Credit EDA Assignment

Pugalenthi Soundararajan (Pugal)

Problem Statement

- Identifying the customers who not likely to repay the loan and reduce the financial loss to the company.
  - 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.
- Identify the customers who likely to repay the loan and provide offers.
- Perform risk analysis using datasets provided (current and previous applications data)

Key Steps

- Data understanding
- Data cleaning and manipulation
- Find missing values
- Impute missing values (if required)
- Prepare dataset for analysis (categorical / numerical and filter data with respect to Target 1 and Target 0)
- EDA – Univariate, Bivariate and Multivariate analysis
- Perform EDA with merged dataset (current and previous application)
- Observations

Data understanding – Application data

- Found missing values of more than 40% in 49 columns
- Dropped missing values columns and created a new data frame (df)
- Info function gives the details of missing values

```
66   AMT_REQ_CREDIT_BUREAU_HOUR   265992 non-null  float64
67   AMT_REQ_CREDIT_BUREAU_DAY    265992 non-null  float64
68   AMT_REQ_CREDIT_BUREAU_WEEK   265992 non-null  float64
69   AMT_REQ_CREDIT_BUREAU_MON    265992 non-null  float64
70   AMT_REQ_CREDIT_BUREAU_QRT    265992 non-null  float64
71   AMT_REQ_CREDIT_BUREAU_YEAR   265992 non-null  float64
```

- Value counts and describe function helps to find out most repeated or mean of the above columns

Data understanding – Application data (fixing missing values)

```
for i in df[missing_val_AMT_REQ_CREDIT].columns:
    print(df[i].value_counts())
```

- even though columns (AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR) data type is float, by running value counts function, its clearly visible these columns are categorical
- From describe function, it is observed the most repeated or mean of these columns are ‘0’.
- Filling the above-mentioned columns with '0’

Data understanding – Filling missing values / Replace / Grouping Category columns

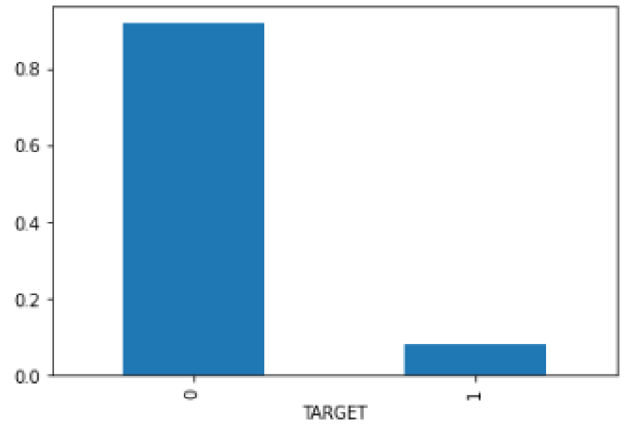
- EXT_SOURCE_3 column's missing values filled mean (based on data distribution using box plot)
- Occupation_Type column have missing values; it was replaced with 'not known'
- Dropping the rows with missing values directly, since 2980 rows have missing values out of 307511 rows. Which is 1% of missing data being dropped.
- Replaced XNA values in Gender column with 'F' being the mode of the column.
- Grouping other_a, other_b and group of people values in NAME_TYPE_SUITE column with other.
- 'ORGANIZATION_TYPE' column - less than 1% organization grouped into 'Other'
- 'ORGANIZATION_TYPE' column – Business, Trade, Transport and Industry names regularized

Data understanding – How balanced?

- Finding out how the data is balanced (Target1 - defaulter / clients have payment difficulty and Target0 - non-defaulters)

```
round(100*df['TARGET'].value_counts(normalize=True),2)
```

```
TARGET
0    91.9
1     8.1
Name: proportion, dtype: float64
```



- Data is not balanced, as the contribution of defaulter and non defaulter is 8% and 92% respectively. Ratio of data imbalance is 11.35.

```
# Calculating ratio of imbalance data
round(100*df['TARGET'].value_counts(normalize=True),2)[0] / round(
                                                        100*df['TARGET'].value_counts(normalize=True),2)[1]
```

```
11.34567901234568
```

Data Cleaning and Manipulation

- Classifying categorical and numerical columns based on data type and unique values in the columns

```
i=0
cat_cols = []
num_cols = []
for col in df.columns:
    if (len(df[col].unique())<=10) or (df[col].dtypes == 'object'):
        print(col,len(df[col].unique()))
        i=i+1
        cat_cols.append(col)
    else:
        num_cols.append(col)
```

Data Cleaning and Manipulation

- Removing the columns which are not useful for analysis
- From the value counts analysis of Flag columns, it is observed that all the columns have two distinct values (0 and 1). And the maximum percentage of the data are '0'. its better to drop as it doesn't provide any insights
- Columns starts with Flag from categorical data
- Columns from numerical data AMT_REQ_CREDIT_BUREAU_MON’, 'AMT_REQ_CREDIT_BUREAU_QRT are removed as the observations shown from box plot (most of the data points were lies at zero)
- Box plot of days columns 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH’ were in negative, so fixed by applying .abs() python function.

Outliers in Numerical columns and Box plot

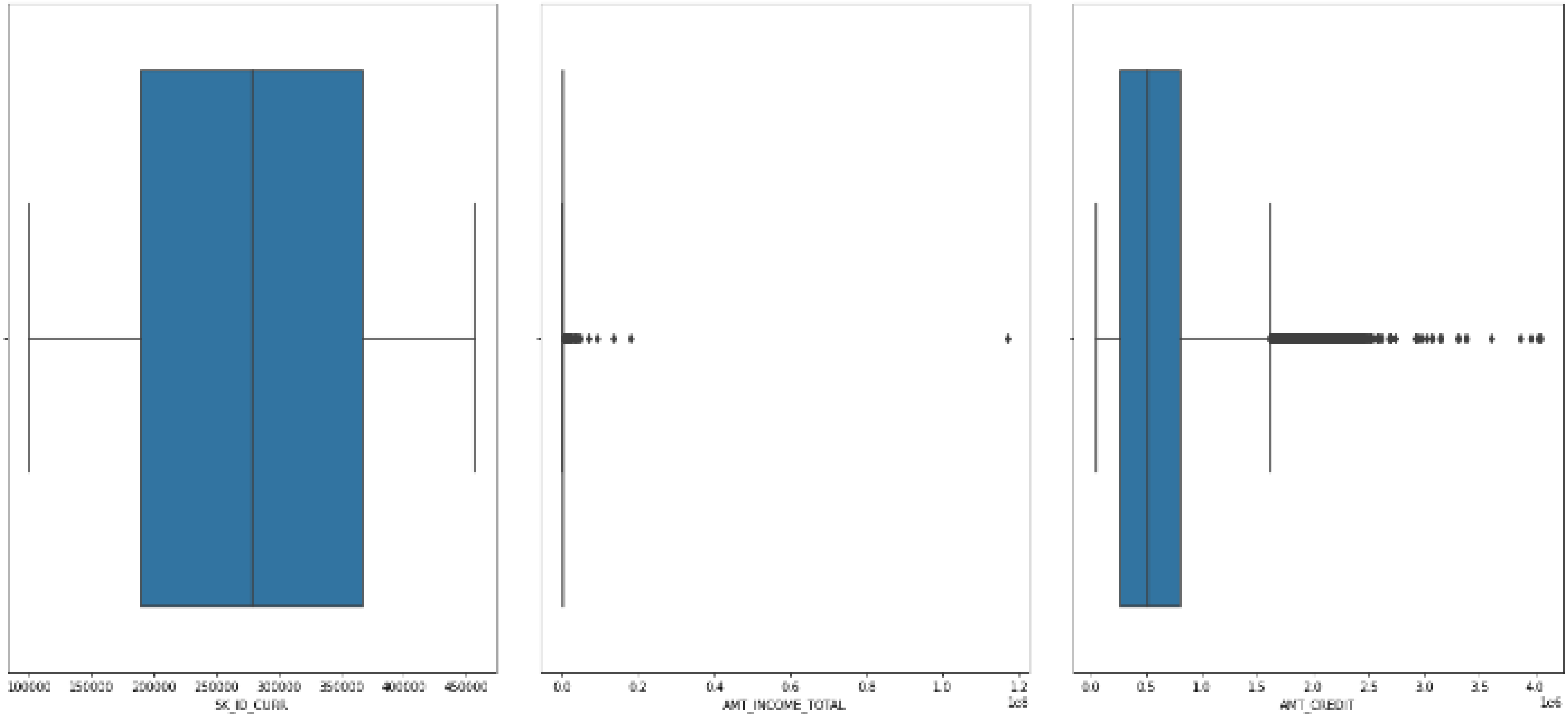
- Found outliers in many of the numerical data columns, Outliers were identified as many data points are fall out of max range in box plot.
- And it is evident that by calculating upper and lower bound values using IQR - Refer the below code.
- Not fixing the outliers as suggested in the problem statement hint in upgrad platform.

```
#Calculating outliers with IQR upper and lower bound
for i in app_num_df.columns:
    q1 = app_num_df[i].describe()['25%']
    q3 = app_num_df[i].describe()['75%']

    iqr = q3-q1

    upper_bound = q3 + 1.5*iqr
    lower_bound = q1 - 1.5*iqr
    print(i,"---", upper_bound,'and', lower_bound)
```

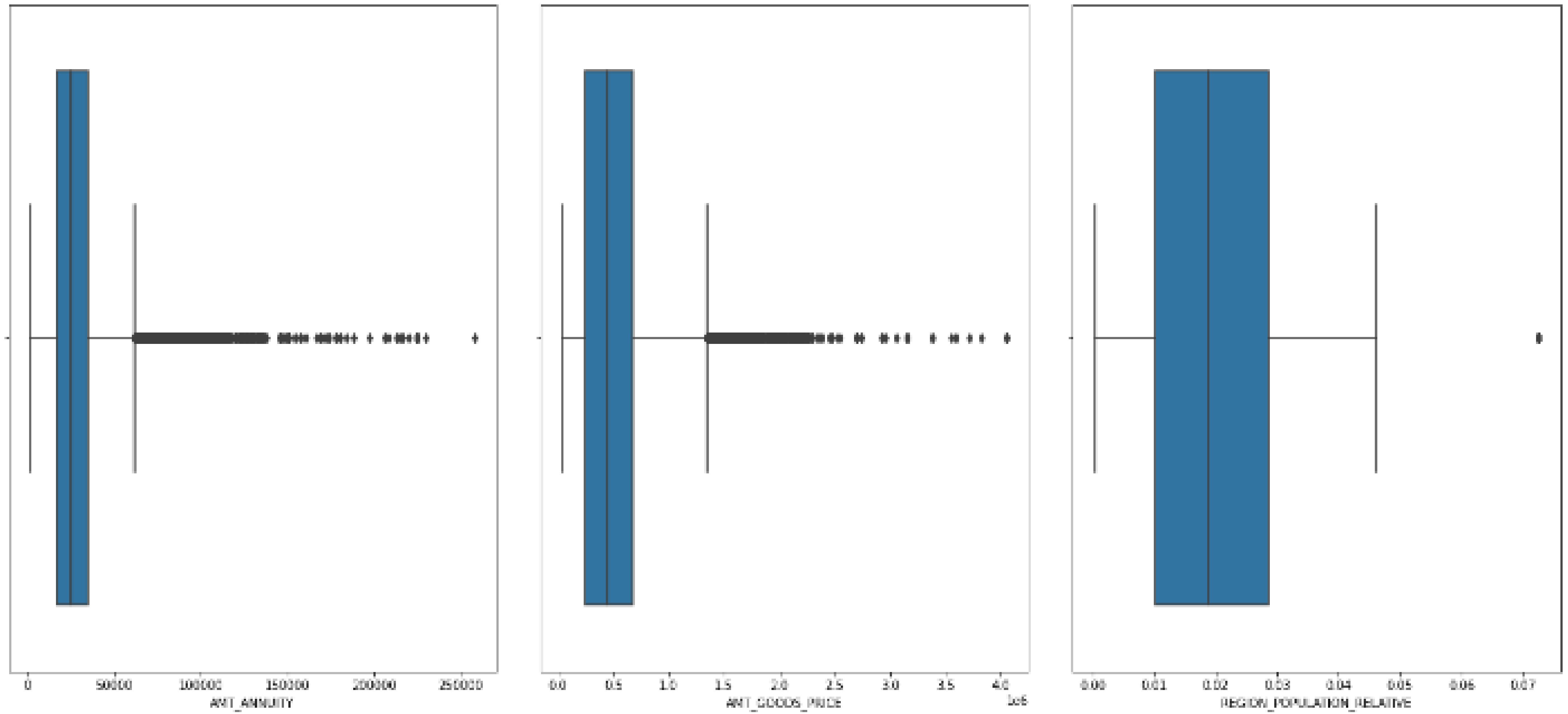
```
AMT_INCOME_TOTAL --- 337500.0 and -22500.0
AMT_CREDIT --- 1616625.0 and -537975.0
AMT_ANNUITY --- 61742.25 and -10527.75
AMT_GOODS_PRICE --- 1341000.0 and -423000.0
REGION_POPULATION_RELATIVE --- 0.06 and -0.02
DAYS_BIRTH --- 30578.0 and 1522.0
DAYS_EMPLOYED --- 12882.75 and -6235.25
DAYS_REGISTRATION --- 15677.0 and -6187.0
DAYS_ID_PUBLISH --- 8166.0 and -2146.0
EXT_SOURCE_2 --- 1.07 and -0.02
EXT_SOURCE_3 --- 0.97 and 0.09
OBS_30_CNT_SOCIAL_CIRCLE --- 5.0 and -3.0
OBS_60_CNT_SOCIAL_CIRCLE --- 5.0 and -3.0
DAYS_LAST_PHONE_CHANGE --- 3516.0 and -1668.0
```

- **Following slides are visualization of numerical data using box plot**

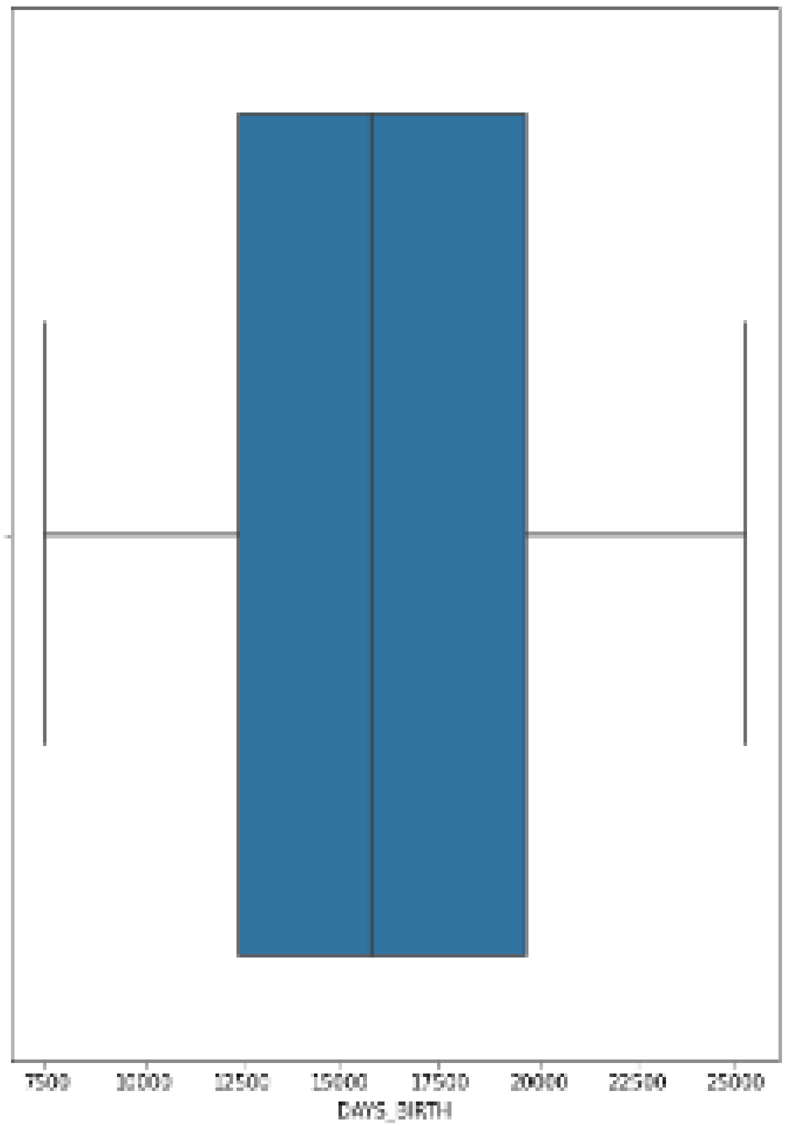
Box plot of Numerical columns

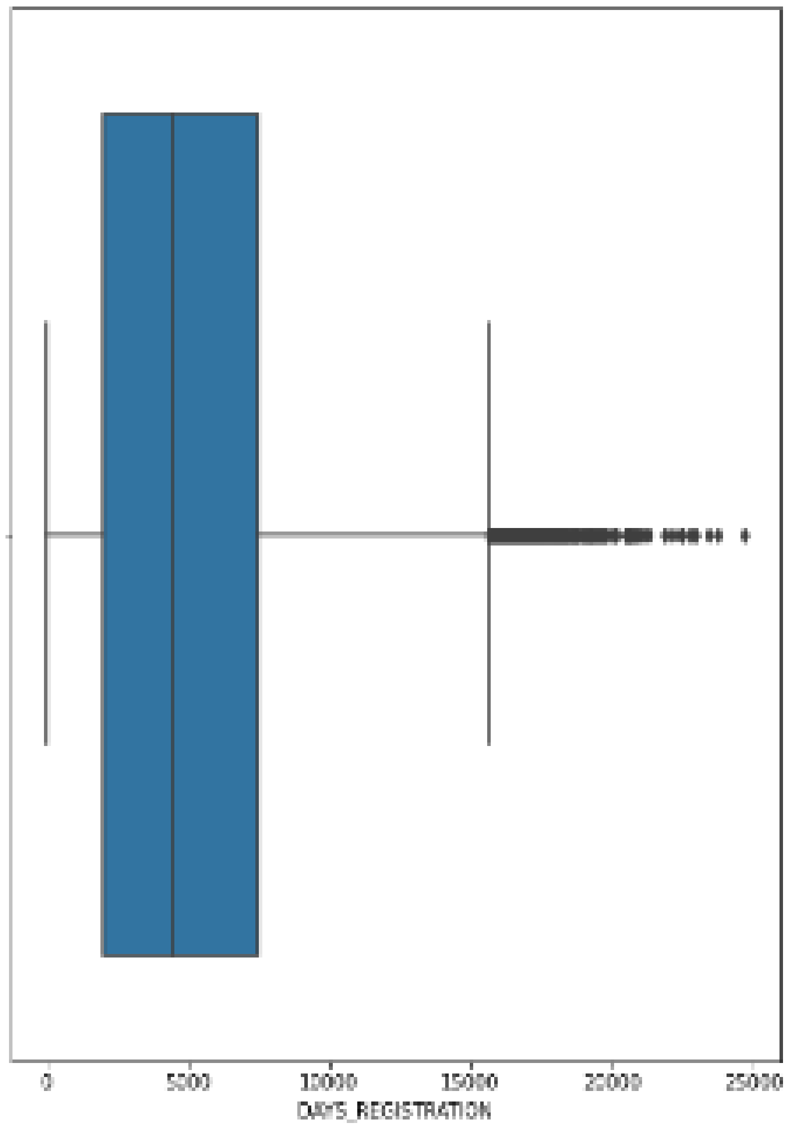


Box plot of Numerical columns

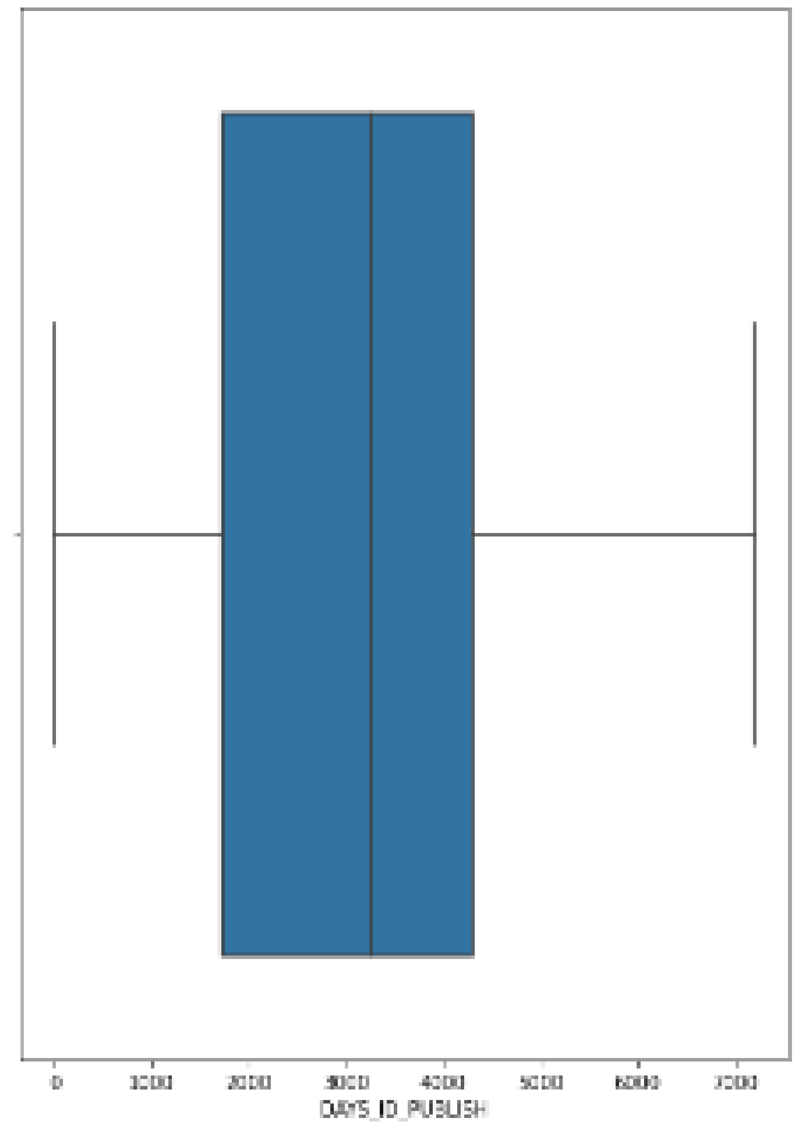


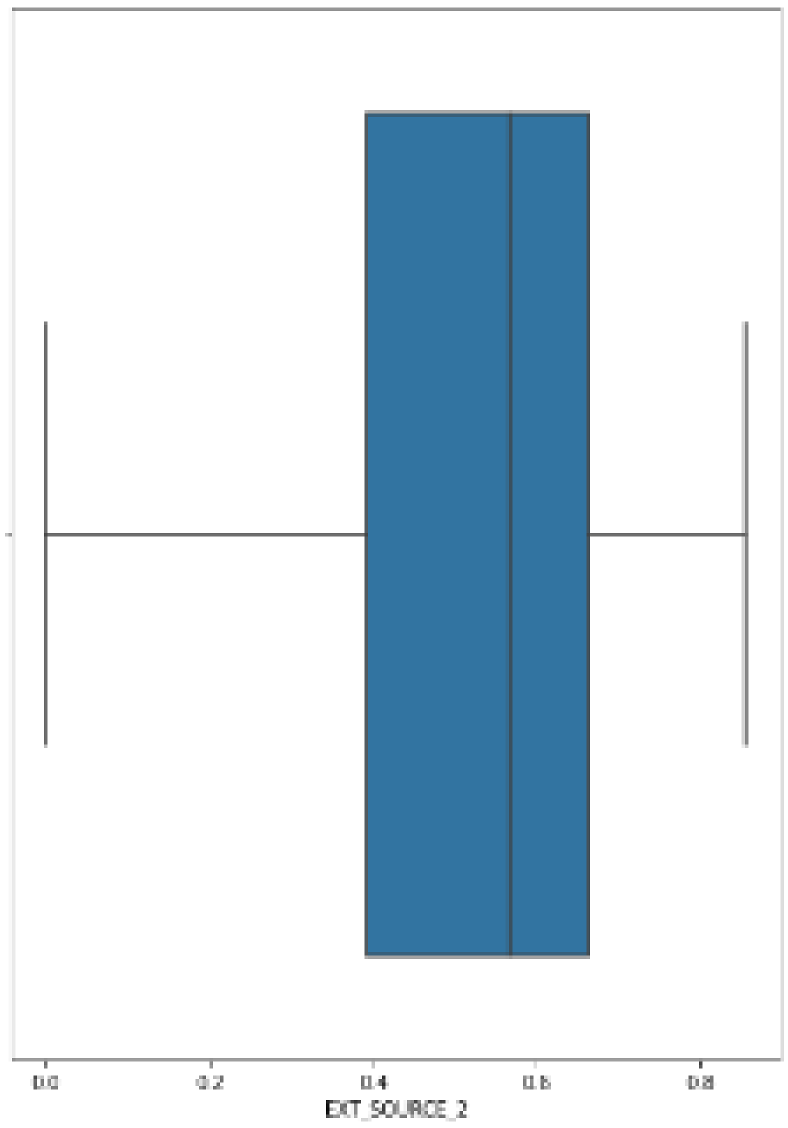
Box plot of Numerical columns

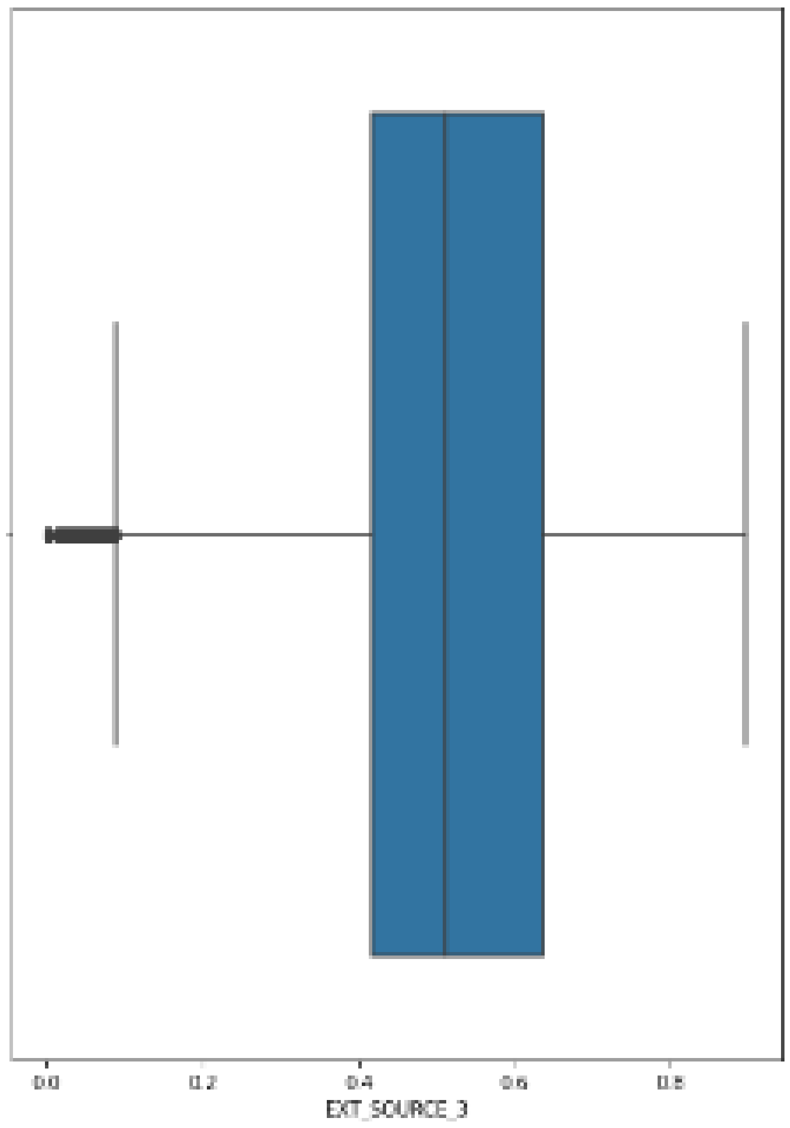




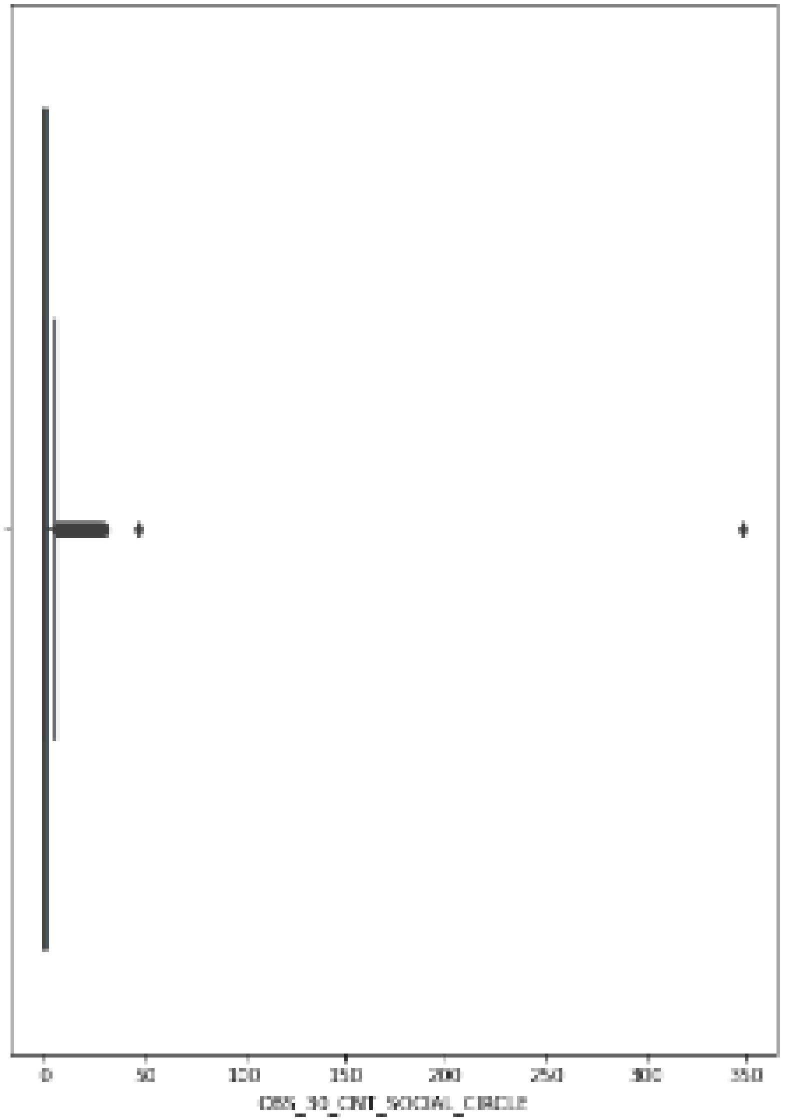
Box plot of Numerical columns

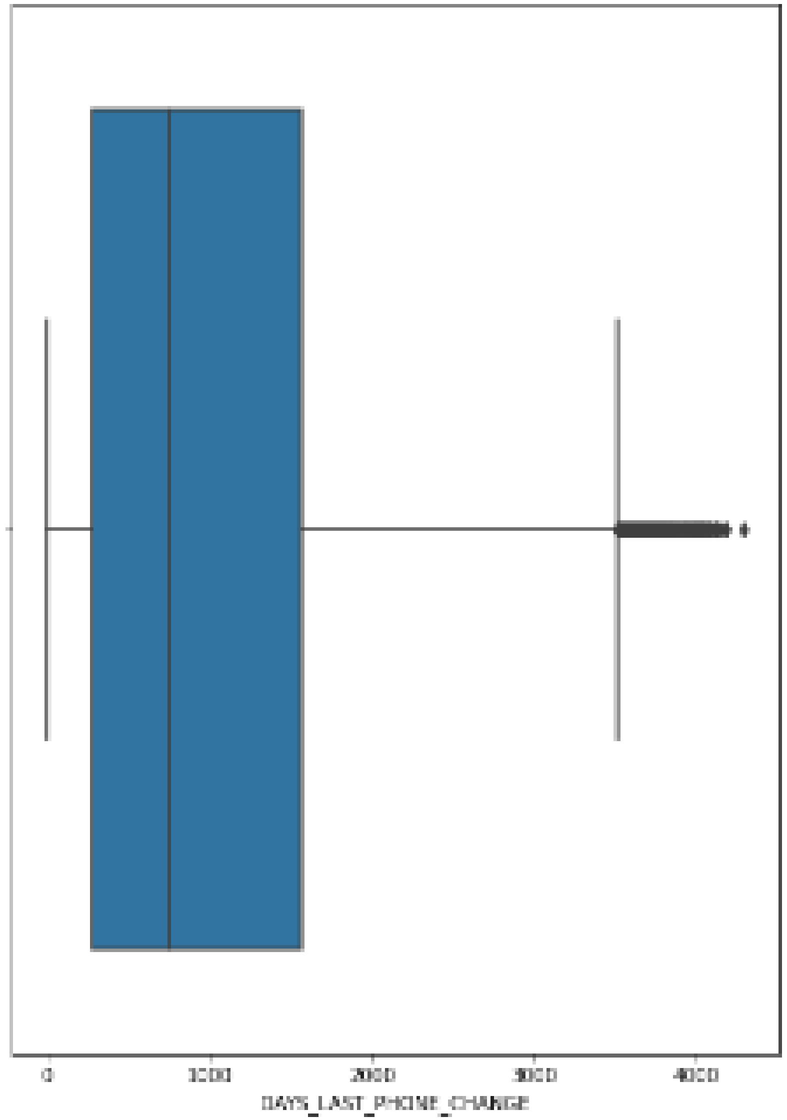






Box plot of Numerical columns



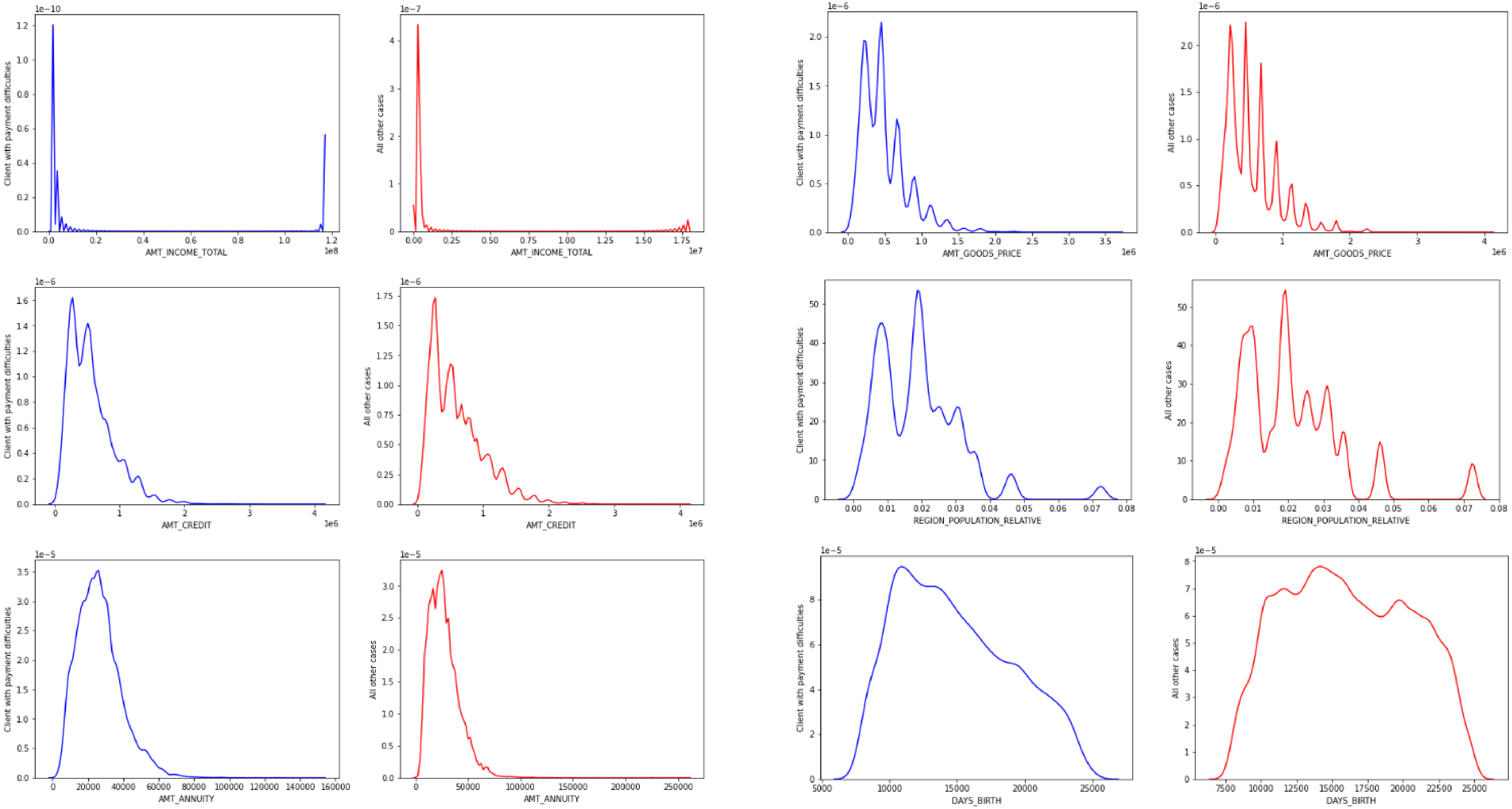


Data analysis – Numerical data (univariate analysis)

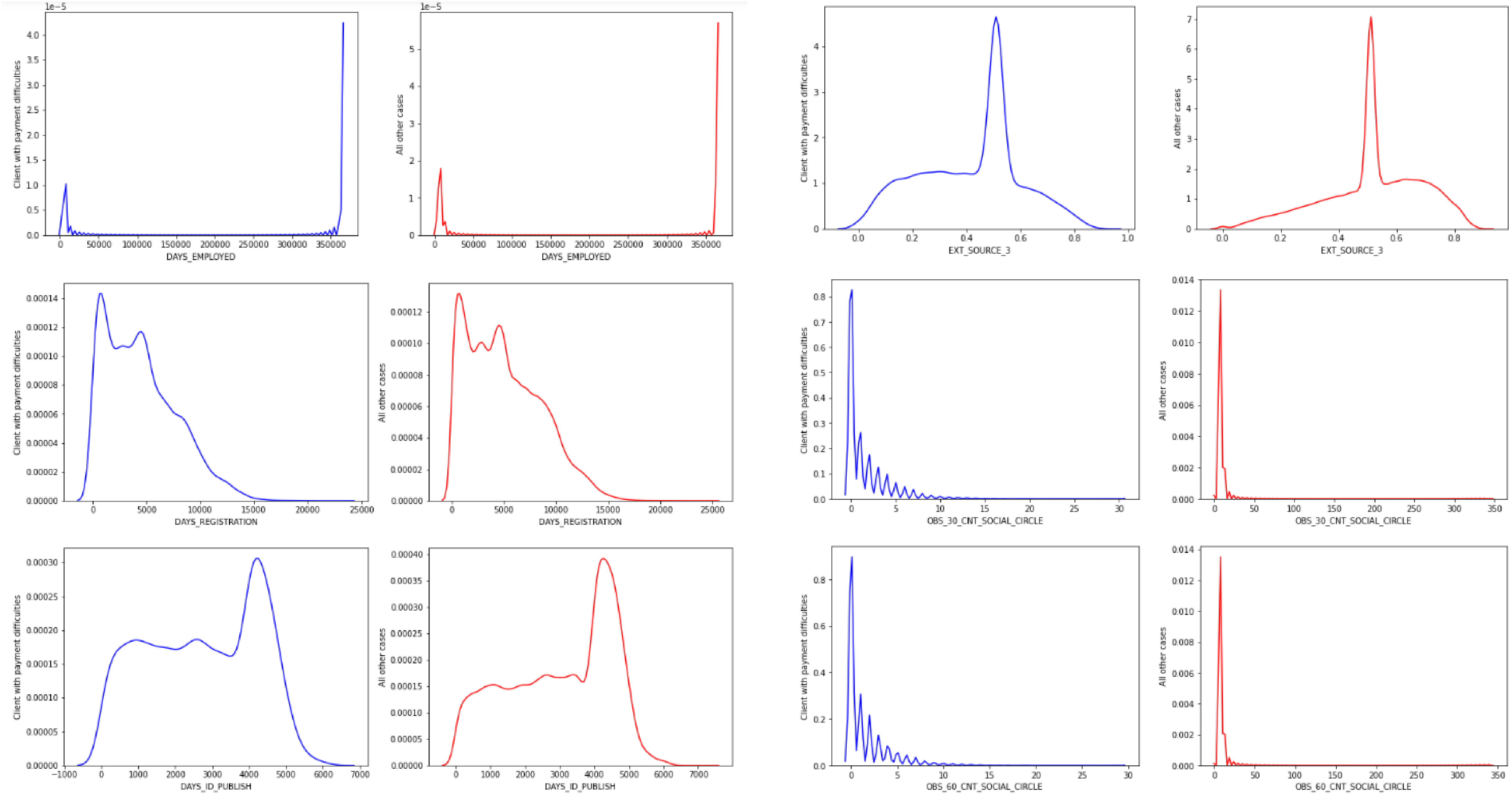
- Further classified numerical data into Target 1 and 0
  - app_num_df_1 = app_num_df[df['TARGET'] == 1]
  - app_num_df_0 = app_num_df[df['TARGET'] == 0]
- Box plot and Dist plot all numerical columns to observe the frequency distribution of the data

```
for i in app_num_df.columns[1:4]:
    plt.figure(figsize=(15,5))
    plt.subplot(1,2,1)
    sns.distplot(app_num_df_1[i], hist=False, color='blue')
    plt.xlabel(i)
    plt.ylabel('Client with payment difficulties')
    plt.subplot(1,2,2)
    sns.distplot(app_num_df_0[i], hist=False, color='red')
    plt.xlabel(i)
    plt.ylabel('All other cases')
    plt.show()
```

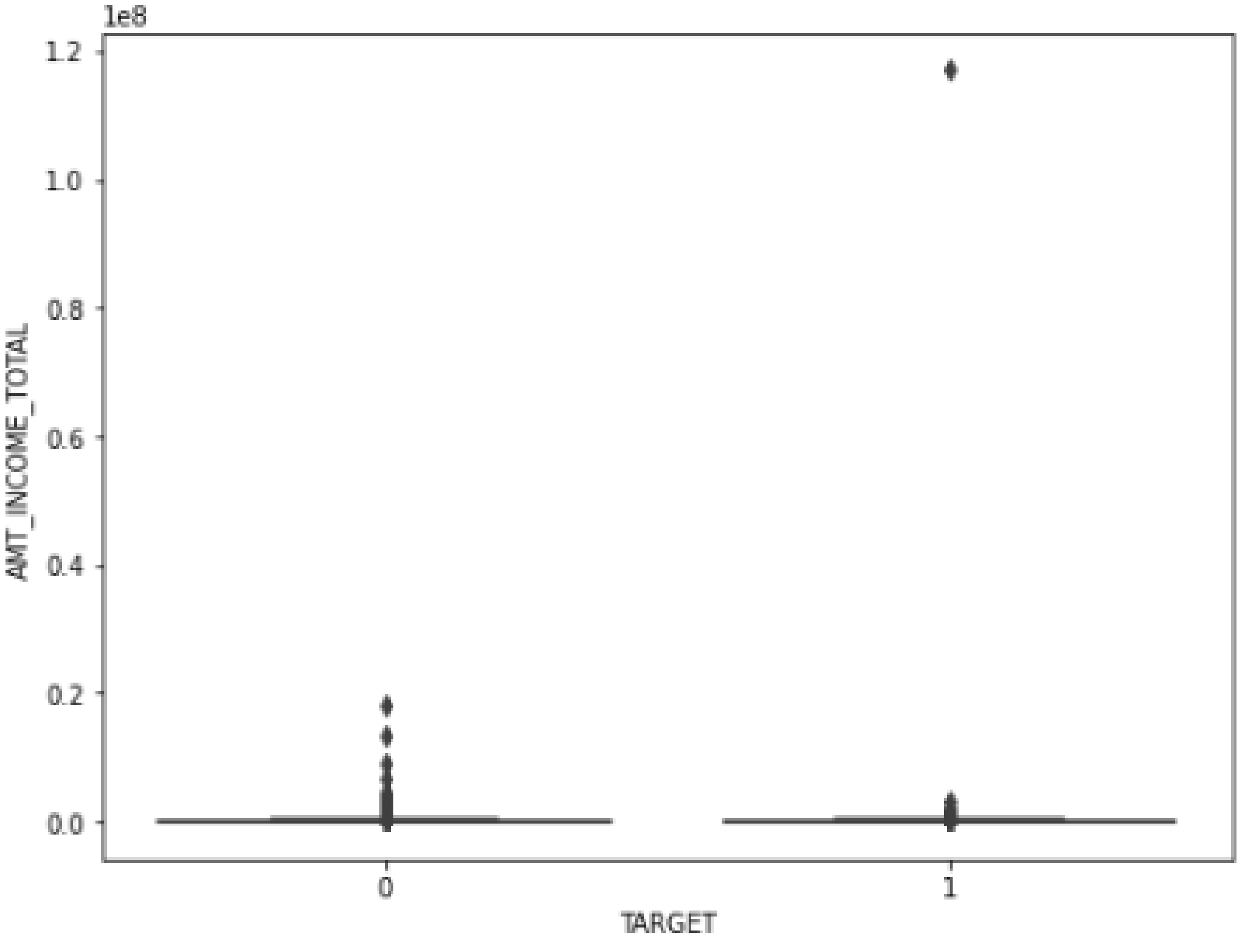
Dist plot of Numerical columns with respect to Target data (‘1’ and ‘0’)



Dist plot of Numerical columns with respect to Target data (‘1’ and ‘0’)



From below Amount Income box plot, it is observed that an outlier in the Target 1 data distribution get skewed and unbale to get any insights from that.



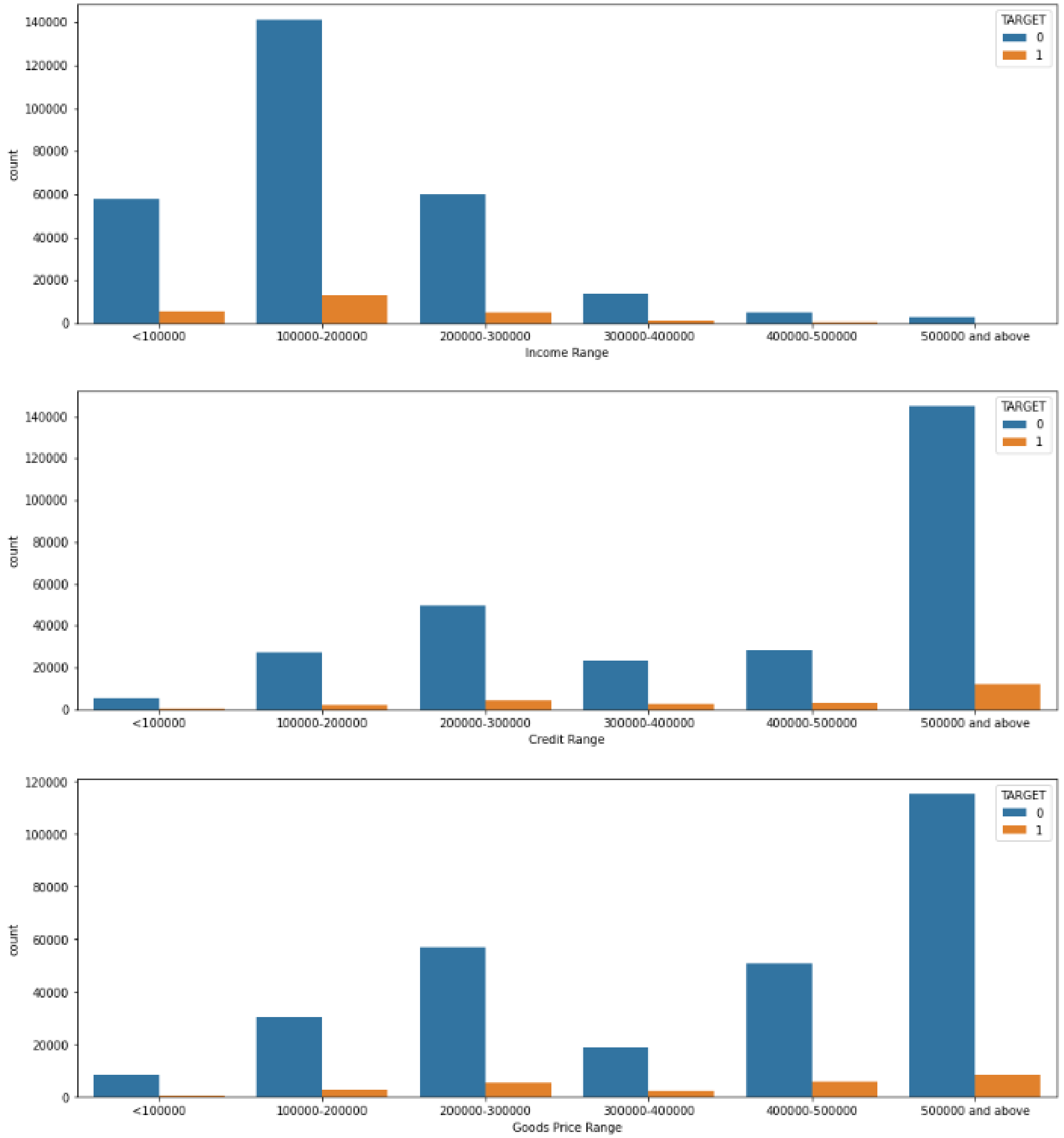
Creating bins for continuous variable column 'AMT_INCOME_TOTAL', 'AMT_GOODS_PRICE' and ‘AMOUNT_CREDIT

```
bins = [0,100000,200000,300000,400000,500000,10000000000]
labels = ['<100000', '100000-200000','200000-300000','300000-400000','400000-500000', '500000 and above

app_num_df['Income Range'] = pd.cut(app_num_df['AMT_INCOME_TOTAL'],bins = bins, labels = labels)
app_num_df['Credit Range'] = pd.cut(app_num_df['AMT_CREDIT'],bins = bins, labels = labels)
app_num_df['Goods Price Range'] = pd.cut(app_num_df['AMT_GOODS_PRICE'],bins = bins, labels = labels)
```

- New columns created from binning - 'Income Range', 'Credit Range', 'Goods Price Range’
- Visualize with count of new columns with respect to Target variable

```
new_cols = ['Income Range', 'Credit Range','Goods Price Range' ]
for i in new_cols:
    plt.figure(figsize=(15,5))
    sns.countplot(app_num_df[i], hue=df['TARGET'])
    plt.xlabel(i)
    plt.show()
```

Count plot of new columns (Income range, Credit range and Goods price range



Insights from Income and credit columns after applying binning method

- Amount Income - '<100000', '100000-200000' contributes for more loan applications and more likely to default
- Amount Income - Clients who have above 300000 less likely to loan repay default
- Amount Credit - Maximum clients have credit range above 500000.

Heat map of all numerical columns

Heat map of all numerical columns with respect to Target data ‘1’

Heat map of all numerical columns with respect to Target data ‘0’

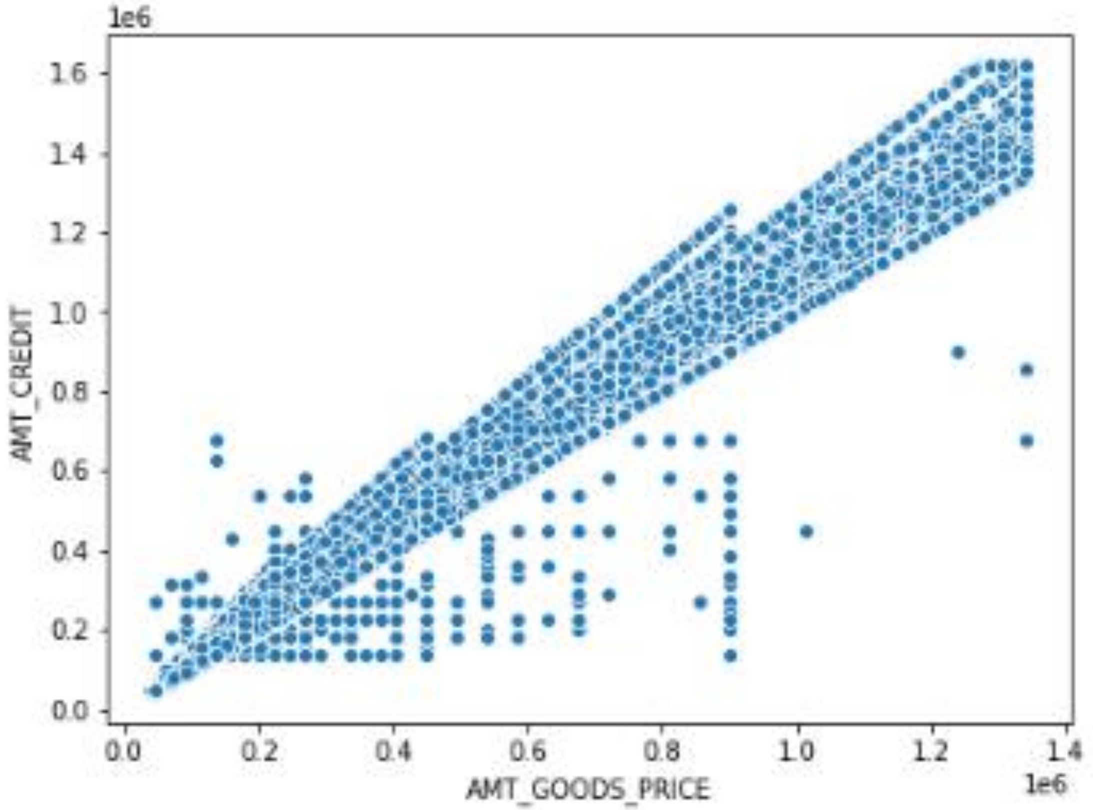
Bivariate / Multivariate analysis of Continuous or Numerical data

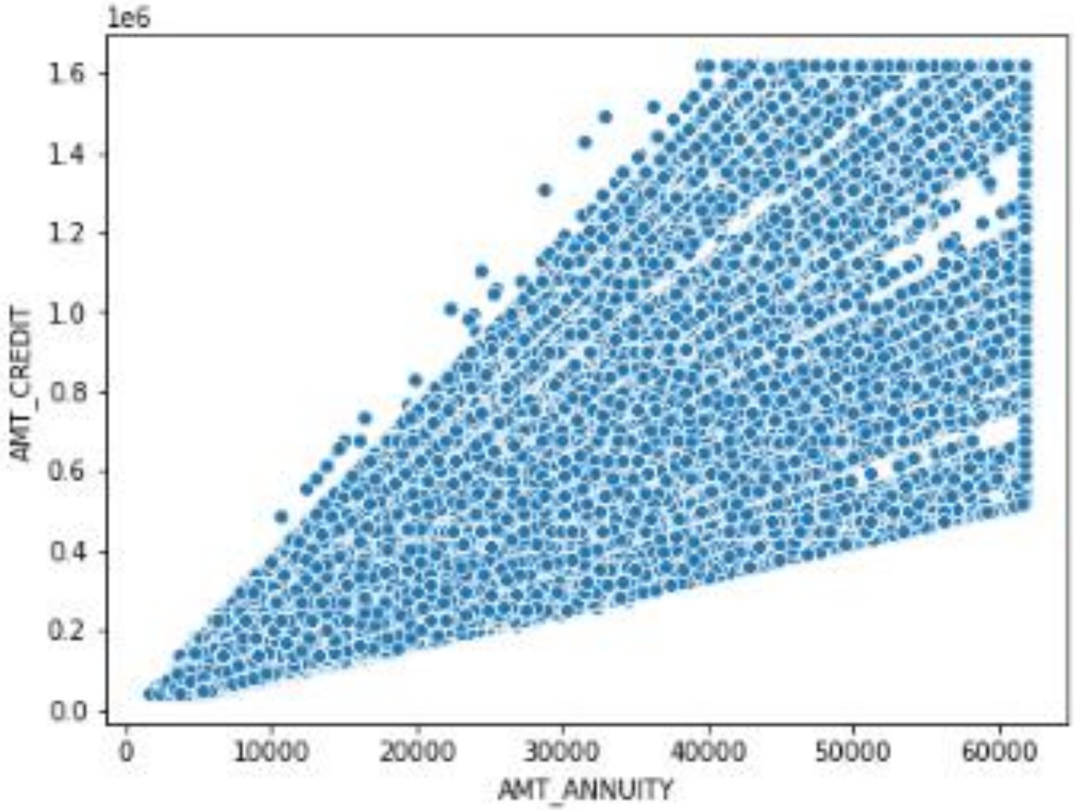
- Plot heatmap of all numerical columns
- Finding Top 10 Correlated values for client with payment difficulties (Target 1) and for all other cases (Target 0)

	col1	col2	Correlation_Target_1
62	AMT_GOODS_PRICE	AMT_CREDIT	0.98
49	AMT_ANNUITY	AMT_GOODS_PRICE	0.75
63	AMT_GOODS_PRICE	AMT_ANNUITY	0.75
111	DAYS_EMPLOYED	DAYS_BIRTH	0.58
98	DAYS_BIRTH	DAYS_REGISTRATION	0.29
99	DAYS_BIRTH	DAYS_ID_PUBLISH	0.25
114	DAYS_EMPLOYED	DAYS_ID_PUBLISH	0.23
220	DAYS_LAST_PHONE_CHANGE	EXT_SOURCE_2	0.21
127	DAYS_REGISTRATION	DAYS_EMPLOYED	0.19
155	EXT_SOURCE_2	REGION_POPULATION_RELATIVE	0.17

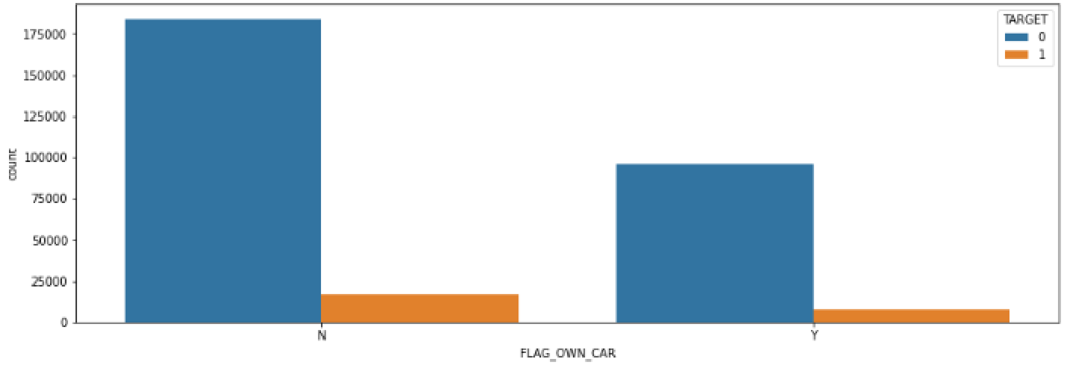
	col1	col2	Correlation_Target_0
62	AMT_GOODS_PRICE	AMT_CREDIT	0.99
63	AMT_GOODS_PRICE	AMT_ANNUITY	0.78
33	AMT_CREDIT	AMT_ANNUITY	0.77
111	DAYS_EMPLOYED	DAYS_BIRTH	0.63
46	AMT_ANNUITY	AMT_INCOME_TOTAL	0.42
61	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.35
31	AMT_CREDIT	AMT_INCOME_TOTAL	0.34
98	DAYS_BIRTH	DAYS_REGISTRATION	0.33
142	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.28
99	DAYS_BIRTH	DAYS_ID_PUBLISH	0.27

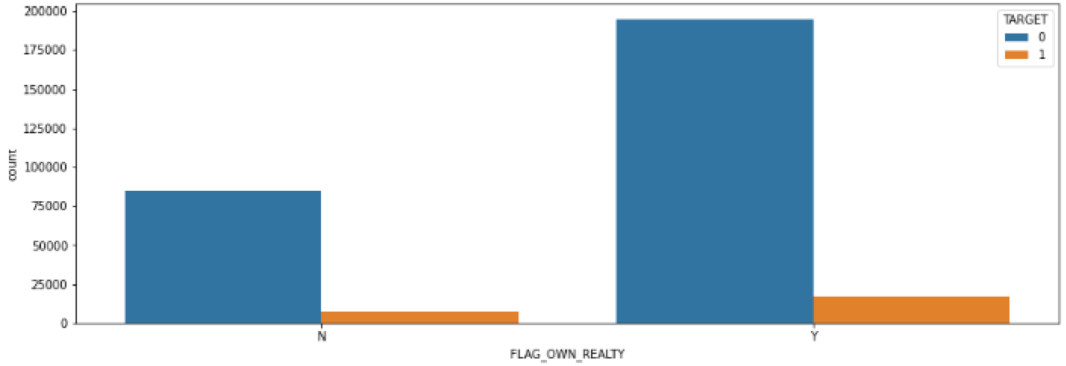
Scatter plot of highly correlated numerical columns

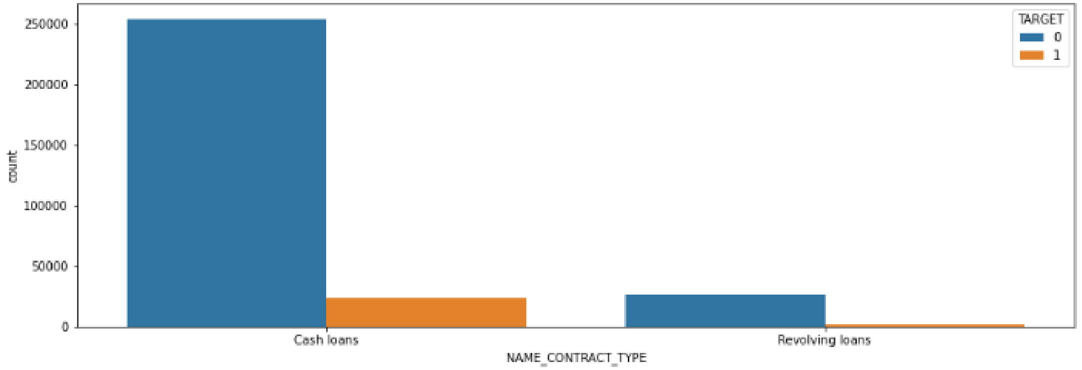


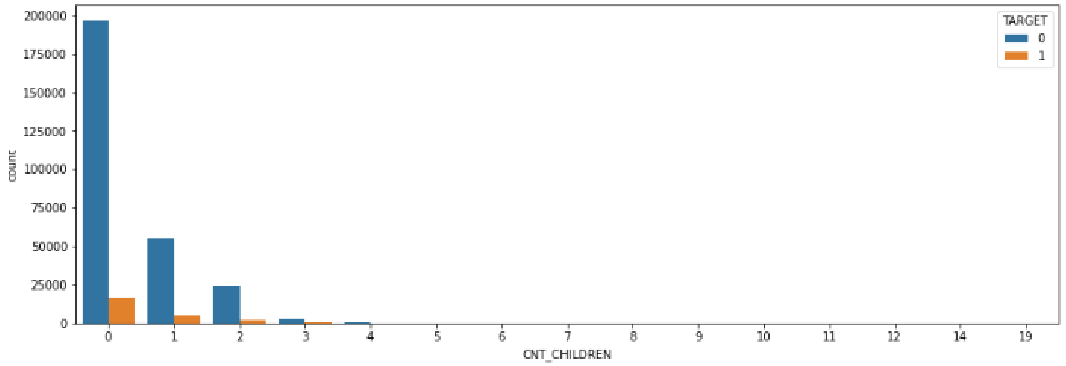


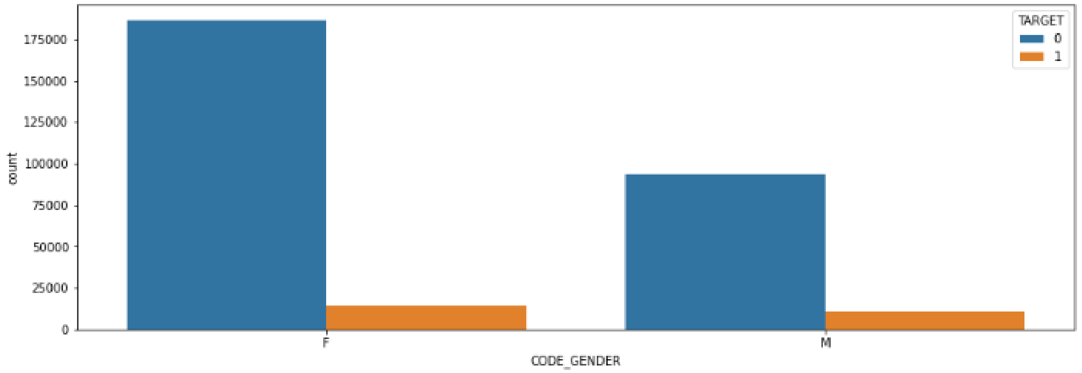
Count plot of Categorical columns with respect to Target data



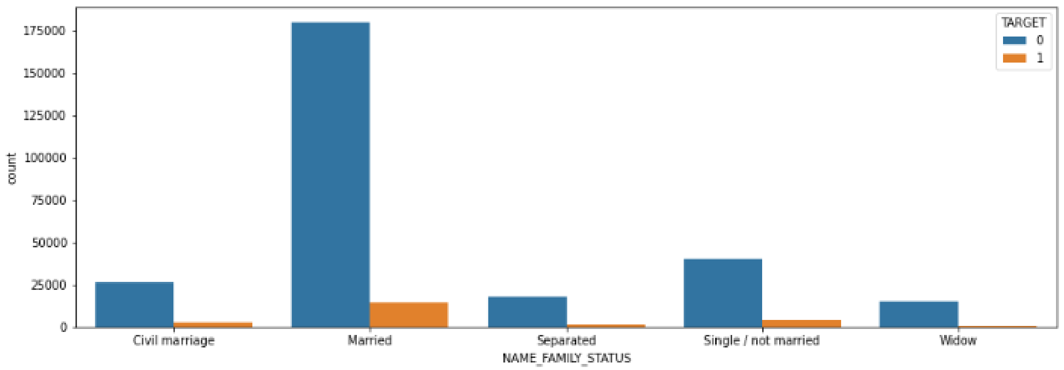


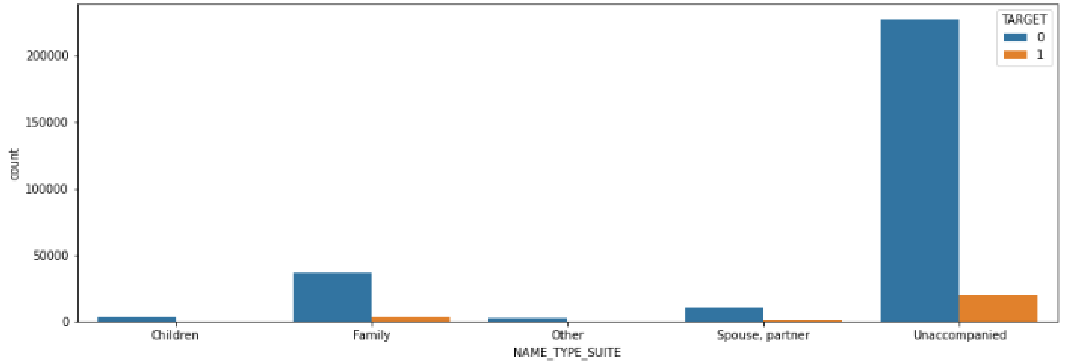


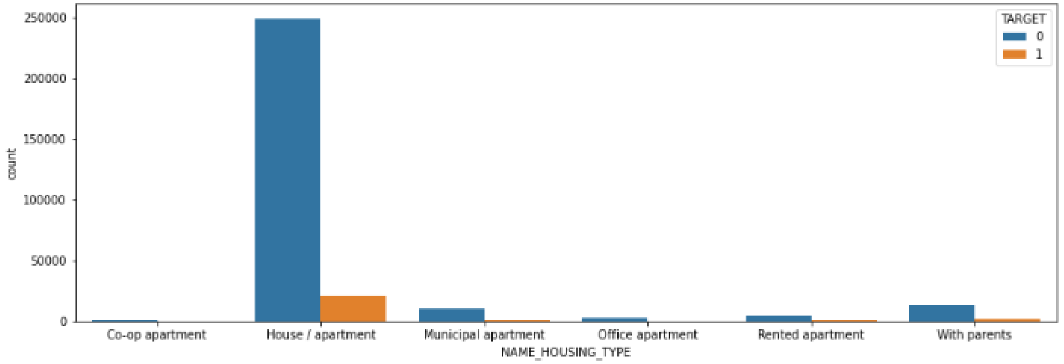


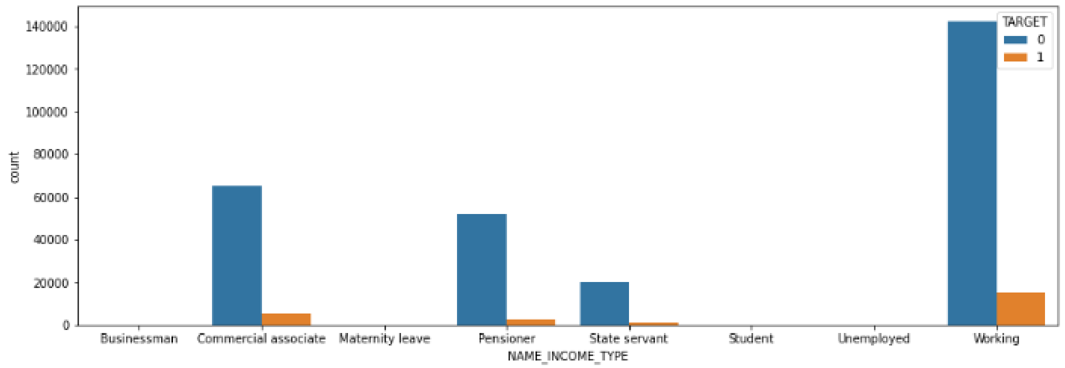


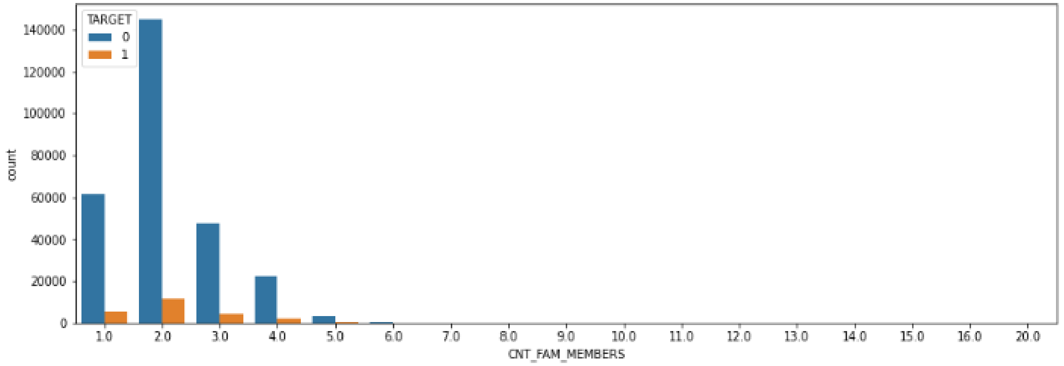
Count plot of Categorical columns with respect to Target data

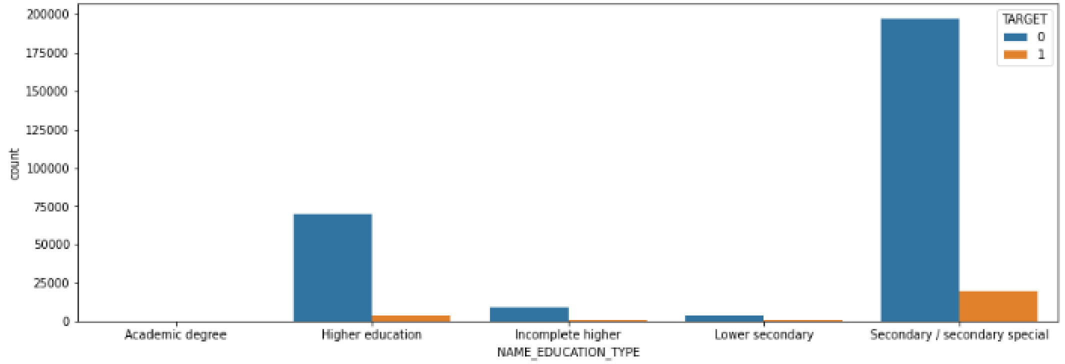




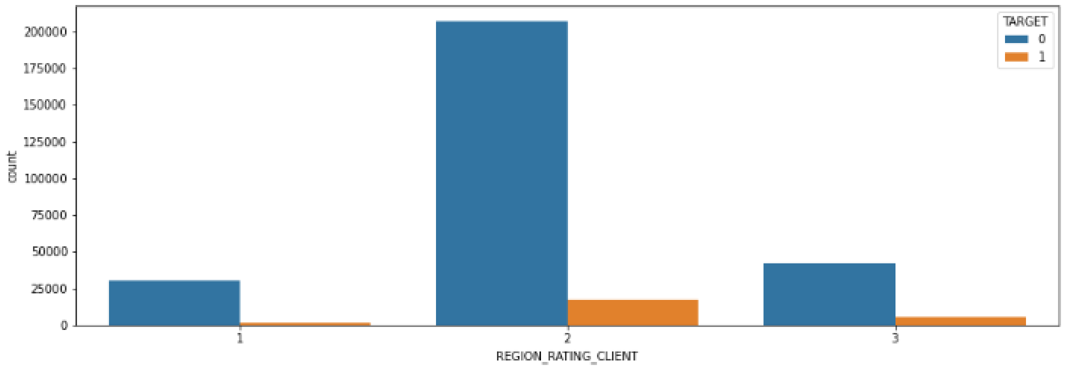


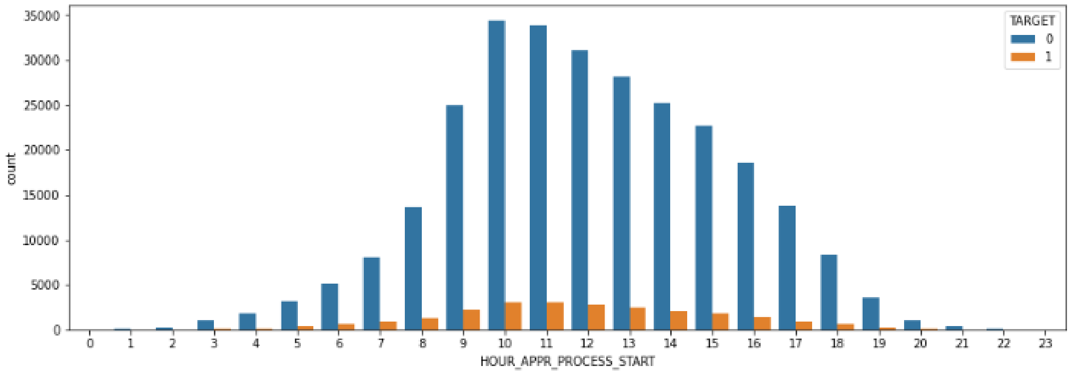


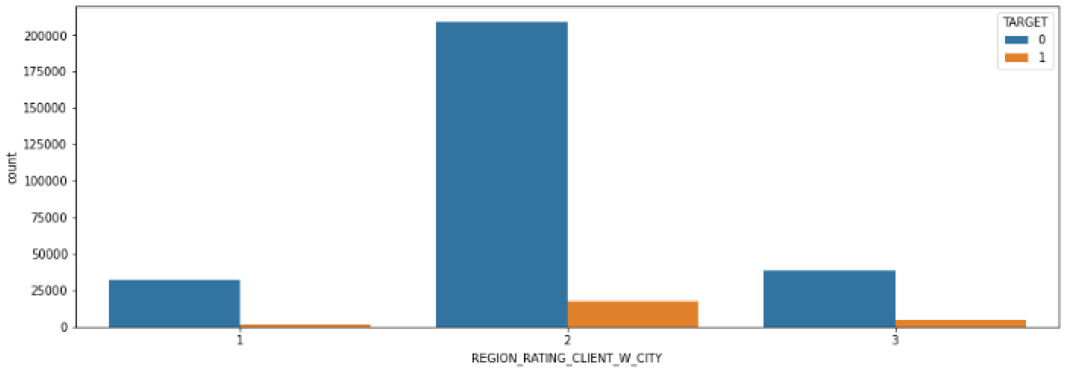


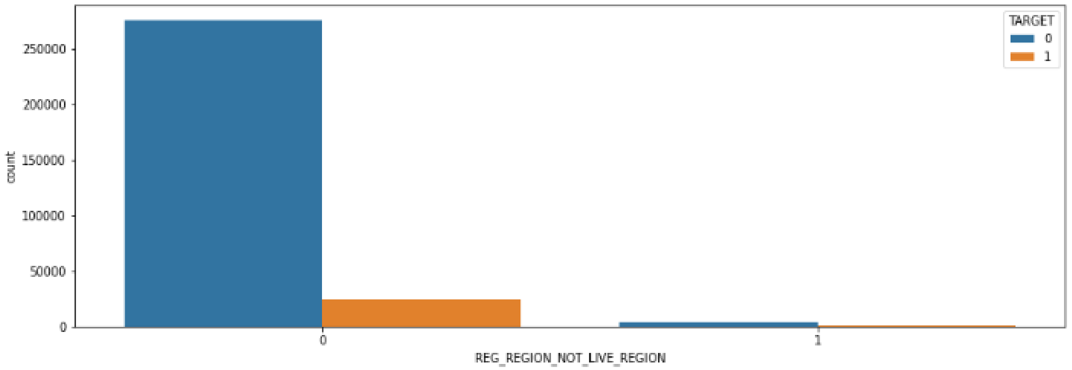


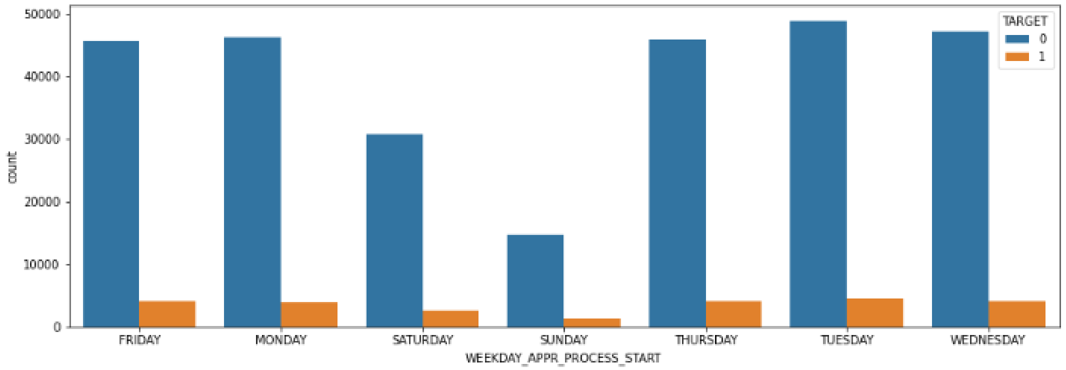
Count plot of Categorical columns with respect to Target data

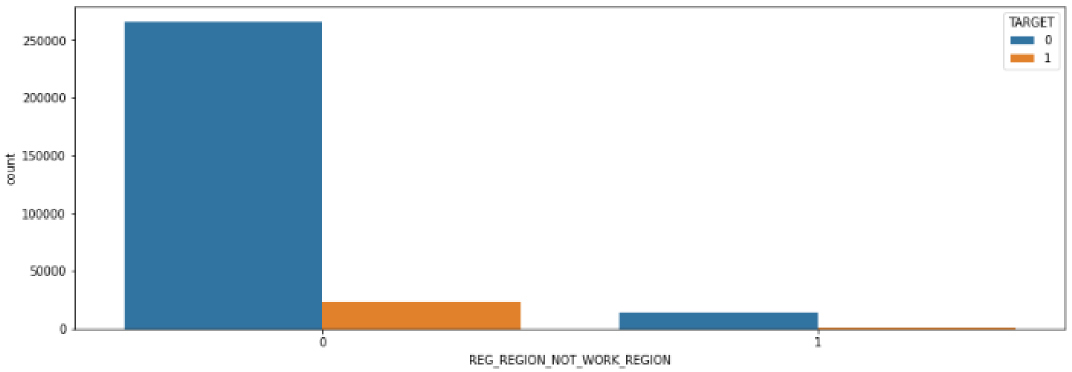




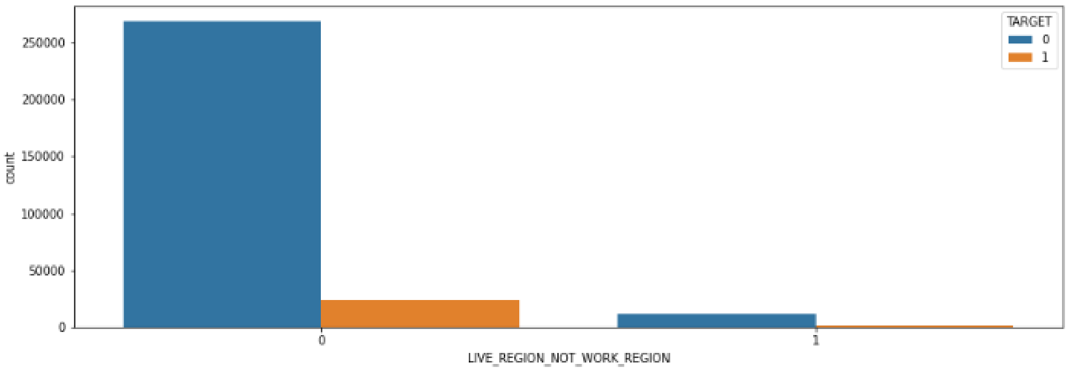


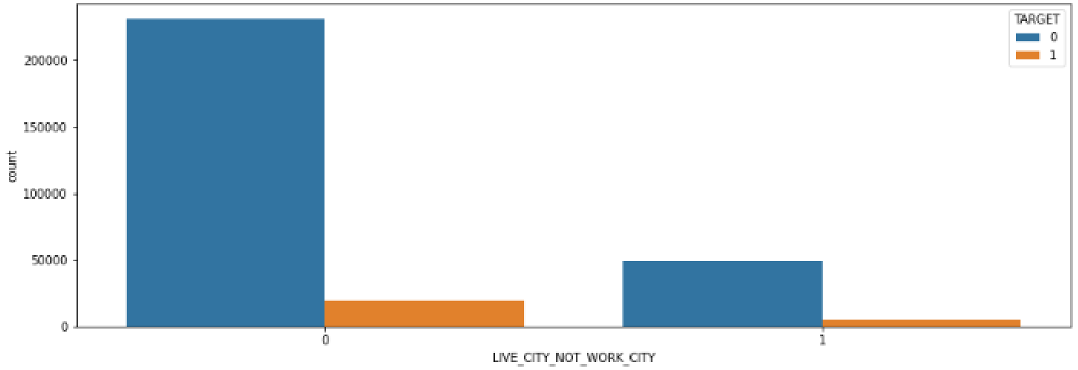


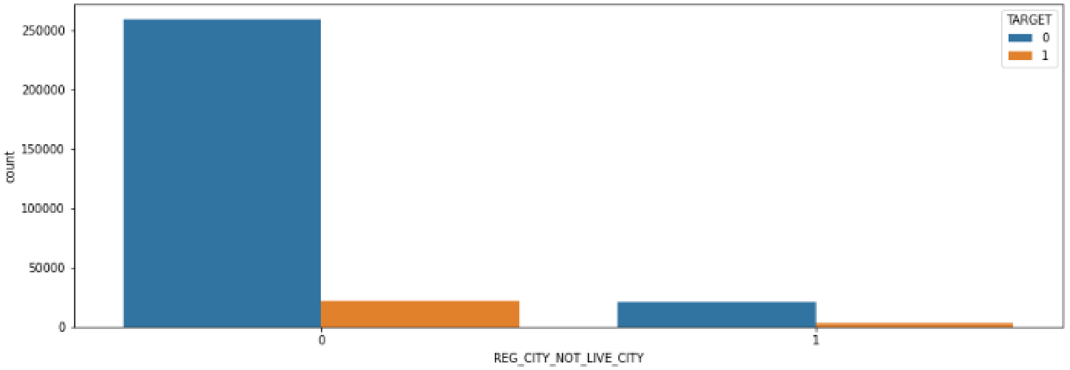


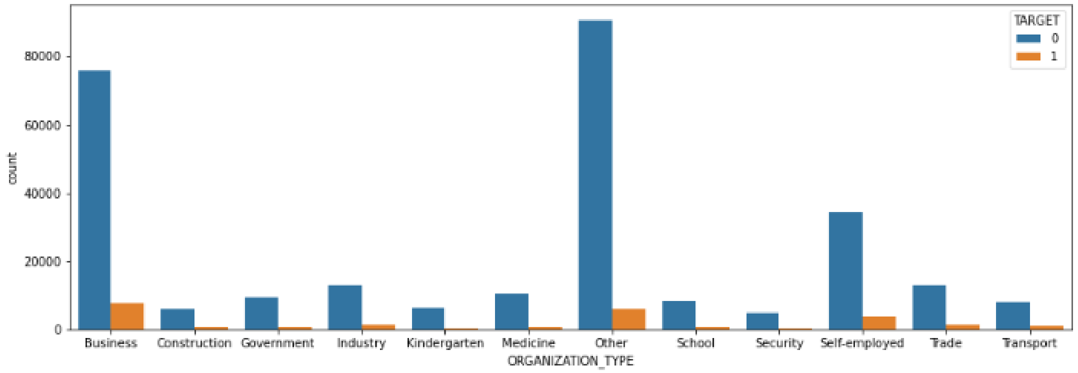


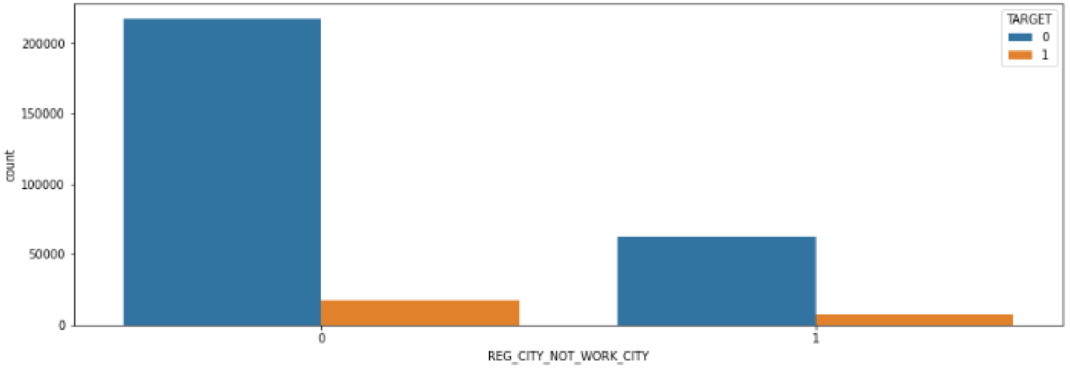
Count plot of Categorical columns with respect to Target data

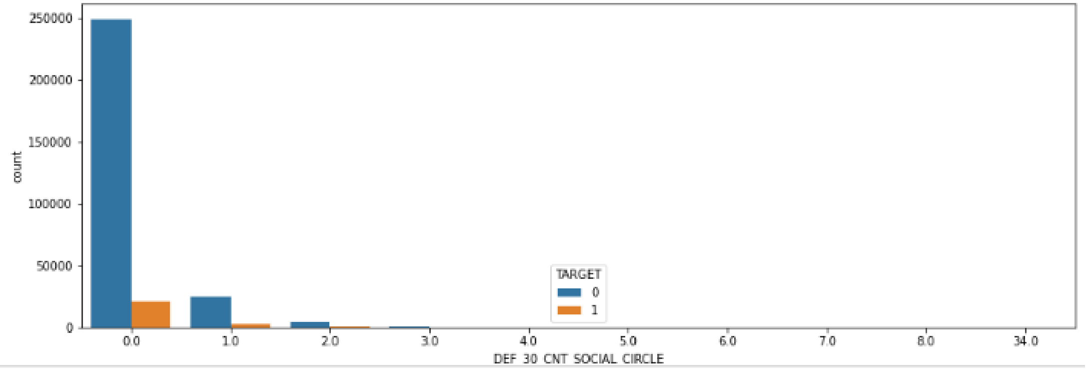




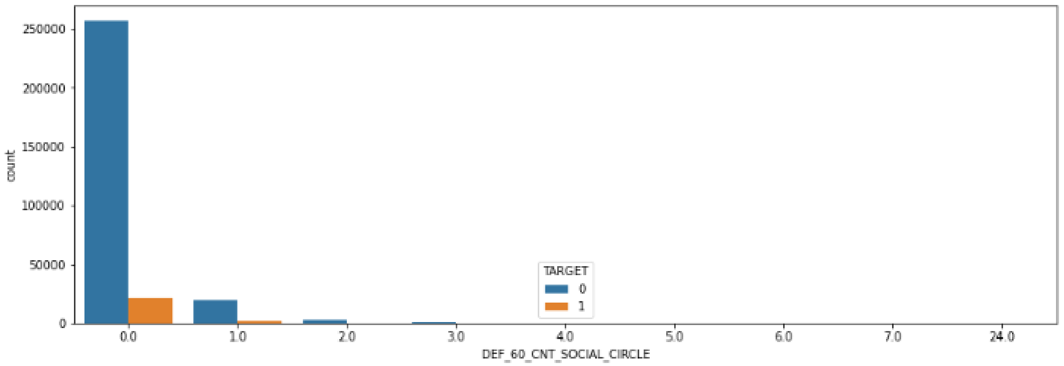


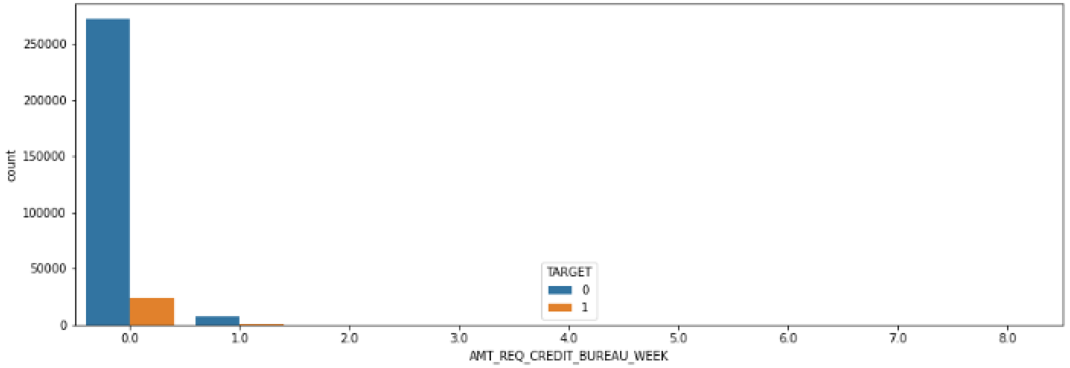


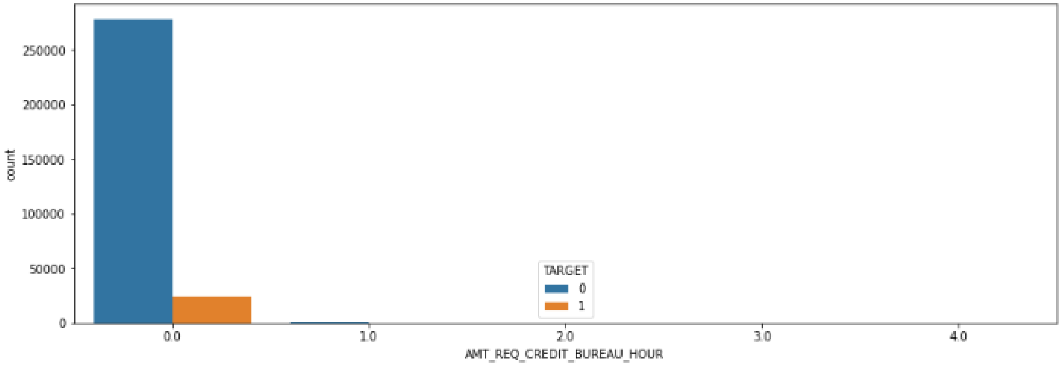


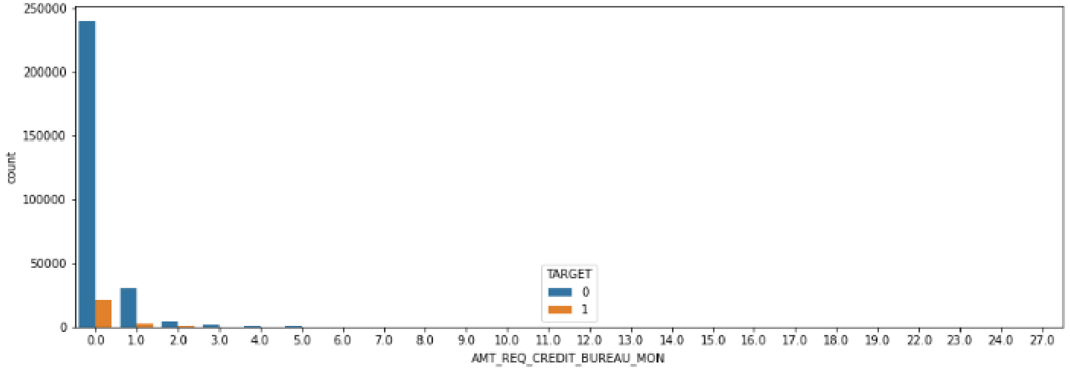


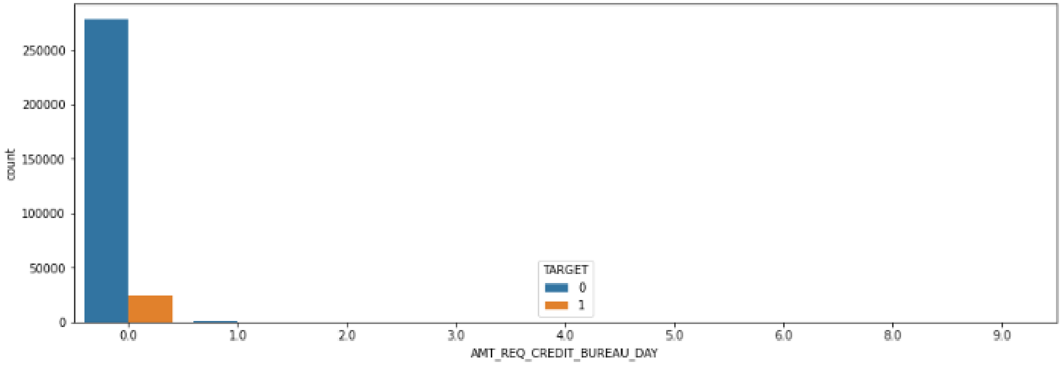
Count plot of Categorical columns with respect to Target data

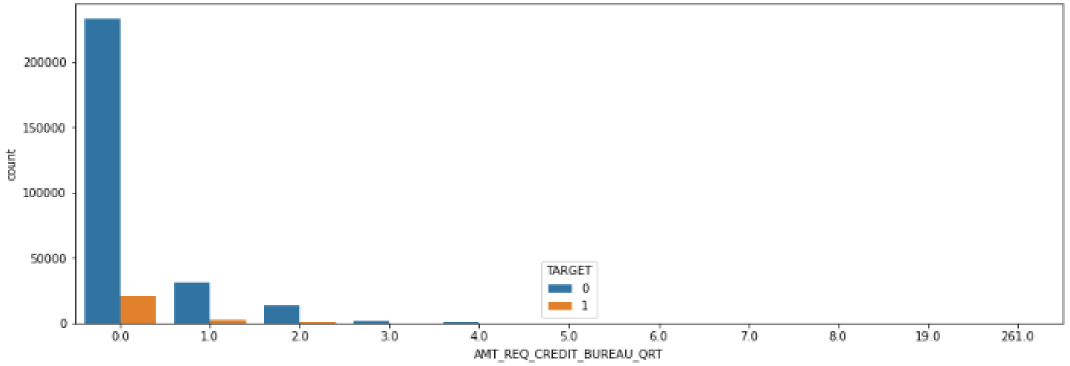




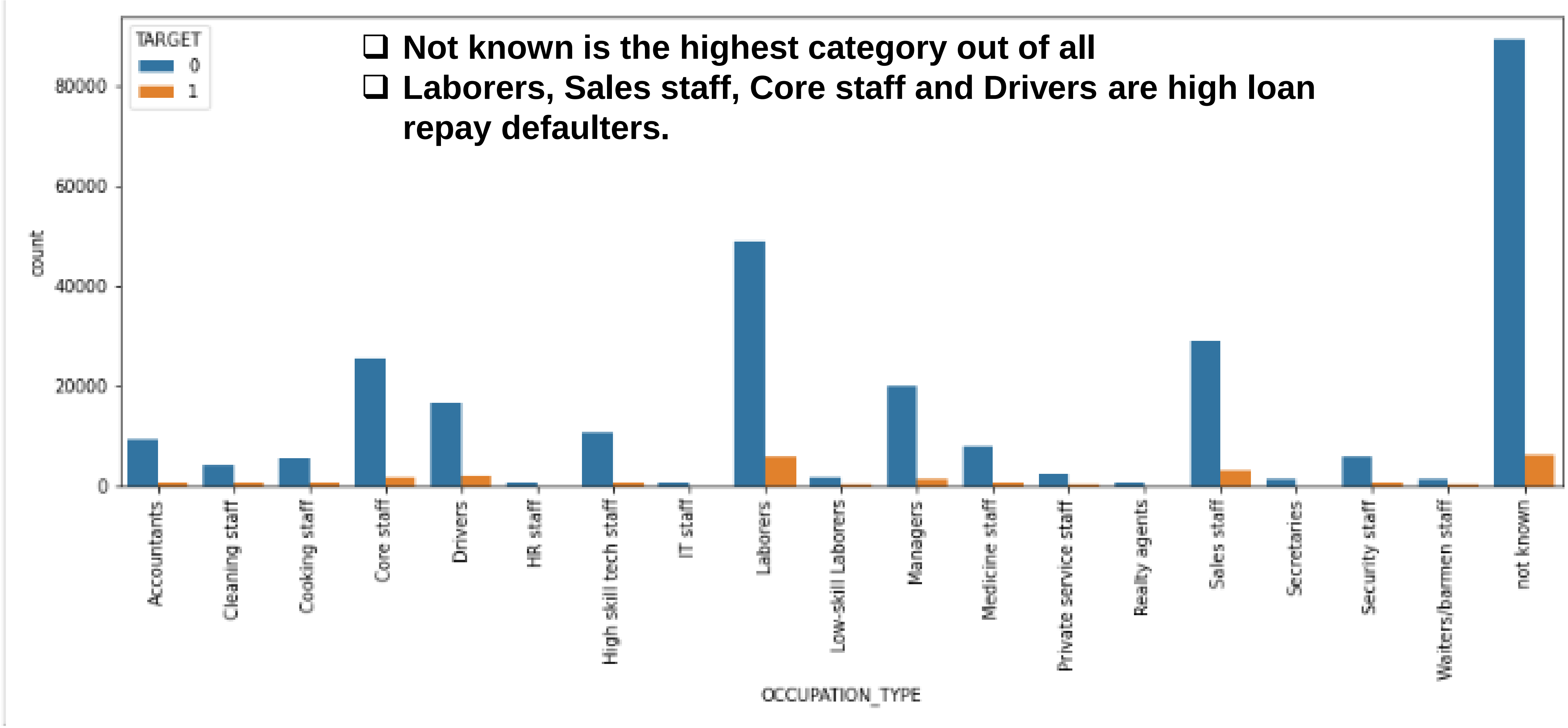








Count plot of Occupation type column with respect to Target data



- **Not known is the highest category out of all**
- **Laborers, Sales staff, Core staff and Drivers are high loan repay defaulters.**

Observations - Occupation type column

Target 0

Laborers	AMT_INCOME_TOTAL	1.000000	0.311138
	AMT_CREDIT	0.311138	1.000000

Target 1

Laborers	AMT_INCOME_TOTAL	1.000000	0.015471
	AMT_CREDIT	0.015471	1.000000

- Above screen shot explains, Occupation type (Laborers) who are loan repay defaulters have very less Amount Income and Amount Credit correlation (less than 2%) and Laborers who are not non-defaulters have better Amount income and Amount credit correlation (31%).

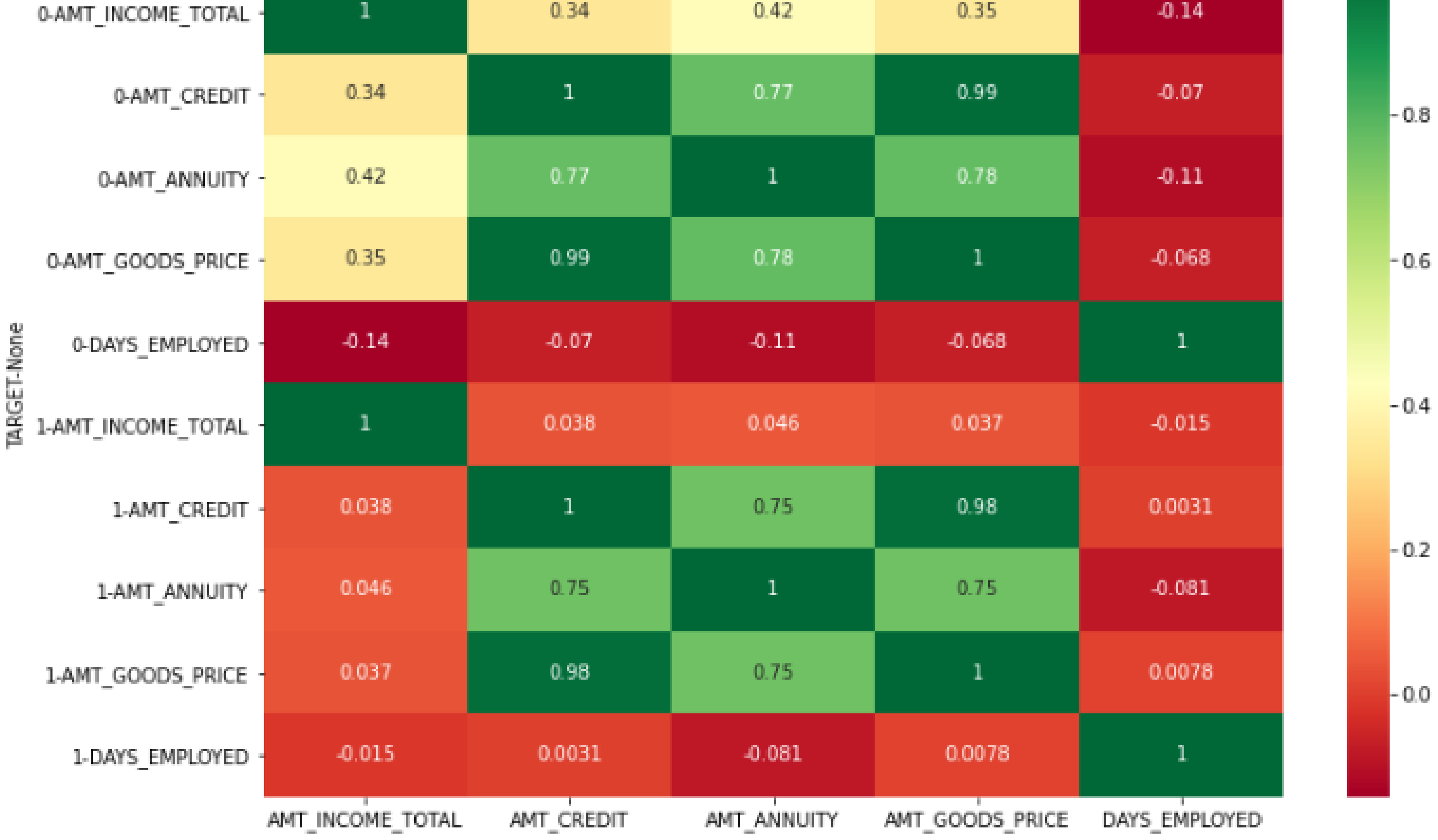
- This could be one of the factor to be considered for Occupation type (Laborers) for Loan approval.

Categorical (Bivariate / Multivariate analysis)

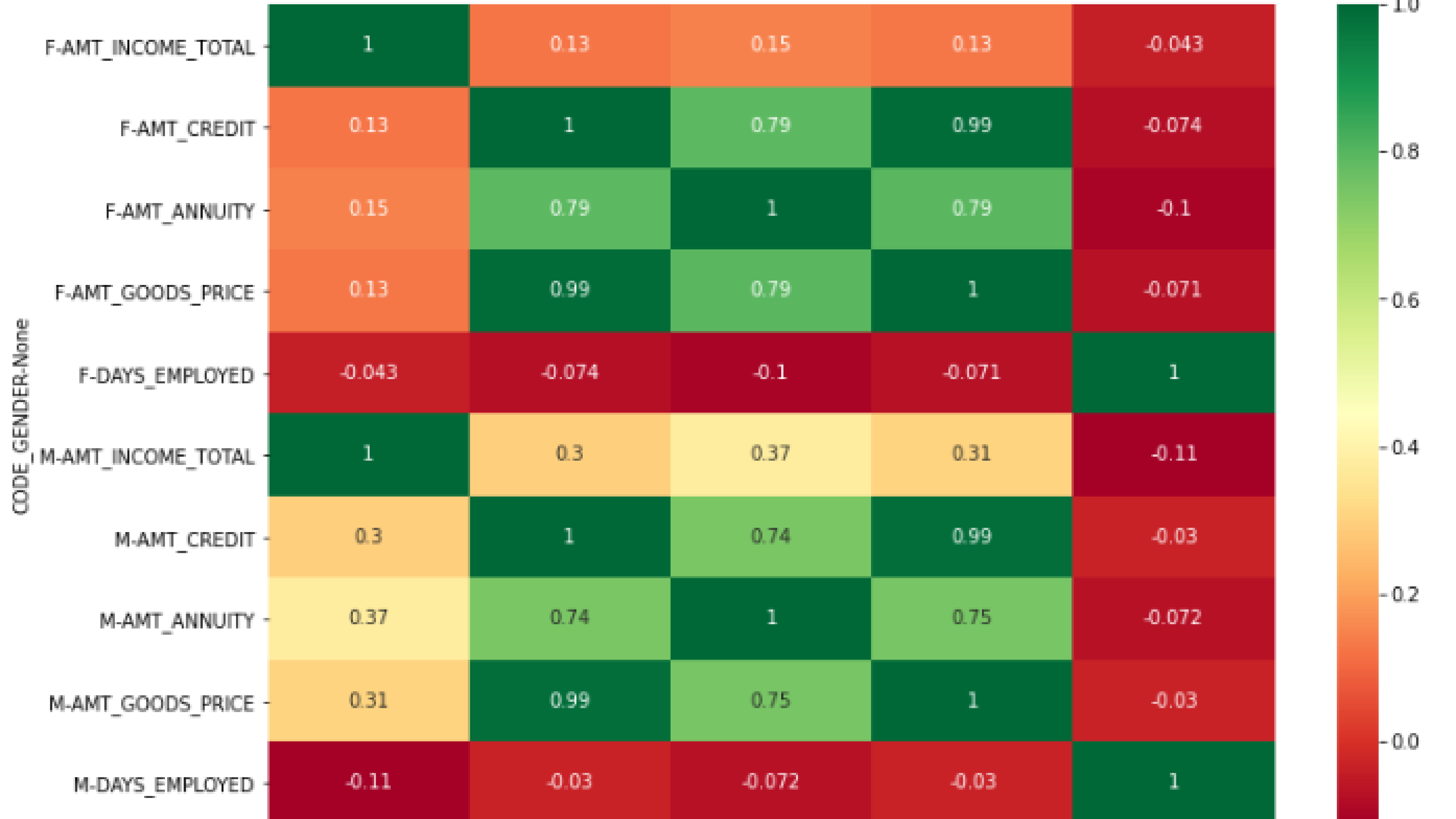
- Created two lists one with numerical and other with categorical columns (critical columns selected based on insights of numerical and categorical analysis)
- Heatmap plot with respect to each categorical column to identified critical numerical columns
- Used pandas groupby function and aggregate to see any findings

```
critical_num_cols = ['AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY','AMT_GOODS_PRICE',
critical_cat_cols = ['TARGET','CODE_GENDER', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE','
for i in critical_cat_cols:
    for j in critical_num_cols:
        print(df.groupby(i)[j].mean())
```

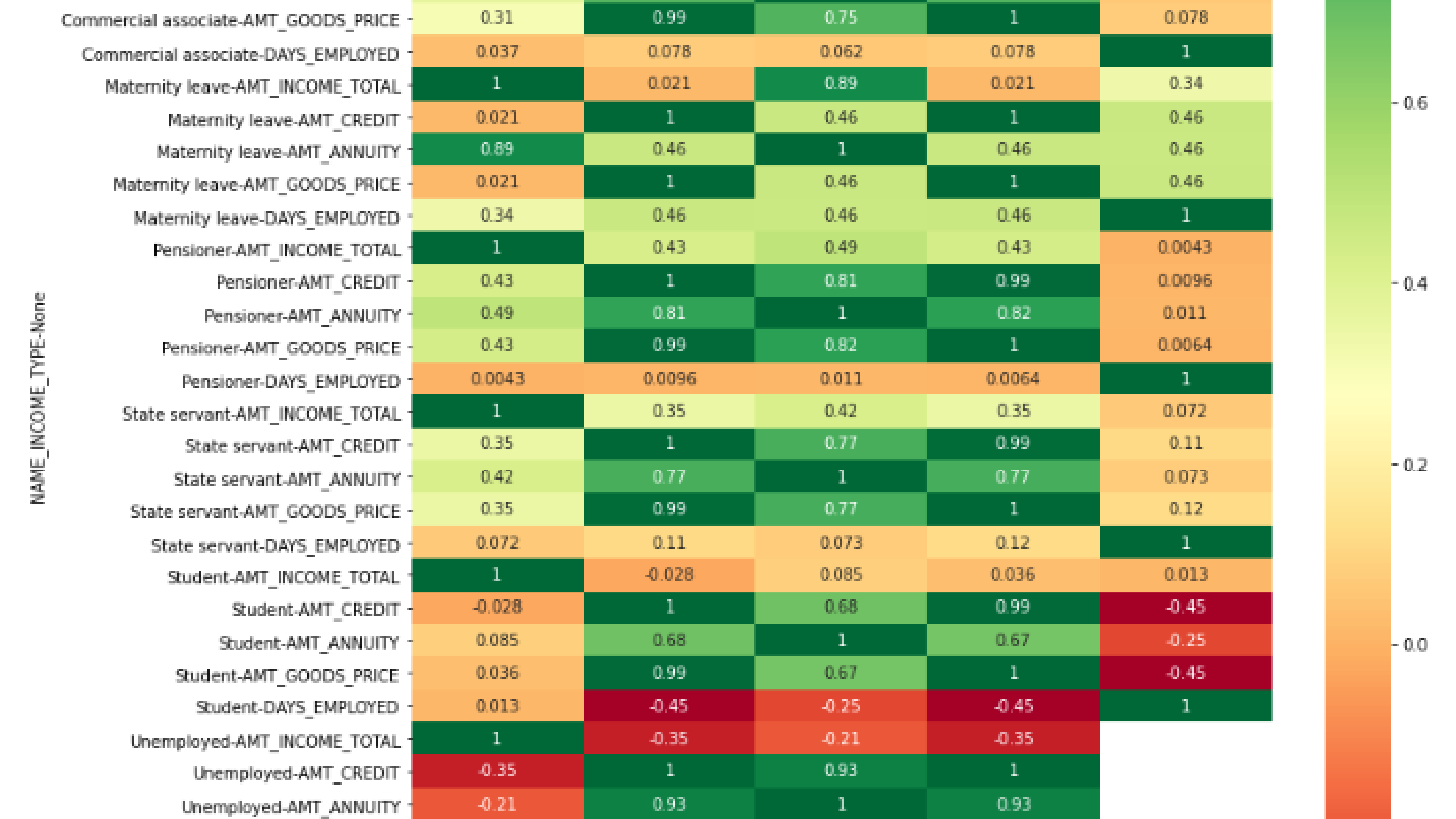
- **Following are the heat map visualization of Categorical Vs Numerical columns**

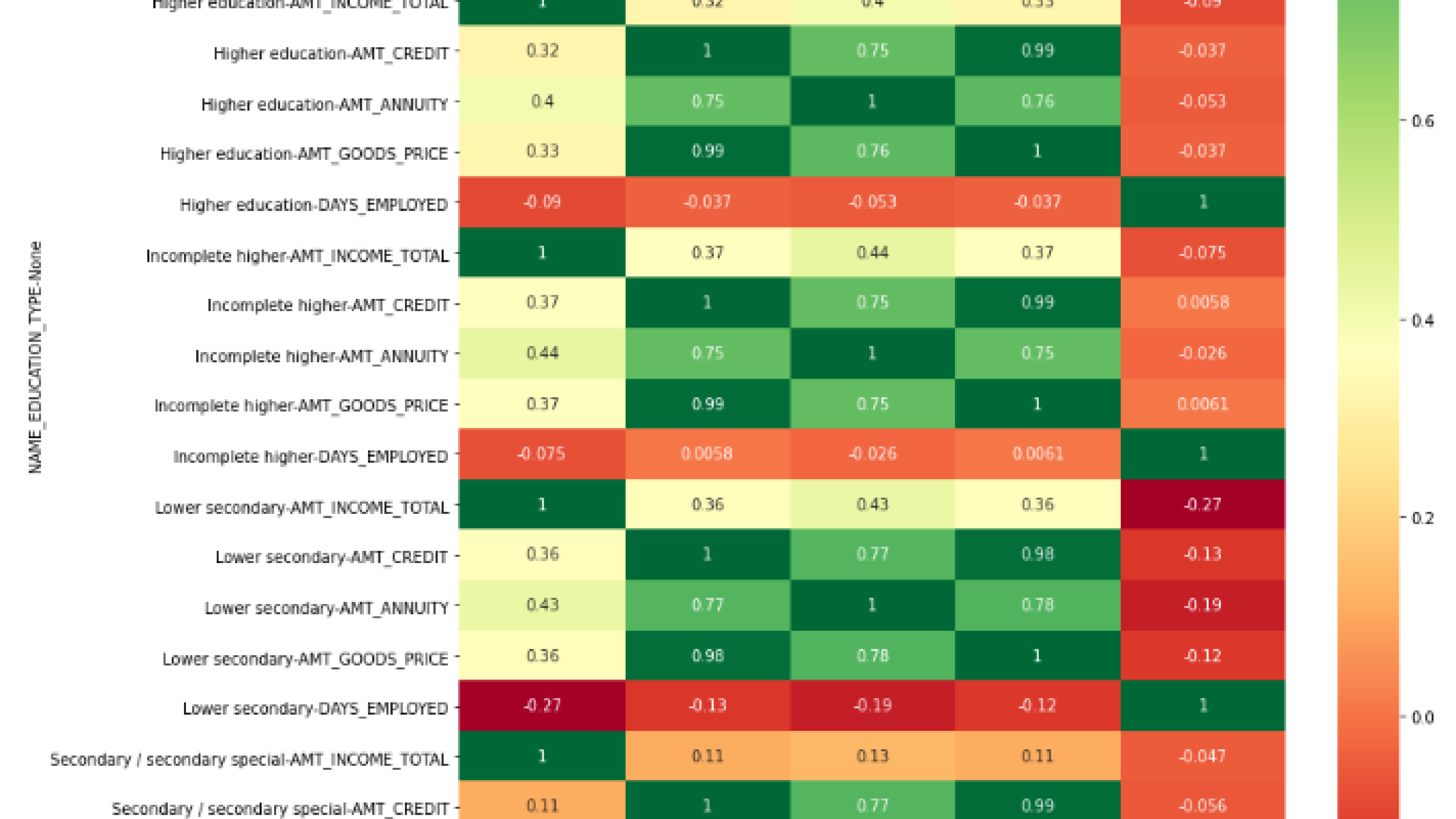


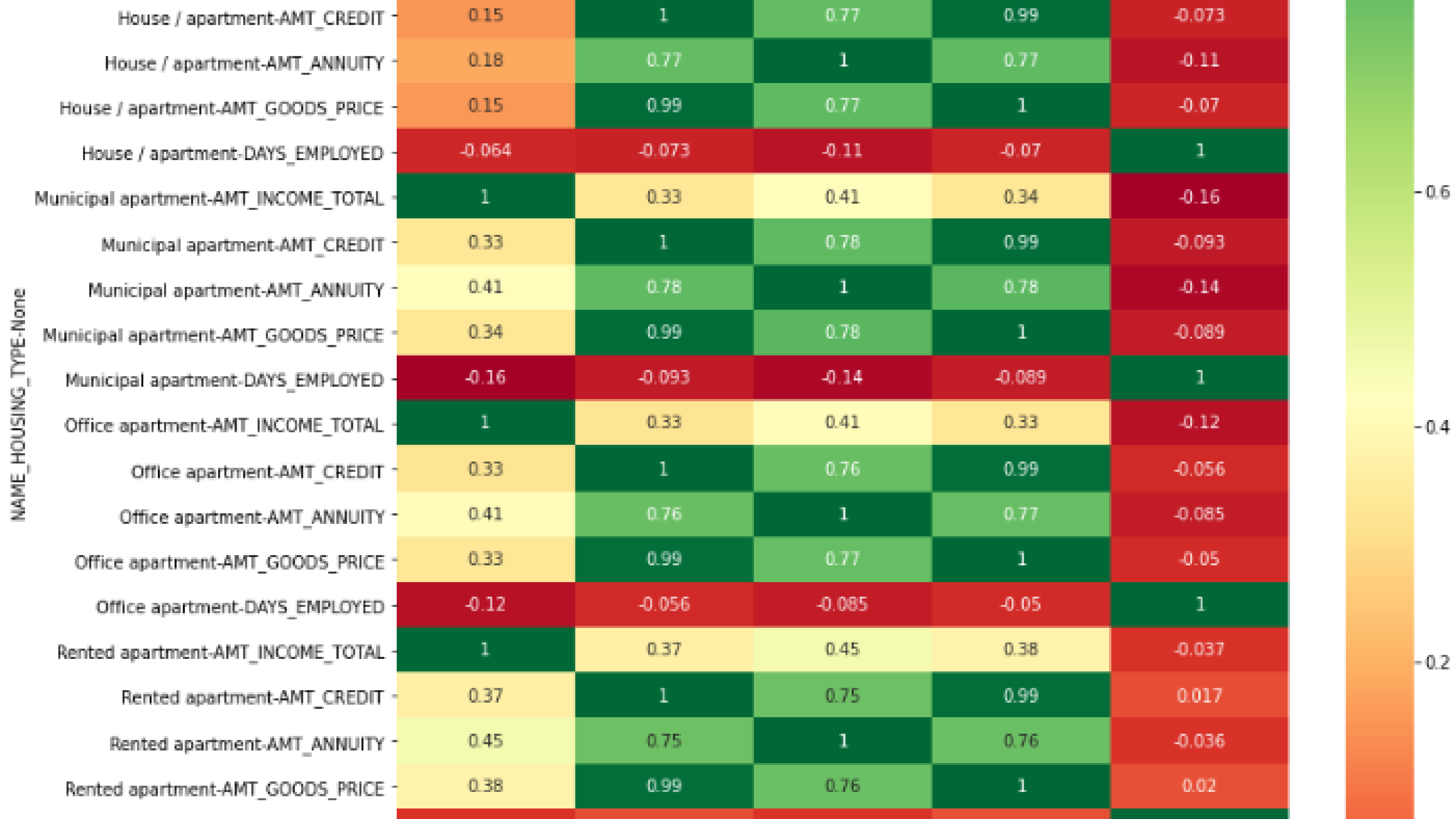
0-AMT_INCOME_TOTAL
0-AMT_CREDIT
0-AMT_ANNUITY
0-AMT_GOODS_PRICE
0-DAYS_EMPLOYED
1-AMT_INCOME_TOTAL
1-AMT_CREDIT
1-AMT_ANNUITY
1-AMT_GOODS_PRICE
1-DAYS_EMPLOYED
TARGET-None
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
DAYS_EMPLOYED
1 0.34 0.42 0.35 -0.14
0.34 1 0.77 0.99 -0.07
0.42 0.77 1 0.78 -0.11
0.35 0.99 0.78 1 -0.068
-0.14 -0.07 -0.11 -0.068 1
1 0.038 0.046 0.037 -0.015
0.038 1 0.75 0.98 0.0031
0.046 0.75 1 0.75 -0.081
0.037 0.98 0.75 1 0.0078
-0.015 0.0031 -0.081 0.0078 1
0.8
0.6
0.4
0.2
0.0

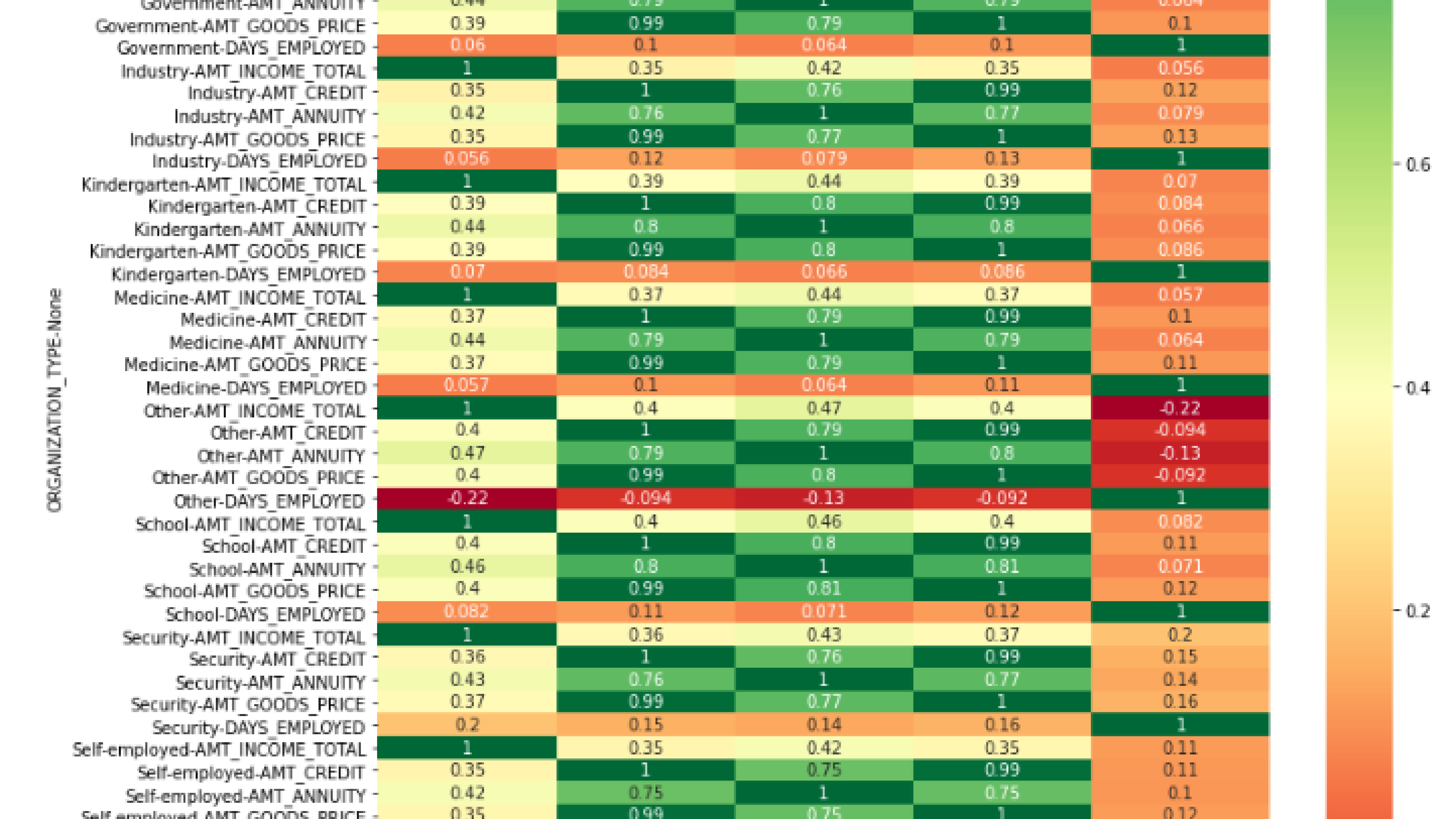


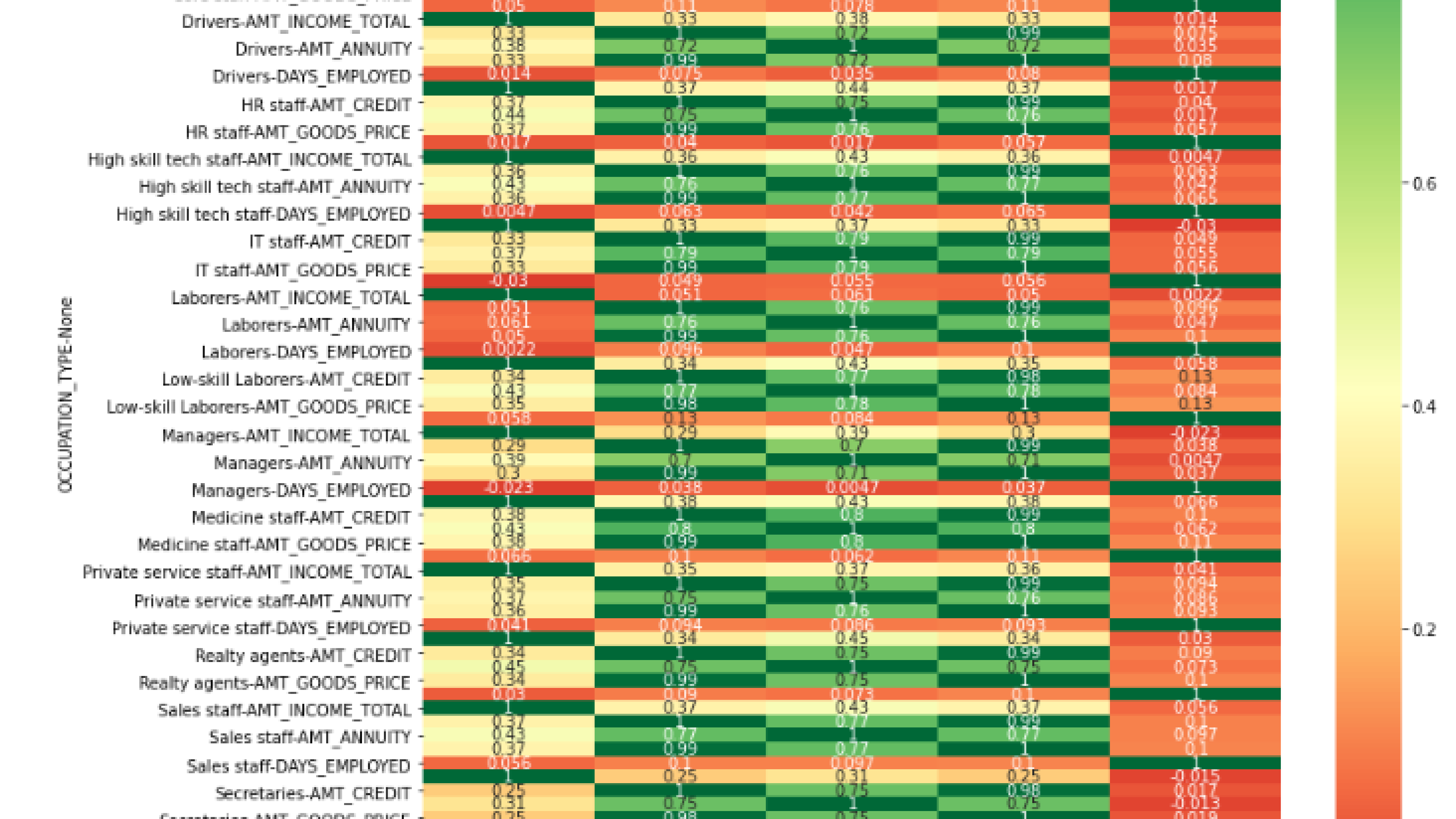
F-AMT_INCOME_TOTAL
F-AMT_CREDIT
F-AMT_ANNUITY
F-AMT_GOODS_PRICE
F-DAYS_EMPLOYED
M-AMT_INCOME_TOTAL
M-AMT_CREDIT
M-AMT_ANNUITY
M-AMT_GOODS_PRICE
M-DAYS_EMPLOYED
CODE_GENDER-None
1 0.13 0.15 0.13 -0.043
0.13 1 0.79 0.99 -0.074
0.15 0.79 1 0.79 -0.1
0.13 0.99 0.79 1 -0.071
-0.043 -0.074 -0.1 -0.071 1
1 0.3 0.37 0.31 -0.11
0.3 1 0.74 0.99 -0.03
0.37 0.74 1 0.75 -0.072
0.31 0.99 0.75 1 -0.03
-0.11 -0.03 -0.072 -0.03 1
1.0
0.8
0.6
0.4
0.2
0.0


NAME_INCOME_TYPE-None
Commercial associate-AMT_GOODS_PRICE
0.31 0.99 0.75 1 0.078
Commercial associate-DAYS_EMPLOYED
0.037 0.078 0.062 0.078 1
Maternity leave-AMT_INCOME_TOTAL
1 0.021 0.89 0.021 0.34
Maternity leave-AMT_CREDIT
0.021 1 0.46 1 0.46
Maternity leave-AMT_ANNUITY
0.89 0.46 1 0.46 0.46
Maternity leave-AMT_GOODS_PRICE
0.021 1 0.46 1 0.46
Maternity leave-DAYS_EMPLOYED
0.34 0.46 0.46 0.46 1
Pensioner-AMT_INCOME_TOTAL
1 0.43 0.49 0.43 0.0043
Pensioner-AMT_CREDIT
0.43 1 0.81 0.99 0.0096
Pensioner-AMT_ANNUITY
0.49 0.81 1 0.82 0.011
Pensioner-AMT_GOODS_PRICE
0.43 0.99 0.82 1 0.0064
Pensioner-DAYS_EMPLOYED
0.0043 0.0096 0.011 0.0064 1
State servant-AMT_INCOME_TOTAL
1 0.35 0.42 0.35 0.072
State servant-AMT_CREDIT
0.35 1 0.77 0.99 0.11
State servant-AMT_ANNUITY
0.42 0.77 1 0.77 0.073
State servant-AMT_GOODS_PRICE
0.35 0.99 0.77 1 0.12
State servant-DAYS_EMPLOYED
0.072 0.11 0.073 0.12 1
Student-AMT_INCOME_TOTAL
1 -0.028 0.085 0.036 0.013
Student-AMT_CREDIT
-0.028 1 0.68 0.99 -0.45
Student-AMT_ANNUITY
0.085 0.68 1 0.67 -0.25
Student-AMT_GOODS_PRICE
0.036 0.99 0.67 1 -0.45
Student-DAYS_EMPLOYED
0.013 -0.45 -0.25 -0.45 1
Unemployed-AMT_INCOME_TOTAL
1 -0.35 -0.21 -0.35
Unemployed-AMT_CREDIT
-0.35 1 0.93 1
Unemployed-AMT_ANNUITY
-0.21 0.93 1 0.93
0.6
0.4
0.2
0.0


NAME_EDUCATION_TYPE-None
Higher education-AMT_CREDIT 0.32 1 0.75 0.99 -0.037
Higher education-AMT_ANNUITY 0.4 0.75 1 0.76 -0.053
Higher education-AMT_GOODS_PRICE 0.33 0.99 0.76 1 -0.037
Higher education-DAYS_EMPLOYED -0.09 -0.037 -0.053 -0.037 1
Incomplete higher-AMT_INCOME_TOTAL 1 0.37 0.44 0.37 -0.075
Incomplete higher-AMT_CREDIT 0.37 1 0.75 0.99 0.0058
Incomplete higher-AMT_ANNUITY 0.44 0.75 1 0.75 -0.026
Incomplete higher-AMT_GOODS_PRICE 0.37 0.99 0.75 1 0.0061
Incomplete higher-DAYS_EMPLOYED -0.075 0.0058 -0.026 0.0061 1
Lower secondary-AMT_INCOME_TOTAL 1 0.36 0.43 0.36 -0.27
Lower secondary-AMT_CREDIT 0.36 1 0.77 0.98 -0.13
Lower secondary-AMT_ANNUITY 0.43 0.77 1 0.78 -0.19
Lower secondary-AMT_GOODS_PRICE 0.36 0.98 0.78 1 -0.12
Lower secondary-DAYS_EMPLOYED -0.27 -0.13 -0.19 -0.12 1
Secondary / secondary special-AMT_INCOME_TOTAL 1 0.11 0.13 0.11 -0.047
Secondary / secondary special-AMT_CREDIT 0.11 1 0.77 0.99 -0.056
0.6
0.4
0.2
0.0


NAME_HOUSING_TYPE-None
House / apartment-AMT_CREDIT
0.15 1 0.77 0.99 -0.073
House / apartment-AMT_ANNUITY
0.18 0.77 1 0.77 -0.11
House / apartment-AMT_GOODS_PRICE
0.15 0.99 0.77 1 -0.07
House / apartment-DAYS_EMPLOYED
-0.064 -0.073 -0.11 -0.07 1
Municipal apartment-AMT_INCOME_TOTAL
1 0.33 0.41 0.34 -0.16
Municipal apartment-AMT_CREDIT
0.33 1 0.78 0.99 -0.093
Municipal apartment-AMT_ANNUITY
0.41 0.78 1 0.78 -0.14
Municipal apartment-AMT_GOODS_PRICE
0.34 0.99 0.78 1 -0.089
Municipal apartment-DAYS_EMPLOYED
-0.16 -0.093 -0.14 -0.089 1
Office apartment-AMT_INCOME_TOTAL
1 0.33 0.41 0.33 -0.12
Office apartment-AMT_CREDIT
0.33 1 0.76 0.99 -0.056
Office apartment-AMT_ANNUITY
0.41 0.76 1 0.77 -0.085
Office apartment-AMT_GOODS_PRICE
0.33 0.99 0.77 1 -0.05
Office apartment-DAYS_EMPLOYED
-0.12 -0.056 -0.085 -0.05 1
Rented apartment-AMT_INCOME_TOTAL
1 0.37 0.45 0.38 -0.037
Rented apartment-AMT_CREDIT
0.37 1 0.75 0.99 0.017
Rented apartment-AMT_ANNUITY
0.45 0.75 1 0.76 -0.036
Rented apartment-AMT_GOODS_PRICE
0.38 0.99 0.76 1 0.02
0.6
0.4
0.2


ORGANIZATION_TYPE-None
Government-AMT_GOODS_PRICE 0.39 0.99 0.79 1 0.1
Government-DAYS_EMPLOYED 0.06 0.1 0.064 0.1 1
Industry-AMT_INCOME_TOTAL 1 0.35 0.42 0.35 0.056
Industry-AMT_CREDIT 0.35 1 0.76 0.99 0.12
Industry-AMT_ANNUITY 0.42 0.76 1 0.77 0.079
Industry-AMT_GOODS_PRICE 0.35 0.99 0.77 1 0.13
Industry-DAYS_EMPLOYED 0.056 0.12 0.079 0.13 1
Kindergarten-AMT_INCOME_TOTAL 1 0.39 0.44 0.39 0.07
Kindergarten-AMT_CREDIT 0.39 1 0.8 0.99 0.084
Kindergarten-AMT_ANNUITY 0.44 0.8 1 0.8 0.066
Kindergarten-AMT_GOODS_PRICE 0.39 0.99 0.8 1 0.086
Kindergarten-DAYS_EMPLOYED 0.07 0.084 0.066 0.086 1
Medicine-AMT_INCOME_TOTAL 1 0.37 0.44 0.37 0.057
Medicine-AMT_CREDIT 0.37 1 0.79 0.99 0.1
Medicine-AMT_ANNUITY 0.44 0.79 1 0.79 0.064
Medicine-AMT_GOODS_PRICE 0.37 0.99 0.79 1 0.11
Medicine-DAYS_EMPLOYED 0.057 0.1 0.064 0.11 1
Other-AMT_INCOME_TOTAL 1 0.4 0.47 0.4 -0.22
Other-AMT_CREDIT 0.4 1 0.79 0.99 -0.094
Other-AMT_ANNUITY 0.47 0.79 1 0.8 -0.13
Other-AMT_GOODS_PRICE 0.4 0.99 0.8 1 -0.092
Other-DAYS_EMPLOYED -0.22 -0.094 -0.13 -0.092 1
School-AMT_INCOME_TOTAL 1 0.4 0.46 0.4 0.082
School-AMT_CREDIT 0.4 1 0.8 0.99 0.11
School-AMT_ANNUITY 0.46 0.8 1 0.81 0.071
School-AMT_GOODS_PRICE 0.4 0.99 0.81 1 0.12
School-DAYS_EMPLOYED 0.082 0.11 0.071 0.12 1
Security-AMT_INCOME_TOTAL 1 0.36 0.43 0.37 0.2
Security-AMT_CREDIT 0.36 1 0.76 0.99 0.15
Security-AMT_ANNUITY 0.43 0.76 1 0.77 0.14
Security-AMT_GOODS_PRICE 0.37 0.99 0.77 1 0.16
Security-DAYS_EMPLOYED 0.2 0.15 0.14 0.16 1
Self-employed-AMT_INCOME_TOTAL 1 0.35 0.42 0.35 0.11
Self-employed-AMT_CREDIT 0.35 1 0.75 0.99 0.11
Self-employed-AMT_ANNUITY 0.42 0.75 1 0.75 0.1
0.6
0.4
0.2


Drivers-AMT_INCOME_TOTAL
Drivers-AMT_ANNUITY
Drivers-DAYS_EMPLOYED
HR staff-AMT_CREDIT
HR staff-AMT_GOODS_PRICE
High skill tech staff-AMT_INCOME_TOTAL
High skill tech staff-AMT_ANNUITY
High skill tech staff-DAYS_EMPLOYED
IT staff-AMT_CREDIT
IT staff-AMT_GOODS_PRICE
Laborers-AMT_INCOME_TOTAL
Laborers-AMT_ANNUITY
Laborers-DAYS_EMPLOYED
Low-skill Laborers-AMT_CREDIT
Low-skill Laborers-AMT_GOODS_PRICE
Managers-AMT_INCOME_TOTAL
Managers-AMT_ANNUITY
Managers-DAYS_EMPLOYED
Medicine staff-AMT_CREDIT
Medicine staff-AMT_GOODS_PRICE
Private service staff-AMT_INCOME_TOTAL
Private service staff-AMT_ANNUITY
Private service staff-DAYS_EMPLOYED
Realty agents-AMT_CREDIT
Realty agents-AMT_GOODS_PRICE
Sales staff-AMT_INCOME_TOTAL
Sales staff-AMT_ANNUITY
Sales staff-DAYS_EMPLOYED
Secretaries-AMT_CREDIT
OCCUPATION_TYPE-None
0.6
0.4
0.2

Numerical / Continuous data (Univariate observations)

- Client with payment difficulties income have majorly distributed between 50000 and 200000
- Credit amount of the loan dist. plot seems similar observation between client with payment difficulties and not.
- Amount Income - '<100000', '100000-200000' contributes for more loan applications and more likely to default
- Amount Income - Clients who have above 300000 less likely to loan repay default
- Amount Credit - Maximum clients have credit range above 500000.

Numerical Bivariate / Multivariate Analysis (Numerical vs Numerical)

- Positive correlation between Credit amount of the loan (AMT_CREDIT) and for consumer loans it is the price of the goods for which the loan is given (AMT_GOODS_PRICE)
- Positive correlation between Credit amount of the loan (AMT CREDIT) and Loan annuity (AMT_ANNUITY)
- AMT_INCOME_TOTAL have better positive correlation with AMT CREDIT, AMT ANNUTIY and AMT GOODS PRICE for the applicants who don't have payment issues than clients with payment difficulties
- Positive correlation between Days employed and Days birth

Categorical Data Analysis - Observations (Univariate)

- Cash loans disbursed more than revolving loans
- Occupation Type - not known is the highest category out of all
- Occupation Type - Laborers, Sales staff, core staff and drivers are high loan repay defaulters.
- Overall, more female taken loan than male and ratio of all other cases and client with payment difficulties seems more for male than female
- Most of the clients don't own a car
- Income Type - Working and commercial associate availed loans than any other category
- Secondary education applicants were greater than any other category of education type
- Most of the applicants were married
- Most of the applicants live in a house or apartment
- Two or less family members were dominant for the applicants
- Business and other organization applicants are more.

Summary - Correlation data (Categorical vs Numerical)

- Clients with less median total income are more likely to default
- Clients with high Credit amount are less likely to have payment difficulty or default
- Clients with greater birth days are less likely to default
- Clients with amount annuity greater than 25000 are less likely to default
- People with house or apartment tend to take more loans
- Married tend to take more Loan as compared to other categories
- Secondary/special educated people are applying loans in high in number
- Occupation type (Laborers) who are loan repay defaulters have very less Amount Income and Amount Credit correlation (less than 2%) and Laborers who are not non-defaulters have better Amount income and Amount credit correlation (31%).

Previous Application Data Analysis

- Consumer loans approved count is greater than other loans and followed by cash loans
- Cash through bank is the most used payment type (Payment method that client chose to pay for the previous application).
- Unaccompanied clients are more likely get loan approval
- Repeater has highest number of approved loans.
- Middle NAME_YIELD_GROUP has highest approval.
- For Medium AMT_INCOME_TOTAL_bin the approval is highest.

Previous application data Observations - Category vs Numerical data

- Amount annuity for previous applicants are less compared to refused loan applicants
- Clients who asked for lesser median credit on the previous application have more approval rate.
- Amount credit previous has highest refused cases and amount credit is similar for all 4 cases.
- Selling area of seller place of the previous application range (0 to 150) have higher loan approval.
- Time spent in unused offer is higher as compared to other categories. So bank should reduce time spent on unused offer.

Merged data - Categorical univariate and multivariate observations

- Gender - More female applicants availed loan than male and ratio of (non-defaulter and defaulter) is higher for male than female
- Income type - working category have higher number of applicants and defaulters
- Family status - Married people are more likely to default than other category
- Housing type - House / apartment category customers are more likely to default.
- Occupation type - Occupation not known for maximum number of clients
- Occupation type - Next to not known, Laborers are more like to deafult
- Organization type - Business clients, other and self employed have more loan applicants
- Income range - less than 300000 income are more likely to default
- Contract type - More Cash and consumer loans availed than revolving loans
- Contract type - Out of consumer and cash loans, cash loans have more loan repay defaulters.
- Client type - More loan applicants from repeater

Summary

- Following are most important parameters to be considered for approving loan application
  - Amount Income Total (lesser income more likely to have payment difficulty) / less median total income are more likely to default.
  - Amount Credit - Clients with high Credit amount are less likely to have payment difficulty or default
  - Days Employed - higher the number more likely to repay the loan
  - Higher positive relationship of (Amount Annuity / Amount Income Total, Amount Goods Price / Amount Income Total and Amount Credit / Amount Income Total) these parameters will help to approve loan.
  - Gender - Male applicants likely to have payment difficulty than female
  - Name Education Type
  - Name Housing Type