Credit EDA Case Study : Exploratory Data Analysis on Bank Loan Data

Presented by

Md Irshad Ali

Problem Statement – I

- **Business Understanding**
- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Problem Statement – II

- **Results Expected**
- Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.
- Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)
- Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
- Identify if there is data imbalance in the data. Find the ratio of data imbalance.

- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
- Find the top 10 correlation for the **Client with payment difficulties** and **all other cases** (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: **Var1, Var2, Var3, Var4, Var5, Target**. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.
- Include visualisations and summarise the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the **clients with payment difficulties with all other cases.**

Two data data set given

- For the purpose of this case study, two data sets were provided namely:
- *application_data.* contains all the information of the client at the time of application.
  The data is about whether a **client has payment difficulties.**
-
- *2. 'previous_application.* contains information about the client’s previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

Step follow:

- First, we took the Application Dataset for analysis.
- Importing all necessary libraries.
- Uploading Application dataset.
- Checking dataset of all rows and columns.
- Checking for values to impute in columns.
- Now, doing binning variables for analysis.
- Checking for imbalance in Target.
- Working on Univariate Analysis Bivariate Analysis.
- Getting the top 10 correlation of the selected columns.
- Working on previous application dataset.
- Doing all the imputing and analysis in previous application dataset.
- Merging the Application data and Previous dataset.
- Analysis the data and observing the pattern.
- Conclusion and Recommendation .

Data Analysis For Application Data

First checking the null values of the columns.

As observed that null is more then 50% in most of the columns.

Now removing all the columns which having more than 50% null values. After removing then got the (307511, 81) rows and columns.

Also check the 15% of missing the values.

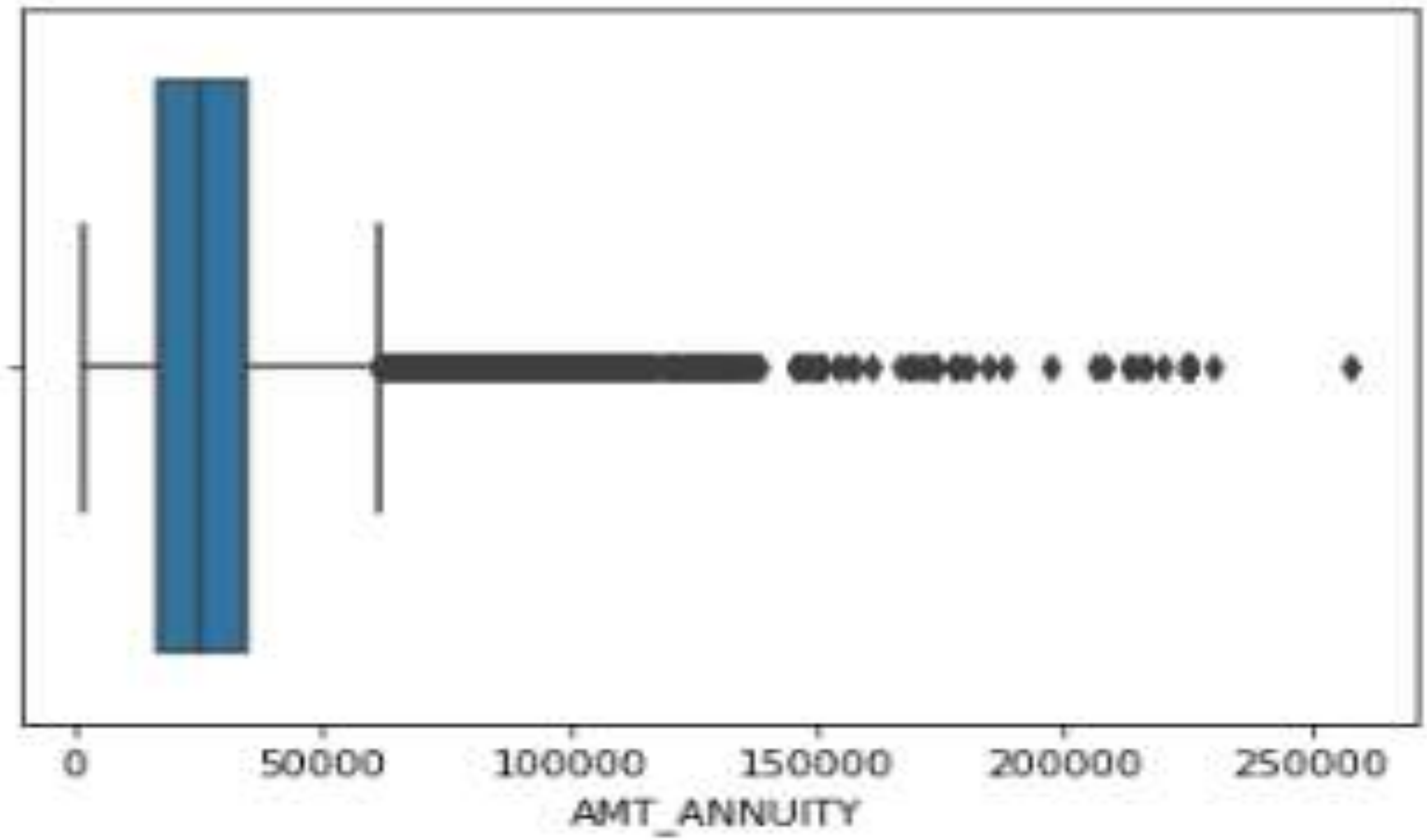
Now describing the data to check the min, median, mode, std-div values.

Checking for values to impute in columns

1. **OCCUPATION_TYPE imputation**
2. Observed that **96391** is total values is counted.
3. Now treating the null values by filling null values with unknow label in rows.

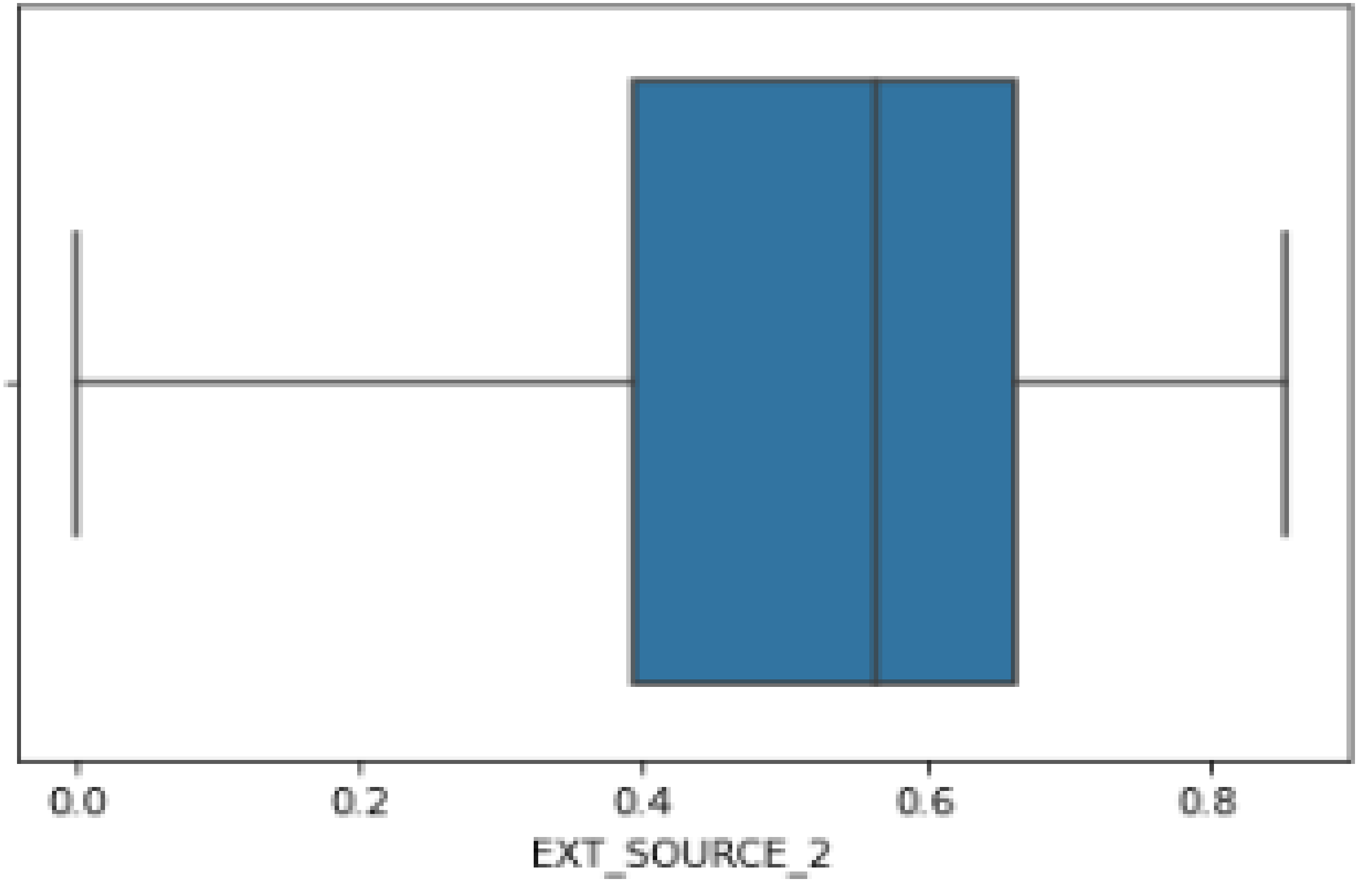
AMT_ANNUITY imputation

- Checking null values which found 12
- Finding the outlier and filling with median value 24903.0



EXT_SOURCE_2 imputation

- Checking null values which is found 12
- No outlier found and filling with mean value 0.51

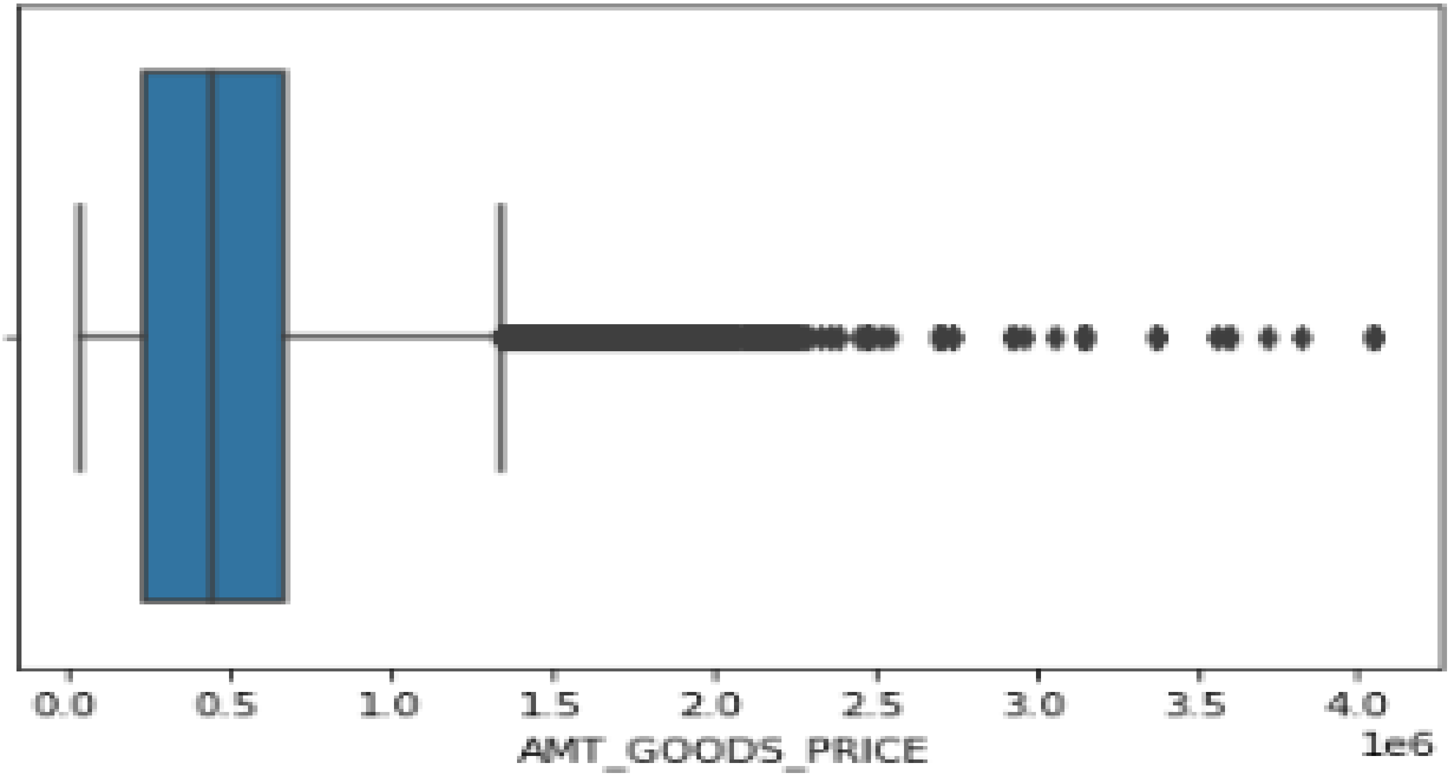


NAME_TYPE_SUITE Imputation

- It is Categorical column.
- Null value is 1292.
- Imputing the mode values because it is categorical column with unaccompanied.

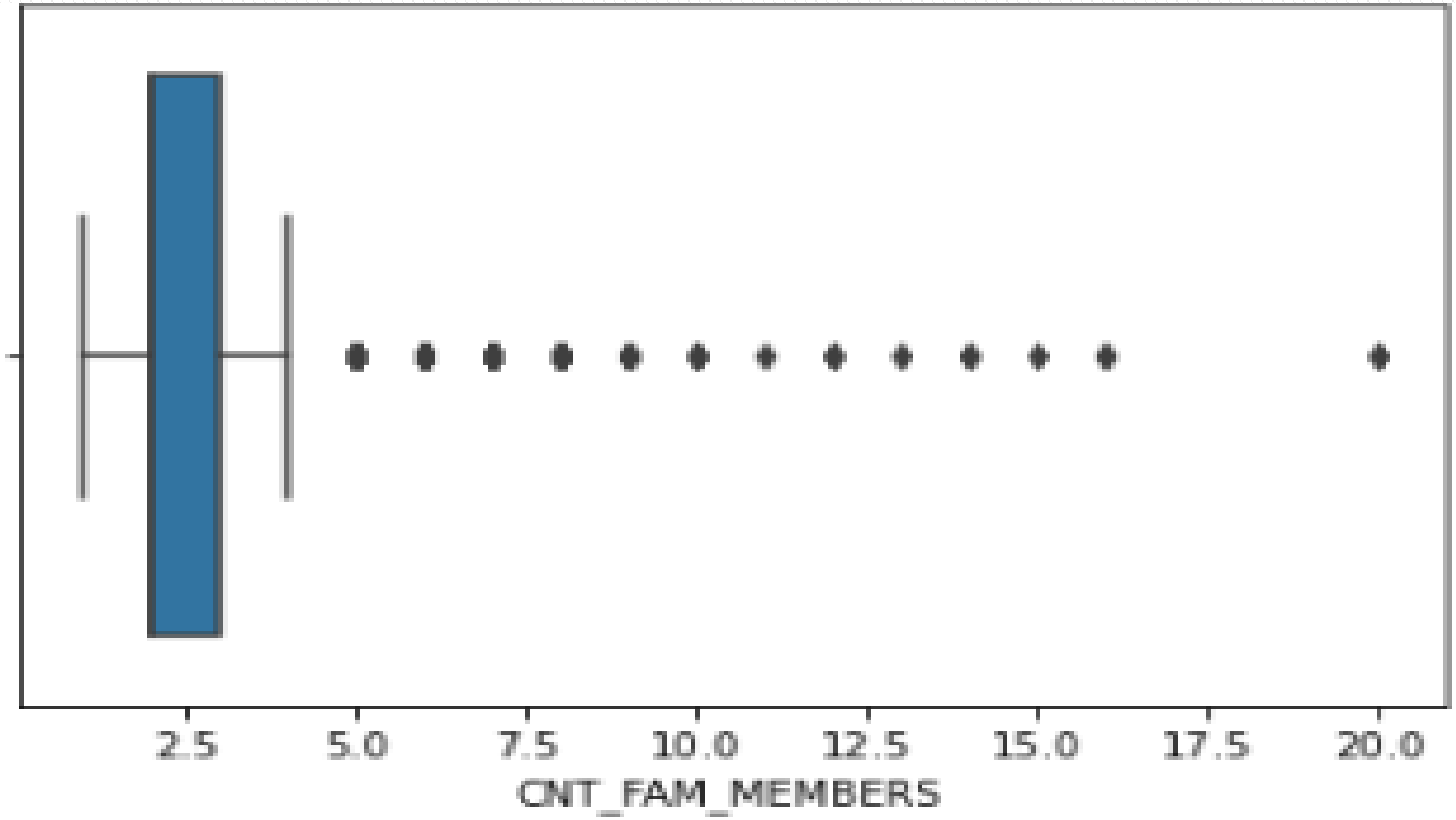
AMT_GOODS_PRICE imputation

- Checking null values which is found **278**
- Outlier found and filling with median value 450000.
- This continuous column so imputing median values

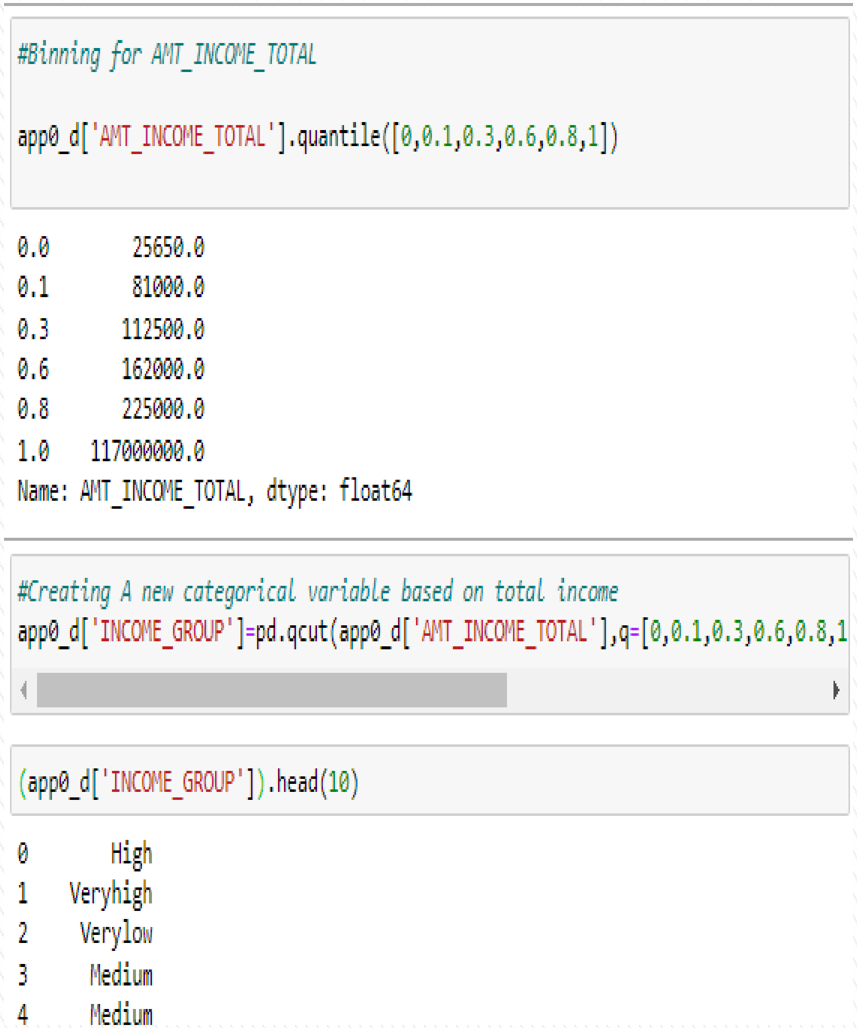


CNT_FAM_MEMBERS Imputation

- Checking null values which is found **2**
- Outlier found and filling with median value.
- This continuous column so imputing median values.



Analysis for Binning variables



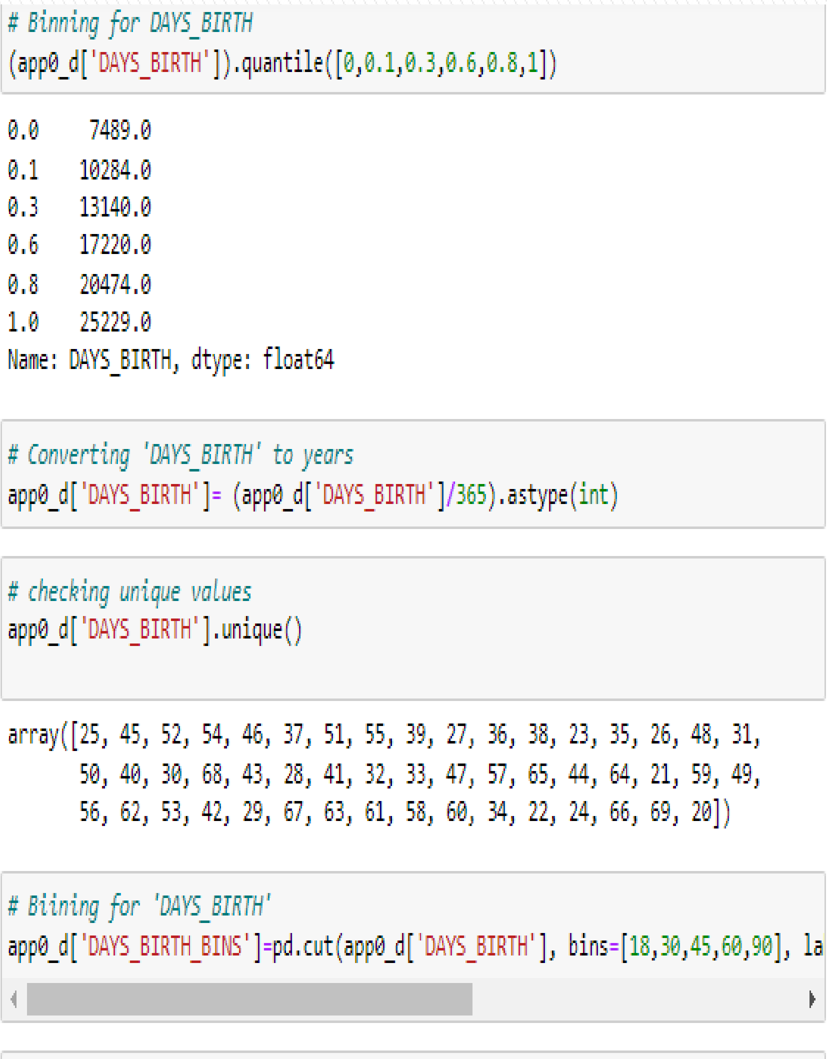
```
#Binning for AMT_INCOME_TOTAL

app0_d['AMT_INCOME_TOTAL'].quantile([0,0.1,0.3,0.6,0.8,1])
```

```
0.0        25650.0
0.1        81000.0
0.3       112500.0
0.6       162000.0
0.8       225000.0
1.0    117000000.0
Name: AMT_INCOME_TOTAL, dtype: float64
```

```
#Creating A new categorical variable based on total income
app0_d['INCOME_GROUP']=pd.qcut(app0_d['AMT_INCOME_TOTAL'],q=[0,0.1,0.3,0.6,0.8,1
```

```
(app0_d['INCOME_GROUP']).head(10)
```

```
0        High
1    Veryhigh
2     Verylow
3      Medium
4      Medium
```



```
# Binning for DAYS_BIRTH
(app0_d['DAYS_BIRTH']).quantile([0,0.1,0.3,0.6,0.8,1])
```

```
0.0     7489.0
0.1    10284.0
0.3    13140.0
0.6    17220.0
0.8    20474.0
1.0    25229.0
Name: DAYS_BIRTH, dtype: float64
```

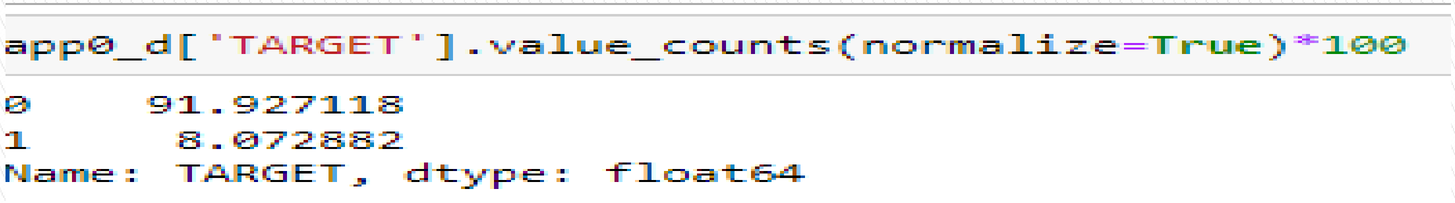
```
# Converting 'DAYS_BIRTH' to years
app0_d['DAYS_BIRTH']= (app0_d['DAYS_BIRTH']/365).astype(int)
```

```
# checking unique values
app0_d['DAYS_BIRTH'].unique()
```

```
array([25, 45, 52, 54, 46, 37, 51, 55, 39, 27, 36, 38, 23, 35, 26, 48, 31,
       50, 40, 30, 68, 43, 28, 41, 32, 33, 47, 57, 65, 44, 64, 21, 59, 49,
       56, 62, 53, 42, 29, 67, 63, 61, 58, 60, 34, 22, 24, 66, 69, 20])
```

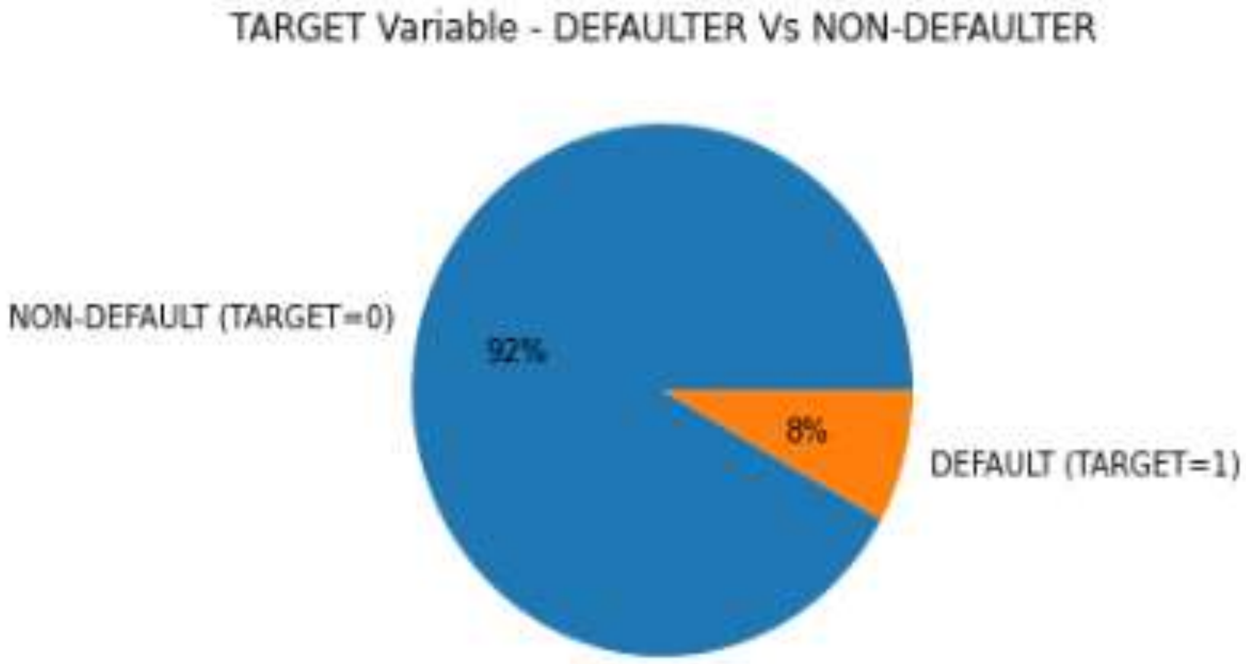
```
# Biining for 'DAYS_BIRTH'
app0_d['DAYS_BIRTH_BINS']=pd.cut(app0_d['DAYS_BIRTH'], bins=[18,30,45,60,90], la
```

Checking for imbalance in Target



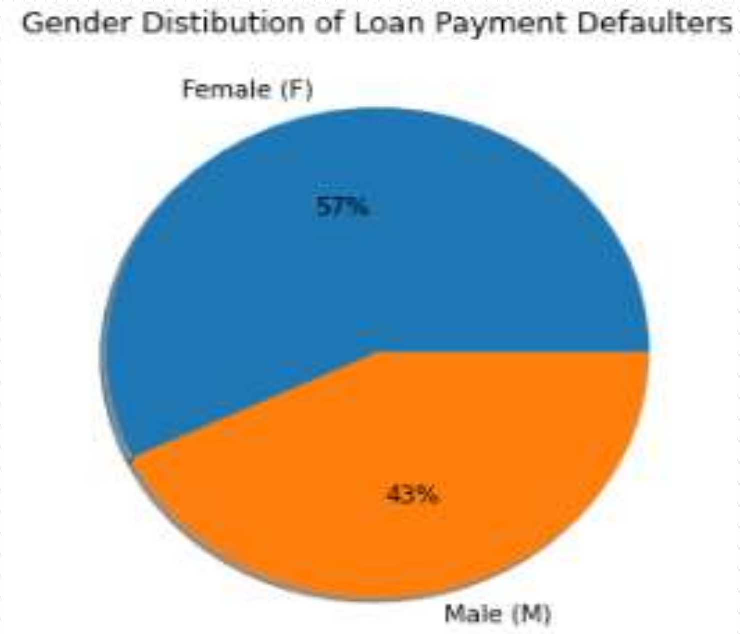
```
app0_d['TARGET'].value_counts(normalize=True)*100
```

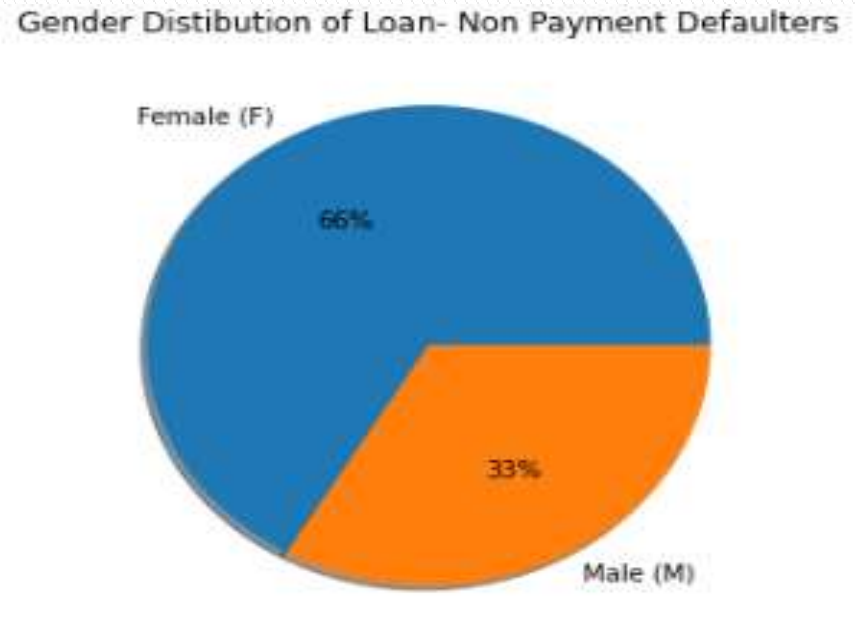
```
0     91.927118
1      8.072882
Name: TARGET, dtype: float64
```



More than 92% of people didn't default as opposed to 8% who defaulted. so it observed imbalance between people who defaulted.

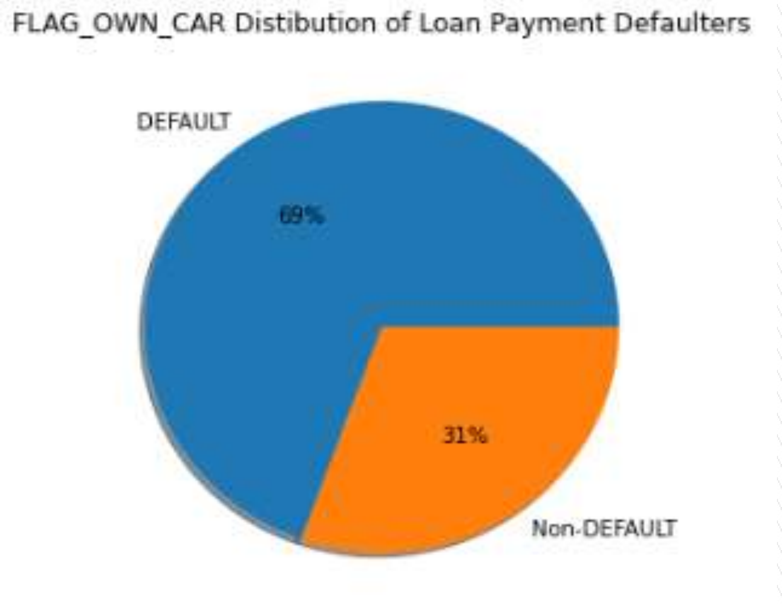
Splitting the original dataset into two different datasets depending upon the target value

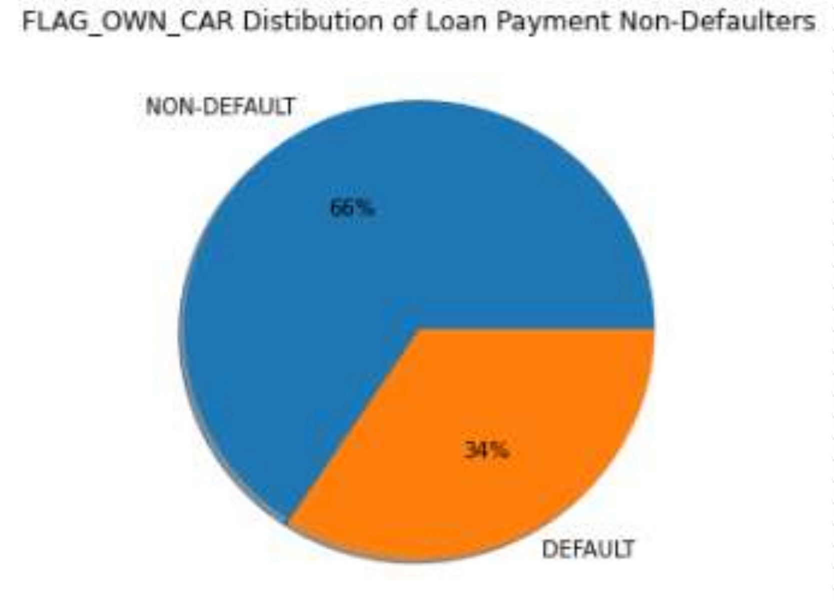




Female contribute 66% to the non-defaulters while 57% to the defaulters. We observed that more
female applying for loans than males and hence the more number of female defaulters as well.
But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts.

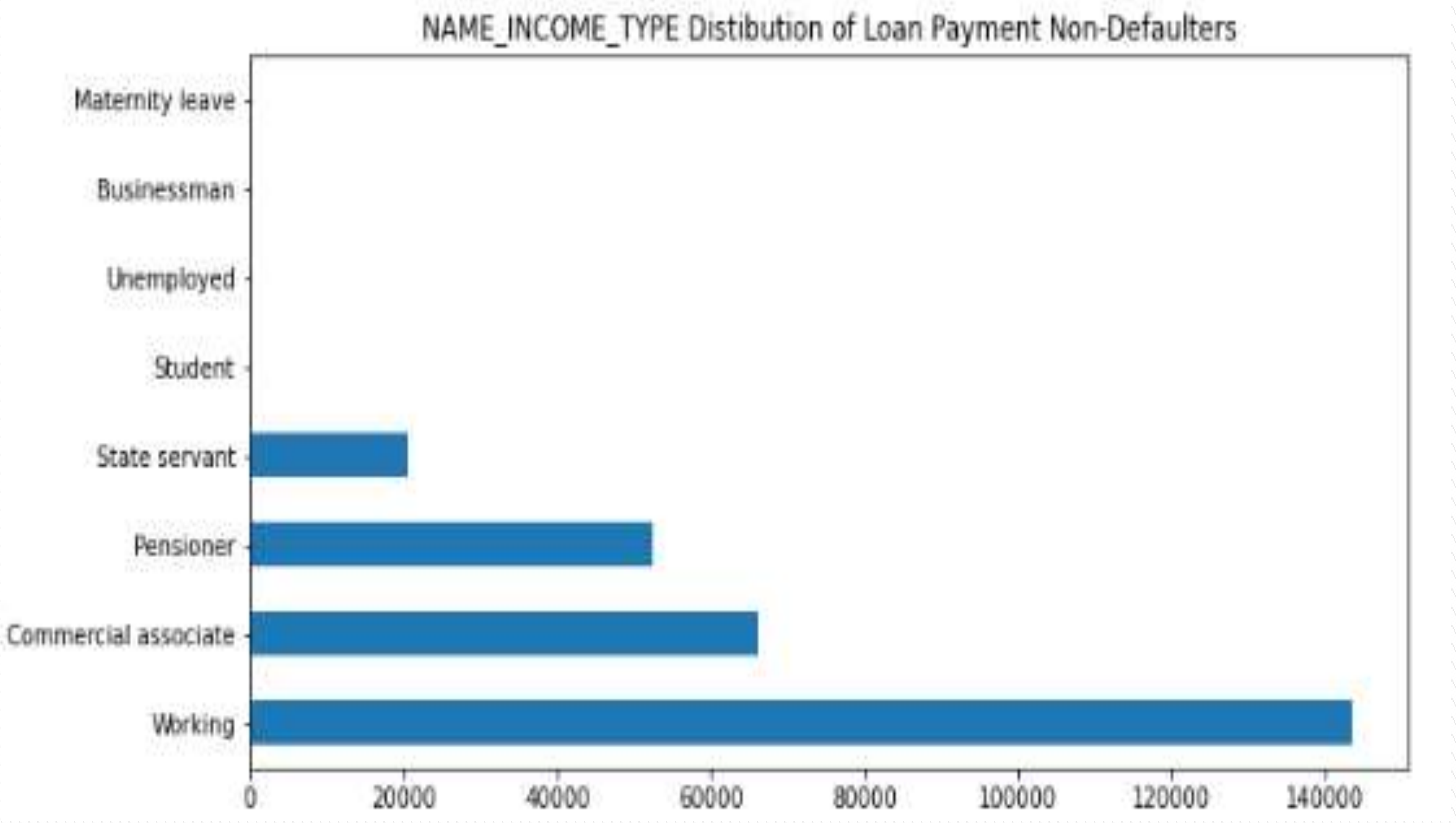
FLAG_OWN_CAR Distibution of Loan Payment Defaulters



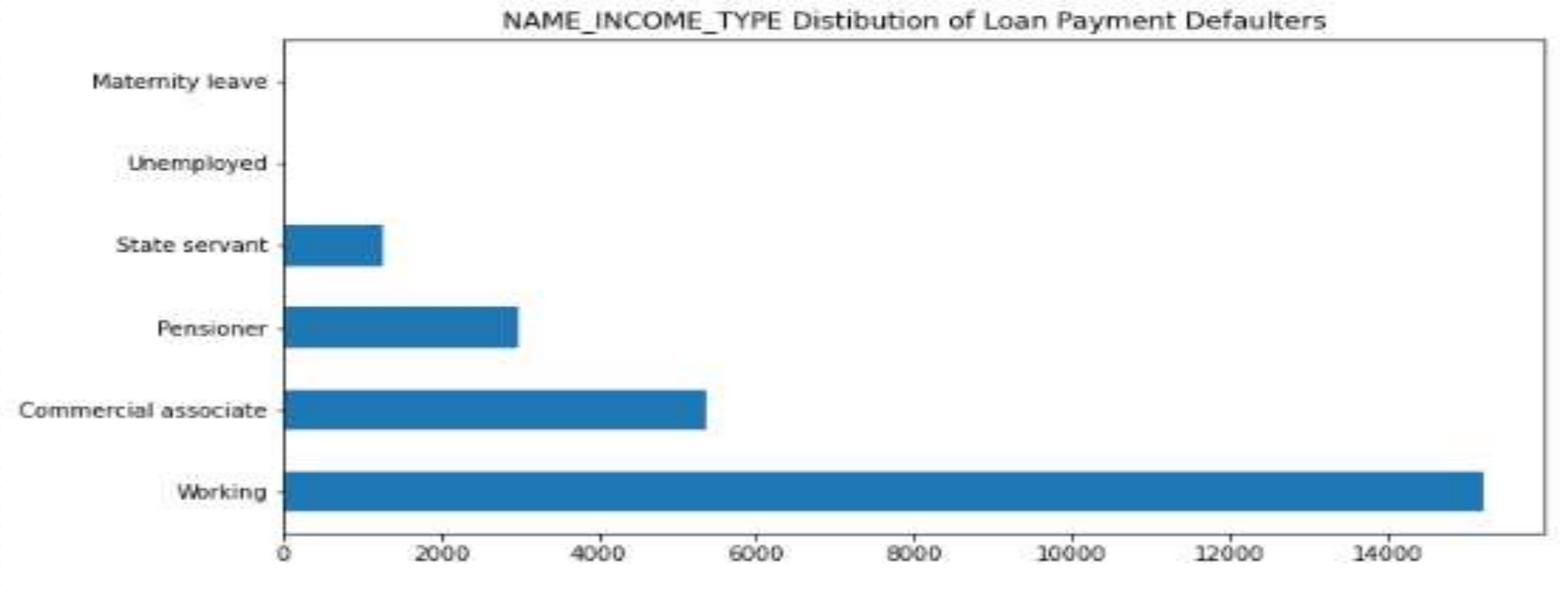


People with cars contribute 66% to the non-defaulters while 69% to the defaulters.While people who have car default more often,the reason could be there are simply more people without cars.
Looking at the percentages in both the charts,
Conclude that the rate of default of people having car is low compared to people who don't have car.

NAME_INCOME_TYPE Distibution of Loan Payment Defaulters



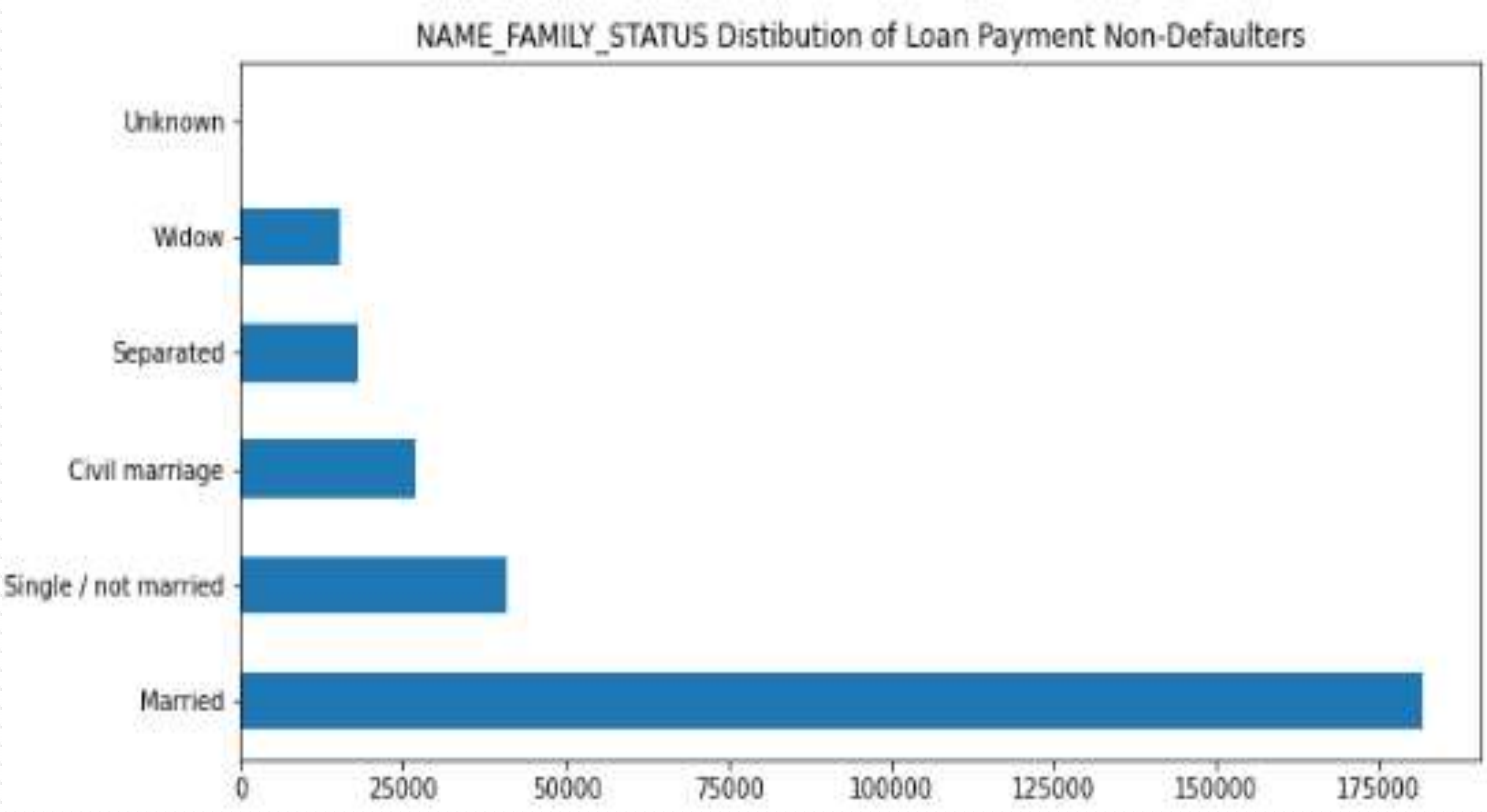
NAME_INCOME_TYPE Distibution of Loan Payment Defaulters



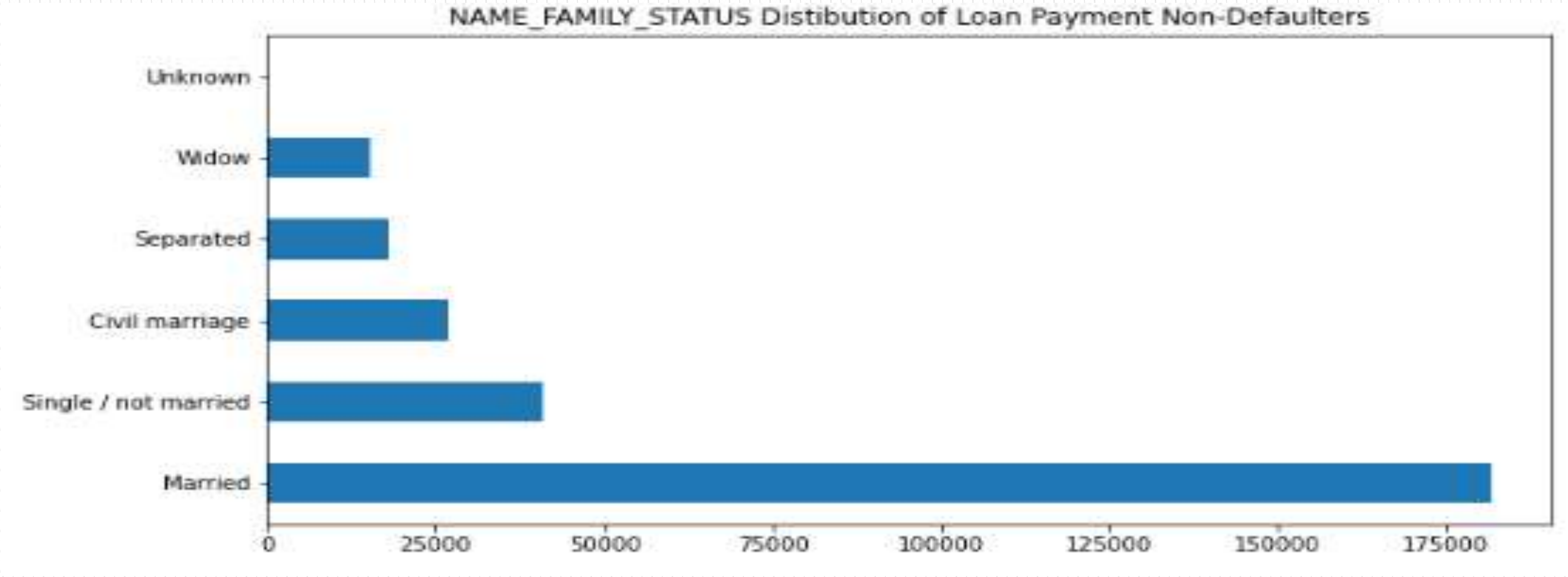
The students don't default. The reason could be they are not required to pay during the time they are students. Observed the Businessman never default. Most of the loans are distributed to working class people.

Also observed that decrease in the percentage of Payment Difficulties who are pensioners and an increase in the percentage of Payment Difficulties who are working.

NAME_FAMILY_STATUS Distibution of Loan Payment Non-Defaulters

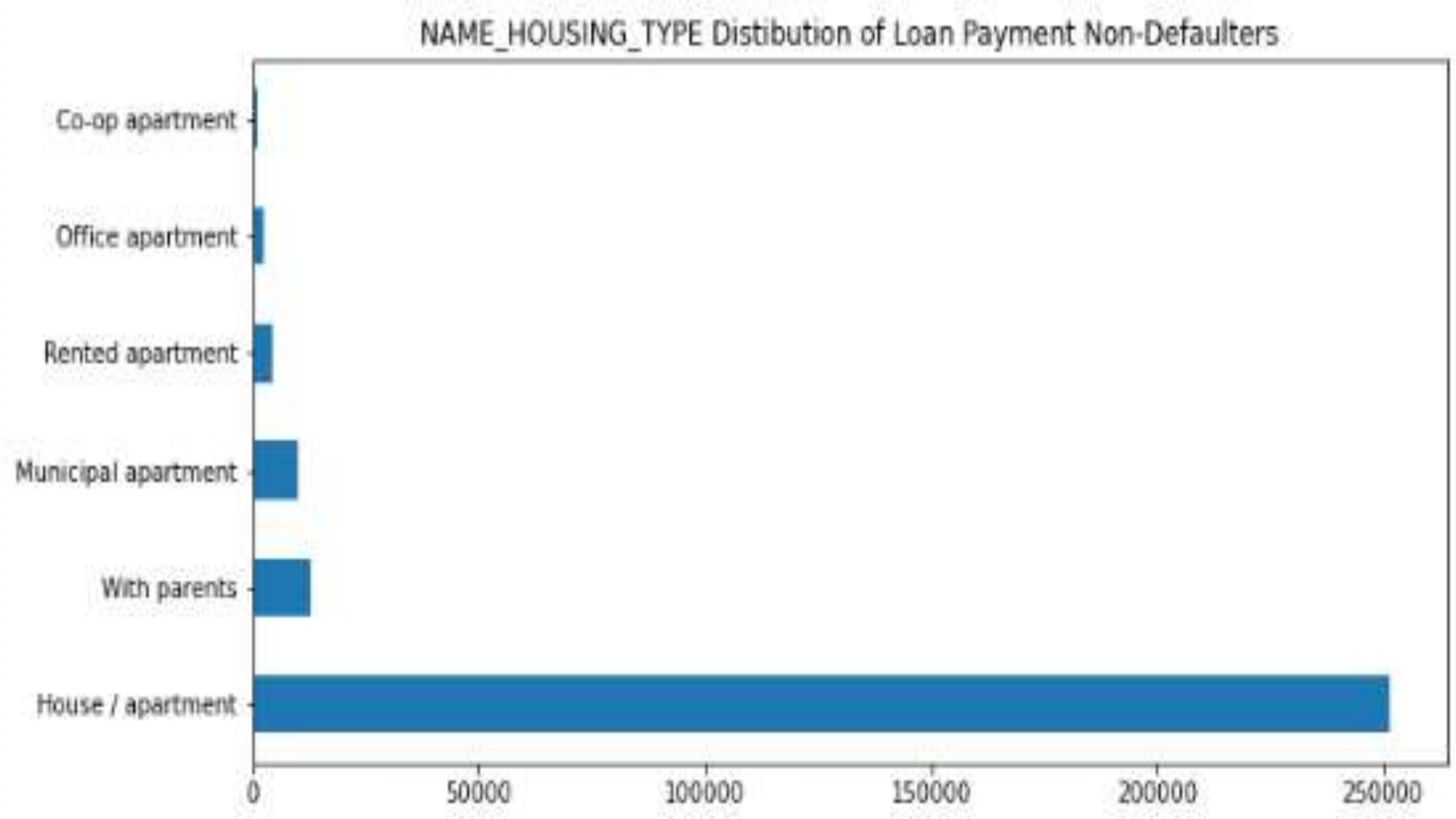


NAME_FAMILY_STATUS Distibution of Loan Payment Defaulters

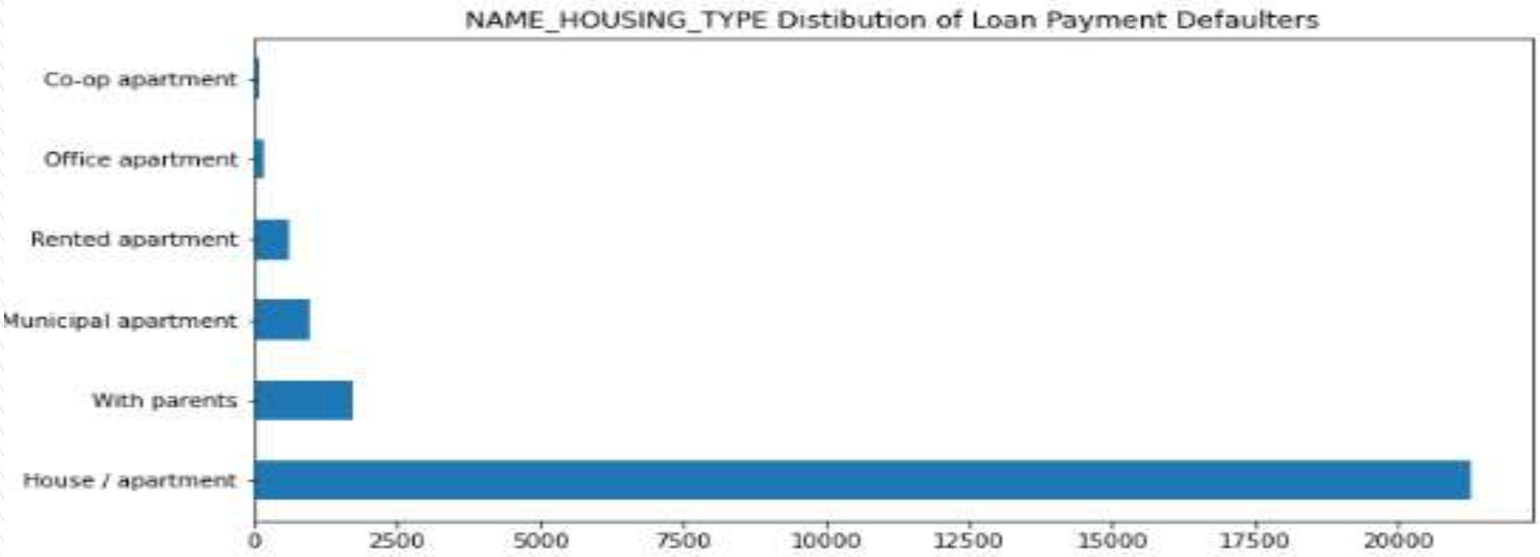


Observing that a decrease in the percentage of married and widowed with Loan Payment Difficulties and an increase in the percentage of single and civil married with Loan Payment Difficulties when compared with the percentages of both Loan Payment Difficulties and Loan Non-Payment Difficulties.

NAME_HOUSING_TYPE Distibution of Loan Payment Non-Defaulters

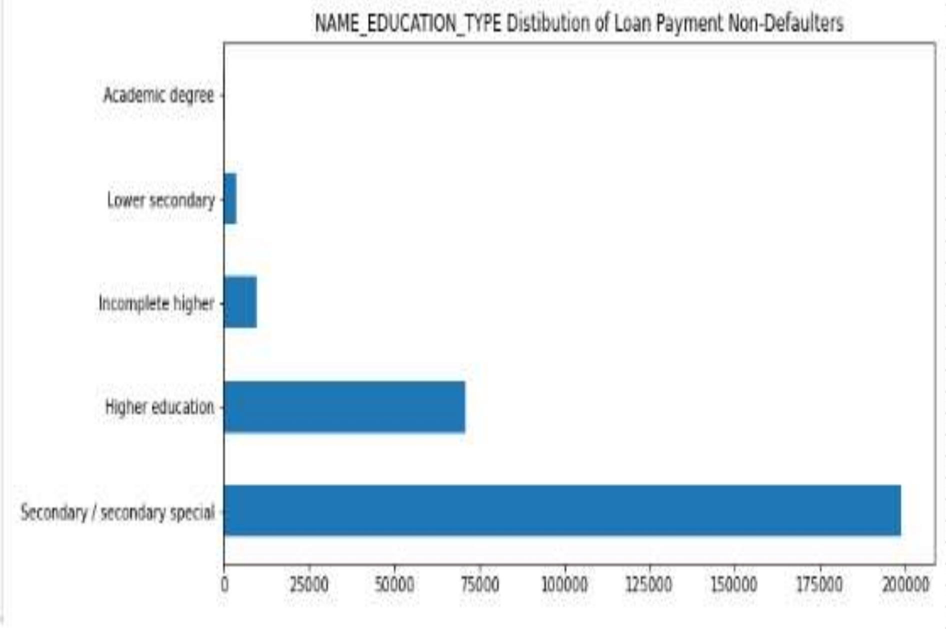


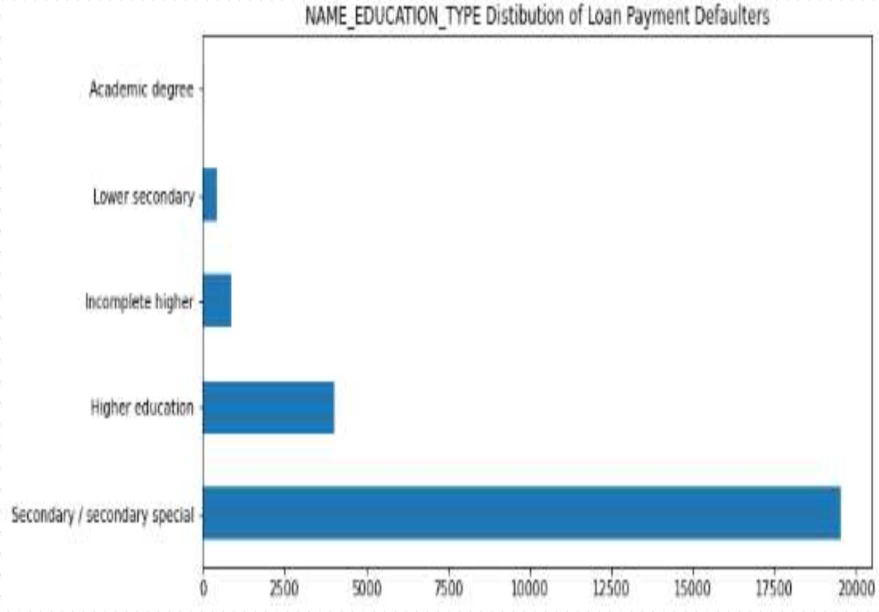
NAME_HOUSING_TYPE Distibution of Loan Payment Defaulters



It observed from the graph that people who have House/Apartment, tend to apply for more loans. People living with parents tend to default more often when compared with other. The reason could be their living expenses are more due to their parents living with them.

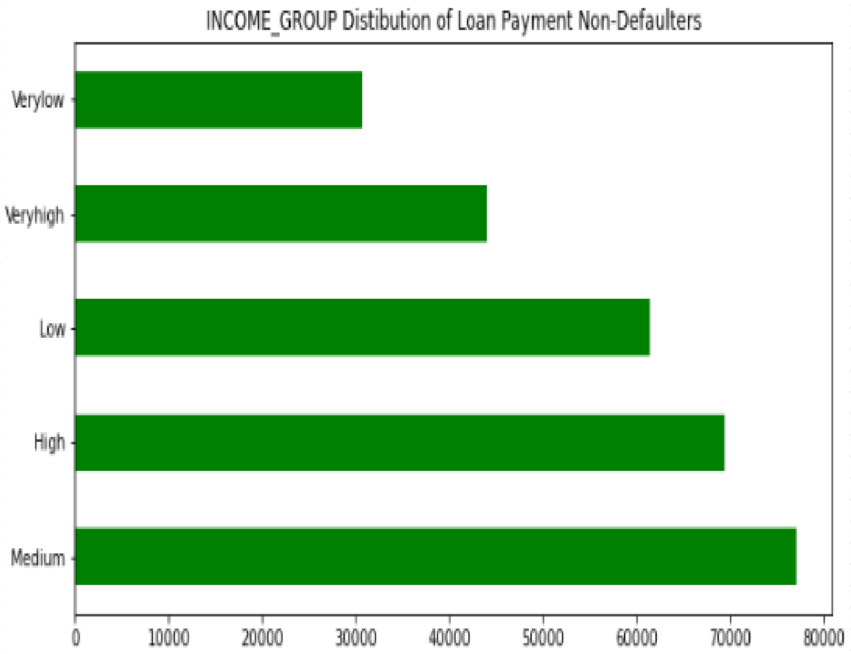
NAME_EDUCATION_TYPE Distibution

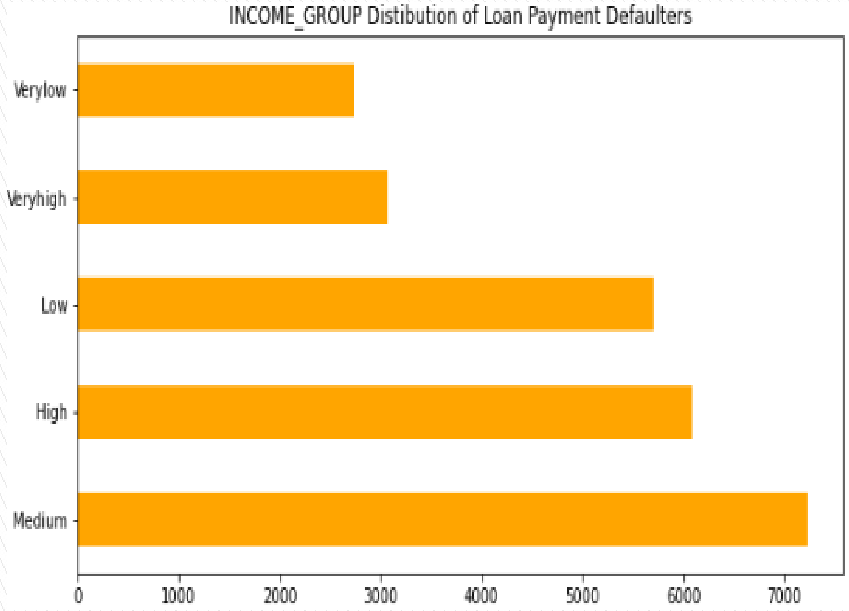




We observe an increase of Loan Payment Difficulties whose educational qualifications are secondary/secondary special and a decrease in the Loan Payment Difficulties who have completed higher education.

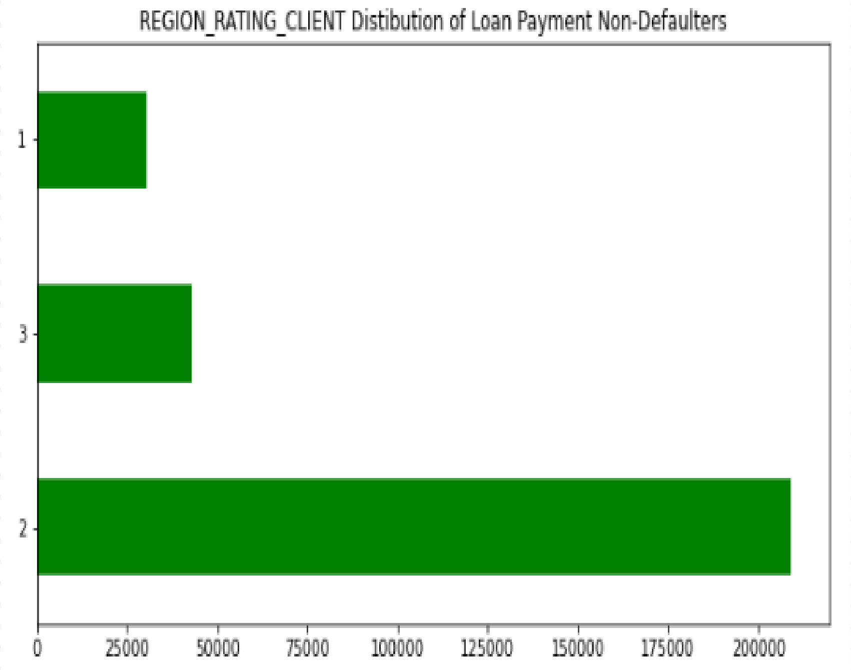
INCOME_GROUP Distribution

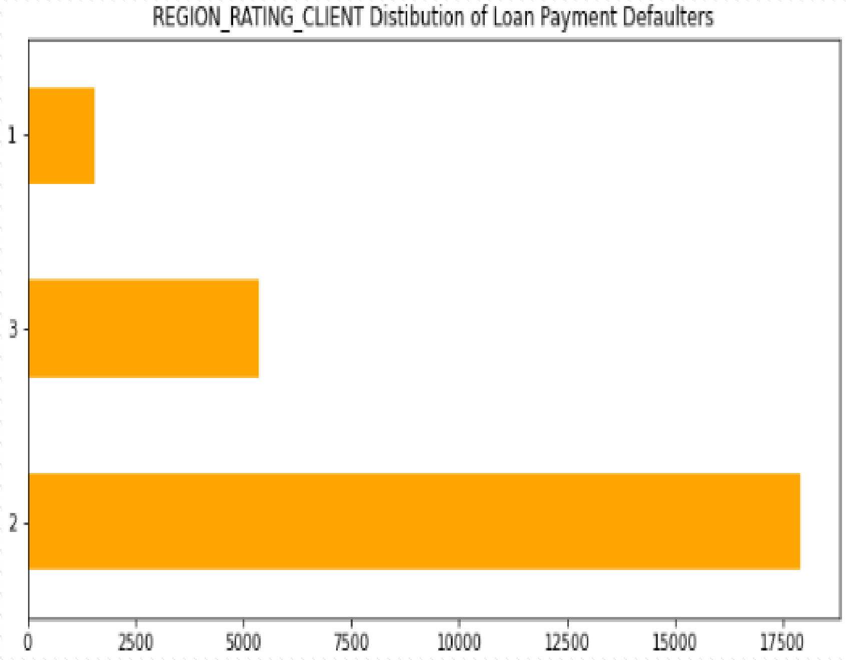




Observed that the Very High income group tend to default less often. They contribute 12.4% to the total number of defaulters, while they contribute 15.6% to the Non-Defaulters.

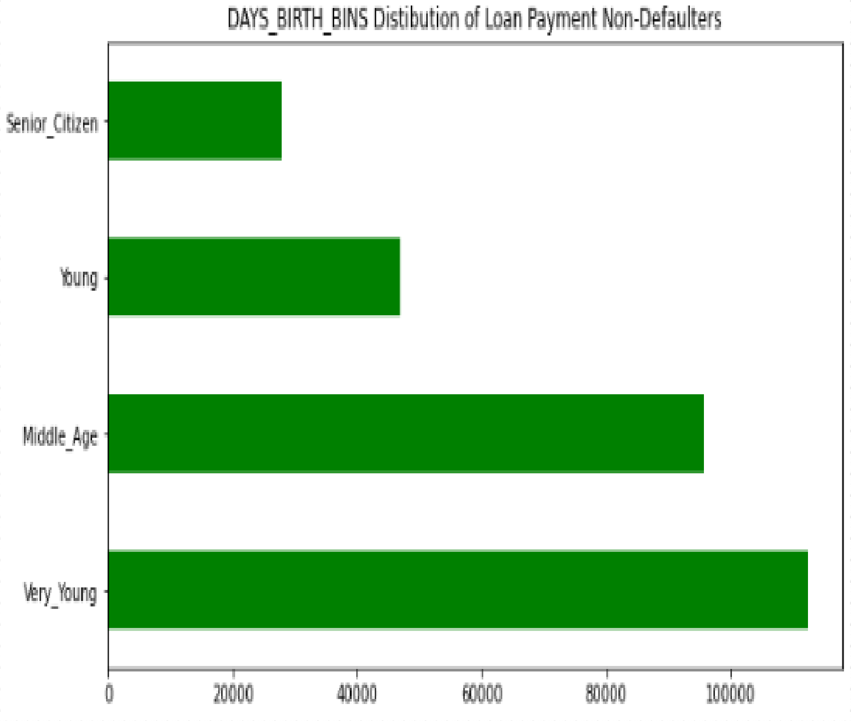
REGION_RATING_CLIENT Distribution

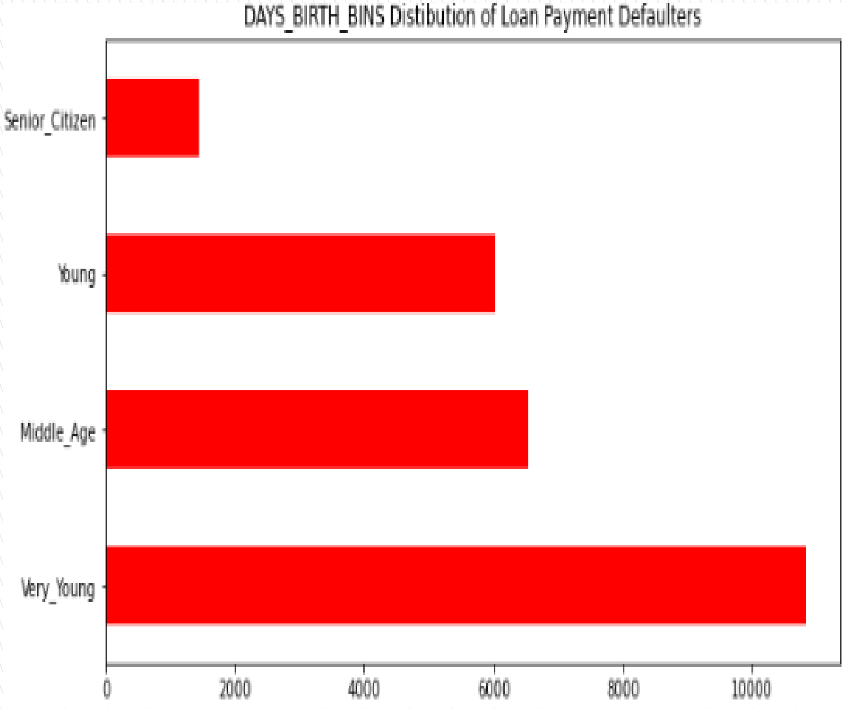




Observed that people from 3rd trie region defaulter increase by 21.6%. More people from 2nd tier regions tend to apply for loans.

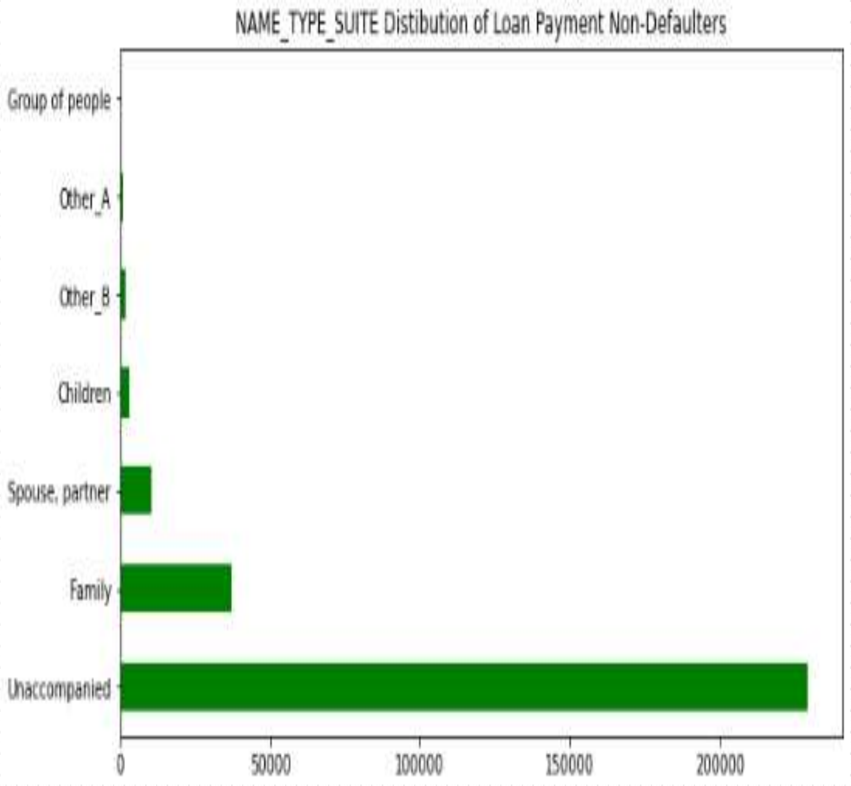
DAYS_BIRTH_BINS Distribution

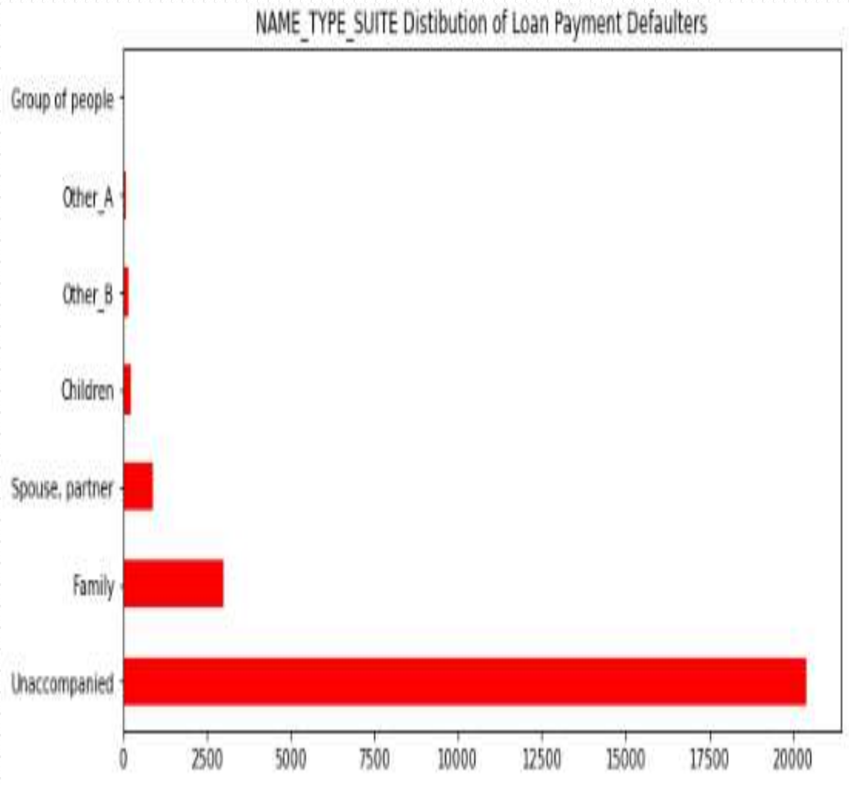




Observed that increase the defaulters in age group very young by 43% and young by 24%.

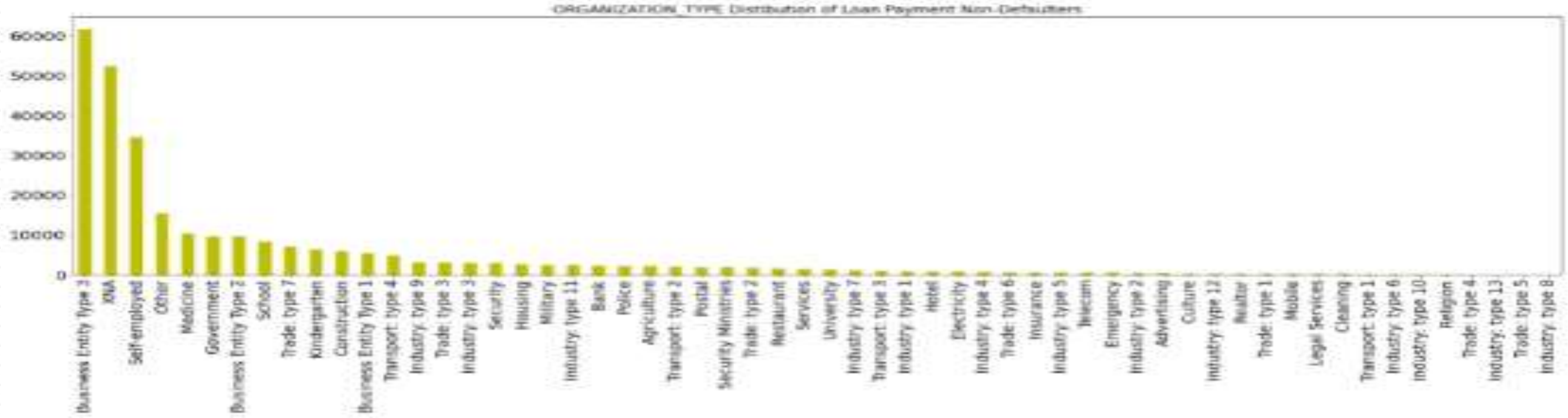
NAME_TYPE_SUITE Distribution

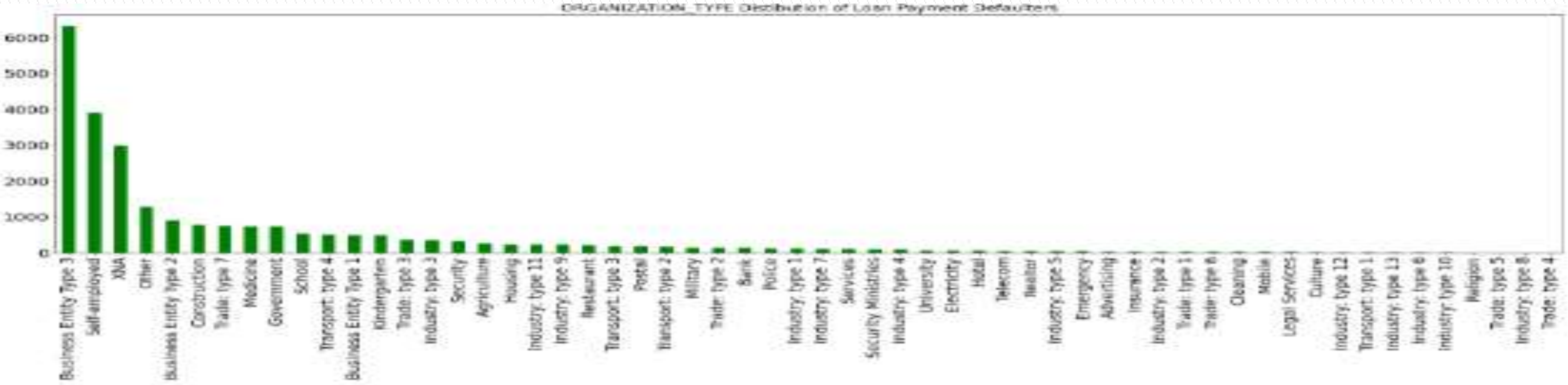




No major changes observed

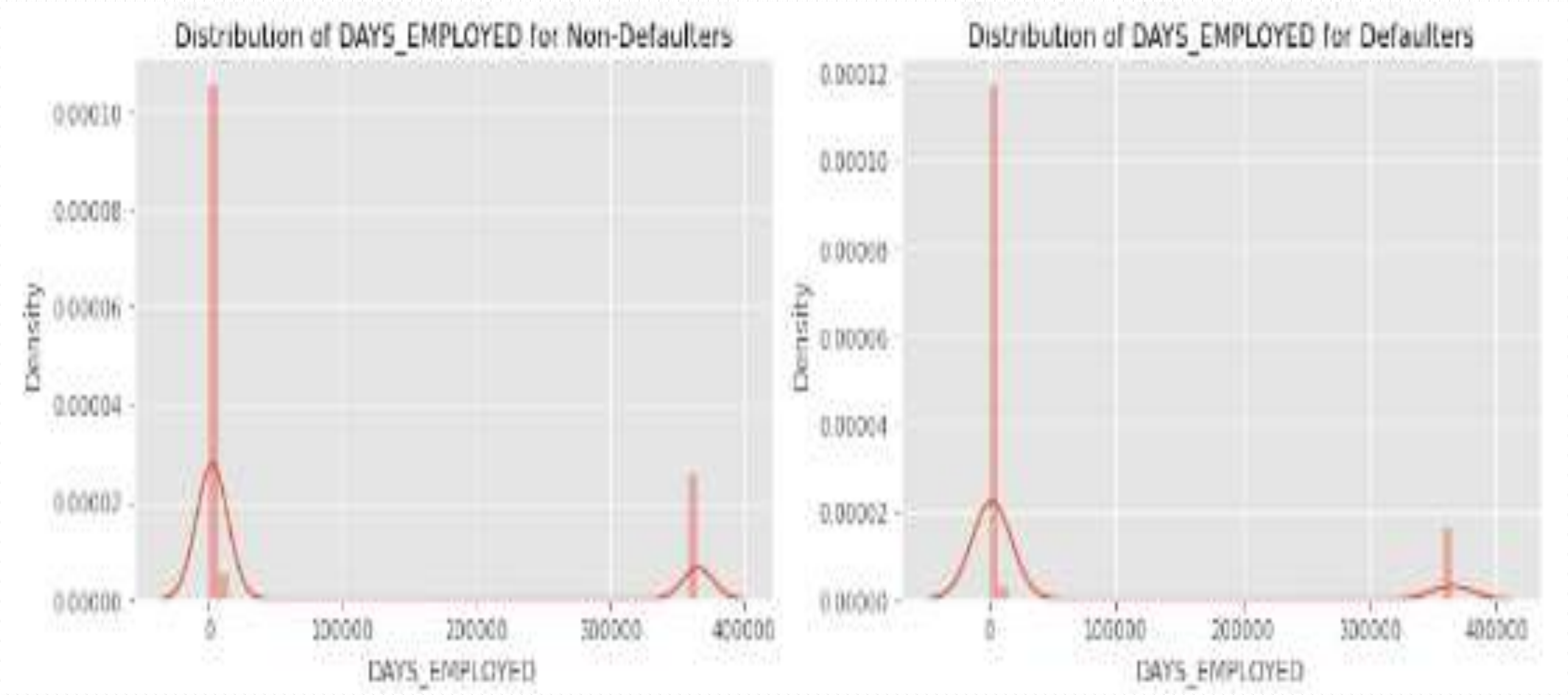
ORGANIZATION_ TYPE Distribution


ORGANIZATION_TYPE Distribution of Loan Payment Non-Defaulters


ORGANIZATION_TYPE Distribution of Loan Payment Defaulters

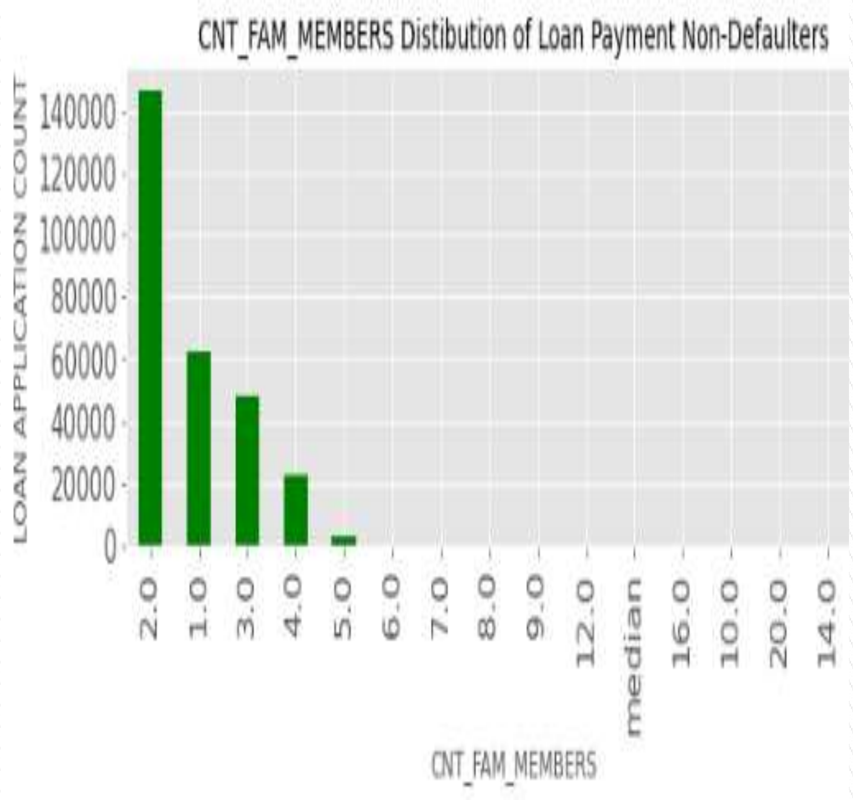
No major changes observed.

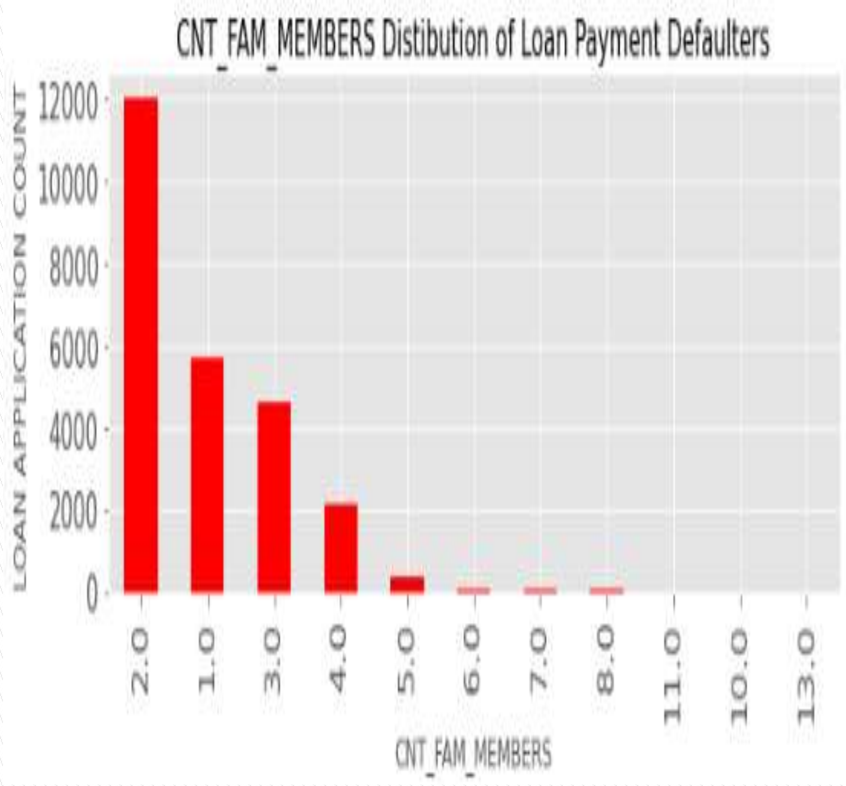
Univariate continuous variable analysis



No major changes

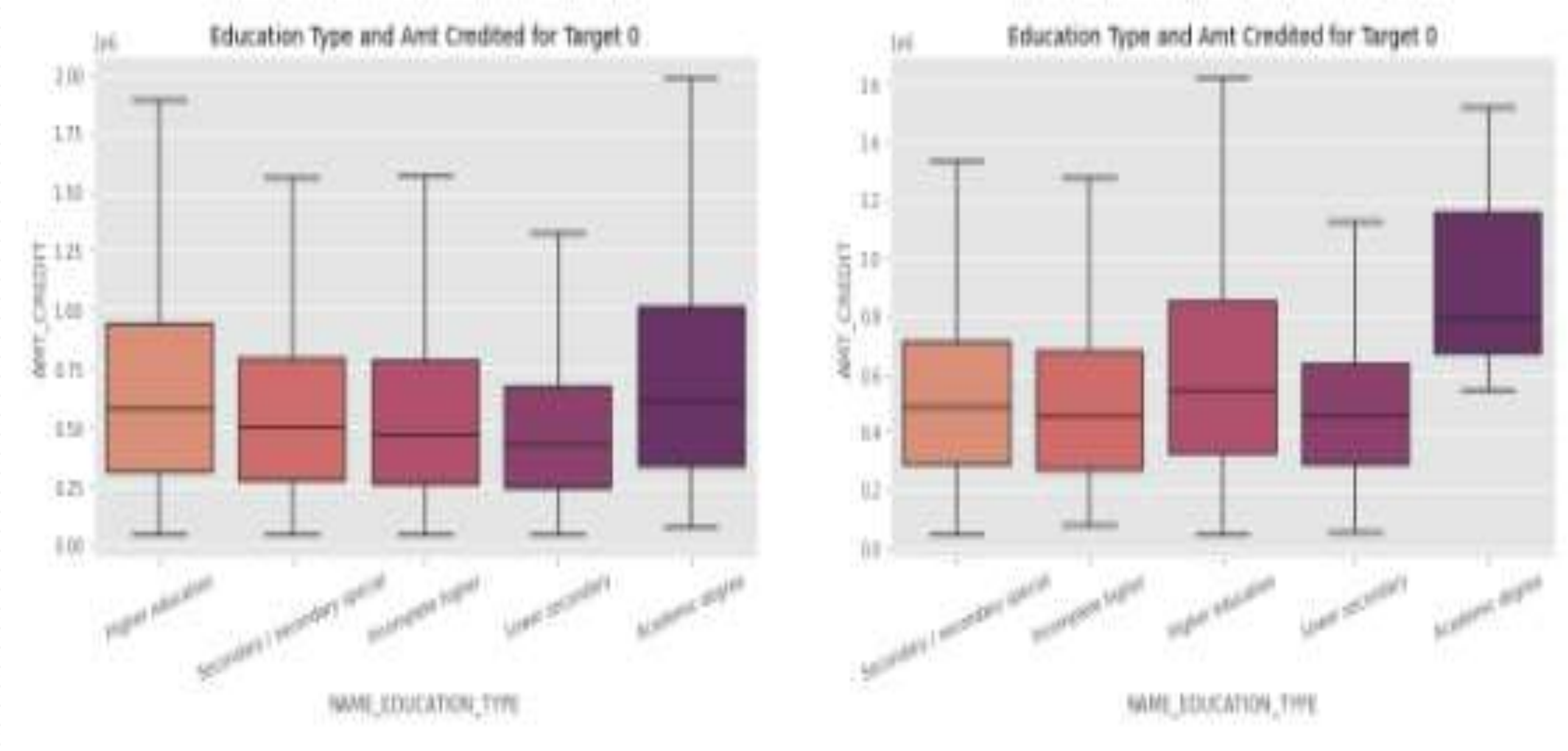
CNT_FAM_MEMBERS Distribution





Observed that a family of 3 applies loan more often than the other families.

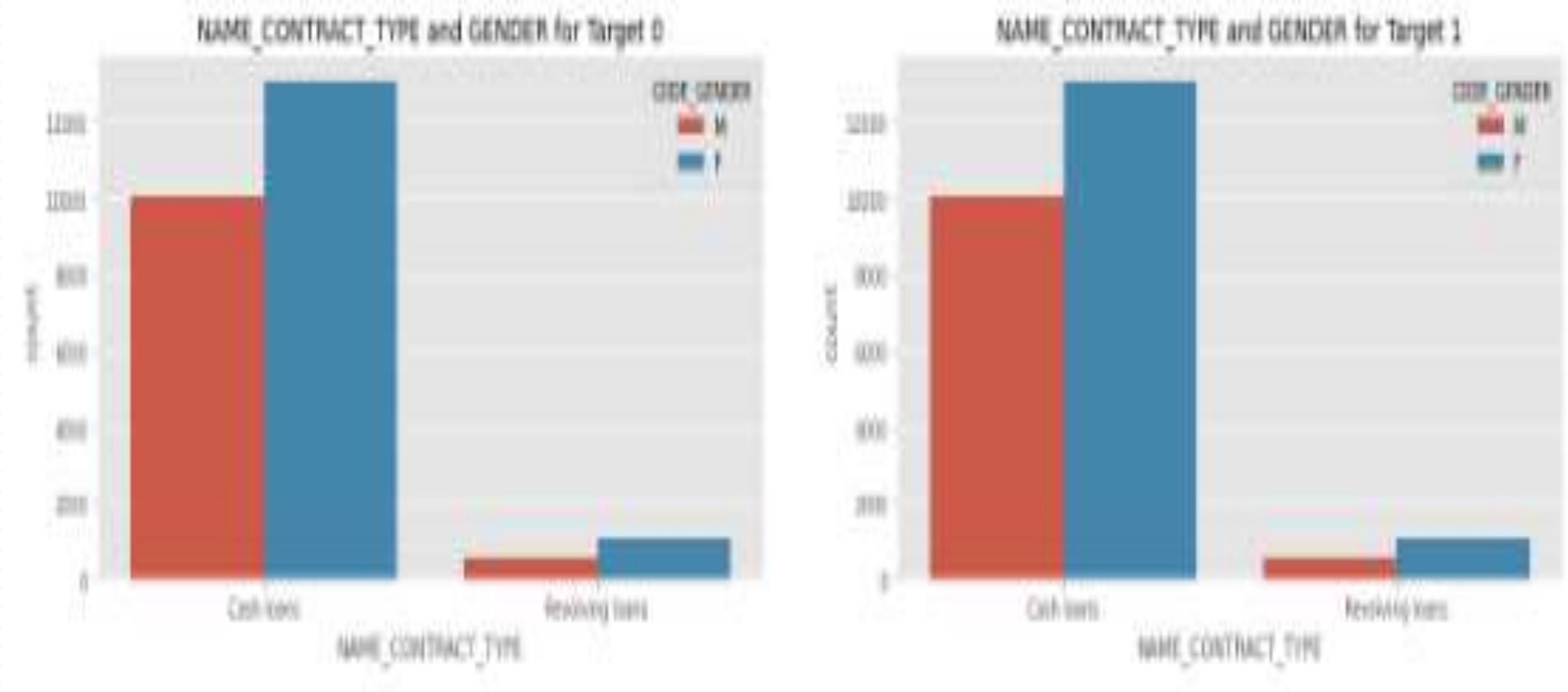
Bivariate Analysis of Categorical to Continuous



Observed that median of Loan values defaulting for Applicants with Academic degree is higher. But in a plot above, no of applicants with academic degree is miniscule and no inference can be drawn from this analysis.

Bivariate analysis categorical and categorical

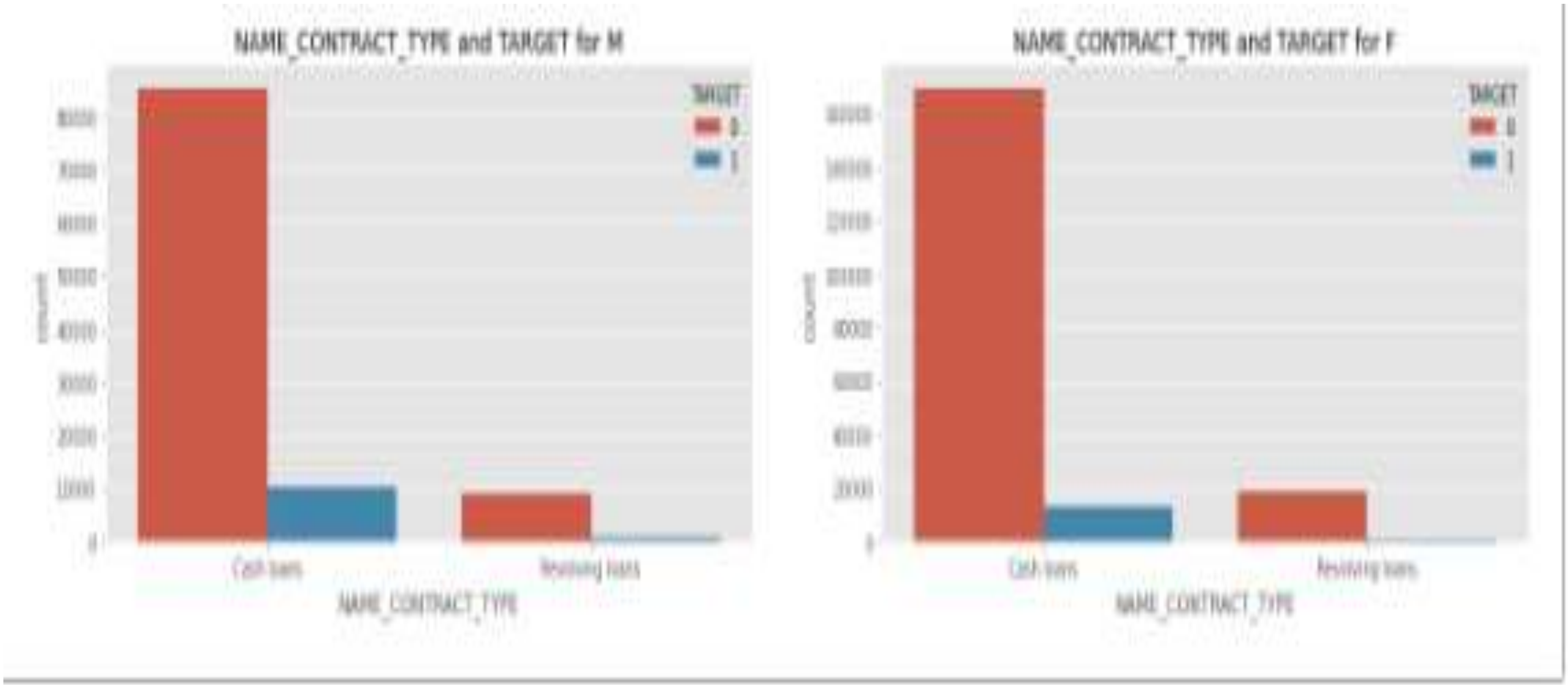
NAME_CONTRACT_TYPE and GENDER for Target 0 and Target 1



Observed females as more loan applicant. As seen in plot above, though male applicants are lower, ratio of male applicants defaulting is higher. Let us check this by another analysis

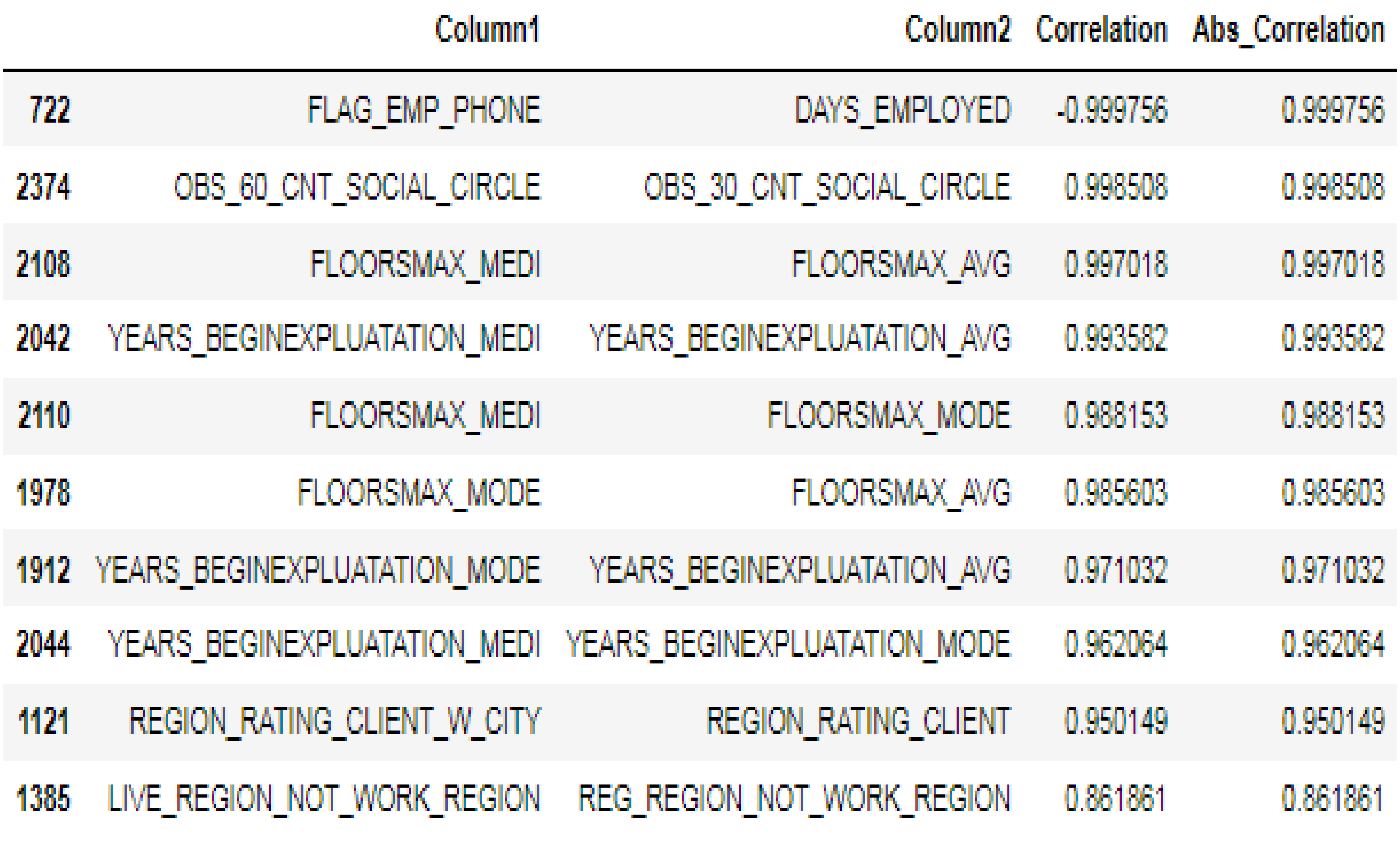
Other way of analysis:

NAME_CONTRACT_TYPE and TARGET for M and F



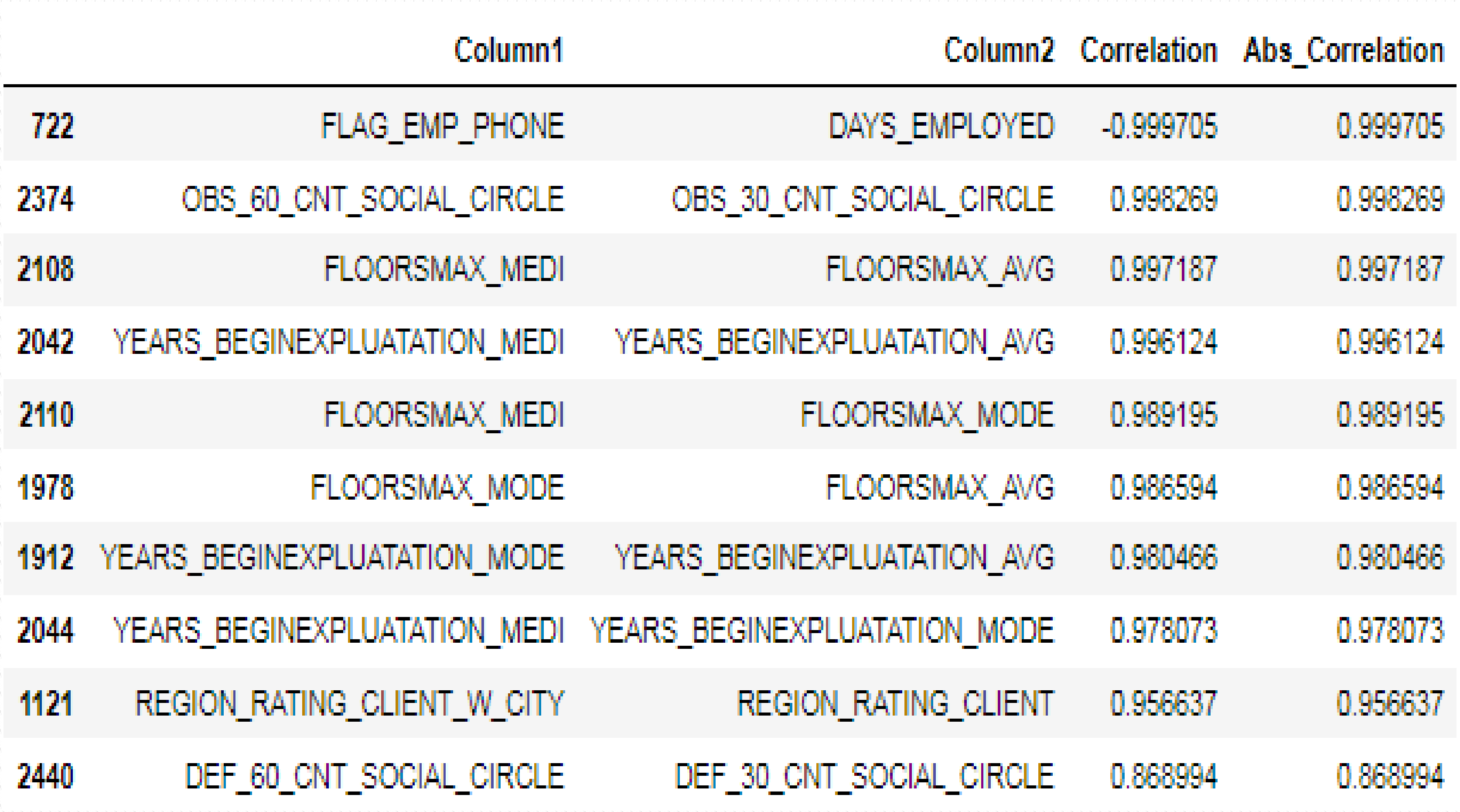
Observerd that male applicants are defaulting more that female applicants.

The top 10 correlation of the selected columns in target 0



	Column1	Column2	Correlation	Abs_Correlation
722	FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999756	0.999756
2374	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508	0.998508
2108	FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997018	0.997018
2042	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.993582	0.993582
2110	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.988153	0.988153
1978	FLOORSMAX_MODE	FLOORSMAX_AVG	0.985603	0.985603
1912	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.971032	0.971032
2044	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.962064	0.962064
1121	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149	0.950149
1385	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861	0.861861

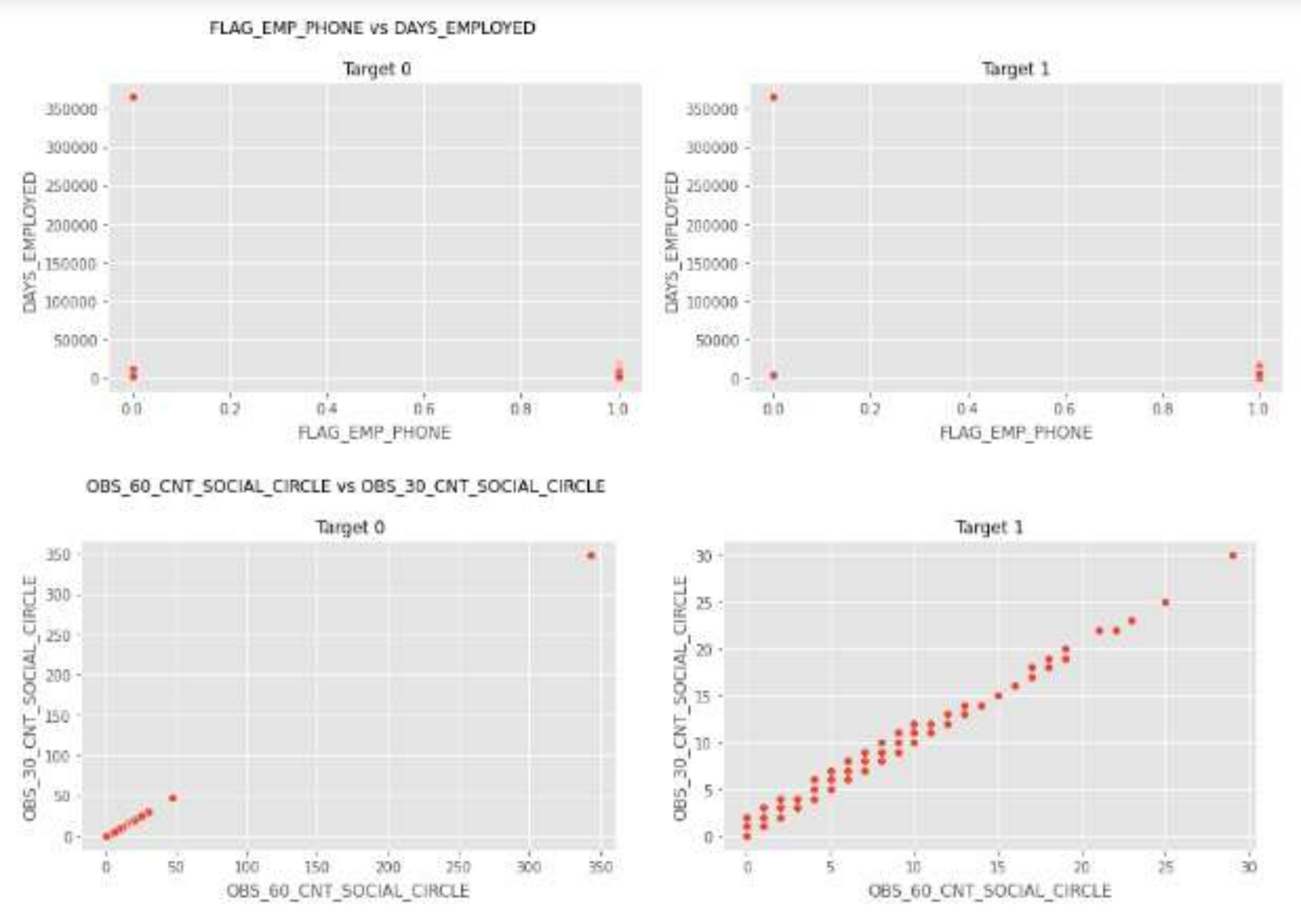
The top 10 correlation of the selected columns in target 1



	Column1	Column2	Correlation	Abs_Correlation
722	FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999705	0.999705
2374	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269	0.998269
2108	FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997187	0.997187
2042	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.996124	0.996124
2110	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.989195	0.989195
1978	FLOORSMAX_MODE	FLOORSMAX_AVG	0.986594	0.986594
1912	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.980466	0.980466
2044	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.978073	0.978073
1121	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637	0.956637
2440	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994	0.868994

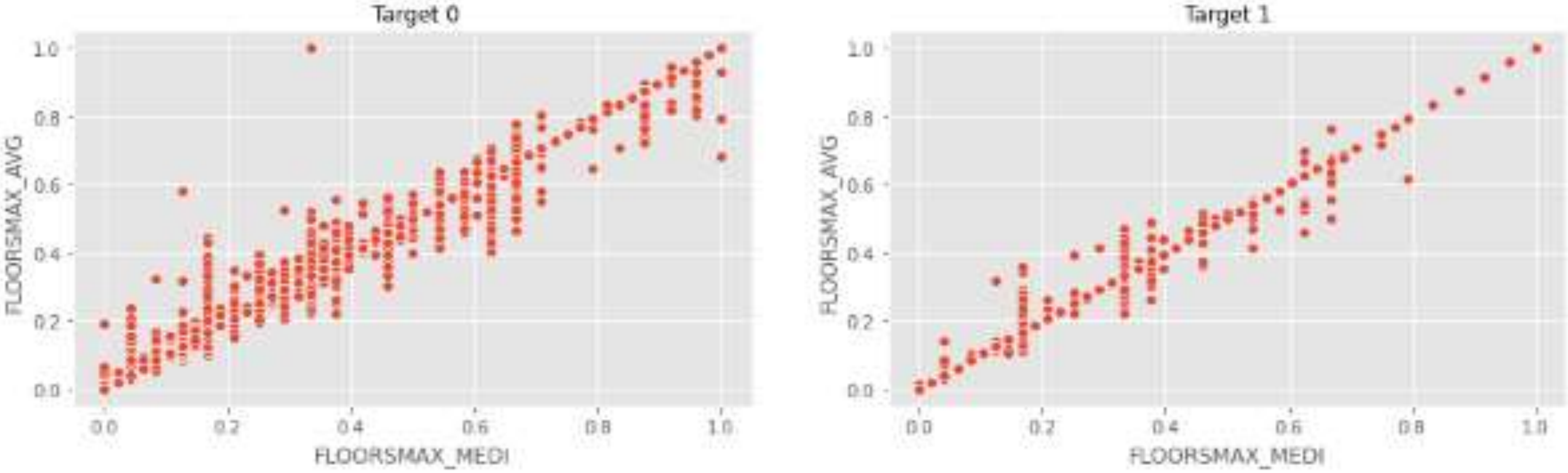
Observed that the Top 10 correlation columns are same for Target 0 and Target 1

Analysis on Continuous Variables with scatter plot

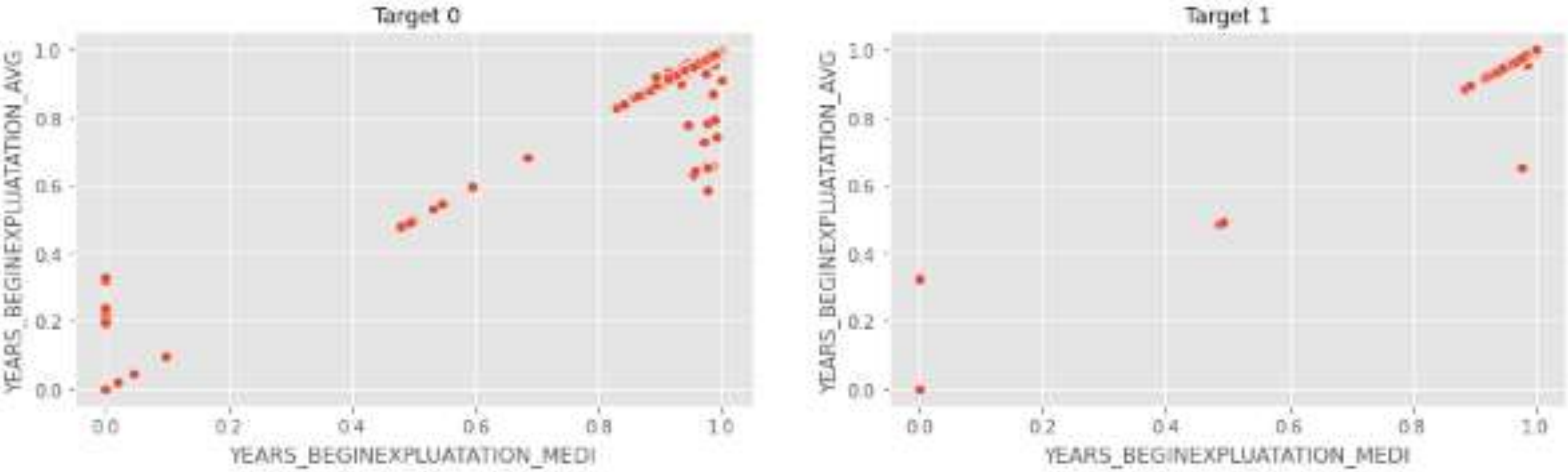


Analysis on Continuous Variables with scatter plot

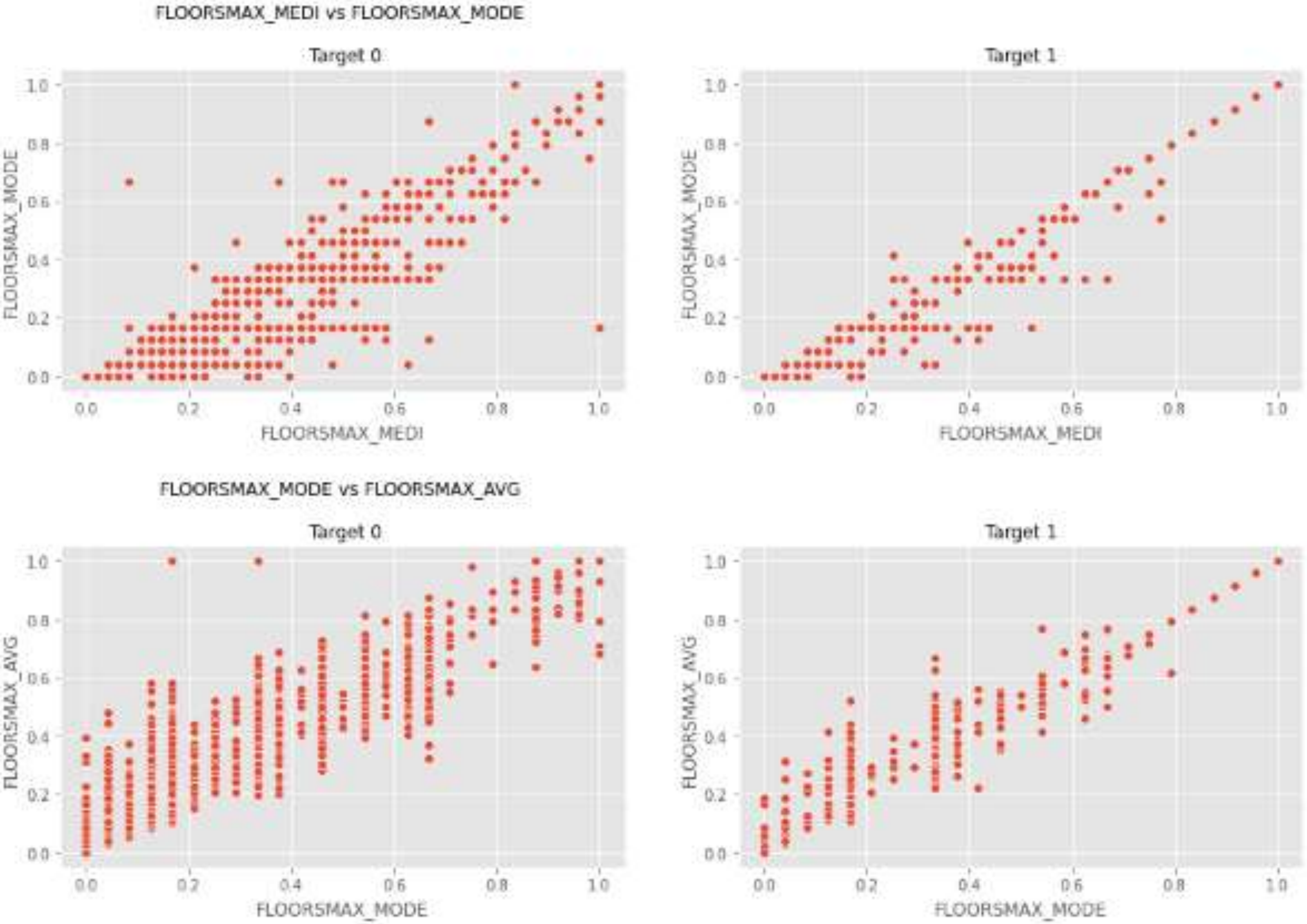
FLOORSMAX_MEDI vs FLOORSMAX_AVG



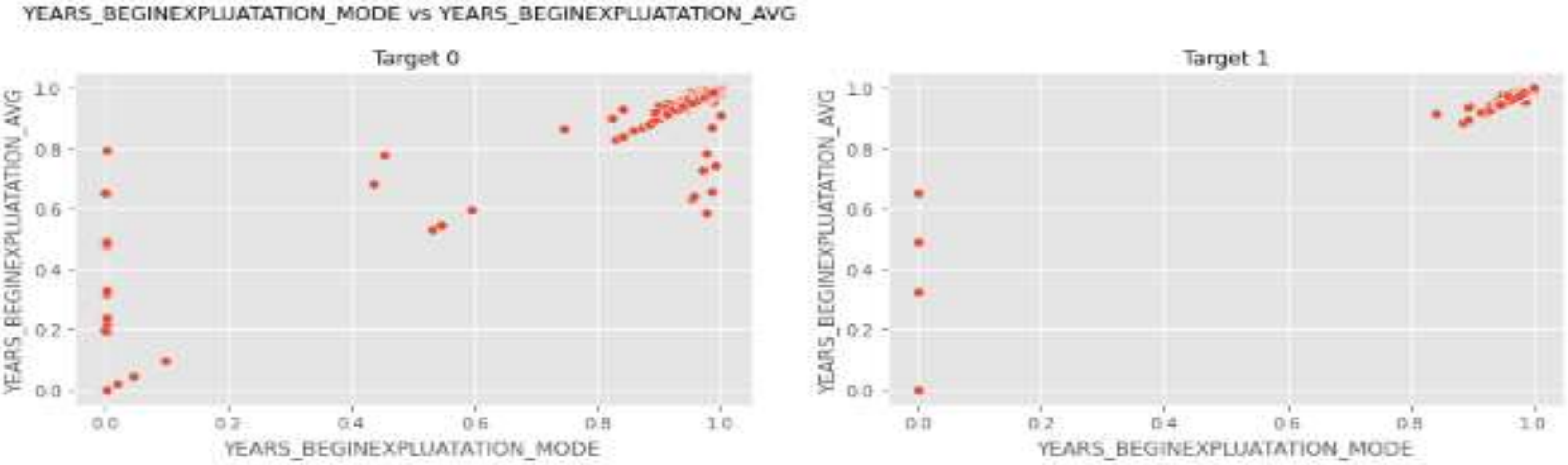
YEARS_BEGINEXPLUATATION_MEDI vs YEARS_BEGINEXPLUATATION_AVG

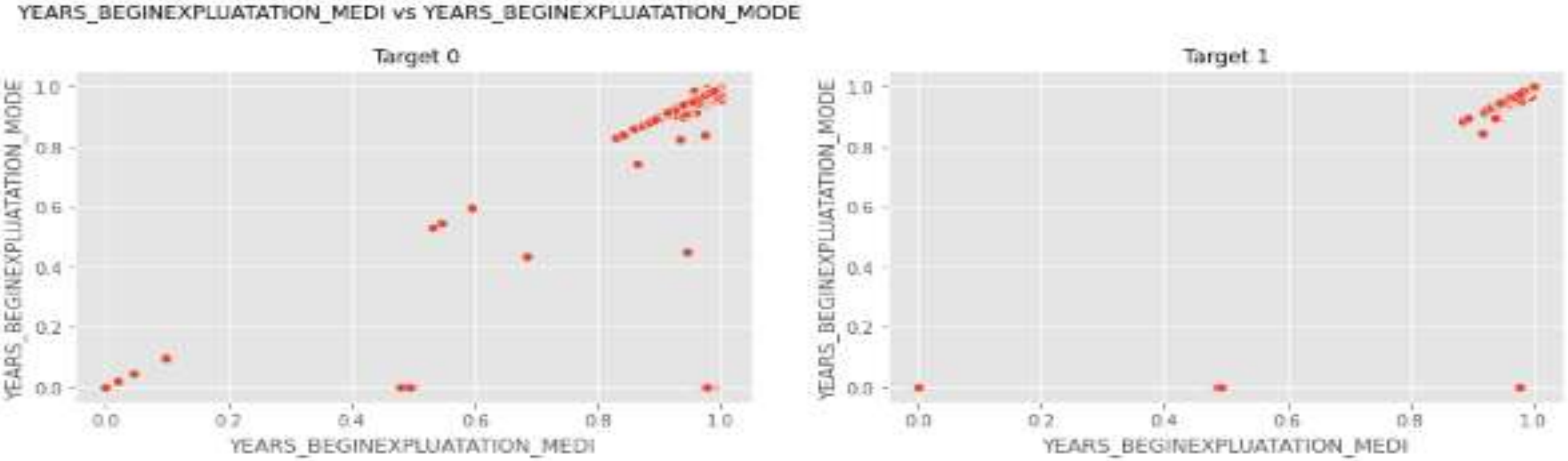


Analysis on Continuous Variables with scatter plot



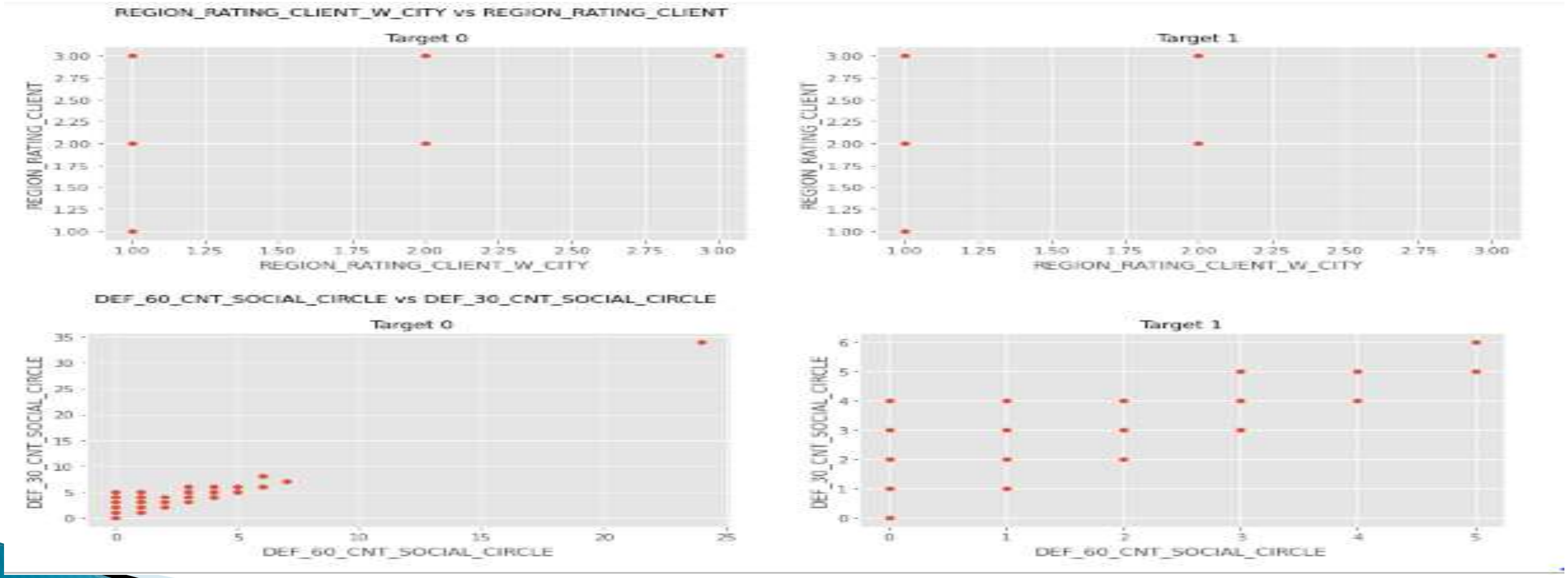
Analysis on Continuous Variables with scatter plot



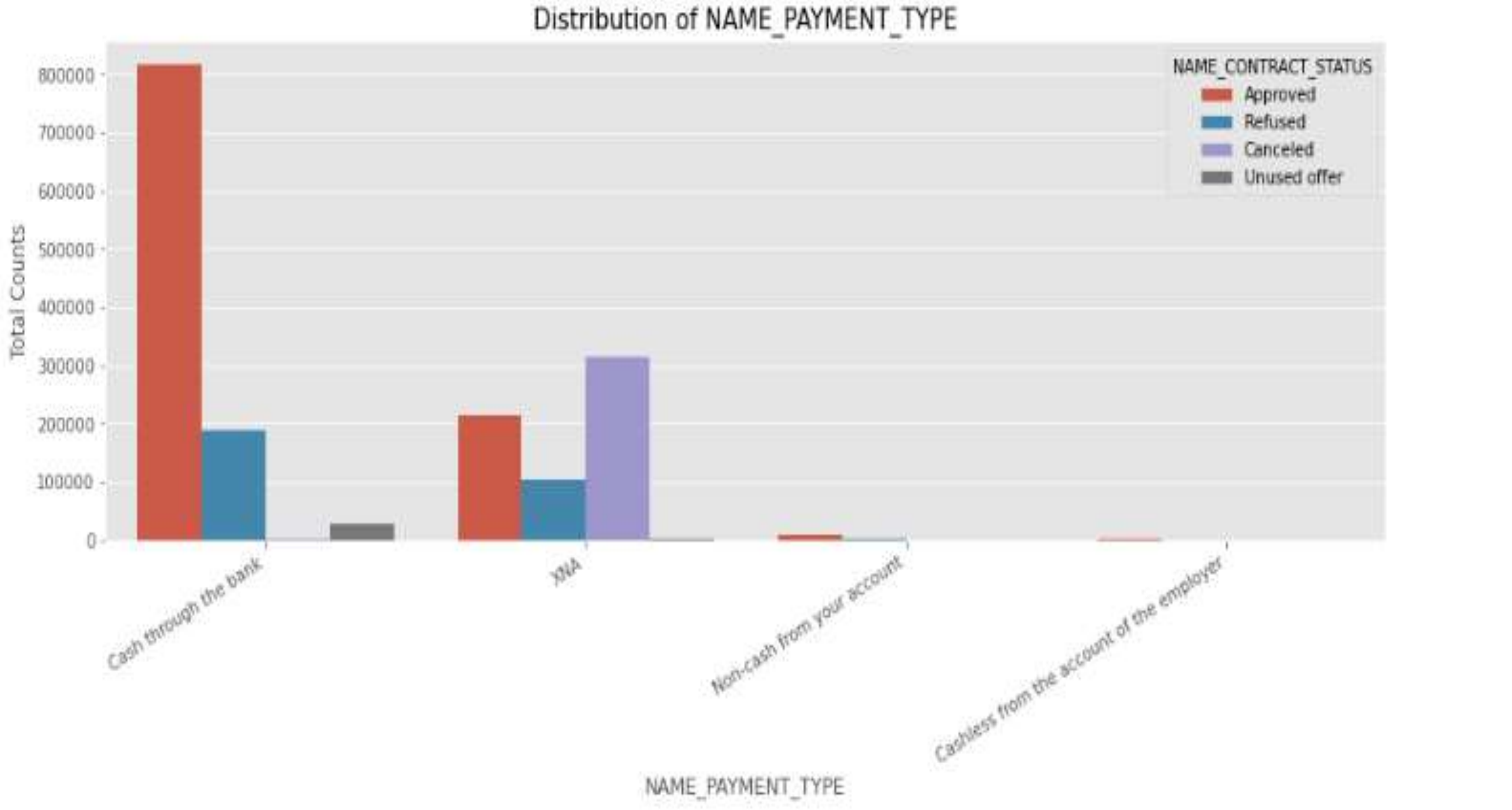


Analysis on Continuous Variables with scatter plot

1. OBS_30_CNT_SOCIAL_CIRCLE',OBS_60_CNT_SOCIAL_CIRCLE' – denote the client's social surroundings with observable 30/60 DPD. These are definetly correlated. We can also see that its higher and steeper for Target 1, signyfying that in approval process this parameter must be strongly looked into. 2.DEF_30_CNT_SOCIAL_CIRCLE – Trend is going up. But Target 1 has lesser data and hence graph is not dense. 3. Years employed has an outlier value of 999 and this is skewing the graph 4.AMT_CREDIt and AMT_GOOD PRICE dont seem to be increasing proportionately with AMT_INCOME for TARGET 1, thus possibly leading to default.

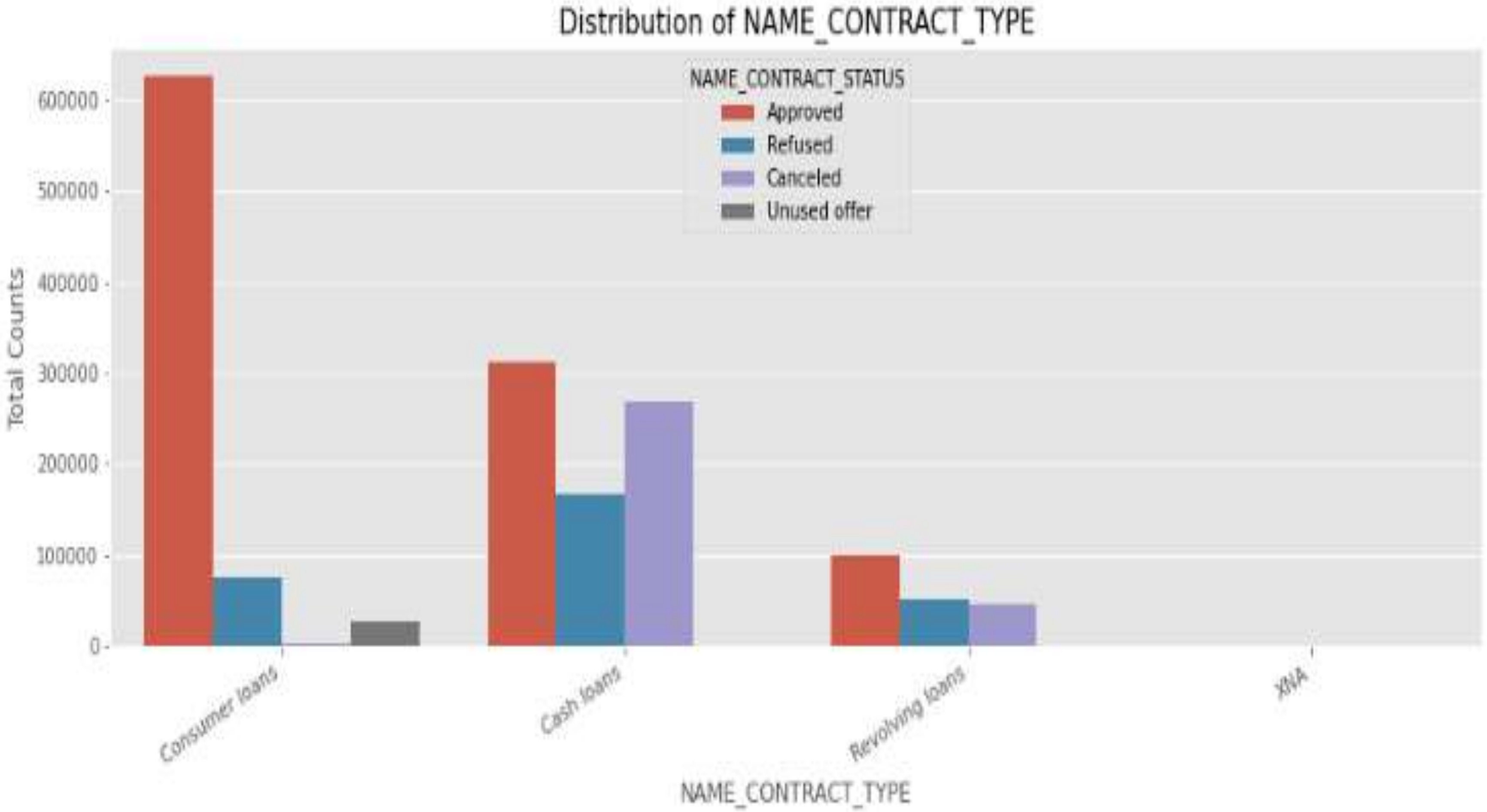


Previous Application Dataset analysis
Univariate analysis for NAME_PAYMENT_TYPE



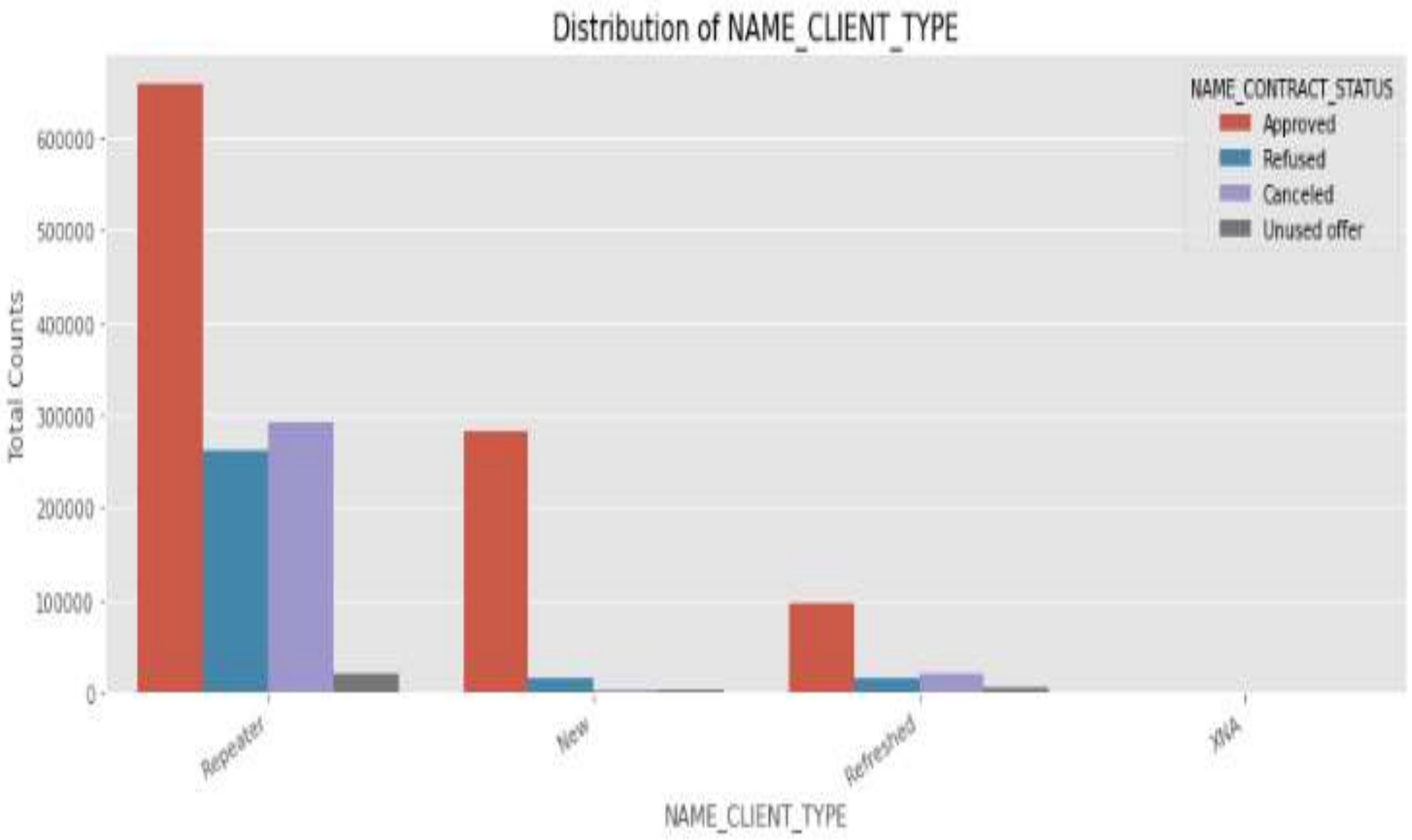
Observed that most of the clients chose to repay the loan using the 'Cash through the bank' option, also observed 'Non-Cash from your account' & 'Cashless from the account of the employee' options are not at all popular in terms of loan repayment amongst the customers.

Univariate analysis for NAME_CONTRACT_TYPE



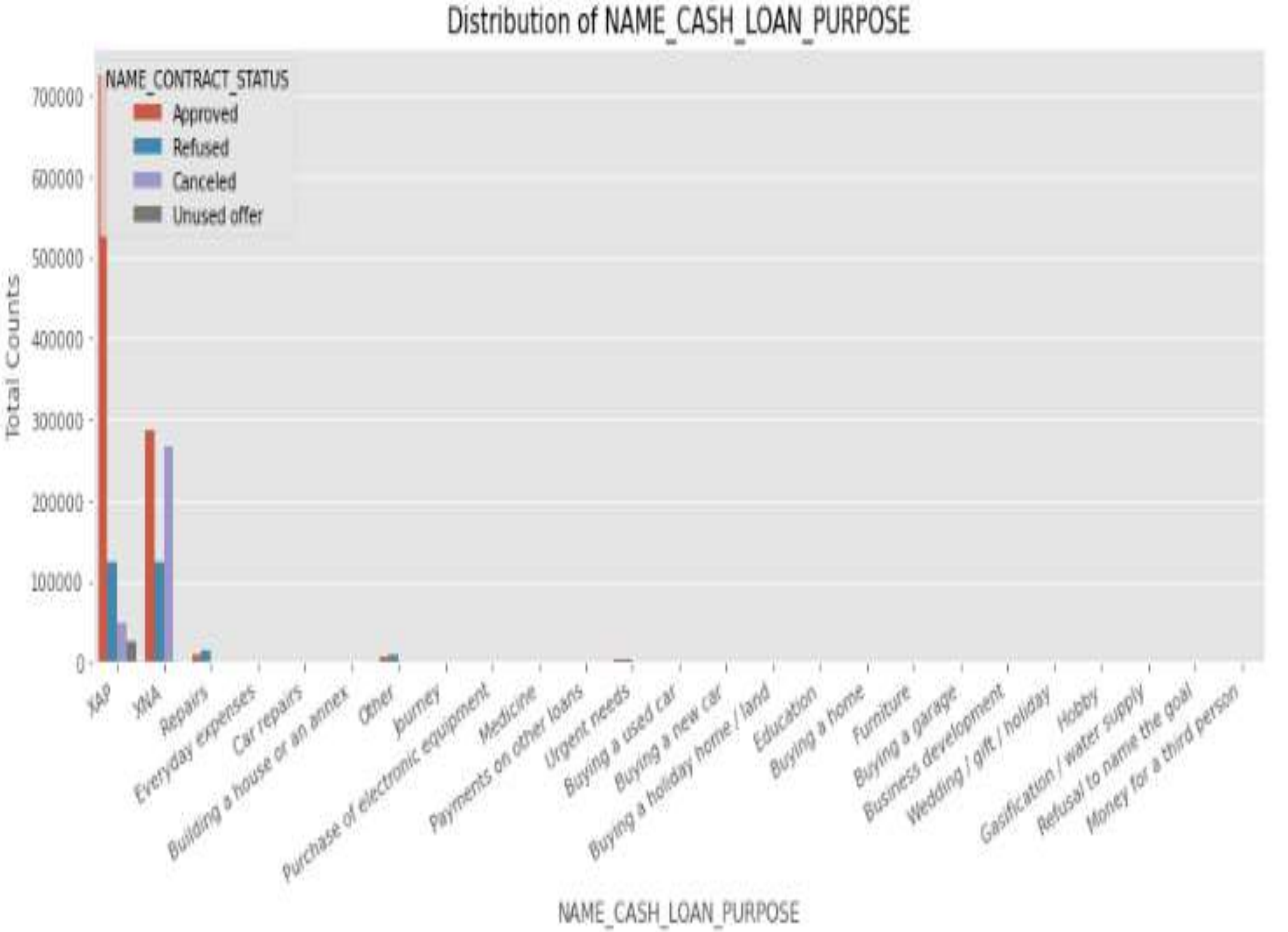
Observed that most of the applications are for 'Cash loan' and 'Consumer loan'. Although the cash loans are refused more often than others.

Univariate analysis for NAME_CLIENT_TYPE

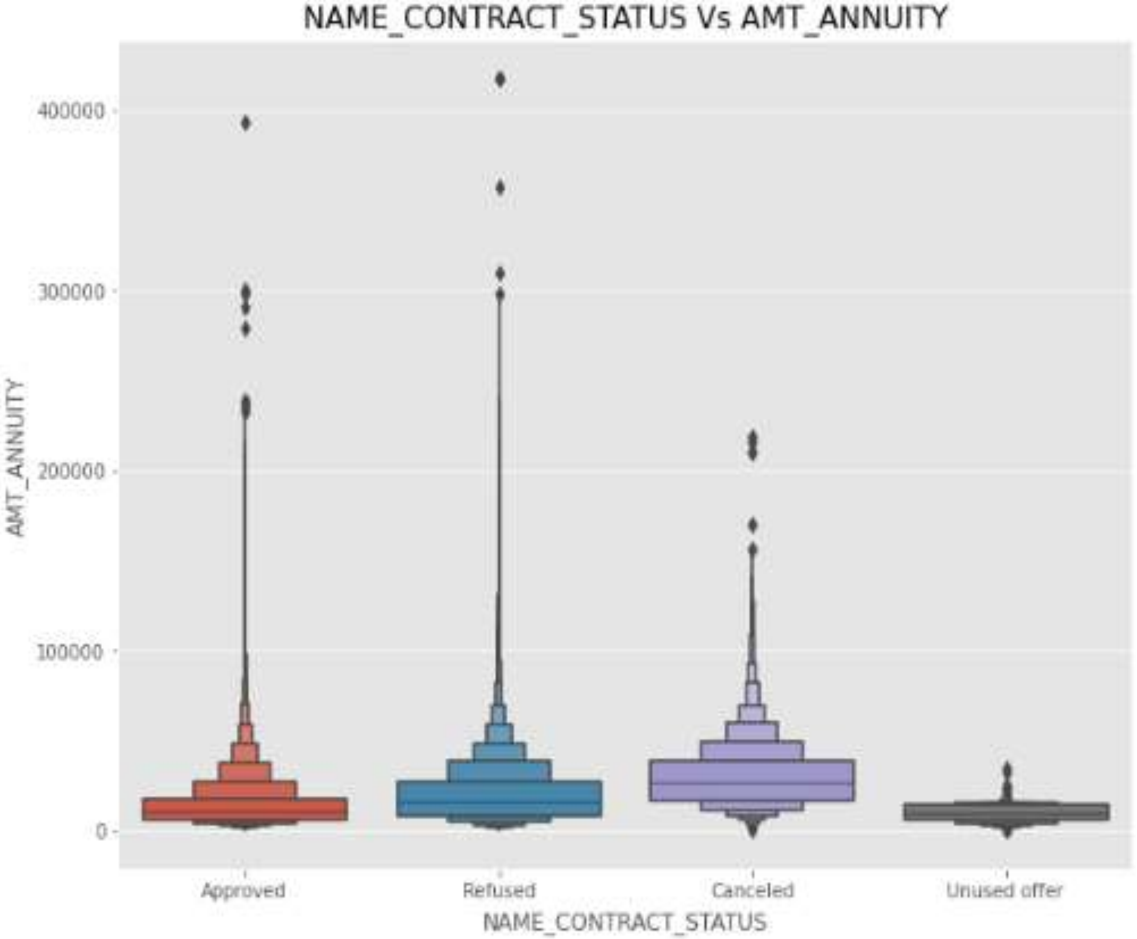


Most of the loan applications are from repeat customers, out of the total applications customers are repeaters. They also get refused most often also.

Univariate analysis for NAME_CASH_LOAN_PURPOSE

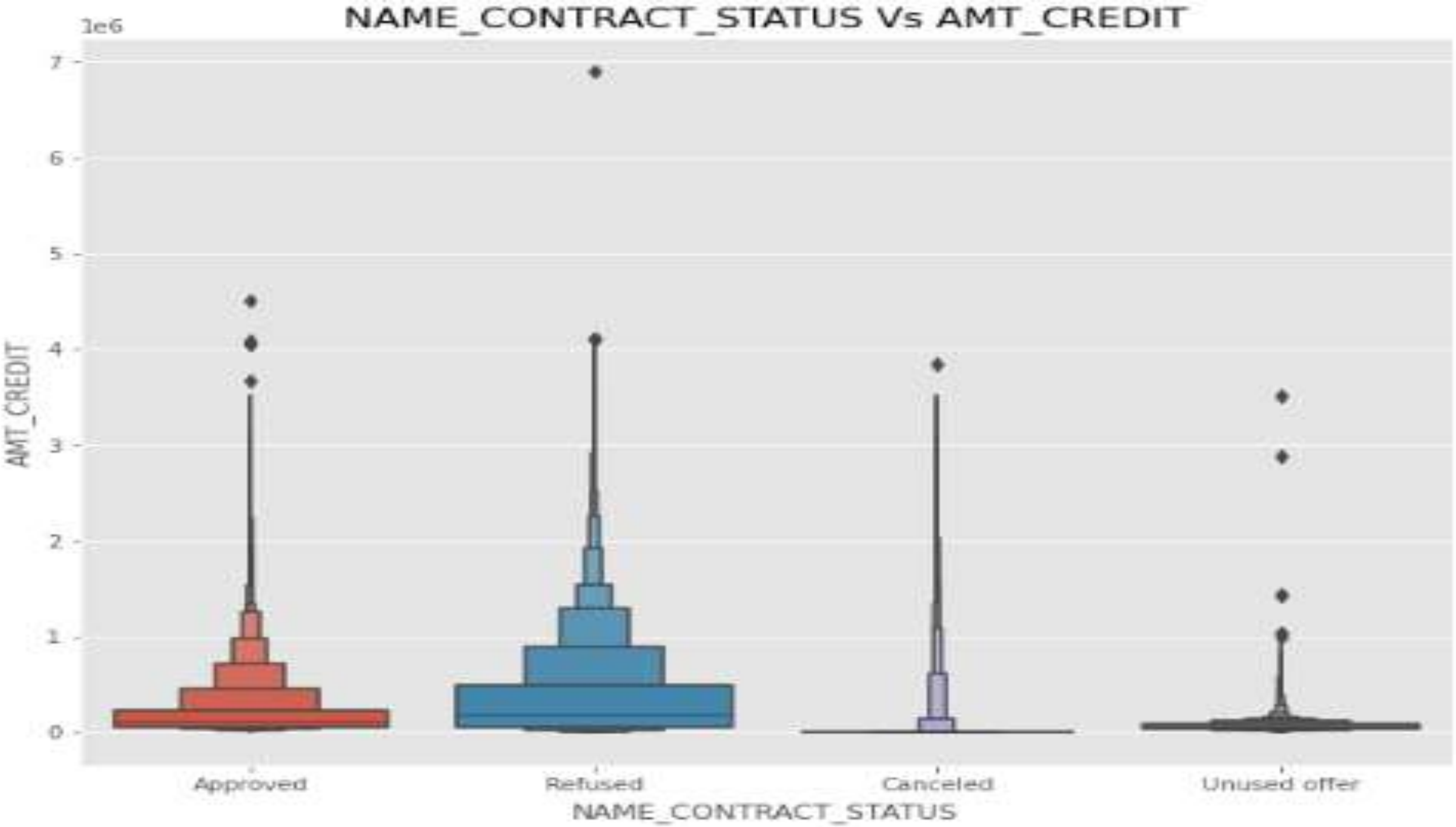


Using box plot to do some more bivariate analysis on categorical vs numeric columns.



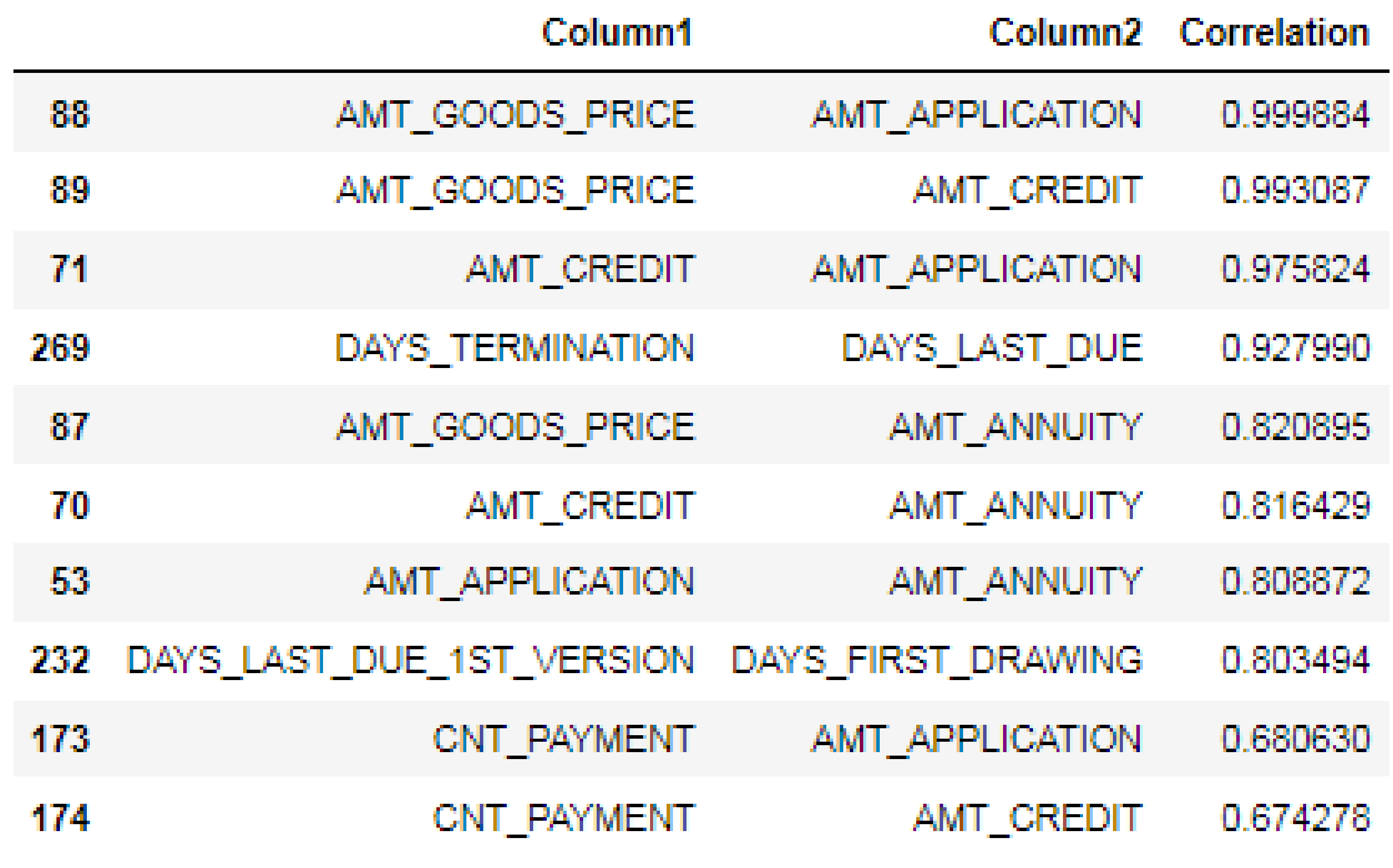
Observed that loan application for people with lower AMT_ANNUITY gets canceled or Unused most of the time,also applications with too high AMT ANNUITY also got refused more often than others.

Analysis NAME_CONTRACT_STATUS vs AMT_CREDIT



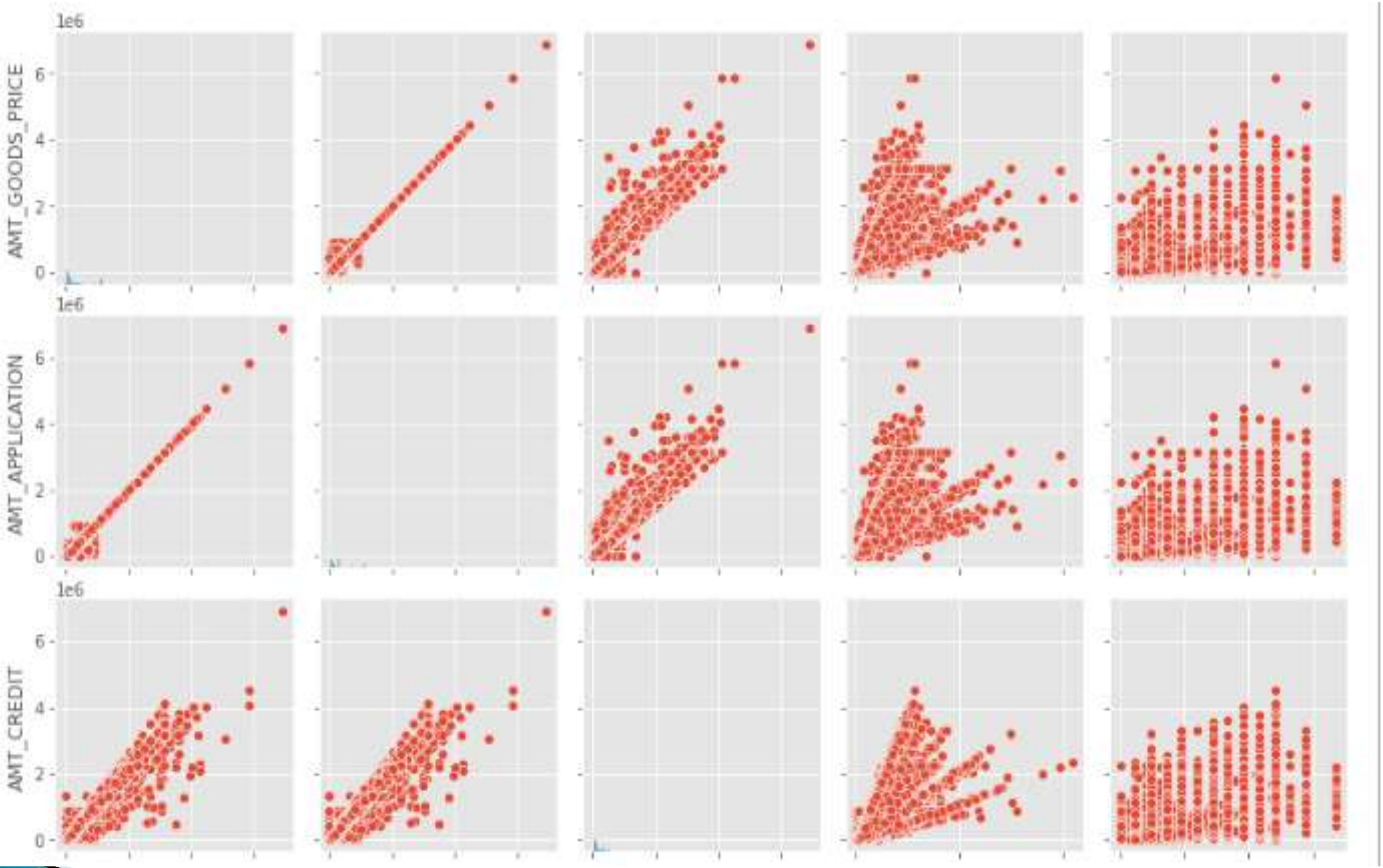
Observed that when the AMT_CREDIT is too low, it get's cancelled/unused most of the time.

Top Correlation of previous application

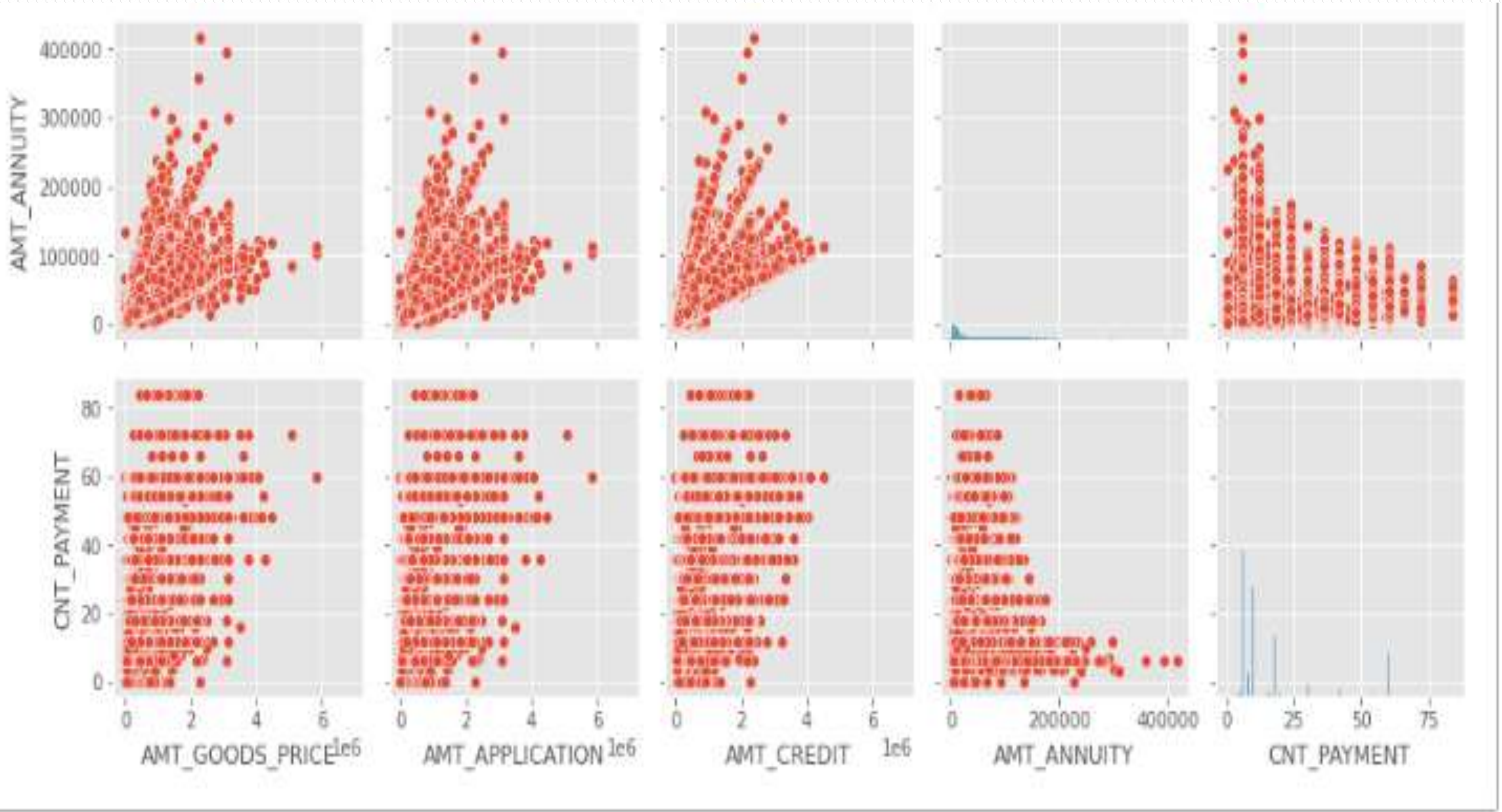


	Column1	Column2	Correlation
88	AMT_GOODS_PRICE	AMT_APPLICATION	0.999884
89	AMT_GOODS_PRICE	AMT_CREDIT	0.993087
71	AMT_CREDIT	AMT_APPLICATION	0.975824
269	DAYS_TERMINATION	DAYS_LAST_DUE	0.927990
87	AMT_GOODS_PRICE	AMT_ANNUITY	0.820895
70	AMT_CREDIT	AMT_ANNUITY	0.816429
53	AMT_APPLICATION	AMT_ANNUITY	0.808872
232	DAYS_LAST_DUE_1ST_VERSION	DAYS_FIRST_DRAWING	0.803494
173	CNT_PAYMENT	AMT_APPLICATION	0.680630
174	CNT_PAYMENT	AMT_CREDIT	0.674278

Plot to show correlation of selected columns

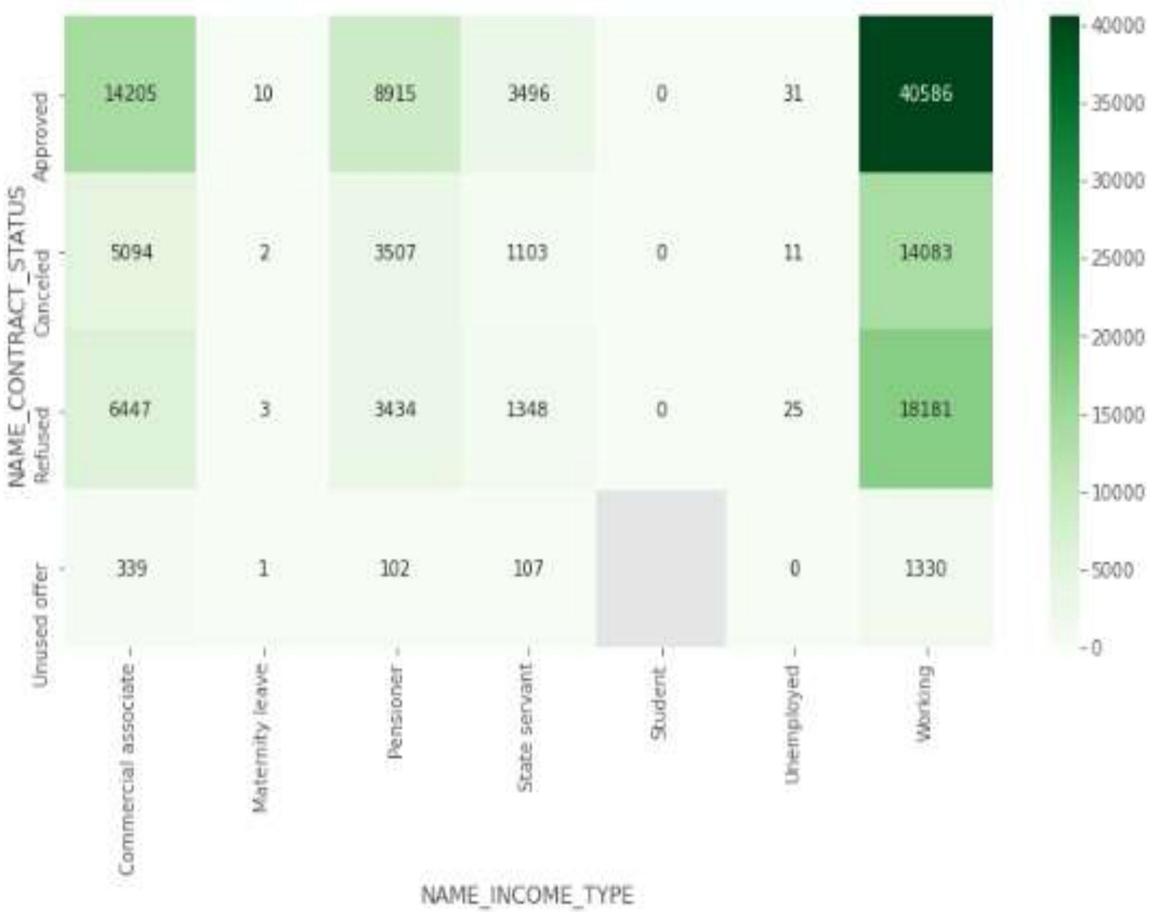


Plot to show correlation of selected columns



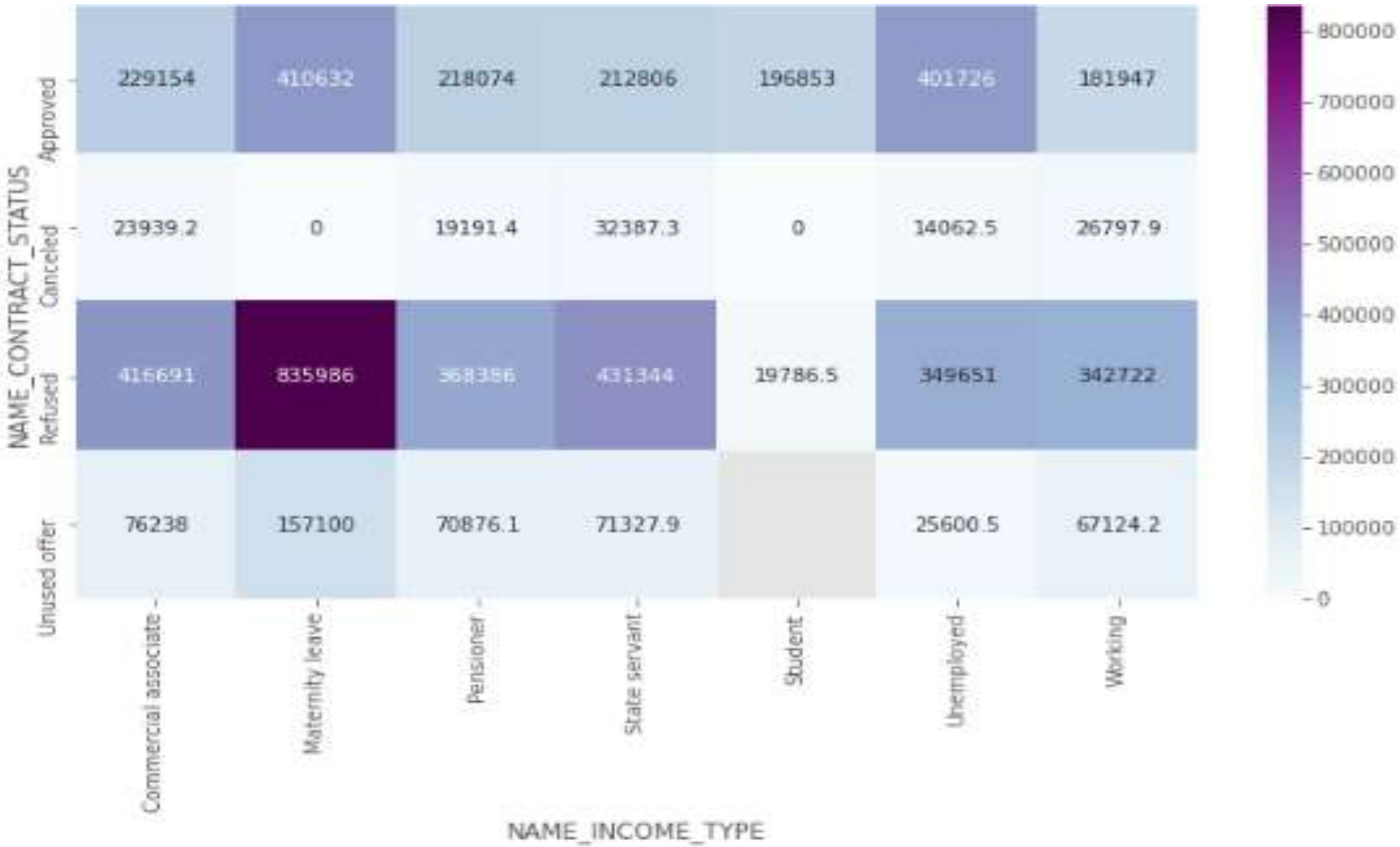
AMT_GOODS_PRICE, AMT_ANNUITY, AMT_APPLICATION - as expected have high correlation. Higher the value of good purchased more there will be need of loan and surely all these will correlate Similary, AMT_Credit to AMT_GOOD_PRICE also the correlation is high Column CNT_Payment ideally should have had a high correlation with AMT_credit, ie higher credit, more the term of loan. But no such correaltion can be seen.

Using Heatmap for checking "NAME_CONTRACT_STATUS", "NAME_INCOME_TYPE" aggregating on Target



Working applicant with Approved status have defaulted in highest numbers.Previous applications with Refused, Cancelled, Unused loans also have default which need to be concern. This indicates that the financial company had Refused/cancelled previous application, but has approved the current and is facing default on these loans.14,204 applicanst of working class were refused earlier and now have defaulted.

Checking "NAME_CONTRACT_STATUS", "NAME_INCOME_TYPE",aggregating on Target



Highely refused in Maternity leave and approved load is 410632

Observation from both the data

- All the below variables were established in analysis of Application data frame as leading to default.
- Checked these against the Approved loans which have defaults, and it proves to be correct
- -Medium income
- -25-35 years old , followed by 35-45 years age group
- -Male
- -Unemployed
- -Laborers, Salesman, Drivers
- -Business type 3
- -Own House - No
- Other IMPORTANT Factors to be considered
- -Days last phone number changed - Lower figure points at concern
- -No of Bureau Hits in last week. Month etc – zero hits is good
- -Amount income not correspondingly equivalent to Good Bought – Income low and good value high is a concern
- -Previous applications with Refused, Cancelled, Unused loans also have default which is a matter of concern. This indicates that the financial company had Refused/Cancelled previous application but has approved the current and is facing default on these.
- *

Credible Applications refused

- Unused applications have lower loan amount. Is this the reason for no usage?
- Female applicants should be given extra weight age as defaults are lesser.
- 60% of defaulters are Working applicants. This does not mean working applicants must be refused. Proper scrutiny of other parameters needed
- Previous applications with Refused, Cancelled, Unused loans also have cases where payments are coming on time in current application. This indicates that possibly wrong decisions were done in those cases.

To major variables to consider for load prediction

- NAME_EDUCATION_TYPE
- DAYS_BIRTH
- AMT_CREDIT
- CODE_GENDER
- AMT_ANNUITY
- NAME_INCOME_TYPE
- NAME_HOUSING_TYPE
- DAYS_EMPLOYED
- AMT_INCOME_TOTAL

THANK YOU