Presentation On Credit EDA

Abstract: In this assignment we are applying some analytical and visualisation techniques on a real business scenario. We will also develop basic understanding as to how data can be used to minimise the risk of losing money while lending to clients.

Problem Statement:

When the company receives a loan application the company has to decide for loan approval based on applicants profile.There are Two types of risks are associated with the bank's decision:

1. f the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. 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.

Objectives:

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

Structure

1. Application data set
    a. Identifying nulls and outliers
    b. Target variable analysis
    c. Univariate analysis
    d. Segmented analysis
    e. Bivariate analysis
    f. correlation
2. Previous application dataset
    a. Identifying nulls and outliers
    b. Univariate analysis
    c. Bivariate analysis
    d.
3. Merged Dataset
    a. Bivariate Analysis
4. Recommendations

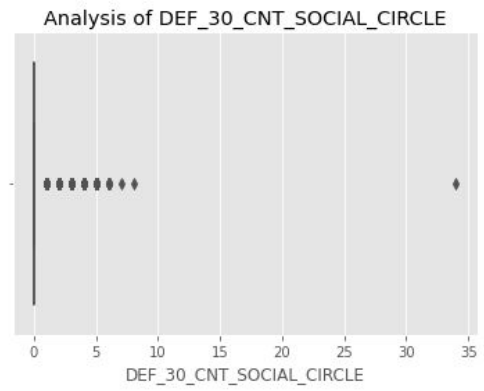
For Application Dataset

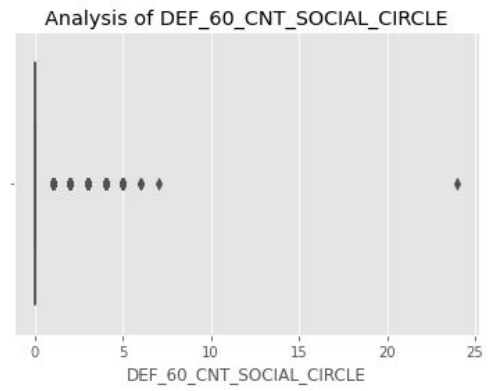
Data Analysis: Identification of null values and outliers

For our analysis we identified null values and where null values were greater than 45% we dropped them

Outliers:

There are many columns with potential outliers and in these columns we used median/mode to impute the missing values.Below are graphs of some potential outliers



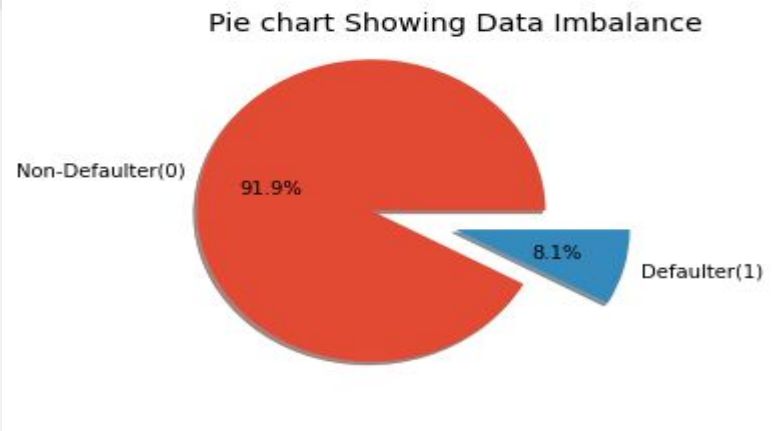


For Application Dataset

Target Variable Analysis:

After performing analysis on target variable we found out there is huge imbalance among defaulters and non-defaulters

The imbalance ratio is 11.39, and as close to 91% are non-defaulters where as 8.1% are defaulters

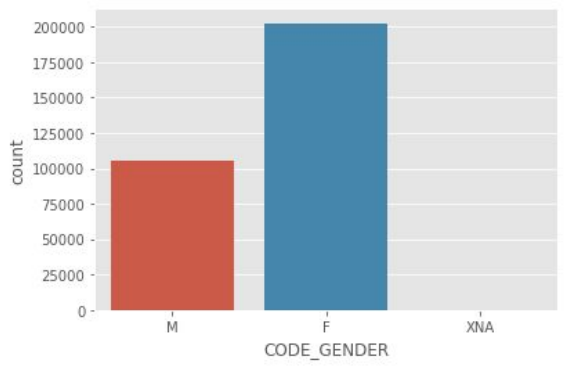


Inference: It is common to have this imbalance in bank dataset, because a bank should always have as much as low percentage of low defaulters.

For Application Dataset

Univariate Analysis:

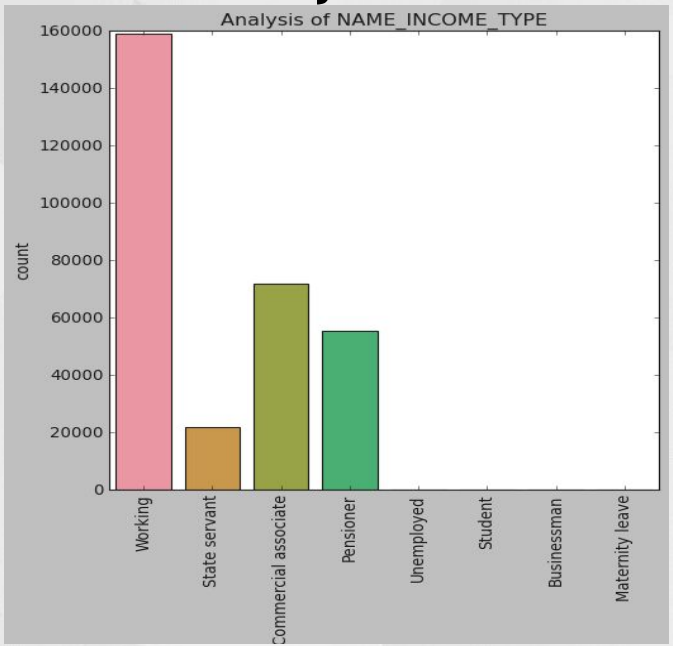
Following shows result of univariate analysis on application dataset



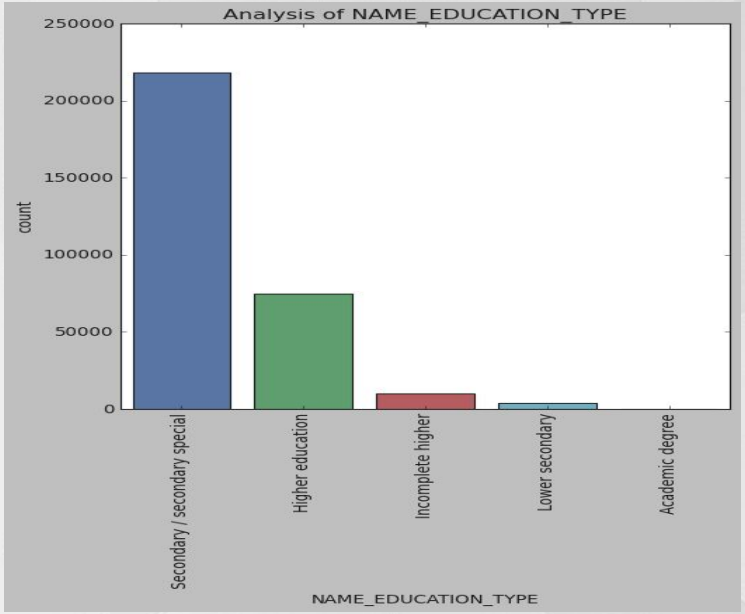
Inference:As there are only 4 (XNA) records in CODE_GENDER, we can impute it with F, as their frequency is much higher

For Application Dataset

Univariate Analysis:



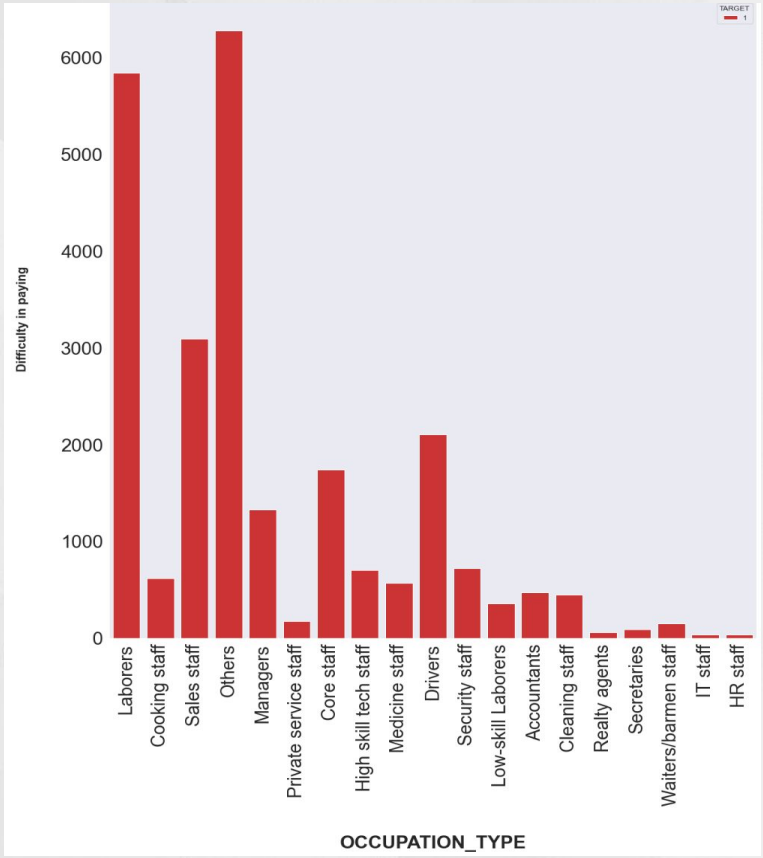
Inference: Working income type have most counts

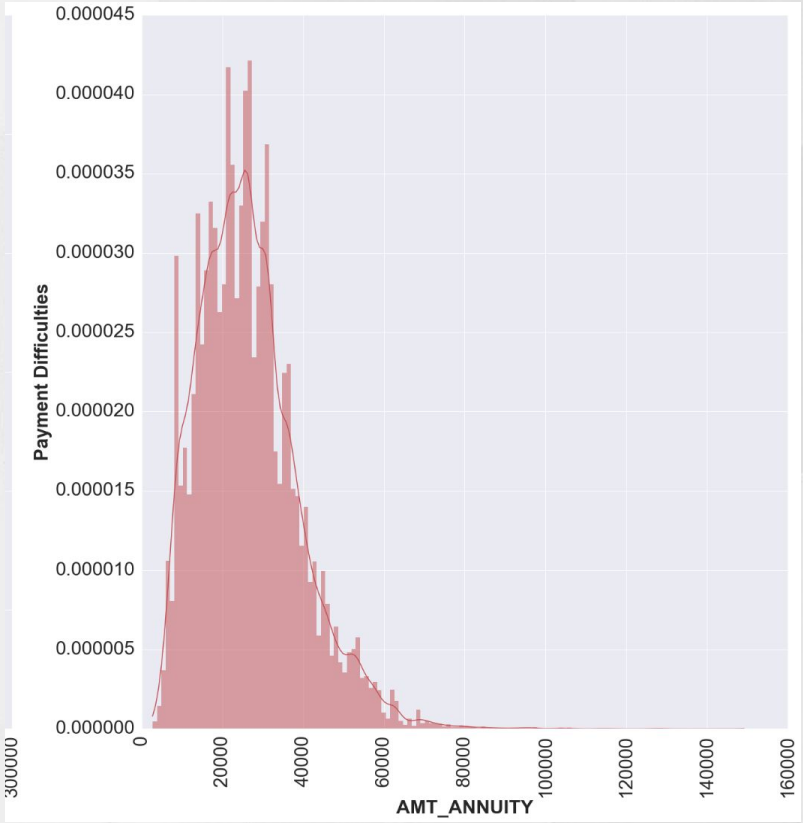


Inference: Education Type having most counts is Secondary/secondary special

For Application Dataset

Segmented Analysis:

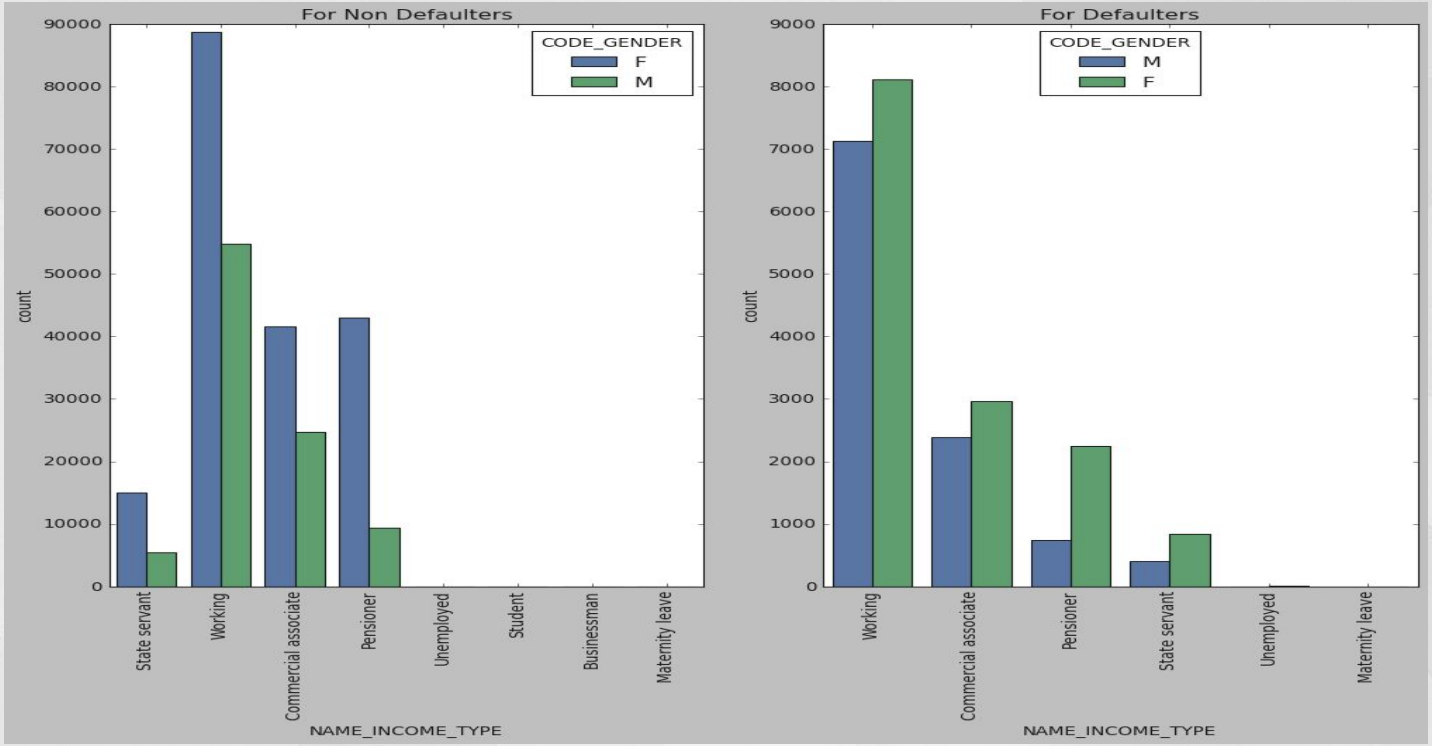




Inference:Labourers and others have high defaults in occupation type while people with difficulty in paying loan among their amt_annuity is more spread

For Application Dataset

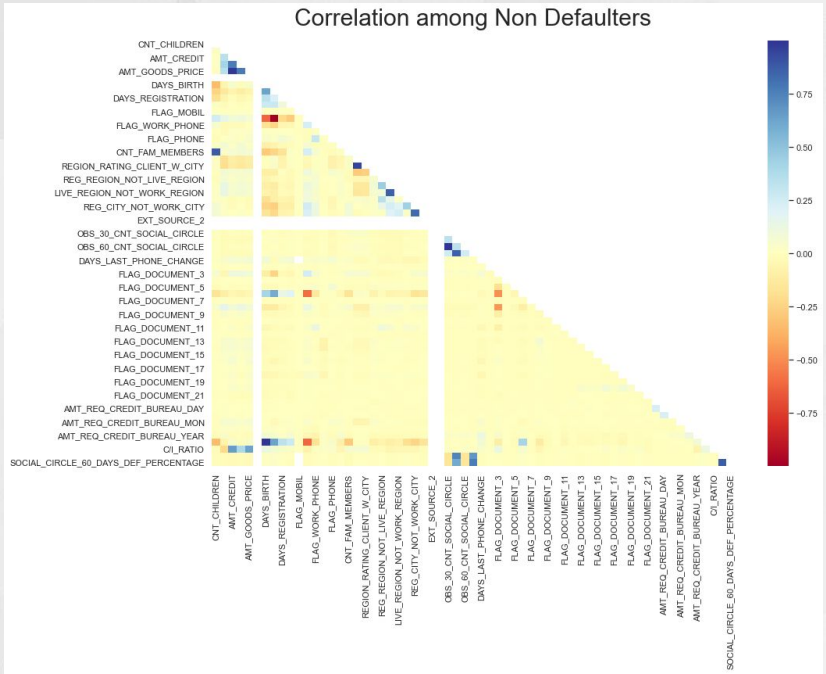
Bivariate Analysis:

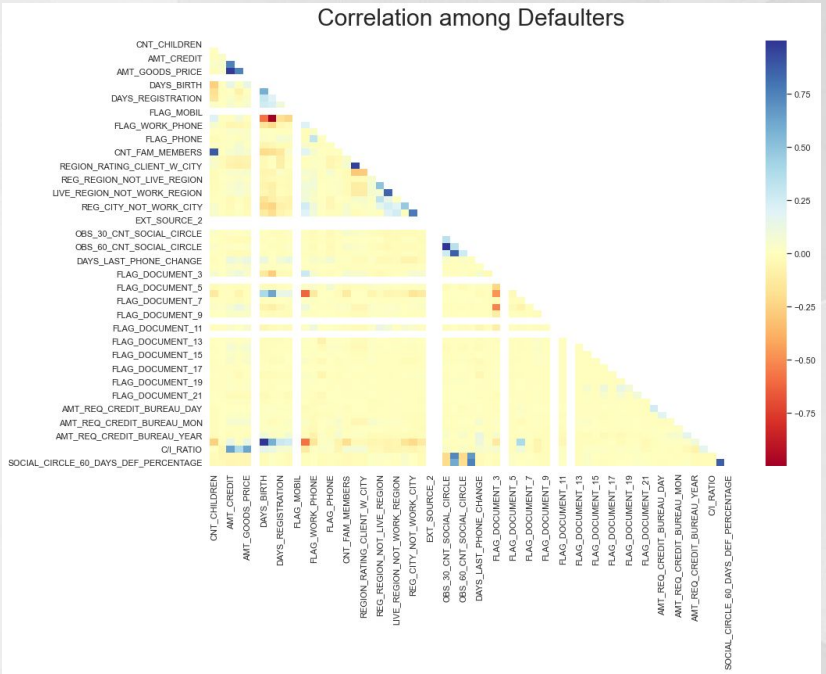


Inference:Working females have both non-defaulter as well as defaulter frequency
For Pensioner the default frequency of women is more

For Application Dataset

Correlation





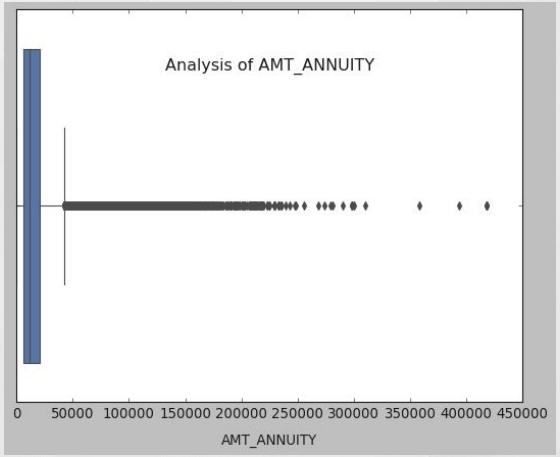
Inference:Correlation looks more or less the same

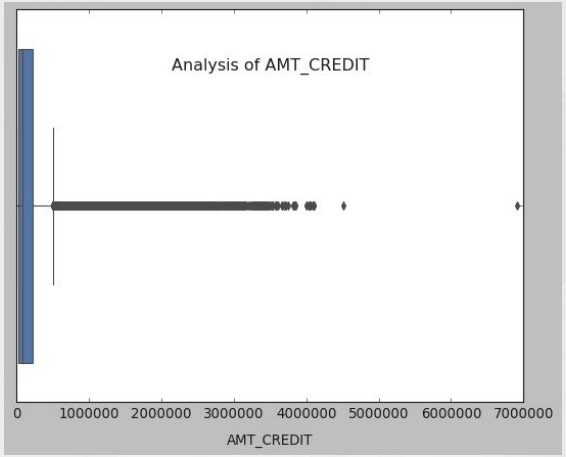
For Previous Application Dataset

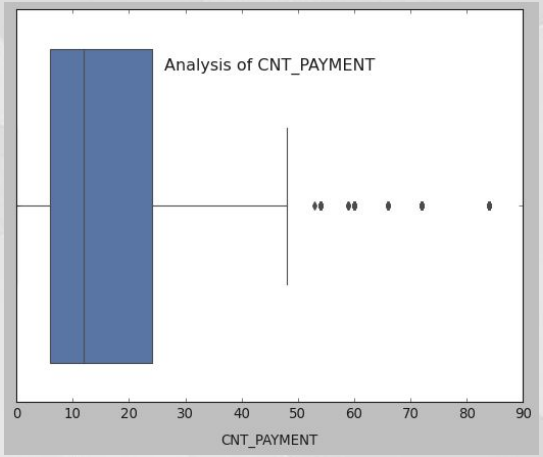
Identification of Null Values and Outliers

We put a capping of 40%, null values greater than 40% are dropped, approximately 11 columns are dropped

Outliers

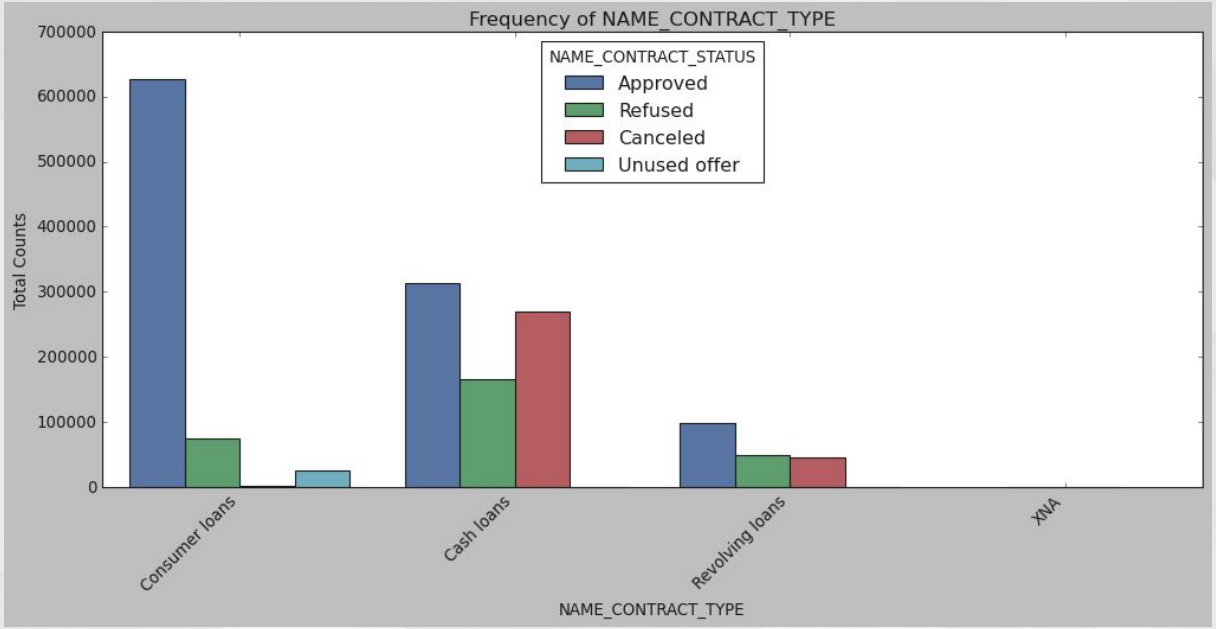






For Previous Application Dataset

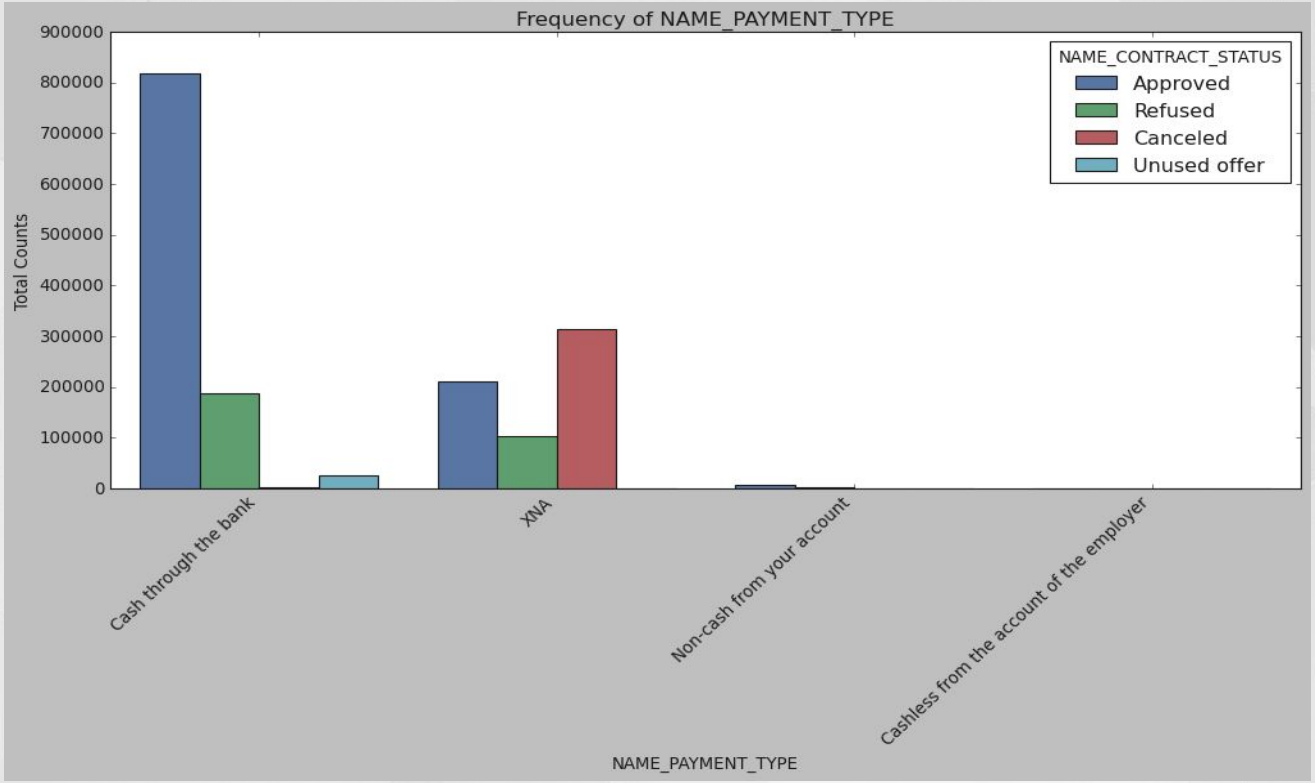
Univariate Analysis



Inference:Frequency of consumer loans is highest and have highest approval too, Cash loans have high frequency but they have the highest cancellation status in graph

For Previous Application Dataset

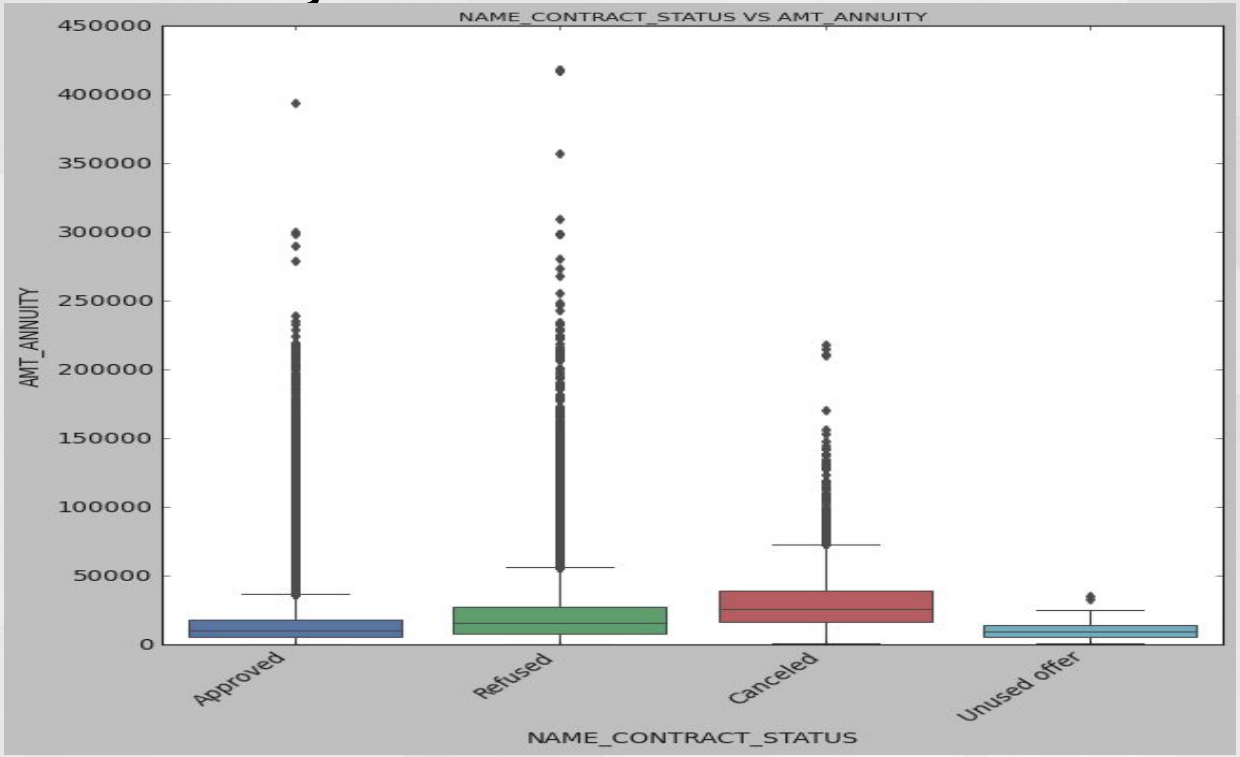
Univariate Analysis



Inference:In the above chart cash through bank is most approved while cashless option is least favourable

For Previous Application Dataset

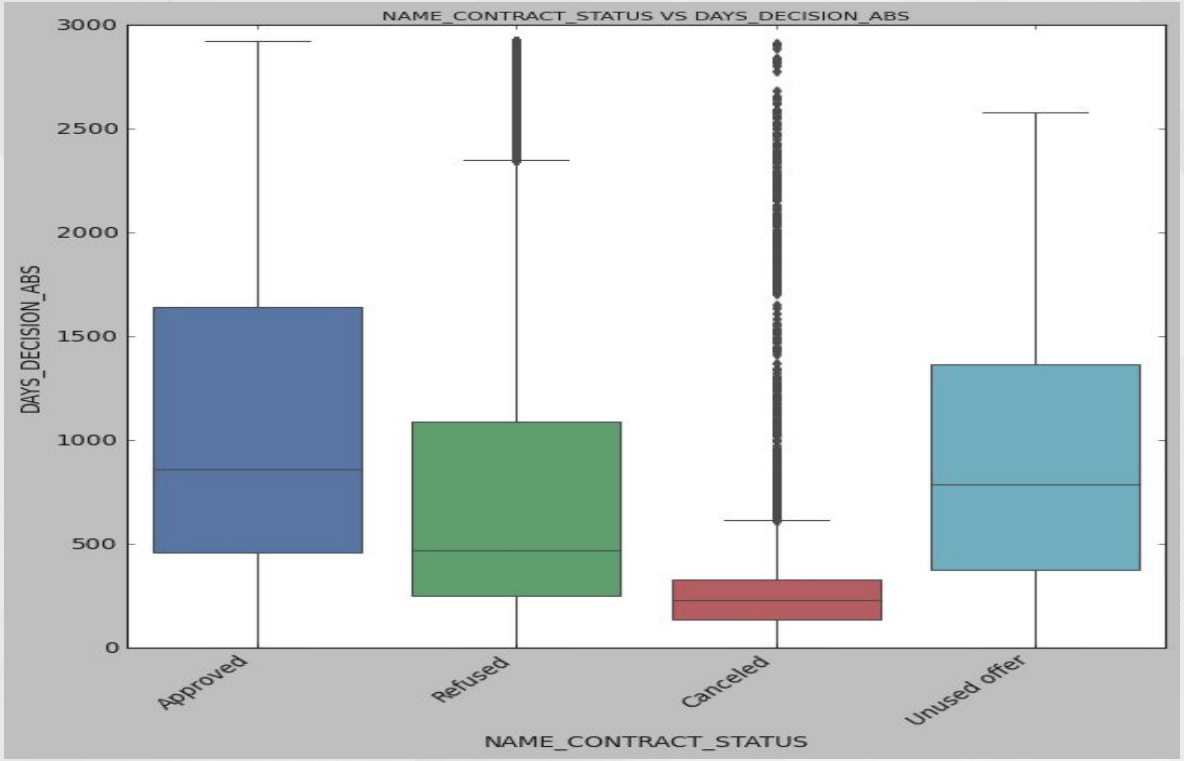
Bivariate Analysis



Inference:People with Low amt annuity dont use their loans, clients with high annuity either are refused or get their application cancelled

For Previous Application Dataset

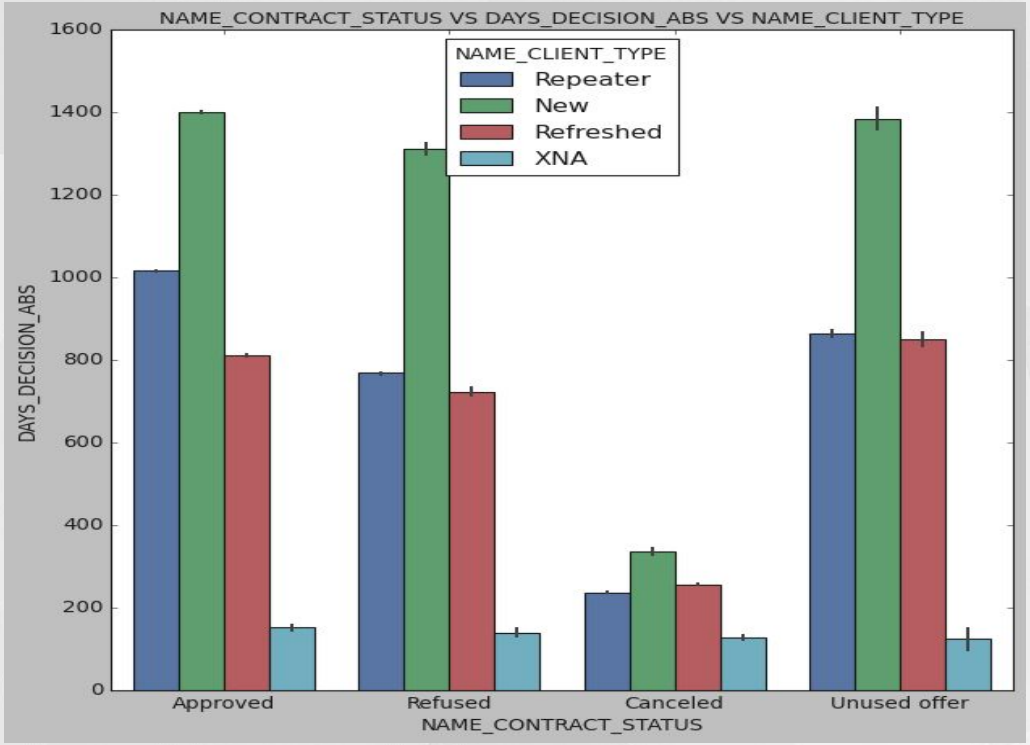
Bivariate Analysis



Inference:Organisation usually takes more days to give approval on loan application, similarly it takes less amount of days to cancel out the application

For Previous Application Dataset

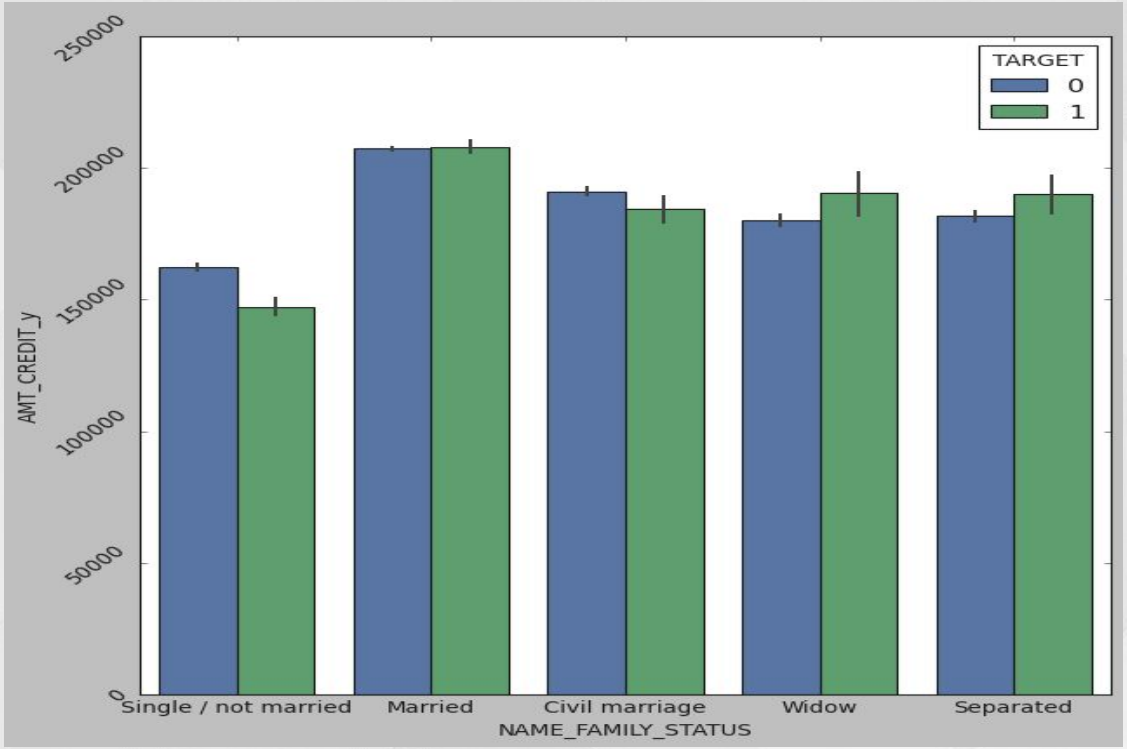
Bivariate Analysis



Inference:We can infer from the barplot that when the client is new , in any case whether the organisation makes a decision it takes the most time for new clients, In case of repeated clients , bank takes more time to give approval than refusing or cancelling

For Merged Dataset

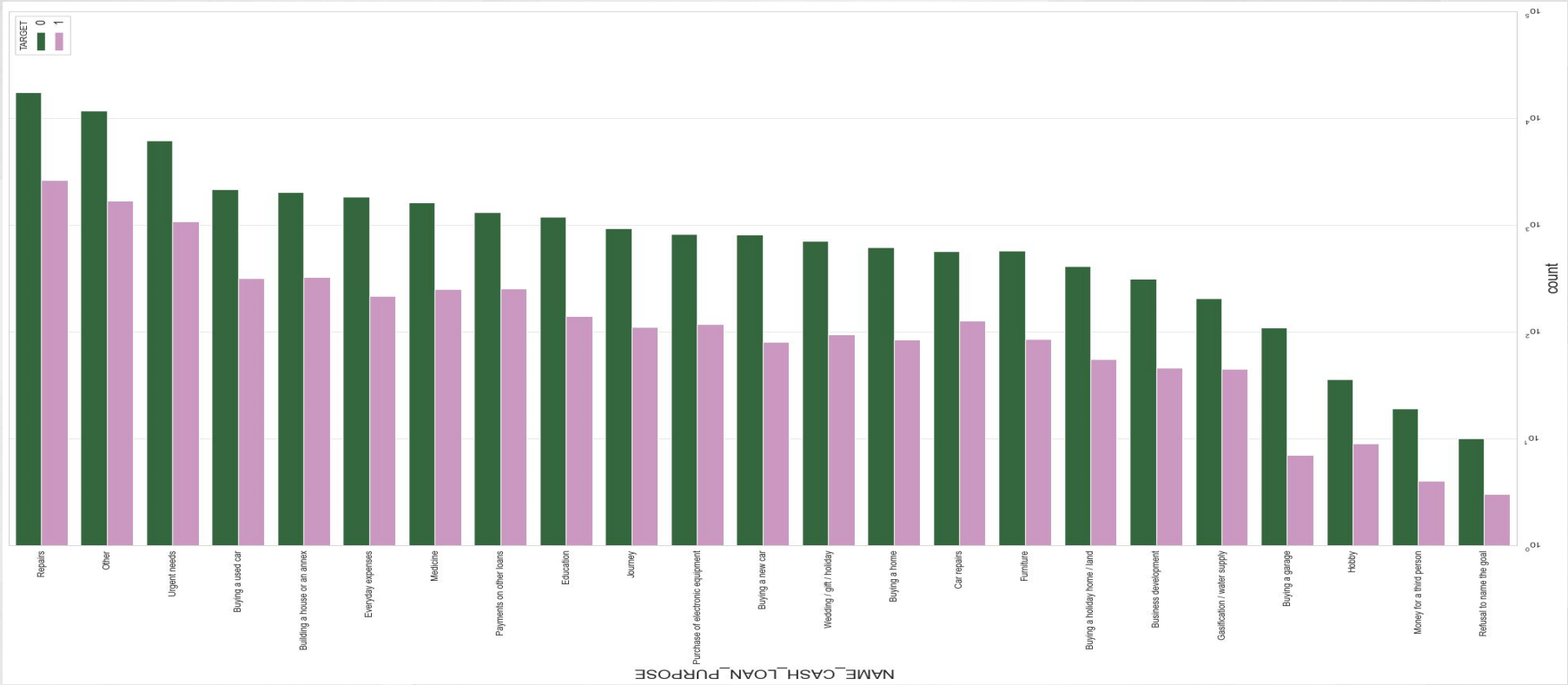
Bivariate Analysis



Inference:Spearated and Widow people have default ratio more along with higher credit amount

For Merged Dataset

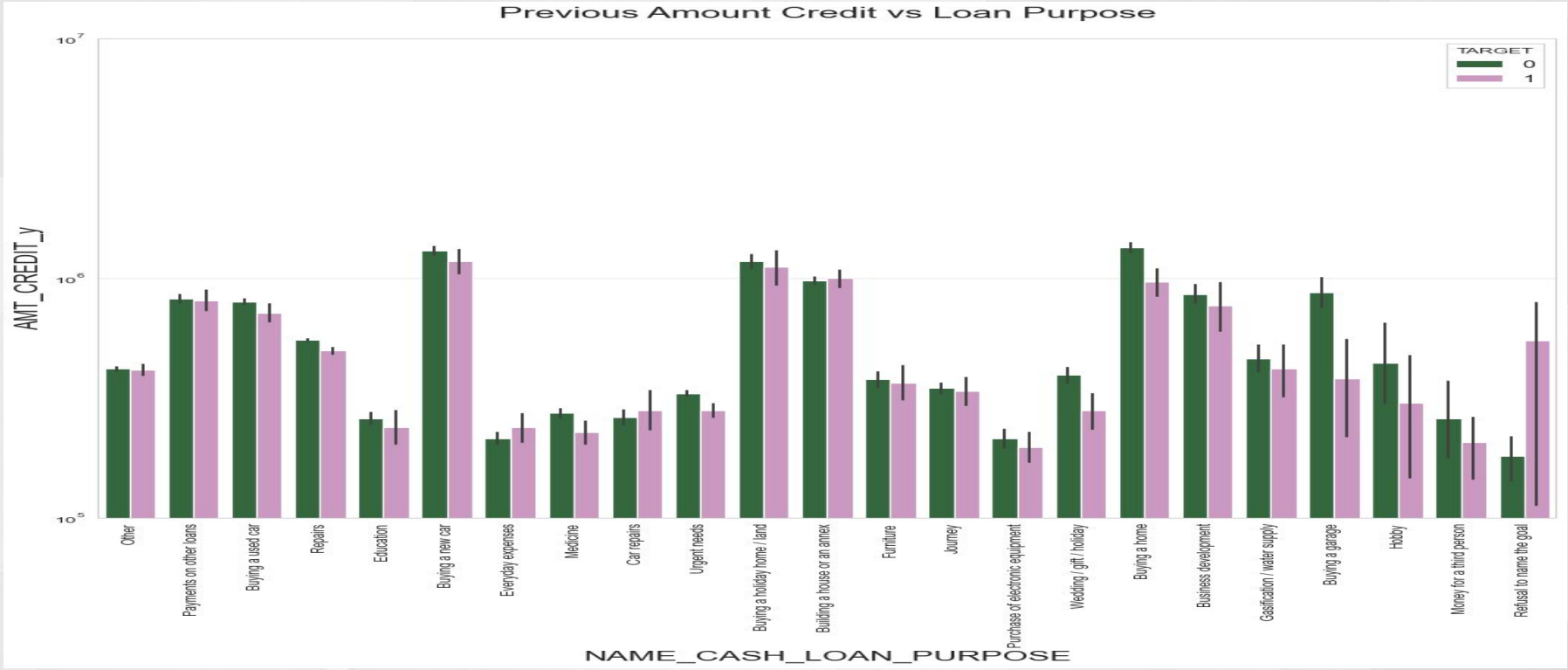
Bivariate Analysis



Inference:Buying a holiday home/land, Business development, Buying a Garage, Buying a Home , wedding/gift/holiday, all these categories have higher non-defaulter and lower defaulter count

For Merged Dataset

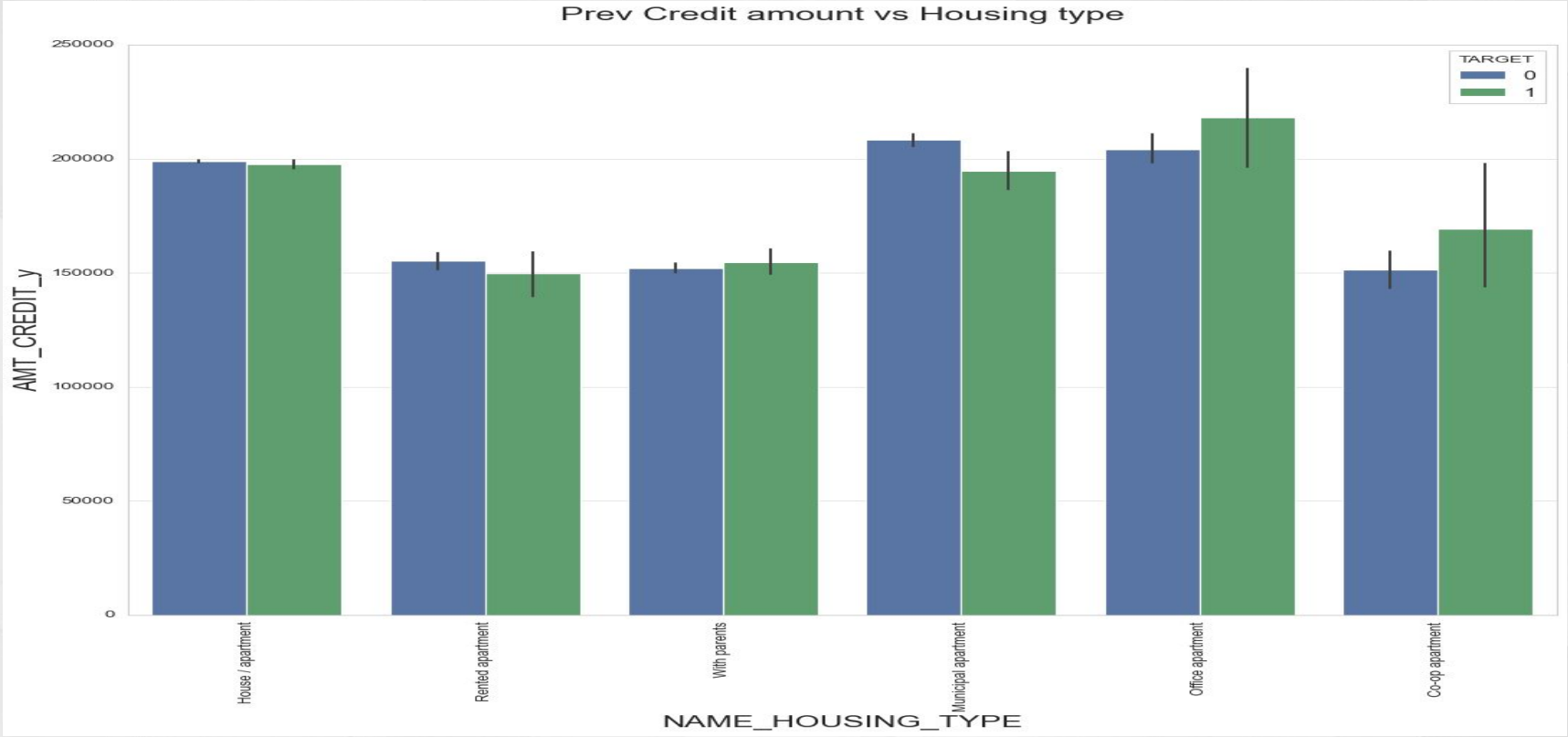
Bivariate Analysis



Inference:Following categories: Refusal to name the goal, building a home or an annex. have higher defaulters especially Refusal to name the goal.

For Merged Dataset

Bivariate Analysis



Inference:Co-op apartment and office apartments have high defaulters

Recommendations

1)Banks should target following cash loan purposes more "Buying a holiday home/land, Business development, Buying a Garage" as they have paid their loans on time

2)Widow and Separated people have high defaults on larger amount credits, so Banks should give them less credit

3)Undisclosed Goals for loan purpose, this category has payment difficulties, so maybe for this category Bank should put a cap on amount credit or increase rejection ratio.

4)Non-Defaulters have slightly higher canceled and refused status, Bank should use this indicator and work on bringing down the ratio for non-defaulters.

5)The banks should have taken more time and do their due diligence on loan applications as defaulters who are either New or Repeaters got their decision in less time than Non-Defaulters.

6)Lastly Banks should put a capping on amount credit on clients who have co-op apartment accommodation or office-apartment, as they have high defaulters.