“ CREDIT EDA CASE STUDY.

— Kamal Thampi

Jan 16, 2024

PROBLEM STATEMENT

The problem is to identify the factors that influence the loan default risk of urban customers who apply for various types of loans from a consumer finance company.

There are two types of risks associated with the bank's decision for loan approval,

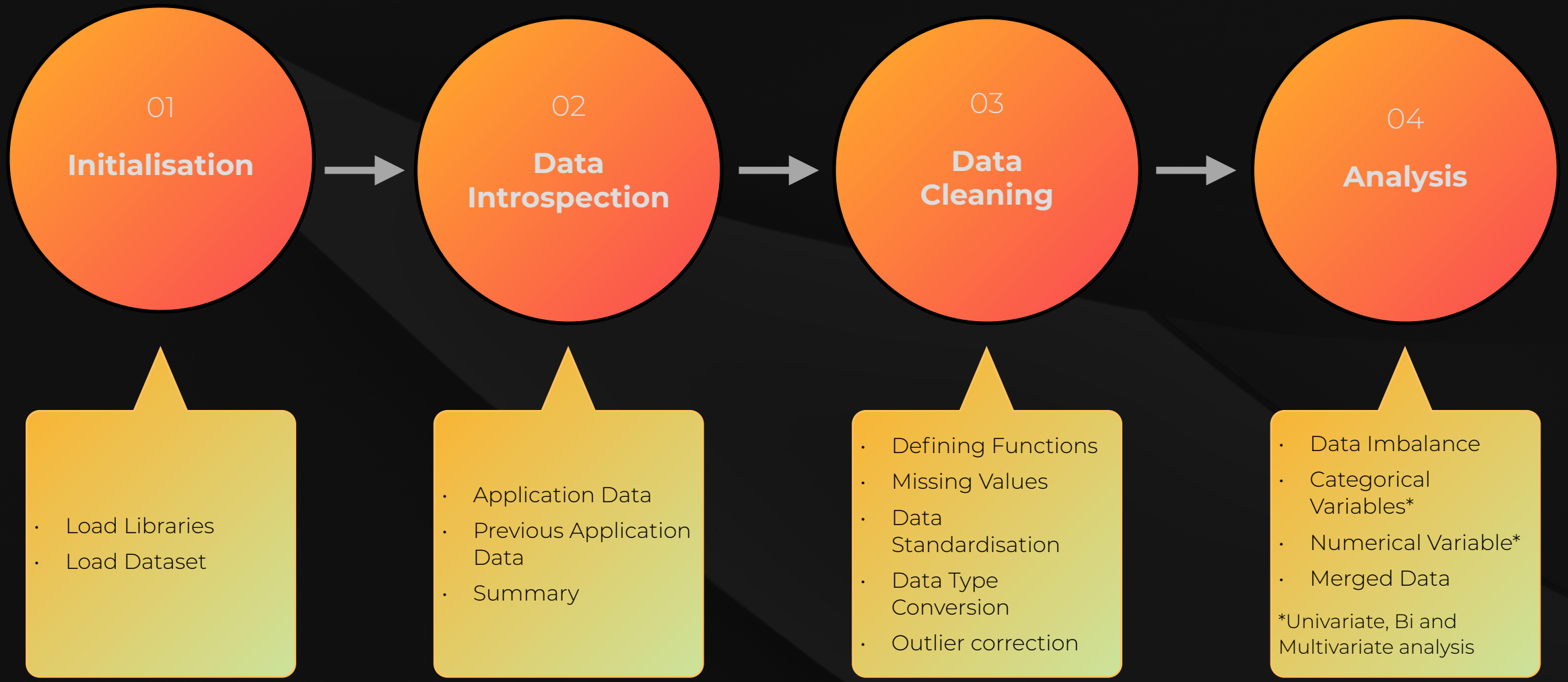
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

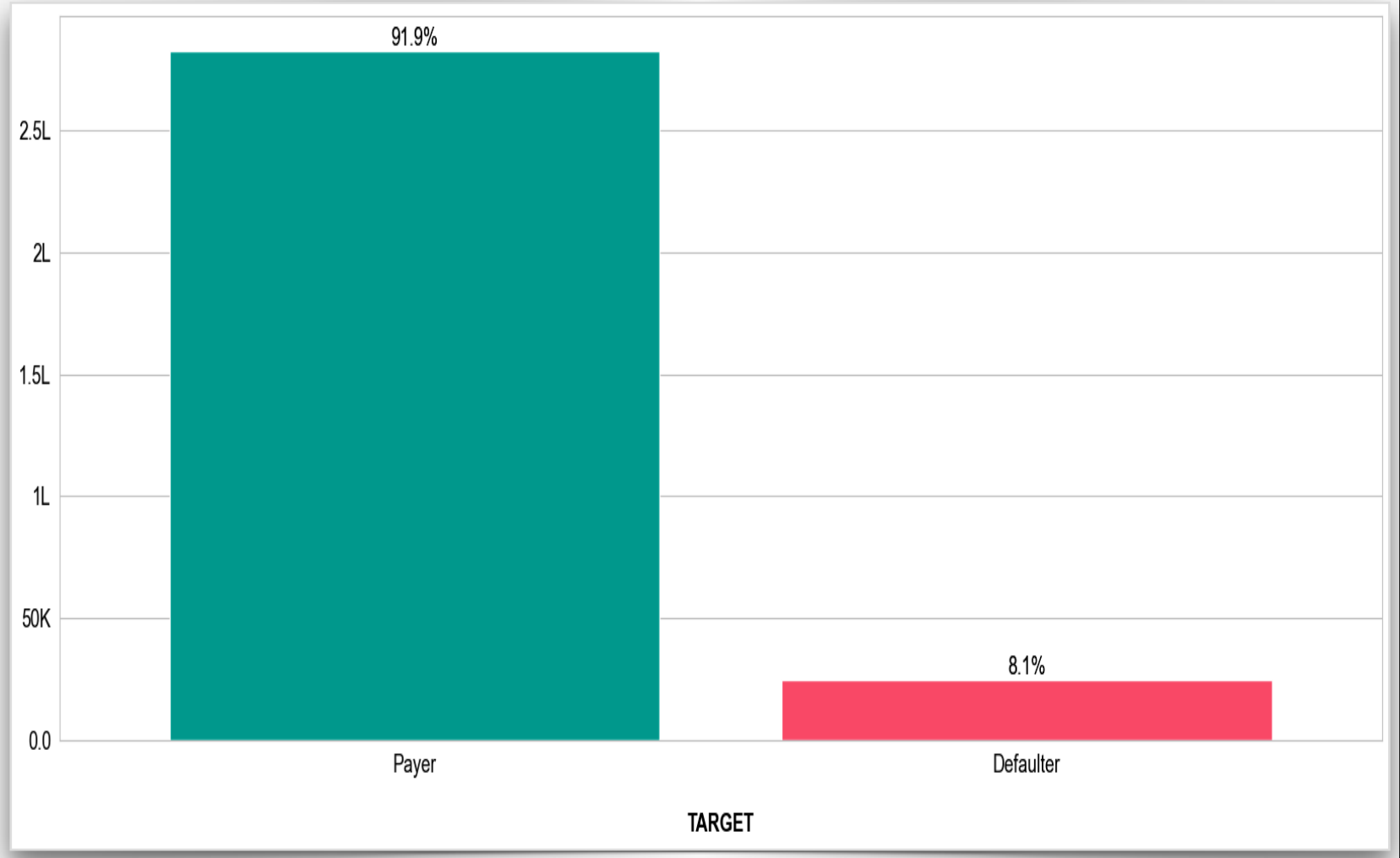
These risks can affect the profitability and reputation of the bank, as well as the credit score and financial situation of the applicant. Therefore, it is important for the bank to assess the credit risk of each applicant carefully and make informed decisions based on the data and the bank's risk policy.

Business Objective

The objective is to use EDA to find patterns in the data that indicate the likelihood of payment difficulties for different loan applicants, and to use this information to improve the loan approval process and minimise the financial losses for the company.

APPROACH


01
Initialisation
02
Data Introspection
03
Data Cleaning
04
Analysis
Load Libraries
Load Dataset
Application Data
Previous Application Data
Summary
Defining Functions
Missing Values
Data Standardisation
Data Type Conversion
Outlier correction
Data Imbalance
Categorical Variables*
Numerical Variable*
Merged Data
*Univariate, Bi and Multivariate analysis

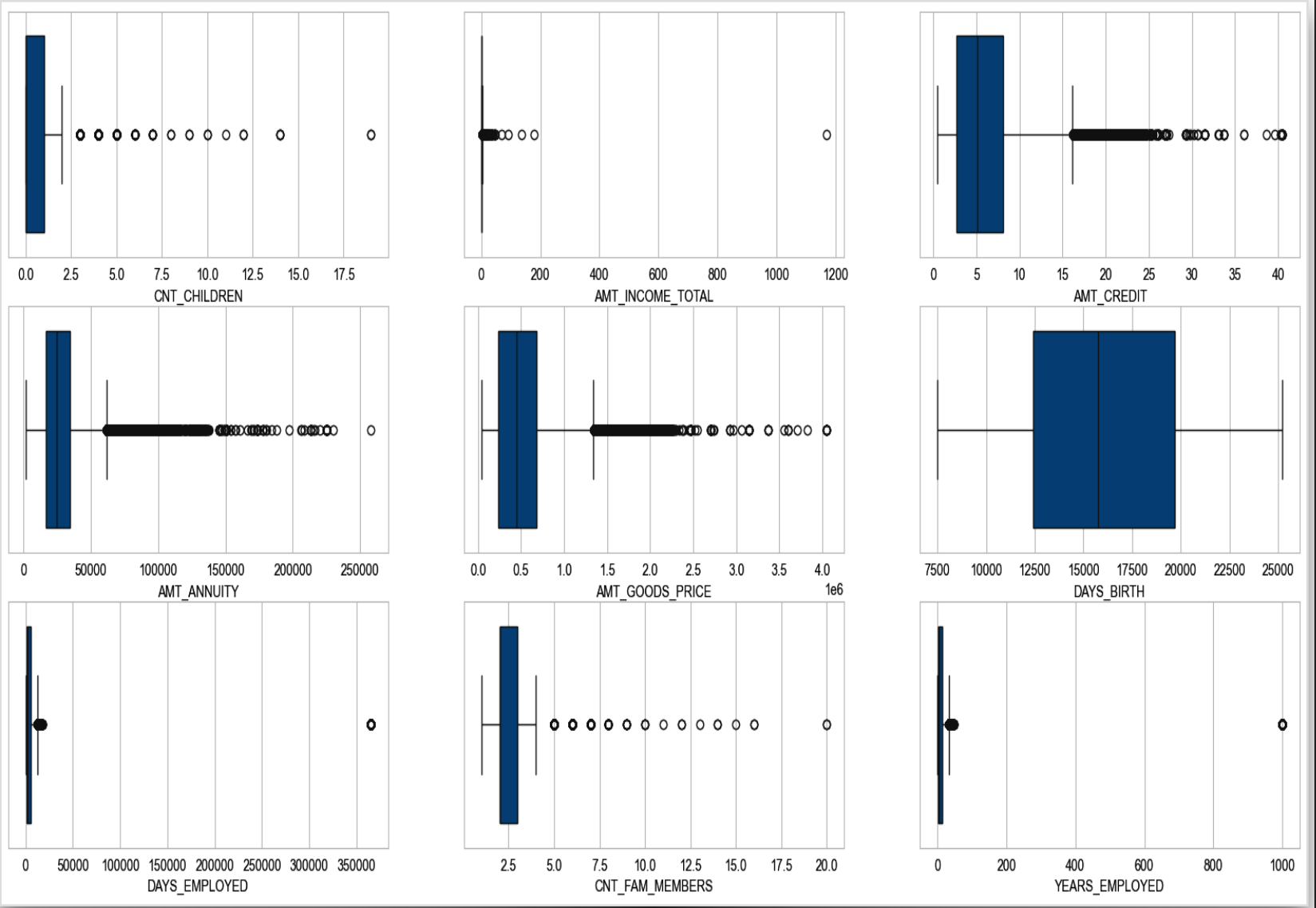


ANALYSIS

DATA IMBALANCE

TARGET VARIABLE

- The ratio of Defaulter is much smaller than the Payer, which means that we have an imbalanced dataset.
- Imbalanced datasets can cause problems for many machine learning algorithms, as they tend to overfit the majority class and underfit the minority class. This can lead to poor generalization and low predictive performance on the defaulters.



ANALYSIS

IDENTIFYING OUTLIERS

APPLICATION DATA

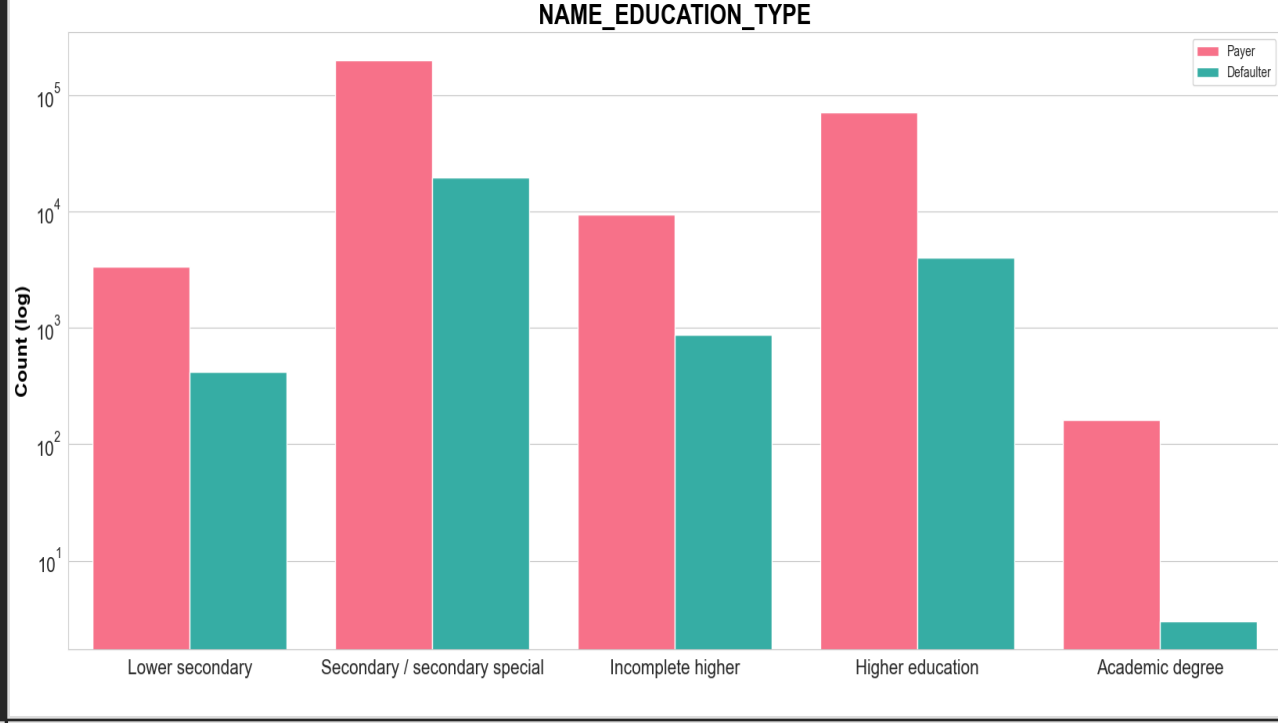
- The data has some incorrect entries in DAYS_EMPLOYED and YEARS_EMPLOYED, which show outlier values of around 350000 days (1000 years).
- AMT_INCOME_TOTAL has a large number of outliers, indicating that some loan applicants have much higher income than others.
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, CNT_CHILDREN and CNT_FAM_MEMBERS also have some outliers, but they are less extreme.
- DAYS_BIRTH has no outliers, suggesting that the data is reliable for this variable.

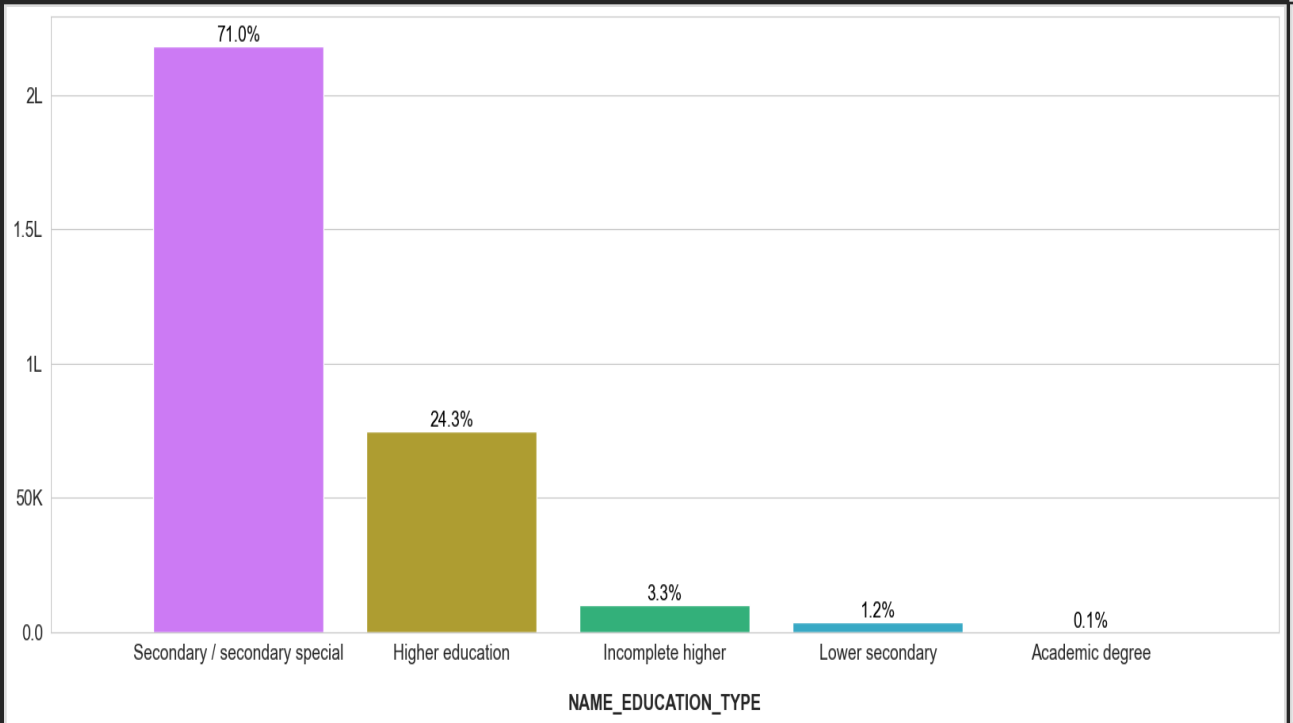
UNIVARIATE ANALYSIS

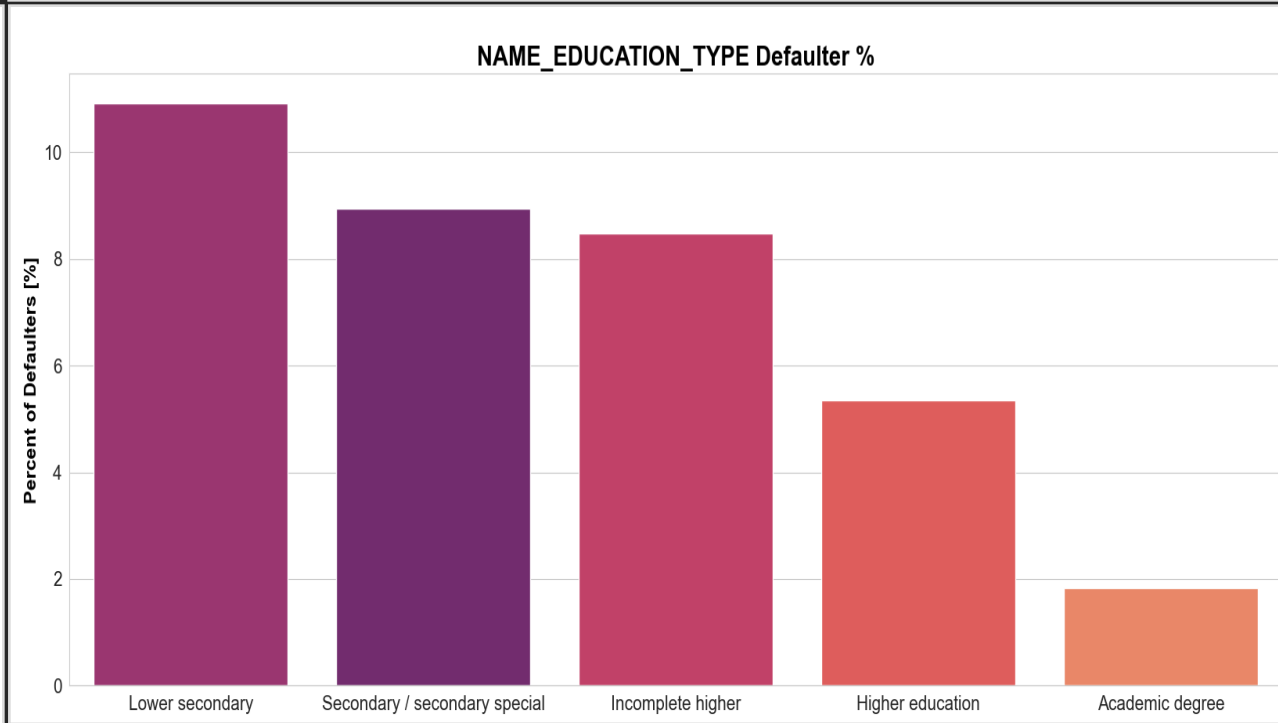
ANALYSIS

EDUCATION

- Applicants with lower secondary or secondary education have the highest defaulting rate of approx. 11% and 9% respectively.
- Applicants with an academic degree have the lowest defaulting rate of less than 2% and the least loan amount of 0.1%.
- Applicants with higher education have a moderate defaulting rate of about 24.3% and a significant loan amount of 71%.



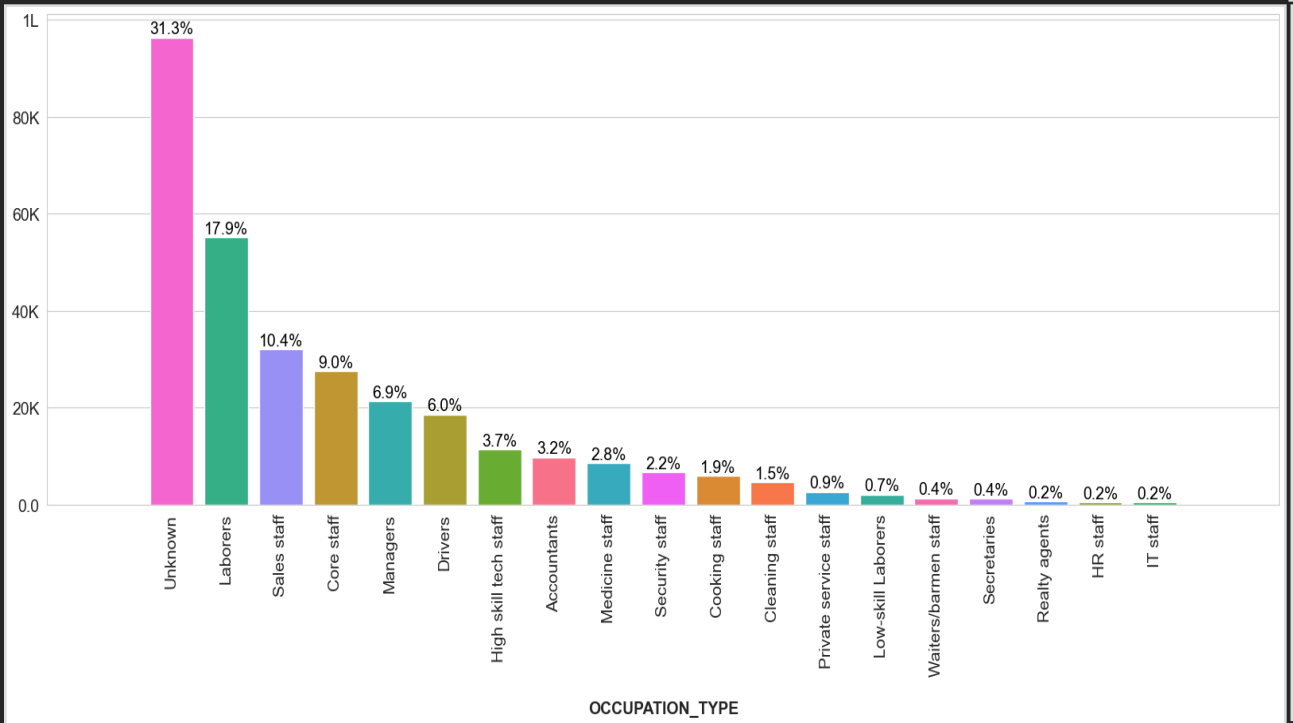


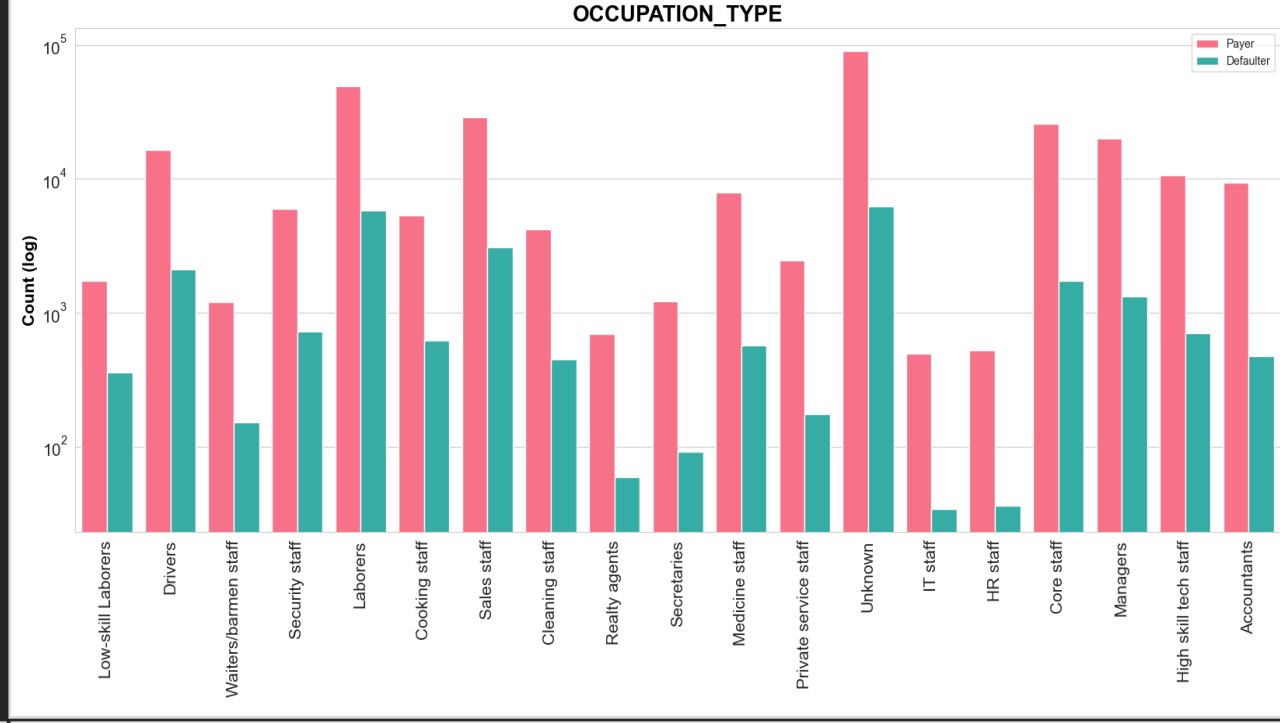


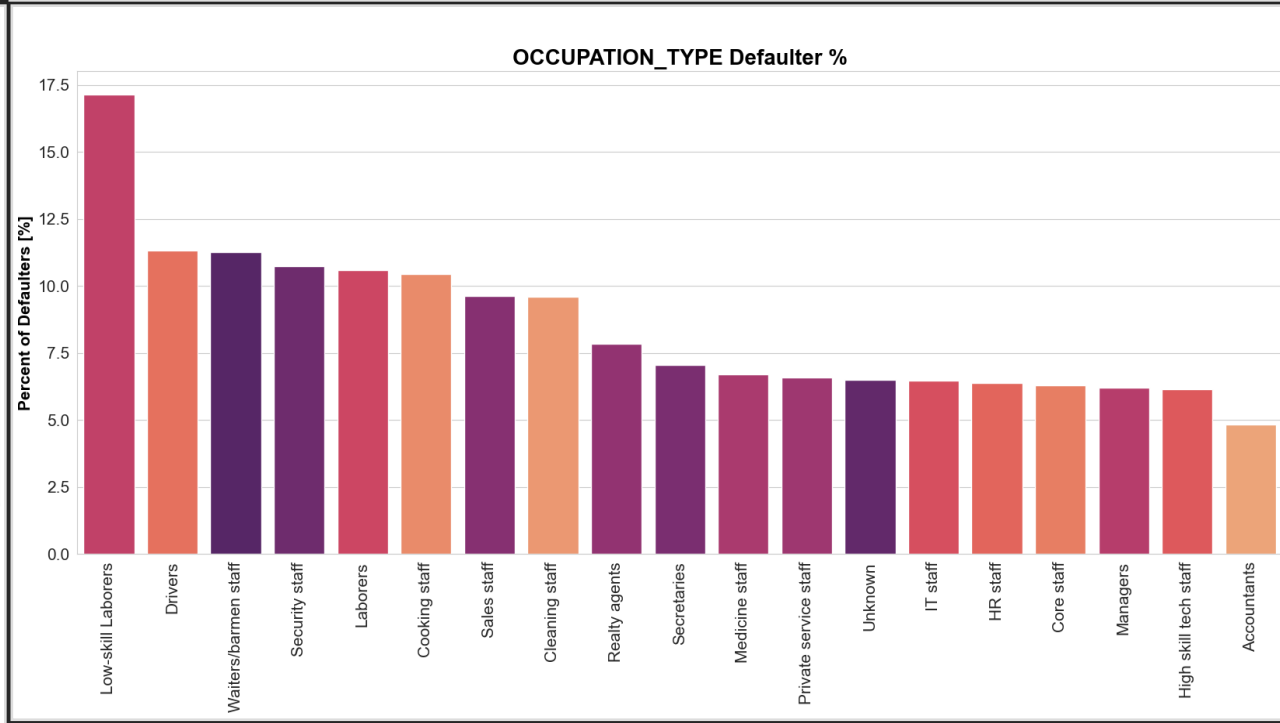
ANALYSIS

OCCUPATION TYPE

- Unknown occupation types account for 31% of the loan applicants, followed by laborers, sales and core staff.
- Low skilled laborers, drivers, waiters/barmen, security staff and cooking staff have the highest loan default rates, ranging from 17% to 10%.



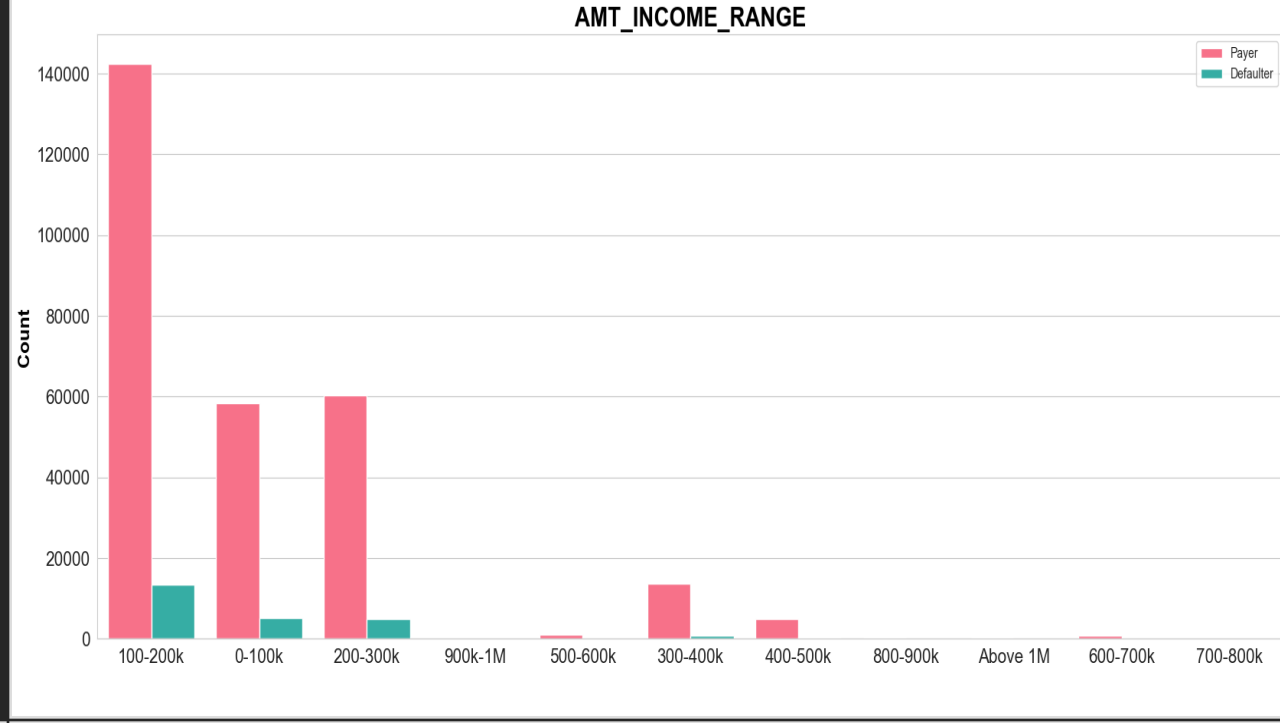


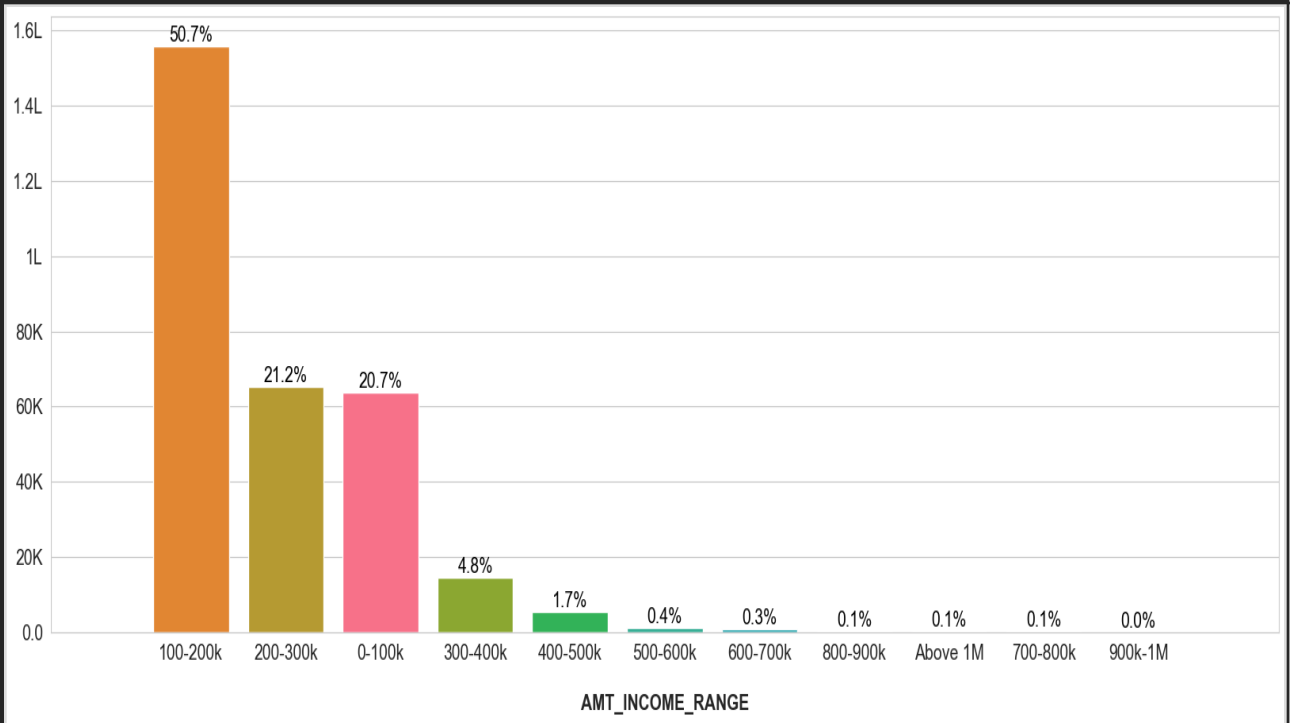


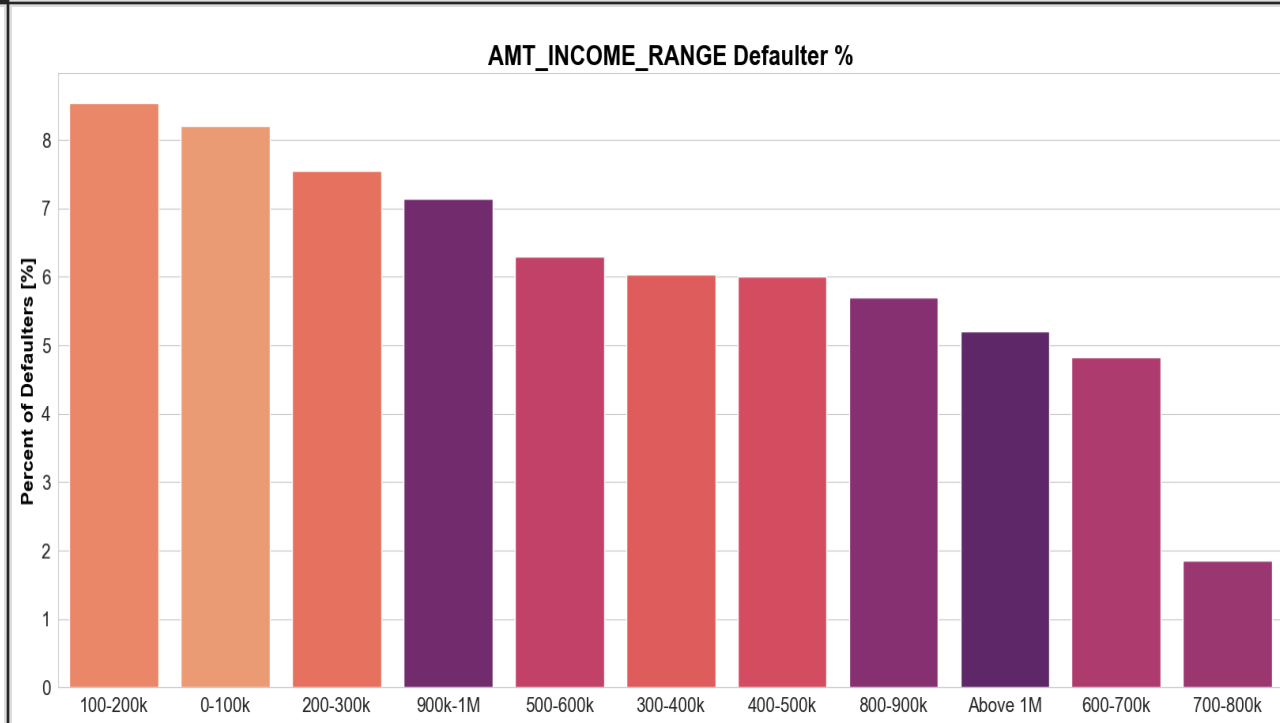
ANALYSIS

INCOME

- The default probability of loan applicants varies according to their income range.
- The highest default risk is among applicants with income between 0 and 300k.
- The lowest default risk is among applicants with income between 700 and 800k.
- About 50% of the applicants have income between 100 and 200k.



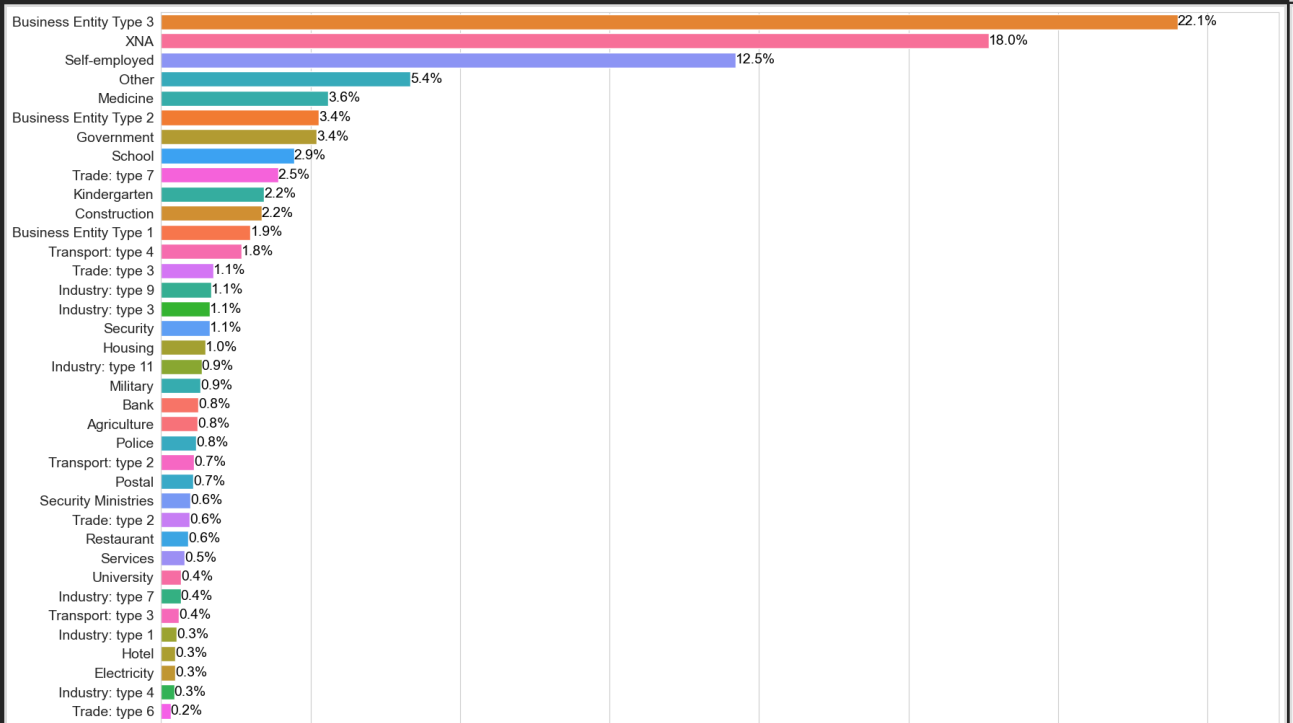


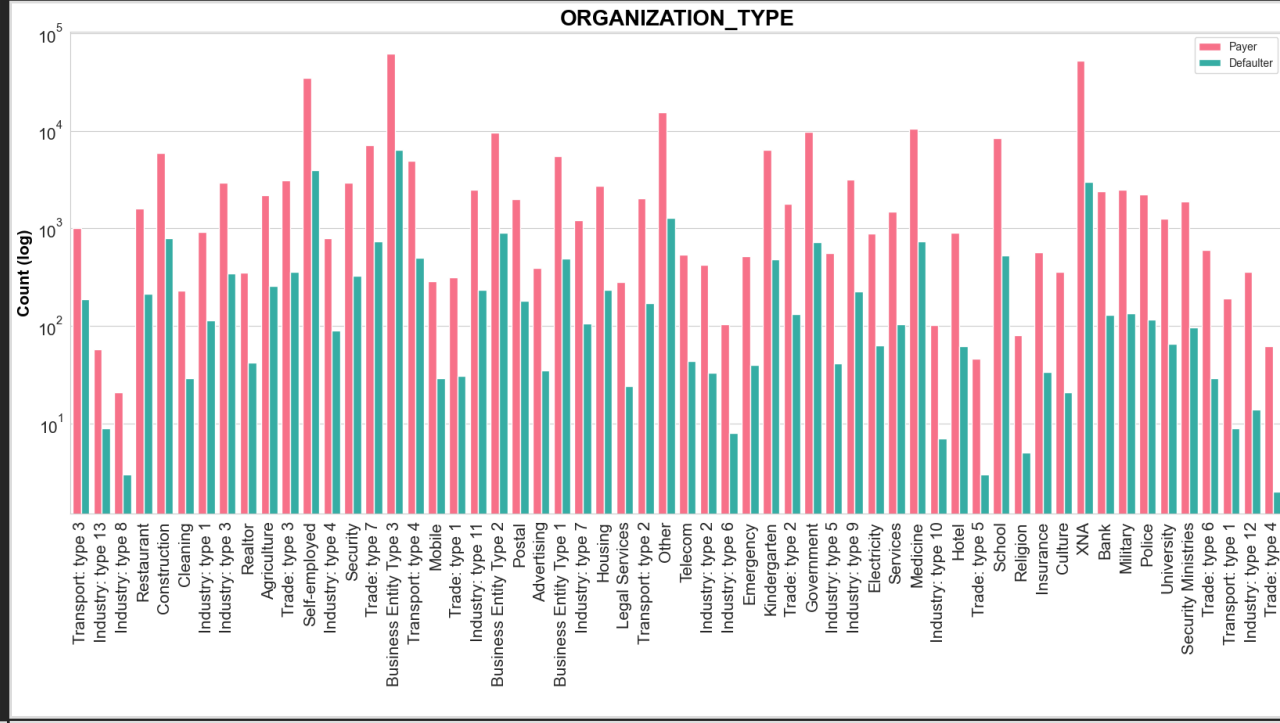


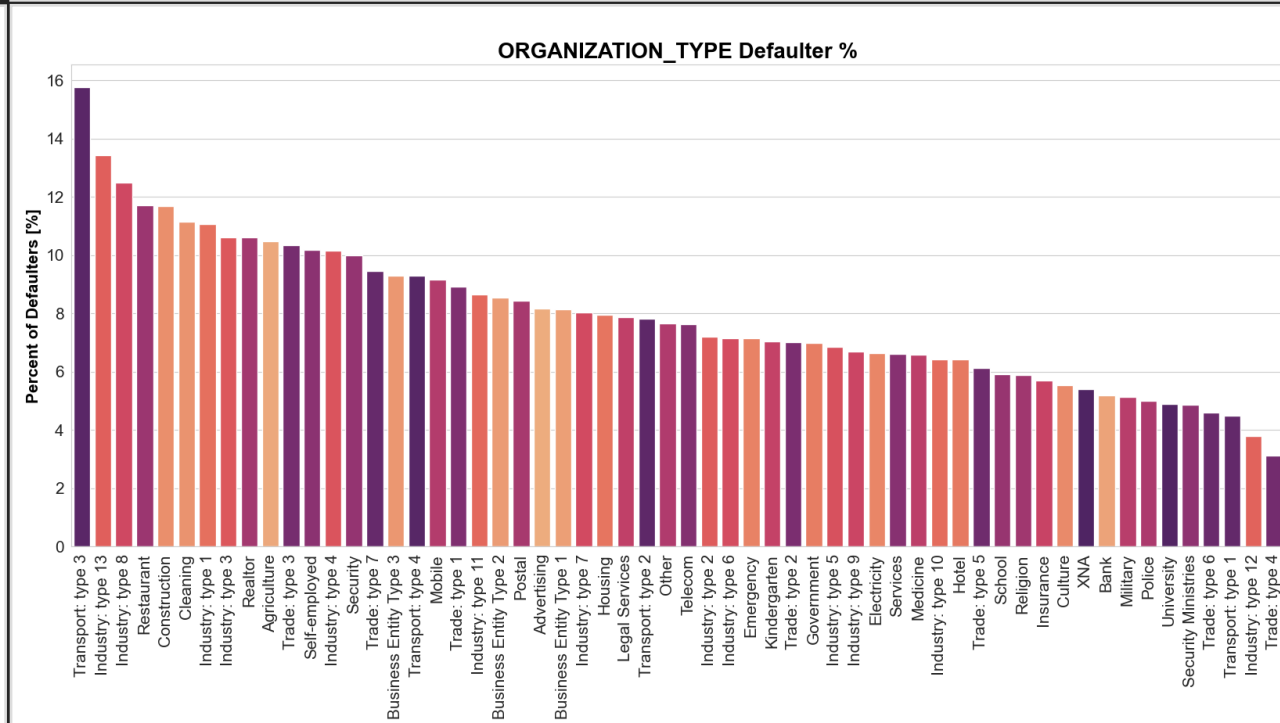
ANALYSIS

ORGANIZATION TYPE

- 18% of the loan applications have missing information (XNA).
- The most frequent organization type among the loan applicants is Business Entity Type 3.
- The organization types with the highest default rates are Transport Type 3 (16%), Industry: Type 13 (13%), Industry: Type 8 (12%), Restaurant (<12%)
- The organization type with the lowest default rate is Trade: Type 4 (<4%), followed by Industry: Type 12 (<4%).
- The Self-employed borrowers have a relatively high default rate (>10%).



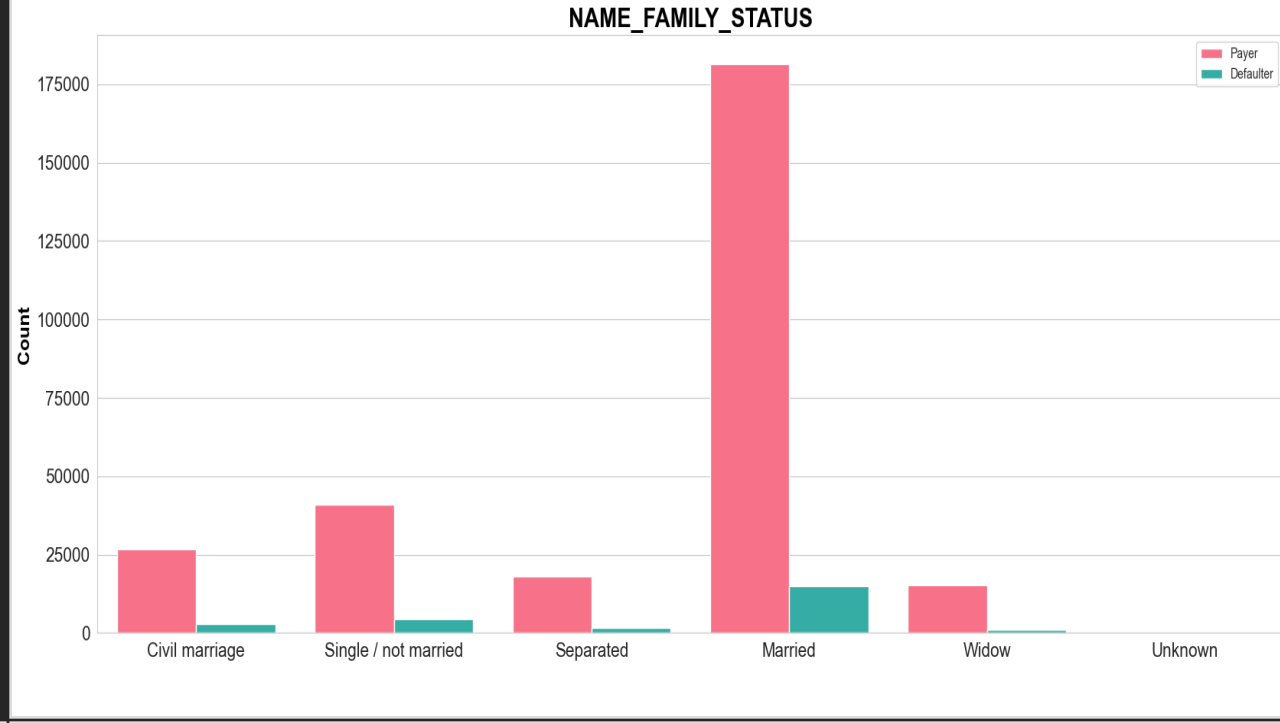


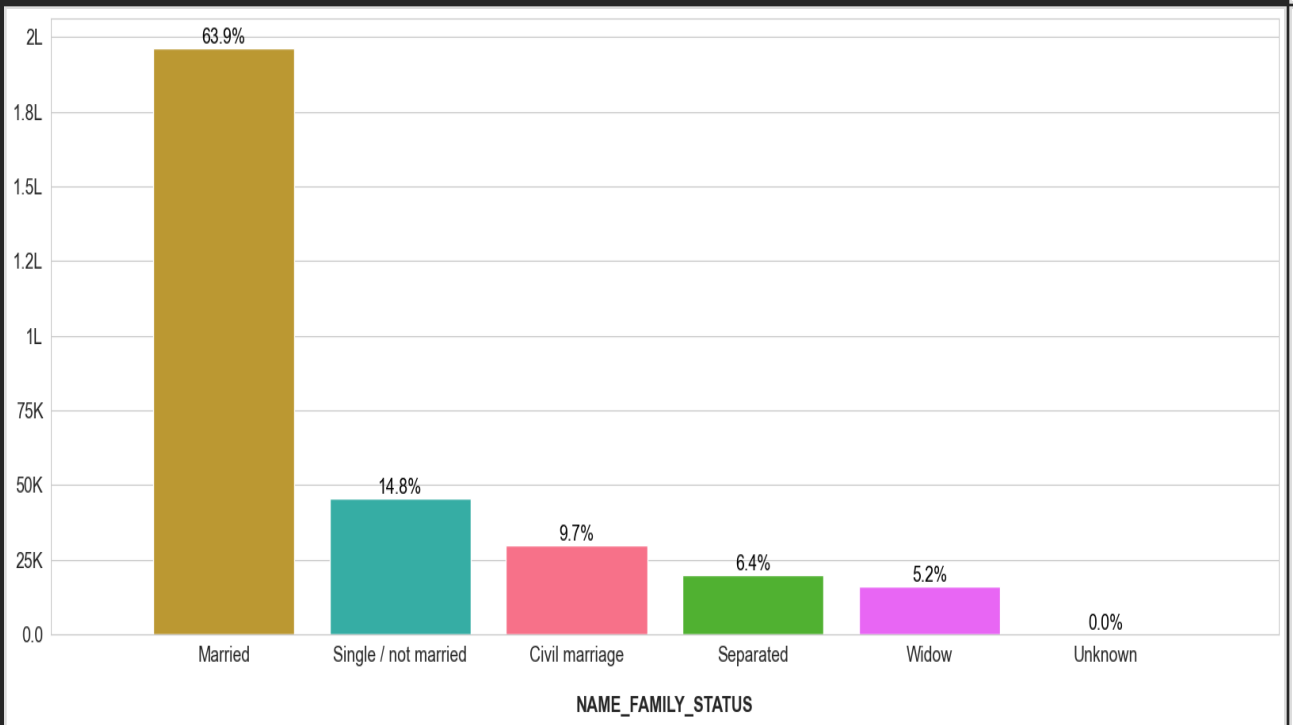


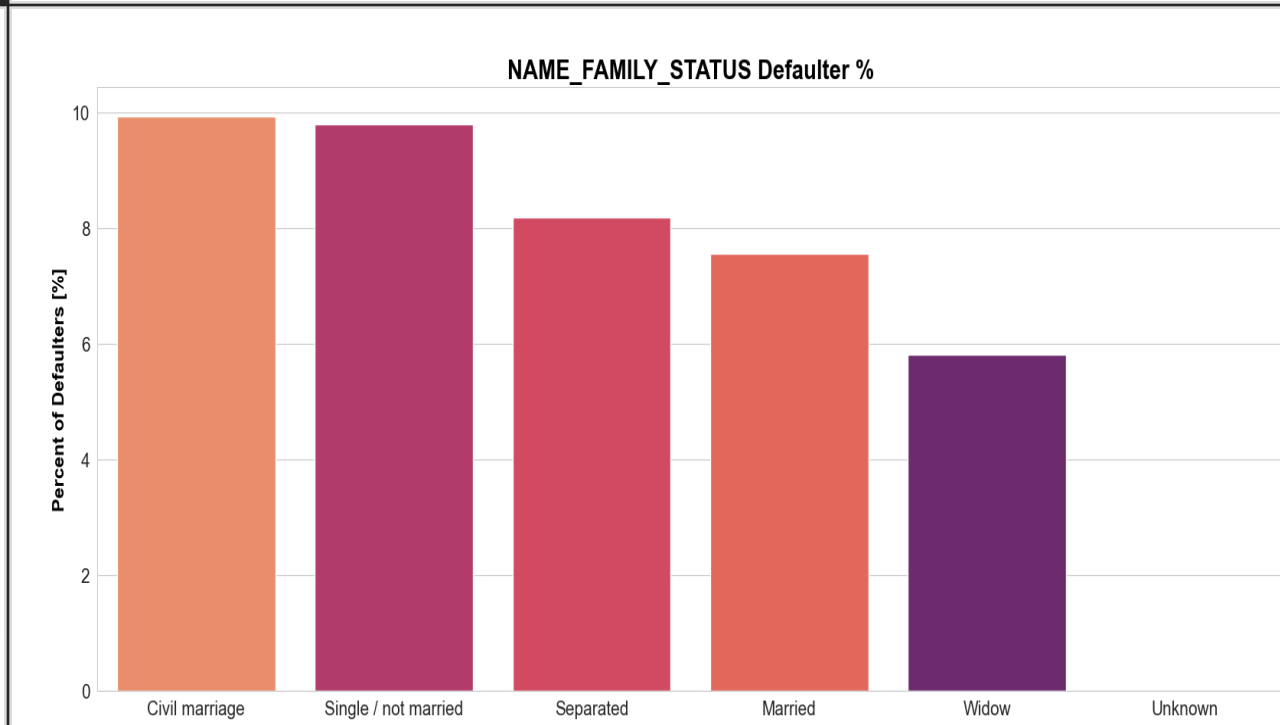
ANALYSIS

MARITAL STATUS

- The majority of loan applicants are married (64%), followed by single (15%) and civil marriage (10%).
- Civil marriage applicants have the highest loan default rate (10%), followed by single applicants.
- Widowers have the lowest loan default rate, except for unknowns.



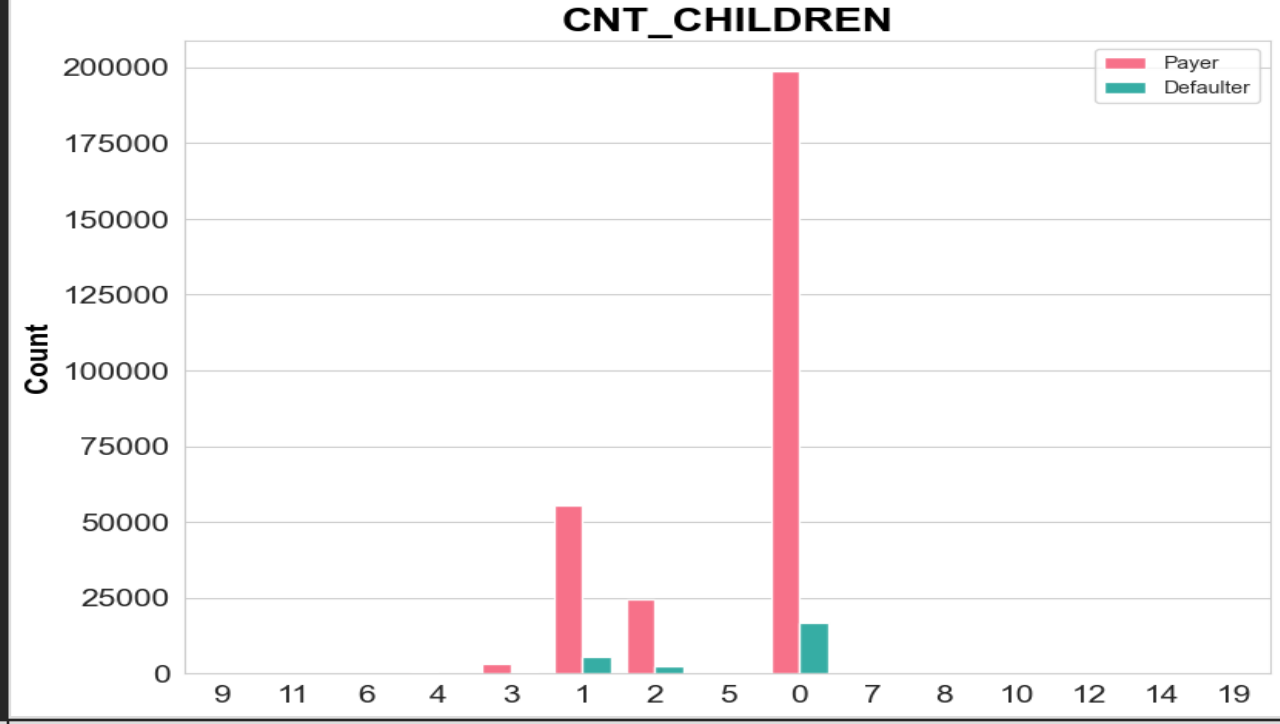


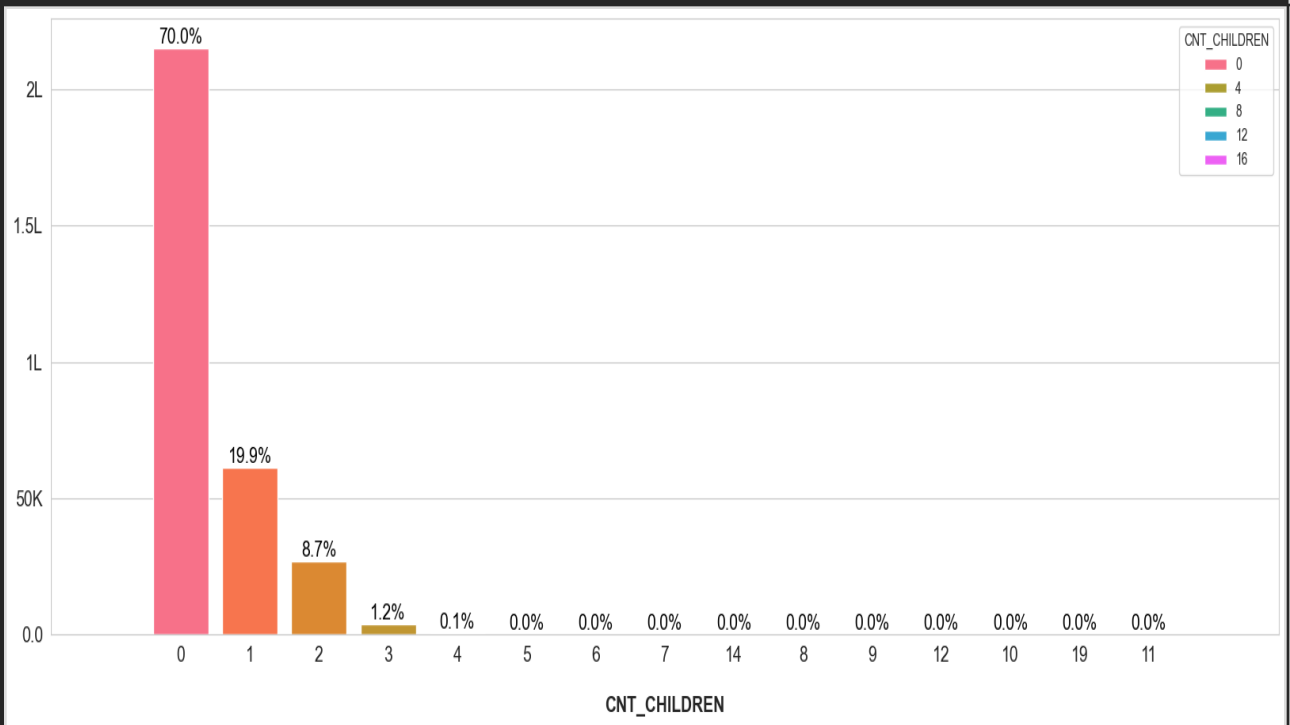


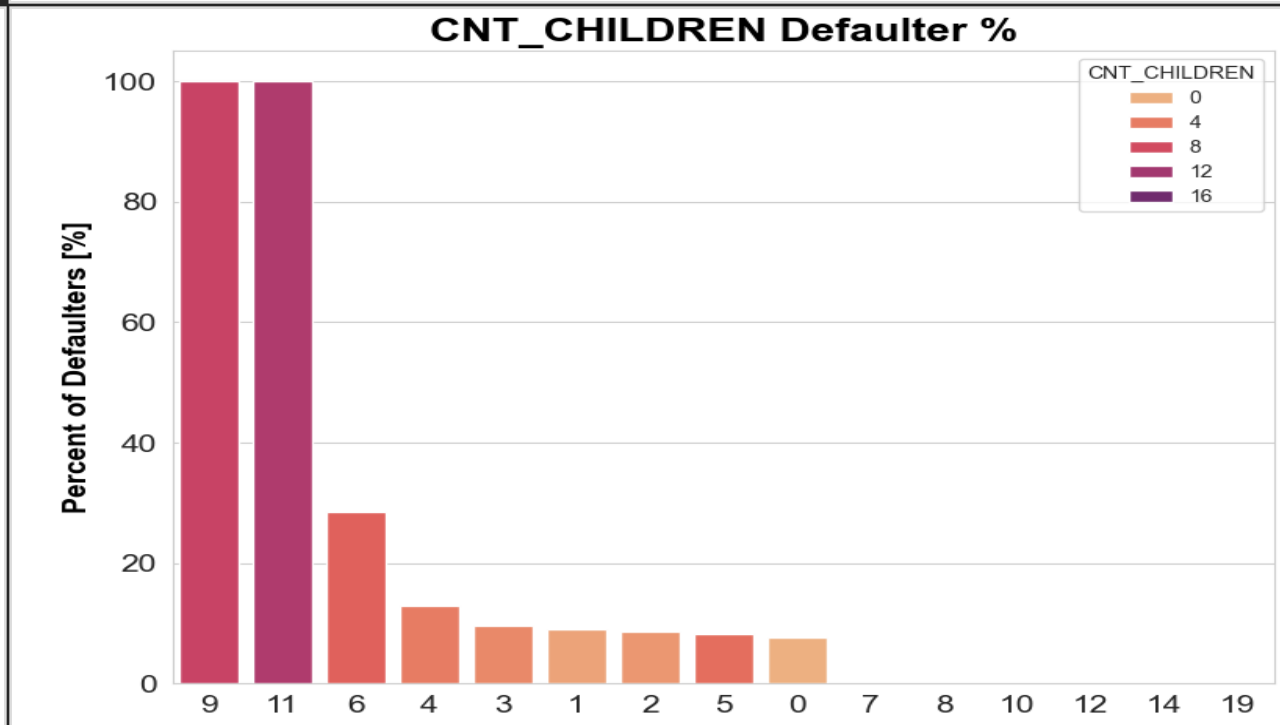
ANALYSIS

CHILDREN COUNT

- The default rate of loan applicants is influenced by their number of children.
- Applicants with more than 4 children have a very high default rate, especially those with 9 or 11 children who have a 100% default rate.
- Applicants with no children make up 70% of the total applicants.
- Applicants with more than 3 children are very rare.



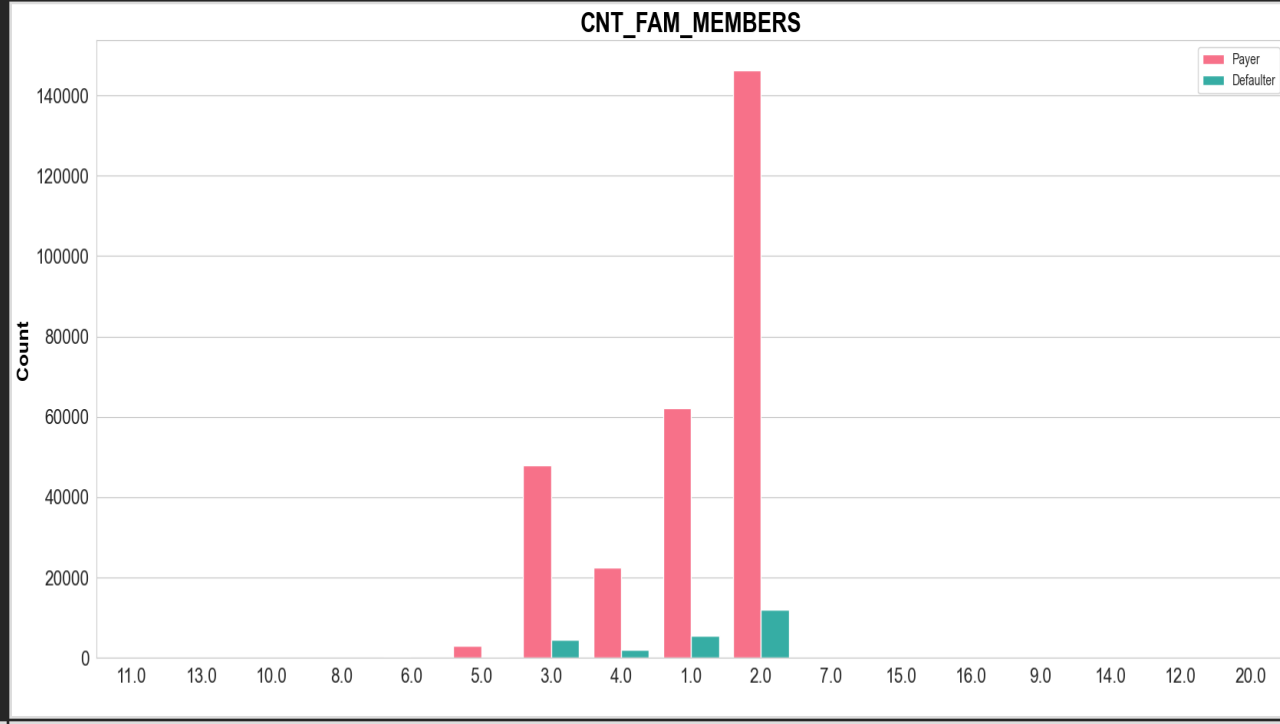


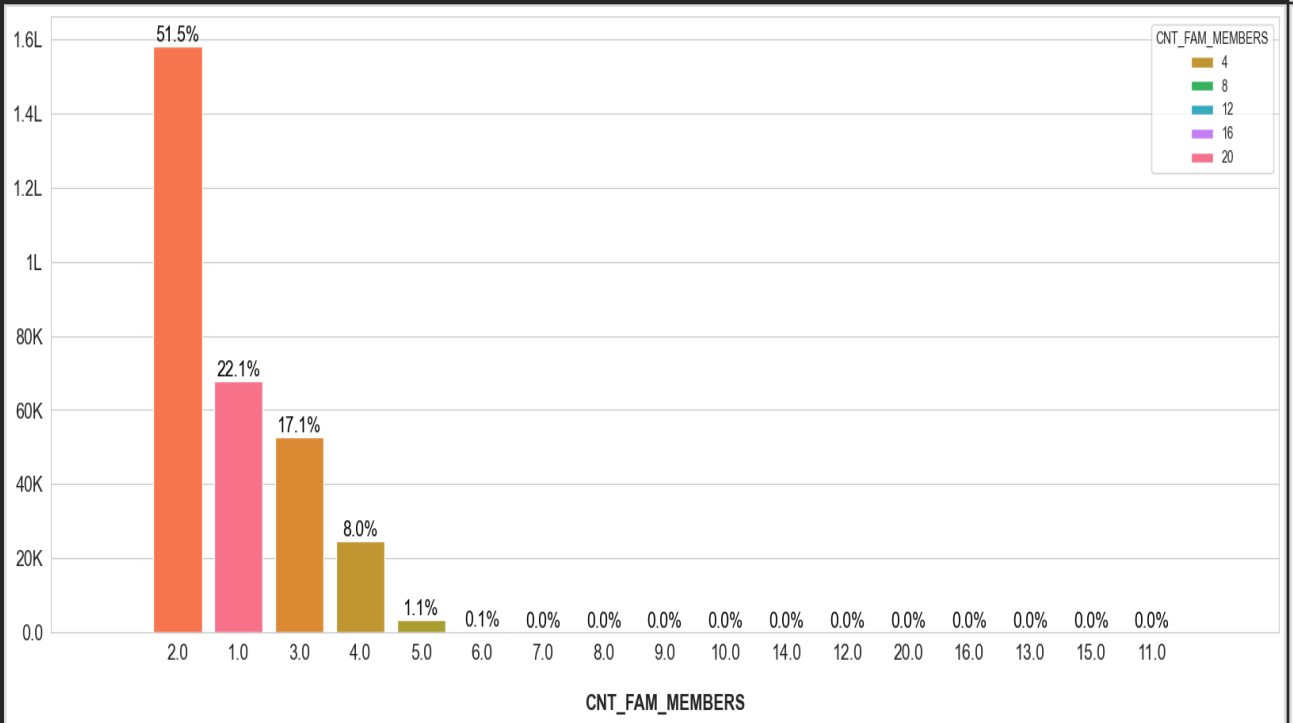


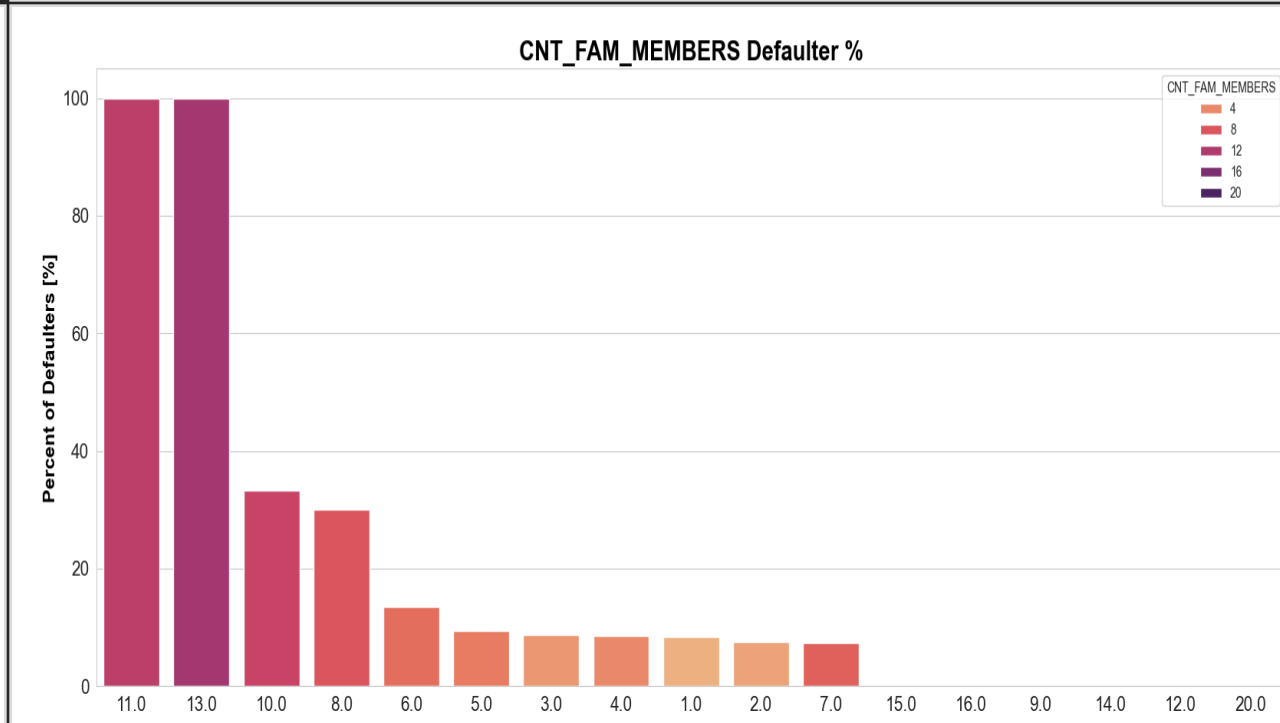
ANALYSIS

FAMILY MEMBERS COUNT

- Similar to the one with children, the defaulter rate increases with the count of family members.



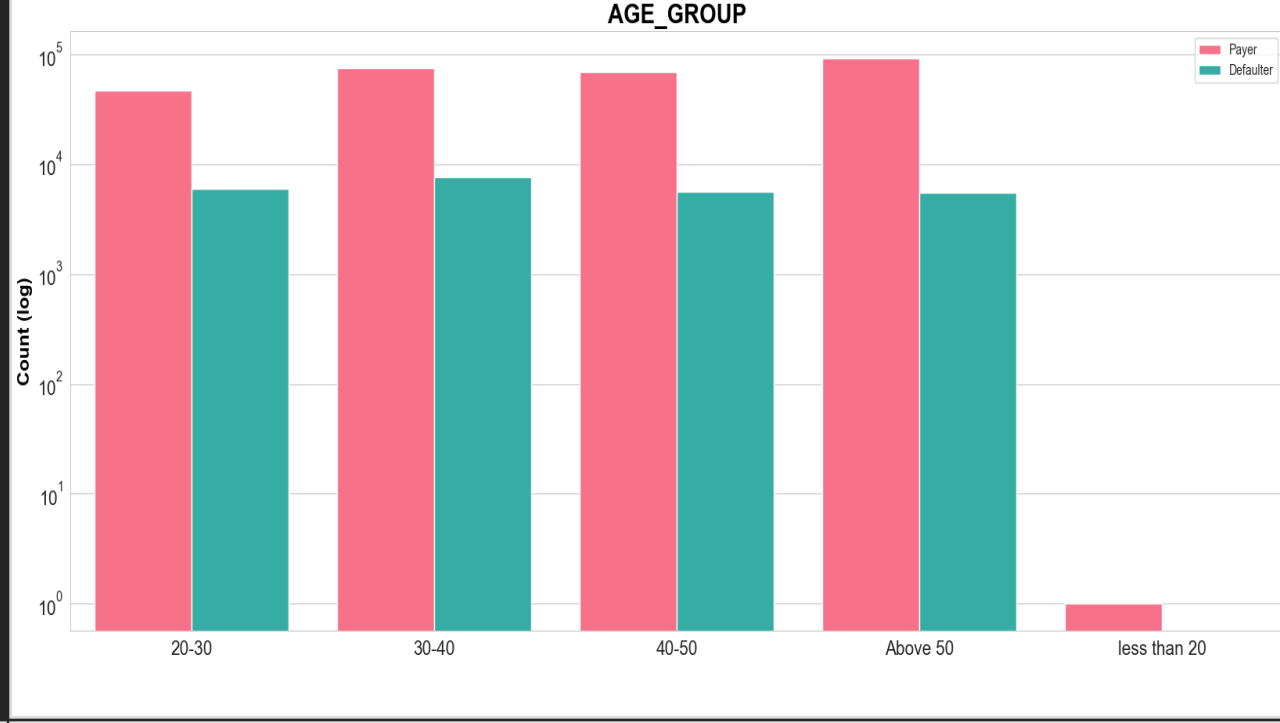


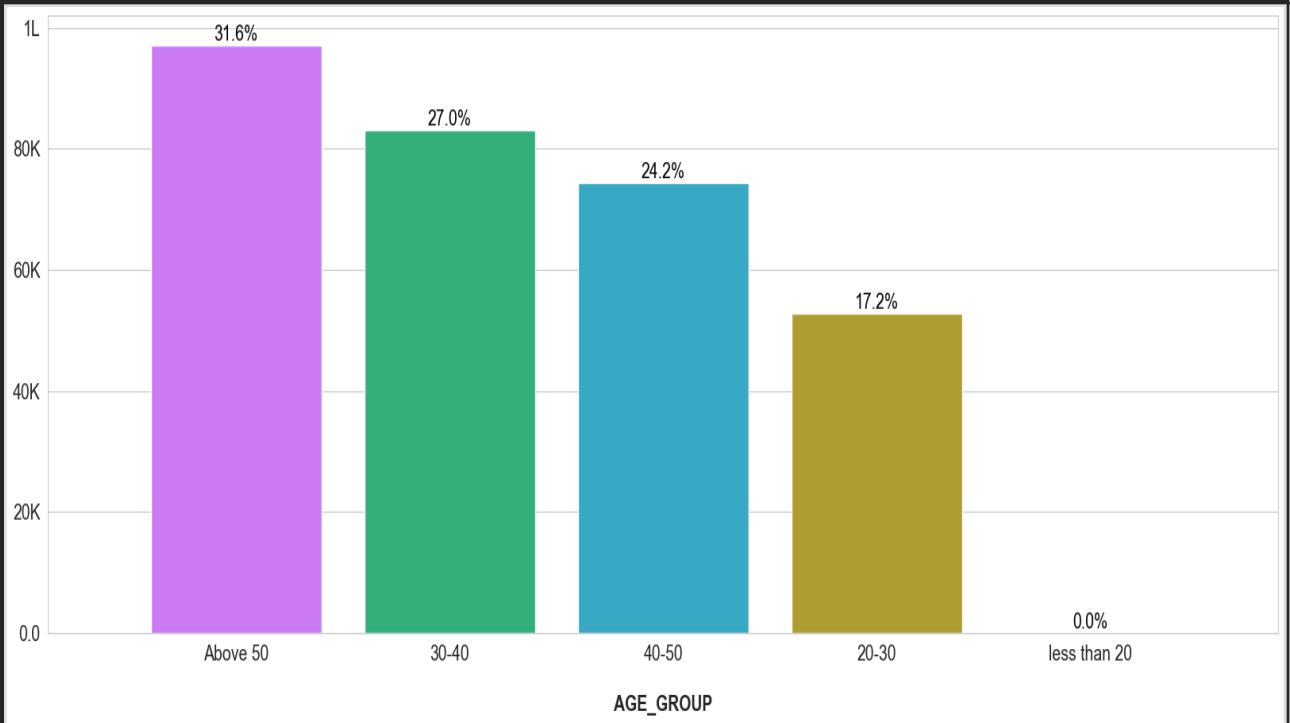


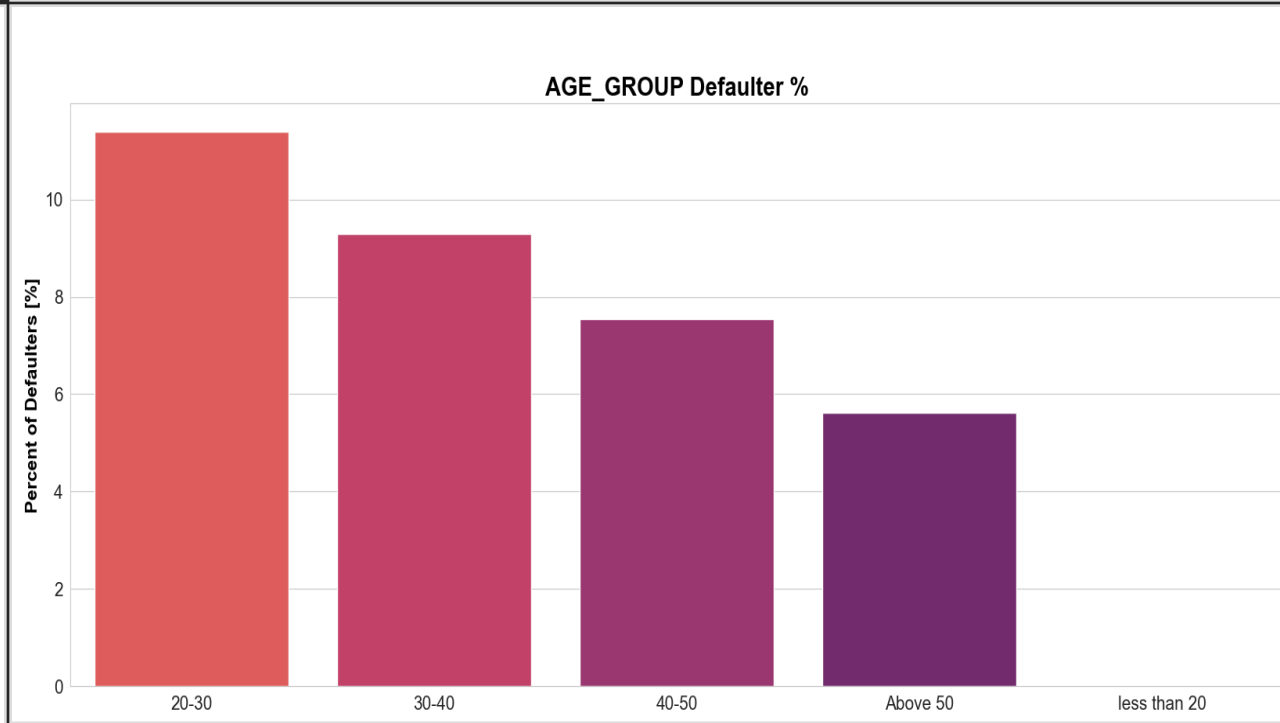
ANALYSIS

AGE

- The loan applicants' age affects their default rate.
- The highest default rate is among applicants aged 20 to 40.
- The lowest default rate is among applicants aged above 50, who are also the most frequent loan applicants.
- The majority of the loan applicants are aged above 30.



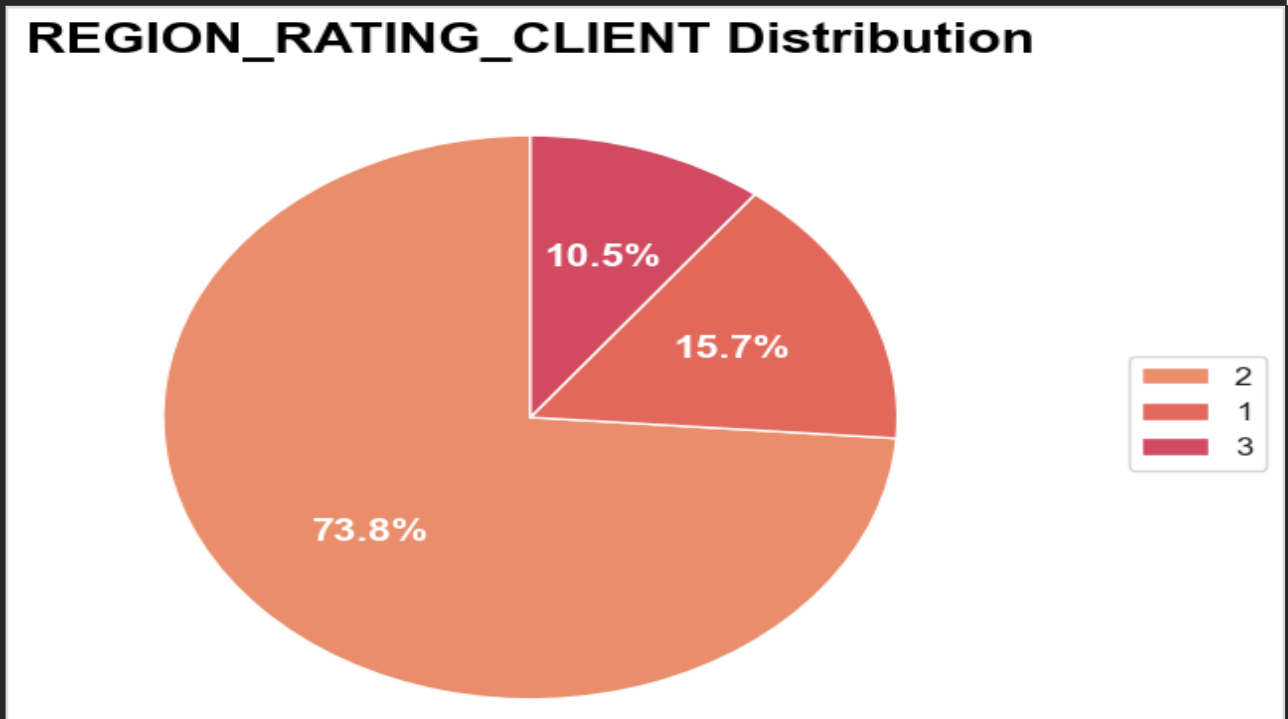


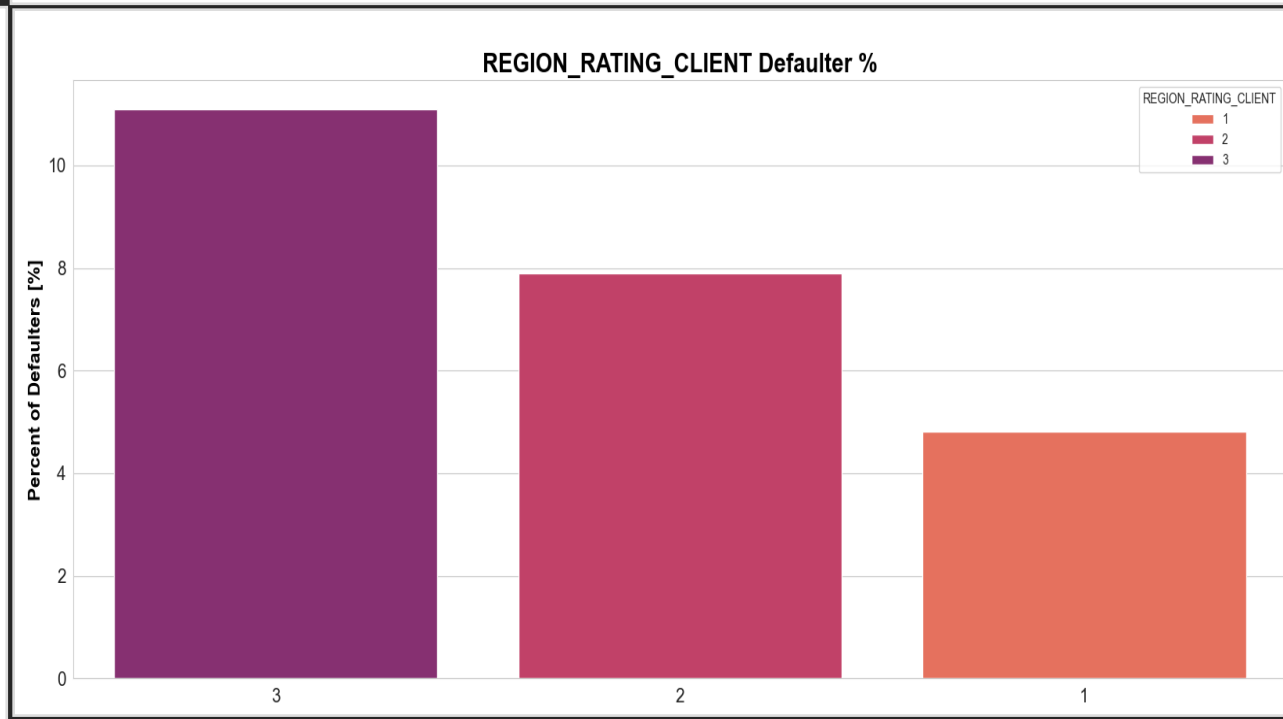


ANALYSIS

REGION RATING

- Focus on applicants living in regions with Rating 1, as they have the lowest defaulting rate of less than 5%.
- Monitor applicants living in regions with Rating 2, as they constitute the majority of the applicants with about 74%.
- Avoid applicants living in regions with Rating 3, as they have the highest defaulting rate of about 11%.

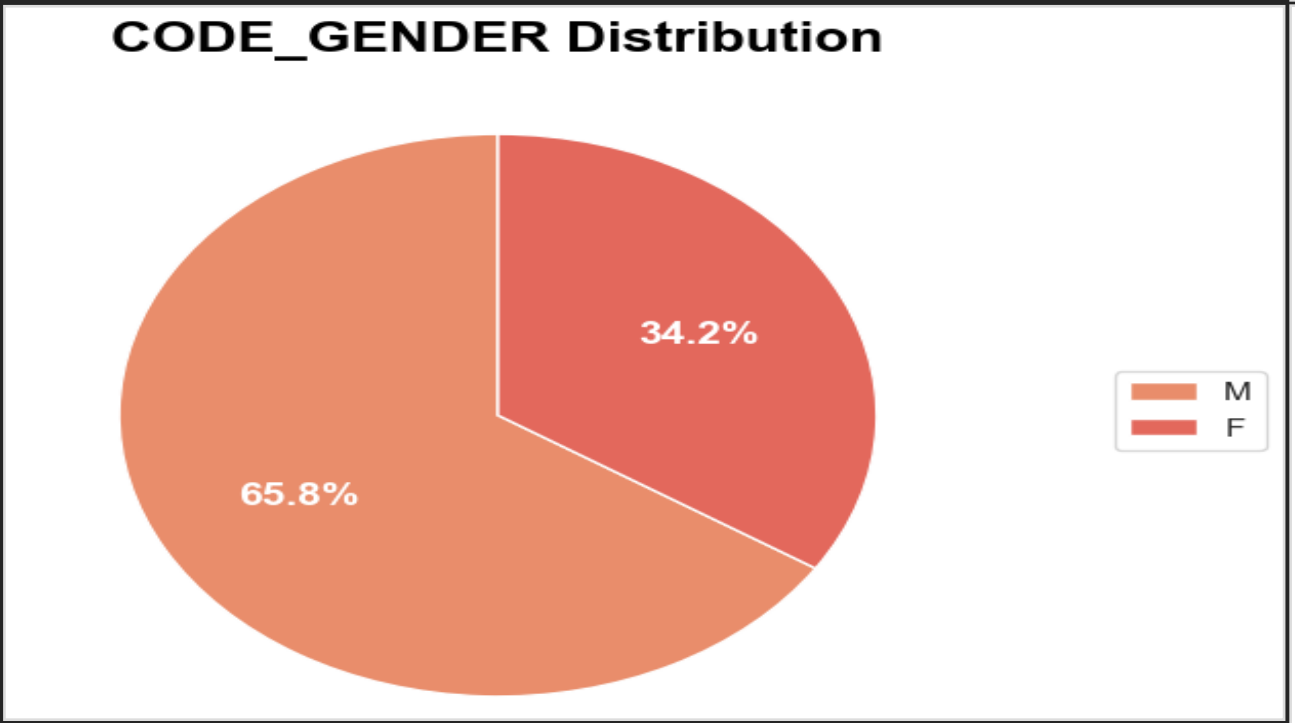


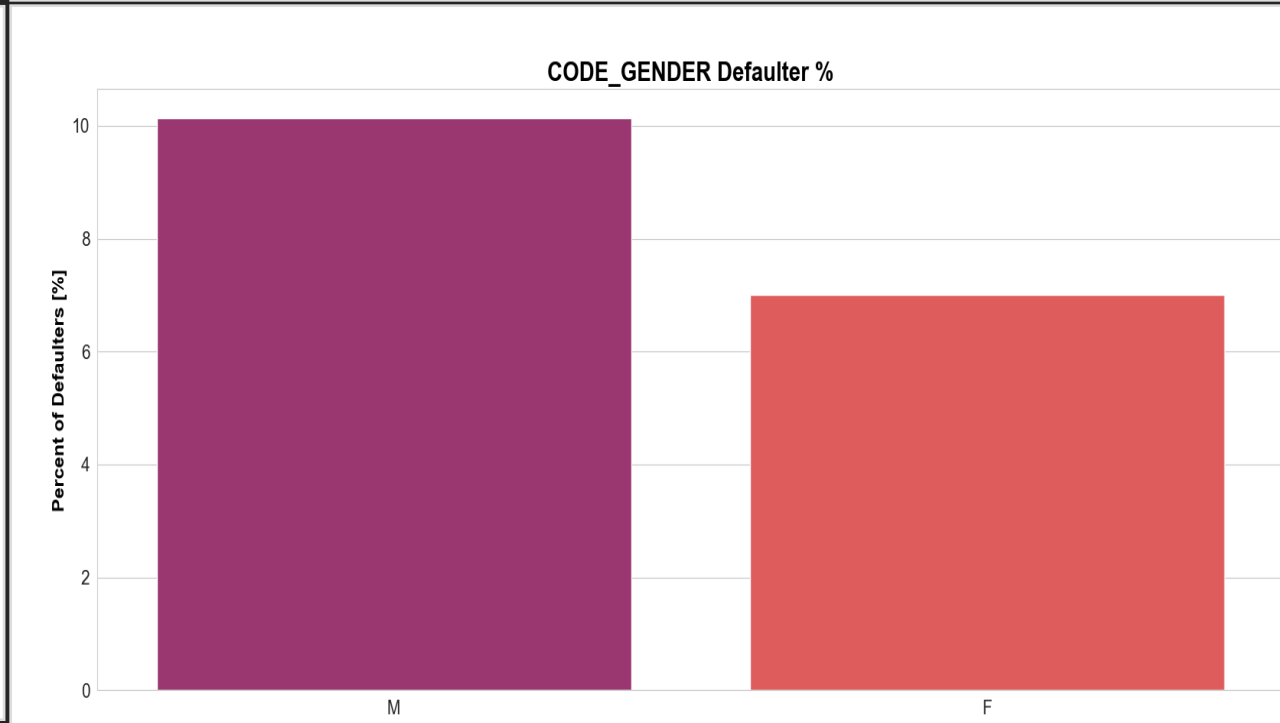


ANALYSIS

GENDER

- Female clients are almost twice than that of Males.
- Further, the proportion of Male defaulters are higher as compared to Females
- Females and Males have a defaulter rate of approximately 7% and 10% respectively.

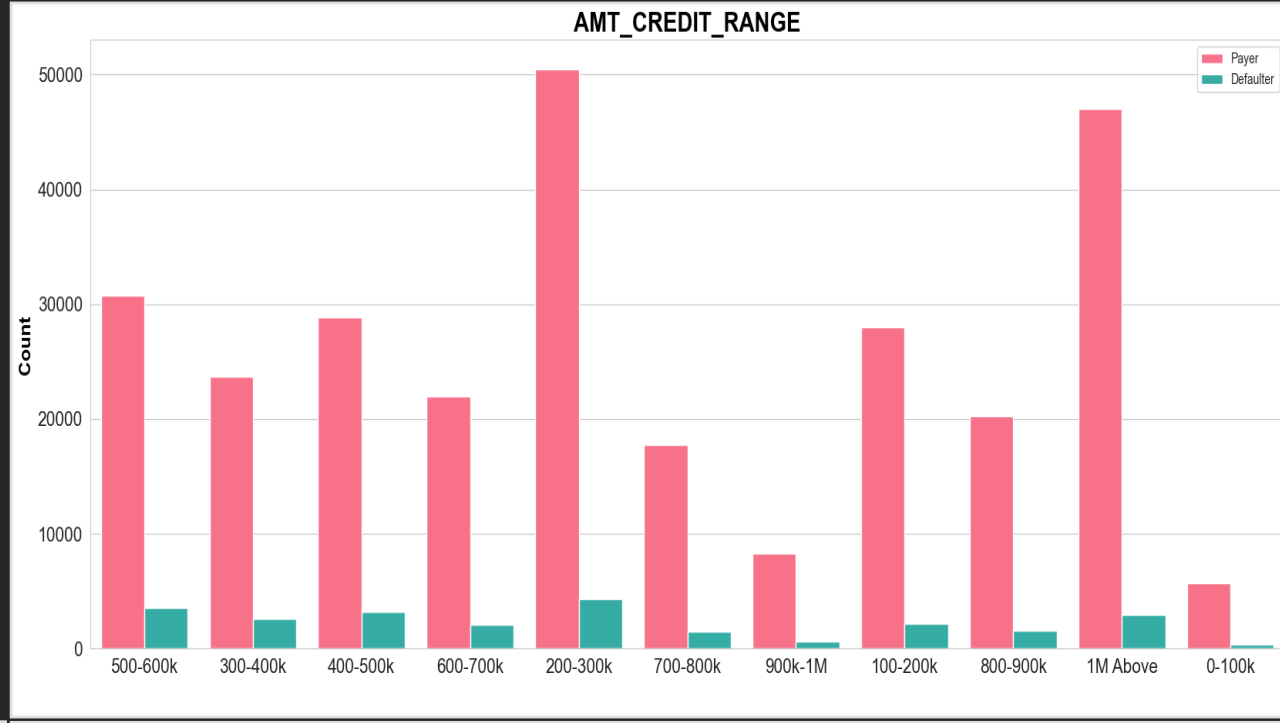


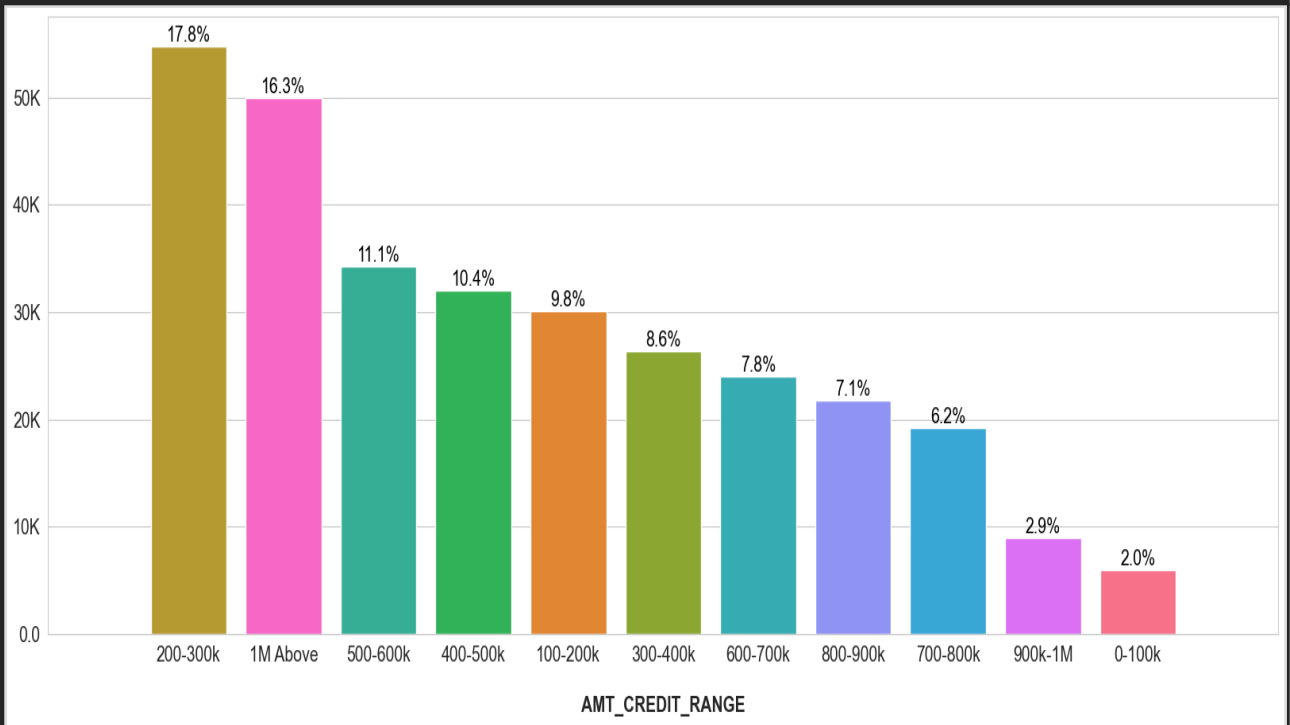


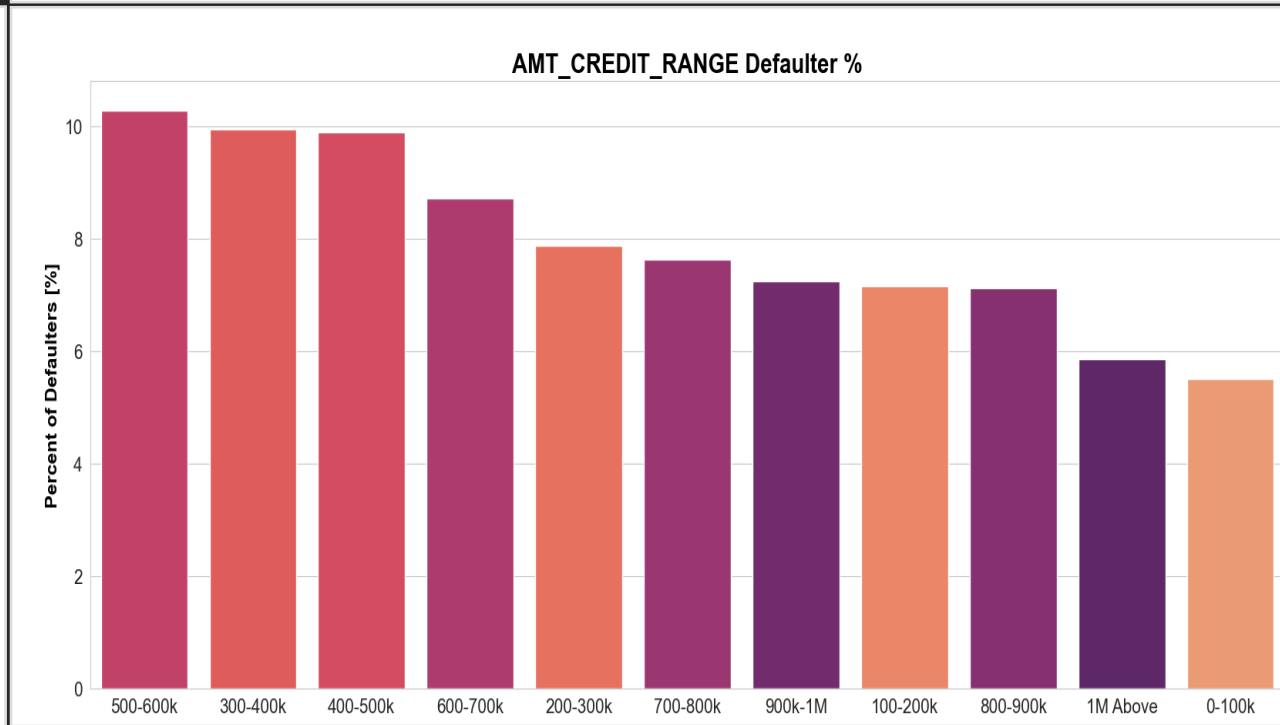
ANALYSIS

CREDIT AMOUNT

- The most common loan amount range is 200K-300K, followed by 1M+.
- The loan amount range with the highest default rate is 300K-600K.



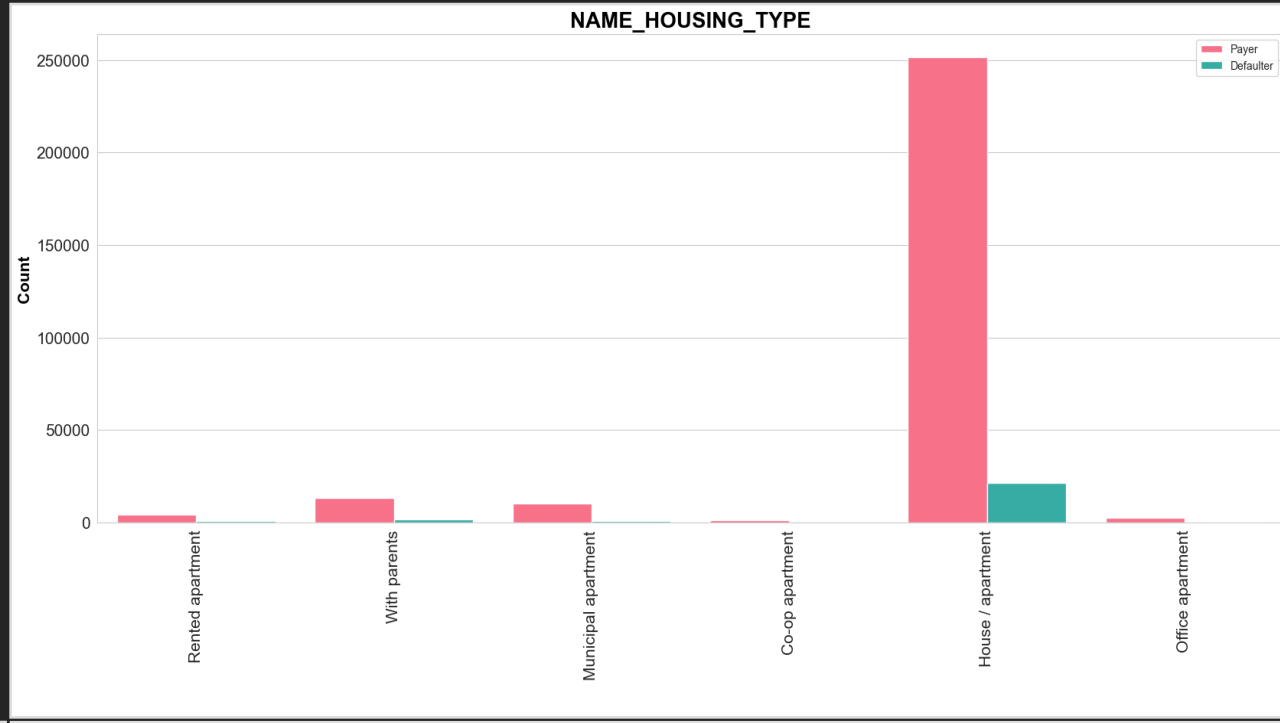


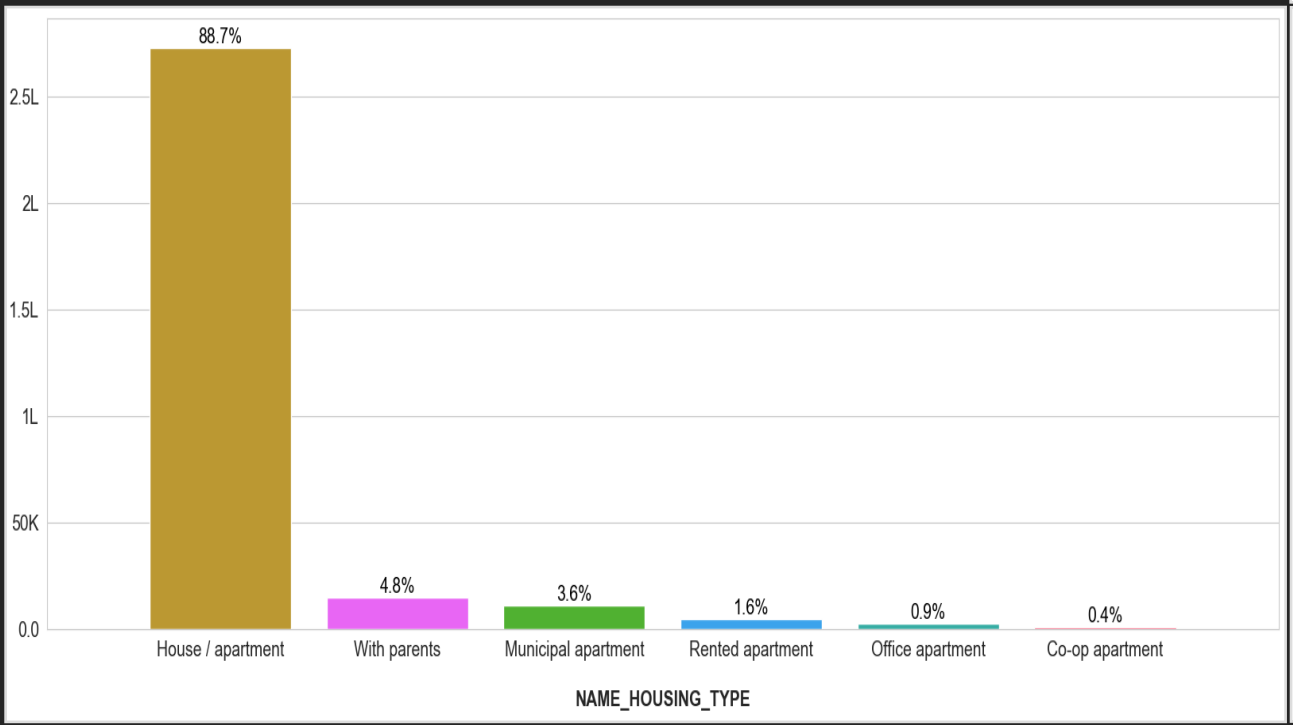


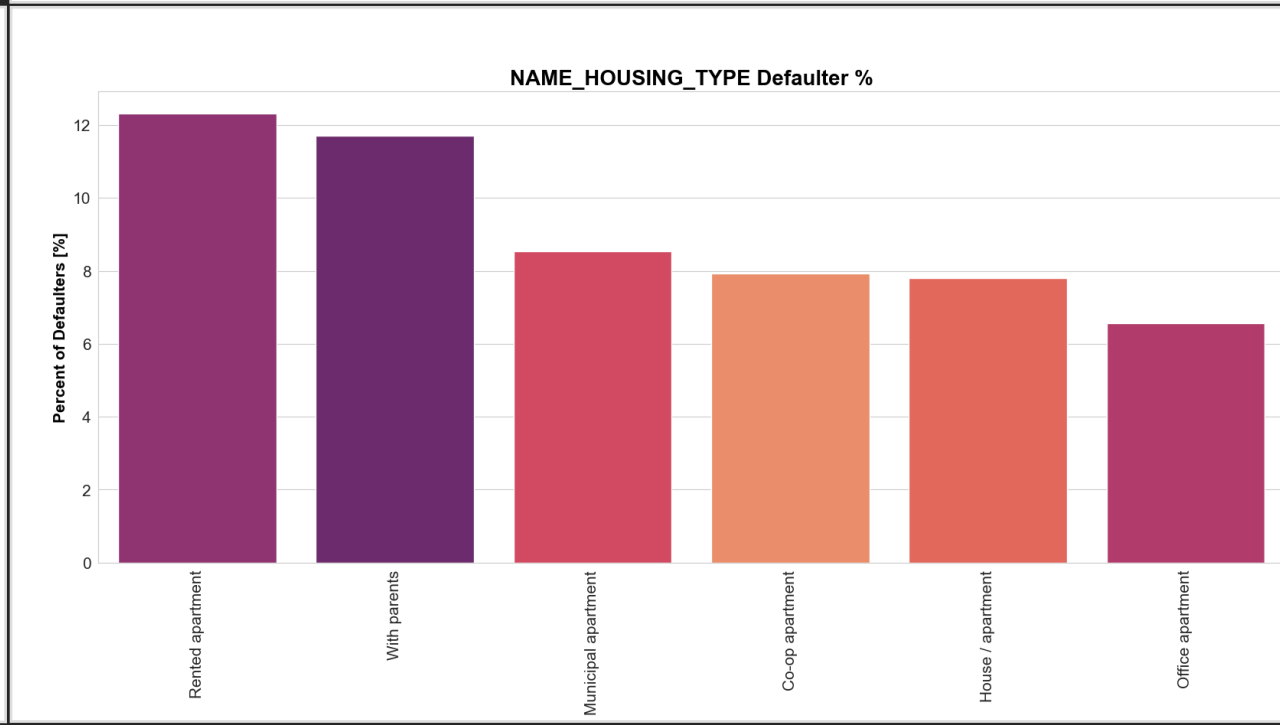
ANALYSIS

TYPE OF HOUSING

- The most prevalent living arrangement is House/Apartment (88% of the borrowers).
- The living arrangement with the lowest default rate is Office Apartment (0.6%), followed by House/Apartment (7.9%).
- The living arrangement with the highest default rate is Rent or with Parents (12%).







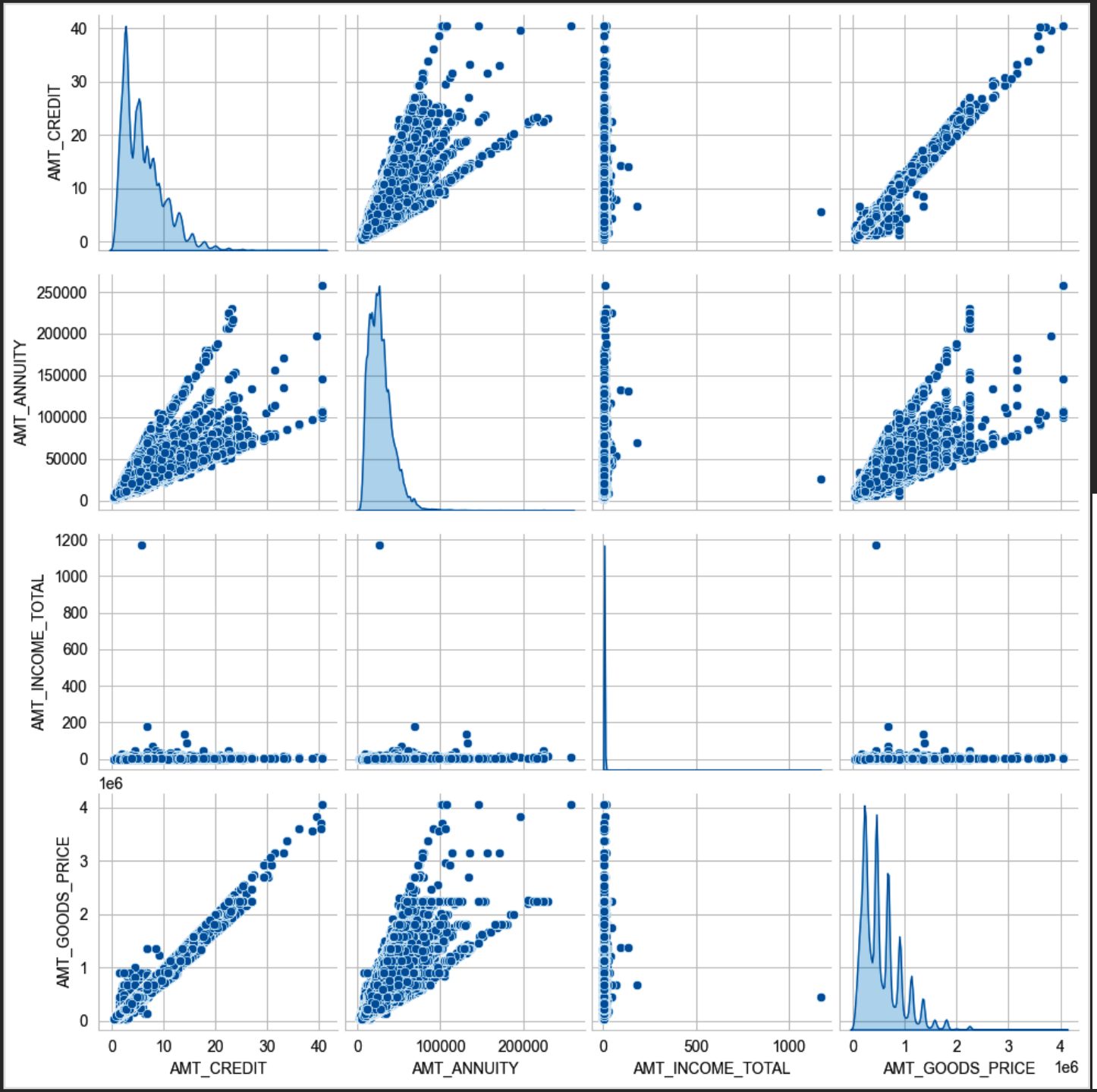
BIIVARIATE AND MULTIVARIATE ANALYSIS

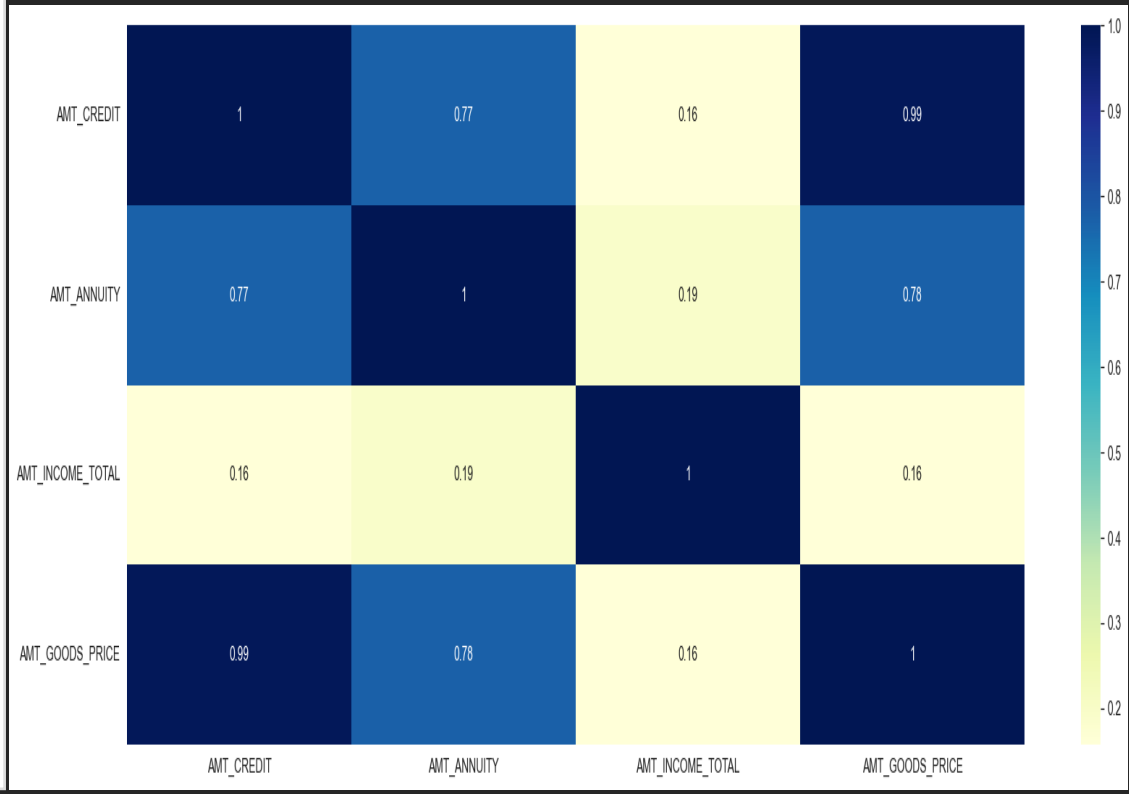
ANALYSIS

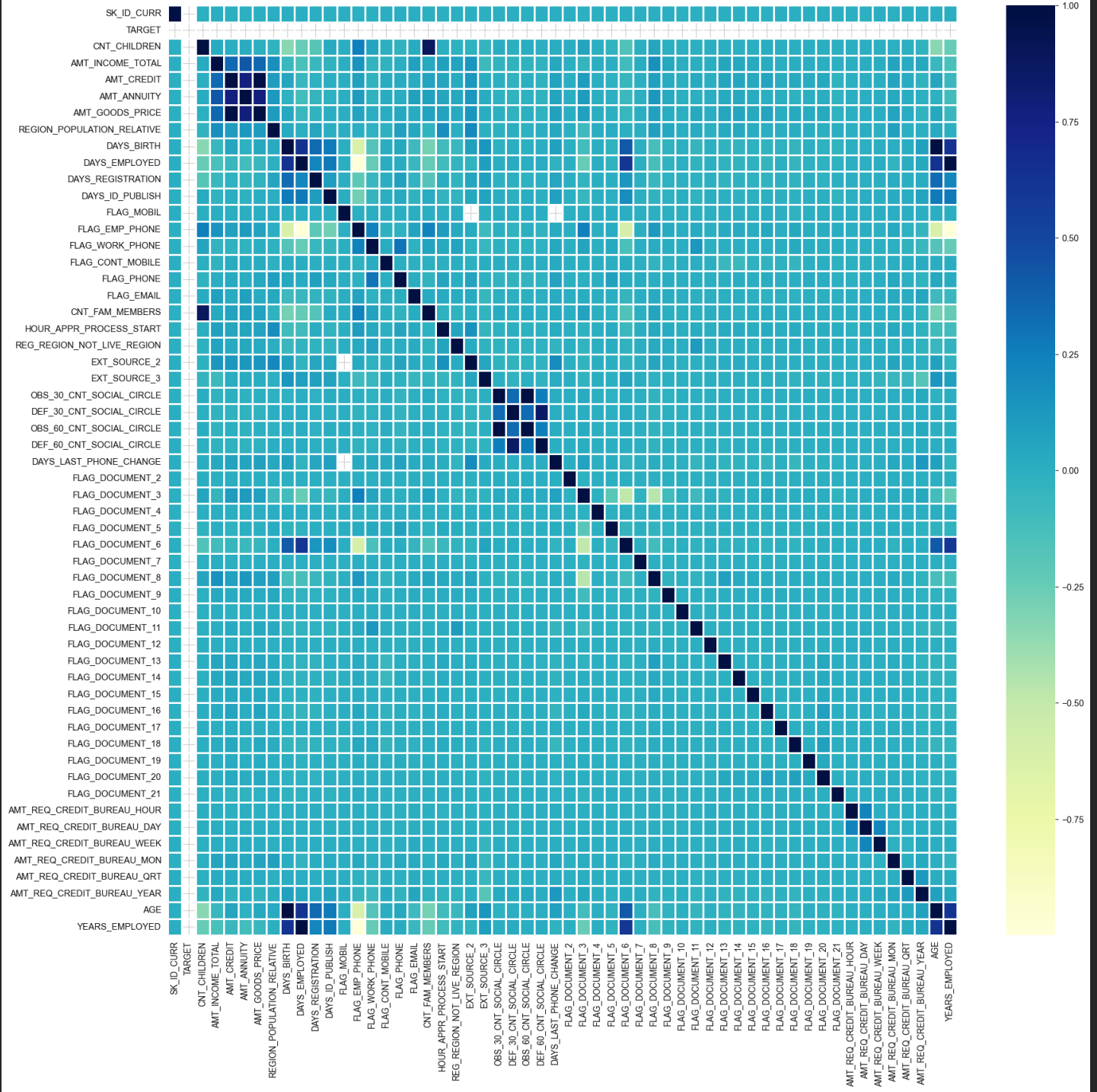
AMOUNT COLUMNS

AMT_CREDIT, AMT_ANNUITY, AMT_INCOME_TOTAL, AMT_GOODS_PRICE

- There is a high linear co-relation between AMT_CREDIT and AMT_GOODS_PRICE.





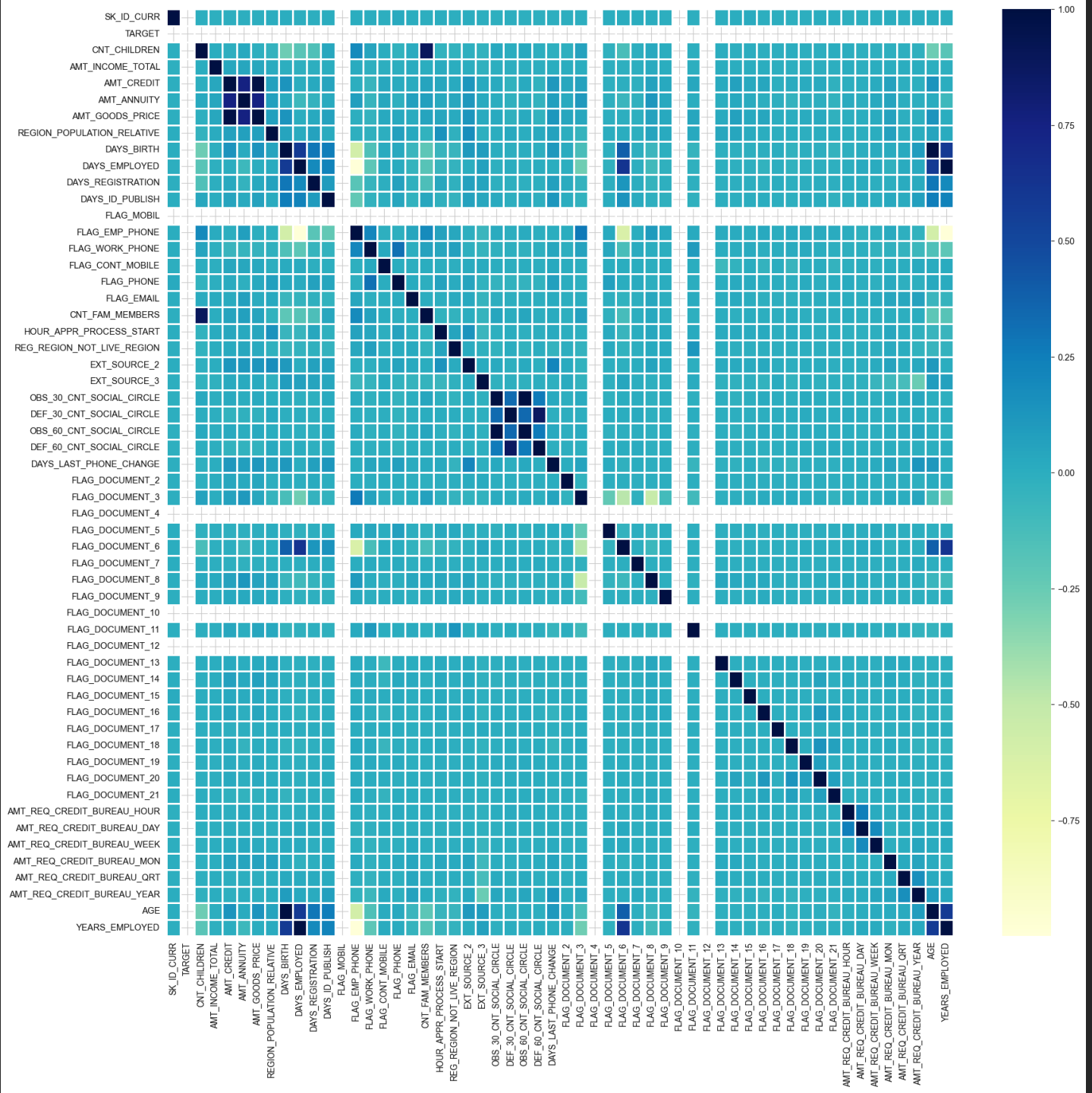


ANALYSIS

NUMERIC COLUMNS

ALL NUMERICAL COLUMNS WHEN TARGET=0 (PAYERS)

- Amongst the payers, there seems to be a high correlation of the Credit amount with the Loan Annuity, Total Income and Goods price.
- Also, we could see a high correlation with the number of Days/Years employed i.e. Experience and Days birth i.e. Age.



ANALYSIS

NUMERIC COLUMNS

ALL NUMERICAL COLUMNS WHEN TARGET=1 (DEFAULTERS)

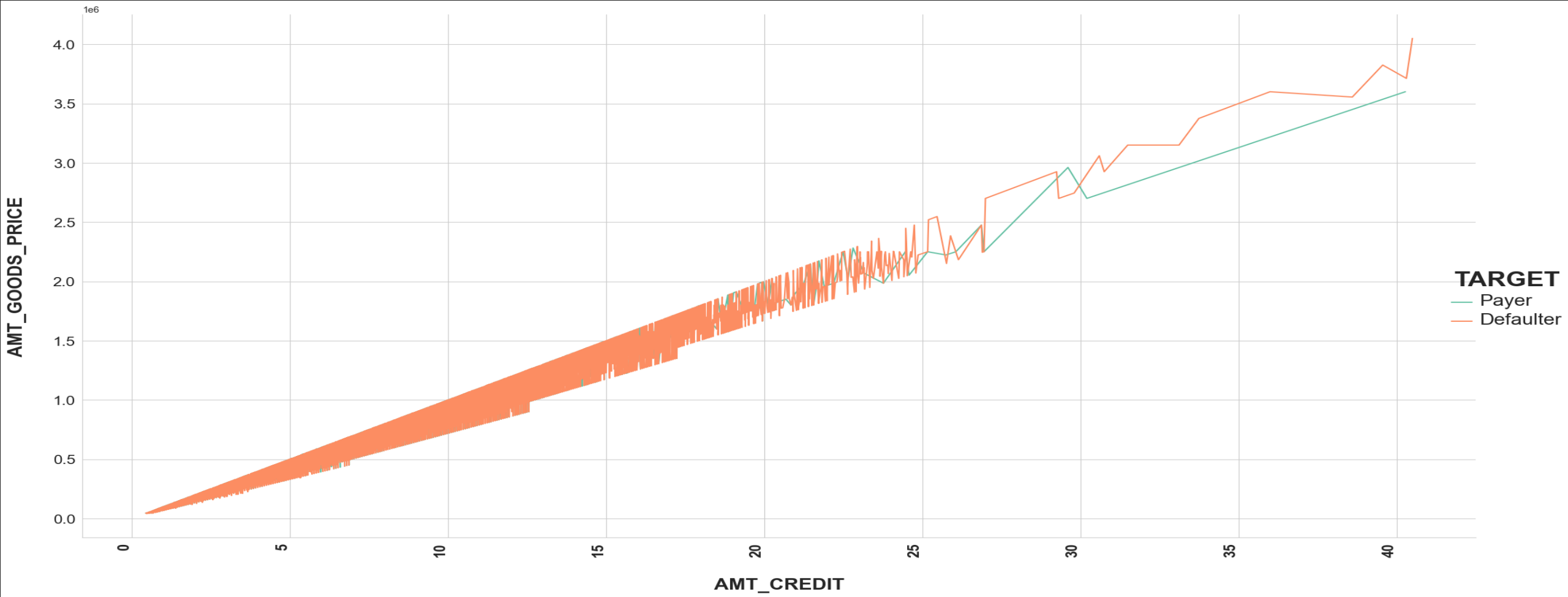
- Even amongst the defaulters, there seems to be a high correlation of the Credit amount with the Goods price.
- However, the amount annuity correlation with the credit amount has slightly reduced for defaulters with 0.75 when compared to Payers with 0.77.
- For defaulters, the correlation between Credit Amount and Income has severely dropped to 0.0381 when compared with 0.3428 for Payers.
- Correlation between Days since birth and count of children has reduced to 0.259 for defaulters when compared with 0.34 of Payers.
- There is a slight increase in defaulted to observed count in social circle among defaulters (0.264) when compared to payers (0.255).

ANALYSIS

CREDIT AMOUNt & GOODS PRICE WITH TARGET

AMT_CREDIT, AMT_GOODS_PRICE, TARGET

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

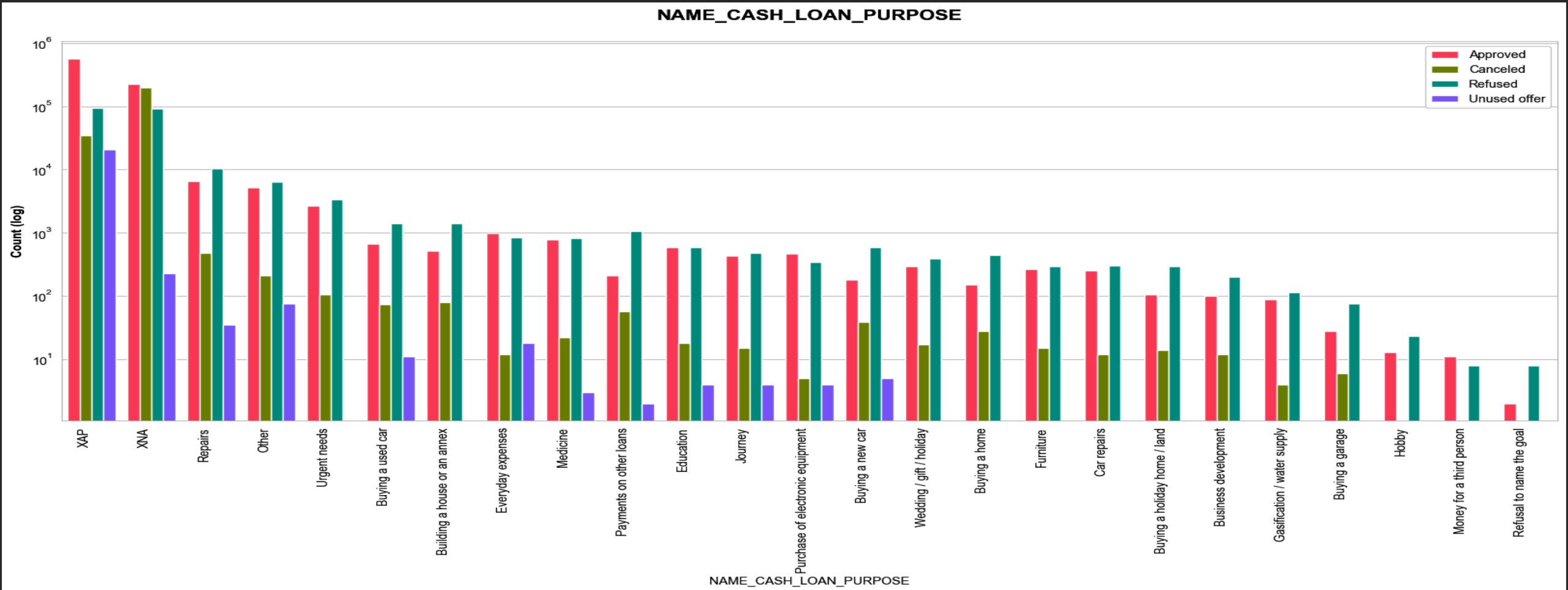


ANALYSIS

MERGED DATAFRAME

NAME_CASH_LOAN_PURPOSE WHEN TARGET=0 (PAYERS)

- Loan purpose has a high number of unknown values (XAP, XNA)

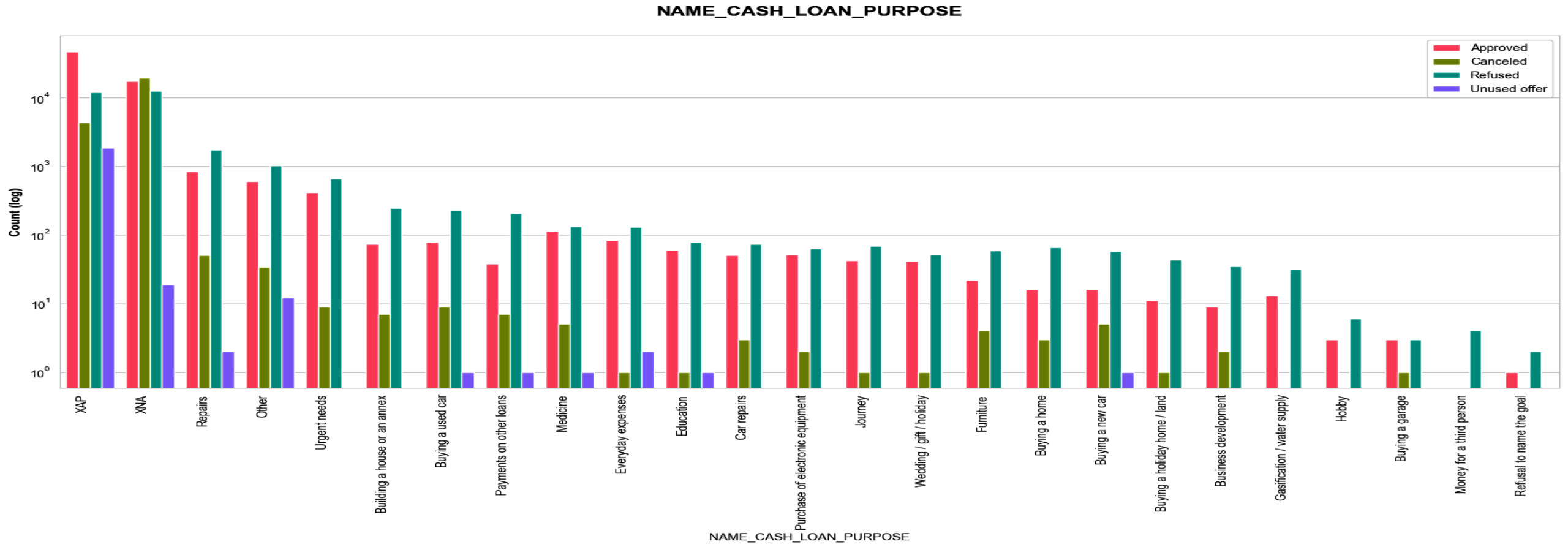


ANALYSIS

MERGED DATAFRAME

NAME_CASH_LOAN_PURPOSE WHEN TARGET=1 (DEFAULTERS)

- Loan taken for repair appears to have the greatest default rate
- A large number of applications with the intention of "repair or other" have been rejected by banks or declined by clients.

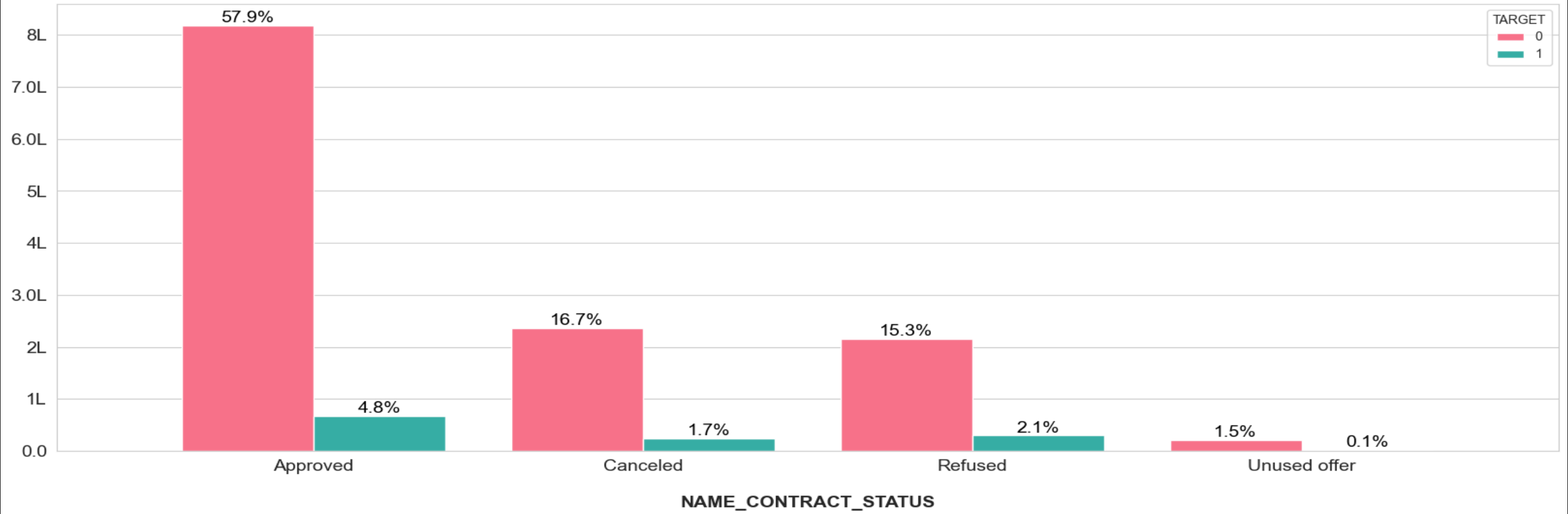


ANALYSIS

MERGED DATAFRAME

NAME_CONTRACT_STATUS, TARGET

- 90% of clients, who had their loans cancelled previously, have actually repaid their loans. By reconsidering the interest rates, the bank could potentially create more business opportunities from these clients.
- 88% of the clients who were previously denied a loan have managed to repay their loans in recent cases.



WHAT DO WE INFER?

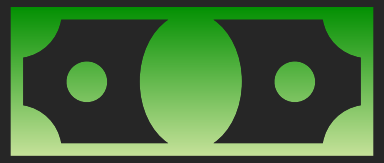
SUMMARY

CASE STUDY

KEY ELEMENTS FOR A BORROWER TO BE CREDITWORTHY - I

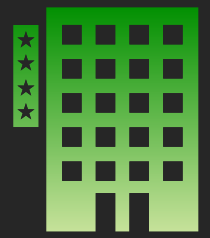
Businessmen are good at repaying their loans

Applicants with extensive work experience (40+ years) have an excellent default rate of less than 1%



Higher income applicants (700,000+) have a strong repayment record

Applicants from regions with Rating 1 are the safest borrowers



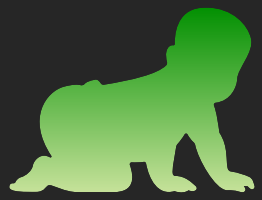
Applicants from Trade Type 4 and 5 and Industry Type 8 are quite trustworthy borrowers, with a default rate of less than 3%

KEY ELEMENTS FOR A BORROWER TO BE CREDITWORTHY - II

Academic degree holders have a low default rate

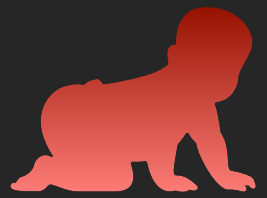
Older applicants (50+ age) are very reliable in repaying their loans

Loans for personal interests or needs, such as hobbies or buying a garage, are mostly repaid on time



Having up to two children does not affect loan repayment

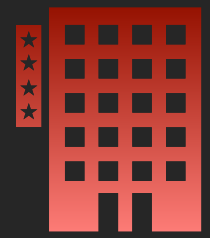
KEY ELEMENTS FOR A BORROWER TO BE DEFAULTER - I



Applicants with 9 or more children have a 100% default rate and should not be approved

Applicants with large families (11 or more members) also have a high default rate and should not be approved

Applicants who are on maternity leave or unemployed have a high default rate and should be avoided

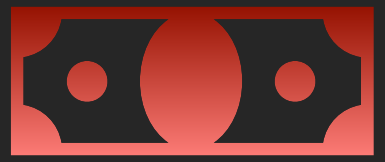


Applicants who work in low-skill or service occupations have a huge default rate and should not be granted loans

Organizations in Transport: type 3, Industry: type 13 Industry: type 8, Restaurant and Self-employed people

KEY ELEMENTS FOR A BORROWER TO BE DEFAULTER - II

Applicants who live in regions with Rating 3 are the most risky borrowers and have the highest defaults



Credit amounts above 3M have a higher default risk

Male applicants tend to default more than female applicants

Young applicants (20-40 years old) have a higher probability of defaulting and should be screened carefully

KEY ELEMENTS FOR A BORROWER TO BE DEFAULTER - III

Applicants with less than 5 years of work experience have a high default rate and should be evaluated thoroughly

Applicants with low or incomplete education levels have a high default rate and should be given lower priority

Applicants who are single or in civil marriage have a high default rate and should be assessed cautiously

END