PDS CODED PROJECT

Austo Motor Company Business Report

DSBA

Submitted By: Maheep Singh
Batch: PGPDSBA.O.AUG24.A

Table of Contents

List of Figures

List of Tables

Business Context & Data Dictionary

Context

Analysts are required to explore data and reflect on the insights. Clear writing skill is an integral part of a good report. Note that the explanations must be such that readers with minimum knowledge of analytics is able to grasp the insight.

Austo Motor Company is a leading car manufacturer specializing in SUV, Sedan, and Hatchback models. In its recent board meeting, concerns were raised by the members on the efficiency of the marketing campaign currently being used. The board decides to rope in analytics professional to improve the existing campaign.

Objective

Austo Motor Company want to analyse the data to get a fair idea about the demand of customers which will help them in enhancing their customer experience. Suppose you are a Data Scientist at the company and the Data Science team has shared some of the key questions that need to be answered. Perform the data analysis to find answers to these questions that will help the company to improve the business.

Data Description

- **Age**: The age of the individual in years.
- **Gender**: The gender of the individual, categorized as male or female.
- **Profession**: The occupation or profession of the individual.
- **Marital_status**: The marital status of the individual, such as married &, single
- **Education**: The educational qualification of the individual Graduate and 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 "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

Problem 1- Data Overview

Load the Dataset

- Import required libraries to load the dataset, perform data manipulation, analysis & visualization tasks to solve the business problem.
- Below is the view of the first 5 rows of the imported dataset to give a quick overview of the data: -

	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

Table 1: First 5 Rows of the Dataset

Understanding the structure of Data & Data-types

- Delving deeper to analyse basic information of data: -

```
<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: 173.1+ KB
```

Figure 1: Basic Information of Dataset

- After analysing the basic information of the data set, the following inferences can be derived: -
  - ✓ There are **1528** rows and **14** columns.

- ✓ No duplicate entries found.
- ✓ It can be observed that there **8 object** datatype columns and **6 numerical** data-type columns.
- ✓ Columns, **Gender & Partner_salary**, don't have 1528 values, which means **there are null values** present in the dataset.
- ✓ Delving into the unique values of all categorical variables to check for any inconsistencies: -

```
----------------------------------------------------------------
Gender :-

Unique Values = ['Male' 'Femal' 'Female' nan 'Femle']

Column Value Count Gender
Male      1199
Female     327
Femal        1
Femle        1
Name: count, dtype: int64

----------------------------------------------------------------
Profession :-

Unique Values = ['Business' 'Salaried']

Column Value Count Profession
Salaried    896
Business    685
Name: count, dtype: int64

----------------------------------------------------------------
Marital_status :-

Unique Values = ['Married' 'Single']

Column Value Count Marital_status
Married    1443
Single      138
Name: count, dtype: int64

----------------------------------------------------------------
Education :-

Unique Values = ['Post Graduate' 'Graduate']

Column Value Count Education
Post Graduate    985
Graduate         596
Name: count, dtype: int64
```

```
----------------------------------------------------------------
Personal_loan :-

Unique Values = ['No' 'Yes']

Column Value Count Personal_loan
Yes    792
No     789
Name: count, dtype: int64

----------------------------------------------------------------
House_loan :-

Unique Values = ['No' 'Yes']

Column Value Count House_loan
No     1054
Yes     527
Name: count, dtype: int64

----------------------------------------------------------------
Partner_working :-

Unique Values = ['Yes' 'No']

Column Value Count Partner_working
Yes    868
No     713
Name: count, dtype: int64

----------------------------------------------------------------
Make :-

Unique Values = ['SUV' 'Sedan' 'Hatchback']

Column Value Count Make
Sedan        702
Hatchback    582
SUV          297
Name: count, dtype: int64
```

Figure 2: Unique Values of Categorical Variables

Clearly, **Gender** has both **null & mis-spelled values** that would require treatment.

- ✓ Summary of the division of all categorical variables: -

Binary	Multilevel
Gender	Make
Profession	
Marital_status	
Education	
Personal_loan	
Home_loan	
Partner_working	

Table 2: Division of Categorical Variables

- After carefully vetting the data, below table summarizes each data-type & initial impression/action: -

Attribute/Column	Inference	Action
Age	✓ Integer Type ✓ No null/missing values	a) No value-treatment required b) Recommended to add another column 'Age_category' for better analysis based on age-brackets
Gender	✓ Object Type ✓ Null values found ✓ Mis-spelled values found	a) Null-value treatment required b) Mis-spelled values need to be replaced with appropriate values
Profession	✓ Object Type ✓ No null/missing values	a) No value-treatment required
Marital_status	✓ Object Type ✓ No null/missing values	a) No value-treatment required
Education	✓ Object Type ✓ No null/missing values	a) No value-treatment required
No_of_Dependents	✓ Integer Type ✓ No null/missing values	a) Recommended to translate into a categorical variable for better analysis since the no. of dependents won't be high
Personal_loan	✓ Object Type ✓ No null/missing values	a) No value-treatment required
House_loan	✓ Object Type ✓ No null/missing values	a) No value-treatment required
Partner_working	✓ Object Type ✓ No null/missing values	a) No value-treatment required
Salary	✓ Integer Type ✓ No null/missing values	a) Recommended to convert to Float data-type for consistency (Partner_salary column is also Float data-type & in case if any operation required between the two, it is recommended to keep the date-type same)
Partner_salary	✓ Float Type ✓ Null values found	a) Null-value treatment required
Total_salary	✓ Integer Type ✓ No null/missing values	a) Recommended to convert to Float data-type for consistency (Partner_salary column is also Float data-type & in case if any operation required between the two, it is recommended to keep the date-type same) b) This is a derived column and it is recommended to re-calculate the column post null value treatment is done for Partner_salary
Price	✓ Integer Type ✓ No null/missing values	a) Recommended to convert to Float data-type for consistency
Make	✓ Object Type ✓ No null/missing values	a) No value-treatment required

Table 3: Data-type Summary & Next Action

Data Treatment: Check for & treat Data-types, Missing-values & Data-irregularities

- Converting Int64 to Float64 data-type for Salary, Total_salary, Price.
- Converting Int64 to Object data-type for No_of_Dependents
- Create a new categorical variable 'Age_category' from 'Age' based on Age-brackets for future analysis: -

```
Age_category
Young (Below 31)        958
Middle-aged (31-45)     488
Old (>45)               135
```

Figure 3: Age-categories based on age-brackets

- Data-type summary after data-type treatment: -

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581 entries, 0 to 1580
Data columns (total 15 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   object 
 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   float64
 10  Partner_salary    1475 non-null   float64
 11  Total_salary      1581 non-null   float64
 12  Price             1581 non-null   float64
 13  Make              1581 non-null   object 
 14  Age_category      1581 non-null   object 
dtypes: float64(4), int64(1), object(10)
memory usage: 185.4+ KB
```

Figure 4: Data-type Summary after data-type Treatment

- Inspecting Null Values: how many & what percentage: -

```
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
```

```
Gender
Male      0.758381
Female    0.206831
NaN       0.033523
Femal     0.000633
Femle     0.000633
Name: proportion, dtype: float64
 --------------------------------------------------
Partner_salary
0.0        0.394054
NaN        0.067046
40500.0    0.018343
40200.0    0.017078
40100.0    0.015813
             ...
45800.0    0.000633
27300.0    0.000633
22900.0    0.000633
23100.0    0.000633
35000.0    0.000633
Name: proportion, Length: 150, dtype: float64
 --------------------------------------------------
```

Figure 5: Null Value Inspection

I. 2 variables require null value treatments – ‘Gender’ & ‘Partner_salary’.
II. 3.35% constitute null values in ‘Gender’ & 6.70% constitute null values in ‘Partner_salary’. Given the percentage of null values, it is not recommended to drop them as it would mean loss of information.
III. Couple of entries in Gender have been mis-spelled & would be required to be replaced with correct spellings for consistency.

- Treating ‘Gender’ null/error-values: -
  I. Gender
     - Replace error-values, ‘Femal’ & ‘Femle’, with the correct value ‘Female’: -

```
Gender
Male      0.758381
Female    0.208096
NaN       0.033523
```

Figure 6: Gender Distribution after correcting mis-spelled values

     - Clearly, the data is dominated by males with 75.83% as against females with 20.81%.
     - Given the high percentage of null values, we would not drop null values, rather, impute them with the mode value ‘Male’ (for categorical variables, it is recommended to fill the missing values with the majority class of that column)
     - Imputing Null values with ‘Male’: -

```
Gender
Male      0.791904
Female    0.208096
```

Figure 7: Gender Distribution after null & error-value treatment

II. Partner_salary

- Out of 106 null values, let's look at the distribution by 'Marital_status': -

Marital_status	
Married	90
Single	16

Figure 8: Partner Salary Null values by Marital Status before Treatment

- Clearly, those who are 'Single', won't have a partner & there is no question of 'Partner_salary'. For these cases, lets impute 'Partner_salary' as 0.

Marital_status	
Married	90

Figure 9: Partner Salary Null values by Marital Status after Treatment

- Out of remaining 90 null values, let's look at the distribution by 'Partner_working': -

Partner_working	
No	74
Yes	16

Figure 10: Partner Salary Null values by Partner_working before Treatment

- Clearly, those who's partners are not working would have 'Partner_salary' as 0. Let's impute with 0 for them: -

Partner_working	
Yes	16

Figure 11: Partner Salary Null values by Partner_working after Treatment

- Remaining working partners would be having, 'Partner_salary' = ('Total_salary' – 'Salary'). Since we know that 'Total_salary' & 'Salary' fields are non-null, we can reverse-compute to derive 'Partner_salary' for these 16 values.

▪ Since, we have treated 'Partner_salary', let's re-compute the Total Salary derived column as a new column 'Total_Household_Salary' & drop 'Total_salary'.

▪ Let's re-run inspection functions to validate if all the intended changes (highlighted in 'Table 3') are completed?

```
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
Price                     0
Make                      0
Age_category              0
Total_Household_Salary    0
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581 entries, 0 to 1580
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     1581 non-null   int64
 1   Gender                  1581 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   object
 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   float64
 10  Partner_salary          1581 non-null   float64
 11  Price                   1581 non-null   float64
 12  Make                    1581 non-null   object
 13  Age_category            1581 non-null   object
 14  Total_Household_Salary  1581 non-null   float64
dtypes: float64(4), int64(1), object(10)
memory usage: 185.4+ KB
```

Figure 12: Null value & Data inspection after Data-treatment

Check Statistical Summary followed by Observations & Insights

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	1581.0	NaN	NaN	NaN	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
Gender	1581	2	Male	1252	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Profession	1581	2	Salaried	896	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Marital_status	1581	2	Married	1443	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Education	1581	2	Post Graduate	985	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_Dependents	1581.0	5.0	3.0	557.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Personal_loan	1581	2	Yes	792	NaN	NaN	NaN	NaN	NaN	NaN	NaN
House_loan	1581	2	No	1054	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Partner_working	1581	2	Yes	868	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	1581.0	NaN	NaN	NaN	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1581.0	NaN	NaN	NaN	19233.776091	19670.391171	0.0	0.0	25100.0	38100.0	80500.0
Price	1581.0	NaN	NaN	NaN	35597.72296	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0
Make	1581	3	Sedan	702	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age_category	1581	3	Young (Below 31)	958	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Total_Household_Salary	1581.0	NaN	NaN	NaN	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0

Figure 13: Statistical Summary of Variables

✓ Observations & Insights can be summarized below in the table: -

Type	Columns	Observations & Insights
Numerical	Age	✓ Customer age ranges between 22 to 54 years, which can imply that most buyers belong to working age group. ✓ Mean is 31.92 & Median is 29. ✓ Standard Deviation is 8.43, which means that 68.2% of the age data falls between 23.49 & 40.35, which implies that majority of the buyers are either young or early middle-aged.
Categorical	Gender	✓ Bivariate variable with 2 values. Gender is predominantly categorized as 'Male' & 'Female'. ✓ 'Male' is the most dominant value occurring 1252/1581 in the Gender data, which implies that Male fraternity tend to buy more than the female fraternity.
Categorical	Profession	✓ Bivariate variable with 2 values. Data is divided between 'Salaried' & 'non-Salaried' (or Business) ✓ 'Salaried' is the most dominant value occurring 896/1581 in the Profession data. Customer-class is almost evenly distributed between 'Salaried' & 'Business'.
Categorical	Marital_status	✓ Bivariate variable with 2 values. Data is divided between 'Married' & 'Single'. ✓ 'Married is the most dominant value occurring 1443/1581 in the Marital_status data, which implies that most of the married individuals prefer to buy a car. This may also be attributed to the feeling of 'settling down' & as a 'need' before raising family to provide for safety.
Categorical	Education	✓ Bivariate variable with 2 values. Data is divided between 'Graduate & 'Post Graduate'. ✓ 'Post Graduate' is the most dominant value occurring 985/1581 in the Education data. Customer-class is almost evenly distributed in terms of Education. Everyone at some point needs a car, regardless of education.
Categorical	No_of_Dependents	✓ Multivariate variable with 5 values between 0 to 4. ✓ Almost one-third of the customers have 3 dependents (i.e. 557/1581).
Categorical	Personal_loan	✓ Bivariate variable with 2 values 'Yes' & 'No'. ✓ Approximately 50% of customers (i.e. 792/1581) have a Personal Loan. Having a personal loan doesn't really impact rationale to buy a car as per the data
Categorical	House_loan	✓ Bivariate variable with 2 values 'Yes' & 'No'. ✓ Approximately 66% of customers (i.e. 1054/1581) don't have a House Loan, which implies that customers having a house loan have less tendency to buy a car.
Categorical	Partner_working	✓ Bivariate variable with 2 values 'Yes' & 'No'. ✓ Approximately 55% of customers (i.e. 868/1581) have their partners working. Having a working partner doesn't really impact the rationale to buy a car as per the data.
Numerical	Salary	✓ Customers buying a car have their salaries ranging from 30K to 99.3K. ✓ Mean is 60.4K while median is 59.5K. ✓ Standard Deviation is 14.7K, which means that 68.2% of the Salary data falls between 45.7K & 75.1K, which implies that majority of the buyers earn fairly well.
Numerical	Partner_salary	✓ Customers buying a car have their Partner-salaries ranging from 0 to 80.5K, which implies that there are customers ranging from non-working partners to high-earning partners. ✓ Mean is 19.2K (less because a lot of partners don't have a salary) while median is 21.5K. ✓ Standard Deviation is 19.7K, which means that 68.2% of the Partner Salary data falls between 0 & 38.9K, which implies that majority of the buyers are either not working or not very high-earning.
Numerical	Price	✓ Customers are buying cars that range from 18K to 70K. ✓ Mean is 36.6K while median is 31.5K ✓ Standard Deviation is 13.6K, which means that 68.2% of the Pricing Data falls between 23K & 50.2K.
Categorical	Make	✓ Multivariate variable with 3 values 'SUV', 'Sedan' & 'Hatchback'. ✓ Approximately 44% of customers (i.e. 702/1581) prefer to buy a 'Sedan' over an 'SUV' & 'Hatchback'.
Categorical	Age_category	✓ Multivariate variable with 3 values 'Young', 'Middle-Age' & 'Old'. ✓ Approximately 61% of customers (i.e. 958/1581) are fairly young, aged below 31 years This also indicates the young now-a-days prefer to buy a car early in their career days itself.
Numerical	Total_Household_Salary	✓ Clientele is very diverse having a wide range of Total Salary between 30K to 171K. People with a minimum Household Salary of 30K prefer to buy a Car. ✓ Mean is 79.6K while median is 78K. ✓ Standard Deviation is 25.5K, which means that 68.2% of the data falls between 54.1K & 105.1K, which implies that majority of the buyers have a good household income to afford a car.

Table 4: Observations & Insights from Statistical Summary

Problem 2- Univariate Analysis

Analysing Numerical Variables

- Use Histograms & Boxplots to analyse each numerical variable: -

    I. Age: -

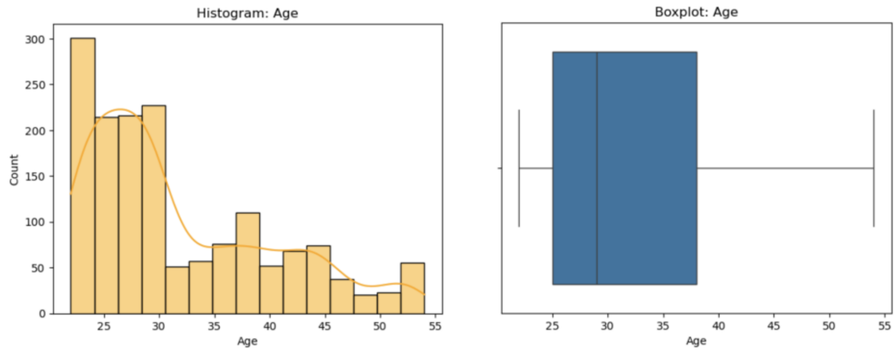


Figure 14: Age - Histogram & Boxplot

Observations: -

- ✓ Distribution of Age is highly right skewed.
- ✓ Unimodal distribution having a single peak with most no. of customers aged 22 years
- ✓ Majority of the customers have Age between 25 to 38 years
- ✓ No Outliers or negative values.
- ✓ No Outlier Treatment required.

    II. Price: -

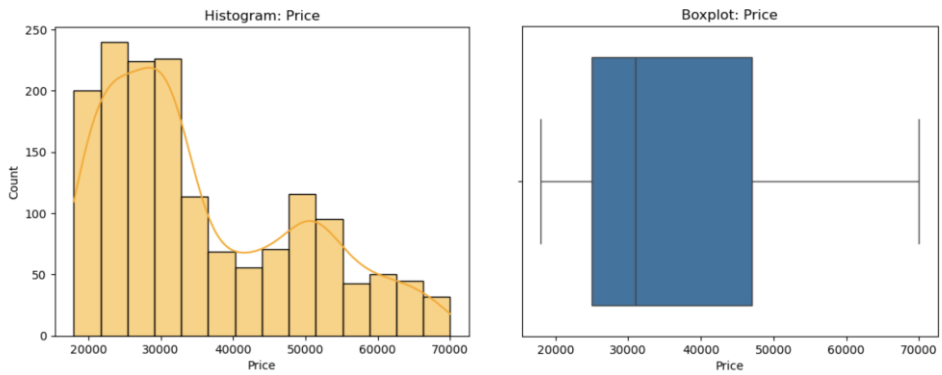


Figure 15: Price - Histogram & Boxplot

Observations: -

- ✓ Distribution of Prices of the cars bought is right skewed.
- ✓ Seems to be bimodal distribution having 2 peaks. Most no. of buyers at a price of $ 30,000, followed by, 50,000.
- ✓ Majority of the customers have bought cars in the range $ 25,000 to $ 48,000.
- ✓ No Outliers or negative values.
- ✓ No Outlier Treatment required.

III. Salary: -

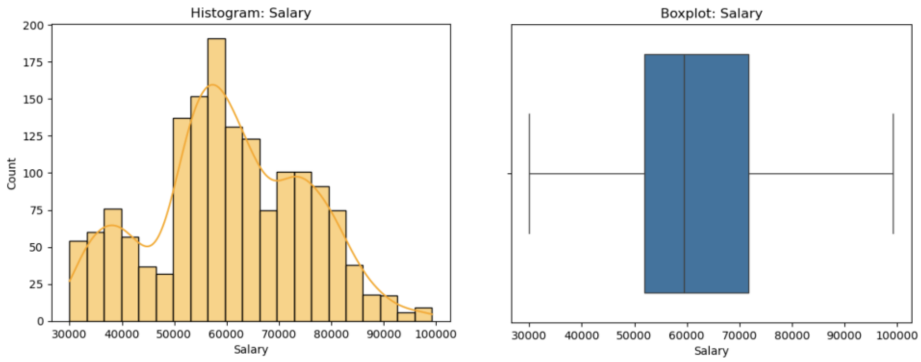


Figure 16: Salary - Histogram & Boxplot

Observations: -

- ✓ Distribution of Salary of customers seems to be symmetric with minimal skewness on either side.
- ✓ Seems to be Multimodal distribution with most of the salary of customers concentrated at $ 60,000, followed by, $ 75,000 & $ 38,000.
- ✓ Majority of the customer salaries range between $ 52,000 to $ 72,000.
- ✓ No Outliers or negative values.
- ✓ No Outlier Treatment required.

IV. Partner Salary: -

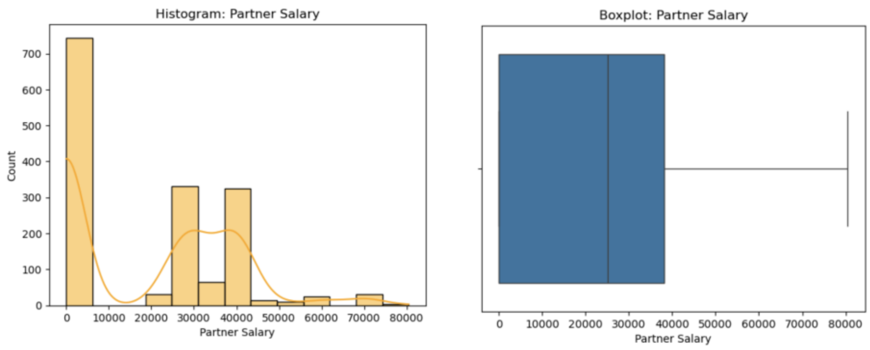


Figure 17: Partner Salary - Histogram & Boxplot

Observations: -

- ✓ Distribution of Partner Salary of customers is highly right-skewed.
- ✓ Seems to be Multimodal distribution with most of the Partner salary of customers concentrated at $ 0 (i.e. not-earning partners), followed by, $ 30,000 & $ 38,000.
- ✓ Majority of the customer Partner salaries range between $ 0 to $ 38,000.
- ✓ No Outliers or negative values.
- ✓ No Outlier Treatment required.

V. Total Household Salary: -

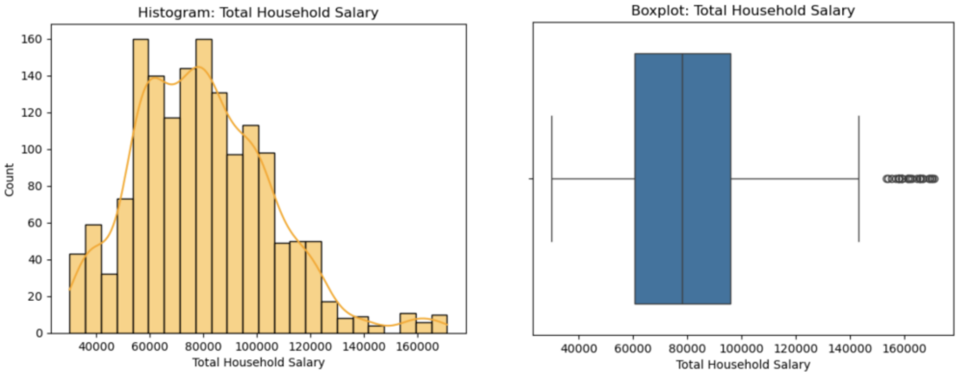


Figure 18: Total Household Salary - Histogram & Boxplot

Observations: -

- ✓ Distribution of Total Household Salary of customers seems to be symmetric with minimal skewness on either side.
- ✓ Bimodal distribution with most of the Total Household Salary of customers concentrated at $ 58,000 & $ 80,000.
- ✓ Majority of the customer Total Household Salaries range between $ 60,000 to $ 96,000.
- ✓ No negative values, but few outliers found with few customers having significantly high Total Household Salaries.
- ✓ Outlier Treatment required.

Outlier Treatment of Numerical Variables

- Based on Boxplot analysis, clearly 'Total_Hosuehold_Salary' requires outlier treatment from the visualization.
- Using the IQR Formula to detect outliers:

Lower Whisker = Q1 – (IQR)*1.5
Upper Whisker = Q3 + (IQR)*1.5

where,

Q1 – 1st Quartile
Q3 – 3rd Quartile
IQR = (Q3-Q1)

Any value below Lower Whisker & above Upper Whisker are considered as Outliers.

```
Lower Wishker at 7400.0 | Upper Whisker at 149000.0
Lower Whisker Outlier Count = 0
Upper Whisker Outlier Count = 27
Total Outlier Count= 27
Outlier Percentage in Total Household Salary= 1.707779
```

Figure 19: Total Household Salary Outlier Count/Percentage

- No outliers on the lower side, while 27 outliers on the upper side.
- Since, the Outlier percentage is high, it is not recommended to drop the rows.
- Further analysing the mean & median of the Outliers: -

```
Mean of Outliers= 162318.5185185185
Median of Outliers= 161800.0
```

Figure 20: Mean/Median of Outliers

- Mean & Median of outlier data are similar at around $ 1,62,000., which indicates individuals with a very high Total Household Salary.
- The Upper Whisker, i.e. $ 1,49,000, is already classified as a very high-income group. Therefore, replacing these 27 outlier values with the Upper Whisker Value of $ 1,49,000 would have a very minimal impact (rather, no impact!) on the overall analysis of 'Total Household Salary'.
- Displaying the Histogram & Boxplot again for the 'Total_Household_Salary' after Outlier-treatment: -

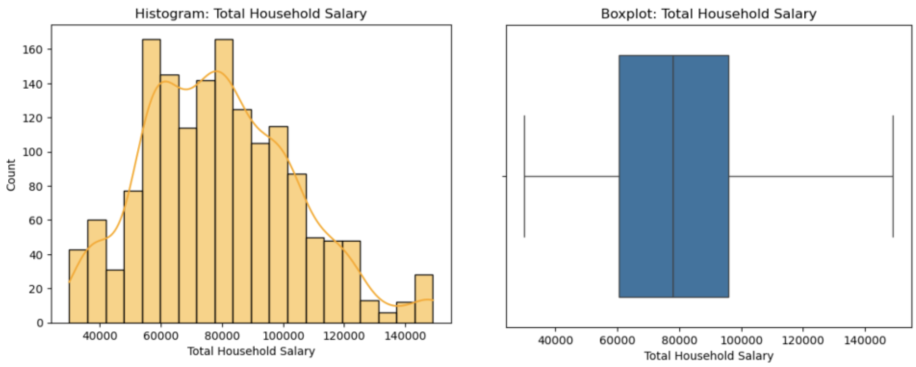
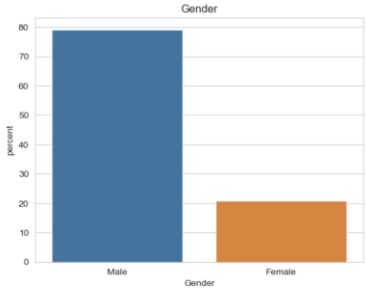
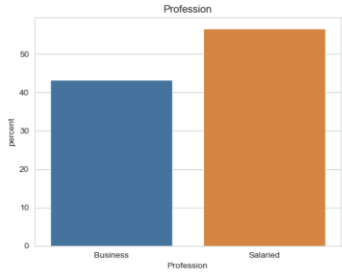


Figure 21: Total Household Salary - Histogram & Boxplot after Outlier Treatment

Analysing Categorical Variables

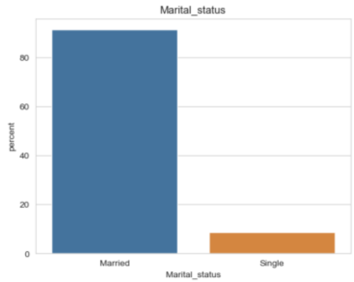
- Below are the countplots of all the categorical variables to analyse the distribution of their values: -

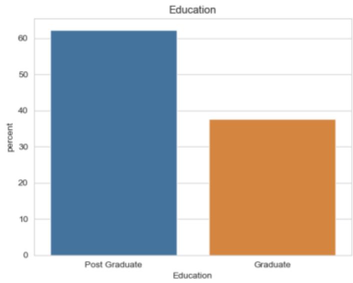





```
Gender :-

Unique Values = ['Male' 'Female']

Column Value Count Gender
Male      1252
Female     329
```

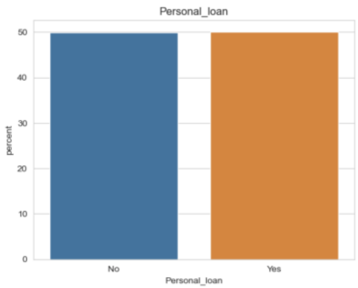
```
Profession :-

Unique Values = ['Business' 'Salaried']

Column Value Count Profession
Salaried    896
Business    685
```

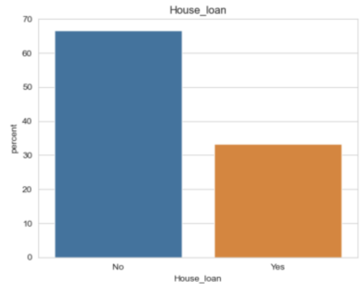





```
Marital_status :-

Unique Values = ['Married' 'Single']

Column Value Count Marital_status
Married     1443
Single       138
```

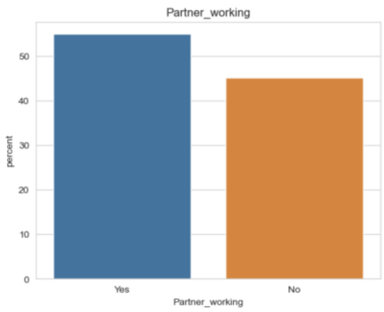
```
Education :-

Unique Values = ['Post Graduate' 'Graduate']

Column Value Count Education
Post Graduate    985
Graduate         596
```

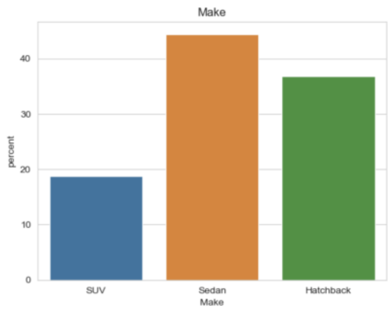





```
Personal_loan :-

Unique Values = ['No' 'Yes']

Column Value Count Personal_loan
Yes    792
No     789
```

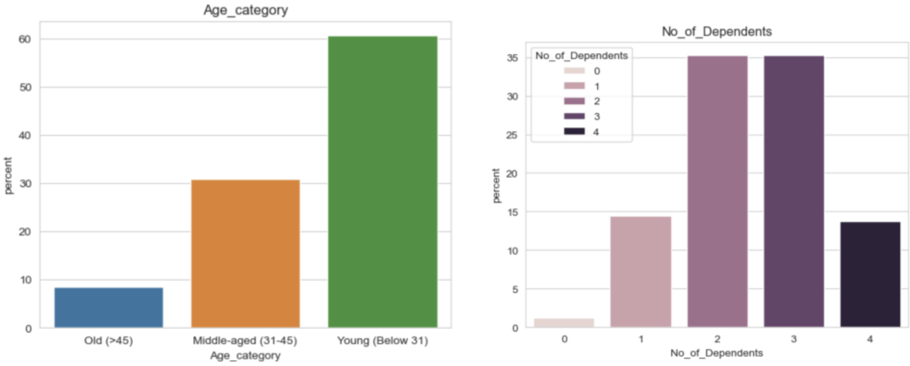
```
House_loan :-

Unique Values = ['No' 'Yes']

Column Value Count House_loan
No     1054
Yes     527
```






```
Partner_working :-

Unique Values = ['Yes' 'No']

Column Value Count Partner_working
Yes    868
No     713
```

```
Make :-

Unique Values = ['SUV' 'Sedan' 'Hatchback']

Column Value Count Make
Sedan        702
Hatchback    582
SUV          297
```




```
---------------------------------------------------------------
Age_category :-

Unique Values = ['Old (>45)' 'Middle-aged (31-45)' 'Young (Below 31)']

Column Value Count Age_category
Young (Below 31)       958
Middle-aged (31-45)    488
Old (>45)              135
```

```
---------------------------------------------------------------
No_of_Dependents :-

Unique Values = [4 3 2 1 0]

Column Value Count No_of_Dependents
3    557
2    557
1    229
4    218
0     20
```

Figure 22: Countplots of Categorical Variables

- Observations & Insights: -

Variable	Observations & Insights
Gender	▪ 90% of the customers who bought cars are males. Clearly, the Male Fraternity has more tendency to own a car.
Profession	▪ The Salaried individuals prefer personal cars over Business individuals, though the difference is not significant. The slight difference may be attributed to the fact that the Business class may be registering their vehicles in the name of the company rather than themselves (only a conjecture but no conclusive evidence!)
Marital_status	▪ Married people are more inclined to buy cars as it is clearly evident in the chart with married customer percentage being more than 90%. This may be because people tend to own a car once the get married, settled & with the feeling a raising a family in future, which would require a car for transportation & in general safety.
Education	▪ Even though the difference between the Graduate & Post-graduate is not significant, Post Graduate people have slightly more tendency to buy cars
Personal_loan	▪ Personal loan doesn't seem like impacting the decision to buy a car as the percentage for those customers with a Personal Loan is almost the same as those without a Personal Loan.
House_loan	▪ Customers who don't have a house loan are almost double the ones having a house loan, which implies, those who have an outstanding house have less tendency to buy a car as most of their salary may be going as House loan EMIs leaving no room for another big spending on car.
Partner_working	▪ More no of customers have been recorded with working partners than with those who don't have their partners working. The difference is not significant (10%), but still suggests that family with both partners working have more tendency to buy a car.
Make	▪ Sedan is the preferred type of car, followed by Hatchback & SUV. This may be attributed to the idea that Sedan positions itself at par with an SUV, but not as expensive as an SUV at the same time.
Age_category	▪ Majority of the customers belong to the 'Young' category (aged below 31 years), followed by 'Middle-aged' (aged between 31 & 45) and 'Old' (aged above 45). This may be suggestive about how the youth priorities are aligned to spend for comforts early rather than a conservative mindset of saving for future. It may also suggest that the youth earning power is fairly good that they can afford a car right at their start of their career.
No_of_Dependents	▪ Most of the car buyers have 2-3 dependents, followed by 1, 4 & 0. Very few customers have 0 dependents which implies that people usually don't prefer to go for a car if there is no dependent in the family.

Table 5: Univariate - Categorical Variable Insights

Problem 3- Bivariate Analysis

Bivariate Analysis of Numerical Variables

- Analysing the relationship & correlation of all the Numerical Variables using a Pairplot & a Heatmap: -

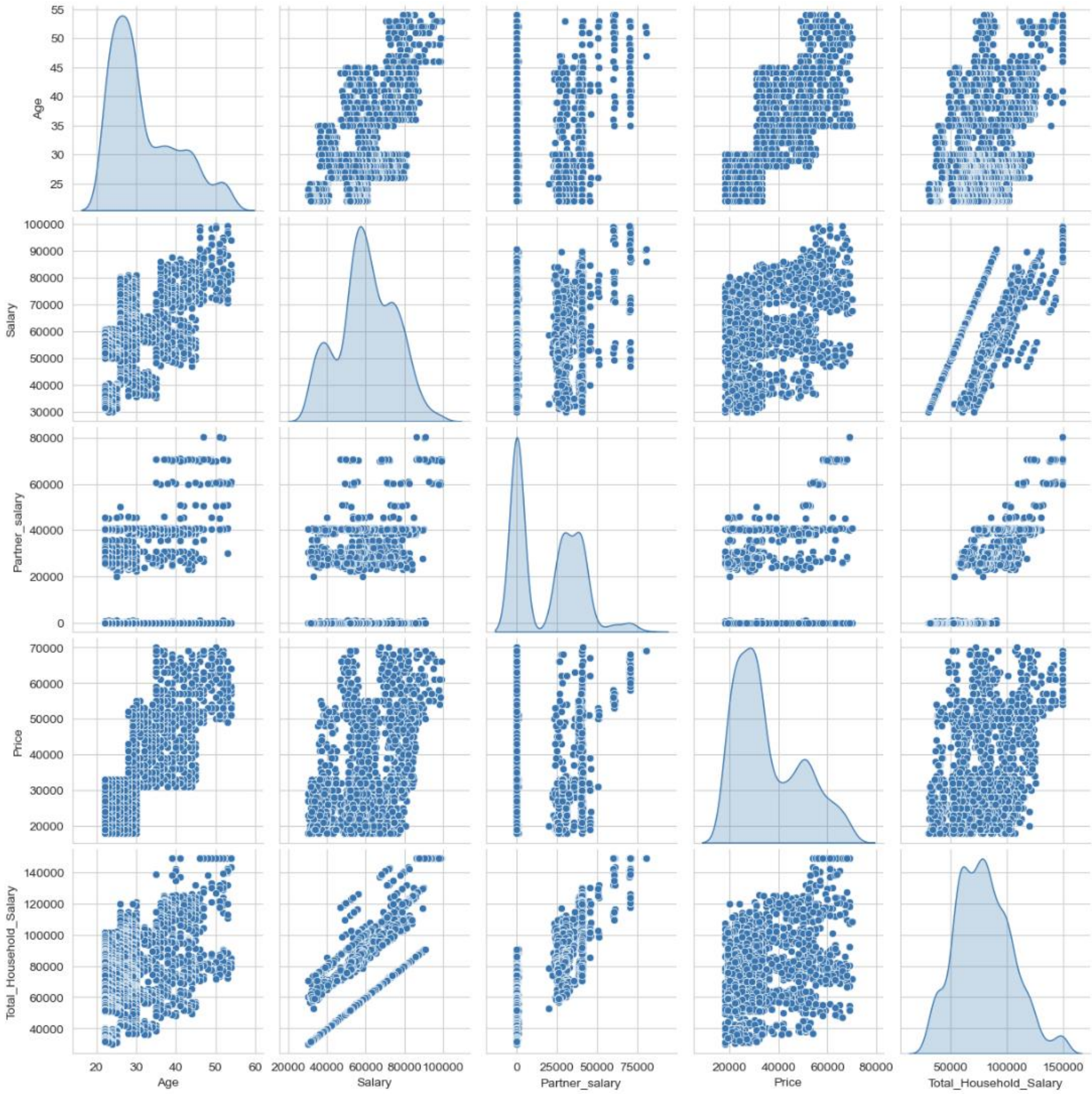


Figure 23: Pairplot of all Numerical Variables

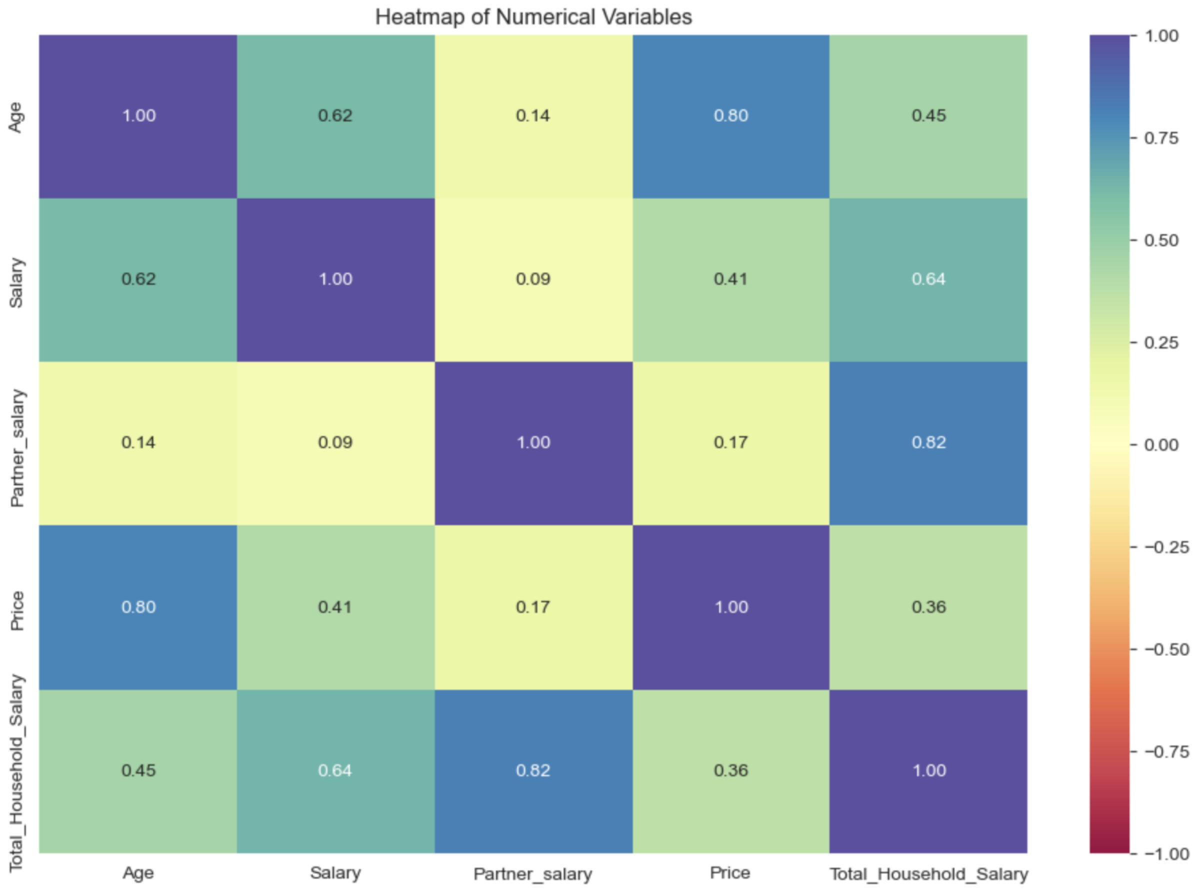


Figure 24: Heatmap of all Numerical Variables

- Observations & Insights: -
    - ✓ Correlation-coefficient (corr.) Key for inference: -
        - |Corr.| > 0.8 → Highly correlated
        - 0.5 < |Corr.| < 0.8 → Moderately correlated
        - |Corr.| < 0.5 → No decisive correlation
        - Corr. > 0 → Positive correlation
        - Corr. < 0 → Negative correlation
    - ✓ **Age Relationship with Salary (0.62):** Age shows positive correlation with Salary, which implies that higher the age higher the salary. This makes logical sense as with age, the compensation of an individual increases as his/her career grows with experience.
    - ✓ **Age Relationship with Price (0.80):** Age shows high positive correlation with the Price of the Car bought. As established above that higher the age, higher the salary; hence, higher the salary, higher the tendency to buy an expensive car. Moreover, with age, an individual tends to move to a more luxury style of living & hence the increase in the price of the car.
    - ✓ **Salary Relationship with Total Household Salary (0.64):** Salary shows positive correlation with Total Household Salary. This is obvious as we are aware that Total Household Salary is a derived attribute which is a summation of 'Salary' and 'Partner_salary'.
    - ✓ **Partner_salary Relationship with Total Household Salary (0.82):** Partner Salary shows positive correlation with Total Household Salary. This is obvious as we are aware that Total Household Salary is a derived attribute which is a summation of 'Salary' and 'Partner_salary'.

Bivariate Analysis of Categorical Variables

- Let’s carry out bivariate analysis of all categorical variables with one another using barplots: -

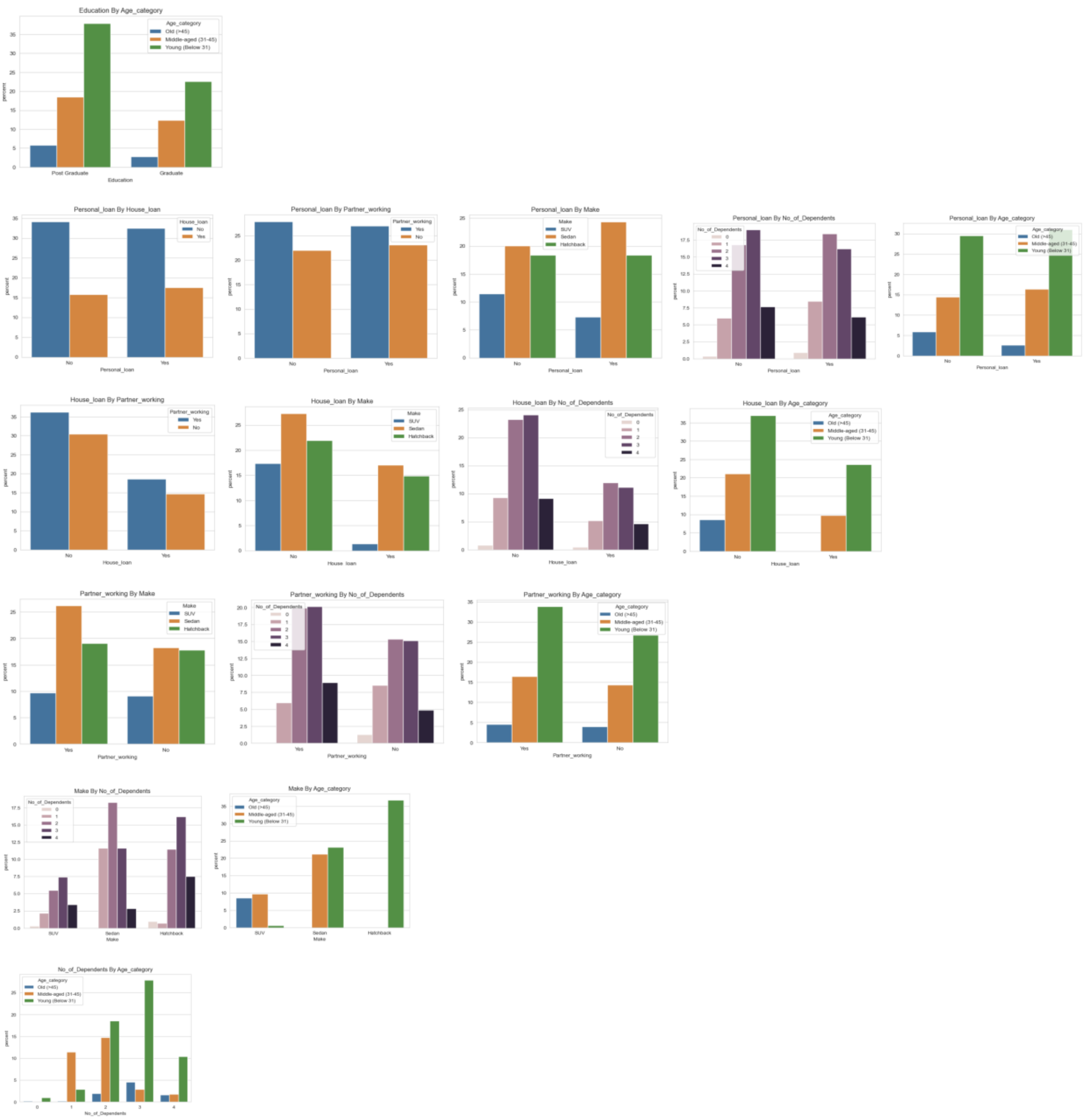


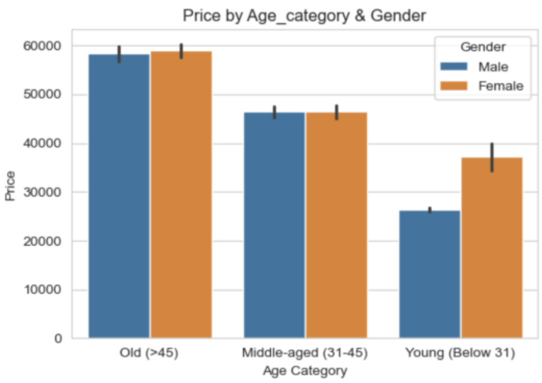
Figure 25: Bivariate Analysis of Categorical Variables

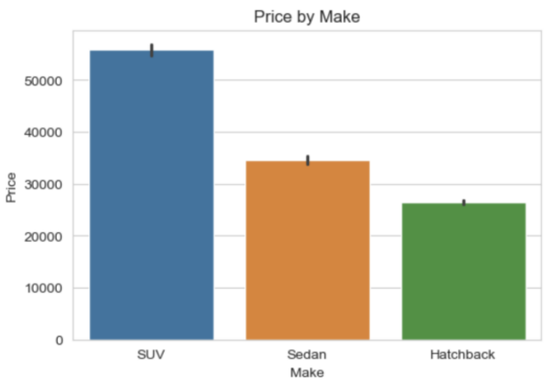
- Observations & Insights: -
  - ✓ Male is the dominant gender, meaning most of the owners are male owners.
  - ✓ Both, male & female, customers are Salaried rather than Business class. Basically, the Salaried tend to purchase cars over Business Class.
  - ✓ Married people, both male & female, tend to buy a car over Single people.
  - ✓ Post-graduate Individuals, both men & women, tend to purchase a car over Graduate individuals.
  - ✓ Males with personal loan show slightly higher tendency to purchase a car over males without personal loan. While, females without personal loan show slightly higher tendency to purchase a car over females with personal loan.
  - ✓ Individuals, men & women, without a home loan show a higher likelihood to purchase a car over individuals with a home loan.
  - ✓ Individuals, men & women, with working partners show a higher likelihood to purchase a car over individuals without a working partner.
  - ✓ Males prefer Sedans & Hatchbacks whereas Females prefer SUVs & Sedans.
  - ✓ Males having 3 dependents & females having 1 dependent show a higher likelihood of purchasing a car.
  - ✓ Male customers are mostly young (below 31 years old) while female customers are middle-aged (between 31 to 45 years old).
  - ✓ Married & post-graduate individuals, regardless of profession, have a high likelihood to purchase a car.

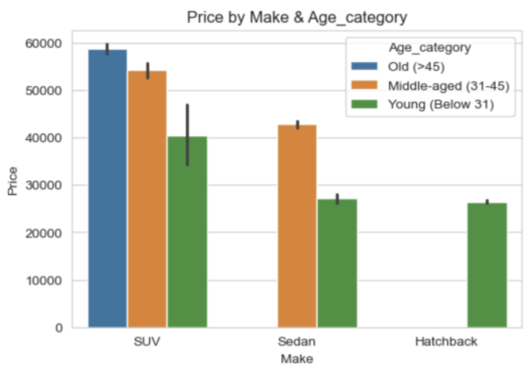
- ✓ Business class with a personal loan have a higher likelihood of purchasing a car. While, Salary class without a personal loan have a higher likelihood of purchasing a car.
- ✓ Regardless of profession, individuals with no house loan prefer to purchase a car over those who have a house loan.
- ✓ Regardless of profession, individuals with a working partner have a high likelihood to purchase a car.
- ✓ Regardless of profession, individuals prefer to buy a Sedan or a Hatchback over a SUV.
- ✓ For either profession, individuals with no of dependents 2-3 have a high likelihood to purchase a car.
- ✓ Regardless of profession, young people (below 31 years) have a higher likelihood to purchase a car.
- ✓ Sedan is the most preferred car-type for married people, whereas, Hatchback is the most preferred car-type for Single people.
- ✓ Those having a house loan, prefer not to purchase a SUV.
- ✓ Old people (above 45 years age) don't have a house loan.
- ✓ Hatchbacks are preferred by young people only (below 31 years of age). Sedans are preferred by middle-aged (between 31 to 45 years of age). SUVs are preferred by old people (above 45 years of age).

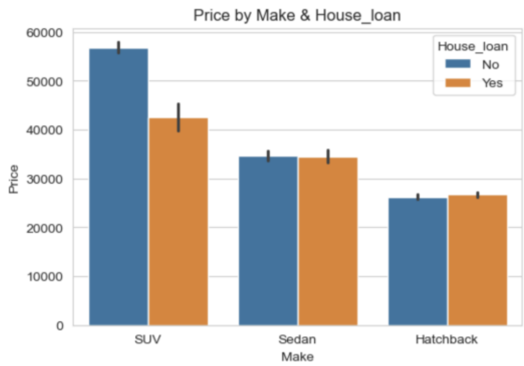
Bivariate Analysis between Categorical Variables & Numerical Variables

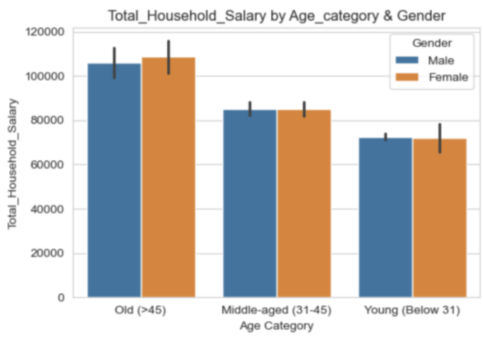
- Let's carry out bivariate analysis between categorical & numerical variables using barplots: -

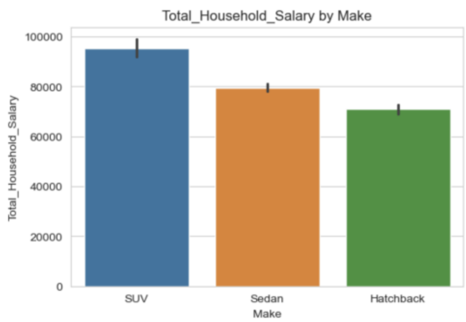












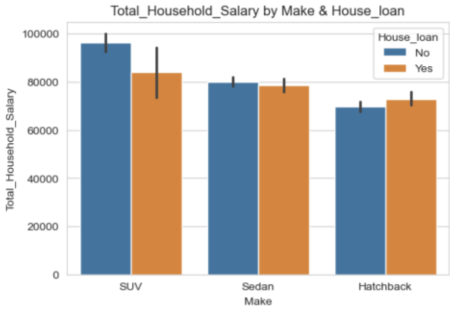


Figure 26: Bivariate Analysis between Categorical & Numerical Variables

- Observations & Insights: -
  - ✓ Overall average spending by Females is higher than that of Males in Car purchases, especially in the SUV segment.
  - ✓ Overall average spends on SUVs is higher than that of Sedans & Hatchbacks. Male spending in SUV is higher than Female, while Females are higher for Sedans. For Hatchbacks, the spending is more or less the same.
  - ✓ Old people only prefer SUVs, while young people only prefer Hatchbacks.
  - ✓ People with House loan usually buy a SUV since the average price for SUV is the highest.
  - ✓ Old people have a high Total Household salary, which is obvious as with experience, people grow in their careers & income increases eventually.
  - ✓ People with high Total household income prefer to buy SUVs over Sedans and Hatchbacks.
  - ✓ People with no housing loan (especially old people) are more likely to buy a SUV over Sedan or a hatchback.

Problem 4- Key Questions

Explore the data to answer the following questions: -

Question 1 – Do men tend to prefer SUVs more compared to women?

- Simply doing a countplot of 'Make' by 'Gender' may not give us the right results as the no. of male & female customers vary in the data set.
- To solve this problem, we would have to compute the proportion of Males & Females buying SUVs.
  - ✓ Proportion of Males who bought SUVs = (No. of Males who bought SUV / Total no. of Males in the dataset)
  - ✓ Proportion of Females who bought SUVs = (No. of Females who bought SUV / Total no. of Females in the dataset)

```
Total Males/Females in Dataset Gender
Male       1252
Female      329
```

```
Total Males/Females who bought SUV in Dataset Gender
Female    173
Male      124
```

```
Male Proportion (Percent)= Gender
Male     9.904153
```

```
Female Proportion (Percent)= Gender
Female    52.583587
```

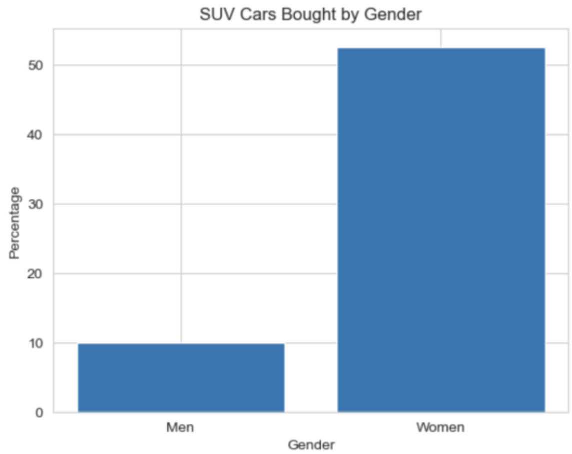


Figure 27: SUV Cars Bought by Gender

- Answer: -
  - ✓ Clearly, Men tend to prefer SUVs less in comparison to Females.
  - ✓ In the existing dataset, 9.9% men prefer SUVs whereas 52.58 females prefer SUVs.

Question 2 – What is the likelihood of a salaried person buying a Sedan?

- Let's see the distribution of Salaried people in the dataset: -

```
Profession
Salaried     896
Business     685
```

Figure 28: Distribution of Salaried People

- We further delve into this data of 896 Salaried people as only data for Salaried people is the subject of interest to solve the problem
- Let's further see the distribution of Car-types bought by the above salaried people: -

```
Make
Sedan        396
Hatchback    292
SUV          208
```

Figure 29: Distribution of Car-types bought by the Salaried People

▪ Plotting the distribution of Car types (Make) for all Salaried people: -

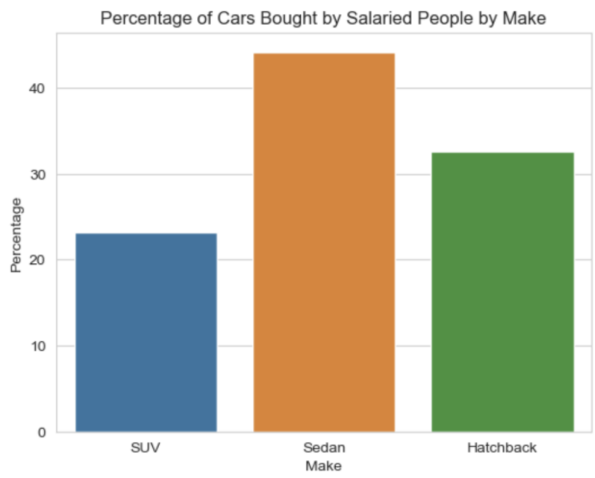


Figure 30: Percentage of Cars bought by Salaried People by Car-type

Sedan Percentage | 396/896 = 44.2%
Hatchback Percentage | 292/896 = 32.6%
SUV Percentage | 396/896 = 23.2%

▪ Let's do similar analysis for Males & Females separately: -

```
Salaried Male:- Profession
Salaried    672
```

```
Salaried Female:- Profession
Salaried    224
```

```
Salaried Male:- Make
Sedan         305
Hatchback     277
SUV            90
```

```
Salaried Female:- Make
SUV          118
Sedan         91
Hatchback     15
```

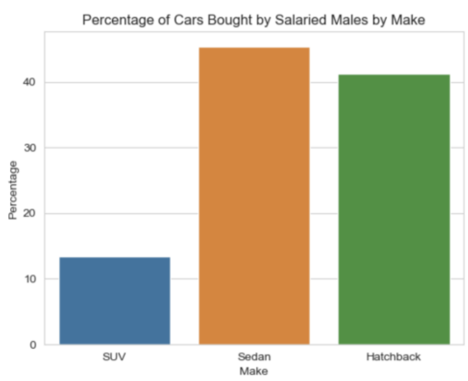


Figure 31: Cars Bought by Salaried Males by Car-type

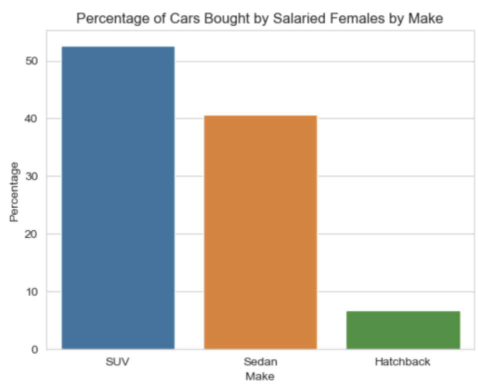


Figure 32: Cars Bought by Salaried Females by Car-type

Male Sedan Percentage | 305/672 = 45.4%
Male Hatchback Percentage | 277/672 = 41.2%
Male SUV Percentage | 90/672 = 13.4%

Female Sedan Percentage | 91/224 = 40.6%
Female Hatchback Percentage | 15/224 = 6.7%
Female SUV Percentage | 118/224 = 52.7%

Answer: -

- ✓ Likelihood of salaried people to buy Sedan is 44.2%.
- ✓ The likelihood of Salaried people buying a Sedan is more than Hatchback & SUV as evident in the visualization also.
- ✓ Male-salaried prefer Sedans with 45.4% likelihood, although hatchback likelihood is not far behind at 41.2%. On the contrary, female-salaried people prefer SUVs with 52.7% likelihood over SUVs & Hatchback.

Question 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?

- Let’s look at the distribution of Salaried Males by Make of the car: -

```
Total Salaried Male = 672
Salaried Male Counts by Make          Salaried Male Percentage by  Make
Sedan         305                     Sedan        0.453869
Hatchback     277                     Hatchback    0.412202
SUV            90                     SUV          0.133929
```

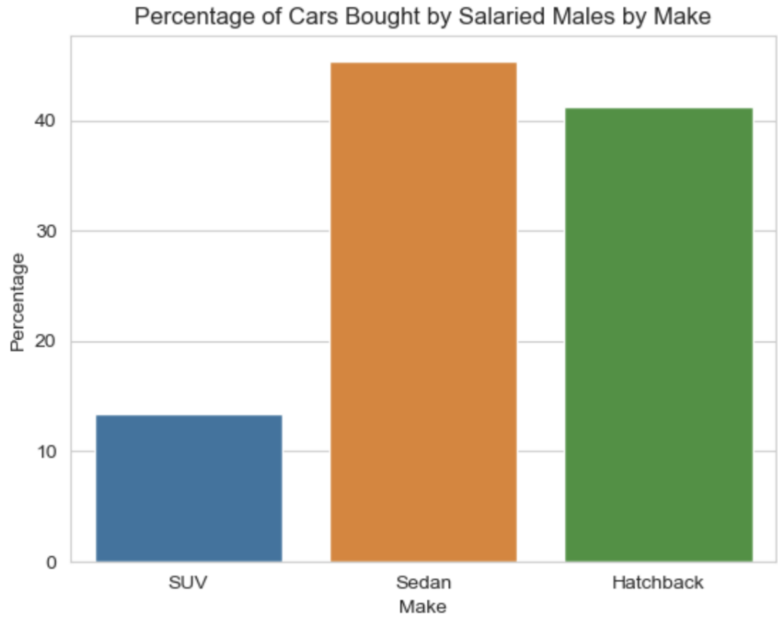
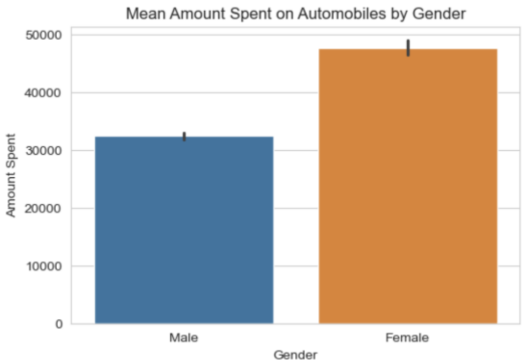


Figure 33: Percentage of Cars Bought by Salaried Males by Car-type

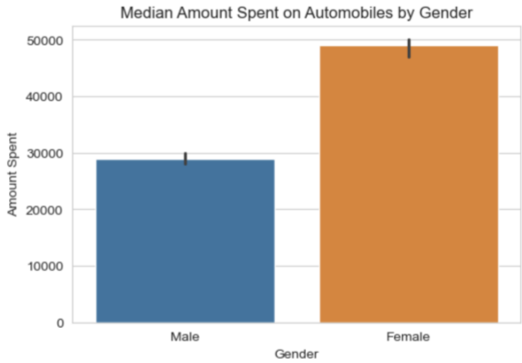
- Answer: -
  - ✓ Historical data suggests that the likelihood of a salaried-male buying a Sedan is 45.4%, which is far more than that of a SUV, which is 13.4%
  - ✓ Hence, the claim made by Sheldon Cooper that a salaried-male is an easier target for a SUV sale over a Sedan sale is not True.

Question 4 – How does the amount spent on purchasing automobiles vary by gender?

- Let's use barplots between Gender & Price to analyse the mean/median amount spent: -




```
Mean Price Spent by  Gender
Female     47705.167173
Male       32416.134185
```

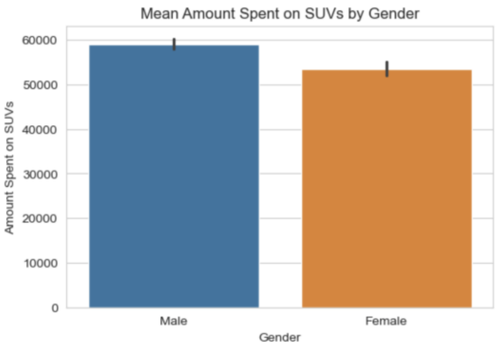



```
Median Price Spent by  Gender
Female     49000.0
Male       29000.0
```

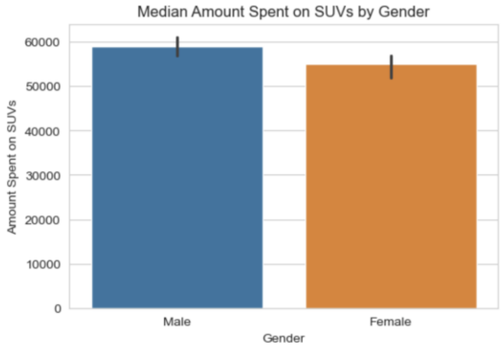
Figure 34: Mean/Median Amount Spent on Automobiles by Gender

- Extending the analysis for each Car type: -

I. SUV



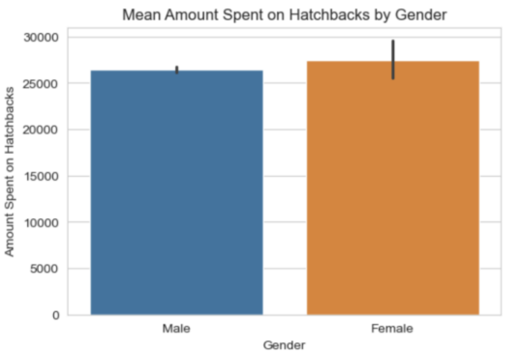

```
Mean Price Spent on SUV by  Gender
Female     53479.768786
Male       59096.774194
```



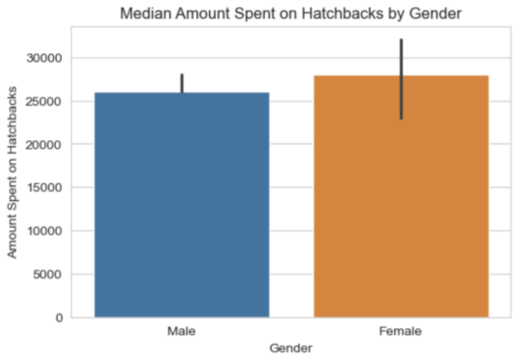

```
Median Price Spent on SUV by  Gender
Female     55000.0
Male       59000.0
```

Figure 35: Mean/Median Amount Spent on SUVs by Gender

II. Hatchback



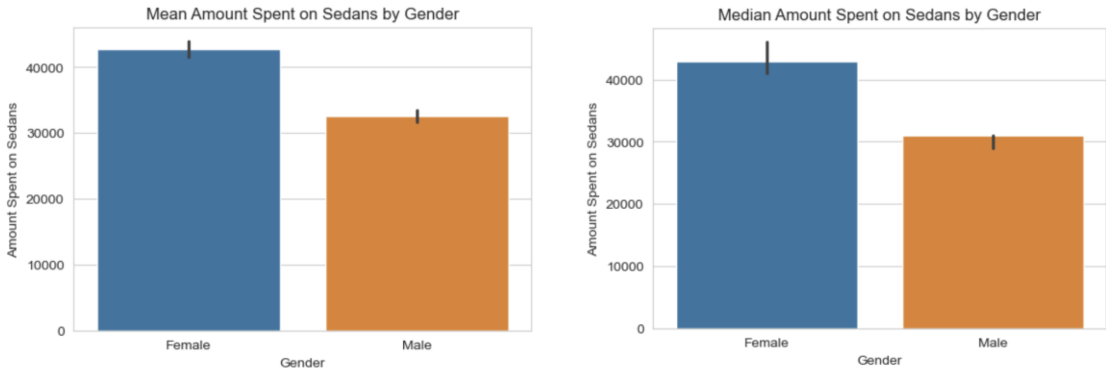

```
Mean Price Spent on Hatchbacks by  Gender
Female     27466.666667
Male       26447.971781
```




```
Median Price Spent on Hatchbacks by  Gender
Female     28000.0
Male       26000.0
```

Figure 36: Mean/Median Amount Spent on Hatchbacks by Gender

III. Sedan




```
Mean Price Spent on Sedans by  Gender
Female     42773.049645
Male       32550.802139
```

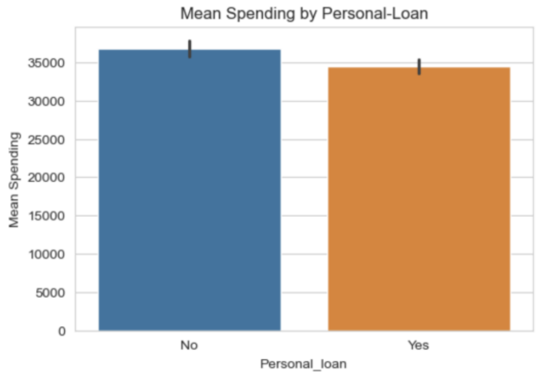
```
Median Price Spent on Sedans by  Gender
Female     43000.0
Male       31000.0
```

Figure 37: Mean/Median Amount Spent on Sedans by Gender

- Answer: -
  - ✓ Mean & Median for the Amount spent by both males & females are similar. Considering the mean amount spent, clearly the female fraternity spends more than the male fraternity [male: approx. $ 33,000 | female: approx. $ 48,000] on automobile as a whole.
  - ✓ Extending the analysis to the type of Cars: -
    - I. SUV: Males ($ 59,100) spend more than Females ($ 53,000) in the SUV segment.
    - II. Hatchback: No significant difference in the spending pattern between Males ($ 26,500) & Females ($ 27,500) in the Hatchback segment (although, the data suggest females spend more by a small amount).
    - III. Sedan: Females ($ 42,800) spend more than Males ($ 32,600) in the Sedan segment.

Question 5 – How much money was spent on purchasing automobiles by individuals who took a personal loan?

- Let's look at the distribution of amount spent between individuals who took Personal Loan as against those who did not: -




```
Count of People:- Personal_loan
Yes    792
No     789
Name: count, dtype: int64
```

```
Mean Spending:- Personal_loan
No     36742.712294
Yes    34457.070707
```

Figure 38: Mean Spending on Automobiles by Personal Loan

- Extending the analysis by Gender for those who took Personal Loan: -

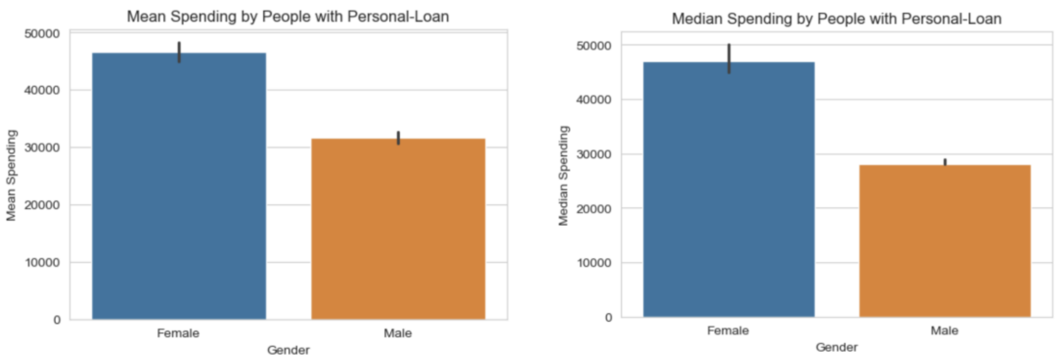


Figure 39: Mean/Median Spending on Automobiles by People with Personal Loan

- Further analysing the spending by Car type: -

Figure 40: Mean/Median Spending on Automobiles by People with Personal Loan by Car-type

- Answer: -
  - ✓ Mean Spending of people without personal loan ($ 36,742) is more than those with personal loan ($ 34457). This is attributed may be to the fact that people with personal loan under a liability avoid to spend heavily on expensive assets like cars.
  - ✓ Amongst the individuals who took personal loan, females tend to spend more than the males as clearly evident in the visualization above.
  - ✓ For both males & females, who took personal loan, the spend on SUVs is the highest, followed by Sedan & Hatchback. This is indicative of idea that SUVs are a luxury class & the most expensive compared to other Car-types, which warrants the need for a personal loan to put more liquidity in the hands of customer

Question 6 – How does having a working partner influence the purchase of higher-priced cars?

- Let's look at the overall trend of how a working partner influences the overall mean spending of an individual for initial hints: -

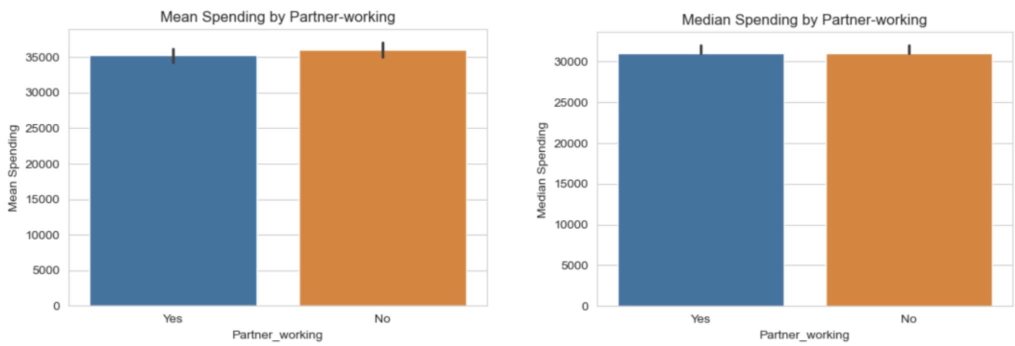
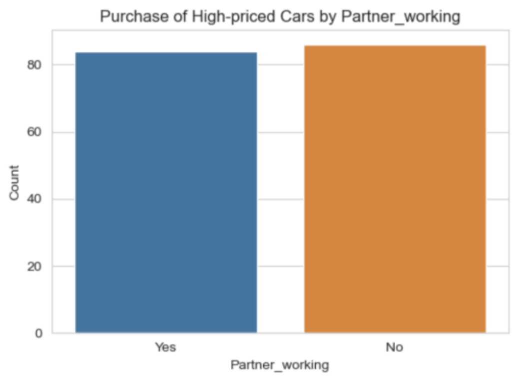


Figure 41: Mean/Median Spending on Automobiles by Partner Working

- Clearly, a working partner has no impact on the overall mean spending on cars. However, lets dig deeper by first defining the range of price for a 'higher-priced car'. By looking at the Statistical Summary of Price, we assume that cars of price above $ 55,000 can be defined as a high-priced car.
- Let's plot a Countplot to understand how many Cars with price above $ 55,000 (i.e. high-priced cars) have been sold & how they are distributed between working vs non-working partners: -




```
High-priced Car Purchases (>Rs. 55,000) by  Partner_working
No     86
Yes    84
```

Figure 42: Purchase of High-priced Cars by Partner Working

- Answer:-
  - ✓ Mean & Median Spending for working & non-working partners are almost the same. This shows that overall there is no impact of partner working on the sales of the Car.
  - ✓ After defining the bracket for high-priced cars (i.e. above $ 55,000), we found the the no. of purchases are similar for both type of individuals (with partner working & without partner working).
  - ✓ Hence, we conclude that there is no influence, a working partner exerts on purchasing a higher-priced car.

Problem 5- Actionable Insights & Recommendations

- Let's segment customers based on Gender & Marital Status so that each segment can be positioned based on their buying behaviour: -

```
Gender  Marital_status  Make
Female  Married         Hatchback     14
                        SUV          166
                        Sedan        127
        Single          Hatchback      1
                        SUV            7
                        Sedan         14
Male    Married         Hatchback    484
                        SUV          115
                        Sedan        537
        Single          Hatchback     83
                        SUV            9
                        Sedan         24
```

```
Gender  Marital_status
Female  Married               SUV
        Single              Sedan
Male    Married             Sedan
        Single          Hatchback
```

Figure 43: Customer Segmentation by Gender & Marital Status

✓ We can segment our customers in 4 basic categories based: (1) Female Married (2) Female Single (3) Male Married (5) Male Single: -

Segment	Actionable Insights & Recommendations
Married Female	✓ Prefer SUVs. ✓ Tailor marketing strategies to target Married Females to increase sales for SUVs, which would also mean higher revenue for the company as SUVs are highly priced
Single Female	✓ Prefer Sedans. ✓ Tailor marketing strategies to target Single Females to increase sales for Sedans
Married Male	✓ Prefer Sedans. ✓ Tailor marketing strategies to target Single Females to increase sales for Sedans
Single Male	✓ Prefer Hatchbacks. ✓ Tailor marketing strategies to target Single Females to increase sales for Hatchbacks

Table 6: Customer Segmentation with Insights/Recommendations

Based on each segment preferences, marketing strategies should be positioned to increase the no. of units sold for each Car-type. This would eventually contribute to increase in market share as the likelihood of a car being bought from the company will increase if the right segment is targeted with the right model & strategy.

- Let's analyse the data by checking the Amount spent for each Car-type: -

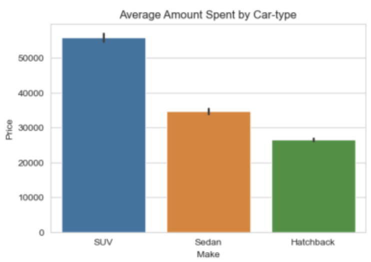


Figure 44: Amount Spent by Car-type

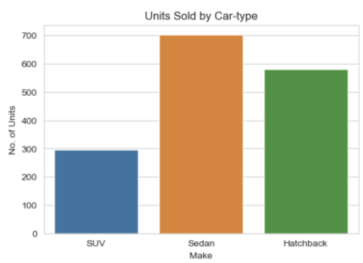


Figure 45: Unit Sold by Car-type

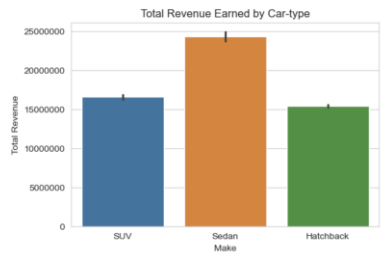


Figure 46: Total Revenue Earned from each Car-type

✓ Clearly, SUV is the most expensive of all the car-types & higher sales of SUV means higher revenue for the company. Since, SUVs are a luxury segment with a high price, the units sold are the lowest. SUV-segment is an untapped segment & increasing sales would mean increase in the revenue for the company. The company should target 'Married Females' (refer Table 6), who prefer SUVs over other car-types, aggressively by campaigning & creating promotional offers for SUVs targeted at married females.

✓ From the visualization it is evident that Sedan segment is the 'Cash-cow' for the company not only in terms of units sold but in total revenue. With reference to Table 6, it is clear that the company should continue to target 'Single Females' & 'Married Males', who have a high likelihood to buy a Sedan. This marketing strategy is required to sustain the success in the Sedan segment, rather strengthen by further adopting promotional offers for these customer segments to sustain the success in the Sedan-class.

✓ Hatchback is the cheapest segment and even though the units sold are decent in comparison to Sedans, still it offers an opportunity area to increase the overall revenue. Hatchback segment should increase its no. of sales so that the company can improve on the overall revenue (increase revenue by increasing the volume of sales). As per Table 6, 'Single Males' are more likely to buy a Hatchback, hence, an aggressive marketing strategy should be launched convincing Single Males to buy cars for their daily commute.

- Let's analyse Car-type by Age-category, Gender & Marital Status: -

```
Age_category         Gender  Marital_status
Middle-aged (31-45)  Female  Married                Sedan
                             Single                 Sedan
                     Male    Married                Sedan
                             Single                 Sedan
Old (>45)            Female  Married                  SUV
                             Single                   SUV
                     Male    Married                  SUV
                             Single                   SUV
Young (Below 31)     Female  Married                Sedan
                             Single                 Sedan
                     Male    Married            Hatchback
                             Single             Hatchback
```

Figure 47: Car-type by Age-category, Gender & Marital Status

- ✓ Above figure shows the likelihood of an individual buying a Car-type by age & gender. Positioning strategies as per the preferences outlined above would increase the likelihood of increasing sales.
- ✓ For Sedans, target middle-aged individuals & young females.
- ✓ For SUVs, target old individuals.
- ✓ For Hatchbacks, target young males.

- Let's analyse Car type by House Loan. Below table shows the likelihood of an individual to buy a car-type, based on an existing house loan: -

```
House_loan  Gender  Age_category
No          Female  Middle-aged (31-45)          SUV
                    Old (>45)                    SUV
                    Young (Below 31)       Hatchback
            Male    Middle-aged (31-45)        Sedan
                    Old (>45)                    SUV
                    Young (Below 31)       Hatchback
Yes         Female  Middle-aged (31-45)        Sedan
                    Young (Below 31)           Sedan
            Male    Middle-aged (31-45)        Sedan
                    Young (Below 31)       Hatchback
```

Figure 48: Car type by House Loan

- ✓ Below Actionable Insights/Recommendations can be drawn from the above analysis: -

Segment	Sub-segment	Actionable Insights & Recommendations
Individuals without House Loan	Females	✓ Target SUVs to middle-aged & old females without a house loan. ✓ Target Hatchbacks to young females without a house loan.
Individuals without House Loan	Males	✓ Target SUVs to old males without a house loan. ✓ Target Hatchbacks to young males without a house loan. ✓ Target Sedans to middle-aged males without a house loan.
Individuals with House Loan	Females	✓ Target Sedans to young & middle-aged females with a house loan ✓ Old females with a house loan seldom buy a car
Individuals with House Loan	Males	✓ Target Sedans to middle-aged males with a house loan. ✓ Target Hatchbacks to young males with a house loan. ✓ Old males with a house loan seldom buy a car

Table 7: Insights/Recommendations on Car-type by House Loan

- Let's analyse the Car-type by Mean of Total Household Salary: -

Make	Gender	
Hatchback	Female	76006.66
	Male	70886.06
SUV	Female	94237.57
	Male	96975.80
Sedan	Female	81734.04
	Male	79044.56

Figure 49: Car-type by Total Household Salary

- ✓ Regardless of the Gender, Total Household Income is linked to the Price of the Car (SUV being the most expensive, followed by Sedan & hatchback).
- ✓ Individuals with Total Household Salary above $ 90,000 prefer SUVs.
- ✓ Individuals with Total Household Salary around $ 80,000 prefer SUVs
- ✓ Individuals with Total Household Salary around $ 70,000-75,000 prefer Hatchbacks