PYTHON FOR DATASCIENCE
PROBLEM STATEMENT
PDS GRADED PROJECT - CODED

AUSTO MOTOR COMPANY DATA ANALYSIS

By: BENITA MERLIN.E
PGP-Data Science and Business Analytics.
BATCH: PGPDSBA.O.MAY24.A

LIST OF CONTENT

LIST OF FIGURES

LIST OF TABLES

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 an analytics professional to improve the existing campaign.

Objective

They 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.

A.Data Dictionary:

Dataset Name: austo_automobile+%282%29+%281%29

This data dictionary provides a clear and concise reference for understanding the dataset's structure and the meaning of each column, ensuring accurate and consistent data handling in analysis.

Some important technical information about the dataset:

The data set contains of 1581 rows & 14 columns.
The data contains 5 Integer, 1 float & 8 Object data type columns,

Table 1. Data Dictionary

SI.No	Column Name	Description	Data Type	Example Values
1	Age	The age of the individual in years.	Integer	25, 34, 50
2	Gender	The gender of the individual, categorized as male or female.	Object	Male, Female
3	Profession	The occupation or profession of the individual.	Object	Salaried,Business
4	Marital status	The marital status of the individual, such as married or single.	Object	Married, Single
5	Education	The educational qualification of the individual.	Object	Graduate, Post Graduate
6	No of Dependents	The number of dependents (e.g., children, elderly parents) that the individual supports financially.	Integer	0, 2, 3
7	Personal loan	A binary variable indicating whether the individual has taken a personal loan.	Object	Yes, No
8	House loan	A binary variable indicating whether the individual has taken a housing loan.	Object	Yes,No
9	Partner working	A binary variable indicating whether the individual's partner is employed.	Object	Yes, No
10	Salary	The individual's salary or income.	Integer	50000, 75000
11	Partner_salary	The salary or income of the individual's partner, if applicable.	Float	45000.00, 60000.00
12	Total_salary	The total combined salary of the individual and their partner (if applicable).	Integer	95000, 135000
13	Price	The price of a product or service.	Integer	25000, 30000
14	Make	The type of automobile	Object	Sedan, SUV, Hatchback

B. STATISTICAL ANALYSIS:

The statistical analysis provides a summary of the key metrics for the numerical columns in the dataset. This includes measures such as the count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile, Q2), 75th percentile (Q3), and maximum values for each column.

Figure 1. Statistical summary

	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
count	1581.000000	1581.000000	1581.000000	1475.000000	1581.000000	1581.000000
mean	31.922201	2.457938	60392.220114	20225.559322	79625.996205	35597.722960
std	8.425978	0.943483	14674.825044	19573.149277	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	38300.000000	95900.000000	47000.000000
max	54.000000	4.000000	99300.000000	80500.000000	171000.000000	70000.000000

Observation:
- The average age is around 32 years, with the majority of individuals aged between 25 and 38 years. The age distribution suggests a relatively young population.
- There are missing values in the gender column. The gender column has 4 unique values, including misspellings that need correction.
- Most of the customers are married. There are 1443 customers who are married
- Customers with post-graduate qualifications are higher in count compared to those with graduate qualifications.
- The majority of the partners are employed.
- The price of the cars ranges from 18,000 to 70,000 units.
- The most common make preferred by customers is the Sedan.
- Customers who have bought cars more frequently are those who do not have a house loan.

C. CHECK FOR DISCREPENCIES AND TREATMENT:

C.1. Unique value in column for object data type:

1.Gender - Male, Femal, Female, nan, Femle.
2. Profession – Business, Salaried
3. Marital status – Married, Single
4. Education-Graduate, Post Graduate
5. Personal Loan – Yes, No
6. House Loan – Yes, No
7. Partner Salary – Yes, No
8. Make – SUV, Sedan, Hatchback

C.2. Gender Column:

Upon reviewing the dataset, the following observations were made regarding the missing values in the partner salary column:

Missing Values: There are 53 missing values in the gender column.

Data Type: The data type for the gender column is object.

- To ensure data accuracy and consistency, the missing values in the gender column were treated by correcting spelling errors.
- Specifically, any occurrences of 'Femal' and 'Femle' were replaced with the correct spelling, "Female". This step was essential to standardize the gender data, facilitating accurate analysis and reporting.
- The remaining missing values in the gender column were addressed using a specific imputation technique.
- The null values were replaced with the 'mode' of the column.
- The mode represents the most frequently occurring value within the column. This approach ensures that the imputed values are consistent with the predominant category, thereby preserving the integrity of the data distribution.

C.3.Partner salary column:

Upon reviewing the dataset, the following observations were made regarding the missing values in the partner salary column:

Missing Values: There are 106 missing values in the partner salary column.

Data Type: The data type for the partner salary column is float.

- **Single Marital Status:** There are 16 missing values in the partner salary column for individuals with a marital status of "Single".
- **Married Working Partner:** There are also 16 missing values in the partner salary column for individuals with a marital status of "Married" and who have a working partner.
- **Married Non-Working Partner:** The remaining 74 missing values in the partner salary column correspond to individuals with a marital status of "Married" and who have a non-working partner.

D. Exploratory Data Analysis (EDA)

D.1.Univariate Analysis

D.1.1. Age
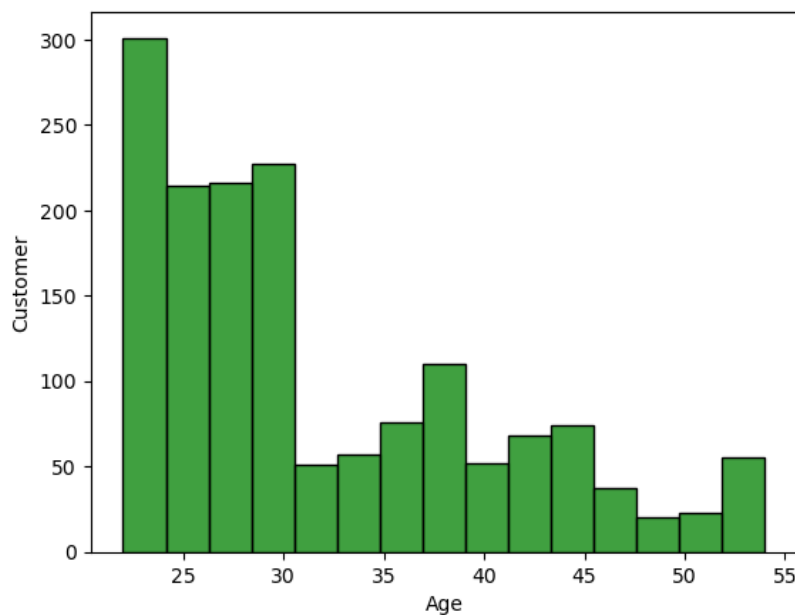
Figure 2. Histogram for Age of the Customer

Observation:
- This plot shows that, there are relatively more customers under the age of 30.

- A histogram skewed to the right has a large number of occurrences on the left side of the plot and a few on the right side of the plot.
- It is a right skewed distribution.

D.1.2 GENDER:

Figure 3. Bar plot for Gender of the Customer

Observation:
- There are 1252 male customers and 329 female customers.

D.1.3 PROFESSION
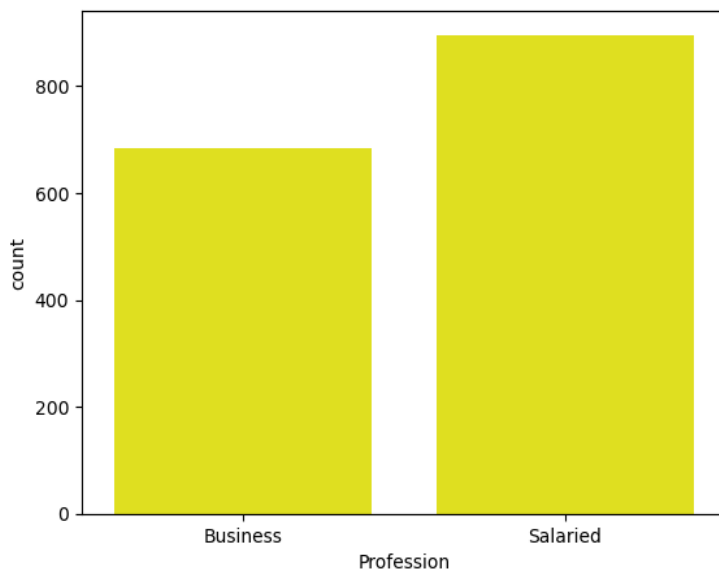
Figure 4. Bar plot for Profession of the Customer

Observation:

- Business customers are lesser than Salaried customers
- There are 896 salaried customers and 685 Business customers.

D.1.4. MARITAL STATUS
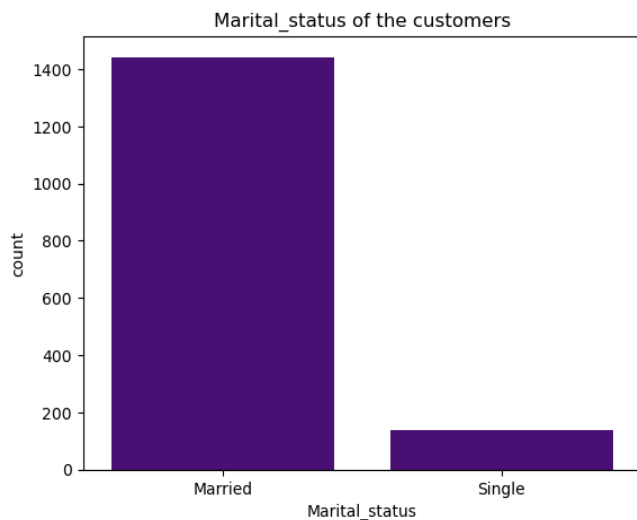
Figure 5. Bar plot for Marital status of the Customer

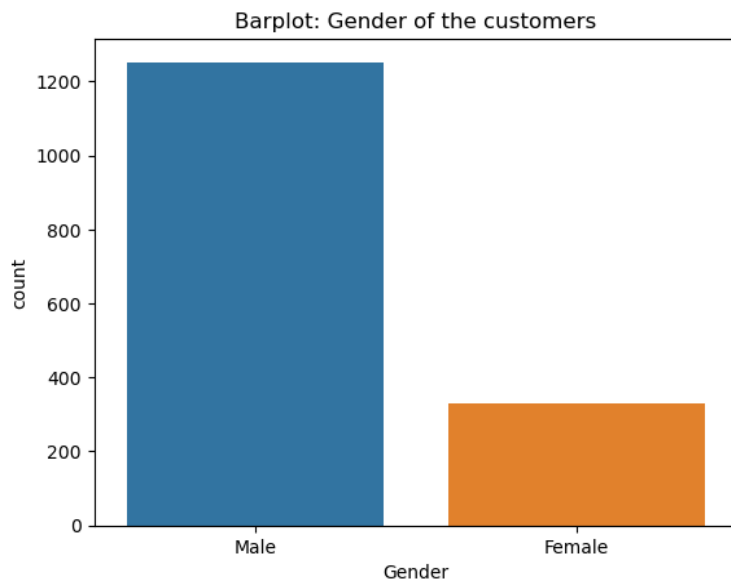Observation:

- There are 1443 married customers and 138 single customers
- Majority of the customers are married

D.1.5. NUMBER OF DEPENDENTS:
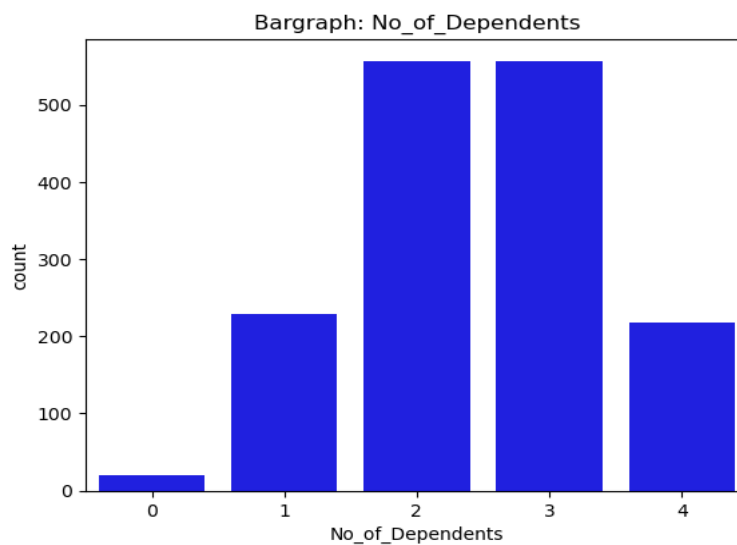
Figure 6. Bar plot for No. of. Dependents of the Customer

Observation:

- 20 customers have 0 dependents.
- 229 customers have 1 dependents.
- 557 customers have 2 dependents.
- 557 customers have 3 dependents.
- 218 customers have 4 dependents.
- This is bimodal distribution.
- 2 and 3 dependents have same count.

D.1.6. BOXPLOT FOR NO. OF. DEPENDENTS OF THE CUSTOMER
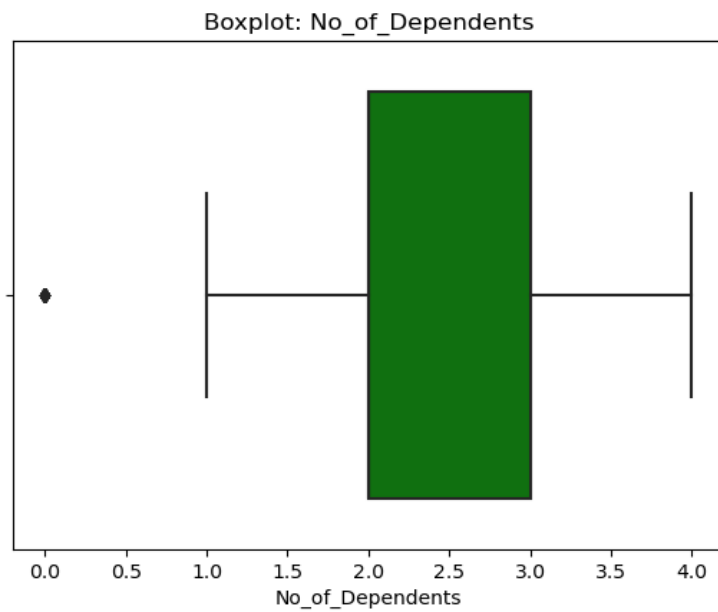
Figure 7. Boxplot of No. of. Dependents of the Customer

- From the above graph we can clearly see the existence of outliers.

D.1.7 REMOVING THE OUTLIER USING IQR METHOD

D.1.7.1. Removing the outliers using IQR method for No. of. Dependents:

FORMULA:

```
# Calculate Q1(25th percentile) and Q3(75th percentile):
Q1= NO. of. Dependents .quantile(0.25)
Q3= NO. of. Dependents .quantile(0.75)

# Calculate the Inter Quartile Range (IQR)
IQR = Q3 - Q1

# Define outlier threshold
Lower bound = Q1 - 1.5* IQR
Upper bound = Q3 + 1.5* IQR

# Replace the values in the No. of.
Dependents column with NaN if they are less than the lower bound or greater than the upper bound.
```

Table 2: Formula to rectify outliers

Output:

Boxplot of No_of_Dependents after removing outliers

Figure .8. Boxplot for No. of. Dependents after removing outliers.

Observation:
- outliers are removed.
- Now the dependents are from 1 to 4.
- Most of the customers have 2 to 3 dependents and they are equal in count

D.1.7.2. Removing the outliers using IQR method for Total salary:

Boxplot: TOTAL SALARY
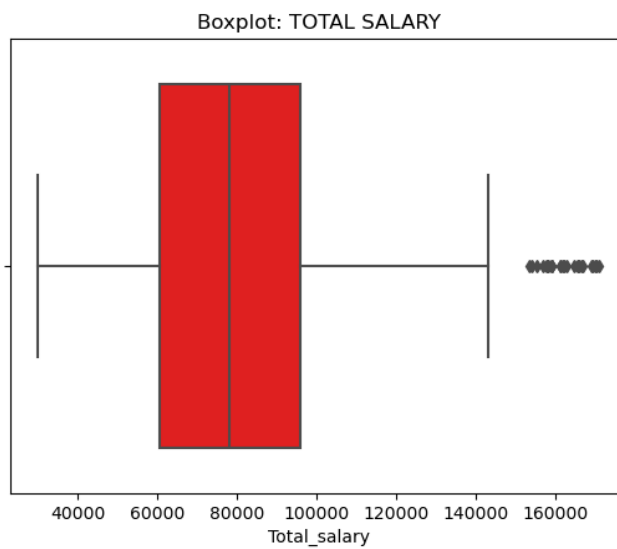
Figure. 9. Boxplot for Total salary

Observation:

* Outliers are seen in Total salary.
* Same like No.of .Dependents, IQR method is used to remove outliers in Total salary.

Output:
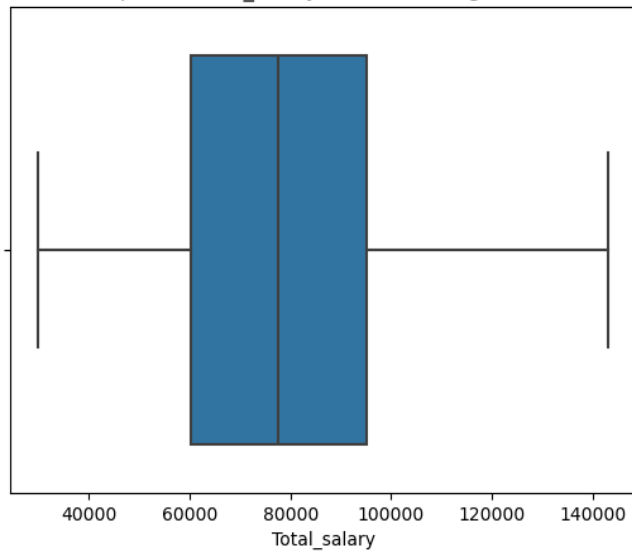
Figure. 10. Boxplot for Total salary

Observation:
- Outliers are removed.
- This is a bimodal distribution.
- Most of the customers have a total salary ranging from 60,000 to 90,000.

D.1.8 PERSONAL LOAN
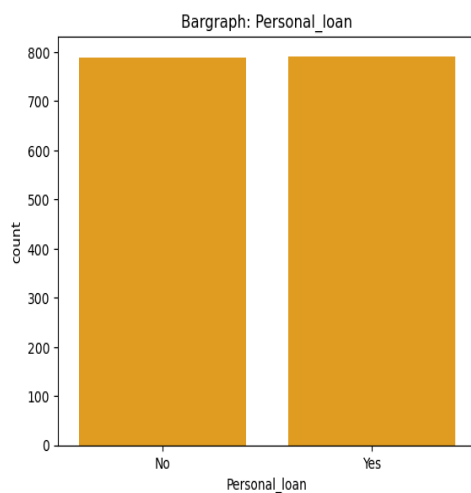
Figure. 11. Bargraph for Personal loan

Observation:

- 789 customers have not taken personal loan whereas 792 customers have taken personal loan.

D.1.9 HOUSE LOAN
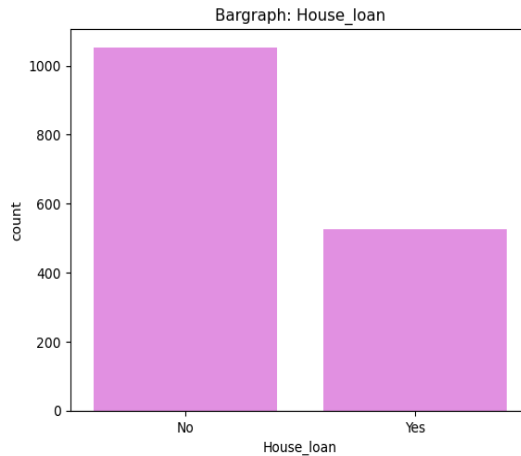
Figure. 12. Bargraph for Personal loan

Observation:
- 527 customers have taken house loan whereas 1054 have not taken house loan.

D.1.10 PARTNER WORKING
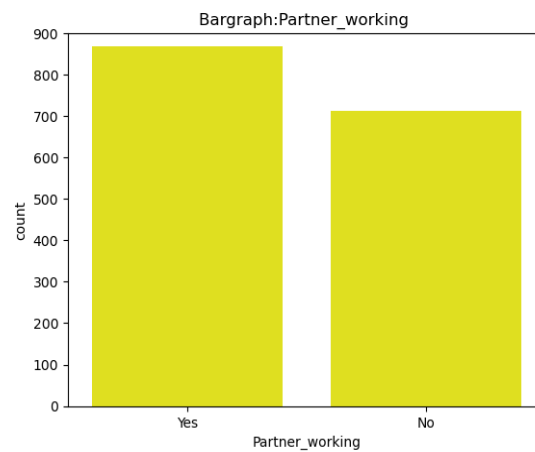
Figure. 13. Bar graph for Partner working

Observation:
- 693 partners are non workers.
- 868 partners are working

D.1.11 PARTNER SALARY:
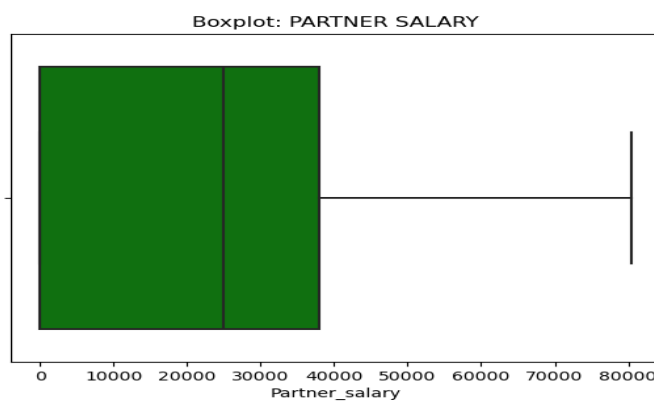
Figure. 14. Boxplot for Partner salary

Observation:
- The median is closer to the right end of the box and there is no whisker on the left end of the box.
- The minimum value in the dataset is equal to the first quartile (Q1).
- The data is likely left-skewed (negatively skewed)
- A left-skewed distribution has a longer tail on the right side, indicating that there are higher values that are more spread out, while lower values are more clustered

D.1.12 MAKE:
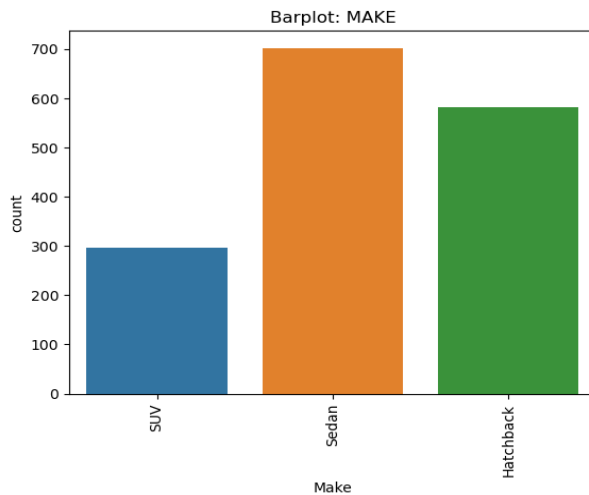
Figure. 15. Bar plot for Make

Observation:

- Sedan: Purchased by 702 customers
- Hatchback: Purchased by 582 customers
- SUV: Purchased by 297 customers
- A large number of customers purchased sedans.

D.2.BIVARIATE ANALYSIS

D.2.1. Categorical vs Categorical: Relationship between Gender and Make
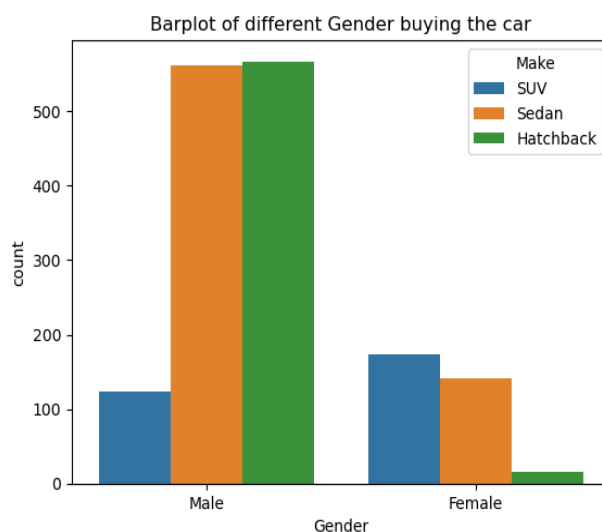
Figure. 16. Bar plot for different Gender buying the car

Observation:

14

- There are more male buyers for sedan and hatchback.
- Women are preferred to buy suv than men
- **Sedan:** Preferred by male customers, with purchase count nearly equal to that of hatchbacks.
- **Hatchback:** Preferred by male customers, with purchase count nearly equal to that of sedans.

D.2.2. Categorical vs Categorical: Relationship between Make and Profession:
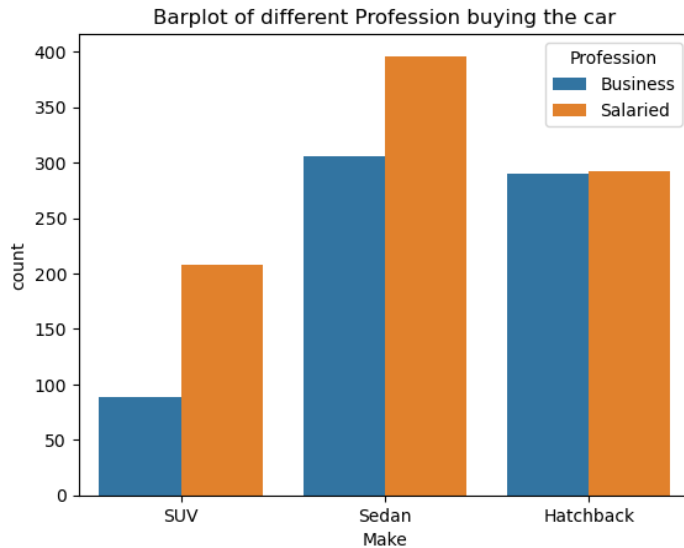
Figure. 17. Bar plot for different Profession buying the car

Observation:
- Salaried proffesionals have bought more SUV than Business professional.
- Salaried professionals have bought more Sedan than Business professionals.
- Both Salaried and Business professionals have bought almost equal number of Hatchback.

D.2.3 NUMERICAL VS. NUMERICAL: RELATIONSHIP BETWEEN AGE AND TOTAL SALARY
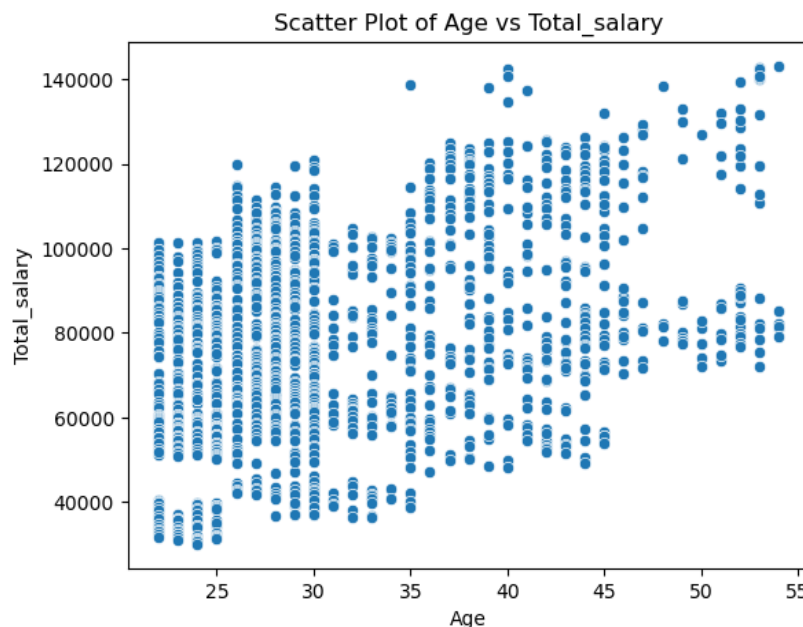
Figure. 18. Scatterplot of Age vs Total salary

- In a scatter plot when the y variable tends to increase as the x variable increases, we say there is a positive correlation between the variables.

15

D.2.4. CATEGORICAL VS CATEGORICAL: RELATIONSHIP BETWEEN AGE AND HOUSE LOAN:
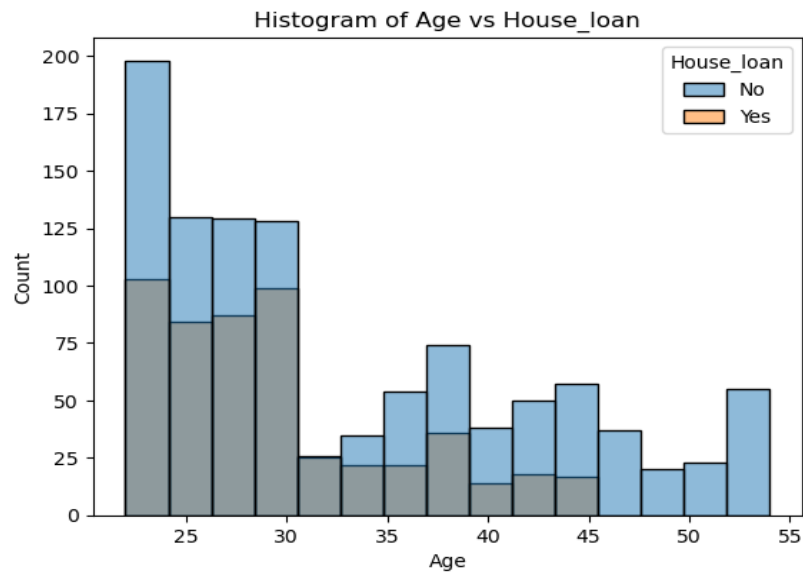
Figure. 19. Histogram of Age vs House loan

Observation:

- Majority of the customers have not taken House loan

D.2.5. CATEGORICAL VS CATEGORICAL: RELATIONSHIP BETWEEN MAKE AND PERSONAL LOAN:
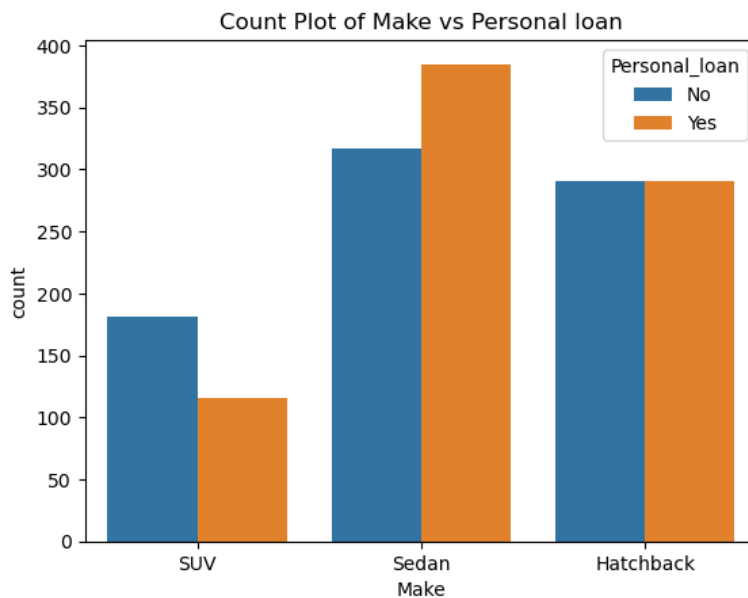
Figure. 20. Count of Make vs Person loan

Observation:

- **SUV:** More customers who have bought SUVs do not have a personal loan.
- **Sedan:** More customers who have bought sedans have a personal loan.
- **Hatchback:** The number of customers who purchased hatchbacks with personal loans is almost equal to those who purchased hatchbacks without a personal loan.

D.2.6. CATEGORICAL VS CATEGORICAL :RELATIONSHIP BETWEEN MAKE AND TOTAL SALARY:
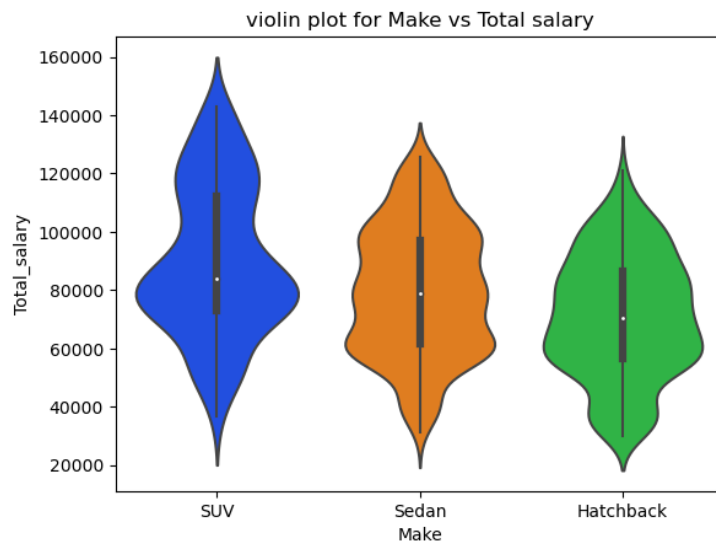
Figure. 21. Violin plot of Make vs Total salary

Observation:
- We can say that SUV is preferred in all salary ranges.
- Those with salary range up to 13500 go with sedan and hatchback.

D.3. MULTIVARIATE ANALYSIS:

D.3.1.Heatmap for Numeric Data:
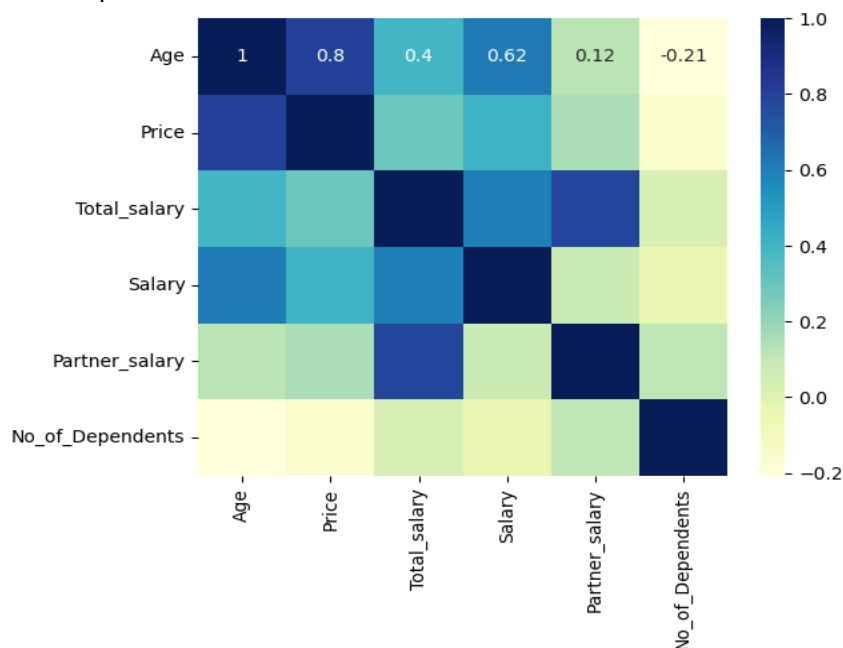
Figure. 22: Heatmap for Numeric Data

Observation:
- Older customers tend to purchase higher-priced cars.
- There is a strong positive correlation of 0.8 between age and price.
- There is a negative correlation between the number of dependents and price, as well as between age and the number of dependents.
- Age and price have a correlation of -0.2, indicating a negative relationship.
- Age and the number of dependents have a correlation of approximately -0.15, indicating a negative relationship.

D.3.2. Numerical vs Categorical : Relationship among Gender, Make, Price, House loan:
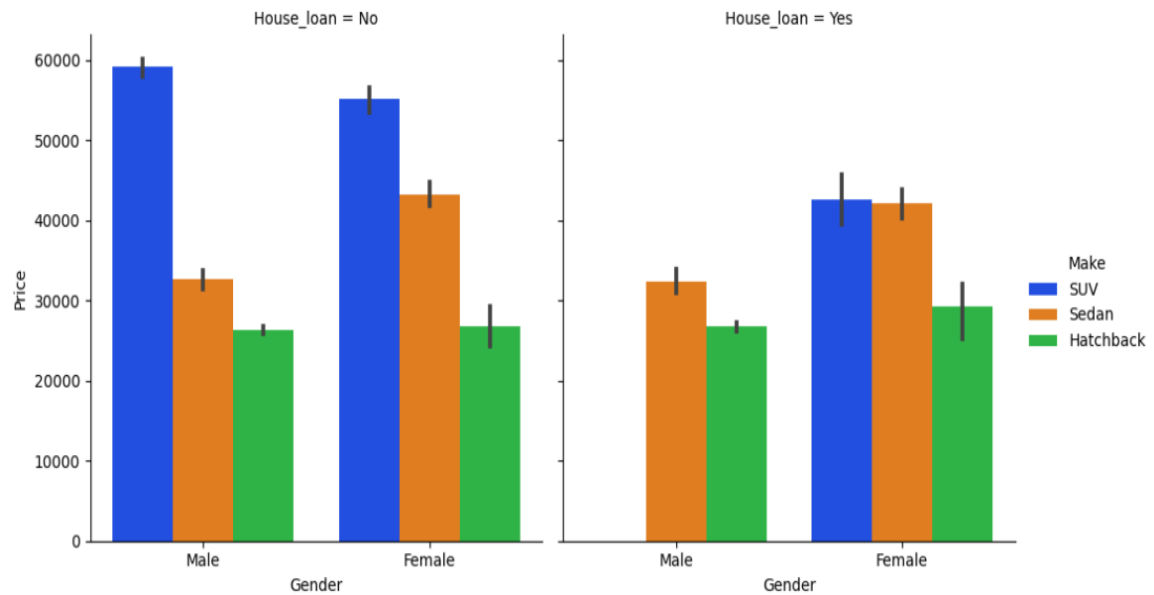
Figure. 23: Cat plot for Gender, House loan, Make and Price.

Observation:

- Men who have house loan have not purchased SUV
- Female who have house loan have purchased suv and sedan for almost equal price.
- Men who donot have house loan have purchased suv for more price.
- Women who donot have house loan have purchased suv for more price.

E. KEY QUESTIONS:

E1: Do men tend to prefer SUVs more compared to women?
Distribution of SUV Preferences: The counts of males and females who prefer SUVs were visualized.
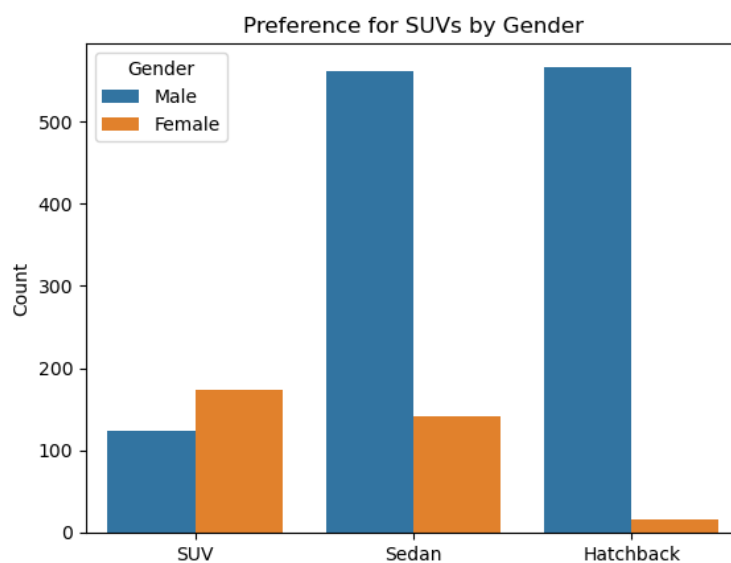
Figure. 24: Bar plot of Make vs Gender

Observation:
- Number of males who prefer SUVs: 124
- Number of females who prefer SUVs: 173

So, Men didn't prefer SUVs more compared to women.

E 2: What is the likelihood of a salaried person buying a Sedan?
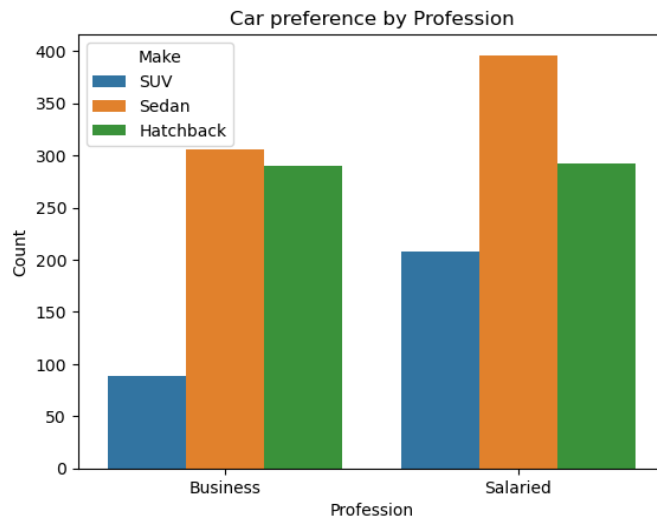
Figure. 25: Bar plot of Profession vs Make

Observation:
- The counts of business and salaried customers who prefer Sedans were visualized.
- Number of Business who prefer Sedan: 306
- Number of Salaried who prefer Sedan: 396
- 396 salaried person liked to buy sedan

From this plot we can say that the salaried professionals showed a higher preference for 'Sedans' compared to other segment because of the affordable Price.
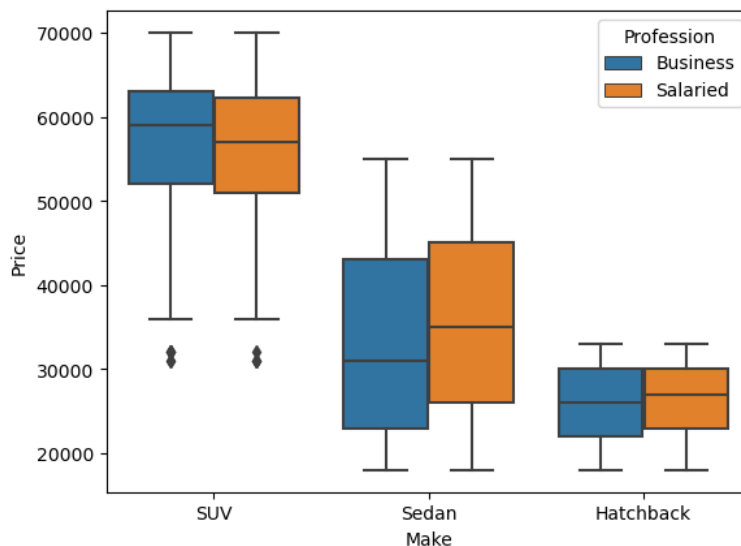
Figure. 26: Box plot of Make vs profession

- From the all above plot we can say that the Sedan is available at various price points, from lower to higher comparatively other cars, making it more likely for a salaried customer to purchase sedan

E 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?
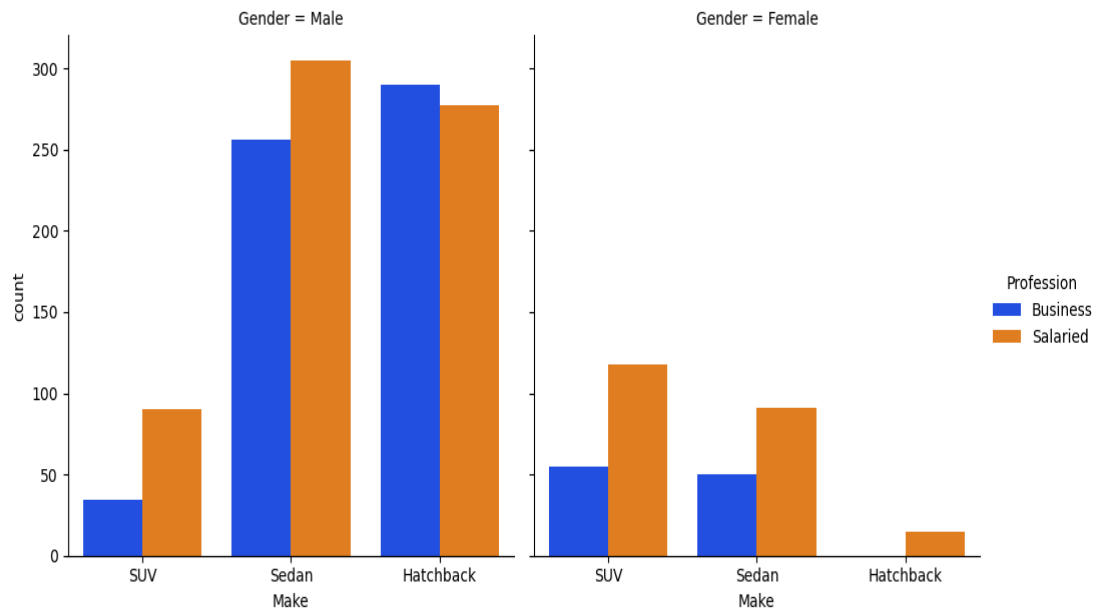
Figure. 27: Cat plot for Gender, Profession, Make

Observation:

- From this plot we can say that the salaried male purchased sedan more than SUV.

So, there is "No" evidence or data supports Sheldon cooper's claim that a salaried male is an easier target for a SUV sale over a Sedan sale.

E 4: How does the amount spent on purchasing automobiles vary by gender?

	Gender	Make	mean	median	sum	count
0	Female	Hatchback	27466.666667	28000.0	412000	15
1	Female	SUV	53479.768786	55000.0	9252000	173
2	Female	Sedan	42773.049645	43000.0	6031000	141
3	Male	Hatchback	26447.971781	26000.0	14996000	567
4	Male	SUV	59096.774194	59000.0	7328000	124
5	Male	Sedan	32550.802139	31000.0	18261000	561

Figure. 26: Mean, Median, Sum, Count of Gender and Make

Observation:
- Females spent more on SUVs than males.
- Males spent more on sedans than females.
- Males also spent more on hatchbacks than females.
- Only 15 females purchased hatchbacks.
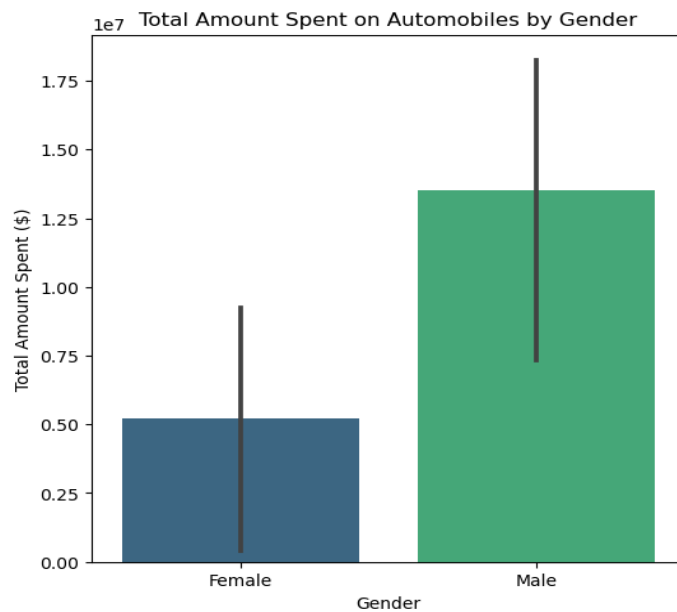
Figure. 28: Bar graph for Gender and Sum

- Sedan and hatchback were purchased by male at high amount than female.
- SUV were purchased by female at high amount than male.

E 5.: How much money was spent on purchasing automobiles by individuals who took a personal loan?

```
   Personal_loan      Make         mean    median        sum  count
0             No  Hatchback  26683.848797   28000.0   7765000    291
1             No        SUV  57309.392265   58000.0  10373000    181
2             No      Sedan  34233.438486   33000.0  10852000    317
3            Yes  Hatchback  26264.604811   26000.0   7643000    291
4            Yes        SUV  53508.620690   53000.0   6207000    116
5            Yes      Sedan  34909.090909   34000.0  13440000    385
```

Figure. 29: Mean, Median, Sum, Count of Personal loan and Make

Observation:
- The amount spent on purchasing SUVs is higher for customers who have not taken a personal loan.
- The amount spent on purchasing hatchbacks is higher for customers who have not taken a personal loan.
- The amount spent on purchasing sedans is higher for customers who have taken a personal loan.

Figure. 30: Bar graph for Personal loan vs Total Amount spent($)

Individuals who did not take a personal loan have spent more on purchasing automobiles

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

	Partner_working	Make	mean	median
0	No	Hatchback	26323.843416	26000.0
1	No	SUV	56173.611111	58000.0
2	No	Sedan	35354.166667	34000.0
3	Yes	Hatchback	26614.617940	28000.0
4	Yes	SUV	55496.732026	57000.0
5	Yes	Sedan	34082.125604	33000.0

Figure. 31: Mean, Median for partner working and Make

Observation:

- The average price for SUVs and sedans is lower, which might indicate that having a working partner does not significantly influence the purchase of higher-priced cars.
- The average price for hatchbacks is higher, which might indicate that having a working partner does significantly influence the purchase of higher-priced cars.
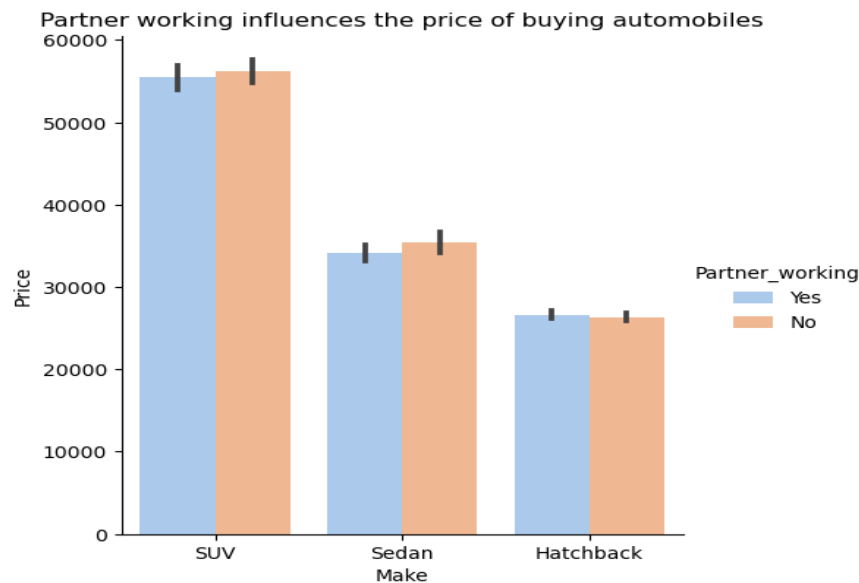
Figure. 32: Cat plot for Make ,Parner working, Price

* From this plot, we can say that SUVs and sedans were purchased by customers whose partners are not working, and these customers tend to purchase at higher prices.
* Hatchbacks were purchased by both customers with working and non-working partners at similar price levels.
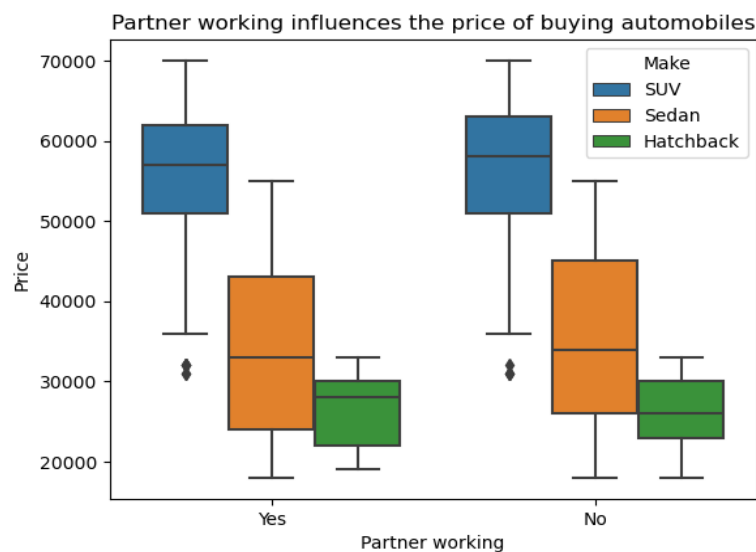
Figure. 33: Box plot of Partner working, Price, Make

* From this boxplot, we can observe that SUVs were purchased at higher prices by customers whose partners are not working.
* Sedans were also purchased at higher prices by customers whose partners are not working.
* Hatchbacks were purchased at higher prices by customers whose partners are working.

we can say Partner working doesn't influences the price of buying automobiles at high price.

F. INSIGHTS AND RECOMMENDATION:

F.1 INSIGHTS:

- While men show a preference for SUVs, women tend to prefer SUVs even more, as evidenced by their higher average spending on this vehicle type.
- This suggests an opportunity for targeting SUV marketing efforts specifically towards women, highlighting features and benefits that resonate with their preferences and needs.
- Salaried professionals demonstrate a stronger preference for sedans, largely due to their affordability compared to other vehicle segments.
- This indicates a market opportunity to further cater to this segment's preferences and needs.
- The salaried males tend to purchase sedans more frequently than SUVs.
- Females have a higher expenditure on SUVs compared to males, indicating a preference or higher investment in this vehicle type.
- Males tend to spend more on sedans than females, suggesting a stronger inclination towards this vehicle category among males.
- SUV and Sedan Purchases: Customers whose partners are not working tend to purchase SUVs and sedans at higher prices. This indicates that these customers might have higher disposable incomes or different spending priorities when it comes to vehicle purchases.
- Both customers with working and non-working partners purchase hatchbacks at similar price levels, suggesting that the decision to buy a hatchback is less influenced by the partner's employment status.
- Customers who have not taken a personal loan tend to spend more on purchasing SUVs.
- This suggests that those without personal loans have more disposable income or prefer to invest in higher-priced SUVs without needing additional financing.
- Hatchback Purchases and Personal Loans: Similarly, customers without personal loans spend more on purchasing hatchbacks.
- This indicates a preference or financial capacity to buy these vehicles outright without resorting to personal loans.
- Sedan Purchases and Personal Loans: Customers who have taken a personal loan tend to spend more on purchasing sedans.
- This implies that sedans are a preferred choice for those who need financing, potentially due to their affordability and the balance between cost and features.

F.2. RECOMMENDATION:

- Develop and launch targeted marketing campaigns that emphasize SUV features that appeal to women, such as safety, comfort, and versatility.
- Increase the variety and availability of sedan models that cater to different price points and features preferred by salaried professionals
- Offer attractive financing options, such as low-interest rates or flexible payment plans, to make sedans even more accessible to salaried professionals.
- Provide clear and detailed information about the benefits and value proposition of sedans through online platforms, social media, and in-person consultations to educate potential customers about their advantages.
- Focus on promoting sedan models to salaried males, emphasizing their affordability, practicality, and features that align with their preferences.
- Implement promotional strategies aimed at increasing hatchback sales among both genders, emphasizing versatility, fuel efficiency, and cost-effectiveness.
- Promote the affordability and practicality of hatchbacks to both customer groups, emphasizing that they offer great value regardless of the partner's employment status.
- Consider introducing features or packages that cater to a wide range of needs to maintain their appeal across different demographics.
- Offer educational content on the benefits of different vehicle types and how they align with various lifestyle and financial situations. This can help customers make informed decisions based on their specific circumstances.
- Flexible Financing for SUVs and Hatchbacks:
- SUVs: Promote flexible financing options and low-interest rate loans for customers interested in SUVs. Highlight the long-term benefits and value of investing in an SUV, making it an attractive option even for those who might need financing.

- Hatchbacks: Offer similar financing options for hatchbacks to encourage purchases among customers who may not have the upfront cash but are interested in these vehicles.
- Special Offers on Sedans:
- Create special financing deals, such as zero down payment or low monthly payments, for sedans. Emphasize these offers in marketing campaigns to attract customers who are likely to take personal loans for purchasing sedans.
- Provide clear and accessible information about the benefits of various financing options for each vehicle type. Use online platforms, brochures, and sales representatives to explain how financing can make higher-priced vehicles more attainable.
- Host promotional events or workshops that explain the financial benefits of purchasing SUVs and hatchbacks without loans. Similarly, hold events focusing on the affordability of sedans through personal loans, providing real-life examples and testimonials from satisfied customers.
- Segment marketing campaigns based on customer financial profiles. For example, target ads for SUVs and hatchbacks towards customers with higher disposable incomes and no personal loans, while promoting sedans with financing options to those who are more likely to need a loan.