SMDM PROJECT (Coded)

R. SUKANYA

Date of Submission: 8th October, 2023

TABLE OF CONTENTS

I. **Data Dictionaries of the two problems**……………………………..........07

II. **Problem 1 – Description**…………………………………………….....09

1. **Data Overview**

1.1 Import the libraries……………………………………………………..09

1.2 Load the Data…………………………………………………………09

1.3 Check the structure of the data………………………………………10

1.3.1 Number of rows and columns in the dataset…………………..10

1.4 Check the types of the data……………………………………………..10

1.5 Check the statistical summary…………………………………...….10

1.6 **Check and treat data irregularities**

1.6.1 Check for duplicates…………………………………………12

1.6.2 Check for missing values……………….……………………12

1.6.3 Check for Bad Data……...……………………………………...13

1.7 **Treatment of missing values, bad data and duplicates**…………..13

1.7.1 Treat missing values……………………………………...13

1.7.2 Treat bad data…………………………………………………14

1.7.3 Treat anomalies………………………………………………....15

1.8 **Observations and Insights**………………………………………15

2. **Data Visualization**

2.1 **Univariate Analysis**……………………………………………16

2.1.1 Numerical Variables……….……………………………...16

2.1.2 Treating outliers…………….………………………………19

2.1.3 Categorical data…………………………………………..... 20

2.2 **Bivariate Analysis**……………………………………………………..25

2.2.1 Numerical vs Numerical…………………………………….....25

2.2.2 Pairplot for Numerical vs Numerical data……………………..28

2.2.3 Correlation between all the numerical variables…….....………30

2.2.4 Categorical vs Numerical……………………………………...30

2.3 Relationship between Categorical and Numerical variables – Insight ………………………………………………………………………..36

3. **Key Questions**……….………………………………...…………..37

3.1 Do men tend to prefer SUVs more compared to women?...................37

3.2 What is the likelihood of a salaried person buying a Sedan?.............38

3.3 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?.................38

3.4 How does the the amount spent on purchasing automobiles vary by gender?..39

3.5 How much money was spent on purchasing automobiles by individuals who took a personal loan?...39

3.6 How does having a working partner influence the purchase of higher-priced cars?……………………………………………..…………40

4 **Actionable Insights**……………………………………………...…40

5 **Business Recommendations**…..……………………………….…42

6 **Problem 2**……………………………………………………..……...43

6.1 Loading the libraries…………………………………………….43

6.2 Initial questions to be raised…………………………………….43

6.3 Top 5 variables…………………………………………………..44

6.4 Business Justification......……………………………………....45

6.5 Conclusion……………………………………………………....49

LIST OF FIGURES

Univariate Variables

Numerical Variables

Fig 2.1.1 a) Histogram/Boxplot of Distribution of Age……………….……..16

Fig 2.1.1 b) Histogram/Boxplot of No of Dependents…..……………………17

Fig 2.1.1 c) Histogram/Boxplot of Distribution of Salary, Partner_salary…....17

Fig 2.1.1 d) Histogram/Boxplot of Distribution of Total_salary………………18

Fig 2.1.1 e) Histogram/Boxplot of Distribution of Price………………………19

Fig 2.1.2 Treating Outliers……………………………………………......20

Categorical Variables

Fig 2.1.3 a) Countplot - Distribution of Gender…………………………....20

Fig 2.1.3 b) Countplot - Distribution of Profession…………………………21

Fig 2.1.3 c) Countplot - Distribution of Marital_status………………………..22

Fig 2.1.3 d) Countplot - Distribution of Education……………………………22

Fig 2.1.3 e) Countplot - Distribution of Personal_loan………………………..23

Fig 2.1.3 f) Countplot - Distribution of House_loan……………………......23

Fig 2.1.3 g) Countplot - Distribution of Partner_working…………………......24

Fig 2.1.3 h) Countplot - Distribution of Make…………………………………24

Bivariate Analysis

Numerical Vs Numerical

Fig 2.2.1 a) Scatter Plot - Age vs No_of_Dependents………………………....25

Fig 2.2.1 b) Scatter Plot - Age vs Salary………………………………………25

Fig 2.2.1 c) Scatter Plot - Age vs Partner_salary……………………………....26

Fig 2.2.1 d) Scatter Plot - Age vs Total_salary…………………………………...26

Fig 2.2.1 e) Scatter Plot - Age vs Price…………………………..…………27

Fig 2.2.1 f) Scatter Plot - Salary vs Total Salary………………………………27

Fig 2.2.1 g) Scatter Plot – Salary vs Price……………………………………..28

Fig 2.2.2 Pairplot of all numerical variables…………………………………..29

Fig 2.2.3 Heatmap of all numerical variables…………………………………30

Categorical Vs Numerical

Fig 2.2.4 Boxplot of Gender vs Age……………………………………………...30

Fig 2.2.4 a) Boxplot of Gender vs all numerical variables…………………….31

Fig 2.2.4 b) Boxplot of Profession vs all numerical variables………………...32

Fig 2.2.4 c) Boxplot of Marital_status vs all numerical variables……………..33

Fig 2.2.4 d) Boxplot of Education vs all numerical variables………………...33

Fig 2.2.4 e) Boxplot of Personal_loan vs all numerical variables……………..34

Fig 2.2.4 f) Boxplot of House_loan vs all numerical variables………………..34

Fig 2.2.4 g) Boxplot of Partner_working vs all numerical variables…………..35

Fig 2.2.4 h) Boxplot of Make vs all numerical variables……….……………..35

Key Questions

Fig 3.1 Countplot of Gender…………………………………………………...38

Fig 3.3 Countplot of Make……………………………………………………..38

Fig 3.6 Countplot of Partner_working…………………………………………40

Problem 2

Top 5 variables

Fig 6.4.1 Scatterplot annual_income_at_source vs avg_spends/cc_limit…….45

Fig 6.4.2 Countplot of otherbank_cc_holding vs Occupation/card_tupe……..46

Fig 6.4.3 Boxplot of avg_spends_l3m……………………………………… ..47

Fig 6.4.4 Heatmap and Countplot of Transactor_revolver……………………48

Fig 6.4.5 Countplot of card_type………………………………………………49

LIST OF TABLES

Table 1: Dataset austo_automobile.csv……………………………………….07

Table 2: Dataset godigt_cc_data…………………………………………….08

Table 3: Statistical summary of numerical variables…………………………10

Table 4: Statistical summary of categorical variables…………………………11

Table 5: Key Question 3.2…………………………………………………..38

I. Data Dictionaries for the two problems

Austo_automobile.csv

Data Description

- **age:** The age of the individual in years.
- **gender**: The gender of the individual, categorized as male or female.
- **profession**: The occupation or profession of the individual.
- **marital_status**: The marital status of the individual, such as married &, single
- **education**: The educational qualification of the individual Graduate and Post Graduate
- **no_of_dependents**: The number of dependents (e.g., children, elderly parents) that the individual supports financially.
- **personal_loan**: A binary variable indicating whether the individual has taken a personal loan "Yes" or "No"
- **house_loan**: A binary variable indicating whether the individual has taken a housing loan "Yes" or "No"
- **partner_working**: A binary variable indicating whether the individual's partner is employed "Yes" or "No"
- **salary**: The individual's salary or income.
- **partner_salary**: The salary or income of the individual's partner, if applicable.
- **Total_salary**: The total combined salary of the individual and their partner (if applicable).
- **price**: The price of a product or service.
- **make**: The type of automobile

Table 1

Godigt_cc_data

Data Description

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_data - Credit card sourcing date

high_networth - Customer category based on their net-worth 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 - Credit Card activity in the last 30 days

cc_active60 - Credit Card activity in the last 60 days

cc_active90 - Credit Card activity in the last 90 days

hotlist_flag - Whether card is hot-listed(Any problem noted on the card)

widget_products - Number of convenience products customer holds (dc, cc, net-banking active, mobile banking active, wallet active, etc.)

engagement_products - Number of investment/loan products the customer holds (FD, RD, Personal loan, auto loan)

annual_income_at_source - Annual income recorded in the credit card application

other_bank_cc_holding - Whether the customer holds another bank credit card

bank_vintage - Vintage with the bank (in months) as on Tthmonth

T+1_month_activity - Whether customer uses credit card in T+1 month (future)

T+2_month_activity - Whether customer uses credit card in T+2 month (future)

T+3_month_activity - Whether customer uses credit card in T+3 month (future)

T+6_month_activity - Whether customer uses credit card in T+6 month (future)

T+12_month_activity - Whether customer uses credit card in T+12 month (future)

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

Table 2

II. 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.

Objective

They want to analyze the data to get a fair idea about the demand of customers which will help them in enhancing their customer experience. Suppose you are a Data Scientist at the company and the Data Science team has shared some of the key questions that need to be answered. Perform the data analysis to find answers to these questions that will help the company to improve the business.

1. DATA OVERVIEW

1.1. Import the libraries:

Import the libraries for data manipulation and data visualization. Libraries like pandas, numpy, matplotlib and seaborn were imported. Check the versions of numpy, pandas and seaborn.

Output:

```
The version of Numpy is:  1.24.3
The version of Pandas is:  1.5.3
The version of Seaborn is:  0.12.2
```

1.2. Load the data:

The data austo_automobile.csv is loaded into the jupyter notebook so that data analysis can be done. Check the working directory and then load the appropriate .csv file.

1.3. Check the structure of the data:

Use the head and tail functions to observe the different fields in the dataset. There are 14 columns, namely, Age, Gender, Profession, Marital_status, Education, No_of_Dependents, Personal_loan, House_loan, Partner_working, Salary, Partner_salary, Total_salary, Price and Make.

1.3.1 Number of rows and columns in the dataset:

Use the shape function to determine the number of rows and columns in the dataset.

Output:

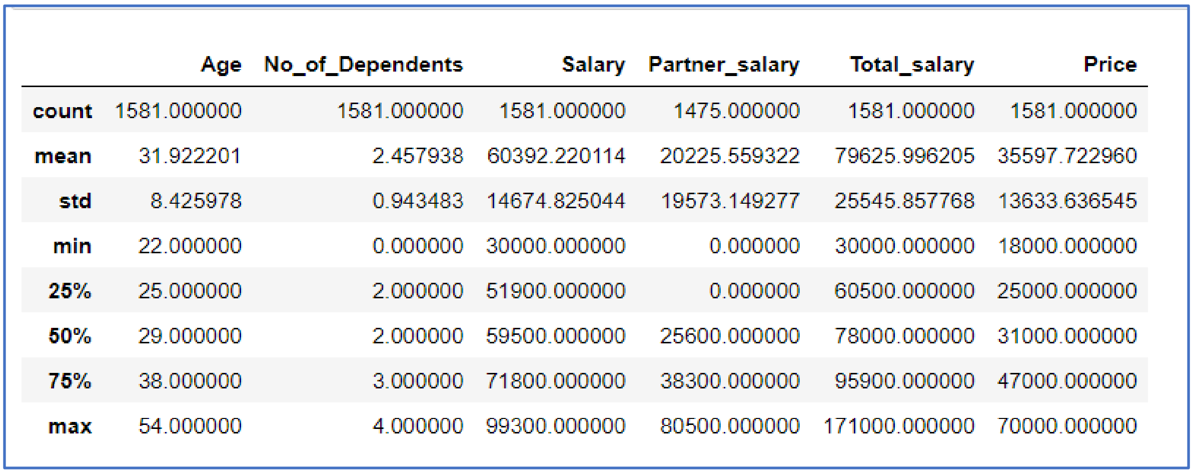
```
Number of rows in the dataset:  1581
Number of columns in the dataset:  14
```

1.4 Check the types of the data:

Use the info() function to get the datatypes of the columns/fields of the dataset. There are 5 fields with integer datatype, 8 fields with object datatype and 1 field with float datatype. That is, 8 fields are categorical in nature while 6 fields are numerical data.

1.5 Check the statistical summary:

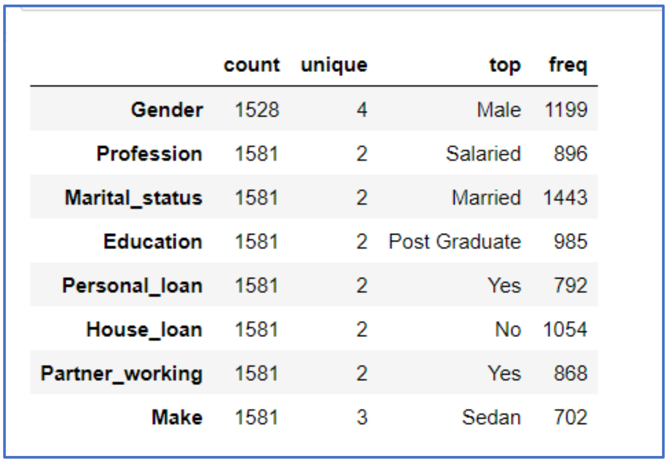
Use the describe() function to list out the statistical summary of numerical fields of the data.



	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
count	1581.000000	1581.000000	1581.000000	1475.000000	1581.000000	1581.000000
mean	31.922201	2.457938	60392.220114	20225.559322	79625.996205	35597.722960
std	8.425978	0.943483	14674.825044	19573.149277	25545.857768	13633.636545
min	22.000000	0.000000	30000.000000	0.000000	30000.000000	18000.000000
25%	25.000000	2.000000	51900.000000	0.000000	60500.000000	25000.000000
50%	29.000000	2.000000	59500.000000	25600.000000	78000.000000	31000.000000
75%	38.000000	3.000000	71800.000000	38300.000000	95900.000000	47000.000000
max	54.000000	4.000000	99300.000000	80500.000000	171000.000000	70000.000000

Table 3

Use the describe() function with the include=object to get the statistical summary of all the categorical fields.



	count	unique	top	freq
Gender	1528	4	Male	1199
Profession	1581	2	Salaried	896
Marital_status	1581	2	Married	1443
Education	1581	2	Post Graduate	985
Personal_loan	1581	2	Yes	792
House_loan	1581	2	No	1054
Partner_working	1581	2	Yes	868
Make	1581	3	Sedan	702

Table 4

Insights from Statistical summary of numerical and categorical fields:

- Age, No_of_Dependents, Salary, Partner_salary, Total_salary and Price are the numerical fields in the dataset.
- Note the mean, minimum and maximum value of all the numerical fields.
- For example, the mean age of people who want to buy a car is around 31 years while the minimum age is 22 years. The maximum age of people wanting to buy a car is 54 years.
- Similarly, the mean salary of a personal wanting to buy a car is 60,392 while the max salary is 99300 and the minimum is 30000.
- The average price of a car is 31000. The highest price of the car is 70000 and lowest is 18000.
- Gender, Profession, Marital_status, Education, Personal_loan, House_loan, Partner_working and Make are categorical in nature.
- Age, No_of_Dependents, Salary, Partner_salary, Total_salary and Price are numerical in nature.
- We observe that out of 1528 people who want to buy a car, 1199 are males and remaining females.

- Married people outweigh the unmarried people who want to buy a car.
- About 985 people out of a total of 1581 people are postgraduates.
- Salaried people wanting to buy a car add up to 896.
- 792 people with personal loan want to buy a car.
- Around 868 people's partners are also working.
- 702 out of a total of 1581 want to buy a sedan.

1.6 Check for and treat (if needed) data irregularities

1.6.1 Check for duplicates:

Check for duplicates using the duplicated() function. We found that the number of duplicate rows is 0.

Output:

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

1.6.2 Check for missing values:

Check for missing values using the isnull() function. Gender has 53 missing values while Partner_salary has 106 missing values.

Output:

```
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
```

1.6.3 Check for Bad data:

Check for bad data using unique() function for all the numerical and categorical fields. We see that the column Gender has bad dat a like nan, Femle and Femal. These values need to be cleaned.

Output:

Gender:

```
array(['Male', 'Femal', 'Female', nan, 'Femle'])
```

Insights for checking duplicates, missing values and bad data

- Dataset has 1581 rows and 14 columns.
- The dataset has float, integer and object datatypes.
- No duplicate values are there in the dataset.
- Gender has 53 missing values while Partner_salary has 106 missing values.
- The Gender column has bad data like nan, Femle and Femal. The Partner_salary field has nan values.

1.7 Treatment of missing values, bad data and duplicates:

1.7.1 Treat missing values:

We fill the missing values in Partner_salary using the mean value of the field Partner_salary using the fillna() function. Now, if we use the isnull() function to check for missing values in this column we don't find any missing value.

Output:

```
 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      0
Total_salary        0
Price               0
Make                0
```

We observe that all the 106 rows of missing values in the Partner_salary field have been filled with the mean value of Partner_salary.

1.7.2 Treat bad data:

On doing a value_counts() on the Gender column which has bad data, we see the following output:

Output:

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

We change the Femal and Femle column names to Female using the replace() function. We have replaced 'Femal' and 'Femle' with 'Female' field. Hence, 1 person each from Femle and Femal have been added to Female field. Now, the Female has 329 counts instead of 327 earlier. On doing a value_counts on the Gender field, we get the following output:

Output:

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

As the mode of the Gender column is male, we replace the missing values in Gender by Male.

Now, we see that there are no null values in any columns in the dataset.

Output:

```
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
```

1.7.3 Treat Anomalies

Find out if Partner_salary is 0 if Partner_working =No.
If the partner is not working and the partner salary is greater than 0, then it is an anomaly. We treat such an anomaly by filling the Partner_salary as 0 when Partner_working is No.

1.8 Observations and Insights

- In this way, we have treated the bad data, missing values and checked for any anomalies.
- The missing values in Gender has been replaced with the mode value of Gender, which is Male.
- The spelling mistakes in the Gender column like Femle or Femal have been replaced with the mode value of Gender.
- The Partner_salary column had 106 missing values too. The missing values of Partner_salary field were replaced with the mean value of the field.
- Thus, the data has been cleaned now and is ready for visualisation.

2.0 DATA VISUALIZATION

2.1 Univariate Analysis

2.1.1 Numerical Variables

a) Age

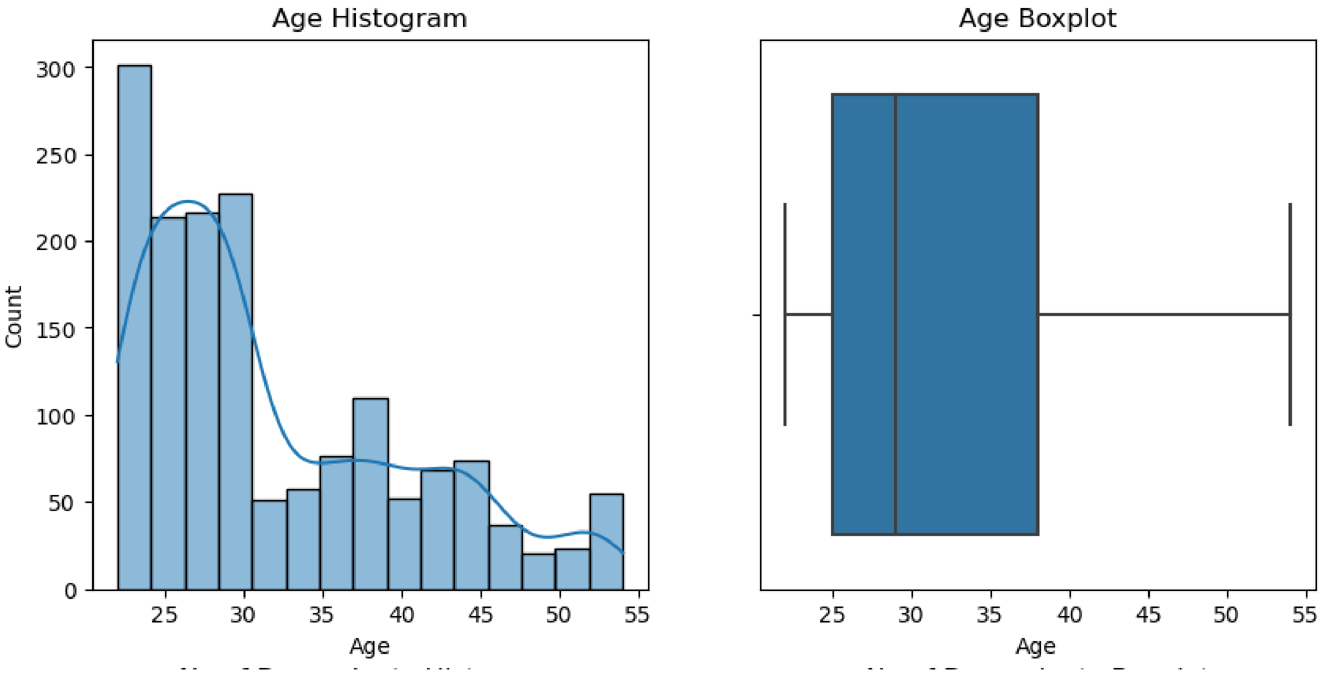


Fig 2.1.1 a)

<u>Insights:</u>

- Age ranges from 22 years to 54 years.
- 75% of the people are of age 38 years.
- Age distribution is right-skewed as seen from the count plot and the boxplot.
- Mean age is 32 years which is higher than the median value of 29 years, so the distribution is right skewed.

b) No. of Dependents

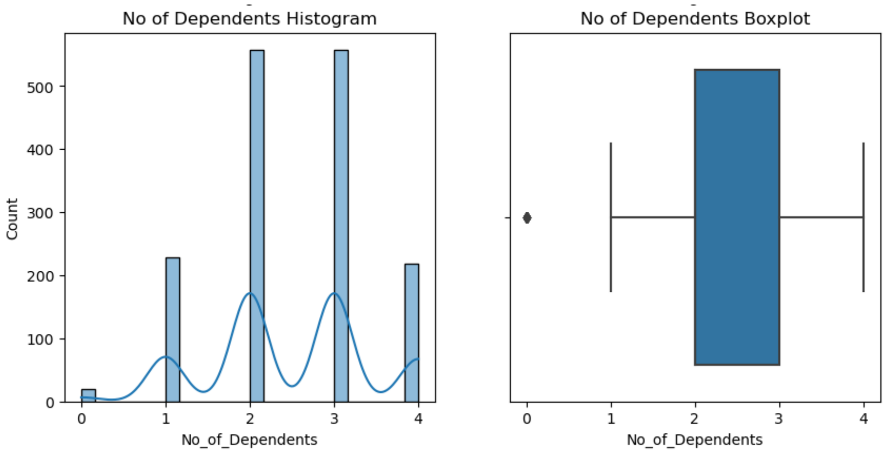


Fig 2.1.1 (b)

Insights:

- The No of Dependents distribution is normally distributed but there is an outlier.
- The number of dependents ranges from 0 to a maximum of 4.
- Mean and Median of No of dependents is almost same, hence the distribution is normally distributed.

c) Salary and Partner_salary:

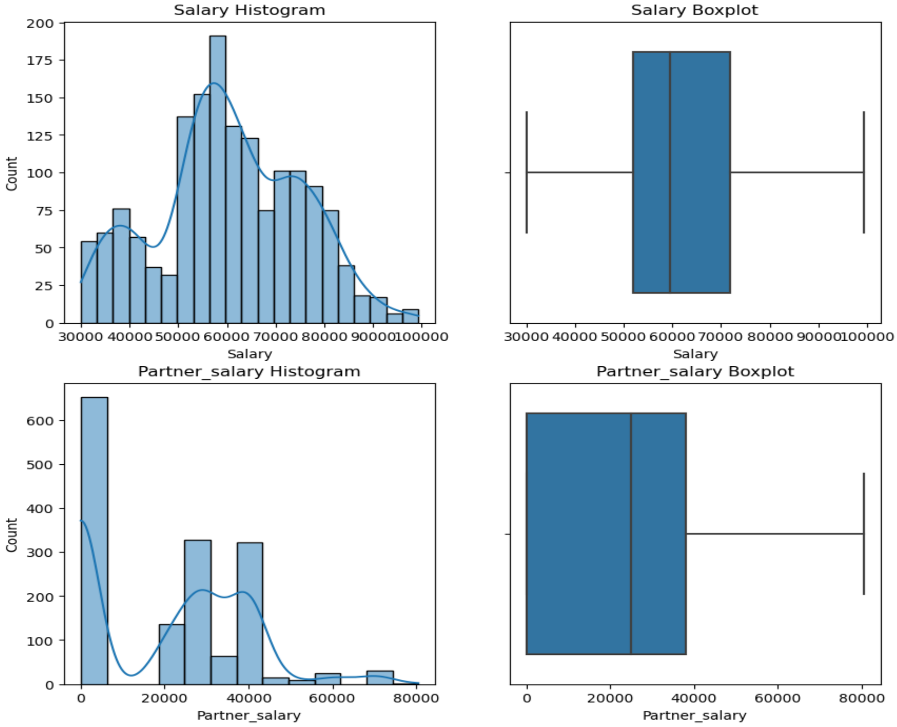


Fig 2.1.1 c)

Insights:

- The Salary field ranges from a minimum of 30000 to a maximum of 99300.
- 75% of the people have a salary of 71800.
- The mean salary is greater than the median salary, indicating a right skewed distribution.
- Partner salary has a minimum and maximum of 0 and 80500 respectively.
- The mean value of Partner salary is 20225 and a median value of 24900. As median is higher than mean, it indicates a left skewed distribution.
- 75% of the people have a partner salary of 38000.
- Salary field seems to be normally distributed with no outliers.
- Partner salary has the minimum and 25% starting at 0, hence there are no left whiskers. No outliers too.

d. Total_salary

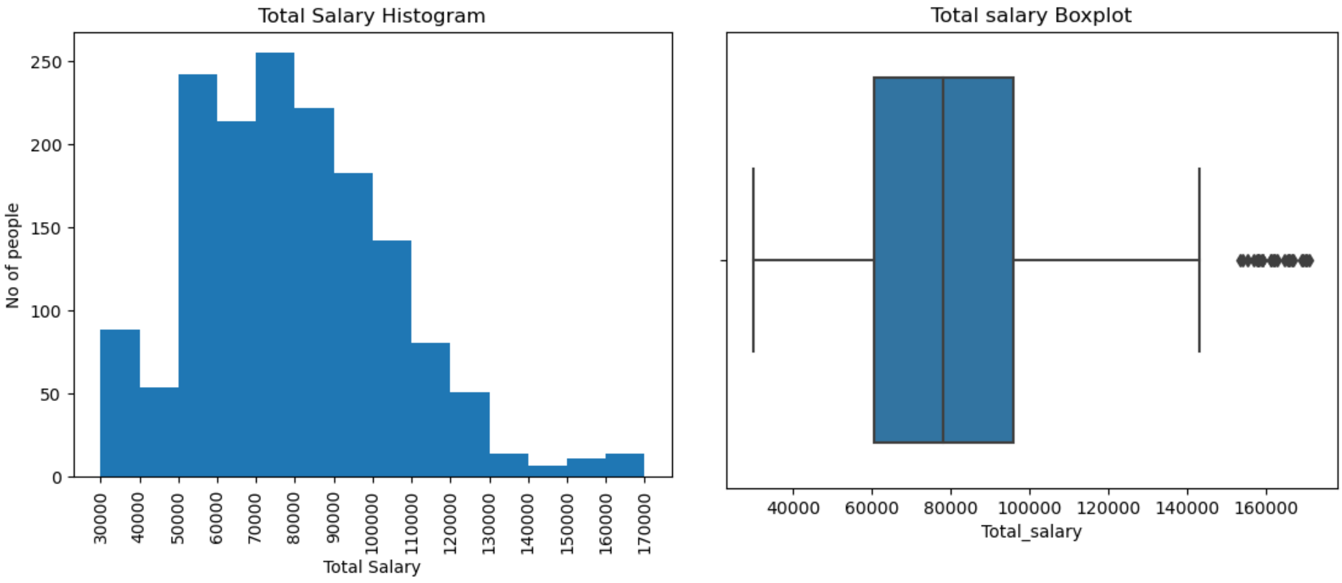


Fig 2.1.1 d)

Insights:

- The minimum salary is 30000 while the maximum salary is 171000.
- The median salary is 78000, while the mean is 79625.

- Both mean and median values are almost the same, hence looks to be normally distributed.
- The maximum salary is way higher than the 75% value, indicating outliers.
- The boxplot clearly shows the presence of outliers in the upper range.

e) Price

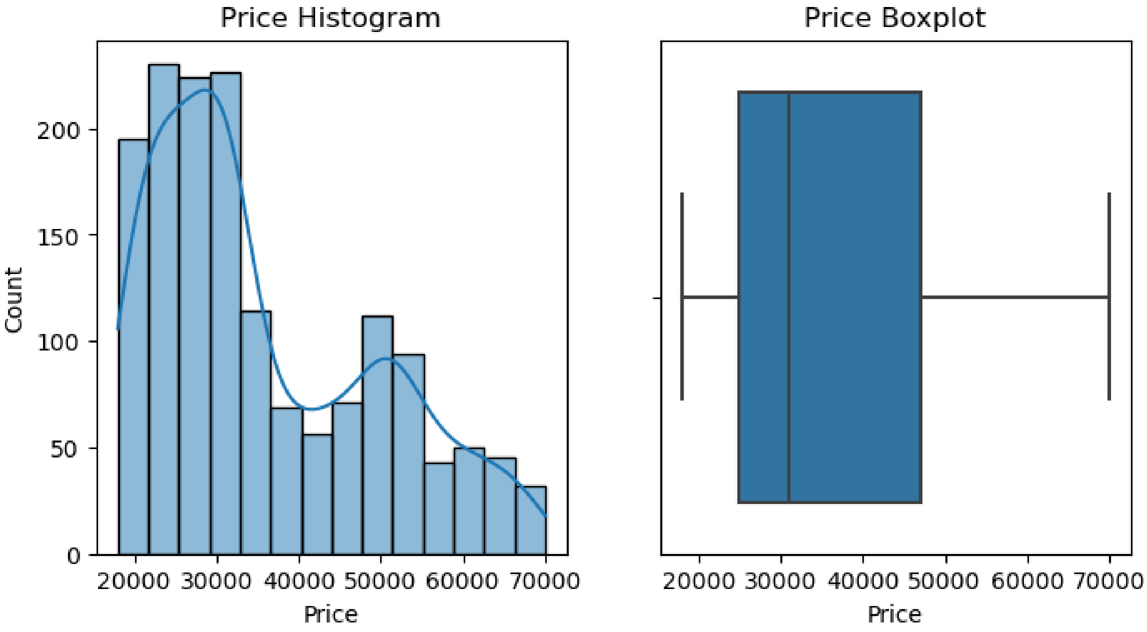


Fig 2.1.1 e)

Insights:

- Since Mean is greater than Median, the distribution is right-skewed as seen in the countplot as well as the boxplot.
- The price of cars ranges from a minimum of 18000 to a maximum of 70000.
- There are no outliers.

2.1.2 Treating Outliers

In order to treat outliers in No_of_Dependents, find the IQR.

The Inter Quartile Range = Q3 – Q1

Find the lower range which is equal to Q1 – 1.5* IQR and the upper range which is equal to Q3 + 1.5* IQR. Create arrays of Boolean values indicating the outlier rows. Remove the outliers by using the drop() function on the upper array and lower array.

After treating the outliers in No_of_Dependents, plot a boxplot to check if there are any outliers left.

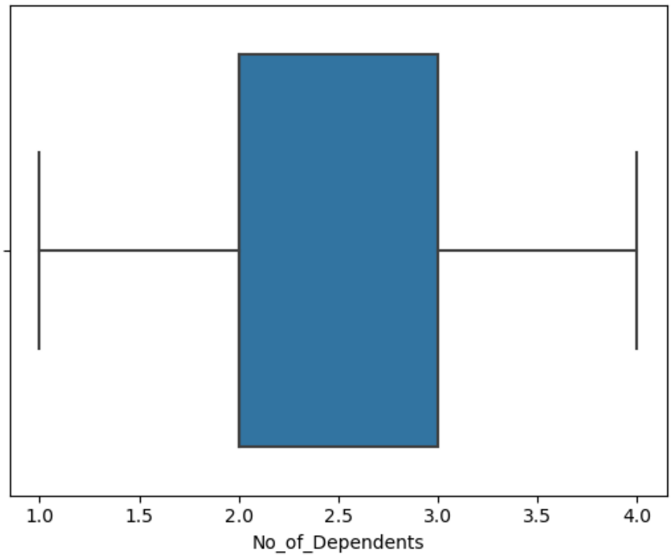


Fig 2.1.2

We observe that there are no outliers left in the column 'No_of_Dependents'.

2.1.3 Categorical Data

a) Gender

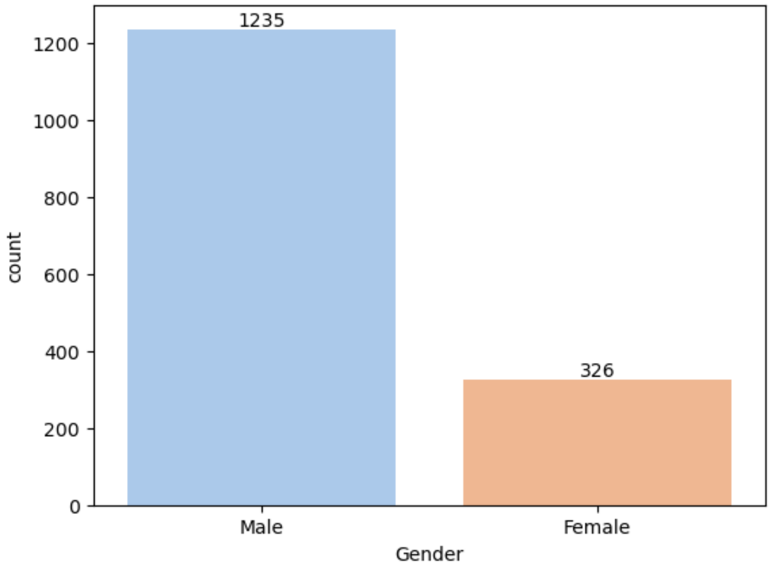


Fig 2.1.3 a)

Insights:

- About 79% of the people are males who buy a car while 20.8% of the people are females who buy a car.
- In numbers, 1252 people who buy a car are males while 329 are females.

b) Profession:

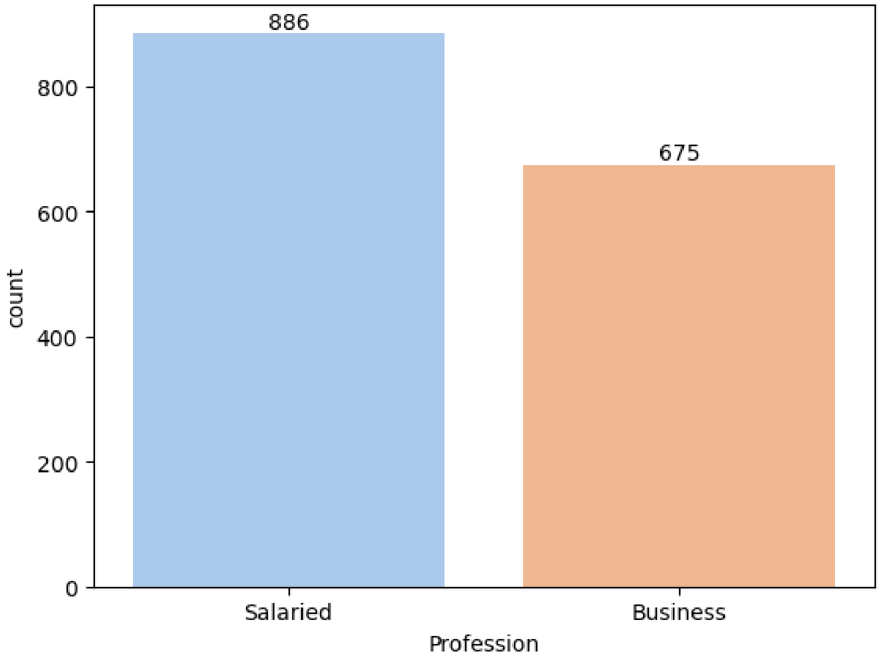


Fig 2.1.3 b)

Insights:

- Clearly, salaried professionals are more than business people who prefer to buy a car.
- 56.6% of the people who buy a car are salaried employees, while 43.3% of the people are from the business class.
- 896 people who buy a car are salaried while 685 are from business class.

c) Marital_status

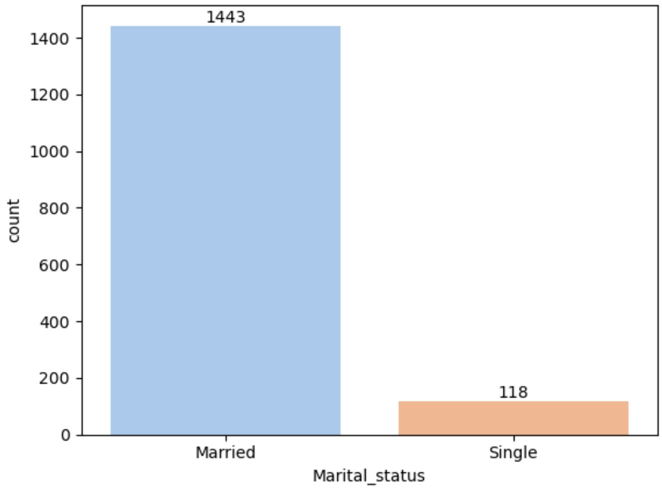


Fig 2.1.3 c)

Insights:

Married people outweigh single unmarried people who buy a car. 91.2% (1443 out of 1581) of the people who buy a car are married while 8% (138)are single.

d) Education

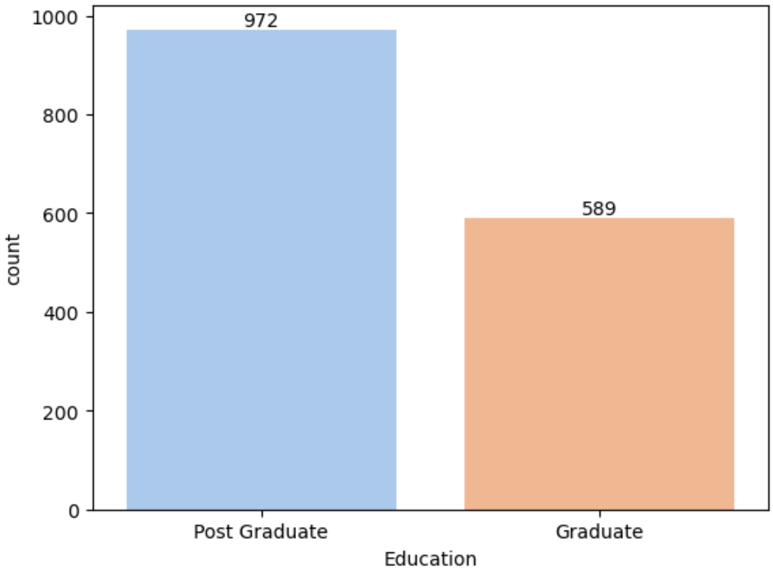


Fig 2.1.3 d)

Insights:

62.3% of the people who buy a car are post-graduates while 37.7% are only graduates. 985 post-graduates buy a car while 596 graduates buy a car.

e) Personal_loan

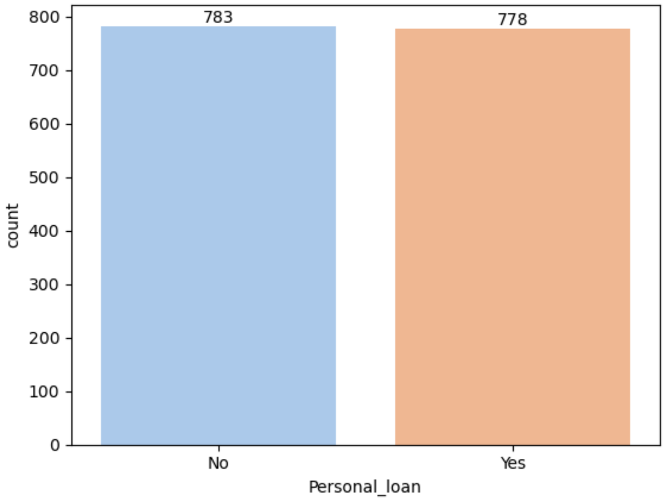


Fig 2.1.3 e)

Insights

The percentage of people who take personal loan and who do not take personal loan to buy a car is almost the same, that is, 792 people who have taken personal loan buy a car and 789 who do not take personal loan also buy a car. It is almost 50% for both.

f) House_loan

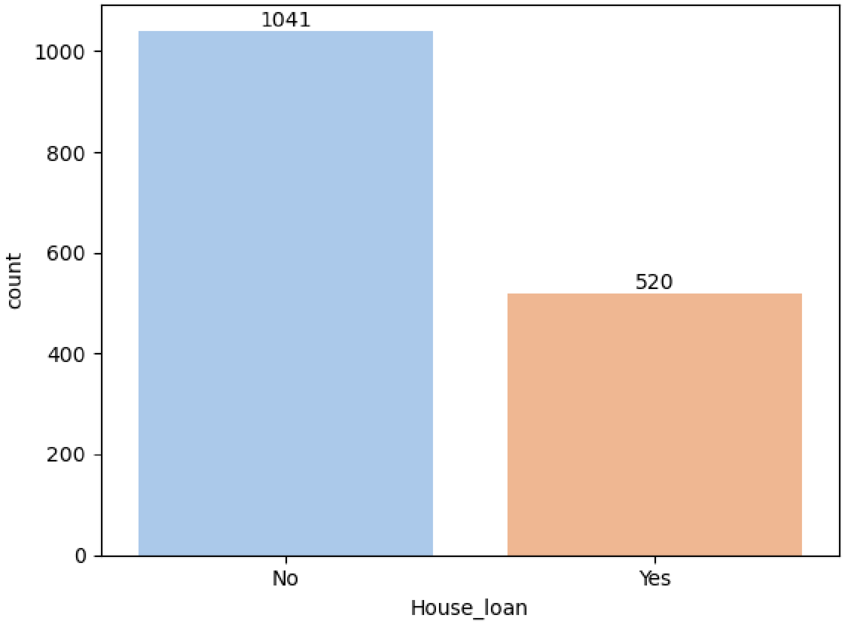


Fig 2.1.3 f)

Insights: The graph shows that 66.6% (1054 in number) of the people who buy cars do not have house loans while 33.3% (527) have house loans.

g) Partner_working

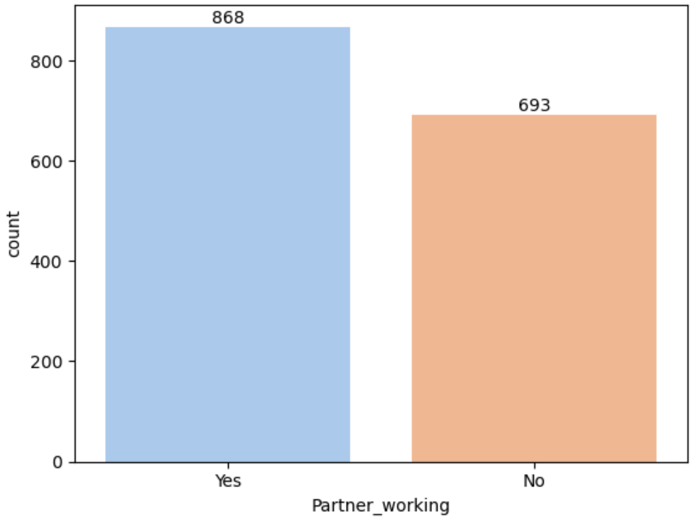


Fig 2.1.3 g)

Insights:

The graph clearly shows that among the people who buy cars, 55% of the people have working partners while 45% have non-working partners. Or, 868 people who buy a car have working partners while 713 have partners who are not working.

h) Make

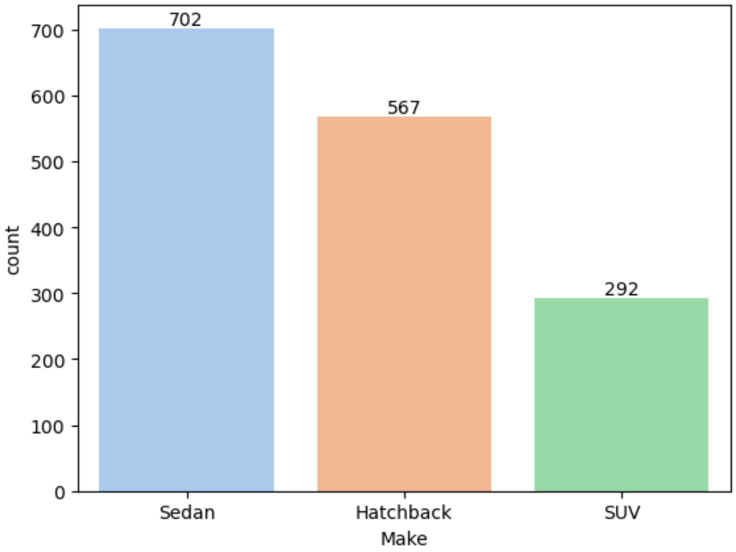


Fig 2.1.3 h)

Insights:

44.4% of the people who buy a car prefer Sedan, 36.8% Hatchback while 18.7% SUV. People prefar Sedan more than hatchback or an SUV. 702 people prefer sedan, 582 prefer Hatchback while 297 prefer SUV.

2.2 Bivariate Analysis

2.2.1 Numerical vs Numerical

a) Age vs No_of Dependents:

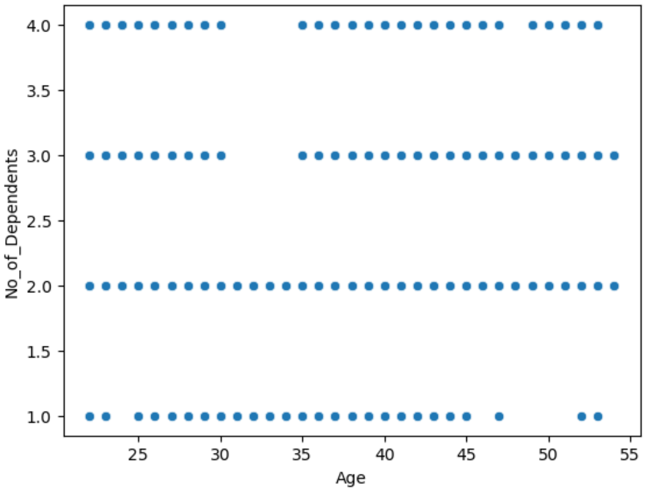


Fig 2.2.1 a)

Insights: There is no correlation between Age and No_of_Dependents as the dots are all scattered all over.

b) Age vs Salary

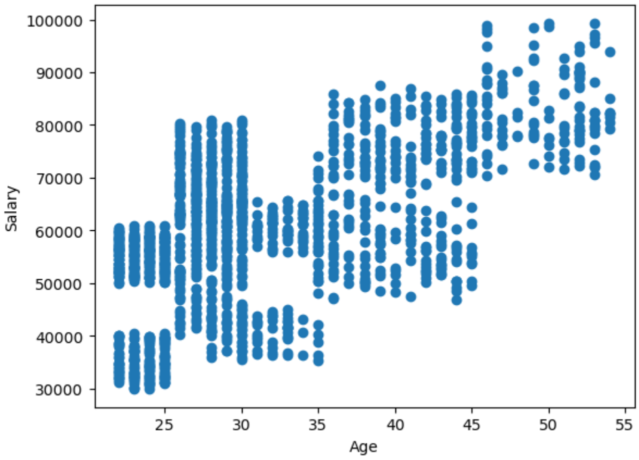


Fig 2.2.1 b)

Insights: There seems to be a positive correlation between Age and Salary. As Age increases, Salary also increases.

c) Age vs Partner_salary

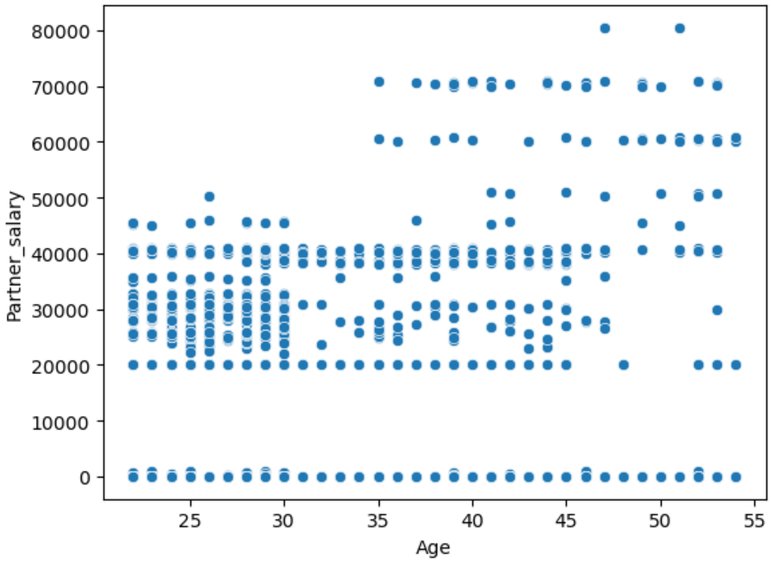


Fig 2.2.1 c)

Insights: There is no correlation between Age and Partner_salary.

d) Age vs Total Salary

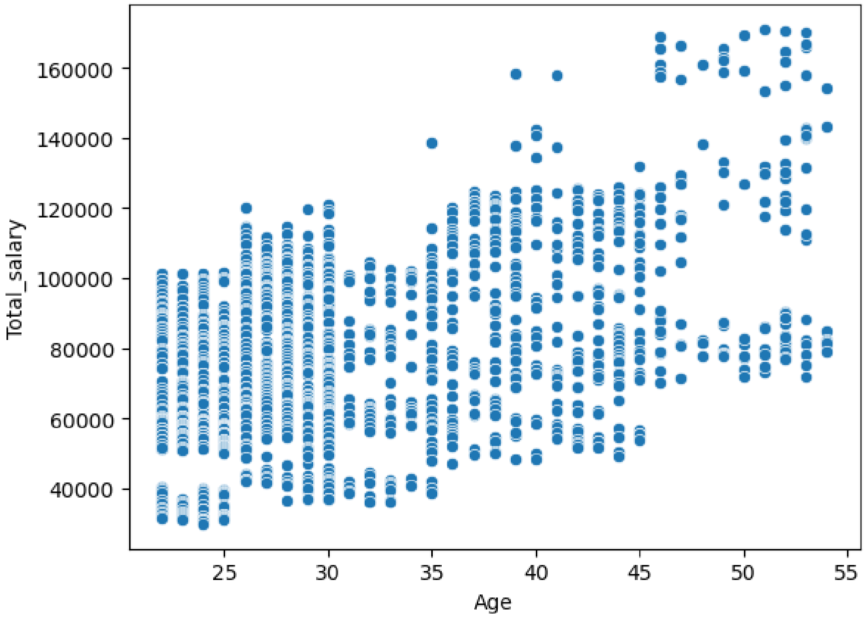


Fig 2.2.1 d)

Insights: We can see a positive correlation between Age and Total Salary. As Age increases, the Total_salary also increases linearly.

e) Age vs Price

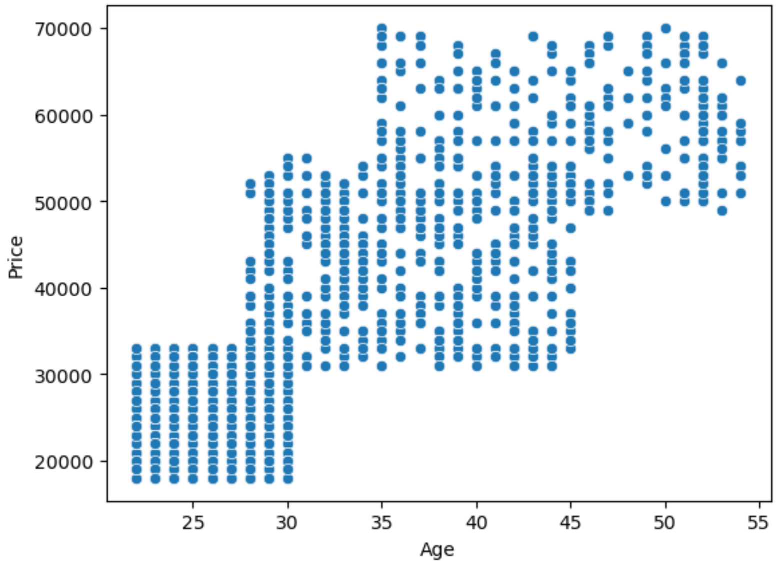


Fig 2.2.1 e)

Insights: There is a positive correlation between Age and Price. As Age increases, they prefer cars of higher prices.

f) Salary vs Total salary

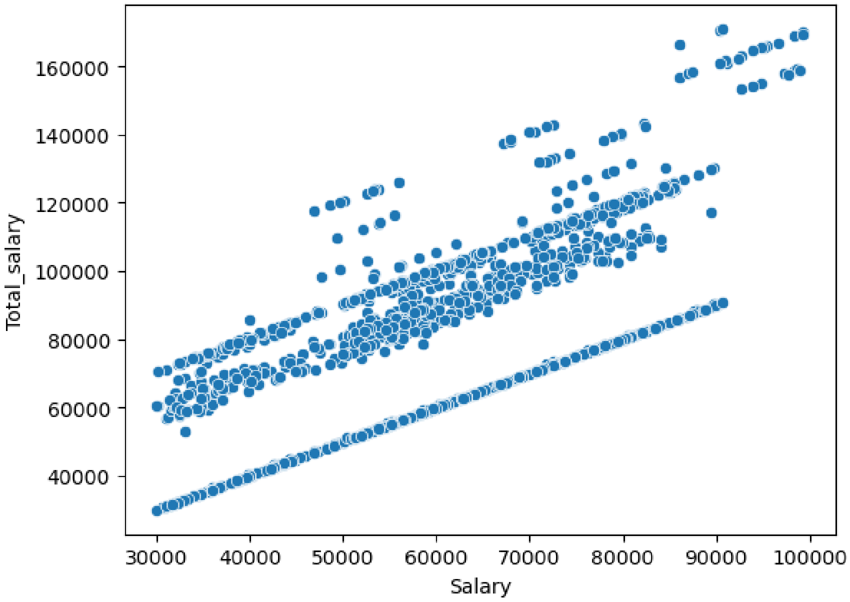


Fig 2.2.1 f)

Insights:

There is a positive correlation between Salary and Total Salary. As salary increases, the total salary also increases.

g) Salary vs Price

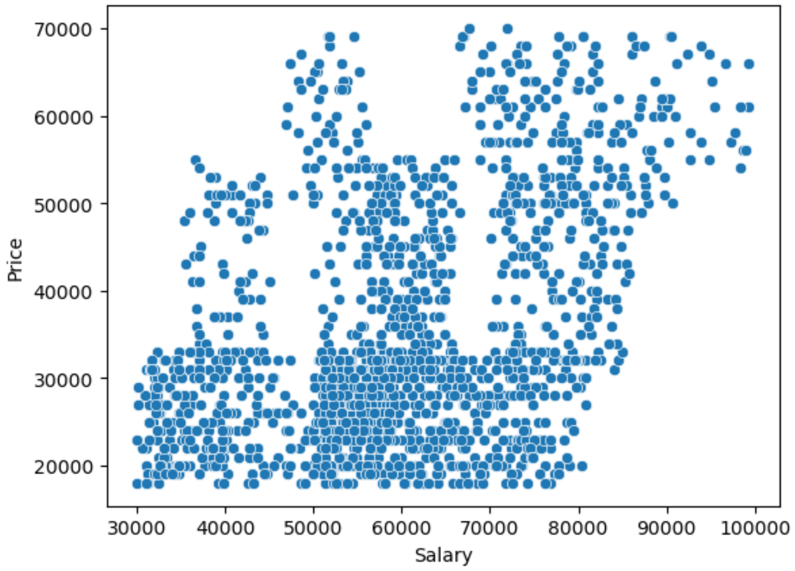


Fig 2.2.1 g)

Insights:

There is some correlation between Salary and Price. We can see that as Salary increases, Somewhat higher price of cars are preferred.

2.2.2 Pairplot for Numerical vs Numerical data:

In order to get a wider picture, we can plot a pairplot which will give us a comparative analysis of each of the numerical variables varying with each other.

Insights of Numerical vs Numerical data:

From the above scatterplots, we can see that there is a positive correlation between:
Age and Price
Age and Total_salary fields
Age and Salary
Salary and Total_salary
Partner_salary and Total_salary
Price and Total_salary
Price and Salary
Price and Age
As one increases, the other also increases indicating positive correlation. There is no correlation between all the other fields.

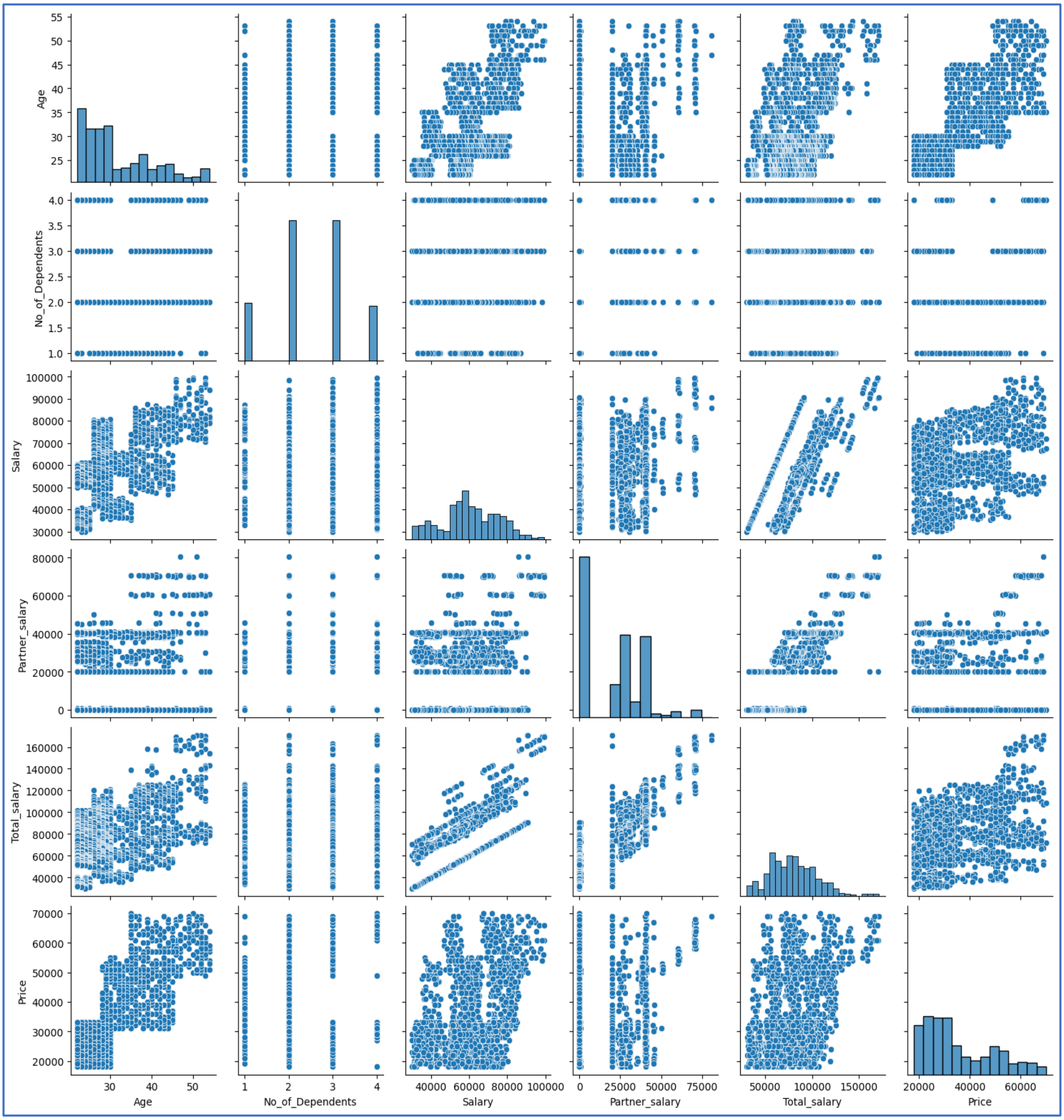

Age
55
50
45
40
35
30
25
No_of_Dependents
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Salary
100000
90000
80000
70000
60000
50000
40000
30000
Partner_salary
80000
60000
40000
20000
0
Total_salary
160000
140000
120000
100000
80000
60000
40000
Price
70000
60000
50000
40000
30000
20000
30
40
50
Age
1
2
3
4
No_of_Dependents
40000
60000
80000
100000
Salary
0
25000
50000
75000
Partner_salary
50000
100000
150000
Total_salary
20000
40000
60000
Price

Fig 2.2.2

2.2.3 Correlation between all the numerical variables

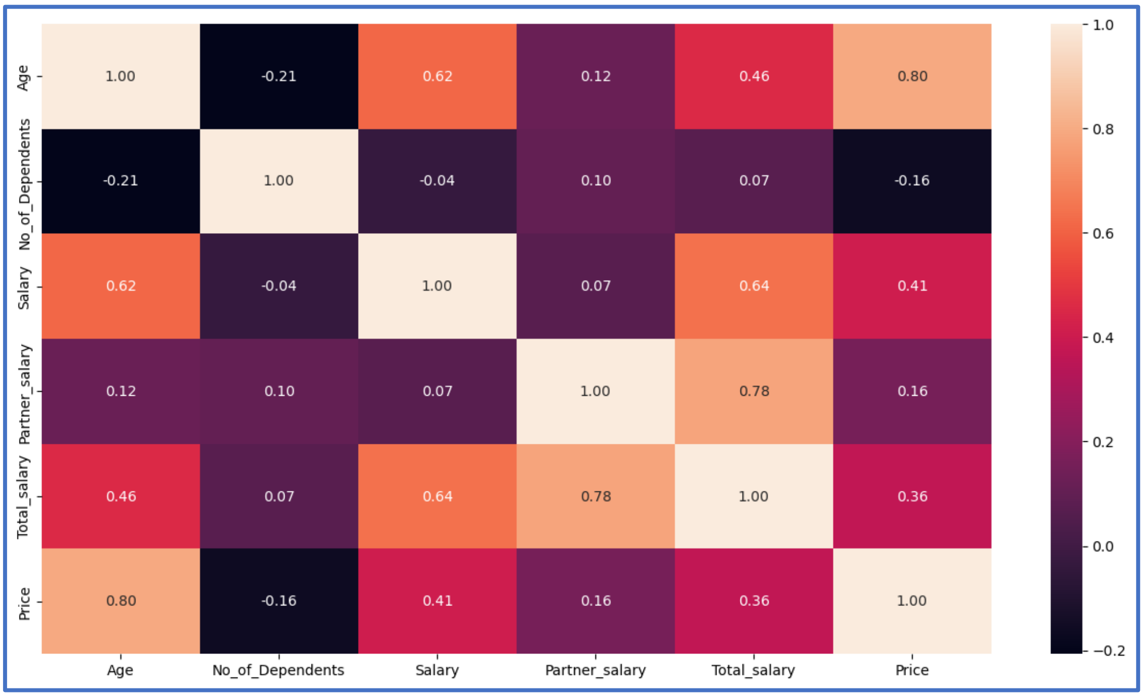


Fig 2.2.3

Insights:

From the heatmap, it is clearly seen that Age and Price fields are very highly correlated. It is also observed that there is a positive correlation between: Age and Total_salary Age and Salary Salary and Total_salary Partner_salary and Total_salary.

2.2.4 Categorical vs Numerical

Use a boxplot to plot categorical variable against a numerical variable to get an insight of the variability of the two variables.

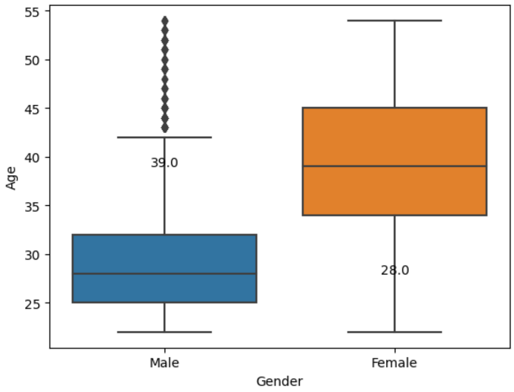


Fig 2.2.4

a) Boxplot for categorical variable ‘Gender’ vs all numerical variables – Age, No_of_Dependents, Salary, Partner_salary, Total_salary and Price.

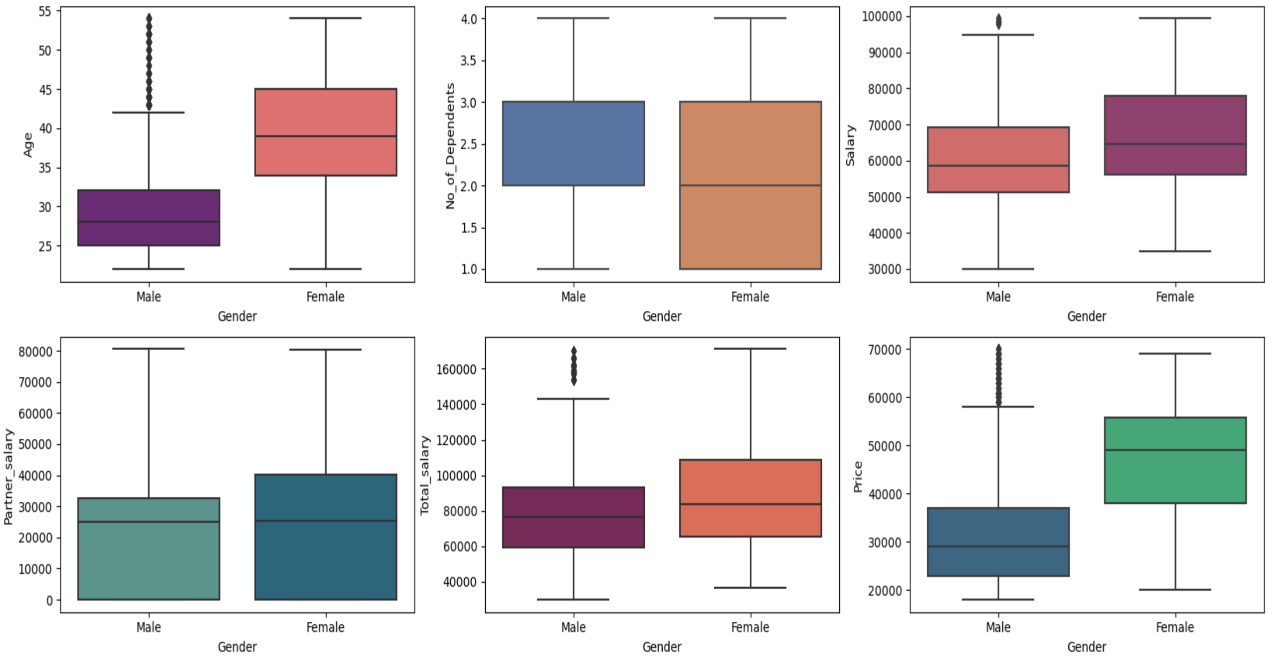


Fig 2.2.4 a)

Insights:

Gender vs Age: The median value of Males who buy a car is much lesser than the median value of females who buy a car. The age of males who buy a car ranges from 22 years to 43 years while that of females, it ranges from 22 years to 54 years. There are outliers when it comes to males who buy a car. There are quite a few males who are more than the upper limit of 42 years who buy a car. The males age group have large number of outliers. Females have more variability as compared to males. The age distribution of males is right skewed.

Gender vs No_of_Dependents: The range of female number of dependents is much more than the male dependents. The maximum age of male and female dependents is the same. The female dependents show more variability than males. There is one low outlier in the male dependents.

Gender vs Salary: The median salary of females buying a car is slightly more than the median salary of males. There is a high outlier in the salary of male.

Gender vs Partner_salary: The median Partner_salary of male and female is the same. The minimum and maximum of both male and female partner_salary is the same.

Gender vs Total_salary: The median total salary of female is more than that of males. The range of female total salary is more than that of males. There are high outliers in total salary of males.

Gender vs Price: The median price is much higher for females as compared to males. The maximum price of cars that females buy is much higher than that of males. There are high outliers in prices for males.

b) Profession Vs all numerical variables

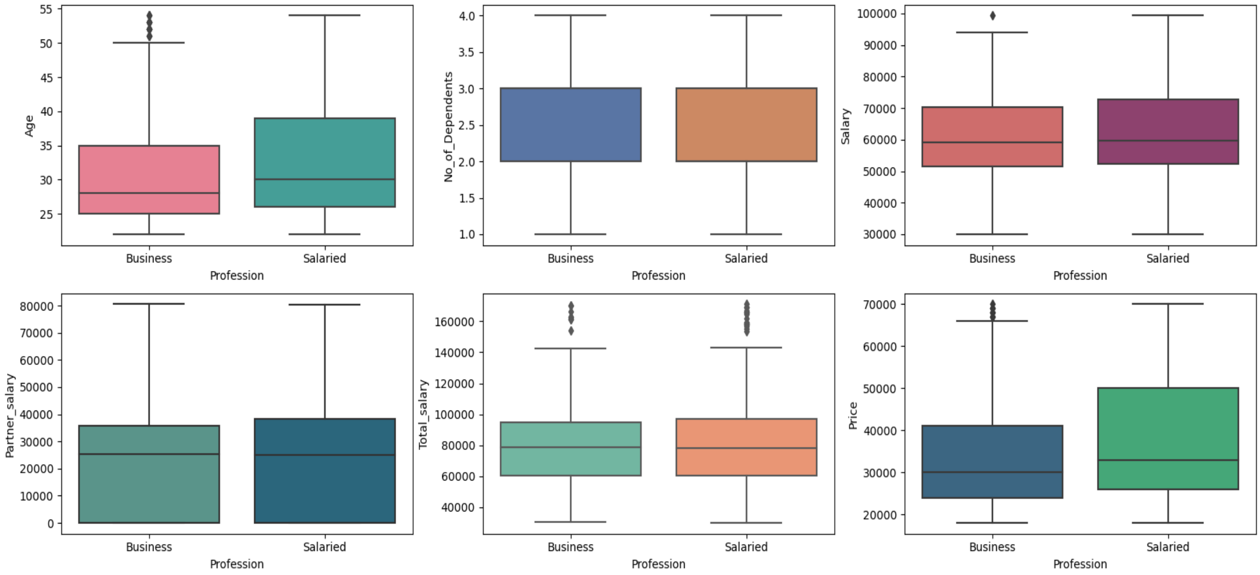


Fig 2.2.4 b)

Insights:

Profession vs Age: The salaried people's ages are right skewed. The business median age is around 27 and have outliers of 50+ age.

Profession vs Salary: The variation of salaried customers is higher. The median salary looks similar. Business customers having outliers in salary.

c) Marital_status Vs all numerical variables

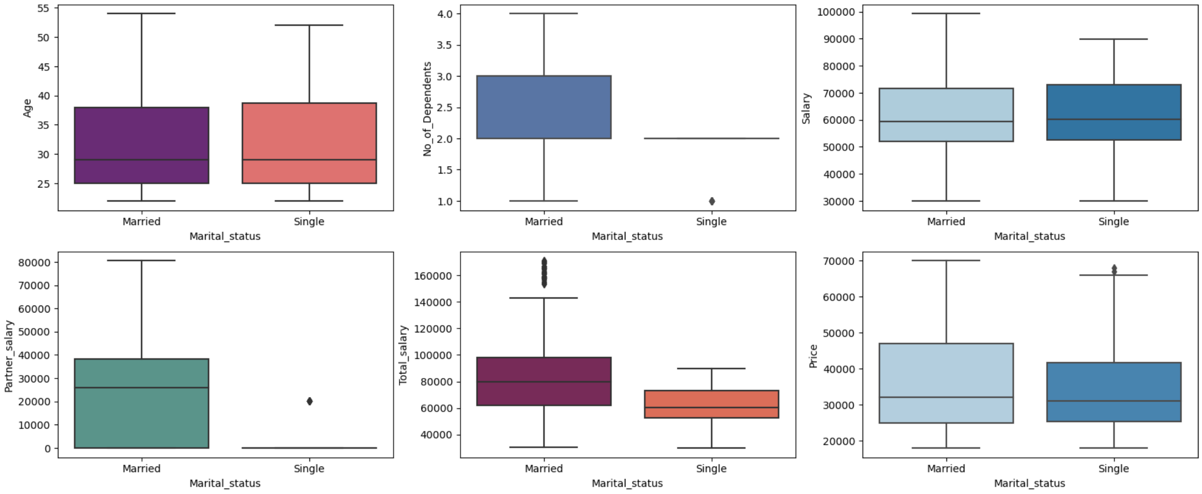


Fig 2.2.4 c)

<u>Insights:</u>

Marital_status vs Total_salary: The married people have slightly higher median salary as compared to single people. Also, the maximum salary is higher in case of married.

Marital_status vs Partner_salary: It is very clearly seen from the boxplot that single people do not have partner salary, which is obvious.

d) Education Vs all numerical variables

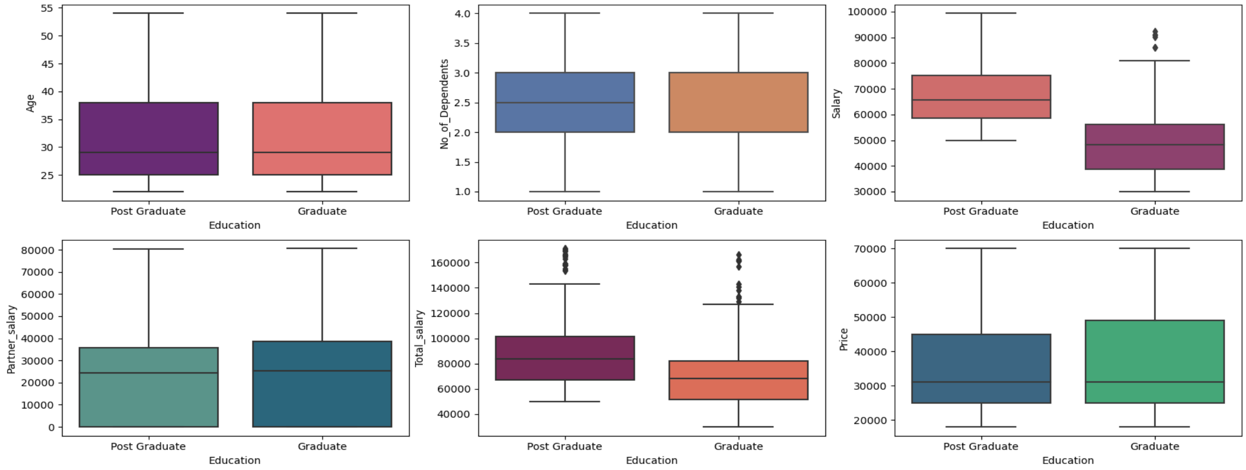


Fig 2.2.4 d)

<u>Insights:</u>

Education vs Salary: The boxplot shows post graduates getting much higher median salary as compared to graduates.

Education vs Age: The median age, minimum and maximum age of graduates and post graduates is almost the same.

e) Personal_loan Vs all numerical variables

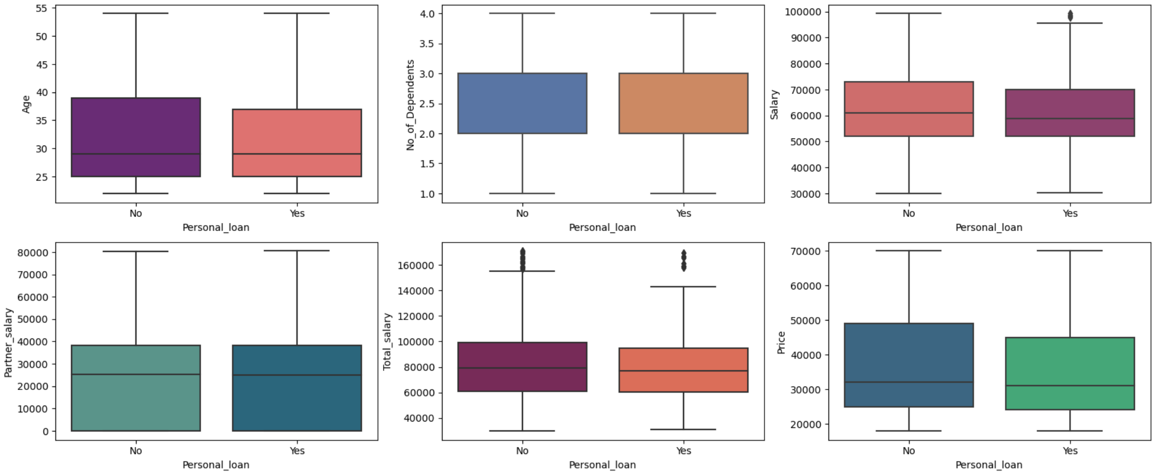


Fig 2.2.4 e)

Insights:

Personal_loan vs Salary: The variability of customers who have not opted for personal loan is higher. The median salary is lesser for customers who opted for personal loans with outliers.

Personal_loan vs Price: The distribution looks right-skewed. The maximum price is slightly higher for customers without a personal loan.

f) House_loan Vs all numerical variables

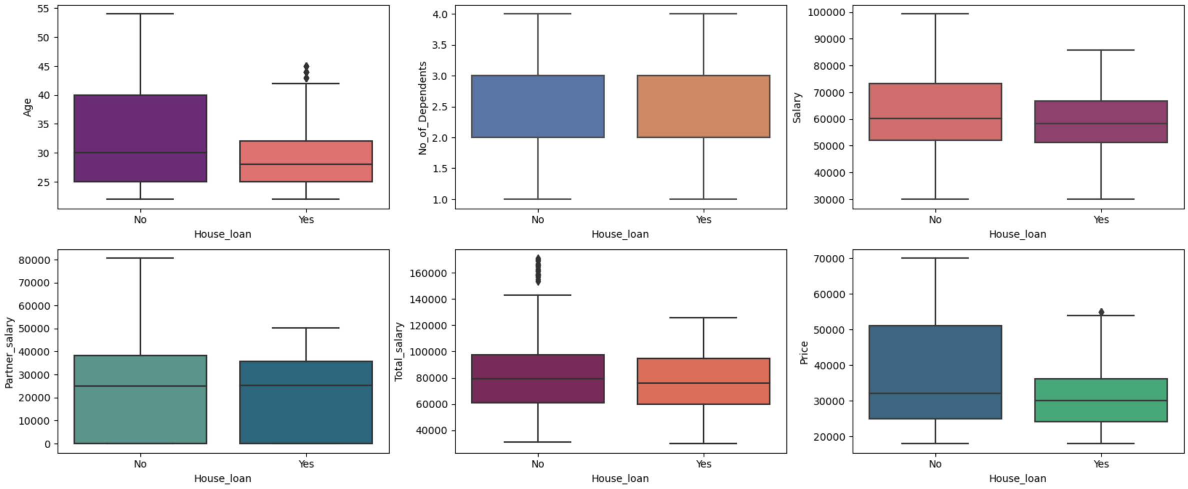


Fig 2.2.4 f)

Insights:

House_loan vs Age: The customers without a house loan seem to have a higher median age and maximum age compared to ones with a house loan.

House_loan vs Salary: The maximum salary differs with house loan as it is slightly lesser than people who do not have house loans.

g) Partner_working Vs all numerical variables

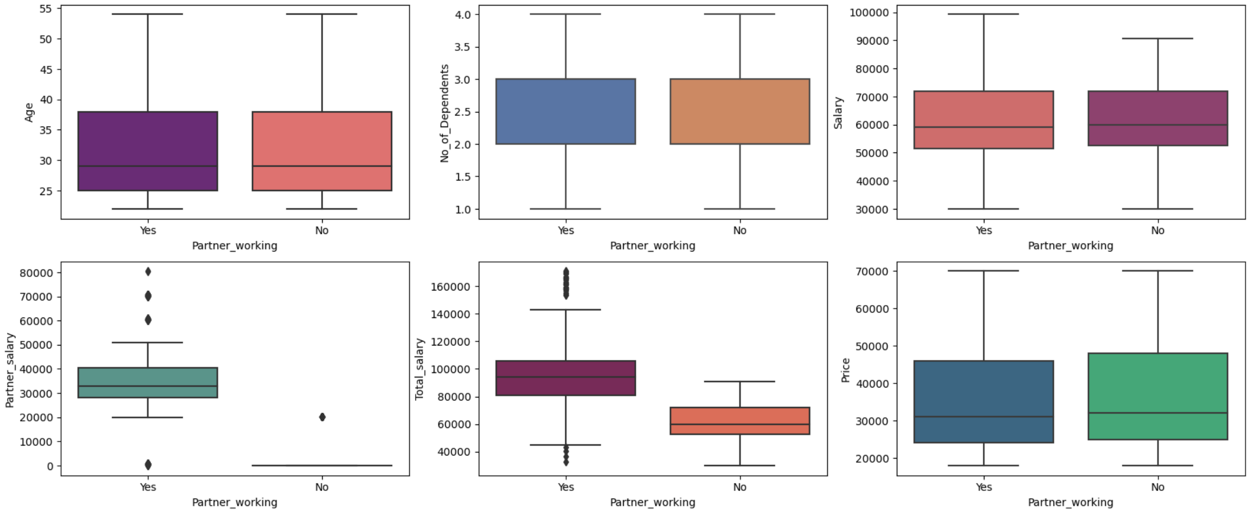


Fig 2.2.4 g)

Insights:

Partner_working vs Partner_salary: As partner is not working, partner salary is also zero, as seen from the box plot. The median age, median salary and median price for working and not working partners is the same.

h) Make Vs all numerical variables

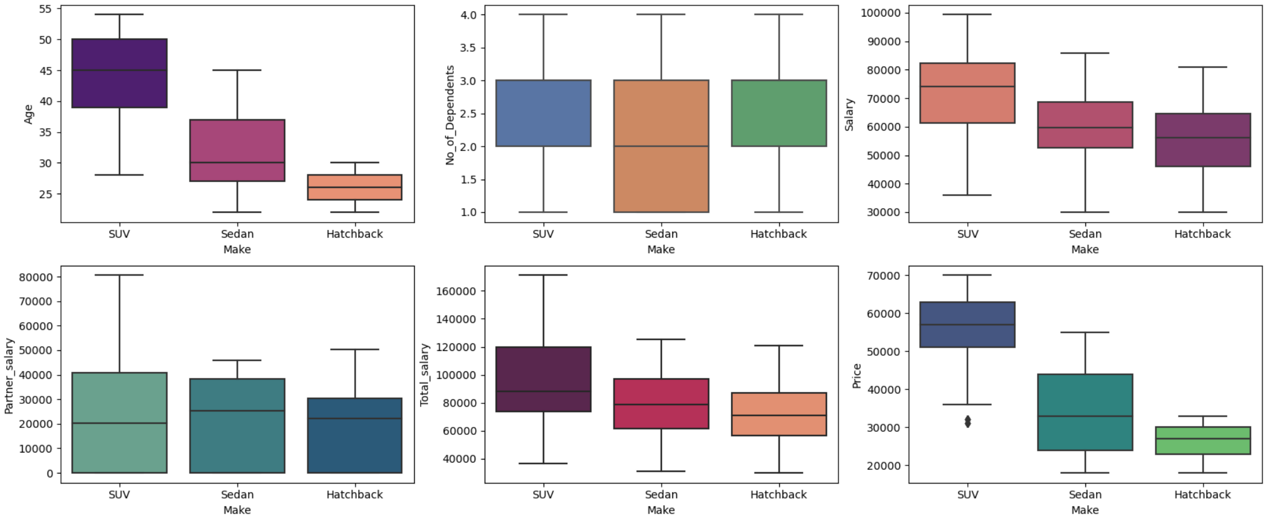


Fig 2.2.4 h)

Insights:

Make vs No_of_dependents: When the number of dependents is 2, the customers prefer Sedans.

Make vs Price: SUVs are very highly priced compared to Sedans and Hatchbacks. Even the minimum salary of SUV is much greater than the median price of Sedans and Hatchbacks.

2.3 Insights – Relationship between Categorical and Numerical variables:

- The boxplots for people taking personal loan and not taking personal loan is almost identical for the numerical categories.
- The minimum and maximum is almost the same for all the boxplots plotting personal loan against the numerical variables.
- The age of people having no house loan shows a lot of variability as compared to people who have a house loan.
- High outliers can be seen in people taking house loans with age, people not taking house loans with their total salary.
- The median of all the box plots for both categories, that is, people taking and not taking house loans is the same.
- Partner_working has outliers in Partner_salary. Outliers can also be seen in Partner_working and Total_salary.
- The median age, median salary and median price for working and not working partners is the same.
- If the partner is not working, the partner does not have any salary.
- The median age of SUV is much greater than the median age of Sedan and Hatchback.
- The SUV is left-skewed while Sedan is right-skewed.
- Both SUV and Sedan show high variability.

- Both SUV and Hatchback have low outliers for no. of dependents.
- Sedan has maximum variability.
- SUV and Hatchback have the same range for no. of dependents.
- Median salary of SUV is more than that of Sedan and Hatchback.
- All the three makes have almost the same range.
- SUV has maximum variability.
- The maximum partner salary is for SUV make.
- Total salary for SUV is right skewed while that for Sedan and Hatchback is normally distributed.
- SUV make has a low outlier.
- The median price for SUV is much greater than that of Sedan or Hatchback.
- Hatchback shows least variability.

3. KEY QUESTIONS

Explore the data to answer the following key questions:

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

Answer: Men don't prefer SUVs more compared to women. In fact, women prefer SUVs more than men, as seen from the plot above. 10.9% women prefer SUVs compared to only 7.8% men. (Refer to the bar plot)

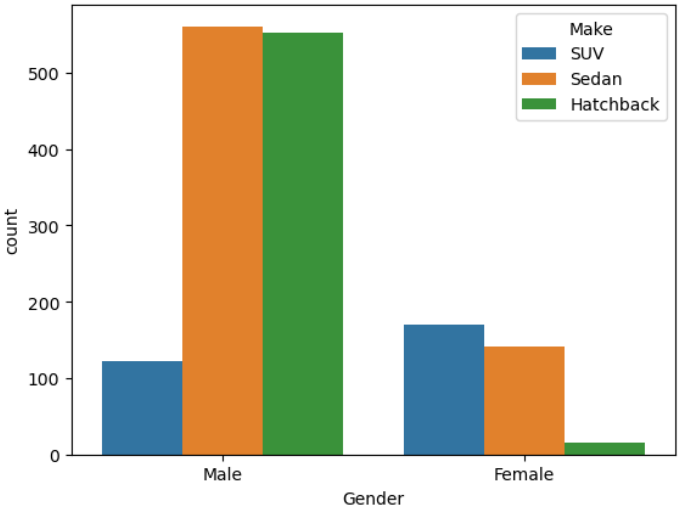
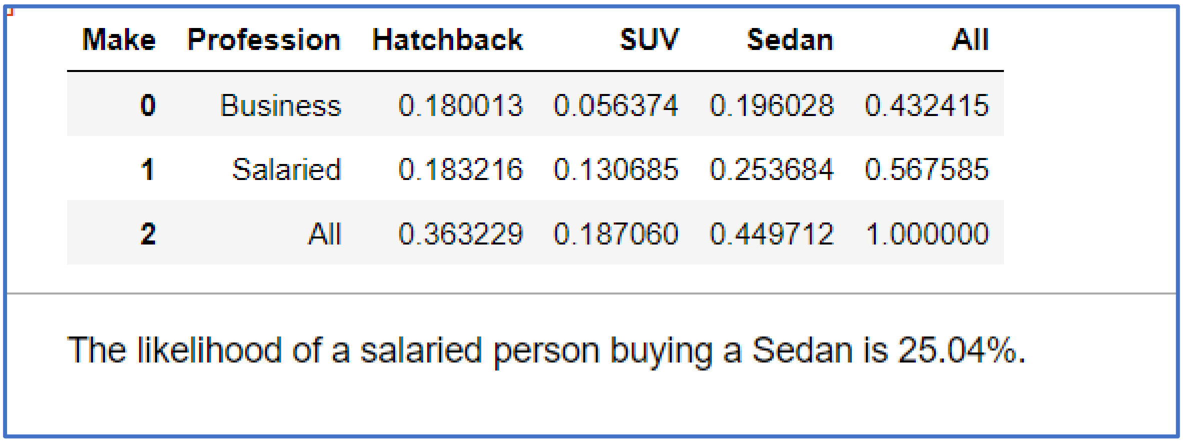


Fig 3.1

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

Answer:



Make	Profession	Hatchback	SUV	Sedan	All
0	Business	0.180013	0.056374	0.196028	0.432415
1	Salaried	0.183216	0.130685	0.253684	0.567585
2	All	0.363229	0.187060	0.449712	1.000000

The likelihood of a salaried person buying a Sedan is 25.04%.

Table 5

3.3. 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?

Answer:

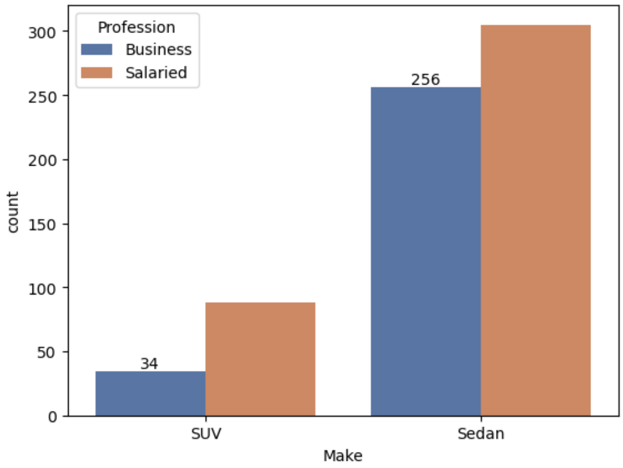
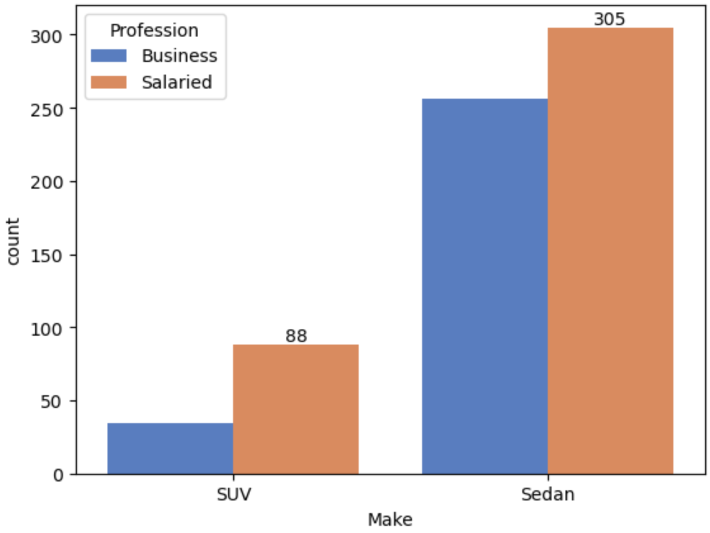


Fig 3.3



From the first plot, the ratio of the business males who buy SUV to Sedan is 34/256 = 0.132

From the second plot, the ratio of the salaried males who buy SUV to Sedan is 88/305 = 0.288.

It is clearly seen from the ratio in the countplot that the ratio of SUV to Sedan of salaried males is highest.

Hence, Sheldon Cooper's claim is right that salaried male is an easier target for SUV sale than Sedan sale.

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

Men spend 40585000 in purchasing automobiles while females spend 15695000. We see that males spend more than 2 times the amount spent by females in buying automobiles.

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

The people who took a personal loan spent 27290000 (2 crore 72 lakh ninety thousand) on purchasing automobiles.

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

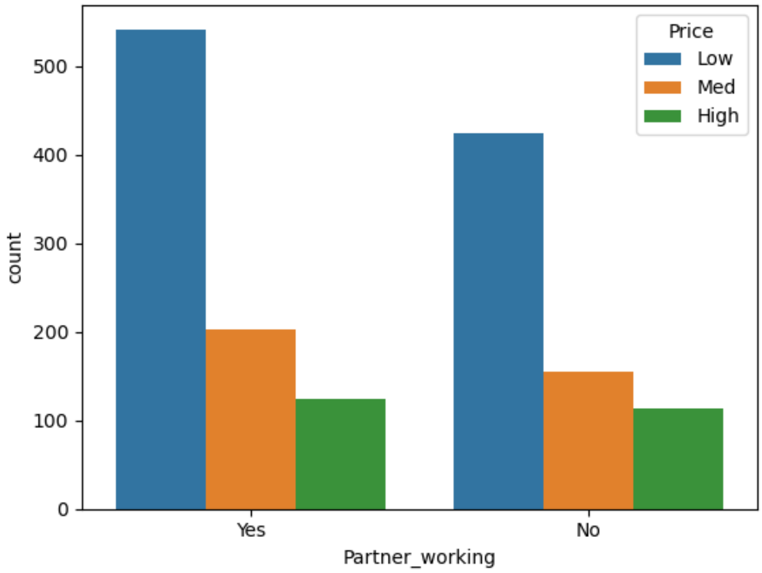


Fig 3.6

- From the bar plot, we can see that the purchase of high priced cars is not dependent on the partner working or not. Irrespective of partner working or not, people are buying high priced cars.
- People with a high total salary(Salary + Partner_salary) greater than 00000 and with their partners working buy higher-priced cars.
- People whose partners are not working are also buying higher priced cars as seen by the orange coloured dots in the price of 50000 to 70000 range.

4. ACTIONABLE INSIGHTS

- Maximum number of people who buy cars are in the age group of 38 years.
- Number of male buyers outweigh the number of female buyers. Males constitute almost 80% of the population which buys cars while females constitute only 20%.
- Salaried class of people prefer to buy cars more than the business class.

- The target audience which buys cars is the married population. The married people buying cars is around 92% while the unmarried people make up the rest of the 8%.
- Education too plays a role in buying cars, with the post graduates buying more cars as compared to the graduate population.
- Another important factor in the car market is the house loan. It is seen that the people who do not have any house loan invest in higher-priced cars.
- The most preferred car is the Sedan followed by the hatchback followed by the SUV.
- We see a trend in age as well. As a person becomes older, he draws more income, as a result, he/she prefers higher-priced cars, but females have a tendency to buy higher priced cars.
- As age increases, people prefer SUVs and Sedans but do not go for Hatchbacks.
- Men don't prefer SUVs more compared to women. In fact, women prefer SUVs more than men.
- 10.9% women prefer SUVs compared to only 7.8% men.
- Men spend more than 2 times the amount spent by females in buying automobiles.
- 50% of car prices lie above 32K~. Customers with above 32k salary can be targeted for high value cars.
- 50% of the customer base has a personal loan. 66.7% of the customer does have a house loan.
- 54.9% of the customers' partners are working who tend to buy high value cars.

5. BUSINESS RECOMMENDATIONS

1. Potential customers for SUVs are the higher age group people.

2. The low-priced cars, which is the hatchback, is mainly bought by males. So, the market for low-priced cars is men.

3. The market for Sedans and SUVs is men and women as both prefer to buy these cars. Female customers contribute more to SUV cars than male.
4. Salaried Male & Business Male contribute to 33.8% of the car sales which is sedan

5. The median price of cars is 50000 for females while it is 30000 for males. Hence, the market for high-priced cars rests with the females.

6. The people have taken house loans go for low-priced cars like Hatchback and do not buy SUVs at all.

7. The sale of Sedan type of cars is correlated with partner working.

8. The youngsters of age group 22 years to 30 years buy low-priced cars, that is, the hatchback.

9. People with higher salaries prefer to buy SUVs, with a median total income of 75000 dollars.

10. First time buyers go for Hatchbacks.

11. Customers who do not have a house loan tend to buy high value cars (18%) and who contribute a high number of car purchases overall. The target customers without a house loan have higher probability for high value cars.

12. Customers with higher total salary where their age is 37+, gender as female and partner working status as yes have a high probability of getting SUV cars with high value.

6. PROBLEM 2:

A bank generates revenue through interest, transaction fees, and financial advice, with interest charged on customer loans being a significant source of profits. GODIGT Bank, a mid-sized private bank, offers various banking products and cross-sells asset products to existing customers through different communication methods. However, the bank is facing high credit card attrition, leading them to reevaluate their credit card policy to ensure customers receive the right card for higher spending and intent, resulting in profitable relationships.

Objective As a Data Scientist at the company and the Data Science team has shared some data. You are supposed to find the key variables that have a vital impact on the analysis which will help the company to improve the business.

6.1 Loading the libraries and dataset:

The dataset godigt_cc.xlsx is loaded using the read_excel() function. The dataset has 8448 rows and 28 columns.

There are 19 numerical variables, 8 categorical variables and 1 datetime field.

Checking for null values:

Check for null values in the dataset using the isnull() function. We observe that the field 'Transistor_revolver has around 38 null values. We are not going to treat any null values as we are going to analyse the dataset as is, and draw some inferences.

6.2 Initial questions that can be raised:

1. How many customers had their current account active in the last 30 days, 60 days and 90 days?
2. Does the annual income of a person affect the credit card limit?
3. Is there any relation between the income earned by the customer and his/her average spend in the last three months?
4. How many cards are issued to the high networth people?
5. Which occupation makes use of credit cards the maximum number of times?
6. Determine the usage of high networth people of different occupations.

7. Does being a Revolver affect the average spend in the last 3 months?

8. Which type of card is the most used?

9. Compare the numerical variables with each other.

10. How does hotlisted card types vary with the annual income?

11. Are cards which are hotlisted used for credit card transactions?

12. Does a Transactor or a Revolver affect the average spend in the last three months?

13. Does holding cards from other banks affect the credit card payments?

14. How are cards affected by a transactor or a revolver?

15. Does a high networth card help in more spendings as compared to other bank credit cards?

16. Does profession of the customer affect holding other bank credit cards?

17. Does occupation determine the transactors and revolvers?

6.3 Top 5 Variables

The top 5 variables that have a vital impact on the analysis which will help the company to improve the business are:

1) annual_income_at_source

2) other_bank_cc_holding

3) avg_spends_l3m

4) Transactor_revolver

5) card type

6.4 Business Justification:

6.4.1 Annual_income_at_source:

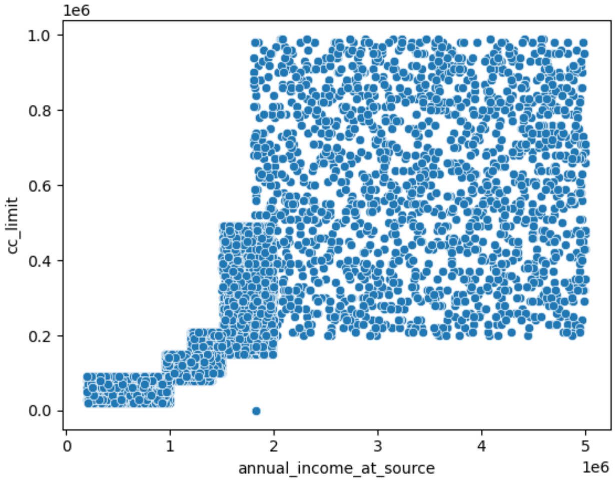
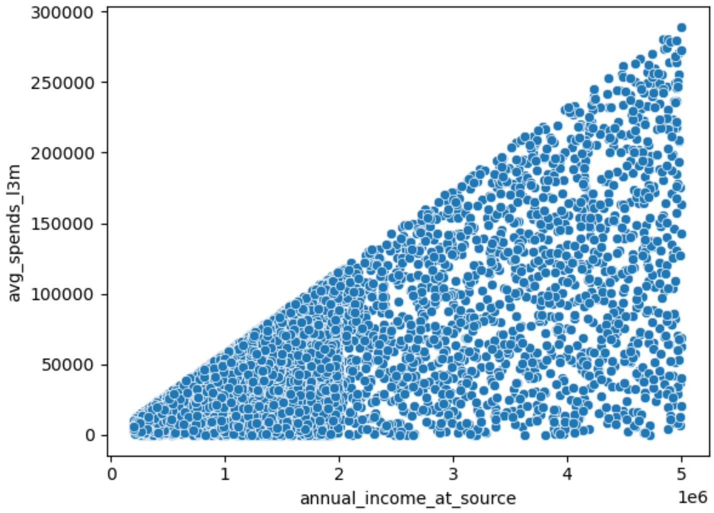


Fig 6.4.1



Insights:

We see that there is a direct correlation between cc_limit and the annual_income_at_source. As annual income increases, the credit card limit of spending also increases. Having said that, people with the same annual income have different credit card limits. We need to dig deeper to understand this aspect.

We see from the scatter plot that as the annual income increases, the average spends in the last 3 months also increases. However, the scatter plot also shows that some people at all income levels have 0 spends in the last three months.

Thus, we need to target people who are not spending enough even though they have higher incomes, maybe because their credit card limit is lesser.

6.4.2 Other_bank_cc_holding:

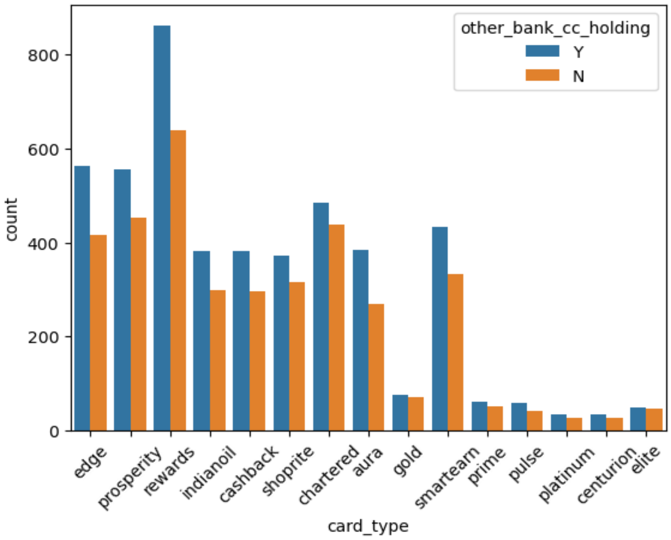
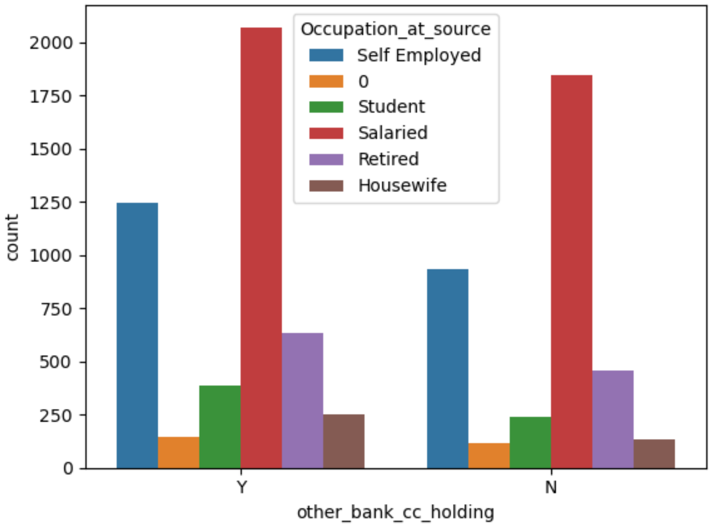


Fig 6.4.2



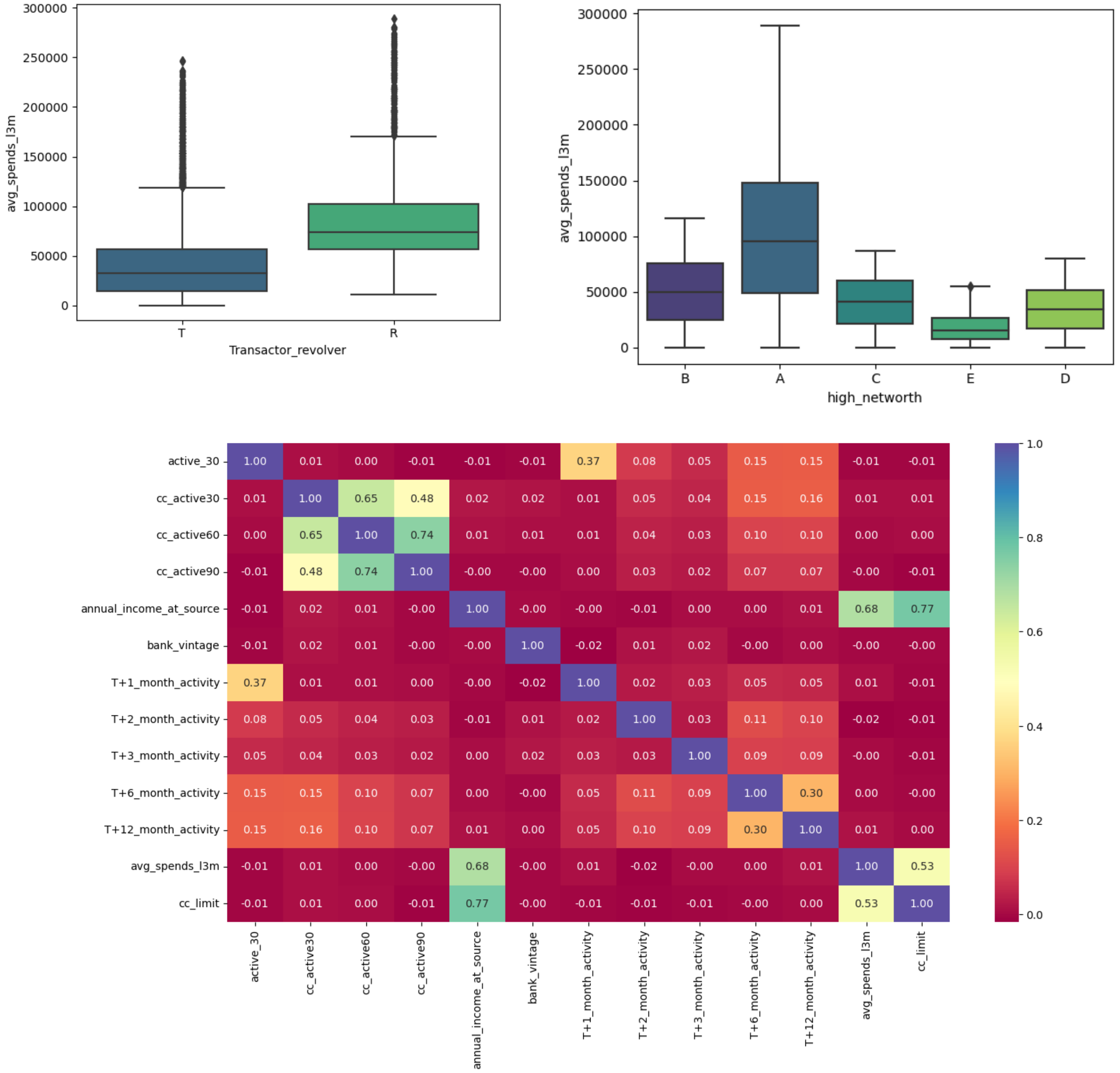
Insights:

We see from the bar plots that other bank credit cards are much more in size than the existing credit cards. The 'rewards' credit card of other banks are the highest among the customers.

We also observe that the salaried professionals have largest number of other bank credit cards followed by the self employed and the retired.

Thus, other bank credit cards influence the spending trend of the existing bank credit card. Maybe, we need to find out what is it that the other banks are offering which we are not able to.

6.4.3 avg_spends_l3m:



	active_30	cc_active30	cc_active60	cc_active90	annual_income_at_source	bank_vintage	T+1_month_activity	T+2_month_activity	T+3_month_activity	T+6_month_activity	T+12_month_activity	avg_spends_l3m	cc_limit
active_30	1.00	0.01	0.00	-0.01	-0.01	-0.01	0.37	0.08	0.05	0.15	0.15	-0.01	-0.01
cc_active30	0.01	1.00	0.65	0.48	0.02	0.02	0.01	0.05	0.04	0.15	0.16	0.01	0.01
cc_active60	0.00	0.65	1.00	0.74	0.01	0.01	0.01	0.04	0.03	0.10	0.10	0.00	0.00
cc_active90	-0.01	0.48	0.74	1.00	-0.00	-0.00	0.00	0.03	0.02	0.07	0.07	-0.00	-0.01
annual_income_at_source	-0.01	0.02	0.01	-0.00	1.00	-0.00	-0.00	-0.01	0.00	0.00	0.01	0.68	0.77
bank_vintage	-0.01	0.02	0.01	-0.00	-0.00	1.00	-0.02	0.01	0.02	-0.00	0.00	-0.00	-0.00
T+1_month_activity	0.37	0.01	0.01	0.00	-0.00	-0.02	1.00	0.02	0.03	0.05	0.05	0.01	-0.01
T+2_month_activity	0.08	0.05	0.04	0.03	-0.01	0.01	0.02	1.00	0.03	0.11	0.10	-0.02	-0.01
T+3_month_activity	0.05	0.04	0.03	0.02	0.00	0.02	0.03	0.03	1.00	0.09	0.09	-0.00	-0.01
T+6_month_activity	0.15	0.15	0.10	0.07	0.00	-0.00	0.05	0.11	0.09	1.00	0.30	0.00	-0.00
T+12_month_activity	0.15	0.16	0.10	0.07	0.01	0.00	0.05	0.10	0.09	0.30	1.00	0.01	0.00
avg_spends_l3m	-0.01	0.01	0.00	-0.00	0.68	-0.00	0.01	-0.02	-0.00	0.00	0.01	1.00	0.53
cc_limit	-0.01	0.01	0.00	-0.01	0.77	-0.00	-0.01	-0.01	-0.01	-0.00	0.00	0.53	1.00

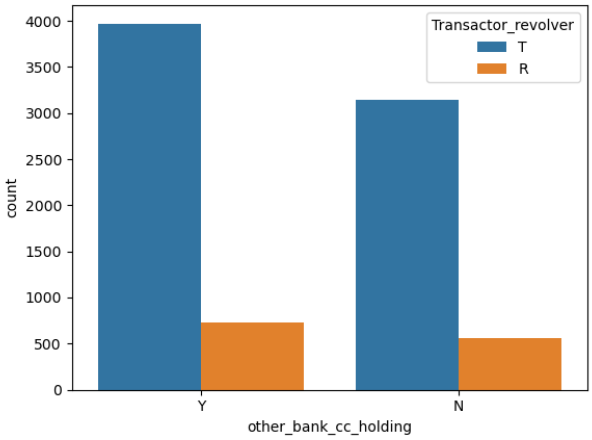
Fig 6.4.3

<u>Insights:</u>

'A', the highest networth category has a median spending in the last 3 months to 100000, while the lowest networth, that is, 'E' has a median spend of around 20000.

Transactor has a lower median spend in the last 3 months as compared to a Revolver. We need to figure out ways to increase the Transactor's spending through credit cards.

6.4.4 Transactor_revolver



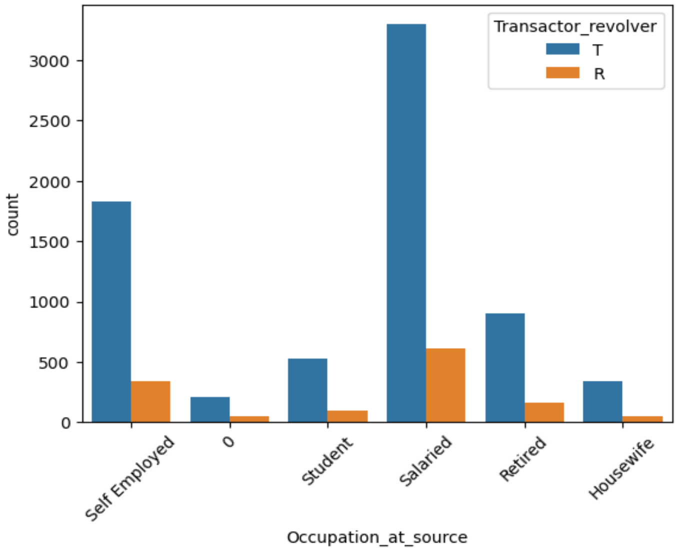
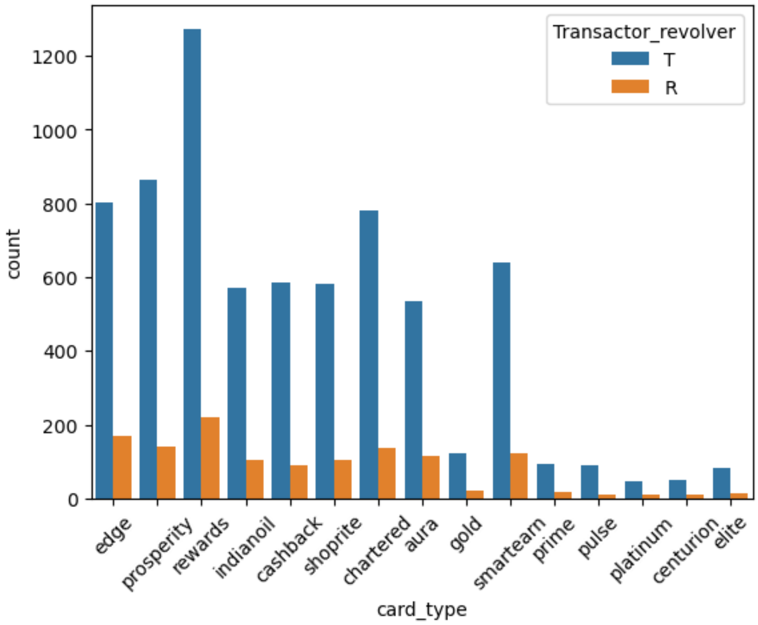


Fig 6.4.4

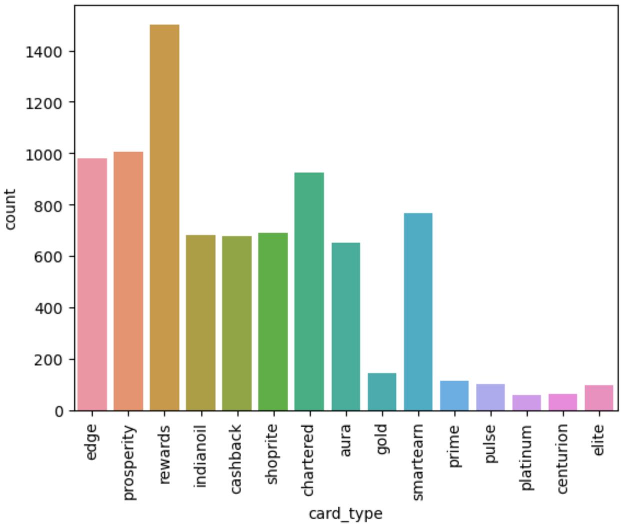


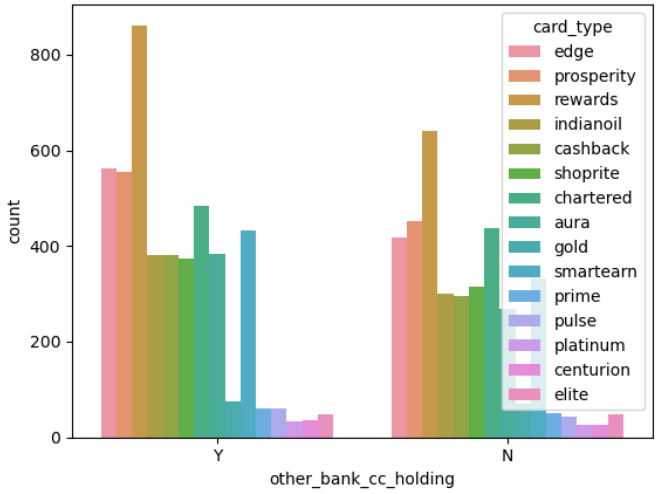
Insights:

As we can clearly see that the salaried professionals are mostly transactors. In the self-employed, students, retired and the housewives, the transactors outnumber the revolvers.

It is clearly seen that the Transactors with other bank credit cards are much more in number as compared to revolvers with other bank credit cards. The transactors with no credit cards from other banks outnumber the revolvers with no credit cards from other banks.

6.4.5 card_type





<u>Insights:</u>

We can see that the 'rewards' card type is the most used by the customers followed by 'prosperity' type. The least used card type is 'platinum' cards. Transactors' spending through credit cards is far more as compared to the revolvers. Also, people holding other bank credit cards is lesser than the people with our bank credit cards.

6.5 CONCLUSION:

Thus, we have identified the five most important variables which are crucial for the credit card business. These are annual_income_at_source, other_bank_cc_holding, avg_spends_l3m, Transactor_revolver and card_type.

****