SMDM PROJECT

Submitted by,
VIDYA V

PGPDSBA.O.2023.B
07.05.2023

CONTENTS

Case 1: AUSTO MOTOR COMPANY

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 task here is to perform a thorough analysis of the data and come up with insights to improve the existing campaign.

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)

On a cursory viewing, the following are the observations about the dataset:

- It contains information about 1581 customers
- 14 different aspects about each customer are available:
  - General attributes- age, gender,profession, marital_status,education, the number of dependants
  - Loan related attributes- whether personal or home loans have been taken by the customer
  - Salary related attributes- whether the partner is working or not, the salary of the customer,partner and the total salary
  - Car related attributes- The car model owned and the price
- Of these, the age, number_of_dependants, salary, partner_salary,total_salary, price are numeric variables
- Others are object type categorical variables, with personal_loan, home_loan, partner_working as boolean variables with yes or no values

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

On conducting a quality check,

- There were missing values in the gender field and partner_salary field.
  - For the gender field, imputing with the mode of the field was done, and it was verified that this imputation did not cause any drastic changes in the data.
  - The partner_salary field can contain null values if the partner is not working. Hence, these were cross checked with the partner_working field, on those fields where the partner was working, but the partner_salary contained null values were imputed with the difference between the total_salary field and the Salary field.
- There were bad values in the gender field. The same were treated
- It made better sense to convert all the salary fields as float type for easy and accurate computations, and hence the same was done.
- Also, the no_of_dependants field contained numeric values, for the sake of the analysis, it was converted to a categorical field.

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

Univariate Analysis:

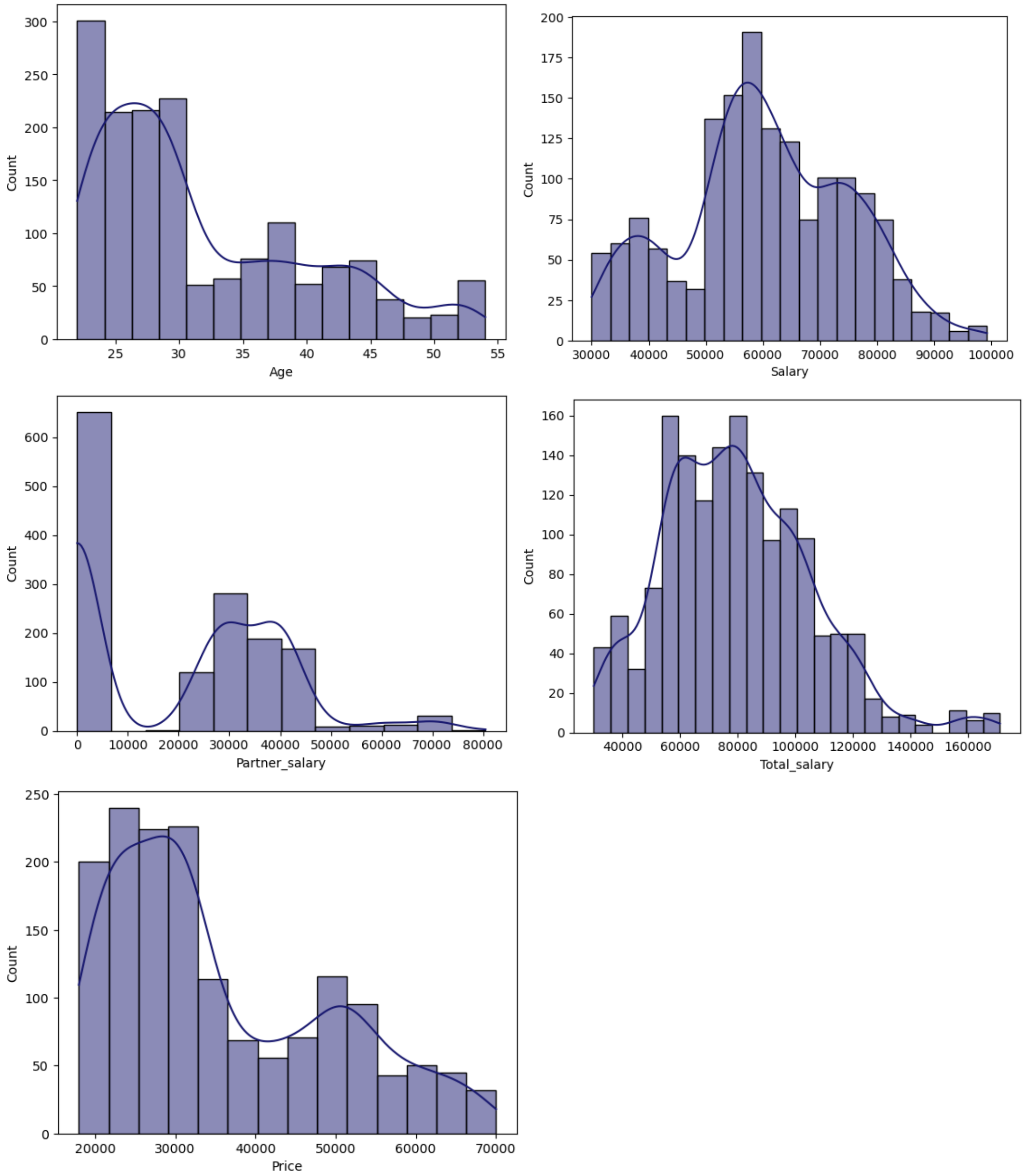
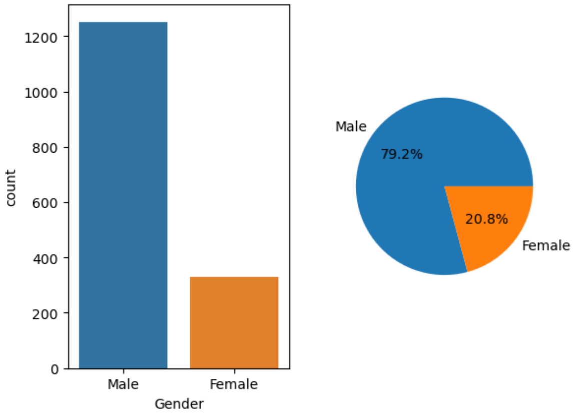
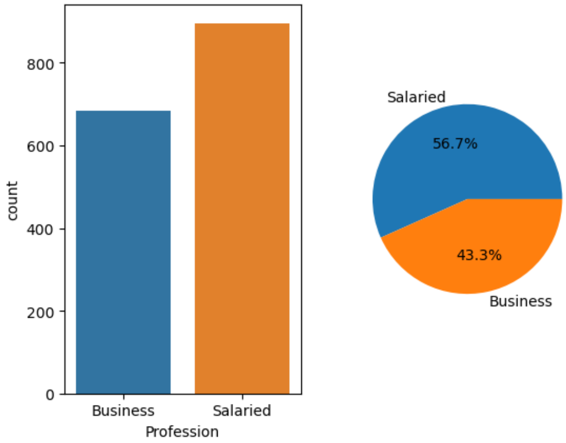


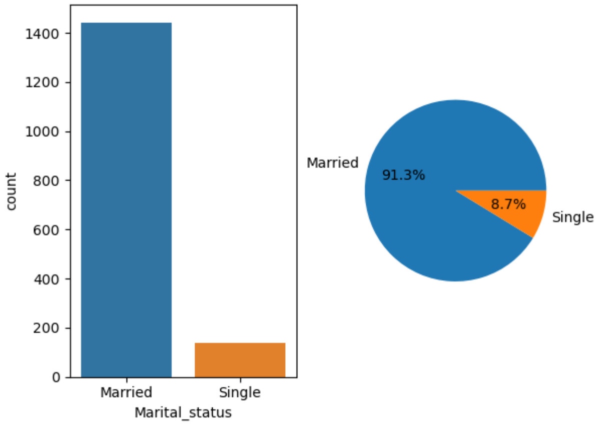
Fig. 1. Histograms of numerical fields in the dataset

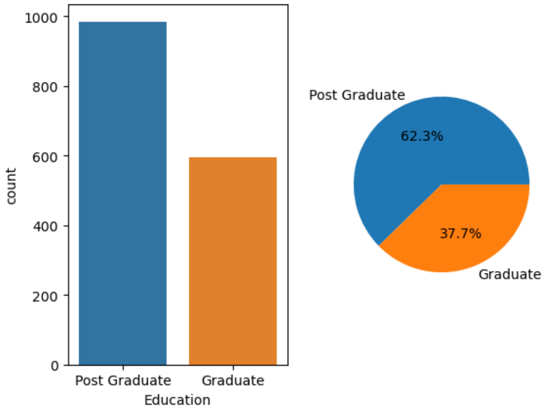
On performing a univariate analysis on the numerical columns, the general observations are:

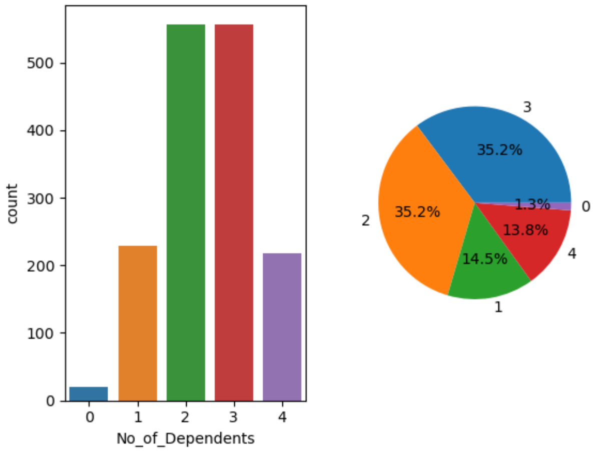
- Most of the customers are young, being 25-30 years old
- Most customers have salary ranging from 50k to 70k
- Considerable number of customers do not have working partners, and those who do have, have salary around 30k or 40k
- The total salary of the customers mostly range from 50k to 100k.
- Most of the cars sold have a price ranging from 20k to 35k.

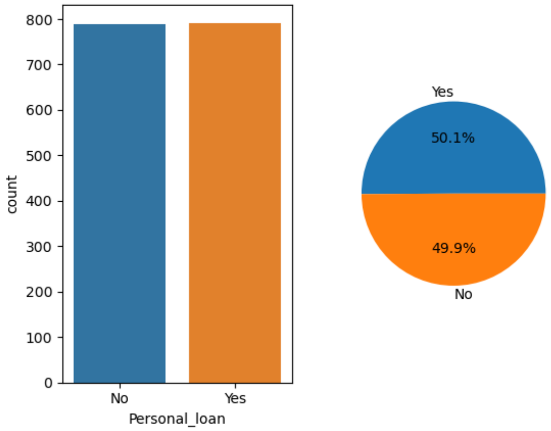

count
0
200
400
600
800
1000
1200
Male
Female
Gender
Male
79.2%
20.8%
Female


count
0
200
400
600
800
Business
Salaried
Profession
Salaried
56.7%
43.3%
Business


count
0
200
400
600
800
1000
1200
1400
Married
Single
Marital_status
Married
91.3%
8.7%
Single


count
0
200
400
600
800
1000
Post Graduate
Graduate
Education
Post Graduate
62.3%
37.7%
Graduate


count
0
100
200
300
400
500
0
1
2
3
4
No_of_Dependents
3
35.2%
2
35.2%
1.3%
0
13.8%
4
14.5%
1


count
0
100
200
300
400
500
600
700
800
No
Yes
Personal_loan
Yes
50.1%
49.9%
No

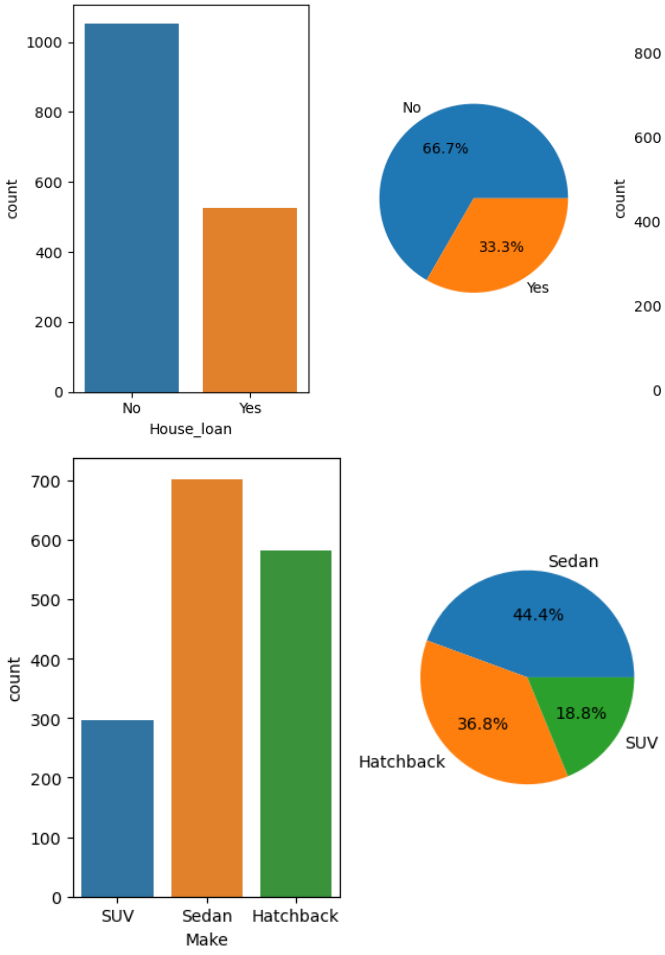


Fig.2. Count and pie distribution plots of categorical columns in the dataset

On performing a univariate analysis on the categorical columns, the general observations are:

- Most of the customers are male, salaried and married
- There are more post graduates in the customer dataset
- More than 70% of the customers have 2 or more dependents
- Most of the customers have not availed a home loan
- Most customers prefer sedans over hatchback and SUV

The following were the questions/insights that developed after the univariate analysis:

- Since most customers are young, could we develop a specific strategy to bring in middle aged and aged customers?
- Are there any strategies that can be developed to target low earning and high earning groups- maybe by taking a look at the preferred cars for different age groups?
- For those who have working partners, can upselling of high-end cars be possible? Can affordability be determined by price vs total salary comparison?
- Specifically, for the customers whose salary>100k, what is the preferred car model? Are they having dependents?
- In spite of having customers with high salary, only the low-priced cars are sold in high numbers. Why?
- Is there a link between salary and education level?

- What model do the customers with working partners and high total salary prefer?
- Since most customers do not have a home loan, their financial obligation is low, and hence may have a higher affordability. Can we upsell high end models to them?
- What is the average price of each car model? Which is the costliest?
- Going by size, the SUVs are the most spacious. Since most customers have more than 2 dependents, their preference should be SUVs, which is not the case here. Is upselling of car make possible here?

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.

Bivariate Analysis:

In this case, the goal is to identify the potential buyers by segmentation and then market the cars. For identification of potential buyers, affordability and necessity become the important parameters.
In general sense, affordability can be determined by:

- Total Salary and working partners: Buying capacity, generally is proportional to the earning capacity, and hence it is reasonable to have an initial assumption that groups with higher salaries can afford premium models and higher priced cars
- Lesser liabilities: The customers who have lesser liabilities, like loans, have better affordability

Necessity can be assessed by:

- Marital status- Married customers might need bigger sized cars than single customers
- No of dependents- Higher the number of dependents, greater is the need for a bigger sized model

In addition to this, a few other comparisons might help in the analysis:

- Education level, salary and price of car owned
- Gender, age and price comparison

If we assume that the profit of the company is proportional to the price of the car, then price becomes the target variable here.

Having made the assumptions and initial insights, the analysis is as follows:

a) Price vs car make comparison

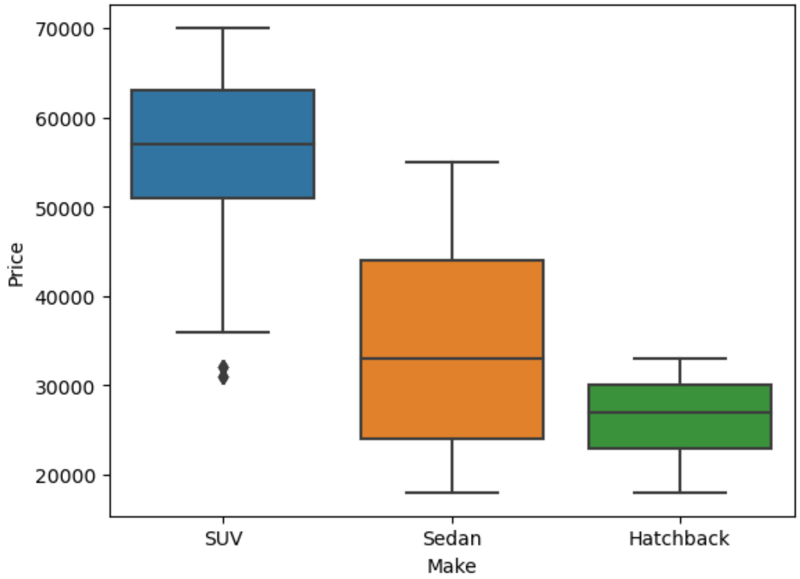


Fig.3. Box plot showing the distribution of price for different makes

b) **Salary vs Education**

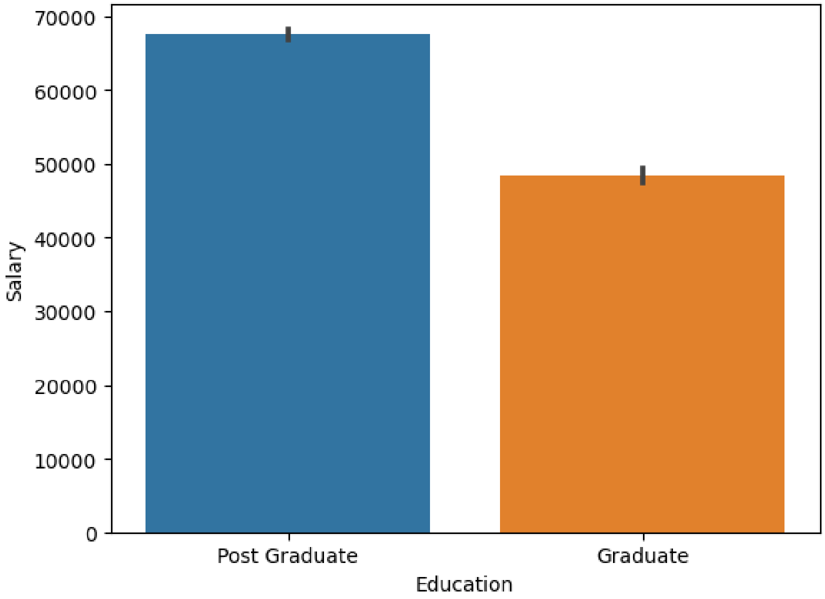


Fig.4. Salary vs Education level bar plot

c) **Price vs Education**

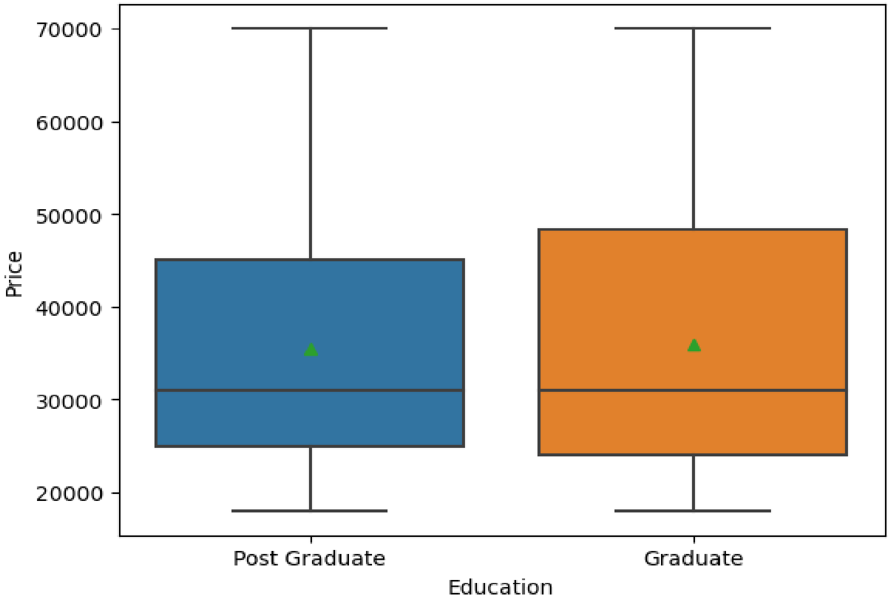


Fig.5. Box plot showing the distribution of price across the two education levels

d) **Total_Salary vs Price and age vs price**

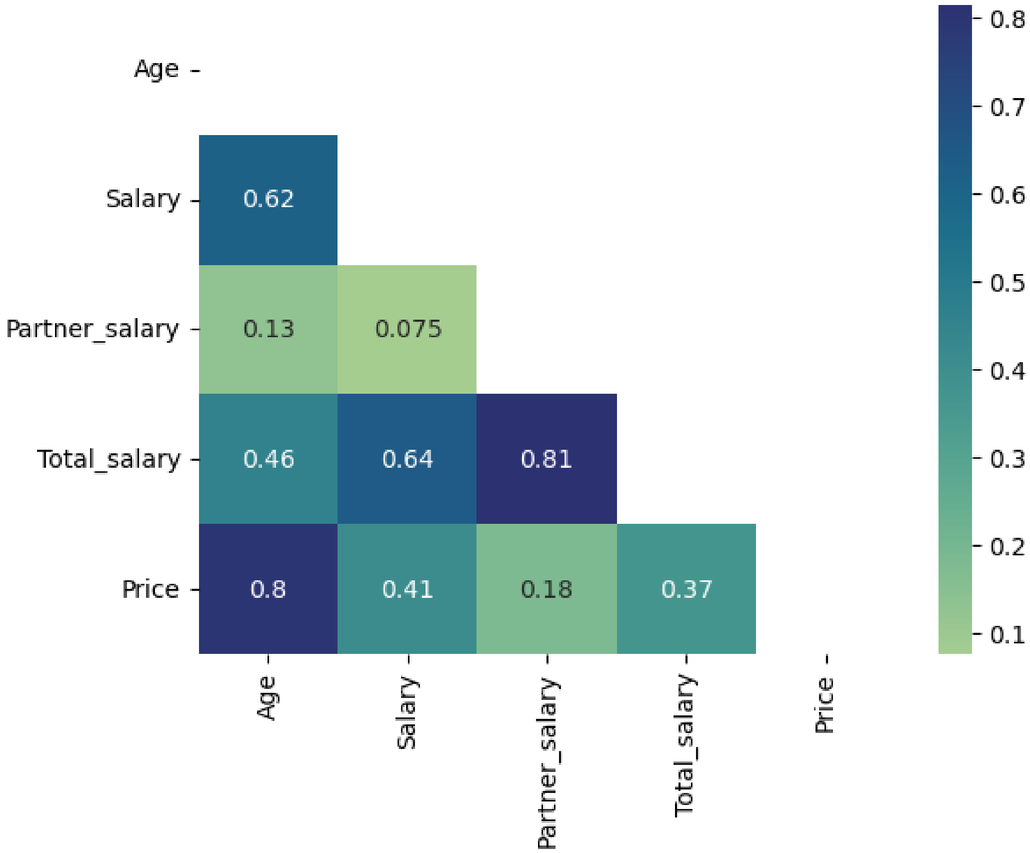


Fig.6. Heatmap of numeric fields

e) **Gender vs price:**

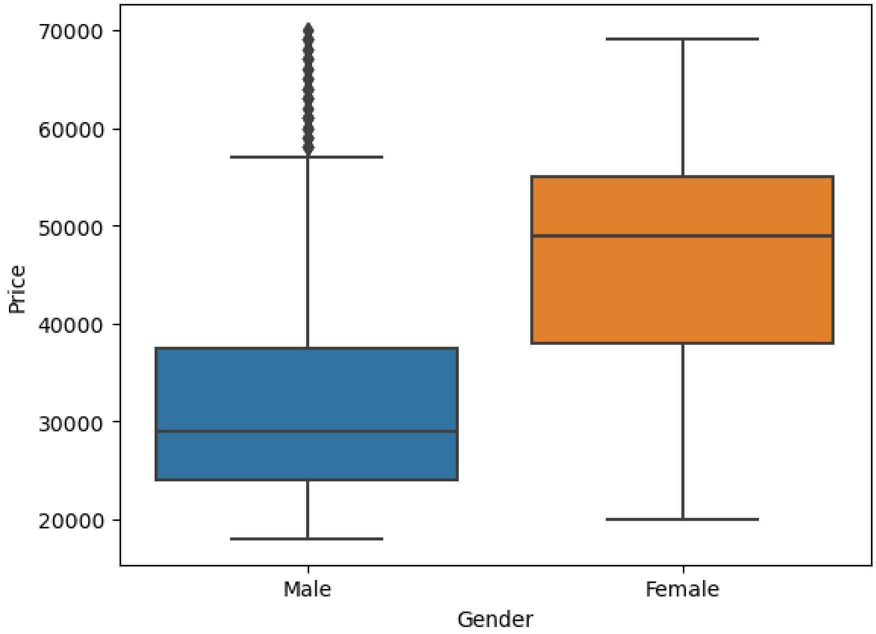


Fig.7. Boxplot price distribution across genders

f) **Marital_status vs Price**

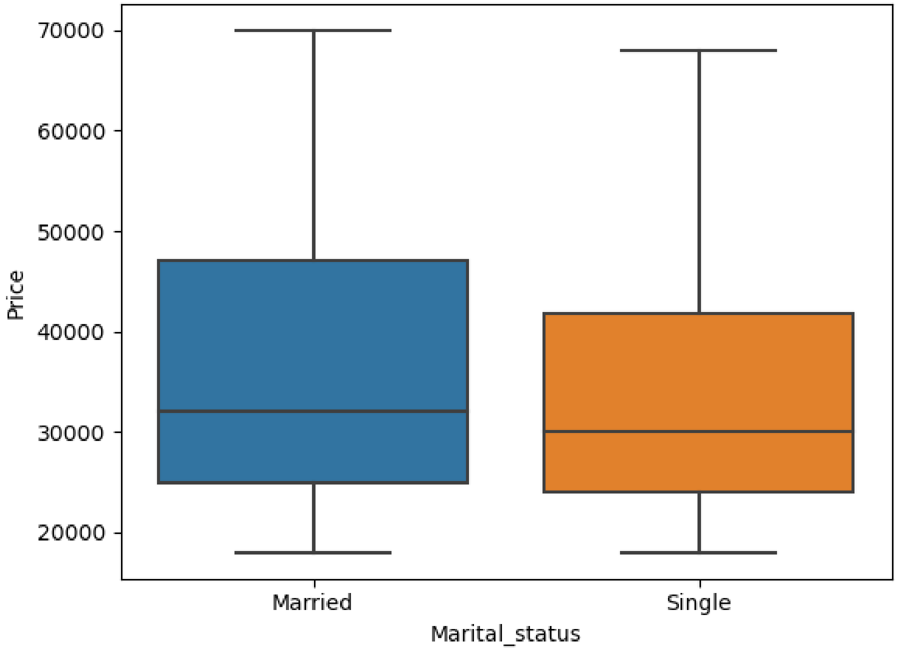


Fig.8. Boxplot price distribution across marital status

g) **No_of_dependents vs Price**

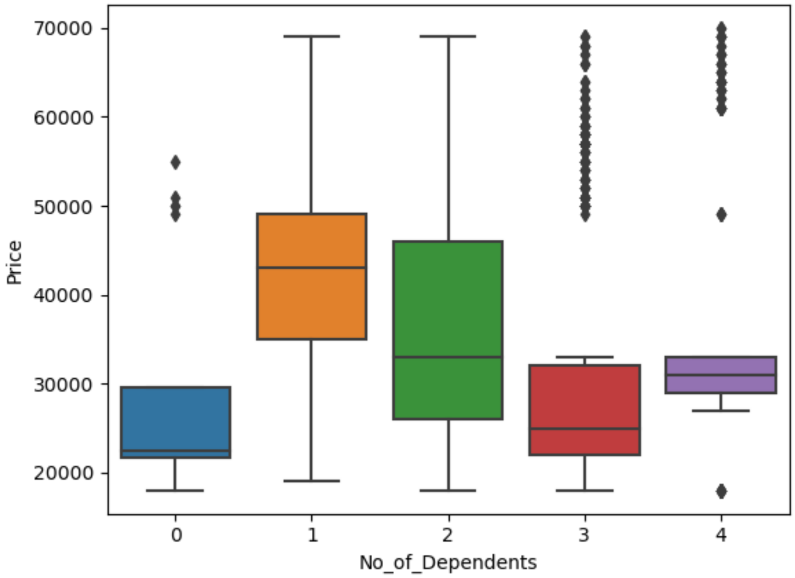


Fig.9. Boxplot price distribution across no of dependents

The observations from the above analysis is as follows:

- The car makes arranged in the decreasing order of price is as follows- SUV, Sedan, Hatchback
- The assumption that the Salary is higher for postgraduates as compared to the graduates is right
- Despite having higher salaries, there is no difference in the price ranges of cars bought by graduates and post-graduates
- There is not much of a correlation between the Total_salary and Price, meaning it's not just about affordability, but necessity and preference as well
- There is a decent correlation between the age and salary fields, and a high correlation between the age and price fields
- The females tend to opt for higher priced cars than the males
- Married customers prefer slightly higher priced cars than single customers
- Customers with one or two dependents buy higher priced cars than the rest

Going back to our initial assumptions and insights, the following are the developments after the bivariate analysis:

<u>Affordability</u>:

- Total_salary has a decent positive correlation with price
- There is no proof to support that the liability, i.e. loans, influences the affordability from the given data

<u>Necessity:</u>

- Marital status has a little influence over the price, but not by a large extent
- No of dependents does seem to make customers buy higher priced cars, but then the affordability also comes into the picture

<u>Additional information gained:</u>

- Females have markedly higher price tolerance than males
- Age also influences the price

Multivariate Analysis:

Going forward, the make preference and price of the groups showing tolerance to high price, i.e.

- Females
- Middle aged and aged customers
- Married
- Having 1 -3 dependents

can be analysed

a) Age vs Price vs Make

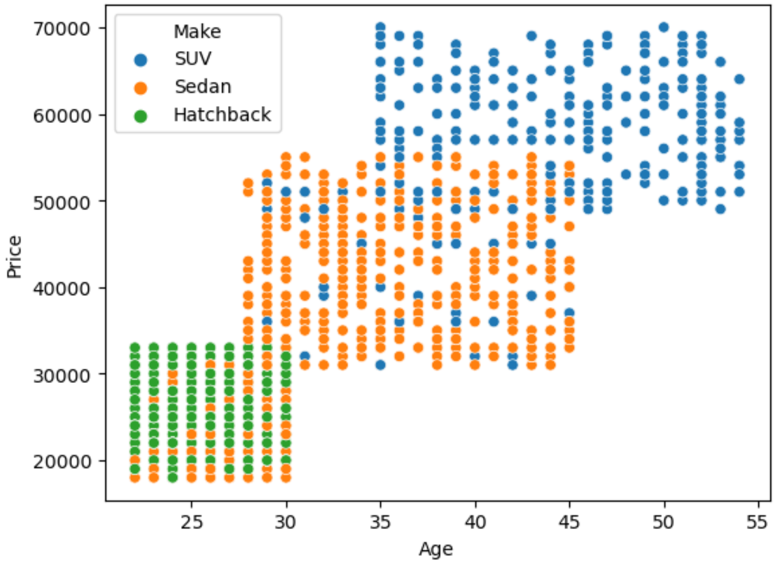


Fig.10. Scatterplot price vs age for different makes

b) Gender vs price vs Make

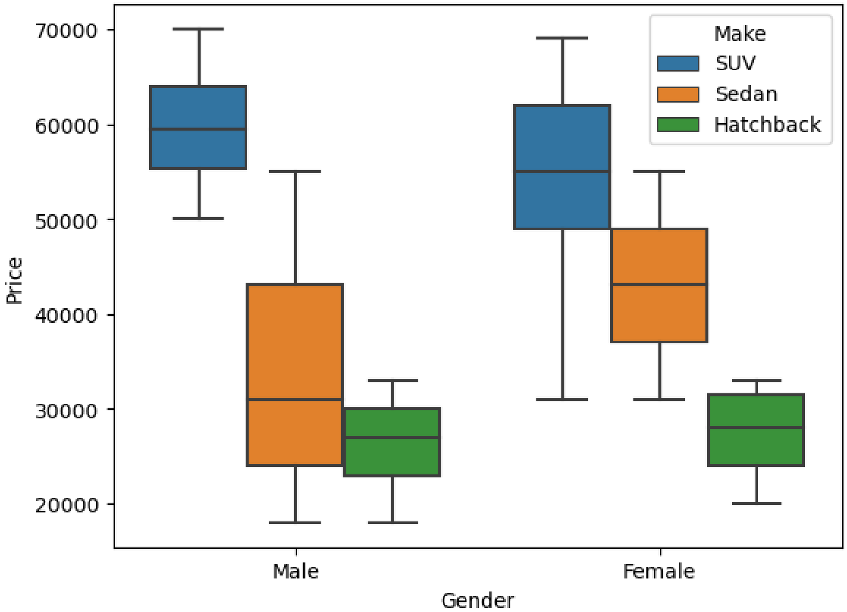


Fig.11. Boxplot price distributions different makes across genders

c) Marital status vs price vs Make

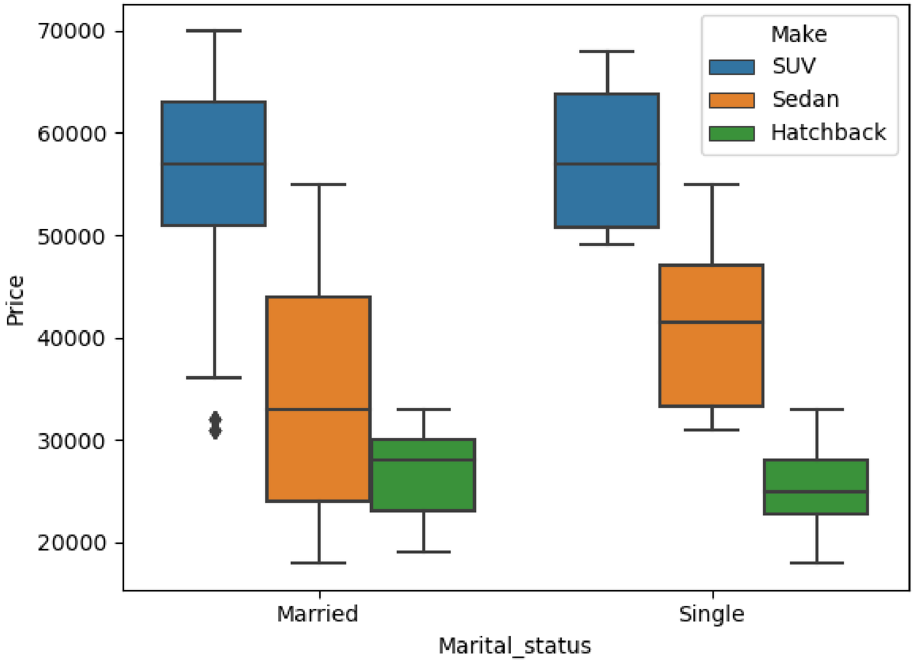


Fig.11. Boxplot price distributions different makes across marital status

d) No of dependents vs price vs make

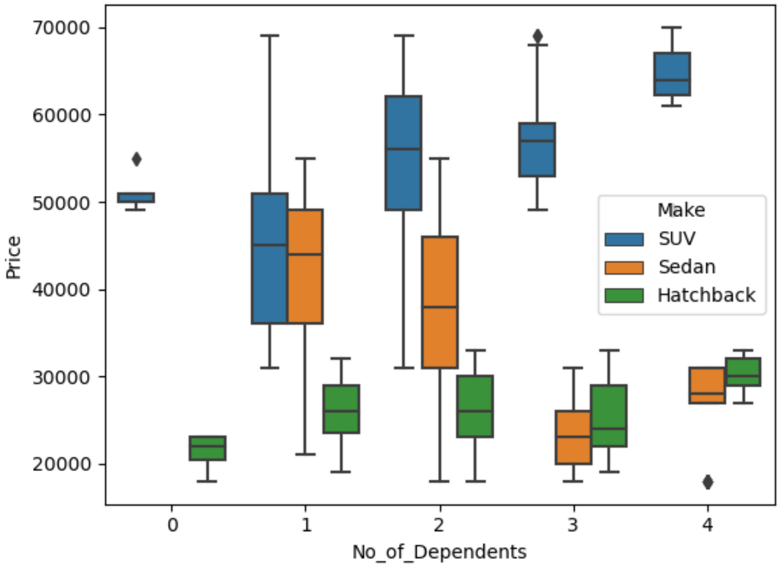


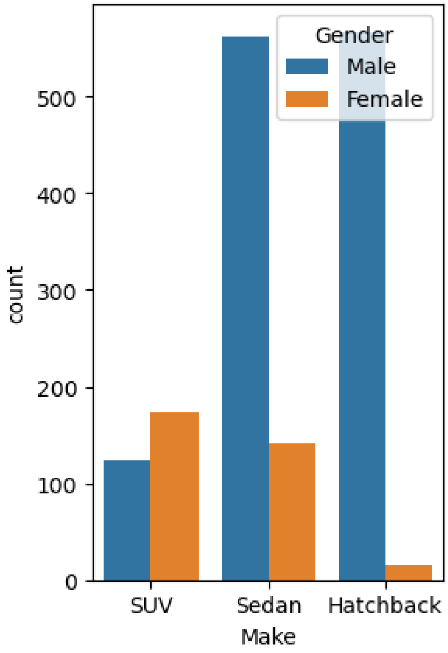
Fig.12. Boxplot price distributions different makes across no of dependents

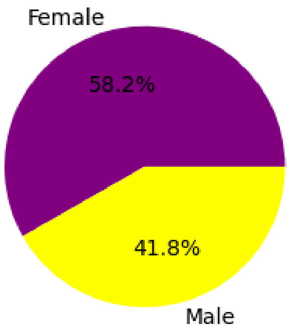
After a thorough analysis, having price as the target variable, the predictors were identified to be gender, age, salary, marital status and no of dependents. The following final insights are the result of the analysis:

- The female customers have shown a significantly higher tolerance for price and tend to prefer high-end makes of cars, particularly high-priced sedans. Hence, they can be our first target group, both for generating new customers, as well as upselling within the existing customer base
- Though the males have shown lesser tolerance to price, they have still opted for higher priced SUVs as compared to the females. Hence, some special offers can be developed for the SUVs and can be pitched to the male customers of the existing database.
- The dataset predominantly has customers belonging to the age groups 25-35. However, we have seen that the higher the age, the higher the salary. Consequently, the affordability is high. Hence, for the creation of new customer base, customers with age 45+ can be considered, as they clearly opt for SUVs. For the existing customer base, identifying the higher aged customers with low-end cars, and upselling pricier version can be done
- There are customers with 3 or more dependents using hatchbacks. Hatchbacks are ideal for single customers or small families. Hence, there is a scope of upselling a pricier model of car, even for a slight increase in price for this category.

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.

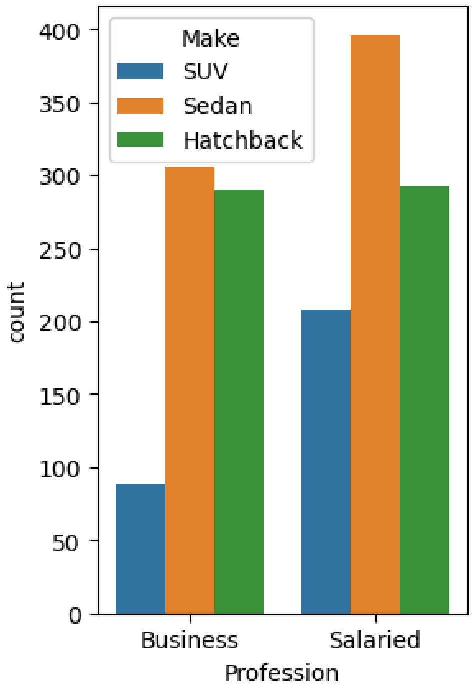
E1) Steve Roger says "Men prefer SUV by a large margin, compared to the women"

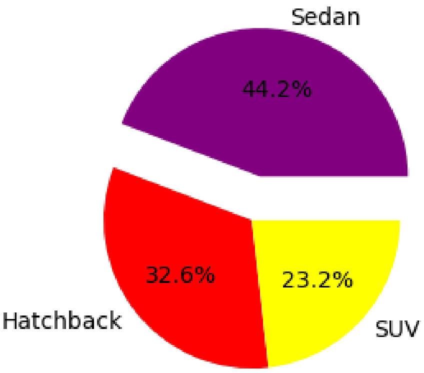




It is evident that the observation made by Steve Roger is incorrect. Amongst the makes, women prefer SUV more than men. From the pie chart above, it can be seen that 58.2% of women prefer SUVs, whereas only 41.8% of the men prefer the same.

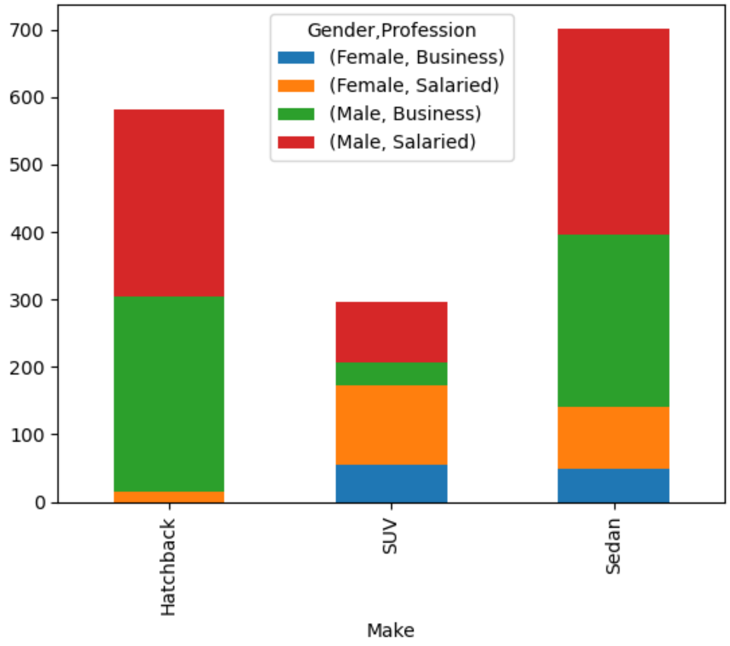
E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.





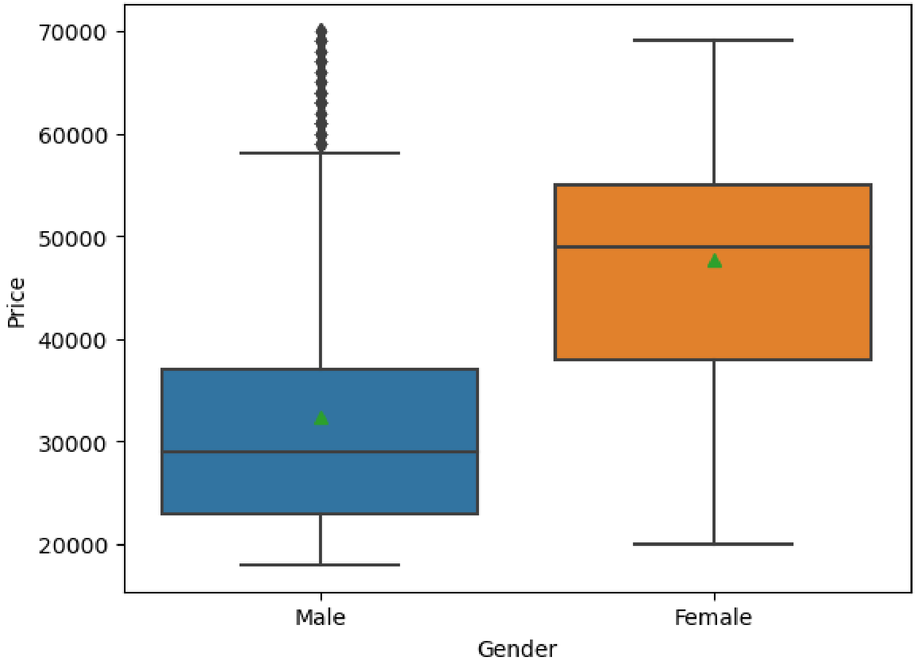
The observation of Ned Stark that a salaried person is most likely to buy a Sedan is right as evident from the above graphs

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.

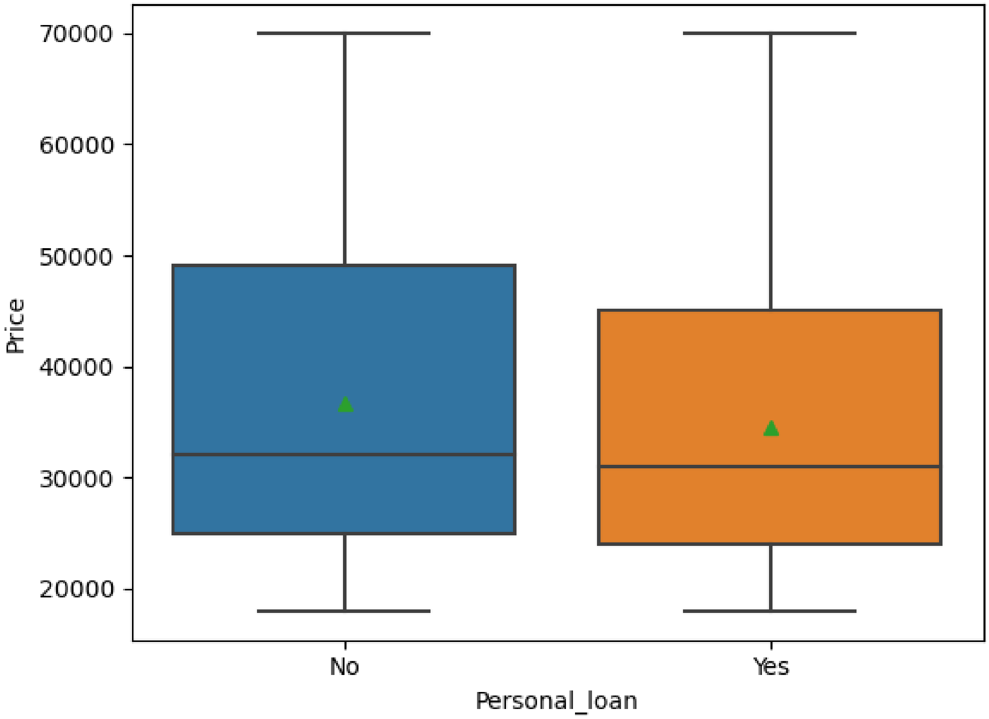


From the above, it can be inferred that the Salaried male prefers Sedan over SUV. Hence, Sheldon Cooper's observation is not true.

F. From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a Business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions.
Give justification along with presenting metrics/charts used for arriving at the conclusions.
F1) Gender
F2) Personal_loan

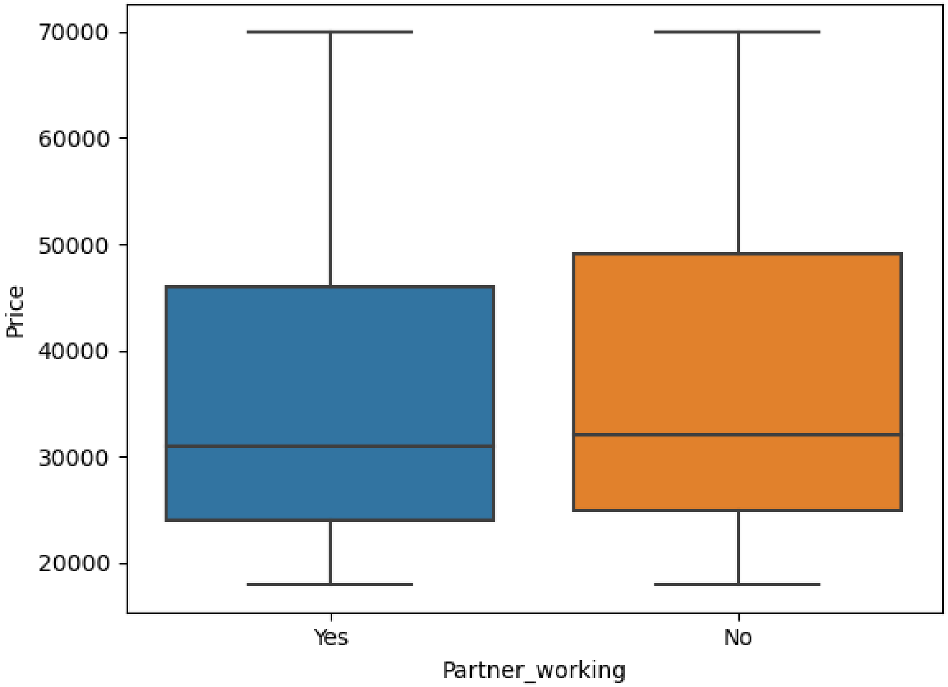


F1) From the above, it can be inferred that the female customers are more likely to spend higher amounts on the purchase of a car, as compared to the male customers. The median value of price for females is 48000, whereas for males, it is 29000. However, when looking at the customer data, there are more males than females. Hence, a special discount or offer for female customers can bring in more business to the store.



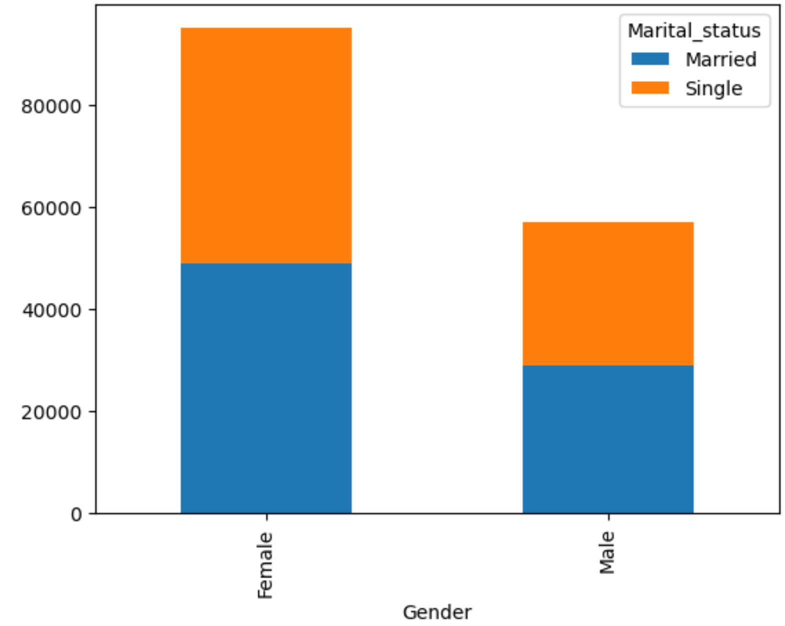
F2) The personal loan availment status doesn't seem to make much of a difference in the price of the car

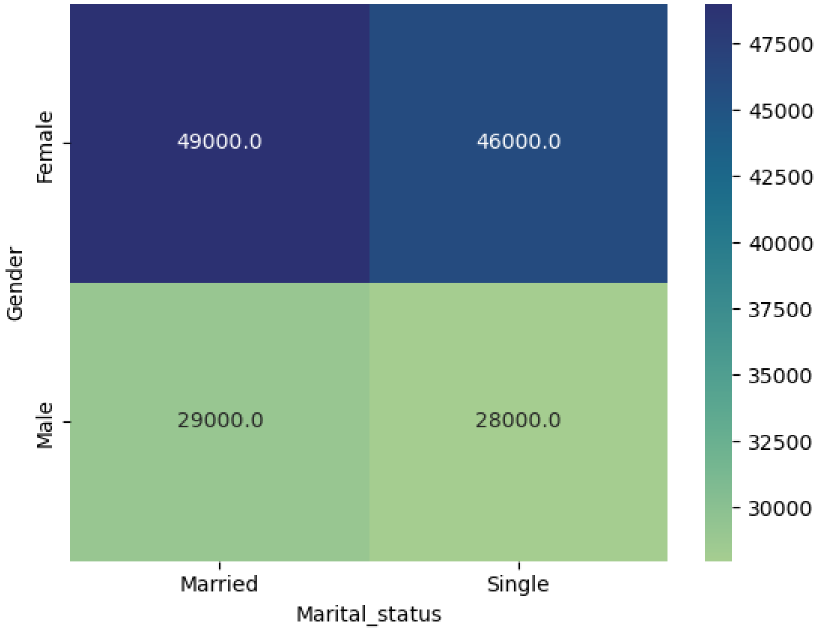
G. From the current data set comment if having a working partner leads to the purchase of a higher-priced car.

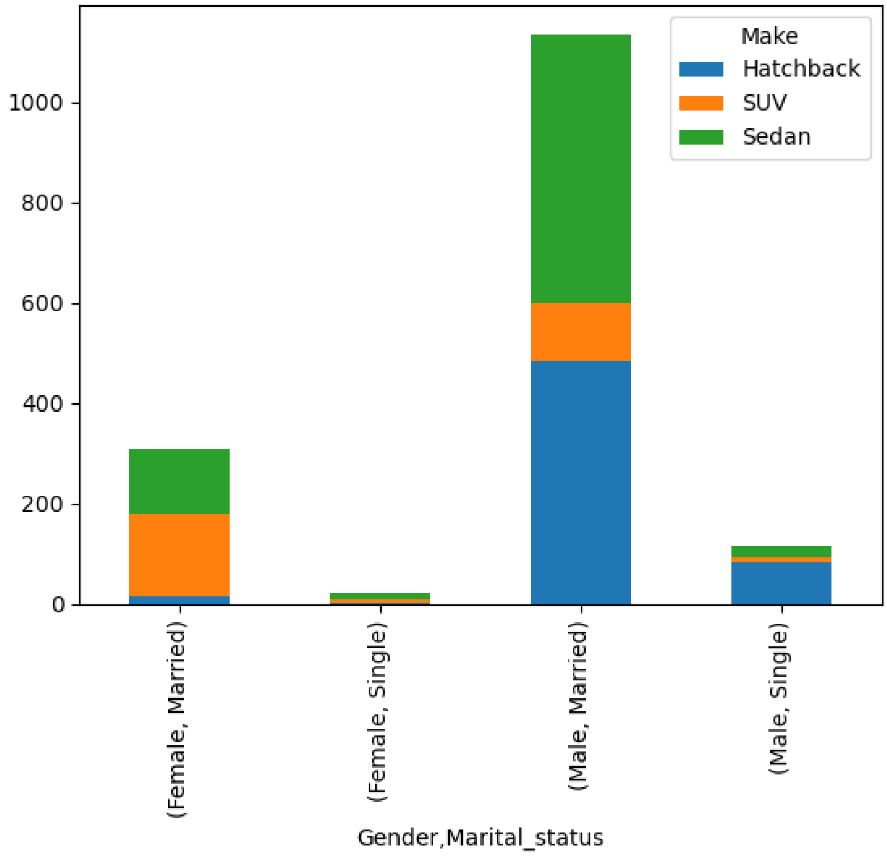


The working status of partner does not seem to influence the price of the car bought as seen from the above barplot

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.







On exploring the data given the following are a few observations and actionable insights, consequent to a deeper analysis of the gender and marital status fields:

- The price field has extreme values and is right skewed. Hence, the median of the price across categories was chosen for the sake of comparison. On examining the above graphs, where a crosstab comparison was made for the different combinations of Gender and marital status, we find that the median price for married females are the highest. This is then followed by single females, married males and single males in that order.
- Assuming profit of the company is proportional to the price, the primary focus is on the married female category. We have already seen that the SUV is the priciest of the three models. From the above bar graph, it is evident that the married females are more interested in buying the SUVs, probably owing to the spacious nature. This category should be the primary target group.
- In general, as per the data, females are more inclined to buy a higher priced car. Hence, the company must focus on special offers and promotions targeting them.
- Based on the marital status, we find that the median price for married customers is slightly higher than the single customers. On exploring the make for these and constructing a crosstab, we find that a significant portion of the married customers, with 3 dependents are using the hatchback model. Hatchbacks are ideal for single persons or a small family and are the least pricey. Hence, efforts could be made to upsell a sedan for these customers.
- On exploring the data, we find that the customer base has only 8.7% of single customers. Efforts could be made to draw in more customers from this category. This is to achieve profit not by price, but by volume of sales.

Case 2: GODIGT BANK CREDIT CARD DATA

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

Problem Statement:

GODIGT Bank 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.)

Solution:

Introduction:

The purpose of the analysis is to improve credit card usage amongst the customers. This can be done in two ways:

- From the existing customer base, identify those with potential for higher usage and then develop strategies that incentivize the target groups to use the card more.
- Identify patterns and groups from the existing base that have higher usage, and target these groups to acquire new customers.

Understanding the dataset:

In the given dataset, there are 27 fields, of which 2 variables are id fields, 1 date field, 2 numerical fields and the rest are categorical in nature, of which 8 are object fields. As per the data dictionary, the fields can be classified into groups:

- ID and sourcing related data- card_no, user_id, card_bin_no, card_source_date
- Customer specific information- high_networth, annual_income at source, Occupation at source
- Card specific information- Issuer, card_type, hotlist_flag, cc limit, Transactor/revolver
- Card Usage related information- cc_active 30,60 & 90, T+1,2,3,6 & 12 months activity, average_spends
- Customer-bank information- active 30,60 & 90, widget_products,engagement_products, bank_vintage
- Miscellaneous information- Other-bank cc holding

Feature Engineering:

Having defined the use case of the analysis and having looked at the available data, it is essential to understand the problem statement and identify target and predictor variables. Here, the usage of the customers is the target. Usage can be interpreted in terms of two factors:

- ➢ Frequency, i.e., the number of times the card is being used

- Amount, i.e., the extent to which the customer spends using the card

From the given variables in the Usage fields, all of cc active fields and T+ activity fields indicate the frequency of usage. However, it might be difficult to perform analysis when they are in the raw form. Hence, we can engineer a field, called ‘card activity’ which is a weighted sum of these variables. The weights were assigned in the increasing order of time. That is, the more recent usage got more weight.

This is represented as,

Card activity= (1*cc_active 30)+(2*cc_active 60)+(3*cc_active90)+(4*T+1_month active)+…(8*T+12_month active)

This variable represents the quantitative aspect of usage. High card activity means frequent and longer duration of usage and vice versa. Thus the bucketization can be achieved by a new column- card usage

To assess the qualitative aspect of usage, the average spends field can be used. However, owing to the width of the range of cc limits in the dataset, it might be better to interpret the amount of usage in terms of a ratio, spend ratio.

Spend Ratio= average spends/cc limit

This gives a scaled metric across the dataset to compare the utilization of limits.

Also, another field that might need to be scaled is the cc limit. Generally, limits are assigned based on the annual income of the customer. Unless the customer specifically asks for a lower limit, the maximum limit is usually a percentage of annual income. Having verified this from the dataset, another ratio can be created, i.e., the limit ratio

Limit ratio= CC limit/Annual income

Similar to the card activity field, the account activity fields can be combined to form a weighted field,

Ac activity=(1*ac_active30)+ (2*ac_active60)+(3*ac_active90)

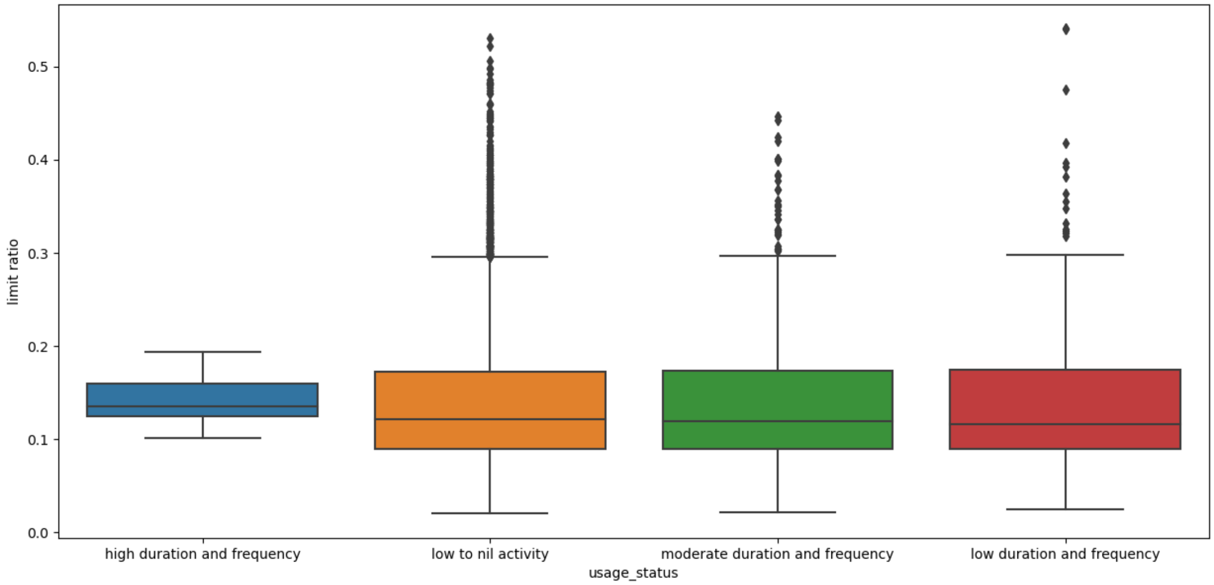
Thus, after engineering the features, the dataset contains the following variables:
Target variables- Spend Ratio, Card activity, Card usage
Predictor Variables- Limit Ratio, cc limit, annual income, card type, issuer, account activity, networth, and occupation, other-bank cc holding, transactor/revolver, widget and engagement products, bank vintage, average spends

Initial Insights:
- Are customers not utilizing the limit here because of the other bank cc?
- Customers who are not using the savings account likely to not use the cc as well. Is this assumption right? Does this mean customer inactivity and attrition at the bank level, and not just the product level?
- Are customers with high widget products more or less likely to have better usage?- Are they opting for other payment options e.g. wallet/netbanking rather than using cc?
- Are the high engagement customers vintage customers?- What is their usage?

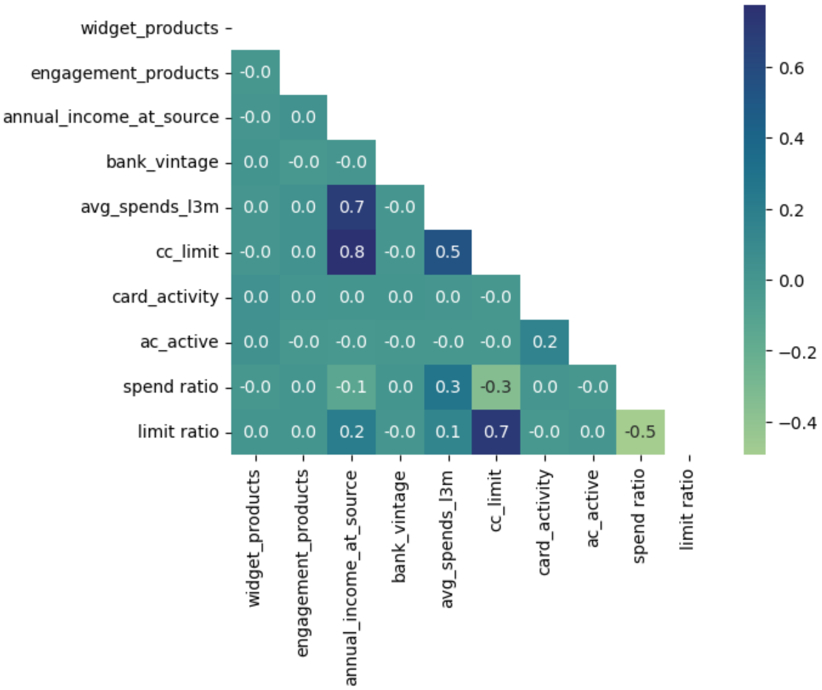
- How does income affect spending?
- Are cc limits being determined by annual income? cc vs income
- Which category spends the most? T or R?
- What is the distribution of T and R? If we assume that T type have better cash flow, they may be using the cc only during certain periods, like when there are cc related offers in shopping etc. Is this the case?
- which is the most used card type for t and r?
- Are HNIs being given the right limit and card type?- HNI vs Limit vs card type
- Which card type holders have maximum utilization? Can we push that particular type of card to everyone else? Is there a particular card type with very minimal spending? If so, why?
- Which occupation type people use the most and least?
- How does the usage vary with bank vintage?

Post-Analysis observations:

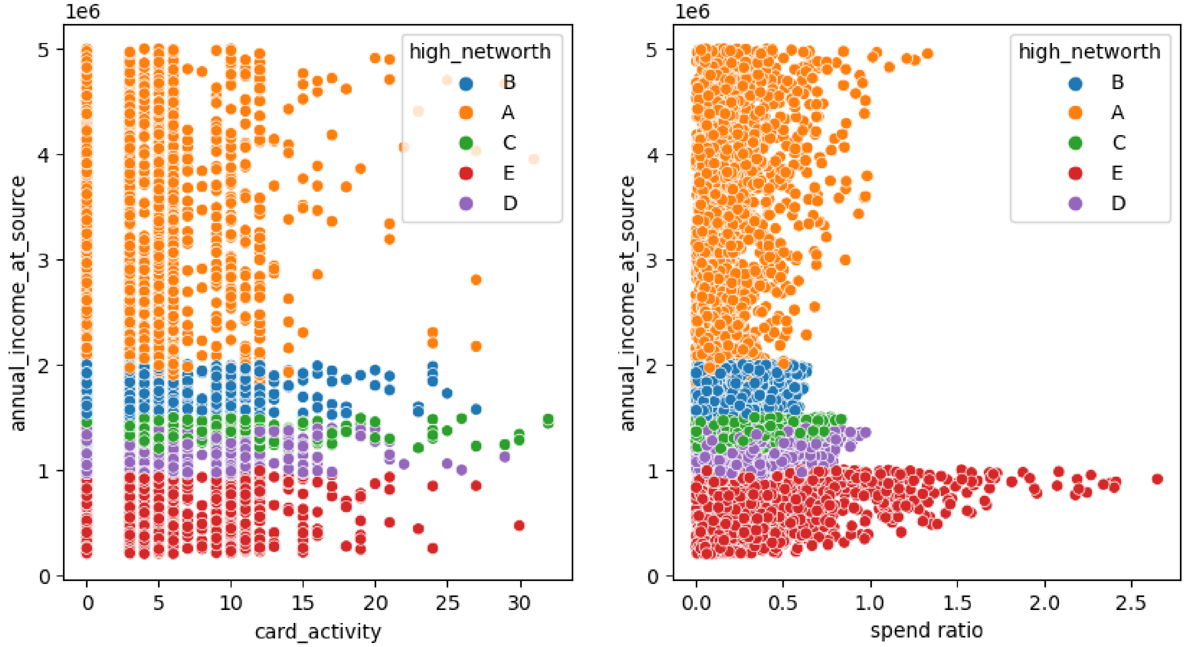
1) Customers with higher limit ratio show higher card activity



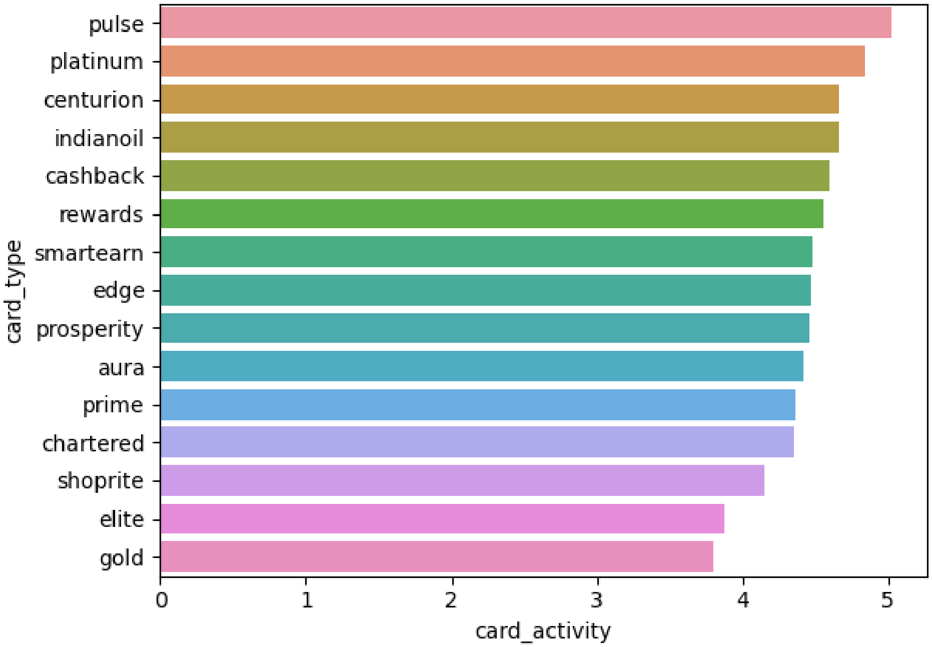
2) Spend ratio and limit ratio have a high correlation

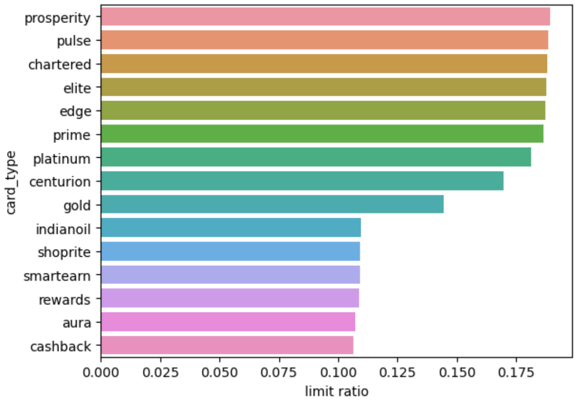


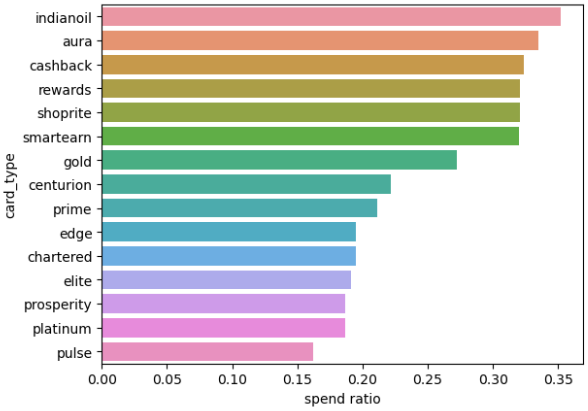
3) E category customers show higher activity and spending than the rest



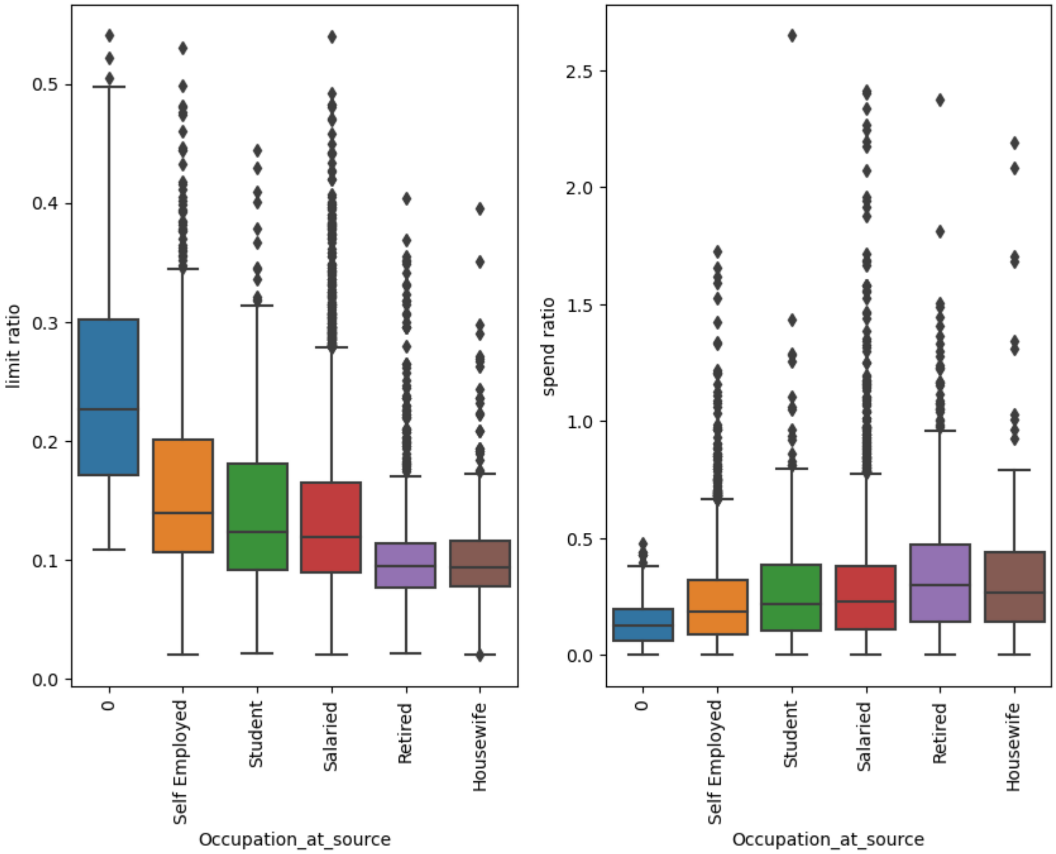
4) Cards like indianoil, aura, shoprite, rewards,cashback etc have high activity and spend ratio, despite having lower limit ratio







5) Retired customers and housewifes are likely to spend more despite having lower limits



Target variables for analysis:

- Spend ratio- spend and limit
- Card Activity- cc active 30,60,90, t+ 1,2,3,6,12 activity
- Card Usage

Important predictor variables identified after analysis:

- CC limit
- Annual Income at source
- High networth
- Card Type
- Occupation

Final insights after analysis:

- The general observation is that the limit ratio tends to increase the usage. Hence, if it is risk-viable, increasing the limits of card can bring about increase in usage for the existing groups. This can be done for groups based on networth, and card type, that show high utilization already
- When sourcing new customers, more emphasis can be laid upon high usage card types like indianoil, rewards, shoprite, aura, cashback etc.
- Retired and housewifes show higher utilization, but lower activity. Hence, special incentives can be giver for more instances of usage
- The groups with lower networth have the highest need for utilization. Hence, targeting those groups for new sourcing, and increasing the limits, if feasible , for existing customers can lead to better usage.