A Business Project Report

on

Statistical Methods for Decision Making (SMDM)

submitted by

Prabhakar Singh

Batch

PGPDSBA.O. MAY23.A

on

14/07/2023

for partial fulfilment of

PG Program in Data Science and Business Analytics

To

Great Learning
POWER AHEAD

Table of Contents

List of figures

List of tables

Problem 1

Introduction

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 analytics professional to improve the existing campaign.

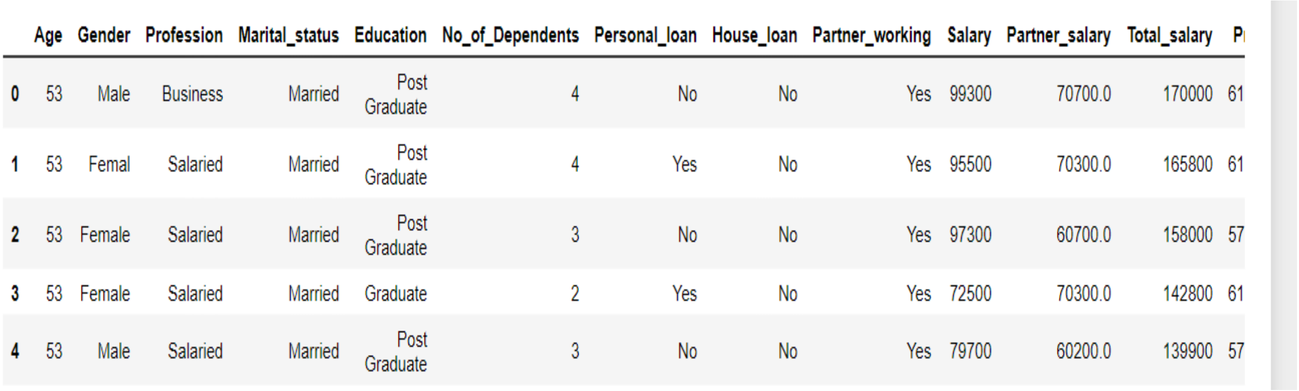
The purpose of this report is to do Austo_automobile data analysis and suggest better ways of marketing campaigns to target different segments of customer present in the dataset to increase car sales.

Data description

Here is list of all the features present in the dataset.

1. Age – age of customers who bought cars
2. Gender- male and female
3. Profession- business and salaried
4. Marital_status- married and single
5. Education- post graduate and graduate
6. No_of_dependants- numbers of member dependant on the customer
7. Personal_loan- yes and no
8. House_loan- yes and no
9. Partner_working- yes and no
10. Salary- salary of the customers (Let's take in dollar)
11. Partner_salary- salary of customer's partner
12. Total_salary- sum of customer salary and partner salary
13. Price- price of the car sold
14. Make- model of the car

Sample of dataset



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

Table 1: - Top five rows of dataset

Problem 1.A

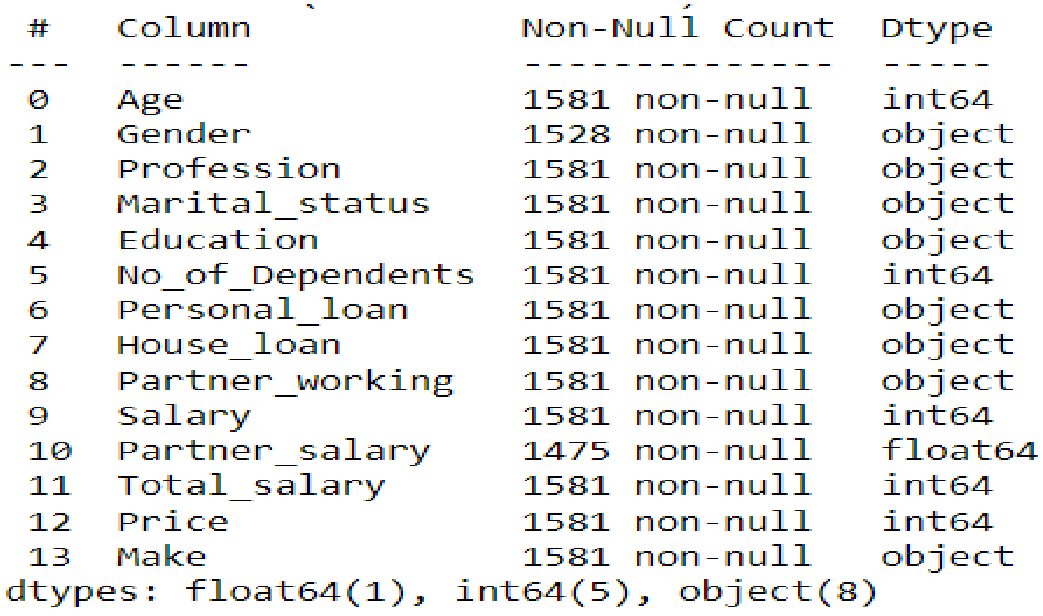
What is the important technical information about the dataset that a database administrator would be interested in?

Solution: -

Here is the list of important technical information about the Austo Motor Company dataset.

A) The given dataset contains 1581 rows and 14 columns.

B) Basic details of dataset



```
 #    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)
```

Table 2: - Basic details of dataset

After looking at the above table it is very much clear that there are null values present in the given dataset.

There are 6 continuous features and 8 categorical features present in the given dataset.

Problem 1.B

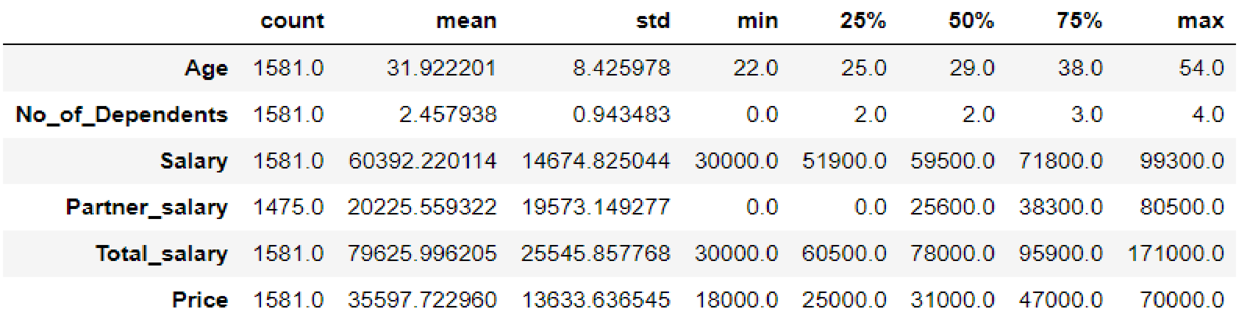
Take a critical look at the data and do a preliminary analysis of the variables. Do a quality check of the data so that the variables are consistent. Are there any discrepancies present in the data? If yes, perform preliminary treatment of data.

Solution: -

A) Checking for presence of any missing and duplicate values

There are **53** missing values present in Gender column and **106** values are present in the Partner_salary column of the dataset. Also, there is no presence of any duplicate values in the dataset.

B) Basic summary of the data



	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1475.0	20225.559322	19573.149277	0.0	0.0	25600.0	38300.0	80500.0
Total_salary	1581.0	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	1581.0	35597.722960	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0

Table 3: - Basic summary of dataset

Distribution of age is looking appropriate with minimum 22 and maximum 54 years with mean age of 32 years. Median age is 29 years.

No of dependants varies from 0 to 4 with mean and median approximately same and hence seems normally distributed.

Salary of the customers varies from $30000 to $99300 with mean salary of $60392.

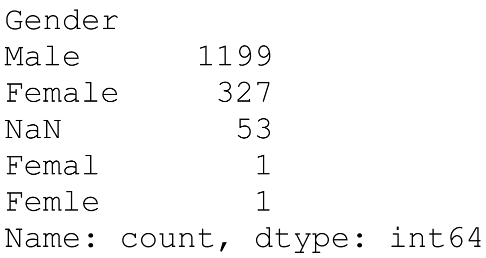
Partner_salary is varying from 0 to $80500 as 0 being no working partner. But here there are some missing values which need to be resolved. As there are no missing values in the total_ salary and salary columns, that's why we can remove missing values from partner_salary column by subtracting salary from total_salary column.

Total_salary is simply sum of salary of customer and partner salary which varies from $30000 to $171000.

Prices of the cars vary from $18000 to $70000.

C) Treatment of missing values

As there are missing values present in the data. So, before detailed analysis of the data, it is very important to treat the missing values to get proper output and insights from data.

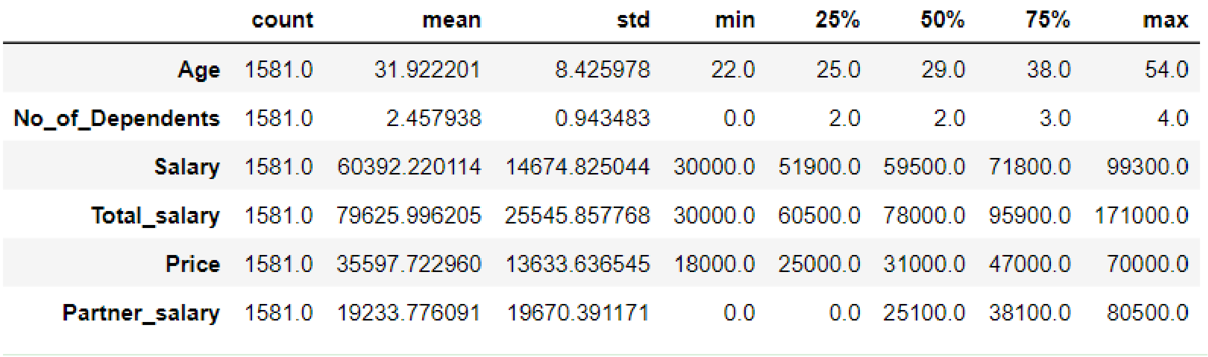


```
Gender
Male       1199
Female      327
NaN          53
Femal         1
Femle         1
Name: count, dtype: int64
```

Table 4: - Details of Gender column

The above table for Gender column shows that in two values there is spelling mistake and maximum occurrence is for male. That's why missing values need to be imputed with mode and two rows with spelling mistake will be corrected and coupled with female.

To treat missing values in Partner_salary column, first Partner_salary column was dropped and the again a Partner_salary column was calculated by subtracting salary from total_salary and summary found was found as below.



	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Total_salary	1581.0	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	1581.0	35597.722960	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0
Partner_salary	1581.0	19233.776091	19670.391171	0.0	0.0	25100.0	38100.0	80500.0

Table 5: - Summary of continuous columns after missing values treatment

Here after treating missing values, Partner_salary mean has changed to $19233.

D) Checking for outliers

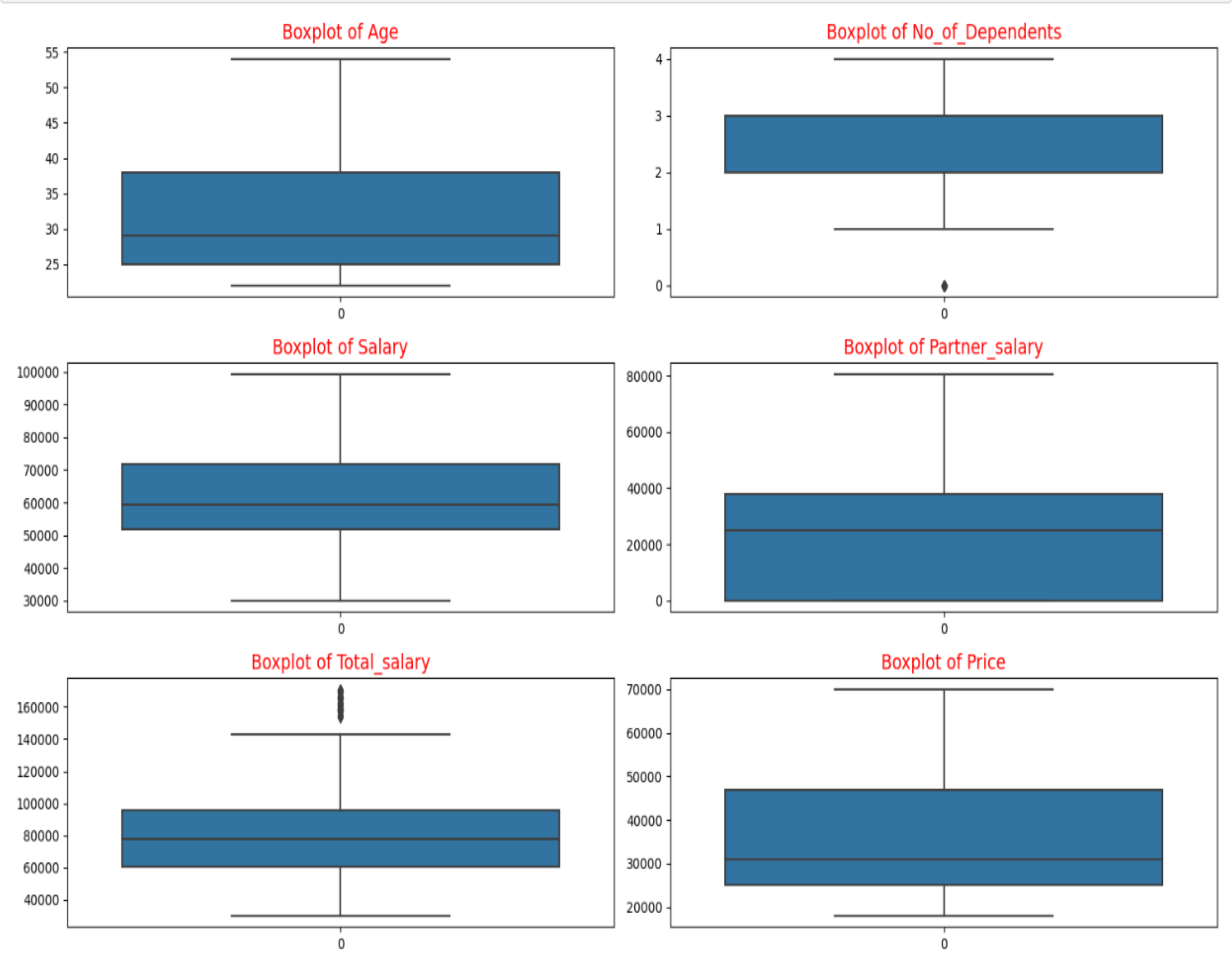


Fig 1: - Boxplot of continuous columns

Fig 1 shows the boxplot of all numerical features. From the figure it is very much clear that No_of_Dependents and Total_salary columns have outliers.

But there is no need of outlier treatment here because in No_of_Dependents column value zero means that there is no dependent for that person and Total_salary column is just the summation of Salary and Partner_salary and both these columns do not have any outliers.

E) Check for any discrepancies in Categorical features

Below fig 2 clearly shows that there is no presence of any discrepancies in the categorical features. All rows have names as described in the data description.

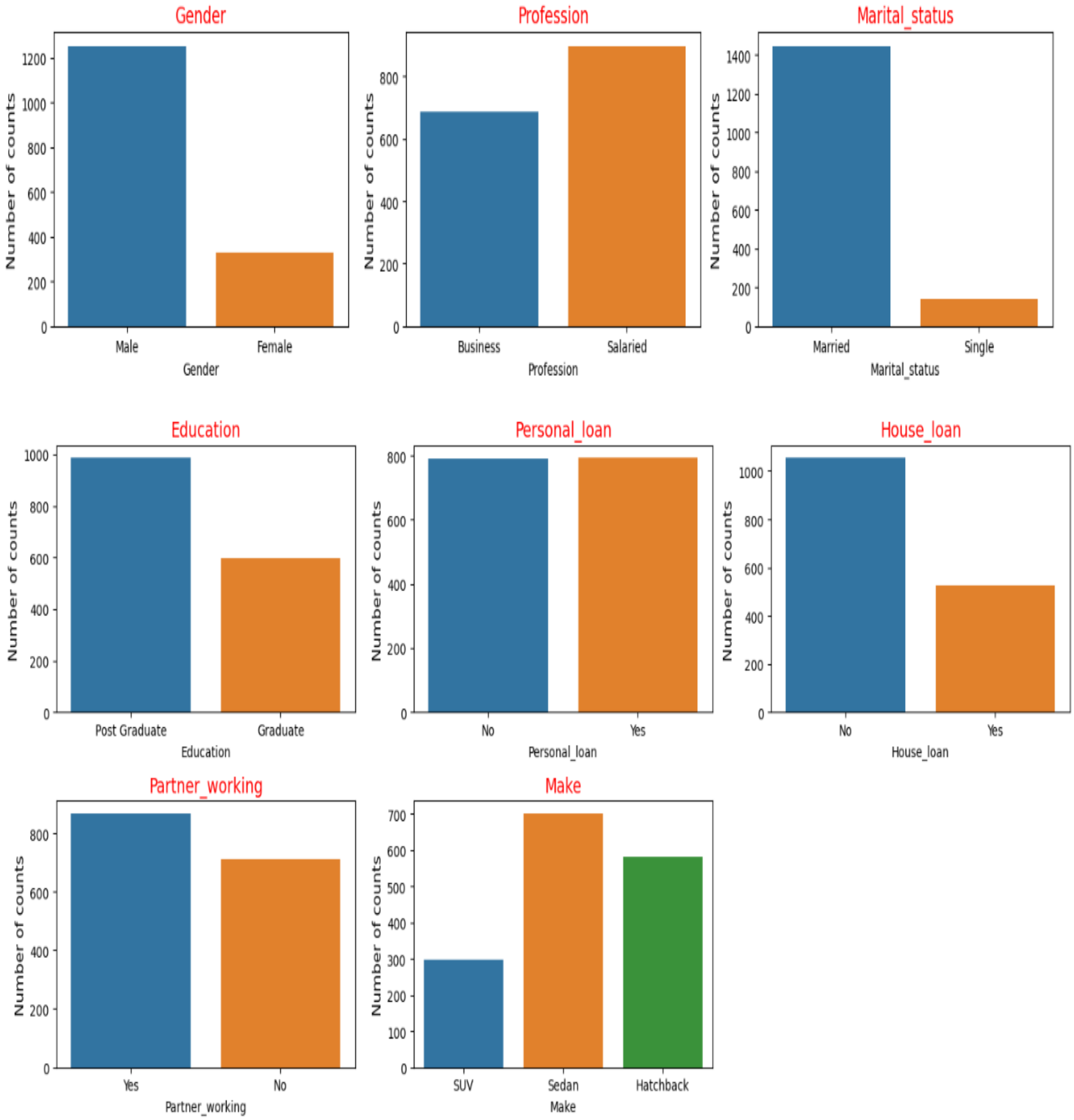


Fig 2:- Countplot of categorical features

One variable named df_single was created containing rows with marital status as single to check presence of any partner_working yes and partner_salary greater than 0. It was found that those who are single have no working partner and partner salary as 0. Hence it is very clear that there is no discrepencies present in the dataset.

Problem 1.C

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

Solution: -

The below plot of age shows that the age group of approximately 22-31 years are buying maximum numbers of cars and it decreases significantly after 31 years. Here we can give suggestions to company to capture more customers between age group of 22-31 and do some market surveys to understand the preferences of customer of higher age group.

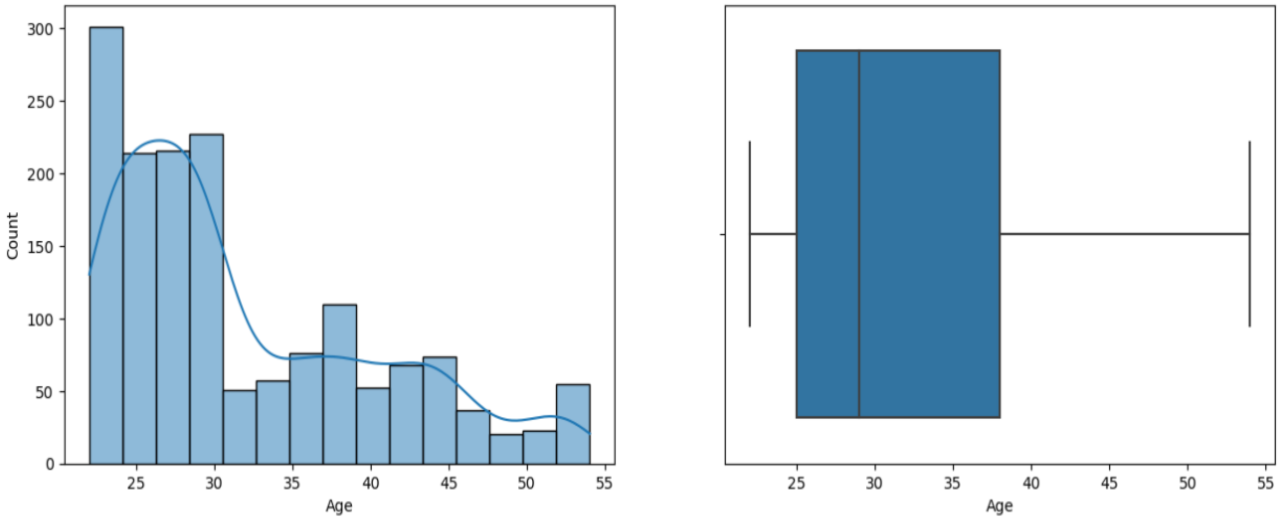


Fig 3: - Histogram and boxplot of Age

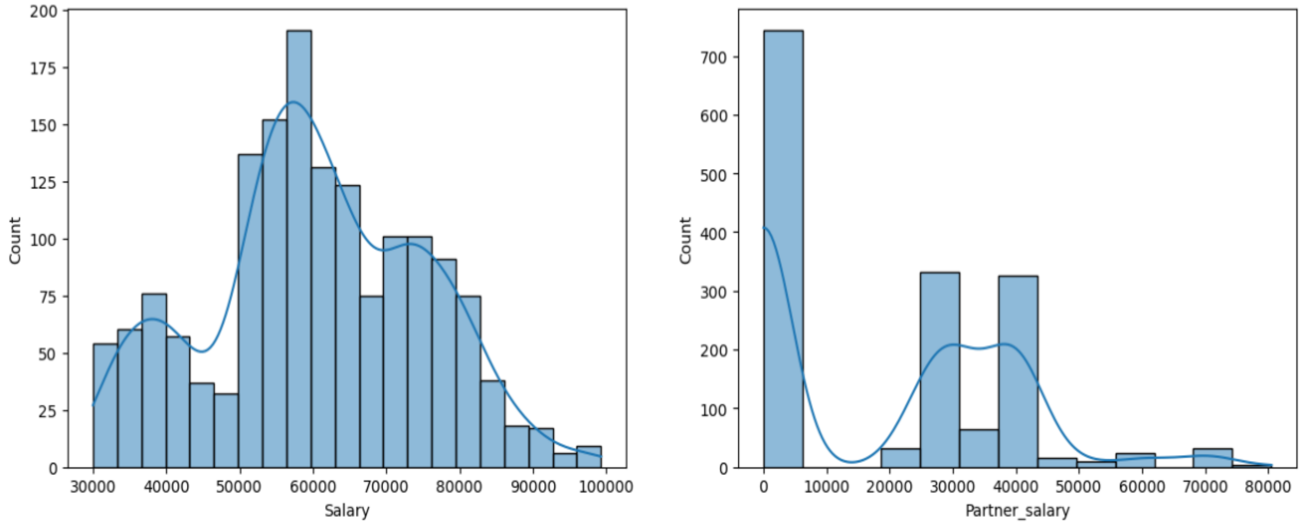


Fig 4: - Histogram of Salary and Partner_salary

Above plot suggests that maximum customers and target customers who are buying cars are having salary between range of $50000 to $80000. Almost 50% of customers are those whose partner_salary is less than $6000 and remaining have salary varying from $20000 to $80000. Hence we can conclude that partner_salary has impact on car sales but not too a large extent. Because out of 1581 customers almost 700 customers have no working partner.

Histogram of Price of cars indicate that almost 65% of the cars lie in price range of $18000-$35000 and 75% of cars have price below $47000. Hence company should focus on production of these price range cars. Company should do marketing campaigns and devise market strategies among high earning groups to increase sales of high-priced cars.

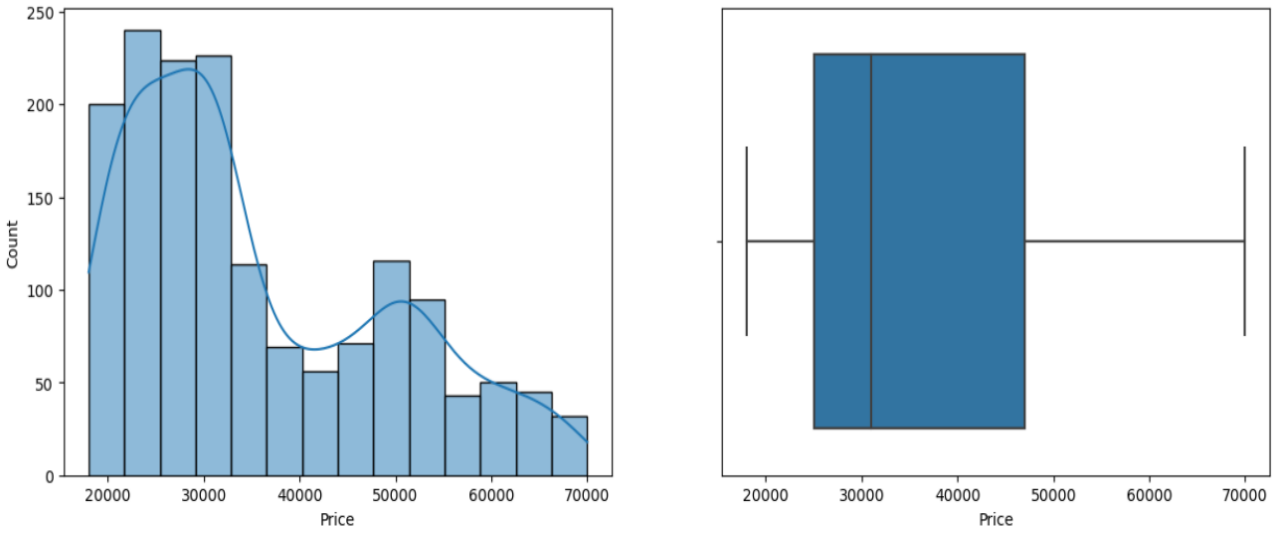


Fig 5: - Histogram and boxplot of Price

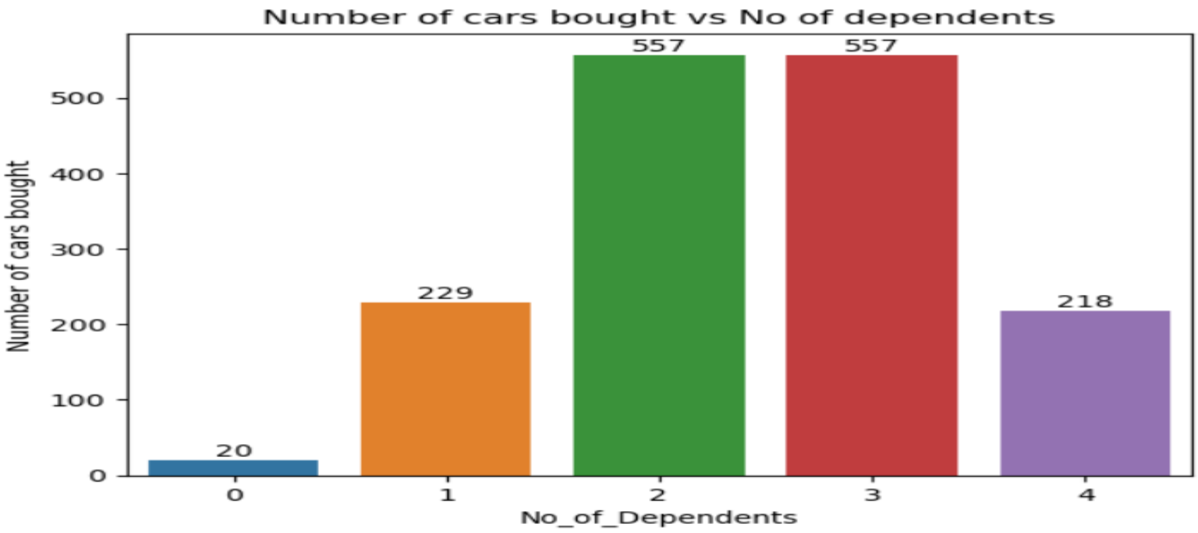


Fig 6: - Countplot of No_of_Dependents

The above plot tells that most numbers of people buying cars have 2 and 3 dependents. The basic reason of this might be a family of 4 and 5 members need cars to carry them.

Business insights and useful information from categorical features of dataset as drawn from fig no. 2 are -

The no. of males buying cars are more than 3 times as compared to females. Car company can encourage females to buy more cars by different modes of advertisement.

In the given dataset salaried persons are buying more cars than business people. In my view, one of the reasons of this might be that most of the businessmen like luxury cars and our dataset contains only three Make of cars in which Sedan have been sold the most and SUV have been sold the least.

Married people are buying almost 90% of the cars. One reason might be that the married people have more members to carry them.

People having higher education are buying more cars because they are earning more.

Number of people taking loans to buy cars is same as number of people not taking loan. Also, people having no home loan are buying more cars which is quite obvious.

Sedan is the most preferred car make and SUV is the least preferred.

Problem 1.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.

Solution: -

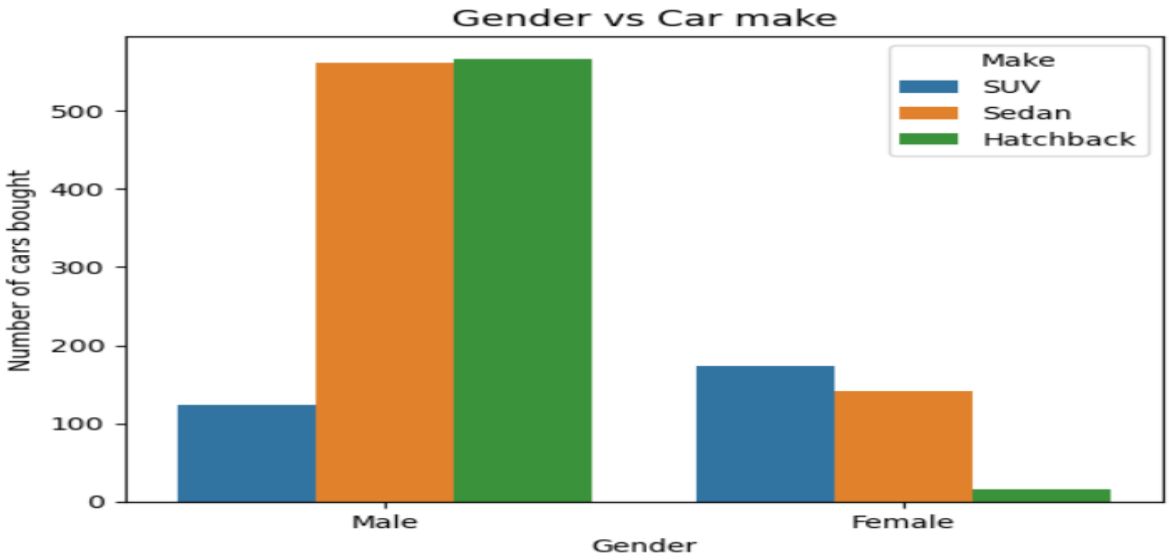


Fig 7: - Countplot of Gender distribution with Make

Male prefers both Sedan and Hatchback whereas female prefers SUV. Numbers of Hatchback bought by female are almost negligible. Also, numbers of SUVs bought by female are more than numbers of SUVs bought by male.

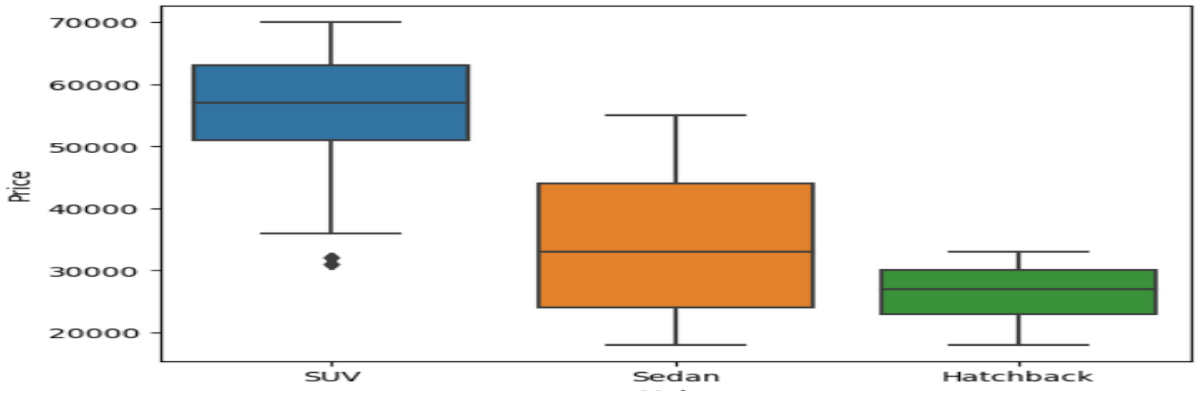


Fig 8: - Boxplot of Price of different make

The above boxplot shows the prices of cars of different Make. SUVs are the costliest and Hatchbacks are the cheapest of car make.

If we talk about car make preference based on different profession, the below plot suggests that salaried persons are buying more number of SUVs as compared to business persons. Sedan is the favourite choice for both business and salaried persons. Number of hatchbacks sold is almost same in both professions.

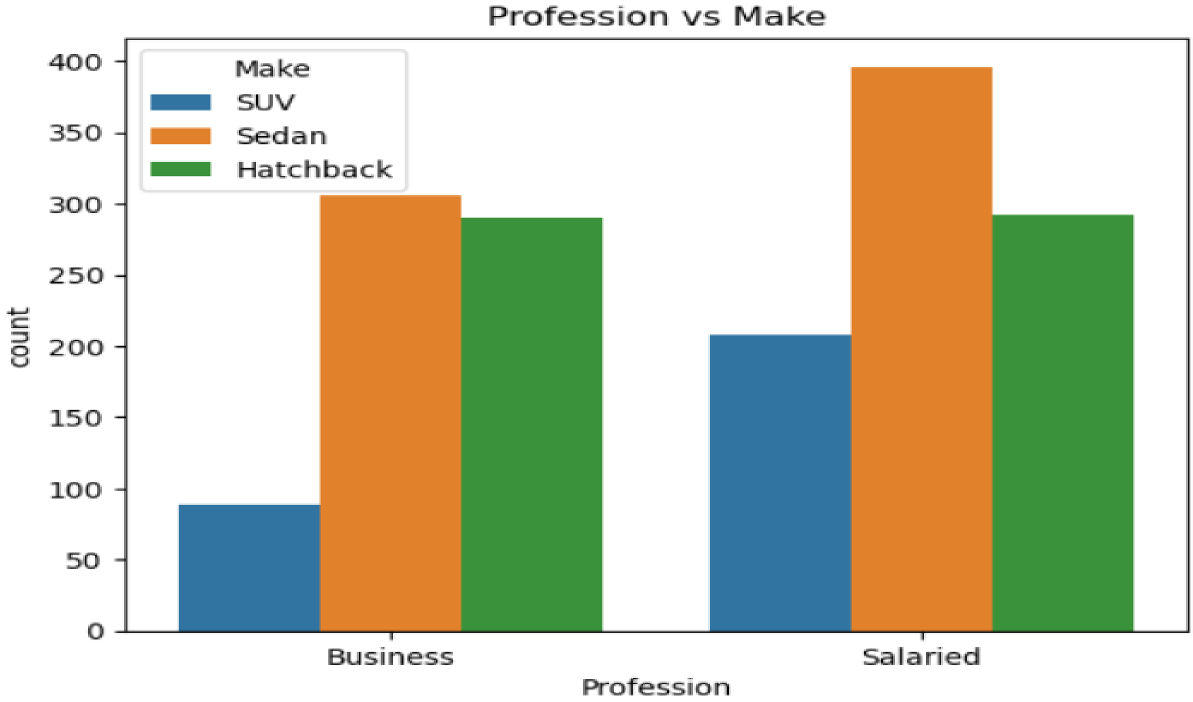


Fig 9: - Countplot of car distribution based on profession and make

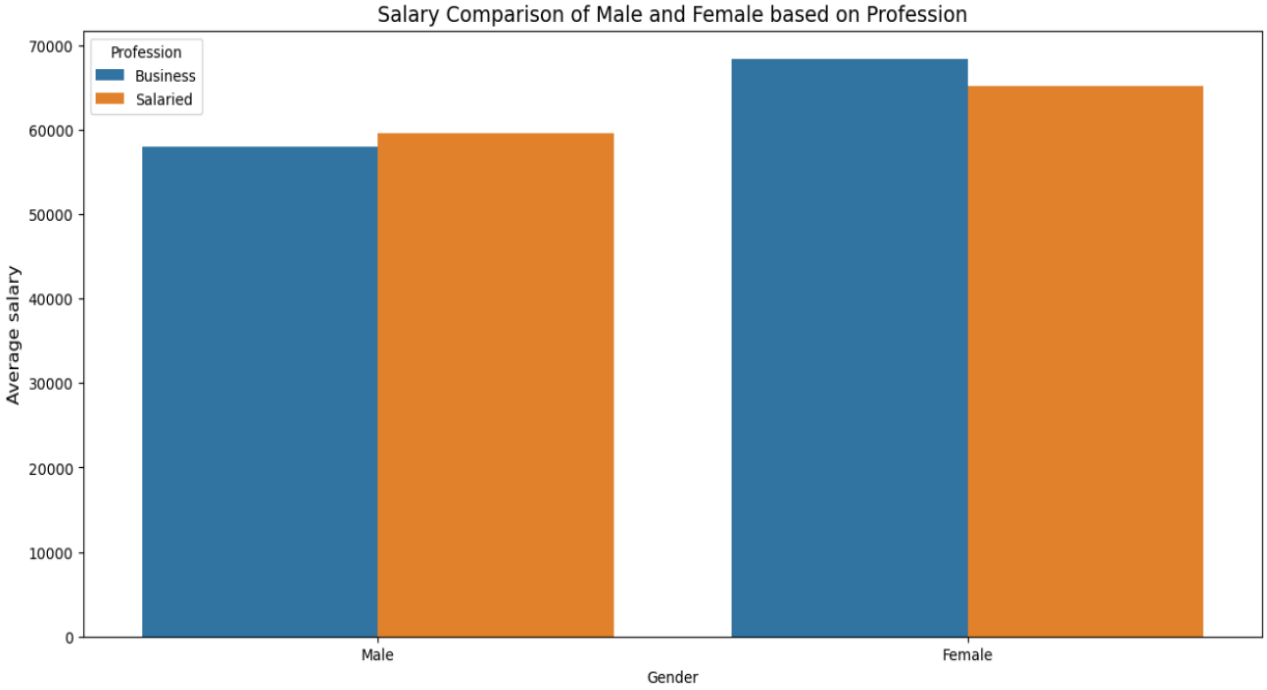


Fig 10: - Barplot of average salary gender wise based on profession

This above plot is very important. This plot tells that female average salary is more than male average salary for both business and salaried persons. This may be one of reasons why females are buying more numbers of SUVs.

To draw below plot, I added new column which was difference of customer's salary and partner salary. Then, I converted positive difference to True and Negative differences to False. Then, I plotted a countplot with Gender as hue. It shows that in 95% of the cases Cars are being bought when customer's salary is greater than partner salary whether it is male or female.

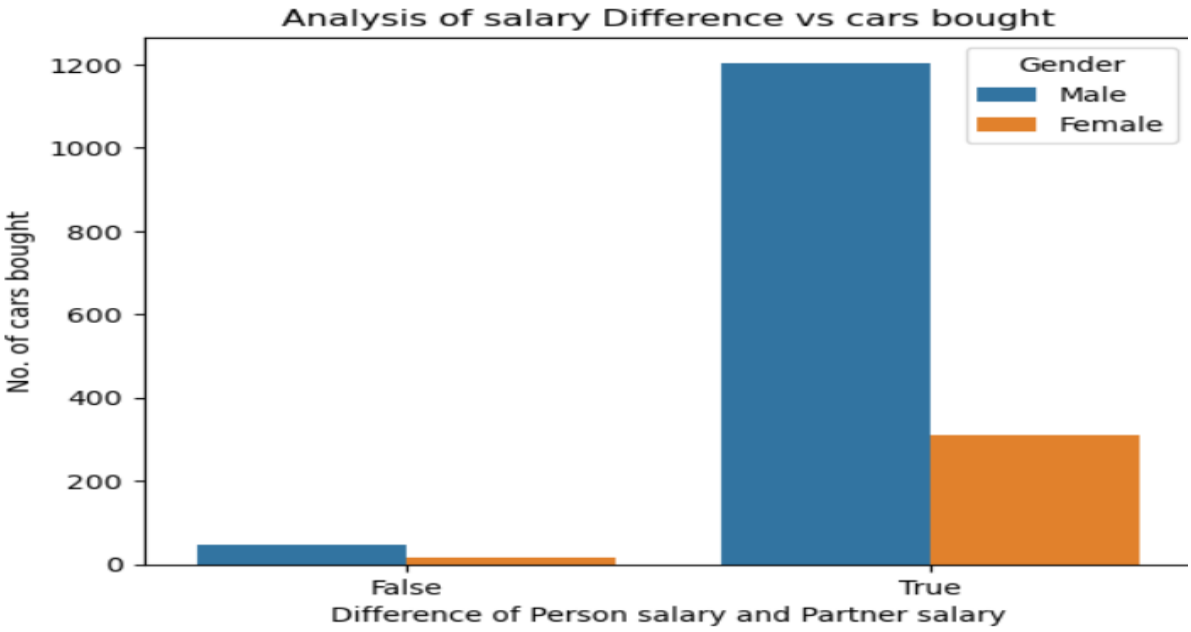


Fig 11: - Countplot of difference of customer and partner salary

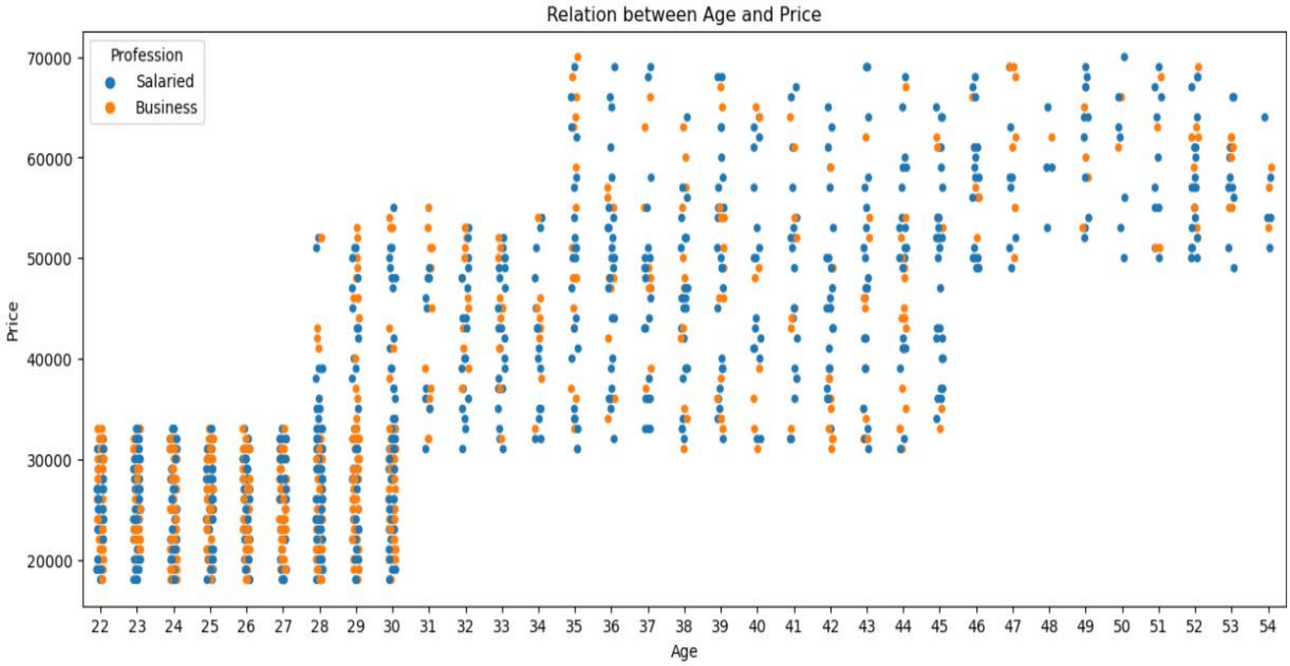


Fig 12: - Strip plot between Price and Age

Fig. 12 shows that with increasing age of customers, prices of cars being bought are also increasing and salaried customers are preferring more to buy high priced cars as compared to business personnel.

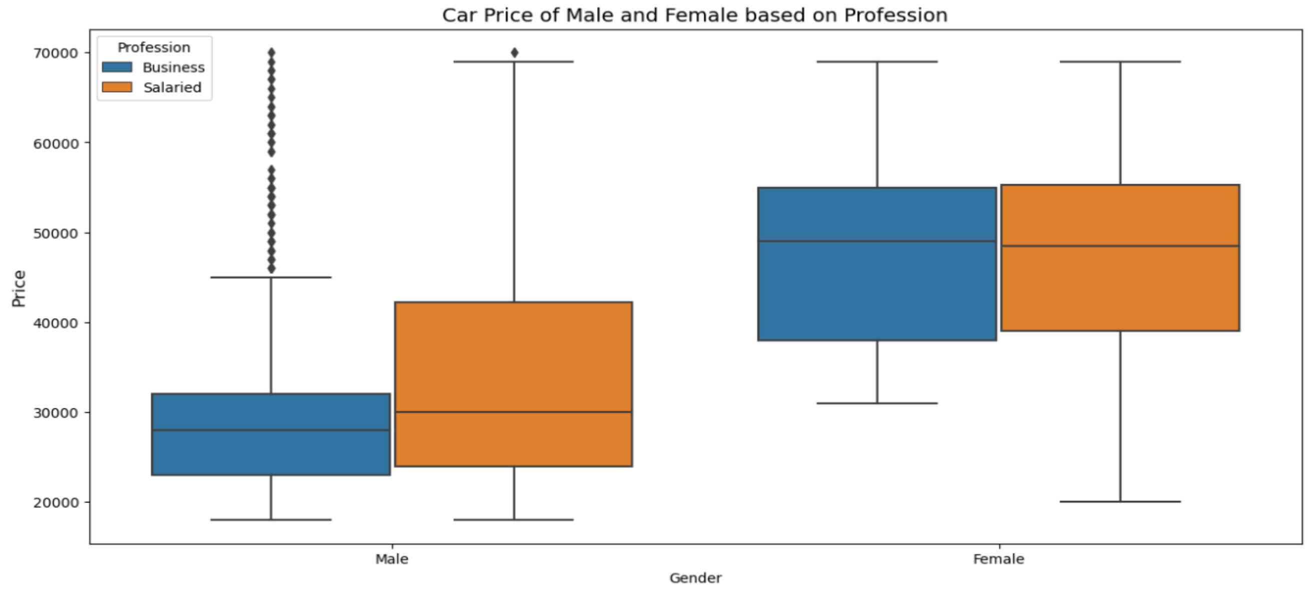


Fig 13: - Boxplot of Price with respect to Gender

Median price of cars being bought by female is much higher than male for both Salaried and Business. For male, salaried and business customers have almost same median price but some of them who are high earning people are buying high priced cars. Automobile company should specifically target female customers to increase sales of high-priced cars.

Problem 1.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”.

Solution: -

Looking at the fig 7 we can clearly say that Women prefer SUV as compared to men. Hence what Steve Roger is telling is wrong.

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

Solution: -

Fig 9 clearly shows that salaried persons are buying more numbers of Sedan than business persons. Hence what Ned Stark believes is true.

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

Solution: -

Fig no. 14 clearly shows that a salaried male is not an easier target for a SUV over a Sedan Sale. Since Sheldon Cooper is not believing in this, hence he is true.

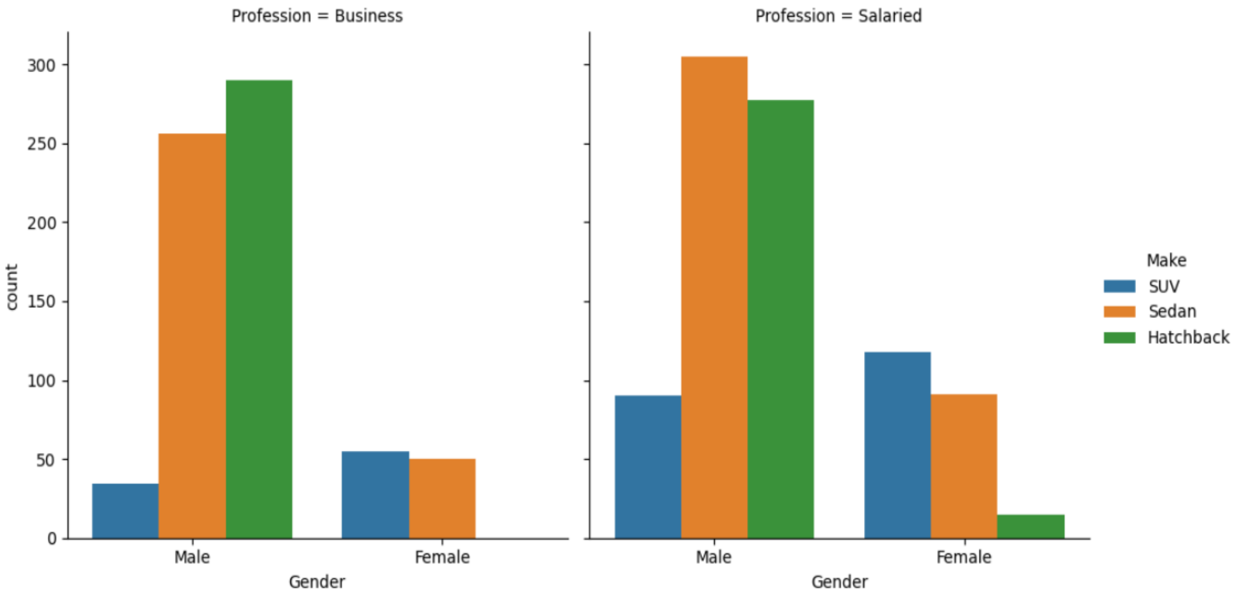


Fig 14: - Plot of distribution of car make gender wise for business and salaried profession

Problem 1.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

Solution: -

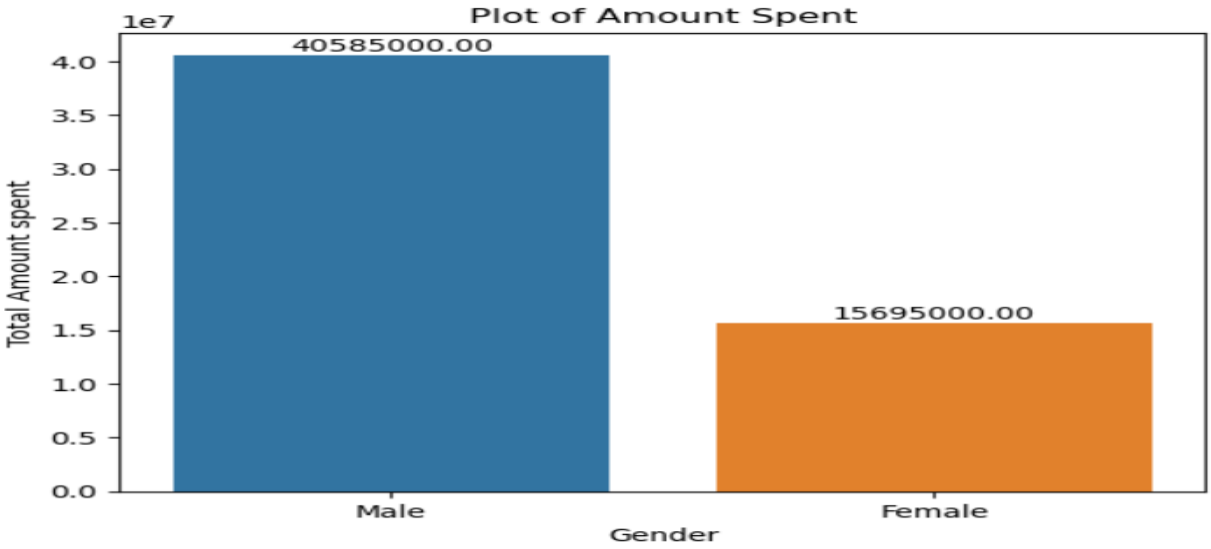


Fig 15: - Gender wise amount spent in buying cars

Total amount spent in buying cars is 56280000 in which 40585000 is spent by male and 15695000 is spent by female.

% of amount spent in buying cars by women= (15695000/56280000) *100 = 27.88%

% of women buying cars= (329/1581) *100= 20.80%

329 women (21% of total) are spending 28% of total amount in car purchasing. This is because women are buying mostly SUVs and their average salary is more than men.

Car company should encourage more women to buy cars to increase sales and hence profit because they are contributing more to sales even if their number is less. Also, women are mostly buying SUVs whose prices are high. Hence SUVs sales will go up and company would capture high-priced cars market.

F2) Personal_loan

Solution: -

Figs 16, 17 and 18 are very important plots of personal loan taken by customers to buy cars.

Total amount of loan taken to buy cars=27290000

% total of loan amount= (27290000/56280000) *100 =48.49 %

Almost 50% of amount that came as sales for car company is from loans taken. Also, one very important information is that out of 527 customers who have home loan on them to pay, 278 customers have taken personal loans to buy cars which is good sign for car company. Car company needs proper advertisements and good marketing strategy to increase sales.

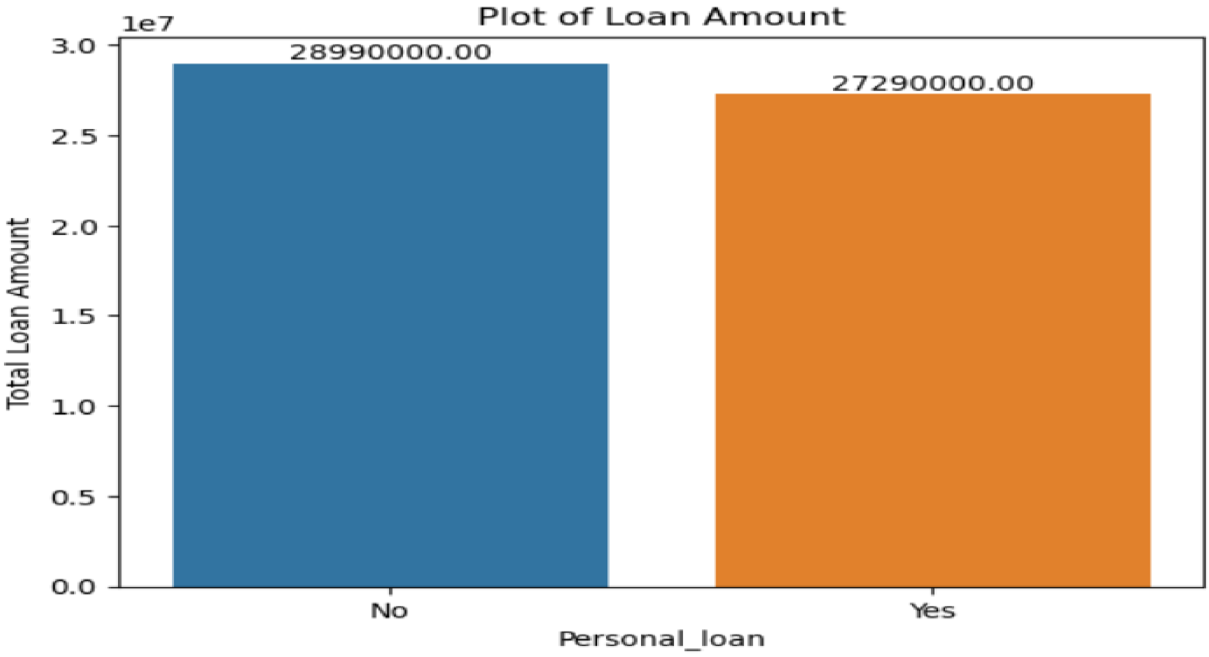


Fig 16: - Plot of total loan amount

1054 of customers do have home loan on them to pay. Out of 1054, only 514 customers have taken personal loans to buy cars.

Customers who have house loan on them pay and also taken personal loan are buying mostly Sedan and customers who do not have house loan on them to pay and also are not taking personal loan are buying Sedan mostly followed by SUVs and Hatchbacks in same quantity.

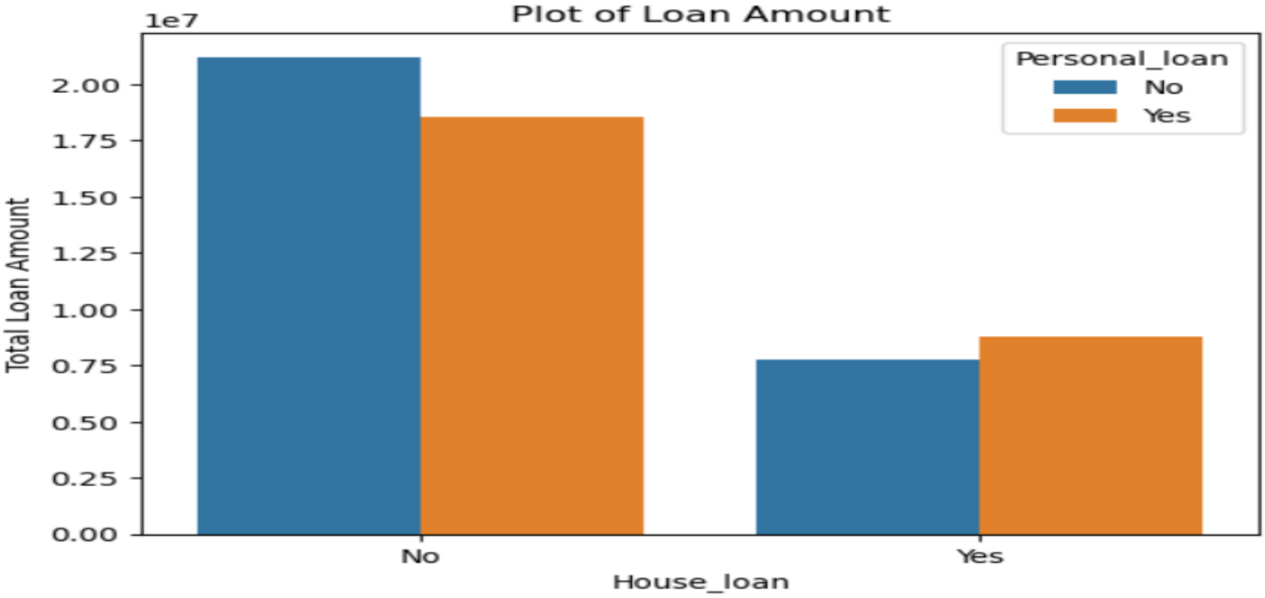


Fig 17: - Plot of total house loan amount vs personal loan

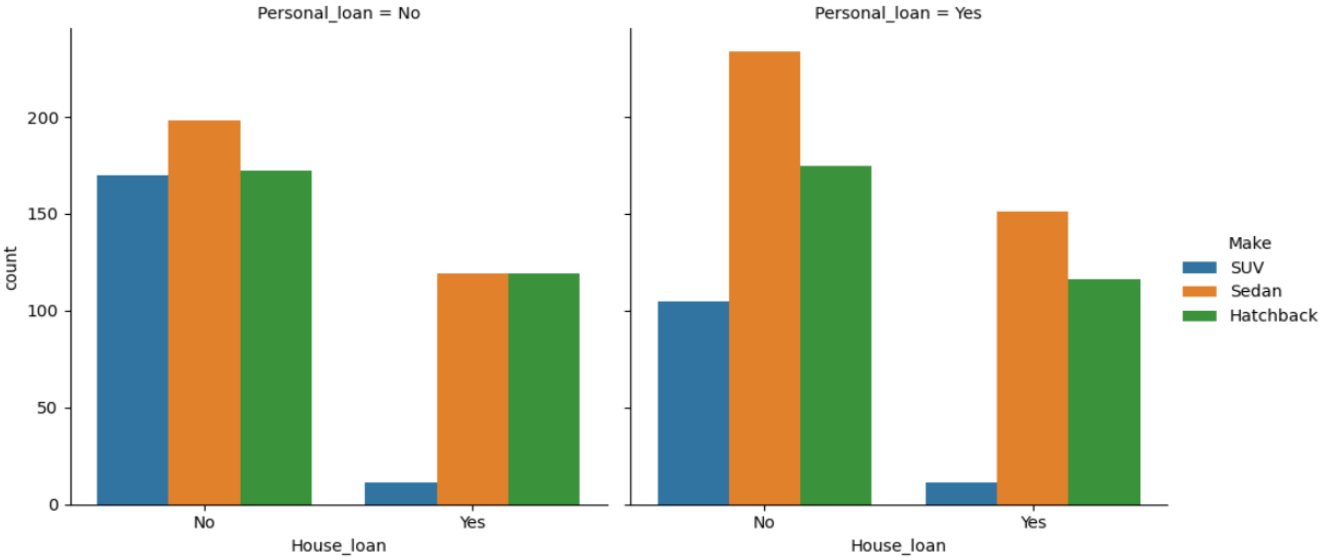


Fig 18: - Plot of Make based on house_loan & personal_loan

Problem 1.G

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

Solution: -

Fig 19 gives clear insights about the type of cars purchased with respect to the working partner. Given dataset contains three types of cars in which SUVs are the costliest and Hatchback are the cheapest. Looking at the plot it is very clear that SUVs sales are not being impacted whether partner is working or not.

But, in case of Sedan, which is medium priced car type, for working partner yes, cars bought are more. Also, there is not much impact of Hatchback car type in both the cases.

So, we can conclude that the working partner yes, leads to purchase of higher-priced car but in very little extent.

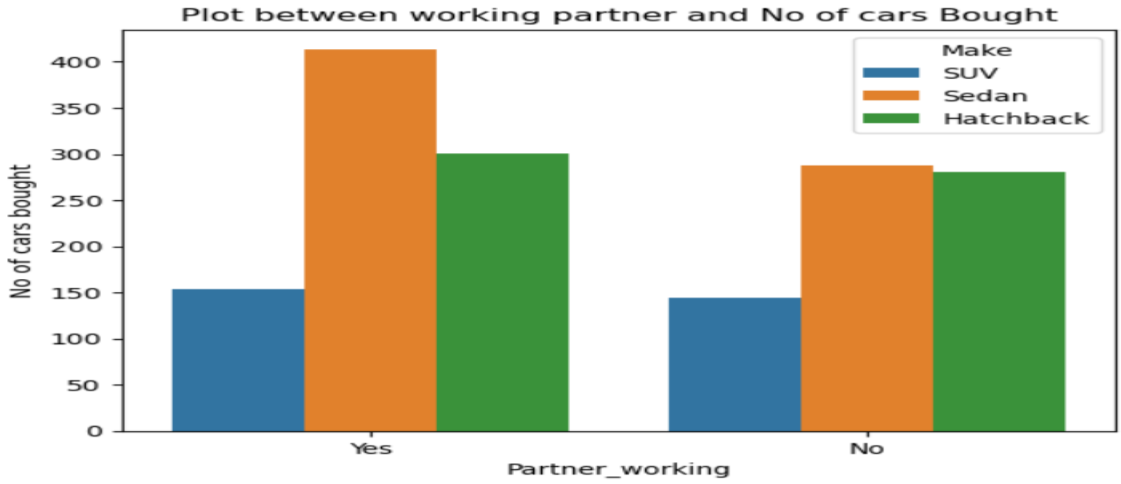


Fig 19: - Plot between working partner and no of cars bought

Problem 1.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.

Solution: -

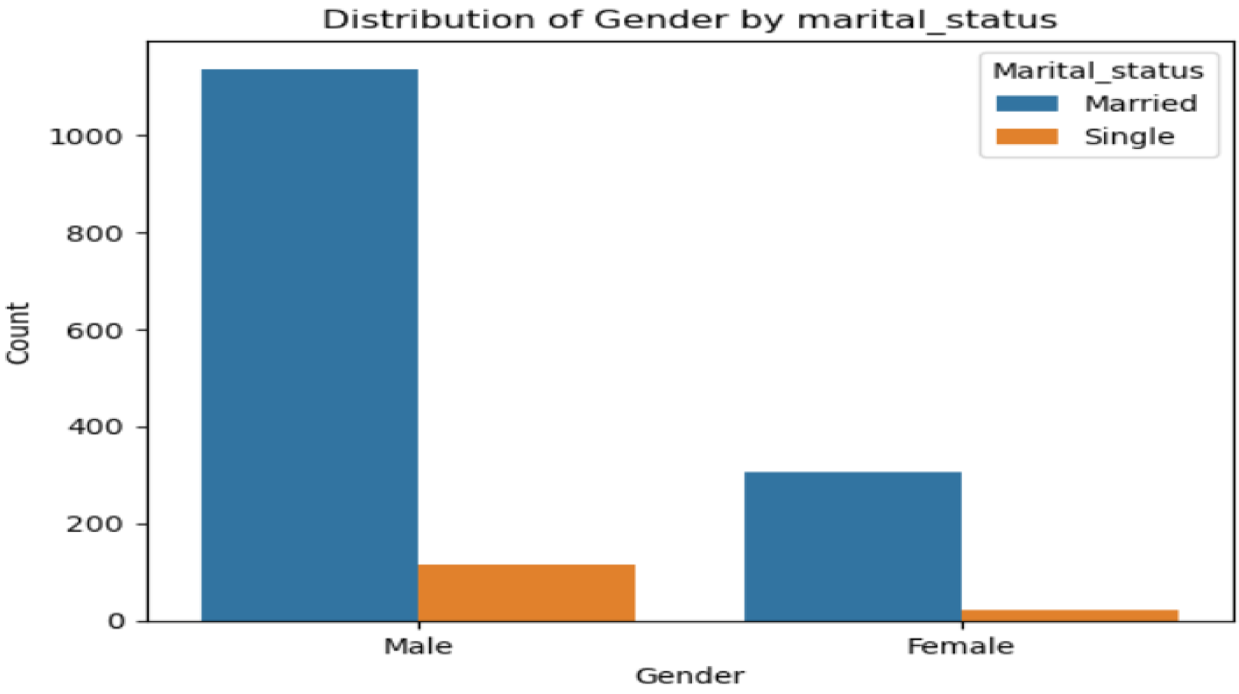


Fig 20: - Distribution of Gender by Marital Status

The above plot gives a clear picture about purchase history of cars sold. Numbers of male who are buying cars is more than three times of numbers of car being bought by female.

```
Gender  Marital_status
Female  Married           307
        Single             22
Male    Married          1136
        Single            116
```

Table 6:- Details of customers based on gender and marital status

% of women who bought cars=((307+22)/(307+22+1136+116))*100 = 20.8 %

% of men who bought cars= 100-20.8= 79.2%

% of married women who bought cars =(307/(307+22))*100= 93.33%

% of married men who bought cars= (1136/(1136+116))*100 = 90.7%

Both married men and married women are buying almost 90 % of cars in male category and female category respectively.

Hence car company should target married men to sell more and more cars and should encourage more women to buy cars. Here are some suggestions for car companies to sell more and more cars.

1. Company should conduct market research to gain insights into the preferences, priorities, and lifestyles of married men and women. Identify their motivations for purchasing a car, such as family needs, transportation convenience, safety, and comfort.

2. To encourage more and more women to buy cars company should open women-oriented showrooms, keep female staffs in the showrooms and provide financial incentives to women.

3. Company staffs should go to male and female social groups and highlight to them about family friendly features of cars, emphasize comfort and conveniences etc.

4. Company should collaborate with influencers or organizations targeting families, partner with influential bloggers, social media influencers, family-oriented organizations to promote cars as suitable choices for married couples. These partnerships can help expand company's reach and credibility among targeted audience.

Problem 2

Introduction

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.)

Data Dictionary

1. Userid: Unique bank customer id

2. card_no: Masked credit card number

3. card_bin_no: Credit card IIN number

4. Issuer: Card network issuer

5. card_type: Credit card type

6. card_source_date: Credit card sourcing date

7. high_networth: Customer category basis their networth value (A: High to E: Low)

8. active_30: Savings/Current/Salary etc account activity in last 30 days (ranges between 0 and 1 with 0 being no activity in account at all and 1 being very high activity in account)

9. active_60: Savings/Current/Salary etc account activity in last 60 days

10. active_90: Savings/Current/Salary etc account activity in last 90 days

11. cc_active30: CC (Credit card) activity in last 30 days

12. cc_active60: CC activity in last 60 days

13. cc_active90: CC activity in last 90 days

14. hotlist_flag: Whether card is hotlisted (Hotlisting is the invalidation of card and putting of restrictions to avoid any unauthorized transactions).

15. widget_products: Number of convenient product customer holds (dc, cc, netbanking active, mobile banking active, wallet active etc). Here in the dataset 0 means customer does not hold any of dc, cc, netbanking active etc. and 7 means customer hold 7 different types of convenient products offered by bank.

16. engagement_products: Number of investment/loan product customer holds (FD, RD, Personal loan, auto loan etc)

17. annual_income_at_source: Annual income recorded in credit card application other_bank_cc_holding: Hold other bank credit card

18. bank_vintage: Vintage (Associated) with the bank (in months) as on Tth month

19. T+1_month_activity: Customer spends in (T+1) month using credit card

20. T+2_month_activity: Customer spends in T+2 month using credit card

21. T+3_month_activity: Customer spends in T+3 month using credit card

22. T+6_month_activity: Customer spends in T+6 month using credit card

23. T+12_month_activity: Customer spends in T+12 month using credit card

24. 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.

25. avg_spends_l3m: Average credit card spends in last 3 months.

26. Occupation_at_source: Occupation recorded at the time of credit card application

27. cc_limit: Current credit card limit

Problem statement

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

Solution: -

A) Basic details of data

Given Dataset contains 8448 rows and 28 columns.

There are 19 continuous columns, 8 categorical columns and 1 datetime column available in the dataset.

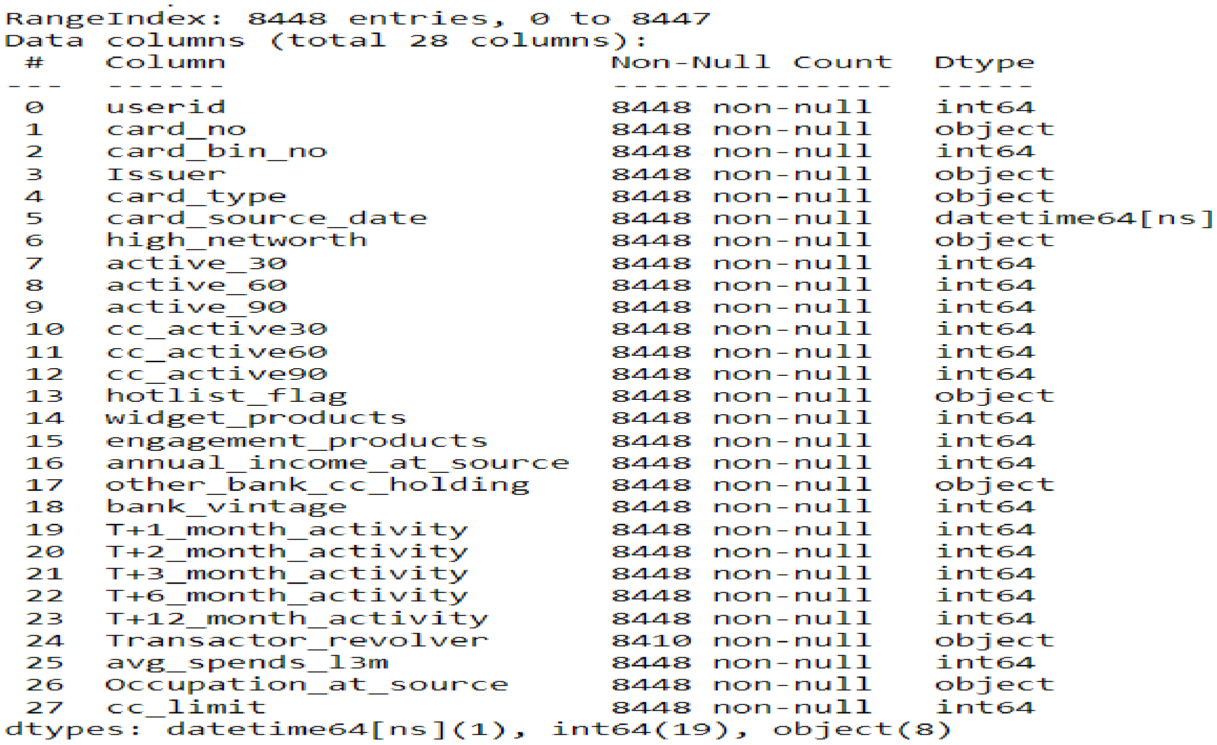
B) Checking of null values with proper treatment

Transactor_revolver column contains 38 missing values. Since missing values number is very very less hence dropping is the best option. After dropping and removing useful columns we have 8410 rows and 24 columns in the dataset.

Also, we can drop userid, card_no, card_bin_no, card_source_date. User id, card_no and card bin no are just unique numbers and do not provide any meaningful insights.

Card source date information is also not useful here for data analysis unless we have to study trend and patterns related to credit card sourcing date.

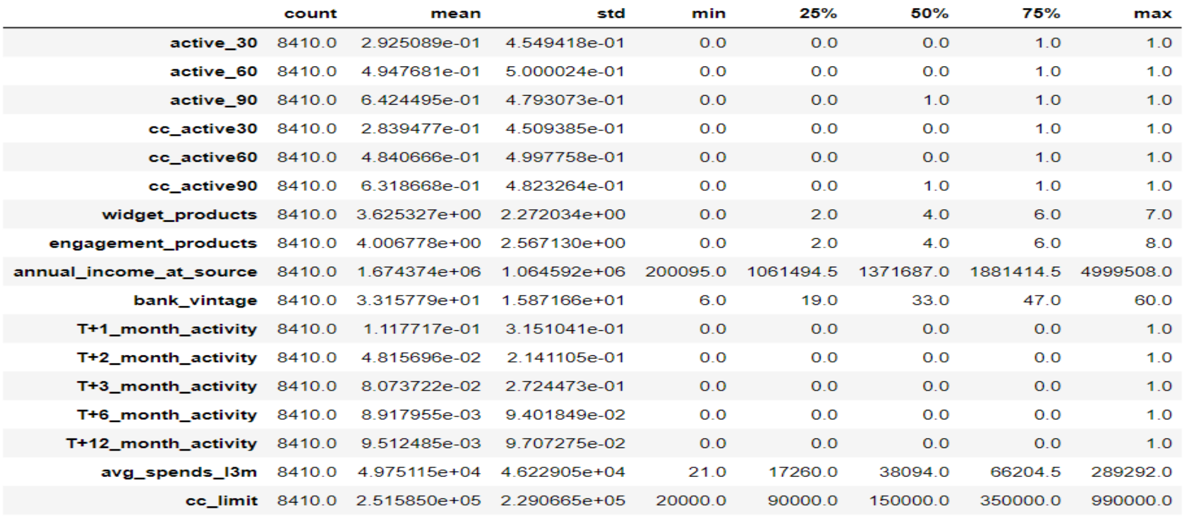
We can also drop hotlist_flag column because it contains only “N” i.e. card being used is not invalidated. Hence every card being used in the given dataset is valid card.



```
RangeIndex: 8448 entries, 0 to 8447
Data columns (total 28 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   userid                    8448 non-null   int64
 1   card_no                   8448 non-null   object
 2   card_bin_no               8448 non-null   int64
 3   Issuer                    8448 non-null   object
 4   card_type                 8448 non-null   object
 5   card_source_date          8448 non-null   datetime64[ns]
 6   high_networth             8448 non-null   object
 7   active_30                 8448 non-null   int64
 8   active_60                 8448 non-null   int64
 9   active_90                 8448 non-null   int64
 10  cc_active30               8448 non-null   int64
 11  cc_active60               8448 non-null   int64
 12  cc_active90               8448 non-null   int64
 13  hotlist_flag              8448 non-null   object
 14  widget_products           8448 non-null   int64
 15  engagement_products       8448 non-null   int64
 16  annual_income_at_source   8448 non-null   int64
 17  other_bank_cc_holding     8448 non-null   object
 18  bank_vintage              8448 non-null   int64
 19  T+1_month_activity        8448 non-null   int64
 20  T+2_month_activity        8448 non-null   int64
 21  T+3_month_activity        8448 non-null   int64
 22  T+6_month_activity        8448 non-null   int64
 23  T+12_month_activity       8448 non-null   int64
 24  Transactor_revolver       8410 non-null   object
 25  avg_spends_l3m            8448 non-null   int64
 26  Occupation_at_source      8448 non-null   object
 27  cc_limit                  8448 non-null   int64
dtypes: datetime64[ns](1), int64(19), object(8)
```

Table 7:- Basic details of dataset

C) Summary of continuous variable in the data



	count	mean	std	min	25%	50%	75%	max
active_30	8410.0	2.925089e-01	4.549418e-01	0.0	0.0	0.0	1.0	1.0
active_60	8410.0	4.947681e-01	5.000024e-01	0.0	0.0	0.0	1.0	1.0
active_90	8410.0	6.424495e-01	4.793073e-01	0.0	0.0	1.0	1.0	1.0
cc_active30	8410.0	2.839477e-01	4.509385e-01	0.0	0.0	0.0	1.0	1.0
cc_active60	8410.0	4.840666e-01	4.997758e-01	0.0	0.0	0.0	1.0	1.0
cc_active90	8410.0	6.318668e-01	4.823264e-01	0.0	0.0	1.0	1.0	1.0
widget_products	8410.0	3.625327e+00	2.272034e+00	0.0	2.0	4.0	6.0	7.0
engagement_products	8410.0	4.006778e+00	2.567130e+00	0.0	2.0	4.0	6.0	8.0
annual_income_at_source	8410.0	1.674374e+06	1.064592e+06	200095.0	1061494.5	1371687.0	1881414.5	4999508.0
bank_vintage	8410.0	3.315779e+01	1.587166e+01	6.0	19.0	33.0	47.0	60.0
T+1_month_activity	8410.0	1.117717e-01	3.151041e-01	0.0	0.0	0.0	0.0	1.0
T+2_month_activity	8410.0	4.815696e-02	2.141105e-01	0.0	0.0	0.0	0.0	1.0
T+3_month_activity	8410.0	8.073722e-02	2.724473e-01	0.0	0.0	0.0	0.0	1.0
T+6_month_activity	8410.0	8.917955e-03	9.401849e-02	0.0	0.0	0.0	0.0	1.0
T+12_month_activity	8410.0	9.512485e-03	9.707275e-02	0.0	0.0	0.0	0.0	1.0
avg_spends_l3m	8410.0	4.975115e+04	4.622905e+04	21.0	17260.0	38094.0	66204.5	289292.0
cc_limit	8410.0	2.515850e+05	2.290665e+05	20000.0	90000.0	150000.0	350000.0	990000.0

Table 8:- Summary of numerical features

The above table gives the summary of numerical features present in the datasets. There are not any discrepancies present in the data.

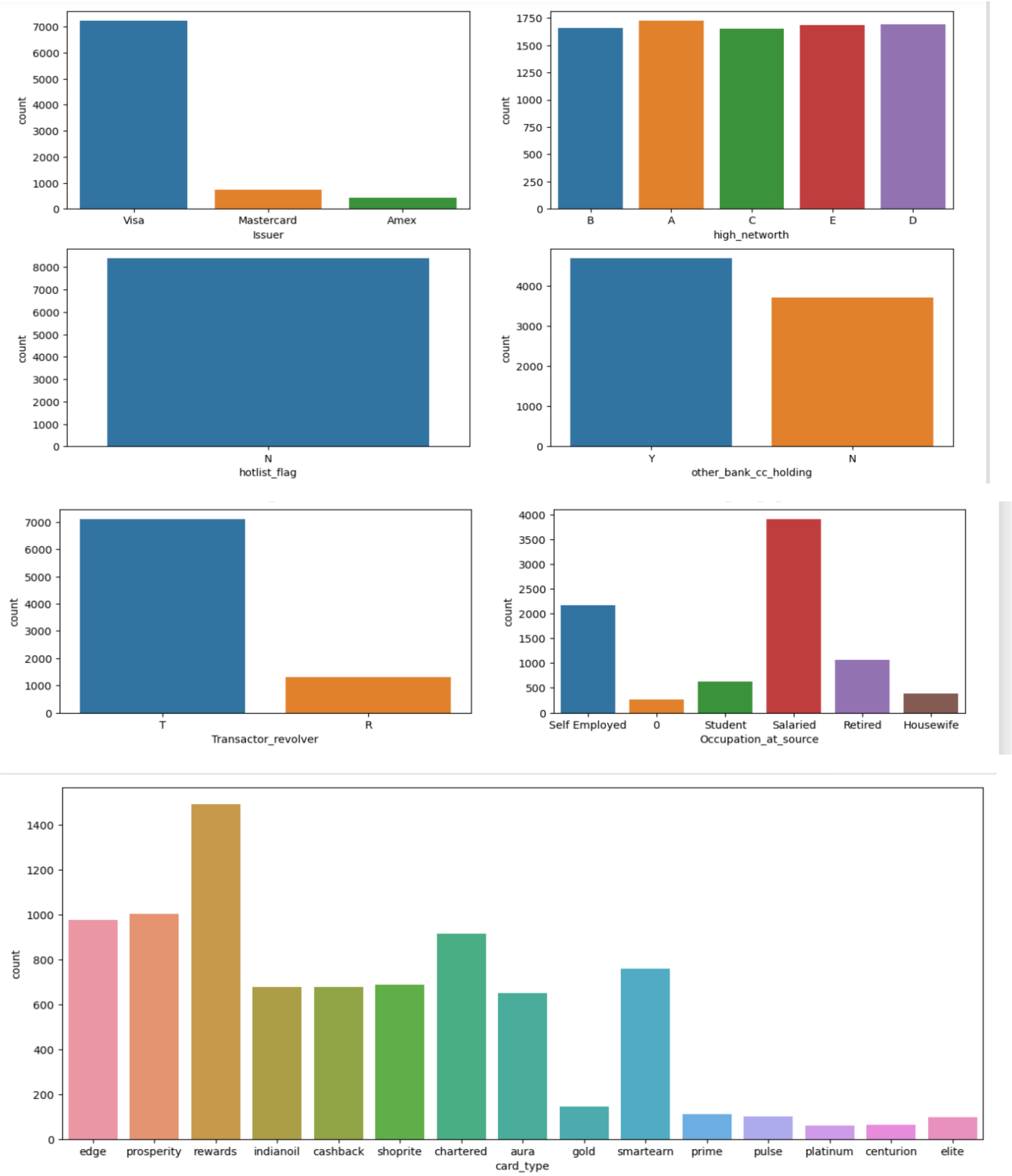


Fig 21: - Countplot of categorical features

There is not any discrepancy present in dataset except in Occupation_at_source. In Occupation_at_source column, there is presence of 0 which is a clear discrepancy. 0 needs to be coupled with Salaried because Salaried is the highest occurring in this column.

D) Checking for outliers

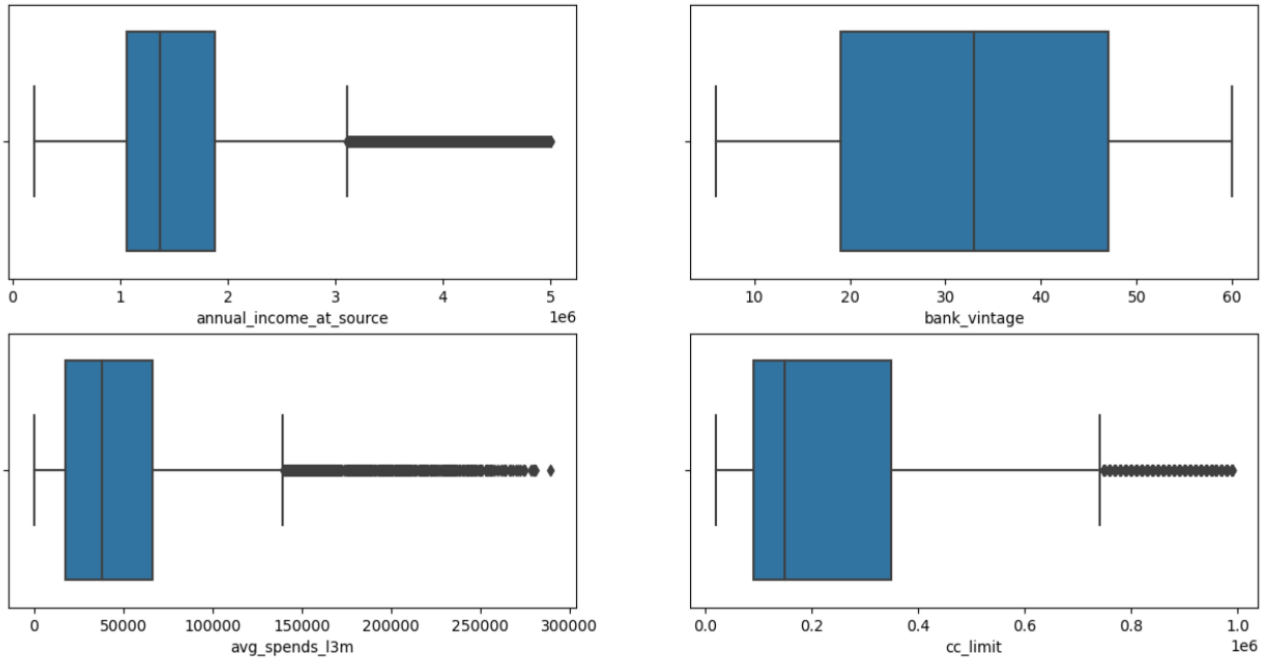


Fig 22: - Boxplot of important continuous features

Annual_income_at_source, avg_spends_l3m and cc_limit are the columns which contain outliers, but outliers' removal will not make any sense here.

E) Selection of top 5 variables using statistical analysis

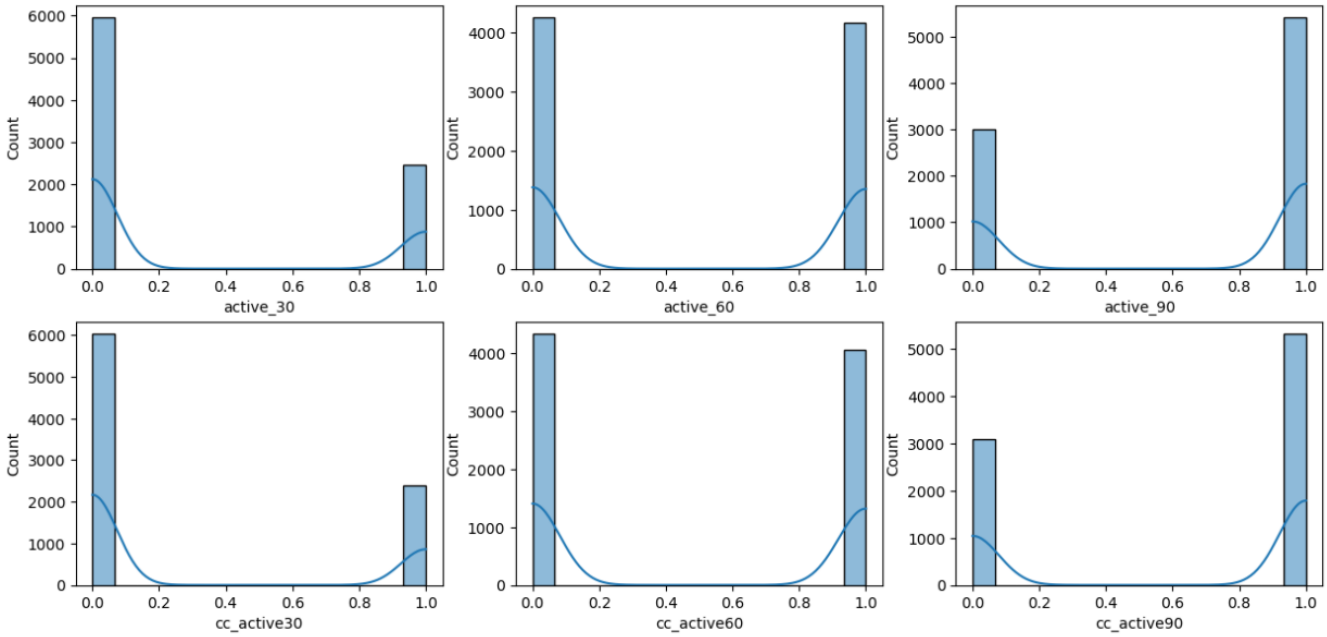


Fig 23: - Histogram of customers activity in bank acc. and in credit card

Above plot clearly shows that in first month almost 30% of customers are active in both bank account and in credit card and in third month almost 65% of customers have been active in both cases. Hence both active_90 and cc_active90 are important features which clearly identify the list of customers who can attrite in coming months.

Hence , suitable schemes and communication can be established with those customers who are not active in bank account transactions and in credit card transactions.

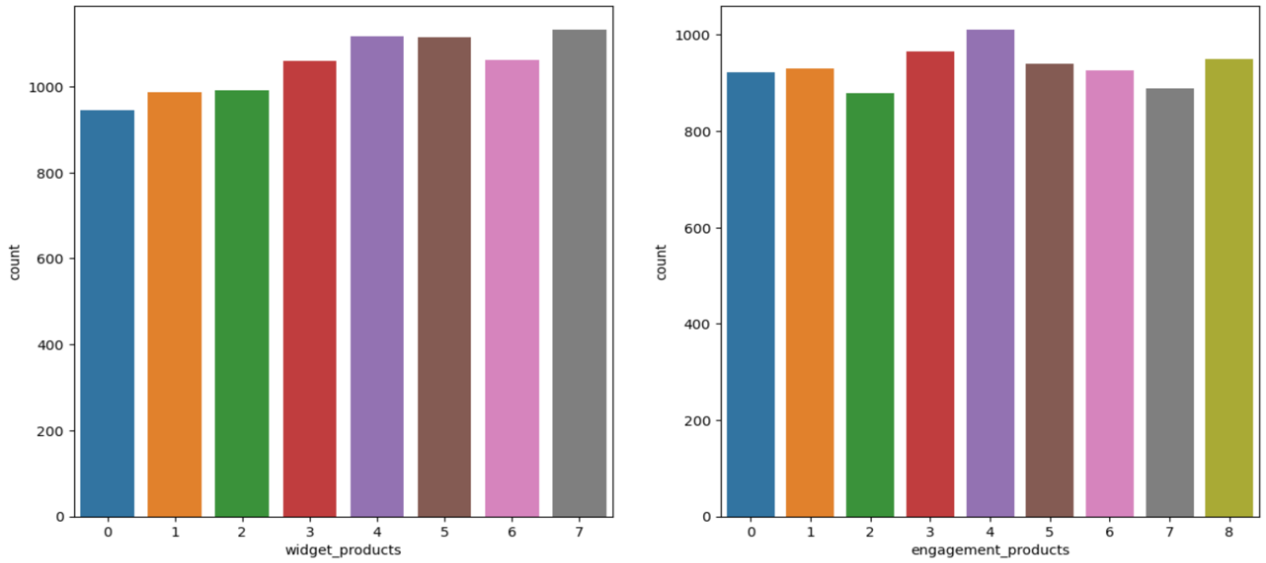


Fig 24: - Plot of distribution of widget_products and engagement_products

Plot no. 24 gives details of widget_products (credit card, debit card, net banking active, mobile banking active etc.) and engagement_products (Fixed deposit, recurring deposits, any loan etc). Customers having more numbers of widget_products and engagement_products are high likely to be associated with bank for longer duration and can maximise bank's profits. Hence these two features are very important.

Annual_income_at_source is important feature. Bank should hold customers having high earning to gain more profits. But customers having high income are very less as compared to customers having medium or less income.

Boxplot of bank_vintage from fig 22 and summary from table 8 clearly indicate that we have customers associated with bank from minimum 6 months to 60 months with mean and median of 33 months which is very uniform.

Fig no 25 is the plot of T+1_month activity, T+2_month_activity, T+3_month activity, T+6_month_activity, T+12_month_activity and Transactor_revolver. This plot gives the transactions which are going to happen in coming future. It gives the customers transactional behaviour in the coming future. Here T+1_month_activity is very important which indicates the customer's spending behaviour in the immediate future.

Transactor_revolver column is also very important feature present in the dataset. This variable categorizes customers into two groups based on their credit card payment behaviour based on customers who pay in full every month and those who carries balances to next month.

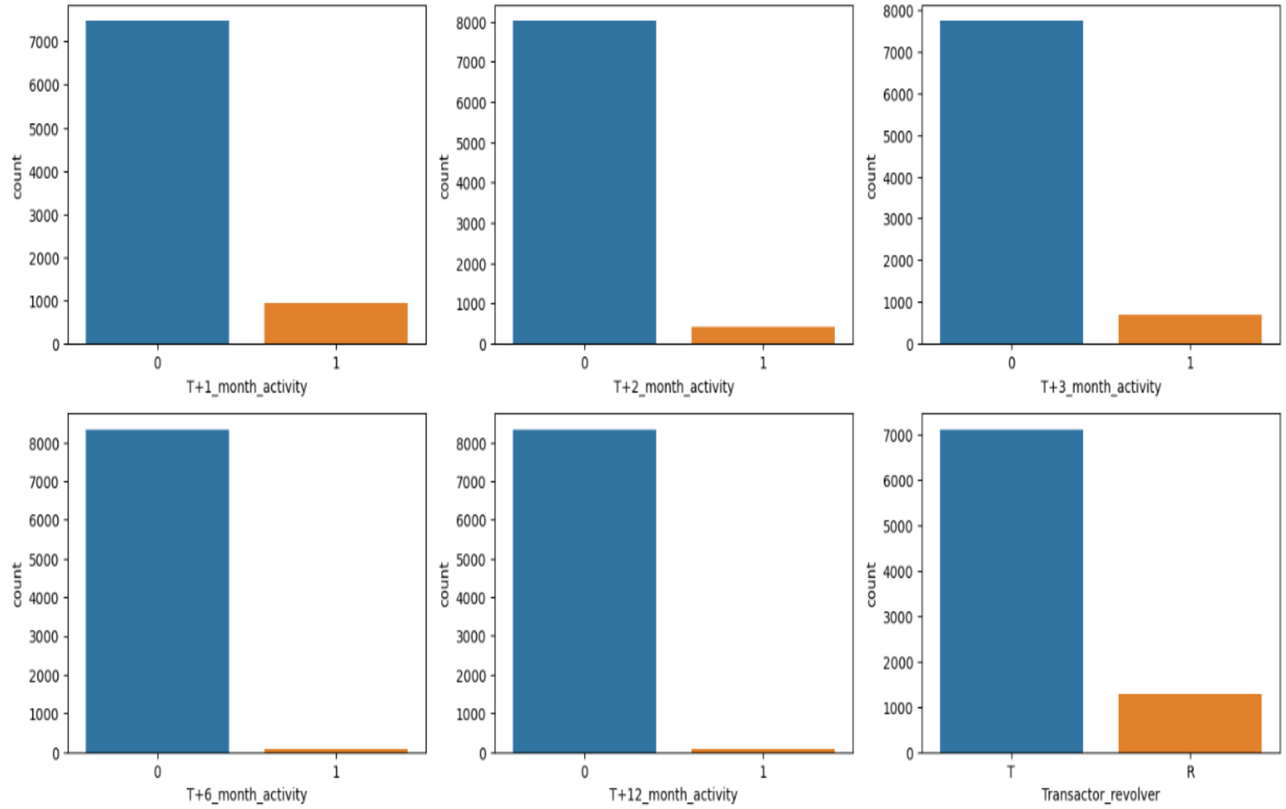


Fig 25: - Plot of distribution of future bank transactions activity

Below plot is the histogram of average credit card spent in last three month and distribution of credit card limit. These two factors are very useful. It shows that maximum customers average credit card spend is between 0 to 100000. Higher average spending is mainly depending on those customers who have higher credit limit.

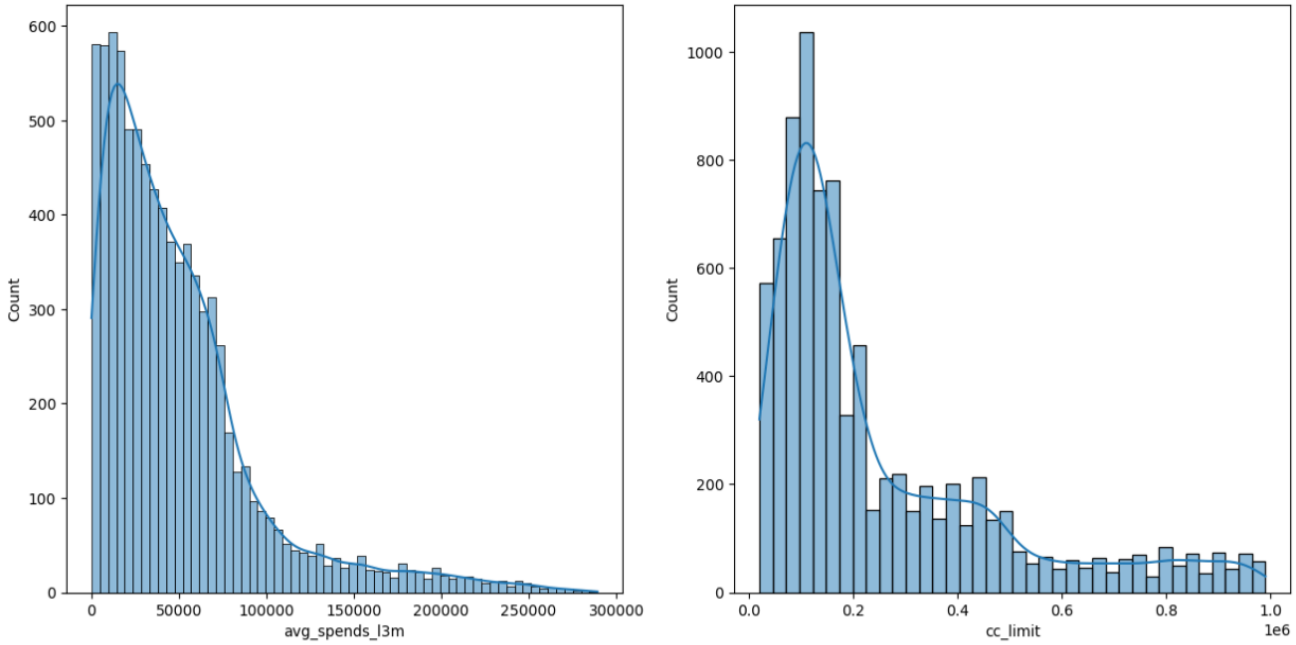


Fig 26: - Plot of distribution of avg_spend_l3m and cc_limit

After analysing all the above feature present in the dataset, here is the list of top five features which are very important to reduce attrition of customers and increase profits.

A) cc_active90:- The credit card activity in the last 90 days can provide insights into customer engagement and usage. Customers who actively use their credit cards indicate higher involvement with the bank's services and products, potentially leading to increased profits. Additionally, customers who are actively using their cards are less likely to churn or attrit.

B) widget_products:- The number of convenient products a customer holds, such as debit cards, credit cards, net banking, mobile banking, and wallet services, indicates the level of engagement and dependency on the bank. Customers who utilize multiple convenient products are more likely to have a stronger relationship with the bank, which can lead to increased profits and lower attrition rates.

C) Transactor_revolver:- Identifying customers who are transactors (paying off balances in full every month) or revolvers (carrying balances over from one month to the next) is crucial for profitability. Transactors tend to generate fewer interest charges for the bank but may have higher customer loyalty and lower attrition rates due to responsible credit card management.

D) avg_spends_l3m:- The average credit card spends in the last 3 months can indicate the level of customer activity and potential profitability. Higher average spends suggest more significant revenue generation for the bank and may also indicate customer satisfaction and engagement. By analysing the average spends, the bank can identify customers who are actively using their credit cards and have a higher intent to spend. Targeting such customers with personalized offers and rewards can encourage them to continue using the credit card and increase their spending.

E) bank_vintage:- The vintage or the association period of the customer with the bank represents the length of the customer's relationship. Longer bank vintages often signify customer loyalty, trust, and stability, which can lead to lower attrition rates. Focusing on retaining long-term customers can contribute to sustained profitability.

These above-mentioned variables capture customer engagement, usage, loyalty, and financial behaviour, all of which can directly impact profitability and attrition rates.

This does not mean that other variables like credit card limit, T+1_month activity, engagements products etc are not important. Most of the variables in the dataset are important and to have complete business insights most of the features are required.