Austo Motor Company
Data Analysis

Team:
Apoorv P
Jitendra S
Kashyap
Preyal C
Rathi S
Sharath A
PGP-AIML June’23
20/10/2023

Table of Contents

Executive Summary

Austo Motor Company is a leading car manufacturer specializing in SUV, Sedan, and Hatchback models. The dataset contains historical purchase data and consists of various details of the customers, who have purchased a car from Austo Motor Company. Based on the different attributes/characteristics of the customers, we need to explore the different attributes of the customer and see if they would buy a car after the campaign is launched.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. Analyze the different attributes of the customer which can help in analyzing the customer behavior and come up with insights to improve the marketing campaign. This exercise should help the student in exploring the summary statistics, contingency tables, and different plots used in EDA and how they can be used to make inferences about the data.

Data Description

1. Age: discrete - ranges from 22 to 54
2. Gender: categorical - male, female
3. Profession: categorical - Business, Salaried
4. Marital_status: categorical - Married, Single
5. Education: categorical - Graduate, Post Graduate
6. No_of_Dependents: discrete - ranges from 0 to 4
7. Personal_loan: categorical - Yes, No
8. House_loan: categorical - Yes, No
9. Partner_working: categorical - Yes, No
10. Salary: discrete - ranges from 30000 to 99300
11. Partner_salary - continuous ranges from 0 to 80500
12. Total_salary: discrete - ranges from 30000 to 171000
13. Price: discrete - ranges from 18000 to 70000
14. Make: categorical - Hatchback, Sedan, SUV

Sample of the Dataset

	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

Fig-1 sample_dataset

Dataset has 14 variables, which describe the various customer attributes.

Exploratory Data Analysis

Let us check the types of variables in the dataframe.

```
 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
```

There are a total 1581 rows and 14 columns in the dataset. Out of 14, 8 columns are of object type and rest 6 are of either integer or float data type.

Check for missing values in the dataset.

```
Age                 0
Gender             53
Profession          0
Marital_status      0
Education           0
```

No_of_Dependents	0
Personal_loan	0
House_loan	0
Partner_working	0
Salary	0
Partner_salary	106
Total_salary	0
Price	0
Make	0

From the above results we can see that there are missing values present in the dataset for the columns namely - ‘Gender’ and ‘Partner_working’:-

- 53 rows does not have a value for the field - ‘Gender’.
- 106 rows does not have a value for the field - ‘Partner_salary’.

To treat the missing values we will proceed with the following:-

- For ‘Gender’, we will correct it using the mode value from the Gender column that is ‘Male’.
- For ‘Partner_salary’, there is another column Total_salary which is a sum of Salary and Partner_salary. We can subtract Salary from it to get Partner_salary wherever it is missing.

Univariate Analysis

We will proceed with Univariate analysis of the features of the dataset starting with Numerical ones. We will look at their descriptive statistical summary as well as Boxplots and Histograms to identify the pattern of each variable - how much is the spread, are there any outliers, etc. We will also infer some insights from the plots.

	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1581.0	19233.776091	19670.391171	0.0	0.0	25100.0	38100.0	80500.0
Total_salary	1581.0	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	1581.0	35597.722960	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0

Fig-2 stastical_summary_of_numerical_variables

Insights from Univariate Analysis of Numerical variables

- No variable is following a proper normal distribution
- 75% percent of customers are below the age of 40

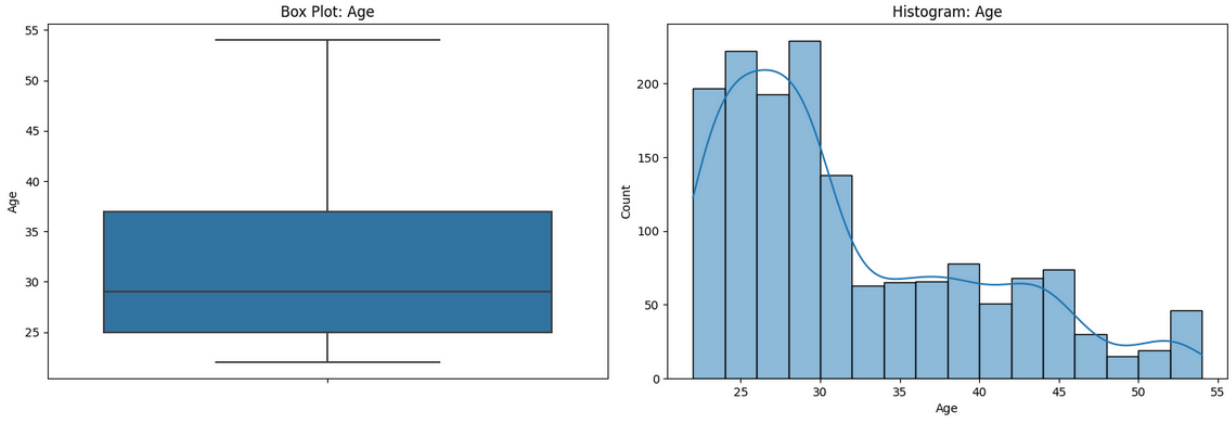

Fig-3 boxplot_histogram_of_Age

- No_of_Dependents is defined as a numerical variable, but it can pass for a categorical variable. The value of 0 is an outlier but that is valid as many of the customers don't have any dependents, which is alright.
- Salary of the customers ranges from 30k to almost 100k.
- Partner_salary is 0 for many customers indicating that their partner doesn't earn.

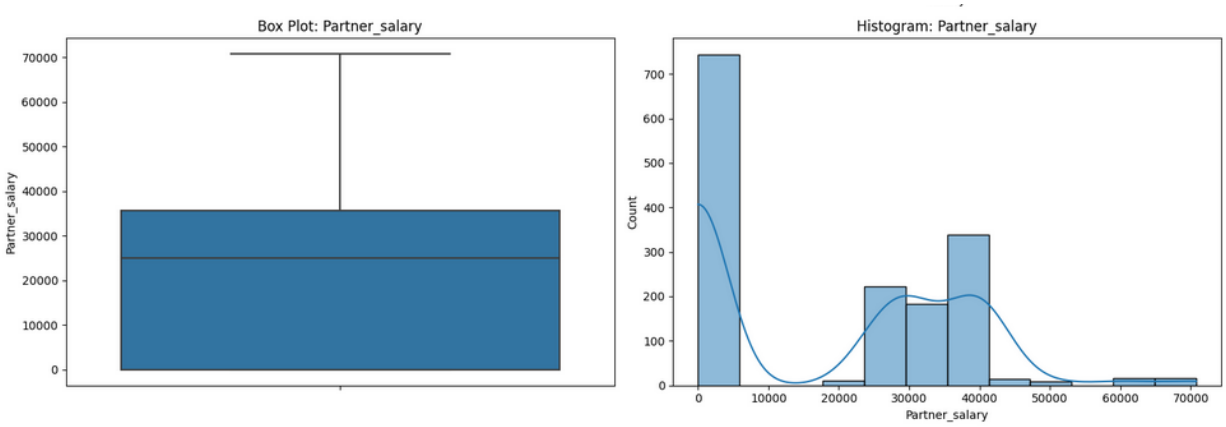

Fig-4 boxplot_histogram_of_Partner_salary

- Total_salary is a deduced variable from Salary and Partner_salary but that has outliers. We will be treating those outliers. To treat the outliers we will use the boxplot method.
    - Find IQR from the 5-point summary
    - Find Lower and Upper range using IQR formula: (**`LR = Q1 - 1.5xIQR, UR = Q3 + 1.5xIQR)`**
    - As the outliers are only 27, we can remove them from the dataset.
    - Post removing, there are no outliers in the dataset.

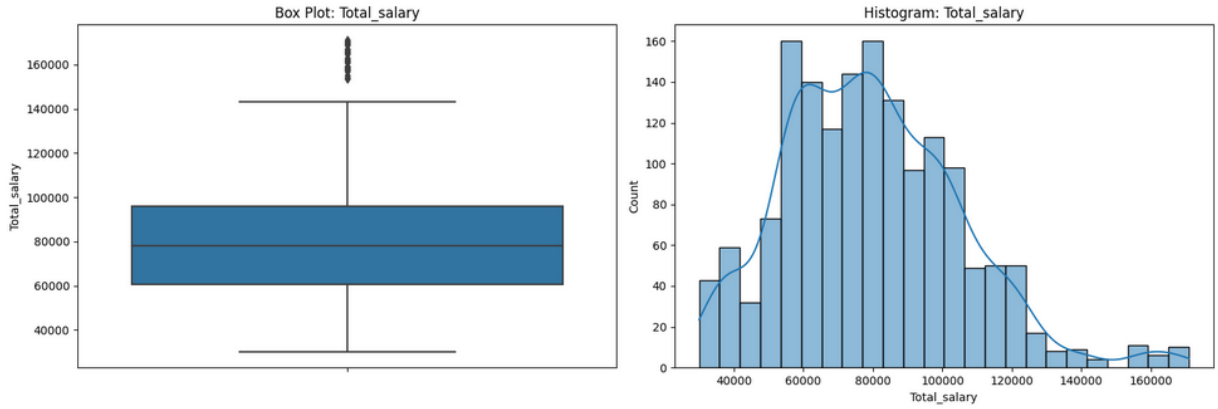

Fig-5 boxplot_histogram_of_Total_salary

- On an average, the price of a car sold by Austo Motors is nearly 36k, but ranges from 18k to 70k.

Insights from Univariate Analysis of Categorical variables

- Gender has an anomalous value of 'Femle' which we have to treat as 'Female'. Also, from the data the majority of the customers are Male.

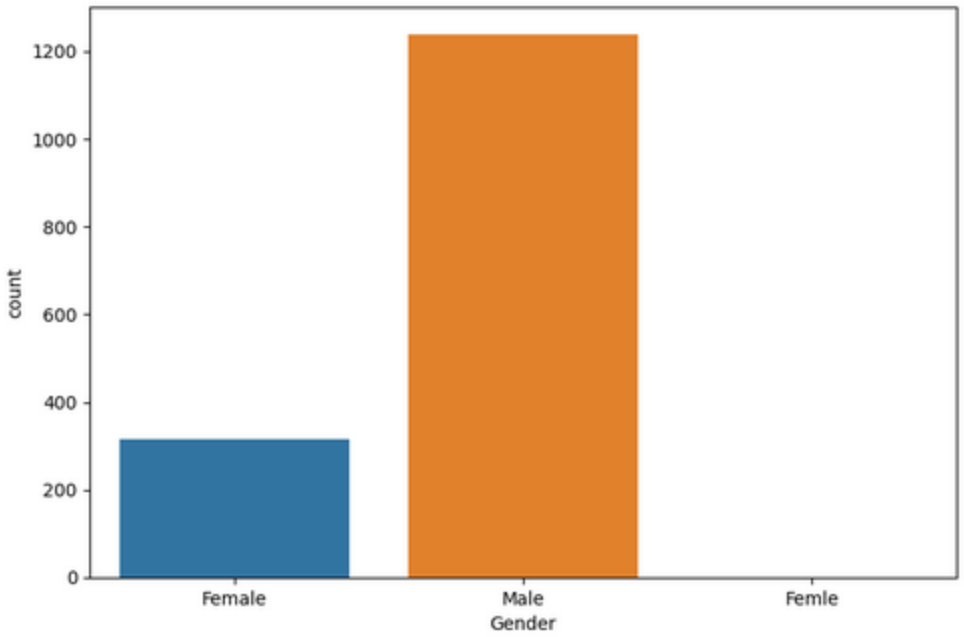

Fig-6 countplot_of_Gender

- The car buying quotient of Married people is much higher than that of Unmarried people.
- There is no impact of a Personal Loan on whether a customer can buy a car or not.

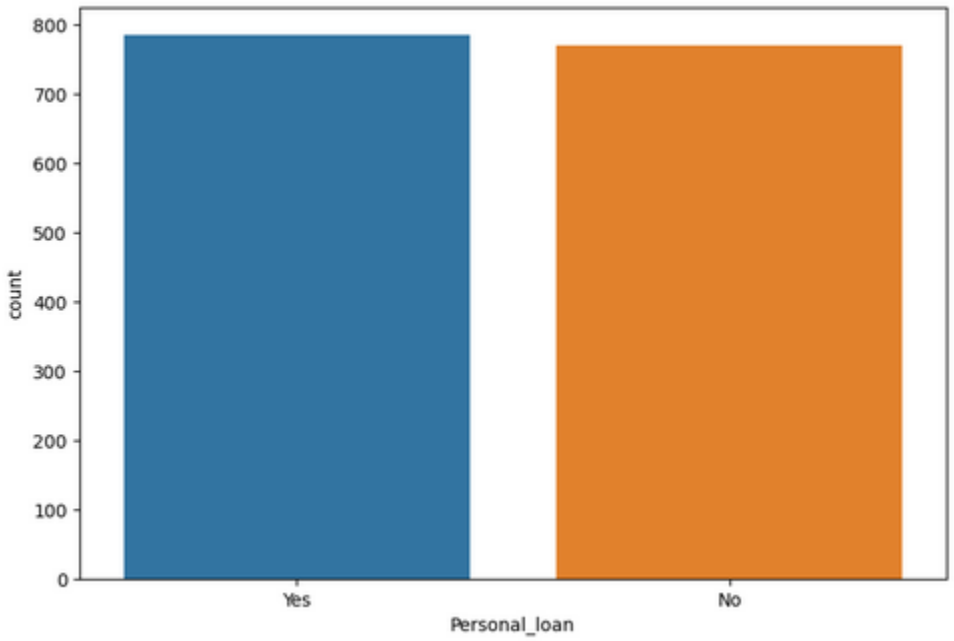

Fig-7 countplot_of_Personal_loan

- Sedan type cars are preferred much over SUV and Hatchback.

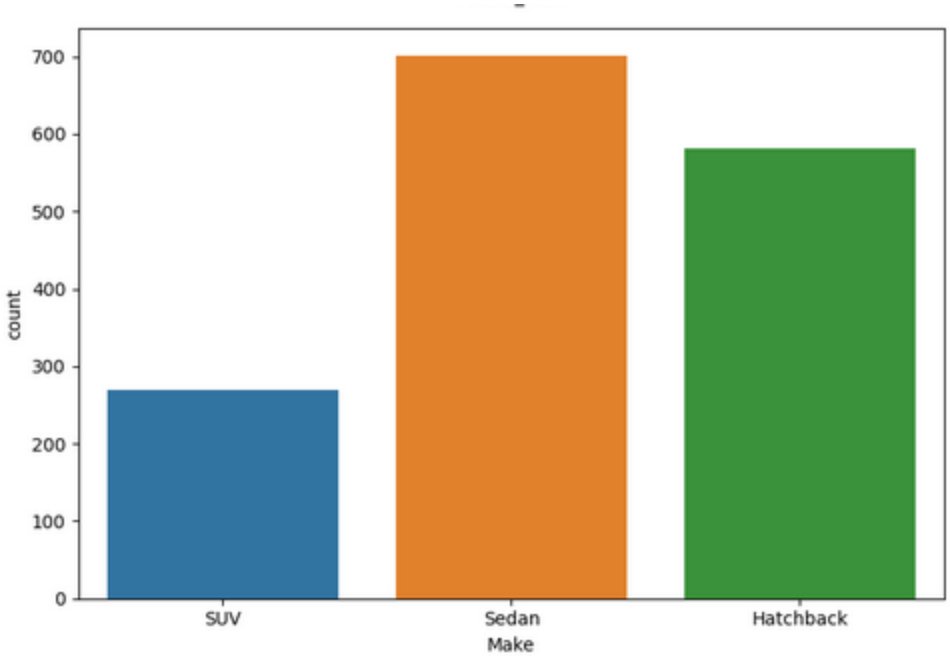

Fig-8 countplot_of_Make

- Having a house loan might impact the car buying capacity of an individual as a lot of investment goes in house loans already. So, people who don't have a house loan purchase a car is more likely.

Bivariate Analysis

Bivariate analysis is done to find correlation between different variables. We will perform bivariate analysis on numerical versus numerical and categorical versus categorical variables, and find insights post analysis.

Numerical v/s Numerical

Fig-9 heatmap_of_correlation_between_numerical_variables

As per the heatmap of numerical variables, we get the following observations:-

- Age and Salary have a moderate positive correlation which tells us that increasing age salary also increases and so does purchasing power.
- But it seems like Age has a higher correlation with buying costly cars than the salary's correlation with price. This may indicate that as one grows older one desires a better car whether or not the salary is high.
- Salary and Partner_salary have high correlation with Total_salary which is not a relevant information as Total_salary is derived from them.

- As we observed in the categorical variable analysis, Partner_working does not play a significant role towards a higher cost car and that shows here as well. Partner_salary is very mildly correlated with Price of the car.

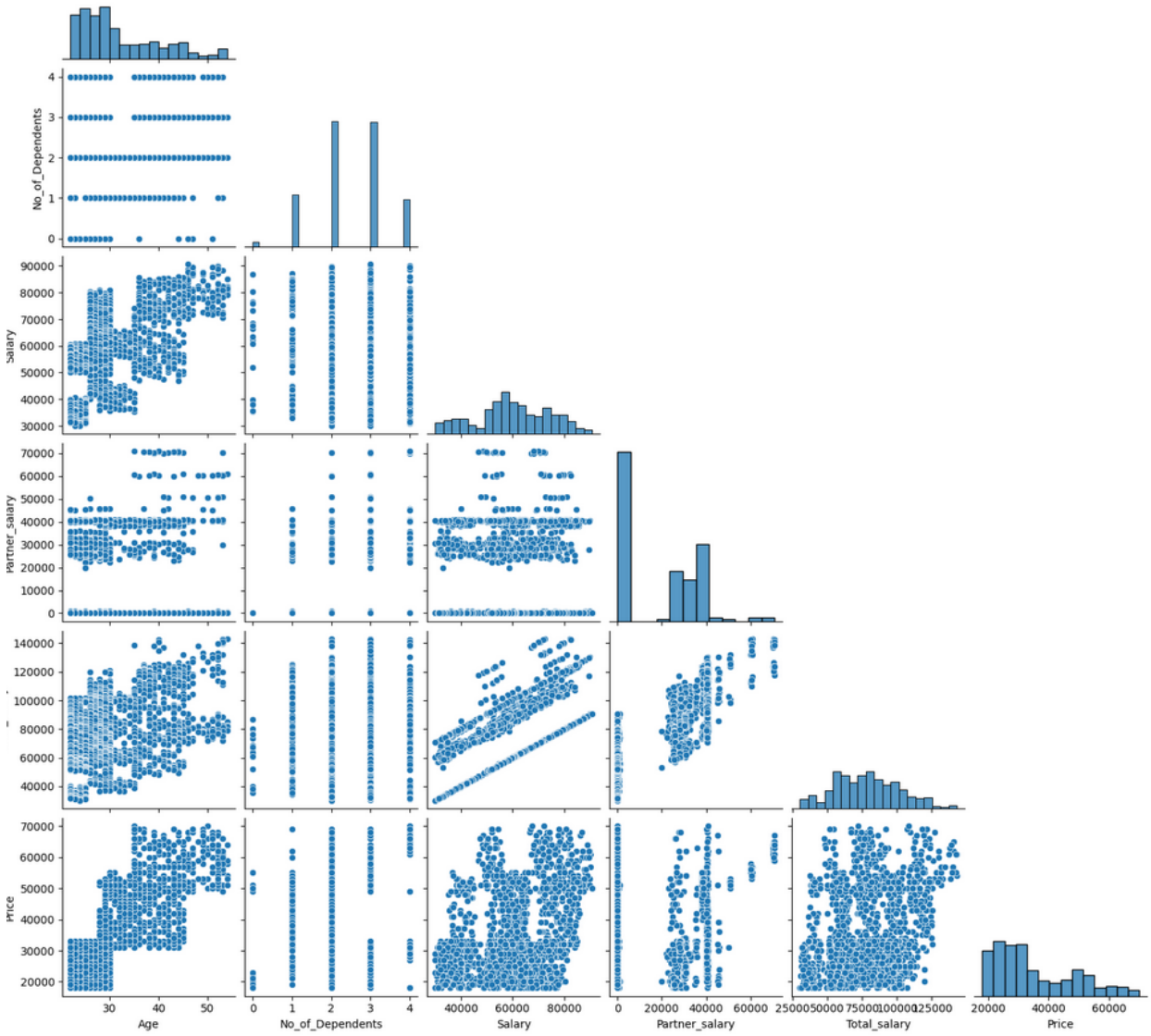

Fig-10 pairplot_for_numerical_variables

Above is the pairplot for the same. This is cornered at the bottom part. The diagonals have spread of the variable using histograms. All off-diagonal graphs are scatter plots between different variables.

Categorical v/s Categorical

- Salaried Women tend to purchase cars at almost double rate than Business Women.

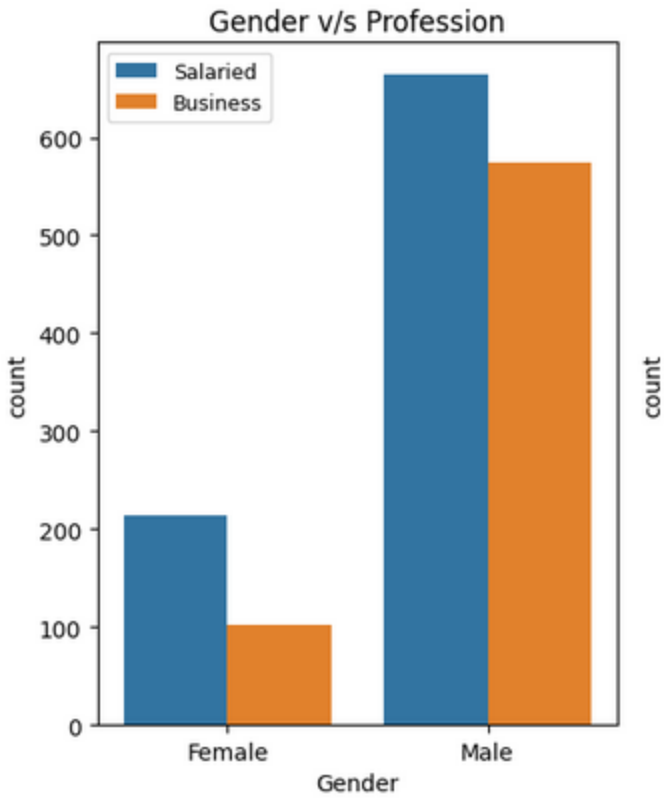

Fig-11 countplot_for_Gender_against_Profession

- Married customers prefer Sedan cars over any other type of cars.

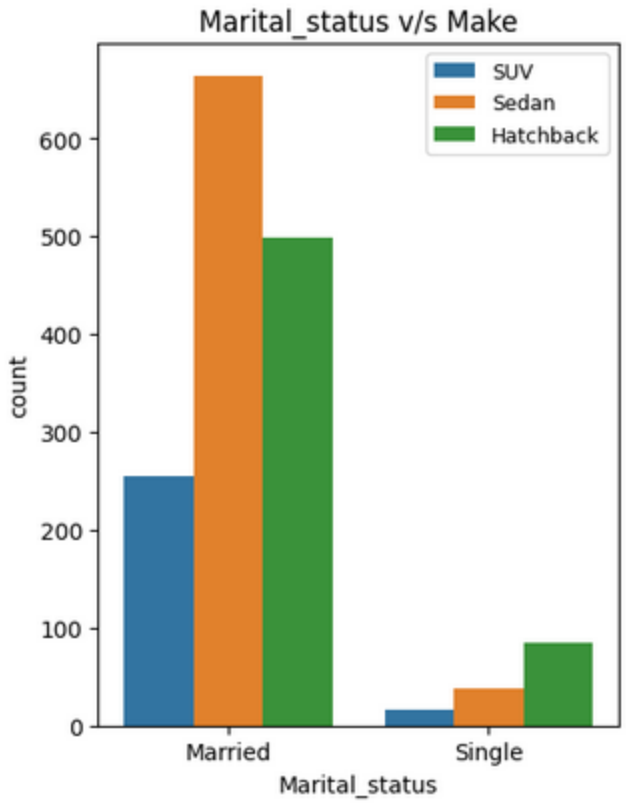

Fig-12 countplot_for_Marital_status_against_Make

- The SUV cars preference in customers who have House Loans is very less given the fact that House Loans demand quite a large investments and SUV cars are the costliest as shown in the barplot below between Make and Price.

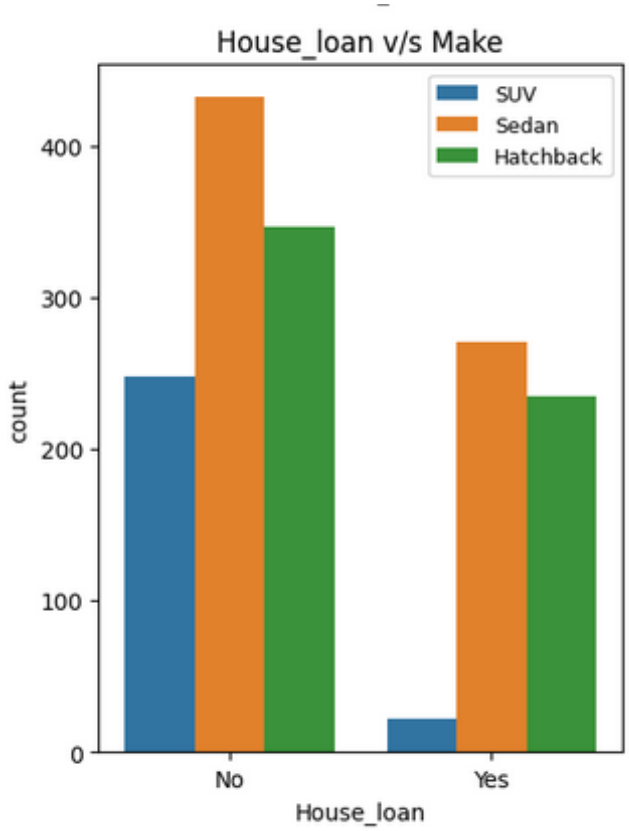

Fig-13 countplot_for_House_loan_against_Make

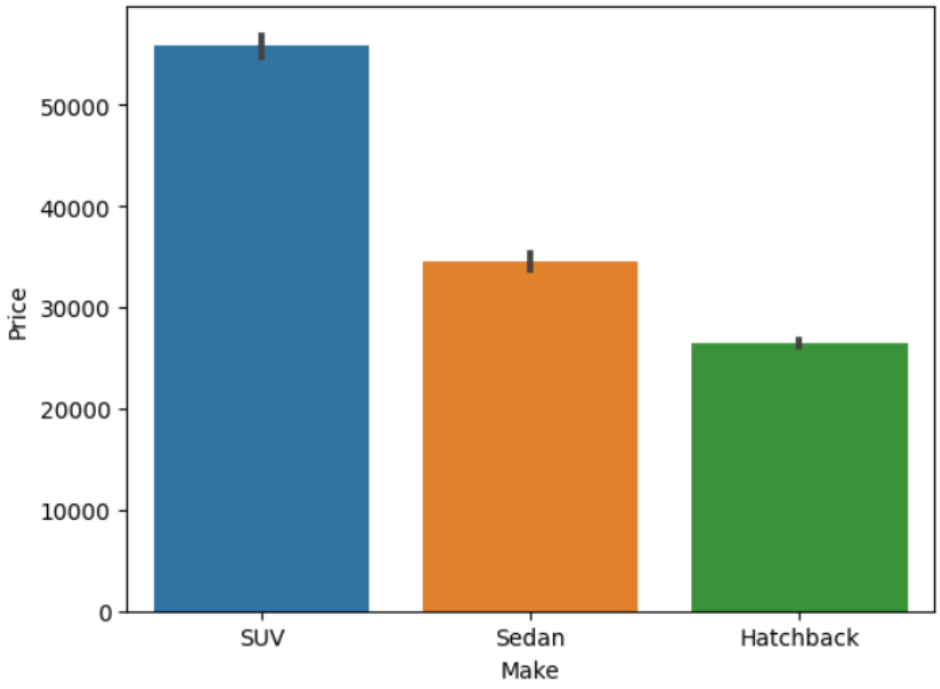

Fig-14 barplot_between_Make_and_Price

Analysis of Remarks made by Employees

1. Steve Roger believes that men prefer SUV by a large margin compared to the women

Make Gender	Hatchback	SUV	Sedan	All
Female	15	159	141	315
Male	567	111	561	1239
All	582	270	702	1554

Fig-15 crosstab_for_Make_and_Gender

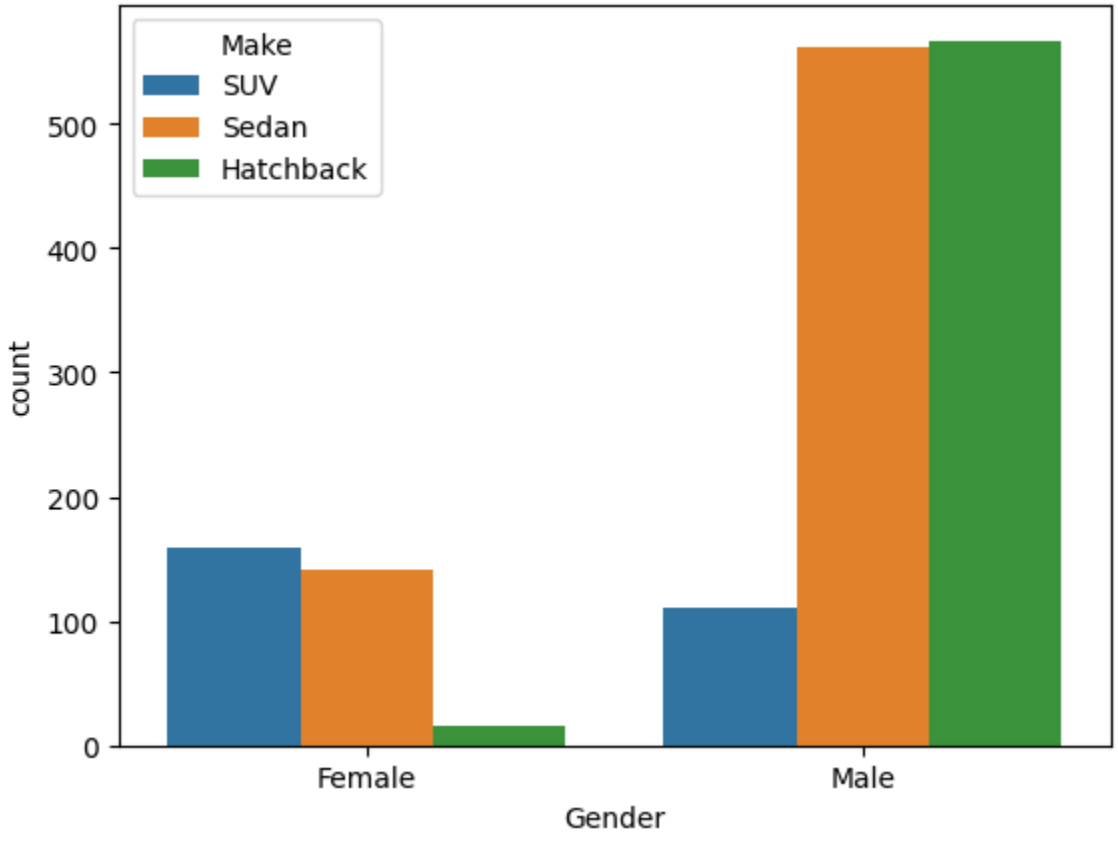

Fig-16 countplot_for_Make_and_Gender

We do not agree with this statement, as from the graph, we can see that Women prefer SUVs over Men and not the other way around. Out of total Female customers i.e. 315, 159 of them preferred SUV against 111 out of 1239 Men.

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

Make Profession	Hatchback	SUV	Sedan	All
Business	290	81	306	677
Salaried	292	189	396	877
All	582	270	702	1554

Fig-17 crosstab_for_Profession_and_make

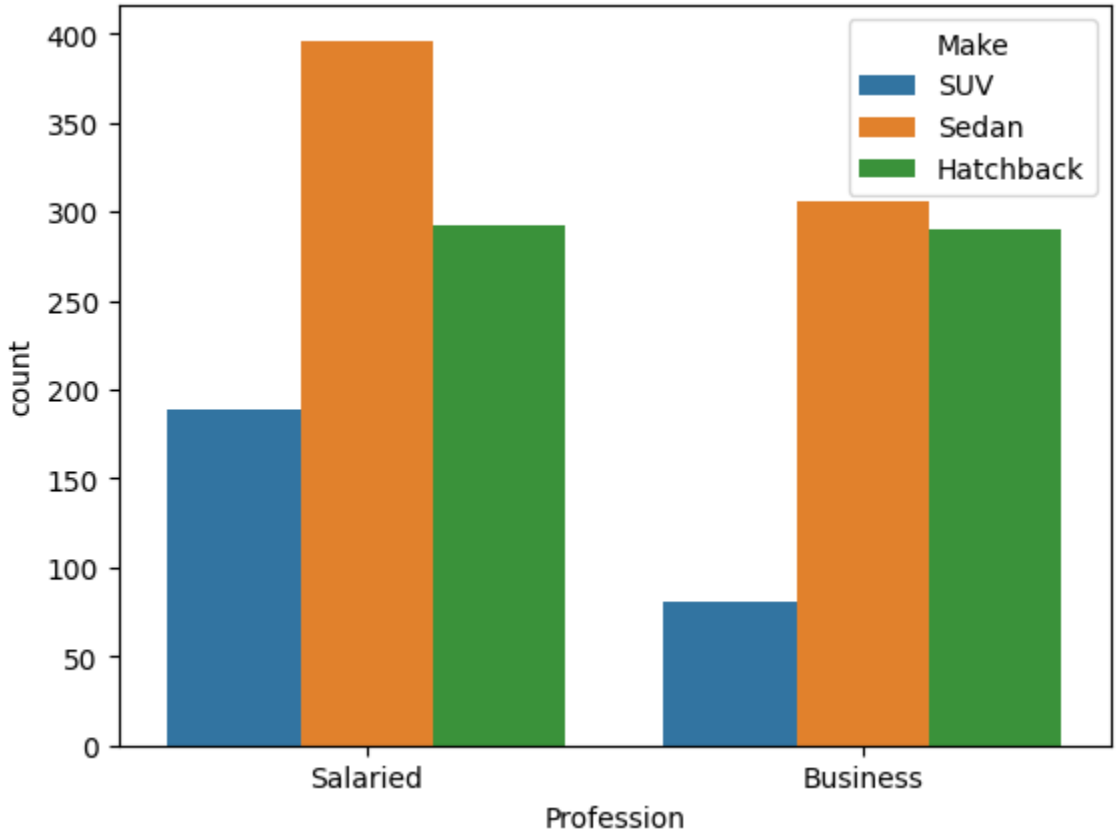

Fig-18 countplot_for_Make_and_Profession

We agree with the above statement, as we notice from the graph that a salaried person preferred Sedan over any other kind of car. As observed, Salaried customers bought 396 Sedan cars against 292 Hatchback and 189 SUV cars.

3. Sheldon Cooper claims that a salaried male is an easier target for a SUV sale over a Sedan sale.

Profession	Business		Salaried		All
Gender	Female	Male	Female	Male	
Make					
Hatchback	0	290	15	277	582
SUV	52	29	107	82	270
Sedan	50	256	91	305	702
All	102	575	213	664	1554

Fig-19 crosstab_for_Make_between_Profession_and_Gender

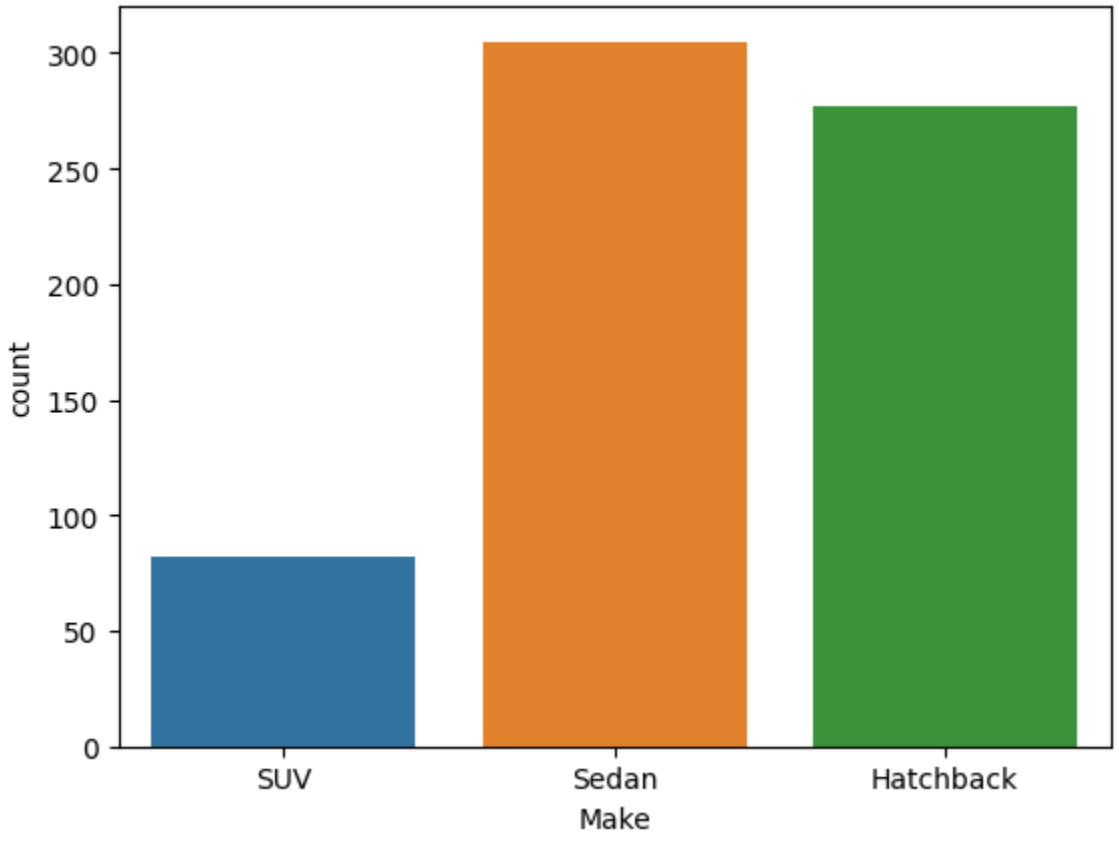

Fig-20 countplot_of_Salaried_Males_and_Make

We do not agree with the statement as a Salaried Male like Sedan more than a SUV as per the Data. Furthermore going by the numbers, Salaried Male Customers have purchased 305 Sedan cars as compared to only 82 SUV cars.

Does the amount spent on purchasing automobiles differ across the following attributes - Gender & Existence of Personal_Loan?

Gender

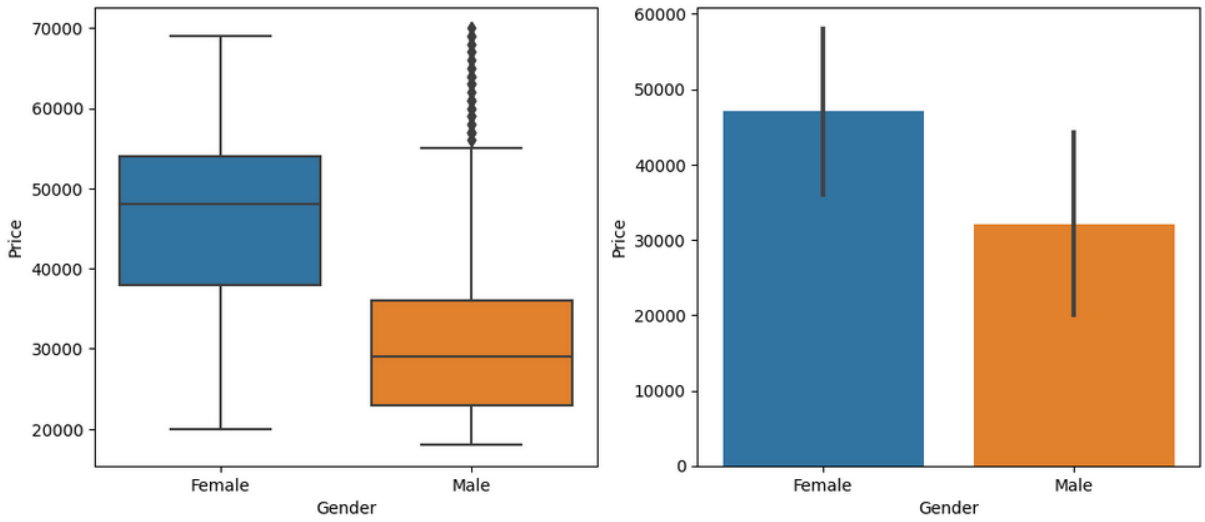

Fig-21 boxplot_and_barplot_of_Gender_and_Price

For the amount spent on automobiles, there is a specific correlation with gender. The graphs show the same. **Females prefer high priced cars than males**.

Gender	count	mean	std	min	25%	50%	75%	max
Female	315.0	47031.746032	10970.963782	20000.0	38000.0	48000.0	54000.0	69000.0
Male	1239.0	32119.451170	12074.764444	18000.0	23000.0	29000.0	36000.0	70000.0

Fig-22 stastical_summary_for_Gender_wise_Price

By the numbers

- The range for price for Females is between 20k to 69k, but for Males is 18k to 70k.
- IQR for both also differ significantly.
- Medians and Means of both are significantly different, almost 15-20k.

Presence of Personal Loan

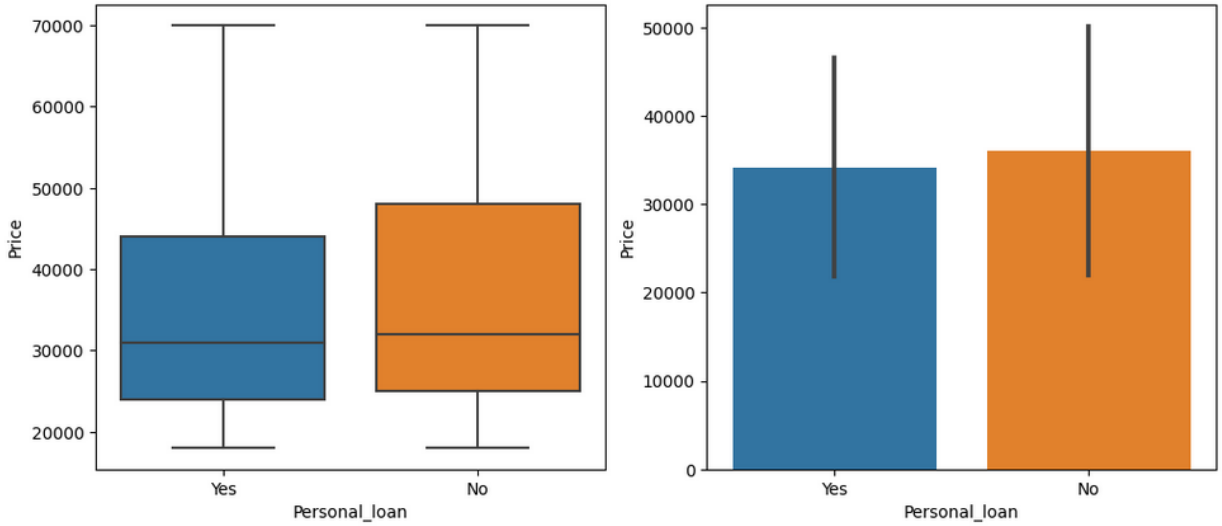

Fig-23 boxplot_and_barplot_of_Personal_loan_and_Price

For the amount spent on automobiles, **there is no significant correlation with the presence of personal loan**. The purchasing power of the customer is slightly higher in absence of a personal loan but cannot be called a factor for not buying a car.

Personal_loan	count	mean	std	min	25%	50%	75%	max
No	769.0	36087.126138	14112.695717	18000.0	25000.0	32000.0	48000.0	70000.0
Yes	785.0	34216.560510	12362.673448	18000.0	24000.0	31000.0	44000.0	70000.0

Fig-24 stastical_summary_for_Personal_loan_wise_Price

By the numbers

- The range for price for presence and absence of Personal Loan is same.
- IQR for both does not differ significantly.
- Medians and Means of both have slight differences of nearly 2k which is small as compared to the scale.

Does having a working partner lead to purchase of a higher priced car?

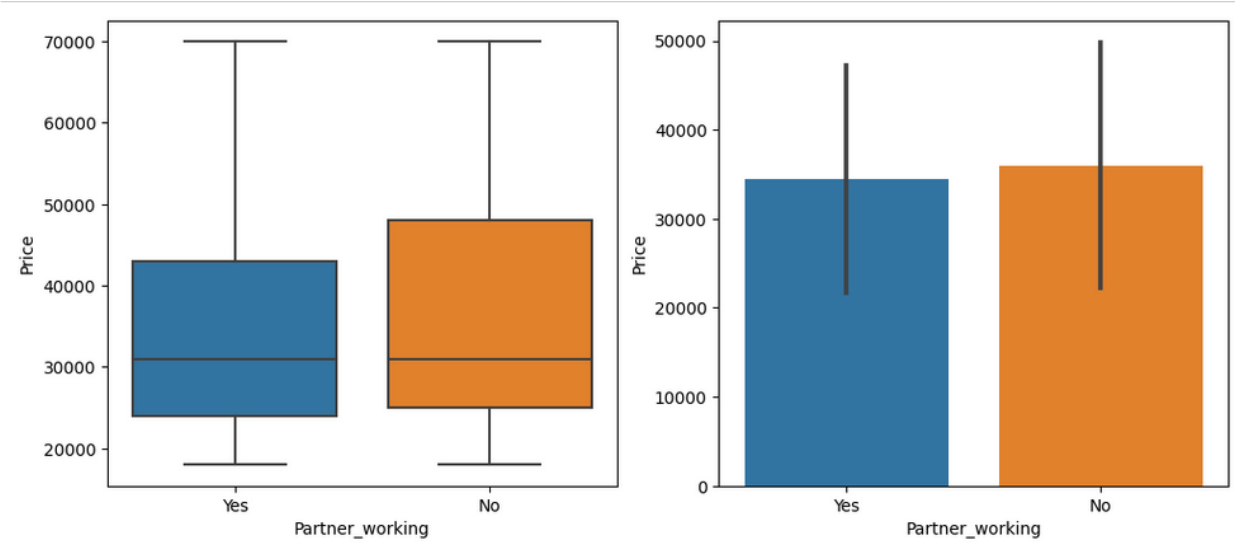

Fig-25 boxplot_and_barplot_of_Partner_working_and_Price

We notice that there is **no correlation between a working partner and the higher price of the car**.

Partner_working	count	mean	std	min	25%	50%	75%	max
No	713.0	36000.000000	13817.734086	18000.0	25000.0	31000.0	48000.0	70000.0
Yes	841.0	34414.982164	12781.691297	18000.0	24000.0	31000.0	43000.0	70000.0

Fig-26 stastical_summary_for_Partner_working_wise_Price

By the numbers

- The price range for the price for a partner working or not is the same.
- IQR differs by only 1k which is small looking at the scale.
- Medians of both are exactly the same and Means of both have slight differences of nearly 1.5k which is small as compared to the scale.

Devise an improved marketing strategy to send targeted information to different groups of potential buyers

For devising improved marketing strategy to send targeted information to different groups of potential buyers, we have used the variables namely - Gender and Martial_status to arrive at specific target groups.

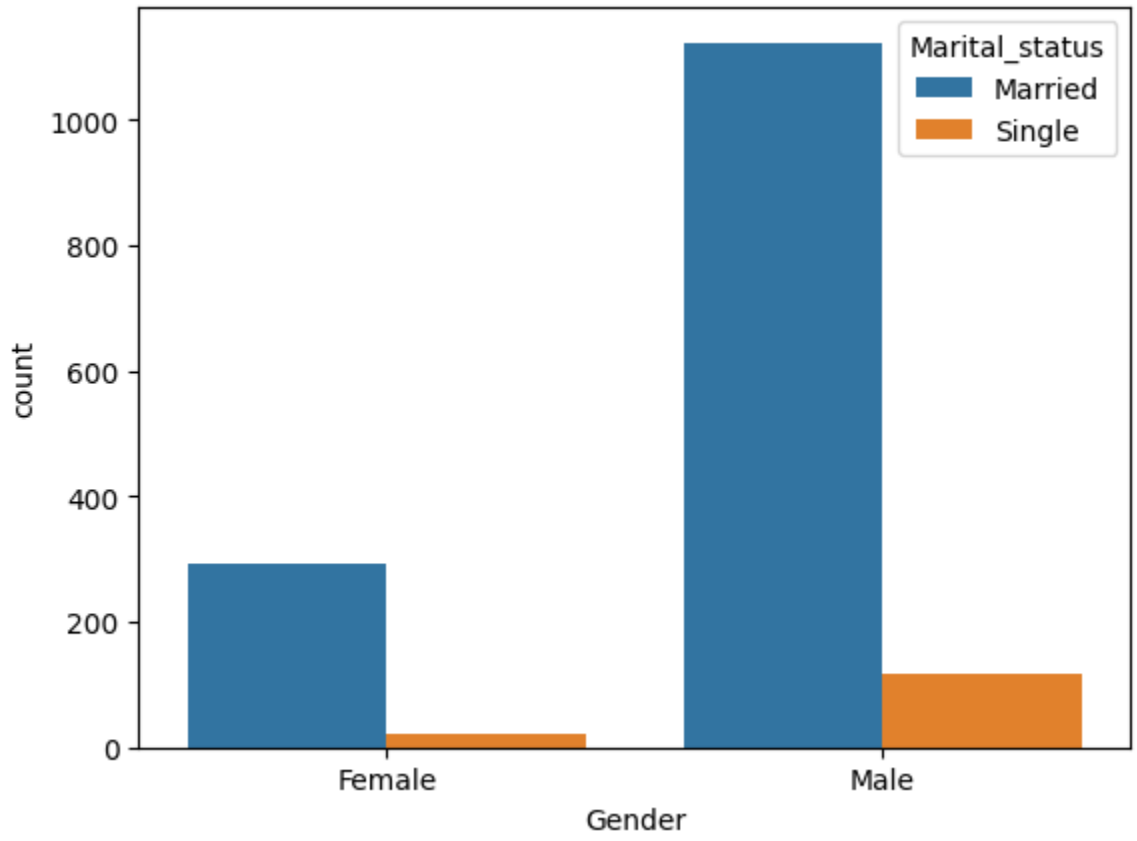

Fig-27 countplot_of_Gender_and_Marital_status

Marital_status	Married	Single	All
Gender			
Female	293	22	315
Male	1123	116	1239
All	1416	138	1554

Fig-28 crosstab_of_Gender_and_Marital_status

We have further analyzed, for this grouping based on Gender and Martial_status, what are the different models the various groups prefer. The details are as follows:

Marital_status	Married		Single		All
Gender	Female	Male	Female	Male	
Make					
Hatchback	14	484	1	83	582
SUV	152	102	7	9	270
Sedan	127	537	14	24	702
All	293	1123	22	116	1554

From the above details, we infer the following details and hence have come up with the specific groups to be targeted, as part of marketing strategy:

- Males, who are married, would definitely buy a car. They can be targeted to buy Sedans. In case, they do not agree for Sedan, then you can target them for Hatchback.
- Females, who are married, can be targeted to buy SUVs.
- Most males, who are single, would definitely buy a hatchback. They can be targeted.
- Most females, who are single, would definitely buy a sedan. They can be targeted.
- Hatchbacks are popular among male customers.
- SUVs are popular among female customers.

THE END!