EDA ANALYSIS

Getting a brief understanding of how a bank operates when it comes to giving out credit. Understand the parameters used while making a decision for Approving or Rejecting a loan application.

Applying EDA in a real business scenario…

- Develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.
- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.
- Using EDA to analyse the patterns present in the data and ensuring that the applicants capable of repaying the loan are not rejected.

Information on the dataset.

- The data frame given contains the information about the loan application at the time of applying for the loan and the application's result i.e. Approved or Rejected
- Any previous application submitted by the client and its result.
- The dataset contains two scenarios that client had payment difficulties. This includes the defaulters on the loan as well as the clients who made late payments on the loan. The other scenario is of the clients who made timely payments as well as squared off their loans.

Decisions by the client/bank

- **Approved:** The Bank has approved loan Application.
- **Cancelled:** The client cancelled the application sometime during approval.
- **Refused:** The Bank rejected the loan.
- **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

EDA Objective

- Identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan etc.
- Understanding the reasons or factors behind the defaulting on NPA's for a bank.
- Utilising the insights drawn from this study to improve risk management and lending powers of the bank.

Understanding the data

- First and foremost step in analysis is to check for missing values.
- In our data there were a number of columns(122 to be precise) and some what 40 columns had null entries present in them.
- Now these null-entries after being calculated were found to be >40% for these 40 columns.
- Hence a decision was taken to drop these columns as they were insignificant for our analysis.

2.2 CHECKING NULL VALUES PRESENT IN THE DATA

```
In [5]: null= app.isnull().sum()/len(app)*100
        null.sort_values(ascending=False)
```

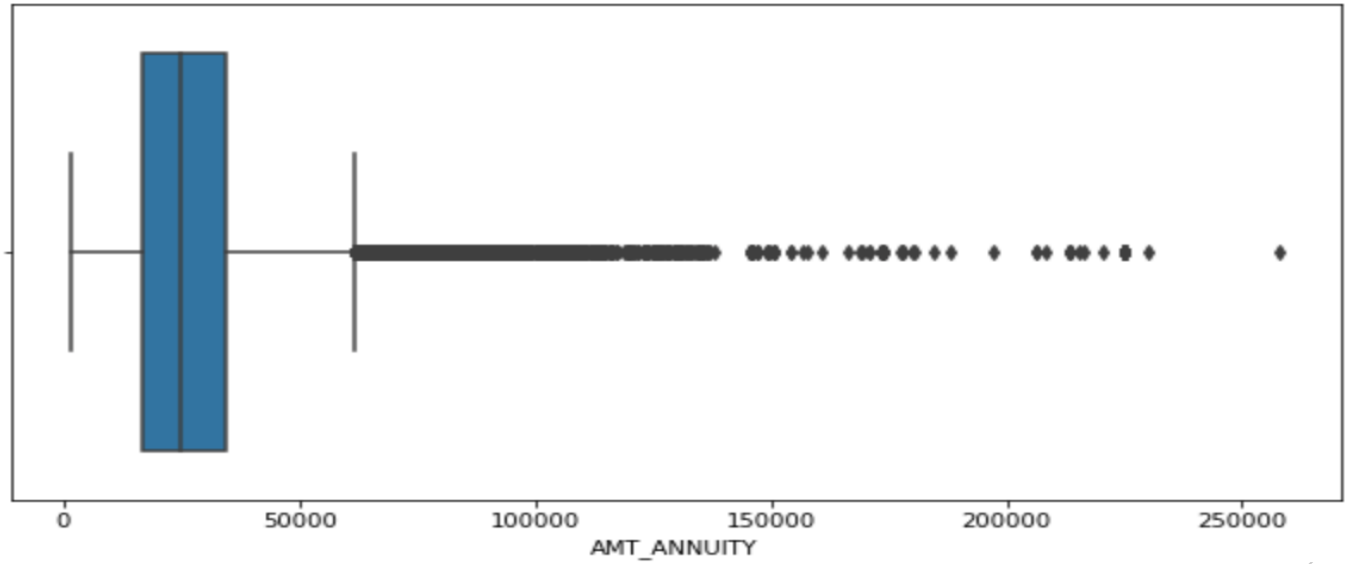
```
Out[5]: COMMONAREA_MEDI                69.872297
        COMMONAREA_AVG                 69.872297
        COMMONAREA_MODE                69.872297
        NONLIVINGAPARTMENTS_MODE       69.432963
        NONLIVINGAPARTMENTS_AVG        69.432963
        NONLIVINGAPARTMENTS_MEDI       69.432963
        FONDKAPREMONT_MODE             68.386172
        LIVINGAPARTMENTS_MODE          68.354953
        LIVINGAPARTMENTS_AVG           68.354953
        LIVINGAPARTMENTS_MEDI          68.354953
        FLOORSMIN_AVG                  67.848630
        FLOORSMIN_MODE                 67.848630
        FLOORSMIN_MEDI                 67.848630
        YEARS_BUILD_MEDI               66.497784
        YEARS_BUILD_MODE               66.497784
        YEARS_BUILD_AVG                66.497784
        OWN_CAR_AGE                    65.990810
        LANDAREA_MEDI                  59.376738
        LANDAREA_MODE                  59.376738
        LANDAREA_AVG                   59.376738
        BASEMENTAREA_MEDI              58.515956
        BASEMENTAREA_AVG               58.515956
        BASEMENTAREA_MODE              58.515956
        EXT_SOURCE_1                   56.381073
        NONLIVINGAREA_MODE             55.179164
        NONLIVINGAREA_AVG              55.179164
        NONLIVINGAREA_MEDI             55.179164
        ELEVATORS_MEDI                 53.295980
        ELEVATORS_AVG                  53.295980
        ELEVATORS_MODE                 53.295980
        WALLSMATERIAL_MODE             50.840783
        APARTMENTS_MEDI                50.749729
        APARTMENTS_AVG                 50.749729
        APARTMENTS_MODE                50.749729
        ENTRANCES_MEDI                 50.348768
        ENTRANCES_AVG                  50.348768
```

Imputation for remaining null value columns

- Imputing the values is a quick and easy way to deal with null values in a column.
- We have performed 2 types of imputation:
    1. With median for columns with outliers.
    2. With mean for columns without outliers.

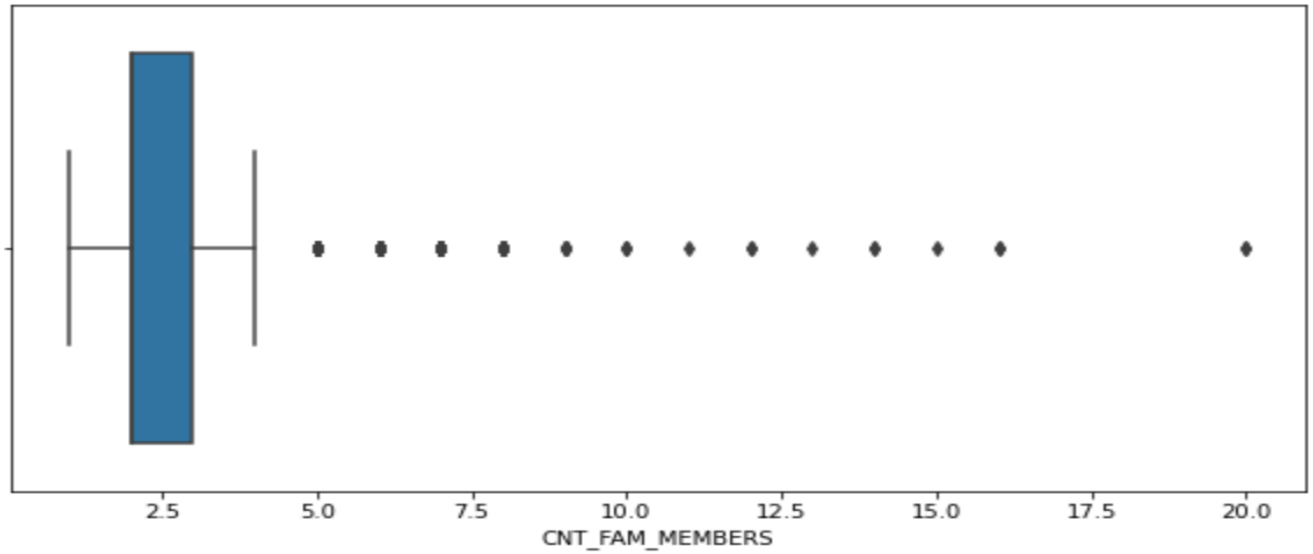
Column with outliers

- A column named AMT_ANNUITY had outliers present.
- We came to know about this using boxplot.
- Thus we imputed it with the median values as the provide a more balanced approach when dealing with null values where a number of values are so high that taking a mean would further push them out.



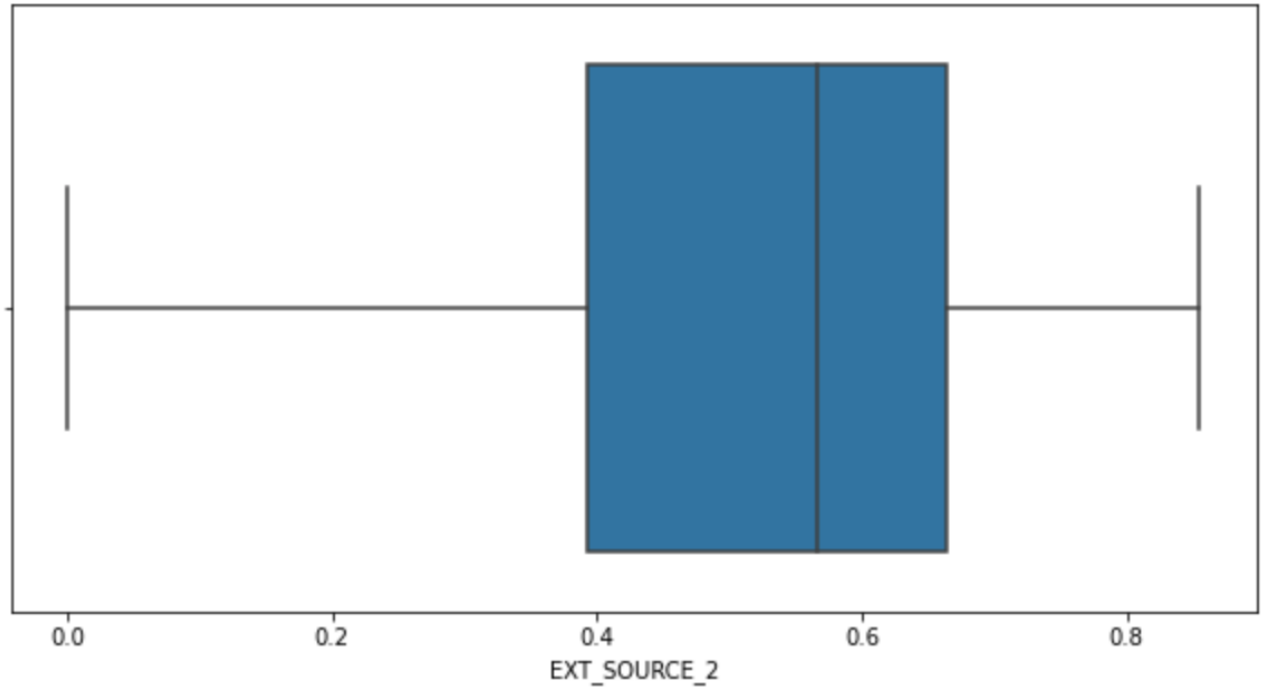
Column with outliers

Same for a column of family members where the max value reached out to 20



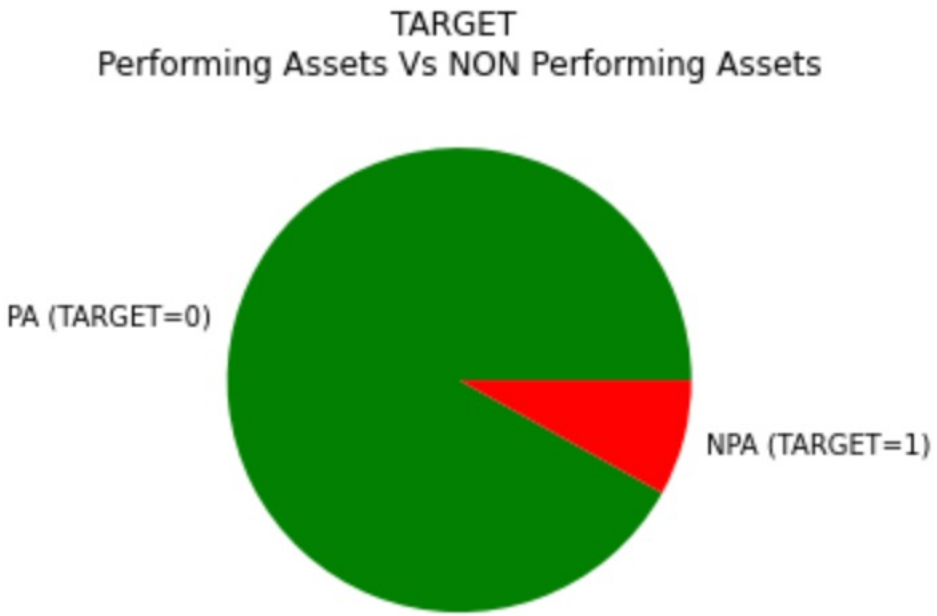
Column without outliers

- A column named EXT_SOURCE_2 had no outliers present.
- We came to know about this using boxplot.
- Thus we imputed it with the mean values as the mean values represent the overall values present in the column.



Imbalance in the data

- There is a certain and significant data imbalance present. As we can see that 90% of the people who took loan did not default.
- Maybe their amounts were very low when compared to the defaulters but that seems highly unlikely.

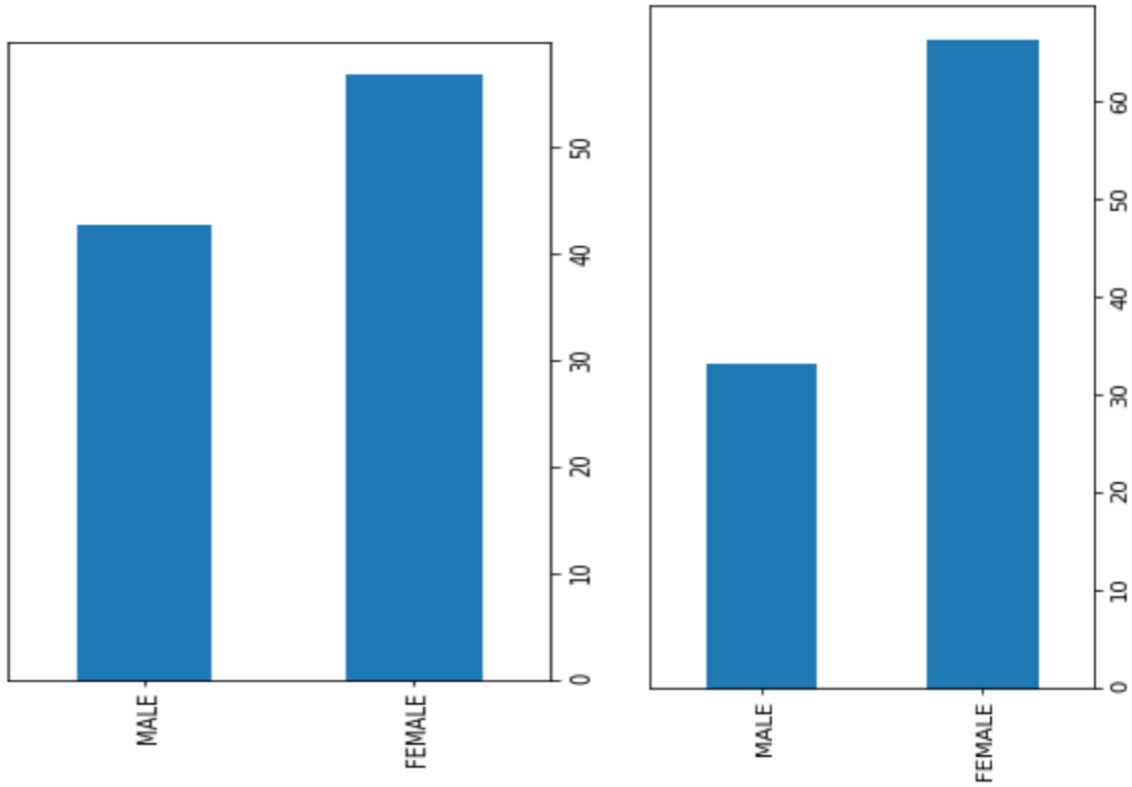


Steps to analyse the imbalance in the data

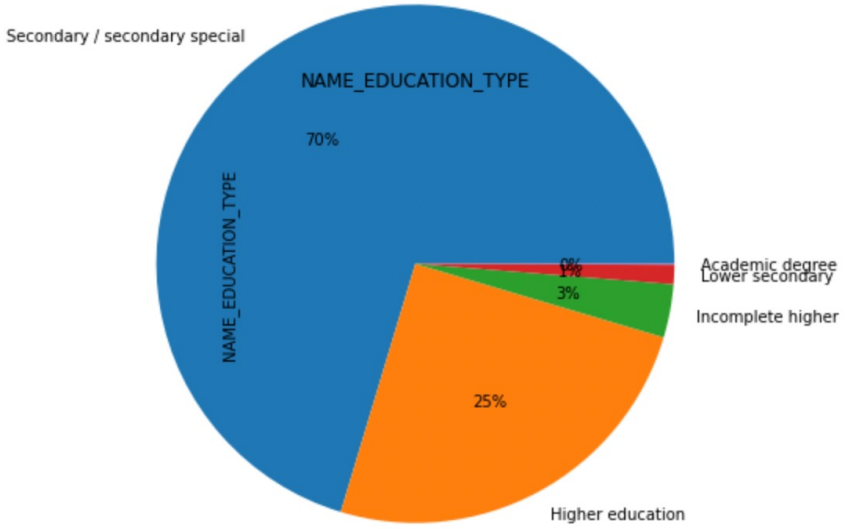
- We split the data into two separate data frames.
- Identified the categorical and numerical columns separately.
- Performed Univariate and Bivariate analysis on them.
- Generate a correlation between the dataset.

Univariate Analysis

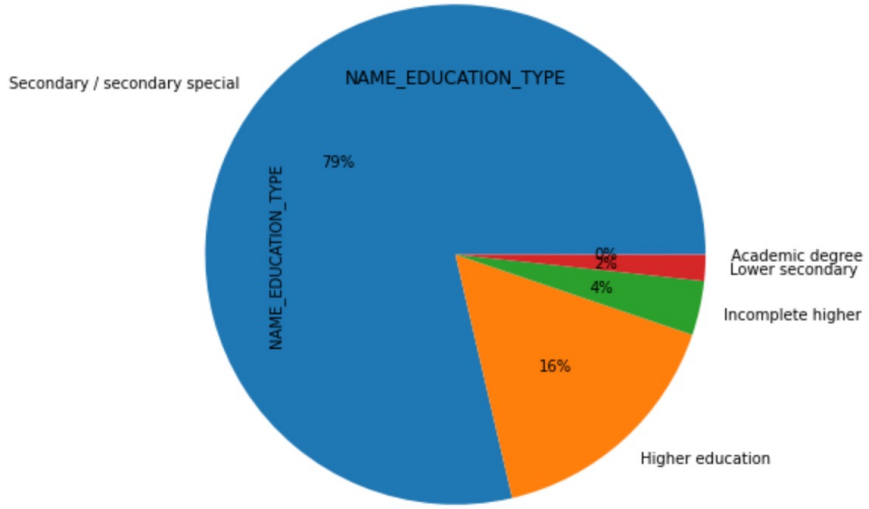
- We can see that Female contribute to the non-defaulters as well as to the defaulters.
- We can conclude that we see more female applying for loans than males and hence the more number of female defaulters as well.
- **But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts.**



Univariate Analysis

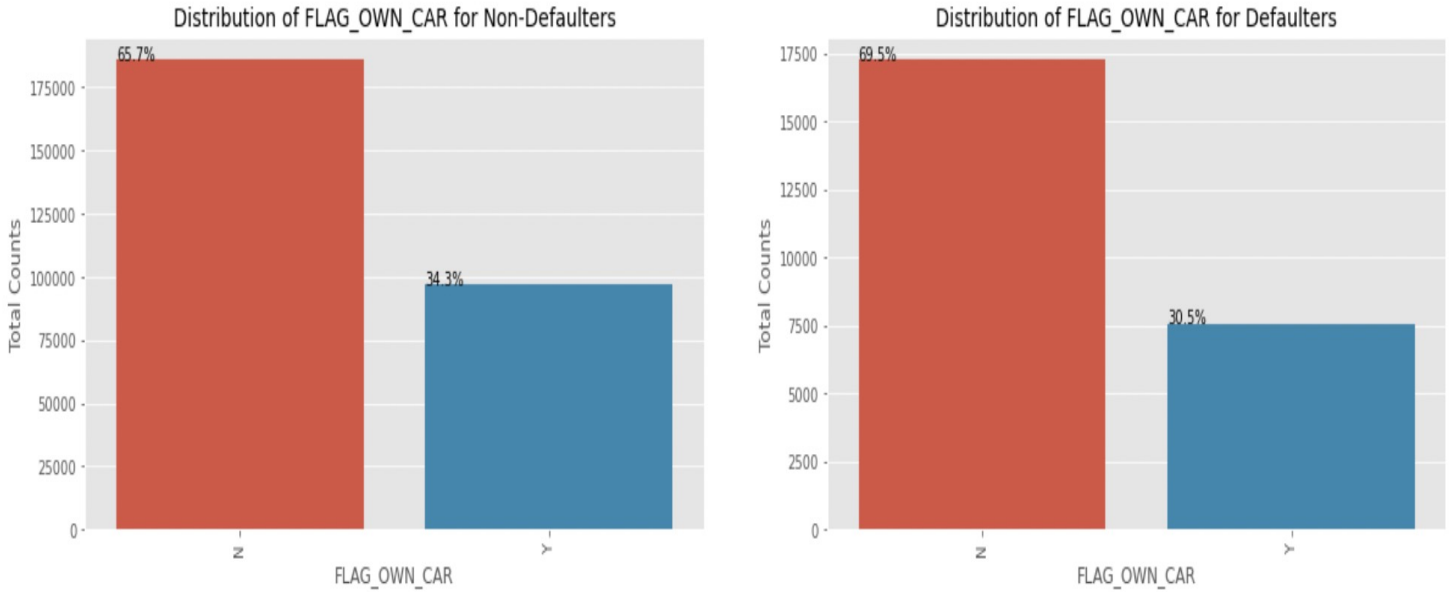


A lot of defaulters and non defaulters from secondary education background.
This shows that a lot of people have availed loans who belong to secondary education background.



Univariate Analysis

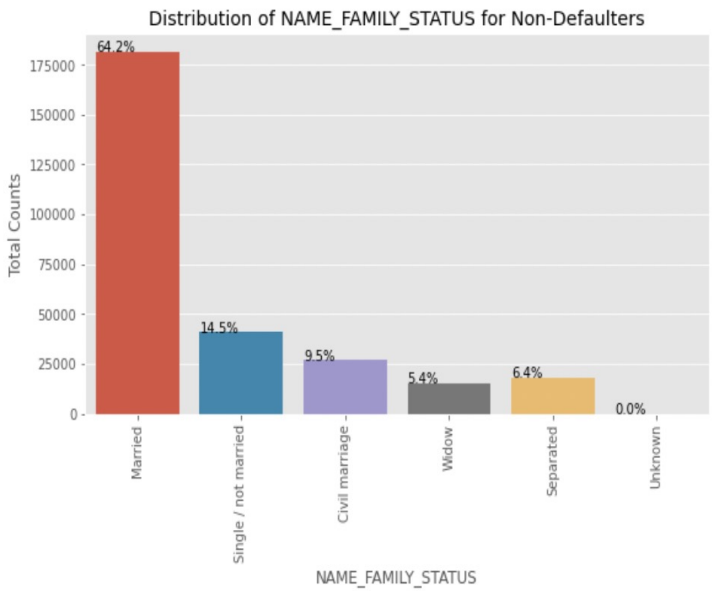
If a person owns a CAR

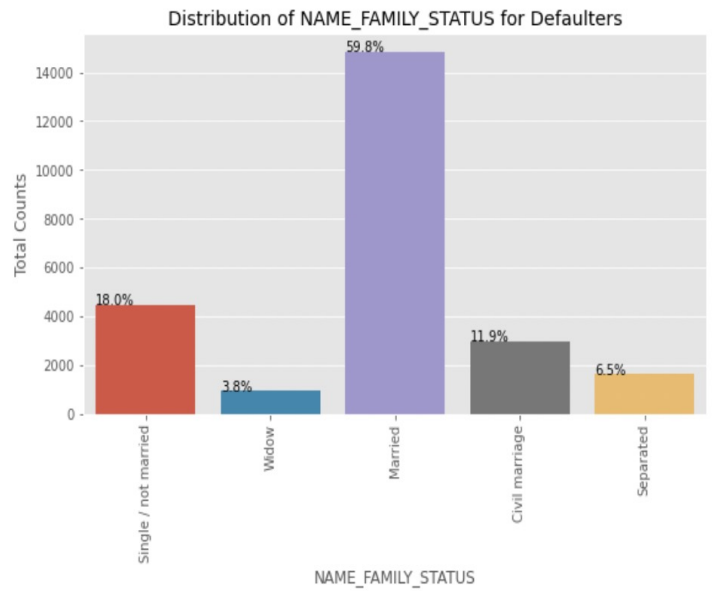


There are simply more people without cars Looking at the percentages in both the charts, we can conclude that the rate of default of people having car is low compared to people who don't.

Univariate Analysis

On the basis of family status

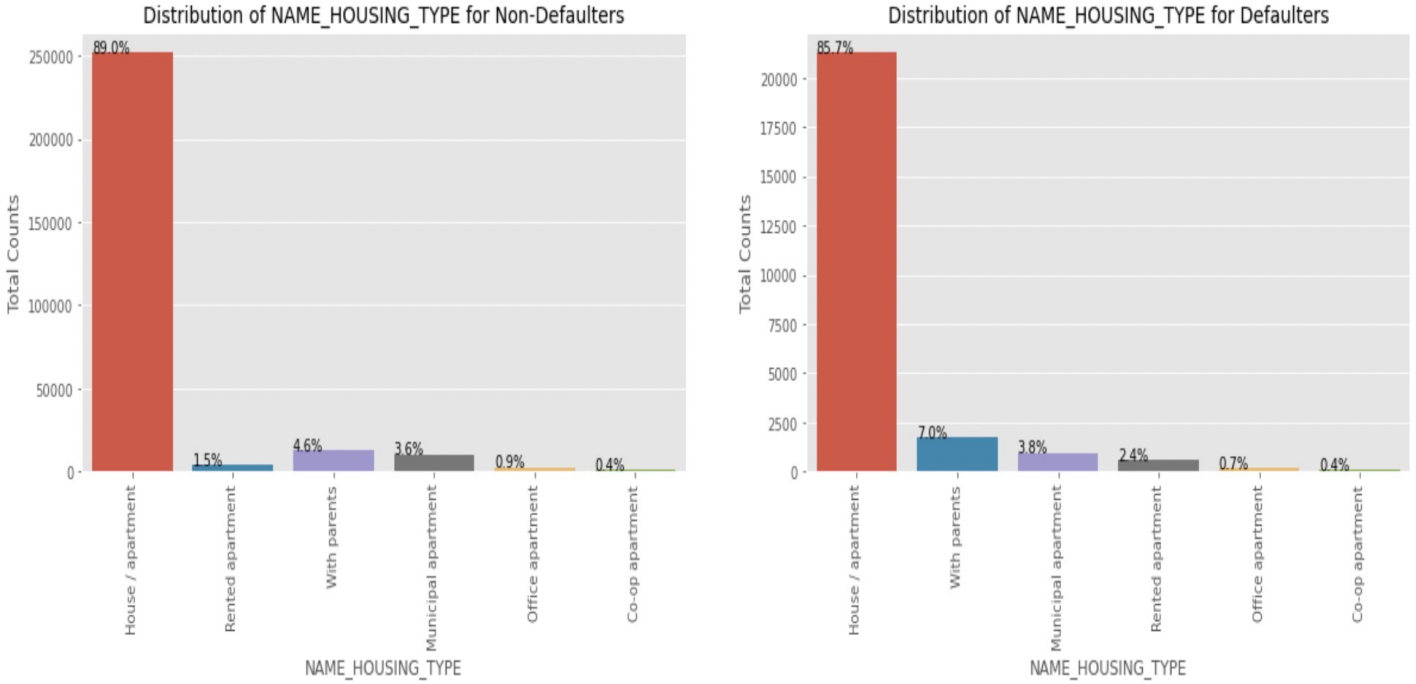




Married persons are in heavy numbers when it comes to applying for loans
It is also clearly visible that they are the one's to defalut most of the loans.
whereas single people **{WIDOW, SEPARATED}**, are very less when it comes to defaulting the loans.

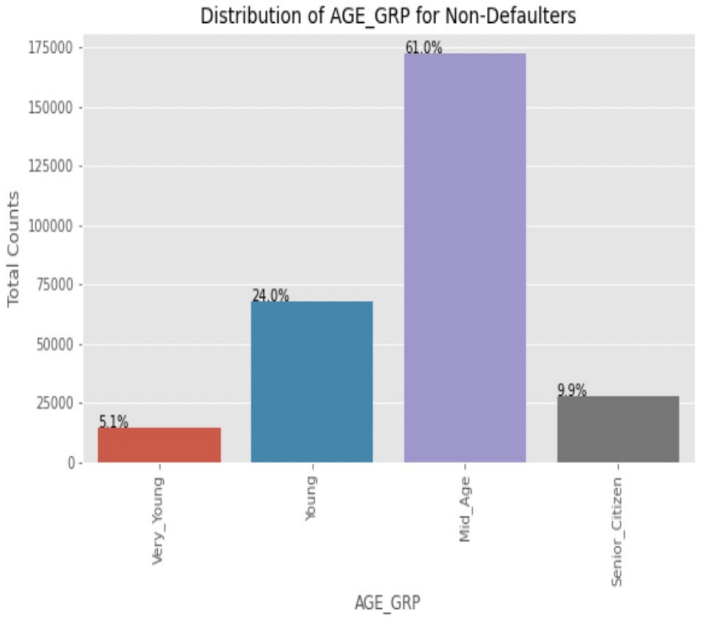
Univariate Analysis

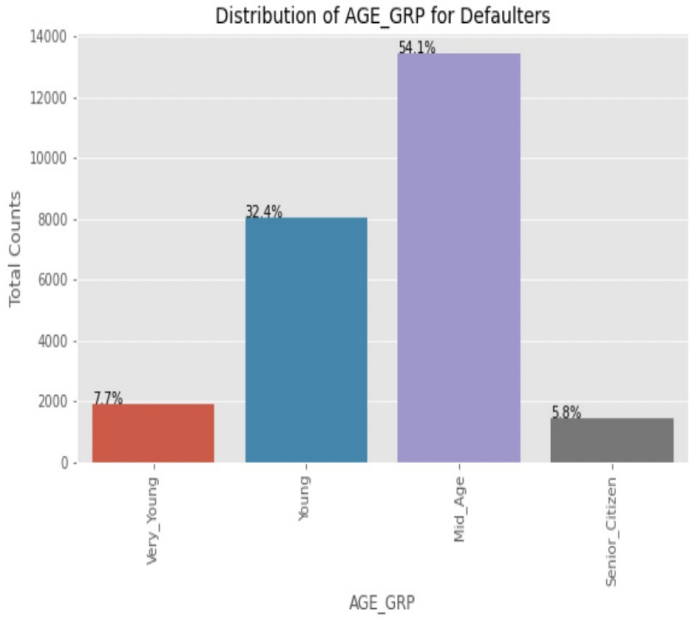
On the basis of their HOME



People with houses/apartments have applied for loans.These loans are probably for their houses themselves **{CONSTRUCTION, BUYING LAND, APARTMENT LOAN Etc.}**

On the basis of AGE

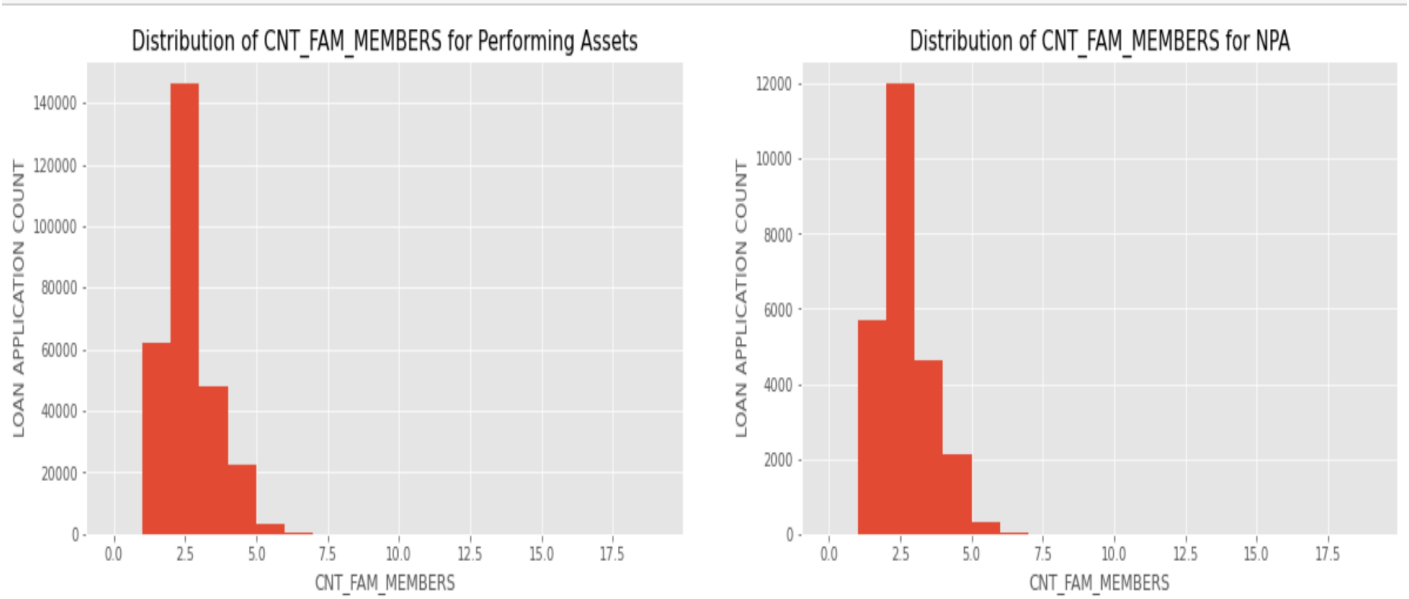




mid-age group tends to default more often

Univariate Analysis

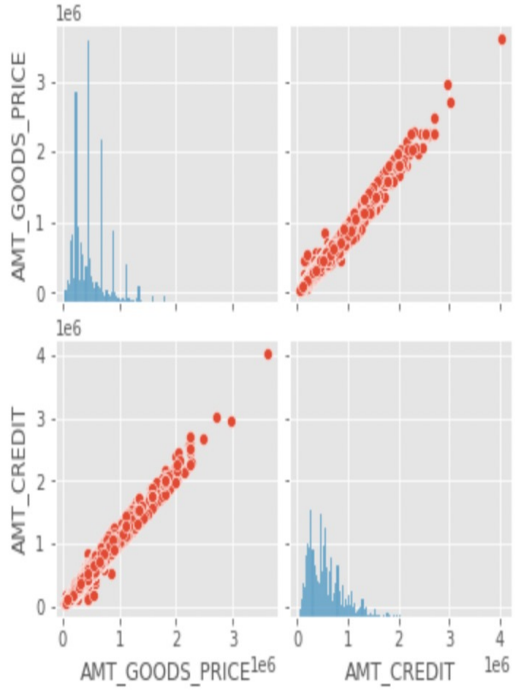
On the basis of number of family members



Most common familie are the families of 3 people.

Generally a married couple and their child who apply most for the loan.

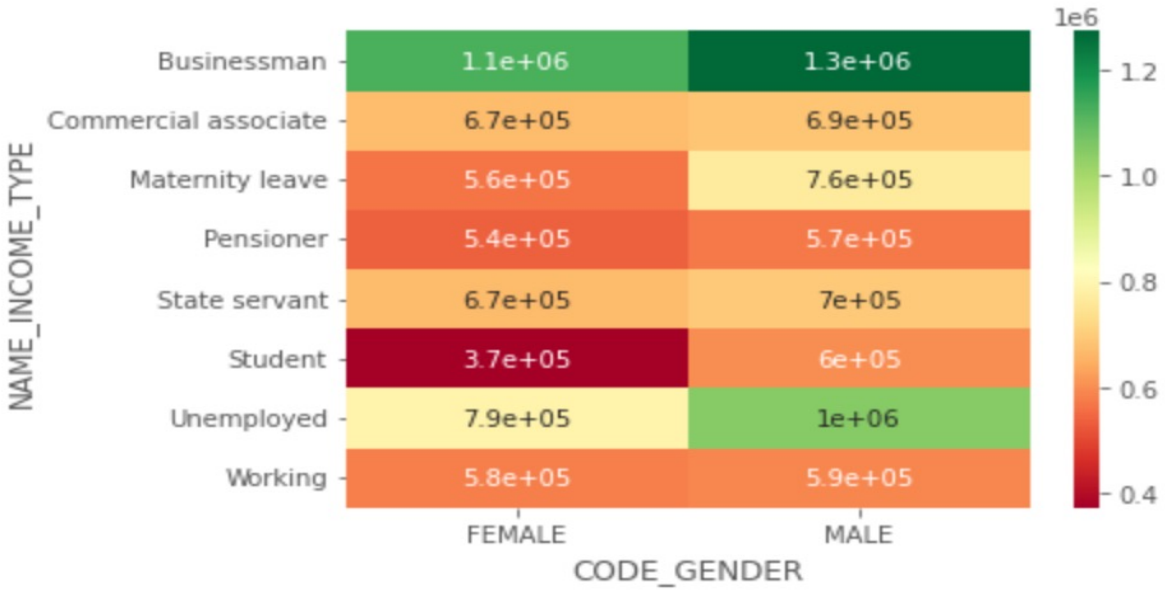
Bivariate Analysis



This is the distribution of credits on the basis of goods price which means the loan amount is decided on the basis of the price of the product which is to be purchased with the loaned amount.

Higher the goods price higher the loan amount and it is same for both the data frames whether the application is from a default account or not.

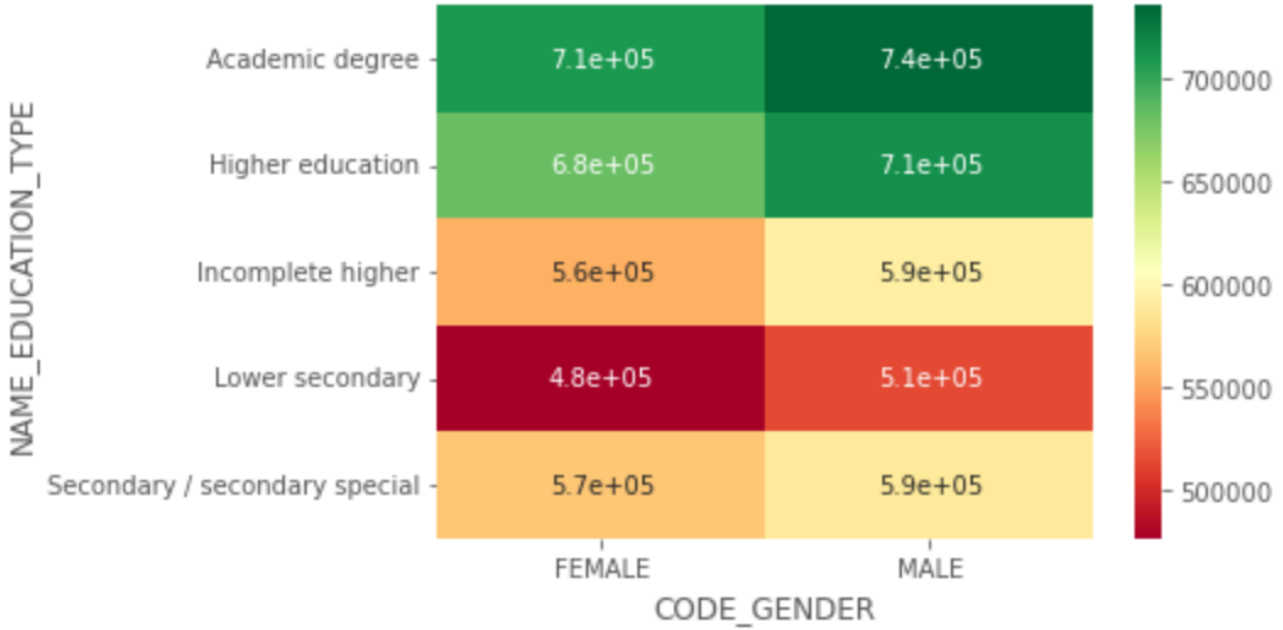
Bivariate Analysis



Businessmen have the highest credit ratio for men as well as women

Distribution of credit to males and females on the basis of income type

Bivariate Analysis

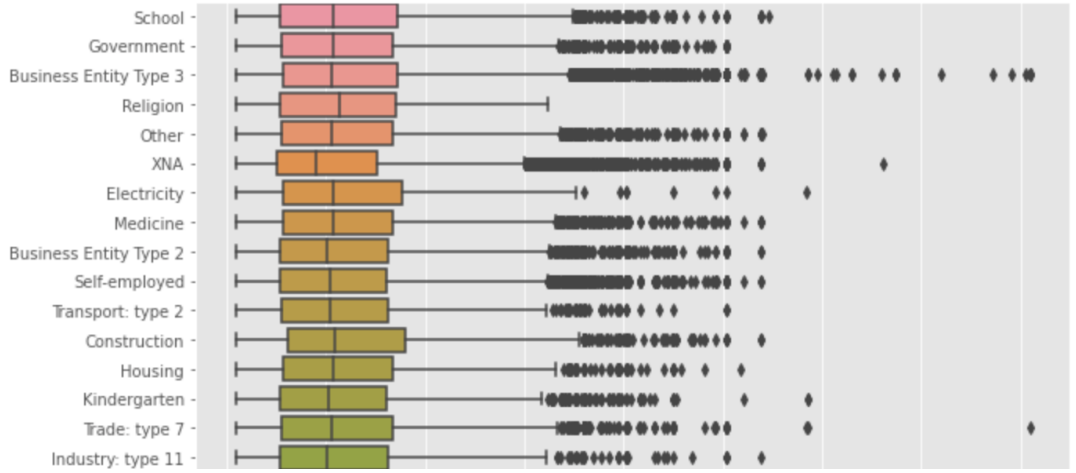


MALES with academic degree take higher loans than their counterparts

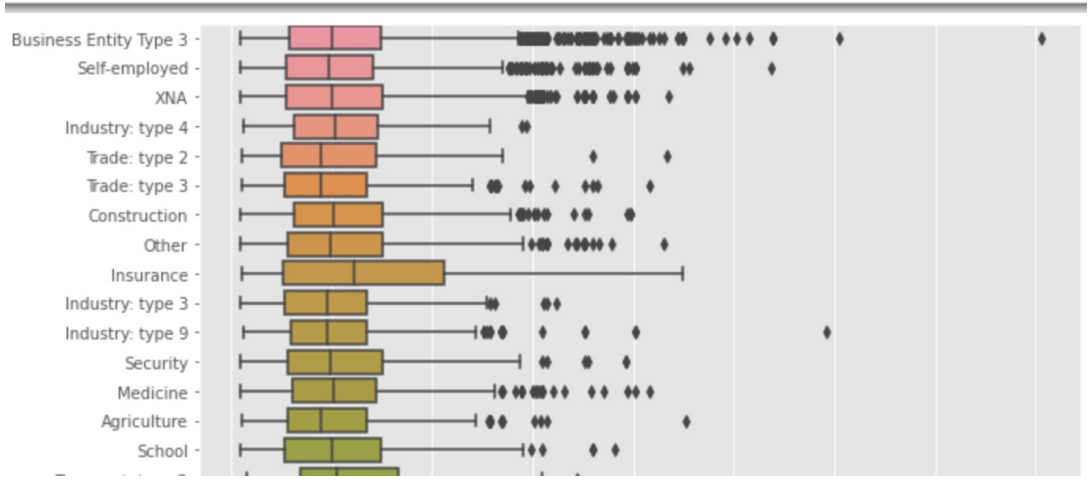
Distribution of credit to males and females on the basis of their education level

Bivariate Analysis

- Business entity type3 has taken most amount of loans and has the most number of defaulters.
- Insurance sector has no outlier values when it comes to credit repayment.



Non-defaulters



Defaulters

Bivariate Analysis

Clients who had occupation listed as managers took the highest number of loans.

Correlation

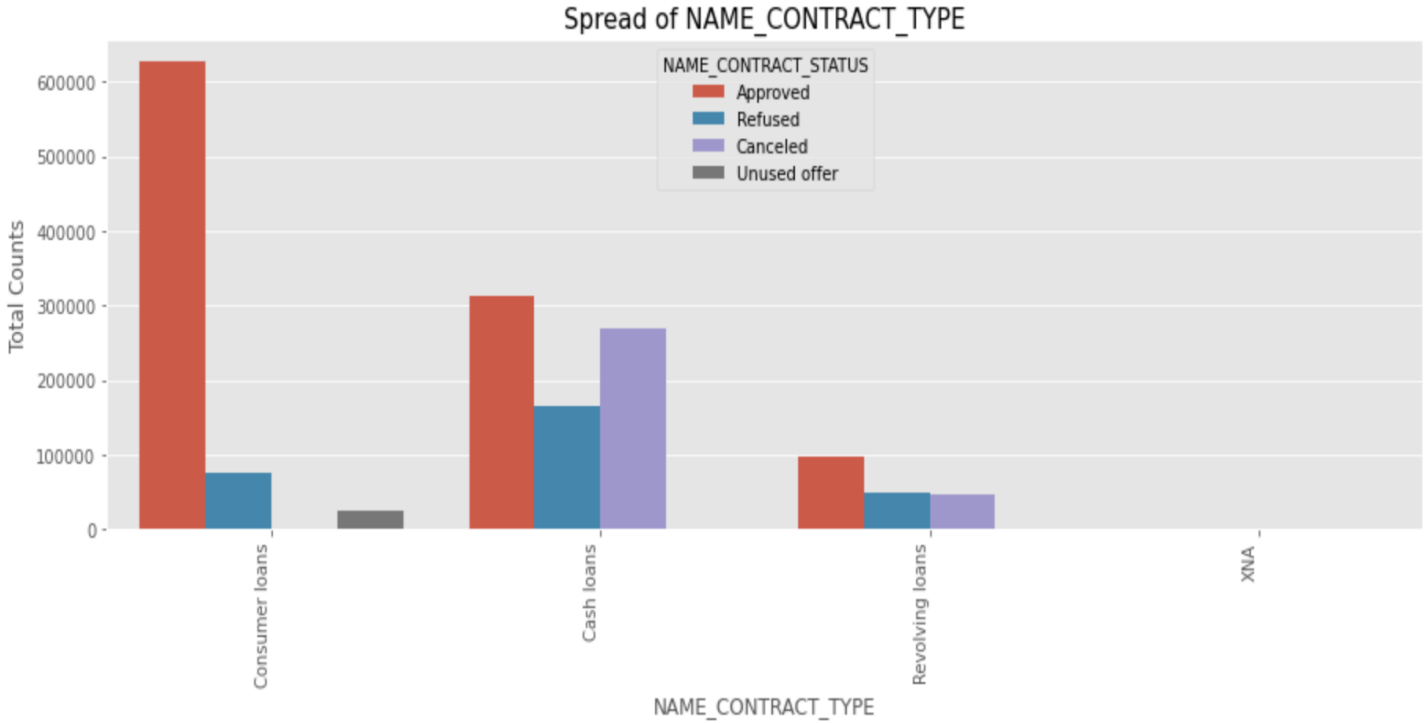
	Criteria_1	Criteria_2	Correlation	Abs_Correlation
1063	YEAR_EMP	DAYS_EMPLOYED	1.000000	1.000000
451	FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999756	0.999756
1067	YEAR_EMP	FLAG_EMP_PHONE	-0.999756	0.999756
1028	AGE	DAYS_BIRTH	0.999711	0.999711
1098	YEAR_REG	DAYS_REGISTRATION	0.999554	0.999554
1133	YEAR_ID_PUBLISH	DAYS_ID_PUBLISH	0.997518	0.997518
208	AMT_GOODS_PRICE	AMT_CREDIT	0.987024	0.987024
664	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950148	0.950148
580	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878569	0.878569
804	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861	0.861861

TOP 10 Correlation for non-defaulters

	Criteria_1	Criteria_2	Correlation	Abs_Correlation
1063	YEAR_EMP	DAYS_EMPLOYED	1.000000	1.000000
451	FLAG_EMP_PHONE	DAYS_EMPLOYED	-0.999705	0.999705
1067	YEAR_EMP	FLAG_EMP_PHONE	-0.999705	0.999705
1028	AGE	DAYS_BIRTH	0.999691	0.999691
1098	YEAR_REG	DAYS_REGISTRATION	0.999479	0.999479
1133	YEAR_ID_PUBLISH	DAYS_ID_PUBLISH	0.997531	0.997531
208	AMT_GOODS_PRICE	AMT_CREDIT	0.982783	0.982783
664	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637	0.956637
580	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
804	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885	0.847885

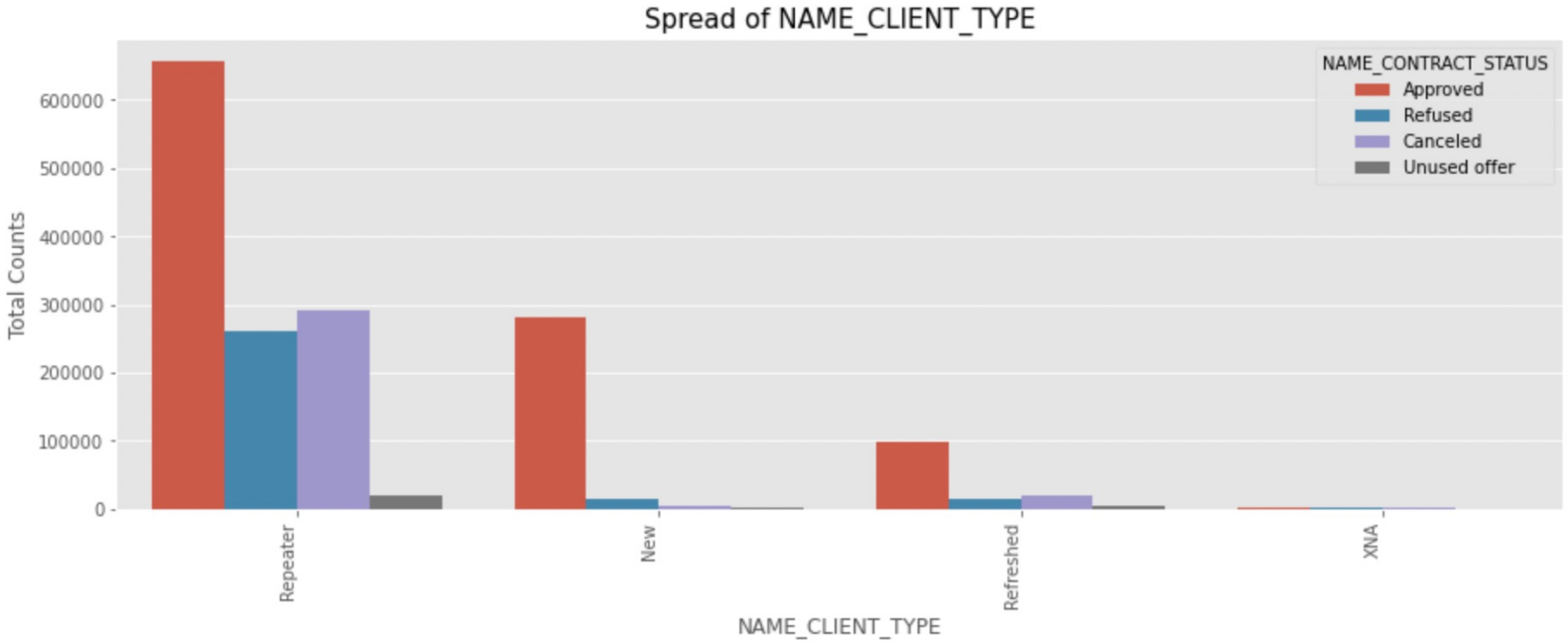
TOP 10 Correlation for defaulters

Pre-App Analysis



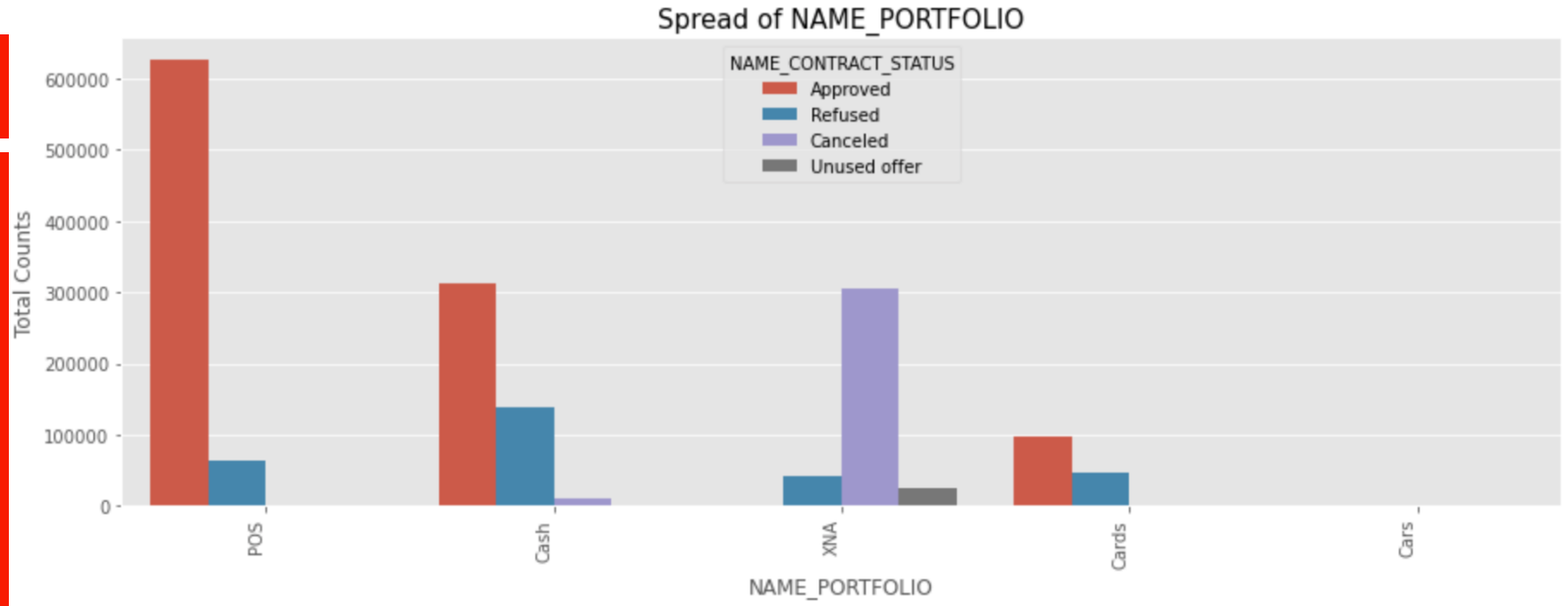
Consumer and cash loans are the most prefferd ones although the approval ratings are a big low for cash loans.

Pre-App Analysis



Repeaters are preffered more than any other client

Pre-App Analysis

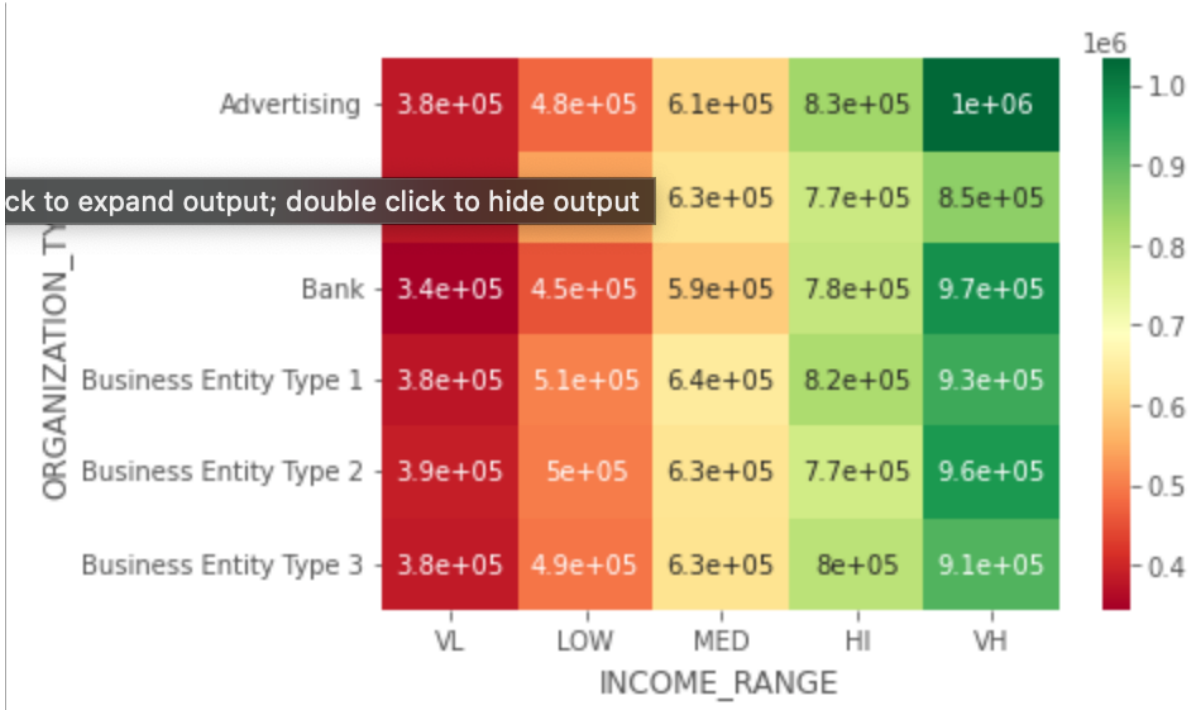


Pos portfolios are highly approved

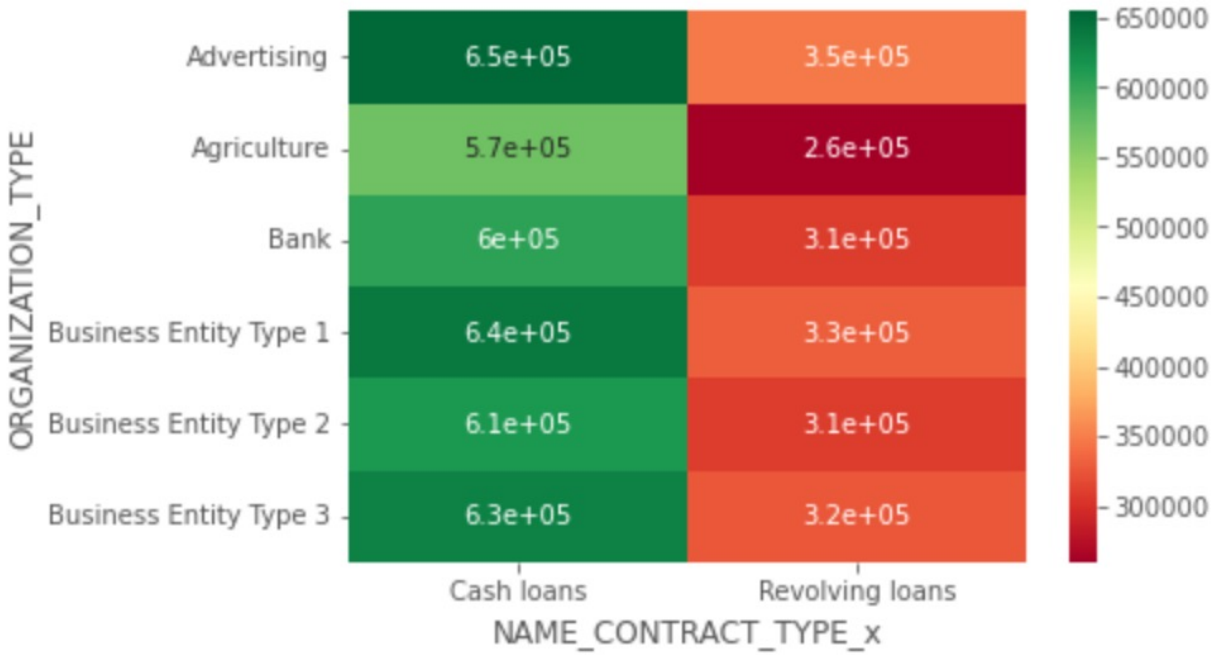
Merging

- **Merging the files previous application and new application and examining the data.**
- **Inspecting and treatment of null-values.**
- **Naming the file Target Clientele.**

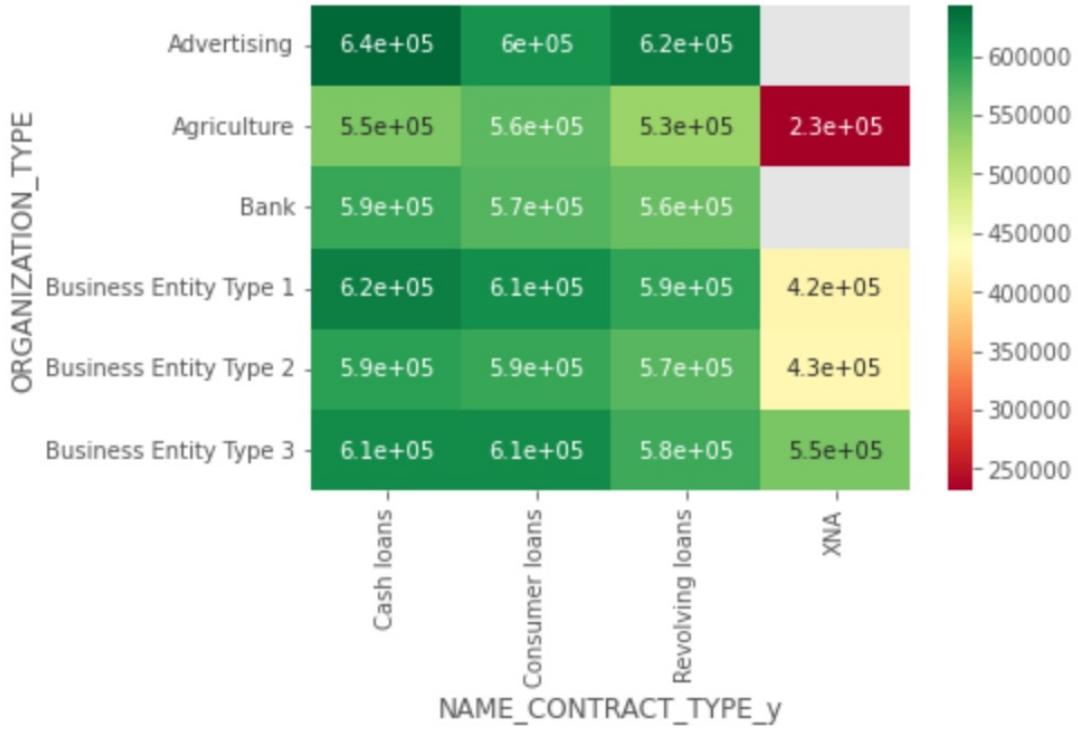
Target_Clientele



Target_Clientele



Target Clientele



Business entity type 3 has the maximum numer of loans applications and it takes moderate credit amount.

> They prefer cash loans rather than revolving loans.

Thankyou.