CREDIT EDA

BANK_LOAN_DEFAULT_RISK_ANALYSIS

CASE STUDY

PROBLEM STATEMENT

- When the Bank receives a loan application, the Bank has to decide for the loan approval based on the applicant's profile.

Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

EDA ANALYSIS APPROACH

- UNDERSTANDING OF PROBLEM STATEMENT
- DATA UNDERSTANDING
- DATA SOURCING
- DATA IMPUTATION
- DATA CLEANING
- OUTLIER ANALYSIS
- DATA ANALYSIS

Understanding of Problem Statement

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.

- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

DATA SOURCING

- Importing Application Data
- Importing Previous Application Data
- Finding Shape, Data types, Info, describe of both the application and previous data’s.

DATA CLEANING

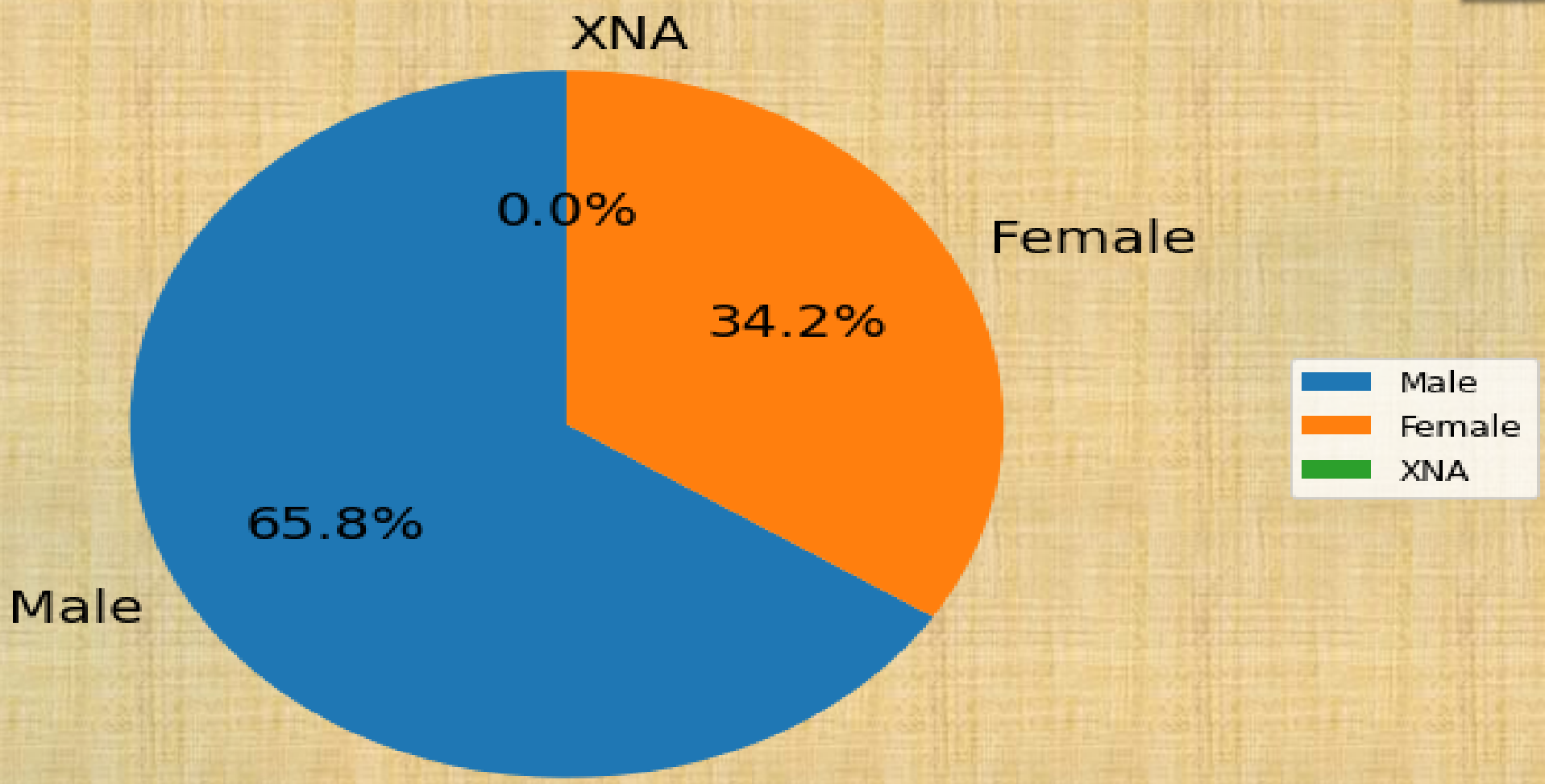
- Finding Null values on the Application Data.

***Report of Null values in Application Data ::: There are 49 coloumns in Application data are missing values. Very high missing values detected.¶

- Finding Null values on the Previous Application Data

- ***Report of Null values ::: There are 11 coloumns in Previous application data are missing values detected.
- **Dropping the missing values taking percentage of >=40%

Distribution of Gender



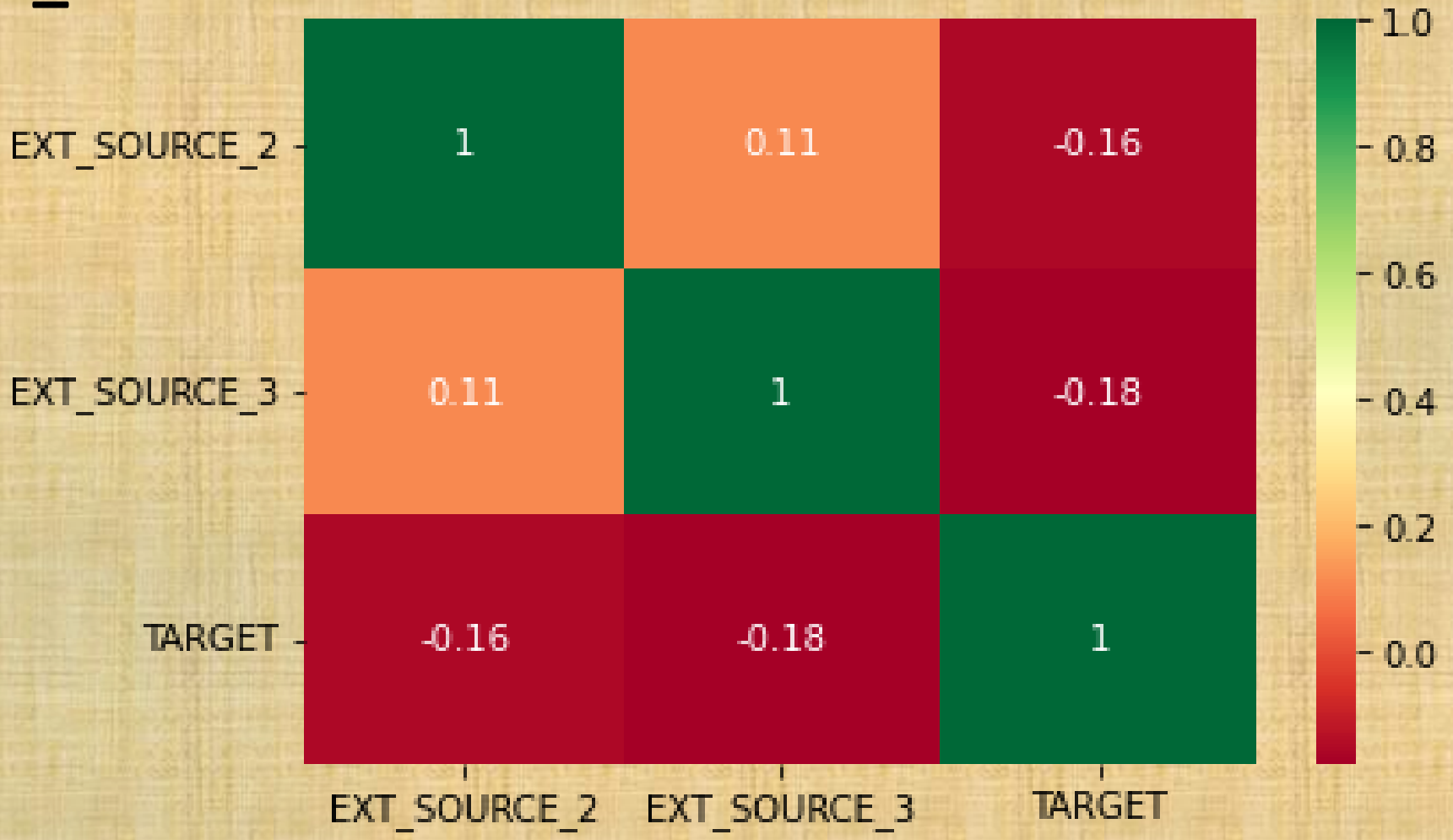
Observations and Suggestion from Distribution of Gender:::

** Based on the Gender Distribution Analysis there are 4'XNA' values in Code_Gender column, those can be imputed by female.

** Mode of the Gender code value is 65% for the toatal records adding this 4'XNA' records doesn't impact our overall anaylsis.

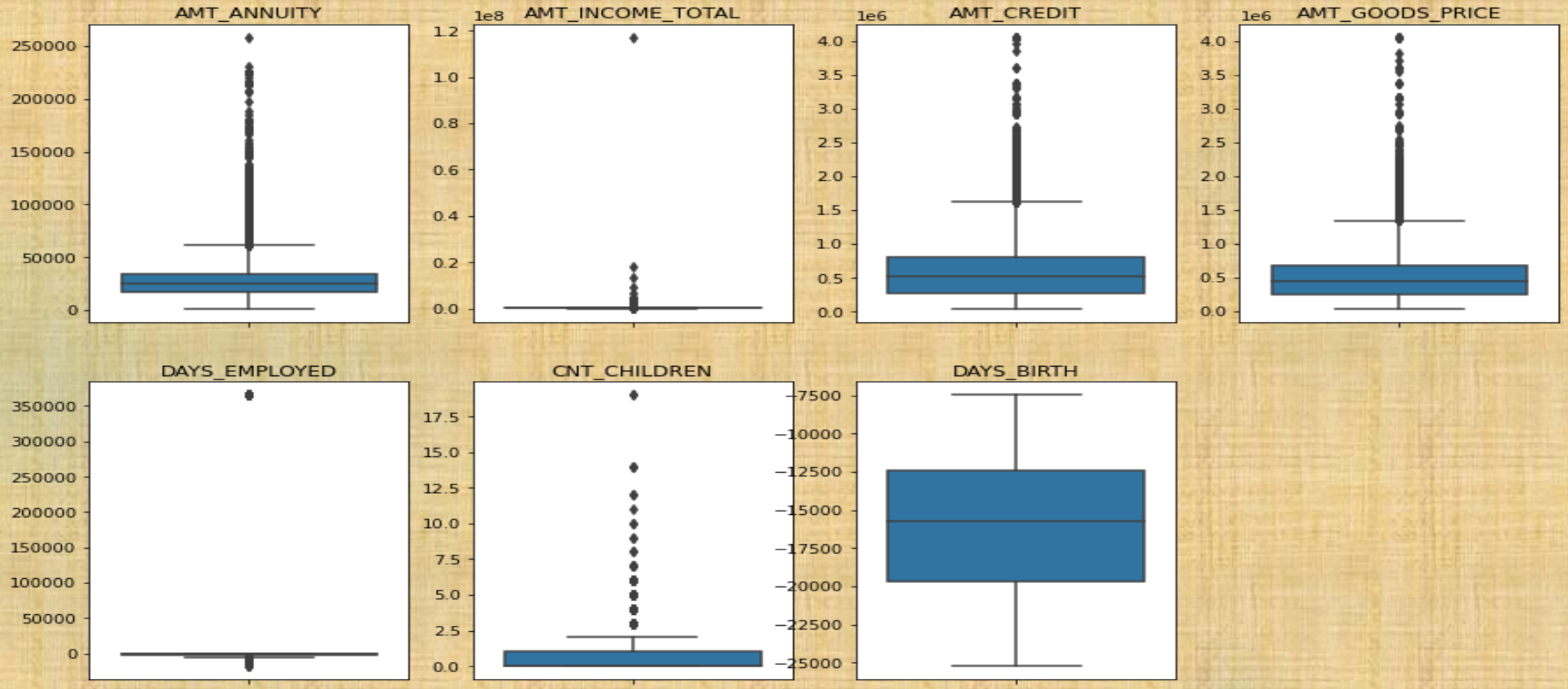
Suggestion: There is no need to impute null values beacause it doesn't affect our analysis.

HEATMAP for variables and columns
EXT_Source Vs. TARGET Varibles



Observations::: Based on the above observation HEATMAP there is almost no corelation between the EXT_SOURCE and TARGET coloumns..

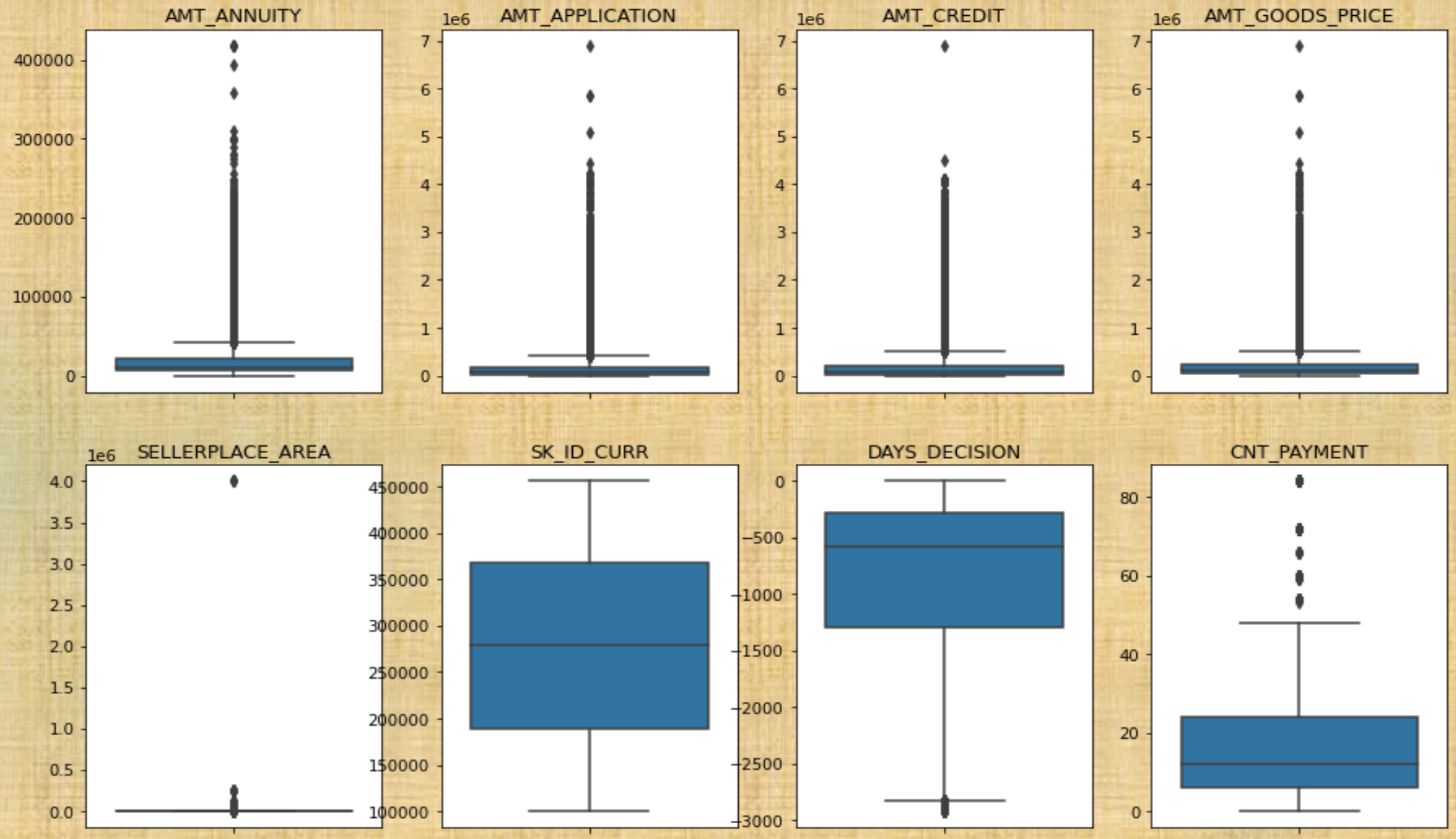
Outliers Analysis
From APPLICATION DATA



- Observations of Outlier Analysis::: From Application data

- 1. AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- 2. AMT_INCOME_TOTAL has huge number of outliers thet indicate that few of the loan applicants have high income when compared to the others.
- 3.DAYS_BIRTH has no outliers means the data availability is Reliable.
- 4. DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.¶

Outlier Analysis
From Previous Application Data

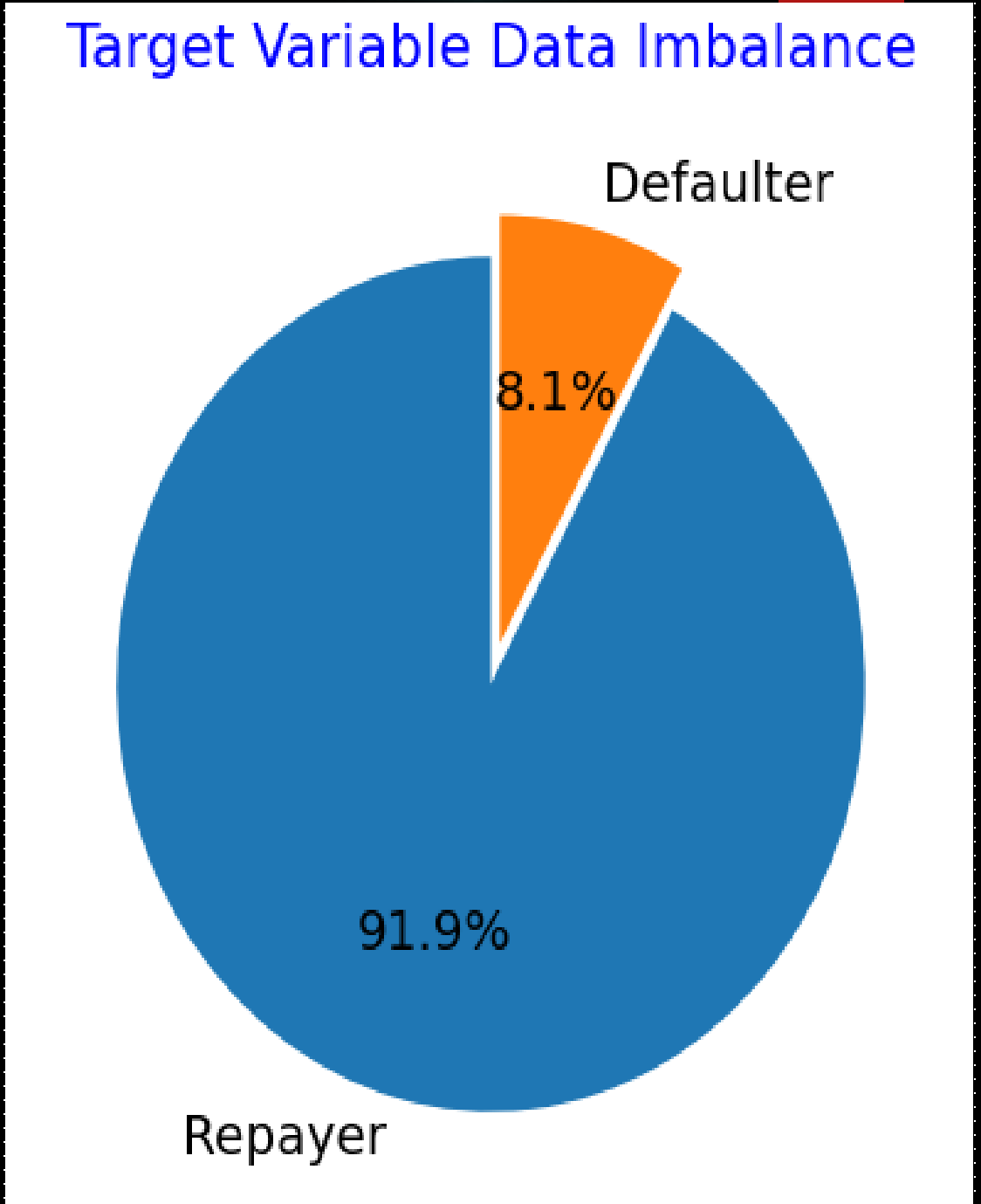


- Observations of Outliers::: From Previous application data::
- 1.AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
- 2.CNT_PAYMENT has few outlier values.
- 3.SK_ID_CURR is an ID column and hence no outliers.
- 4.DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.

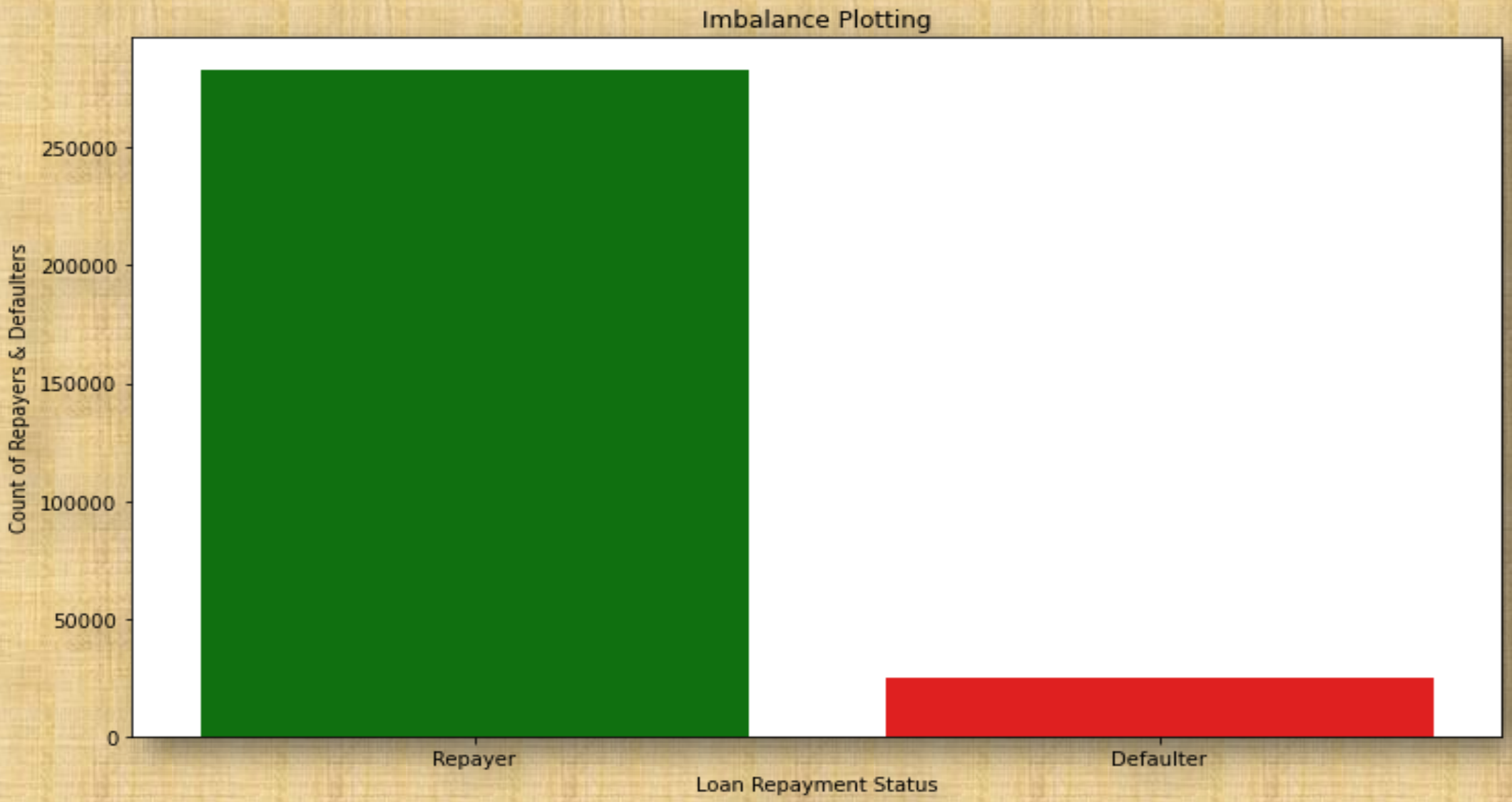
DATA ANALYSIS
IMBALANCE ANALYSIS

Observations:::

Repaying Loan the REPAYERS percentage is approx. 91.9% and who are not paying the loan DEFAULTERS is approx.8.1%

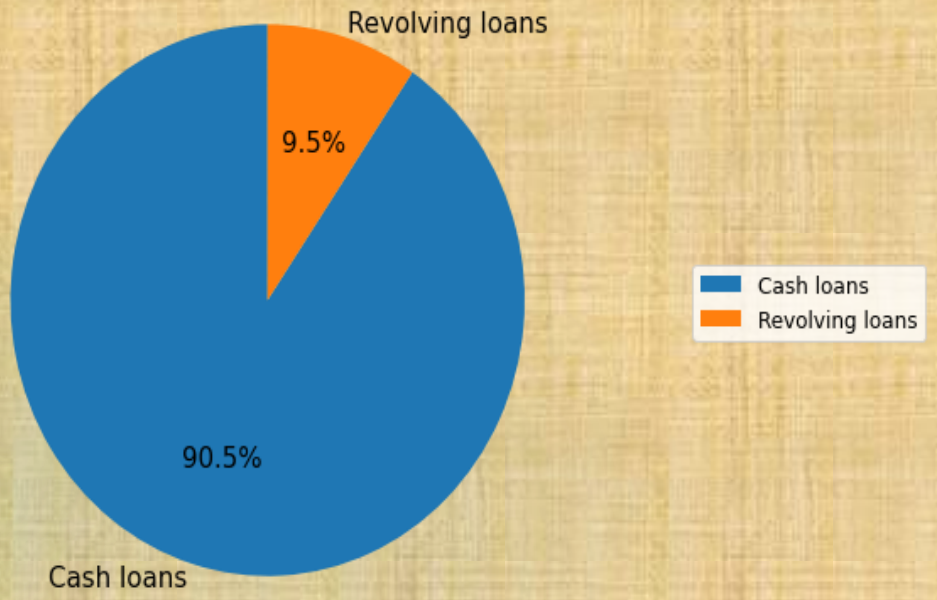


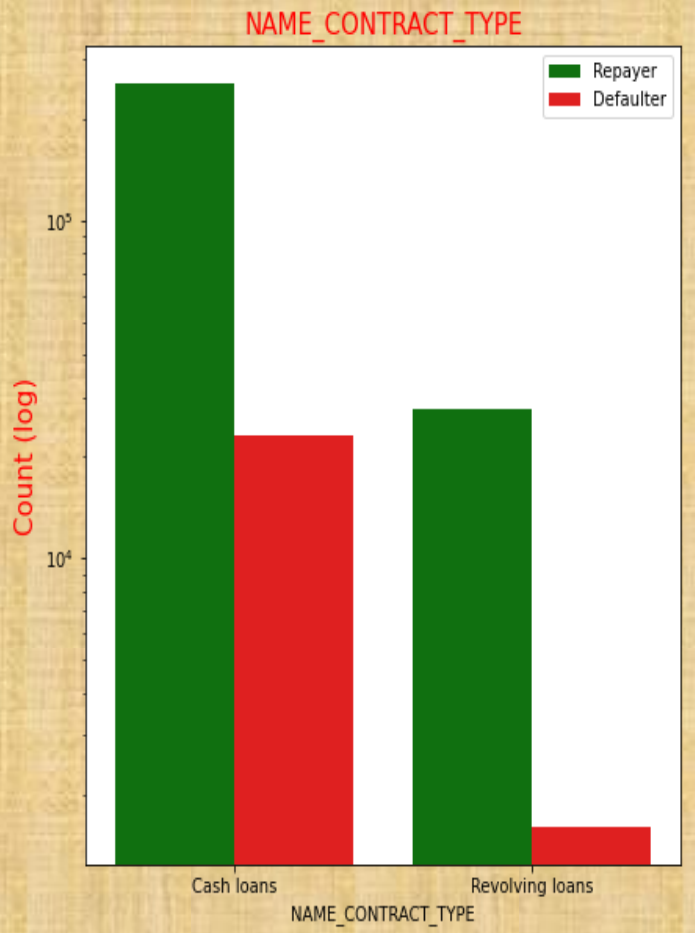
Showing Imbalance in Bar Plot

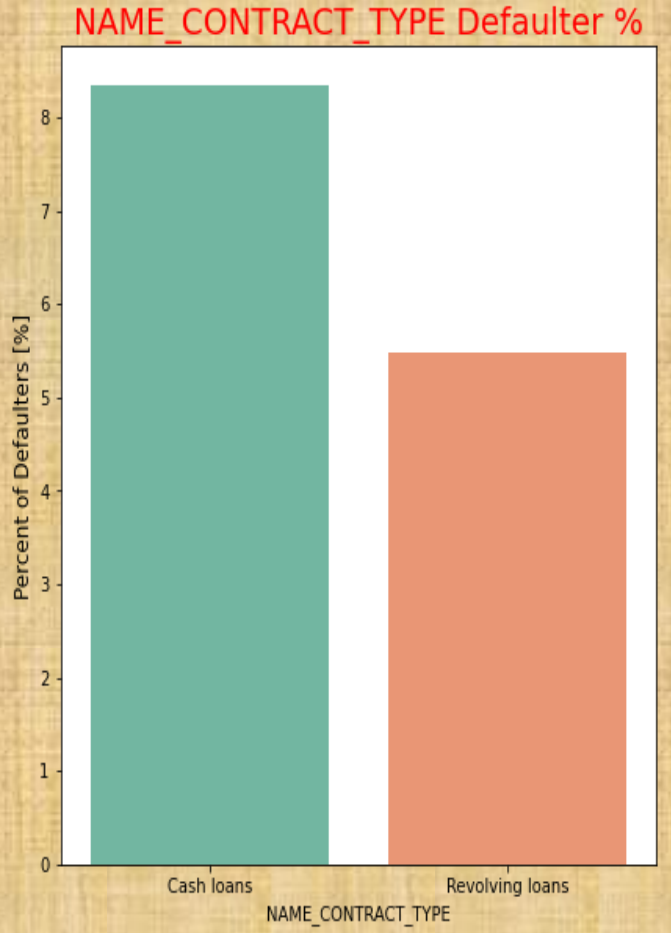


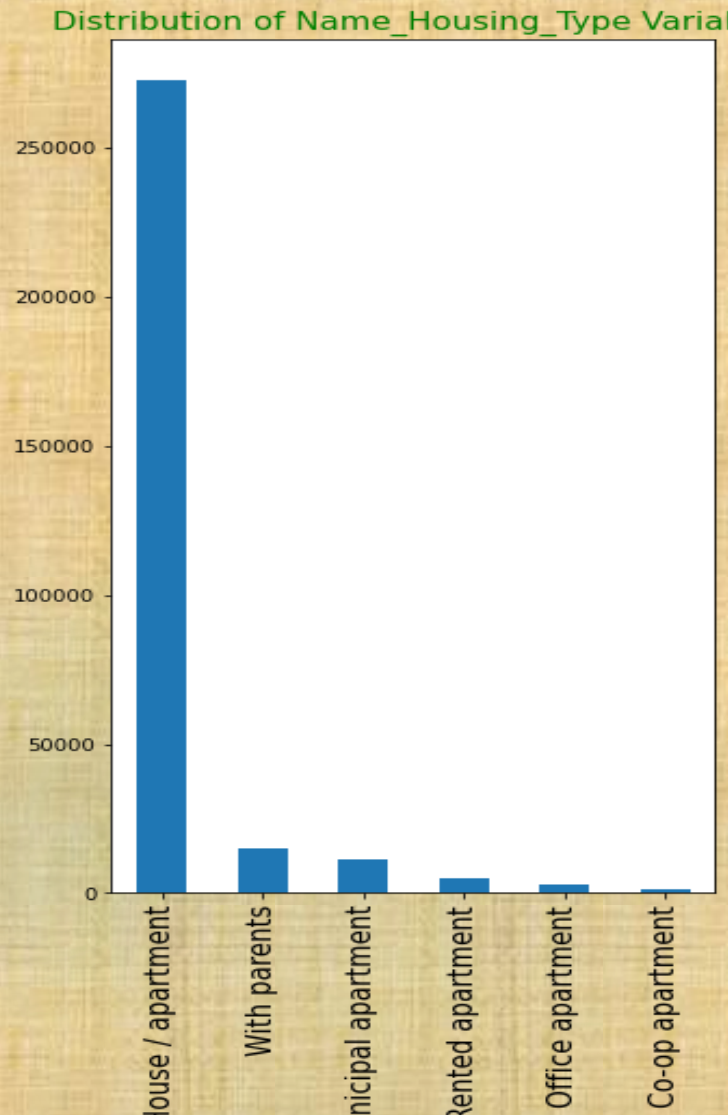
Categorical Variable Analysis
SEGMENTED Univariate Analysis

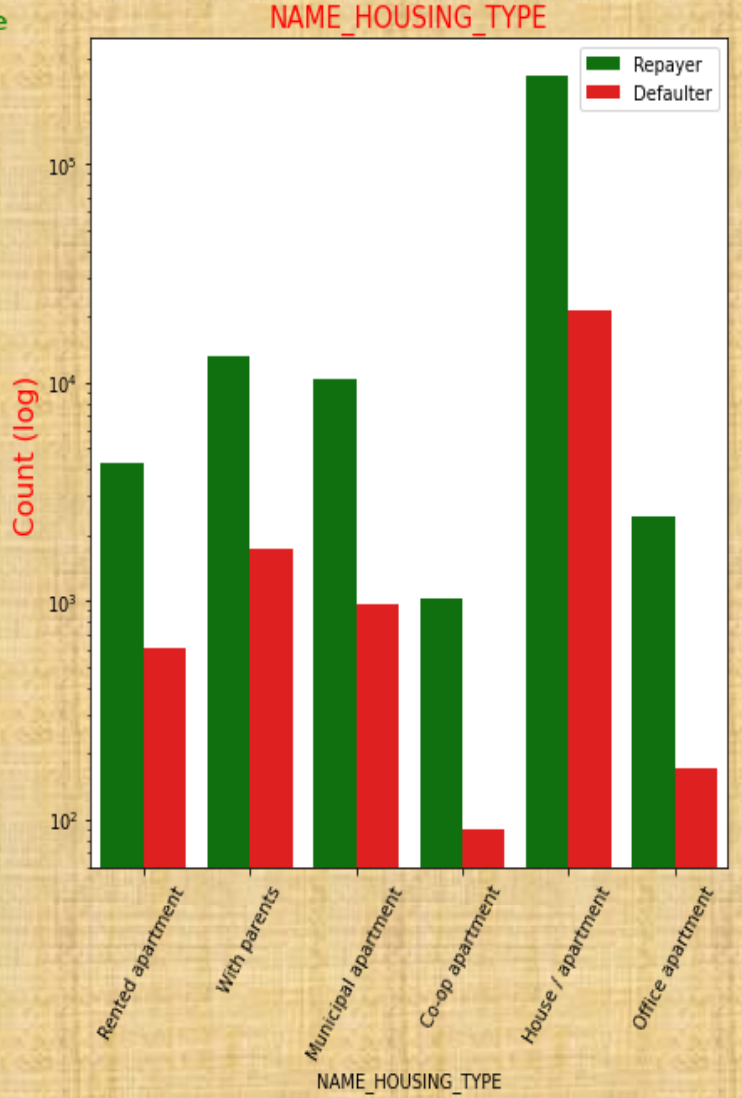
Distribution of Name_Contract_Type Variable

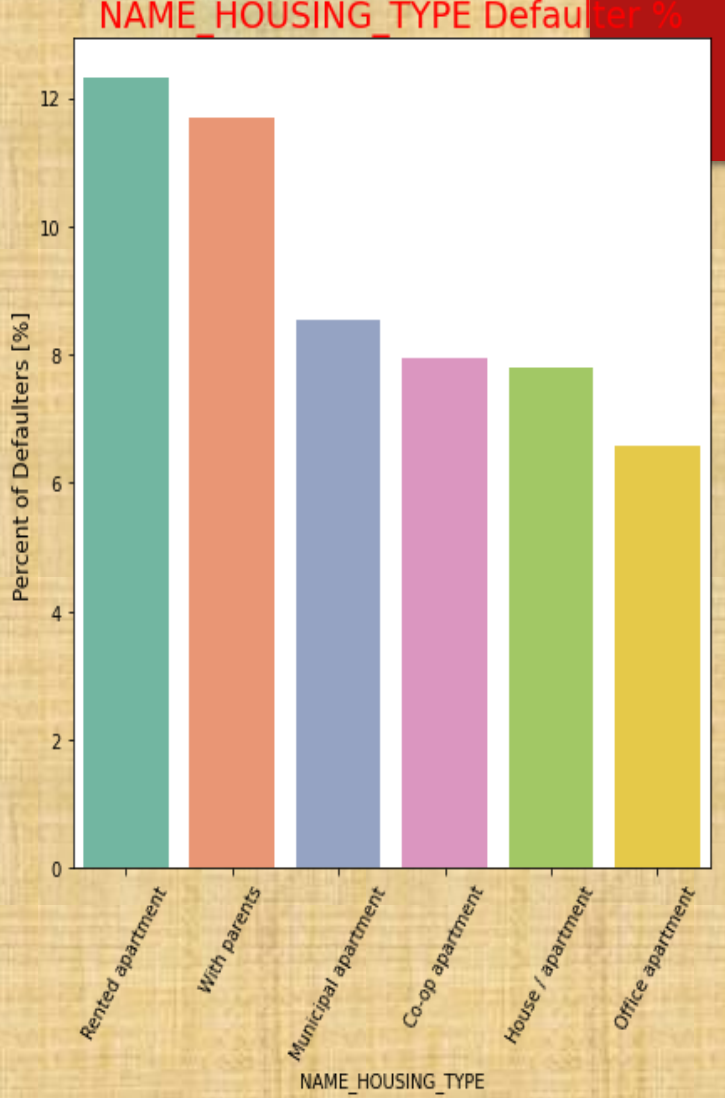




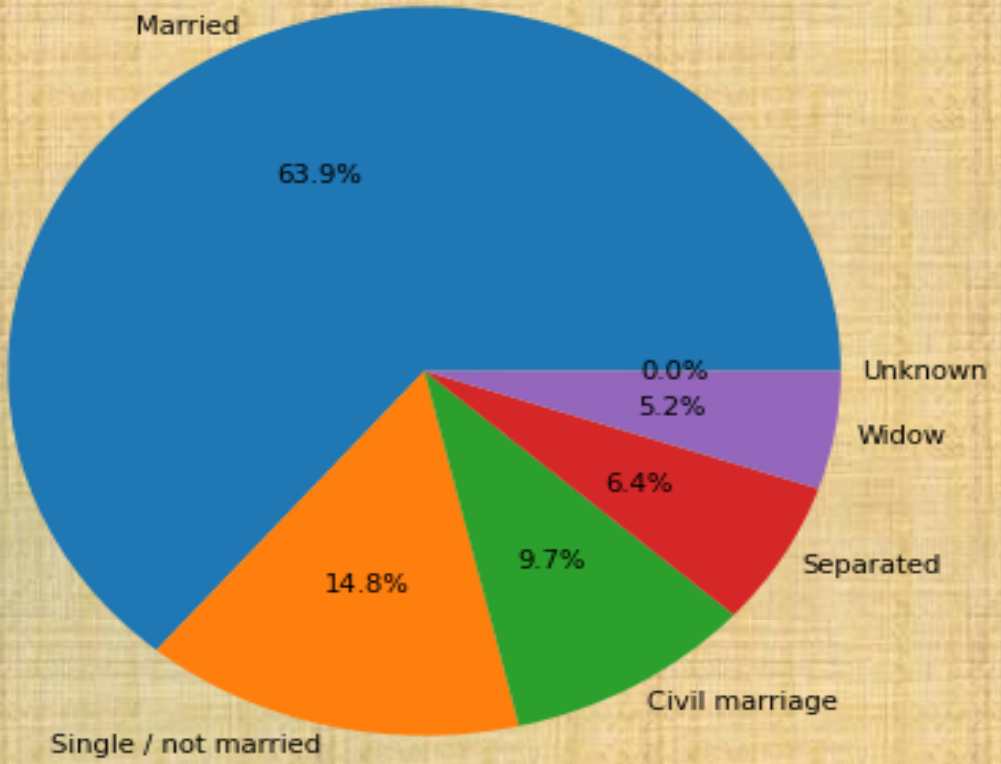



Distribution of Name_Housing_Type Variable
250000
200000
150000
100000
50000
0
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment

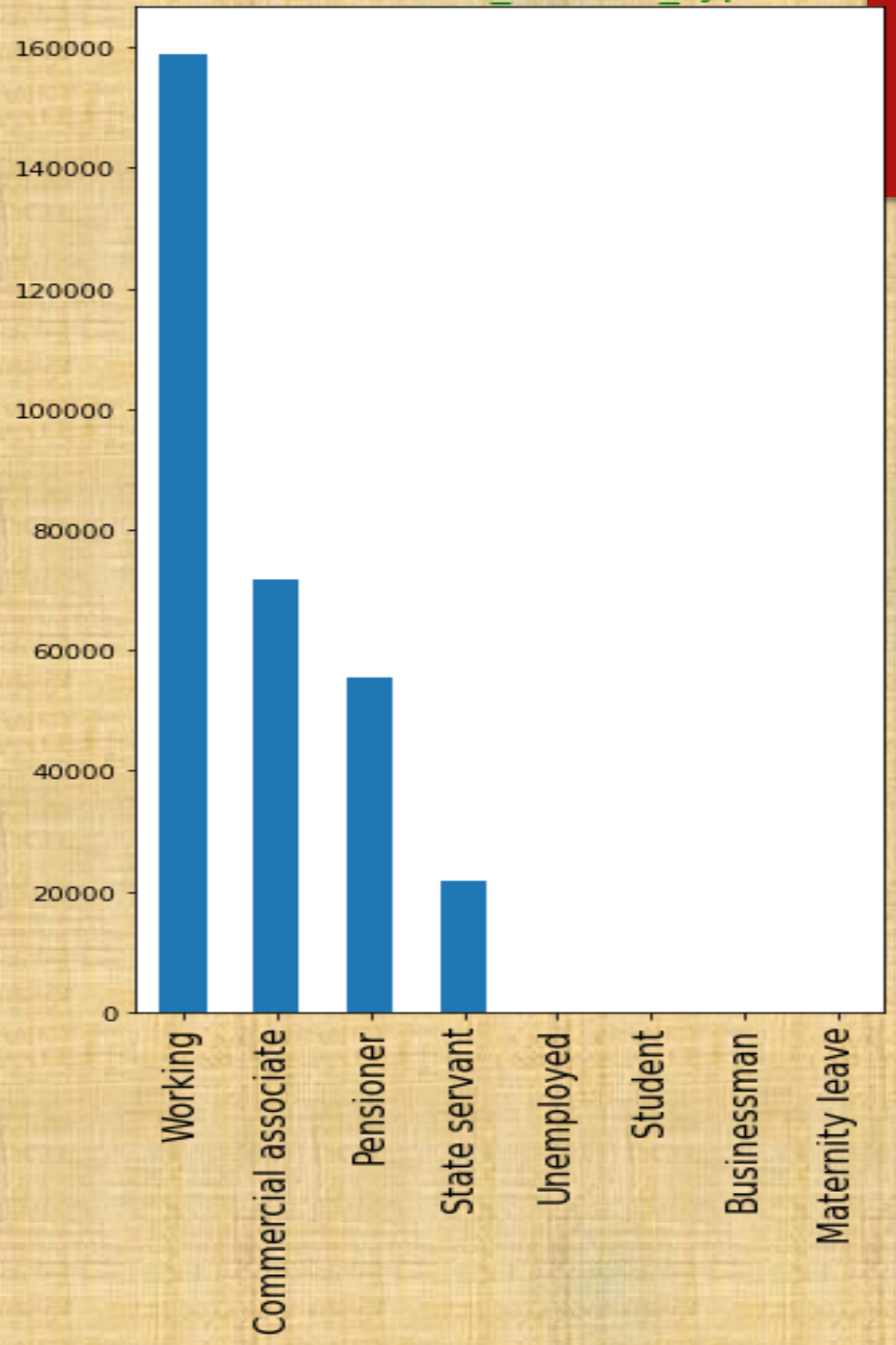

NAME_HOUSING_TYPE
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
Rented apartment
With parents
Municipal apartment
Co-op apartment
House / apartment
Office apartment
NAME_HOUSING_TYPE

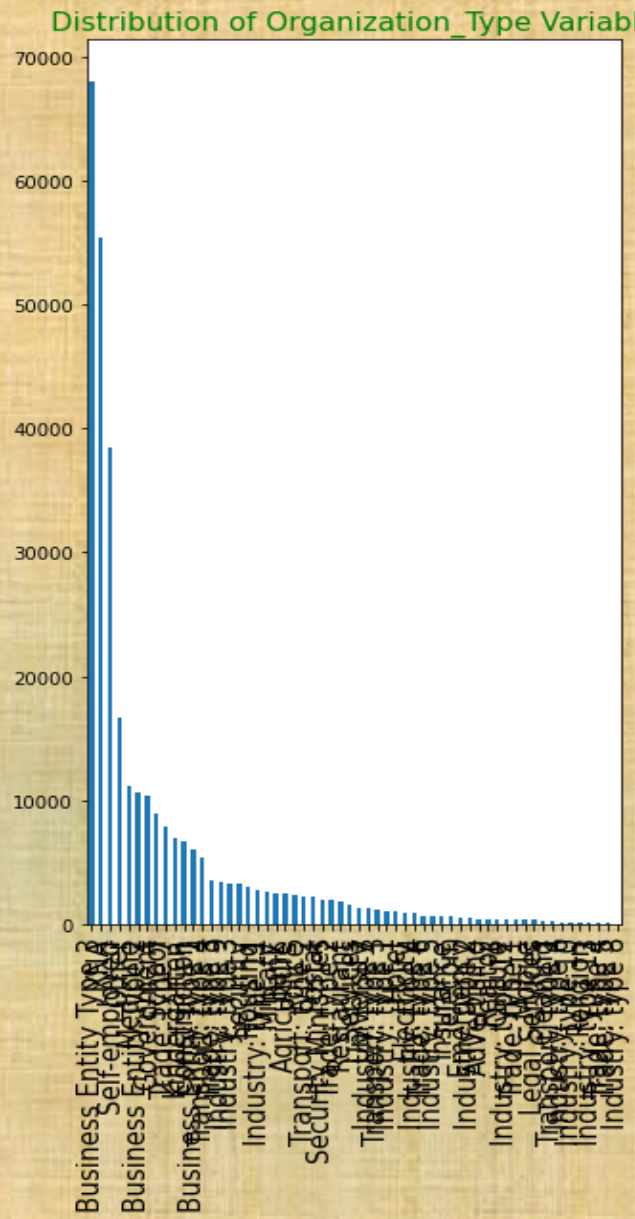

NAME_HOUSING_TYPE Defaulter %
Percent of Defaulters [%]
12
10
8
6
4
2
0
Rented apartment
With parents
Municipal apartment
Co-op apartment
House / apartment
Office apartment
NAME_HOUSING_TYPE

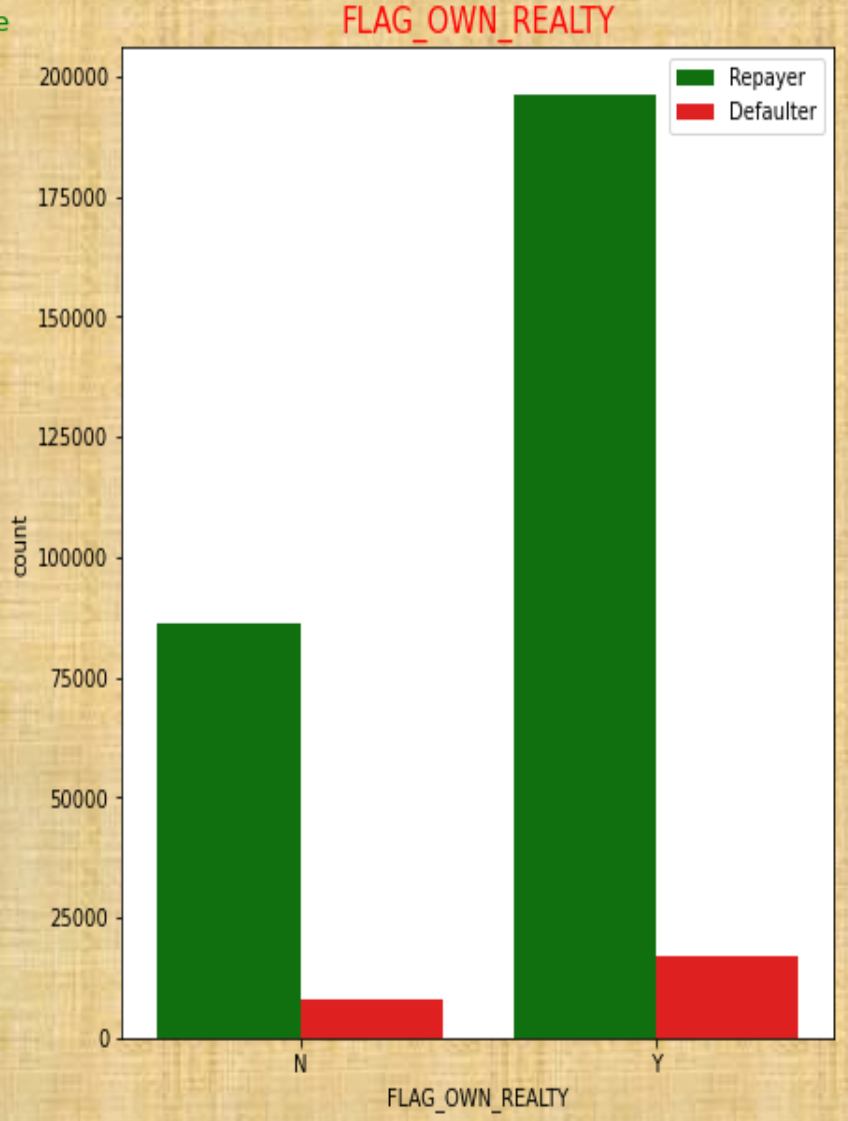
Distribution of Name_Family_Status Variable

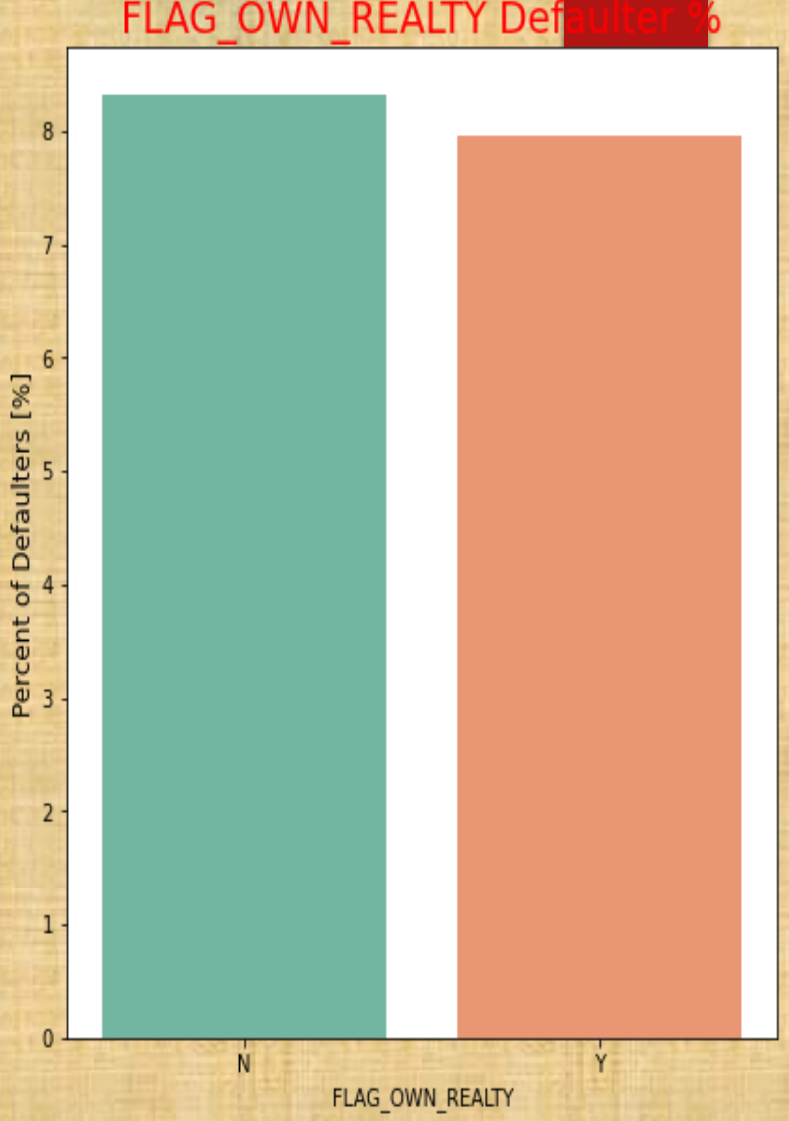


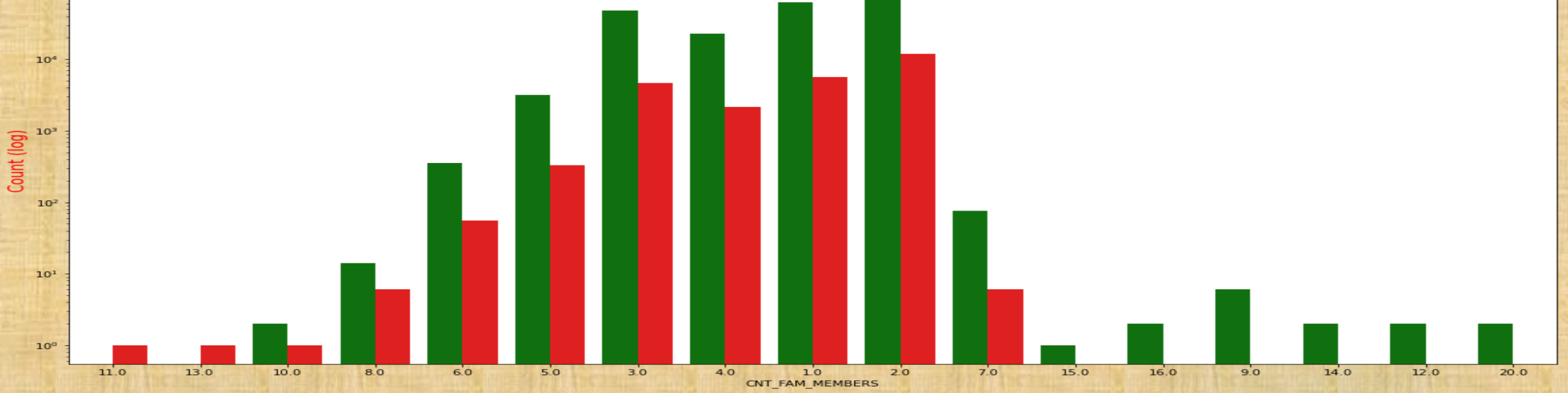
Distribution of Name_Income_Type Variable

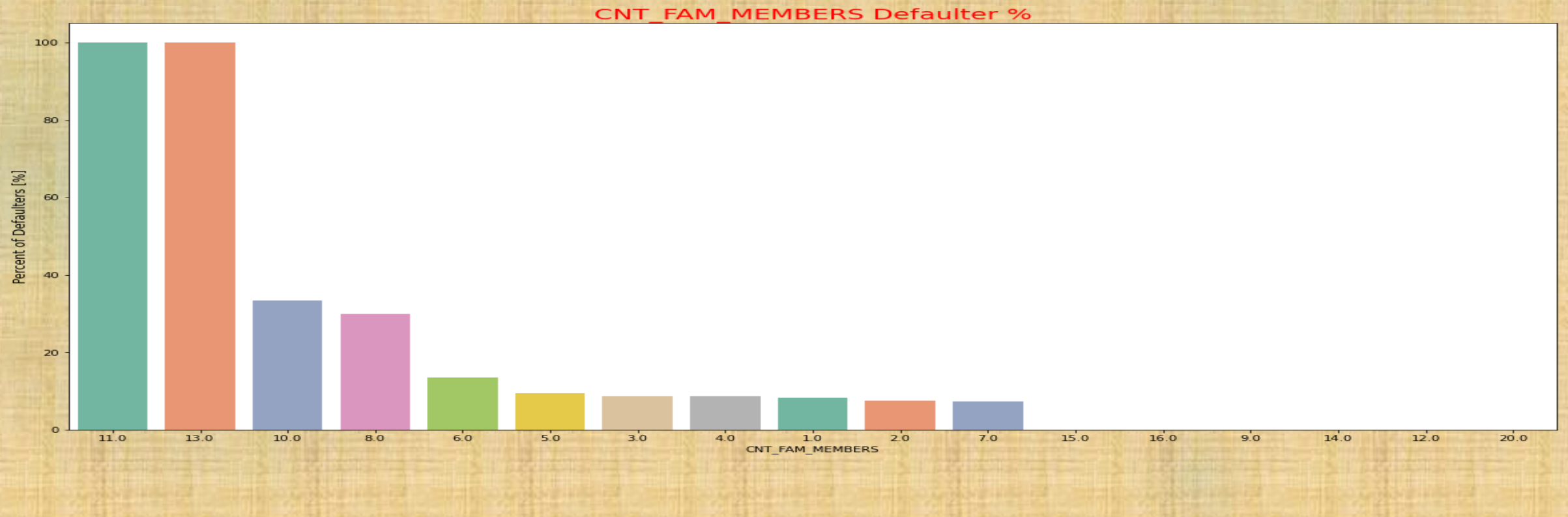



Distribution of Organization_Type Variable
70000
60000
50000
40000
30000
20000
10000
0
Business Entity Type 3
Self-employed
Business Entity Type 2
Business Entity Type 1
Industry: type 1
Industry: type 2
Agriculture
Transport: type 4
Security Ministries
Industry: type 3
Trade: type 7
Industry: type 11
Industry: type 4
Industry: type 5
Industry: type 9
Industry: type 12
Industry: type 6
Industry: type 10
Industry: type 13
Industry: type 8


FLAG_OWN_REALTY
Repayer
Defaulter
count
200000
175000
150000
125000
100000
75000
50000
25000
0
N
Y
FLAG_OWN_REALTY

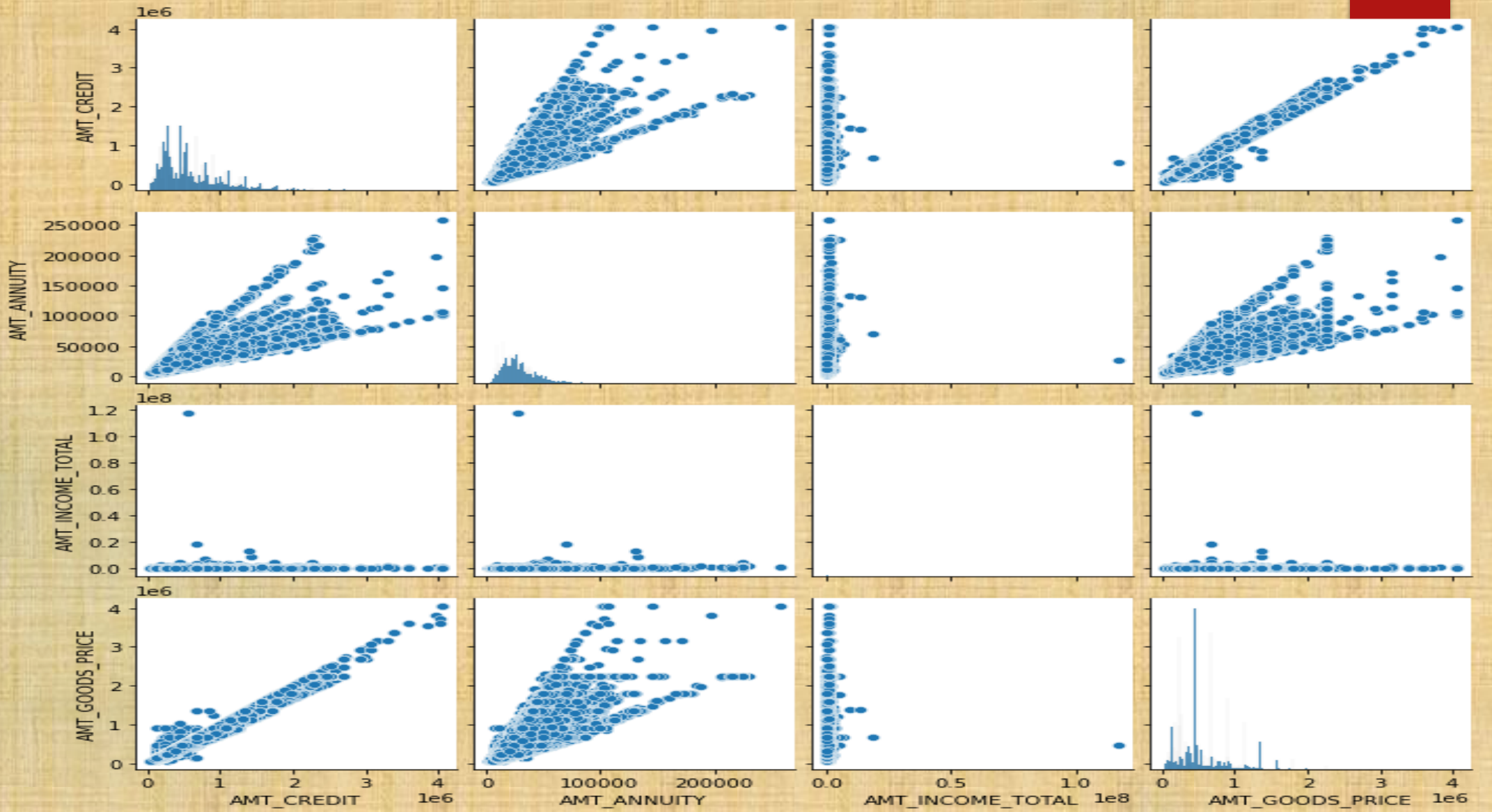

FLAG_OWN_REALTY Defaulter %
Percent of Defaulters [%]
8
7
6
5
4
3
2
1
0
N
Y
FLAG_OWN_REALTY


Count (log)
10^4
10^3
10^2
10^1
10^0
11.0
13.0
10.0
8.0
6.0
5.0
3.0
4.0
1.0
2.0
7.0
15.0
16.0
9.0
14.0
12.0
20.0
CNT_FAM_MEMBERS


CNT_FAM_MEMBERS Defaulter %
Percent of Defaulters [%]
100
80
60
40
20
0
11.0
13.0
10.0
8.0
6.0
5.0
3.0
4.0
1.0
2.0
7.0
15.0
16.0
9.0
14.0
12.0
20.0
CNT_FAM_MEMBERS

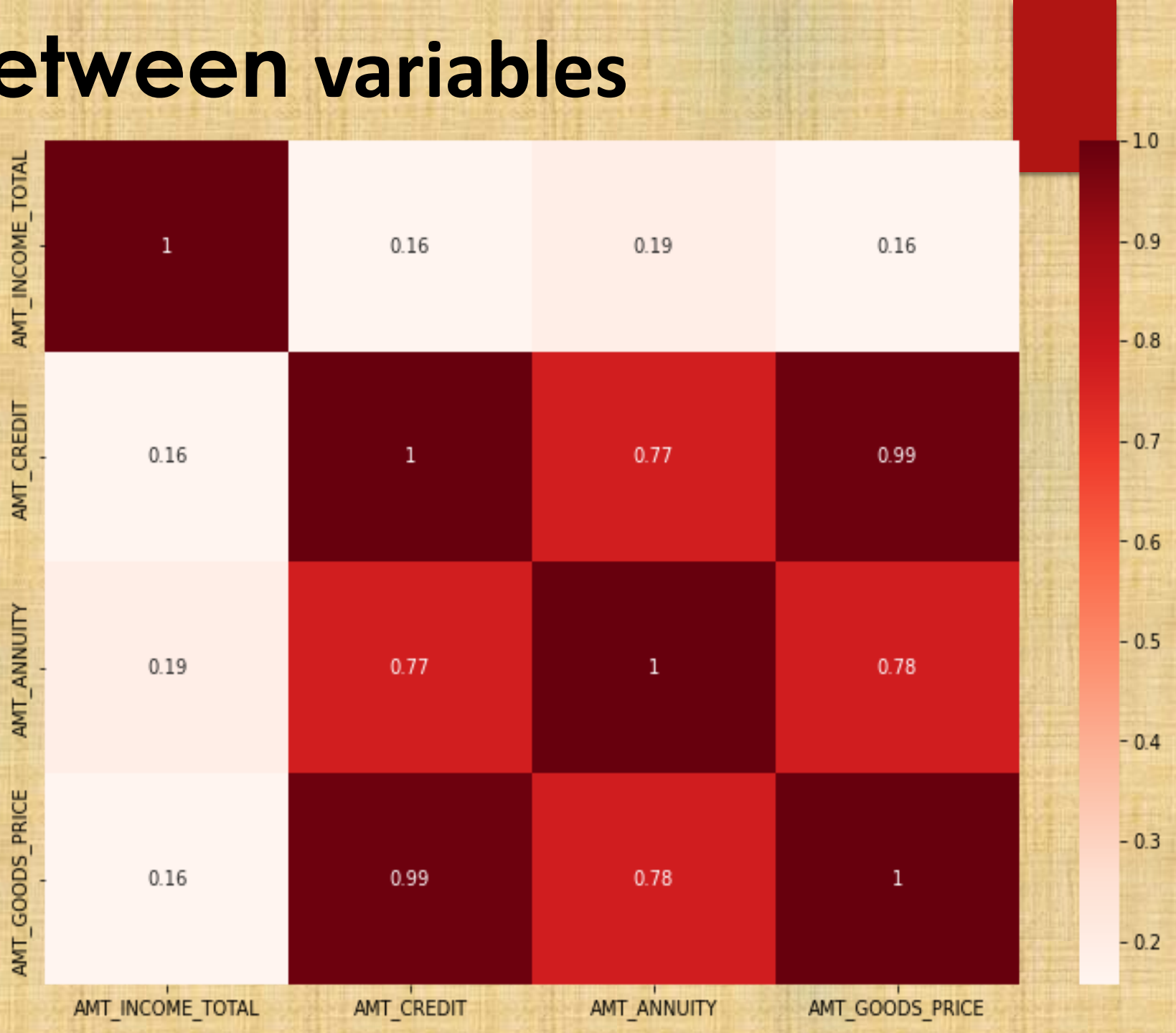
Categorical Bivariate/Mltivariate Analysis

PAIR PLOT

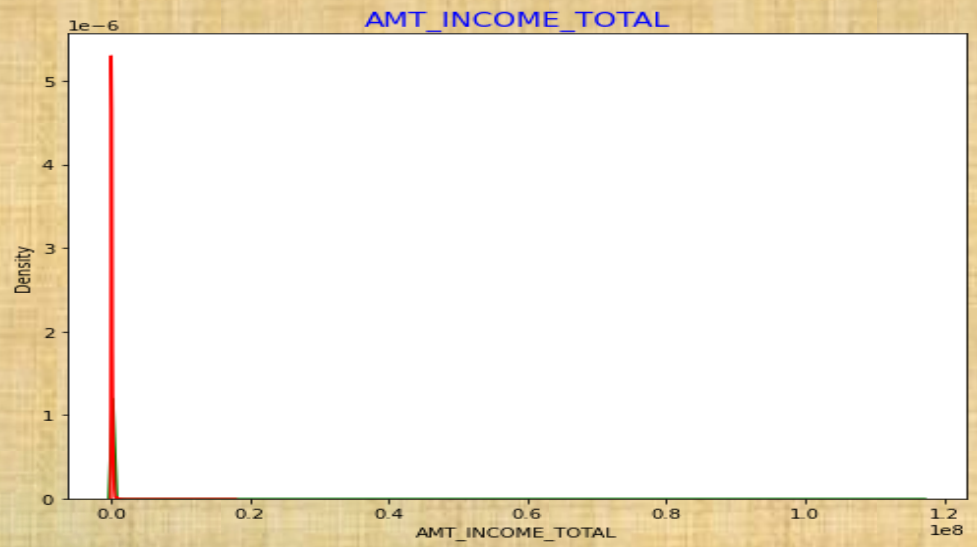


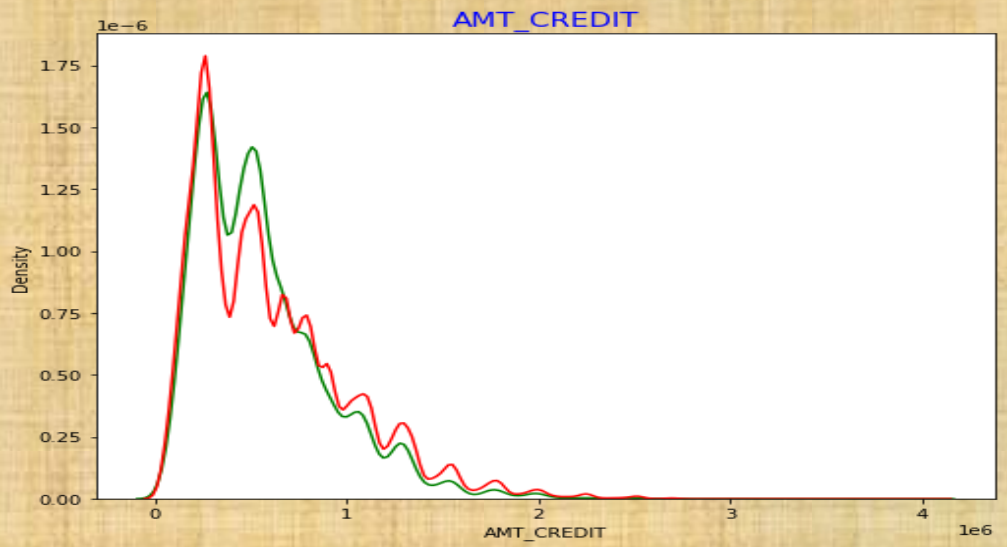
Correlation between variables

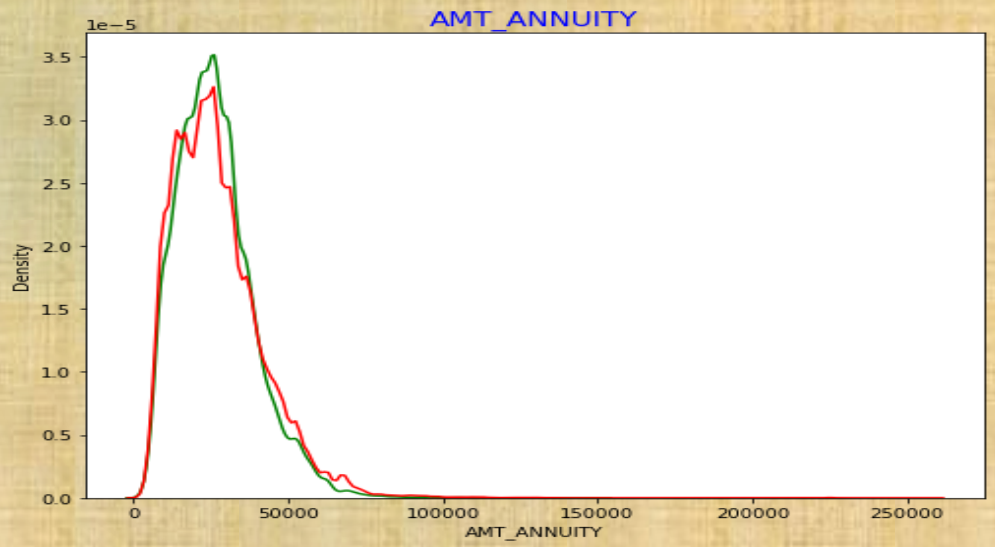
Insights of Heatmap::: Found very High Correlation between AMT_CREDIT and AMT_GOODS_PRICE. Applicants who are owning goods of high value can take loans of higher amounts...

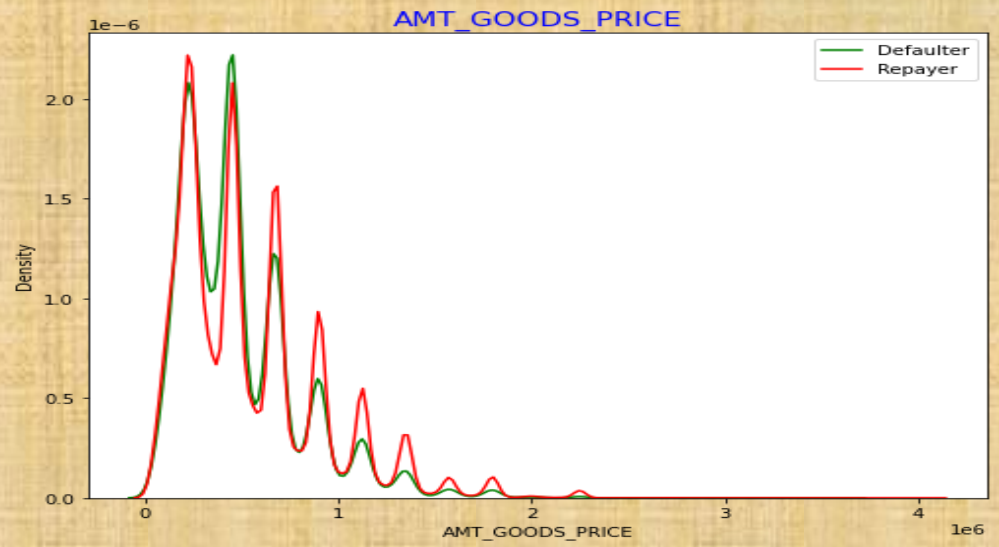


Numerical Univariate Analysis

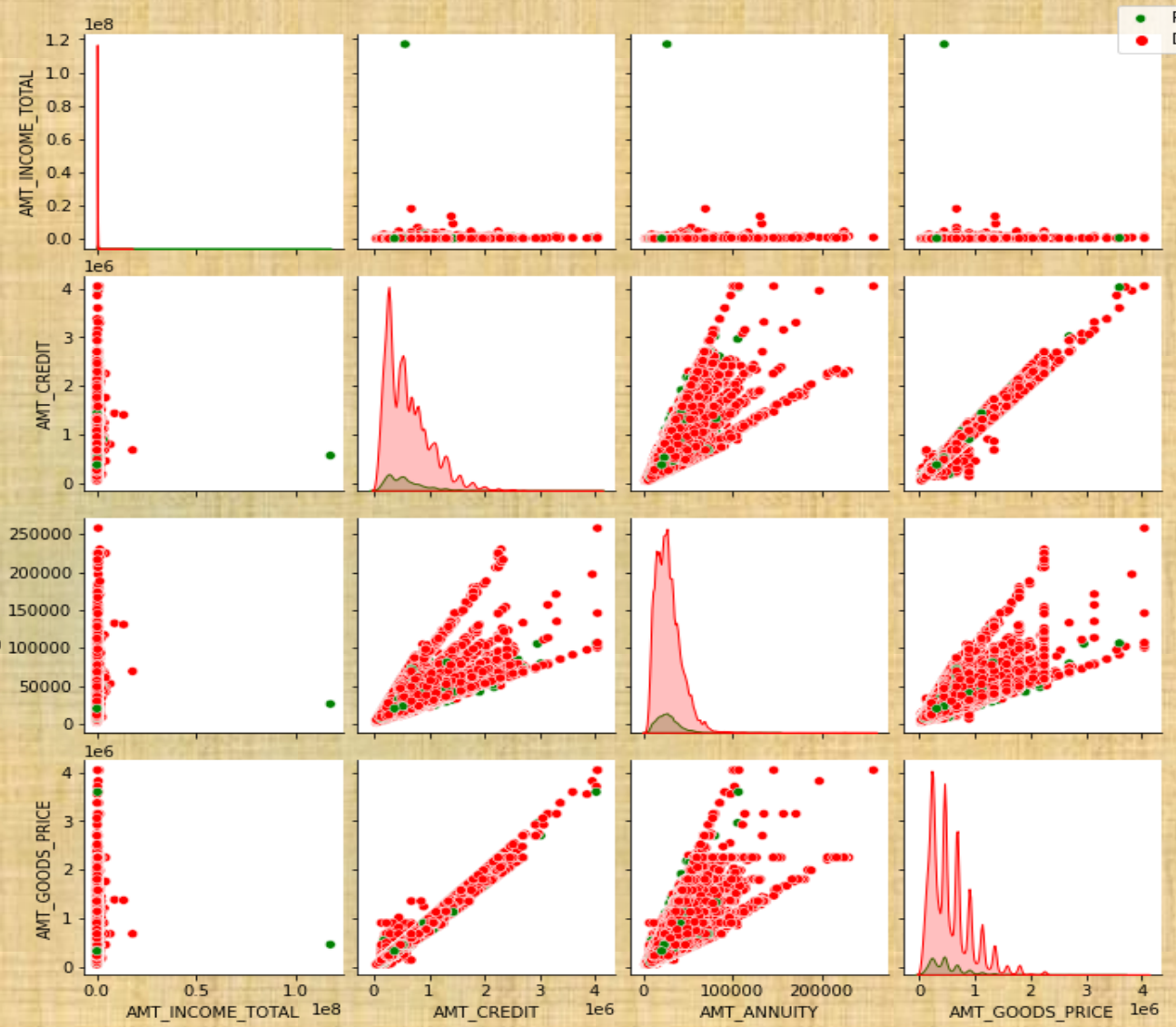


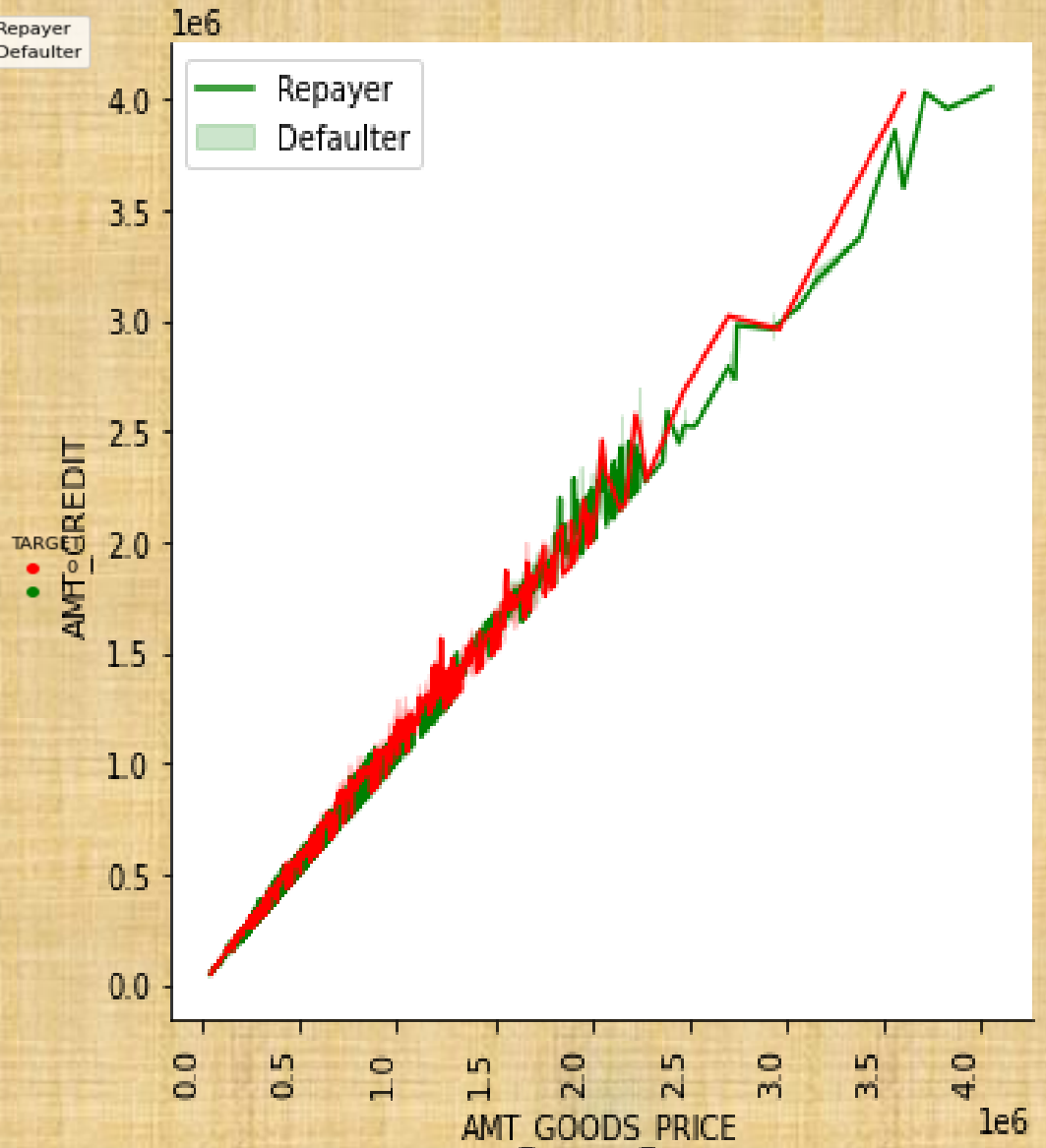




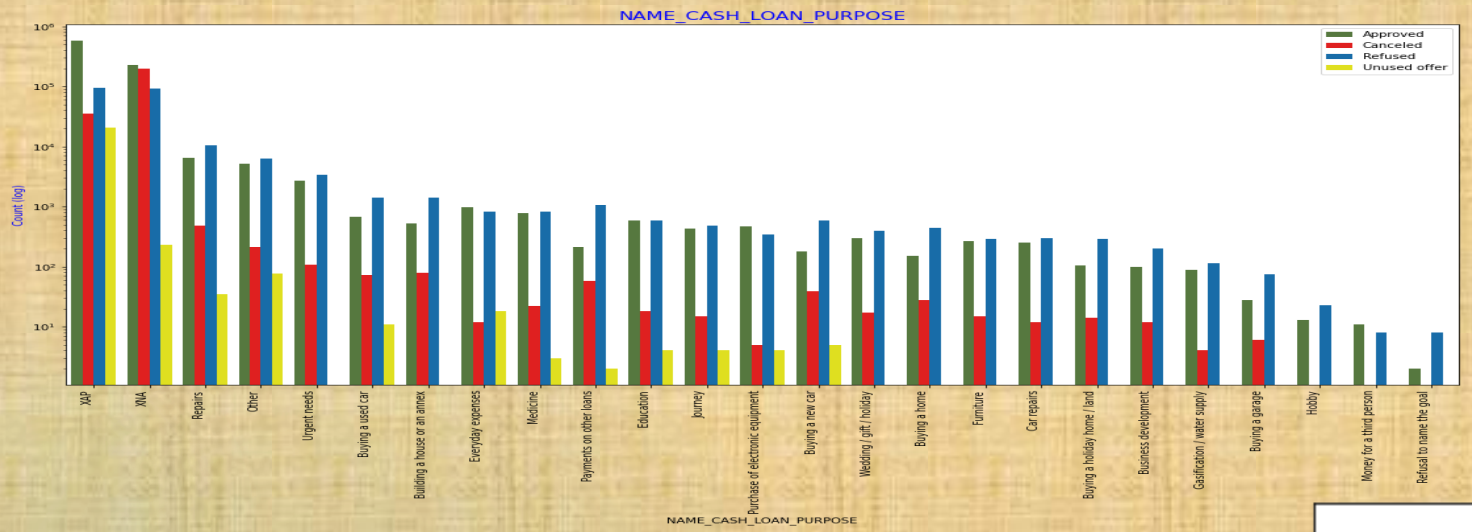


NUMERICAL BIVARIATE ANALYSIS

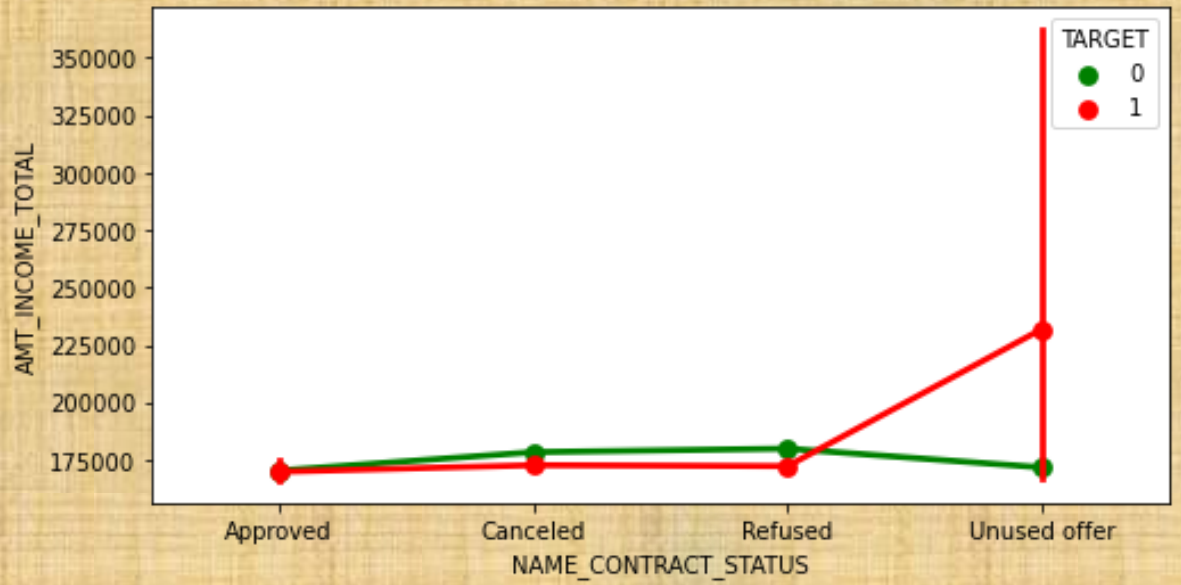




MRGED DATA FRAMES UNIVARIATE ANALYSIS



BIVariate Analysis¶ __



Final Conclusions from the Case Study:::

1. After analysing the datasets, found there are few attributes of a client with which the bank would be able to identify if they will repay the loan or not.

2.The analysis is considered as below with the contributing factors and categorization.

::::The Resultant Factors for whether an applicant will be a Re-Payer, hence the Apllications can be approved:::::

I. DAYS_BIRTH: The low probability of defaulting for the applicants age above 50.

II. DAYS_EMPLOYED: The Clients who are with 40+ years experience having less than 1% default rate.

III. CNT_CHILDREN: Clients who are having with zero to two children tend to repay the loans.

IV. AMT_INCOME_TOTAL: Applicants who are with Income more than 7lacs. are less likely to be default.

V. NAME_CASH_LOAN_PURPOSE: The Applicants of the Loans who are bought for Hobby, Buying garage are being repayed mostly.

VI. NAME_INCOME_TYPE: They have no defaults for Student and Businessmen.

VII. NAME_EDUCATION_TYPE: Academic degree has less defaults.

VIII. ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%

IX. REGION_RATING_CLIENT: Applicants who live in areas with Region Rating 1 are safe borrowers.

::::The Resulatant Factors whether an applicant will be a Potential Defaulter, hence the applications can be Rejected::::

I. CODE_GENDER: The Male applicants are under relatively higher default rate.

II. DAYS_BIRTH: Avoiding of young people who are in age group of 20-40 as they have higher probability of defaulting.

III. DAYS_EMPLOYED: Applicants who are having less than 5years of employment have high default rate.

IV. OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is very high.

V. NAME_FAMILY_STATUS : Applicants who have civil marriage or who are single default a lot.

VI. NAME_EDUCATION_TYPE: Applicants with Lower Secondary & Secondary education

VII. NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.

VIII.AMT_GOODS_PRICE: When the credit amount goes beyond 3M, there is an increase in defaulters.

IX. CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.

Suggestions from Analysis:::

**** Previously cancelled clients have actually repayed the loan of 90%. Record the reason for cancellation which might help the bank to determine and negotiate terms with these repaying customers in future for increase business opportunity.

****88% of the clients who were refused by bank for loan earlier have now turned into a repaying client. Hence documenting the reason for rejection could mitigate the business loss and these clients could be contacted for further loans.

THANK YOU