SMDM Project

Submitted by : Taniya Dubey

INDEX

List of figures

Problem 1: Austro Motor Company Problem

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.

A. What is the important technical information about the dataset that a database administrator would be interested in?

Dataset has 1581 rows and 14 columns.

Table 1: Top five rows of the dataset

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

Table 2: Basic Information of the dataset

```
<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: 173.0+ KB
```

The info table of dataset tells the following details:

- Numerical variables: 6
- Categorical variables: 8
- Null values : Partner_salary & Gender
- Duplicate: 0

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?

- Null Values:
  i. Gender: 53 nulls
  ii. Partner_salary: 106 nulls
- Handling of null values:
  i. Gender: the mode of Gender which is Male is assigned to null values because Gender is a categorical variable so mode will be used to impute value and as the number of null values are very less so column drop will not be a good option.
  ii. Partner_salary: As the variables are inter related so treatement will be done accordingly.

If Partner_working = 'No' then Partner_salary = 0

If Partner_working = 'Yes' then Partner_salary = Total_salary - Salary

- The Summary Statistics of the Dataset (Numerical fields):

Table 3: Numerical summarization of the dataset

	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

Table 4: Skewness of variables

```
Age                0.893087
No_of_Dependents  -0.129808
Salary            -0.011571
Partner_salary     0.338255
Total_salary       0.609706
Price              0.740874
dtype: float64
```

The above data about the dataset tells the following information:

- Customer's age group: 22-54 years old

  This is a group og people belong to working age group. Average age of people are 31.92 years, median age of people are 29 years, indicating age distribution is positively skew. The value of skewness is 0.89.

- Salary Ranges: 30k – 99.3k

  Average salary: 60392.22, Median: 59500. the distribution is symmetric. The mean and the median values are very close and skewness is very close to 0.

- Total salary Range: 30k-171k

  It does not show a high degree of skewness.

- Price of the purchased automobile: minimium=18k, maximum=70k

  Price has a skewness of + 0.74 indicating moderate skewness. This indicates a small number of high-priced purchases were made.

➢ Checking for anomalous values in categorical variables:

Table 5: Value Counts of the Categorical variables

```
Gender
 Male      1252
Female     327
Femal        1
Femle        1
Name: Gender, dtype: int64

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

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

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

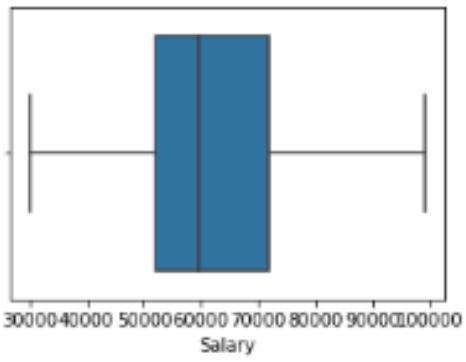
```
House_loan
 No      1054
Yes      527
Name: House_loan, dtype: int64

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

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

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

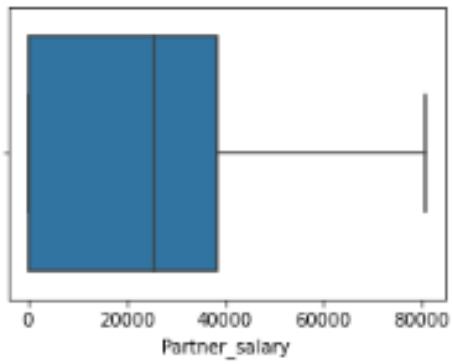
There are currently 4 categories in Gender that is Male, Female, Femal, Femle.

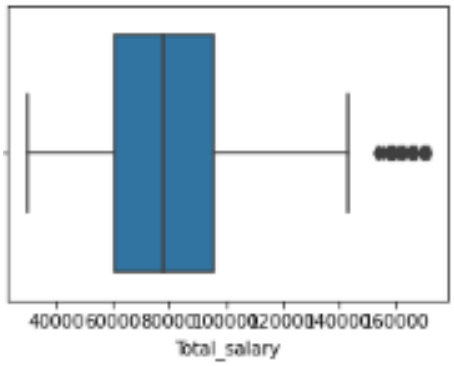
We can see the spelling of female is incorrect because of which two extra columns are made. So, to make the data cleaner and more useable we will assign femle and femal to female. after which we have two categories I.e male and female.

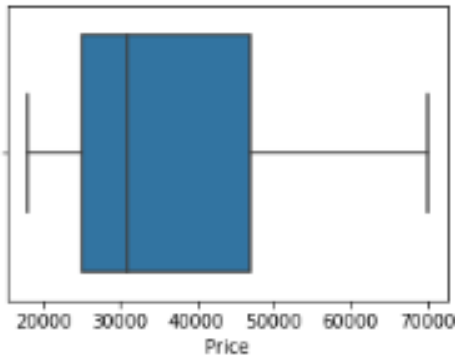
Rest of the categorical fields seem to be free from any such issues.

- Inspecting continuous fields for anomalies/extreme values –









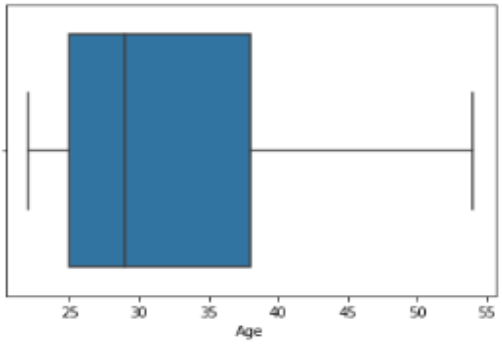


Figure 1: Boxplots of the numerical variables

There are no negative values present in the numerical fields.

We can also see there are outliers present in Total_salary.

Total_Salary- A total of 27 outlier values are present in the variable.

To handle the outliers in Total_salary, we can choose any of the following two approaches -

1) We can treat the outlier values using Winsorization. However, this may lead to loss of valuable information hence should be used with caution.

2) We do not treat the outlier values, and see if analysing them separately can give us some more insights.

For the current study, we will implement both the approach and will analyse how results/ inferences have changed due to the Outlier Treatment.

Creating two datasets to create solution for the two approach –

df- dataset without outlier treatment

data_out_treat – dataset with outlier treatment

Outlier Treatment in dataset data_out_treat -

Outliers were treated by using Winsorization, i.e. bringing the larger outliers (Data points above the Q3 + 1.5 * IQR value) to the upper whisker value and bringing the smaller outliers (Data points below the Q1 - 1.5 * IQR value) to the lower whisker.

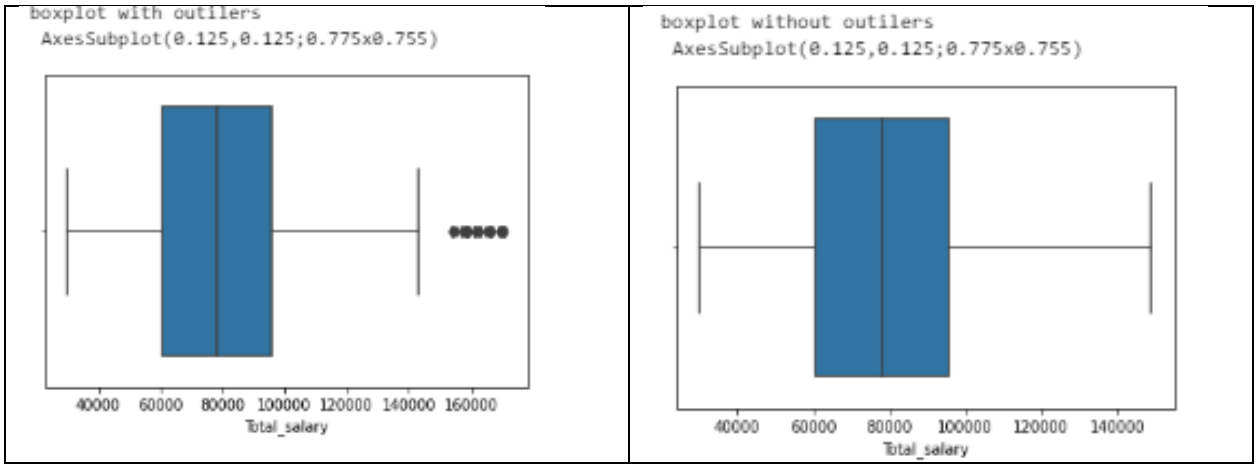


Figure 2: Effect of outlier treatment on Total_salary variable

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

Univariate Analysis of Numerical fields –

For performing Univariate analysis we will take a look at the Boxplots and Histograms to get better understanding of the distributions.

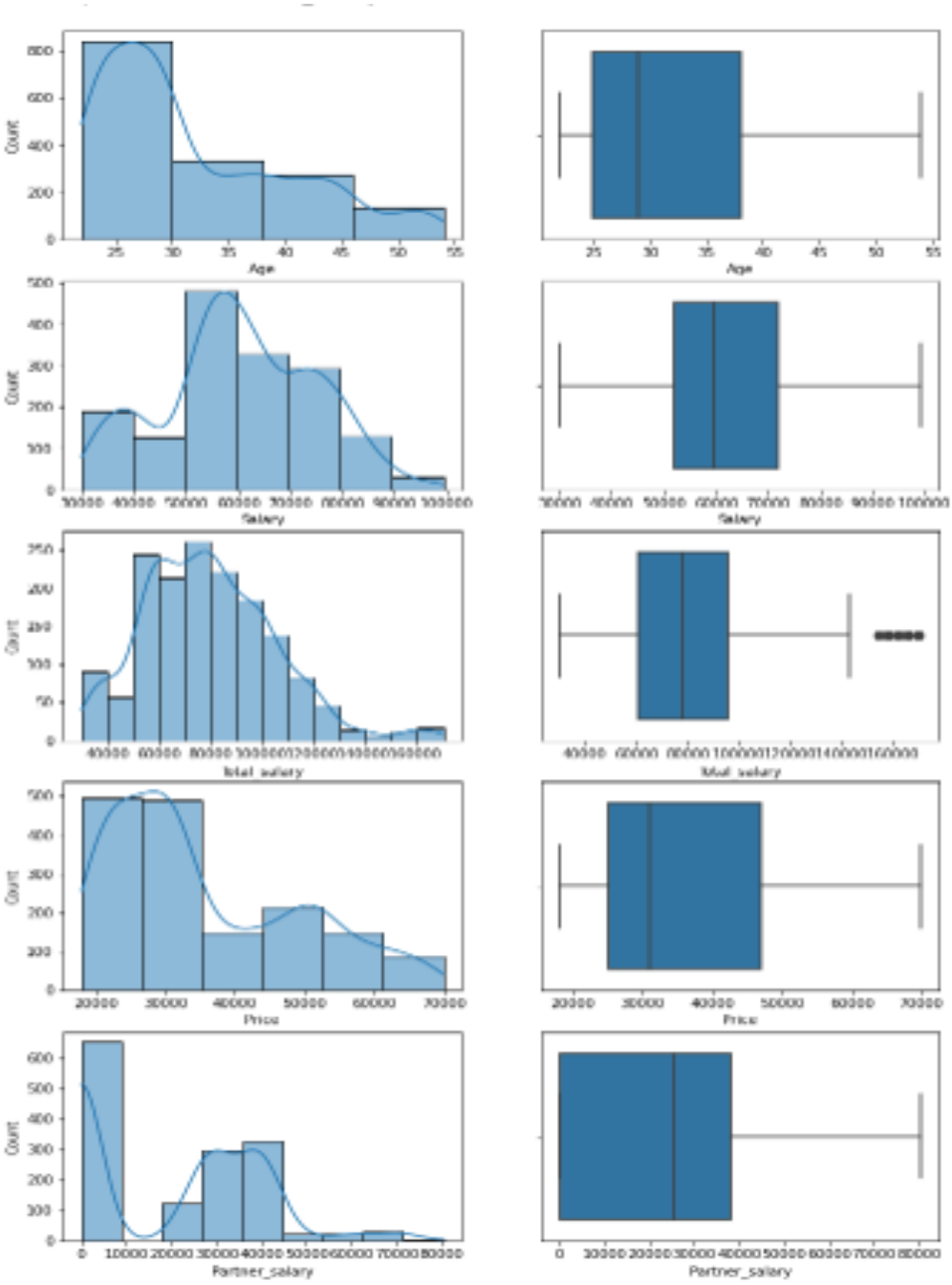


Figure 3: Univariate analysis of numerical variables

```
Total_salary after outlier treatment -

<AxesSubplot:xlabel='Total_salary'>
```

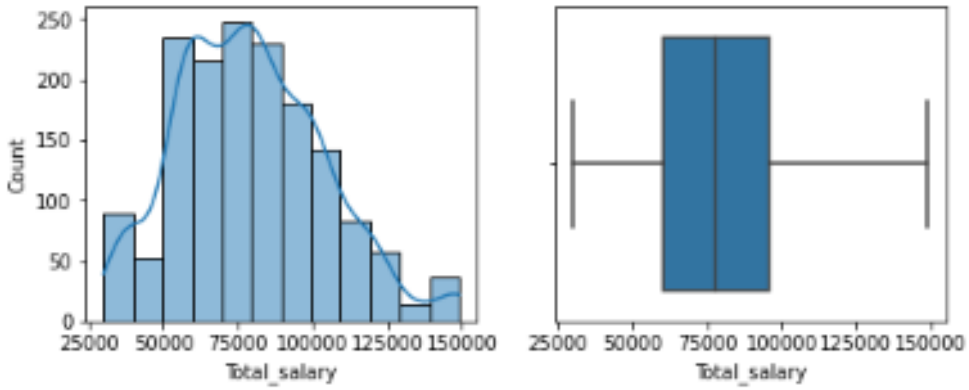
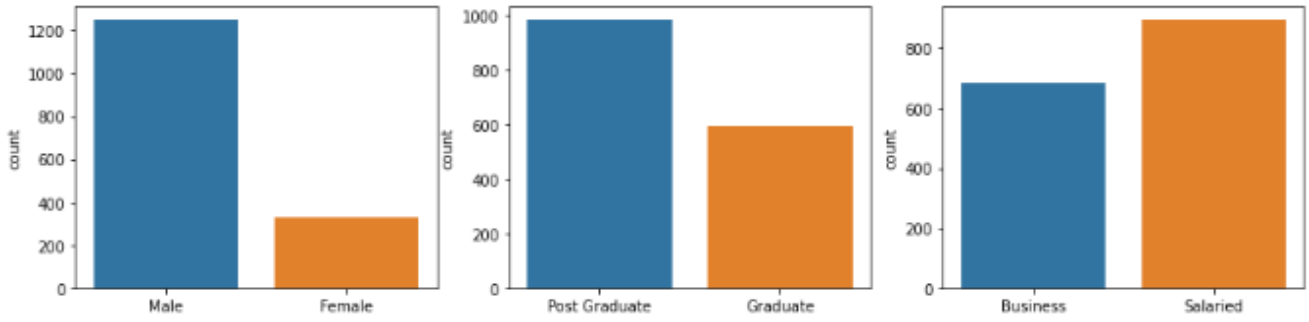


Figure 4: Univariate analysis of Total_salary post outlier treatment

Inferences –

1. Salary has a multimodal distribution, with bulk of data points in the range 50K to 70K.
2. Price seems to have a Bi-modal distribution, and a positive skew of 0.74.
3. Age seems to have a multimodal distribution, and has the highest positive skew of 1.14 among all the fields.
4. Skewness of Total_salary has reduced significant post outlier treatment. The distribution seems to be multimodal, with bulk of data points in the range of 60K to 100K.
5. Almost all the variables have some skewness present, thus none of them follow a Normal distribution. Total_salary can be considered Near-Normal distribution with fair bit of approximation.

Univariate analysis of Categorical variables –



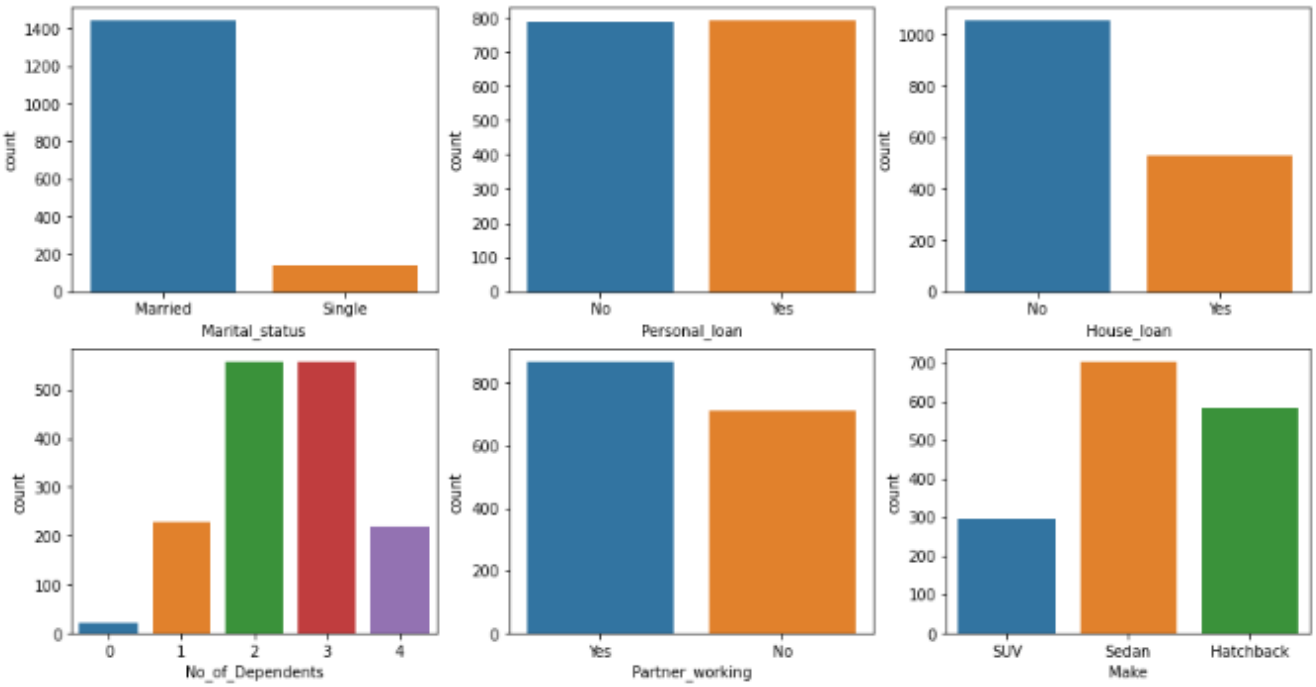


Figure 5: Univariate analysis of categorical fields

Inferences –

- The customers having working partner are more that the customers not having working partner. Total customer 713 with working partner, Total customer with no working partner 138.
- Preferred Purchase a Sedan followed by Hatchback and SUV
- Salaried customers are more that business customers
- Most customers are Post Graduate.
- The majority of the customer have either 2-3 dependents, followed by 1-4. Very few customers have zerono of dependensts
- Number of customer who did not take a house loan is almost double the customer who took a house loan

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.

Bivariate analysis of Numerical variables:

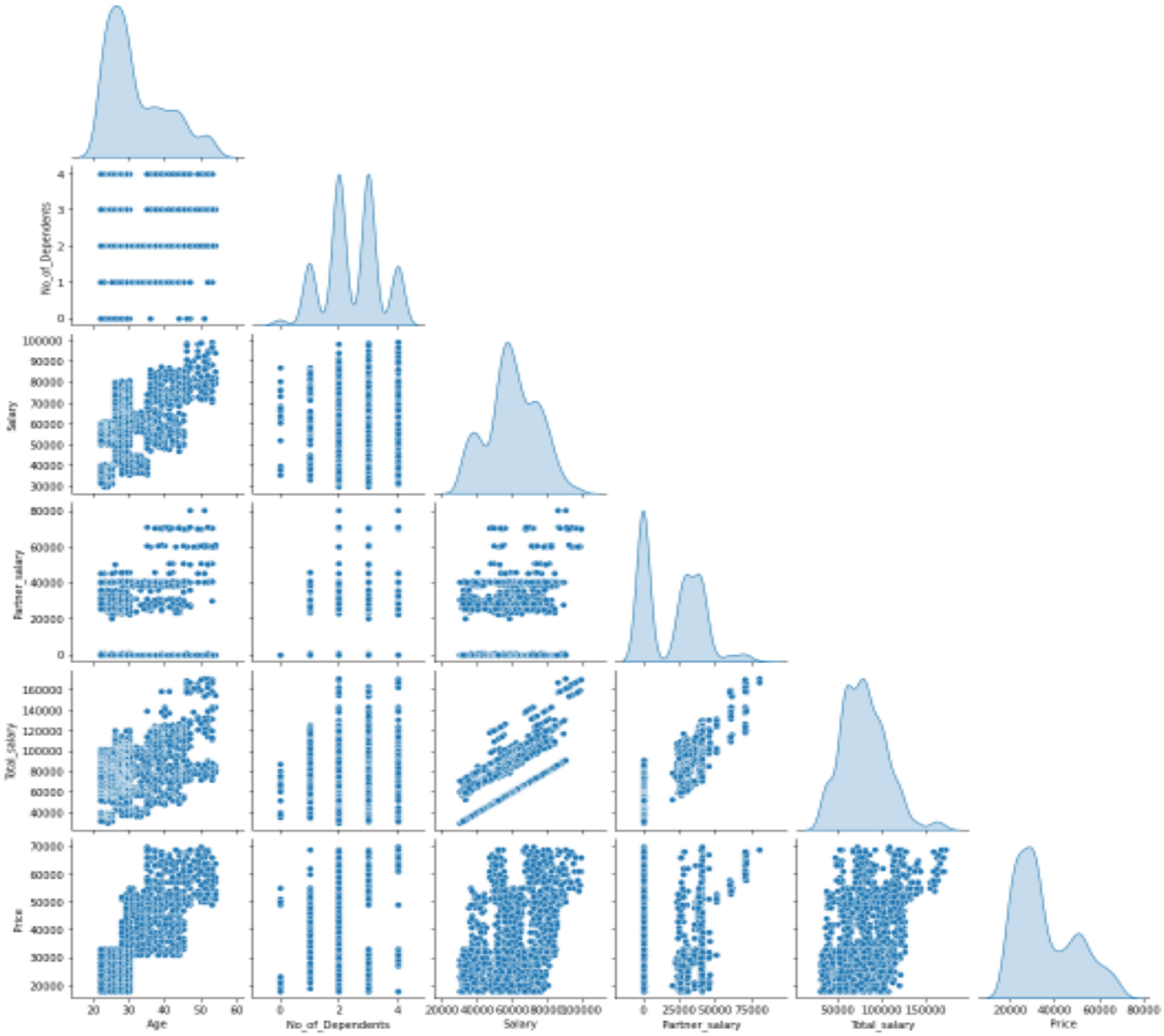


Figure 6: Pair plot on the Data set without treating outliers in Total_salary

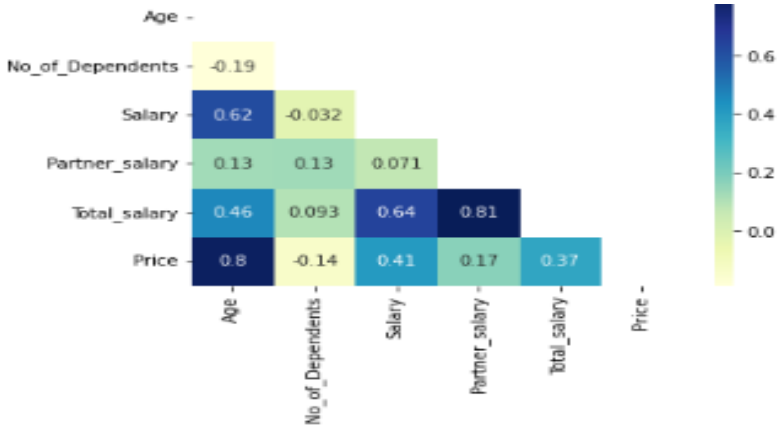


Figure 7: Correlation Heatmap without treating outliers in Total_salary

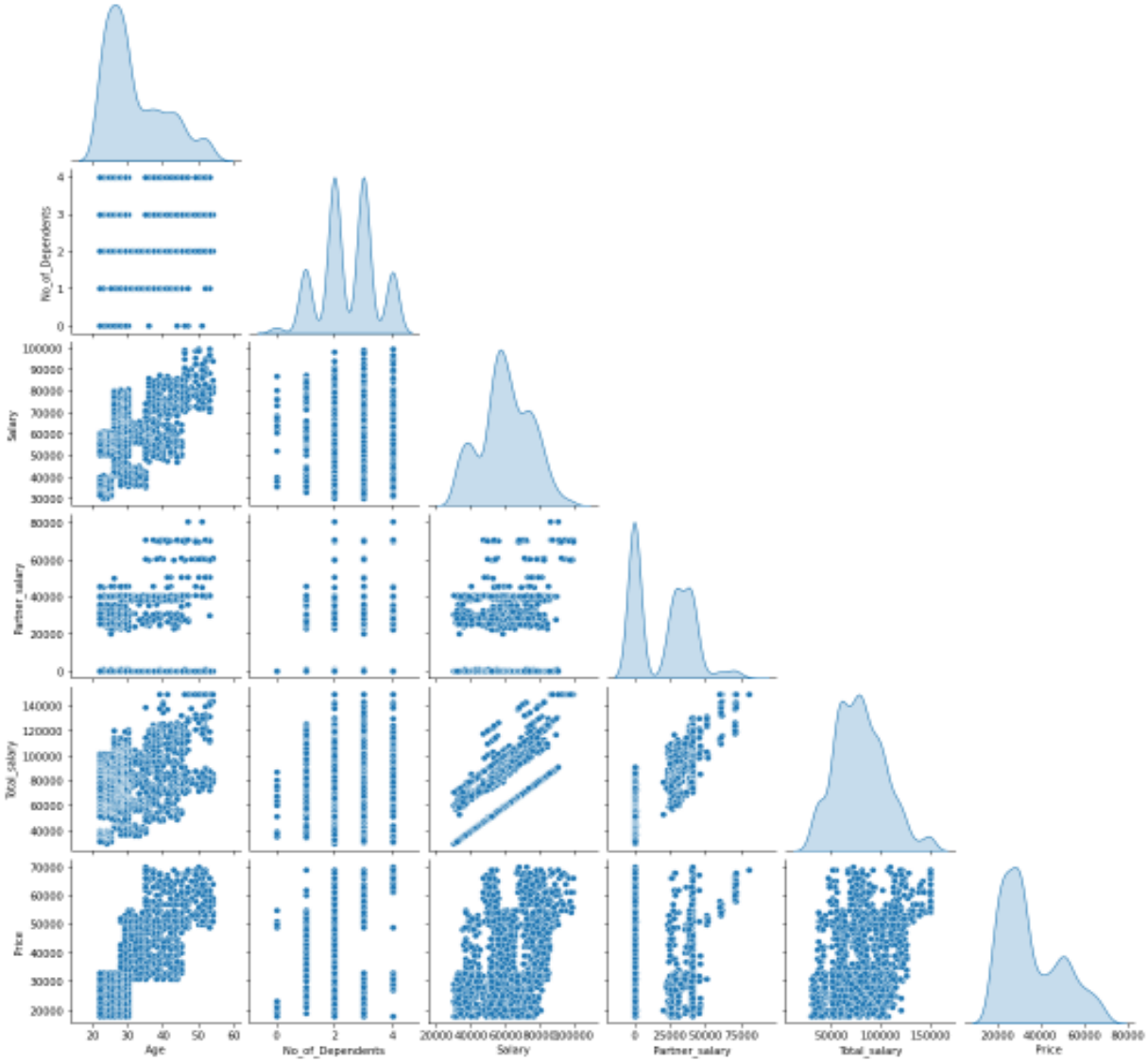


Figure 8: Pair plot on the dataset post outlier treatment in Total_salary

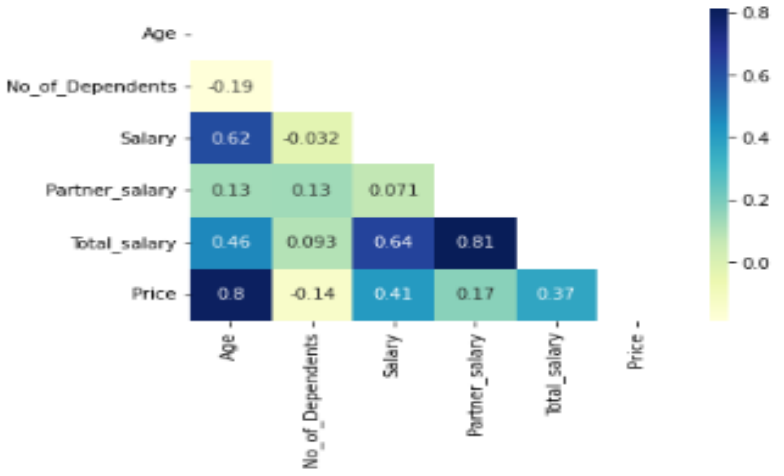


Figure 9: Correlation Heatmap post outlier treatment in Total_salary

Inferences –

1. Rarely any linear relationship present among fields.
2. Highest positive correlation exists between Price and Age (Post Outlier Treatment), and Total_salary and Partner_salary (without outlier treatment)

Bi- Variate analysis of Categorical vs Categorical variables –

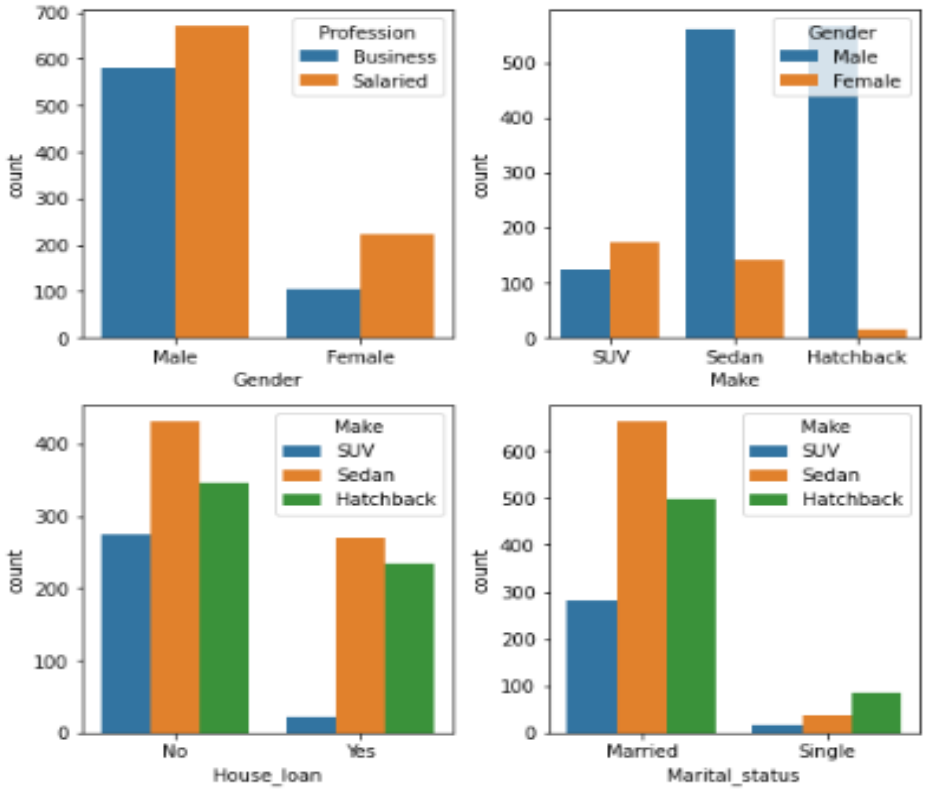


Figure 10(a): Proportion plots for Categorical vs Categorical fields

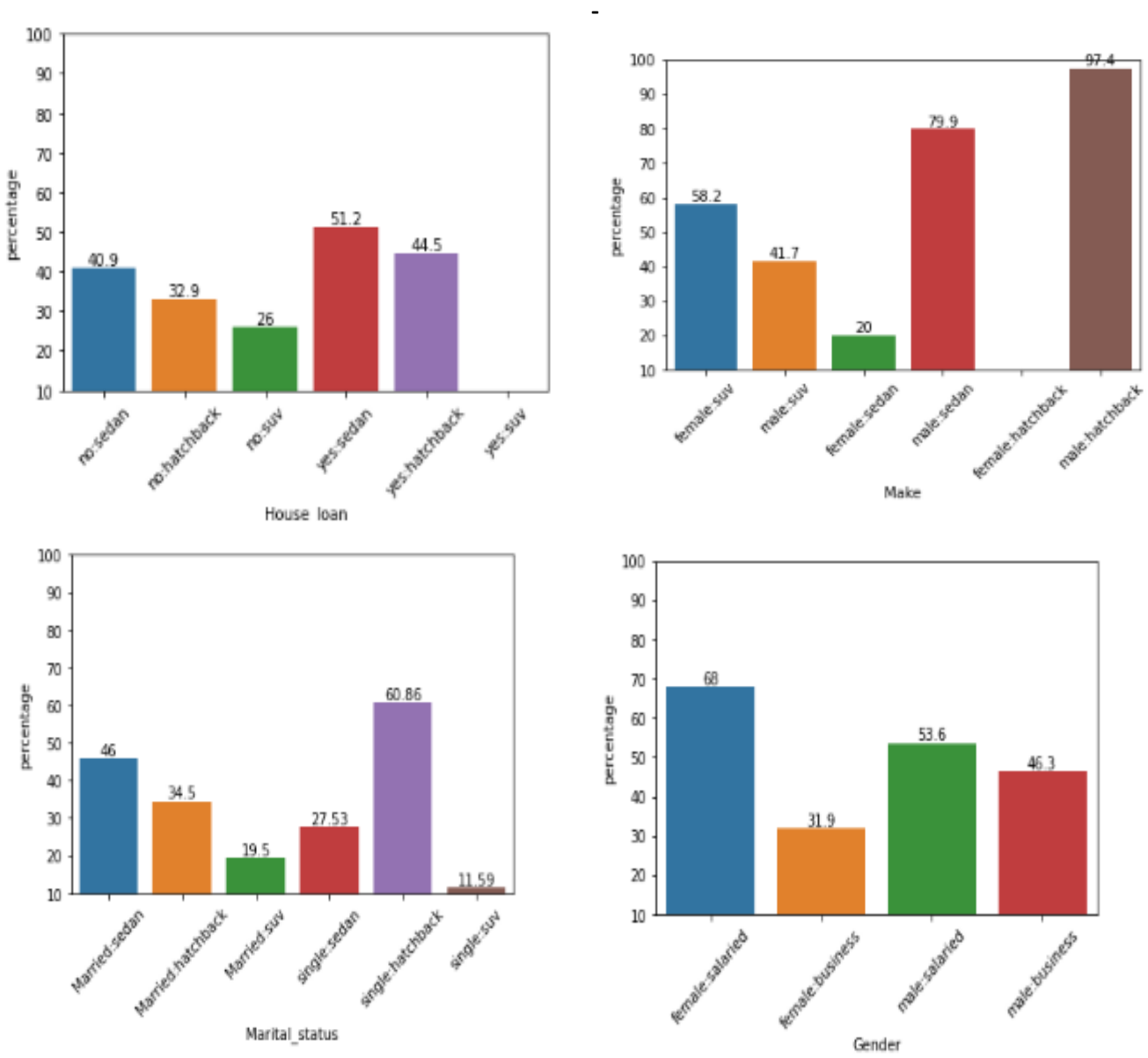


Figure 10(b): Proportion plots for Categorical vs Categorical fields

Inferences –

1. Customers having house loan are not likely to but a SUV. Sedan is the most preferred car among all categories.
2. Female prefer SUV and least likely to but a Hatchback. Male prefers Sedan or Hatchback and least likely to buy SUV.

3. Married customers prefer Sedan whereas single customers prefer Hatchbacks.

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.

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

Analysing the ratio of SUV purchases for both the Genders, we get:

```
Proportion of females buy SUV=  0.5258358662613982  (no. of females bought SUV / Total no. of females)
Proportion of males buy SUV=  0.09904153354632587  (no. of males bought SUV / Total no. of males)
```

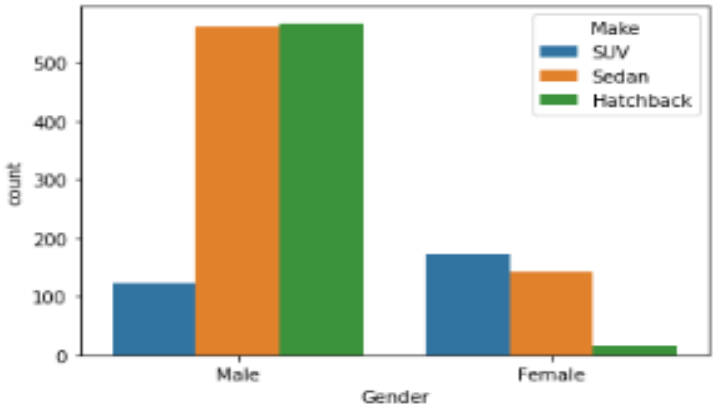


Figure 11: Count Plot of Gender vs Make

As we can see men least prefer SUV. Hence the statement made by Steve Rogers is incorrect.

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

Analysing the Proportion of Car Make purchases for salaried customers, we get:

```
Proportion of Hatchbacks purchased =  0.32589285714285715  (Total Hatchbacks bought by salaried / Total Cars purchased by salaried)
Proportion of SUV purchased =  0.23214285714285715  (Total SUVs bought by salaried / Total Cars purchased by salaried)
Proportion of Sedan purchased =  0.4419642857142857  (Total Sedans bought by salaried / Total Cars purchased by salaried)
```

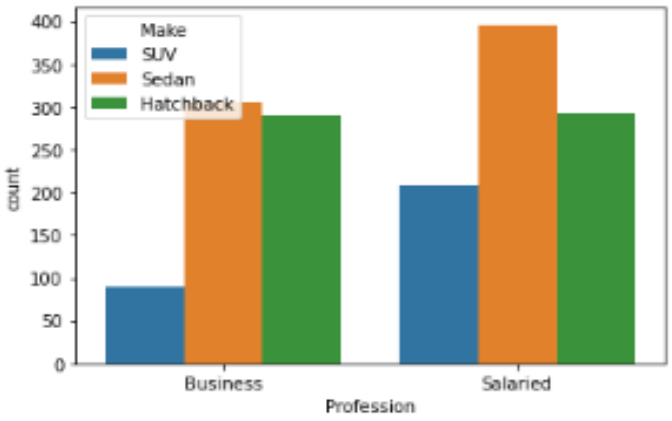


Figure 12: Count Plot of Profession vs Make

From the above results and chart, it is evident that salaried person is more likely to buy a Sedan.

Hence the statement made by Ned Stark is correct.

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.*

Calculating Total number of Cars purchased by Salaried Male Customers for each Make, we get

```
Proportion of Hatchbacks purchased=  0.41220238095238093 ( Hatchbacks bought by salaried male/ Cars purchased by salaried male)
Proportion of SUV purchased =  0.13392857142857142  ( SUVs bought by salaried male / Cars purchased by salaried male)
Proportion of Sedan purchased = 0.4538690476190476 ( Sedans bought by salaried male/ Cars purchased by salaried male)
```

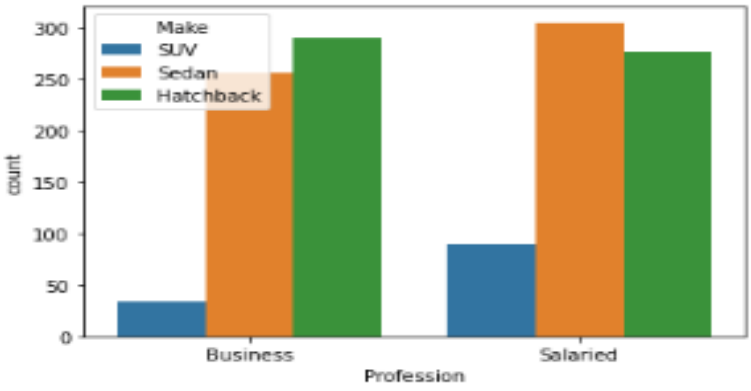


Figure 13: Count Plot of Profession vs Make (For Male customers)

From the above results and chart, it is evident that Salaried male prefers Sedan over SUV.

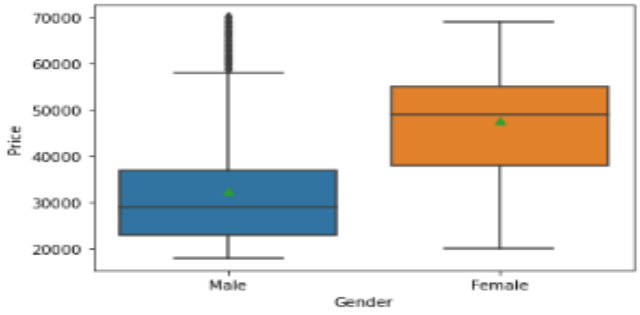
Hence the statement made by Sheldon Cooper is incorrect.

F) From the given data, comment on the amount spent on purchasing automobile 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.

F1) *Gender*

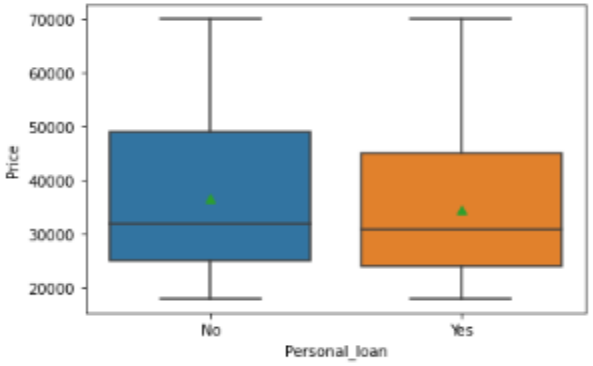
Females are more likely to buy SUVs and on an average spend more on cars than males 47705 Units against 32416 Units. Mean and Median of female are more than Male.

```
Mean Gender
Female    47705.167173
Male      32416.134185
Name: Price, dtype: float64

 Median Gender
Female    49000.0
Male      29000.0
Name: Price, dtype: float64

<AxesSubplot:xlabel='Gender', ylabel='Price'>
```



F2) *Personal loan*

```
Mean Personal_loan
No      36742.712294
Yes     34457.070707
Name: Price, dtype: float64

 Median Personal_loan
No      32000.0
Yes     31000.0
Name: Price, dtype: float64

<AxesSubplot:xlabel='Personal_loan', ylabel='Price'>
```

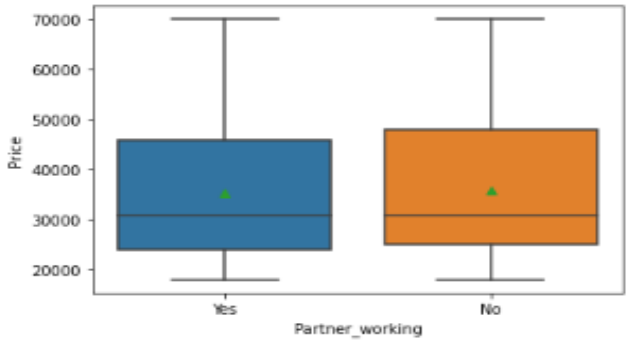


Mean and Median of Price for purchase made by customers without a Personal loan is slightly higher than customers who have a Personal Loan.

To ensure increased spend of customers with Personal loans, the business can look to make the interest rate cheaper (for Automobile purchase) or ease down the repayment terms.

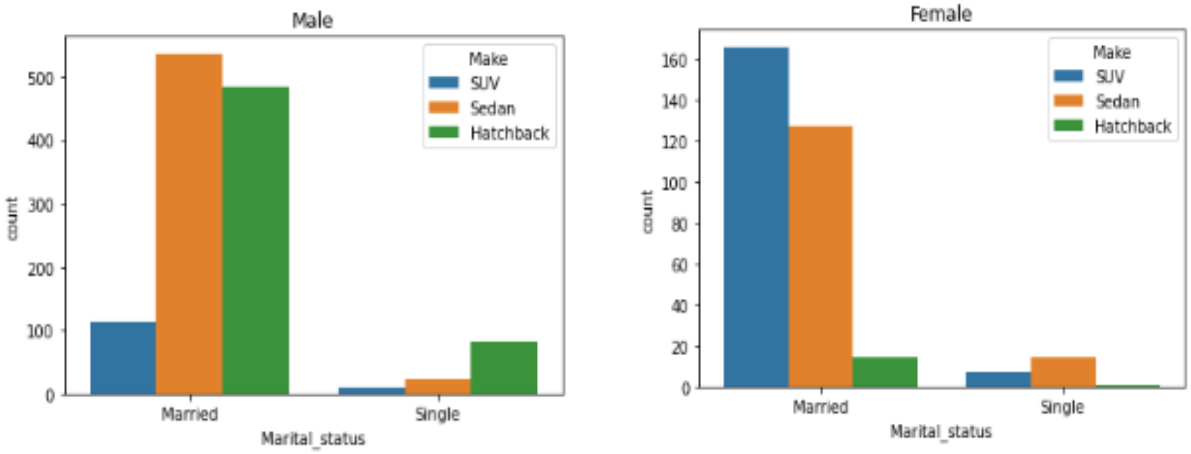
G. From the current data set comment if having a working partner leads to the purchase of a higher-priced car.

```
Partner_working
No     36000.000000
Yes    35267.281106
Name: Price, dtype: float64

 Partner_working
No     31000.0
Yes    31000.0
Name: Price, dtype: float64

<AxesSubplot:xlabel='Partner_working', ylabel='Price'>
```



The Mean and Median price of the purchased automobile is almost similar across the Partner_working category, thus indicating that partner working or not has no effect on the Purchase made by the customer

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.



From the above diagrams and tables , we can make four groups and assign a Car make to each group basis the most frequently occurring value in the past –

A. Married Female – SUV

B. Married Male - Sedan

C. Single Female – Sedan

D. Single Male – Hatchback

Problem II: GODIGT Bank

A bank can generate revenue in a variety of ways, such as charging interest, transaction fees and financial advice. Interest charged on the capital that the bank lends out to customers has historically been the most significant method of revenue generation. The bank earns profits from the difference between the interest rates it pays on deposits and other sources of funds, and the interest rates it charges on the loans it gives out.

GODIGT Bank is a mid-sized private bank that deals in all kinds of banking products, such as savings accounts, current accounts, investment products, etc. among other offerings. The bank also cross-sells asset products to its existing customers through personal loans, auto loans, business loans, etc., and to do so they use various communication methods including cold calling, e-mails, recommendations on the net banking, mobile banking, etc.

GODIGT Bank also has a set of customers who were given credit cards based on risk policy and customer category class but due to huge competition in the credit card market, the bank is observing high attrition in credit card spending. The bank makes money only if customers spend more on credit cards. Given the attrition, the Bank wants to revisit its credit card policy and make sure that the card given to the customer is the right credit card. The bank will make a profit only through the customers that show higher intent towards a recommended credit card. (Higher intent means consumers would want to use the card and hence not be attrite.)

Analyze the dataset and list down the top 5 important variables, along with the business justifications.

Following are the top five variables, along with the business justification:

1. **`avg_spends_13m:`**

   It is Average credit card spends in the last 3 months. Avg_spends_13m talks about the customer spending behavior. It tells about the high or low spending of the customer which helps in knowing the type of customer I.e., high spend indicates primary account whereas lower spend would mean secondary account. Campaigns can be rolled out on the basis of the customer preference; customized offers can be given to lure customers into using the credit account more frequently.

2. **`T+1_month_activity:`**

T+1+month_activity tells whether a customer uses the credit cards in the T+1 month or not. As T+1_month_activity can be used to plan out campaigns and promotional offers so as to increase activity in the credit card.

3. **cc_limit:**

   It is Current credit card limit. A Credit Card limit for customers basis their attributes (such as income, CIBIL Score, etc.) is part of the Risk Management practice wherein the banks try to minimize the number of defaulters. The banks seek a quantifiable answer to the query “How much is too much?”

4. **annual_income_at_source:**

   It is Annual income recorded in credit card application. Annual income is the total value of income earned during a year by a person. It tells about the purchasing power of a person so hence an important piece of information. Income helps banks to make better decision regarding risk profiling, targeted ads, campaigns, offers, loan limits etc.

5. **cc_active30:**

   It is Credit card activity in last 30 days. Cc_active30 talks about how frequently the customer use credit card. This helps in identifying what are the reasons that the usages of credit card have decline or increased for customers.