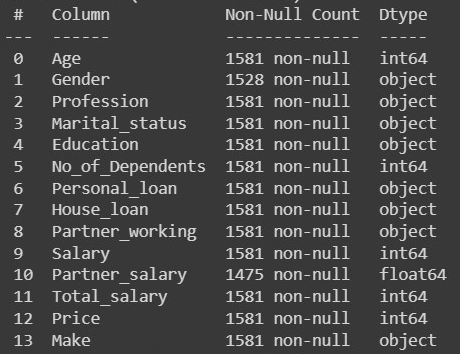
Report for problem 1



* The shape of dataset is , i.e. dataset contains columns and records.
* The size of dataset is , i.e. dataset has cells with values.
* The columns of dataset are as follows:

1. Age
2. Gender
3. Profession
4. Marital\_status
5. Education
6. No\_of\_Dependents
7. Personal\_loan
8. House\_loan
9. Partner\_working
10. Salary
11. Partner\_salary
12. Total\_salary
13. Price
14. Make

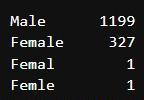
* Here we have some basic information about columns of dataset:



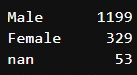
* Above, Column represents the name of columns in dataset. Non-Null Count represents the Number of records present in data and Dtype represents, what type of data is present in respective column.
* From Non-Null Count we can observe that “Gender” column has values and “Partner\_Salary” column has values and all other column has values, i.e. values are missing from “Gender” column and values are missing from “Partner\_Salary” column.
* The Dtype ‘int64’ shows that respective columns have integers values whereas ‘float64’ shows that respective column has numbers with decimal values. Dtype ‘object’ shows that respective column has information in the form of text.



* There are no duplicated fields in dataset.
* There are some null values i.e. some missing values which is mentioned above.
* In “Gender” column, there is majorly two categories ‘Male’ and ‘Female’, but due to spelling mistakes and missing values (nan), it has four categories.



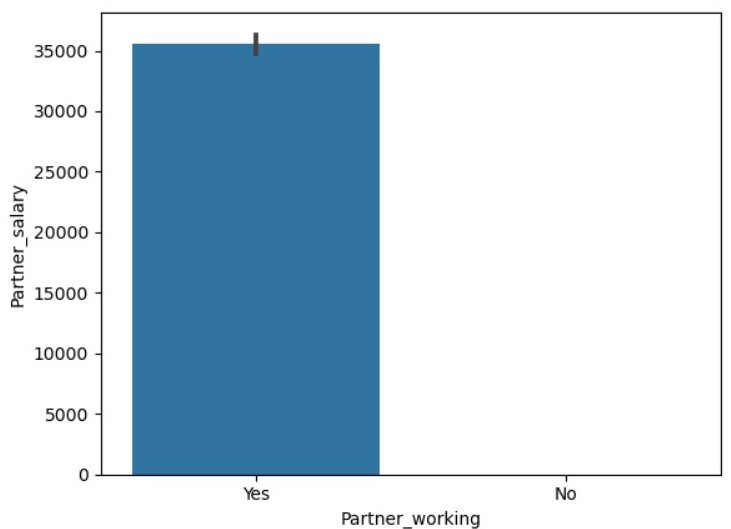
* To clean this, we will replace ‘Femal’ , ‘Femle’, to ‘Female’



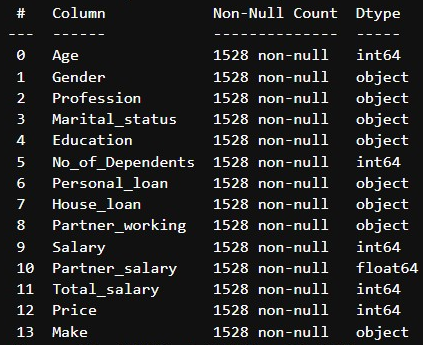
* In “Gender” column, there are missing values.
* In dataset there are rows of data, and “Gender” column has missing values so tackle with this discrepancy we drop those rows with missing genders. So finally

dataset will have rows of data left.

* So after cleaning the “Gender” column, the shape of dataset is .
* In “Partner\_salary” column out of rows of data values are missing. Initially in dataset there were values of “Partner\_salary” were missing but since we drop the rows with missing value in “Gender” column the rows that were commonly had missing values in both “Partner\_salary” and “Gender” columns was dropped that’s why missing values in “Partner\_salary” column.
* We have, “Total\_salary” column and “Salary” column.
* “Total\_salary” is sum of “Salary” and “Partner\_Salary” and there is no missing values in “Total\_salary” and “Salary” columns.
* So we can fill the missing values in “Partner\_Salary” columns with the help of “Total\_salary” and “Salary” columns.
* “Partner\_salary” = “Total\_salary” – “salary”



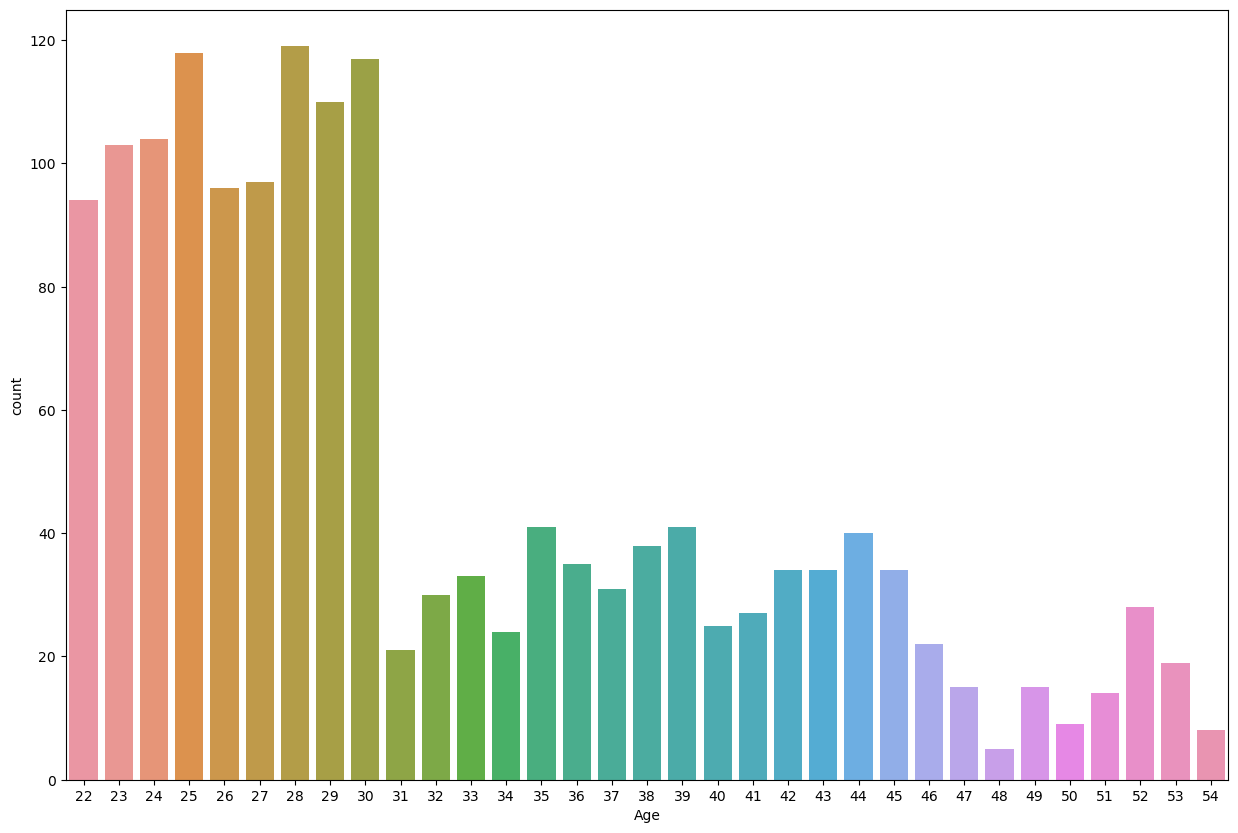
* From graph it can be concluded that, if the Partner is not working that the Partner salary is 0.
* And if Partner is working than Partner salary is total salary minus the salary.



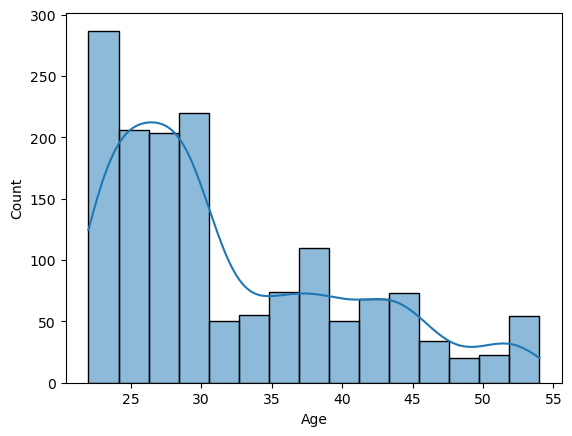
* All the missing values is removed.
* Here is the basic statistical information for those columns that consist of numerical data.



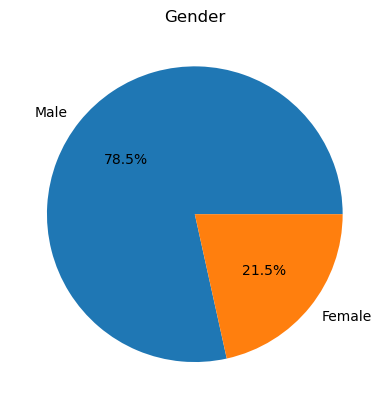
* Here is count-plot for “Age” column, it is numerical column.



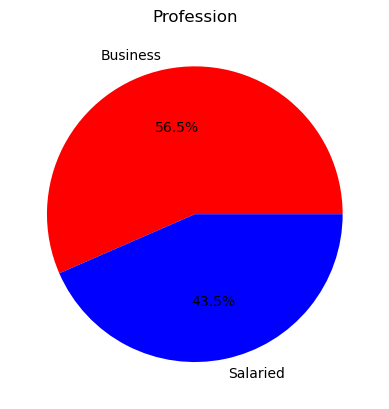
* Max age is 54
* Min age is 22
* Average age from dataset is 32
* More than 75% of data consist of information about age group less than 38
* From this it can be concluded that targeted audience should be one with age less than or equal to 38-40.
* Here is histogram plot of “Age”



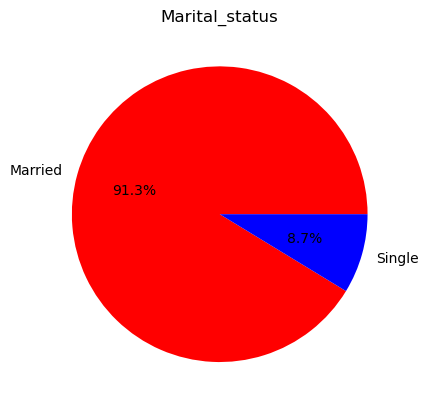
* The data in “Age” column is negatively skewed.
* Here is a pie chart of “Gender” column, it is a categorical column.



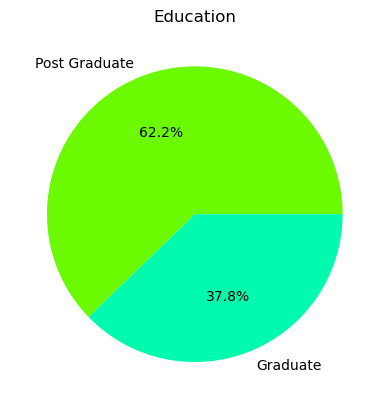
* In given dataset we have 78.5% male and 21.5% female.
* Males are supposed to buy vehicles more than women.
* Here is pie chart for “Profession” column, it is categorical column



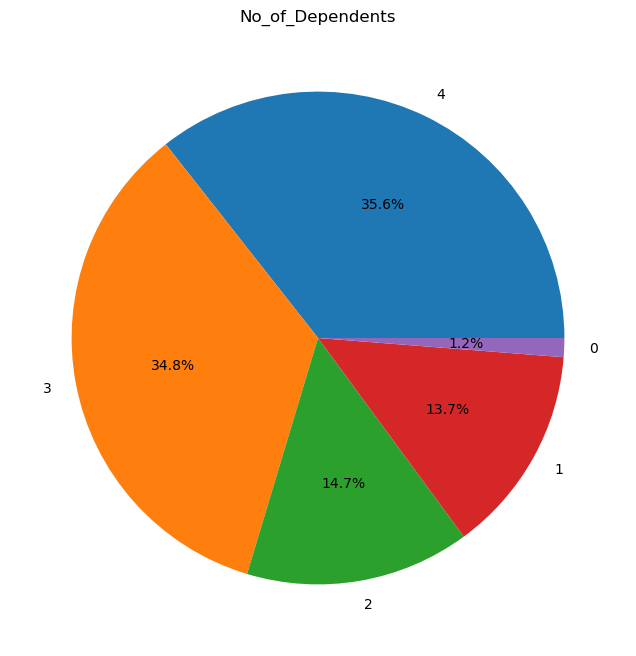
* About 56.5% of people have Business as their profession.
* About 43.5% of people are Salaries.
* There is not huge difference in buying vehicle decision based on profession.
* Here is pie chart for “Marital\_status” column, it is categorical column.



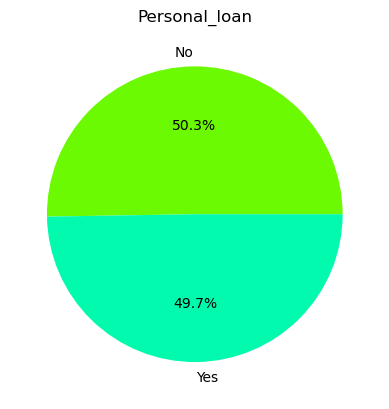
* 91.3% of peoples in dataset are married
* 8.7% people are single.
* Married people tends to buy vehicle more than single peoples.
* Here is pie chart of “Education” column, it is categorical column



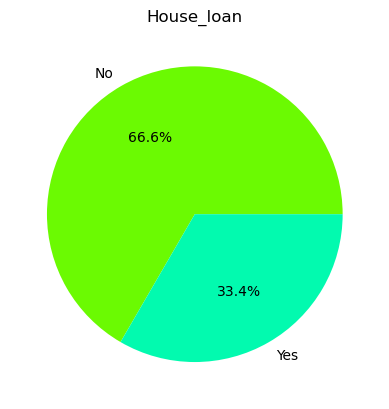
* 62.2% of people are Post graduate
* 37.8% of people are Graduate
* The people with higher education qualifications are tends to buy vehicle more than lower ones.
* Here is pie chart of “No\_of\_dependents”, it is categorical column with numeric data.



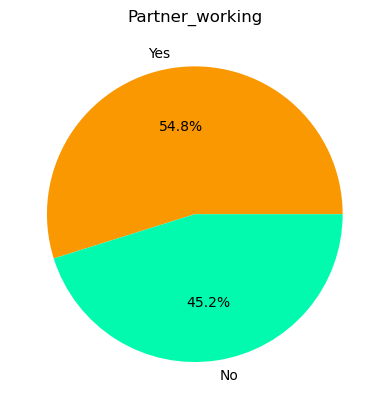
* There are 35.6% of people with 4 dependents on them.
* There are 34.8% of people with 3 dependents on them.
* There are 14.7% of people with 2 dependents on them.
* There are 13.7% of people with 1 dependent on them.
* There are 1.2% of people with no dependents on them.
* Here is pie chart for “Personal\_loan” column, it is categorical column.



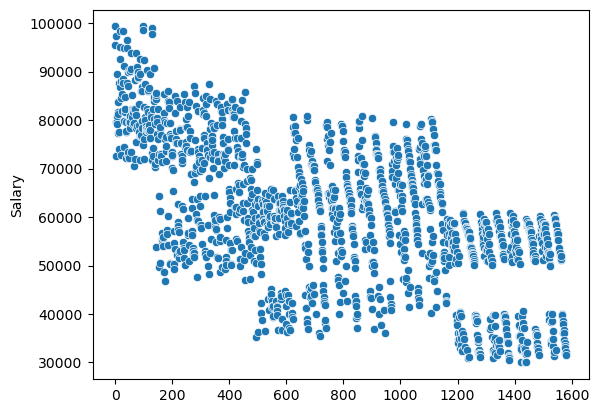
* 49.7% of people are having personal loan.
* 50.3% of people are with no loans.
* Here is pie chart for “House\_loan” column, it is categorical column.



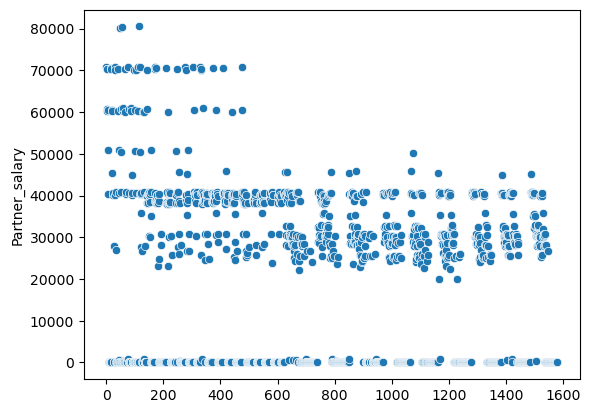
* 66.6% of people do not have House loan.
* 33.4% of people do have house loan.
* Here is pie chart for “Partner\_working” column, it is categorical column



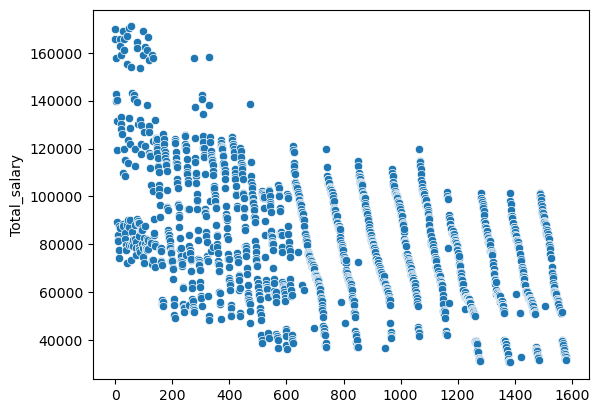
* About 54.8% people have working partners.
* 45.4% of people do not have working partners.
* Here is scatterplot of “Salary” column, it is numerical column



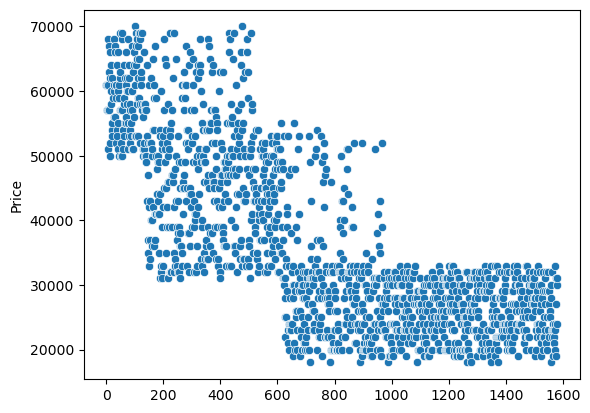
* On X-axis we have bin size of 200, which shows number of peoples.
* On Y-axis we have Salary information.
* Max salary in dataset is 99300
* Min salary in dataset is 30000
* Average salary is 60453.40
* 75% of people in dataset has salary less than equal to 72000.
* Here is scatterplot for “Partner\_salary”, it is numerical column.



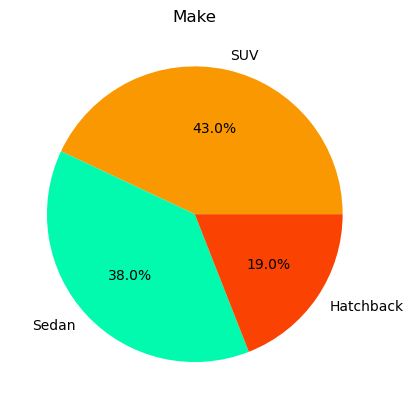
* On X-axis we have bin size of 200, which shows number of peoples.
* On Y-axis we have Partner\_salary information.
* Max Partner\_salary is 85000
* Min Partner\_salary is 0.
* Average Partner\_salary is 19516.36.
* Here is scatterplot for “Total\_salary” column. It is numerical column



* On X-axis we have bin size of 200, which shows number of peoples.
* On Y-axis we have Total\_salary information.
* Max Total\_salary is 171000.
* Min Total\_salary is 30600.
* Average Total\_salary is 79969
* 75% of people in dataset has Total\_salary less than equal to 79969.
* Here is a scatterplot of “Price” column, it is numerical column.



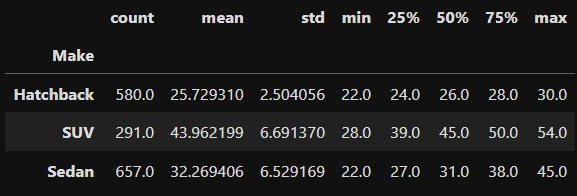
* On X-axis we have bin size of 200, which shows number of peoples.
* On Y-axis we have Price information.
* Max Price is 70000.
* Min Price is 18000.
* Average Price is 36022.
* Here is pie chart of “Make” column, it is categorical column

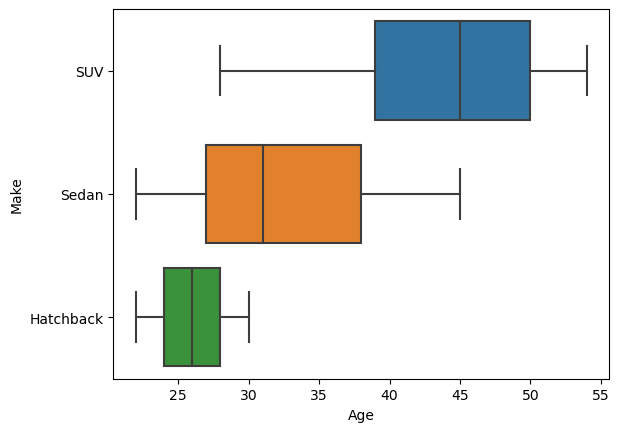


* 43.0% can make SUV
* 38.0% can make Sedan
* 19.0% can make Hatchback

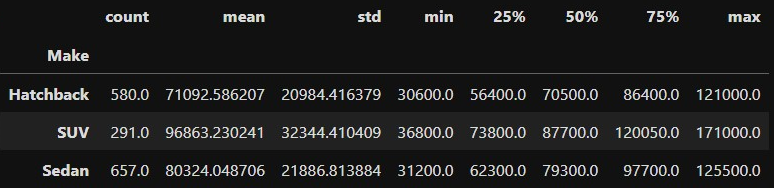
D)

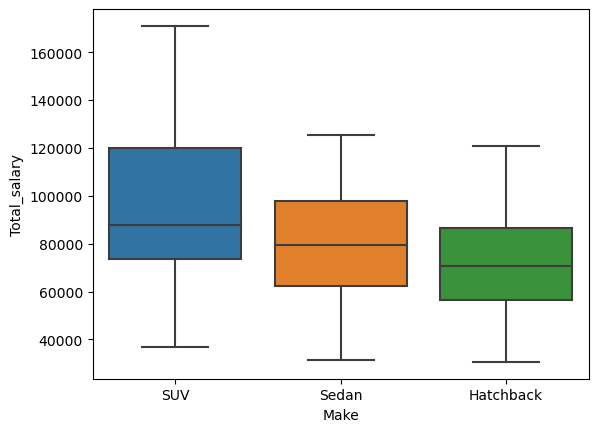
* Relationship between “Age” and “Make”



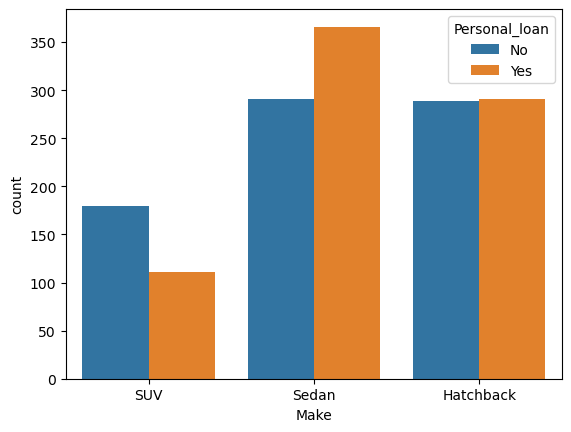


* Mean age of people that can be target customer for ‘Hatchback’ is 25-26.
* Mean age of people that can be target customer for ‘SUV’ is 43-44.
* Mean age of people that can be target customer for ‘Sedan’ is 32-33.
* Younger people are interested in ‘Hatchback’.
* As the age is increasing the interest is shifting towards ‘SUV’.
* Relationship between “Make” and “Total\_salary”

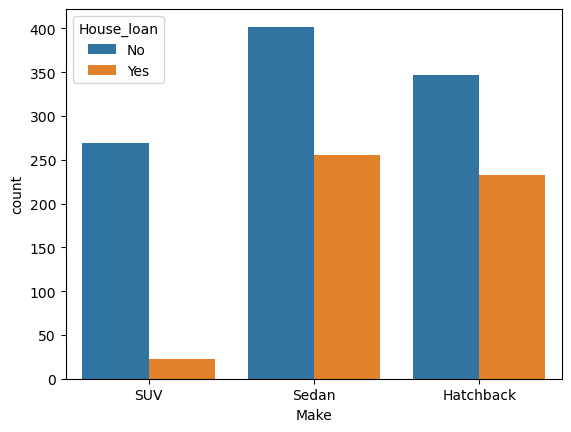




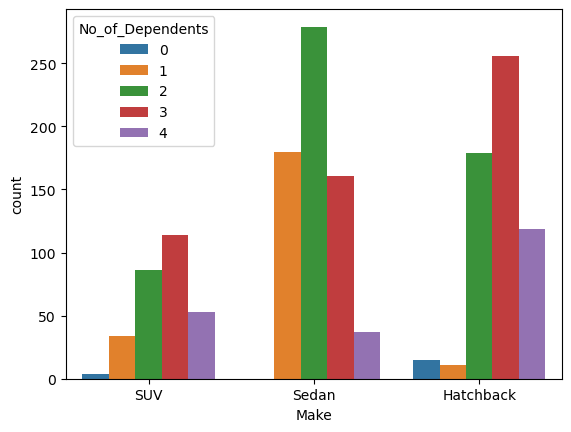
* If the median total\_salary is around 87700 than there is more chance that customer shows more interest in SUV.
* If the median total\_salary is around 79300 than, people can be targeted for ‘Sedan’.
* If the mdian total\_salary is around 70500 than he/she is more interested in ‘Hatchback’.
* Relationship between “Personal\_loan” and “Make”



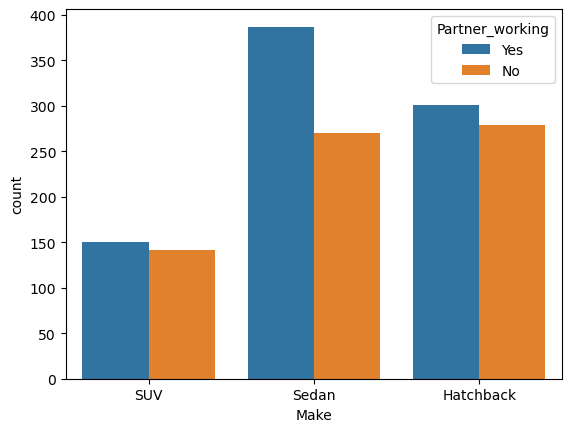
* If the person having personal loan there is not much affect on buying the vehicle.
* When talking about ‘Sedan’ and ‘Hetchback’, people with personal loan are tends to buy slightly more than with no personal loan.
* But the case is inverse for ‘SUV’.
* Relationship between “House\_loan” and “Make”



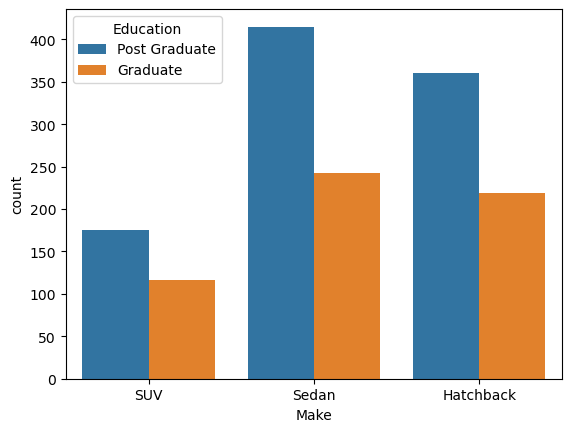
* The people with house loan buys the vehicle less than that one who do not have house loan.
* In case of ‘SUV’ if the person have house loan the chance that he will buy ‘SUV’ is very less.
* Relationship between “Number\_of\_Dependents” and “Make”



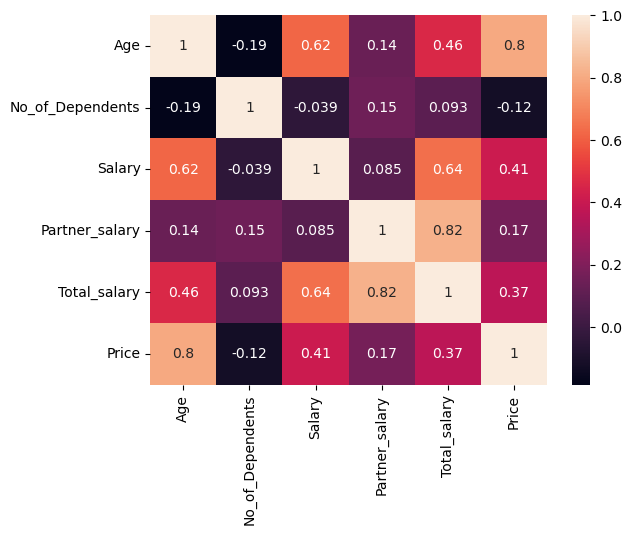
* Relationship between “Partner\_working” and “Make”



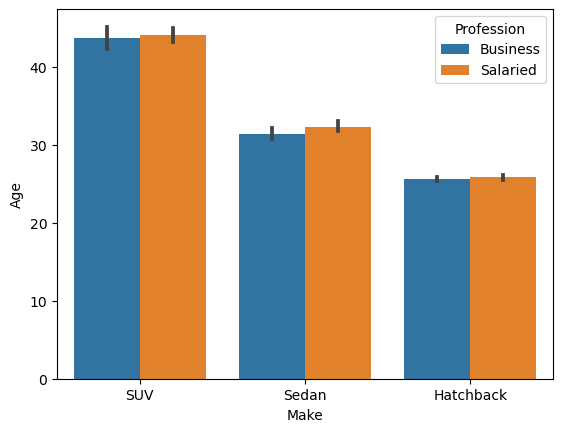
* If the partner is working than there is comparatively more chance that person buys vehicle.
* Relationship between “Education” and “Make”



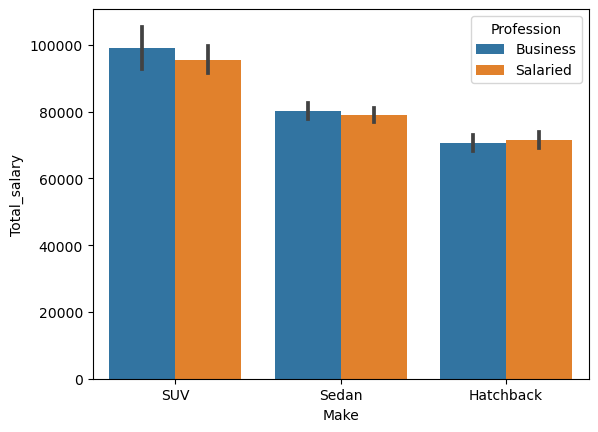
* Correlation between numerical columns.



* Relationship between “Make”,”Age” and “Proffession”

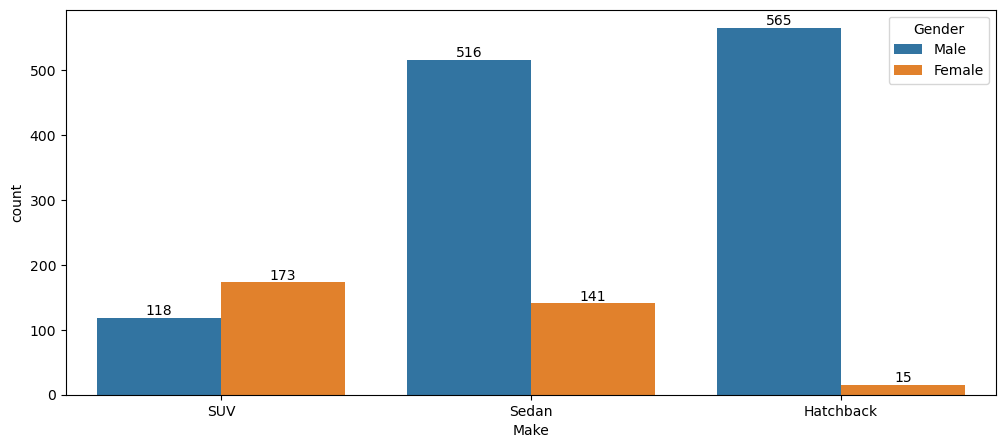


* Relationship between “Make”,”Total\_salary” and “Proffesion”

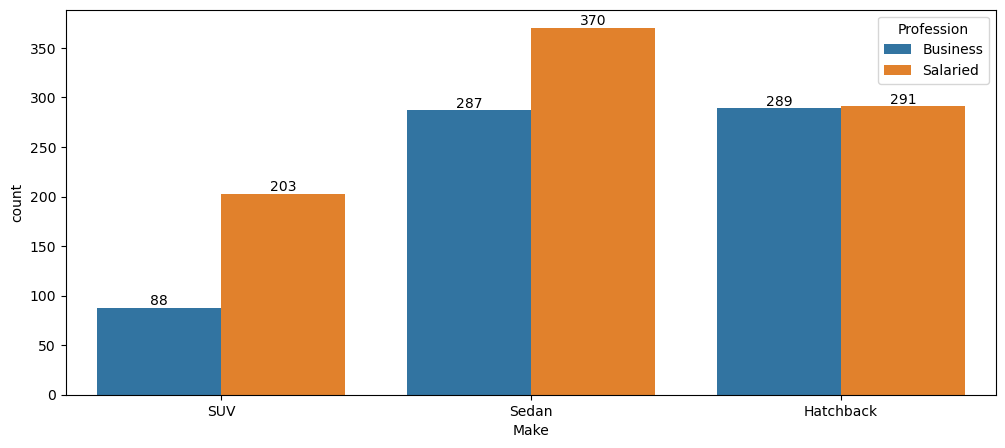


E)

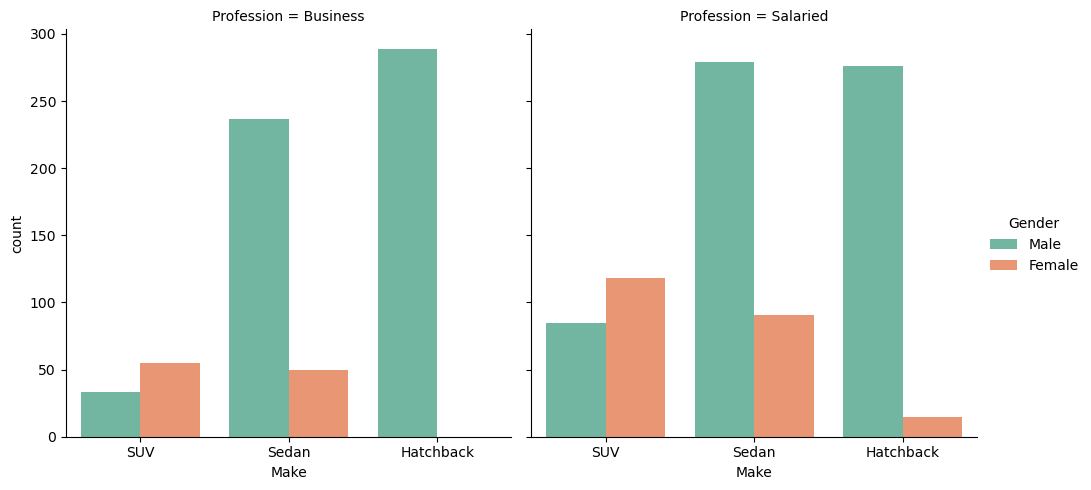
* E1]



* I disagree with Steve Rogers, from the data it can be concluded women prefer SUV more than man. But the difference is not much.
* E2]



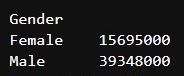
* I agree with Ned Stark salaried person is more likely to buy a Sedan
* E3]



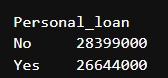
* I disagree with Sheldon Cooper.

F)

* F1]

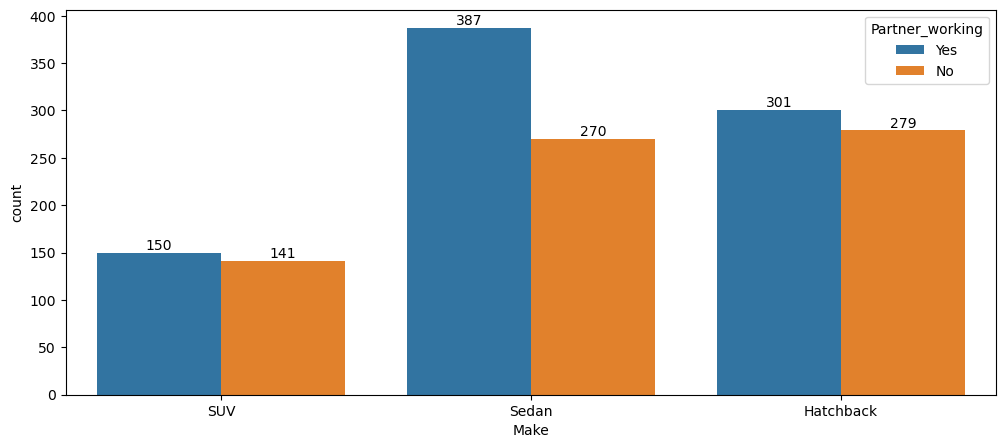


* Grouping the dataset with respect to gender and then summing up the price we get.
* Female spends total: 15695000
* Male spends total: 39348000
* Total amount spend by male is more than total amount spend by female for purchasing automobiles.
* Business can focus on Male customer more as their target audience.
* F2]



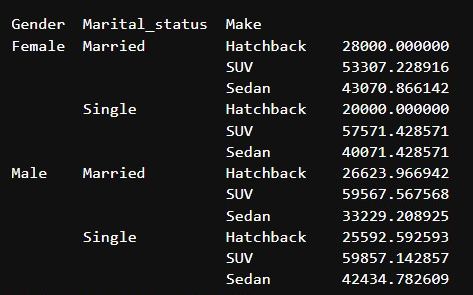
* Grouping the dataset with respect to personal loan and then summing up the price we get.
* People with personal loan spends total: 26644000
* People with no personal loan spends total: 28399000
* Total amount spend by person with no personal loan is more than one with personal loans.

G)



* From above graph we can conclude that having working partner leads to purchase of higher-priced car.

H)



* Here we have calculated average of each category.
* The average spending on purchasing ‘Sedan’ by Married Female is same as average spending on purchasing ‘Sedan’ by Single Male.
* For Married Male and Single Male the average spending on purchasing ‘SUV’ is almost the same.

Report for problem 2

* The shape of dataset is (8448, 28), i.e. 28 columns and 8448 rows of data.
* The Top 5 important variables are-

1. Annual Income

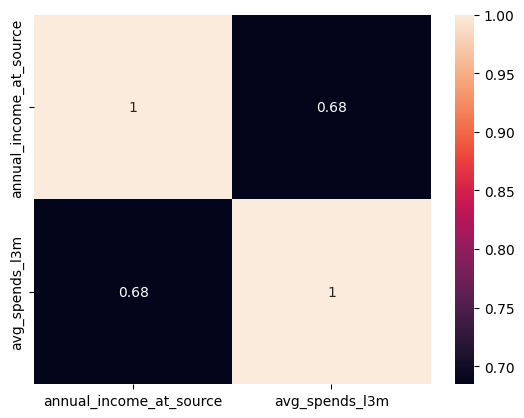
2. Transactor & Revolver

3. avg\_spends\_l3m

4. Occupation\_at\_source

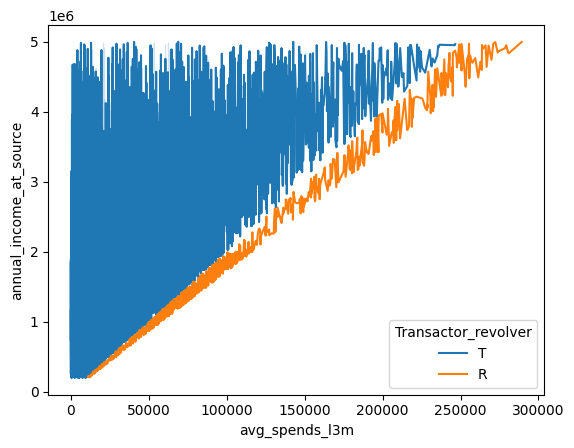
5. CC limit





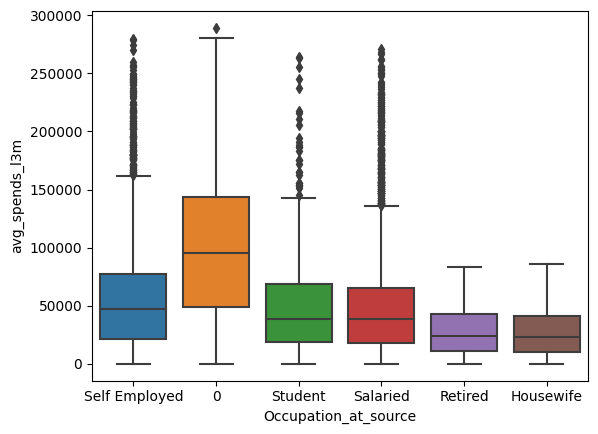
(correlation heatmap between avg\_spends\_l3m and annual\_income\_at\_source)

* There is positive correlation between avg\_spends\_l3m and annual\_income\_at\_source.
* More the annual income more the spending by credit card.



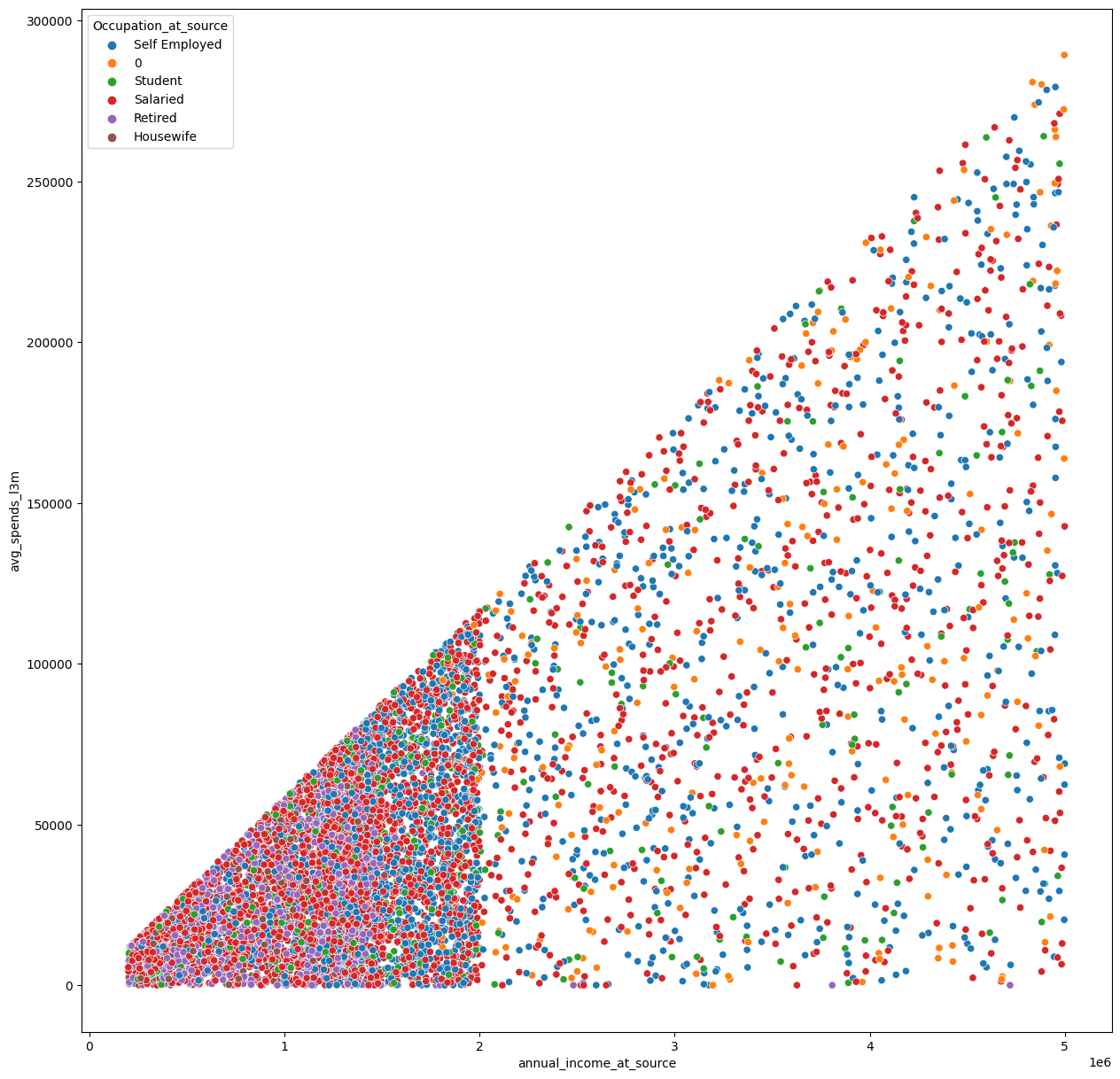
(line-plot with avg\_l3m and annual\_income\_at\_source with hue Transactot\_revolver)

* We can observe that Transactor are more in case of low average salary while revolver are also seen in case of people who spend more and their annual income is also high.
* Company should focus more on such people that why they are carrying balances over from one month to the next.



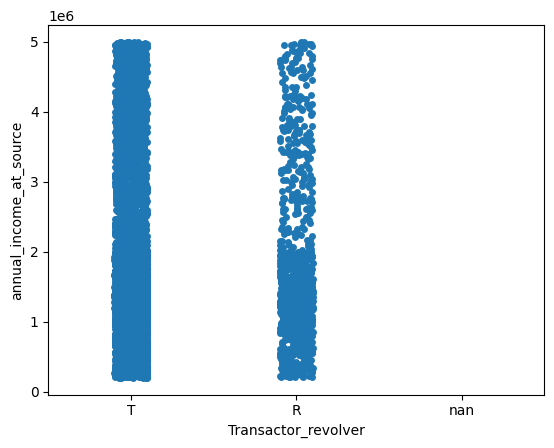
(Boxplot between Occupation\_at\_source and avg\_spends\_l3m)

* We can clearly observe that there are many outlier for ‘Salaried’, ’Self Employment’, ‘Student’. These people are spending more. They should be targeted as customer for credit card.



(Scatter plot for annual\_income\_at\_source and avg\_spends\_l3m with hue Occupation\_at\_source)

* As the annual income is increasing the average spending is also increasing.
* Average spend by salaried person is more.
* Average is spending is more clustered but amount spend is comparatively less for people with lower annual income.
* Most of them are Salaried person and Self\_employed.



(Strip plot between Transactor\_revolver and annual\_income\_at\_source)

* Revolvers are more in case of low income group there are only a few revolvers at high income group.
* Bank should assign maximum limit on credit cards of low income group people so that there is less attrition.