Project Report – SMDM

Problem – 1:

Question. A: What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables)

Answer: The provided dataset **'austo_automobile.csv'** is about the **Austo Motor Company**, which is a leading car manufacturer in the market, specializing in **Sedan**, **SUV,** and **Hatchback** models. All technical specifics of the dataset that I used to make my findings are listed below. A database administrator would be interested in this information since it gives them an overview of the dataset and helps them comprehend the nature of the data. Please go through the screenshot as well for the reference:

- The dataset has **1581 rows** and **14 columns**.
- The data has **5 integers**, **1 float**, and **8 object** data type columns.
- The **memory usage** of the provided csv file (i.e., austo_automobile.csv) is **173.1 KB**.

```
<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
```

```
df.shape
```

```
(1581, 14)
```

- Head function is used to gain a brief idea about the dataset.

```
df.head().T
```

	0	1	2	3	4
Age	53	53	53	53	53
Gender	Male	Femal	Female	Female	Male
Profession	Business	Salaried	Salaried	Salaried	Salaried
Marital_status	Married	Married	Married	Married	Married
Education	Post Graduate	Post Graduate	Post Graduate	Graduate	Post Graduate
No_of_Dependents	4	4	3	2	3
Personal_loan	No	Yes	No	Yes	No
House_loan	No	No	No	No	No
Partner_working	Yes	Yes	Yes	Yes	Yes
Salary	99300	95500	97300	72500	79700
Partner_salary	70700.0	70300.0	60700.0	70300.0	60200.0
Total_salary	170000	165800	158000	142800	139900
Price	61000	61000	57000	61000	57000
Make	SUV	SUV	SUV	SUV	SUV

Question. B: Take a critical look at the data and do a preliminary analysis of the variables. Do a quality check of the data so that the variables are consistent. Are there any discrepancies present in the data? If yes, perform preliminary treatment of data.

Answer: After critically examining the dataset, there are several discrepancies were seen in my preliminary findings which were treated accordingly to get them ready for further analysis. The steps taken for the same are as follows:

1. To check for any duplicate data, I used the 'duplicated' function and found no duplicate values in the data.

```
Total Duplicate values :  0
```

2. There are **53** and **106 null values** in the **'Gender'** and **'Partner_Salary'** columns respectively.

```
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
```

3. All columns that have object datatype are examined for unique values. The findings are as follows:

Column Name	Unique values
Gender	Male, Femal, Female, nan, Femle
Profession	Business, Salaried
Marital_status	Married, Single
Education	Graduate, Post Graduate
Personal_loan	Yes, No
House_loan	Yes, No
Partner_working	Yes, No
Make	Hatchback, Sedan, SUV

4. In the 'Gender' column, Female is spelled incorrectly in several entries which have created 2 more additional unique values, for example, Femal and Femle.

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

5. The ‘replace’ function is used to remove the spelling error:

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

6. The data shows that there are 1199 Males and 329 Females in the **Gender** column. Apart from this, there are 53 Nan (Null values) from which we have to get rid of to do further analysis.

```
Gender
Female     329
Male      1199
Name: Gender, dtype: int64
```

7. To remove null values from the **Gender** column, I followed the approach where we replaced the null values with the most common element. The following screenshot shows the increased number of Males to 1252.

```
Gender
Female     329
Male      1252
Name: Gender, dtype: int64
```

8. To remove null values from the **Partner_salary** column, I checked for any relationship or connection between the **Salary**, **Partner_salary,** and **Total_salary**.

```
Is Salary + Partner_salary = Total_salary ?  [ True]
```

I found out the Total_salary is the sum of Salary and Partner_salary. Once the connection was established, it was a piece of cake to calculate the Partner_salary to replace null values.

```
Total Null values in Partner_salary :  0
```

9. After that, there are no null values and discrepancies found in any columns.

```
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
```

10. The statistical analysis of the data is as follows; This data seems to be fine with no abnormalities.

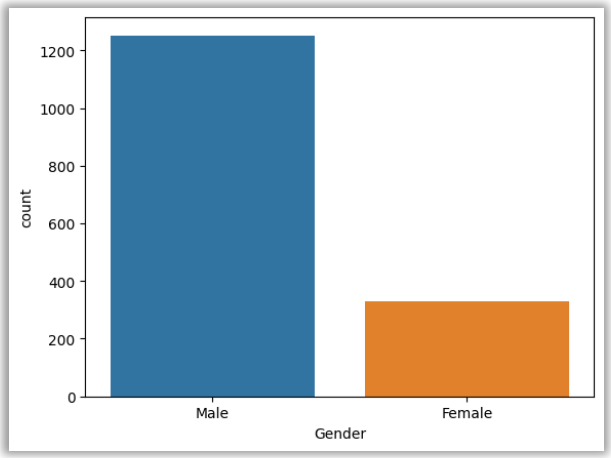
	mean	std	min	25%	50%	75%	max
Age	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	19233.776091	19670.391171	0.0	0.0	25100.0	38100.0	80500.0
Total_salary	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	35597.722960	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0

Question. C: Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.

Answer:

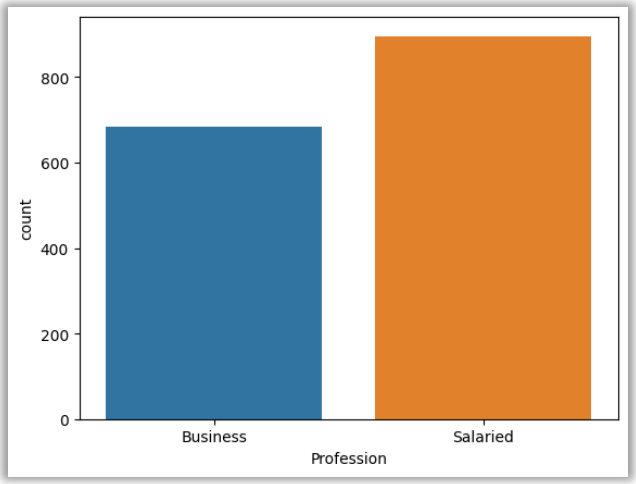
- The number of **Male** customers is much **higher** than Female customers.

```
Gender
Male      1252
Female     329
Name: count, dtype: int64
```



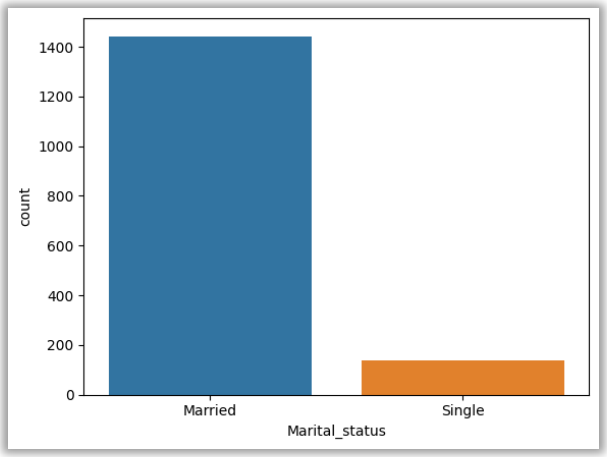
- The **Salaried** customers are **higher** in numbers than Business customers.

```
Profession
Salaried    896
Business    685
Name: count, dtype: int64
```



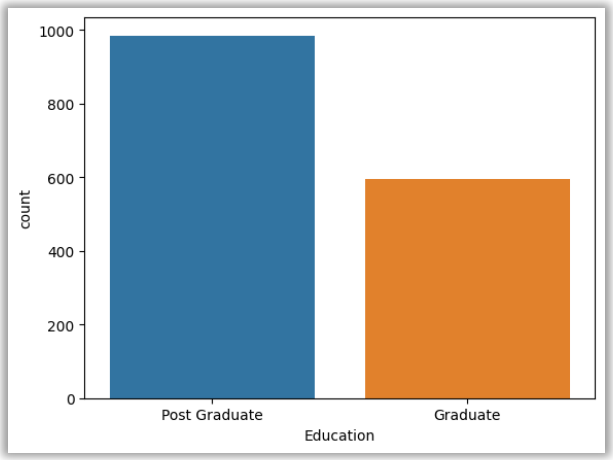
- The **Married** customers are **more** than Single customers with a huge difference.

```
Marital_status
Married    1443
Single      138
Name: count, dtype: int64
```



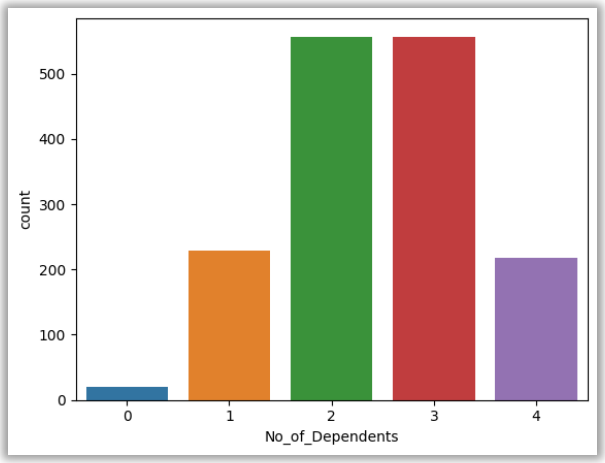
- Customers whose education is **Post Graduates** are buying **more** cars than Graduates.

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



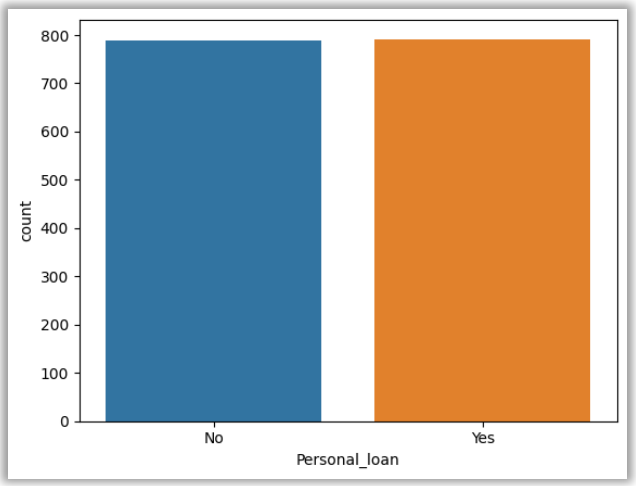
- Customers who have **2 & 3** number of **dependents** are buying **more** cars.
- Customers who have **no dependents** are **least** interested in buying cars.

```
No_of_Dependents
3    557
2    557
1    229
4    218
0     20
Name: count, dtype: int64
```



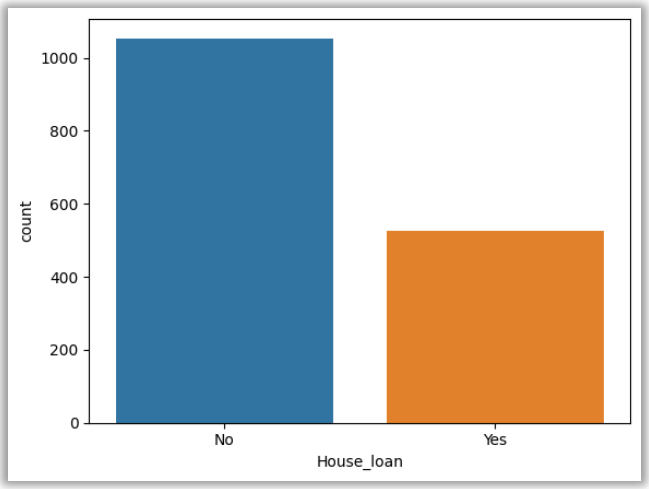
- There is a **minute difference** between the number of customers who **have personal loans** and those who **haven't**.

```
Personal_loan
Yes    792
No     789
Name: count, dtype: int64
```



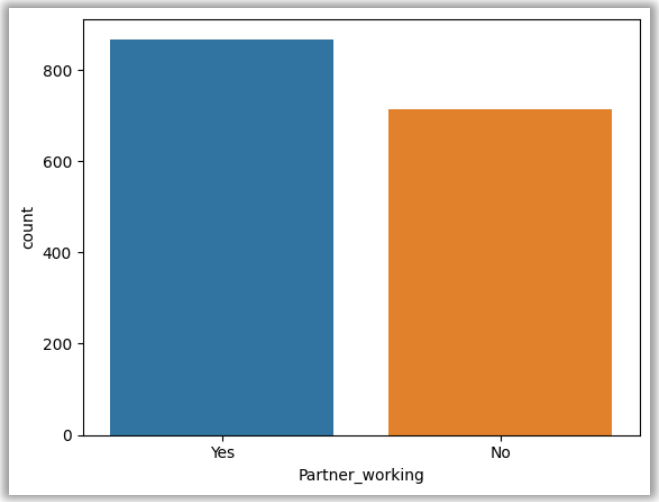
- Customers who have **no house loan** are **purchasing more** cars than the customers who are paying house loans.

```
House_loan
No     1054
Yes     527
Name: count, dtype: int64
```



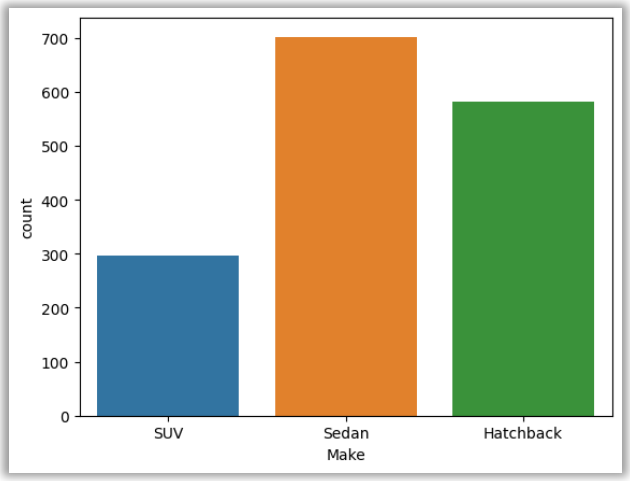
- Customers whose **partners** are **working** are **buying more** cars.

```
Partner_working
Yes    868
No     713
Name: count, dtype: int64
```



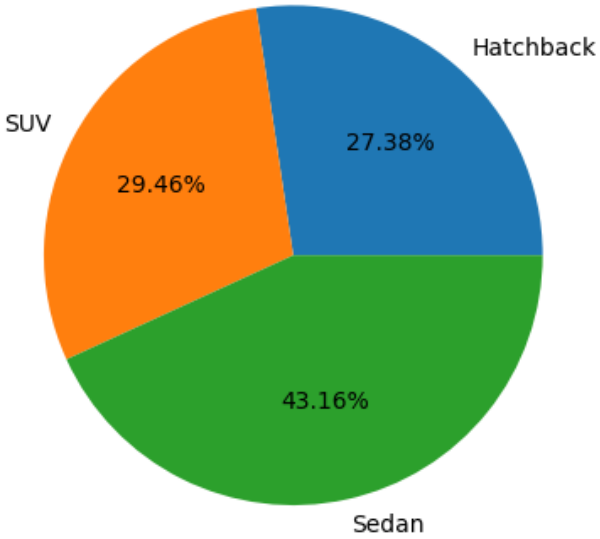
- Among all the types, **Sedans** are the **top choice** for customers followed by **Hatchbacks**.
- Although **SUVs** are the **least choice** but still make a **reasonable sale**.

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



- **Sedans** generates most of the company's revenue, with **SUVs** and **Hatchbacks** coming in second and third, respectively.

```
Make         Total_Revenue
Hatchback      15408000
SUV            16580000
Sedan          24292000
Name: Price, dtype: int64
```

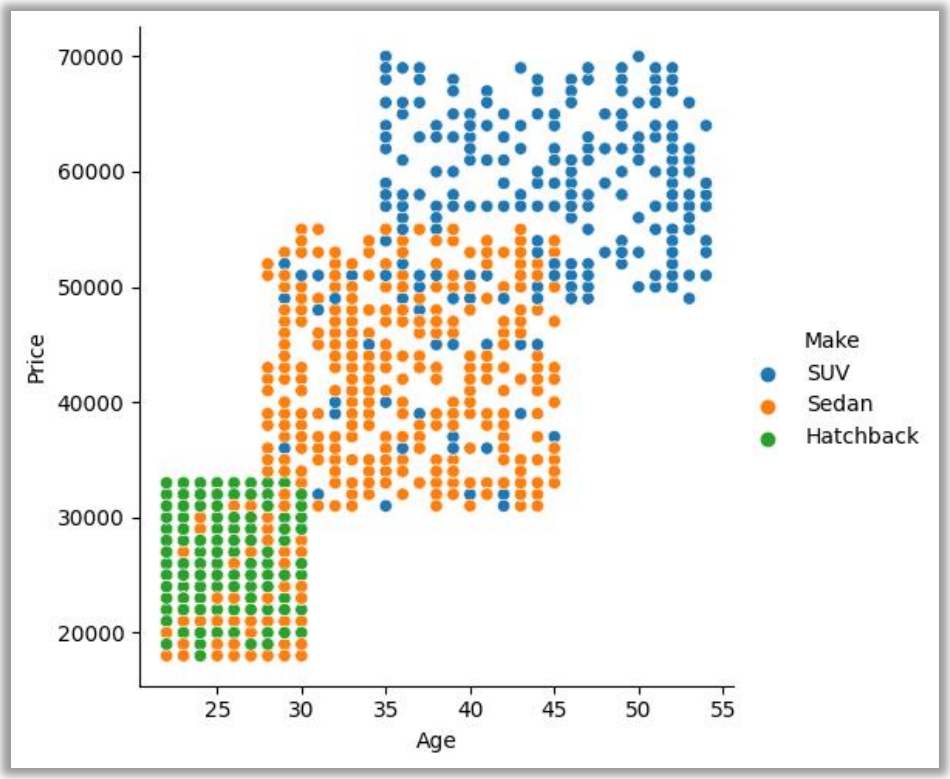


Question. D: Understanding the relationships among the variables in the dataset is crucial for every analytical project. Perform analysis on the data fields to gain deeper insights. Comment on your understanding of the data.

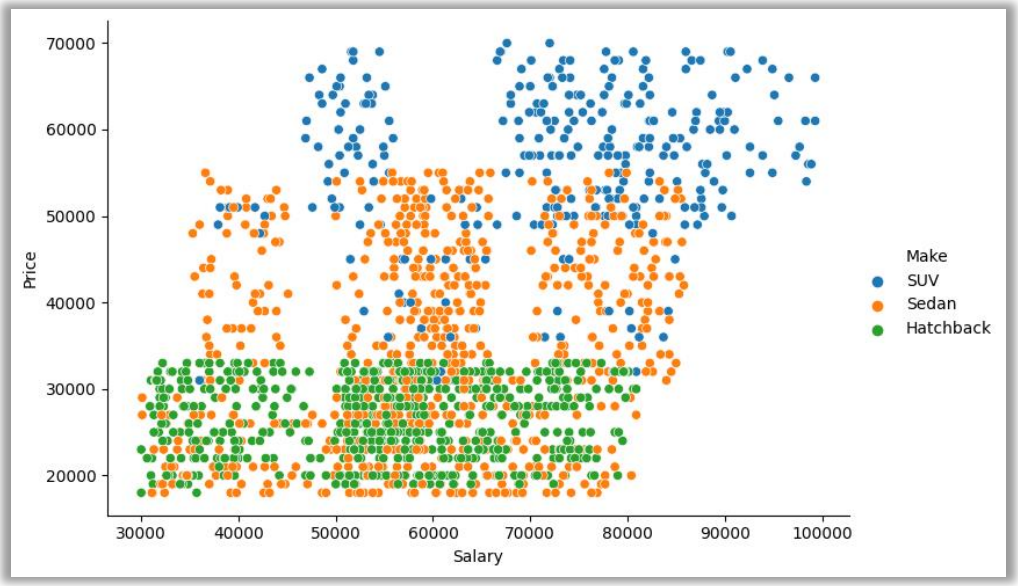
Answer:

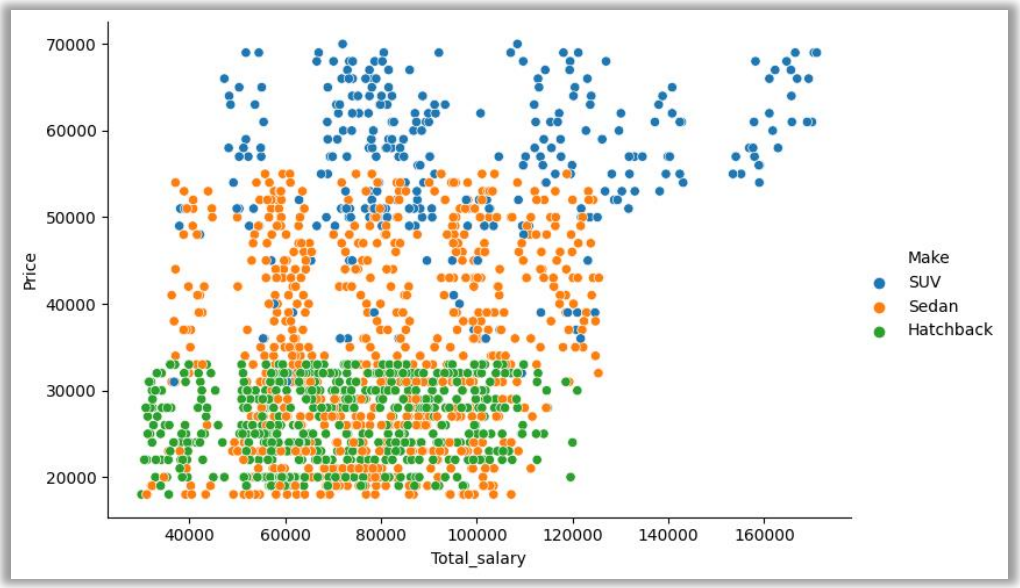
While performing analysis on the dataset to gain deeper insights, I found some interesting facts that are mentioned below in the graphs:

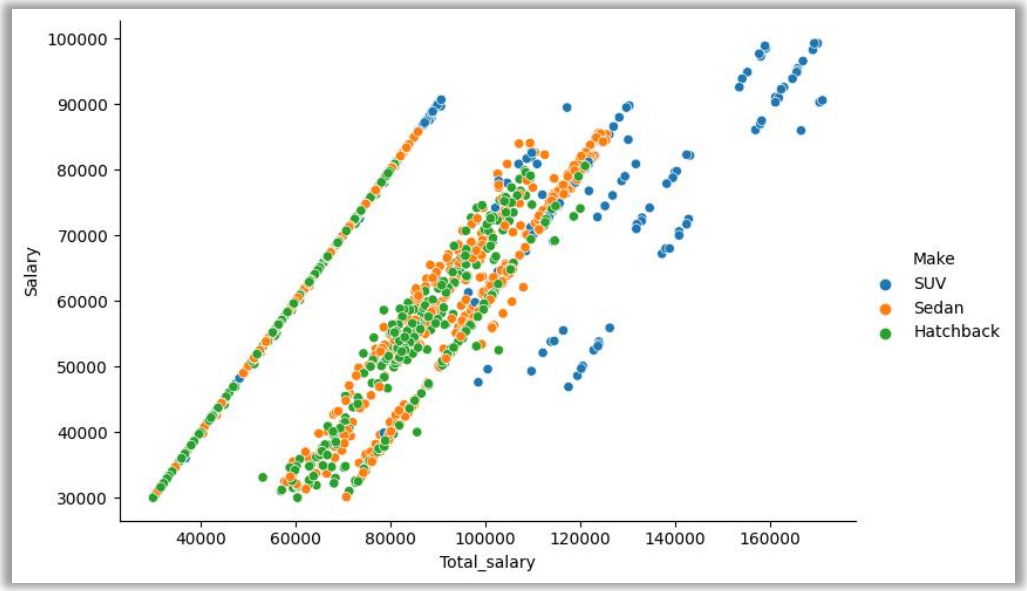
- The **age** group between **20 to 30** years prefers **Hatchbacks** and **Sedans.** The lowest sales of **SUVs** can be seen in this age group**.**
- The **age** group between **31 to 45** years prefers **Sedans and SUVs,** but **no sale** for **Hatchback** can be seen.
- The **age** group between **46 to 55** years prefers **SUVs only** and shown **no** interest in **Hatchbacks** and **Sedans.**



- The customers whose salary ranges between **30k to 45k** majorly purchased **Hatchbacks** and **Sedans**, and very few purchased **SUVs.**
- The **majority of buyers** are those whose salary ranges between **46k to 85k** and purchased cars from every segment (i.e. **Sedan, SUV, and Hatchback**).
- The customers whose salary ranges between **86k to 1 Lakh** majorly purchased **SUVs**, very few purchased **Sedans**, and shown **no** interest in buying **Hatchbacks**.
- The sales of more expensive cars are increased where the Partner is also working.
- There is a strong relationship can be seen between **Salary** and **Total Salary.**


Price
Salary
Make
SUV
Sedan
Hatchback


Price
Total_salary
Make
SUV
Sedan
Hatchback

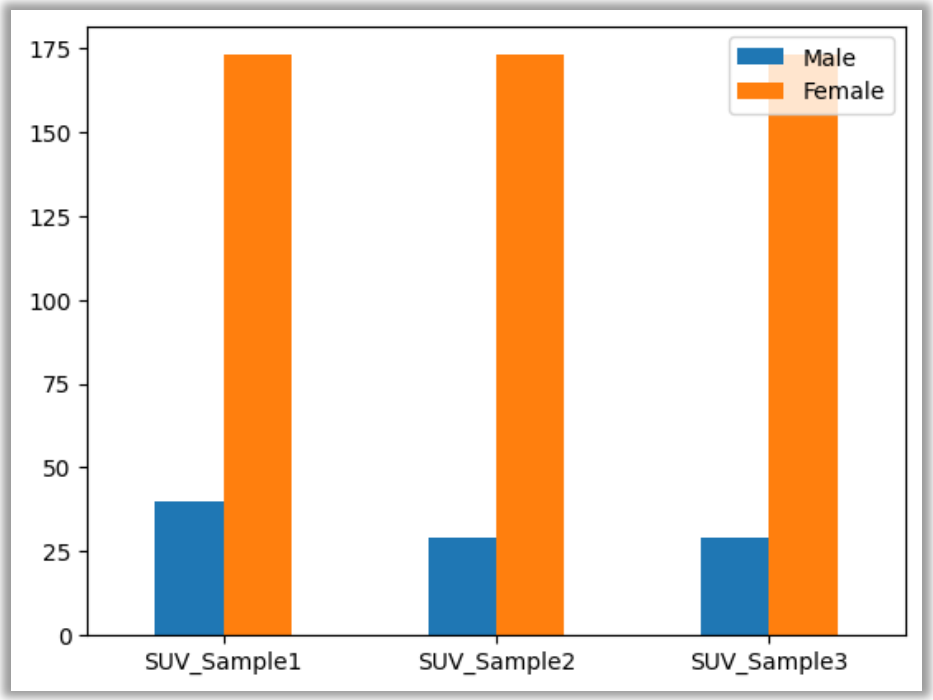

Salary
Total_salary
Make
SUV
Sedan
Hatchback

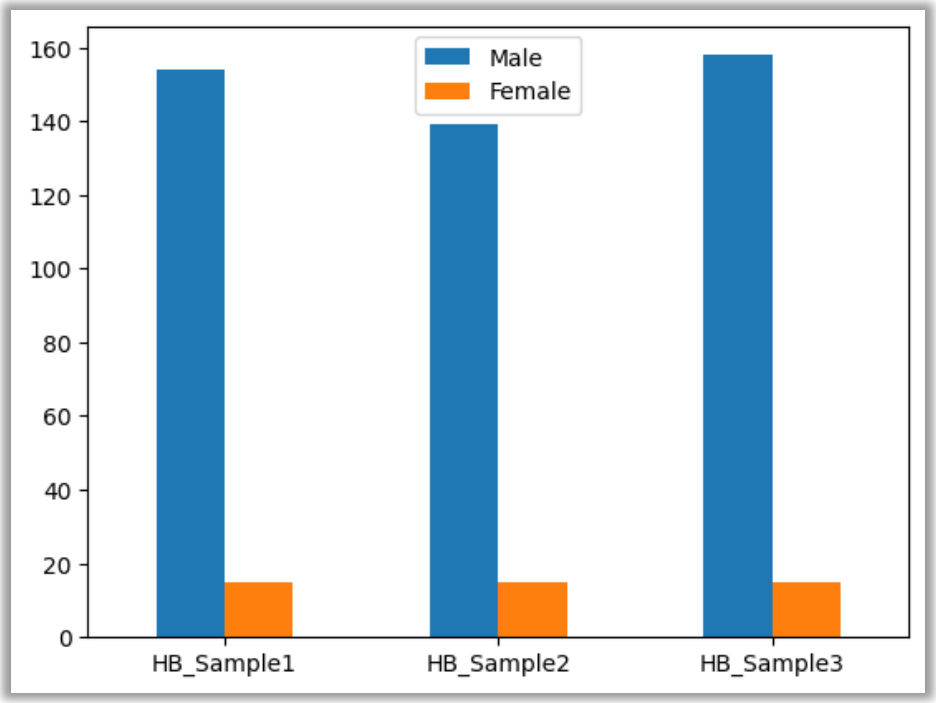
Question. E: Employees working on the existing marketing campaign have made the following remarks. Based on the data and your analysis state whether you agree or disagree with their observations. Justify your answer Based on the data available.

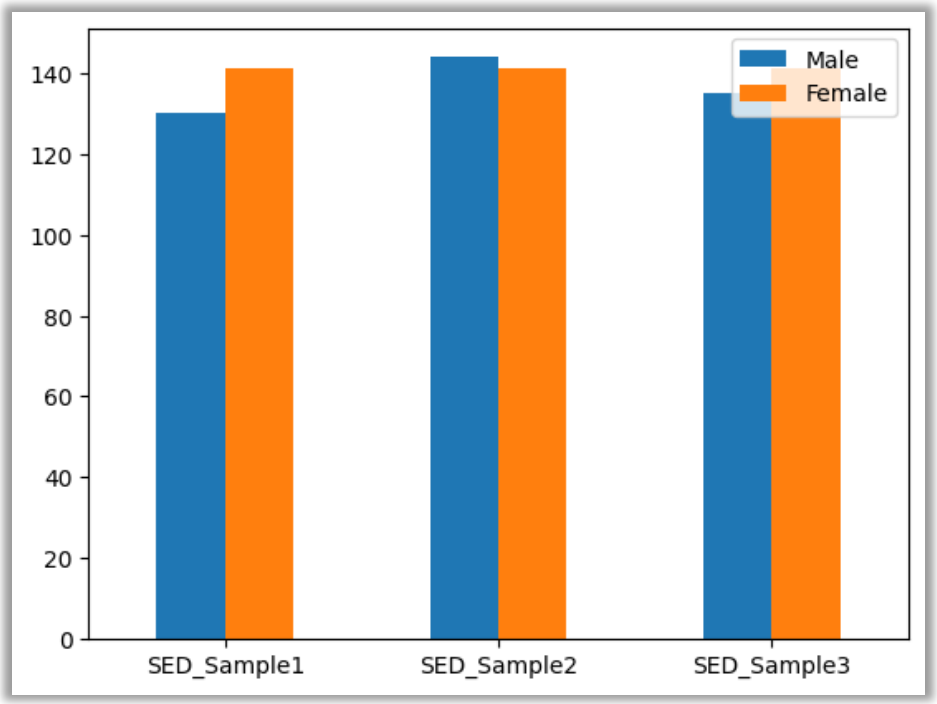
Answer:

E1) Steve Roger says "Men prefer SUV by a large margin, compared to the women".

I strongly disagree with Steve Roger as the graph shows women prefer SUVs more than men. In fact, Men are more predilected to Hatchback. At the same time, Sedan is the most prevalent among both Men and Women.



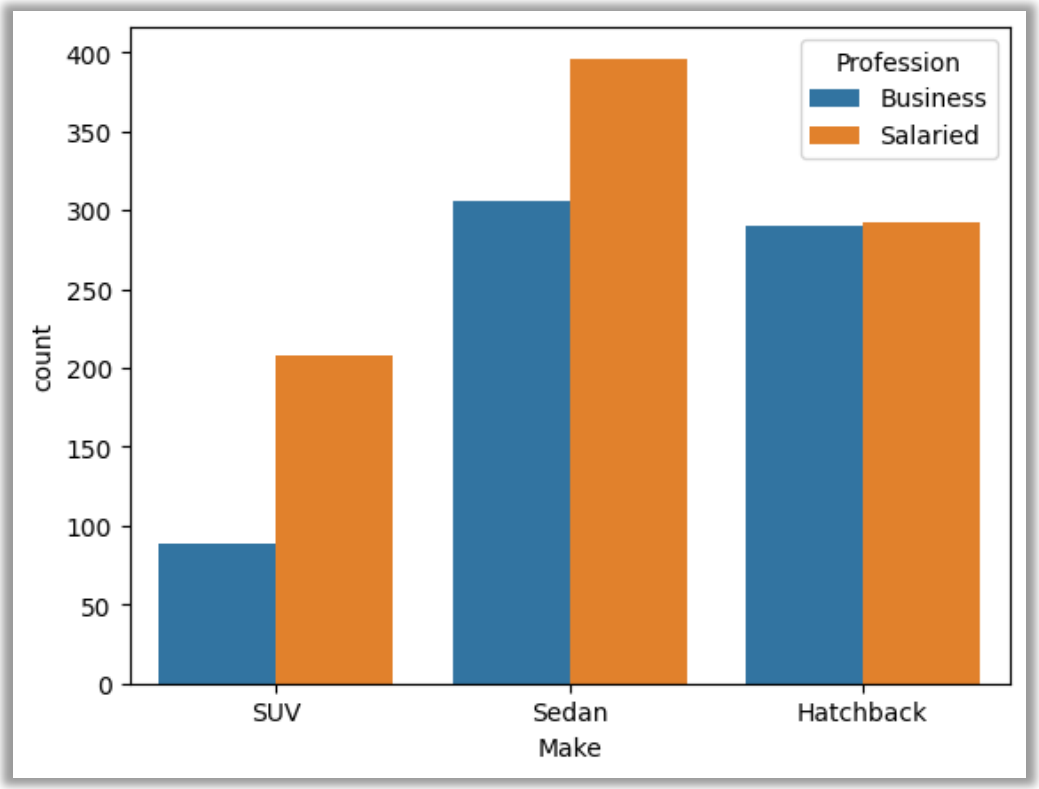



Male
Female
140
120
100
80
60
40
20
0
SED_Sample1
SED_Sample2
SED_Sample3

E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.

Yes, I concur with Ned Stark's belief, which is backed by the table and count plot below, showing that salaried people purchase more sedans than business professionals.

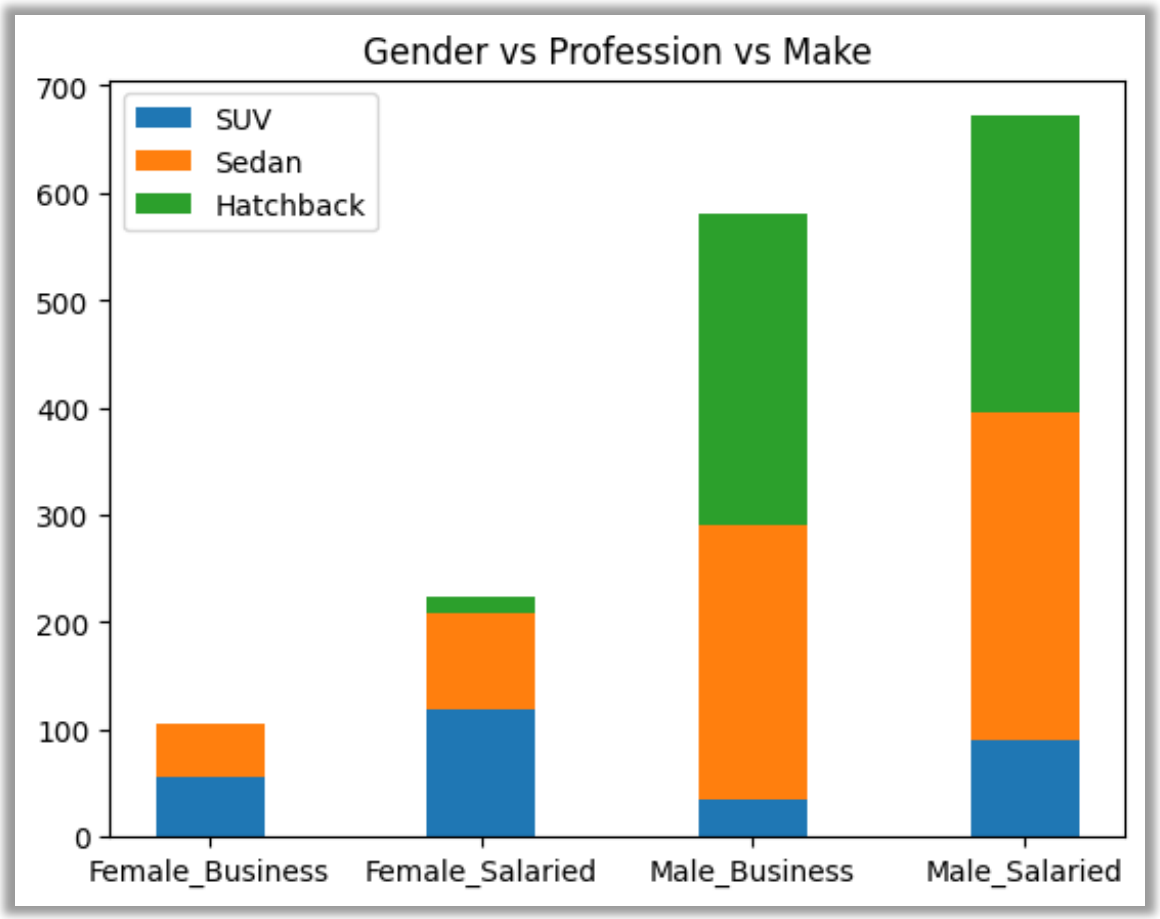
```
Profession  Make
Business    Hatchback    290
            SUV           89
            Sedan        306
Salaried    Hatchback    292
            SUV          208
            Sedan        396
Name: Make, dtype: int64
```



E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale.

No, Sheldon Cooper is mistaken, as shown by the facts below; Salaried Males prefer to purchase Sedans over SUVs.

```
Gender  Profession  Make
Female  Business    SUV          55
                    Sedan        50
        Salaried    Hatchback    15
                    SUV         118
                    Sedan        91
Male    Business    Hatchback   290
                    SUV          34
                    Sedan       256
        Salaried    Hatchback   277
                    SUV          90
                    Sedan       305
Name: Make, dtype: int64
```

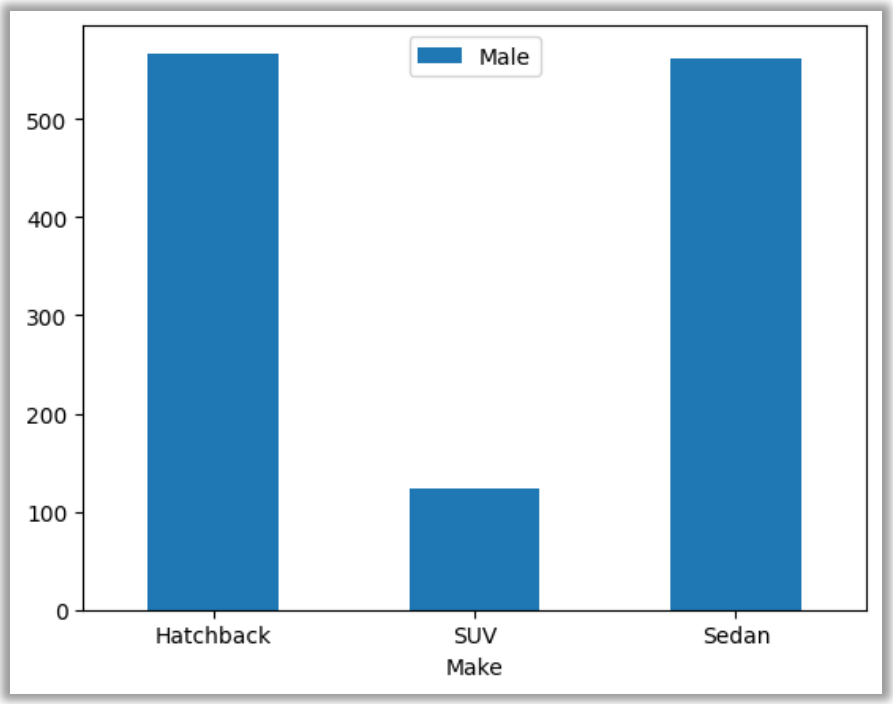


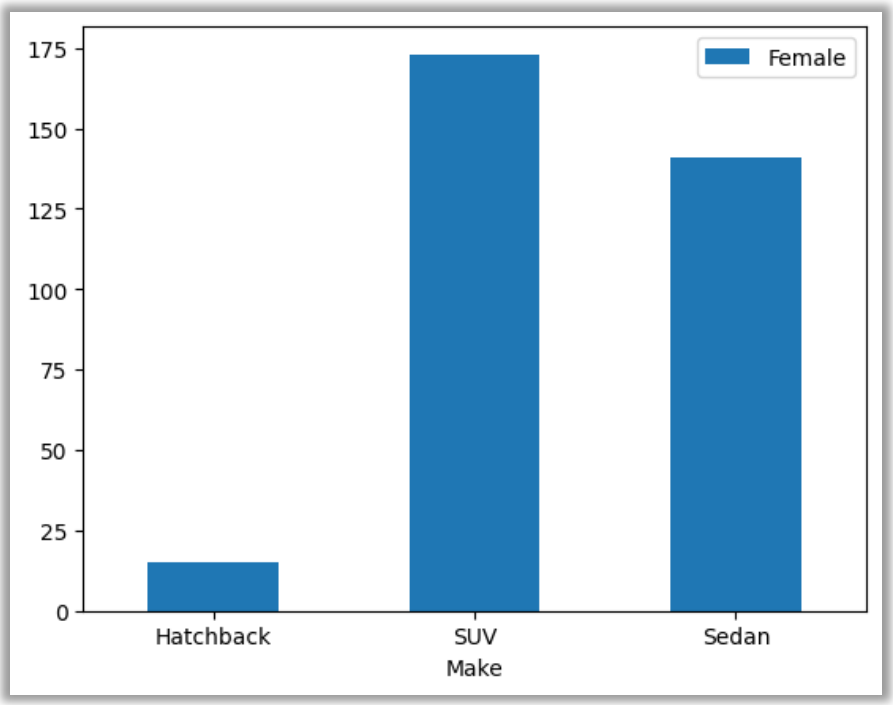
Question. F: From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a Business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions.

Answer:

F1) Gender

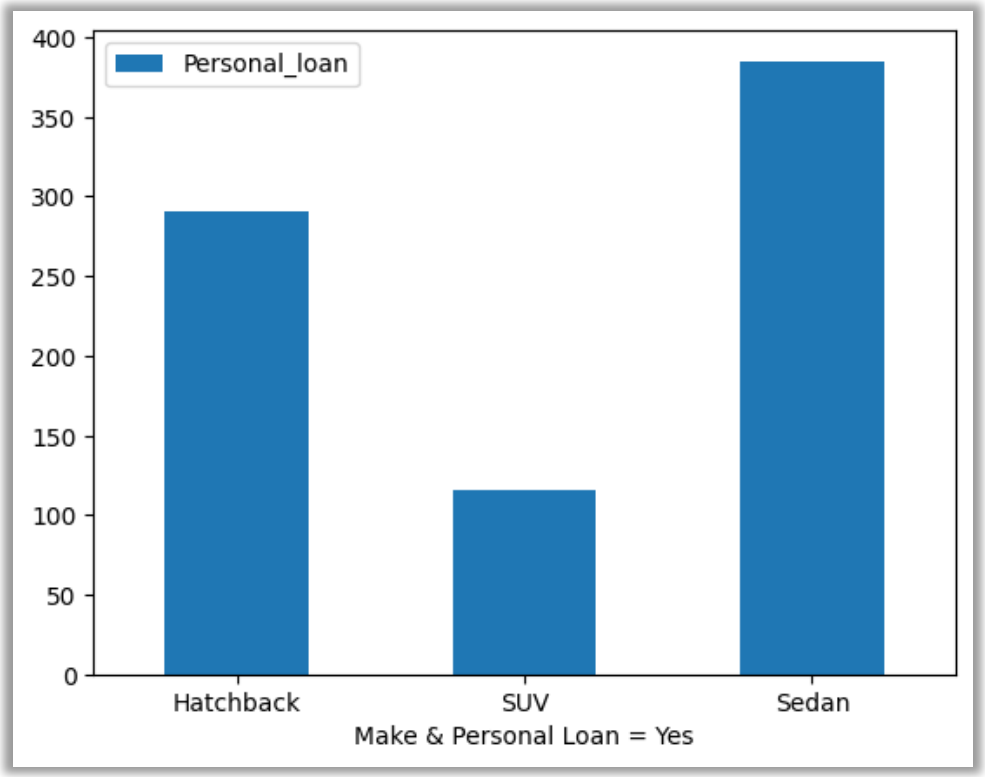
Males are pretty much interested in purchasing the Sedan and Hatchback rather than the SUV. However, Females are least interested in Hatchback, and prefers to go with SUV as their top priority after that their second choice is Sedan.

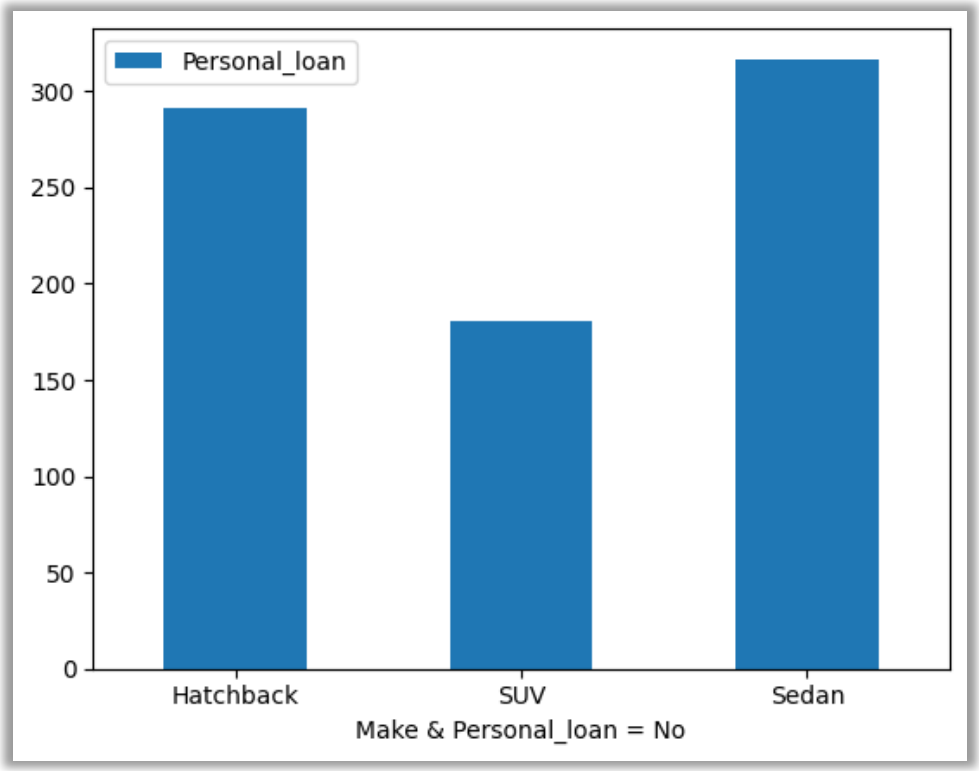




F2) Personal_loan

Based on the graphs below, we may infer that customers tend to choose **Sedans over SUVs or Hatchbacks**, whether or not they have personal loans. Customers who do **not have personal loans**, however, have a **stronger predilection for SUVs**. The choice to buy a **more expensive car** may be **deterred by current obligations** resulting **from personal loan** debt.





Question. G: From the current data set comment if having a working partner leads to the purchase of a higher-priced car.

Answer:

I started by figuring out the pricing ranges for the various car categories, which can indicate the more expensive car. The maximum and minimum numbers in the table below illustrate that:

- **SUVs** are more expensive than **Sedans** and **Hatchbacks**.
- **Hatchbacks** are the least expensive cars.

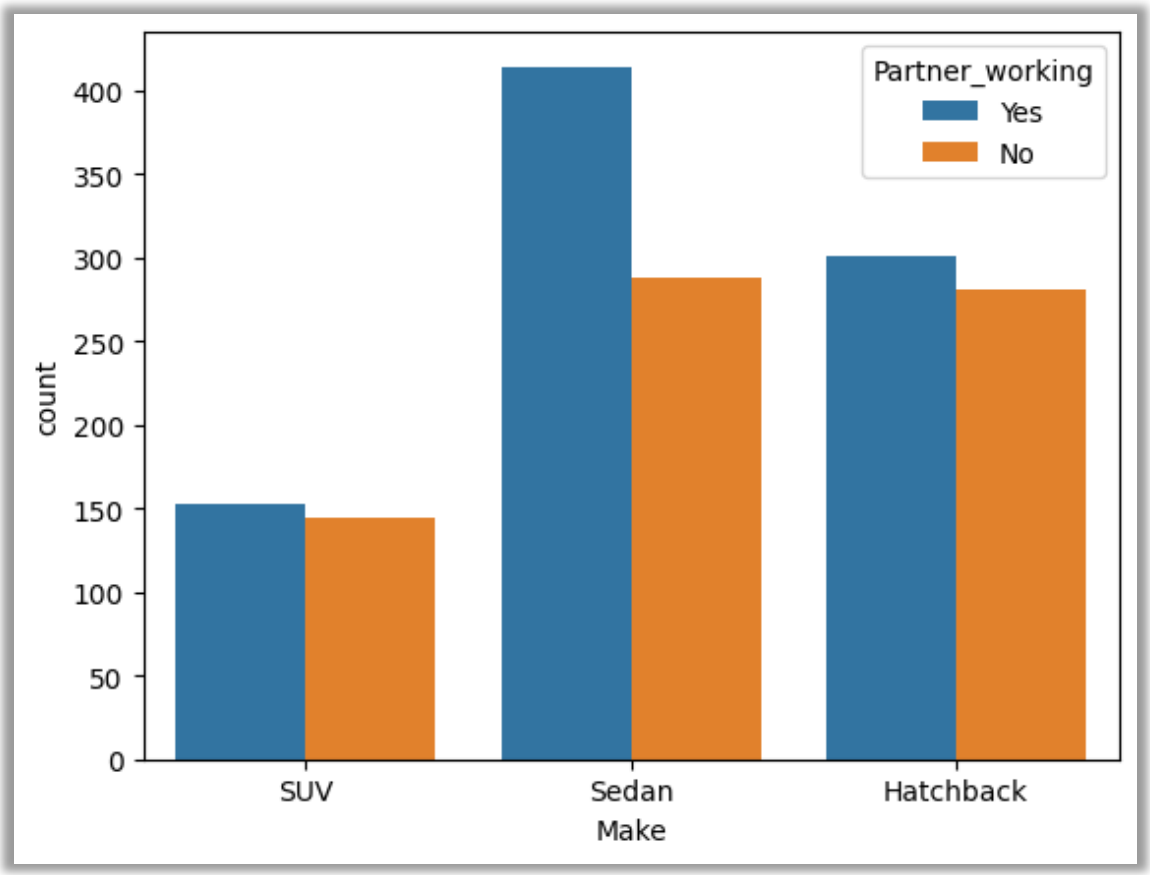
Min. Price

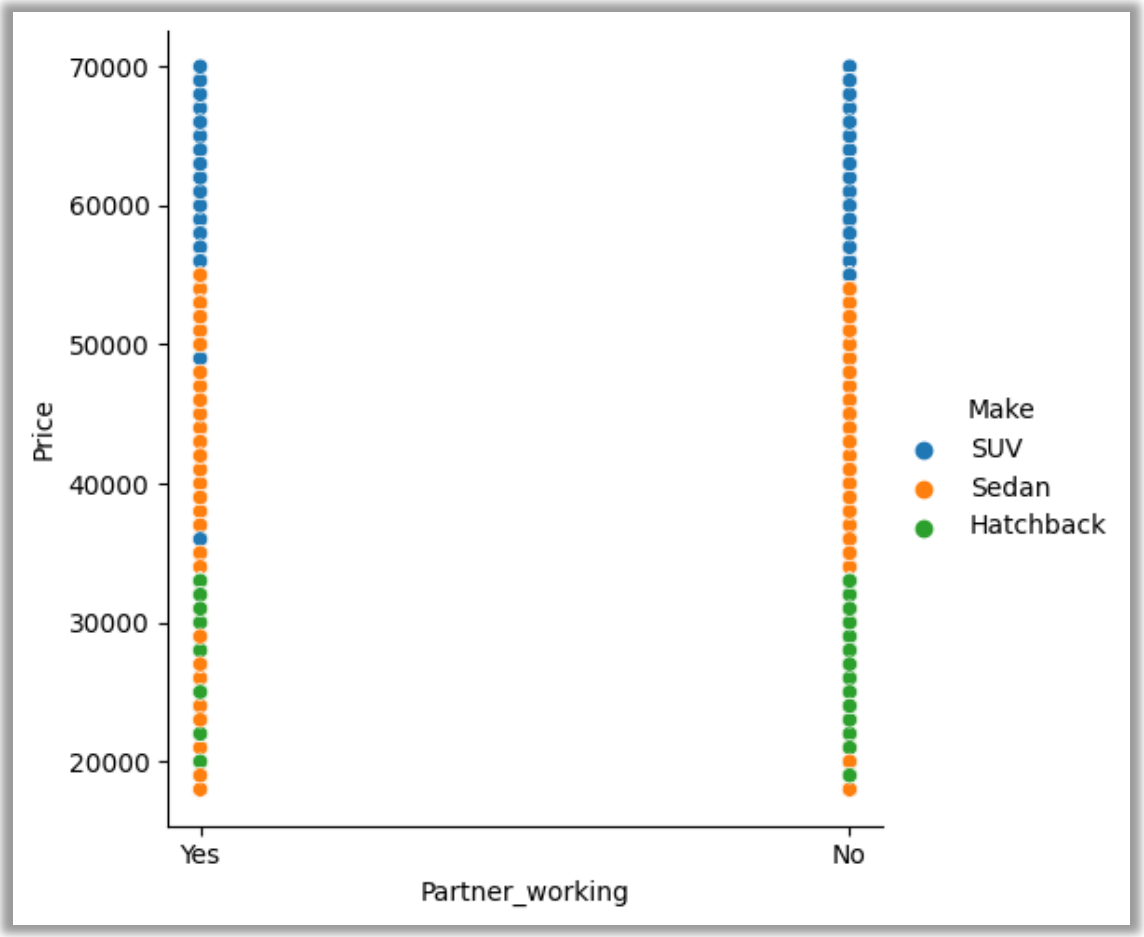
```
Make
Hatchback    18000
SUV          31000
Sedan        18000
Name: Price, dtype: int64
```

Max. Price

```
Make
Hatchback    33000
SUV          70000
Sedan        55000
Name: Price, dtype: int64
```

- The graphs below indicate that having **a working partner doesn't necessarily result** in buying a more **expensive** car.
- A **marginal difference** is observed **between working and not working** in the purchase of **SUVs** and **Hatchbacks**.
- The number of customers of **Sedans** is pretty much higher whose partner is working.





Question. H: The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use the Gender and Marital_status - fields to arrive at groups with similar purchase history.

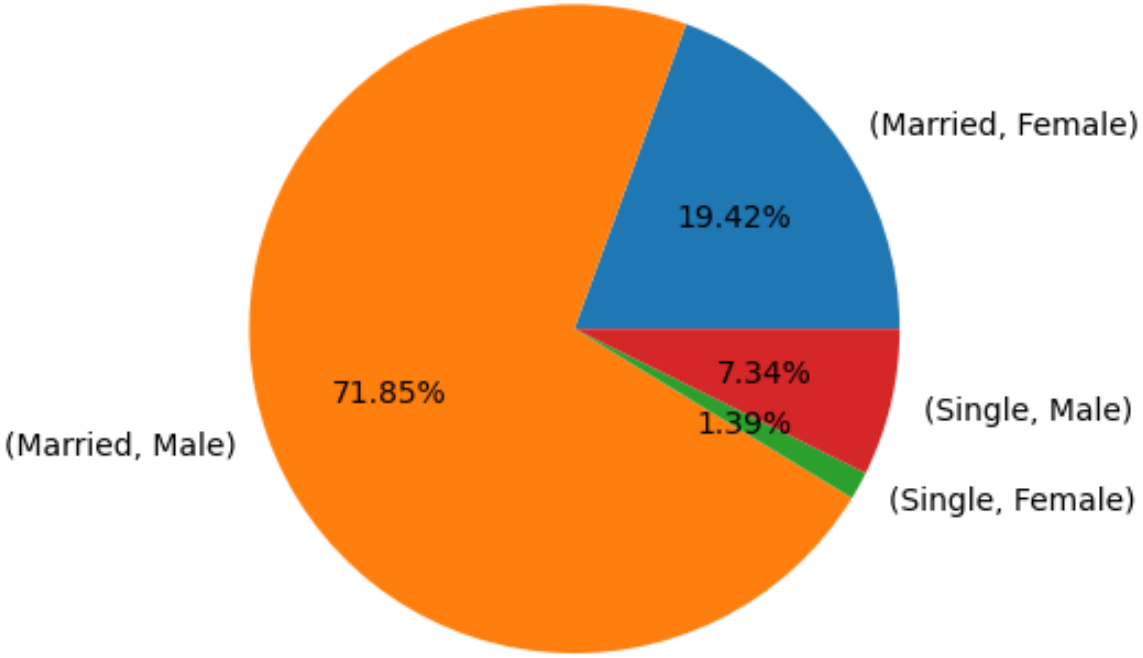
Answer:

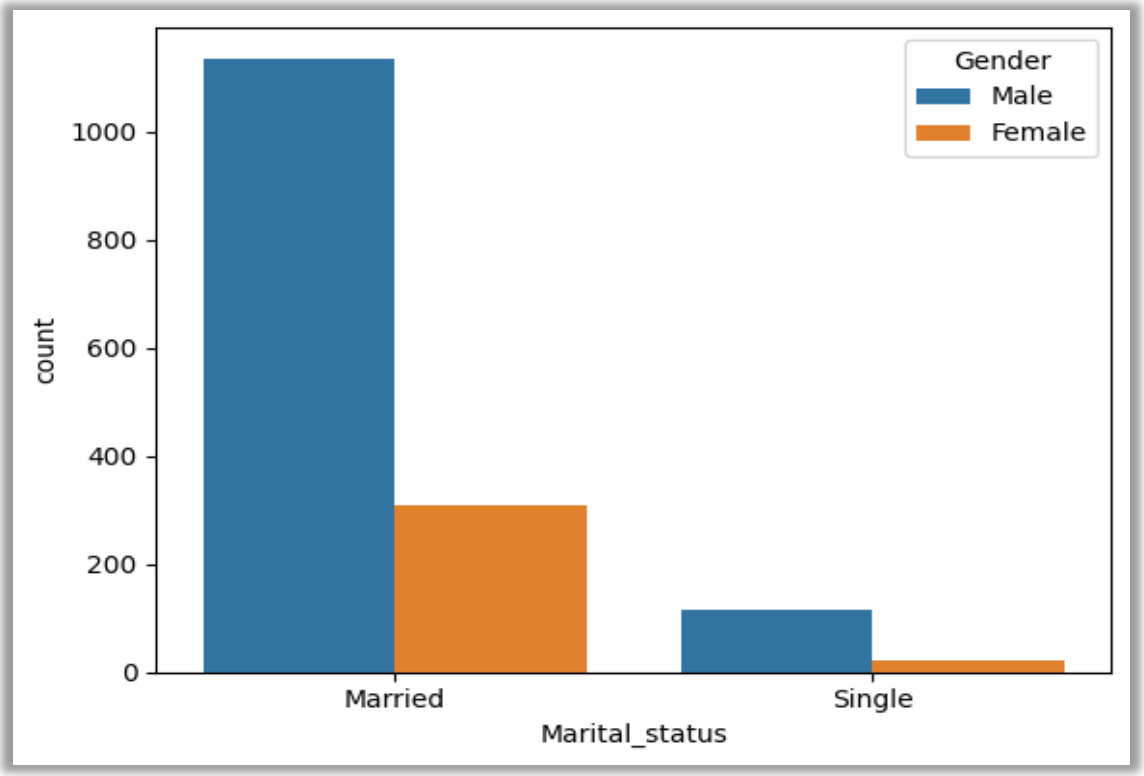
From the tables and graphs in this analysis, we can come up with the conclusion that:

- **Married Males** are the major contributor to sales with a percentage of **71.85 %** as compared to **Single Males** who stand with only **7.34 %**.
- Only **19 %** of **Married Females** are interested in purchasing cars. Whereas **Single Females** have recorded the least interest in purchasing cars with an overall percentage of **1.39 %**.

Marital_status	Gender	Percentage
Married	Female	19.418090
	Male	71.853257
Single	Female	1.391524
	Male	7.337128

Name: Gender, dtype: float64

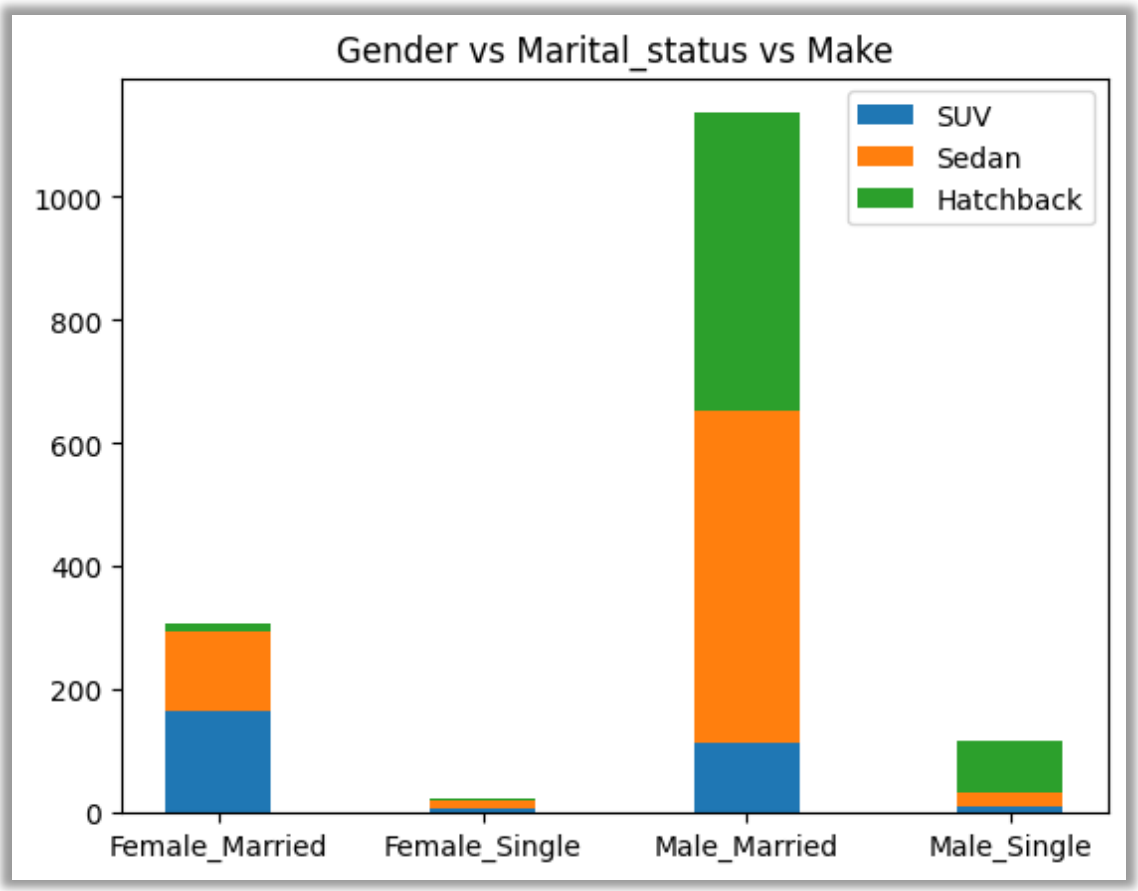




- **Married Males** have a keen interest in buying **Sedans** with **35.07 %** followed by **Hatchbacks** with **31.61 %**. The last choice for them is **SUVs** with only 7.51 %.
- **Married Females** have shown their interest in buying **SUVs** with **10.84 %** followed by **Sedans** with **8.29 %.** The last choice for them is **Hatchbacks** with only 0.91 %.

Gender	Marital_status	Make	Count	Percentage
Female	Married	Hatchback	14	0.914435
		SUV	166	10.842587
		Sedan	127	8.295232
	Single	Hatchback	1	0.065317
		SUV	7	0.457218
		Sedan	14	0.914435
Male	Married	Hatchback	484	31.613325
		SUV	115	7.511430
		Sedan	537	35.075114
	Single	Hatchback	83	5.421293
		SUV	9	0.587851
		Sedan	24	1.567603

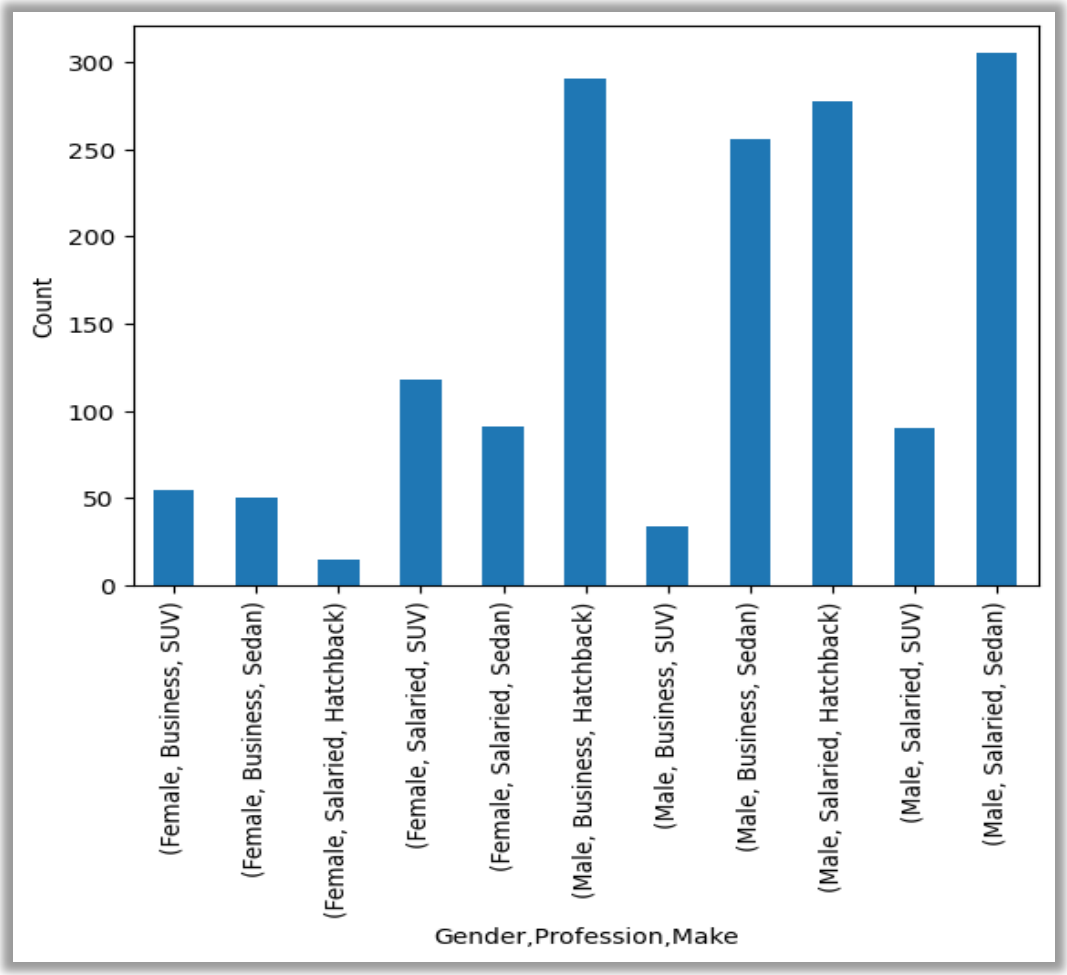
Name: Make, dtype: int64

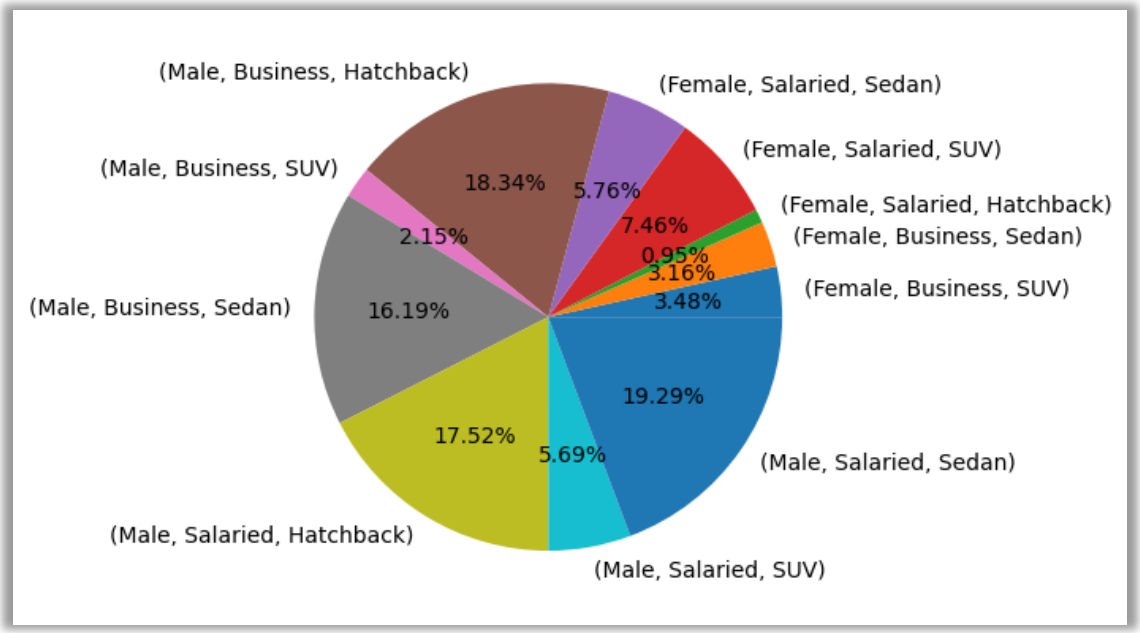


- **Females** who are **Business** professionals have marked **zero sales** for the **Hatchback** cars. At the same time, their interest in the **SUVs** and **Sedans** can be seen.
- **Females** who are **Salaried** professionals have purchased **SUVs** more than **Sedan** and **Hatchback.**

Gender	Profession	Make	Count	Percentage
Female	Business	SUV	55	3.478811
		Sedan	50	3.162555
	Salaried	Hatchback	15	0.948767
		SUV	118	7.463631
		Sedan	91	5.755851
Male	Business	Hatchback	290	18.342821
		SUV	34	2.150538
		Sedan	256	16.192283
	Salaried	Hatchback	277	17.520557
		SUV	90	5.692600
		Sedan	305	19.291588

Name: Make, dtype: int64





On the basis of my overall analysis, I would like to conclude that only the youth **aged 30 or below** are **purchasing Hatchbacks,** and those **above 45** are **interested in SUVs only**. Also, we can see that the **Male** segment both **Single** and **Married**, finds **SUVs** to be of **very low popularity**. In order to increase their topline for **SUVs** in the Male category, the corporation can further explore and try to determine the causes of the same. In a similar vein, the **Sedan** appears to be the **preferred vehicle** among **Married men**, whereas the **Hatchback** is **preferred by Unmarried men**.

Based on the information and analysis, the **company can customize and provide targeted information** regarding **festive offers, deals**, and **new launches** to the identified segments for their preferred car make, **in order to increase their top line and profitability**.