SMDM Business Report

Contents:-

Problem:- 1 Austo Motor Company problem

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)

B. 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.

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

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

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

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

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

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

sale over a Sedan SaleF. 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.

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

F1) Gender

F2) Personal_loan

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

Problem:- 2 Framing An Analytics Problem

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

List Of Figures

List of Tables

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.

1. **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.**

The instructions below are given to help you complete the project

Table 1 Top 5 rows of the data set are as follows

Age	Gender	Profession	**Marital**_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	Price	Make

53	Male	Business	Married	Post Graduate	4	No	No	Yes	99300	70700	170000	61000	SUV
53	Femal	Salaried	Married	Post Graduate	4	Yes	No	Yes	95500	70300	165800	61000	SUV
53	Female	Salaried	Married	Post Graduate	3	No	No	Yes	97300	60700	158000	57000	SUV
53	Female	Salaried	Married	Graduate	2	Yes	No	Yes	72500	70300	142800	61000	SUV
53	Male	Salaried	Married	Post Graduate	3	No	No	Yes	79700	60200	139900	57000	SUV

```
 no. of rows: 1581
no. of columns: 14
```

Table-2 Basic 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
```

Observations

-Data has been loaded into the pandas dataframe

-There are 1581 rows and 14 columns

-There are 6 numerical and 8 categorical variables present

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

Table:- 3 Checking the Null values

```
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
```

Total Null Values in Gender= 53,

-Total Null values in Partner_salary=106

Handling the Null Values

1. Deleting Rows

This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set.

2. Replacing With Mean/Median/Mode

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values

3. Assigning An Unique Category

A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values

4. Predicting The Missing Values

Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance.

5. Using Algorithms Which Support Missing Values

KNN is a machine learning algorithm which works on the principle of distance measure. This algorithm can be used when there are nulls present in the dataset. While the algorithm is applied, KNN considers the missing values by taking the majority of the K nearest values.

For categorical values we have use imputing the Null values with majority class and for continuous values we have use KNN imputer to treat the null values

Inspecting the duplicates

```
No of duplicate rows= 0
```

Table :- 4 Numerical summarization of the data

	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price

count	1581.000000	1581.000000	1581.000000	1581.000000	1581.000000	1581.000000
mean	31.922201	2.457938	60392.220114	20225.559322	79625.996205	35597.722960
std	8.425978	0.943483	14674.825044	18905.183912	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	24900.000000	78000.000000	31000.000000
75%	38.000000	3.000000	71800.000000	38000.000000	95900.000000	47000.000000
max	54.000000	4.000000	99300.000000	80500.000000	171000.000000	70000.000000

Table- 5 Skewness and kurtosis of the dataset

Variable	Skewness	kurtosis
Age	**0.89**	**-0.24**
No_of_Dependents	**-0.12**	**-0.54**
Salary	**-0.11**	**-0.51**

Partner_salary	**0.35**	**-0.74**
Total_salary	**0.60**	**0.64**
Price	**0.74**	**-0.57**

Observations:-

1) We have observed from the above dataset that the minimum age is 22 years whereas the maximum age is 54years and the average age is 29 years with the positively skewness 0.89 and kurtosis -0.24

2) We have observed from the above dataset that the minimum No_of_dependents is 0 whereas the maximum No_of_dependents is 4 and the average No_of_dependents is 2 with the negatively skewness -0.12 and kurtosis -0.54

3)We have observed from the above dataset that the minimum Salary is 30k whereas the maximum Salaryts is 99k and the average Salary is 51k with the skewness very cloase to 0

4)We have observed from the above dataset that the minimum Total_salary is 30k whereas the maximum Total_salary is 171k and the average Total_salary is 60k with the skewness are kurtosis is almost equal

We have observed from the above dataset that the minimum price of automobile is 18k whereas the maximum price of automobile is 70k and the average price of automobile is 25k with the moderate skewness of 0.74

Table :- 6 Checking for anomalous values in categorical values

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

Salaried    896
Business    685
Name: Profession, dtype: int64

Married    1443
```

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

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

Yes    792
No     789
Name: Personal_loan, dtype: int64

No     1054
Yes     527
Name: House_loan, dtype: int64

Yes    868
No     713
Name: Partner_working, dtype: int64

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

From the value counts it is observed that the categorical fields are free from anomalies

Now inspecting the anomalies as follows:

Figure:-1 Boxplot of Price

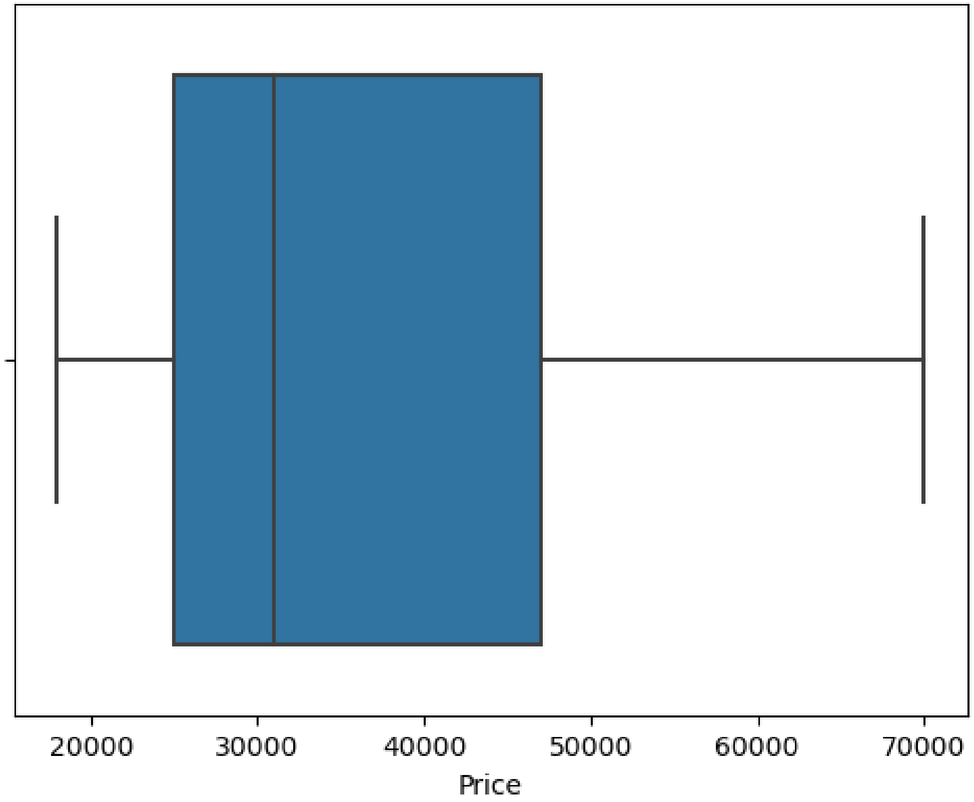


Figure:-2 Boxplot of Total Salary

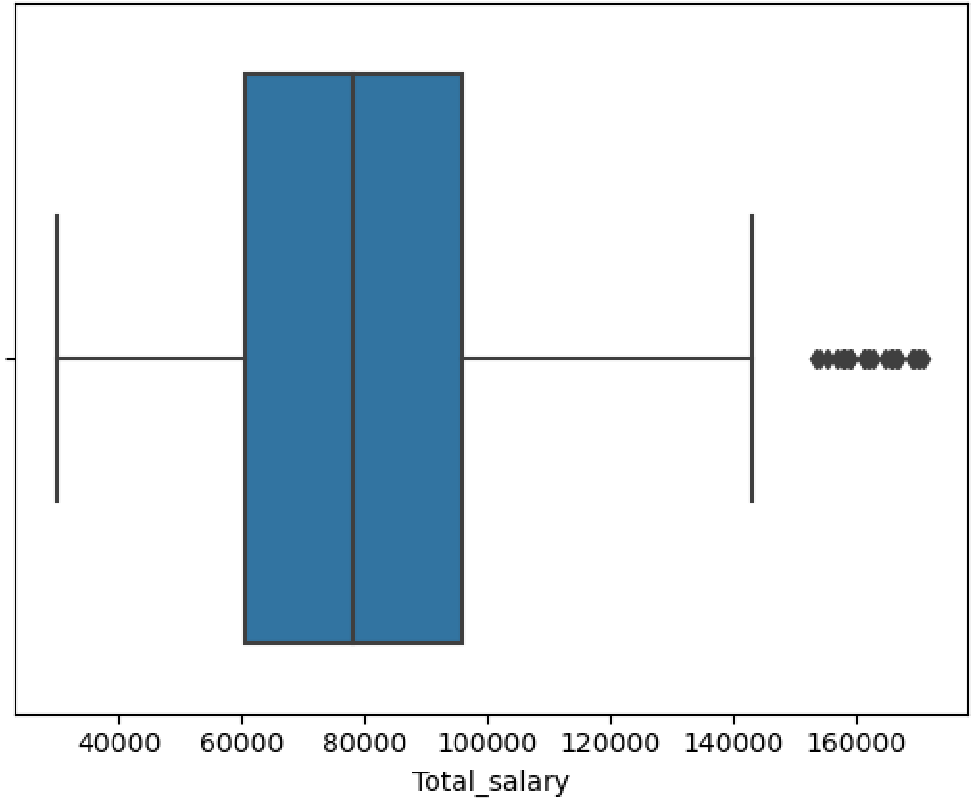


Figure:-3 Boxplot of Partner Salary

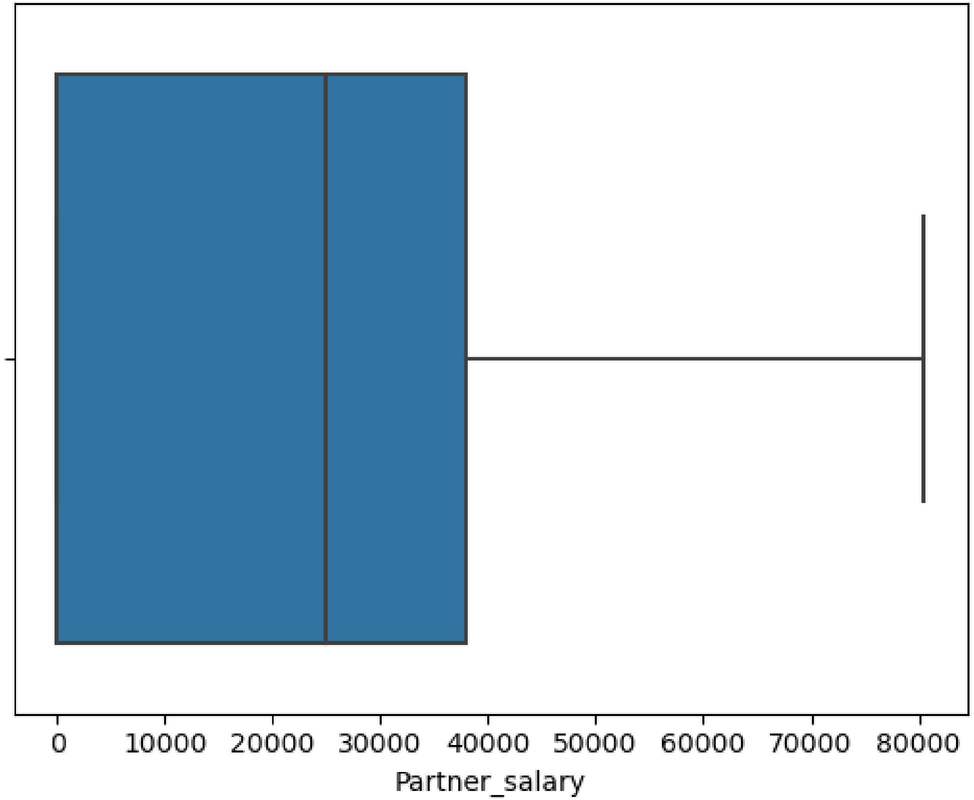


Figure:-4 Boxplot of Salary

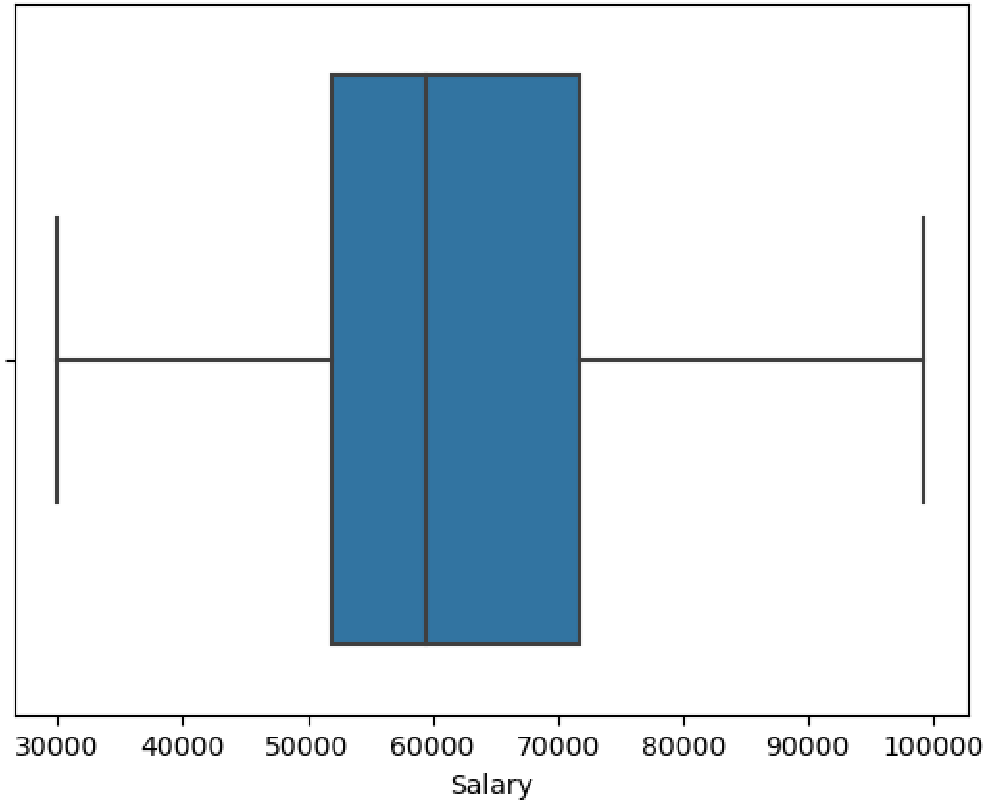
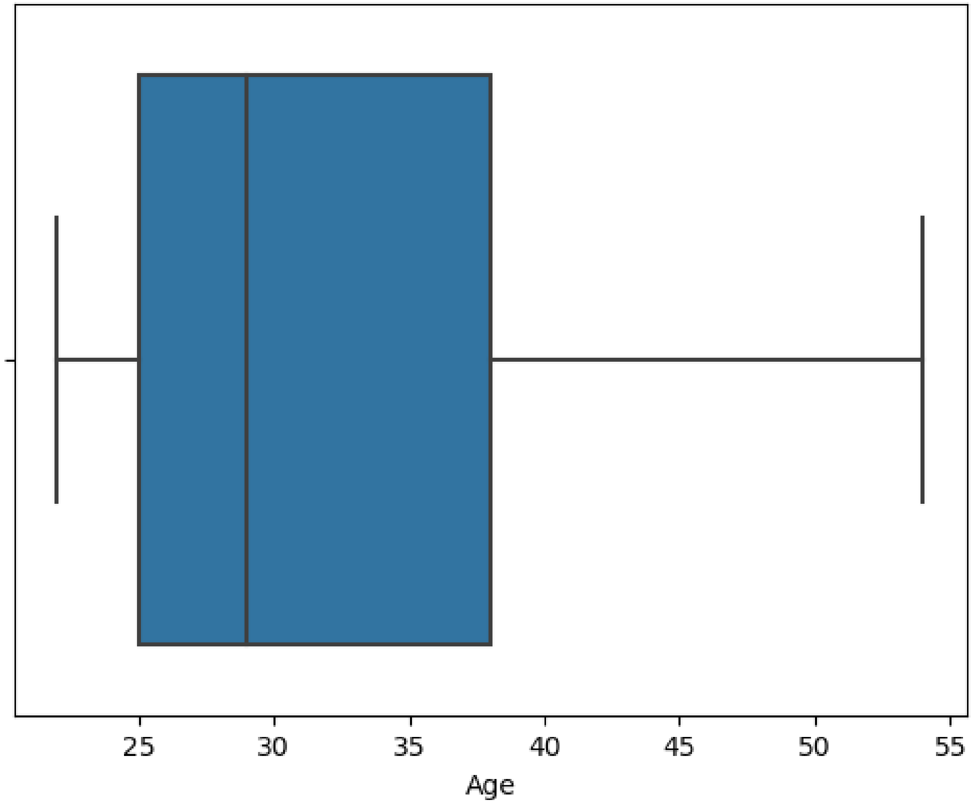


Figure:-5 Boxplot of Age



Observations:-

1)From the above boxplots it is observed that there are no negative values present in the numerical fields

2)Outliers are present in the Total_salary

Now Outliers are treated by using the IQR method

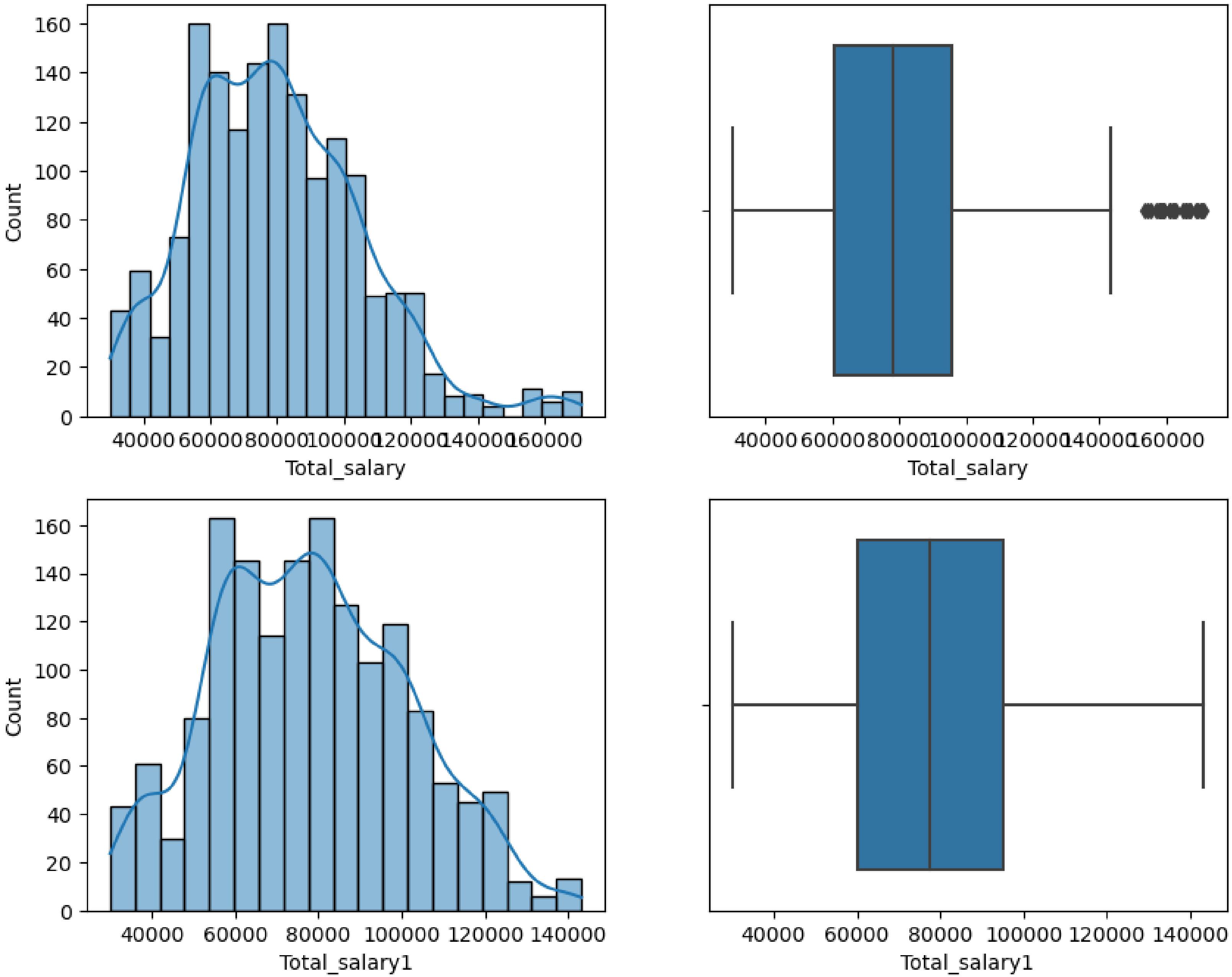
percentile25= 60500.0
percentile75= 95900.0

```
iqr=percentile75-percentile25= 35400.0

upper_limit=percentile75+1.5*iqr= 149000.0

lower_limit=percentile25-1.5*iqr= 7400.
```

Figure:-6 Comparison of the Outliers with Histplot and Boxplot


Count
0
20
40
60
80
100
120
140
160
40000
60000
80000
100000
120000
140000
160000
Total_salary
40000
60000
80000
100000
120000
140000
160000
Total_salary
Count
0
20
40
60
80
100
120
140
160
40000
60000
80000
100000
120000
140000
Total_salary1
40000
60000
80000
100000
120000
140000
Total_salary1

From the above graph it is clearn that outliers of the Total_salary has been treated

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 Values

Figure:-7 Comparison of the Age & Salary with Histplot and Boxplot

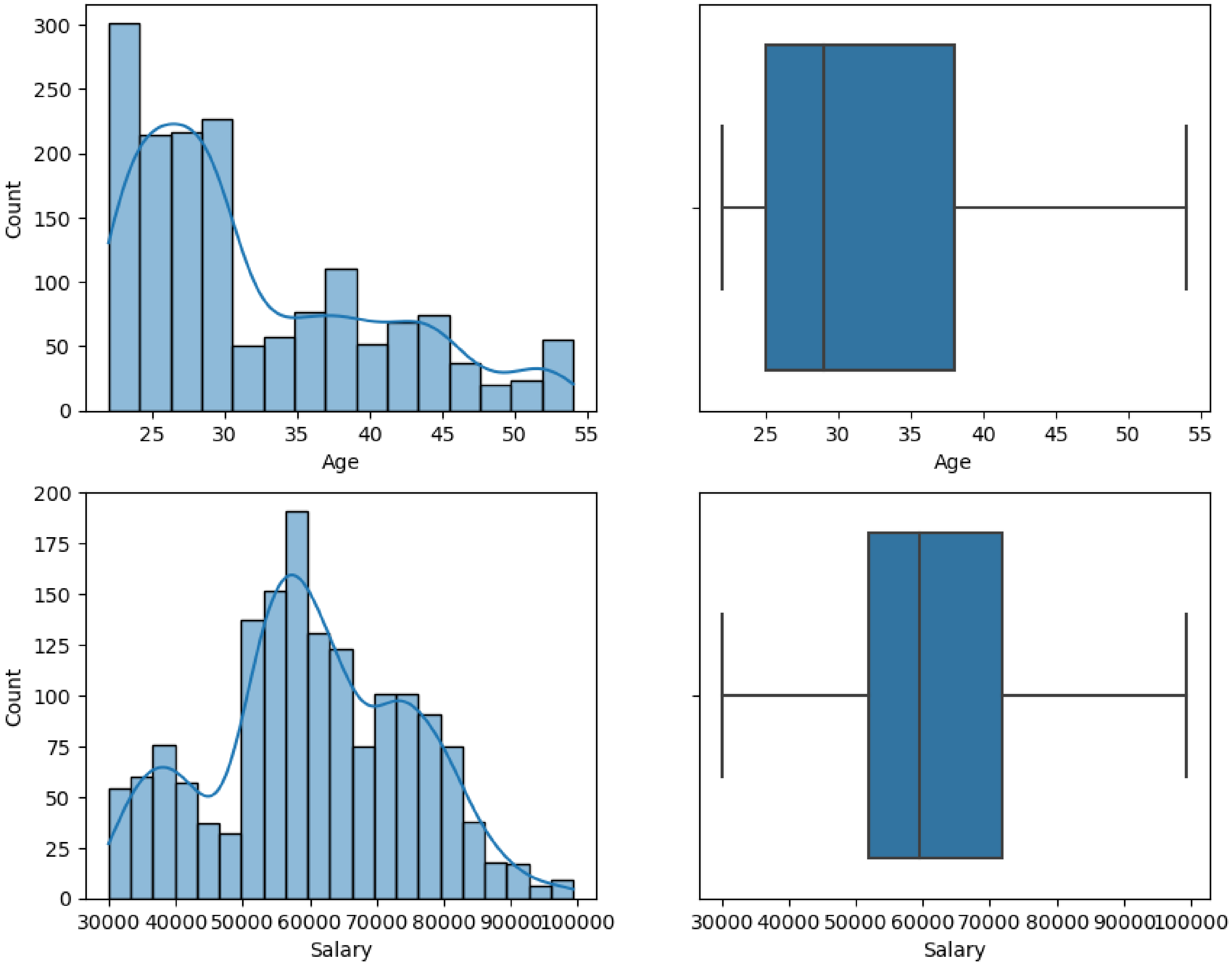
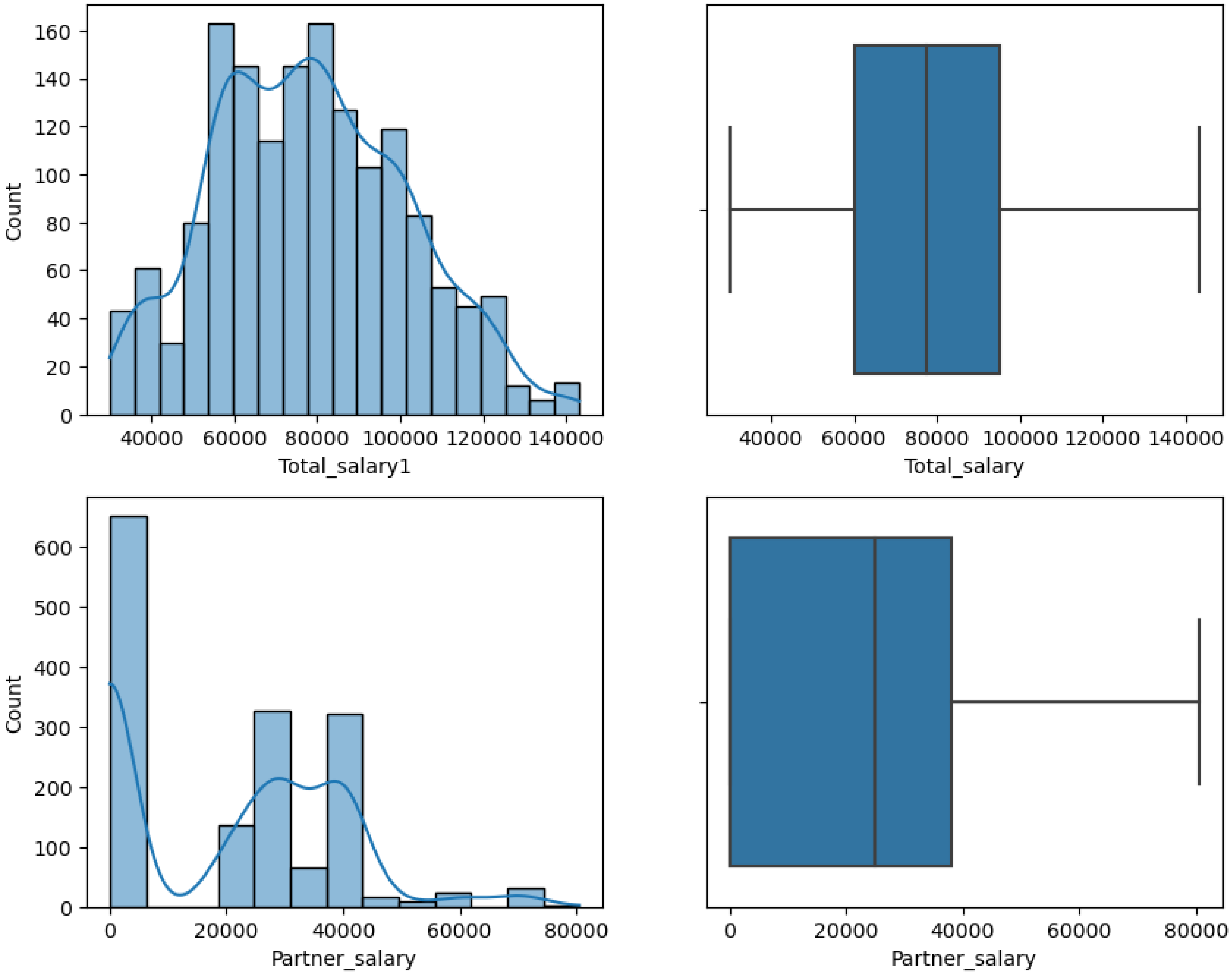

Count
300
250
200
150
100
50
0
25
30
35
40
45
50
55
Age
25
30
35
40
45
50
55
Age
Count
200
175
150
125
100
75
50
25
0
30000
40000
50000
60000
70000
80000
90000
100000
Salary
30000
40000
50000
60000
70000
80000
90000
100000
Salary

Figure:-7 Comparison of theTotal Slary & Partner Salary with Histplot and Boxplot


Count
160
140
120
100
80
60
40
20
0
40000
60000
80000
100000
120000
140000
Total_salary1
40000
60000
80000
100000
120000
140000
Total_salary
Count
600
500
400
300
200
100
0
0
20000
40000
60000
80000
Partner_salary
0
20000
40000
60000
80000
Partner_salary

Insights

From the univariate analysis of the numerical fields we have observed following things:-

1) Age of customer seems to be multimodal distribution generally lies betwen 25 years to 40 years

2) Salary seems to be in the range 50k to 70k with multimodal distribution

3)The skewness of the Total_salary seems to be reduced and it ranges 60k to 100k

4) Partner-salary lies between 0k to 40k

Univariate analysis of Categorical Values

Figure:-8 Count plot of Gender

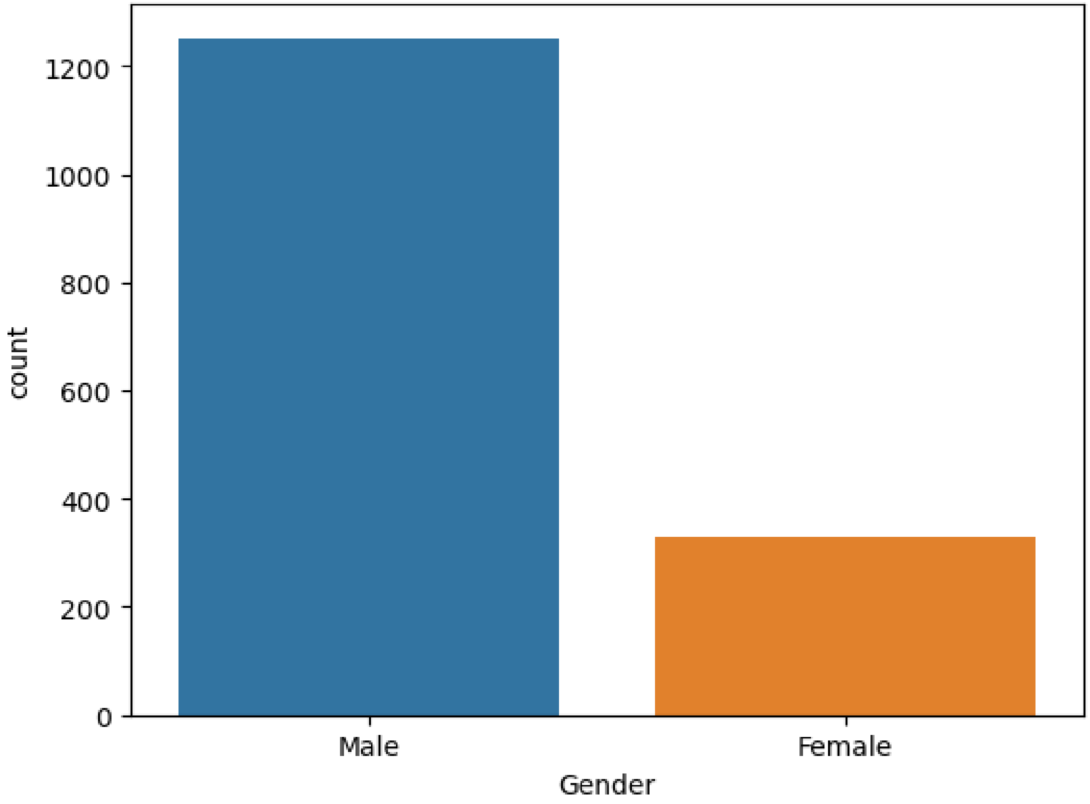


Figure:-9 Count Plot of Education

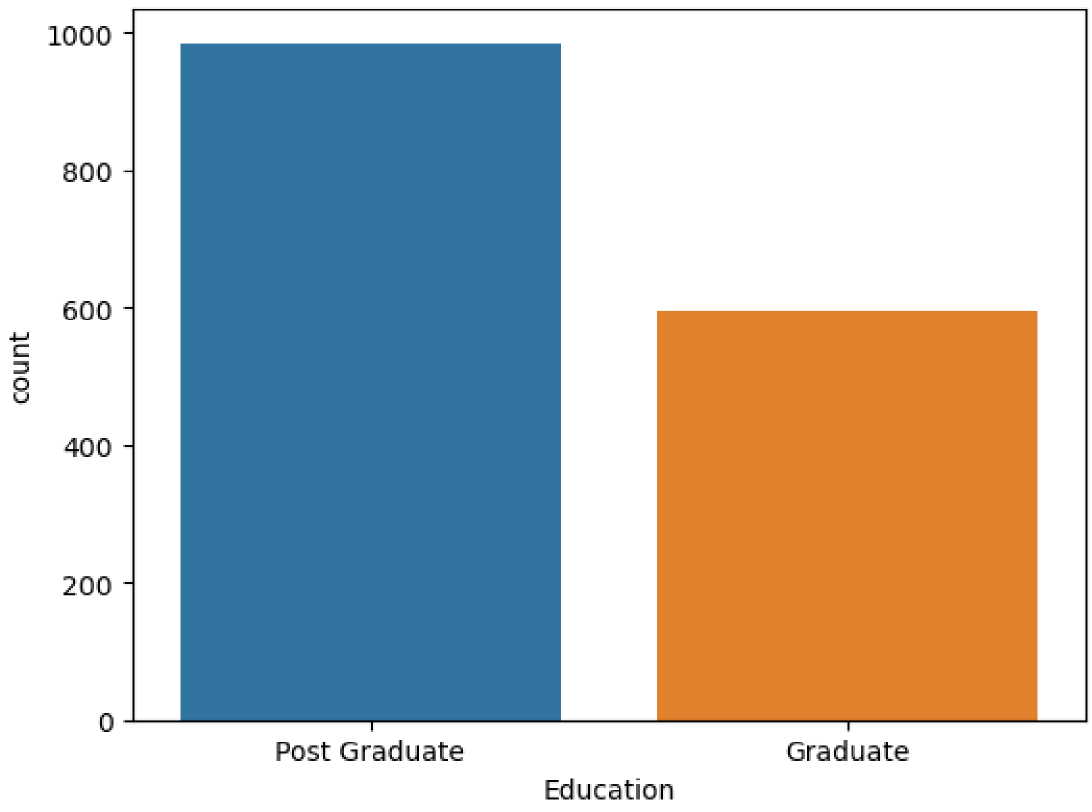


Figure:-10 Count Plot of Profession

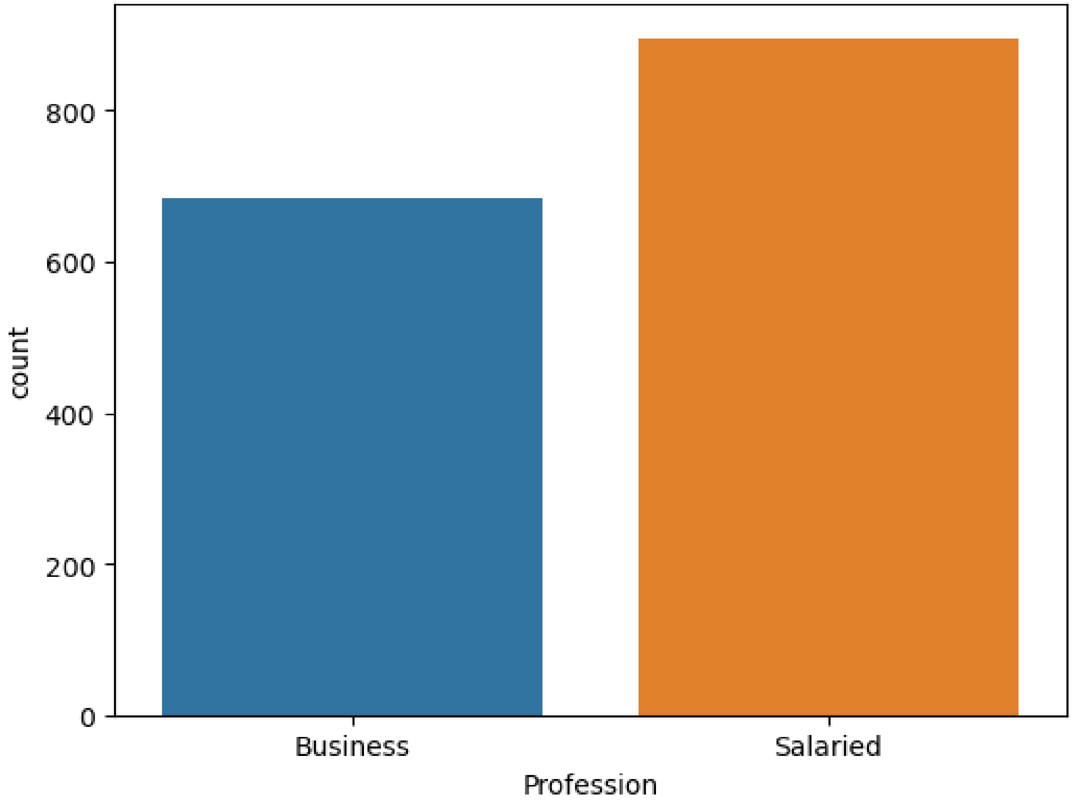


Figure:-11 Count Plot of Marital Status

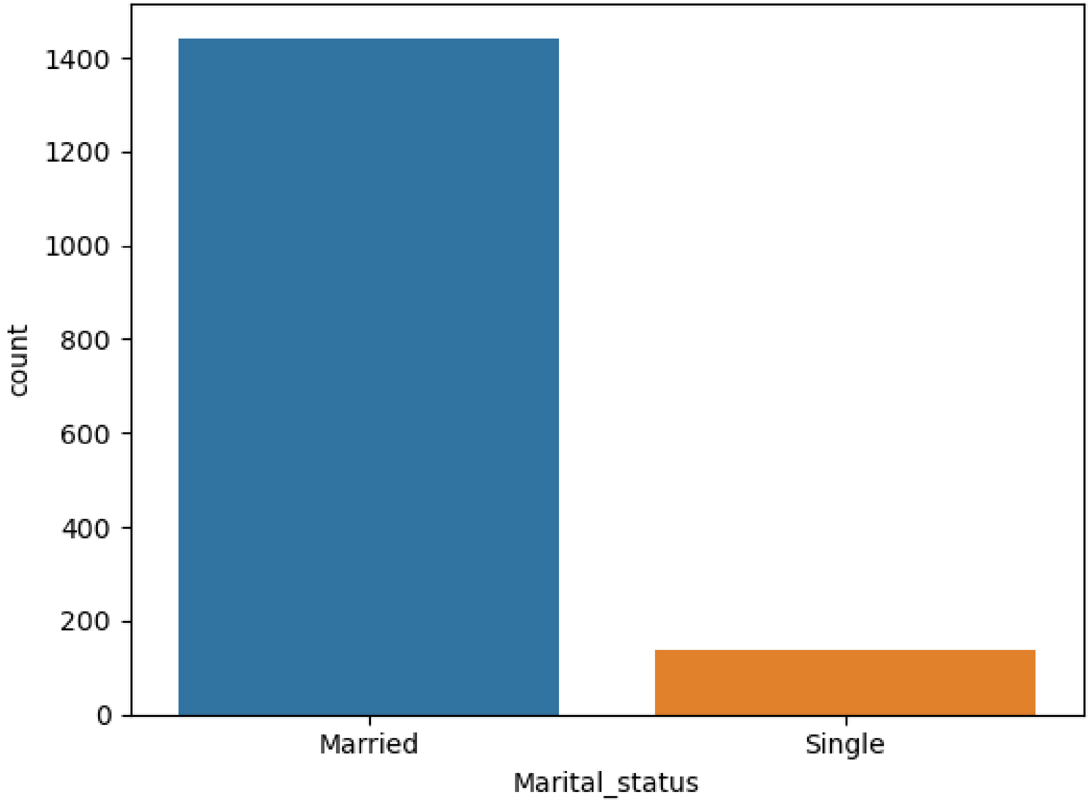


Figure:-12 Count Plot of Personal loan

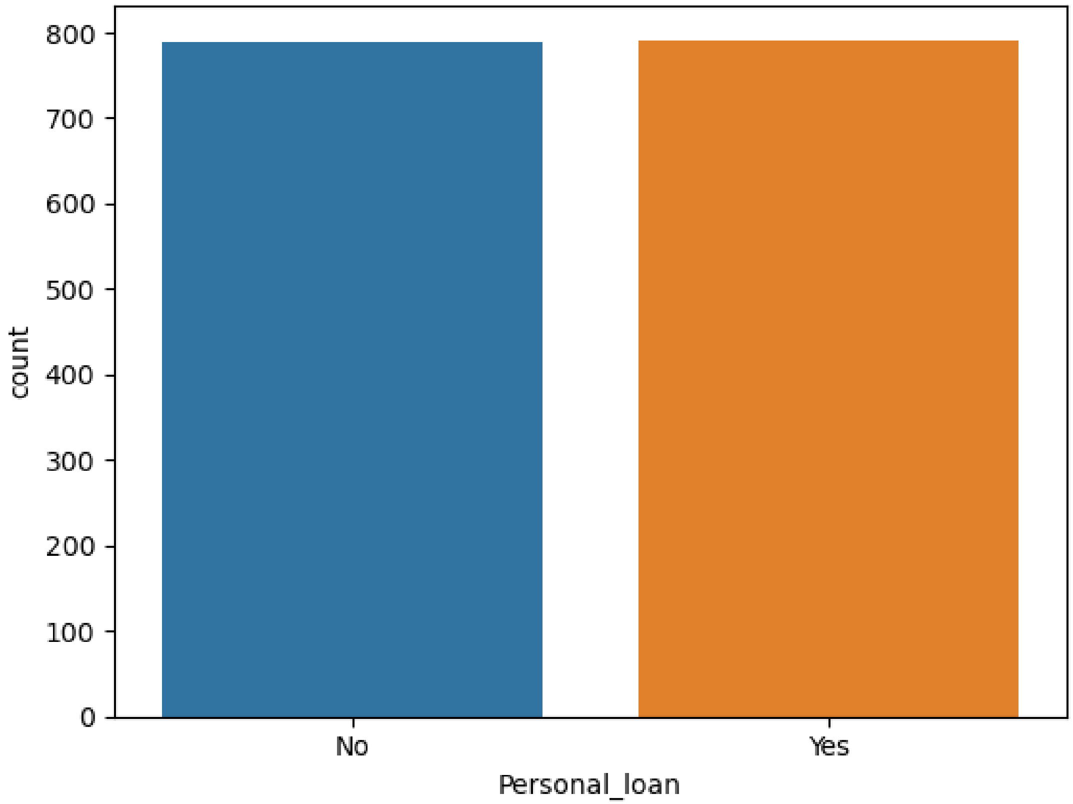


Figure:-13 Count Plot of House Loan

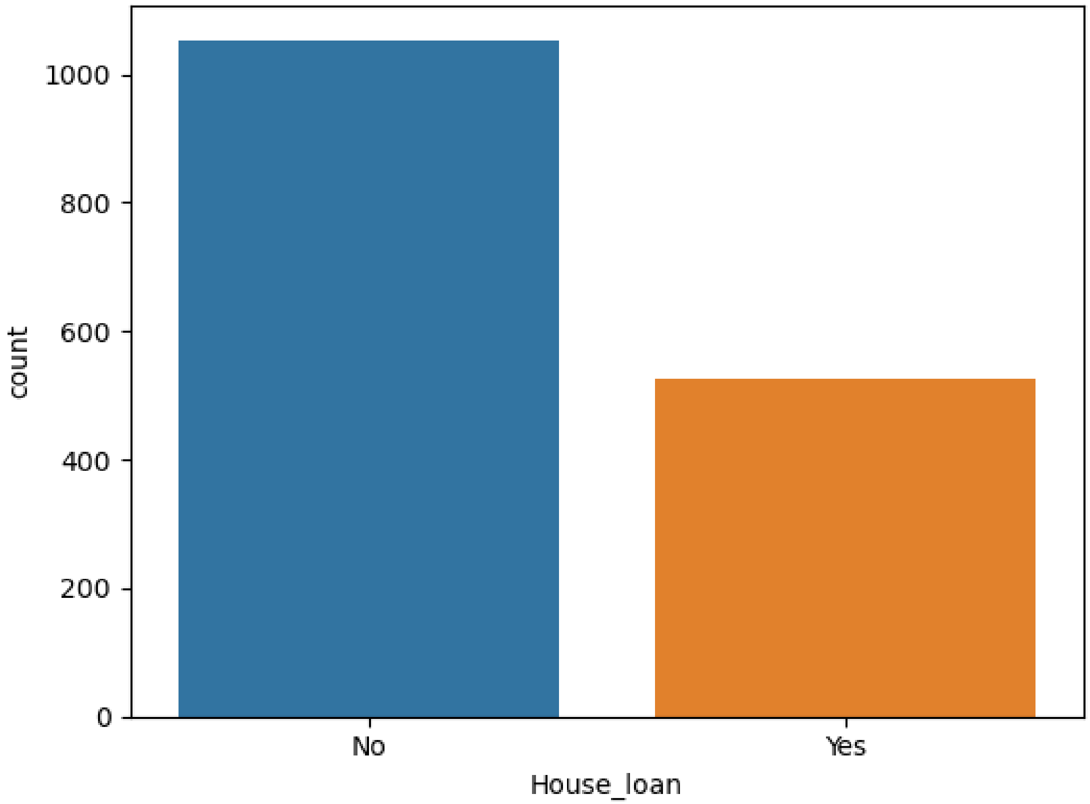


Figure:-14 Count Plot of No of Dependants

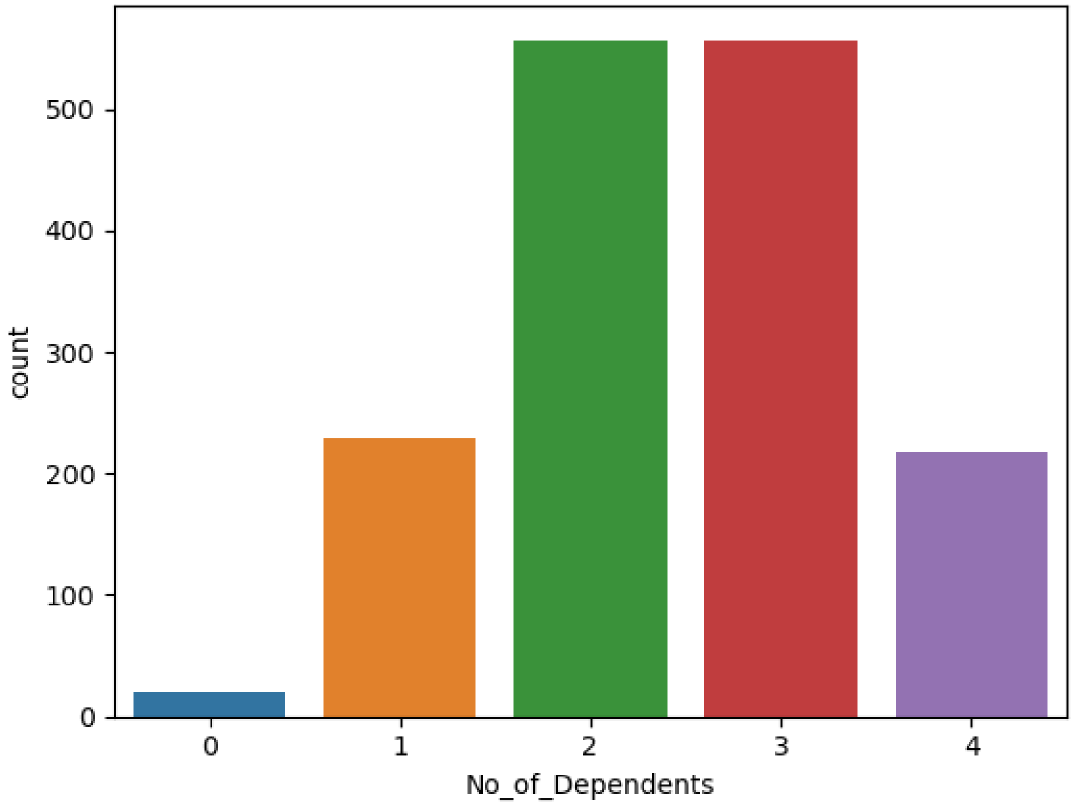


Figure:-15 Count Plot of Partner Working

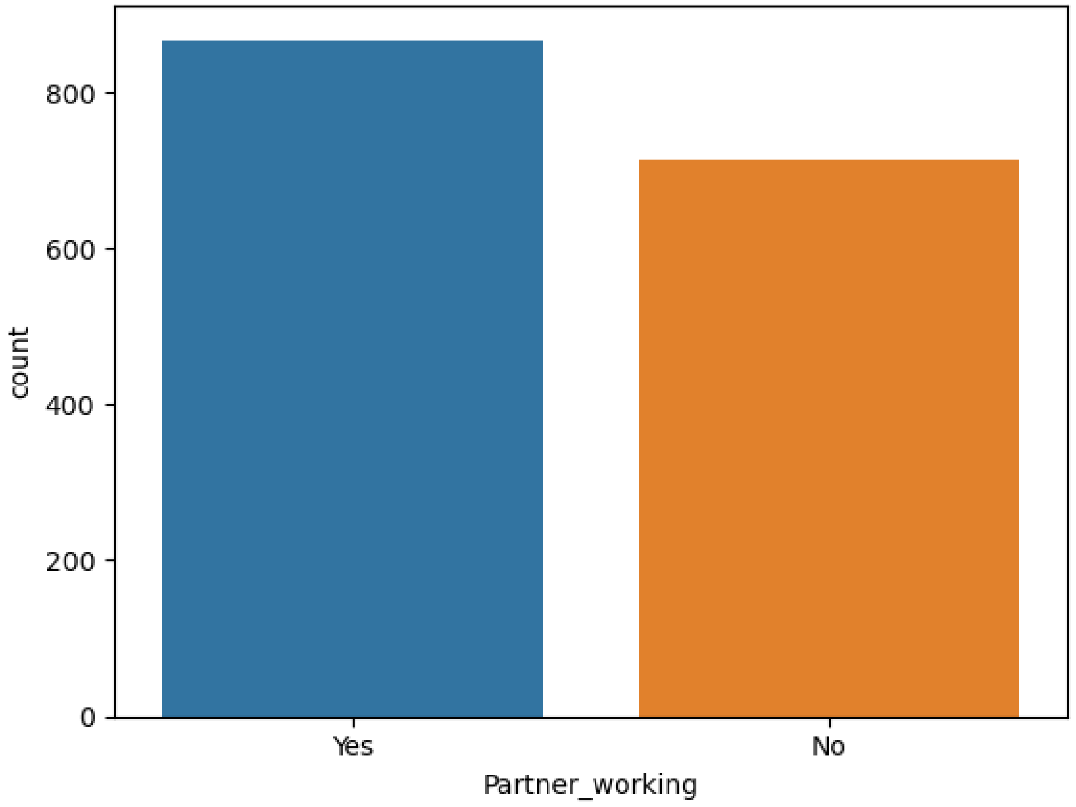
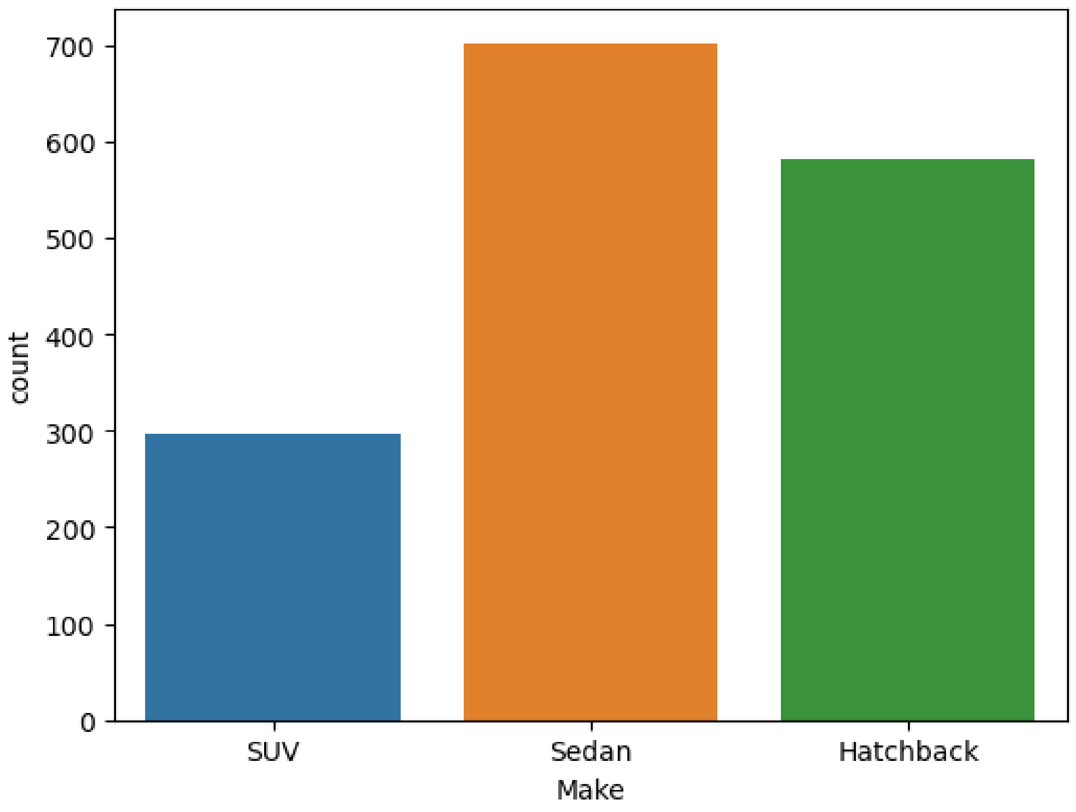


Figure:-16 Count plot of Make



Insights

1) Male customers are higher than female customers

2) Salaried customers are slightly higher than business cutomers

3) Married customers are higher than single customers

4)Post graduate customers are having higher majority

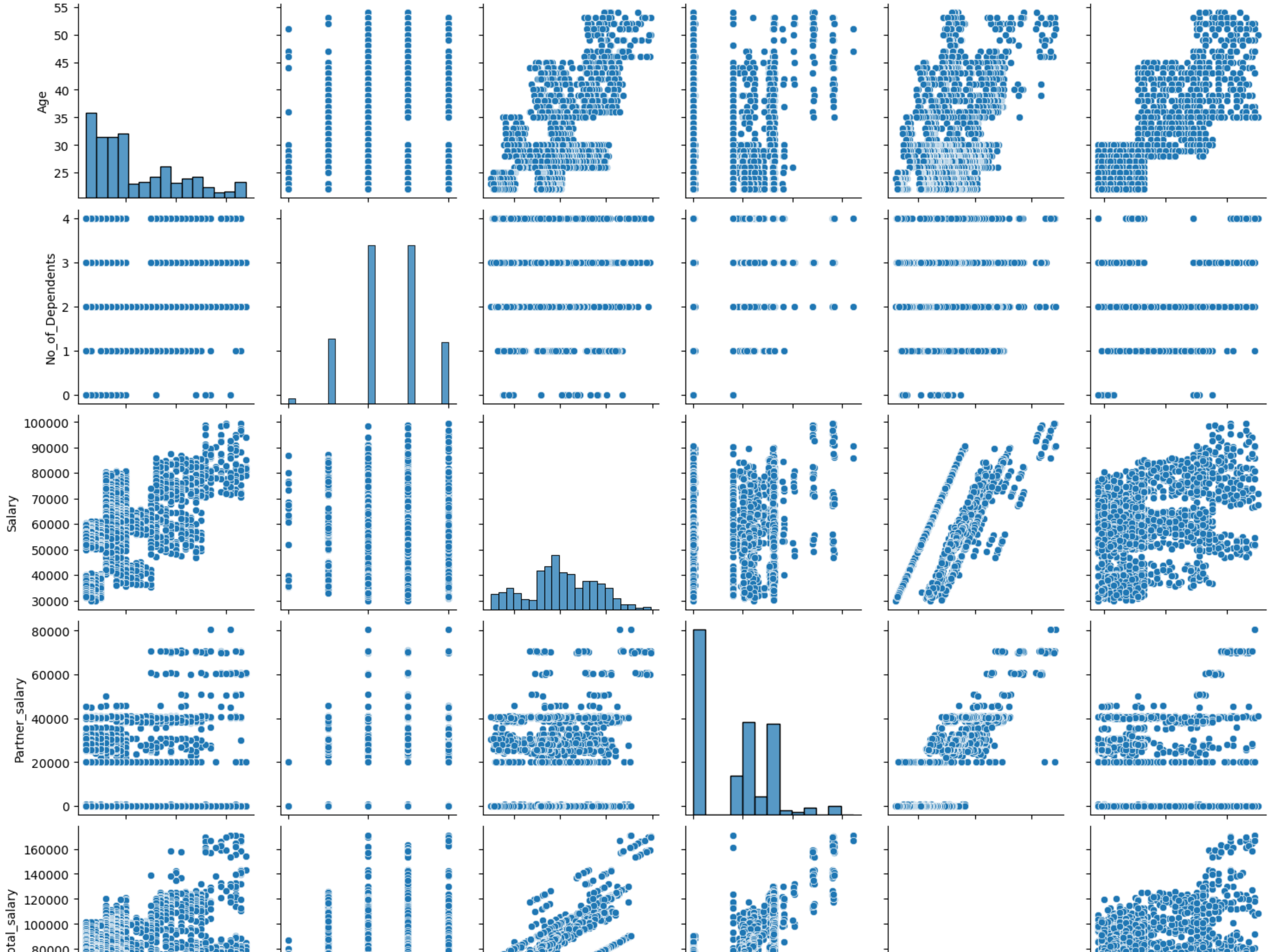
5) majority of customers preferred Sedan than Hatchback than SUV

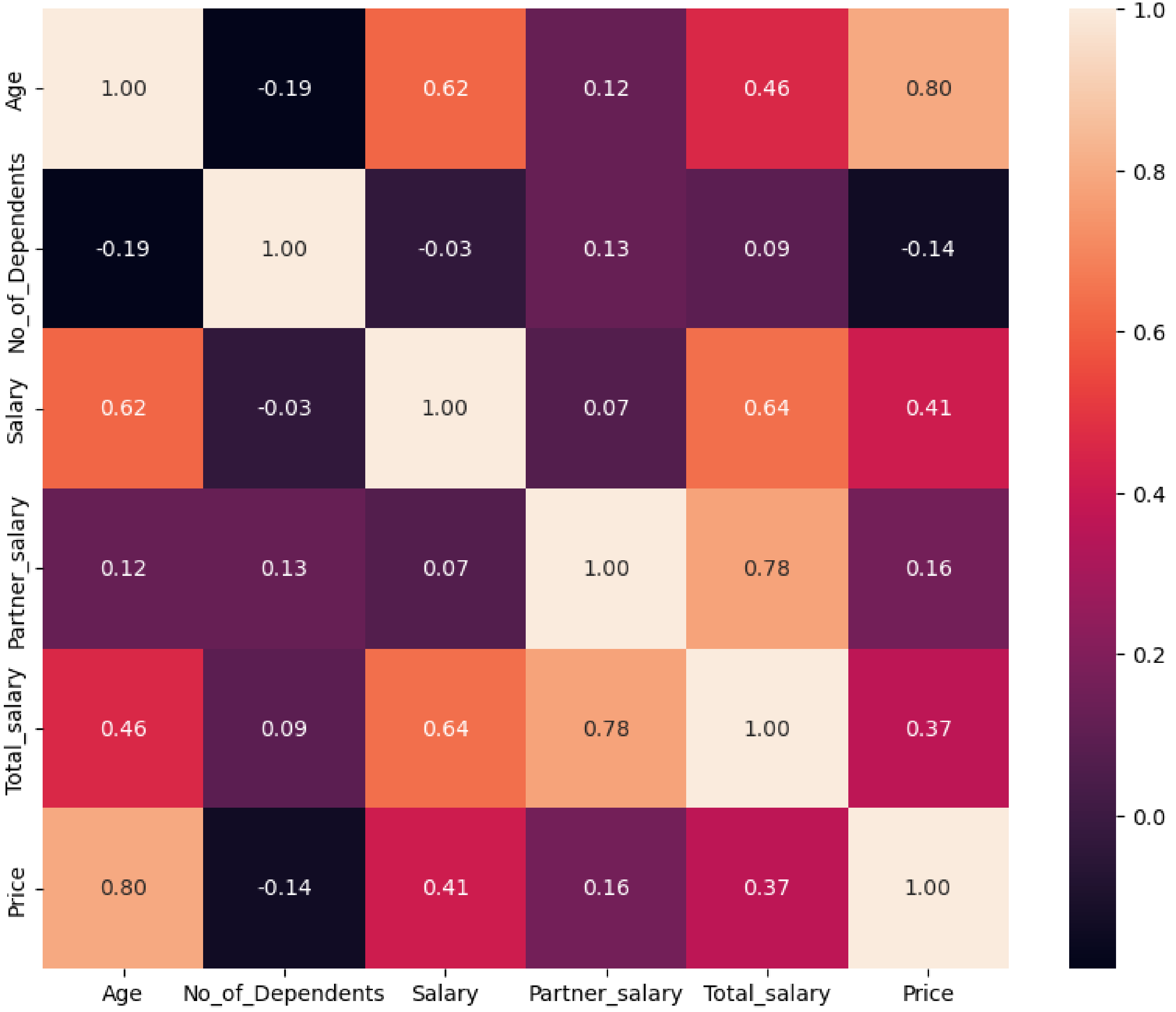
6)From the graph we have observed than there is a slight difference between customers working and non working

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 of Numerical vs Numerical value

Figure:- 17 Pair Plot of Numerical Values


Age
55
50
45
40
35
30
25
No_of_Dependents
4
3
2
1
0
Salary
100000
90000
80000
70000
60000
50000
40000
30000
Partner_salary
80000
60000
40000
20000
0
otal_salary
160000
140000
120000
100000
80000



	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
Age	1.00	-0.19	0.62	0.12	0.46	0.80
No_of_Dependents	-0.19	1.00	-0.03	0.13	0.09	-0.14
Salary	0.62	-0.03	1.00	0.07	0.64	0.41
Partner_salary	0.12	0.13	0.07	1.00	0.78	0.16
Total_salary	0.46	0.09	0.64	0.78	1.00	0.37
Price	0.80	-0.14	0.41	0.16	0.37	1.00

Insights

1) High correlation exists between Price & Age, Tota_salary & Age and Salary & Age

2)There is no linear relationship exits among variables

Bivariate analysis of Categorical vs Categorical value

Figure 18- Count Plot of Make vs gender

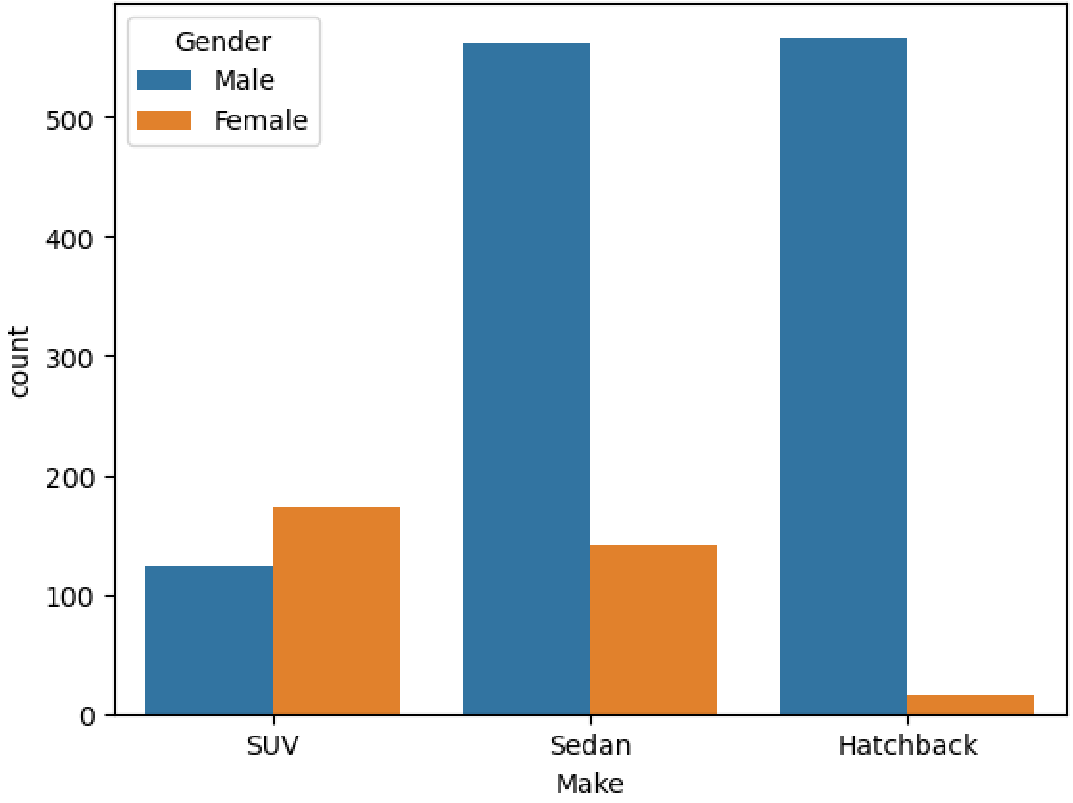


Figure 19- Count Plot of Make vs Marital Status

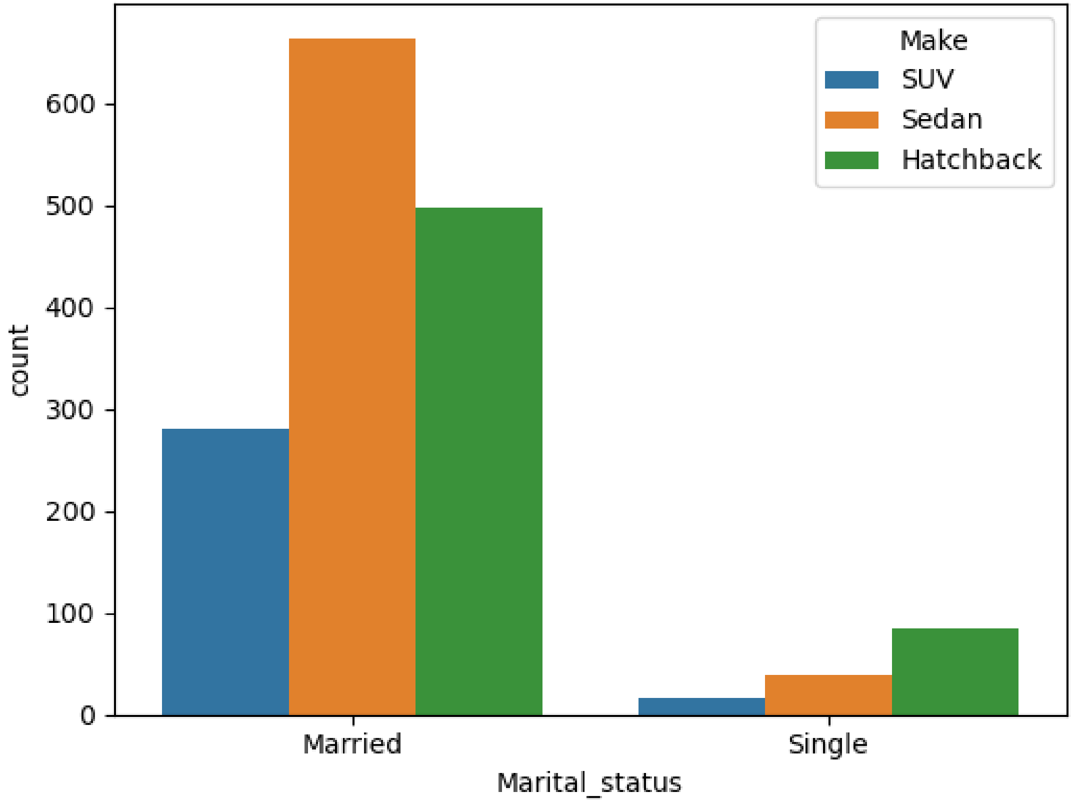


Figure 20- Count Plot of Gender vs Profession

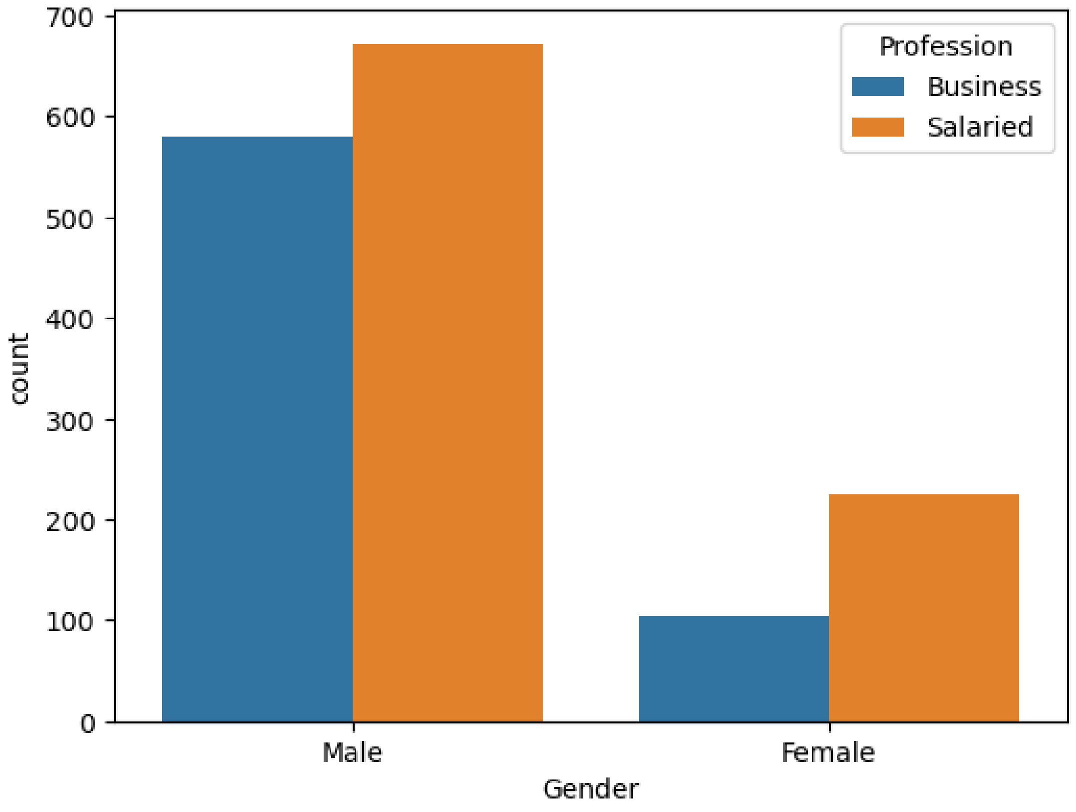
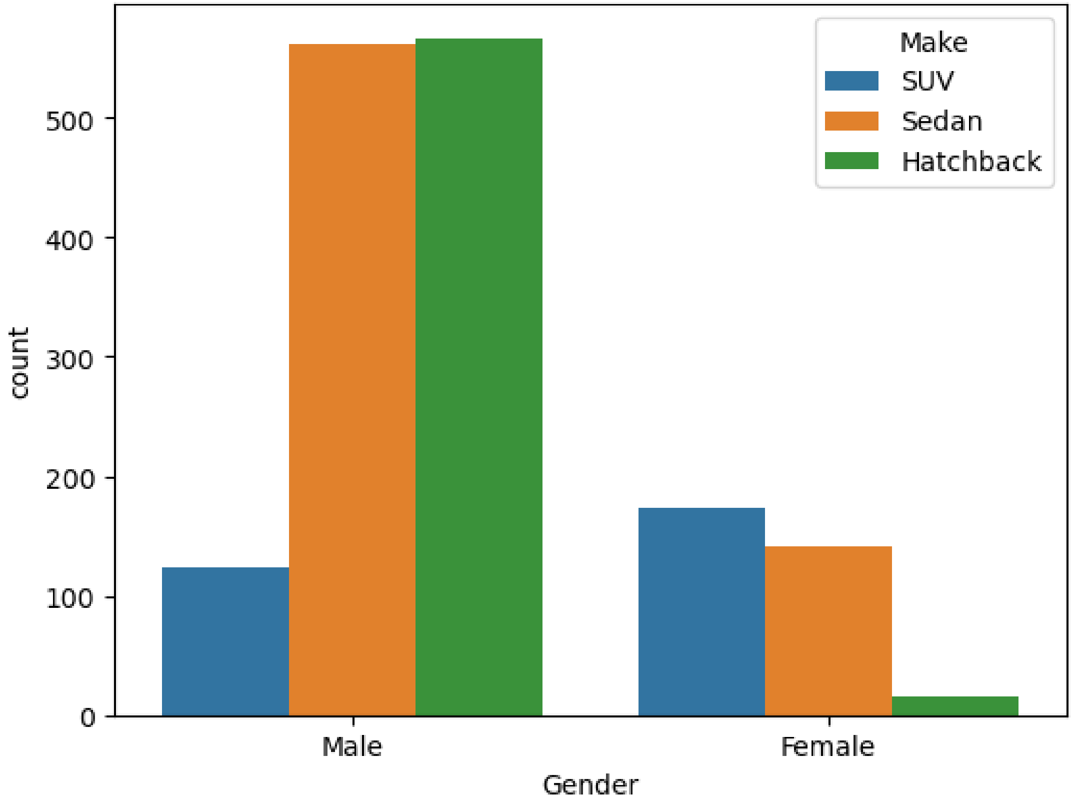


Figure 21- Count Plot of Gender vs Make


Make
SUV
Sedan
Hatchback
500
400
300
200
100
0
count
Male
Female
Gender

Insights

1) It is observed from the graph that male customers prefer sedan where as female customers prefer SUV

2)Married customers purchase more cars than single customers where as married customers purchase sedan slightly higher than singles where as single customers prefer hatchback

3) salaried customers purchase more car than business customers

4) Males purchase more car than females and they generally prefer sedan and 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.

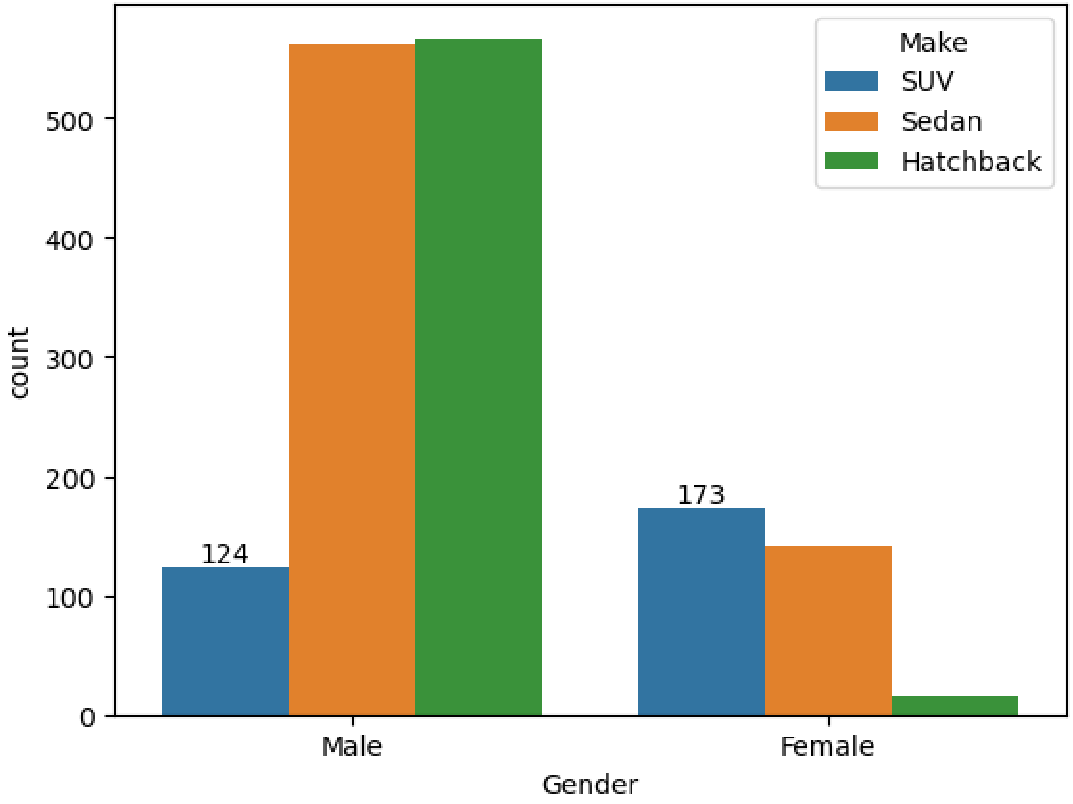
E1) Steve Roger says "Men prefer SUV by a large margin, compared to the women"

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

Proportion of female buying SUV=(No of females brought SUV/Total no of females)=(173/329)=0.52

Proportion of male buying SUV=(No of males brought SUV/Total no of males)=(124/1252)=0.09

Figure22 :- Count plot of Gender vs Make



It is observed from the above graph than females prefer more SUV than male. Thus statement made by Steve Roger is false

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

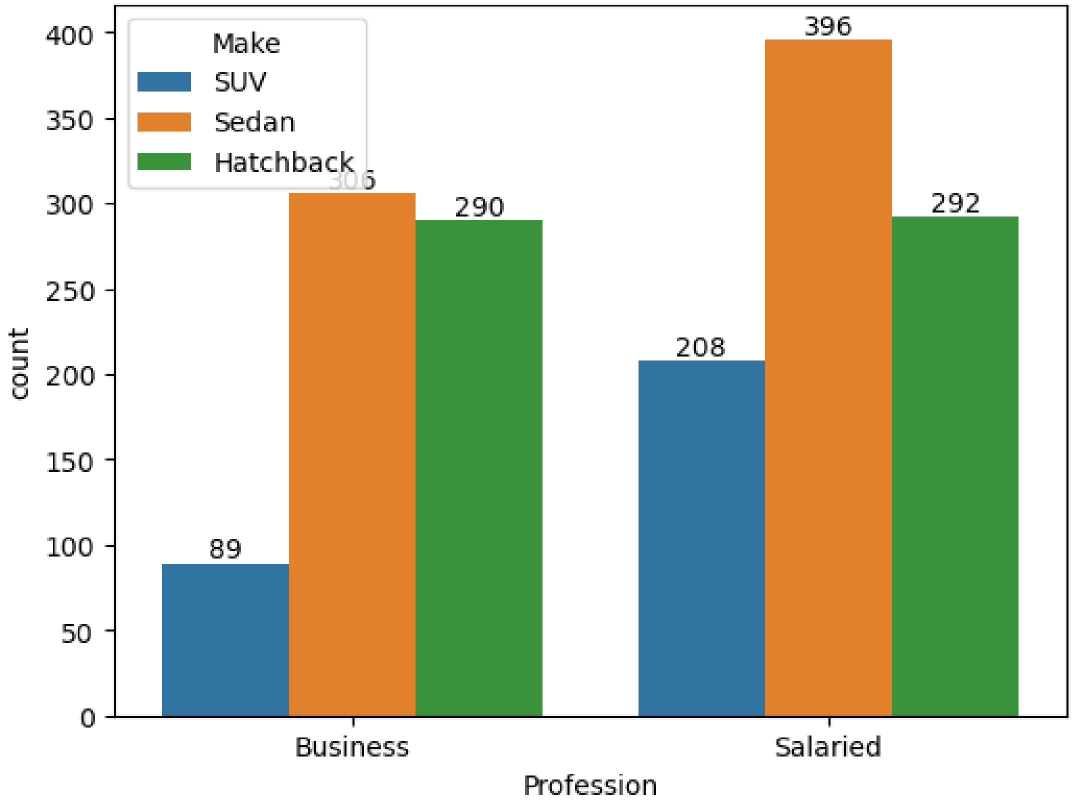
```
Salaried    896
Business    685
Name: Profession, dtype: int64
```

Proportion of Hatchbacks purchased=0.32

Proportion of SUV purchased=0.23

Proportion of Sedan purchased=0.44

Figure 23:- Count plot of Profession vs Make



From the above graph visualization we have observed that salaried person is more likely to buy a sedan.Hence statement made by Ned Stark 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.

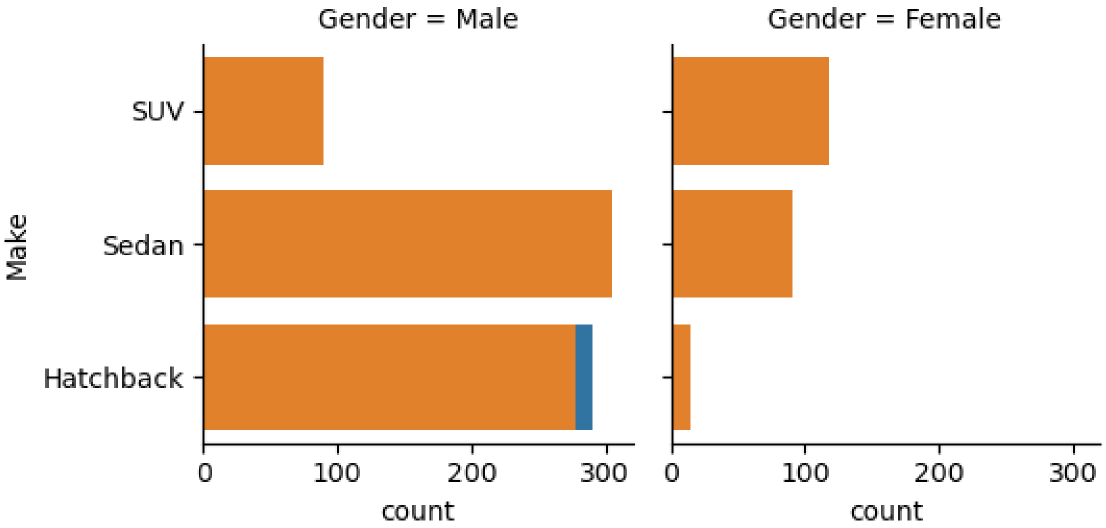
Analysing the Total cars purchased by salaried male customer are as follows:

Proportion of Hatchbacks purchased=0.41

Proportion of SUV purchased=0.13

Proportion of Sedan purchased=0.45

Figure-24 Profession vs Make for Male & Profession vs Make for Female



It is observed that salaried male prefers Sedan. Hence statement made by Sheldon Cooper 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.

F1) Gender

Table : 7 Gender-Mean and Median

Gender	**Mean**	**Median**
Male	32416	29000.0
Female	47705.1	49000

From the above calculation it is clear that females are more likely to buy than males

F2) Personal_loan

Table :8 Personal Loan -Mean and Median

Personal_loan	**Mean**	**Median**
No	36742.712294	32000

Yes	34457	31000

It is observed from the above table that the purchase made by customers who have personal loan is slightly higher

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

Table : 9 Partner Working Mean and Median

Partner_working	**Mean**	**Median**
No	36000	31000
Yes	35267	31000

From the above calculation we have observed that mean and median value for partner working is same thus we can say that partner working has no impact on purchasing car

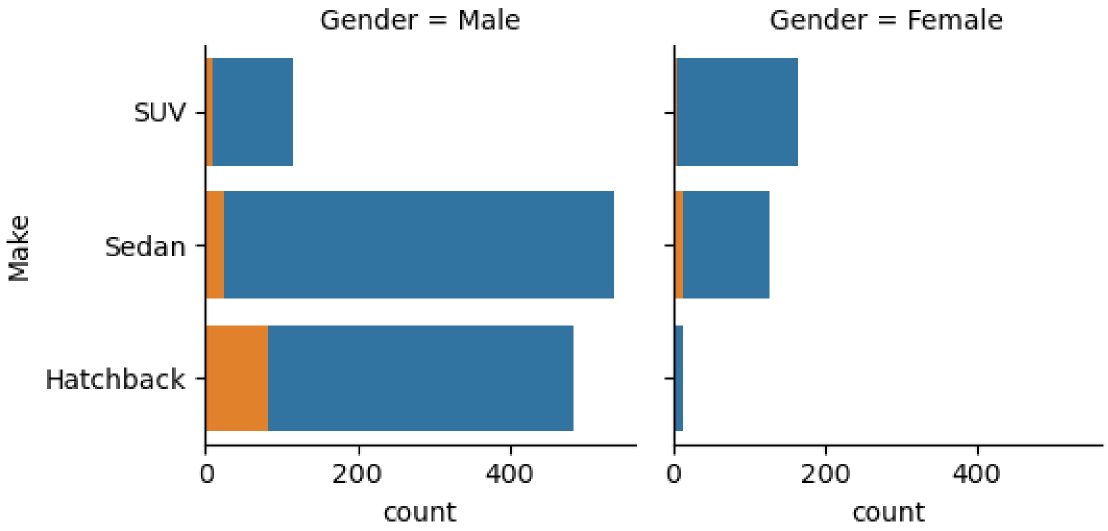
H.) The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use the Gender and Marital_status - fields to arrive at groups with similar purchase history.

Table:- 10 Analysis Gender and Marital_status

```
 Marital_status  Gender  Make
Married         Male    Sedan        537
               Hatchback       484
     Female  SUV          166
          Sedan        127
     Male    SUV          115
```

```
Single          Male    Hatchback     83
          Sedan           24
Married         Female  Hatchback     14
Single          Female  Sedan         14
        Male    SUV            9
        Female  SUV            7
            Hatchback      1
              dtype: int64
```

Figure-25 Marital Status vs Make for Male & Marital Status vs Make for Male



From the above visualization we have observed:

Married: Female- SUV

Married: Male - sedan

Single: Female - sedan

Single: Male- hatchback

Problem 2

*****Framing An Analytics Problem*** Analyse the dataset and list down the top 5 important variables, along with the business justifications.**

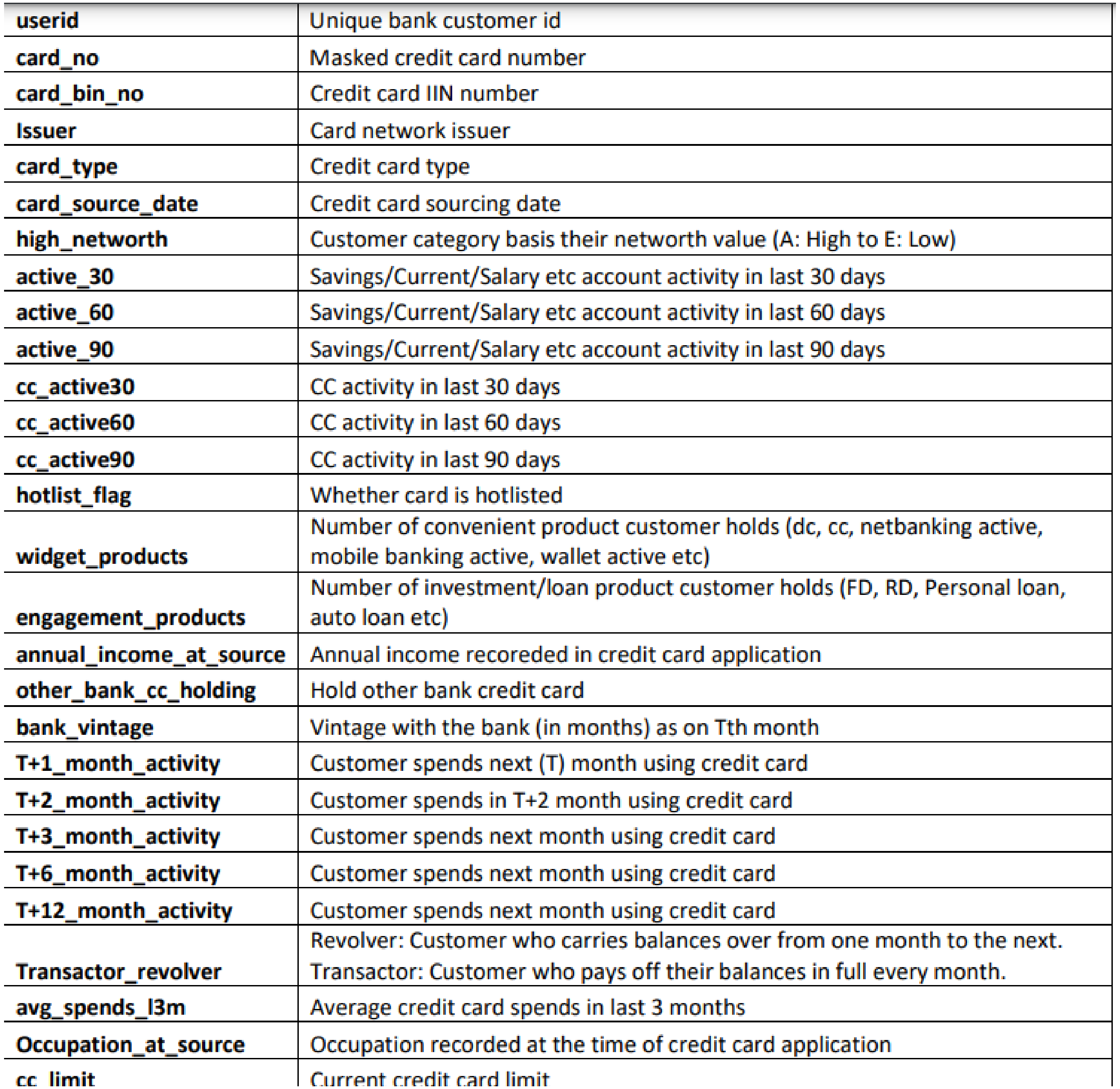
A bank can generate revenue in a variety of ways, such as charging interest, transaction fees and financial advice. Interest charged on the capital that the bank lends out to customers has historically been the most significant method of revenue generation. The bank earns profits from the difference 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.)

Question: (Analyze the dataset and list down the top 5 important variables, along with the business justifications.

Data dictionary:



userid	Unique bank customer id
card_no	Masked credit card number
card_bin_no	Credit card IIN number
Issuer	Card network issuer
card_type	Credit card type
card_source_date	Credit card sourcing date
high_networth	Customer category basis their networth value (A: High to E: Low)
active_30	Savings/Current/Salary etc account activity in last 30 days
active_60	Savings/Current/Salary etc account activity in last 60 days
active_90	Savings/Current/Salary etc account activity in last 90 days
cc_active30	CC activity in last 30 days
cc_active60	CC activity in last 60 days
cc_active90	CC activity in last 90 days
hotlist_flag	Whether card is hotlisted
widget_products	Number of convenient product customer holds (dc, cc, netbanking active, mobile banking active, wallet active etc)
engagement_products	Number of investment/loan product customer holds (FD, RD, Personal loan, auto loan etc)
annual_income_at_source	Annual income recoreded in credit card application
other_bank_cc_holding	Hold other bank credit card
bank_vintage	Vintage with the bank (in months) as on Tth month
T+1_month_activity	Customer spends next (T) month using credit card
T+2_month_activity	Customer spends in T+2 month using credit card
T+3_month_activity	Customer spends next month using credit card
T+6_month_activity	Customer spends next month using credit card
T+12_month_activity	Customer spends next month using credit card
Transactor_revolver	Revolver: Customer who carries balances over from one month to the next. Transactor: Customer who pays off their balances in full every month.
avg_spends_l3m	Average credit card spends in last 3 months
Occupation_at_source	Occupation recorded at the time of credit card application
cc_limit	Current credit card limit

All above data has been recorded as on T th month excluding T+1_month_activity, T+2_month_activity, T+3_month_activity, T+6_month_activity, T+12_month_activity

Top 5 important variables, along with the business justifications are as follows:-

1) **cc_limit**

Credit card limit is basically a Risk Management practice used by the banks to reduce the number of bad loans. This is calculated based on the customer's income, their CIBIL score etc.

2) **avg_spends_l3m**

It helps in identify how frequently customer use the credit card and also identify if the customer faces any issue

3) **cc_active30**

CC_Active30 provides the information regarding customer's credit card usage frequency. If the credit card is not used frequently, bank can reach out to the customer with new deals and offers. Also if customer is facing issues in using the credit card then they can help the customer with their concerns

4) **annual_income_at_source**

Annual income provides an insight into the purchasing capacity of the customer and is a very crucial information. When making decisions related to risks involved, offers to send to a customer, loan limit for the customer etc. correct information related to the Annual income can make a big difference

5) **T+12_month_activity**

It play very crucial role for bank. It helps bank to identify the areas where customer is more interested to use credit cards and the areas where customer uses less credit card. With this information bank can focus on areas where credit card is less use and can attract customers by giving various offers etc to increase its profitability

2)