A Project On

Bank Loan Case Study

By

Swastik Kumar Pal

A student of

Jalpaiguri Government Engineering College

Submitted to

Trainity

PROJECT DESCRIPTION

Problems Faced by the Loan Providing Companies

1. Loan providing companies find it hard to provide the loans to the applicants due to their insufficient and non existing credit history.

2. Some of the clients are using it as their advantage by become a defaulter. For this reason the companies are facing losses.

3. Due to this reasons sometime clients capable of paying the installments are getting rejected. So it also impact the business of the companies.

4. When the companies are providing the loan to the clients who are not capable of paying the loans then it may lead to the financial losses.

To find the solution of this problems and identifying the patterns of becoming a defaulter I need to use EDA to analyse from the provided data.
This analysis will help the loan providing companies to ensure that the applicants capable of repaying the loan are not getting rejected.

Goals of this analysis

1. Identifying the patterns of the clients who are not capable of repaying the loan. So the companies can deny the loan, reduce the loan amount or providing the loan amount at higher interest rate.

2. The clients who are capable of paying the loan are not getting rejected.

3. The companies also want to understand the driving factors behind loan default, which variables are the strong indicator of a loan default.

Approach to the analysis

1. At first I will do some data overview of the provided data.

2. Then I will deal with the missing data and data cleaning to start my analysis.

3. After that I will do an outlier treatment in my data.

4. Then I will perform the univariate analysis, segmented univariate analysis and the bivariate analysis to find the solution of the problem.

5. I will try to provide charts and visualizations for better understanding of the analysis and to find the results.

6. I will also describe the reasons why I came up with this solutions.

USED TECH STACKS

I used WPS office Excel and Google sheets to perform the given tasks.

Reasons behind using **WPS office Excel** and **Google Sheet**.

1. Calculation: Excel is also a powerful calculator. It can perform complex calculations, such as financial calculations or statistical analysis. It can also be used to create formulas and functions to automate repetitive calculations.

2. Data analysis: Excel also provides a range of functions and tools for analyzing data. These include charts and graphs, pivot tables, and conditional formatting. With these tools, you can quickly visualize and make sense of large amounts of data.

3. Reporting: Excel can be used to create professional-looking reports and presentations. You can format data, add graphics and charts, and create tables to summarize your findings. This can be useful for presenting data to colleagues, clients, or stakeholders

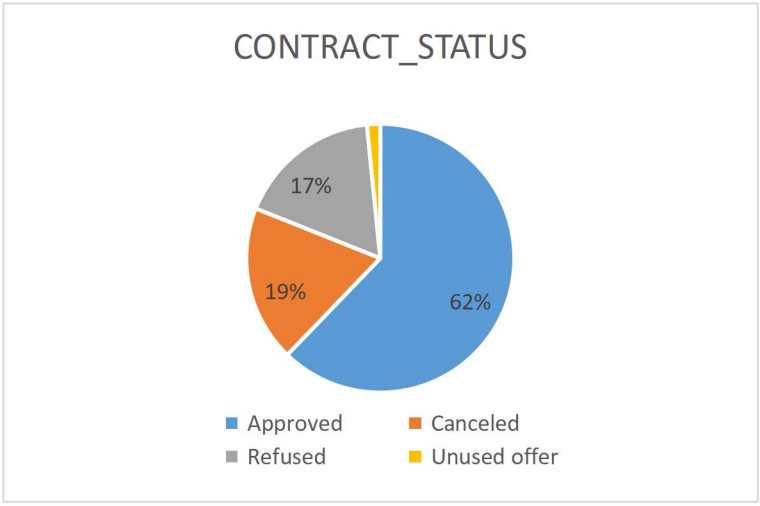
Insights

Data Overview

To analyse the clients capable of paying the loan, at first I will do some data set overview like

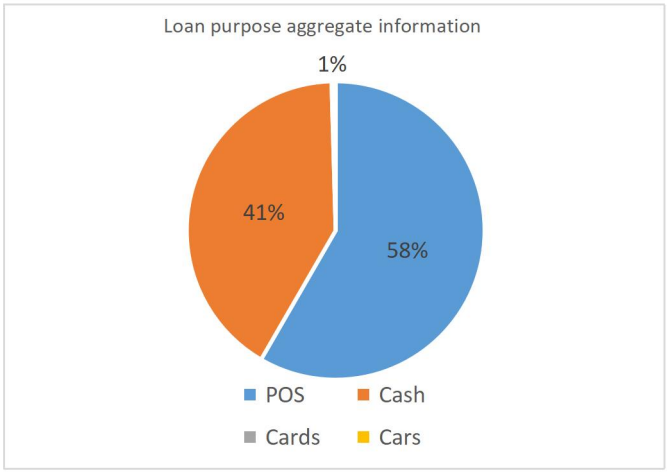
1. Contract statuses of the clients in their previous application.
2. For what reason the clients are taking the loan
3. Level of interest rates in the previous application.
4. Client type

So we will look into the **contract statuses of the clients** in their previous application

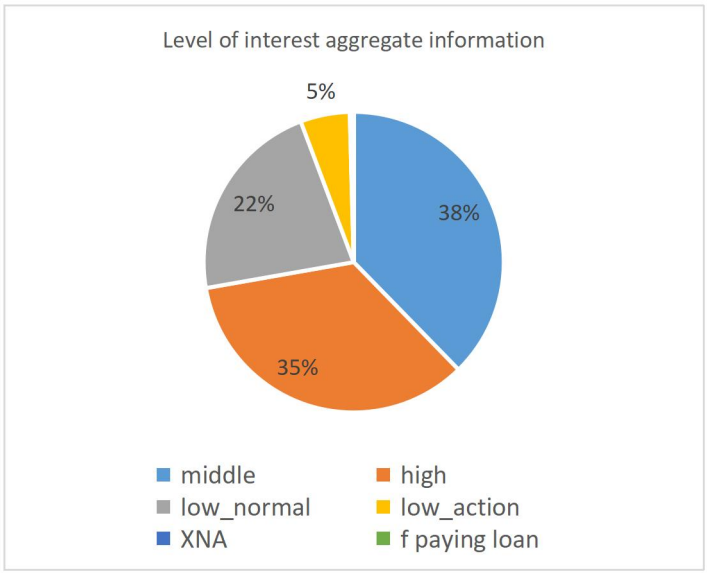


Here 62% of the clients got approval for their loan from the bank. So we will be only more focused on the clients who got the approval in their previous application.

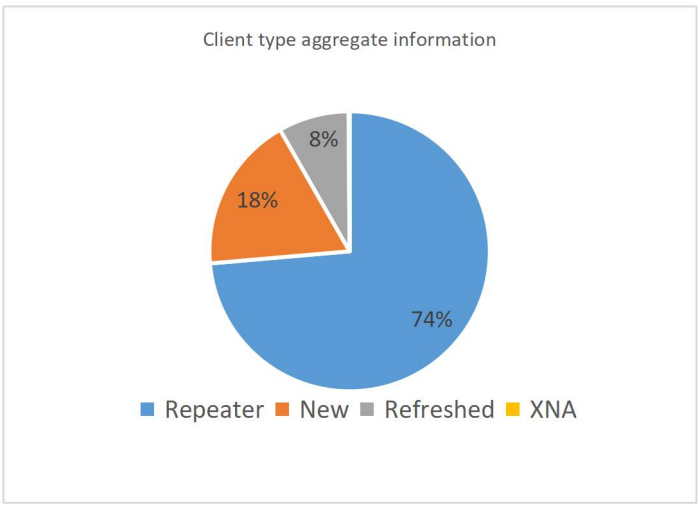
Now we will take a look for what purpose the approved clients were seeking for a loan in their previous application from the bank.



After that we will get a overview of level of interest rates of their approved previous application.



Now a aggregate overview of the client types in the previous application data set



Data Cleaning

Feature distribution:

- Loan Characteristics such as loan amount, annuity, credit amount, target, purpose, interest rate which shows the information about the loan that will help us finding loan defaulter pattern.
- Demographic variables such as age, relationship status which shows the information about the borrower profile which is not useful for us.
- Behavioral variables such as next payment date, EMI which shows the information which is updated after providing the loan which in our case is not useful as we need to decide whether we should approve the loan or not by default analysis.

So, in the data cleaning process we will remove all the **demographic variables** and the columns which has **blank values greater than 40%**.

I removed this columns from THE data set CURRENT APPLICATIONS to perform my analysis:-

EXT_SOURCE_2
EXT_SOURCE_3
APARTMENTS_AVG
BASEMENTAREA_AVG
YEARS_BEGINEXPLUATATION_AVG
YEARS_BUILD_AVG
COMMONAREA_AVG
ELEVATORS_AVG
ENTRANCES_AVG
FLOORSMAX_AVG
FLOORSMIN_AVG
LANDAREA_AVG
LIVINGAPARTMENTS_AVG
LIVINGAREA_AVG
NONLIVINGAPARTMENTS_AVG
NONLIVINGAREA_AVG
APARTMENTS_MODE
BASEMENTAREA_MODE
YEARS_BEGINEXPLUATATION_MODE
YEARS_BUILD_MODE
COMMONAREA_MODE
ELEVATORS_MODE
ENTRANCES_MODE
FLOORSMAX_MODE
FLOORSMIN_MODE
LANDAREA_MODE

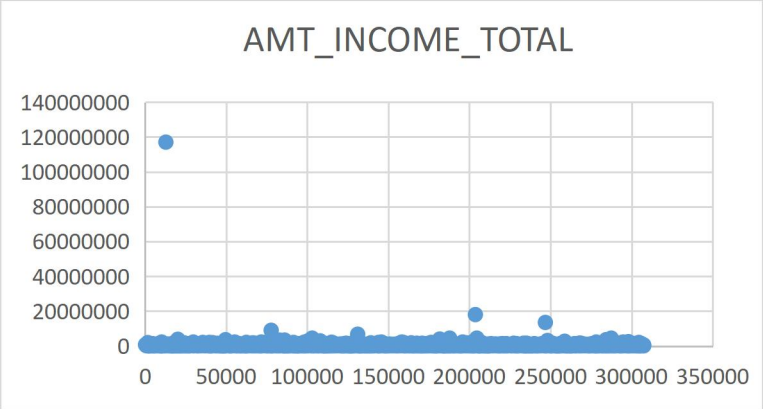
LIVINGAPARTMENTS_MODE
LIVINGAREA_MODE
NONLIVINGAPARTMENTS_MODE
NONLIVINGAREA_MODE
APARTMENTS_MEDI
BASEMENTAREA_MEDI
YEARS_BEGINEXPLUATATION_MEDI
YEARS_BUILD_MEDI
COMMONAREA_MEDI
ELEVATORS_MEDI
ENTRANCES_MEDI
FLOORSMAX_MEDI
FLOORSMIN_MEDI
LANDAREA_MEDI
LIVINGAPARTMENTS_MEDI
LIVINGAREA_MEDI
NONLIVINGAPARTMENTS_MEDI
NONLIVINGAREA_MEDI
FONDKAPREMONT_MODE
HOUSETYPE_MODE
TOTALAREA_MODE
WALLSMATERIAL_MODE
EMERGENCYSTATE_MODE

Outlier Treatment

Now we need to identify the outliers for numerical columns.
For the outlier analysis the numerical columns we will focus on

- AMT_INCOME_TOTAL
- AMT_ANNUITY
- AMT_CREDIT
- AMT_GOODS_PRICE

Outlier analysis for AMT_INCOME_TOTAL



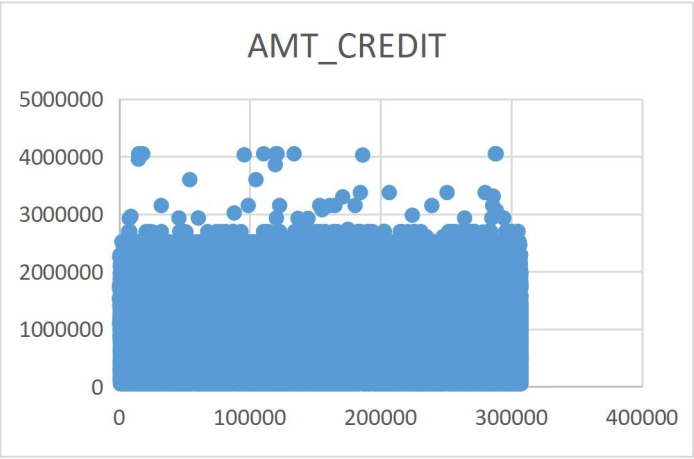
I used a scattered chart plot for the visualization of the outliers.
To treat the outliers I used the upper limit technique.
In this plotting, not so much of outliers are present.
For this reason I took quantile percentage as 90%.

The upper limit value of the AMT_INCOME_TOTAL is 270000.

Outlier analysis for AMT_CREDIT

In this plotting many outliers are present .
So I took quantile percentage as 80%.

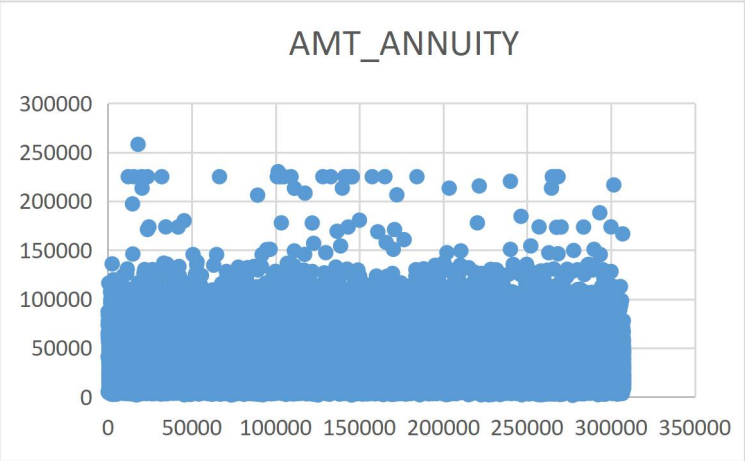
The upper limit value of the AMT_CREDIT is 900000.



Outlier analysis for AMT__ANNUITY

In this plotting many outliers are present. So I took quantile percentage as 80%.

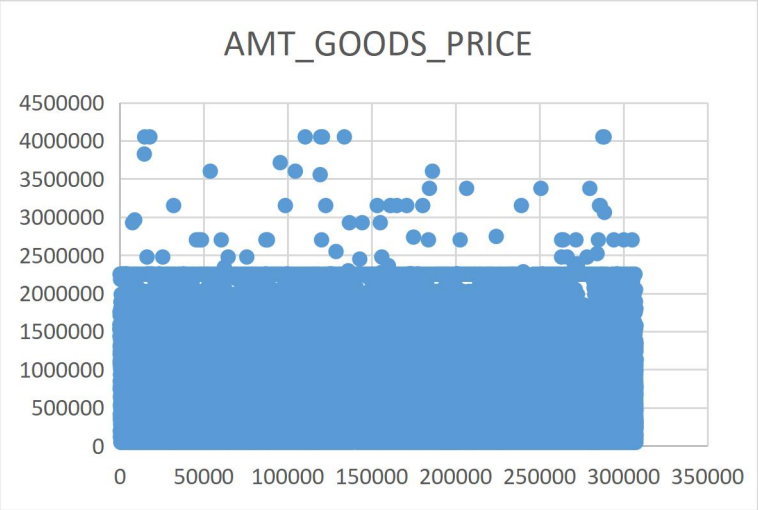
The upper limit value of the AMT_ANNUITY is 37543.



Outlier analysis for AMT_GOODS_PRICE

According to the plotting of the Outliers I took the quantile percentage as 80%.

The upper limit value of the AMT_GOODS_PRICE is 814500.



Data Imbalance

The provided data set of the clients of the current application for the loan Is imbalanced.

So total number of defaulters in data set (Target =1) is 24804 (8%) and the total number applicants whose loans got granted are 282417 (92%).

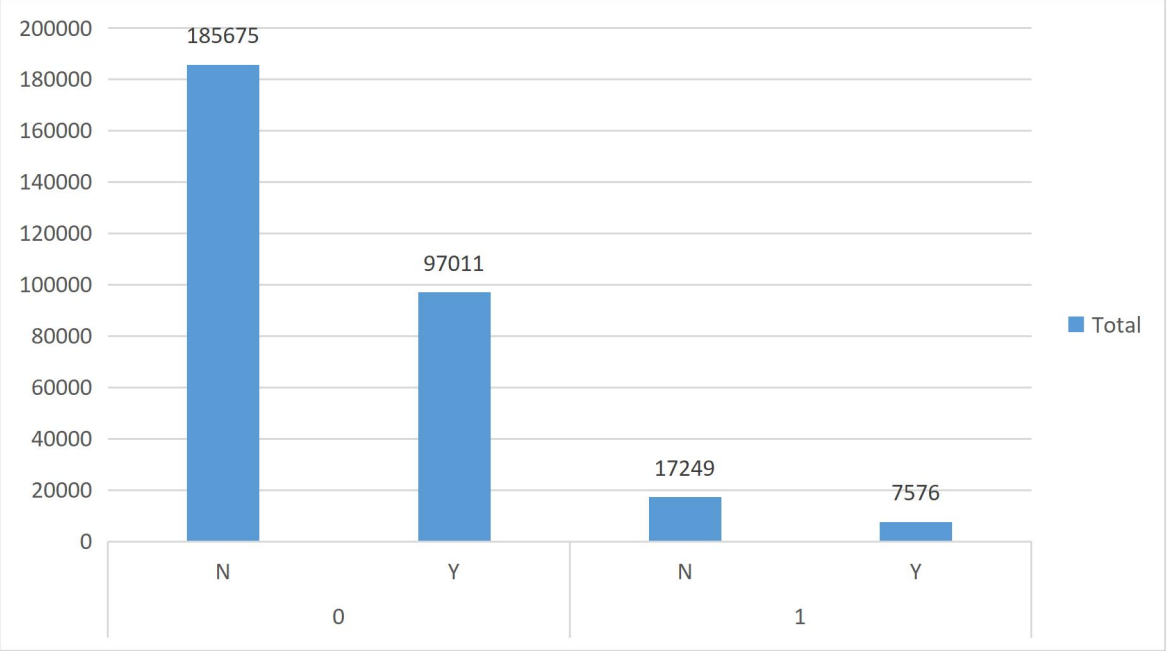
The ratio of data imbalance is 2 : 23.

Univariate analysis of Categorical variable

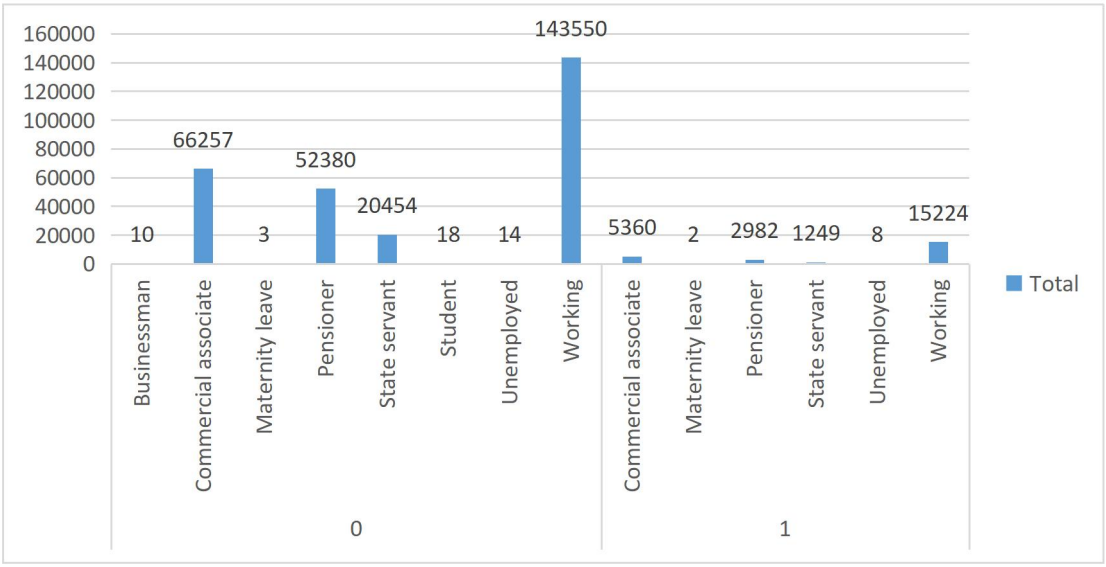
Under Univariate analysis, we will look at distribution of values of categorical variable.

NAME_CONTRACT_TYPE :

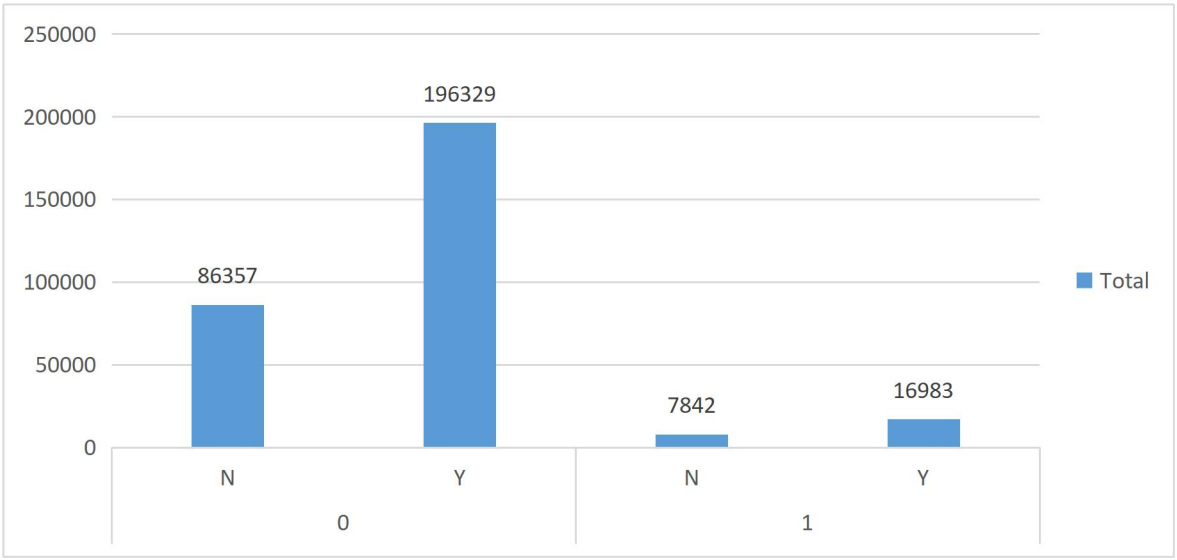
FLAG_OWN_CAR :



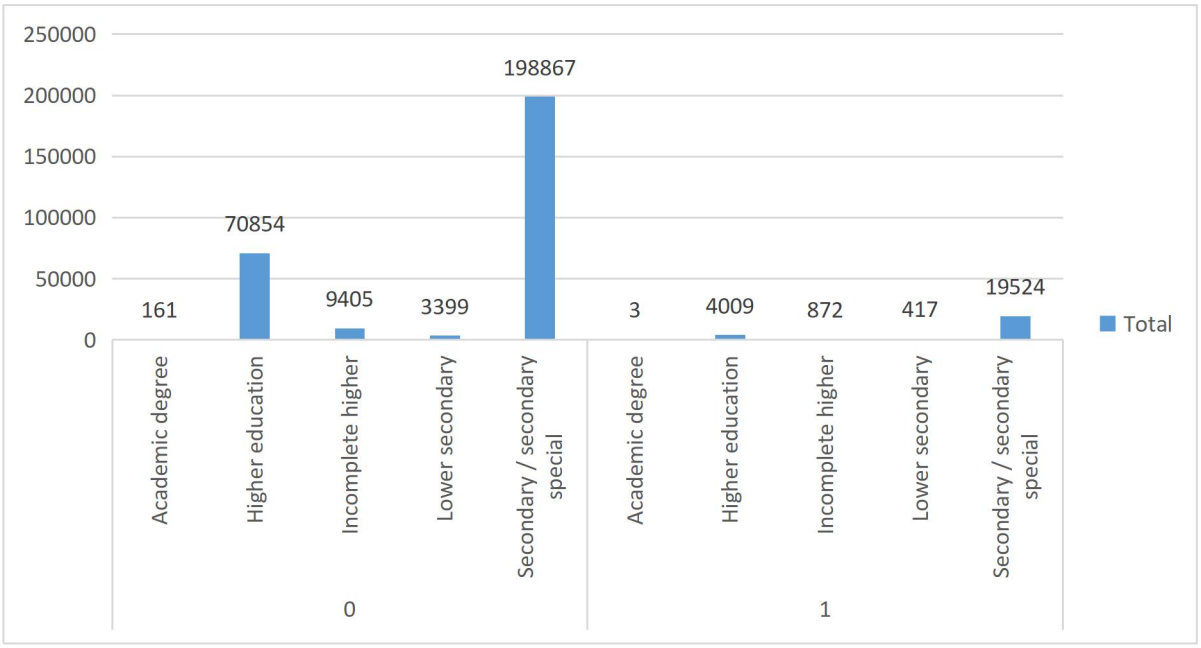
NAME_INCOME_TYPE :



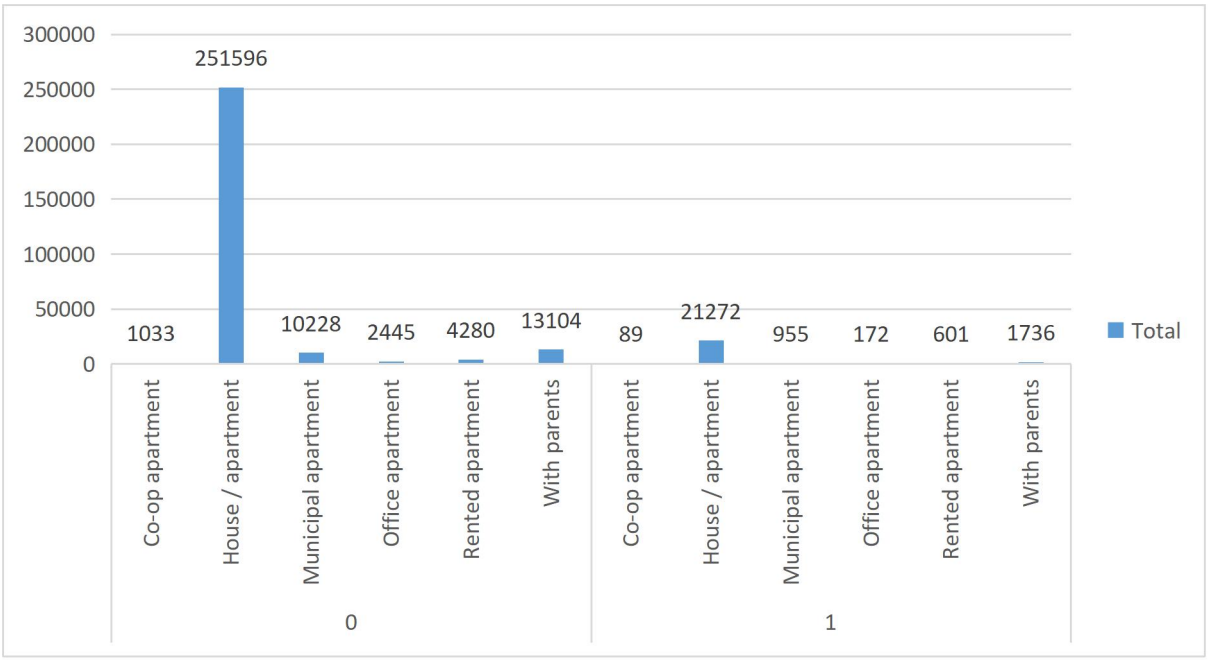
FLAG_OWN_REALTY :



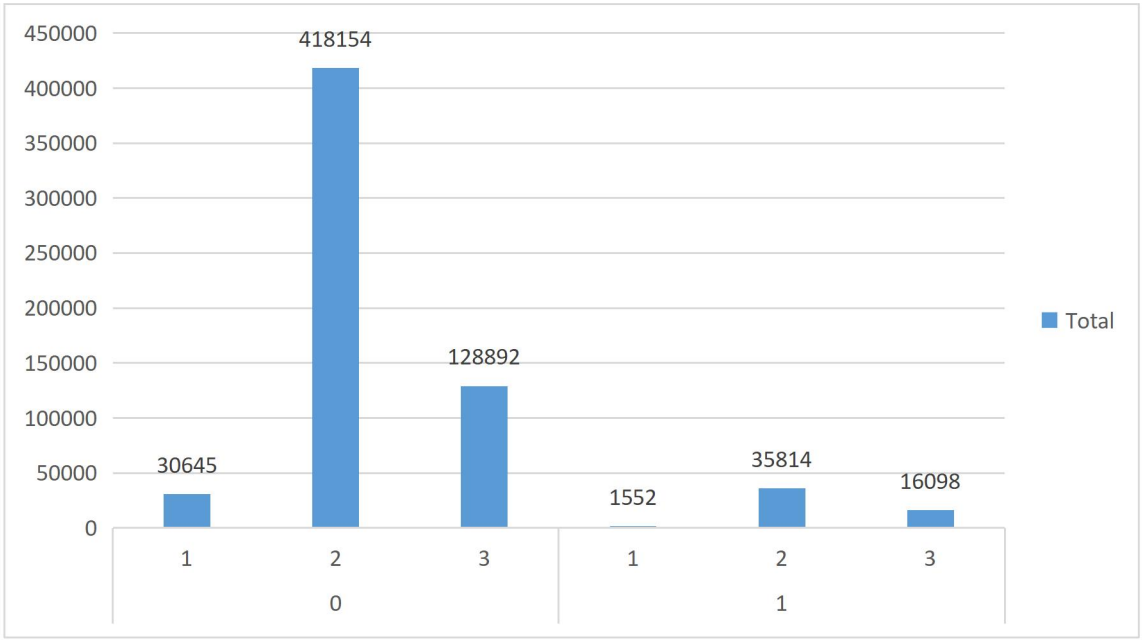
NAME_EDUCATION_TYPE:



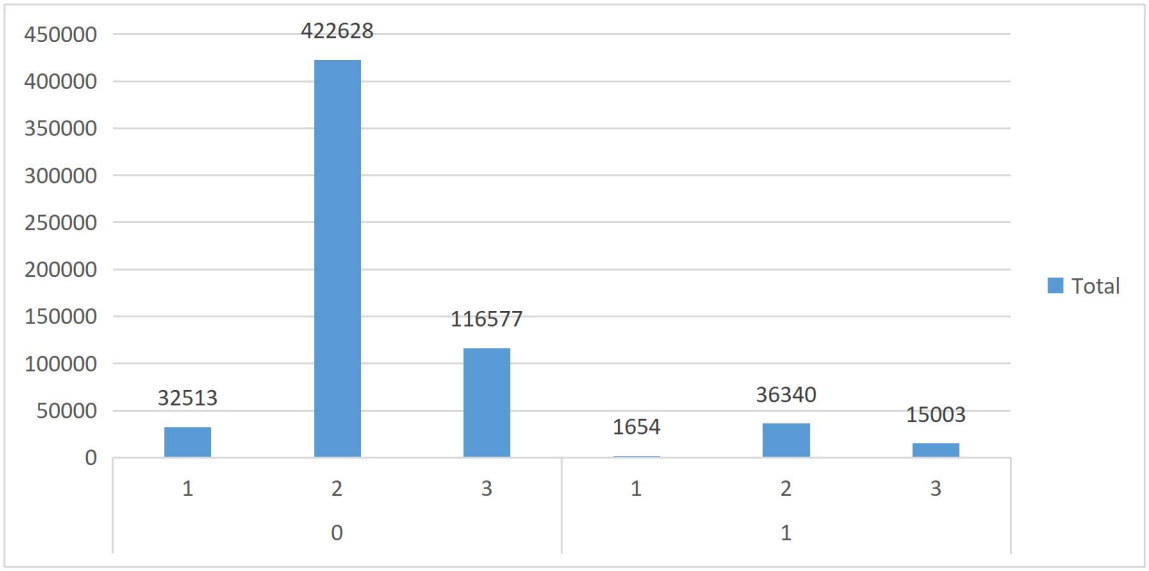
NAME_HOUSING_TYPE :



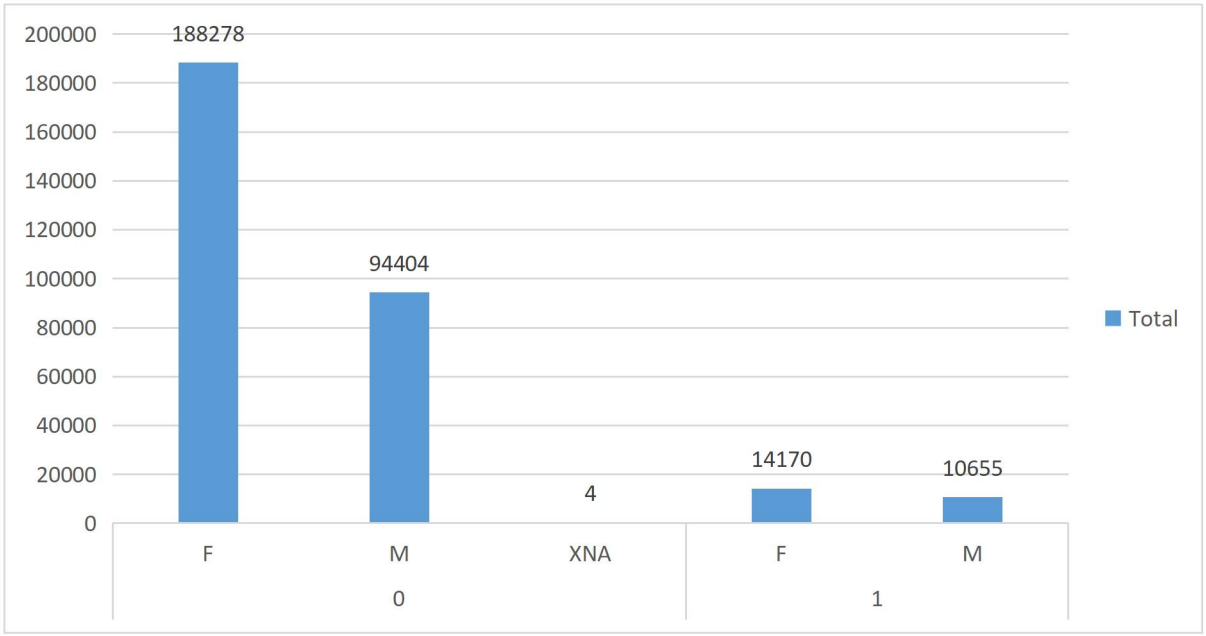
REGION_RATING_CLIENT :



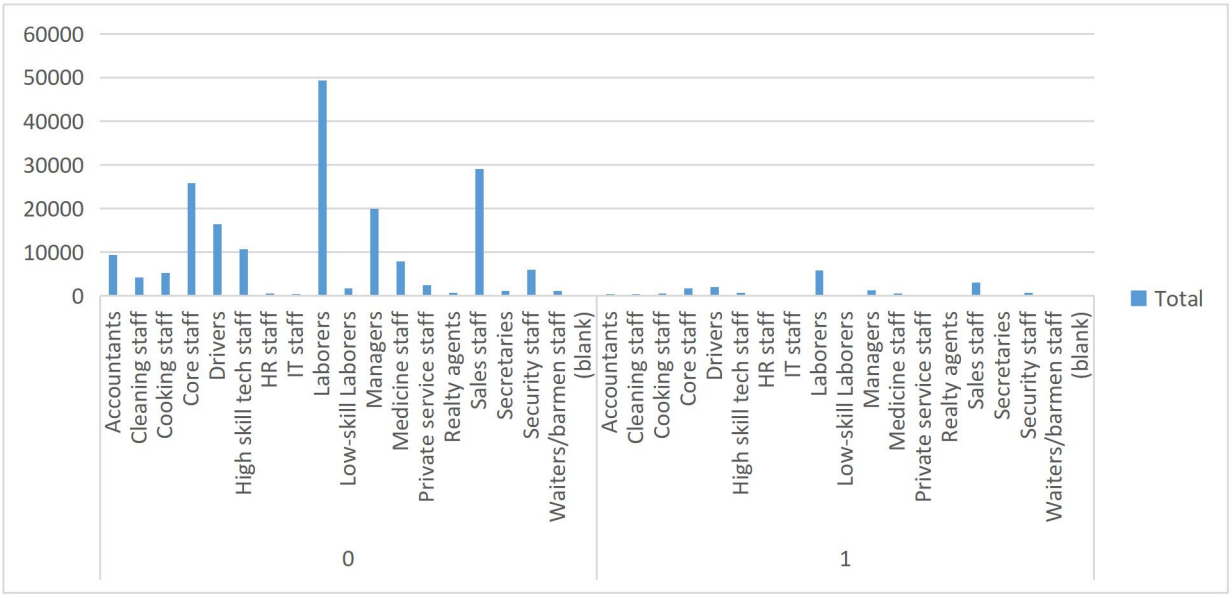
REGION_RATING_CLIENT_WITH_CITY :



CODE_GENDER



OCCUPATION_TYPE :



The above charts shows the distribution of clients across categorical variable for both target 0 & 1.

Key interpretation from Univariate analysis of categorical values.

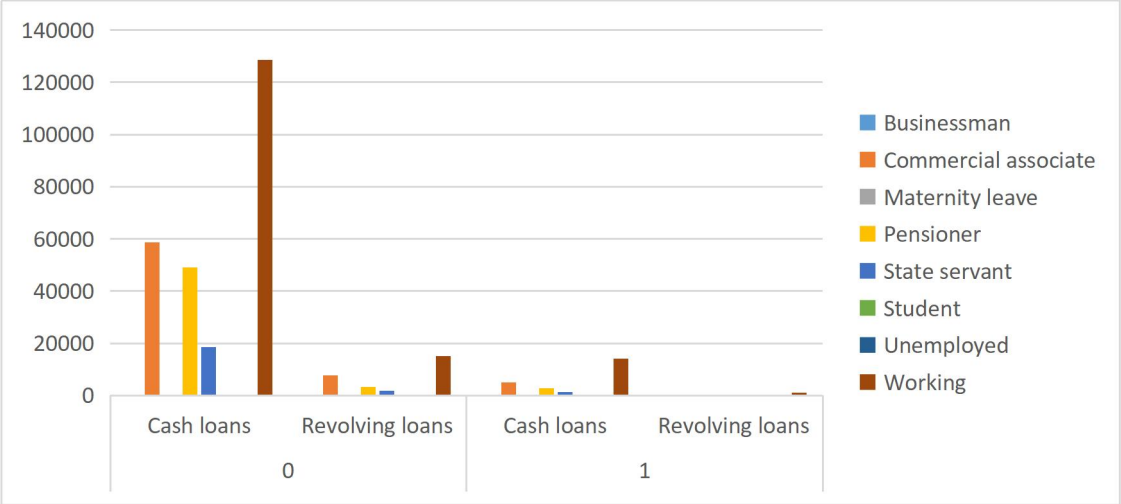
In this part, we will highlight the variables having significant difference between target=0 & target=1

- **OCCUPATION_TYPE :** Laborers has higher percentage of becoming a defaulter(Target=1).
- **NAME_HOUSING_TYPE :** Those who owns house/apartments has more contribution of becoming a defaulter (Target=1).
- **CODE_GENDER :** Defaulter has a higher percentage of female customers (Target=1).
- **NAME_INCOME_TYPE:** Defaulter has a higher percentage in working income type (Target=1).
- **NAME_CONTRACT_TYPE :** Those who took the cash loans has greater percentage of defaulters (Target=1).
- **FLAG_OWN_CAR :** Those who don't own cars has greater percentage of becoming a defaulter (Target=1).

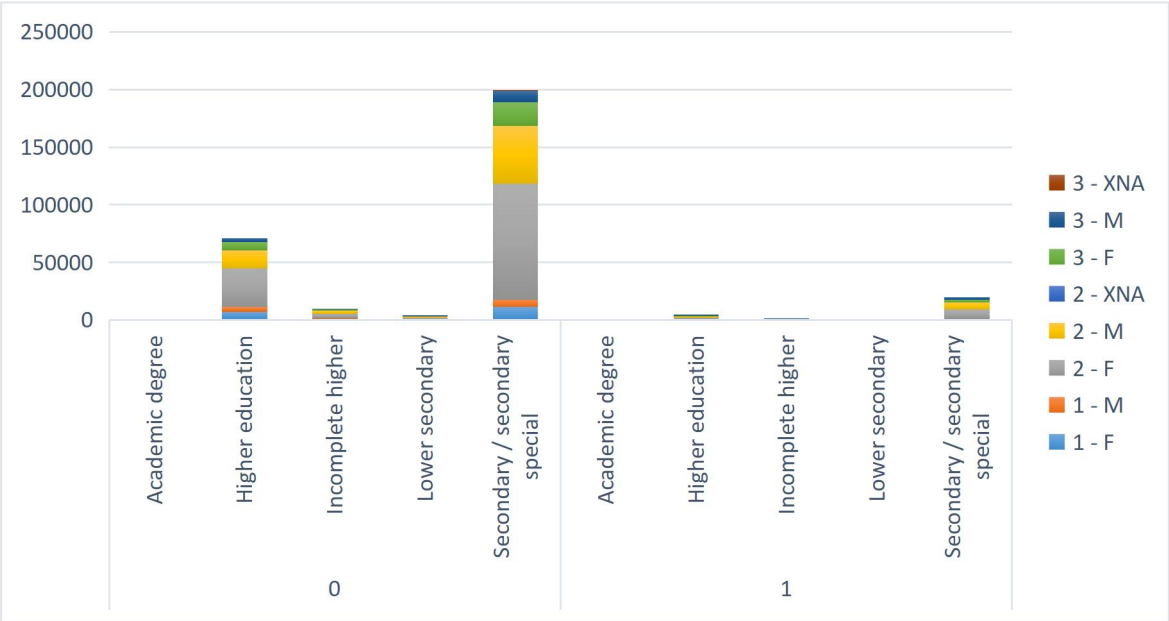
Correlation between Categorical variables by bivariate analysis

Under Bivariate analysis, we will look at distribution of values of categorical variable.

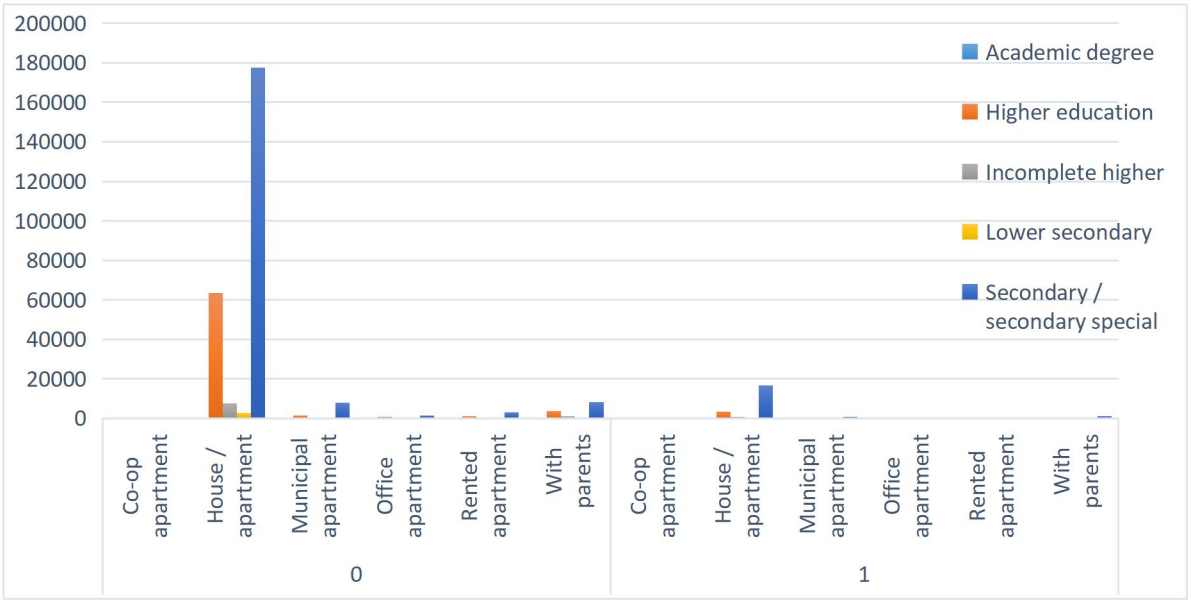
CONTRACT_TYPE & NAME_INCOME_TYPE :



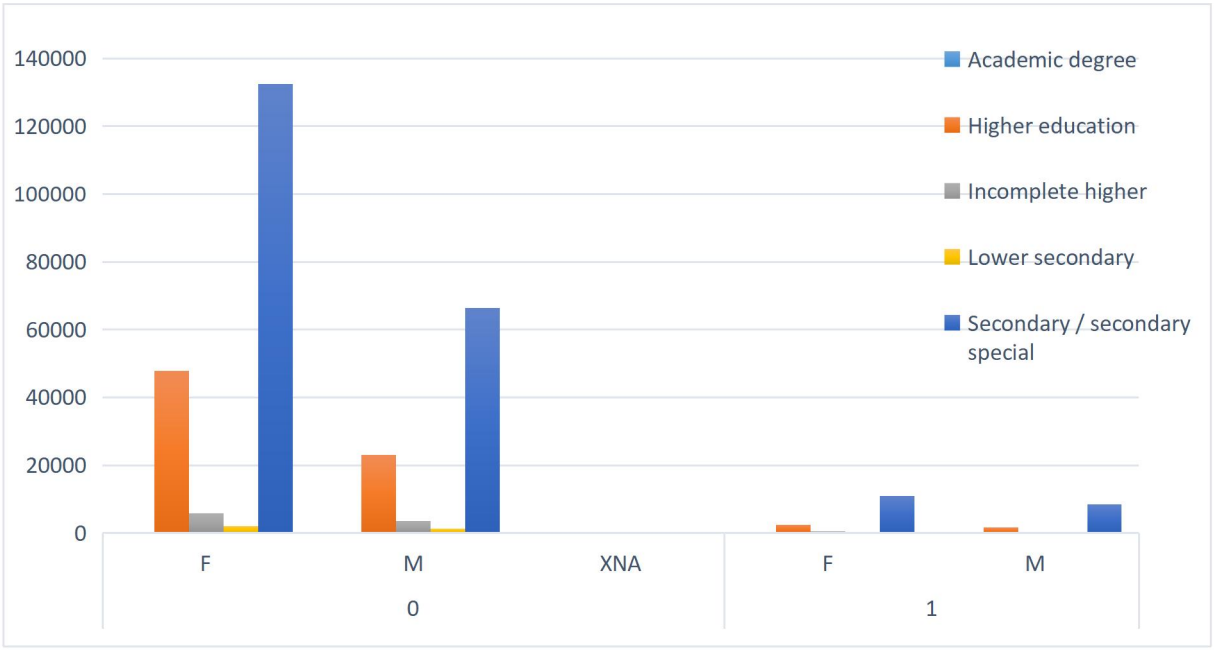
EDUCATION_TYPE , CODE_GENDER & REGION_RATING_CLIENT:



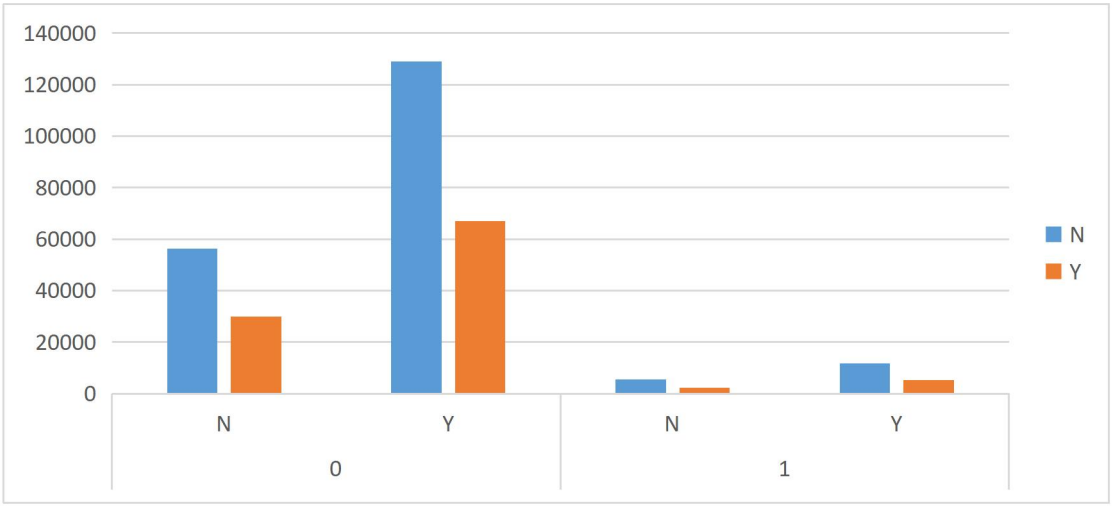
NAME_HOUSING_TYPE & NAME_EDUCATION_TYPE :



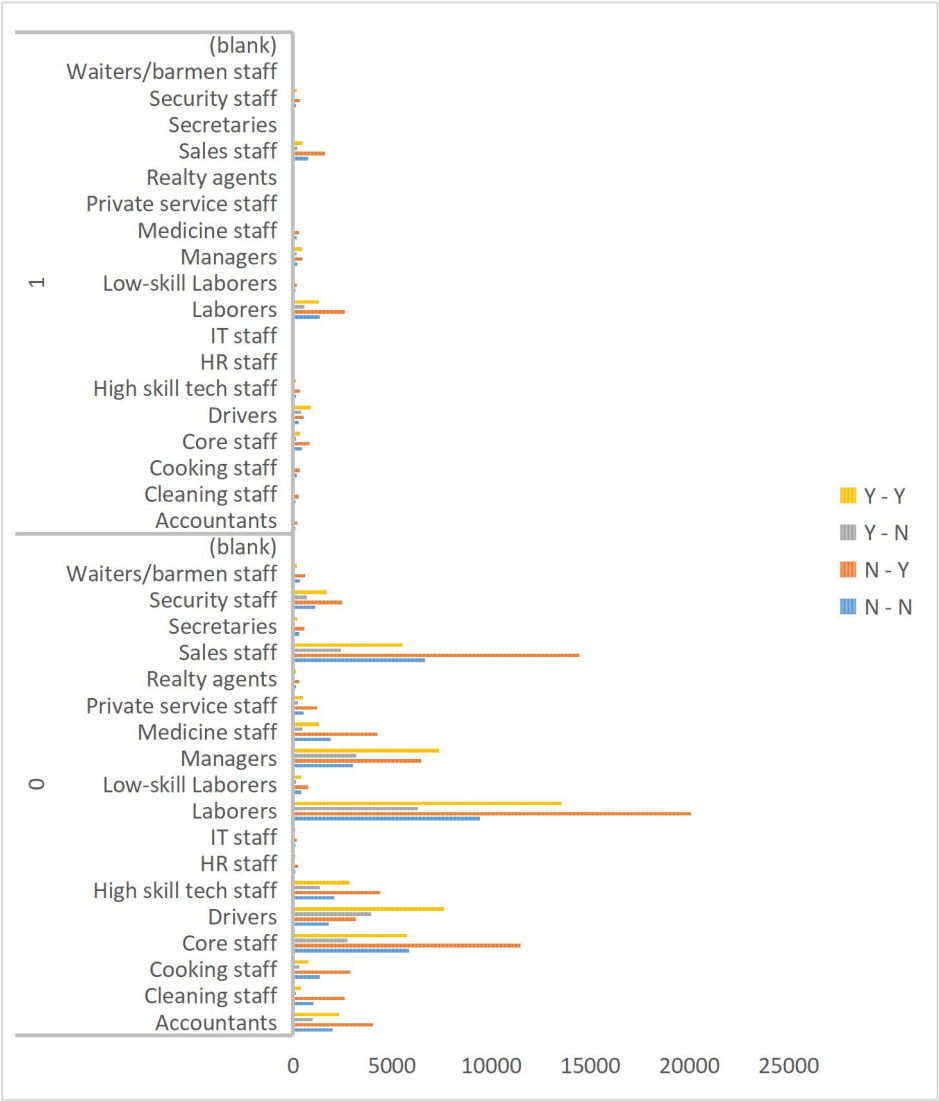
CODE_GENDER & NAME_EDUCATION_TYPE :



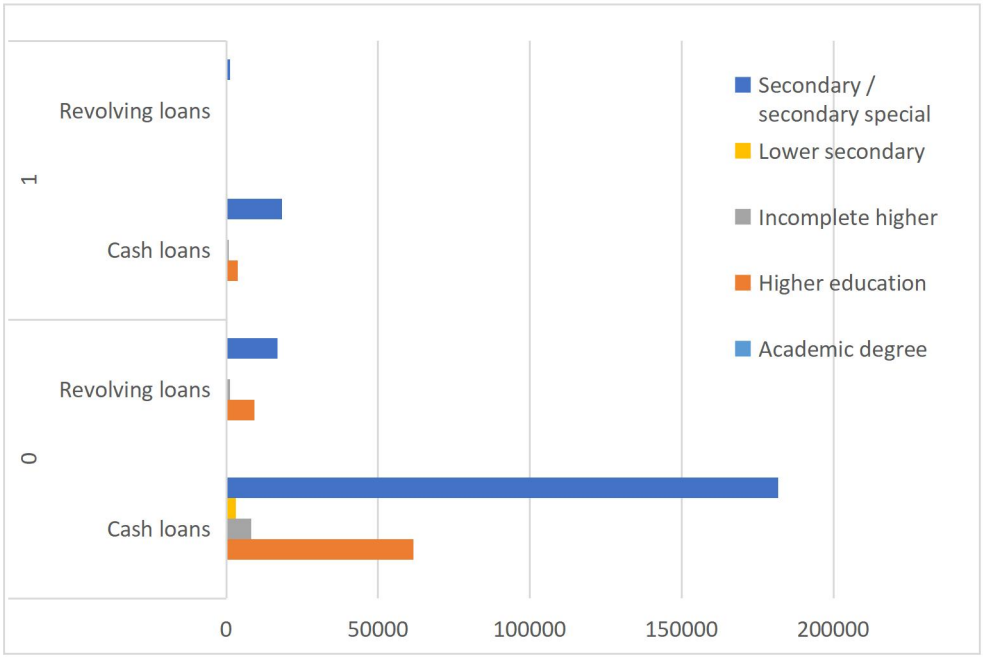
FLAG_OWN_REALTY & FLAG_OWN_CAR :



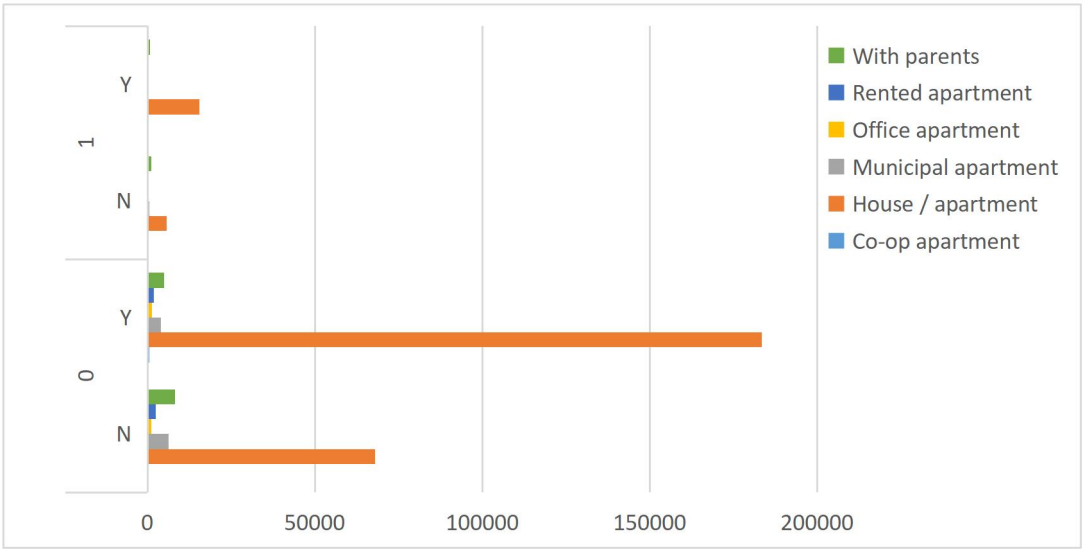
OCCUPATION_TYPE, FLAG_OWN_CAR & FLAG_OWN_REALTY



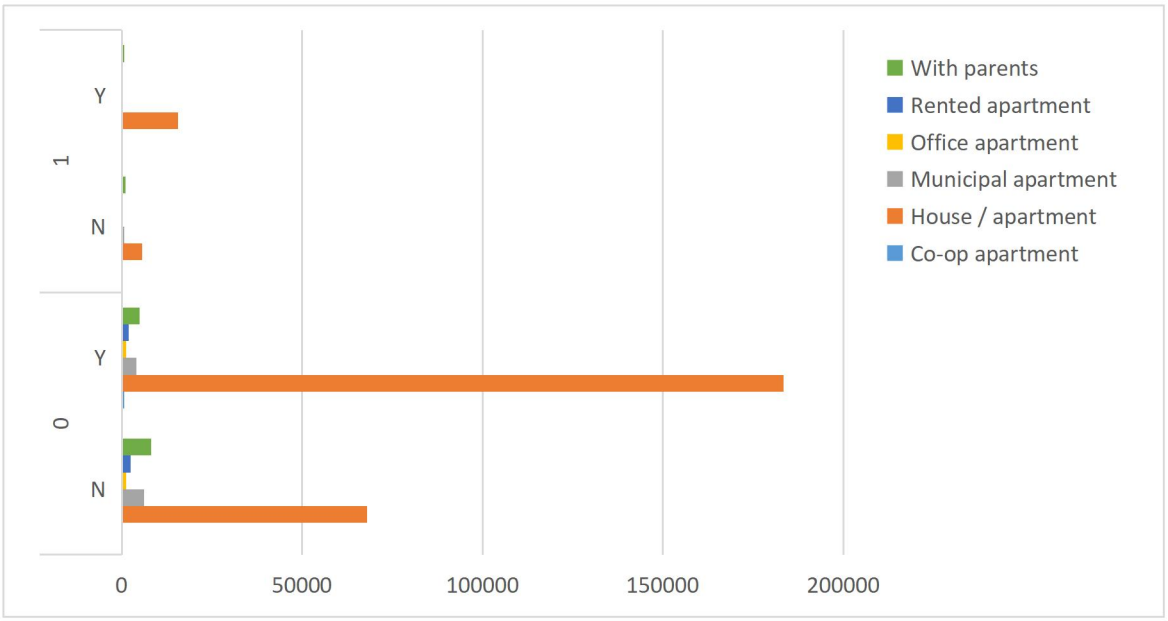
NAME_EDUCATION_TYPE & NAME_CONTRACT_TYPE :



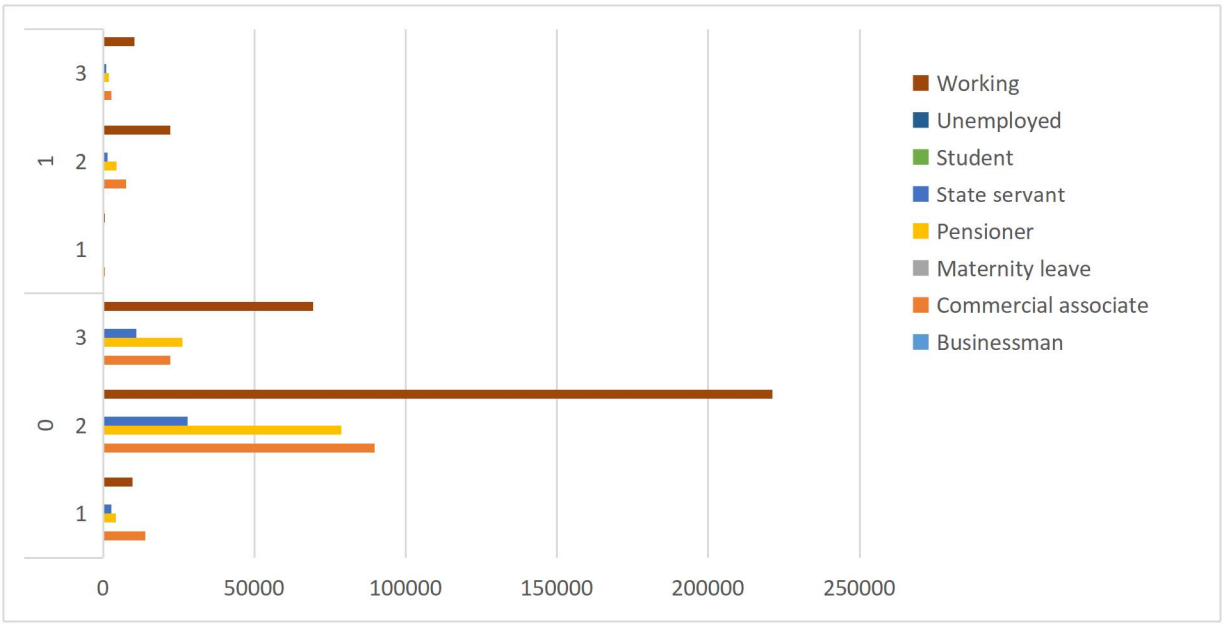
FLAG_OWN_REALTY & NAME_HOUSING_TYPE:



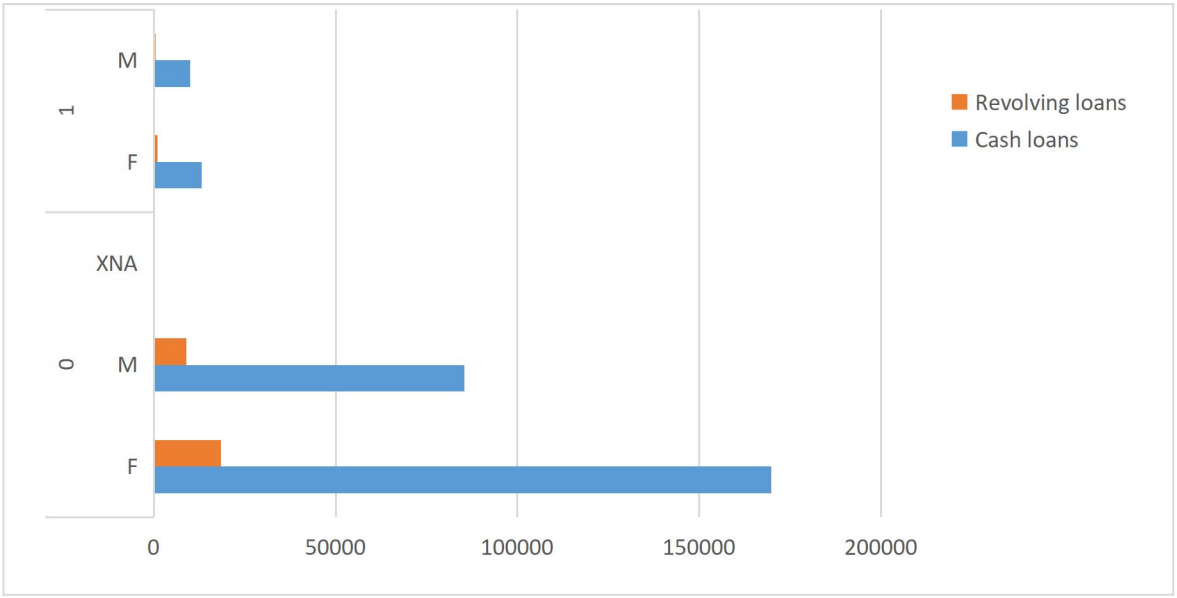
FLAG_OWN_CAR & NAME_HOUSING_TYPE:



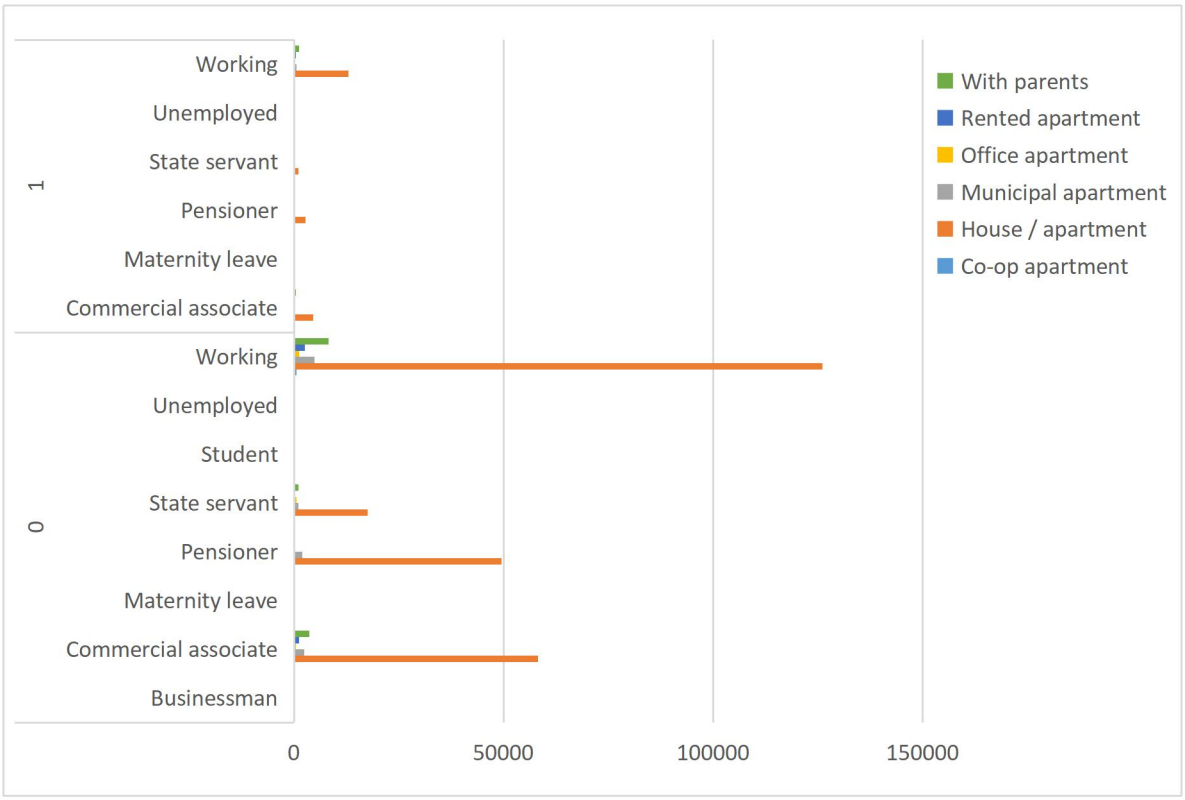
REGION_RATING_CLIENT & INCOME_TYPE:



CODE_GENDER & NAME_CONTRACT_TYPE



NAME_INCOME_TYPE & NAME_HOUSING_TYPE



The above charts shows the distribution of clients across categorical variable for both target 0 & 1 by bivariate analysis.

Key interpretation from Bivariate analysis of categorical values.

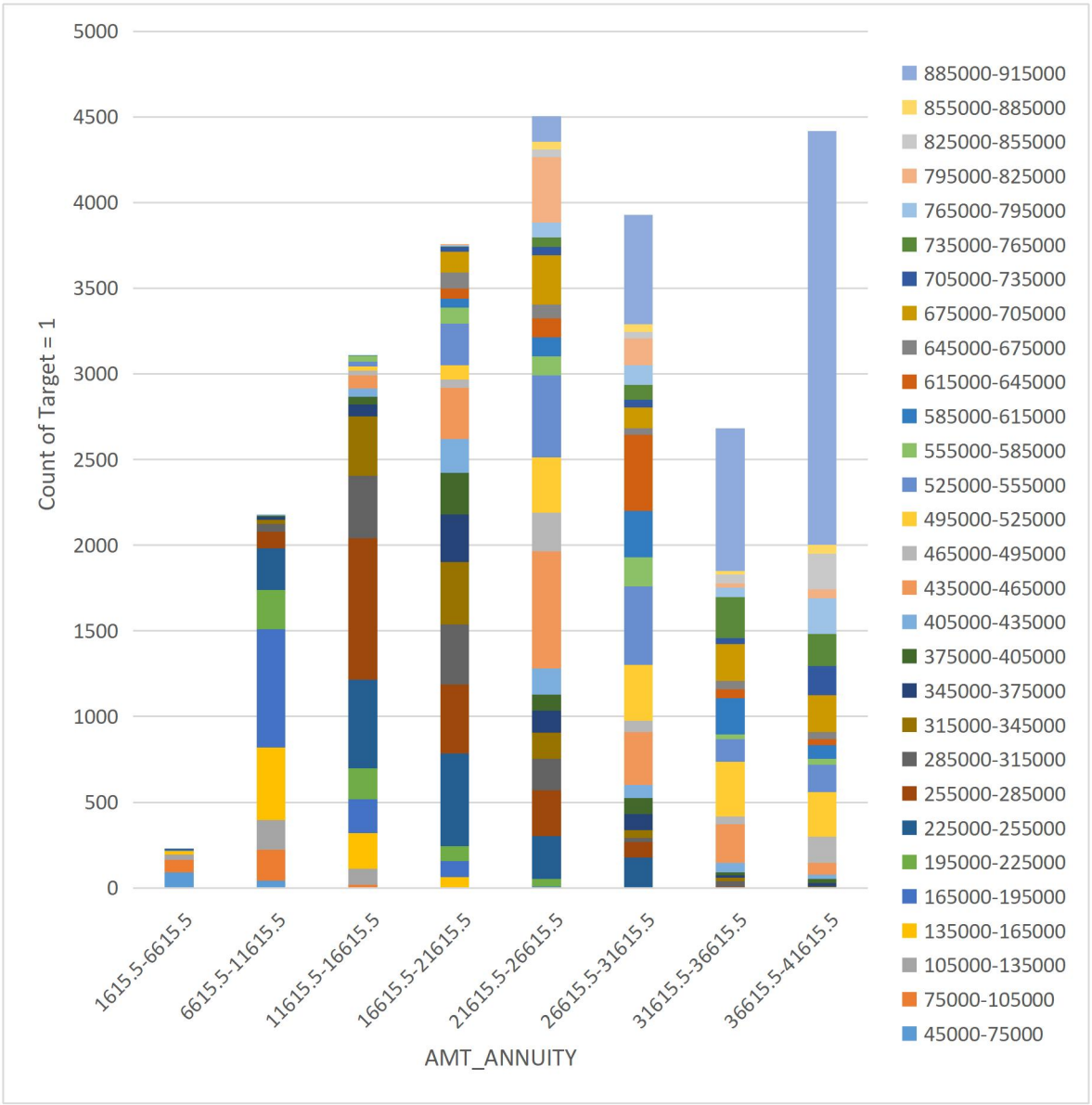
In this section, I will highlight the correlation of variables having significant difference between target=0 & target=1

- **CODE_GENDER & NAME_EDUCATION_TYPE :** Female with Secondary/secondary special education has higher percentage of becoming a defaulter.
- **NAME_INCOME_TYPE & NAME_HOUSING_TYPE :** The clients with income type of working and who owns a house/apartment has higher percentage of defaulter.
- **REGION_RATING_CLIENT & NAME_INCOME_TYPE :** The clients with rating 2 and income type of working is high percent of defaulter.
- **FLAG_OWN_REALTY & FLAG_OWN_CAR :** The clients who own realty but doesn't own car are most of them defaulters.
- **HOUSING_TYPE & EDUCATION_TYPE :** The clients with house/apartments and education type secondary has more contribution to become a defaulter.

Correlation between Numerical variables

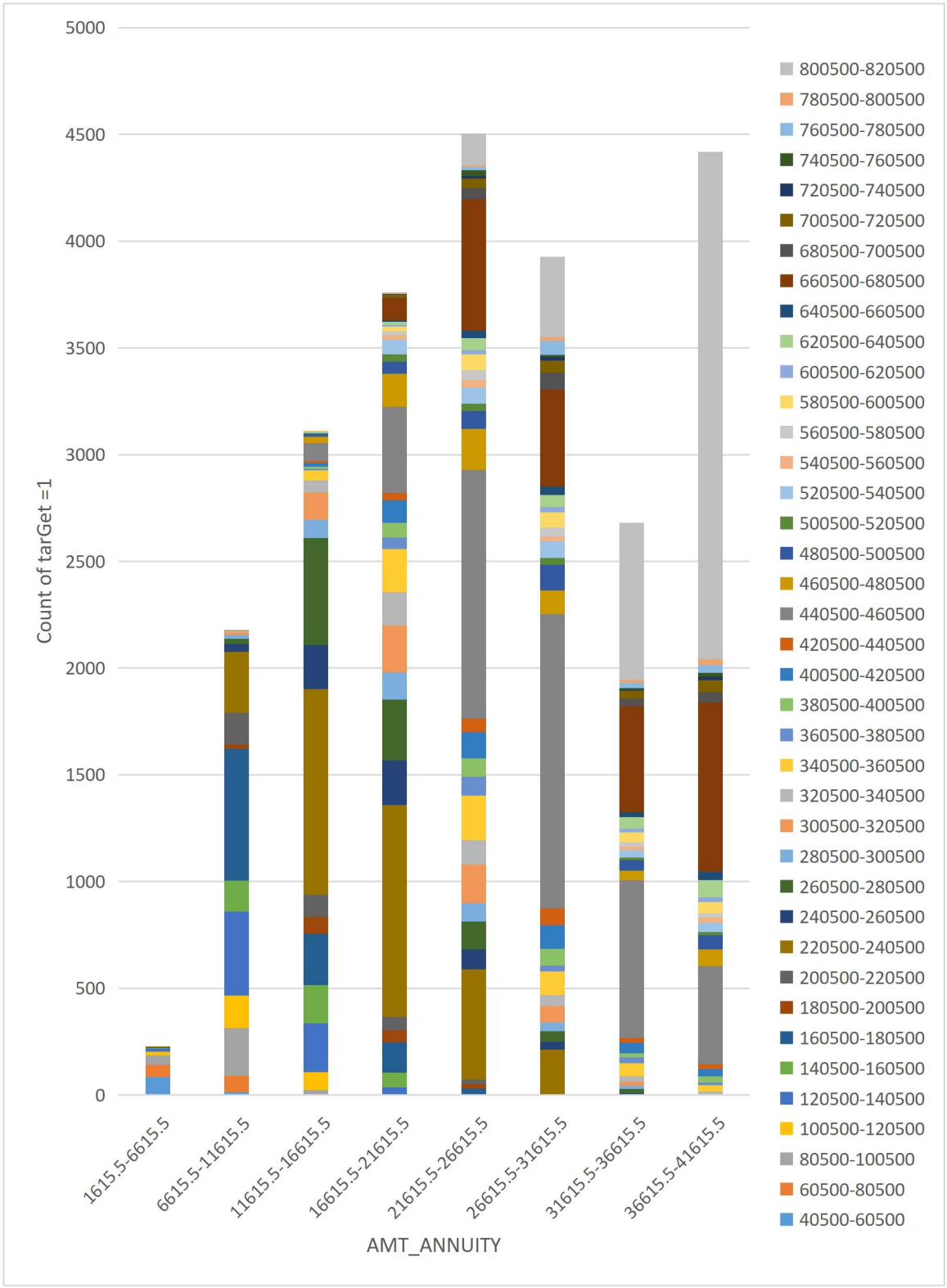
AMT_ANNUITY VS AMT_CREDIT

The below column chart represents correlation between AMT_ANNUITY & AMT_CREDIT of the defaulters Target =1.



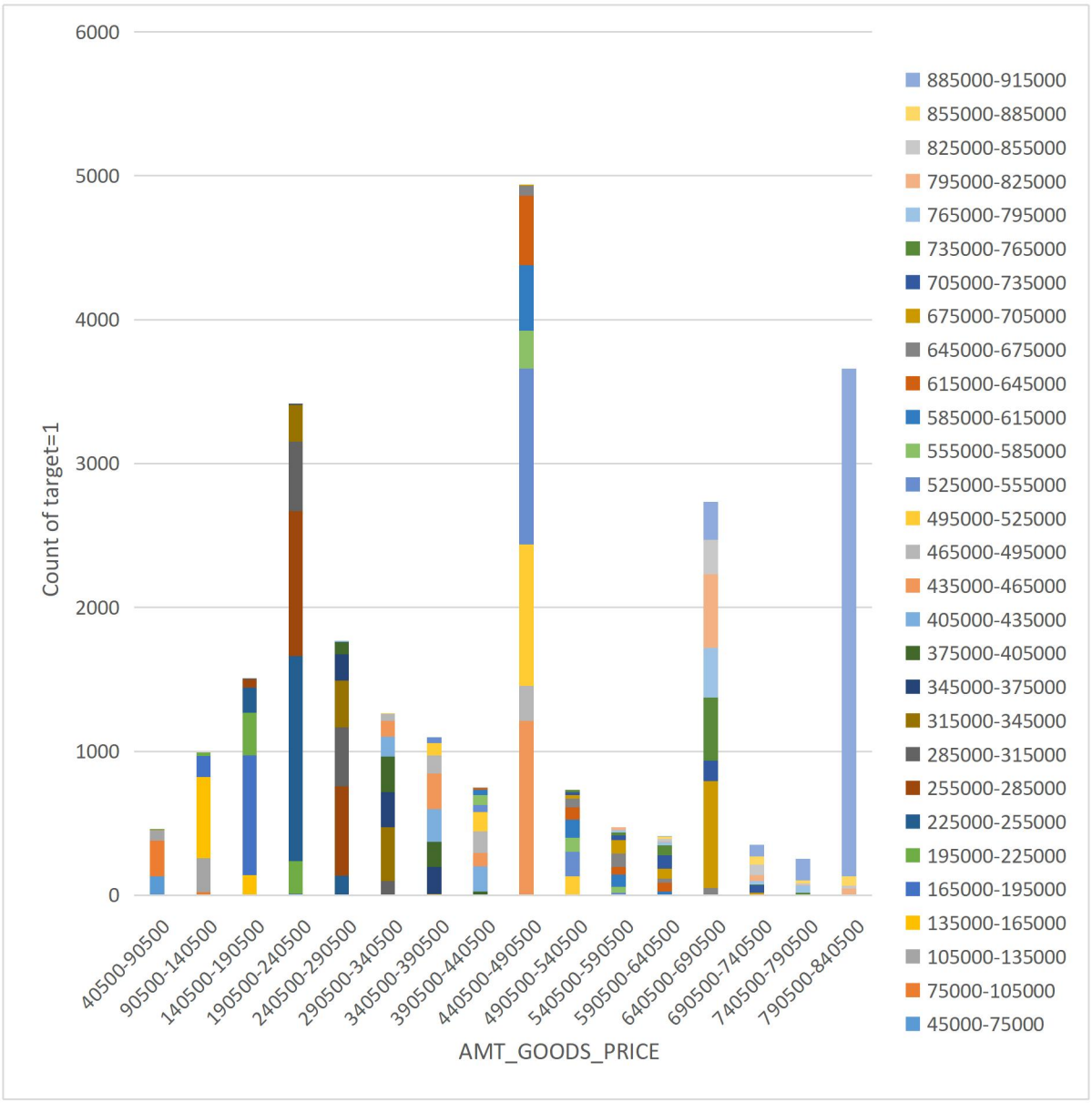
Legends represents the AMT_CREDIT.

AMT_ANNUITY VS AMT_GOODS_PRICE



Legends represent the AMT_GOODS_PRICE

AMT_GOODS_PRICE VS AMT_CREDIT



Legends represent the AMT_CREDIT

Key interpretation from Bivariate analysis of numerical values.

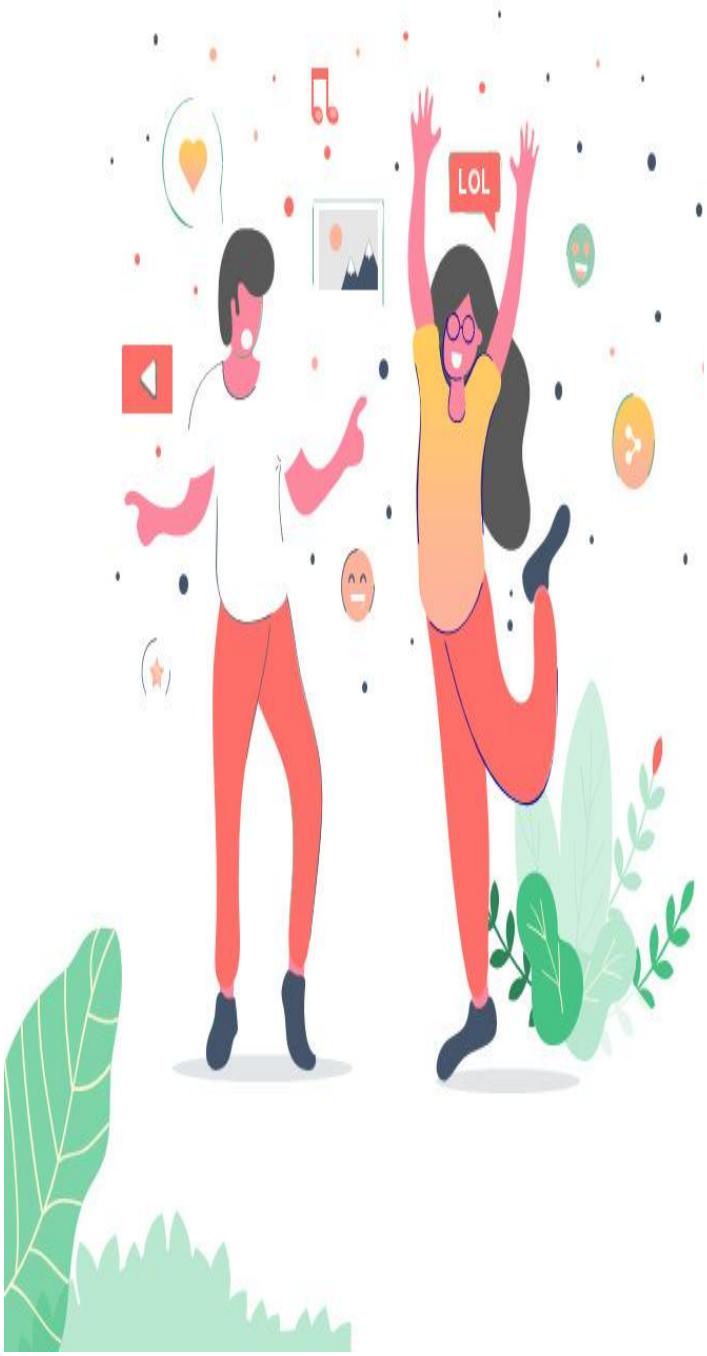
In this section, I will highlight the correlation of numerical variables by bivariate analysis having significant number of defaulters (target=1).

Ignoring the outlier values, I got this results

- **AMT_GOODS_PRICE VS AMT_CREDIT:** The clients who has amount of the goods price ranging from 790500-840500 and the credited amount ranging from 885000-915000 are the most number of defaulters.

- **AMT_ANNUITY VS AMT_GOODS_PRICE :** The clients who has amount annuity ranging from 36615-41615 and amount of the goods price ranging from 800500-820500 are the most number of defaulters.

- **AMT_ANNUITY VS AMT_CREDIT :** The clients who has amount annuity ranging from36615-41615 and amount credited ranging from 885000-915000 are the most number of defaulters.



Conclusion

This Bank Loan Case study project really helped me in different ways.

1. This project gave me more clarity about the actual industry requirements of a Excel to analyze the data and proper visualization of the large datasets.

2. To do the tasks I need to use some functions and charts/graphs to get the required insights. During that I came to know more about the advanced feature of the excel and more easy way to represent the data to the clients.

3 . It also introduced me with the more advanced features about the Excel I Used.

4.It helped me to have more clear concepts about the concepts and functions of the Excel.