Credit EDA Assignment

By:- Manish Kumar Lamoria

Introduction

1. Business Understanding

- Companies that provide loan also receives application from people who are requesting for the time, so it’s a tough decision for the companies to provide loans to people with insufficient credit-history. EDA can be used to analyse the patterns to ensure that the applicant can repay the loan.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:
  - ➢ If the applicant is likely to repay the loan, then rejecting the applicant 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.

2. Business Objectives

- The aim here is to identify patterns which indicate if a client has difficulty paying their instalments, after that further steps could be taken like rejecting the applicant, reducing the loan amount etc. This will also ensure that the applicants which can repay the loan are not rejected.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

3. Data Understanding

- *'application_data.csv'* contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties.**
- *'previous_application.csv'* contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
- *'columns_description.csv'* is data dictionary which describes the meaning of the variables.
- *'application_data.csv'* is named as df1 while reading the csv file in python.
- '*Previous_application.csv'* is named as df2 while reading the csv file in python.

4. Approach

1. Reading the csv files :- After carefully reading the meaning of each of the variables present in both the datasets with the help of data dictionary. The datasets are uploaded in jupyter notebook using the pandas library function.

2. Data Cleaning :-

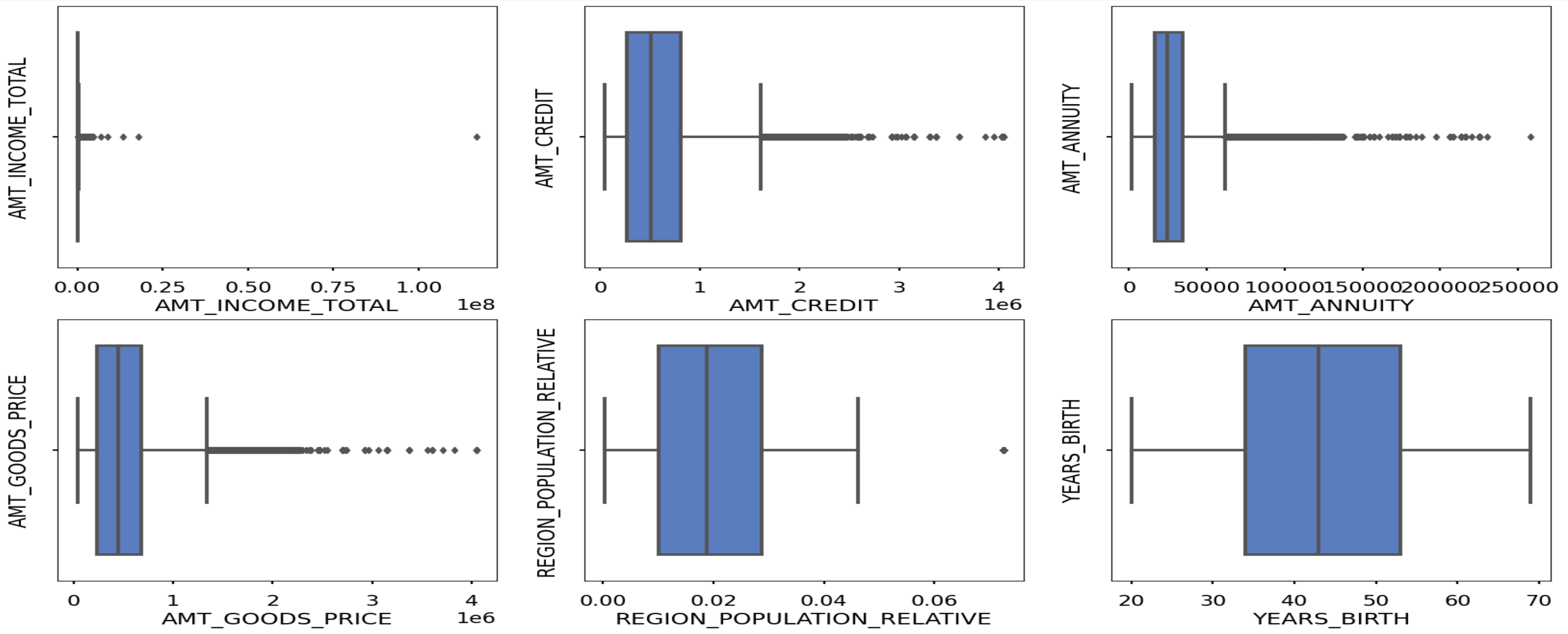
   - Checking the summary and statistical summary of the dataframe(df1) for better understanding the variables.
   - Dropping unnecessary variables from the dataframe.
   - Checking the null values present in the dataframe and removing the variables which consists of null values more than 45%.
   - OCCUPATION_TYPE consists of around 31% null values and 18% of XAN values, but can't impute with any other occupation, it'll affect the analysis of other occupation categories, so leaving it as null.
   - Checking null values again for imputation. Null values in EXT_SOURCE_2 ,EXT_SOURCE_3 , AMT_GOODS_PRICE and AMT_ANNUITY are replaced with median while others are replaced with mode.

3. Data Standardization

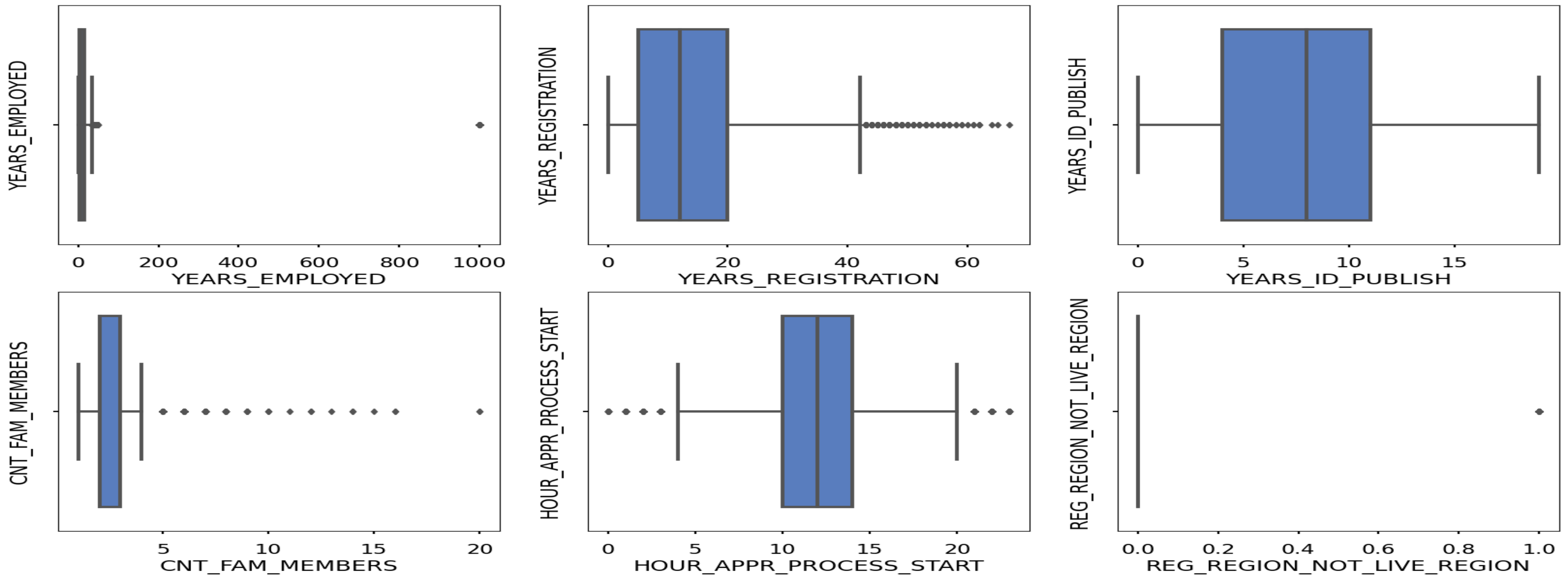
- Stored list of variable names stored in 'c' and value_counts performed using a for loop for detecting the irregularities. Few variables found with negative and XNA values. XNA values present in CODE_GENDER are imputed with mode.
- Negative values present in the variables removed using the abs() function. Variables consisting of days are converted to year and their datatypes are changed to int for ease of analysis and renamed accordingly like DAYS_BIRHT to YEARS_BIRTH.

4. Analysing Numerical columns for outliers i.e univariate numerical analysis

- Numerical variables present in the df1 are analysed by plotting boxplots and using describe() method. And certain variables are found with outliers.



- AMT_INCOME_TOTAL has outliers i.e. exceptionally high income and considerable gap between 95 and 99th quantile. Most of the income is present between 1lakh to 2 lakh.
- Outliers can be observed in AMT_CREDIT , AMT_ANNUITY, AMT_GOODS_PRICE variables.



- YEARS_EMPLOYED and YEARS_REGISTRATION consists of outliers. ex: employed for 1000 years or registered for 67 years.
- CNT_FAM_MEMBERS has outliers. ex:-having more than 5 or 6 children.
- HOUR_APPR_PROCESS_START does not have any outliers. The hour at client applied for loan could vary.

- Approach for handling outliers for the necessary variables.

  1. Outliers present in the AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE can be handled by binning the values.
  2. CNT_FAM_MEMBERS outliers can be handled by capping the values i.e having children less than 4 or 5.

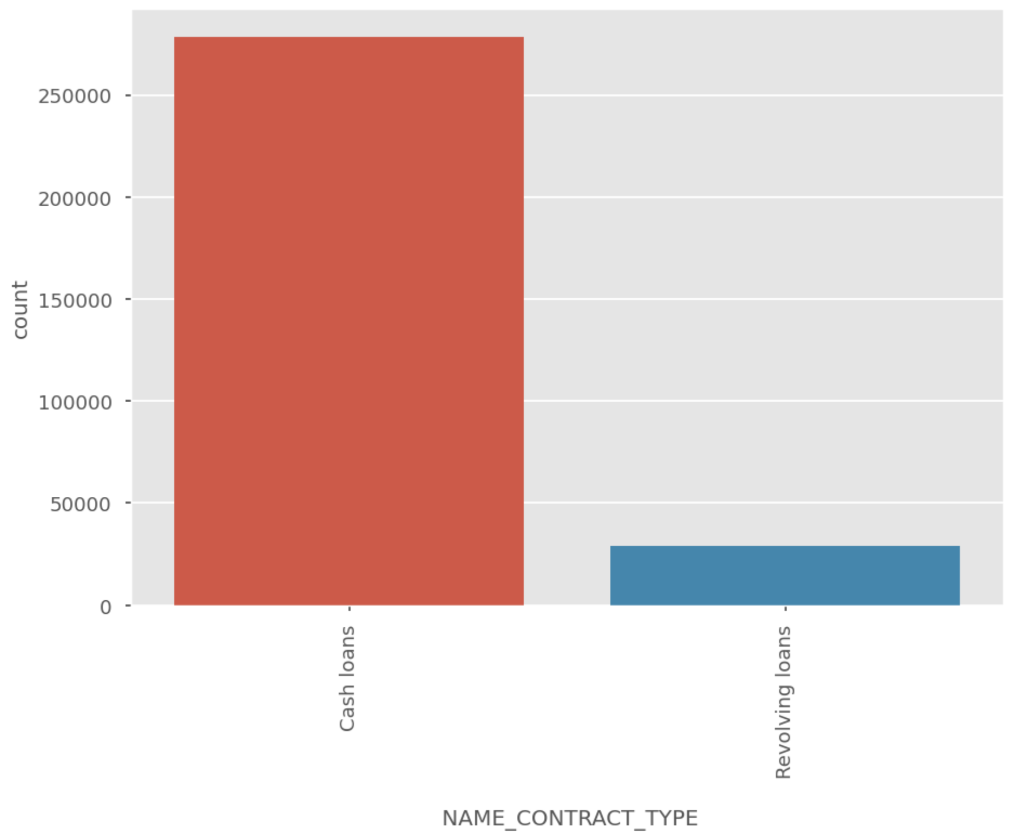
- Binning YEARS_BIRTH,AMT_INCOME_TOTAL,AMOUNT_CREDIT & AMT_GOODS_PRICE using pd.cut and pd.qcut for better analysis.

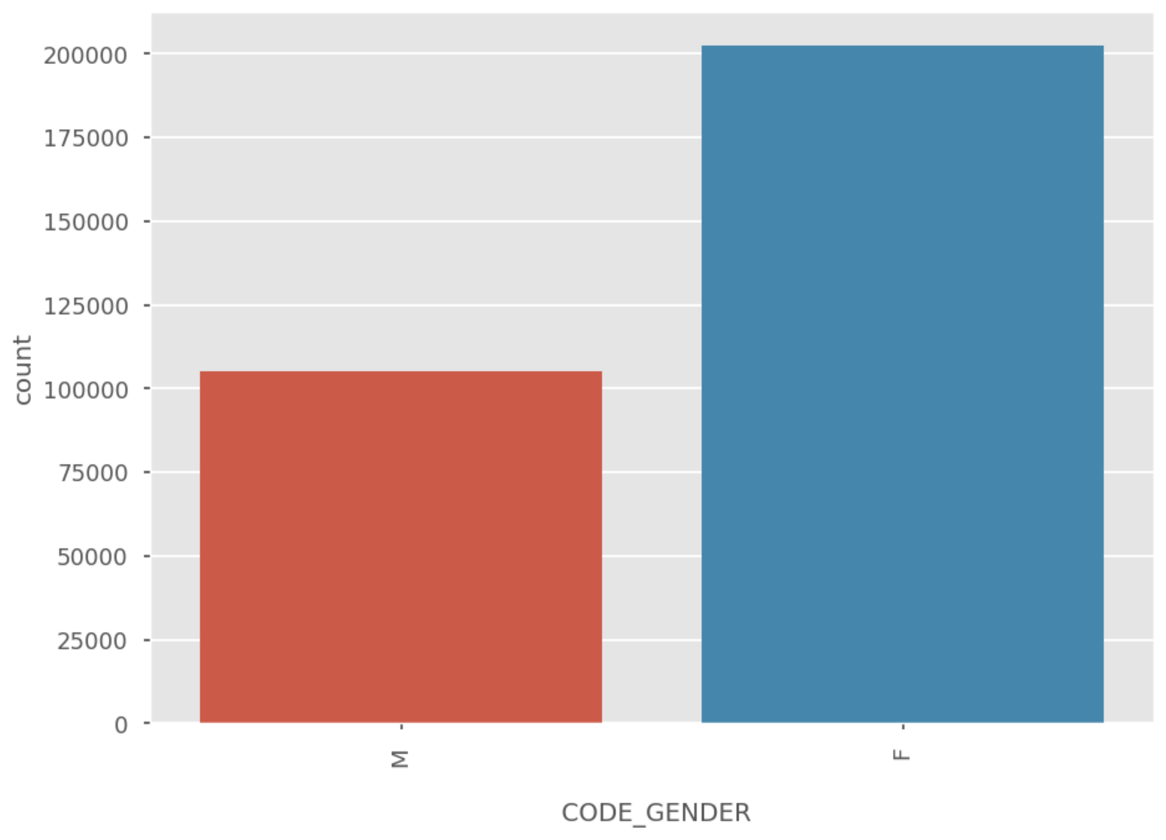
5. Data Imbalance - Analysis of Target Variable(0- No installment payment difficulty and 1- having difficulty with payment)

   - After the analysis of Target variable the ratio is 1:11. i.e. 1 in every 11 person has payment difficulty.

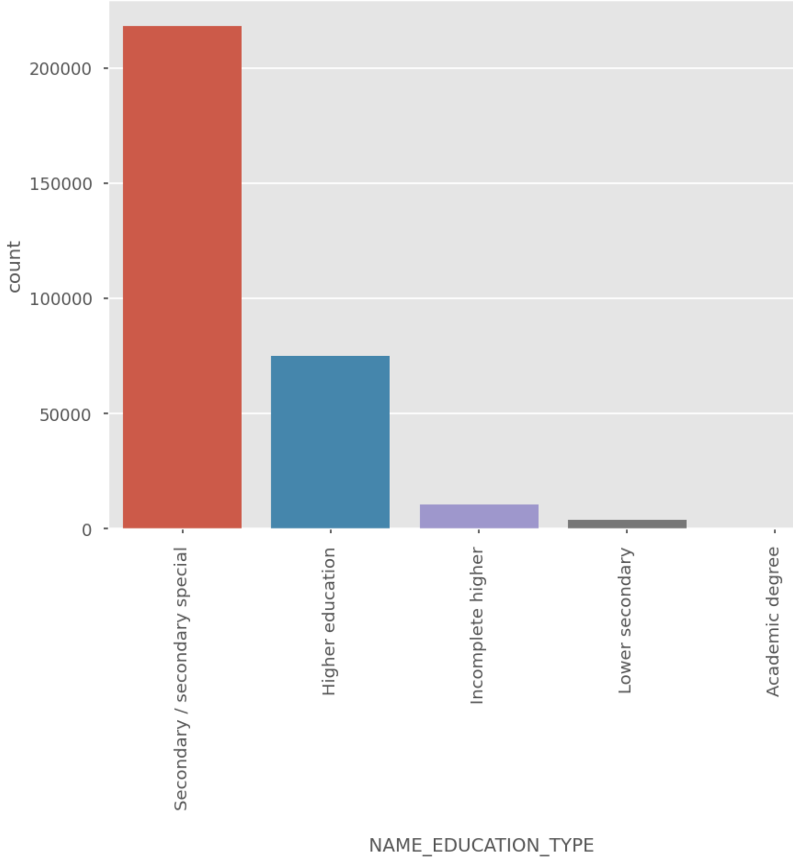
6. Univariate Analysis

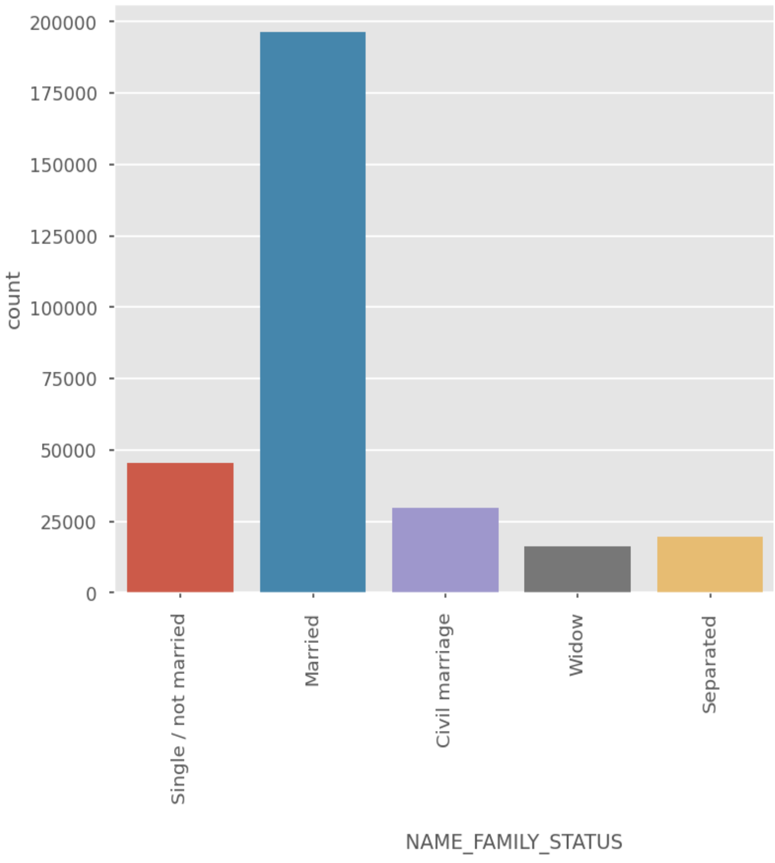
   - Plotted countplot for categorical variables.

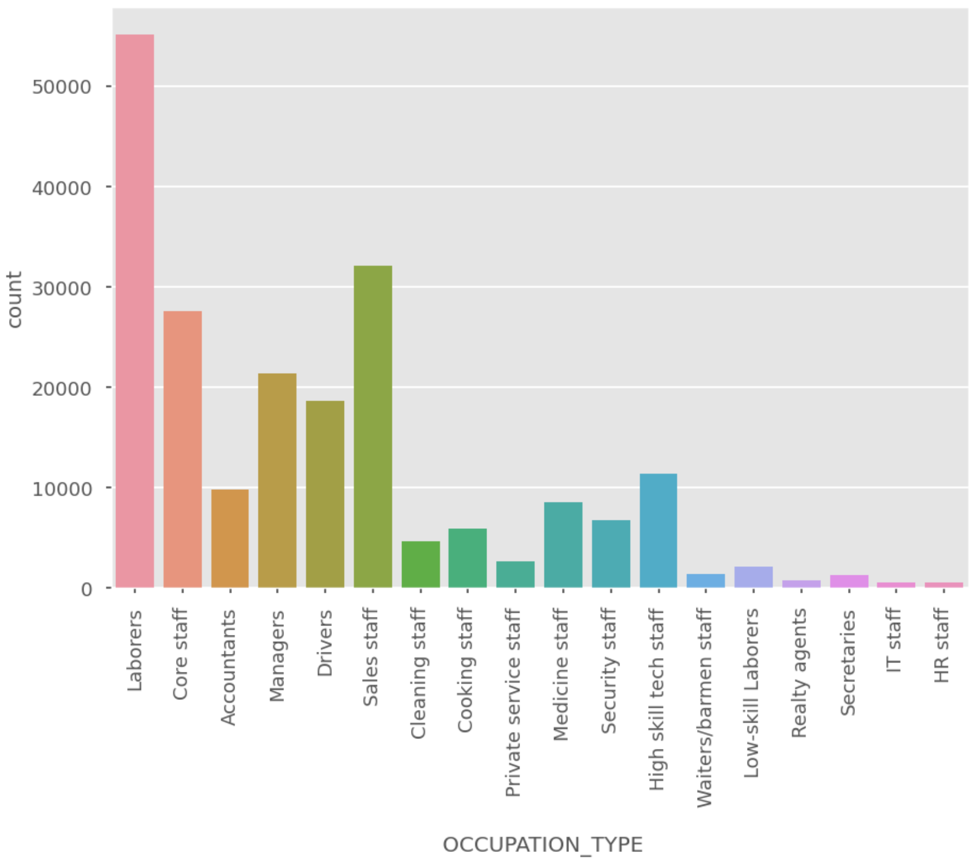




- NAME_CONTRACT_TYPE - Majority of the loans are cash loans.
- CODE_GENDER – No. of Female clients who applied for loan is much higher than Male clients.



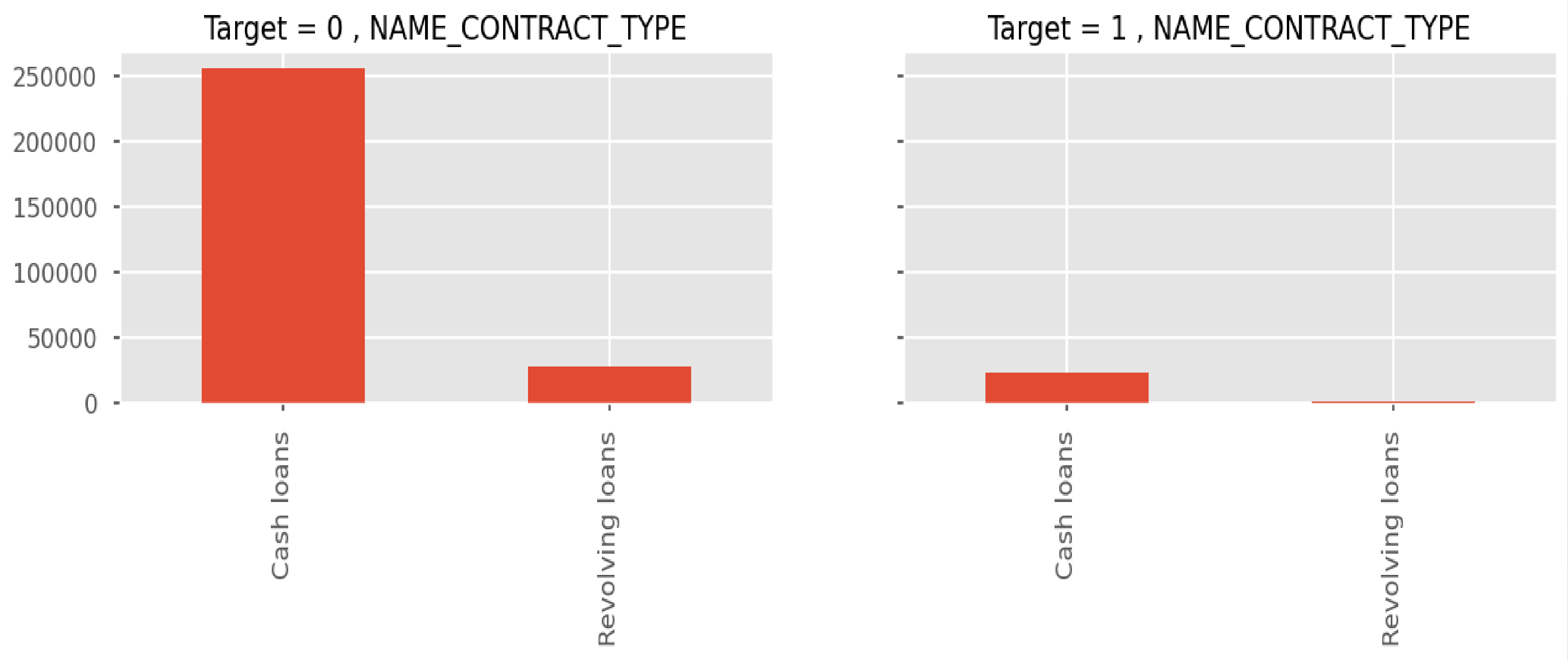




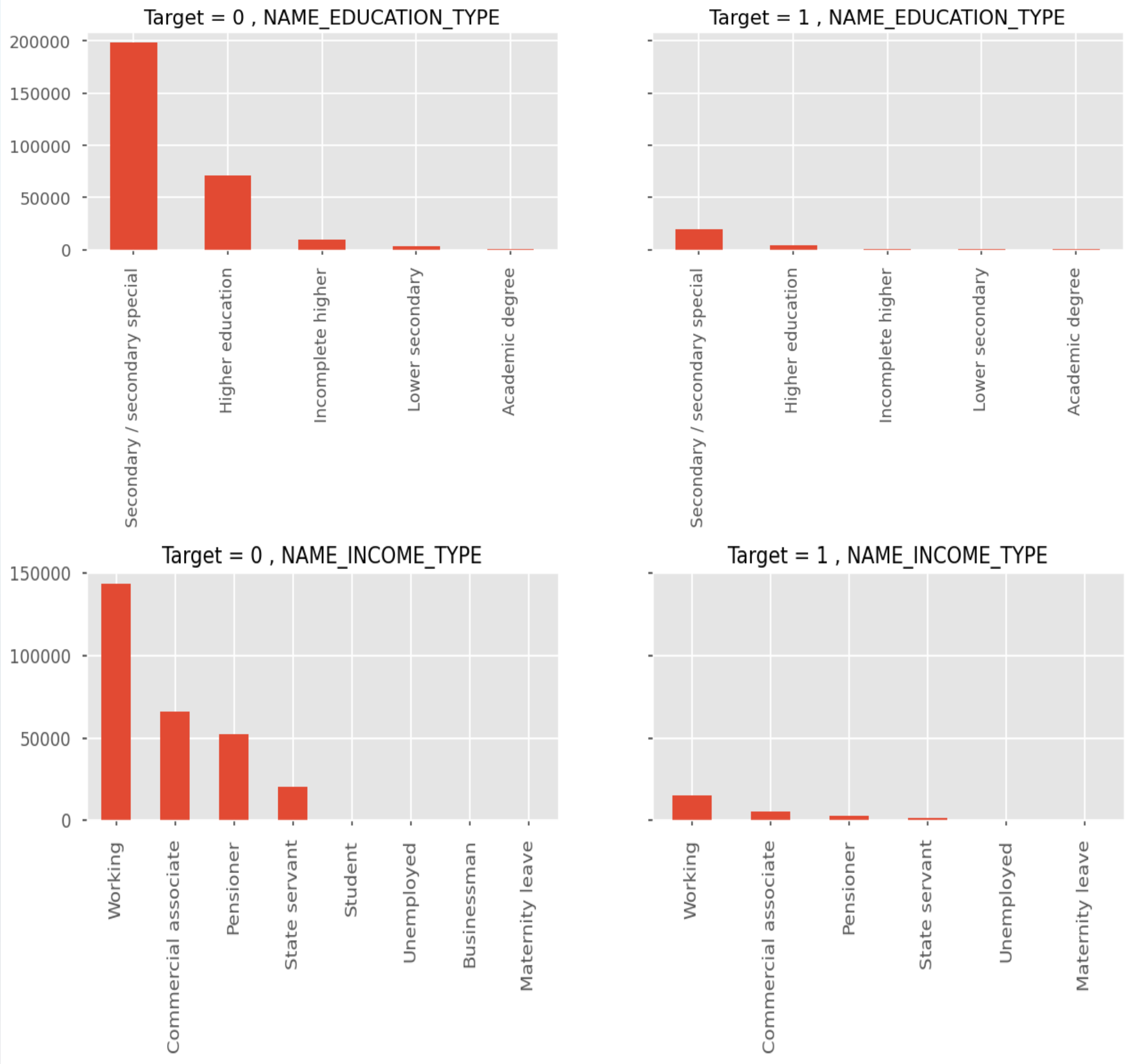
- NAME_EDUCATION_TYPE - Significant no of client have Secondary Education who applied for loan.
- NAME_FAMILY_STATUS - Most of the applicants are married.
- OCCUPATION_TYPE - Majority of the loan applicants have 'Laborers' as occupation, followed by 'sales staff’ and ‘core staff’.

7. Segmented univariate Analysis

- For segmented analysis the dataframe df1 is divided into 2 dataframes df1_T0, df1_T1 based on target value i.e. 0 and 1, 0 being no payment difficulty and 1 being difficulty in payment.
- Bar graph is plotted for comparison between both the target cases.

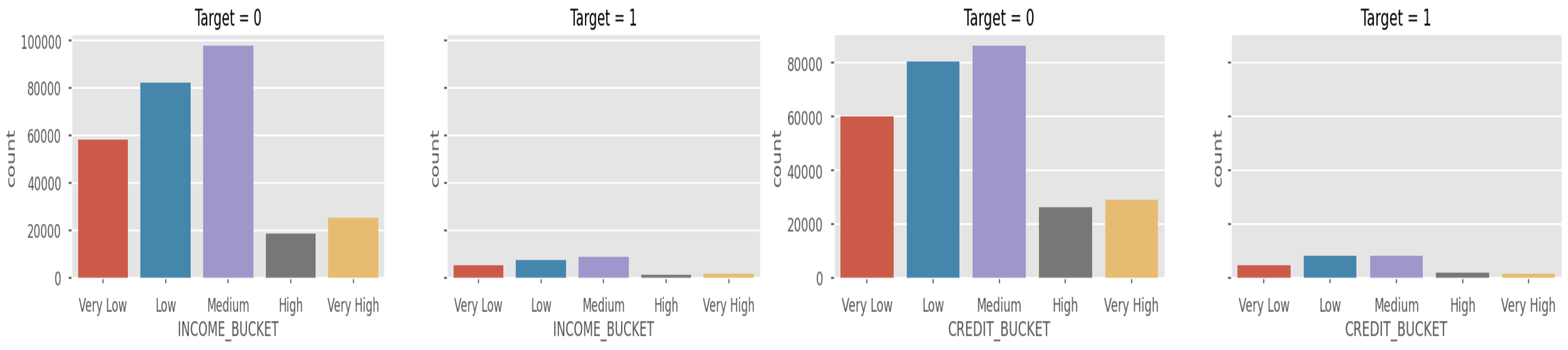


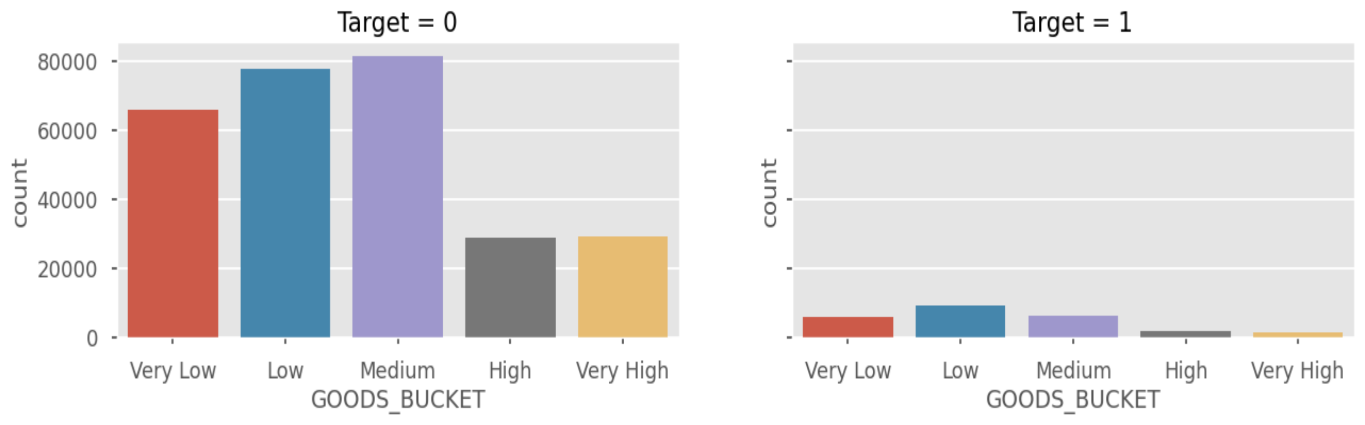
- The No. of people facing repayment difficulties with cash loan is low. Revolving loan is in small number but majority of clients having revolving loan face no difficulty in repayment.



- Among the education type category people with secondary education face no difficulty in repayment but some people from the same category face difficulty in repayment.

- Working people have least difficulty in payments followed by Commercial associates and pensioners.

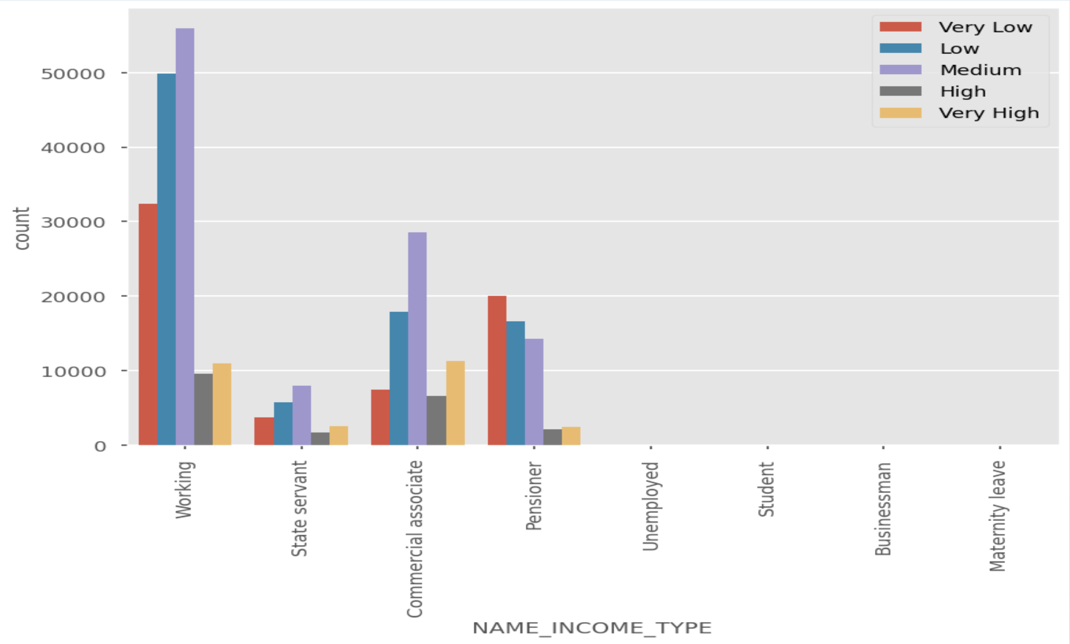


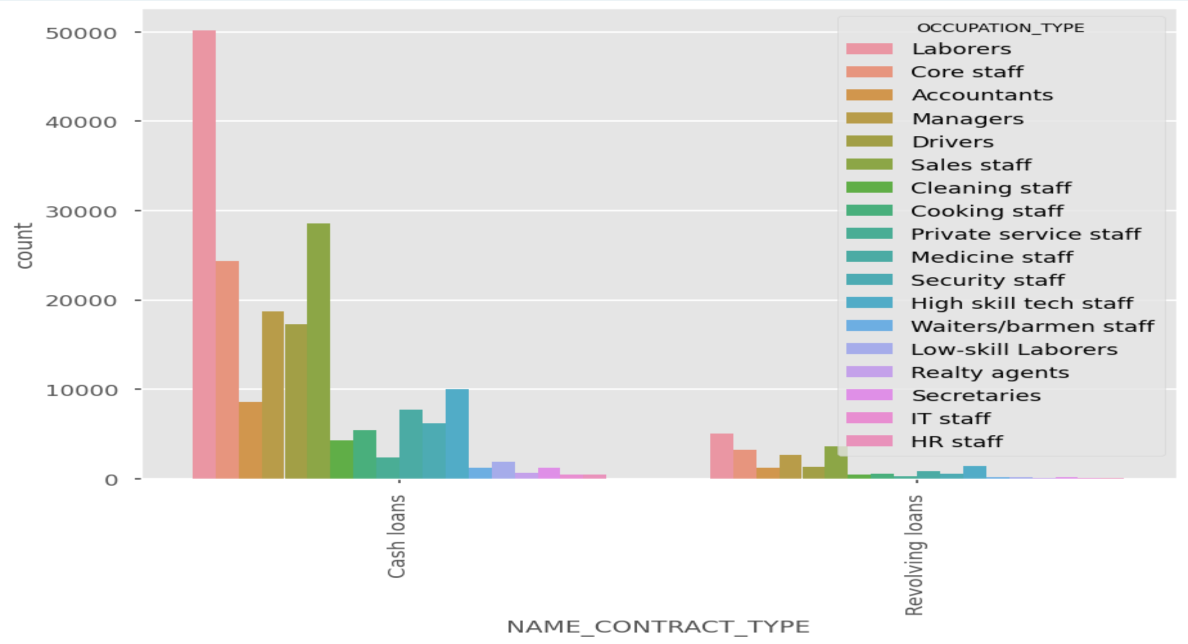


- Understandably, the INCOME, CREDIT and GOODS price is low for people having payment difficulties as compared to people having no payment difficulties

8. Bivariate Analysis

- Categorical – Categorical analysis

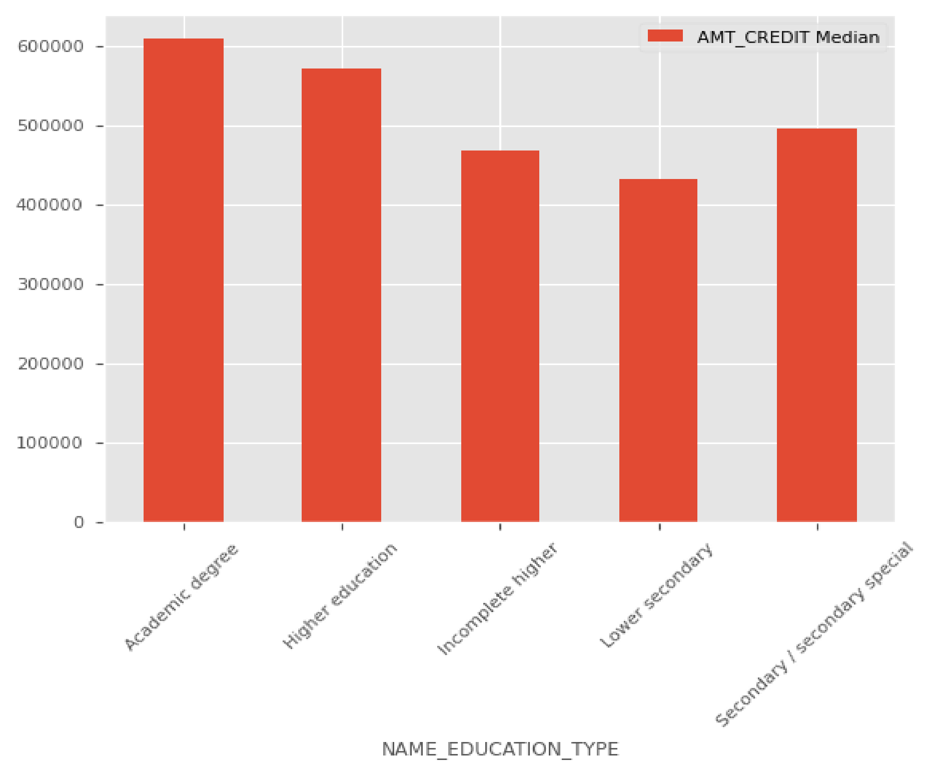


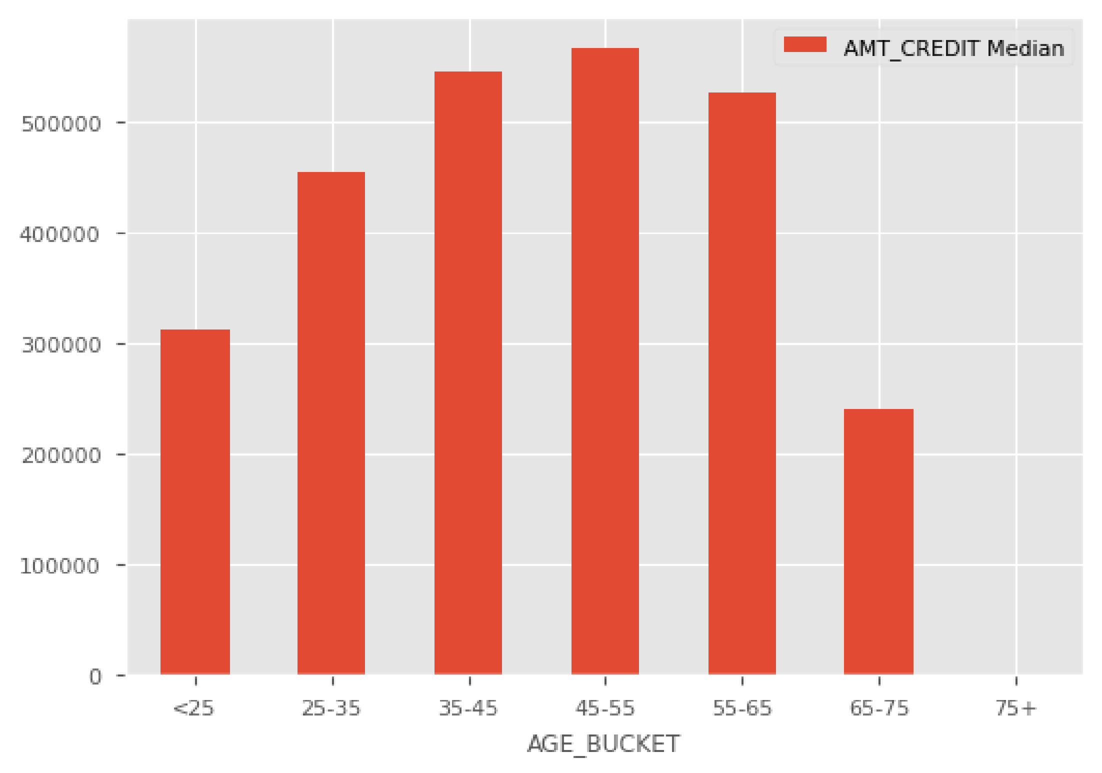


- Working class has salary varied between very-low , low and medium. most of the people from working class are in medium and low salary bracket. similar to commercial associate

- Cash loan are highest for laborers and sales staff followed by core staff.
- No. of Revolving loan is low as compared to cash loans but it is highest for laborers and sales staff.

- Numerical - Categorical Analysis

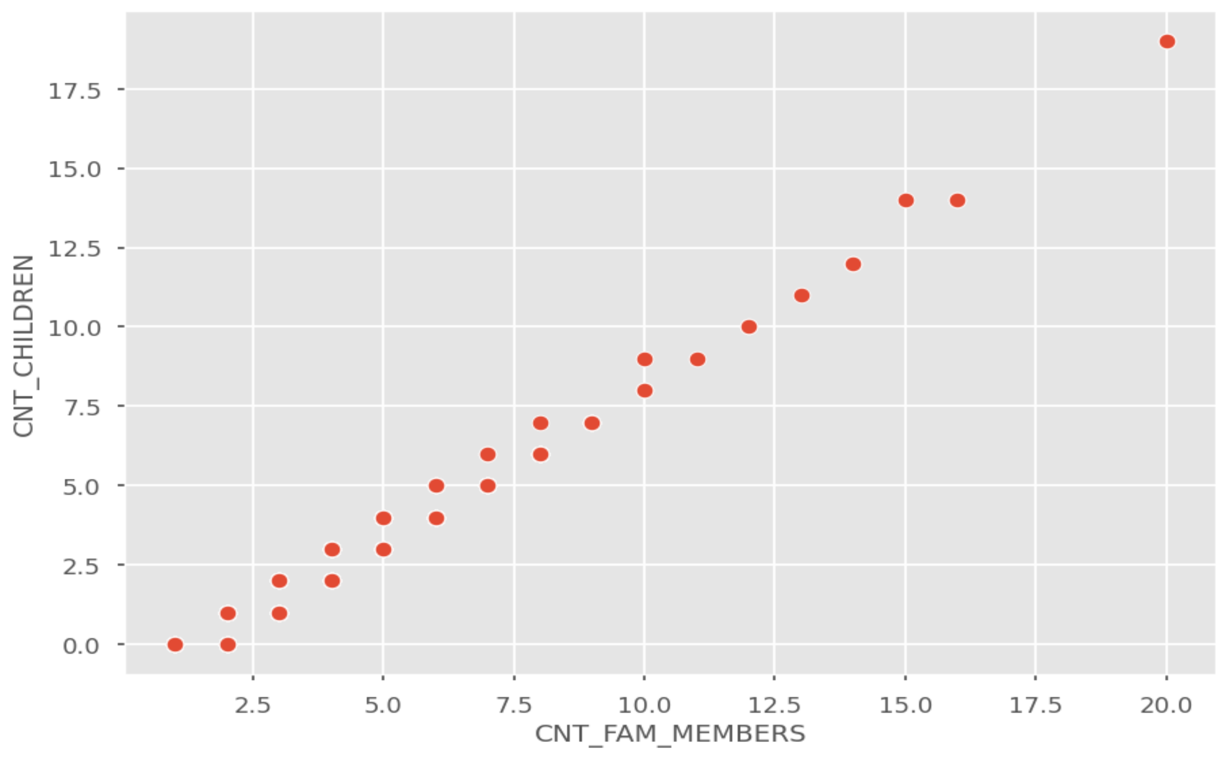


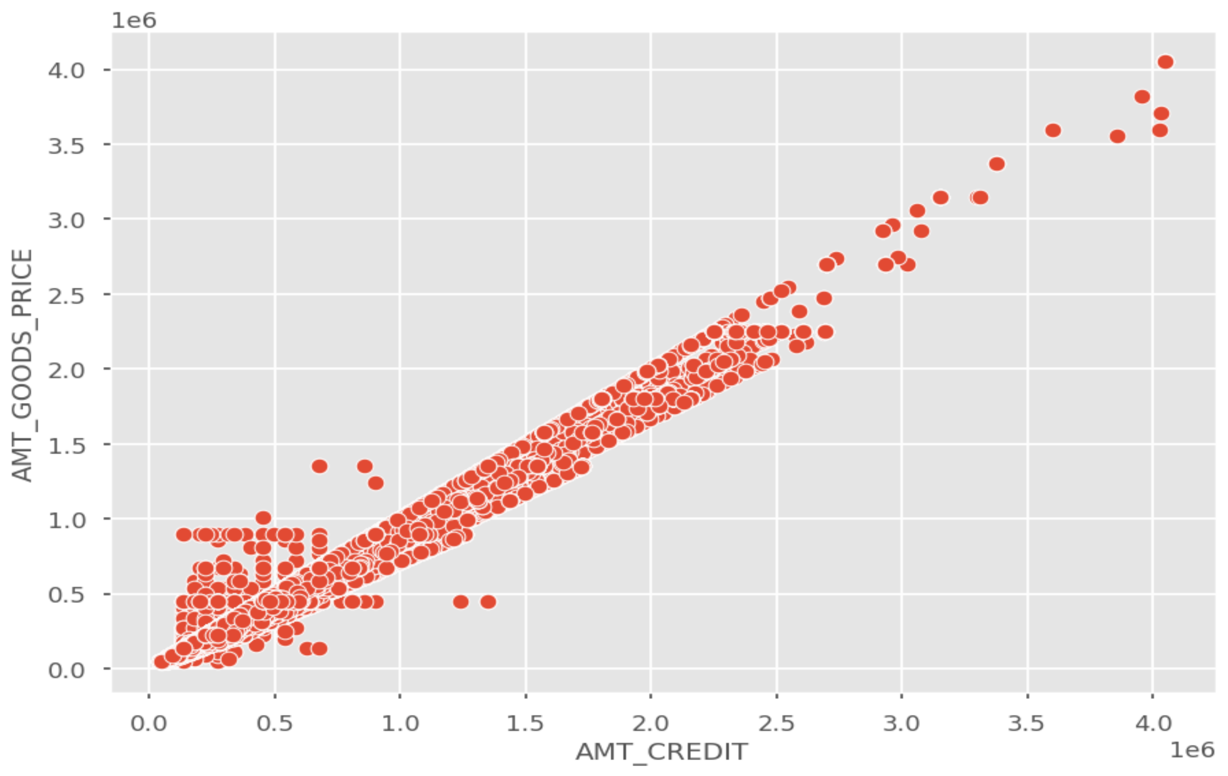


- Credit amount is highest for the people with academic degree followed by higher education.

- The credit is highest in the age group 45-55 followed by 35-45.

- Numerical – Numerical Analysis





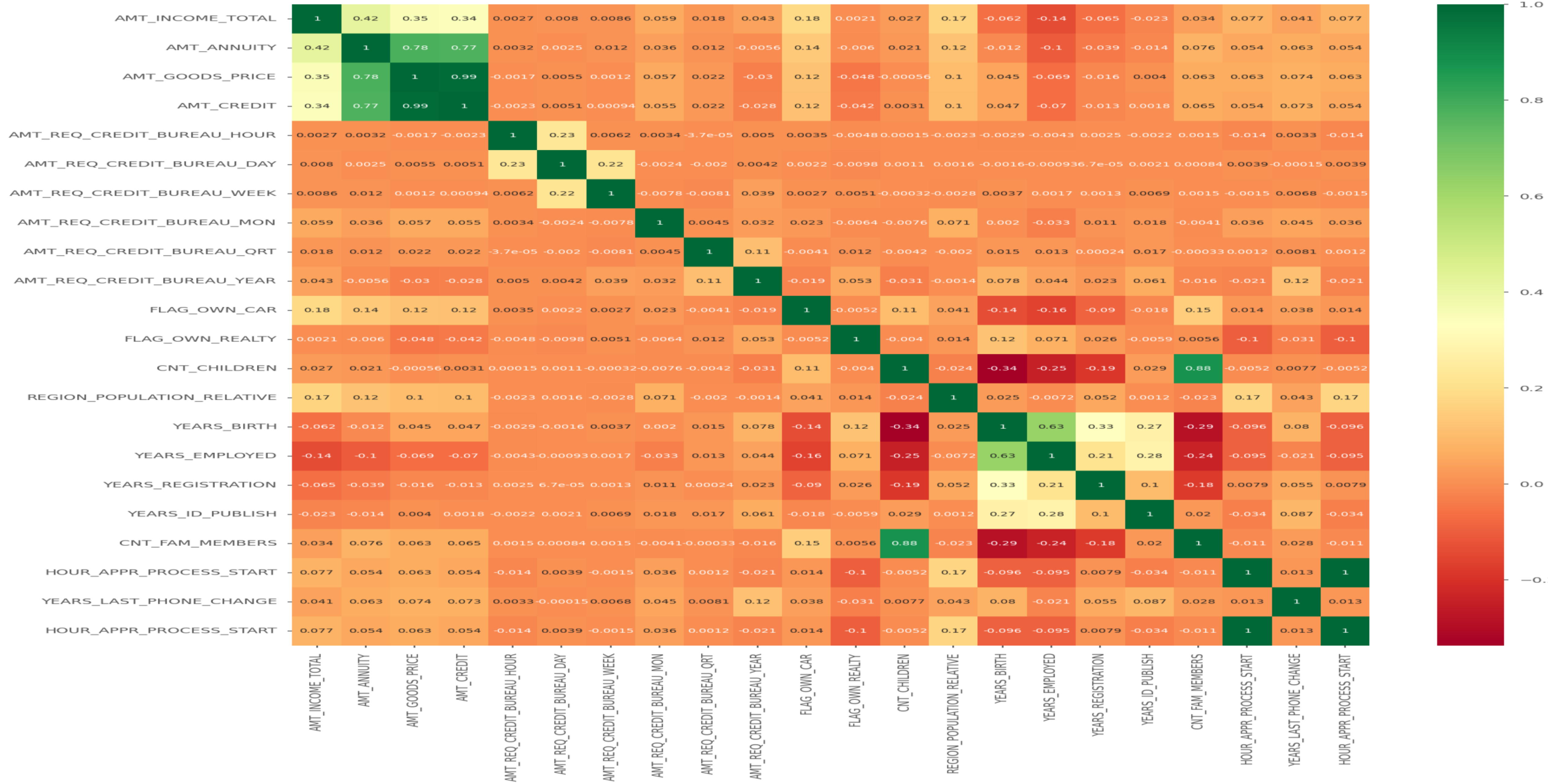
- The no. of family members increases proportionally with increase in no. of children.

- Credit amount increases with increase in Goods price.

9. Correlation

- For top correlation, heatmap is plotted for the numeric variables(dataframe =df1_T0 i.e. Target =0) .
- Top 10 Correlation :-

1. CNT_FAM_MEMBERS and CNT_CHILDREN = 0.88
2. YEARS_EMPLOYED and YEARS_BIRTH = 0.63
3. AMT_ANNUITY and AMT_INCOME_TOTAL = 0.42
4. AMT_GOODS_PRICE and AMT_INCOME_TOTAL = 0.35
5. AMT_CREDIT and AMT_INCOME_TOTAL = 0.34
6. YEARS_EMPLOYED and YEARS_ID_PUBLISH = 0.28
7. YEARS_BIRTH and YEARS_ID_PUBLISH = 0.27
8. YEARS_REGISTRAION and YEARS_BIRTH = 0.33
9. AMT_INCOME_TOTAL and REGION_POPULATION_RELATIVE = 0.17
10. AMT_ANNUITY and FLAG_OWN_CAR = 0.14



	AMT_INCOME_TOTAL	AMT_ANNUITY	AMT_GOODS_PRICE	AMT_CREDIT	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_ID_PUBLISH	CNT_FAM_MEMBERS	HOUR_APPR_PROCESS_START	YEARS_LAST_PHONE_CHANGE	HOUR_APPR_PROCESS_START
AMT_INCOME_TOTAL	1	0.42	0.35	0.34	0.0027	0.008	0.0086	0.059	0.018	0.043	0.18	0.0021	0.027	0.17	-0.062	-0.14	-0.065	-0.023	0.034	0.077	0.041	0.077
AMT_ANNUITY	0.42	1	0.78	0.77	0.0032	0.0025	0.012	0.036	0.012	-0.0056	0.14	-0.006	0.021	0.12	-0.012	-0.1	-0.039	-0.014	0.076	0.054	0.063	0.054
AMT_GOODS_PRICE	0.35	0.78	1	0.99	-0.0017	0.0055	0.0012	0.057	0.022	-0.03	0.12	-0.048	-0.00056	0.1	0.045	-0.069	-0.016	0.004	0.063	0.063	0.074	0.063
AMT_CREDIT	0.34	0.77	0.99	1	-0.0023	0.0051	0.00094	0.055	0.022	-0.028	0.12	-0.042	0.0031	0.1	0.047	-0.07	-0.013	0.0018	0.065	0.054	0.073	0.054
AMT_REQ_CREDIT_BUREAU_HOUR	0.0027	0.0032	-0.0017	-0.0023	1	0.23	0.0062	0.0034	-3.7e-05	0.005	0.0035	-0.0048	0.00015	-0.0023	-0.0029	-0.0043	0.0025	-0.0022	0.0015	-0.014	0.0033	-0.014
AMT_REQ_CREDIT_BUREAU_DAY	0.008	0.0025	0.0055	0.0051	0.23	1	0.22	-0.0024	-0.002	0.0042	0.0022	-0.0098	0.0011	0.0016	-0.0016	-0.00093	6.7e-05	0.0021	0.00084	0.0039	-0.00015	0.0039
AMT_REQ_CREDIT_BUREAU_WEEK	0.0086	0.012	0.0012	0.00094	0.0062	0.22	1	-0.0078	-0.0081	0.039	0.0027	0.0051	-0.00032	-0.0028	0.0037	0.0017	0.0013	0.0069	0.0015	-0.0015	0.0068	-0.0015
AMT_REQ_CREDIT_BUREAU_MON	0.059	0.036	0.057	0.055	0.0034	-0.0024	-0.0078	1	0.0045	0.032	0.023	-0.0064	-0.0076	0.071	0.002	-0.033	0.011	0.018	-0.0041	0.036	0.045	0.036
AMT_REQ_CREDIT_BUREAU_QRT	0.018	0.012	0.022	0.022	-3.7e-05	-0.002	-0.0081	0.0045	1	0.11	-0.0041	0.012	-0.0042	-0.002	0.015	0.013	0.00024	0.017	-0.00033	0.0012	0.0081	0.0012
AMT_REQ_CREDIT_BUREAU_YEAR	0.043	-0.0056	-0.03	-0.028	0.005	0.0042	0.039	0.032	0.11	1	-0.019	0.053	-0.031	-0.0014	0.078	0.044	0.023	0.061	-0.016	-0.021	0.12	-0.021
FLAG_OWN_CAR	0.18	0.14	0.12	0.12	0.0035	0.0022	0.0027	0.023	-0.0041	-0.019	1	-0.0052	0.11	0.041	-0.14	-0.16	-0.09	-0.018	0.15	0.014	0.038	0.014
FLAG_OWN_REALTY	0.0021	-0.006	-0.048	-0.042	-0.0048	-0.0098	0.0051	-0.0064	0.012	0.053	-0.0052	1	-0.004	0.014	0.12	0.071	0.026	-0.0059	0.0056	-0.1	-0.031	-0.1
CNT_CHILDREN	0.027	0.021	-0.00056	0.0031	0.00015	0.0011	-0.00032	-0.0076	-0.0042	-0.031	0.11	-0.004	1	-0.024	-0.34	-0.25	-0.19	0.029	0.88	-0.0052	0.0077	-0.0052
REGION_POPULATION_RELATIVE	0.17	0.12	0.1	0.1	-0.0023	0.0016	-0.0028	0.071	-0.002	-0.0014	0.041	0.014	-0.024	1	0.025	-0.0072	0.052	0.0012	-0.023	0.17	0.043	0.17
YEARS_BIRTH	-0.062	-0.012	0.045	0.047	-0.0029	-0.0016	0.0037	0.002	0.015	0.078	-0.14	0.12	-0.34	0.025	1	0.63	0.33	0.27	-0.29	-0.096	0.08	-0.096
YEARS_EMPLOYED	-0.14	-0.1	-0.069	-0.07	-0.0043	-0.00093	0.0017	-0.033	0.013	0.044	-0.16	0.071	-0.25	-0.0072	0.63	1	0.21	0.28	-0.24	-0.095	-0.021	-0.095
YEARS_REGISTRATION	-0.065	-0.039	-0.016	-0.013	0.0025	6.7e-05	0.0013	0.011	0.00024	0.023	-0.09	0.026	-0.19	0.052	0.33	0.21	1	0.1	-0.18	0.0079	0.055	0.0079
YEARS_ID_PUBLISH	-0.023	-0.014	0.004	0.0018	-0.0022	0.0021	0.0069	0.018	0.017	0.061	-0.018	-0.0059	0.029	0.0012	0.27	0.28	0.1	1	0.02	-0.034	0.087	-0.034
CNT_FAM_MEMBERS	0.034	0.076	0.063	0.065	0.0015	0.00084	0.0015	-0.0041	-0.00033	-0.016	0.15	0.0056	0.88	-0.023	-0.29	-0.24	-0.18	0.02	1	-0.011	0.028	-0.011
HOUR_APPR_PROCESS_START	0.077	0.054	0.063	0.054	-0.014	0.0039	-0.0015	0.036	0.0012	-0.021	0.014	-0.1	-0.0052	0.17	-0.096	-0.095	0.0079	-0.034	-0.011	1	0.013	1
YEARS_LAST_PHONE_CHANGE	0.041	0.063	0.074	0.073	0.0033	-0.00015	0.0068	0.045	0.0081	0.12	0.038	-0.031	0.0077	0.043	0.08	-0.021	0.055	0.087	0.028	0.013	1	0.013
HOUR_APPR_PROCESS_START	0.077	0.054	0.063	0.054	-0.014	0.0039	-0.0015	0.036	0.0012	-0.021	0.014	-0.1	-0.0052	0.17	-0.096	-0.095	0.0079	-0.034	-0.011	1	0.013	1

10. Data cleaning (Previous application.csv)

- Checking the summary and statistical summary of the dataframe(df2) for better understanding the variables.
- Checking the null values present in the dataframe df2 and removing the variables which consists of null values more than 40%.

11. Imputing Values

- Imputing values in AMT_GOODS_PRICE,AMT_ANNUITY,CNT_PAYMENT with median
- Imputing values in PRODUCT_COMBINATION with mode.

12. Data Standardization

- Stored list of variable names in ‘b' and value_counts performed using a for loop for detecting the irregularities. 10 variables found with XAP and XNA values. Variables having majority of XAP and XNA values are dropped.
- Negative values present in DAYS Decision and SELLERPLACE_AREA are removed using the abs() function. Variables consisting of days are converted to year and their datatypes are changed to int for ease of analysis and renamed accordingly like DAYS_DECISION to YEARS_DECISION.
- FLAG_LAST_APPL_PER_CONTRACT values in flag variable changes from Y and N to 1 and 0 respectively.

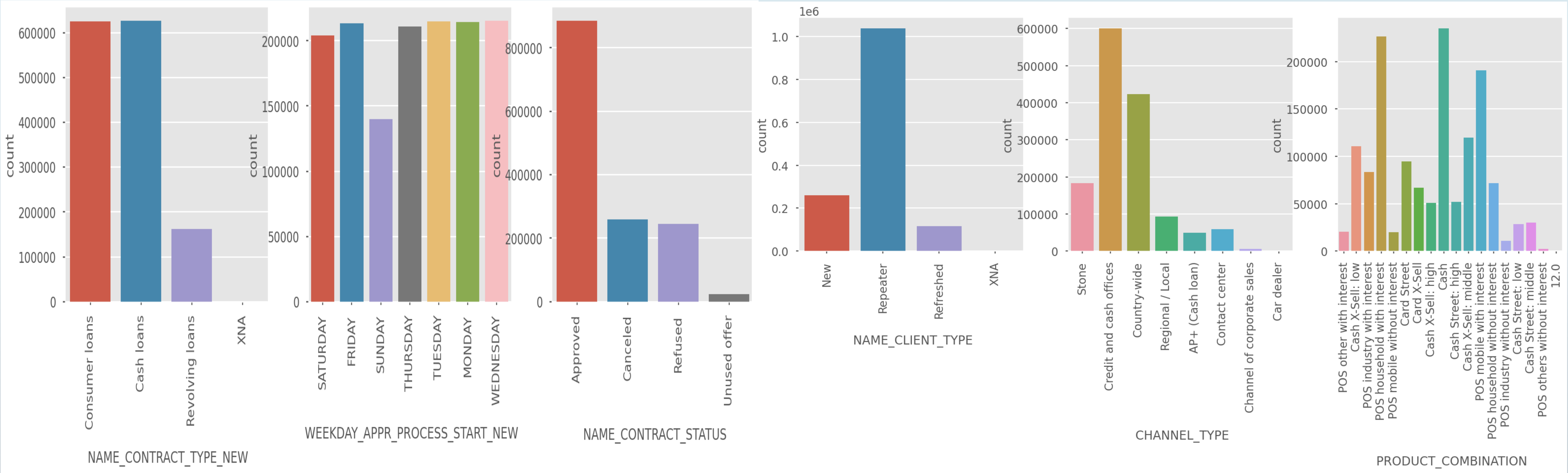
13. Analysing Numerical columns for outliers i.e. univariate numerical analysis.

- AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE have values present outside the upper extreme. i.e. outliers are present.
- HOUR_APPR_PROCESS_START does not have any outliers.
- AMT_APPLICATION, SELLERPLACE_AREA and CNT_PAYMENT too have values outside the upper extreme. outliers are there.
- Outliers can be handled by binning values in AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE.

14. Merging both the dataframes df1 and df2 i.e. application data and previous application.

- Merging both the data frames with inner join with suffixes '_OLD' and '_NEW' , new dataframe is named as df3.

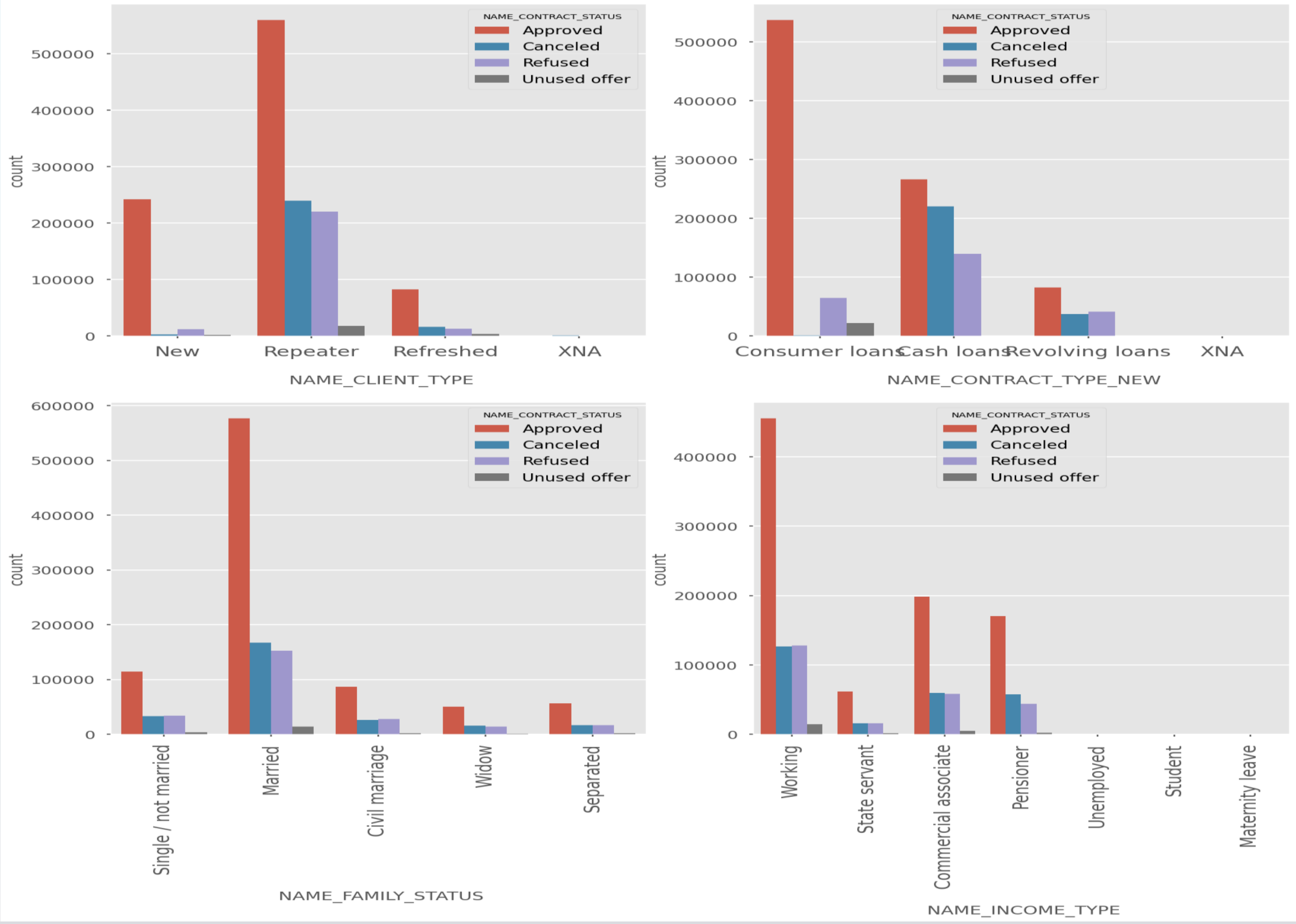
15. Univariate Categorical analysis

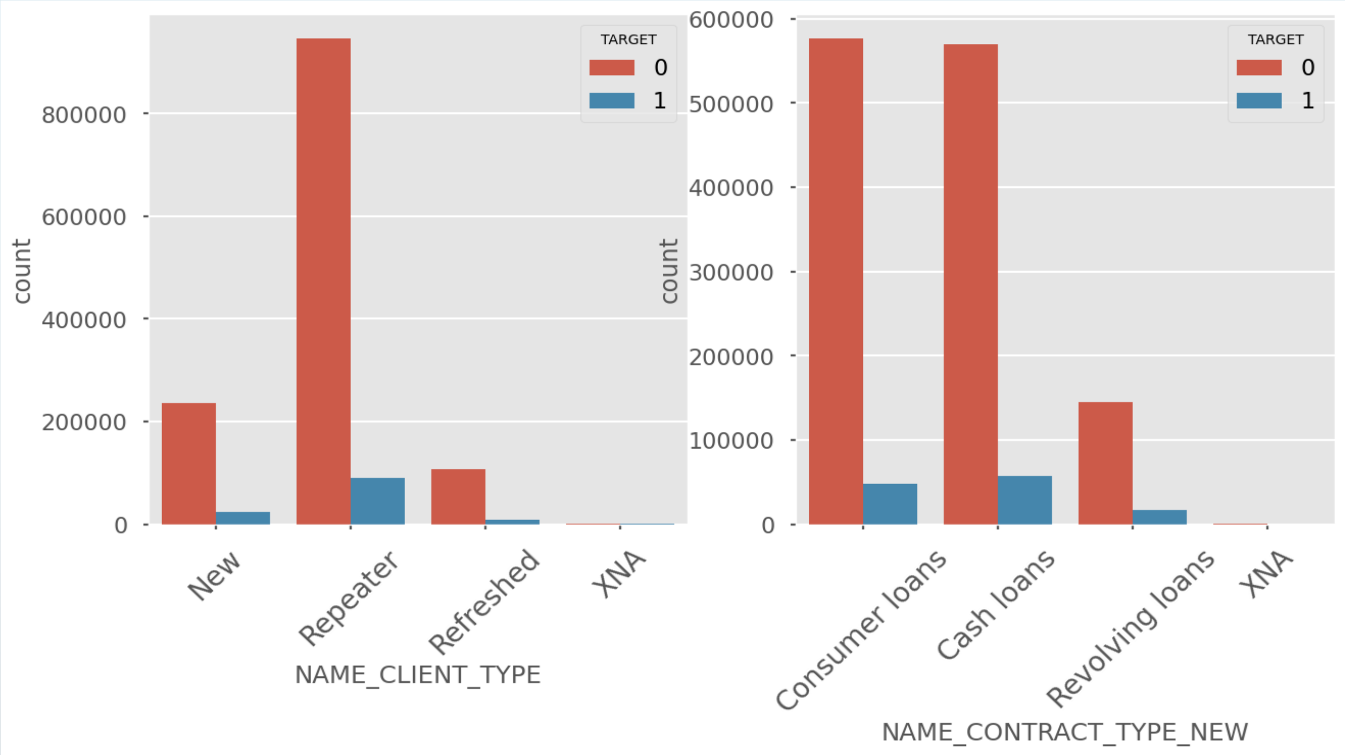


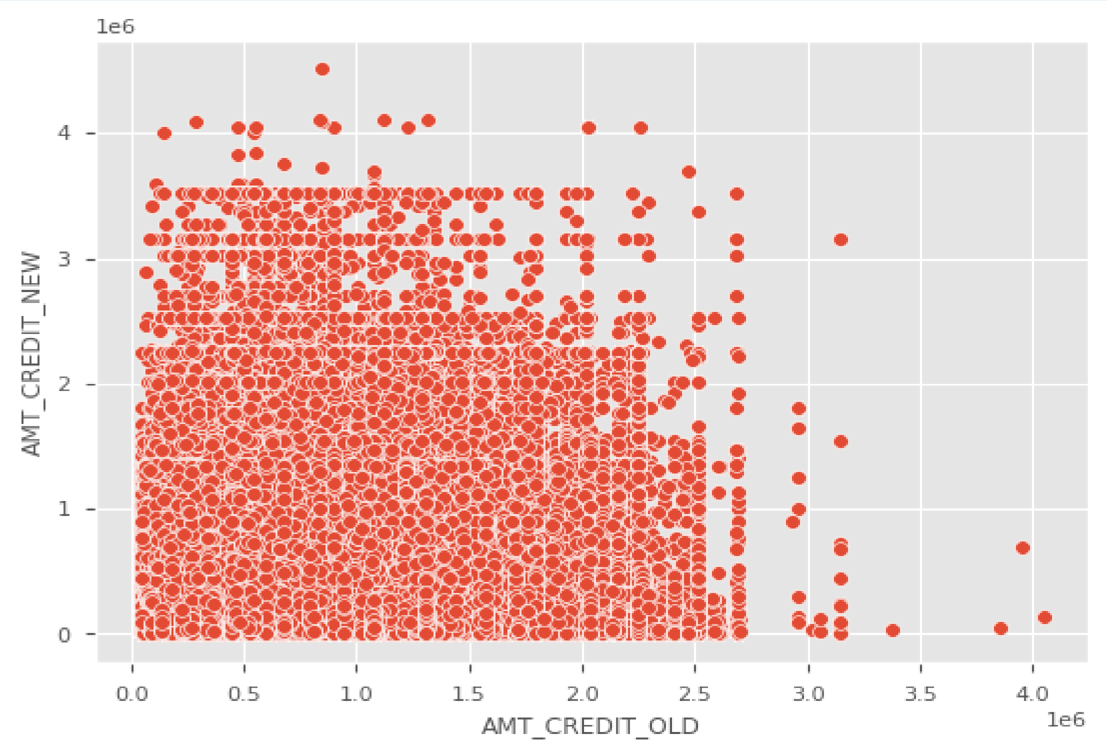
- The no. of Consumer loans and cash loans are nearly identical. The no. of Revolving loans is far less than both the consumer loans and cash loans.
- Clients applied for the loan majorly on the first 4 days of the week and Majority of the loans are approved.
- Majority of clients are repeater and Credit/Cash offices prove to be the most efficient channel for client acquisition.
- POS household with interest, cash and POS mobile with interest are top 3 Product combinations.

16. Bivariate Analysis

- Loan for new client, Repeater and Refreshed are mostly approved. but the no. of New clients, Repeater and Refreshed vary.
- Almost all the Consumer loans are approved except for a few refused and unused offers.
- The Approval, Refusal and Unused status are approximately evenly distributed for cash and revolving loans.
- Married and Working clients have their loans approved as compared to other categories.







- Most of New , Repeater and Refreshed clients has no difficulty in installment payments.
- Similarly most of the clients with consumer, cash and revolving loans has no difficulty in installment payments
- The Analysis of AMT_CREDIT_NEW with AMT_CREDIT_ OLD. What amount the client applied for and what amount the client received during approval.
- The spread of credit amount in the previous loan application to current application is mostly around 2-3 Lakhs.

5. Conclusion

- Most of the loans are Cash and Consumer loan. Consumer loan has the highest approval rate. Cash loans are repayed on time.
- In the cash loans people with occupation as Laborers and core staff applied the most.
- In terms of gender Female clients earn more, face less difficulty in repayment as compared to male clients.
- Working clients are in great numbers, their loans are approved mostly and they face less difficulty in repayment.
- Married client’s loans get approved more as compared to others.
- People working, having secondary education and working in Business entity type face less difficulty.
- People who are applying again have high chances of getting their loan approved.
- Credit and cash offices, country wide prove to be the most efficient channel for client acquisition.
- The product in demand are cash, POS industry with interest, POS mobile with interest. but most of the cash products are rejected.
- The income spread is around 1-2 lakhs.

6. Recommendations

1. The loan providing company should focus more on cash product as the demand for this product is as much as POS industry with interest and POS mobile with interest. But most of the cash product are getting rejected.
2. Working clients, people having Laborers , core staff as occupation, married clients, clients working in business entity type 3, having secondary education face no difficulty in loan repayments. The company should focus more on these clients who certainly benefit the business.
3. Company should focus more on Female clients as they earn more, apply more for loan and face less difficulty in repayment as compared to male clients.
4. In addition to the second point, people who are commercial associate, Pensioner, state servant, people with higher education, those who are married, prove to be beneficial for the business.
5. Loan providing company could set up more credit/ cash office and country wide channel, as these are the main source of client acquisition.
6. The company could target people having income around 1-2 lakhs, as they apply for loan more and they face less difficulty in repayments.