Chetan R Deshpande

SMDM Project Report

Contents – **Page No.**

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

The technical information that an administrator would be interested are the shape and datatype of the variables.

In the given dataset, there are **1581 rows and 14 columns** out of which -

1 is type – float

5 are type – int

8 are type – object

*This was achieved by using the functions – shape & info().

1B. 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.

From the Jupiter Notebook, we can look at the preliminary analysis of the variables.

After the initial analysis using functions like – head(), tail(), shape, info(), describe() we see that we have a clear understanding of what the dataset consists of using head() and tail(). We can also understand the number of columns and rows is equal to 14 and 1581 respectively. The info() function has shown that there are 3 types of datatypes in the given dataset and using the decribe() function we see the statistical summary of the dataset in a simplified table.

Further we have performed quality check using - df.isnull().sum() & df.nunique(), from which we have tracked the **existing inconsistencies.**

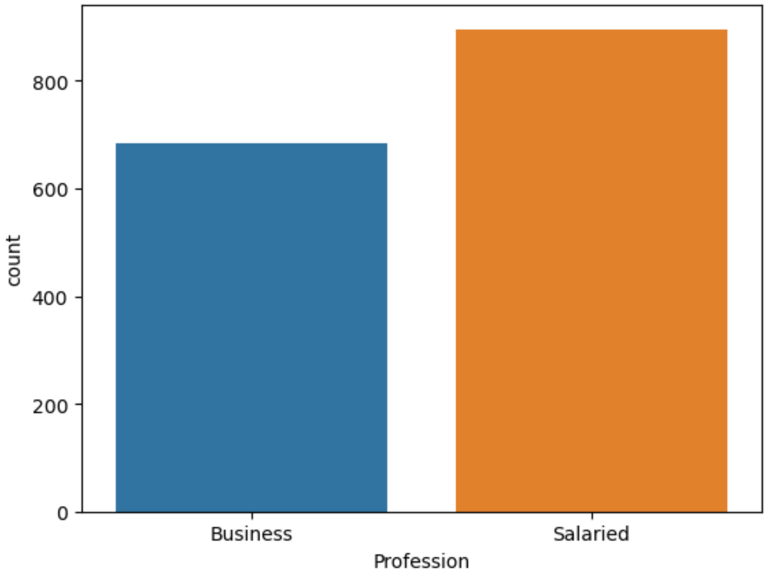
```
Age                      0
Gender                  53
Profession               0
```

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

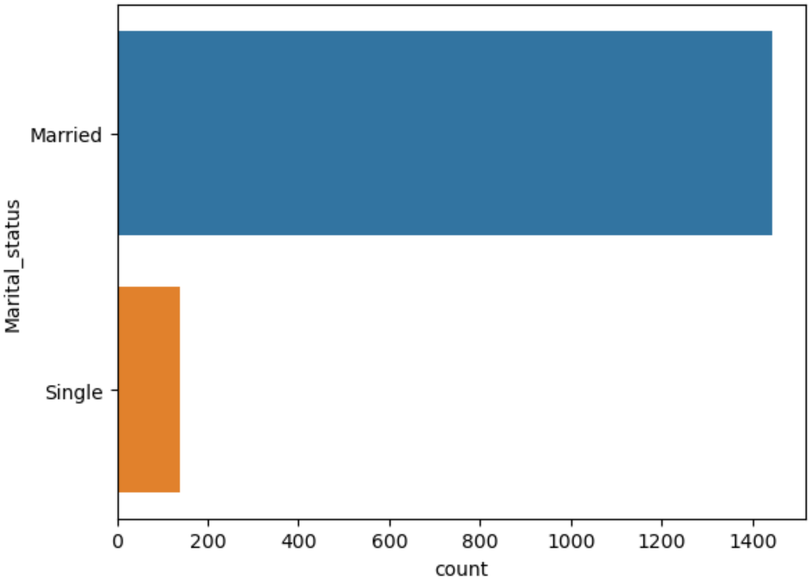
Here, for the inconsistencies such as spelling errors for Female like Femal and Femle has been changed to Female and the Null value have been changed to Male as the mode was Male using replace() and fillna(). Later, the missing values in Partners Salary have been treated with the mean and the data cleaning was achieved.

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

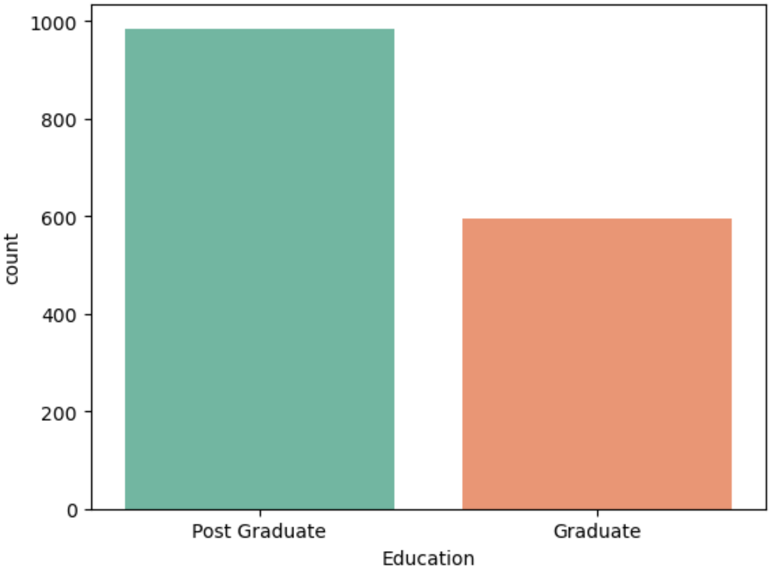
1C. Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.



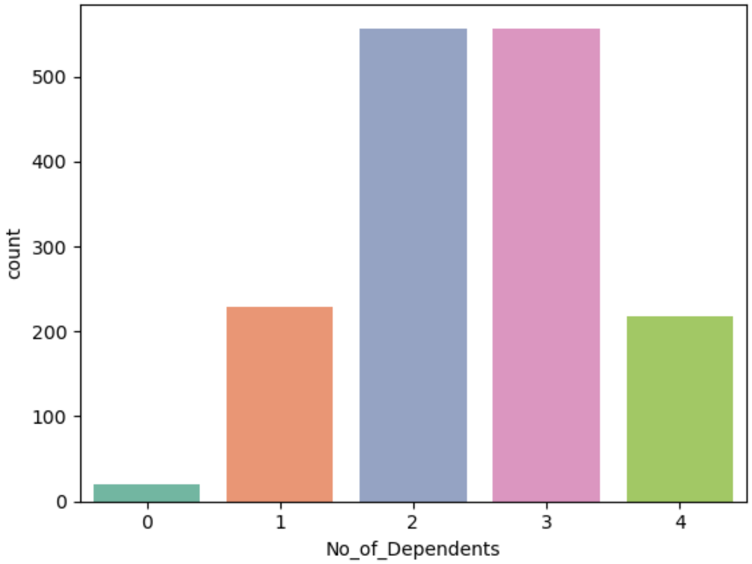
Here we can see that the number of salaried employees are more than the number of people who own a business.



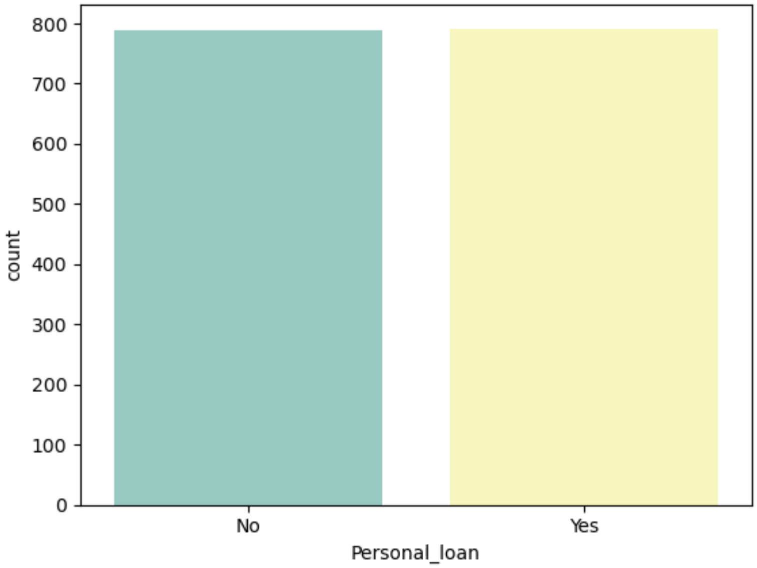
From the above graph we can see that there are a greater number of married people than single people.



From the above graph we can infer that we have over 980 Post Graduates and around 600 graduates.



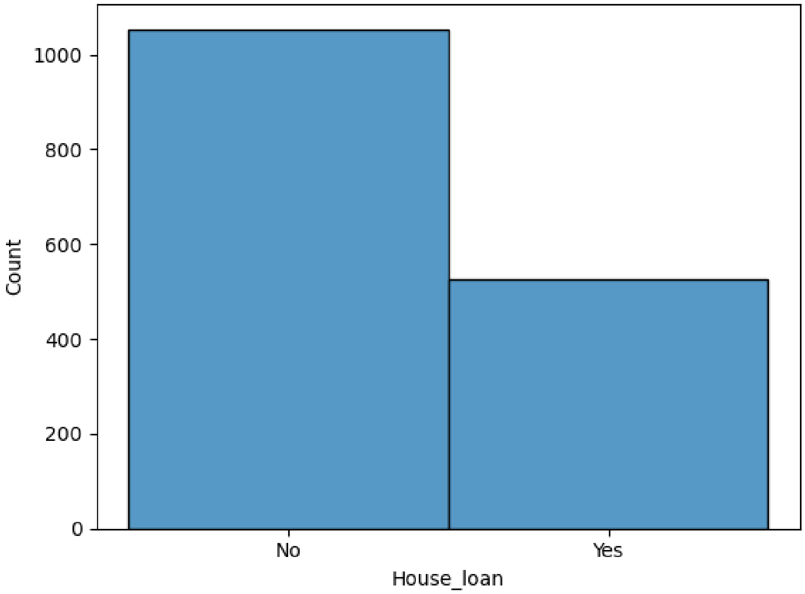
The above graph represents the number of dependents per person, and we can see that highest dependents are 2 & 3 and lowest being 0.



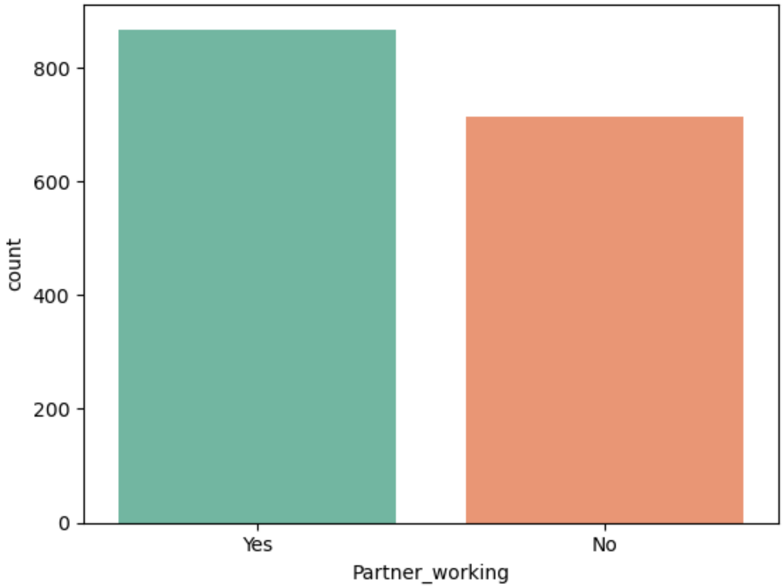
Here, we can see that the number of buyers who have opted for a loan is almost same to the buyers who have not opted for a loan.

We have verified this by using the value_counts(), where the output was -

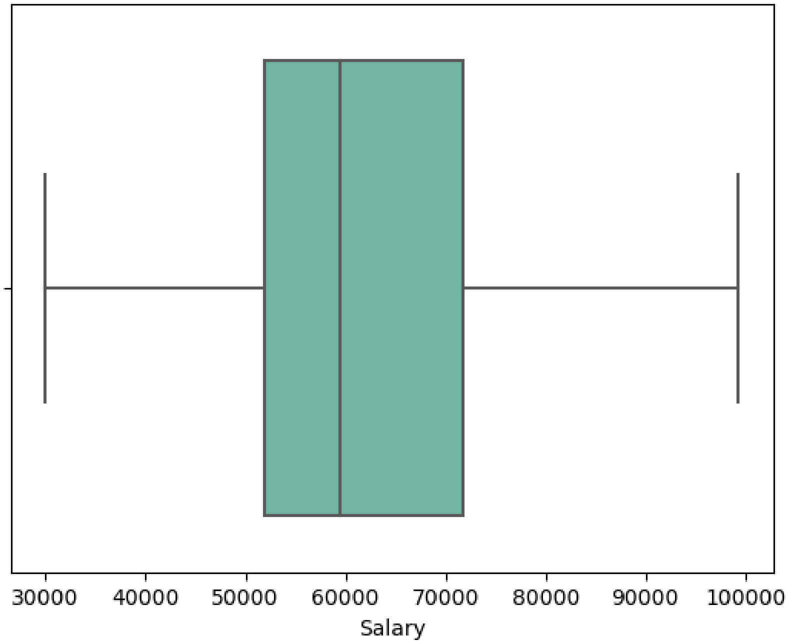
```
Yes    792
No     789
Name: Personal_loan, dtype: int64
```



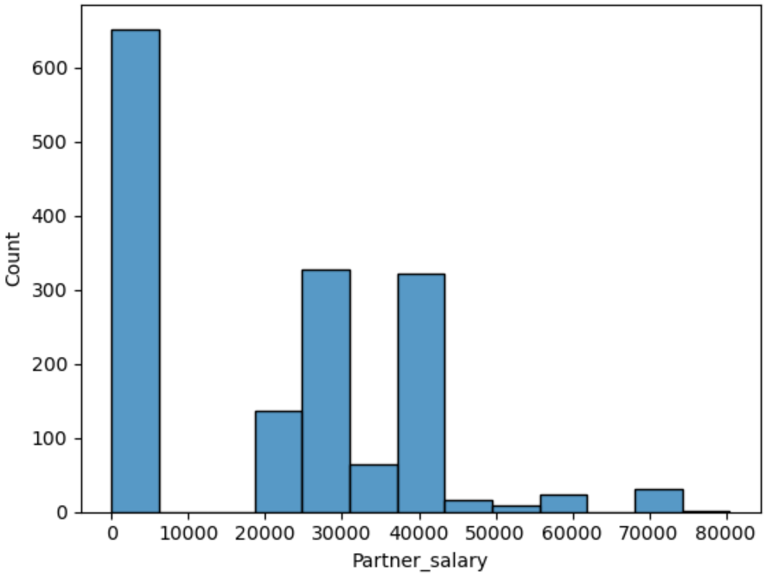
Here from the above bar plot, we can see that there are more number of buyers who don't have a house loan and the number of buyers who have a house loan are around half of those who don't have a house loan.



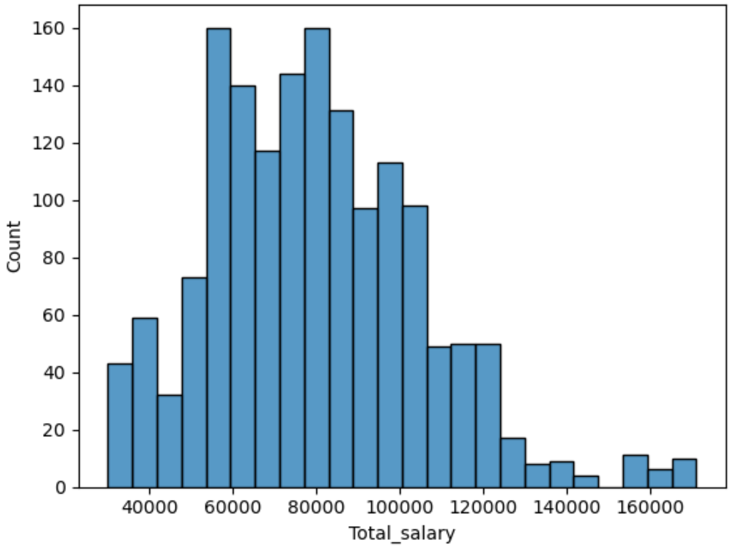
From the above graph, we see that the buyers having a working partner is higher and the buyers who don't have a working partner is just slightly lesser and there's no significant difference.



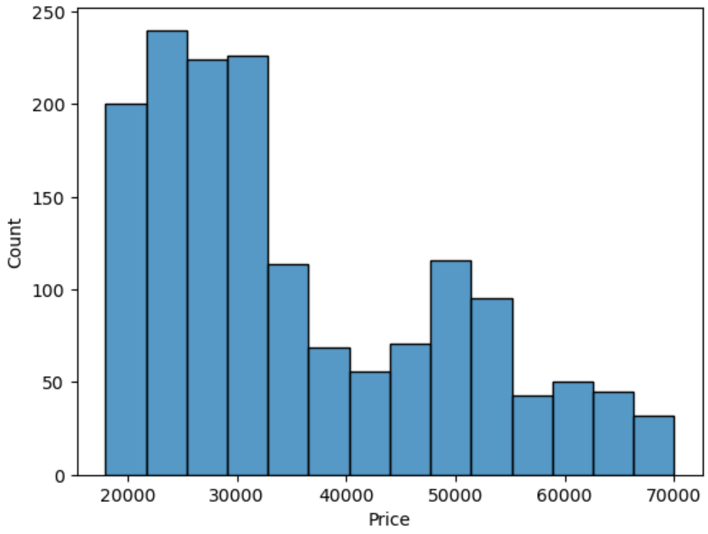
From the above boxplot, we can see the Q1, Mode and Q3 values along with min and max values. Majority of the salary lies between 50000 and 80000.



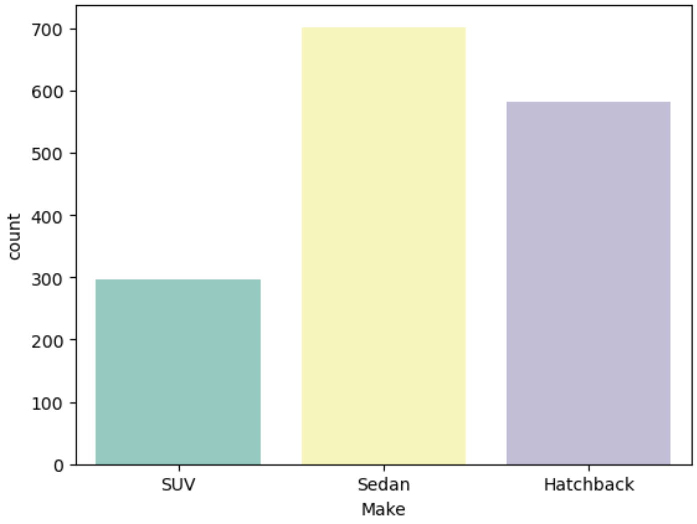
This graph shows that most of the partners’ salary is 0 and shows the other counts till 80000.



From the above histplot, we can see the distribution of the total salary of the buyers when the buyers’ salary is added with the partners’ salary if they are married. Here, the highest total salary lies near 60000 and 80000.

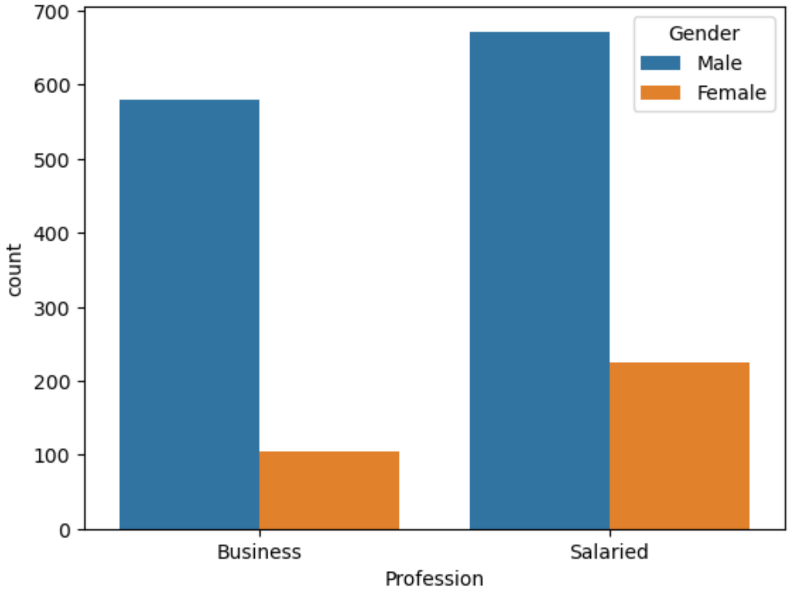


From this graph we can understand the distribution of the price of the car which ranges from 20000 to 70000 where the highest number of buyers have spent around 20000 to 30000.

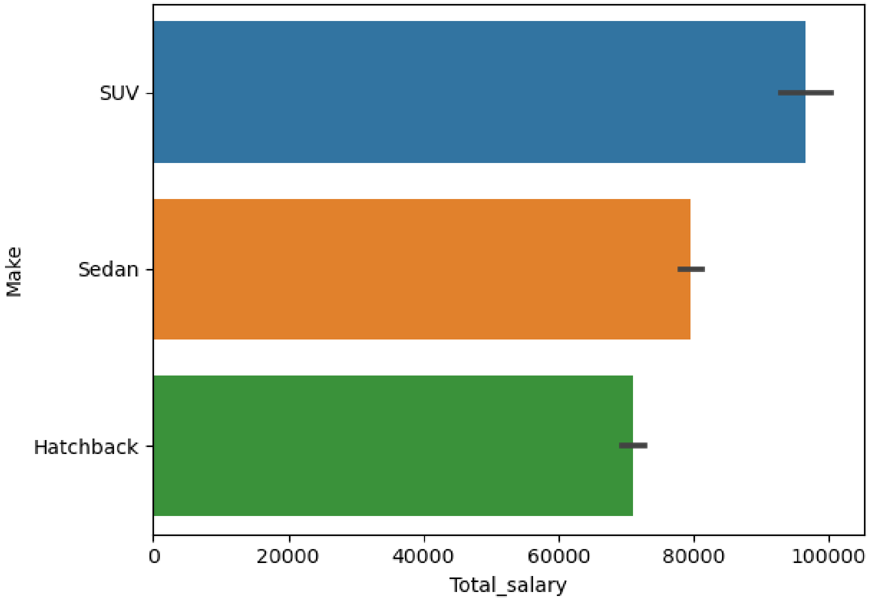


From this graph we see that the most sold car type is Sedan and the least is SUV.

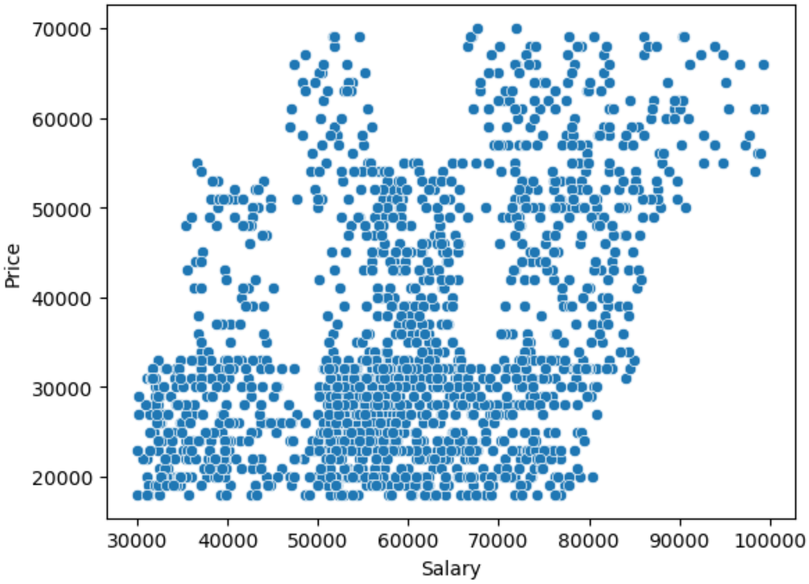
1D. Understanding the relationships among the variables in the dataset is crucial for every analytical project. Performing analysis on the data fields to gain deeper insights. Commenting on our understanding of the data.



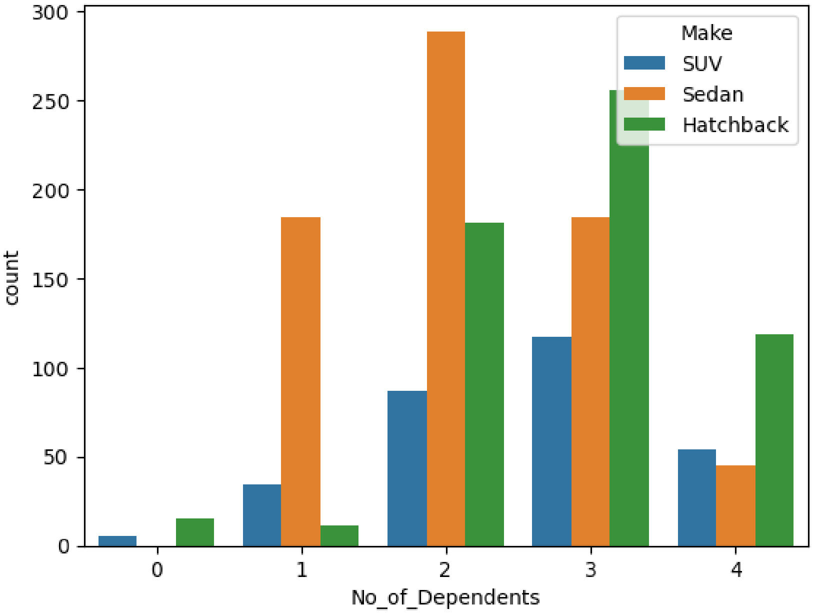
From the above plotted graph, we can understand that the number of male counts outstand female counts who own a business and are salaried.



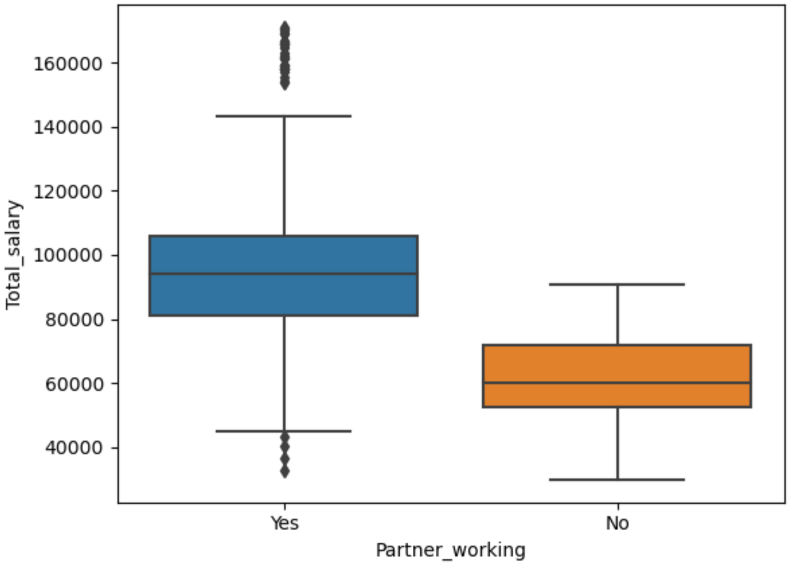
From the above graph we can infer that people with higher drawn total salary tend to buy SUV, whereas comparatively lower salaried people tend go for Hatchback and the people who lie in between this range prefer Sedan.



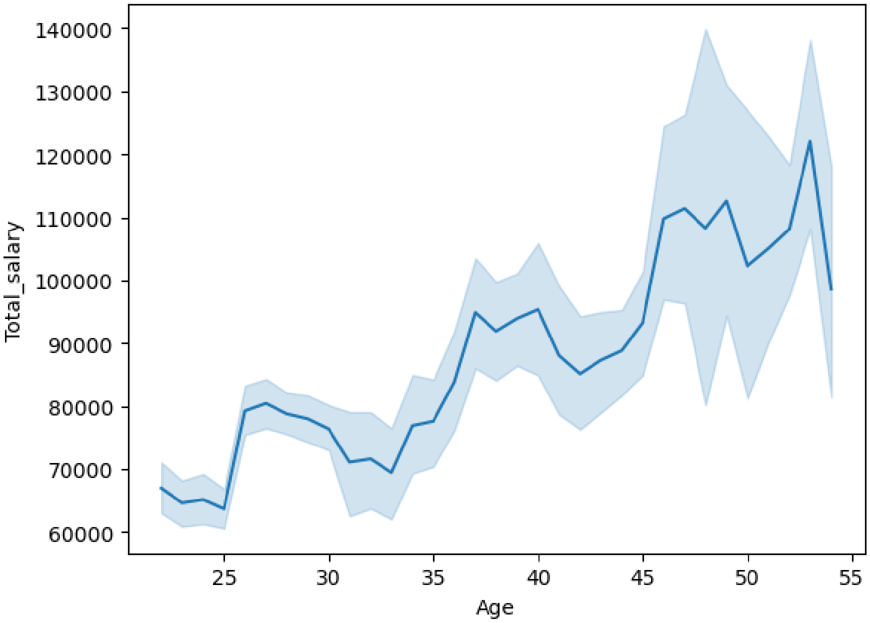
From this graph we can see that the price of the car which is directly depending on which type of car they buy, does not depend on the income of the person.



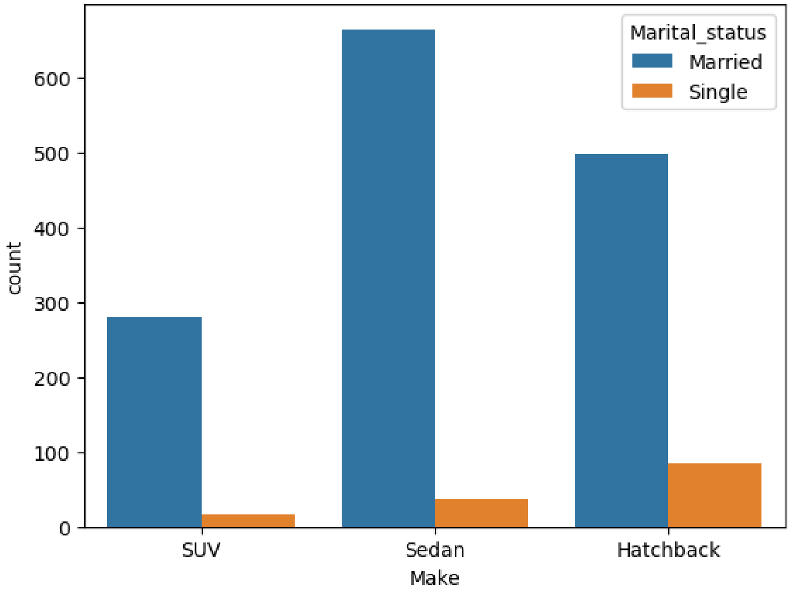
From this chart, we can see the preference of the car type is depending on the number of dependents.



From the above boxplot, we can see that the buyer who has a partner has a greater total salary and the buyer whose partner is not working has comparatively lesser salary.



This line graph shows how the total salary has an uneven growth as the age increases.



This count plot shows the distribution of the types of cars which is directly proportional to the marital status, and we can analyze that most of the car buyers are married and prefer sedan and hatchback.

Verifying if our analysis is right -

```
Married    1443
Single      138
Name: Marital_status, dtype: int64
```

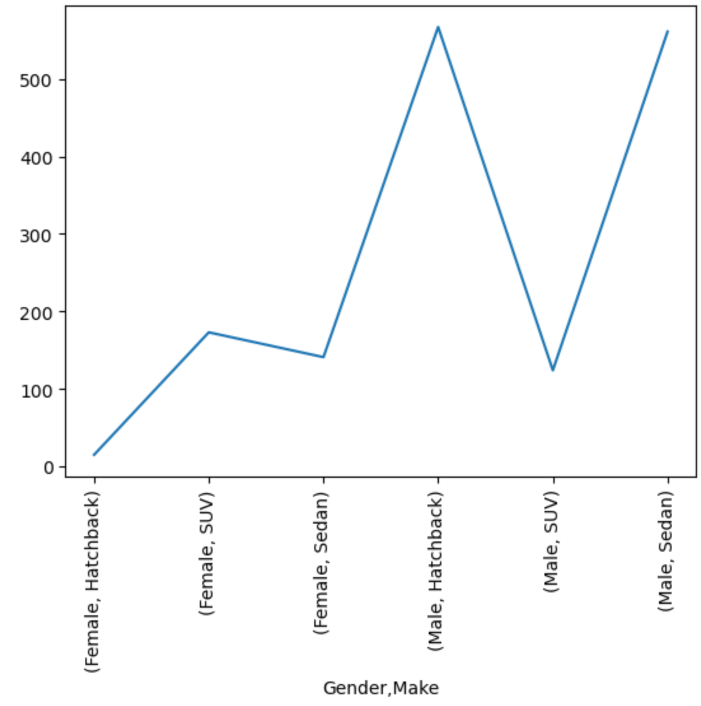
1E. Employees working on the existing marketing campaign have made the following remarks. Based on the data and your analysis state whether you agree or disagree with their observations. Justify your answer based on the data available.

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

When the analysis was done, we have seen that there are a greater number of females who use SUV than the number of males. Below details are the proof to prove Steve Roger's words wrong. Males prefer Hatchback more than any other type of car. So, I disagree with Steve Roger.

```
Gender      Make
Female      Hatchback     15
            SUV           173
            Sedan         141
Male        Hatchback     567
            SUV           124
            Sedan         561
dtype: int64
```

To support the analysis, below is the attached graph to visualize the words of Steve Roger -



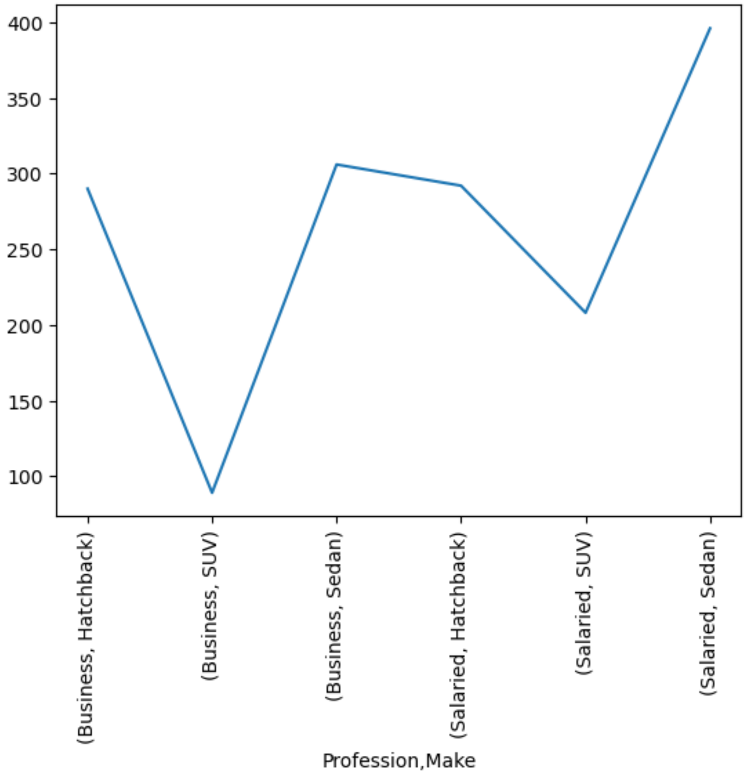
E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.

After a methodological analysis, we can conclude that Ned Stark was correct, and I completely agree with him. From the operations, we can see that there are more number of salaried person who is more likely to buy a sedan than a hatchback or a SUV.

Below is the output of the operations performed to check if Ned Stark was right.

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

To support the above analysis, below is the attached graph to verify the words of Ned Stark -

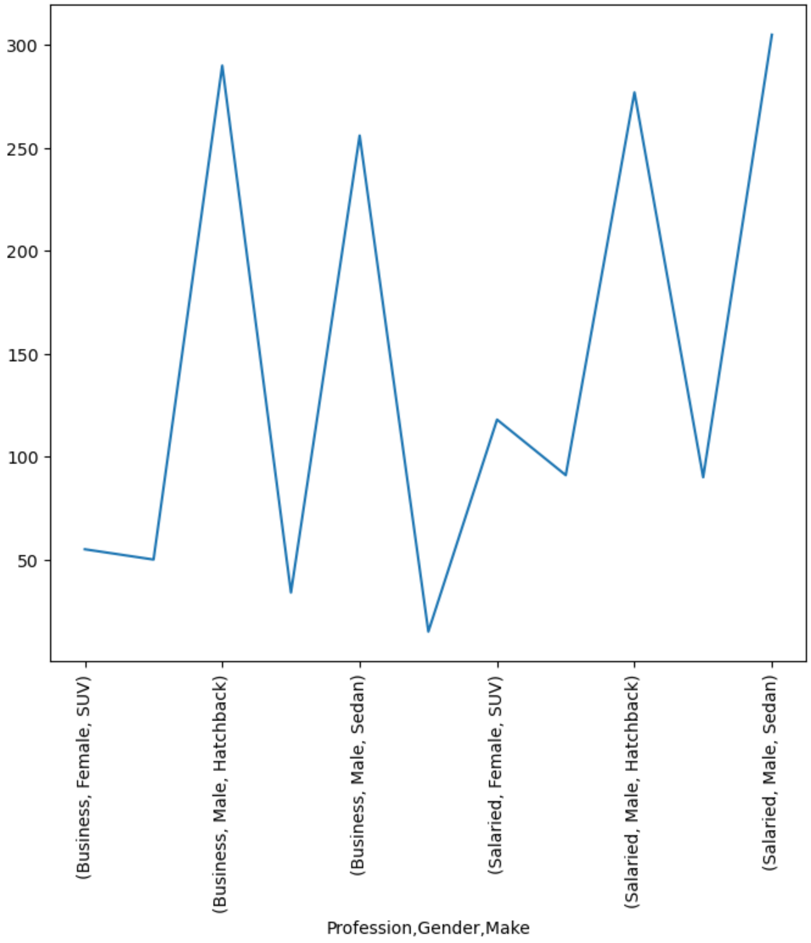


E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale.

After taking a deep look into what Sheldon Cooper said, we can see that a salaried male is not an easier target for a SUV sale over a Sedan Sale. Salaried males prefer Sedan over the other 2 car types. And when it specifically comes to Salaried male, SUV is the least preferred. So, I can confidently say that Sheldon Cooper is wrong, and I disagree with him. The below output is the result of our analysis.

```
Profession    Gender      Make
  Business    Female       SUV          55
                           Sedan        50
              Male         Hatchback    290
                           SUV          34
                           Sedan        256
  Salaried    Female       Hatchback    15
                           SUV          118
                           Sedan         91
              Male         Hatchback    277
                           SUV           90
                           Sedan         305
  dtype: int64
```

To verify the above conclusion, below is a line graph which clearly shows that the salaried men prefer Sedan over hatchback and SUV.



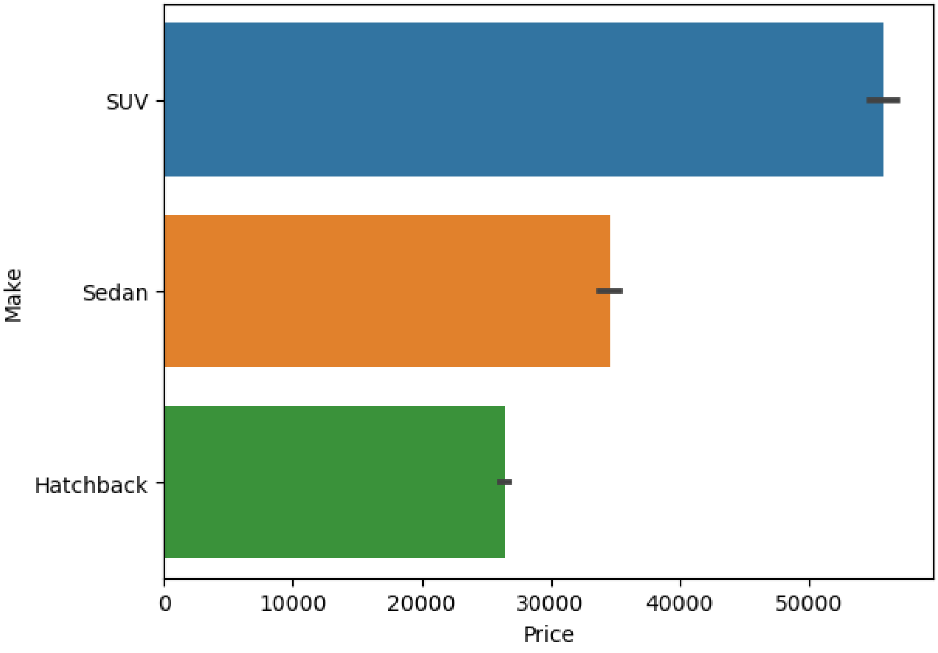
1F. From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions. Give justification along with presenting metrics/charts used for arriving at the conclusions. ***F1) Gender ***F2) Personal_loan

The most amount spent on cars was for SUVs when compared to hatchback and sedan.

The business can use these analytical conclusions to increase their reach towards their targeted customers. They can strategize a campaign which creates 3 sub-campaigns which targets all the 3 types of potential customers for the 3 different types of cars.

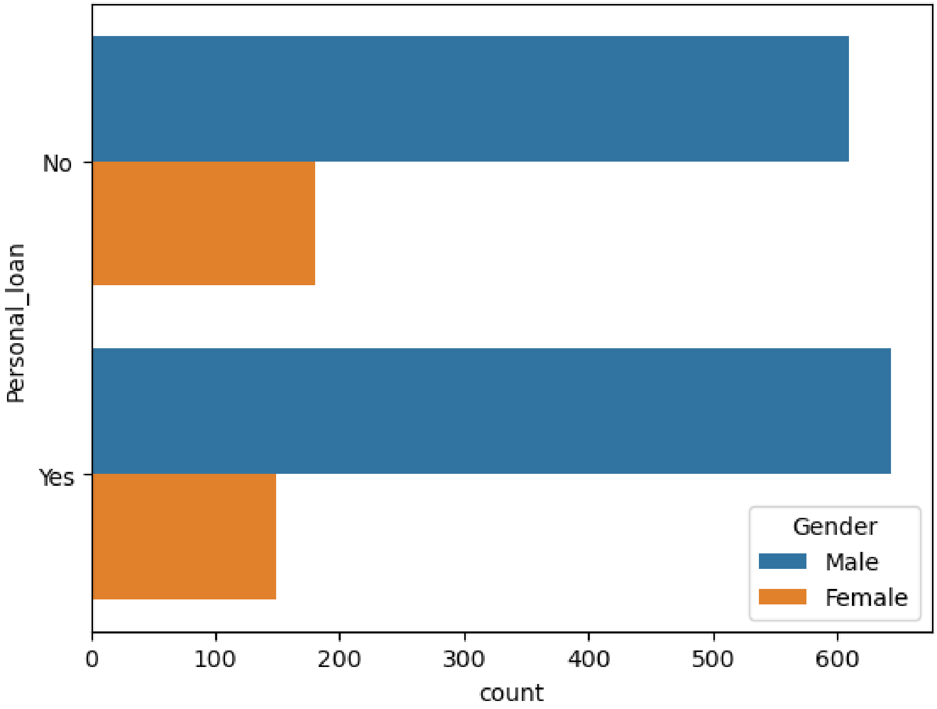
Here, we see the total cars purchased in each type.

```
Make
Hatchback     582
SUV            297
Sedan          702
```



From the above graph we can understand that, even though the number of SUVs are lesser than Sedan and Hatchback, it has outstood the others when the sale amount is calculated. From this we can derive that, a single SUV is expensive when compared to the others.

1F2. Analyzing with personal loan -



This graph helps to analyze the demographics when gender is taken into consideration while checking if they buyers had chosen to take a personal loan. So, here we can see that most male buyers have almost equal distribution who have taken loan to who haven't.

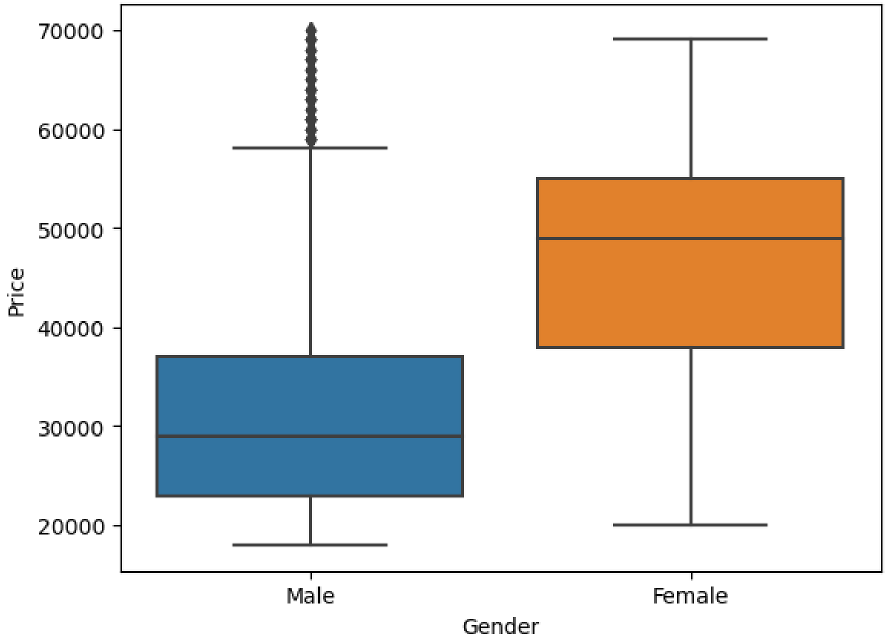
When it comes to female buyers, there are few more buyers who have not taken any personal loan when compared to female buyers who have taken a personal loan.

1F2. Analyzing with gender -

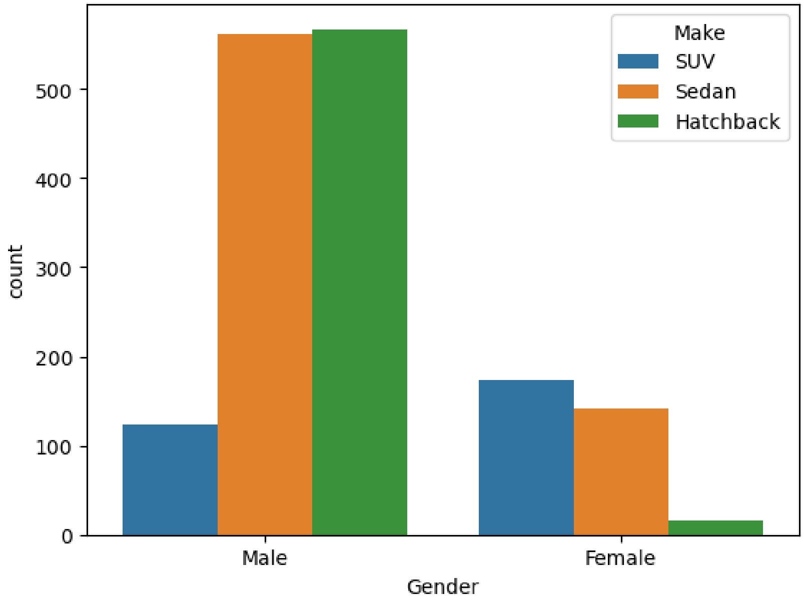
From a marketing campaign perspective, I think the company should focus more on the female audience because, as of now there are very smaller number of female buyers to male buyers where the ratio roughly stands at 1:4.

Male	1252
Female	329

To support the above statement, below is boxplot which gives us a clearer understanding.



From this boxplot, we can see that women have female buyers have higher purchasing potential than the male buyers. So, firstly the company’s marketing team should focus on their outreach on the female audience who have a higher purchasing potential than the male buyers.

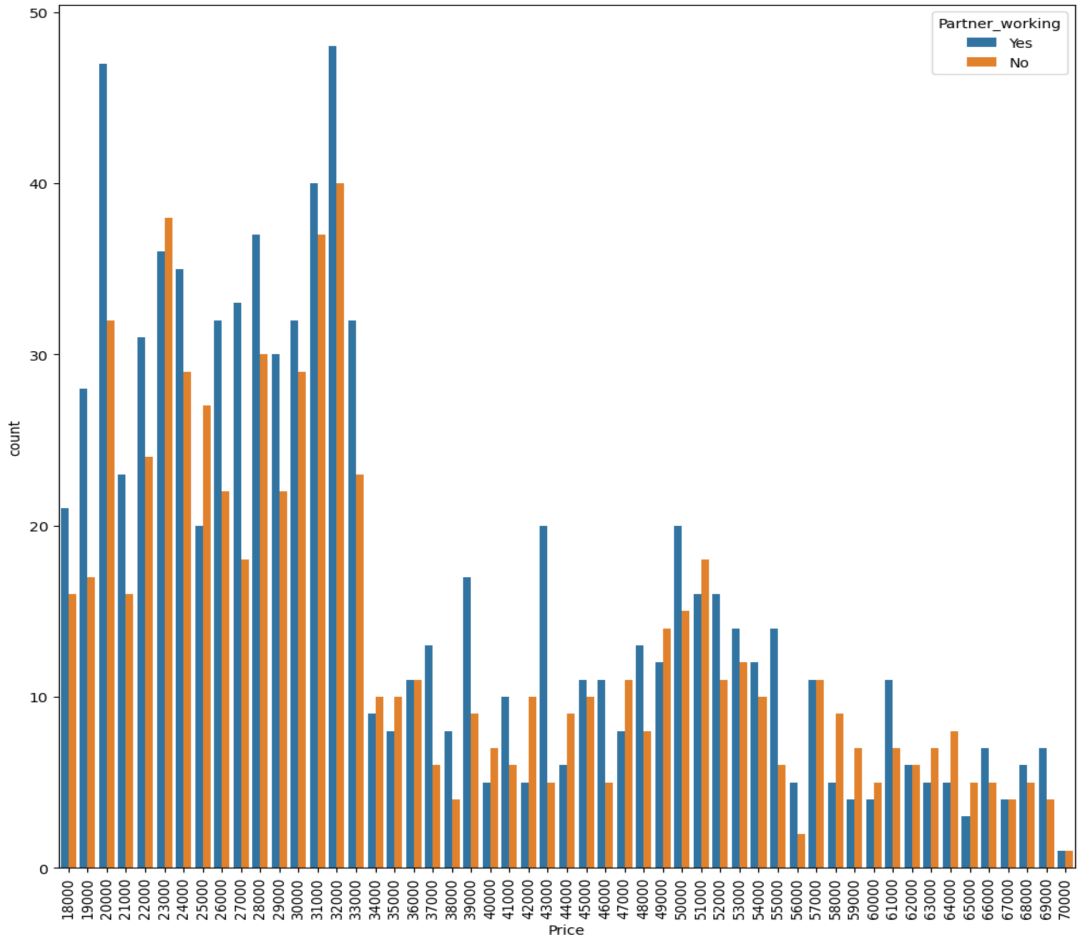


Secondly, the marketing team should also try pushing the SUVs for male audience as SUV is the most profitable type even though the number of cars sold were less.

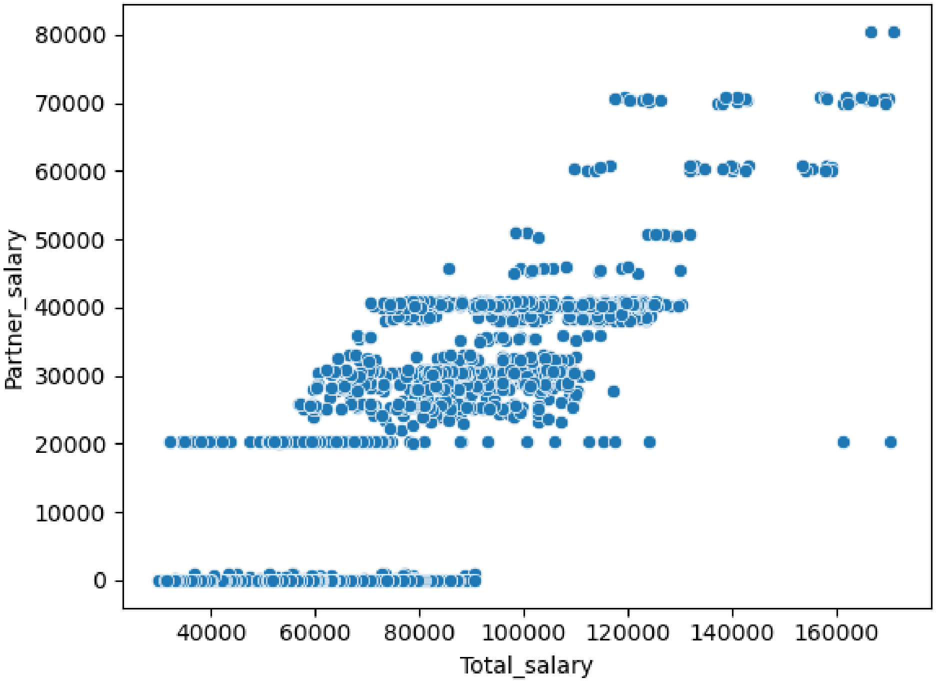
Thirdly, though the female audience is being targeted, SUV and Sedan buyers should be reached specifically.

And finally, and most importantly, hatchback should be considered sensitive for female audience as they have very less sales and cannot let it drop down to negative.

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

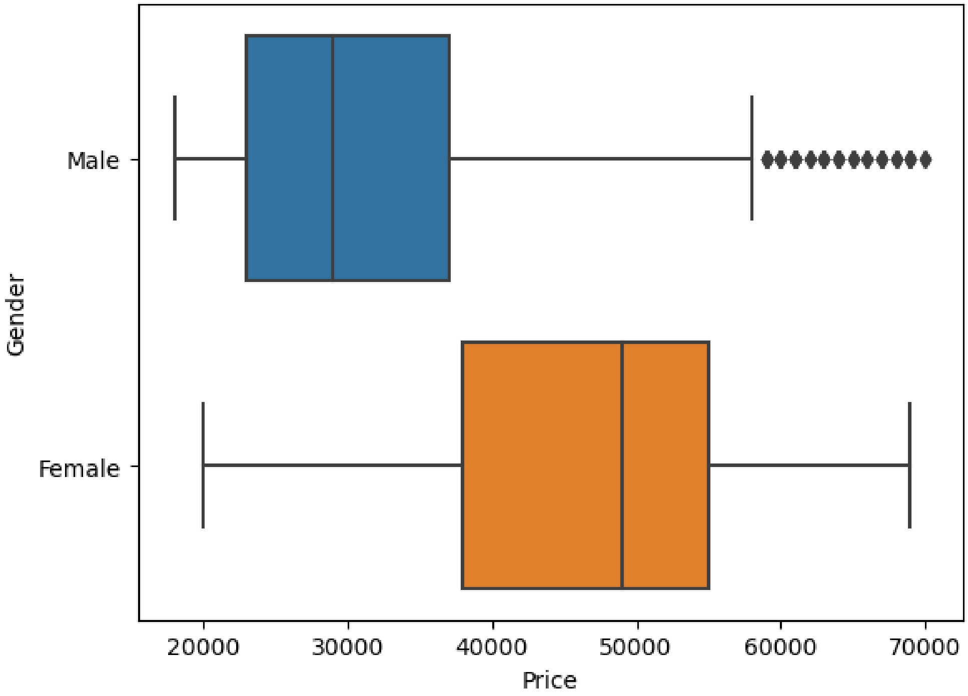


From the above chart, we can agree that having a working partner leads to the purchase of a higher-priced car.



In addition, this graph shows there is a positive correlation where the partner’s salary is adding to the total salary of the buyers’ which directly increases the budget of the buyer.

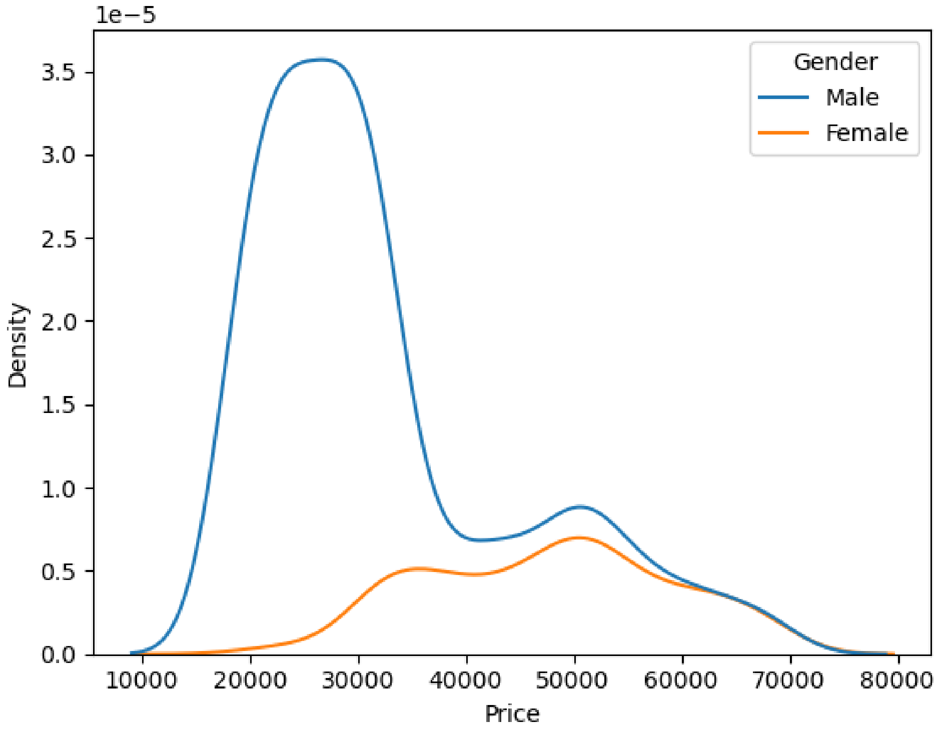
1H. The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use the Gender and Marital_status - fields to arrive at groups with similar purchase history.



The above visualization represents the price spent on purchasing based on **Gender**.

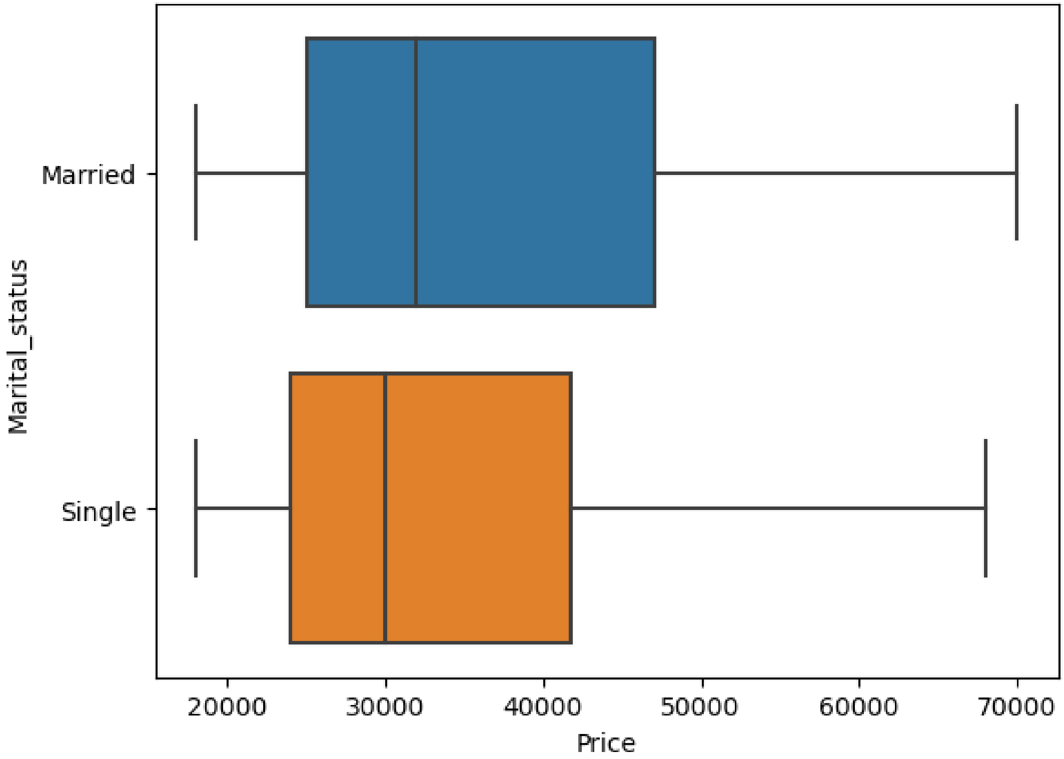
From the above plot, we can see most of the female buyers lie in between 40000 and 55000, so we can group these buyers as one, who spend not greater amount nor are spending very less. Similarly, we can group the female buyers with spending of 20000 to 40000 as less budgeted buyers and 55000 to 70000 as higher budgeted buyers.

On the other hand, we can see that mal buyers have comparatively lesser potential in spending. The male buyers can be grouped according to similar purchasing history as less spending group from 20000 to 25000, average spending buyers from 25000 to almost 40000 and finally higher budgeted buyers from 40000 to 60000.



Male	1252
Female	329

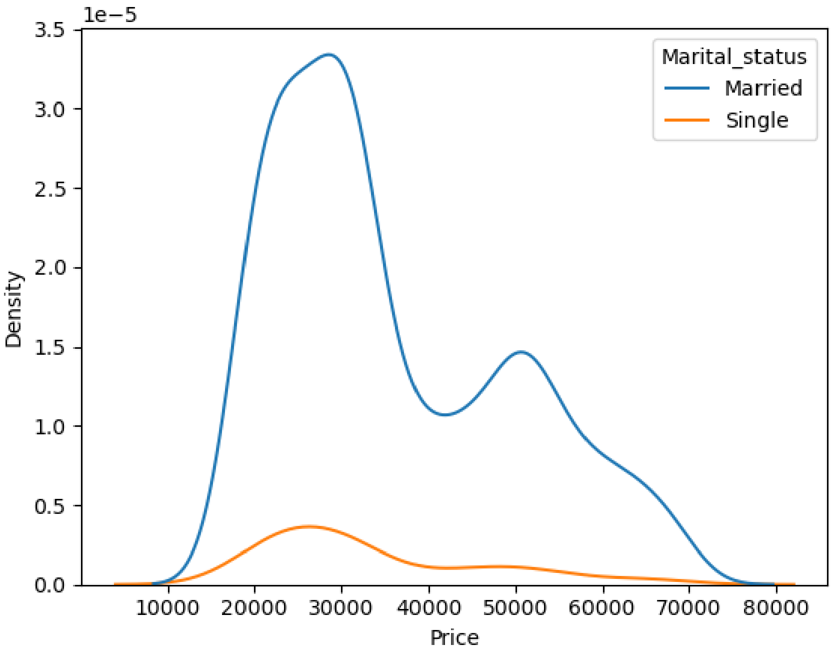
Through this density plot, we can cross verify our grouped buyers based on gender.



The above visualization represents the price spent on purchasing based on **Marital Status**.

From this, we can see that married people can be grouped in to 3 categories. The ones who spent less than 25000, the ones who expenditure lies in between 25000 to 50000 and the buyers who spend above 50000.

Similarly, we can see that for buyers who are single, the 3 categories are ones who have spent less than 25000, the ones who lie in between 25000 to around 45000, and finally the ones who spend more the 45000.

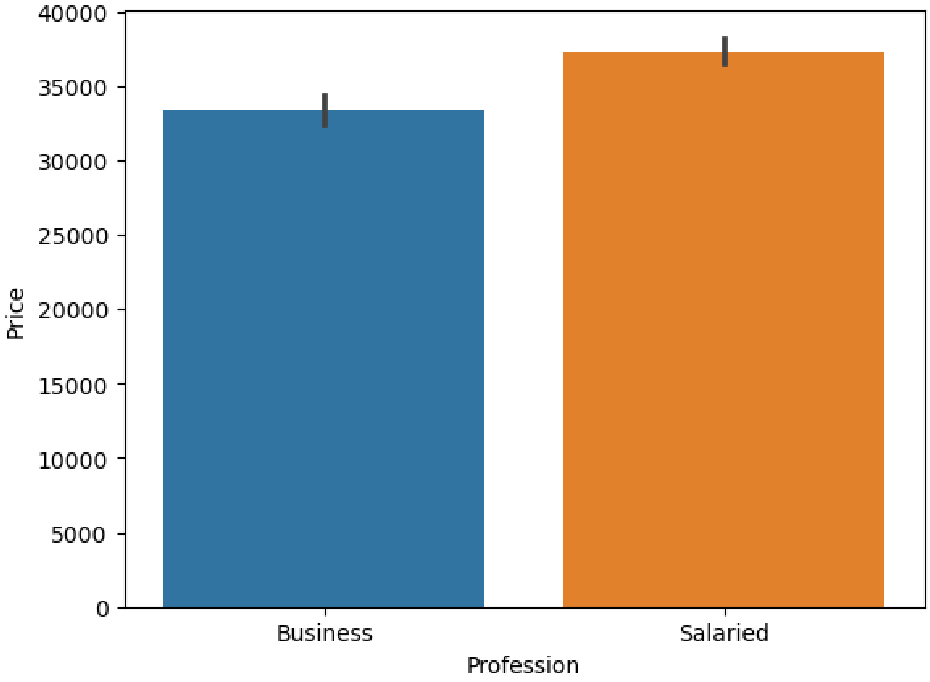


Married	1443
Single	138

Through this density plot, we can cross verify our grouped buyers based on marital status.

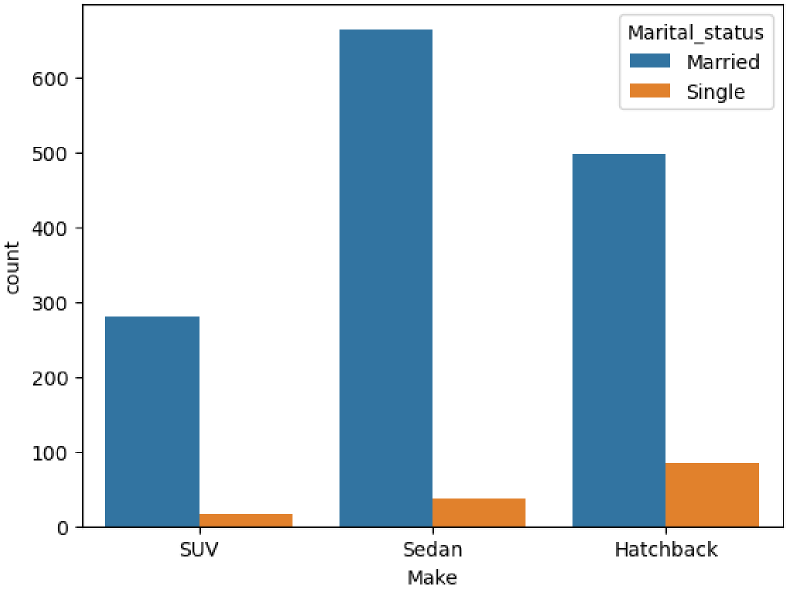
1I. Analyzing the dataset and listing down the top 5 important variables, along with the business justifications.

1. When profession is taken into consideration, the number of salaried employees are more than the number of people who own a business. Also, the number of male counts outstand female counts who own a business and are salaried.



Here, the salaried buyers out beat the business owned buyers. So, I would suggest the marketing team to focus comparatively more on the business owned buyers so that eventually the balanced ratio is achieved further to which they can come up with a rigorous upselling strategy.

2. The marketing team should focus more on reaching out to the targeted customers for SUV, specifically male buyers who have not been preferring SUVs. This can be marketed in a very strategic way and will help the company to increase the revenue as the price of one SUV is higher than a hatchback and sedan according to the previous performed analysis and continue pushing the sales for hatchback and sedan.



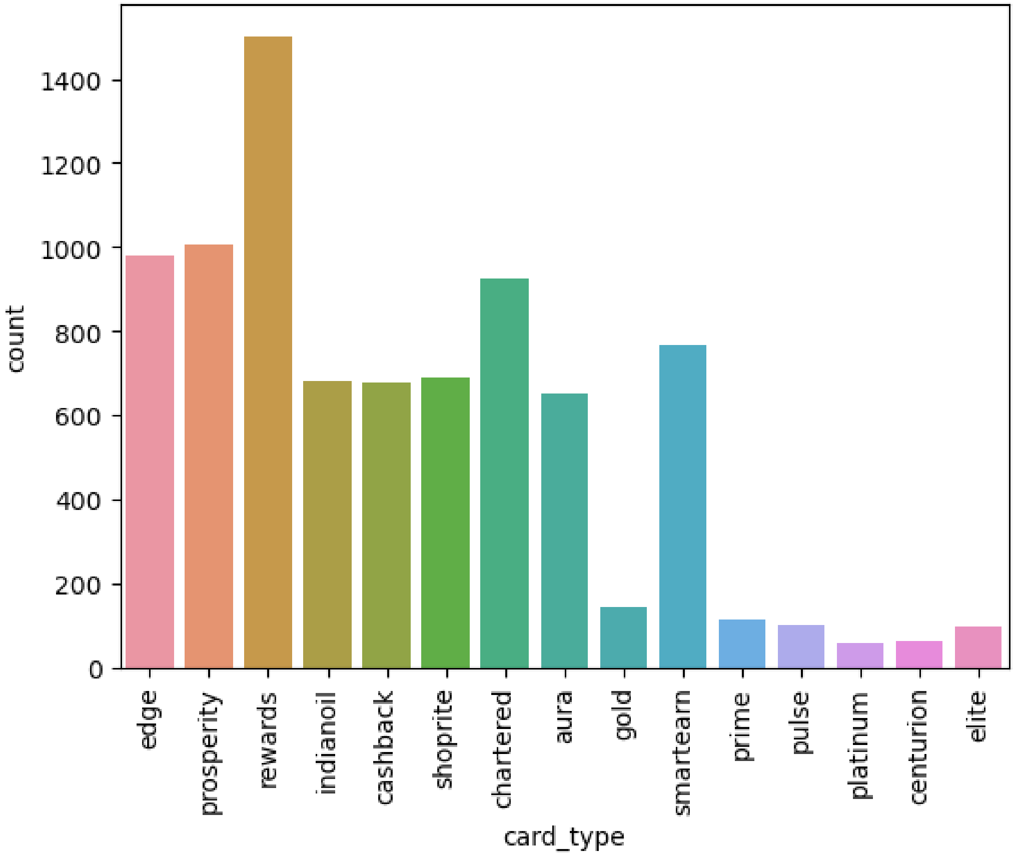
3. The female buyers should be the marketing team's next targeted customers as the ratio of male to female buyers is 4:1. So to balance this and increase the revenue the team should focus more on selling hatchback cars to females as it is the least preferred and should push the existing sales of SUV and Sedan.

4. The marketing team should target the youth as the number of buyers here are higher as well as target the buyers above 40 whose purchase amount is higher. So, if these 2 aspects are balanced, it will lead to a higher revenue generation.

5. There must be a holistic marketing strategy to target the buyers who are single as married buyers are almost 6 times higher. So, if this the single buyers are increased to at least 6 times more than now, then the revenue will easily increase.

2. **Analyzing the dataset and listing down the top 5 important variables, along with the business justifications. (GODIGT Bank)**

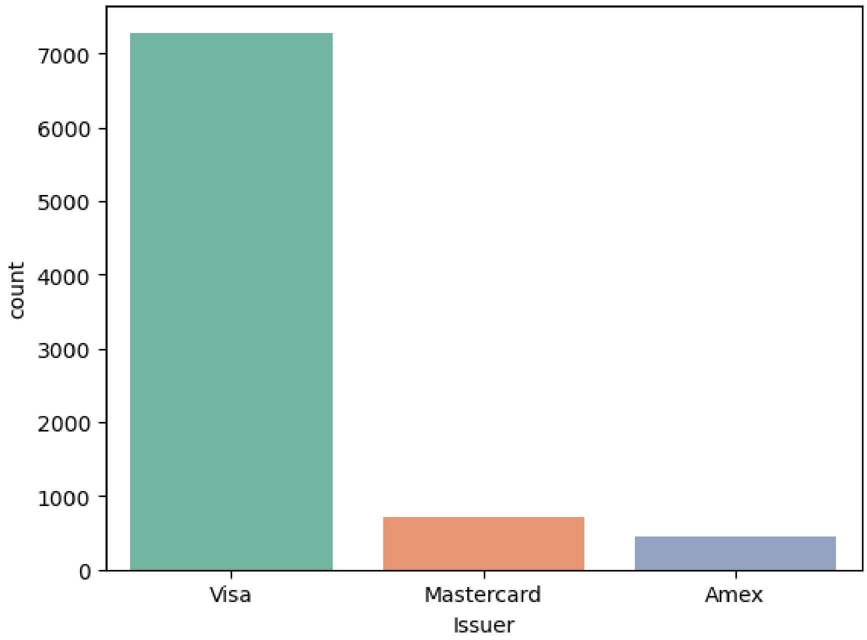
1.



The bank has 8448 users and has 15 different card types issued to its users. Here, we can see that 'rewards' is the type of card, which is used the most, followed by prosperity and edge.

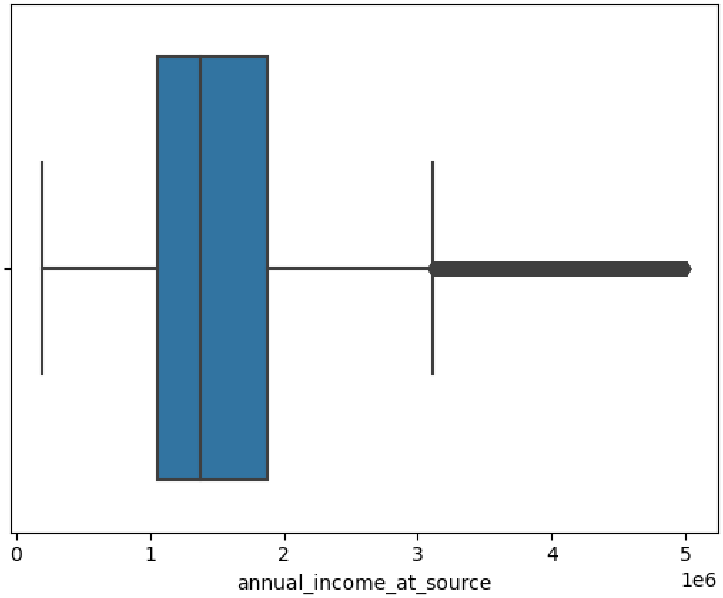
Premium and centurion are the 2 least used type of cards.

2.



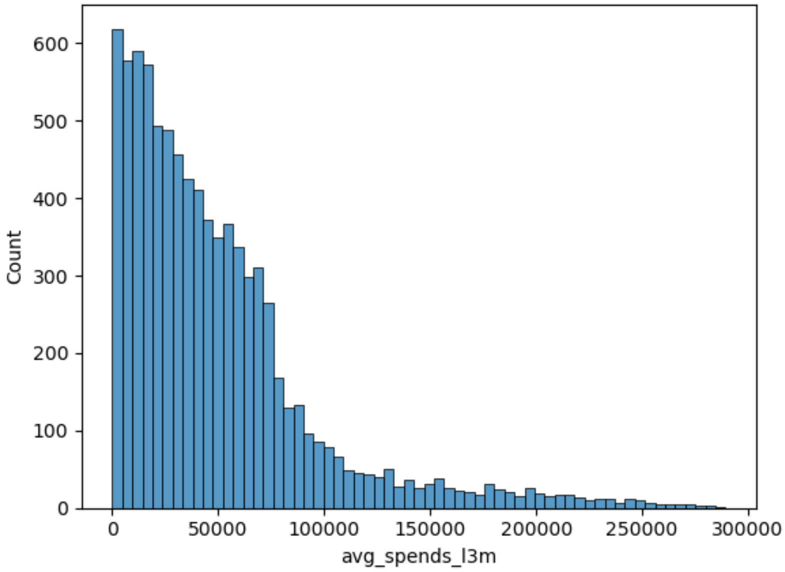
Here, when we take issuer as the base for our analysis, we can see that most of the cards has been issued by vise to the users. It is approximately 6 times more than the MasterCard and Amex users.

3.



Here, most of the annual income lie in-between 1000000-2000000.

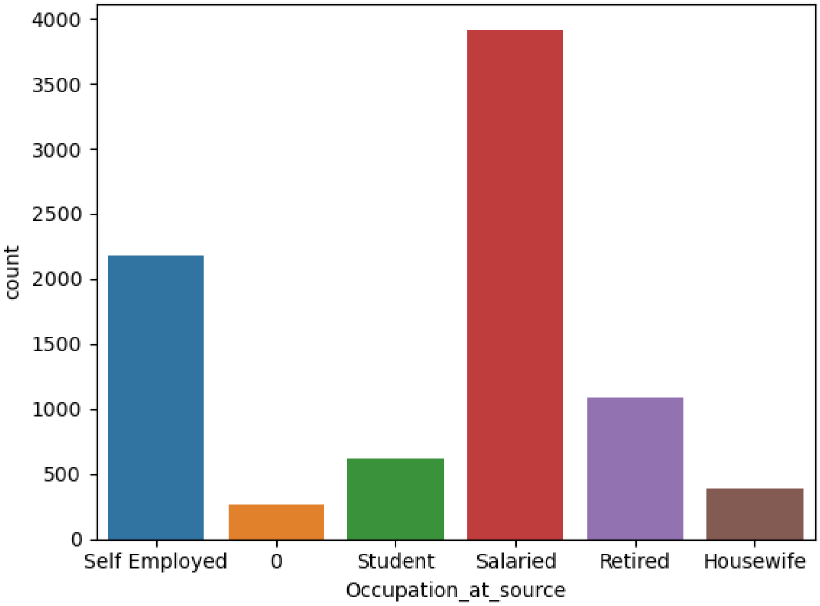
4.



Here, most of the average spending lies in between 0 to 80000.

We can see a dip in the expenditure as it gets expensive.

5.



Here, we can see that most salaried users are the highest, and housewives are the lowest.

Business justifications -

1. Finally, the bank should focus more on housewives and students to increase their revenue.
2. They can come up with strategies and loyalty-points system for the people who spend higher as the higher spending users are very minimum.
3. The business should reach up to higher incomed users to make solid profits.
4. Mastercard and Amex should target more potential customers and come up with a holistic sales approach.
5. Premium and centurion cards should come up with strategized benefits to increase their users.

THANK YOU