EDA case study

By
Pradeep Rajendran,
Santosh Kumar Sahoo

Objective:

- Conduct EDA analysis on the given bank datasets and get the insights.
- Identify the patterns and driving factors behind the loan defaults.
- Analysing the risks associated with the bank's decisions.
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Importing the libraries and datasets(Application_data):

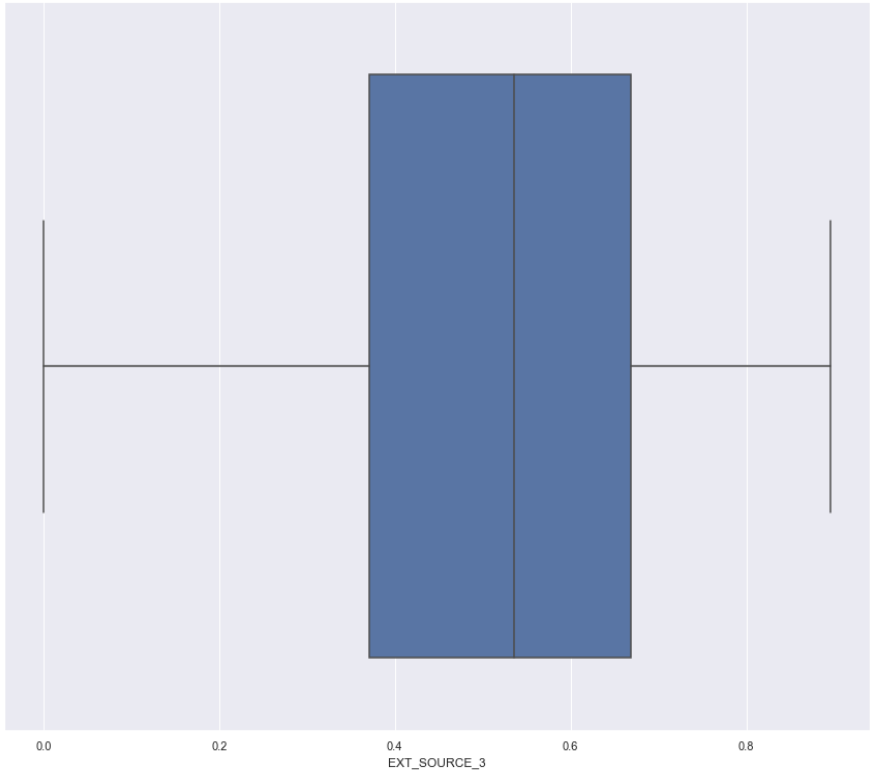
- Importing the needed libraries numpy, pandas, pyplot from matplotlib and seaborn.
- Reading the dataset ‘Application_data.csv file’ from the local excel file to python data frame.
- Checking the dataset’s shape and data imported are complete.
- Application_data.csv file is having 307511 rows and 122 columns.

Treating missing values(Application_data):

- Checking the null values percentage in each column.
- Removing the 49 columns from the data frame which are having the missing value percentage more than 40%.
- The column have ‘OCCUPATION_TYPE’ have missing value 31% and replacing this blank cells with the value ‘Unknown’.
- Deleting the complete rows of the missing values for columns which are having missing values less than 1% which will not affect the analysis.
- 7 columns have missing values and used box plot to impute the values.

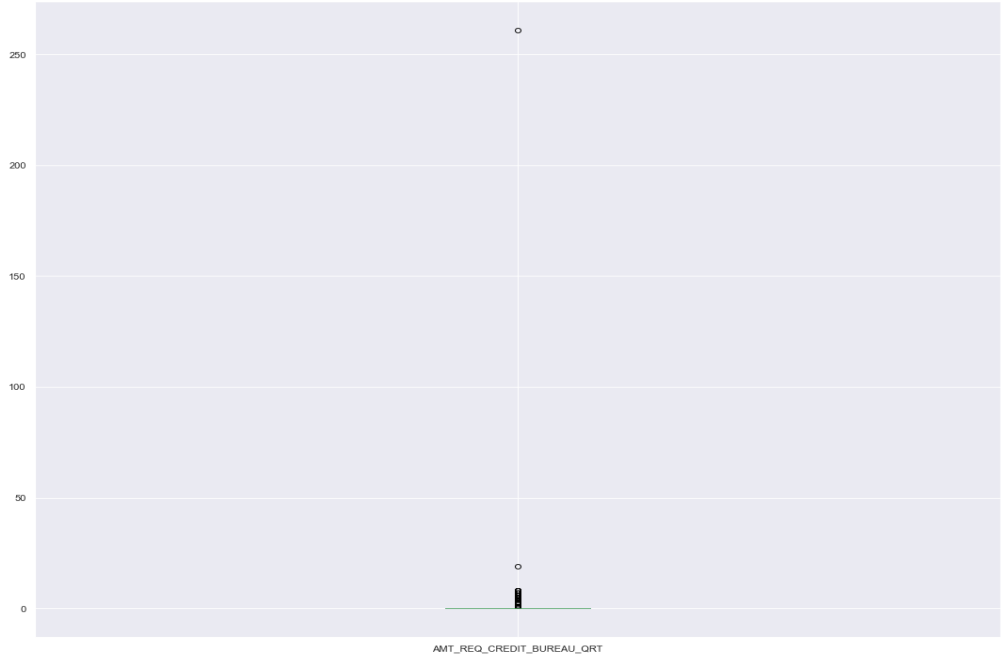
Treating missing values (cont):

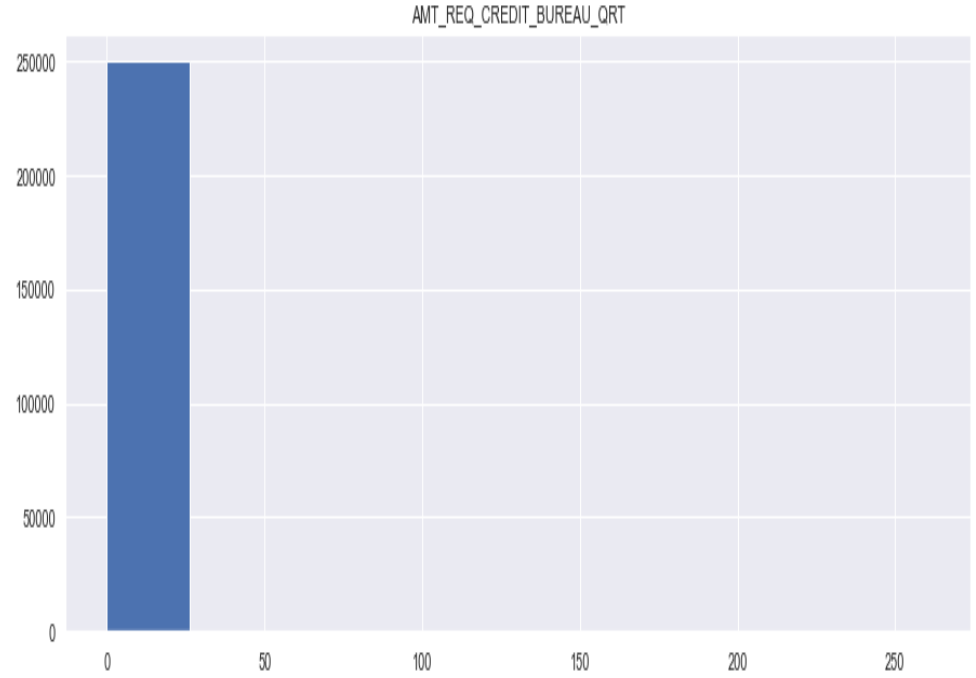
- The column ‘EXT_SOURCE_3’ have 19% missing value and replaced with the mean value. Because mean and media values are nearly same.



Treating missing values (cont):

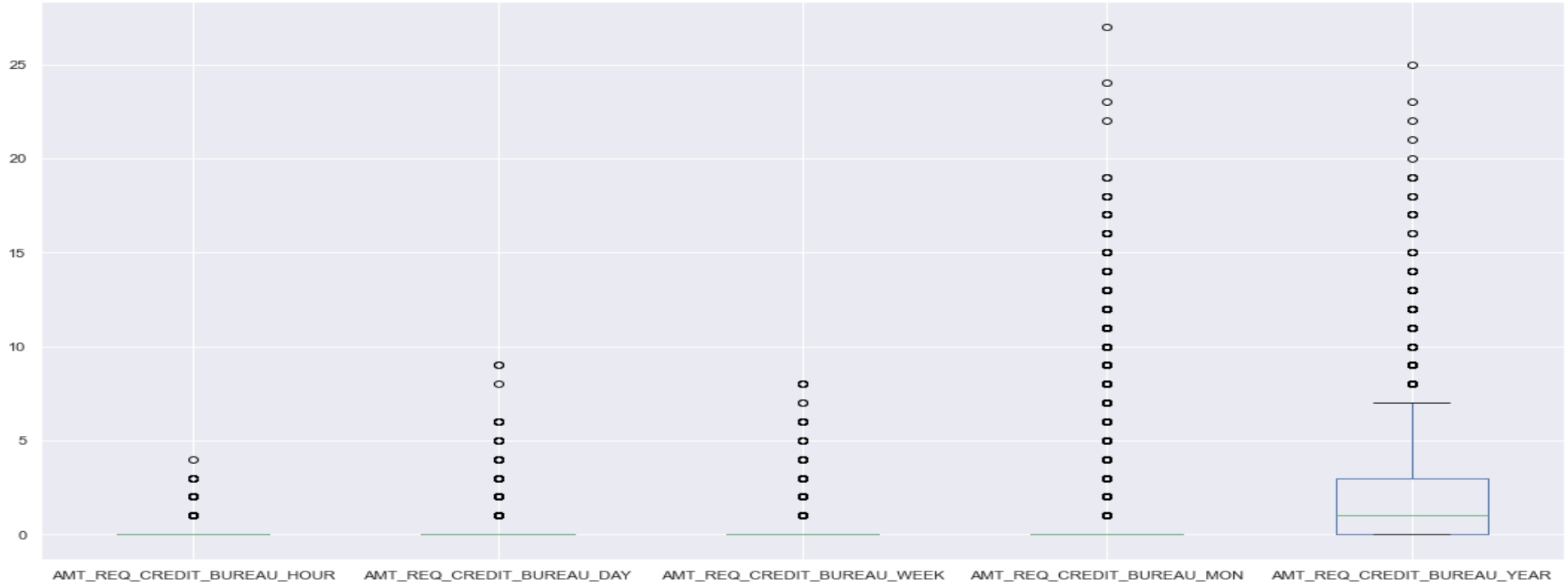
- The column ‘AMT_REQ_CREDIT_BUREAU_QRT’ have an outlier having the value more than 250 which is out of normal range.
- So, we can replace this missing values with median.





Treating missing values (cont):

- The columns 'AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON' and'AMT_REQ_CREDIT_BUREAU_YEAR' are having the missing value around 13%.
- These are some outliers are present in each column. So, median values of each column replaced with the missing values of the respective columns.



Checking non-numeric columns:

- There are 15 non-numeric columns.
- Some rows in the columns 'CODE_GENDER' and 'ORGANIZATION_TYPE’ have the value 'XNA’.
- We have 'XNA’ as value in 4 rows (less than 1%) in 'CODE_GENDER' and 54852 rows (18%) in 'ORGANIZATION_TYPE’. So, removing these rows.

Treating negative values:

- Column names 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH' and 'DAYS_LAST_PHONE_CHANGE' are having the values in negative.
- Changing the negative values to positive values using the abs() function.

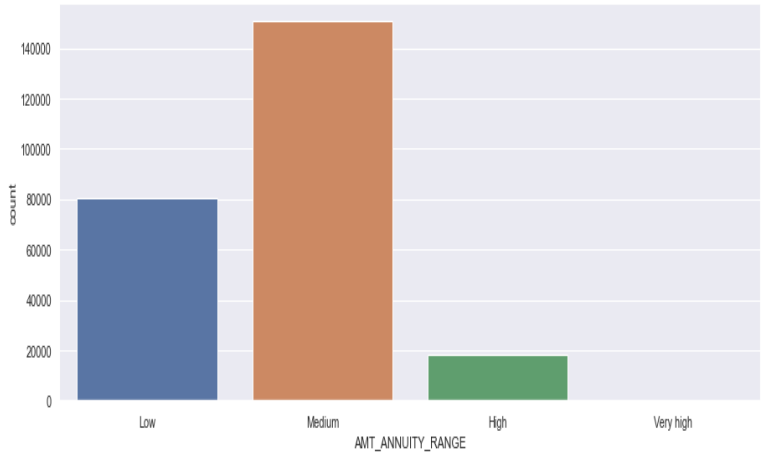
Removing unnecessary columns:

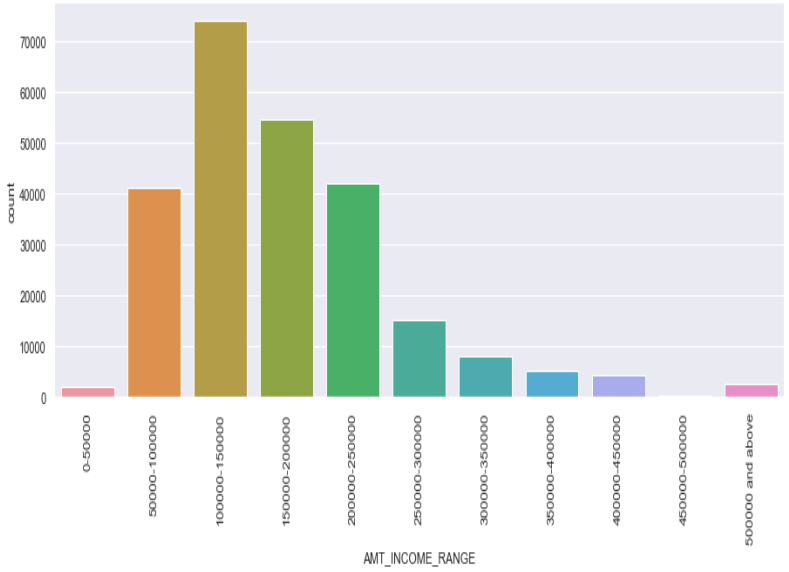
- We have 73 columns in the data frame.
- Removing the unnecessary 35 columns which are not having important information for the analysis.

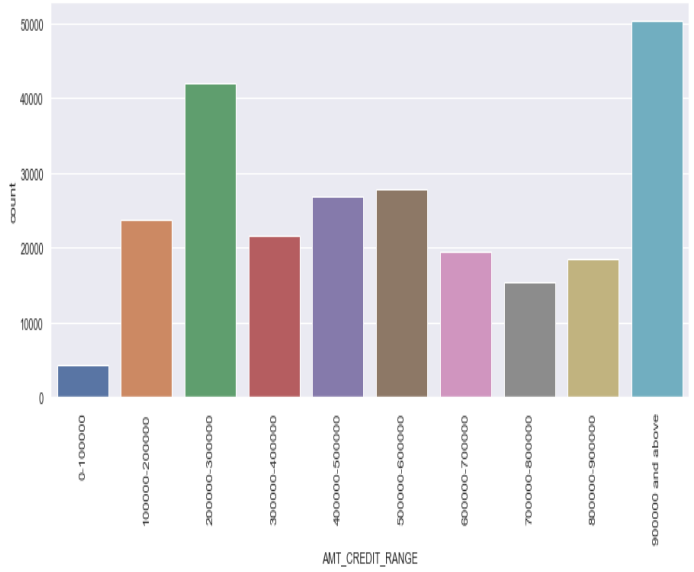
- Now we don’t have the missing values and incorrect data in the data set.
- We have 249675 rows and 38 columns in the data set.

Converting amount columns to categorical:

- Creating new columns for the existing columns ‘AMT_ANNUITY’, ‘AMT_INCOME_TOTAL’ and ‘AMT_CREDIT’ with bins to covert the numerical to categorical values (‘AMT_ANNUITY_RANGE’, ‘AMT_INCOME_RANGE’ and ‘AMT_CREDIT_RANGE’).
- Converting the numerical data into categorical values will help us to perform the analysis in better way.







Finding imbalance in target column:

- Target column having the huge imbalance with the ratio of around 10:90.
- For the target value 0, we have 227975 rows and 41 columns.
- For the target value 1, we have 21700 rows and 41 columns.
- We are having huge imbalance in the target column. So, separating the current data frame (app_df) into two data frames ‘target0_df’ and ‘target1_df’.

Univariate analysis:

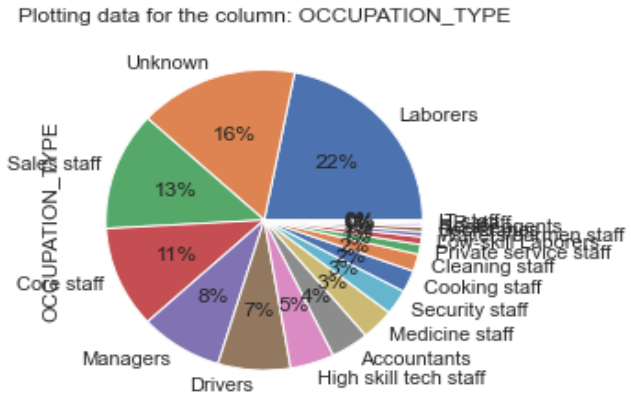
- **Categorical columns:**
  - To avoid recreating the same block of code, creating a function called ‘plotting’ to generate a pie plot, a count plot and a bar plot for the given input column.
  - Using for loop, created plot for all categorical columns to get the insights.

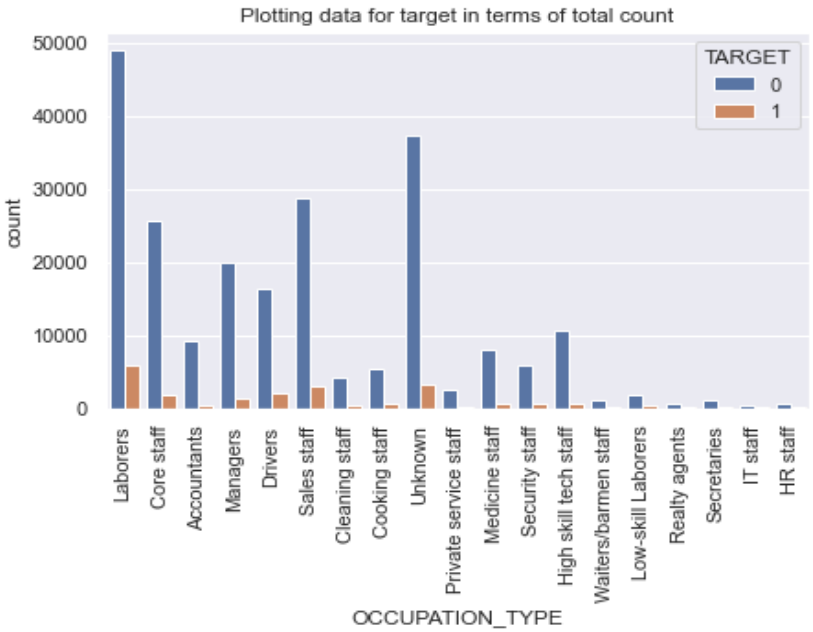
  - Observations as per the plots:
    - People who are working in organization type Business Entity Type 3 and Self-employed are highly applying for the loans.
    - Among these two categories, ‘Business Entity Type 3’ are having high numbers in payment difficulties.
    - Cash loans are highly preferred by the clients than revolving loans and we have payment difficulty numbers high in this category, which is obvious.
    - Females are applying high number of loans but, the default percentage is higher for males.
    - People who are having Secondary/Secondary special education status are applying a higher number of loans and have higher default counts.
    - People who are having an income range of 50,000 to 2,25,000 are applying for high number of loans and have payment difficulties too.
    - Married people applying high number of loans.
    - People who own house/apartment are the most applying for loans.
  - Plots are added in the next slides for each columns.

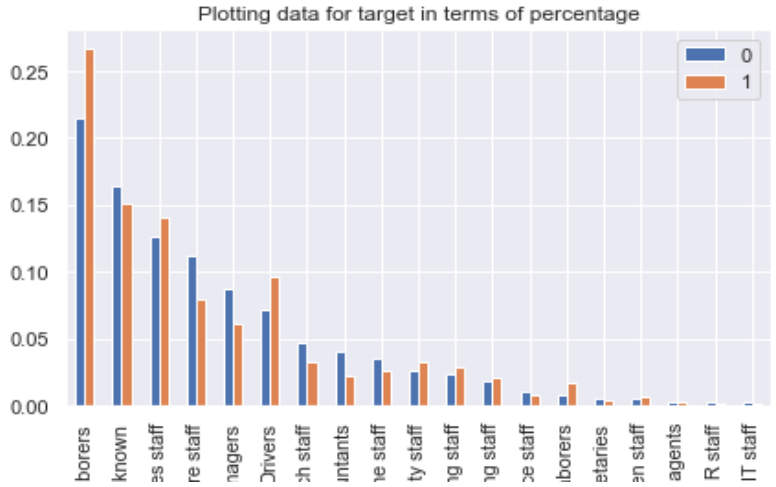
Univariate analysis (Categorical columns):

Column name:

OCCUPATION_TYPE



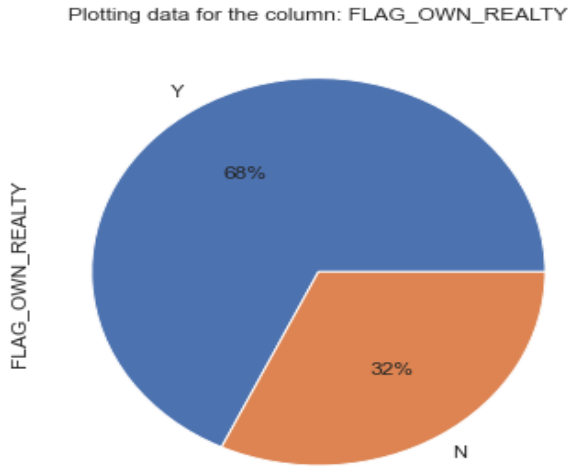


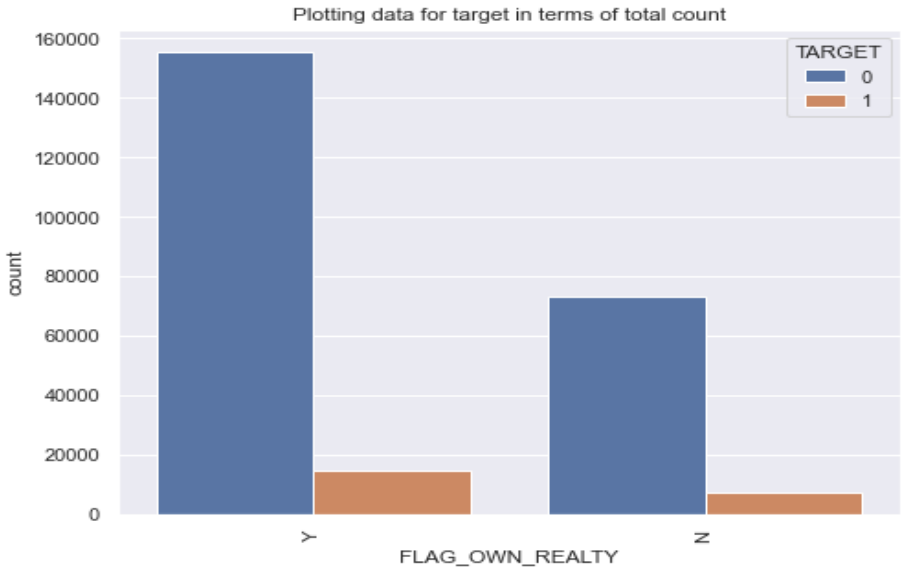


Univariate analysis (Categorical columns):

Column name:

FLAG_OWN_REALTY

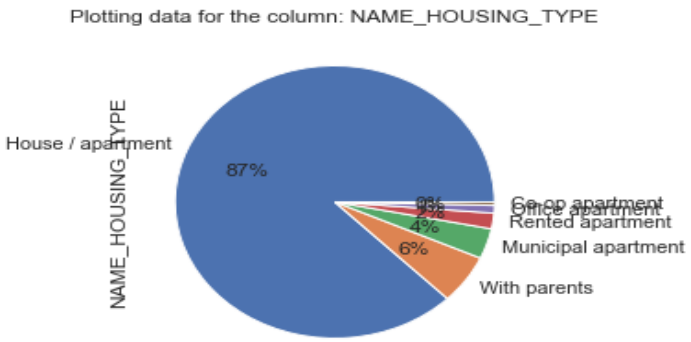


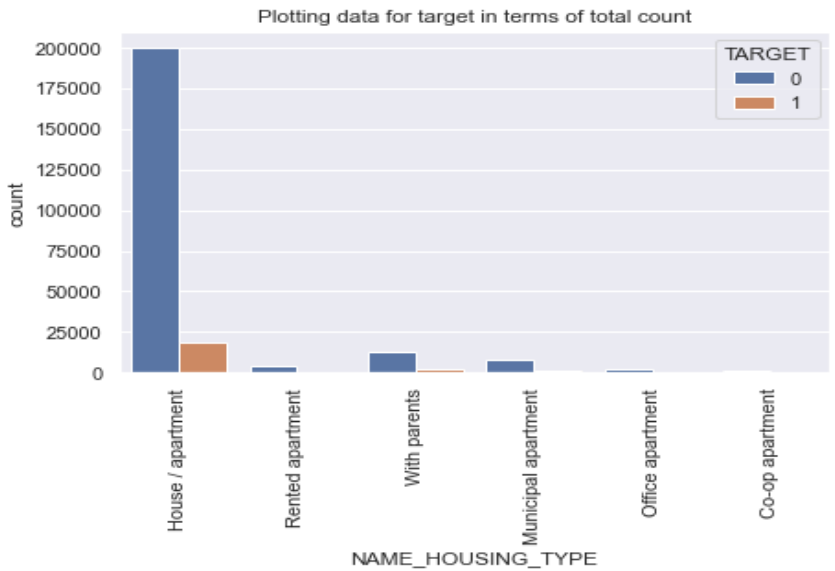


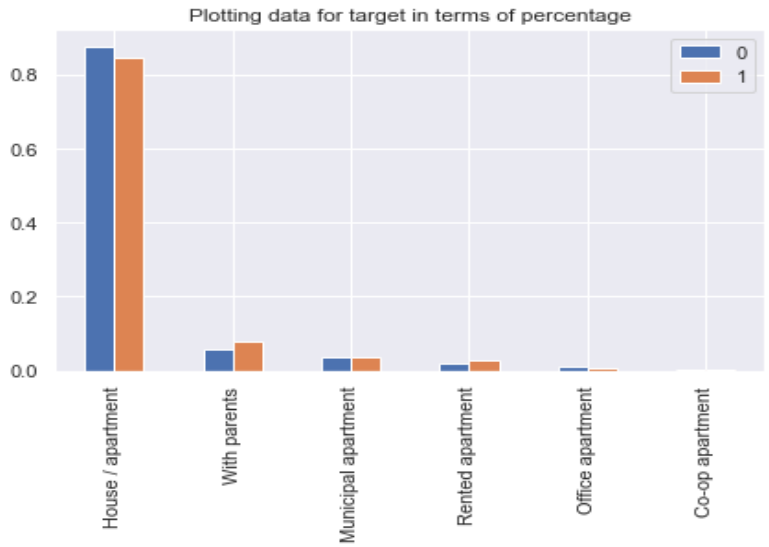
Univariate analysis (Categorical columns):

Column name:

NAME_HOUSING_TYPE



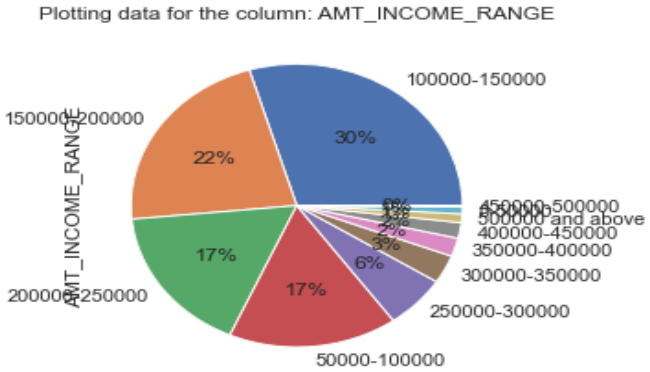


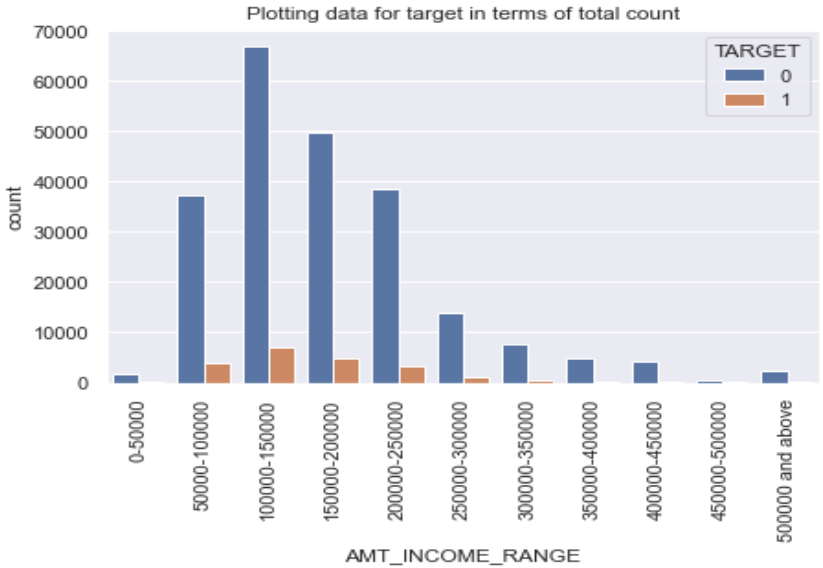


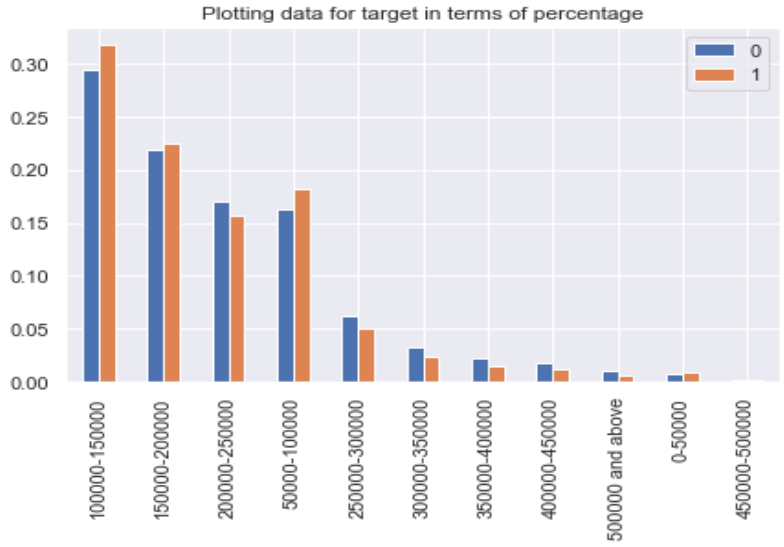
Univariate analysis (Categorical columns):

Column name:

AMT_INCOME_RANGE



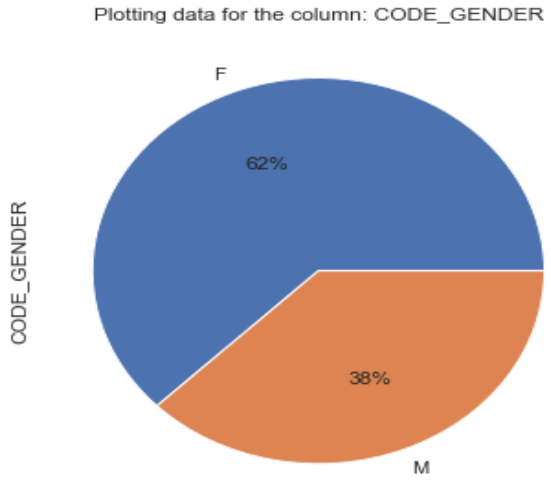


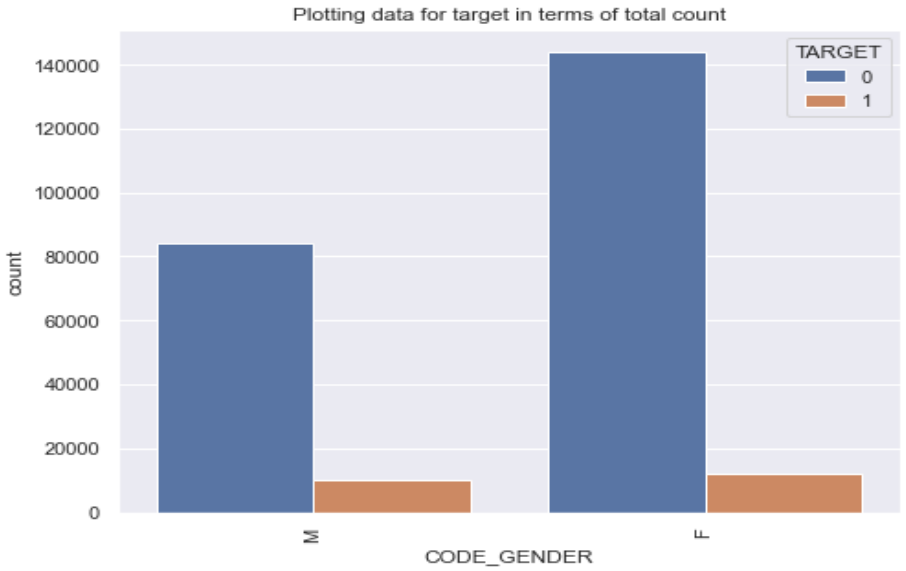


Univariate analysis (Categorical columns):

Column name:

CODE_GENDER

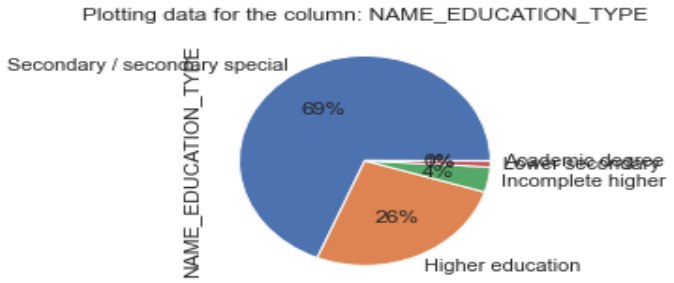


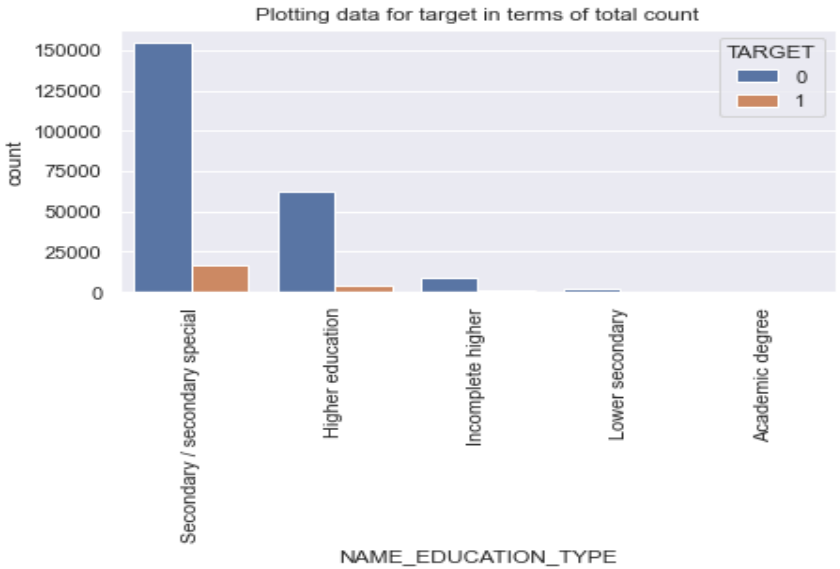


Univariate analysis (Categorical columns):

Column name:

NAME_EDUCATION_TYPE

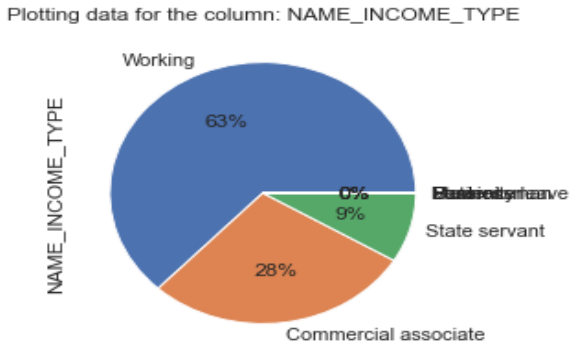


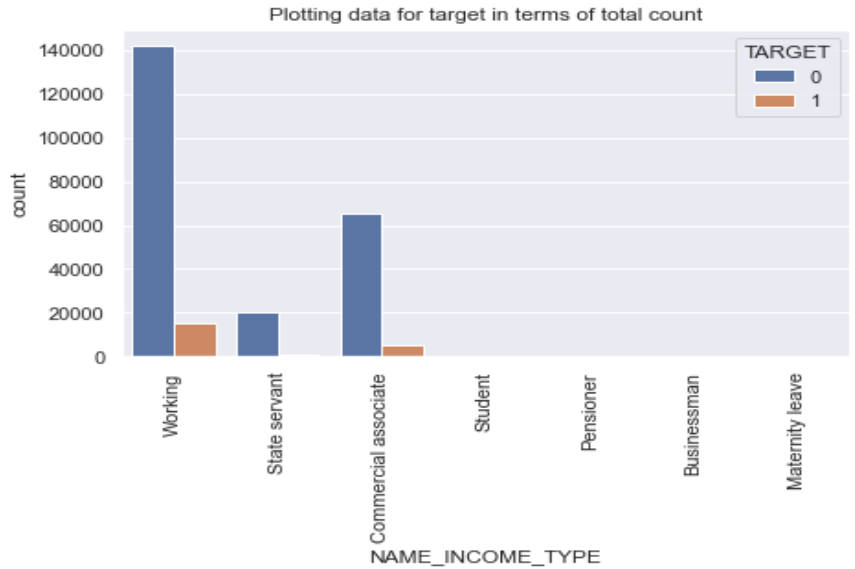


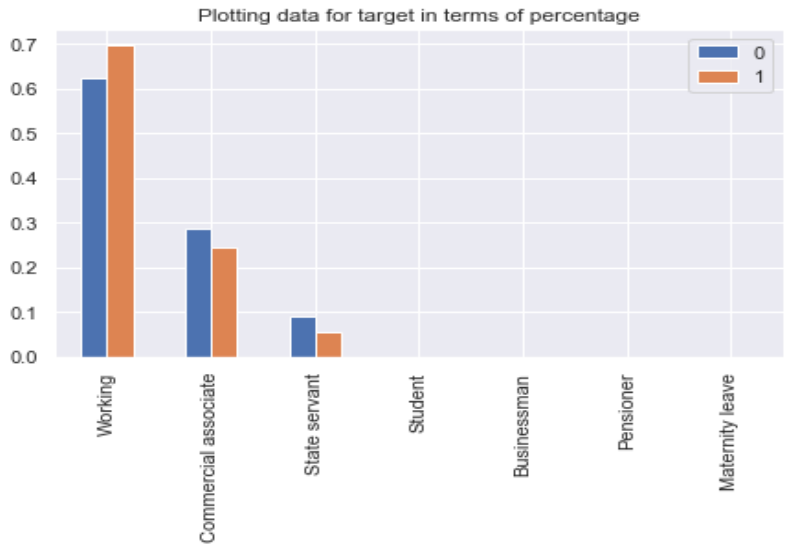
Univariate analysis (Categorical columns):

Column name:

NAME_INCOME_TYPE



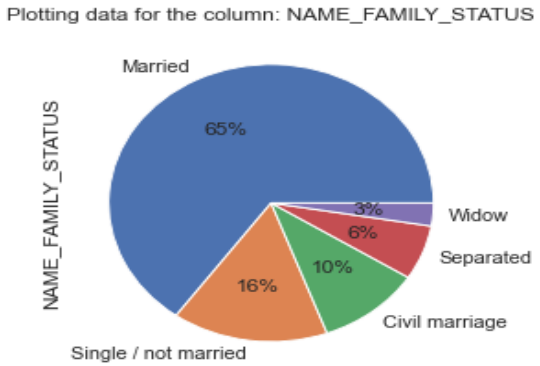


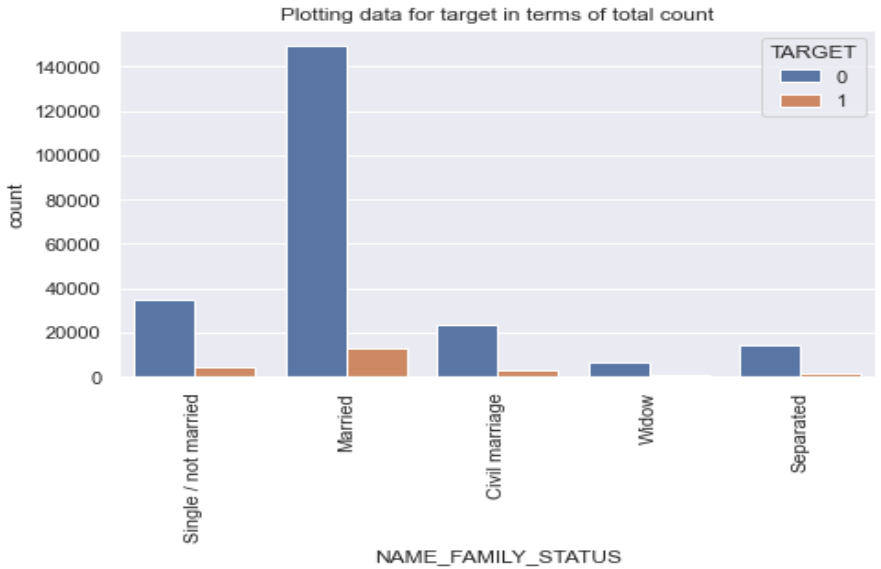


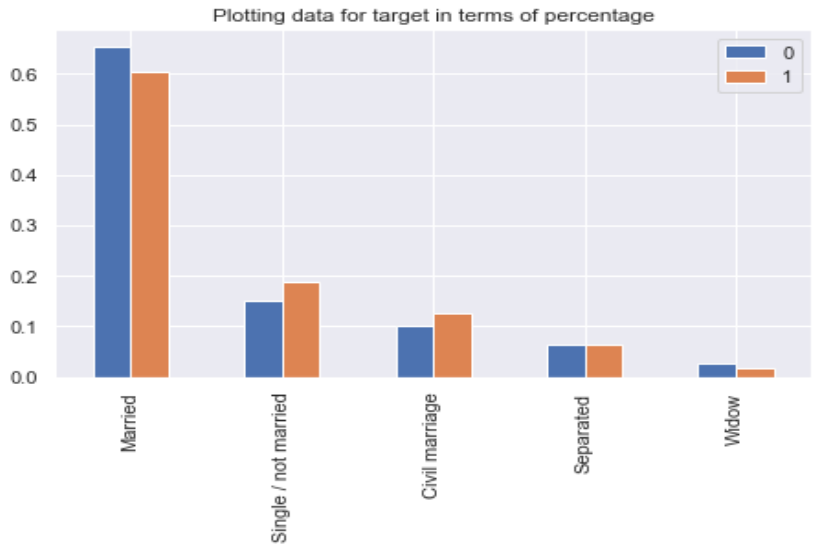
Univariate analysis (Categorical columns):

Column name:

NAME_FAMILY_STATUS



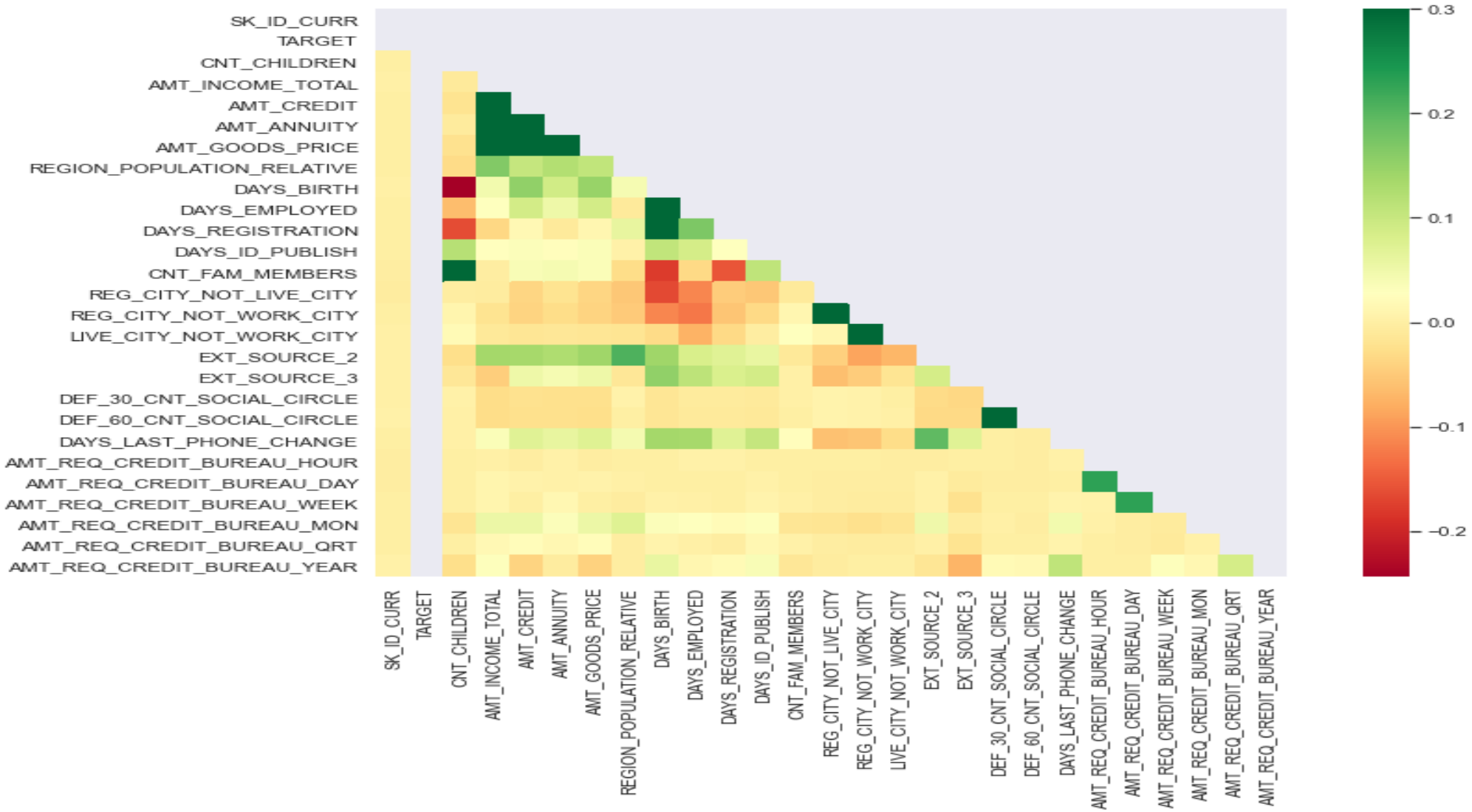




Univariate analysis:

- **Numerical columns:**
  - There are 27 numerical columns.
  - Find the correlation between each columns using the two data frames which are created as per the target value 0 and 1.
  - Observations for the target value 0:
    - Income of the client is directly proportional to the region's population.
    - Client’s Income amount is directly proportional to the goods price and amount annuity.
    - Client’s income is directly proportional to the score of external source 2.
    - Credit amount, Annuity amount and goods price are highly proportional.
    - Client changing his registration is inversely proportional to the count of the family members.
    - Client’s permanent address does not match with the contact address is directly proportional with the client's permanent address does not match with the work address.
    - Client's social surroundings defaulted on 30 days past due directly proportional to 60 days past due.
    - Credit amount is directly proportional to the region's population.

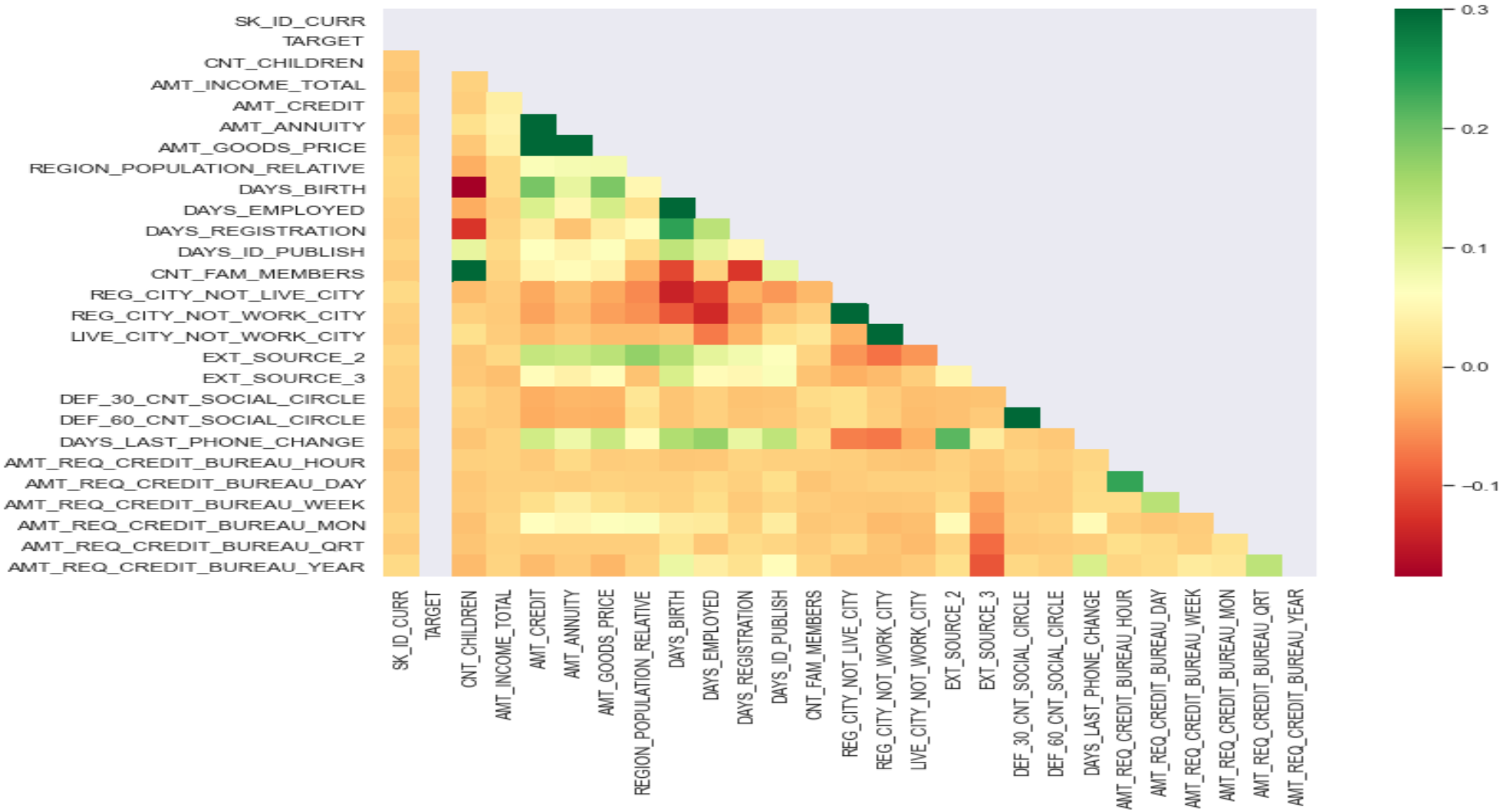
Correlation heatmap (target = 0):



Univariate analysis:

- **Numerical columns:**
  - Observations for the target value 1:
    - The amount credited is directly proportional to the good price and amount annuity.
    - The client’s permanent address does not match the contact address is directly proportional to the client's permanent address does not match the work address.
    - The client's social surroundings defaulted on 30 days past due directly proportional to 60 days past due.
    - Count of children inversely proportional to days birth and days registration.

Correlation heatmap (target = 1):



Top 10 correlation variables:

- Target value = 0:

VARIABLE_1	VARIABLE_2
AMT_CREDIT	AMT_GOODS_PRICE
CNT_FAM_MEMBERS	CNT_CHILDREN
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY
AMT_GOODS_PRICE	AMT_ANNUITY
AMT_CREDIT	AMT_ANNUITY
REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY
AMT_ANNUITY	AMT_INCOME_TOTAL
DAYS_EMPLOYED	DAYS_BIRTH
AMT_GOODS_PRICE	AMT_INCOME_TOTAL

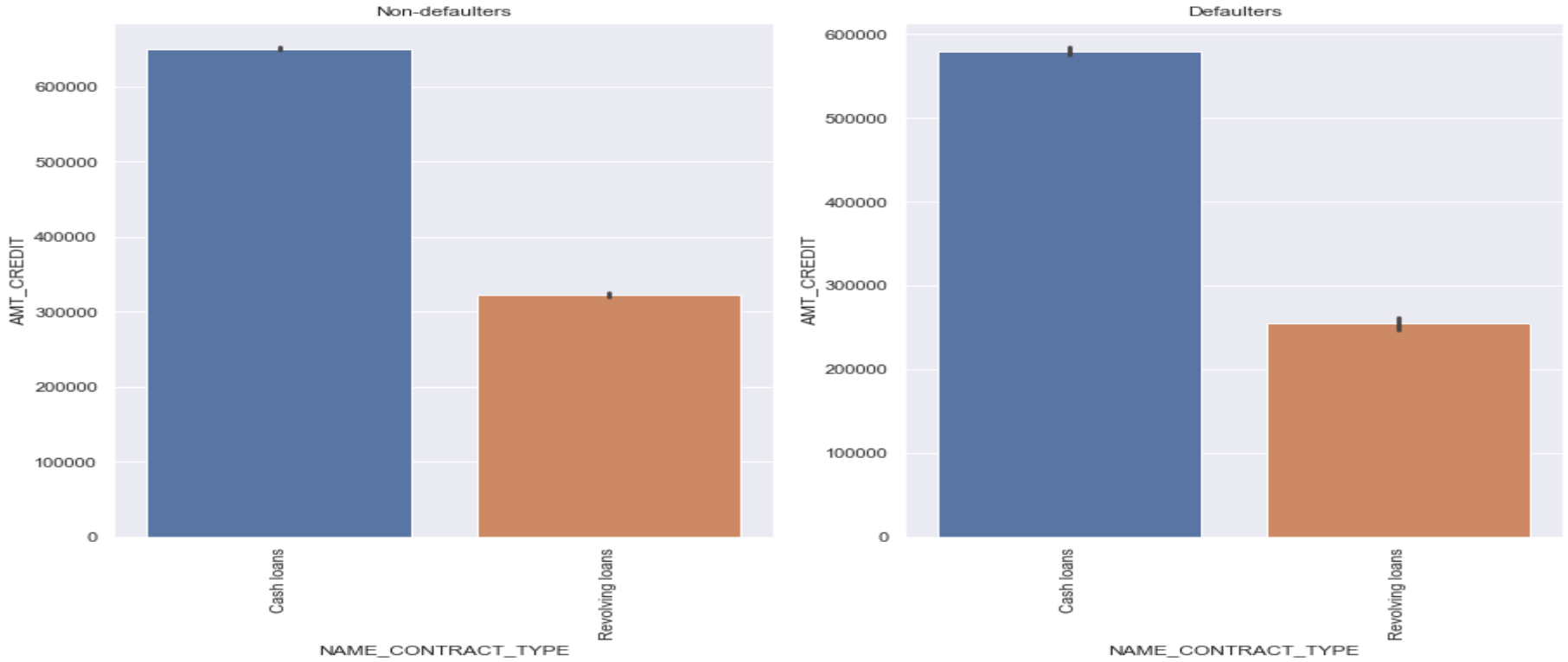
- Target value = 1:

VARIABLE_1	VARIABLE_2
AMT_CREDIT	AMT_GOODS_PRICE
CNT_FAM_MEMBERS	CNT_CHILDREN
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY
AMT_GOODS_PRICE	AMT_ANNUITY
AMT_CREDIT	AMT_ANNUITY
REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY
DAYS_EMPLOYED	DAYS_BIRTH
DAYS_REGISTRATION	DAYS_BIRTH
AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY

Segmented univariate analysis:

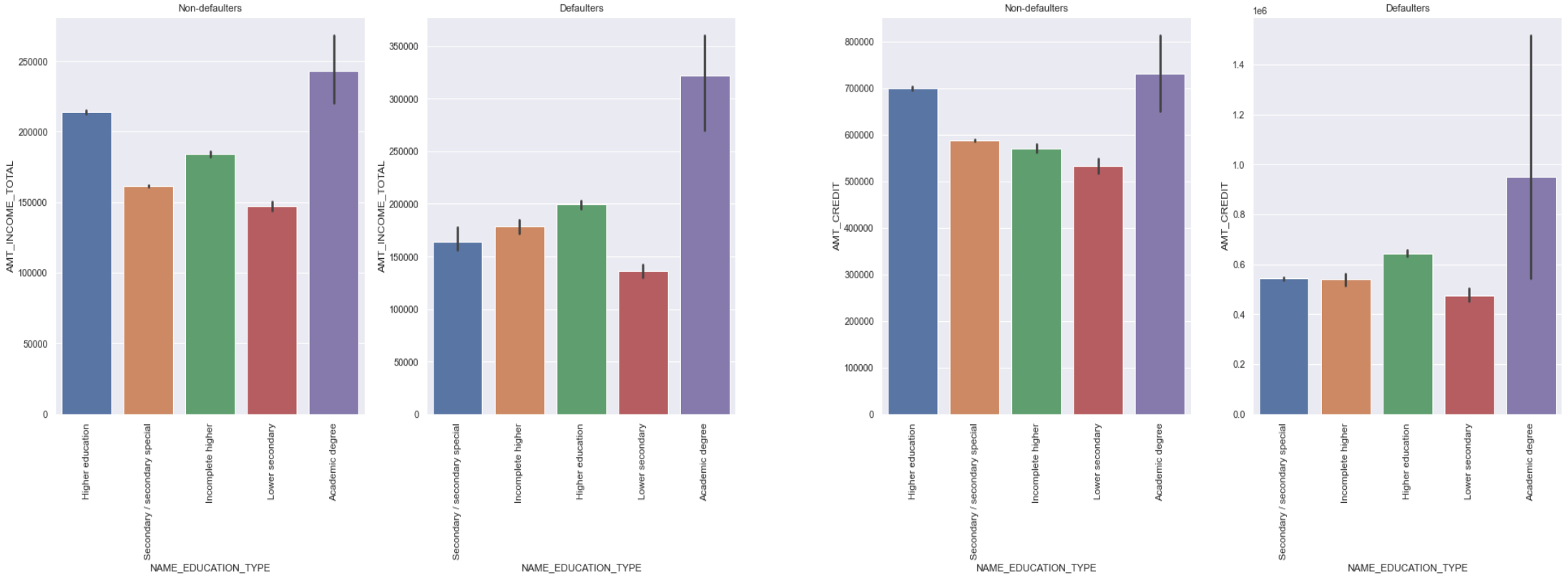
- To avoid recreating the same block of code, creating a function called ‘plotting_seg’ to generate two bar plots for target values 0 and 1 as per the given columns.

  - **'NAME_CONTRACT_TYPE’ vs 'AMT_CREDIT’:**
    - The clients in both categories prefer cash loans to revolving loans.



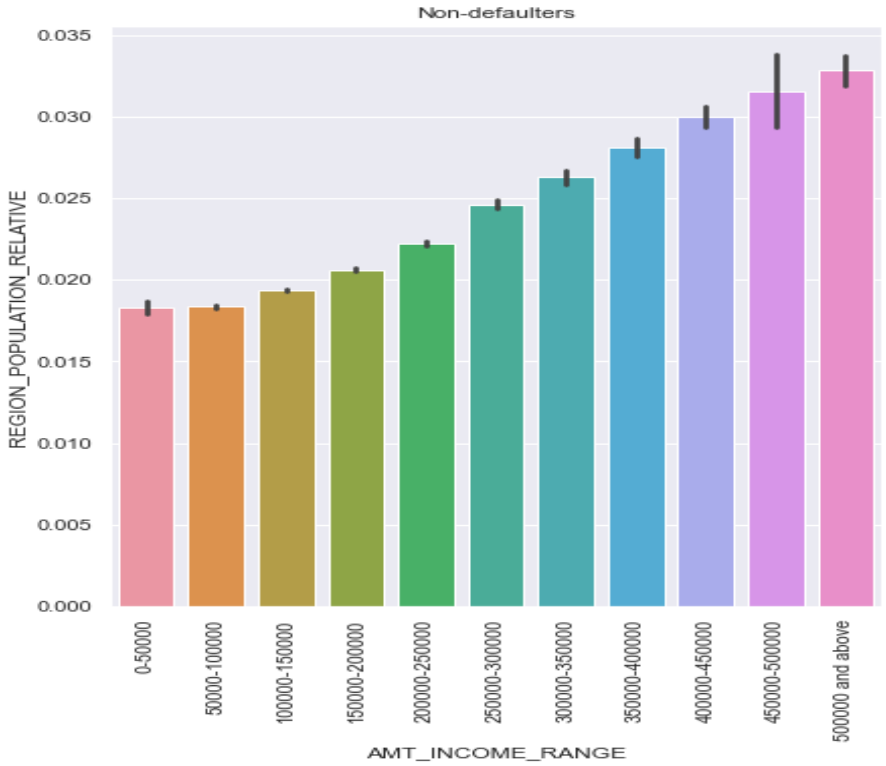
Segmented univariate analysis:

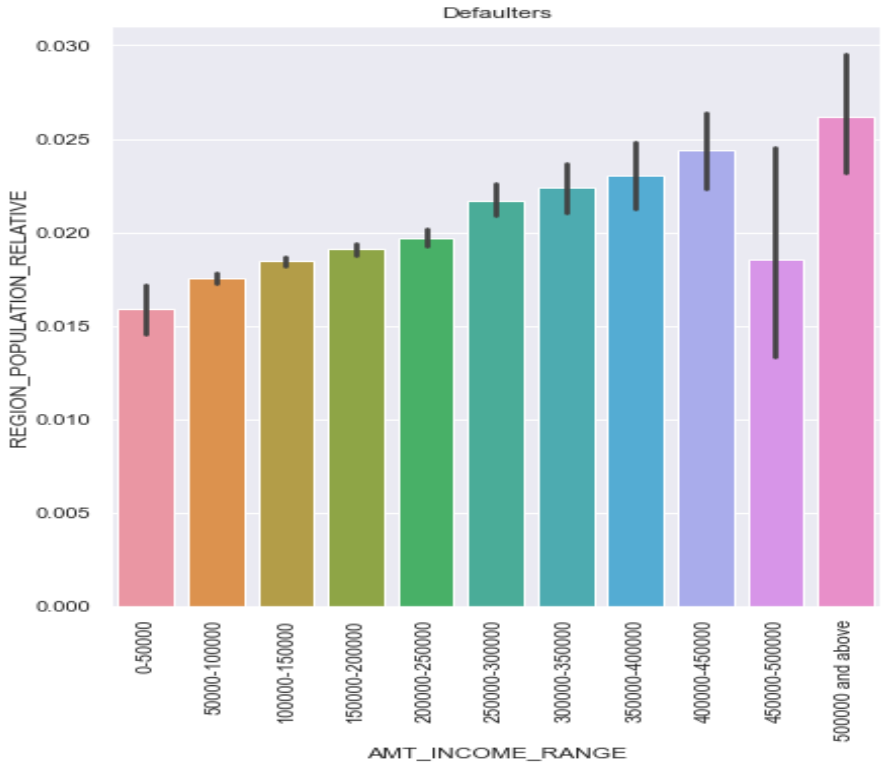
- **'NAME_EDUCATION_TYPE' vs 'AMT_INCOME_TOTAL' and 'AMT_CREDIT':**
  - Clients who are having an academic degree or higher education get higher income and the loan value is high. Also, clients who have an academic degree are having high defaulted value.



Segmented univariate analysis:

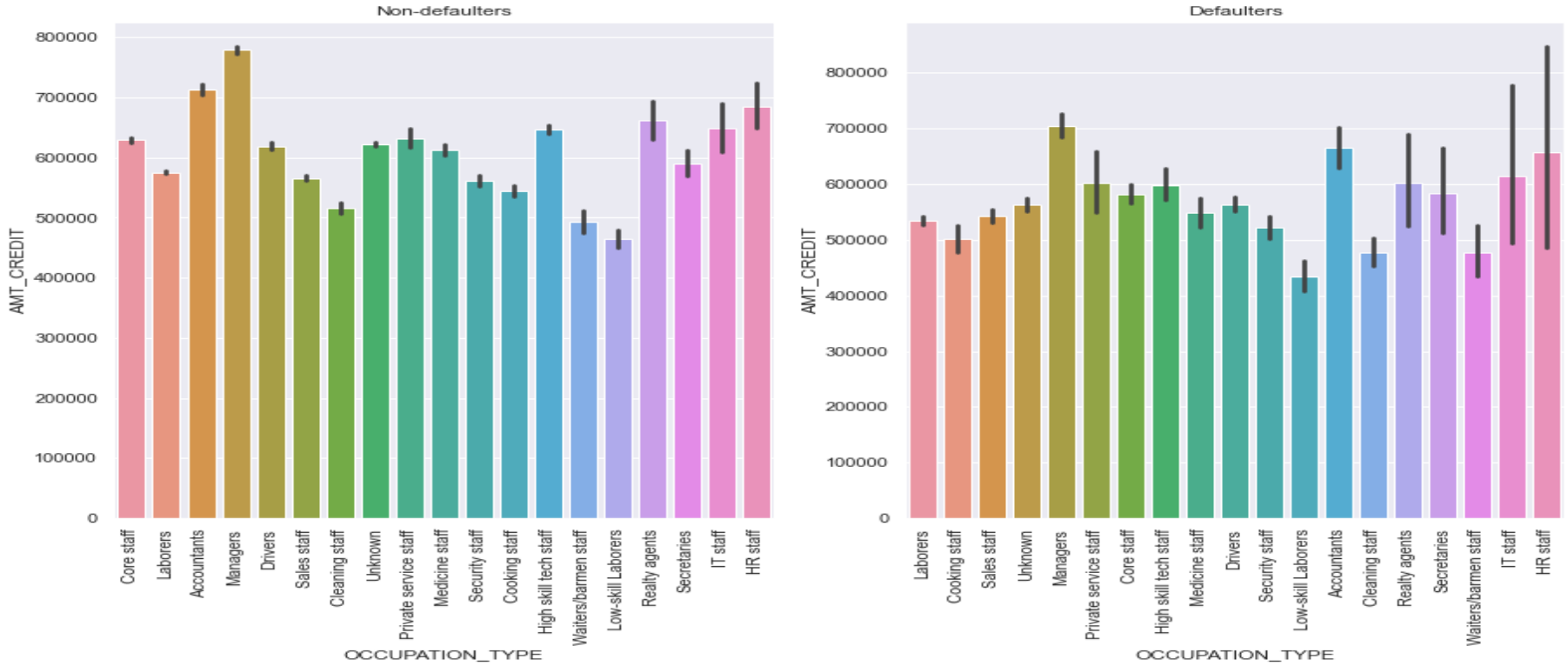
- **'AMT_INCOME_RANGE','REGION_POPULATION_RELATIVE':**
  - The clients who have high incomes are living in highly populated regions.





Segmented univariate analysis:

- **'OCCUPATION_TYPE','AMT_CREDIT':**
  - Managers, Accountants and HR staff are having high credit amounts.

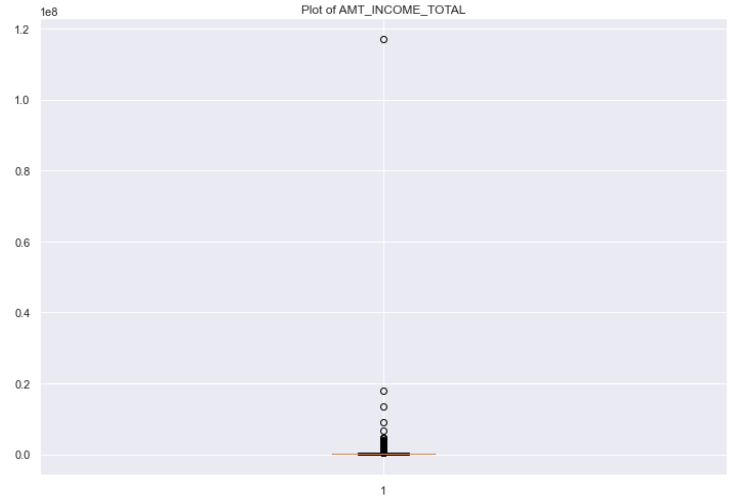


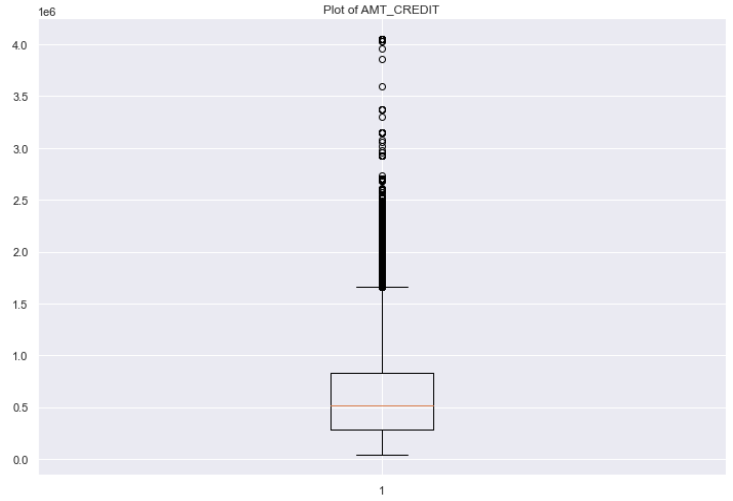
Outlier analysis:

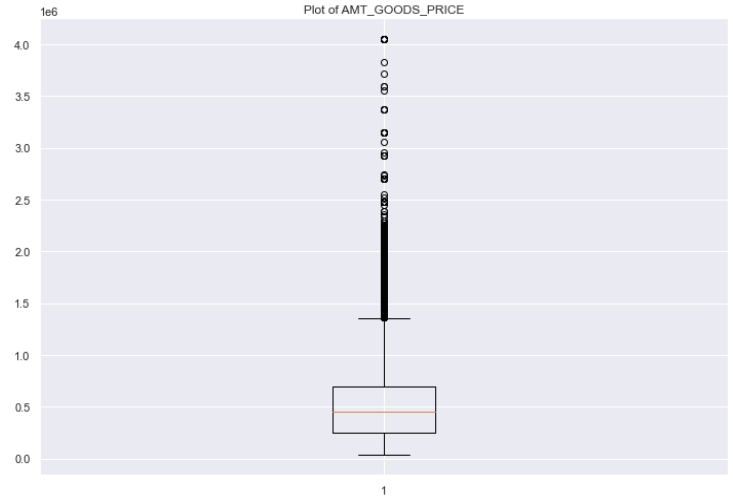
- Plotting the numerical data columns against the index and analysing for the outliers using boxplots.

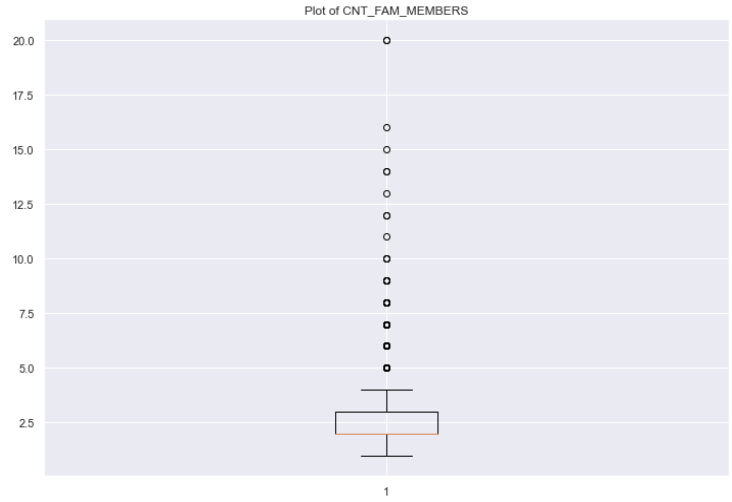
- **Observations from the boxplots:**
  - AMT_INCOME_TOTAL: Total income has an outlier which is having a value of nearly 12,00,00,000. Some outliers are having values in the range between 1 Crore to 2 Crore.
  - AMT_CREDIT: There are many outliers present in this column in the range between 20,00,000 to 40,00,000. First quartile is smaller than the fourth quartile.
  - AMT_ANNUITY: This column also have some outliers. First quartile is smaller than the fourth quartile.
  - AMT_GOOD_PRICE: This column has outliers between the range between 15,00,000 to 40,00,000. First quartile is smaller than the fourth quartile.
  - REGION_POPULATION_RELATIVE: This columns is only having one outlier with a value of around 0.07. Chart showing that the data value is almost equally spread.
  - CNT_FAM_MEMBERS: Most of the family is having the total count below 4. Highest family count in the data is 20 people. Which is rare. Very few families have a total count more than 5 people.

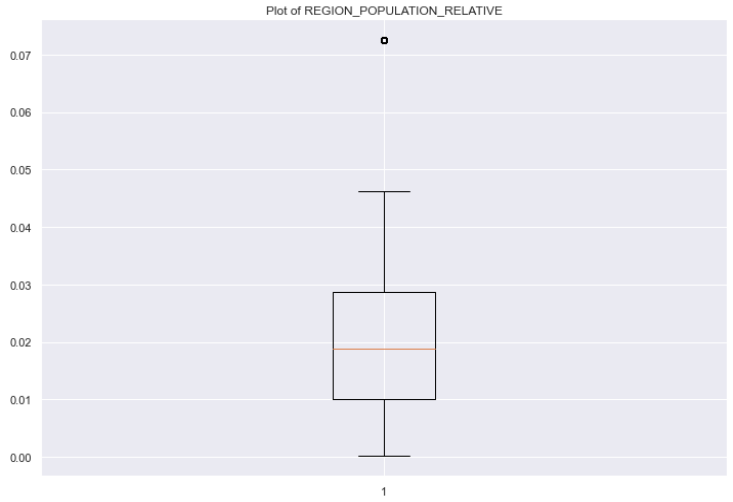
Outlier analysis - plots:

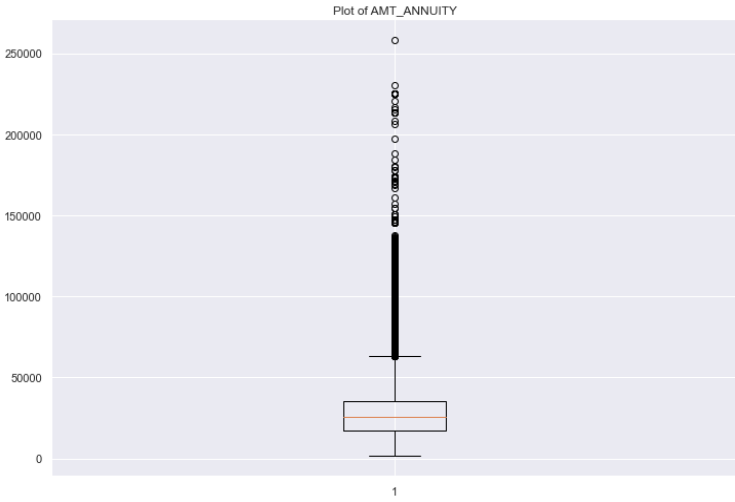












Importing the libraries and datasets (previous_application):

- Reading the dataset ‘previous_application.csv’ from the local excel file to python data frame.
- Checking the dataset’s shape and data imported are complete.
- previous_application.csv file is having 1670214 rows and 37 columns.

Treating missing values(previous_application):

- Checking the null values percentage in each column.
- Removing the 11 columns from the data frame which are having the missing value percentage more than 40%.
- Deleting the complete rows of the missing values for columns 'AMT_CREDIT' and 'PRODUCT_COMBINATION' which are having missing values less than 1% which will not affect the analysis.
- The balance columns are having missing values of around 22% and they are denoting amount and term details. So, we can remove the missing columns to perform the analysis without any biasing.

Checking non-numeric columns:

- There are 15 non-numeric columns.
- Some rows in the columns 'NAME_YIELD_GROUP', 'NAME_PAYMENT_TYPE', 'NAME_PRODUCT_TYPE', 'NAME_CASH_LOAN_PURPOSE', 'NAME_CLIENT_TYPE', 'NAME_GOODS_CATEGORY', 'NAME_SELLER_INDUSTRY' and 'CODE_REJECT_REASON’ have the value 'XNA’.
- Also, some rows in the columns ‘'NAME_CASH_LOAN_PURPOSE' and 'CODE_REJECT_REASON’ have the value ‘XPA’.
- The columns 'NAME_CASH_LOAN_PURPOSE' is important to conduct the analysis. So, we can remove the rows for the values 'XNA' and 'XAP’ in each columns.

Treating negative values:

- In the column 'SELLERPLACE_AREA' we only have '-1' as negative data. This might be mentioned to denote no selling area mentioned in the previous application. So, we can leave this column as it is.

- In the column 'DAYS_DECISION', all values are negative and we are using abs() function to change them into positive values.

Merging two datasets:

- Using inner join function we are merging the both datasets ‘app_df’ and ‘prev_df’ with reference to the column ‘SK_ID_CURR’.
- Now we have 50881 rows and 62 columns in the merged dataset (‘merged_df’).

Finding imbalance in target column:

- Target column having the huge imbalance with the ratio of around 13:87.
- For the target value 0, we have 43927 rows and 62 columns.
- For the target value 1, we have 6954 rows and 62 columns.
- We are having huge imbalance in the target column. So, separating the current data frame (merged_df) into two data frames ‘target0_merged_df’ and ‘target1_ merged _df’.

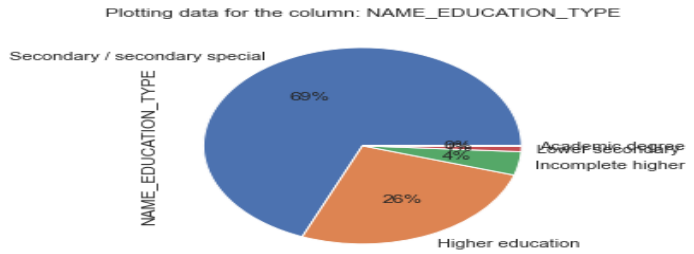
Bivariate analysis:

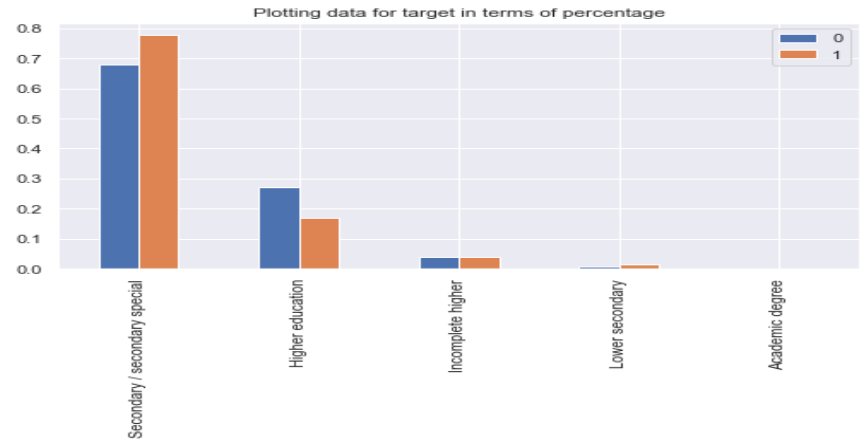
- Creating a user defined function called ‘plotting_bi’ to avoid repeating the same block code.
- In this function, we plot a pie plot which will show the count of each values in the given column.
- A bar plot to show the percentage of counts in each category and separating the data as per the target value 0 and 1.
- Also, two count plots showing data as per the target value 0 and 1 for the given column.

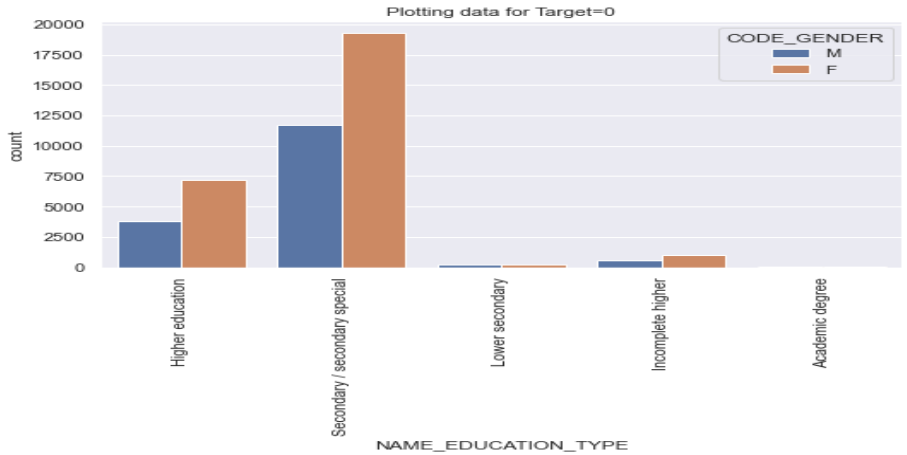
Bivariate analysis:

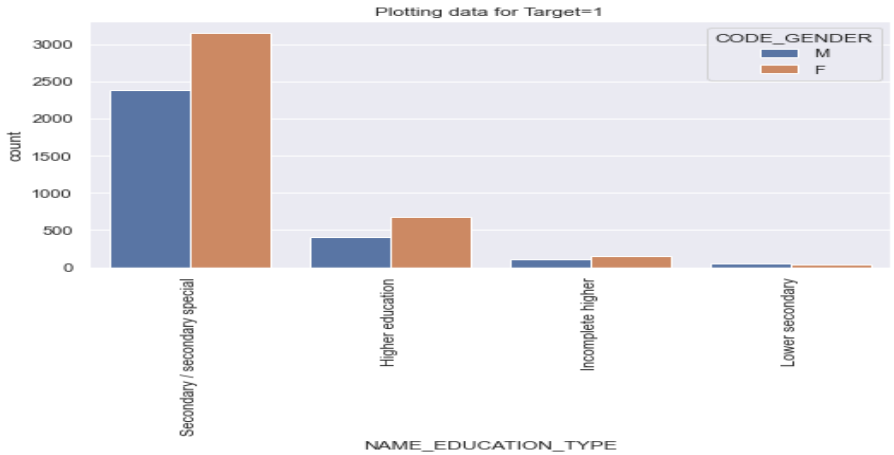
'NAME_EDUCATION_TYPE' and 'CODE_GENDER':

- The clients who have secondary/secondary special education apply for more loans. In this Female gender apply for more loans.





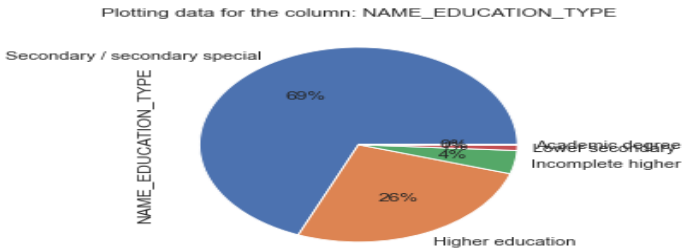


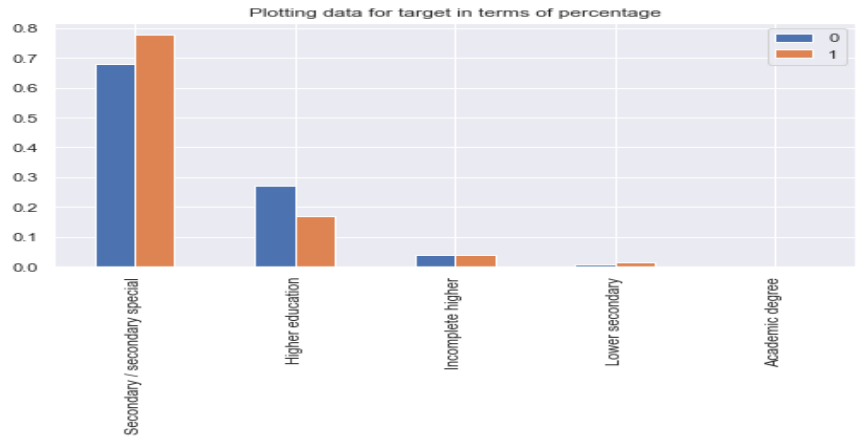


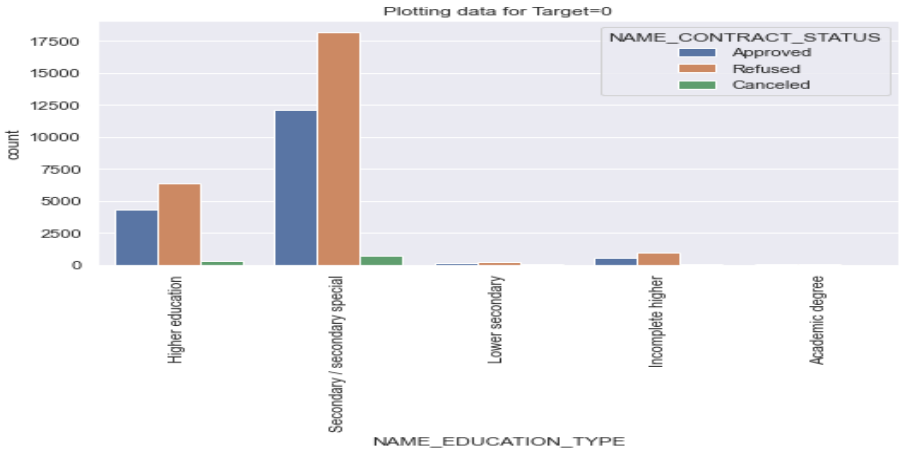
Bivariate analysis:

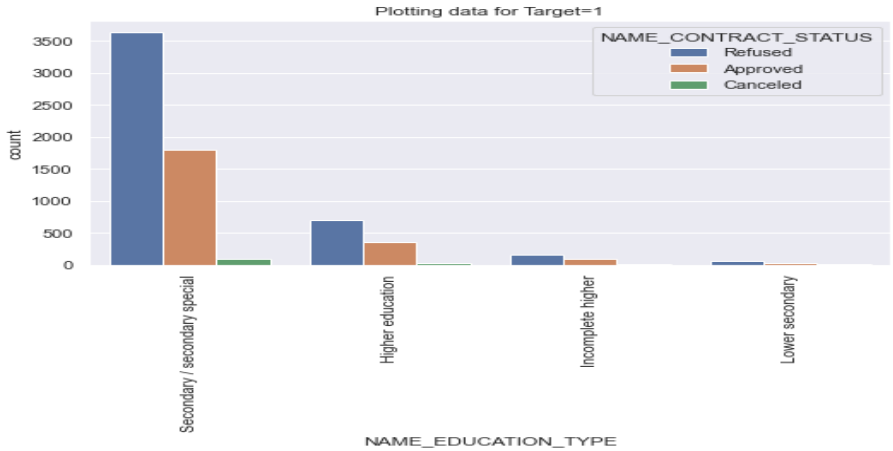
'NAME_EDUCATION_TYPE' and 'NAME_CONTRACT_STATUS':

- The clients have education status higher education and secondary/secondary special are applying the high number of loans and the refusal counts are higher than approved count.





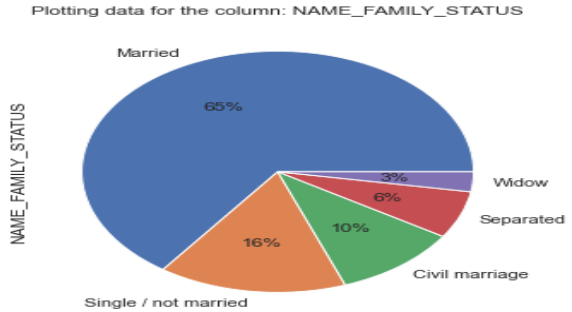


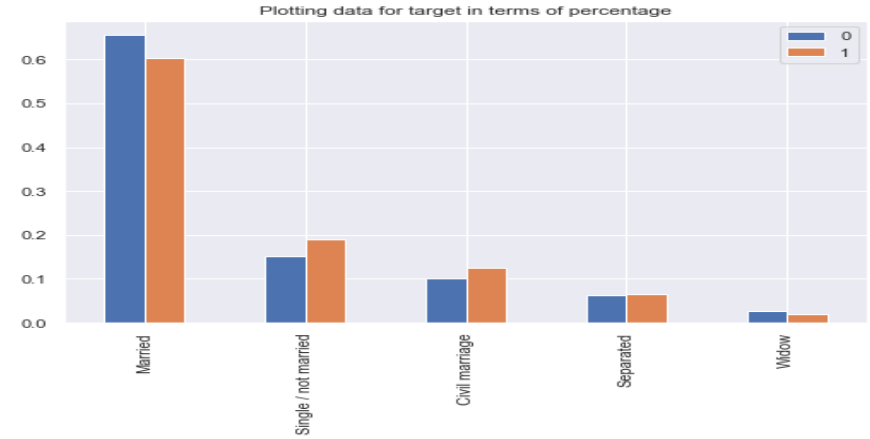


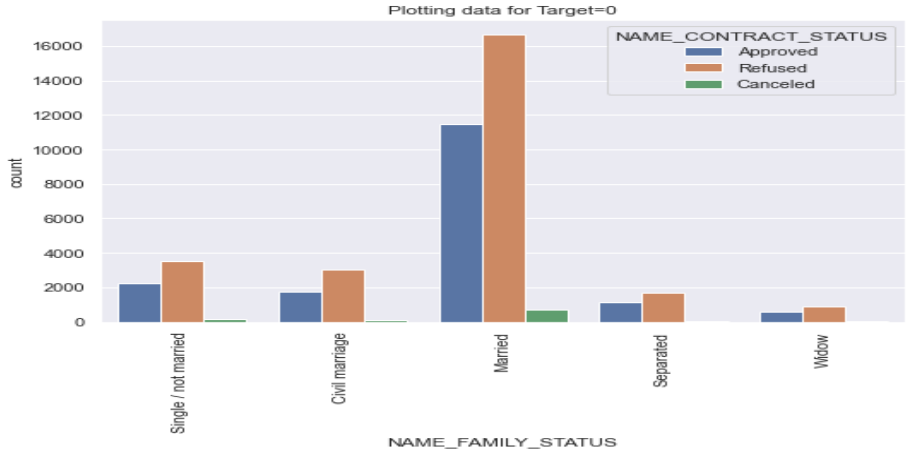
Bivariate analysis:

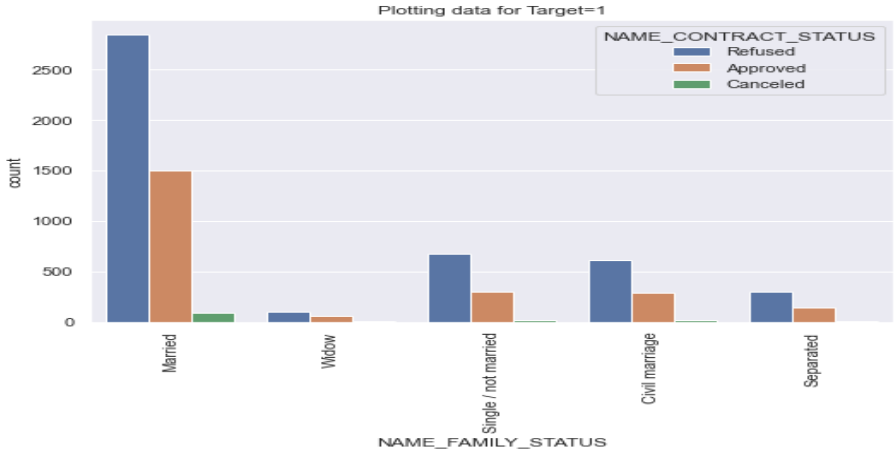
'NAME_FAMILY_STATUS' and 'NAME_CONTRACT_STATUS':

- Higher number of loans are applied by married people.
- Client's who are not married & civil marriage having a high defaulter percentage.





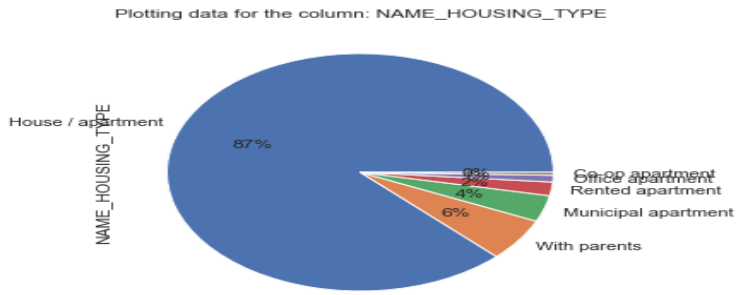


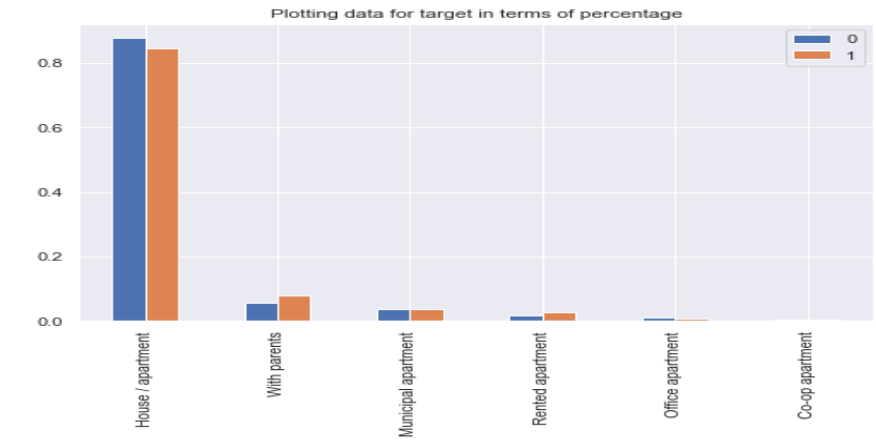


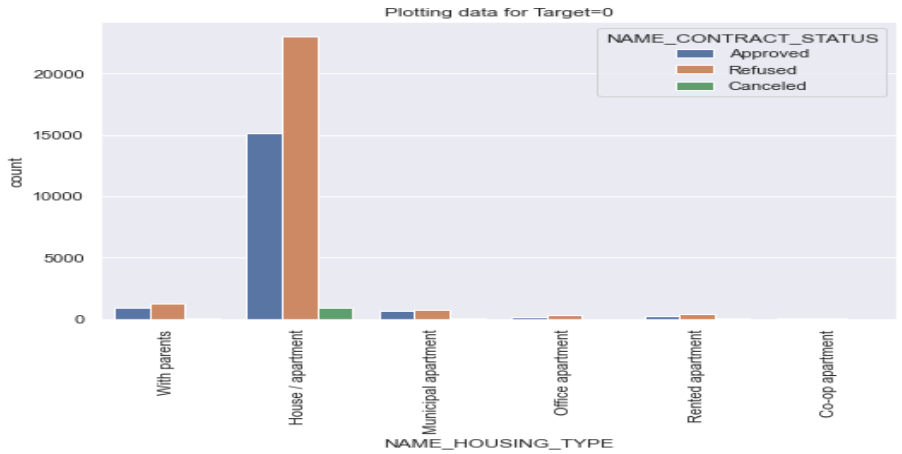
Bivariate analysis:

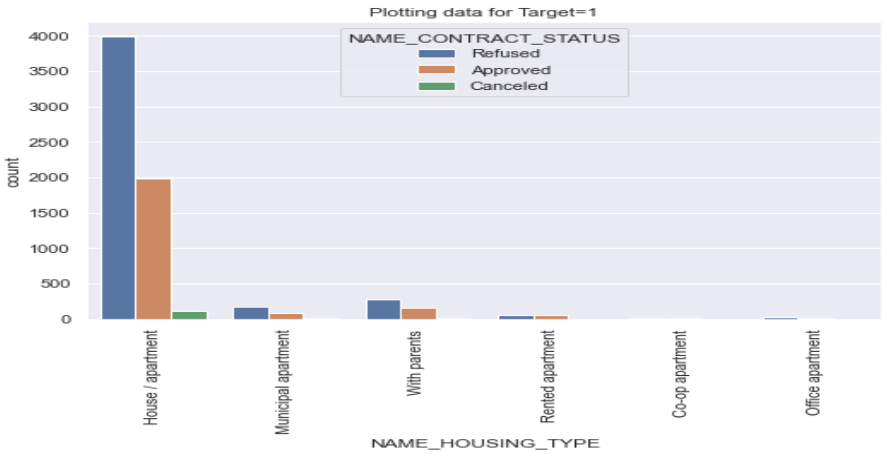
'NAME_HOUSING_TYPE' and 'NAME_CONTRACT_STATUS':

- Clients who are having house/apartments are applying the high number of loans.







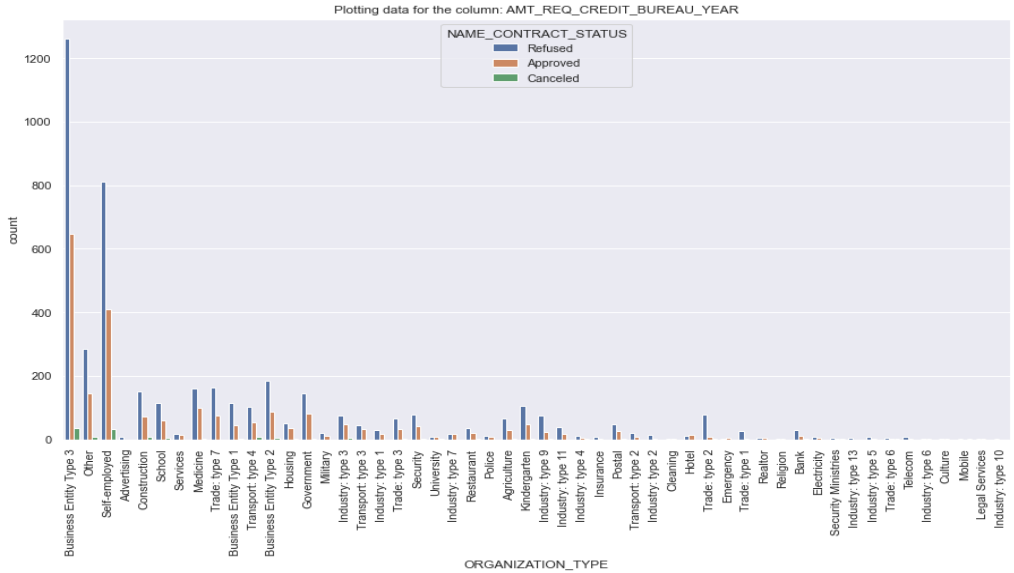


Bivariate analysis:

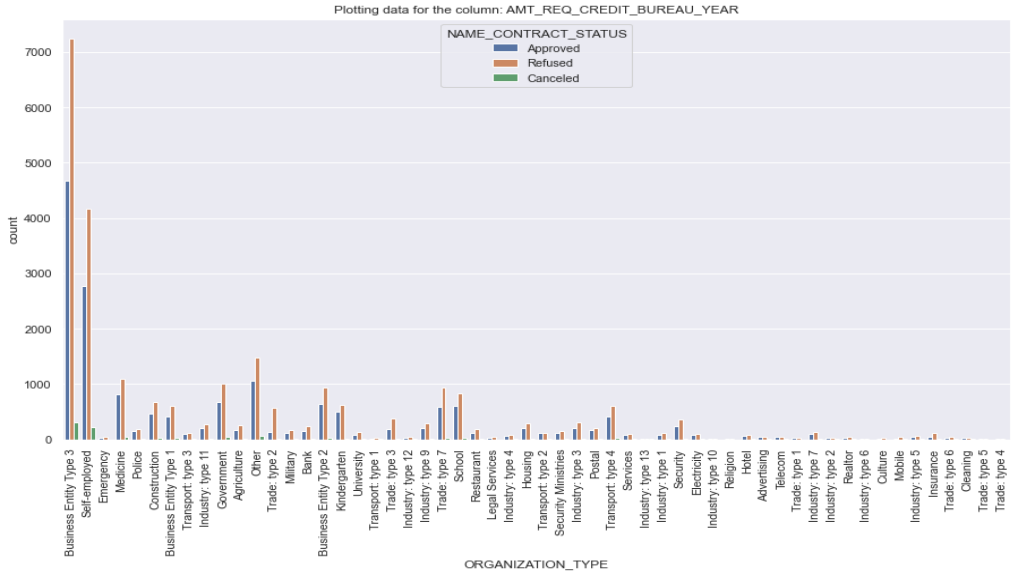
ORGANIZATION_TYPE' and 'NAME_CONTRACT_STATUS':

- Clients who are having organization type as the business entity type-3 and self-employed are applying for more loans and having more defaulters count.

Target value 1



Target value 0



Conclusion:

- Clients who are having secondary education, married, having house or apartment, working in business entity type-3 (Manager, HR and accountants) and self employed who are having the income range between 25K to 225K highly applying for loans.
- Clients who are the gender male, working in business entity type-3 as accountant with academic degree and higher count in family members more likely to have difficulties in payment of the loan.

Thank you