Project Python for Data Science
Austo Motor Company Data Analysis
PDS Coded Project

By: Agnes Raja Kumari. E

PGP-Data Science and Business Analytics

PGPDSBA.O.MAY24.A

List of Content

List of Figure

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.

1. Data Analysis

1.1 Data Description

The data is all about the Austo Motor Company, a leading car manufacturer specializing in SUV, Sedan, and Hatchback models. The data has 1581 rows and 14 columns. Below are the Data Dictionary.

```
 #   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)
```

Fig:1.1 Data Description.

Observations on Data Description

Column 1 indicates the serial number, which should be in integer format.

Column 2 indicates the datasets with 14 columns.

Column 3 shows the count of non-null elements, which should be in integer format, but is currently in float format.

Column 4 indicates the type of data.

The given dataset includes 8 categorical data columns and 6 numerical data columns.

1.2 Statistical summary of the data

	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	1528	4	Male	1199	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	NaN	NaN	NaN	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
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	1475.0	NaN	NaN	NaN	20225.559322	19573.149277	0.0	0.0	25600.0	38300.0	80500.0
Total_salary	1581.0	NaN	NaN	NaN	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
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

Fig:1.2 Statistical summary of the data

Observations on Statistical summary of the data

Age: 75% of car owners are aged 38 or younger.

Gender: There are 4 unique gender values in the dataset, but the data description indicates there should be only 2. This discrepancy needs to be addressed.

Profession: Salaried individuals are more likely to purchase cars compared to business owners.

Marital_status: Married individuals buy more cars compared to single individuals.

Education: Postgraduates are more inclined to purchase cars compared to graduates.

No_of_Dependents: 75% of car owners have 3 or fewer dependents. In the data description, it states that the "No_of_Dependents" column is of integer data type, but in the count column, it shows as float data type. We need to find and address this discrepancy to ensure data consistency.

Personal_loan: Individuals with a personal loan tend to purchase more cars.

House_loan: Individuals with a house loan are less likely to purchase cars.

Partner_working: People whose partners are working are more likely to buy cars.

Salary: The average salary of car owners is 60,392 dollars.

Partner_salary: Few data are misssing and we need to find and address this discrepancy to ensure data consistency.

Total_salary: In the data description, it states that the "No_of_Dependents" column is of integer data type, but in the count column, it shows as float data type. We need to find and address this discrepancy to ensure data consistency.

Price: Price ranges from 18000 to 70000 dollars.

Make: The most common type of car owned is a Sedan.

1.3 Discrepancy in Categorical column

```
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
```

Fig: 1.3 Discrepancy in Categorical column

Observation on Discrepancy in Categorical column

In Gender 53 values and in Partner_salary 106 values are missing.

1.3.1 Gender

```
array(['Male', 'Femal', 'Female', nan, 'Femle'], dtype=object)
```

Fig: 1.3.1 Unique value

Observations on Gender

In the "Gender" column, there are 4 unique values. The spelling of "Female" is incorrectly spelled as "Female" and "Femle." We need to address this by correcting the spelling. For the remaining missing NaN values, we will use the most frequent method (mode) to replace the NaN values.

1.3.2 Partner_salary

```
array([70700., 70300., 60700., 60200., 60500., 50800., 40400.,     0.,
       70600., 60600., 60300., 45500., 40200., 40700.,   600.,   700.,
       27800., 70000., 40100., 40900., 27000.,    nan, 40500., 50400.,
       80400., 60900., 60100., 70200., 30000., 40300., 40800., 70800.,
         900., 45000., 40600., 50700., 80500., 27700., 35800., 26600.,
       60000., 27900., 60800., 40000., 70100., 38100., 38200., 30200.,
       38500., 50900., 35100., 38700., 38300., 38900., 23200., 24700.,
       30800., 28100., 38400., 38000., 23100., 30100., 25700., 70400.,
         200., 30900., 26100., 45700.,   400., 28200., 70900., 38800.,
       45200., 26800., 60400., 30300., 25800.,   800., 38600., 24500.,
       30700., 28500., 24900., 70500., 35900., 30500., 28900., 27200.,
       45900., 25300., 35600., 25000., 25200., 26300., 35700., 23800.,
       32700., 30600., 45600., 28000., 32600., 28600., 25100., 26700.,
       24200., 25900., 22100., 25400., 28400., 28300., 25500., 24000.,
       32400., 28800., 32300., 32900., 27600., 35500., 23500., 45400.,
       28700., 32800., 23700., 32500., 35300., 45800., 27300., 22900.,
       29800., 30400., 35400., 24300., 29200., 24600.,   100., 50300.,
       32000., 32200., 24400., 32100., 22600., 26900.,   300., 20000.,
       26200., 22300., 25600.,   500., 35200., 35000.])
```

Fig: 1.3.2 Partner_salary

Observations on Partner_salary

In the "Partner_salary" column, we can see the string 'nan' in some values, which needs to be corrected. There are 16 rows in the "Partner_salary" column where the individual's "Marital_status" is "Single" and the value is 'nan'. Therefore, it is evident that individuals who are single have no "Partner_salary". We will change these 'nan' values to zero. Now, the number of missing values has reduced from 106 to 90.

There are 74 rows in the "Partner_salary" column where the individual's "Marital_status" is "Married" and "Partner_working" is "No", and the value is 'nan'. Therefore, it is evident that individuals whose partners are not working have no "Partner_salary". We will change these 'nan' values to zero. From 90 missing values, it has now reduced to only 16 missing values.

Of the remaining 16 values in the "Partner_salary" column, these represent actual missing data. To address this, I will replace zeros with NaN values. There are no missing values in the "Partner_salary" column.

1.3.3 Let's check the missing values in the data

```
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
Total_salary       0
Price              0
Make               0
dtype: int64
```

Fig: 1.3.3 Missing Values

Observations on missing values in the data

No Missing value is found in the given datasets.

2. Exploratory Data Analysis (EDA)

2.1 Univariate Analysis

2.1.1 Gender

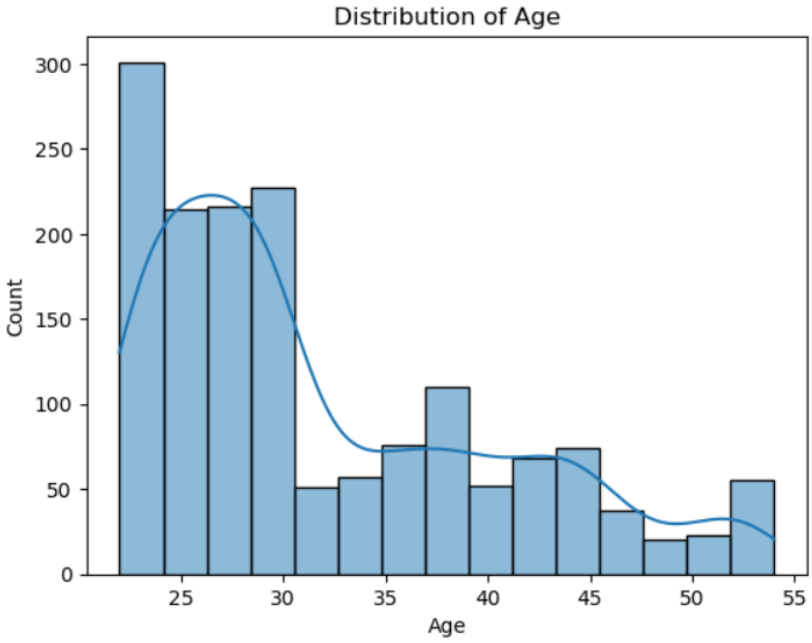

Fig: 2.1.1 Gender

Observations on Age.

The distribution is skewed towards right.

It is clear that age around 20 to 30 years individual are buying more cars than others.

2.1.2 No_of_Dependents

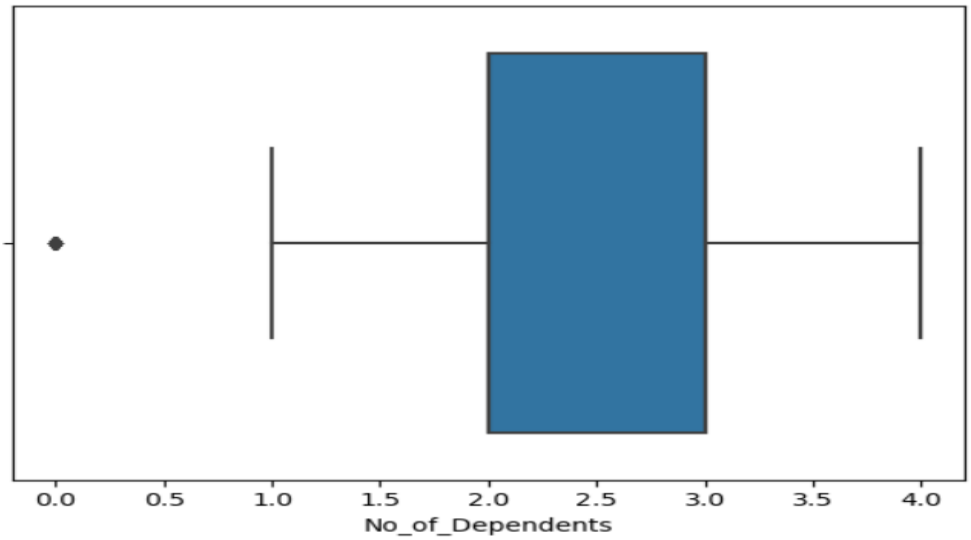

Fig 2:1.2 No_of_Dependents

Observations on No_of_Dependents

75% of car owners have 3 or fewer dependents.

The number of people with 2 and 3 dependents is the same, with each group having more than 500 people.

The number of people with 1 and 4 dependents is also the same, with each group having more than 200 people.

There are 30 people with 0 dependents.

From the above box plot, it is clear that there are outliers.

Rectifying the Outliers

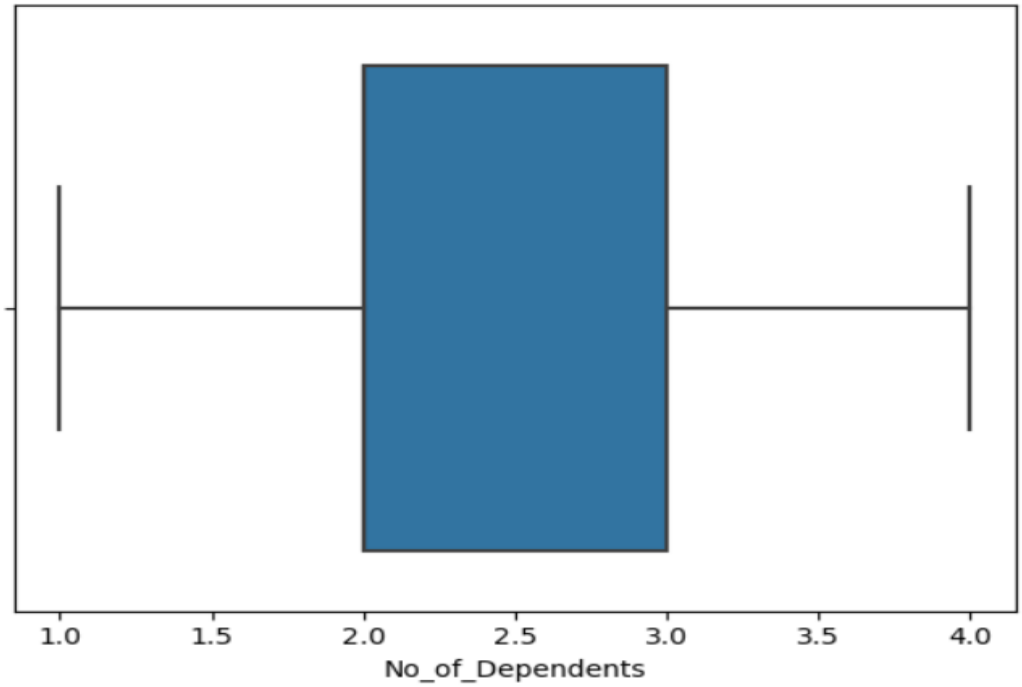

Fig: 2.1.2.1 No_of_Dependents

Observations on No_of_Dependents after rectifying the outliers

After rectifying the outliers, the above plot clearly shows that there are no outliers.

2.1.3 Total_salary

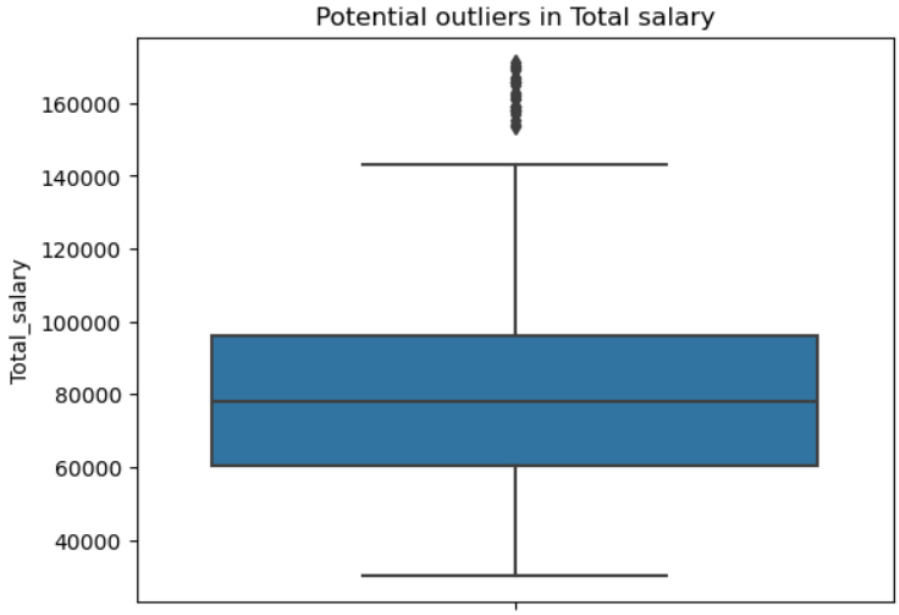

Fig: 2.1.3 Total_salary

Observations on Total_salary

From the above box plot it is clear that Total_salary has some outliers.

Rectifying the Outliers

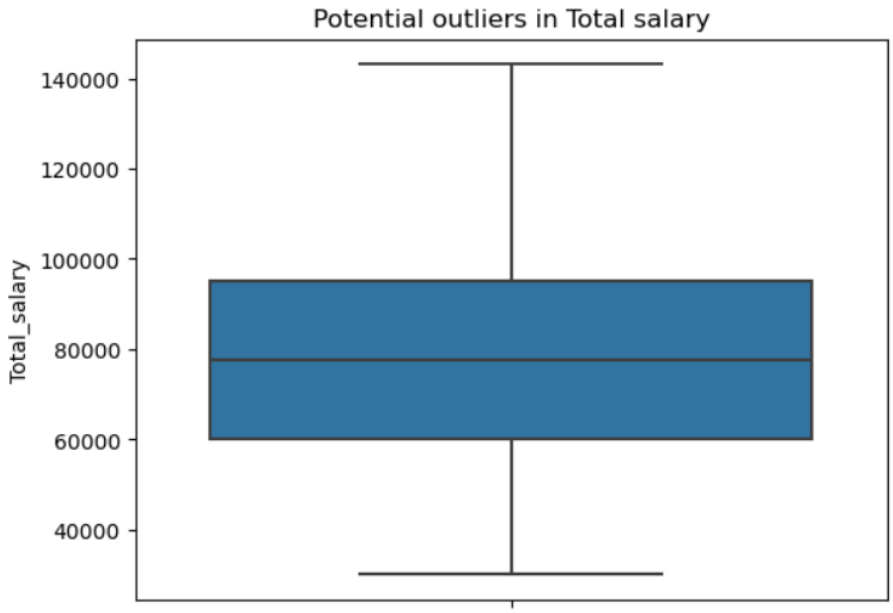

Fig: 2.1.3.1 Total_salary

Observations on Total_salary after rectifying the outliers

There is no outlier in the above plot and tha data are symmentrical.

2.1.4 Make

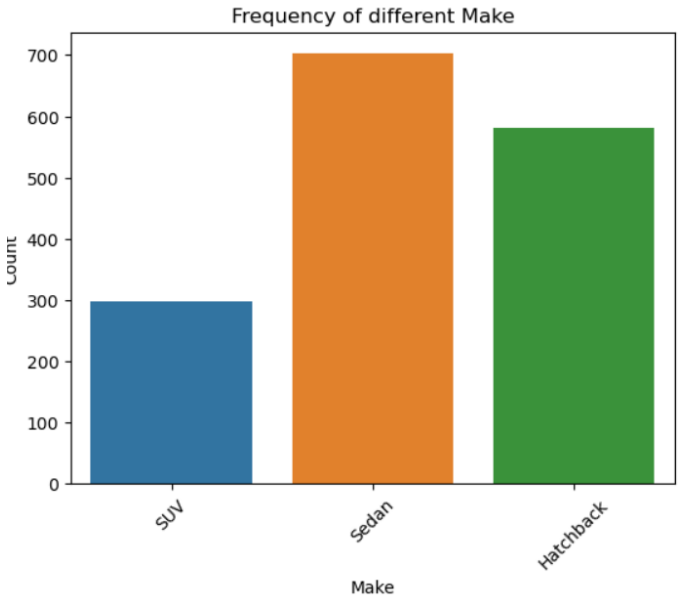

Fig: 2.1.4 Total_salary

Observations on Make

The most common type of car owned is a Sedan.

2.2 Bivariant Analysis

2.2.1 Numerical vs. Numerical: Relationship between Age and Salary

Fig: 2.2.1 Numerical vs. Numerical: Relationship between Age and Salary

Observations on Relationship between Age and Salary

From the above graph it is clear that,there is a positive correlation between the variables.

2.2.2 Categorical vs. Numerical: Relationship between Make and Total_salary

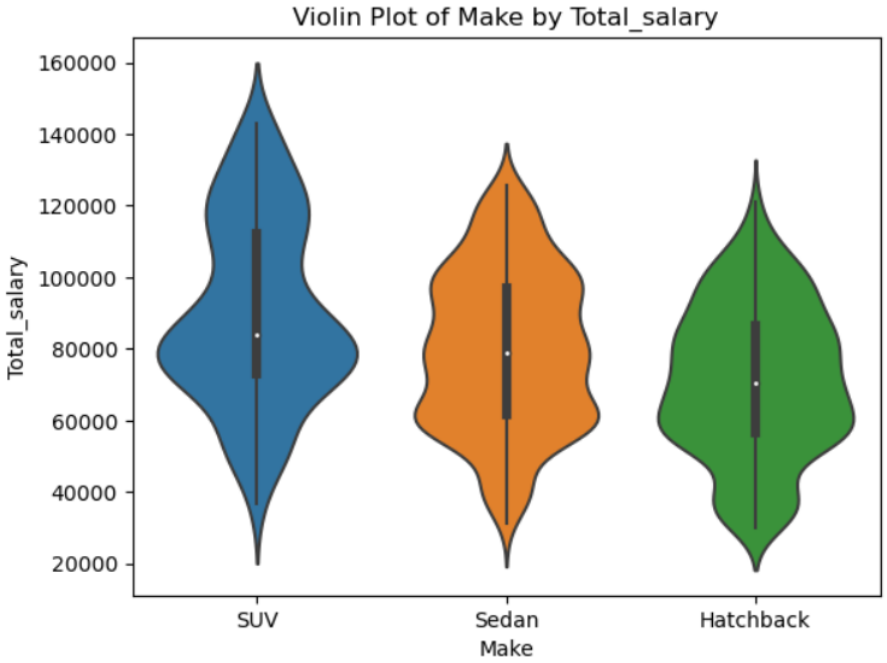

Fig: 2.2.2 Categorical vs. Numerical: Relationship between Make and Total_salary

Observations on Relationship between Make and Total_salary

From the above plot, it is clear that, there is no outliers. Based on the Total_salary, suv is the highest Preference among the individual.

2.2.3 Categorical vs. Categorical: Relationship between Marital_status and Partner_salary

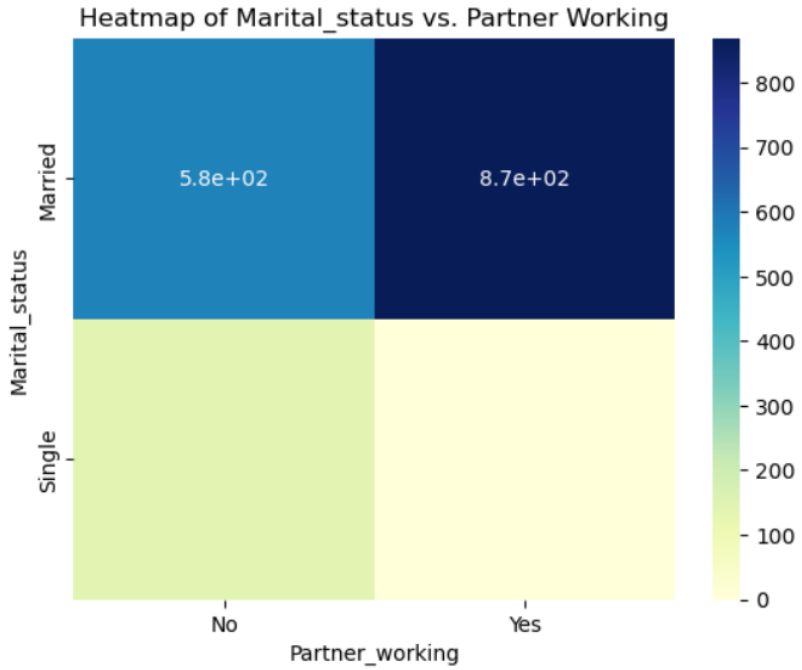

Fig: 2.2.3 Categorical vs. Categorical: Relationship between Marital_status and Partner_salary

Observations on Relationship between Marital_status and Partner_salary

From the above heat map, it is clear that individual who married and their partner who working purchase more car than others.

2.3 Multivariant Analysis

2.3.1 Relationship between Marital status, Salary and Profession.

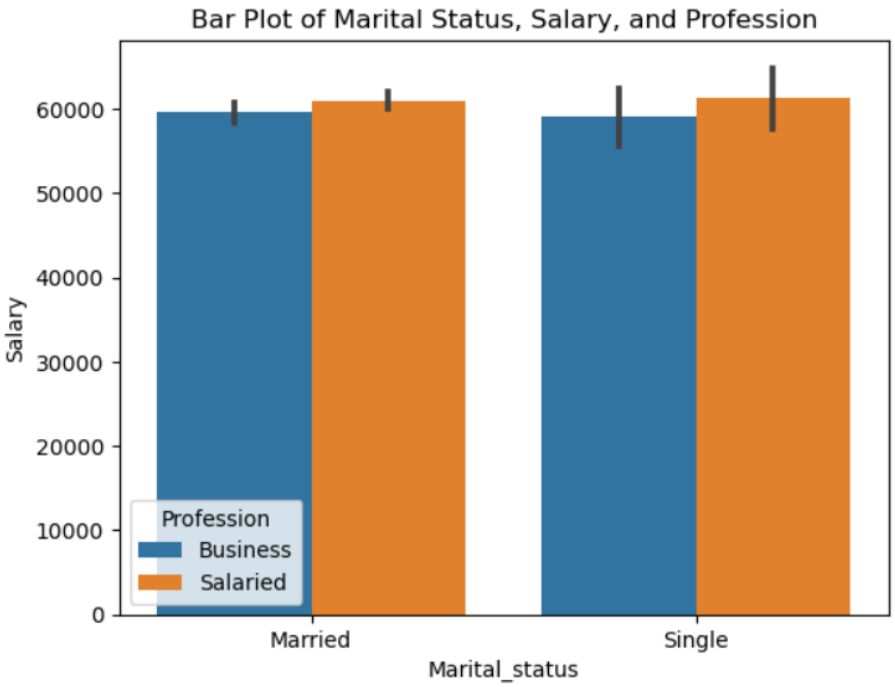

Fig :2.3.1 Relationship between Marital status, Salary and Profession.

Observation on Relationship between Marital status, Salary and Profession.

Comparing to single, Married people perfer more to buy a car.

Married people working professional more in salaried than business.

2.3.2 Relationship between Age, Salary, Partner_salary, Total_salary and Price

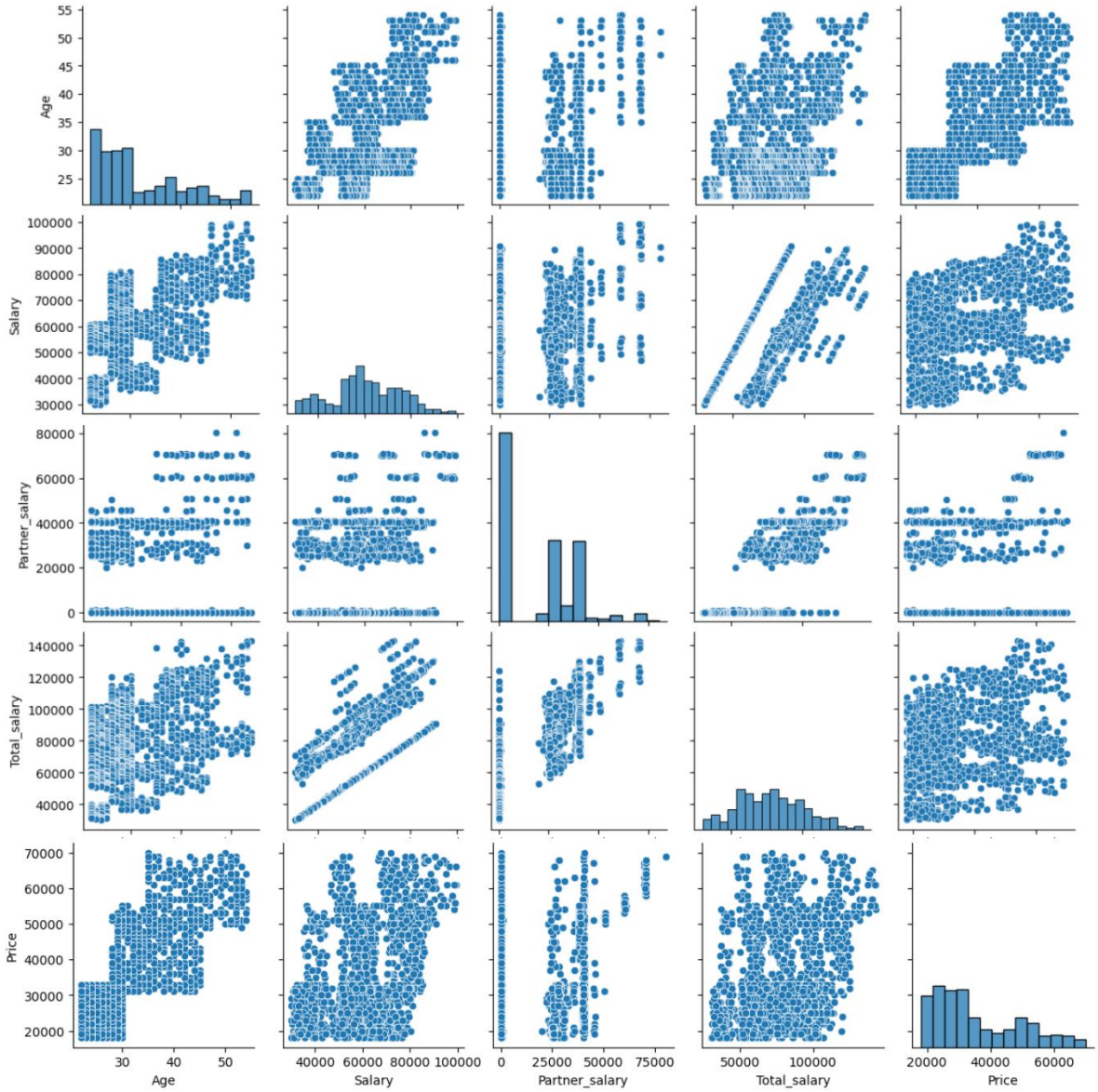

Fig: 2.3.2 Relationship between Age, Salary, Partner_salary, Total_salary and Price

Observation on Relationship between Age, Salary, Partner_salary, Total_salary and Price

From the pairplot above, the Age, salary and Price columns are highly Positively correlated.

3 Key Questions

3.1 Problem 1

Do men tend to prefer SUVs more compared to women?

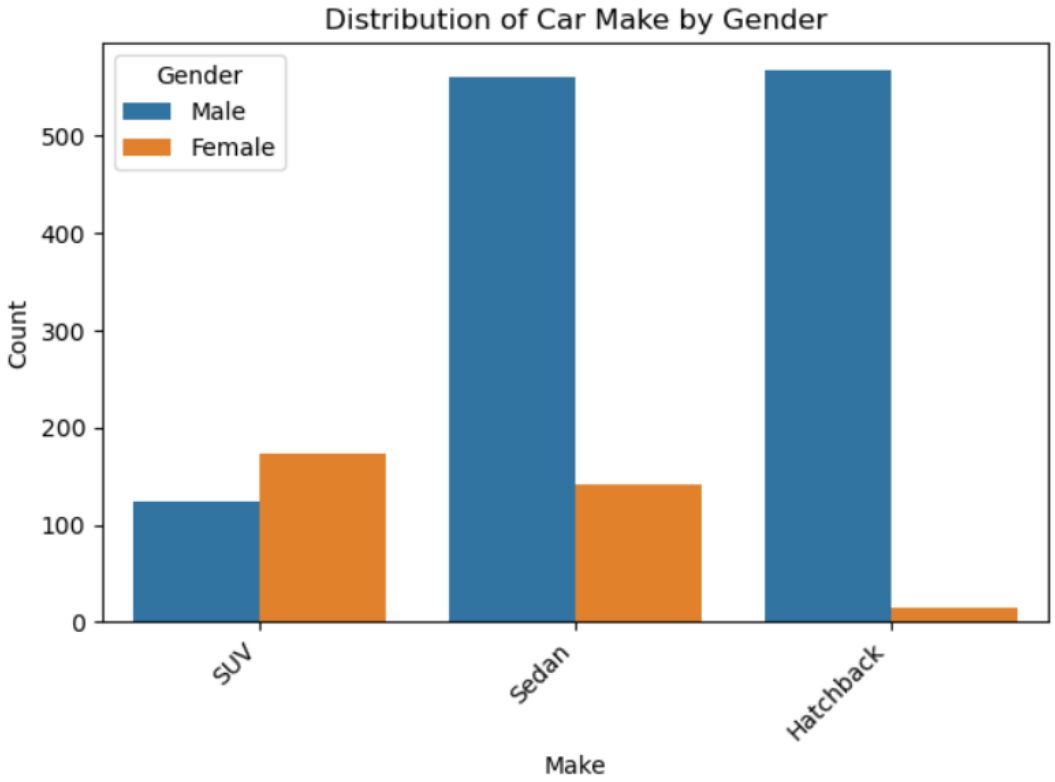

Fig: 3.1 Distribution car by Gender.

Observation

No. Men prefer SUVs less compared to women.

3.2 Problem 2

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

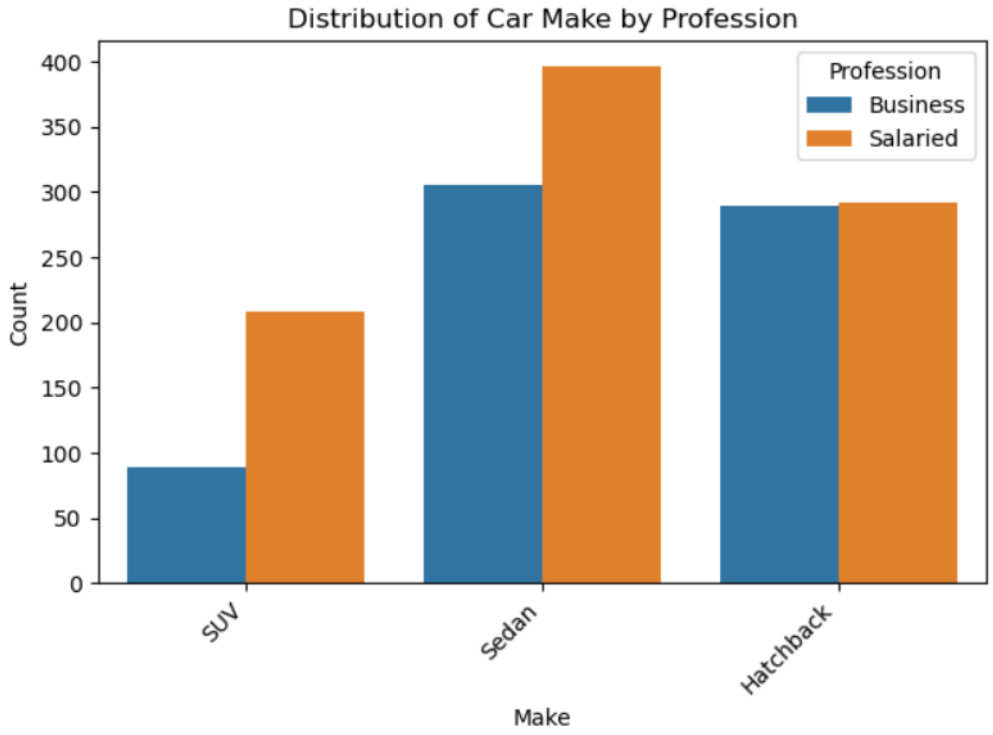

Fig: 3.2 Distribution car by Profession.

Observation

It is clearly visible from the bar graph that salaried person is more likely to buy a sedan.

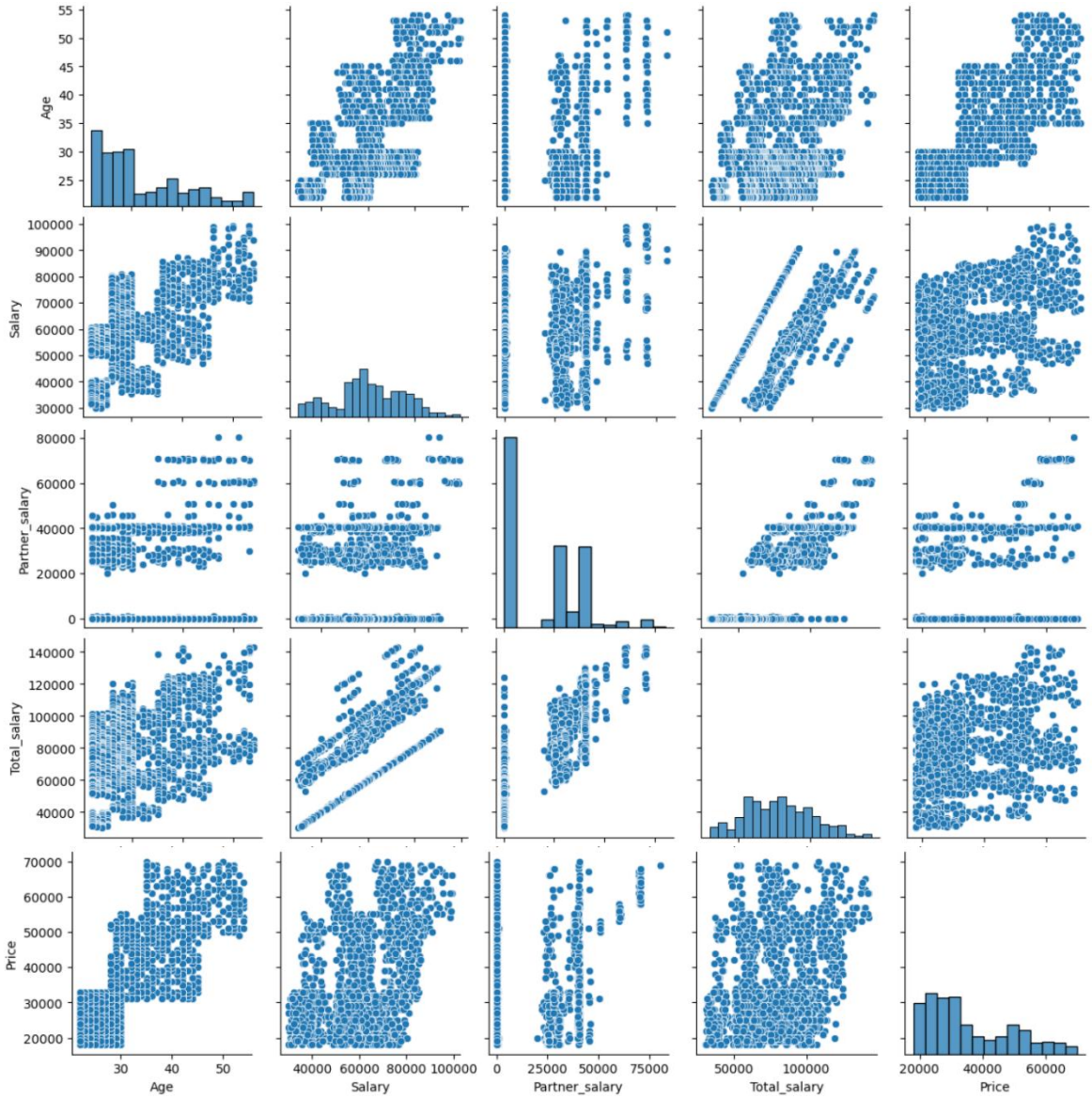

Fig: 3.2.1 Relationship of Data using pair plot.

Observation

Our findings indicate that age and price significantly increase the likelihood of buying a Sedan among salaried individuals. These insights can help automotive companies and marketers target the right demographic segments more effectively.

3.3 Problem 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?

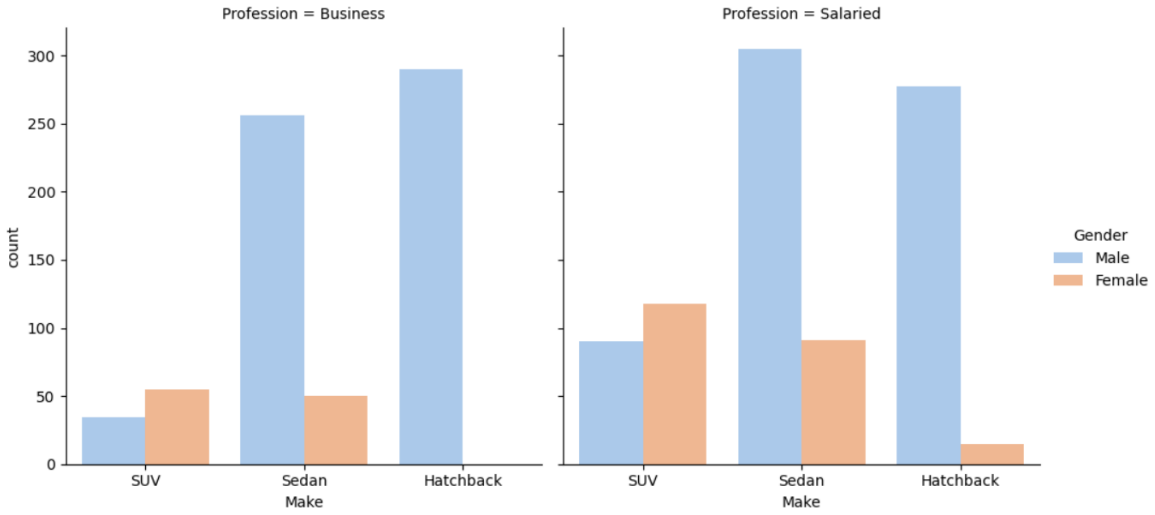

Fig :3.3 Relation between Make and Profession

Observation

It is clearly visible from the chart above that salaried males prefer 'Sedans' much more than 'SUVs'.

3.4 Problem 4

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

```
Make       Gender
Hatchback  Female      412000.0
           Male      14996000.0
SUV        Female     9252000.0
           Male       7328000.0
Sedan      Female     6031000.0
           Male      18261000.0
Name: Price, dtype: float64
```

Fig: 3.4 The amount spent on purchasing automobiles vary by gender

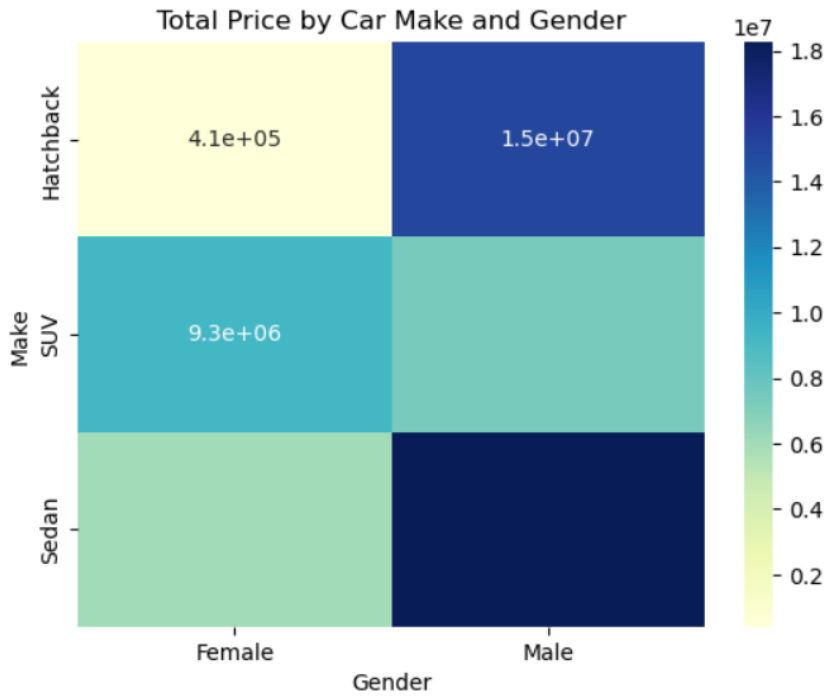

Fig 3.4.1 Heat map for Make Gender and Price

Observation

The amount spent on 'SUV' by females is more than that spent by males.

The amount spent on 'Hatchback' is more than triple that spent by females.

The amount spent on 'Sedan' is almost triple that spent by females.

3.5 Problem 5

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

```
Personal_loan  Make
No             Hatchback     7765000.0
               SUV          10373000.0
               Sedan        10852000.0
Yes            Hatchback     7643000.0
               SUV           6207000.0
               Sedan        13440000.0
Name: Price, dtype: float64
```

Fig: 3.5 Money was spent on purchasing automobiles by individuals who took a personal loan

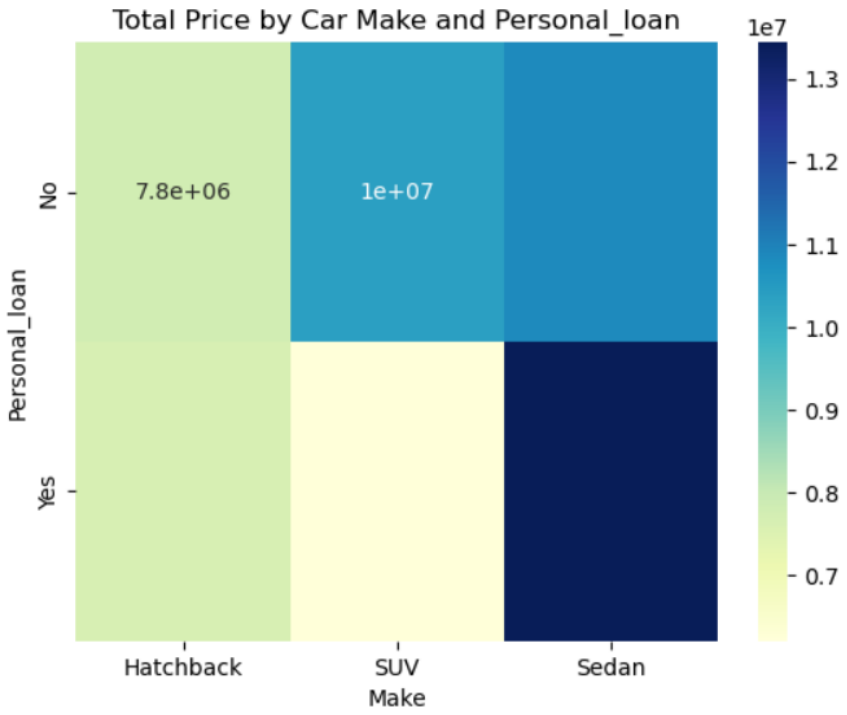

Fig 3.5.1 Heat map for Make Personal_loan and Price

Observation

The amount spent on SUVs by females is higher than that spent by males.

The amount spent on Hatchbacks is more than triple for females compared to males.

The amount spent on Sedans is almost triple for females compared to males.

3.6 Problem 6

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

```
Partner_working  Make
No               Hatchback    26323.843416
                 SUV          56173.611111
                 Sedan        35354.166667
Yes              Hatchback    26614.617940
                 SUV          55496.732026
                 Sedan        34082.125604
Name: Price, dtype: float64
```

Fig: 3.6 Working partner influence the purchase of higher-priced cars

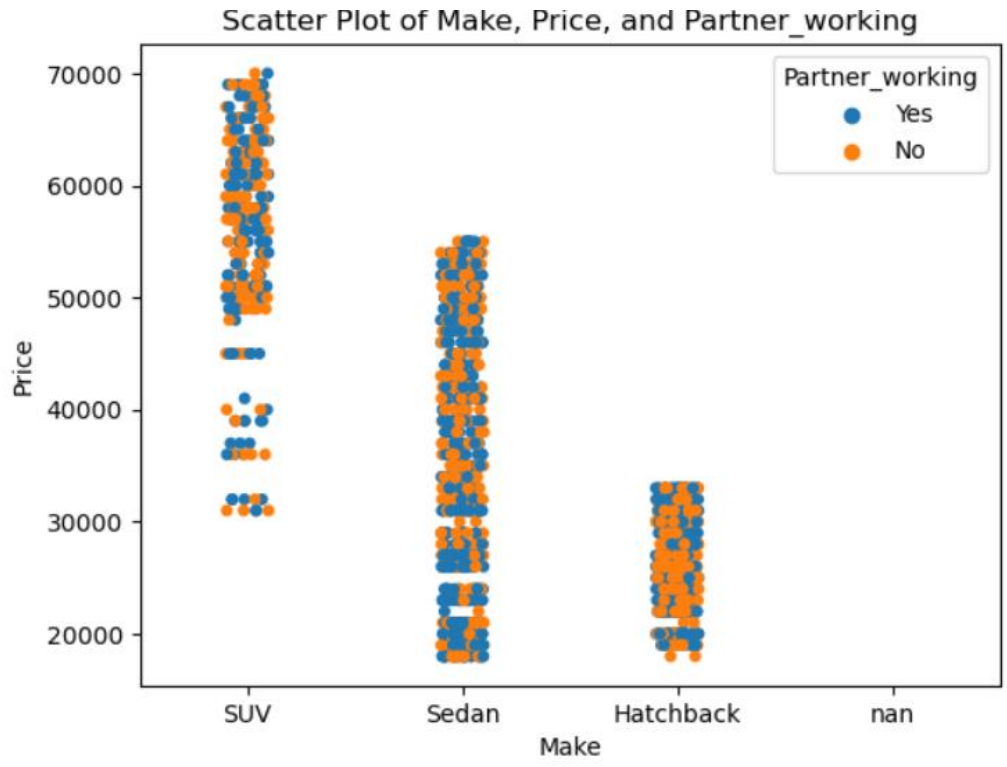

Fig: 3.6.1 Scatter plot of Make price and Partner_working

Observation

Yes, based on the analysis of the data, it seems that buyers with a working partner tend to purchase cars with a higher price. This trend is evident from the patterns observed in the data.

4 Actionable Insights

Develop marketing campaigns that highlight the features and benefits of SUVs to women, emphasizing their preferences and potential for higher sales in this segment.

Consider developing SUV models that cater specifically to women's preferences, such as offering more customization options, safety features, and interior design elements that appeal to them.

Segment the market based on gender preferences and tailor marketing strategies accordingly. For example, focus more on promoting sedans to men and SUVs to women.

Offer special promotions and discounts on SUVs targeted towards women to encourage them to consider these vehicles over sedans.

Enhance the customer experience for women buying SUVs by providing personalized services, such as test drives, financing options, and after-sales support, to increase satisfaction and loyalty.

Collaborate with influencers or organizations that cater to women's interests to promote SUVs as a stylish and practical choice for them.

5 Recommendations

Marketing Strategy: Tailor marketing efforts to highlight the features of SUVs that appeal to women, such as safety, comfort, and versatility. Use targeted advertising channels and messages to reach this demographic effectively.

Product Offering: Expand the range of SUV models available to cater to different preferences among women. Consider offering more customization options and packages that appeal to women's tastes.

Sales Approach: Train sales staff to understand and address the specific needs and preferences of female customers when selling SUVs. Offer test drives and demonstrations that showcase the features that are most important to women.

Promotions and Incentives: Create special promotions and incentives to attract women to SUVs, such as discounts, financing offers, and loyalty programs. Highlight the value and benefits of SUVs for women in promotional materials.

Customer Experience: Enhance the overall customer experience for women purchasing SUVs by providing a welcoming and supportive environment. Offer amenities and services that cater to women's needs, such as child-friendly facilities and female sales representatives.

Market Research: Conduct further market research to gain a deeper understanding of women's preferences and purchasing behaviours related to SUVs. Use this information to refine product offerings and marketing strategies to better meet the needs of female customers.