DS-GA-1007 (Fall 2020) - Data Science Project

Course Professor- Prof. Milan Bradonjic

August 2020 - December 2020

Contents

1. Introduction

Our project aims to apply Exploratory Data Analysis (EDA) in a real world business scenario. In the following case study, we will 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.

1.1. Problem

Banks find it hard to give loans to the people due to their insufficient or non-existent credit history and because of that, some consumers use it as their advantage by becoming a defaulter. Inversely, Banks often find it difficult to target the right people for loan advertisements.

Secondly, COVID-19 has stirred up health and safety concerns among people. Due to this, banks have been losing walk-in customers, which has affected the number of new loan applications.

1.2. Business Implications

Having a predictive model that would determine the chances of a person missing a payment on their loan and better understand the customer's repayment capabilities, would be useful for the Bank.

Alleviating the impact of health and safety concerns of COVID-19 and improving regulations during epidemics through analysis on customer data/feedback would prove to attract more bank loan applications. This could help banks develop a more accurate and customized advertisement campaign.

2. Dataset and Setup

We are working with Python 3.8 and our compute platform is GoogleColaboratory.

2.1. Dataset

The dataset chosen provides us granular details about people visiting banks like Age, Sex, Address, Education level, Income level, Family members etc. The Dataset is available on Kagle.com.

It contains information about 308,000 people who applied for loans and accounts for minute attributes that prove useful on various accounts.

2.2. Libraries

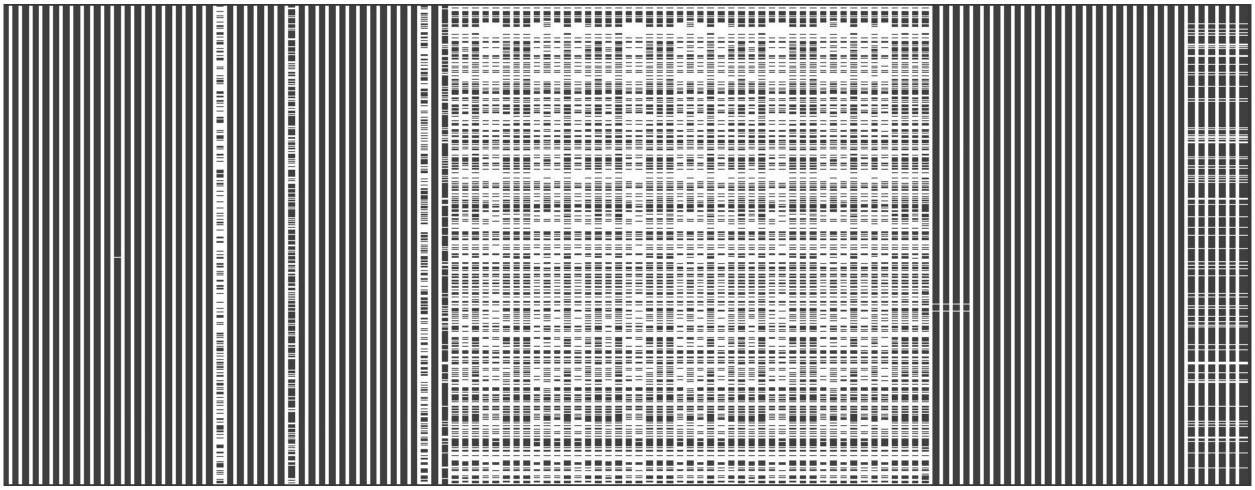
numPy - The fundamental package for scientific computing with Python
Pandas - fast, flexible data structures designed to make working with structured easy.
OS - Miscellaneous operating system interfaces
Intertools - Functions for creating iterators for efficient looping.
Pyplot - functions for creating and analysing graphs
Seaborn - data visualization library based on matplotlib
Sklearn - Simple and efficient tools for predictive data analysis

3. Data Preparation and preprocessing

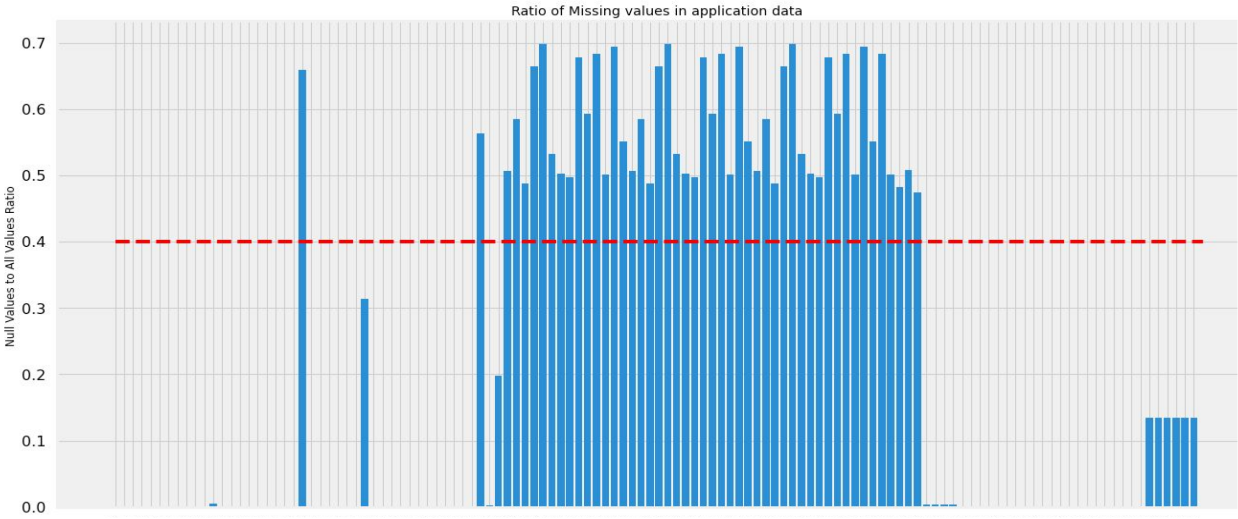
Dealing with a huge data set requires understanding and cleaning the data before useful information can come out of it. We used the following strategies to prepare our data.

3.1. Null Value calculations

Using the missingno library, we were able to get a visual representation of missing values in our huge dataset. The following image shows the missing values in our huge dataset.



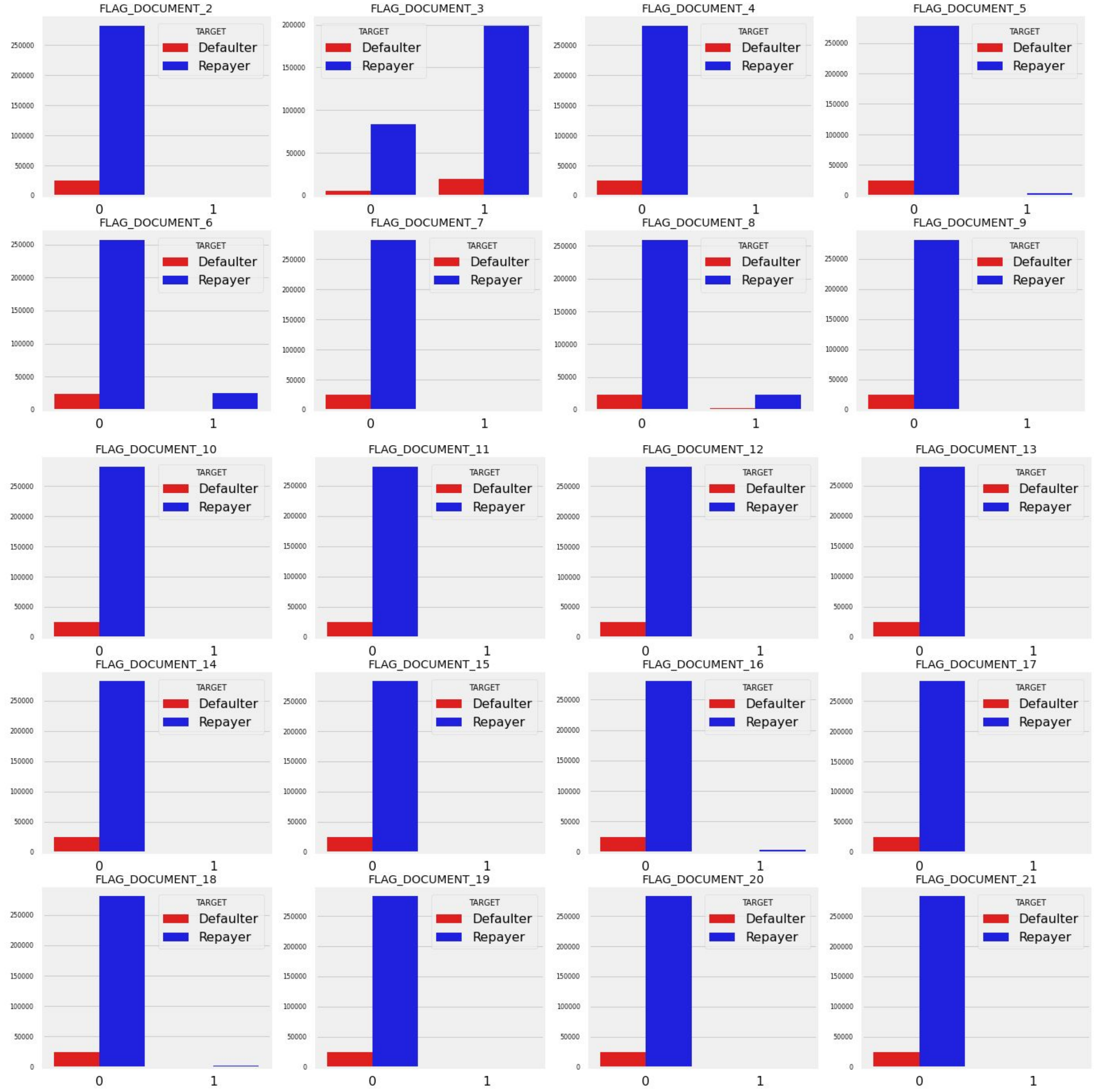
Here we notice that a lot of features have values missing that won't be useful for us. For this project we decided to keep the threshold for missing values to be 40 percent.



3.2. Correlation anatomy

In a dataset this big, Several features in our dataset correspond to general information/ submission of certain documents etc. We wanted to make sure if keeping them for further processing made sense. To figure out which feature contributed significantly to a customer being a defaulter, we tried to narrow down correlations between each of these features and a customer being a defaulter.

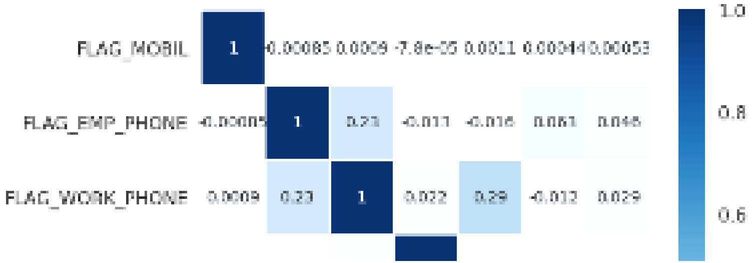
3.3. Document Submission and Defaulting



From these graphs we conclude that, even with a high number of people not submitting the document, the number of defaulters is very low.

However, the same cannot be said for DOCUMENT_3, where proper submission of this document meant less number of defaulters.

3.4. General Information and Defaulting

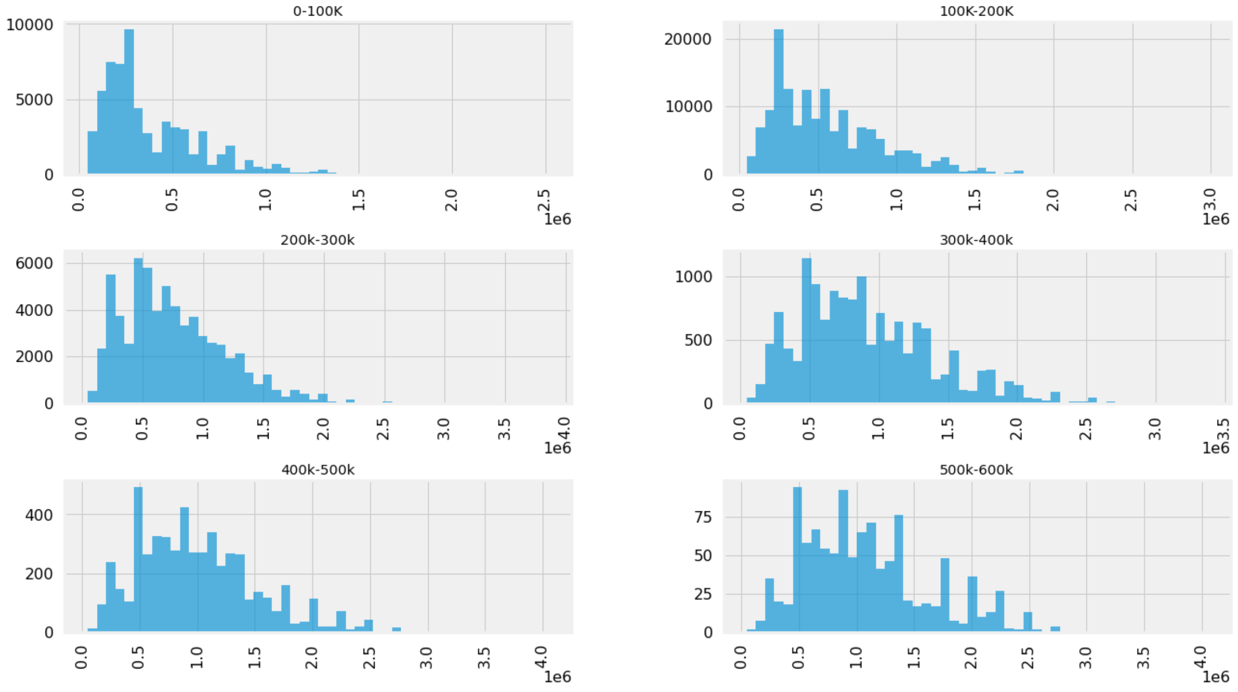


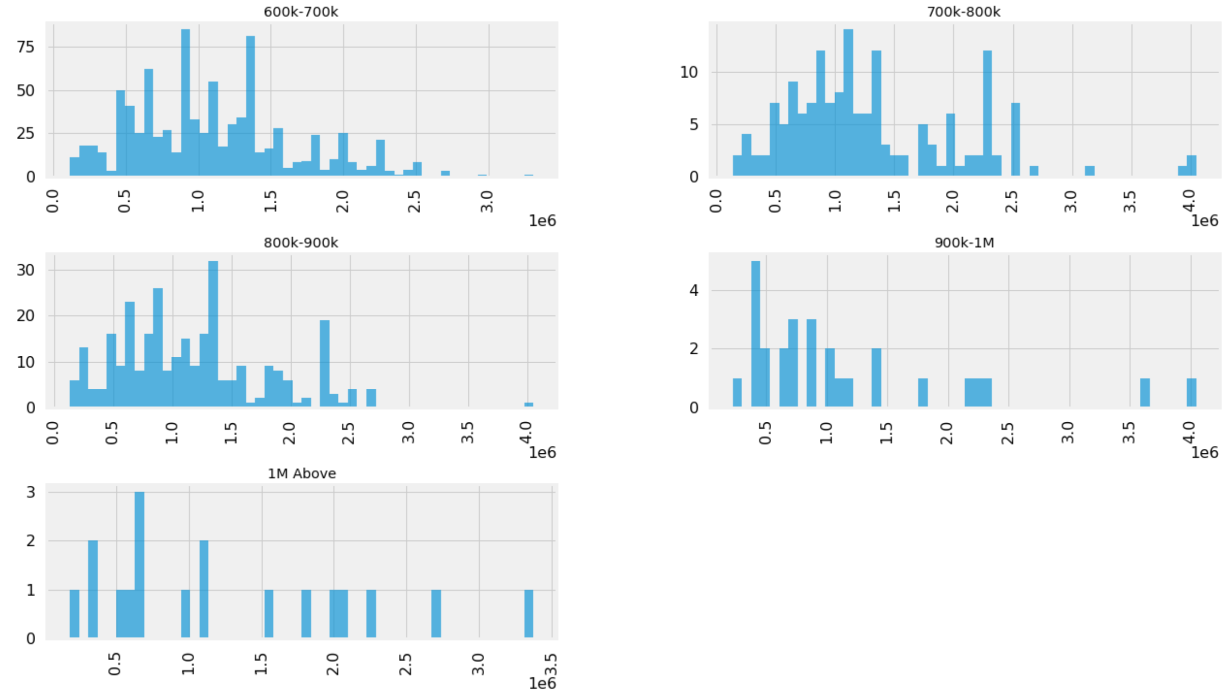
From the following heat map, we see that general information features(Phone number, email, contact details) have no correlation with a person being a defaulter.

3.5. Standardizing values

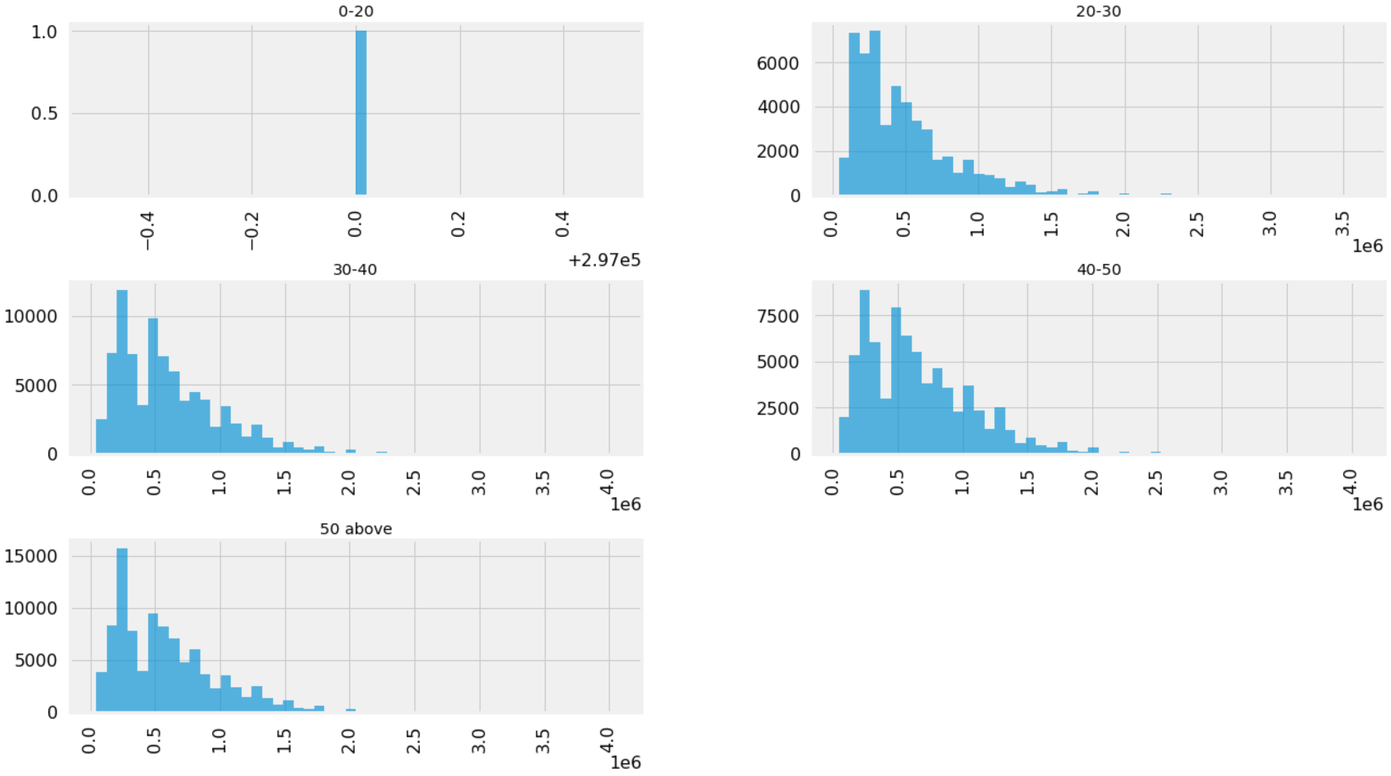
We created bins for AMT_INCOME_RANGE, thus binning numerical columns to create categorical columns.

The graph shows the loan amount requested in each category. Analysing it we found out that more than 50% of the loan applicants have an income in the range of 100k- 200k and almost 92% of the applicants have an income less than 300k.

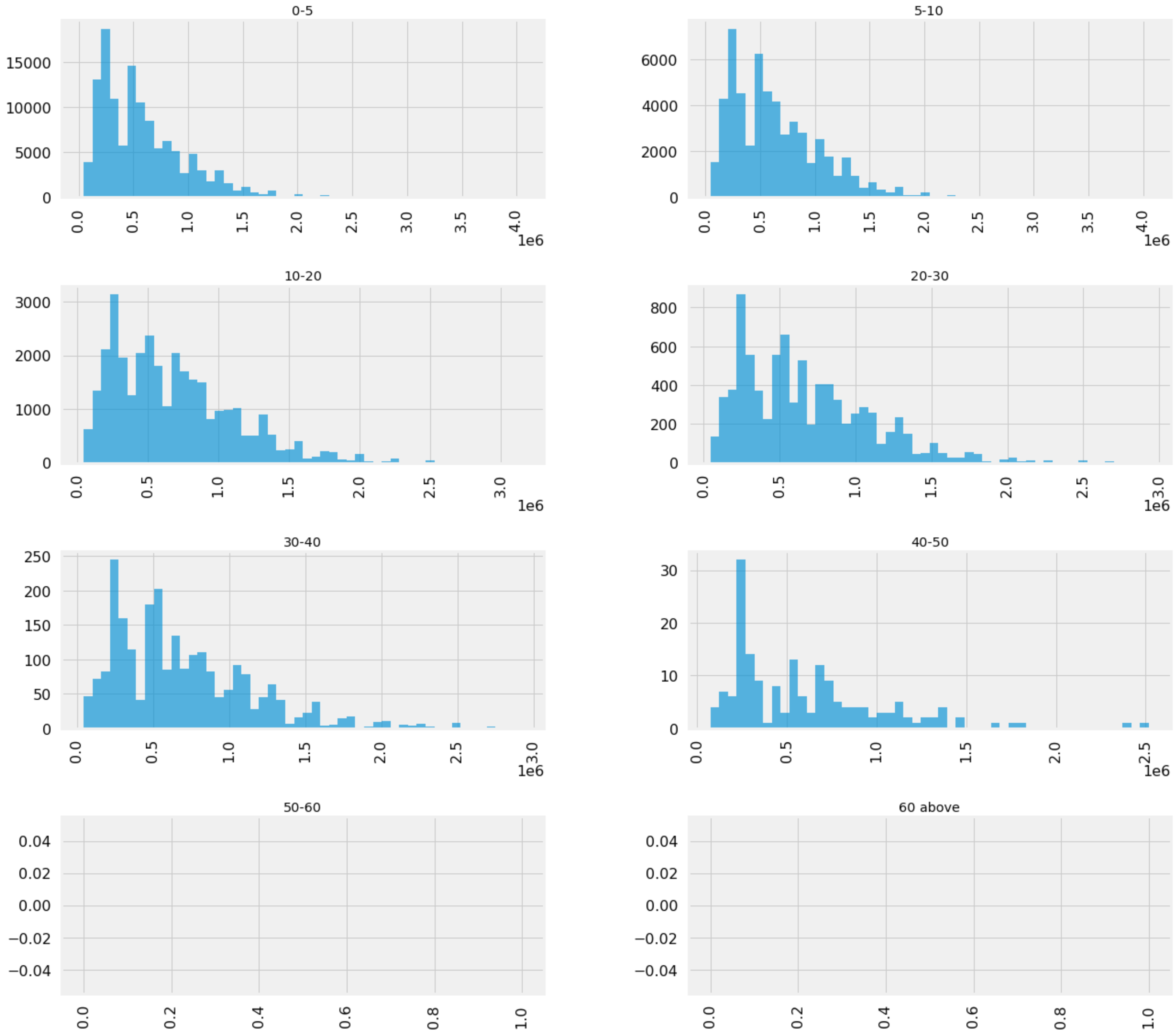




We created bins for AGE_GROUP, and plotted a graph for the same. Upon studying the graph we found that 31% of the loan applicants have an age above 50 and more than 55% of the loan applicants have an age over 40.



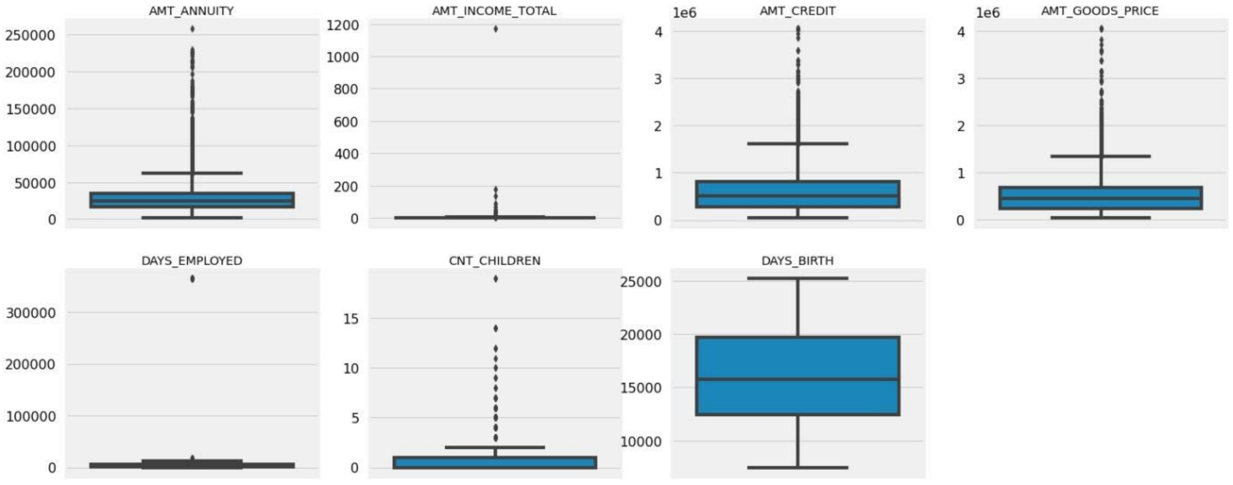
We created bins for EMPLOYMENT_YEAR, and plotted a graph for the same. Upon studying the graph we found that more than 55% of the loan applicants had an experience of 0-5 years and almost 80% of them have less than 10 years of experience.



3.6. Data Type conversion and Null Value data imputation

- First we checked the column type if they are in the correct data type using the obtained standardized values. Further we converted the object and numerical columns into categorical columns.
- We checked the null value percentage of each column in the dataframe.
- To impute null values in categorical variables which have a lower null percentage, we used mode() to impute the most frequent items. The impute categorical variable NAME_TYPE_SUITE has lower null percentage(0.42%)with the most frequent category using mode()[0]
- To impute null values in categorical variables which have higher null percentage, a new category is created. The impute categorical variable OCCUPATION_TYPE has a higher null percentage(31.35%) with a new category as assigning to any existing category might influence the analysis.
- To impute null values in numerical variables which have lower null percentage, we used median as there are no outliers in the columns, which we observed from the result of describe() and the mean returned the decimal values and the columns represents are the number of enquiries made which cannot be decimal.

3.7. Outlier identification



We have plotted a graph to find out information about outliers. On studying the graph obtained we get to know that:

- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE and CNT_CHILDREN have some outliers.

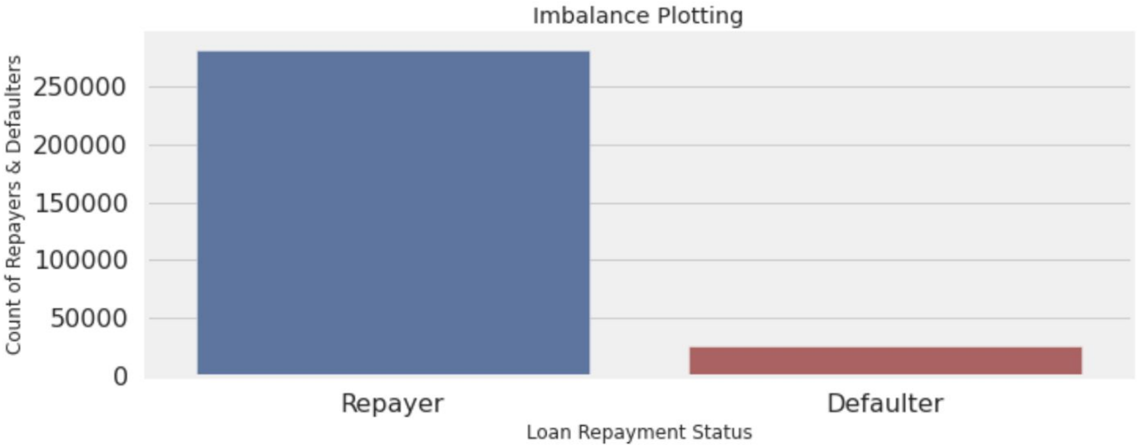
- AMT_INCOME_TOTAL has a huge number of outliers which indicates that few of the loan applicants have high income when compared to the others.
- DAYS_BIRTH has no outliers which means the data available is unreliable.
- DAY_EMPLOYED has outliers values around 350,000 (days) which is around 958 years which is impossible hence this has to be an incorrect entry.

4. Data Analysis

We performed data analysis by plotting various functions and drawing correlations between various fields, and the target variable.

4.1. Imbalanced data

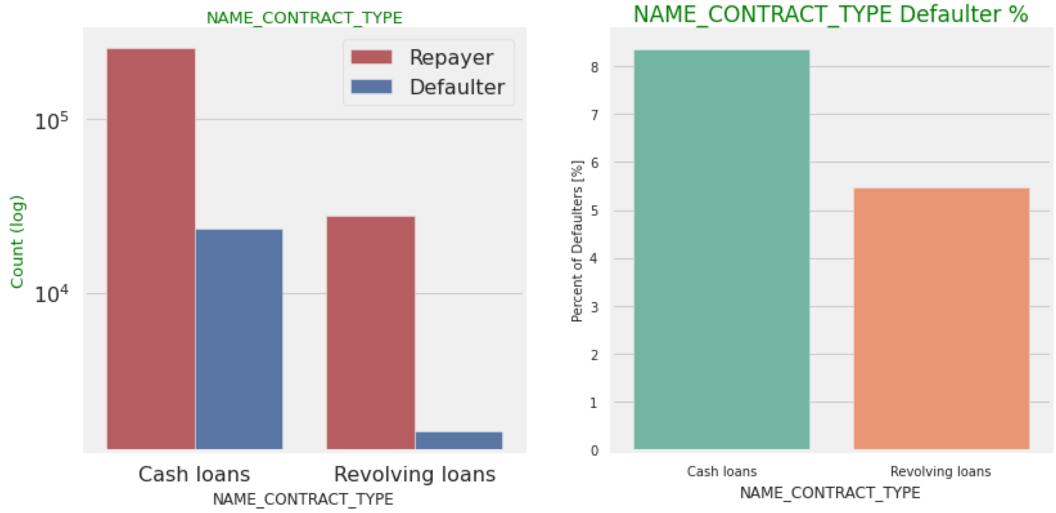
- In this section of Analysis, we plot the Target value (which is a binary data) of our dataset and look for any type of imbalance. After analyzing the “Target” column, we see that the number of re-payers are more as compared to the defaulters.
- This can be clearly visualized through the graph below.



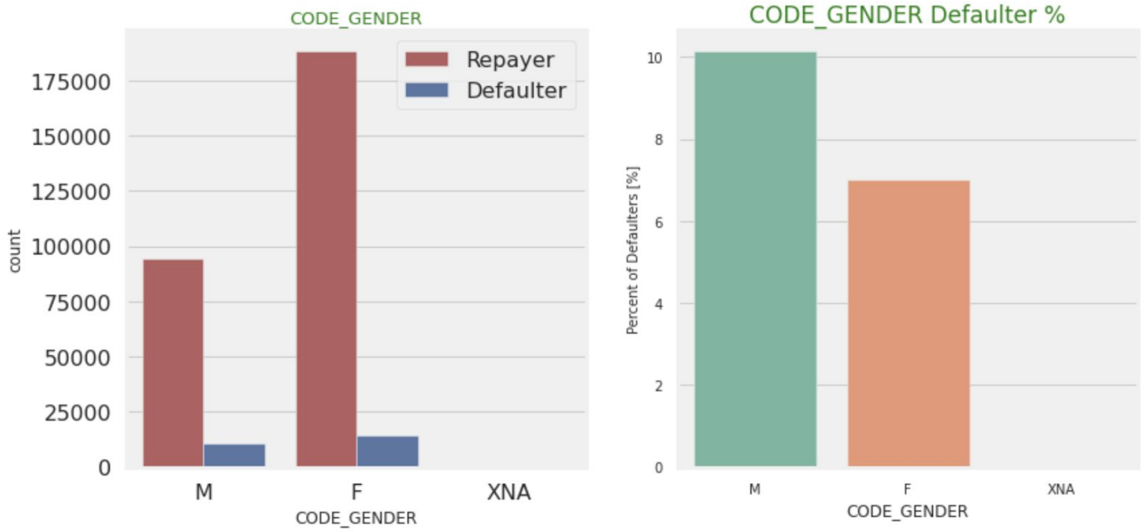
Analysis Drawn : We have calculated the ratios of imbalance with respect to repayers and defaulters data.

4.2. Categorical Variable Analysis

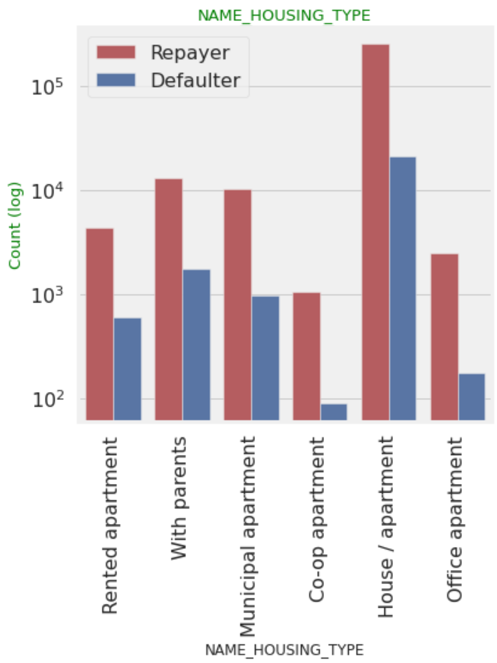
- Below are the univariate categorical plots for columns: [NAME_CONTRACT_TYPE,CODE_GENDER,NAME_HOUSING_TYPE,NAME_EDUCATION_TYPE,NAME_INCOME_TYPE,ORGANIZATION_TYPE,AGE_GROUP]

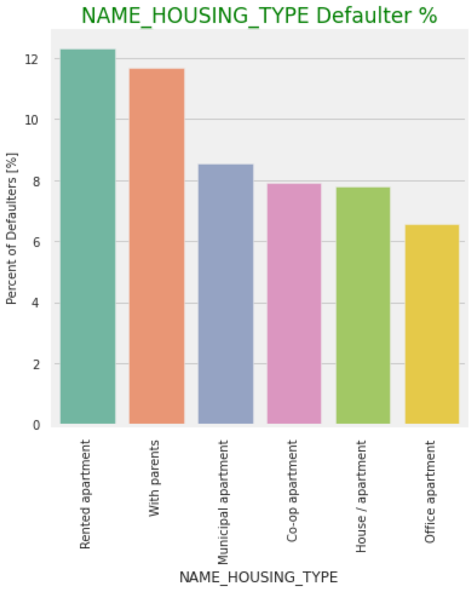


Analysis Drawn : Revolving loans are just a small fraction (~10%) from the total number of loans. In the same time, a larger amount of Revolving loans (comparing with their frequency), are not repaid.

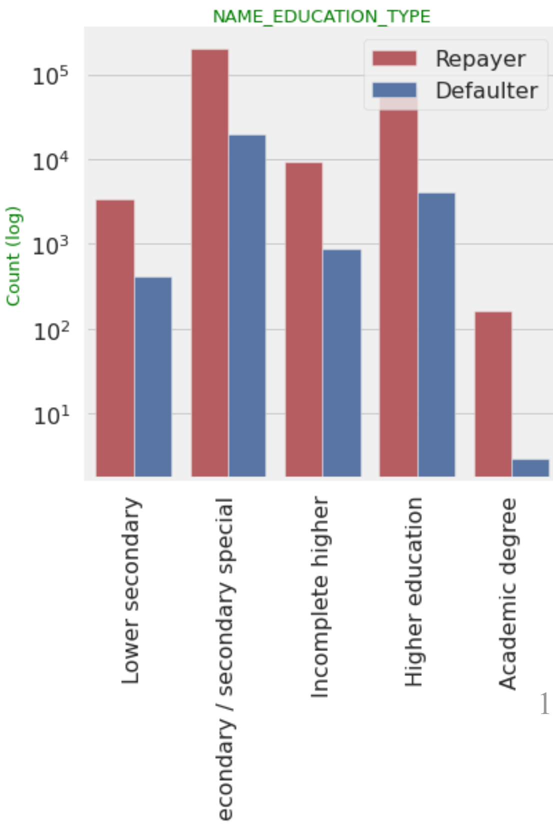


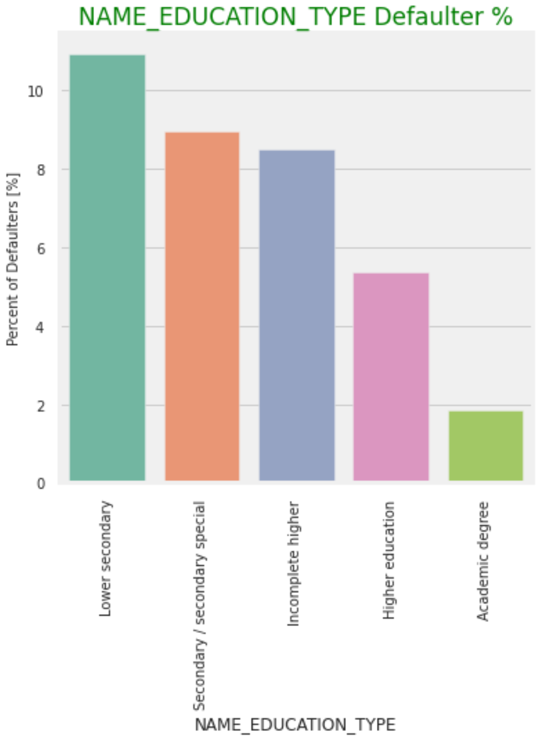
Analysis Drawn : The number of female clients is almost double the number of male clients. If we look at the percentage of defaulted clients, males have a higher chance of not returning their loans (~10%), as compared with women (~7%).



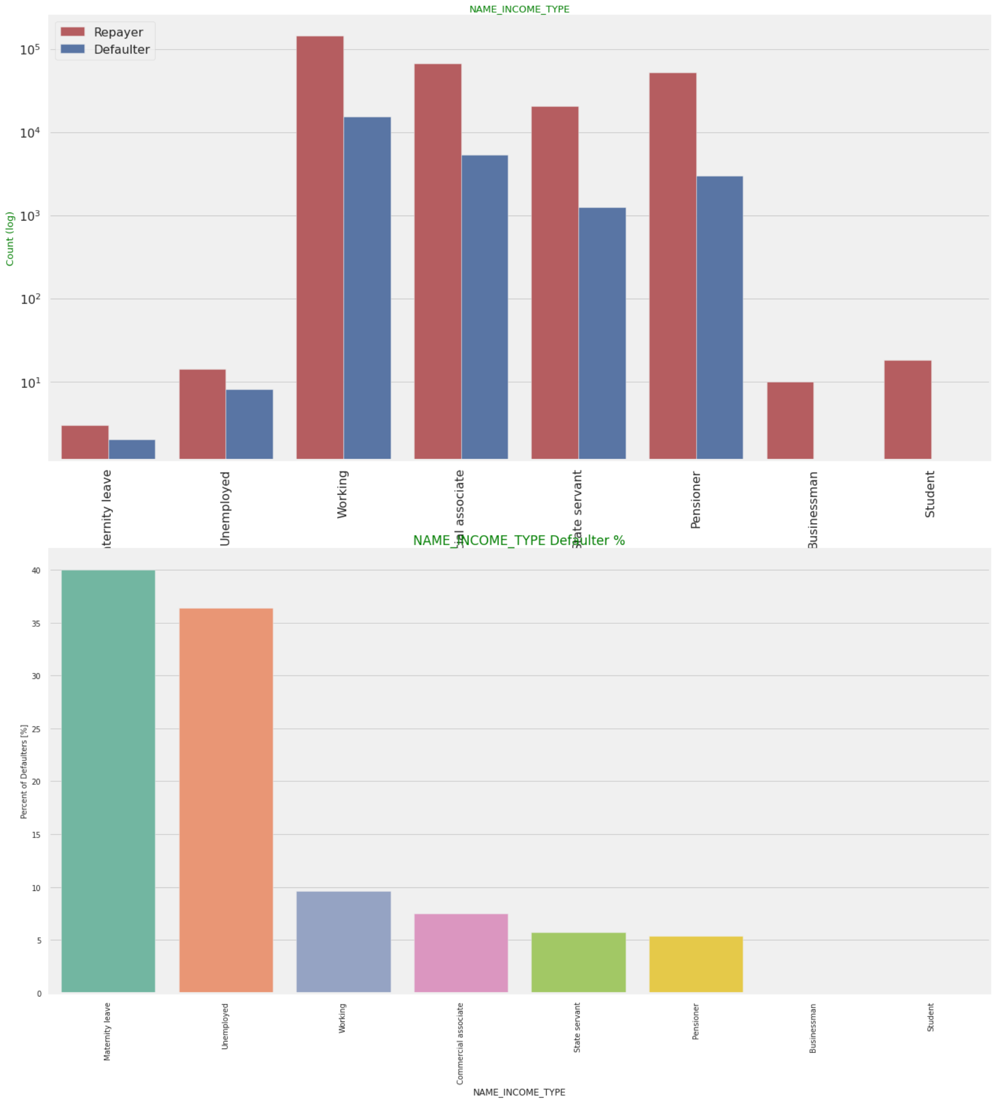


Analysis Drawn : People living in office apartments have the lowest default rate. While people living with parents and living in rented apartments have higher probability of defaulting as most of the people live in houses/apartments.

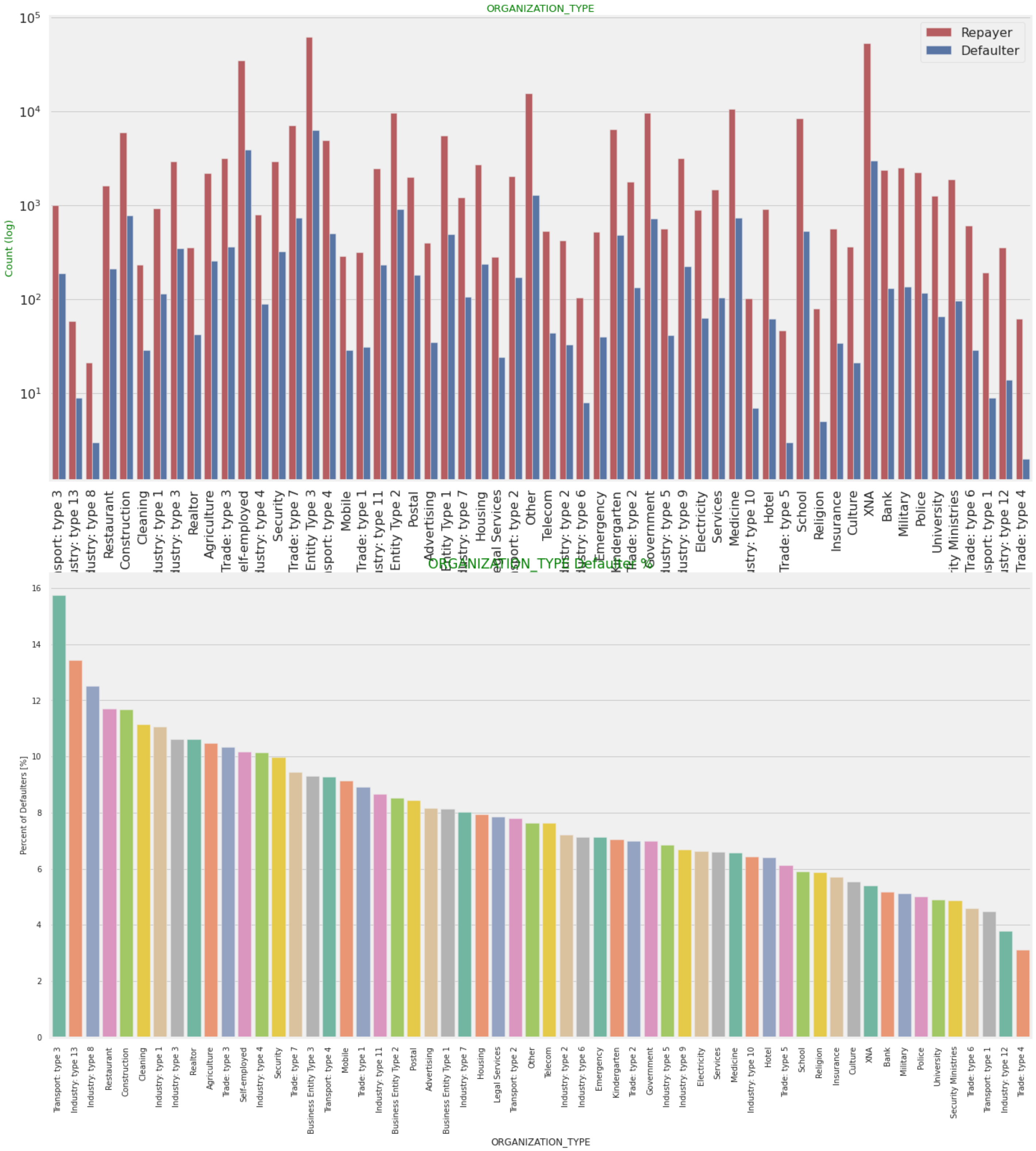




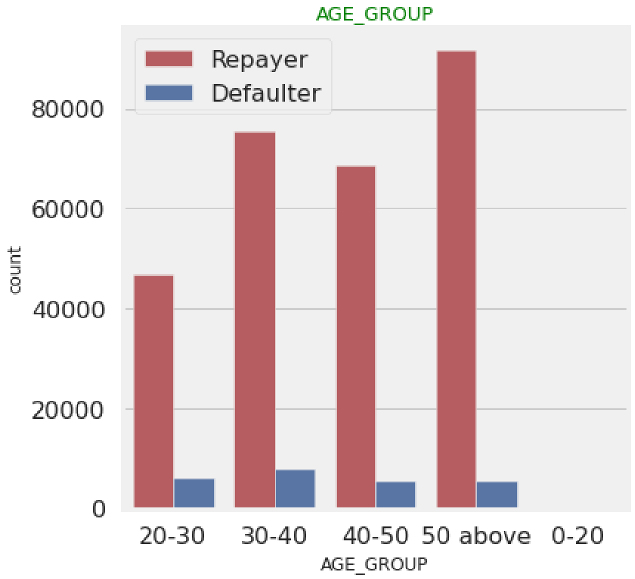
Analysis Drawn : A very small number of clients have an academic degree. The lower secondary category, have the largest rate of not returning the loan. The people with academic degrees have less than 2% defaulting rate.

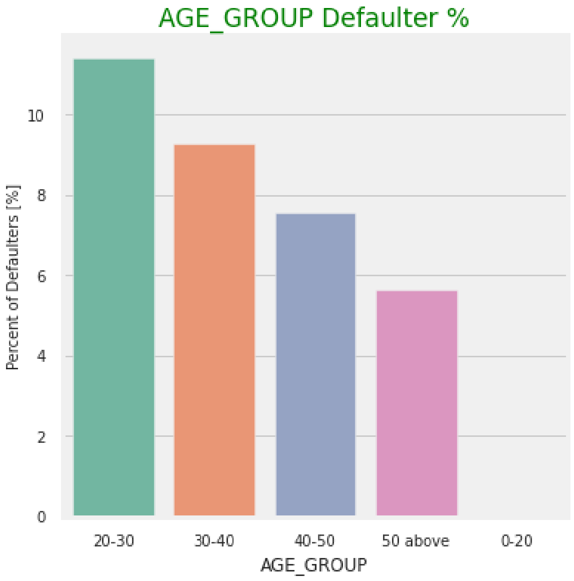


Analysis Drawn : Most applicants who take loans have income type as Working, followed by Commercial associate, Pensioner and State servant. The applicants with the type of income 'Maternity leave' have ~40% ratio of being defaulters in returning the loans, followed by 'Unemployed' (~37%). Remaining types of incomes are under the average ~10% for being defaulters. Students and Businessmen are safest for providing loans as they do not have a high defaulting rate.

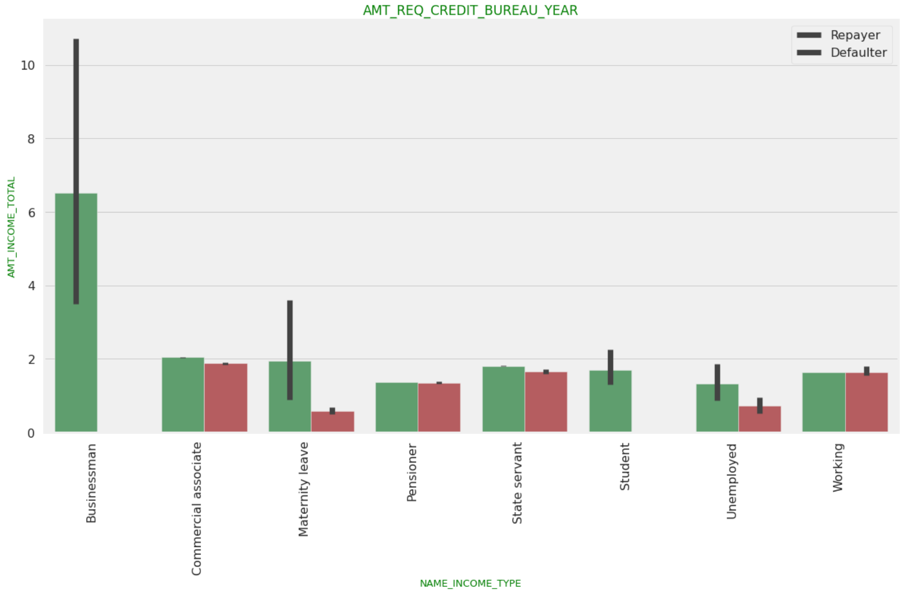

ORGANIZATION_TYPE
Count (log)
10^5
10^4
10^3
10^2
10^1
Repayer
Defaulter
sport: type 3
lustry: type 13
lustry: type 8
Restaurant
Construction
Cleaning
lustry: type 1
lustry: type 3
Realtor
Agriculture
Trade: type 3
elf-employed
lustry: type 4
Security
Trade: type 7
Entity Type 3
sport: type 4
Mobile
Trade: type 1
istry: type 11
Entity Type 2
Postal
Advertising
Entity Type 1
lustry: type 7
Housing
egal Services
port: type 2
Other
Telecom
lustry: type 2
lustry: type 6
Emergency
Kindergarten
Trade: type 2
Government
lustry: type 5
lustry: type 9
Electricity
Services
Medicine
istry: type 10
Hotel
Trade: type 5
School
Religion
Insurance
Culture
XNA
Bank
Military
Police
University
rity Ministries
Trade: type 6
sport: type 1
istry: type 12
Trade: type 4
ORGANIZATION_TYPE Defaulter %
Percent of Defaulters [%]
16
14
12
10
8
6
4
2
0
Transport: type 3
Industry: type 13
Industry: type 8
Restaurant
Construction
Cleaning
Industry: type 1
Industry: type 3
Realtor
Agriculture
Trade: type 3
Self-employed
Industry: type 4
Security
Trade: type 7
Business Entity Type 3
Transport: type 4
Mobile
Trade: type 1
Industry: type 11
Business Entity Type 2
Postal
Advertising
Business Entity Type 1
Industry: type 7
Housing
Legal Services
Transport: type 2
Other
Telecom
Industry: type 2
Industry: type 6
Emergency
Kindergarten
Trade: type 2
Government
Industry: type 5
Industry: type 9
Electricity
Services
Medicine
Industry: type 10
Hotel
Trade: type 5
School
Religion
Insurance
Culture
XNA
Bank
Military
Police
University
Security Ministries
Trade: type 6
Transport: type 1
Industry: type 12
Trade: type 4
ORGANIZATION_TYPE

Analysis Drawn : Most of the loan applicants are from Business Entity Type 3. Self employed people have a very high defaulting rate, thus are not safe for providing loans, also transport, industry and restaurant organizations have high defaulting rates. Trade Type 4 and 5, and Industry type 8 are safest for providing loans.



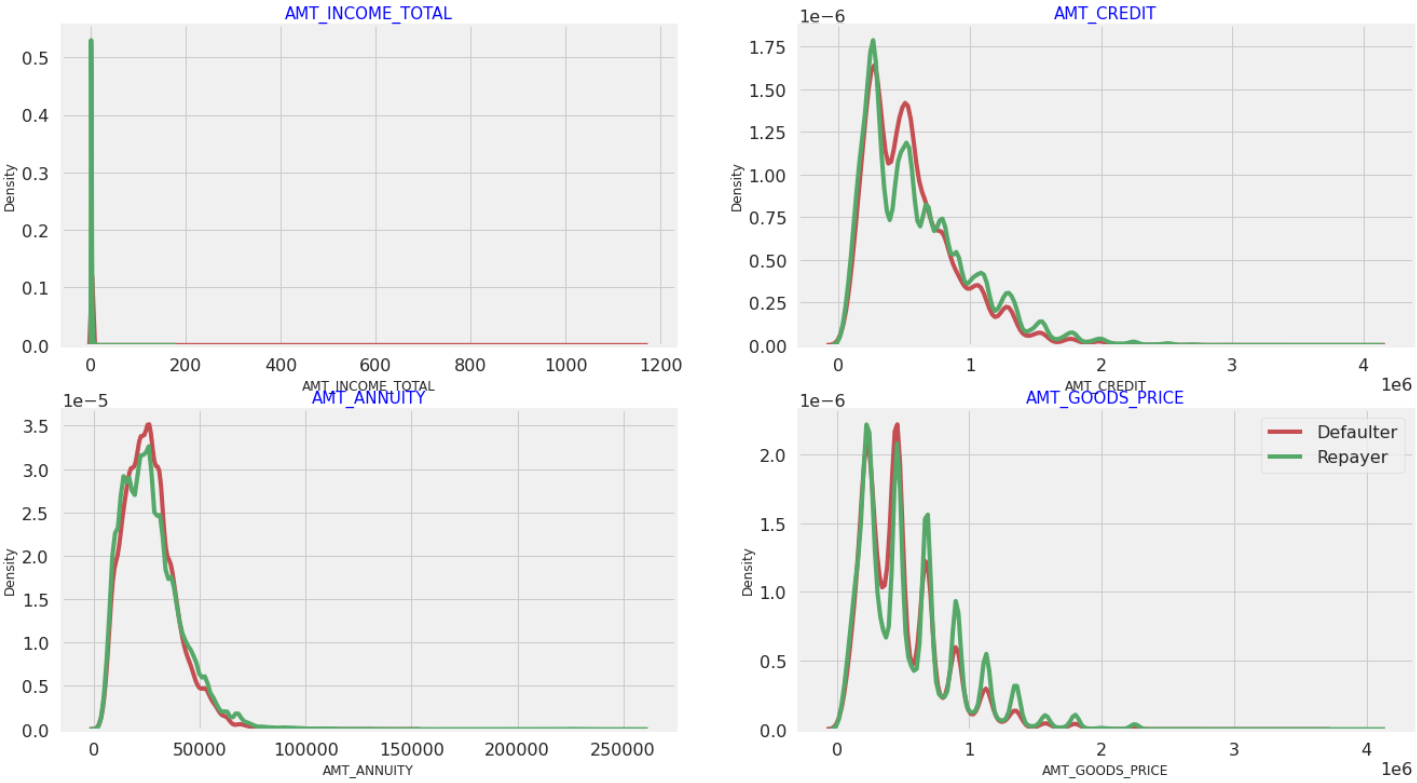


Analysis Drawn : People in the age group of 20-40 have a higher rate of defaulting while people with age > 50 have low rate of defaulting.

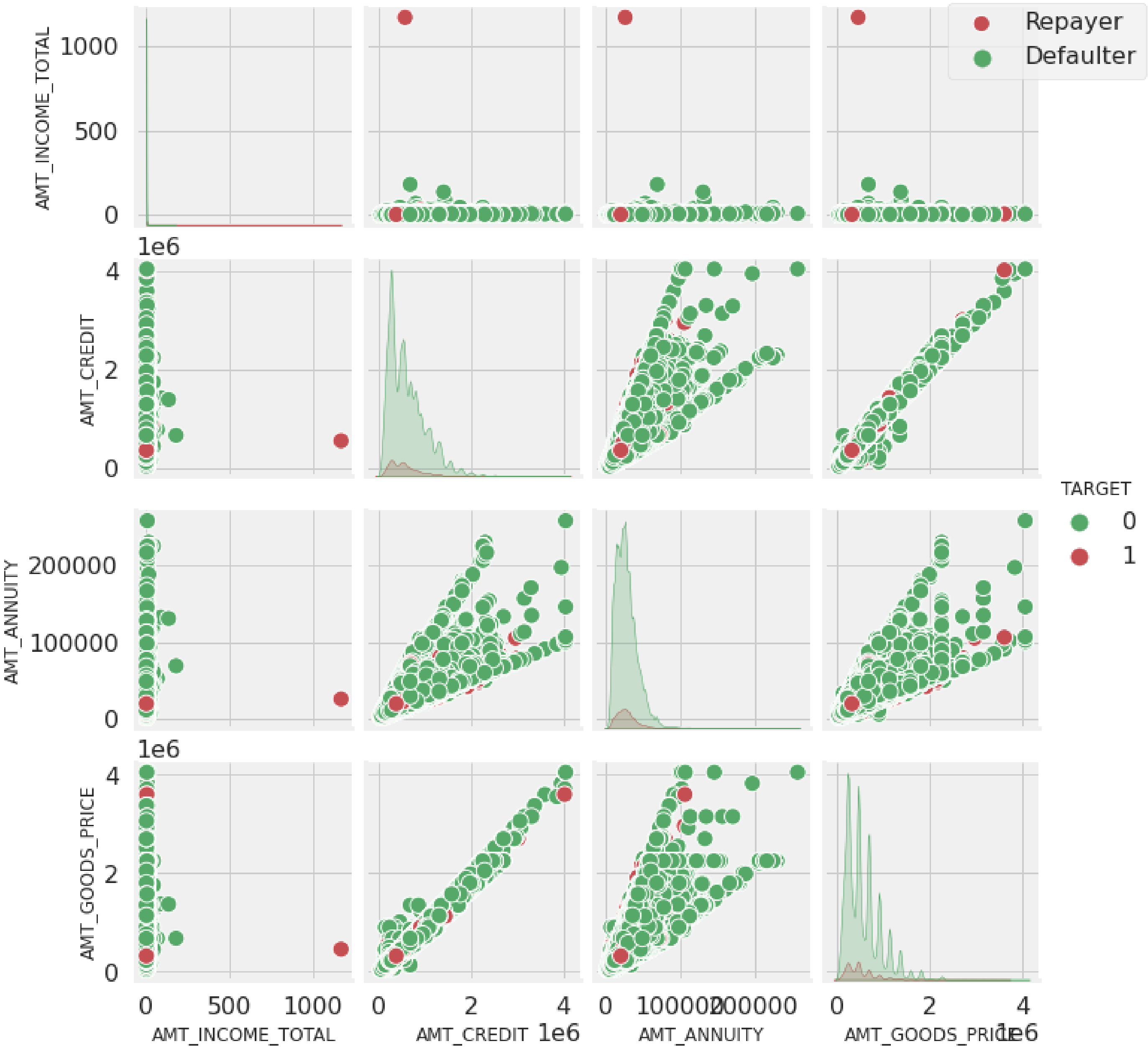


- Below are the multivariate plots for columns NAME_INCOME_TYPE,AMT_INCOME_TOTAL, 'AMT_INCOME_TOTAL', 'AMT_CREDIT','AMT_ANNUITY', 'AMT_GOODS_PRICE', 'AMT_GOODS_PRICE','AMT_CREDIT'.

Analysis Drawn : We saw that business man's income is the highest and confidence of greater than 95% of repaying the loan, income of a business man could be in the range of $ 400,000 (4 hundred thousand) - $ 1,000,000 (1 million).



Analysis Drawn : Most of the loans are given for value less than $1,000,000 (1 million), credit amount of loan is less than $1,000,000 (1 million). As we can see from the plots, repayers and defaulters distribution overlap and hence, they can not be used for decision making.



Analysis Drawn : We observe that if amt_annuity >15000, amt_goods_price> 3M, there are lesser chances of defaulters. The number of defaulters for AMT_CREDIT > 3M are very less.

5. Dimensionality Investigation

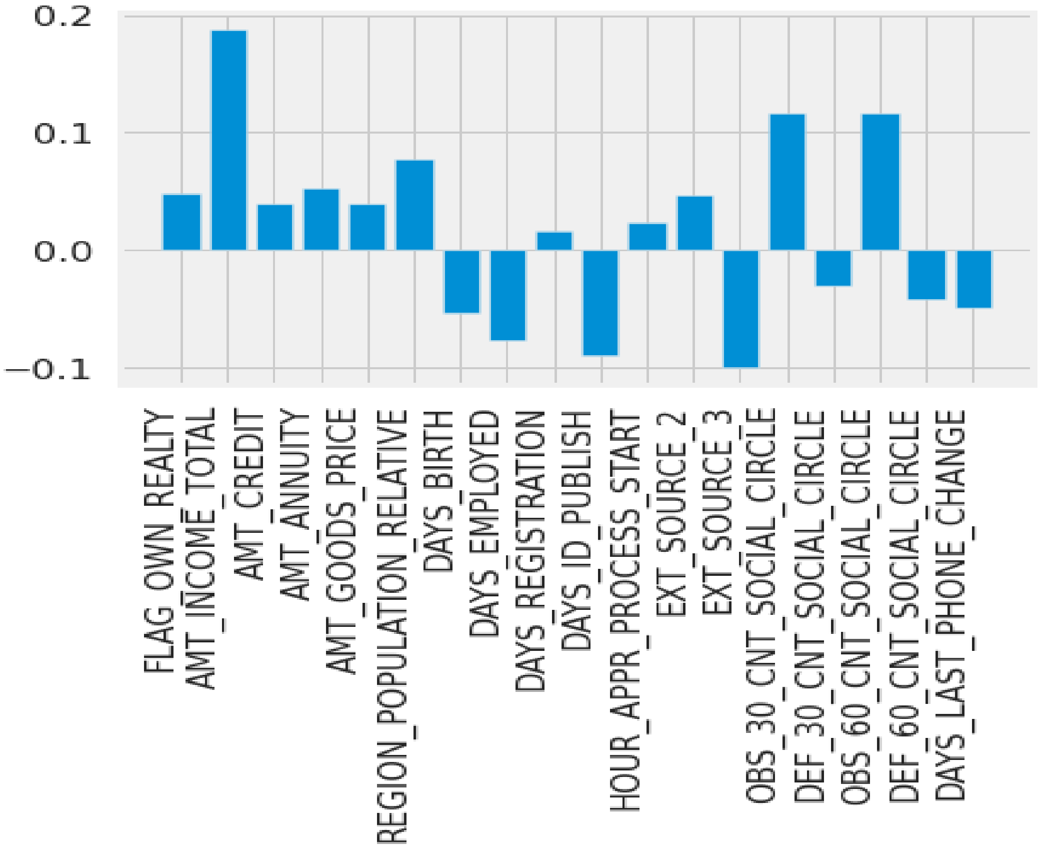
5.1. Principal Component Analysis (PCA)

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

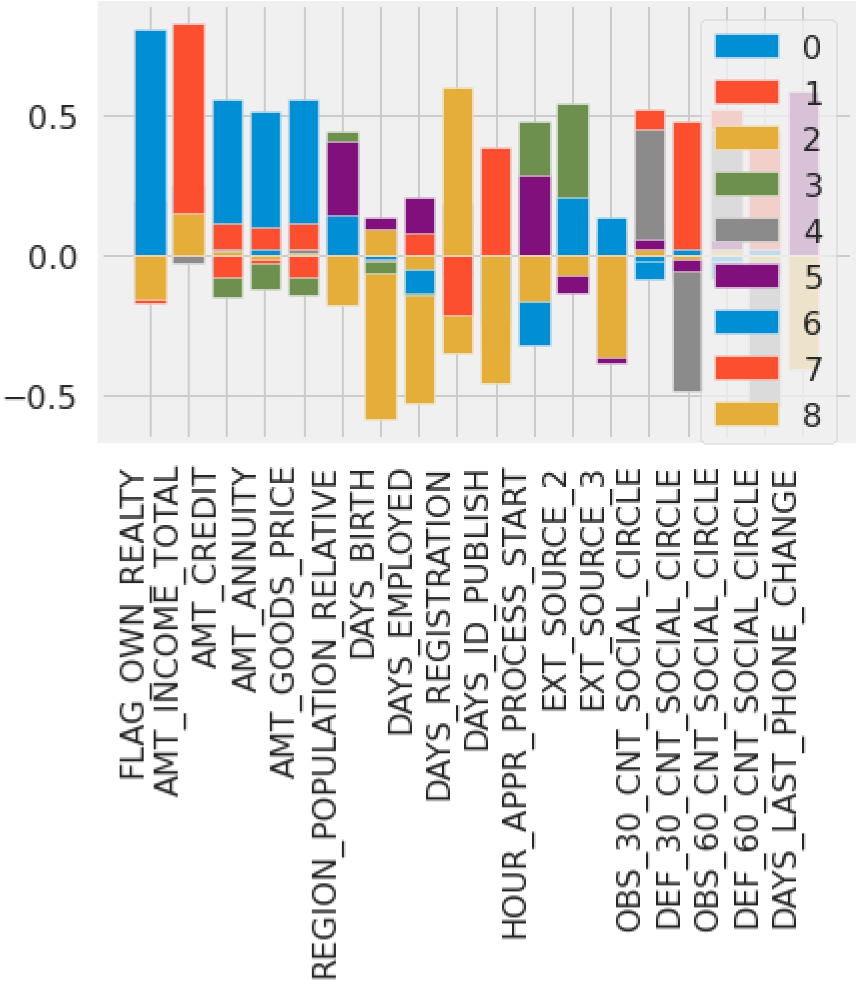
- For this section, we have picked 20 numerical features to reduce the dimensionality of our problem.
- We will be reducing 20 features to 4 PCA features.

5.2. Observations and Results

Analysis Drawn : We see that OBS_60_CNT_SOCIAL_CIRCLE (How many observation of client's social surroundings with observable 60 DPD (days past due) default) and



OBS_30_CNT_SOCIAL_CIRCLE (How many observation of client's social surroundings with observable 60 DPD (days past due) default) are more positively correlated with our target variable i.e. loan granted or not. However 'DAYS_ID_PUBLISH' (How many days before the application did the client change the identity document with which he applied) and 'EXT_SOURCE_2' (Normalized score from external data source) are negatively correlated to our Target variable. We can ignore the rest of the columns as they do not provide much information as compared to ones above.



Analysis Drawn : With 50% reduction in dimensions, we managed to extract 80% of the information.With 9 PCA components, we managed to capture about 80% of information from 20 features.

6. Clustering Algorithm Exploration

In this section, we will explore similar groups of users in the data with 2 unsupervised clustering algorithms: K-Nearest Neighbors (KNN), and K-Means clustering (K-Means).

6.1. Preprocessing for Clustering

Before applying clustering, categorical data in the dataset has to be transformed into numerics for compatibility with the distance metrics. A correctly established numerical representation of the categories will prevent information loss in the transformation and produce more accurate results. To do so, the categorical data are divided into 3 different types with which each treated differently. The 3 types are as follows:

6.1.1. Binary categorical data

Binary categorical data are transformed into 1s and 0s through a predefined dictionary. An example of such data would be the feature "NAME_CONTRACT_TYPE" and "FLAG_OWN_CAR".

6.1.2. Ordinal categorical data

Ordinal categorical data have an innate order of its categories. When transforming these features into numerical data, it is important to inherent the order of these categories. Then each category is ranked from lowest to highest, and mapped into non-negative integers correspondingly. An example would be "CODE_GENDER".

6.1.3. Non-ordinal categorical data

As the counterpart of ordinal categorical data, the non-ordinal data does not have a clear innate order among its categories. In this case, a simple integer mapping will induce false distance between the categories. To solve this problem, non-ordinal categorical data are treated with one-hot encoding, where each category is taken out as one binary feature. For example, "NAME_FAMILY_STATUS" feature will be divided into several columns where they are named after the possible values in the original column. Thus, the distance between categories will be evenly distributed.

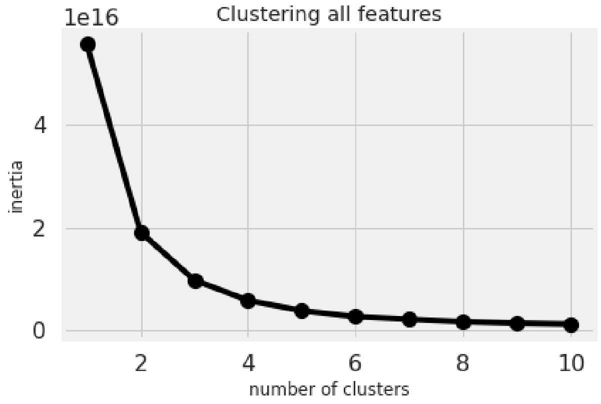
6.2. K-nearest neighbors (KNN)

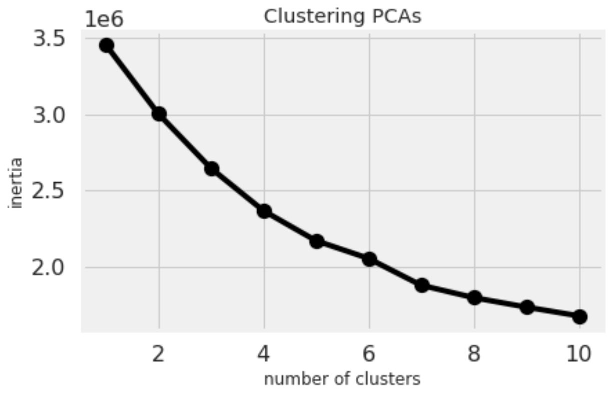
The first approach of clustering is k-nearest neighbors clustering. In this approach, the distance between 2 data points are measured with Euclidean distance (L2), and the number of neighbors to consider is set to 10000. This yields a total of 18 clusters. The 2 model parameters are exposed in the code, and can be tuned in the future for specific interests.

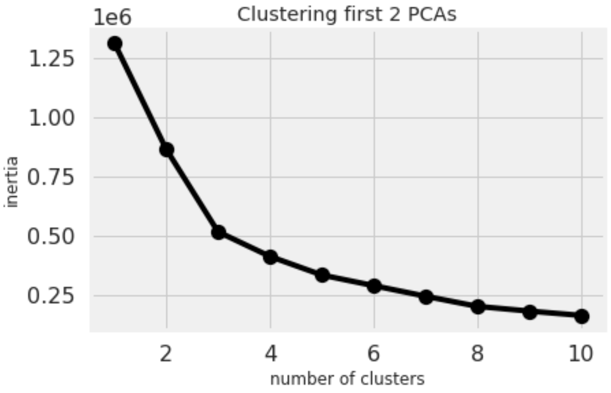
To better understand the distances between data points in the clusters, we attempted to extract distance information from using sklearn kneighbors library, but it turned out undoable due to the restriction of the RAM.

6.3. K-means clustering

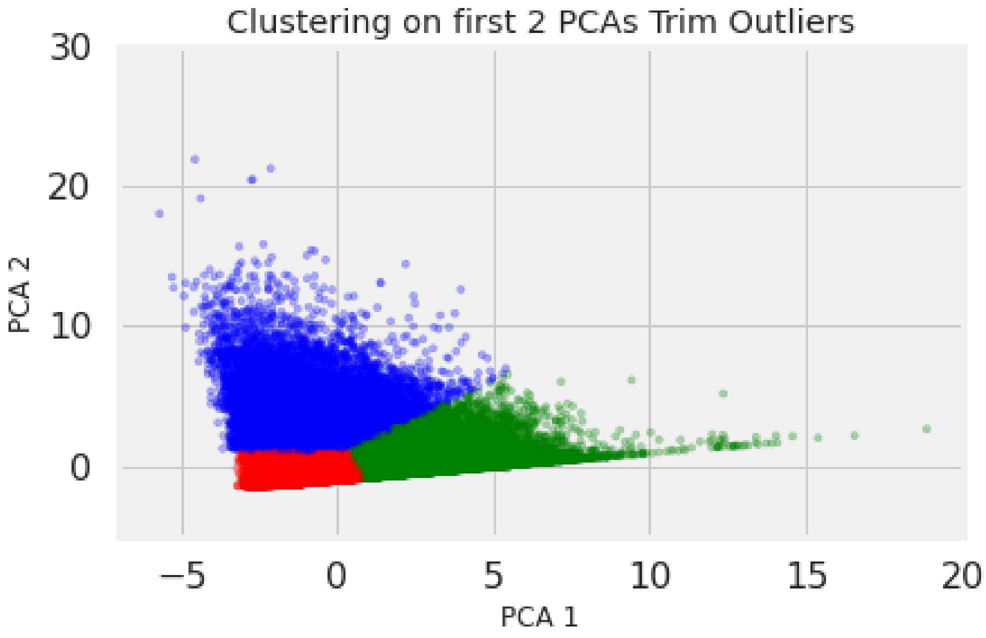
In the K-means clustering approach, the algorithm learns the centers of clusters by minimizing the mean distance within each cluster. The model takes K, the number of clusters to infer, as the parameter to construct the clusters. When clustering with all features included or the first 2 PCAs only, k = 3 produces clusters with the most plausible inertia reduction. Therefore the same clustering algorithm with k = 3 is used on both all features and the PCAs.







From k-means clustering, an array of integer labels will be generated. These labels indicate the learned cluster labels for each data point in the dataset. By grouping the data by labels, will yield data frames that resemble similar customers, which will be useful in more targeted analysis.



7. Covid Study

Our aim is to find safe times in a week when a person can visit the banks for loan applications with the lowest risk of being affected with COVID-19.

We perform this analysis via a heatmap where we take 2 columns - WEEKDAY_APPR_PROCESS_START and HOUR_APPR_PROCESS_START.

Steps involved :

- We groupby the 2 columns - WEEKDAY_APPR_PROCESS_START and HOUR_APPR_PROCESS_START and make a 2D matrix of it.
- We finally convert this matrix into a heatmap by analyzing the traffic on all days with all combinations of times.
- We set a slider as a reference which shows the foot traffic in the range of 1000-6000, 100 being less populated and 6000 being highly populated.

Analysis Drawn: By plotting the heatmap, we draw the following conclusions :

- On all the working days and working hours, banks are most crowded between 10 am to 1 pm.
- On all the working days and working hours, banks are relatively less crowded at 9 am and 4 pm - 5 pm.
- Considering all the days and all the times, we can observe that out of all the working days, banks are most crowded on monday and thus, mondays are unsafe to visit banks.
- We can observe that people generally prefer going to banks in-person and applying for loans as compared to online loan applications.

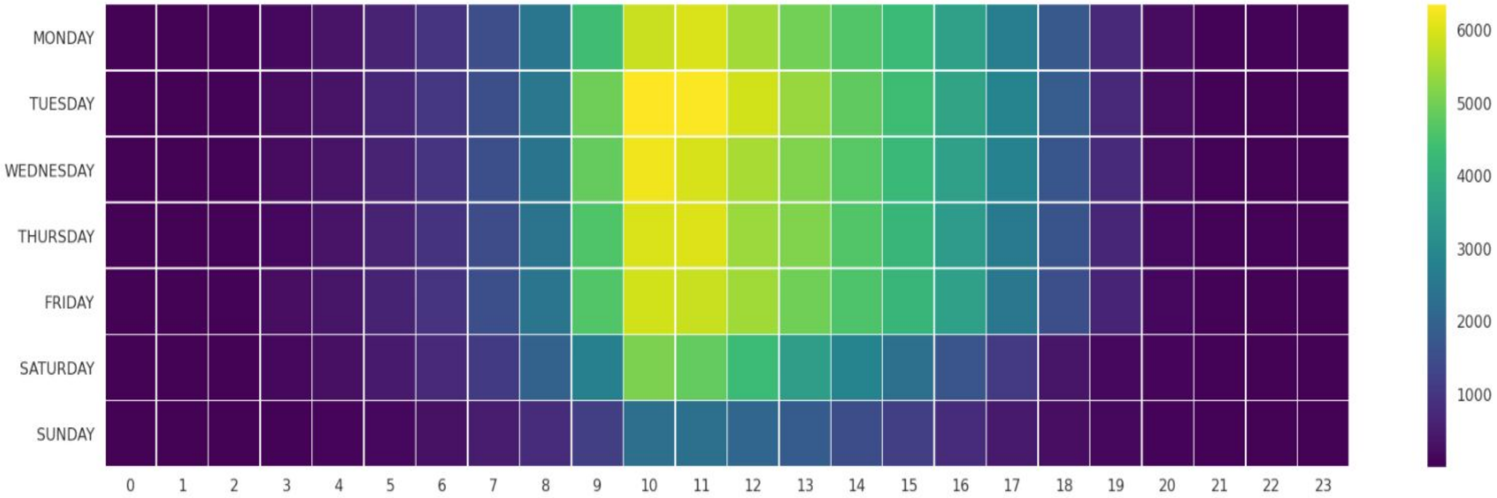


Fig 1

8. Conclusion

a. **Data Analysis** : We performed analysis on our dataset using various plots and we were able to do the following :
   - identified all the correlations between different fields in the dataset.
   - identified all the important fields that had an impact in the decision making.
   - identified the type of people who take the most number of loans, people in various subcategories who'd be able to repay their loans
   - identified the favourable set of people who'd be able to repay the loans and could be targeted by banks.

b. **Principal Component Analysis :** By using PCA we were successfully able to reduce the dimensionality of our dataset. By considering 20 columns containing only numerical data we reduce the dimension from to 4. These 4 columns are highly correlated to our Target variable, i.e. whether the bank approves loan for a particular customer or not.

c. Clustering : Using KNN and K-means clustering methods on further processed data, we were able to partition the users into groups with most similarities. These labels learnt from these 2 unsupervised algorithms will also allow analysis in a specific target user group.

d. **Heat Maps**: From the heat maps, we learn which times are preferred by customers to visit the bank.

- In the light of recent events, this can be used to advise people about the potential visiting hours, so that people feel safe considering that the foot traffic is less during those suggested hours.
- Moreover, we observe that very few people apply for loans on Sundays or after working hours. This clearly indicated that people still prefer going to banks for loan applications. This could be indicative of the process being too confusing for customers to do themselves.
- Banks can use this information to make the online application process much easier or launching an ad campaign that makes the customers feel more secure about going to the bank. Thereby, increasing the number of customers.

9. Future Scope

We can use multiple correspondence (MCA) analysis for dimensionality reduction instead of principal component analysis. In statistics, multiple correspondence analysis is a data analysis technique for nominal categorical data, used to detect and represent underlying structures in a data set.

In the preprocessing process, we discarded features with over 40% null values, and imputed the null values in other features with their means. However, there could be valuable information that could be mined out of these discarded columns in our analysis. As the database will grow increasingly larger in the future, the newly populated rows will potentially be with better quality. As a result, the proportion of null values will likely drop to a usable level in the currently discarded features. Thus, the analysis will be able to include more features of potential importance and boost the authenticity and amount of the information mined.

As our project preprocessed and analyzed the bank dataset, our results could be easily deployed as the baseline for classification models, with which banks will be able to determine the chances of a person missing a payment on their loan and better understand the customer's repayment capabilities. Models such as decision trees, gradient boosting could be plausible immediate next steps.

10. References

- Loan Defaulter Dataset provided by Gaurav Dutta.
- "A One-Stop Shop for Principal Component Analysis" is a towards data science article by Matt Brems.
- "PCA using Python" is an article by Michael Galarnyk.
- https://stackoverflow.com/questions/2397141/ -- Python 2D Arrays
- https://seaborn.pydata.org/generated/seaborn.heatmap.html -- Python Heat Map
- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html -- Sklearn PCA
- https://scikit-learn.org/stable/modules/clustering.html -- Clustering
- https://www.google.com/search?q=pyplot+python&rlz=1C5CHFA_enUS925US925&oq=pyplot+python&aqs=chrome..69i57j0l7.3665j0j4&sourceid=chrome&ie=UTF- - Python PyPlot