Credit EDA Assignment

- By Srishti Pandey

Objective

To identify patterns wherein the client is not able to repay the loan or does not take up loan, so that the bank can decide on :

- Under what conditions can a loan be sanctioned to an applicant
- Launching a new scheme
- Denying the loan

Assumptions Made

- **Target Variable :** We have a Target Column in the dataset which defines if an applicant has had difficulty is repaying the loan or not. Has values 0 and 1.

1 - if an applicant has difficulty in repaying the loan

0- all the other cases

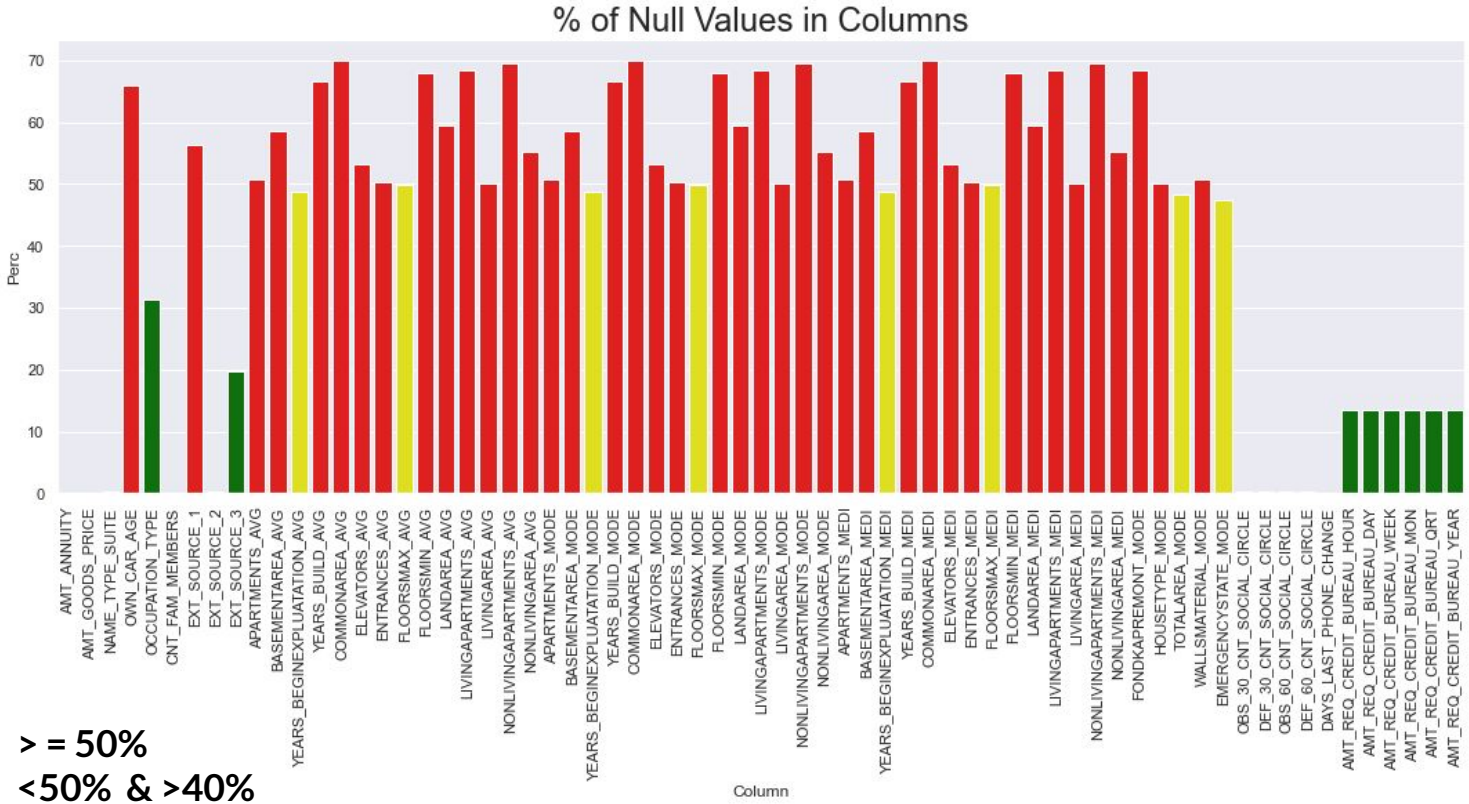
- After merging data of Previous Year is stated with suffix ‘_y’ and that of Application Data is stated with suffix ‘_x’ .
- In the Analysis section the ‘orange’ graph refers to data from Application Data while the ‘blue’ graph refers to data from Previous Year Data.

Overall Approach

Note : The *Overall Approach* for both the datasets has been same.

- **Sanity Checks :**
  - Analysing data at hand using shape, info(), describe(), head()
  - Looking at the overall null values present in the data set
- **Data Cleansing :**
  - Creating a dataframe for storing names of null-columns and the number of null values
  - Converting the number of null values into percentages
  - Dropping the rows having 0% of null values
  - Utilizing bivariate visualization technique to understand what is the upper and lower limit for % null values.
  - Setting threshold value
  - Dropping the columns exceeding the threshold

Percentage of Null Values



> = 50%

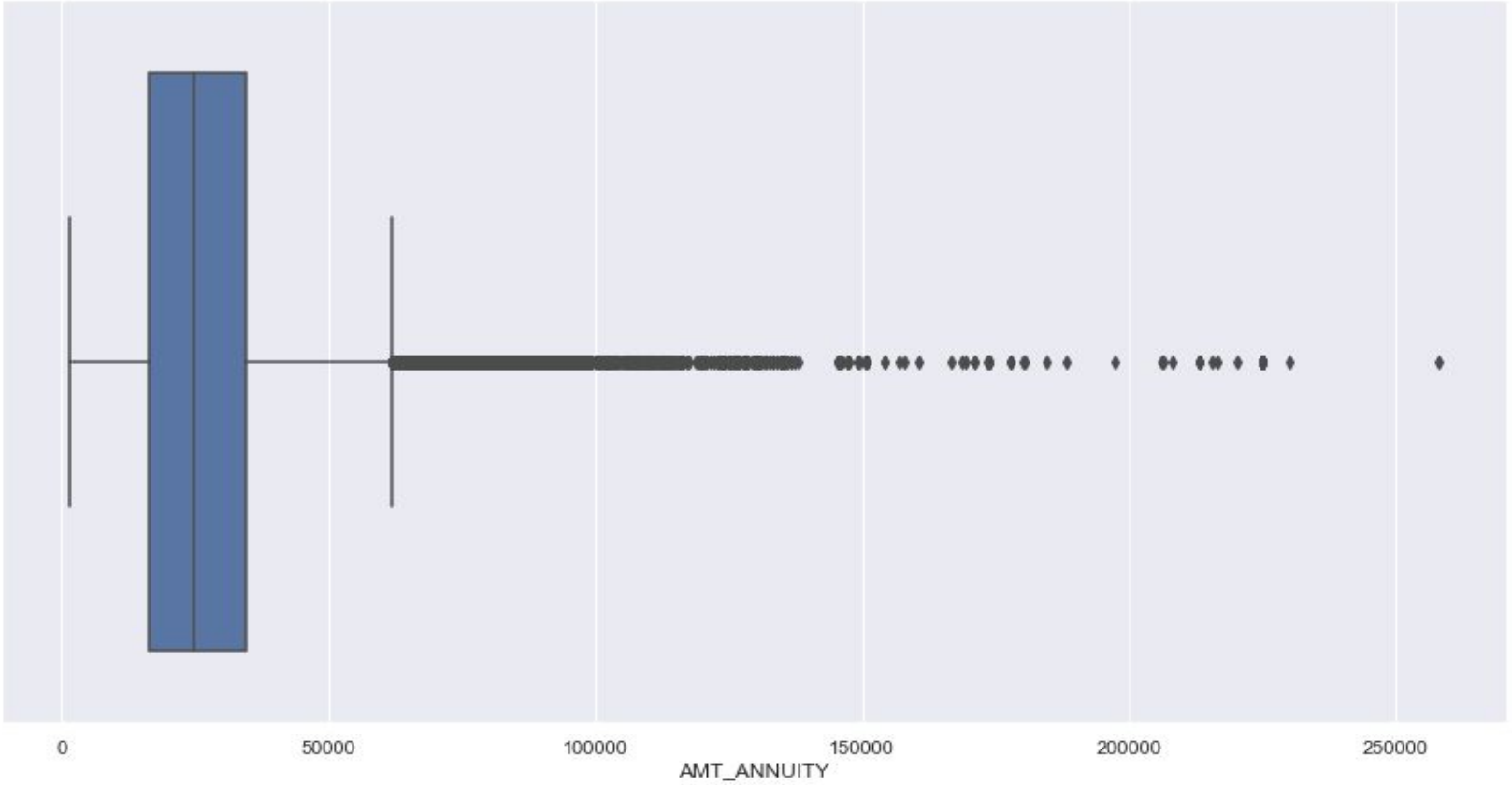
<50% & >40%

< 40%

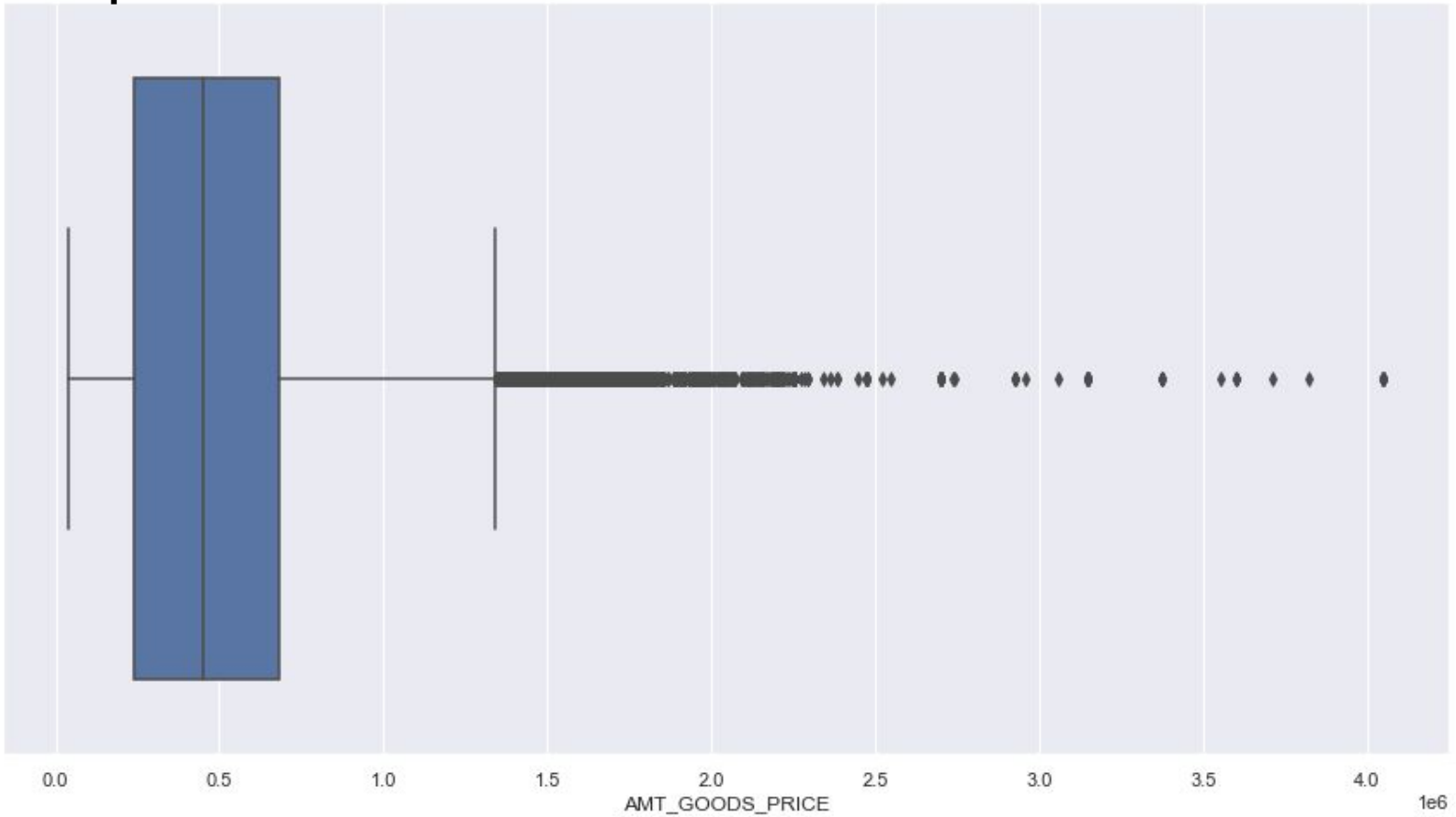
- **Checking for Outliers:**

    - Plotting boxplots and histogram for various features for checking the presence of outliers.

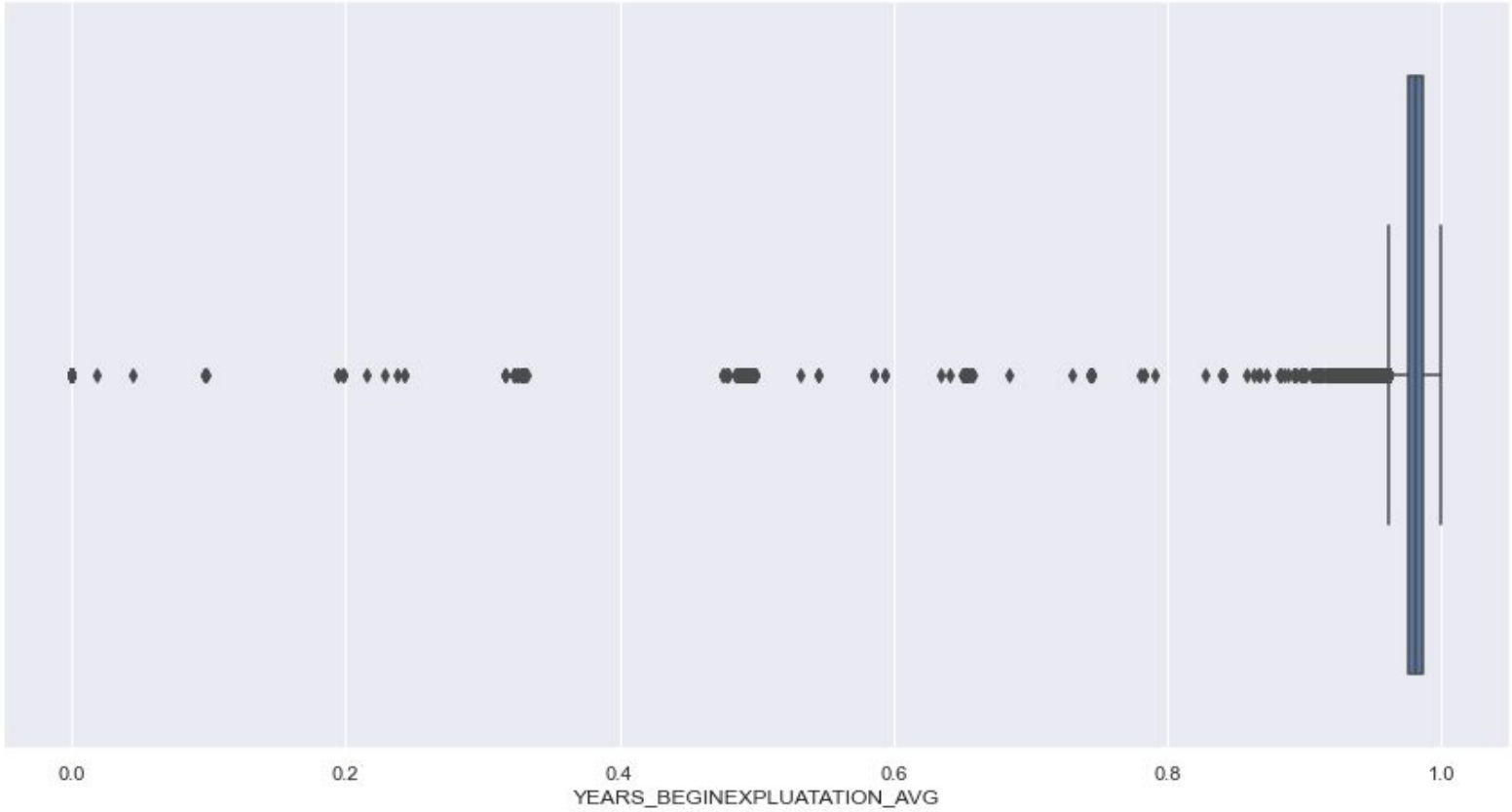
- **AMT_ANNUITY : Annual amount to be paid in order to repay the loan has outlier at 258025. To impute we use median.**



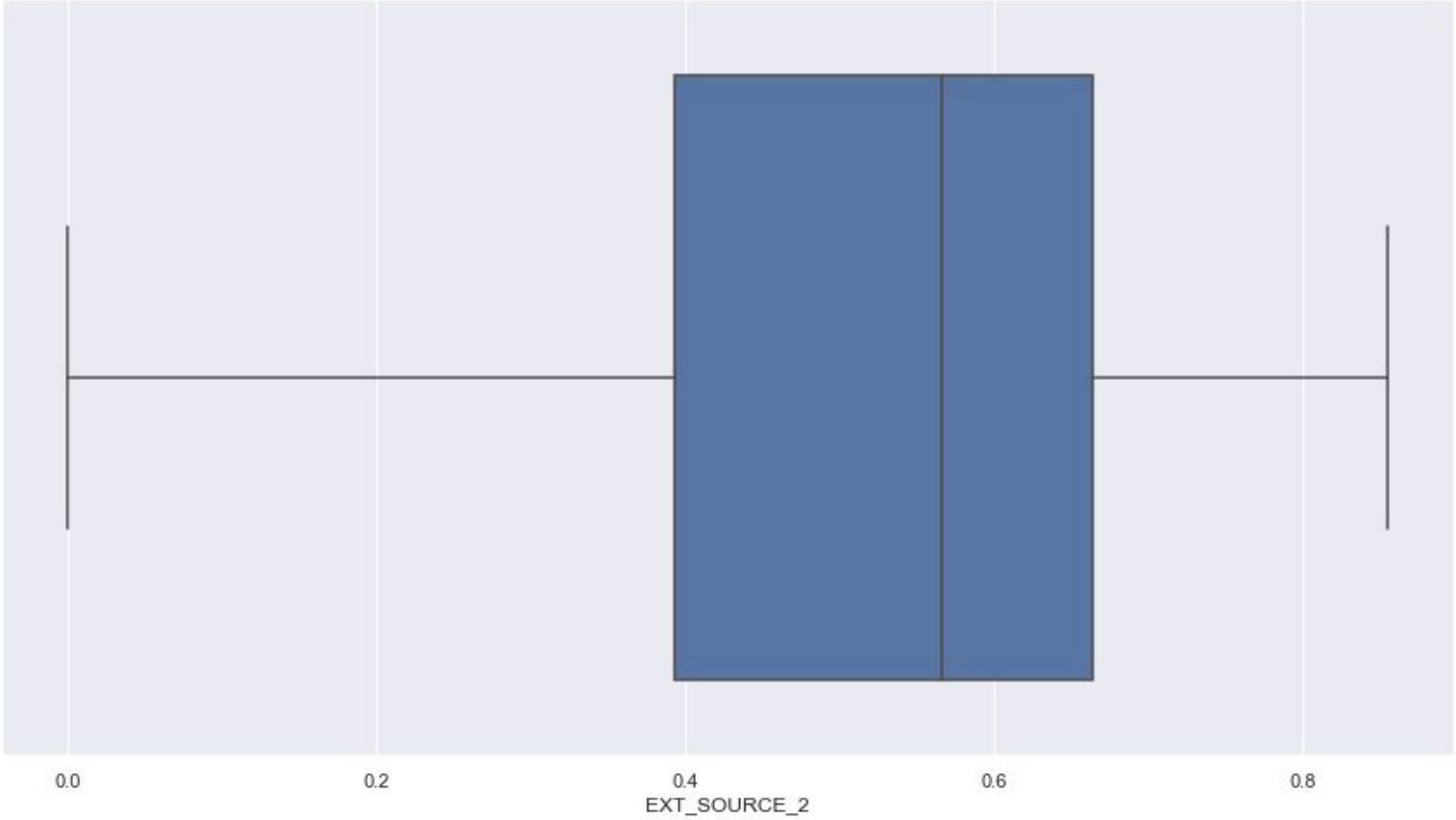
- **AMT_GOODS_PRICE : The Value of product for which loan has been taken. To impute we use median**



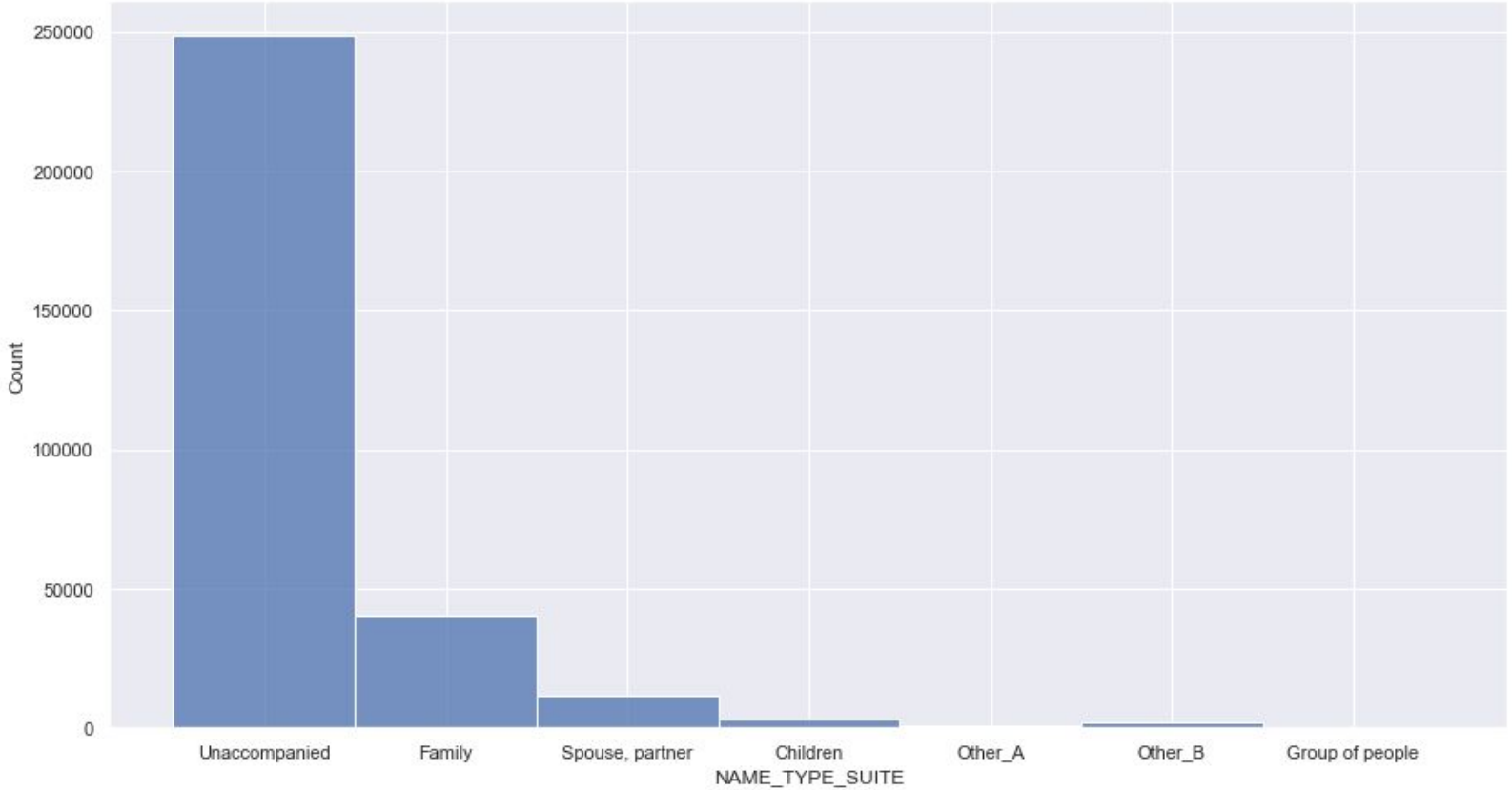
- **YEARS_BEGIN_EXPLUATATION_AVG : Normalized information about building where the client lives. Has too much variability, i.e has more than 40% of data missing. So we drop this column instead.**



- **EXT_SOURCE : Normalized information from external source has no outliers**



- **NAME_TYPE_SUITE : To identify who accompanied client in the previous application. We choose mode for imputing this and other categorical**

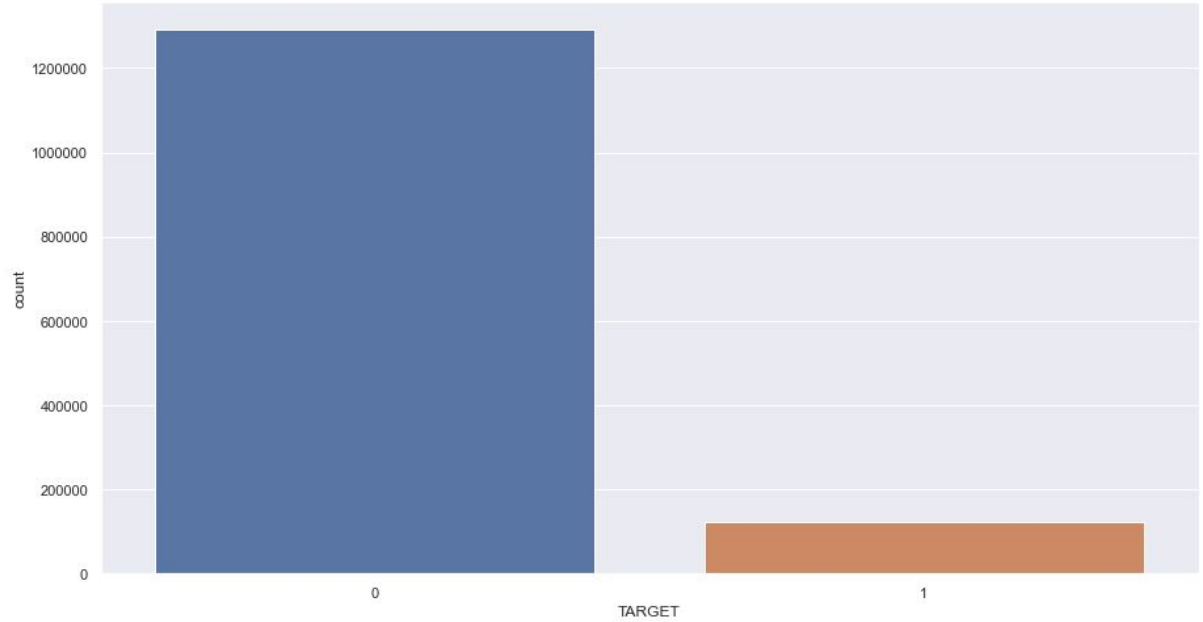


- **Imputing Missing Values :**

    - Checking if the featured having null values have categorical or continuous data.
    - Certain visualization techniques like histogram used for clear identification of categorical / continuous data.
    - For categorical impute mode, for continuous impute median.

Imbalance Target

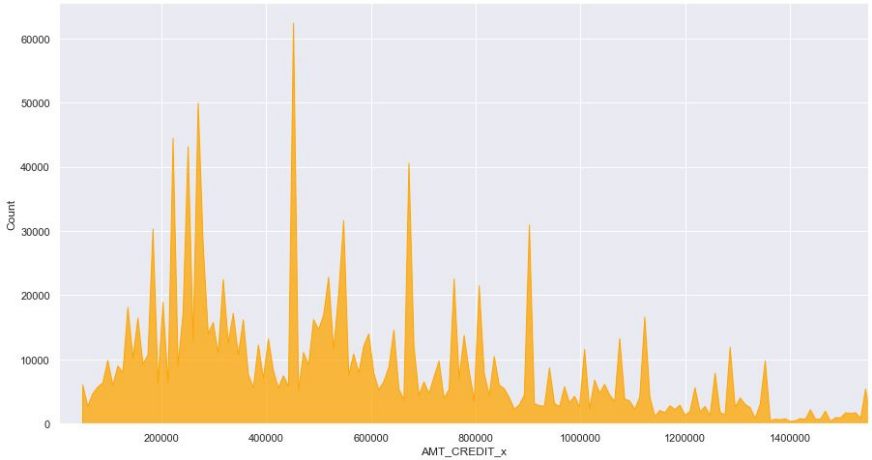
We can see that the Imbalance in the Target variable is too much, ie the people who paid the loan late on at least one of the first installments of loans taken is way too high. These people fall in high risk category for defaulting the loan payment.

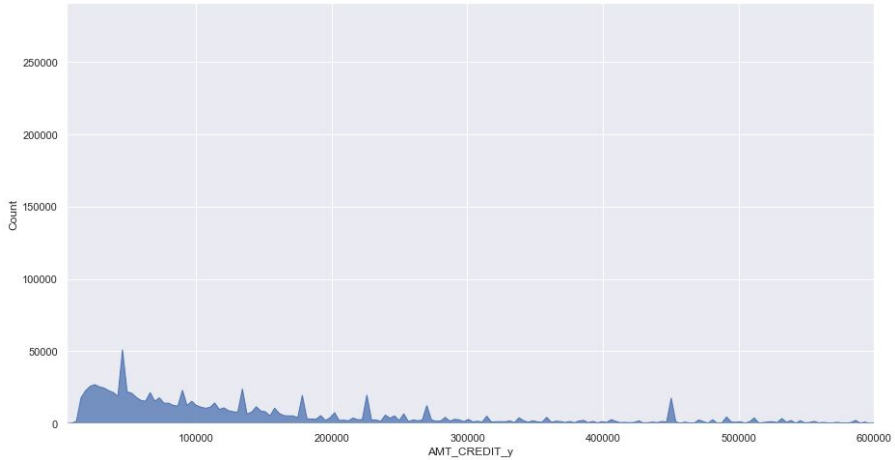


Univariate Analysis

Note : The ‘orange’ graphs are for Application data and the ‘blue’ graphs are for Previous Data.

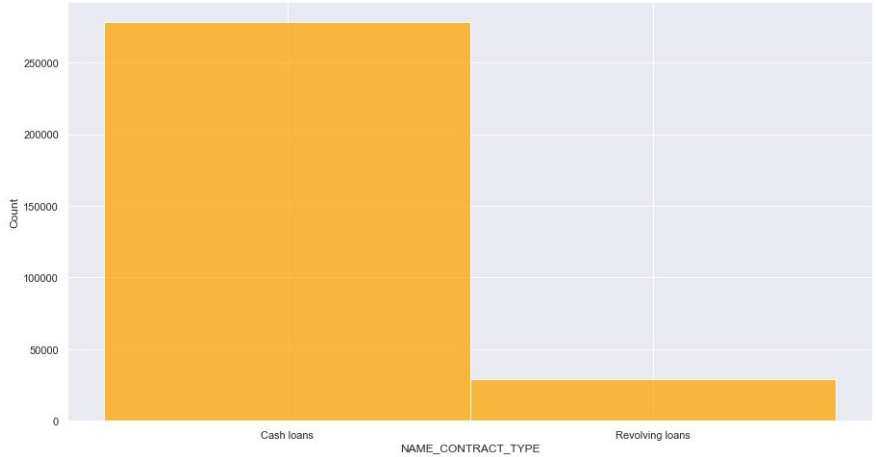
Amount_Credit : When verified, it has been observed that, Customers who have low credit amount are more likely to pay back the loan.

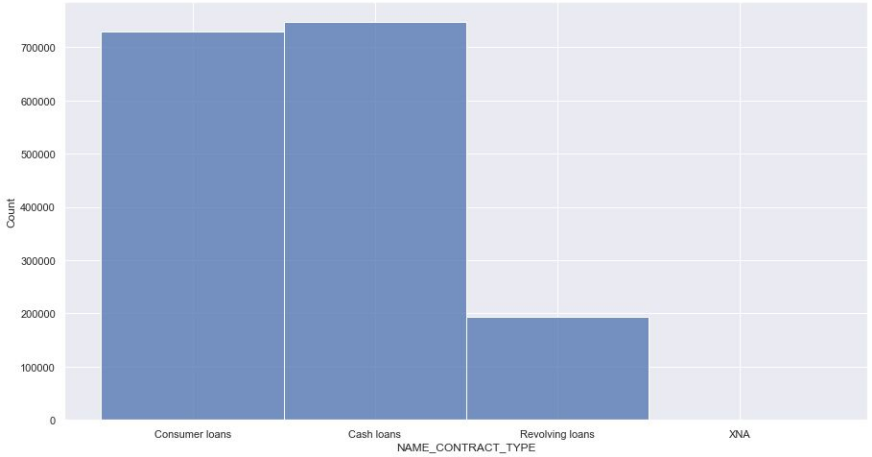




Univariate Analysis

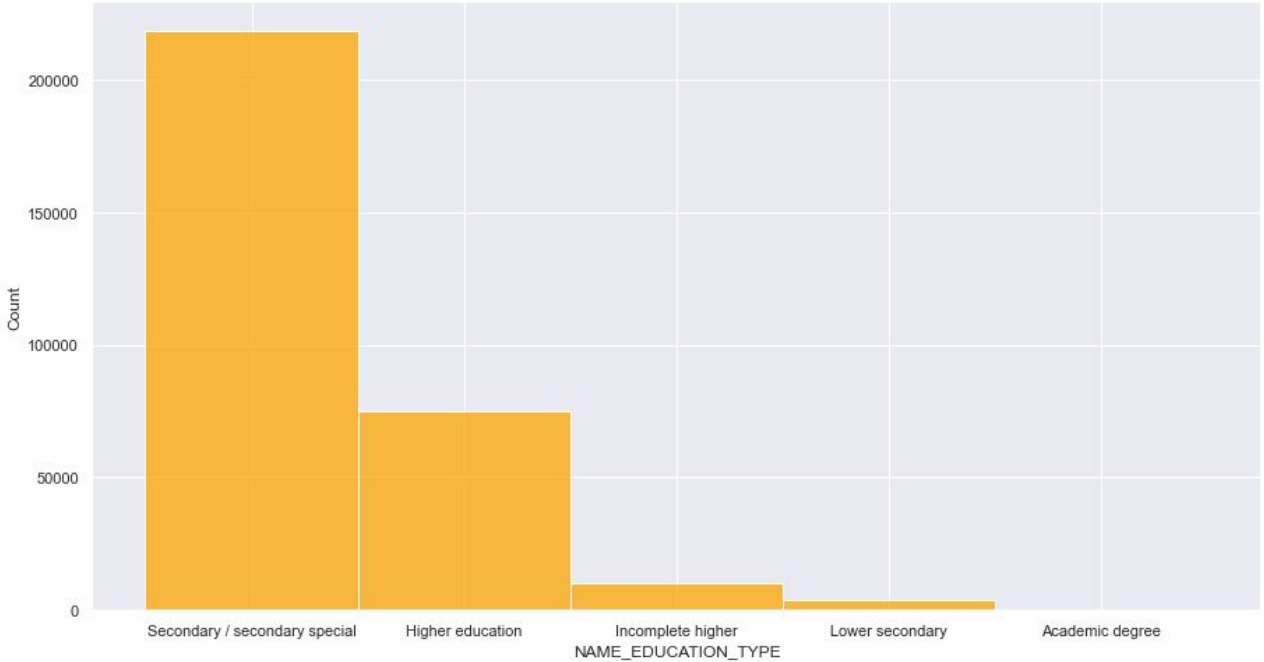
Name_Contract_Type : We can see that we have had different kinds of loans in previous and current year. Cash loans have spiked in current year while revolving loans have decreased.





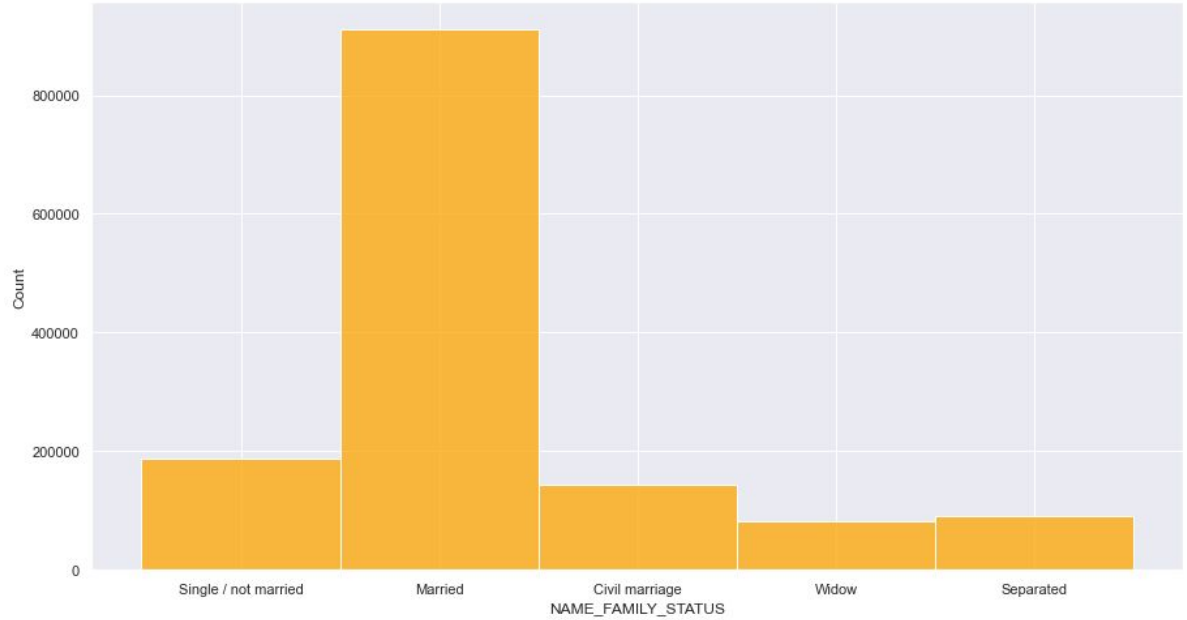
Univariate Analysis

Education_Type : We can see that people who have senior /secondary level education have taken more likely to take loan.



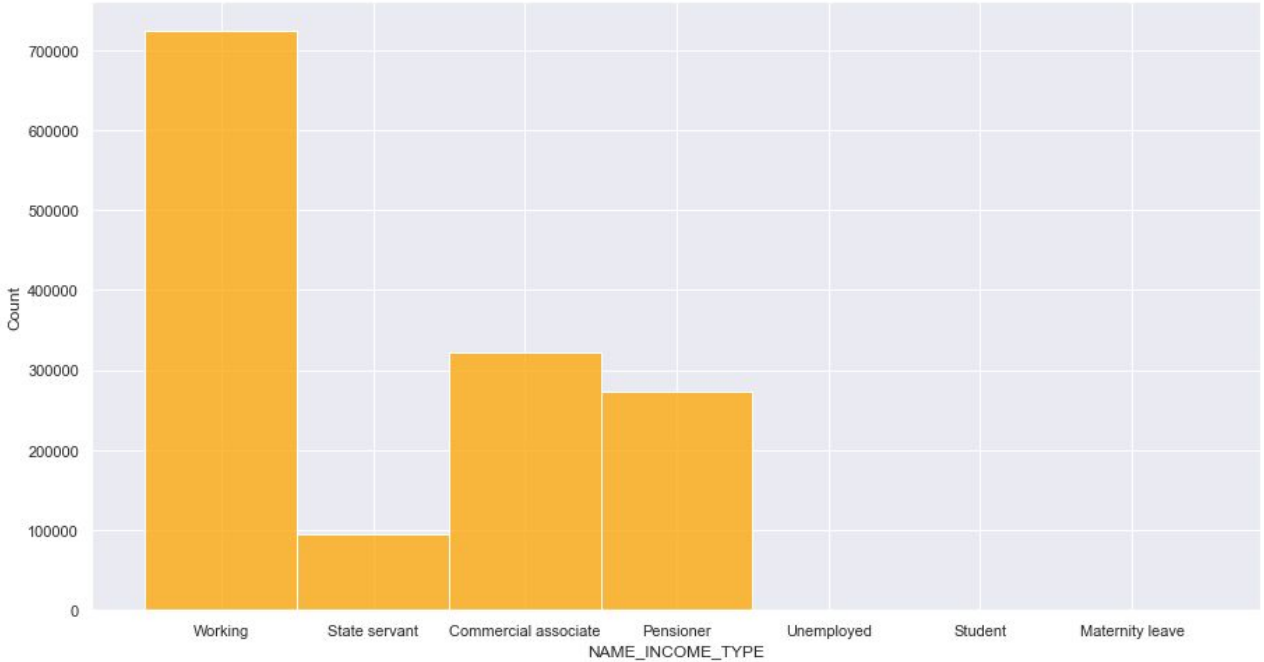
Univariate Analysis

Family_Staus: We can see that more married people take up loan



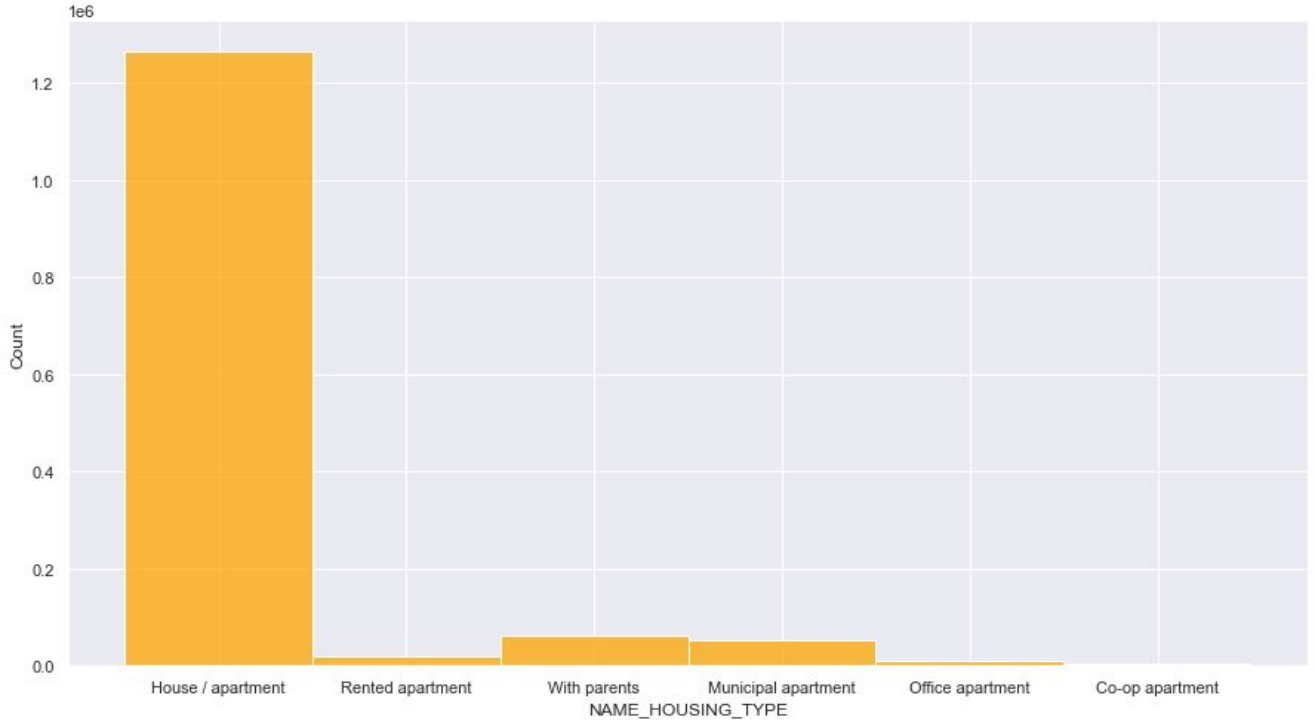
Univariate Analysis

Income_Type : More working people take loan.



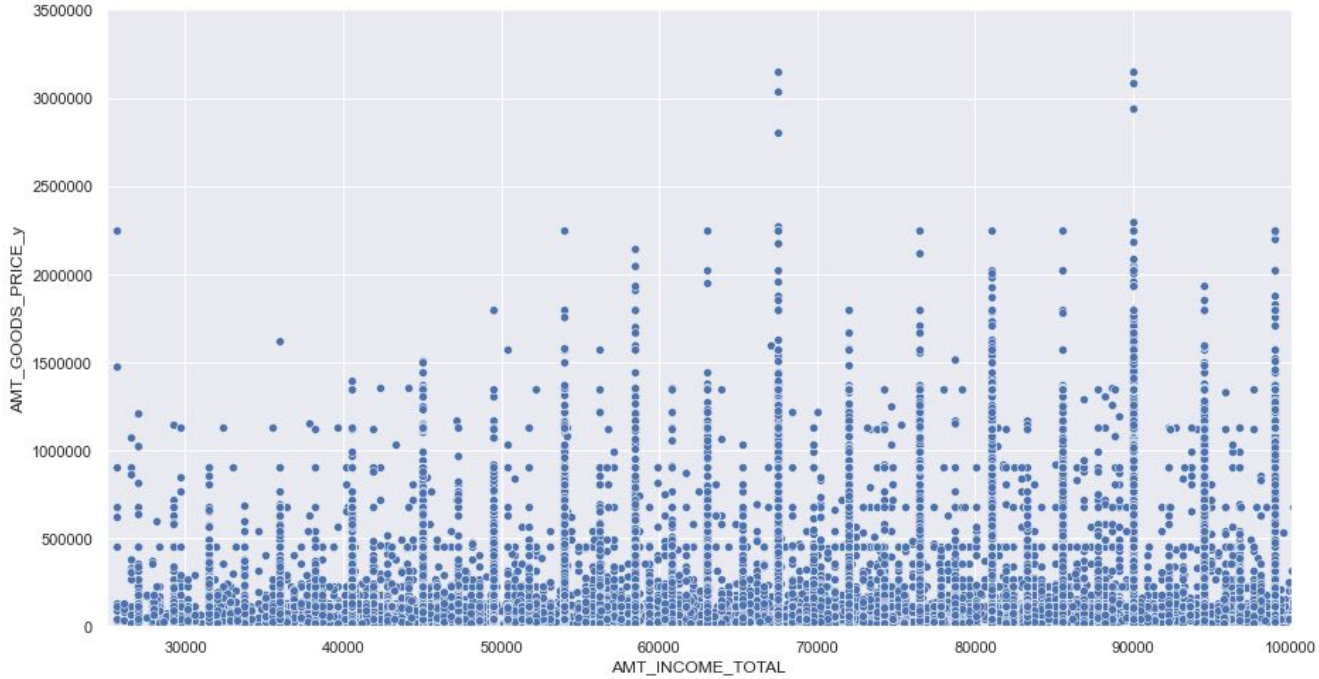
Univariate Analysis

Housing_Type : People who live in their own houses / apartments are more likely to take up loans



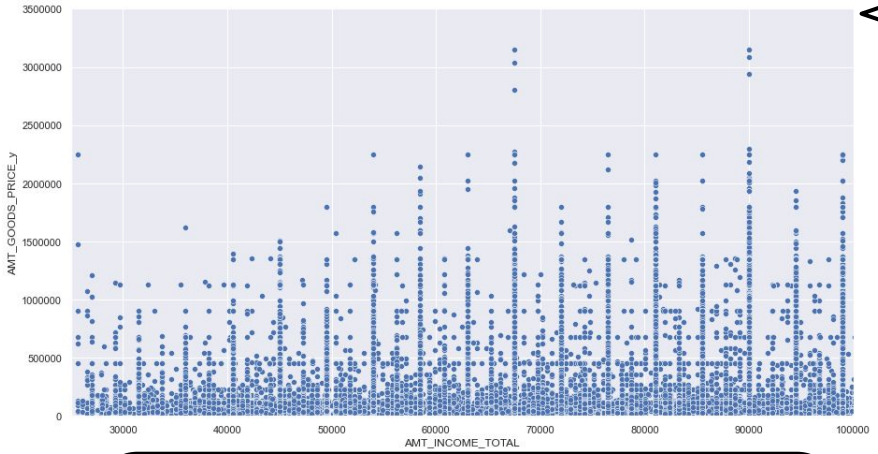
Bivariate Analysis

Amt_Goods_Price_Y -vs- Amt_Income_Total : We can see that Last Year there was an increasing trend of people taking loans of higher value as the income increases with the mean value of good taken at loan being around 50,000.



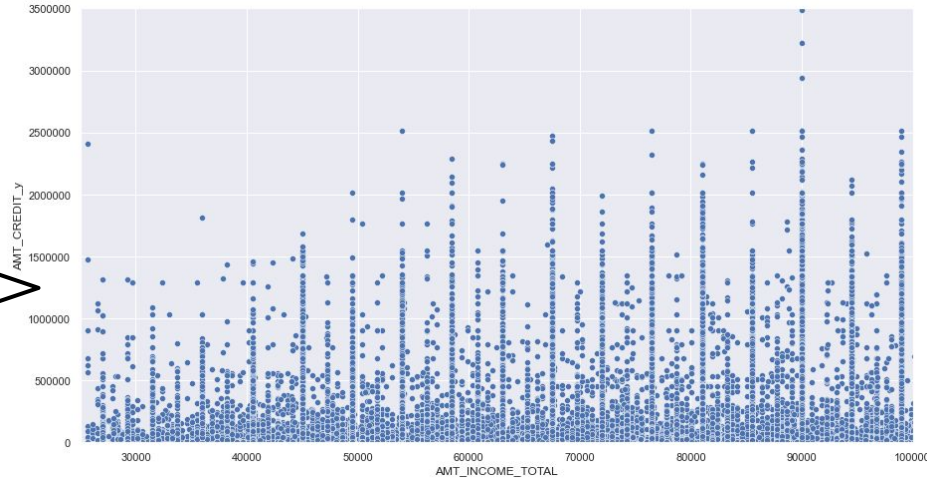
Bivariate Analysis

This trend is very similar to the amount for which credit was taken



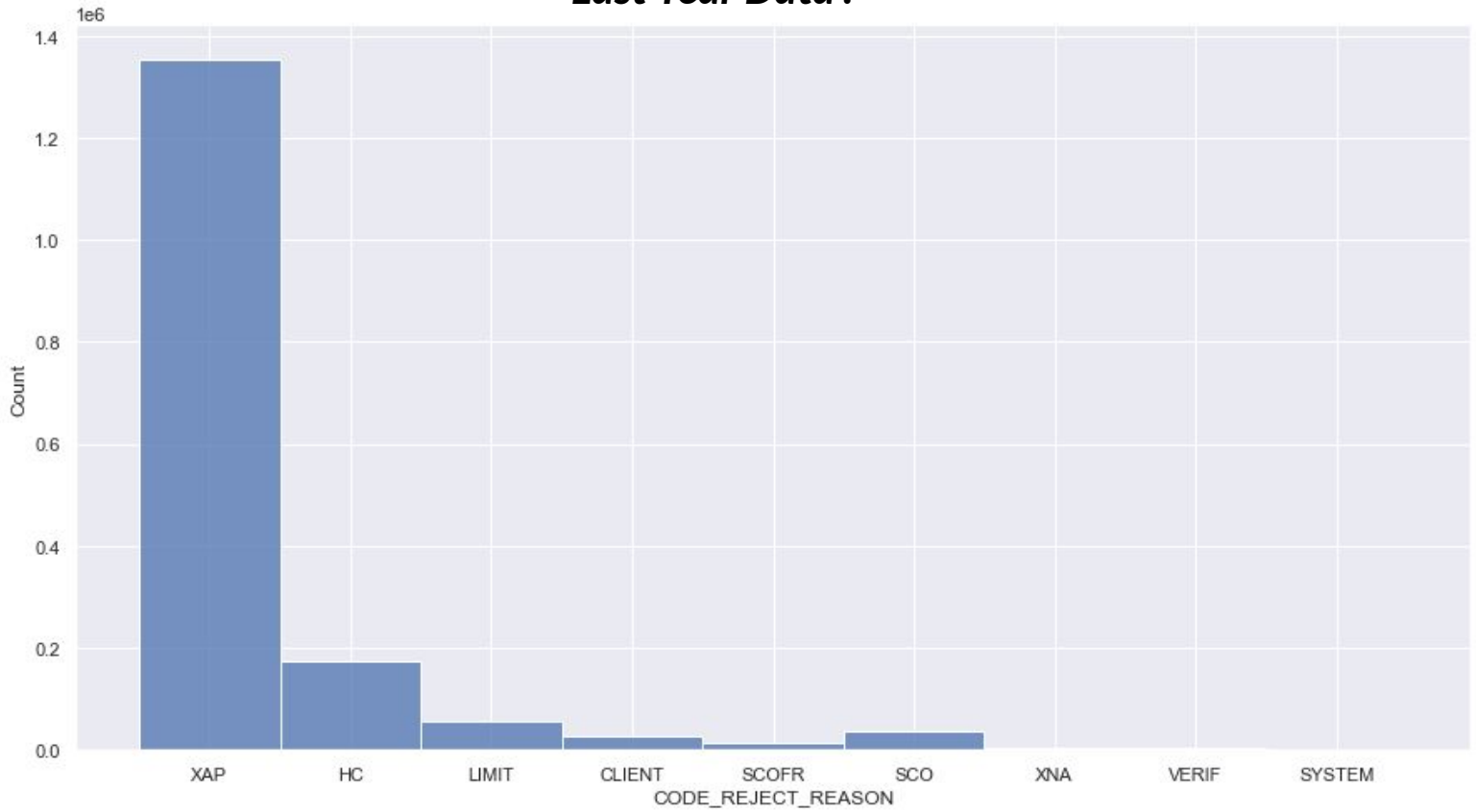
Amount for which loan was applied

Amount for which loan was actually taken



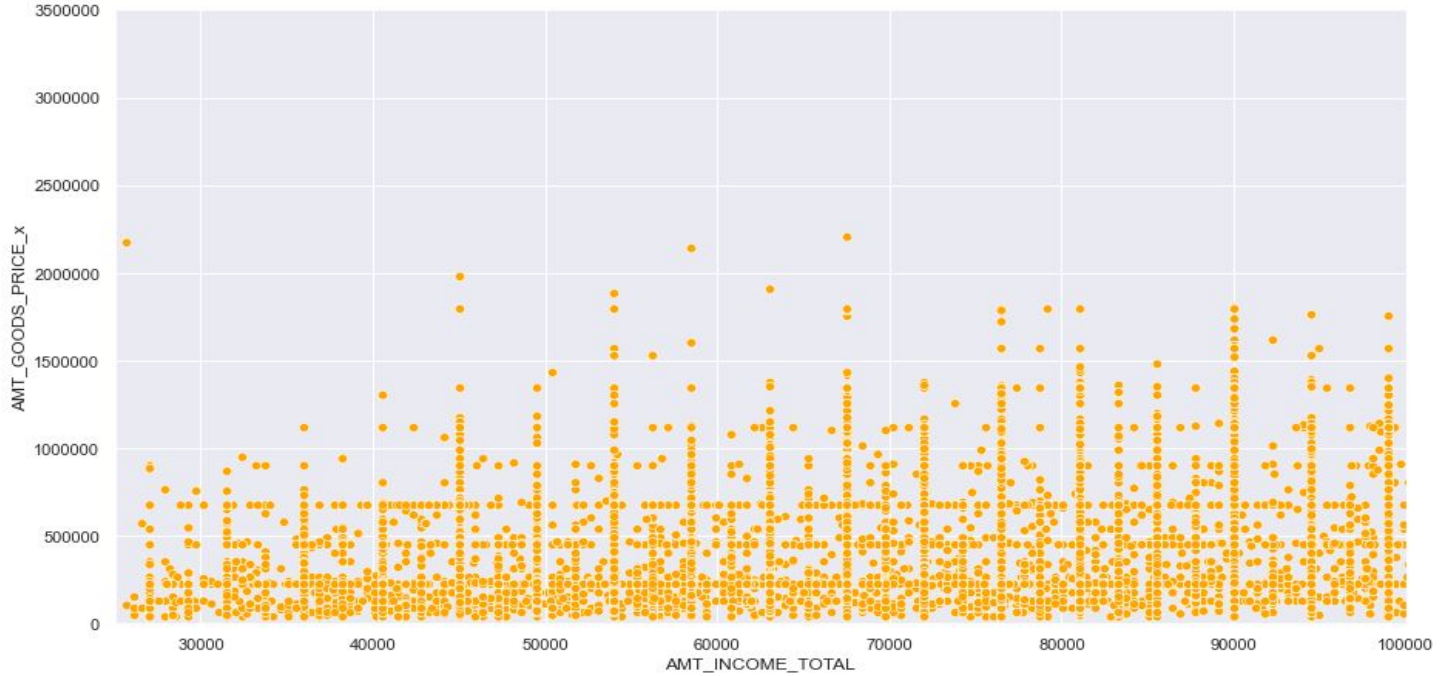
Application Reject Reason

Last Year Data :



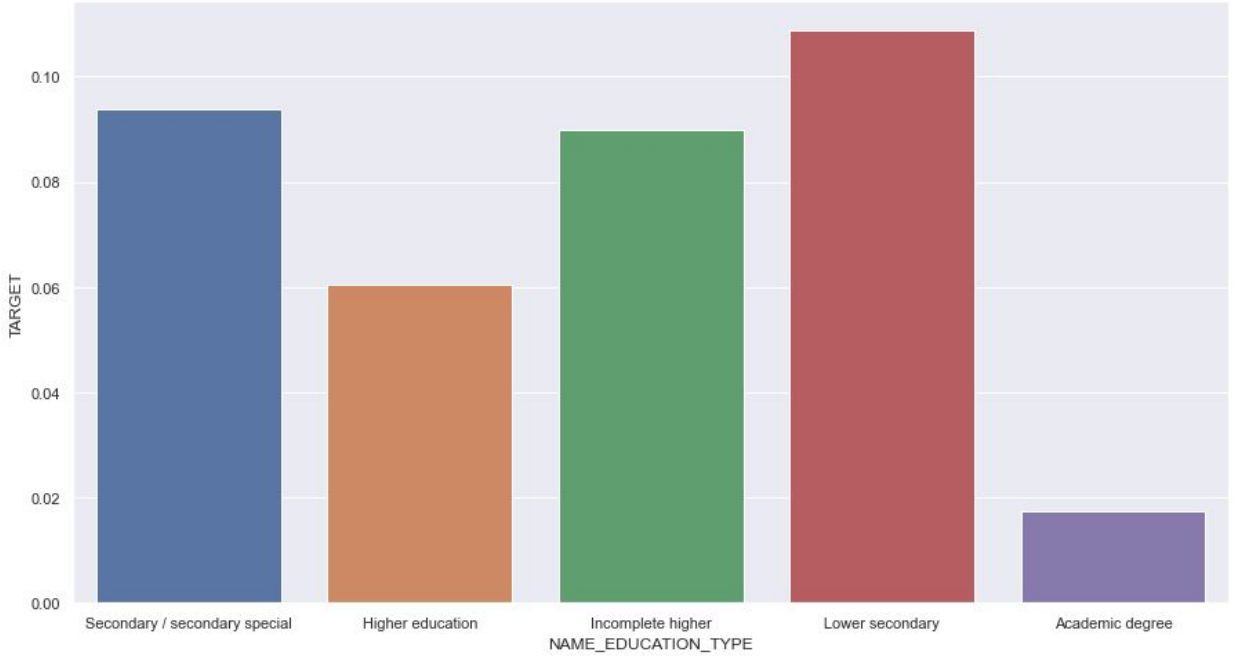
Bivariate Analysis

Amt_Goods_Price_X -vs- Amt_Income_Total : We can see a trend similar to last year though we have lesser applicants this year. We can see more than one point on y-axis having a cluster of points, namely - 25000 and 55000. Certainly the mean value of goods taken on loan has increased.



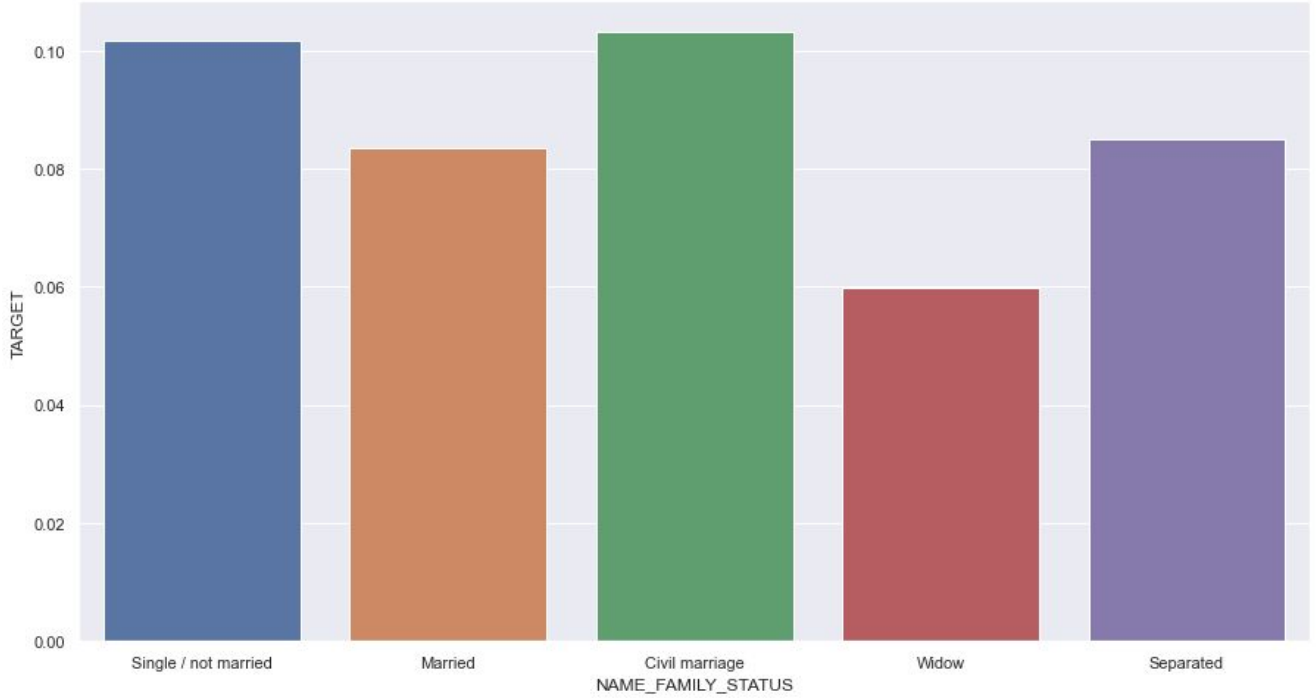
Bivariate Analysis

The following graphs are bivariate analysis of application data Features , conclusions regarding which are given at the end.



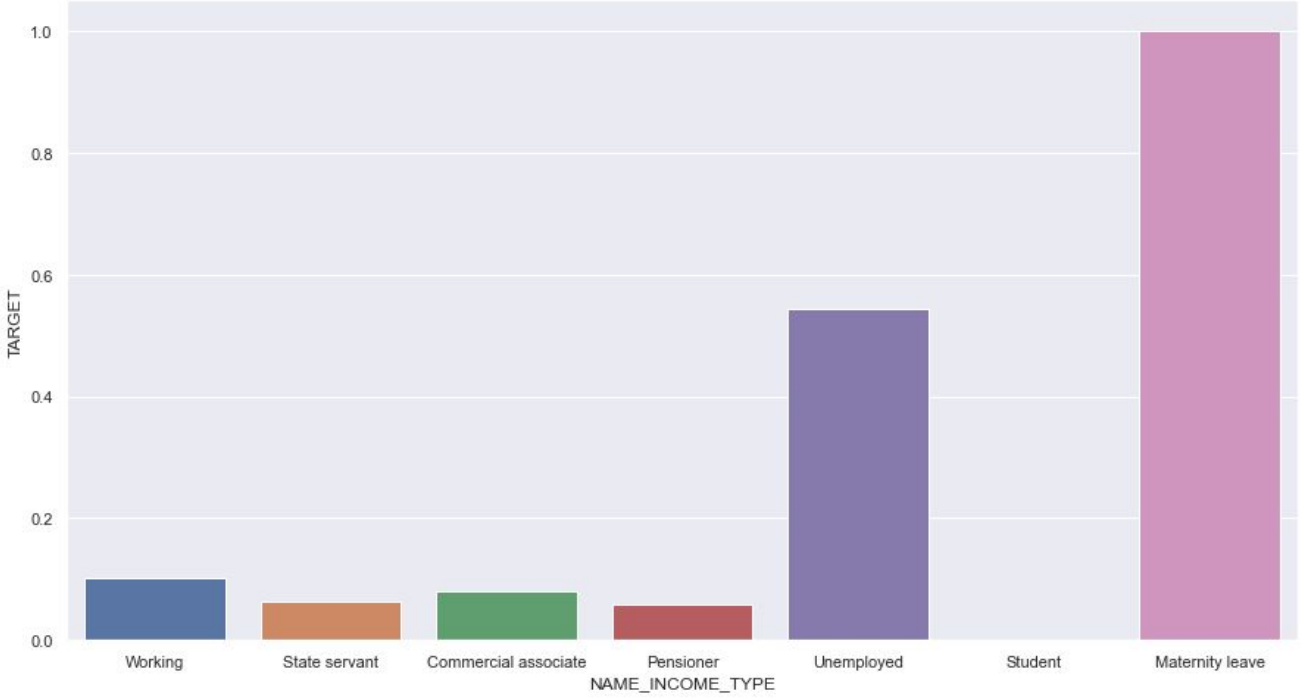
Bivariate Analysis

The following graphs are bivariate analysis of application data Features , conclusions regarding which are given at the end.



Bivariate Analysis

The following graphs are bivariate analysis of application data Features , conclusions regarding which are given at the end.



Some Conclusions from entire analysis:

- On Verifying, we see that small credits are mostly paid back. Bank should focus on leveraging smaller accounts better loaning terms and conditions.
- Unmarried people and Married people have almost equal risk of being the defaulter in 1st installment, but married people have slightly higher risk. Also we have more applications from married people.
- More people take loans in the credit range of 25000 - 50000, hence bank should modify or make their policies which suits people belonging to this credit bracket. They can consider additional parameters like married, working class etc.

Thank You.