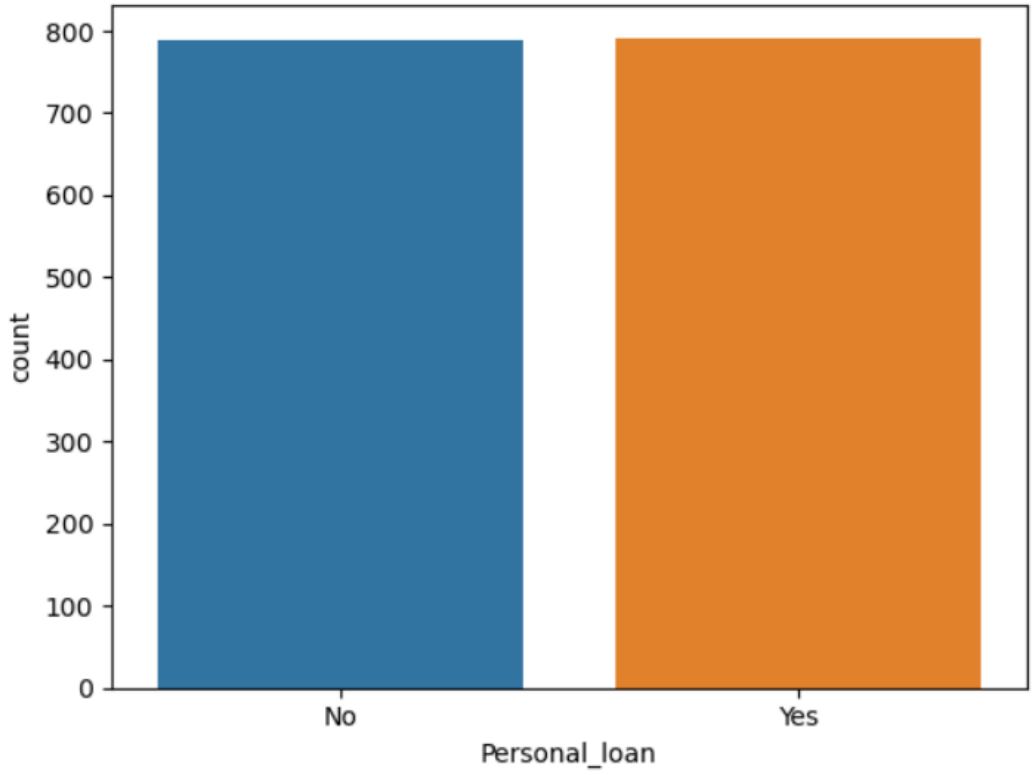

```
<Axes: xlabel='Marital_status', ylabel='count'>
```



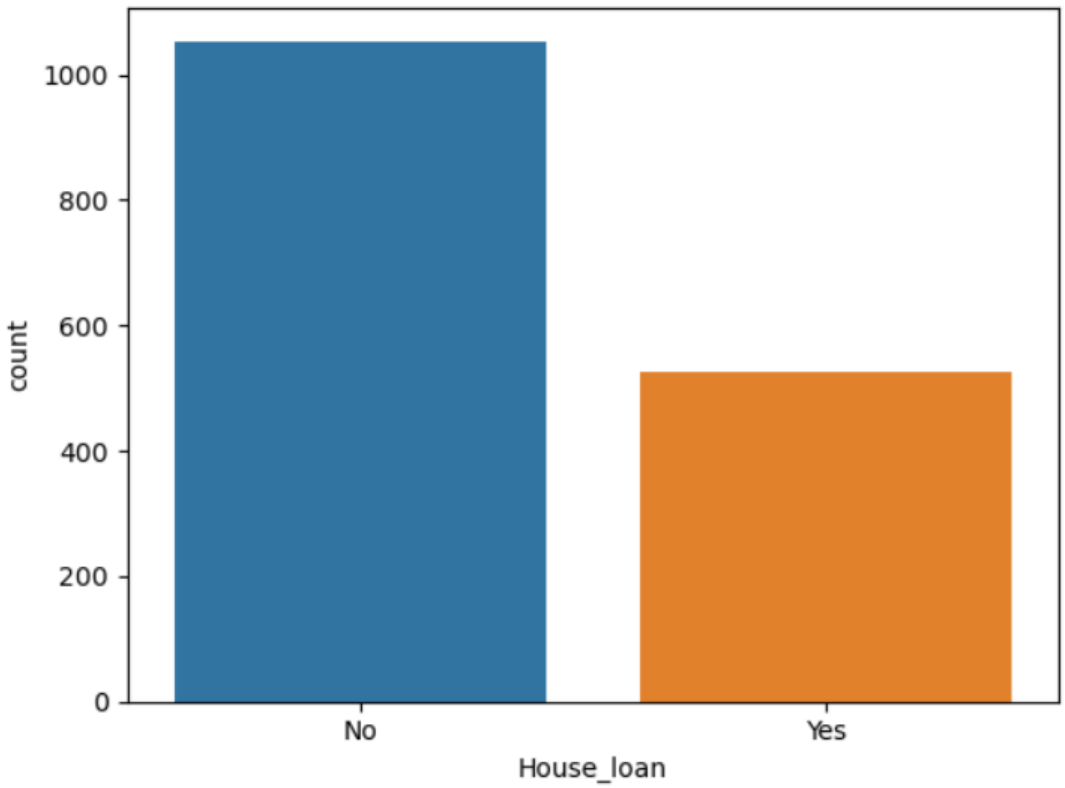
```
<Axes: xlabel='Education', ylabel='count'>
```



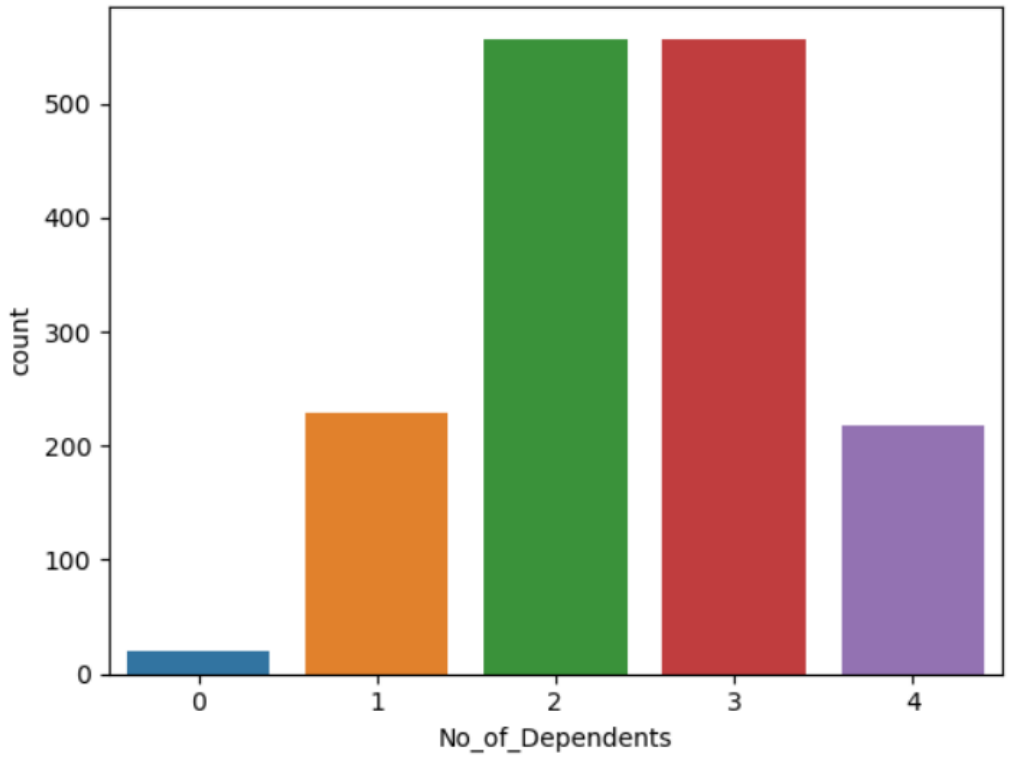

```
<Axes: xlabel='Personal_loan', ylabel='count'>
```



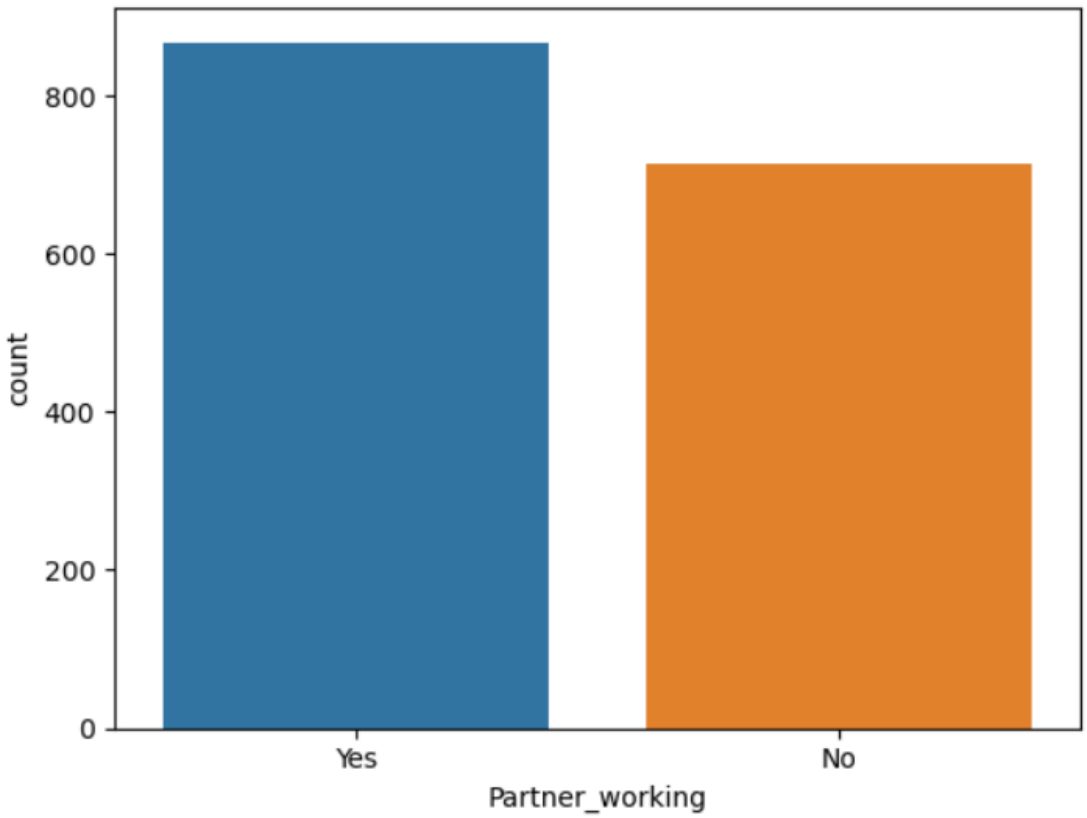
```
<Axes: xlabel='House_loan', ylabel='count'>
```




```
<Axes: xlabel='No_of_Dependents', ylabel='count'>
```



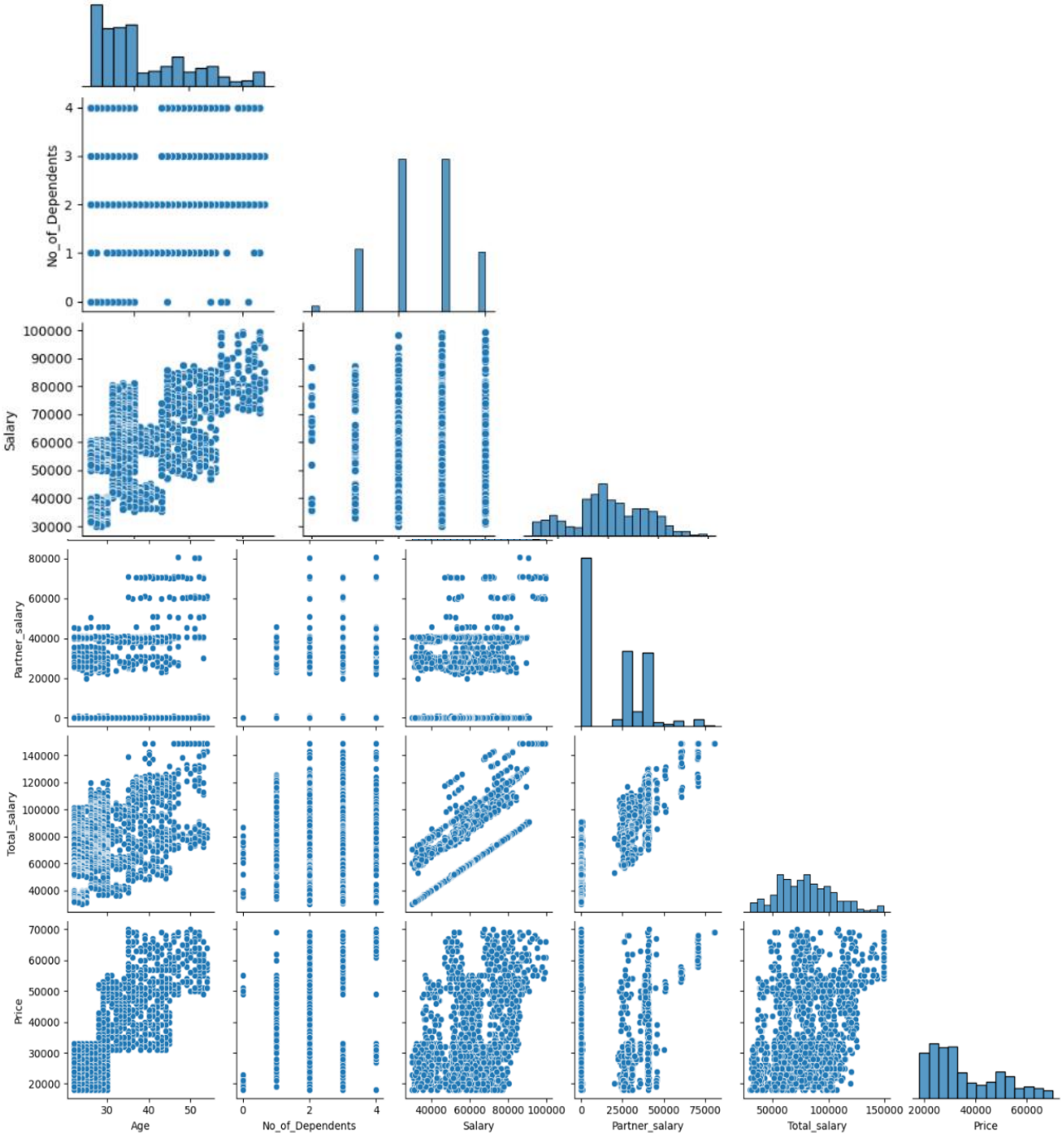
```
<Axes: xlabel='Partner_working', ylabel='count'>
```



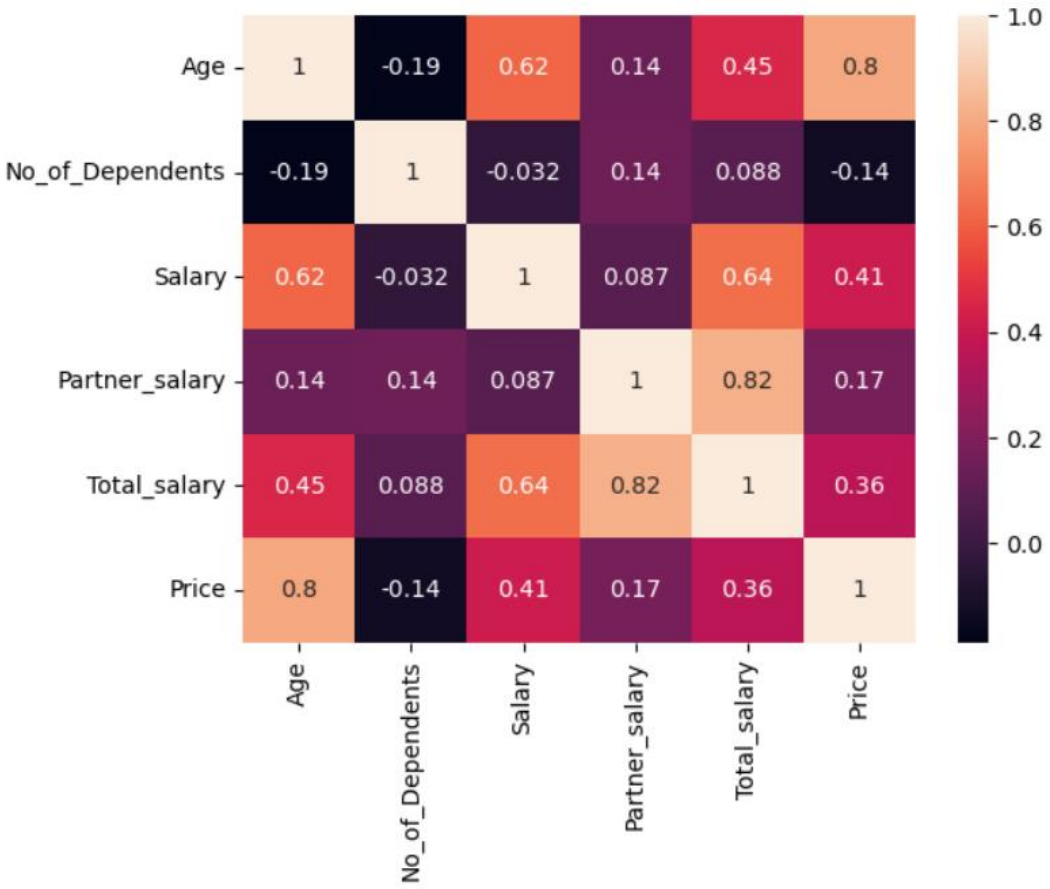
- Working gender is Male is more than Female
- Salaried people are larger number than business people
- Education wise post-graduation people are higher than graduation people
- Marital status has more married people than single people
- Personal loan doesn't show much difference
- House loan is less in number
- Partner working is more working people are there
- Sedan is more manufactured than the Hatchback and 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.

By using the pair plot we have got the graph for complete data.



<Axes: >

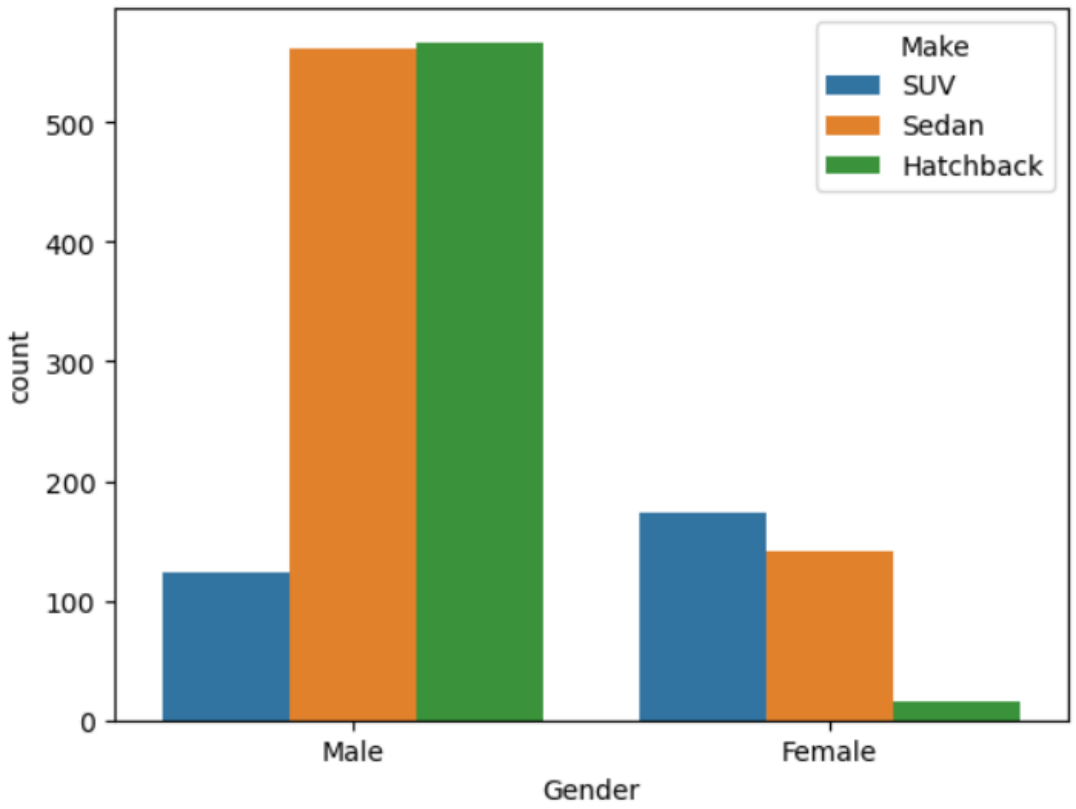


- As we can see from the heatmap high correlation between Total_salary and Partner_salary is there.
- Price and Age are also highly correlated with each other.

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”

```
<Axes: xlabel='Gender', ylabel='count'>
```

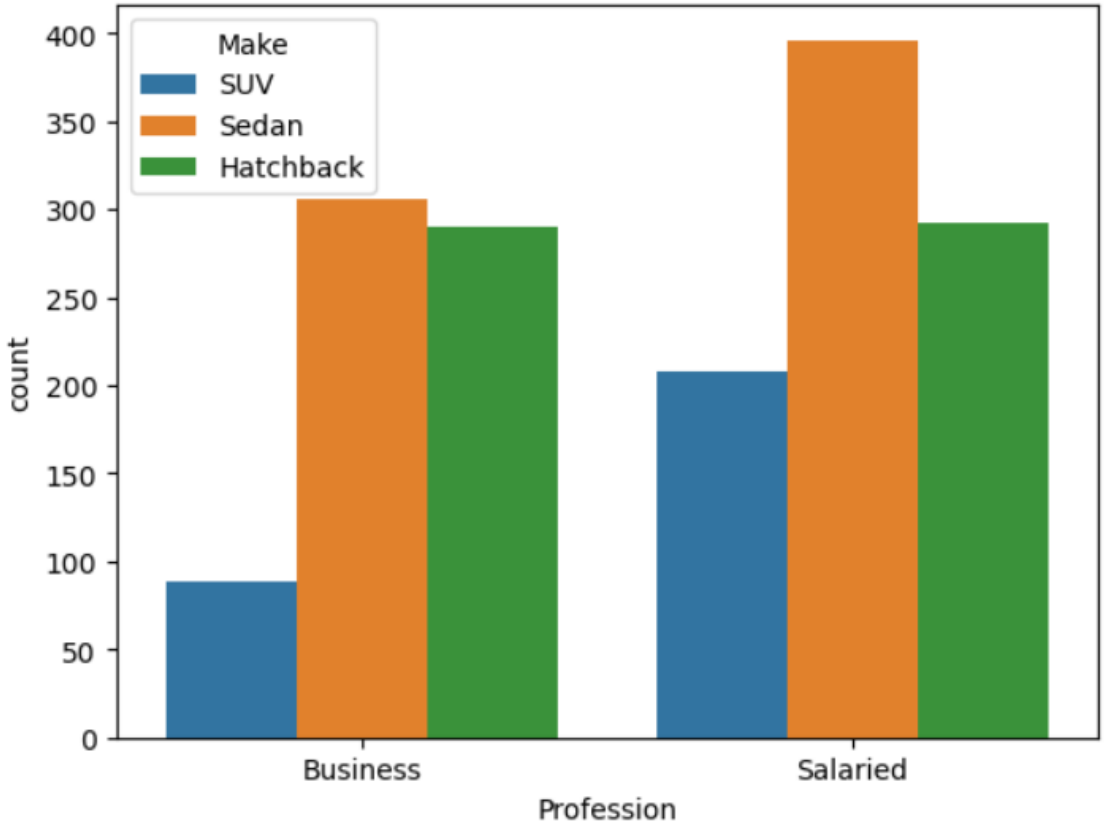



```
Make       Gender
Hatchback  Male      567
           Female     15
SUV        Female    173
           Male      124
Sedan      Male      561
           Female    141
Name: Gender, dtype: int64
```

- The proportion of SUV bought by Female is more than Male
- Female bought SUV are 173 whereas Male bought SUV are 124
- Therefore, Steve Roger statement is False.

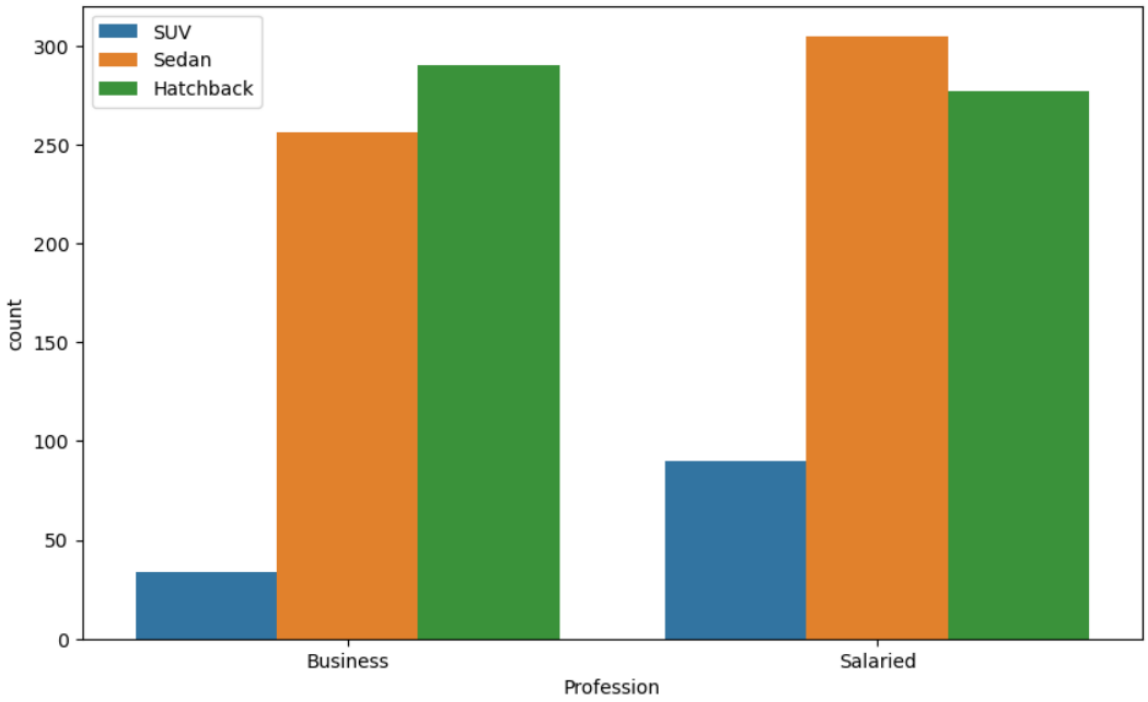
E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.

```
<Axes: xlabel='Profession', ylabel='count'>
```



- Proportion of Salaried people who bought Hatchback is 0.36
- Proportion of Salaried people who bought SUV is 0.18
- Proportion of Salaried people who bought Sedan is 0.44
- Therefore, Ned Stak statement is True.

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 graph we see that Salaried man only preferred Sedan.
- Therefore, Sheldon Copper's statement is False.

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

Here we calculate the average and median of purchasing automobile Gender wise.

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

```
Gender
Female    49000.0
Male      29000.0
Name: Price, dtype: float64
```

- Female has spent 47705.16 avg amount on purchasing automobile
- Male has spent 32416.13 avg amount on purchasing automobile.
- Therefore, we can say Female spent more money than Male on buying/purchasing automobile.

Gender	Mean	Median
Female	**47705.16**	**49000**
Male	**32416.13**	**29000**

F2) Personal_loan

```
Personal_loan
No     36742.712294
Yes    34457.070707
Name: Price, dtype: float64
```

```
Personal_loan
No     32000.0
Yes    31000.0
Name: Price, dtype: float64
```

- As we can see from above the Personal_loan spent on purchasing automobile.
- Business can be utilize this data for more sales, many people bought this personal loan. They can purchase car on easy repayments and lower interest rate, longer loan payments

Personal Loan	Mean	Median
Yes	34457.07	31000
No	36742.71	32000

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

- Mean price of the Partner_working
  1) Partner_working (Yes) = 35267.28
  2) Partner_working (No) = 36000

- Median price of Partner_working
  1) Partner_working (Yes) = 31000
  2) Partner_working (No) = 31000

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.

```
Gender  Marital_status  Make
Female  Married         SUV          166
                        Sedan        127
                        Hatchback     14
        Single          Sedan         14
                        SUV            7
                        Hatchback      1
Male    Married         Sedan        537
                        Hatchback    484
                        SUV          115
        Single          Hatchback     83
                        Sedan         24
                        SUV            9
Name: Make, dtype: int64
```

To make marketing strategy efficient as we previously saw this data has been frequent variable, Make, Gender and Marital_status

Group	Gender	Marital_status	Make
1	Male	Married	SUV
2	Female	Married	Sedan
3	Male	Single	Hatchback
4	Female	Single	Sedan

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.)

These are the variables given for analysis.

```
 #   Column
---  ------
 0   userid
 1   card_no
 2   card_bin_no
 3   Issuer
 4   card_type
 5   card_source_date
 6   high_networth
 7   active_30
 8   active_60
 9   active_90
 10  cc_active30
 11  cc_active60
 12  cc_active90
 13  hotlist_flag
 14  widget_products
 15  engagement_products
 16  annual_income_at_source
 17  other_bank_cc_holding
 18  bank_vintage
 19  T+1_month_activity
 20  T+2_month_activity
 21  T+3_month_activity
 22  T+6_month_activity
 23  T+12_month_activity
 24  Transactor_revolver
 25  avg_spends_l3m
 26  Occupation_at_source
 27  cc_limit
```

Card type customers use a different types of various variety card, we can get business analysis from this most type of cards are used.

Cc_active is used to check the last active user of the cards, how many times they are used.

T+1, T+2, T+3 has the better analysis how much transaction are gone through in the month activity.

T+6 and T+12 should be excluded from details, because may times cards used for transactions.

As per my analysis the top 5 important variables can be used analysis.

Cc_active 30 is used to check the last active of the card, how many time they are used and transaction recently or not.

Annual income source used for estimating their income, directly spending power of the customer.

T+1, T+2, T+3 are have better analysis how much transactions are gone through in the month activity.

Avg_spends_3m is used for estimation how much customer spend in the 3 month wise.

Cc_limit is used for the find the limit of the card and current limit of the card available, depending up on the spending of the customer, high spending customer of the credit like to admit more they spend on the future, lower spending customer of the credit help to focus on the right card can be suggested for the customer to changes variety of credit card.