Austo Motor Company

Business Report

PGP - DSBA

Moumit Manuel Dingdoh

Content

List of Figures

List of Tables

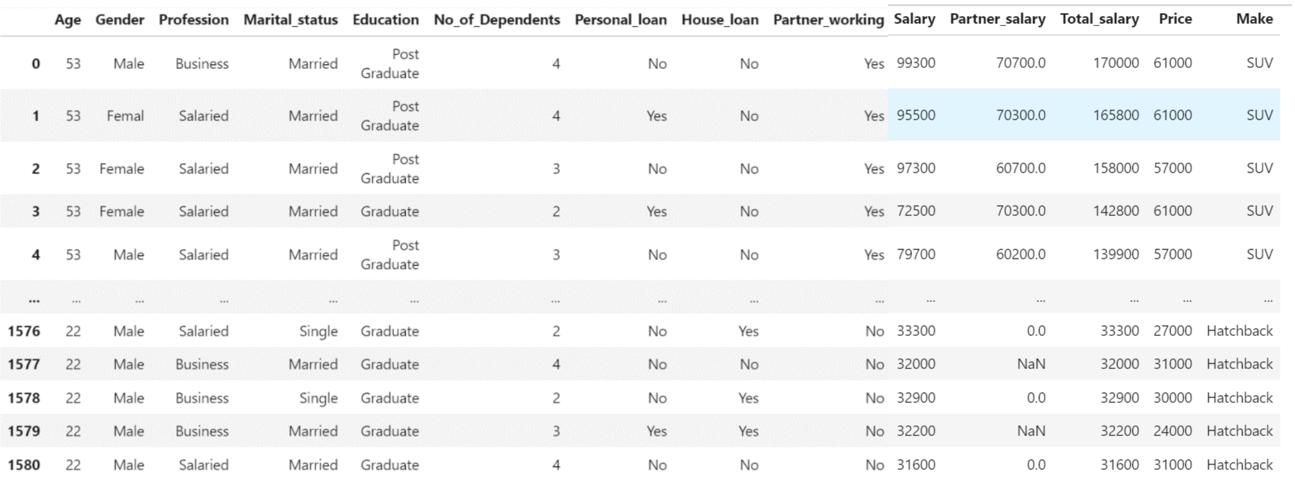
3. Data Descriptions and Data Overview

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.

Data Description

- **Age**: The age of the individual in years.
- **Gender**: The gender of the individual (male or female).
- **Profession**: The profession of the individual.
- **Marital_status**: The marital status (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/cars

3.1. Check the structure of the data

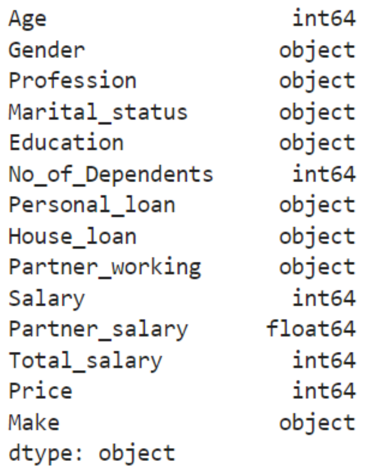


	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	Price	Make
0	53	Male	Business	Married	Post Graduate	4	No	No	Yes	99300	70700.0	170000	61000	SUV
1	53	Femal	Salaried	Married	Post Graduate	4	Yes	No	Yes	95500	70300.0	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1576	22	Male	Salaried	Single	Graduate	2	No	Yes	No	33300	0.0	33300	27000	Hatchback
1577	22	Male	Business	Married	Graduate	4	No	No	No	32000	NaN	32000	31000	Hatchback
1578	22	Male	Business	Single	Graduate	2	No	Yes	No	32900	0.0	32900	30000	Hatchback
1579	22	Male	Business	Married	Graduate	3	Yes	Yes	No	32200	NaN	32200	24000	Hatchback
1580	22	Male	Salaried	Married	Graduate	4	No	No	No	31600	0.0	31600	31000	Hatchback

Tab. 3.1 Top 5 and Bottom 5 records

- There are 1581 number of rows and 14 number of columns present in the dataset.
- In the Fig 3.1, it shows top 5 and bottom 5 records of the dataset.

3.2. Check the type of the data



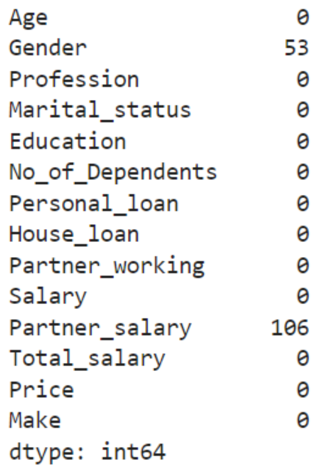
```
Age                    int64
Gender                object
Profession            object
Marital_status        object
Education             object
No_of_Dependents       int64
Personal_loan         object
House_loan            object
Partner_working       object
Salary                 int64
Partner_salary       float64
Total_salary           int64
Price                  int64
Make                  object
dtype: object
```

Tab. 3.2 Datatypes

- Only 6 variables (Age', 'No_of_Dependents', 'Salary', 'Partner_salary', 'Total_salary', 'Price') were numerical.

- Only 8 variables ('Gender', 'Profession', 'Marital_status', 'Education', 'Personal_loan', 'House_loan', 'Make', 'Partner_working') were categorical.

3.3. Check for and treat (if needed) missing value



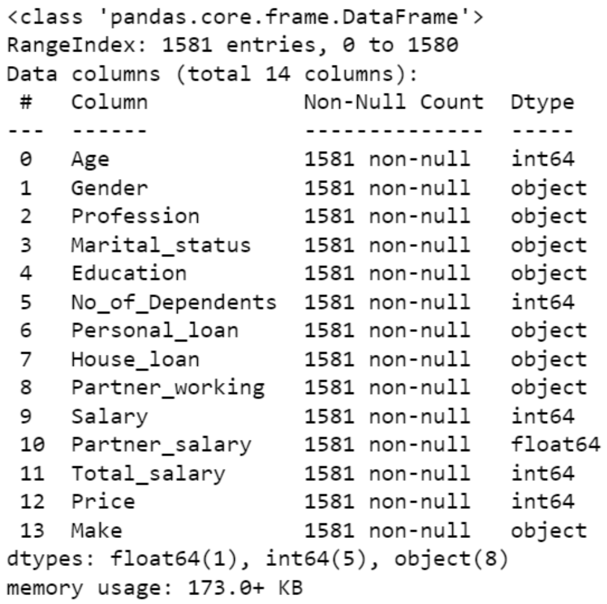
```
Age                   0
Gender               53
Profession            0
Marital_status        0
Education             0
No_of_Dependents      0
Personal_loan         0
House_loan            0
Partner_working       0
Salary                0
Partner_salary      106
Total_salary          0
Price                 0
Make                  0
dtype: int64
```

Tab. 3.3 Missing values before treatment

- There were 53 missing values in Gender and 106 missing values in Partner_salary variables.

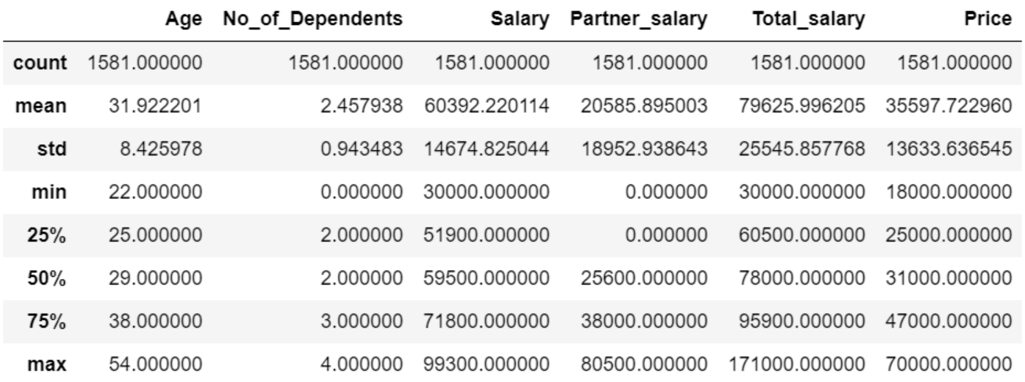
- To treat missing values of the dataset, the missing values of the Gender were replaced with mode() function and the missing values of the Partner_salary were replaced with median() function.

3.4. Check the statistical summary



```
<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            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   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    1581 non-null   float64
 11  Total_salary      1581 non-null   int64  
 12  Price             1581 non-null   int64  
 13  Make              1581 non-null   object 
dtypes: float64(1), int64(5), object(8)
memory usage: 173.0+ KB
```

Tab. 3.4 Information of the data

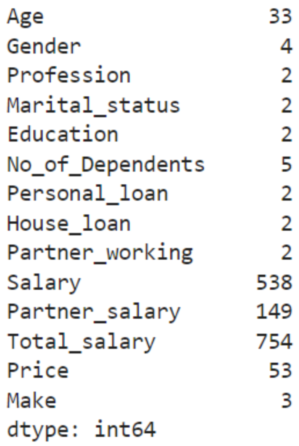


	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
count	1581.000000	1581.000000	1581.000000	1581.000000	1581.000000	1581.000000
mean	31.922201	2.457938	60392.220114	20585.895003	79625.996205	35597.722960
std	8.425978	0.943483	14674.825044	18952.938643	25545.857768	13633.636545
min	22.000000	0.000000	30000.000000	0.000000	30000.000000	18000.000000
25%	25.000000	2.000000	51900.000000	0.000000	60500.000000	25000.000000
50%	29.000000	2.000000	59500.000000	25600.000000	78000.000000	31000.000000
75%	38.000000	3.000000	71800.000000	38000.000000	95900.000000	47000.000000
max	54.000000	4.000000	99300.000000	80500.000000	171000.000000	70000.000000

Tab. 3.4.1 Statistical descriptive of the numerical variables

- The Age of the individuals in the dataset ranges between 22 and 54.
- The average ranges of the No_of_dependents between 2-3.
- The 25% of the individual’s Salary is 51,900.
- The 75% of the Partner_salary is 38,000.
- The 50% of the individual’s Total_salary is 78,000.
- The minimum Price of a car is 18,000 and the maximum price of a car is 70,000.
- The distribution of the Age lies between 25% and 75% percentiles, indicating the individuals were younger.

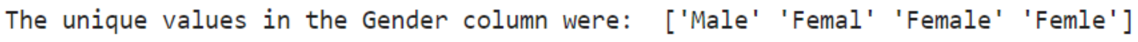
3.5. Check for and treat (if needed) data irregularities



```
Age                   33
Gender                 4
Profession             2
Marital_status         2
Education              2
No_of_Dependents       5
Personal_loan          2
House_loan             2
Partner_working        2
Salary               538
Partner_salary       149
Total_salary         754
Price                 53
Make                   3
dtype: int64
```

Tab 3.5 Checking data irregularities

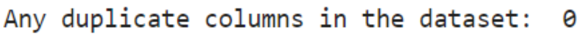
- In the Tab 3.5, Gender column consists of 4 unique values, which was incorrect as the dataset must contains only 2 unique values, i.e., Male and Female.
- To check the values of the Gender column, the unique values consist of Male, Female, Femal, and Femle. Refer the Tab. 3.5.1.



```
The unique values in the Gender column were:  ['Male' 'Femal' 'Female' 'Femle']
```

Tab 3.5.1 Unique values of the Gender column

- To treat the data irregularity, the values (Femle and Femal) was replaced with Female in the Gender column.



```
Any duplicate columns in the dataset:  0
```

Tab 3.5.2 Checking for duplicate values

- There are no duplicate values in the dataset and now, the dataset is ready for the further analysis.

3.6. Observations and Insights

- The Numerical columns were ‘Age’, ‘No_of_Dependents’, ‘Salary’, ‘Partner_salary’, ‘Total_salary’, and ‘Price’.
- The 8ategorical columns were ‘Gender’, ‘Profession’, ‘Marital_status’, ‘Education’, ‘Personal_loan’, ‘House_loan’, ‘Partner_working’, and ‘Make’.
- Missing values were observed in Gender and Partner_salary. Columns.
- There were no duplicate values present in the dataset.
- Data irregularities were present in Gender columns and was treated with the correct value.
- The dataset was structured, handled all the missing values of the dataset and corrected the irregularities of the data to perform the analysis in the next further step.

4. Univariate Analysis

4.1. Explore all the variables (categorical and numerical) in the data

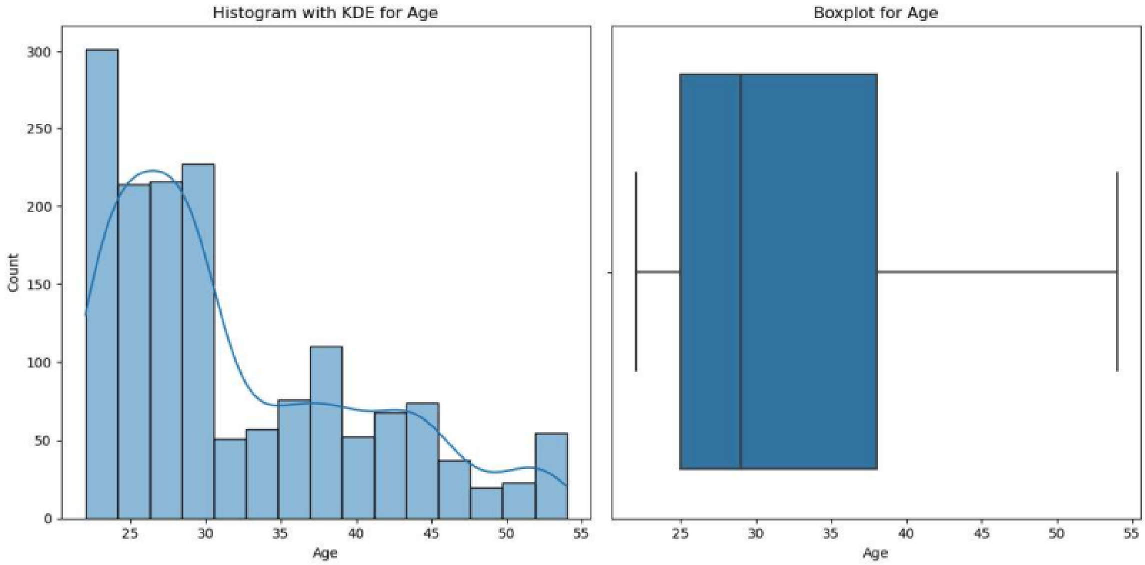


Fig. 4.1 Histogram and Boxplot of Age

- The Age in the Histogram plot shows that maximum number of individuals lies up to 30 years of age.
- It is right-skewed distribution with the mean lies between 25 and 30 as shown in the Boxplot.

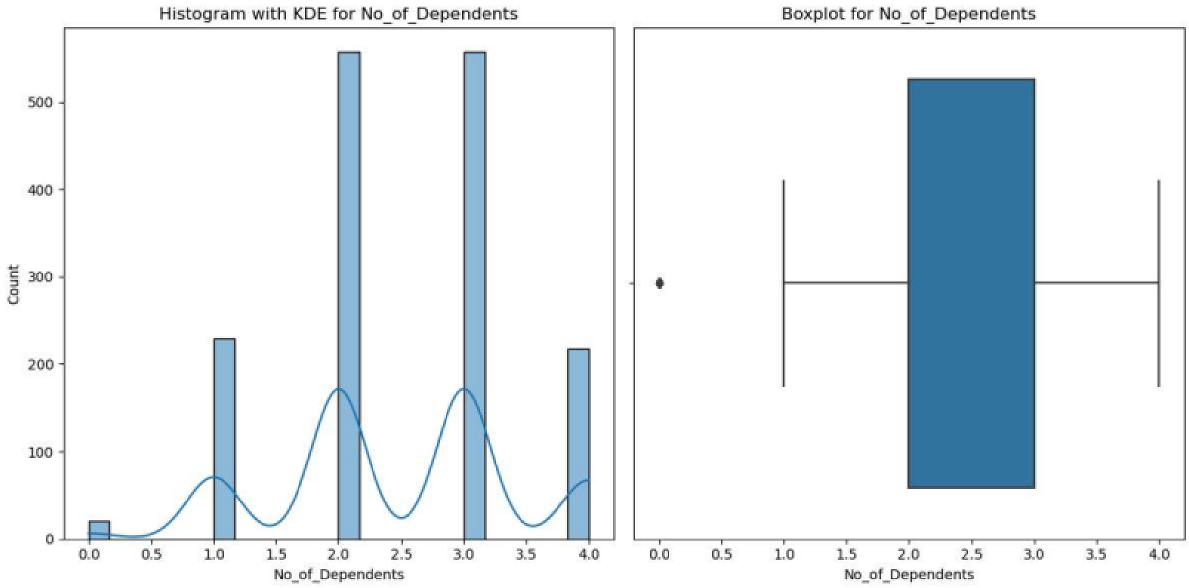


Fig. 4.1.2 Histogram and Boxplot of No_of_Dependents

- The maximum number of dependents is 2 and 3 as shown in the Histogram plot.
- The Boxplot shows that it is a left-skewed distribution.
- An outlier was detected on the lower whiskers in the Boxplot.

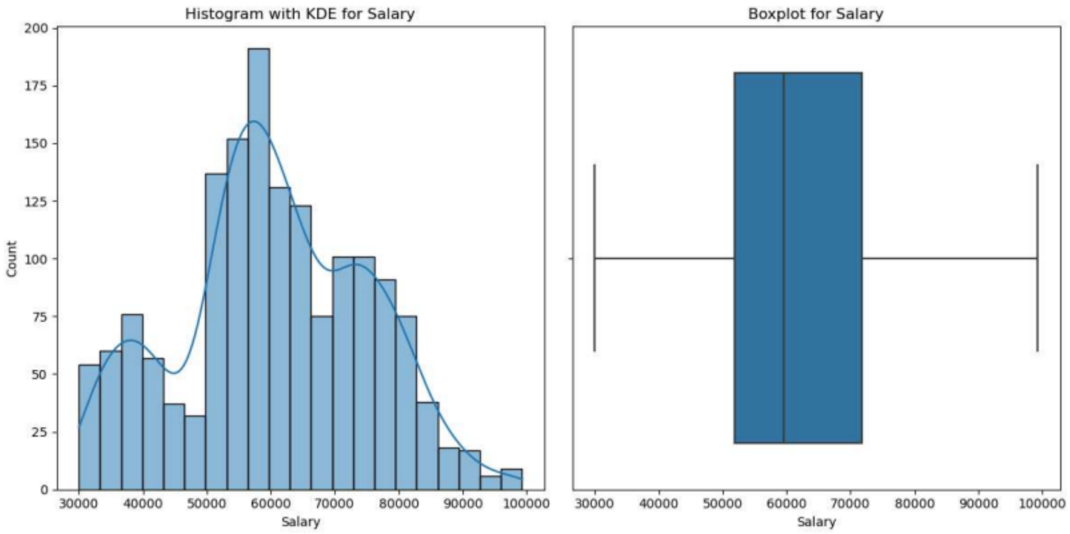


Fig. 4.1.3 Histogram and Boxplot of Salary

- In the Boxplot, it is right-skewed distribution.
- The maximum salary in the Histogram lies between 50,000 to 65,000.
- There is no outlier in the Salary column.

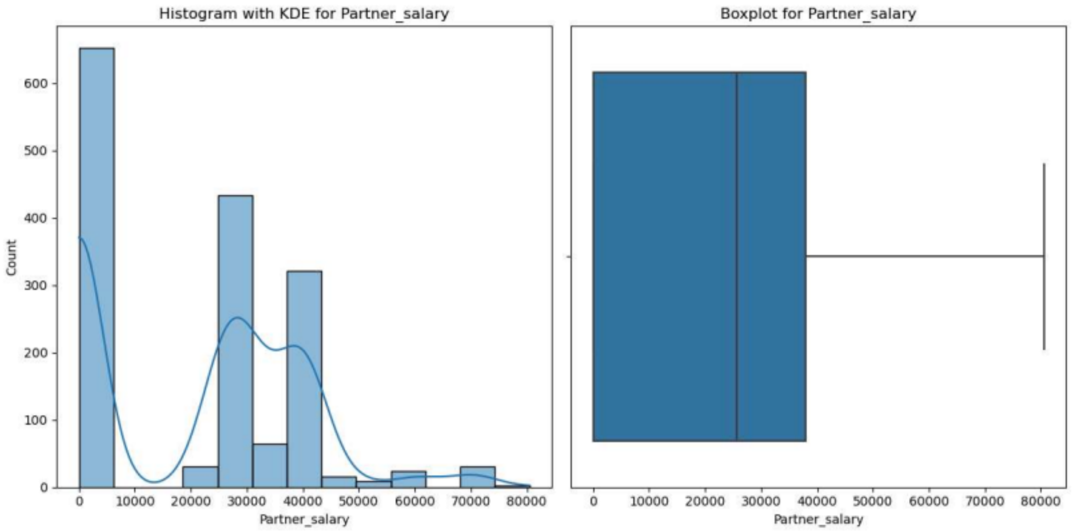


Fig. 4.1.4 Histogram and Boxplot of Partner_salary

- In the Boxplot, it is right-skewed distribution as shown in the Fig 4.4.
- The maximum number of values in the Partner_salary lies in 0, 30,000 and 40,000.
- No outlier was detected in the plot.

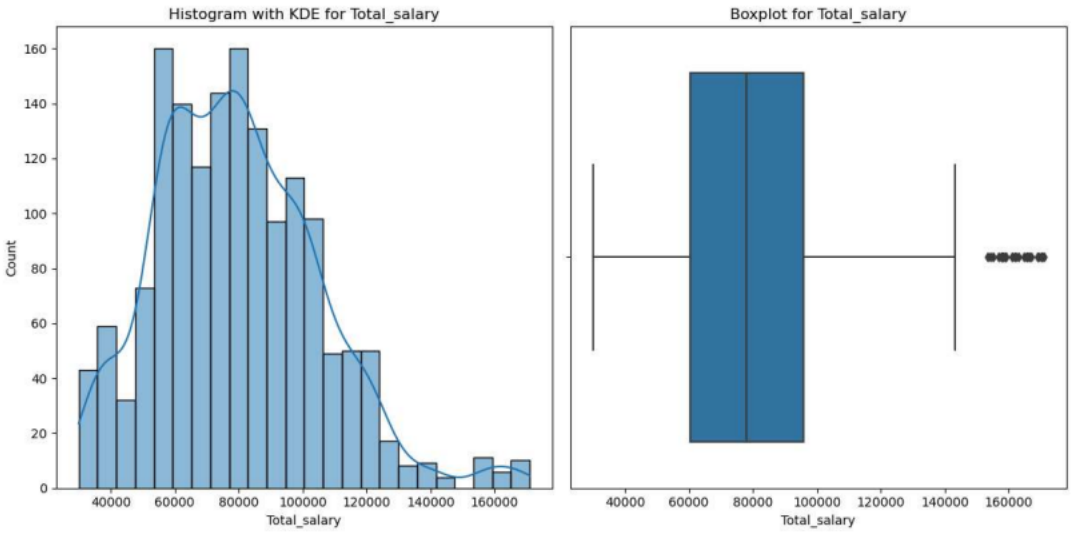


Fig. 4.1.5 Histogram and Boxplot of Total_salary

- The Total_salary column in the Boxplot is right-skewed distribution.
- The maximum Salary including the partner's income lies between 60,000 and 90,000.
- Multiple outliers were detected on the upper whiskers in the Boxplot.

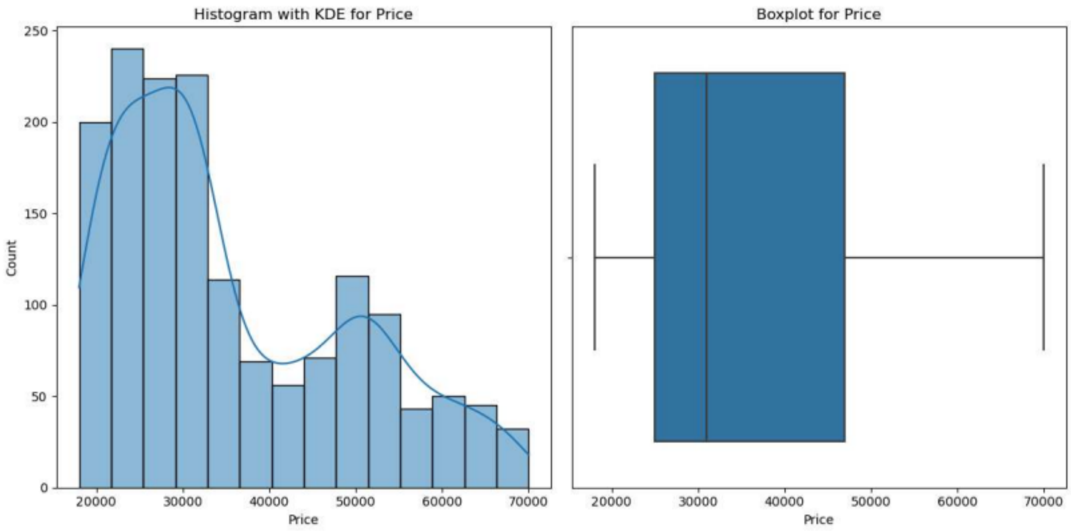


Fig. 4.1.6 Histogram and Boxplot of Price

- In the Boxplot, it is right-skewed distribution.
- The maximum price lies between 20,000 and 35,000.
- No outlier was detected in the Boxplot.

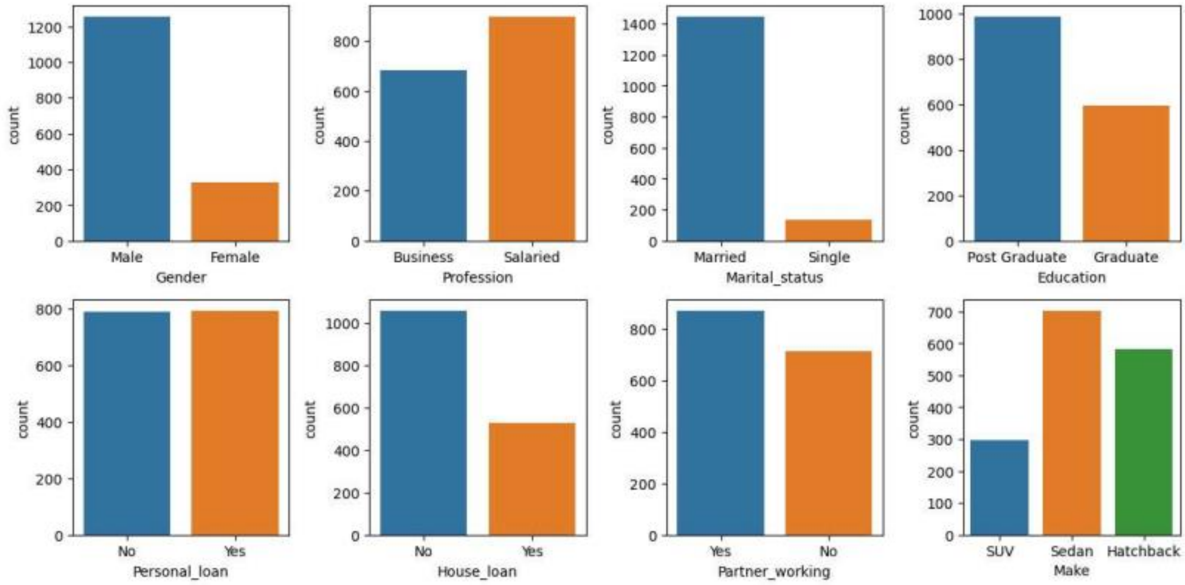
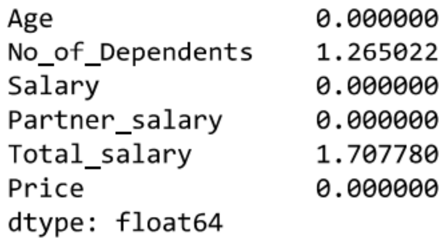


Fig. 4.1.7 Countplots of all Catagorical columns

- Maximum number of Male purchased the automobiles in compared with the Female.
- Salaried individuals tend to purchase more than the person who do Businesses.
- Maximum number of the car purchased is Married.
- Individuals who completed the Post graduation likely to buy more than Graduates.
- Equal distribution of the individuals who has Personal loan (Yes/No) purchased a car.
- Those individuals who does not have House loan purchased the automobiles.
- More their Partner’s working, more the individuals purchased the automobiles.
- The company manufactures more Sedan and Hatchbacks in compare with the SUV.

4.2. Check for and treat (if needed) outliers



```
Age                 0.000000
No_of_Dependents    1.265022
Salary              0.000000
Partner_salary      0.000000
Total_salary        1.707780
Price               0.000000
dtype: float64
```

Tab. 4.2 Check the outlier percentage

- The outliers were present in the No_of_Dependents and Total_salary columns.
- The outlier percentages of those 2 columns were 12.65% and 17.07%.
- The treatment of the outliers was done in the next step.

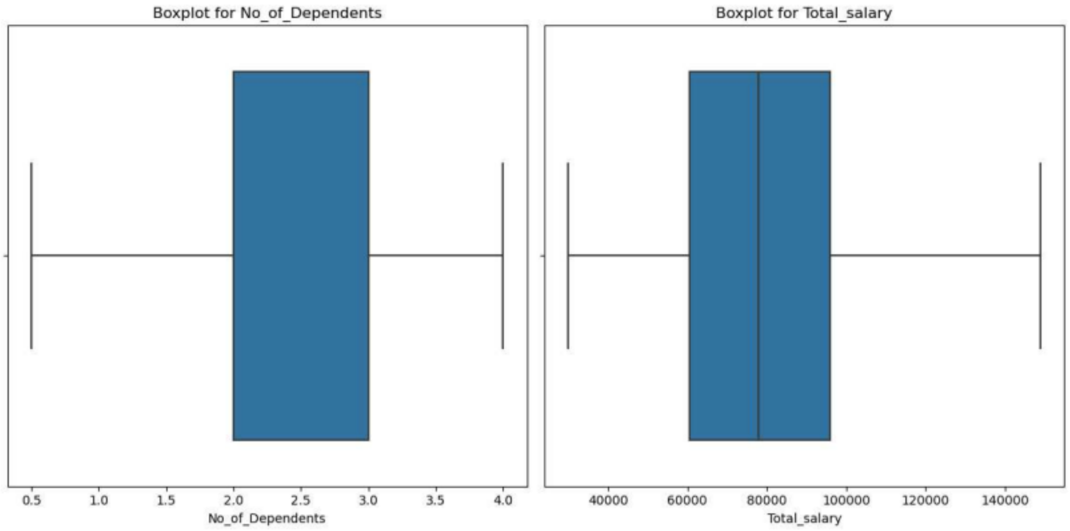


Fig. 4.2.1 Boxplots of No_of_Dependents and Total_salary post Outliers treatment

- The treatment of the outliers of both the columns were executed using Clipping/capping approach.
- The outliers were included in the box-whiskers as those outlier values is important for the further data analysis and insights.

4.3. Observations and Insights

- The Age variable is right-skewed without any outliers
- The No_of_Dependents is normal distribution with outliers present on lower whiskers
- The distribution for Partner_salary, Salary, Price and Total_salary with a right-skewed without having outliers but, Total_salary is having outliers at the upper whiskers.
- In catagorical variables, Male has more in count than female and the Personal_loan (Yes/No) are almost equaly distributed.
- The Salaried professions who have completed Post graduations, got Married and their Partners are working, and having House_loan were significantly higher.
- The automobile company manufacture a greater number of Sedan in compare with SUV and Hatchback.
- The outlier treatment was done using Clipping/Capping approach to analyse the whole dataset in the next steps to ensure more accurate data analysis.

5. Bivariate Analysis

5.1. Explore the relationship between all numerical variables

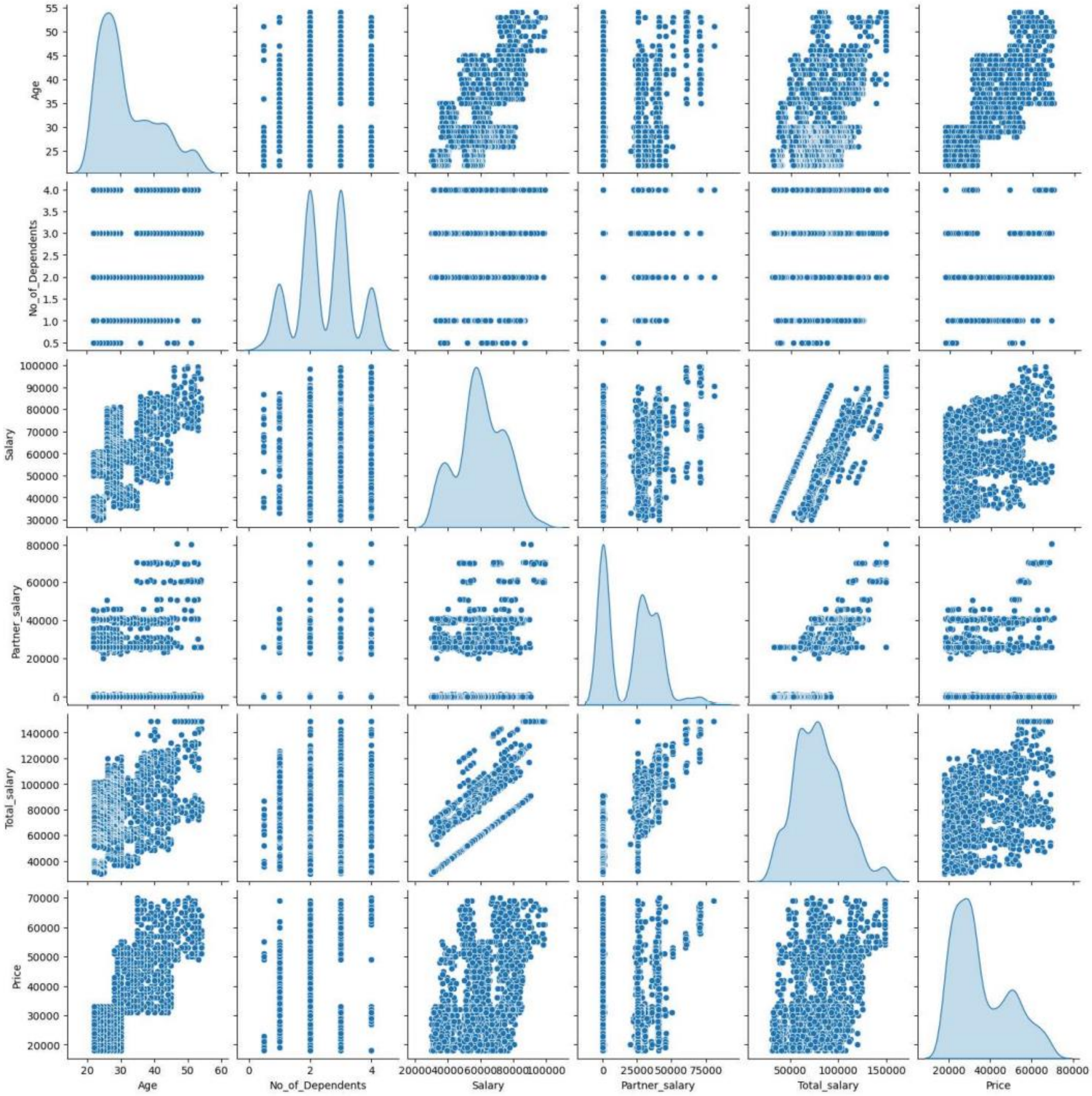
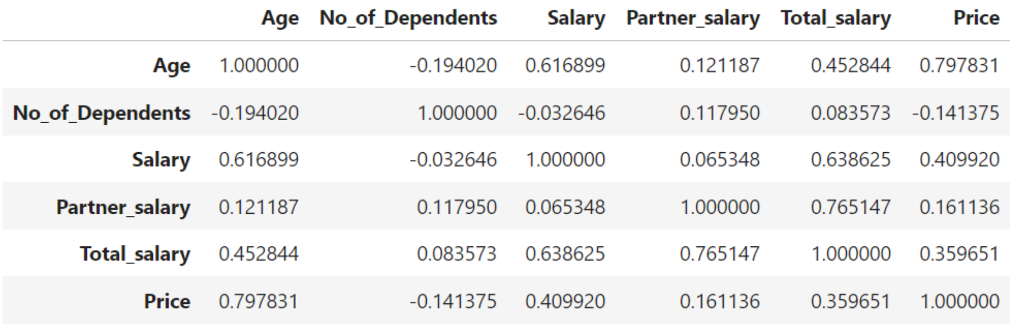


Fig. 5.1 Pairplot of all numerical variables

- Age is a right-skewed distribution with more individuals in the younger age.
- No_of_dependents is displaying maximum values which lies between 2 and 3.
- Salary and Partner_salary shows a positive corelations.
- Salary and Total_salary indicating positive corelations with each other.
- Price shows the right-skewed distribution where the maximum ranges lies approximately between 20,000 to 30,000.

5.2. Explore the correlation between all numerical variables



	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
Age	1.000000	-0.194020	0.616899	0.121187	0.452844	0.797831
No_of_Dependents	-0.194020	1.000000	-0.032646	0.117950	0.083573	-0.141375
Salary	0.616899	-0.032646	1.000000	0.065348	0.638625	0.409920
Partner_salary	0.121187	0.117950	0.065348	1.000000	0.765147	0.161136
Total_salary	0.452844	0.083573	0.638625	0.765147	1.000000	0.359651
Price	0.797831	-0.141375	0.409920	0.161136	0.359651	1.000000

Tab. 5.2 Corelation between the numeric variables

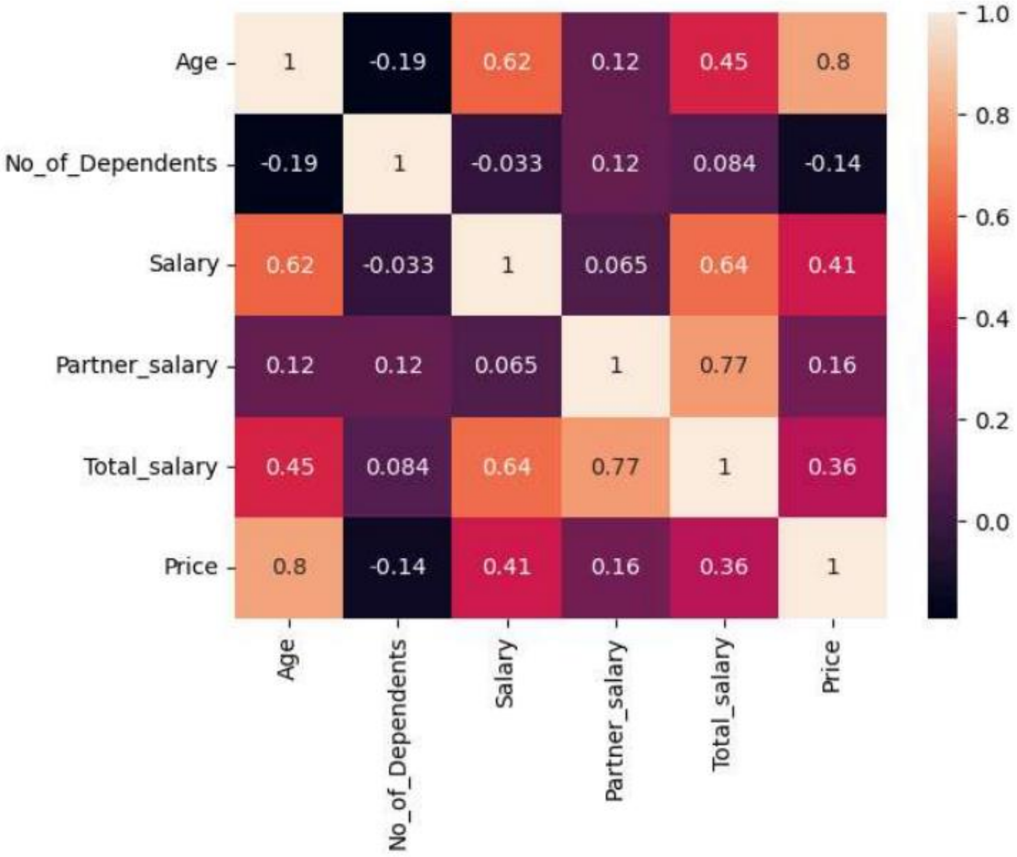


Fig. 5.2.1 Heatmap of the numeric variables corelations

- It shows that the increase of the Age of the individuals tends to purchase high-priced range cars.
- The Total_salary also increase significantly with the increase in either Partner_salary and Salary of the individuals.
- The No_of_dependents have a weak negative corelations with the values of the Price columns.

5.3. Explore the relationship between categorical vs numerical variables

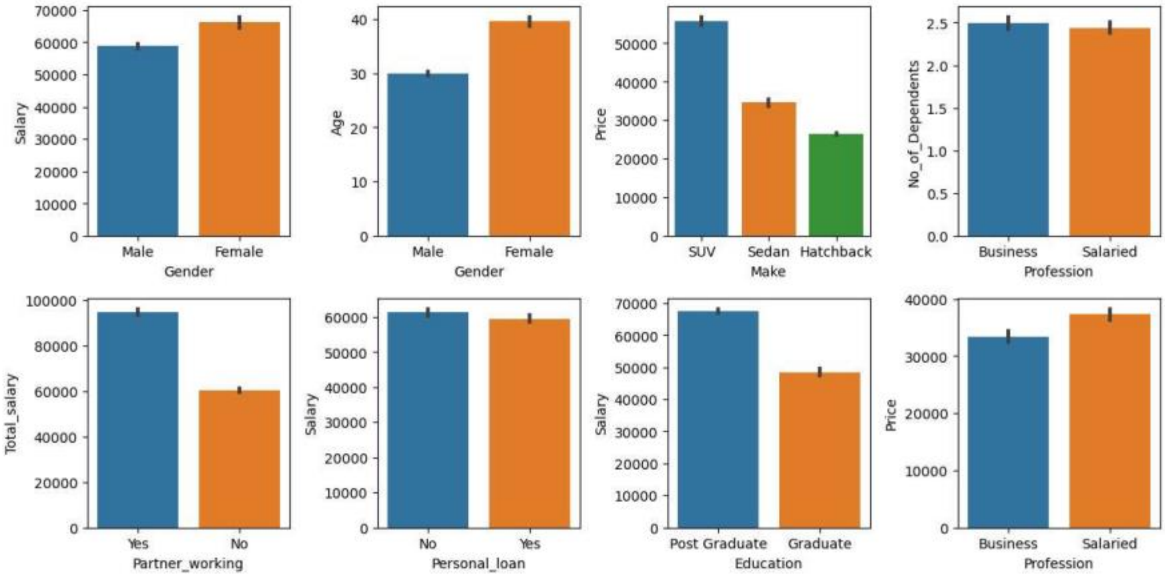


Fig. 5.3 Barplots between categorical and numerical variables

- Female tend to have higher Salary than Male.
- Male are generally younger in average than Female.
- SUV has the higher Price than the Sedan and Hatchbacks.
- Business individuals has more No_of_dependents than the Salaried individuals.
- Individuals with working partners has more Total_salary than the individuals whose partner are not working.
- Individuals without Personal_loan tends to have higher salary than the individuals having Personal_loan.
- Post graduate individuals earns more salary than the Graduates.
- Salaried individuals tend to purchase high-priced cars than the Business individuals.

6. Key Questions

6.1. Do men tend to prefer SUVs more compared to women?

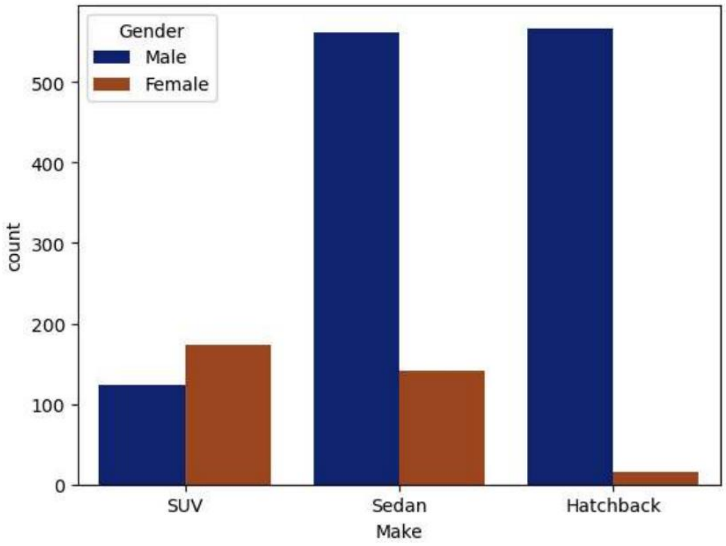


Fig. 6.1 Countplot of Make vs Gender

- The answer is No.
- As per the Fig. 6.1, the plots clearly shows that the Female tends to purchase more SUV than Male.

6.2. What is the likelihood of a salaried person buying a Sedan?

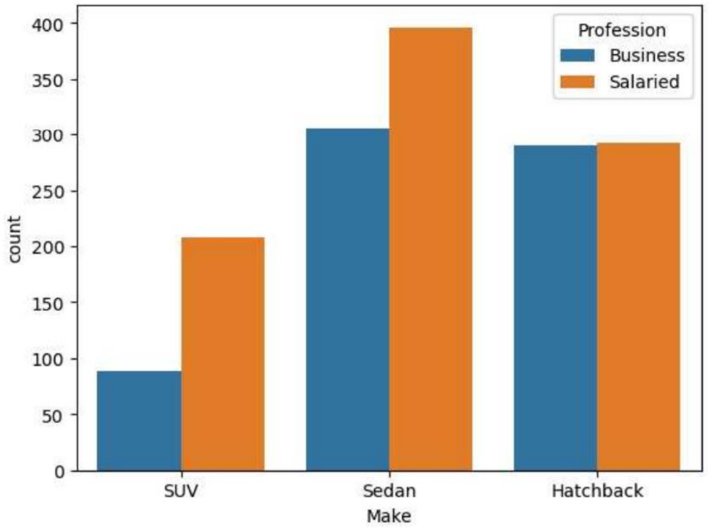


Fig. 6.2 Countplot of Make vs Professions

- The answer is Yes.
- As shown in the above countplot, the Salaried individuals have the tendency to purchase Sedan cars, in compare with SUV and Hatchbacks, is higher than the Business individuals.

6.3. 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?

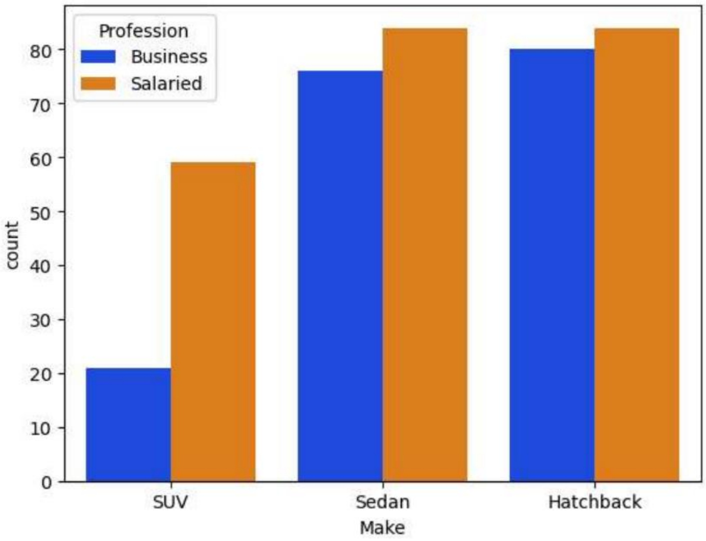


Fig. 6.3 Countplot of Make vs Profession for Sheldon Cooper’s claim

- The answer is No.
- Sheldon Cooper’s claim failed as; the Salaried male is easier to target on Sedan sale in compare with SUV cars.

6.4. How does the amount spent on purchasing automobiles vary by gender?

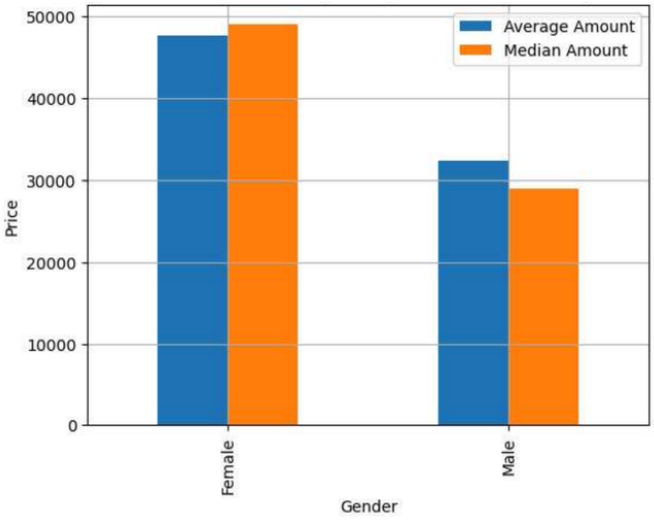
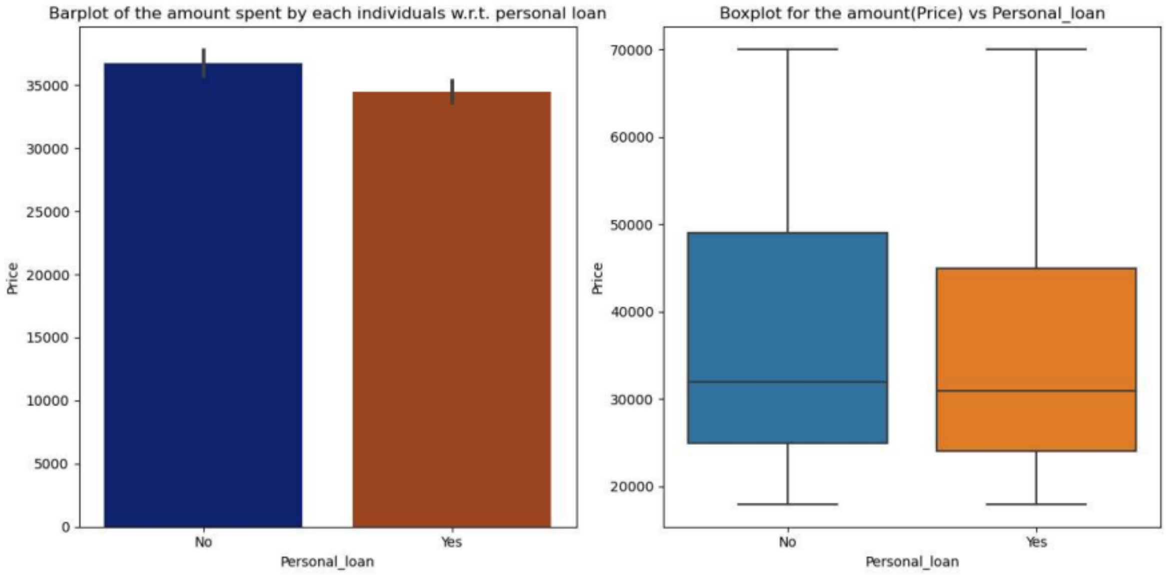


Fig 6.4 Barplot of Gender vs Price

- As shown in the above plot, the Female tend to purchase high price automobiles in average compared with the Male.

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

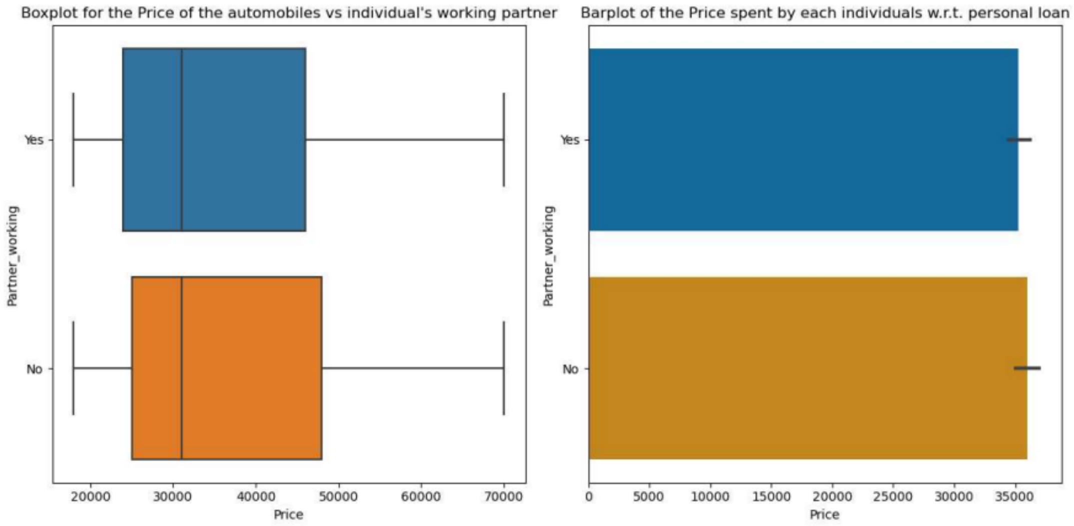

Barplot of the amount spent by each individuals w.r.t. personal loan

Boxplot for the amount(Price) vs Personal_loan

Fig. 6.5 Barplot and Boxplot of Personal_loan vs Price

- As shown in the 2 plots above, the individuals who does not have Personal_loan tends to purchase high-priced cars.

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


Boxplot for the Price of the automobiles vs individual's working partner

Barplot of the Price spent by each individuals w.r.t. personal loan

Fig. 6.6 Boxplot and Barplot of Price vs Partner_working

- As shown in the above plots, the individuals whose partner is working has a slightly higher chances to purchase high-priced cars.

7. Actionable Insights & Recommendations

7.1. Actionable Insights

- Male prefers Sedan and Hatchbacks more than SUV.
- Salaried person is preferred to buy Sedan more than SUV and Hatchbacks.
- Disagreed with the Sheldon Cooper's claim as, the salaried male is more likely to targeted on Sedan and Hatchback car types.
- The Average amount spent by Female to purchase the automobiles is more than Male.
- Individuals who did not take personal loans were more likely to purchase higher-priced cars.
- There is mostly no difference whether their partner is working or not to purchase the high-priced cars.

7.3. Business Recommendations

- A campaign must be executed for Male to buy SUV and Female to purchase all 3 types of cars.
- Create a marketing plan and strategies for the Salaried individuals to purchase more SUV and Hatchbacks.
- Regularly analyze the customer data and refine the marketing plans for more profitable business.
- Focus on double-income individuals whose partners are working to build a solid strategy to purchase maximum high-priced cars.
- Marketing plan should not be executed whether the individual’s partner is working or having personal loan as, it is slightly difference in the data for the high-priced automobile purchases.