Contents

Problem 1 (a):

Problem 1 - Data Overview- Import the libraries - Load the data - Check the structure of the data - Check the types of the data - Check for and treat (if needed) missing values - Check the statistical summary - Check for and treat (if needed) data irregularities - Observations and Insights

Code Snippet of data Overview:

Data Overview

```
<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
```

1. There are total six (6) numerical fields in the data set and total seven (8) categorical fields in data set

2. The names of numeric fields and description of numeric Fields about them are as follow

S.No.	Name	Description of numeric Fields
1	Age	The age of customer who bought the automobile
2	No_of_Dependents	The number of dependents (i.e the person who are dependent on customer)
3	Salary	The salary of customers
4	Partners_Salary	The partners salary of customers in case he/she is working
5	Total_salary	The total salary of customers
6	Price	The price of the automobile bought by the customers.

3. The names of categorical field description of numeric Fields about them are as follow

S.No.	Name	Description of Categorical Fields
1	Gender	The gender of customer
2	Profession	The profession of customer
3	Maritial_status	Is the customer married or not
4	Education	What is the education of the customer
5	Personal_Loan	Whether the customer has taken personal Loan or not
6	House_Loan	Whether the customer has taken house loan or not
7	Partner_working	Whether the partners is working in case partner is working or not.
8	Make	The make of Car Bought

Major Insights from Above Data

Part A: Categorical data

Code Snippets -Unique Values of Categorical Data

```
uniques values of Gender ['Male' 'Femal' 'Female' nan 'Femle']
uniques values of Profession ['Business' 'Salaried']
uniques values of Marital_status ['Married' 'Single']
uniques values of Education ['Post Graduate' 'Graduate']
uniques values of Personal_loan ['No' 'Yes']
uniques values of House_loan ['No' 'Yes']
uniques values of Partner_working ['Yes' 'No']
uniques values of Make ['SUV' 'Sedan' 'Hatchback']
```

1. The gender of the customer can classified into two categories which are male & female. **The incorrect values as show above is a) Femal b) Femle c) nan/Nil need to be imputed using correct values.**
2. The profession of the customer can be classified into two categories which are Business & Salaried.
3. The marital status of the customer can be classified into two categories which are Married & Single
4. The Education status of the customer can be classified into two categories which are Post Graduate & Graduate.
5. The Personal Loan can be classified into two categories which are No & Yes
6. The Housing Loan status of the customer can be classified into two categories which are No & Yes.
7. The Partner working status of the customer can be classified into two categories which are No & Yes.
8. The Make of the customer can be classified into three categories which are SUV, Sedan Make & Hatch Back
9. **Missing Values code Snippet**

```
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
```

10. For Gender as its is a categorical values we need to imputed it by mode.

Part B: Numerical Data

Code Snippets-Description of Numerical Data

	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1475.0	20225.559322	19573.149277	0.0	0.0	25600.0	38300.0	80500.0
Total_salary	1581.0	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	1581.0	35597.722960	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0

1. The Total Number of records are complete except in case of partners salary.
2. The number of missing records in Partners salary are 106.
3. Out of which 16 values has partners working as yes but Partners salary is show as Nan

]:

	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary
43	52	Male	Salaried	Married	Post Graduate	3	No	No	Yes	87600	NaN	88200
49	52	Female	Business	Married	Post Graduate	4	No	No	Yes	90300	NaN	170400
59	54	Male	Salaried	Married	Graduate	3	Yes	No	Yes	80600	NaN	81000
111	48	Female	Business	Married	Graduate	3	No	No	Yes	90300	NaN	161100
209	43	Female	Salaried	Married	Graduate	4	Yes	No	Yes	53400	NaN	123900
284	41	Female	Business	Married	Post Graduate	2	No	No	Yes	70500	NaN	105800
339	39	Male	Salaried	Married	Post Graduate	2	Yes	No	Yes	76800	NaN	115400
376	38	Male	Salaried	Married	Post Graduate	2	Yes	No	Yes	79000	NaN	117400
424	37	Female	Salaried	Married	Graduate	2	Yes	Yes	Yes	62000	NaN	100700
444	36	Male	Business	Married	Post Graduate	2	No	No	Yes	72300	NaN	112400
554	33	Male	Salaried	Married	Graduate	2	No	No	Yes	41600	NaN	70100
654	30	Male	Business	Married	Post Graduate	4	Yes	Yes	Yes	64700	NaN	93200
779	29	Male	Salaried	Married	Graduate	3	No	Yes	Yes	59000	NaN	87700
1345	24	Male	Salaried	Married	Graduate	3	Yes	No	Yes	34600	NaN	58800
1349	24	Male	Business	Married	Graduate	3	Yes	No	Yes	32400	NaN	58200
1546	22	Male	Business	Married	Graduate	3	Yes	No	Yes	32600	NaN	59300

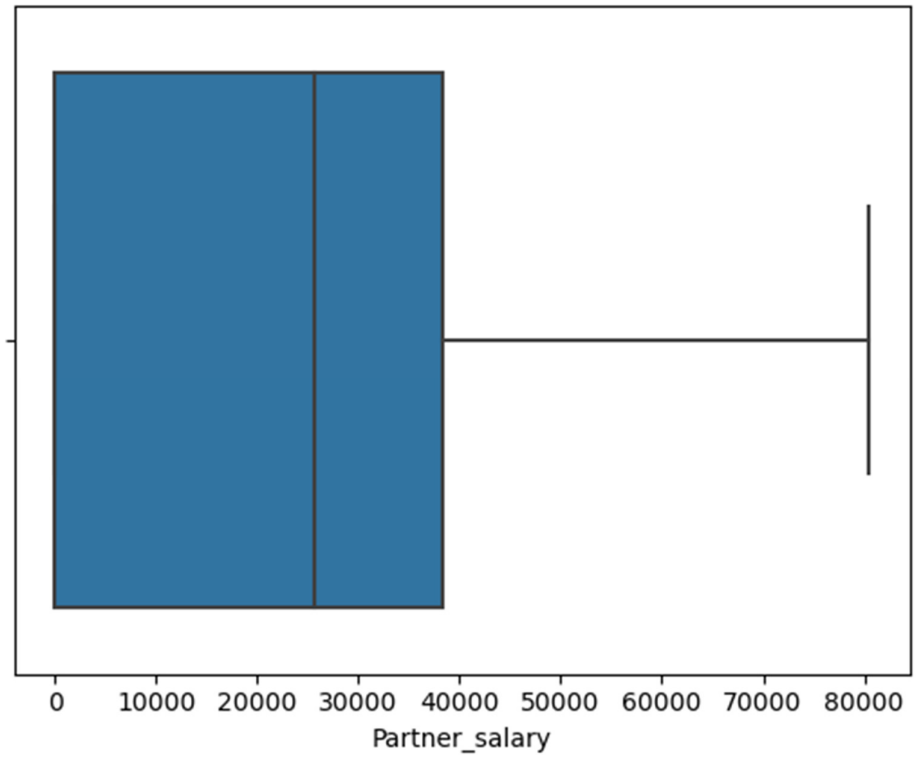
4. Balance 90 Values have shown as Partners working as No & Salary shown as Nan

]:

	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary
40	53	Female	Salaried	Married	Graduate	1	Yes	No	No	72100	NaN	72100
115	48	Female	Salaried	Married	Post Graduate	3	No	No	No	78000	NaN	78000
163	45	Male	Salaried	Married	Post Graduate	1	Yes	Yes	No	71300	NaN	71300
164	45	Male	Business	Married	Graduate	1	Yes	No	No	56700	NaN	56700
165	45	Male	Salaried	Married	Graduate	2	No	No	No	55100	NaN	55100
...	...	...	...	...	...	...	...	...	...	...	...	...
1559	22	Male	Business	Married	Post Graduate	3	Yes	No	No	52100	NaN	52100
1567	22	Male	Salaried	Single	Graduate	0	Yes	Yes	No	39700	NaN	39700
1568	22	Male	Salaried	Married	Graduate	3	No	Yes	No	38000	NaN	38000
1577	22	Male	Business	Married	Graduate	4	No	No	No	32000	NaN	32000
1579	22	Male	Business	Married	Graduate	3	Yes	Yes	No	32200	NaN	32200

90 rows × 14 columns

5. **Imputation is needed as below**
   - **A. The first 18 values not be Imputed by Mean Values as there are no outliners as shown in boxplot.**
   - **B. The Balance 90 Values needs to be imputed by 0 as partners is not working.**



Code Snippet after Imputation

	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1581.0	19074.199209	19477.709066	0.0	0.0	24900.0	38000.0	80500.0
Total_salary	1581.0	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	1581.0	35597.722960	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0

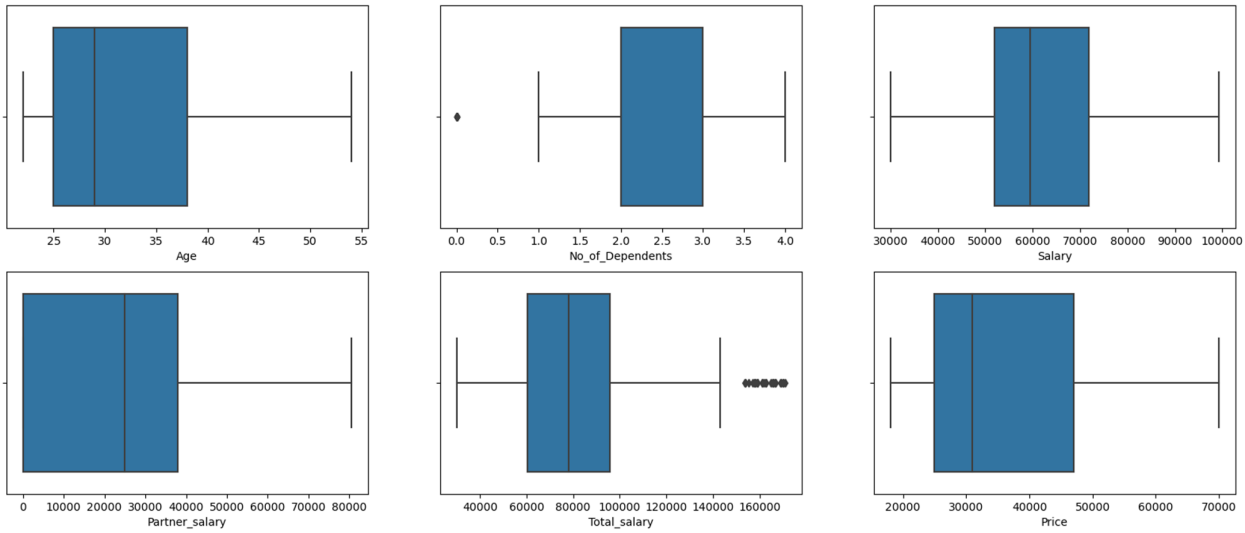
```
uniques values of Gender ['Male' 'Female']
uniques values of Profession ['Business' 'Salaried']
uniques values of Marital_status ['Married' 'Single']
uniques values of Education ['Post Graduate' 'Graduate']
uniques values of Personal_loan ['No' 'Yes']
uniques values of House_loan ['No' 'Yes']
uniques values of Partner_working ['Yes' 'No']
uniques values of Make ['SUV' 'Sedan' 'Hatchback']
```

```
<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   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
```

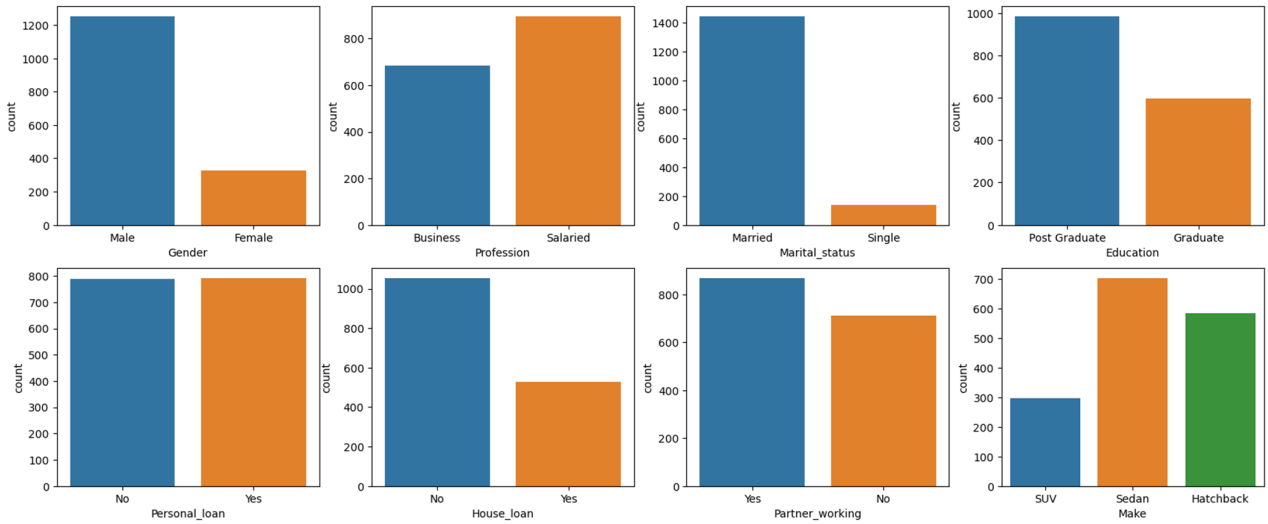
Problem 1 (b):

Problem 1 - Univariate Analysis- Explore all the variables (categorical and numerical) in the data - Check for and treat (if needed) outliers - Observations and Insights

Code Snippet 1 : Numerical Figures



Code Snippet 2 : Categorical Figures



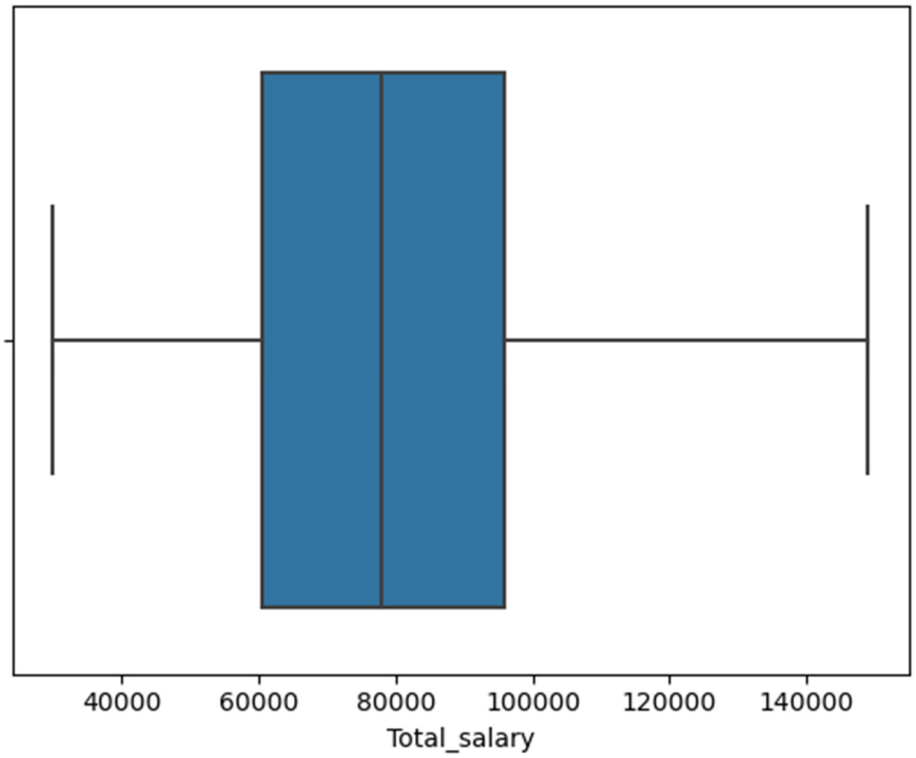
Major Insights and Observations (Including Outliners Treatment)

Part A: Outliners Treatment

1. There are no outlines in numerical values except
   A. Total Salary. It needs to be corrected.
   B. Number of dependents: Needs not be corrected as it can be 0 for persons who are single (As per below code snippet)

	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary
93	51	Male	Salaried	Single	Post Graduate	0	Yes	No	No	86900	0.0	86900
128	47	Female	Business	Single	Graduate	0	Yes	No	No	73300	0.0	73300
138	46	Female	Salaried	Single	Post Graduate	0	Yes	No	No	80200	0.0	80200
203	44	Male	Salaried	Single	Post Graduate	0	Yes	No	No	68600	0.0	68600
462	36	Female	Salaried	Single	Post Graduate	0	No	No	No	67500	0.0	67500
701	30	Male	Business	Single	Post Graduate	0	Yes	No	No	67000	0.0	67000
826	29	Male	Salaried	Single	Post Graduate	0	Yes	Yes	No	62300	0.0	62300
912	28	Male	Business	Single	Post Graduate	0	Yes	Yes	No	76600	0.0	76600
936	28	Male	Business	Single	Post Graduate	0	No	Yes	No	66300	0.0	66300
1020	27	Male	Salaried	Single	Post Graduate	0	No	No	No	76000	0.0	76000
1049	27	Male	Business	Single	Post Graduate	0	Yes	No	No	63600	0.0	63600
1133	26	Male	Business	Single	Post Graduate	0	Yes	No	No	67500	0.0	67500
1220	25	Male	Salaried	Single	Post Graduate	0	Yes	No	No	60700	0.0	60700
1369	24	Male	Business	Single	Graduate	0	Yes	No	No	38200	0.0	38200
1372	24	Male	Salaried	Single	Graduate	0	No	Yes	No	35400	0.0	35400
1374	24	Male	Business	Single	Graduate	0	No	No	No	35700	0.0	35700
1477	23	Male	Business	Single	Graduate	0	No	Yes	No	35700	0.0	35700
1562	22	Male	Salaried	Single	Post Graduate	0	Yes	Yes	No	51900	0.0	51900
1567	22	Male	Salaried	Single	Graduate	0	Yes	Yes	No	39700	0.0	39700
1570	22	Male	Business	Single	Graduate	0	Yes	No	No	38000	0.0	38000

2. Imputation needs to be done using IQR Method
3. Boxplot after Imputations



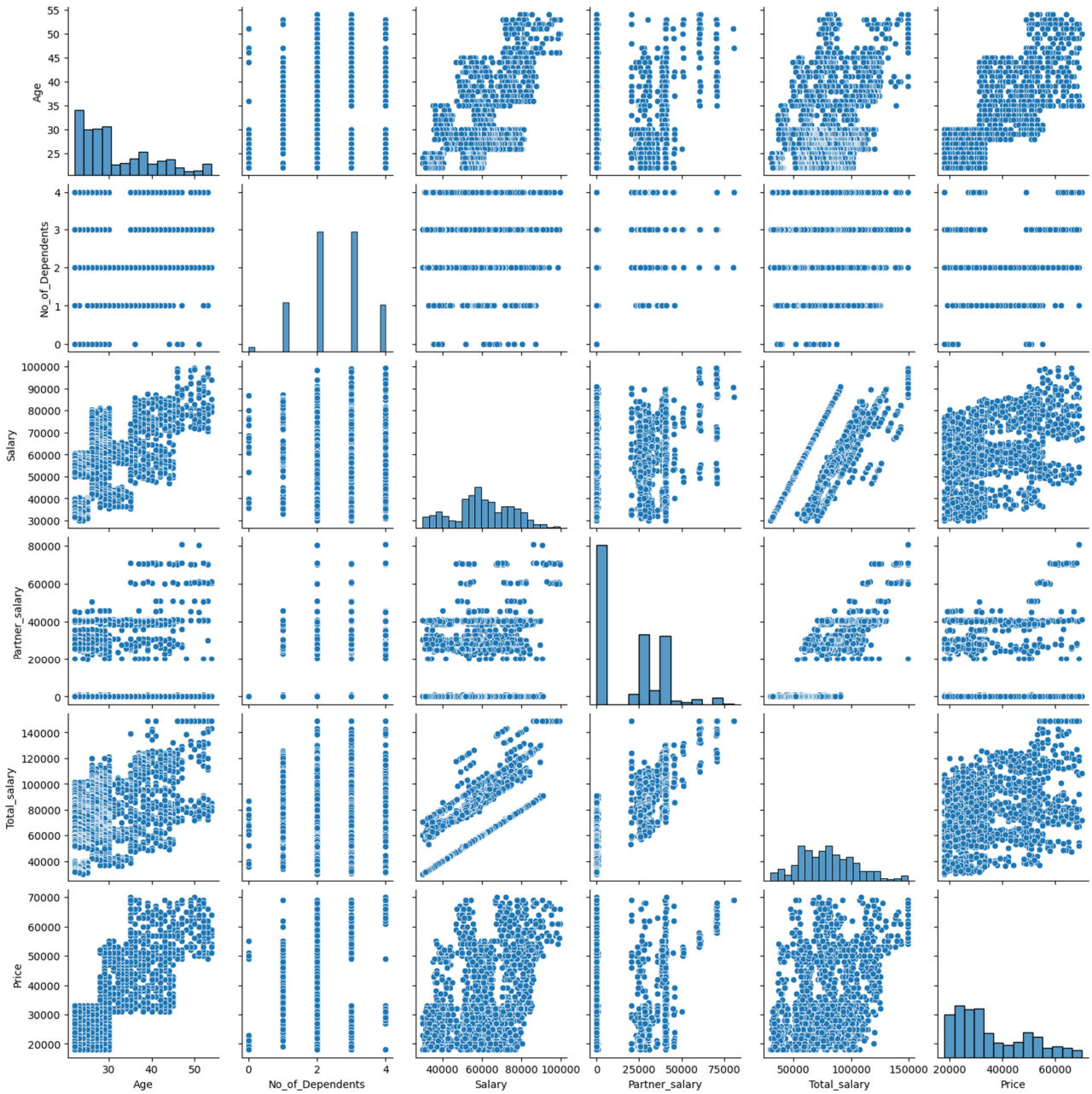
Part B Major Insights

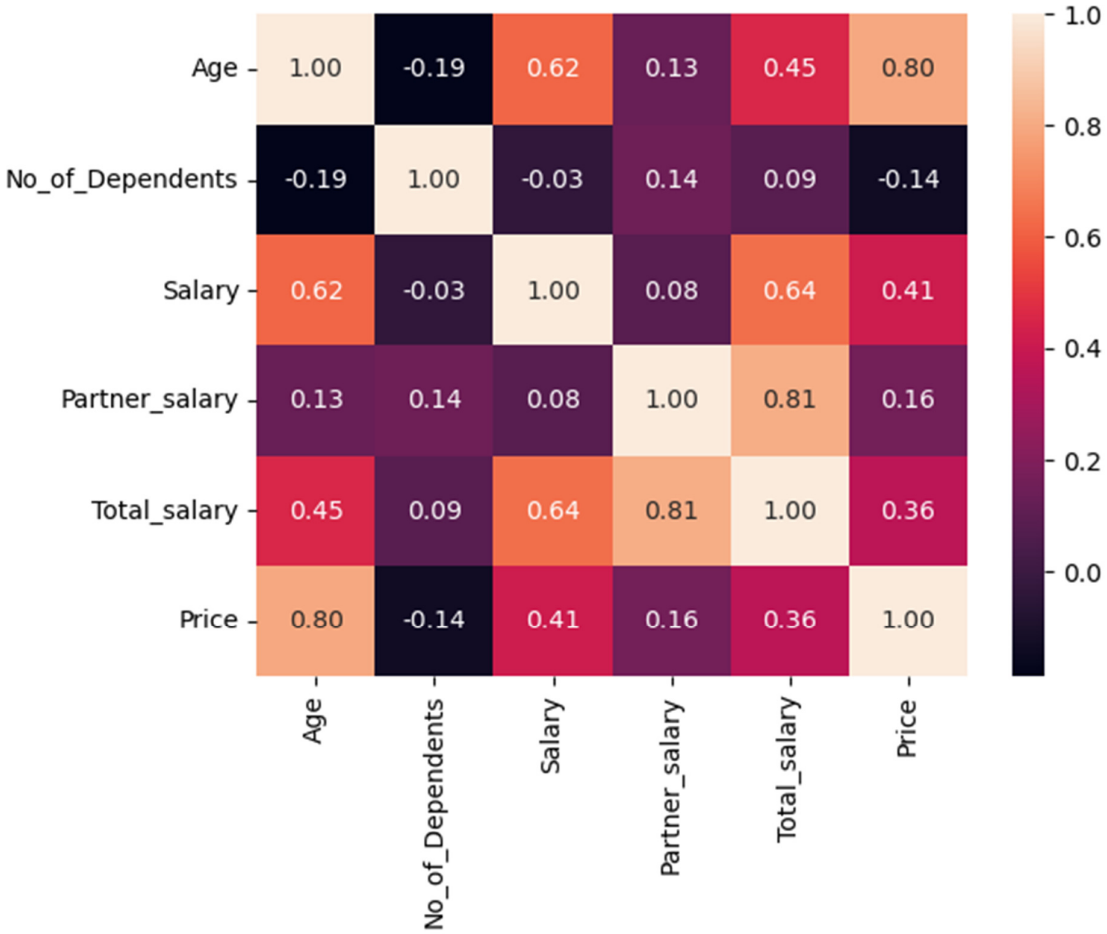
4. The number of male customers are more in comparison to female customers.
5. The number of Married customers are more in comparison to Single customers.
6. The number of customers having housing loan is less than number of customers than those are have a pre existing housing loan
7. The sales of Sedan is maximum closed follower by hatchback
8. The sales of SUV is very less in comparison to other makes.
9. Most of the customers are post-graduate.

Problem 1.3:

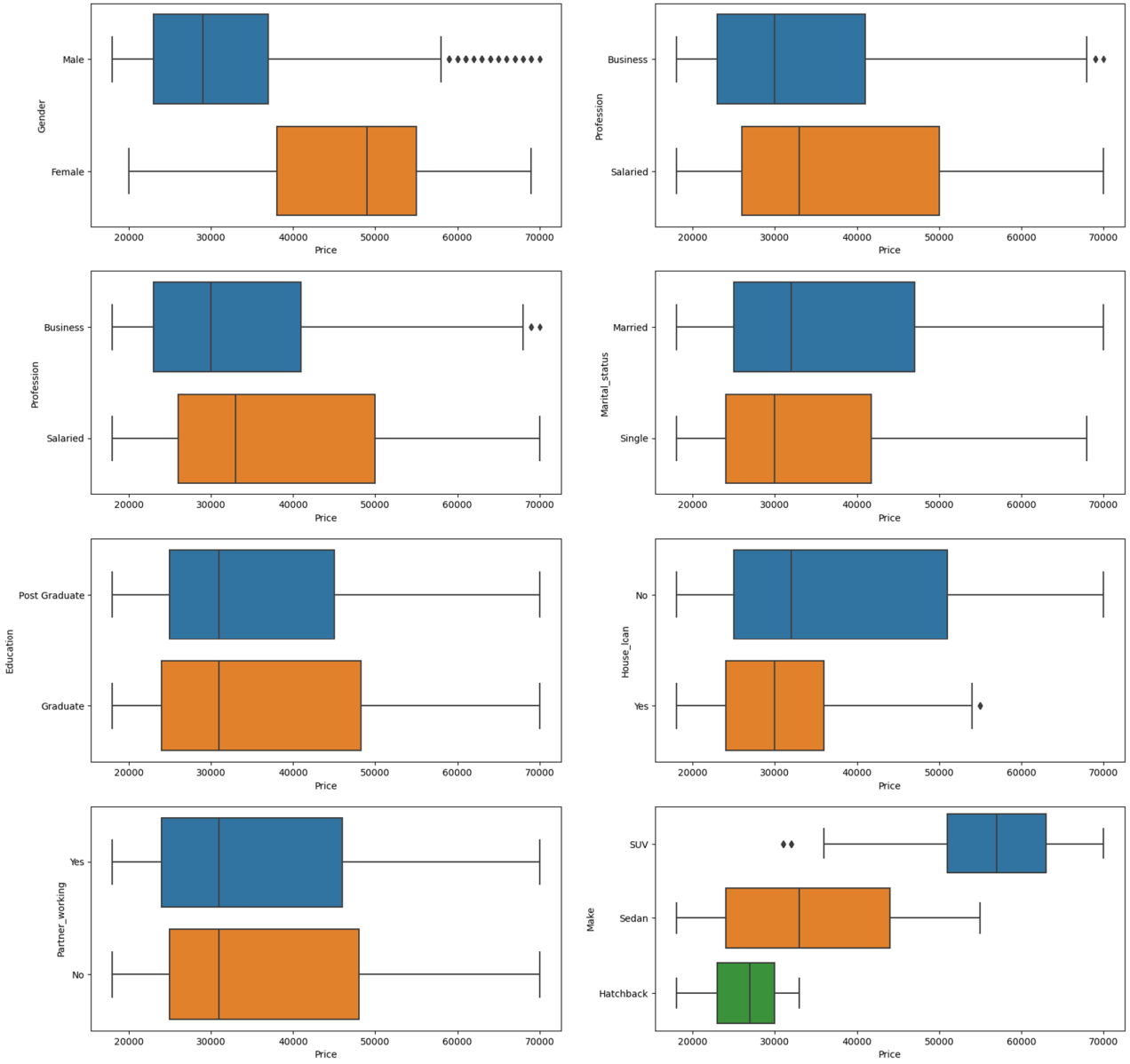
Problem 1 - Bivariate Analysis- Explore the relationship between all numerical variables - Explore the correlation between all numerical variables - Explore the relationship between categorical vs numerical variables

1. Explore the relationship between all numerical variables & Explore the correlation between all numerical variables




Age
No_of_Dependents
Salary
Partner_salary
Total_salary
Price
1.00 -0.19 0.62 0.13 0.45 0.80
-0.19 1.00 -0.03 0.14 0.09 -0.14
0.62 -0.03 1.00 0.08 0.64 0.41
0.13 0.14 0.08 1.00 0.81 0.16
0.45 0.09 0.64 0.81 1.00 0.36
0.80 -0.14 0.41 0.16 0.36 1.00
Age
No_of_Dependents
Salary
Partner_salary
Total_salary
Price
1.0
0.8
0.6
0.4
0.2
0.0

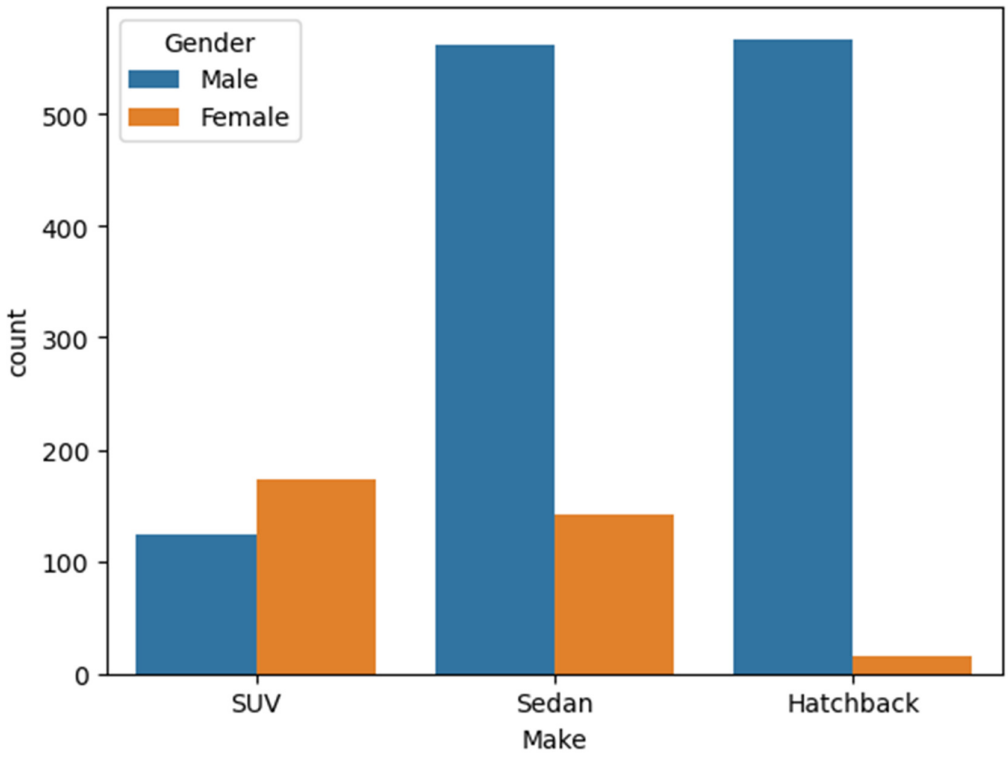
Explore the relationship between categorical vs numerical variables.



Problem 1.4 (a) (i) : Do men tend to prefer SUVs more compared to women?

Answer :

No, Female tends to purchase SUVs More than Men



Problem 1.3 (a) (ii) : What is the likelihood of a salaried person buying a Sedan?

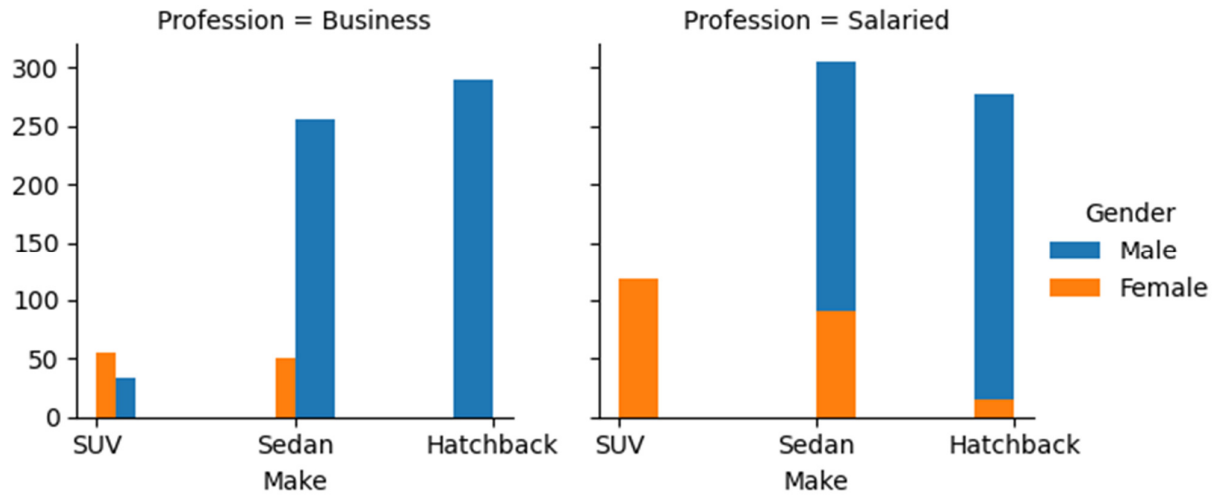
The Likelihood of salaries person buying a sedan is 56.41%.

```
likelihood of a salaried person buying a Sedan 56.41
```

Problem 1.4 (a) (iii) : 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?

Answer

As per Below Code Snippet there is no evidence to support that a salaried male is an easier target for a SUV sale over a Sedan sale.

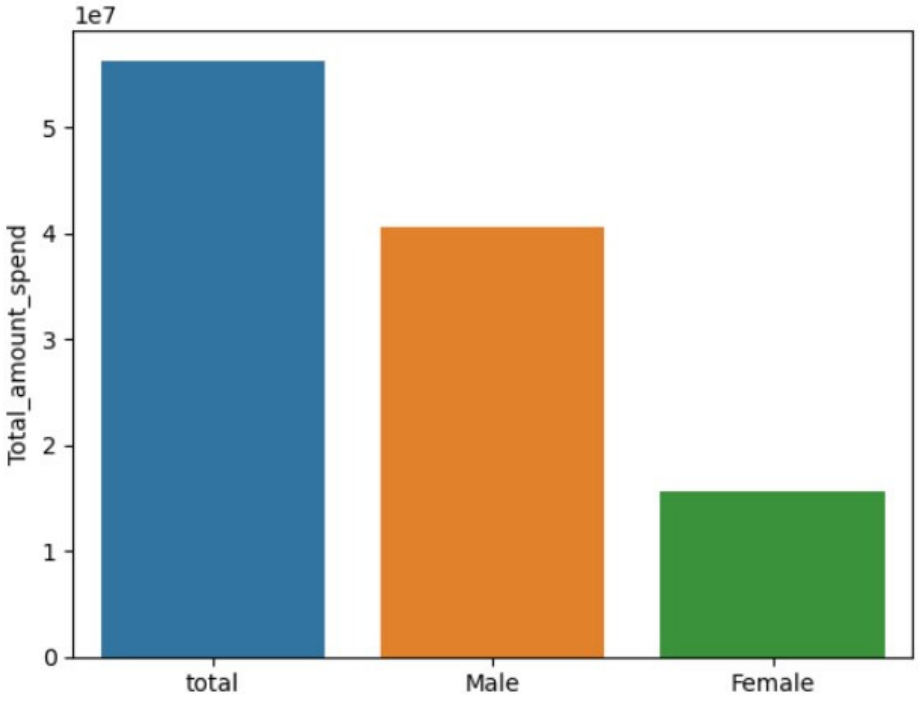


Problem 1.4 (a) (iv) : How does the amount spent on purchasing automobiles vary by gender?

The total amount spent by male is Rs. 4,05,85,000 and The total amount spend by female is Rs.1,56,95,000.

Hence the total amount spent by male is significantly greater than female.

```
Gender
Female    15695000
Male      40585000
Name: Price, dtype: int64
```



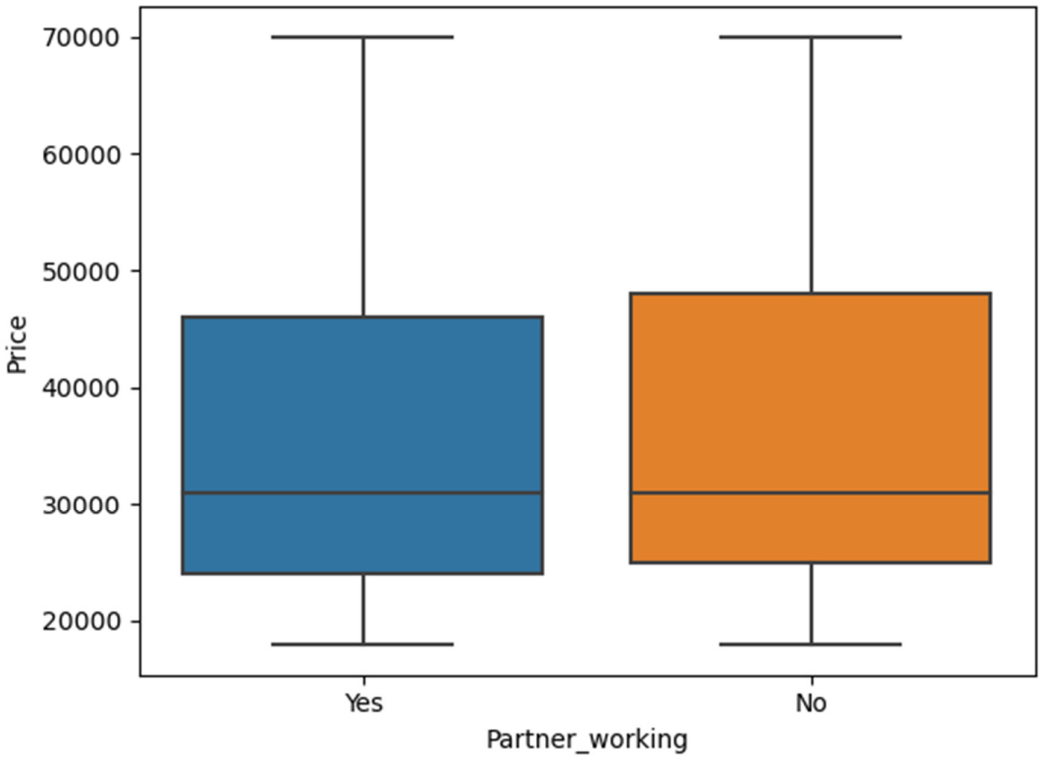
Problem 1.4 (a) (v) : How much money was spent on purchasing automobiles by individuals who took a personal loan?

The amount spent on purchasing automobiles by individuals who took a personal loan is Rs. 2,72,90,000/-

```
Personal_loan
No     28990000
Yes    27290000
Name: Price, dtype: int64
```

Problem 1.4 (a) (vi): How does having a working partner influence the purchase of higher-priced cars?

Answer: There is no influence on working partner on the purchase of higher-priced cars.



Problem 1.5

Based on your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective

Recommendations: -

1. The company can provide special discounts, offers, Cashbacks for graduates & persons of young age.
2. The company can provide special discounts, offers, Cashbacks for Single Persons, Females
3. The company can provide special discounts, offers, Cashbacks on make of Suvs
4. The company can improve its advertising campaigns to specifically target graduates, persons of young age, Single Persons, & Females
5. The company can improve its advertising campaigns to tell the quality of Suvs vs Other Makes
6. The company can tie up with banks to provide loans to business class.
7. The company can tie up with banks to provide loans to business class.
8. The company can tie up with banks to provide special concessions to persons having housing loans.

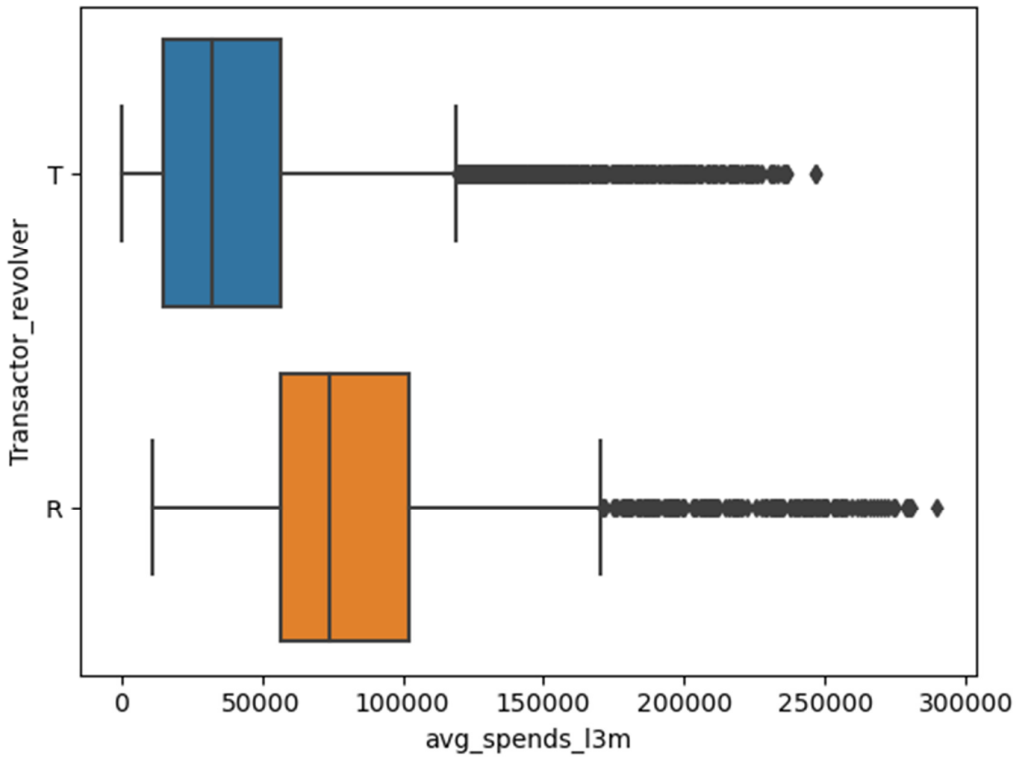
PROBLEM 2:

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

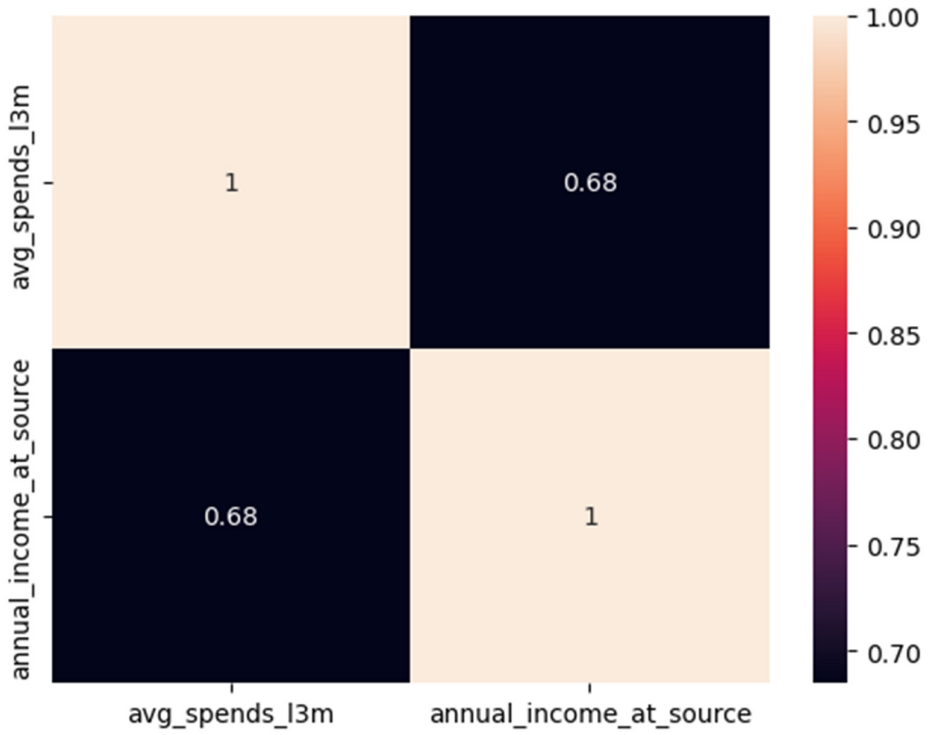
Top 5 Variables with Business Justifications

Top 5 Variables:-

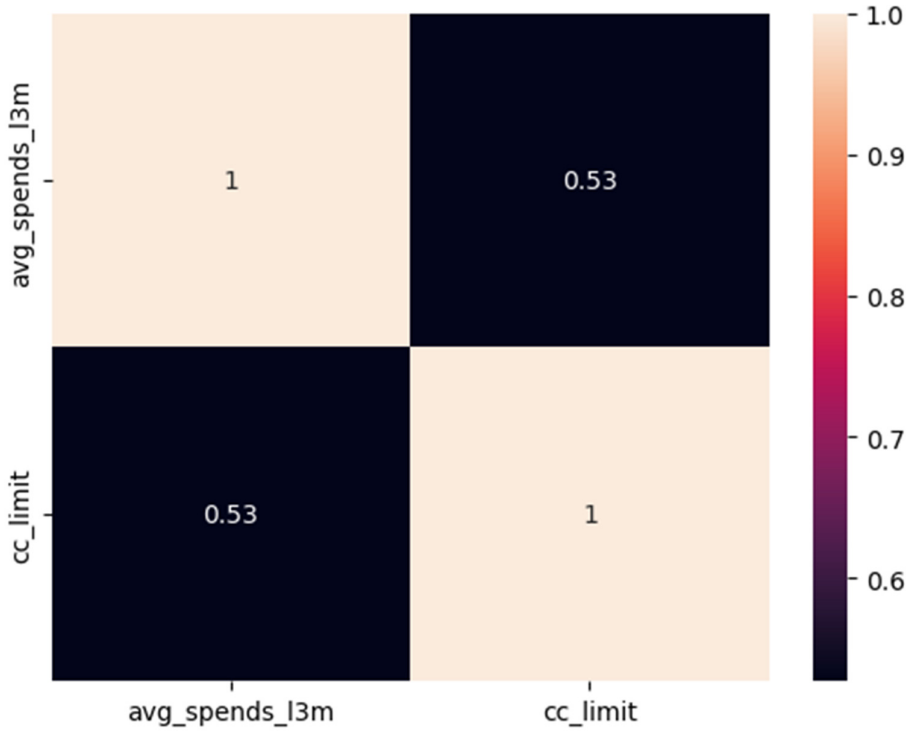
1. Transactor revolver : Whether the customer has revolved the balance next months helps us in retaining the customer



2. Average Spend in Last 3 Months
   If then amount of Average Spend in Last 3 Months is more there is more chances of retaining the customers.

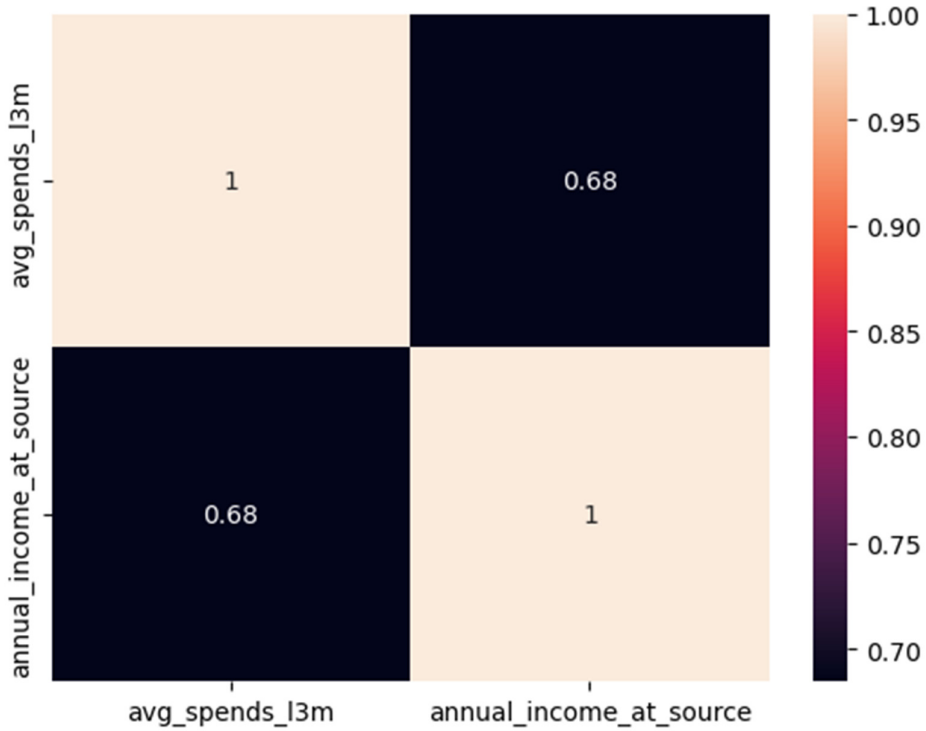


3. CC Limit
   If limit of CC is higher. Chances is customer will be retained



4. Annual Income at Source

If limit of Annual Income at Source is higher. Chances is customer will be retained



5. Hotlist Flag

As per Chart below if there is Hotlist flag on CC higher is limit.