upGrad

Credit EDA Case Study

- Sourabh Somvanshi and Nikhil Bari

Problem statement and Requirement

Usually Banks face problems while deciding giving loans to applicants.

In this ‘Credit EDA Case Study’ we have to analyse provided data which will help banks to decide to whom the loan should be given, so that the Bank remains profitable.

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 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.

Given Data

We have used two Datasets provided:

1. Application Data
2. Previous application Data
3. Columns_description

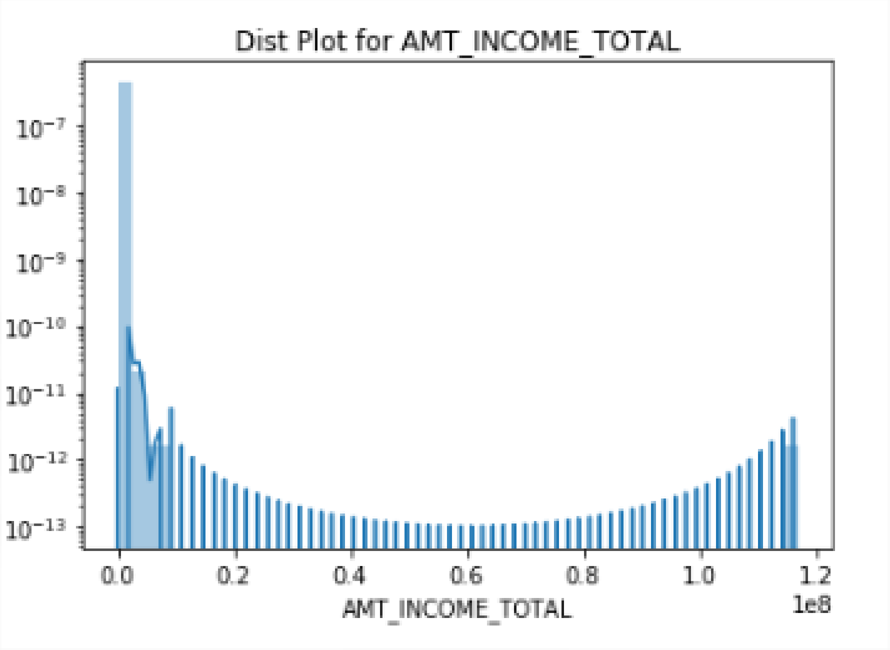
Data Quality Check and Missing values

Following steps were performed on Application Data:

1. Calculated the percentage of null values
2. Remove columns with high missing percentage (more than 14%)
3. Imputed the columns with missing value less than 13%
4. Changed the datatypes of required columns

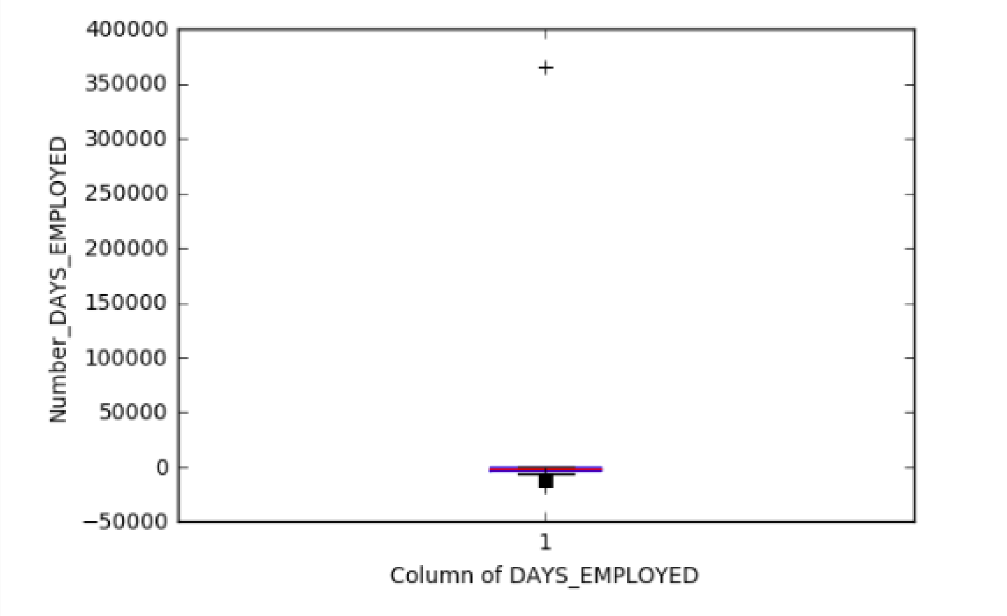
Checking for Outliers and Imputng

1. Refering to the column 'AMT_INCOME_TOTAL' the max value is 1.170000e+08 which is way above the mean and 75% . Hence treating 1.170000e+08 as an outlier.



2. Finding and Analysing Outliers: DAYS_EMPLOYED

Refering to the column 'DAYS_EMPLOYED' the max value is 365243 days of employment (which is an outlier)



Binning of continuous variables.

Created bins for column ‘CNT_CHILDREN’

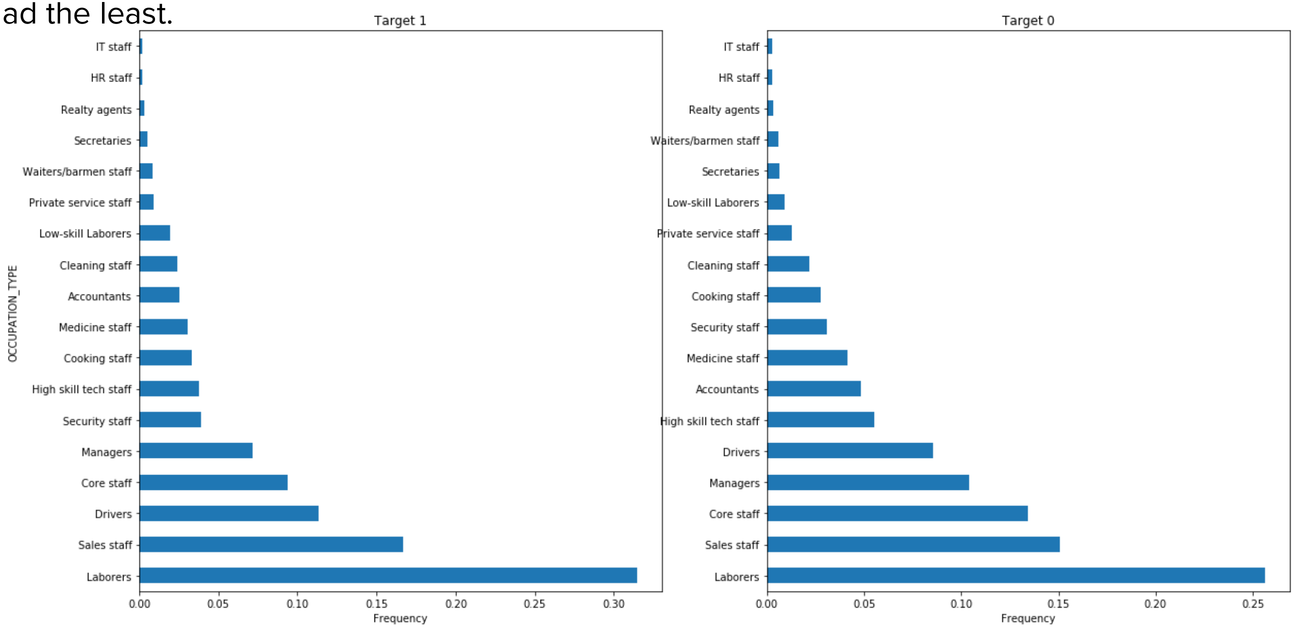
Bin interval [0,5,10,15,20,100]

Analysis

Divided the data into two sets, i.e. Target=1 and Target=0 and performed univariate and Bivariate analysis on following Columns:

OCCUPATION_TYPE –

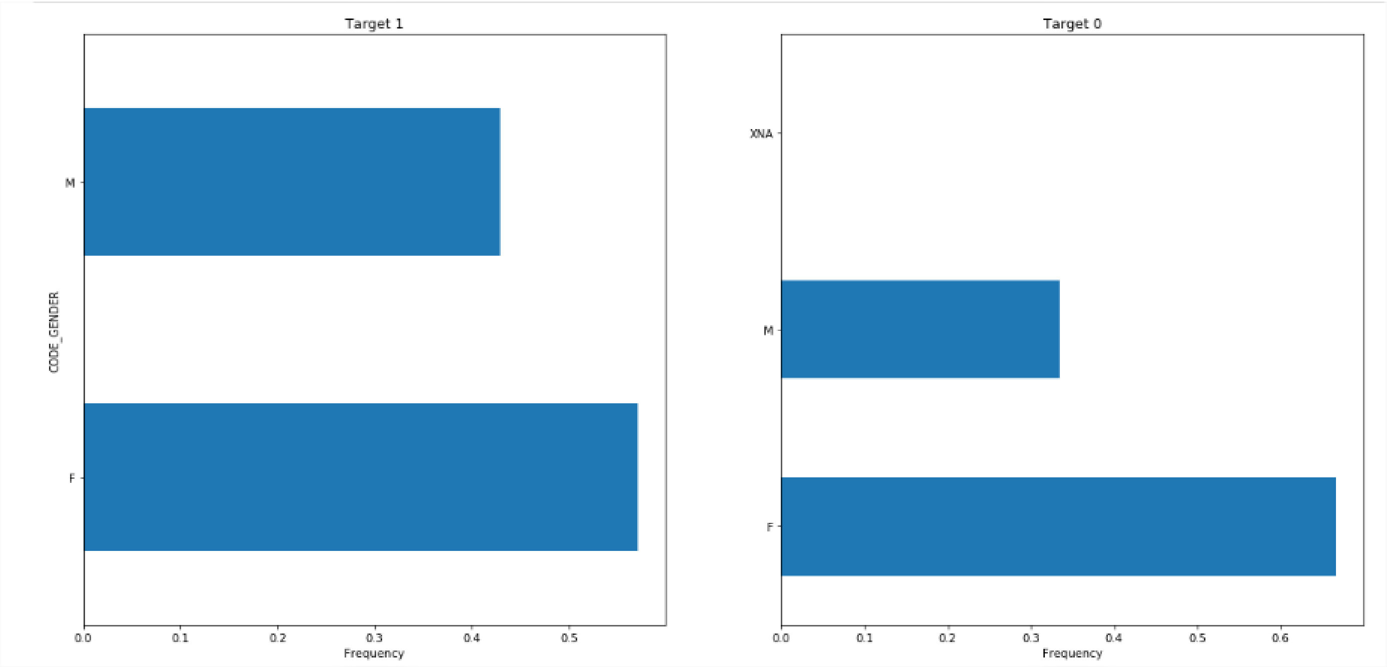
From the below 'OCCUPATION_TYPE' plot we can observe that 'Laborers' had most payment difficulty whereas 'IT staff' had the least.



upGrad

CODE_GENDER

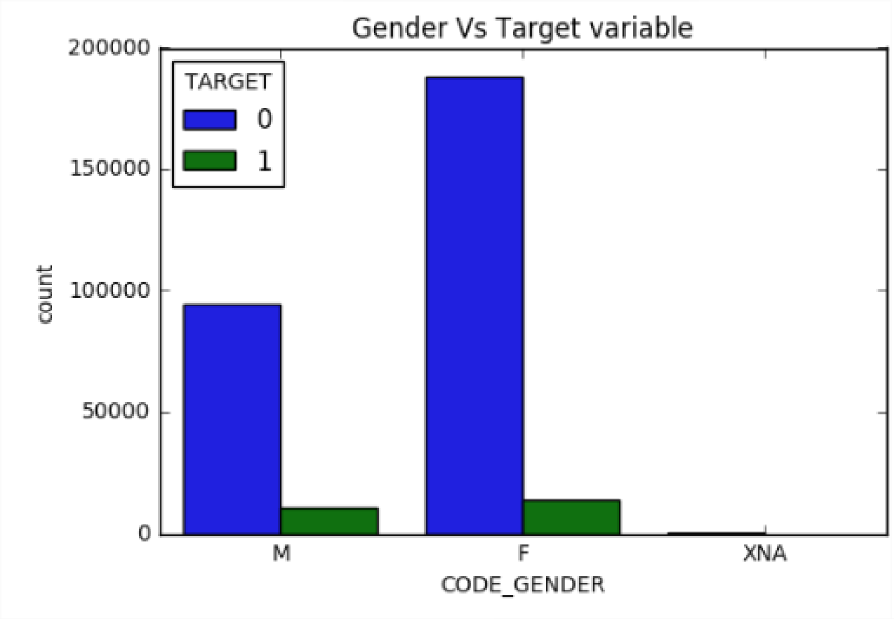
From the 'CODE_GENDER' plot we can observe that 'Gender: Female' had most payment difficulty than Male applicants.

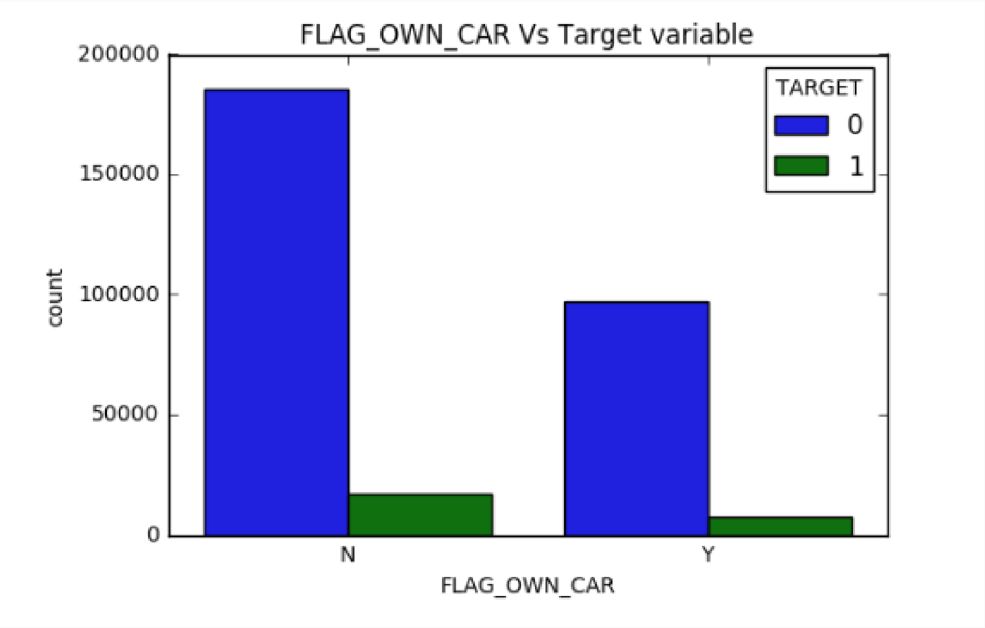


Perform Bivariate analysis for categorical- categorical variables for both 0 and 1.

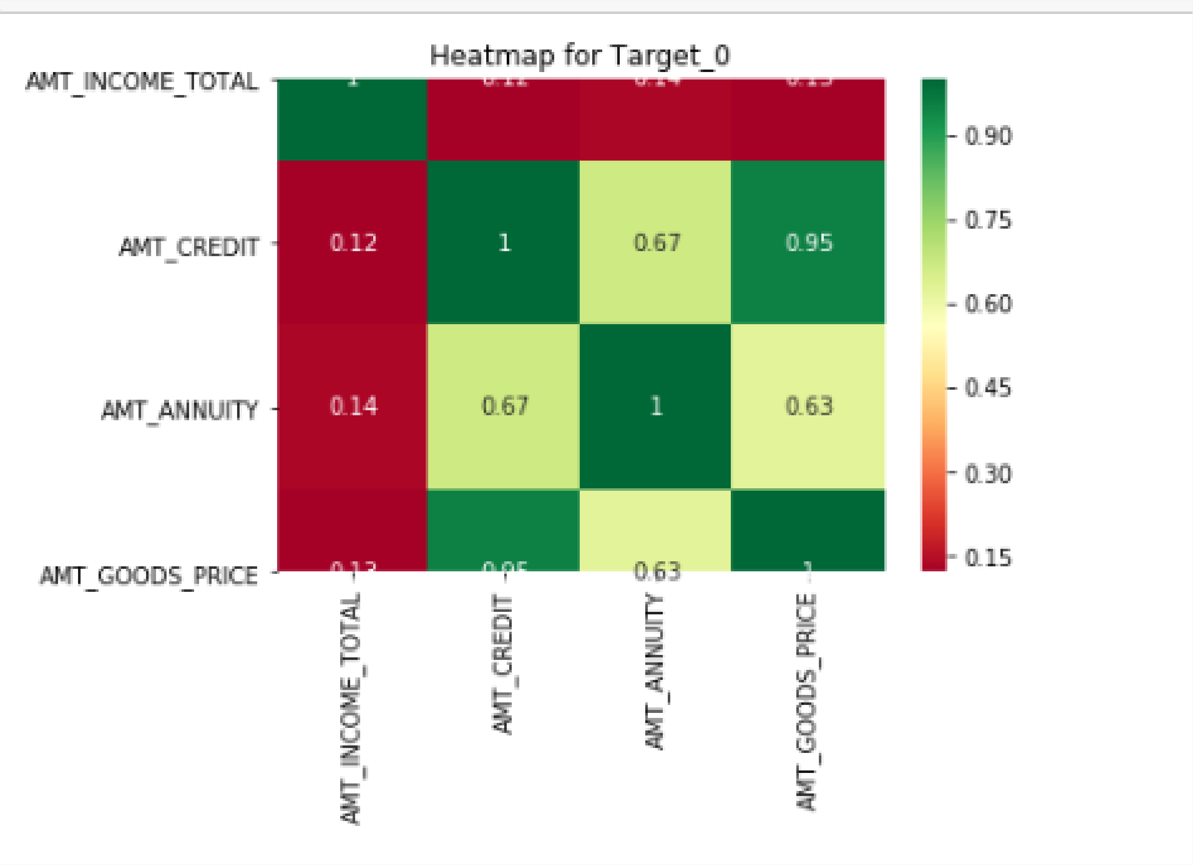
upGrad

From 'Gender Vs Target variable' plot we can see that Target 0 i.e 'all other cases' count is higher for Male,Female,XNA than Target 1





Correlation for numerical columns



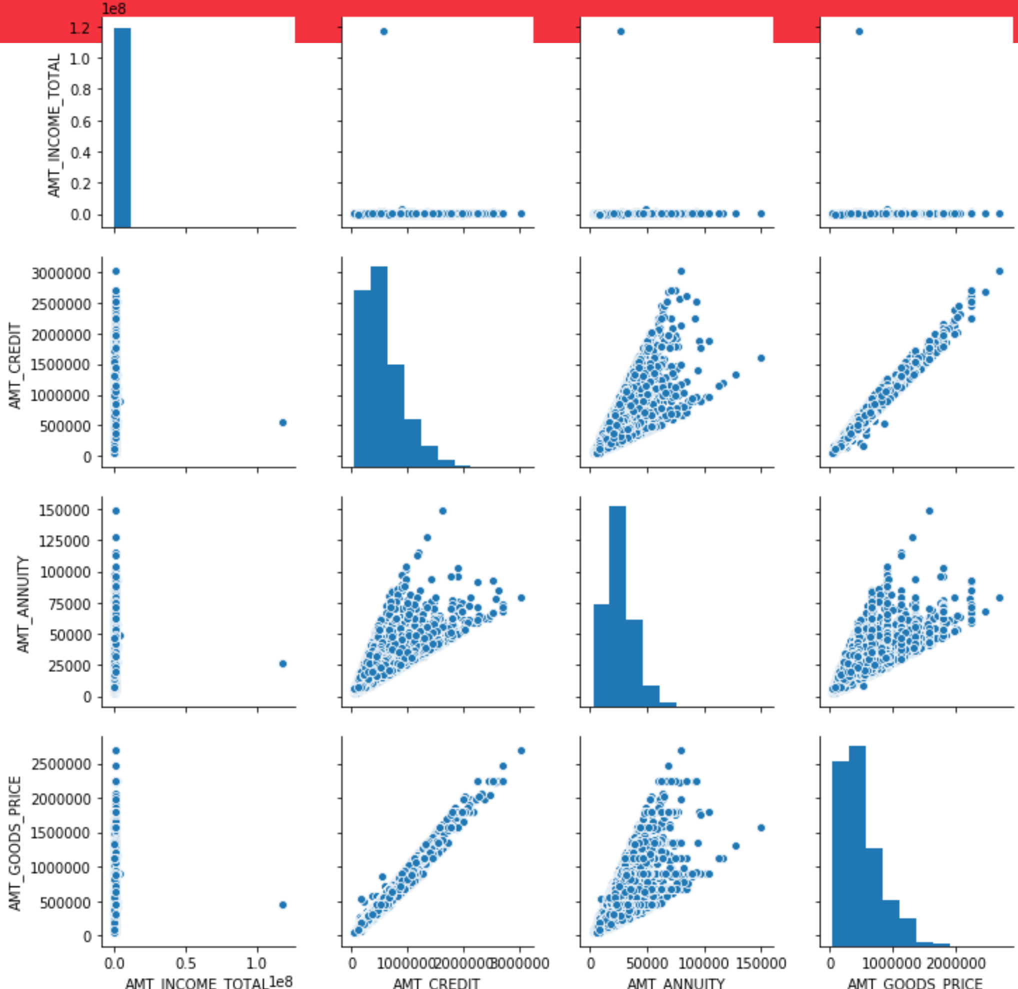
From the above heat maps of Target 1 and Tarrget 0

The highest correlation observed is between 'AMT_GOODS_PRICE' and 'AMT_CREDIT'

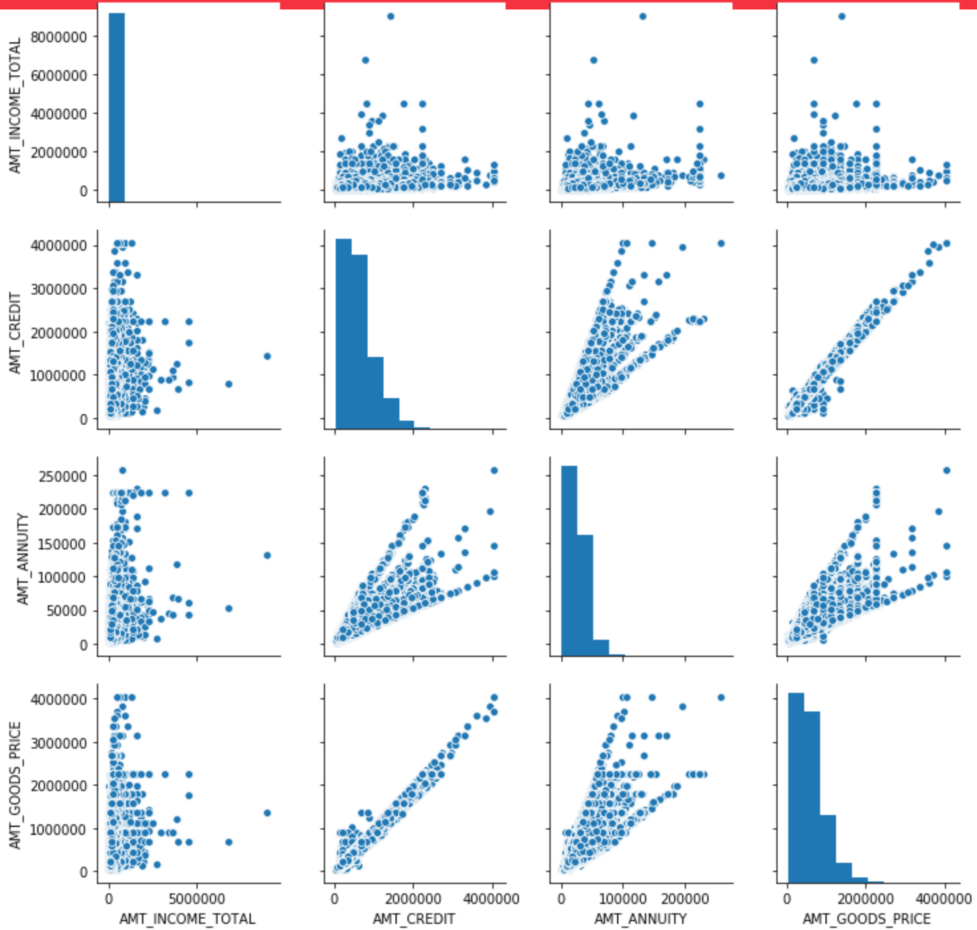
For Target 1 it is 0.91 ,whereas for Target 0 it is almost 1

Therefore, the highest correlation is not same.

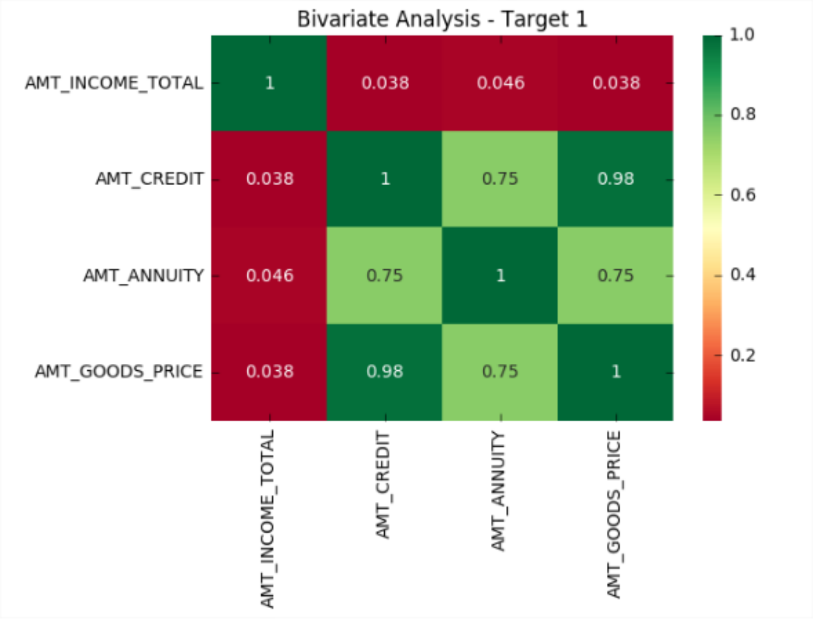
Bivariate Analysis - Target 1

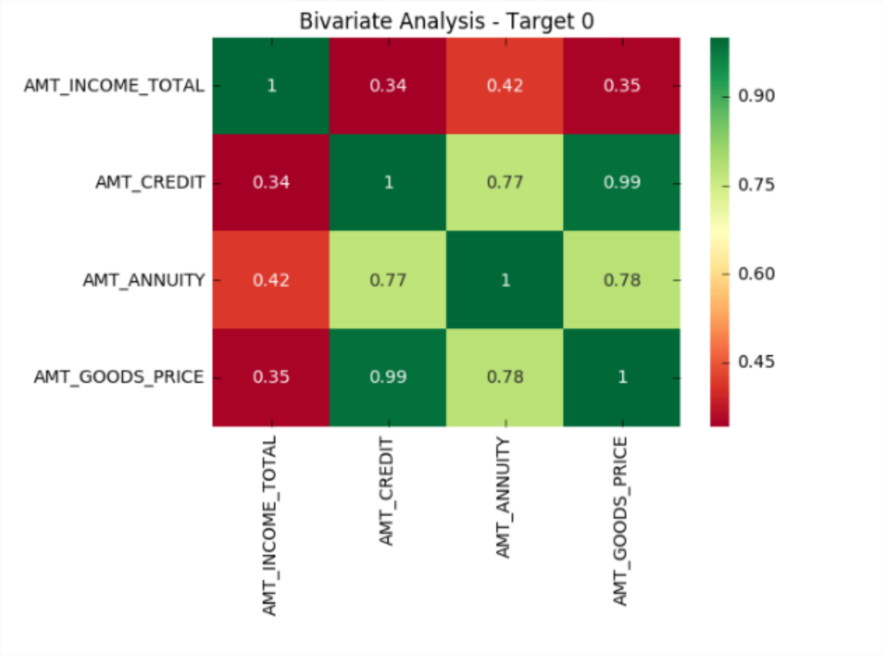


Bivariate Analysis - Target 0



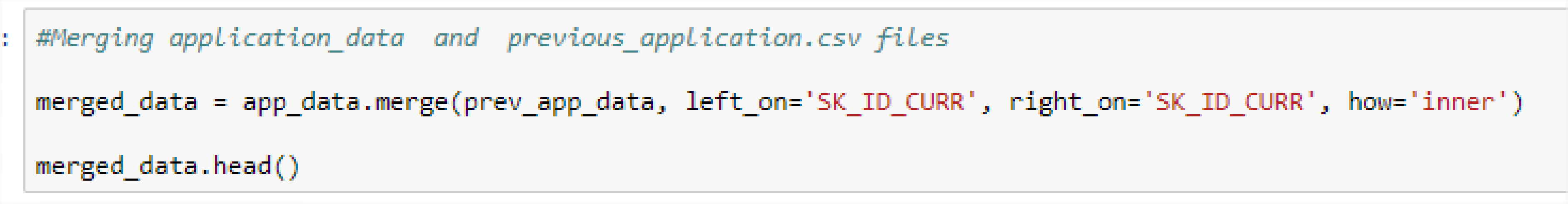
Heat Map for Bivariate Analysis - Target 1 and 0


Bivariate Analysis - Target 1


Bivariate Analysis - Target 0

Merged Application Data and Previous Application

We merged the given data as shown in below snippet:



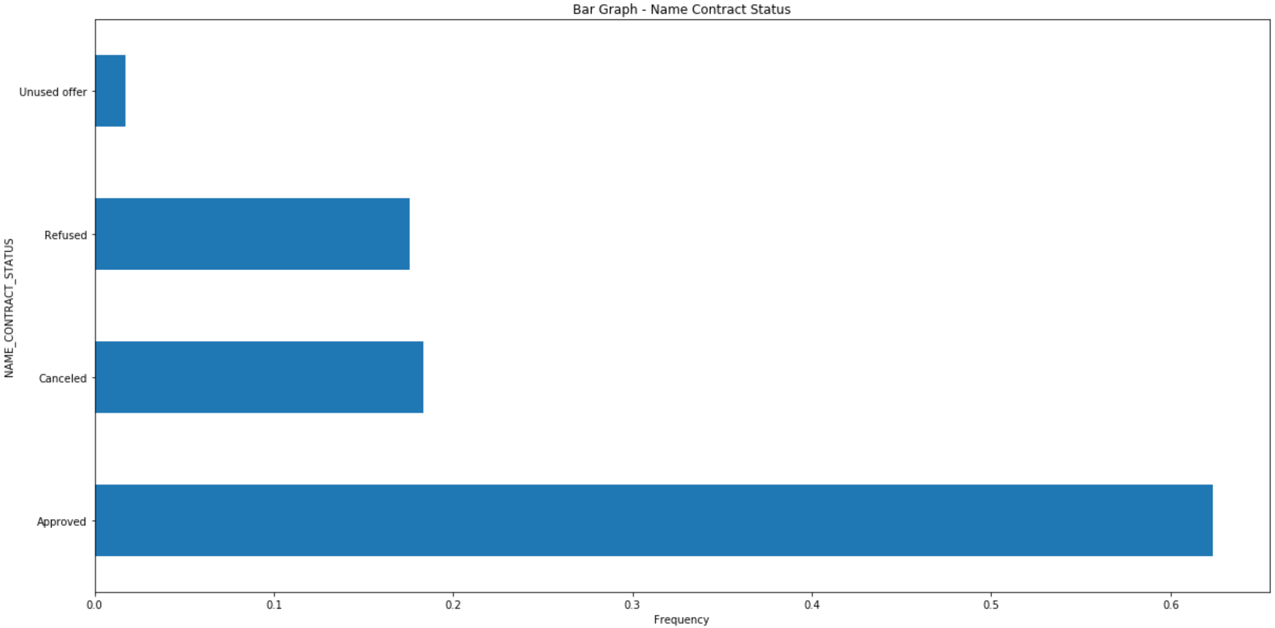
```
#Merging application_data  and  previous_application.csv files

merged_data = app_data.merge(prev_app_data, left_on='SK_ID_CURR', right_on='SK_ID_CURR', how='inner')

merged_data.head()
```

Performed univariate and bivariate analysis to find some pattern

Univariate Analysis



Univariate Analysis

upGrad

From the 'Bar Graph - Name Contract Status' plot we can observe that Approved percentage is more than around 65%

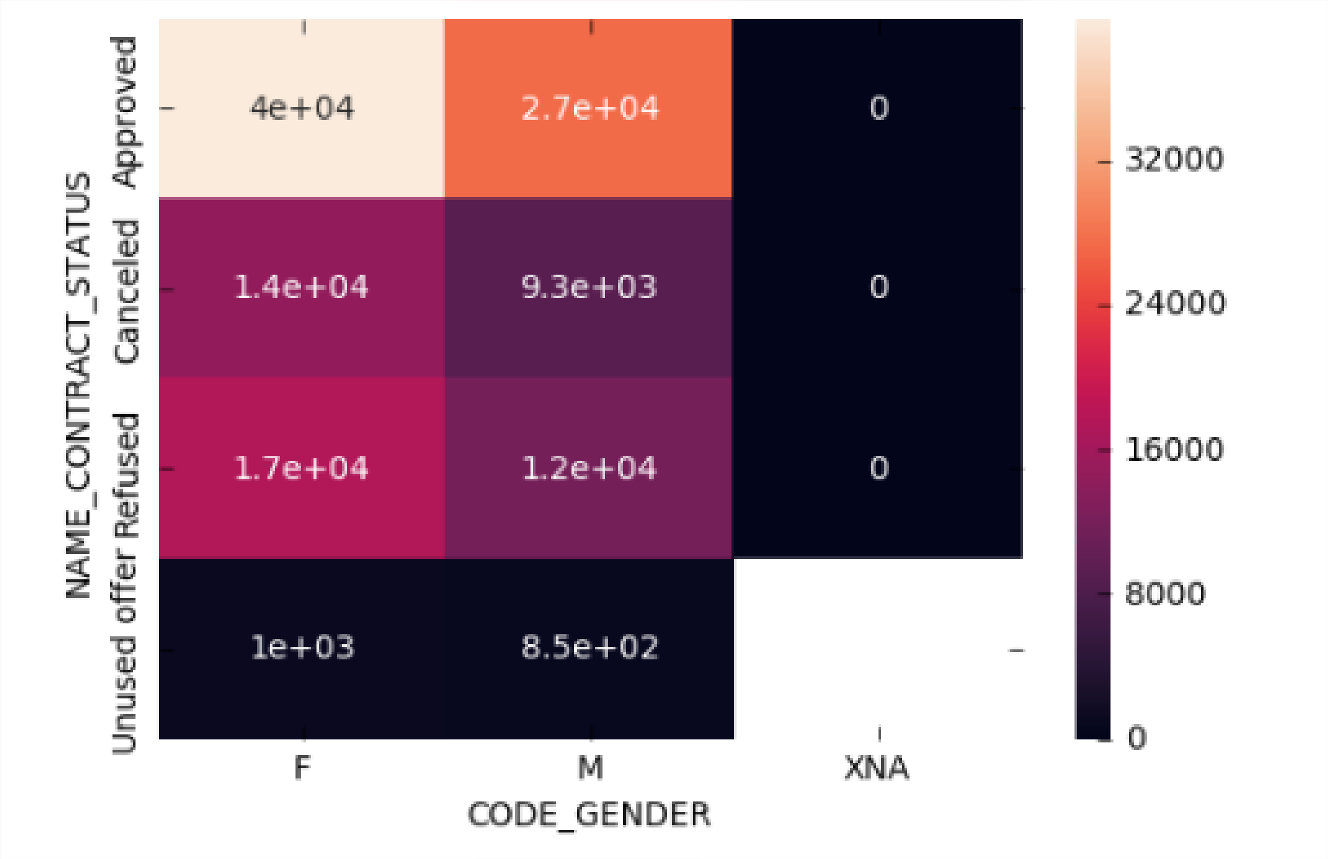
Canceled percentage is around 18%

Refused percentage is around 15%

Unused offer percentage is around 5%

Bivariate Analysis

upGrad



Bivariate Analysis

The heatmap is plotted with respect to Gender and NAME_CONTRACT_STATUS having sum of Target 1 (client with payment difficulties)
From the heatmap we can infer following points:

1) Approximately 40,000 Female applications were Approved who had late payments on at least one of the first installments of the loan in our sample. (Highest value in the heatmap)

2) Approximately 9,300 Male applications were Canceled who had late payments on at least one of the first installments of the loan in our sample.

3) And approximately 1000 Female applications were under 'Unused Offer' category