2022

SMDM PROJECT REPORT

DSBA

Girish Chadha

18/09/2022

Contents

List of figures :

EXECUTIVE SUMMARY

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 are able to grasp the insight.

Austo Motor Company is a leading car manufacturer specializing in SUV, Sedan, and Hatchback models. In its recent board meeting, concerns were raised by the members on the efficiency of the marketing campaign currently being used. The board decides to rope in an analytics professional to improve the existing campaign.

1. You as an analyst have been tasked with performing a thorough analysis of the data and coming up with insights to improve the marketing campaign.

Sample of Dataset:

	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary
0	53	Male	Business	Married	Post Graduate	4	No	No	Yes	99300	70700.0	170000
1	53	Female	Salaried	Married	Post Graduate	4	Yes	No	Yes	95500	70300.0	165800
2	53	Female	Salaried	Married	Post Graduate	3	No	No	Yes	97300	60700.0	158000
3	53	Female	Salaried	Married	Graduate	2	Yes	No	Yes	72500	70300.0	142800
4	53	Male	Salaried	Married	Post Graduate	3	No	No	Yes	79700	60200.0	139900
...	...	...	...	...	...	...	...	...	...	...	...	...
1576	22	Male	Salaried	Single	Graduate	2	No	Yes	No	33300	0.0	33300
1577	22	Male	Business	Married	Graduate	4	No	No	No	32000	NaN	32000
1578	22	Male	Business	Single	Graduate	2	No	Yes	No	32900	0.0	32900
1579	22	Male	Business	Married	Graduate	3	Yes	Yes	No	32200	NaN	32200
1580	22	Male	Salaried	Married	Graduate	4	No	No	No	31600	0.0	31600

1581 rows × 14 columns

Introduction to problem 1

From the above dataset we can conclude that the dataset has 1581 rows and 14 columns which has data of salaries of Male & Female Workers of different professions for sale of three car models i.e. Hatchback,Sedan & SUV.

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

Solution:

Importing all the libraries and data into the jupyter notebook

The important technical information about the dataset that a database administrator would be interested in is the size of the dataset and nature of variables.

Size of dataset-
1581 rows and 14 columns

Nature of Variables-

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581 entries, 0 to 1580
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Age               1581 non-null   int64
 1   Gender            1528 non-null   object
 2   Profession        1581 non-null   object
 3   Marital_status    1581 non-null   object
 4   Education         1581 non-null   object
 5   No_of_Dependents  1581 non-null   int64
 6   Personal_loan     1581 non-null   object
 7   House_loan        1581 non-null   object
 8   Partner_working   1581 non-null   object
 9   Salary            1581 non-null   int64
 10  Partner_salary    1475 non-null   float64
 11  Total_salary      1581 non-null   int64
 12  Price             1581 non-null   int64
 13  Make              1581 non-null   object
dtypes: float64(1), int64(5), object(8)
```

There are 8 object data types,5 int64 types & 1 float64 types of variables.

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?

Consistency of data can be viewed in many ways including stability & uniformity.
Typical measures of data consistency include statistics such as the range (i.e., the largest value minus the smallest value among a distribution of data), the variance (i.e., the sum of the squared deviations of each value in a distribution from the mean value in a distribution divided by the number of values in a distribution) and the standard deviation (i.e., the square root of the variance).
If one is evaluating the consistency of data drawn in a sample from a population, the standard error of the mean (i.e., the standard deviation of the sampled population divided by the square root of the sample size) is often examined.

Using the Describe function to show the range & standard deviation of the variables.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	1581.0	NaN	NaN	NaN	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
Gender	1528	4	Male	1199	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Profession	1581	2	Salaried	896	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Marital_status	1581	2	Married	1443	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Education	1581	2	Post Graduate	985	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_Dependents	1581.0	NaN	NaN	NaN	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Personal_loan	1581	2	Yes	792	NaN	NaN	NaN	NaN	NaN	NaN	NaN
House_loan	1581	2	No	1054	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Partner_working	1581	2	Yes	868	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	1581.0	NaN	NaN	NaN	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1475.0	NaN	NaN	NaN	20225.559322	19573.149277	0.0	0.0	25600.0	38300.0	80500.0
Total_salary	1581.0	NaN	NaN	NaN	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	1581.0	NaN	NaN	NaN	35597.72296	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0
Make	1581	3	Sedan	702	NaN	NaN	NaN	NaN	NaN	NaN	NaN

IMPORTANT INSIGHTS :

Range of Age column is 54-22=32 Years

Range of Salary Column is 99300-30000= Rs 69300

Range of Price Column is 70000-18000= Rs 52000

The gender ‘Male’ has the highest number of counts indicating that males buy cars more than females.

The Profession ‘Salaried’ has the highest number of counts indicating that Salaried employees are more interested in buying cars from Austo Motor Company.

Personal loan is preferred by most of the customers.

Company should provide offers on Personal loans so that they can achieve more sales and hence marketing will be more fruitful as Personal loan is preferred by most of the customers.

Sedan make is preferred by most of the customers. Company should provide more attractive offers for Sedan Cars as it's most preferred which will in turn increase their sales and revenue.

More Marketing campaigns of Sedan Cars will lead to high returns.

The Company should also focus on marketing of Hatchback & SUV models to increase their sales in the market.

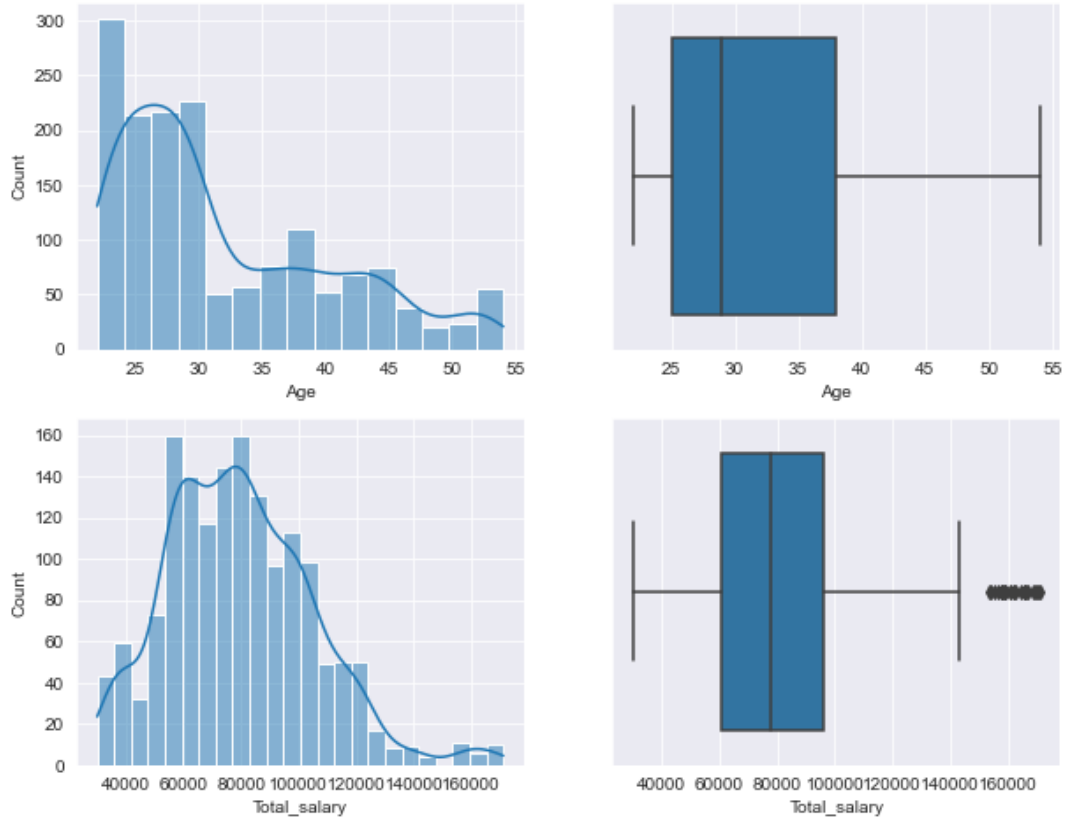
Mean Total salary is Rs 79625 which is higher than the median value(Rs 78000) indicating that the distribution is right tailed.

Standard deviation of Age Column is 8.42 Years.

Standard deviation of Total Salary Column is Rs 25,545.(high)

The Standard deviation(spread of the data) is high for Total salary indicating the presence of outliers.

Using the histogram & boxplot to visualise some of the variables & show presence of outliers.



Age & Total salary Histplot & Boxplot(Figure 1)

The total salary column has many outliers as shown in the box plot.

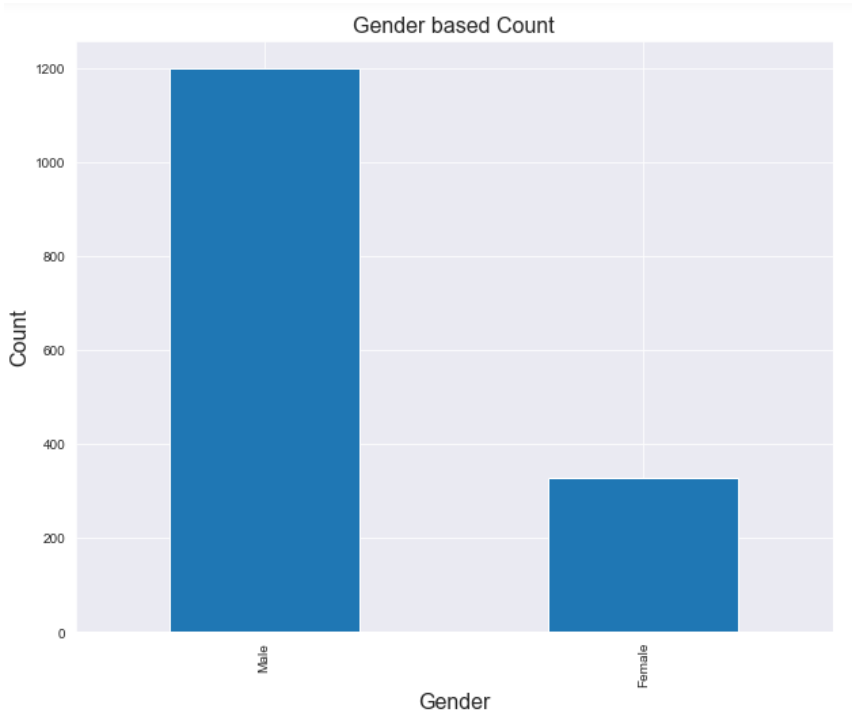
Discrepancies present in the data

1. The Gender ‘Female’ is misspelt as Femal & Femle in the data. Using the str.replace function, the data is replaced by Female.
2. Using the isnull function there are 53 missing values in 'Gender' & 106 missing values in 'Partner Salary' column.
3. We can find the missing entries in Partner Salary column by finding out the difference between Total Salary and Salary Column.

```
Age                  0
Gender              53
Profession           0
Marital_status       0
Education            0
No_of_Dependents     0
Personal_loan        0
House_loan           0
Partner_working      0
Salary               0
Partner_salary     106
Total_salary         0
Price                0
Make                 0
```

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

Using the bar plot it is clear that the Male Gender Customers are more as compared to Female customers.

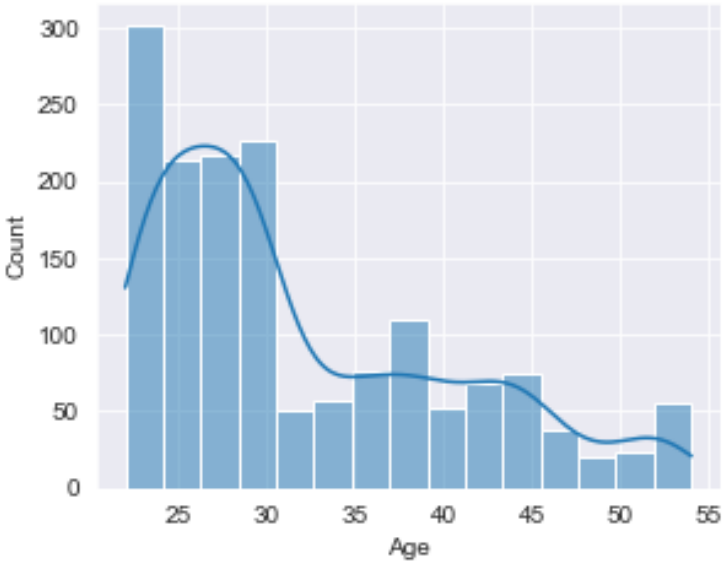


Gender Based Count(Figure 2)

Using the histplot on Age, it can be observed that the most buyers are from the age group of 25 to 30 years indicating that company are targeting more on young adults rather they should target more on the buyers of age group of more than 30 years to earn good amount of revenue.

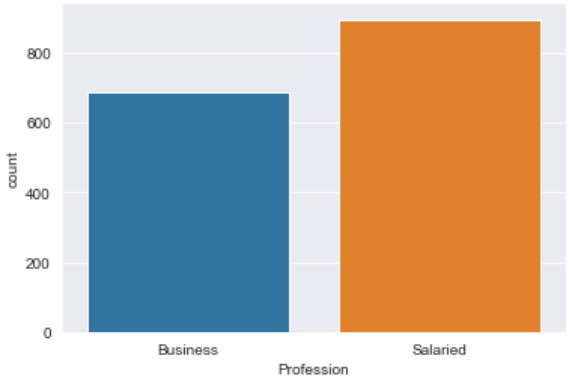
To improve the efficiency of the Marketing Campaign, old adults should be targeted more. Only 39 years of age seems to have higher frequency in older adults.

Marketing Ads should target older adults as older adults seem to have higher salary which will in turn generate more revenue for the company.



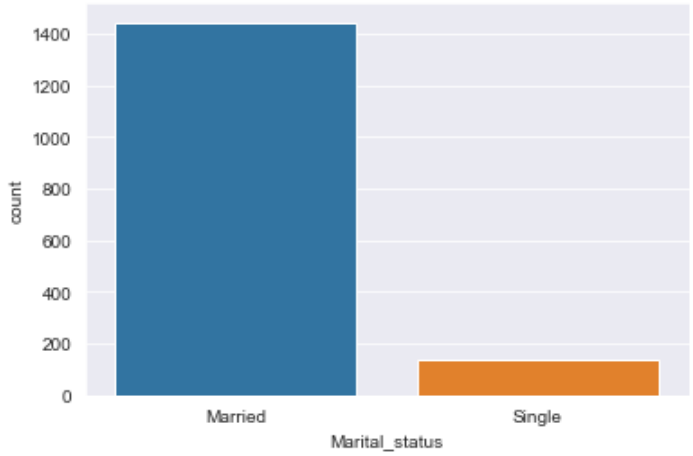
Age Based Count(Figure 3)

Salaried people are more as compared to business owners. Company should give more offers to business owners to increase their sale.



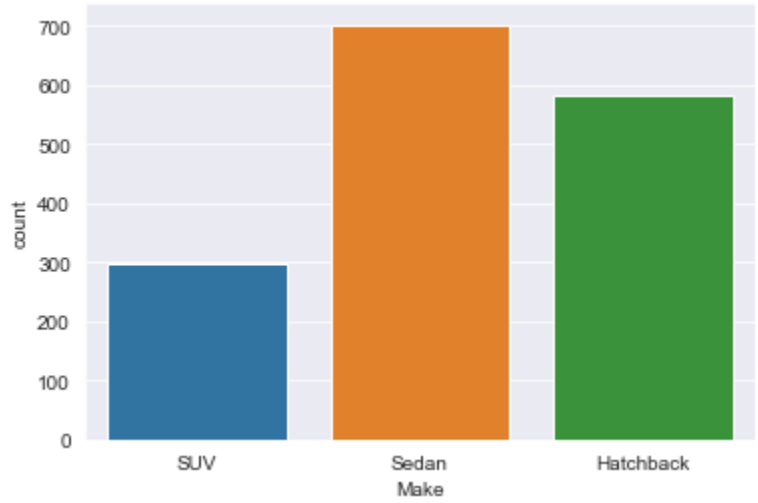
Profession Based Count(Figure 4)

Married Couples are more as compared to single individuals. More offers for single people can increase their sales to generate more revenue.



Marital Status (Figure 5)

Sedan Cars selling rate is high as compared to SUV & Hatchback models. Marketing campaigns for SUV cars should be targeted to increase their sales.



Make Based Count(Figure 6)

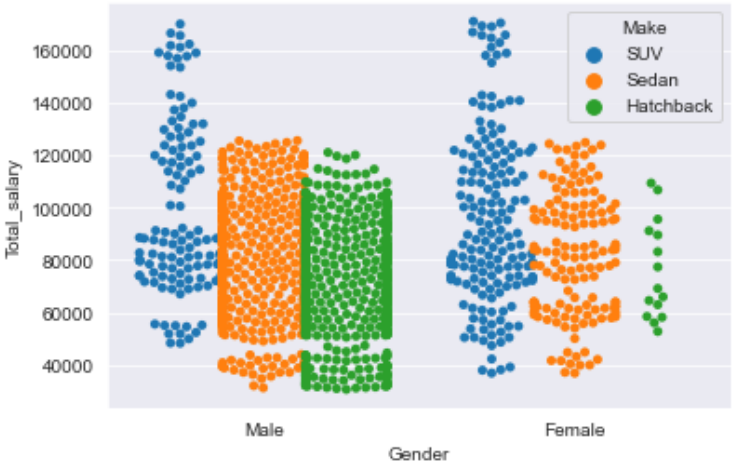
The salary of age group 50 to 55 years is higher and they prefer more SUV cars .

Company should focus more on manufacturing luxury sedan cars to increase their sales.

Sedan Cars & Hatchback cars marketing should be focused more to increase their sale among old adults.

Salary, Age & Make based Line Plot(Figure 7)

More females are buying SUV models. Male prefer more Sedan & Hatchback models and Customers with high salary prefer SUV models.Very few females buy Hatchback models. Company should focus more on manufacturing hatchback models more attractive for females to increase their spread . More marketing campaign should be launched to target females.



Salary, Gender & Make based Swarm Plot(Figure 8)

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.

To find relationships among variables we can use scatterplot, pairplot and heatmap to show if there's a positive, negative or no correlation between the variables.

- In a scatter plot when the y variable tends to increase as the x variable increases, we say there is a **positive correlation** between the variables.
- Again, when the y variable tends to decrease as the x variable increases, we say there is a **negative correlation** between the variables.
- If the points on the scatter plot seem to be scattered randomly, we say that there is **no correlation** between the variables.

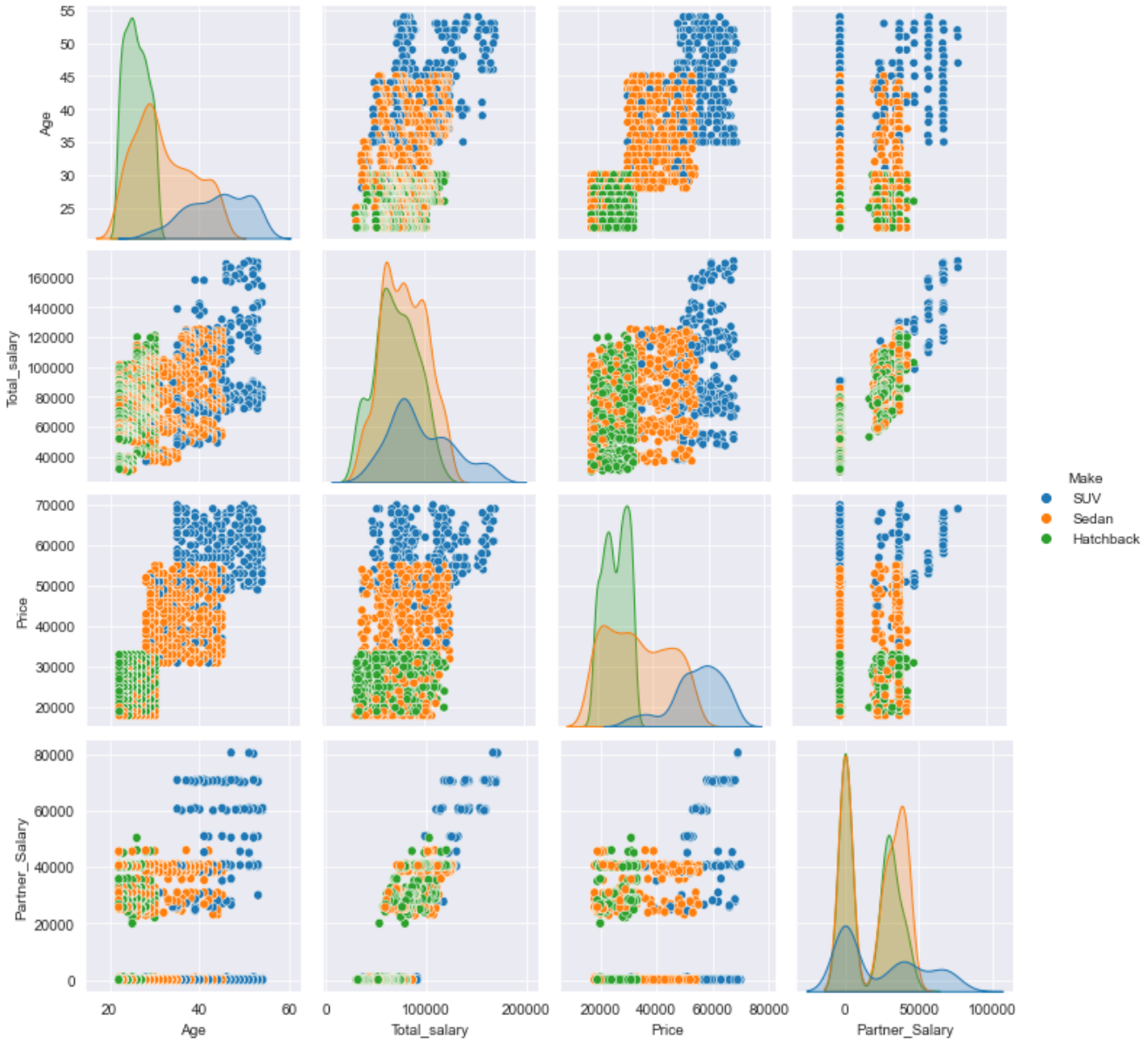
Price and Age show a high positive relationship indicating that as age increases, buyers are more likely to buy high priced cars.

Price & Age Scatterplot(Figure 9)

Total Salary and Age seem to show a low positive correlation.

Total Salary & Age Scatterplot(Fig 10)

The pairplot among Age,Total salary,Price & Partner salary shows us various scatterplot with each numerical column. Partner Salary & age seem to show close to no correlation which is also the case of price & partner salary.



Pairplot(Figure 11)

A **heatmap** is a graphical representation of data as a color-encoded matrix.

It is a great way of representing the correlation for each pair of columns in the data.

The *heatmap()* function of seaborn helps us to create such a plot.

We can see a high positive correlation between Age and Price, Partner salary and Total Salary variables .

Heatmap (Figure 12)

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

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

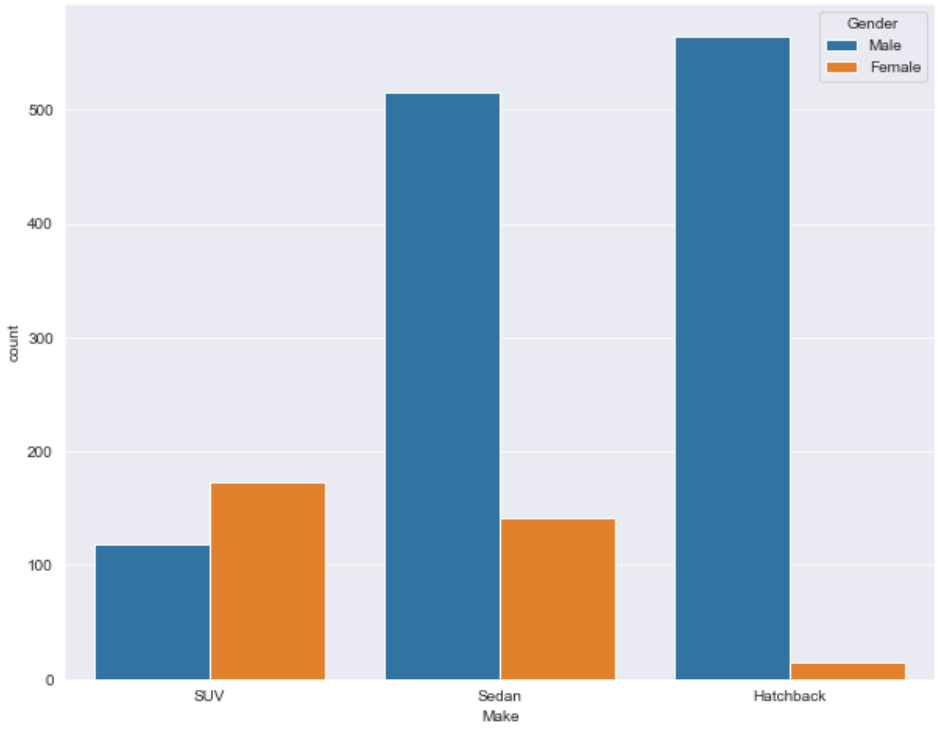
E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.

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

Solution:

E1) We can use groupby to show Gender & Make Counts. Below Countplot shows that Steve Roger observation is not correct. Female actually prefer SUV more as compared to men.

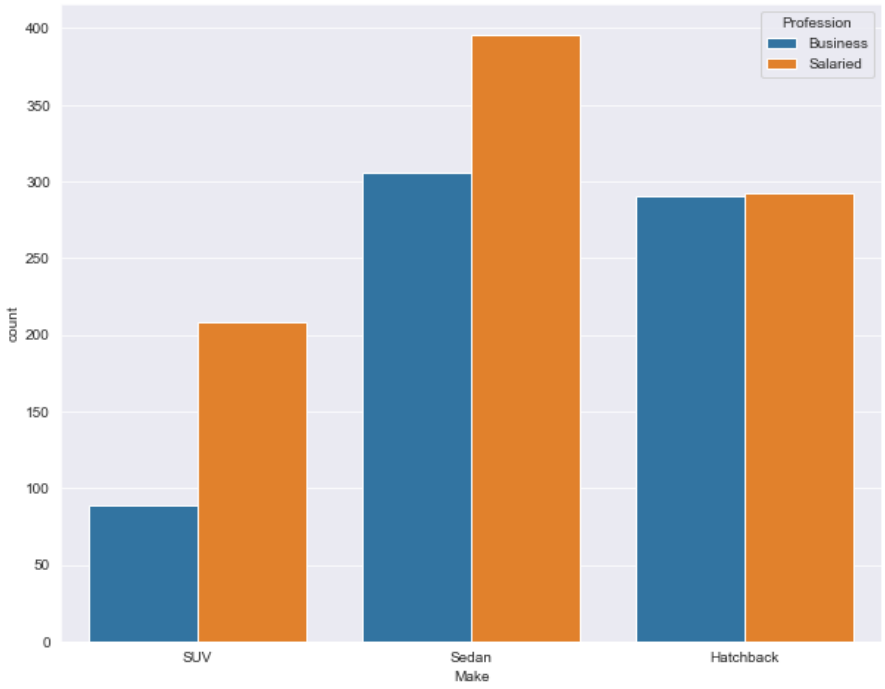
```
Gender  Make
Female  Hatchback     15
        SUV          173
        Sedan        141
Male    Hatchback    565
        SUV          118
        Sedan        516
dtype: int64
```



Gender & Make Countplot(Figure 13)

E2)This observation is true by Ned Stark as can be seen from the countplot that salaried people are more inclined towards buying sedan cars.

```
Profession  Make
Business    Hatchback    290
            SUV           89
            Sedan        306
Salaried    Hatchback    292
            SUV          208
            Sedan        396
dtype: int64
```



Profession & Make based countplot(Figure 14)

E3) We can use groupby on profession,Make and Gender Columns to arrive at the conclusions.

```
Profession  Gender  Make
Business    Female  SUV          55
                    Sedan        50
            Male    Hatchback   289
                    SUV          33
                    Sedan       237
Salaried    Female  Hatchback    15
                    SUV         118
                    Sedan        91
            Male    Hatchback   276
                    SUV          85
                    Sedan       279
dtype: int64
```

The observation is not correct as Salaried male is not an easier target for a SUV sale rather salaried male is an easier target for a Sedan Car sale.

F. From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a Business can utilise the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions.

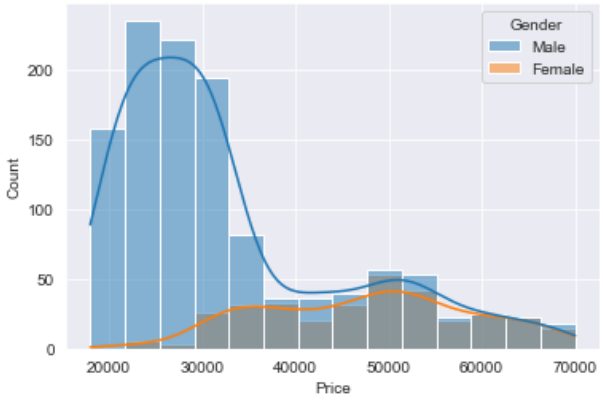
Give justification along with presenting metrics/charts used for arriving at the conclusions.

F1) Gender

F2) Personal_loan

Solution:

F1) Price and Gender distribution seems to be right tailed as shown in the following histplot.Male spends more on purchasing automobiles as compared to females. However, Male tend to spend more on cars that are less expensive i.e. within the price range of Rs 20000 to Rs 30000.Business can focus more on female buyers.



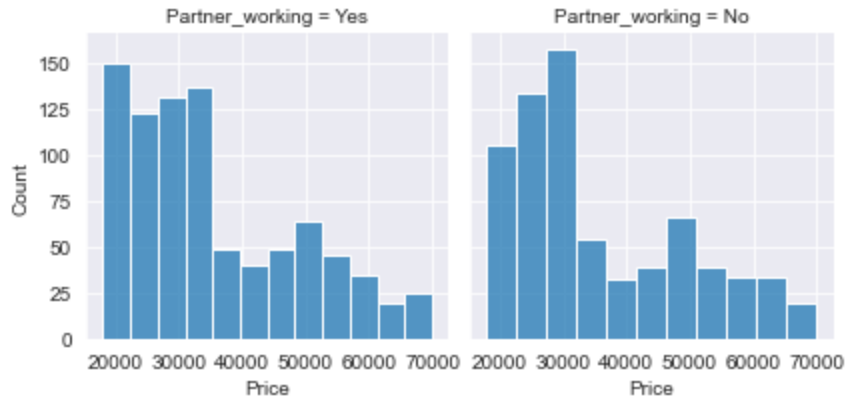
Gender & Price Histplot(Figure 15)

F2) People who take personal loans are more likely to buy less expensive cars as can be seen from the following histplot. However some of the buyers who take personal loan buy high priced cars in the range of Rs40000 to Rs50000. Business can focus more providing personal loan for high priced cars.

Personal loan and price histplot(Fig 16)

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

As can be seen from the below graph that having a working partner is not leading to purchase of a high priced car as both situations show almost same histplot.



Price & partner working hisplot(Fig 17)

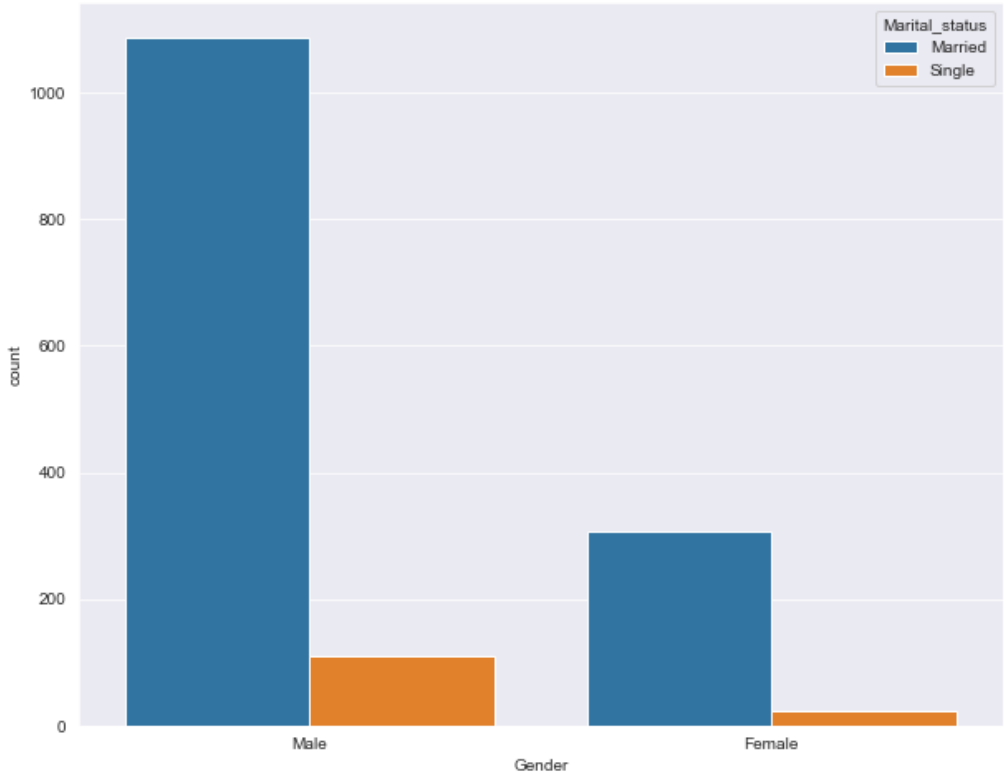
H. The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use the Gender and Marital_status - fields to arrive at groups with similar purchase history.

Solution:

Married males are tending to show greater purchase history as compared to females.

To improve business marketing strategy more offers for married couples should be made available. Company should also focus on increasing their sales among singles which tend to be low as compared to married couples.

```
Gender  Marital_status
Female  Married          307
        Single            22
Male    Married          1088
        Single            111
dtype: int64
```



Gender & Marital status based countplot(fig 18)

As can be seen from the countplot, married males are more as compared to Single males.

EXECUTIVE SUMMARY

Problem 2

A bank can generate revenue in a variety of ways, such as charging interest, transaction fees and financial advice. Interest charged on the capital that the bank lends out to customers has historically been the most significant method of revenue generation. The bank earns profits from the difference between the interest rates it pays on deposits and other sources of funds, and the interest rates it charges on the loans it gives out.

GODIGT Bank is a mid-sized private bank that deals in all kinds of banking products, such as savings accounts, current accounts, investment products, etc. among other offerings. The bank also cross-sells asset products to its existing customers through personal loans, auto loans, business loans, etc., and to do so they use various communication methods including cold calling, e-mails, recommendations on the net banking, mobile banking, etc.

GODIGT Bank also has a set of customers who were given credit cards based on risk policy and customer category class but due to huge competition in the credit card market, the bank is observing high attrition in credit card spending. The bank makes money only if customers spend more on credit cards. Given the attrition, the Bank wants to revisit its credit card policy and make sure that the card given to the customer is the right credit card. The bank will make a profit only through the customers that show higher intent towards a recommended credit card. (Higher intent means consumers would want to use the card and hence not be attrite.)

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

Sample of Dataset:

	userid	card_no	card_bin_no	Issuer	card_type	card_source_date	high_networth	active_30	active_60	active_90	...	bank_vintage	T+1_month_activity
0	1	4384 39XX XXXX XXXX	438439	Visa	edge	2019-09-29	B	0	1	1	...	27	0
1	2	4377 48XX XXXX XXXX	437748	Visa	prosperity	2002-10-30	A	1	1	1	...	52	0
2	3	4377 48XX XXXX XXXX	437748	Visa	rewards	2013-10-05	C	0	0	0	...	23	1
3	4	4258 06XX XXXX XXXX	425806	Visa	indianoil	1999-06-01	E	0	1	1	...	49	0
4	5	4377 48XX XXXX XXXX	437748	Visa	edge	2006-06-13	B	1	1	1	...	21	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8443	8444	4262 41XX XXXX XXXX	426241	Visa	chartered	2010-01-10	A	1	1	1	...	47	0

The dataset has 8448 rows × 28 columns. The column card source date has been formatted in the correct format to arrive at conclusions.

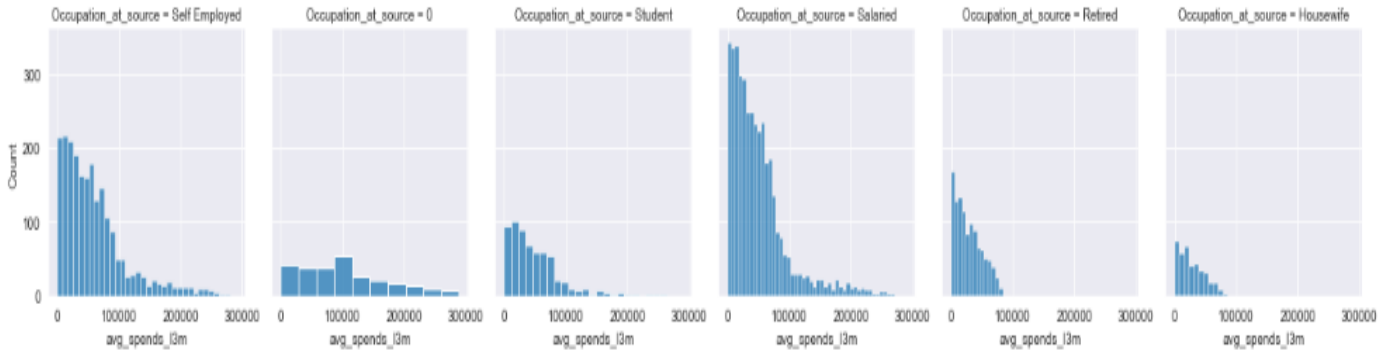
The Top 5 important variables are-

1. Annual Income
2. Transactor & Revolver
3. Average credit card spending

4. CC limit
5. Occupation at Source

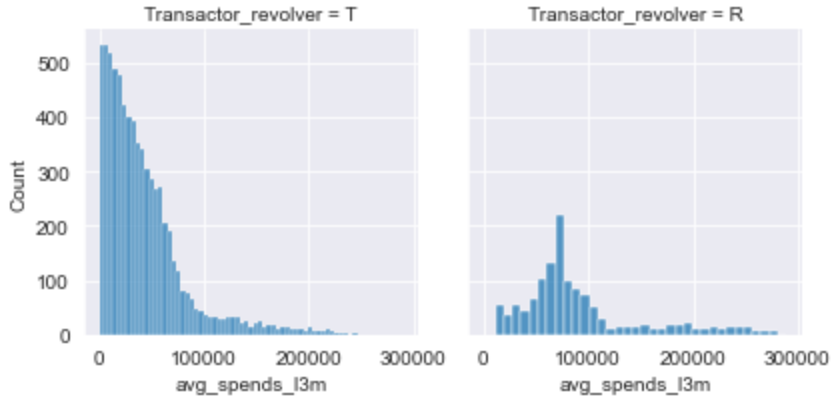
Company should concentrate on retaining most profitable customers through active engagement and campaigns.

As can be seen from the below subplots of histplot which is showing the different average spending by people of different occupations. Salaried people are more likely to spend more using credit cards.



Average spending & occupation based histplot(fig 19)

The people whose average spending is less are more likely to be transactor rather than revolver as can be seen from the below graph.

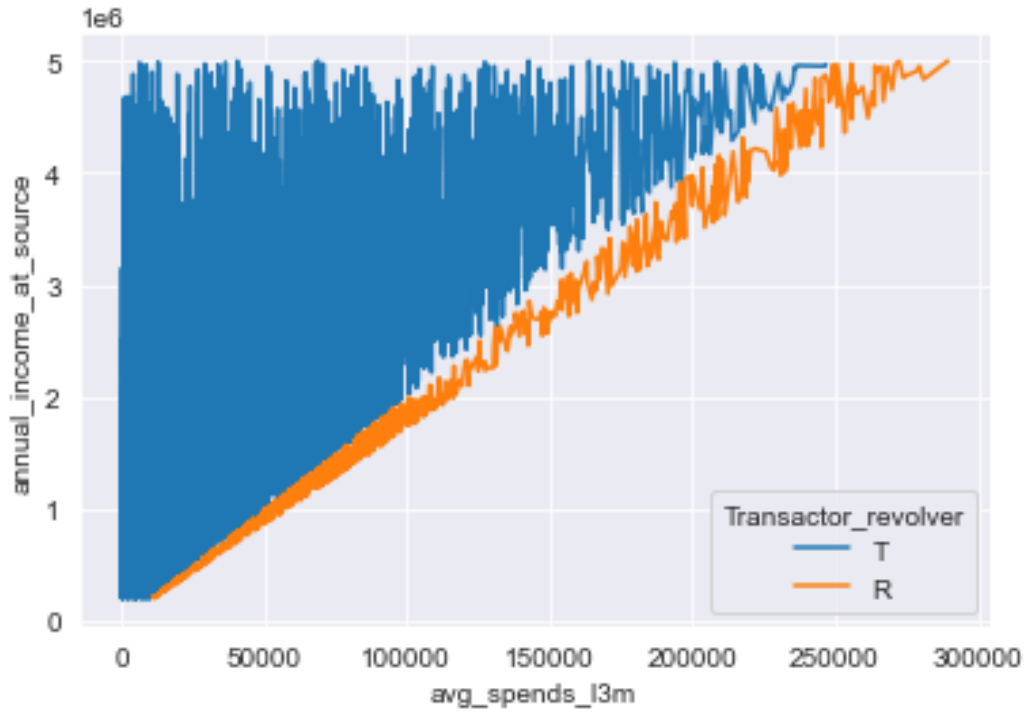


Average spending & Transactor Revolver histplot(Figure 20)

Annual income and average spending seems to have a high positive correlation which indicates that as income increases, average spending by customers are also increasing.

By using hue parameter we can see that transactors are more in case of low average spending however revolver are also seen in case of people who spend more and their annual income is also high.

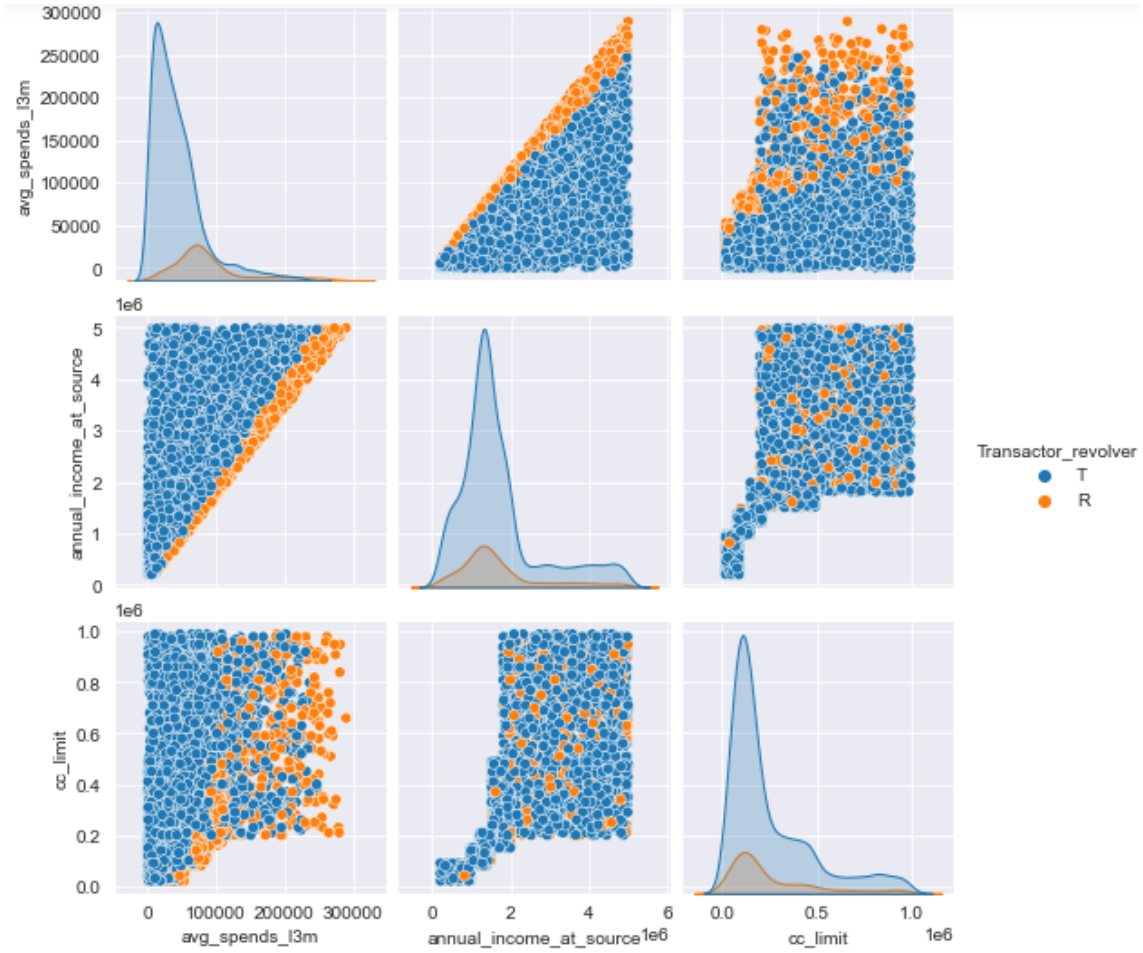
Company should focus more on such people that why they are carrying balances over from one month to the next.



Annual income & average spending lineplot(figure 21)

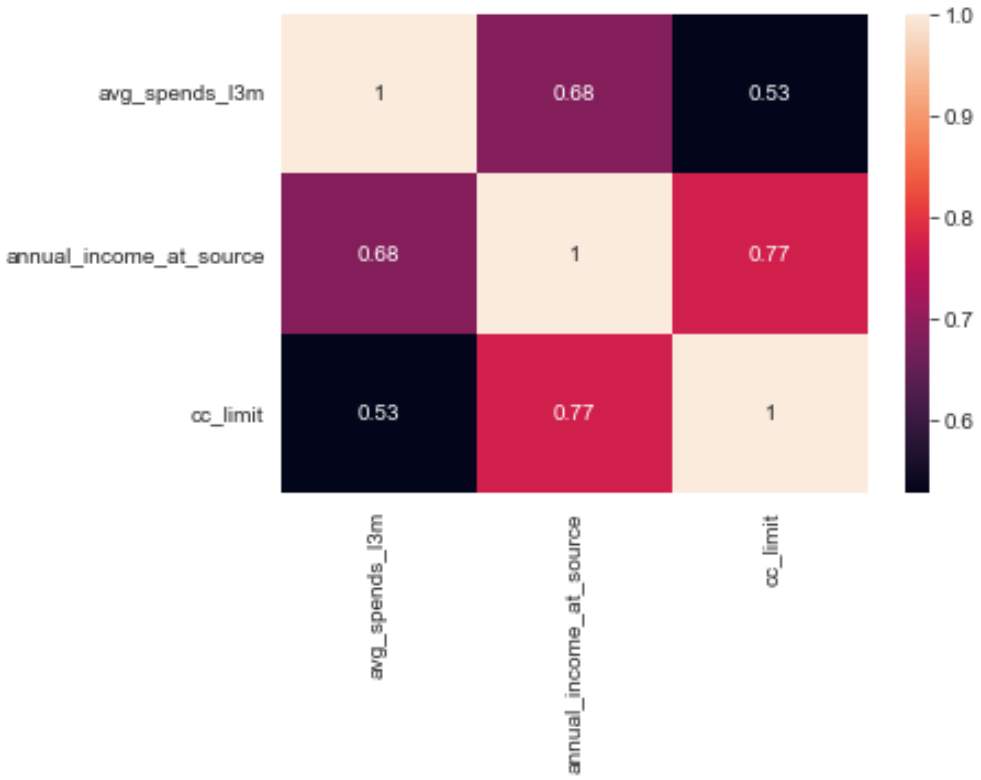
The pairplot among various numerical column can be seen below showing positive correlation between average spending and cc limit. It shows that business should focus more on revolver whose average spending are more so that high attrition is not seen.

The cc limit and annual income are also showing high positive correlation. Indicating that as annual income increases, cc limit also increases.



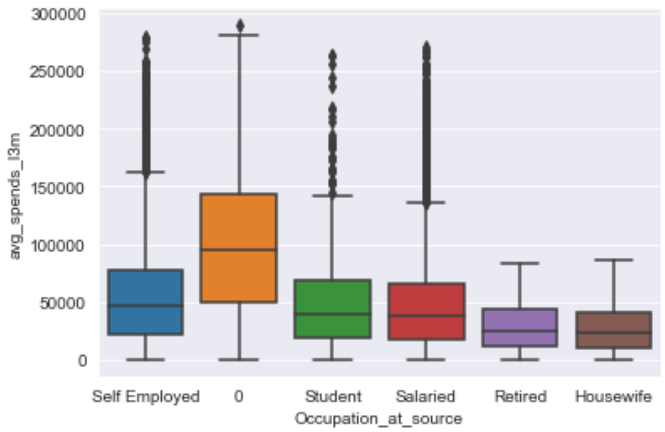
Pair plot (figure 22)

The heatmap shows the amount of correlation between the numerical variables ranging from 0 to 1(1 being high positive correlation and 0 being no correlation).



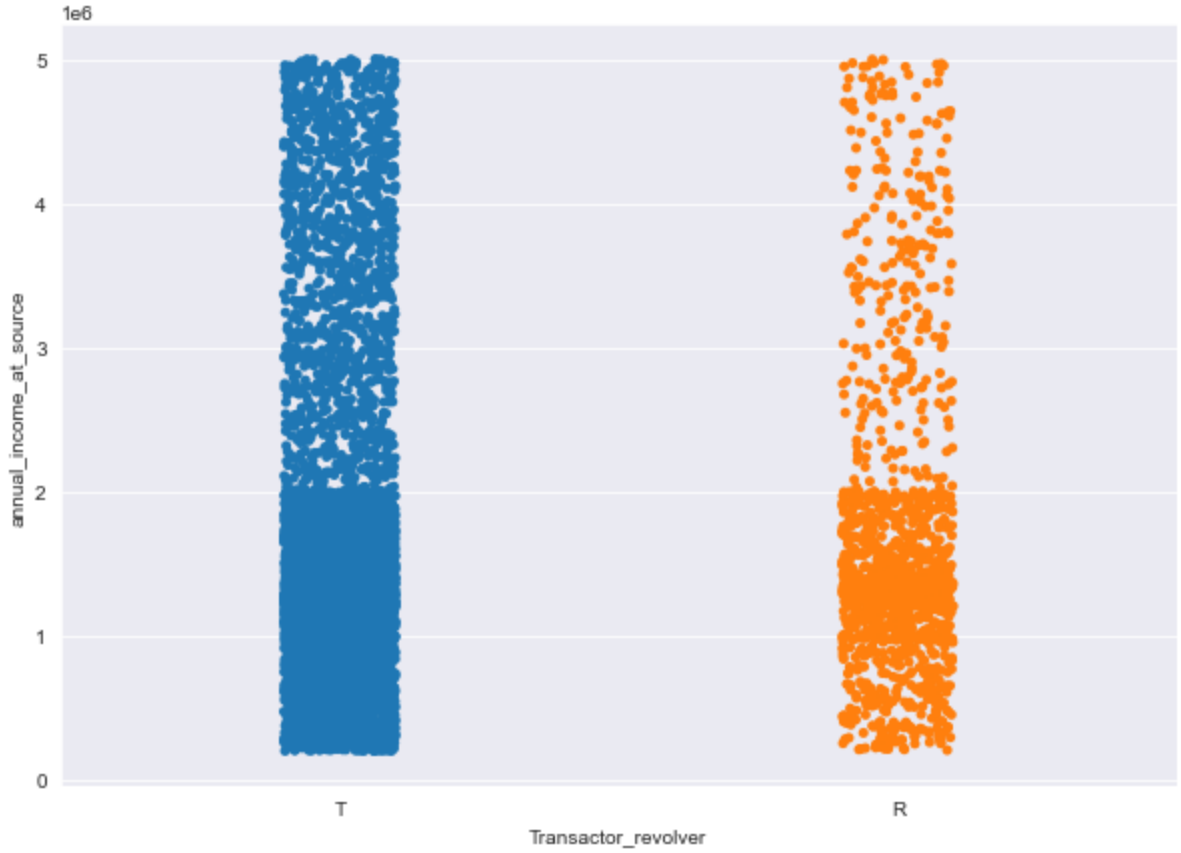
Heatmap (Figure 23)

The boxplot is showing presence of many outliers in different occupations. So many self employed,students & salaried people are spending more which is in turn leading to high attrition.Company should focus more on providing credit card to such customers.



Average spending & Occupation boxplot(Figure 24)

Using strip plot it can be observed that Revolvers are more in case of low income group there are only a few revolvers at high income group. It indicates the business should assign a maximum limit on the credit cards of low income group people so that there is less attrition.



Annual income & Transactor Revolver Strip Plot(figure 25)

THANK YOU