Credit EDA Assignment
Case Study

By: Thulasiram Saravanan

Initial dropping of columns

- The Columns in the image have rows which has “null values greater than 50%”.
- We are Dropping these columns since these will be creating wrong correlation with the target column.

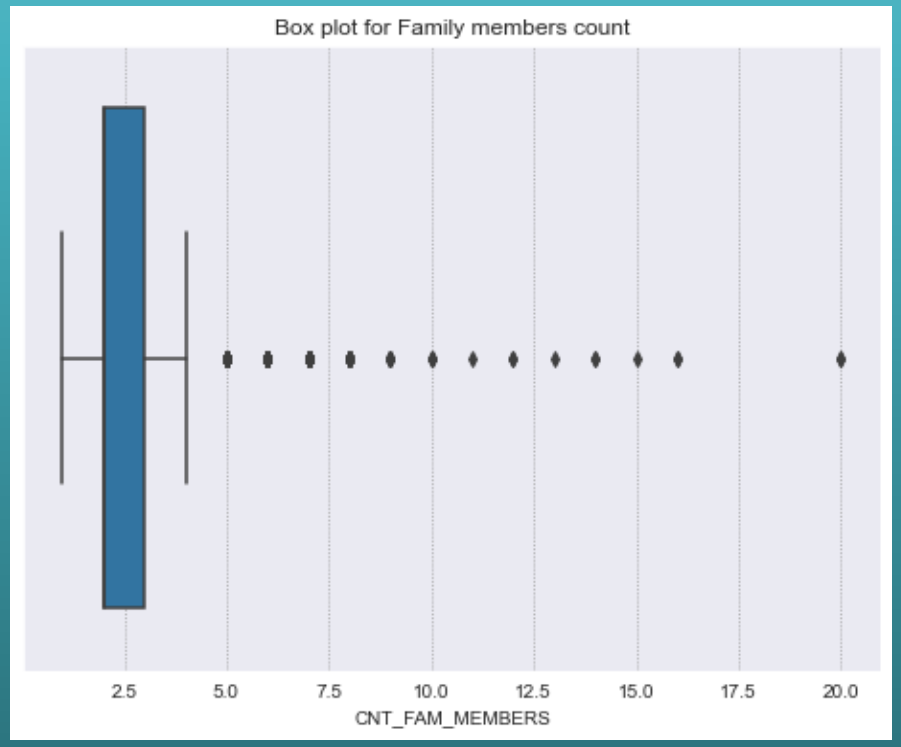
OWN_CAR_AGE	202929
EXT_SOURCE_1	173378
APARTMENTS_AVG	156061
BASEMENTAREA_AVG	179943
YEARS_BUILD_AVG	204488
COMMONAREA_AVG	214865
ELEVATORS_AVG	163891
ENTRANCES_AVG	154828
FLOORSMIN_AVG	208642
LANDAREA_AVG	182590
LIVINGAPARTMENTS_AVG	210199
LIVINGAREA_AVG	154350
NONLIVINGAPARTMENTS_AVG	213514
NONLIVINGAREA_AVG	169682
APARTMENTS_MODE	156061
BASEMENTAREA_MODE	179943
YEARS_BUILD_MODE	204488
COMMONAREA_MODE	214865
ELEVATORS_MODE	163891
ENTRANCES_MODE	154828
FLOORSMIN_MODE	208642
LANDAREA_MODE	182590
LIVINGAPARTMENTS_MODE	210199
LIVINGAREA_MODE	154350
NONLIVINGAPARTMENTS_MODE	213514
NONLIVINGAREA_MODE	169682
APARTMENTS_MEDI	156061
BASEMENTAREA_MEDI	179943
YEARS_BUILD_MEDI	204488
COMMONAREA_MEDI	214865
ELEVATORS_MEDI	163891
ENTRANCES_MEDI	154828
FLOORSMIN_MEDI	208642
LANDAREA_MEDI	182590
LIVINGAPARTMENTS_MEDI	210199
LIVINGAREA_MEDI	154350
NONLIVINGAPARTMENTS_MEDI	213514
NONLIVINGAREA_MEDI	169682
FONDKAPREMONT_MODE	210295
HOUSETYPE_MODE	154297
WALLSMATERIAL_MODE	156341
dtype: int64	

Outlier Detection

Family members count

Inference

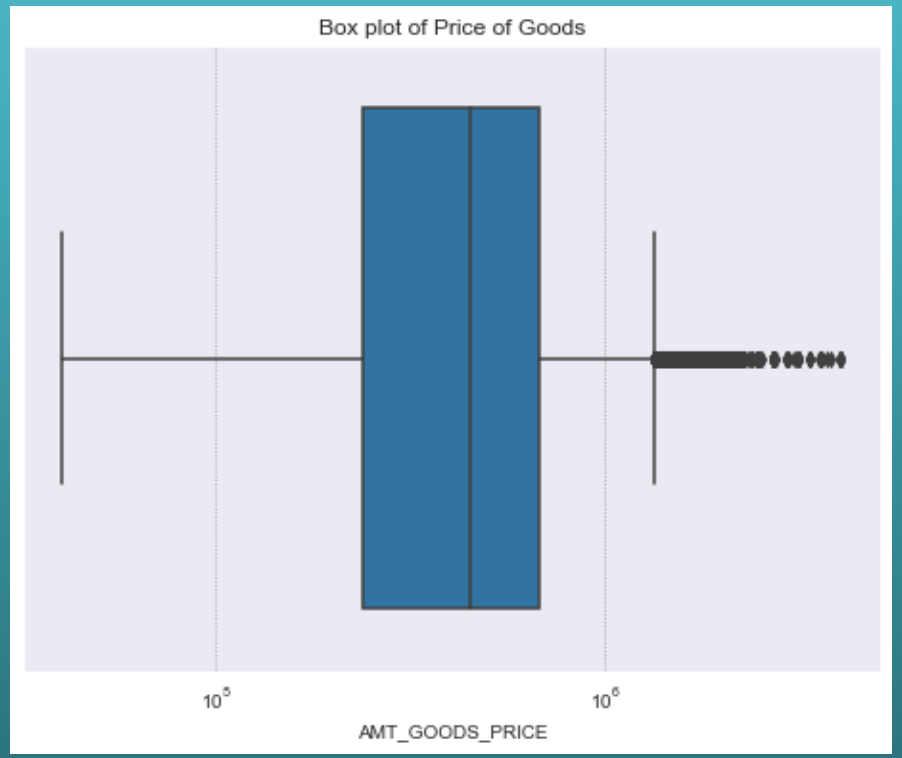
- From the above plot we could see the majority count of family members is less than 4 or 5. The count above 5 are like the joint families where in todays world its very rare to see such families and hence this is a valid case that there are outliers.



Price of Goods

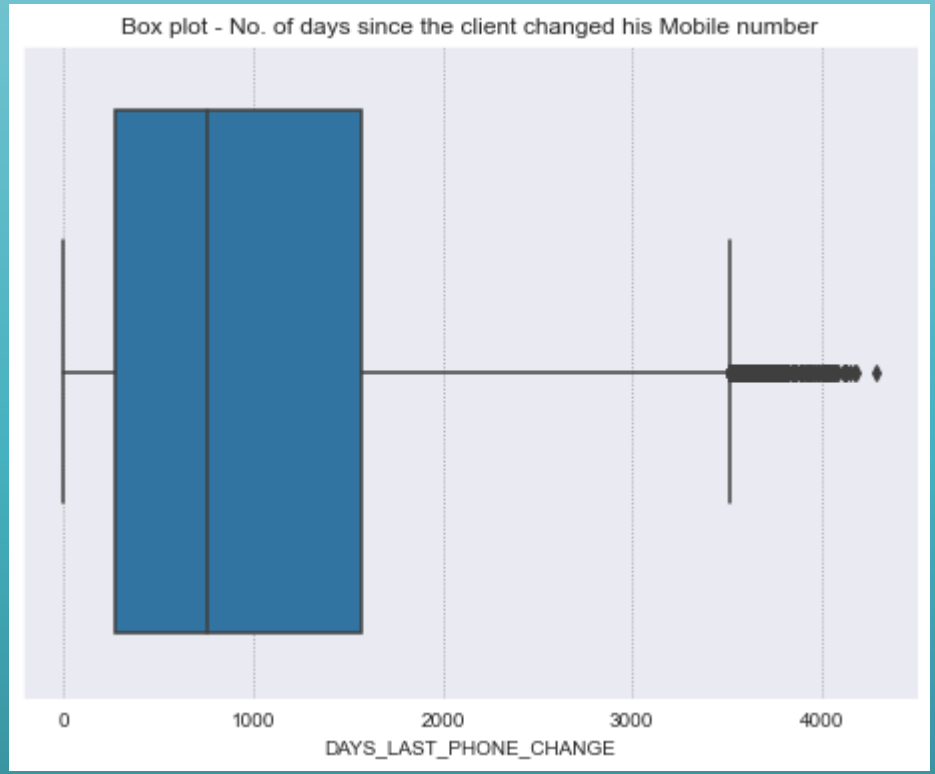
Inference

- There are many points more than range of 12 Lakhs, But the median lies around 5 to 6 Lakhs. Reason is many people have applied for goods ranging less than 6 Lakhs only.



No. of days since the client changed his Mobile number

- Initially this column were of negative values which was understandable because it tells us that before so many days the number was changed and that negative –ve symbol supports before
- For calculation purpose this is changed as positive variable



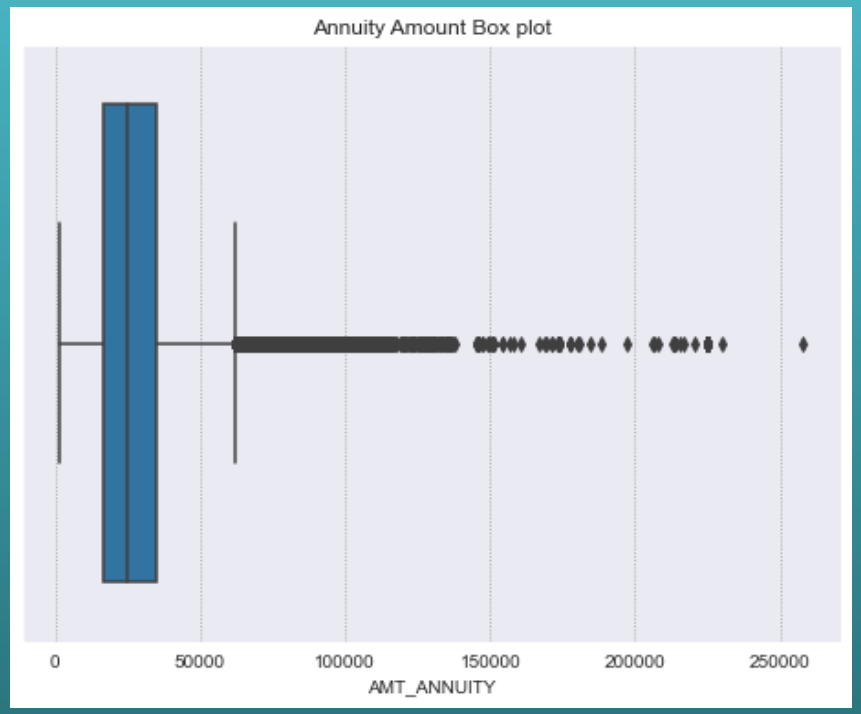
Inference:

- This conveys that for almost like past 4 to 5 years most of the people are using the same number and the outliers are due to the reason that only few would have not changed the mobile in their life.

Annuity Amount

Inference

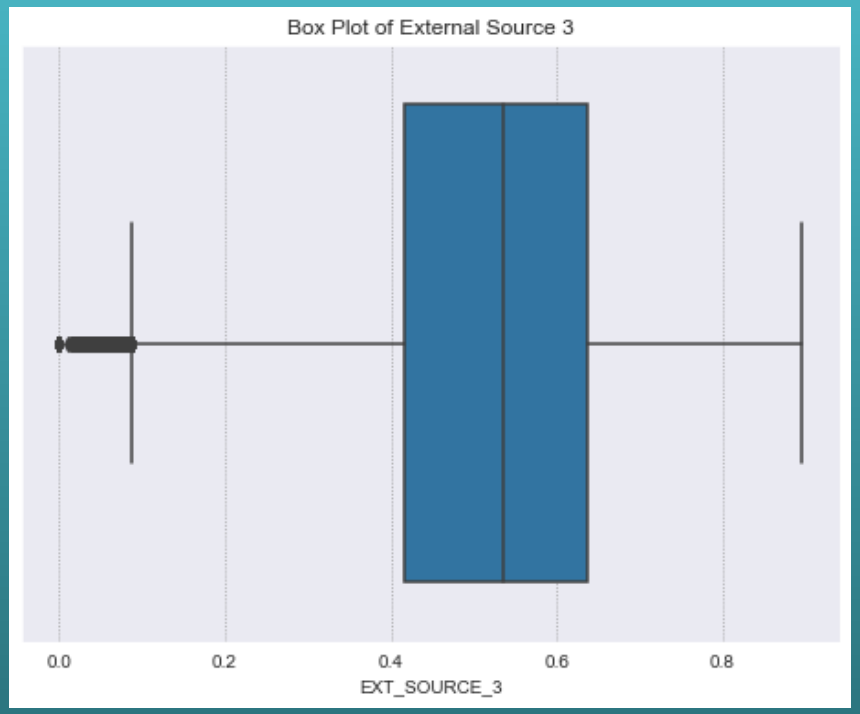
- The Annuity seems to have almost like 25000 as median. there are many outliers for the sole reason that most of the lower class and middle class people will be able to pay only that much as annuity and only few people in the list would have capacity to pay more. So its a unusual case but still valid one.



Price of Goods

Inference:

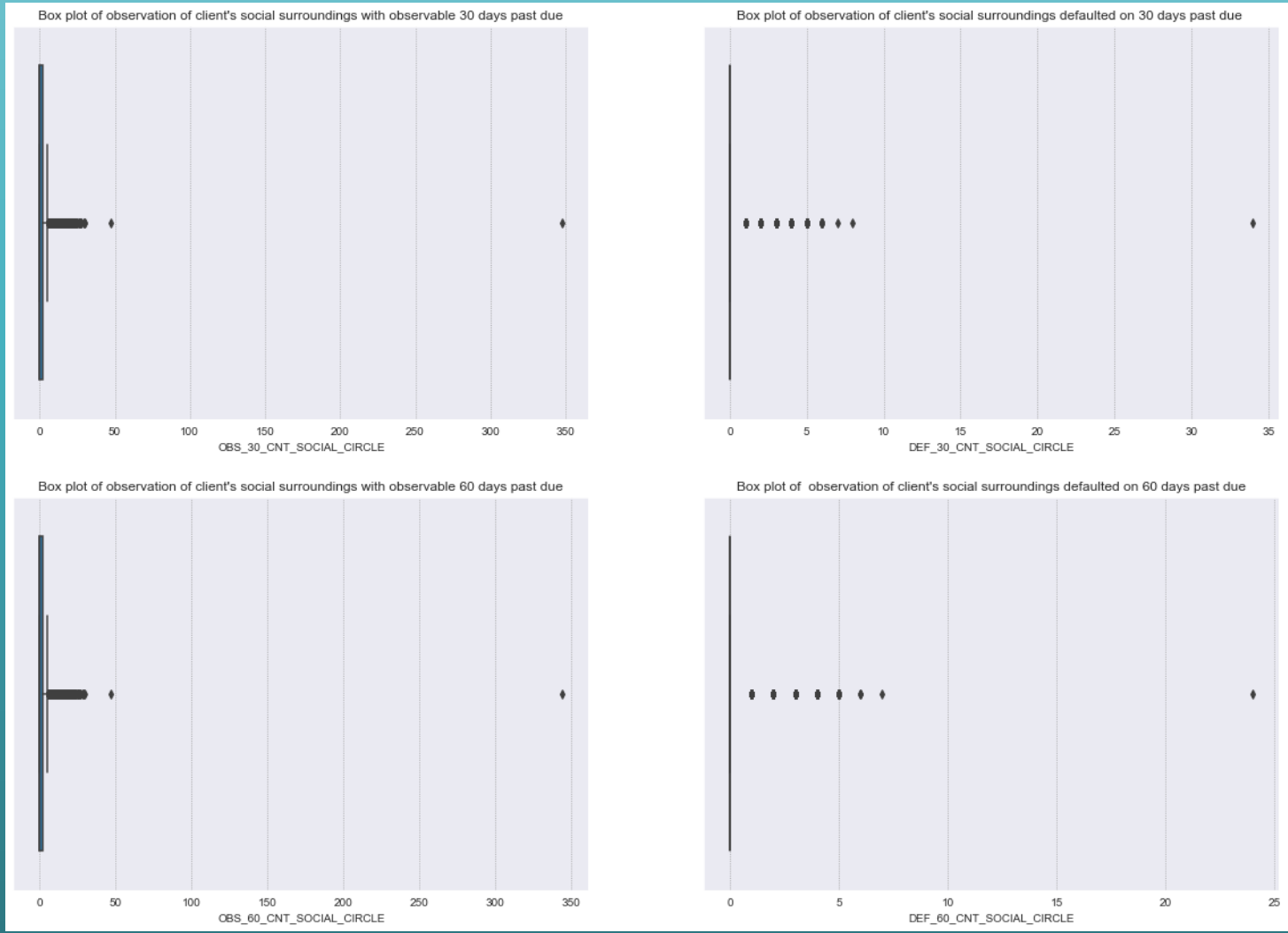
- Only of the external data sources would have given reviews as low and that's the reason it was treated as outlier.



Client's social surroundings observation

Inference:

- There is no proper evidence to use the column since the median of all the 4 are almost or very near to 0.

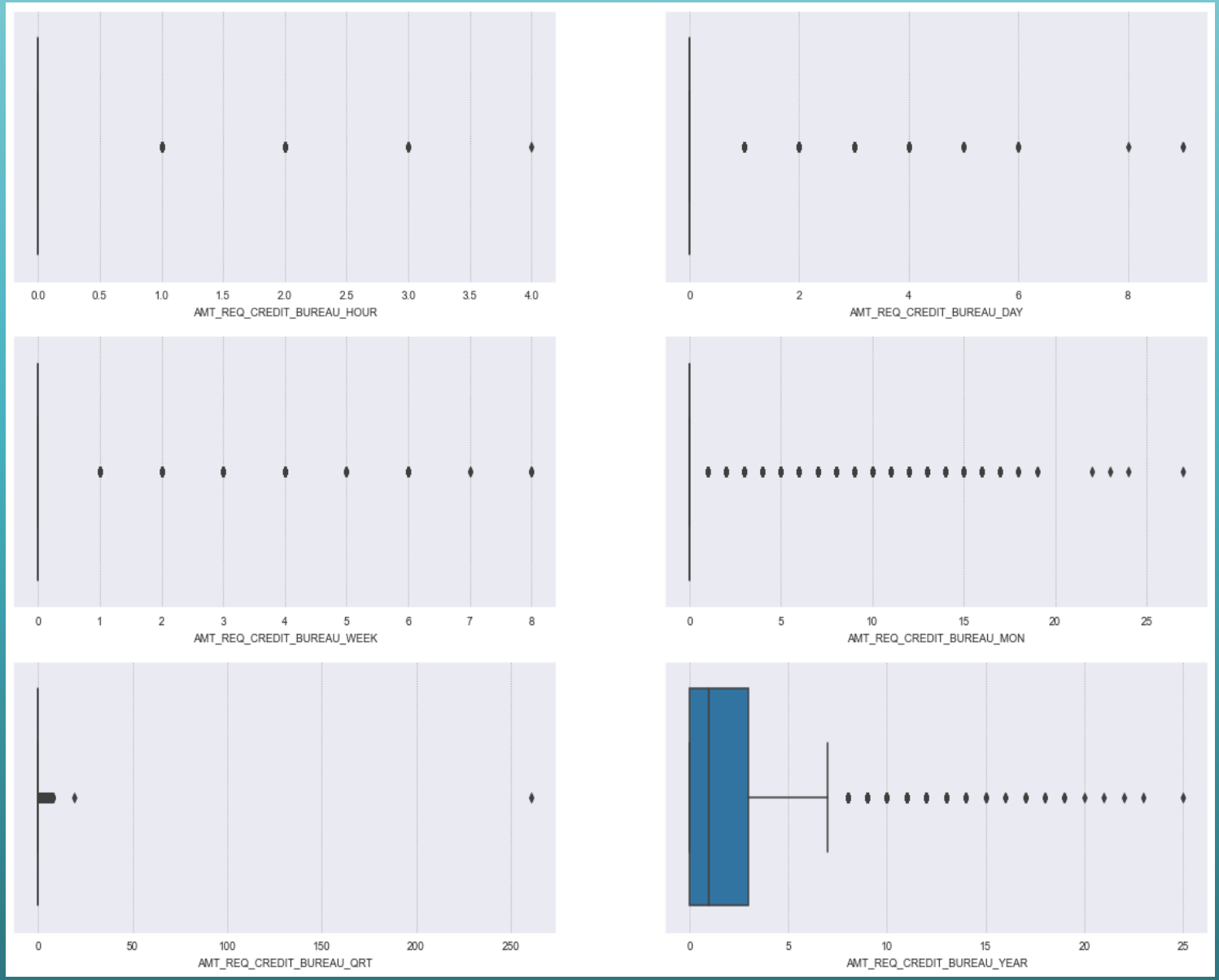


Number of enquiries to Credit Bureau

Inference:

- There is no proper evidence to use the columns since the median of all the 5 are almost or very near to 0.
- The 6th plot also have median close to zero along with high amount of outliers.
- The null values were filled with 0 (mode). Still it's the same as before.

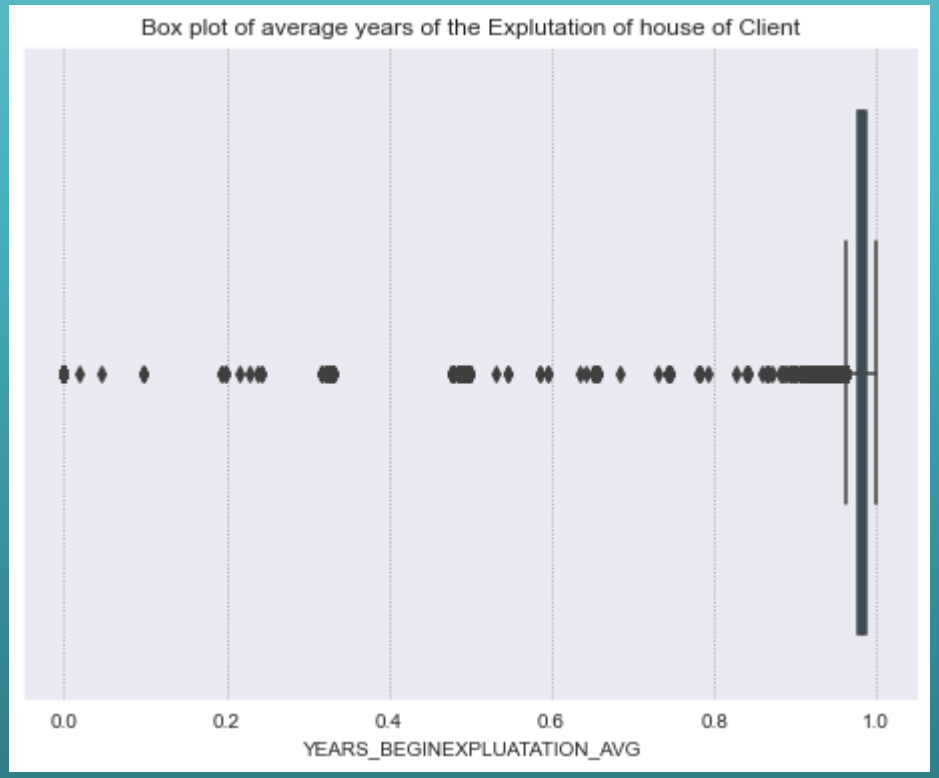
- These data columns will not be contributing to the Target variable

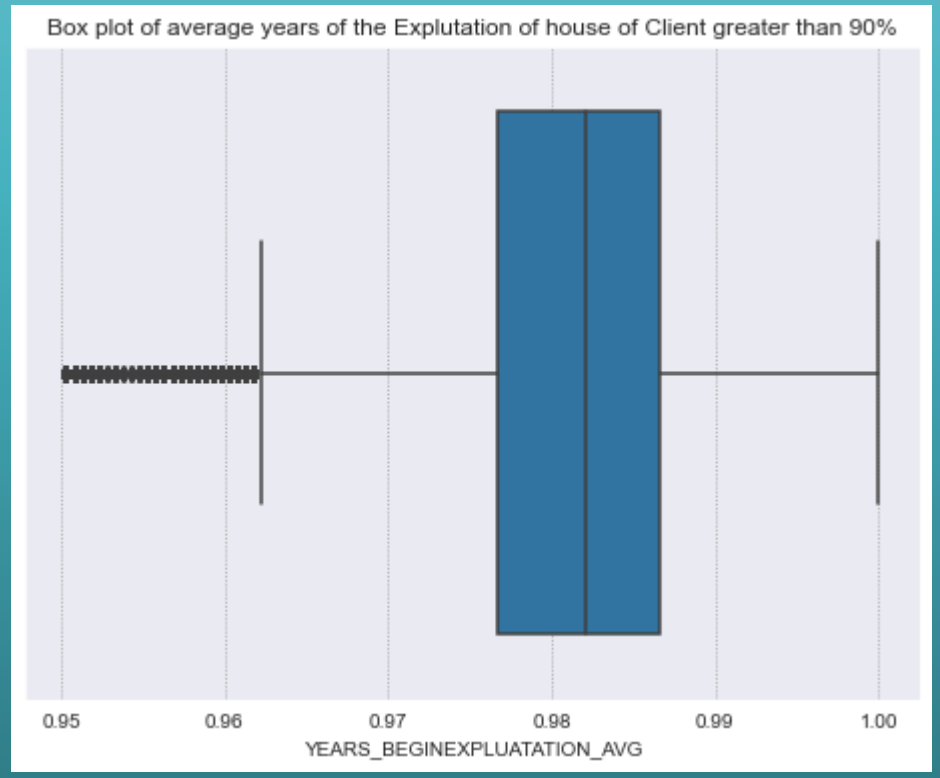


Average Expluatation of the Building of Client

Inference:

- Building Expluatation does not seemed to be a useful factor with over 95% data point breadth as outlier

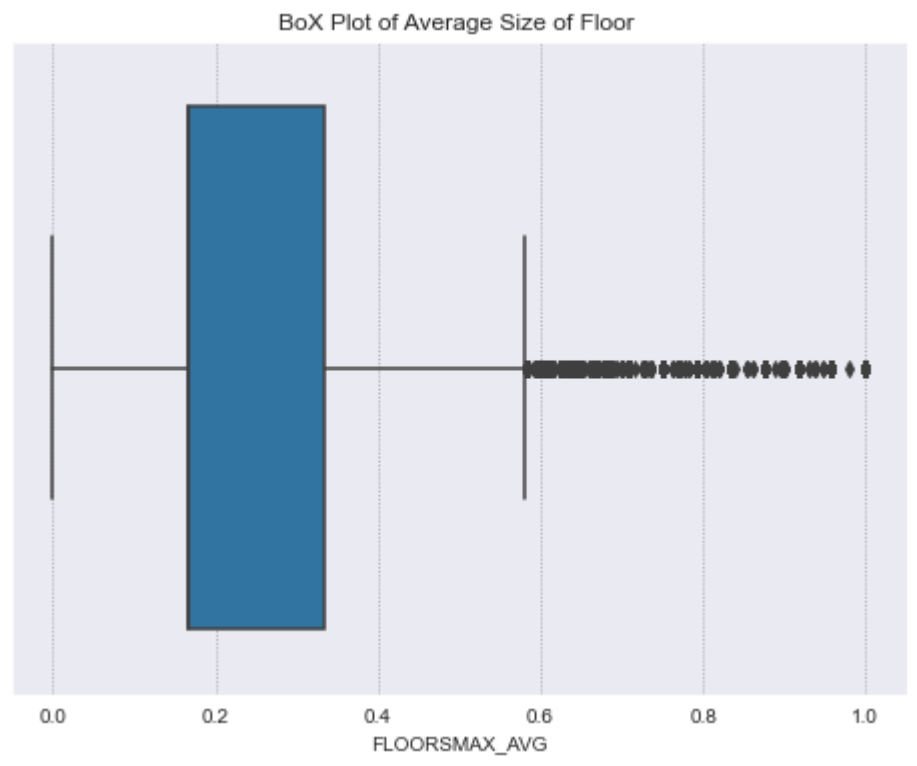


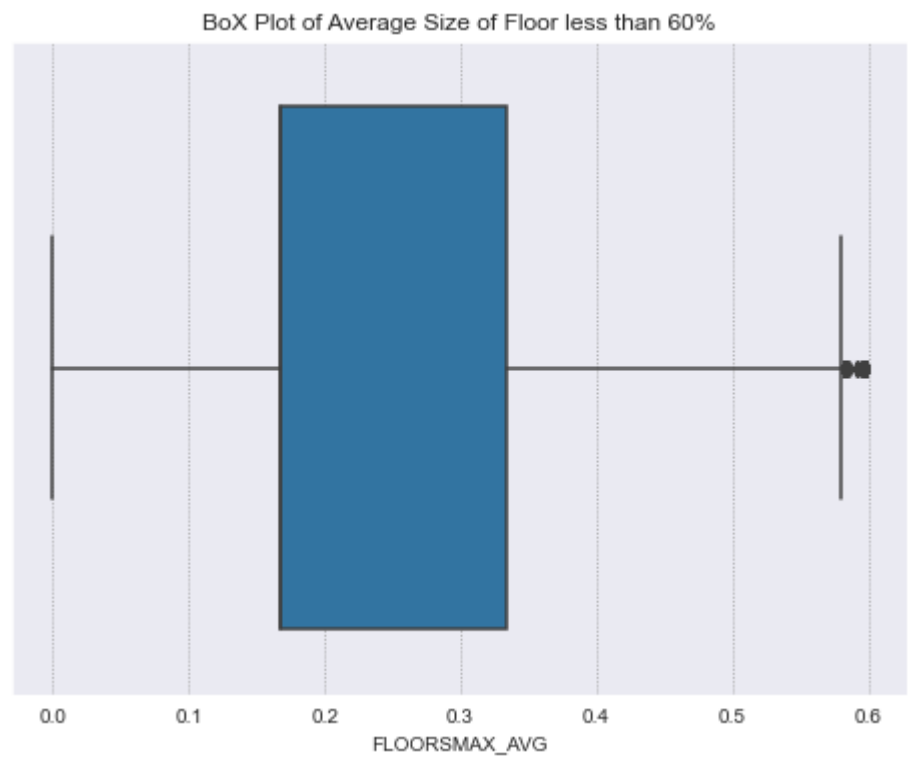


Average Expluatation of the Building of Client

Inference:

- Average floor size has more outliers since most of the houses will be of medium sizes which is depended on the monthly income and only few will be of upper class.
- Hence if we have to remove top 35% of data we are able to get this column with almost no outliers.

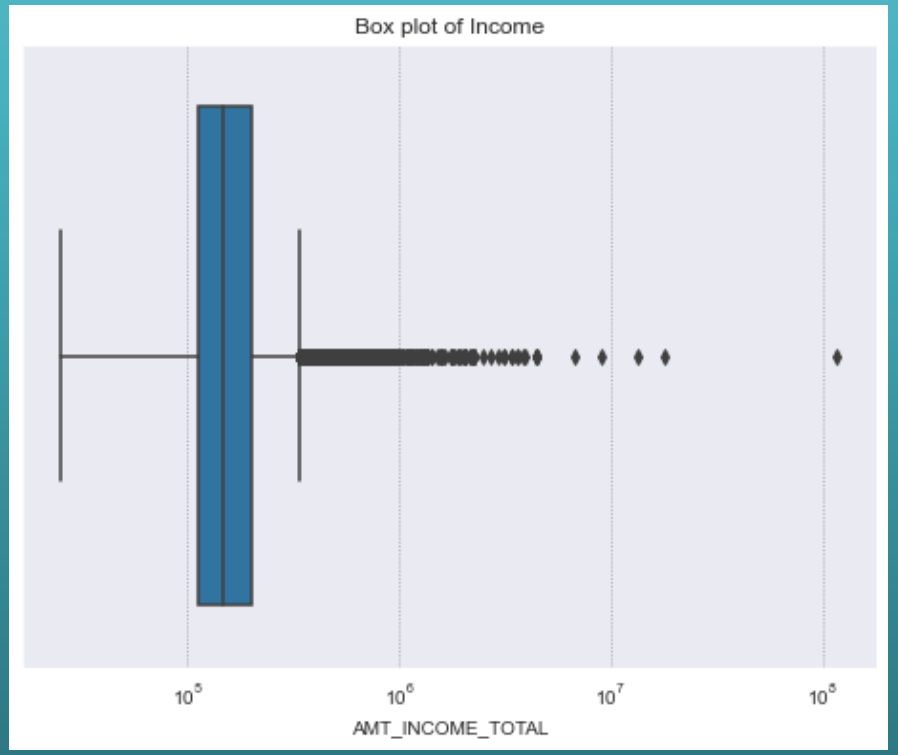




Income of the Client

Inference:

- Most of the people will be getting lower salary only (applying for loans).
- Very few only will be getting a high Salary. Hence the outlier case.

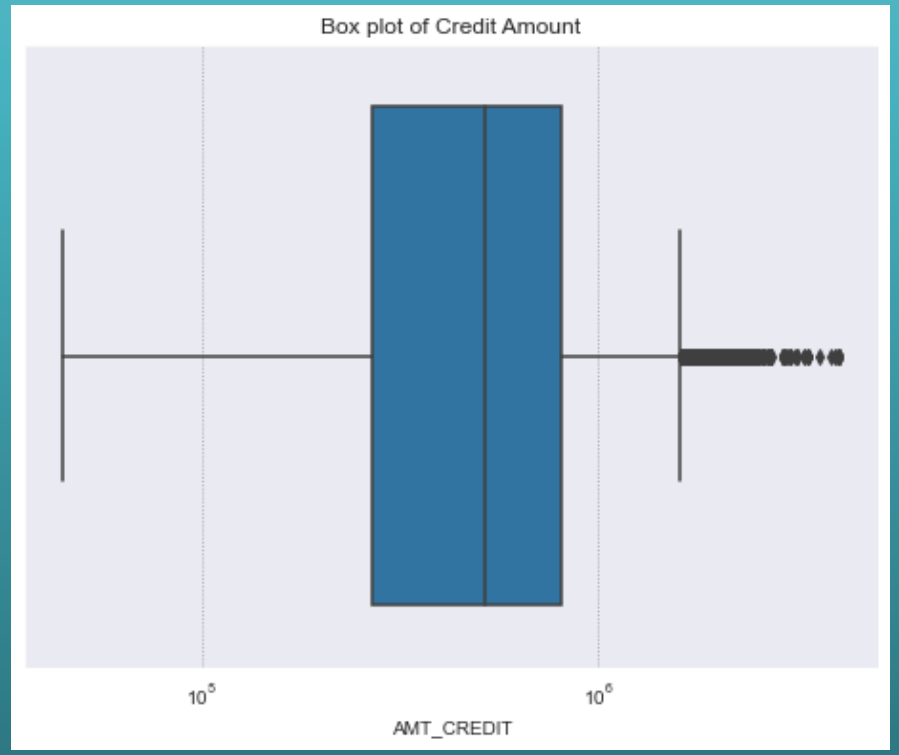


Credit Amount

Inference:

- Credit amount is some what having mean and median as same.

```
count    304531.00000
mean     599559.23833
std      402145.31390
min       45000.00000
25%      270000.00000
50%      517266.00000
75%      808650.00000
max     4050000.00000
Name: AMT_CREDIT, dtype: float64
```

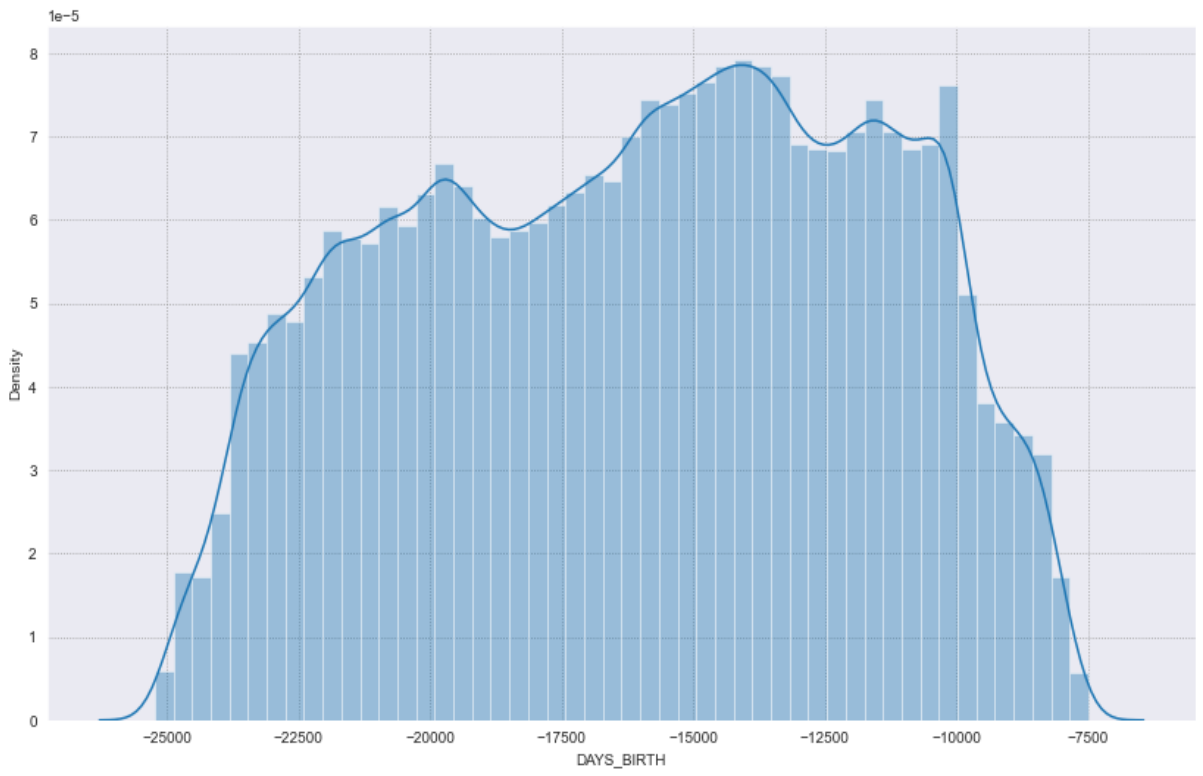


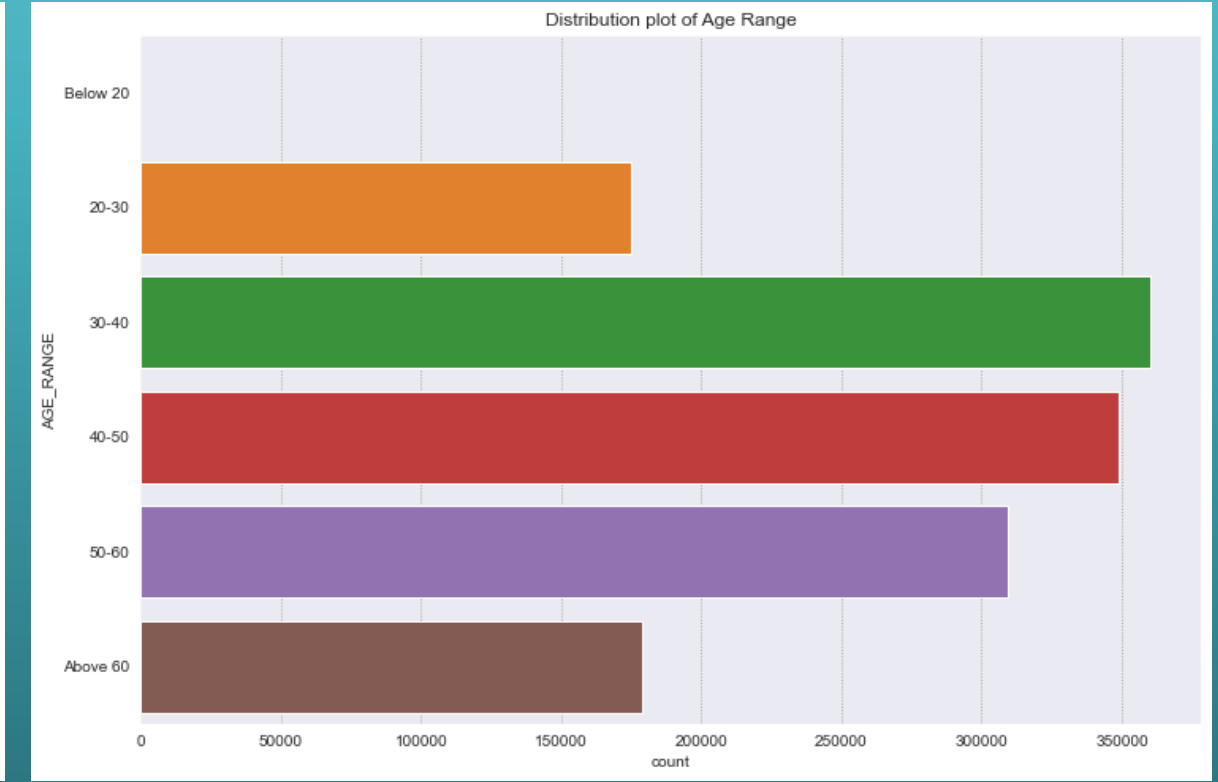
Distribution plots of important Data Columns

Age Range - Distribution

Inference:

- The Age column has negative values since it describes like how many days before the person has born.
- For visualization purpose we are converting it to positive values.
- We are also grouping it into range of values.

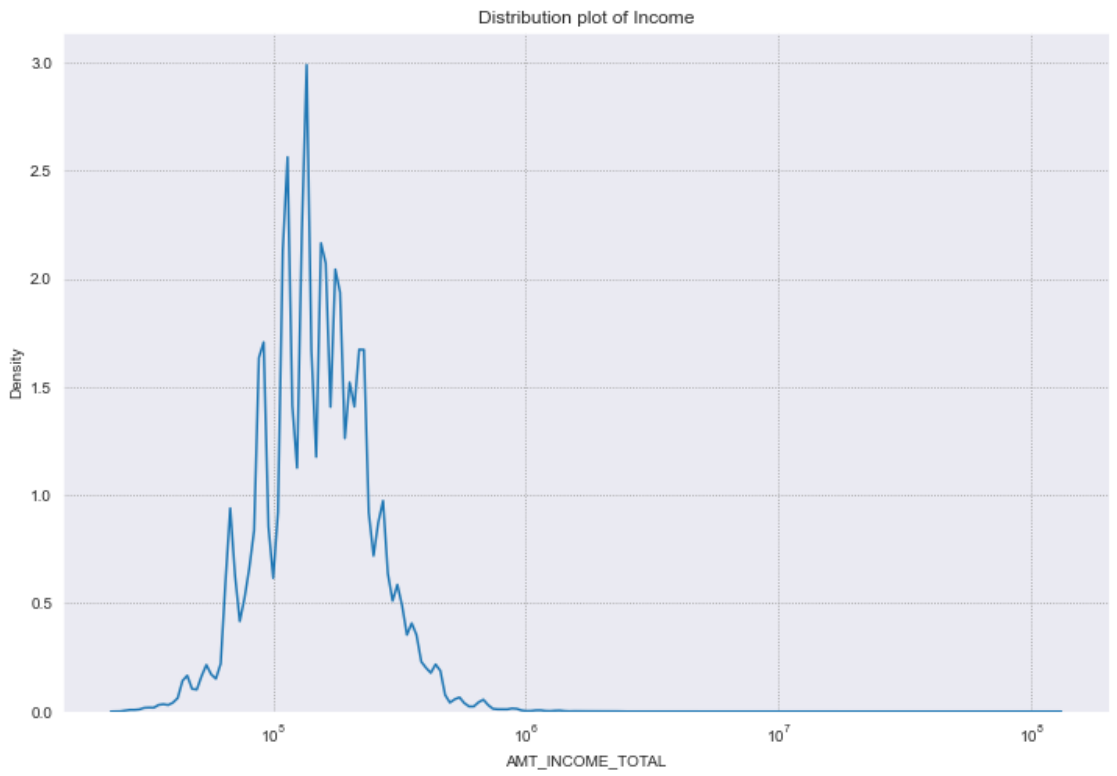


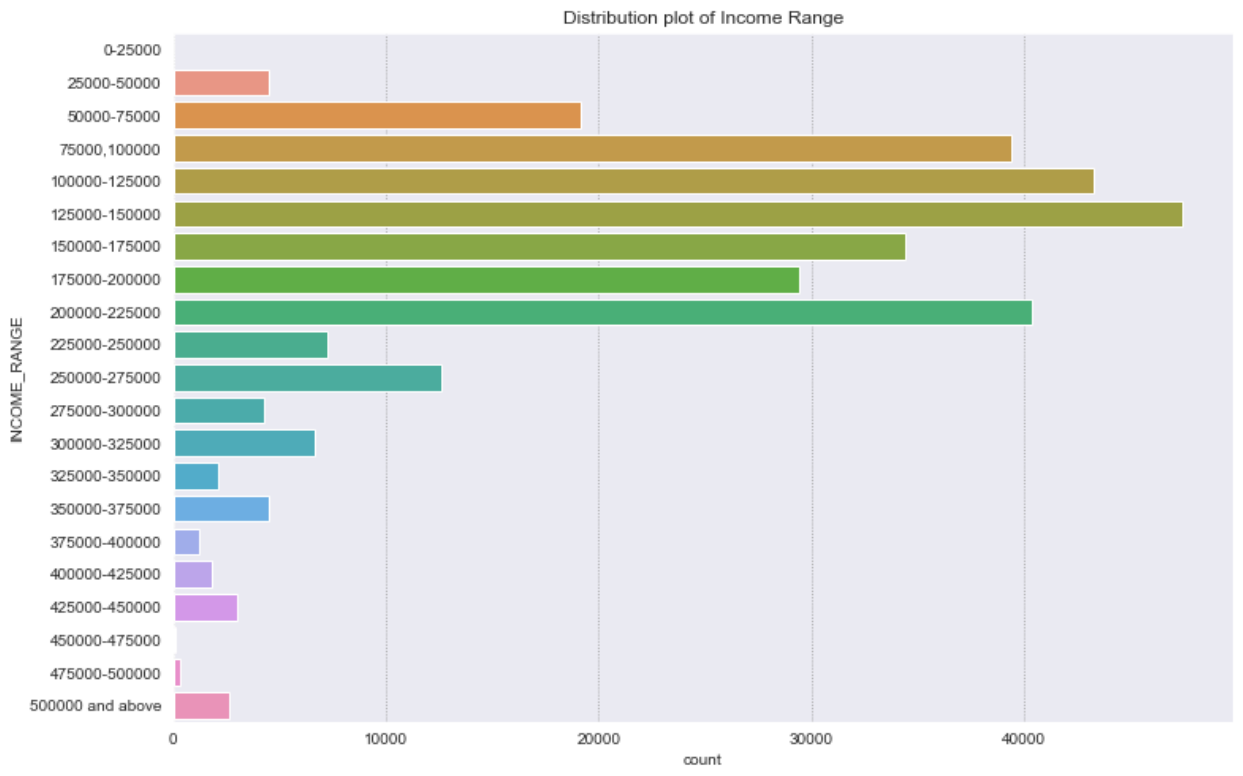


Income Range - Distribution

Inference:

- Many of the people will be having income ranging between 50,000 and 2,25,000.
- But it has a wide range of distribution from 25000 to 70,00,000
- We are also grouping it into range of values.

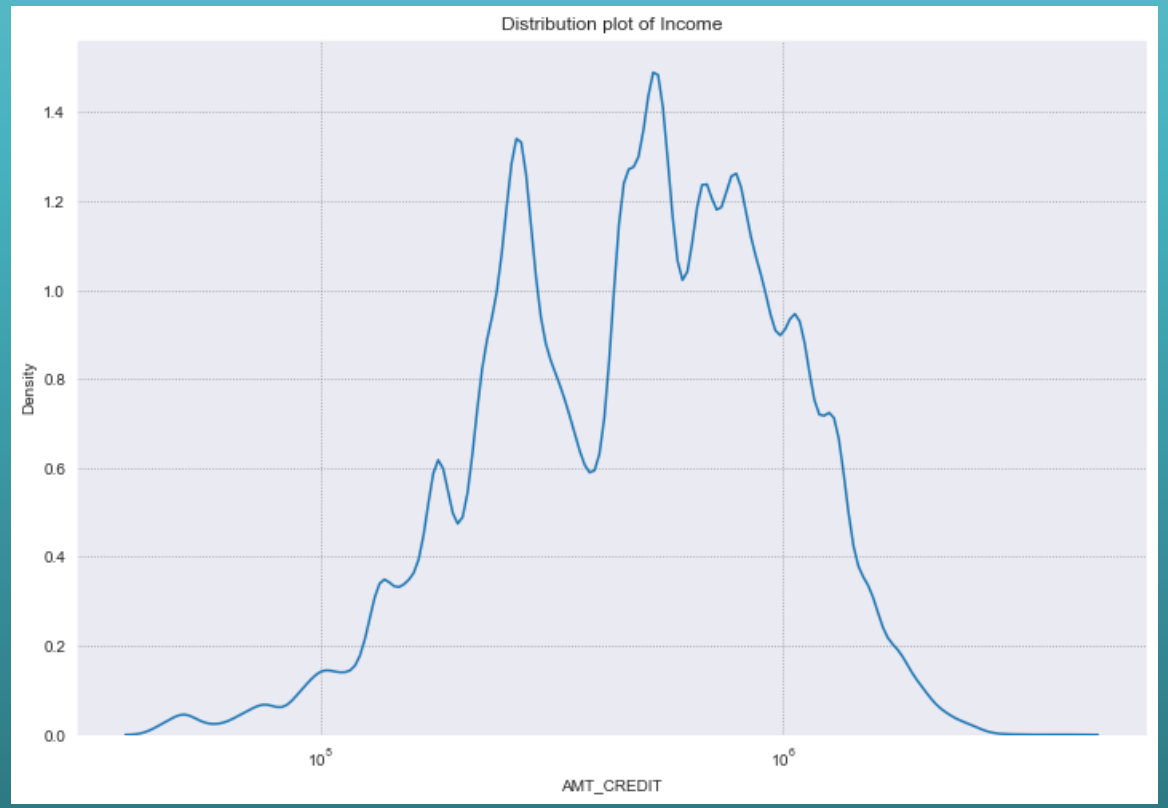


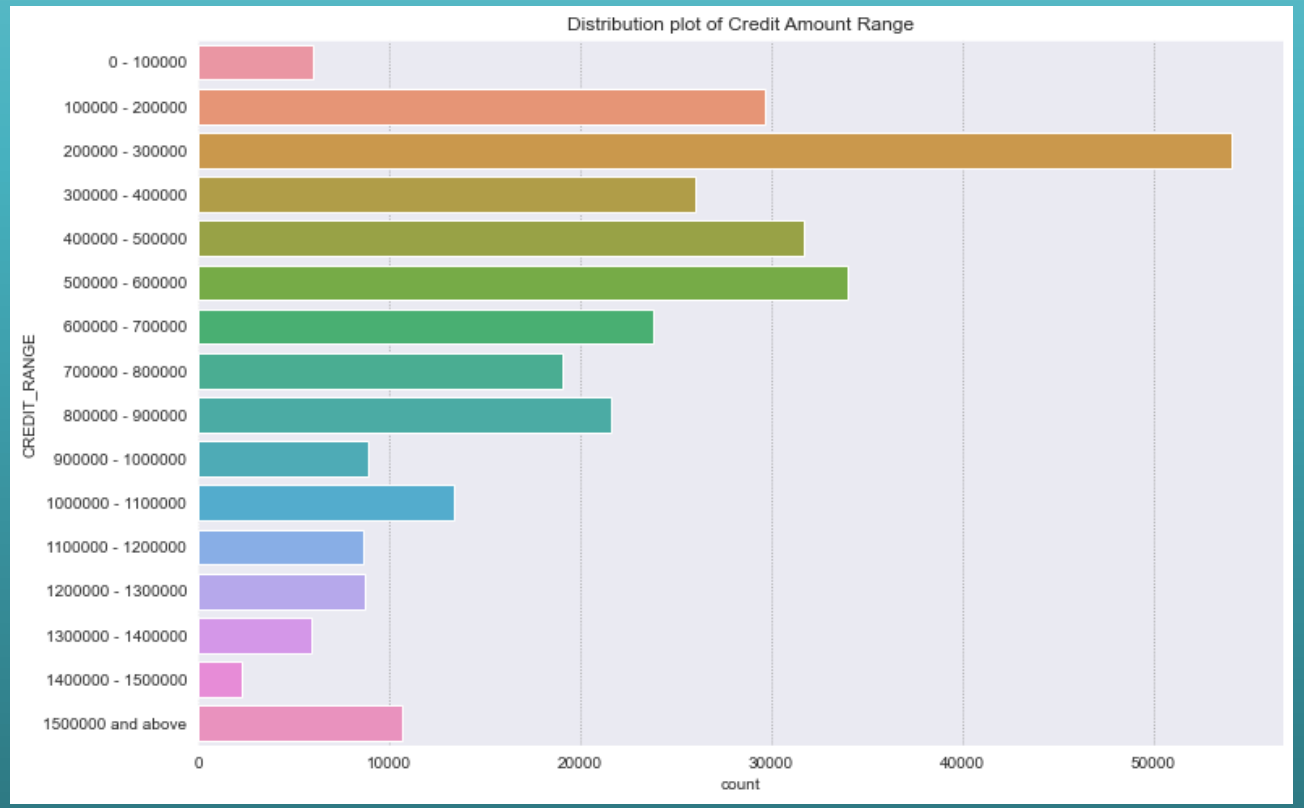


Credit Amount Range - Distribution

Inference:

- Most of the people have been credited by the loan amount ranging between 1,00,000 and 9,00,000.
- But it has a wide range of distribution from 1,00,000 to 1,50,00,000.
- We are also grouping it into range of values.

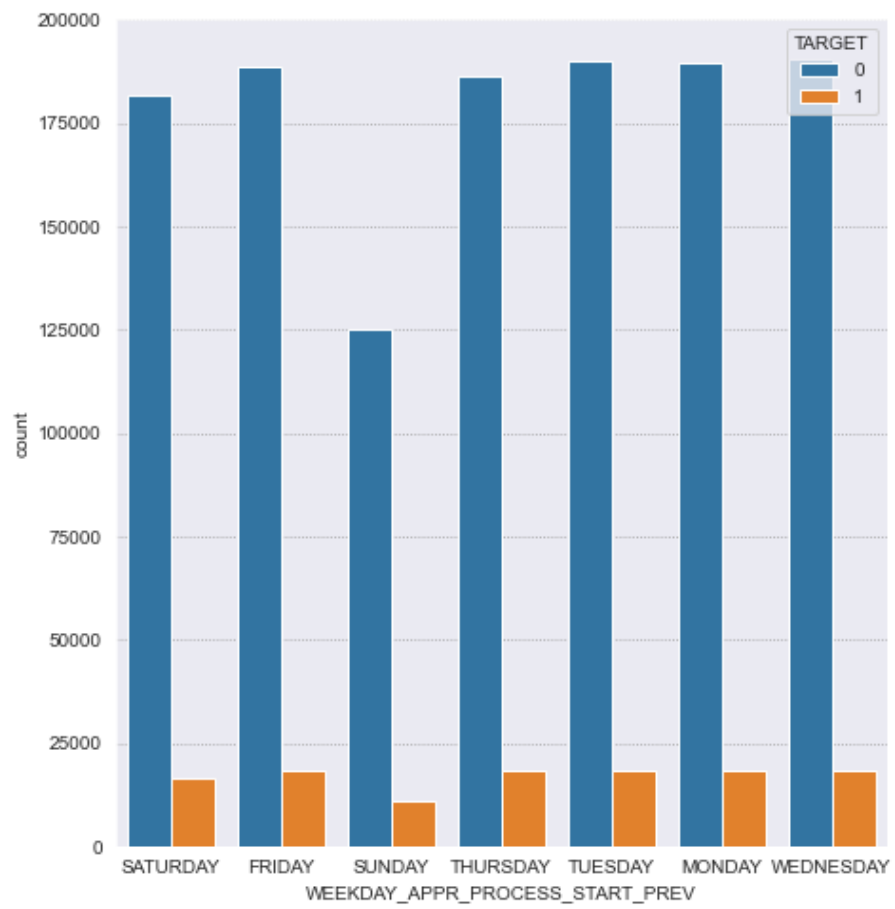


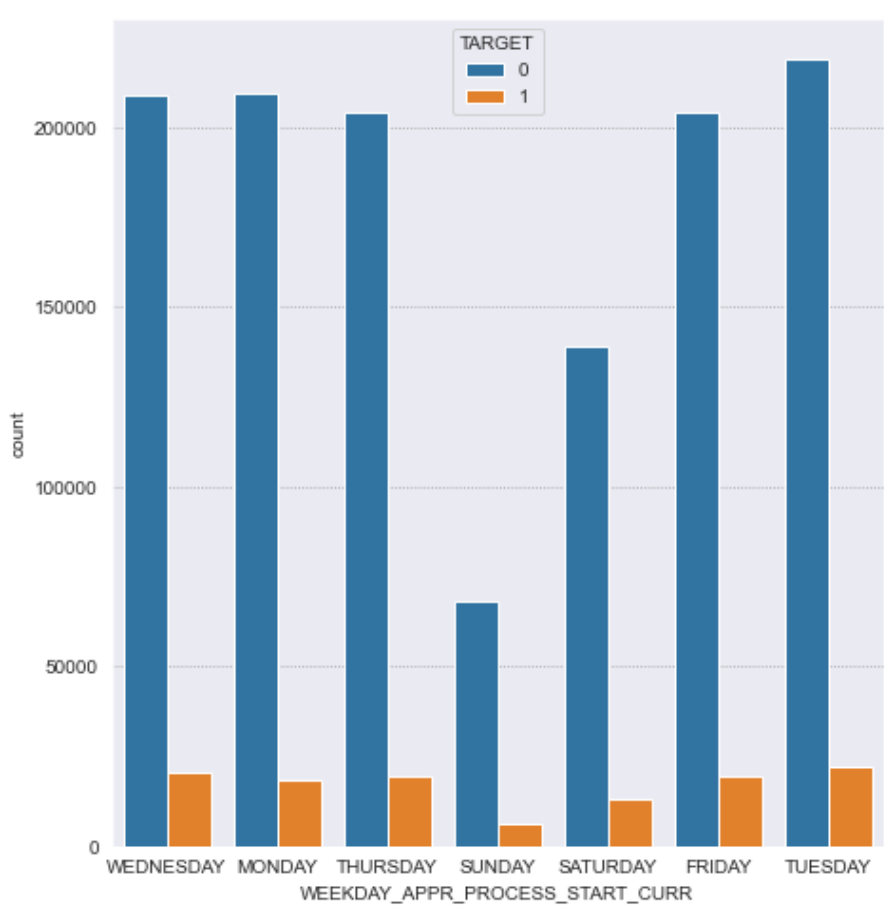


Day of application for loan

Inference:

- Both Previous and Current are almost Equal and this might be a direct correlation for Target column.

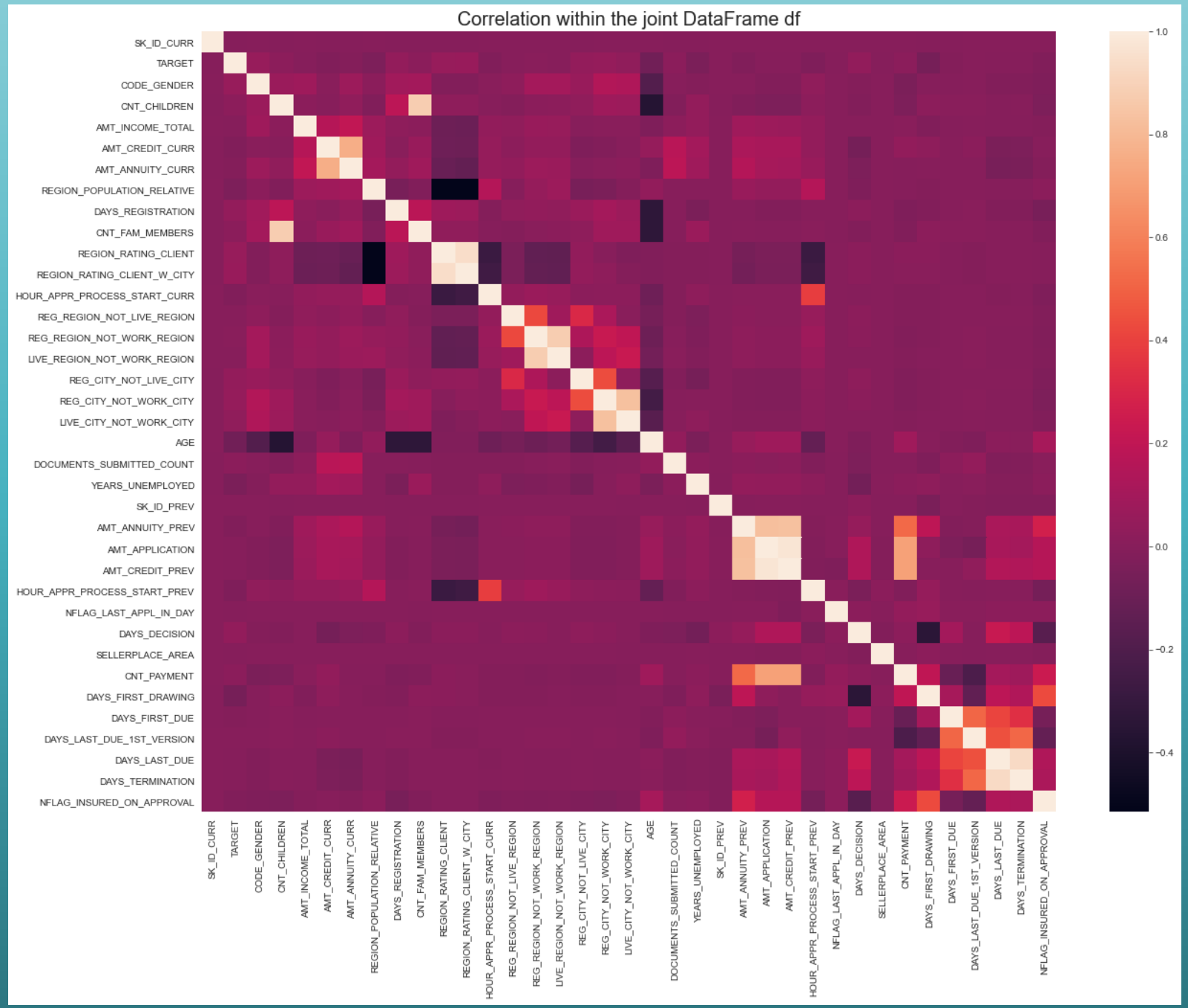




Correlation within the joint Dataframe

Inference:

- There is greater dependencies between the Previous dataset.
- Also there is greater correlation between the region living status in the center of the heatmap.
- The important factor Target depends on Count of family members, Region rating and days registration(little).
- The Amount annuity also depends on Credit amount and income and vice versa which makes more sensible.
- There are few columns seemingly having no direct relationship with other columns. Those can be removed.



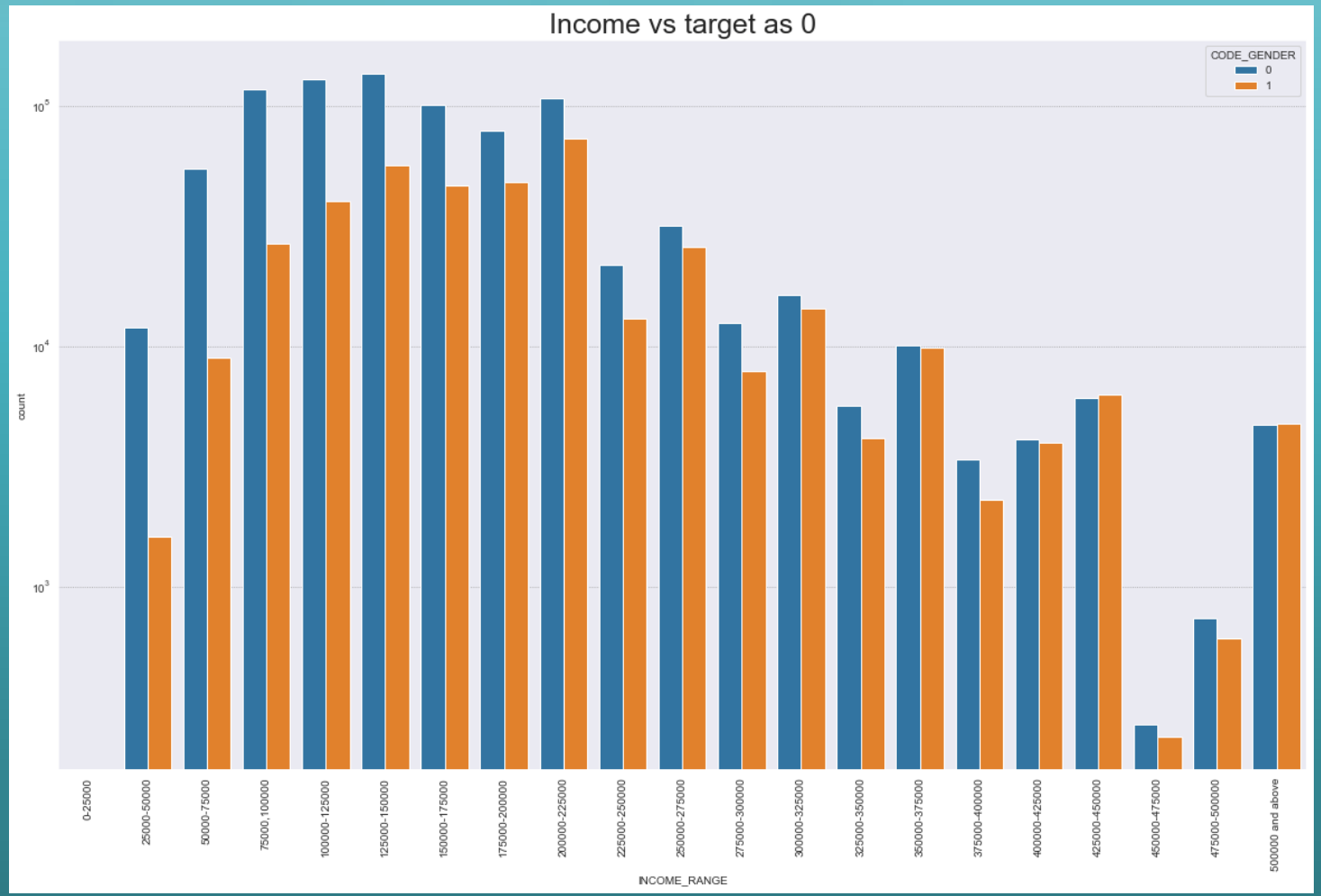
Univariate Analysis

Income range with hue as Gender

INFERENCE :

1. Seems like the Female gender having high counts and this can be expected since the data has high Female counts like twice as Male count. So Keeping in mind we could see that for lower Scale Salary the rejection possibilities seems to be equal since the repayment difficulties are higher.
2. Same when Salary gets higher the count gets reduced since for higher salary people will be able to repay the loan. But then here the count of men and women are equal (though the data is initially skewed towards female gender).
3. Point 2 means that Female might have a little consideration than Male.

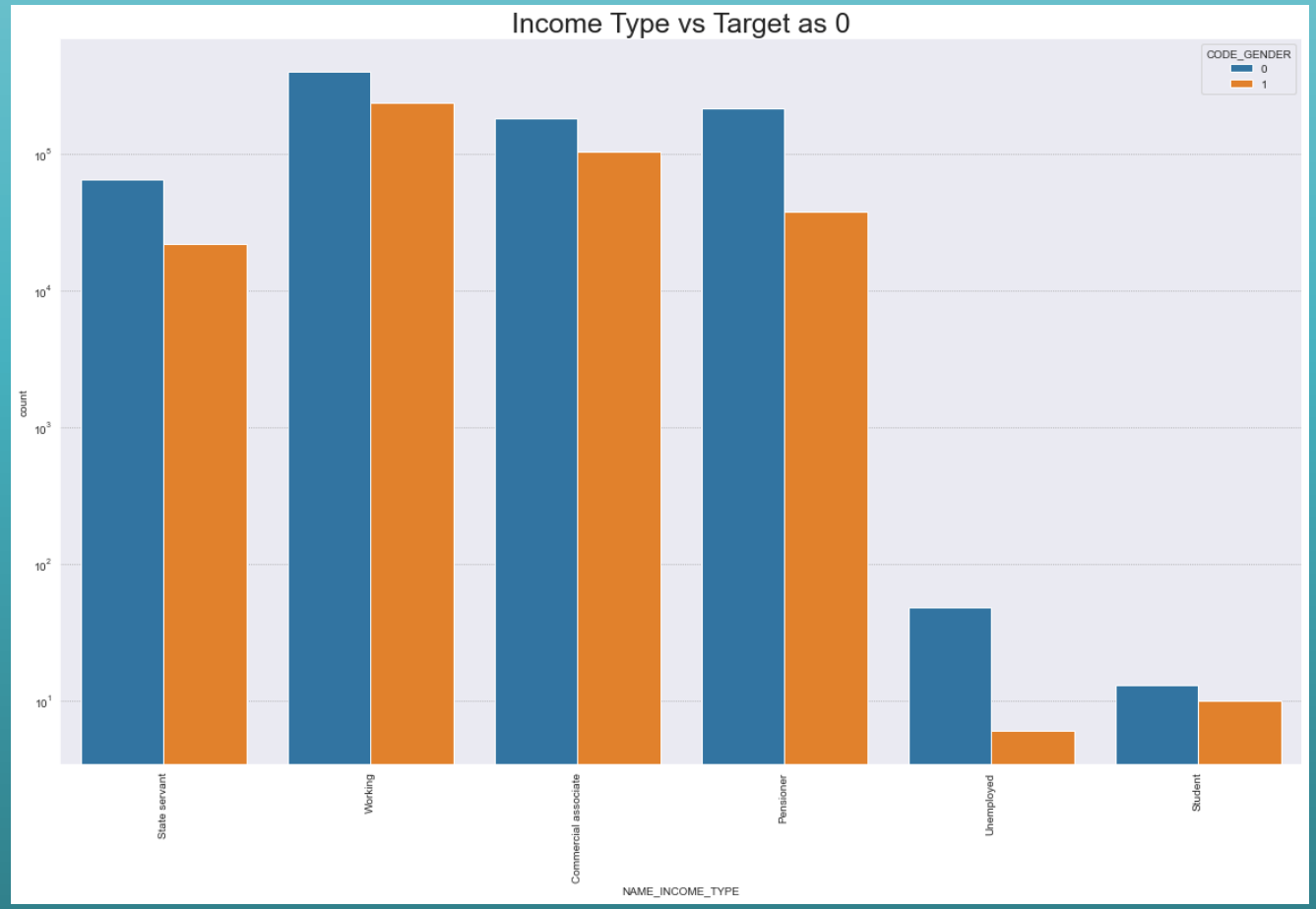
Note: Code 0 – Female
1 – Male



Income Type vs Target as 0

INFERENCE :

1. For State Servant the chances of repaying difficulties are low when compared to others since they might have many benefits and offers as Government staff.
2. For student loans its not easily predictable, only prediction we could do is with their current education status.
3. For Pensioner the loan amount that can be allowed will be less and hence the difficulties are lower than the working people.

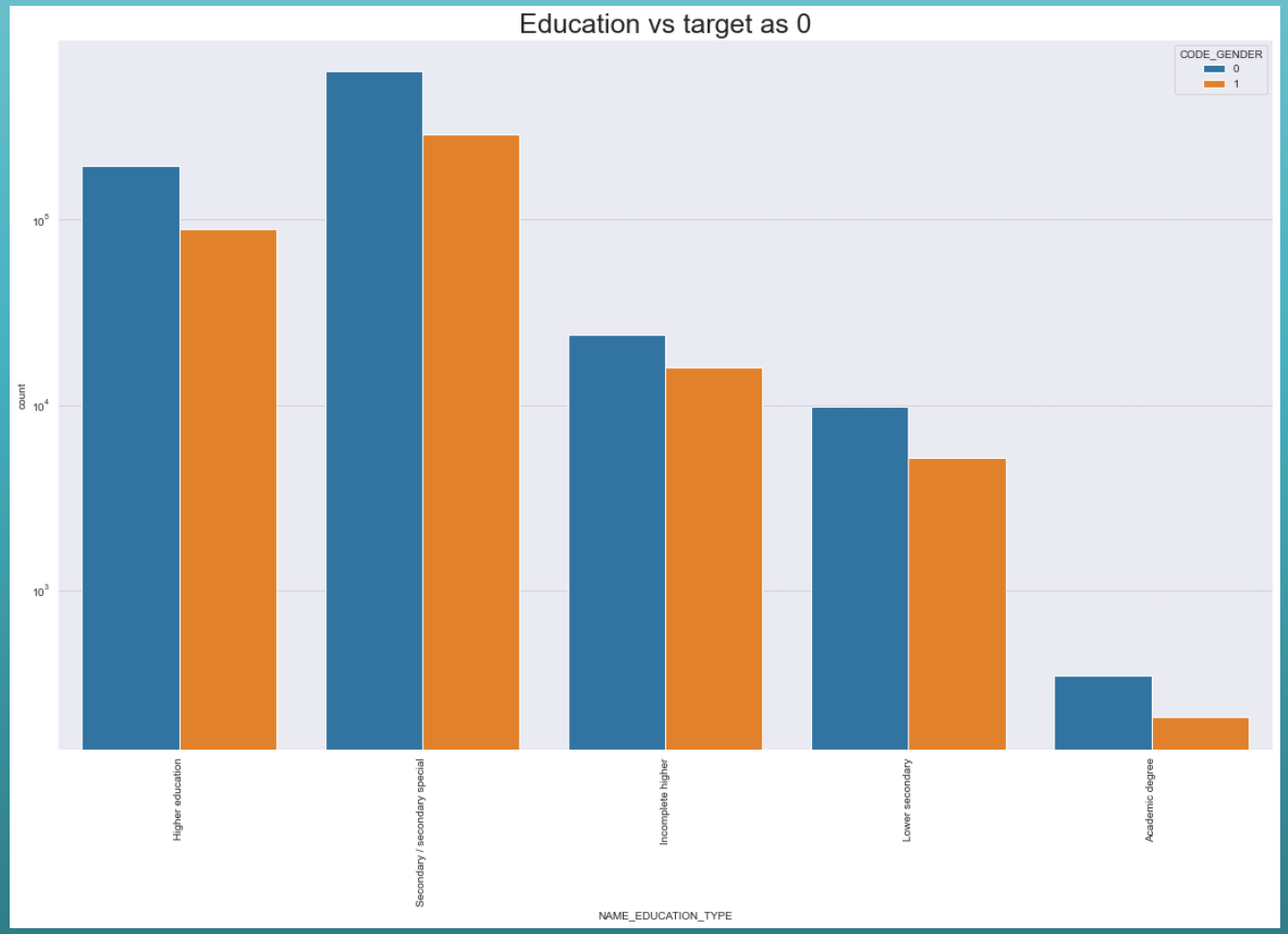


Note: Code 0 – Female
1 – Male

Education vs target as o

INFERENCE :

1. The repayment difficulties are low for the people studying or completed Acedemic degree(highest of all in the options).
2. While the Secondary degree will have difficulties higher than Acedemic degree which is very much relatable and acceptable.



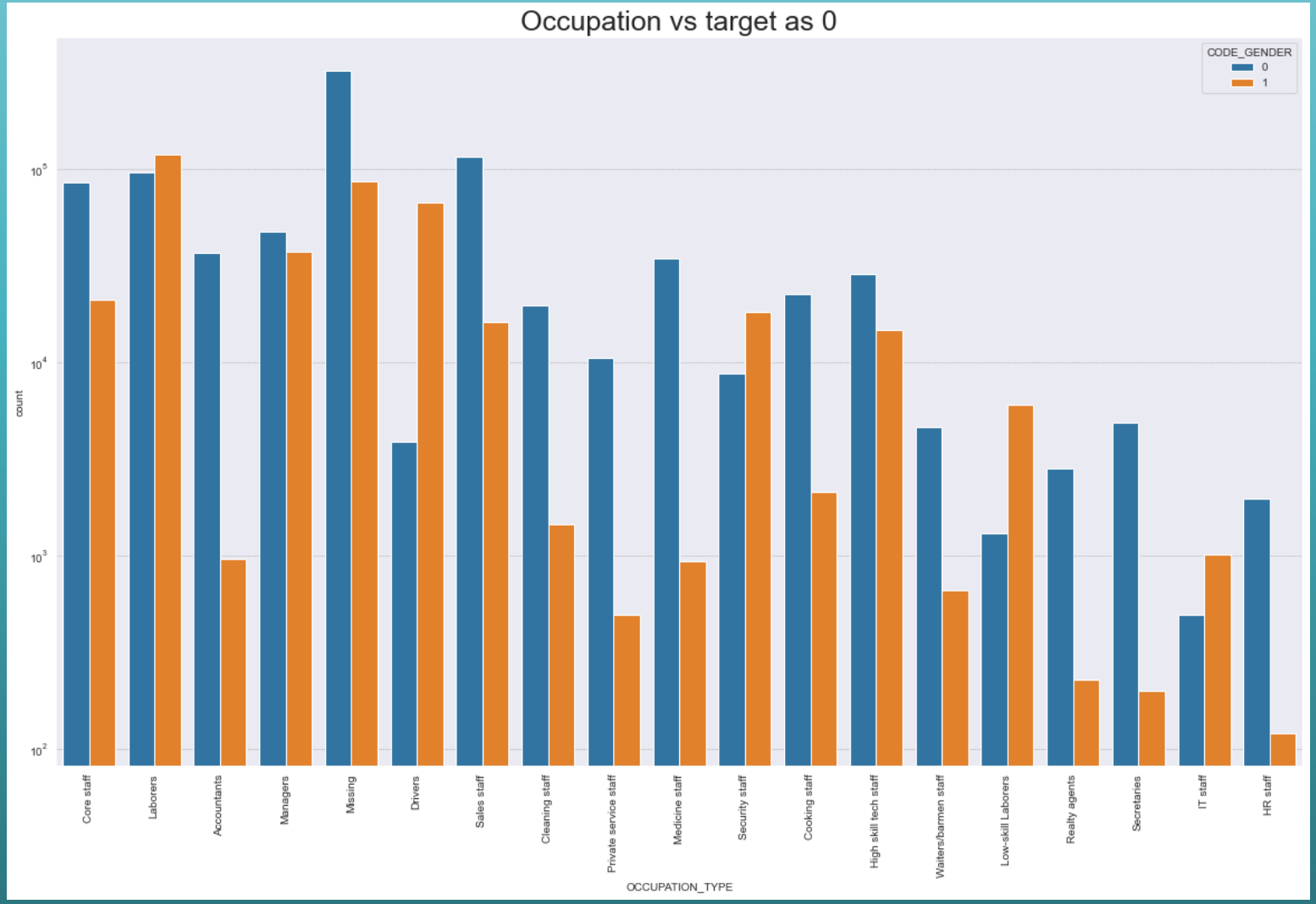
Note: Code 0 – Female
1 – Male

Occupation vs target as o

INFERENCE:

1. There are lot of missing data for the Occupation so we might need to neglect that for our current inference.
2. One of the reasons that Female have high Missing Occupation is they might be a homemaker or doing a small scale business which might not be mentioned.
3. The Laborers, Sales staff have difficulties in repaying since their annual income be less when compared to others.
4. For Drivers there might be more men in that occupation than women and hence thats relatable.

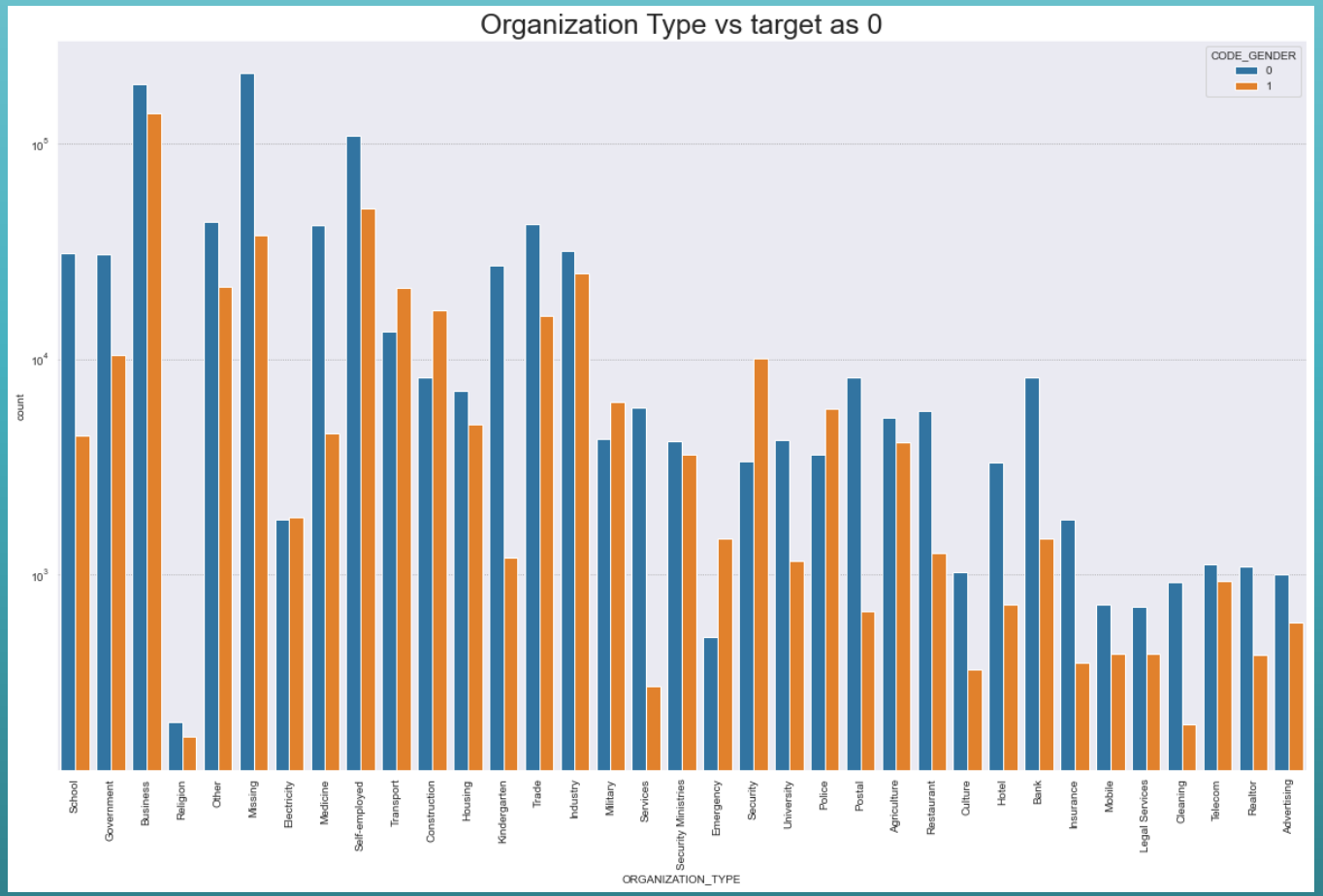
Note: Code 0 – Female
1 – Male



Organization Type vs target as o

Inference:

- Neglecting the Missing Type
- For the Business the applications are more and also the risk that they might not pay back is more because of unstable income returns from Business.
- The Same goes for Self-Employed also.
- For other kind where there are stable income like Industry, Services, Restaurant, etc have a less chances that they might not pay back.

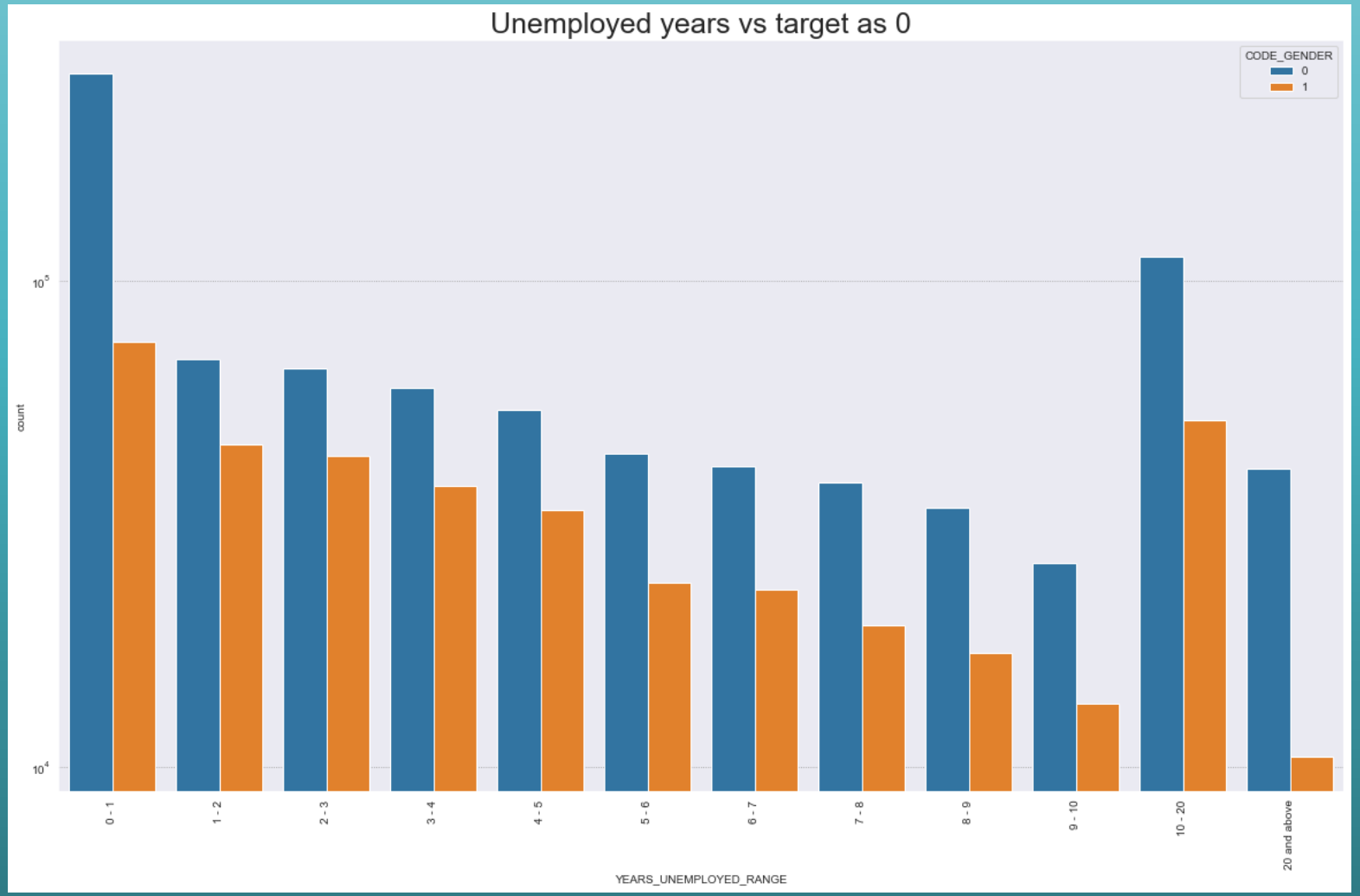


Note: Code 0 – Female
1 – Male

Unemployed years vs target as o

Inference:

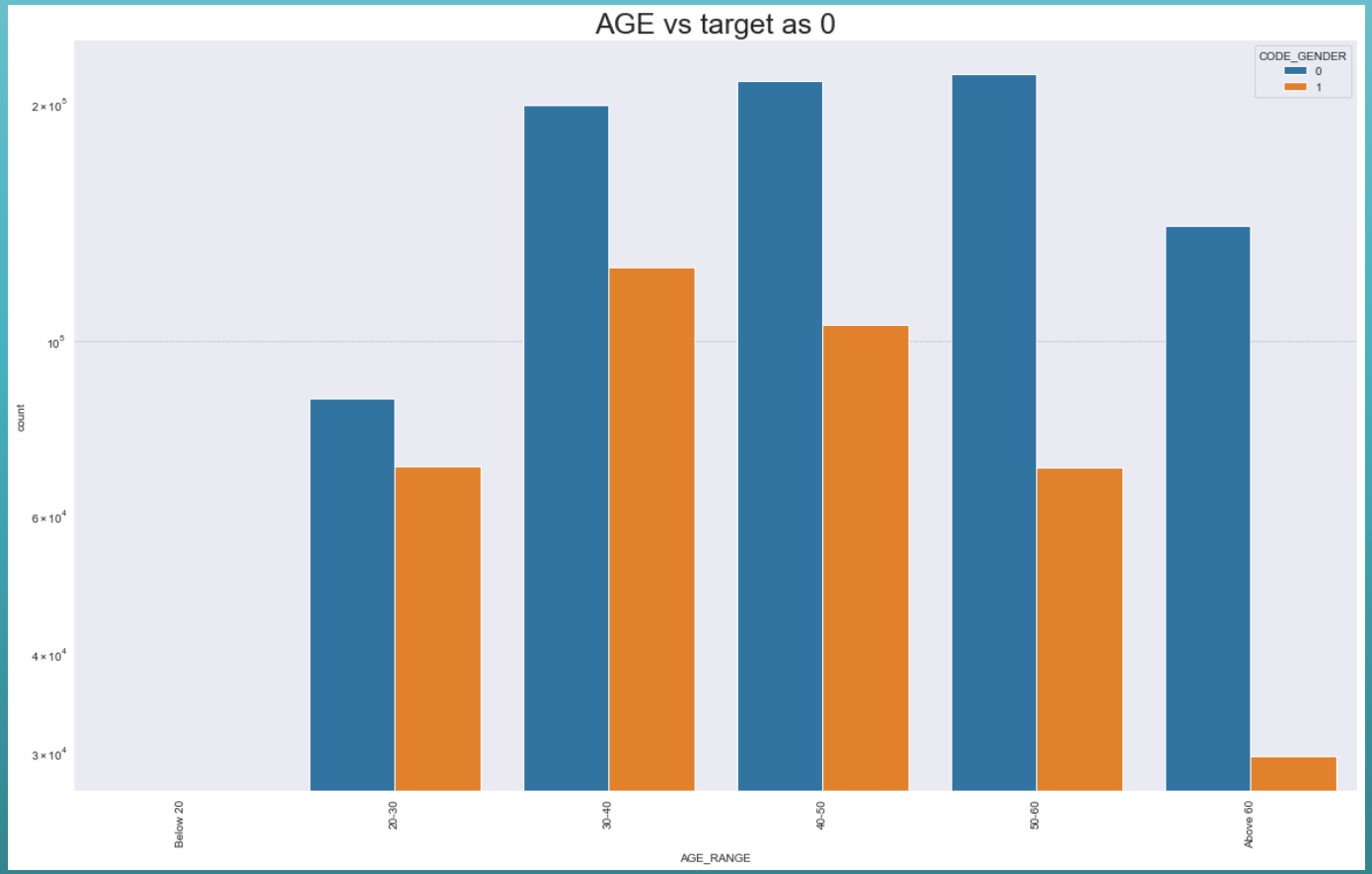
- There are many applications from the Female who have been unemployed less than a year. but there is also a high risk for male as well as female.



Age vs target as o

Inference:

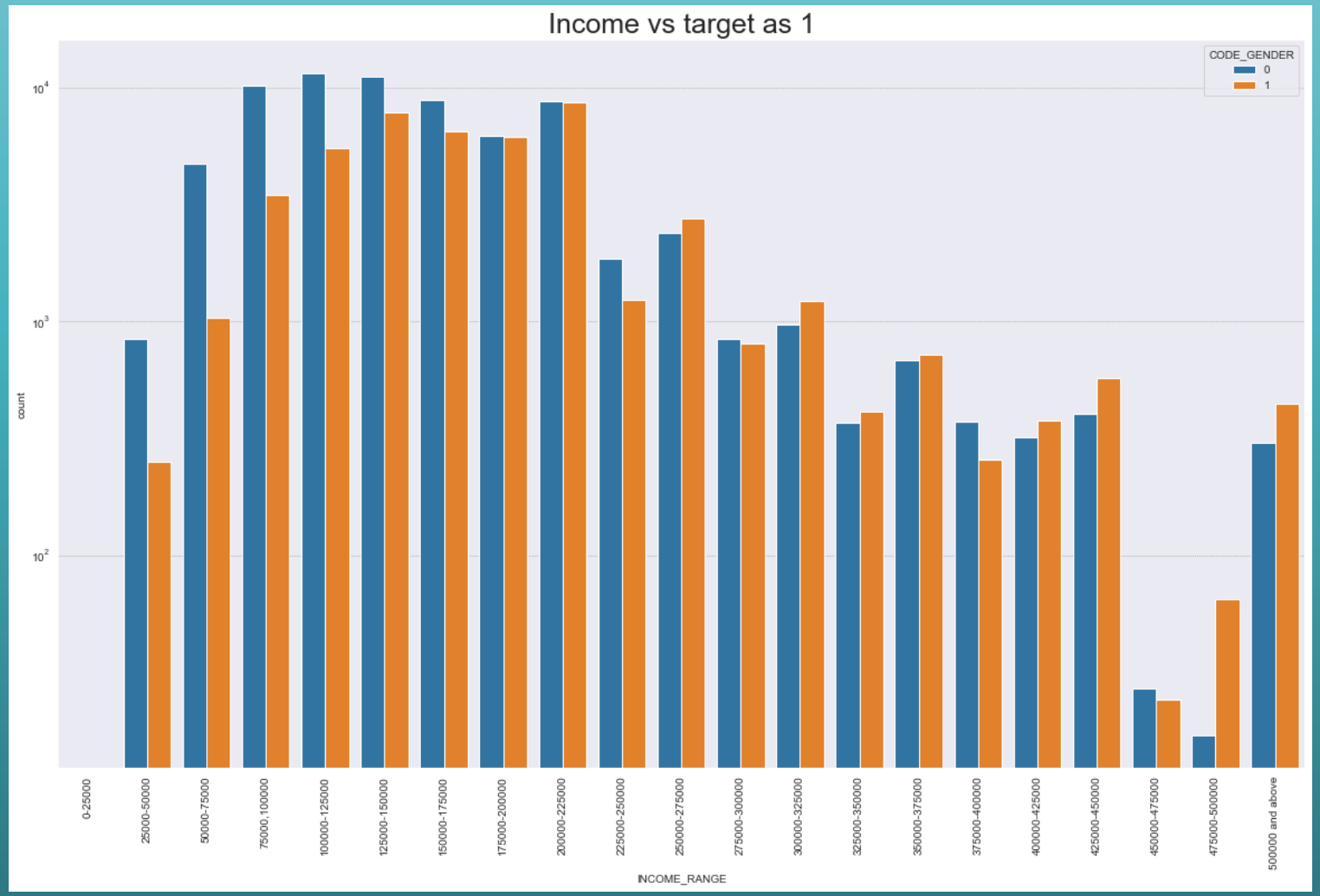
- The Age graph is somewhere like bell curve where the people of center age like above 50 nearing the retirement age they have high chances that they might not pay back,
- Keeping in mind the Male count is half less than the female count in dataset, we can get to know that here almost with age both genders have risk equally.



Income vs target as 1

Inference:

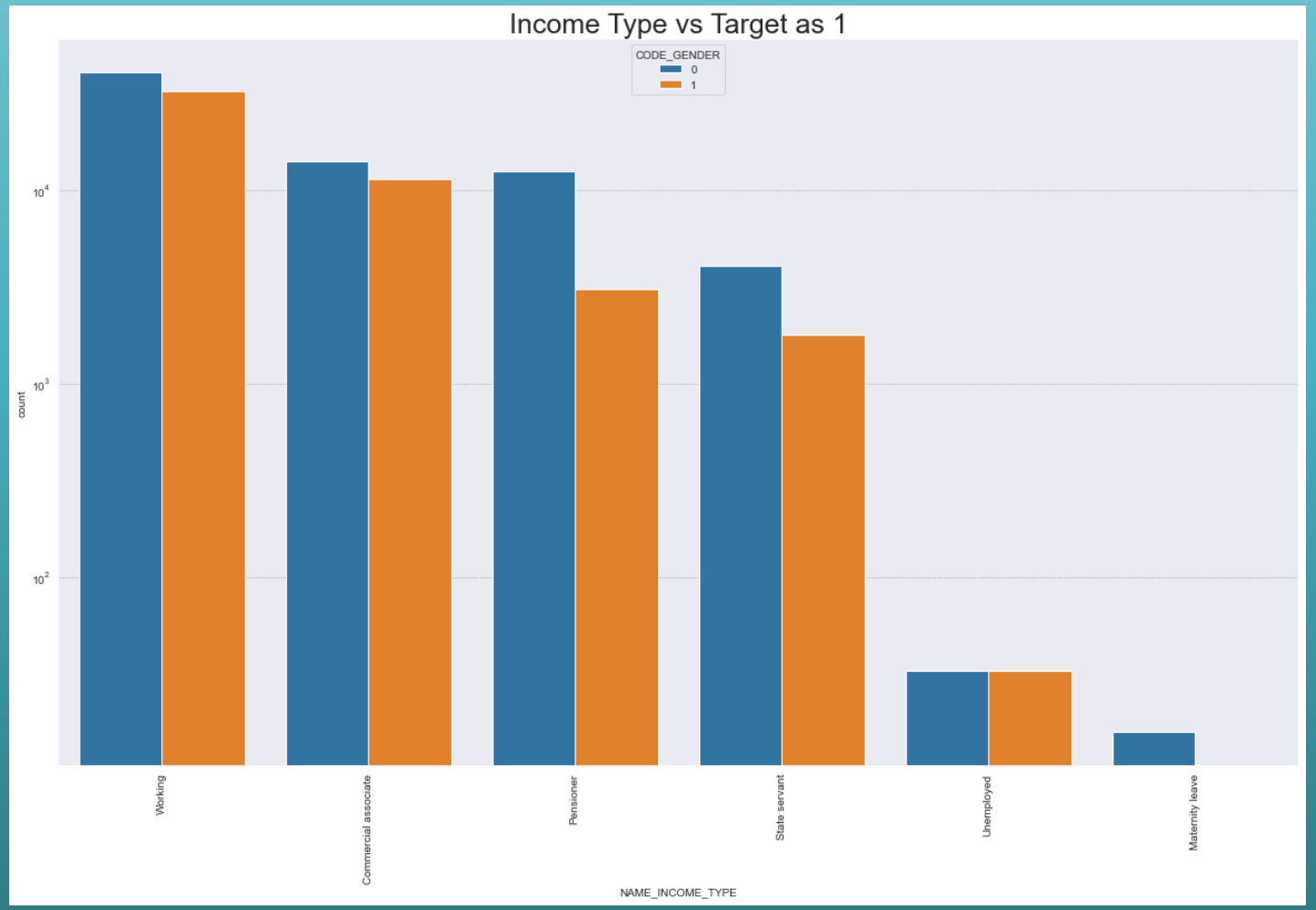
- The More Successful payers are from the category who earn between 1Lakh and 2.5Lakh, since they might be in mentality to get loan for what they earn and they have more time to pay back.
- People earning above 4L might not need a loan and the applications are itself less in number may be.



Income Type vs Target as 1

Inference:

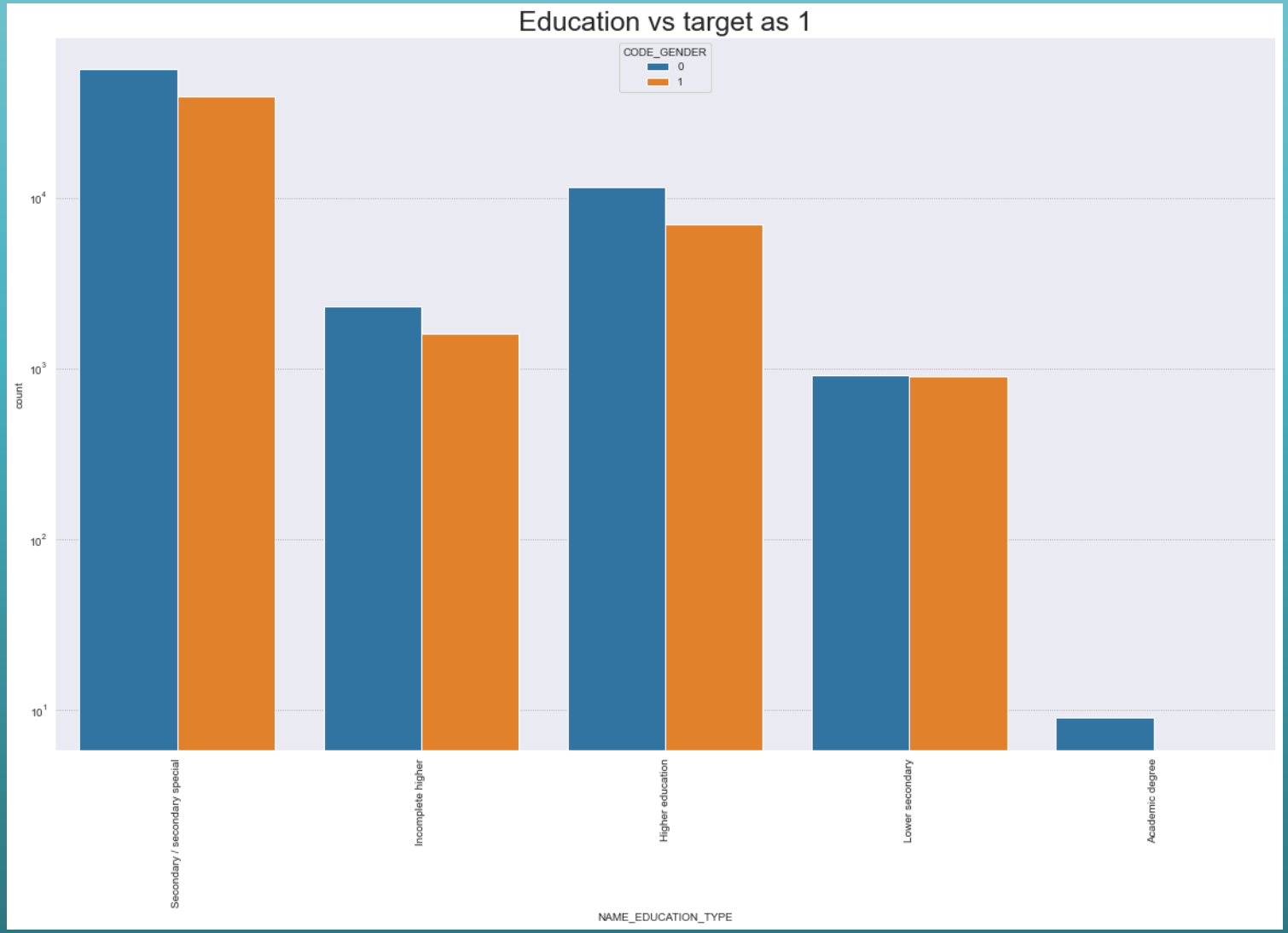
- Working people have high chances that they repay because of stable income.
- Other hand unemployed will not have good opportunity to pay back.



Education vs target as 1

Inference:

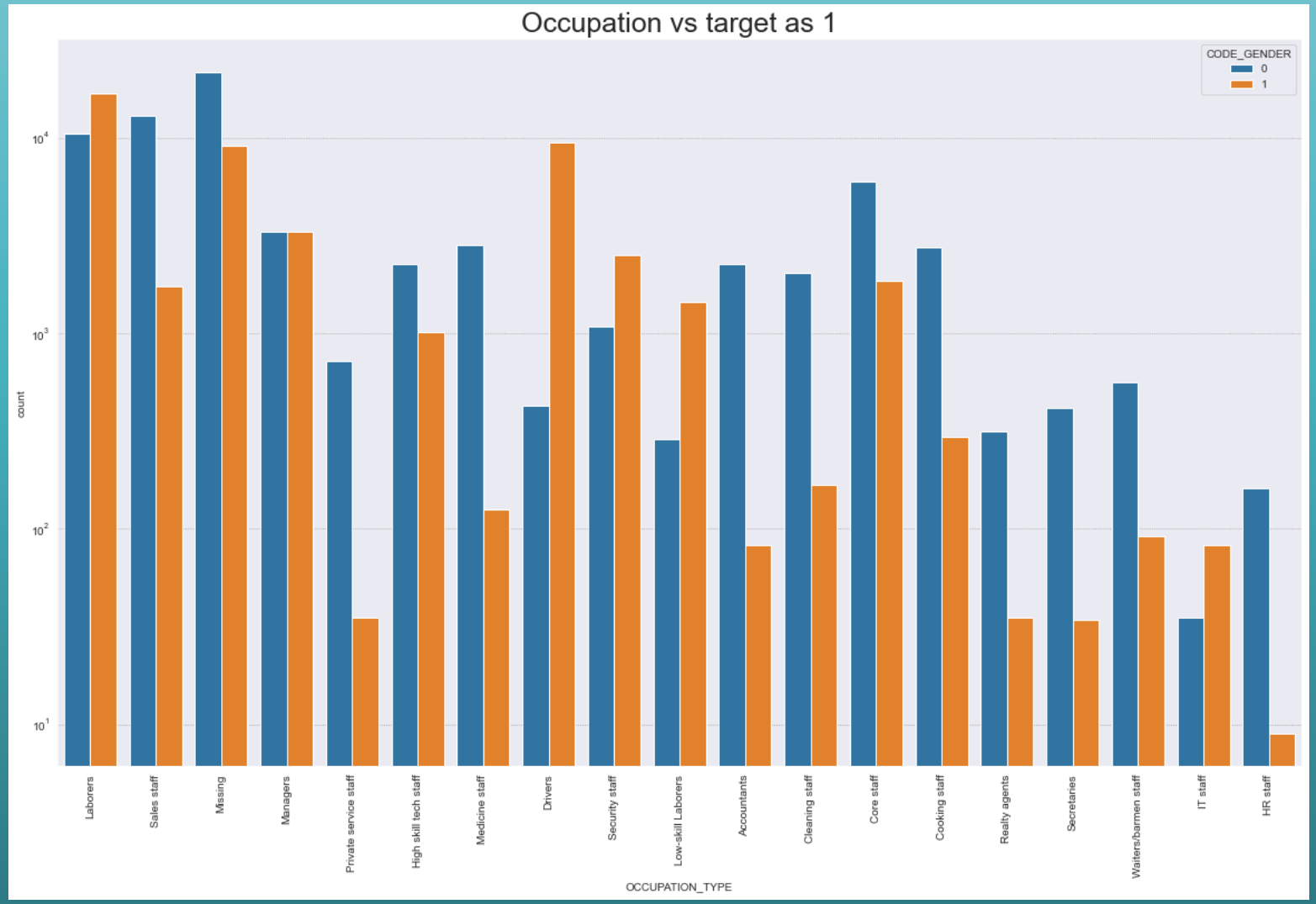
- The count of people applying for loan will have majorly completed at least a secondary degree.
- The chances of paying the loan for a person incomplete in his studies are lower.



Occupation vs target as 1

Inference:

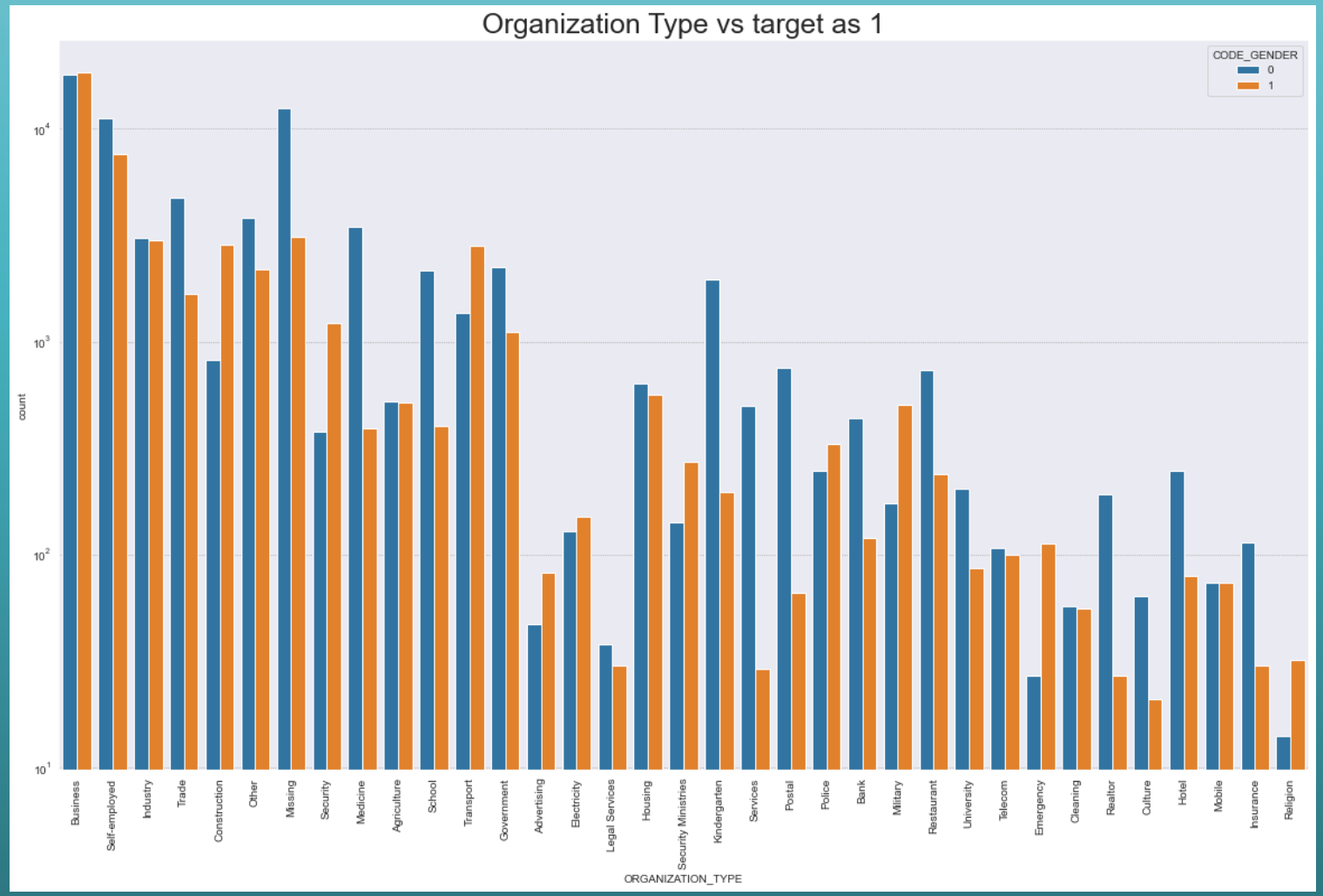
- There are lot of missing data for the Occupation so we might need to neglect that for our current inference.
- One of the reasons that Female have high Missing Occupation is they might be a homemaker or doing a small scale business which might not be mentioned.
- For Drivers there might be more men in that occupation than women and hence that's relatable.



Organization Type vs target as 1

Inference:

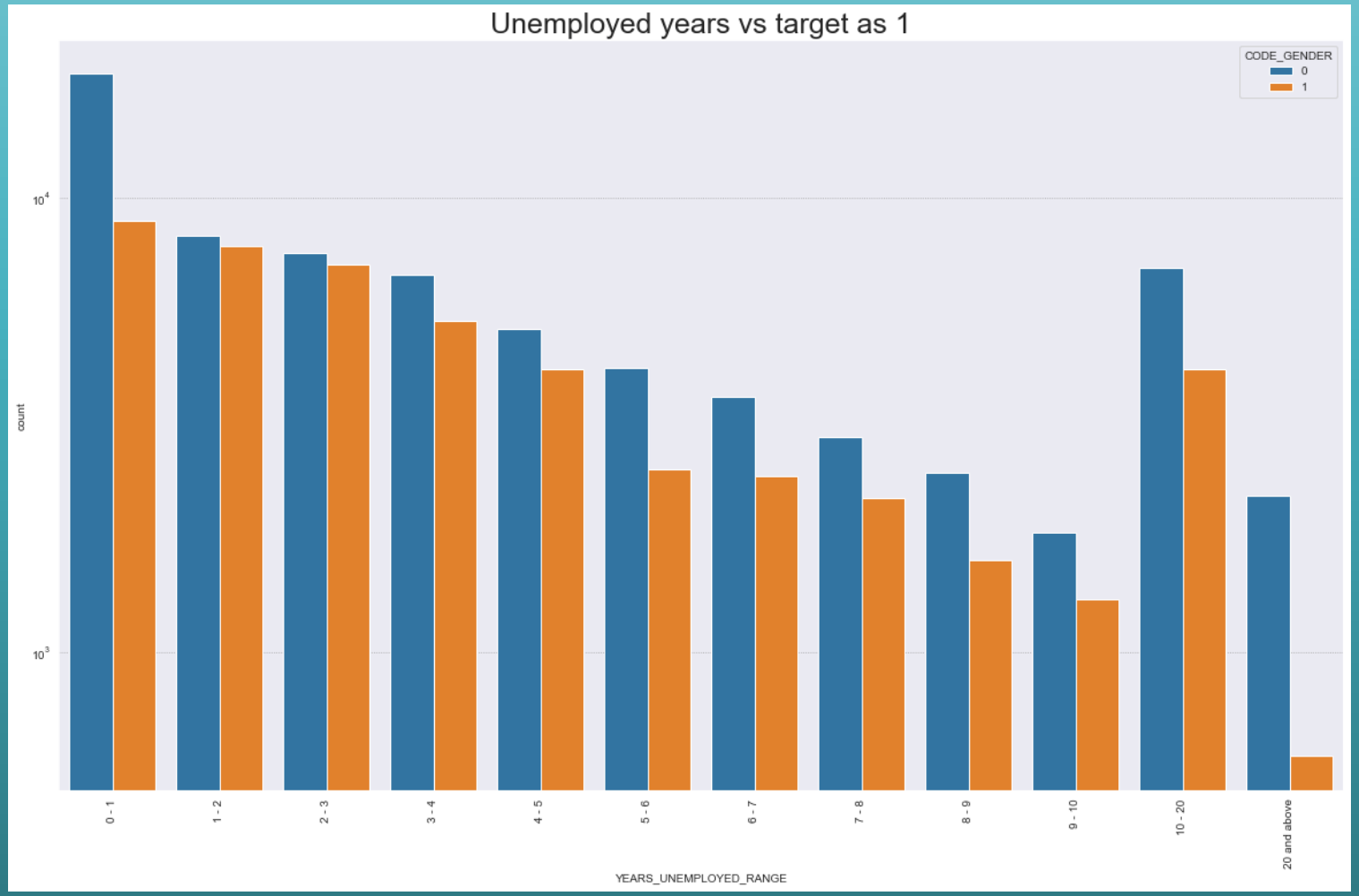
- For the People doing business if the plan and other factors seems to be okay there are high chances that they pay back loan.
- Same goes for Self-employed. They only required to start a small business for themselves.
- People working in Industry will be getting a fixed income and loans will be given accordingly low only since they cant pay back very soon might need long term duration.



Unemployed years vs target as 1

Inference:

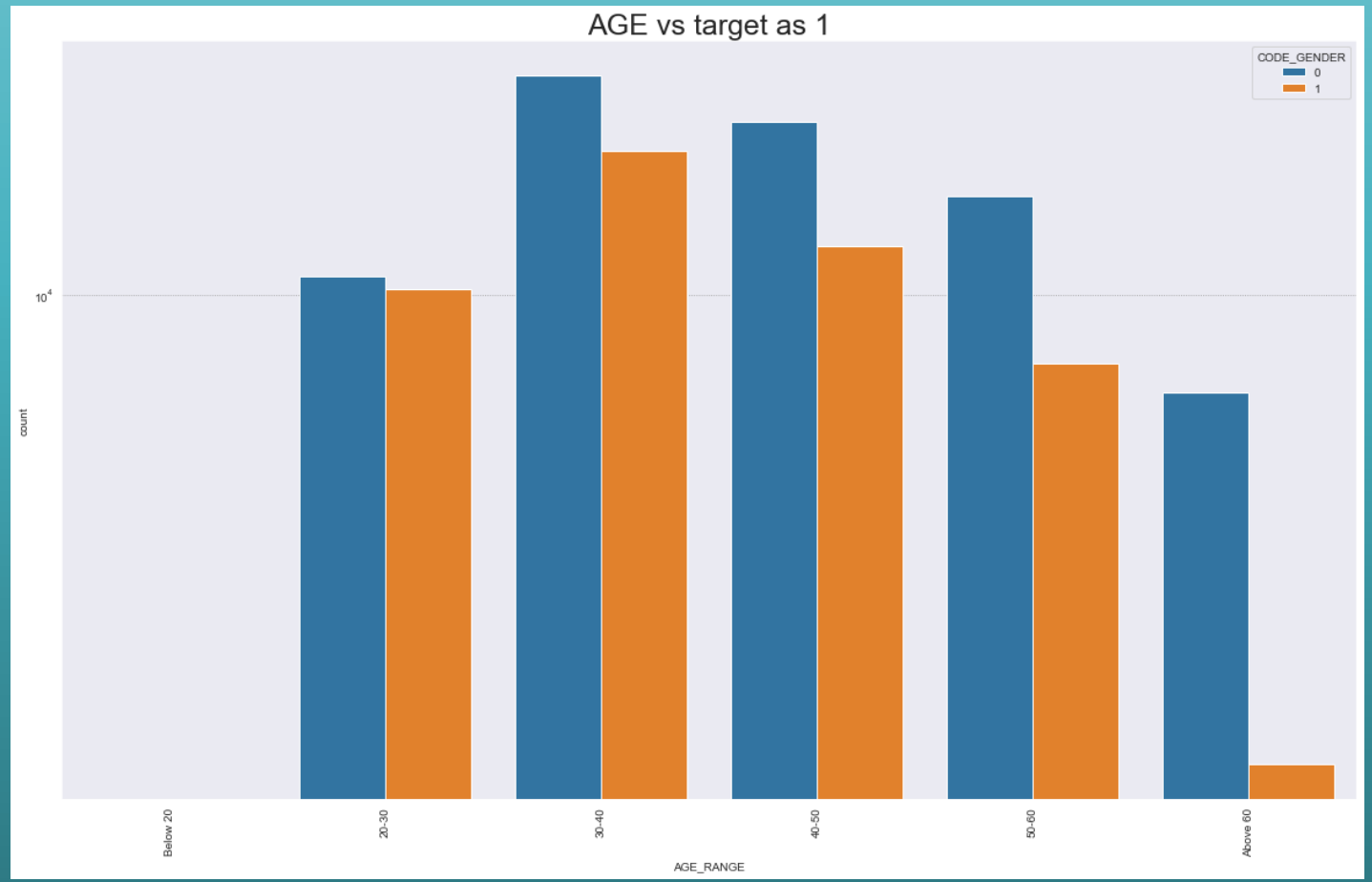
- As the number of unemployed years increases the changes of risk of repayment is high also they wont be provided with loans and so application is itself low in numbers.



AGE vs target as 1

Inference:

- The Age graph is somewhere like bell curve where the people of center age like above 50 nearing the retirement age they have low need of loans.
- Keeping in mind the Male count is half less than the female count in dataset, we can get to know that here almost with age both genders have risk equally.

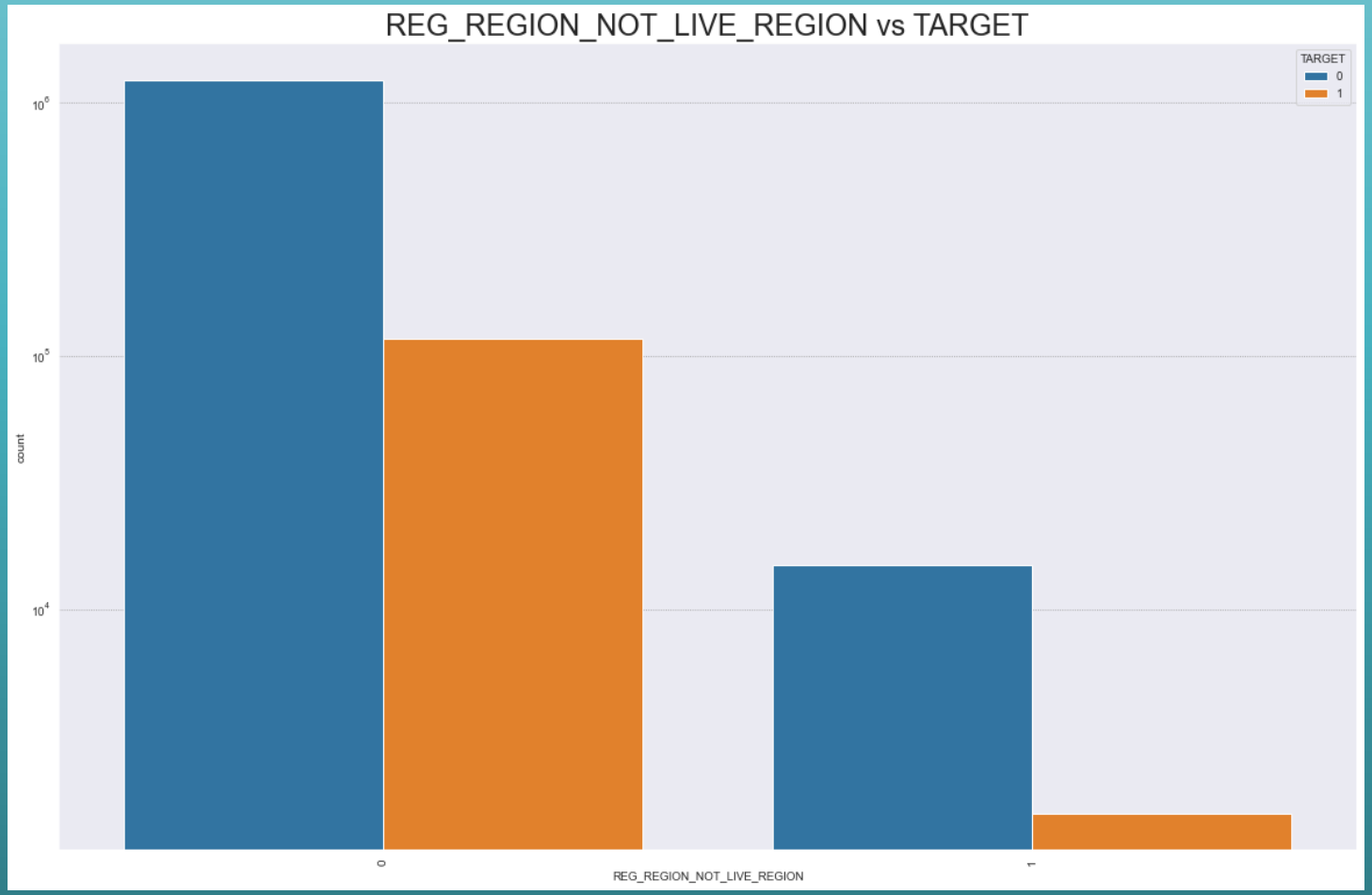


Columns vs Target

Given address not same as current address

Inference:

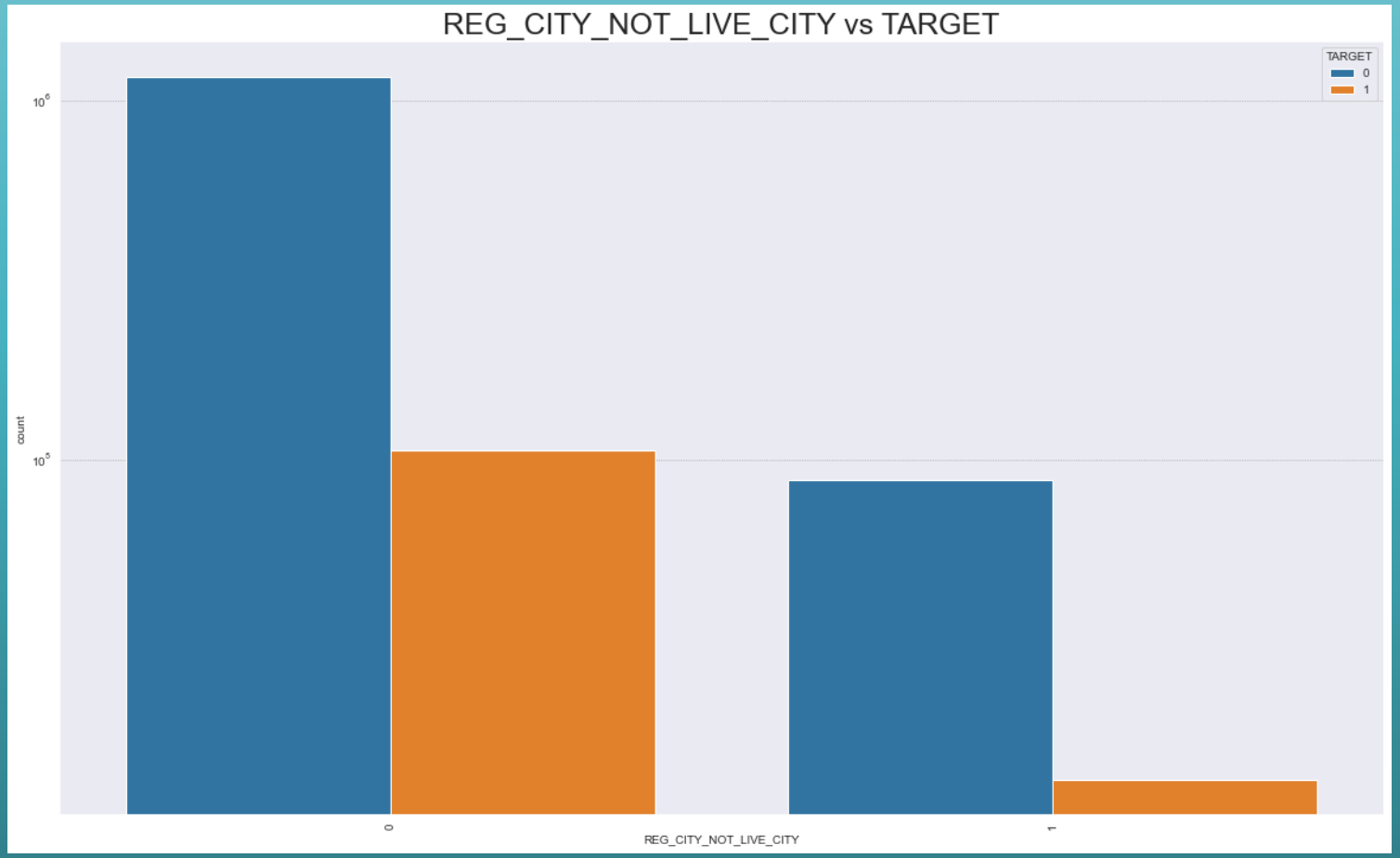
- This can confirm that when your current address not same as the given address, there might be some kind of fraudulent activity and hence the chances of it being 0 is obvious than 1.



Given City not same as current City

Inference:

- When the Permanent address (Region/City) does not match with contact address it says that there is high chances of risk and loan process will not be able to progress.



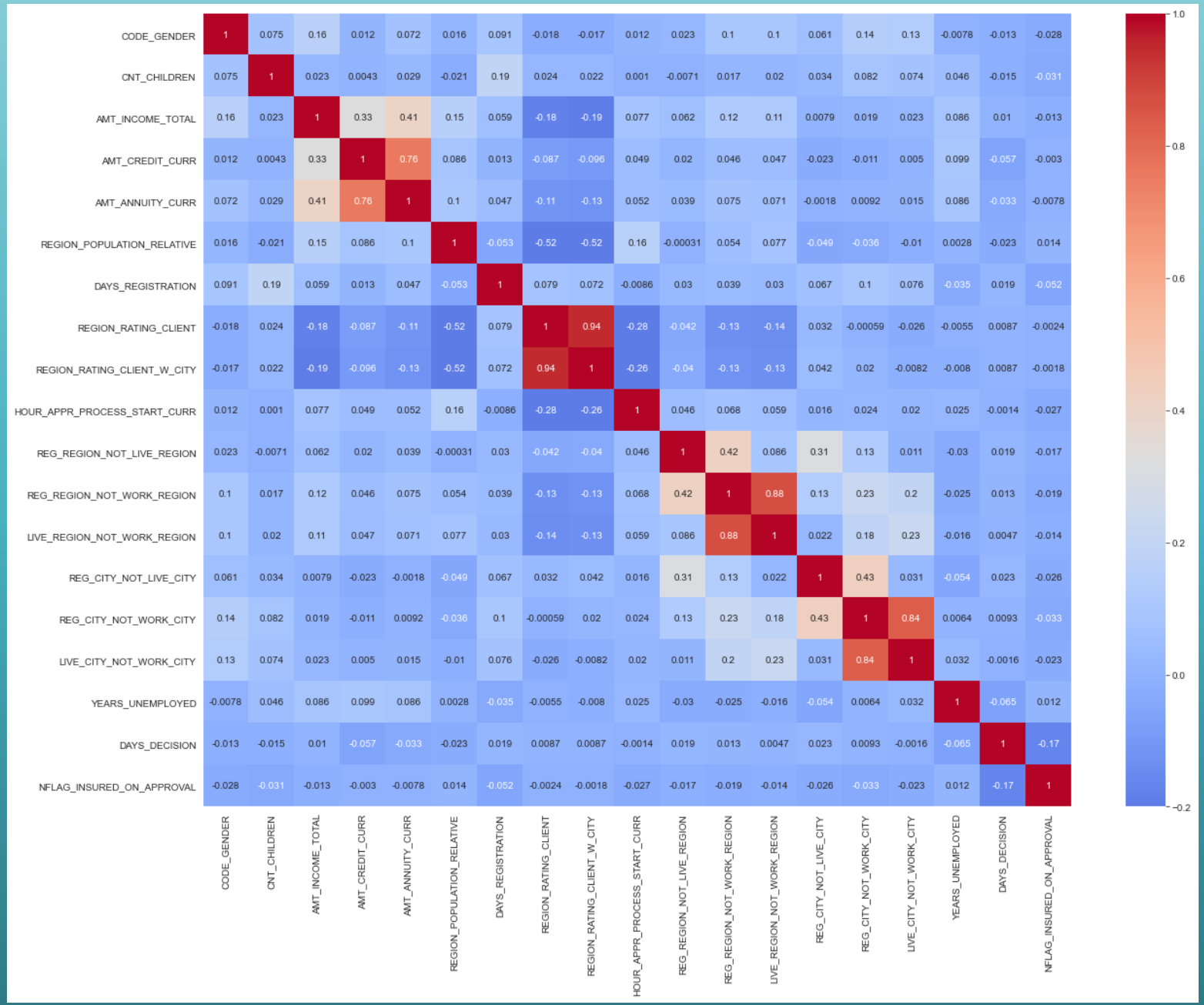
Correlation of Target 0 data

- Due to Data imbalance in the Target column the total data is split into two by the column target by 0's in one table and by 1's as another table

Inference:

- From the above we could see that there is no gender bias for application of loan or rejection of loan.
- The majority of the correlation occurs at Income amount, Credit amount and annuity current & previous
- The other group of high correlation is the region where the people live.
- So the above are the factors contributing relationship with each other.

Note : The Value is on Gender



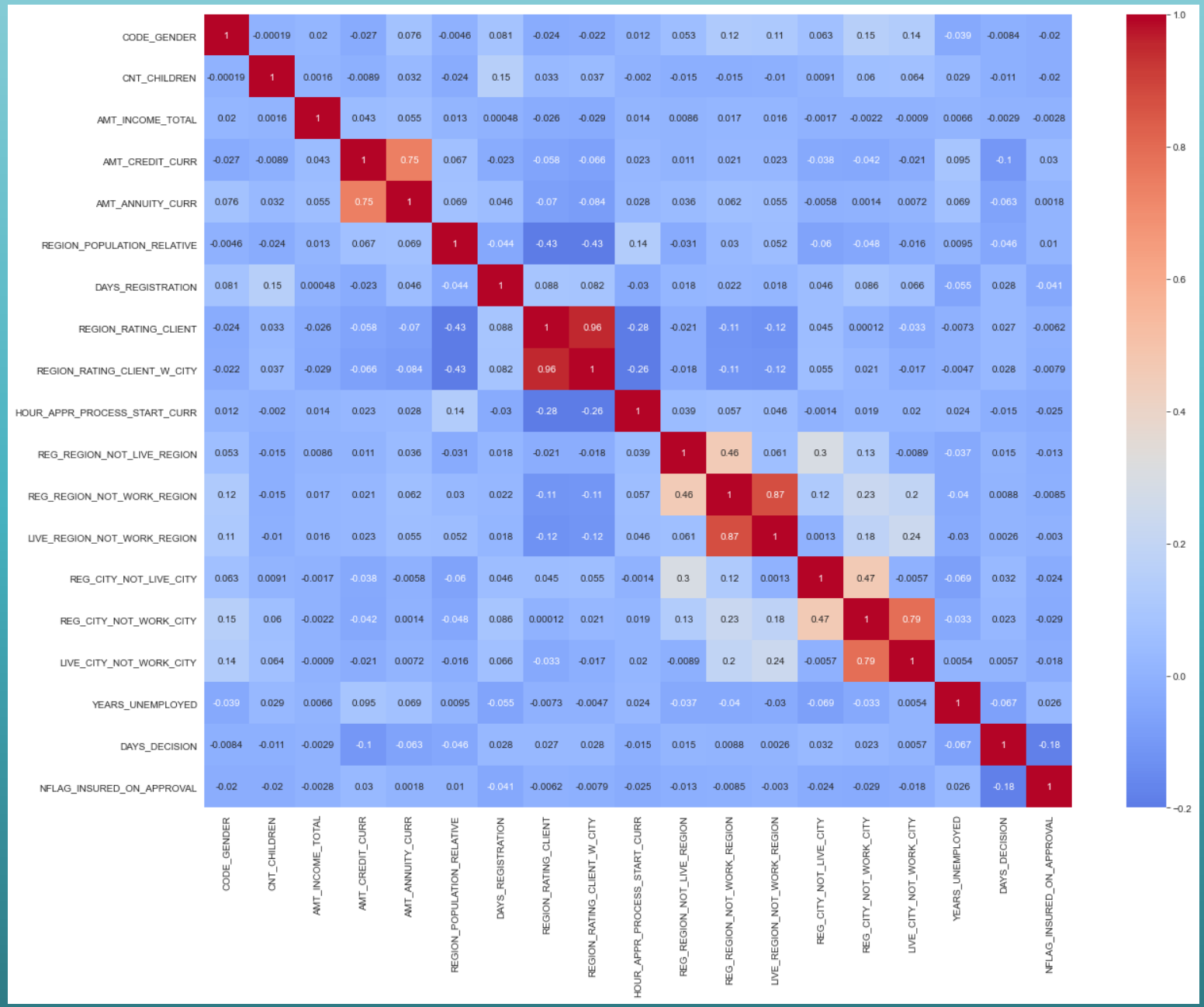
Correlation of Target 1 data

- Due to Data imbalance in the Target column the total data is split into two by the column target by 0's in one table and by 1's as another table

Inference:

- From the above we could see that there is no gender bias for application of loan or rejection of loan.
- The majority of the correlation occurs at Income amount, Credit amount and annuity current & previous
- The other group of high correlation is the region where the people live.
- So the above are the factors contributing relationship with each other.

Note : The Value is on Gender



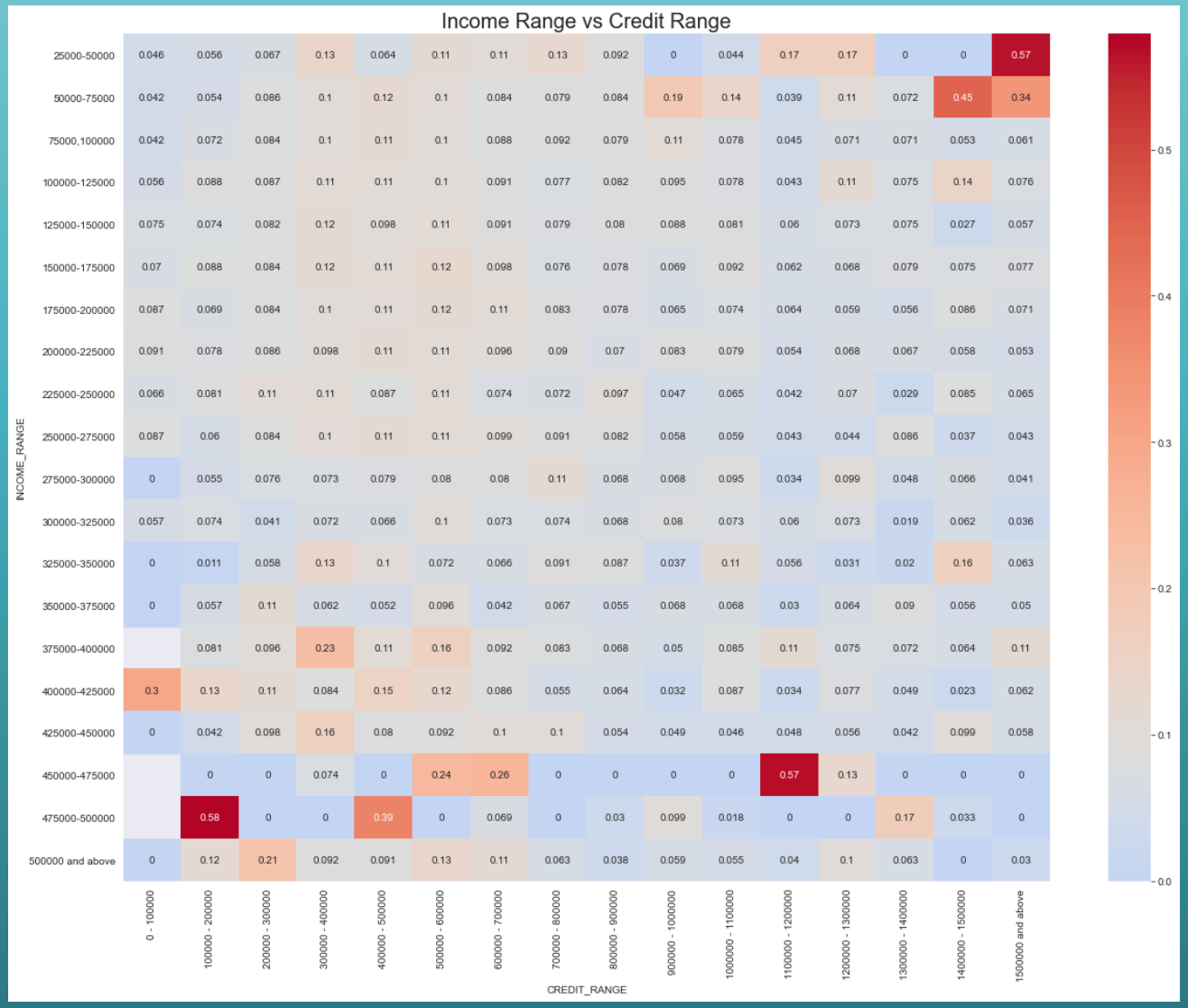
Multivariate analysis

Income Range vs Credit Range

Inference:

- This Plot shows that if the credit amount is low then there are high changes that the target is 1.
- Also for the people with high income the Target 1 changes are high.

Note : Pivot on Target



Occupation Type vs Credit Range

Inference:

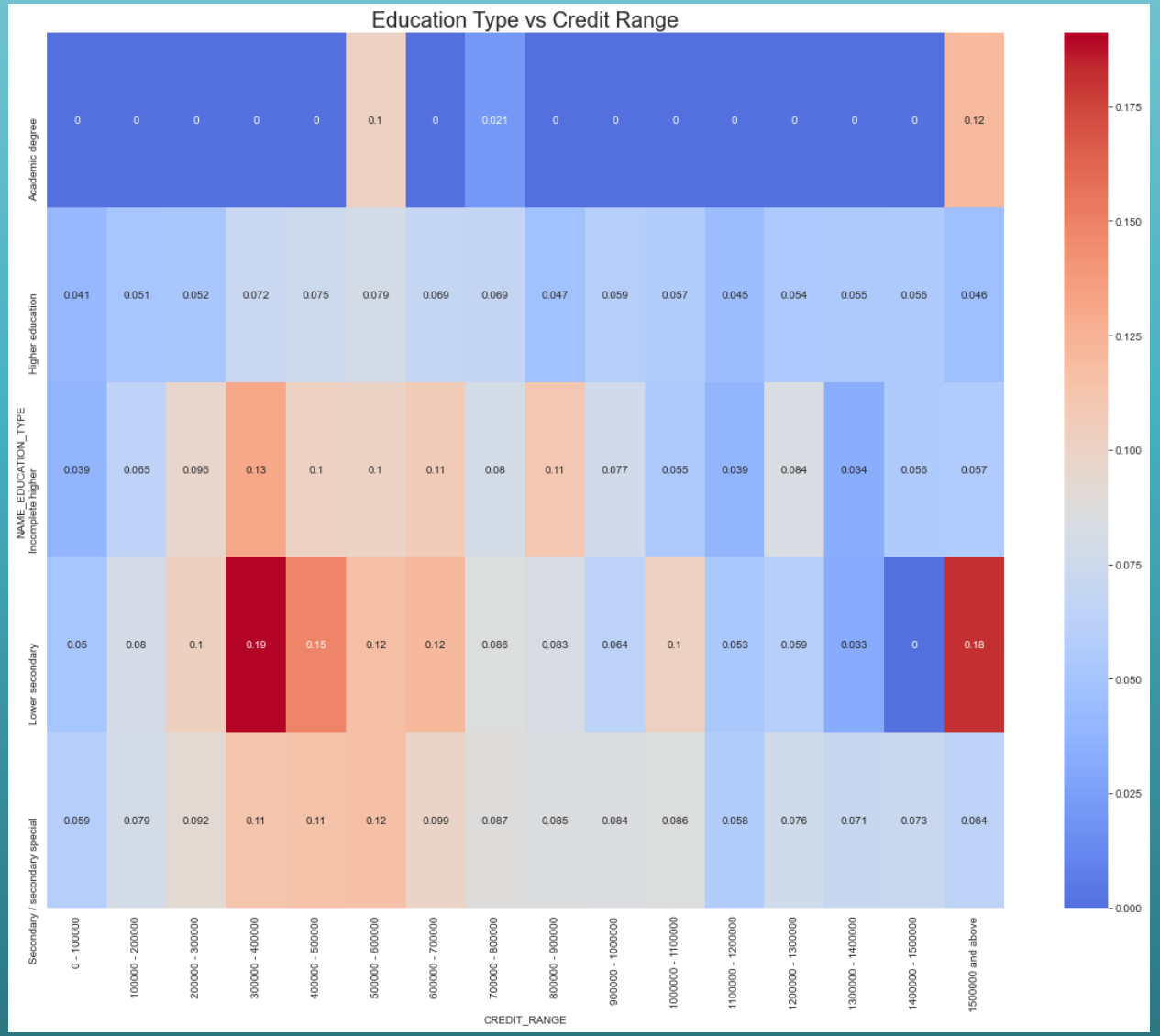
- Since the Target values are imbalance mode by 10 times we could see that there is no proper relation displayed.
- But we could observe a pattern of IT staff, and HR staff has little higher probability that risk is low.
- Also again the lower the credit rate lower the risk.

Note : Pivot on Target

Education Type vs Credit Range

- There is high correlation for the lower secondary education with the credit range overall.
- The people having an Academic degree seem to have less correlation with Credit amount on Target variable.

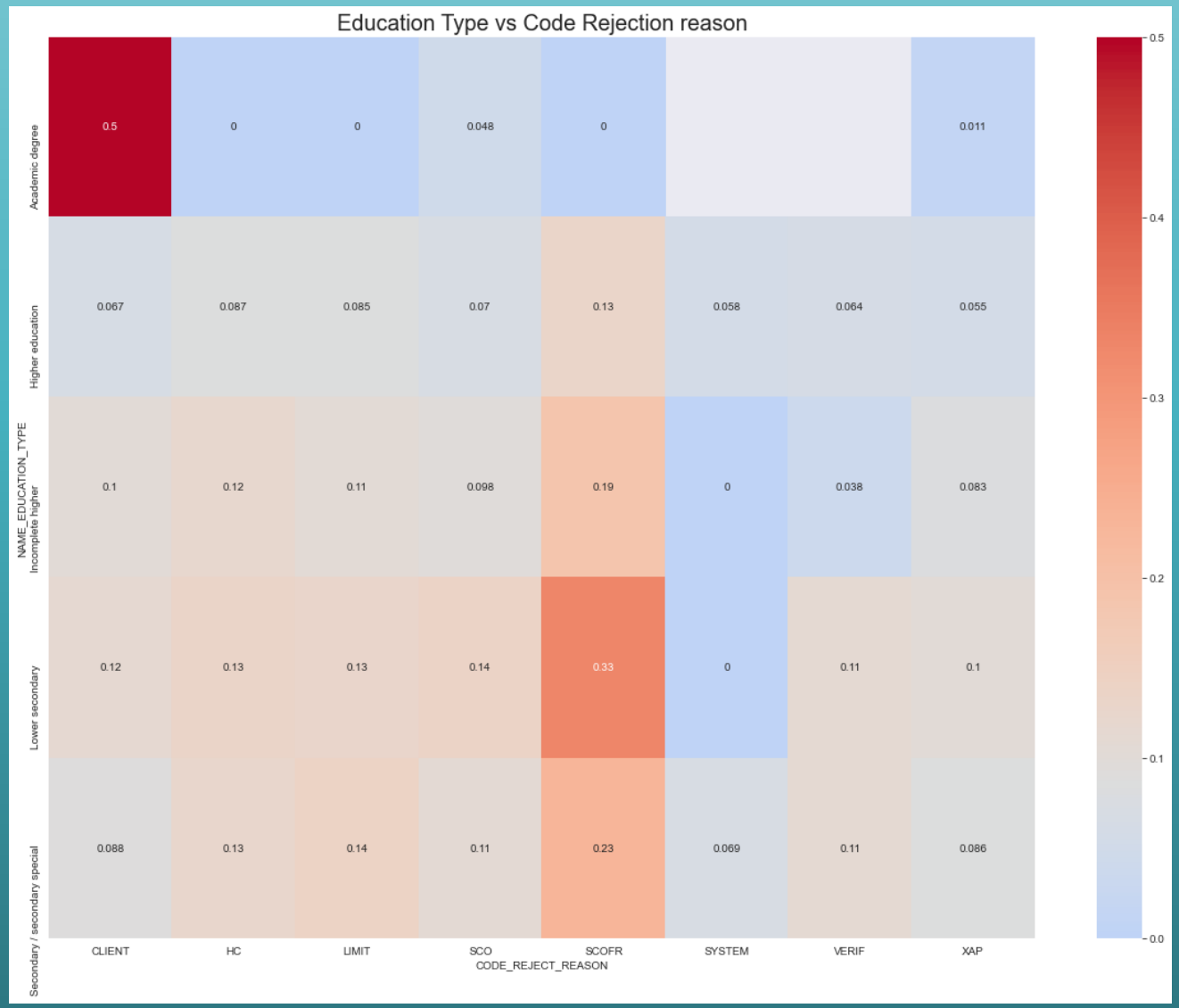
Note : Pivot on Target



Education Type vs Code Rejection reason

Inference:

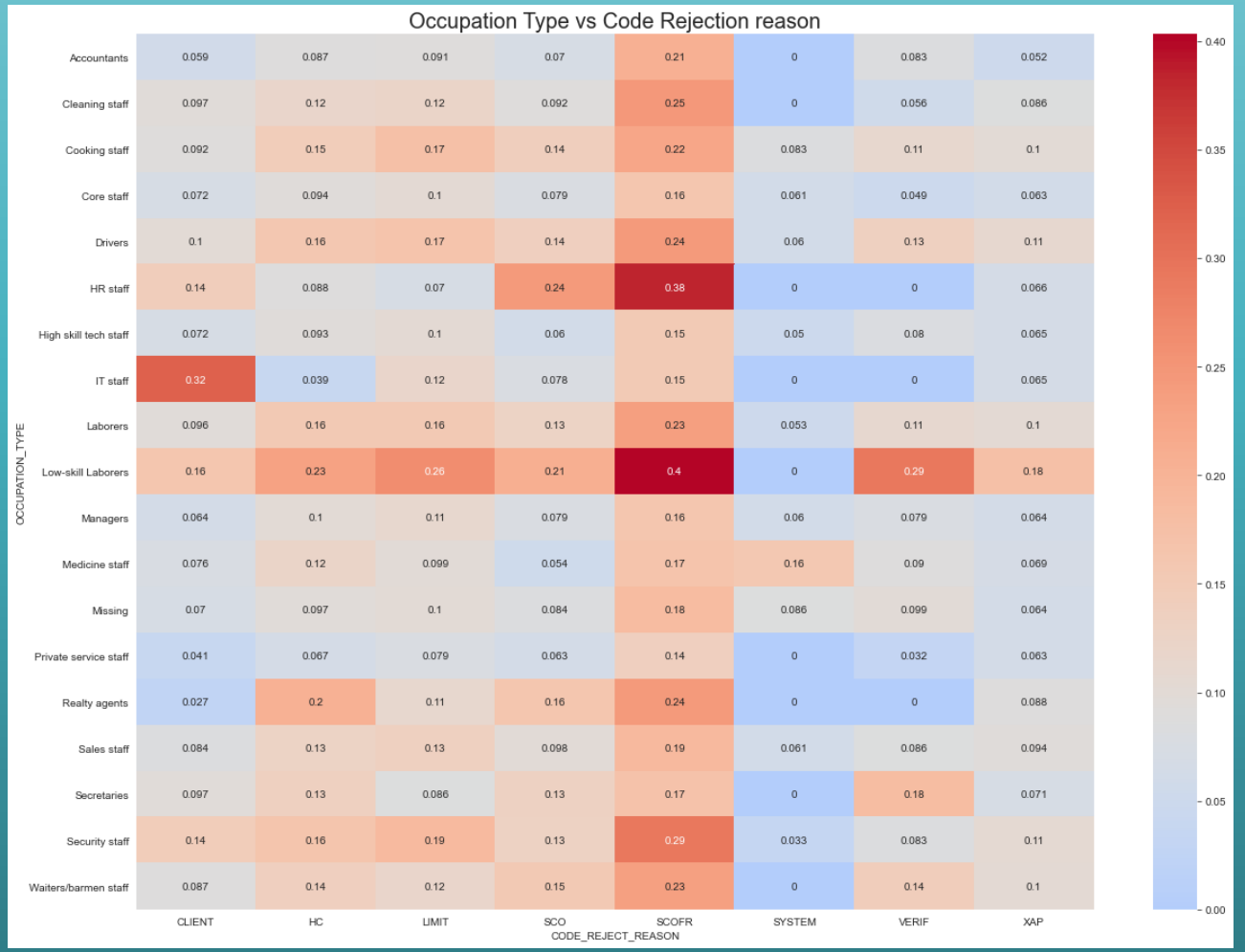
- The lower secondary education have high correlation with code reject reason SCOFR.
- Also same with Academic degree with Client the highest correlation.



Note : Pivot on Target

Occupation Type vs Code Rejection reason

- There is High correlation for the SCOFR(cibil score) with all occupation type.
- Lower Left corner has high correlation of over all plot.

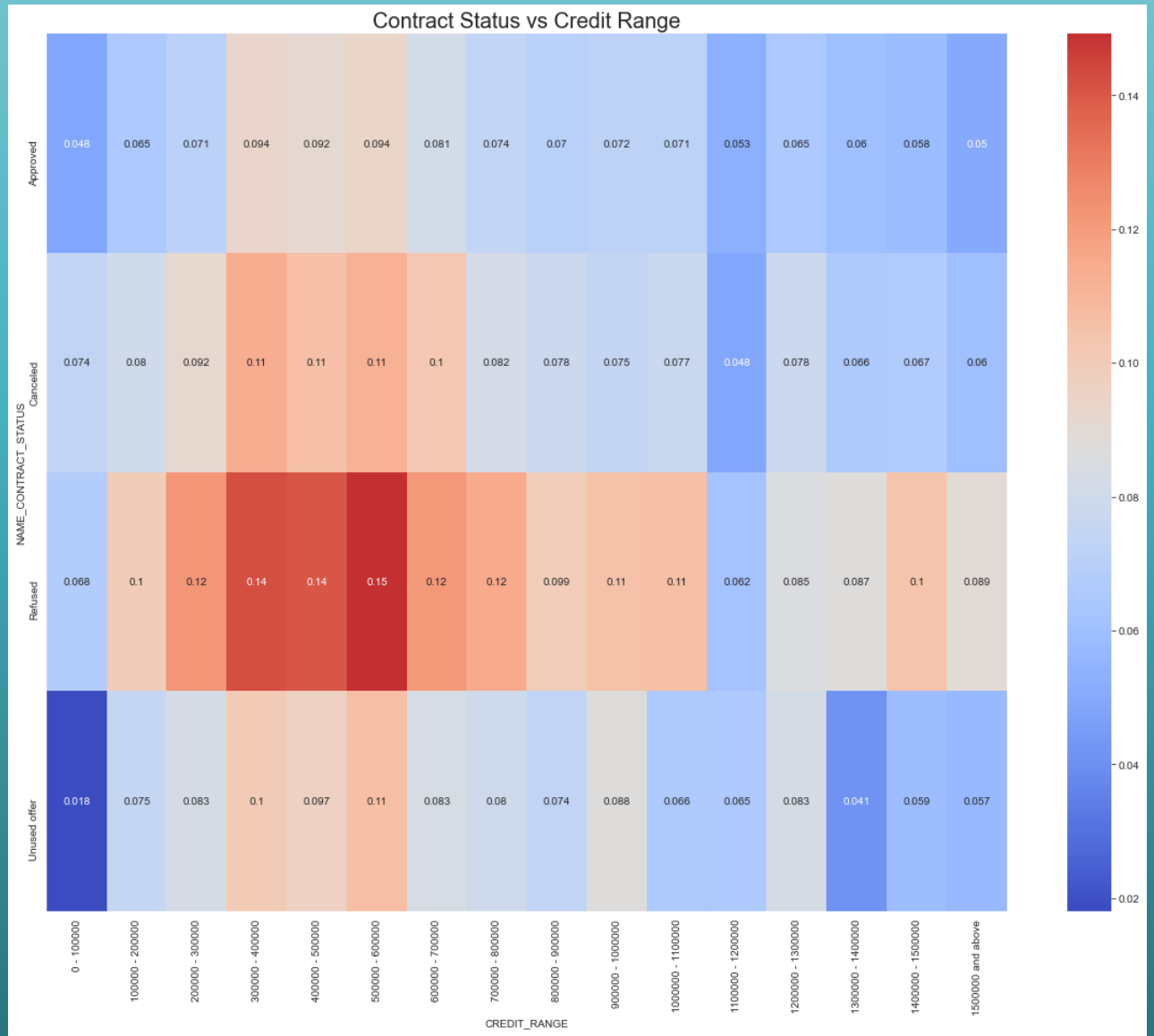


Note : Pivot on Target

Contract Status vs Credit Range

Inference:

- The Refused candidates have high correlation with all kind of credit range.
- High correlation is almost in the middle only.



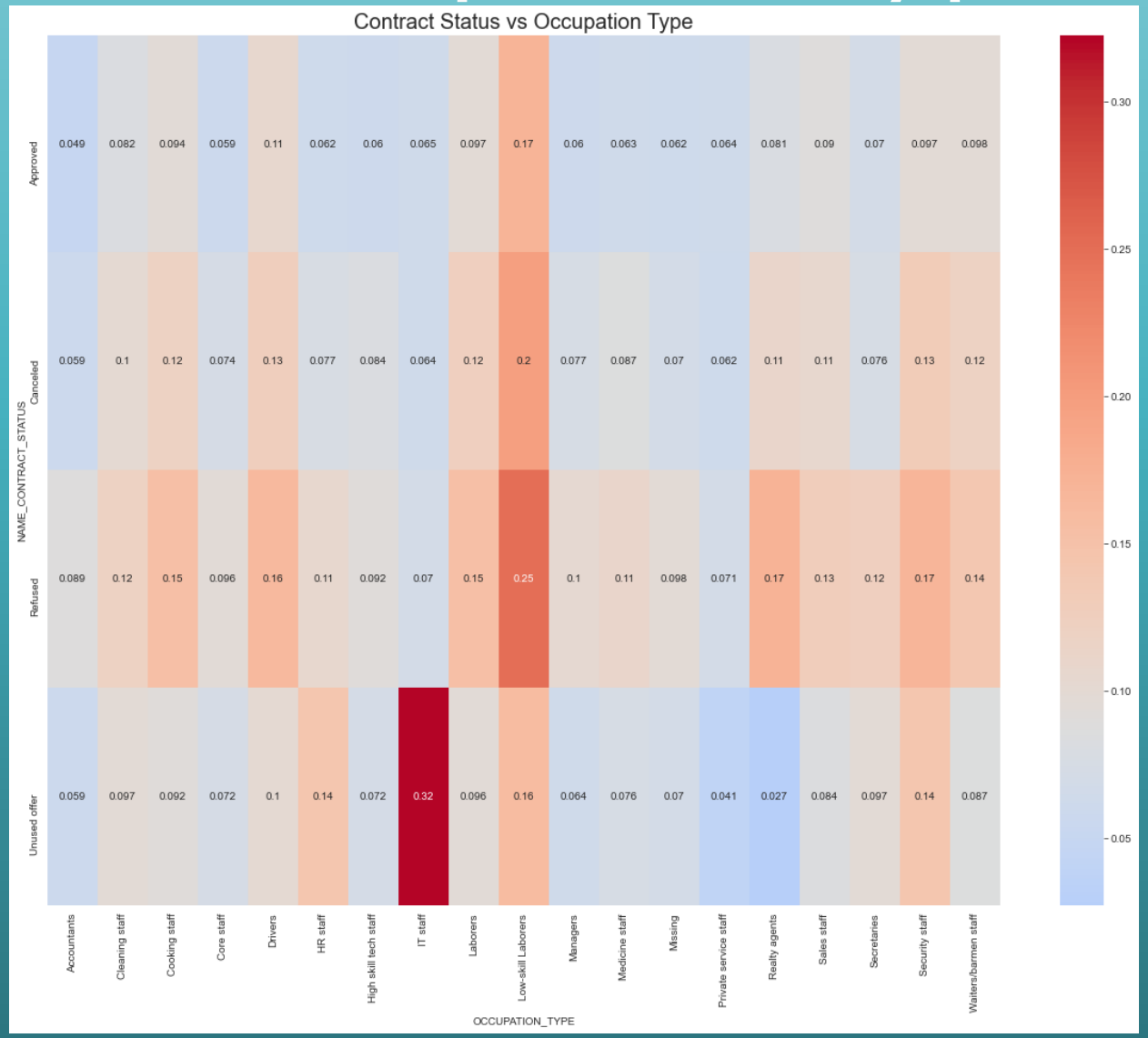
Note : Pivot on Target

Contract Status vs Occupation Type

Inference:

- Low Skill labourer's have high correlation with all the contract status giving target 1.
- The same goes for Refused category in the contract status against occupation type.

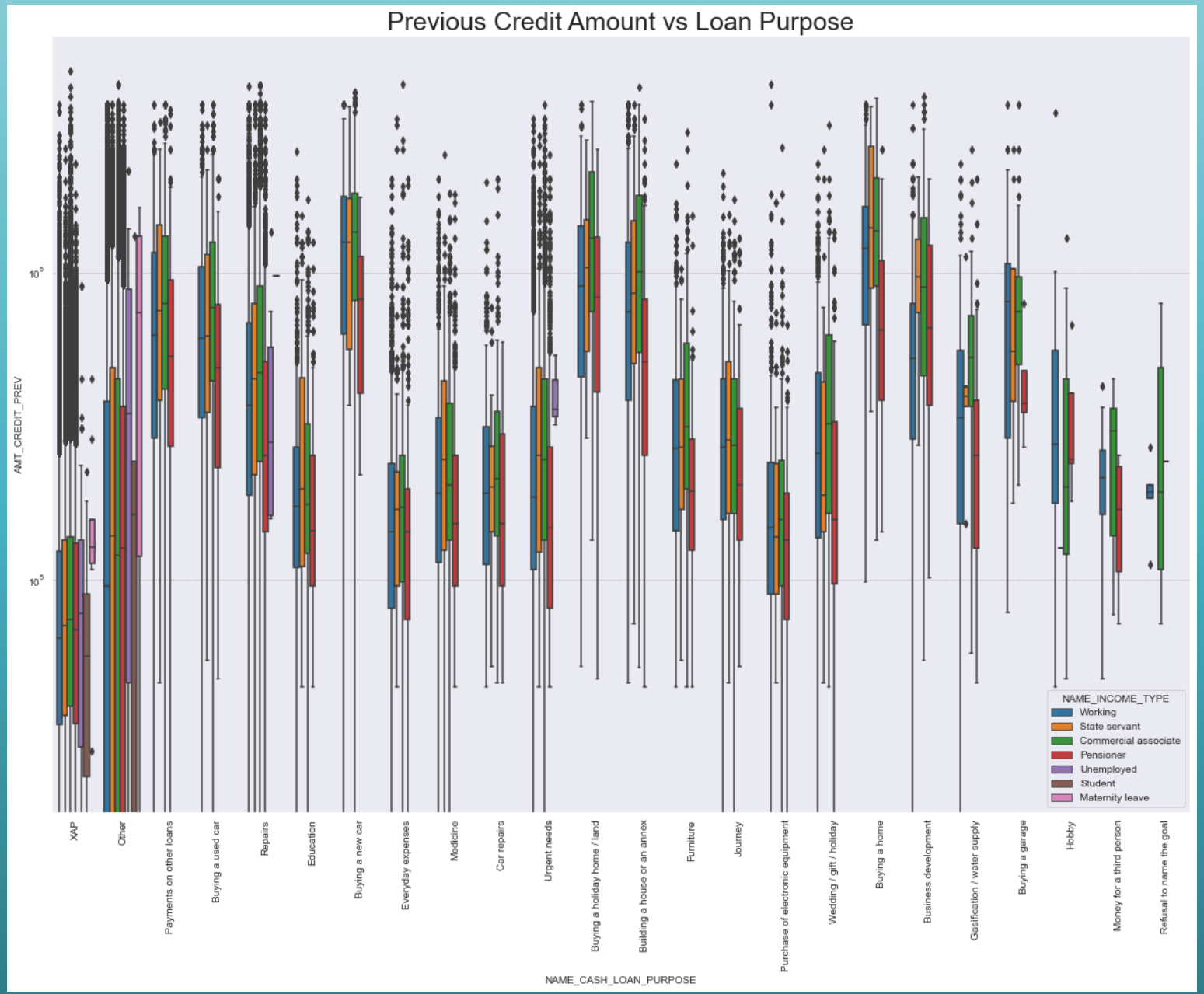
Note : Pivot on Target



Previous Credit Amount vs Loan Purpose

Inference:

- For the cash loan purpose Other the mean seems to be low but it has a wide spread amount credit.
- Mean's of Buying a car, Buying a Home, Furniture, Business development are other necessary items whose Amount credit range is high for all income type.



CONCLUSION

- Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ for successful payment and focus less on income type ‘Working’ as they are having most number of unsuccessful payments.

- Also with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.

- For Academic range education also there seems to be less loan approval rate and that can also be considered fro bank side for successful loan rate incomes.

Thankyou