Business Report

SMDM Project
Coded

MARCH 10, 2024

Authored by: Kartik Trivedi

Table of Contents

Executive Summary

PROBLEM 1

Business Problem

In the board meeting of Austo Motor Company, questions were raised over the effectiveness of companies marketing campaign, to offer recommendations and actionable insights aimed at enhancing and optimizing the company's existing campaign strategies this analysis was carried out.

Key Takeaways

1. By demography, customers are aged between 22 and 54, however, 50% customers are below 29 years of age who mostly men who prefer buying hatchbacks.
2. If we go by the make sedan are most preferred followed by hatchbacks while SUVs are least preferred.
3. By gender while men prefer sedans and hatchbacks more women prefer more SUVs and what this means is that on an average women purchase more high-priced cars as SUVs have the highest price amongst three.
4. While young people buy more cars, they prefer buying low priced and as the age of buyer increases though the volume comes down, average purchase price moves up.

Recommendations

1. Marketing campaigns, point of sale designing and staff training should be done considering differences in demographic behavior to improve customer's buying experience.
2. Special campaign specifically targeting women should be run to increase vehicle sales to women as they prefer purchasing high priced vehicles.
3. There are multiple biases in customer behavior based on their occupation, age, gender, sales team should be trained about these biases to achieve higher sales and better realization per sale.

PROBLEM 2

Key Takeaways

Based on the analysis top 5 variables identified that could help identify high attrition in credit card usage are

cc_active90, 60, 30: They help identify how the usage of credit card has dipped in last 90, 60, 30 days.

Occupation at source: It helps identify how attrition varies with occupation of the customer.

cc_limit: It helps identify that most of the customer who stopped using credit card have low credit limit.

PROBLEM 1

1.1 BACKGROUND INFORMATION

Austo Motor Company, a prominent car manufacturer renowned for its SUV, Sedan, and Hatchback models, recently discussed concerns regarding the effectiveness of its current marketing campaign during a board meeting. In response, this project has been approved to offer recommendations and actionable insights aimed at enhancing and optimizing the company's existing campaign strategies.

1.2 PROBLEM STATEMENT

The objective of this analysis is to carry out an in-depth study of the provided dataset in order to answer the relevant questions shared by the data science team and generate actionable insights which can be used to understand the demand of customers and help company improve customer experience.

1.3 METHODOLOGY

1. **Data Collection**: Customer data was provided by the data engineering team which contains customers demographic information along with details regarding make and price of the car bought.
2. **Data Cleaning and Pre-processing**: Dataset was checked for duplicates, missing values and bad data. Missing values, bad data and outliers were found in the dataset amongst which missing values and bad data were treated as per the procedure.
3. **Univariate Analysis**: Individual variables were analyzed using boxplot and histogram to understand distribution, central tendency and variability of variables.
4. **Bivariate Analysis**: Various demographic variables were examined to ascertain their correlation with car price and model type, with the aim of gaining deeper insights into customer preferences and behaviors when engaging with the company.
5. **Visualization Techniques:** In the report we have used histograms and boxplot for univariate analysis, in bivariate analysis, to understand correlation between numeric variables heatmap, pair plot, scatter plot and regression plot are used and line plot and bar plot are used to understand relationship between categorical and numeric variables.
6. **Tools and Software:** We have carried out the analysis using programming language python on Jupyter notebook. For this analysis Python libraries Numpy, Pandas, Matplotlib and Seaborn were used.
7. **Assumptions and Limitations:** During the univariate analysis, outliers were found in ‘No of Dependents’ and ‘Total salary’ variables, however, for ‘No of Dependents’ though the data type is numeric but we have considered it as a categoric variable as it can be used to cluster the data. As far as, ‘Total salary’ is concerned we found that it is a dependent variable which is sum of ‘Partner salary’ and ‘Salary’ and since both of them do not have any outliers so we did not treat the outliers for ’Total salary’.

1.4 DATA OVERVIEW

1.4.1 Data Description

Dataset has 1581 rows and 14 columns amongst which 8 columns, Gender, Profession, Marital Status, Education, Personal loan, House loan, Partner working and Make are of object type categorical in nature and 6 columns are numeric type of which Age, Salary, Partner salary, Total salary and Price are continuous and No of dependents is categorical.

1.4.2 Data Pre-processing

Data had missing values and some bad data for which following steps were carried out:

1. **Bad data treatment**: When checking for unique values in variables with object type data, Gender column has some incorrect entries with Femal and Femle entered in some cases. These entries were rectified by replacing them with the correct spelling, "Female."
   Statistical summary of 'Gender'

```
count     1528
unique       2
top       Male
freq      1199
Name: Gender, dtype: object
```

2. **Missing Values**: Gender and Partner salary had missing values, here, for gender, data type is object and for Partner salary it is numeric, method used to replace the missing values for different variables depend upon the data type. In case of object type mode value ie. Most frequent occurrence, is used which in this case it was male so 53 cells with missing values in Gender column were replaced with ‘Male’.
   In case of numeric values, we use mean that is average if data in that column is normally distributed but if data is skewed then we use median value. In our case, additionally, we observed that there is a variable from which we can know whether partner is working or not and if partner is not working then their salary is 0, so we segregated the data further by imputing 0 as partner salary in case partner is not working and for working partner median value for Partner salary of working partner were used since the data is right skewed.

   statistical summary of 'Partner salary' grouped by 'Partner working'

Partner_working	count	mean	std	min	25%	50%	75%	max
No	623.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
Yes	852.0	35014.906103	12046.487544	100.0	28400.0	32900.0	40300.0	80500.0

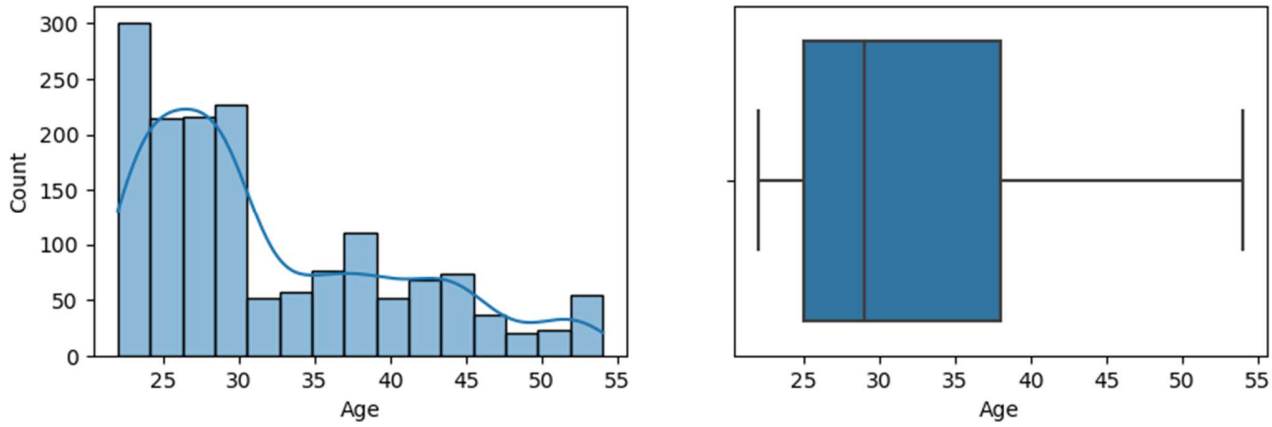
1.4.3 Statistical Summary

	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	1581	2	Male	1252	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	1581.0	NaN	NaN	NaN	19202.466793	19526.571322	0.0	0.0	25100.0	38000.0	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

1.5 ANALYSIS AND FINDINGS

1.5.1 Univariate Analysis

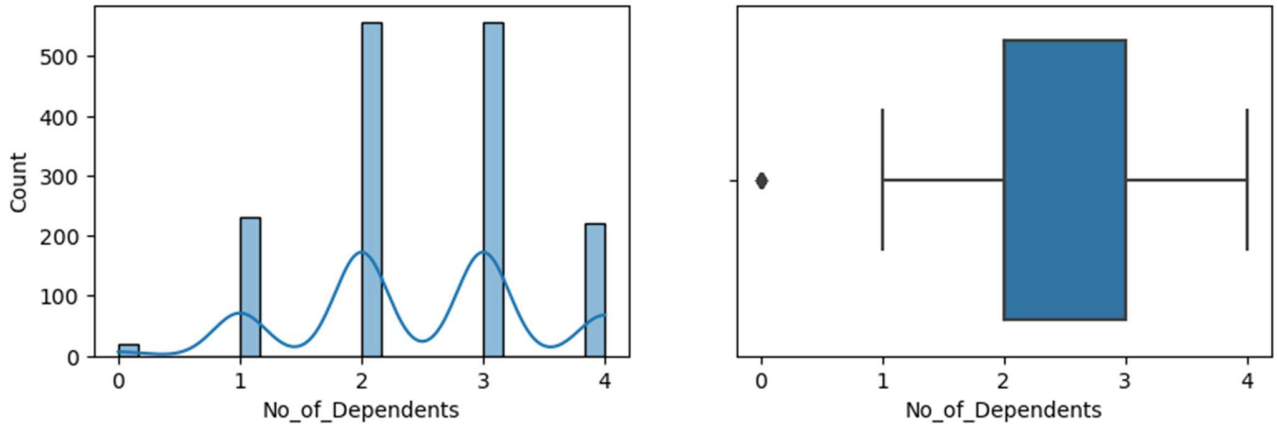
1.5.1.1 Distribution of Age



Observation

As per the above plot most people in the dataset are aged below 30 with median value of 29 and highest number aged around 22 and 23. This data does not have any outliers and from the plot we can conclude that our targeted demography is young people aged between 22 and 30, maybe, first time car buyers, which could be explored further.

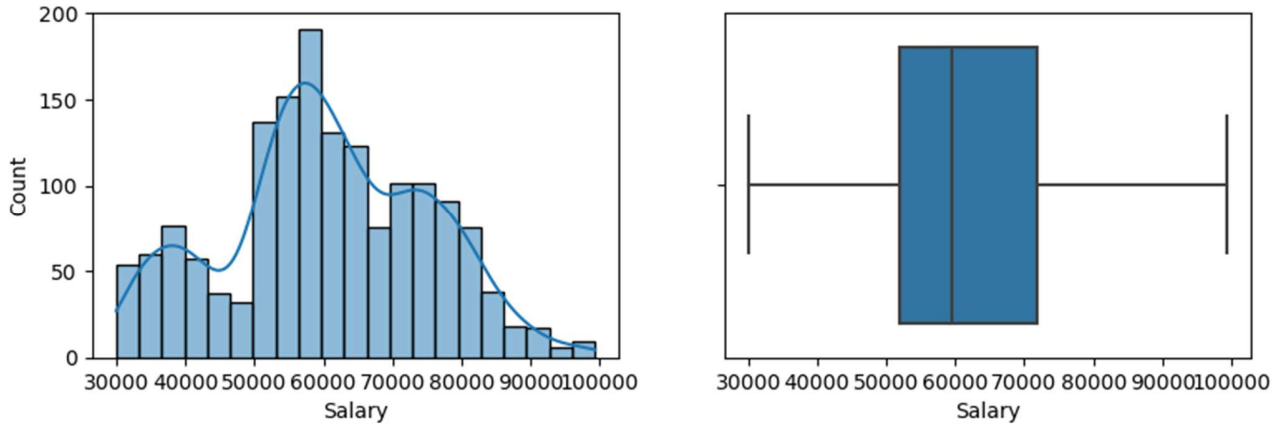
1.5.1.2 Distribution of Number of dependents



Observation

Since No of Dependents has numeric values so it got plotted, however, it is an categorical column as data in it cannot be continuous. The plot indicates that the majority of customers have dependents, with no dependents being an outlier. Both the first quartile and median values are 2, and multiple modes are observed at 2 and 3. Understanding who these dependents are could offer valuable insights aligned with our business objectives.

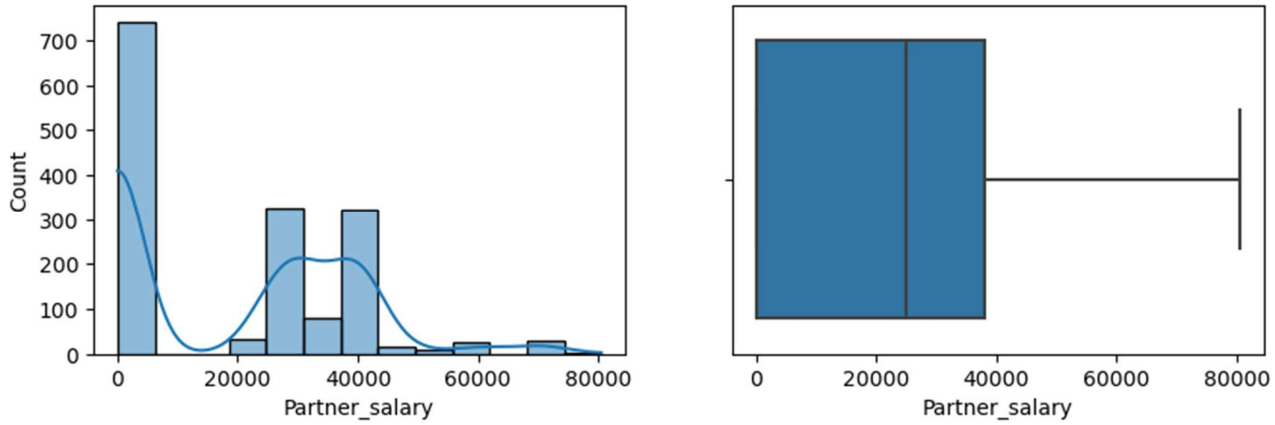
1.5.1.3 Distribution of Salary



Observation

The plot looks like a normal distribution with median value of around 60000 and IQR range of 50000 and 71000.

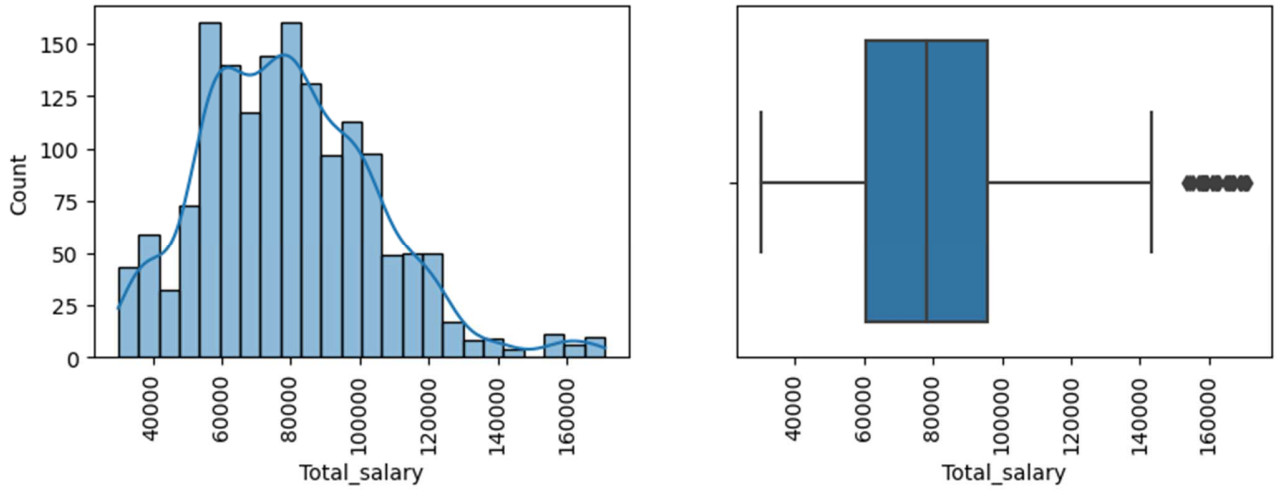
1.5.1.4 Distribution of Partner salary



Observation

From the Plot it's evident that more than 40% of customers have partners with minimal or zero earnings, contributing to a slight right skew in the data. Notably, there are no outliers observed. For individuals with earning partners, their partners salary exhibits multiple modes, first clustered around 25000 and second around 40000.

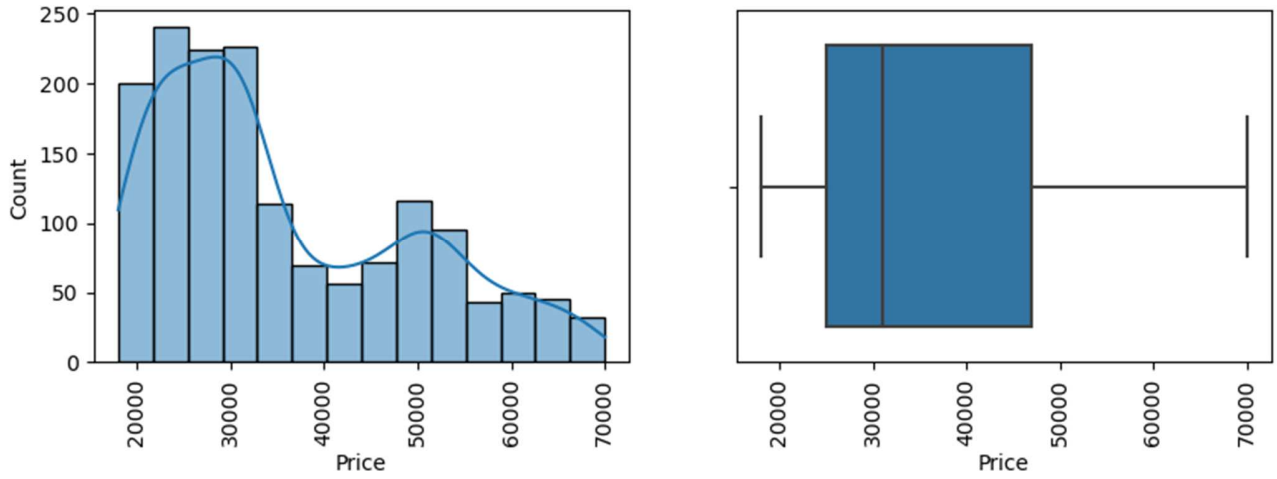
1.5.1.5 Distribution of Total salary



Observation

Total salary has a median value of around 80000 with two modes one coming around 60000 and another around 75000. From the boxplot we can clearly imply that the plot is right skewed having outliers.

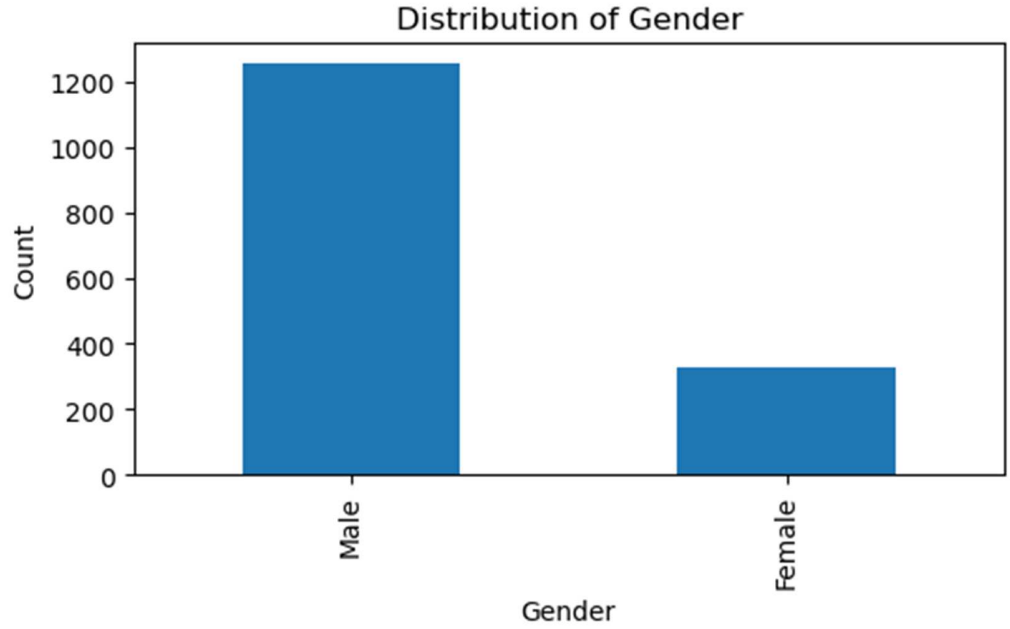
1.5.1.6 Distribution of Price



Observation

For Austo Motor Company most sales come for the cars priced in the range of 20000 and 30000 with a median value of around 31000 and there are no outliers.

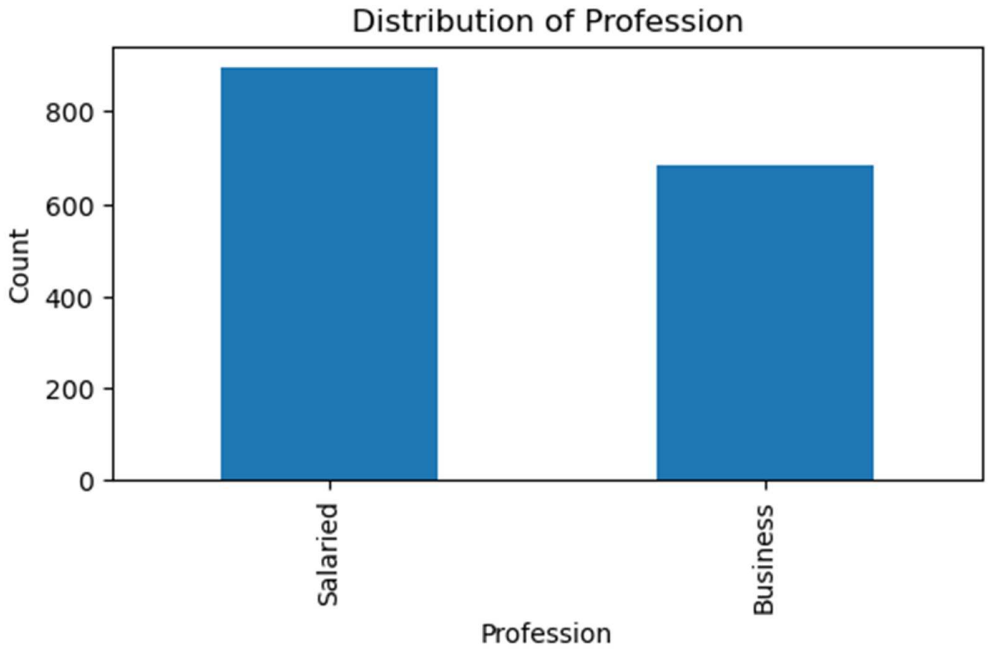
1.5.1.7 Distribution of Gender



Observation

In our data over 75% of customers are male.

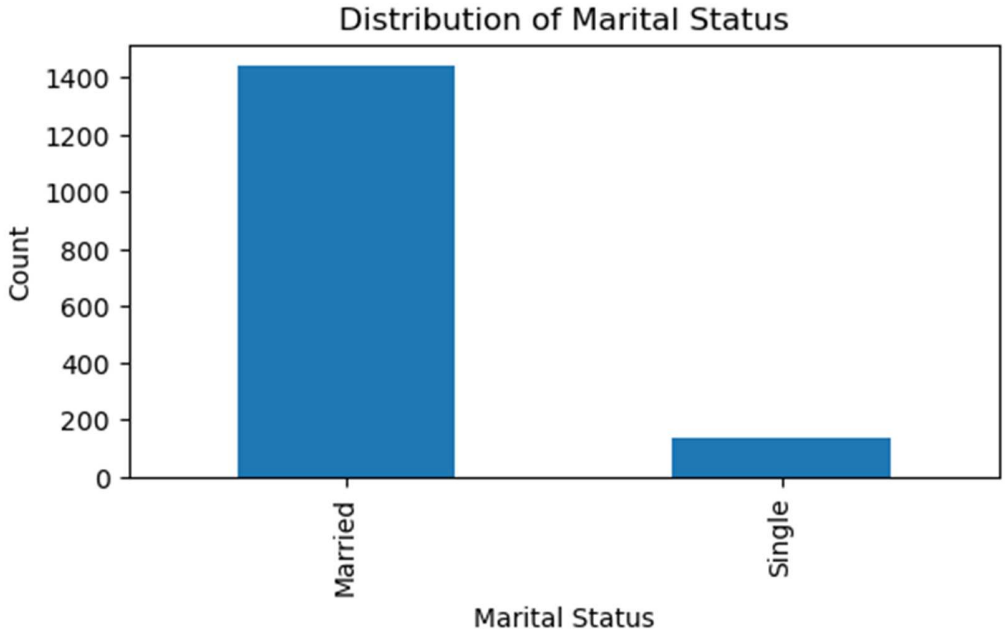
1.5.1.8 Distribution of Profession



Observation

The majority of customers are employed individuals, comprising approximately 60% of the dataset.

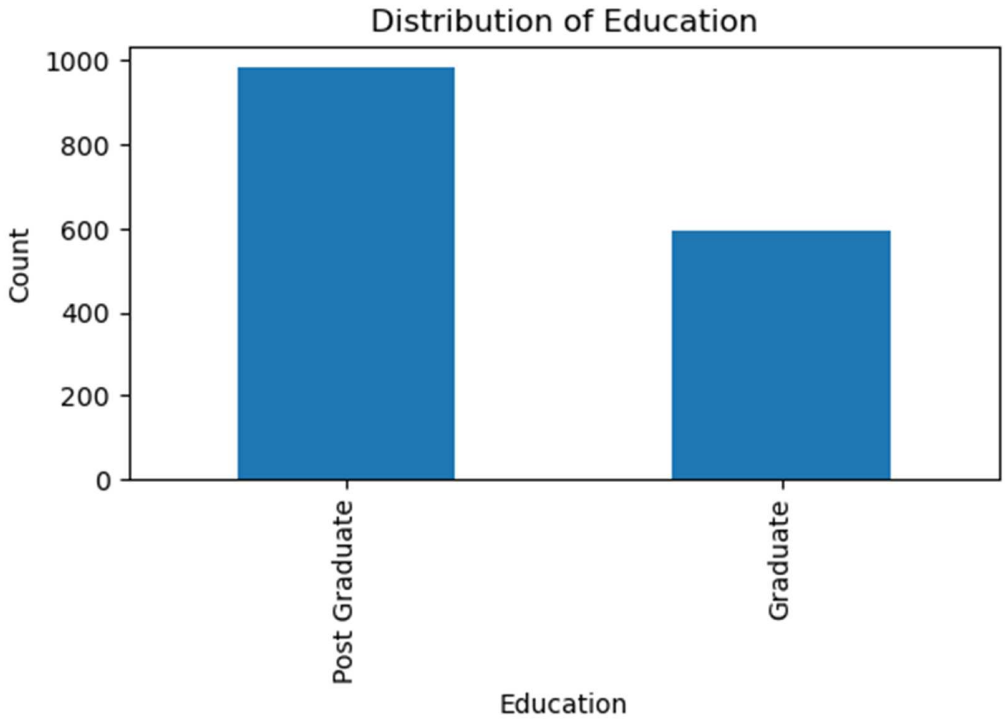
1.5.1.9 Distribution of Marital Status



Observation

Almost 90% of customers in the dataset are married.

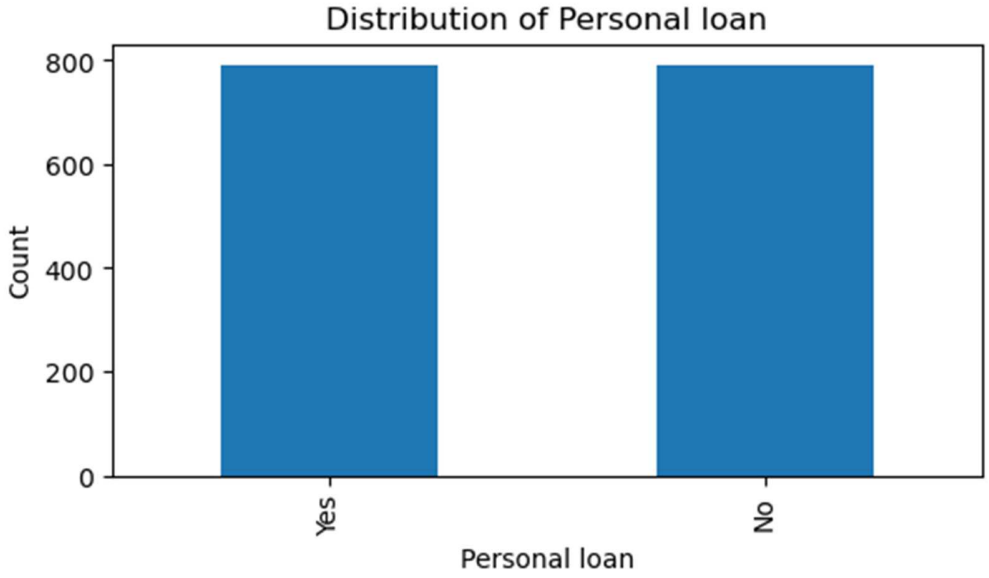
1.5.1.10 Distribution of Education



Observation

Every customer in the dataset holds either a graduate or postgraduate degree, with approximately 60% having completed postgraduate studies.

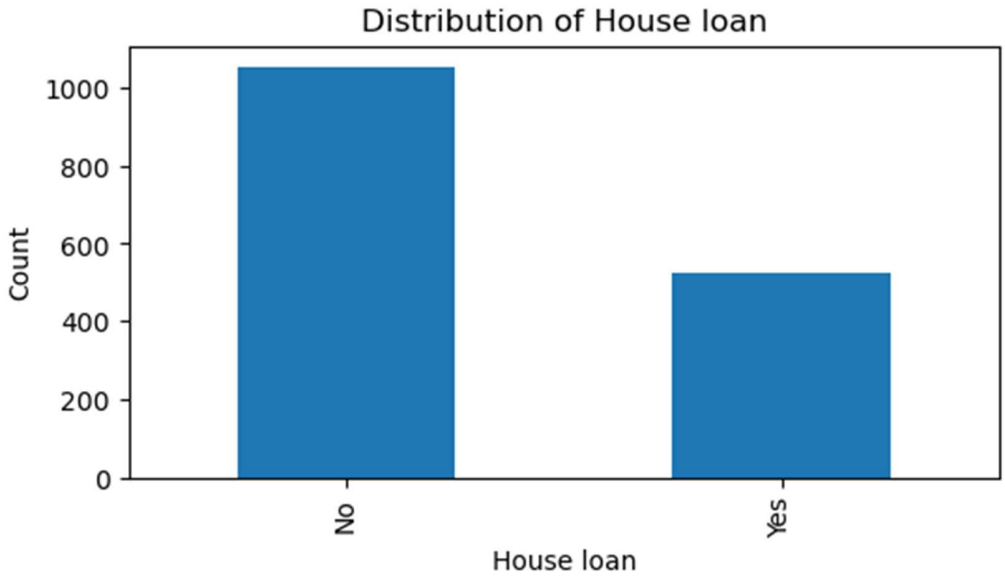
1.5.1.11 Distribution of Personal loan



Observation

In the dataset, the number of individuals who have taken a personal loan is equal to the number of those who have not.

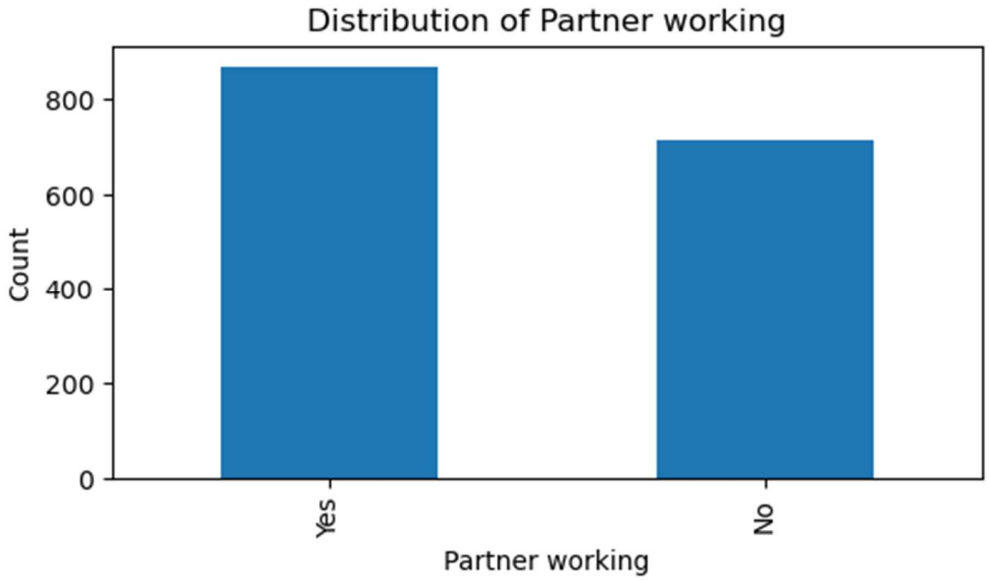
1.5.1.12 Distribution of House loan



Observation

Roughly two-thirds of customers do not carry any housing loan liability.

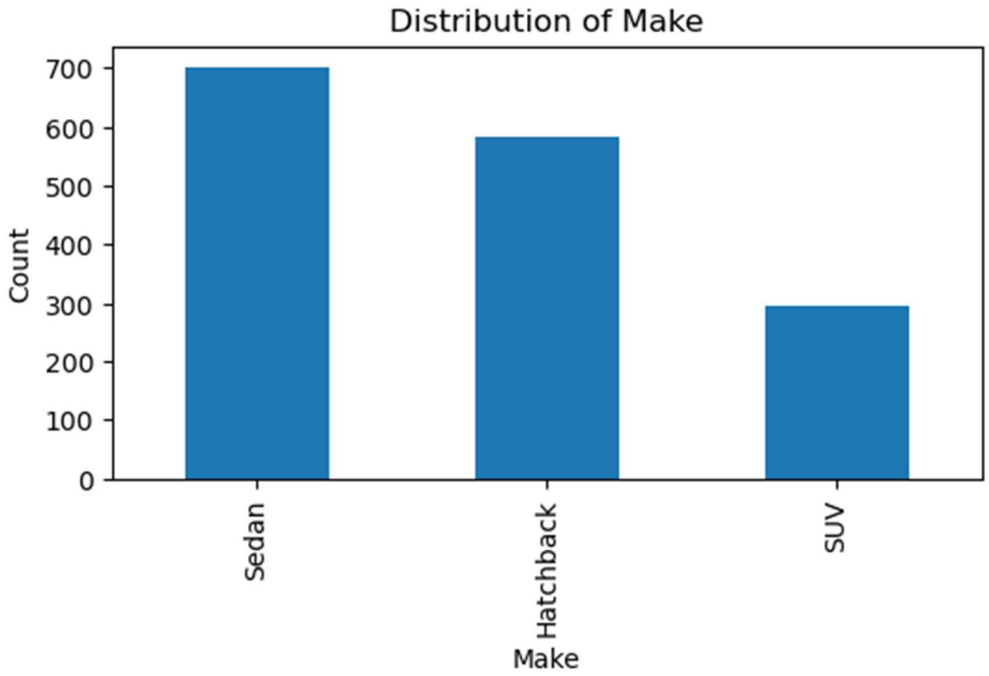
1.5.1.13 Distribution of Partner working



Observation

More than 50% of the customers have working partners.

1.5.1.14 Distribution of Make



Observation

Sedans are the most preferred car types followed by hatchbacks while SUV are the least preferred accounting for less than 20% in the dataset.

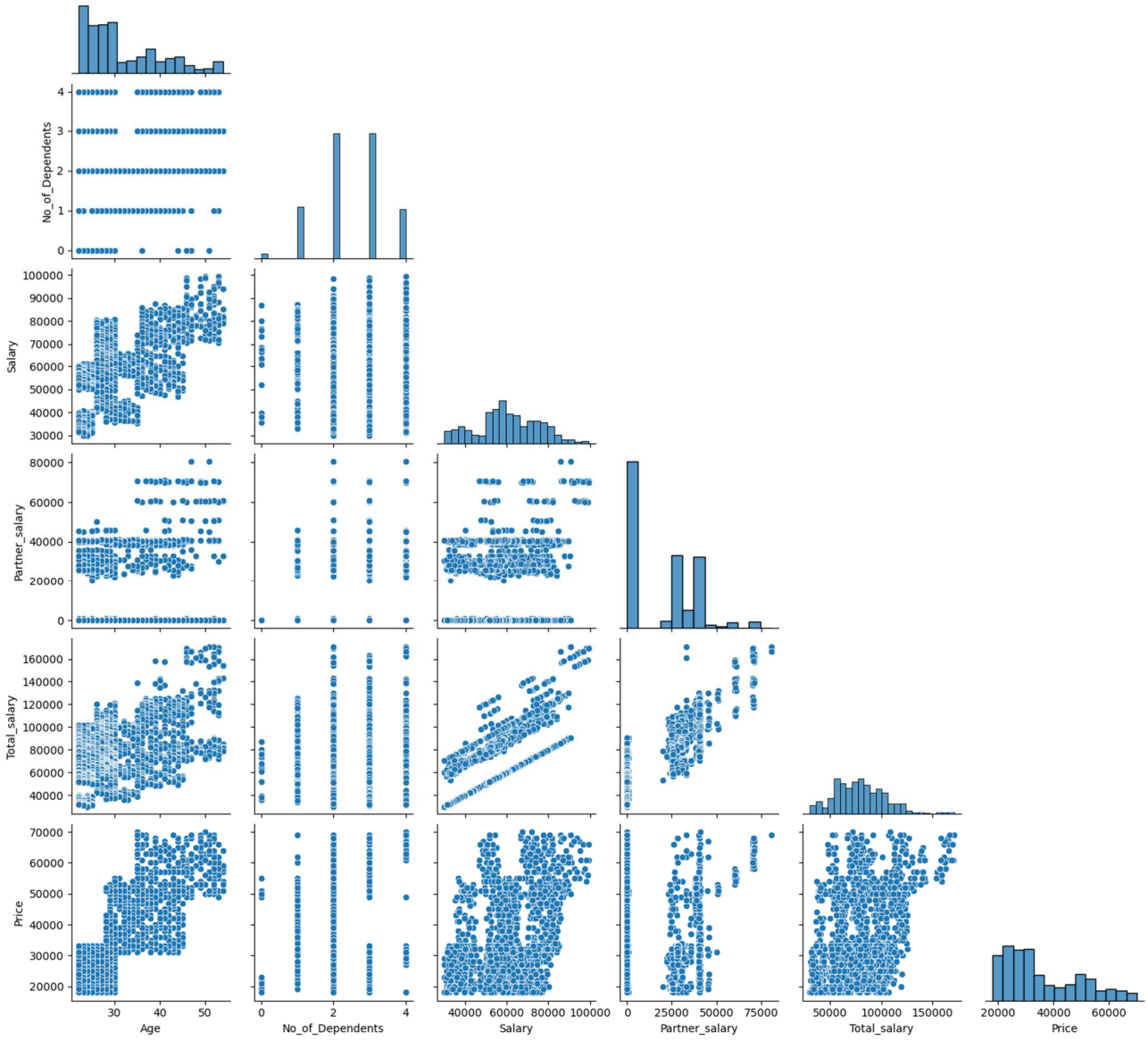
1.5.1.15 Key Takeaways of Univariate Analysis

1. Over 50% of our car buyers are aged between 22 amd 30 with median value of around 29 and 75% aged below 38 which means that our cars are more applealing towards young car buyers.
2. Over 90% of the car buyers are married and over 75% are male.
3. Most preferred car make is sedans while SUV's are least preferred.

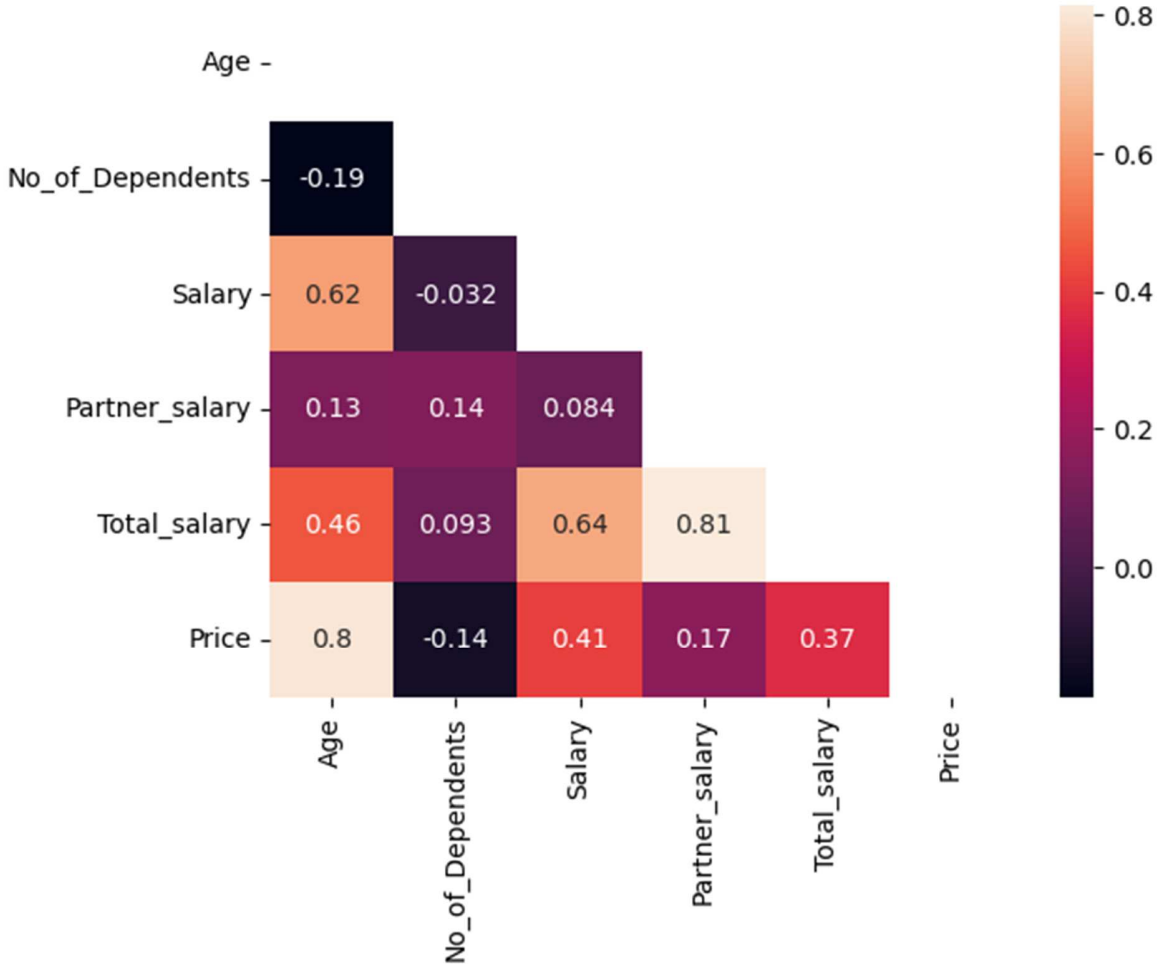
1.5.2 Bivariate Analysis

1.5.2.1 Correlation Analysis of numeric variables

1.5.2.1.1 Pair plot



1.5.2.1.2 Heat map

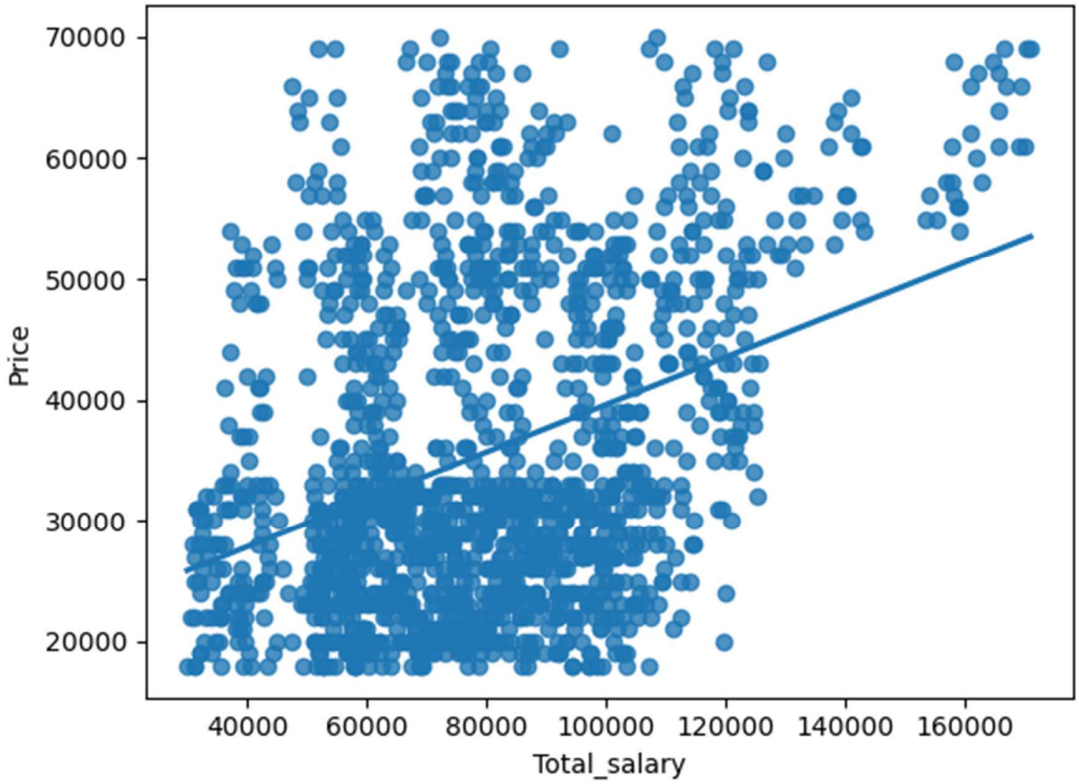


Observation

Based on the above plots we can conclude that amongst the numeric variables only the following variable sets show considerable correlation. Age and Salary and Salary and Total salary have strong positive correlation and Age and Price and Partner salary and Total salary have very strong positive correlation, correlation shows the probability of a variable moving up or down based on change in the value of another variable, however, correlation does not mean causation.

1.5.2.2 Relationship between Total salary and Price

We had decided not to treat the outliers for Total salary column as we considered it a dependent column from which we might gather some additional information which might have some business relevance which we try to find here.



Observation

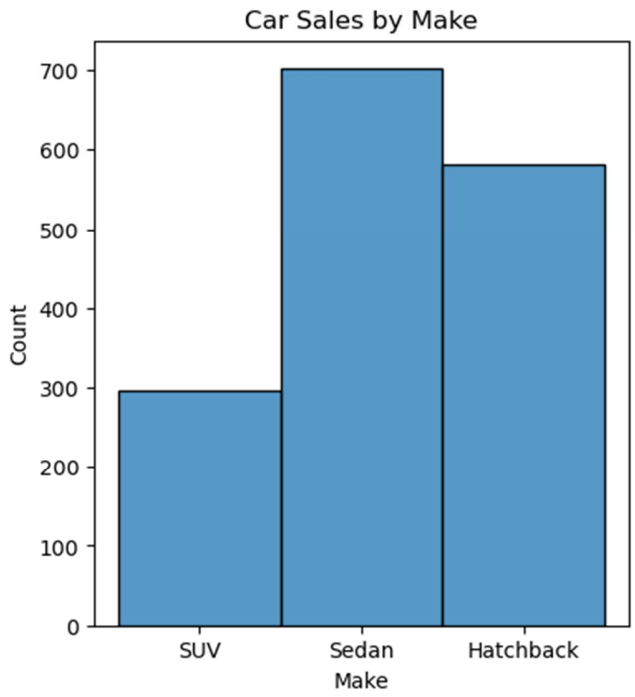
From the above plot we can clearly interpret that as the income increases, customers purchase more expensive cars.

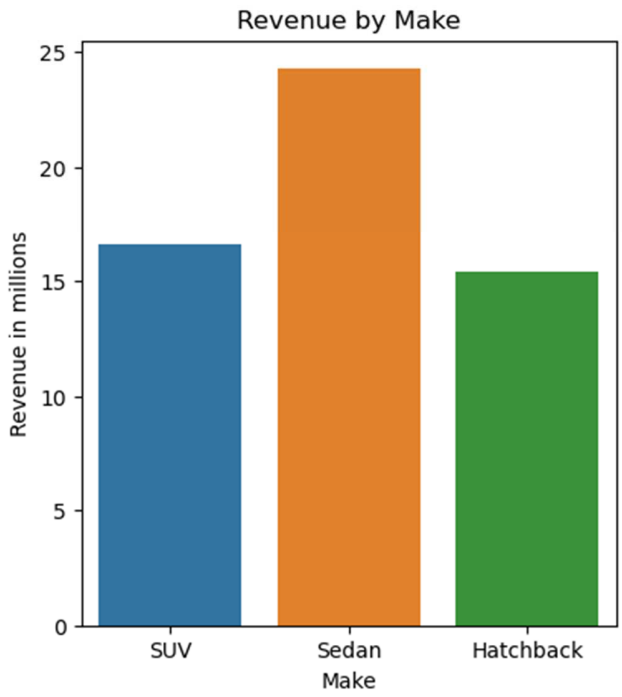
Customers with a total income exceeding 140,000 are considered outliers. From the plot provided, it's apparent that these customers have purchased cars priced above 50,000, which falls into the higher price range for the company.

1.5.2.3 Relationship between categorical and numeric variables

In this section, our focus will primarily be on the "Price" and "Make" columns, as the outcomes of bivariate analysis using these two columns will directly align with our business objectives.

1.5.2.3.1 Sales Volume and Revenue by Make

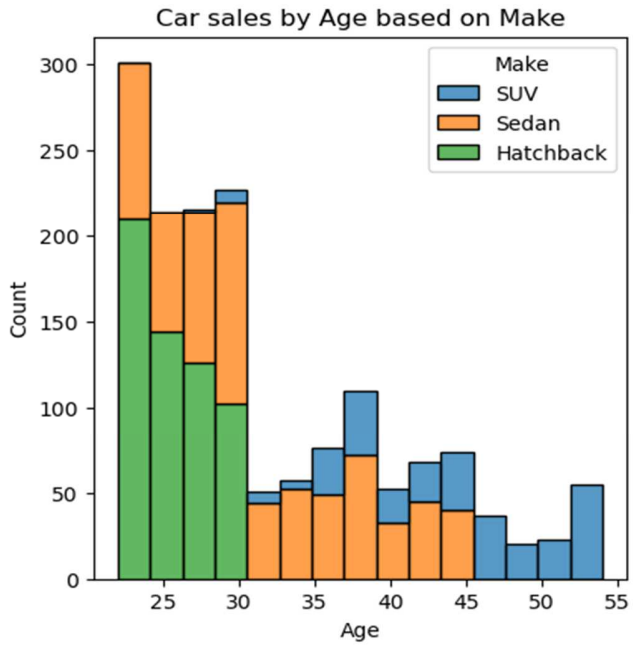


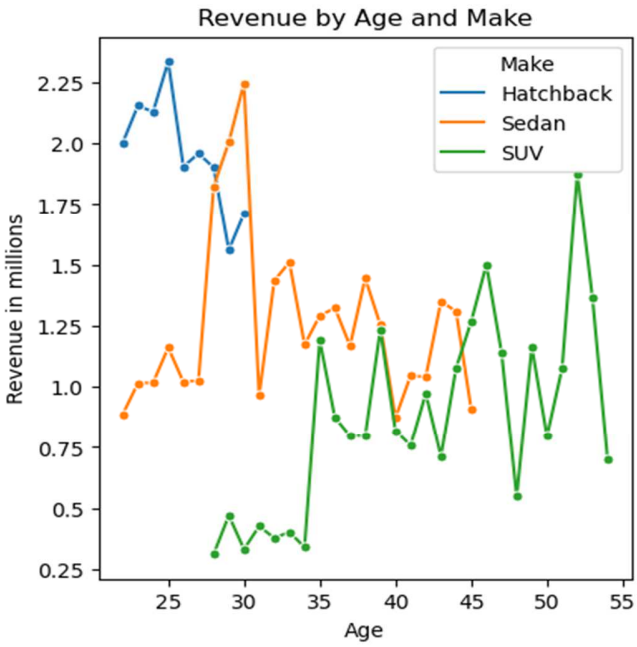


Observations

1. Sales by revenue and volume are highest for sedans.
2. By volume hatchback sales are almost twice that of SUV's, however, since SUVs are higher priced the revenue generated by sales of SUV's is slightly higher than that from hatchbacks.

1.5.2.3.2 Sales Volume and Revenue based upon customer Age and Make

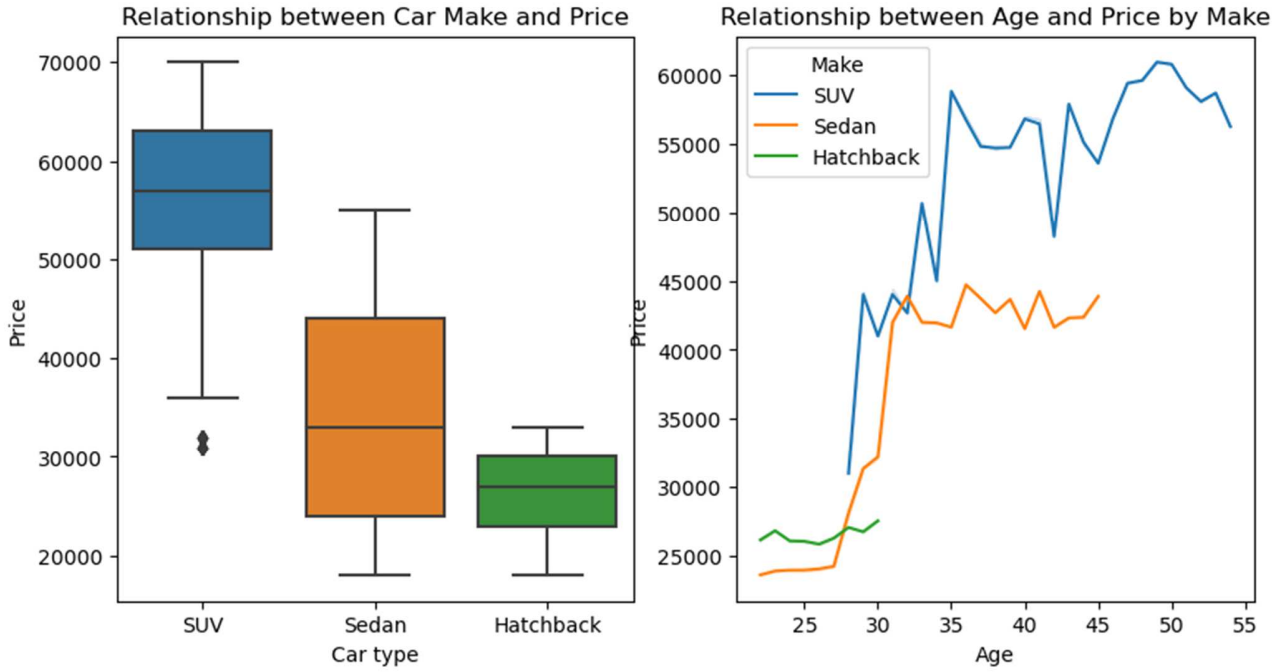




Observation

Demographically, sedans exhibit the highest acceptance and most consistent performance in terms of both volume and revenue. Conversely, hatchbacks are favored by younger individuals, with preferences shifting towards larger vehicles such as SUVs as age increases.

1.5.2.3.3 Average Price by Age and Make

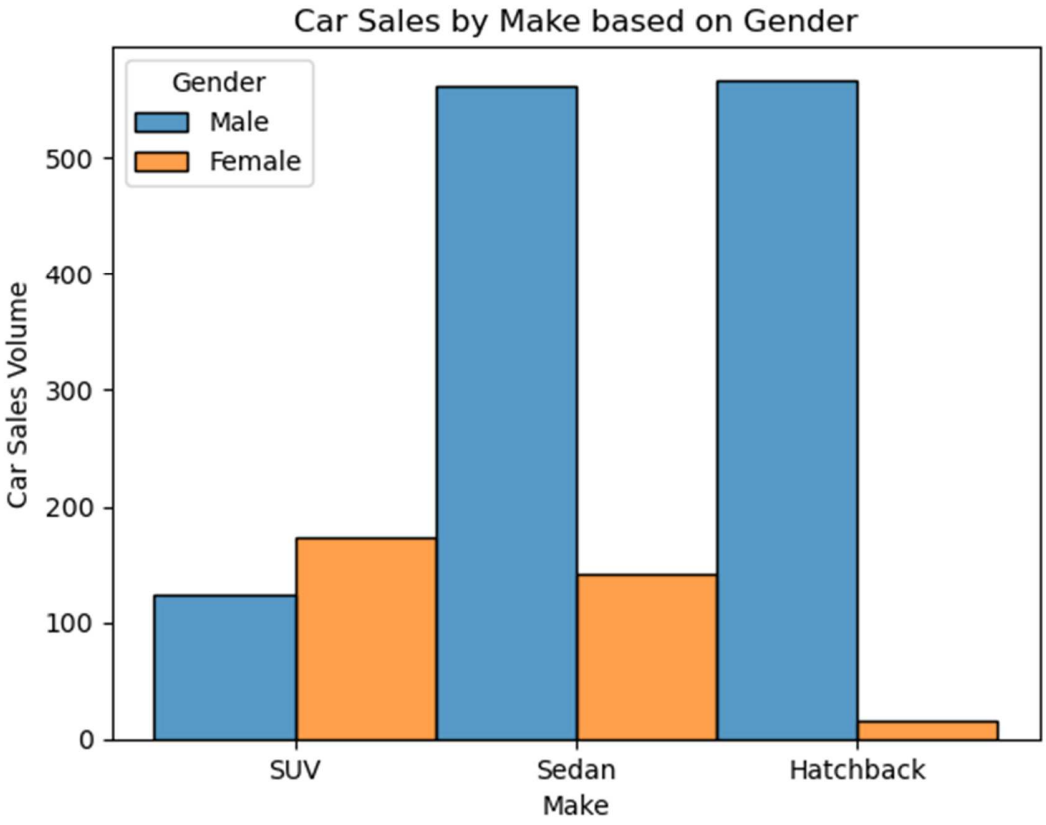


Observations

1. SUV's are priced highest followed by sedans and hatchbacks.
2. Hatchbacks are preferred by younger people aged under 30 while sedans are preferred by those aged under 45 and SUV's by those aged above 30.
3. As age increases people tend to buy higher priced cars.

1.5.3 Key Question Answers

1.5.3.1 Do men tend to prefer SUVs more compared to women?



Observations

1. As per the above plot it is very evident that while around 125 men bought the SUVs the figure for women is around 175 so we can conclude that women tend to prefer SUVs more than men.
2. Additionally, we have another important takeaway from this plot, we know from earlier plots that young people under 30 years prefer hatchbacks while older people over 30 years prefer SUVs and in the above plot number of men mostly buy hatchbacks and sedans while women mostly buy SUV's and Sedans. Based upon Gender and Age preference we can conclude that men tend to buy cars at an younger age when compared to women, however, since cost of SUV is higher than Hatchbacks and Sedans so average revenue per transaction is higher for women than men.

```
average age of car purchase be gender

Gender
Female    39.525836
Male      29.924121
Name: Age, dtype: float64
```

```
average price of car by gender

Gender
Female    47705.167173
Male      32416.134185
Name: Price, dtype: float64
```

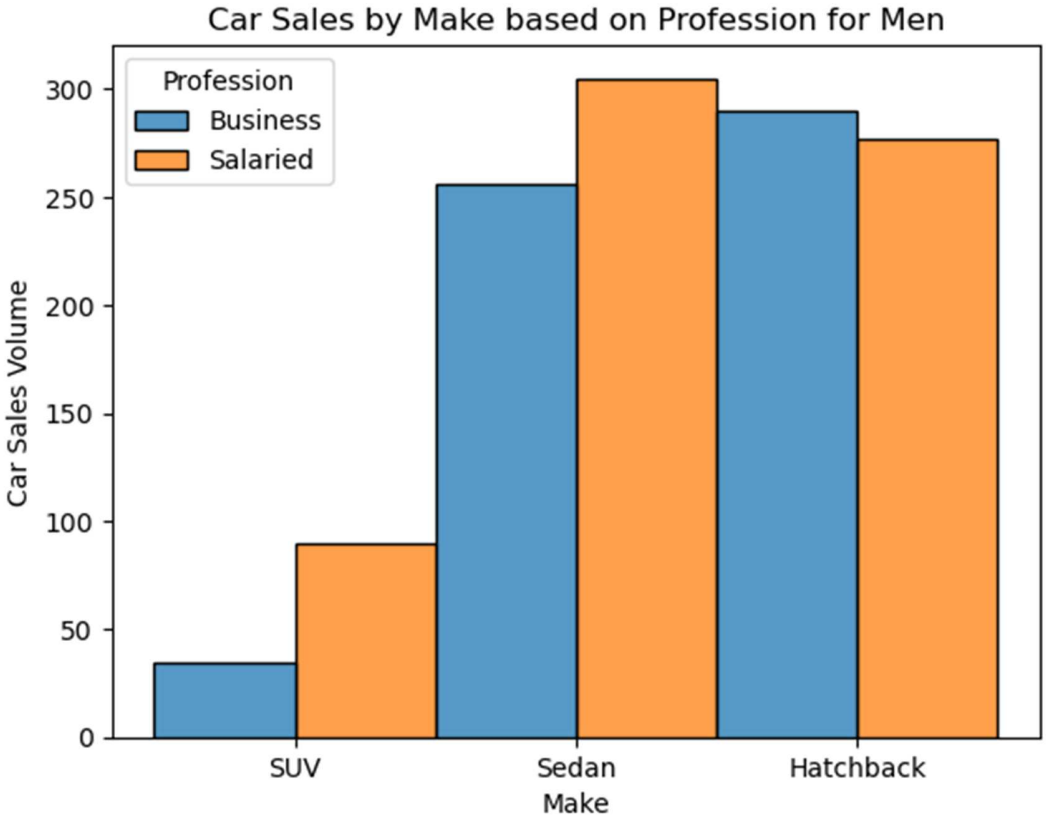
1.5.3.2 What is the likelihood of a salaried person buying a Sedan?

```
probability of buying different makes for salaried individuals

Make
Sedan        0.441964
Hatchback    0.325893
SUV          0.232143
```

Observation

Salaried individuals mostly prefer to buy sedans, as per our data there is 44% of salaried individuals prefer buying sedans followed by 32.5% for hatchbacks and only 23% purchase SUV.

1.5.3.3 What evidence or data supports Sheldon Cooper's claim that a salaried male is an easier target for a SUV sale over a Sedan sale?



Observation

The data contradicts the claim that salaried male is an easier target for SUV sale over Sedans as under 100 salaried men opted for SUV against 300 Sedans.

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

```
Gender
Female    47705.167173
Male      32416.134185
Name: Price, dtype: float64
```

Observation

Women tend to spend around 1.5 times more as compared to men on purchasing automobiles. However, it is important to also note that men tend to buy automobiles at younger age than woman.

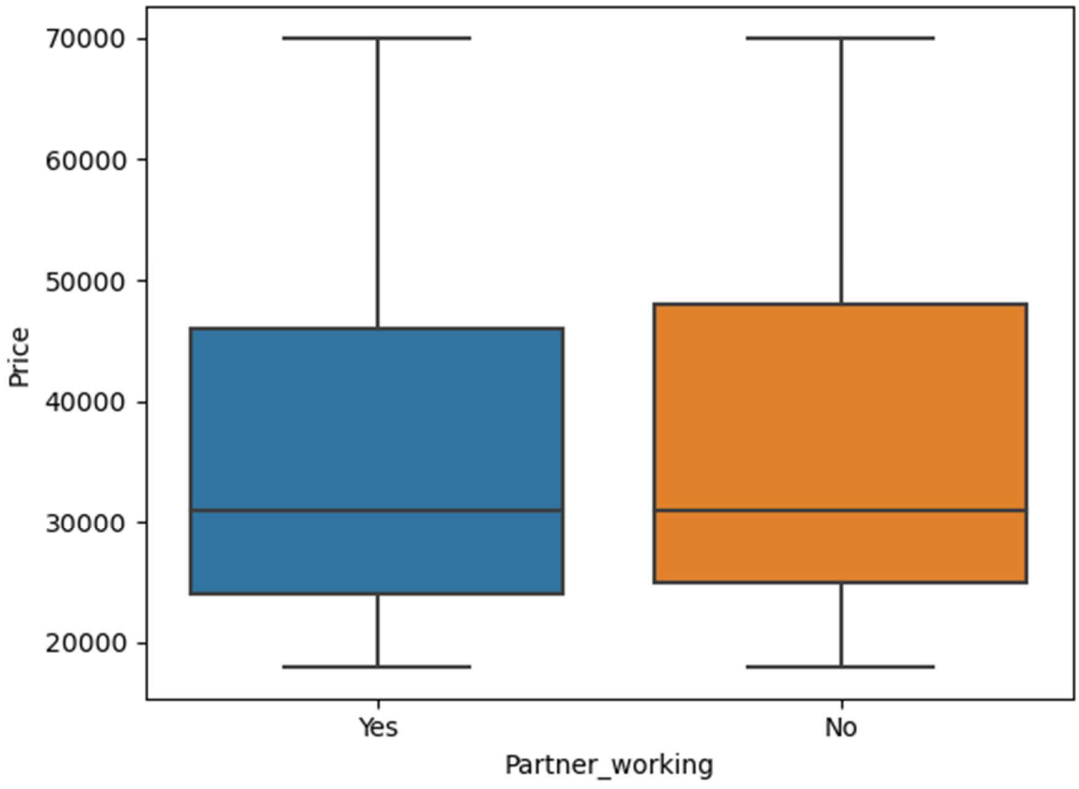
1.5.3.5 How much money was spent on purchasing automobiles by individuals who took a personal loan?

```
Amount spent on purchasing cars for individuals who have personal loan

Personal_loan
No     28990000
Yes    27290000


Average amount spent on purchasing cars for individuals who have personal loan

Personal_loan
No     36742.712294
Yes    34457.070707
```

Observation

Individuals with personal loans tend to spend marginally less on purchasing automobiles with an average purchase spend of 34457 compared to those without personal loans who spend 36743.

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



statistical summary of car price based on working partner

Partner_working	count	mean	std	min	25%	50%	75%	max
No	713.0	36000.000000	13817.734086	18000.0	25000.0	31000.0	48000.0	70000.0
Yes	868.0	35267.281106	13479.532555	18000.0	24000.0	31000.0	46000.0	70000.0

Observation

Those who do not have working partner have slightly higher probability of purchasing higher-priced cars, as the median price for both cases is same, however, 75 percentile value for those without partner working is 48000 while those with working partners is 46000 which means those without working partner have spent more on higher price cars.

1.6 INSIGHTS AND RECOMMENDATIONS

1. While the demographic spread of customers spans from 22 to 54 years old, a significant portion—over 50%—falls under the age of 29, primarily comprising men who exhibit a preference for purchasing hatchbacks and lower-priced sedans. Conversely, older customers tend to favor higher-priced cars, with price demonstrating a strong correlation, resulting in increased revenue per sale. Tailoring marketing campaigns and optimizing point-of-sale strategies to cater to these demographics could effectively attract more customers.

   Moreover, investing in training for sales staff and ensuring the presence of team members who can readily connect with customers, understand their needs, and effectively address their requirements could significantly enhance the sales conversion rate and provide customers with a more pleasant buying experience.

2. Running a marketing campaign tailored at women could be initiated, though women constitute around only 25% of the total customers, they prefer buying more high-priced cars especially SUVs which are highest priced cars manufactured by Austo Motor Company. This could be a great growth opportunity especially in revenue terms.

3. Initiating a tailored training plan for the sales team could prove beneficial in helping them understand customer biases. For instance, recognizing that salaried men often gravitate towards sedans, while businessmen may prefer purchasing hatchbacks, can enable the team to effectively cater to these preferences. Similarly, understanding how customers' age influences their preference for different model types allows for targeted sales efforts, potentially leading to increased sales of higher-priced models. By tapping into these biases, the sales team can enhance their ability to meet customer needs and drive revenue growth.

1.7 SCOPE OF FURTHER STUDY

1. Most car buyers have dependents, collecting data about them and understanding who these dependents are could be beneficial.

2. Understanding the average car replacement cycle can significantly enhance our ability to target customers whose vehicles are approaching this replacement age. This data holds immense value, especially considering that individuals typically opt for higher-priced cars as they age. By leveraging this information, we can effectively tailor our marketing efforts to capture this segment of the market and capitalize on their purchasing behavior.

3. Gathering data on whether a partner owns a car or not would be advantageous. Our findings indicate that individuals with working partners tend to purchase fewer higher-priced cars compared to those without working partners. However, in cases where partners are employed, the total household income—evidenced by the fact that the mean and median values of the total salary are 30% higher than those of the individual salary—suggests the possibility that both partners own cars or are considering purchasing them.

Problem 2

2.1 BACKGROUND INFORMATION

GODIGT Bank, a private mid-size bank, is facing high attrition rate in its credit card usage which is having adverse effect on bank business as interest from credit card is one of the important sources of income for the bank. Due to this bank is reevaluating its credit card policy so that customers receive right card for higher spending and intent, resulting in profitable relationship.

2.2 PROBLEM STATEMENT

GODIGT Bank is facing high credit card attrition, necessitating the identification of key variables that could help identify the underlying causes and subsequently reduce attrition.

2.3 METHODOLOGY

1. **Data Overview:** Credit card data was provided by data engineering team which contains information about customers like credit card limit, occupation at time of applying, annual income at the time of applying, number of convenience and loan/investment products held, credit card issuer name, credit card type etc.
2. **Data Cleaning and Pre-processing:** Data was checked for duplicates, there we some columns which were not of use in this analysis, they were deleted and missing values were imputed.
3. **Univariate Analysis:** Univariate analysis was done for all variables to identify variable that would be most useful in understanding the reasons behind declining credit card usage.
4. **Bivariate Analysis:** Bivariate analysis was carried since not all the relevant variables could not be identified in univariate analysis.
5. **Visualization Techniques:** Histogram, boxplot, pair plot and heat map are used in the analysis.
6. **Tools and Software:** This analysis is done using programming language python on Jupyter notebook. Libraries used includes Numpy, Pandas, Matplotlib and Seaborn.
7. **Assumptions and Limitations:** Transactor_revolver column had 38 missing values, these 38 rows belonged to those credit card users who have become inactive so for this analysis we considered them to be transactors and T was imputed in place of null value.

2.4 VARIABLES IMPACT ANALYSIS

2.4.1 Relevance

Five relevant variables identified are

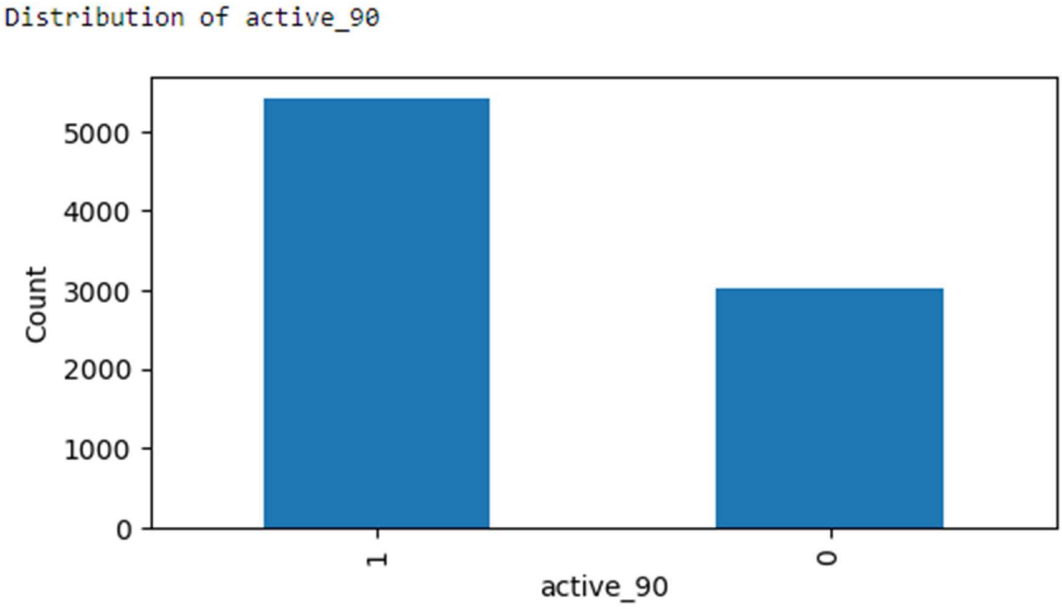
1. cc_active90
2. cc_active60
3. cc_active30

4. Occupation_at_source
5. cc_limit

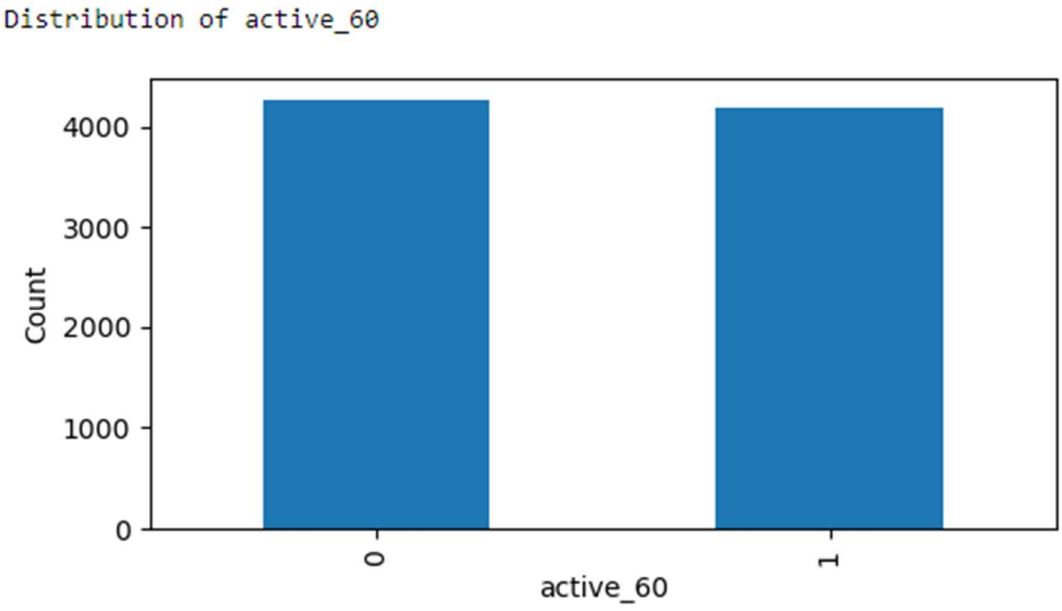
The above five variables are insights regarding the magnitude of attrition, how it is taking place and why it is taking place.

2.4.2 Data Insights

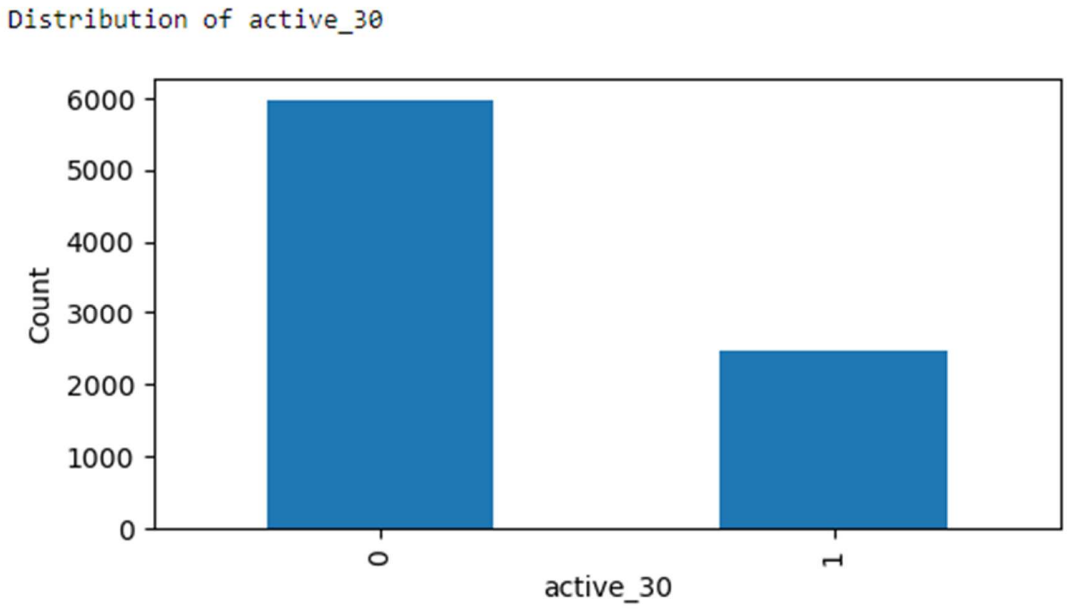
1. cc_active90



2. cc_active60



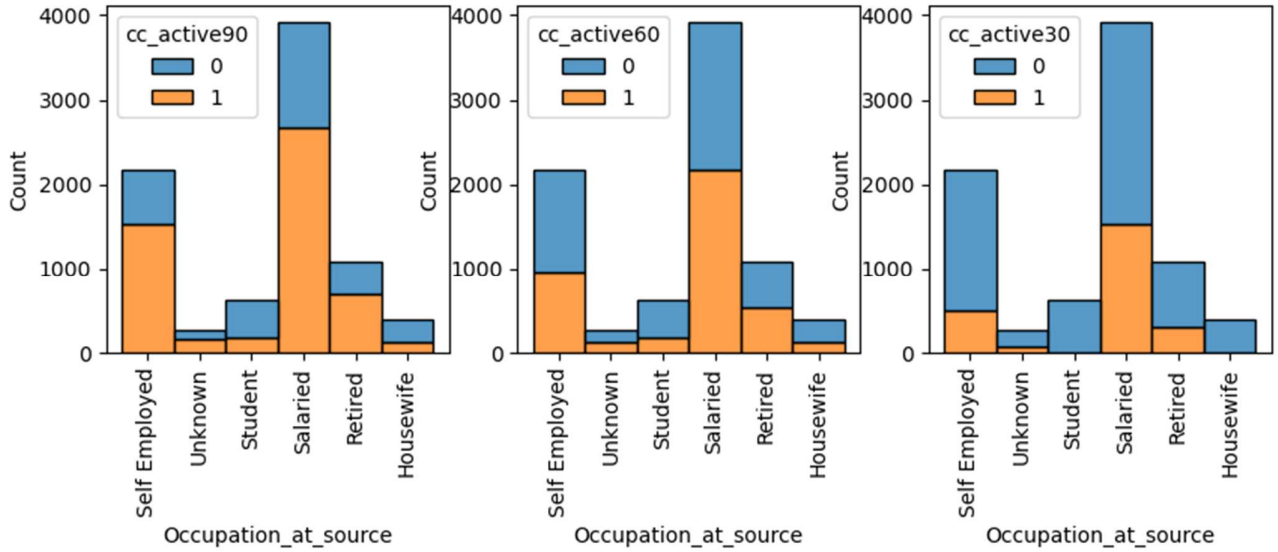
3. cc_active30



Observation

These three variables together helped in identifying how customer attrition has increased over the 90 days period, when they are analyzed along with other variables, they helped identify which variables have crucial information available regarding the attrition.

4. Occupation_at_source

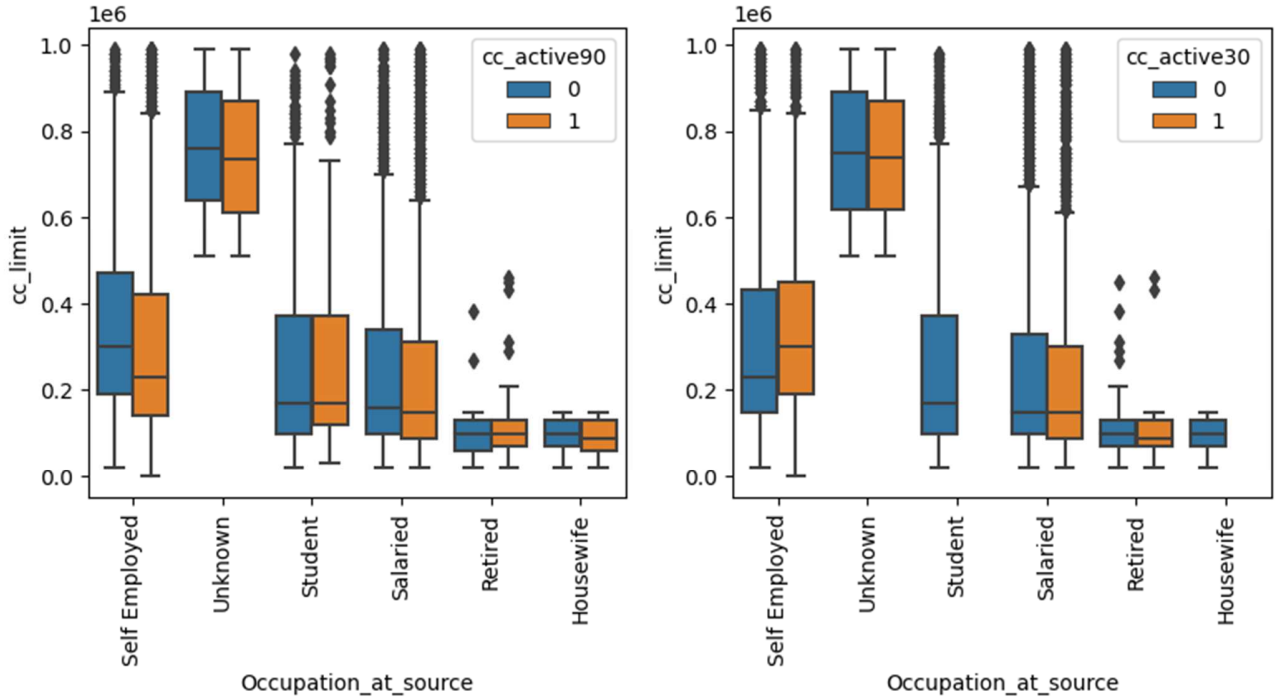


Observation

When histogram for Occupation at source was plotted side by side stacked by cc_active90, 60 and 30, we was evident that while attrition was universal, the magnitude was different based on

occupation which is evident from the fact that while students and housewife have ceased using their credit card in the last 30 days the effect is considerably less for salaried individuals

5. cc_limit



Observation

When credit card limit was observed in relation with occupation and credit card activity over last 30 and 90 days, we found that in most cases median value and in some cases even the entire box has moved up for active user in last 30 days when compared to the box plot for last 90 days and vice versa has happened for inactive user which means that most customers who have become inactive are those with lower credit limit and there is a high probability they have switched to another bank which might be providing them with higher credit limit.

2.4.3 Business Impact

These variables have helped us identify how attrition has taken place over the past 90 days, how for different occupations the attrition was different and by analyzing the distribution and central tendency of cc limit by occupation over credit card activity for last 90 and 30 days we were able to identify that customers with low credit limit were the ones transiting. These variables would help identify the changes the credit card usage by days, credit limit and demography of the customers