PGP-DSBA PROJECT REPORT

PDS – Coded Project

BY

ISHAAN SHAKTI JAYARAMAN

PGPDSBA.O.JULY24.A

Contents

DATA OVERVIEW

1.1 Introduction

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.

1.2 Understanding the objective

They want to analyze the data to get a fair idea about the demand of customers which will help them in enhancing their customer experience. The Data Science team has shared some of the key questions that need to be answered. We will perform data analysis to find answers to these questions that will help the company to improve the business.

1.3 Understanding the structure of the data

We are presented with a dataset (austo_automobile.csv) to analyze and publish our observations and insights. The dataset contains information about customers who have purchased from Austo automobile

- The dataset consists of 1581 rows and 14 columns, this tells us that we have information for 1581 customers
- The columns consist of 8 object datatypes and 6 numerical datatypes
- Age, No of Dependents, Salary, Total Salary are integer datatypes
- Partner Salary is a float datatype
- Gender, Profession, Martial status, Education, Personal loan, House loan, Partner working and Make are object datatypes

1.4 Description of the columns

- **Age**: The age of the customer in years.
- **Gender**: The gender of the individual, categorized as male or female.
- **Profession**: The occupation or profession of the individual, categorized as business or salaried.
- **Marital_status**: The marital status of the individual, categorized as married or single
- **Education**: The educational qualification of the individual, categorized as Graduate or Post Graduate
- **No_of_Dependents**: The number of dependents (e.g., children, elderly parents) that the individual supports financially.

- **Personal_loan**: A binary variable indicating whether the individual has taken a personal loan categorized as Yes or No
- **House_loan**: A binary variable indicating whether the individual has taken a housing loan "Yes" or "No"
- **Partner_working**: A binary variable indicating whether the individual's partner is employed "Yes" or "No"
- **Salary**: The individual's salary or income.
- **Partner_salary**: The salary or income of the individual's partner, if applicable.
- **Total_salary**: The total combined salary of the individual and their partner (if applicable).
- **Price**: The price of a product or service.
- **Make**: The type of automobile categorized as SUV, Sedan or Hatchback

1.5 Preliminary analysis of the dataset

Conducting a preliminary analysis of the dataset is important to identify and treat any data irregularities such as missing values, duplicate values and spelling mistakes.

Here are the observations

- There are a total of 159 null values present in the dataset. Any cell which is blank or has the number “0” will be considered as a null value.
- There are 53 null values in the “Gender” column and 106 null values in the “Partner_salary” column.
- In the “Gender” column the word “Female” has been spelt incorrectly as “Femle” and “Femal” in some locations.
- There are no duplicate values in the dataset.

From these observations we can start treating all the anomalies of the dataset.

1.5.1 Treatment of null values

There are two ways to treat data with null values such as Imputation or Dropping the column. Dropping the column can be considered when a large portion of the data in the columns is missing. Imputation is a method that is used when we would like to replace the missing values with either the mean, median or mode value of the column so that there is minimal data loss.

The null values in the “Gender” column are treated by imputing the missing values with the “Mode” of the column as it is a categorical datatype. The occurrences of “Male” are more than the occurrences of “Female” so the mode of the column is “Male” which will be used to replace the missing/null values.

The null values in the “Partner_salary” column is a unique case as such there are multiple columns that can help us to ascertain a value for it. Going by the logic that

Total_salary = Salary + Partner_salary

Then,

Partner_salary = Total_salary – Salary

We will also take into consideration the value in “Partner_working” column. With these findings we can impute the values based on the Rule that

If value in “Partner_working” = “No”, “Partner_salary” = 0
If value in “Partner_working” = “Yes”, “Partner_salary” = “Total_salary” – “Salary”

Following these steps, we may observe now that all null values have been treated successfully without any loss in data. The “Partner_salary” column has also been converted to an integer datatype from its previous float datatype for consistency.

1.5.2 Treatment of spelling errors

The spelling errors in the “Gender” column can be easily treated by replacing the incorrect spelt words with the correct word. In this case replacing all instances of “Femle” and “Femal” with “Female”

1.6 Statistical summary of the dataset

Table 1 Statistical Summary

	count	mean	std	min	25%	50%	75%	max
Age	1581.000	31.922	8.426	22.000	25.000	29.000	38.000	54.000
No_of_Dependents	1581.000	2.458	0.943	0.000	2.000	2.000	3.000	4.000
Salary	1581.000	60392.220	14674.825	30000.000	51900.000	59500.000	71800.000	99300.000
Partner_salary	1581.000	19233.776	19670.391	0.000	0.000	25100.000	38100.000	80500.000
Total_salary	1581.000	79625.996	25545.858	30000.000	60500.000	78000.000	95900.000	171000.000
Price	1581.000	35597.723	13633.637	18000.000	25000.000	31000.000	47000.000	70000.000

Following all corrections to the data, we may begin analyzing the dataset deeper.

- The average age of the customers is 31.92 years while the median age is 29.
- On average customers may have 2 number of dependents.
- The median salary in our dataset is 59,500 and salaries range from 30,000 to 99,300
- The average of total salary is 78,000 which is higher than the average salary, this will help us finding a connection between partner salary and type and price of the car purchased.
- The average price of a car sold is 35,598. The prices of the car range from 18,000 to 70,000

1.7 Detection of outliers or data irregularities in the dataset

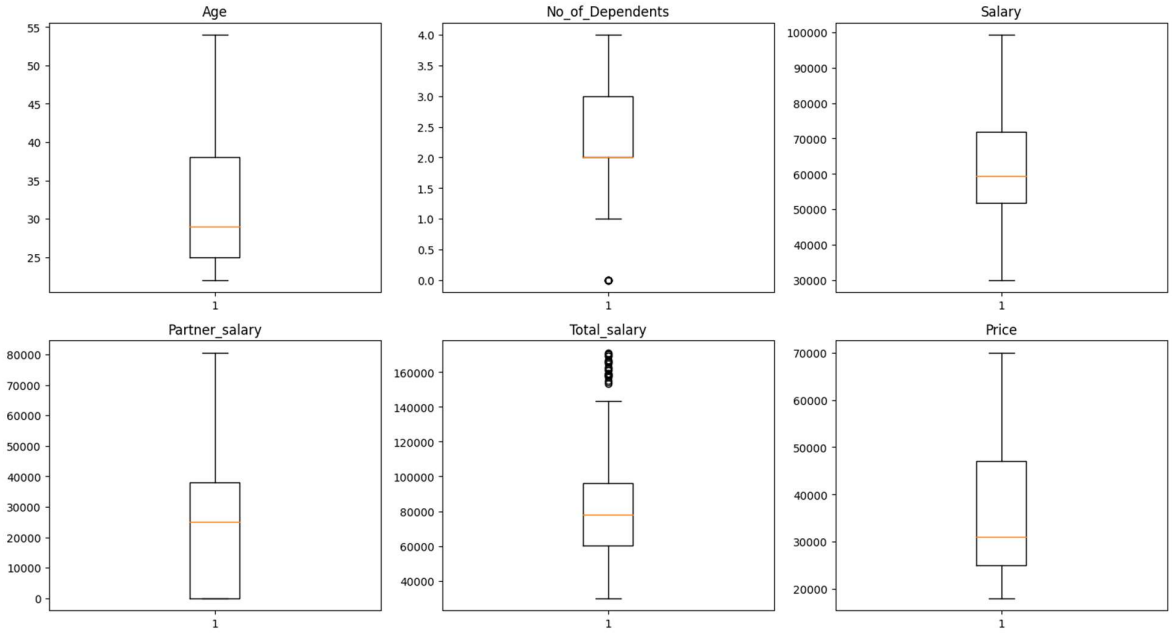
An outlier is a data point that is completely distant from the rest of the data. As these distant values can affect our observations of the dataset. Depending on the conditions we may impute the values with the upper or lower whisker of the dataset.

Upper whisker = Q3 + 1.5 * IQR
Lower whisker = Q1 – 1.5 * IQR

Or we can drop these specific data points or we may also replace these values with null values which will be ignored when analyzing the data or if the outliers are genuine points, we may keep them untouched. It is important to utilize the best method to treat outliers

1.7.1 Outliers in numerical variables

Figure 1 Outlier detection using boxplot for all numerical variables

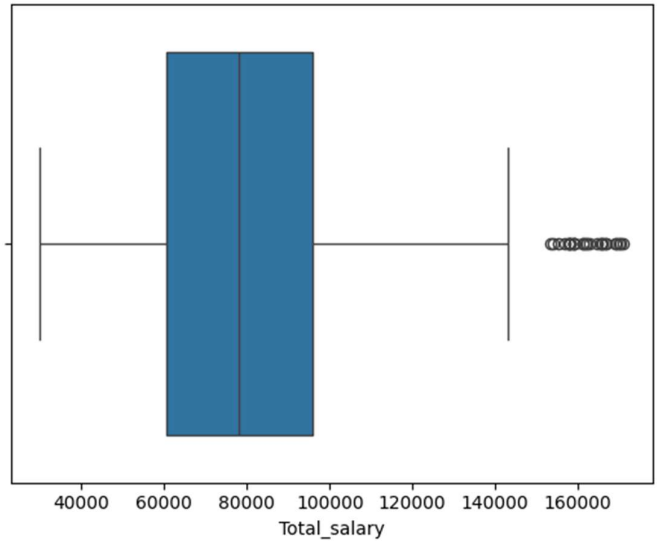


From the above picture we may observe few outliers in “No_of_Dependents” and in “Total_salary”.

No_of_Dependents tells us the number of dependents a person supports. 0 is a genuine value as there exist people that may not have any dependents. This can be ignored.

Total_salary has multiple outliers has multiple outliers so we may observe the chart closely.

Figure 2 Outlier detection in Total_salary



Based on the above boxplot, there are 27 observations of total salaries that are considered as outliers (total salaries greater than 150,000) These total salaries are very high number as the salary of the partner is also relatively high which increases the total. The outliers can be replaced with the upper whisker however the total salary values are genuine, it may help us generate insights if there is any correlation between high total salaries and make of the car purchased so we will consider the outliers as genuine and not change the values.

1.7.2 Data irregularities in the dataset

Figure 3 Value counts in categorical variables

```
Gender
Male      1252
Female     329
Name: count, dtype: int64
-------------------------
Profession
Salaried    896
Business    685
Name: count, dtype: int64
-------------------------
Marital_status
Married    1443
Single      138
Name: count, dtype: int64
-------------------------
Education
Post Graduate    985
Graduate         596
Name: count, dtype: int64
```

```
Personal_loan
Yes    792
No     789
Name: count, dtype: int64
-------------------------
House_loan
No     1054
Yes     527
Name: count, dtype: int64
-------------------------
Partner_working
Yes    868
No     713
Name: count, dtype: int64
-------------------------
Make
Sedan        702
Hatchback    582
SUV          297
Name: count, dtype: int64
```

After analyzing all the values in the categorical variables after already correcting the incorrect values in the “Gender” column, we may observe that are no irregularities present in the dataset.

UNIVARIATE ANALYSIS

A univariate analysis explores all variables in a dataset separately so that we may identify any patterns. In this dataset we will explore the numerical and categorical variables individually.

2.1 Exploring all numerical variables in the dataset

2.1.1 Age

Figure 4 Histogram of "Age"

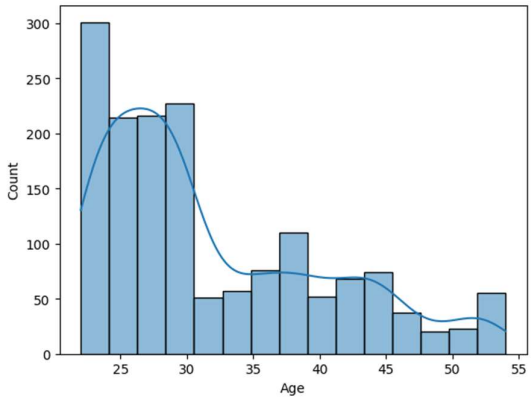
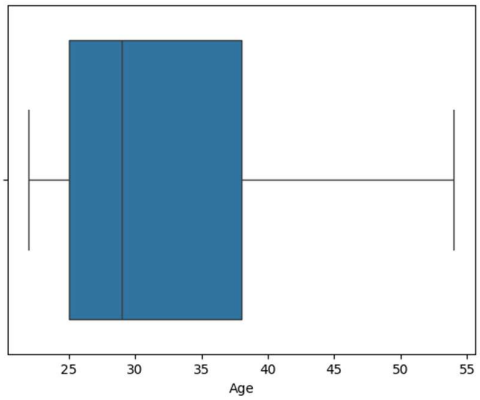


Figure 5 Boxplot of "Age"



Insights:

- The distribution of “Age” is positively skewed.
- Majority of buyers are aged between 20-30, which means that the younger group forms a major part of our customer base.

2.1.2 Salary

Figure 6 Histogram of "Salary"

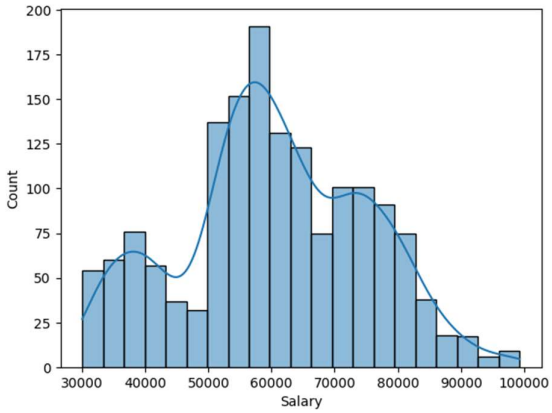
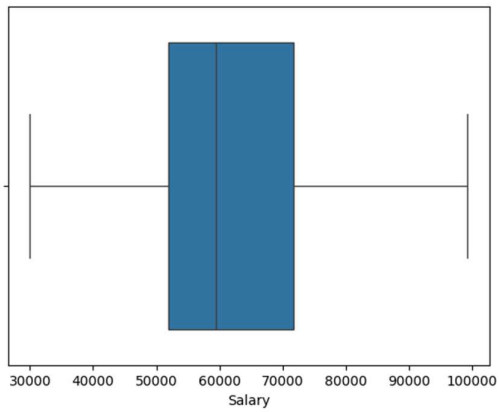


Figure 7 Boxplot of "Salary"



Insights:

- Majority of buyers are people with salaries ranging from around 52,000 to 72,000

2.1.3 Partner Salary

Figure 8 Histogram of "Partner_salary"

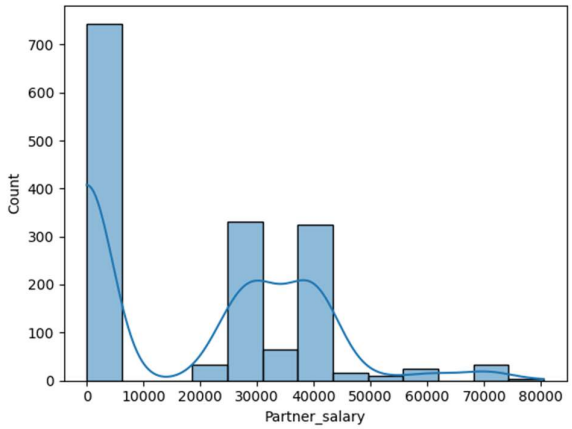
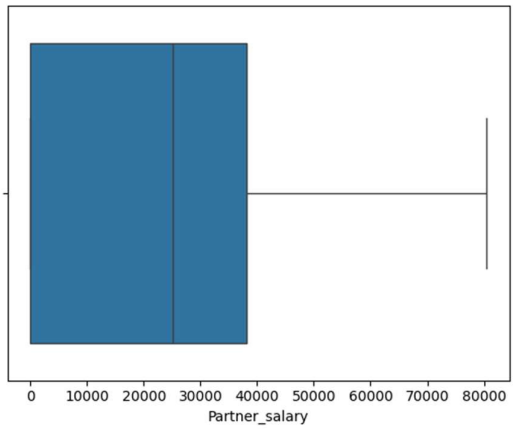


Figure 9 Boxplot of "Partner_salary"



Insights:

- Majority of buyers are those without a working partner in other words the partner salary is zero
- There are a significant number buyers with working partners whose partner salary range from 25,000 to 45,000 approximately.

2.1.4 Total Salary

Figure 10 Histogram of "Total_salary"

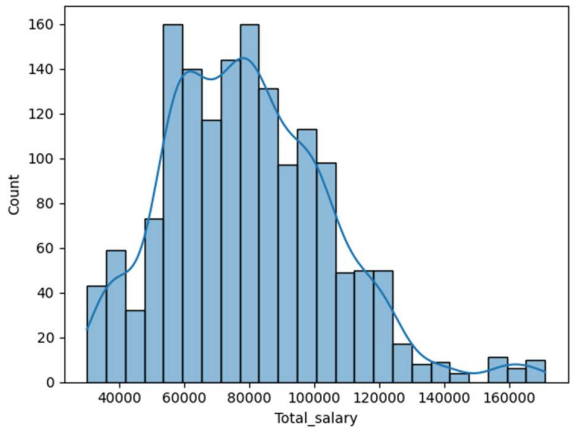
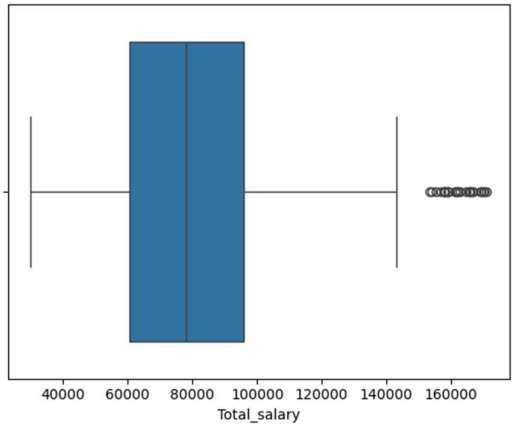


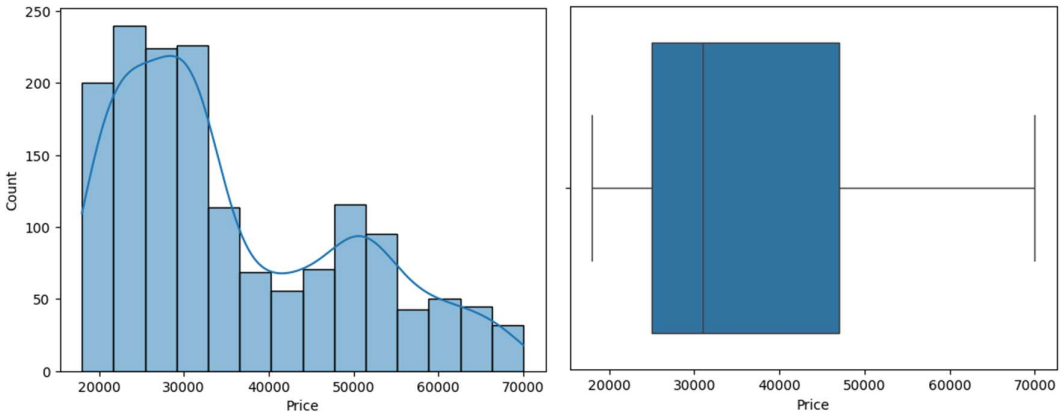
Figure 11 Boxplot of "Total_salary"



Insights:

- The bulk total salary range of buyers lies between 50,000 to 100,000

2.1.5 Price

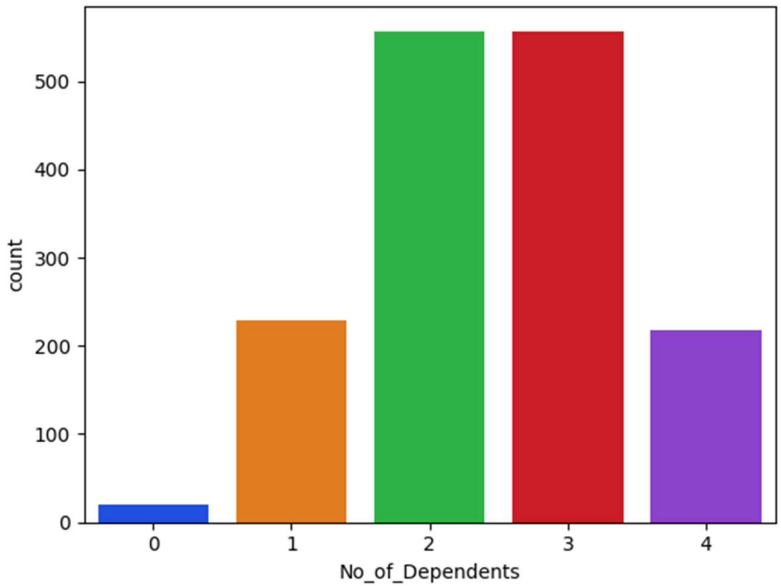


Insights:

- The distribution of "Price" is positively skewed
- Majority of cars sold are priced between 18,000 to 33,000 approximately

2.1.6 No of Dependents

Figure 12 Barplot of "No_of_Dependents"



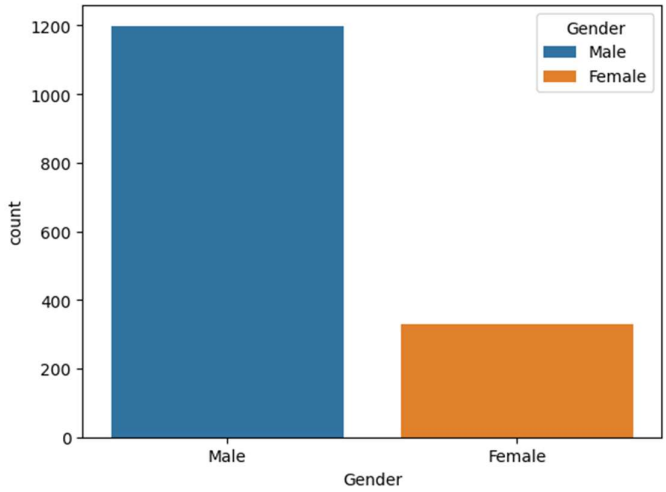
Insights:

- Large number of buyers of the cars are people with 2 or 3 no of dependents.
- No of dependents may affect the type of car they may purchase, having more dependents may push the buyer to go for an SUV or Sedan over a Hatchback

2.2 Exploring all categorical variables in the dataset

2.2.1 Gender

Figure 13 Barplot of "Gender"

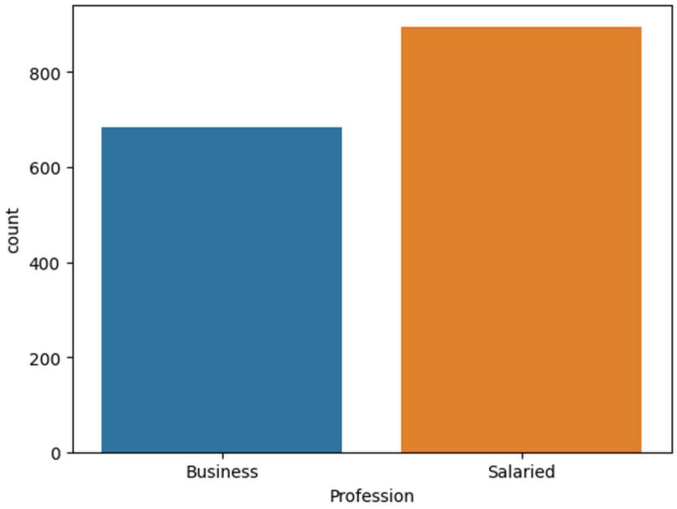


Insights:

- Majority of customers are “Male”

2.2.2 Profession

Figure 14 Barplot of "Profession"

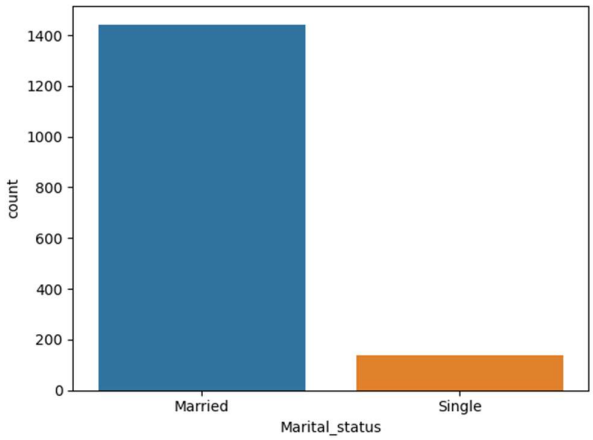


Insights:

- There are higher chances of a salaried individual purchasing a car over someone who runs a business.

2.2.3 Marital Status

Figure 15 Barplot of "Marital Status"

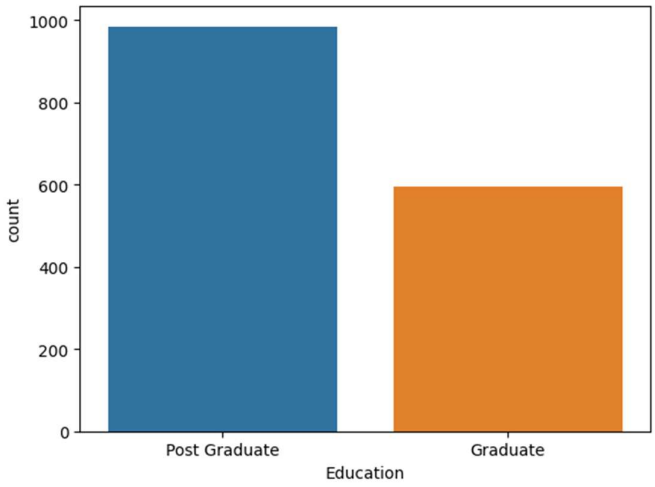


Insights:

- A large number of buyers are married individuals. More than 1400 people in the dataset are married.

2.2.4 Education

Figure 16 Barplot of "Education"

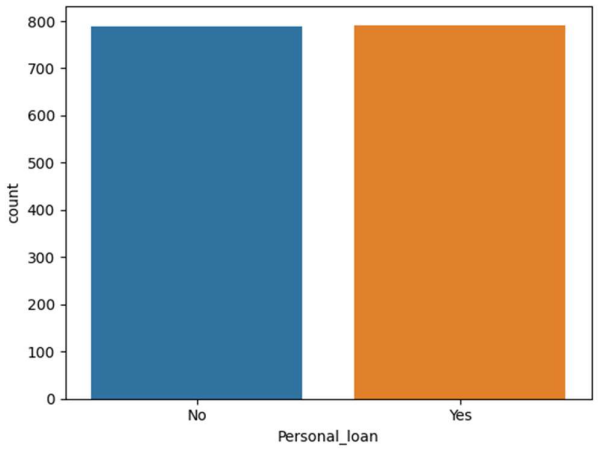


Insights:

- A person holding a post-graduate degree is more than likely to purchase a car over a person holding a graduate degree.

2.2.5 Personal Loan

Figure 17 Barplot of "Personal_loan"

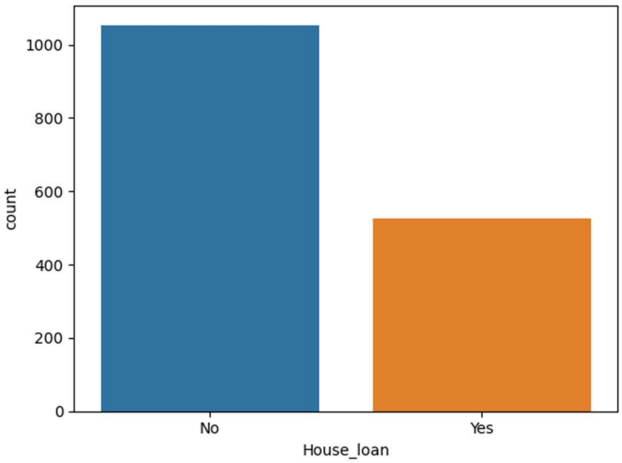


Insights:

- There are almost equal instances of people purchasing cars regardless of whether they have taken a personal loan or not.

2.2.6 House Loan

Figure 18 Boxplot of "House Loan"

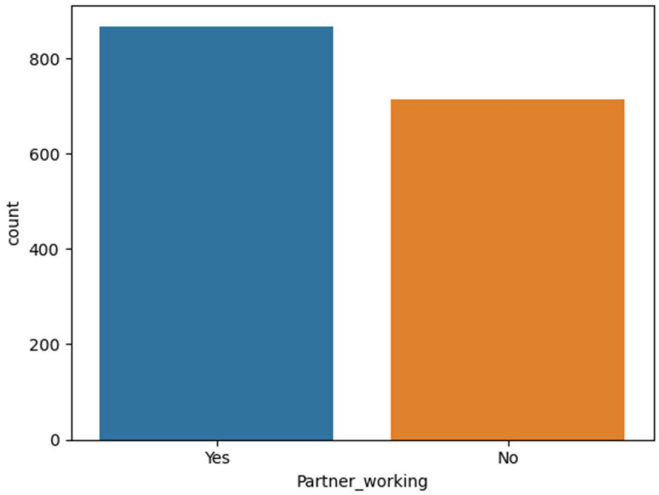


Insights:

- There are a large number of people who have not taken a House loan while purchasing a car, this value is almost double compared to those who have taken a house loan.

2.2.7 Partner Working

Figure 19 Boxplot of "Partner_working"

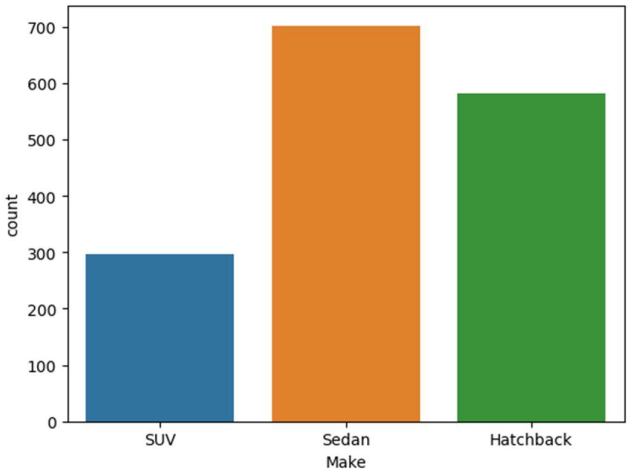


Insights:

- The number of buyers who have a working partner are slightly more than the number of buyers who do not have one.

2.2.8 Make

Figure 20 Boxplot of "Make"



Insights:

- Majority of buyers have purchased the Sedan
- Hatchback is the second most purchased and SUV is the least purchased of all the car types

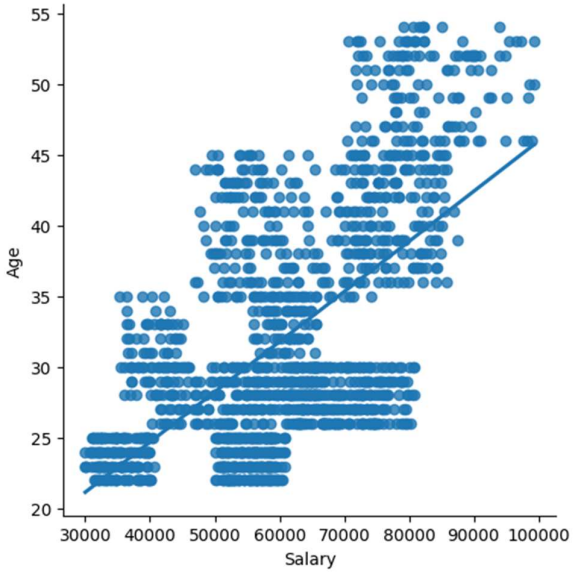
BIVARIATE ANALYSIS

A bivariate analysis explores the relationship between two or more variables in a dataset so that we may gain deeper insights. We will explore the relationship and correlation between all numerical variables as well as the relationship between numerical and categorical variables.

3.1 Exploring the relationship between numerical variables

3.1.1 Salary vs Age

Figure 21 Lmplot of "Salary" vs "Age"

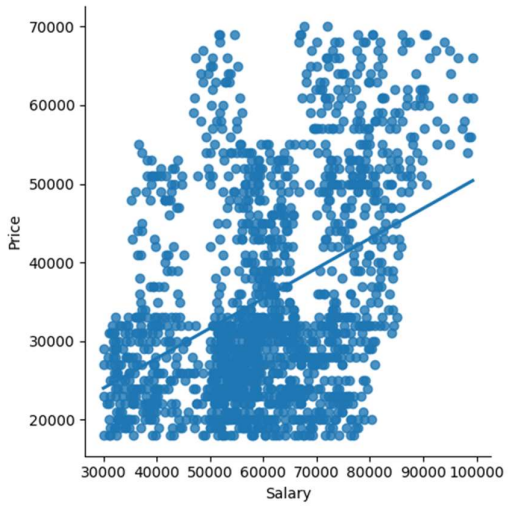


Insights:

- We observe that there is a positive correlation between salary and age.
- As a person ages the salary received by them also increases.
- A person who receives a high salary may have a higher chance of purchasing a car

3.1.2 Salary vs Price

Figure 22 Lmplot of "Salary" vs "Price"

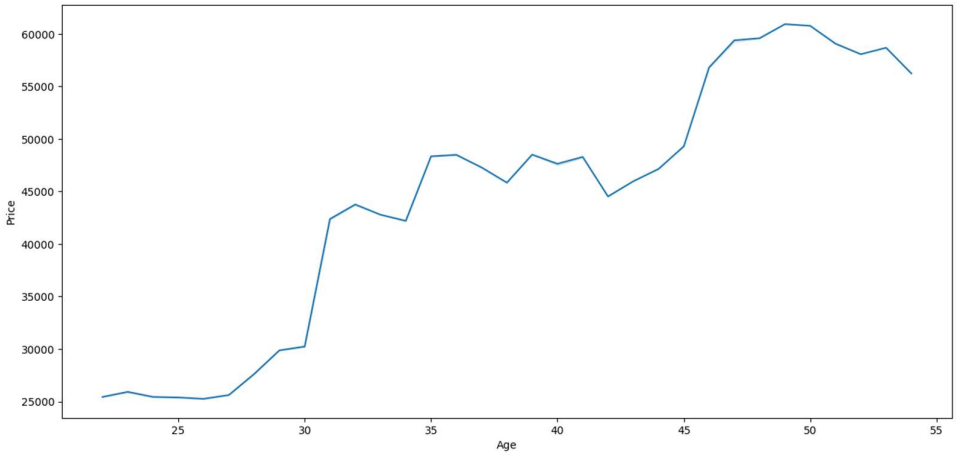


Insights:

- There is a positive correlation between salary and price.
- A person who receives a higher salary is willing to purchase cars that are priced higher than others.

3.1.3 Age vs Price

Figure 23 Lineplot of "Age" vs "Price"



Insights:

- There is a positive correlation between Age and Price.
- An older person is more willing to purchase a car that is priced higher as he earns more salary.

3.2 Exploring the correlation between all numerical variables

Figure 24 Heatmap of all numerical variables

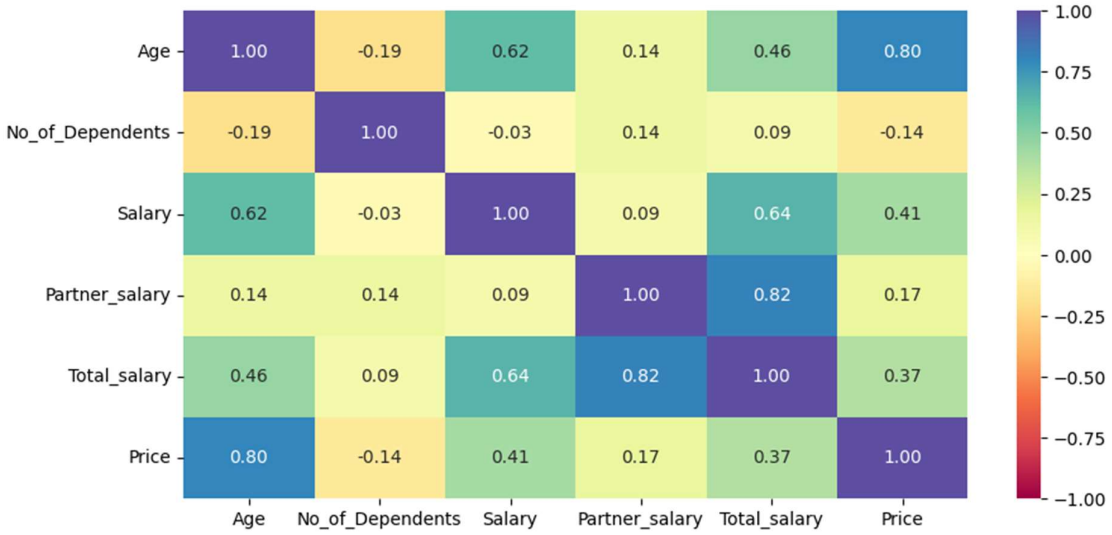
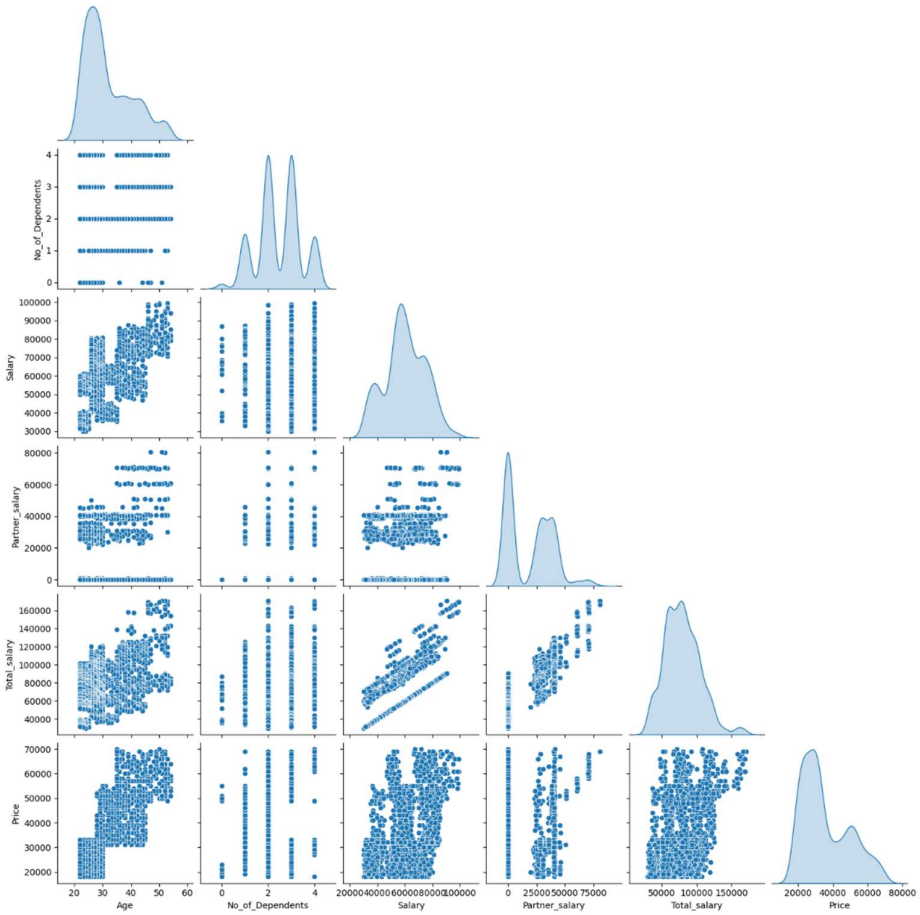


Figure 25 Pairplot of all numerical variables



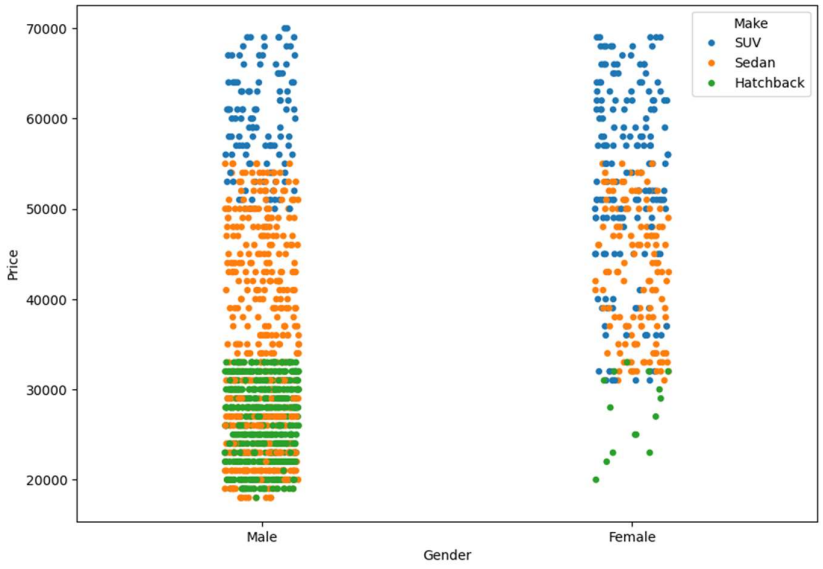
Insights:

- The highest correlation is between Partner salary and Total salary, logically total salary increases as partner salary increases.
- There is also high correlation between Price and Age, Salary and Age, and Total Salary and Salary
- There is little correlation between all other numerical variables

3.3 Exploring relationship between categorical vs numerical variables

3.3.1 Relationship between Gender, Price and Make

Figure 26 Stripplot of "Gender", "Price" and "Make"

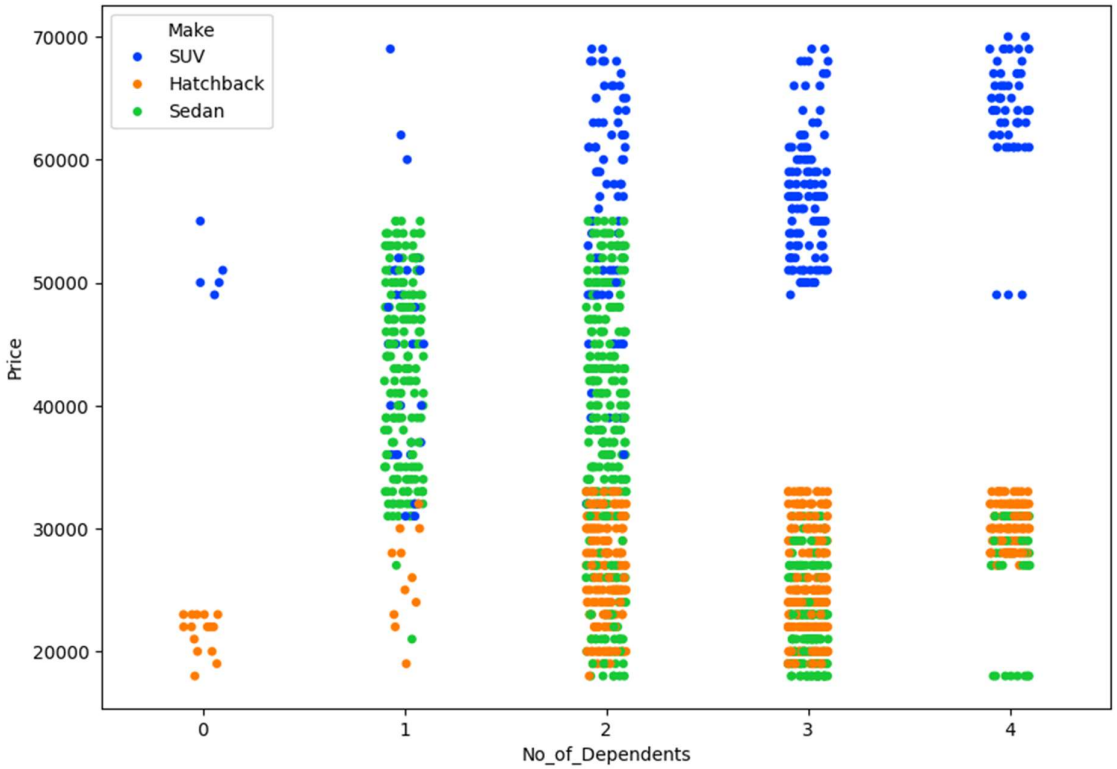


Insights:

- Both Men and Women purchase cars at all prices, high or low
- Men purchasing cars under 32,000 strongly prefer hatchbacks
- There are more Women that purchase SUVs at medium to high prices

3.3.2 Relationship between No of Dependents, Make and Price

Figure 27 Stripplot of "No_of_Dependents", "Make" and "Price"

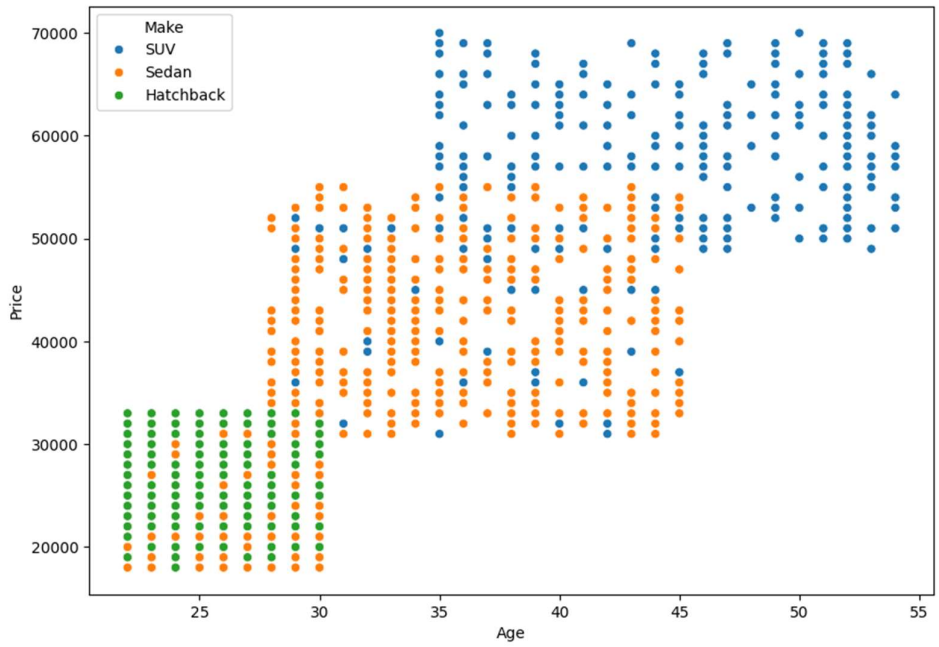


Insights:

- Sedans are more preferred by the buyers when the no. of dependents is 1 or 2
- SUVs are preferred by the buyers who are willing to pay a higher price when the no. of dependents is greater than 2
- Hatchback is preferred by the buyers who have 2 or more dependents but are not willing to pay a price more than 33,000 approximately.

3.3.3 Relationship between Age, Price and Make

Figure 28 Scatterplot of "Age', "Price" and "Make"



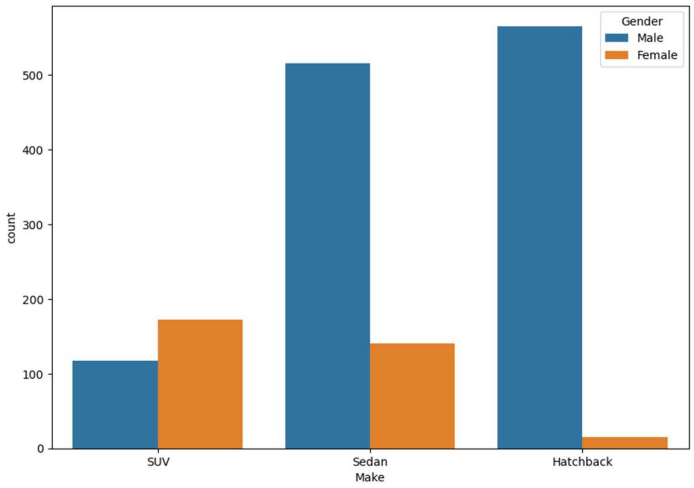
Insights:

- Younger age groups (between 18-30) prefer to buy hatchbacks
- Middle-aged group mostly prefer to buy Sedans
- Older and some middle-aged group prefer to buy SUVs
- There is a positive correlation between age and salary. we can say that as salaries become higher, Sedans and SUVs will more preferred by the individual.

KEY QUESTIONS

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

Figure 29 Barplot of "Make"

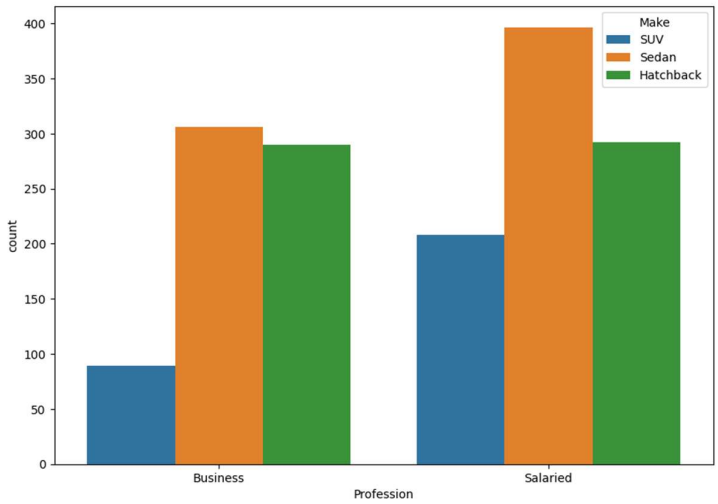


Insights:

- Based on chart, Men do not prefer SUVs more compared to women
- Men prefer Sedans and Hatchbacks
- Women prefer SUVs

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

Figure 30 Barplot of "Profession"

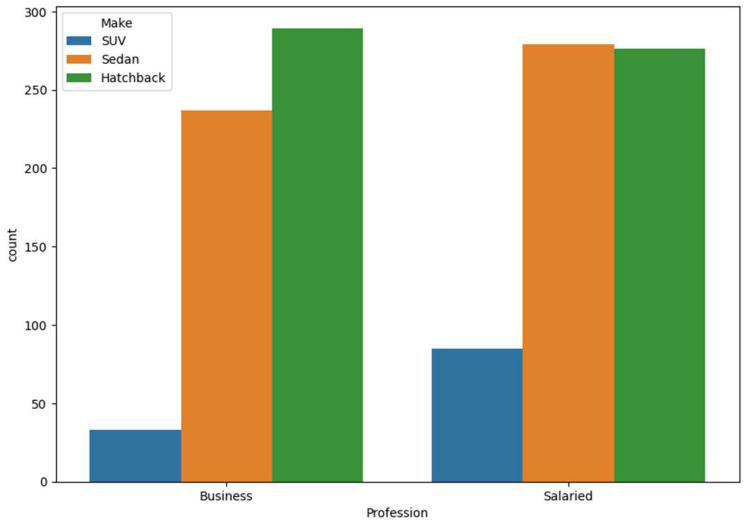


Insights:

- Based on the chart, a salaried person is more likely to buy a sedan than an SUV or a Hatchback

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

Figure 31 Barplot of "Profession" - Male Only

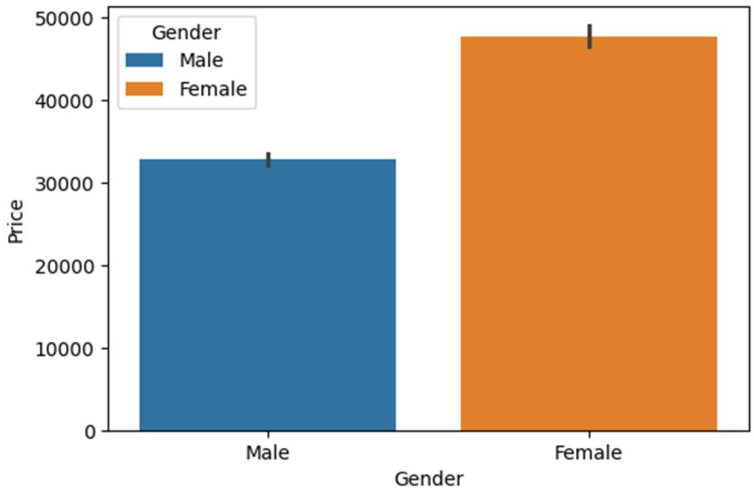


Insights:

- Based on the chart a salaried male prefers Sedans and/or Hatchbacks.
- SUVs are least preferred by salaried males

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

Figure 32 Barplot of "Gender" and "Price"

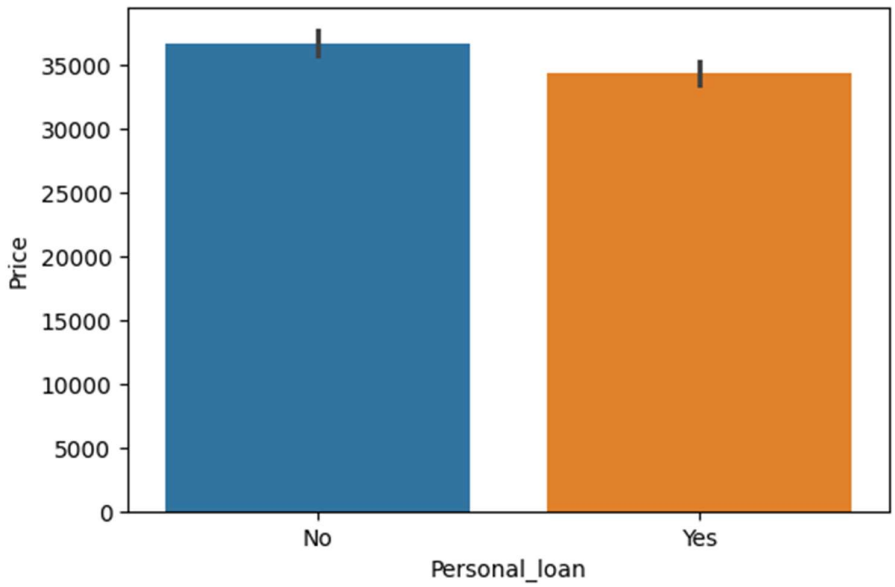


Insights:

- Based on the chart, on average Women spend more on cars when compared to Men
- Women have spent around 47,705 whereas Men have spent 32,416

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

Figure 33 Barplot of "Personal_loan" and "Price"

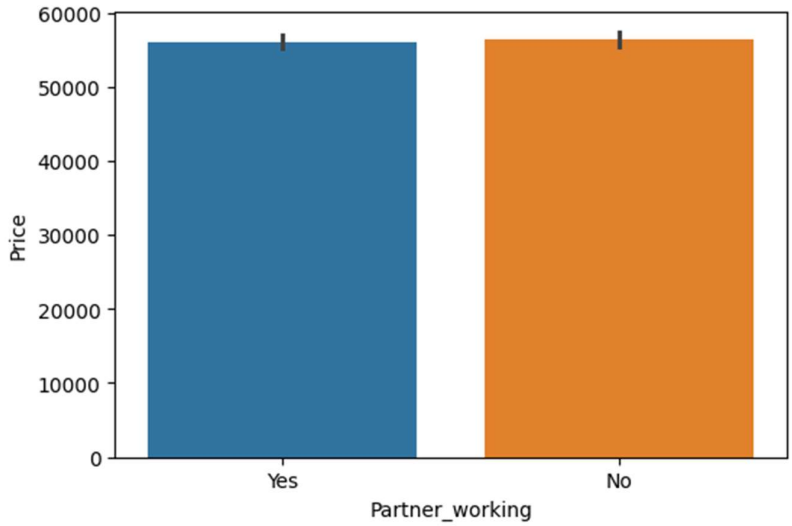


Insights:

- Based on the chart, people who haven't taken a personal loan on average have spent more on automobiles compared to those who have taken a loan.
- Those who have taken a loan have spent an average of 34,457 whereas those who have not have spent 36,743

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

Figure 34 Boxplot of "Partner_working" and "Price"



Insights:

- Based on the chart, having a working partner does not influence the decision to purchase a higher priced car.
- The mean and median values are almost comparable to each other if there is or isn't a working partner.

ACTIONABLE INSIGHTS & RECOMMENDATIONS

Based on our complete analysis of the dataset, we may infer and recommend the following

- Men with salaries under 32,000 form majority of sales for Hatchbacks
- Women on average spend more than Men on cars even if they are a minor part of the sales and prefer SUVs
- Single women prefer sedans while married women prefer SUVs
- Married men prefer sedans while single men prefer hatchbacks

Table 2 Mode of "make" grouped by Gender and Marital Status

Gender	Marital Status	Mode of "Make"
Female	Married	SUV
	Single	Sedan
Male	Married	Sedan
	Single	Hatchback

- The marketing team must prioritize marketing hatchbacks to the younger males who are salaried.
- Sedans are more popular with the middle-aged married men as their no. of dependents are likely higher and their total salary is also higher.
- Sedans are also popular among single women who are willing to purchase a higher priced car.
- The target audience for SUVs are married women and older-aged group people as likely the very high salary and high no. of dependents warrant the need to purchase larger cars to accommodate more people.