# CREDIT EDA CASE STUDY

**By**

**Simran Singh**
**and**
**Sowmya Vijeth**

# OVERVIEW

Companies that provide loan find it hard to give loans to the customers due to their insufficient or non-existent credit history.

Two types of risks associated with bank's decision-

1. If the customer is likely to pay the loan - Not approving the loan results in a loss of business for the company.
2. If the customer is not likely to pay the loan - Approving the loan may lead to financial loss for the company.

**The aim of this case study is to:**

- To identify patterns if a client had difficulty paying their installments.
- To ensure customers capable of repaying the loan are not rejected.
- To understand how consumer attributes and loan attributes influence the tendency of default.

# DATA UNDERSTANDING AND PREPARATION

The following csv files were used to explore and perform EDA on

1. 'Application_data.csv' - contains all the information of the client at the time of application.

2. 'previous_application.csv' contains information about the client's previous loan data.

3.'columns_description.csv' is a data dictionary which describes the meaning of the variables.

**Data preparation**:

The following two data sets were imported to jupyter notebook and basic data inspection was done using pandas

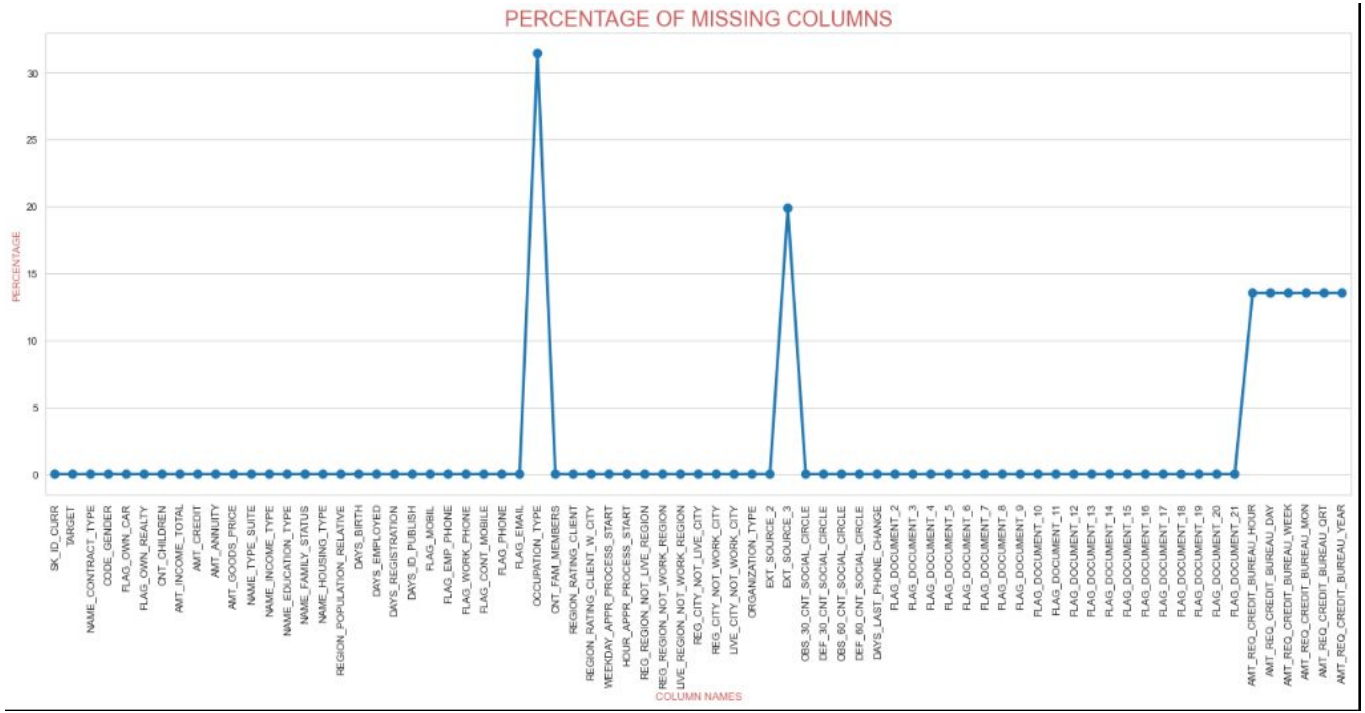
- Application_data (app_data)
- Previous_application(prev_data)

# DATA CLEANING

The process of data cleaning involved the following steps:

1. Checking the shape of the dataframe.
2. Checking the datatypes and typecasting the required attributes.
3. Checking the numerical columns using describe function.
4. Checking the null values and their percentage.
5. Negative values of few attributes were handled by taking absolute values.

# Missing columns in application data




We decided to drop the columns that had more than 40% missing data.

This graph indicates percentage of missing values in each column.

# Missing value imputation and suggestions

Columns which had > 13% missing data :

Suggestions have been provided for few columns with more than 13% of data being missing.

Columns which had  < 13% missing data :

Imputation method:

- Mean -  when its a normal distribution
- Median - when the we observe skewness in the graph.
- Mode - when its a categorical variable

# Duplicates, missing data and binning data

1. No duplicate records were found.
2. As per industry standards we decided to drop attributes which had 40%~50% missing data. Keeping these records might lead to difficulties in drawing inferences.
3. Binning was performed on couple of continuous variables like income and work experience of the clients
4. We even separated Categorical and Numerical Columns for ease during analysis

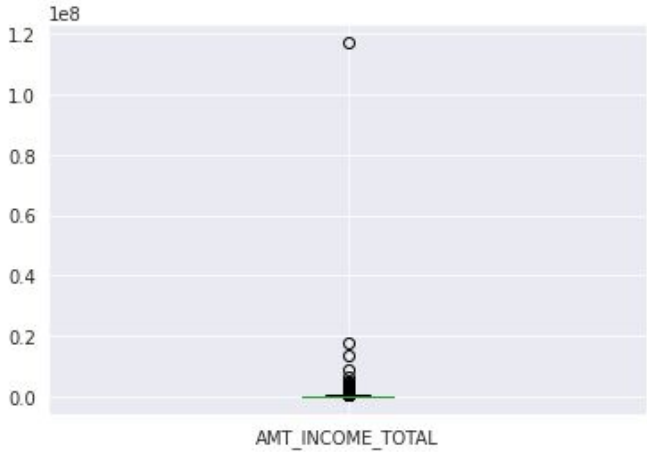
# OUTLIER DETECTION METHOD

**For outlier detection we have used:**

- Box plot

- Describe function

- Quantile function




Example of an outlier in the dataset:

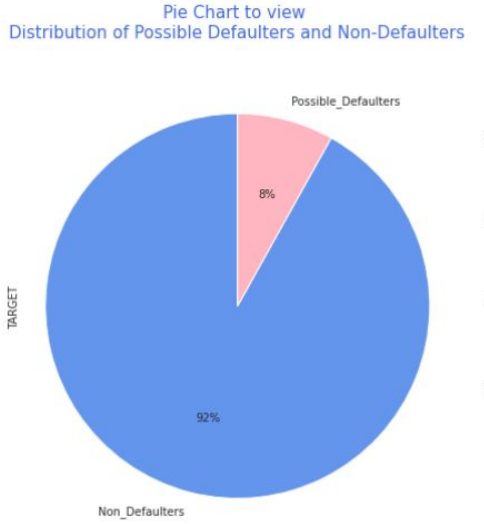
The income column has one value which is way beyond all the other values.

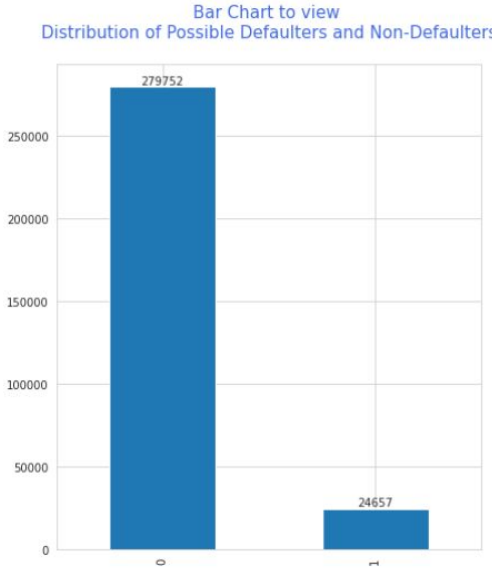
# OUTLIER TREATMENT

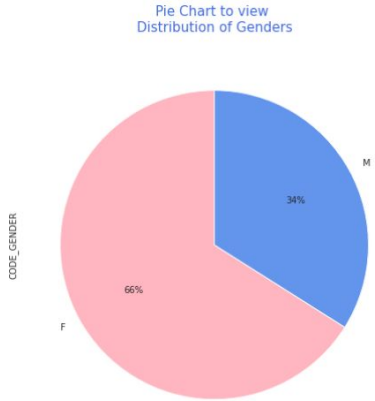
**Example of Inference and Treatment:**

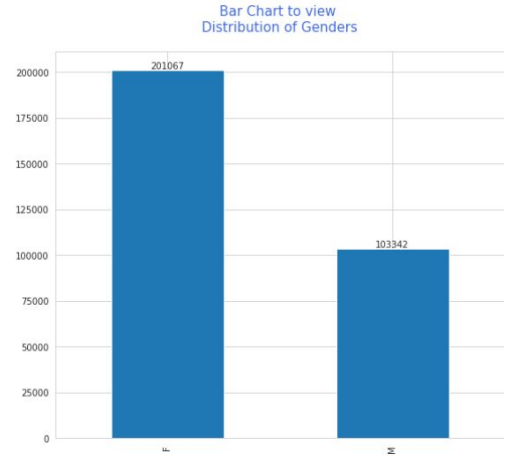
- We can visually see these outliers through the boxplot.
- We can see that there is a value of '117000000' in the AMT_INCOME_TOTAL column that is an outlier much farther than the rest of the values. This could be an error or it could be valid value.
- There is large variation in data between 99th percentile and max value. In this column for income, it is possible to see variations like this.
- We can choose not to treat them and instead just be aware of the situation and let them be left untreated. Thus whenever we are aggregating/analysing these values, instead of looking at the mean, we can look at percentiles/medians so we don't get affected by outliers.
- Such a decision is dependant on a case by case basis according to the business problem.

**In this case study we have decided to cap outliers at 99th percentile.**



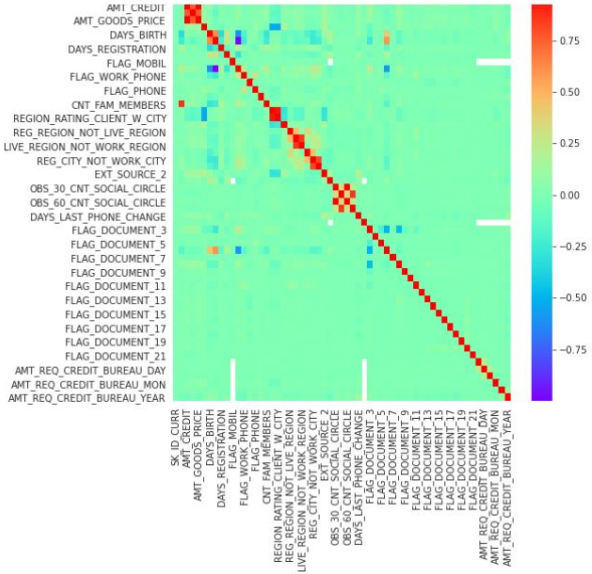




# IMBALANCE RATIO

- **The first distribution shows the data imbalance in regards to Target variable as 92% Non-defaulters and 8% Possible-Defaulters**

- **The second distribution shows data imbalance in regards to gender as 66% Female and 34% Male**

# MULTIVARIATE ANALYSIS USING HEATMAP




# Segmenting application _data with respect to TARGET variable

The dataframe has been separated with respect to TARGET variable.

- df1 - Applicants who are likely to be defaulters
- df0 - Applicants who are non-defaulters

# TOP 10 CORRELATED VARIABLES - For Possible Defaulters

| | | |
|---|---|---|
| OBS_60_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 1.00 |
| OBS_30_CNT_SOCIAL_CIRCLE | OBS_60_CNT_SOCIAL_CIRCLE | 1.00 |
| AMT_GOODS_PRICE | AMT_CREDIT | 0.98 |
| AMT_CREDIT | AMT_GOODS_PRICE | 0.98 |
| REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | 0.96 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.96 |
| CNT_CHILDREN | CNT_FAM_MEMBERS | 0.89 |
| CNT_FAM_MEMBERS | CNT_CHILDREN | 0.89 |
| DEF_60_CNT_SOCIAL_CIRCLE | DEF_30_CNT_SOCIAL_CIRCLE | 0.87 |
| DEF_30_CNT_SOCIAL_CIRCLE | DEF_60_CNT_SOCIAL_CIRCLE | 0.87 |

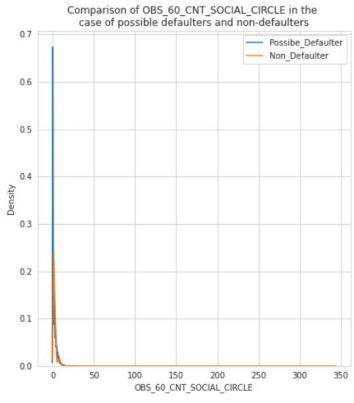
dtype: float64

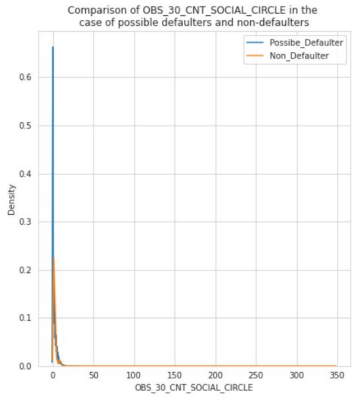
# TOP 10 CORRELATED VARIABLES - FOR NON-DEFAULTERS

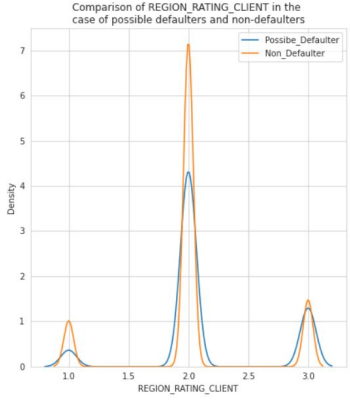
| | | |
|---|---|---|
| OBS_30_CNT_SOCIAL_CIRCLE | OBS_60_CNT_SOCIAL_CIRCLE | 1.00 |
| OBS_60_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 1.00 |
| AMT_CREDIT | AMT_GOODS_PRICE | 0.99 |
| AMT_GOODS_PRICE | AMT_CREDIT | 0.99 |
| REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | 0.95 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.95 |
| CNT_FAM_MEMBERS | CNT_CHILDREN | 0.88 |
| CNT_CHILDREN | CNT_FAM_MEMBERS | 0.88 |
| REG_REGION_NOT_WORK_REGION | LIVE_REGION_NOT_WORK_REGION | 0.86 |
| LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.86 |

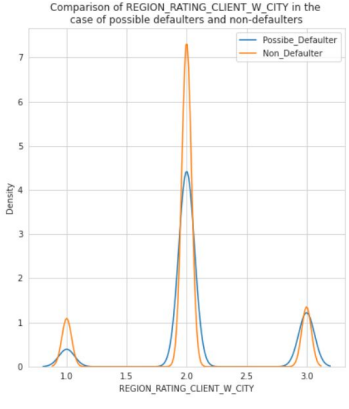
dtype: float64

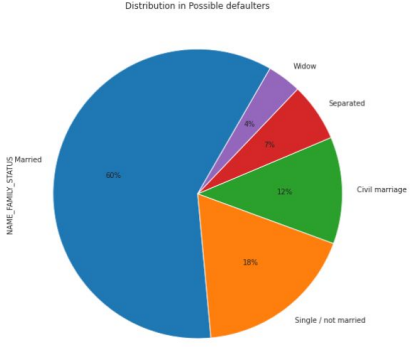
# UNIVARIATE ANALYSIS - FEW OF THE CHARTS

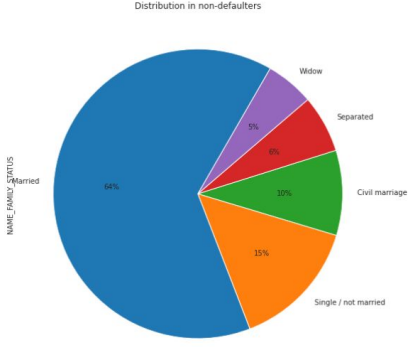


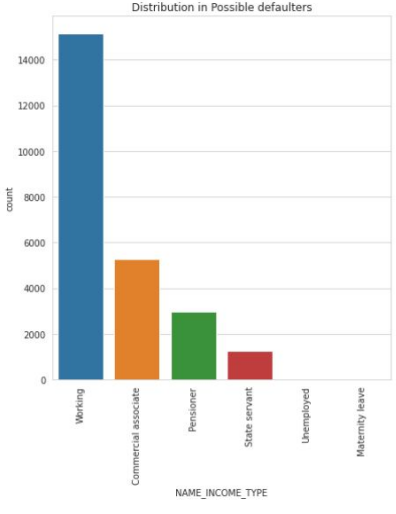



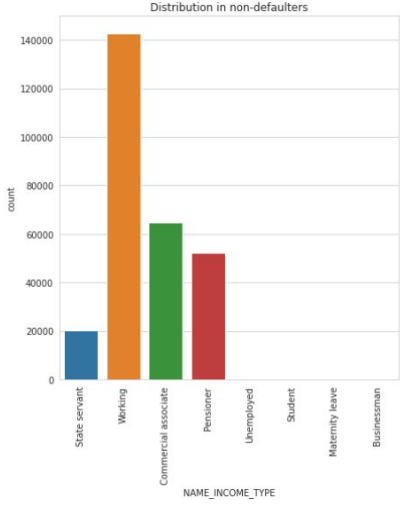



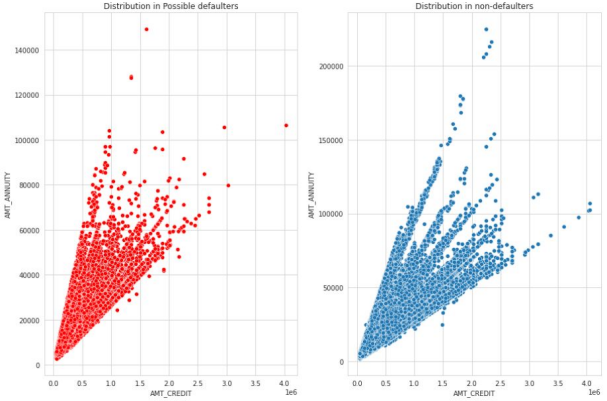



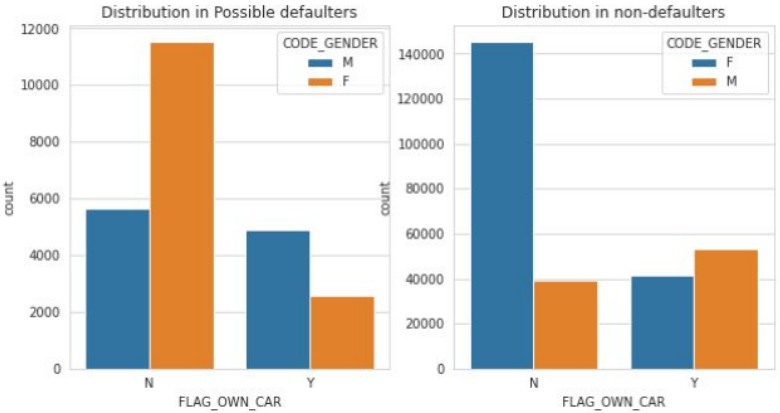



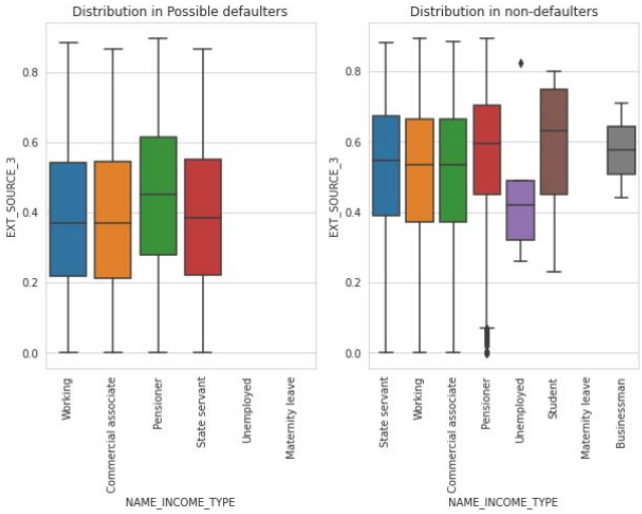
# UNIVARIATE ANALYSIS - INFERENCES

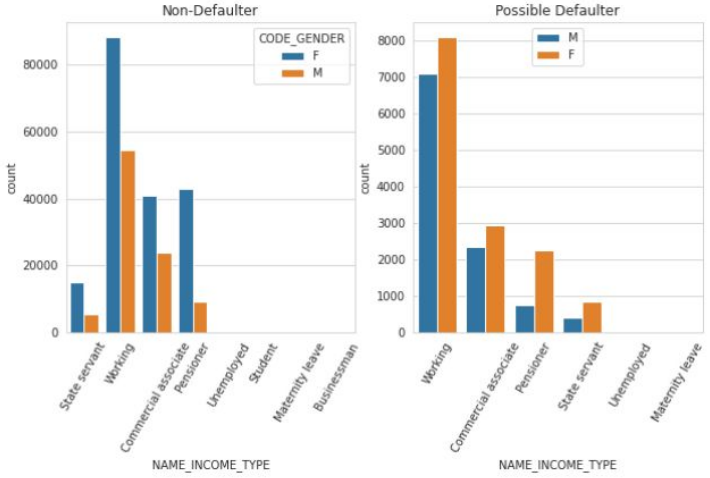
- High number of observations have been seen in the possible defaulters' social circle regarding late payments of interest; as compared to non-defaulters.
- Ratings of the regions where possible defaulters' live are quite low; as compared to non-defaulters.
- Possible defaulters' permanent address does not match work address and contact address does not match work address.
- As amount credit and amount goods price increase, we see more observations from the possible defaulters.
- Possible defaulters have to pay higher loan annuity than non defaulters.
- 60% of possible defaulters' are married and 57% of possible defaulters' are Female
- Among both types of clients highest proportion of Occupation Type is Laborers, Income Type is Working, Education Type is Secondary/secondary special.

# BIVARIATE ANALYSIS - FEW OF THE CHARTS



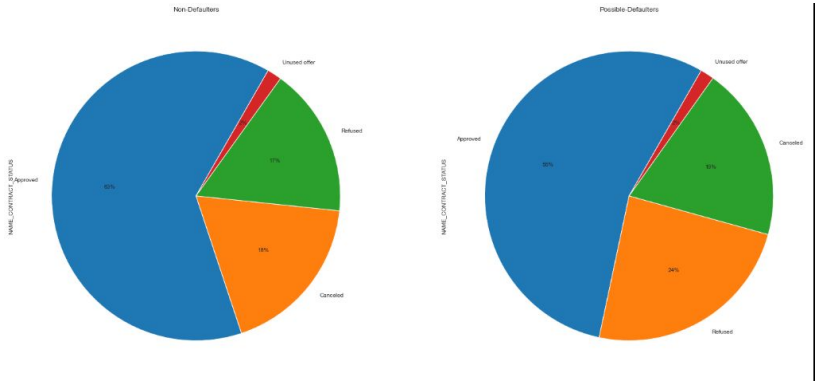
# Bivariate analysis- inferences

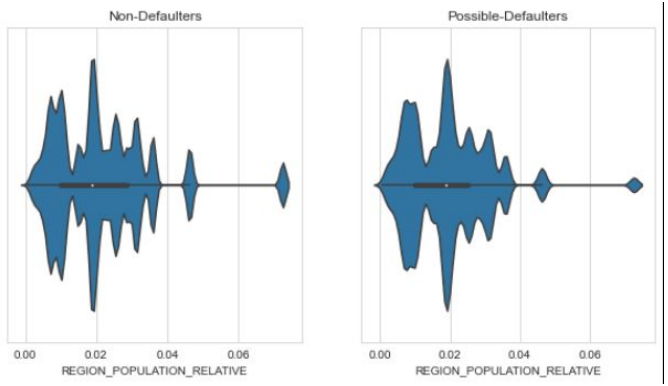
- Credit amount of the loan, Goods Price and Loan Annuity amount are highly correlated Variables.
- It seems majority of possible defaulters have taken higher loans than the non-defaulters across education types.
- Possible defaulters have low credit scores (mostly between .2 and .55) while the non-defaulters have higher credit scores (mostly between .3 and .75) across income types.
- Possible defaulters have changed their phones more recently across both genders this shows less trustworthiness.
- For possible defaulters across the kinds of family status the credit scores are low.
- Applicants with family status is married and have higher education are most likely to default on loans

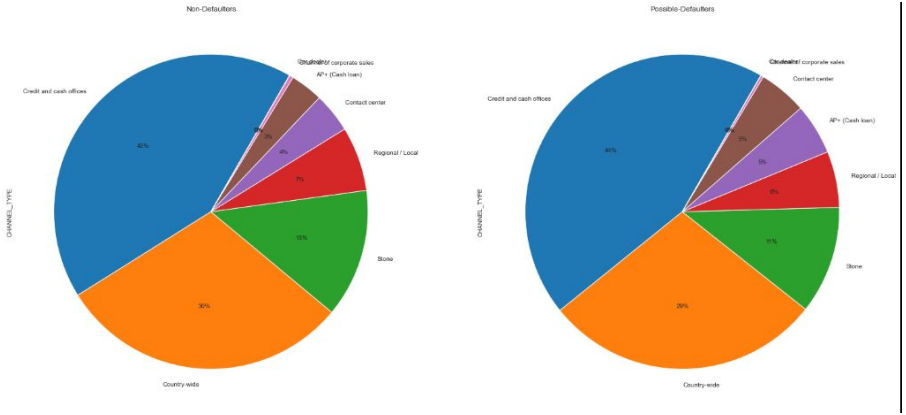
# PREVIOUS _APPLICATION DATASET

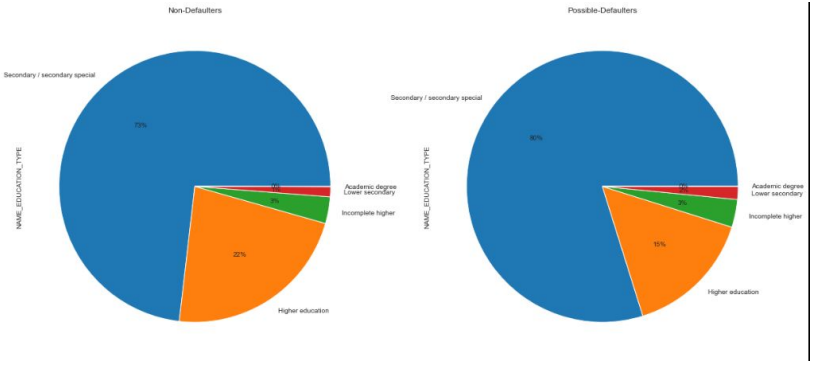
- We performed Data Understanding and Preparation & Data Cleaning and Manipulation steps on previous_application dataset similar to how we did them on application_dataset.
- Next we merged both the datasets and created combined_df
- Then we made a heatmap of  combined_df to identify all highly correlated variables
- Then we checked the data imbalance of combined_df and found out the correlation between variables using heatmap
- Finally we segmented the dataframe on basis of Target variable so we could begin analysis.
- We began the analysis by listing top 10 correlated variabled

# UNIVARIATE ANALYSIS - FEW OF THE CHARTS



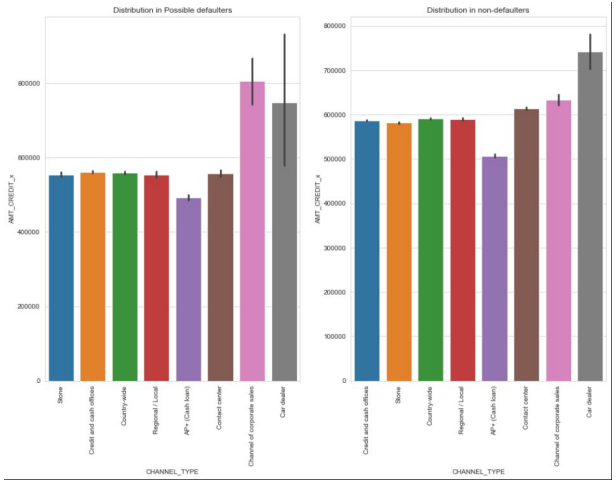



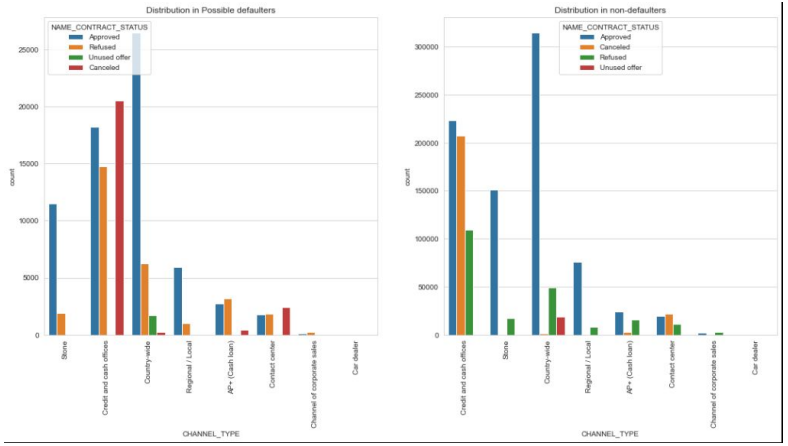



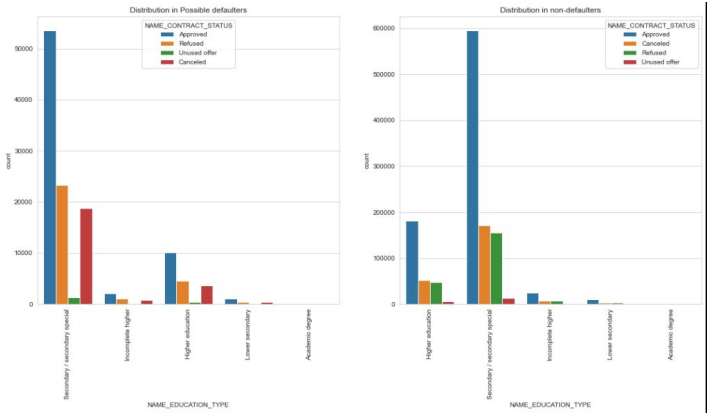
# UNIVARIATE ANALYSIS - INFERENCES

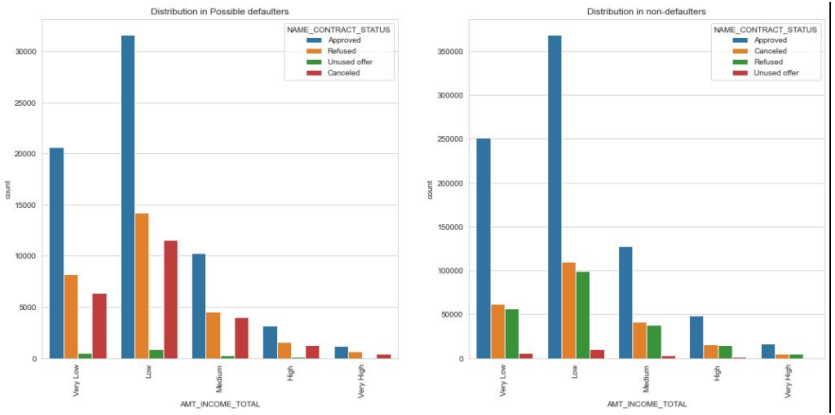
- Non-defaulters cancelled applications are more than refused and Possible-Defaulters refused applications are more than cancelled
- Credit and cash offices are types through which clients were previously acquired
- In region with high population density, clients are less likely to default on loans
- Secondary education applicants are 7% more likely to be possible - defaulters when compared with non-defaulters
- Higher Education applicants are 7% more likely to be non-defaulters

# BIVARIATE ANALYSIS - FEW OF THE CHARTS



# BIVARIATE ANALYSIS - INFERENCES

- We can see that Car dealer loan of extremely high credit are among non-defaulters. Whereas those clients acquired through corporate sales who have taken extremely high credit are likely to be defaulters. Bank should make a note of this.
- The non-defaulters cancellation is higher through credit and cash offices. Bank should see why these clients are cancelling the application.
- We can see that a large number of persons in the non-defaulters category are getting their applications cancelled, the bank needs to look further in this matter to avoid interest loss.
- Applicants who have Low income are most refused in non-defaulters, this needs to be looked into to avoid interest loss.

# CONCLUSION AND RECOMMENDATIONS

The bank should pay attention to the following to identify **possible defaulters**

1. Those who take extremely high amount of credit
2. Those who have low credit scores (mostly between .2 and .55)
3. Those who changed their phones more recently
4. Those whose persons in social circle have done late payments
5. Those who live in lower rated regions or whose work, contact, permanent addresses don't match

The bank should pay attention to the following to make sure those **capable of repaying loans are not rejected**

1. Bank should look into why more applications of non-defaulters were cancelled.
2. Applicants who have Low income are most refused in non-defaulters, this needs to be looked into to avoid interest loss.
3. We can see that Car dealer loan of extremely high credit are among non-defaulters. Whereas those clients acquired through corporate sales who have taken extremely high credit are likely to be defaulters. Bank should make a note of this.