# Credit EDA Case Study

By
SYED ASHIK

# Problem statement

- This case study aims to identify patterns which indicate if a client is eligible for a loan that he requested based on some categories and to look that the client is more interested in repaying the loan and not being a defaulter. It's main goal is in 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.

- The company wants to understand the driving factors behind loan default, i.e., the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

# About the dataset

This dataset has 3 files as explained below:

1. '**application_data.csv**' contains all the information of the client at the time of application which holds the data is about whether a client has payment difficulties.

2. '**previous_application.csv**' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

3. '**columns_description.csv**' is data dictionary which describes the meaning of the variables.

# Major Steps in Analysis

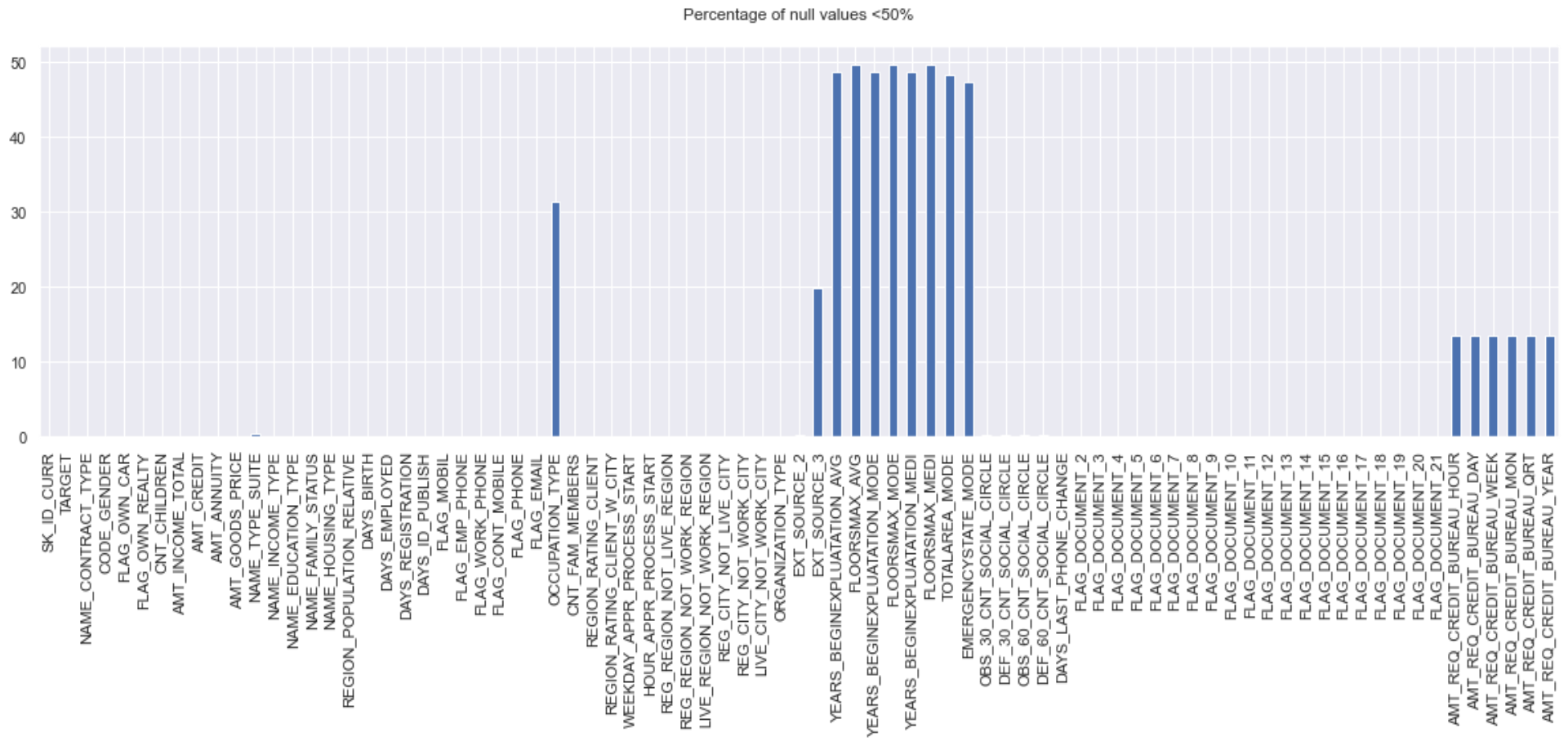
- ★ Data Sourcing
- ★ Data Understanding
- ★ Checking and Handling Missing values in the data
- ★ Handling Data Errors
- ★ Outlier Identification and Analysis
- ★ Univariate Analysis
- ★ Bivariate and Multivariate Analysis
- ★ Finding Top 10 Correlated Features those Support Target Column.

The above steps are done on both the Datasets and the results are shown.

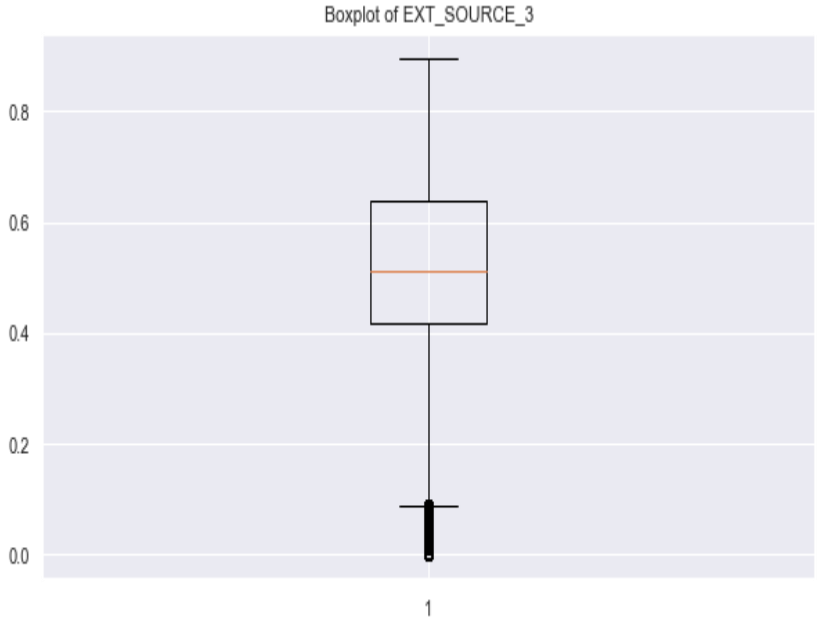
# Results

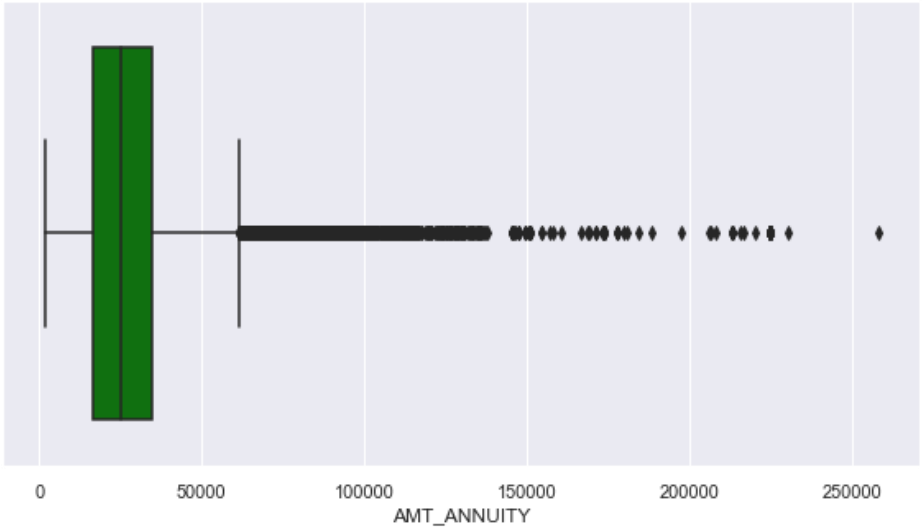
## application_data.csv Dataset

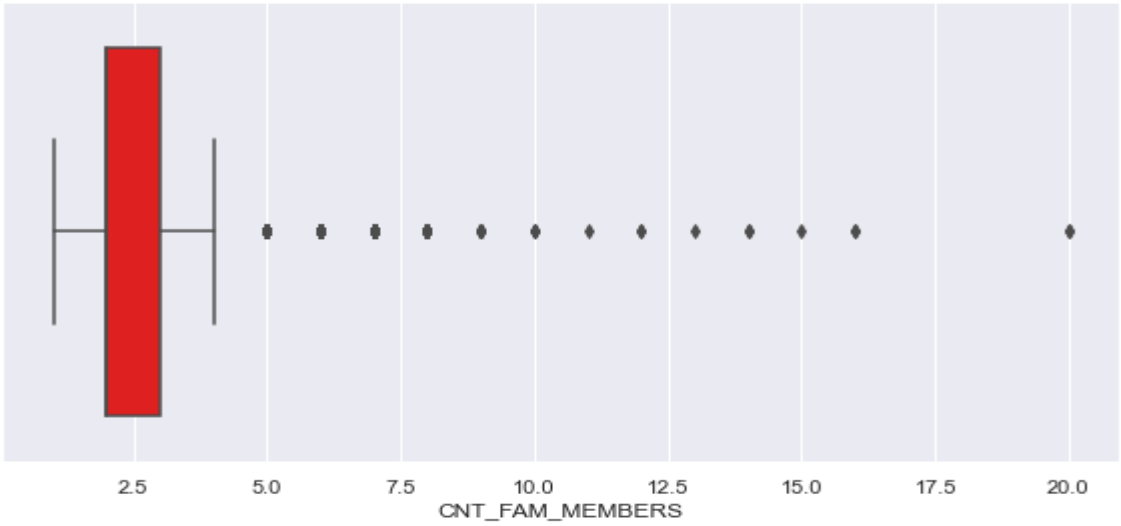
# Handling Missing Data

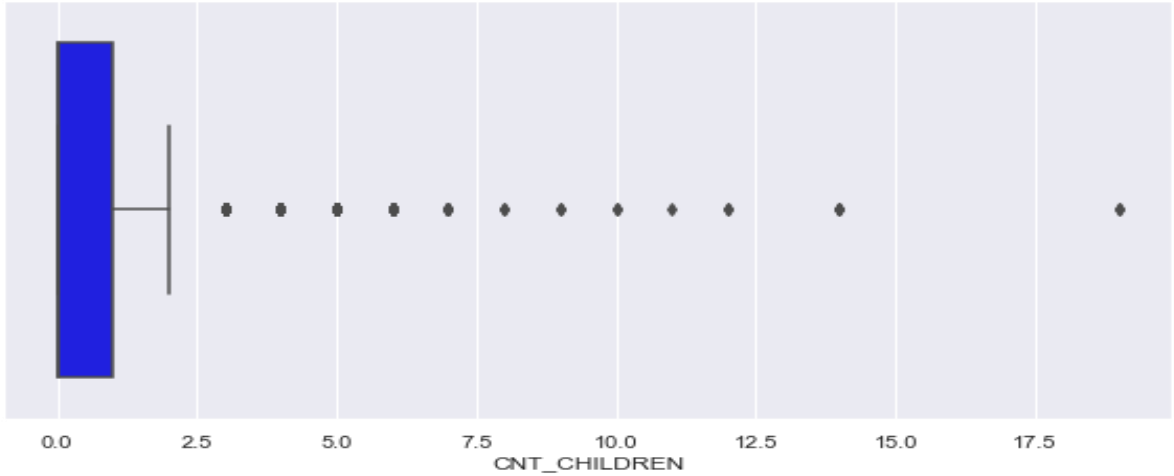


# Outlier Analysis

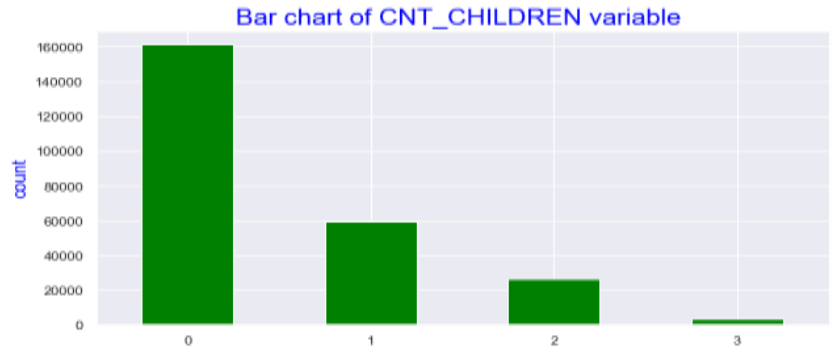


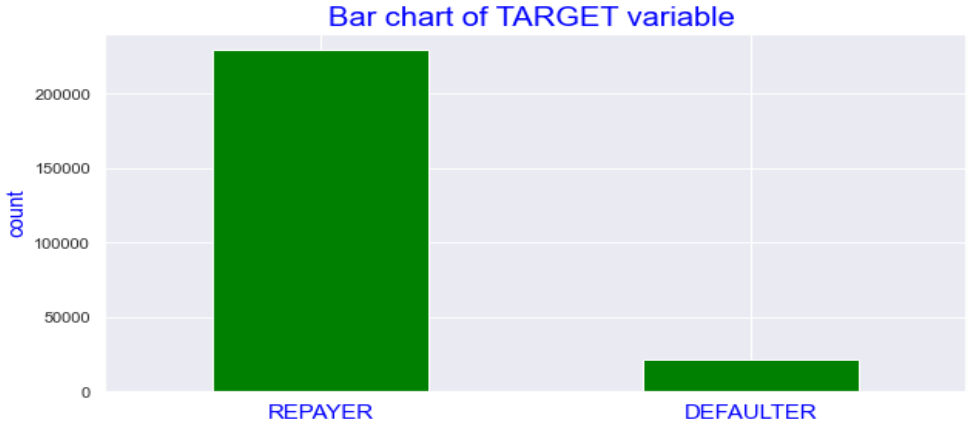



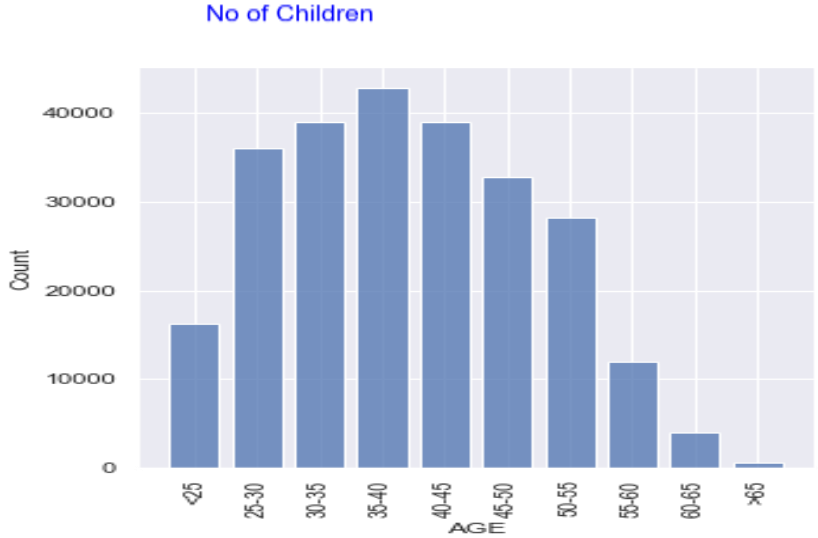

2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
CNT_FAM_MEMBERS

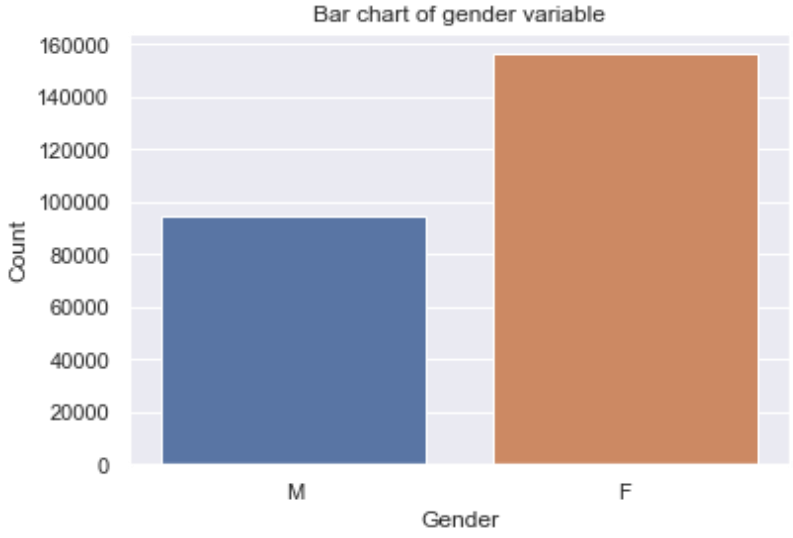

0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
CNT_CHILDREN

# Univariate Analysis

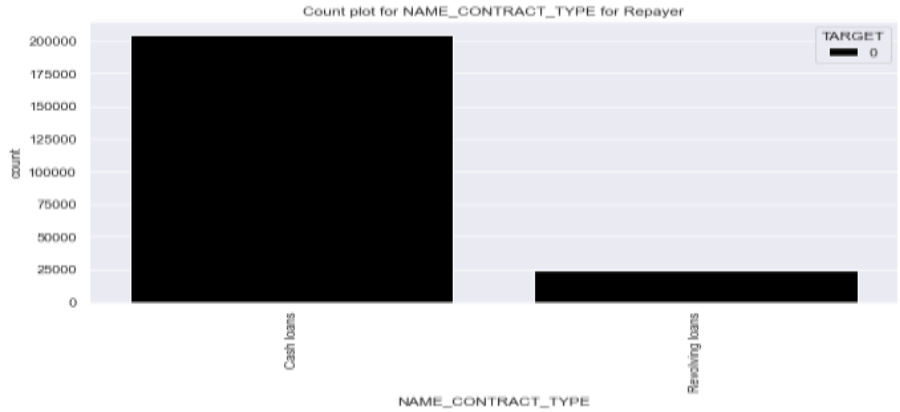


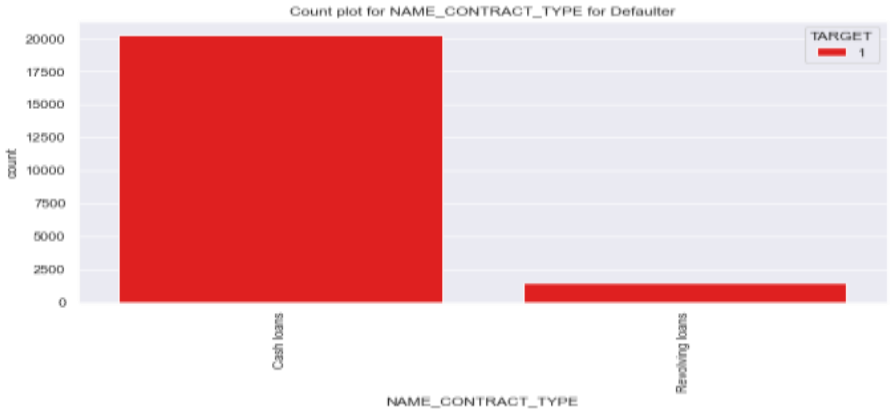



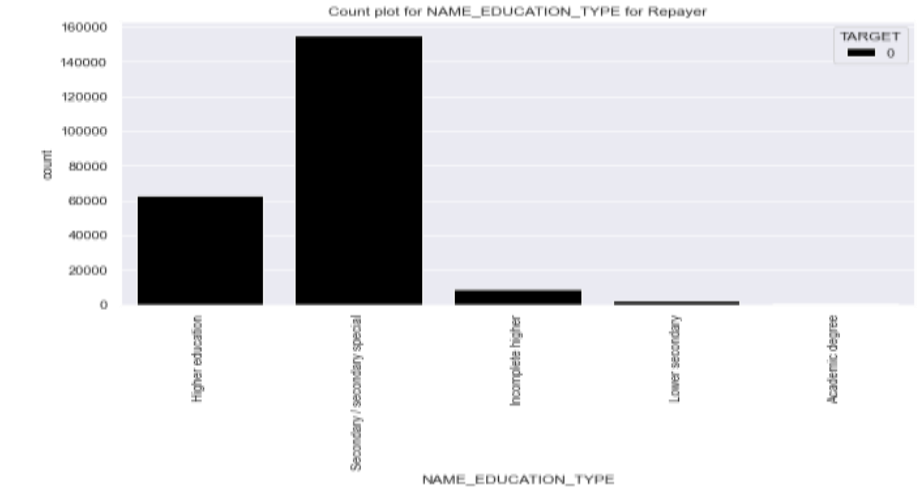



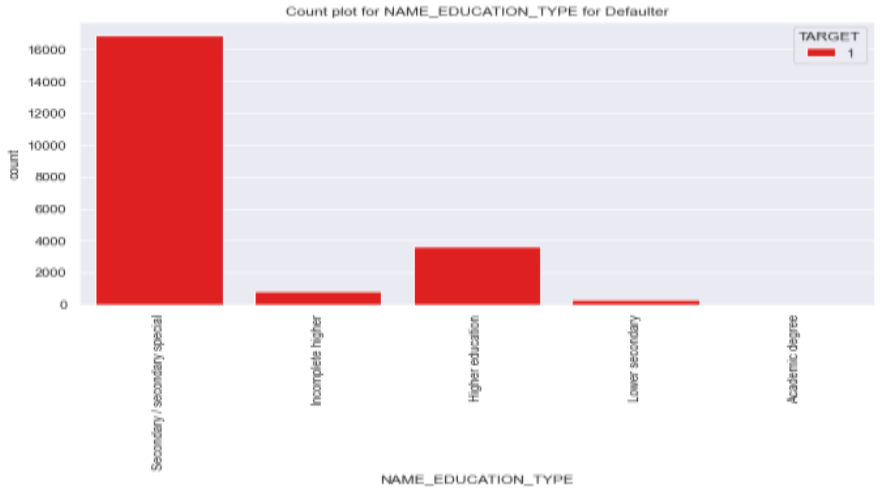



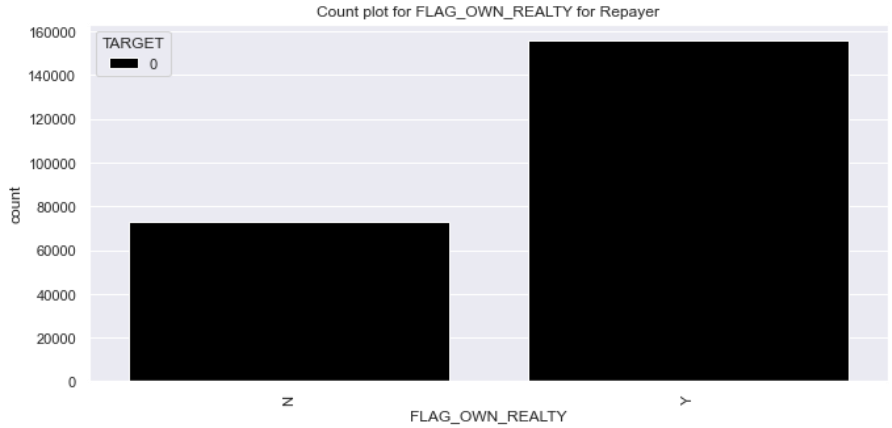
# Bivariate Analysis

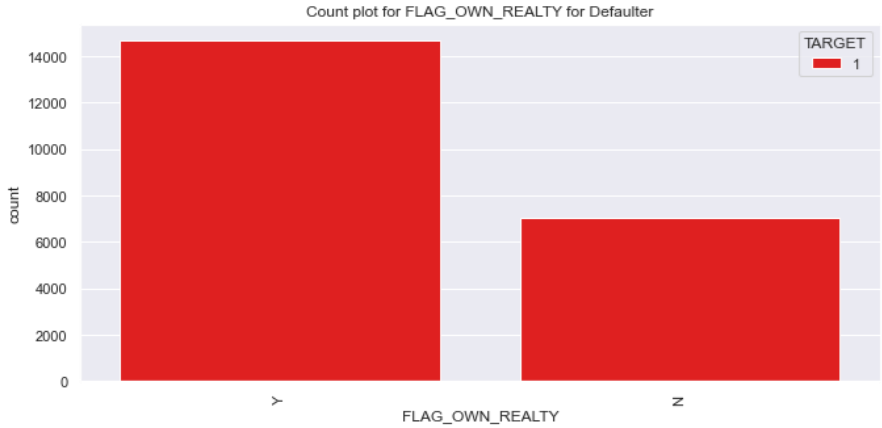


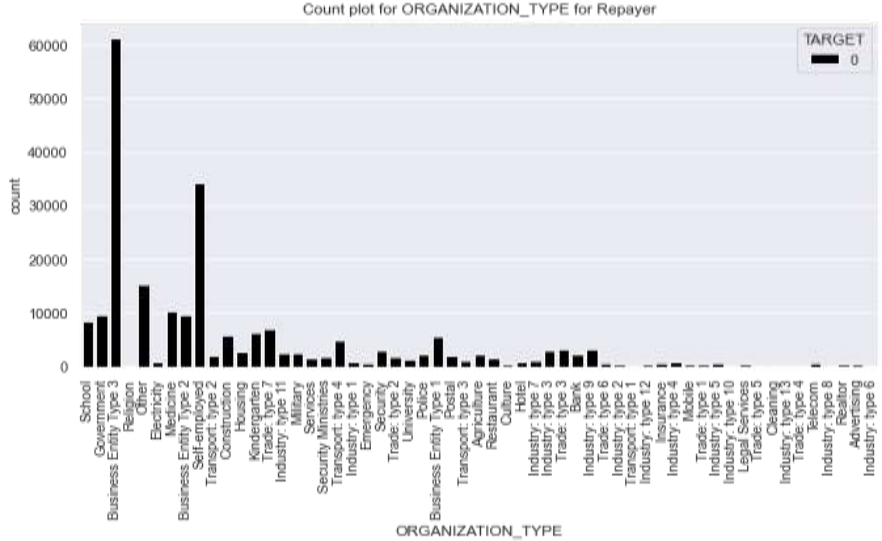



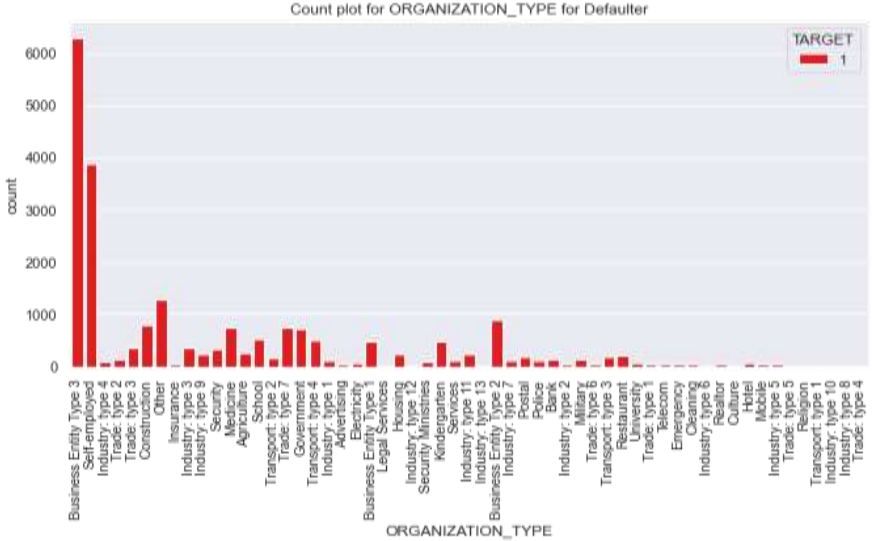




Count plot for FLAG_OWN_REALTY for Repayer
TARGET
0
count
FLAG_OWN_REALTY
N
Y


Count plot for FLAG_OWN_REALTY for Defaulter
TARGET
1
count
FLAG_OWN_REALTY
Y
N


Count plot for ORGANIZATION_TYPE for Repayer
TARGET
0
count
ORGANIZATION_TYPE

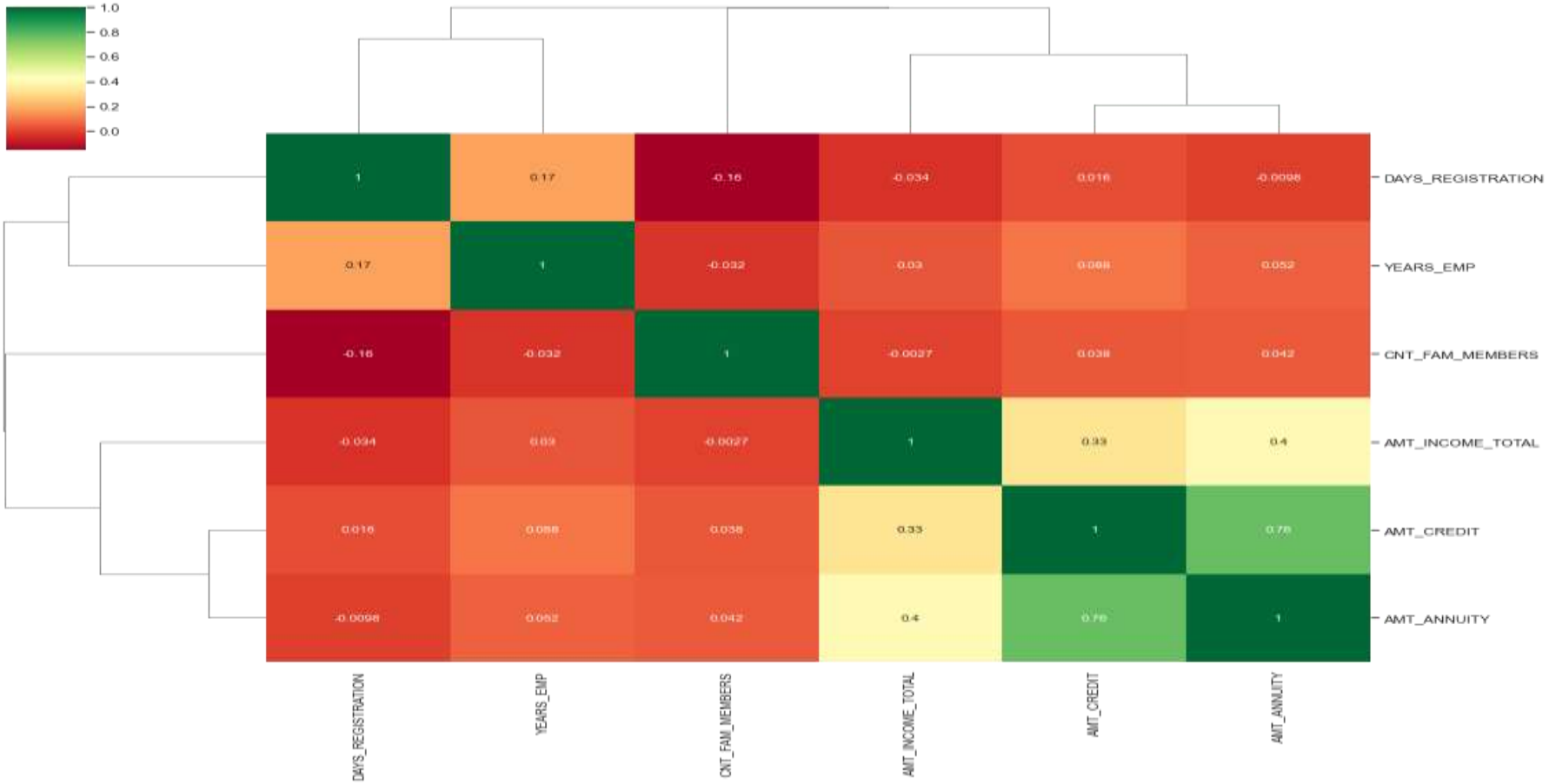

Count plot for ORGANIZATION_TYPE for Defaulter
TARGET
1
count
ORGANIZATION_TYPE

# Multivariate Analysis

# Top 10 Correlated Features for Repayers

| | variable_1 | variable_2 | correlation_coeff | abs_correlation_coeff |
|---|---|---|---|---|
| **13** | AMT_ANNUITY | AMT_CREDIT | 0.761985 | 0.761985 |
| **12** | AMT_ANNUITY | AMT_INCOME_TOTAL | 0.399959 | 0.399959 |
| **6** | AMT_CREDIT | AMT_INCOME_TOTAL | 0.325003 | 0.325003 |
| **34** | YEARS_EMP | DAYS_REGISTRATION | 0.172128 | 0.172128 |
| **27** | DAYS_REGISTRATION | CNT_FAM_MEMBERS | -0.156058 | 0.156058 |
| **31** | YEARS_EMP | AMT_CREDIT | 0.087949 | 0.087949 |
| **32** | YEARS_EMP | AMT_ANNUITY | 0.052049 | 0.052049 |
| **20** | CNT_FAM_MEMBERS | AMT_ANNUITY | 0.042150 | 0.042150 |
| **19** | CNT_FAM_MEMBERS | AMT_CREDIT | 0.037947 | 0.037947 |
| **24** | DAYS_REGISTRATION | AMT_INCOME_TOTAL | -0.034258 | 0.034258 |

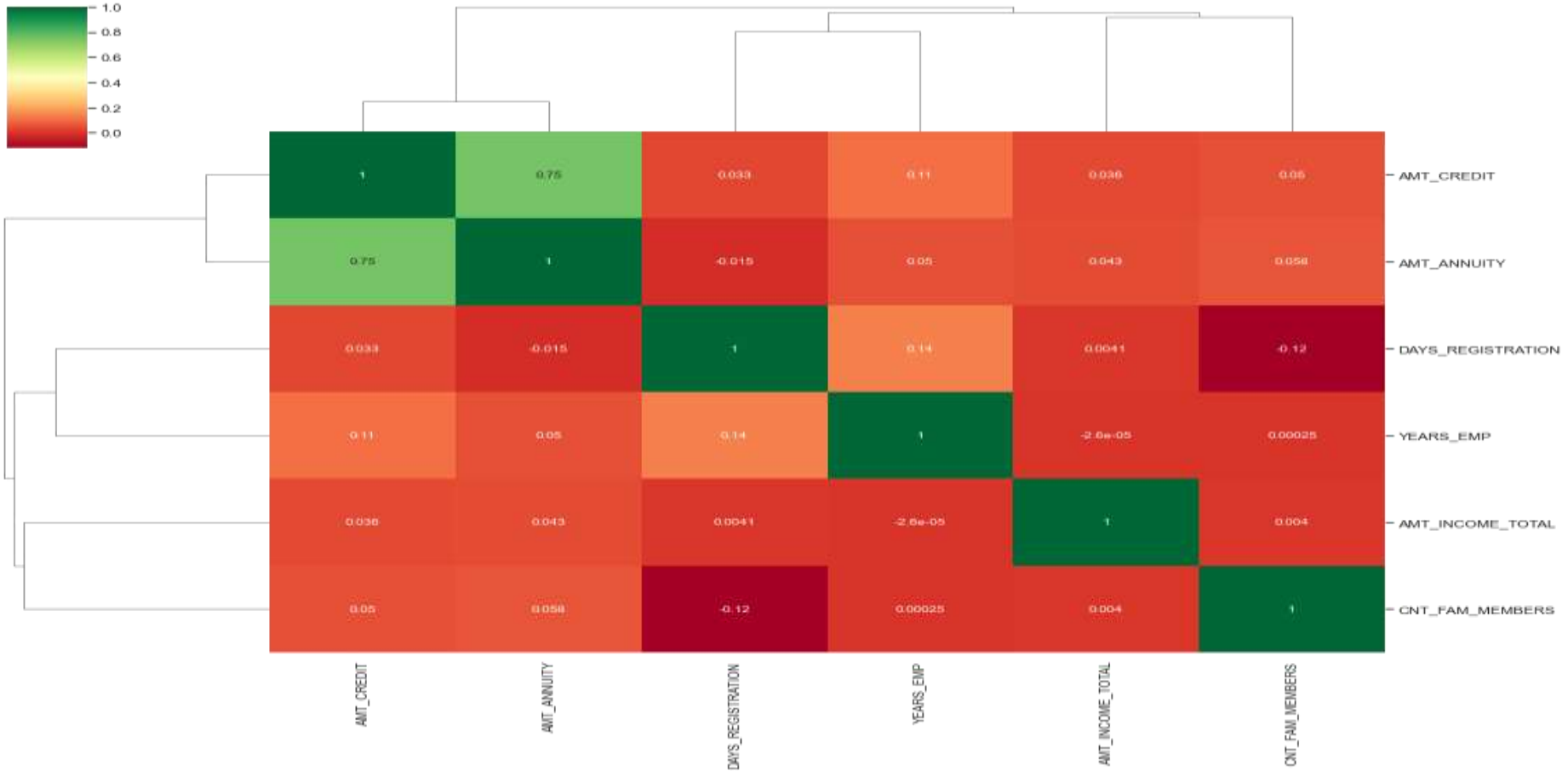
# Cluster Map for Repayer



# Top 10 Correlated Features for Defaulters

| | variable_1 | variable_2 | correlation_coeff | abs_correlation_coeff |
|---|---|---|---|---|
| **13** | AMT_ANNUITY | AMT_CREDIT | 0.748585 | 0.748585 |
| **34** | YEARS_EMP | DAYS_REGISTRATION | 0.138431 | 0.138431 |
| **27** | DAYS_REGISTRATION | CNT_FAM_MEMBERS | -0.122175 | 0.122175 |
| **31** | YEARS_EMP | AMT_CREDIT | 0.107943 | 0.107943 |
| **20** | CNT_FAM_MEMBERS | AMT_ANNUITY | 0.057776 | 0.057776 |
| **19** | CNT_FAM_MEMBERS | AMT_CREDIT | 0.050353 | 0.050353 |
| **32** | YEARS_EMP | AMT_ANNUITY | 0.049652 | 0.049652 |
| **12** | AMT_ANNUITY | AMT_INCOME_TOTAL | 0.043134 | 0.043134 |
| **6** | AMT_CREDIT | AMT_INCOME_TOTAL | 0.036289 | 0.036289 |
| **25** | DAYS_REGISTRATION | AMT_CREDIT | 0.032948 | 0.032948 |

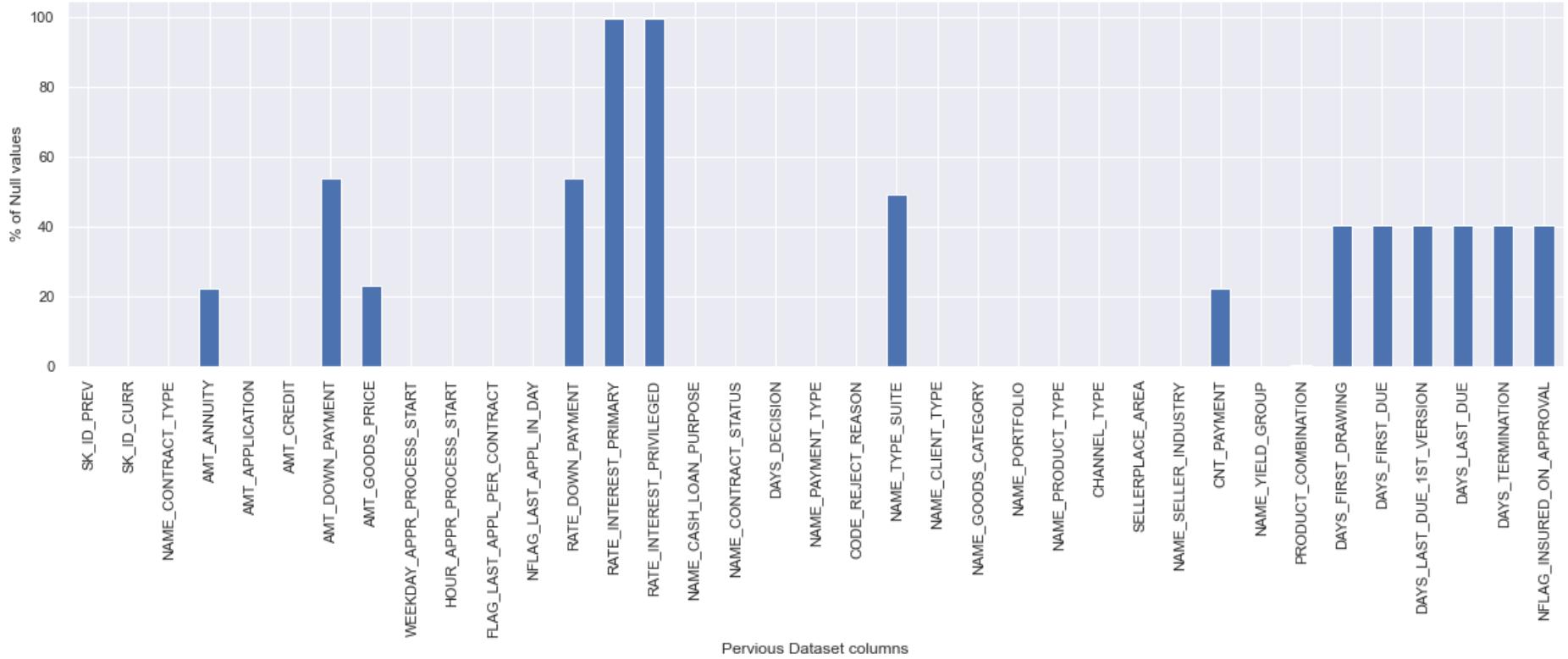
# Cluster Map for Defaulter



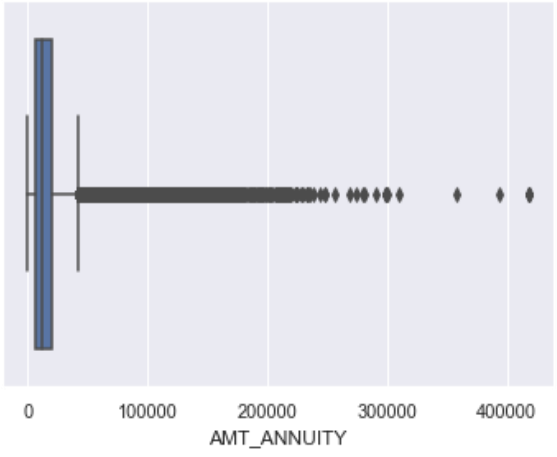
# Results

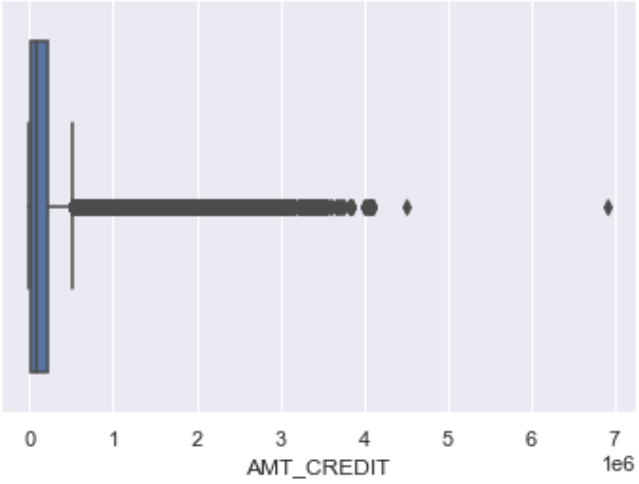
## previous_application.csv Dataset

# Handling Missing Data

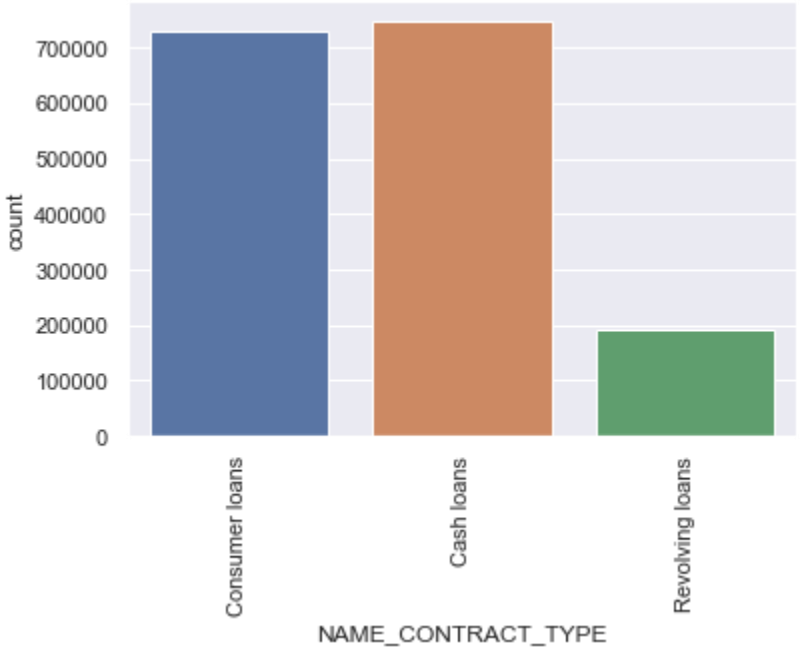


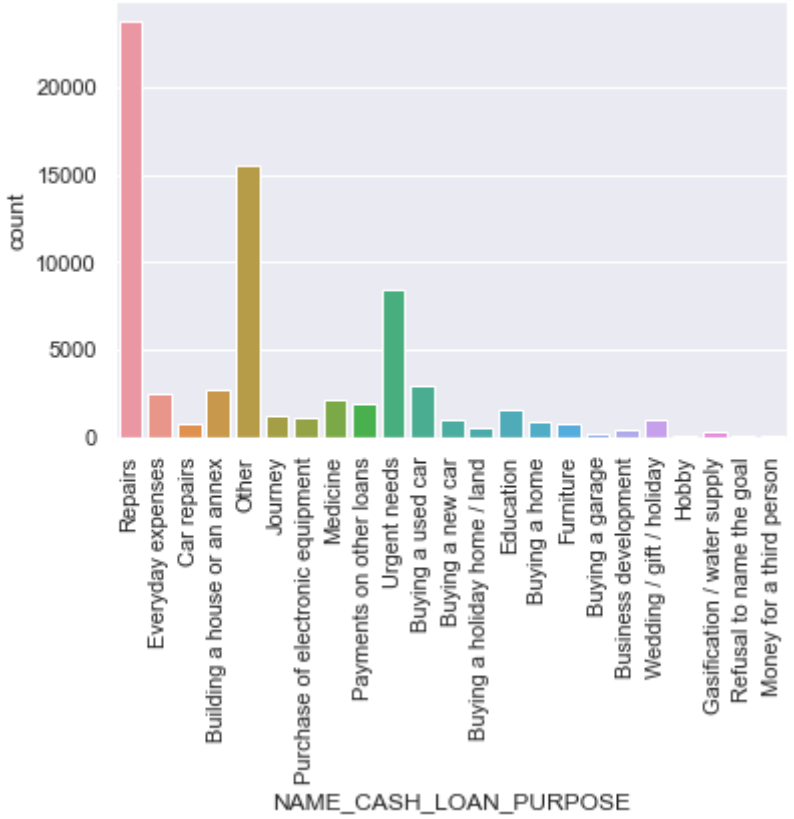
# Outlier Analysis



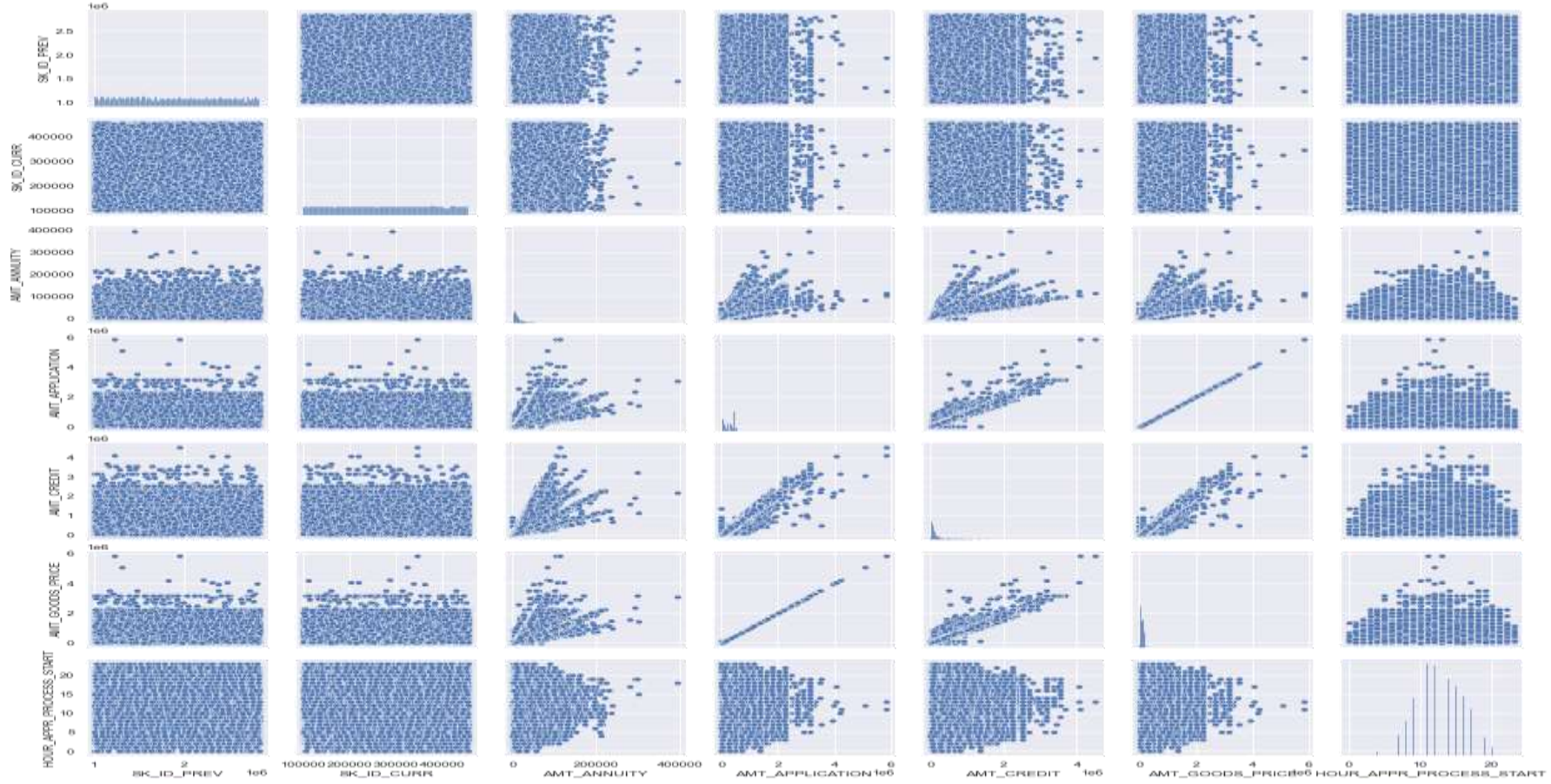




# Univariate Analysis



# Bivariate Analysis




SK_ID_PREV
SK_ID_CURR
AMT_ANNUITY
AMT_APPLICATION
AMT_CREDIT
AMT_GOODS_PRICE
HOUR_APPR_PROCESS_START
1 -0.0021 0.011 0.012 0.012 0.011 0.0046
-0.0021 1 -0.0015 -0.00088 -0.00079 -0.00079 0.0036
0.011 -0.0015 1 0.9 0.9 0.9 0.04
0.012 -0.00088 0.9 1 1 1 0.044
0.012 -0.00079 0.9 1 1 0.99 0.044
0.011 -0.00079 0.9 1 0.99 1 0.042
0.0046 0.0036 0.04 0.044 0.044 0.042 1
1.0
0.8
0.6
0.4
0.2
0.0

# Conclusion

- **The people belongs to Maternity leave (>98%) and Unemployed(>55%) categories are the one who were tend to be defaulters and it is wise to check once again if such people approach bank to lend loans.**
- **Students are the one who are very less tend to default.**
- **The clients who are Refused to name the goal were the majority people who were being defaulters.**
- **The people who lend a loan to Buying a garage were very less to be in defaulters.**

- **The SCOFR reason category is the most CODE_REJECT_REASON which got highest percentage of Defaulters.**
- **The people who belongs to Transport:type 3 were the majority people in defaulters.**
- **The people who belongs to Industry:type 12 were the very less to be default.**
- **Low-skilled Laborers are the one who tend to being like defaulters.**
- **The people who belongs Accountants category are very fewer to be in defaulters.**
- **Highest percentage of Defaulters belongs to Insurance and Vehicles category.**


Thank You