CREDIT EDA CASE STUDY

PGDDS C26 November 2020

Presented By

RAKESH YADAV & KAVITA MALI

Introduction

1 Objective

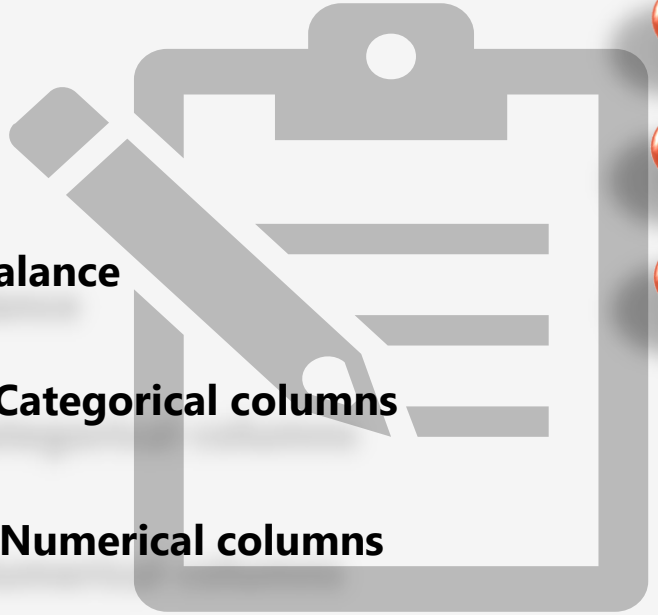
- When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- 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.

2 Data set used

- Current applications → "application_data.csv"
- Previous applications → "previous_application.csv"

Steps Involved

1. Loading Data
2. Inspecting Data
3. Data Analysis
4. Checking the Data Imbalance
5. Univariate analysis for Categorical columns
6. Univariate Analysis for Numerical columns
7. Bivariate Analysis : Numerical – Categorical
8. Bivariate Analysis : Numerical – Numerical
9. Correlations
10. Loading Data "previous_application.csv"
11. Final Conclusions



2.Loading Data / EDA

EDA ANALYSIS

Reading Data

1. Reading Dataset **application_data.csv.**
2. Reading Dataset **previous_application.csv.**

Inspecting Data frame

1. *Inspecting and understanding Data*
   - Checking few records of Dataset such as .shape, .info(), .describe() .
2. *Data Cleaning*
   - Checking the percentage of null values in the data frame in descending order.
   - Analyzing number of null columns
   - Dropped columns having null values > 35%
   - Imputing Columns having null values ≤ 19% with Mode values for numeric columns except for continuous numeric columns we imputed with Median value.
3. *Handling Errors in Data types and Data*
   - Checking the percentage of null values in the data frame in descending order.
   - Analyzing number of null columns : **49**
   - Dropped columns having null values > 35%
   - Imputing Columns having null values ≤ 19% with Mode values for numeric columns except for continuous numeric columns we imputed with Median value.
   - After observing data frame, we find columns: ***'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH' and 'DAYS_LAST_PHONE_CHANGE'*** which had negative or mixed values, So we imputed them with absolute values for our analysis.
   - Then we changed the values of columns **FLAG_OWN_CAR** and **FLAG_OWN_REALTY** from 'Y' and 'N' to 1 and 0 respectively for convenience in analysis.

- We found that the column CODE_GENDER having value 'XNA' which simply means *Not Available* . So, we imputed those values with the most frequent value(mode value) i.e., F.
- Then we found that column **ORGANIZATION_TYPE** also have 18% 'XNA' values. So firstly, we checked that whether the values are *Missing Completely at Random(MCAR), Missing at Random(MAR) or Missing Not at Random(MNAR).After* comparing the values of ORGANIZATION_TYPE with third variables i.e., values of column NAME_INCOME_TYPE , we found that Clients who are pensioner were having XNA values in ORGANIZATION_TYPE. So, we replaced "XNA" with "Pensioner".
- Similarly, we found the same relation between OCCUPATION_TYPE and NAME_INCOME_TYPE , so we imputed null values with "Pensioner".
- For convenience to our analysis, we made binning based on the quantiles for the columns mentioned below:
  AMT_INCOME_TOTAL , AMT_CREDIT *AND AGE_DAYS*

3.Analysis

3.1 Data Types Conversion's for better understanding of Variables

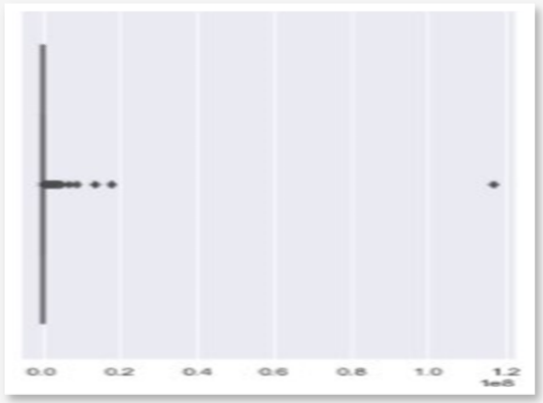
In data frame, we observed some columns having d-type “object” ,which can be converted to d-type “category” which will be convenient for our analysis as well as it will save memory usage.

There are so many columns in data frame. We’ll remove columns which we don’t need for further analysis.

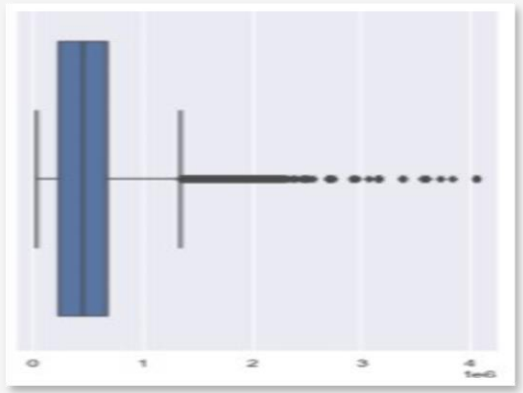
3.2 Finding and Analyzing outliers

We made list of all numerical columns .And then plot Boxplots for each numerical columns.

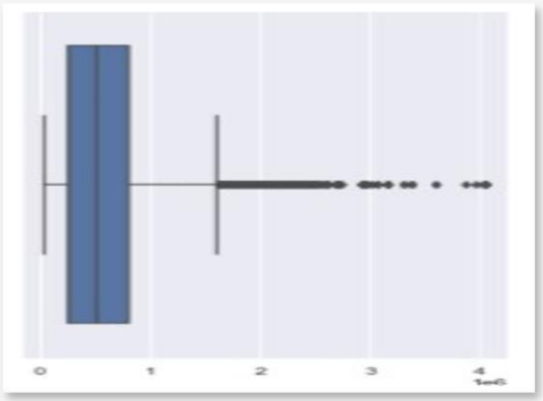
- IQR for **AMT_INCOME_TOTAL** is very slim, and it have many outliers. It has Maximum value of 117000000 which is a huge variation from 75th percentile.
- Third quartile of **AMT_GOODS_PRICE, AMT_CREDIT** is larger as compared to First quartile which means that most of the **Credit amount of the loan** of customers are present in the third quartile. Here maximum value is 4050000 which varies a lot from 75th percentile.
- **DAYS_EMPLOYED** have maximum value at 375000 ,which vary a lot from mean and 75th percentile.
- Visual Representation by Boxplot is shown



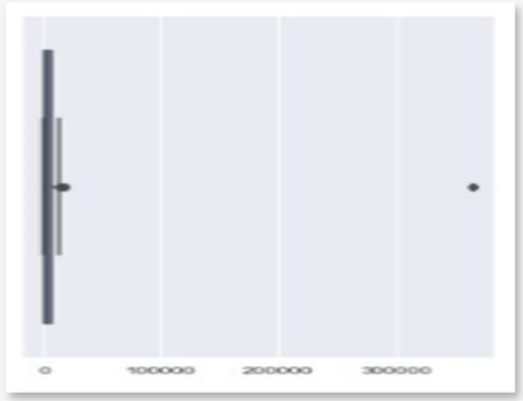
AMT_INCOME_TOTAL



AMT_GOODS_PRICE



AMT_CREDIT



DAYS_EMPLOYED

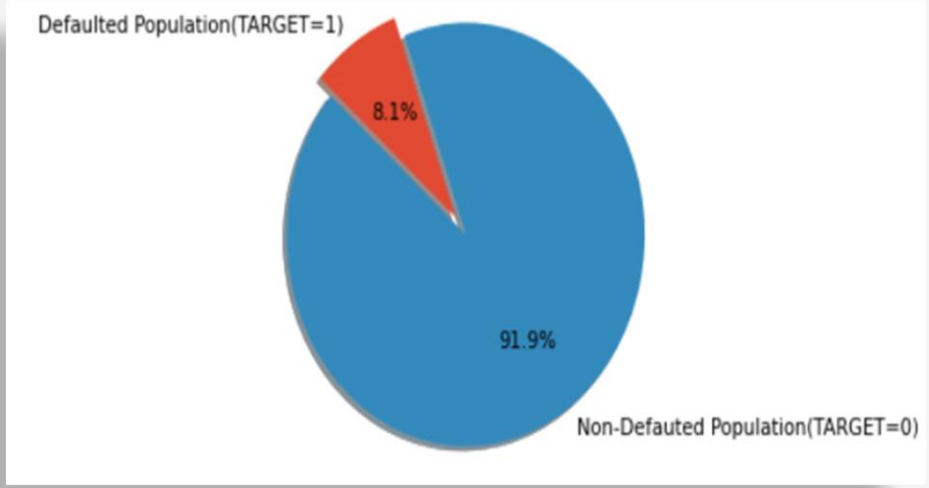
4. Checking Imbalance

4.1 Checking Data Imbalance for Target Variable

Since there is a huge imbalance between the TARGET variables 0 and 1, it makes more sense to divide data frame into two sub datasets then continue our analysis.

We have splits data frame as follows:

- **Target0 :** (Non-Defaulted Population) Clients without Payment Difficulties.
- **Target1 :** (Defaulted Population) Clients with Payment Difficulties.



Conclusion

We have found that -

Ratio of Data Imbalance is "**11.3**"

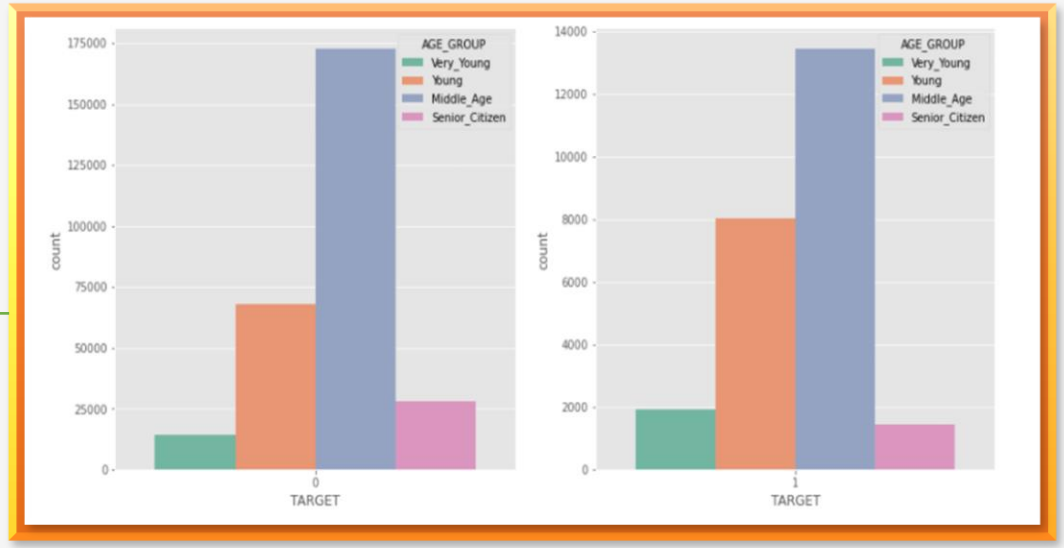
In order to analyze the imbalance and various aspects of data we will perform various types of analysis such as:

- Univariate analysis , Bivariate analysis , Multivariate analysis

5.Univariante Analysis Categorical

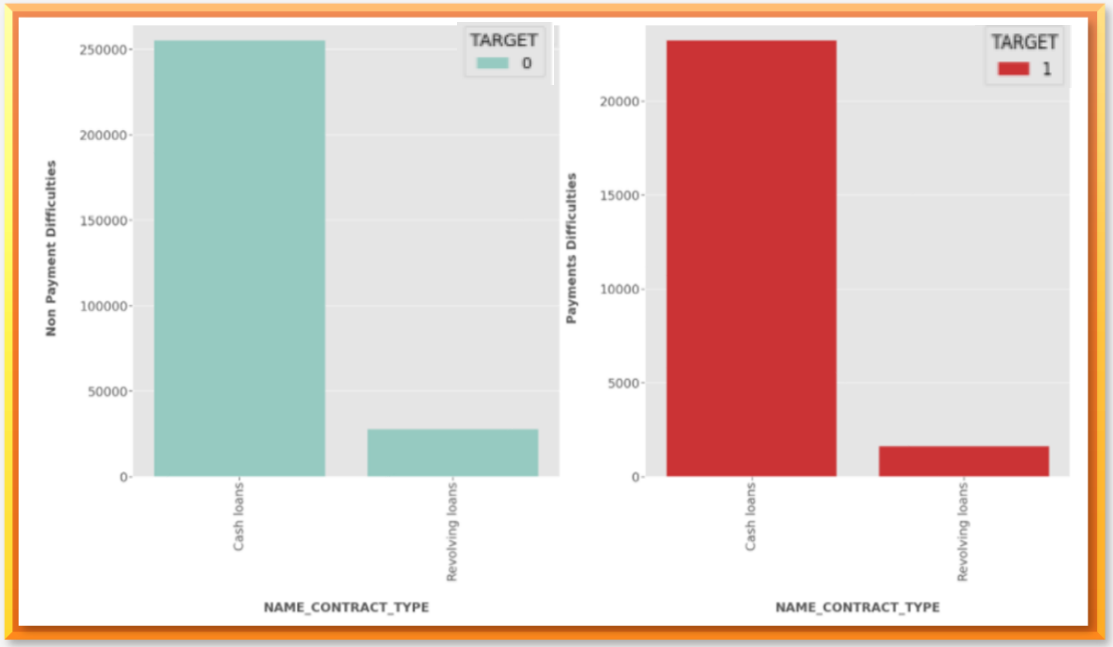
5.1 Gender/Age Group with respect to Target variables

- It seems like Female clients applied higher than male clients for loan.
- 66.6% Female clients and 33.4% male clients are payment difficulties.
- 57% Female clients and 42% male clients are with payment difficulties.

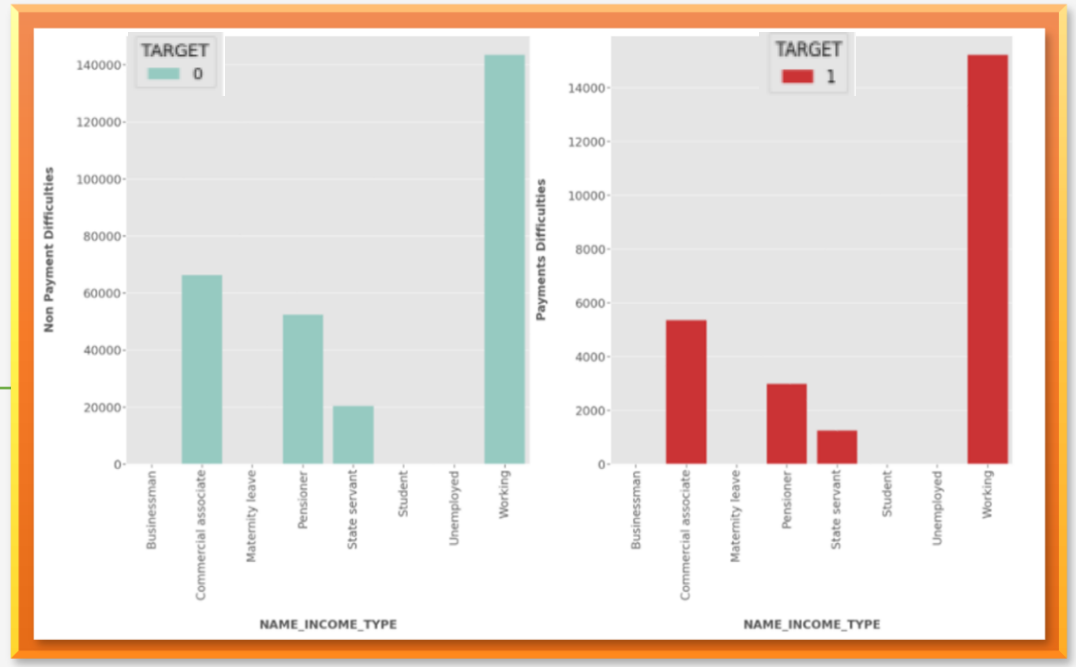


- Middle Age Group(35 – 60) have applied most and have higher payment difficulties amongst all.
- While Senior Citizens(60 - 100) and Very young (19 - 25) age group facing fewer paying difficulties as compared to other age groups

5.2 CONTRACT/INCOME Type with respect to Target variables

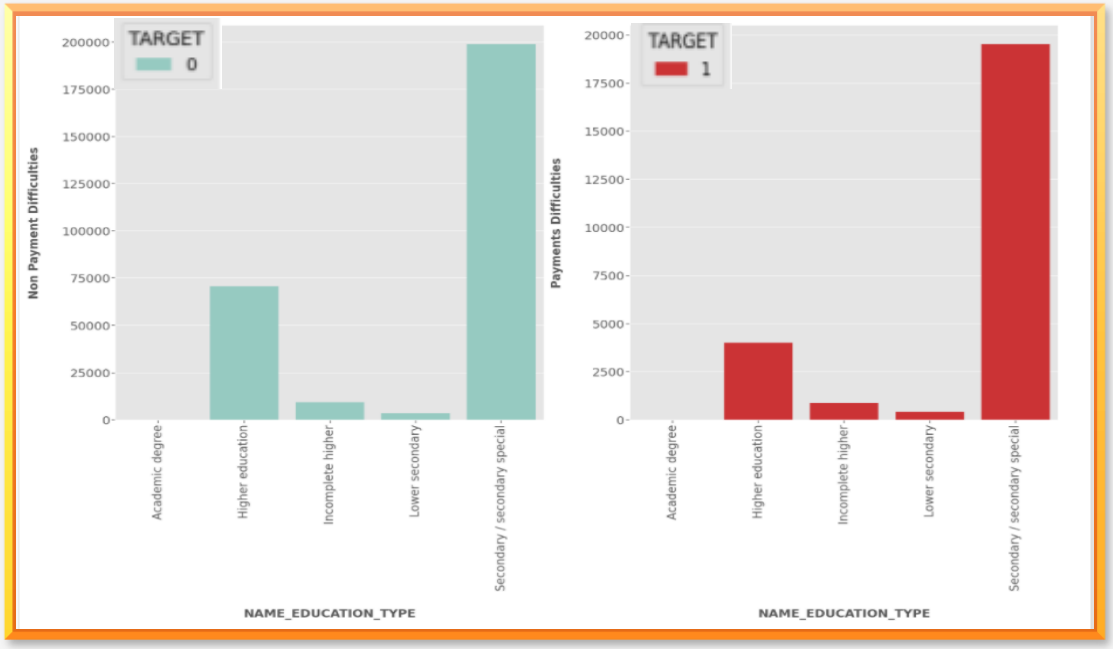


- Most of the clients applied for the Cash Loan while a very small proportion applied for Revolving Loan.
- But Clients applied for Cash Loan have higher payment difficulties.

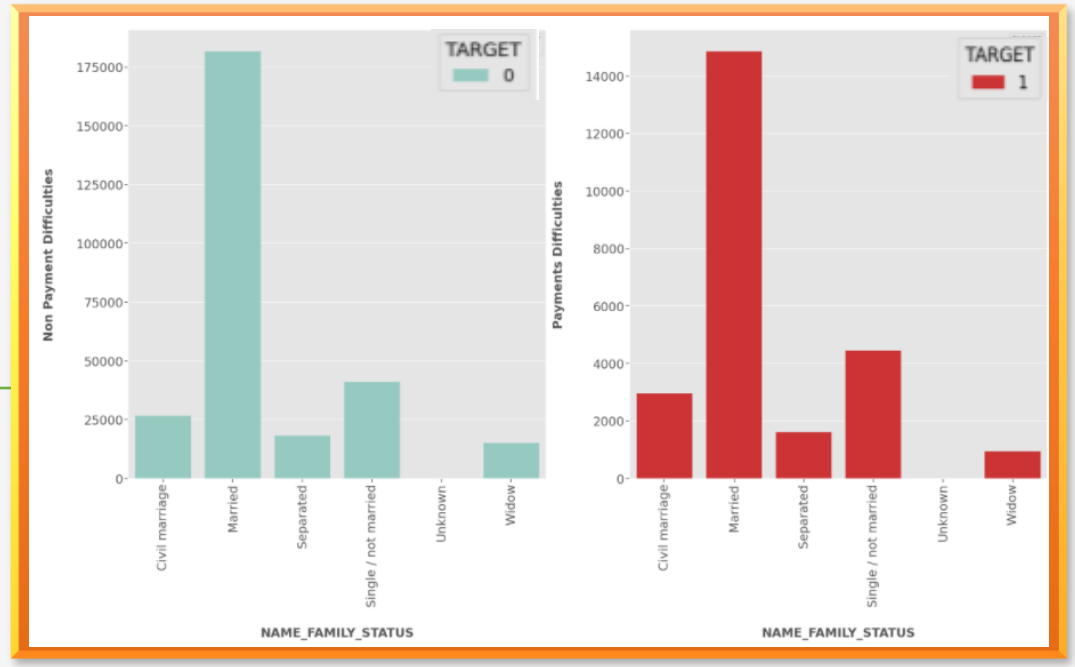


- Clients who applied for loans were getting income by Working, Commercial associate and Pensioner are more likely to apply for the loan, highest being the Working-class category .
- Businessman, students and Unemployed less likely to apply for loan .
- Working category have high risk to default.
- State Servant is at Minimal risk to default.

5.3 Education/Family Type with respect to Target variables

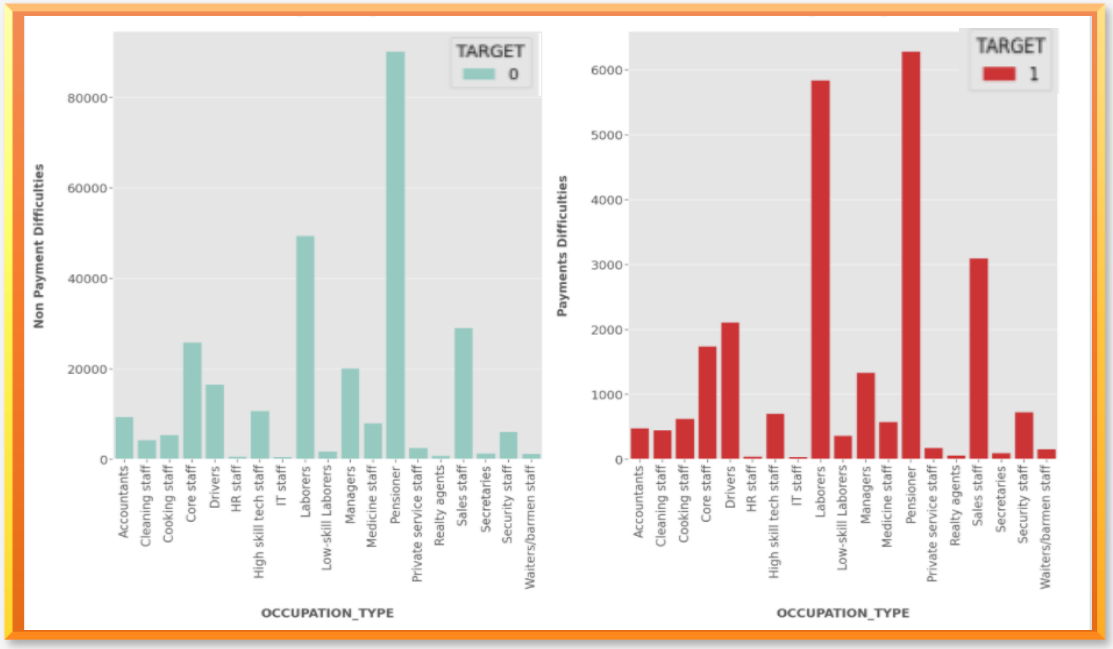


- Clients having Secondary/Secondary Special Education are more likely to apply for the Loan.
- Also, clients having Secondary/Secondary Special Education are facing higher payment difficulties ,so they have high risk to default. Other education types are at minimal risk

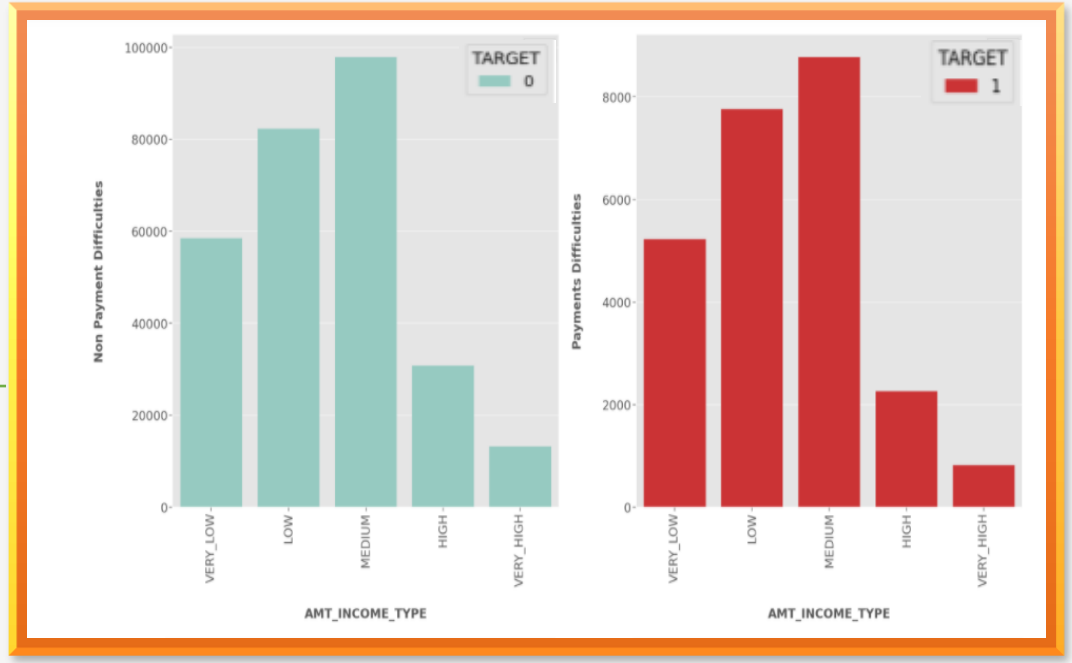


- It seems that Married clients applied most for the loan and have higher payment difficulties.
- Widows are less likely to apply for the loan and have minimal risk.
- Clients with the single relationship have minimal risk to default i.e., have less payment difficulties.

5.3 Occupation/Income Range with respect to Target variables

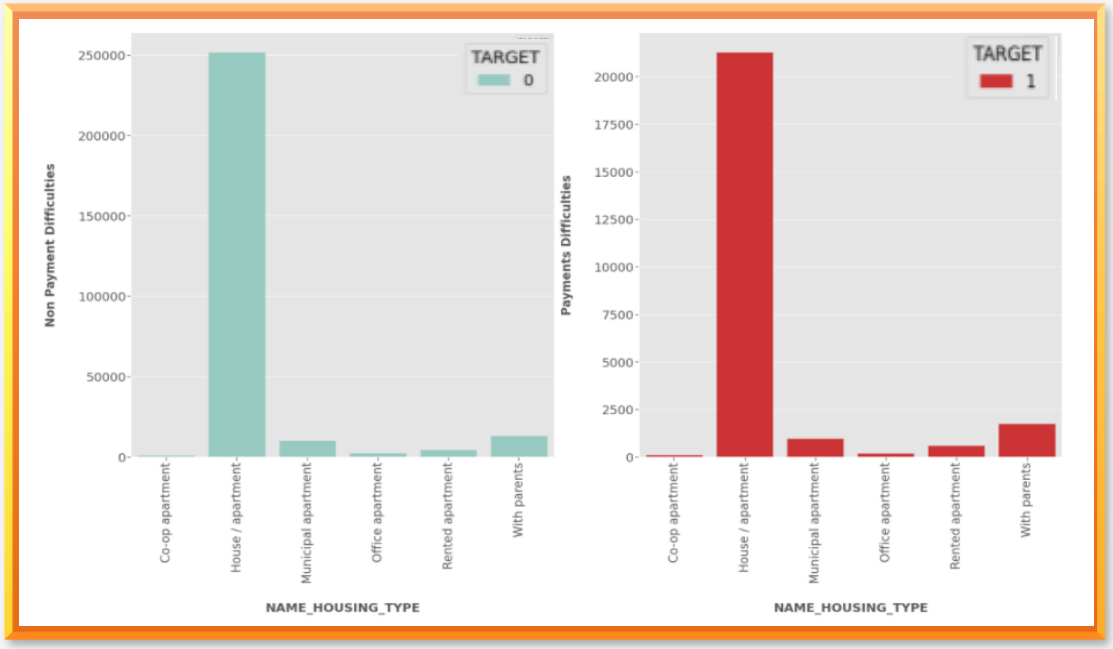


- Pensioners and Laborers have applied the most for the loan.
- Pensioner being highest followed by laborers have higher payment difficulties , so have high risk to default.

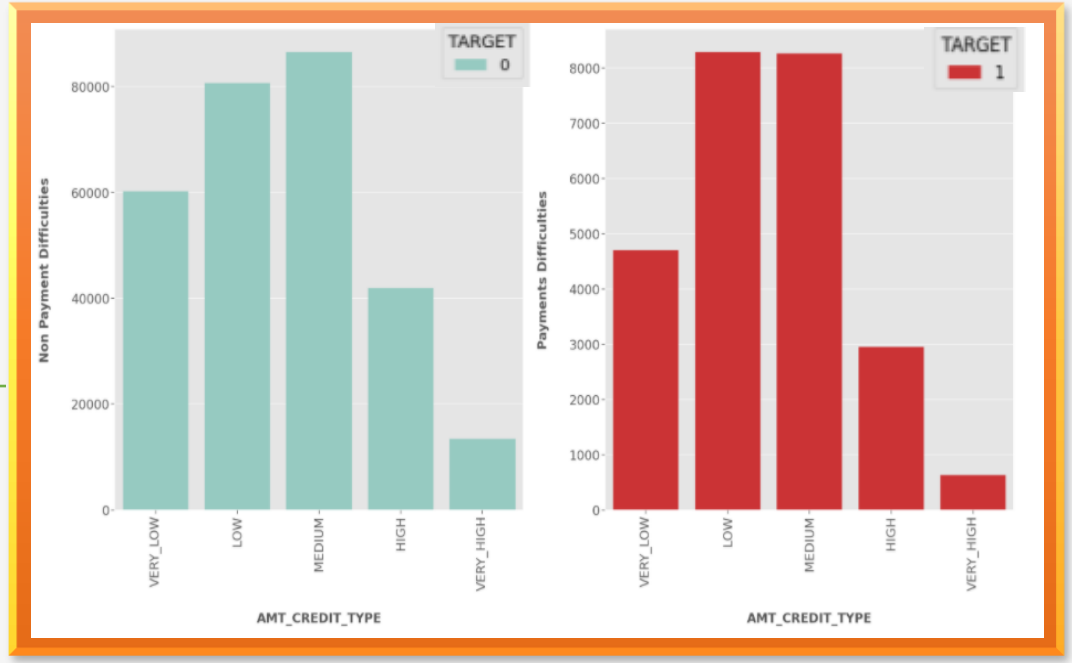


- Clients having **Medium salary range** are more likely to apply for the loan. And have higher payment difficulties.
- Clients having low and medium income are at high risk to default.
- Clients having high salaries are at minimal risk.

5.3 Housing Type/Credit Range with respect to Target variables



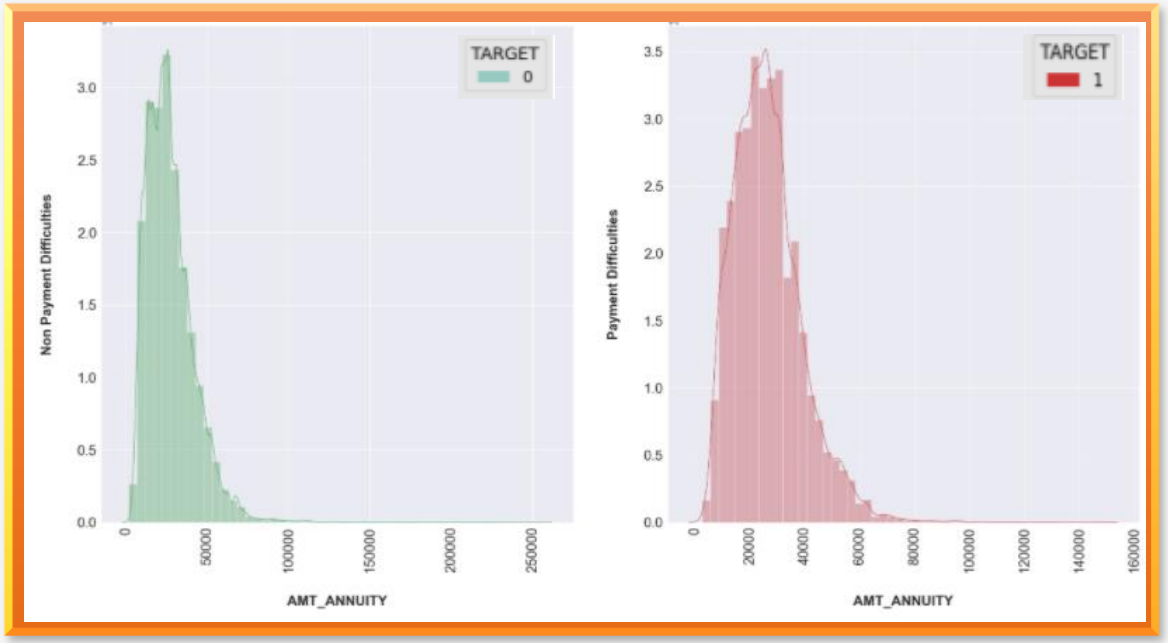
- Most of the clients applied for the loan owns a house/apartment and have a higher payment difficulties.
- Other clients have less payment difficulties with low applications for loan.



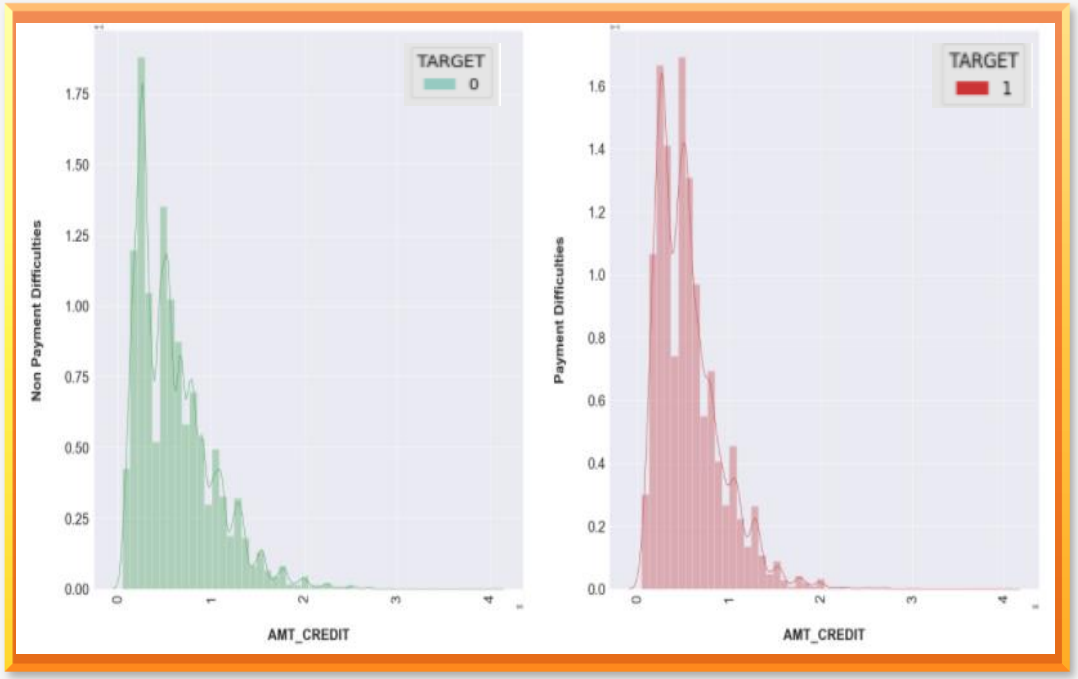
- Most of the clients applied for Medium Credit Amount for the loan .
- Clients applying for medium and low credit have high payment difficulties and have high risk to default.

6.Univariante Analysis Numerical

6.1 Annuity/Credit with respect to Target variables

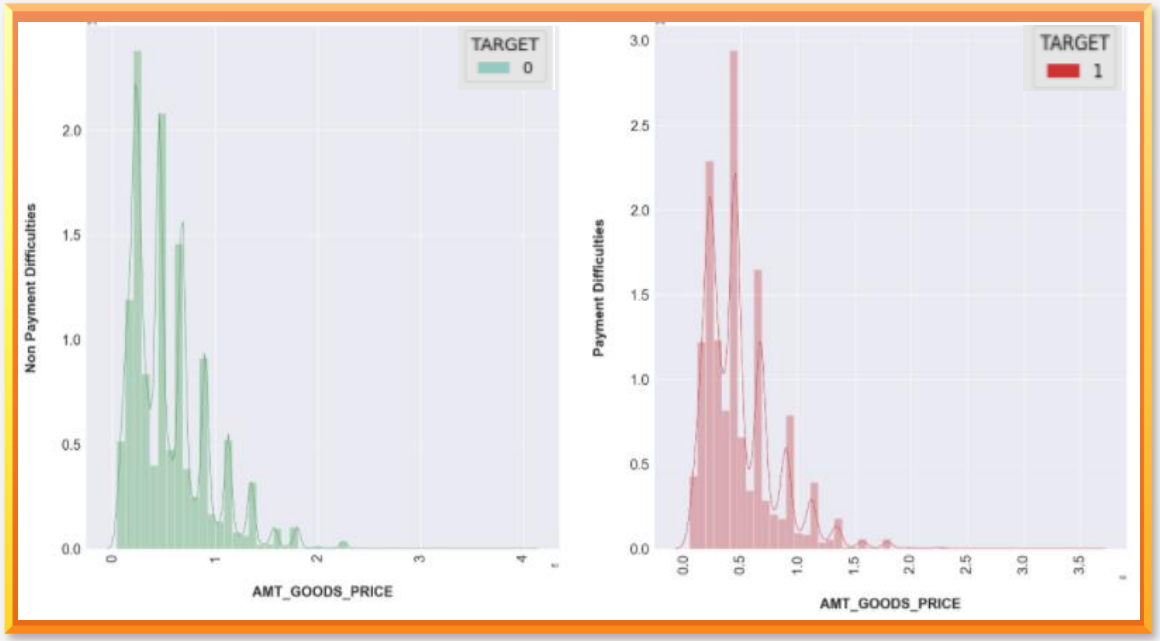


- Distribution of AMT_ANNUITY for Target1 is broader than Target0

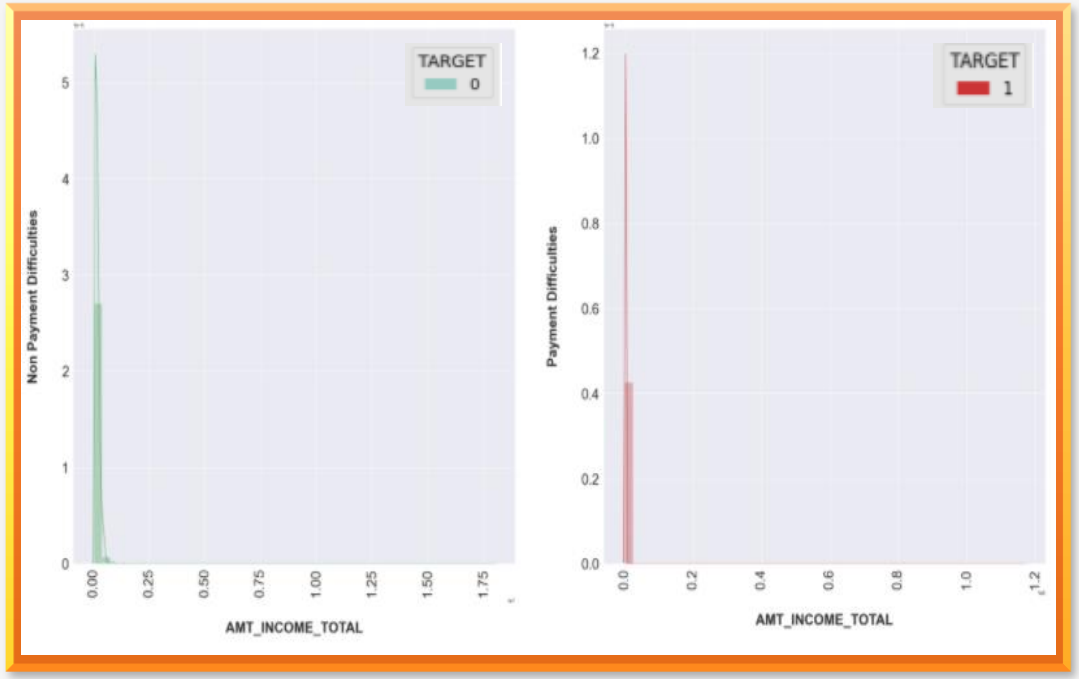


- Distribution of AMT_CREDIT for Target1 and Target0 is similar.

6.2 Goods Price/ Amt Income with respect to Target variables



- Distribution of AMT_GOODS_PRICE for Target1 and Target0 is similar.



- Distribution of _INCOME_TOTAL for Target1 and Target0 is similar.

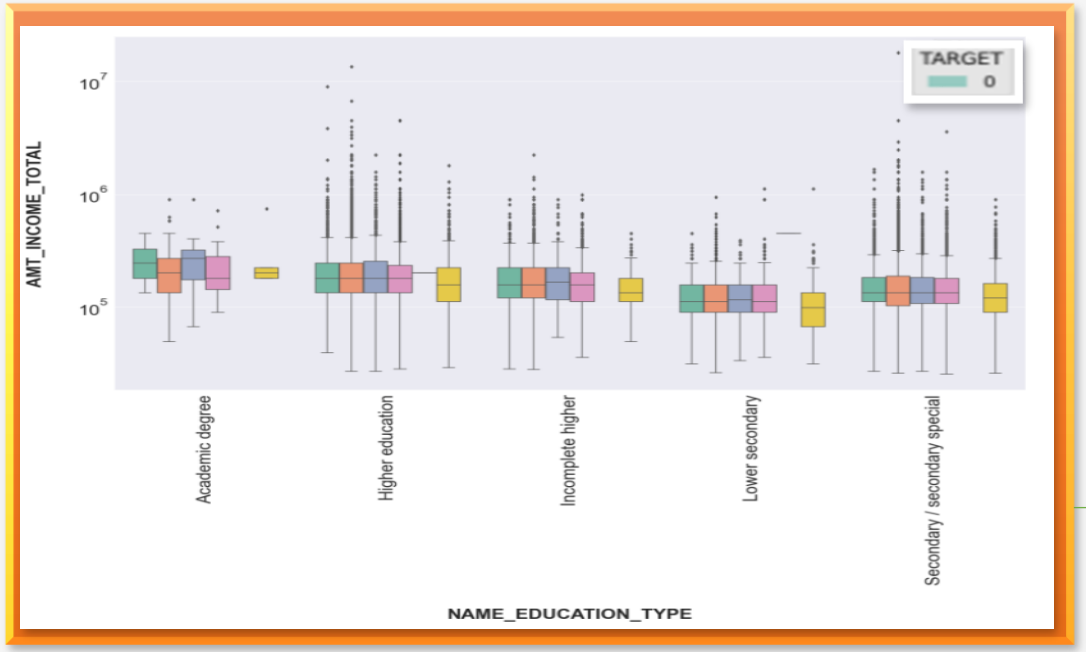
6.3 Points To Conclude

Notes

- People with target one has largely staggered income as compared to target zero. Dist. plot clearly shows that the shape in Income total, Annuity, Credit and Good Price are similar for Target 0 and similar for Target 1.
- The plots are also highlighting that people who have difficulty in paying back loans with respect to their income, loan amount, price of goods against which loan is procured and Annuity.
- Dist. plot highlights the curve shape which is wider for Target 1 in comparison to Target 0 which is narrower with well defined edges.

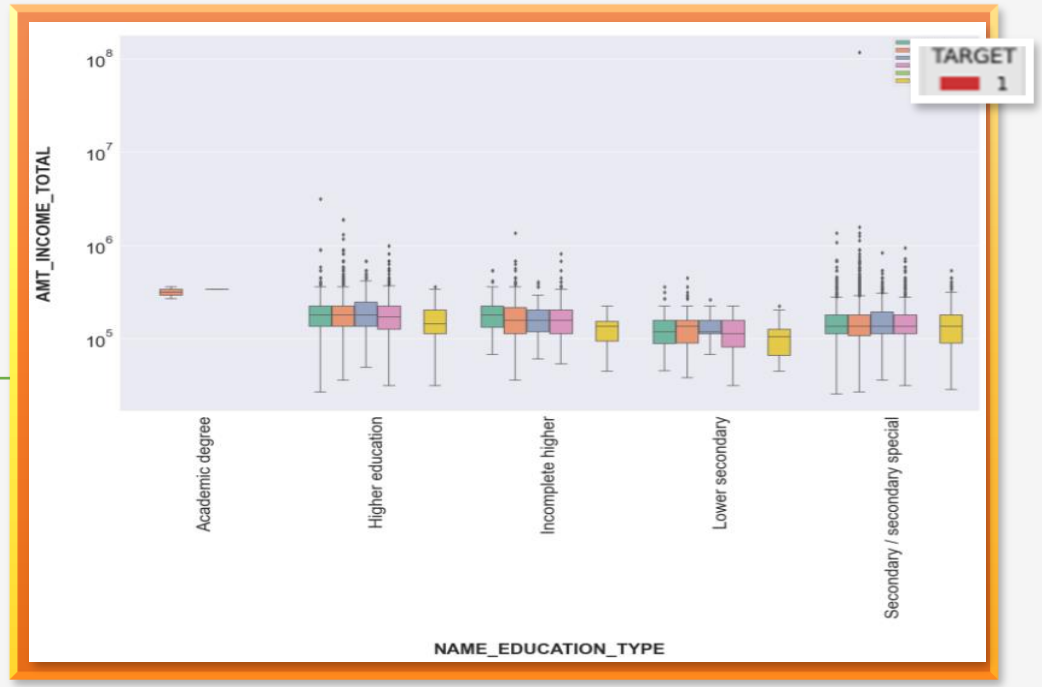
7. Bivariate Analysis
Numerical V/s Categorical

7.1 Income_Amount Vs Education_Status Vs FAMILY_Status among Target Variables

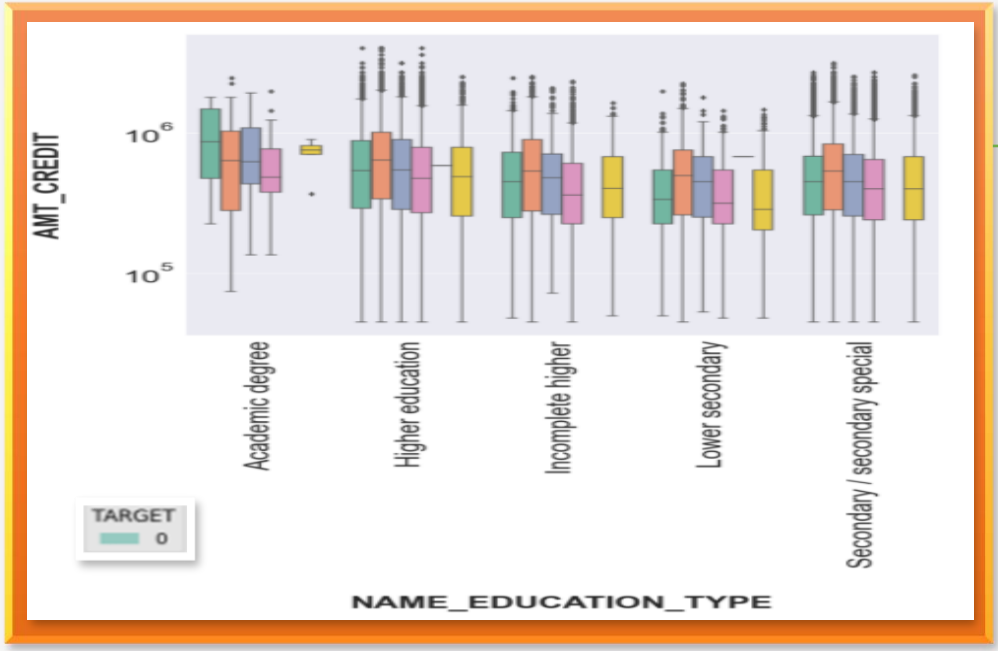


- Widow Client with Academic degree have a very few outliers and doesn't have First and Third quartile. Also, Clients with all type of family status having academic degree have very less outliers as compared to other type of education.
- Income of the clients with all type of family status having rest of the education type lie Below the First quartile i.e., 25%.

- Income amount for Married clients with academic degree is much lesser as compared to others.
- (Defaulter) Clients have relatively less income as compared to Non-defaulters.

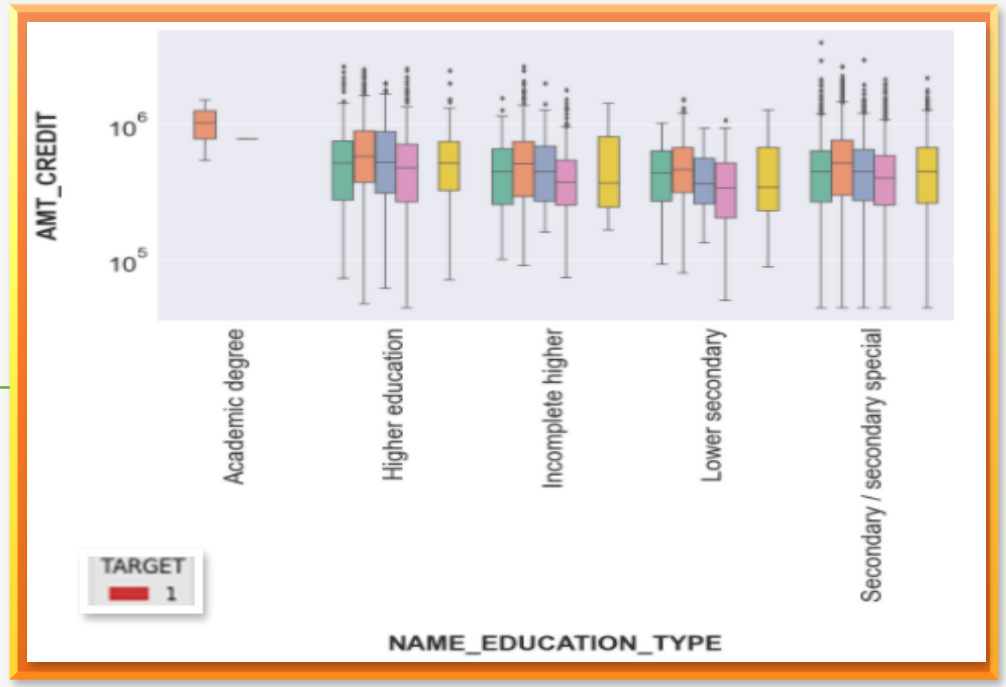


7.2 Credit Vs Education_Status Vs FAMILY_Status among Target Variables



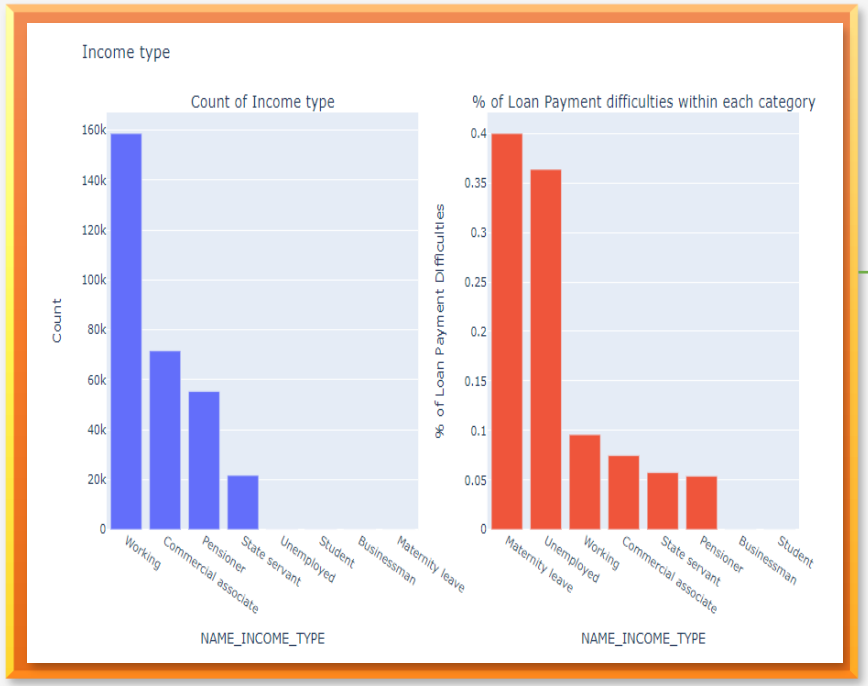
- Clients with all Education type except Academic degree have large number of outliers
- Most of the population i.e., clients credit amounts lie below 25th percentile.
- Clients with Academic degree and who is a widow tend to take higher credit loan.
- Some of the clients with Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special Education are more likely to take high amount of credit loan.

- Married client with academic applied for higher credit loan. And doesn't have outliers. Single clients with academic degree have a very slim boxplot with no outliers.
- Some of the clients with Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special Education are more likely to take high amount of credit loan.

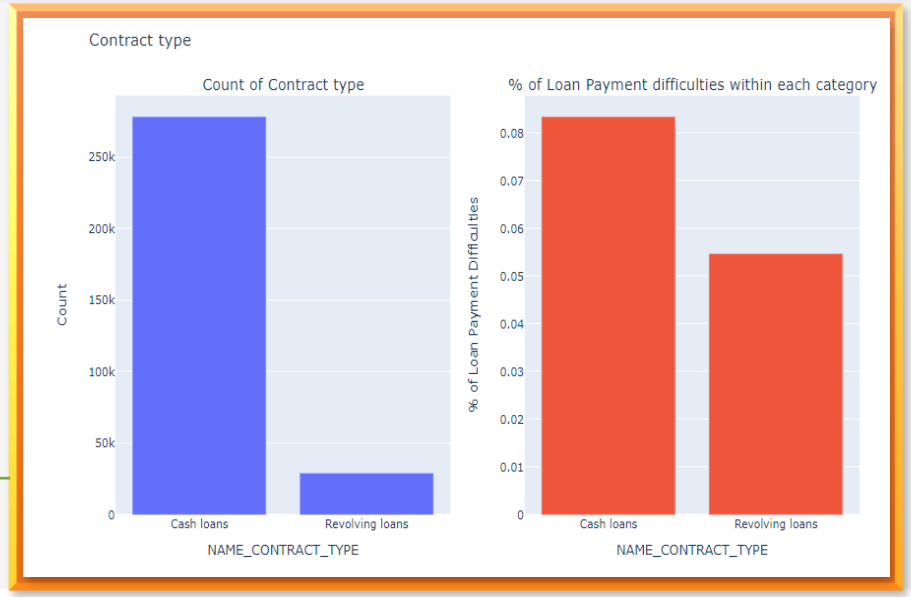


8. Bivariate Analysis
Categorical V/s Categorical

8.1 Categories having Maximum % of Risk of Default using “Biplots”



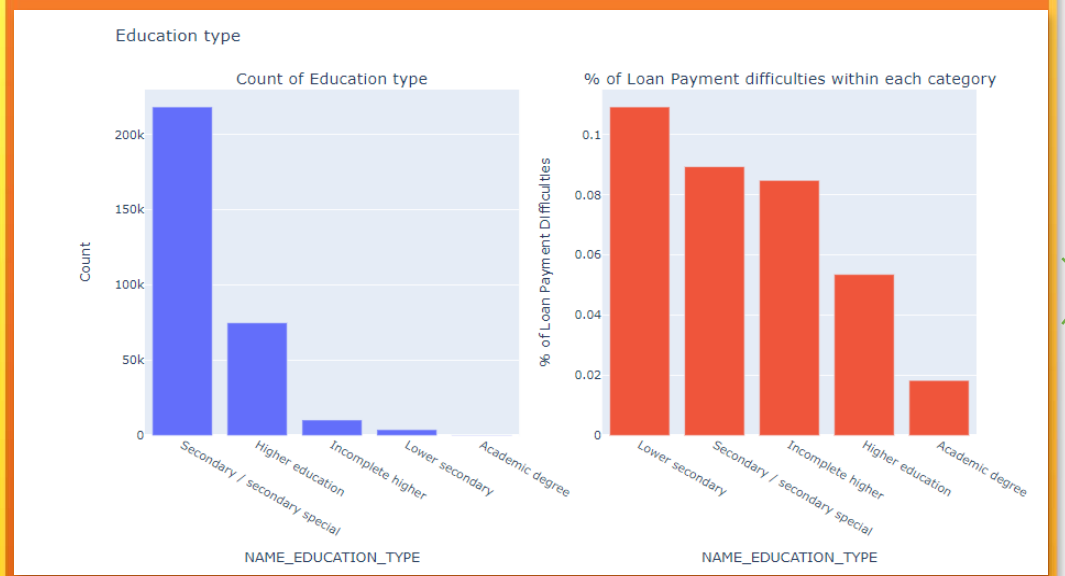
- **Though count of working clients applying for loan is significantly high , risk to default in payments is less as compared to others**
- **Count of clients with income type Maternity leave is only 5, but risk to default in payments for those is minimum among all the income types.**
- **Same condition is observed in case of unemployed. Though count is very low, risk to default in payments is low.**
- **Pensioner, State servant and Commercial associate have higher risk to default.**



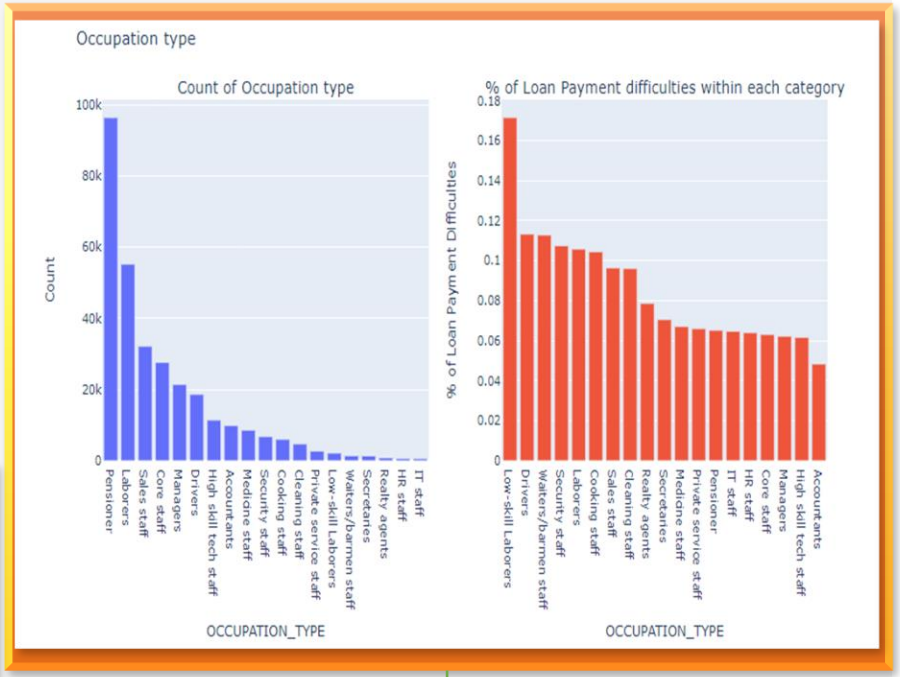
- Cash loans have higher risk to default, Revolving loans have comparatively lower risk for the same .

8.2 Categories having Maximum % of Risk of Default using "Biplots"

- Clients having Low income have high risk to default followed by clients with medium and very low income.
- Clients with high salaries have minimal risk to default.

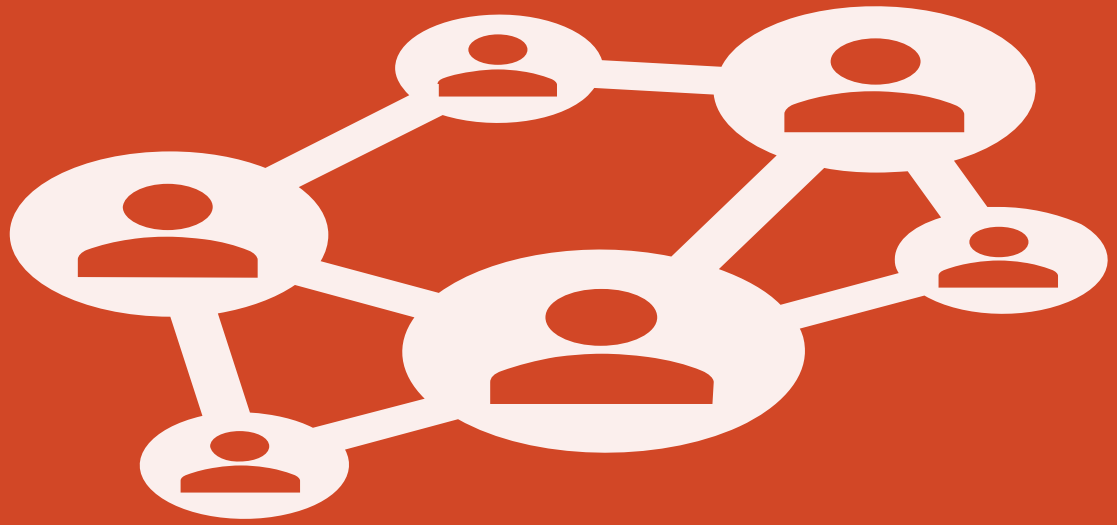


- Clients having Academic Degree and higher Education have lower risk to default.
- Clients having Lower Secondary , Secondary/Secondary Special Education have very high risk to default.



- Low-skill Laborers have higher risk to default
- Managers, High skill tech staff and Accountants have relatively lower risk to default

8.Correlation’s

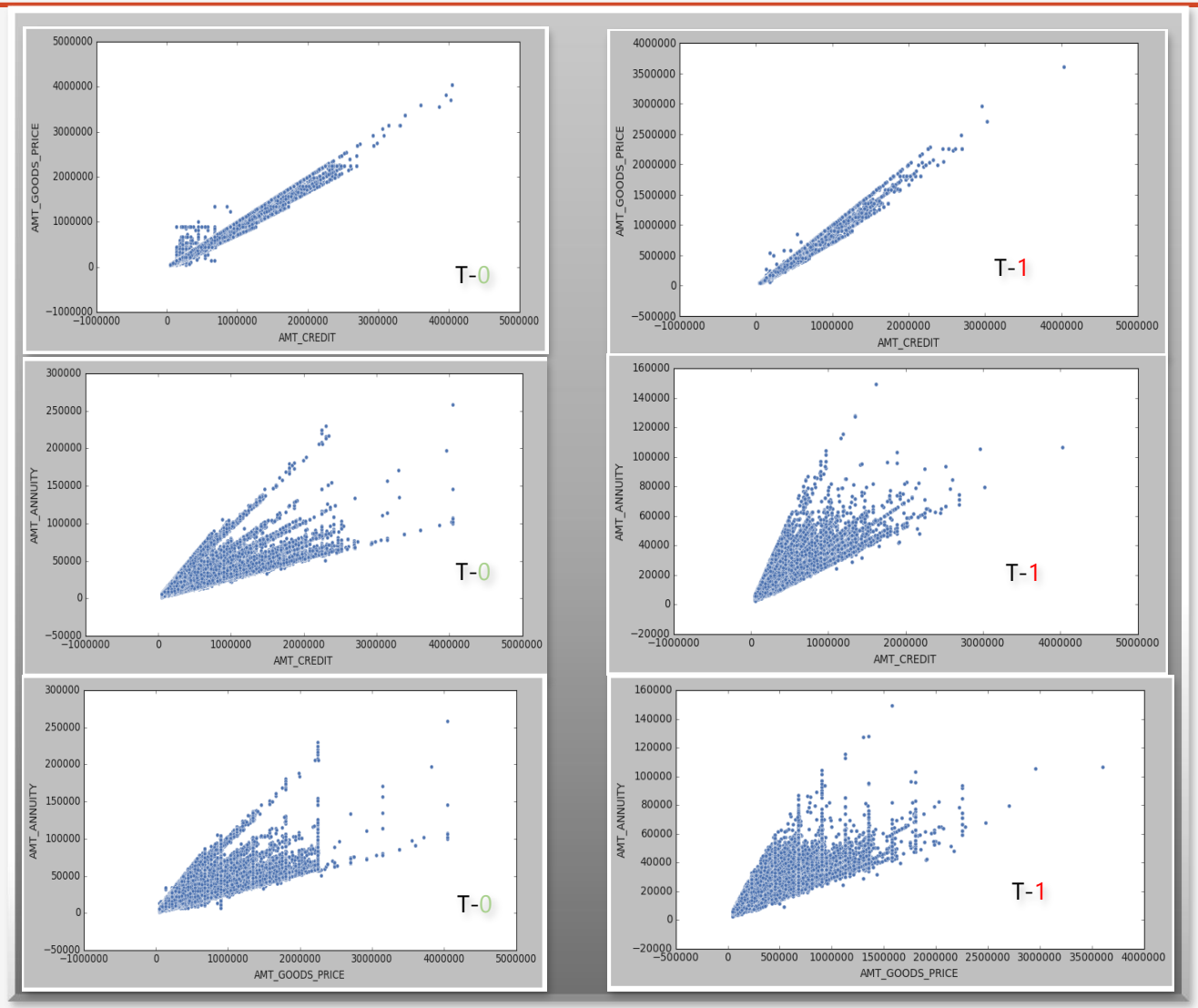


8.1 Correlations between numerical variables Using "Pair Plots "

GOODS PRICE Vs CREDIT AMOUNT

CREDIT AMOUNT Vs ANNUITY

ANNUITY Vs GOODS PRICE



Points to Mark:

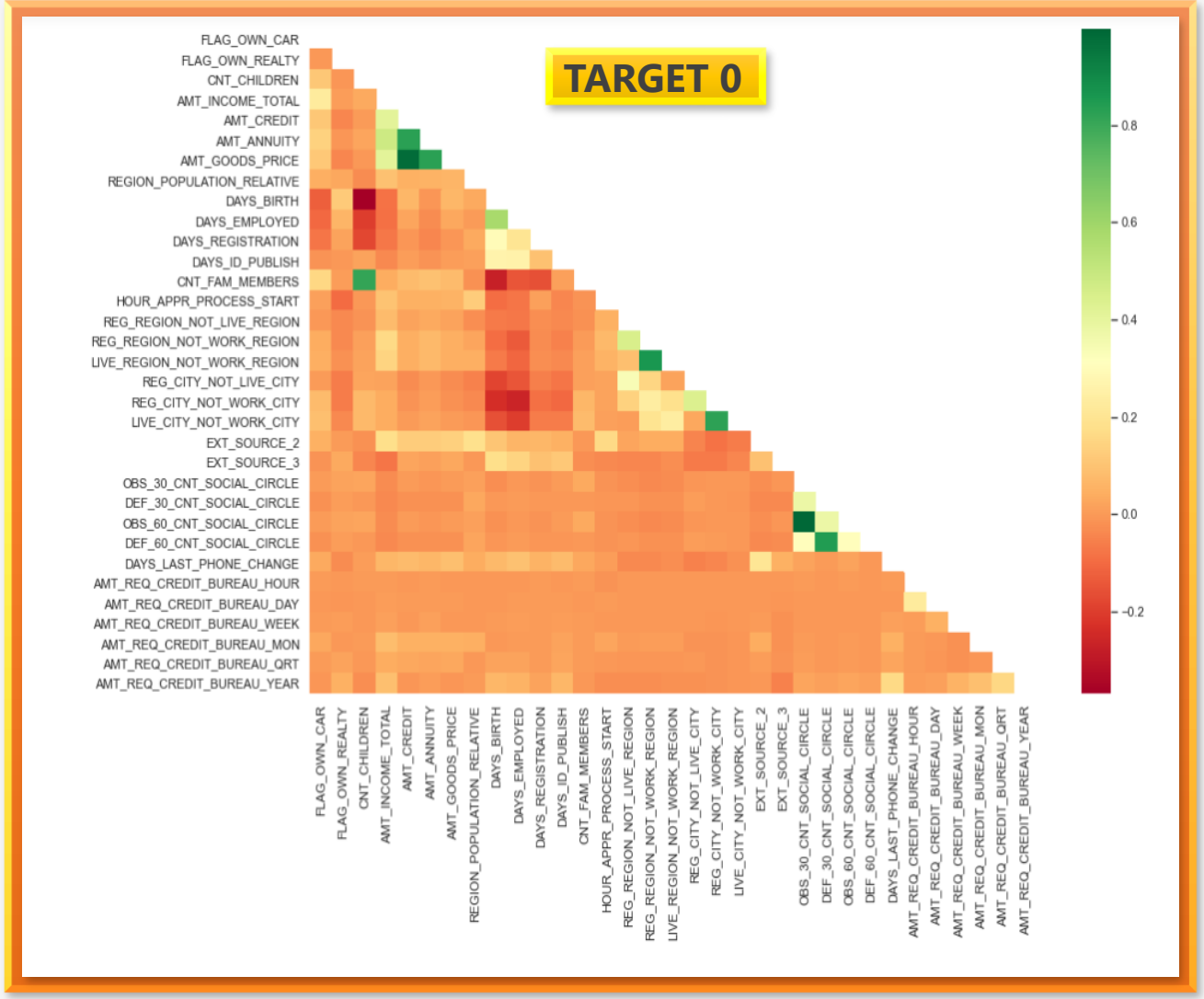
1. High correlated variables for both defaulters and non - defaulters. So as the home price increases the loan amount also increases

2. High correlated variables for both defaulters and non defaulters . So as the home price increases the EMI amount also increases which is logical

Conclusion

- All three variables Are highly correlated for both defaulters and non-defaulters, which might not give a good indicator for defaulter detection

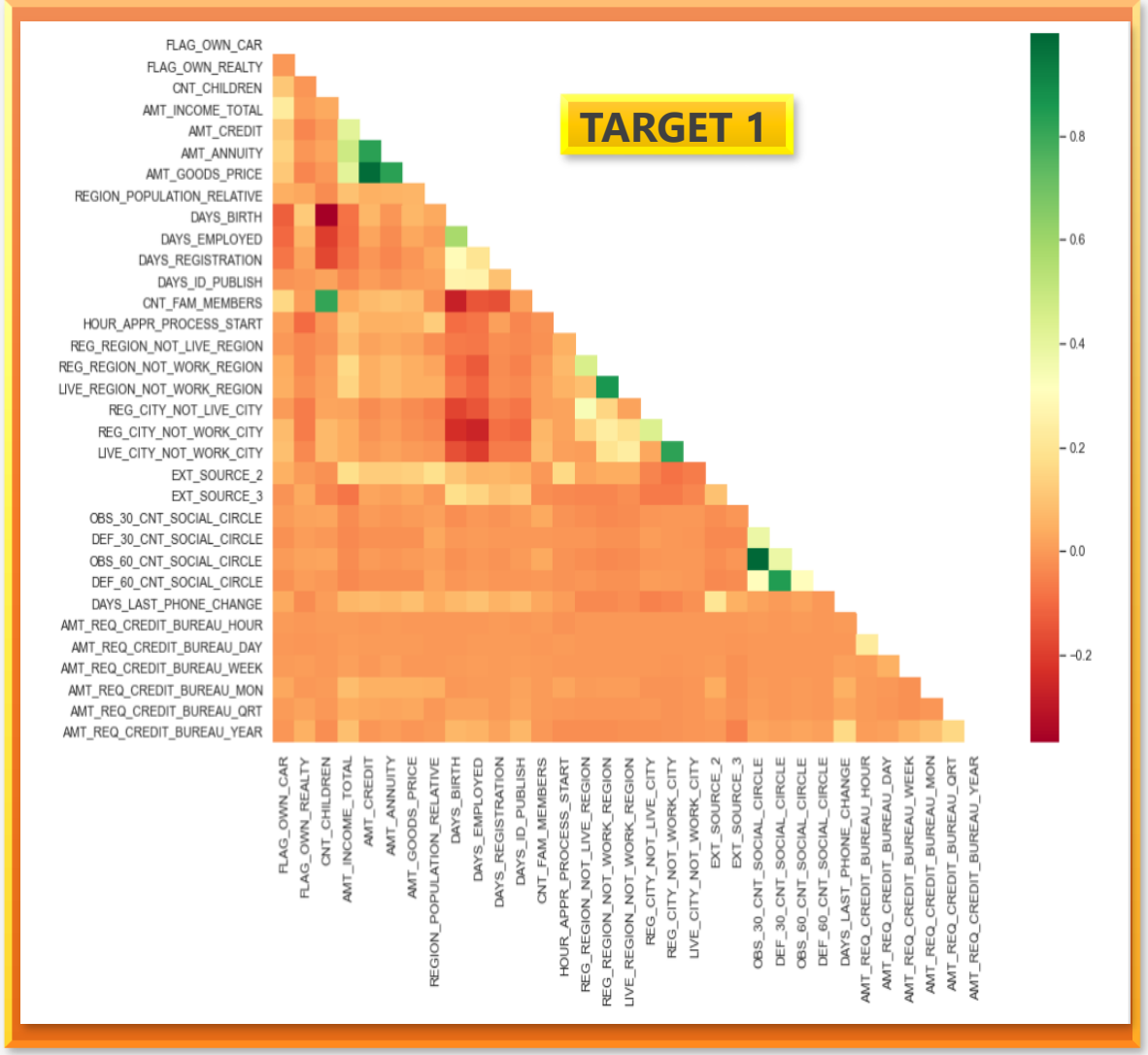
8.2 Correlations between numerical variables using "Heatmap's"



As we can see from correlation heat map for TARGET-0, There are number of observation we can point out

- **AMT_CREDIT is inversely proportional to the DAYS_BIRTH, peoples belongs to low-age group taking high Credit amount and vice-versa**
- **AMT_CREDIT is inversely proportional to the CNT_CHILDREN, means Credit amount is higher for less children count client have and vice-versa.**
- **AMT_INCOME_TOTAL is inversely proportional to the CNT_CHILDREN, means more income for less children client have and vice-versa.**
- **Less CNT_CHILDREN client have in densely populated area.**
- **AMT_CREDIT is higher to densely populated area.**
- **AMT_INCOME_TOTAL is also higher in densely populated area.**

8.2 Correlations between numerical variables using "Heatmap's"



- AMT_CREDIT is inversely proportional to the DAYS_BIRTH, peoples belongs to low-age group taking high Credit amount and vice-versa
- AMT_CREDIT is inversely proportional to the CNT_CHILDREN, means Credit amount is higher for less children count client have and vice-versa.
- AMT_INCOME_TOTAL is inversely proportional to the CNT_CHILDREN, means more income for less children client have and vice-versa.
- Less CNT_CHILDREN client have in densely populated area.
- AMT_CREDIT is higher to densely populated area.
- AMT_INCOME_TOTAL is also higher in densely populated area.

Conclusion

This heat map for Target 1 is also having quite a same observation just like Previous Target 0. But for few points are different. They are listed below.

- **The client's permanent address does not match contact address are having less children.**
- **The client's permanent address does not match work address are having less children.**

9. Top 10 Correlation’s

9.1 Correlation Remarks

- Top 10 correlations between both Default (TARGET 1) and non default(TARGET 0) clients are almost at the same level for different variables

TARGET 0

VAR1	VAR2	CORRELATION	CORR_ABS
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998510	0.998510
AMT_GOODS_PRICE	AMT_CREDIT	0.987250	0.987250
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571	0.878571
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861	0.861861
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859371	0.859371
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381	0.830381
AMT_GOODS_PRICE	AMT_ANNUITY	0.776686	0.776686
AMT_ANNUITY	AMT_CREDIT	0.771309	0.771309
DAYS_EMPLOYED	DAYS_BIRTH	0.626028	0.626028
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.446101	0.446101

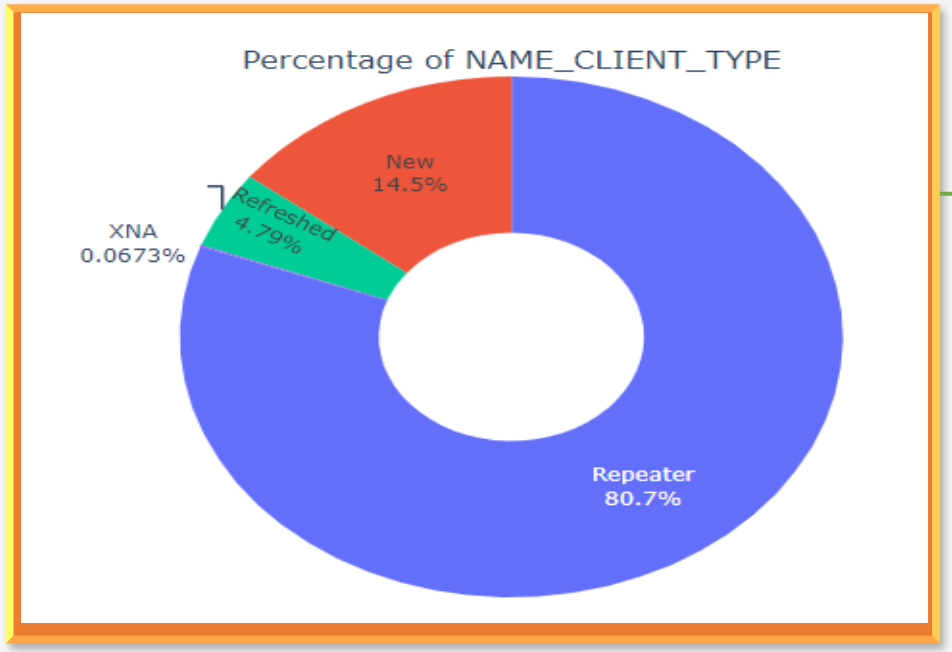
V/S

TARGET 1

VAR1	VAR2	CORRELATION	CORR_ABS
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998270	0.998510
AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.987250
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.878571
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885	0.861861
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869016	0.859371
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540	0.830381
AMT_GOODS_PRICE	AMT_ANNUITY	0.752699	0.776686
AMT_ANNUITY	AMT_CREDIT	0.752195	0.771309
DAYS_EMPLOYED	DAYS_BIRTH	0.582441	0.626028
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.497937	0.446101

10. Loan Distributions and Purposes

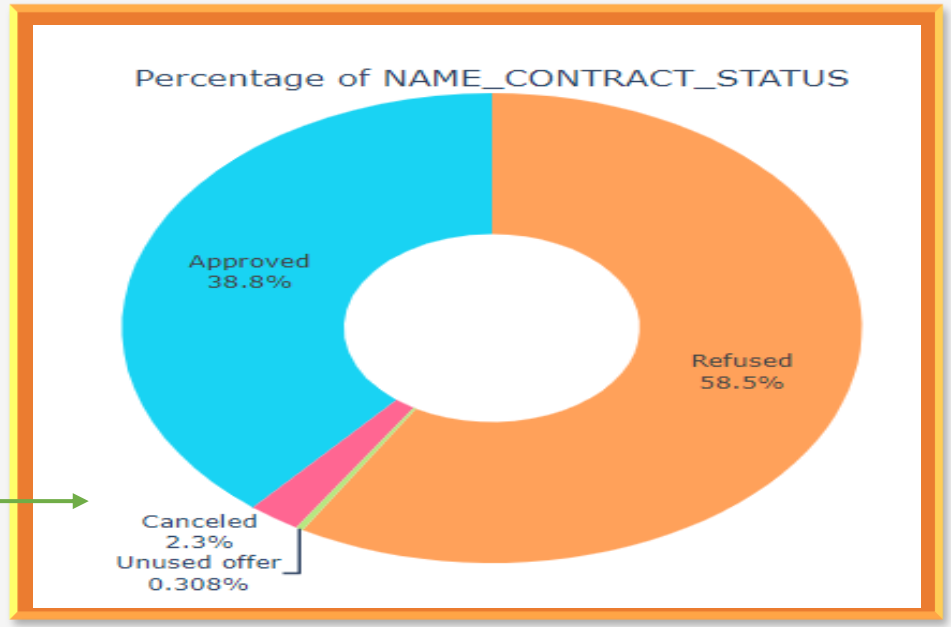
10.1 Percentage of `NAME_CONTRACT_STATUS` and `NAME_CLIENT_TYPE`



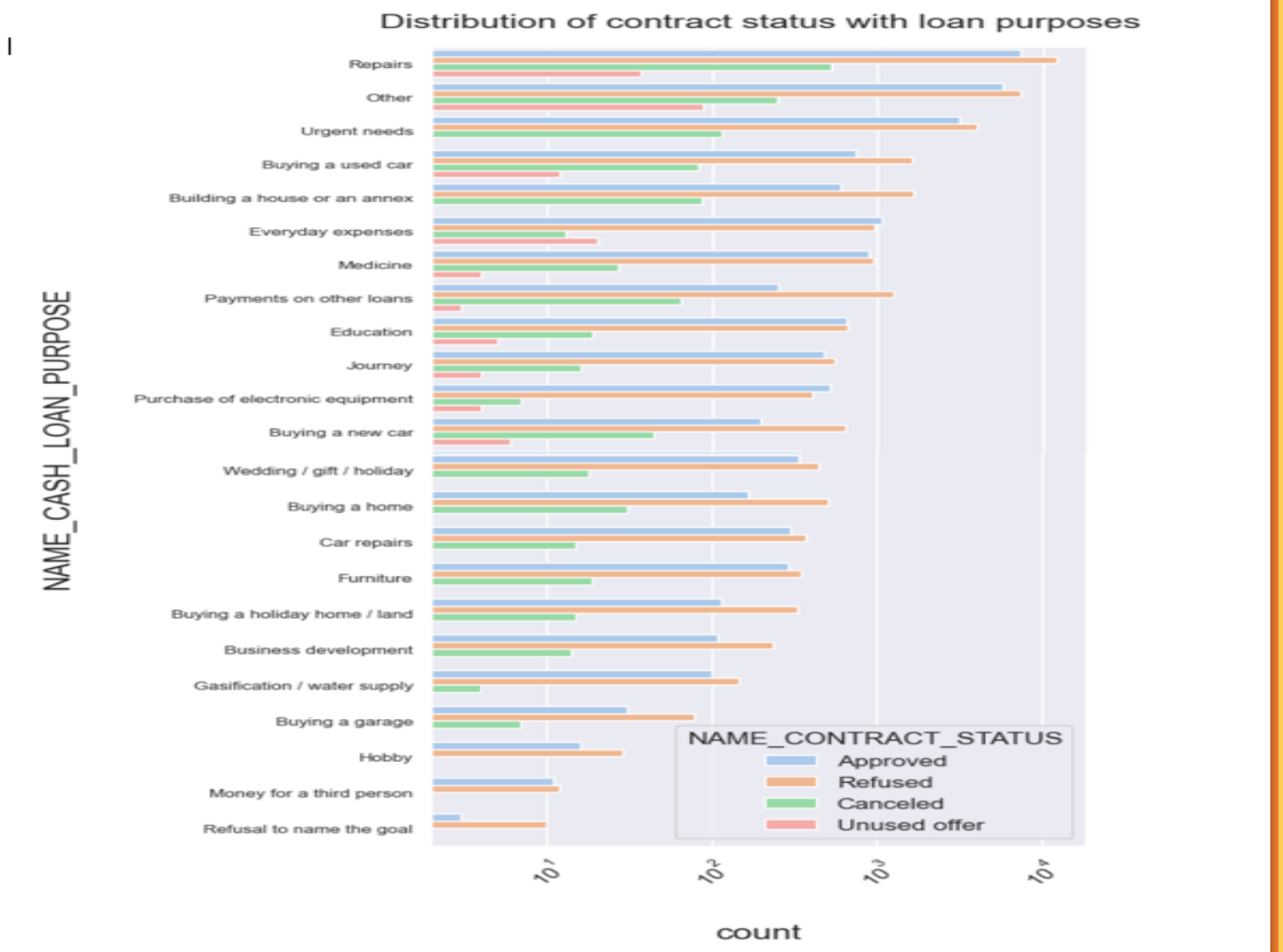
- Around 80.7% clients were repeaters applying for loan.
- 14.5% clients are new applying for the loan.

Percentage of contracts approved or not in previous applications

- Approved :- 38.8%
- Refused :- 58.5%
- Canceled :- 2.3%
- Unused offer :- 0.31%

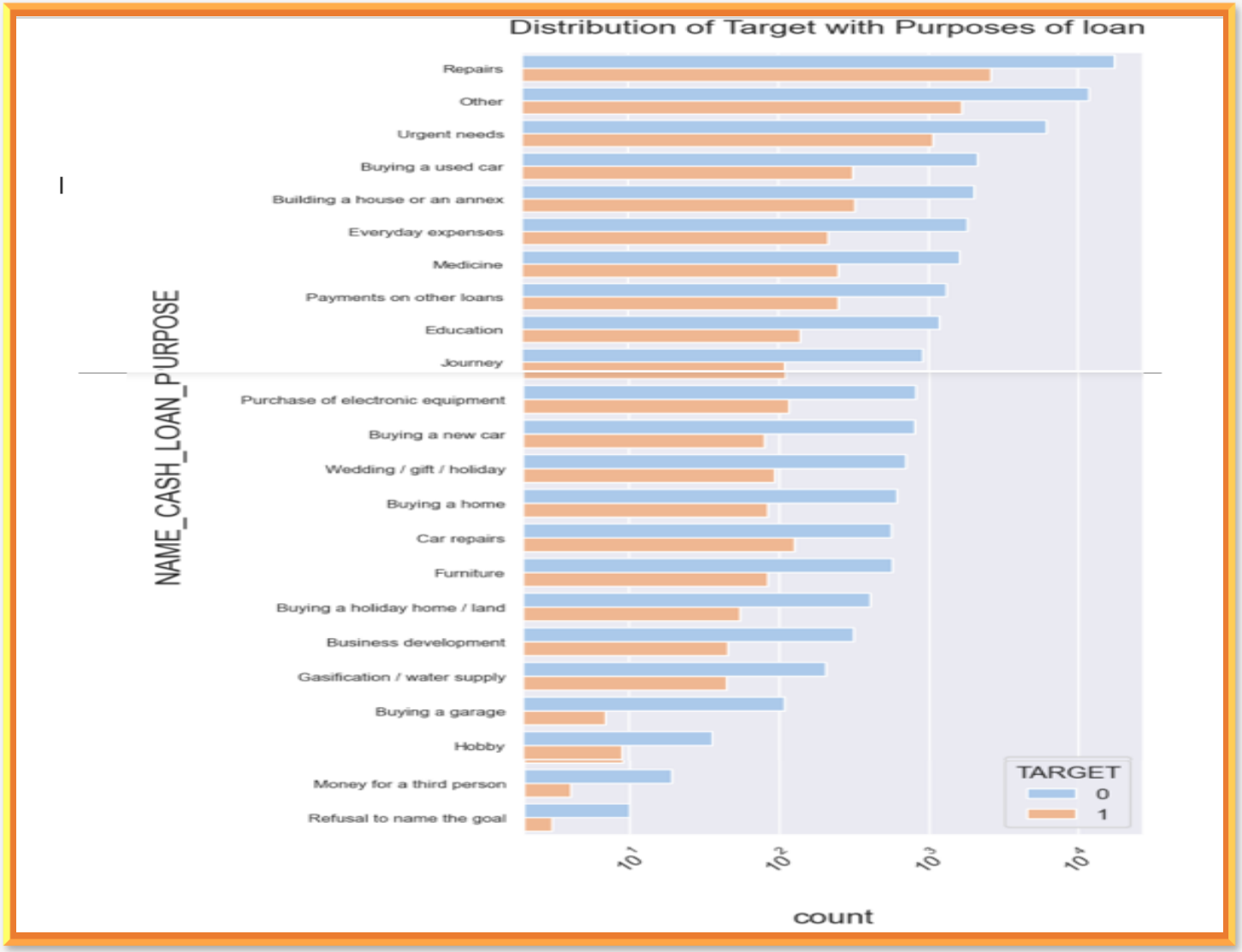


10.2 CONTRACT_STATUS V/s LOAN_PURPOSE



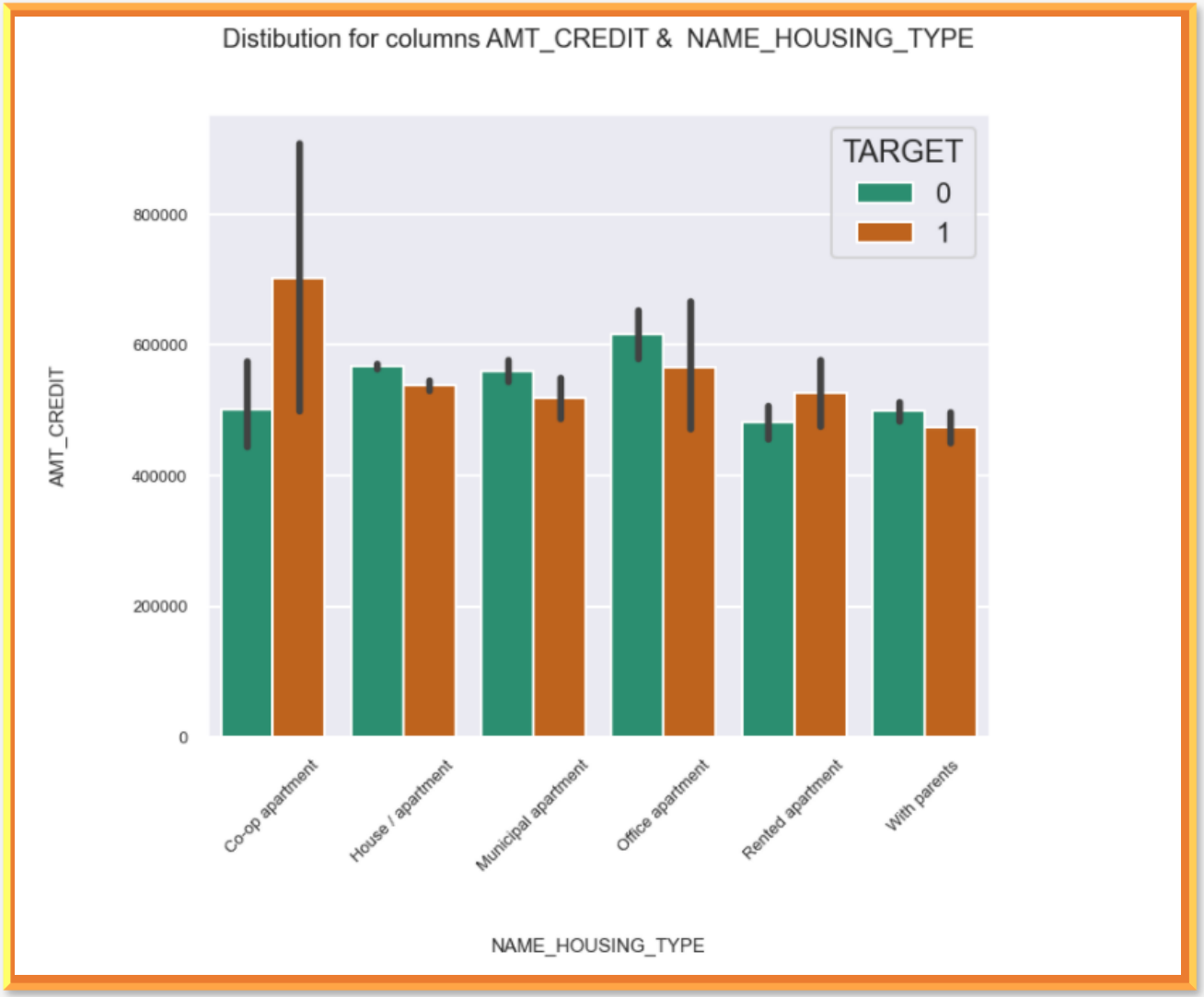
- Most rejection of loans came from purpose "Repairs".
- For "Education" & "Medicine" purposes we have equal number of approves and rejection
- “Paying other loans” and "Buying a new car`` is having significant higher rejection than approves.

10.3 TARGET V/s LOAN_PURPOSE



- Here also there is a high variation for "Repairs" for both the targets
- Comparing "Education" & "Medicine" purposes , Medicine there is high no. of clients having difficulties for re payment the loan amount compare to "Education"
- "Buying used Car" and "Building purpose" client having difficulties in payment have equal ratio

10.4 AMT_CREDIT vs NAME_HOUSING_TYPE



Here for Housing type

- **Office apartment is having higher credit of target 0 and co-op apartment is having higher credit of TARGET 1.**

Conclusion

So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House or municipal apartment for successful payments.

Conclusions

Final Insights

- We observe a decrease in the percentage of Payment Difficulties who are pensioners and an increase in the percentage of Payment Difficulties who are working when compared the percentages of both Payment Difficulties and non-Payment Difficulties.

- We observe a decrease in the percentage of married and widowed with Loan Payment Difficulties and an increase in the 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

- We observe an increase in percentage of Loan Payment Difficulties whose educational qualifications are secondary/secondary special and a decrease in the percentage of Loan Payment Difficulties who have completed higher education when compared with the percentages of Loan Payment Difficulties and Loan Non-Payment Difficulties

Final Insights

- The count of ‘Low skilled Laborers’ in ‘OCCUPATION_TYPE’ is comparatively very less and it also has maximum % of payment difficulties- around 17%. Hence, client with occupation type as ‘Low skilled Laborers’ are the driving factors for Loan Defaulters.

- The count of ‘Lower Secondary’ in ‘NAME_EDUCATION_TYPE’ is comparatively very less and it also has maximum % of payment difficulties- around 11%. Hence, client with education type as ‘Lower Secondary’ are the driving factors for Loan Defaulters.

- Banks should focus more on contract type Student ,pensioner and Businessman with housing type other than Co-op apartment, Office apartment for successful payments.

- Banks should focus less on income type Working as they are having the greatest number of unsuccessful payments.

Final Insights

- Applicants living in House/Apartments has the highest number of loan application. While we see that Rented apartment and applicants living with parents have very high percentage to default
- Get as much as clients from housing type With parents as they are having least number of unsuccessful payments
- Also, with loan purpose Repair is having higher number of unsuccessful payments on time.

THANK YOU