Empowering Financial Inclusion: Addressing the Credit Gap through Innovative Lending Solutions

Project Objective:

Our goal with EDA is to better understand the factors associated with bad credit. The analysis will help financial institutions make more decisions during loan approval and reduce the risk of crime and bankruptcy while ensuring that the loan is made to reliable applicants. Finally, the program aims to encourage accounting and credit transactions that are beneficial to comp anies and loan applicants.

Prepared by Pamal Mondal

The process includes

- Importing the data
- Check the structure of the data, datatypes of columns, shape of the data
- Missing value check
- Check outlier
- Perform Univariate Analysis
- Perform Bivariate Analysis
- Perform Multivariate Analysis
- Conclusion

Importing Libraries and Data

- **import numpy as np** – For structuring the data
- **import pandas as pd** - To work with datasets and cleaning
- **import seaborn as sns** – To make statistical graphics
- **import matplotlib.pyplot as plt** – For data visializations
- **import scipy.stats as ss** – To solve mathematical problems
- **import warnings** - To handle warning

- Here the raw data has been imported ass app_df (Application Dataframe)
- Here the previous data has been imported ass prev_df (Previous Dataframe)

- Here I have restricted the number of rows and columns while viewing to make work environment friendly.

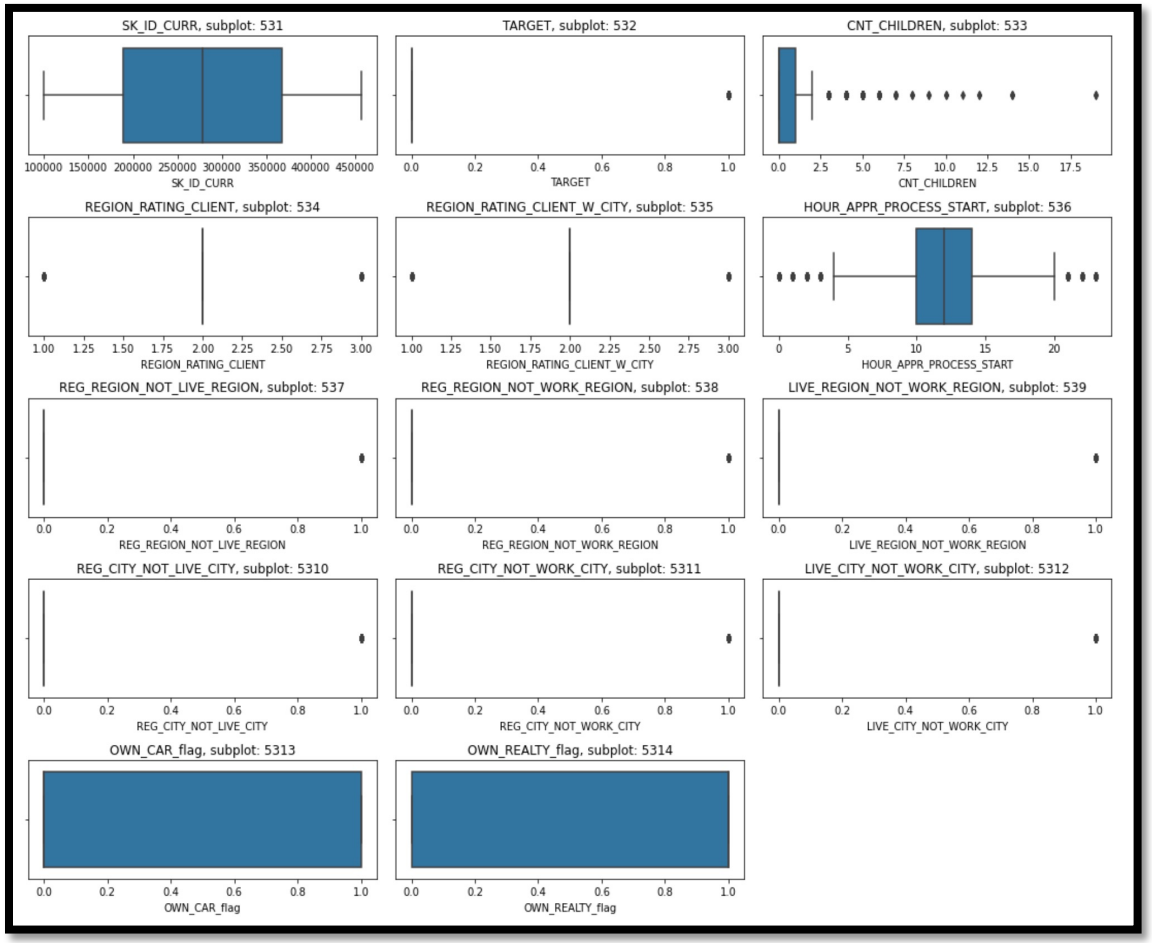
Checking of the data

- Once the data has been imported, the first step in the Exploratory Data Analysis (EDA) process is to ensure that the data has been imported correctly and examine its basic characteristics. We will perform the following checks:
    1. **Check Head and Tail**: By inspecting the head and tail of the data, we can quickly assess the structure and content of the dataset.
    2. **Check DataFrame Shape**: This step involves verifying the dimensions of the DataFrame, i.e., the number of rows and columns, which helps us understand the dataset's size.
    3. **Check Data Types**: We will examine the data types of individual columns to ensure they have been appropriately interpreted during the import process.
- These preliminary checks are crucial for getting an initial understanding of the data and identifying any potential issues or anomalies that may require further investigation.
- After the initial checks, we can proceed with the actual EDA, including data cleaning, visualization, feature engineering, and correlation analysis, as described in the previous short note.
- By conducting a comprehensive EDA, we aim to gain valuable insights that will help in making informed decisions and developing predictive models to identify loan default tendencies and minimize risks for the consumer finance company.

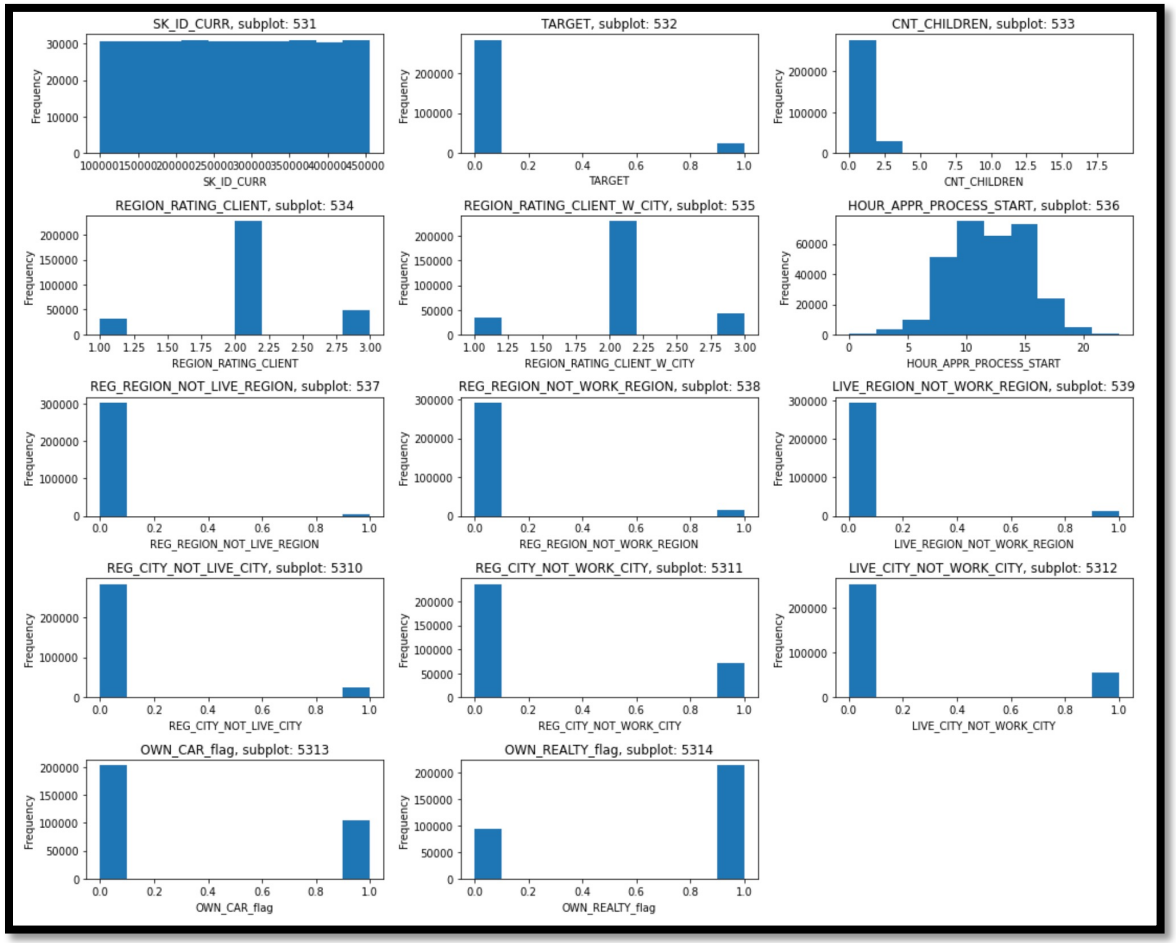
Data Handling & Checking

- **The first few steps involve making sure that there are no missing values or incorrect data types before we proceed to the analysis stage. These aforementioned problems are handled as follows:**
- For Missing Values: Some common techniques to treat this issue are
  - Dropping the rows containing the missing values
  - Imputing the missing values
  - Keep the missing values if they don't affect the analysis
- Incorrect Data Types:
  - Clean certain values
  - Clean and convert an entire column

Data Correction



Box Plot Analysis



Histogram Analysis

Outliers

- Outliers observed in 'AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE','YEARS_EMPLOYED','YEARS_REGISTRATION','OWN_CAR_AGE', DAYS_LAST_PHONE_CHANGE

```
count    3.075110e+05
mean     1.687979e+05
std      2.371231e+05
min      2.565000e+04
25%      1.125000e+05
50%      1.471500e+05
75%      2.025000e+05
max      1.170000e+08
Name: AMT_INCOME_TOTAL, dtype: float64
```

```
135000.0    0.116256
112500.0    0.100871
157500.0    0.086358
180000.0    0.080384
90000.0     0.073113
              ...
117324.0    0.000003
64584.0     0.000003
142897.5    0.000003
109170.0    0.000003
113062.5    0.000003
Name: AMT_INCOME_TOTAL, Length: 2548, dtype: float64
```

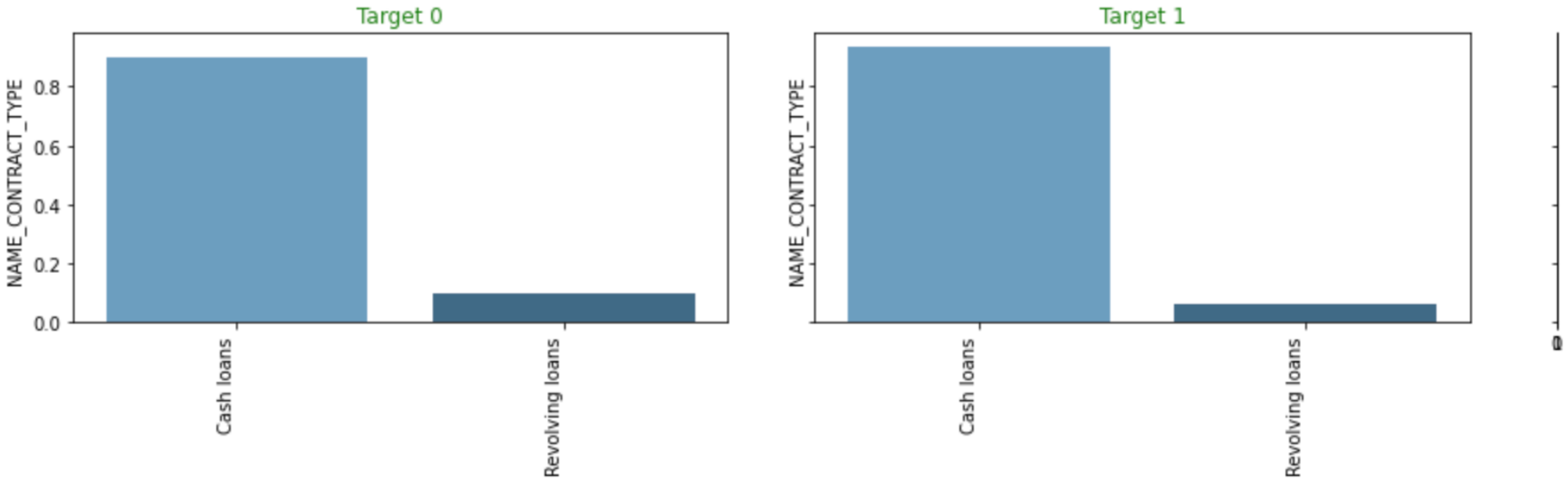
```
count    3.075110e+05
mean     5.990260e+05
std      4.024908e+05
min      4.500000e+04
25%      2.700000e+05
50%      5.135310e+05
75%      8.086500e+05
max      4.050000e+06
Name: AMT_CREDIT, dtype: float64
```

```
450000.0     0.031573
675000.0     0.028867
225000.0     0.026542
180000.0     0.023876
270000.0     0.023547
               ...
487318.5     0.000003
630400.5     0.000003
1875276.0    0.000003
1395895.5    0.000003
1391130.0    0.000003
Name: AMT_CREDIT, Length: 5603, dtype: float64
```

```
count    307499.000000
mean      27108.573909
std       14493.737315
min        1615.500000
25%       16524.000000
50%       24903.000000
75%       34596.000000
max      258025.500000
Name: AMT_ANNUITY, dtype: float64
```

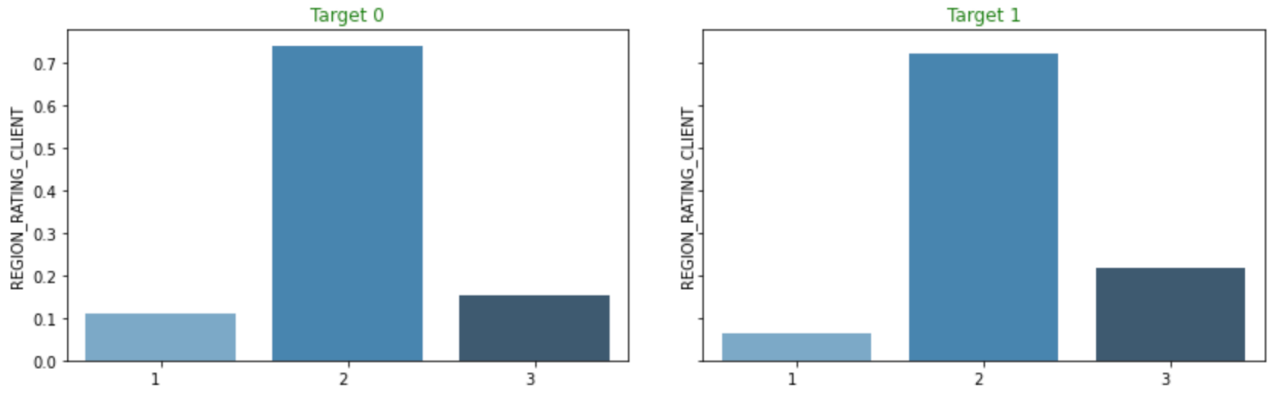
```
9000.0       0.020763
13500.0      0.017931
6750.0       0.007411
10125.0      0.006618
37800.0      0.005210
               ...
79902.0      0.000003
106969.5     0.000003
60885.0      0.000003
59661.0      0.000003
77809.5      0.000003
Name: AMT_ANNUITY, Length: 13673, dtype: float64
```

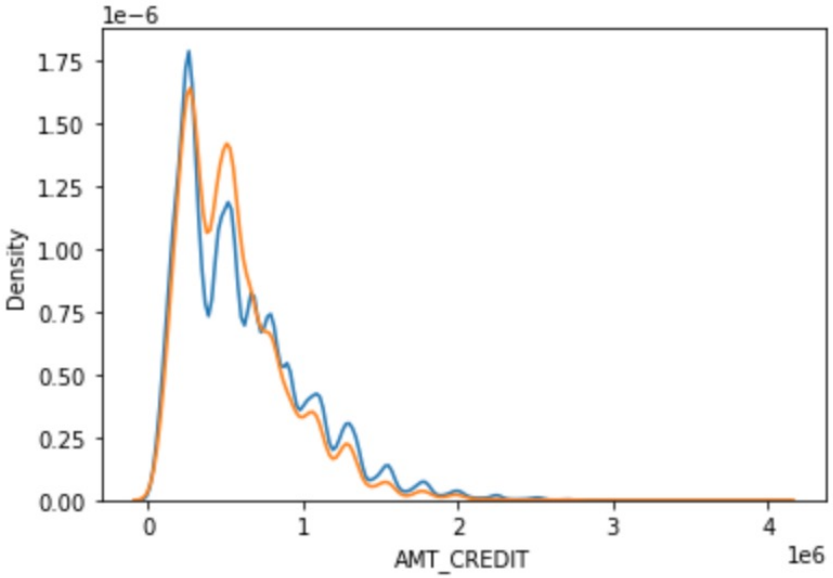
Analysis of Categorical Nominal Variables



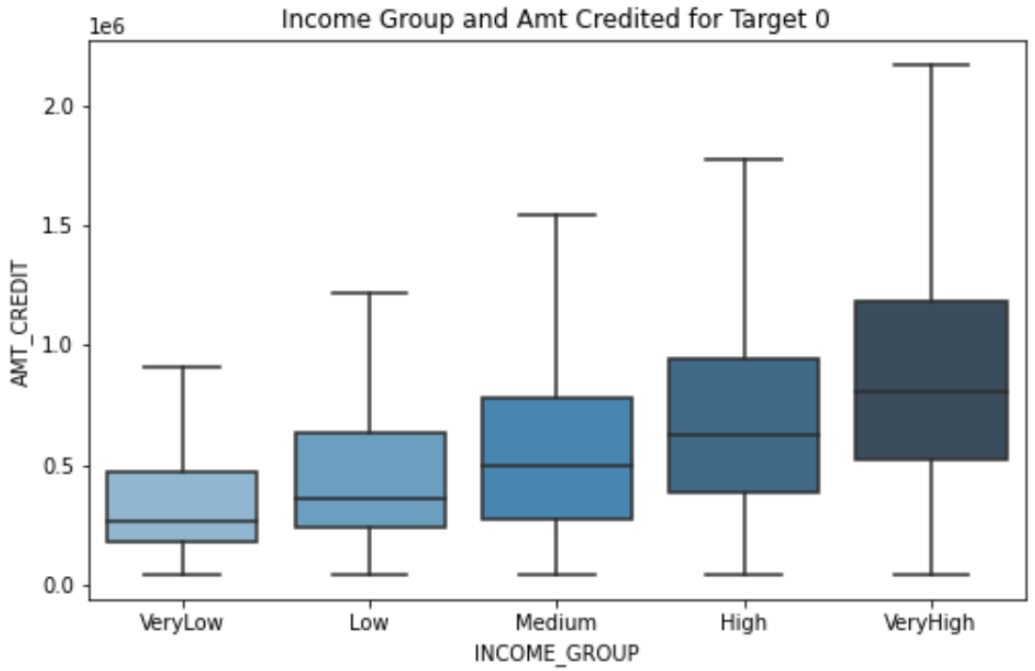
Univariate Analysis

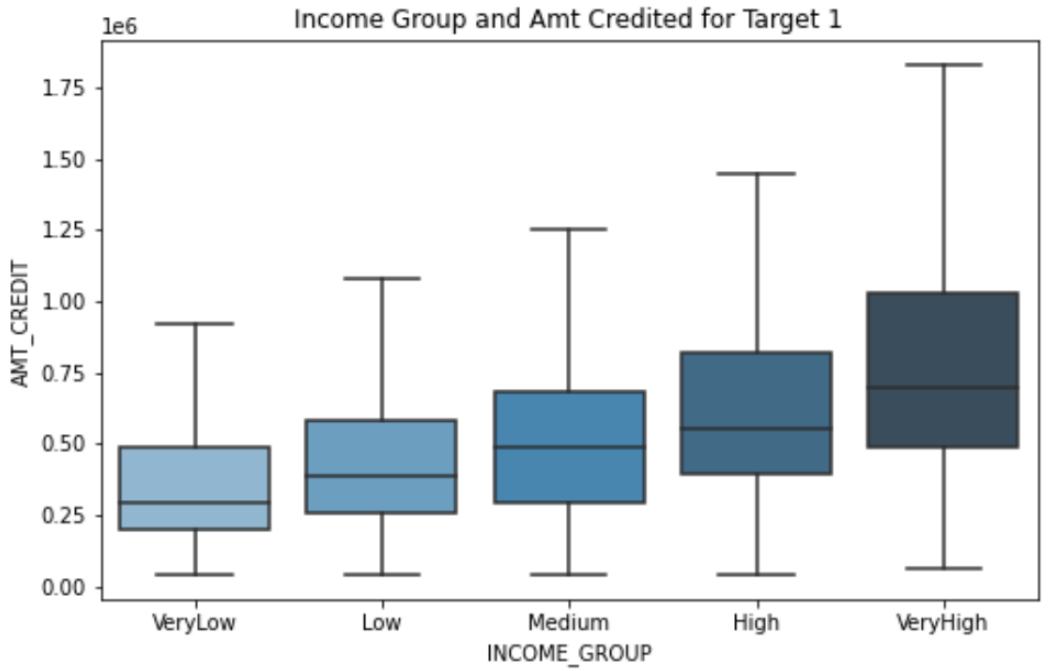
- Univariate Analysis on Categorical Ordered
- Univariate Analysis on Continuous Variables





Bivariate Analysis on Categorical and Continuous Variable





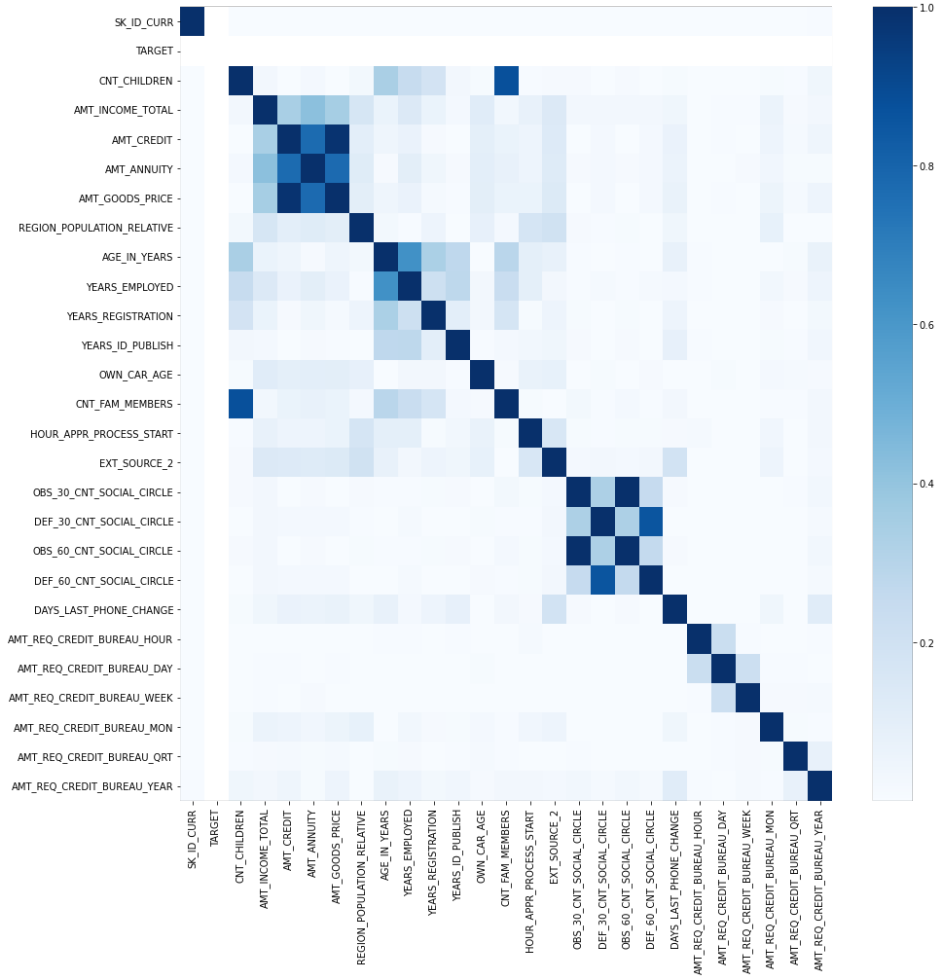
1. OBS_30_CNT_SOCIAL_CIRCLE',OBS_60_CNT_SOCIAL_CIRCLE' - denote the client's social surroundings with observable 30/60 DPD.

These are definetly correlated. We can also see that its higher and steeper for Target 1, signyfying that in approval process this parameter must be strongly looked into.

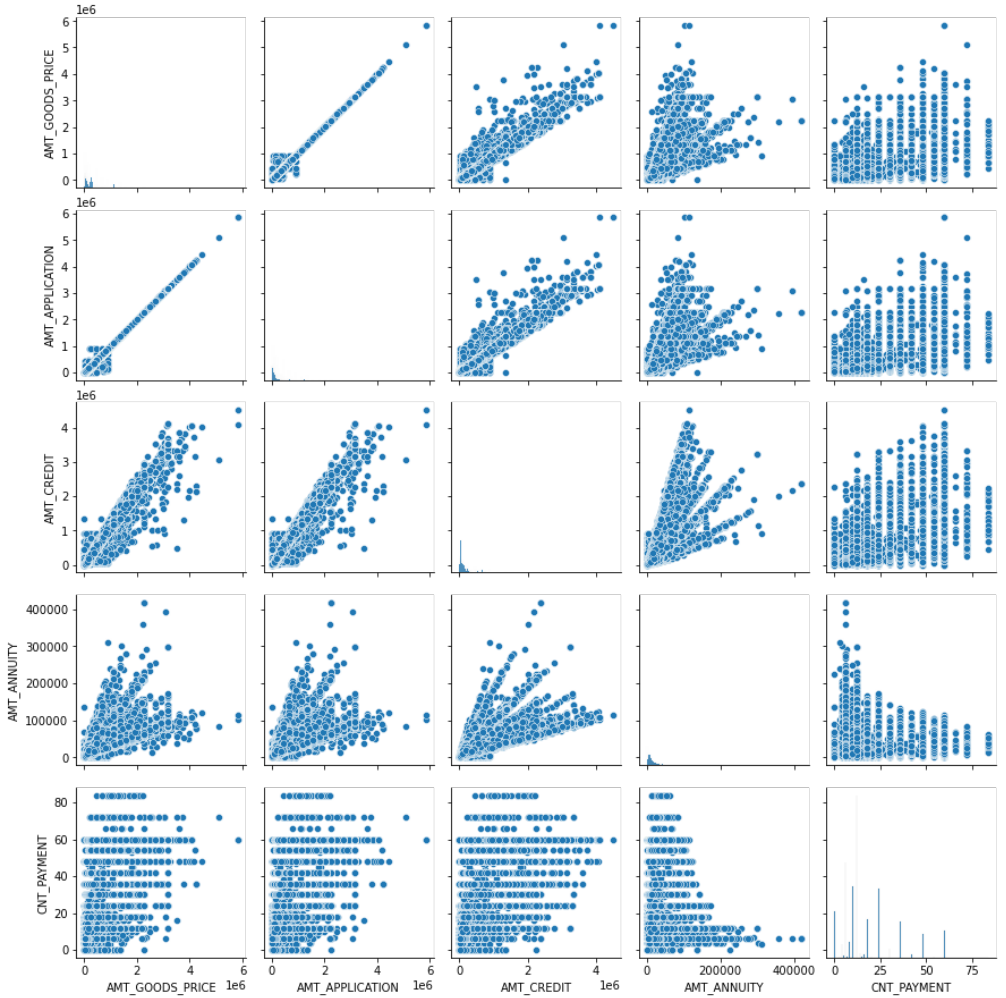
2.DEF_30_CNT_SOCIAL_CIRCLE - Trend is going up. But Target 1 has lesser data and hence graph is not dense.

3. Years employed has an outlier value of 999 and this is skewing the graph

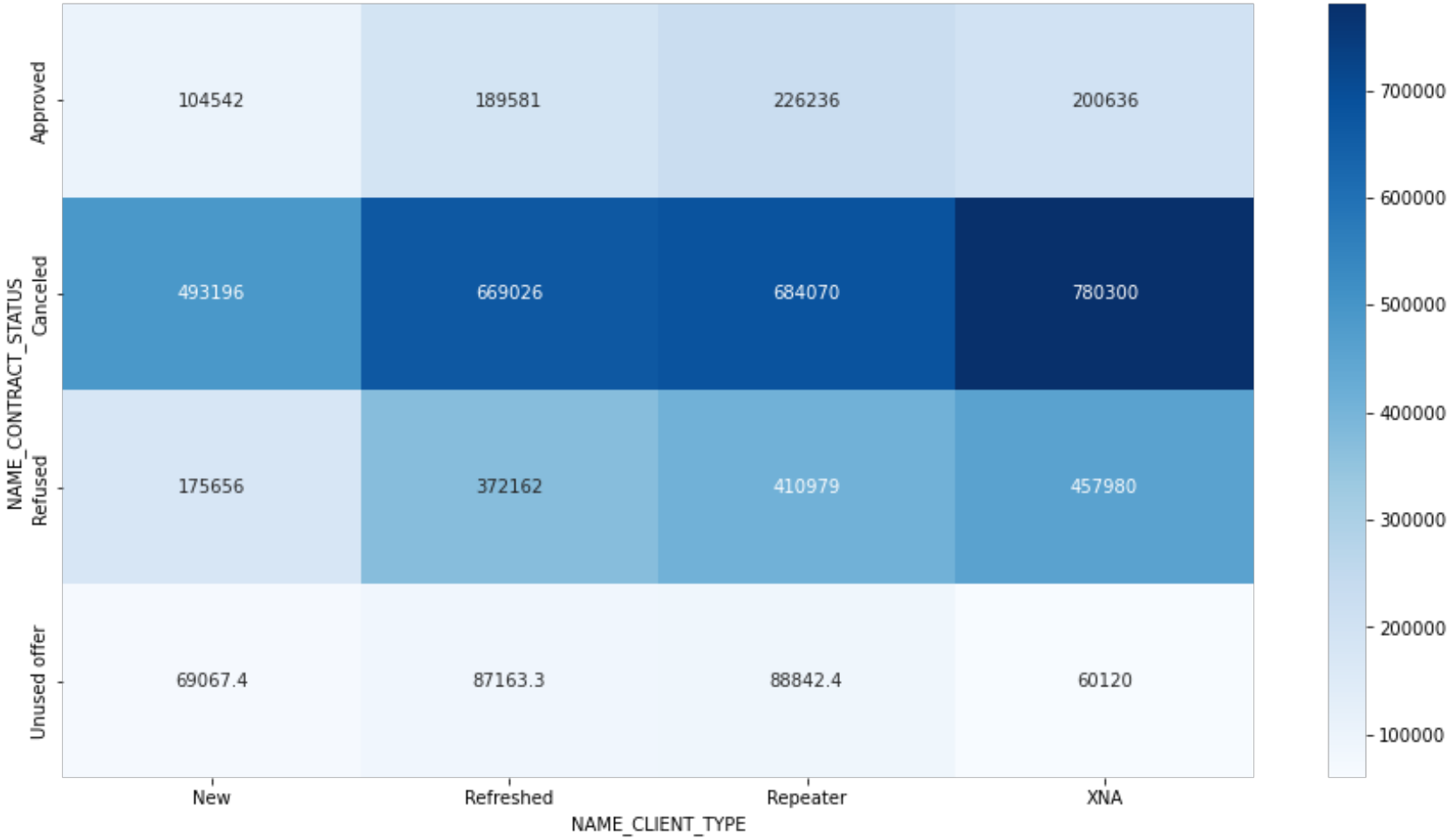
4.AMT_CREDIt and AMT_GOOD PRICE dont seem to be increasing proportionately with AMT_INCOME for TARGET 1, thus possibly leading to default



Top Correlations from previous_data



Multivariate Analysis



Summary

- All of the following variables are defined as variables that cause errors in the dataframe analysis application.
  Check and validate unapproved loans
  - Average income
  - Age group 25-35, followed by age group 35-45
  - Male
  - Unemployed
  - Worker, seller, driver
  - Job type 3 444 Housing - None
  Consider Other important factors to consider are
  - Phone number change deadline - Lower number to display
  - Roundup Office Clicks. month vs. - zero hits are good
  - income does not correspond to "good buy" - low income and high price are problems
  -
  previously rejected, abandoned, not using credit cards also have errors of concern. This indicates that the financial institution rejected/rejected the previous application but approved the current application and defaulted.
- Application Not Approved
  _The application was not used with a lower loan amount. Is this a reason not to use it?
  _More weight should be given to female candidates because there are less by default.
  _60% of defaulters are job applicants. This does not mean that applicants should be rejected.
  Other measures need comprehensive review
  _Previous applications rejected, cancelled, unused loans also have timely payments in current practice. This shows that bad decisions can be made in these situations.