Credit EDA Case Study

Problem Statement:

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.

Data Exploration

- This case study contains following files.
    1. application_data.csv.
    2. previous_application.csv.
    3. columns_description.csv.
- Application data contains all the information about the client at the time of application. The data is about whether the **client has payment difficulties.**
- Previous application contains information about client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
- Columns description file explains about the meaning of the variable.
- Based on the problem statement we have to concentrate on Target variable column with other columns from the application_data.csv.
- Shape of the application data (307511,122).

Data Cleaning and Manipulation

- From the observation of application data file shape there are 122 columns and 307511 rows from this we can clean un necessary data to not affect the analysis.

Data Cleaning:

- Find sum of null values and their percentage for all the columns in application data.
- Remove columns which the null value percentage is greater than 45.
- Check for duplicate rows and drop duplicates.

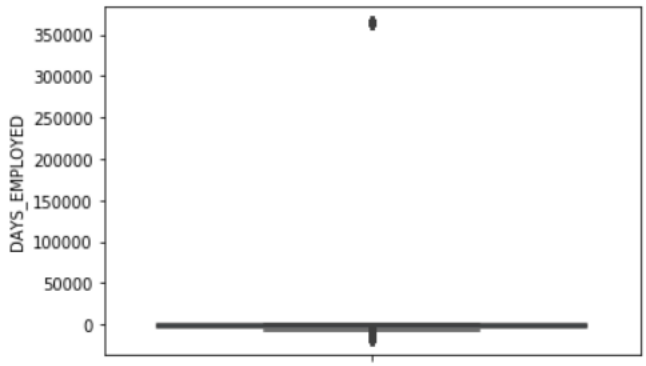
Manipulation:

- In application file the DAYS_BIRTH column has in-proper format of days count, so make another column as Applicant Age with proper format.

Identify Outliers

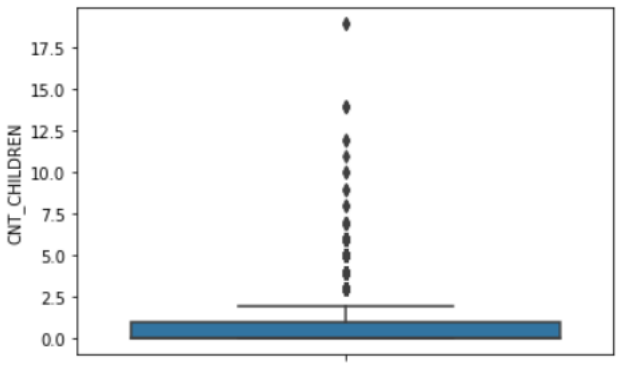
- Identity outliers in the application data columns
- Don’t remove the outliers but just to find if there are any outliers in the dataset.
- Why we call the Outlier? An outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error.

- **DAYS_EMPLOYED column:**



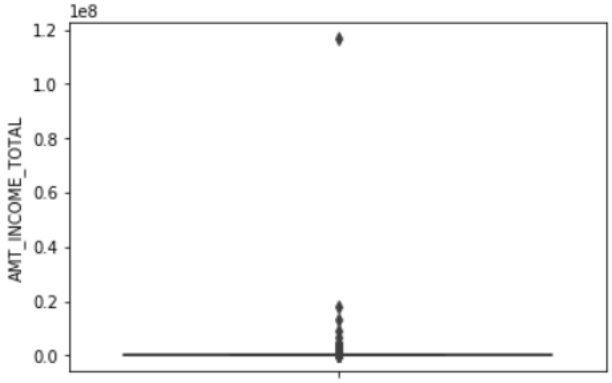
- If we observe the above graph for days employed column there is an outlier where the value is out of range compared to the values in same column.
- The outlier value is 365243 for DAYS_EMPLOYED column.

- **CNT_CHILDREN column:**



- The outlier value is 19.0 for the CNT_CHILDREN column.

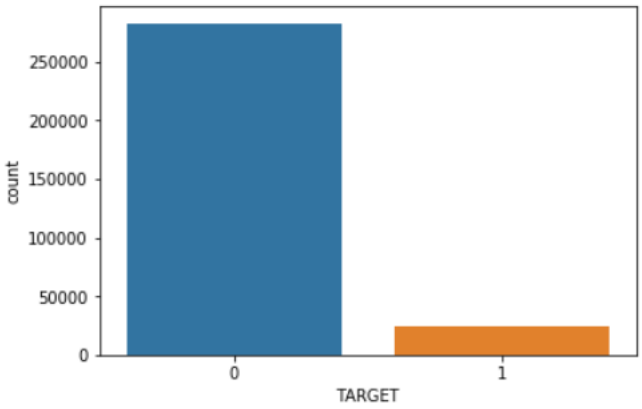
- **AMT_INCOME_TOTAL column:**



- Outlier value from AMT_INCOME_TOTAL is 117000000.

Data Imbalance

- As described in the problem statement to find the analysis for target variable in application data file.
- Find the imbalance ratio for the Target variable.
- The Target Variable contains:
    1. Target 0 for All other cases.
    2. Target 1 for Client with payment difficulties.

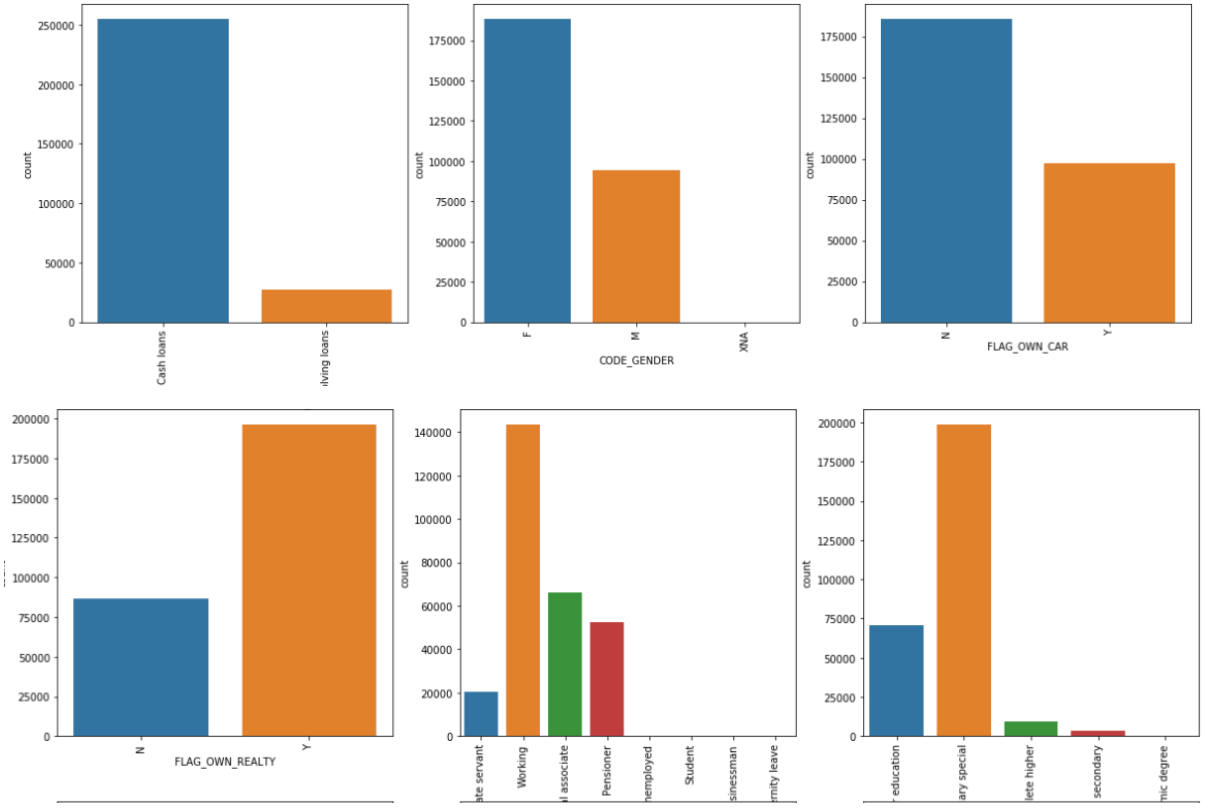


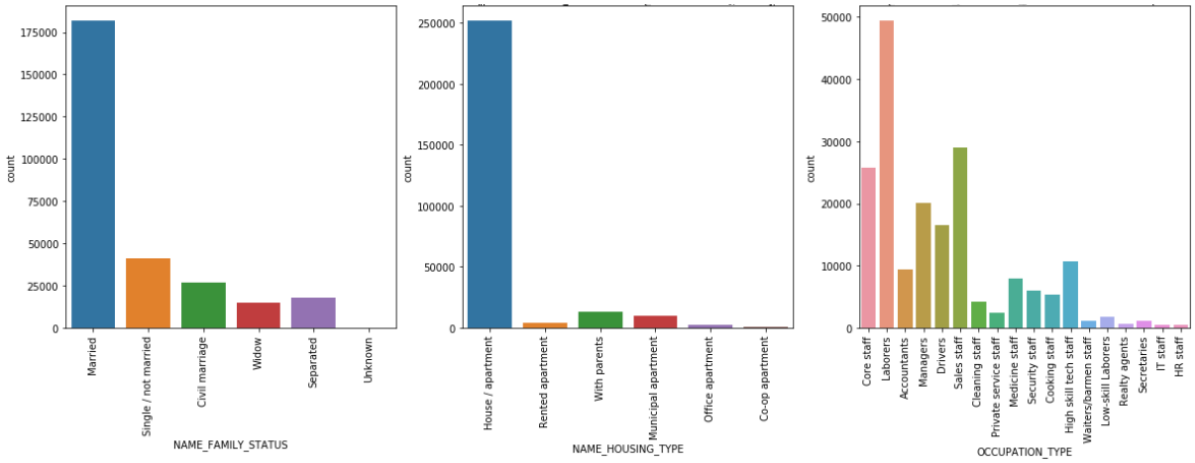
- If we observe the above graph there is data imbalance for the target variable.
- The Imbalance ratio is:
    1. Percentage of All other cases we get 91.93%.
    2. Percentage of Client with payment difficulties 8.07%
- From this we can say that client with payment difficulties has far less percentage when compare to all other cases.

- To find the analysis between **Client with payment difficulties** and **All other cases.** We need to divide data into two parts as target 0 and target 1.
- Then perform analysis for Univariate, Bi-Variate, Segmented Univariate analysis for both client with payment difficulties and all other cases for understanding the data.
- Select only few columns which are important for the valuable insights.
- ['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'DAYS_BIRTH', 'CNT_FAM_MEMBERS','DAYS_EMPLOYED', 'OCCUPATION_TYPE'].

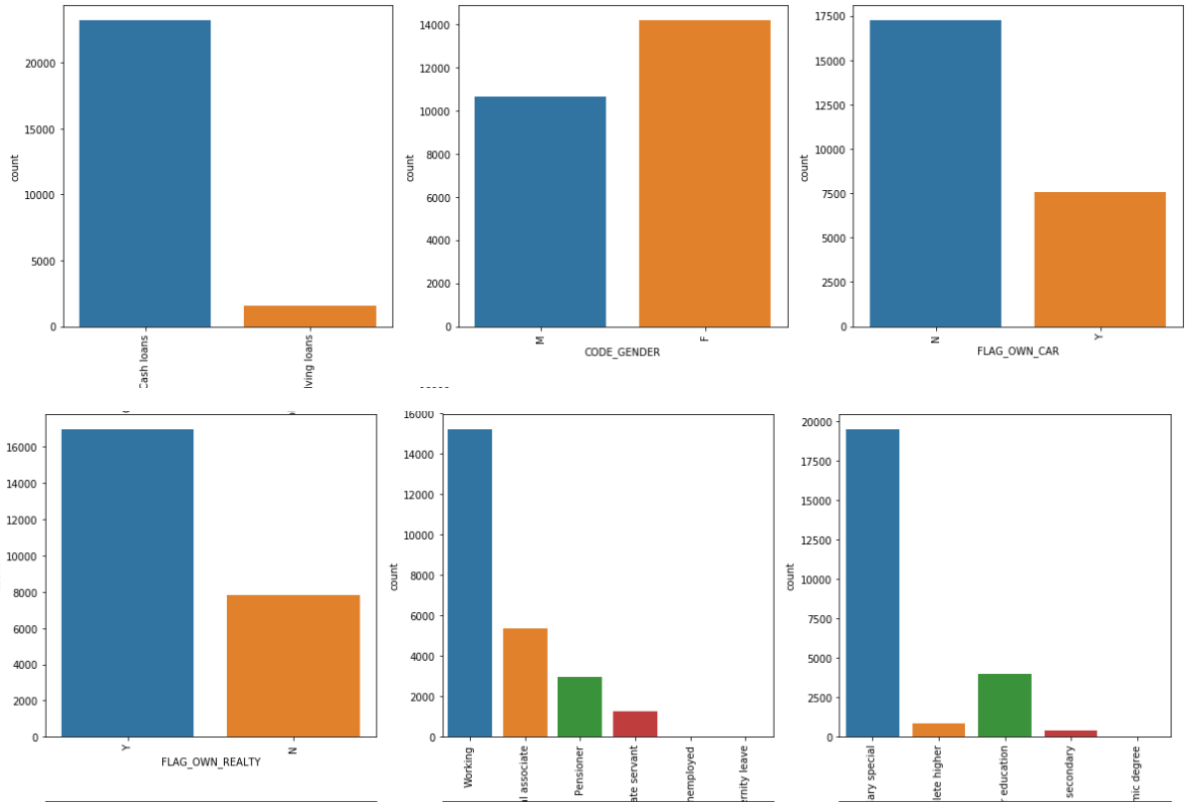
Univariate Analysis

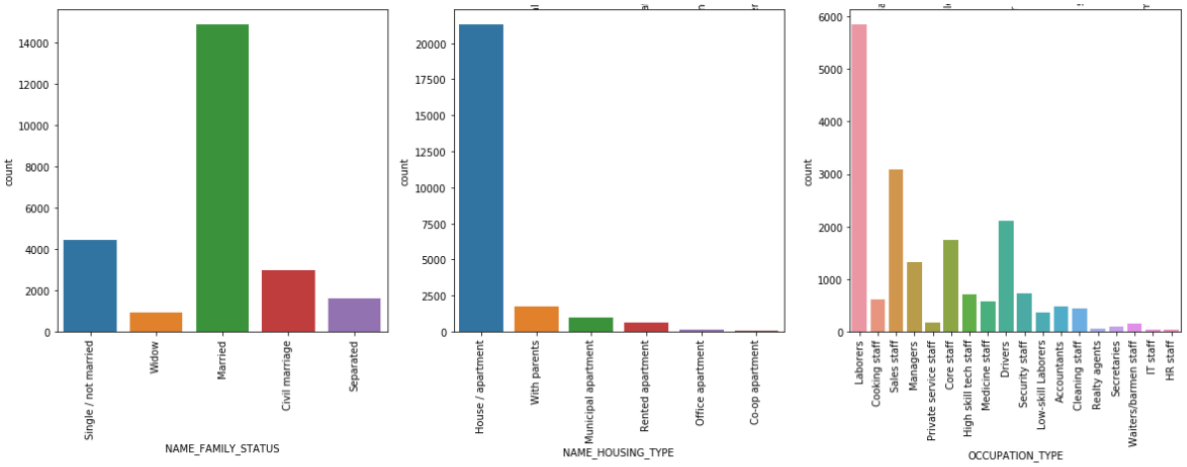
- **Univariate analysis for Target variable 0 (All other cases**).
- We observe that Target variable All other cases is having highest in Cash loans (CONTRACT_TYPE) and so on.





- **Univariate Analysis for Target Variable 1 (Clients with payment difficulties).**
- From the target variable 1 we can observe that clients with payment difficulties is higher in Females (CODE_GENDER), Not Having Reality (FLAG_OWN_REALITY), and Married clients (NAME_FAMILY_STATUS) are more etc.



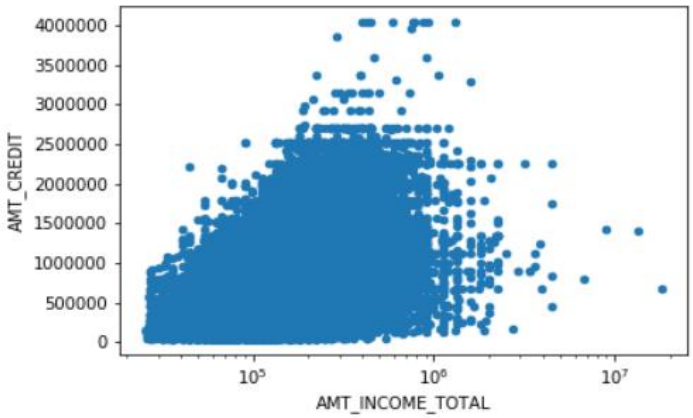


Bi-Variate Analysis

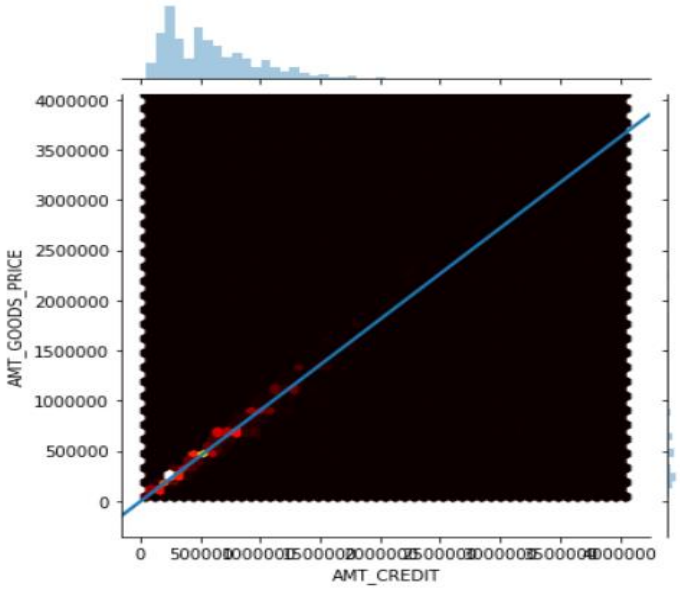
- Bi-Variate analysis is used to find the relationship between two sets of values.
- Let us find if there is any relationship between variables in application data.

- **Bi-Variate analysis for Target variable 0 (All Other cases):**

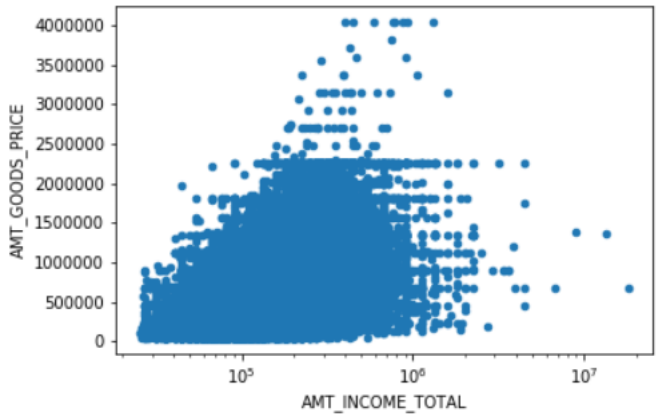
1. Bi-Variate Analysis between AMT_INCOME_TOTAL and AMT_CREDIT



2. Bi-Variate Analysis between AMT_CREDIT and AMT_GOODS_PRICE

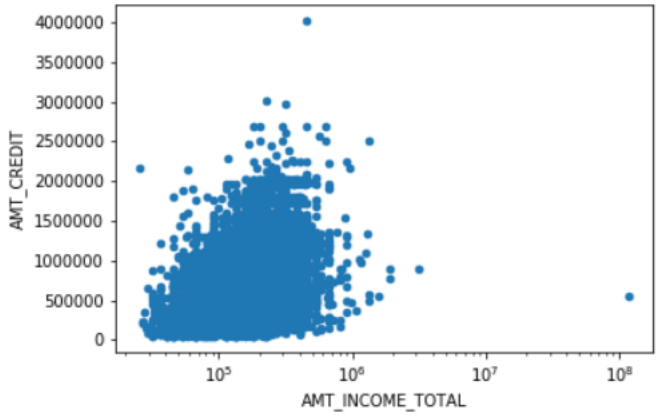


3. Bi-Variate Analysis between AMT_GOODS_PRICE and AMT_INCOME_TOTAL

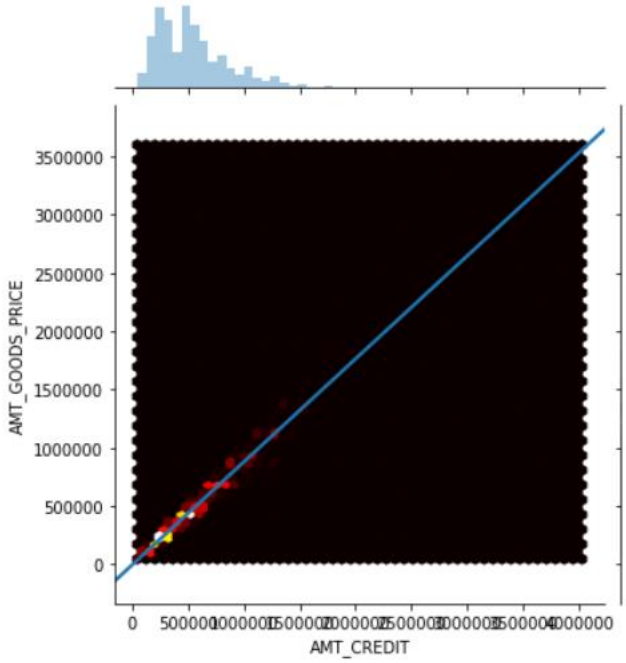


- **Bi-Variate analysis for Target variable 1 (Clients with payment difficulties):**

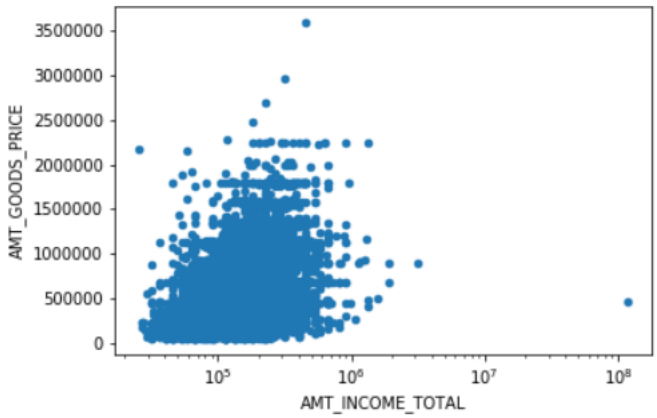
4. Bi-Variate analysis for AMT_INCOME_TOTAL and AMT_CREDIT.



5. Bi-variate analysis for AMT_CREDIT and AMT_GOODS_PRICE



6. Bi-Variate analysis for AMT_INCOME_TOTAL and AMT_GOODS_PRICE.

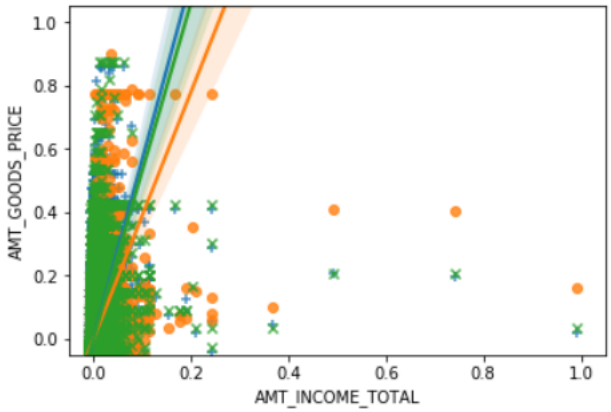


Multi-Variate Analysis

- Multi-variate analysis is used to find the relationship between multiple variables.

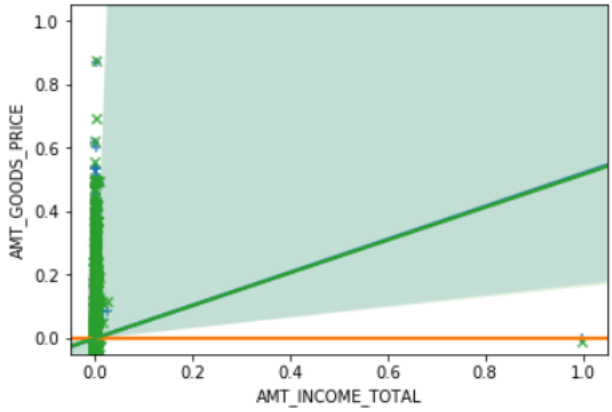
- **Multi-Variate analysis for Target Variable 0**:

- Analysis between AMT_ANNUITY, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE
- Where “+” is between AMT_INCOME_TOTAL and AMT_CREDIT
- “o” is between AMT_INCOME_TOTAL and AMT_ANNUITY
- “x” is between AMT_INCOME_TOTAL and AMT_GOODS_PRICE



- **Multi-Variate analysis for Target Variable 1**:

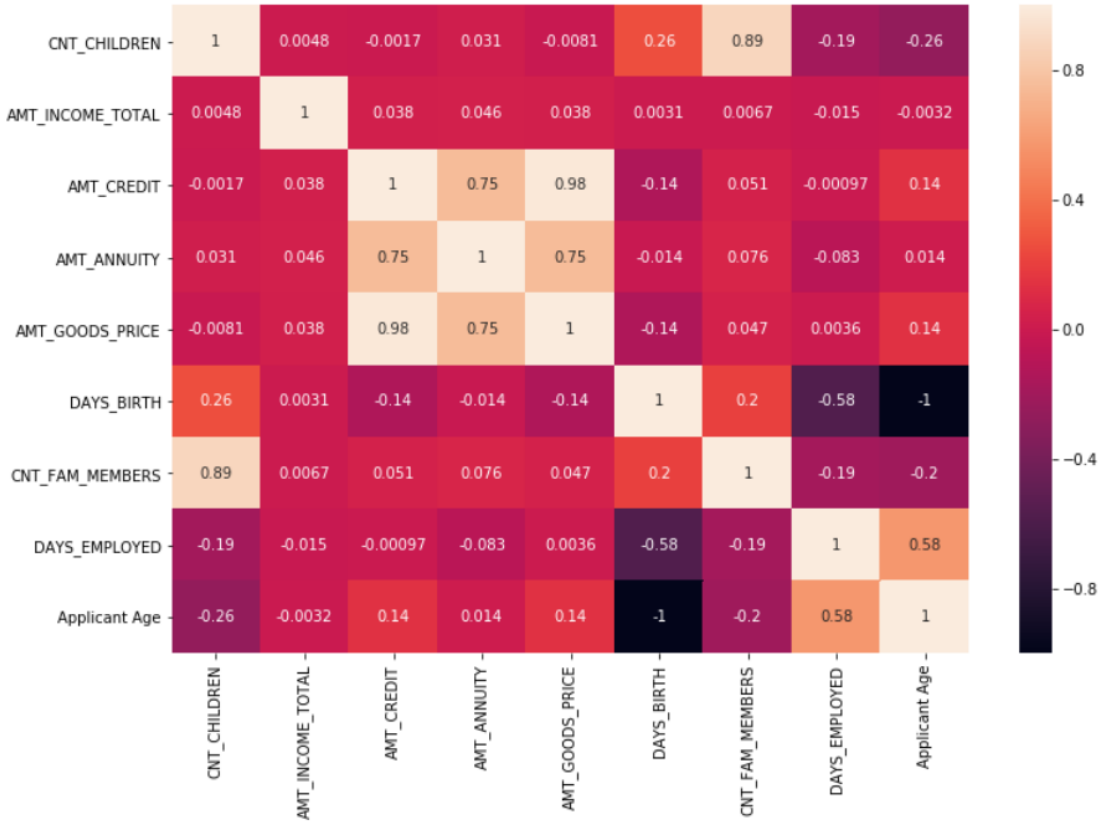
- Analysis between AMT_ANNUITY, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE
- Where “+” is between AMT_INCOME_TOTAL and AMT_CREDIT
- “o” is between AMT_INCOME_TOTAL and AMT_ANNUITY
- “x” is between AMT_INCOME_TOTAL and AMT_GOODS_PRICE



Finding Correlation

- **Correlation for Target Variable 0**:
- From this heat map we observe that AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE are correlated.

- **Correlation for Target Variable 1**:
- From this heat map we observe that AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, are correlated.



- From the Target variable 0 and 1 heat maps we can clearly observe that the variables AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, are correlated in both target variable 0 and 1 heat maps.

Merging the Data

- To view the insights between current application and previous application status we need to merge the application_data.csv and previous_application.csv
- Joining these two files based on the common id SK_ID_CURR.
- Shape of the Merged data is (1413701,109).

Data Cleaning and Manipulating

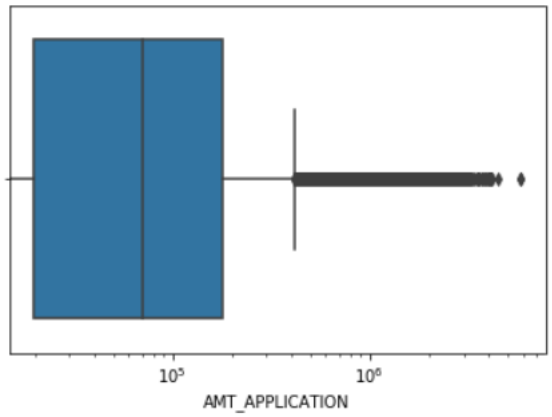
- **Data Cleaning**:
    I. Find sum of null values and their percentage for all the columns in application data.
    II. Remove columns which the null value percentage is greater than 45.
    III. Check for duplicate rows and drop them.

Identify Outliers

- Identity outliers in the merged data columns.

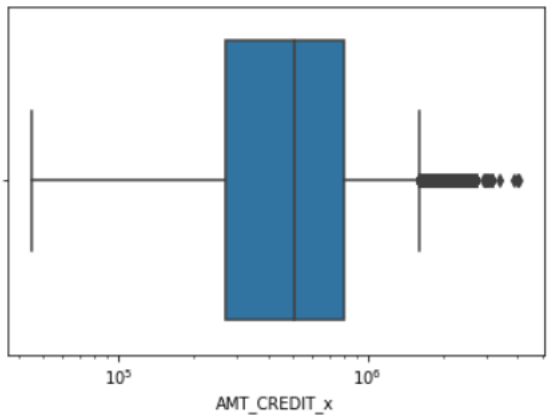
- **AMT_APPLICATION column**:

  From the below graph, we can clearly observe a, outlier point in AMT_APPLICATION column the outlier value is 5850000



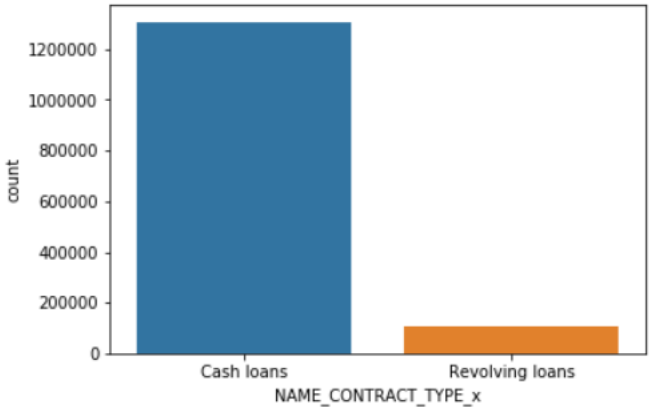
- **AMT_CREDIT column**:

  From the below graph, we can observe there is a outlier point in AMT_CREDIT column and the outlier value is 4050000

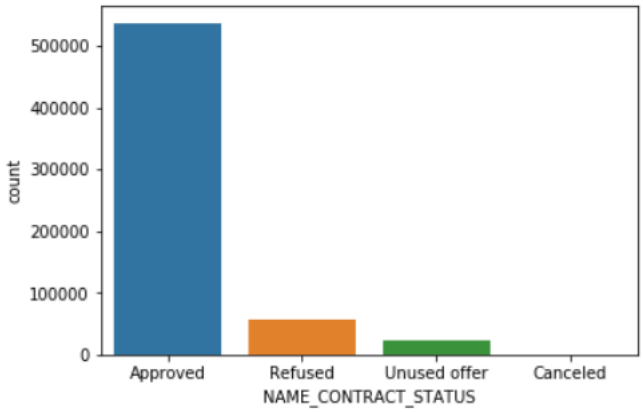


Uni-Variate Analysis

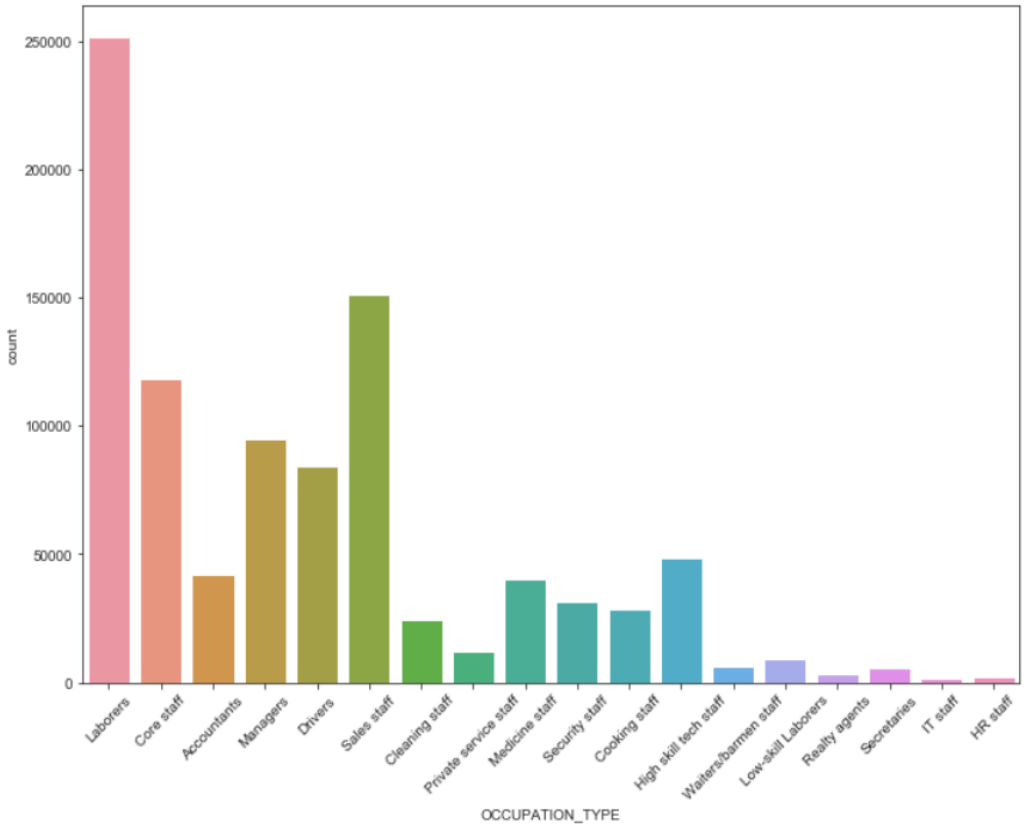
- Univariate analysis for CONTRACT_TYPE column.



- Univariate analysis for CONTRACT_STATUS column.
- From the below analysis we observe that most of the loans are approved by the bank.

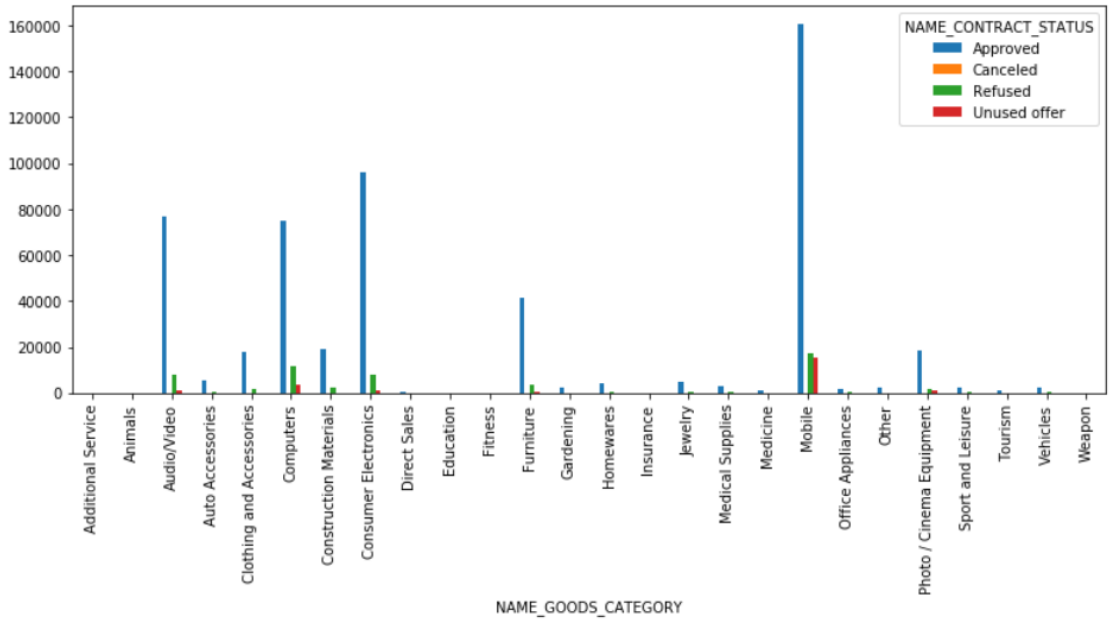


- Univariate analysis for OCCUPATION_TYPE column.
- From the below analysis occupation type of Laborers are high in applying for the loan.

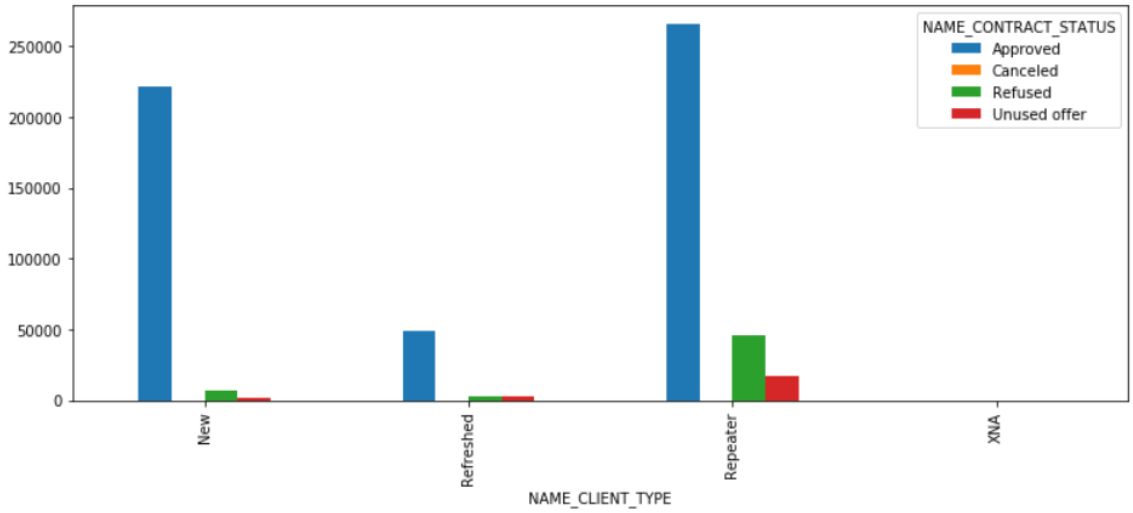


Bi-Variate Analysis

- Finding some insights for NAME_GOODS_CATEGORY and NAME_CONTRACT_STATUS.
- From the below graph, we can say that Mobile category has most approved status and next to that consumer electronics.



- Analysing which client type and contract status has most approved.
- We can see that Repeater has more approved status. A repeater will get loan easily and followed by that New customer will also get approved easily.



Summarizing the Analysis

1. Younger applicants [Age between 20-30] are more likely to not repay the loan! The rate of failure to repay is above 10% for the youngest three age groups and below 5% for the oldest age group [60-70].
2. This above information that could be directly used by the bank: because younger clients are less likely to repay the loan, maybe they should be provided with more guidance or financial planning tips. This does not mean the bank should discriminate against younger clients, but it would be smart to take precautionary measures to help younger clients pay on time.
3. Males are less defaulters in terms of Females.
4. Married people have a greater number of defaults.
5. It is mostly better for Banks to approve loans for Mobile/Electronics as people tend to repay their loan in most of the cases.
6. People with real estate should be preferred as they tend to repay their loan and if at all they fail the bank can sell them to recover the money.
7. People with more work experience a have a good chance of repaying the loan and it also comes from the fact that more work experience means more age and we have seen that as clients get older, they tend to repay their loans on time more often.