SMDM PROJECT

REPORT

DSBA

Contents

Problems

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)

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.

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

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.

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

   E1) Steve Roger says “Men prefer SUV by a large margin, compared to the women”
   E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.
   E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale.

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.

   F1) Gender
   F2) Personal_loan

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

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

I. Framing An Analytics Problem. Analyse the dataset and list down the top 5 important variables, along with the business justifications.

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)

As a database administrator, some important technical information about a dataset that should be considered includes:

1) Size of the dataset: We should consider the number of records (rows) and the number of variables (columns) in the dataset. This should be the basic thing to check for given any datasets.
2) Data types and variable formats: We would want to know the data types of the variables in the dataset (e.g., integer, float, string, date) and the format in which they are stored (e.g., YYYY-MM-DD for date variables). This information will help us to analyse the data and come up with relevant problem statement and accurate solutions for the same.
3) Missing values: We would want to know if the dataset contains any missing values and if so how they are represented in the data (egs :NaN, 0,"?"..). This information will help us in determining how to handle missing values when loading the data and doing analysis on it.
4) Categorical variables: We would want to know if the dataset contains any categorical variables and how they are encoded (e.g., as strings, integers, or one-hot vectors). This information will help us in determining how to store and query categorical variables in a database.
5) Cardinality of variables: We would want to know the number of unique values and the distribution of values for each variable in the dataset
6) Indexes: We would want to know if the dataset contains any columns that should be indexed for efficient querying.

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.

We will explore the austo automobile dataset and perform the exploratory data analysis on the dataset. The major EDA done are:

- Removing duplicates
- Missing value treatment
- Outlier Treatment
- Univariate Analysis
- Bivariate Analysis

Basic Data Exploration:

In this step, we will perform the below operations to check what the data set comprises of. We will check the below things:

- head of the dataset
- shape of the dataset
- info of the dataset
- summary of the dataset

Head of the dataset

```
In [4]: df.head()
```

Out[4]:

	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	Price	Make
0	53	Male	Business	Married	Post Graduate	4	No	No	Yes	99300	70700.0	170000	61000	SUV
1	53	Femal	Salaried	Married	Post Graduate	4	Yes	No	Yes	95500	70300.0	165800	61000	SUV
2	53	Female	Salaried	Married	Post Graduate	3	No	No	Yes	97300	60700.0	158000	57000	SUV
3	53	Female	Salaried	Married	Graduate	2	Yes	No	Yes	72500	70300.0	142800	61000	SUV
4	53	Male	Salaried	Married	Post Graduate	3	No	No	Yes	79700	60200.0	139900	57000	SUV

Here we can view the head of the dataset.

Shape of the dataset

```
In [31]: df.shape
Out[31]: (1581, 14)
```

Here shape if the data tells number of rows and columns: We can understand that the dataset has 1581 rows/records and 14 columns.

Info of the dataset

Here from the info provided below we can understand each of the columns /variables in the dataset, their datatypes and their non-null count.

From the below table we can understand that the Partner_Salary and Gender columns contains null values as there are only 1475 and 1581 non null count whereas, the other variables have a total non-null count of 1528. Hence we can understand that there are missing values in those columns.

```
In [8]: df.info()
```

```
<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
```

info() is used to check the Information about the data and the datatypes of each respective attributes

Summary of the dataset

```
In [42]: df.describe()
```

Out[42]:

	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
count	1581.000000	1581.000000	1581.000000	1475.000000	1581.000000	1581.000000
mean	31.922201	2.464263	60392.220114	20225.559322	79398.545225	35597.722960
std	8.425978	0.928532	14674.825044	19573.149277	24849.147996	13633.636545
min	22.000000	0.500000	30000.000000	0.000000	30000.000000	18000.000000
25%	25.000000	2.000000	51900.000000	0.000000	60500.000000	25000.000000
50%	29.000000	2.000000	59500.000000	25600.000000	78000.000000	31000.000000
75%	38.000000	3.000000	71800.000000	38300.000000	95900.000000	47000.000000
max	54.000000	4.000000	99300.000000	80500.000000	149000.000000	70000.000000

Removing duplicates

Here we understand there are no duplicate records in the dataset.

```
In [9]: df.duplicated().sum()

Out[9]: 0
```

Checking for missing values

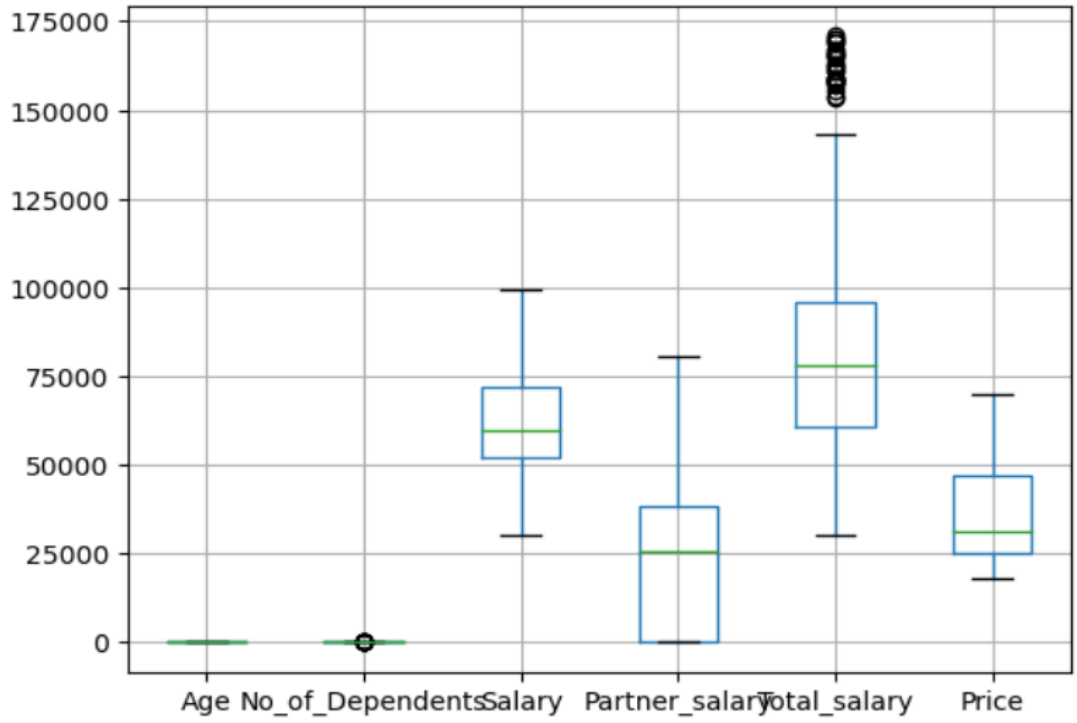
```
: df.isnull().sum()[df.isnull().sum()>0]

: Gender            53
  Partner_salary    106
  dtype: int64
```

Here we understand there are 53 missing values in Gender and 106 missing values in Partner_salary.

We impute the Gender column with the mode values as it is Categorical variable and Partner salary with Total_Salary-Salary for all the null values

Outlier Treatment



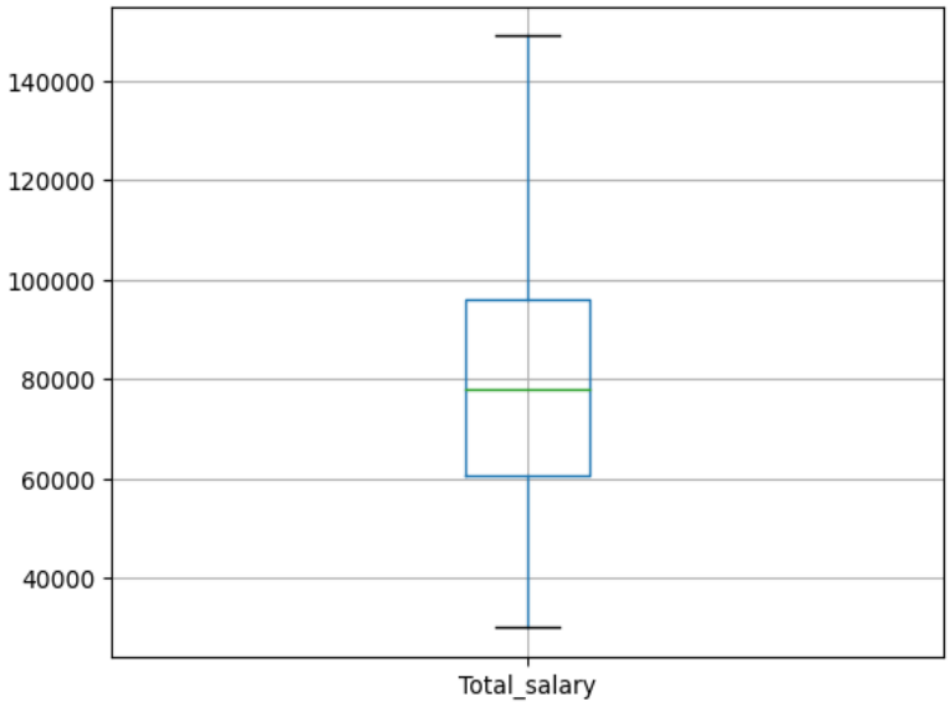
Here we can see that we have mainly 2 outliers

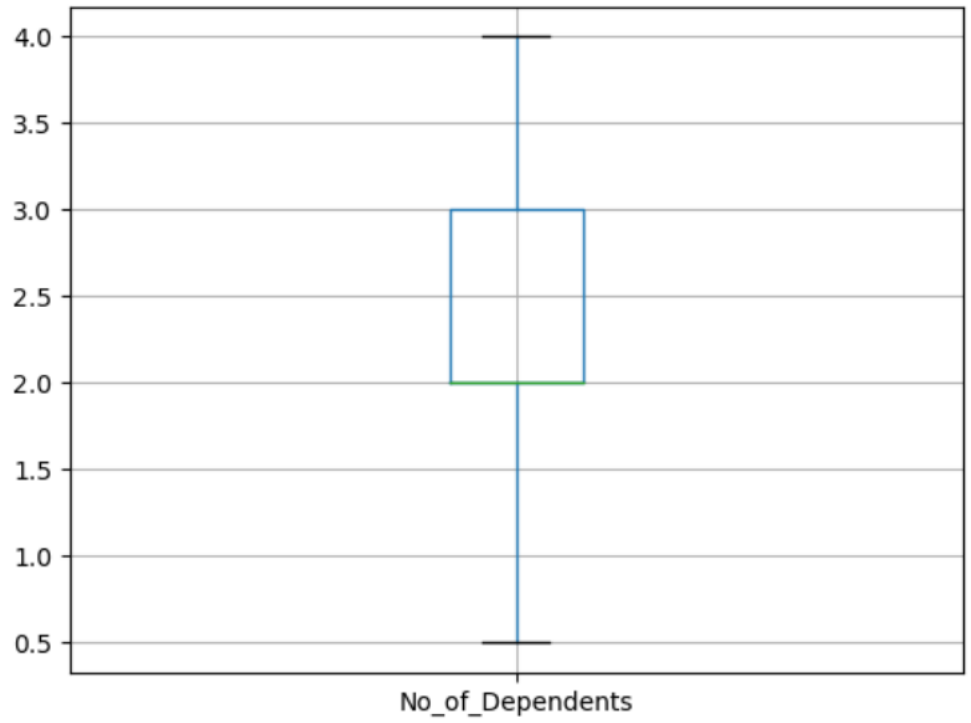
- No_of_Dependents
- Total_Salary

Here we are treating outliers like:

- Drop the outlier value
- Replace the outlier value using the IQR

After Outlier Treatment



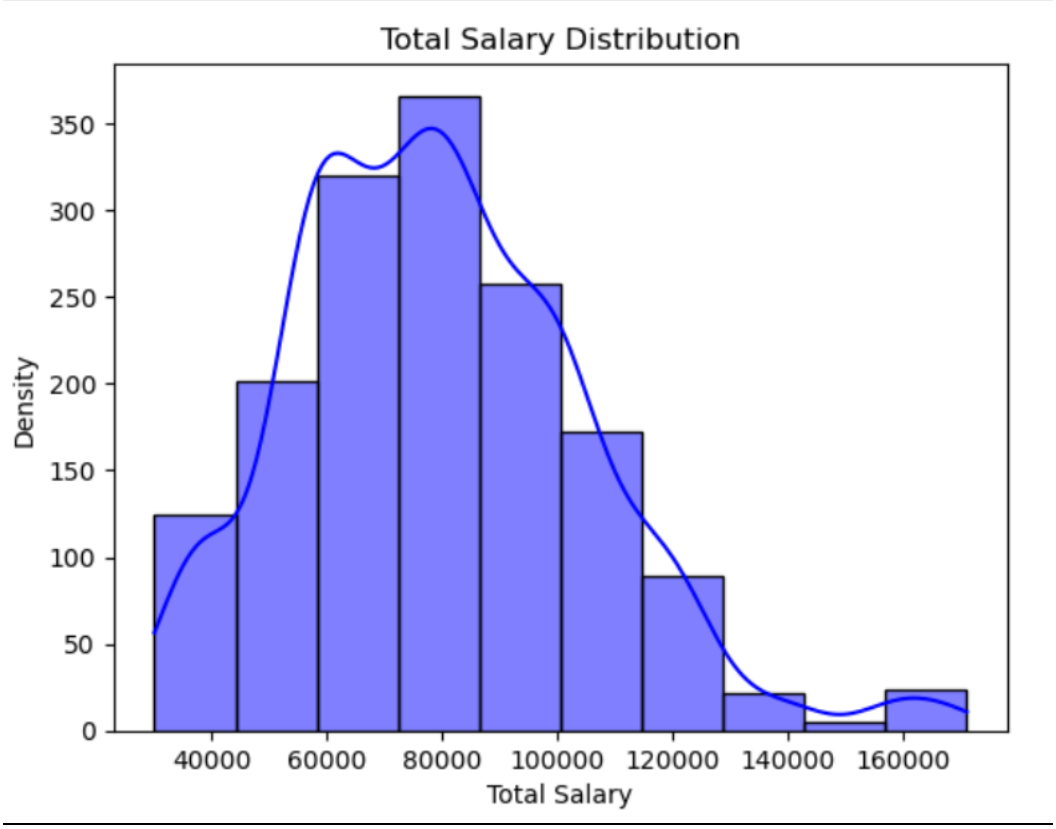


If we look at the box plots above, post treating the outlier there are no outliers in all these columns.

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

I. Total Salary Distribution



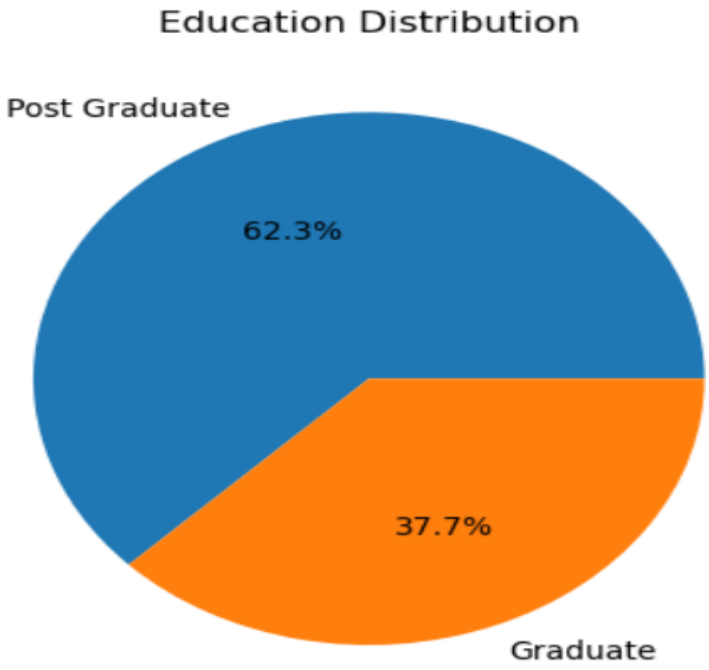
Analysis

- So from this graph we can see that the maximum people fall into the category of having total salary from 60,000 to 90,0000.
- Hence we can say that mostly the cars are bought by middle class family having salary range within 60,000 to 90,000. Hence they should be target market in order to improve the car sales.

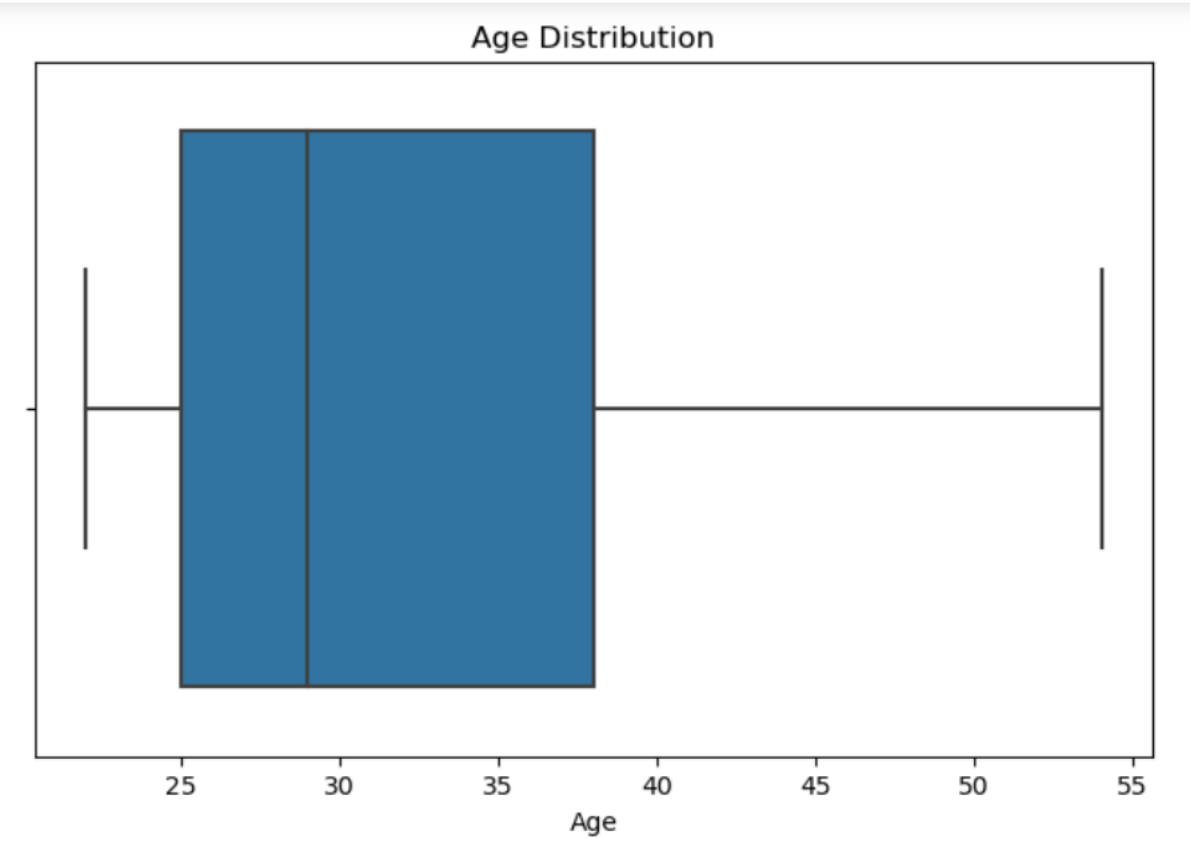
II. Education Distribution

Analysis

- From the below pie chart graph of education distribution we can conclude that 62.3% people have Post Graduation and 37.7% have Graduation.
- Hence it is evident that almost 62% car buyers are PG holders hence it will be much beneficial if we increase our marketing campaign more to the PG holders as from the dataset it is evident that they buys the car more compared to Graduate holders.



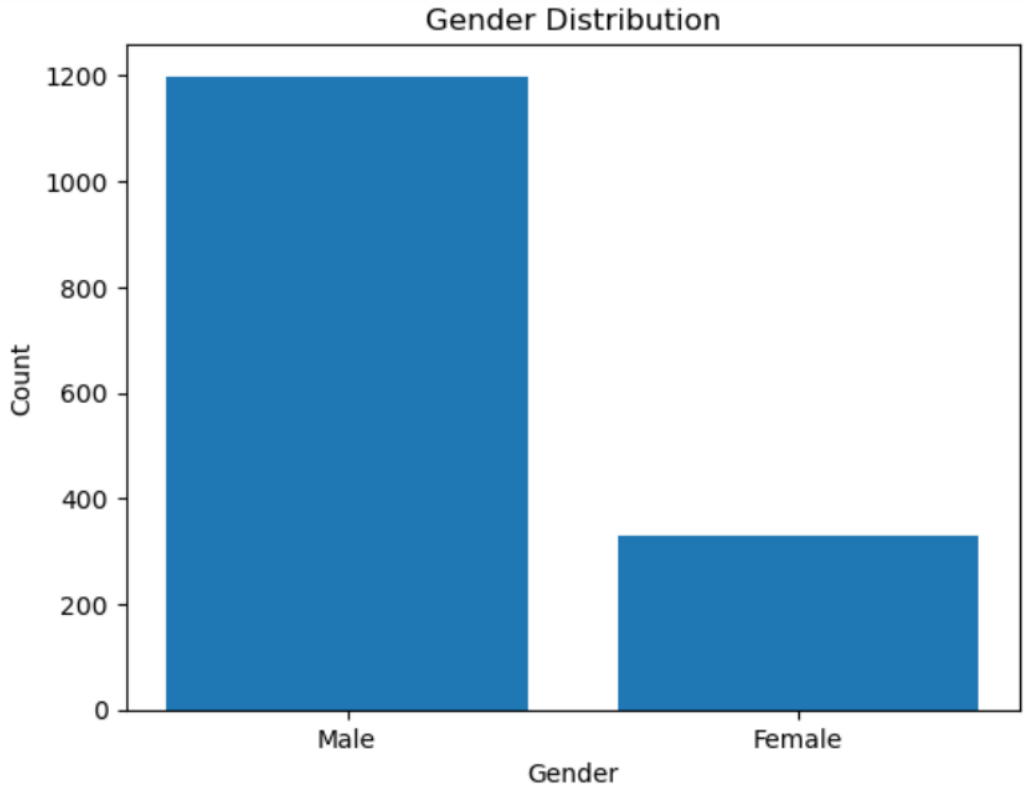
III. Age Distribution



Analysis

- Here from looking at the boxplot we understand that it is a right-skewed age distribution boxplot.
- It signifies that the majority of the age values are concentrated on the left-hand side of the distribution and there are some outliers on the right-hand side of the distribution.
- Here majority of individuals are younger, that is 50% of the population belongs to the age group from 25 to 38 years old having a median value of 30.
- Hence our targeted age group for car marketing campaign is ideal to be within 25 to 38 years old

IV. Gender Distribution

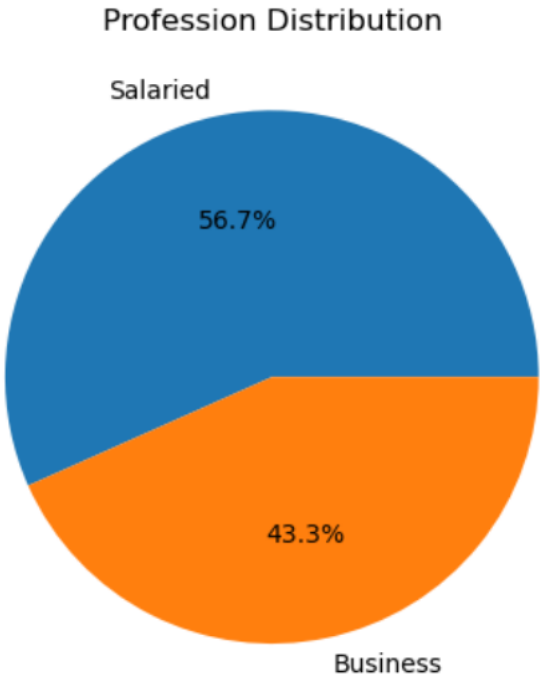


- From the gender distribution plot we identify that the majority of the customers are Male (almost 1200) and Female comes around 300.
- Here Males are almost 4 times the size of Females in the given dataset. Hence we should target more males in our marketing campaign as we can see that they buys more cars than the Females.

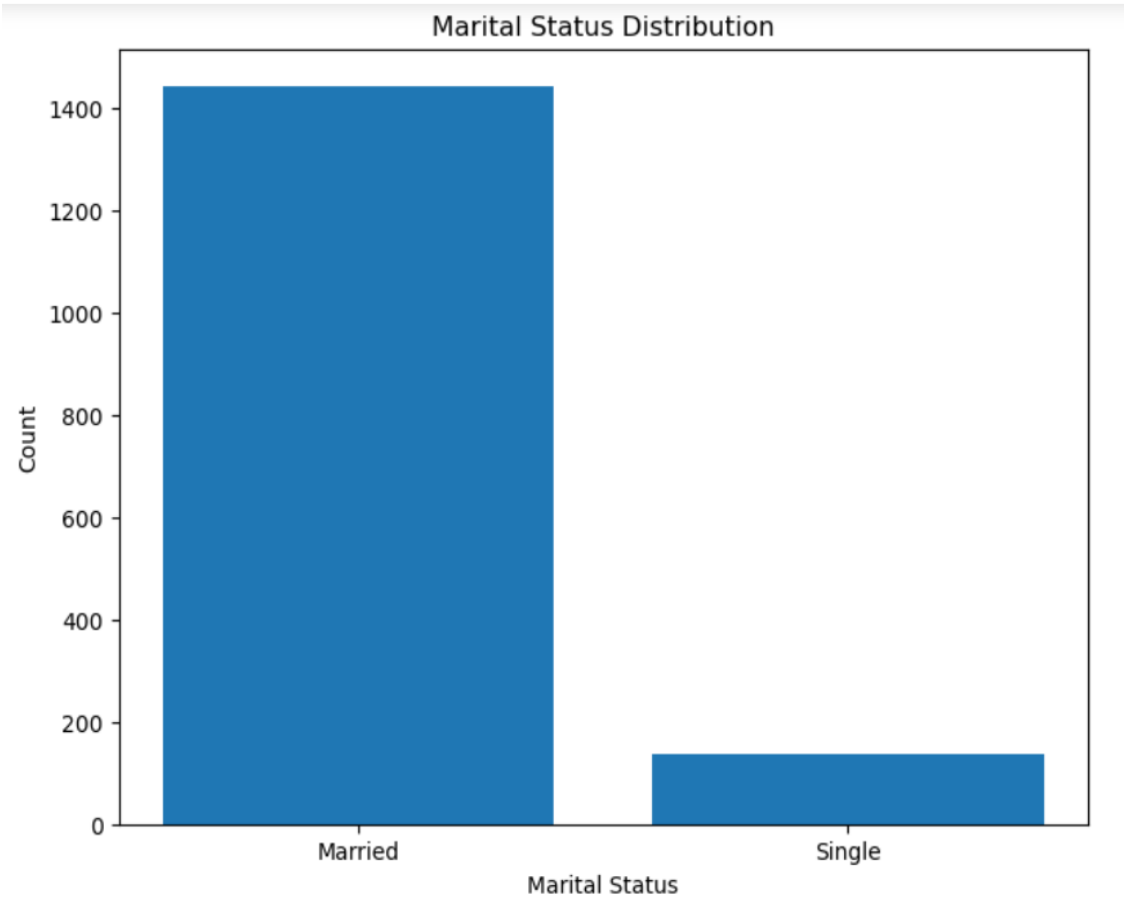
V. Profession Distribution

- From the below profession Distribution pie chart we can identify that in the given dataset of people purchasing car, among the profession almost 56.7% people are salaried working professional whereas almost 43.3% are doing business
- Since there is no much huge difference between the working professionals and the business class we can consider both of them as our target customers and

can devise various marketing strategy to both the groups.

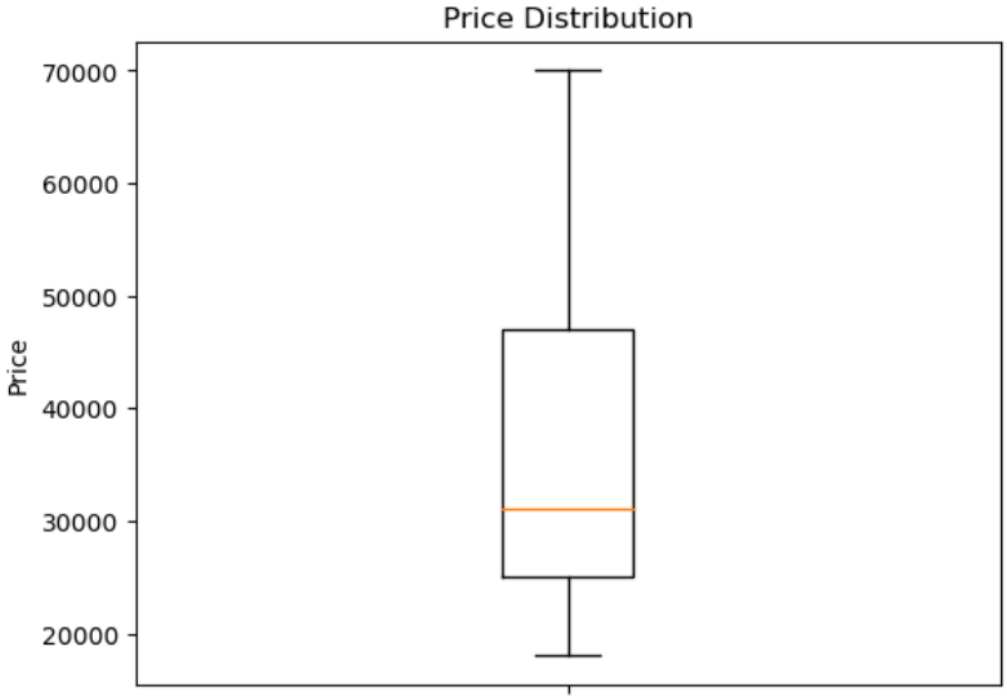


VI. Marital Status



- From the above Marital Status distribution we can identify that the majority chunk of customers are married almost nearing to more than 1400 whereas the single population is as low as 100.
- Hence we can understand that the majority of the customers who will buy the car would be married. Hence we should increase our marketing campaign to more married customers than to bachelors.

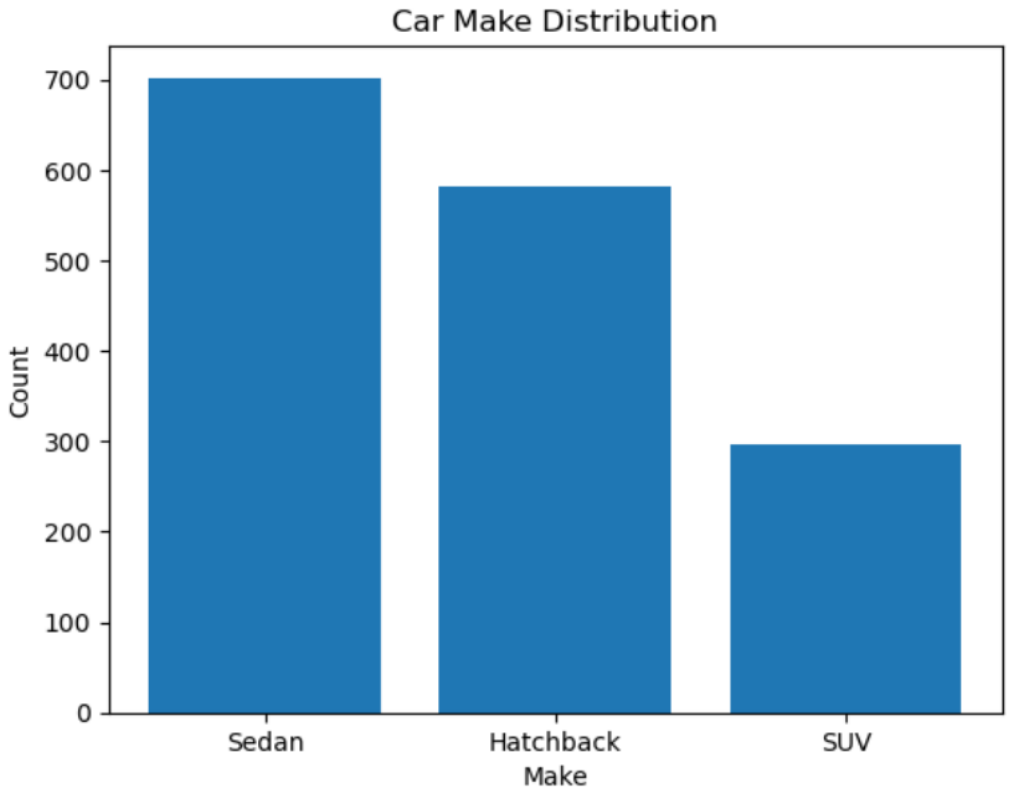
VII. Price Distribution



- From the price distribution we can understand that 50% of the price distribution of the cars are within $25,000 to almost $50,000 with the median value coming around 30,000.
- This indicate that the cars which gets more sales are the one's having price within the range $25,000 to $50,000. Hence we should market the cars within this price range more.

VIII. Car Make Distribution

- From the car make distribution we can find that the most number of sales happened for Sedan, Hence it is the most popular one with almost 700 count. Then comes the Hatchback make with second position with a count of almost around 600 and then the least preferred is SUV.
- Hence when we market we should give maximum preference to Sedan as it's the most popular car which gets maximum sales then the Hatchback Make and then the SUV.



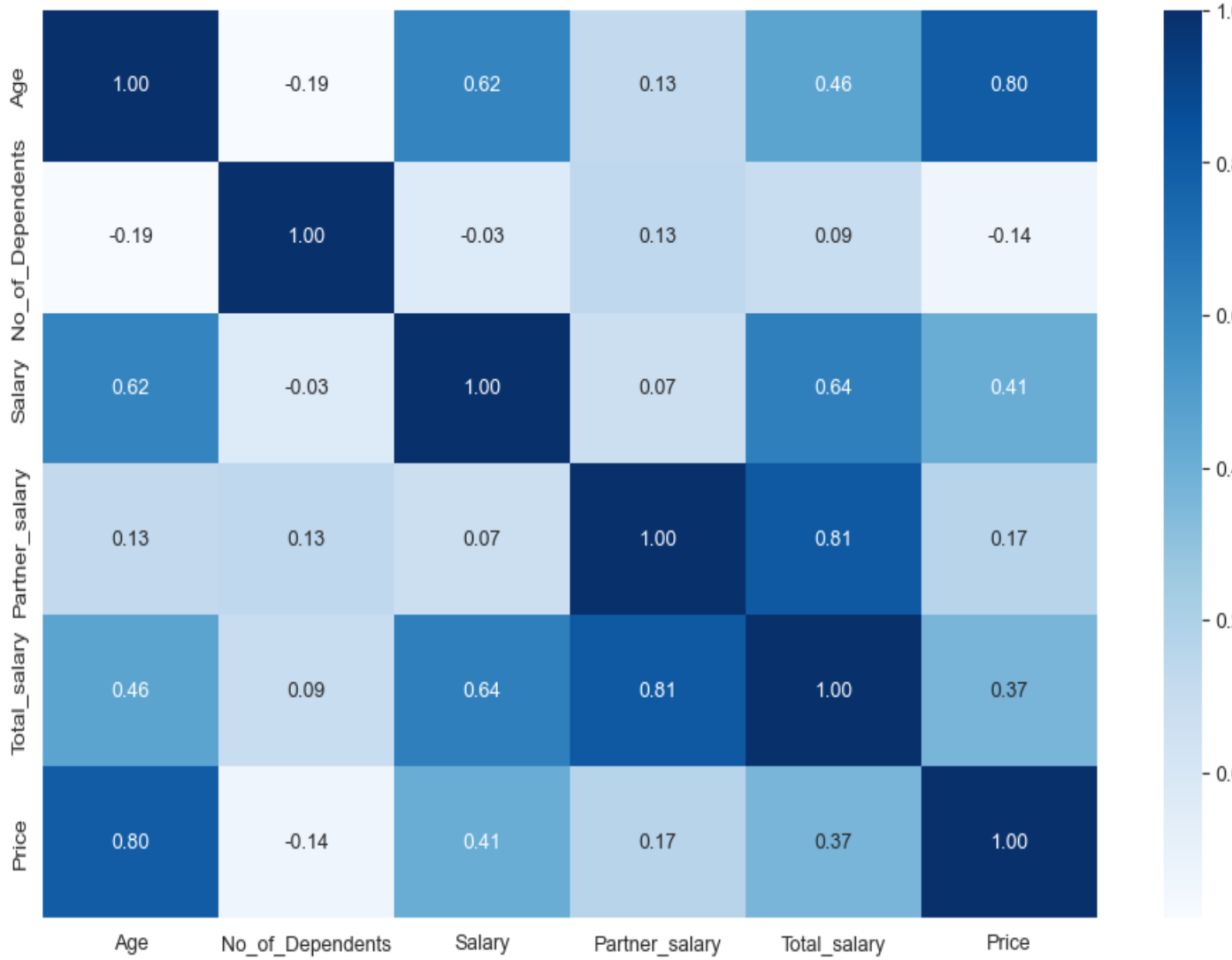
D. Understanding the relationships among the variables in the dataset is crucial for every analytical project. Perform analysis on the data fields to gain deeper insights. Comment on your understanding of the data.

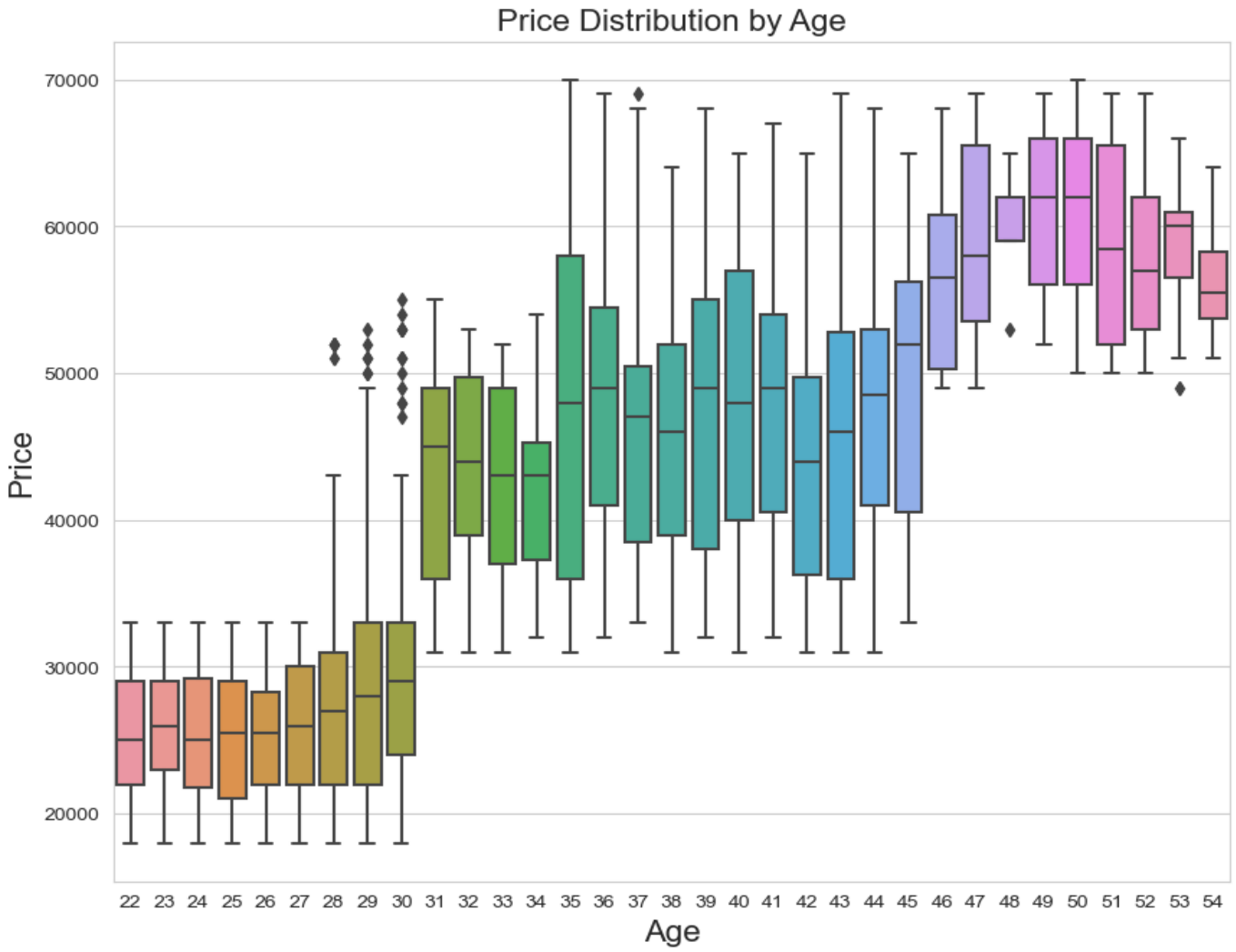
Bivariate and Multivariate analysis:

From the above correlation map and pair plot, we can derive the following:

- Age and Salary have a moderate positive correlation (0.62). This indicates that as Age increases, Salary tends to increase as well.
- Age and Price have a strong positive correlation (0.80). This indicates that as Age increases, the Price of the car tends to increase as well. Hence we can deduct that as age increases there are more chance that the person will buy a more expensive car.
- No of Dependents and Total salary have a weak negative correlation (-0.14). This indicates that as the number of dependents increases, Total salary tends to decrease slightly.
- Partner salary and Total salary have a strong positive correlation (0.80). This indicates that as Partner salary increases, Total salary tends to increase as well.
- Salary and Total salary have a moderate positive correlation (0.64). This indicates that as Salary increases, Total salary tends to increase as well.

- Price and Total salary have a moderate positive correlation (0.36). This indicates that as Total salary increases, the Price of the car tends to increase as well. Hence we can assume that a person who will have a high total salary will buy more pricey cars. Hence we can target people who are having high total salary to buy more expensive cars.

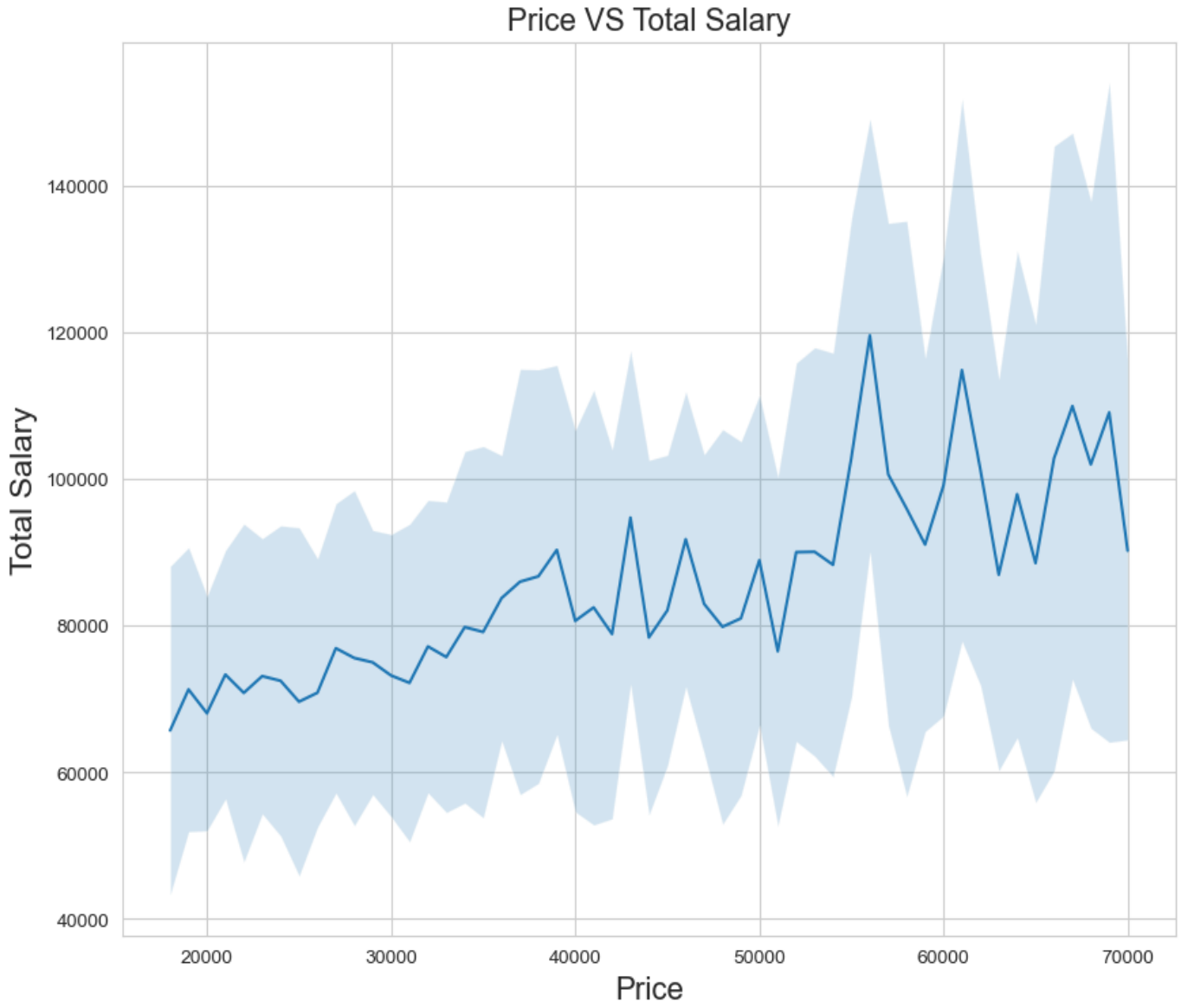




Price Distribution Vs Age

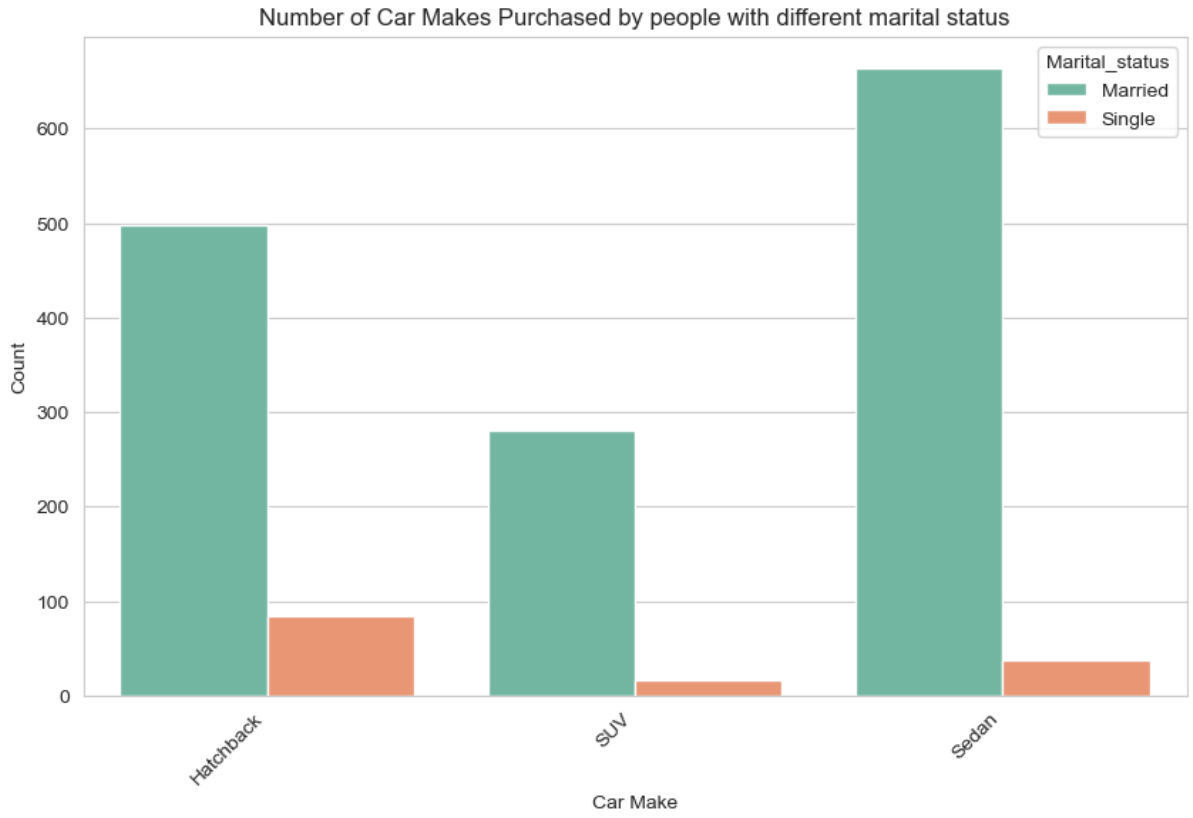
- Here we can understand that between the age of 20-30 people tend to go for low priced cars that is from ($20,000 to max around $32,000).
- In the mid 30's people tend to go for higher priced cars that is median price values ranging from $45,000 to max values reaching to almost $65,000 to $70,000.
- As the age increases as people tend to reach in their 50's they are buying cars having median values around $60,000 to max values reaching $70,000
- Hence we can say that as age increases people's purchasing power increases as accordingly we can market low priced cars to people in their 20's and more expensive cars as one ages.

Price Distribution Vs Total Salary



Here also we can understand that there is a positive correlation between the price of the car bought and the Total Salary received by the customers.

We can see that there is a steady increase meaning as the Total Salary increases the ability of the customer to buy higher priced car increases. Hence we can look upon the salary of the customers and market them the car accordingly. For a person having higher average salary we can advertise higher priced cars and so on.



Analysis

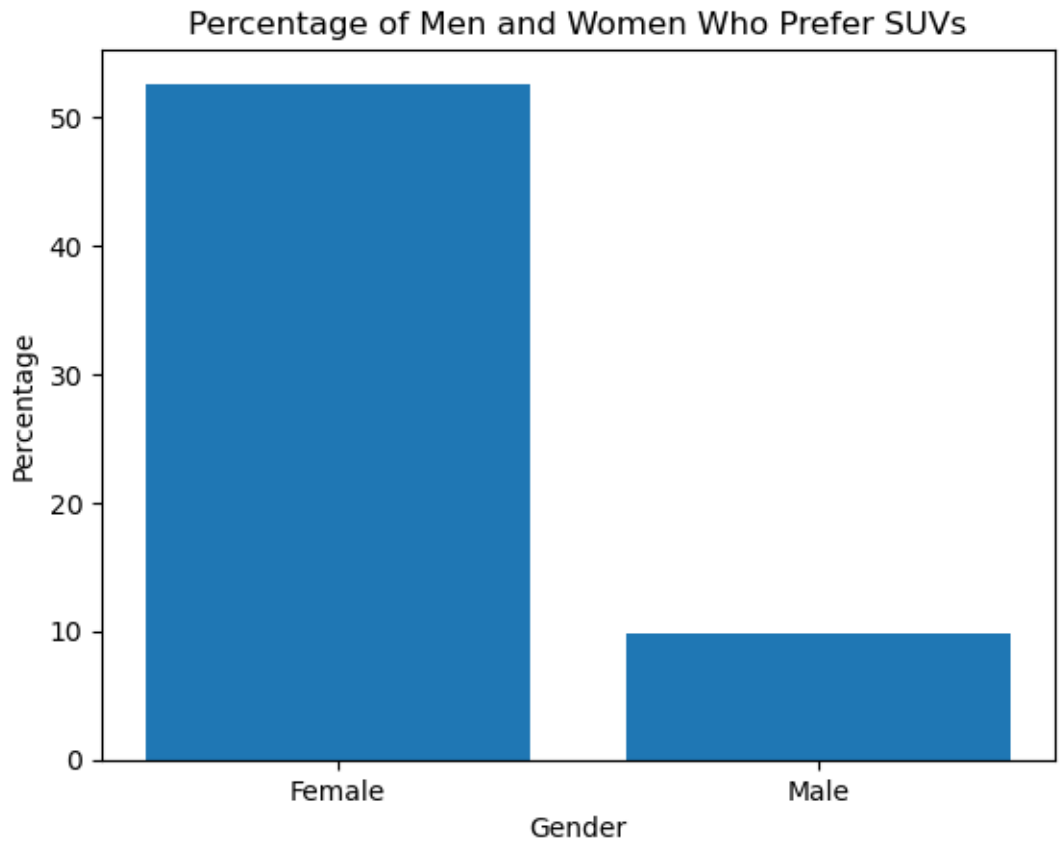
- From the above analysis we understand that the most preferred car make among married couple is Sedan and hence it should be advertised more among the married couples followed by Hatchback.
- Least favoured car make would be SUV around married couples so we can devise new marketing strategy to boost its sales among couples.
- Finally we can analyse that the number of car getting sold among singles are very low compared to married couples but even then Hatchback is most preferred among singles so we can advertise Hatchback models more among the singles followed by Sedan.
- Least favoured car make would be SUV among singles so we can devise new marketing strategy to boost its sales among singles.

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

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

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

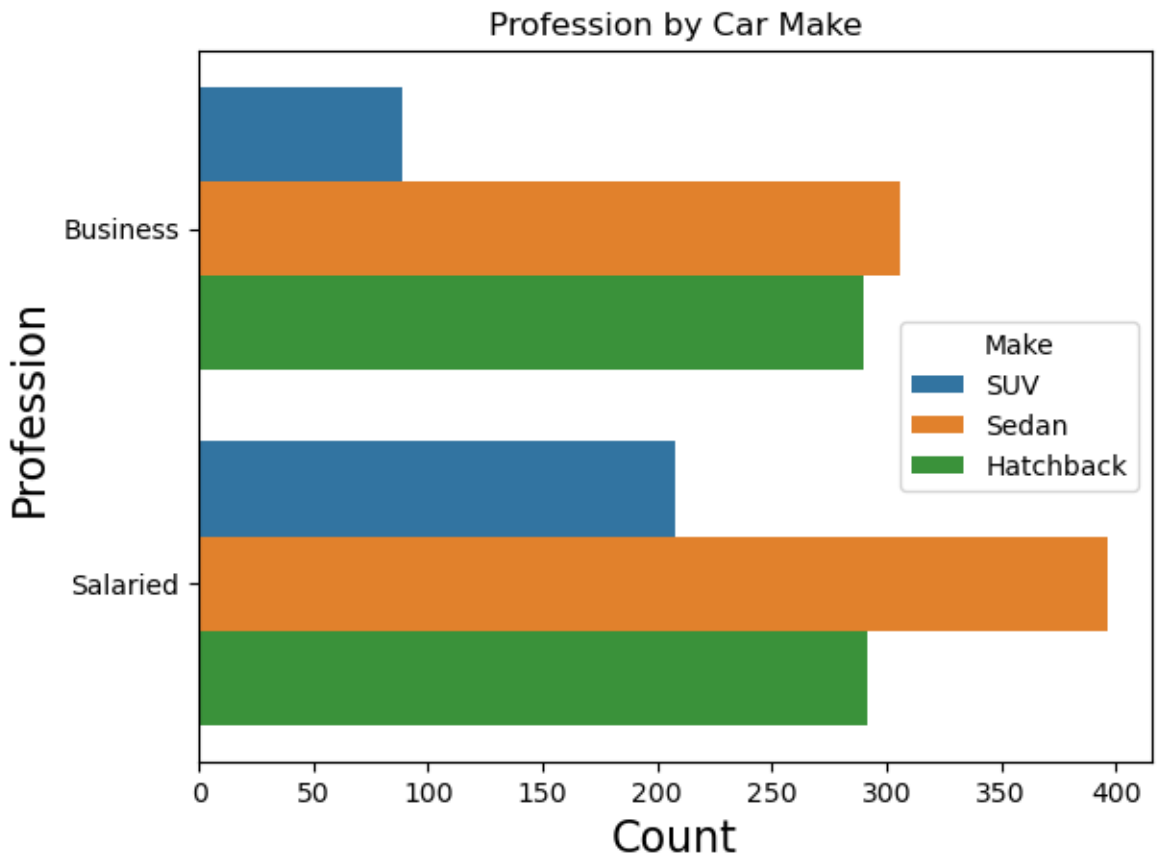
E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale.



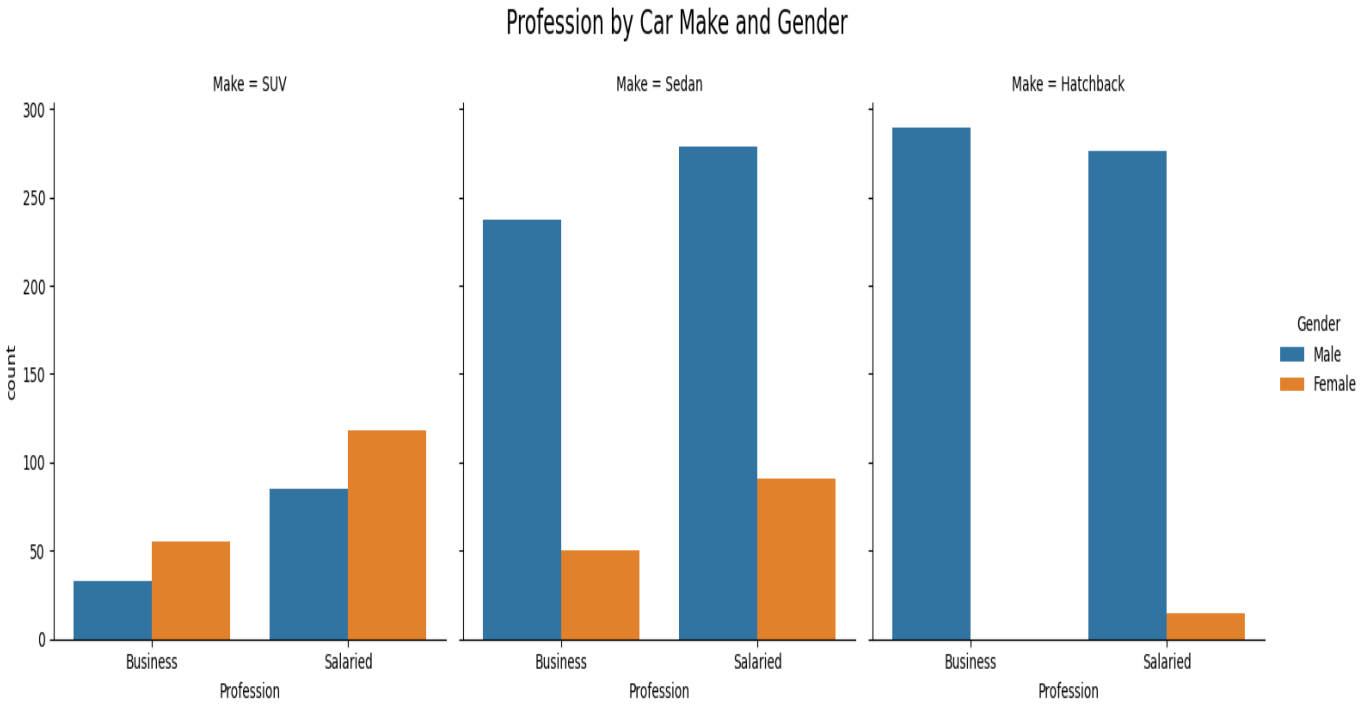
From the above plot it is clearly visible that almost 50% of Women prefer SUV whereas only a smaller percentage of Men prefer buying SUV’s (10%).

Hence Steve Roger says “Men prefer SUV by a large margin, compared to the women” is false and the truth according to the data given is ““Women prefer SUV by a large margin, compared to the Men”

E2) Here from the below graph we can find that the count of salaried professionals who are buying Sedans comes around 400, whereas a business professional who buys Sedan comes to 300. Hence Ned Stark belief that a salaried person is more likely to buy a Sedan is true



E3) Here from the graph we can find that Salaried Male are easier targets for Sedan or Hatchback than the SUV's as we can see that the number of Salaried Male opting for either Sedan or Hatchback comes around 275, whereas the same Salaried Male opting for SUV is just around 90. Hence Sheldon Cooper's claim that a salaried male is an easier target for a SUV sale over a Sedan Sale is False or wrong.

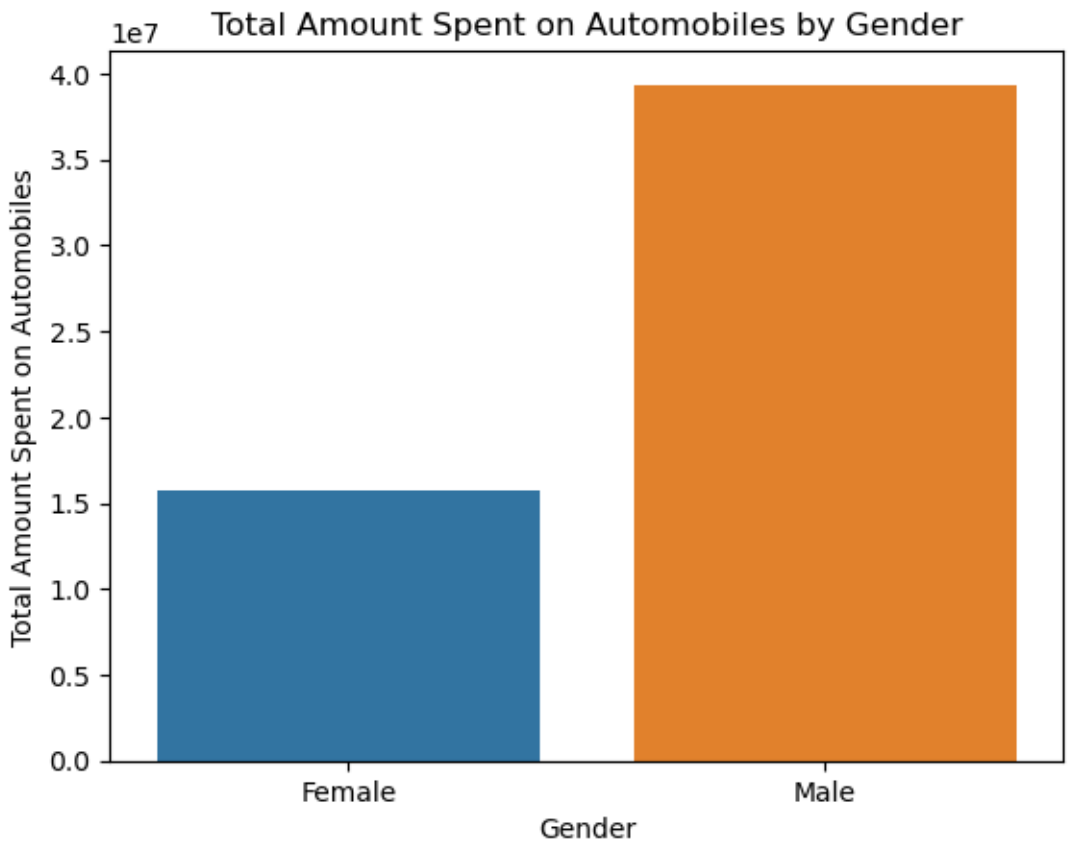


F. From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a Business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions.

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

F1) Gender

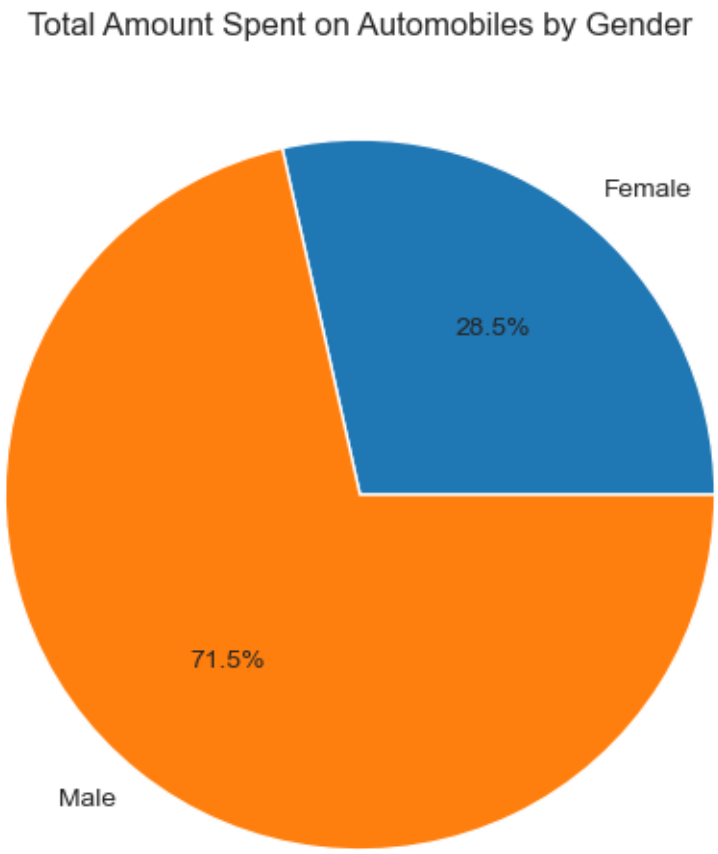
F2) Personal_loan



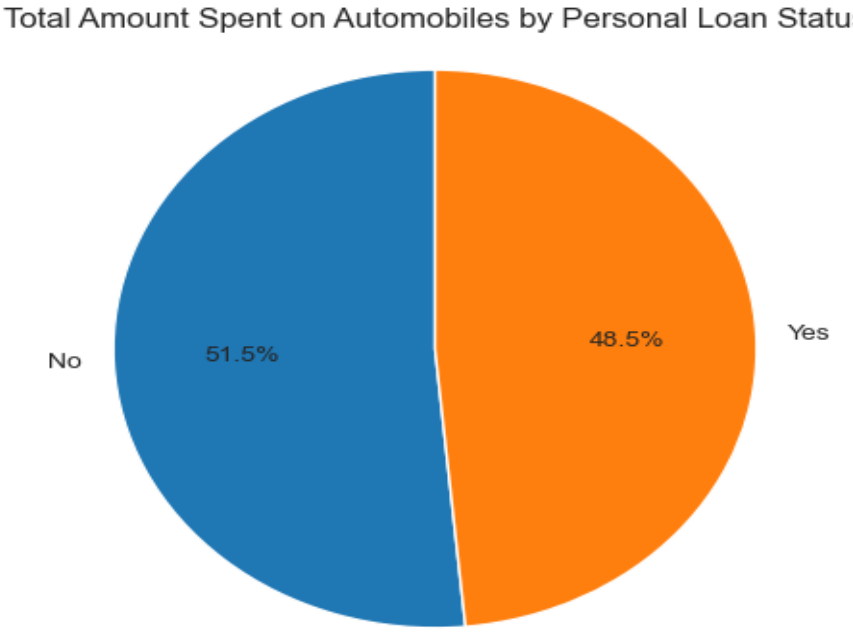
Here we can observe that more males are purchasing automobiles compared to females. They are spending more on purchasing cars than females.

Here we can say that business can tailor its marketing efforts towards men more as they have been purchasing more cars .

Hence we can say that in other words more marketing strategy should be focussed around men more than women as they have been seen spending on purchasing cars more than women.



Here we can see that among the total amount spend on Automobiles the major portion comes from Male (71.5%) whereas women have a smaller percentage (28%.) Hence we can say that more marketing campaigns should focus around men.

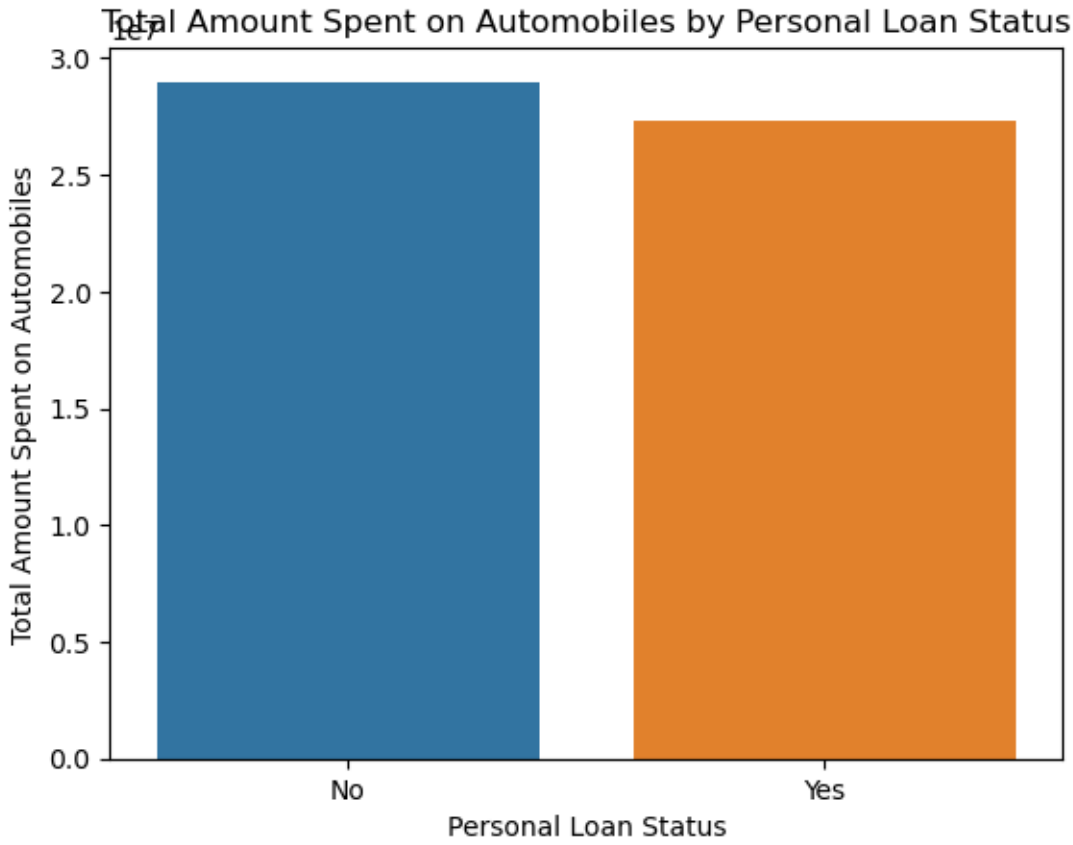


F2) Personal_loan

From the above distribution we can see that people without personal loans have spent a slightly higher percentage (51.5%) on purchasing cars compared to those with personal loans (48.5%). This indicates that personal loan status may have a small impact on the amount spent on purchasing automobiles.

Here we can utilize these results by targeting potential customers who do not have personal loans and marketing various car models to them. We could also offer financing options to this group to increase sales.

The company can also target customers without personal loans with marketing campaigns that emphasize the benefits of purchasing a car without taking a loan. For example, the company can advertise the benefits of owning a car outright, such as lower long-term costs, avoiding interest payments, and having full ownership and control of the car.



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

Here from the boxplot we understand that having a working partner or not does not have much influence on the purchase of a higher-priced car.

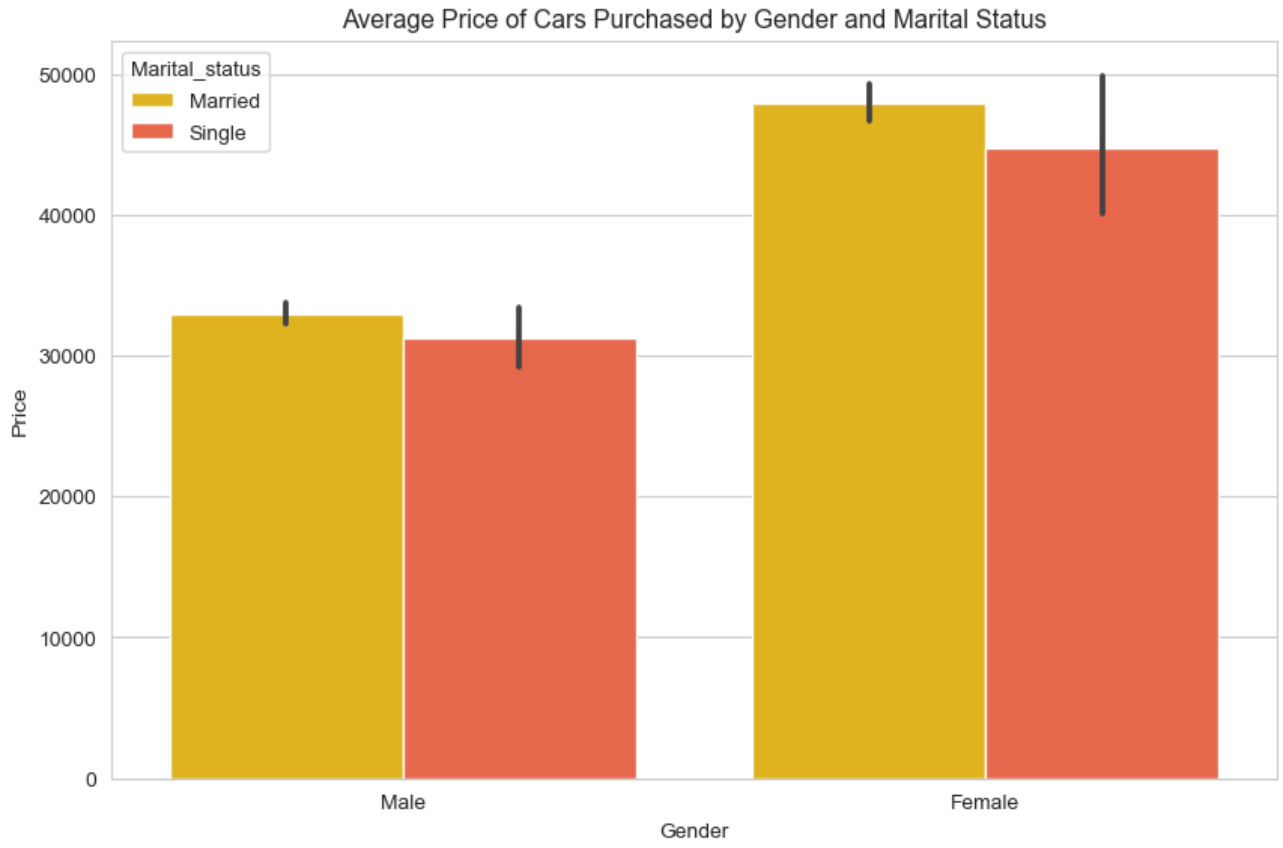
From the above boxplot it is evident that the max priced car that both group of customers whether there partner is working or not goes for $70,000, and 50% of the customers whether there partner is working or not goes for car within the range of approx.. ($25,000 to $45,000).

Hence we can say that having a working partner or not does not leads to the purchase of a higher-priced car.

Here also we can see that having a working partner or not does not leads to the purchase of a higher-priced car.

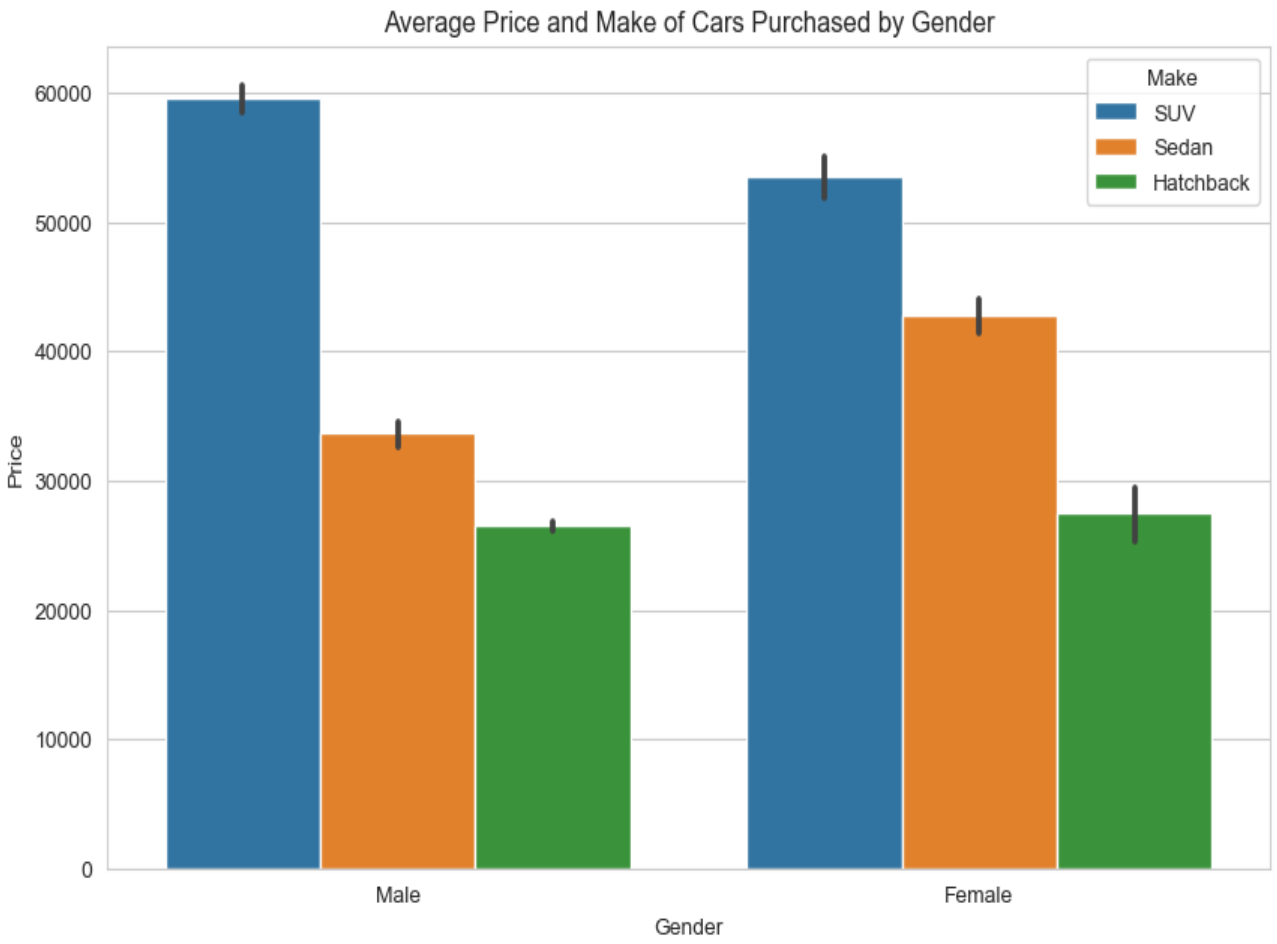
As the bar plot clearly don't show much difference between the prices of car bought by customers having working partner or not.

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.



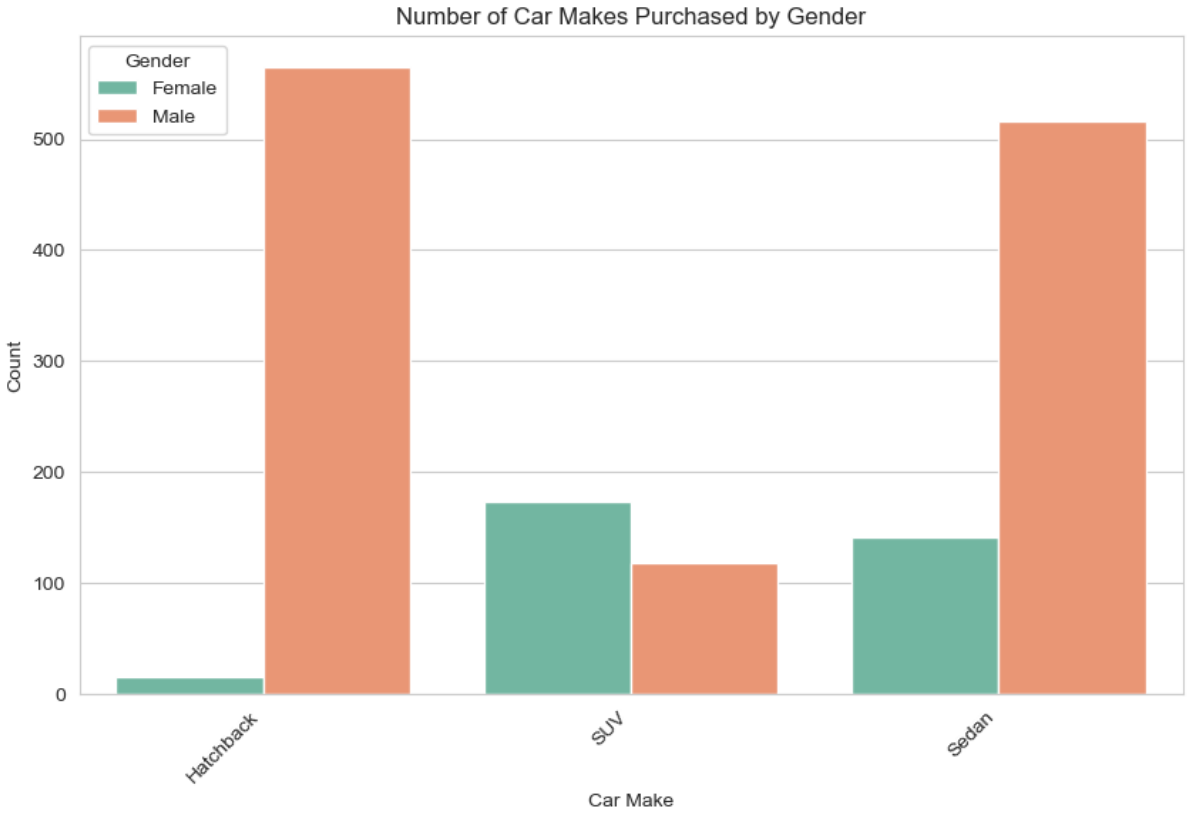
Analysis

- From the above plot we can see that Females tend to buy more priced cars than men.
- Among the Females, married women tend to buy cars with an average price ranging from approx.$45,000 to $50,000 whereas single women go for a car with an average pricing range of $45,000
- On the contrary, males tend to go for a much lesser priced cars ranging $30,000.
- We can see that there is no much difference between the prices of the cars bought by either married or single males.
- From this we can make an improved marketing strategy to target females (both married and single) with more pricey cars as the analysis show that they tend to buy cars worth $45,000 to $50,000 than the males.
- On the other hand we can target males (both married and single) with an improved marketing strategy to show them middle priced cars within $30,000 as from the analysis they tend to buy cars within that price range more.



Analysis

- Here we can see that SUV is having highest avg price range among the other car make, hence we can say that more marketing campaign should be done with SUV car make having price range within $50,000 to $60,000 among both males and females.
- Also since price range of SUV's that are getting sold is high that is more than $50,000. Hence we can market more SUV's within avg price range of above $50,000 to both males and females.
- Hatchback is having the lowest avg price range within $25,000 and is both bought within the same price range by both males and females. Hence we say that more marketing campaign should be done around Hatchback's having avg price range around $25,000 to both males and females.
- The Sedan model within the price range $42,000 is mostly bought by females. Hence Sedan model within that price range should be marketed more to females and then Sedan models within an avg price range within $32,000 is mostly sold around Male hence it should be marketed more around men.



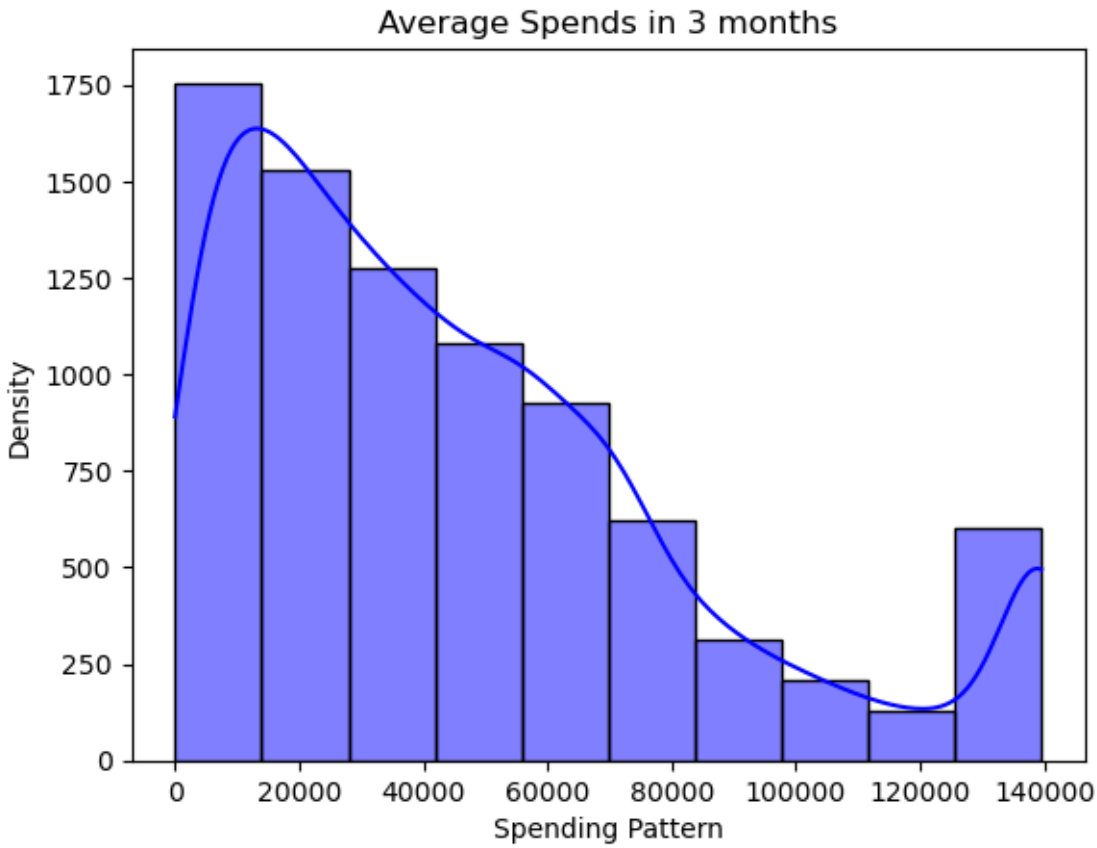
- Here we can say that Hatchback is most preferred by Males as it has the highest count of sales recorded. Hence it should be more marketed around men followed by Sedan and the least favoured by men is SUV.
- Hence Hatchback should be more marketed around men followed by Sedan and then SUV.
- Among females the most preferred one is SUV hence it should be marketed more followed by Sedan and then the Hatchback.

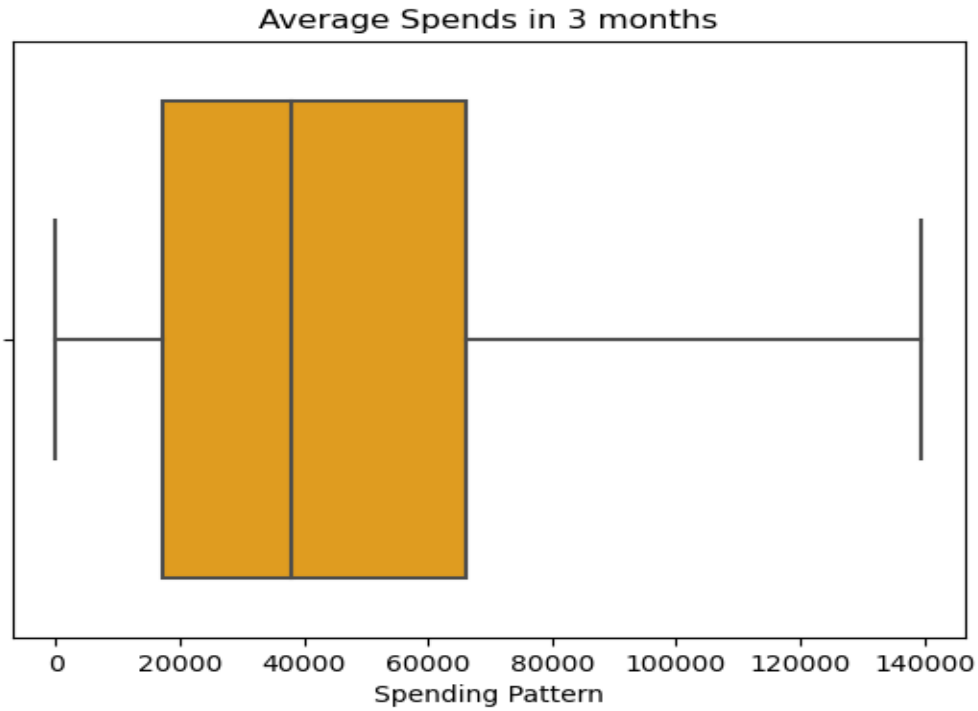
I) Analyse the dataset and list down the top 5 important variables, along with the business justifications.

The objective of the problem is to identify the top 5 important variables from the given dataset to revisit GODIGT Bank's credit card policy. Since the bank is observing high attrition in credit card spending, the bank wants to make sure that the card given to the customer is the right credit card to ensure customers spend more on credit cards, and the bank makes a profit only through customers who show higher intent towards a recommended credit card. Therefore, the selected variables should be able to help the bank identify customers with higher intent towards a recommended credit card.

The selected variables should be relevant to the bank's credit card policy and business objectives. The following are the top 5 important variables from the dataset, along with their business justifications:

avg_spends_l3m: This variable represents the average spends of a customer in the last 3 months. This variable can help the bank identify customers who are likely to spend more on their credit cards and target them for cross-selling other credit card products. Customers who have higher spends are more likely to have a higher intent towards a recommended credit card.

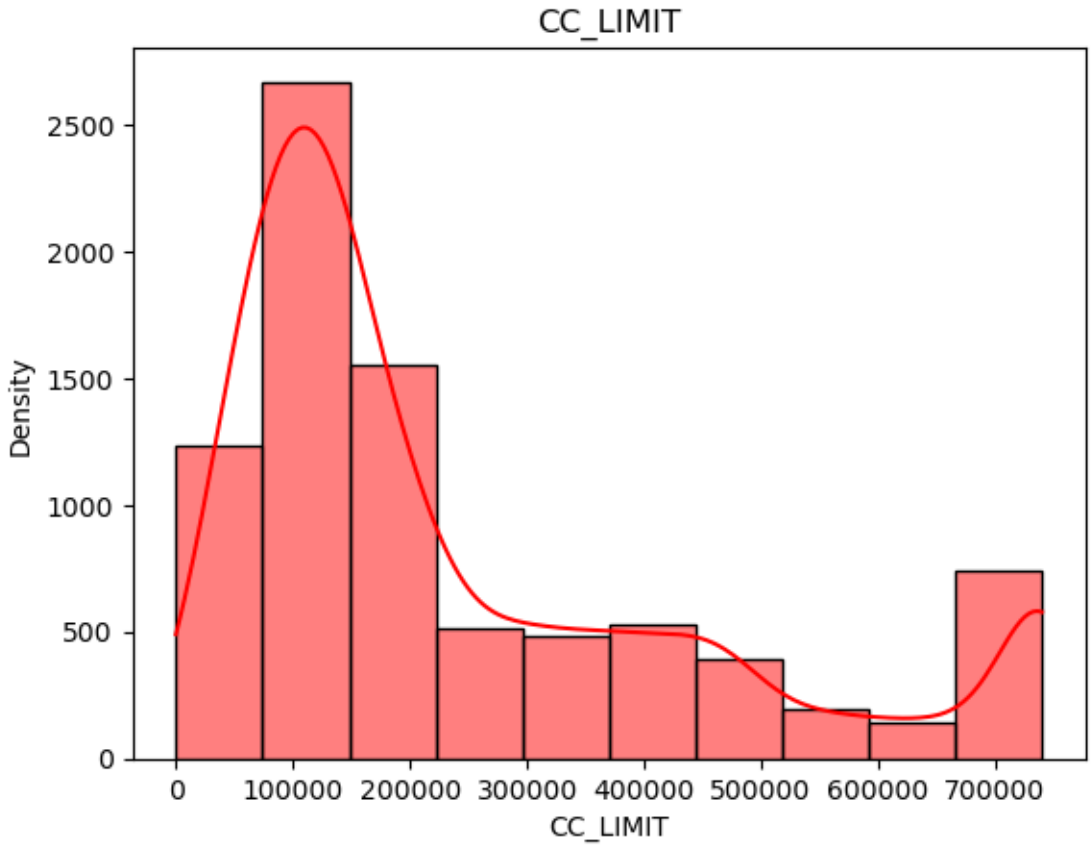




Here we can see that 50% of the customers tend to be spending on average of $20,000 to $60,000, with max value amounting to $140000.

Hence we can target/market credit card with cc_limit of within this range more to the customers as the customers will be much benefitted with this credit limit as this is there avg spending pattern and hence the chance of keeping it more.

cc_limit: This variable represents the credit limit assigned to the customer. This variable can help the bank identify customers who have the financial capacity to spend more on their credit cards. Customers who have higher credit limits are more likely to have a higher intent towards a recommended credit card.



Here we can see that there is a big density curve over cc_limit within $10,000 to $20,000 range hence we can understand that the credit with cc_limit within this range is more . Hence we can advertise such cards more to our future customers and there are more chances that they will keep this card.

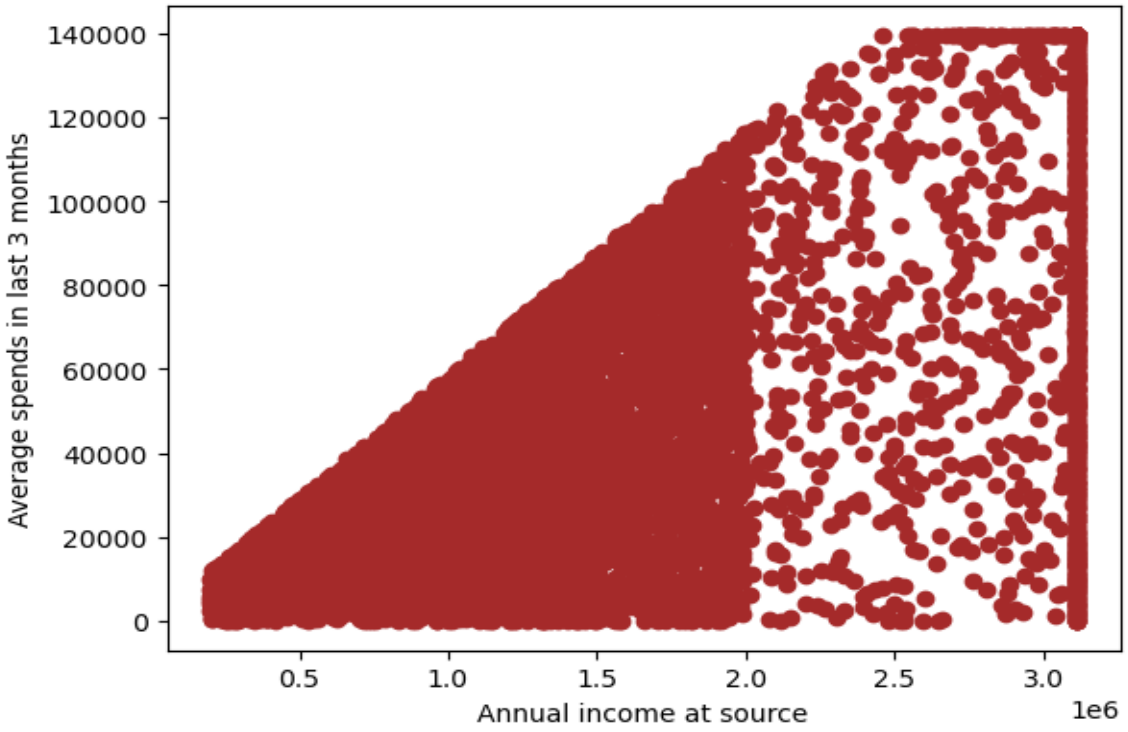
Here we can see that as the amount increases the density decreases meaning as the cc_limit increases less number of cards or people are using such cards but there is a slight increase in cards having cc_limit $70,000 meaning there are some high profile customers who will be using such high credit limits.

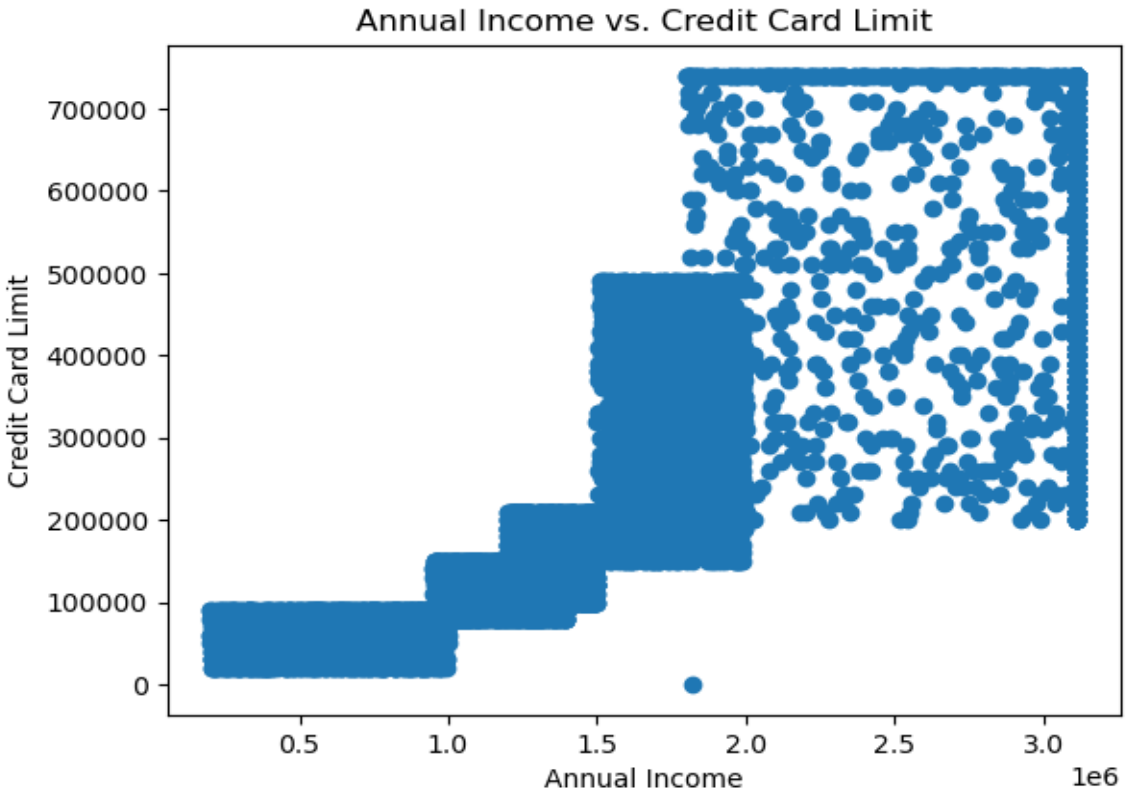
annual_income_at_source:

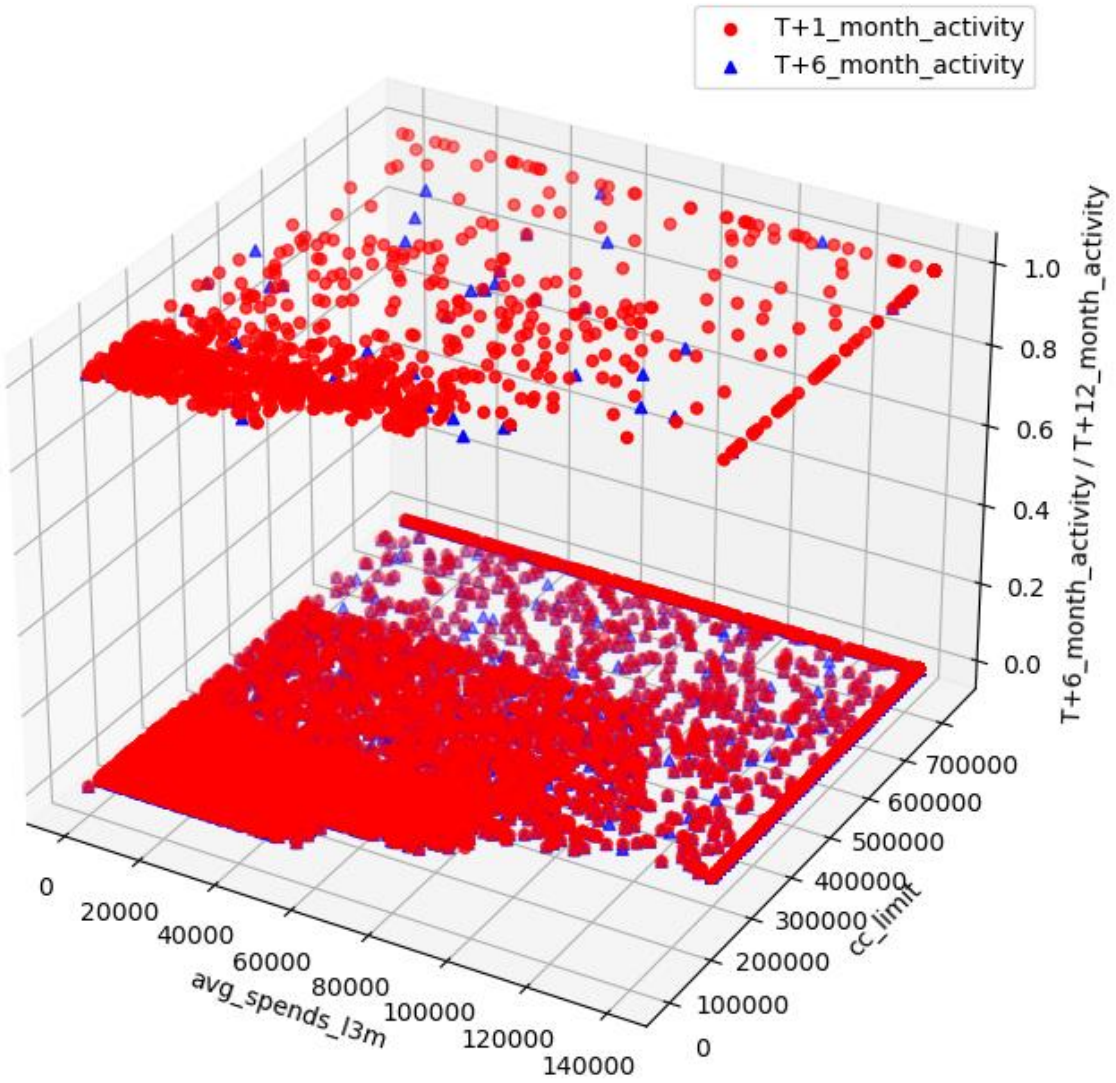
From the below graph we can identify that as the annual income of the customers increases there spending patterns increase as well.

Also we can see that as the annual income increases customers are able to keep credit card with higher cc_limit.

Hence we can easily target customers having higher annual income to purchase credit card with higher cc_limit as there from our analysis we can see that customers who have higher annual income will have higher average spends and will be able to keep the cc_limit as well


140000
120000
100000
80000
60000
40000
20000
0
Average spends in last 3 months
0.5
1.0
1.5
2.0
2.5
3.0
Annual income at source
1e6

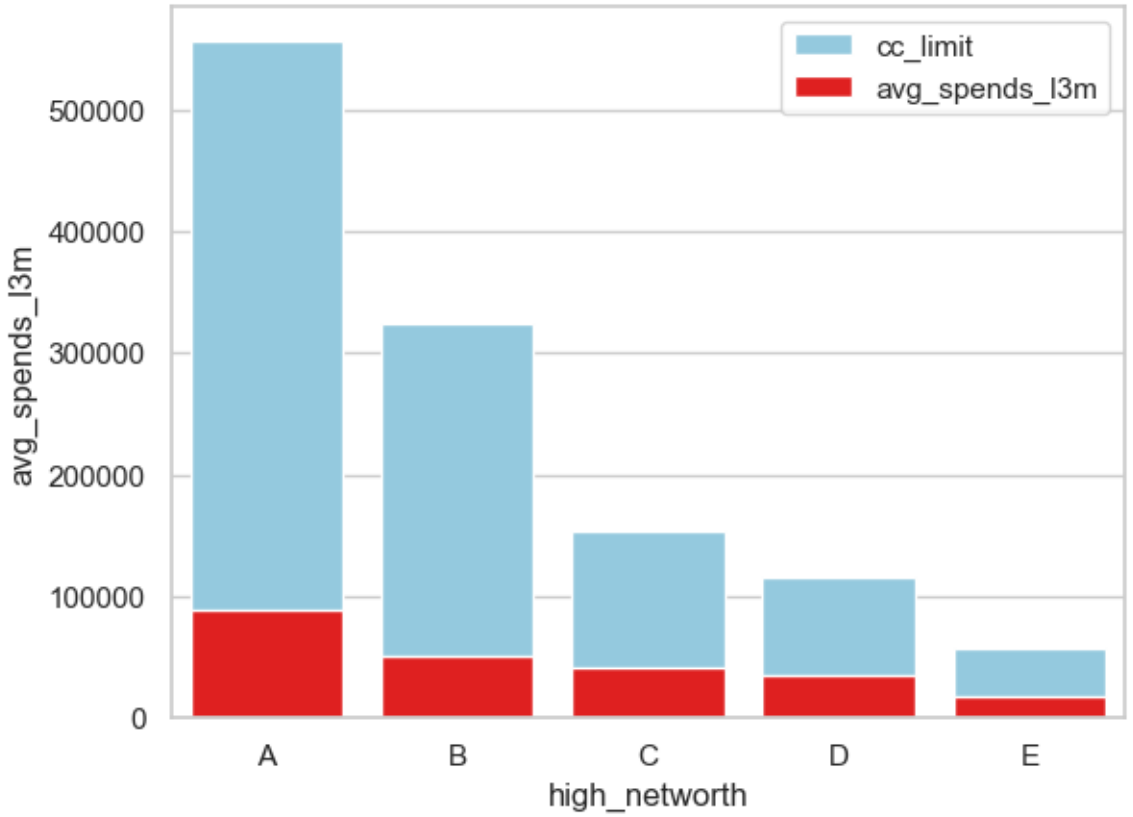

Annual Income vs. Credit Card Limit
700000
600000
500000
400000
300000
200000
100000
0
Credit Card Limit
0.5
1.0
1.5
2.0
2.5
3.0
Annual Income
1e6



T+1_month_activity: This variable represents the customer's activity in the first month after acquiring the credit card. This variable can help the bank identify customers who have been actively using their credit cards and are more likely to have a higher intent towards a recommended credit card. Customers who have been active in the first month after acquiring their credit card are more likely to be receptive to cross-selling credit card products.

From the above 3D plot we can see that there are more number of people who are active at the first month but as the month passes we can see that they are not spending on their credit cards or they have either stopped using it. Hence we can check on the reasons as why people are stopping after the first month and we can provide credit card according to people’s spending pattern.

high_networth



It is actually customer category based on their net worth value from (A:Highest to E:lowest)

Here we can see that customers with high profile tend to have higher average spends and they tend to keep their cc_limit high. Hence we can target people with higher profile to purchase credit cards with higher cc_limits.

We can see that the cc_limit and average spend on the credit card tend to decrease from A to E meaning as their networth decreases their credit card limit and spend tends to decrease as well.

Hence we can check our customer's net worth profile and suggest them credit cards accordingly.