Exploratory Data Analysis
Bank Loan Data

Submitted by – Gaziya Gani Khan

Course - M.S. DS

Approach of Analysis

- Problem statement:

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history hence use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

The dataset 3 files :

- *'application_data.csv"*
- *‘ previous_application.csv‘*
- *'columns_description.csv'*

Exploratory data analysis: EDA refers to a set of process of initial investigation in order to spot anomalies, to test hypothesis with the help of summary statistics and graphical representation.

The EDA of bank loan dataset involves following steps:

- **Data Reading Process:** This is a data reading step where application data in csv file format has been uploaded on Google colab interpreter. The data set has 307511 rows and 119 columns having observations in multiple data types as float, integer and object format.

- **Fixing rows and columns**: In this step unnecessary rows and columns have been deleted. After reading the data dictionary it seems as few columns irreverent for the analysis, thus delete unnecessary colums with name as EXT_SOURCE_1','EXT_SOURCE_2','EXT_SOURCE_3.

• **Data Cleaning:** Data cleaning is a very necessary and important step of exploratory data analysis where data is being prepared in order to proceed further with a error free data.
This data set has large number of observations and having more than 40 percent missing data or incorrect data in some columns which can negatively influence the result of analysis. To clean this data set a standard missing values limit has been set which is 40 percent. Columns with more than 40 percent of missing values is deleted as imputing more than 40 percent missing values will definitely impact the accuracy of analysis.

There are total 48 columns having more than 40 percent of missing values. Thus these columns is being deleted.
There are other columns as well with less than 40 percent missing values which will be handled.

- Imputing/Removing missing values: There are rows where missing value is missing completely at random thus deleting these row . There are some missing values which are missing at random such as where amount of goods price is missing they belong to revolving sub category of loan type. However banks allows borrowers to transfer money in other bank amount with no apparent reason and credit card loans are most famous example of such loan type where loan has been provided over no goods. This lead to impute missing values with the median value of the column.
- same approach has been followed to impute missing values in other columns with appropriate statistical values.

  There are rows with missing values in similar columns hence delete these rows.

• **Handling incorrect data types:** There are multiple types of data types available in the data set which are not suitable for further analysis.
Following are some of variables:
-DAYS_REGISTRATION – changed to absolute value
- EMPLOYED – binned into different slots etc.

• **Handling Outliers**: Outliers are those extreme values which are far beyond the rest of the data set. There are several ways to identify outliers. In application data set outlier detection has been made using describe function and plotting box plot of some variables.

Following are columns:

- CNT_Children : Some values beyond upper fence are continuous as having maximum 10 children can be accepted in some exceptions but more than ten such as 19 children is not factual this proved that these are global outliers .

-Amount_Credit: This Colum has some outliers which are continuous to certain extent which cannot be treated as outliers but after that there are some extreme values as well which are contextual outliers because applicants with low to medium income cannot apply for loan for very high amount where the credit-income ratio is beyond the limit set bank and people with higher income doesn't apply for loans hence extreme loan amount has been targeted as outlier.

- AMT_ANNUNITY: This column has outliers but as they continuous to some extent hence will be included into normal value through binning but after that are still some global outliers that seem un factual as higher annuity means higher debt-to-income ratio which is not usual.

- Days_Employed: This column has global outlier as it is far beyond the other data points and no one can employed for that high number of days.

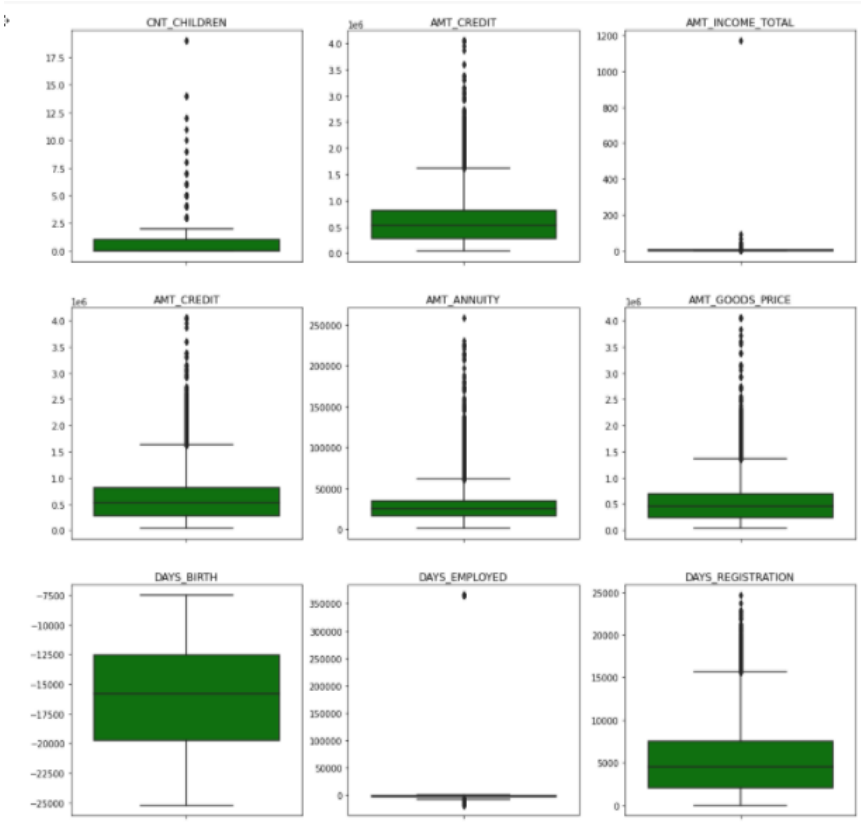
- Days_Registration: this column has contextual outliers as generally people register 1-2 days before for a loan application

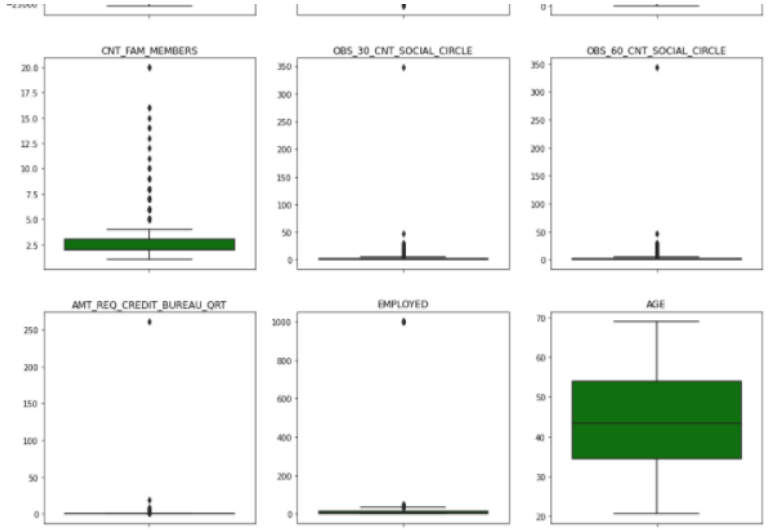
- Days_FAM_MEMBERS: This columns continuous outlier values that can be binned but there are some global outliers as well which should be deleted.

- OBS_30CNT_CIRCLE & OBS_60CNT_CIRCLE : These column have outlier as some values are beyond the range of distant limit.

- AMT_REQ_CREDIT_BUREAU_QRT: This Colum has global outliers because some observations are much beyond the other data points.

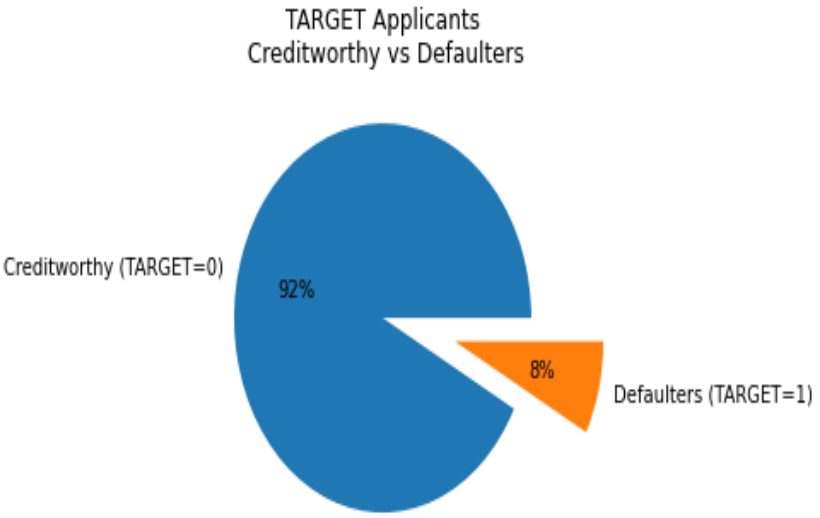
Following are the outliers in bank dataset.

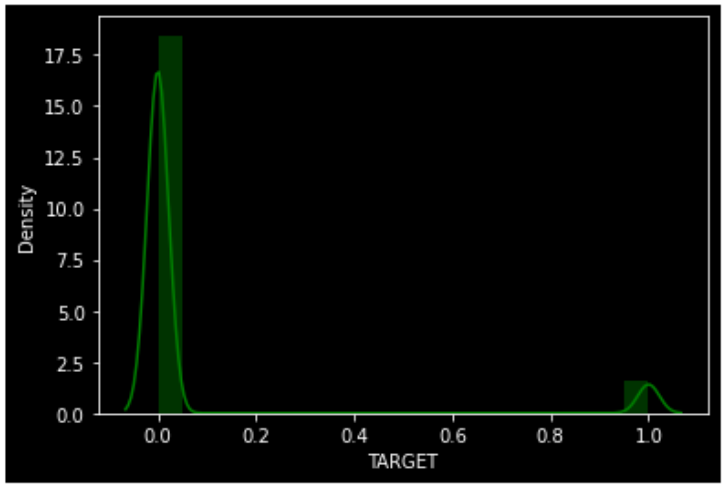


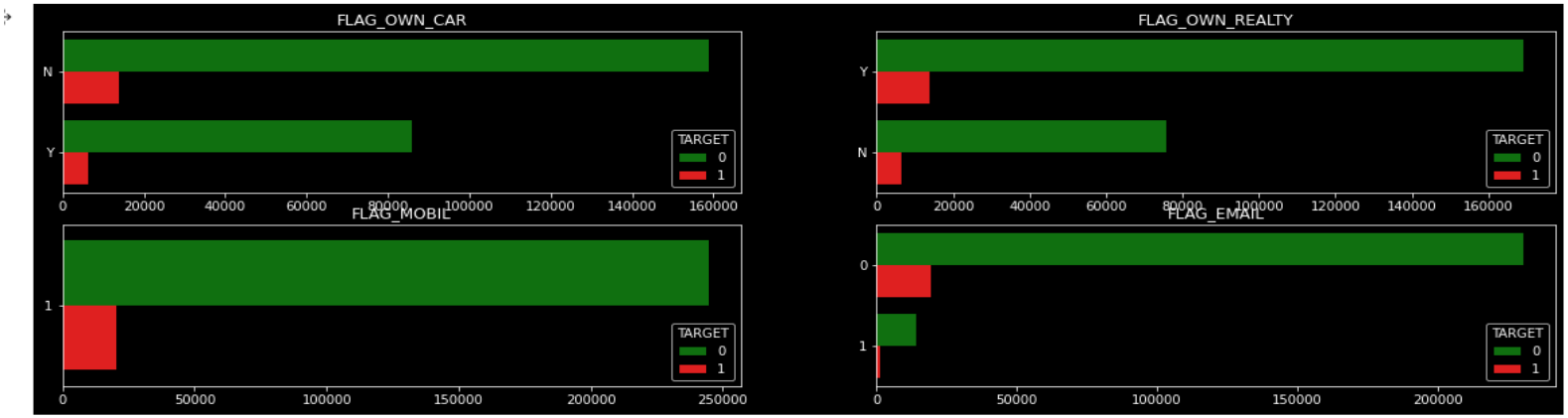


- **Imbalance Data Check:** In order to check the whether the data is imbalance with respect to Target column, the percentage of target column values have been counted which shown that majority of data in the dataset belongs non defaulter applicants. The percent of non defaulters is 98 where as percent of defaulters is 2. Re sampling is an option to deal with imbalance dataset. can be done to test the accuracy of data but that is quite time consuming. In order to analyze the imbalanced data the dataset should be divided into two halves and then analyzed separately.

Following are the figures speaking about imbalance dataset.

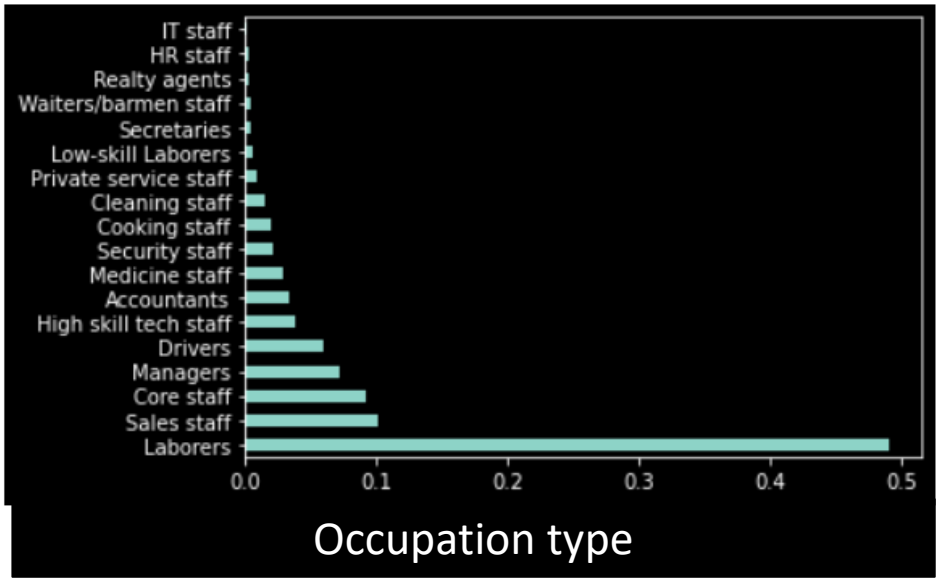




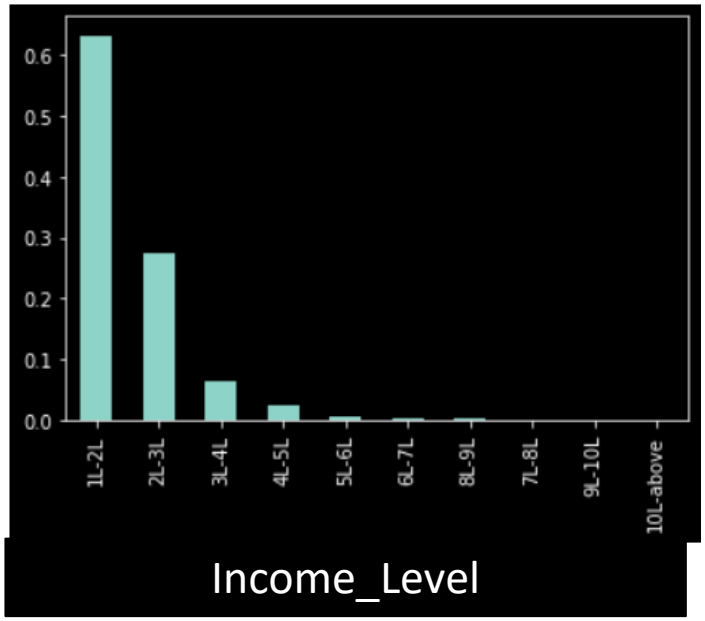


These graphics representing the uneven distribution of data in the dataset.

• Univariate Analysis: Univariate analysis is the process of analyzing each feature of dataset individually in order to identify its distribution.



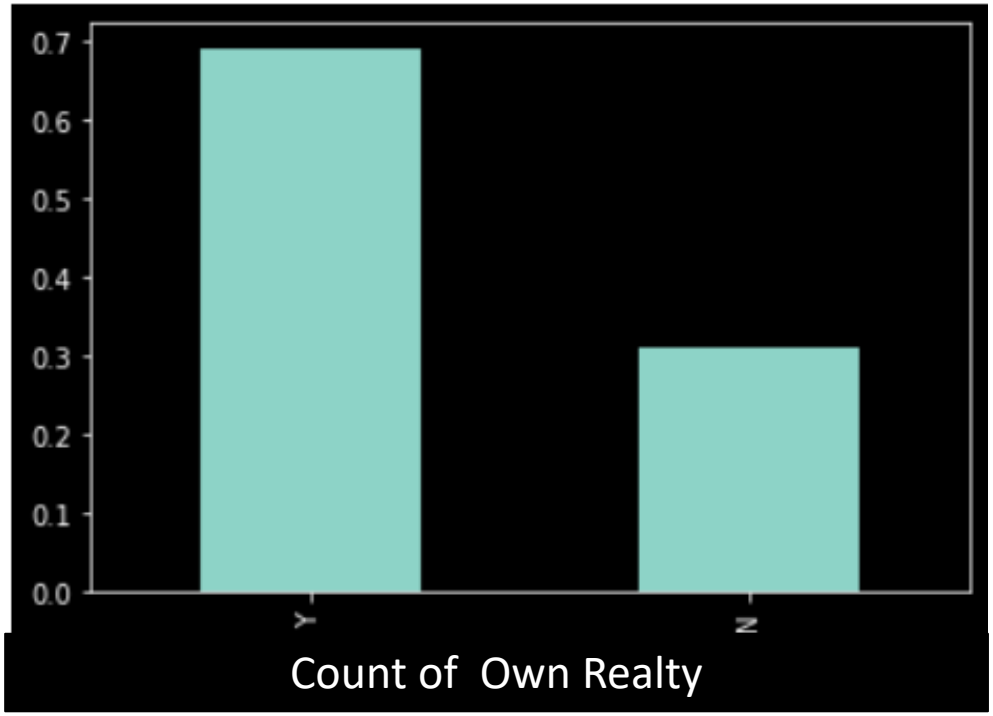
Occupation type



Income_Level

•maximum number of loan applicants involve in low paying job type
•applications for loan are decreasing with the increase in high paying job

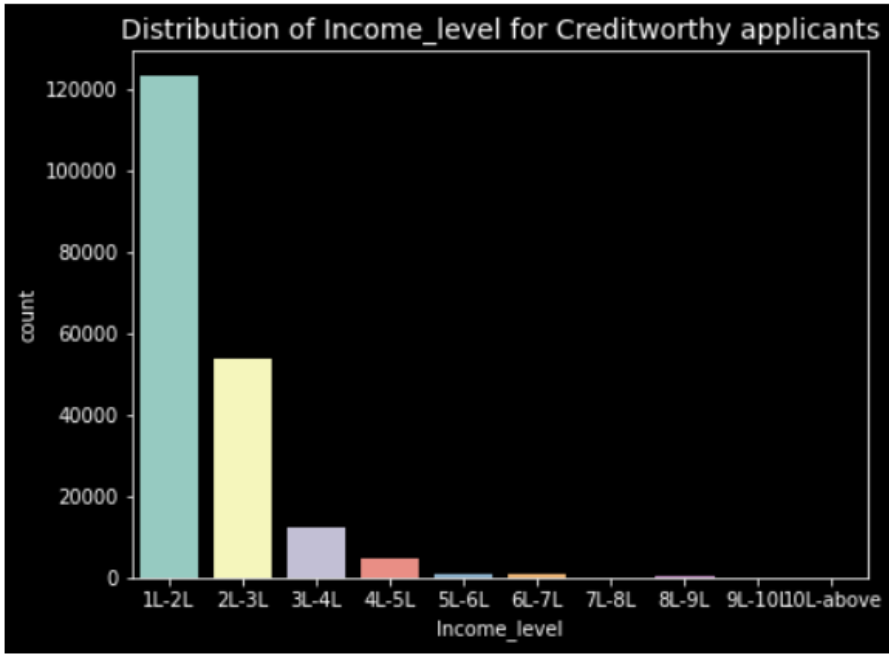
•People with low income applying for loans in high number
•applications for loan are decreasing with the increase in income level

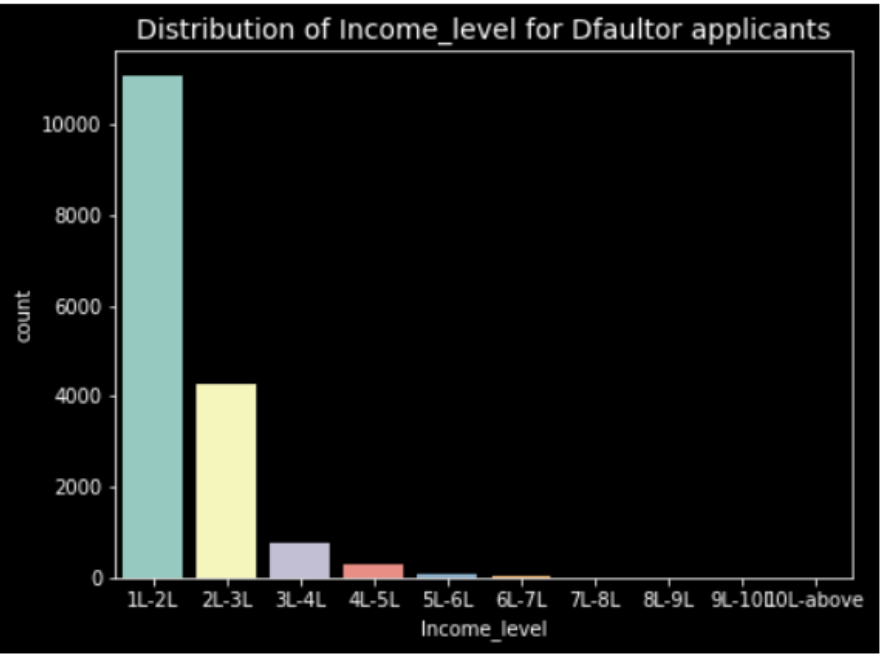


•People who own an asset are double than who don't own any asset
•Banks have higher number of applicants who can give security for loan in terms of their realty.

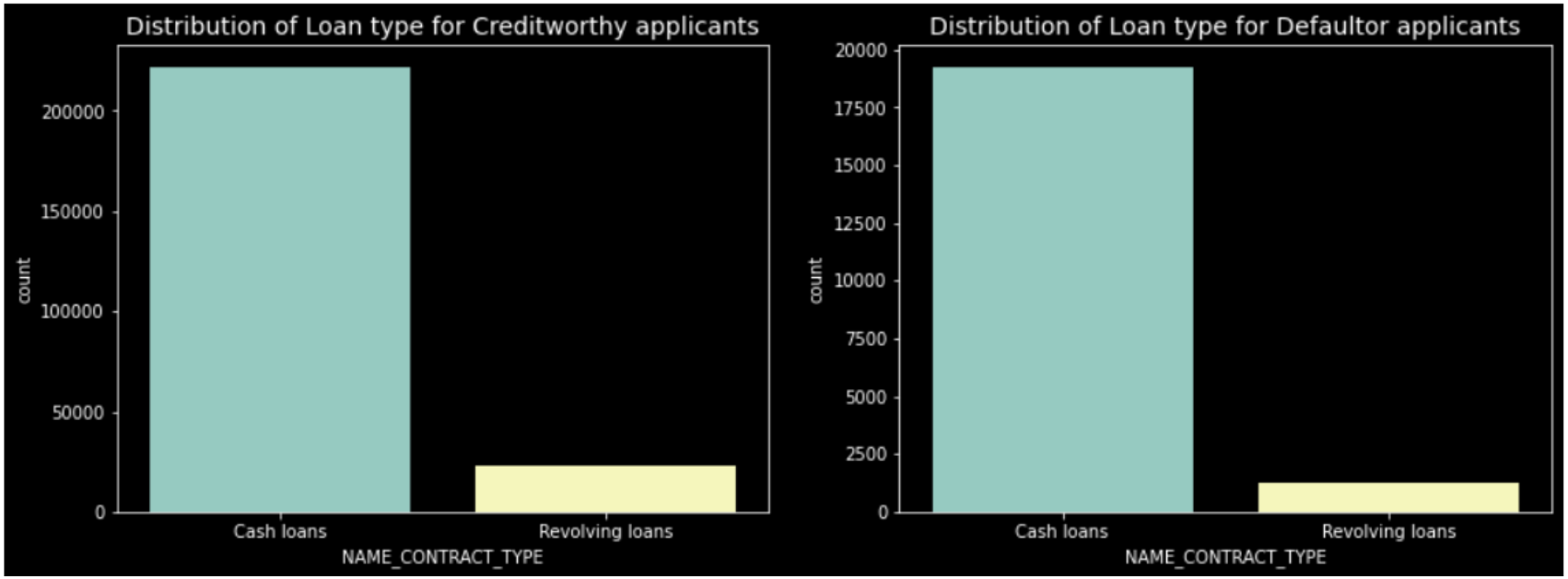
Segmented Analysis: This is an analysis which has been done after segmenting data into two halves with respect to target variable which is Target column. To perform segmented analysis the data set has been divided into two part one consist dataset where applicants did not face difficulty in repaying loan installments and the second dataset consist of information about applicants who faced difficulty in repaying loan installments.

- Univariate Analysis: There is a univariate analysis of income level in both dataset has been performed

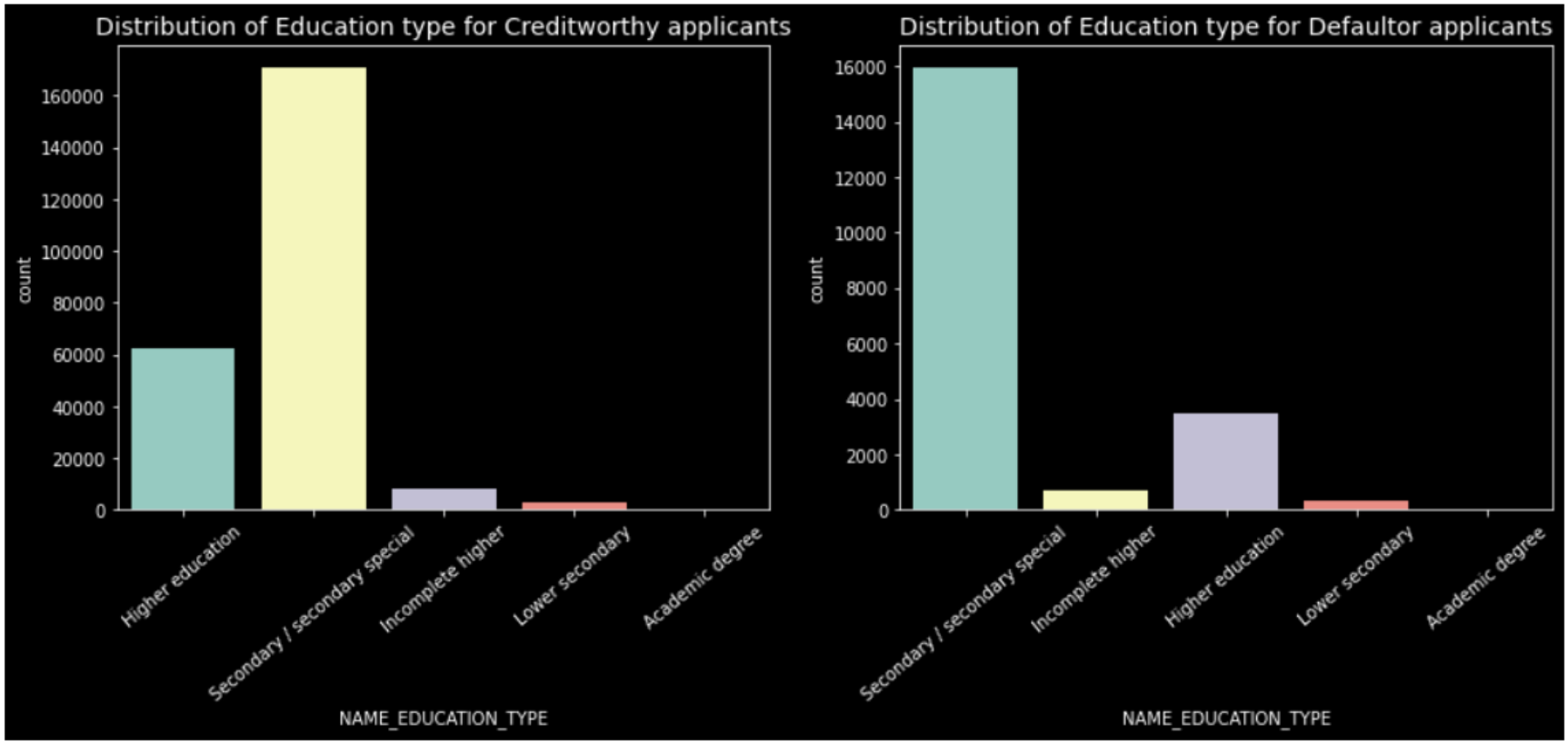




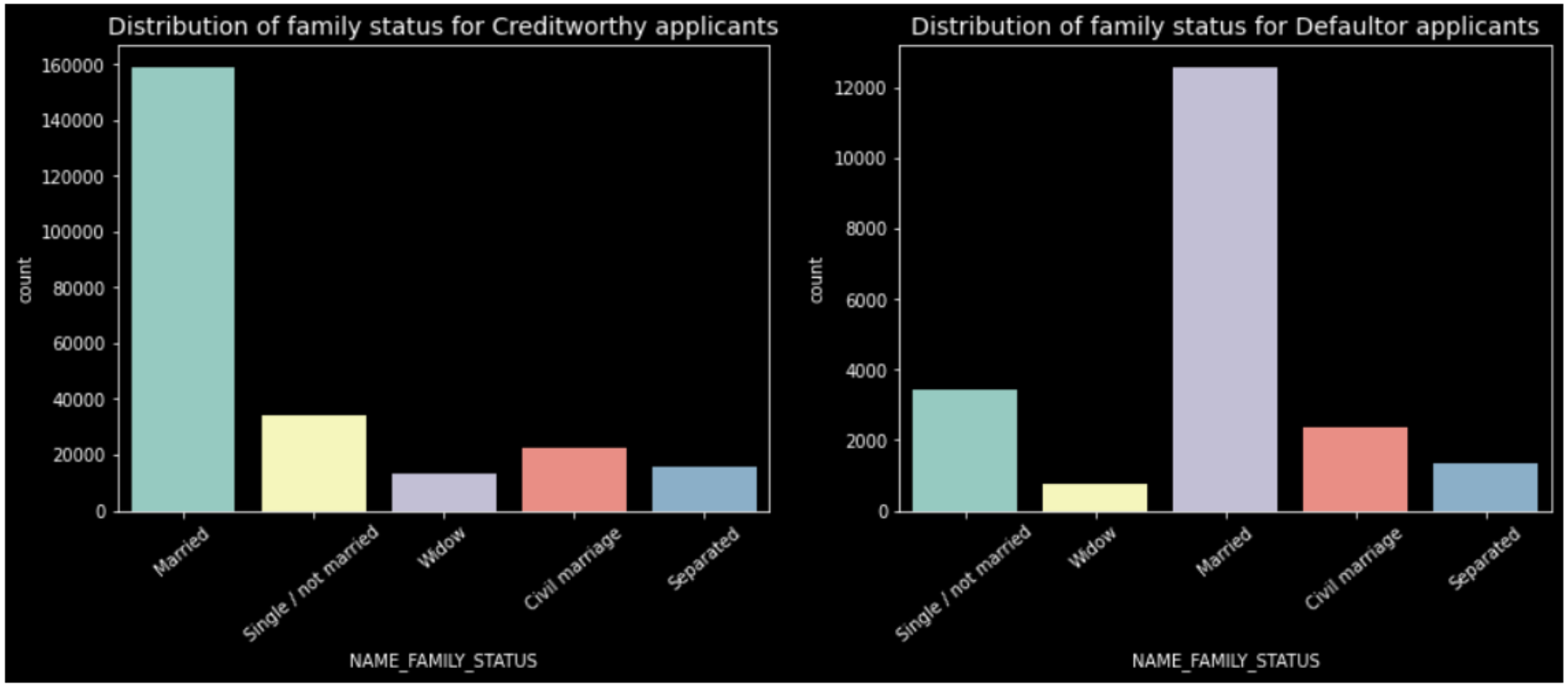
- proportion of low to medium income level applicants is low in defaulter applicants
- proportion of low-medium income level is high in non defaulter applicants



• Non defaulter applicants apply for revolving loans more than defaulter applicants

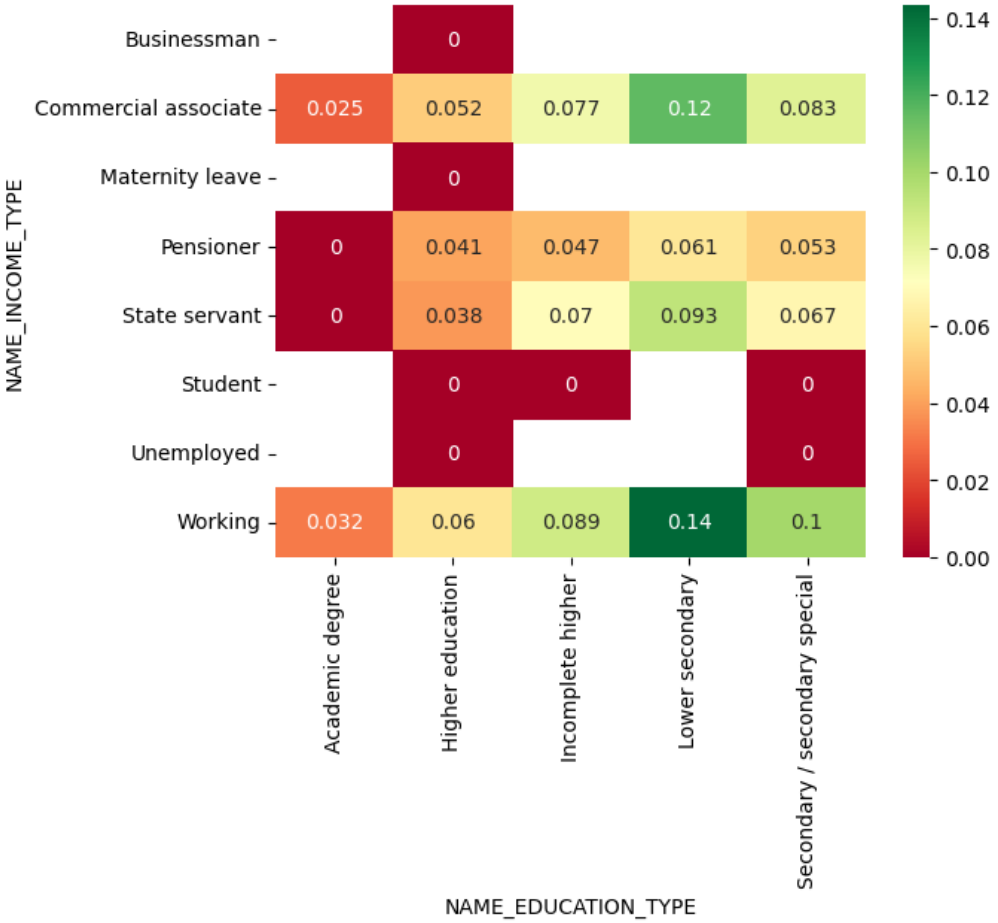


• The ratio of higher education in creditworthy applicants is quite high and very low in defaulters. While the ratio of secondary education is same in both applicants.



• The proportion of married applicants is high in both cases while single status is comparatively higher in defaulters.

Bivariate/Multivariate Analysis : This analysis involves analysis of more than more variable together, it helps identify relation between the variables



- Businessman, Maternity leave, student, unemployed with higher education are non defaulter applicants
- prisoners, state servants with academic degree are creditworthy applicants
- student and unemployed applicants with secondary education are also creditworthy
- students and unemployed with secondary education are non defaulters

• Identify top ten correlation with non defaulters: Following features are highly correlated in non defaulter applicants.

- AGE- DAYS_BIRTH
- FLAG_EMP_PHONE - DAYS_EMPLOYED
- EMPLOYED - FLAG_EMP_PHONE
- OBS_60_CNT_SOCIAL_CIRCLE - OBS_30_CNT_SOCIAL_CIRCLE
- AMT_GOODS_PRICE - AMT_CREDIT
- REGION_RATING_CLIENT_W_CITY - REGION_RATING_CLIENT
- CNT_FAM_MEMBERS - CNT_CHILDREN
- LIVE_REGION_NOT_WORK_REGION - REG_REGION_NOT_WORK_REGION
- DEF_60_CNT_SOCIAL_CIRCLE- DEF_30_CNT_SOCIAL_CIRCLE

The correlation does not indicate any decisive factors which help identify non defaulters

• Identify top ten correlation with defaulters: Following features are highly correlated in non defaulter applicants.

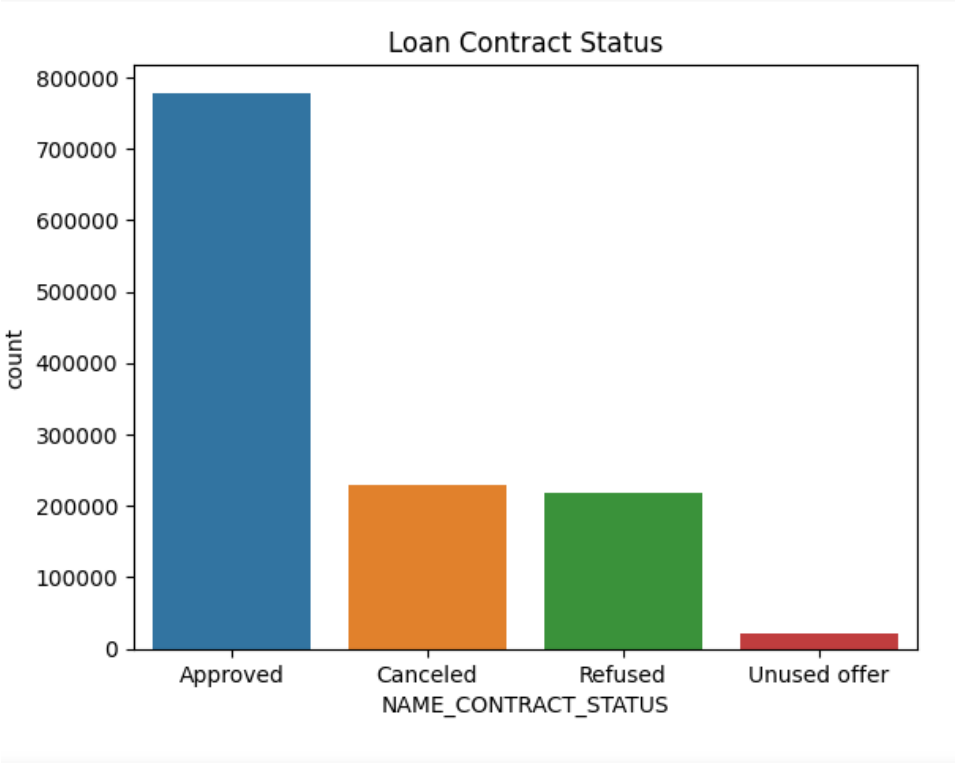
- AGE- DAYS_BIRTH
- FLAG_EMP_PHONE - DAYS_EMPLOYED
- EMPLOYED - FLAG_EMP_PHONE
- OBS_60_CNT_SOCIAL_CIRCLE - OBS_30_CNT_SOCIAL_CIRCLE
- AMT_GOODS_PRICE - AMT_CREDIT
- REGION_RATING_CLIENT_W_CITY - REGION_RATING_CLIENT
- CNT_FAM_MEMBERS - CNT_CHILDREN
- LIVE_REGION_NOT_WORK_REGION - REG_REGION_NOT_WORK_REGION
- DEF_60_CNT_SOCIAL_CIRCLE- DEF_30_CNT_SOCIAL_CIRCLE

The correlation does not indicate any decisive factors which help identify defaulters

• **Reading Previous application data:** Data related with previous application of loan applicants stored in another data frame. At this stage reading previous application data

• The data set has 1670214 rows and 37 columns
• missing values are less than 10 percent in few of the columns of data set

• **Merging both data set**: data with new application has been merged with previous data over the common column in both data which is 'SK_ID_CURR' on left



In merged data set the number of application approved are highest which shows that maximum number of applicants are non defaulters.