CREDIT EDA ASSIGNMENT

Divya N L August 28th, 2022

CONTENTS

01. Problem statement

02. Steps involved in Analysis

03. Conclusions & Recommendations

Problem Statement

- This case study aims to identify patterns if an loan applicant has difficulty in repaying his/her loan in a Finance Based Company.

- Also to ensure that applicants capable of repaying are not denied the application request and applicants who are incapable of repaying are approved/accepted the application request.

- Bank wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

- The bank can utilize this knowledge for its *portfolio and risk assessment*.

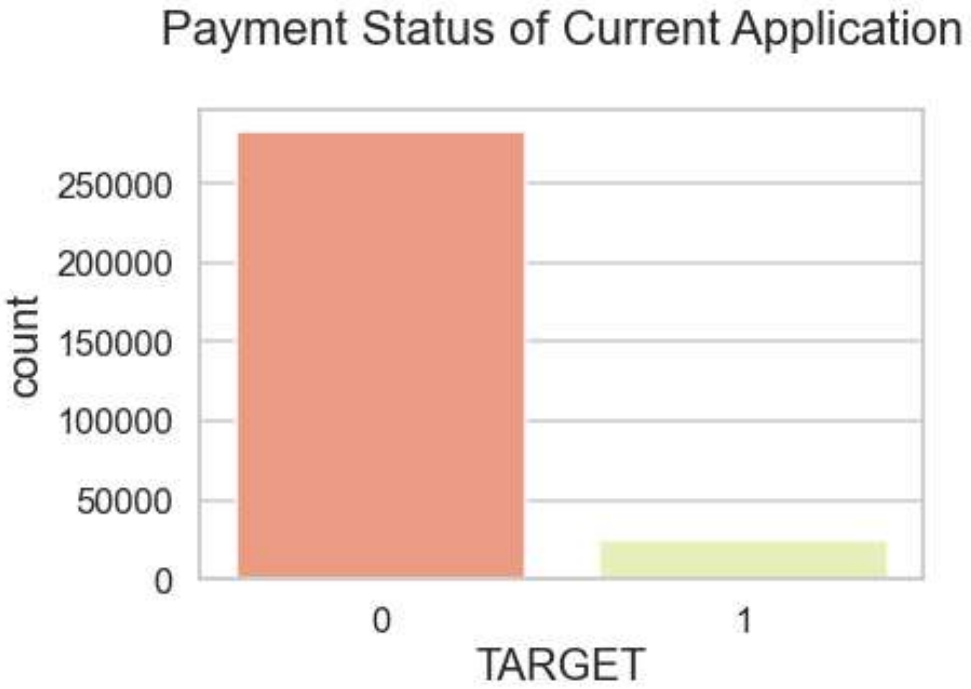
Steps involved in analysis

- Import Warnings and Libraries
- Read the Datasets
- Analyze, Clean the Dataset
  - Eliminating null values and Identifying Outliers
- Check the Imbalance Ratio
- Univariate, Segmented and Bivariate Analysis
- Summary and Recommendations

IMBALANCE RATIO

OBSERVATIONS :

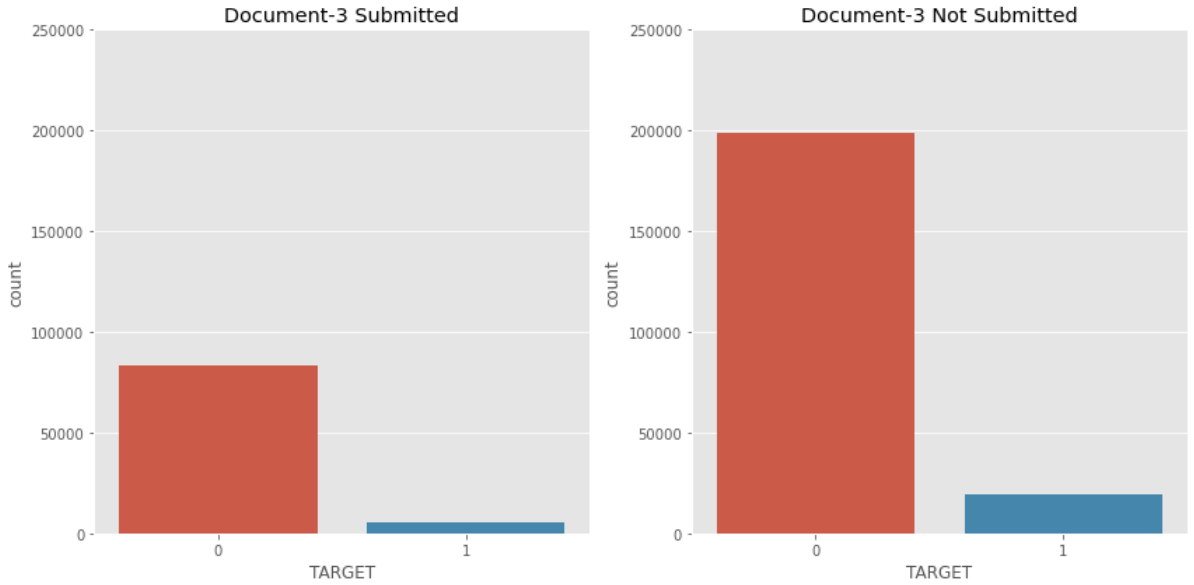
- The data is highly imbalanced.
- Data Imbalance Ratio 8:92
- Number of defaulters(Target = 1): 24825
- Number of non-defaulters (Target = 0): 282686



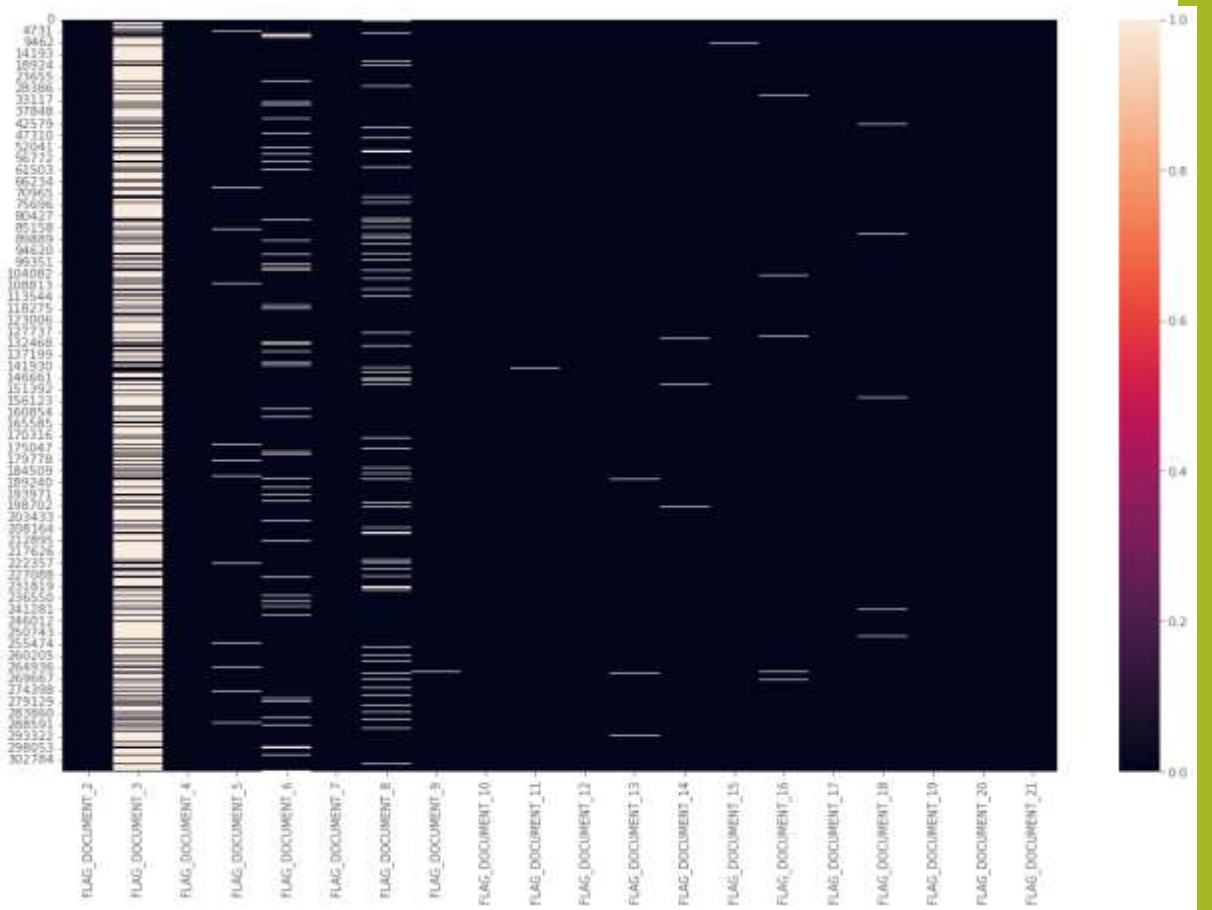
Segment wise analysis

Documents submitted by applicants:

- except Document_3 , other documents were not provided by applicants in majority of the cases.
- Document_3 shows similar trends for both defaulters and non-defaulters.



Heat Map

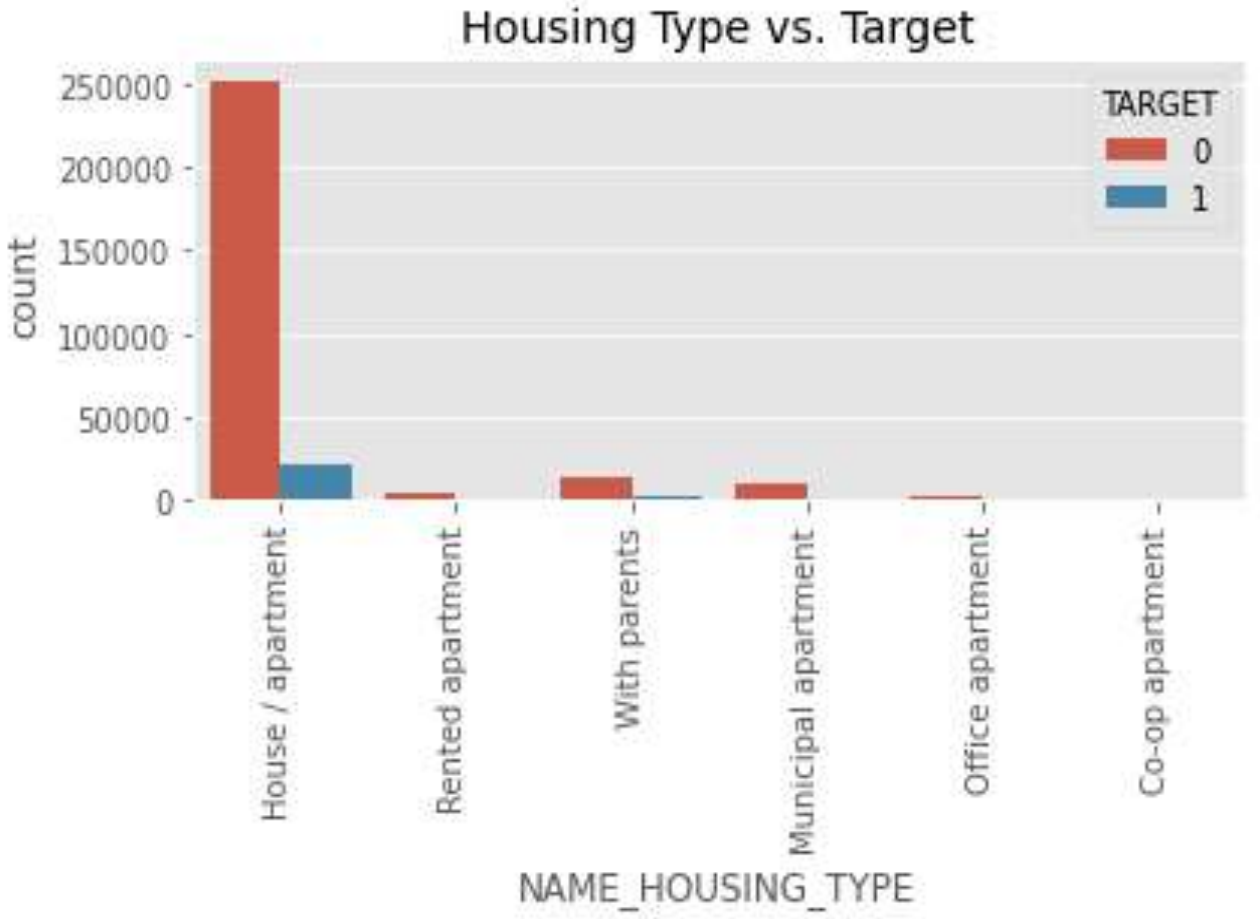


Housing Type

- <u>OBSERVATIONS:</u>

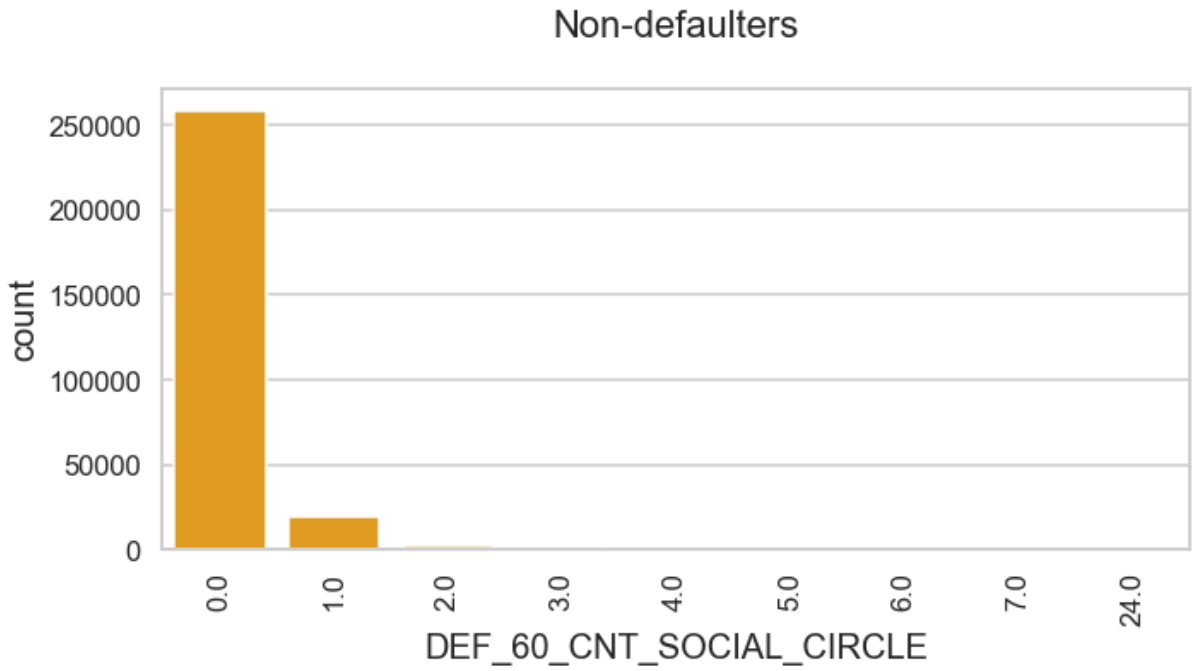
- Most of the loan applicants live in House/Apartment

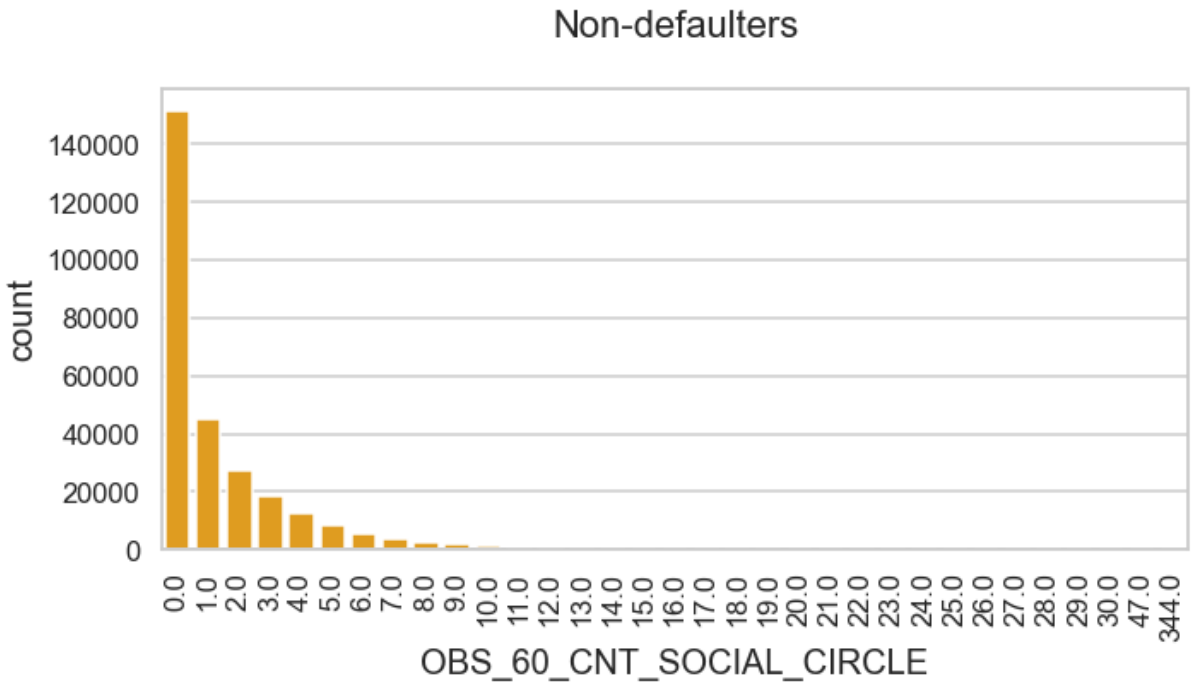
- The most defaulters live in rented apartment or live with their parents



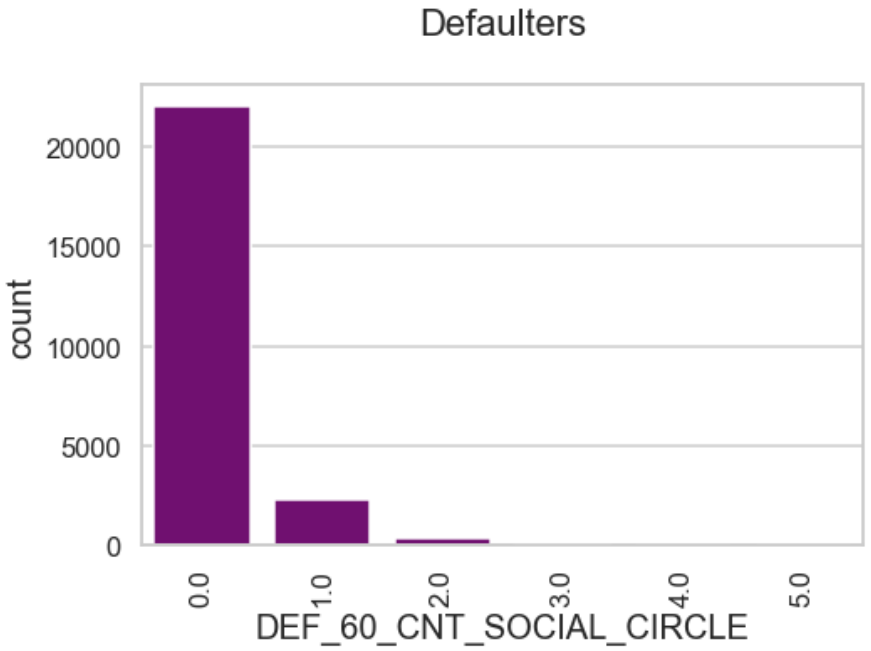
SOCIAL CIRCLE

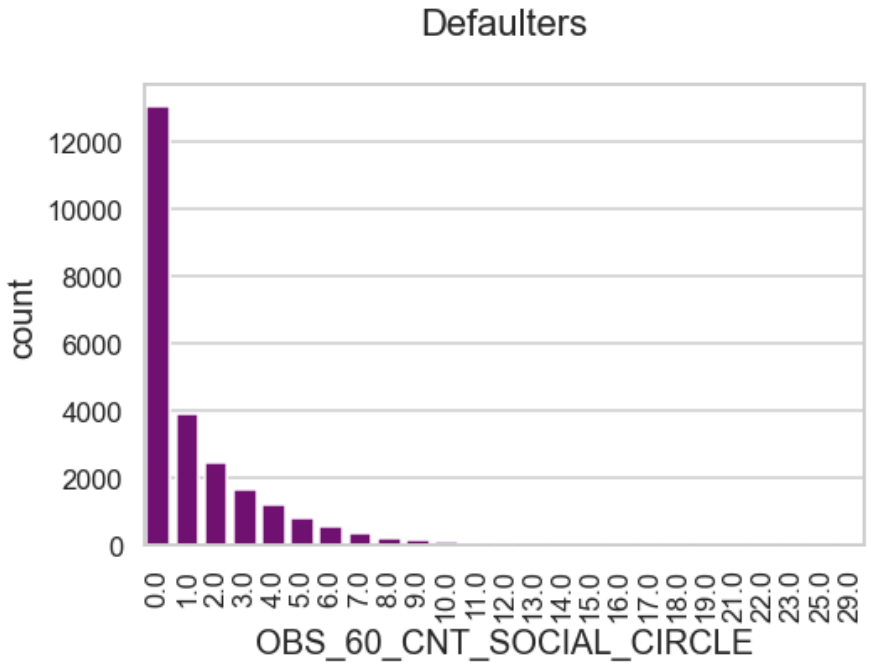
- For defaulters and non-defaulters 'DEF_60_CNT_SOCIAL_CIRCLE','OBS_60_CNT_SOCIAL_CIRCLE' features show similar trend.





- DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE are highly correlated
- OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE are identical column

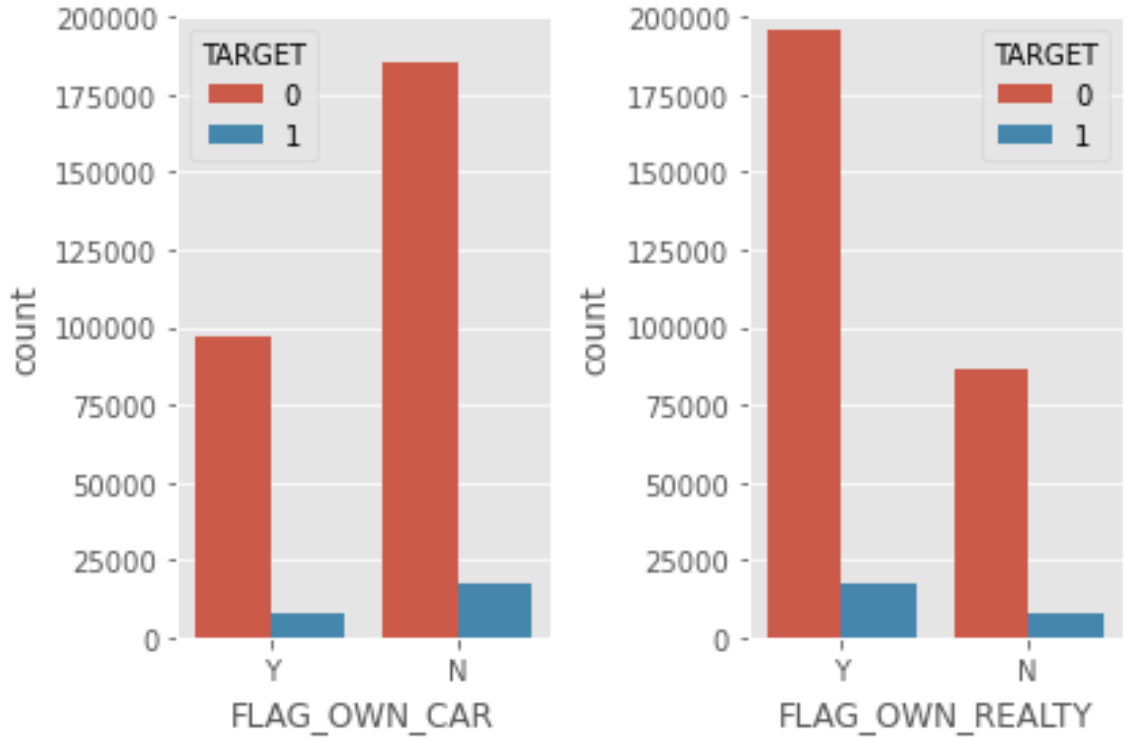




ASSET DETAILS

OBSERVATIONS:

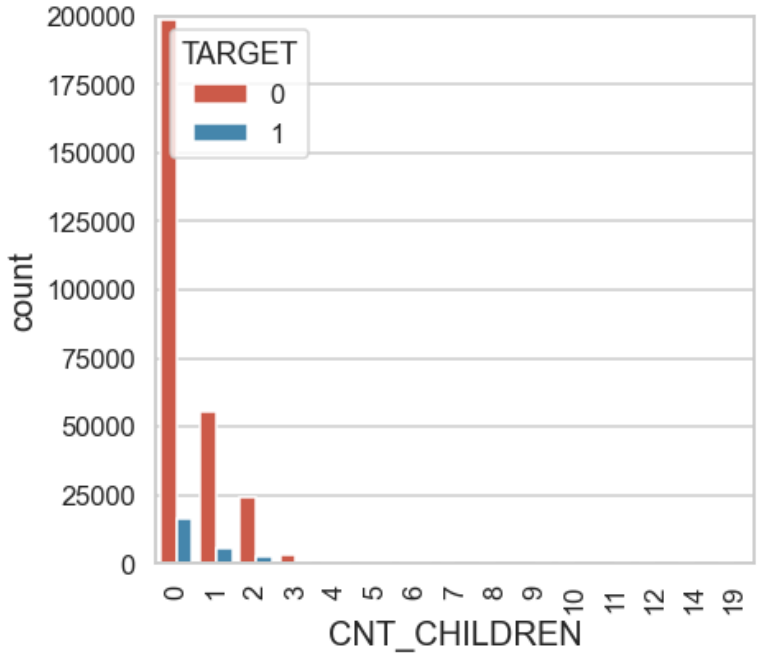
- People not owning realty and car and have a slightly higher default rate than the people who own reality and car.

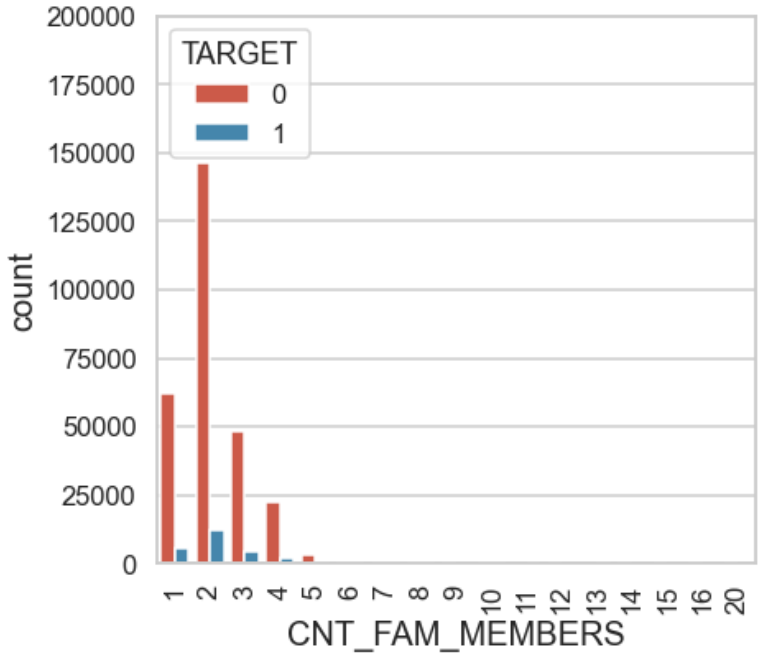


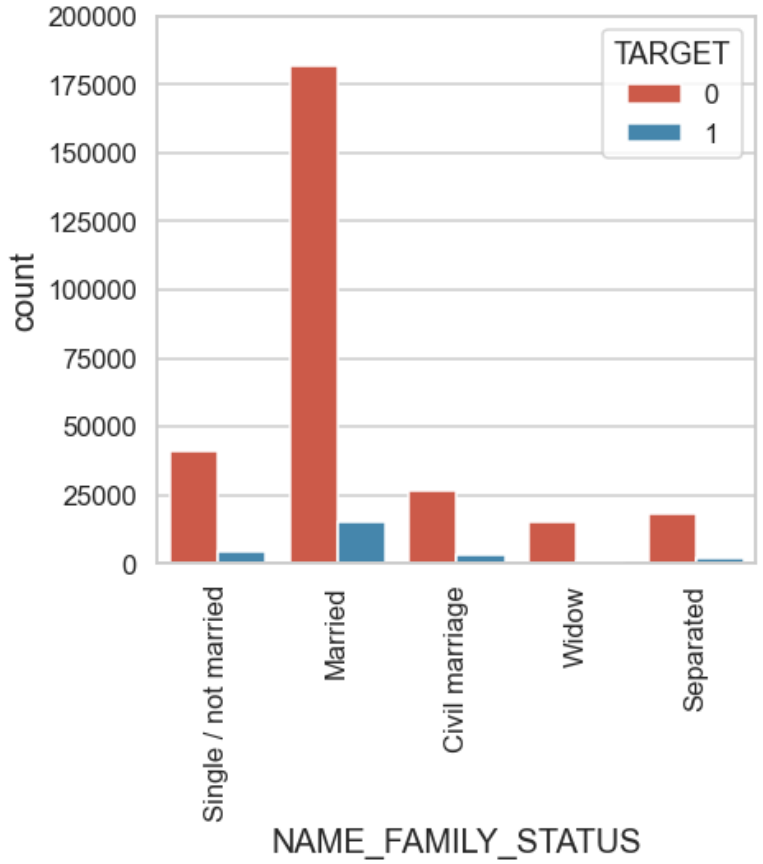
FAMILY RELATED INFO

OBSERVATIONS:

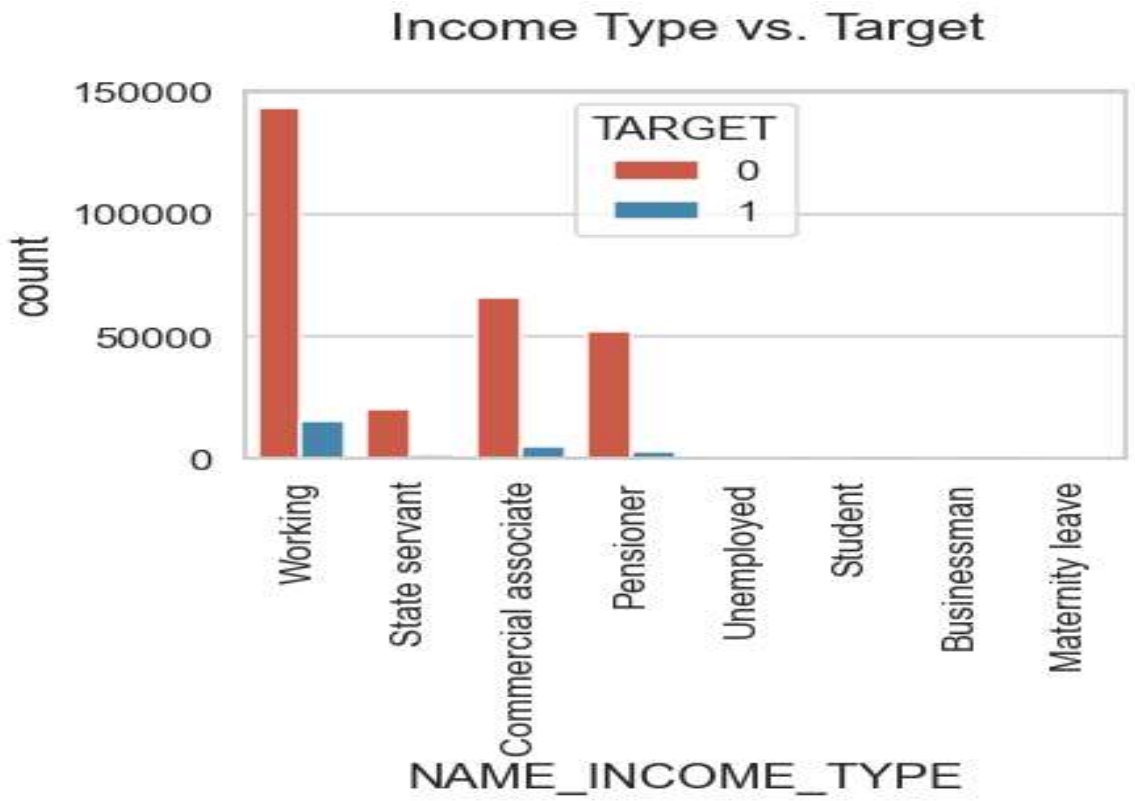
- Default rate is highest for Civil Marriage and Single applicants
- Most of the applicants are married (and/or) no children (and/or) 2 family members.
- Applicants with relatively more number of children (and/or) family members have higher default percentage.
- For some of the cases where count children/family members is high, and the default rate is very high or very low. These cases cannot be taken as a conclusion as number of applicants having a large family is very low.


TARGET
0
1
count
200000
175000
150000
125000
100000
75000
50000
25000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 14 19
CNT_CHILDREN


TARGET
0
1
count
200000
175000
150000
125000
100000
75000
50000
25000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 20
CNT_FAM_MEMBERS


TARGET
0
1
count
200000
175000
150000
125000
100000
75000
50000
25000
0
Single / not married
Married
Civil marriage
Widow
Separated
NAME_FAMILY_STATUS

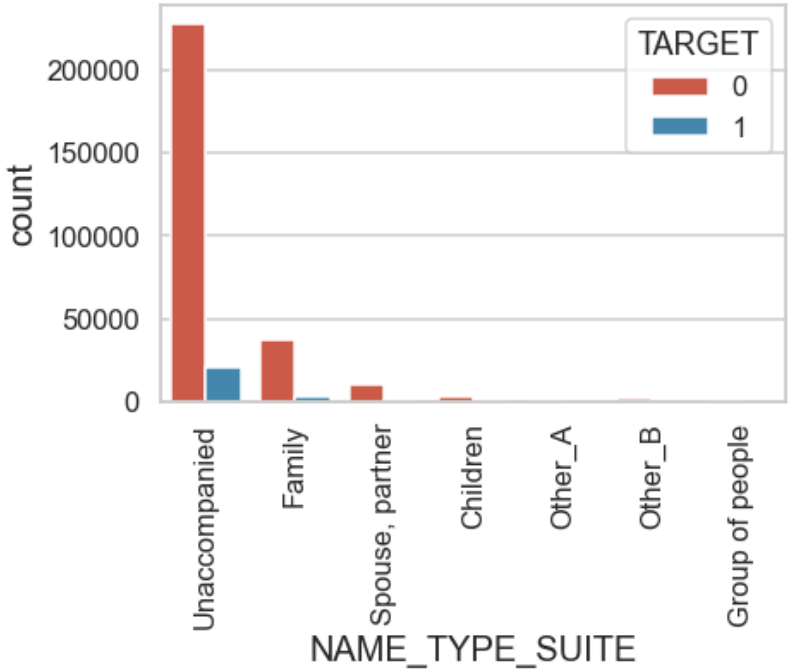
Income Type and gender

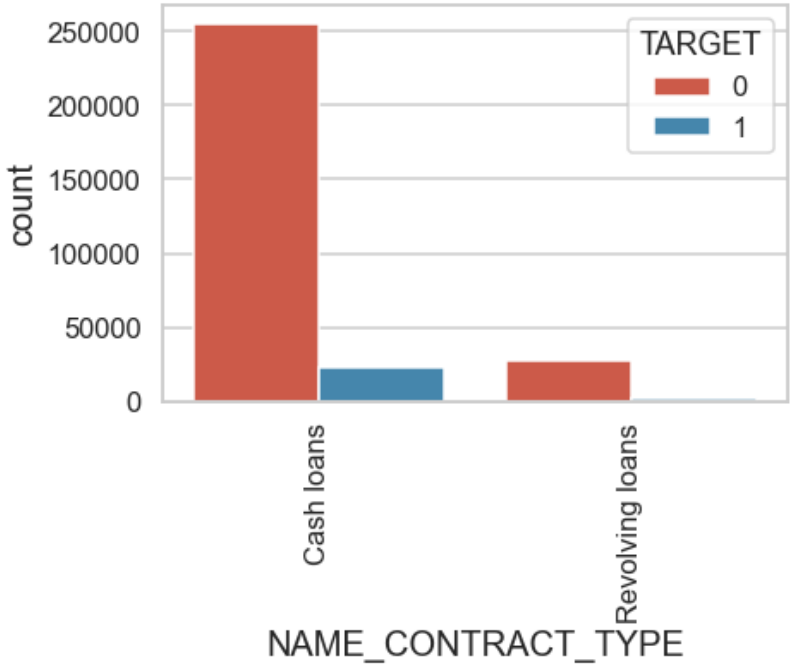


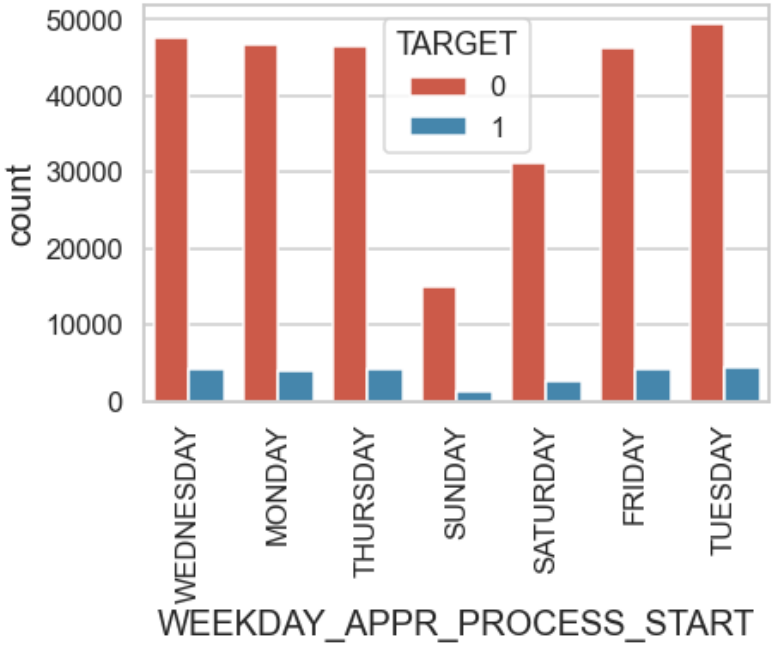
OBSERVATIONS

- Female applicants are more than male applicants
- Defaulter percentage is higher for male applicants
- Applicants having "Lower secondary" education have highest percentage of Defaulter. Applicants having 'Higher education' and ' academic degree' have lowest percentage of defaulters
- Applicants on Maternity Leave and Unemployed has highest percentage of Defaulter
- Businessman have lowest (0) percentage of Defaulter However, if we see the count, the applicants of income type('Unemployed', 'Student', 'Businessman', 'Maternity leave') are very few in the dataset to contribute in the analysis. Most of the applicants are working.

- Most of the applicants are unaccompanied while applying for loan
- Number Cash loans is quite higher than Revolving Loans
- All weekdays have similar number of applicants than weekend(Saturday and Sunday)

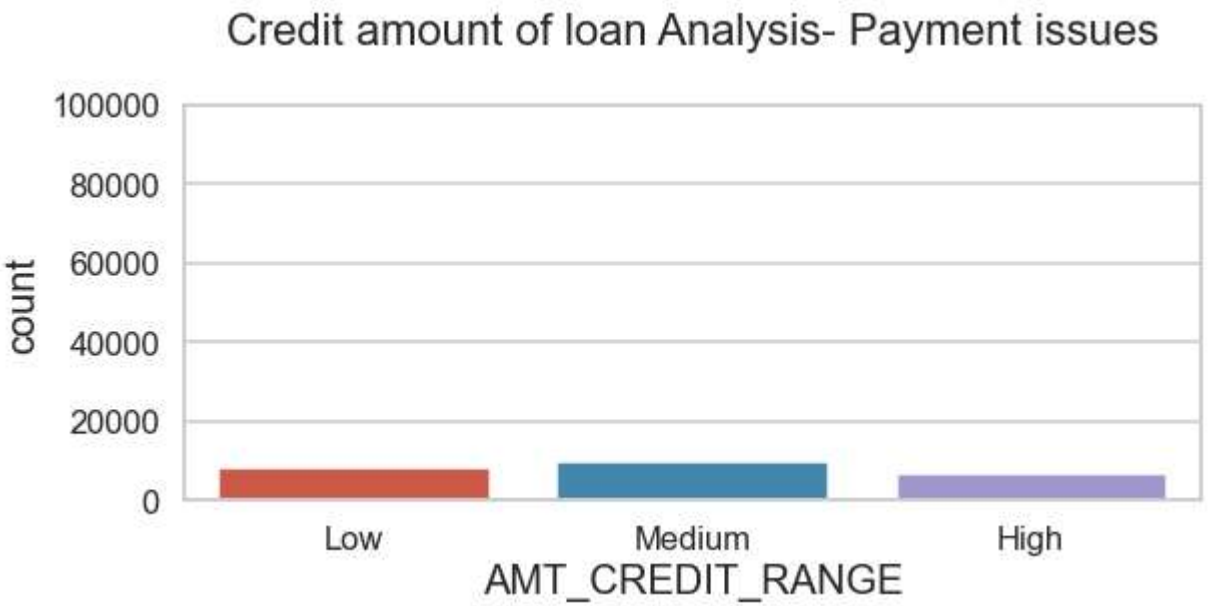


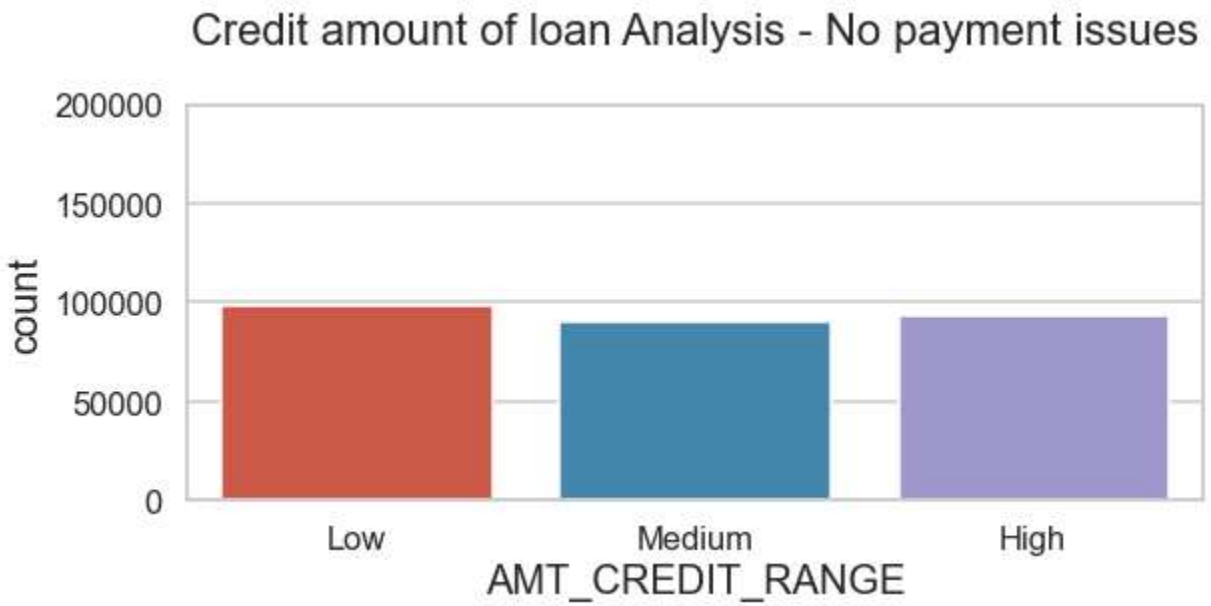




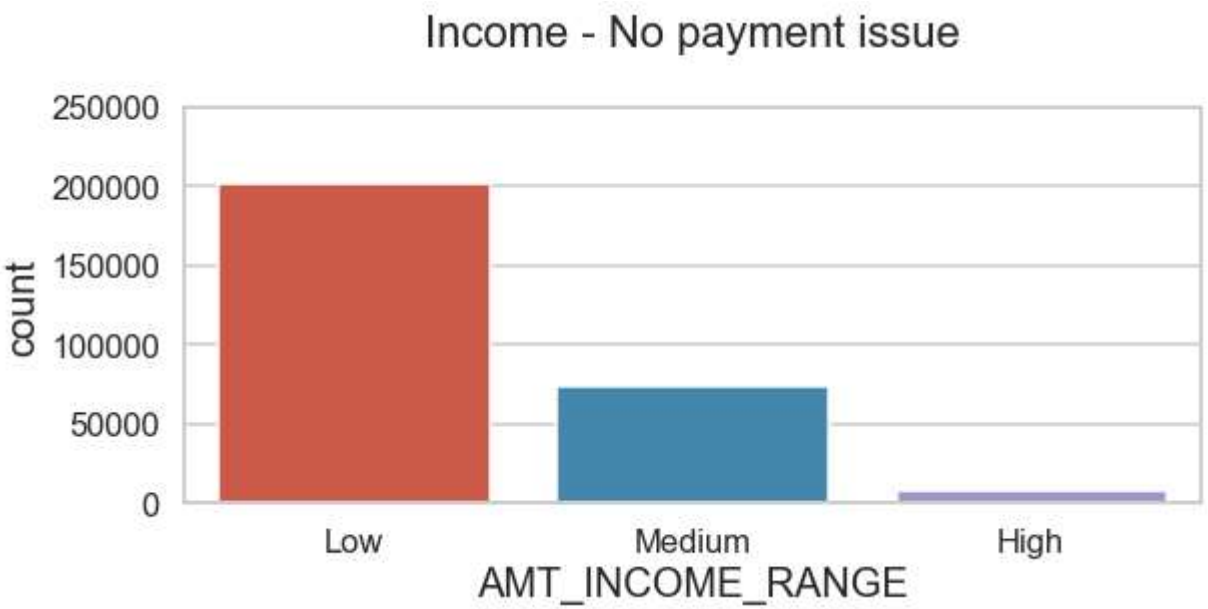
Credit Amount Analysis:

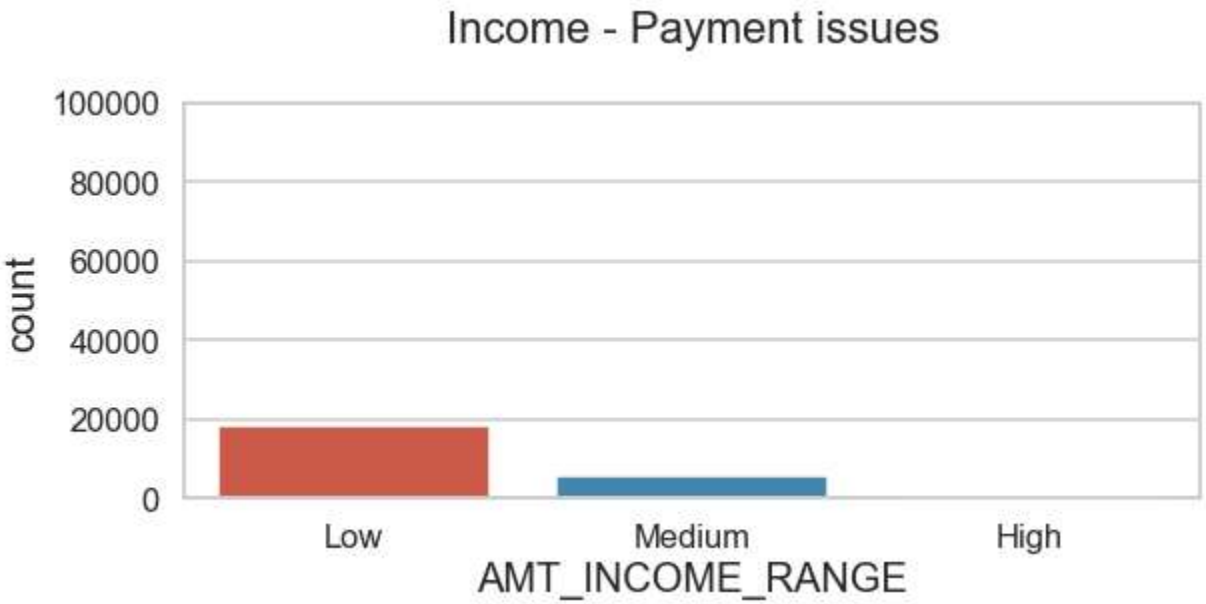
Customers who have low credit amount are more likely to pay back the loan

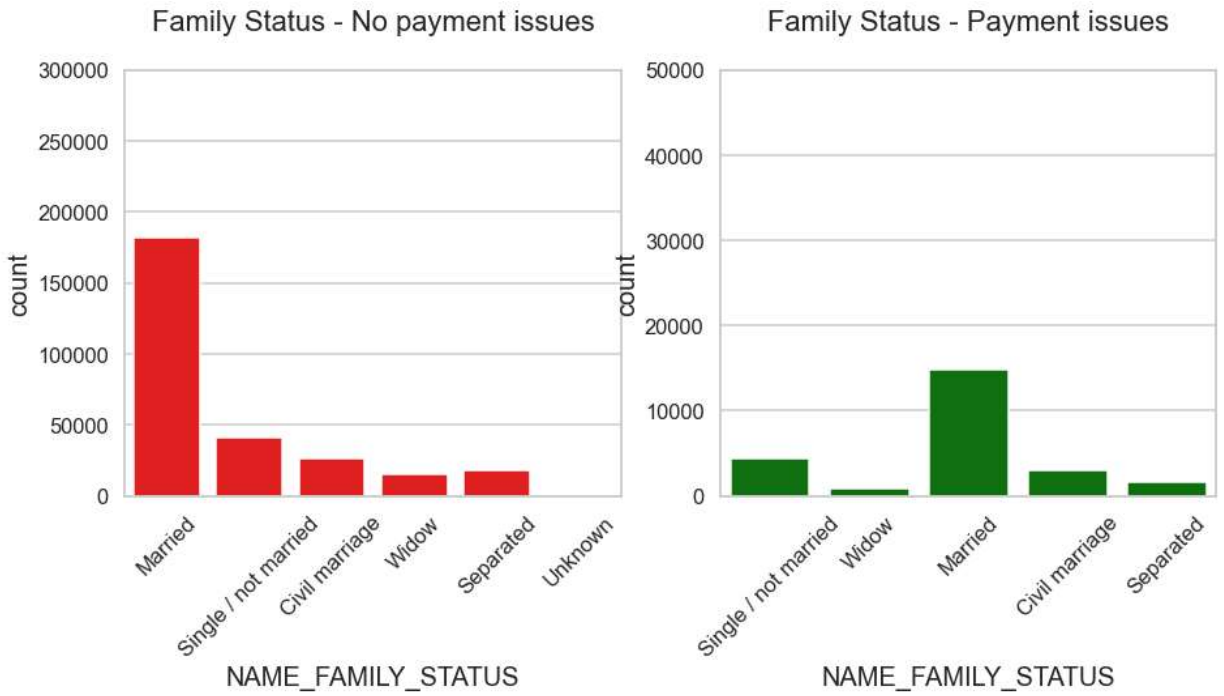


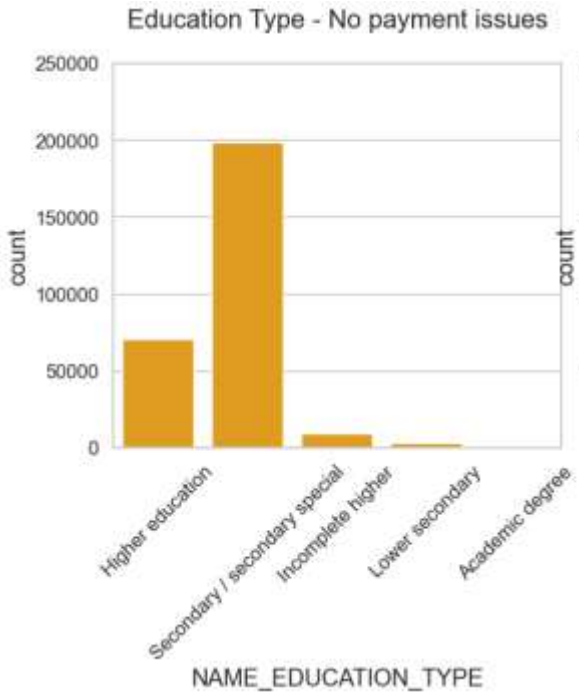


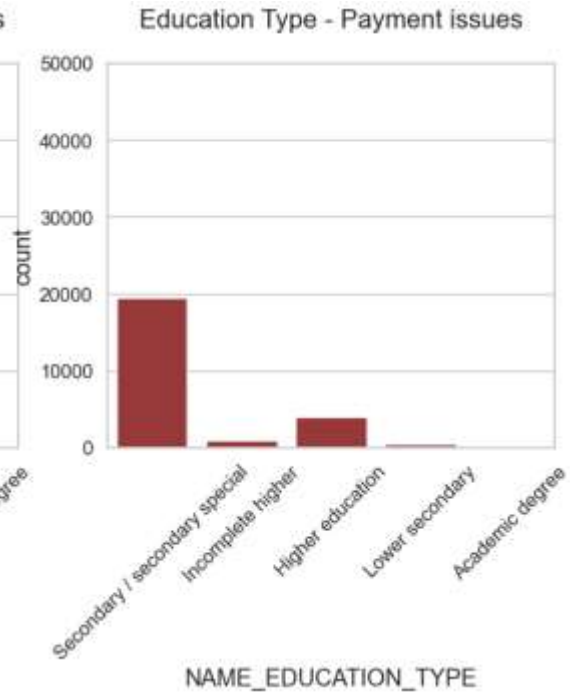
- As compared to the other categories, clients having low income are more likely to repay the loan.







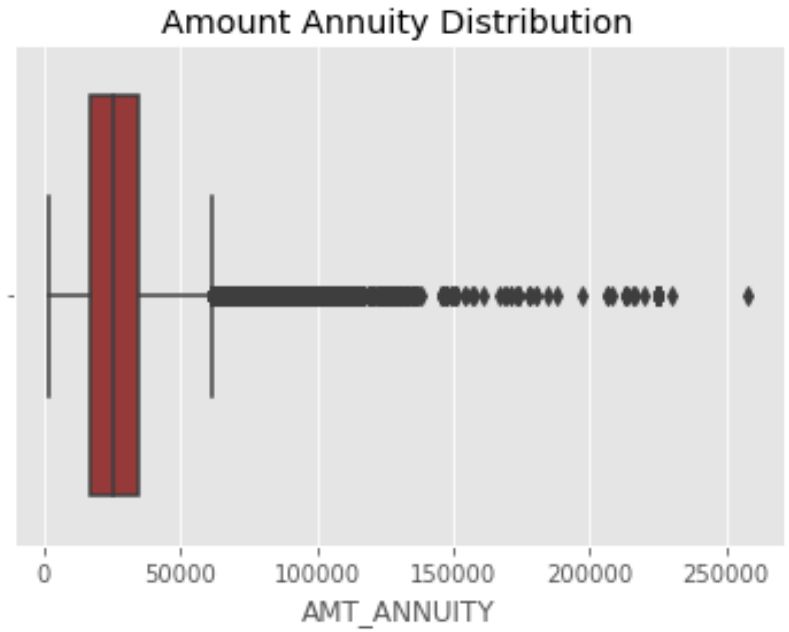


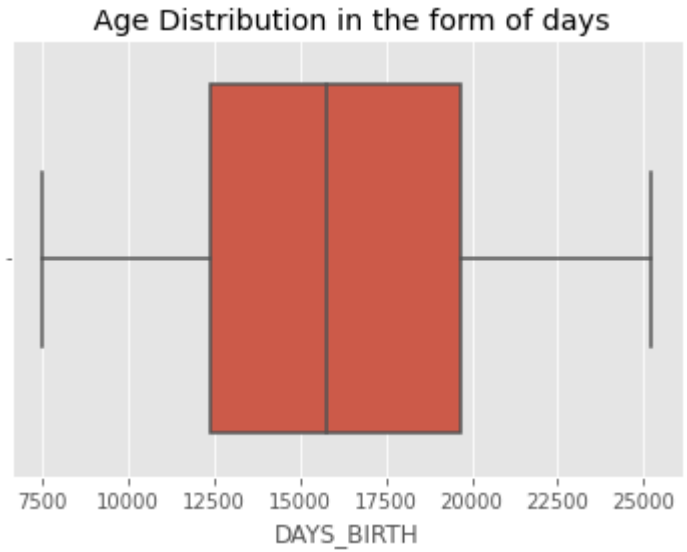


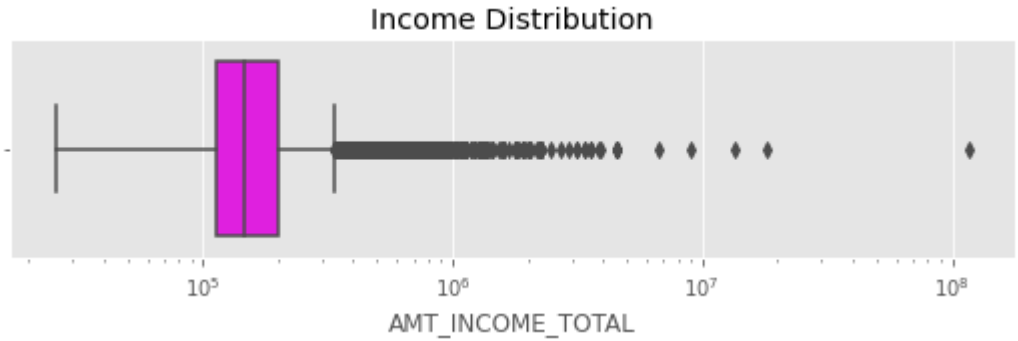
- People most likely to repay the loan have secondary/Secondary special education status whereas clients having an academic degree are the most defaulters.

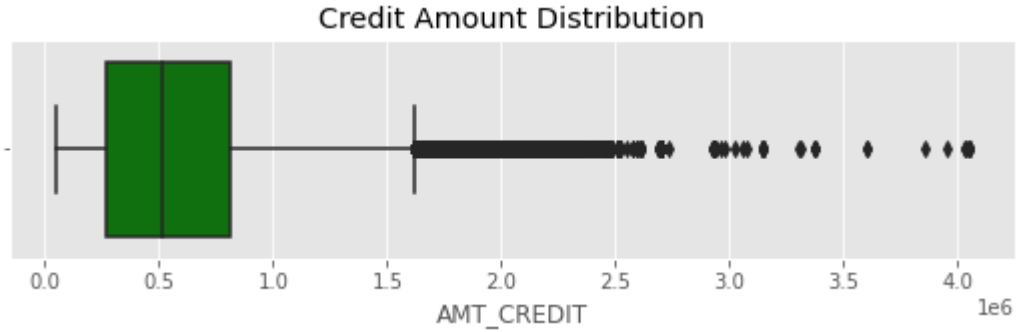
- Widows are the least likely category to repay whereas as seen here, Married category is most likely to repay the loan amount.

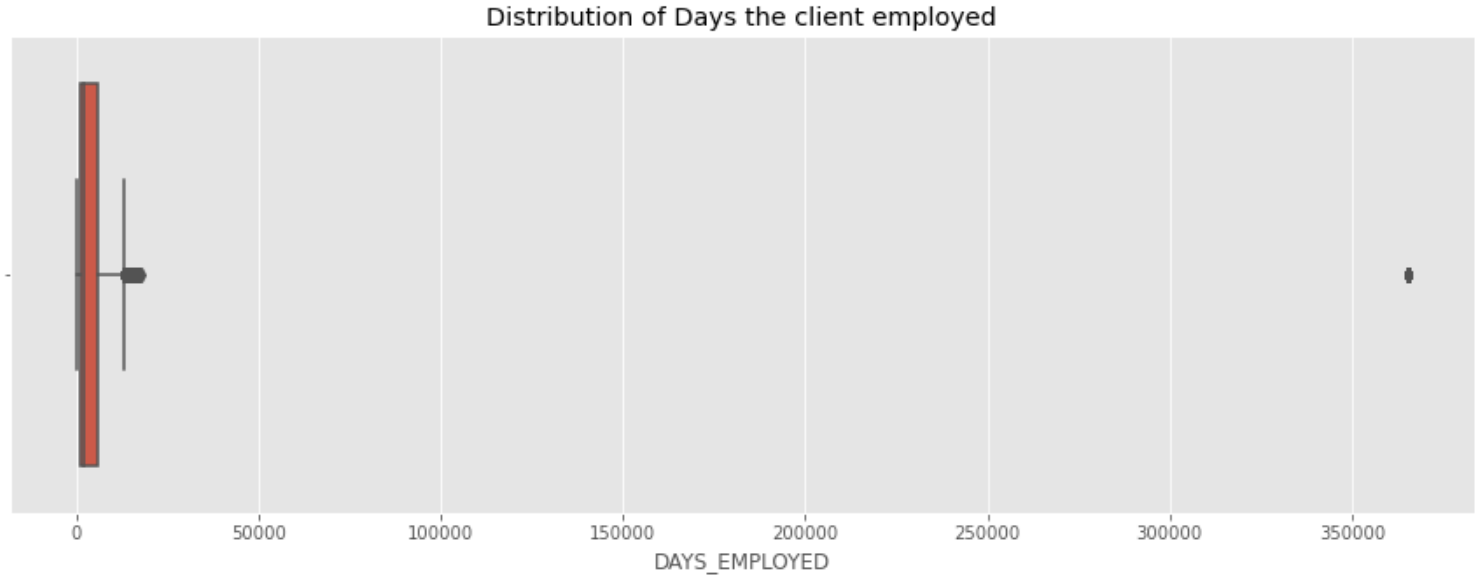
Identifying Outliers







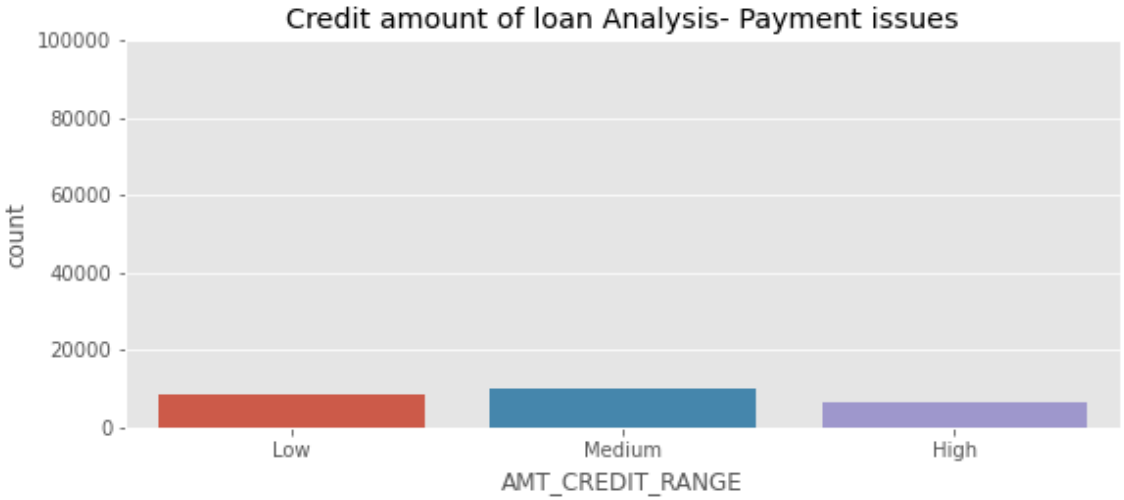




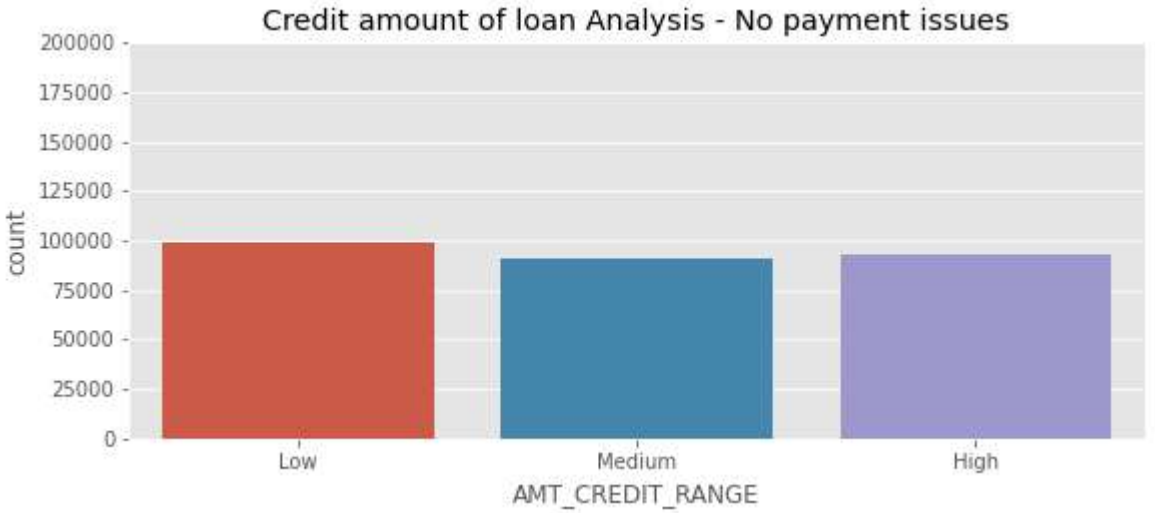
OBSERVATIONS:

- Amount_Annuity : outlier is present at 258025
- Here in 'AMT_INCOME_TOTAL' variable outlier values stands at 1.17x10^8. As the 95th and 99th quantile values differ largely, we can conclude the presence of outliers in the data set.
- Amt_credit : The outliers here are present after the 99th quantile.
- No outliers are present in Days_birth.
- we have an outlier at 365243 in Days_employed.

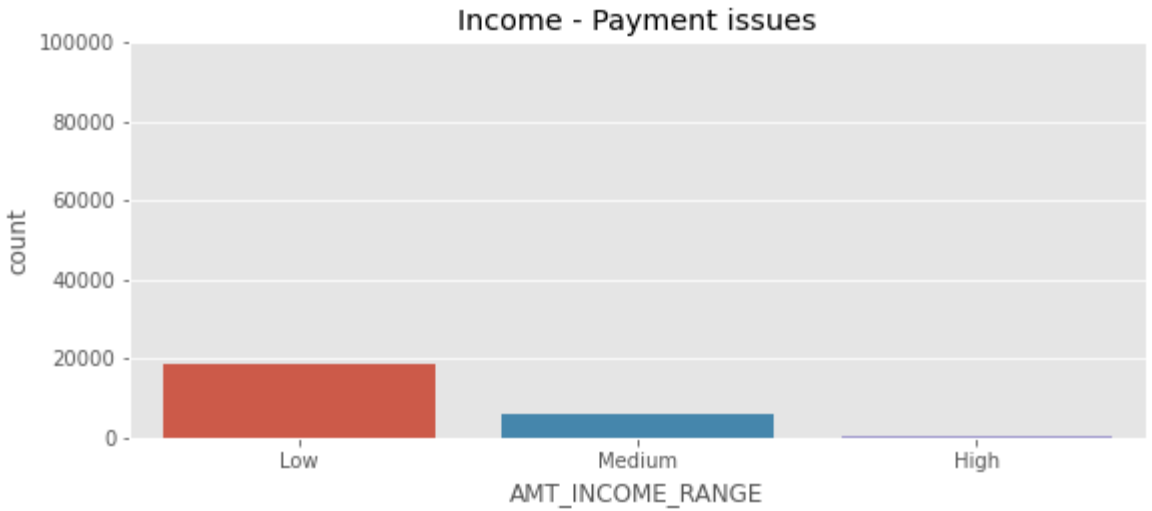
Univariate Analysis - Numeric variables



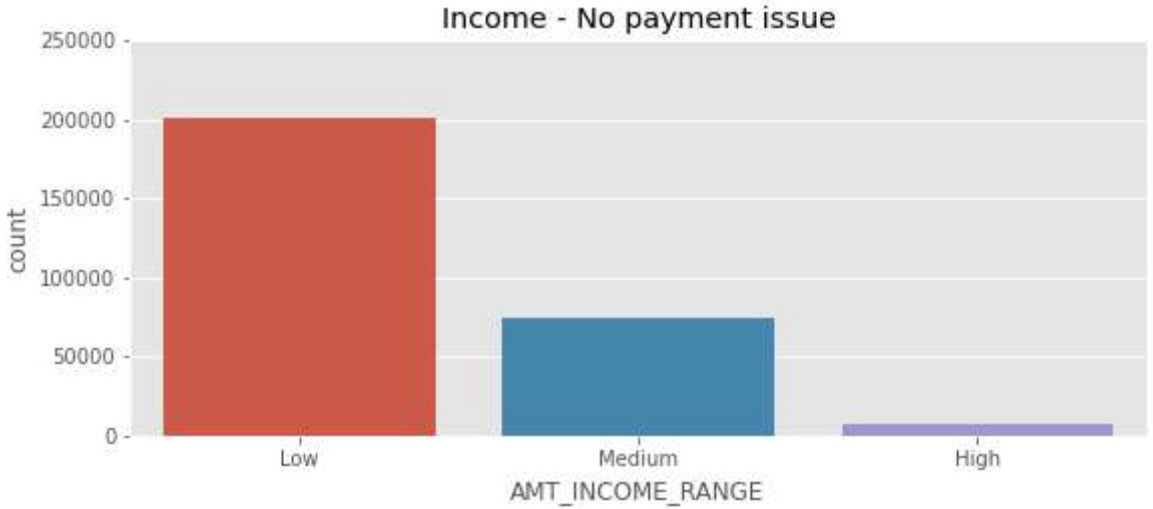
Defaulters



Non Defaulters

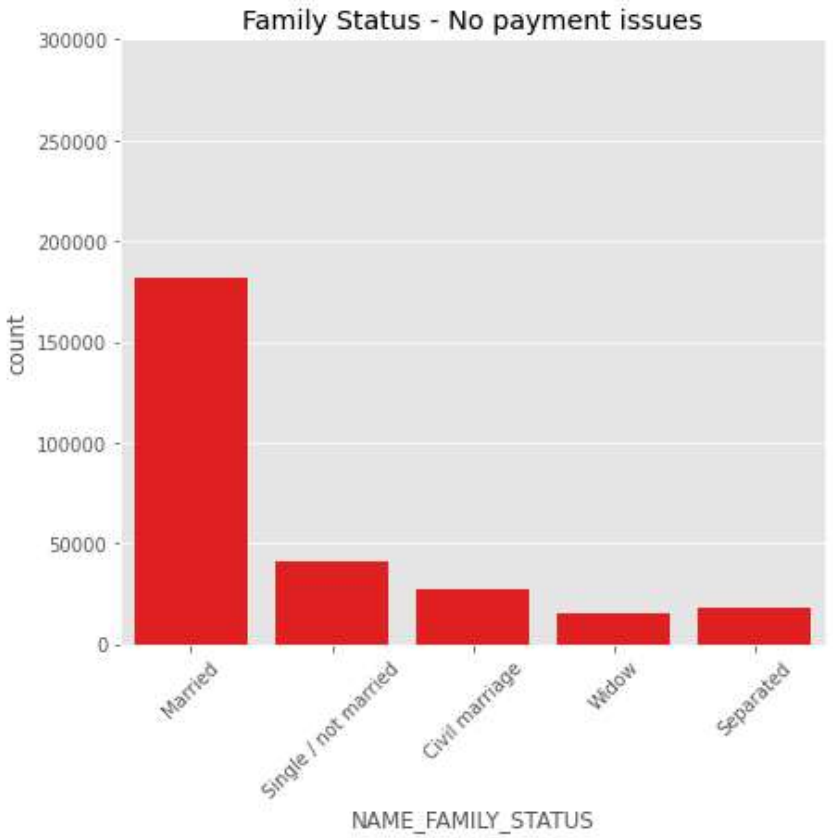


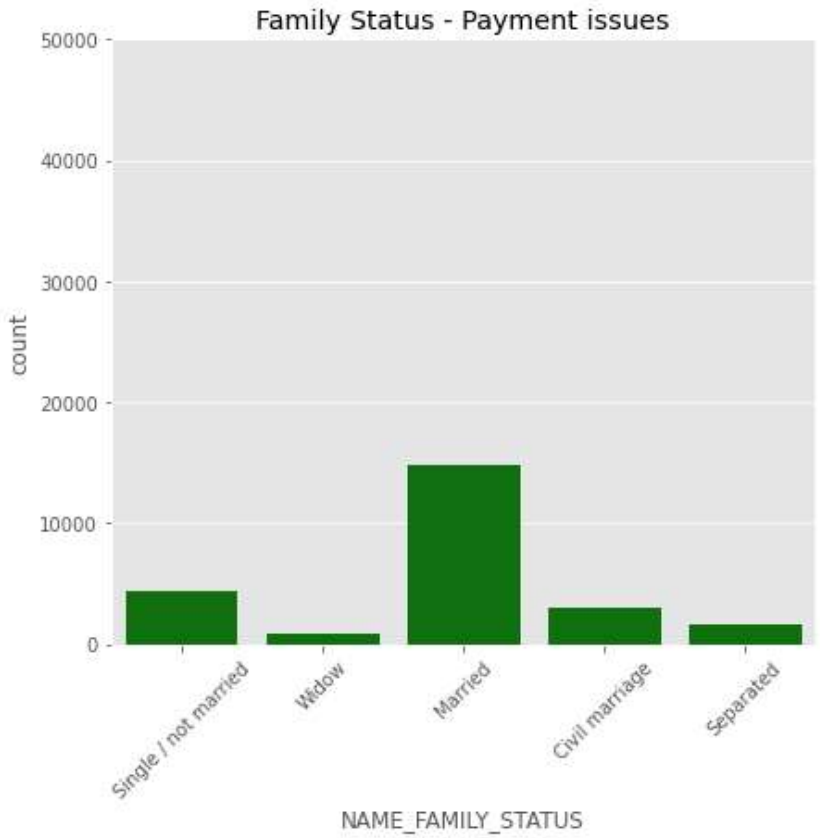
Defaulters

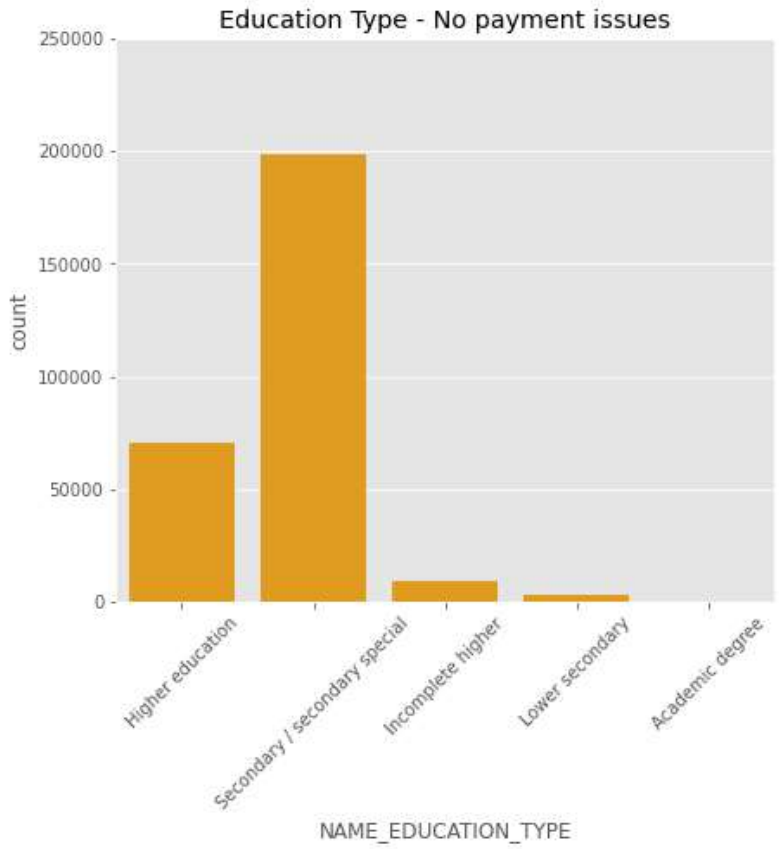


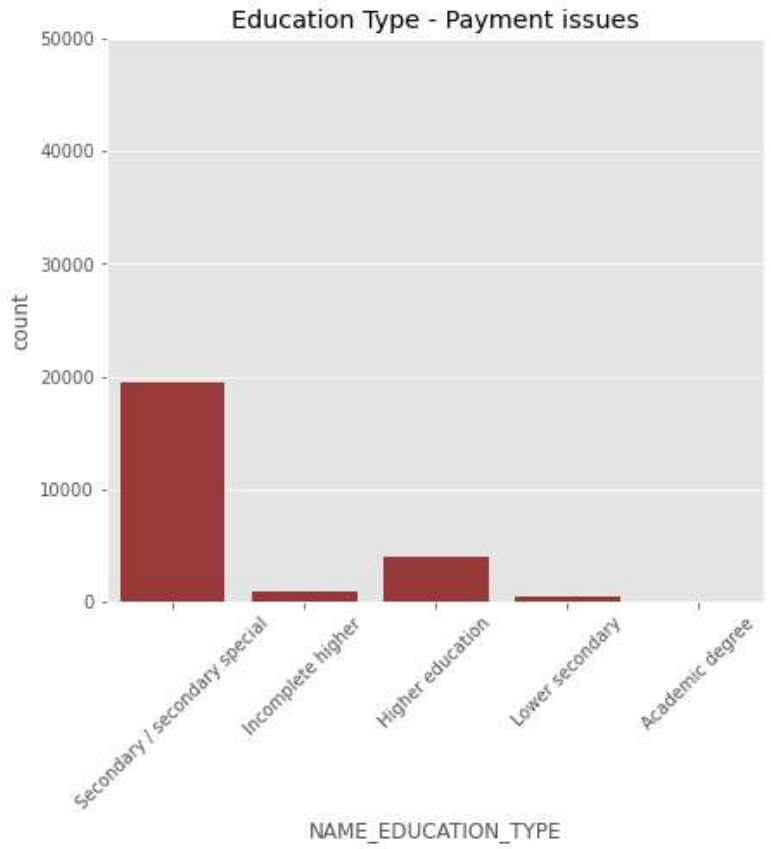
Non Defaulters

Univariate analysis - Categorical variables





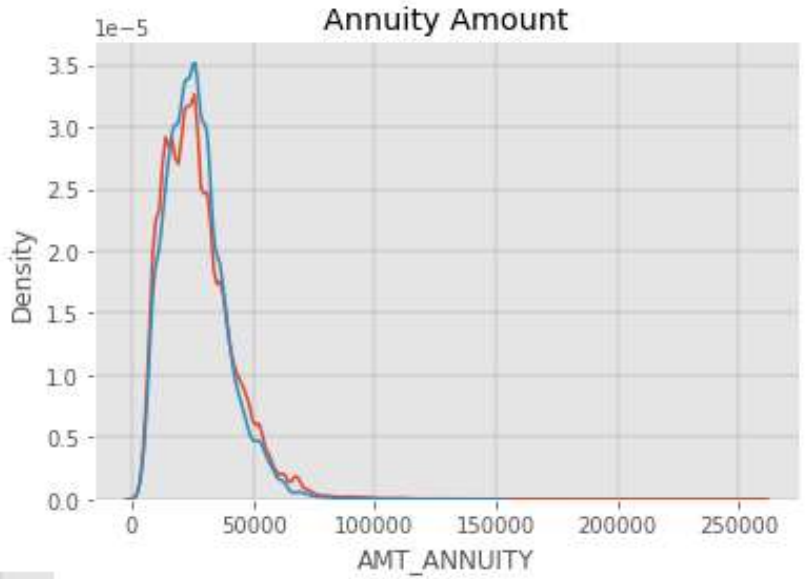


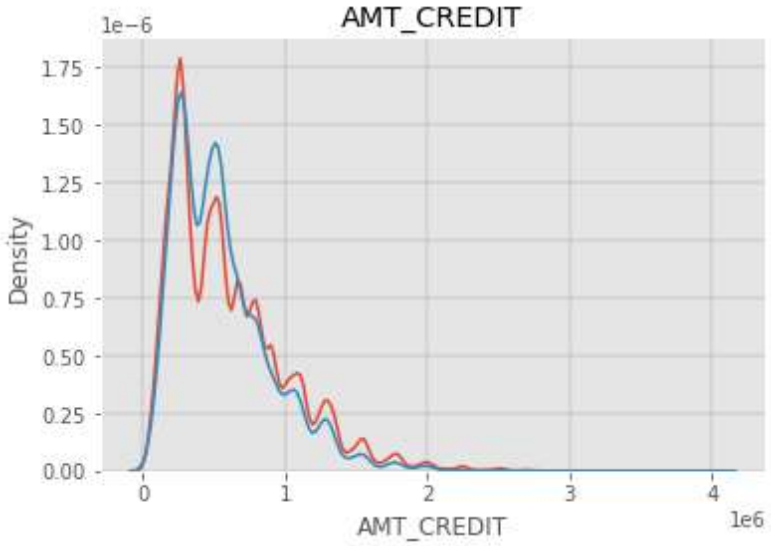


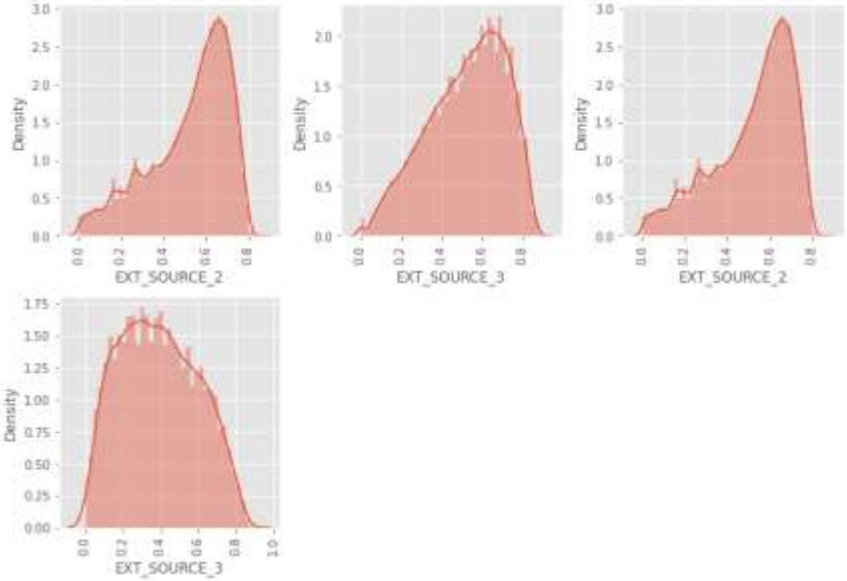
- People most likely to repay te loan have secondary/Secondary special education status whereas clients having an academic degree are the most defaulters.

- Widows are the least likely category to repay whereas as seen here, Married category is most likely to repay the loan amount.

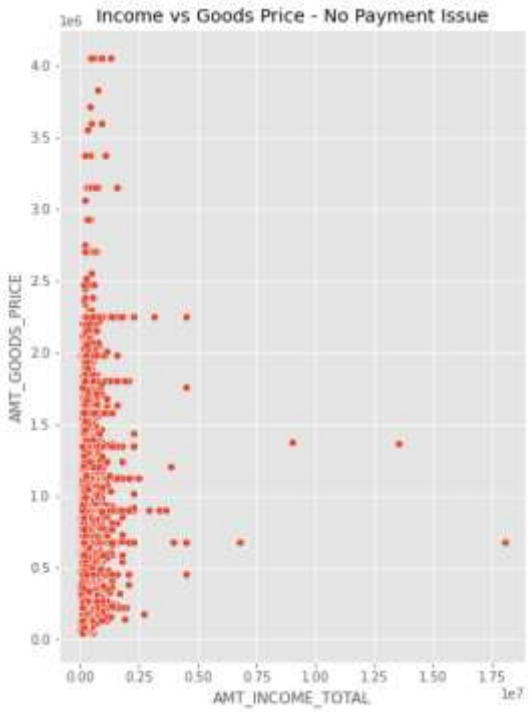
Univariate Analysis - Continuous variables

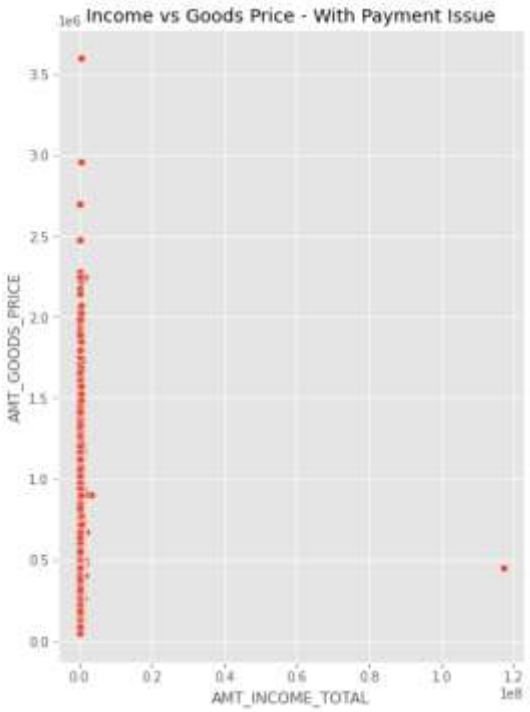


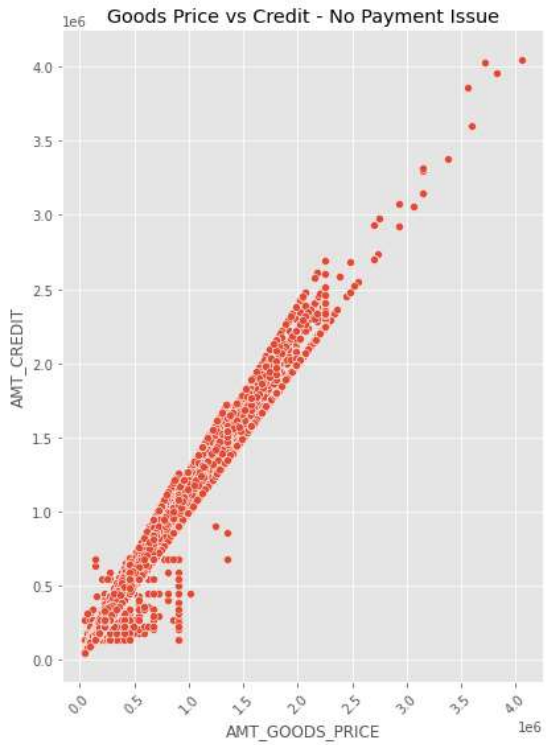


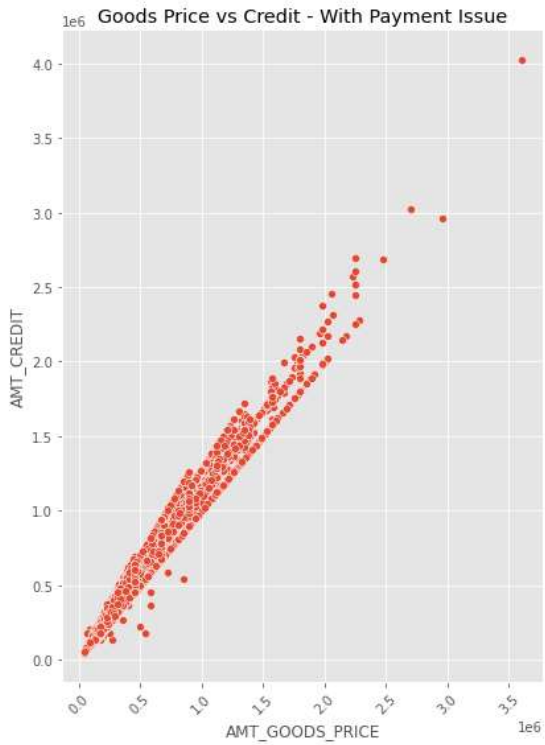


Bivariate Analysis

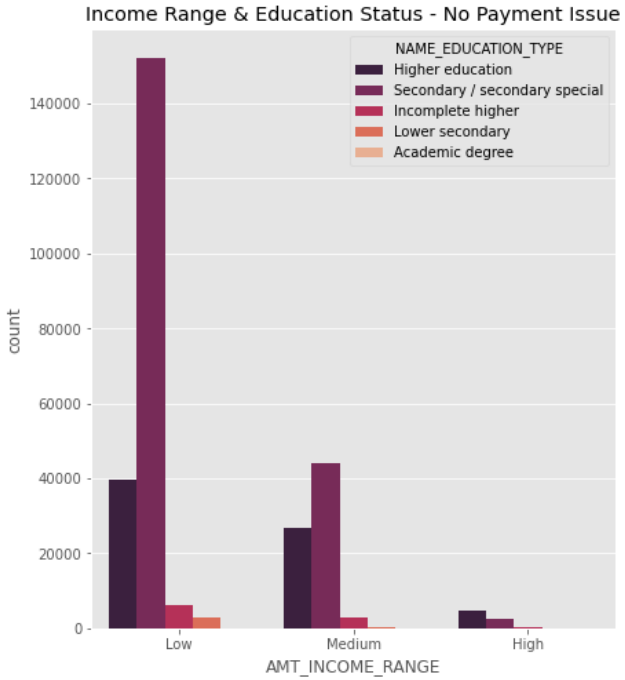


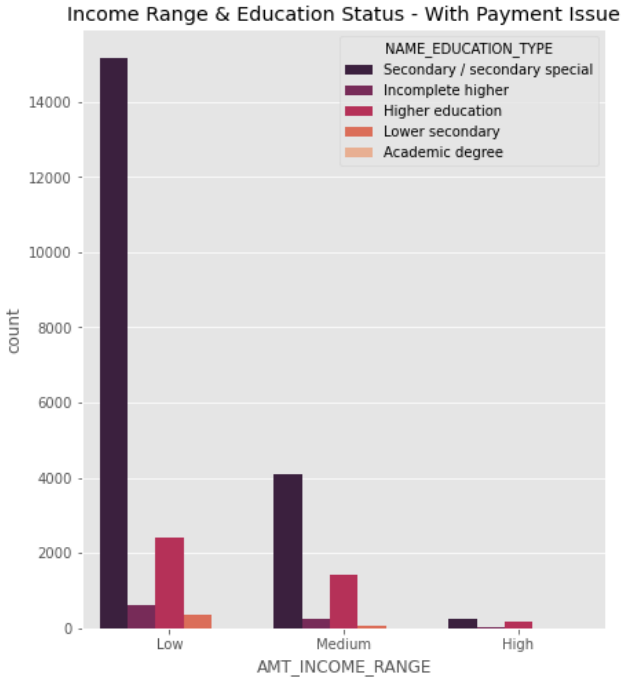


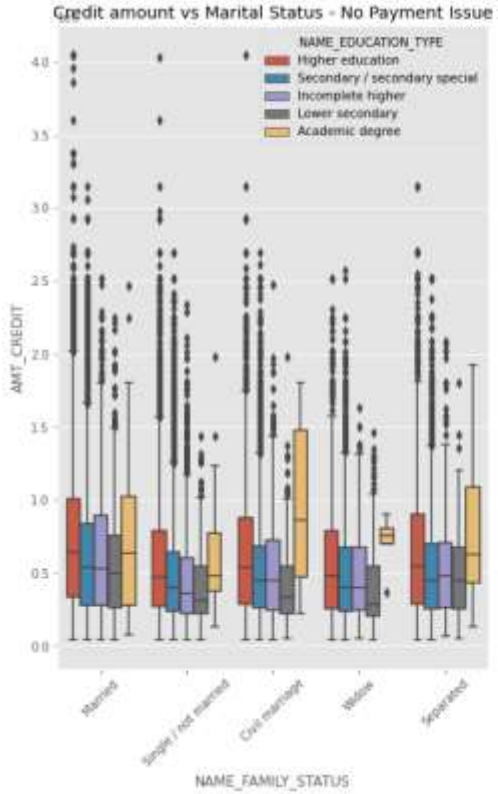


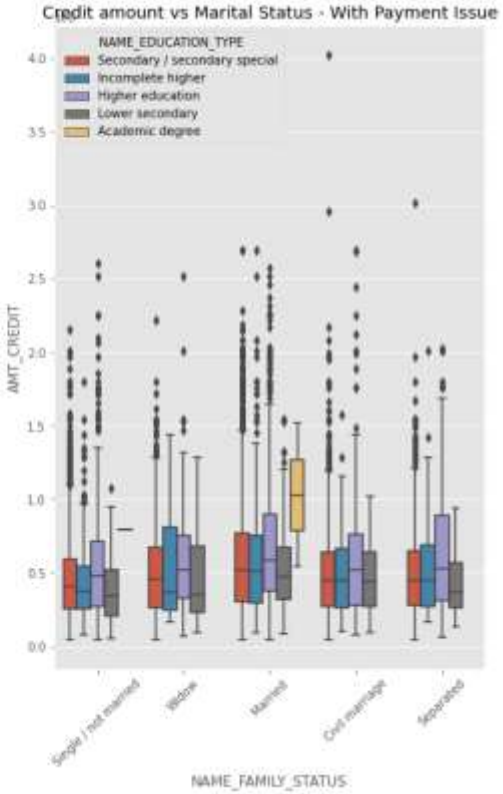


Numeric - Categorical

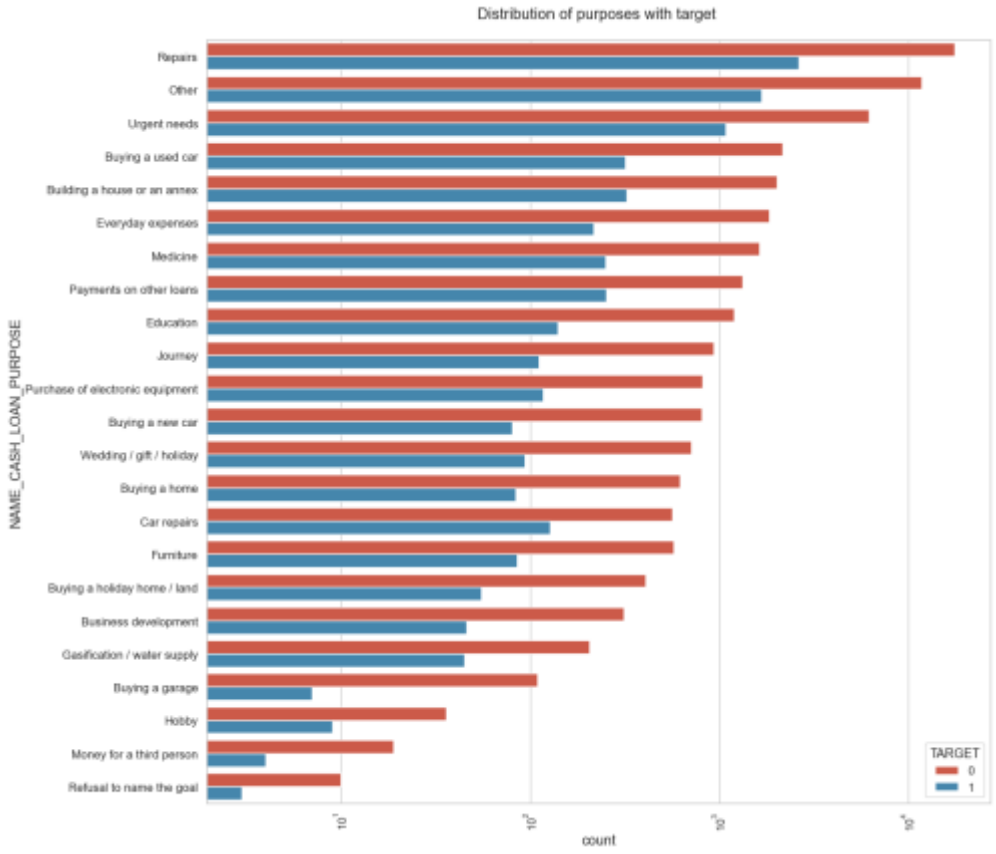






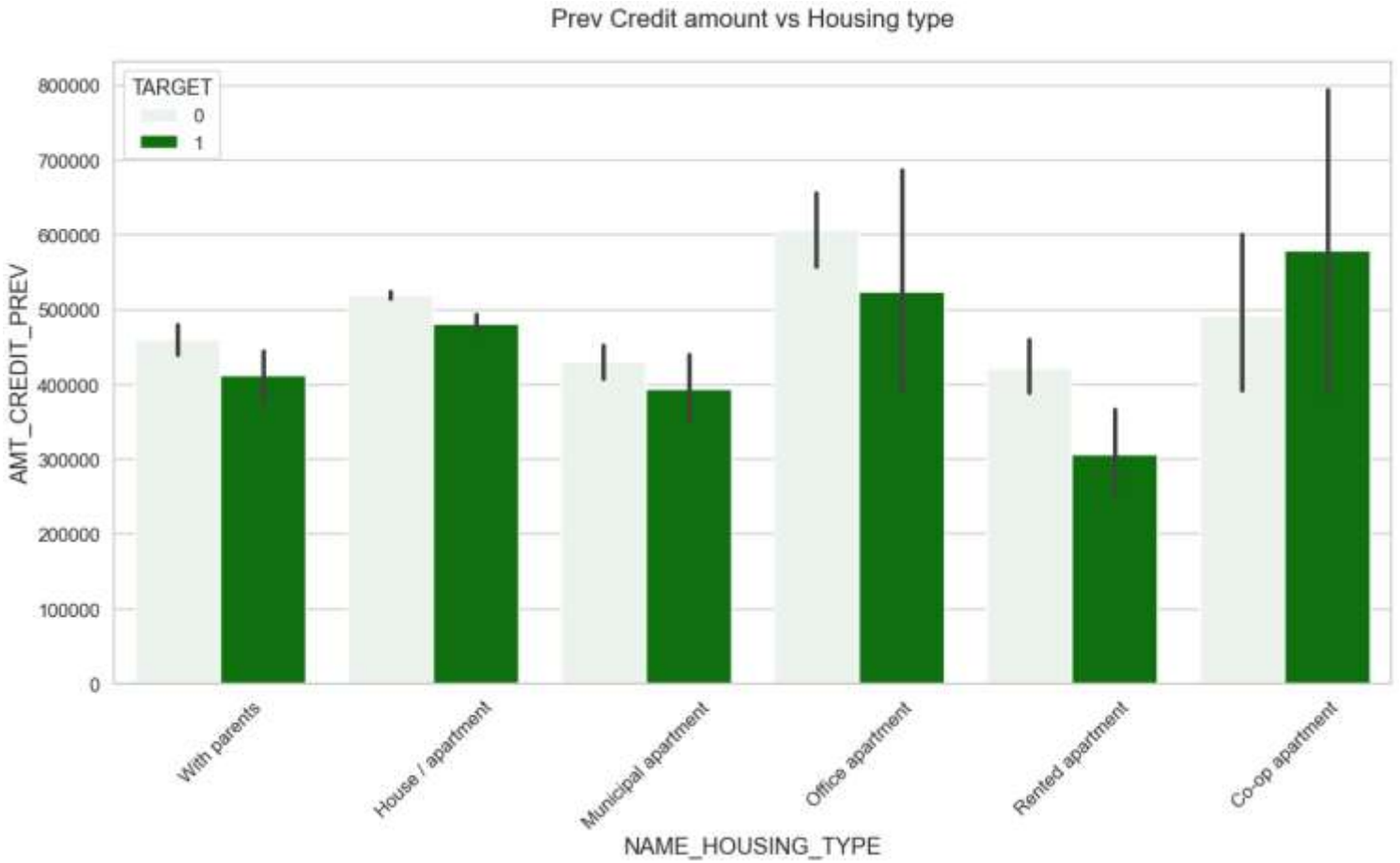


Univariate analysis on merged data



- Categories like 'Buying a garage', 'Money for a third person', etc are the categories who have a high chance of successful loan repayment and should be prioritized.

Bivariate analysis on merged data

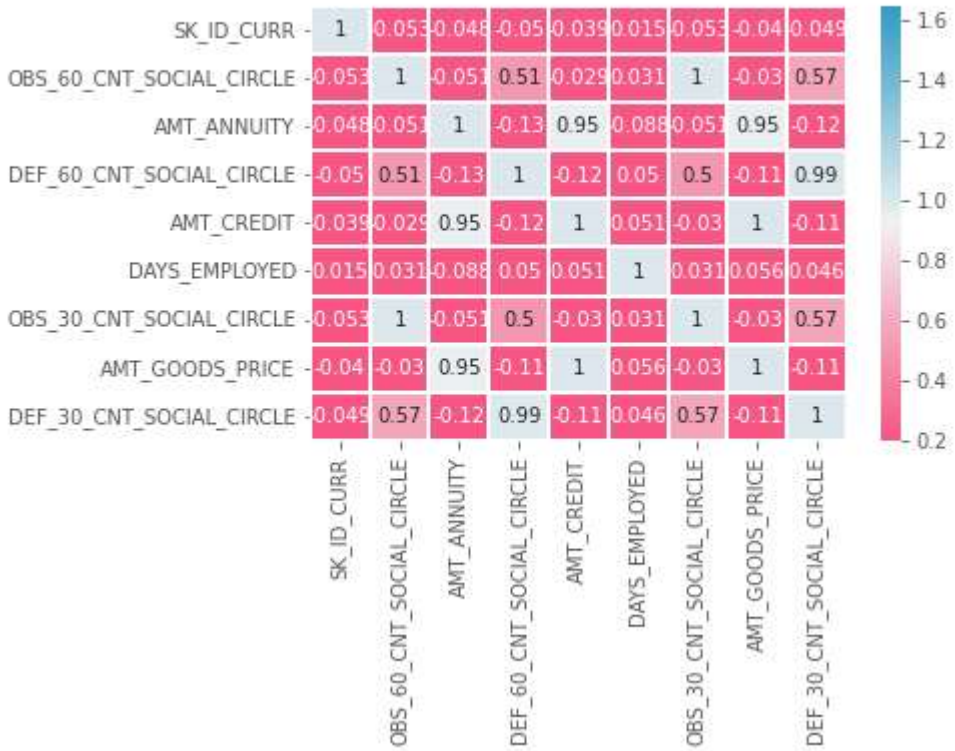


- When it comes to the ease of payment, "office apartment" category has the higher credit as compared to others. Also, bank should be careful while approving loans for 'co-op apartment' categories.

Top 10 correlation for Defaulters

SK_ID_CURR	SK_ID_CURR	1.000000
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269
AMT_GOODS_PRICE	AMT_CREDIT	0.983103
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.847885
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
AMT_ANNUITY	AMT_GOODS_PRICE	0.752699
AMT_CREDIT	AMT_ANNUITY	0.752195
DAYS_EMPLOYED	FLAG_DOCUMENT_6	0.617307

CORELATION MATRIX FOR DEFAULTERS



Top 10 correlation for Non Defaulters

SK_ID_CURR	SK_ID_CURR	1.000000
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
AMT_GOODS_PRICE	AMT_CREDIT	0.987250
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878571
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.859332
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.830381
AMT_ANNUITY	AMT_GOODS_PRICE	0.776686
FLAG_EMP_PHONE	DAYS_BIRTH	0.622073

CORELATION MATRIX FOR NON DEFAULTERS

Summary and Recommendations

- People most likely to repay the loan have **secondary/Secondary special education** status and a **lower income range**.
- Be careful while approving loan to low skilled laborers, male applicants and widows, and those who belong to ‘co-apartment’ categories. Suggestion is to give ***low credit amount*** to these category people, because as per our analysis, ‘**Customers who have low credit amount are more likely to pay back the loan.’**
- While approving, we can ignore the social circle, family having children or not, though married people repay the loan at a higher rate.
- Cash loans is quite higher than Revolving Loans
- Applicants who are working are more likely to apply for loan.
- Working category people apply for loans when compared to unemployed, maternity leave and businessmen. Hence, **bank can create schemes to encourage other category people to apply for loans.**