EDA CASE STUDY: _Credit risk analysis_

BY KAMESH VISHWAKARMA AND A V RAMYA KEERTHANA

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Problem Statement

The company has to decide for loan approval based on an applicant's profile. There are two types of risks 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.

Aim

- Identify patterns in the dataset to ensure that the applicants are capable of repaying the loan are not rejected.
- Understand the influence of consumer and loan attributes on the tendency of default.
- understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default

Approach

We perform Exploratory Data Analysis (EDA) on the given datasets with the help of the following information:

Loan payment status as per the 'application_data.csv' file -

- Client with payment difficulties (‘0’) - Defaulter
- All other cases (‘1’) – Re-payer

Decision on loan application as per the 'previous_application.csv' file -

- **Approved** - by the Company
- **Cancelled** - by the Client
- **Refused** - by the Company
- **Unused offer** - by the Client

Steps of this EDA

1. **Data sourcing** -

    a. Importing and Reading the data -

    b. Data Inspections

    c. Inspecting data frames

    - variable types
    - dimensions
    - dataframe info
    - Statistical description

Data Cleaning and Manipulation

2. **Data cleaning -**

    - handling null values
    - Standardize values
    - Null value data imputation
    - identifying outliers
    - deleting unnecessary columns
    - Identifying outliers

Data Sourcing

3. Data Analysis -

-Imbalance Analysis

- Categories Dataframe into as Defaulters and Repayers

- Categorical Analysis :

  a. univariate categorical analysis

  b. Bivariate/Multivariate categorical analysis

- Numerical Analysis :

  a. univariate numerical analysis

  b. Bivariate/Multivariate numerical analysis

Merged Dataframe Analysis and Conclusion

4. Merged Dataframe Analysis -

- merged dataset reading and understanding
- categorize dataset into Defaulter and Repayer for analysis
- Columns univariate and Multivariate analysis
- driving short inferences from merged dataframe

5. Conclusion –

- deriving conclusions from full EDA process.

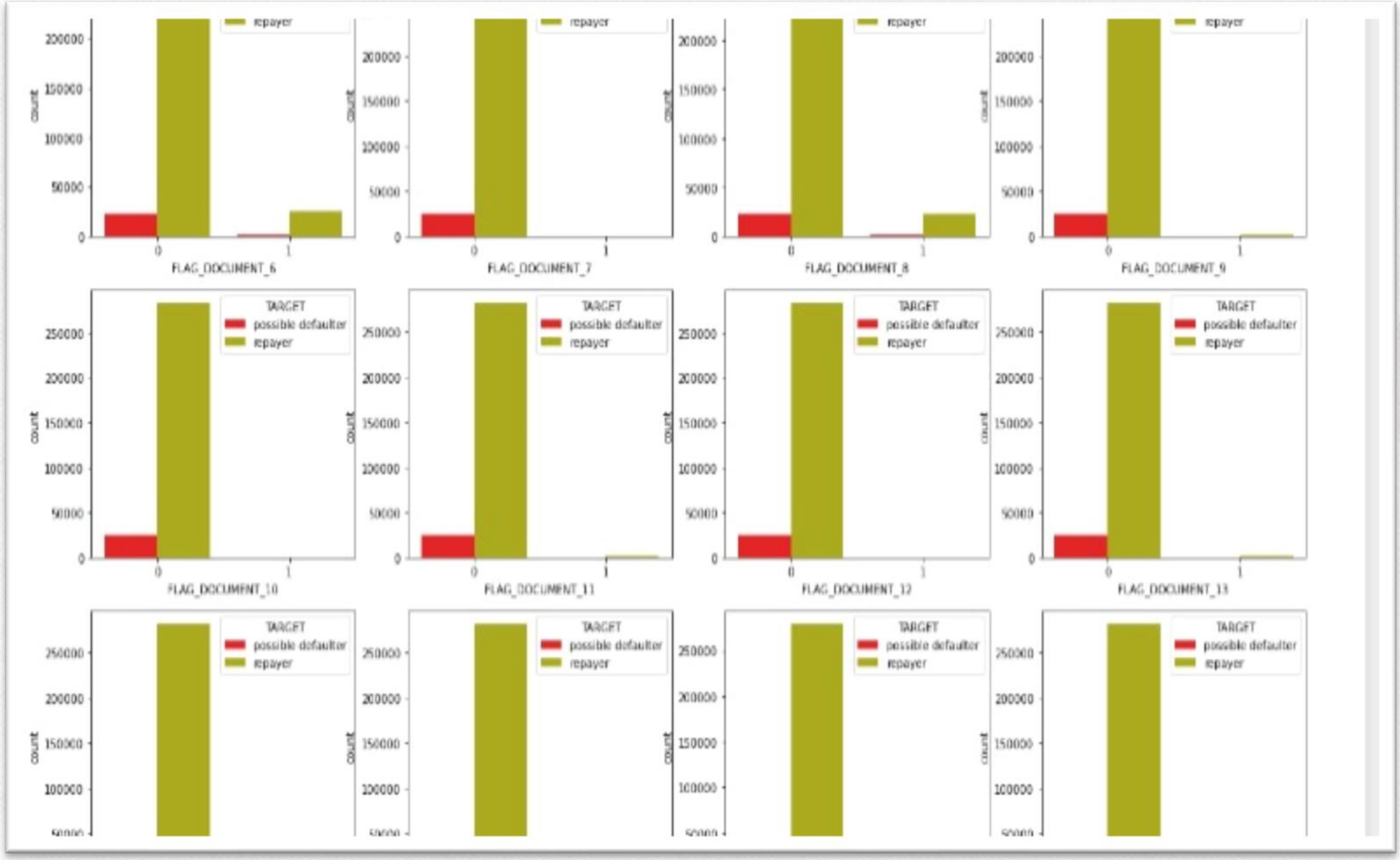
Data Cleaning

Handling Null Values - (appdata)

- The % of missing values in each column of appdata is found.
- It is observed that there are many columns with high missing values (more than 40%). Such columns are handled by either dropping them others can be managed by imputing values in them based on their relevance.
- There are 49 such columns found and most of them are related to different area sizes or apartment owned/rented by the loan applicant. As such information is not of much significance for this analysis we store these columns in a list (app_nulldata_40 & PREV_NULLDATA_40) to be dropped from the data frame.

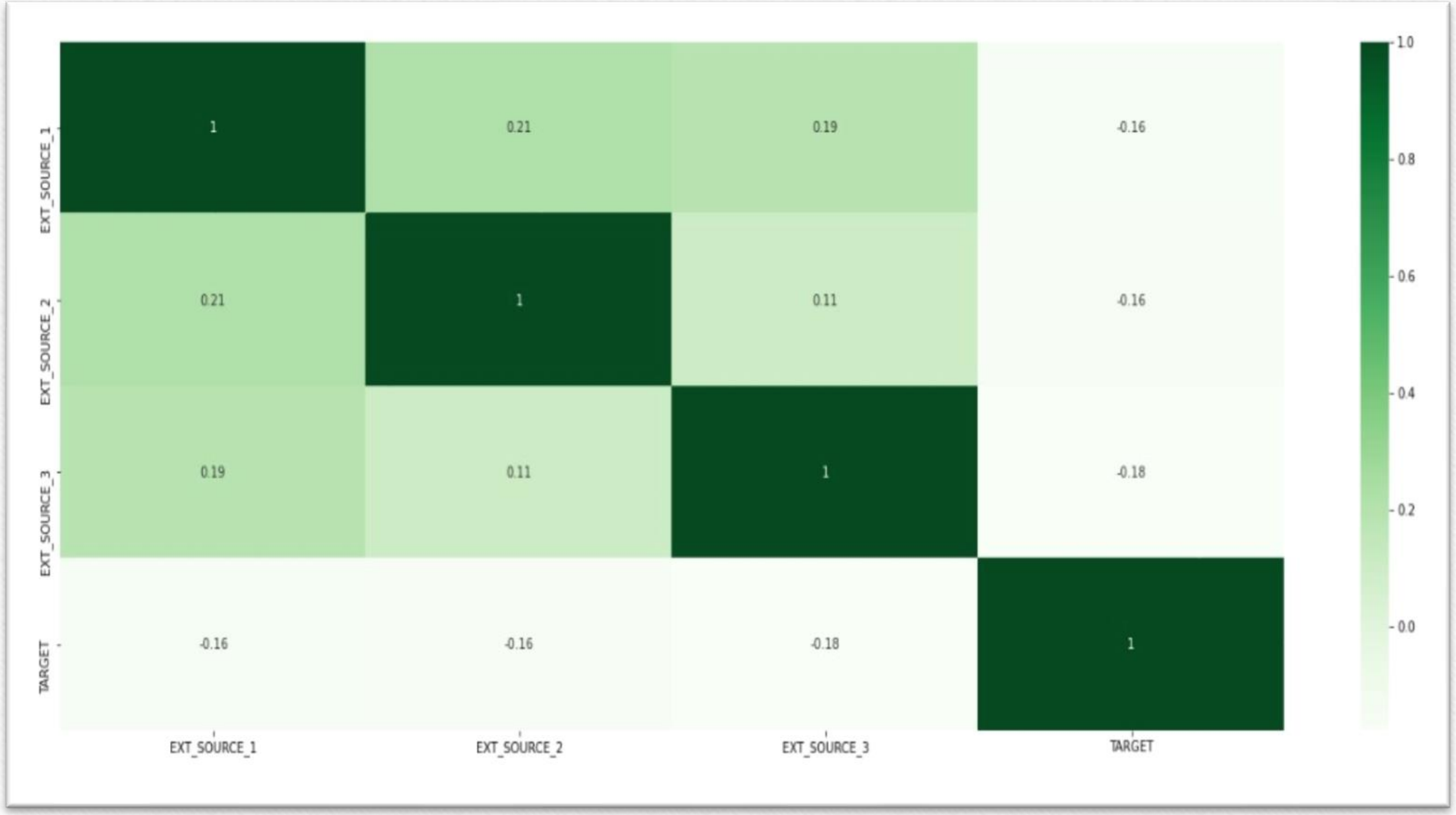
Flag Document –

- It is observed from the count plots that most applicants have not submitted FLAG_DOCUMENT_X except FLAG_DOCUMENT_3. Also, such applicants have a lesser chance of defaulting the loan. Hence, among them only FLAG_DOCUMENT_3 is of relevance and the rest can be dropped.

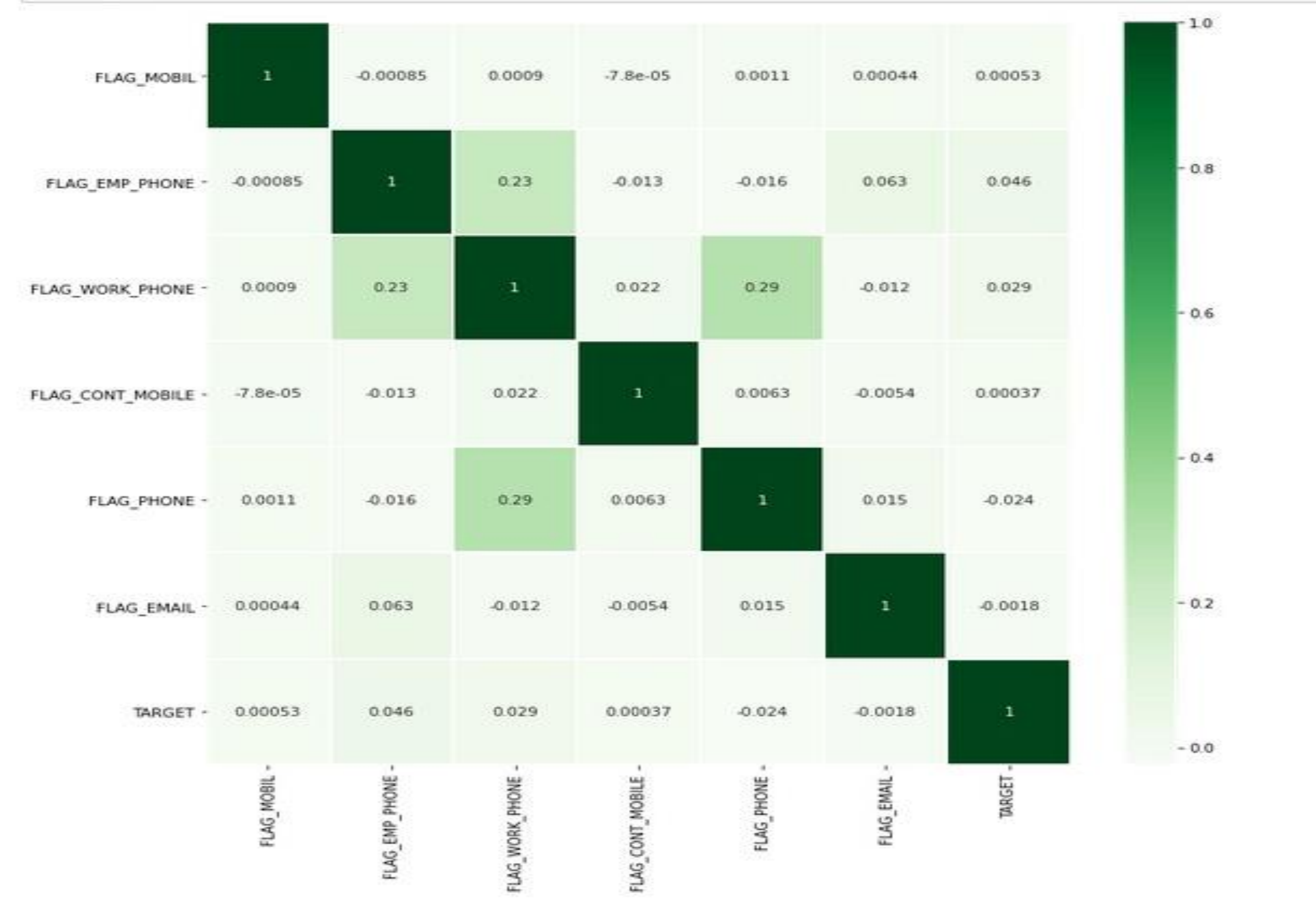


EXT_Source -

- It is observed from the heat map that there is almost no correlation between EXT_SOURCE_X columns and Target column. So these columns can be dropped as well



Contact Flag –



	FLAG_MOBIL	FLAG_EMP_PHONE	FLAG_WORK_PHONE	FLAG_CONT_MOBILE	FLAG_PHONE	FLAG_EMAIL	TARGET
FLAG_MOBIL	1	-0.00085	0.0009	-7.8e-05	0.0011	0.00044	0.00053
FLAG_EMP_PHONE	-0.00085	1	0.23	-0.013	-0.016	0.063	0.046
FLAG_WORK_PHONE	0.0009	0.23	1	0.022	0.29	-0.012	0.029
FLAG_CONT_MOBILE	-7.8e-05	-0.013	0.022	1	0.0063	-0.0054	0.00037
FLAG_PHONE	0.0011	-0.016	0.29	0.0063	1	0.015	-0.024
FLAG_EMAIL	0.00044	0.063	-0.012	-0.0054	0.015	1	-0.0018
TARGET	0.00053	0.046	0.029	0.00037	-0.024	-0.0018	1

- It is observed from the heat map that there is no correlation between Contact flag columns and Target column. Similarly, these columns can also be dropped.

Handling Null Values - (prevappdata)

○ Similarly the % of missing values in each column of 'previousDF' is found.

○ It is observed that there are many columns with high missing values (more than 40%). Such columns are handled by either dropping them or imputing values in them based on their relevance.

○ There are 11 such columns found and most of them are related to interest rates and days of due payment for the loan. These columns are stored in a list (Unwanted_previous) to be dropped from the data frame.

○ Other columns that are not necessary for this analysis are also added to the list Unwanted_previous for them to be dropped.

○ Thus there are only 22 columns remaining after dropping the non relevant columns from the previous_df for this analysis.

Standardize Values –

- The columns related to count of days are converted to their absolute values as days cannot be negative.
- The significant numerical columns are converted to categorical columns by grouping them into bins.
- Datatypes are changed for certain categorical variables.

Null Value Data Imputation - (appdata)

- Checking null value % of the remaining columns and imputing/ignoring them accordingly.

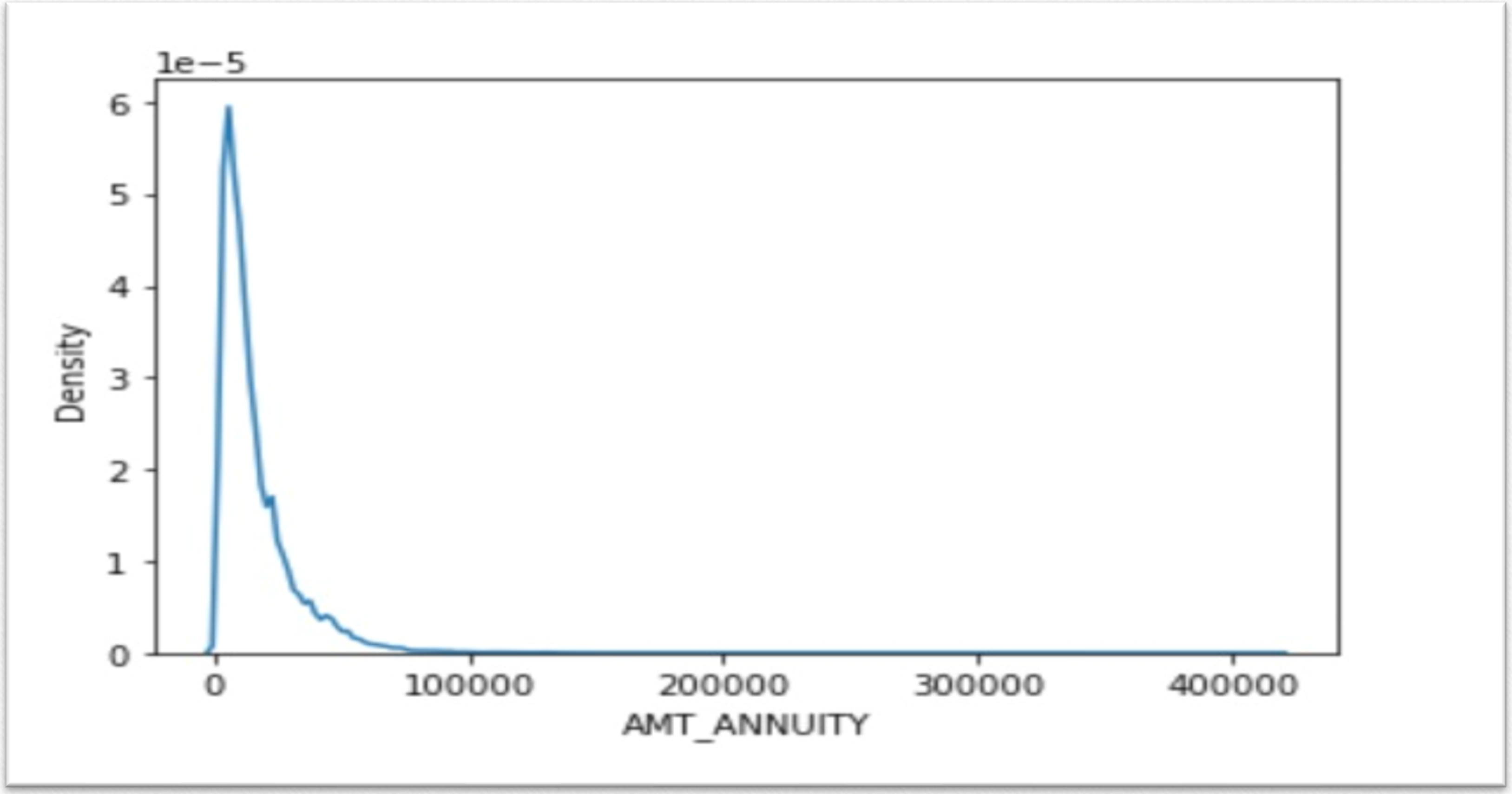
Records in columns with **very low % of missing values are ignored** for this analysis.

- The categorical variable **'NAME_TYPE_SUITE'** which has **lower null percentage(0.42%)** is imputed with the most frequent category (**Mode**).
- The categorical variable **'OCCUPATION_TYPE'** which has **higher null percentage(31.35%)** is imputed with a **new category (Unknown)** as assigning any existing category might influence the analysis.
- The columns representing the **number of enquiries made** are imputed with their r**espective median values** as there are **no outliers** as per their summary statistics. But as their respective **means are in decimal,** they cannot be used to impute count of enquiries.

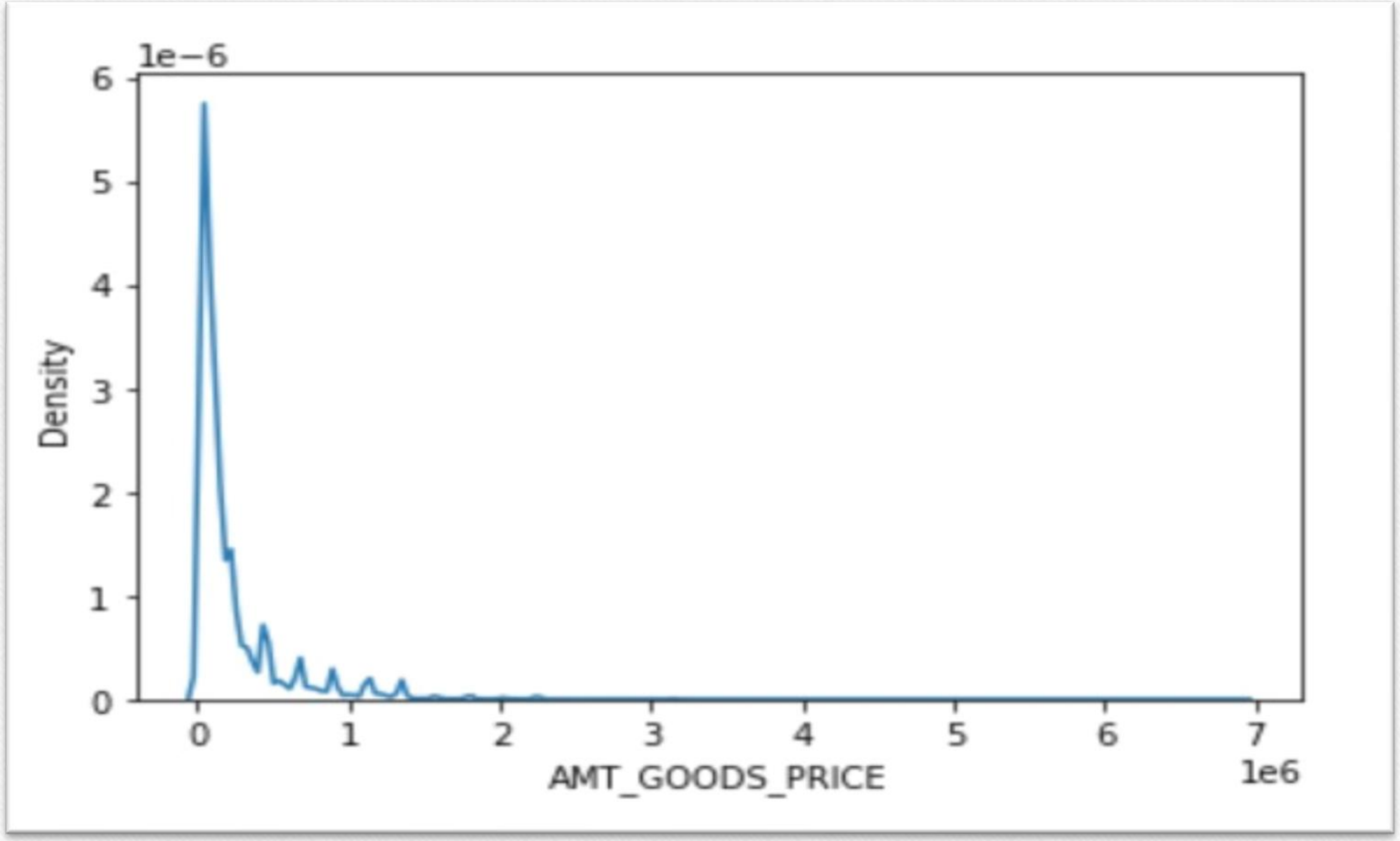
Null Value Data Imputation - (prevappdata)

- Checking null value % of the remaining columns and imputing/ignoring them accordingly.
- ‘PRODUCT_COMBINATION’ column has very low % of missing values and thus such records are ignored for this analysis.
- Missing values for ‘CNT_PAYMENT’ are imputed with 0 as NAME_CONTRACT_STATUS for these indicate that most of these loans were not started.

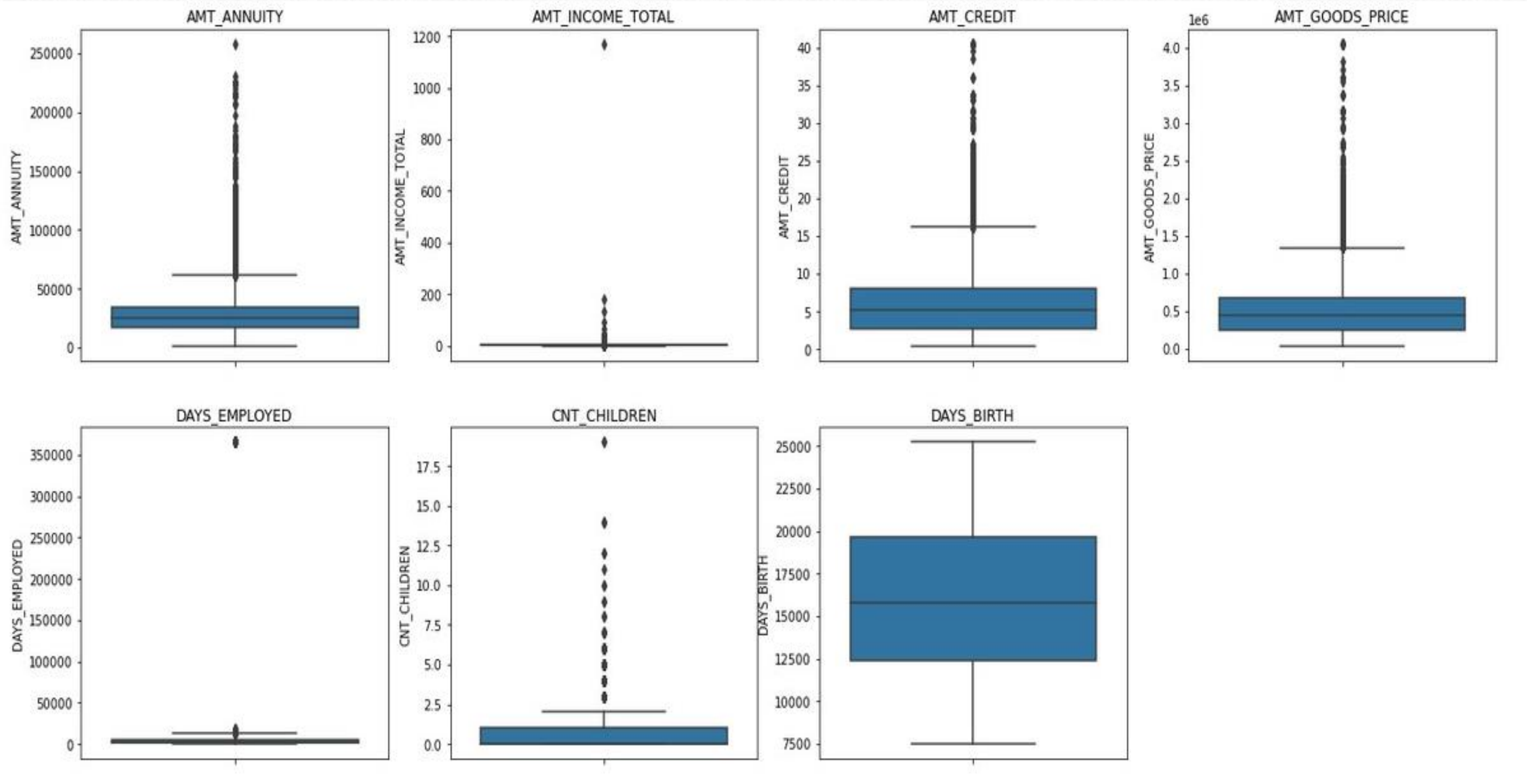
A single peak at the left side for the AMT_ANNUITY distribution indicate skewness, i.e., presence of outliers and are thus imputed with median values to avoid exaggeration of data.



It is observed the distribution of AMT_GOODS_PRICE data is closer to its distribution when imputed with Mode and hence it is imputed accordingly.



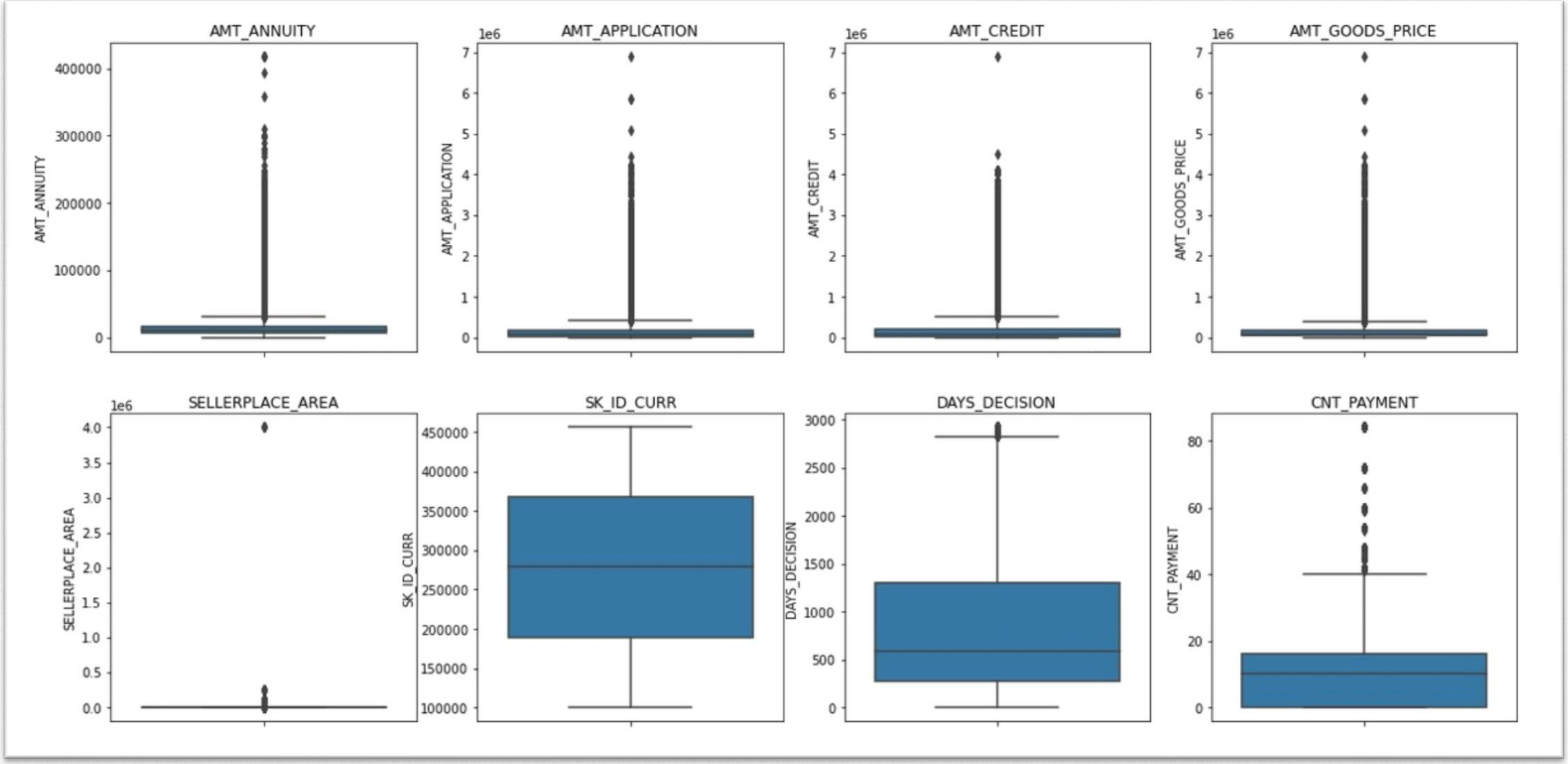
Identifying Outliers - (applicationDF)



Boxplots are used to identify outliers for relevant columns and the following observations are made:

- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, CNT_CHILDREN, CNT_FAM_MEMBERS have some outliers.
- AMT_INCOME_TOTAL has a large number of outliers which indicates that few of the loan applicants have higher income as compared to others.
- DAYS_BIRTH has no outliers which means the available data is reliable.
- DAYS_EMPLOYED has extreme outlier values at around 350,000 days (i.e. about 958 years), which is not possible and hence such values have to be considered as incorrect entry.

- ## Identifying Outliers - (previousDF)

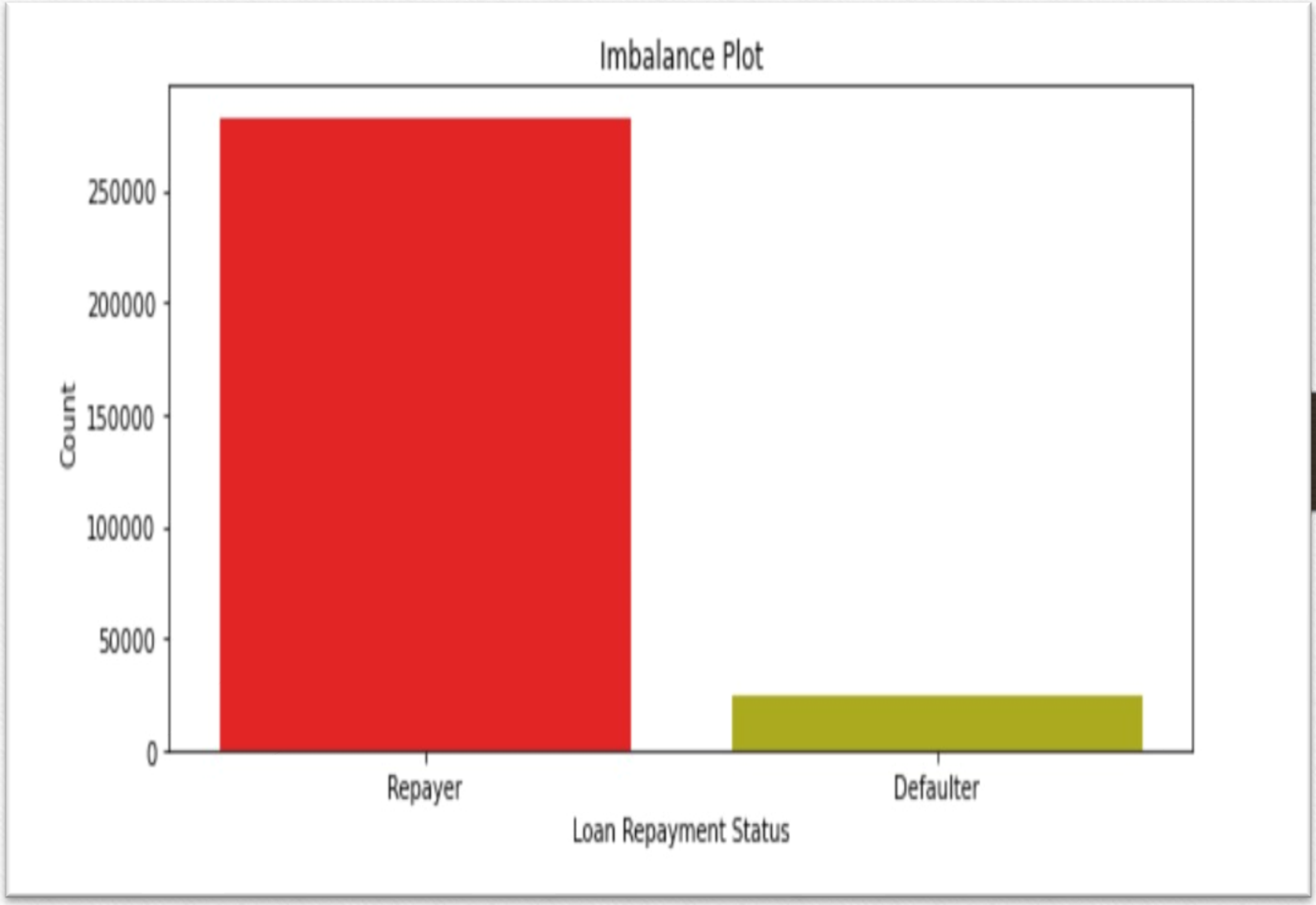


Similarly, Boxplots are used to identify outliers for relevant columns. The following observations are made:

- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have a large number of outliers.
- CNT_PAYMENT has fewer outlier values.
- DAYS_DECISION has very few outliers indicating that decisions for these previous applications were taken long back.

Data Analysis

- Imbalance Data – Count of the target variable is plotted to determine the percent ratio of Re-payers to Defaulters, which is found to be : Repayer - 91.93 % and Defaulter - 8.07 %.
- Imbalance Ratio in relative with respect to Repayer and Defaulter data is 11.39:1

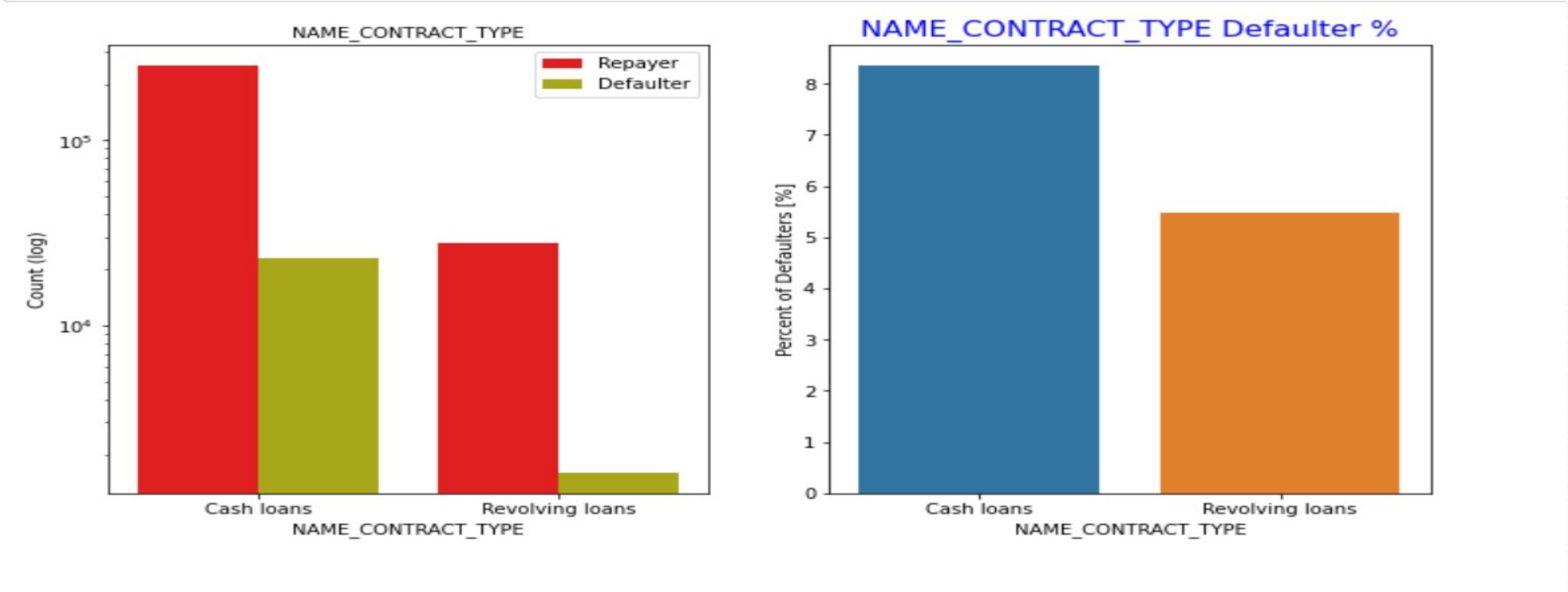


Univariate Categorical Analysis -

It is performed based on repayment loan status (TARGET) with the help of count plot and bar plot for percentage of defaulters of such columns. The observation made are as below.

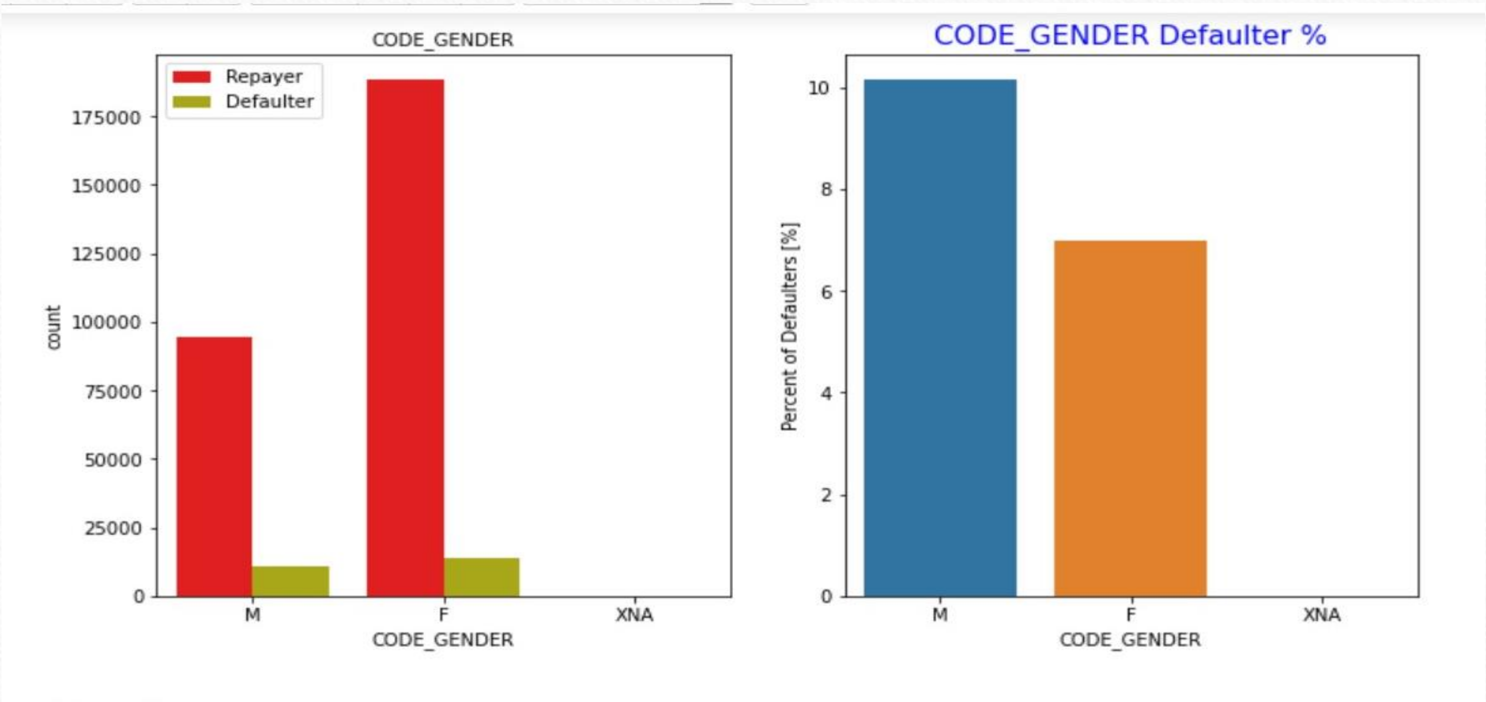
NAME_CONTRACT_TYPE -

- The Contract type - 'Revolving loans' are just a small fraction of the total number of loans. Also, a larger
- amount of Revolving loans when compared to their frequency, are not being repaid.



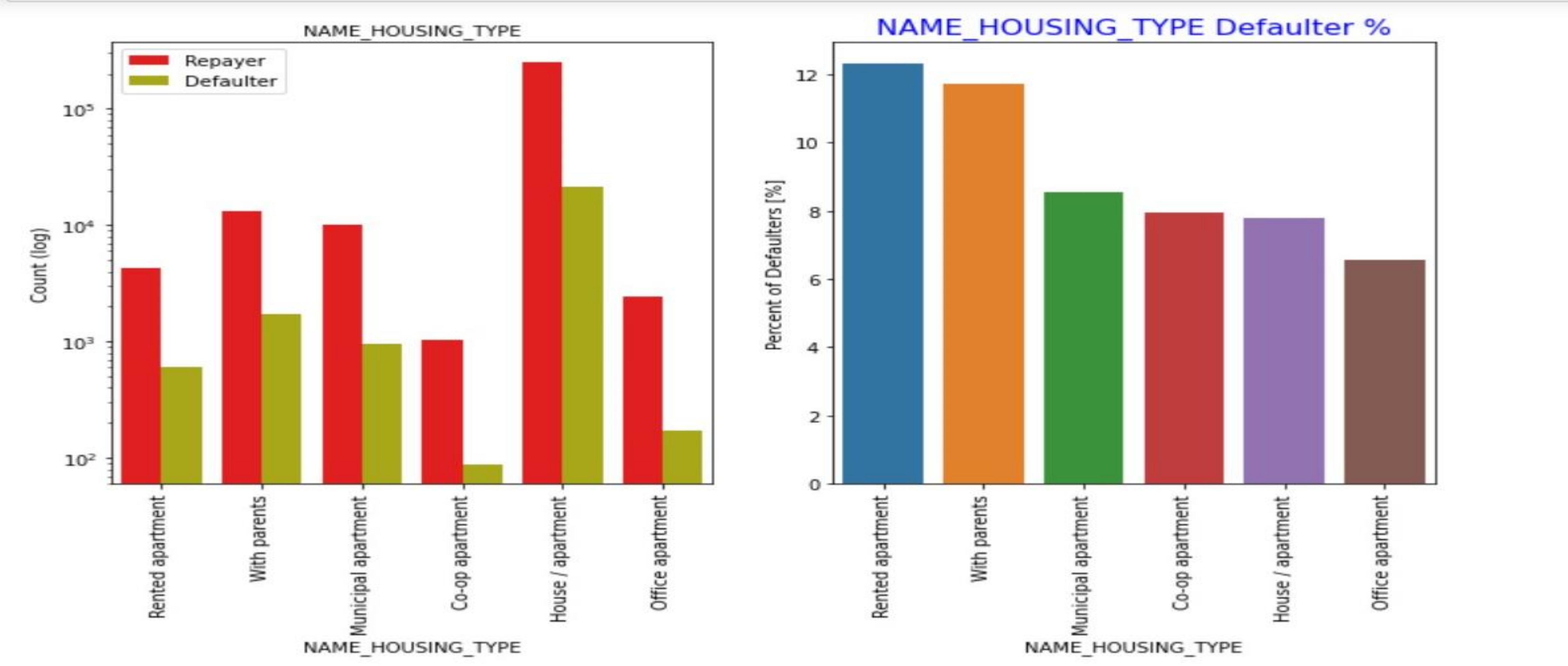
CODE_GENDER –

- The number of female clients is almost twice the number of male clients. Based on the % of defaulted credits, males are more likely to Default as compared to females.



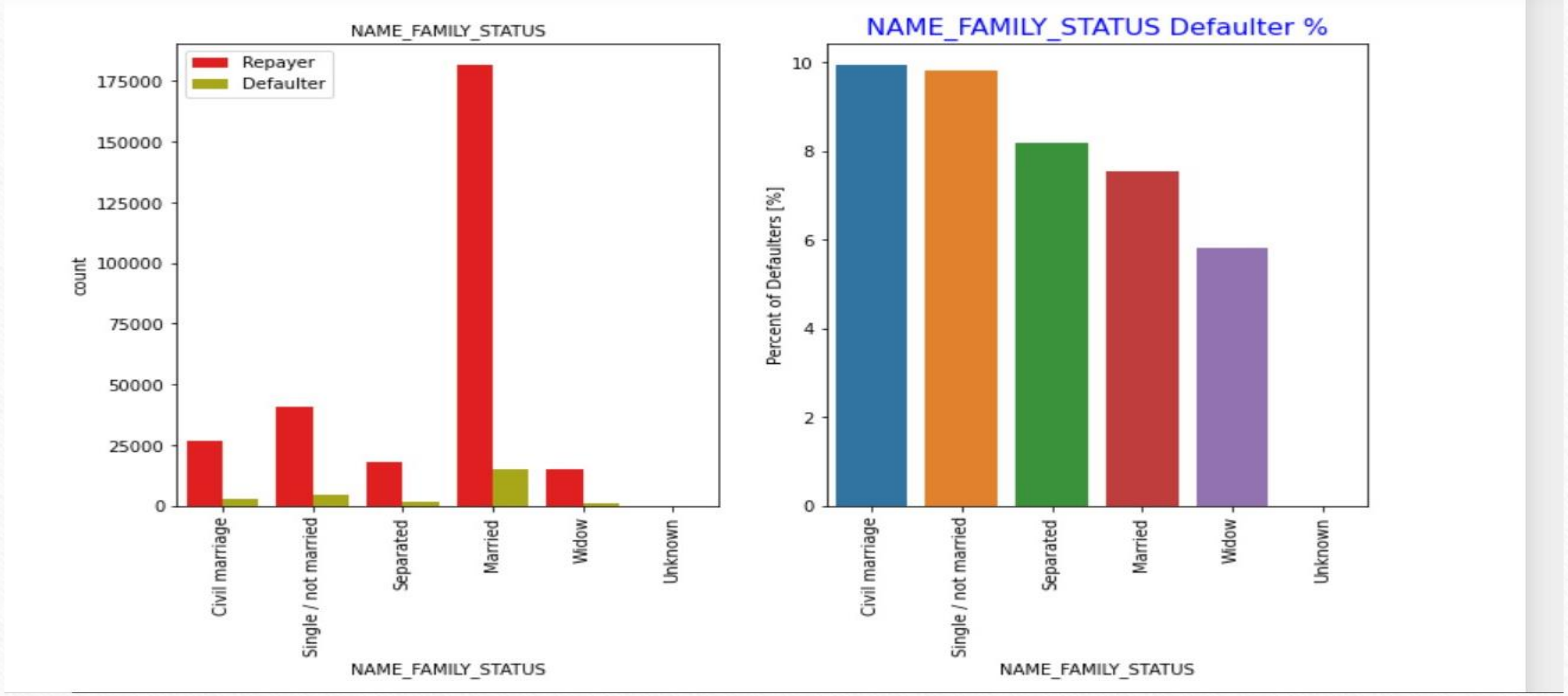
NAME_HOUSING_TYPE -

- **Most clients** live in **House/Apartment.**
- **Clients living in Rented Apartments** are **most likely to be Defaulters(>12%)**.
- **Clients living With parents** also have a **higher % (almost 12%) to be Defaulters**.
- **Clients living in the Office apartment** are **least likely to be Defaulters.**



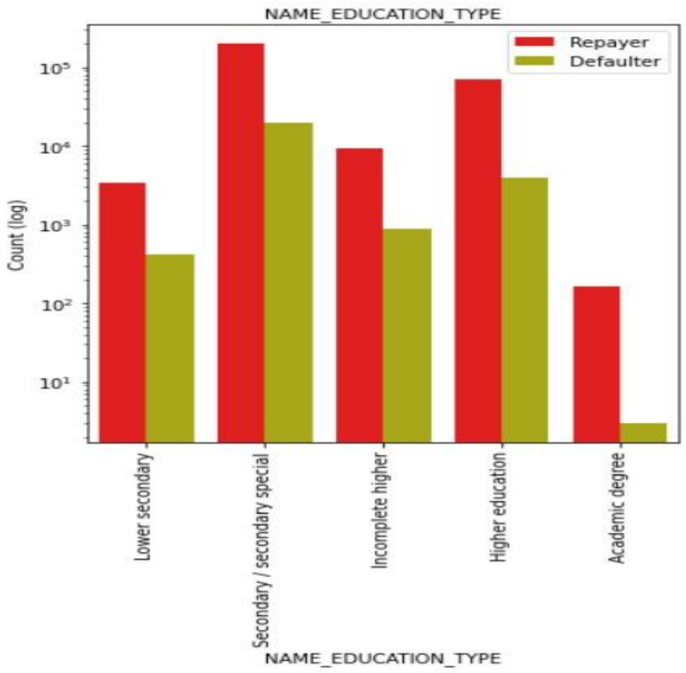
NAME_FAMILY_STATUS -

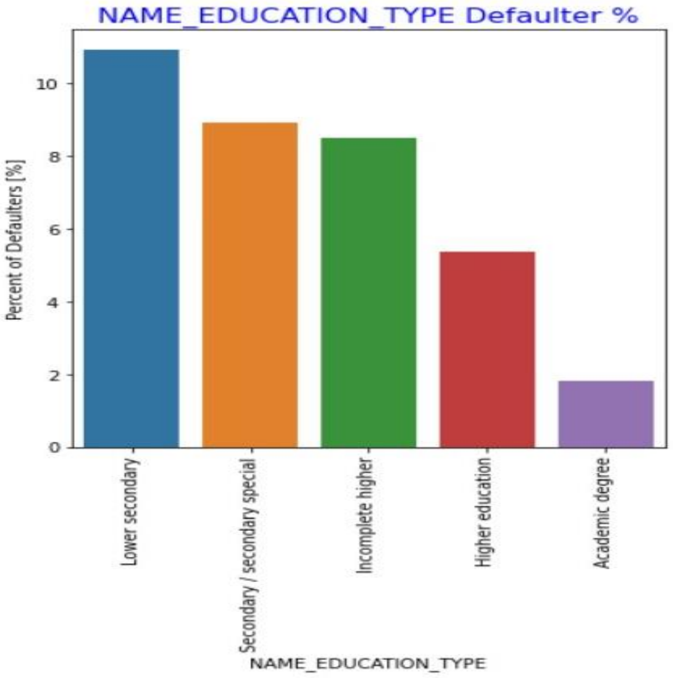
- Most clients are Married.
- Clients with Civil marriage are having high Defaulter% (10%) and then Single/not married (almost 10%).
- Widow clients are the least Defaulters.



NAME_EDUCATION_TYPE -

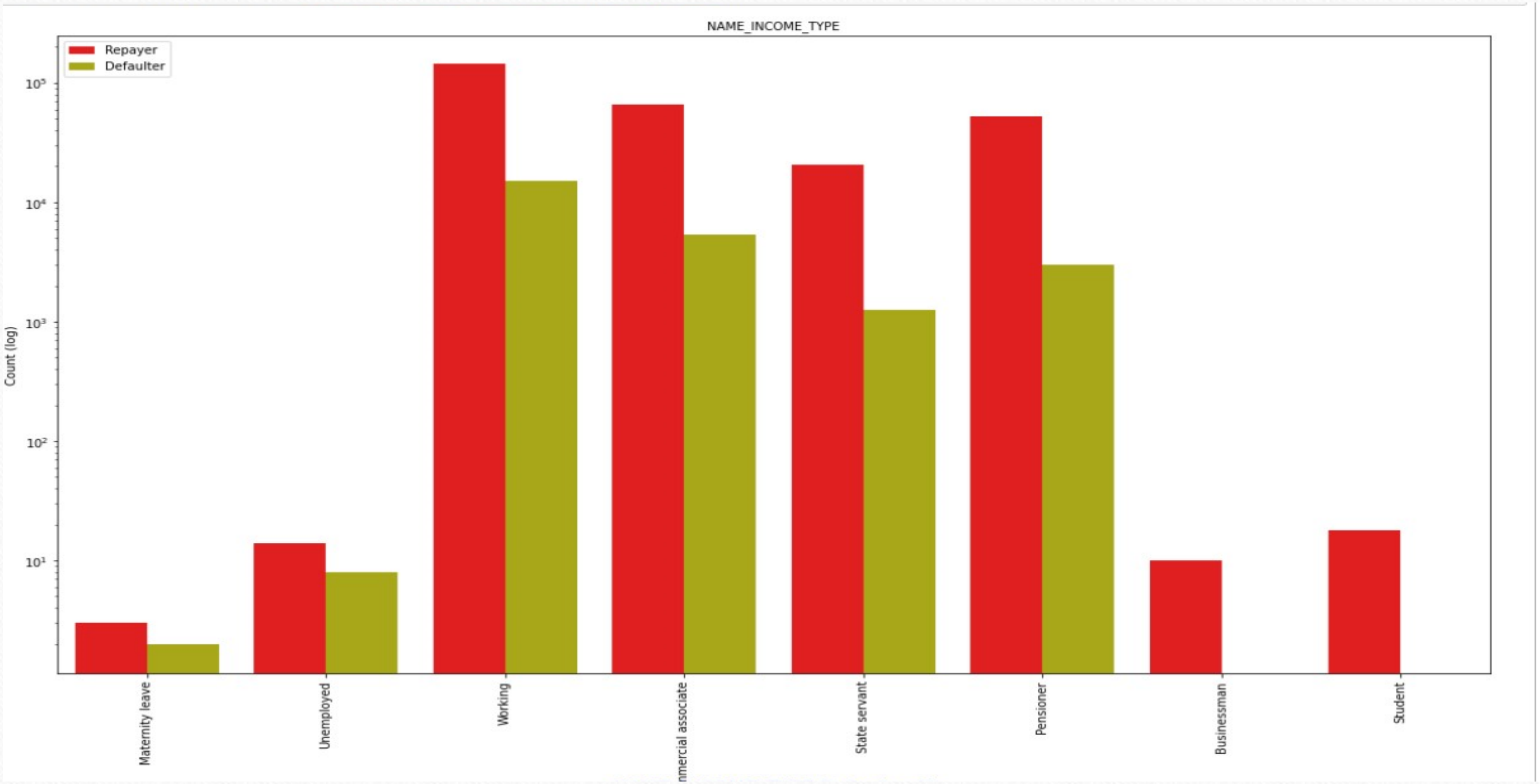
- Most clients are having Secondary/secondary special education.
- Clients with lower secondary (even though they are of very low numbers) have high Defaulter% (>10%) and those with Secondary/secondary special education (about 9%).
- Academic degree clients are the least Defaulters.

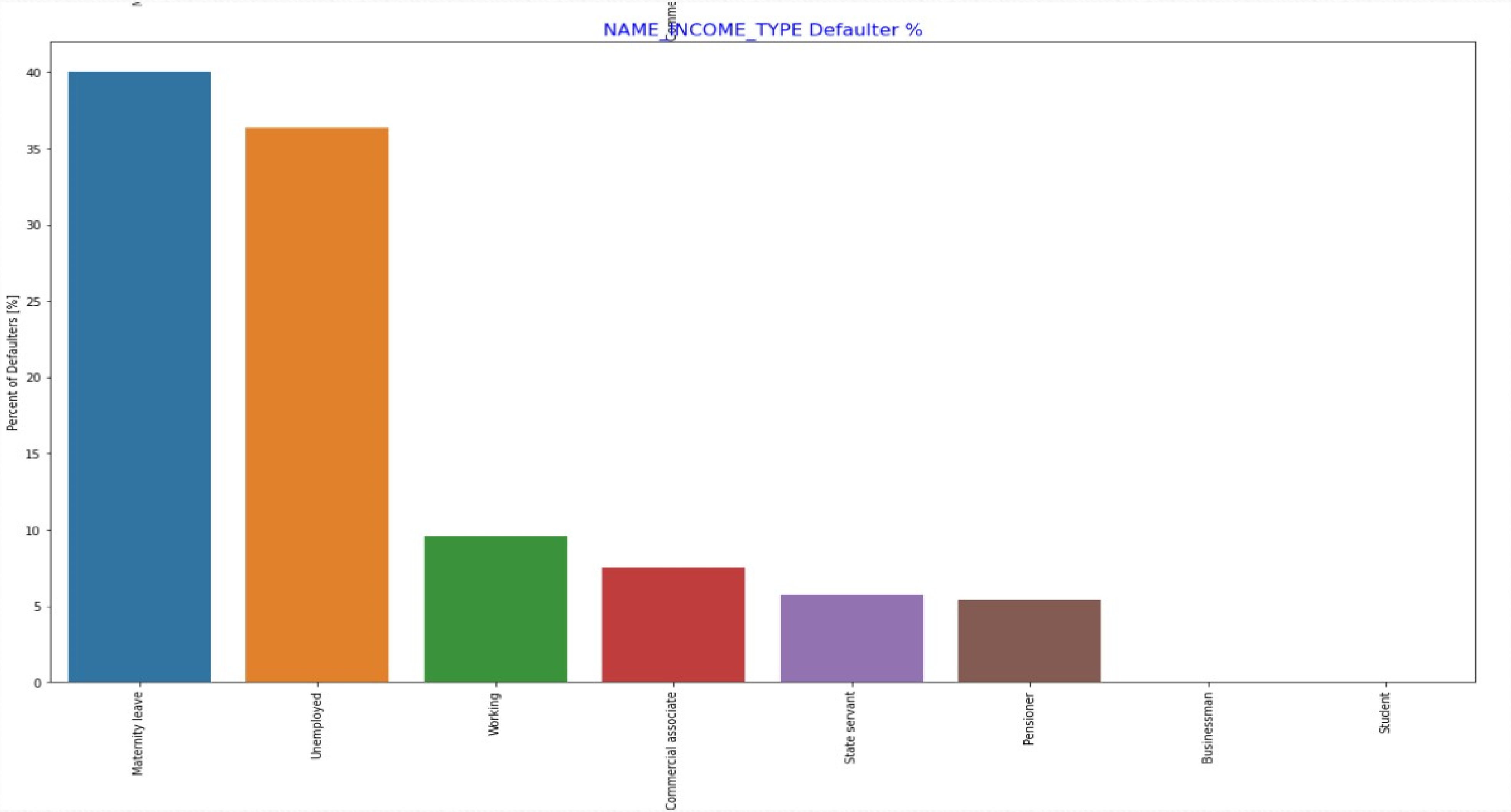




NAME_INCOME_TYPE –

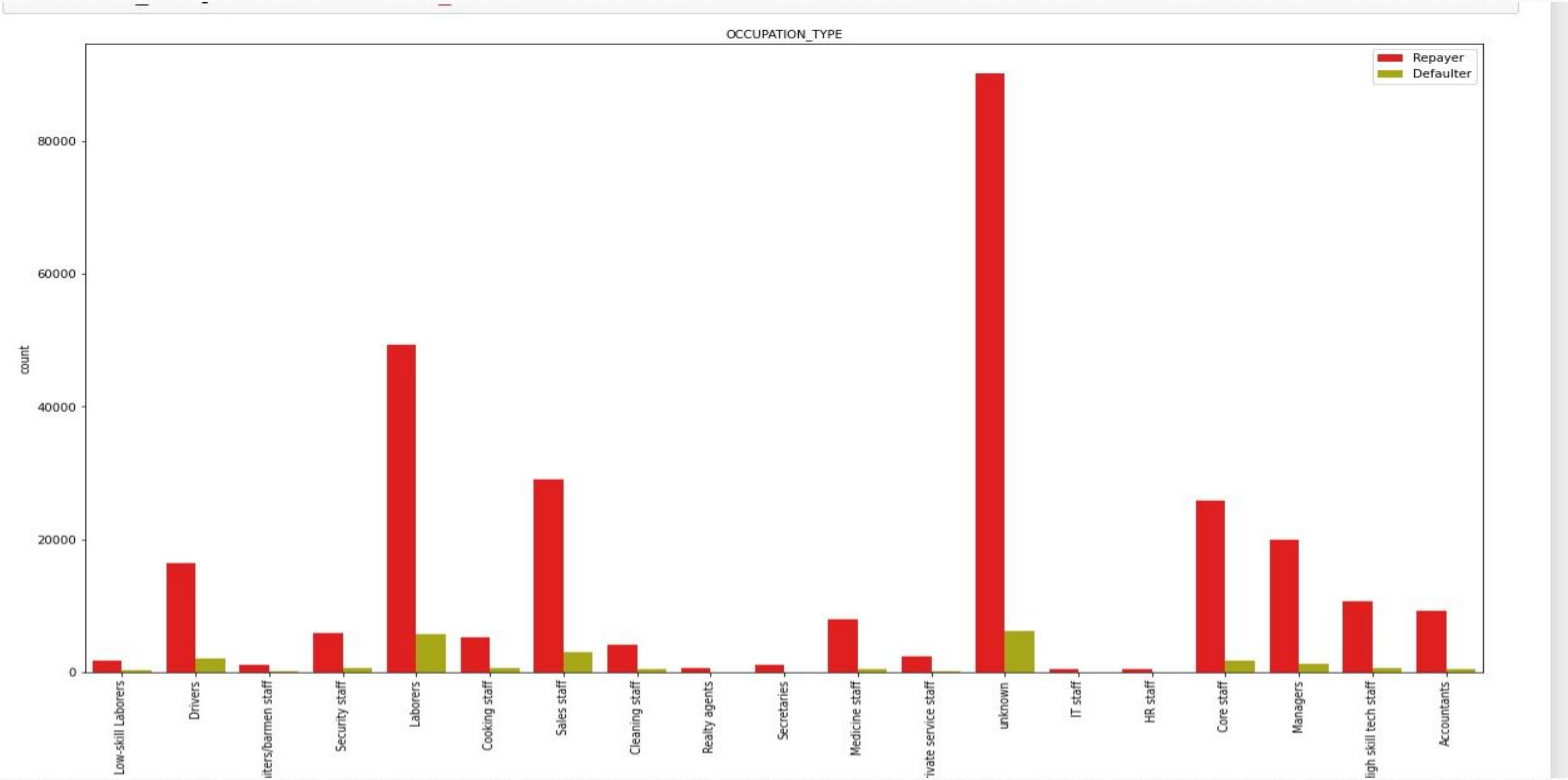
- Most clients have INCOME Category as Working, Commercial associate, Pensioner and State servant.
- Clients on Maternity leave (even though they are of very low numbers) have high Defaulter% (almost 40%) and those who are Unemployed (>35%).
- Pensioner clients are the least Defaulters.
- Students and Businessmen, though less in numbers have no Default record. Thus these two categories are safest for providing loan.


NAME_INCOME_TYPE
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
10^1
Maternity leave
Unemployed
Working
nmercial associate
State servant
Pensioner
Businessman
Student

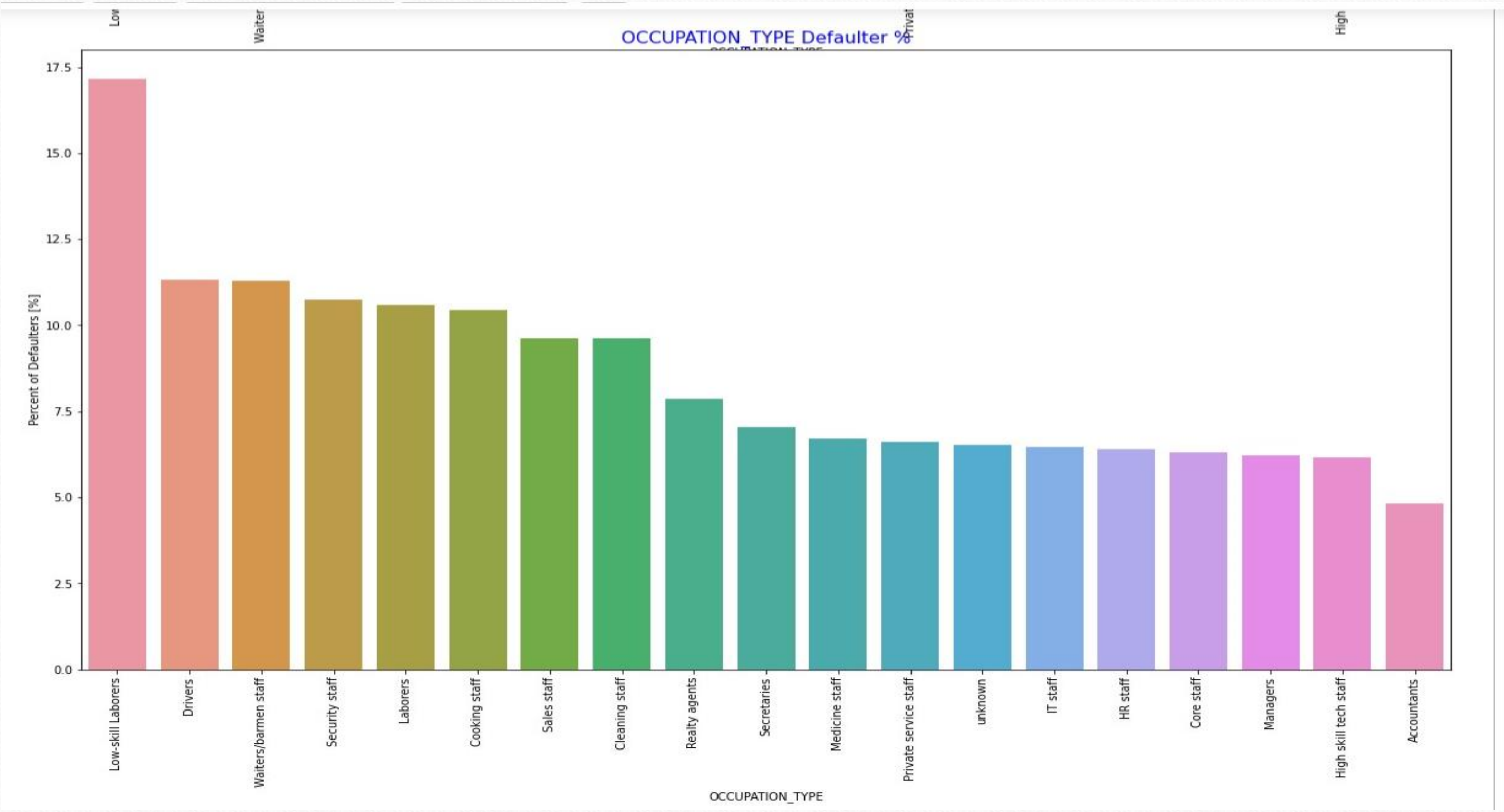

NAME_INCOME_TYPE Defaulter %
Percent of Defaulters [%]
0
5
10
15
20
25
30
35
40
Maternity leave
Unemployed
Working
Commercial associate
State servant
Pensioner
Businessman
Student

OCCUPATION_TYPE -

- Most Clients haven't mentioned their Occupation Type.
- Low Skill Laborers are the highest Defaulters (even though they are rare clients) with >17%, Drivers and Waiters/barmen, Security staff, Laborers, each with >10%.
- Accountants are the least Defaulter Clients with <5%.

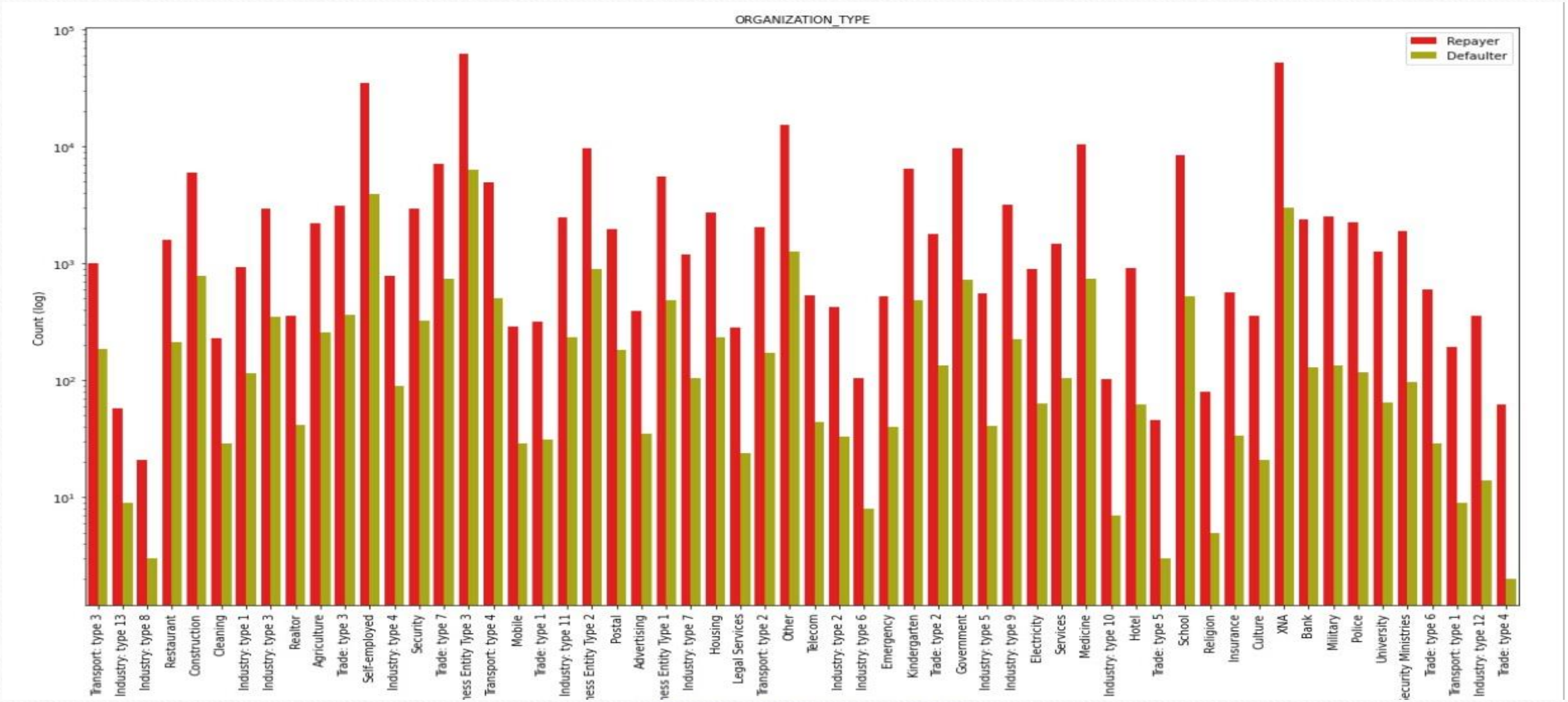


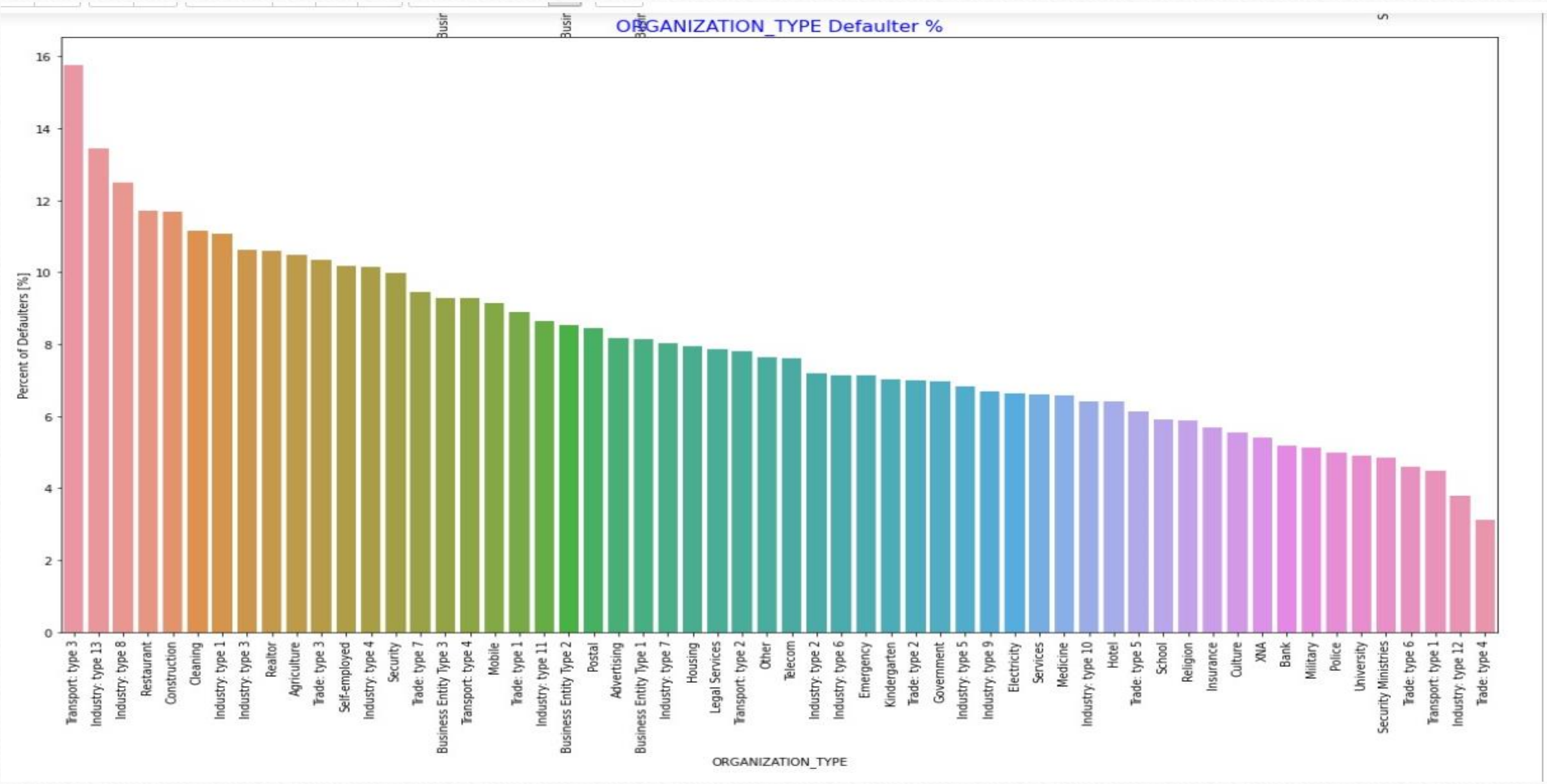
OCCUPATION_TYPE
Repayer
Defaulter
count
0
20000
40000
60000
80000
Low-skill Laborers
Drivers
iters/barmen staff
Security staff
Laborers
Cooking staff
Sales staff
Cleaning staff
Realty agents
Secretaries
Medicine staff
ivate service staff
unknown
IT staff
HR staff
Core staff
Managers
ligh skill tech staff
Accountants


Low
Waiter
Privat
High
OCCUPATION_TYPE Defaulter %
Percent of Defaulters [%]
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Low-skill Laborers
Drivers
Waiters/barmen staff
Security staff
Laborers
Cooking staff
Sales staff
Cleaning staff
Realty agents
Secretaries
Medicine staff
Private service staff
unknown
IT staff
HR staff
Core staff
Managers
High skill tech staff
Accountants
OCCUPATION_TYPE

ORGANIZATION_TYPE -

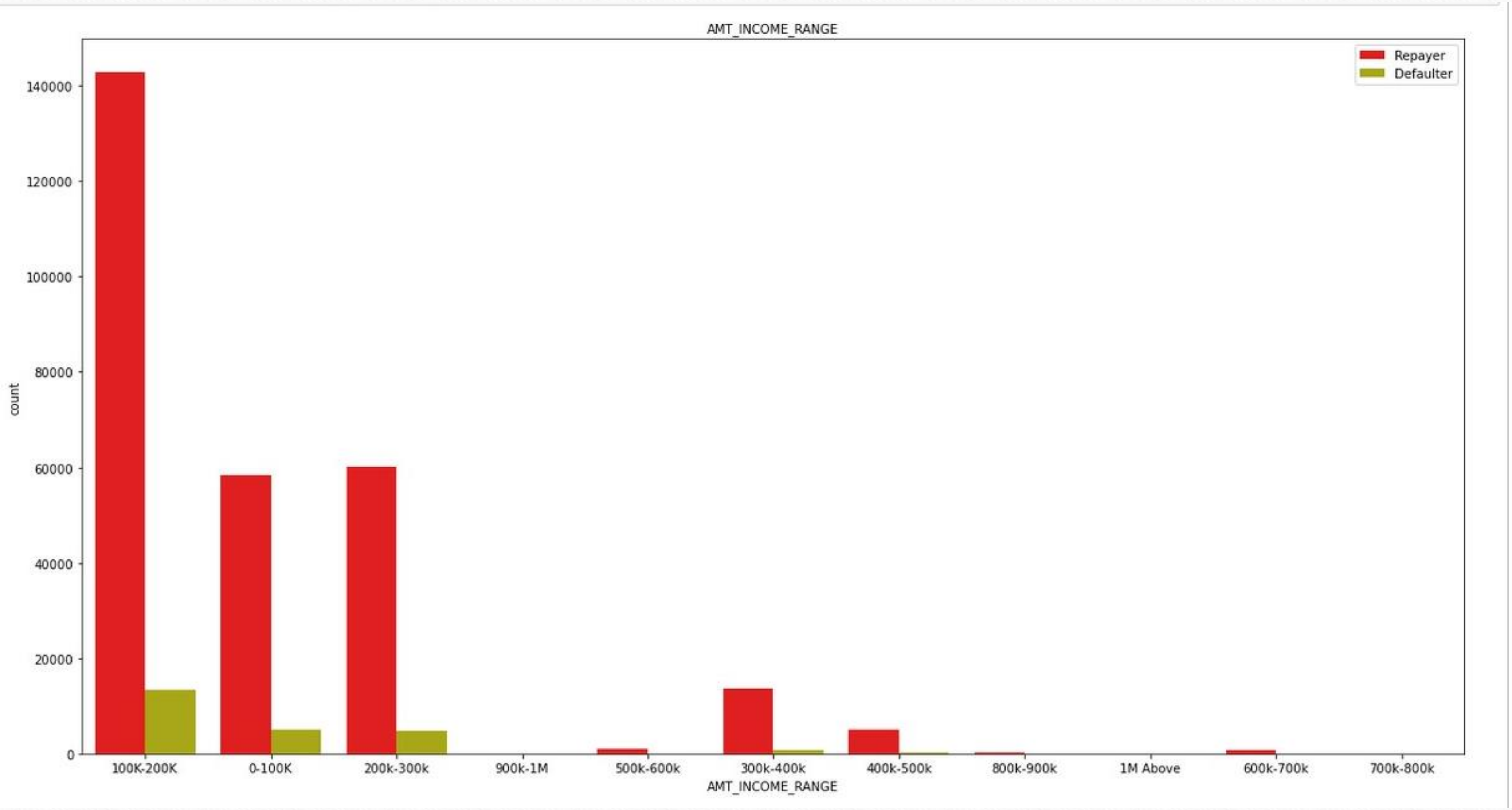
- Organizations with highest percent of loans not repaid are Transport: type 3 (almost 16%), Industry: type 13 (about 13.5%), Industry: type 8 (about 12.5%), Restaurant and Construction (almost 12% each).
- Self employed people have relative high Default%.
- Most clients are from Business Entity Type 3.
- Information is unavailable for a large number of clients.
- Trade: type 4 and Industry: type 12 are the least Defaulters

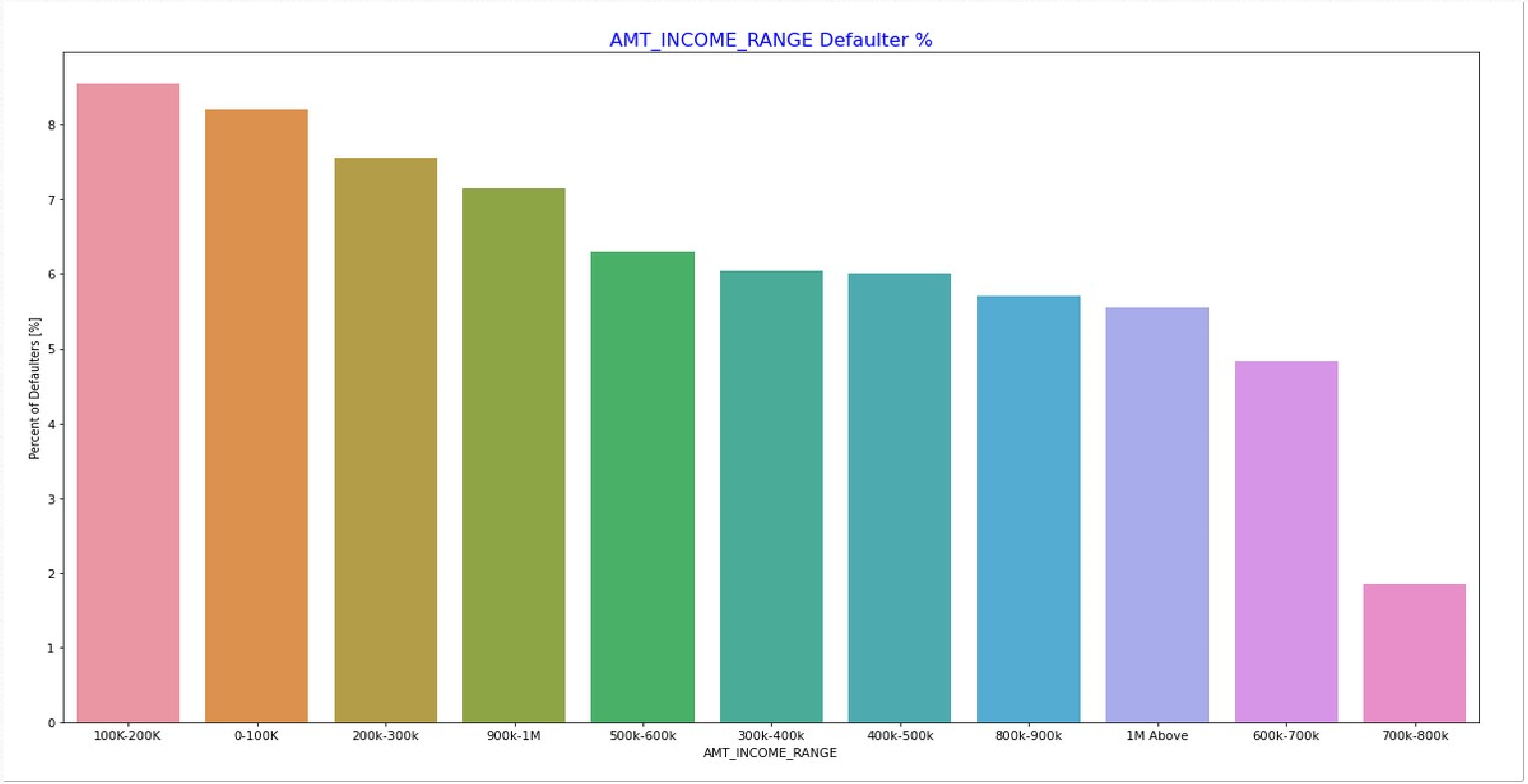

ORGANIZATION_TYPE
Repayer
Defaulter
Count (log)
10^5
10^4
10^3
10^2
10^1
Transport: type 3
Industry: type 13
Industry: type 8
Restaurant
Construction
Cleaning
Industry: type 1
Industry: type 3
Realtor
Agriculture
Trade: type 3
Self-employed
Industry: type 4
Security
Trade: type 7
ness Entity Type 3
Transport: type 4
Mobile
Trade: type 1
Industry: type 11
ness Entity Type 2
Postal
Advertising
ness Entity Type 1
Industry: type 7
Housing
Legal Services
Transport: type 2
Other
Telecom
Industry: type 2
Industry: type 6
Emergency
Kindergarten
Trade: type 2
Government
Industry: type 5
Industry: type 9
Electricity
Services
Medicine
Industry: type 10
Hotel
Trade: type 5
School
Religion
Insurance
Culture
XNA
Bank
Military
Police
University
ecurity Ministries
Trade: type 6
Transport: type 1
Industry: type 12
Trade: type 4


ORGANIZATION_TYPE Defaulter %
Percent of Defaulters [%]
ORGANIZATION_TYPE
0
2
4
6
8
10
12
14
16
Transport: type 3
Industry: type 13
Industry: type 8
Restaurant
Construction
Cleaning
Industry: type 1
Industry: type 3
Realtor
Agriculture
Trade: type 3
Self-employed
Industry: type 4
Security
Trade: type 7
Business Entity Type 3
Transport: type 4
Mobile
Trade: type 1
Industry: type 11
Business Entity Type 2
Postal
Advertising
Business Entity Type 1
Industry: type 7
Housing
Legal Services
Transport: type 2
Other
Telecom
Industry: type 2
Industry: type 6
Emergency
Kindergarten
Trade: type 2
Government
Industry: type 5
Industry: type 9
Electricity
Services
Medicine
Industry: type 10
Hotel
Trade: type 5
School
Religion
Insurance
Culture
XNA
Bank
Military
Police
University
Security Ministries
Trade: type 6
Transport: type 1
Industry: type 12
Trade: type 4

AMT_INCOME_RANGE –

- 90% of the clients have total Income < 300k.
- Clients with Income < 300k has higher Defaulting rates.
- Clients with Income > 700k are less likely to Default.


AMT_INCOME_RANGE
Repayer
Defaulter
count
0
20000
40000
60000
80000
100000
120000
140000
100K-200K
0-100K
200k-300k
900k-1M
500k-600k
300k-400k
400k-500k
800k-900k
1M Above
600k-700k
700k-800k
AMT_INCOME_RANGE


AMT_INCOME_RANGE Defaulter %
Percent of Defaulters [%]
0
1
2
3
4
5
6
7
8
100K-200K
0-100K
200k-300k
900k-1M
500k-600k
300k-400k
400k-500k
800k-900k
1M Above
600k-700k
700k-800k
AMT_INCOME_RANGE

Bivariate Analysis

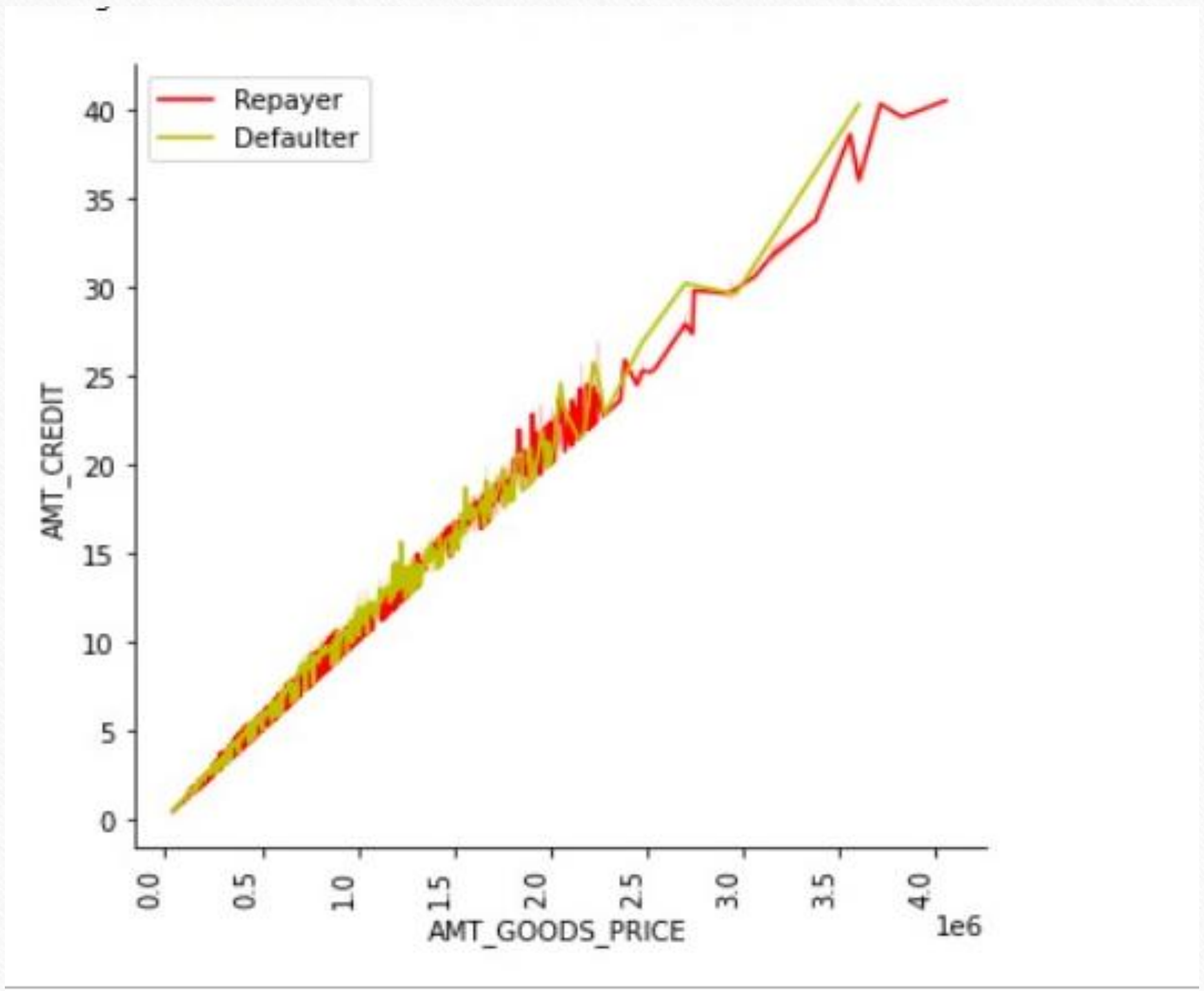
- Top 10 correlations are found for relevant columns and the correlating factors are identified for the Re-payers and Defaulters with the help of heat map and pair plot.

- For Re-payers, it is observed that:
  - ■ Credit amount is highly correlated with amount of goods price, loan annuity & total income
  - ■ Re-payers have high correlation in number of days employed.

- For Defaulters, it is observed that:
  - ■ Credit amount is highly correlated with amount of goods price which is similar as for Re-payers.
  - ■ The loan annuity correlation with credit amount and correlation for Employment days have slightly reduced for Defaulters.

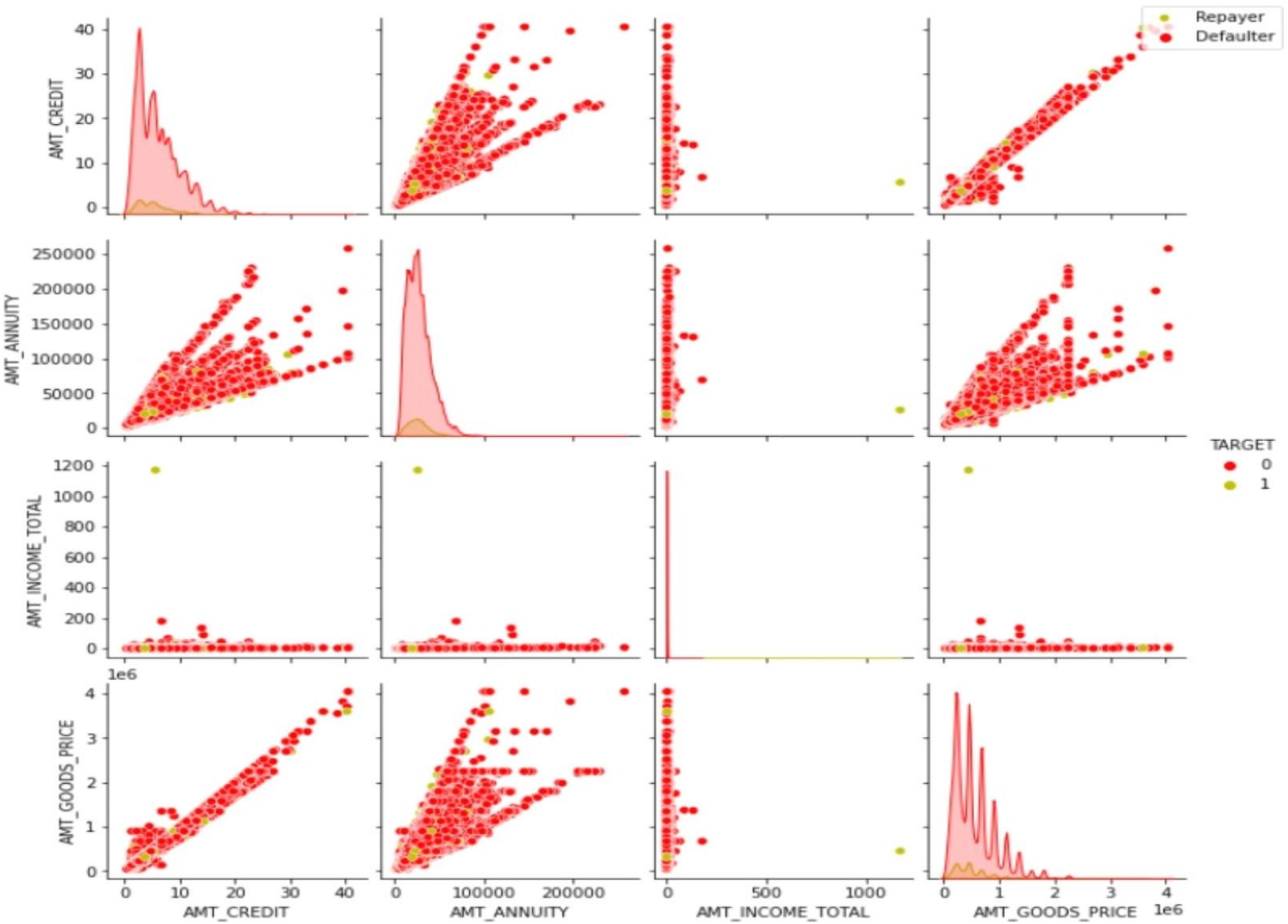
Relational plot for Credit Amount and Good Price –

- It is observed that when the credit amount exceeds 3 millions for amount goods price, there is an increase in Defaulters.



Pair plot between Amount variables –

- When AMT_ANNUITY >15000 AMT_GOODS_PRICE> 3M, there is a lesser chance of defaulters
- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line
- There are very less defaulters for AMT_CREDIT >3M
- Inferences related to distribution plot has been already mentioned in previous distplot graphs inferences section


Repayer
Defaulter
TARGET
0
1
AMT_CREDIT
AMT_ANNUITY
AMT_INCOME_TOTAL
AMT_GOODS_PRICE
0
10
20
30
40
0
50000
100000
150000
200000
250000
0
200
400
600
800
1000
1200
1e6
0
1
2
3
4
0
100000
200000
0
500
1000
1e6

Merged Data frame Analysis –

- Both ‘application_df’ and ‘previous_df’ are merged based on the current application id and common records

are analyzed.

- Re-payer and Defaulter records are segregated into two separate data frames and impact of the Decisions taken for previous loans are analyzed for relevant records.

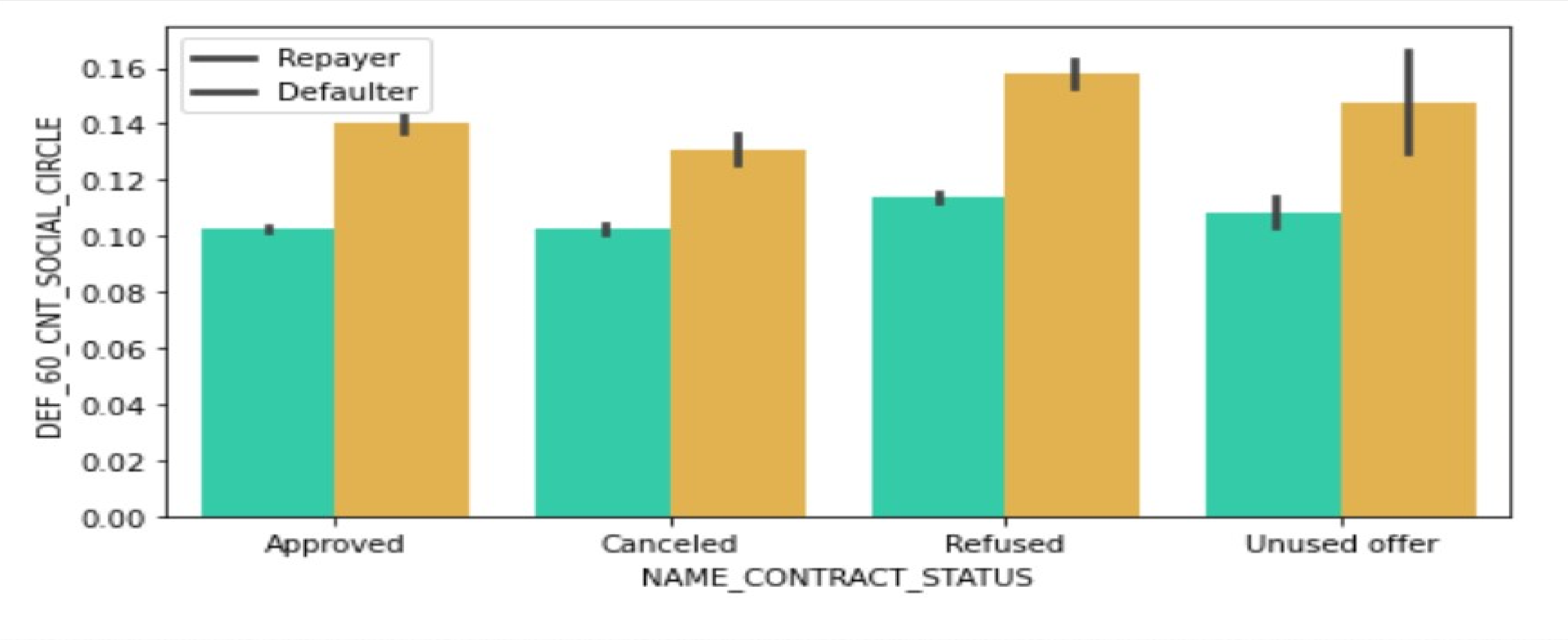
- Observations for decisions taken based on loan purpose:
  - ■ Loan purpose has high number of unknown values (XAP, XNA).
  - ■ Loans taken for the purpose of Repairs seems to have highest Default rate.
  - ■ A very high number of applications for the purpose of 'Repairs' and 'Others' has been Refused by Company or Canceled by Client.

Observations for decisions taken to identify business or financial loss –

- About 90% of the loans have been repaid for cases where the Client Canceled their application.
- 88% of the Clients who have been previously Refused a loan by the Company, have repaid the Current loan.

Relationship between total income and contract status –

- It is observed that the Clients with Unused offer earlier have Defaulted even when their average income is higher than others.



Conclusion

The following factors indicate that an applicant will be a Re-payer :

○ Academic degree clients have less Defaults.
○ Students and Businessmen have no Defaults.
○ Clients with Trade: type 4 and Industry: type 12 have Defaulting rate of less than 4%.
○ Clients above the age of 50 years have low probability of Defaulting.
○ Clients with 40+ years of employment have less than 1% Default Rate.
○ Clients with Income more than 700,000 are less likely to Default
○ Loans applied for Hobby, Buying garage are being repaid in most cases.
○ Clients with zero to two children tend to repay the loans.

The following factors indicate that an applicant will be a Defaulter:

- Male Clients have relatively higher Defaulting rate.
- Clients who have Civil marriage or are Single have higher Default rate.
- Clients with education level of Lower Secondary & Secondary/secondary special have high Default rate.
- Clients who are either on Maternity leave or are Unemployed have high Default rate.
- Low skill Laborers, Drivers and Waiters/barmen, Security staff, Laborers and Cooking staff should be avoided as Clients as their Defaulting rate is high.
- Organizations with highest percent of loans not repaid are Transport: type 3 (almost 16%), Industry: type 13 (about 13.5%), Industry: type 8 (about 12.5%), Restaurant and Construction (almost 12% each). Self employed people have relative high Default%, and thus should be avoided while approving for loan or charged higher interest rate to mitigate the risk of Defaulting.
- Clients in the age group of 20-40 years should be avoided as they have higher probability of Defaulting.
- Clients with less than 5 years of employment have high Default rate.
- Clients with more than 8 children have a Default rate of 100% and hence their applications should not be approved.
- When the Credit amount goes beyond 3 million, there is an increase in Defaulters.

Insights From Previous Applications:

○ About 90% of the loans have been repaid for cases where the Client Canceled their application previously. Thus, recording the reason for cancelation can help the Company to determine and negotiate terms with these repaying Customers in future for increasing their business opportunity.

○ 88% of the Clients who have been previously Refused a loan by the Company, have now turned into a repaying Client. Hence, documenting the reason for rejection can mitigate the business loss and these clients could be contacted for future loans.