SMDM PROJECT SAMPLE REPORT

DSBA

tarunbh2394@gmail.com
gbpu0073

Contents

List of Figure

List of Tables:

Problem 1

I. 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.
You as an analyst have been tasked with performing a thorough analysis of the data and coming up with insights to improve the marketing campaign.

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

Load the required libraries to gain access to their respective functions and classes, to perform data manipulation, analysis and visualization tasks, check the versions of preinstalled environment, set the working environment and load the data file and create a data frame.

View the sample of the row using head() and tail() function for top five and last five dataset.

Table 1: Top 5 Rows of 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	33300	27000	Hatchback
1	53	Femal	Salaried	Married	Post Graduate	4	Yes	No	Yes	95500	70300.0	32000	31000	Hatchback
2	53	Female	Salaried	Married	Post Graduate	3	No	No	Yes	97300	60700.0	32900	30000	Hatchback
3	53	Female	Salaried	Married	Graduate	2	Yes	No	Yes	72500	70300.0	32200	24000	Hatchback
4	53	Male	Salaried	Married	Post Graduate	3	No	No	Yes	79700	60200.0	31600	31000	Hatchback

Table 2: last 5 Rows of dataset

	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	Price	Make
1576	22	Male	Salaried	Single	Graduate	2	No	Yes	No	33300	0.0	33300	27000	Hatchback
1577	22	Male	Business	Married	Graduate	4	No	No	No	32000	NaN	32000	31000	Hatchback
1578	22	Male	Business	Single	Graduate	2	No	Yes	No	32900	0.0	32900	30000	Hatchback
1579	22	Male	Business	Married	Graduate	3	Yes	Yes	No	32200	NaN	32200	24000	Hatchback
1580	22	Male	Salaried	Married	Graduate	4	No	No	No	31600	0.0	31600	31000	Hatchback

Important technical information of the dataset are as follows:

Table 3: Basic information of 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 above table about the dataset tells us that there are 1581 Rows and 14 Columns in the given dataset. There are 8 categorical variables (object datatype) and 6 numeric variables (float and int datatype).

Number of duplicate rows =0 using duplicated() method.

Categorical data: 8

Gender, Profession, Marital_status, Education, Personal_loan, House_loan, Partner_working, Make.

Table 4: Categorical data division

Binary	Multilevel
Gender	Make
Personal_loan	Education
Marital_status	Profession
House_loan	
Partner_working	

gbpu0073

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.

There are disparancies in the two of the columns mentioned.
Null values are there in Gender column and Partner_salary column.
Null in Gender = 53
Null in Partner_salary =106

After identifying the null
Access the extent of missing data for which calculate the count of percentage of null in each missing column which will help to prioritize which column to focus on for handling missing values.
As per the nature of our dataset can follow various strategies to handle the null:
If the missing values are limited and do not significantly impact the dataset, you can choose to delete the rows or columns containing null values using methods like .dropna().
If the missing values are significant and deleting them would result in loss of valuable information we can fill or impute the missing values,

1. In case of continuous variable, common computation technique include mean, median or mode or other methods like regression analysis.
2. In case of categorical data we fill the missing values with the majority class of that column.

Here in Gender column, 'Male' was the majority class so replacing null of Gender with Male.

To impute the null values in Partner_salary column we can use the method in part 1 but as here the three variables are related so we can compute the Partner_salary as below.

Total_salary = Partner_salary + Salary

Case 1: Partner working, Null will be replaced by Total_salary- Salary
Case2: Partner not working, Null will be replaced by 0.

Table 5: 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            1581 non-null   object
 2   Profession        1581 non-null   object
 3   Marital_status    1581 non-null   object
 4   Education         1581 non-null   object
 5   No_of_Dependents  1581 non-null   int64
 6   Personal_loan     1581 non-null   object
 7   House_loan        1581 non-null   object
 8   Partner_working   1581 non-null   object
 9   Salary            1581 non-null   int64
 10  Partner_salary    1581 non-null   int64
 11  Total_salary      1581 non-null   int64
 12  Price             1581 non-null   int64
 13  Make              1581 non-null   object
dtypes: int64(6), object(8)
memory usage: 173.0+ KB
```

Table 6: Statistical summary of dataset

	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
count	1581.000000	1581.000000	1581.000000	1581.000000	1581.000000	1581.000000
mean	31.922201	2.457938	60392.220114	19233.776091	79625.996205	35597.722960
std	8.425978	0.943483	14674.825044	19670.391171	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	25100.000000	78000.000000	31000.000000
75%	38.000000	3.000000	71800.000000	38100.000000	95900.000000	47000.000000
max	54.000000	4.000000	99300.000000	80500.000000	171000.000000	70000.000000

Table 7: Skewness of Numerical data

Age	0.896087
No_of_Dependents	-0.129808
Salary	-0.011571
Partner_salary	0.441069
Total_salary	0.609706
Price	0.740874
dtype: float64	

Conclusion:

1. The minimum and maximum age in the dataset is 22 and 54 years respectively. The mean age is 31.92 and the median age is 29 which depicts Age column is positively skewed with the skewness of 0.8.
2. The salary of the customer and the total salary have mean and median so close to each other with the skewness of -0.011 and 0.61 respectively.
3. Minimum and maximum price are 18000 and 70000 respectively. Price is positively skewed with mean greater than the median with the skewness of 0.74. Positive skewness suggests that there are a few sales of high-end or luxury cars that contribute to the longer tail on the right side of the sales distribution.

Checking for outliers, anomalies/ extreme values for categorical dataset:

Table 8: Value counts for categorical variables

Value counts for Gender:	Male 1199 Female 327 Femal 1 Name: Gender, dtype: int64
Value counts for Profession:	Salaried 896 Business 685 Name: Profession, dtype: int64
Value counts for Marital_status:	Married 1443 Single 138 Name: Marital_status, dtype: int64
Value counts for Education:	Post Graduate 985 Graduate 596 Name: Education, dtype: int64
Value counts for Personal_loan:	Yes 792 No 789 Name: Personal_loan, dtype: int64

Value counts for House_loan:	No 1054 Yes 527 Name: House_loan, dtype: int64
Value counts for Partner_working:	Yes 868 No 713 Name: Partner_working, dtype: int64

In Gender column, Female is spelled incorrectly for the two records

Table 9: Value count for Gender

Male	1199
Female	327
Femal	1
Femle	1

After imputing by using .replace method

Table 10: Value count for Gender

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

Checking for outliers, anomalies/ extreme values for categorical dataset:

Figure 1: Boxplot for categorical dataset

Here we can see the outliers for two columns Total_salary and No_of_Dependents. But outliers in No_of_Dependents can be neglected as it analyzing them separately can give us some more insights and it analyses the expenses of salaried person with no dependent.
There are no negative values here and outliers are few in numbers.

Number of outliers: 27
Outliers:
[170000 165800 158000 165700 162900 159000 169000 165600 161100 166900
155200 170400 171000 154100 164700 161800 153500 169300 159100 162300
161100 166500 156900 158900 157700 157900 158200]

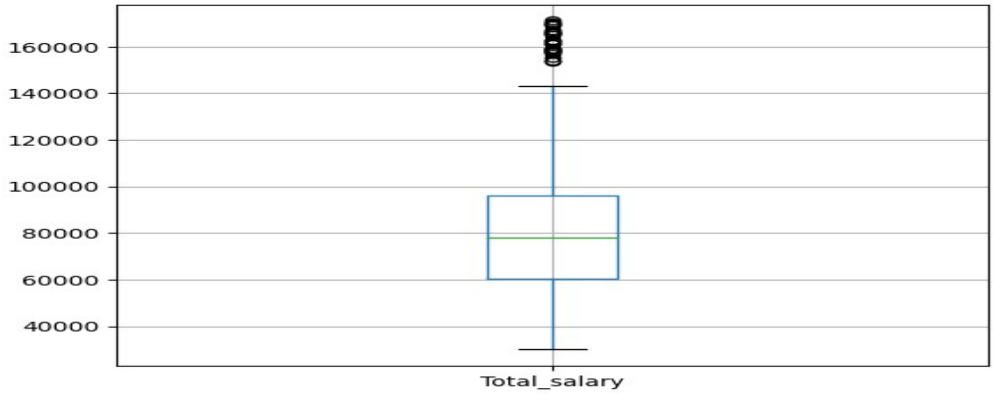

Figure 2: Boxplot of Total_salary with outliers

Treating outliers:
There are various methods of treatment of outliers

1. Data transformation: Using logarithmic, square root or other transformations.
2. Capping and flooring: Replacing outliers with predefined maximum or minimum value.
3. Winsorization: Variation of capping and flooring ie., replacing outliers with the nearest non outlier value.
4. Statistical tests: Using various test such as Z- score, modified Z- score or IQR to detect them and then investigating them further.
5. Trimming: Removing outliers entirely.
6. Modelling techniques: Less sensitive to outliers.

Here in our analysis we used winsorization technique.
Lower range: 7400.0 and Upper range: 149000.0

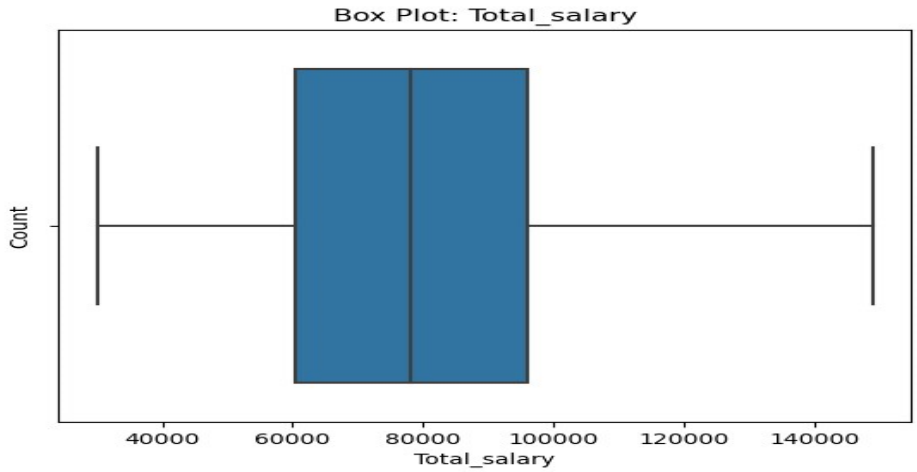

Figure 3: Box plot of Total_salary after winsorization technique

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: As we need to explore the data separately here so let's do the univariate analysis of categorical and numerical variables respectively.

Univariate analysis for categorical fields:

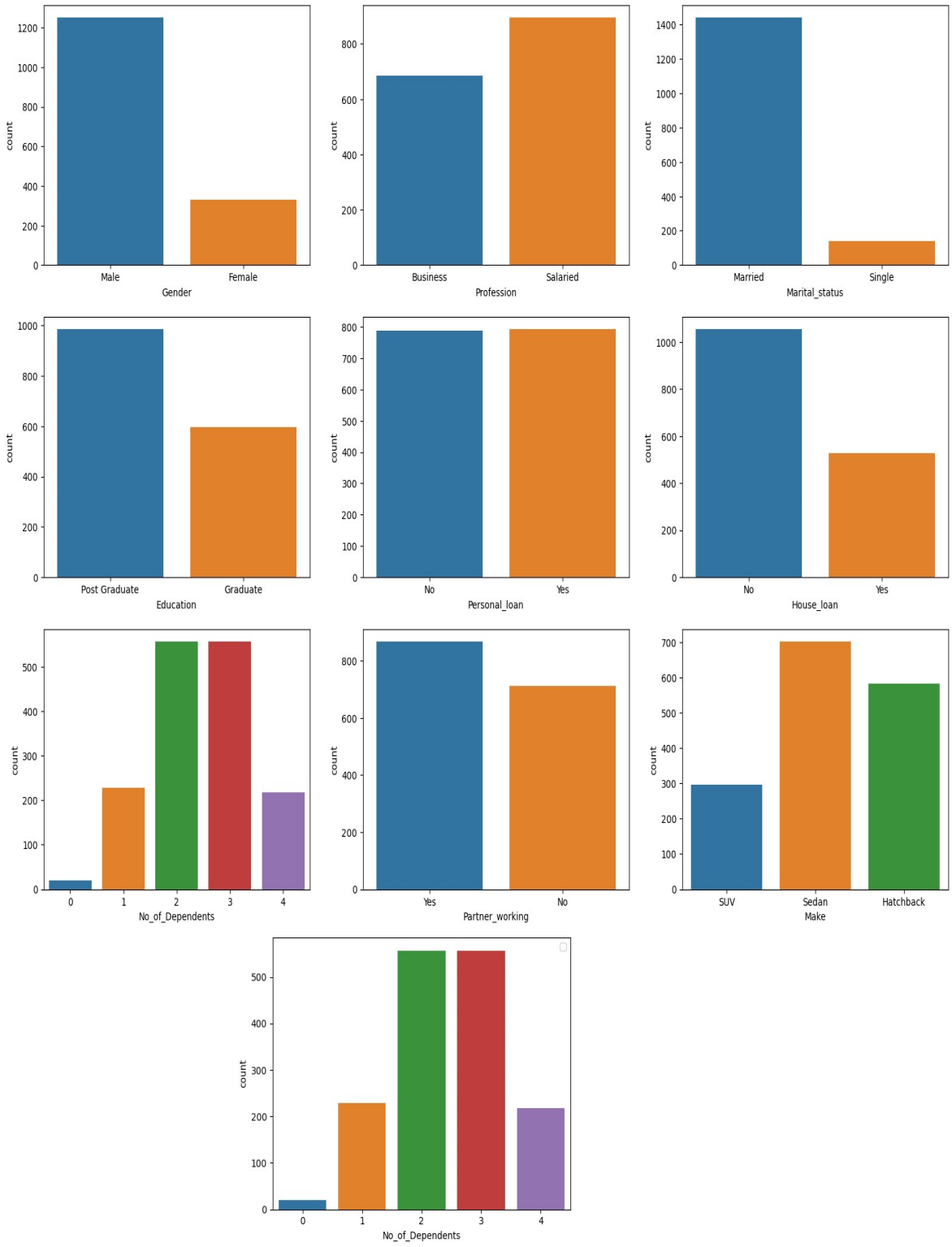

Figure 4: Univariate analysis of categorical variables

Table 11: Value count for Categorical variable

Gender: Male 1252 Female 329 Name: Gender, dtype: int64	**Education:** Post Graduate 985 Graduate 596 Name: Education, dtype: int64
Personal Loan: Yes 792 No 789 Name: Personal_loan, dtype: int64	**Profession:** Salaried 896 Business 685 Name: Profession, dtype: int64
Partner Working: Yes 868 No 713 Name: Partner_working, dtype: int64	**Marital Status:** Married 1443 Single 138 Name: Marital_status, dtype: int64
House Loan: No 1054 Yes 527 Name: House_loan, dtype: int64	**Make:** Sedan 702 Hatchback 582 SUV 297 Name: Make, dtype: int64

Conclusion:

1. Count of salaried profession is more than that of Business profession.
2. Number of customers with their partner working are more than the number of customers with their partner not working. Value count of Partner_working with a Yes is 863 and with a No is 713.
3. Number of people with house loan are almost half of the number of people without house loan.
4. Customers with Marrital_status as married have a count of 1143 which is in majority when compared to that of not married with a count of 138.
5. Majority of the customer are post graduate with the count 985 and customers being Graduate are 596.
6. Most preferred car body style is Sedan followed by Hatchback which in turn is followed by SUV.
7. From the plot it can be concluded that customer with 2 and 3 dependents are in majority followed by 1 and 4 dependents and customer with zero dependent are very few in number.

Univariate analysis of Numerical variable:

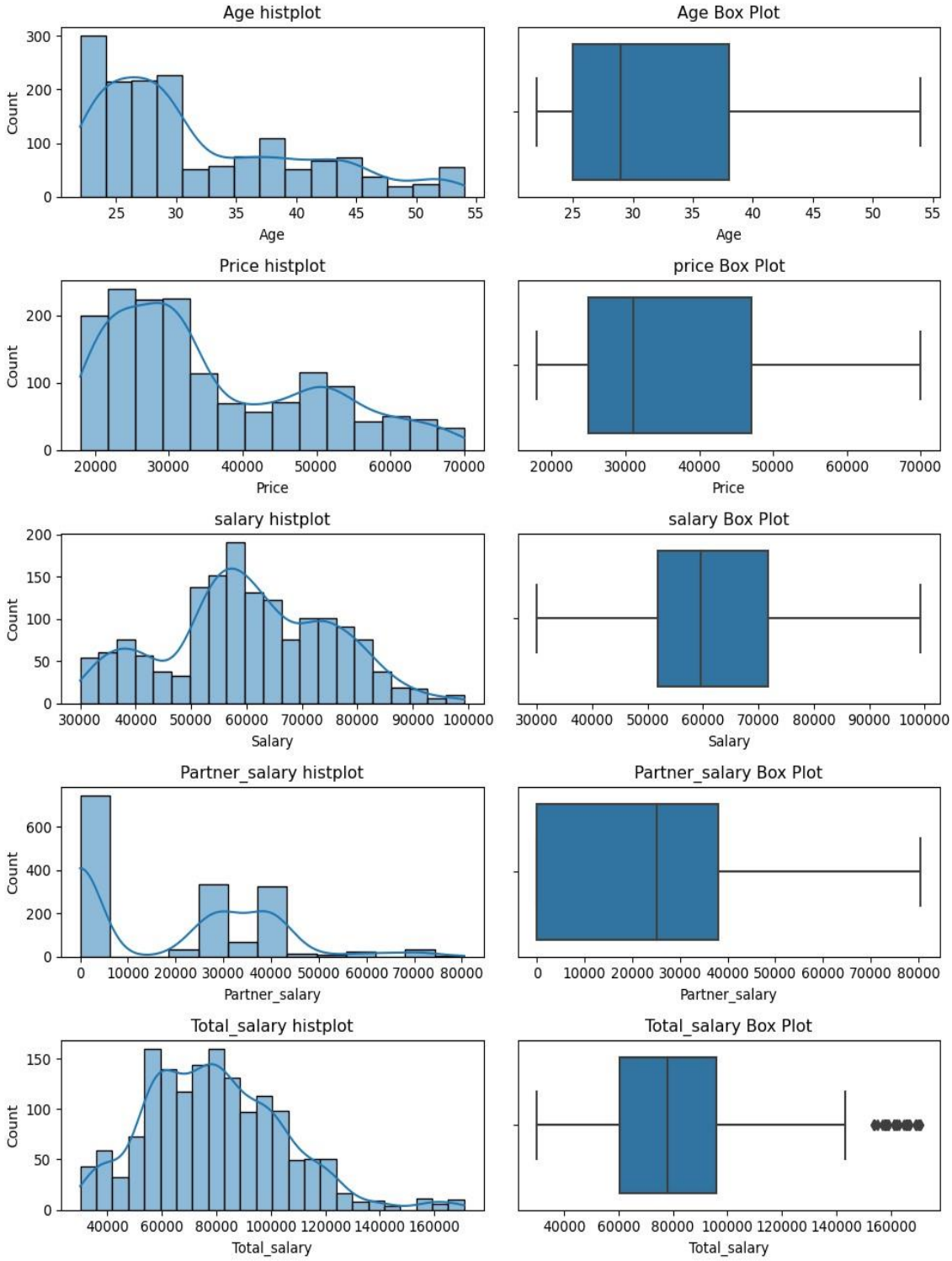

Figure 5: Histplot and boxplot of numerical variables before outlier treatment

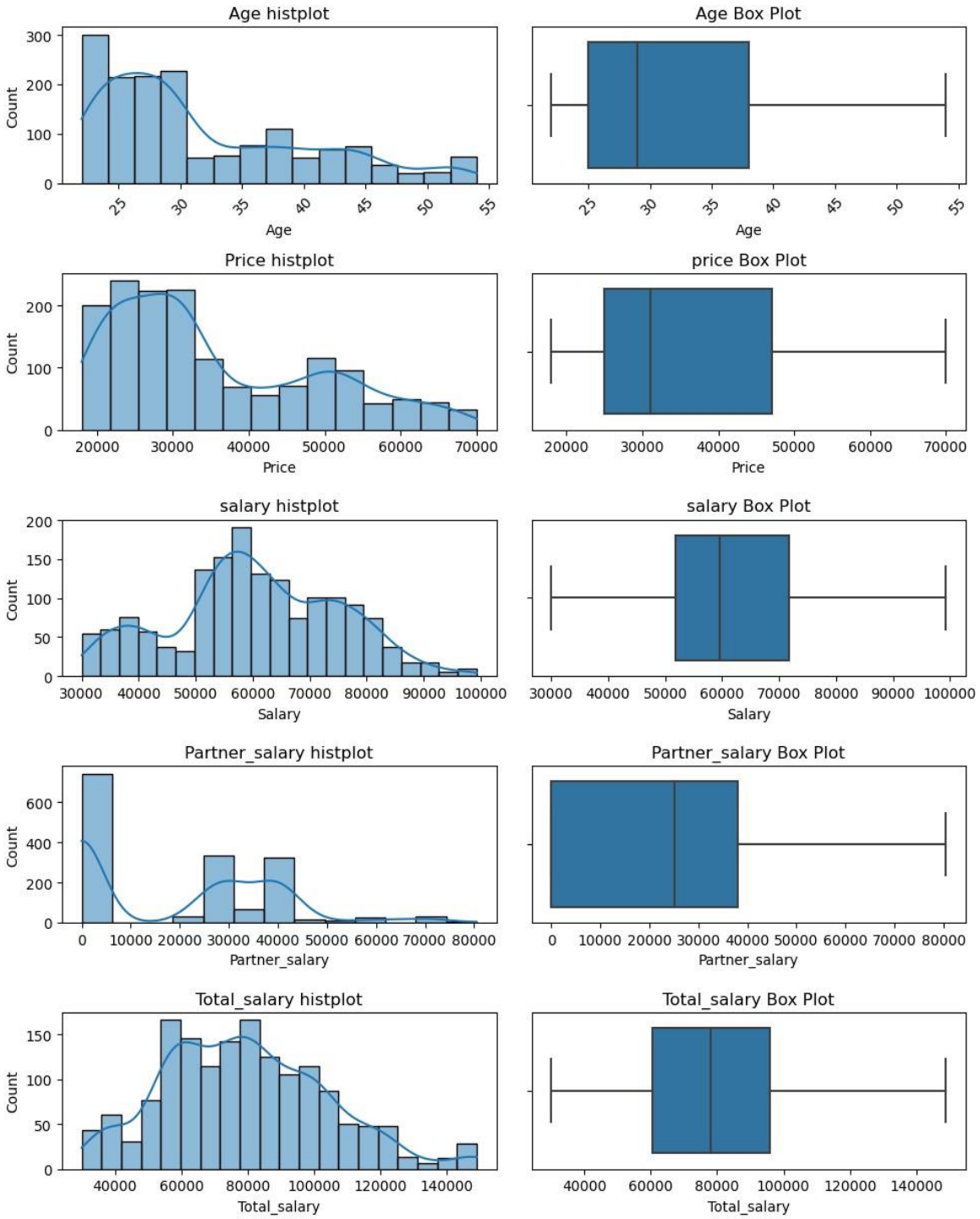

Figure 6: Histplot and boxplot post outliers treatment of Total_salary

Conclusion:

1. The distribution of Total_salary post outlier treatment saw a significant drop in skewness from 0.609706 to 0.424412.
2. Age seems to have highest skewness (positive) with a magnitude of 0.893087.
3. Price has a positive skewness of 0.740874.
4. There seems a decrease in the skewness of Total_salary post outlier treatment from 0.609706 to 0.424412.

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.

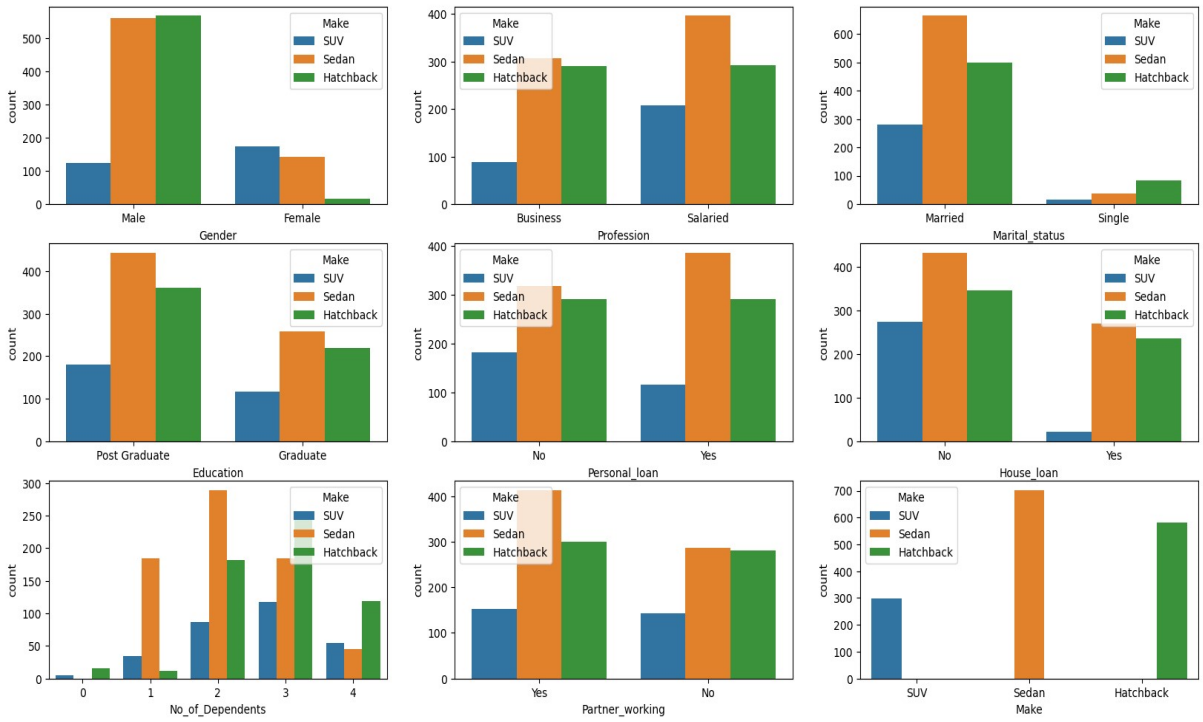

Figure 7: Countplot of categorical variable with Make as Hue

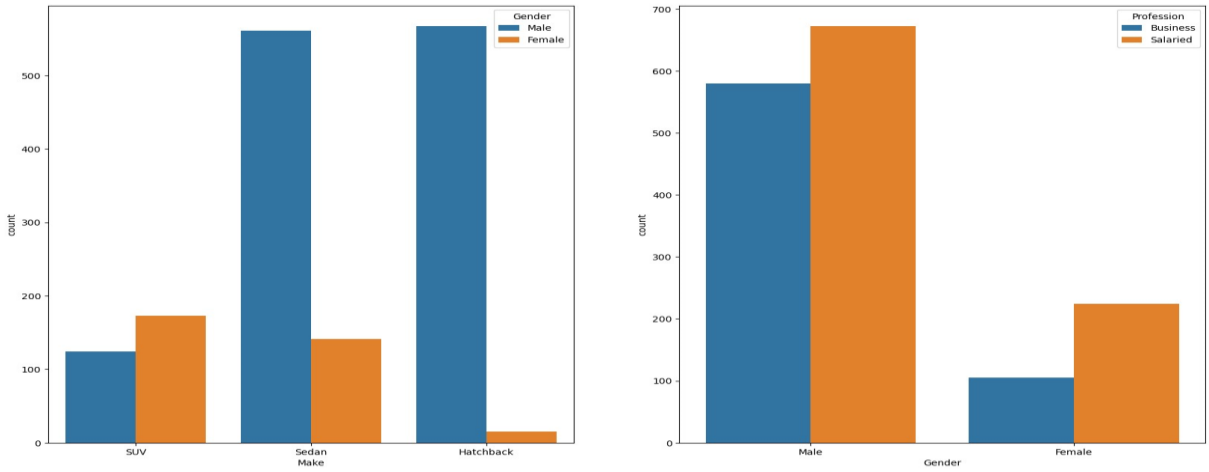

Figure 8: Countplot of Make and Gender

Conclusion:

1. Male prefer hatchback or sedan and Female prefer SUV. SUV is least preferred by male and hatchback is least preferred by female.
2. Business person prefer Sedan or hatchback and salaried person prefer Sedan. SUV is the least preferred by both of them.
3. Single person prefer Hatchback and married person prefer Sedan.
4. A person with house loan or no house loan both prefer Sedan and are not likely to buy SUV. Person with personal loan or no personal loan also has the same trend of preference.
5. Graduate or post graduate both preferred Sedan and SUV is the least preferred.

6. Sedan is preferred by both 1 and 2 number of dependents and Hatchback and SUV are least preferred by number of dependents 1 and 2 respectively. Person with zero dependent does not prefer to buy Sedan.
7. Sedan is preferred by the person if partner working and Sedan or Hatchback is preferred by the person whose partner not working. SUV is the least preferred by the both.
8. Male preferred Sedan and Hatchback and SUV is mostly preferred by female.

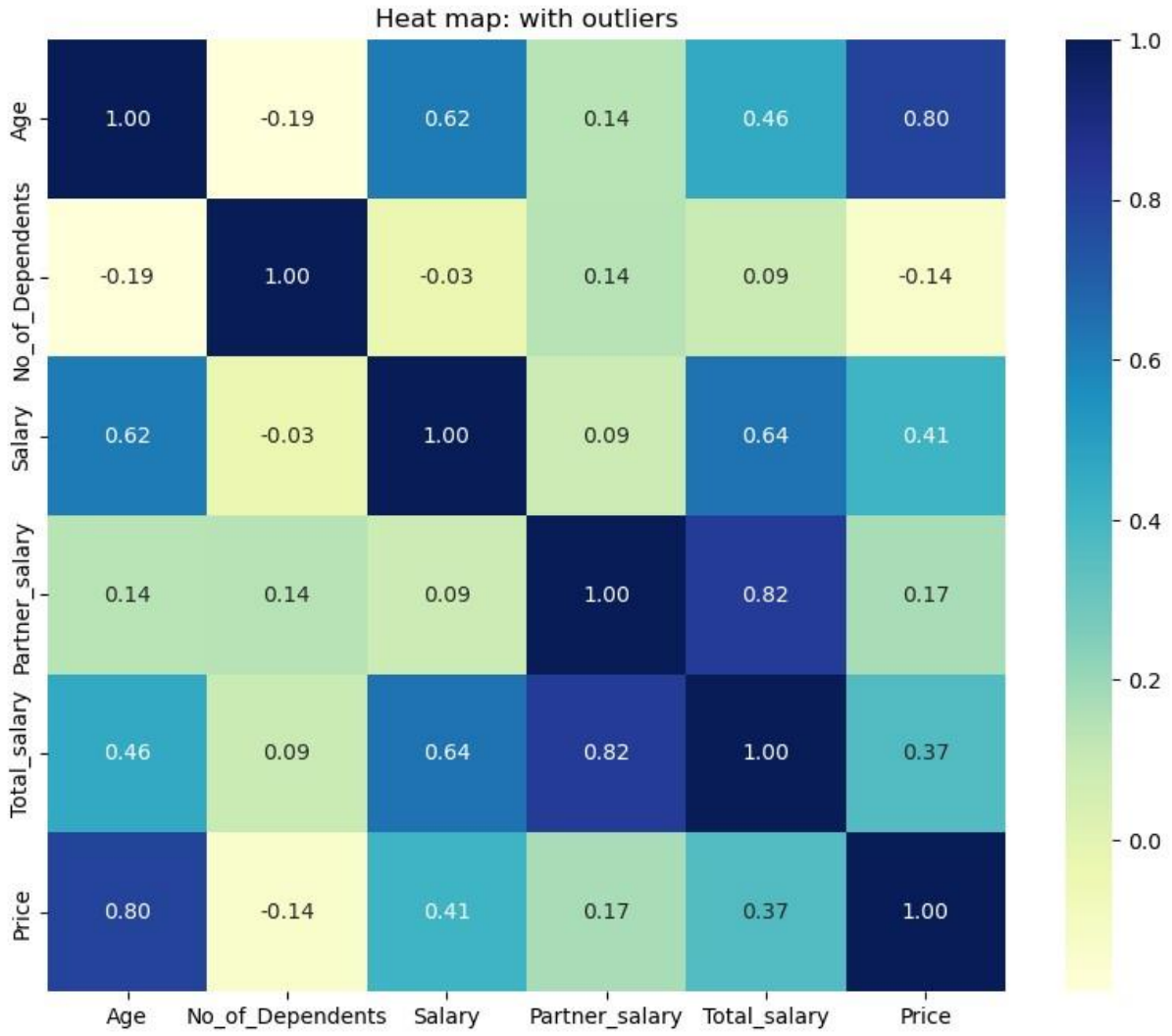

Figure 9: Correlation Heatmap with outliers in Total_salary

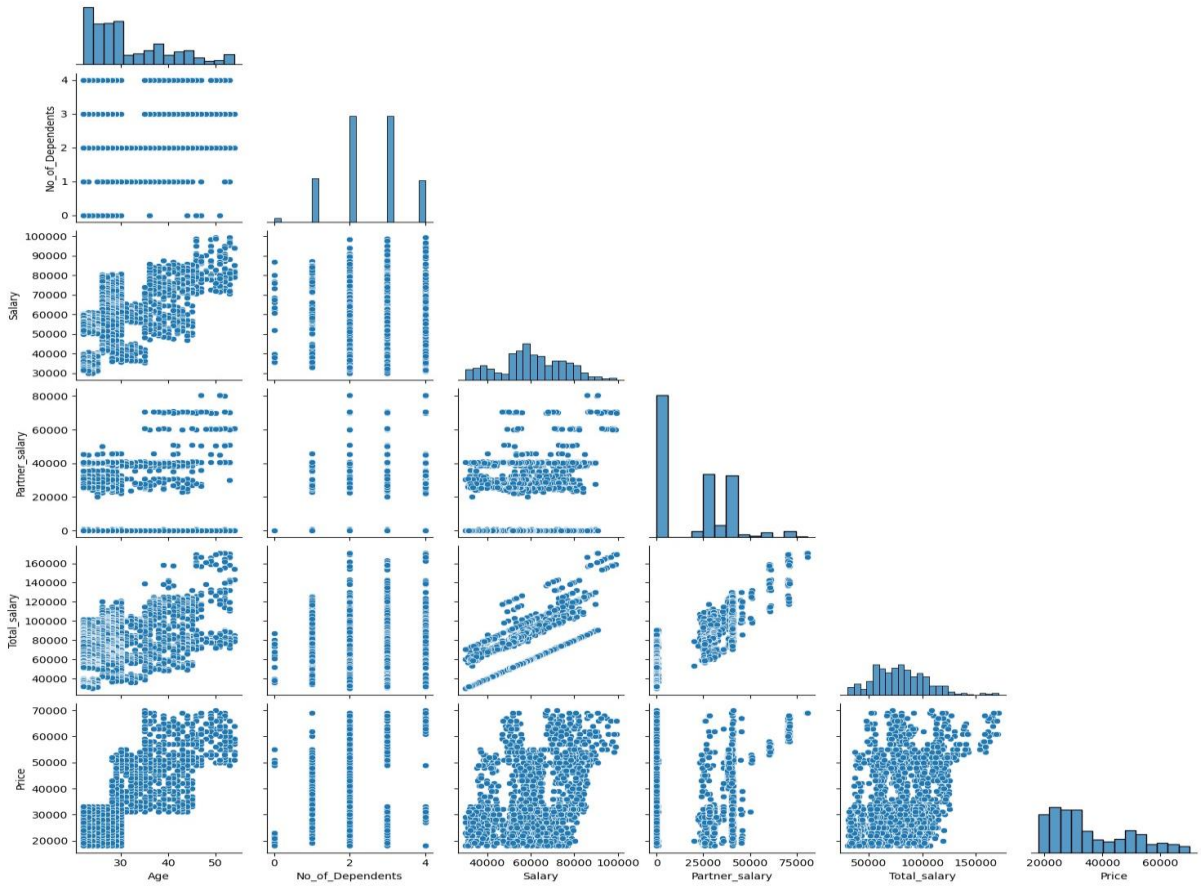

Figure 10: Pair plot with outliers in Total_salary

Post outliers treatment of Total_salary:

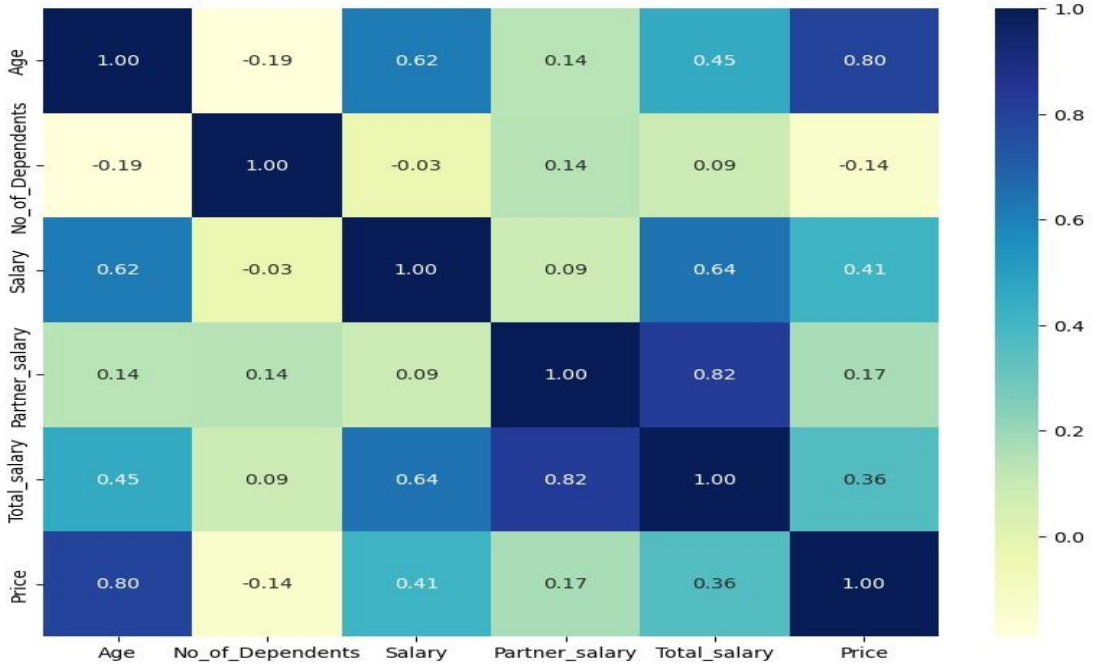

Figure 11: Correlation Heatmap post outlier treatment of Total_salary

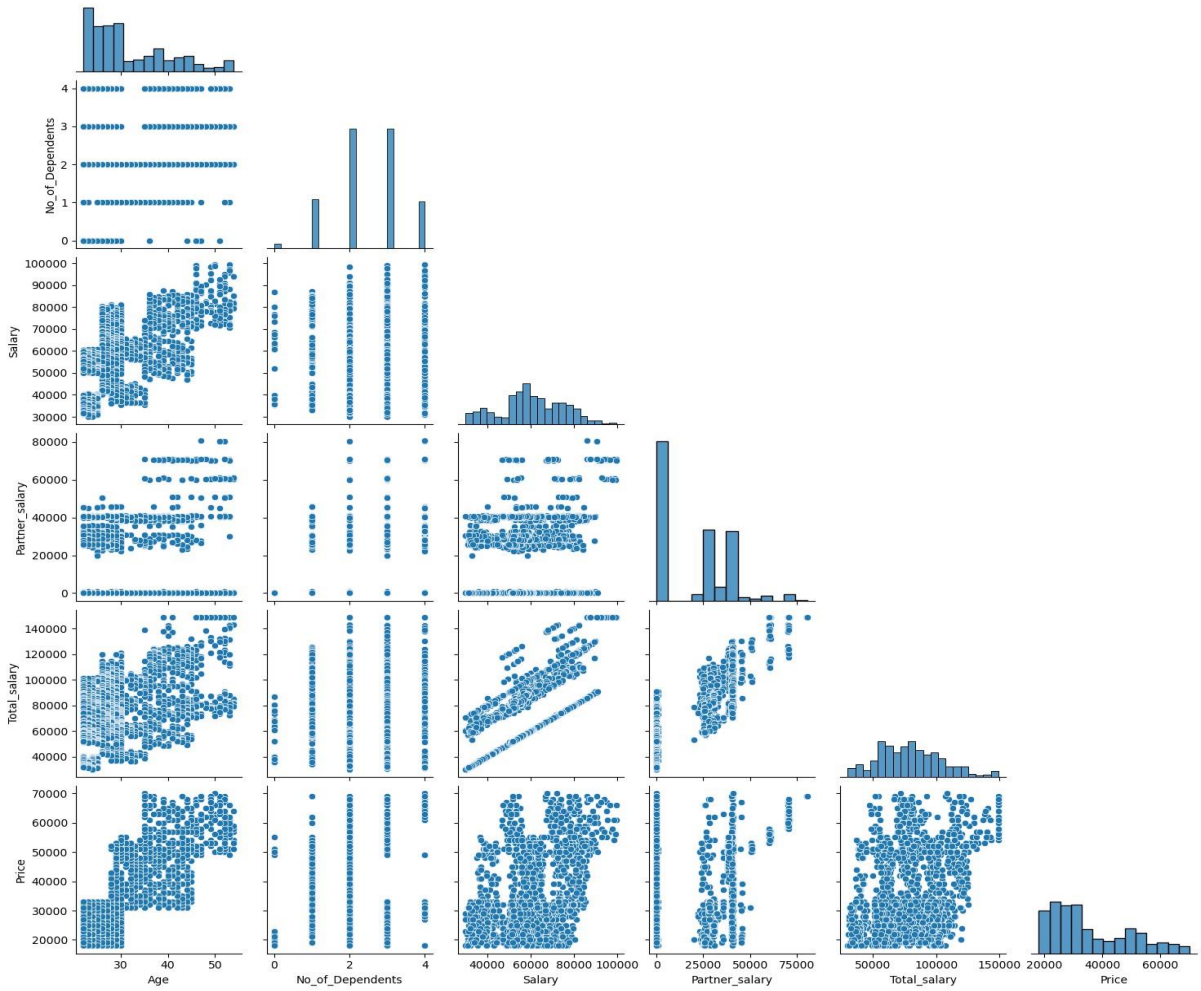

Figure 12: Pairplot post outlier treatment of Total_salary

Conclusion:

1. Maximum correlation is between Total_salary and Partner_salary with outliers and post outliers treatment the maximum correlation is between Age and Price.
2. There is not much correlation between Total_salary and Price of automobile purchased.
3. Based on Age group, price range increases in same proportion.

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”

Male count: 1252
Number of male who bought an SUV: 124

Proportion of number of males buying SUV: 9.90%
Female count: 329

Number of females who bought an SUV: 173
Proportion of number of females buying SUV: 52.58%

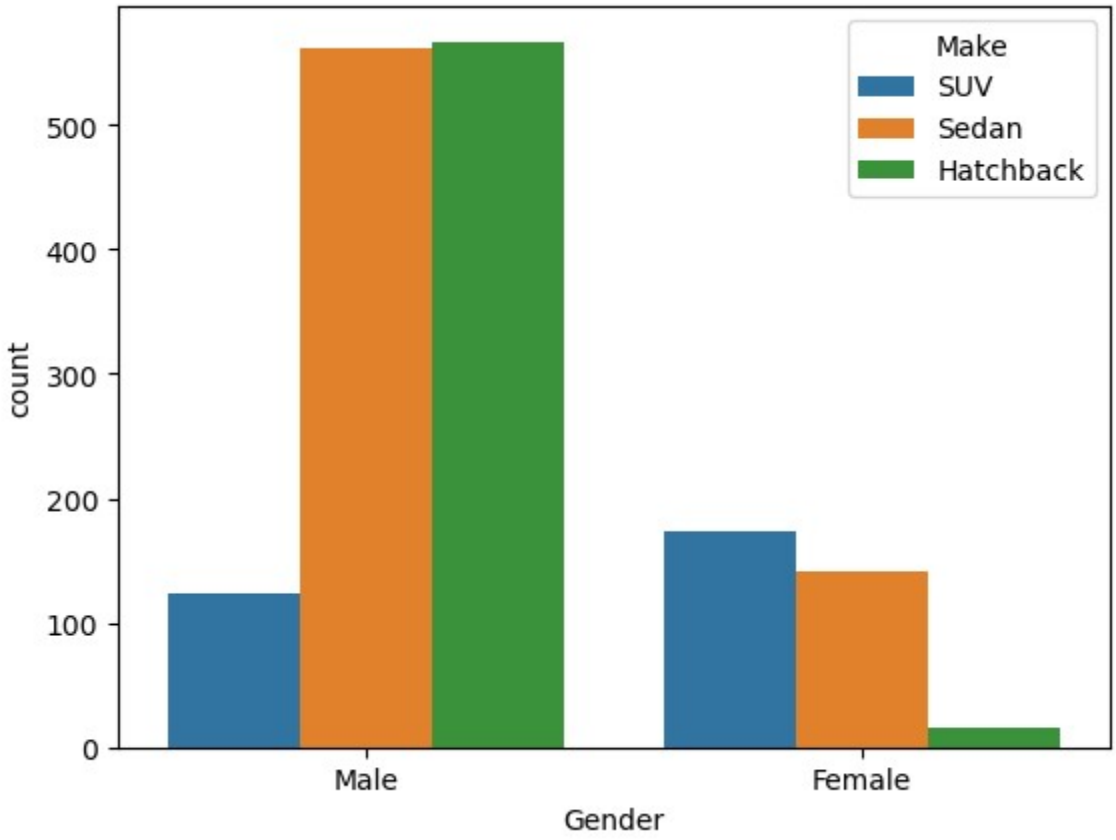

Figure 13: Countplot Of Gender with Make as Hue

From the above analysis we can conclude that statement of Steve roger is wrong and proportion of females buying SUV is 52.58% which is quite large when compared to number of males buying SUV.

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

Count of salaried person who purchased car: 896

Count of salaried persons who bought a Sedan: 396
Percentage of Salaried person who buy a sedan: 44.20%

Count of salaried persons who bought a hatchback: 292
Percentage of Salaried person who buy a hatchback: 32.59%

Count of salaried persons who bought a SUV: 208
Percentage of Salaried person who buy a SUV: 23.21%

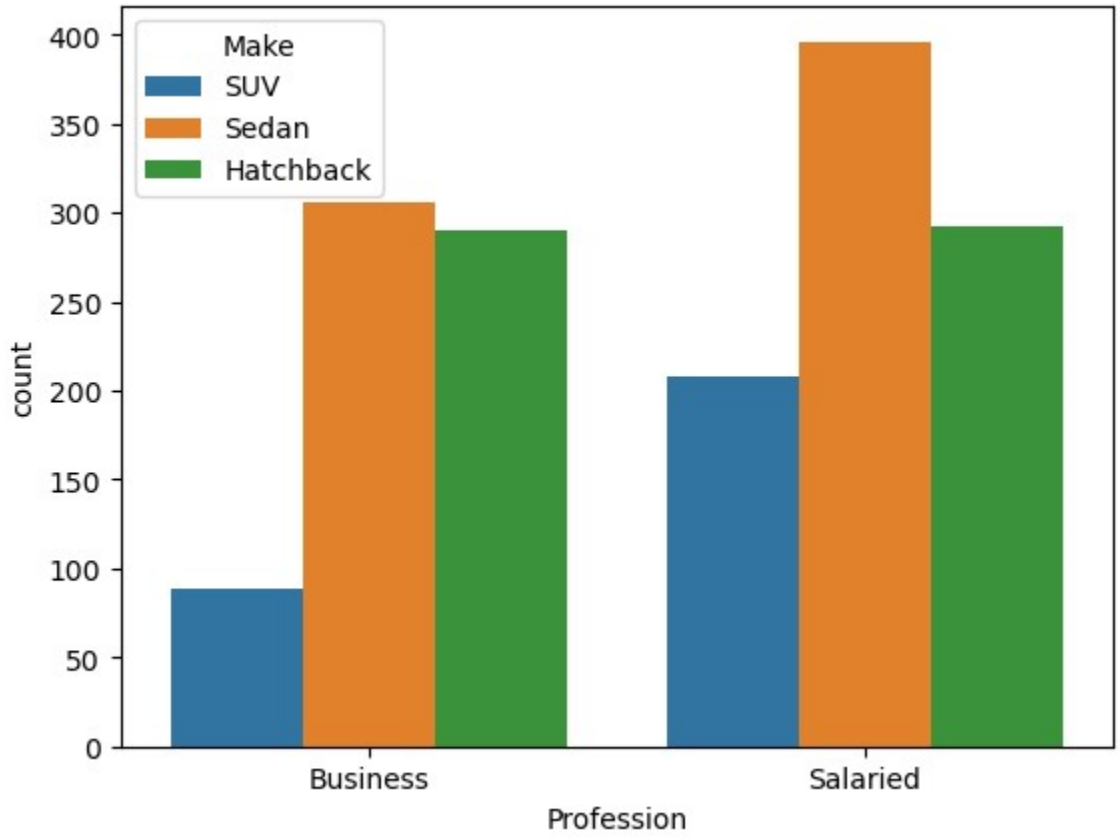

Figure 14: Countplot of Profession with Make as a Hue

From the able result and Countplot it can be concluded that salaried person is more likely to buy a Sedan. Hench 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.

Count of salaried male who purchased car: 672
Count of salaried male who bought a Sedan: 305
Percentage of Salaried male who buy a sedan: 45.39%

Count of salaried male who bought a hatchback: 277
Percentage of salaried male who buy a hatchback: 41.22%

Count of salaried male who bought a SUV: 90
Percentage of Salaried male who buy a SUV: 13.39%

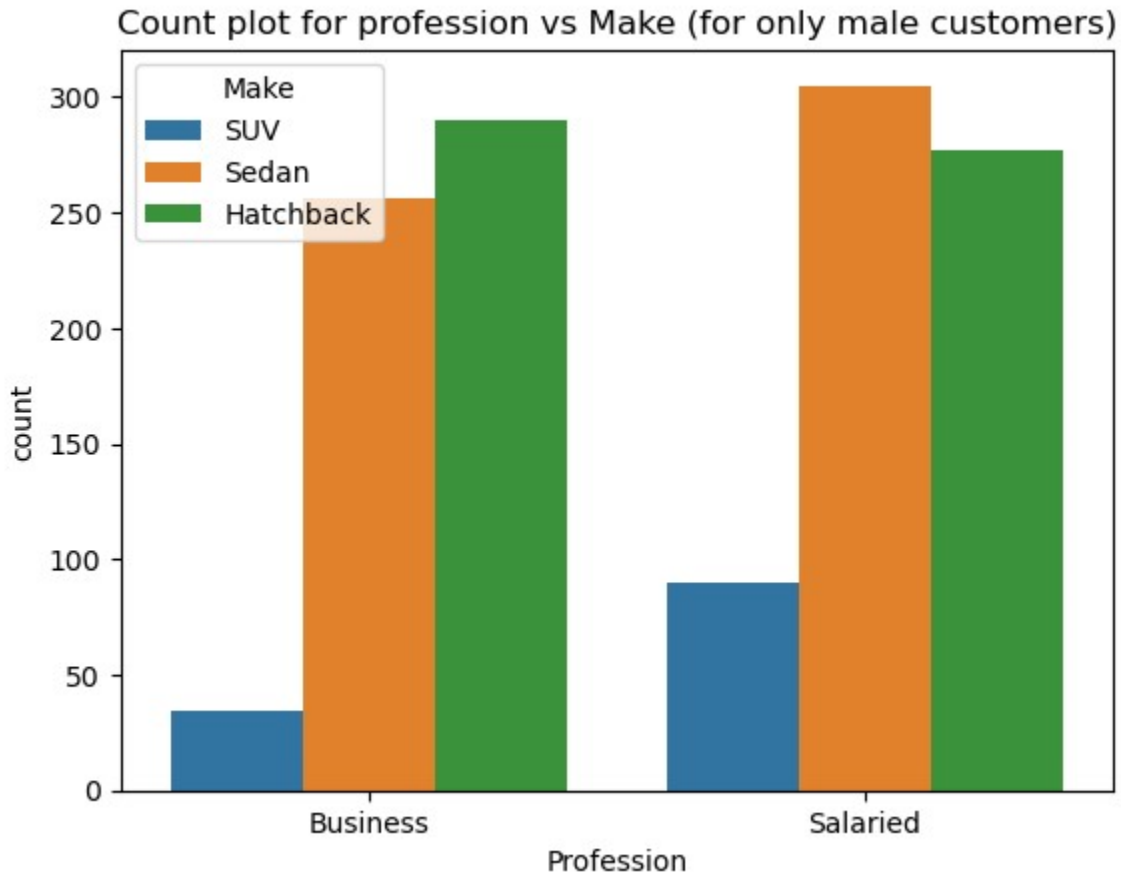

Figure 15: Countplot of Profession vs Make (only for males)

From the result of above analysis and plot of profession with Make as a Hue we arrive to the conclusion that salaried male are easier target for Sedan sale over SUV sale. Hence, statement of Sheldon Cooper is not true.

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.
Give justification along with presenting metrics/charts used for arriving at the conclusions.

F1) Gender

Male count: 1252
Female count: 329
Number of male who bought an SUV: 124
Number of females who bought an SUV: 173

Proportion of number of males buying SUV: 9.90%
Proportion of number of females buying SUV: 52.58%

Number of male who bought a hatchback: 567
Number of females who bought a hatchback: 15

Proportion of number of males buying hatchback: 1.20%
Proportion of number of females buying hatchback: 4.56%

Number of male who bought a sedan: 561
Number of females who bought a sedan: 141

Proportion of number of males buying hatchback: 44.81%
Proportion of number of females buying sedan: 42.86%

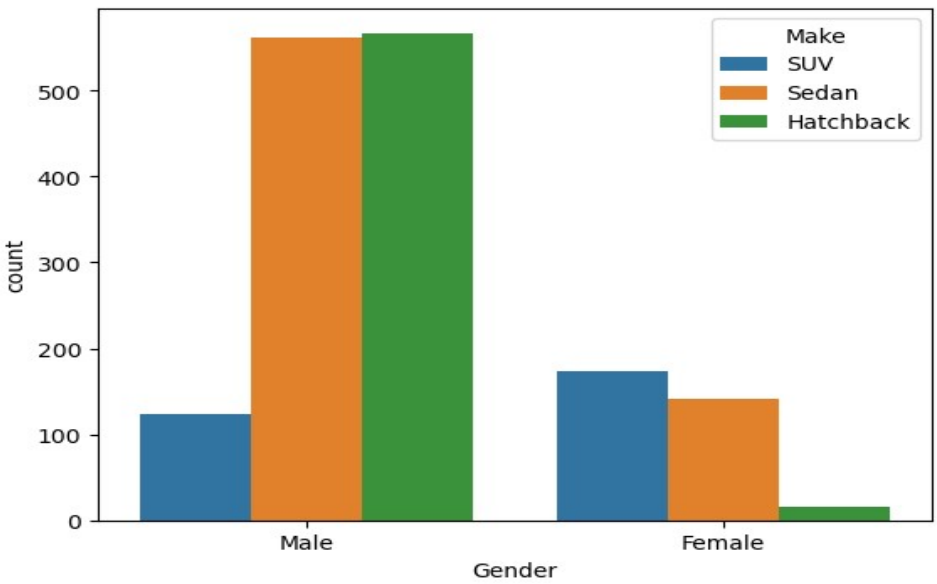

Figure 16: Countplot of Gender with Make as Hue

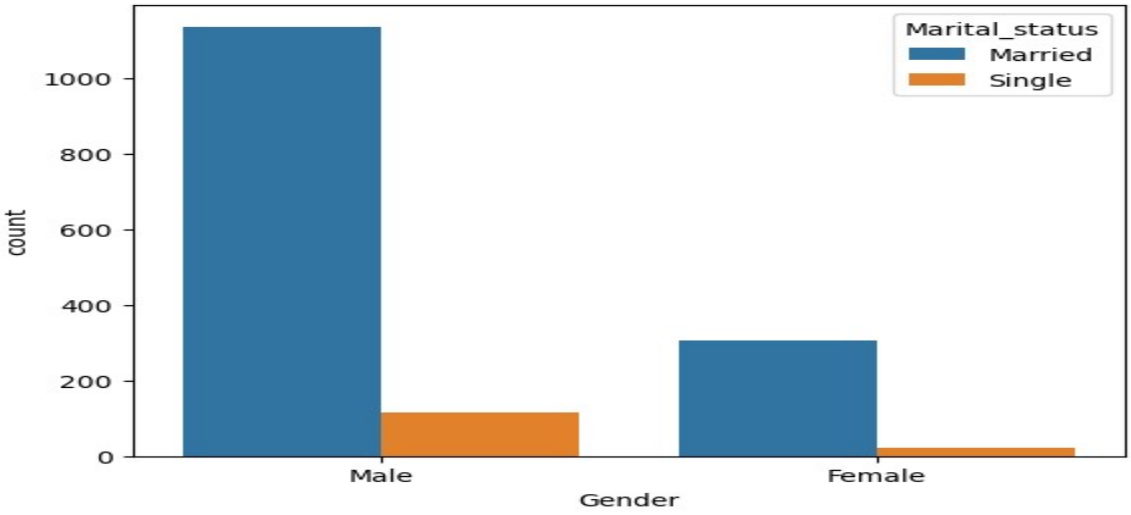

Figure 17: Countplot of Gender with Marital_status as Hue

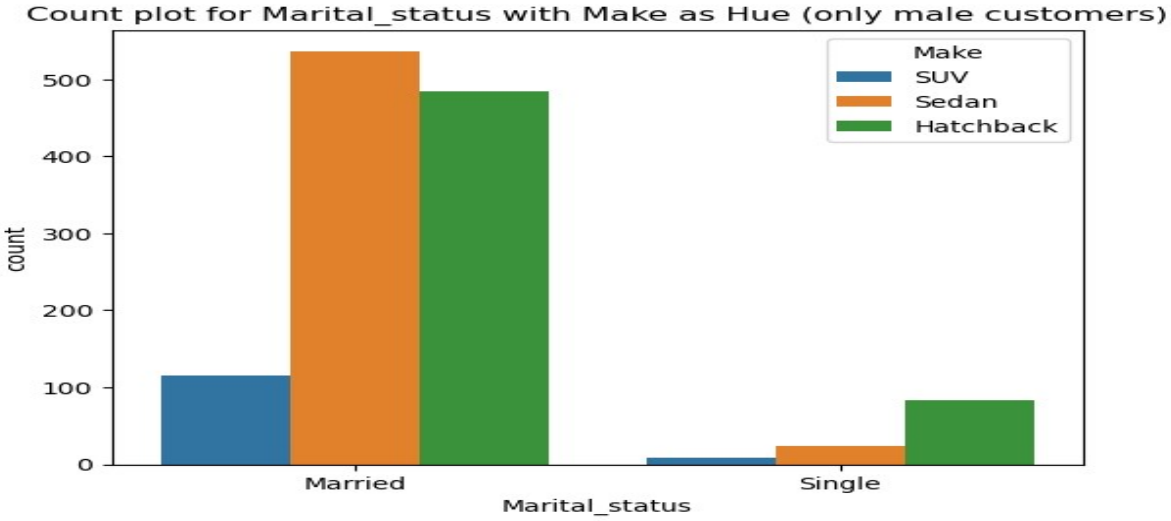

Figure 18: Countplot of Marital_status with Make as Hue (only male customers)

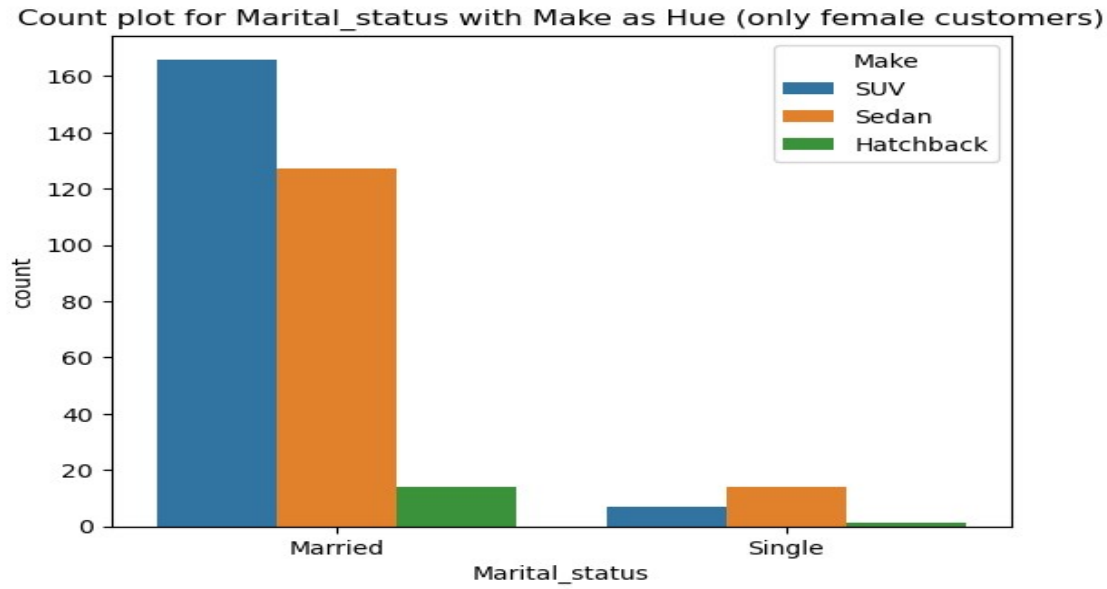

Figure 19: Countplot of Marital_status with Make as Hue (only female customers)

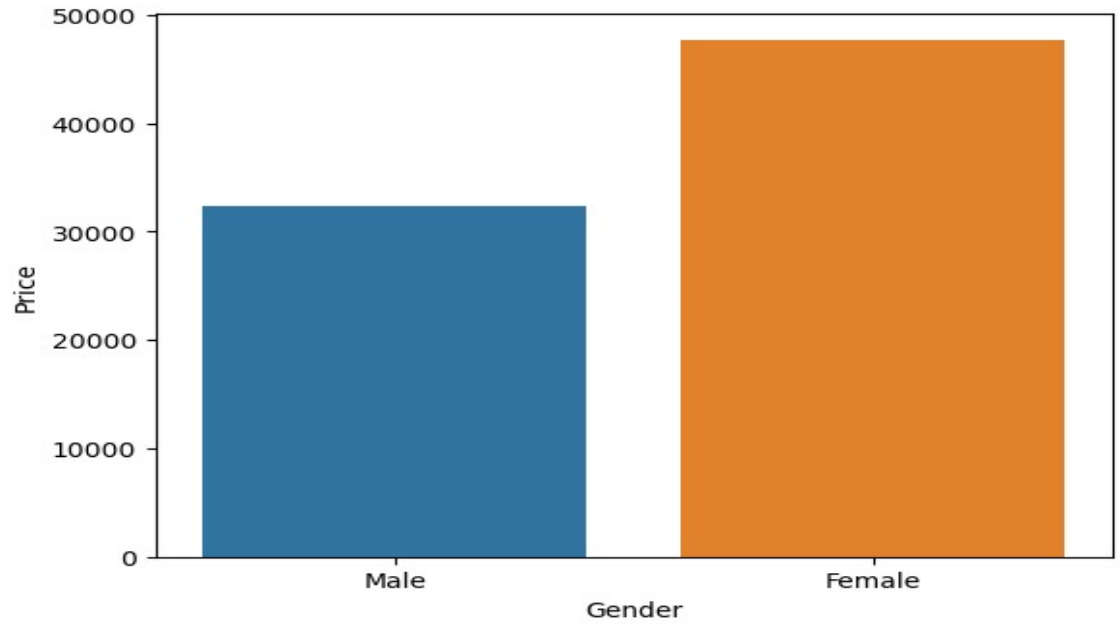

Figure 20: Barplot of Gender

F2) Personal loan

Table 12: Mean and Median of Male and Female Customer (personal loan)

Personal_loan	Mean	Median
No	36742.71	32000.0
Yes	34457.07	31000.0

Table 13: Mean and Median of Male and Female Customers (separated by personal loan status)

Gender	Personal loan	Mean	Median
Female	No	48677.78	50000.0
	Yes	46530.20	47000.0
Male	No	33215.11	30000.0
	Yes	31659.41	28000.0

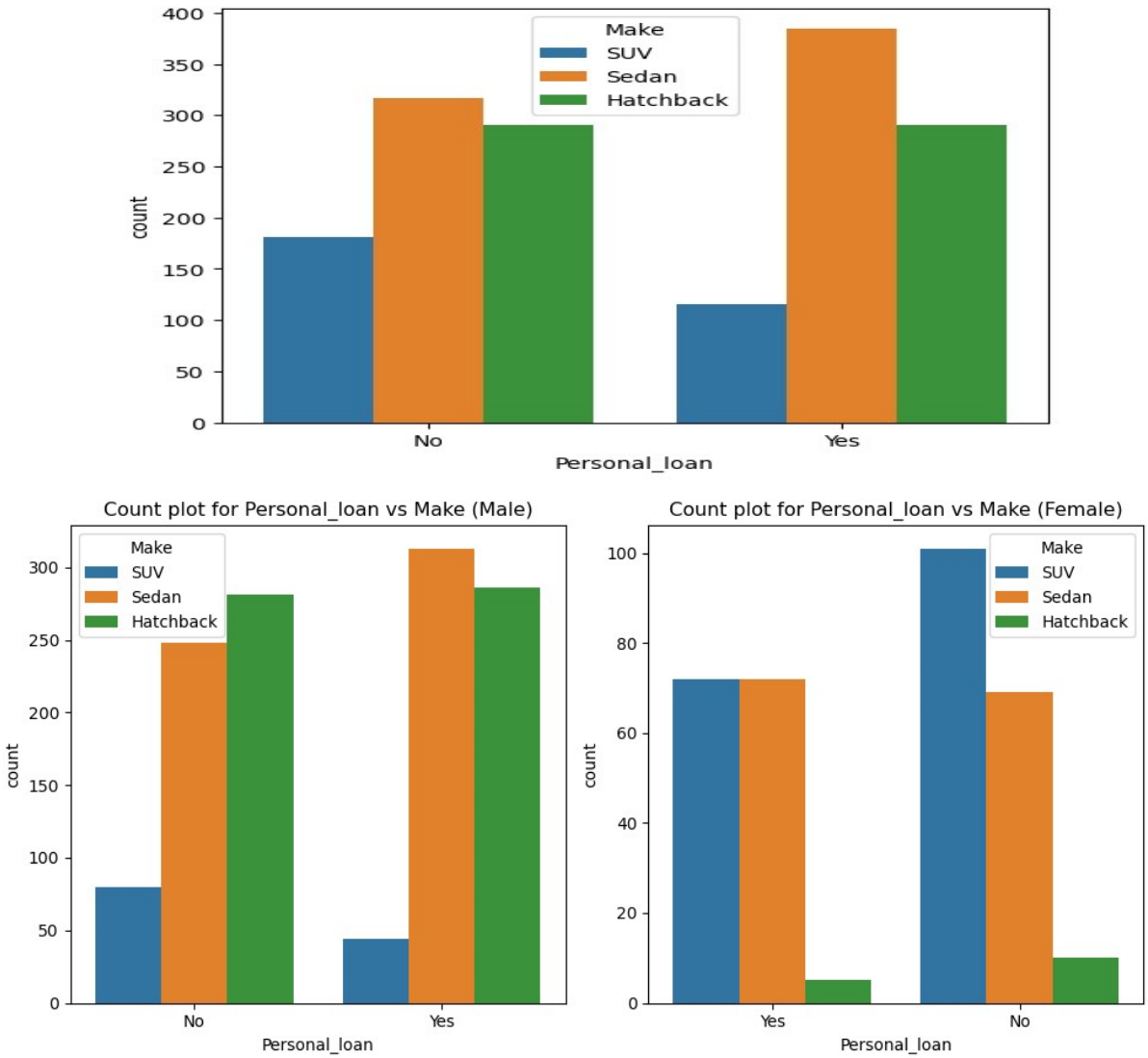

Figure 21: Countplot of Personal_loan for Gender, Male and Female

Conclusion:

From Table() and Figure() we arrive to the result that Customers without a personal loan has a slightly high Mean and Median as compared to customers with a personal loan. This is because customer with no personal loan prefer Hatchback and Sedan over SUV, SUV being the costlier among the three.

From Table() and Figure() on analyzing the Male and Female customer of gender separately we arrive to the result that Male and Female customers without personal loan have higher Mean and Median.

Also mean and Median of female customer is quite high as compared to Male customers.

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

Table 14: Mean and Median of Male and Female Customers

Partner working	Mean	Median
No	36000.00	31000.00
Yes	35267.28	31000.00

Table 15: Mean and Median of Male and Female Customers (separated by partner working):

Gender	Partner working	Mean	Median
Female	No	47218.54	49000.0
	Yes	48117.98	49000.0
Male	No	32985.77	29000.0
	Yes	31952.17	29000.0

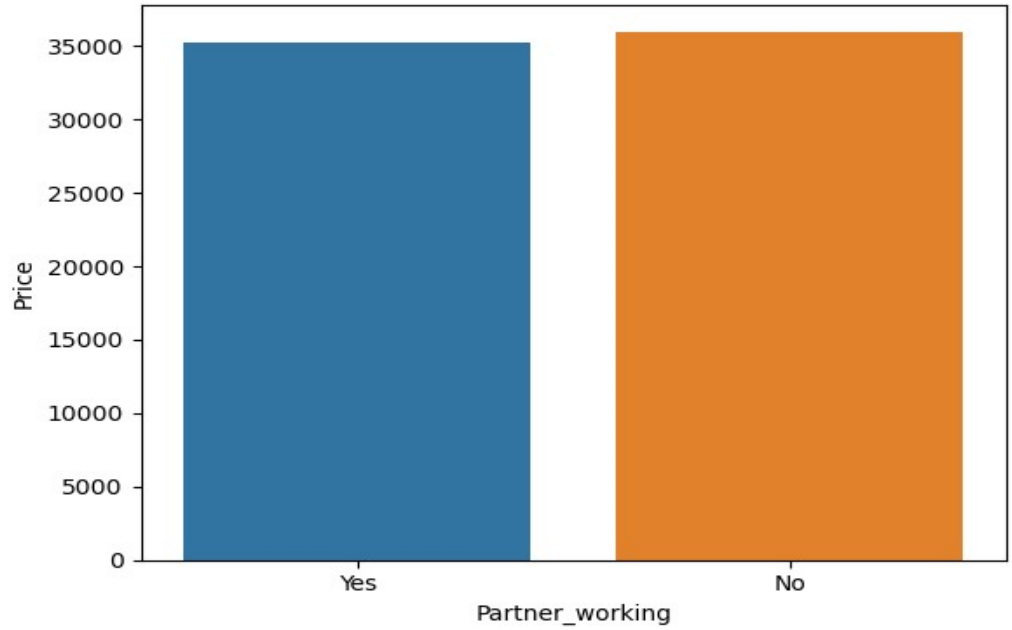

Figure 22: Bar plot of Partner_working

The above result suggest that mean and median of purchase of car is independent of whether the partner is working or not for both male and female and it does not lead to the purchase of costly car.

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.

Table 16: Mean and Median of Male and Female Customers on the basis of Marital status

Marital_status	Mean	Median
Married	35800.42	32000.0
Single	33478.26	30000.0

Table 17: Mode of make- SUV, sedan and Hatchback

Gender	Marital_status	Make
Female	Married	SUV
	Single	Sedan
Male	Married	Sedan
	Single	Hatchback

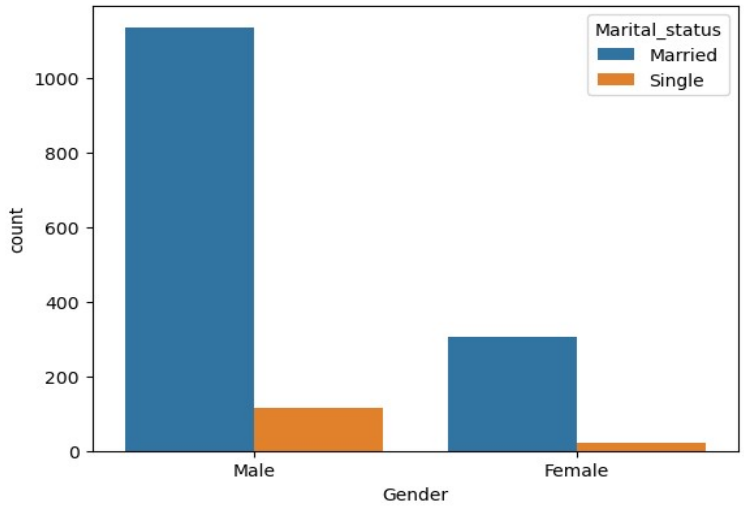

Figure 23: Countplot of Gender: Hue as Marital_status

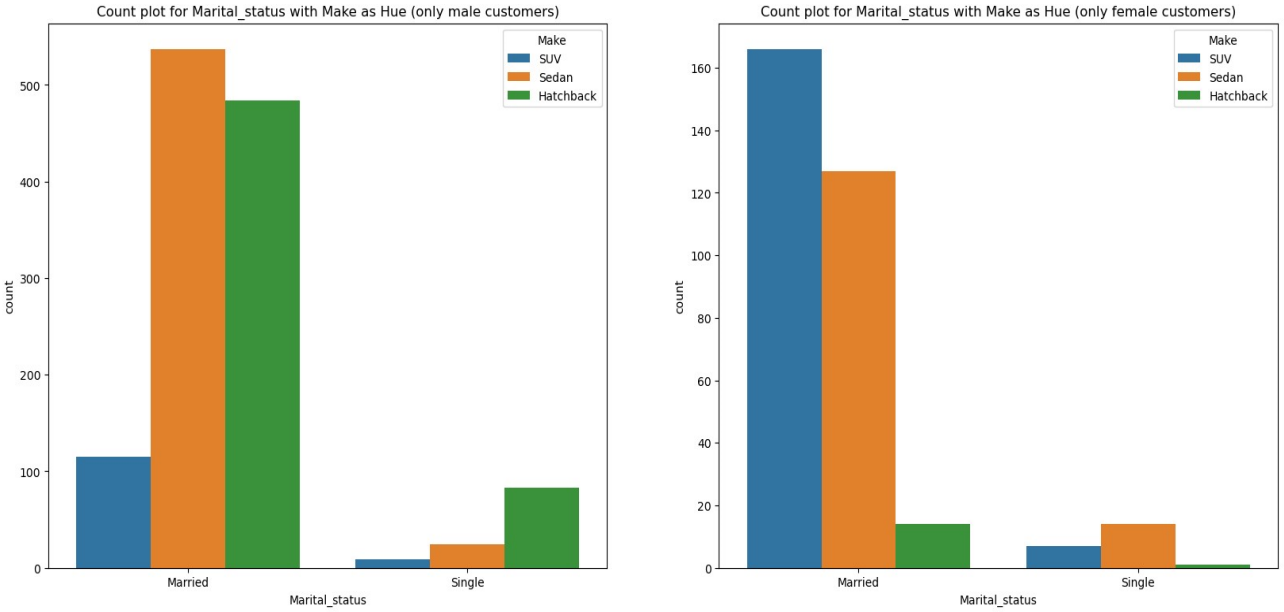

Figure 24: Countplot of Marital_status: Hue as Make

To improve the marketing strategy to send targeted information to different groups of potential buyers present in the data we need to analyse the recent historical data based on Gender and marital status. From the above Table-16 we arrive to the conclusion that Married female are most likely to prefer SUV and Single female and married male prefer Sedan and Single male are most likely to prefer Hatchback.

- Married Female : SUV
- Single Female : Sedan
- Married Male : Sedan
- Single Male : Hatchback

Problem 2

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.)

Problem 2 Question: (Analyze the dataset and list down the top 5 important variables, along with the business justifications.

tarunbh2394@gmail.com

Top 5 important variables are

1. **annual_income_at_source**: It is one of the important factor to get the information about the purchasing power of an individual and bank can utilize this information for targeted advertisement, deciding the maximum cap on the loan amount, deciding interest rate, deciding the offers etc.
2. **T+1_month_activity** : It can be utilized for customer segmentation, targeted marketing compains, promotions and offers based on their spending habits, income levels to help identify potential credit card customers likely to benefit from their offering.
3. **avg_spends_l3m:** This information can be utilized by Bank in several ways such as Customized credit card offers, rewards and benefits alignment as per the customers spending categories, tailored credit card by using average credit card spending to determine appropriate credit limit, specialized credit card recommendations, Upselling and cross selling opportunities and Targeted marketing campaigns.
4. **Occupation_at_source:** The occupation_at_source can give important insights on income assessment and so that credit limit can be determined credit card can be customized based on applicant occupation. Targeted marketing compains to promote credit card effectively, affinity programs for specific occupations can be established. Also information of occupation can be used as a part of risk assessment.
5. **cc_limit:** The cc limit can act as an attractive feature to potential customers. This is an essential feature for the risk assessment to evaluate the financial stability. Bank considers the spending capacity of customer while setting the credit card limit to minimize the defaulters. Credit card limit also affects credit utilization ratio.

tarunbh2394@gmail.com
gbpu0073