SMDM PROJECT REPORT

DSBA

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.

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 given dataset Austo Motor Company, we import numpy, pandas, matplotlib, Seaborn in jupyter notebook, df.head (), df.tail (), df.describe ()

In [4]:
```
print("First few rows of the dataset:")
df.head()
```

```
First few rows of the dataset:
```

Out[4]:

	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

In [10]:
```
print("\nStatistical summary of numerical columns:")
df.describe()
```

```
Statistical summary of numerical columns:
```

Out[10]:

	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

In [12]:
```
print("\nShape of the dataset:")
print(df.shape)
```

```
Shape of the dataset:
(1581, 14)
```

```
In [13]: print("\nInformation about the dataset:")
         print(df.info())
```

```
Information about 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
None
```

Before Analysis the data we are going see the shape of the data, how many rows and columns are there, we are using the five dataset to see as head(), Shape() function for how many rows and columns are present in the dataset 1581 rows and 14 columns

We are going to see variables types using info () , where we have 5 numerical variables and 8 categorical data, there is no duplicate data here, null is there in gender and partner salary

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 value in variable, gender has 53 null count, partner salary has 106 null count values. Missing values can be imputed.

If larger records are missing we can drop the columns or rows. We describe () the function we see the mean, std, min,max

```
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
```

```
Statistical summary of numerical columns:
```

Out[10]:

	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

From the datasets we are described all the mean, median, standard, max, min value, here we get the whole summary of the data. The gender age between 22 to 54 are belong to working people, 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

Treating Anomalies

The value_counts () used to check the each categorical value in the data, help to check the issues.

```
Value Counts of 'Gender' column before treatment:
Male       1199
Female      327
Unknown      53
Femal         1
Femle         1
Name: Gender, dtype: int64
```

```
Value Counts of 'Education' column before treatment:
Post Graduate    985
Graduate         596
Name: Education, dtype: int64
```

```
Value Counts of 'Profession' column before treatment:
Salaried    896
Business    685
Name: Profession, dtype: int64
```

```
Value Counts of 'House_loan' column before treatment:
No     1054
Yes     527
Name: House_loan, dtype: int64
```

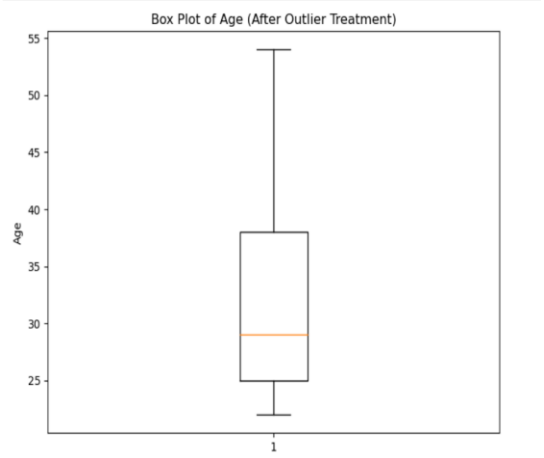
```
Value Counts of 'Personal_loan' column before treatment:
Yes    792
No     789
Name: Personal_loan, dtype: int64
```

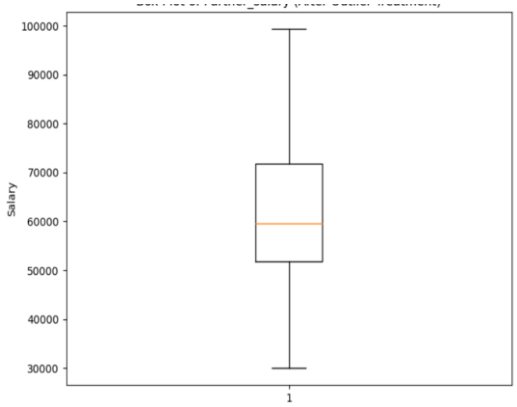
```
Value Counts of 'Make' column before treatment:
Sedan        702
Hatchback    582
SUV          297
Name: Make, dtype: int64
```

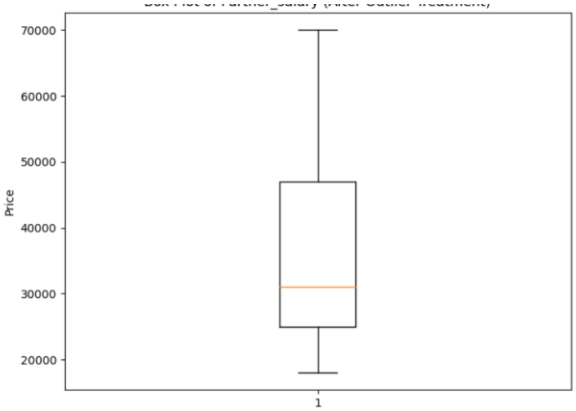
```
Value Counts of 'Marital_status' column before treatment:
Married    1443
Single      138
Name: Marital_status, dtype: int64
```

From the value counts we have from the data is Gender column has misspelled the ‘Femal’, and ‘Feml’, the rest of the categorical columns free from the issues.

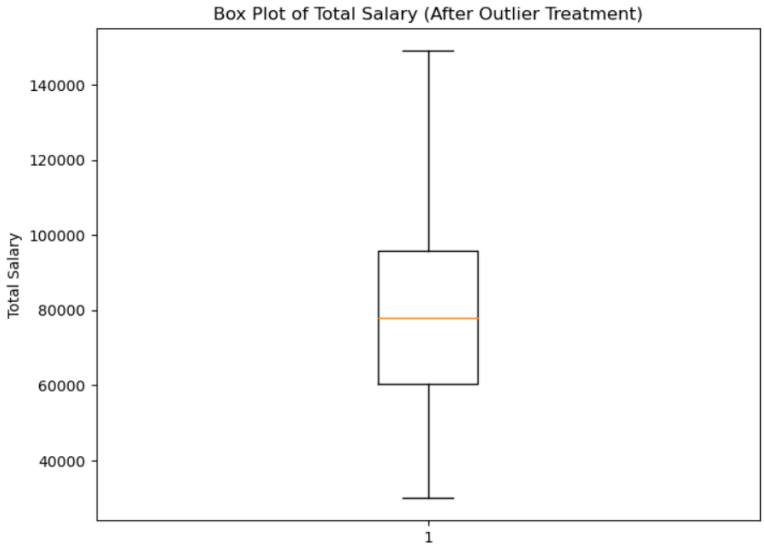
We can impute the data “Femal” and “Feml” in to “Female” in the data.







Here we can see the data, as far as we see the boxplot graph we don’t see any outliers in the age, partner_salary, salary, price



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

We have treated the outliers by using IQR rule

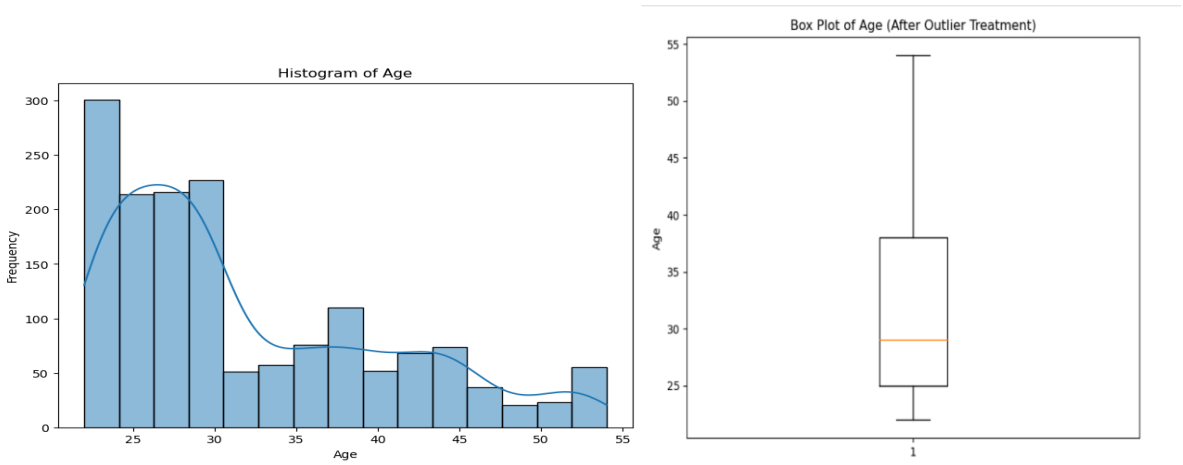
(Q1-1.5 * IQR) is used to treat the lower value

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

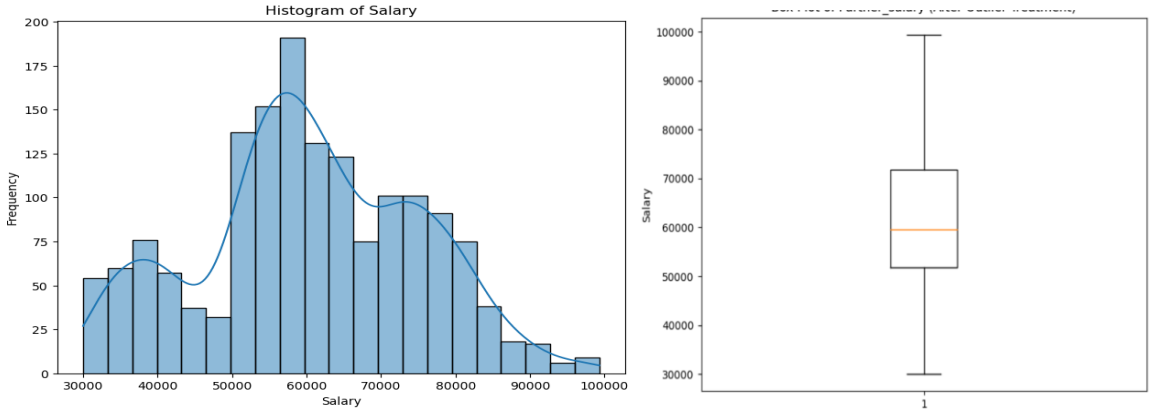
Using this formula , The data has been imputed

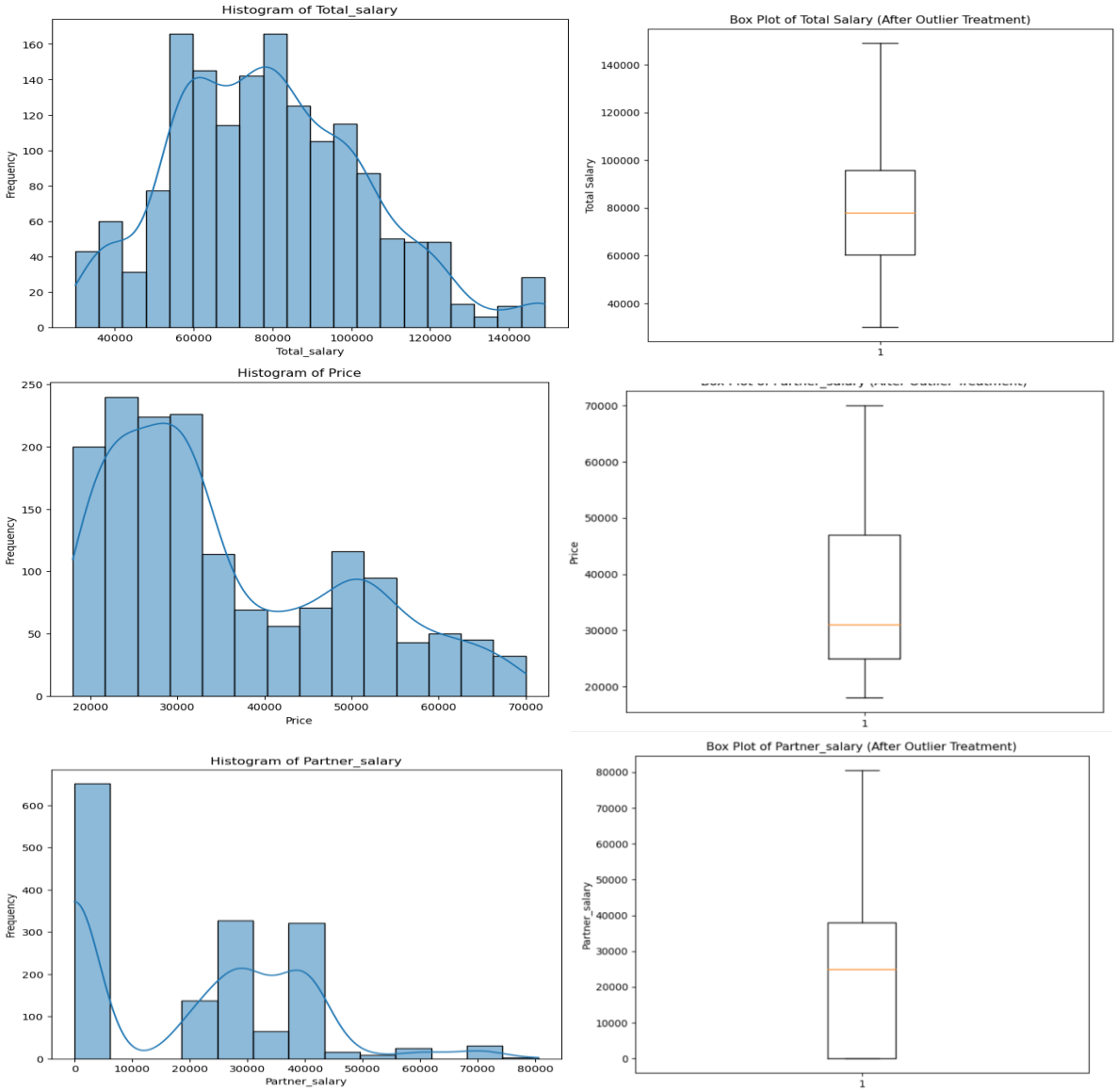
C. Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.

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



For better Visualization we are going use Histogram and Boxplot for better understanding The Graph show the output from the preloaded data





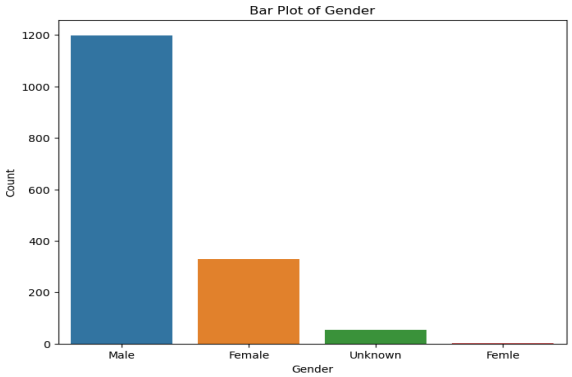
Age is multi distribution from the 50,000 to 80,000.

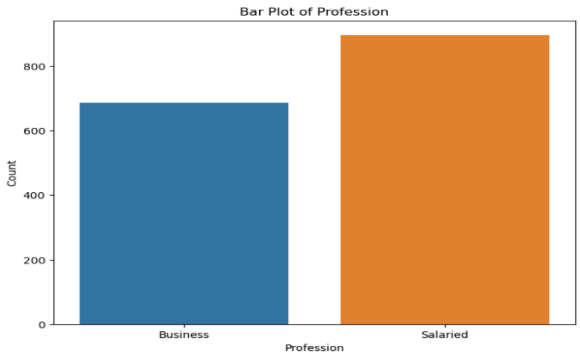
Salary has bulk distribution from the 50,000 to 70,000. Total salary has some outliers in this distribution, which can be imputed for analysis

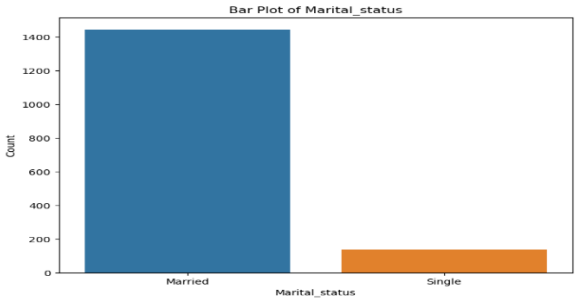
Price seems to be positive skew of 0.74

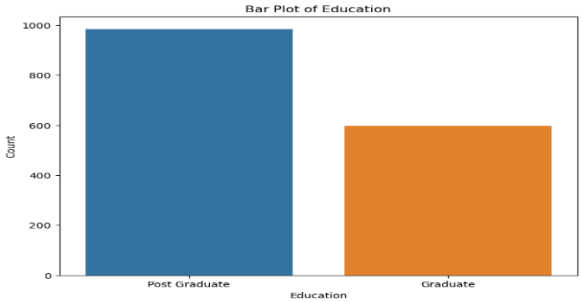
From the above items there is not any normal distribution in this data

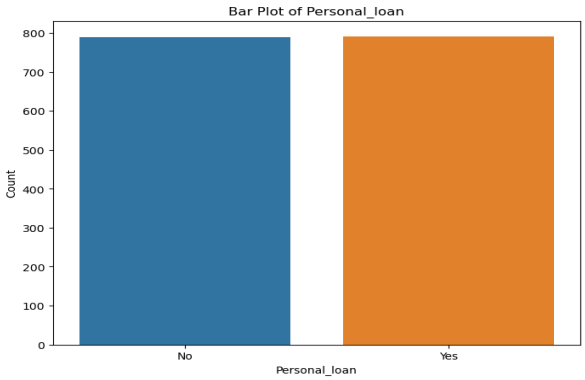
Univariate analysis for Categorical value

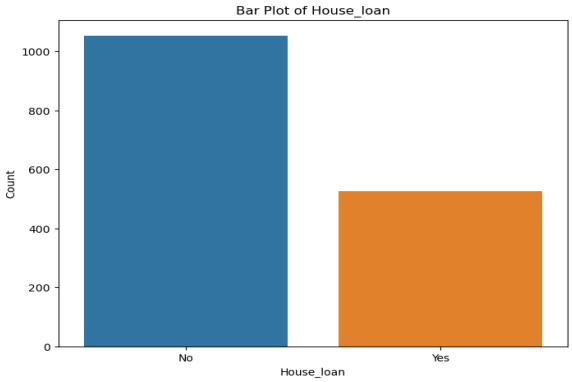


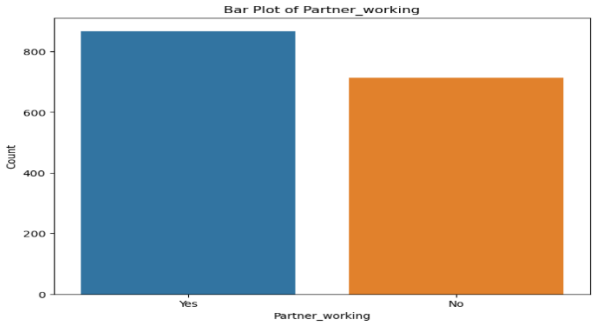


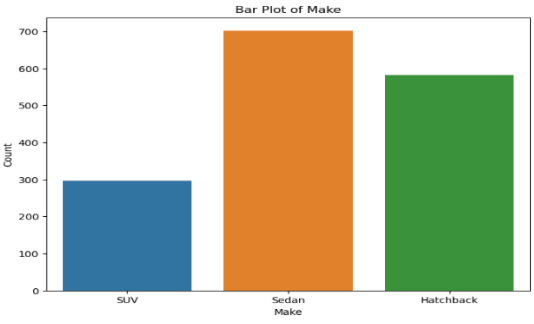












Working gender is Male is more than Female.

Salaried people are larger number than business people

Education wise post-graduation people are higher than graduate

Marital status has more married people than single people

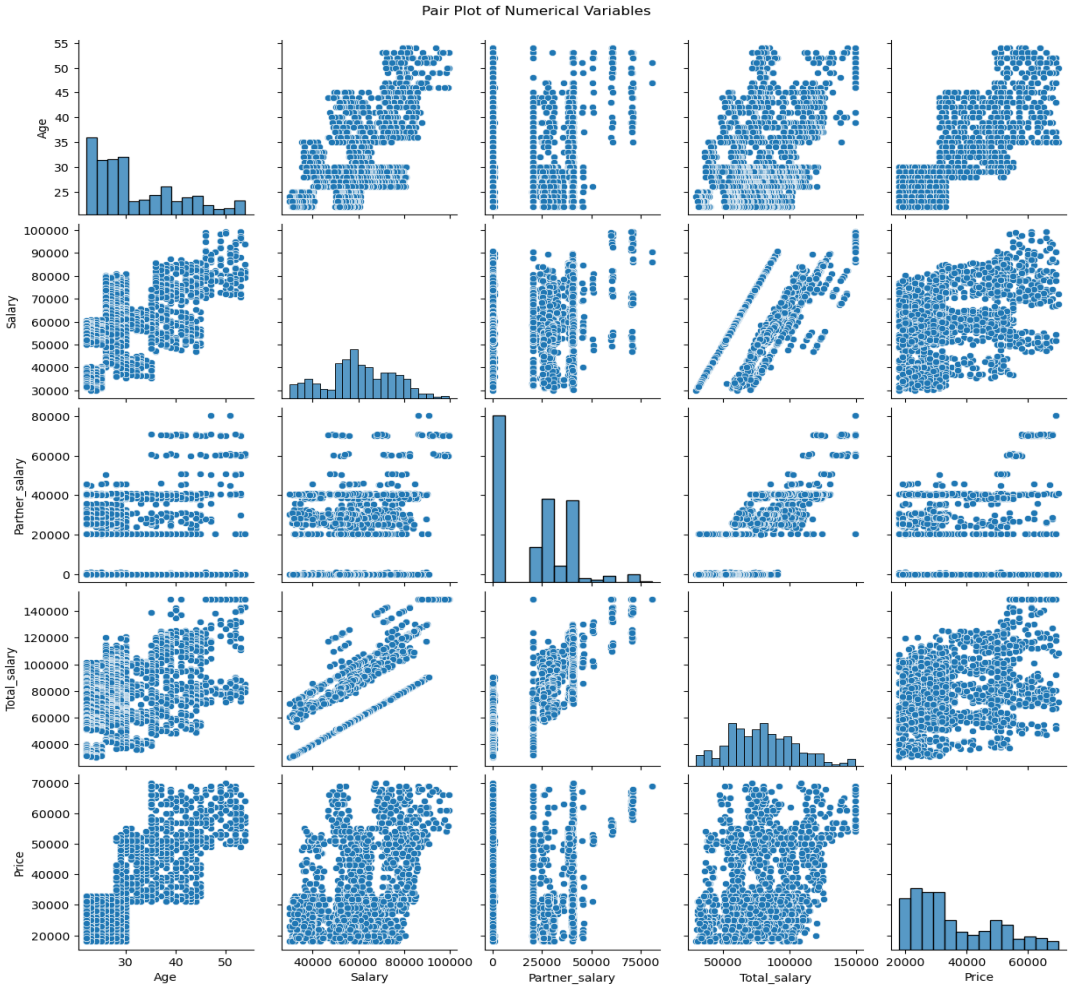
Personal loan doesn’t show much difference

House loan is less in number

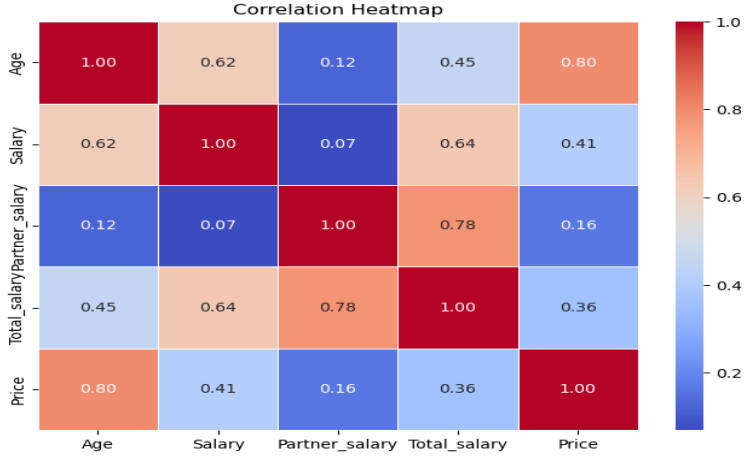
Partner working is more working people are there

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 the complete data



High correlation between the Total_salary and Partner_salary as shown in heat map Price and age is also highly correlated

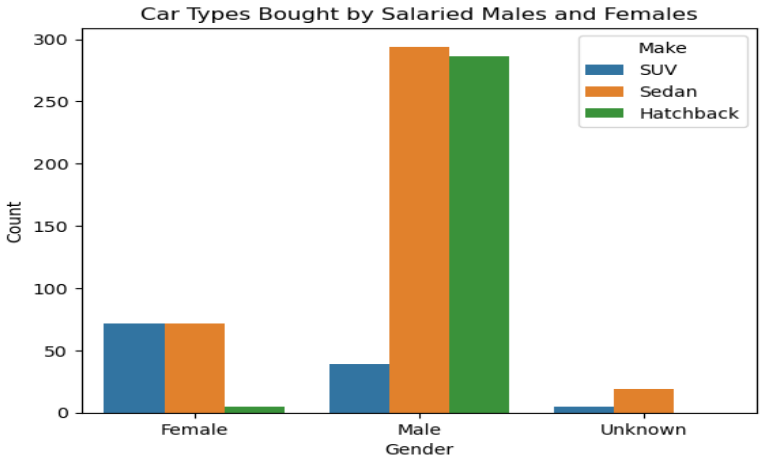
E. Employees working on the existing marketing campaign have made the following remarks. Based on the data and your analysis state whether you agree or disagree with their observations. Justify your answer Based on the data available.

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 Sale.

E1.



The proportion are SUV is bought by more female than man

Female bought SUV is 0.707317

Male bought SUV is 0.234362

Steve roger statement is False

E2.

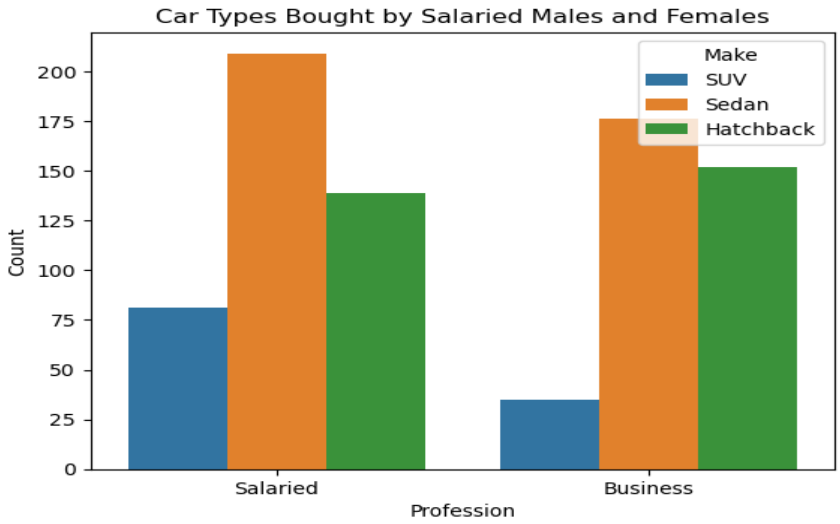
Proportion of salary people bought Hatchback is 0.265152

Proportion of salary people bought SUV is 0.314394

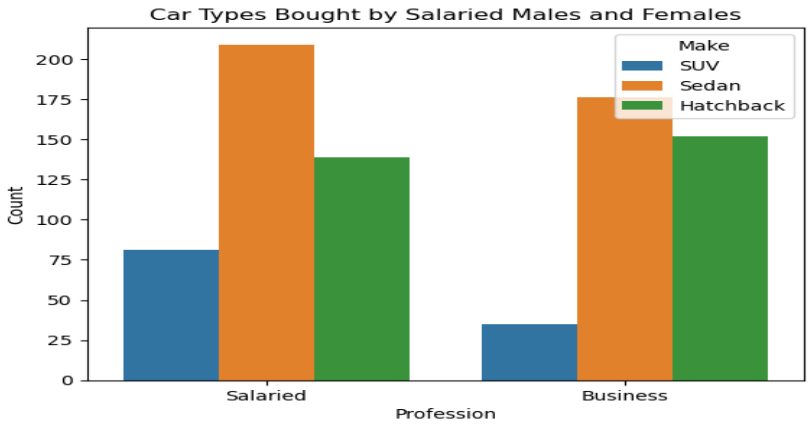
Proportion of salary people bought Sedan is 0.420455

Ned stark statement is true

For visualization we are using countplot, x parameter is profession, hue is Make



E3.



From the above graph and value, we can understand salary man only prefer Sedan than SUV Sheldon cooper 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

F2) Personal_loan

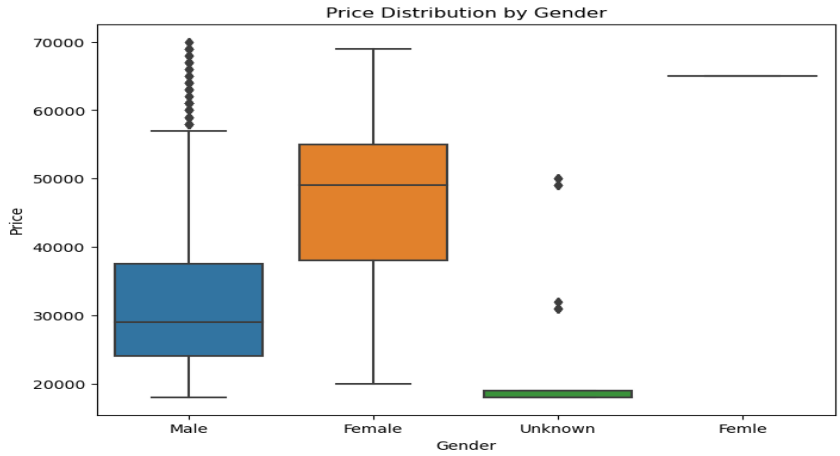
F1.

Female has spent 47611 than male

Male has spent 32867 in the cars

Hence Female bought more cars

Mean value	Median value
Female value is 47652	Female value is 49000
Male value is 32817	Male value is 29000



F2.

We are going to see the Personal_loan spent on purchasing automobiles

People purchased on Personal_loan

Mean value of Personal_loan

Personal loan Yes = 34457

Personal loan No = 36742

Median value of Personal_loan

Personal loan Yes = 31000

Personal loan No = 32000

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

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

Partner_working yes = 34457

Partner_working No = 36742

Median price of the Partner_working

Partner_working yes =31000

Partner_working no=31000

Maximum of Partner_working yes/no =70000

The data does not show much difference in the partner working

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.

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

```
Pivot Table (Grouped by Gender and Marital_status with Make Counts):
Make                    Hatchback  SUV  Sedan
Gender  Marital_status
Female  Married                14  165    127
        Single                  1    7     14
Femle   Married                 0    1      0
Male    Married               484  111    493
        Single                 81    7     23
Unknown Married                 0    4     44
        Single                  2    2      1
```

We can use the crosstab function in the to merge the data, to get the specific function or answer and analysis this data, from the above data most purchased car is SUV.

We have to treat the outliers in data, for better analysis.

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.)

Solution

These are variable given for analysis

card_type
active_30
active_60
active_90
cc_active30
annual_income_at_source
T+1_month_activity
T+2_month_activity
T+3_month_activity
T+6_month_activity
T+12_month_activity
avg_spends_l3m
Occupation_at_source
cc_limit

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

Cc_active is used to check the last active user of the card, how many time they are used

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

T+6 and T+12 should be excluded from details, because many times card used for transactions ,

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

cc_active30
annual_income_at_source
T+1_month_activity
T+2_month_activity
T+3_month_activity
avg_spends_l3m
cc_limit

Cc_active 30 is used to check the last active user of the card, how many time they are used and transactions 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 credit help to focus on the right card can be suggested for the customer to changes variety of credit card