Credit EDA Assignment

Loan application data analysis

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 to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to 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 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 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.
-
- The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:
- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.
-
- When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** Loan has been cancelled by the client but at different stages of the process.
- In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

Business Objectives-

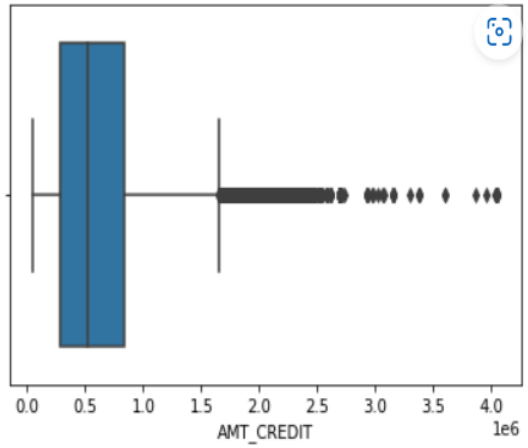
- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers 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 utilise this knowledge for its portfolio and risk assessment.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Handle missing values-

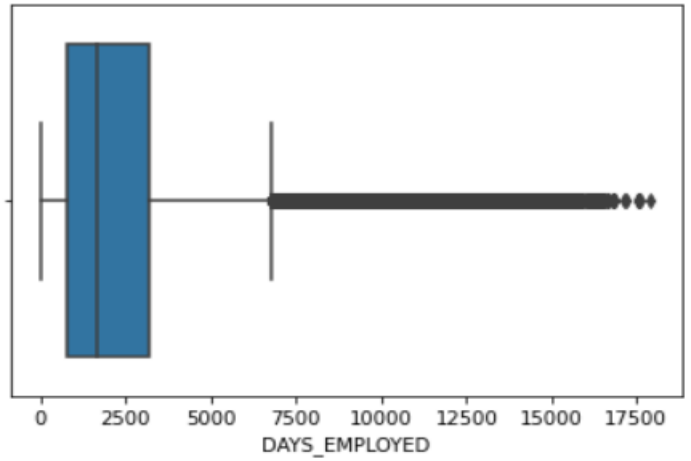
- After analysing null values, found out there were many null values.
- I tried to impute them by automated approach.
- For categorical columns, imputed the value with mode.
- For continuous columns, imputed the value with mean.
- In some columns there were XNA values too. Handled them and replaced them with the most frequent one.
- Did univariate and bivariate analysis over some columns.
- Plotted heatmap for finding top correlated columns.

Identification of outliers-

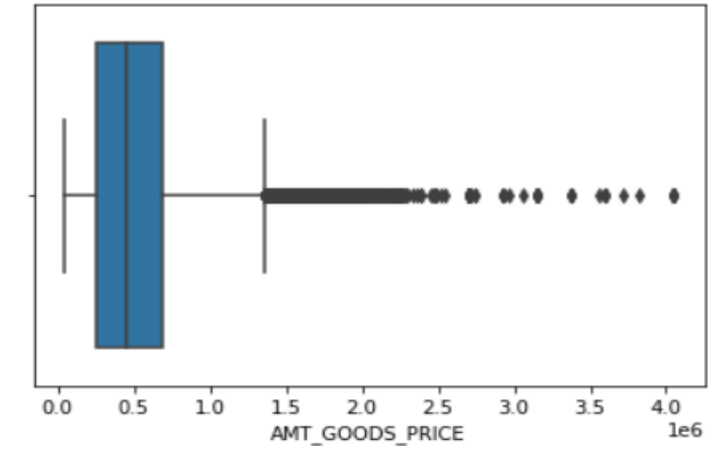
Boxplot of AMT_CREDIT



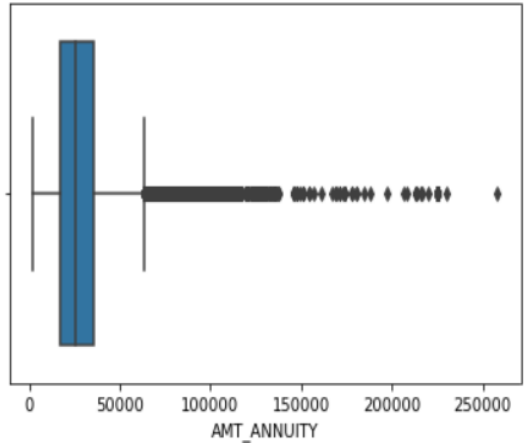
Boxplot of DAYS_EMPLOYED



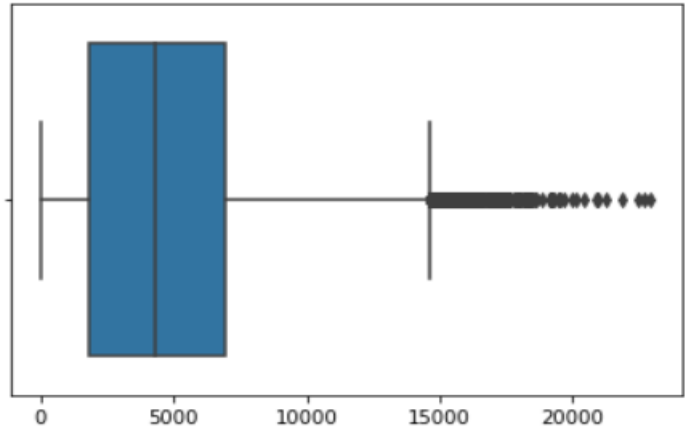
Boxplot of AMT_GOODS_PRICE



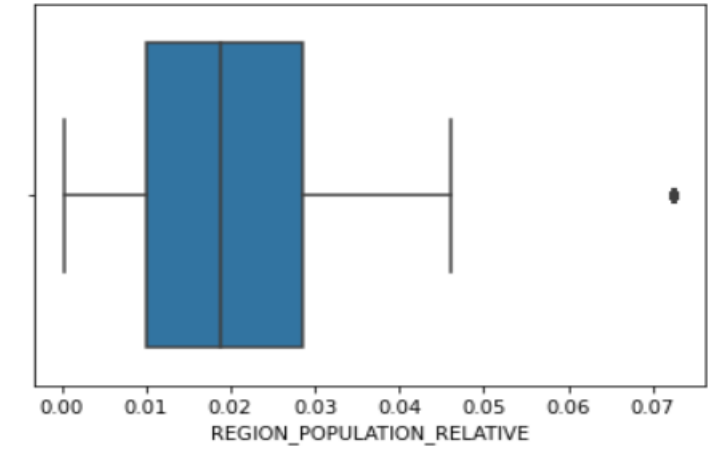
Boxplot of AMT_ANNUITY



Boxplot of DAYS_REGISTRATION

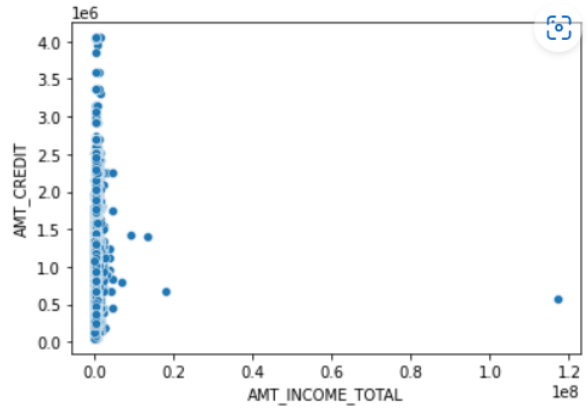


Boxplot of REGION_POPULATION_RELATIVE

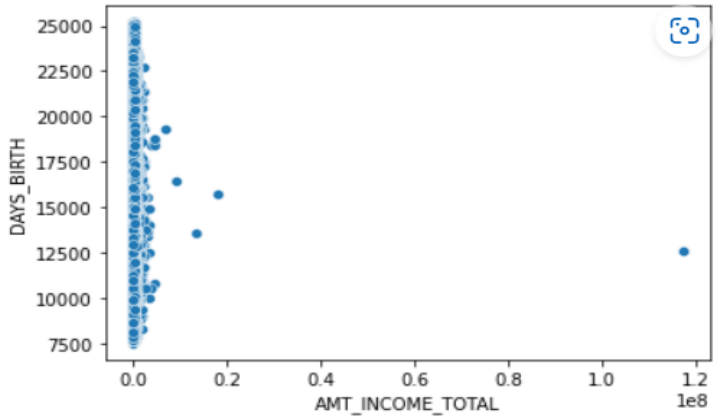


Bivariate analysis over continuous columns using scatter plot-

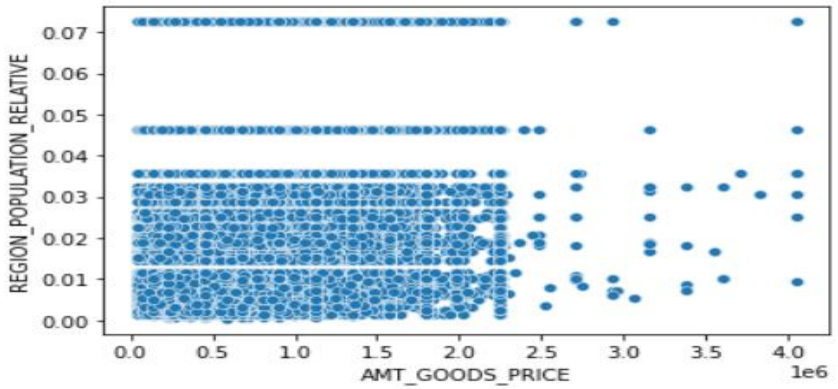
Scatterplot AMT_INCOME_TOTAL Vs AMT_CREDIT



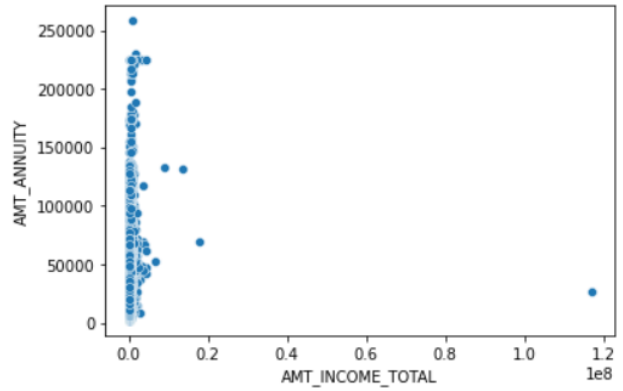
Scatterplot AMT_INCOME_TOTAL Vs DAYS_BIRTH



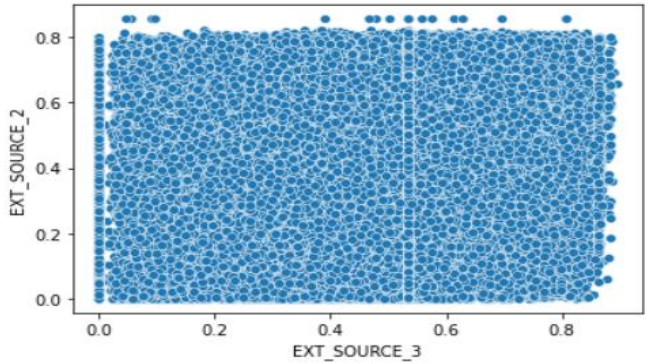
Scatterplot AMT_GOODS_PRICE Vs REGION_POPULATION_RELATIVE



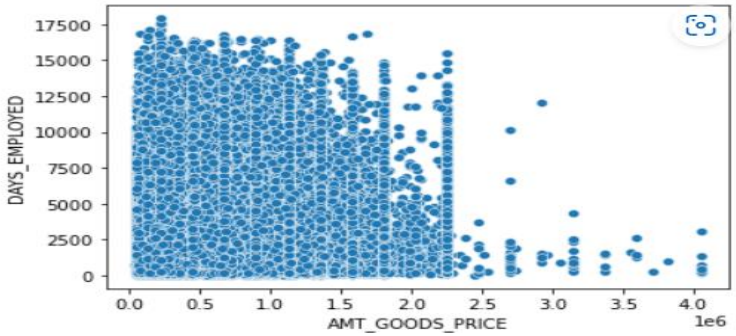
Scatterplot AMT_INCOME_TOTAL Vs AMT_ANNUITY



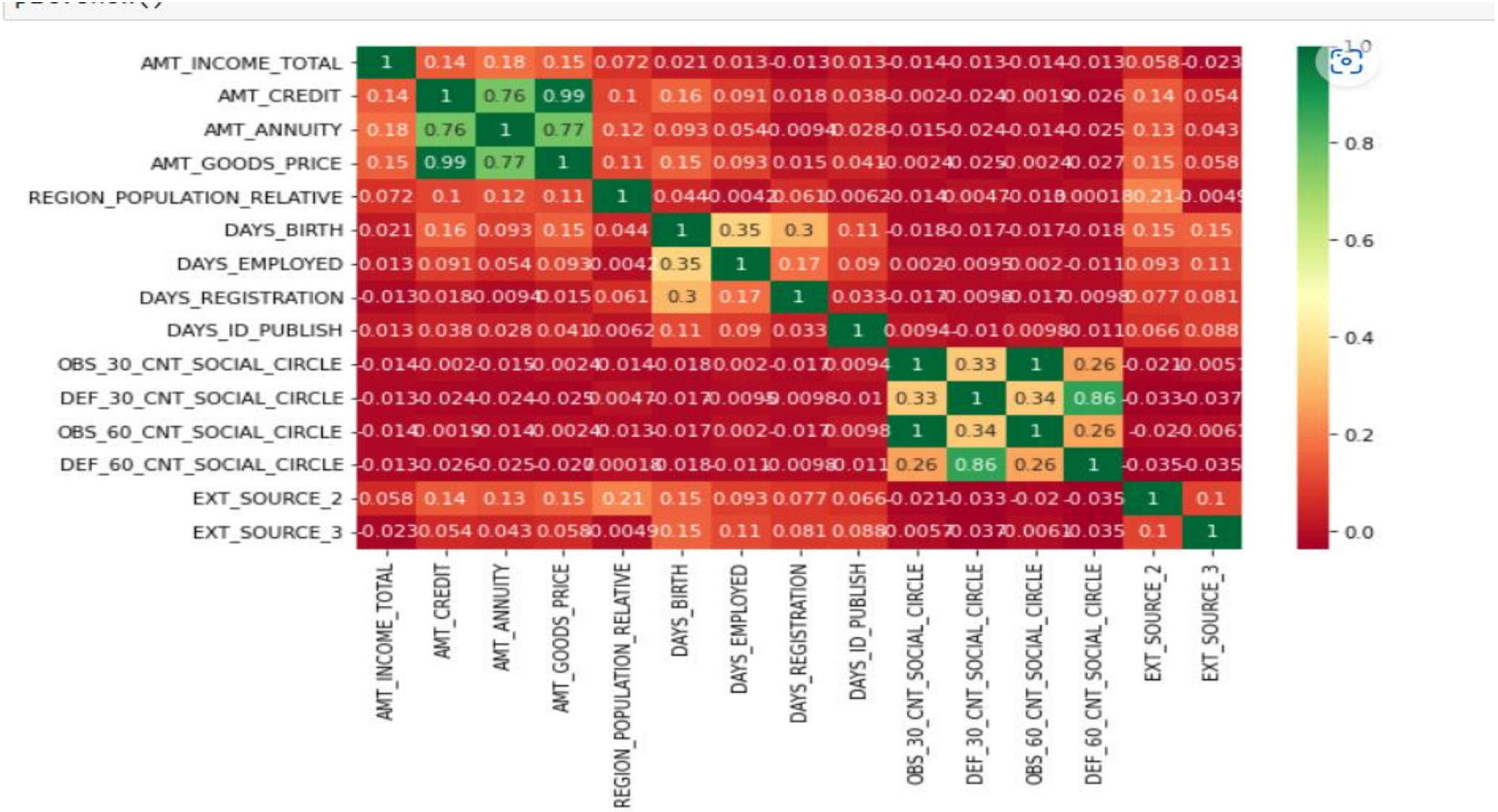
Scatterplot EXT_SOURCE_3 Vs EXT_SOURCE_2



Scatterplot AMT_GOODS_PRICE Vs DAYS_EMPLOYED



Performed Heatmap over application data-



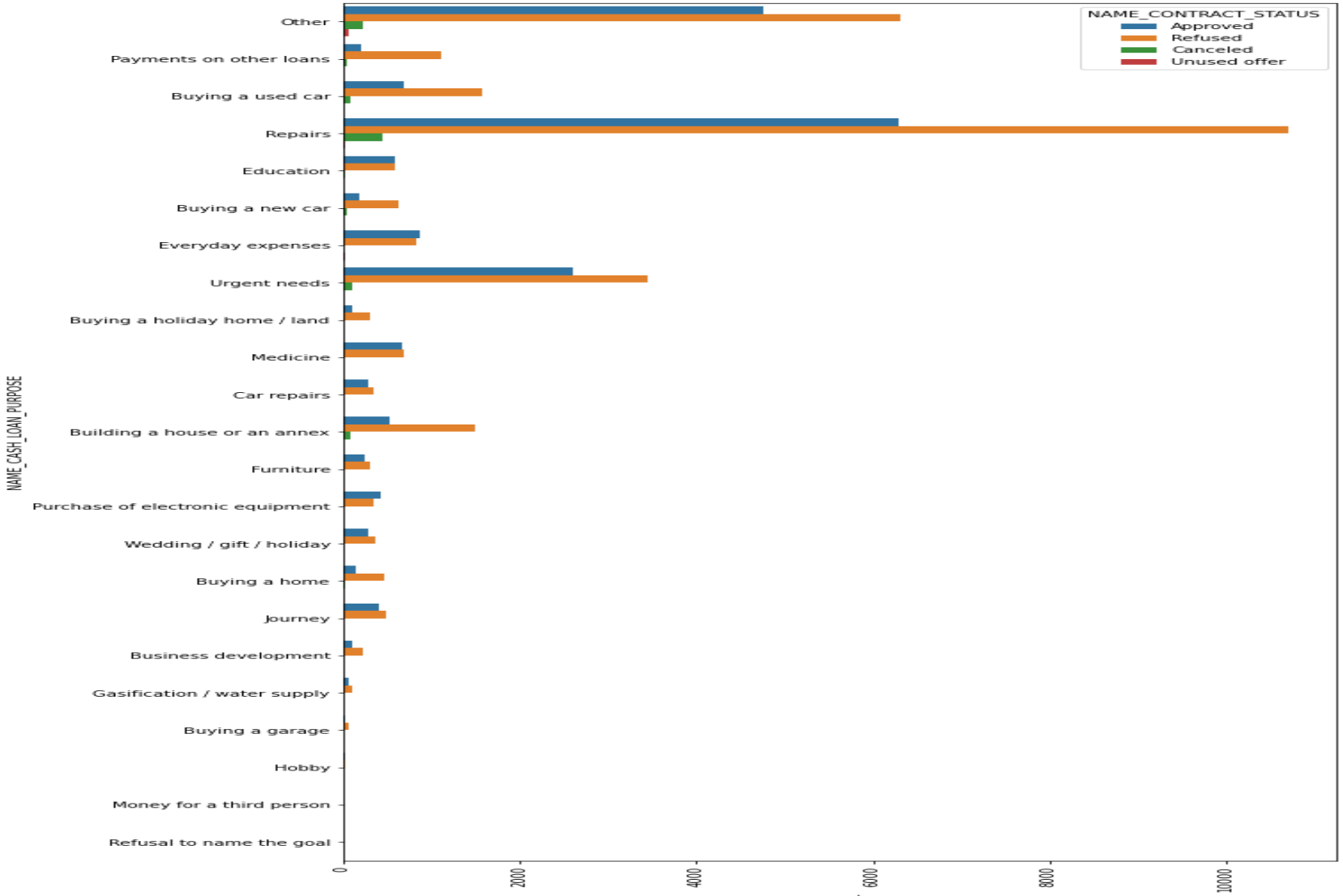
Highly Correlated columns-

- AMT_CREDIT and AMT_ANNUITY (0.74)
- AMT_CREDIT and AMT_GOODS_PRICE (0.99)
- AMT_ANNUITY and AMT_GOODS_PRICE (0.77)

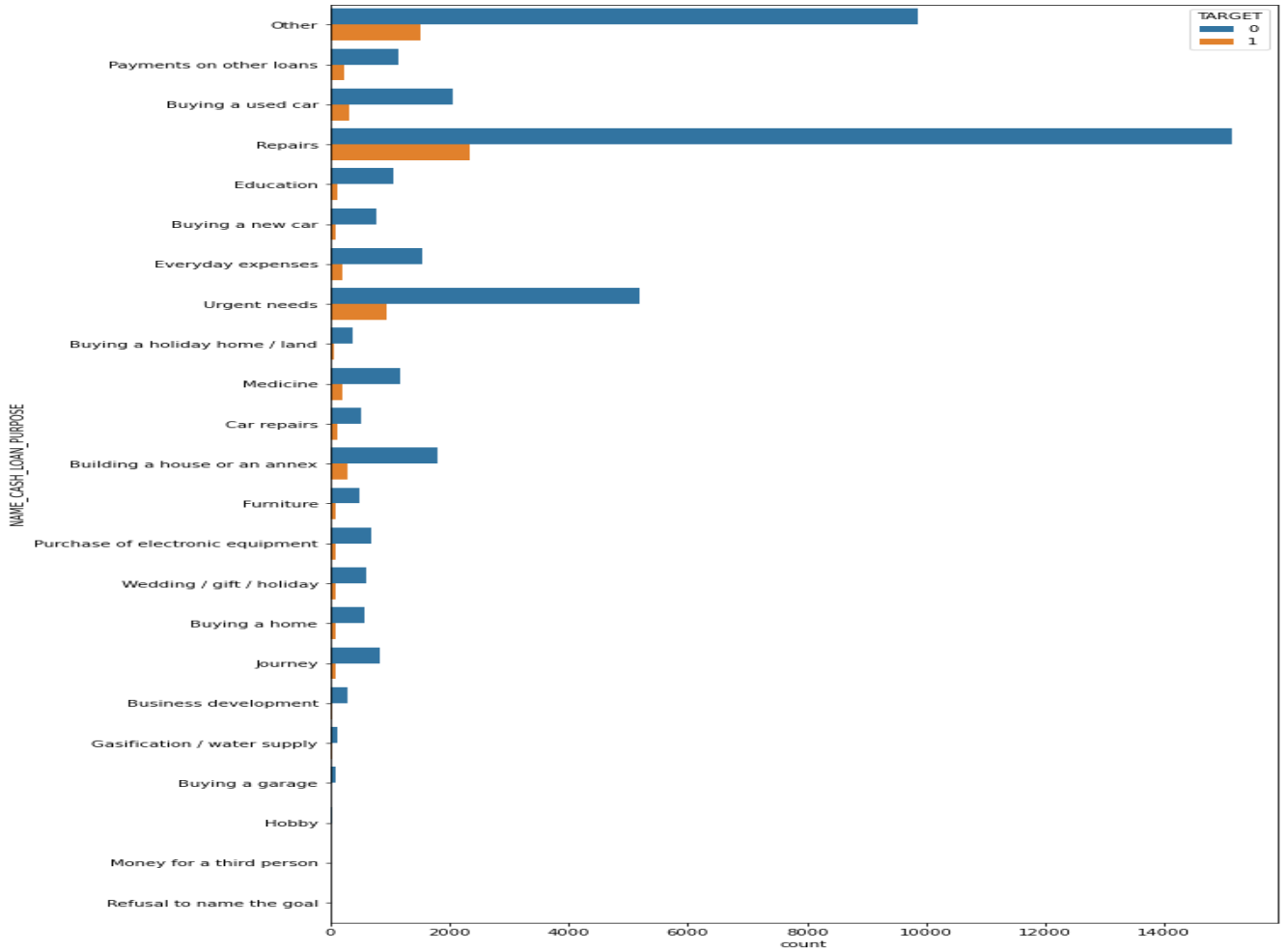
Merged the data-

- After cleaning and analysing application data and previous application data, merget both data set.
- In final data, remove unwanted columns, then perform univariate and bivariate analysis over them.

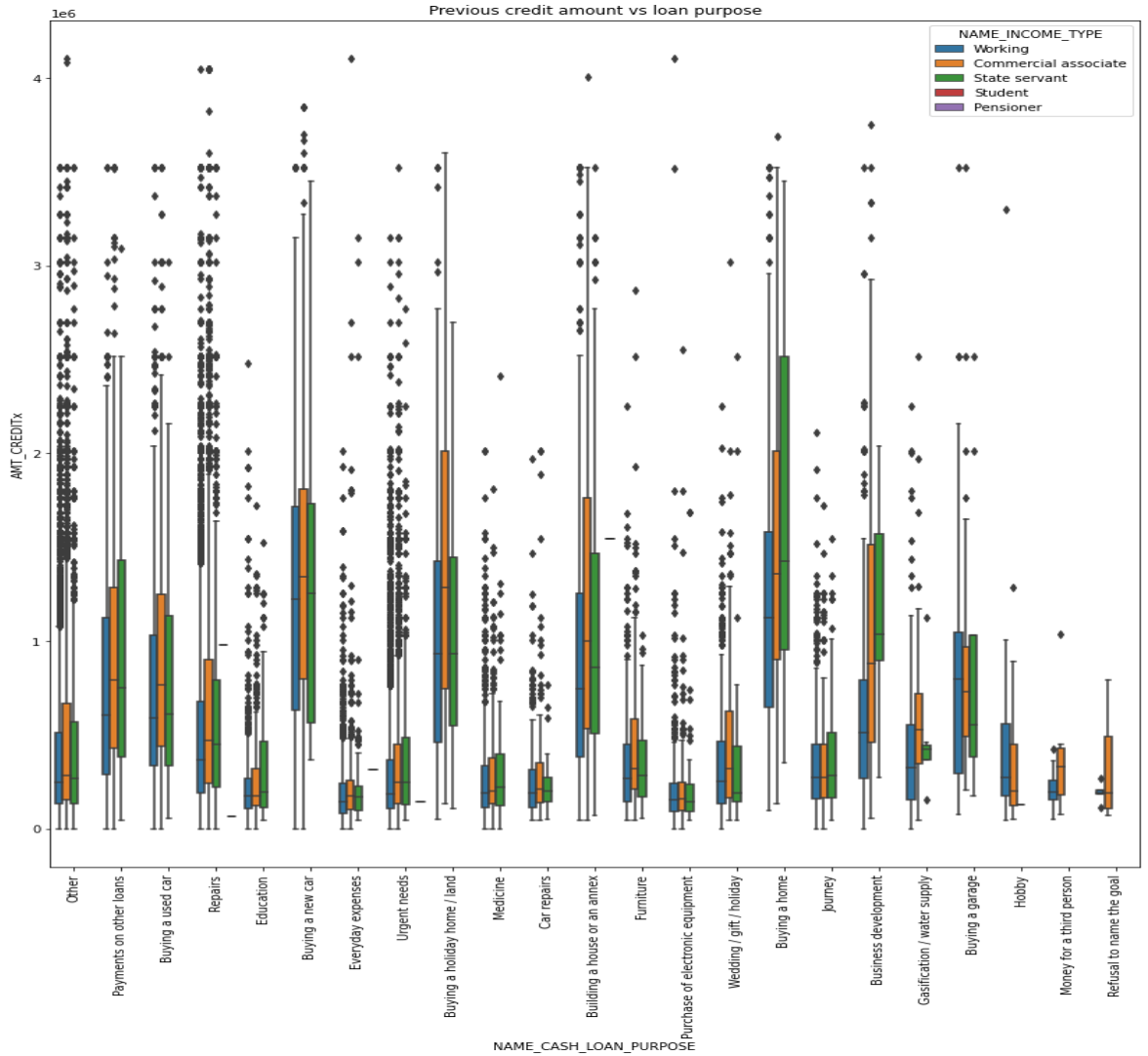
Univariate anlysis over contract status-



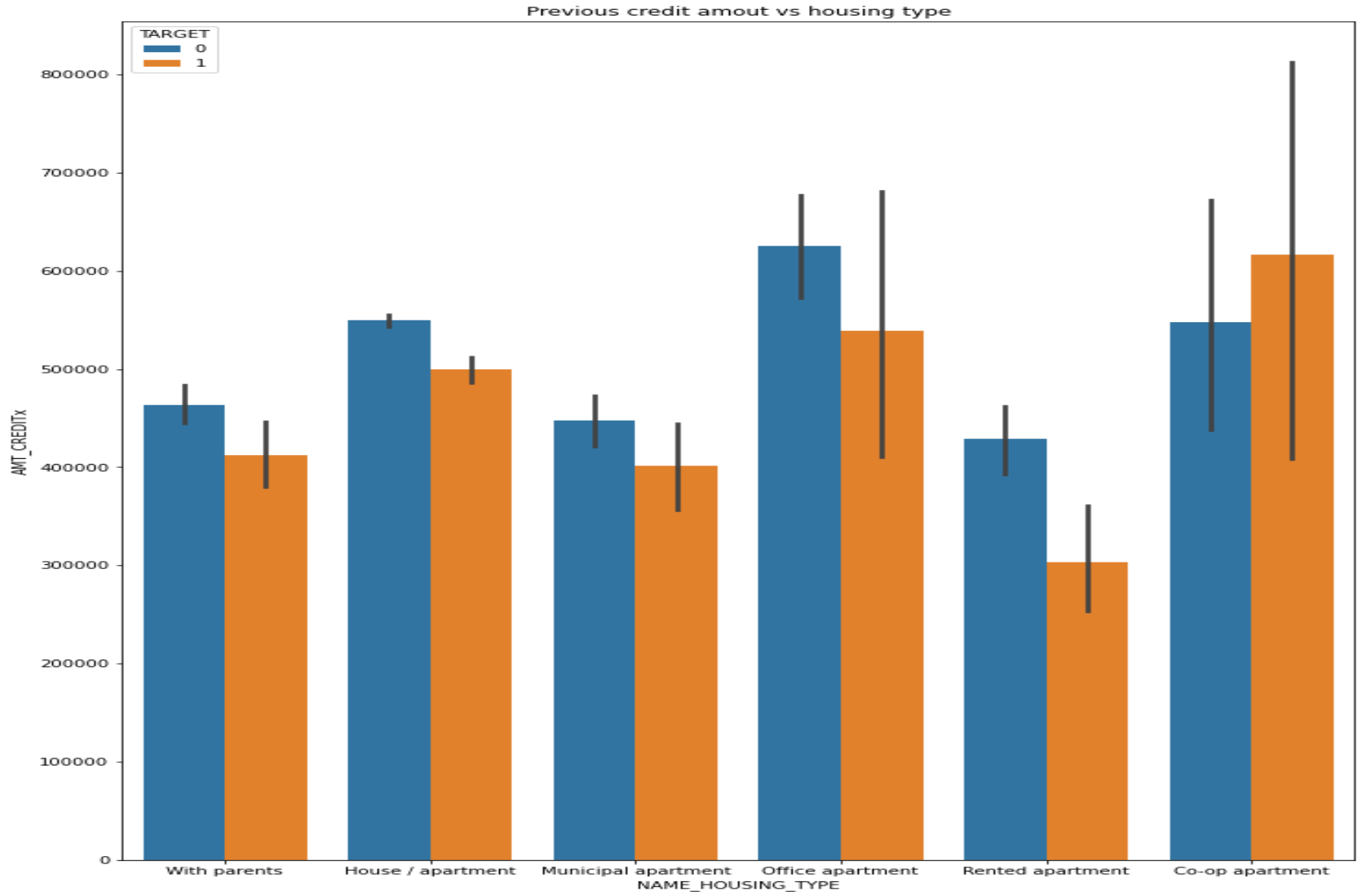
Univariate analysis over TARGET-



Bivariate analysis over income type -



Bar plotting for previous credit amount vs housing type-



On time payments-

- DEF_30_CNT_SOCIAL_CIRCLE- DEF_60_CNT_SOCIAL_CIRCLE 0.86
- AMT GOODS PRICE AMT_CREDIT 0.99
- AMT ANNUITY -AMT GOODS PRICE 0.77
- AMT ANNUITY -AMT CREDIT 0.77

Conclusion-

- *Bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment.*
- *Bank can focus on housing type with parents or House\appartment or miuncipal appartment for successful payments*.
- *Also with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.*