Credit EDA Case Study

By:

Hemant Sawakare

Roll Number - DDS1930106

and

Nirup Sundar Chaudhury

Roll Number - DDS1930188

Date: 02/09/2019

Problem Statement

- Analyze two types of risks associated with the company’s decision i.e., “loss of business” and “financial loss” using EDA.
- Identify patterns indicating clients’ difficulties for repayment of loans.
- Identify the driver variables and understand their significance which are strong indicators of loan default.
- Identify the outliers, if any, in the dataset and justify the same.
- Identify imbalanced data, if any, in the dataset and calculate the ratio of data imbalance.
- Conduct Univariate analysis to understand data (Metadata, data distribution) provided.
- Conduct Segmented Univariate analysis to determine/understand behavior of variables across various segments.
- Conduct bivariate analysis to find relationship between two variables.
- Explain the results of univariate, segmented univariate and bivariate analysis in business terms.
- Identify the top correlation for the client with payment difficulties.
- Summarize the findings using visualizations.

Data Exploration

- ‘**application_data.csv**’ contains all the information about the client at the time of application.
    - The data is about whether a client has payment difficulties.
    - This file contains 307511 rows and 122 columns

- ‘**previous_application.csv**’ contains information about the client’s previous loan data.
    - It contains the data whether the previous application has been approved, cancelled, refused or unused offer.
    - There are 1670214 rows and 37 columns in this dataset.

- **‘columns_description.csv’** is data dictionary which describes the meaning of the variables present in both the datasets i.e., ‘application_data’ and ‘previous_application’.

Data Cleaning

Application Data :

- There are approximately 49 columns with more than 45% missing values, we will be filtering them out for our analysis.
- FLAG_DOCUMENT and EXT_SOURCE do not provide any analytical value and hence we will remove it as well.
- Final shape of application dataframe after cleaning is 307511 rows * 51 columns.
- **Impute Data -** We are going to impute below columns with 0 since Min, 25, 50 and 75% is 0 for these columns
  1. AMT_REQ_CREDIT_BUREAU_MON
  2. AMT_REQ_CREDIT_BUREAU_WEEK
  3. AMT_REQ_CREDIT_BUREAU_DAY
  4. AMT_REQ_CREDIT_BUREAU_HOUR
  5. AMT_REQ_CREDIT_BUREAU_QRT
  6. DEF_30_CNT_SOCIAL_CIRCLE
  7. DEF_60_CNT_SOCIAL_CIRCLE

Previous Application Data :

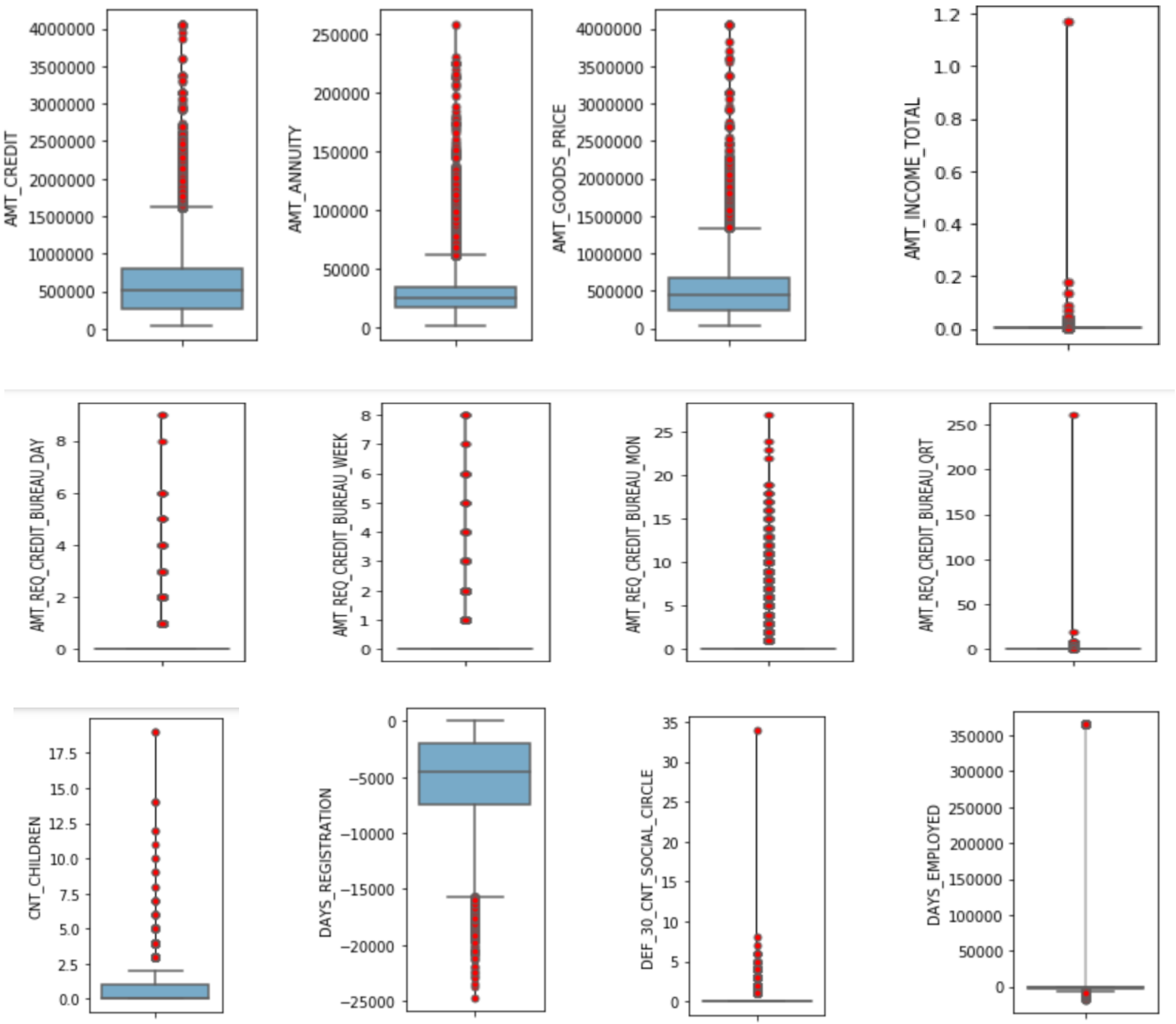
- There are a couple of with more than 90% missing values, we will be filtering them out for our analysis.
- Final shape of application dataframe after cleaning is 1670214 rows and 37 columns
- It will be incorrect to impute values in previous application data since it has data for all the contract status. We will not impute value for previous application data.

Derived Metrics

- Few columns has been derived from application data (‘application_clean’ dataframe) and previous application data (‘previous_clean’) in order to normalize, simplify and derive more meaningful insights from the data.
- Derived metrices are also used to optimize the data. For example, we can convert a variable containing number of days in 1000s to years which will be easy to interpret.
- Following columns have been derived from ‘application_clean’ dataframe;
  - AGE_YEAR for Age in years
  - EMPLOYED_YEAR for Employed in years
  - INCOME_CREDIT_RT for INCOME to CREDIT Ratio
  - LTV_RT (Loan-to-value) for GOODS_PRICE to CREDIT Ratio
- Following columns have been derived from ‘previous_clean’ dataframe;
  - pr_LTV_RT (Loan-to-value) for GOODS_PRICE to CREDIT Ratio

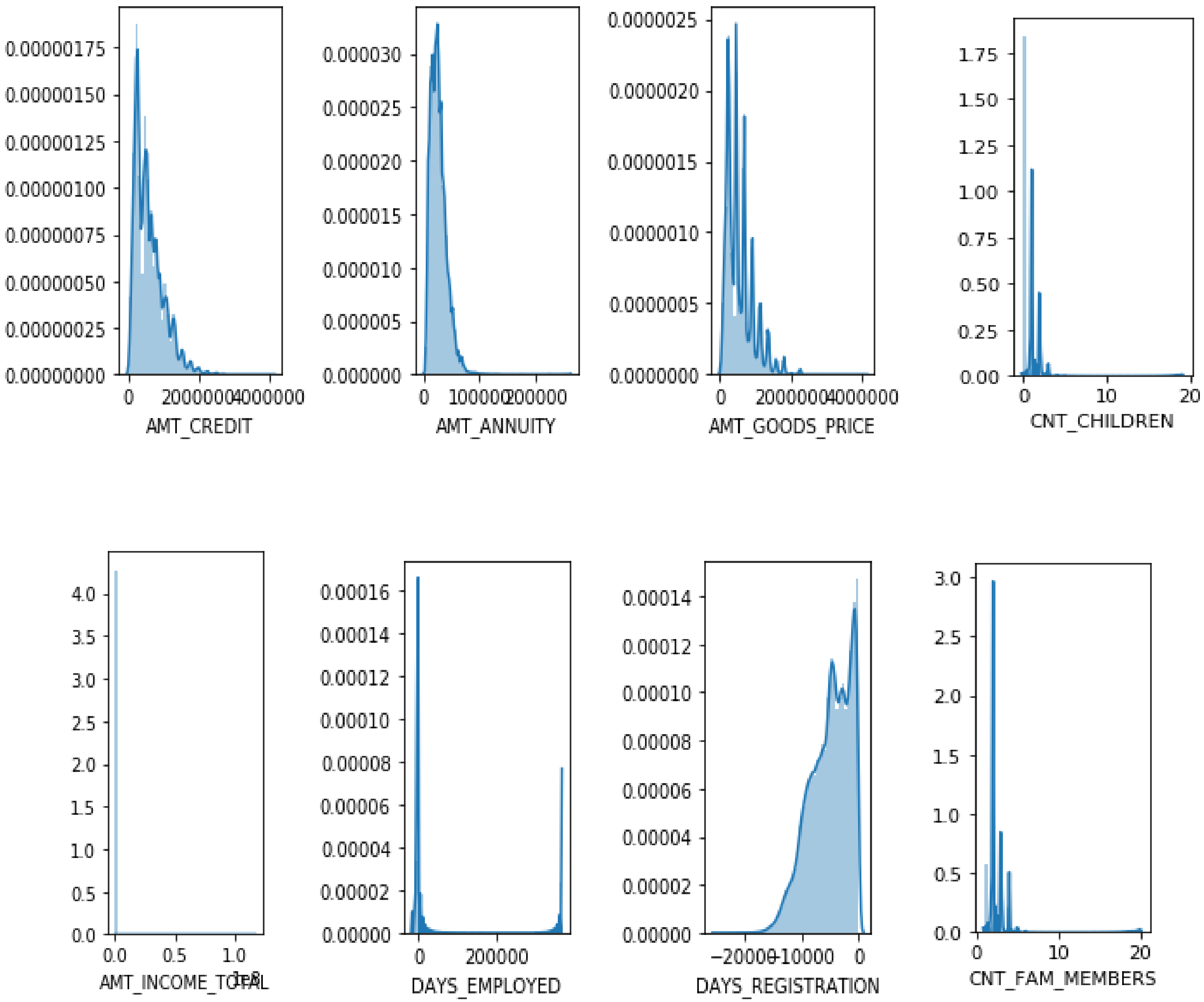
Univariate Analysis – Application Data Outliers

- Univariate analysis revealed data distribution and outliers in application data. Key columns where outliers were identified are:-
    - a. CNT_CHILDREN
    - b. AMT_INCOME_TOTAL
    - c. AMT_CREDIT
    - d. AMT_ANNUITY
    - e. AMT_GOODS_PRICE
    - f. DAYS_EMPLOYED
    - g. DAYS_REGISTRATION
    - h. CNT_FAM_MEMBERS
    - i. DEF_30_CNT_SOCIAL_CIRCLE
    - j. AMT_REQ_CREDIT_BUREAU_*
- Outliers result in incorrect metrices and hence we will be removing them from our dataset for analysis purposes.



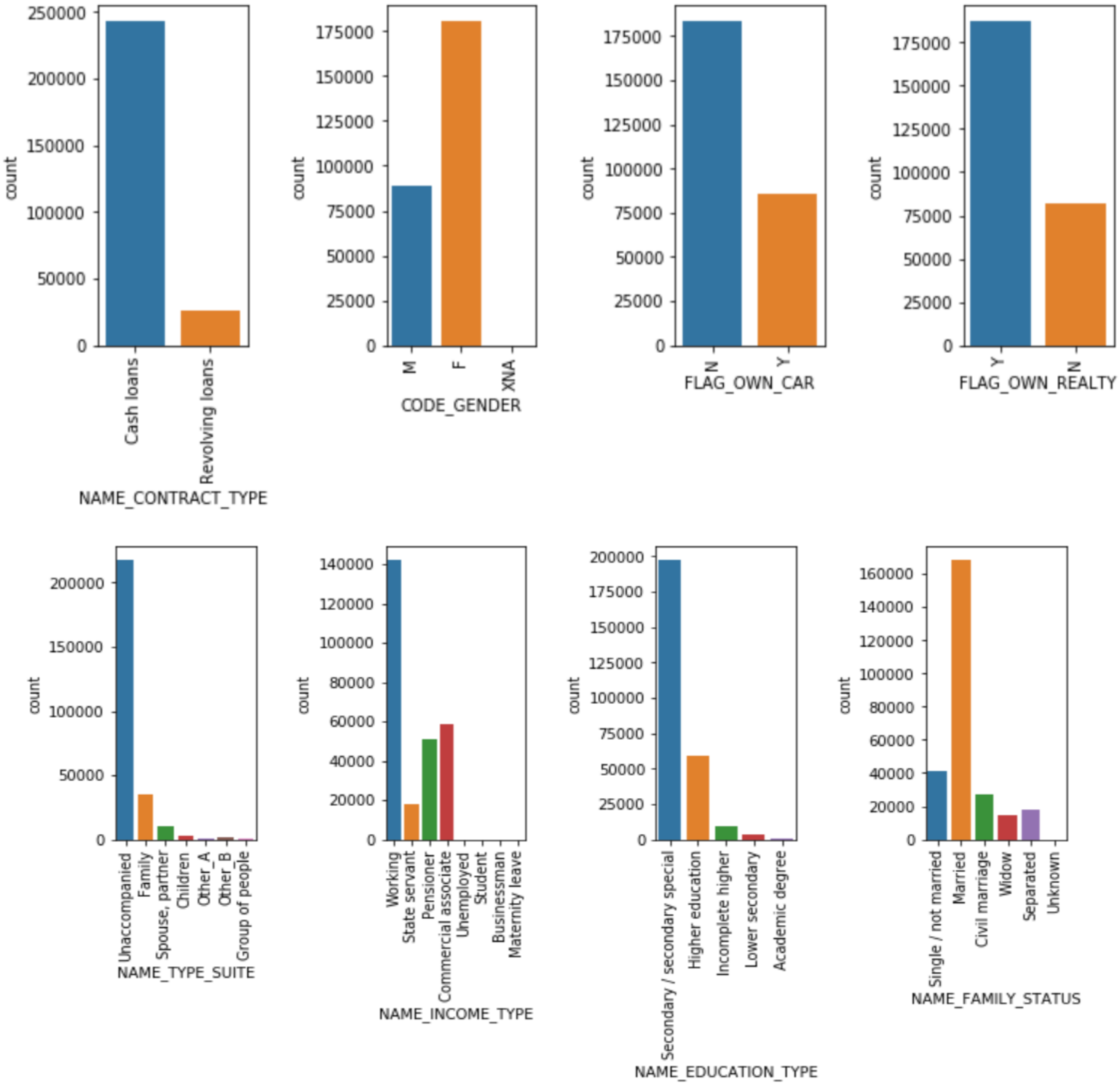
Univariate Analysis – Application Data (Quantitative variables) – Remove Outliers

- Normal distribution of key columns is show here.
- We will below logic to determine and remove outliers.
  - Calculate Inter Quantile Region (75th Percentile - 25th Percentile)
  - Upper cutoff = 75th percentile + 1.5 * (IQR)
  - Lower cutoff = 25th percentile - 1.5 * (IQR)
- Note - We will not remove null values since it might result in removing entire categorical value. eg. Removing null values from AMT_CREDIT from previous_application results in removing "CANCELLED, REFUSED...." Status.
- **Note** - DAYS_EMPLOYED is populated as "365243" for below condition
  - NAME_INCOME_TYPE = "Pensioner and Unemployed"
  - ORGANIZATION_TYPE = "XNA"



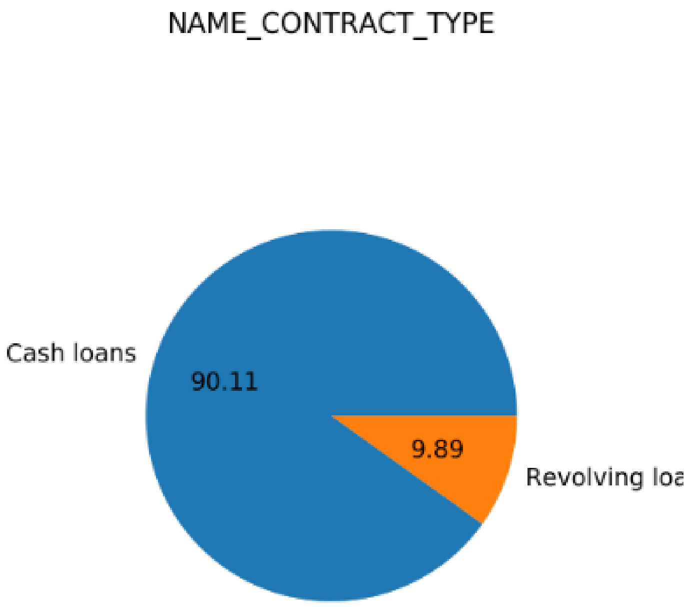
Univariate Analysis – Application Data (Categorical variables)

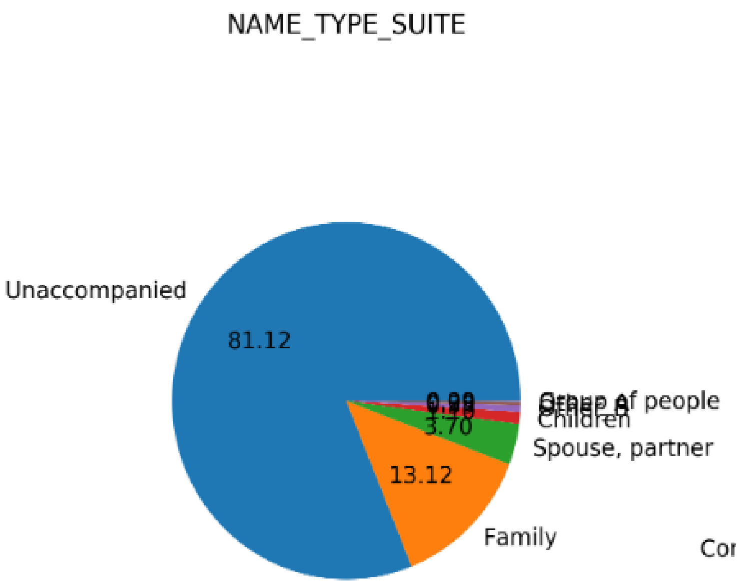
- Univariate analysis has been performed on categorical variables present in application data.
- Number of cash loans are abysmally large with compare to revolving loans.
- Female candidates are applying approximately 60% more in numbers than that of male.
- More than half of the current application numbers are from married clients.
- The clients having profession (Income Type) within ‘working’ class has applied the most.
- More than 70% of population requesting for loan have secondary/ secondary special education.

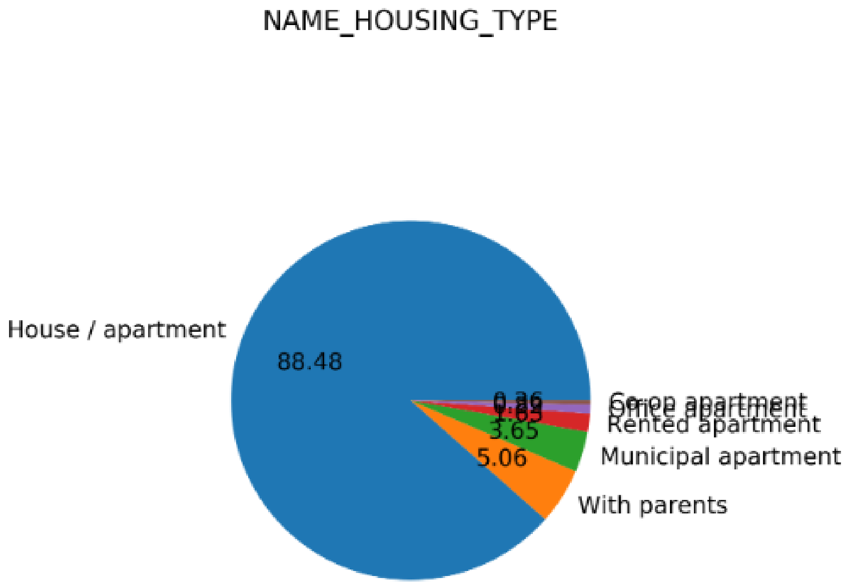


Univariate Analysis – Application Data - Data Imbalance

- If there are two classes, then balanced data would mean 50% points for each of the class. For most Exploratory Data Analysis, little imbalance is not a problem.
- For our case study we will consider any data with 80% of population of column for 1 value as Data Imbalance.
- Quick review of count plot and pie plot indicate clear imbalance for below columns.
  - CONTRACT_TYPE - Over 90% of the population is for 'Cash loans'
  - NAME_TYPE_SUITE - Over 81% of the population is for 'Unaccompanied'
  - NAME_HOUSING TYPE - Over 81% of the population is for 'House / apartment'
- **How will we deal with it?**
  - All analysis on these categorical variables will be using percentage/average instead of absolute value.
  - Value for which contribution is significantly less will not be used to derive any conclusion.







Univariate Analysis – Previous Data Outliers

- Univariate analysis revealed outliers in application data. Key columns where outliers were identified are:-

  a. AMT_CREDIT

  b. AMT_ANNUITY

  c. AMT_GOODS_PRICE

  f. AMT_ APPLICATION

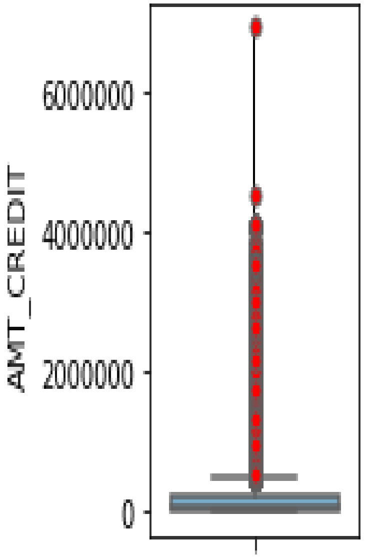
  g. AMT_DOWN_PAYMENT

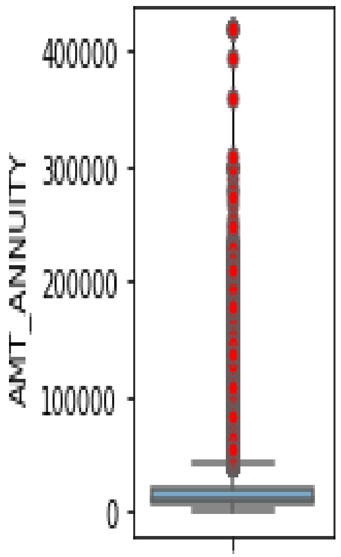
  h. RATE_DOWN_PAYMENT

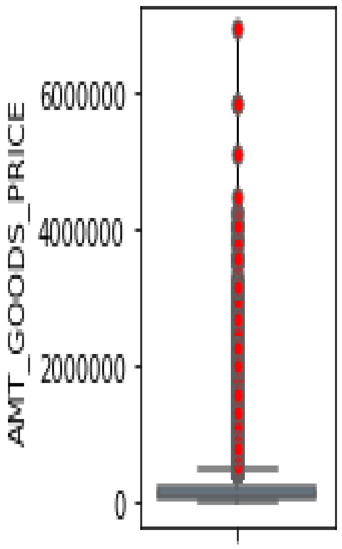
  i. CNT_PAYMENT

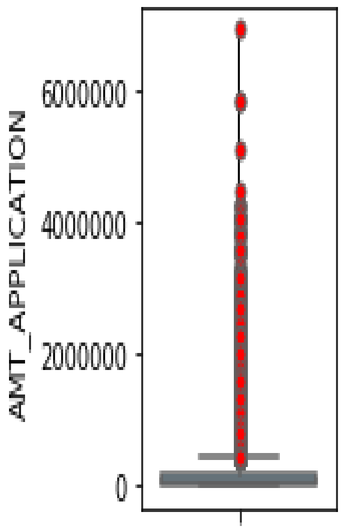
  J. pr_LTV_RT

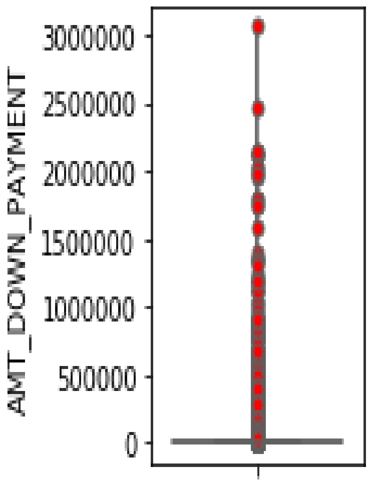
- Outliers result in incorrect metrices and hence we will be removing them from our dataset for analysis purposes.

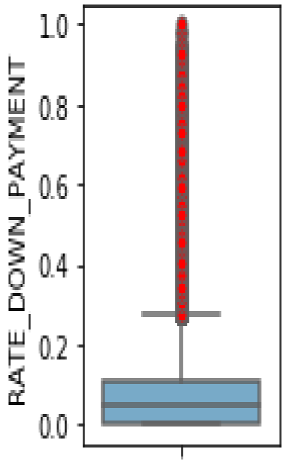


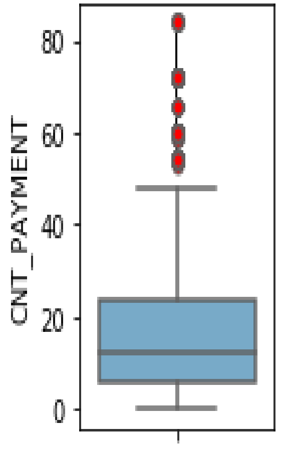


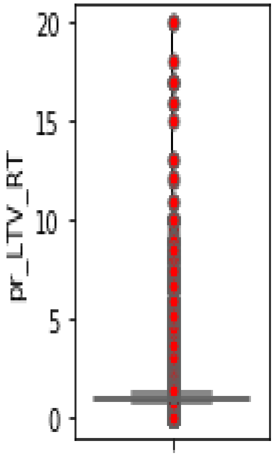






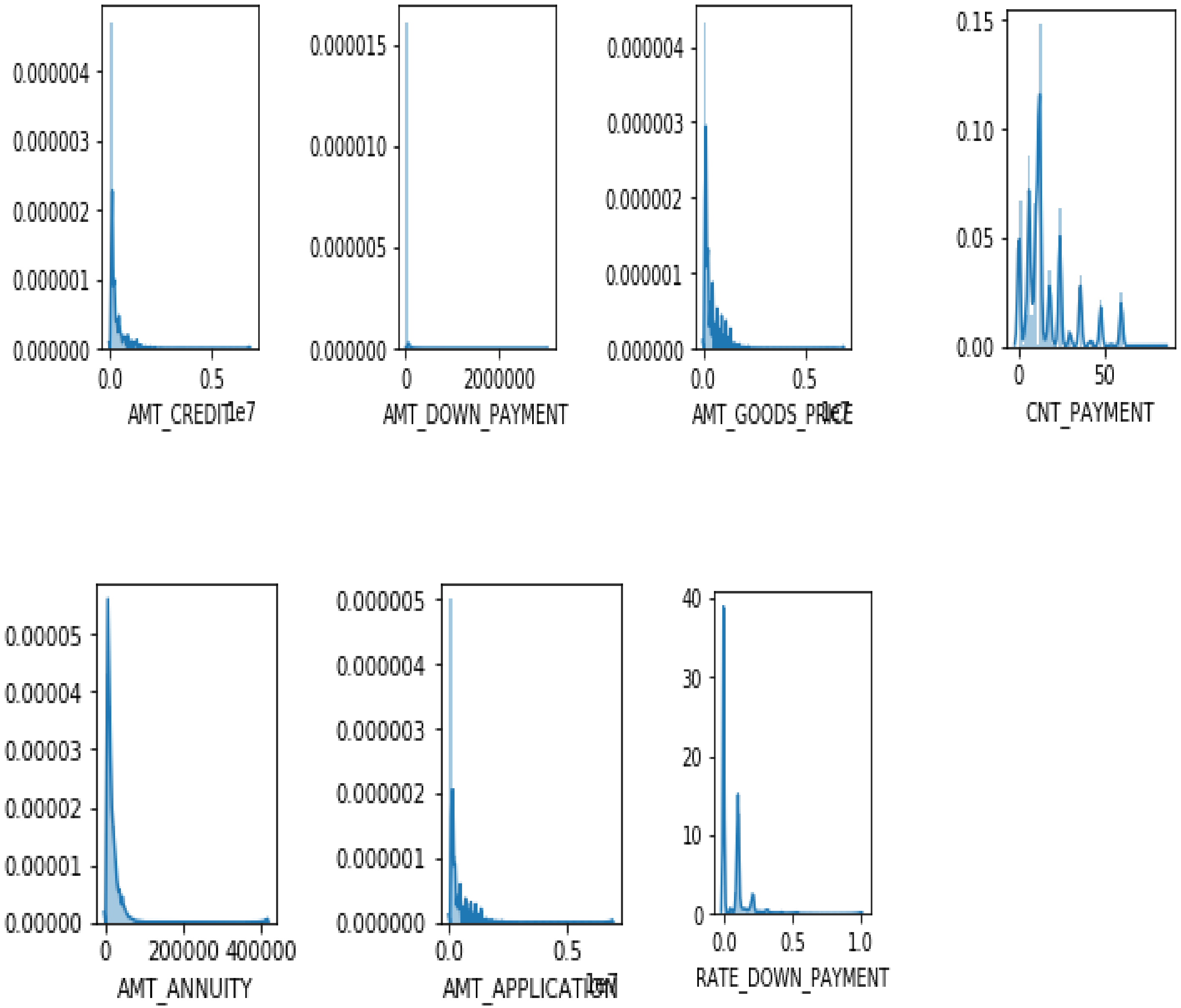






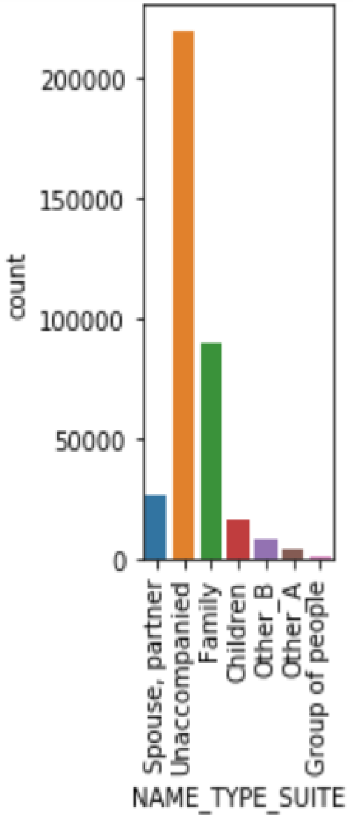
Univariate Analysis – Previous Data (Quantitative variables) – Remove Outliers

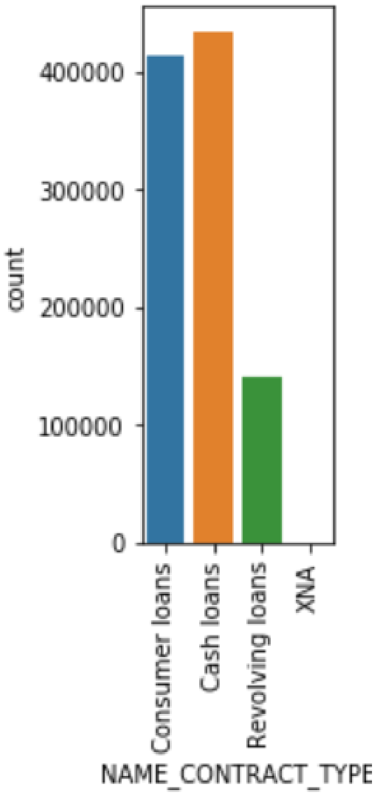
- Univariate analysis revealed outliers in the data.
- Normal distribution of key columns is show here.
- We will below logic to determine and remove outliers.
  - Calculate Inter Quantile Region (75th Percentile - 25th Percentile)
  - Upper cutoff = 75th percentile + 1.5 * (IQR)
  - Lower cutoff = 25th percentile - 1.5 * (IQR)
- Note - We will not remove null values since it might result in removing entire categorical value. eg. Removing null values from AMT_CREDIT from previous_application results in removing "CANCELLED, REFUSED...." Status.

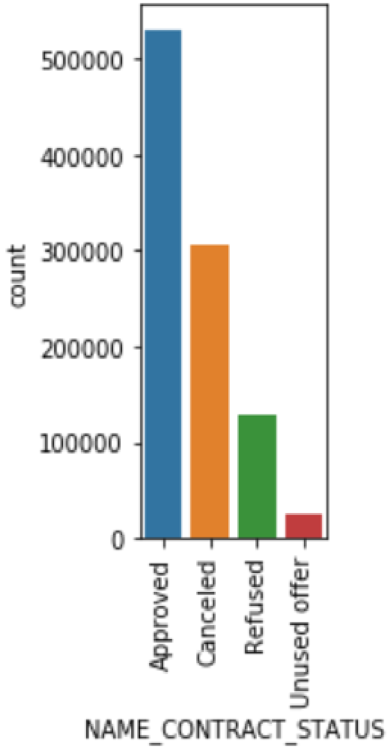


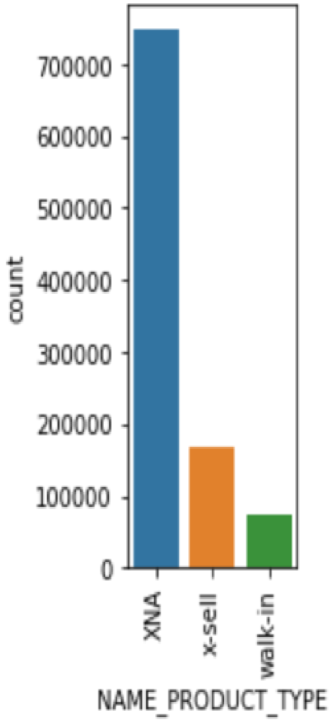
Univariate Analysis – Previous Data (Categorical variables)

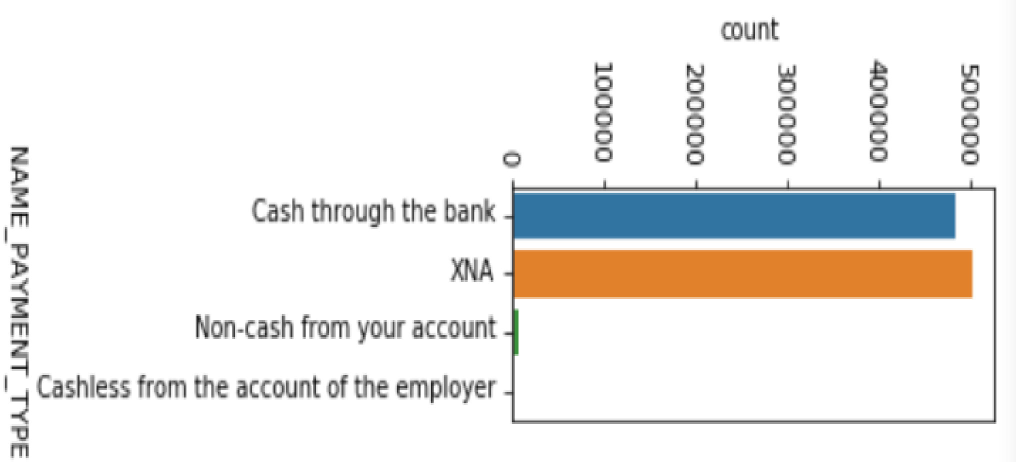
- Univariate analysis has been performed on categorical variables present in previous data.
- 75% of the clients were ‘repeater’ which means they had previously availed the loan.
- Cash loans were the most commonly used contract type followed by consumer loan and revolving loan.
- In previous data, least number of applications were through walk-in.
- In terms of percentage, 53% of the loan applications were approved according to the previous data.
- Most of the clients were not accompanied by anybody while applying for the loan.

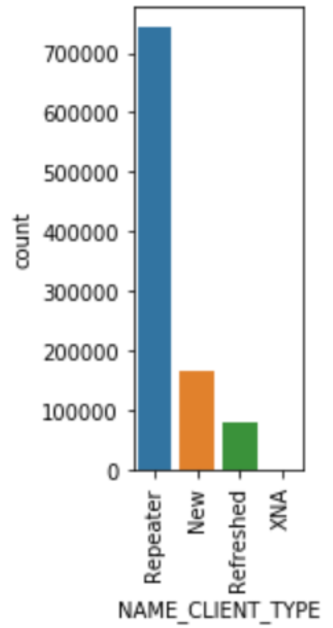












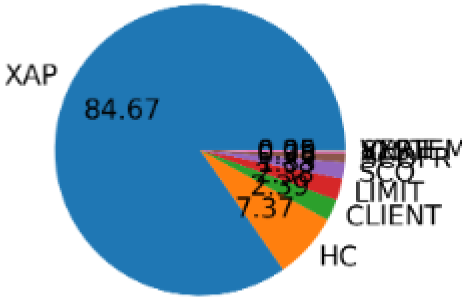
Univariate Analysis – Previous Data - Data Imbalance

- Quick review of countplot and pie plot indicate clear imbalance for below columns.
  - NAME_CLIENT_TYPE - Over 75% of the population is old consumers.
  - NAME_CONTRACT_STATUS - Over 53% of the applications are being approved.
  - NAME_TYPE_SUIT - Over 60% of the population is unaccompanied.
  - FLAG_LAST_APPL_PER_CONTRACT - Over 99% of the population is for 'Y'
  - CODE_REJECT_REASON - Over 81% of the population is for 'XAP'

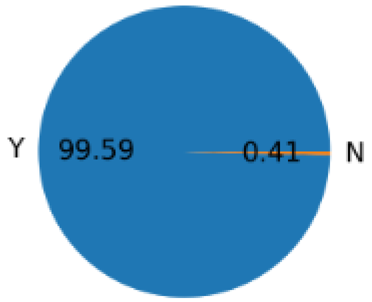
How will we deal with it?

- All analysis on these categorical variables will be using percentage/average instead of absolute value.
- Value for which contribution is significantly less will not be used to derive any conclusion.

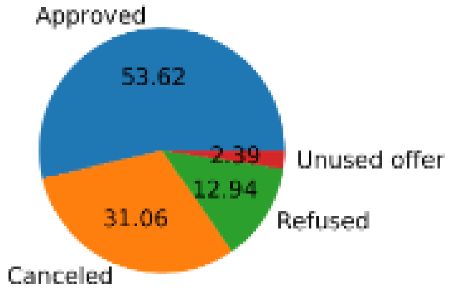
CODE_REJECT_REASON



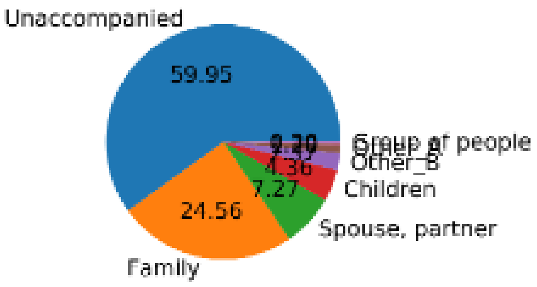
FLAG_LAST_APPL_PER_CONTRACT



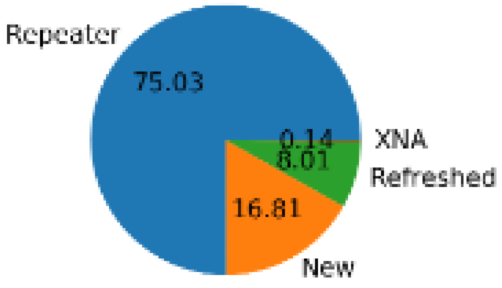
NAME_CONTRACT_STATUS



NAME_TYPE_SUITE

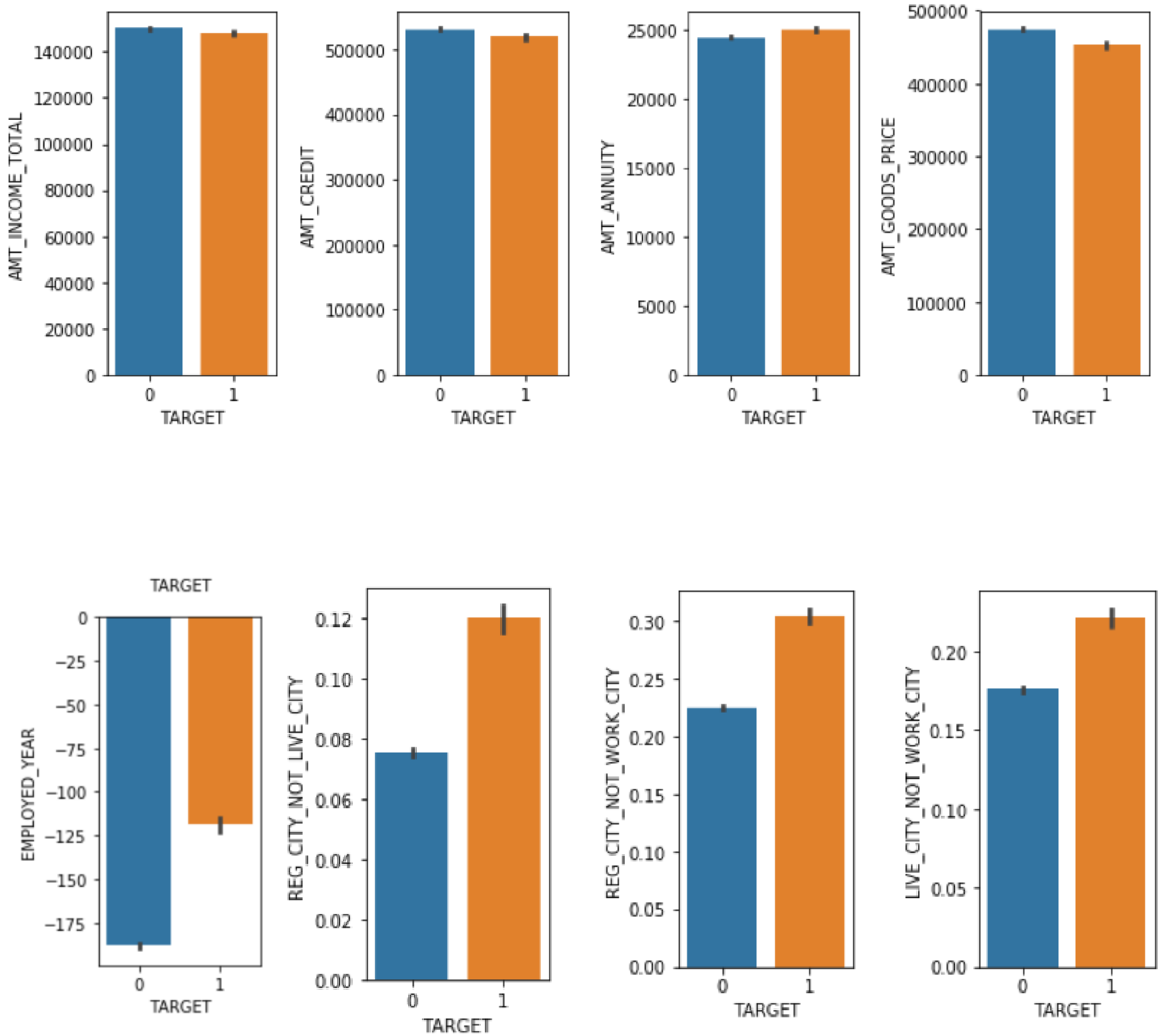


NAME_CLIENT_TYPE



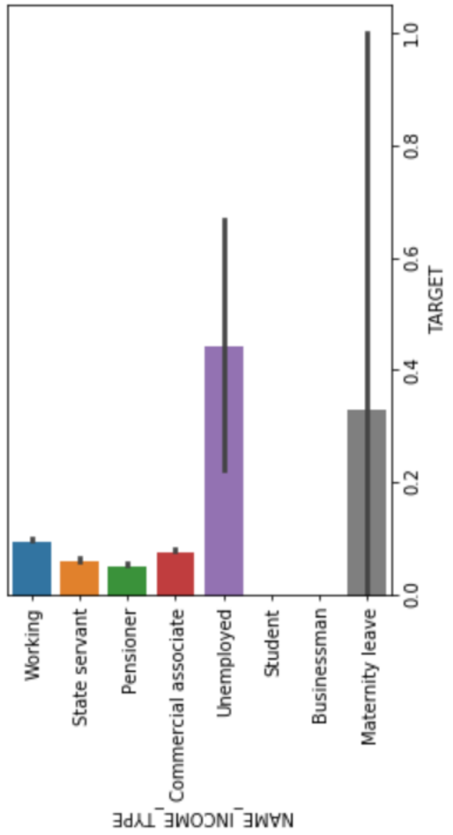
Segmented Univariate Analysis - Quantitative variables (current application data)

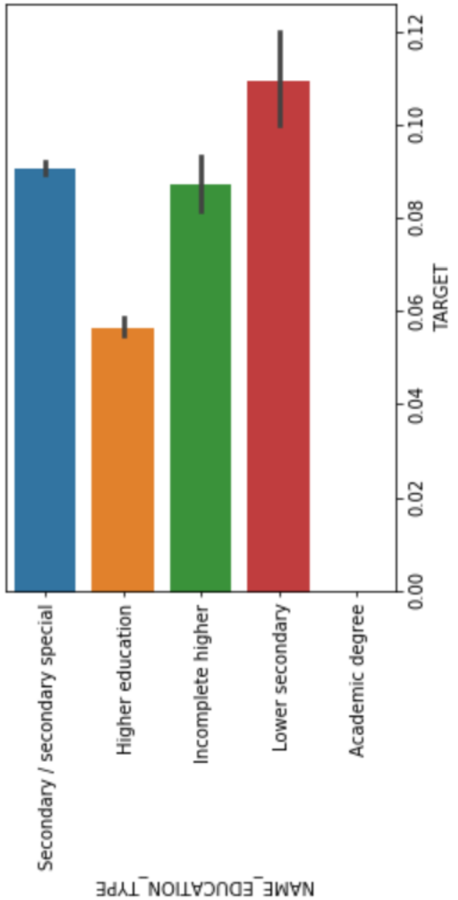
- Segmented univariate analysis has been done using different quantitative variables.
- People who face difficulty in payment have lesser employment history i.e., they have been employed for lesser time compared to group who do not have difficulty in payment.
- There isn’t significant change to below variables with respect to TARGET (i.e. difficulty in payment)
  - AMT_CREDIT
  - AMT_ANNUITY
  - AMT_GOODS_PRICE
- Value of “1” in below variables raises concern of fraud since contact address does not fall under same city as either permanent address or work address
  - REG_CITY_NOT_LIVE_CITY
  - REG_CITY_NOT_WORK_CITY
  - LIVE_CITY_NOT_WORK_CITY

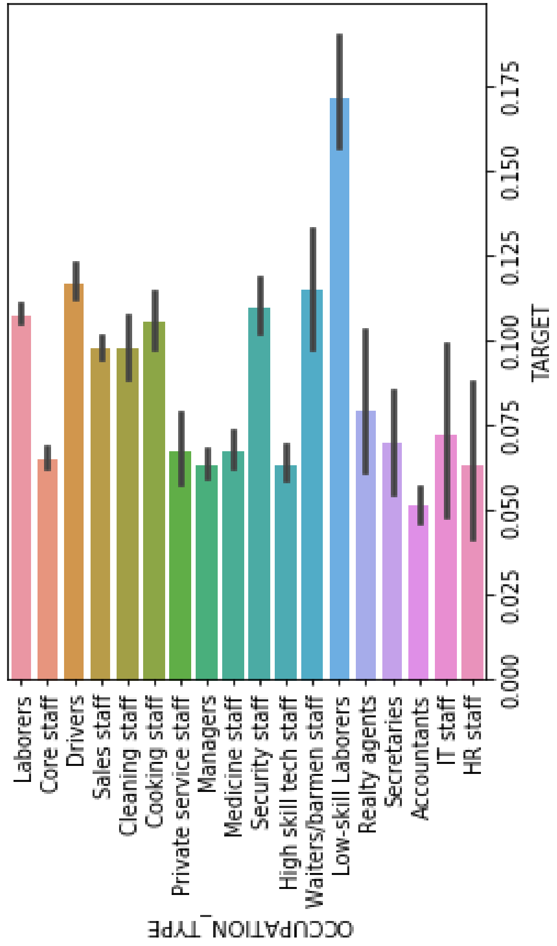


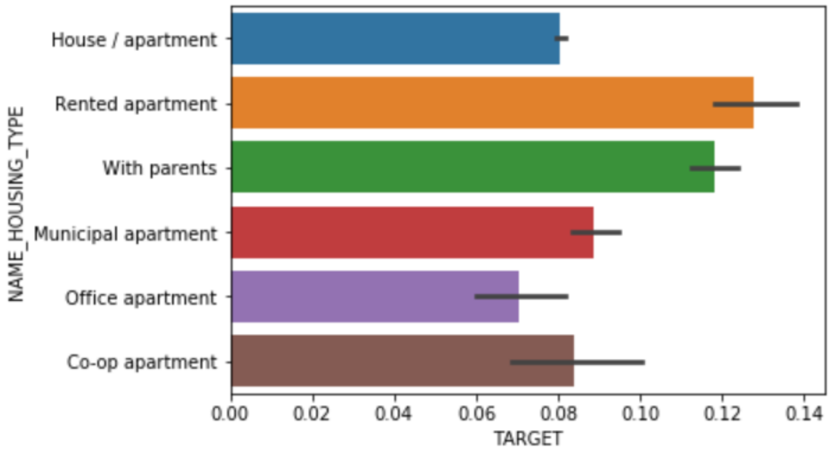
Segmented Univariate Analysis - Categorical variables (current application data)

- Segmented univariate analysis has been done using different categorical variables of current application data.
- The highest proportion of defaulters belong to low-skill laborer category. Accountants professionals default the least.
- Group which is either unemployed or on maternity leave face greater difficulty in payment.
- Population staying in “Rented apartment” or “With parents” have greater difficult in payment.
- People with Lower Secondary education face greater difficulty in payment.



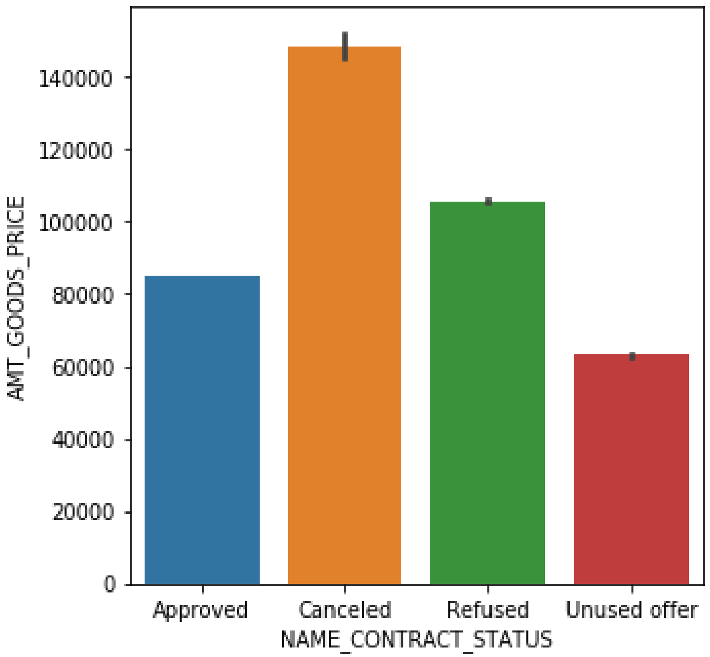


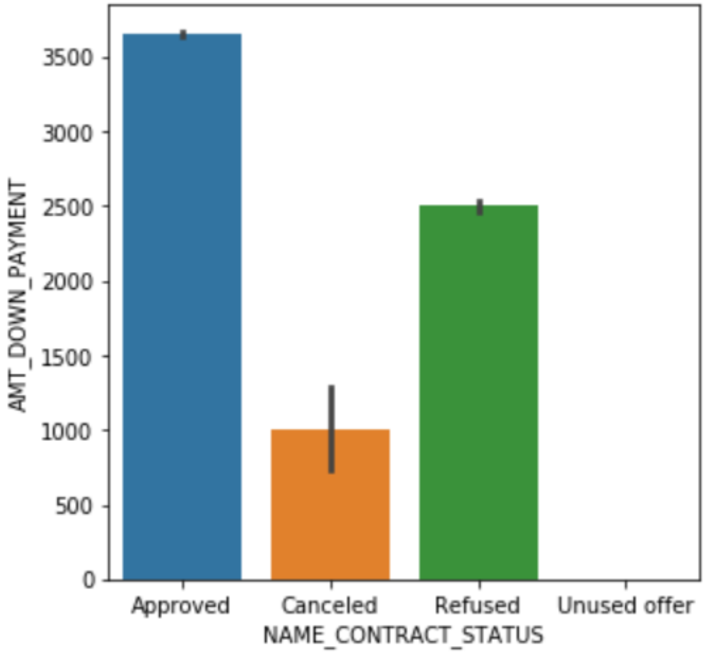


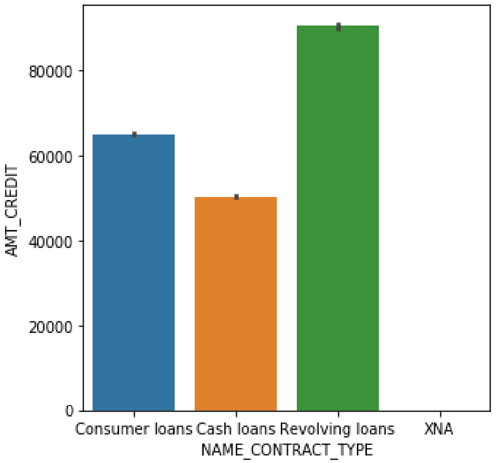


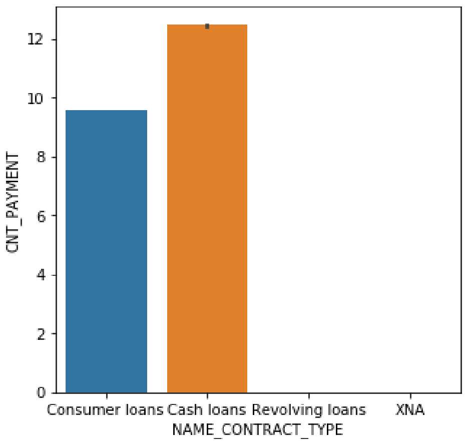
Segmented Univariate Analysis (previous application data)

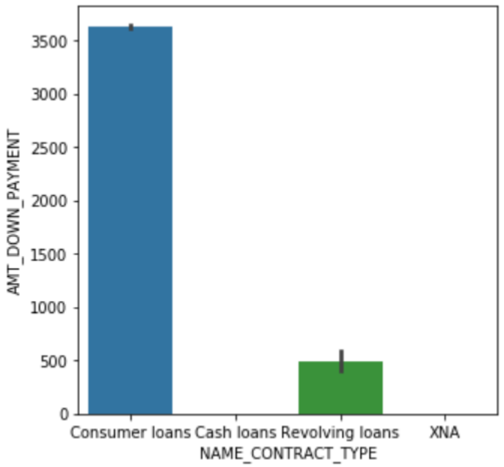
- Customers who had Cancelled their loan had requested loan against high valued goods.
- Customers who had their loan Approved had made greater down payment.
- Amount of credit is highest for revolving loans and lowest for cash loans.
- Cash loans are likely to have longer term of credit(Number of payments).
- Down payment for consumer loan is higher than that for revolving loan.





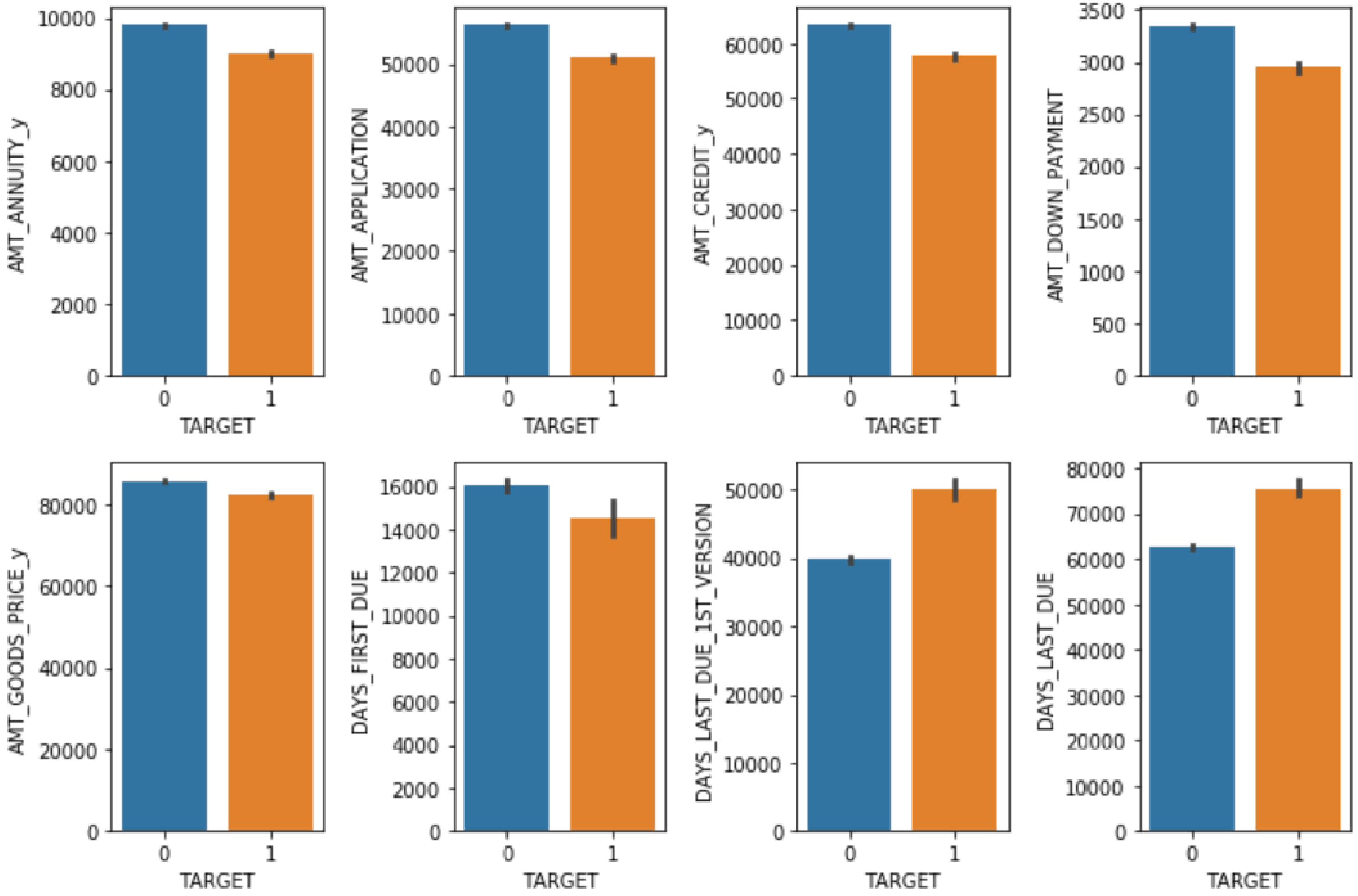






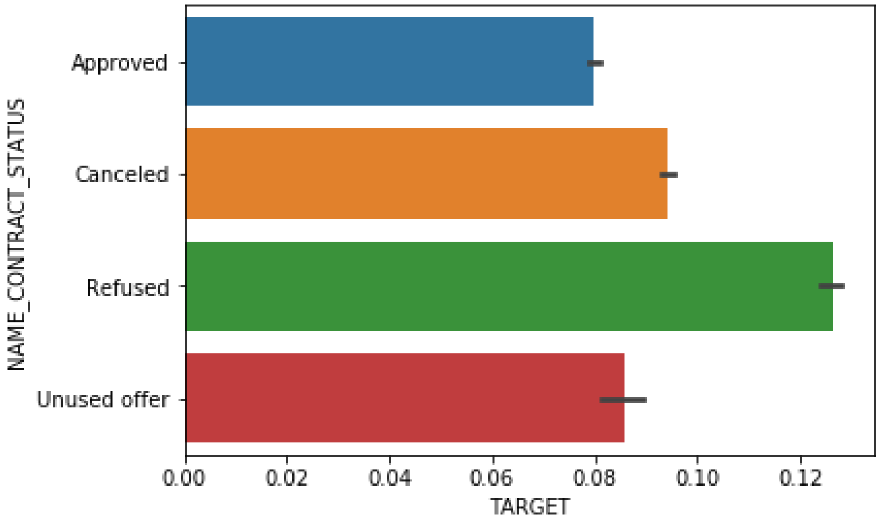
Segmented Univariate Analysis (all app data)

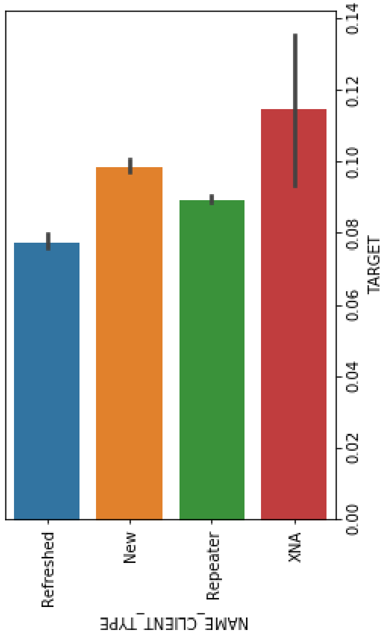
- Segmented univariate analysis has been performed on merged data (combination of previous data and current application data) to study the impact of ‘target’ on quantitative variables.
- Group with greater difficult in payment has lesser down payment on previous application.
- Low down-payment amount is associated with more bad loans compare to high down-payment amount.
- Longer the duration gap of last due date of previous application higher the chances of default.

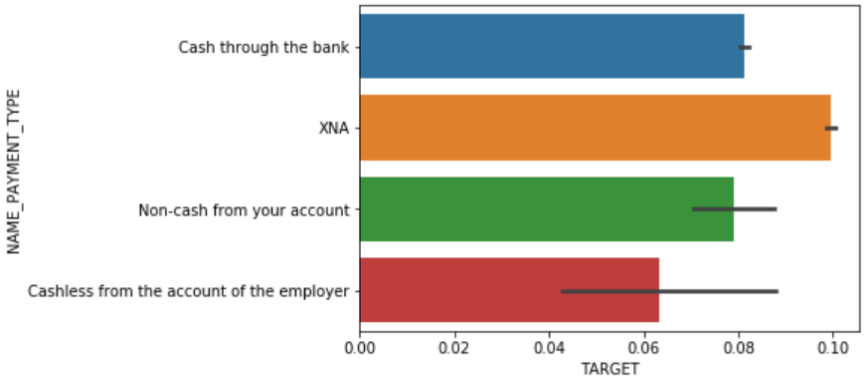


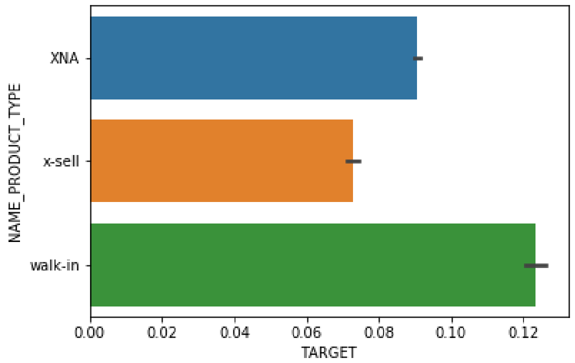
Segmented Univariate Analysis (all app data)

- Segmented univariate analysis has been performed on merged data (combination of previous data and current application data) to study the impact of ‘target’ on categorical variables.
- The clients having no previous record or applying for the first time are more likely to default.
- The clients with greatest difficult in payment had their previous application "Refused" whereas Group with least difficult in payment were "Approved“.
- The clients with least difficult in payment were offered previous loan via cross-sell(x-sell)
- Group that chose “cashless from account of the employer” has least difficult in payment.



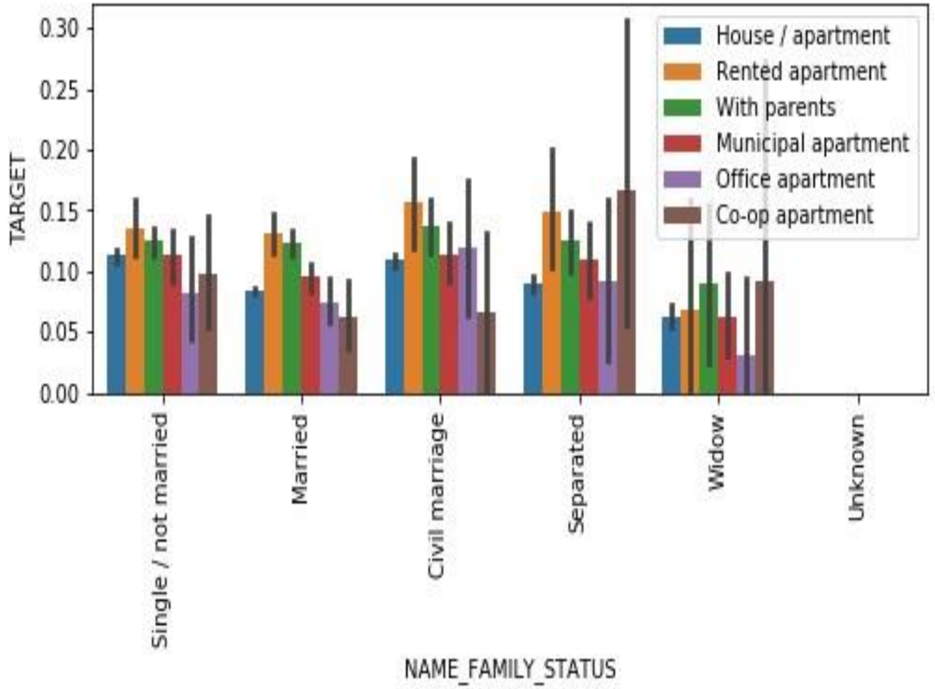


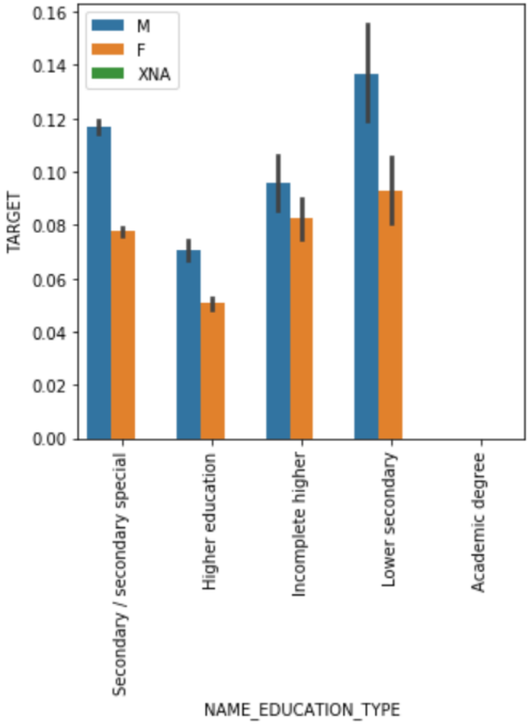


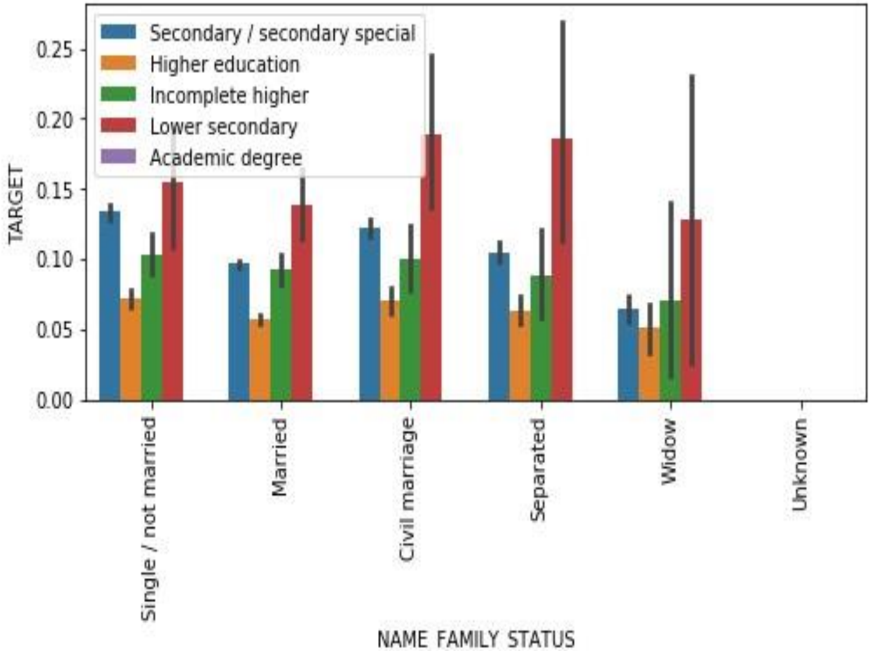


Bivariate Analysis: Categorical variables

- **Impact of Family Status and Housing type on TARGET -** People who are separated and living in co-operative apartments have greater difficulty in payment.
- **Impact of Family Status and Education type on TARGET –**
  - Clients who are widow and highly educated have less difficulty in payment.
  - Clients who are separated or with civil marriage with qualification as lower secondary have more difficulty in payment while higher educated clients can repay easily.
- **Impact of Gender and Education type on TARGET –** Across gender people with lower secondary education show greater difficulty in payment.

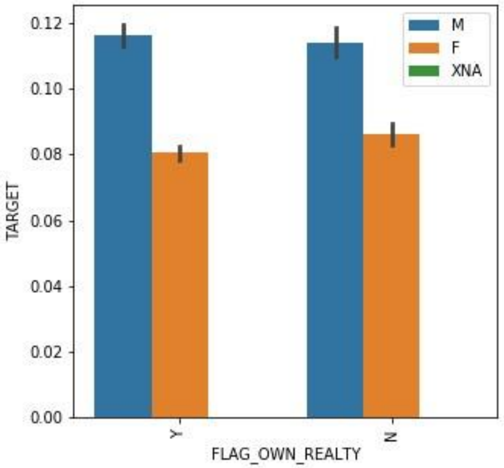


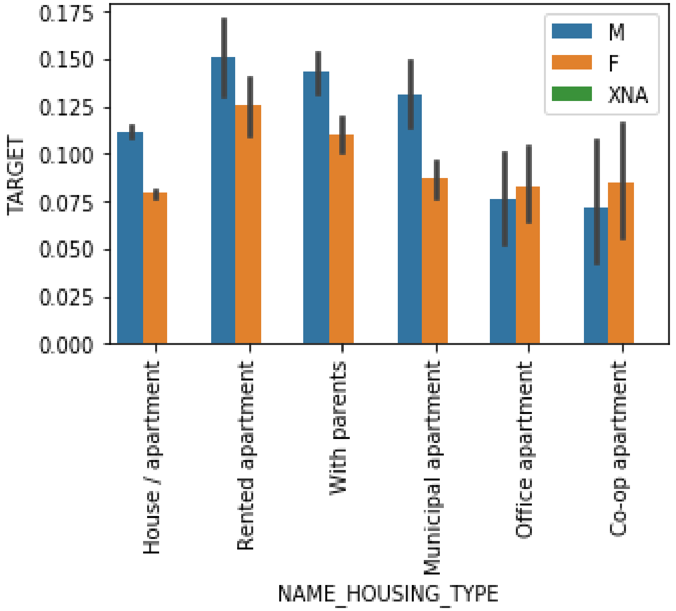


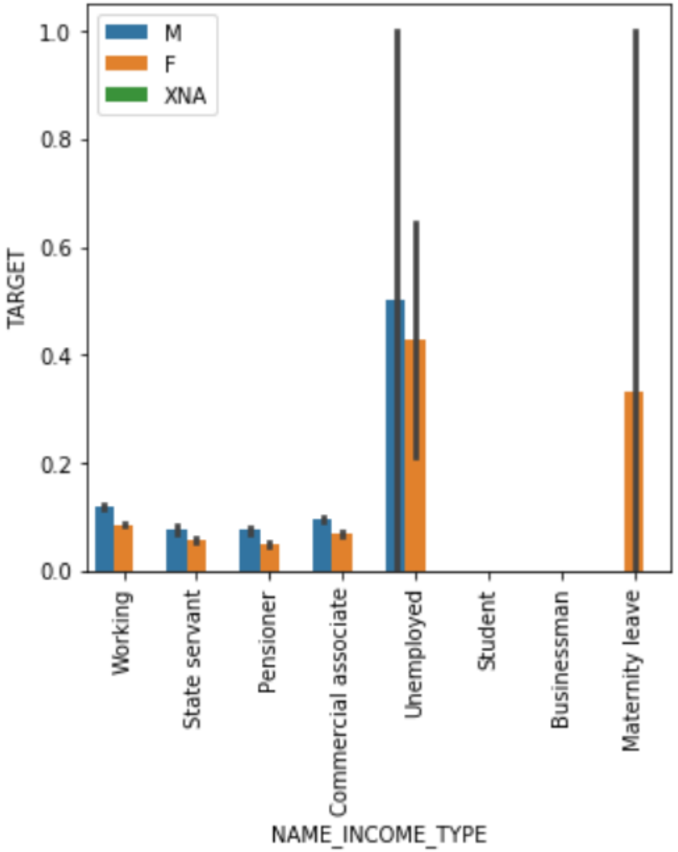


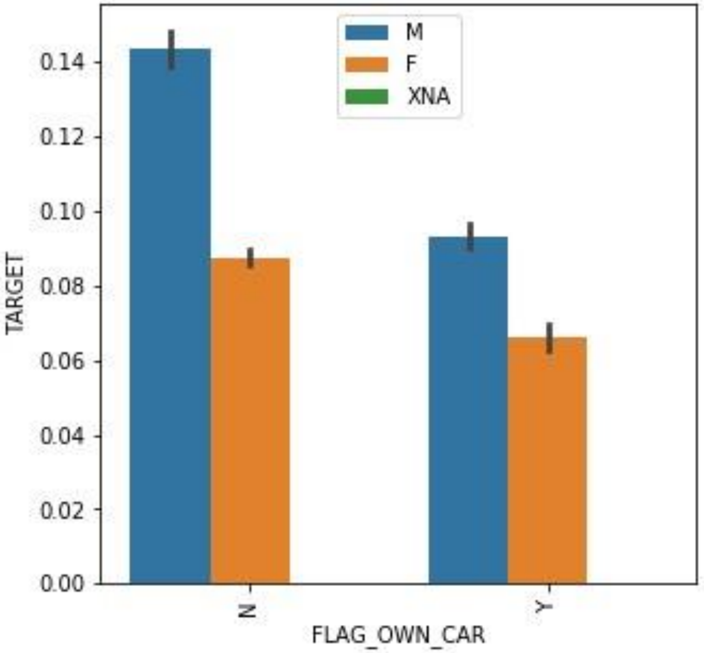
Bivariate Analysis: Categorical variables

- **Impact of Owning car and Gender on TARGET –** Across gender clients owning car are less likely to default.
- **Impact of Housing Type and Gender on TARGET –** Across gender clients living in rented apartment or with parents have greater difficult in payment.
- **Impact of Owning Realty and Gender on TARGET –** Owning realty does not appear to influence default/payment difficult across gender.
- **Impact of Income Type and Gender on TARGET –** Across gender clients who are unemployed or on Maternity leave are more to default.



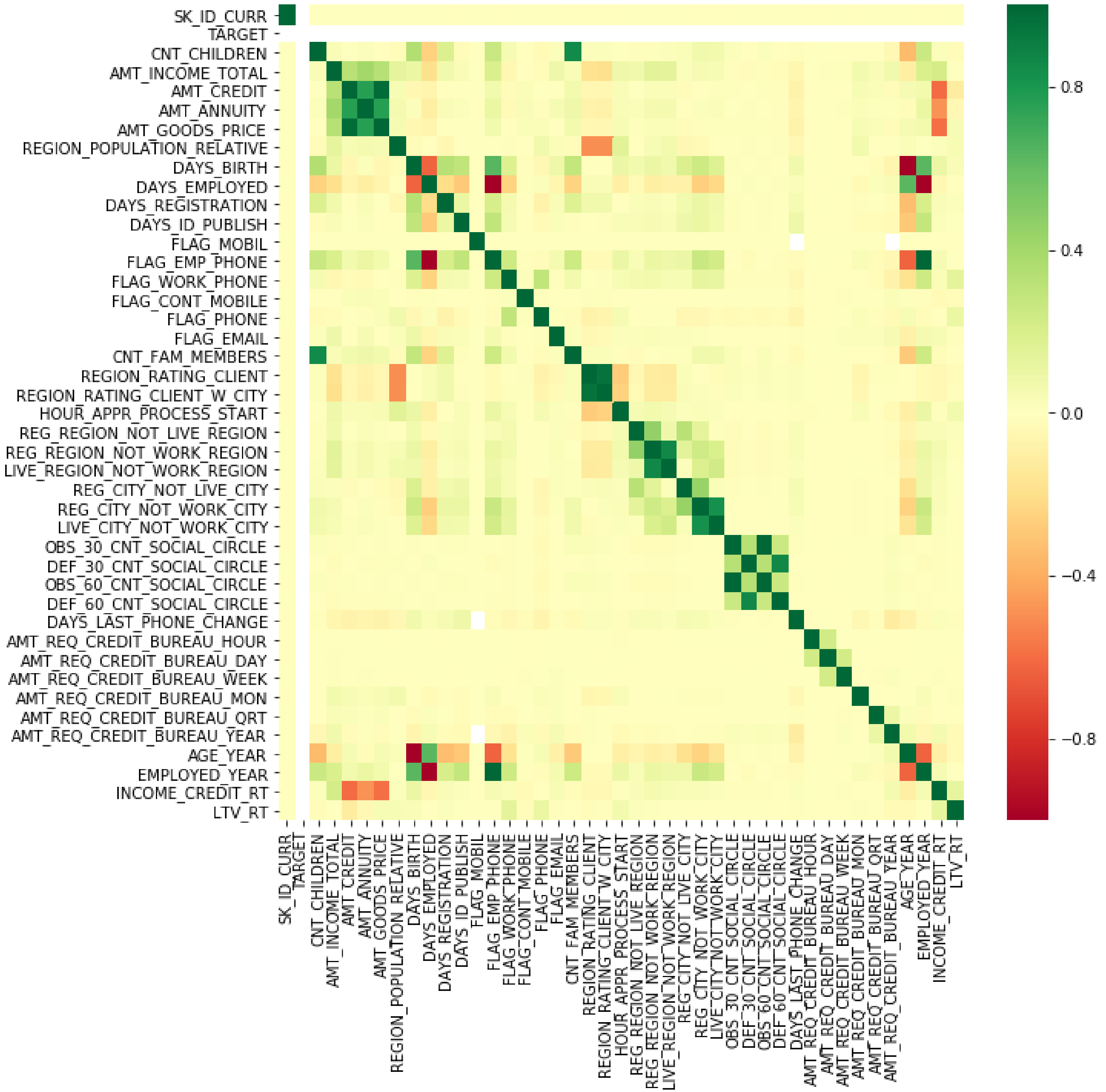






Bivariate Analysis: Correlation (quantitative variables -TARGET=0)

- This heatmap refers to the correlation for key quantitative variables from application_clean with TARGET=0.
- Quantitative variables like AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE are positively correlated with each other.
- Similarly, there is a positive correlation between OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_ CNT_SOCIAL_CIRCLE, OBS_60_ CNT_SOCIAL_CIRCLE and DEF_60_ CNT_SOCIAL_CIRCLE.
- INCOME_CREDIT_RT has a negative correlation with AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE.



Bivariate Analysis: Top Correlation (Clients with No difficulty in payments)

Top Positive correlation for key variables are noted below is observed for below variables:

1. FLAG_EMP_PHONE and EMPLOYED_YEAR (Positive – 0.999788)
2. AMT_CREDIT and AMT_GOODS_PRICE (Positive – 0.982263)
3. CNT_FAM_MEMBERS and CNT_CHILDREN (Positive – 0.85339)
4. AMT_CREDIT and AMT_ANNUITY (Positive – 0.763073)
5. AMT_GOODS_PRICE and AMT_ANNUITY (Positive – 0.762625)
6. FLAG_EMP_PHONE and DAYS_BIRTH (Positive – 0.632847)
7. DAYS_BIRTH and EMPLOYED_YEAR (Positive – 0.628915)

•**There are other strong correlation between below variables:**

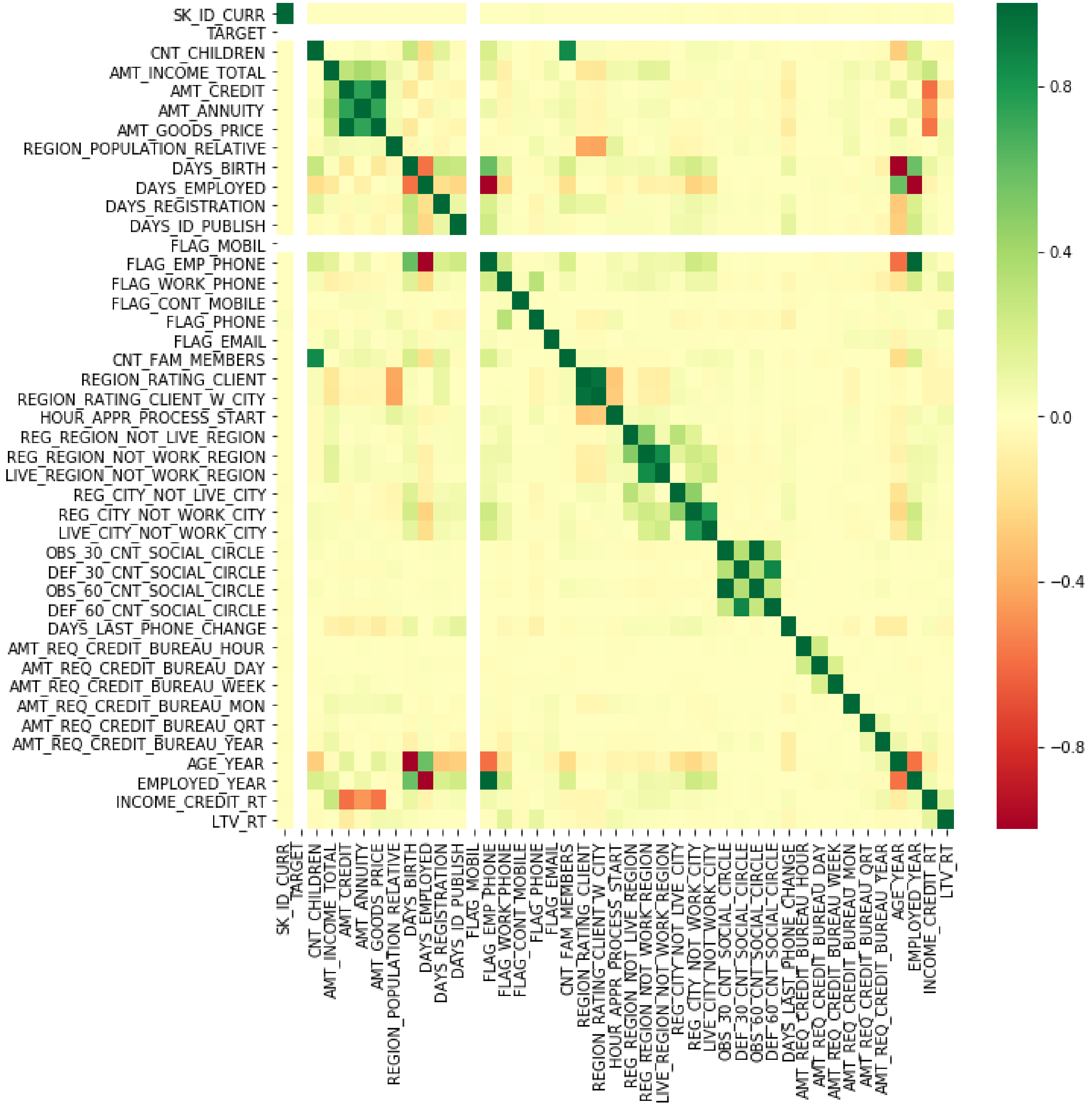
1. OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE (Positive – 0.998513)
2. REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY (Positive – 0.95083)
3. REG_REGION_NOT_WORK_REGION and LIVE_REGION_NOT_WORK_REGION (Positive – 0.860335)
4. REG_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY (Positive – 0.827893)

•**Below are few key negative correlation:**

1. AMT_INCOME_TOTAL and AGE_YEAR (Negative – -0.628839)
2. EMPLOYED_YEAR and AGE_YEAR (Negative – -0.102875)

Bivariate Analysis: Correlation (quantitative variables -TARGET=1)

- This heatmap refers to the correlation for key quantitative variables from application_clean with TARGET=1.
- Like in previous heatmap, Quantitative variables like AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE are also positively correlated with each other.
- Similarly, there is a positive correlation between OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_ CNT_SOCIAL_CIRCLE, OBS_60_ CNT_SOCIAL_CIRCLE and DEF_60_ CNT_SOCIAL_CIRCLE.
- FLAG_EMP_PHONE is negatively correlated with DAYS_EMPLOYED.



Bivariate Analysis: Top Correlation (Clients with No difficulty in payments)

Top Positive correlation for key variables are noted below is observed for below variables:

1. AMT_CREDIT and AMT_GOODS_PRICE (Positive – 0.978397)
2. CNT_FAM_MEMBERS and CNT_CHILDREN (Positive – 0.857971)
3. AMT_CREDIT and AMT_ANNUITY (Positive – 0.743573)
4. AMT_GOODS_PRICE and AMT_ANNUITY (Positive – 0.74107)
5. FLAG_EMP_PHONE and EMPLOYED_YEAR (Positive – 0.586961)
6. DAYS_BIRTH and EMPLOYED_YEAR (Positive – 0.583584)
7. AMT_INCOME_TOTAL and AMT_GOODS_PRICE (Positive – 0.302838)
8. AMT_INCOME_TOTAL and AMT_CREDIT (Positive – 0.302917)

- **There are other strong correlation between below variables:**

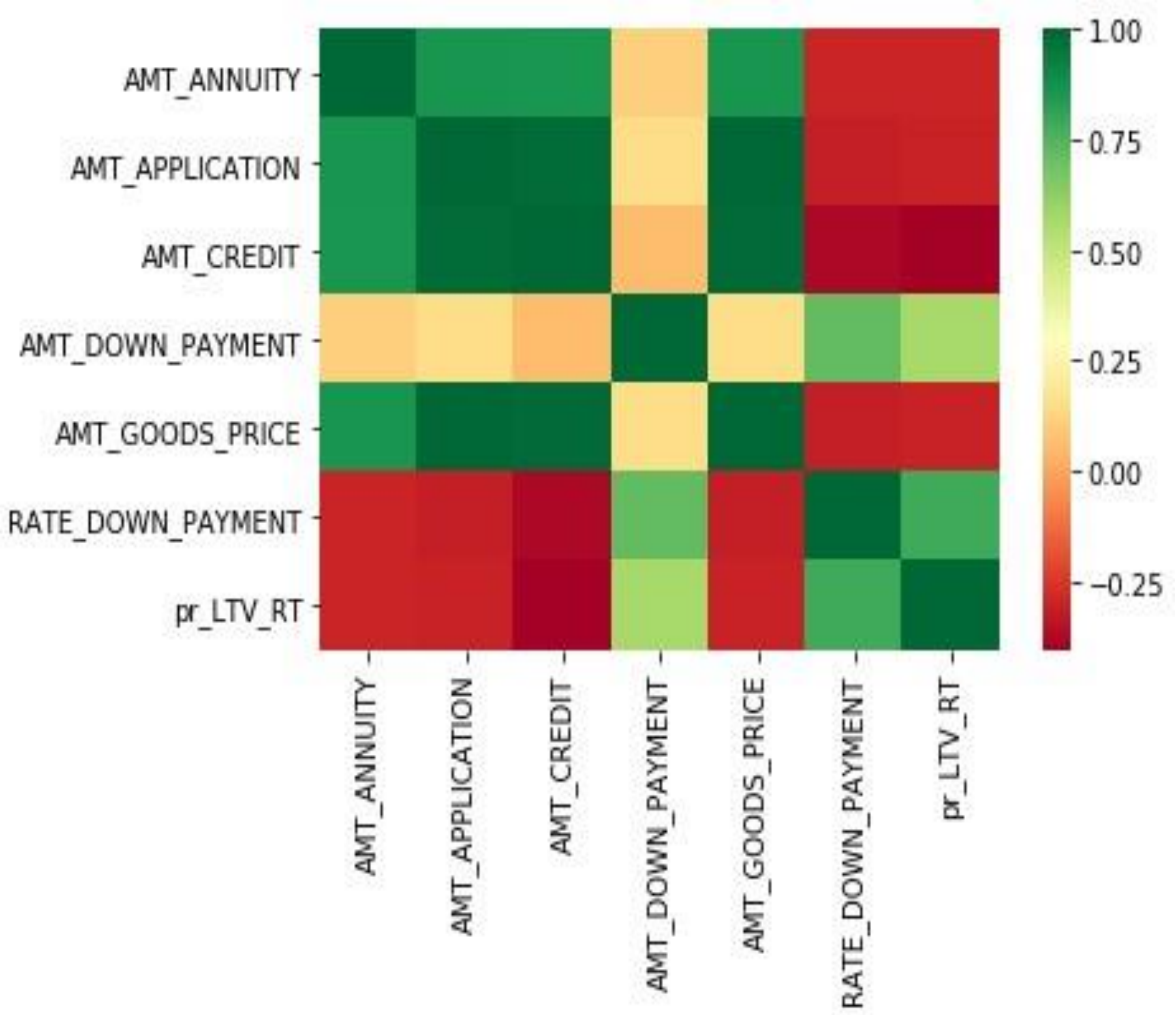
1. OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE (Positive – 0.998224)
2. REG_REGION_NOT_WORK_REGION and LIVE_REGION_NOT_WORK_REGION (Positive – 0.85011)

- **Below are few key negative correlation:**

1. CNT_CHILDREN and AGE_YEAR (Negative – -0.272567)
2. AMT_INCOME_TOTAL and REGION_RATING_CLIENT_W_CITY (Negative – -0.155777)

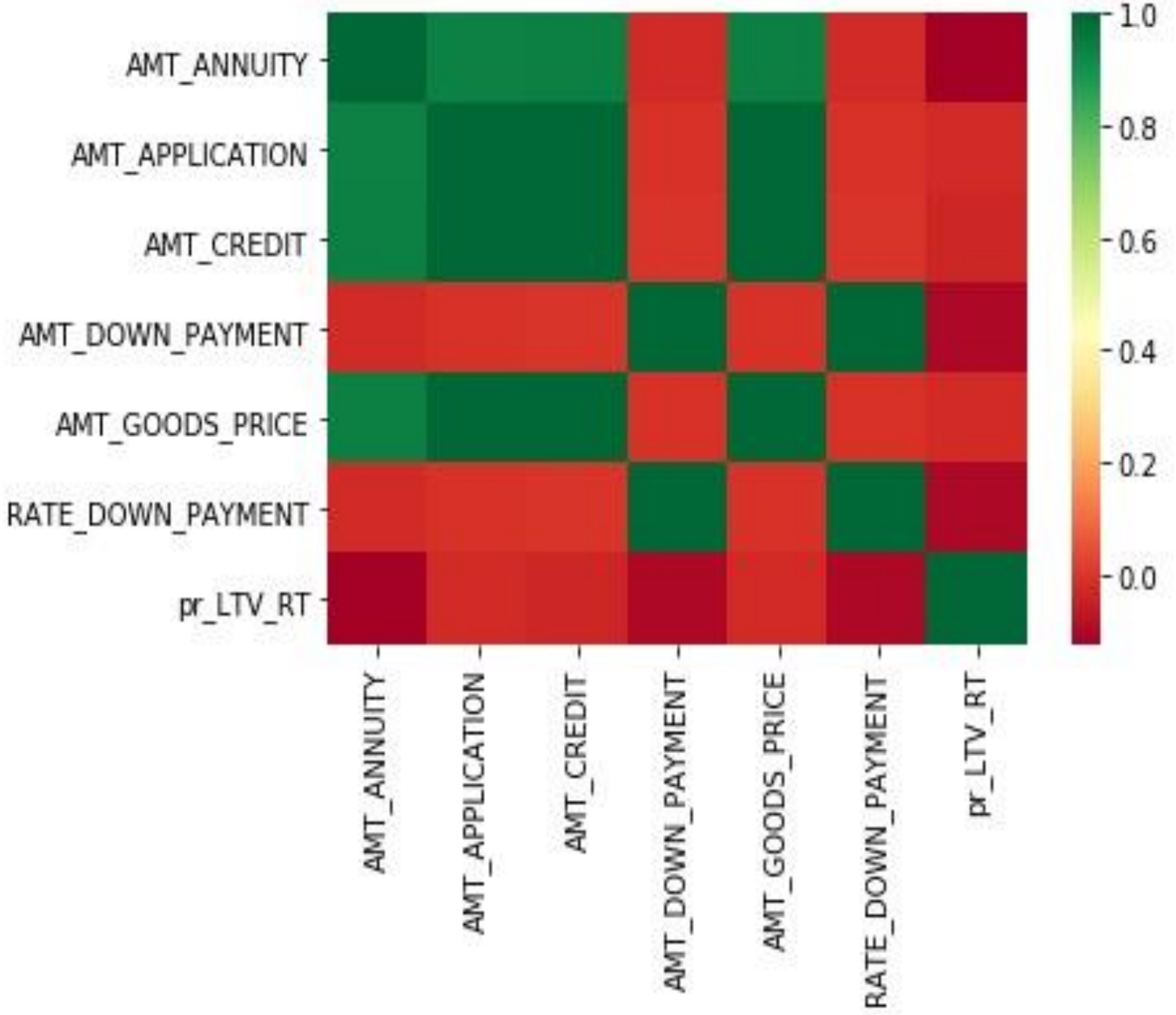
Bivariate Analysis: Correlation (previous_approved)

- The range of color coding used in these correlation plots is Red to Green. Where Red denotes negative correlation and Green denotes positive correlation. Here, the variables belong to 'previous_unused' section of 'previous_app'.
- Followings are some key interpretation of the correlation plot:-
  - Amount annuity is strongly correlated with application amount and amount credit.
  - Amount credit has slightly negative correlation with down payment. Usually client making higher down payment for smaller loans.
  - Goods price is strongly correlated with amount annuity, loan application amount and amount credit. Clients who wanted to purchase high valued goods had applied loan for larger amount.
  - pr_LTV_RT (loan amount / price of goods) has a strong negative correlation with credit amount.
  - Similarly, down-payment rate increases when credit amount decreases and vise-verse.



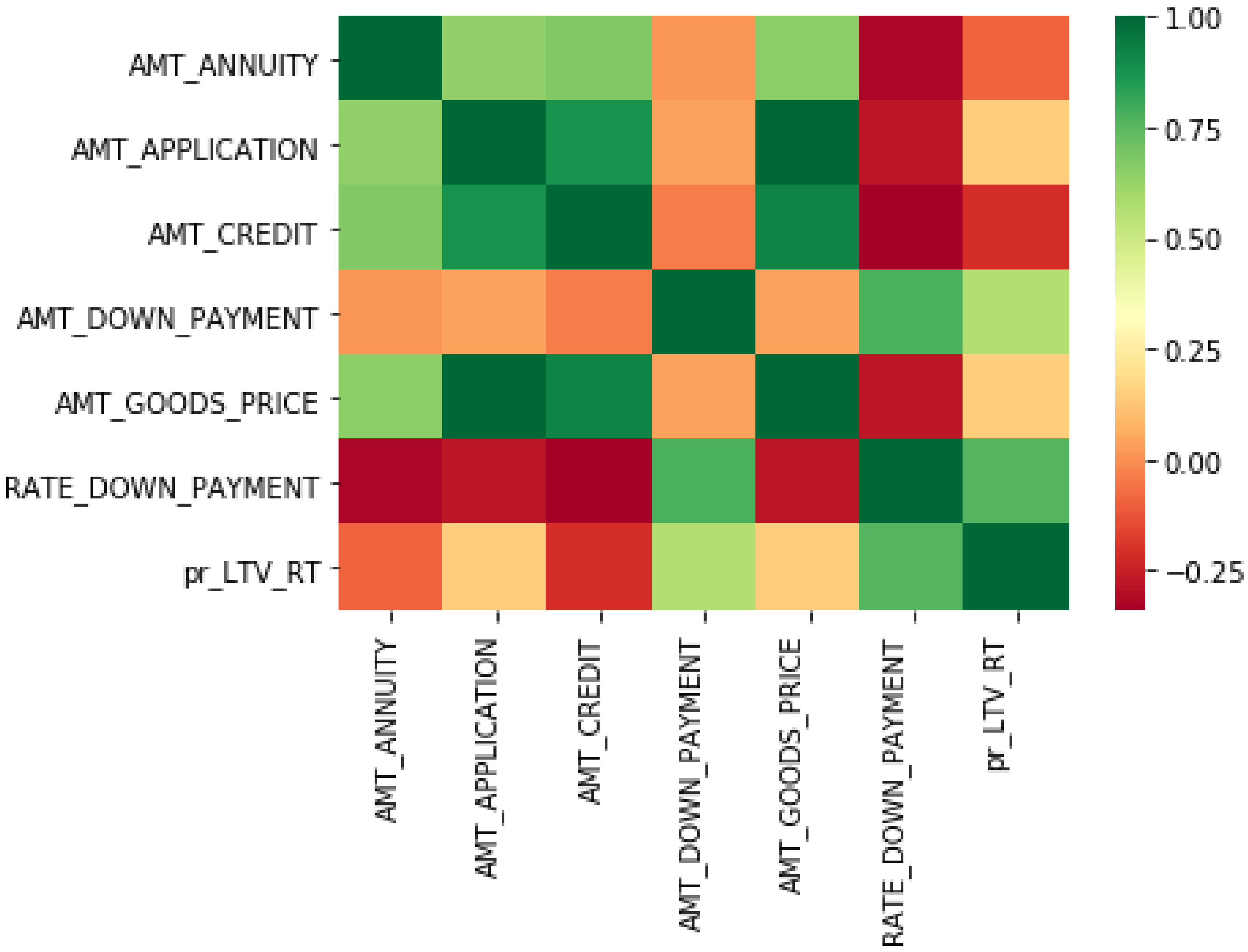
Bivariate Analysis: Correlation (previous_unused)

- This correlation plot helps us to understand the degree of relation between two variables. The scale of correlation ranges from -1 to 1. Where '-1' signifies a strong negative correlation and '+1' signifies a strong positive correlation between two variables. Here, the variables belong to 'previous_approved' section of 'previous_app'
- Followings are some key interpretation of the correlation plot:-
  - Loan annuity increases with increase in the amount of credit and price of goods.
  - The annuity amount decreases when the rate of down-payment increases.
  - There is a negative correlation between the amount asked by the client and the down-payment rate. Which means the client asking for higher loan amount is likely to make a low down-payment.
  - Similarly, price of goods is negatively correlated to the down-payment rate.



Bivariate Analysis: Correlation (previous_refused)

- This bivariate analysis has been performed on ‘previous_refused’ data means those applicants who had been refused to give loan.
- The heatmap shows that there is a strong positive correlation between application amount and goods price.
- Amount annuity and Credit amount is strongly correlated to Goods Price.
- Amount annuity is negatively correlated to down payment.
- Similarly, Credit amount (approved loan amount) is having a strong negative correlation with down payment.



Summary

About Data Distribution:

- Data available for this Exploratory Data Analysis has outlier which were removed for analysis.
- Data imbalance in both the dataset (application and previous application) were noted but were not treated. However care has been taken to ensure that conclusion/result is not because of data imbalance.

Factors which tend to influence greater chances of default:

- Age of employment – Lesser the age of employment greater is the chances of default
- Income Type – Unemployed/Maternity leave have greater chances of default
- Education Type – Lower secondary have greater chances of default (Almost double of Higher Education)
- Occupation Type – Low-skilled labor have greater chances of default
- Housing Type – People stating in Rented Apartment and with parents have greater chances of default.

Factors which tend to influence approval:

- Greater down payment
- Consumer loan has highest approval

Correlation

- Amount annuity is strongly correlated with application amount and amount credit.
- Amount credit has slightly negative correlation with down payment. Usually client making higher down payment request for lesser loan.
- Goods price is strongly correlated with amount annuity, loan application amount and amount credit. Clients who wanted to purchase high valued goods had applied loan for larger amount.
- Loan annuity increases with increase in the amount of credit and price of goods. The annuity amount decreases when the rate of down-payment increases.
- There is a negative correlation between the amount asked by the client and the down-payment rate. Which means the client asking for higher loan amount is likely to make a low down-payment.

Risks:

- Value of “1” in below variables raises concern of fraud since contact address does not fall under same city as either permanent address or work address (In cases, where permanent address, work location address and contact address of a client are distinct then there is a higher probability of defaulting a loan)
  - REG_CITY_NOT_LIVE_CITY
  - REG_CITY_NOT_WORK_CITY
  - LIVE_CITY_NOT_WORK_CITY
- There is always higher risk associated with the clients whose applications has been refused previously.