Credit EDA Assignment

By – Pankti Nilamkumar Patel

Project Description

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding:
The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You must use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

Project Description

• 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

Business Objectives:
It aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants' using EDA is the aim of this case study. 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 utilize this knowledge for its portfolio and risk assessment.

Approach and Methodology

1. Data Understanding and preparation
    - 'Application_data.csv' - Contains the information of the loan applicant and weather the applicant has made late payments in the "Target" column
    - 'Previous_application.csv' - Previous loan applications and their status for the client in the application data. a A client represented by SK_ID_CURR can have multiple previous loan applications
    - columns_description.csv describes the columns in the data tables.
    - Use pandas libraries to load and clear data
    - Use various visualization libraries to create graphs to understand patterns.

2. Data Cleaning and manipulation.

    * Identify variable with maximum empty data-variable.

    * Analyze each variable from the existing variable.

Approach and Methodology

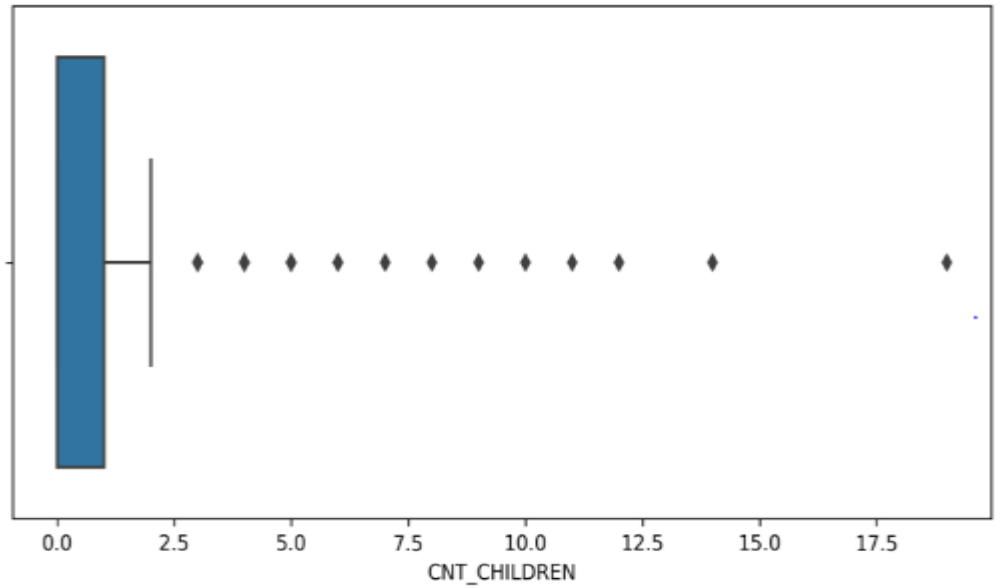
3. Data Analysis

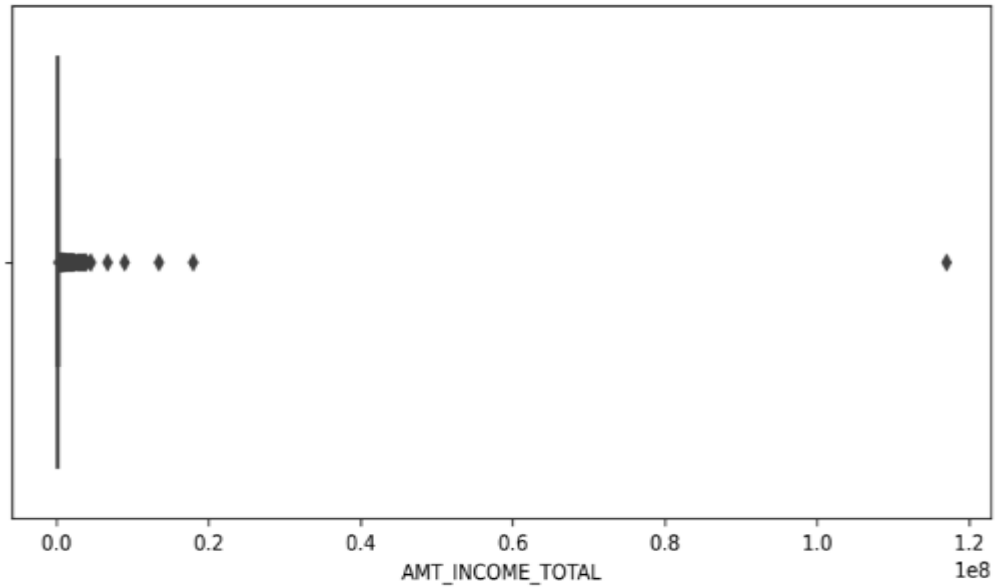
- Univariate Analysis
  - Used count plot in loops to examine the frequency distribution of categorical variables.
  - Used box plot to see distribution of numerical variables.
- Bivariate Analysis
  - Used scattered plot for numerical bivariate analysis

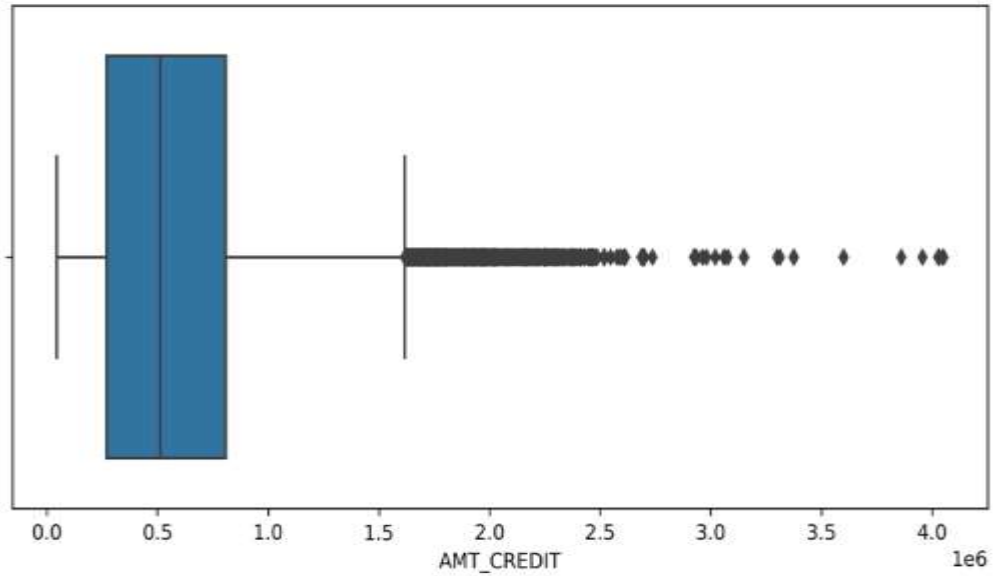
4. Combined Data analysis

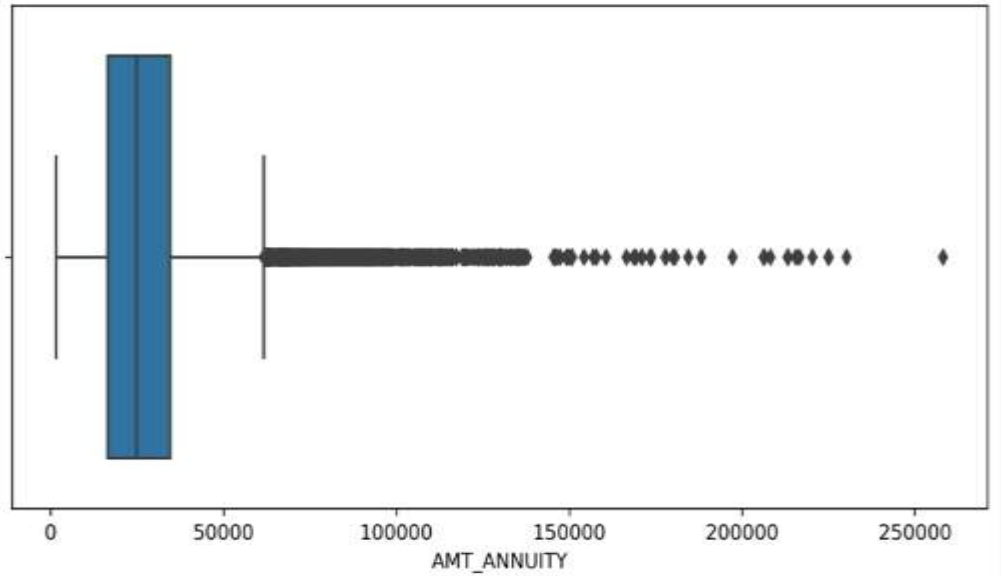
- Joined both data set on SK_ID_CURR.
- Created Heat map for different combination of variables

Finding Outlier For Numerical Values

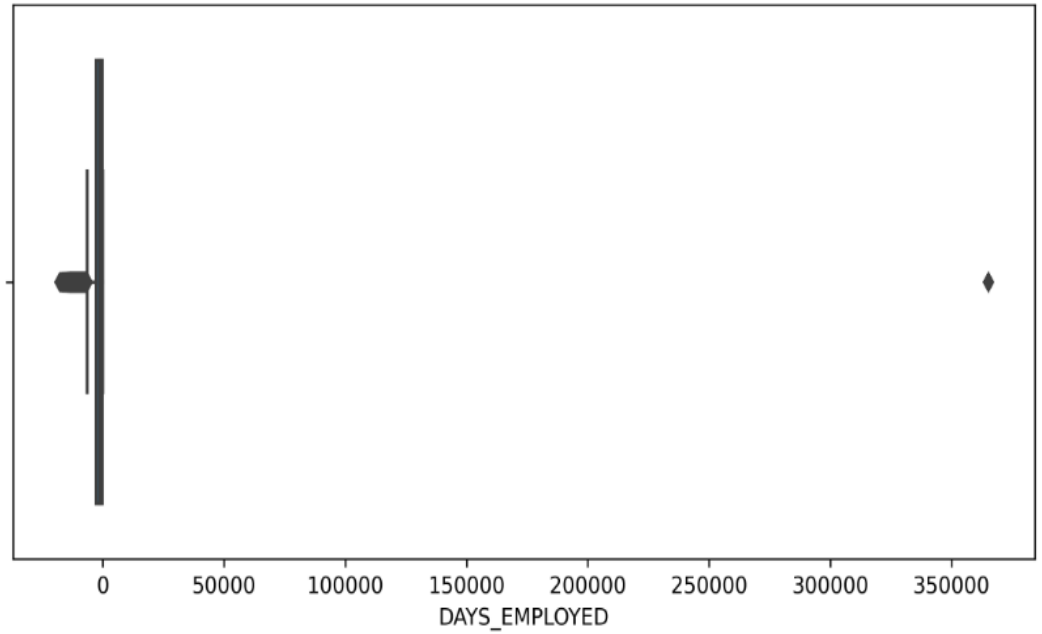


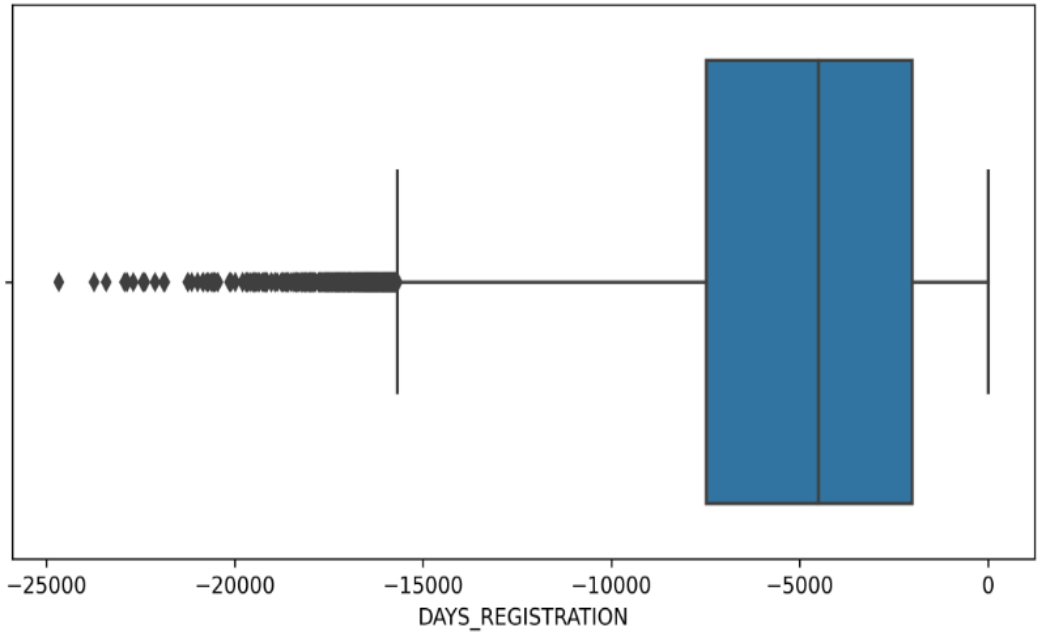






Finding Outlier For Numerical Values



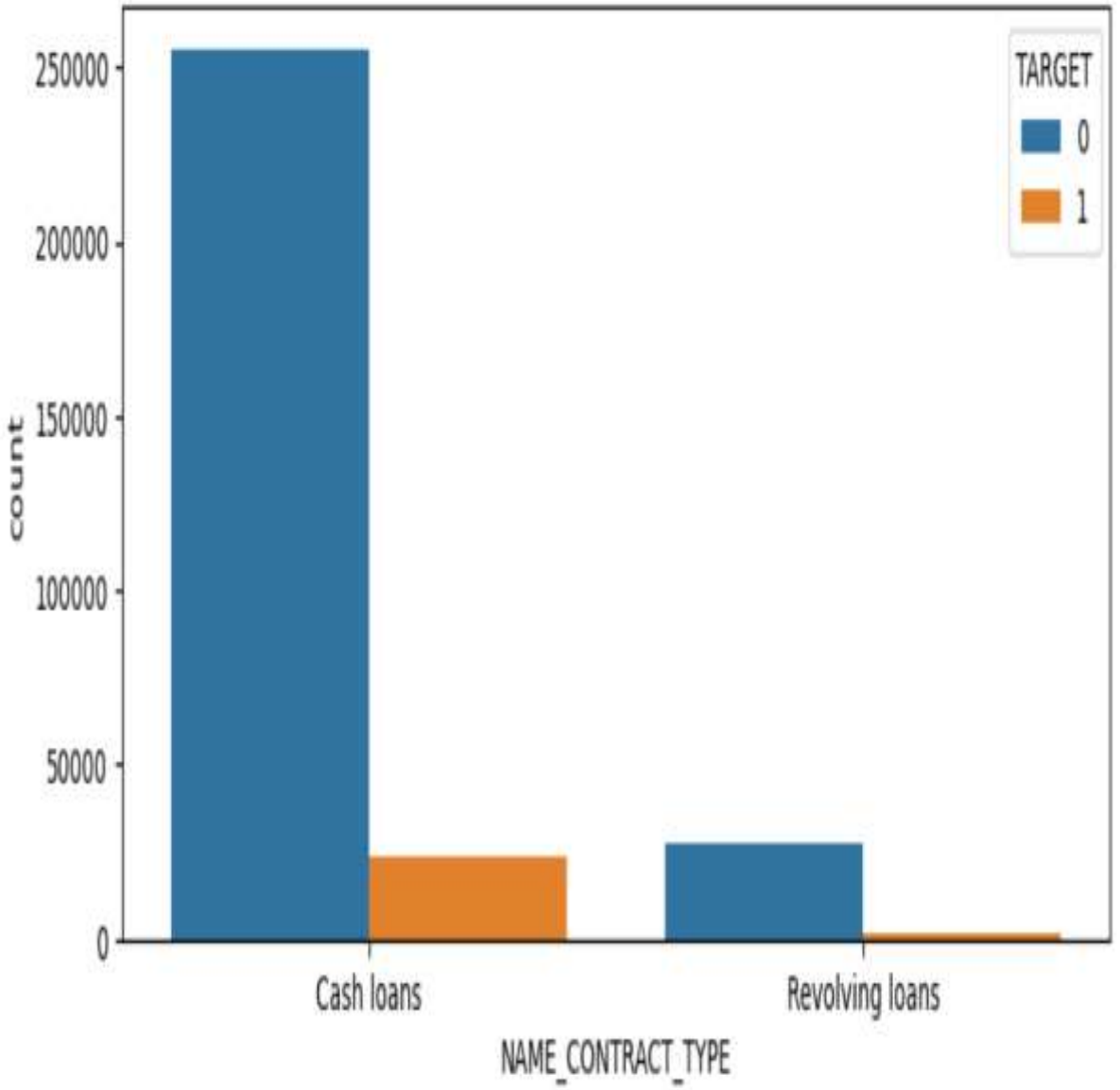


- **CNT_CHILDREN:** The figure shows that certain values can reach as high as 19, which is not feasible in the majority of cases.Hence, an outlier
- **AMT_INCOME_TOTAL:** As we can see from the plot, one number is very high when compared to the rest. As a result, it an outlier
- **AMT_ANNUITY:** As we can see from the plot the outlier value is above 250000 loan which makes the case for it being an outlier
- **DAYS_EMPLOYED:** As we can see from the plot the outlier value is 350000 days which makes the case for it being an outlier

Univariate Analysis for categorical variable in Application Data

NAME_CONTRACT_TYPE

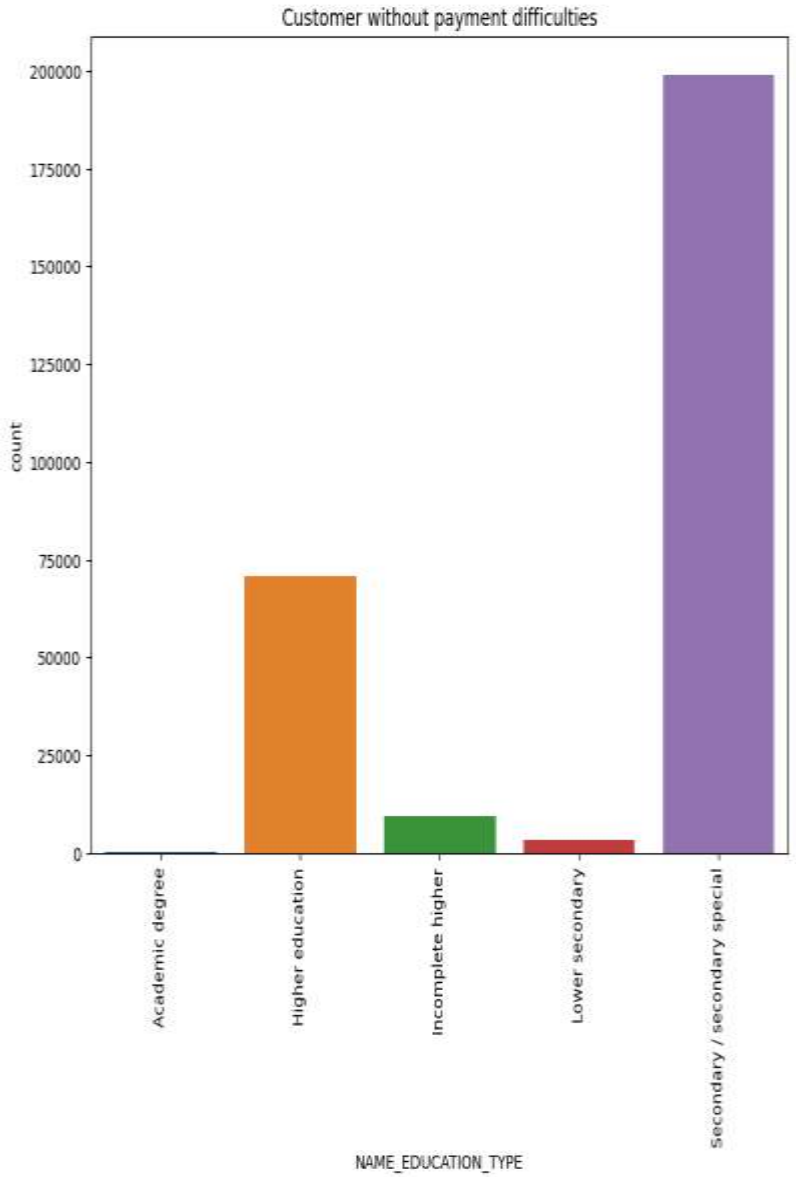
- we can see that the customer without payment and customer with payment difficulties both are taking cash loans rather than revolving loan

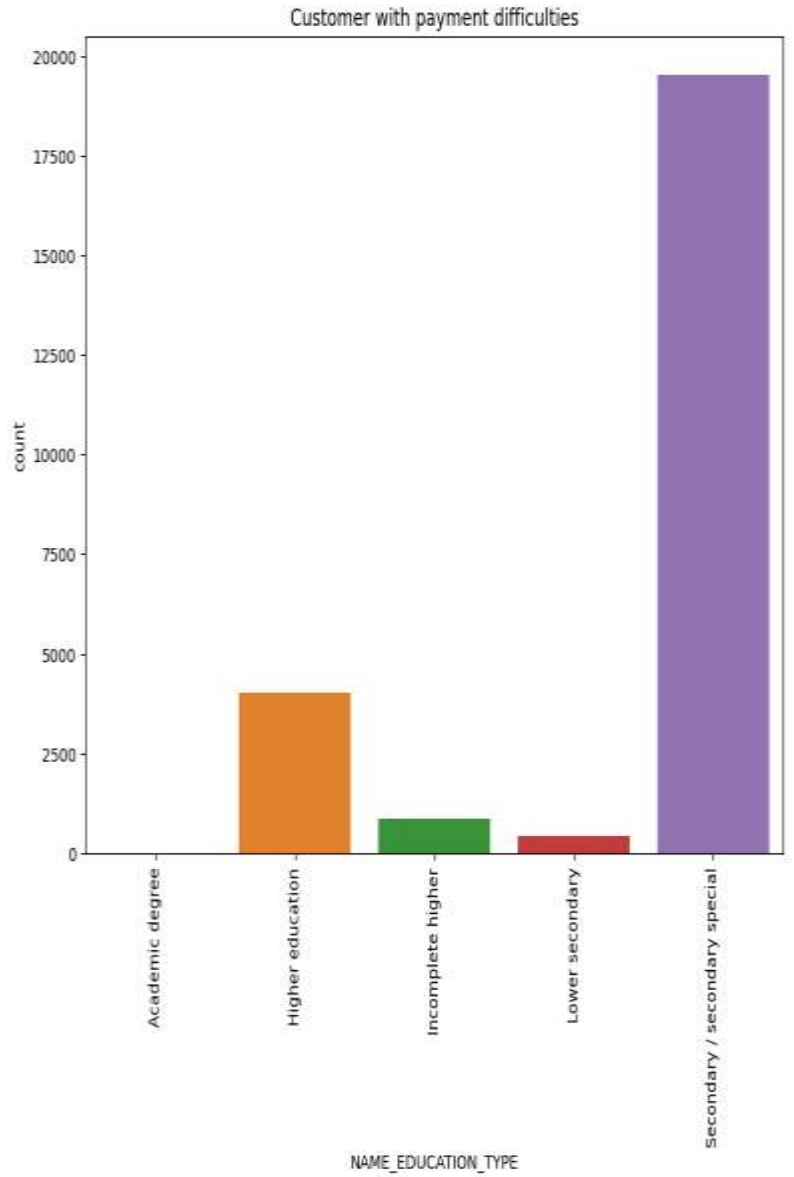


Univariate Analysis for categorical variable in Application Data

NAME_EDUCATION_TYPE

- we can see that the customer having payment difficulties in secondary/ secondary special in both the cases.

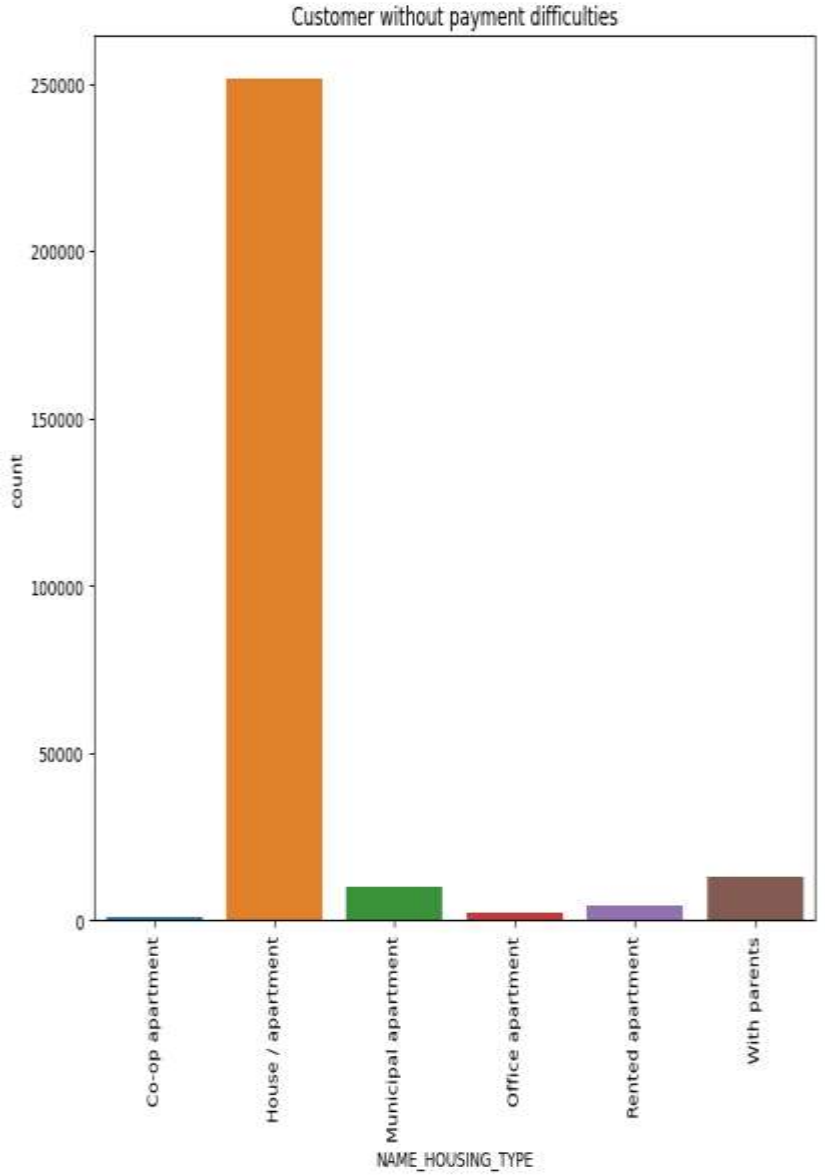


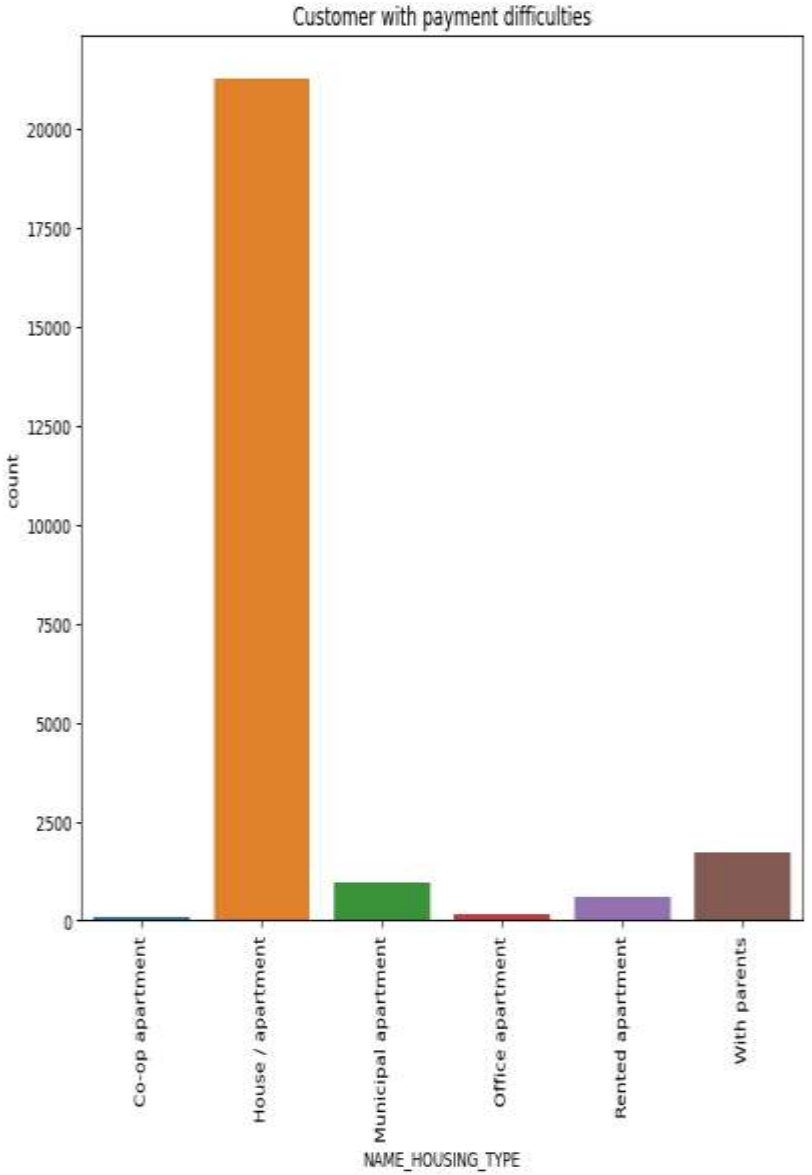


Univariate Analysis for categorical variable in Application Data

NAME_HOUSING_TYPE

- we have the payment difficulties in home/ apartment in both the cases. And we can also say that customers take loan for house/ apartment in compare to others.

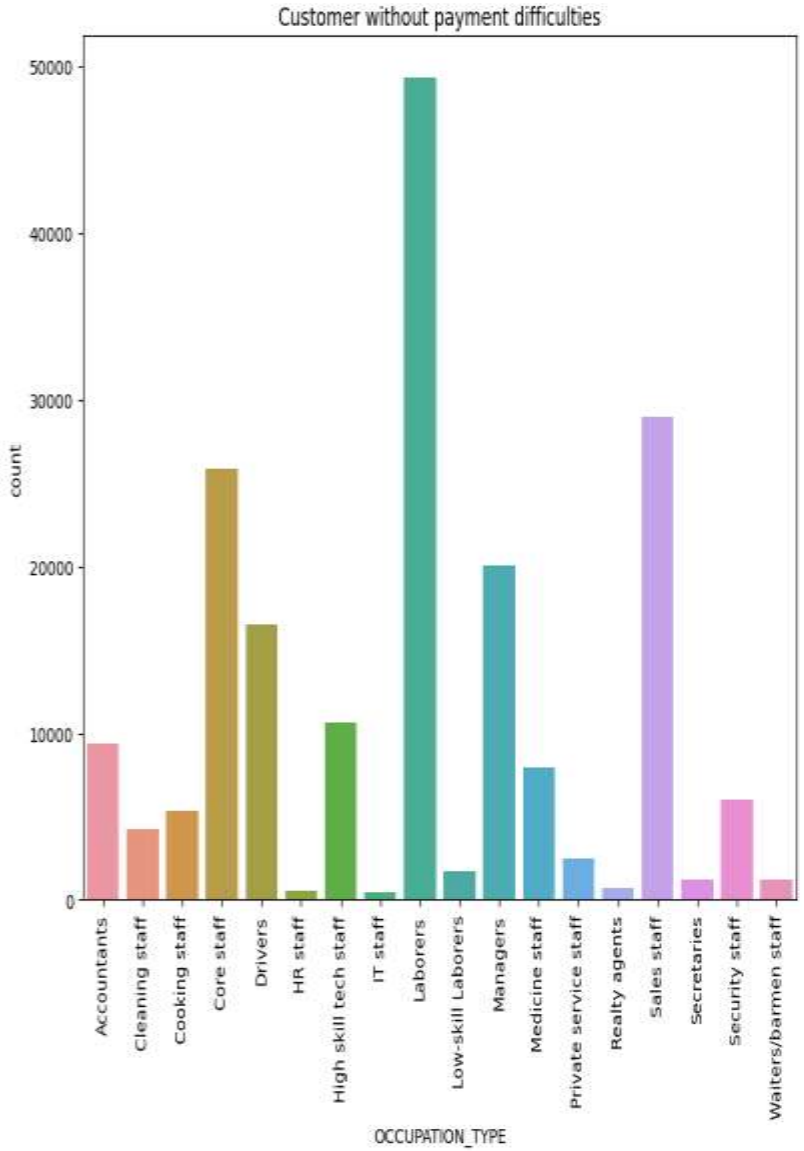


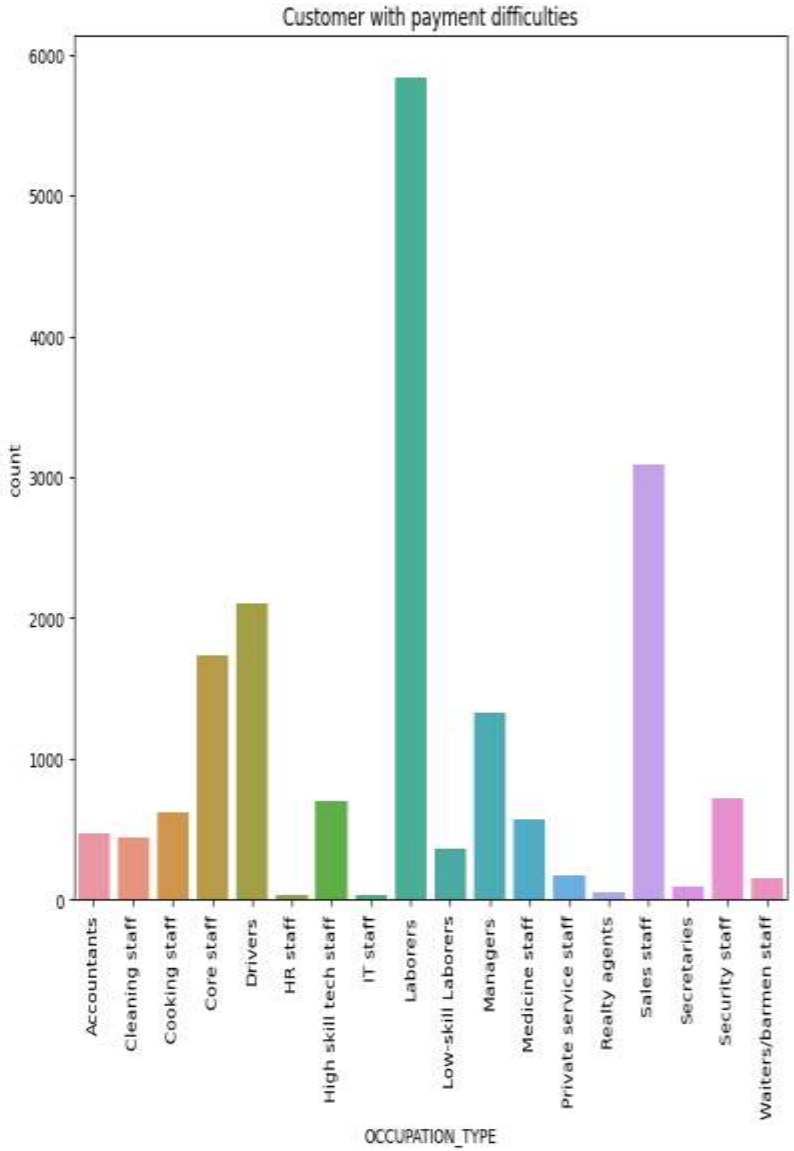


Univariate Analysis for categorical variable in Application Data

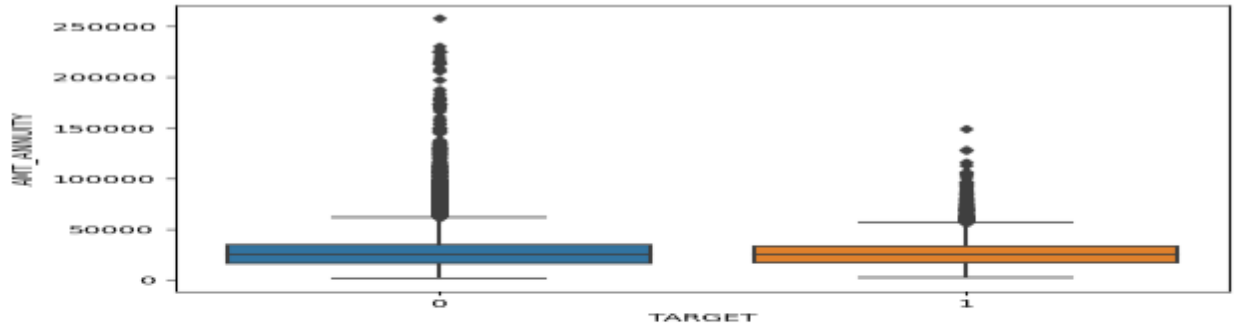
OCCUPATION_TYPE

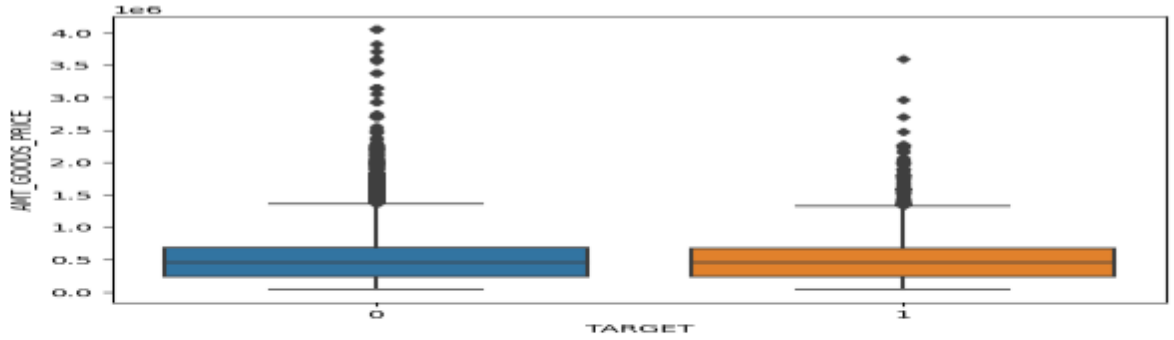
- we can see that laborers as well as the core staff and the sales staff, are having more difficulty repaying the loan. But in the case of laborers those wo have without payment is way more then with having the payment.

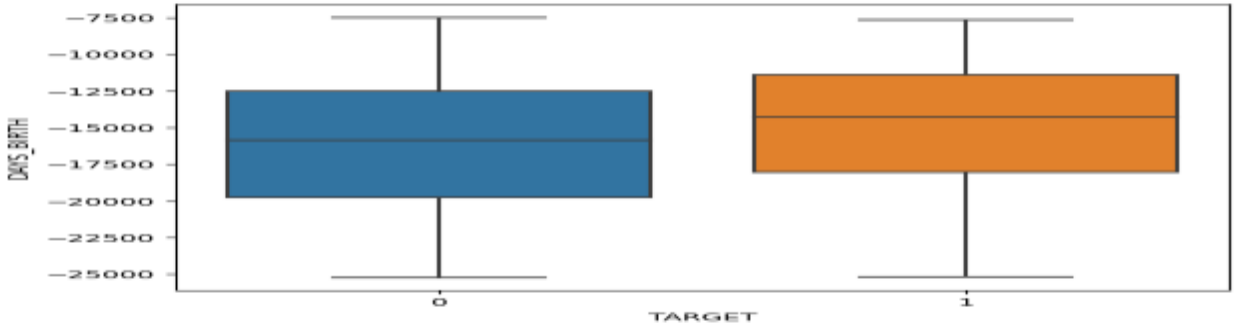


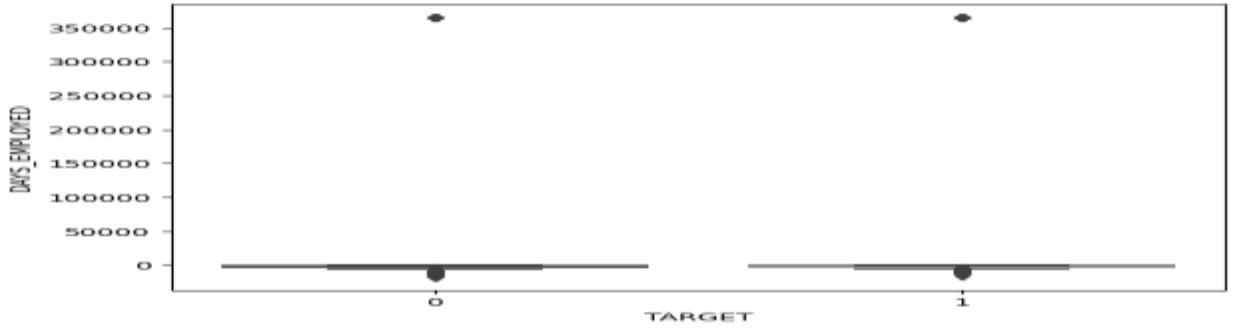


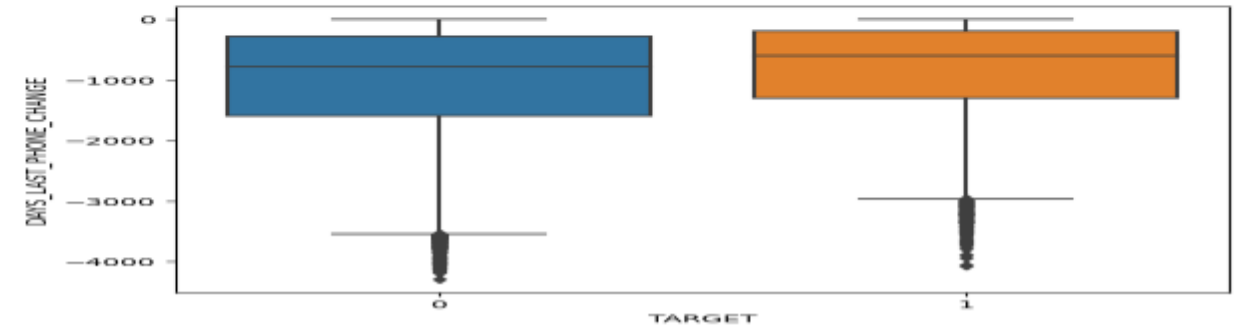
Univariate Analysis for numerical variable in Application Data

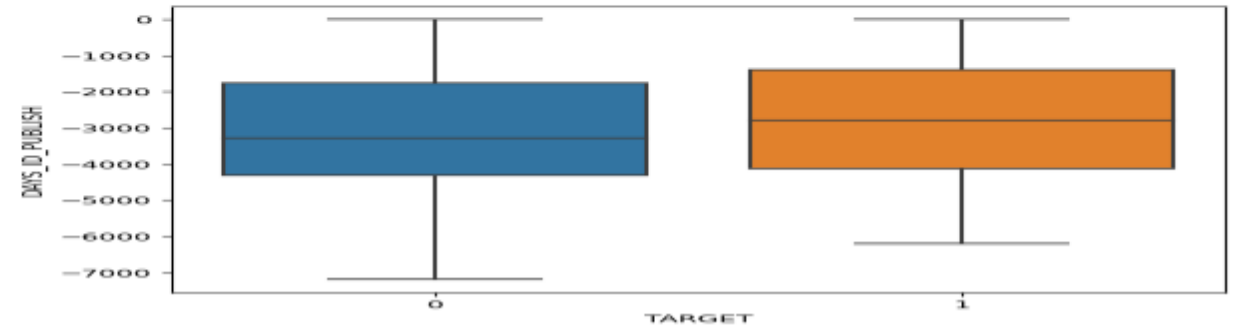








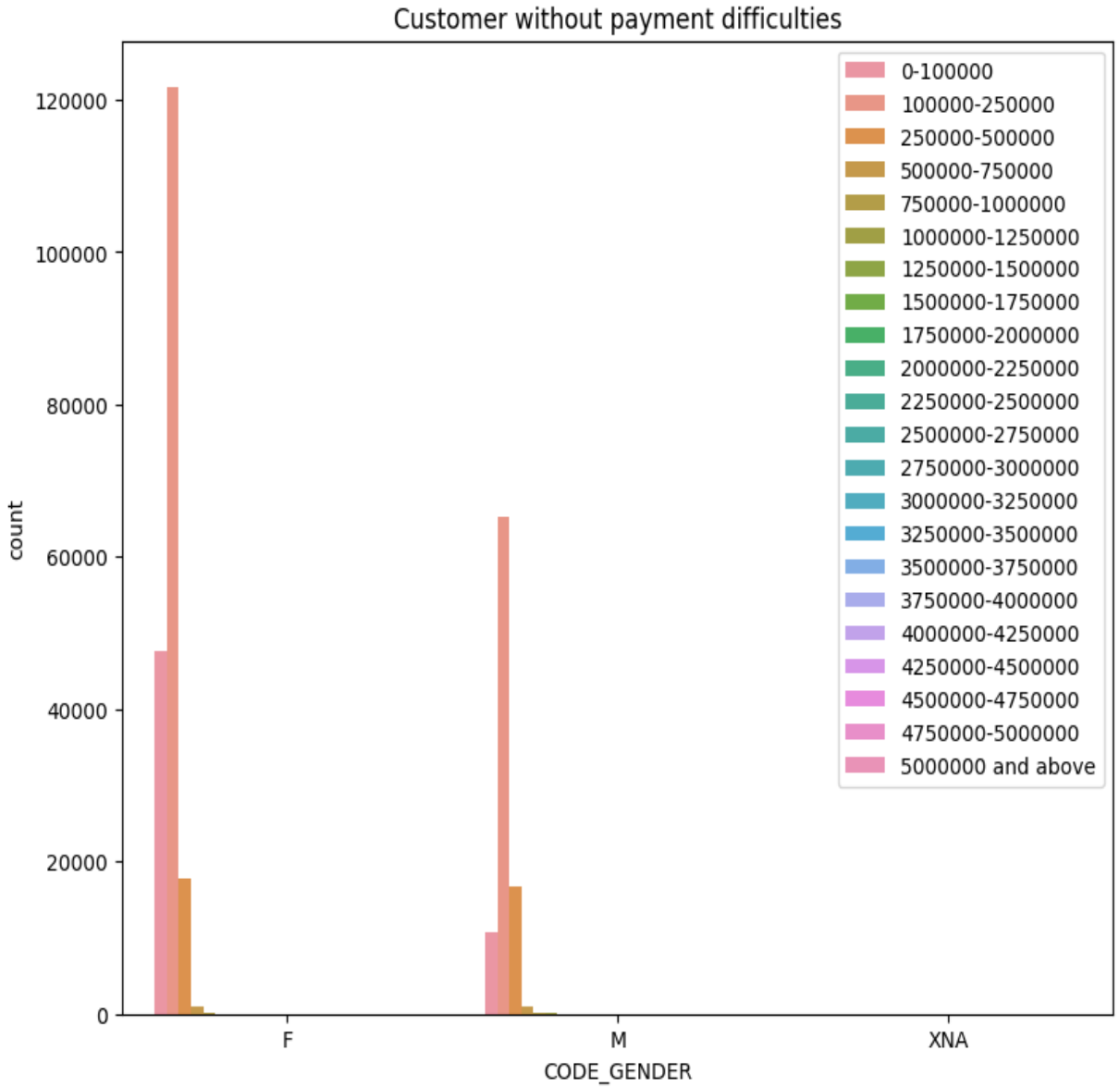


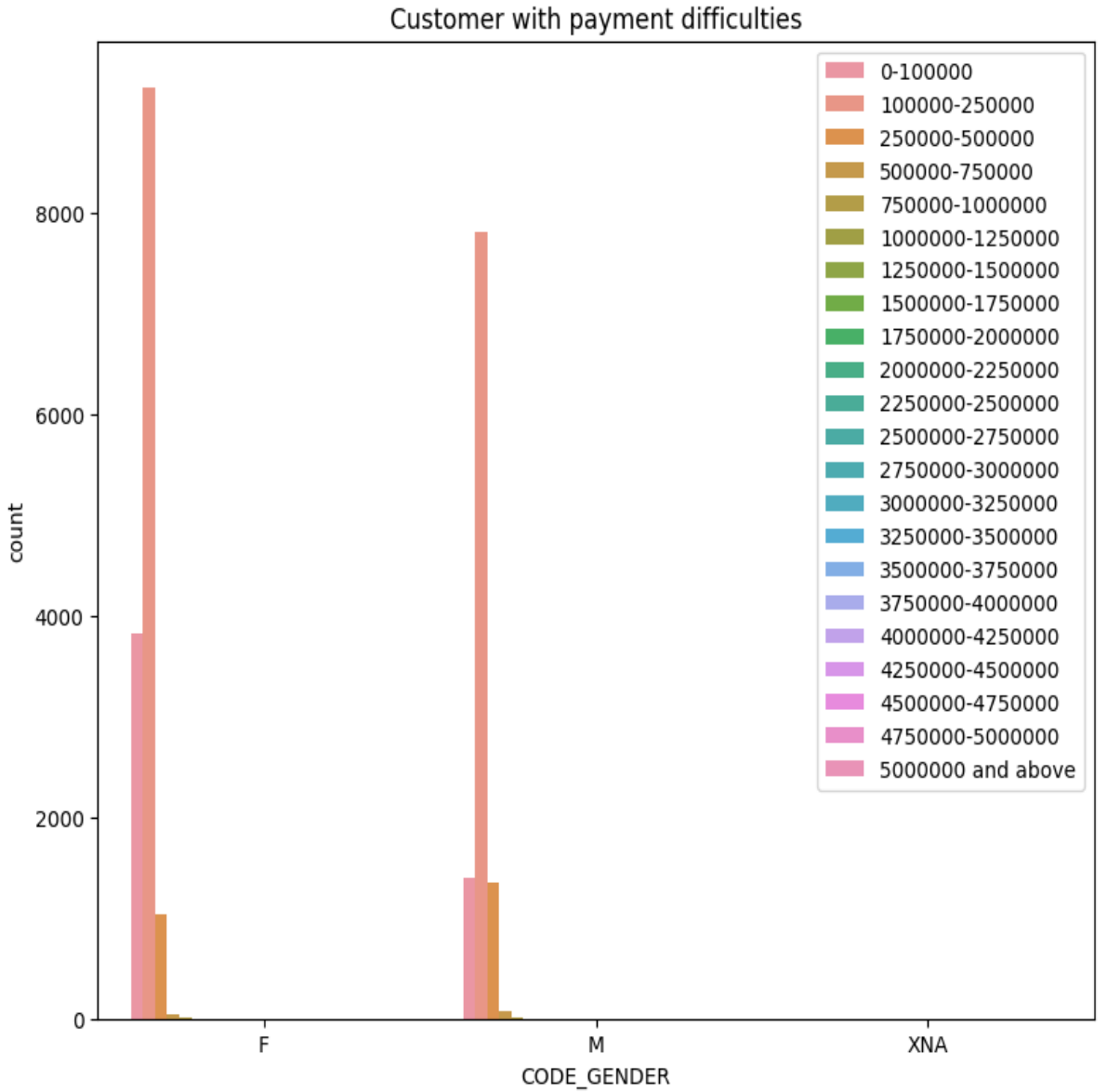


Univariate Analysis for numerical variable in Application Data

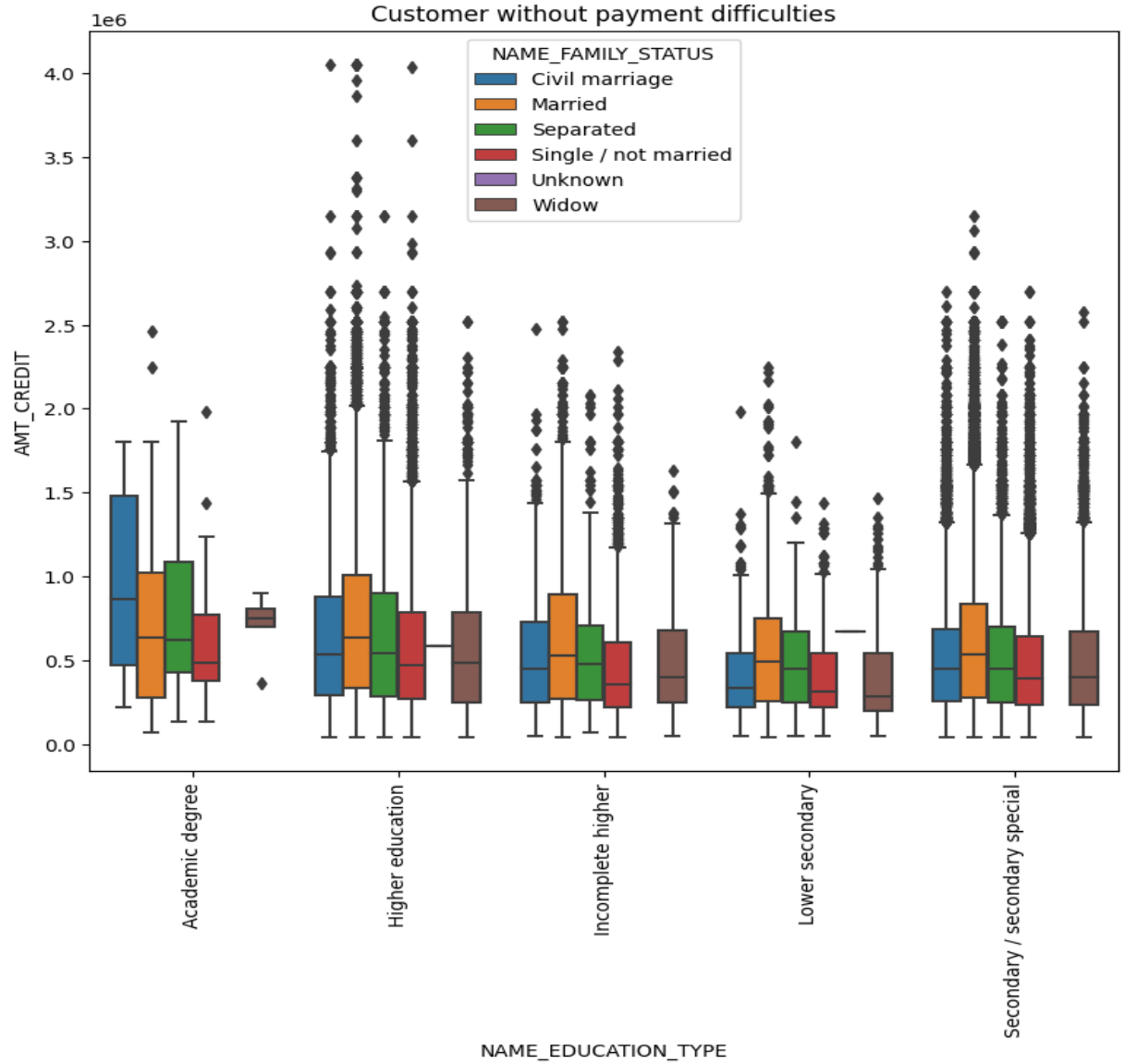
- In **AMT_ANNUITY** People without payment difficuties take more credit for the annuity that they have
- In The **AMT_GOODS_PRICE** boxplot we can see that the customer without payment difficulties lies in between 0.3 to 0.7 and the customer with payment difficulties lies in between the same as of the without payment 0.3 to 0.7 and also both are having the mid value about 0.5.
- In **DAYS_BIRTH** people having higher age are having higher probability of repayment.
- There is single high value data point as outlier present in **DAYS_EMPLOYED**.
- Less outlier observed in **DAYS_ID_PUBLISH**
- In **DAYS_ID_PUBLISH** People who have recently changed their IDs are more likely to be defaults.

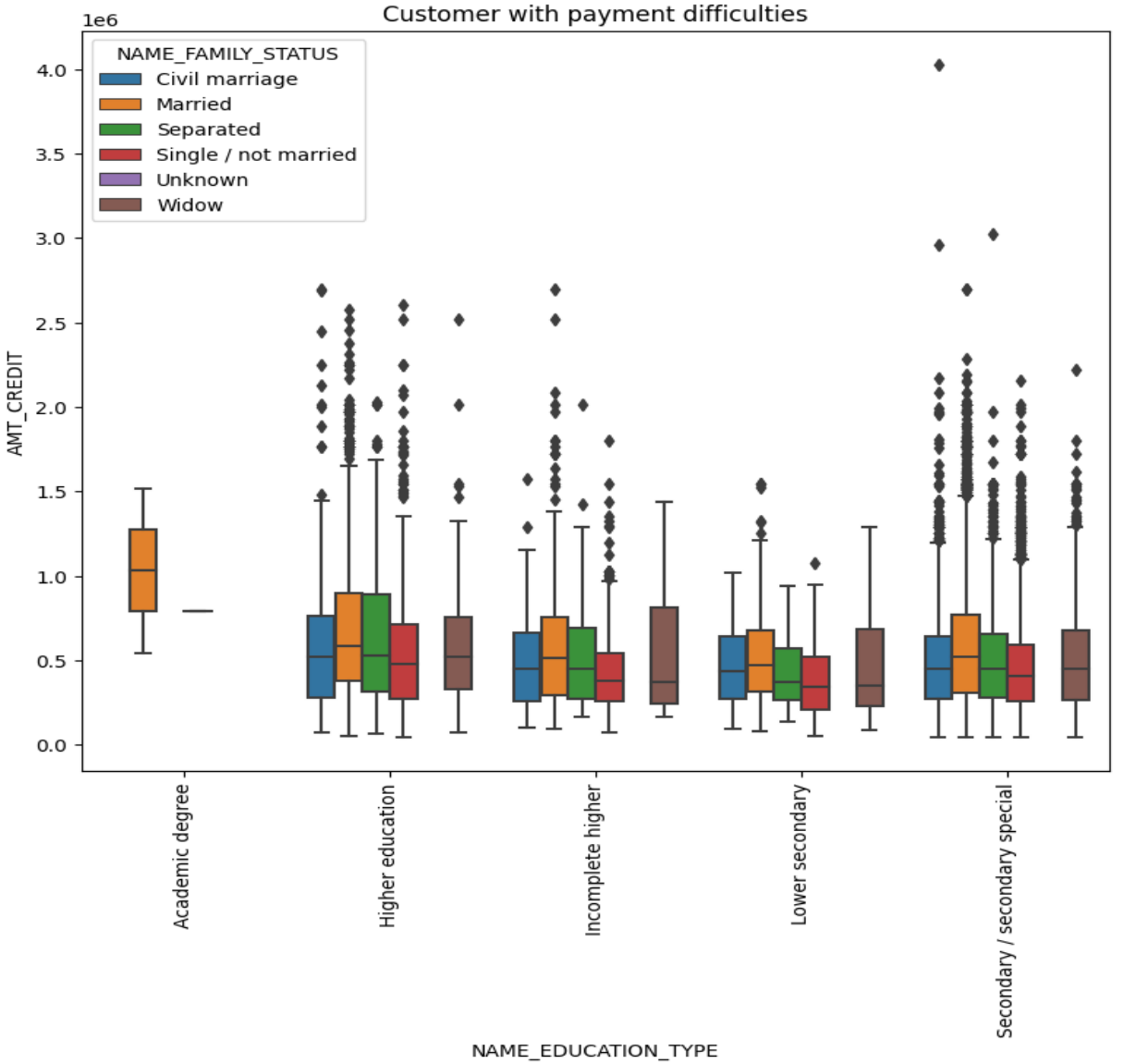
Bivariate Analysis for categorical variable in Application Data





Bivariate Analysis for numerical variable in Application Data





Top 10 Correlations

Highly corelate columns for non defaulters

1. AMT_CREDIT and AMT_GOODS_PRICE (0.98)
2. AMT_CREDIT and AMT_ANNUITY (0.77)
3. AMT_ANNUITY and AMT_GOODS_PRICE (0.77)

	VAR1	VAR2	Correlation_Value	Corr_abs
34	AMT_GOODS_PRICE	AMT_CREDIT	0.987250	0.987250
35	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686	0.776686
23	AMT_ANNUITY	AMT_CREDIT	0.771309	0.771309
71	YEARS_EMPLOYED	YEARS_BIRTH	0.626114	0.626114
22	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953	0.418953
33	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462	0.349462
11	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799	0.342799
82	YEARS_REGISTRATION	YEARS_BIRTH	0.333151	0.333151
94	YEARS_ID_PUBLISH	YEARS_EMPLOYED	0.276663	0.276663
93	YEARS_ID_PUBLISH	YEARS_BIRTH	0.271314	0.271314

Correlations

Top 10 Correlations

Highly corelate columns for defaulters

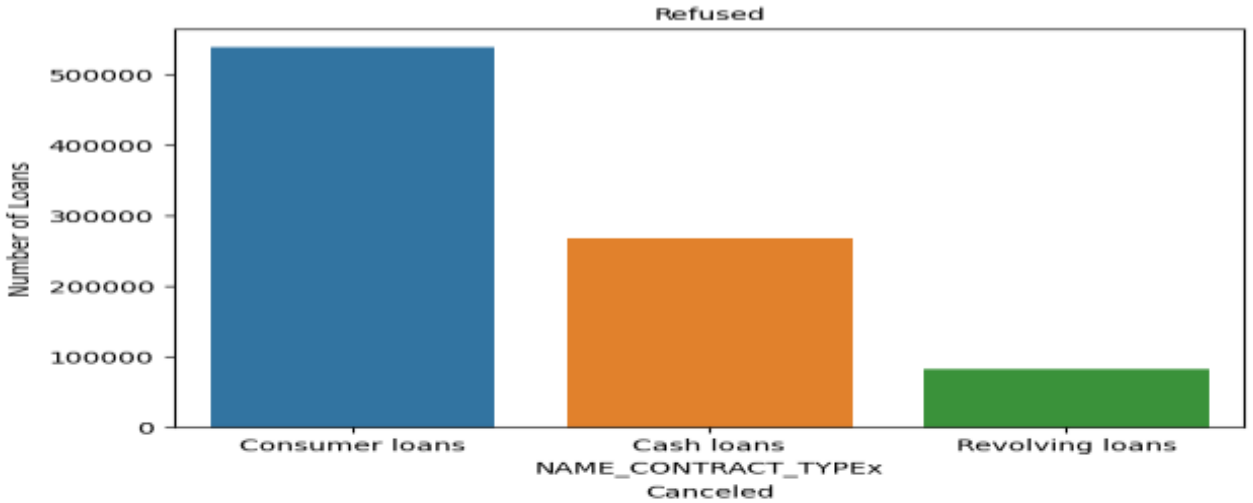
1. AMT_GOODS_PRICE and AMT_CREDIT (0.98)
2. AMT_CREDIT and AMT_ANNUITY (0.75)
3. AMT_GOODS_PRICE and AMT_ANNUITY (0.75)

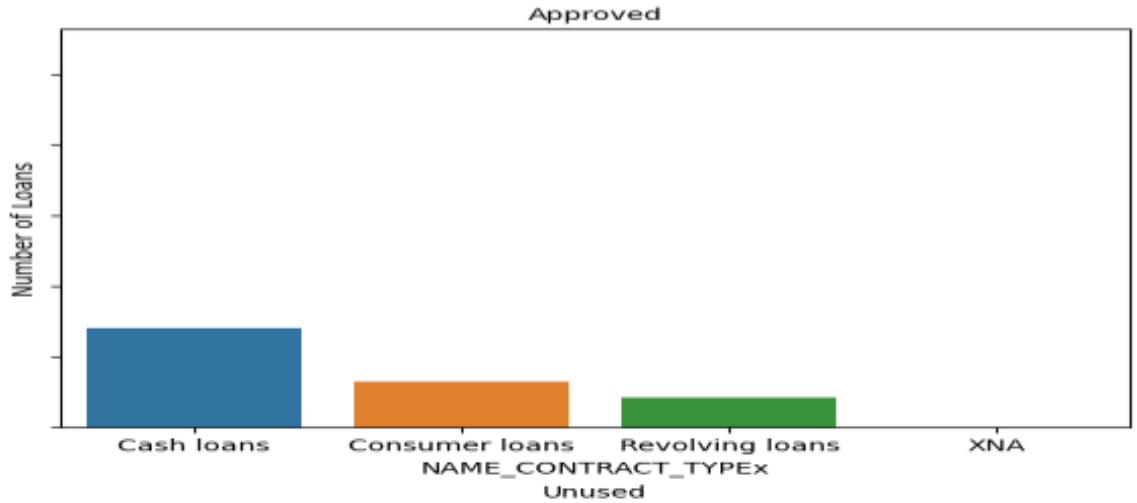
	VAR1	VAR2	Correlation_Value	Corr_abs
34	AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
35	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699	0.752699
23	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
71	YEARS_EMPLOYED	YEARS_BIRTH	0.582185	0.582185
82	YEARS_REGISTRATION	YEARS_BIRTH	0.289114	0.289114
93	YEARS_ID_PUBLISH	YEARS_BIRTH	0.252863	0.252863
94	YEARS_ID_PUBLISH	YEARS_EMPLOYED	0.229090	0.229090
83	YEARS_REGISTRATION	YEARS_EMPLOYED	0.192455	0.192455
115	EXT_SOURCE_3	YEARS_BIRTH	0.171621	0.171621
103	EXT_SOURCE_2	REGION_POPULATION_RELATIVE	0.169751	0.169751

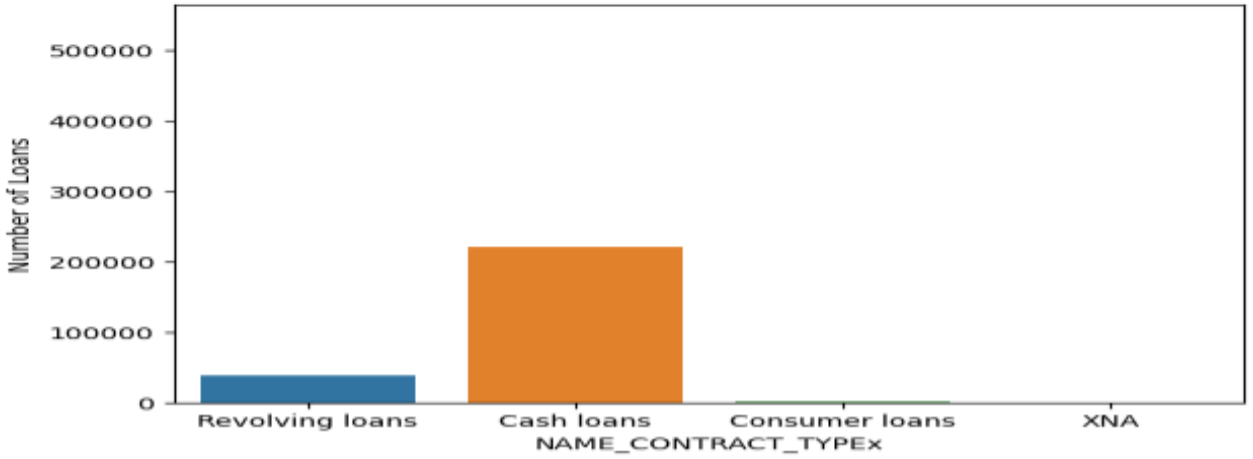
Correlations

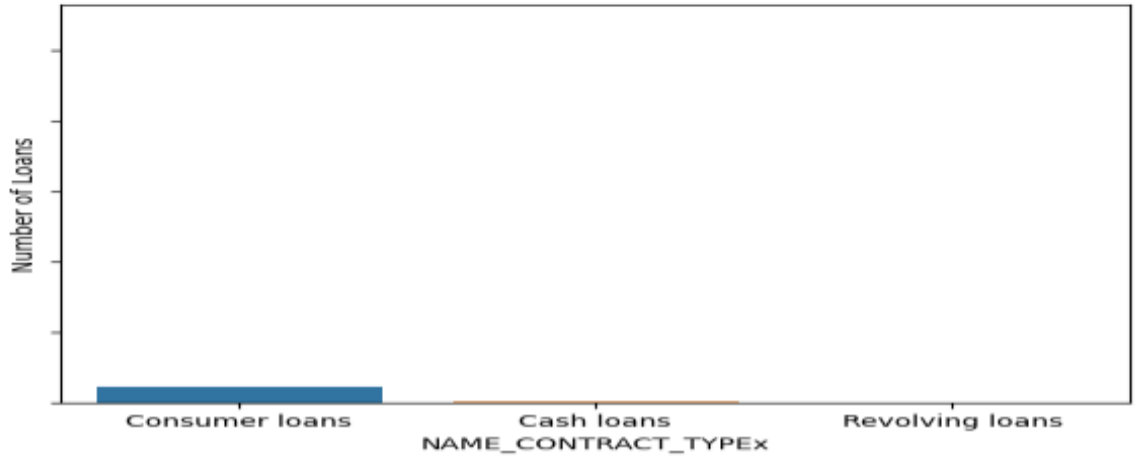
Merged Data

We combined the previous application data, which contained details regarding previous loans made to the client. It comprises information on the previous application's status (whether it was approved, cancelled, refused, or not utilized) as well as information about the current application to do integrated analysis.









Merged Data

- Here we can see that the Revolving loan is much more acceptable as compare to the cash and consumer loans.
- As we can see that to visualize 4 plots we wrote same code multiple times. so to avoid redundancy, and to save our time, we will put the above code in a function and generalize it for our following plots, so that its easy to visualize and saves time.

Conclusion

- For effective payments, banks should place a greater emphasis on contract types such as "student," "pensioner," and "businessman," as well as housing types other than "co-op apartment."
- Banks should focus less on the income category 'Working' because it has the highest proportion of failed payments.
- Also, the loan objective 'Repair' has a larger number of failed payments on time.
- Get as many consumers from the dwelling category 'With parents' as possible because they have the fewest failed payments.

Thank You