SMDM PROJECT REPORT

DSBA

Prepared By : Renjith K P

Contents

2. Austo Motor Company Problem

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.

1. A . What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables)

The dataset contains information about the car sales data of Austo Motor Company. The dataset has a volume of 1581 entries and 14 variables within it. Information that would be of interest to a database administrator from this database would be the good number of customer details that are available and the variables within each which would be of great advantage to the administrator to reach meaningful insights to improve the campaign. The 14 variables can be classified into Categorical and Continuous variables. Below is the classification of the same.

Categorical Variables		Continuous Variables
Binary	**Multi-Level**	
Profession	No_of_Dependents	Salary
Marital_status	Make	Partner_salary
Education		Total_salary
Personal_loan		Price
House_loan		Age
Partner_working		
Gender		

Data administrators would be interested in what each variable implies. Below is the data dictionary having details of each variable.

Data Dictionary

Variable	Description
Age	Age of the customer
Gender	Gender of the customer
Profession	Profession of the customer
Marital_status	Marital status of the customer
Education	Education of the customer
No_of_Dependents	Number of dependents of the customer
Personal_loan	Whether the customer is having personal loan or not
House_loan	Whether the customer is having house loan or not
Partner_working	Whether the customer's partner is working or not
Salary	Salary of the customer
Partner_salary	Salary of customer's partner

Total_salary	Total salary earned by customer and his/her partner
Price	Price of the car purchased
Make	Make of the car purchased

First Five Rows of the Data

	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	Femal	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

Basic Information of the Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581 entries, 0 to 1580
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Age               1581 non-null   int64
 1   Gender            1528 non-null   object
 2   Profession        1581 non-null   object
 3   Marital_status    1581 non-null   object
 4   Education         1581 non-null   object
 5   No_of_Dependents  1581 non-null   int64
 6   Personal_loan     1581 non-null   object
 7   House_loan        1581 non-null   object
 8   Partner_working   1581 non-null   object
 9   Salary            1581 non-null   int64
 10  Partner_salary    1475 non-null   float64
 11  Total_salary      1581 non-null   int64
 12  Price             1581 non-null   int64
 13  Make              1581 non-null   object
dtypes: float64(1), int64(5), object(8)
memory usage: 173.0+ KB
```

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?

There are 1581 customer purchase details are available in this dataset. After a detailed analysis, some entries were found to be missing from the dataset.

```
Age                   0
Gender               53
Profession            0
Marital_status        0
Education             0
No_of_Dependents      0
Personal_loan         0
House_loan            0
Partner_working       0
Salary                0
Partner_salary      106
Total_salary          0
Price                 0
Make                  0
dtype: int64
```

Gender & Partner Salary variables have null values. Gender has 53 null values while Partner_Salary has 106.

Handling Nulls –

Nulls are usually handled by the following techniques –

a) If the proportion of Null values is more than 60 % of the total number of records in a column, then drop the column. Here you assume that the column is uninformative.

b) If any row is missing a large amount of records across columns then that row may also be dropped.

c) Otherwise, the missing values may be imputed.

For the given data, neither (a) nor (b) is applicable since the proportion of null values in any column is small and no row contains a large number of missing observations.

Simple rules for imputation:

a) For categorical variables we can impute the Nulls with the majority class. For the current dataset, Null values in 'Gender' field are imputed with 'Male' (Male being the majority class).
b) For continuous variables it is possible to impute the Null values with mean/median of the variable depending upon the nature of the distribution. However, more efficient imputation is possible if variables are internally related.

Salary + Partner_salary = Total_salary

Also, non-null values in Partner_salary field is possible only if the Binary variable Partner_working is YES. Hence for this data we do a rule based imputation instead of the mean/median imputation –

If Partner_working = 'No' then Partner_salary = 0

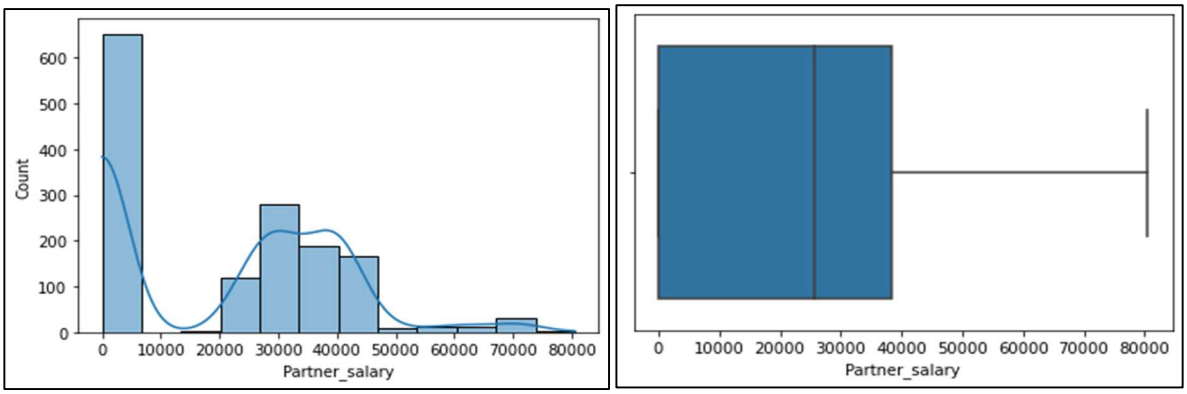
If Partner_working = 'Yes' then Partner_salary = Total_salary – Salary

Below is the transposed description of the dataset showing the statistical summary of numerical variables.

Out[33]:

	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

Below is the histogram and boxplot depicting the distribution of Partner_salary.



The mean of Partner_salary is greater than the median indicating the distribution is right-tailed.

After a detailed analysis of the statistical summary, histogram, and box plot, it was observed that _salary data is having entry input as 0 when the partner is not working or the customer is unmarried. These entries are also considered for calculating mean and quartiles which could lead to misinterpretation of the data.

1) The customers are between 22 and 54 years old. It may be said that they belong to working age group. Mean age is 31.92 while median age is 29 years, indicating age distribution is positively skew. The value of skewness is 0.89.

2) The salary of the customers ranges between 30K and 99.3K and the distribution is symmetric. The mean and the median values are very close and skewness is very close to 0.

3) Total_salary ranges between 30K and 171K and does not show a high degree of skewness.

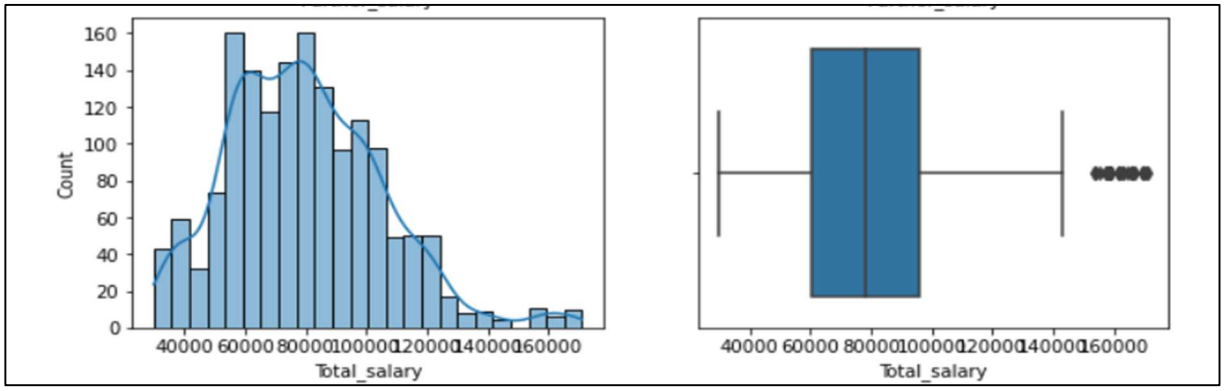
4) The minimum price of the purchased automobile is 18K, whereas max is 70K. Price has a skewness of + 0.74 indicating moderate skewness. This indicates a small number of high priced purchases were made.

The categorical variables also show some discrepancies in variables.

```
Out[138]: Male      1199
          Female     327
          Femal        1
          Femle        1
          Name: Gender, dtype: int64
```

Above is the count of entries in the categorical variable 'Gender', here two entries are written as 'Femal' & 'Femle' which is a typo error. These are both typo error entries which should be replaced with 'Female'.

To find the outliers in the data, boxplots were drawn for each continuous variables.



From the above plots, we can see that the Total_salary is having outliers present in the data.

From the boxplots we can observe outlier values are present in Total_salary variables. (Using the 1.5 * IQR rule)

Note – If more than 25-30% of the total records lie beyond the range defined by the 1.5 times IQR rule, then to avoid a significant loss of data 3 times IQR rule is considered.

Total_Salary- A total of 27 outlier values are present in the variable.

To handle the outliers in Total_salary, we can choose any of the following two approaches -

1) We can treat the outlier values using Winsorization. However, this may lead to loss of valuable information hence should be used with caution.

2) We do not treat the outlier values, and see if analysing them separately can give us some more insights.

Here we are not treating outliers.

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

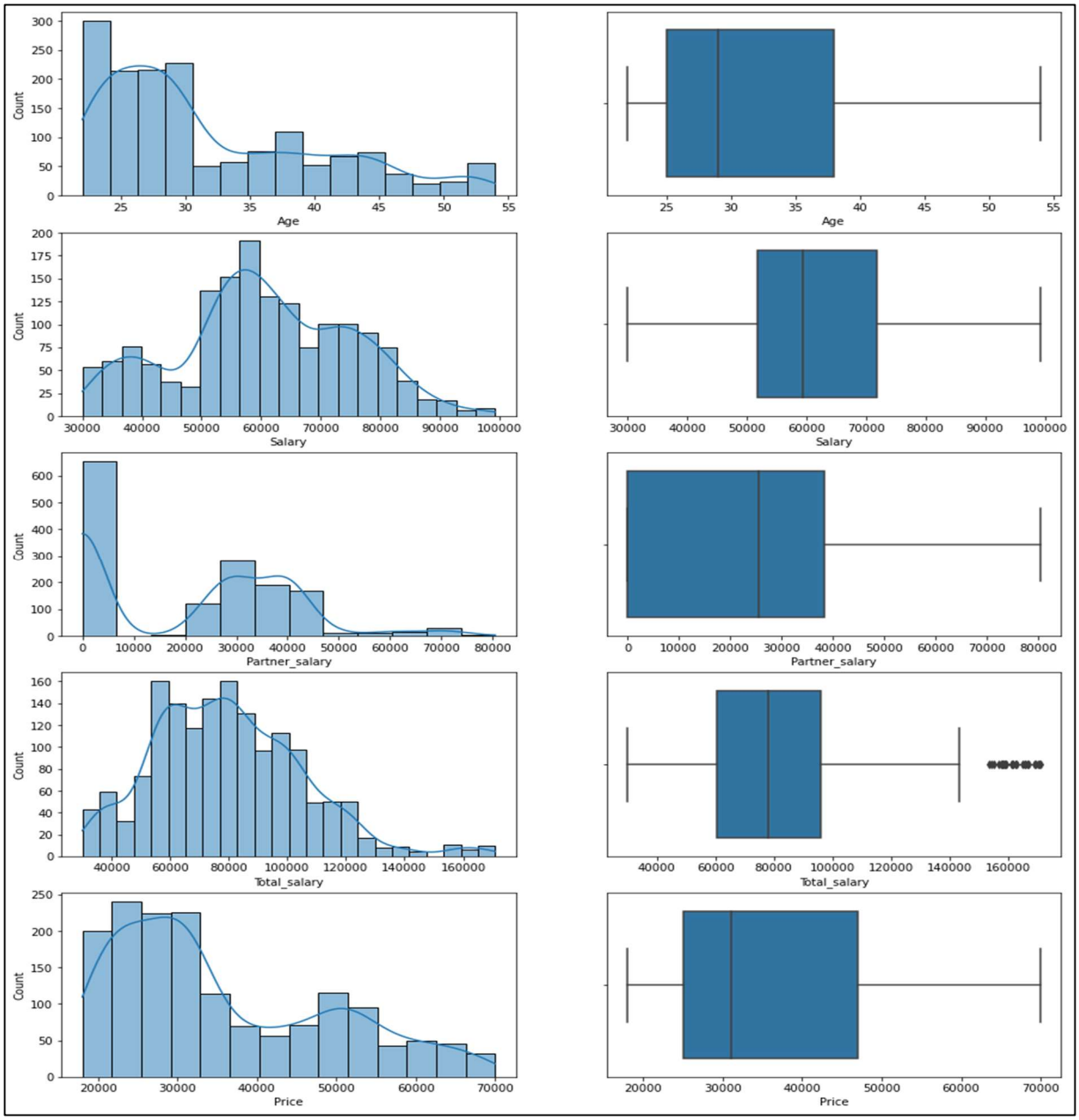
Let's start analyzing the data with the numerical(continuous) variables first.

Below is the statistical summary of continuous variables.

Out[165]:

	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
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

For performing Univariate analysis we will take a look at the Boxplots and Histograms to get better understanding of the distributions.



Inferences:

1) Salary has a multimodal distribution, with bulk of data points in the range 50K to 70K.

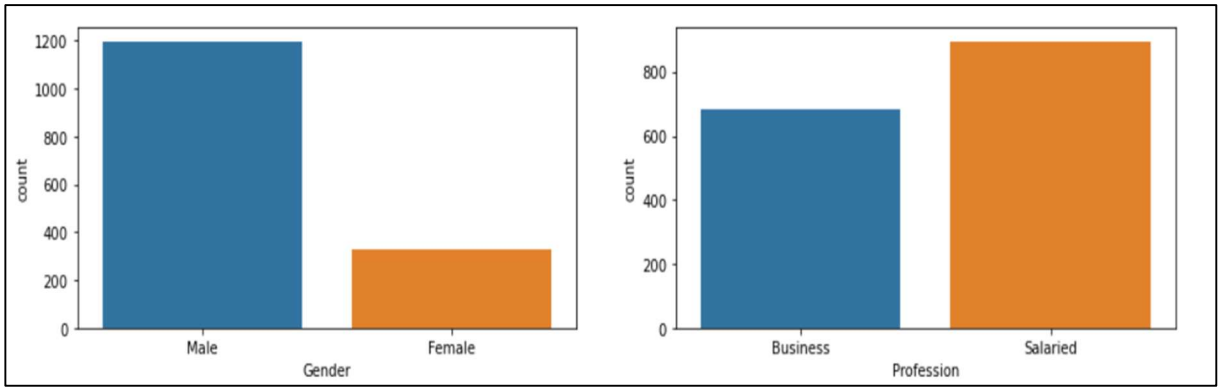
2) Price seems to have a Bi-modal distribution, and a positive skew of 0.74.

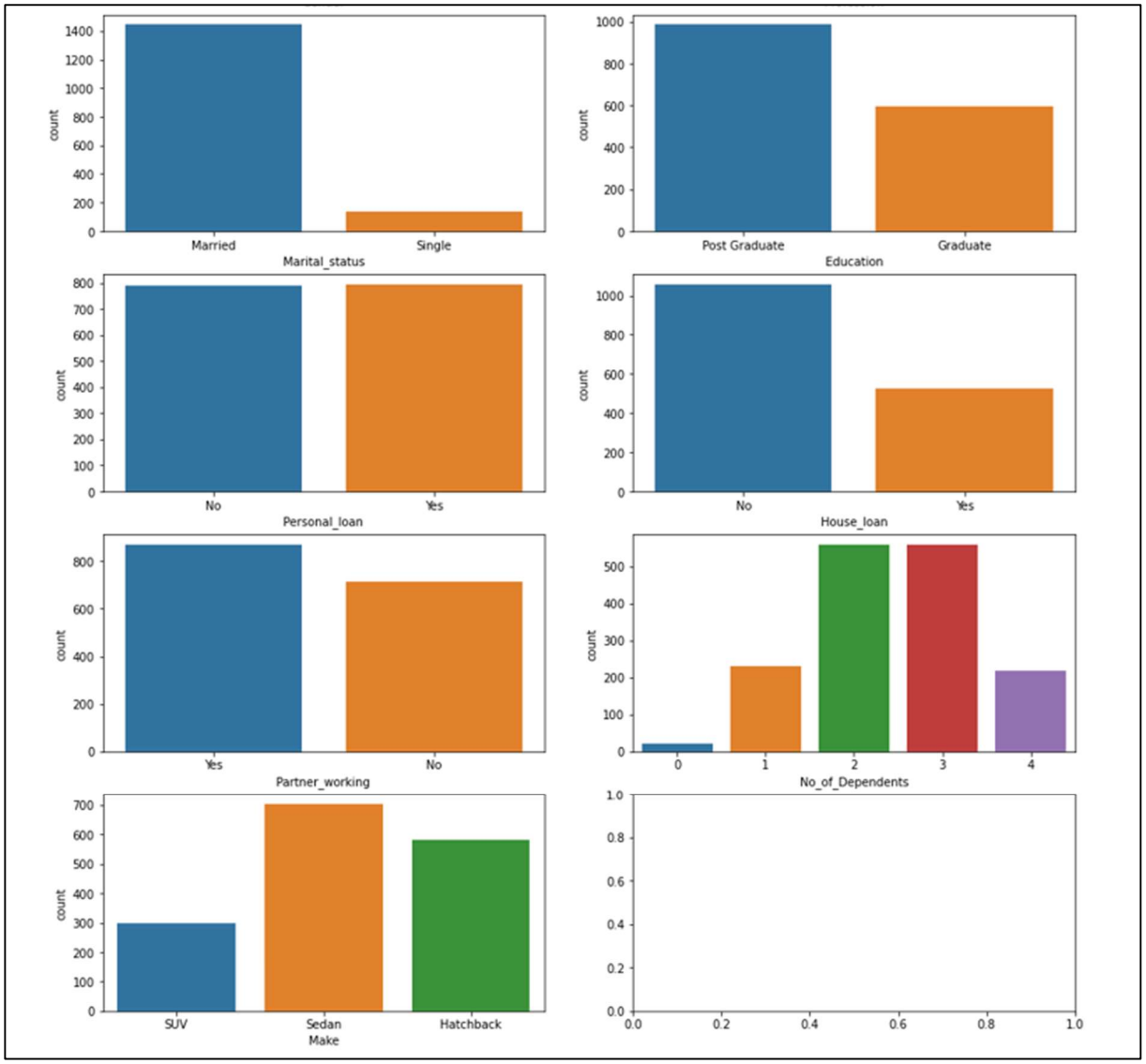
3) Age seems to have a multimodal distribution, and has the highest positive skew of 1.14 among all the fields.

4) Skewness of Total_salary has reduced significant post outlier treatment. The distribution seems to be multimodal, with bulk of data points in the range of 60K to 100K.

5) Almost all the variables have some skewness present, thus none of them follow a Normal distribution. Total_salary can be considered Near-Normal distribution with fair bit of approximation.

Now we will look into the continuous variables





Inferences

1) Sedan is the most preferred purchase, followed by Hatchback and SUV.

2) The number of customers having a working partner are slightly higher than customers with nonworking partner or singles. There are a total of 713 customers with Partner_working variable as 'No', out of which 138 customers are 'Single'.

3) Number of Customers who did not take a House Loan is almost double the customers who took a House Loan.

4) The data consists of very small proportion of Single customers when compared to married customers.

5) Count of Salaried customers is slightly higher than that of Business customers.

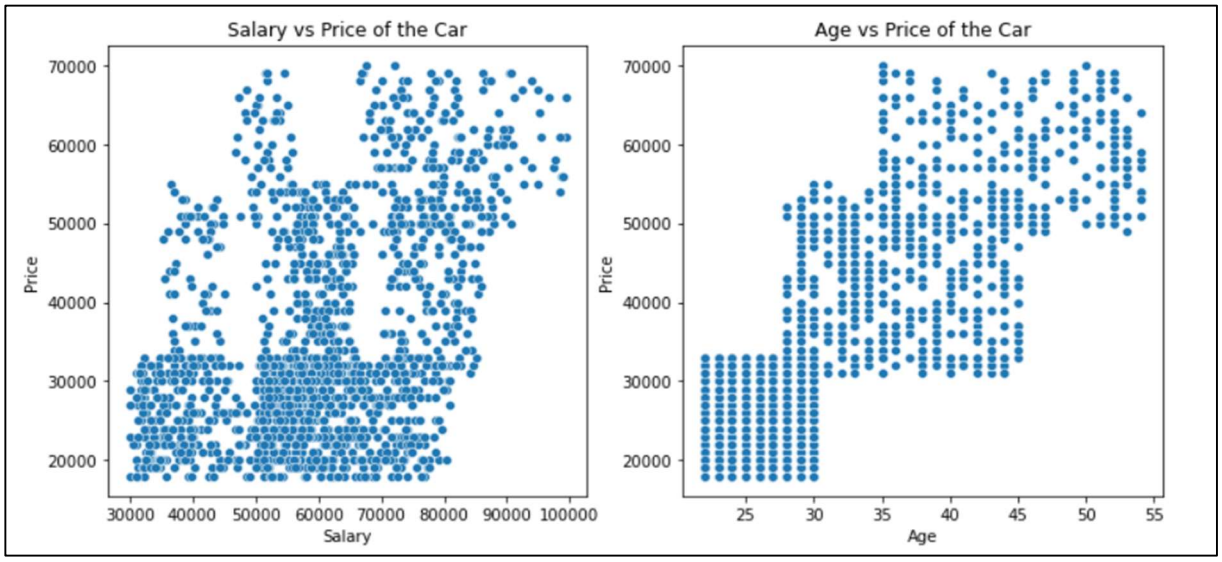
6) Majority of the customers in the dataset are Post Graduate.

7) From the Barplot of No_of_dependnts variable we can infer that majority of the customers have either 2 or 3 dependents, followed by 1 or 4 dependents. Very few customers have zero no of

dependents.

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.

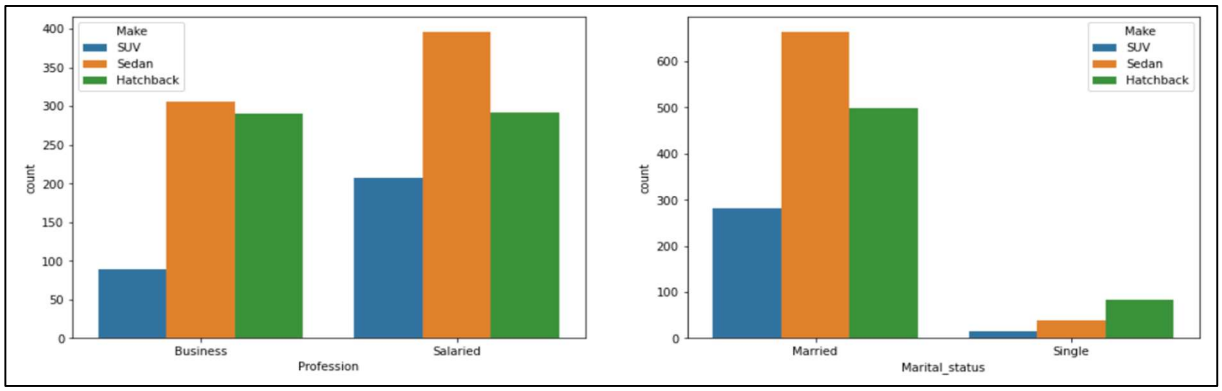
First we will check the relationships between numerical(continuous) variables. Different scatter diagrams were plotted to identify the relationship between them.

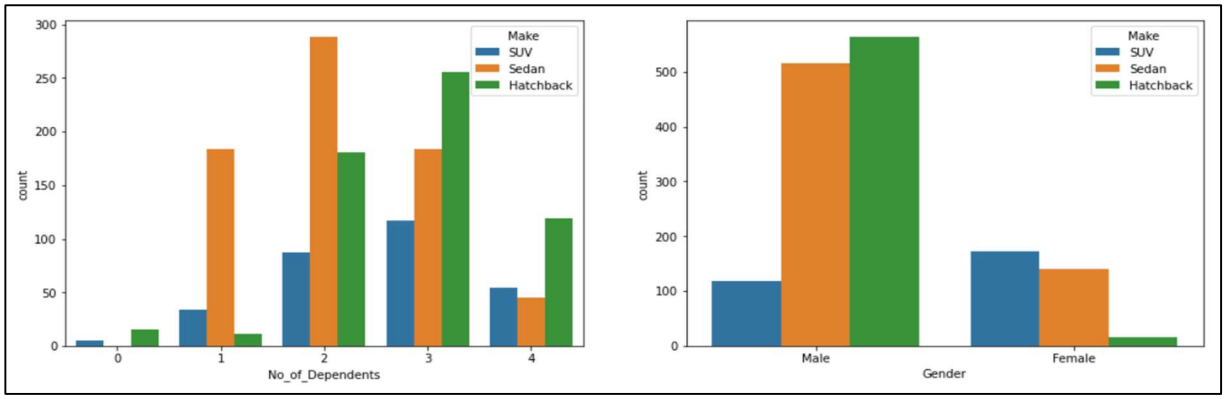


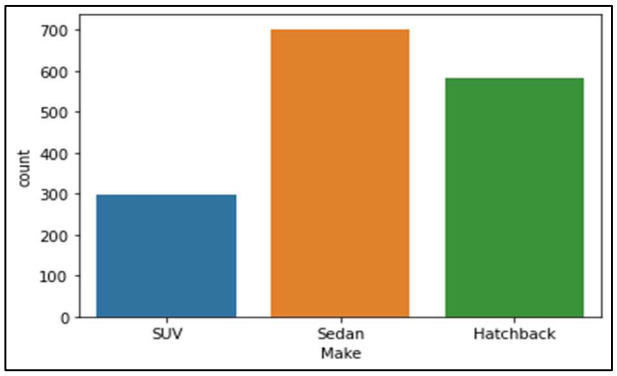
Inferences

1) From the above plot we can see that price of the car purchased increased with the age of the customer.
2) With the increase in salary (Salary and Total_salary) there is an increase in price of the car purchased, but the correlation is weak.

Now we will analyze the relationship between the categorical variables.





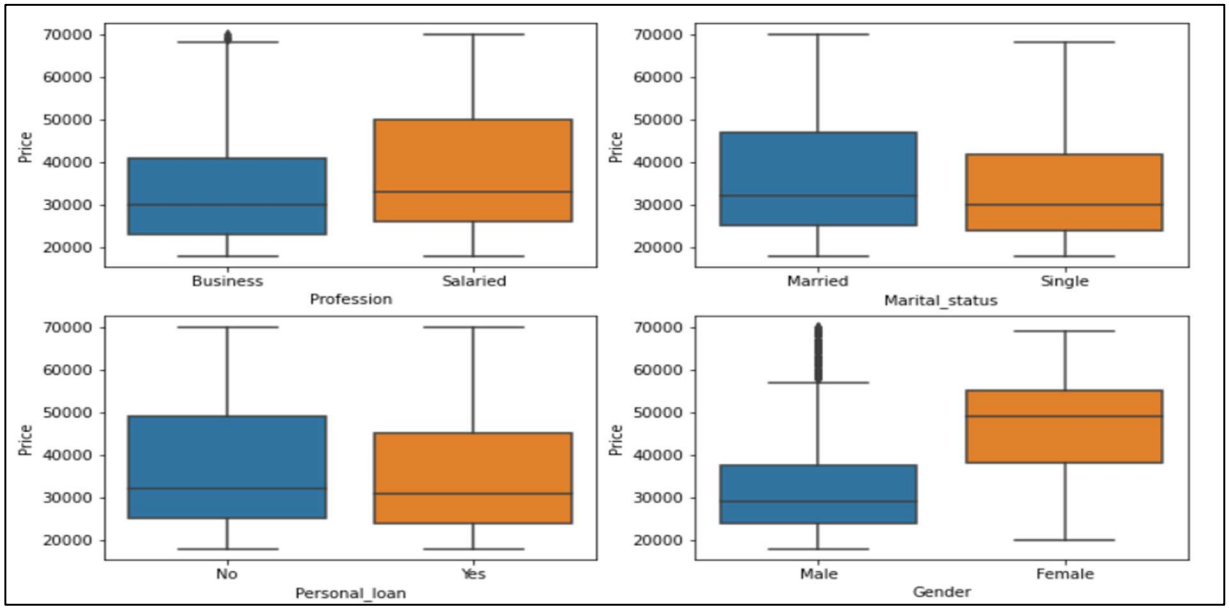


Inferences

1) Customers who have a house loan are not likely to buy an SUV (which is the most costly make among the three). Sedan is most preferred across both the categories.

2) Females prefer SUV and are least likely to buy a Hatchback, whereas Male prefer Sedan or hatchback. SUV is least preferable among males

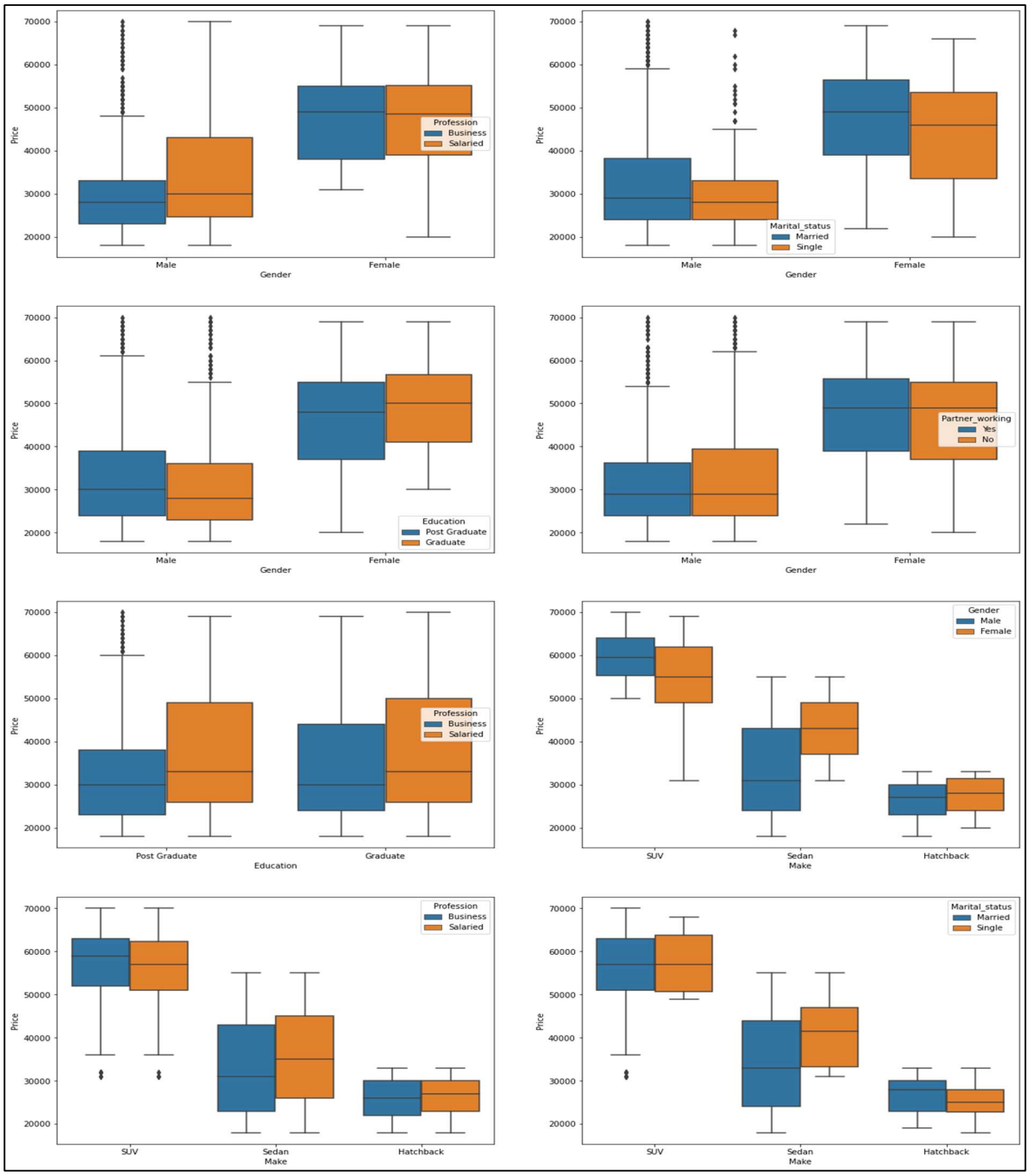
3) Married customers prefer Sedan whereas single customers prefer Hatchbacks. Now we will analyse the relationship between categorical and continuous variables



Inferences

1) Salaried customers bought slightly higher priced cars as compared to customers doing business
2) Married customers bought higher priced cars compared to unmarried customers
3) Customers without house loan and personal loan bought higher priced cars compared to customers with house loan & personal loan
4) Female customers purchased higher priced cars compared to male customers

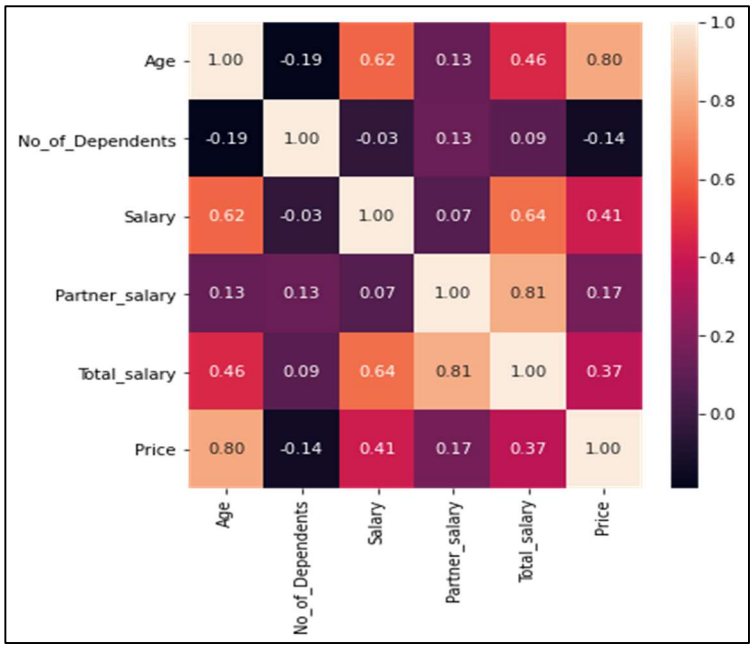
Now we will do multivariate analysis



Inferences

1) Regardless of profession and marital status females bought higher priced cars
2) Salaried males bought higher priced cars compared to male customers doing business
3) Married customers are purchasing higher prised cars than single customers
4) Graduate female customers are purchasing higher priced cars compared to Post graduate female customers. Whereas in the case of male customers it is opposite.
5) In the case of customers bought SUV, male customers spend higher price compared to female customers
6) In case of customers who bought sedan & hatchback, female customers have spent higher price compared to male customers
7) In case of customers who bought SUV, customers doing business have spent a higher price as compared to salaried customers
8) In the case of customers who bought sedan & hatchback, salaried customers have spent higher price compared to customers doing business
9) In the case of customers who bought sedan, unmarried customers spend higher price compared to married customers
10) In the case of customers bought hatchback, married customers spend higher price compared to unmarried customers

To understand the relationships across variables, a heatmap with correlation factor is plotted.



Inferences

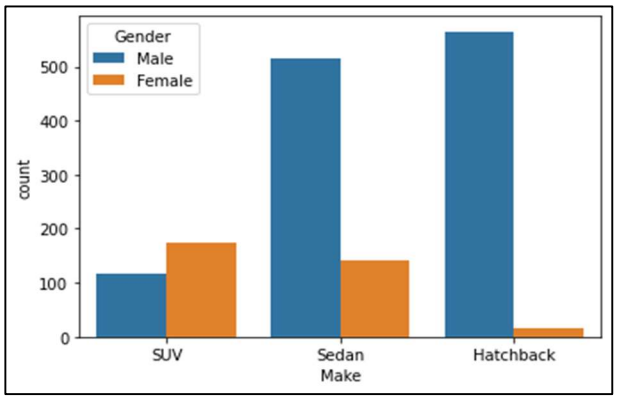
There is some strong correlations observed between a few fields. 'Total_salary' is highly correlated to 'Partner_salary'. Also, there is a positive relationship between age and salary.

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.

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

I don’t agree with observation of Steve Rogers. After analyzing the data, I have reached the conclusion that females prefer SUV more compared to men.

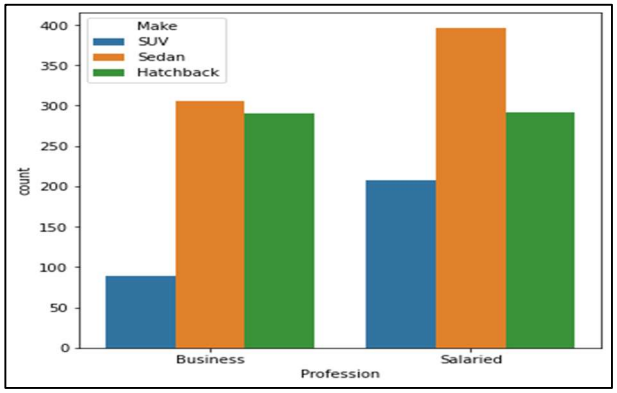
It’s clearly depicted in the graph that follows.



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

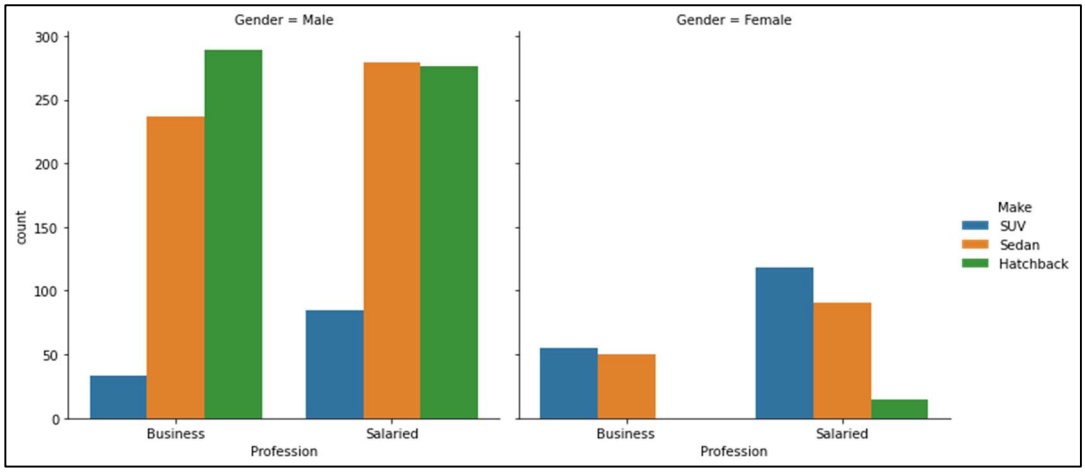
I agree with Ned Stark’s observation. From my analysis of the data, I found my observation matches with Ned’s.

It’s clearly depicted in the graph below.



1.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.

I disagree with Sheldon Cooper. Based on the dataset salaried male customers prefer to buy sedan cars first followed by hatchback cars and SUVs are least preferred by them. Hence Sheldon Cooper’s observations are not valid here. Below graph shows the same.

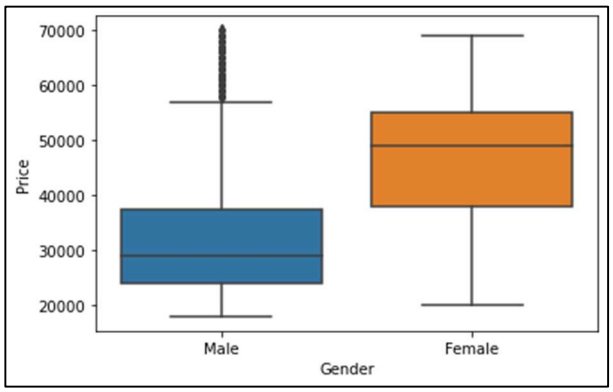


1.F. From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a Business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions.

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

1.F1) Gender

The average amount spent on purchasing automobiles across gender is higher for female customers as compared to males. Here we have compared the amount spent by male and female customers using box plots



We could find lot of outliers for male customers.

Out[260]:

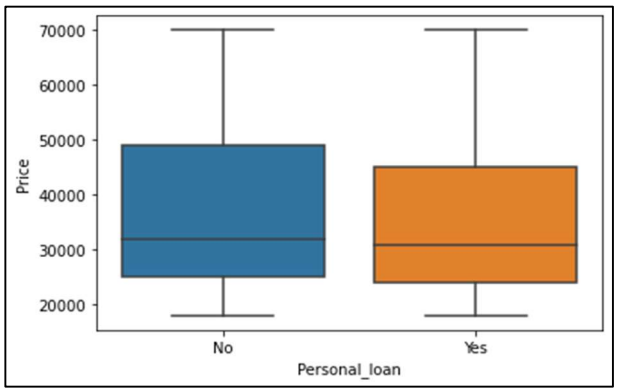
Gender	mean Price	median Price
Female	47705.167173	49000
Male	32817.347790	29000

Here if we compare the mean and median, we could see that the average amount spent by female customers for purchasing cars is always higher than male customers. We cannot compare the sum of amount spent since the number of male customers bought cars is very high compared to the female customers.

Business can utilize this conclusion to prepare marketing strategy that can target selling high priced cars to female customers. Low- and medium-priced cars can be targeted for selling to male customers. Information's and promotions about the higher priced cars need to be targeted at female customers.

1.F2) Personal_loan

Customers with personal loan have bought cars having slightly lower price compared to the customers who doesn't have a personal loan. It is visible in below boxplot.



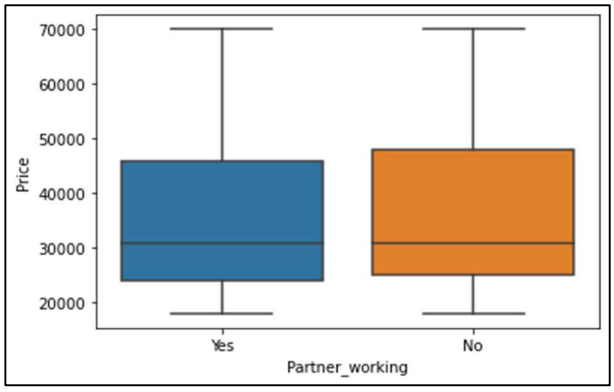
Out[263]:

Personal_loan	mean Price	median Price	sum Price
No	36742.712294	32000	28990000
Yes	34457.070707	31000	27290000

Here we can compare the mean price and sum of the amount spent since there are no outliers present in the data. When we compare mean amount spend for purchasing cars, we could see that customers without personal loan have spent a higher amount. Since the number of customers having personal loan and not having personal loan is almost same, the sum of the amount spend also can be compared. This comparison also leads to the conclusion that customers without personal loan have spent a higher amount for purchasing cars.

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

Having a working partner is not leading to the purchase of a higher-priced car. Below box plot shows the comparison of amount spent by customers with working partner and without.



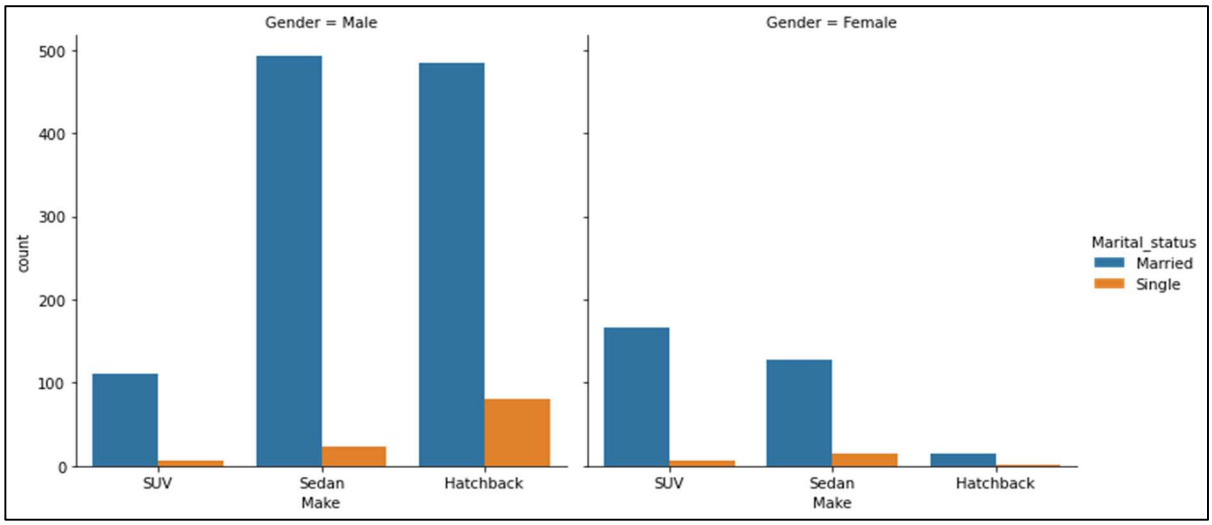
Also the mean and median comparison for the same.

Out[266]:

Partner_working	mean Price	median Price
No	36000.000000	31000
Yes	35267.281106	31000

Both box plot and the comparison of mean and median indicates that having a working partner is not leading to the purchase of a higher-priced car. Median is same for both the cases and mean is slightly higher for customers not having a working partner.

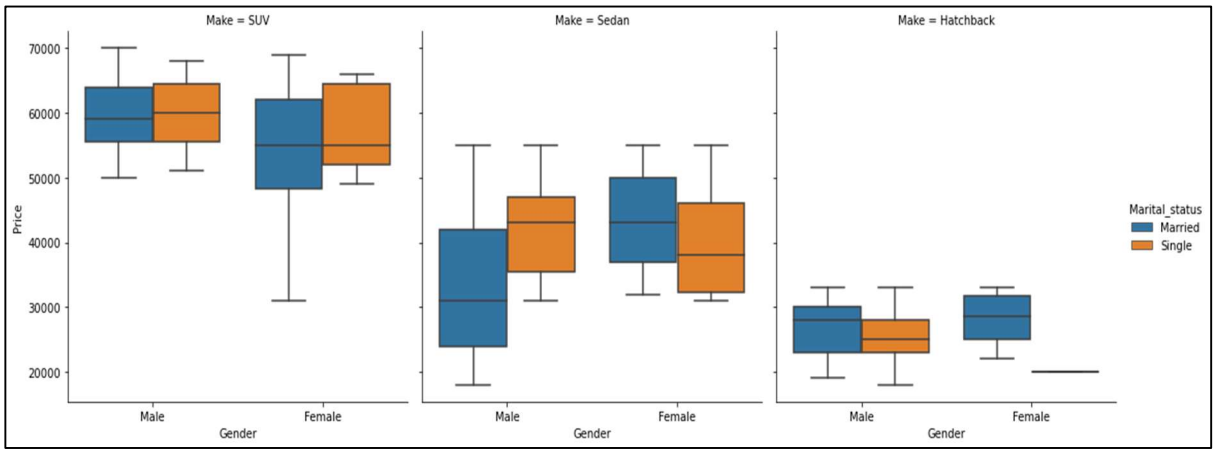
1.H. The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use the Gender and Marital_status - fields to arrive at groups with similar purchase history.



Inferences

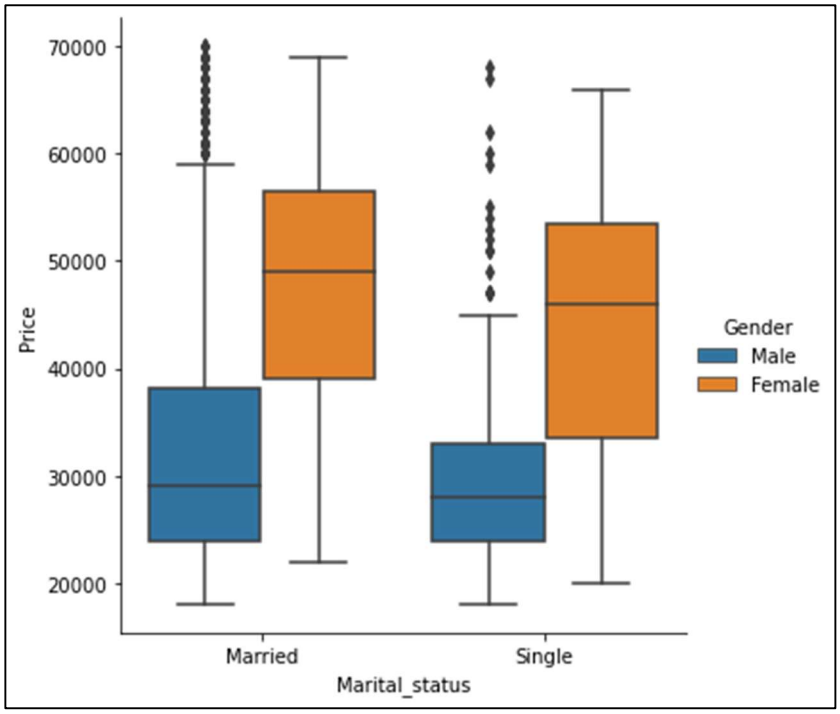
1) Married male customers purchased more sedan cars followed by hatchback cars
2) Married female customers purchased more SUV cars followed by Sedan cars
3) Unmarried male customers purchased more hatchback cars followed by sedan cars

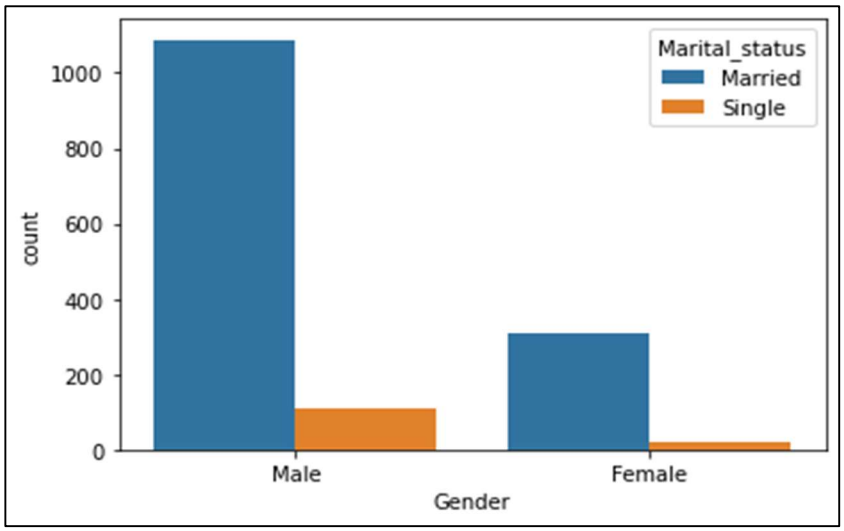
4) Unmarried female customers purchased more sedan cars followed by SUV cars



Inferences

1) Slightly higher priced SUVs are purchased by male customers compared to female customers irrespective of the marital status
2) Unmarried male customers purchased higher priced sedan cars compared to married male customers whereas married female customers purchased higher priced sedan cars compared to unmarried female customers
3) Very low number of female customers only bought hatchback cars. When comparing the female customers based on their marital status, married customers bought higher priced hatchback cars
4) Married male customers bought higher priced hatchback cars compared to unmarried male customers





Inferences

1) Female customers bought high priced cars.
2) For both male and female customers married cutomers bought cars having high price
3) Married male customers purchased more cars compared to married female customers
4) Single male customers purchased more cars compared to single female customers

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.)

Analyze the dataset and list down the top 5 important variables, along with the business justifications. (10 Points)

Top 5 important variables

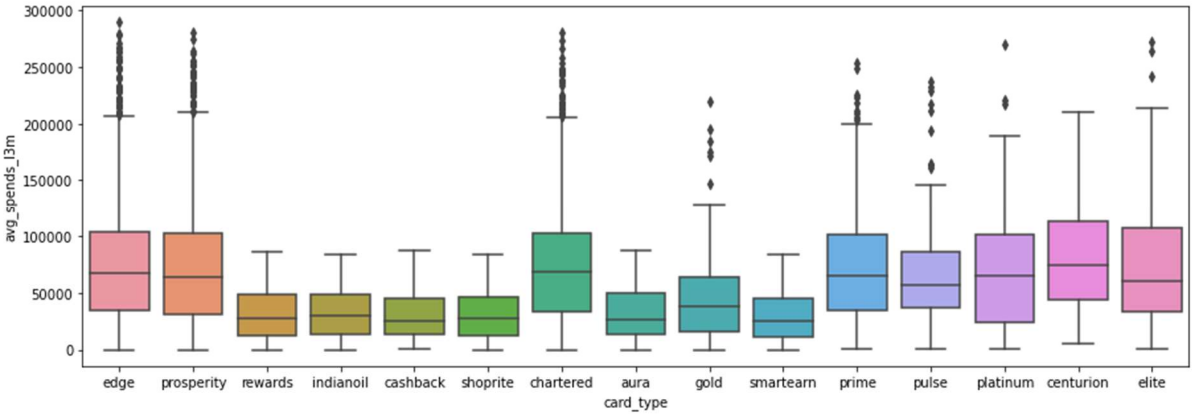
1. card_type – Credit card type
2. Occupation_at_source - Occupation recorded at the time of credit card application
3. annual_income_at_source - Annual income recorded in credit card application
4. avg_spends_l3m - Average credit card spends in last 3 months
5. Cc_limit – credit card limit

Top 5 important variables business justifications

1. Card_type

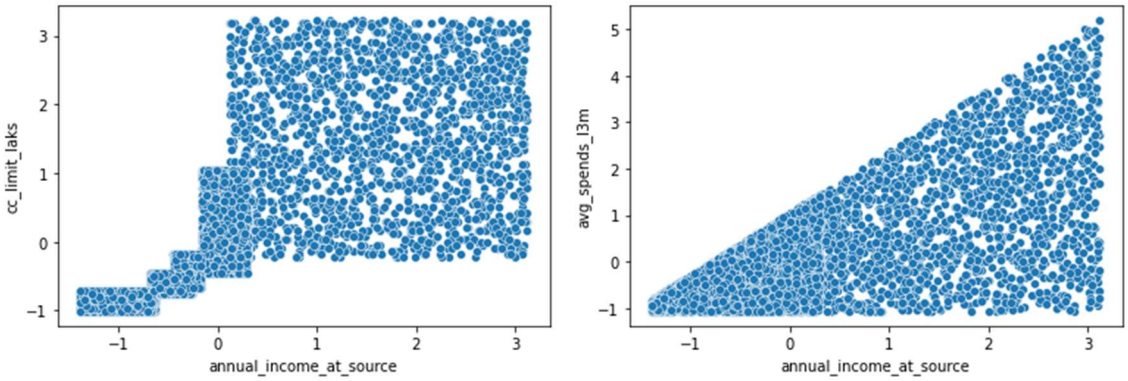
Different card types serves different purposes. For example there are card types for shopping rewards, cashbacks, fuel cashbacks or points,etc.. Bank should understand the need and issue the right card to the customer. When the customer is having the right card, usage of that card will be high. When a customer who is not having any vehicle but he/she interested in shopping issued a fuel savings card, then the usage of the card will be very minimum. Hence the card_type is important variable to improve the business.

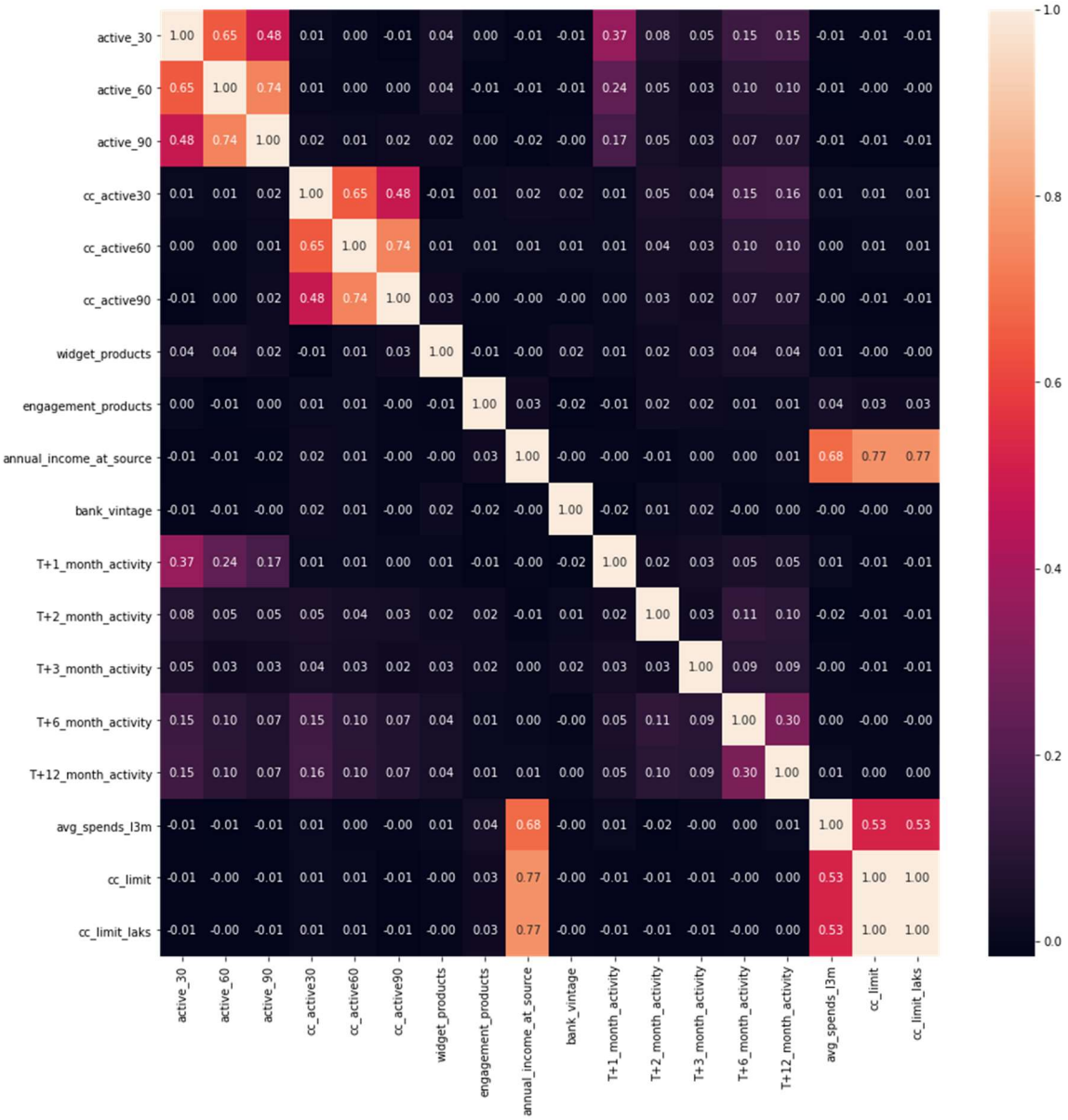
Below boxplot clearly shows that different card_types have different average 3 months spendings. Which indicates that the customer with particular types of credit cards spends more.



2. annual_income_at_source

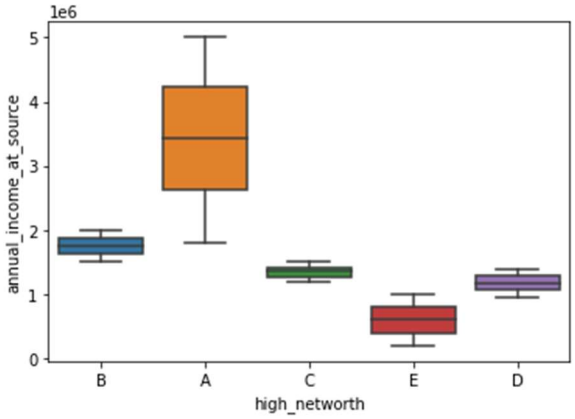
Assessment of customer's financial level is as important as issuing right card to them. The card limit offered to the customers must be decided based on their annual income. Expected average spending must be calculated depending on that and the card offered to them must have enough credit limit accordingly .





From above scatterplots and heatmap we can observe that there is strong positive correlation between the annual income of the customer and average 3 months spending on the credit card also with credit card limit. This variable is important for the business to decide the credit lit as well as to issue right credit card to an important customer.

Below boxplot is indicating that the annual income data is also considered for deciding the high_networth and how important the customer to the bank.



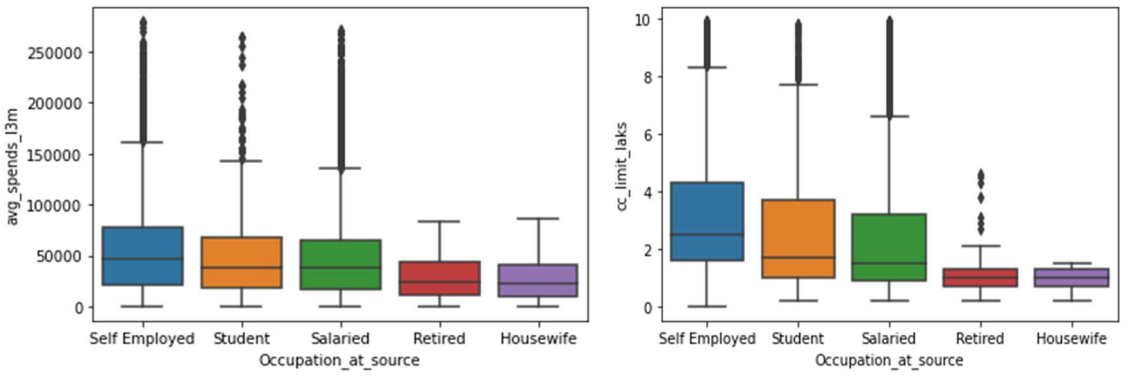
3. avg_spends_l3m

Average spend using the credit card is an important data for bank. This data can be used to analyse the customers spendings in different areas like shopping, fuel &entertainment. Depends on this the credit limit can be changed, new offers can be given to the customers or a new credit card can be offered. All these will help in increasing the usage of credit card.

Average spending is the data which shows how much a customer is using the credit card. This can be used as a variable to find relation with other variables. This relations can be used to improve the spendings and business of the bank.

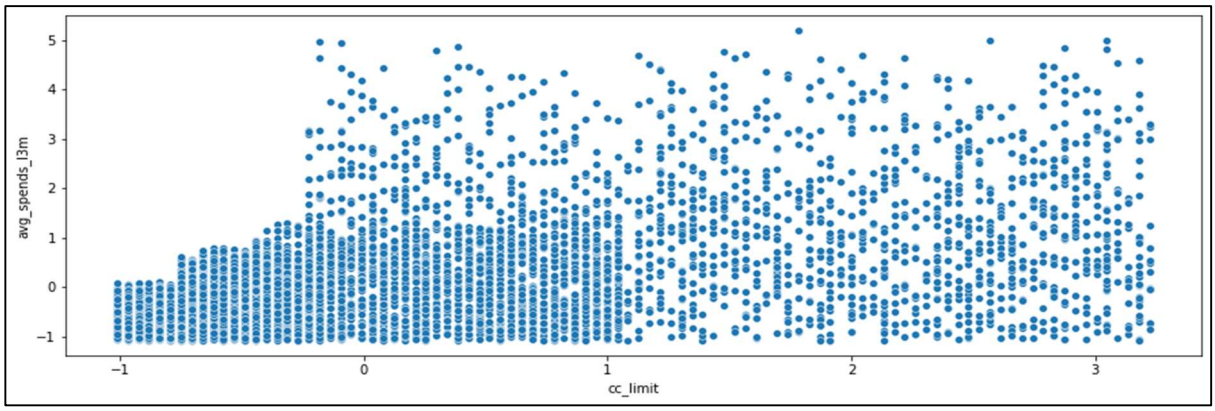
4. Occupation_at_source

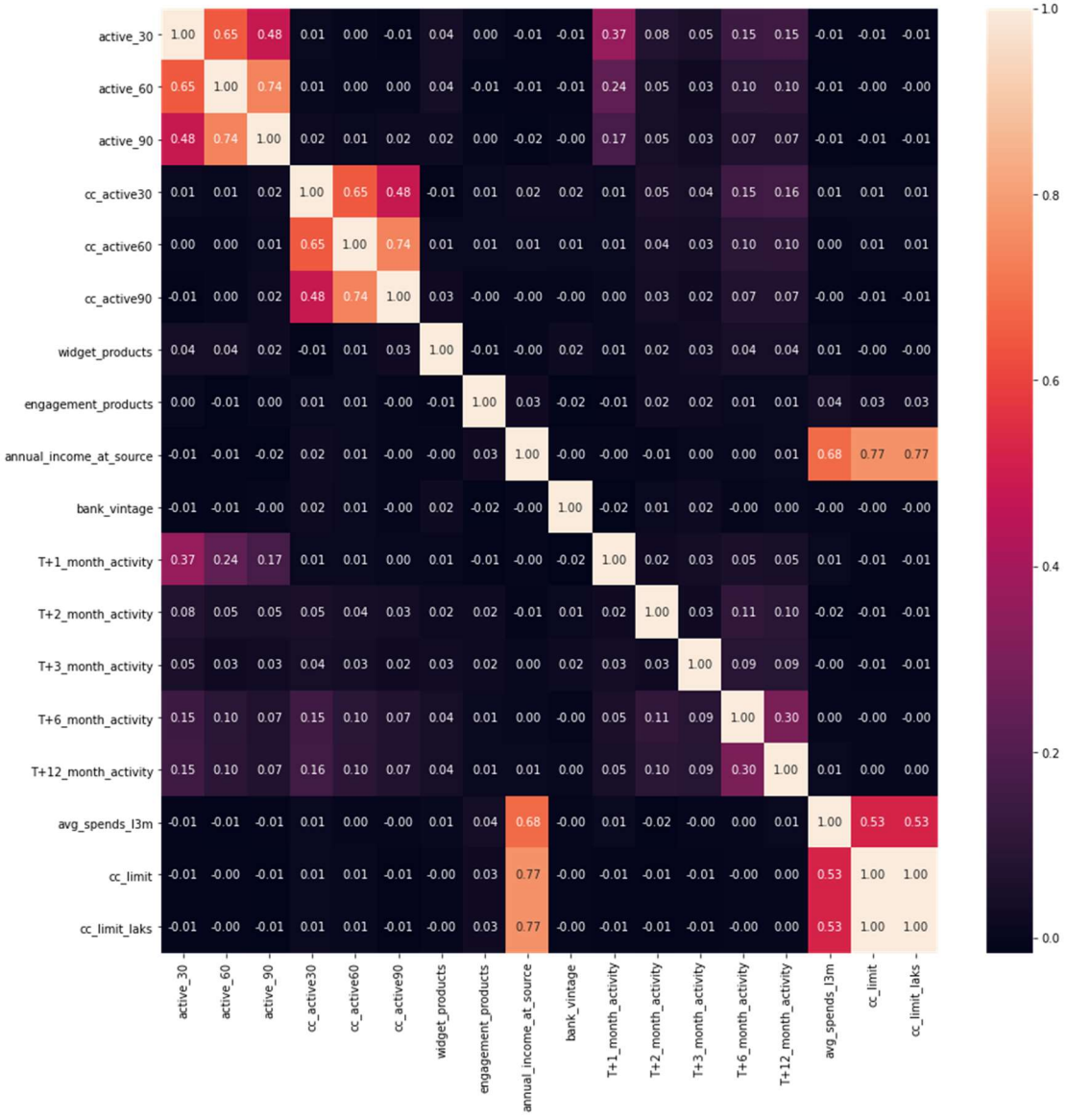
Occupation of the customers is an important factor in deciding the credit limit and the type of card to be offered to the customer. Salaried customers will be spending more compared to others, hence they must be provided with adequate credit limit and good offers like rewards or cash backs. Retired customers spending might not come to the level of salaried customers, retired customers can be offered with fuel savings cards or such. Occupation assessment is also important in determining the repay capability of the customers. This variable is important for the business.



Above box plot shows the difference in average spendings for different occupations of cutomers. Self Employed customers spendings are higher followed by students spendings. This data is important for business. Also we could see the credit limit is also changing with occupation of the customer.

5. Cc_limit


avg_spends_l3m
cc_limit
5
4
3
2
1
0
−1
−1
0
1
2
3


active_30 1.00 0.65 0.48 0.01 0.00 -0.01 0.04 0.00 -0.01 -0.01 0.37 0.08 0.05 0.15 0.15 -0.01 -0.01 -0.01
active_60 0.65 1.00 0.74 0.01 0.00 0.00 0.04 -0.01 -0.01 -0.01 0.24 0.05 0.03 0.10 0.10 -0.01 -0.00 -0.00
active_90 0.48 0.74 1.00 0.02 0.01 0.02 0.02 0.00 -0.02 -0.00 0.17 0.05 0.03 0.07 0.07 -0.01 -0.01 -0.01
cc_active30 0.01 0.01 0.02 1.00 0.65 0.48 -0.01 0.01 0.02 0.02 0.01 0.05 0.04 0.15 0.16 0.01 0.01 0.01
cc_active60 0.00 0.00 0.01 0.65 1.00 0.74 0.01 0.01 0.01 0.01 0.01 0.04 0.03 0.10 0.10 0.00 0.01 0.01
cc_active90 -0.01 0.00 0.02 0.48 0.74 1.00 0.03 -0.00 -0.00 -0.00 0.00 0.03 0.02 0.07 0.07 -0.00 -0.01 -0.01
widget_products 0.04 0.04 0.02 -0.01 0.01 0.03 1.00 -0.01 -0.00 0.02 0.01 0.02 0.03 0.04 0.04 0.01 -0.00 -0.00
engagement_products 0.00 -0.01 0.00 0.01 0.01 -0.00 -0.01 1.00 0.03 -0.02 -0.01 0.02 0.02 0.01 0.01 0.04 0.03 0.03
annual_income_at_source -0.01 -0.01 -0.02 0.02 0.01 -0.00 -0.00 0.03 1.00 -0.00 -0.00 -0.01 0.00 0.00 0.01 0.68 0.77 0.77
bank_vintage -0.01 -0.01 -0.00 0.02 0.01 -0.00 0.02 -0.02 -0.00 1.00 -0.02 0.01 0.02 -0.00 0.00 -0.00 -0.00 -0.00
T+1_month_activity 0.37 0.24 0.17 0.01 0.01 0.00 0.01 -0.01 -0.00 -0.02 1.00 0.02 0.03 0.05 0.05 0.01 -0.01 -0.01
T+2_month_activity 0.08 0.05 0.05 0.05 0.04 0.03 0.02 0.02 -0.01 0.01 0.02 1.00 0.03 0.11 0.10 -0.02 -0.01 -0.01
T+3_month_activity 0.05 0.03 0.03 0.04 0.03 0.02 0.03 0.02 0.00 0.02 0.03 0.03 1.00 0.09 0.09 -0.00 -0.01 -0.01
T+6_month_activity 0.15 0.10 0.07 0.15 0.10 0.07 0.04 0.01 0.00 -0.00 0.05 0.11 0.09 1.00 0.30 0.00 -0.00 -0.00
T+12_month_activity 0.15 0.10 0.07 0.16 0.10 0.07 0.04 0.01 0.01 0.00 0.05 0.10 0.09 0.30 1.00 0.01 0.00 0.00
avg_spends_l3m -0.01 -0.01 -0.01 0.01 0.00 -0.00 0.01 0.04 0.68 -0.00 0.01 -0.02 -0.00 0.00 0.01 1.00 0.53 0.53
cc_limit -0.01 -0.00 -0.01 0.01 0.01 -0.01 -0.00 0.03 0.77 -0.00 -0.01 -0.01 -0.01 -0.00 0.00 0.53 1.00 1.00
cc_limit_laks -0.01 -0.00 -0.01 0.01 0.01 -0.01 -0.00 0.03 0.77 -0.00 -0.01 -0.01 -0.01 -0.00 0.00 0.53 1.00 1.00
active_30
active_60
active_90
cc_active30
cc_active60
cc_active90
widget_products
engagement_products
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
cc_limit_laks
1.0
0.8
0.6
0.4
0.2
0.0

Above scatterplot and heatmap shows there is moderate positive correlation with credit card limit and average spends in three months. This means that when the credit limit increases the average spending also goes up. Hence cc_limit is very important for the business.