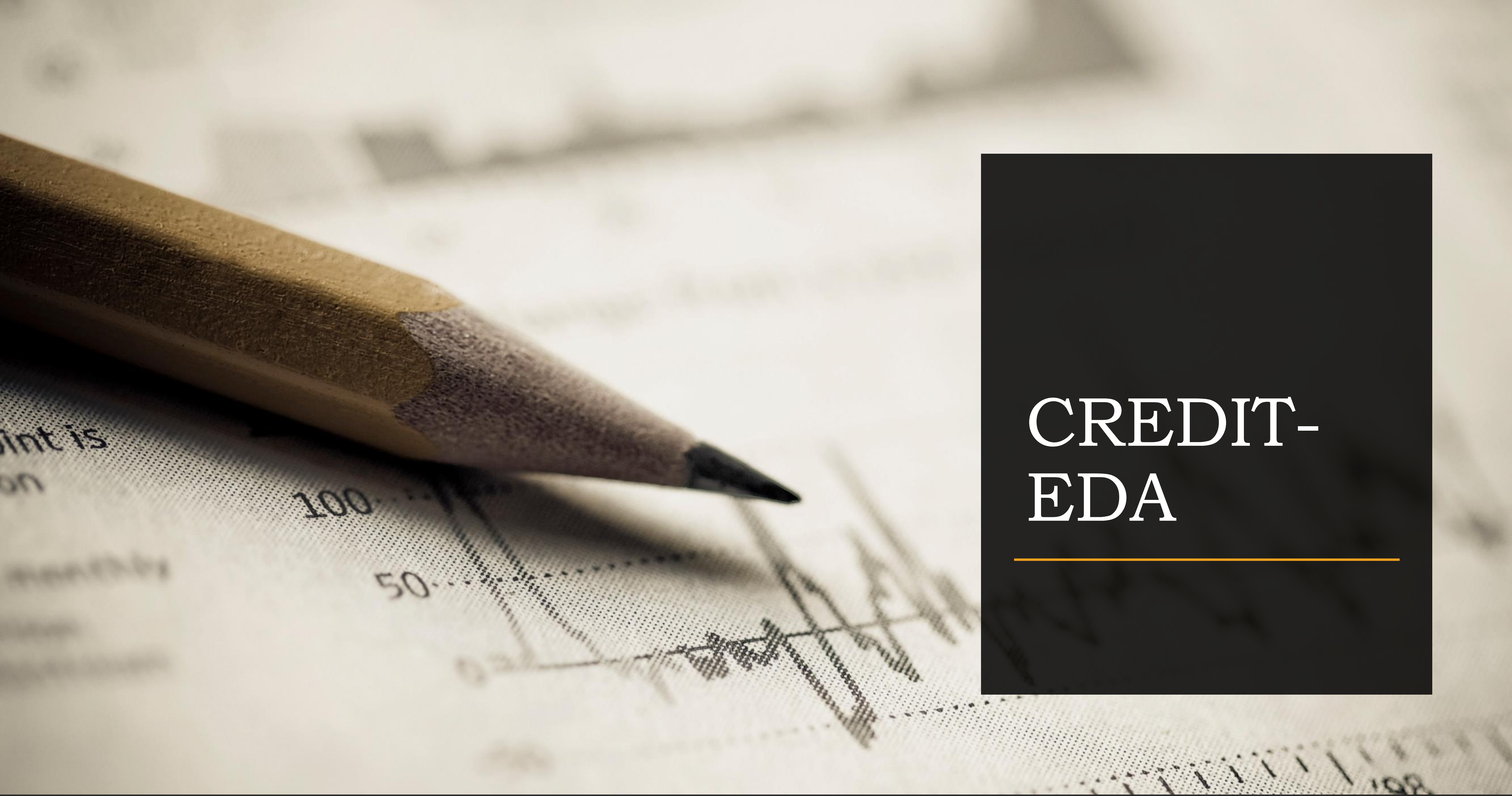


# CREDIT-EDA

# A Table of Contents

- Business objective
- Aim of the study
- Data Engineering
- Exploratory Data Analysis
- Analysis
- Conclusion

# Business Objective

The objective of this case study is to uncover discernible patterns that serve as indicators of clients experiencing challenges in paying their instalments. These patterns will enable us to take appropriate actions, including but not limited to denying loans, reducing loan amounts, and offering loans to risky applicants at higher interest rates. By leveraging exploratory data analysis (EDA), we aim to identify such applicants, ensuring that deserving consumers capable of repaying their loans are not unjustly rejected.

# Aim of Study

We will be taking  note of the following things during our field inspection:

- Prepare dataset for exploratory data analysis
    - Evaluate missing values
    - Treatment of outliers
    - Feature selection

- Perform Exploratory Data Analysis (EDA)
    - Identification of faulty applicants
    - Understand features with target variable
    - Find patterns and trends related to faulty applicants

# Data Engineering

Credit dataset has data of 307511 clients with 121 features and target variable.

The features present in the dataset were divided into numerical and categorical data.

| Data Type | Numerical | Categorical |
|---|---|---|
| Count | 106 | 16 |

The data engineering process for this project involved the utilization of Python libraries such as Pandas, NumPy, matplotlib, and seaborn. These powerful libraries were instrumental in performing various data manipulation, transformation, and visualization tasks, enabling us to effectively engineer the data for analysis.

# Missing values Treatment

Objective data type had 6 features with more than 0% missing data.

Highly missing features were dropped, followed by descriptive statistics-based removal of imbalanced and insignificant variables.

Numerical data type had 47 features with more than 40% missing data.

Missing data was examined for patterns of Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). Through meticulous analysis of descriptive statistics, irrelevant data was subsequently eliminated. Additionally, flag columns were classified to facilitate more detailed feature analysis.

Dataset was compressed to 307492 Clients and 70 significant features.

# Outliers Treatment

Three numeric columns were identified to contain outliers through the application of the boxplot method. The distribution of the underlying features was extensively analyzed to confirm the presence of outliers. Based on this analysis, informed decisions were made regarding the appropriate approach for handling these outliers.

Rows with outliers which accounted for 0.5% of whole data were dropped from 'AMT_TOTAL' feature. Remaining features were skewed features with relevant clients data.

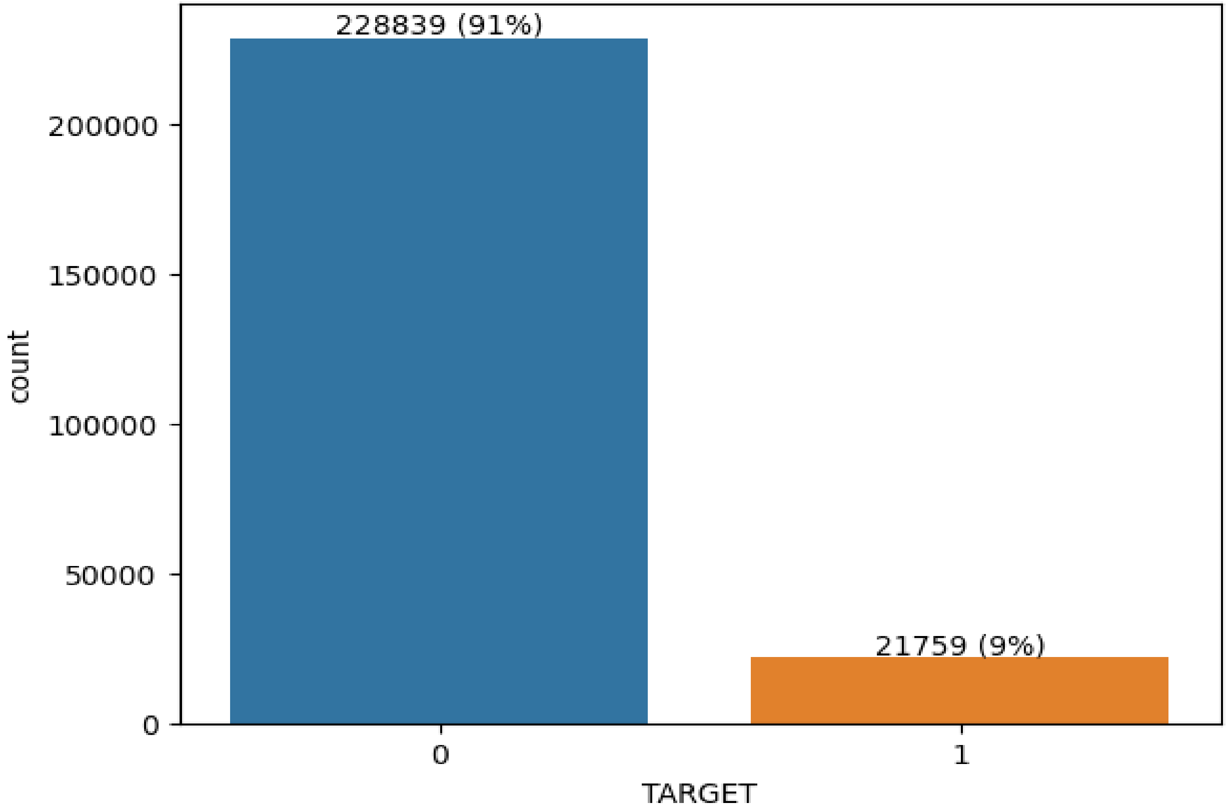
Final data set had 250598 Clients data with 70 significant features.

On similar note 'previous_data' was cleaned. Summary statistics for the previous data is as

| Data Cleaning | Rows | Columns |
|---|---|---|
| Before | 1670214 | 37 |
| After | 1621741 | 17 |

# Univariate Analysis

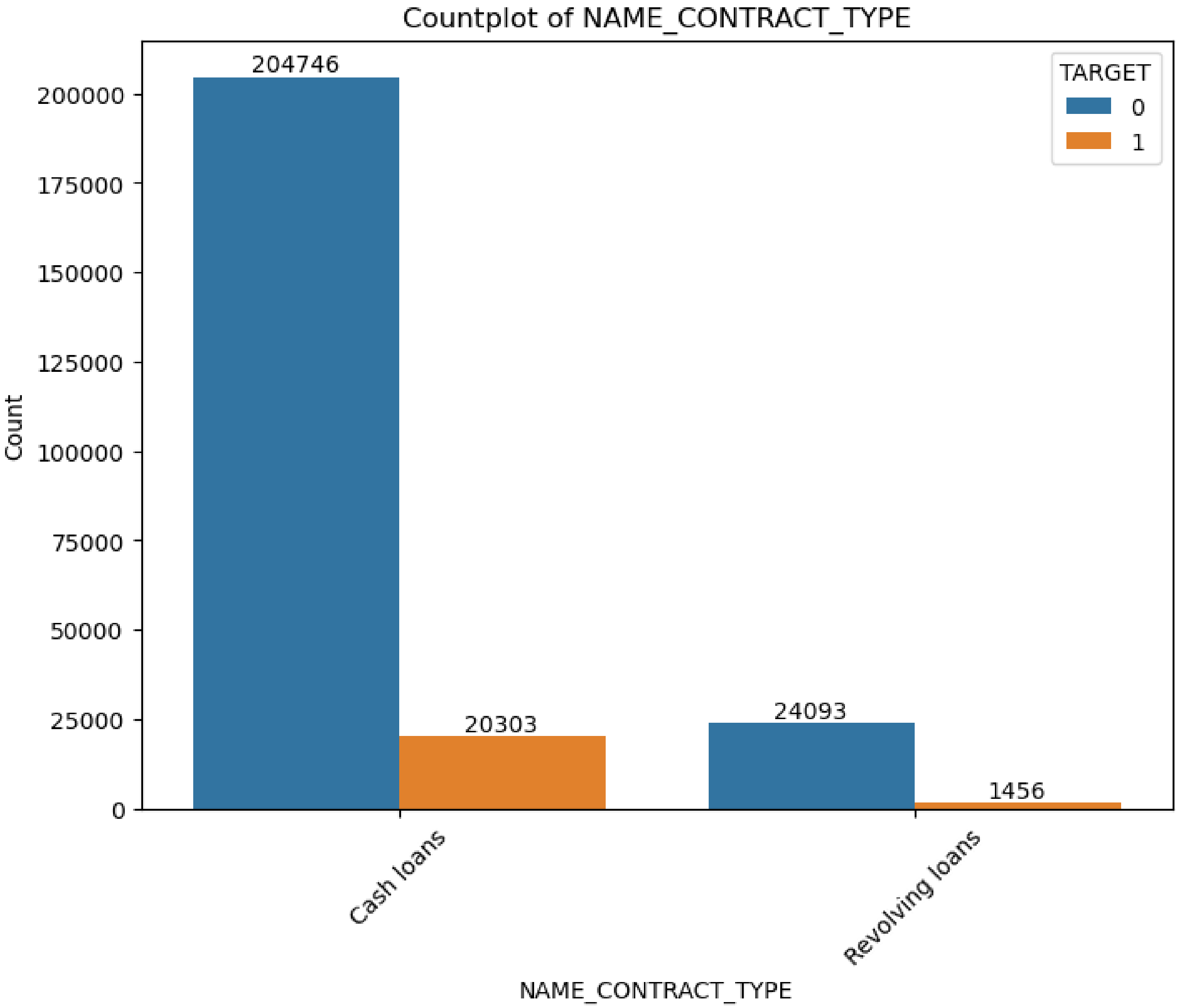
Distribution of Target Variable



Insight

From total clients there are 21759 (9%) individuals who have payment difficulties and 228839 (91%) individuals repaid their loan.

# Distribution of Contract Type

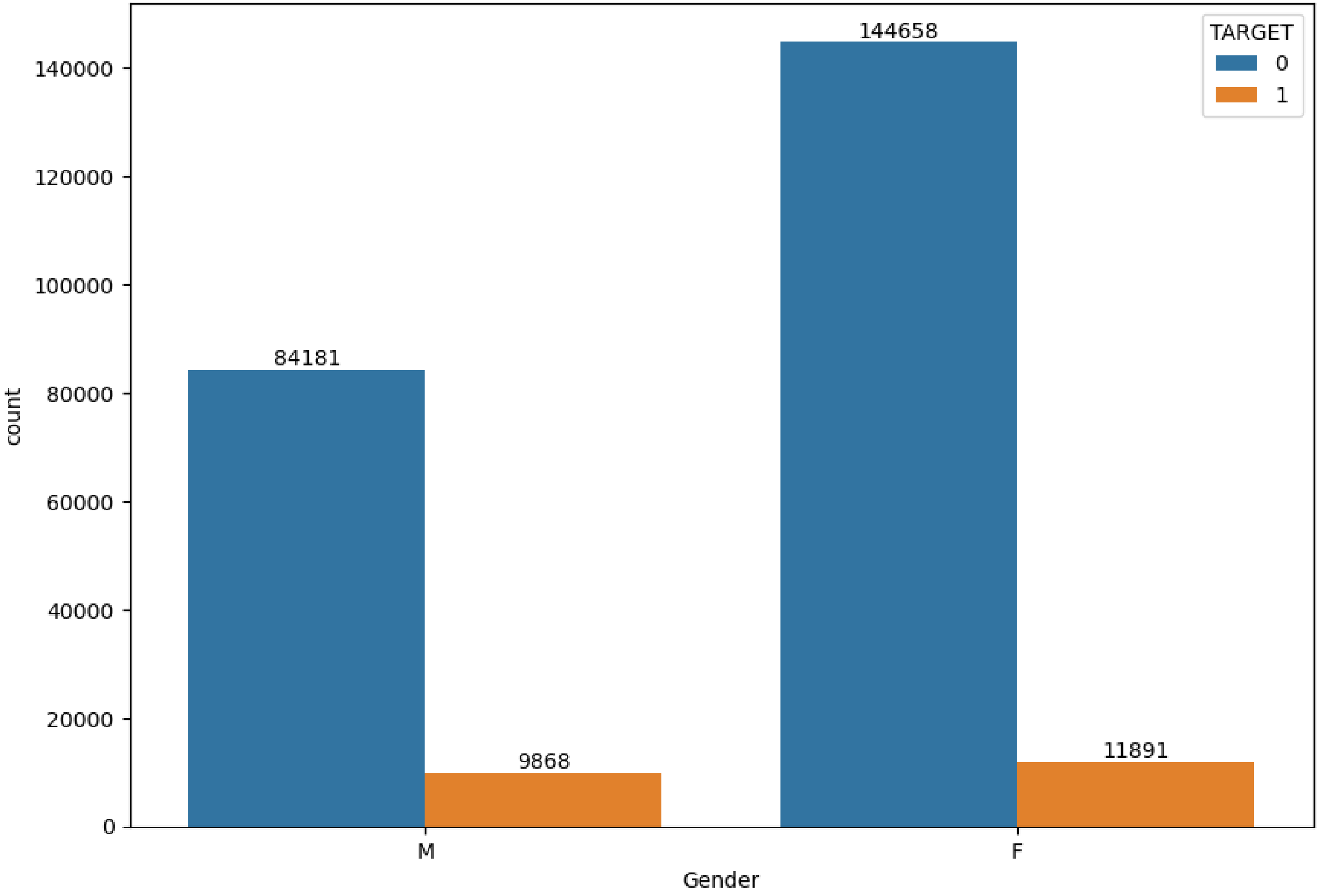


## Insight

Upon meticulous examination, it was revealed that the majority of clients, comprising 88%, held cash loans. In contrast, clients with revolving loans were distributed more diversely throughout the dataset.

When focusing on defaulters, a fascinating pattern emerged. Among those with cash loans, defaulters accounted for 9% of the total, while defaulters with revolving loans represented 5.6%.

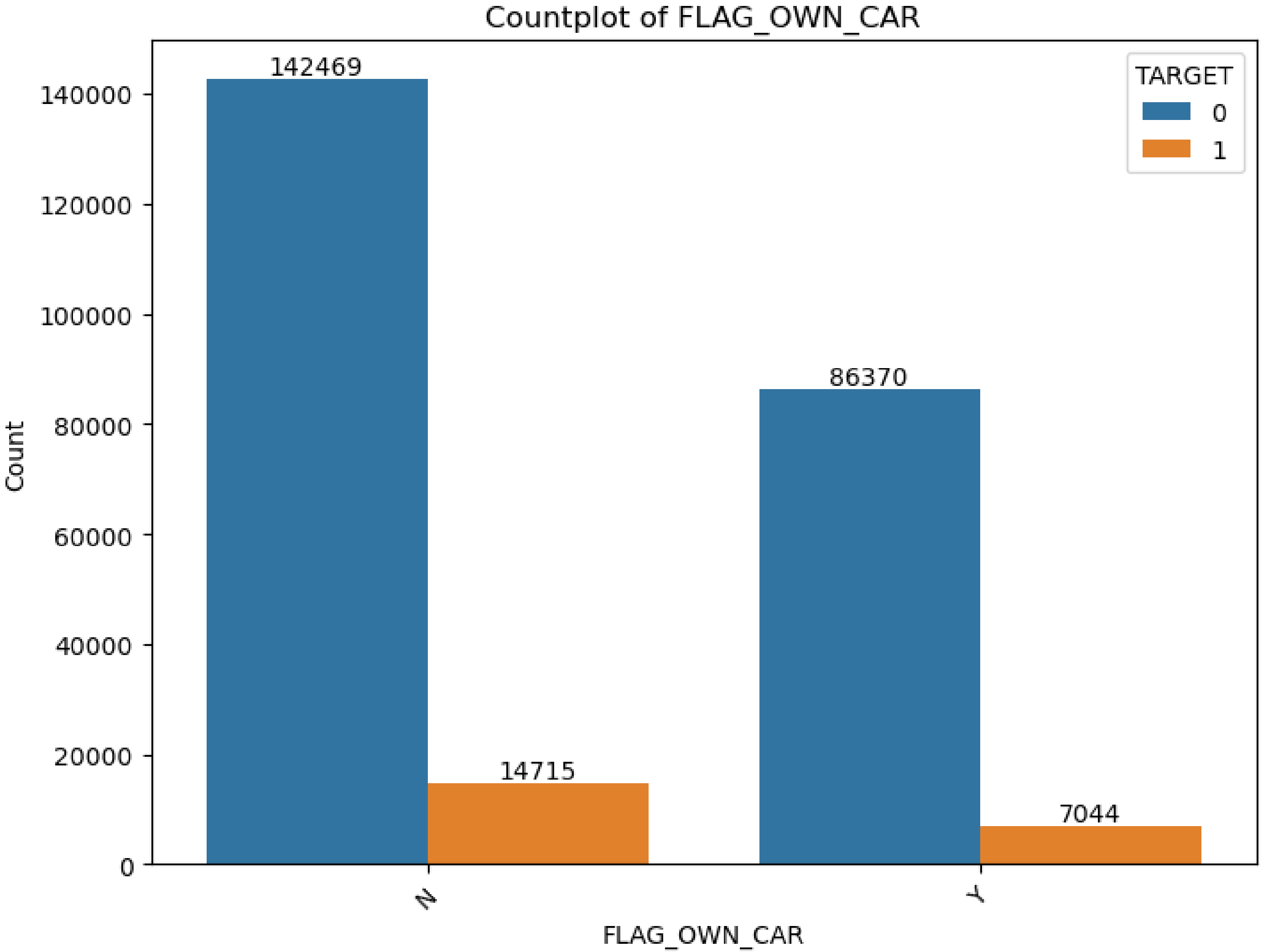
# Distribution of clients over gender



## Insight

Within the overall distribution, an intriguing gender disparity emerged, with 62% of clients identifying as female and 38% as male. However, when examining the subset of defaulters, a captivating shift in percentages was discovered. Female clients accounted for 55% of defaulters, while male clients represented 45%, indicating a slight increase in defaulters among the male population compared to the overall distribution.

# Distribution of clients owning own car

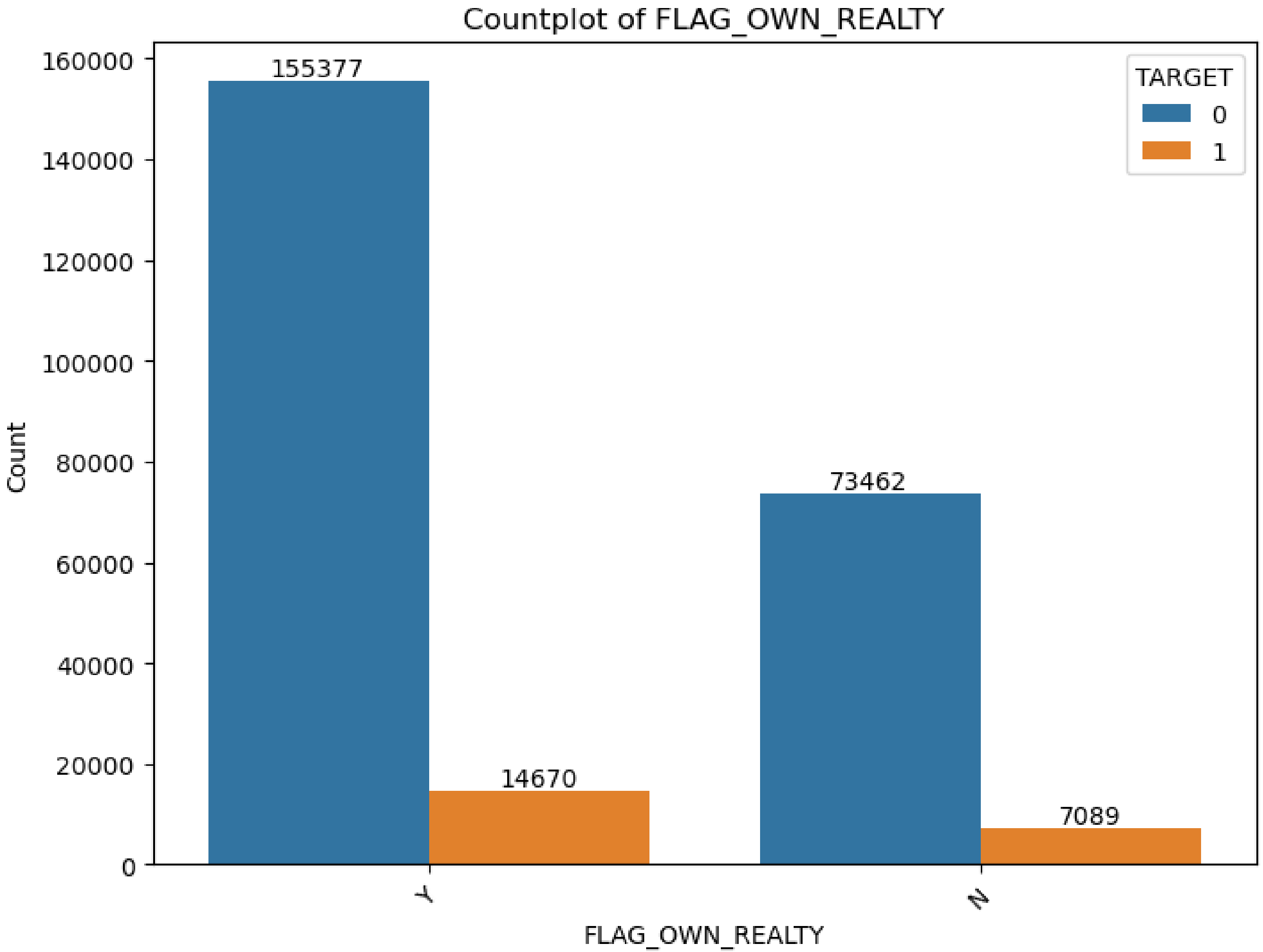


Insight

Upon meticulous examination, it was revealed that the majority of clients, comprising 63%, did not own car. In contrast, clients with own car were distributed more diversely throughout the dataset.

When focusing on defaulters, a fascinating pattern emerged. Among those with no own car, defaulters accounted for 9.3% of the total, while defaulters with revolving loans represented 7.5%.

# Distribution of Clients Owning Reality

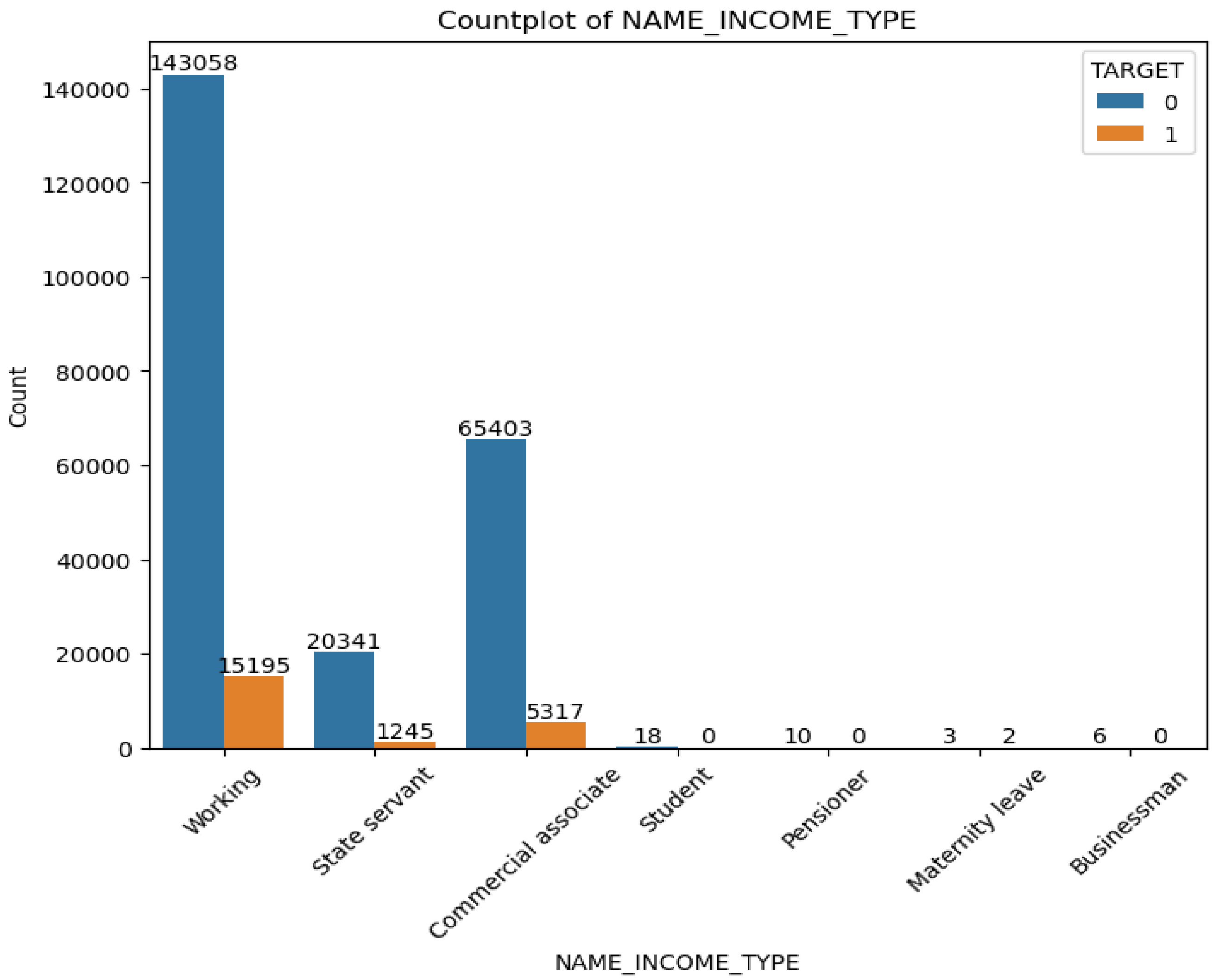


## Insight

Upon meticulous examination, it was revealed that the majority of clients, comprising 68%, own realty. In contrast, clients with no own reality were distributed more diversely throughout the dataset.

When focusing on defaulters, a fascinating pattern emerged. Among those with own reality, defaulters accounted for 8.6% of the total, while defaulters with revolving loans represented 8.8%.

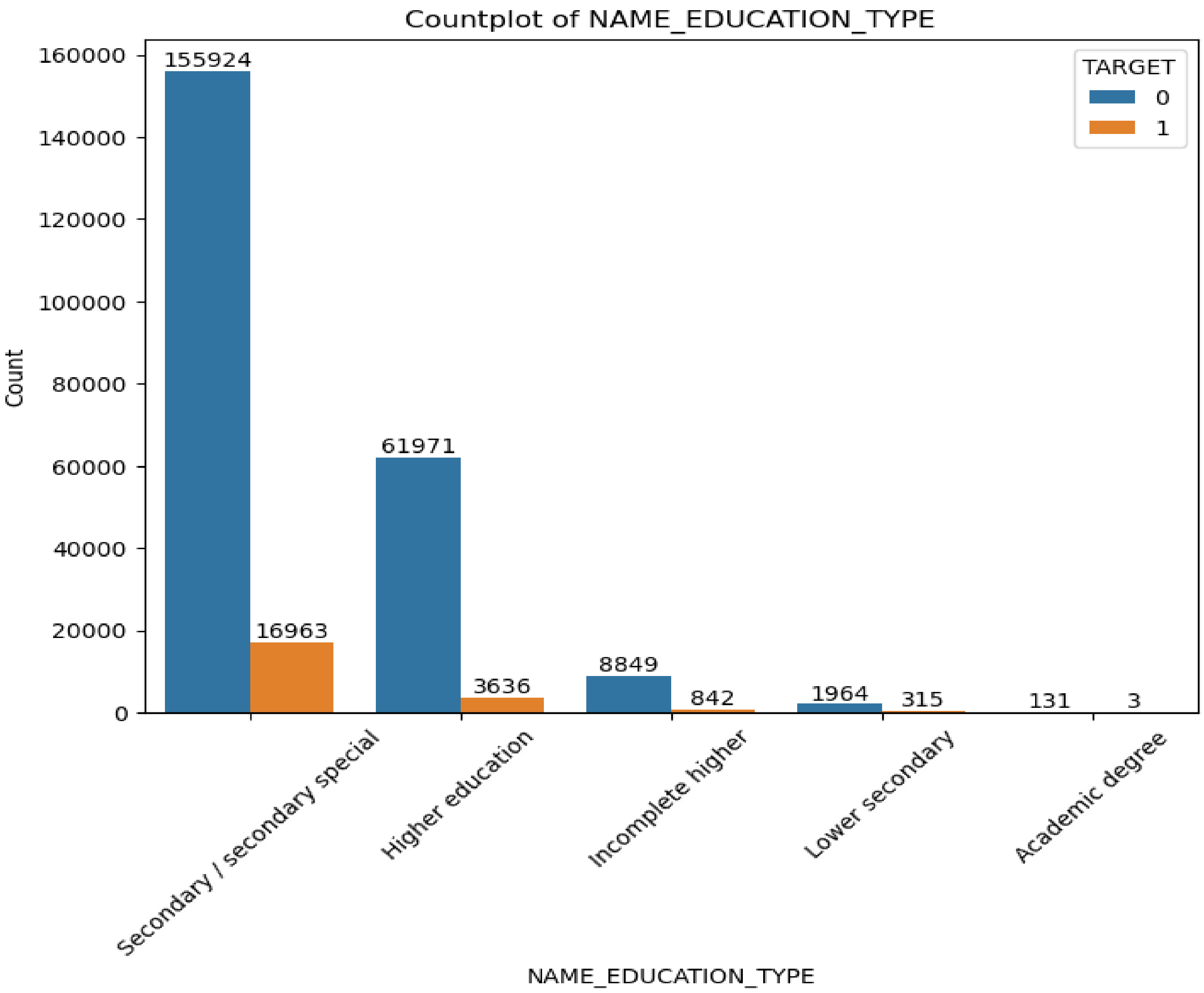
# Distribution of Clients Income Type



Insight

Working account for 63%, state servants 9% and commercial associate account for 28%. State servants have less proportion of defaulters compared to working and commercial associate. Working type clients have 7 % defaulters while 5% defaulters in commercial associate.

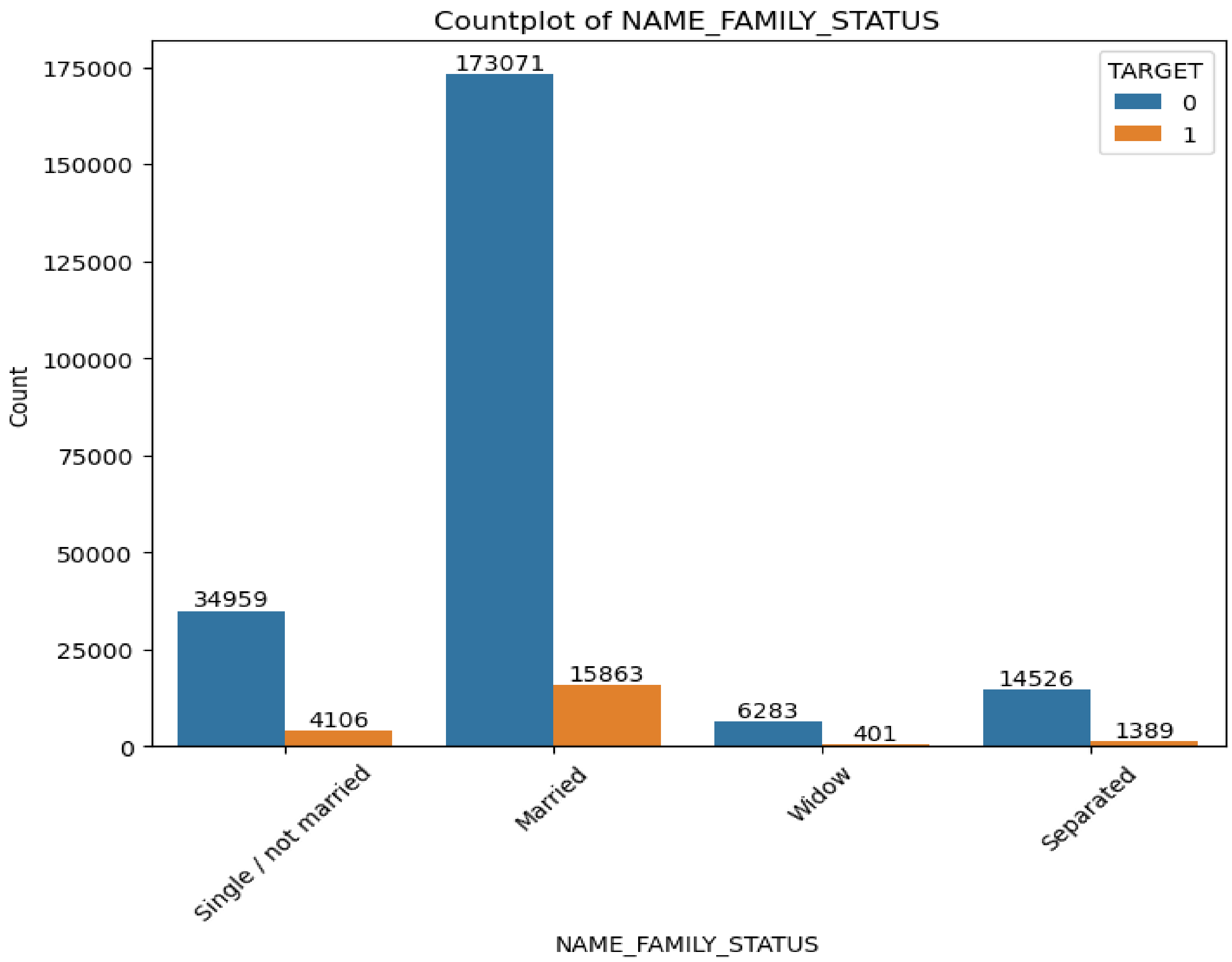
# Distribution of Education Type



Insight

Clients with secondary education, higher education and incomplete higher education have high ratio. While secondary and higher education clients account for 10% defaulters each.

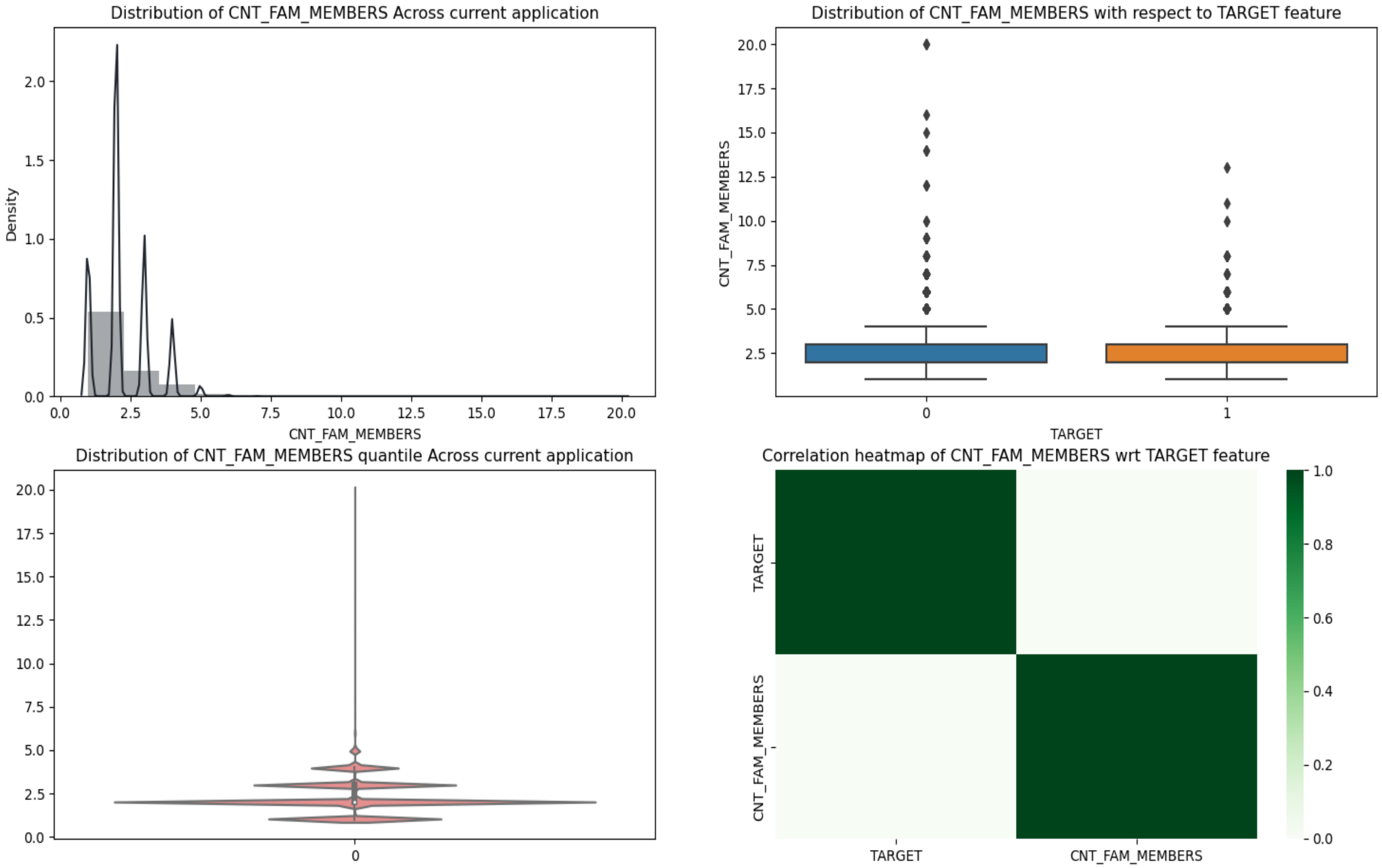
# Distribution of Family Status



## Insight

Married clients account for 75% followed by single clients with 1% distribution and widow account for 6% of total population.
Single client seems to have 3% increased distribution for defaulters.

# Bivariate Analysis



Distribution of CNT_FAM_MEMBERS Across current application

Distribution of CNT_FAM_MEMBERS with respect to TARGET feature

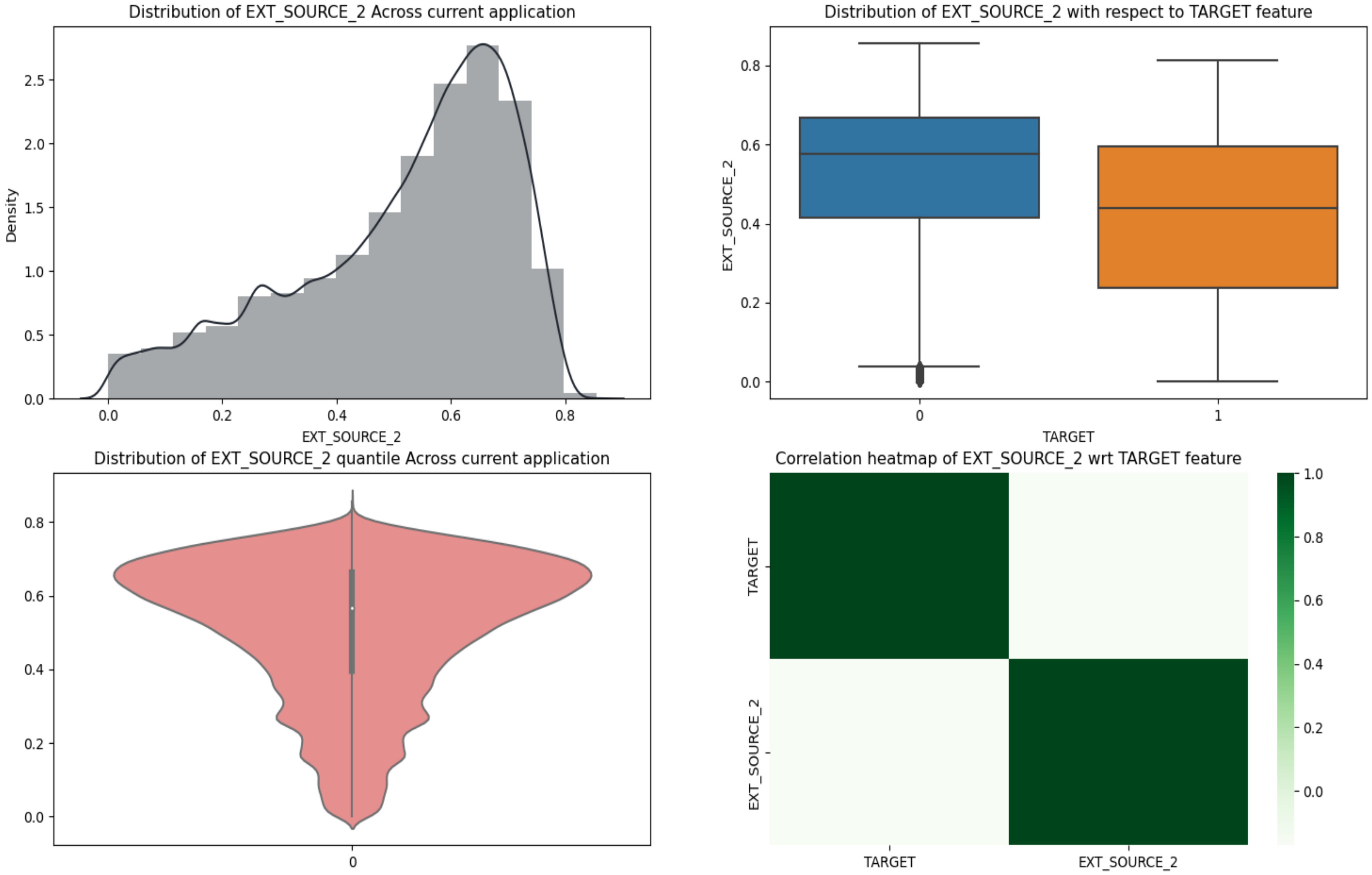
Distribution of CNT_FAM_MEMBERS quantile Across current application

Correlation heatmap of CNT_FAM_MEMBERS wrt TARGET feature

## Insight

In Family members 2 seems to be the most common number, showing most of the client most likely recently married couples. Also considering the trend seen in children, We have peak density around 2,3,4 and 5 members. ( 2 parents + 0,1,2 or 3 kids)

# Distribution of External Source Score_2 of Clients



Distribution of EXT_SOURCE_2 Across current application

Distribution of EXT_SOURCE_2 with respect to TARGET feature

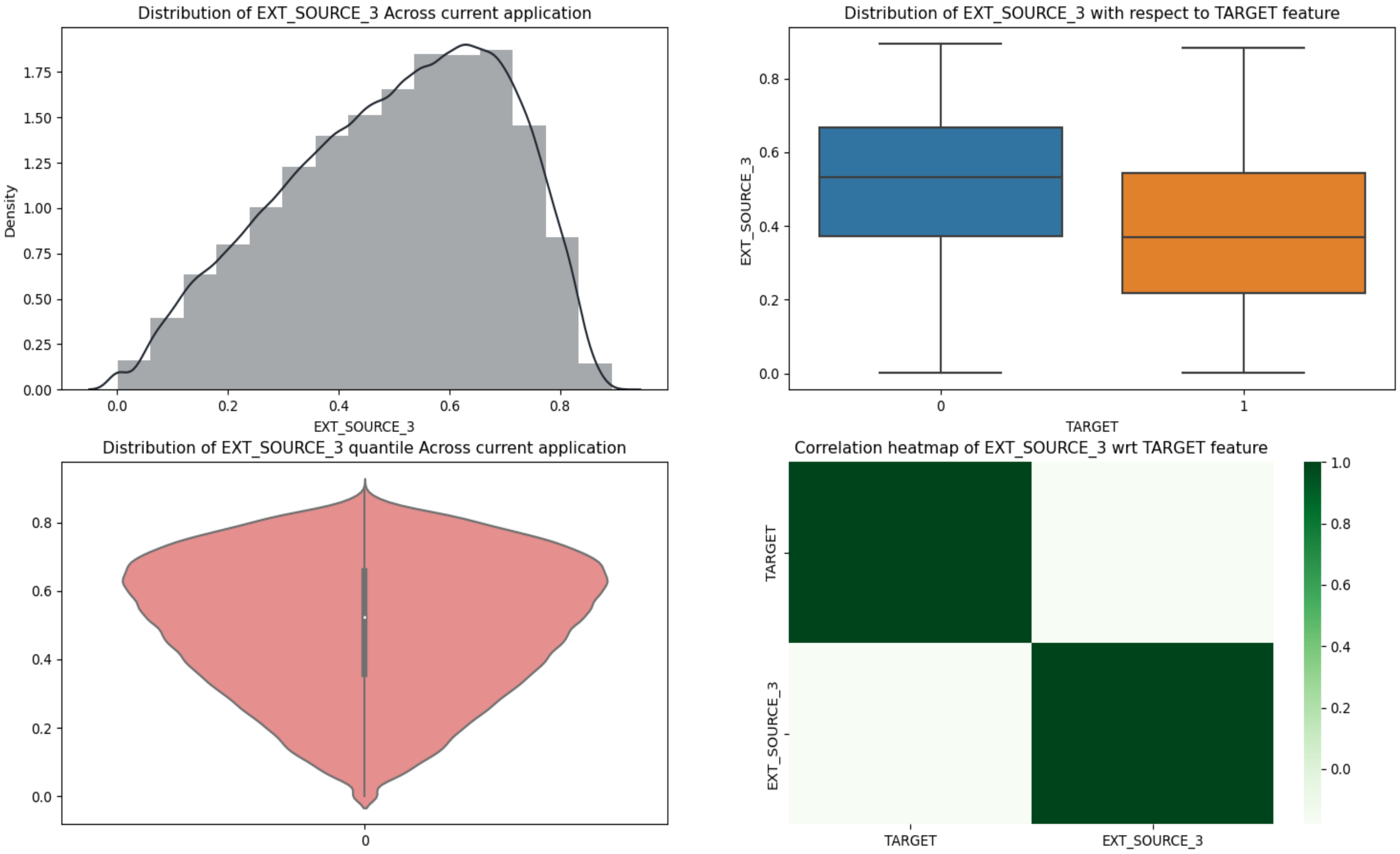
Distribution of EXT_SOURCE_2 quantile Across current application

Correlation heatmap of EXT_SOURCE_2 wrt TARGET feature

## Insight

The median values for EXT_SOURCE_2 seems higher if the TARGET is 0 ( means good client). and median value lower if TARGET is 1 ( bad client ). Clients with low score ranging from 0.2 to 0.4 are more likely to have paying difficulties.

# Distribution of External Source Score_2 of Clients


Distribution of EXT_SOURCE_3 Across current application

Distribution of EXT_SOURCE_3 with respect to TARGET feature

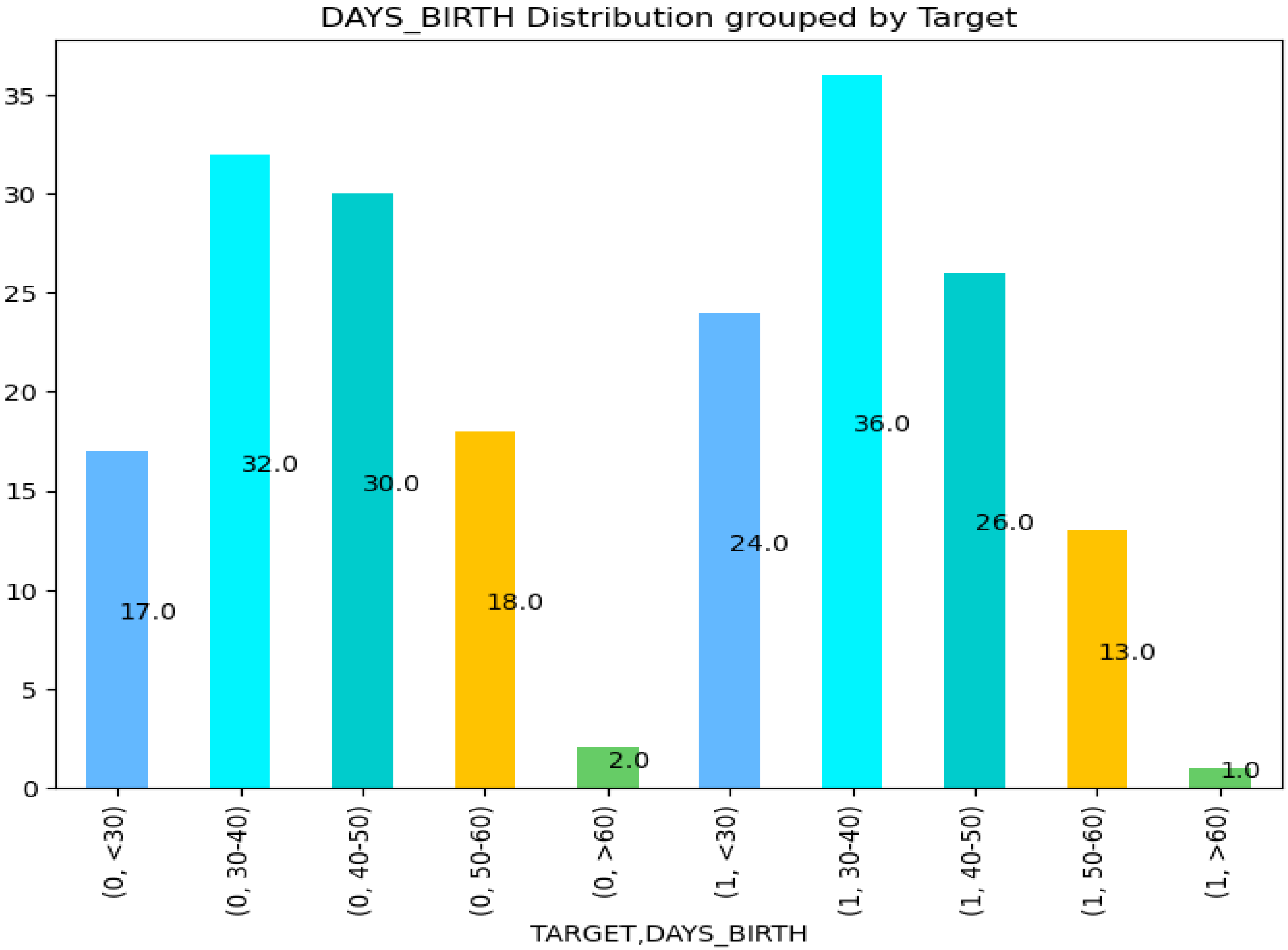
Distribution of EXT_SOURCE_3 quantile Across current application

Correlation heatmap of EXT_SOURCE_3 wrt TARGET feature

## Insight

The median values for EXT_SOURCE_2 seems higher if the TARGET is 0 ( means good client), and median value lower if TARGET is 1 ( bad client ). Clients with low score ranging from 0.2 to 0.4 are more likely to have paying difficulties.

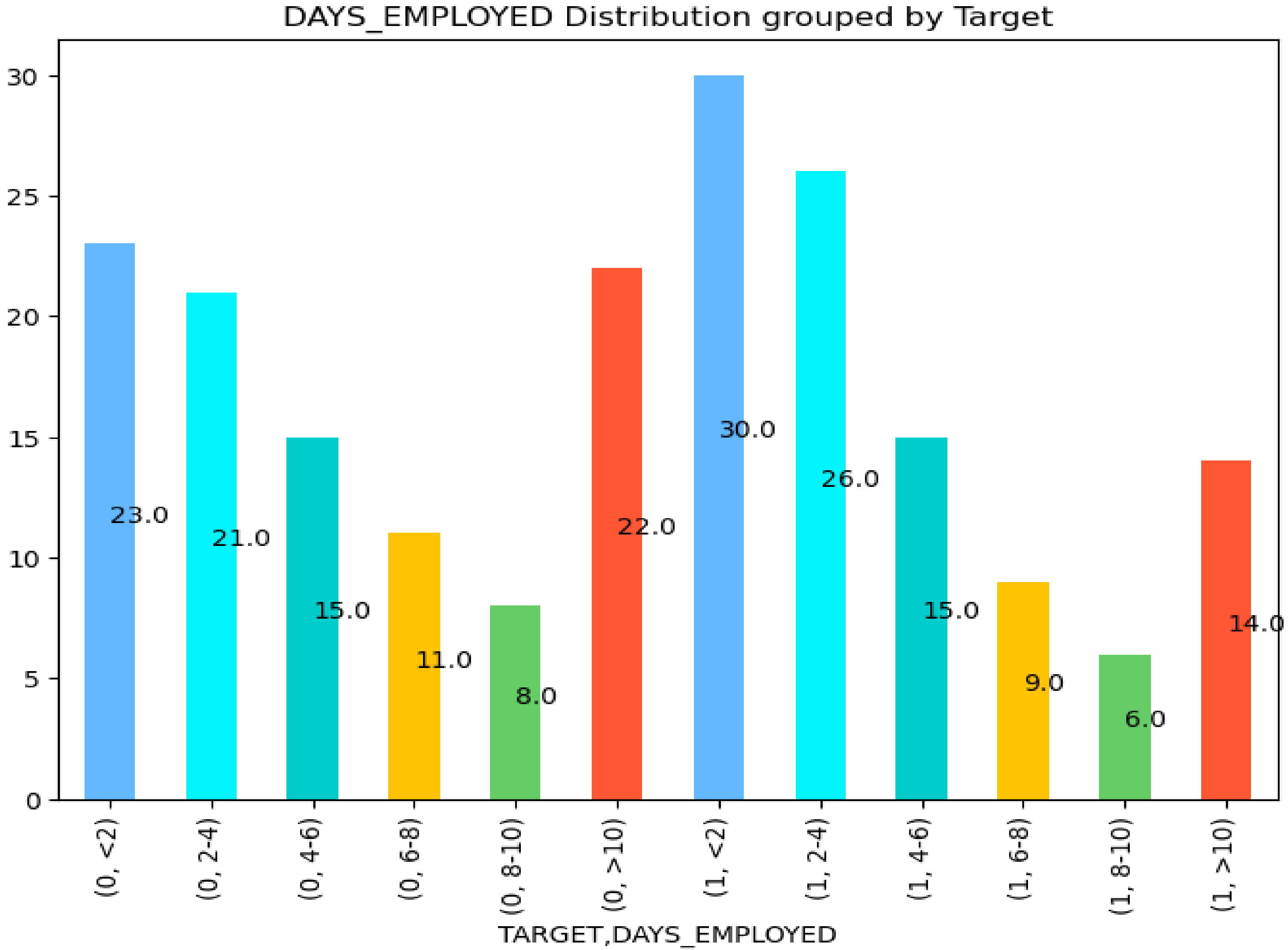
# Distribution of Clients Age Grouped with Target



Insight

- Majority of the clients are actually of the age group 30-50.
- It is also noticed that the clients with less age tend to represent more in TARGET 1 group
  - <30 years went up 6%,
  - 30-40 years increased 4%
  - but 40-50 years went down 4%, 50-60 years by 5%
- 60+ years went half in representation.

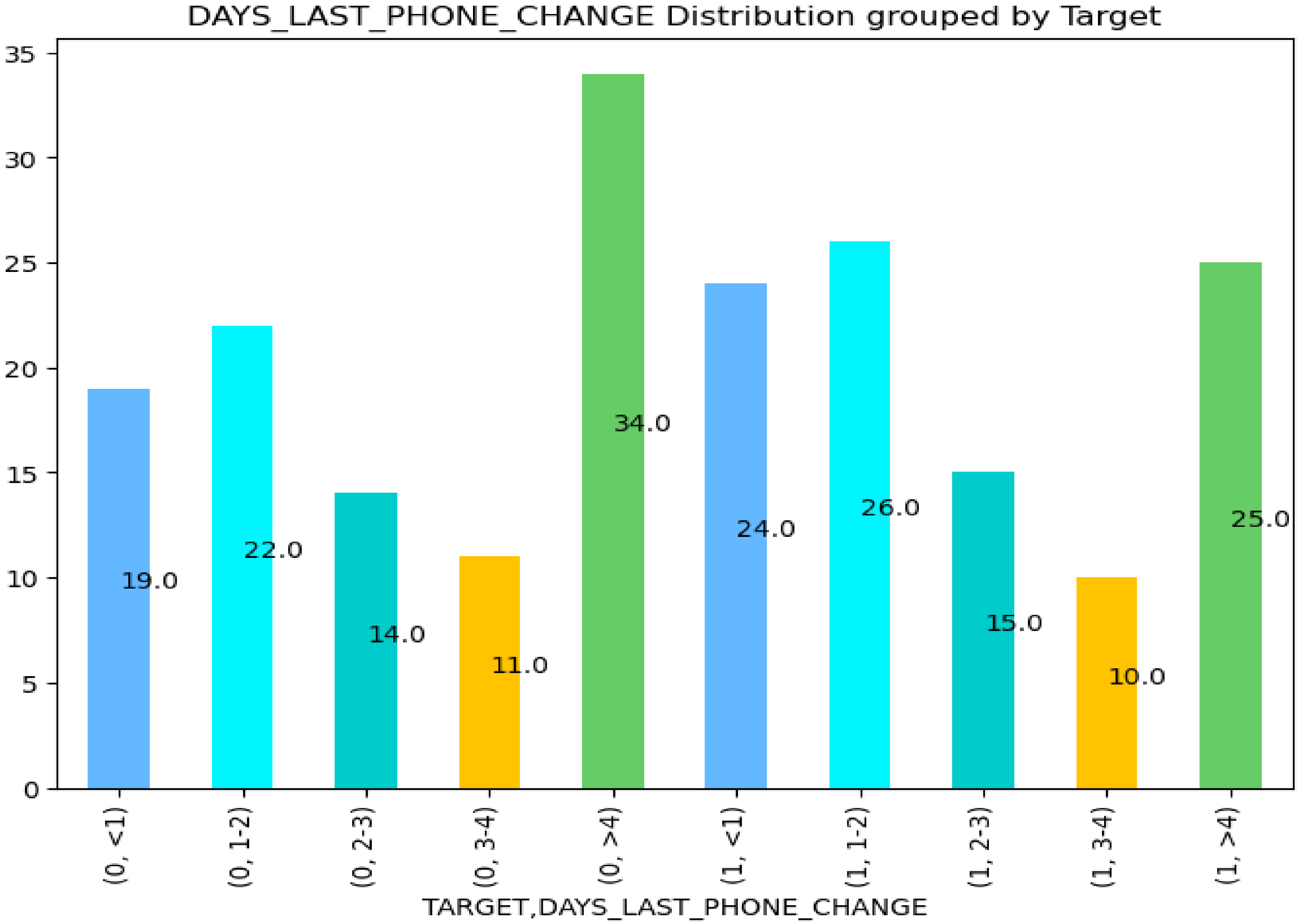
# Distribution of Clients Working Experience Grouped with Target



Insight

- The experience of the employee, which is also highly correlated with age.
- The number of clients reduces as experience increases.
- Similar to age(and highly correlated), clients with less experience tend to represent more in TARGET 1 group:
    - <2 went up 6%, 2-4 years 4%
    - 4-6 years remains the same as 15%
    - As experience grew, 6-8,8-10,10+, the representation in TARGET group 1 reduces even up to 8%.

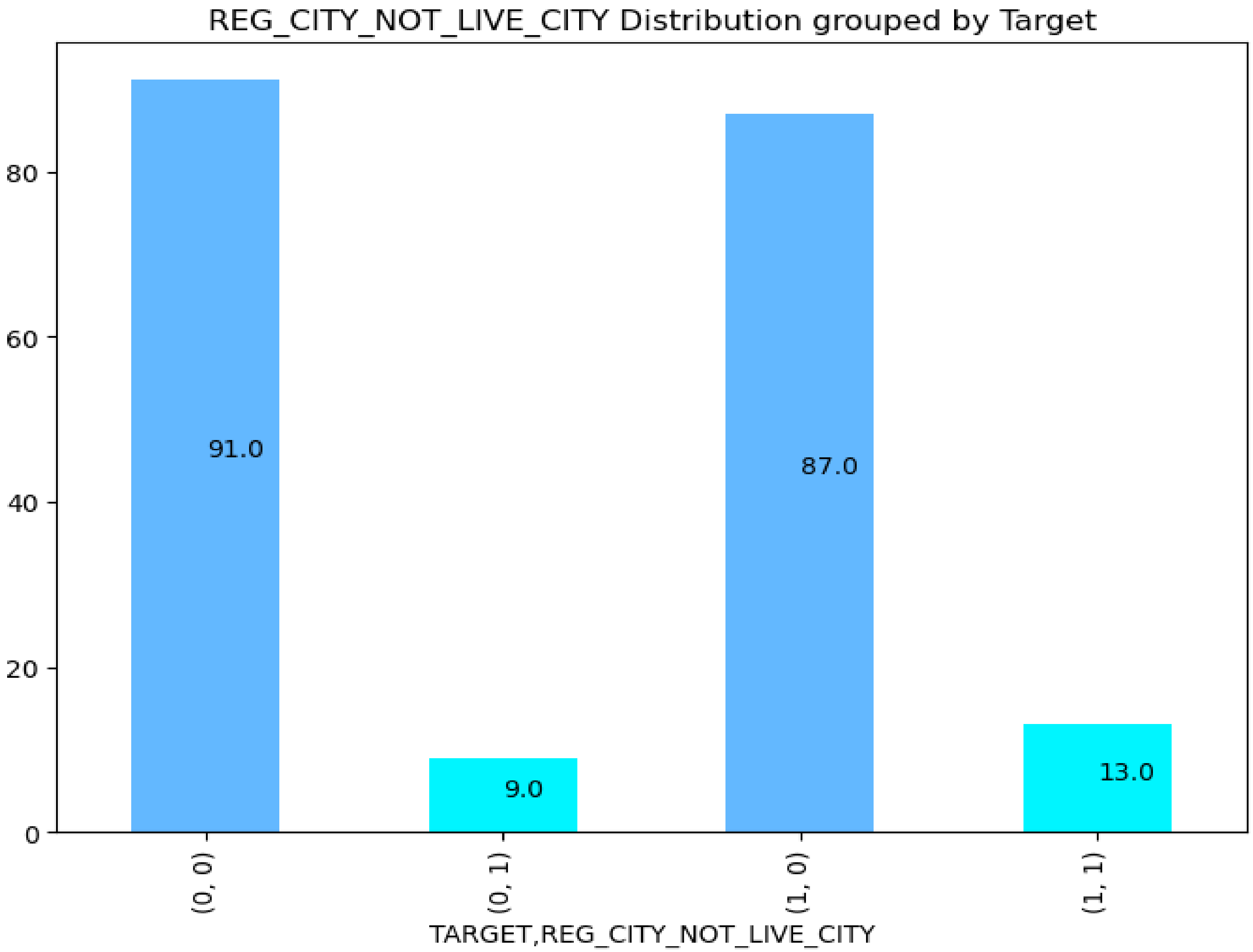
# Distribution of Clients Days Since Last Phone Change Grouped with Target



Insight

- If the Phone is changed within last couple of years there's an increased chance of being a TARGET 1 group client
  - 6% increased risk if phone changed less than a year.
  - 4% if phone changed between last 1-2 years.

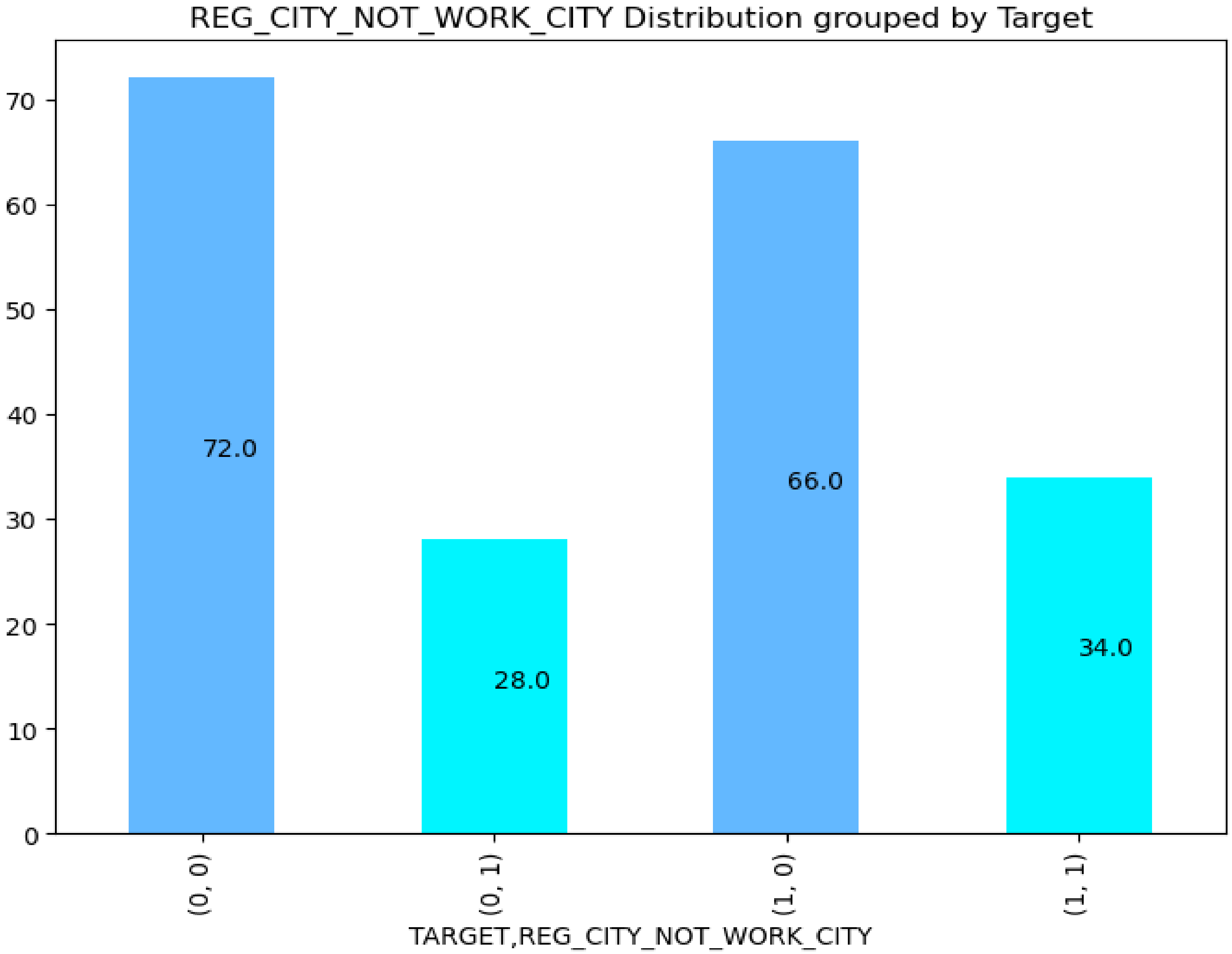
# Distribution of Clients Permeant Address and Contact Address Grouped with Target



Insight

- Clients whose permanent address is different from contact address are more likely to be defaulters.
- When contact address is not the registered address, there is 4% of increased risk of clients default.

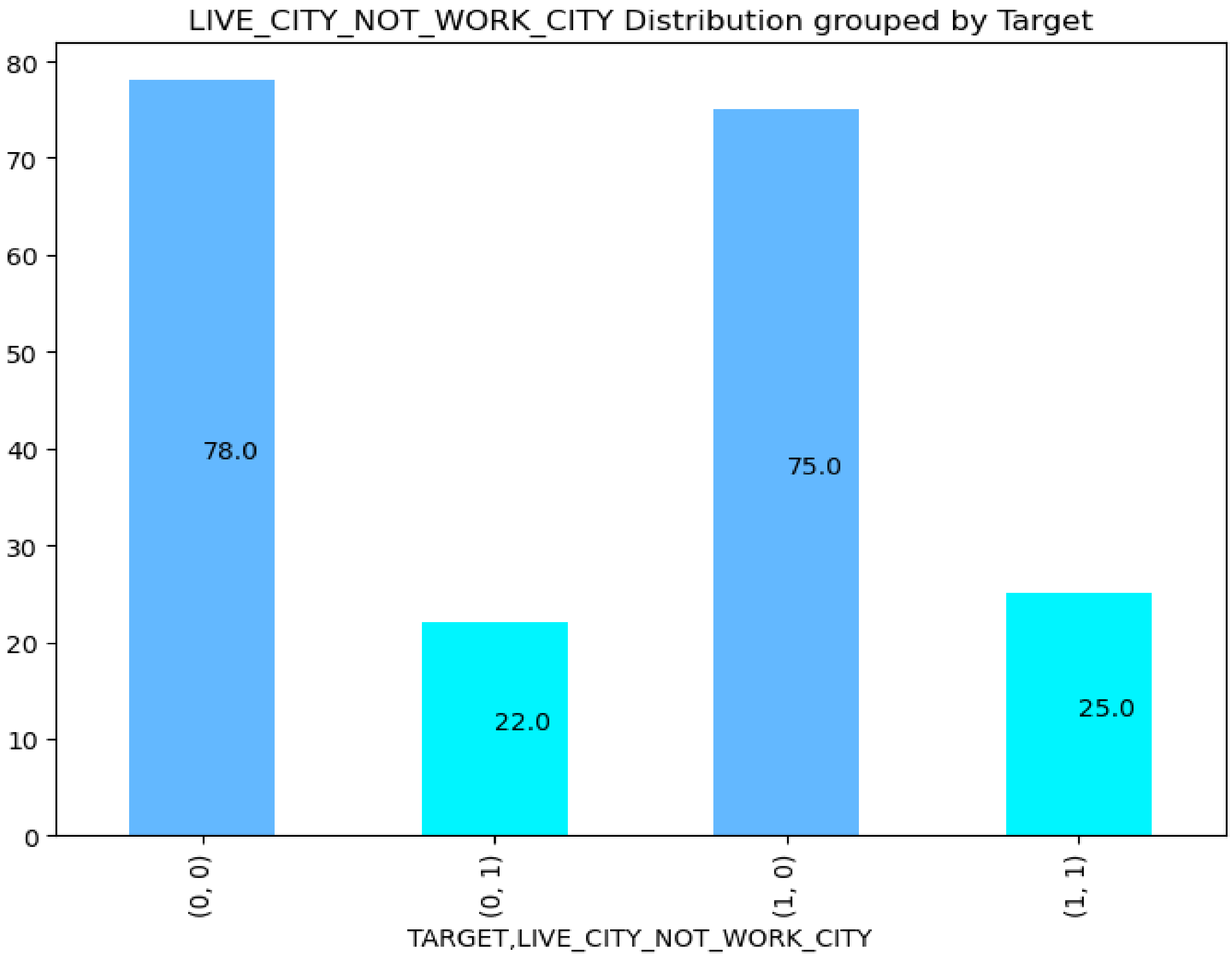
# Distribution of Clients Permeant Address and Work Address Grouped with Target



Insight

- Clients whose permanent address is different from working address are more likely to be defaulters.
- When permanent address is not the working address, there is 6% of increased risk of clients default.

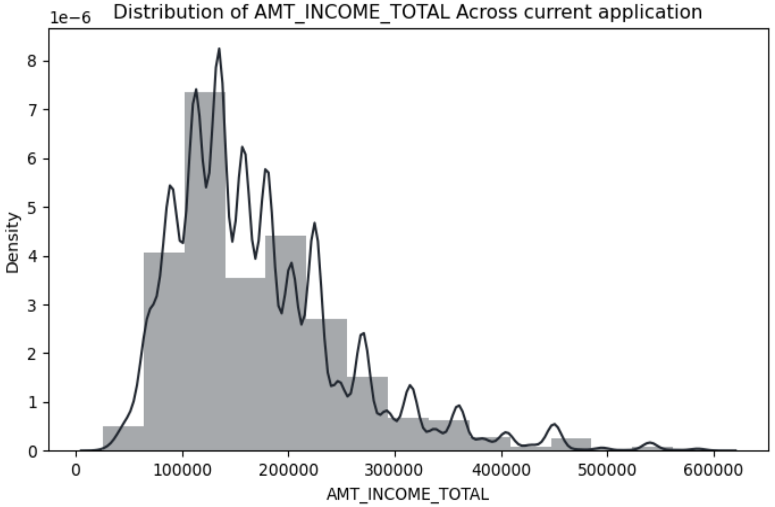
# Distribution of Clients Contact Address and Work Address Grouped with Target

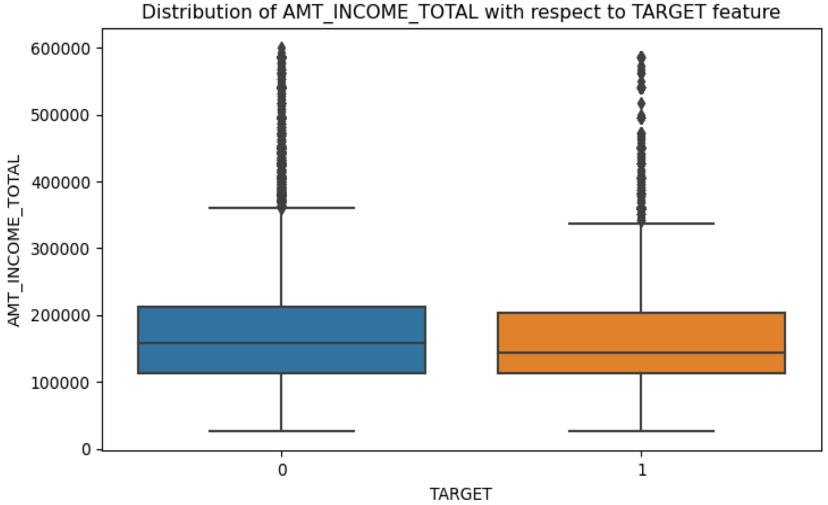


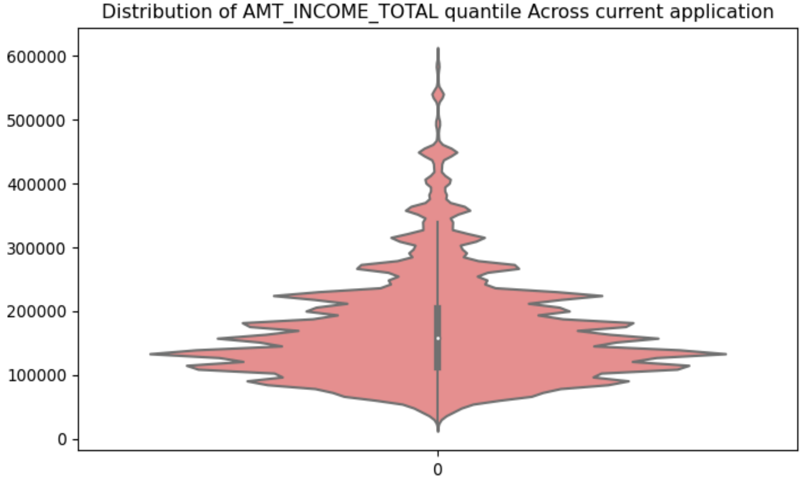
Insight

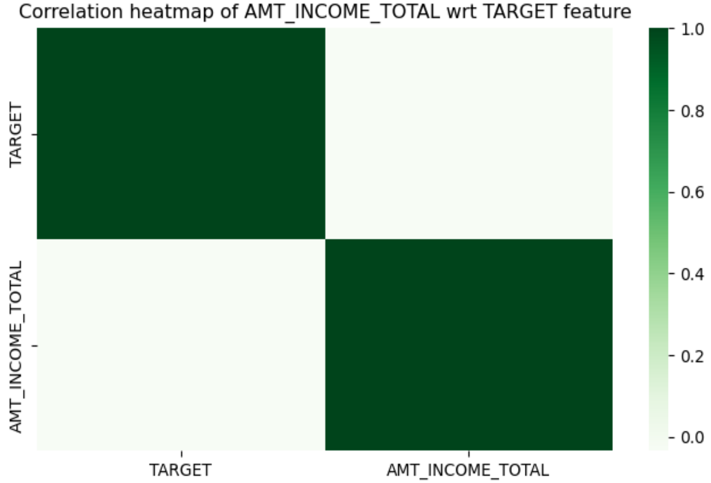
- Clients whose contact address is different from working address are more likely to be defaulters.
- When contact address is not the working address, there is 3% of increased risk of clients default.

# Distribution of Income Amount Grouped with Target


Distribution of AMT_INCOME_TOTAL Across current application


Distribution of AMT_INCOME_TOTAL with respect to TARGET feature

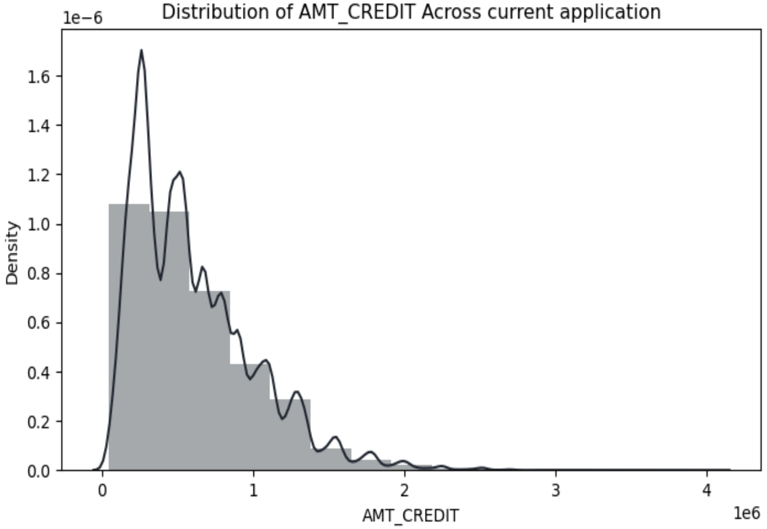

Distribution of AMT_INCOME_TOTAL quantile Across current application

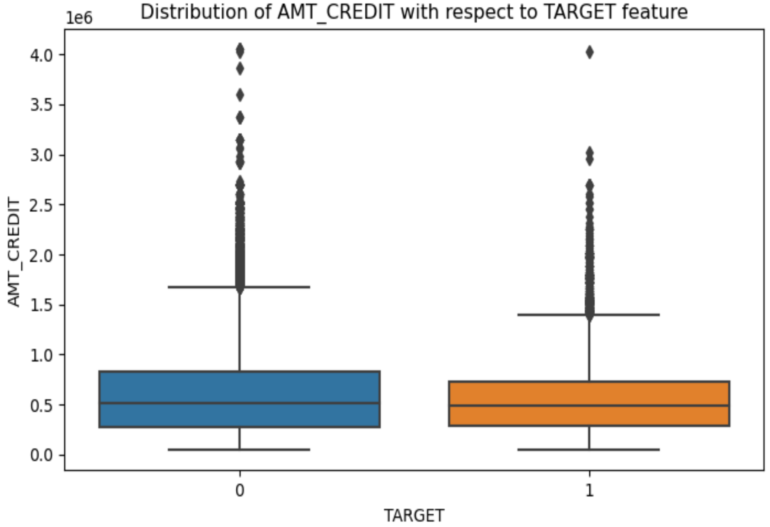

Correlation heatmap of AMT_INCOME_TOTAL wrt TARGET feature

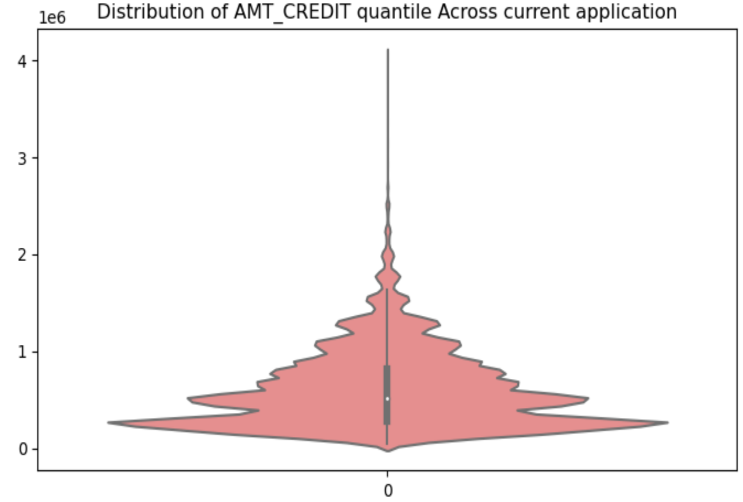
## Insight

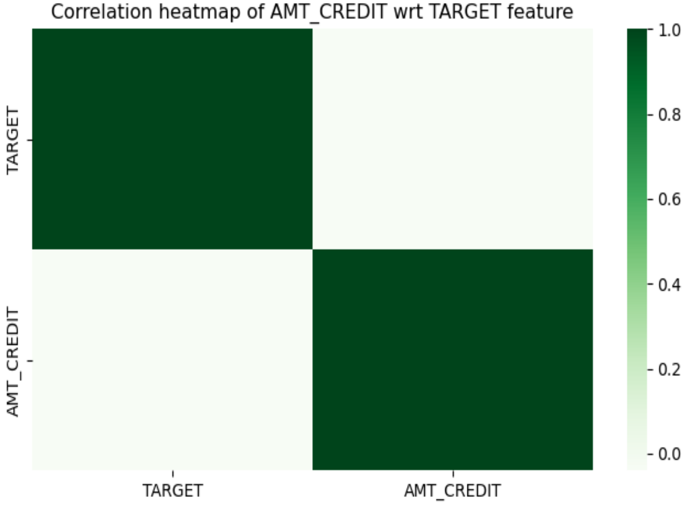
Majority of the client has salary between 100k and 250k Rupees.

# Distribution of Loan Amount Grouped with Target


Distribution of AMT_CREDIT Across current application


Distribution of AMT_CREDIT with respect to TARGET feature

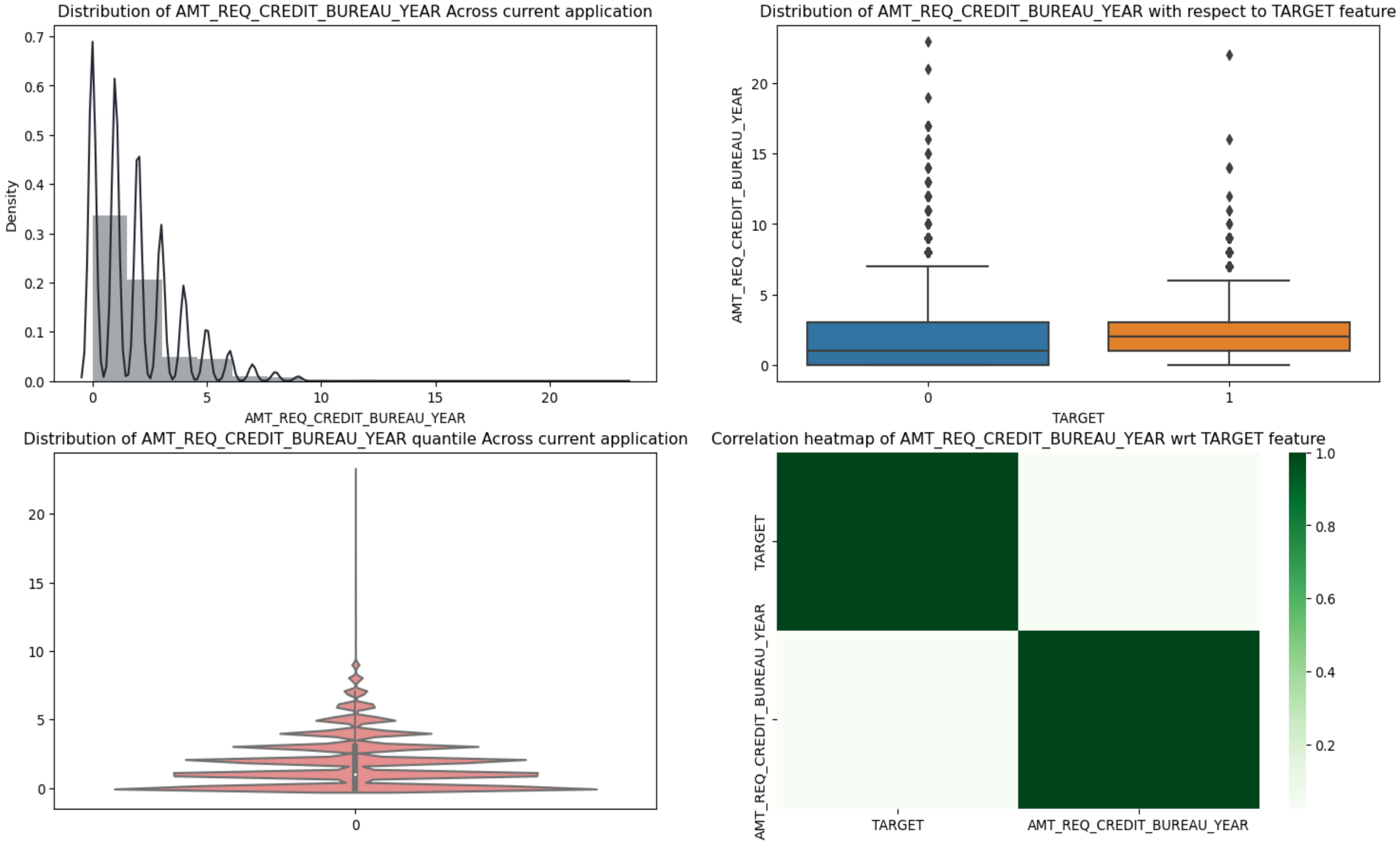

Distribution of AMT_CREDIT quantile Across current application


Correlation heatmap of AMT_CREDIT wrt TARGET feature

## Insight

- Most of the loan amounts are less than 800k Rupees, though we have outliers till 4M Rupees.
- There's a very weak trend in loan amount and the TARGET 1 clients seems to have less loan amount according to boxplot.

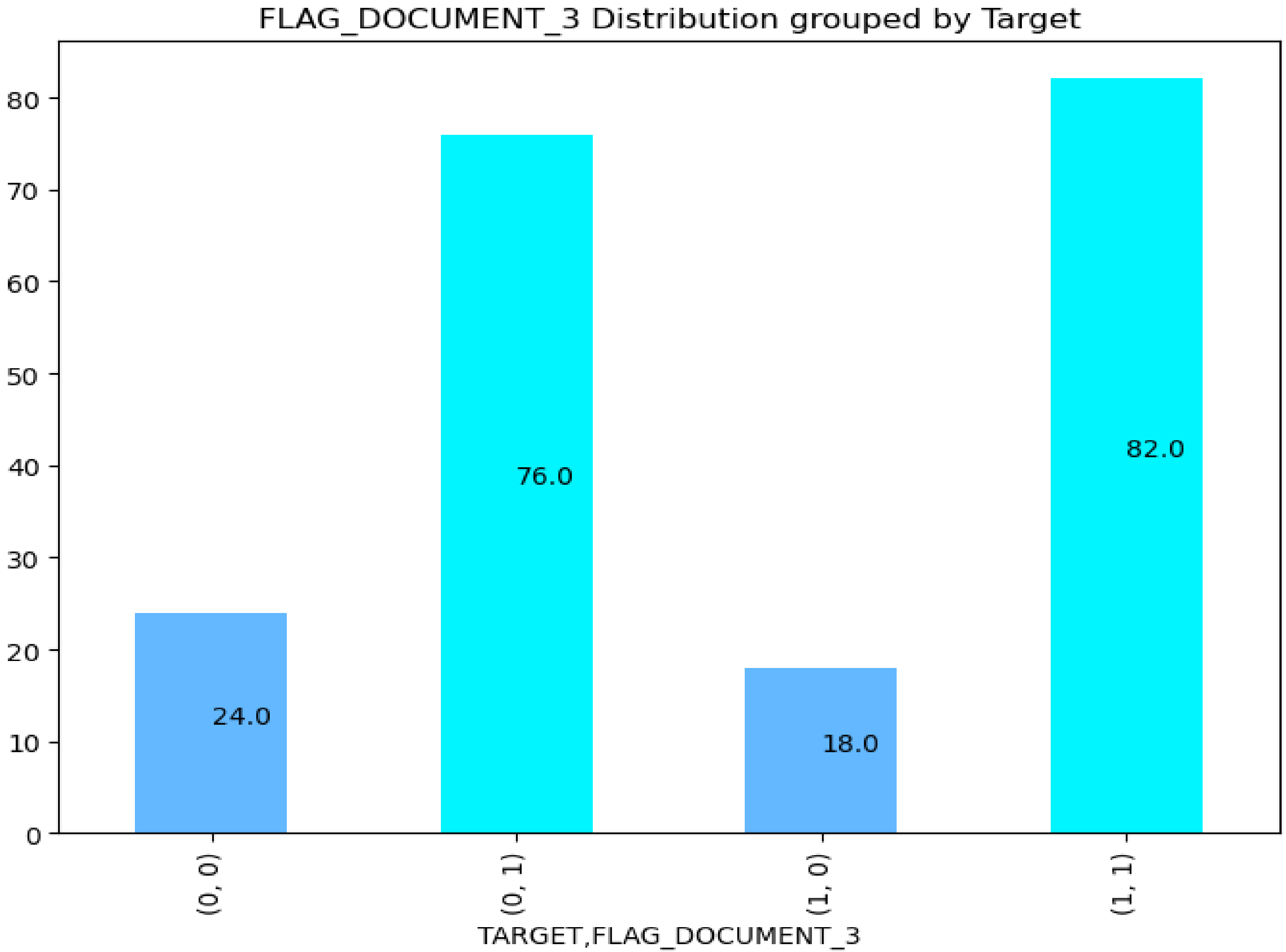
# Distribution of Number of Enquiries to Credit Bureau Grouped with Target



## Insight

- In 1 year span, the TARGET 1, seem to have higher minimum and 25% as compared to TARGET 0.
- Due to the same median also seem to be high.

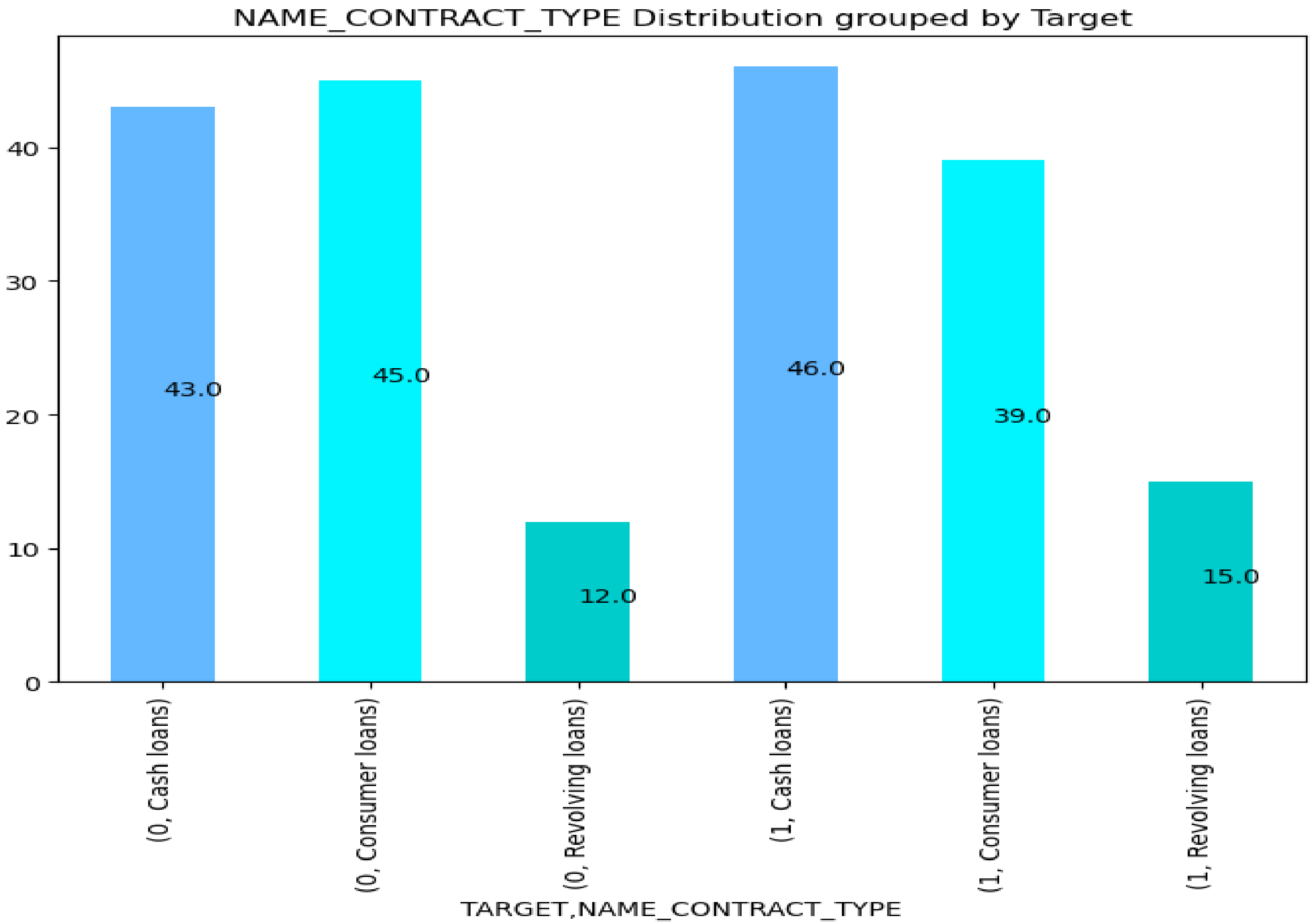
# Distribution of Clients Who Provided Document 3 Grouped with Target



Insight

- Clients who had provided Document 3 are 6% greater in TARGET 1 group as compared to clients who had not provided the document 3.

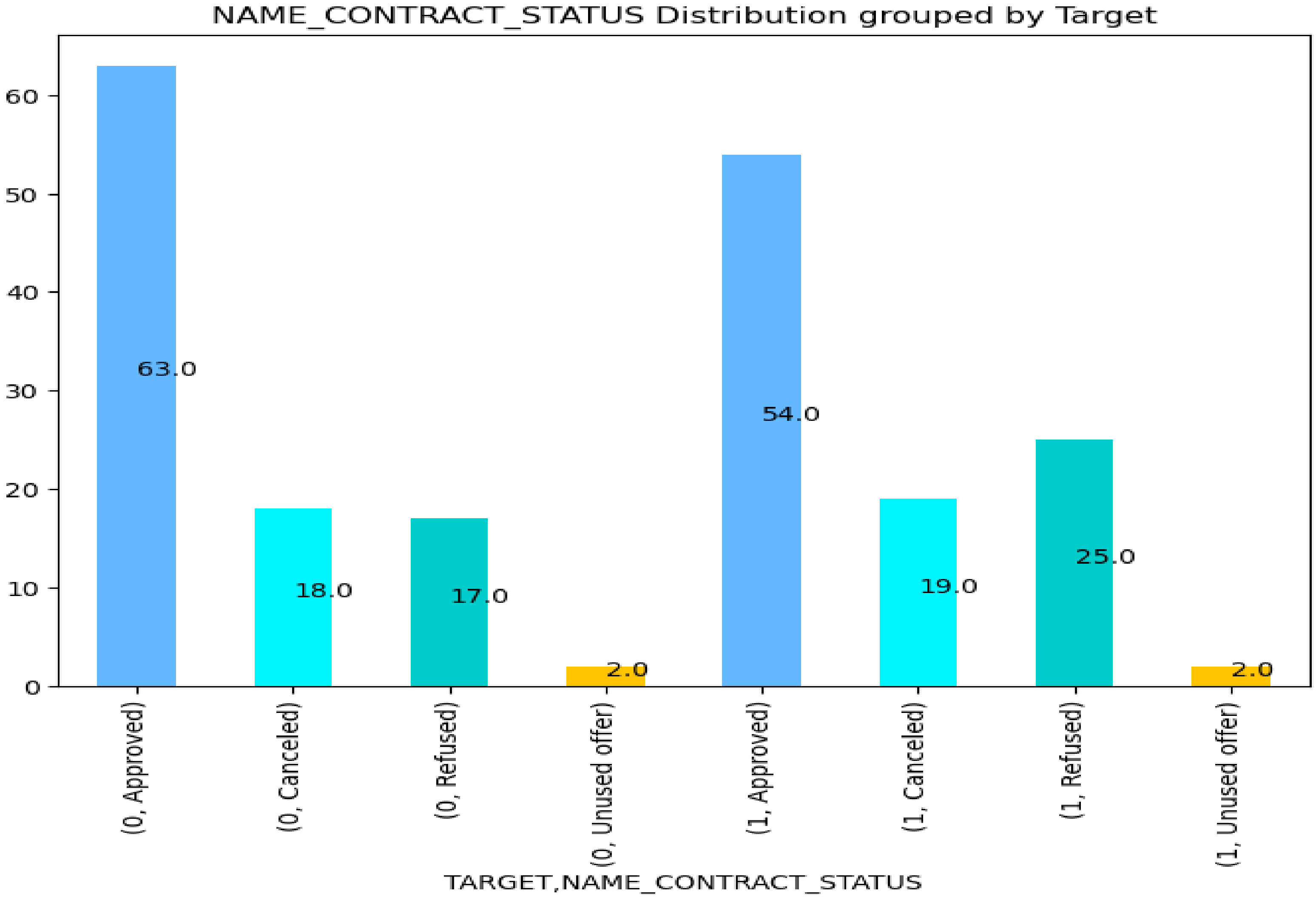
# Distribution of Contract Type of Loan Grouped with Target



## Insight

- Cash loans and Revolving roles have slightly increased risk of having TARGET 1 group.
- Clients who had taken consumer loans have less percentage of defaulters.

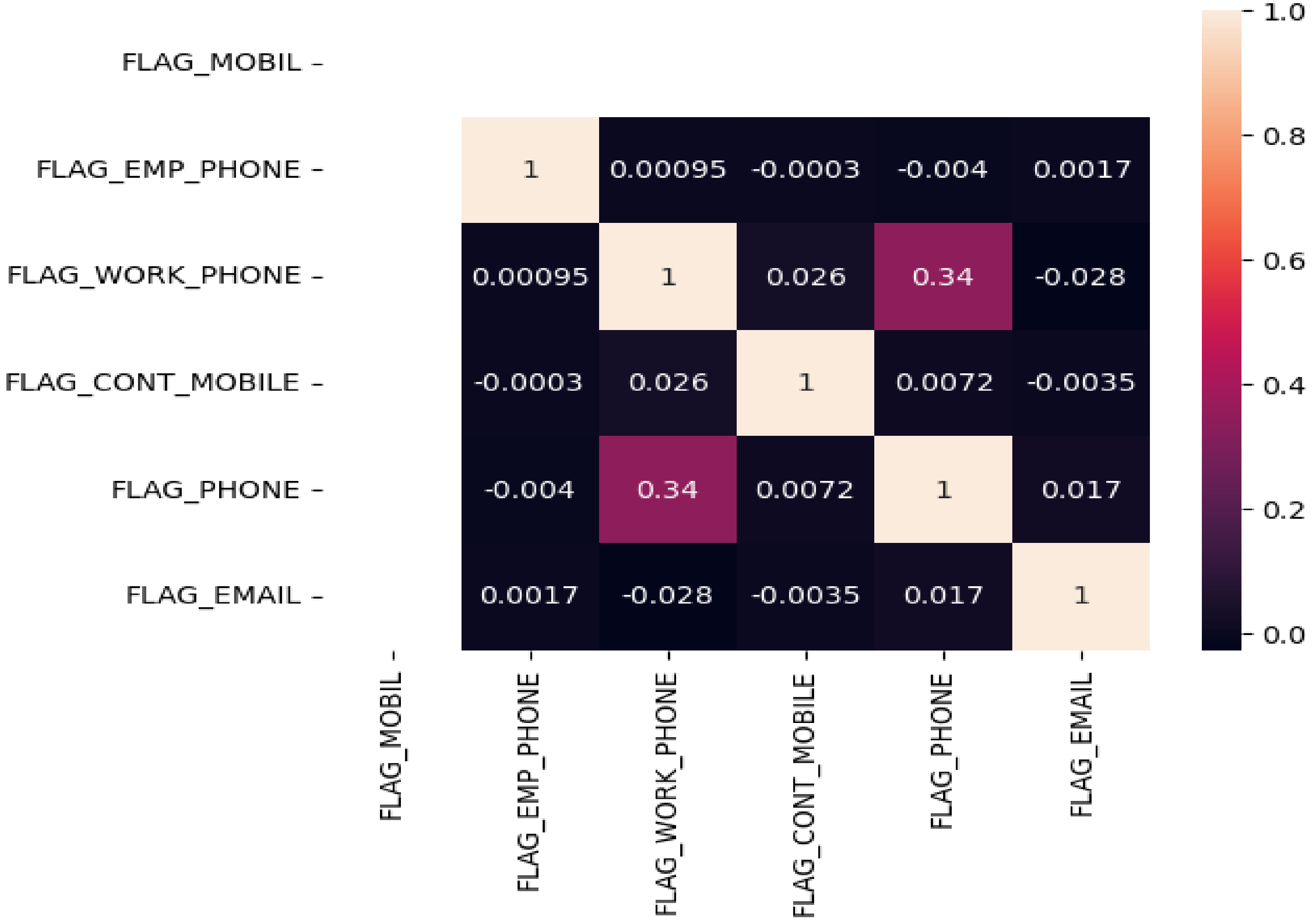
# Distribution of Clients Previous Contract Status Grouped with Target



Insight

Previously refused loan seems to have 8% more increased risk, since the previous refusal might be due to being in TARGET group 1 already.

# Correlation Plot  of Flag Variables

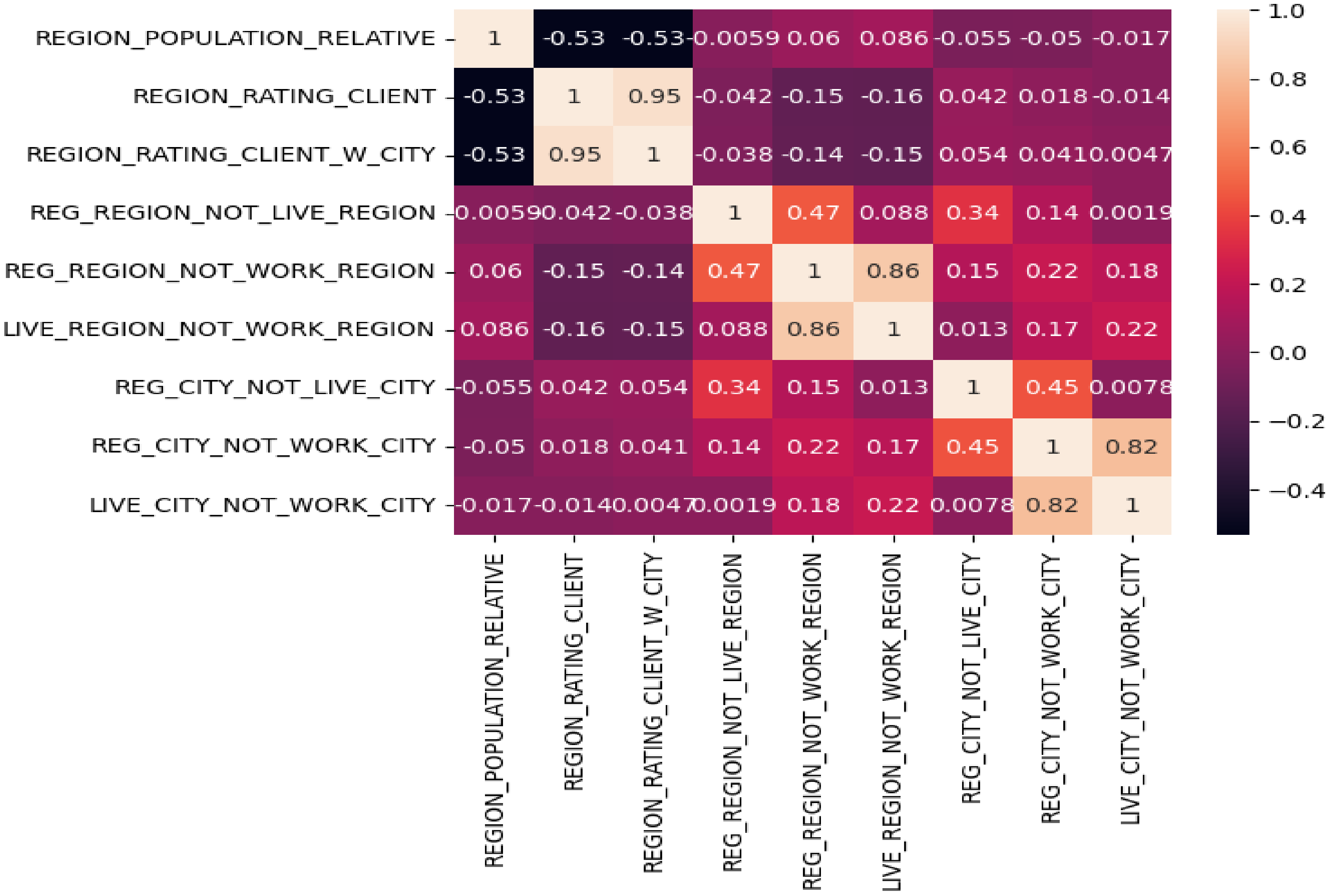


| | FLAG_MOBIL | FLAG_EMP_PHONE | FLAG_WORK_PHONE | FLAG_CONT_MOBILE | FLAG_PHONE | FLAG_EMAIL |
|---|---|---|---|---|---|---|
| FLAG_MOBIL | | | | | | |
| FLAG_EMP_PHONE | | 1 | 0.00095 | -0.0003 | -0.004 | 0.0017 |
| FLAG_WORK_PHONE | | 0.00095 | 1 | 0.026 | 0.34 | -0.028 |
| FLAG_CONT_MOBILE | | -0.0003 | 0.026 | 1 | 0.0072 | -0.0035 |
| FLAG_PHONE | | -0.004 | 0.34 | 0.0072 | 1 | 0.017 |
| FLAG_EMAIL | | 0.0017 | -0.028 | -0.0035 | 0.017 | 1 |

Insight

A fascinating correlation has been discovered between clients who have provided their home phone numbers. It appears that the presence of a home phone number among clients is positively correlated with the likelihood of other clients also providing their home phone numbers

# Correlation Plot  of Flag Variables

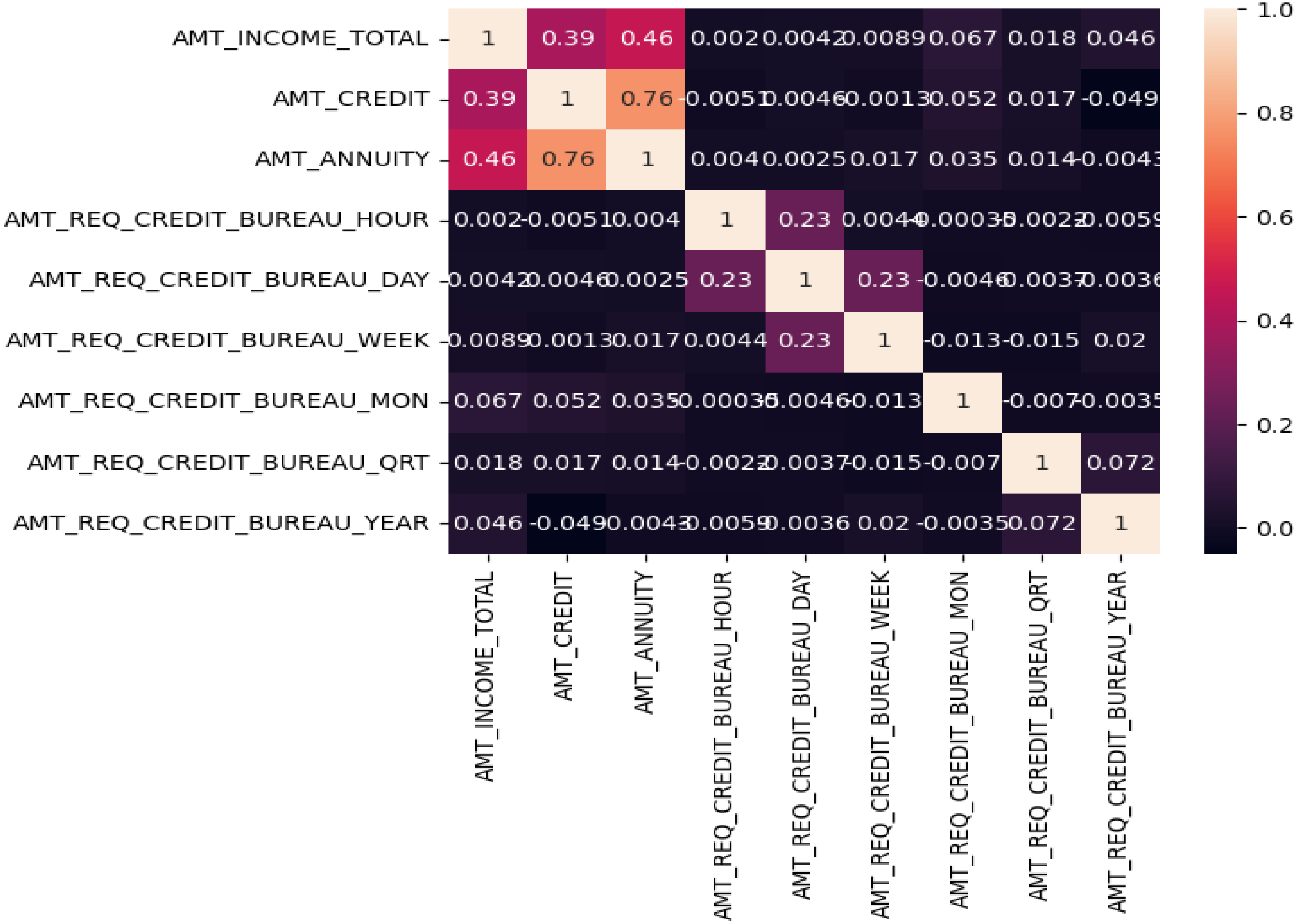


| | REGION_POPULATION_RELATIVE | REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | REG_REGION_NOT_LIVE_REGION | REG_REGION_NOT_WORK_REGION | LIVE_REGION_NOT_WORK_REGION | REG_CITY_NOT_LIVE_CITY | REG_CITY_NOT_WORK_CITY | LIVE_CITY_NOT_WORK_CITY |
|---|---|---|---|---|---|---|---|---|---|
| REGION_POPULATION_RELATIVE | 1 | -0.53 | -0.53 | -0.0059 | 0.06 | 0.086 | -0.055 | -0.05 | -0.017 |
| REGION_RATING_CLIENT | -0.53 | 1 | 0.95 | -0.042 | -0.15 | -0.16 | 0.042 | 0.018 | -0.014 |
| REGION_RATING_CLIENT_W_CITY | -0.53 | 0.95 | 1 | -0.038 | -0.14 | -0.15 | 0.054 | 0.041 | 0.0047 |
| REG_REGION_NOT_LIVE_REGION | -0.0059 | 0.042 | -0.038 | 1 | 0.47 | 0.088 | 0.34 | 0.14 | 0.0019 |
| REG_REGION_NOT_WORK_REGION | 0.06 | -0.15 | -0.14 | 0.47 | 1 | 0.86 | 0.15 | 0.22 | 0.18 |
| LIVE_REGION_NOT_WORK_REGION | 0.086 | -0.16 | -0.15 | 0.088 | 0.86 | 1 | 0.013 | 0.17 | 0.22 |
| REG_CITY_NOT_LIVE_CITY | -0.055 | 0.042 | 0.054 | 0.34 | 0.15 | 0.013 | 1 | 0.45 | 0.0078 |
| REG_CITY_NOT_WORK_CITY | -0.05 | 0.018 | 0.041 | 0.14 | 0.22 | 0.17 | 0.45 | 1 | 0.82 |
| LIVE_CITY_NOT_WORK_CITY | -0.017 | -0.014 | 0.0047 | 0.0019 | 0.18 | 0.22 | 0.0078 | 0.82 | 1 |

Insight

- Rating of Clients registered city is highly correlated with clients living region.
- Clients who have different permanent address and contact address are somewhat corelated to clients whose permanent address is not living address.
- Positive relation seen in clients having different permeant address and living address with clients whose permanent address in not working address.

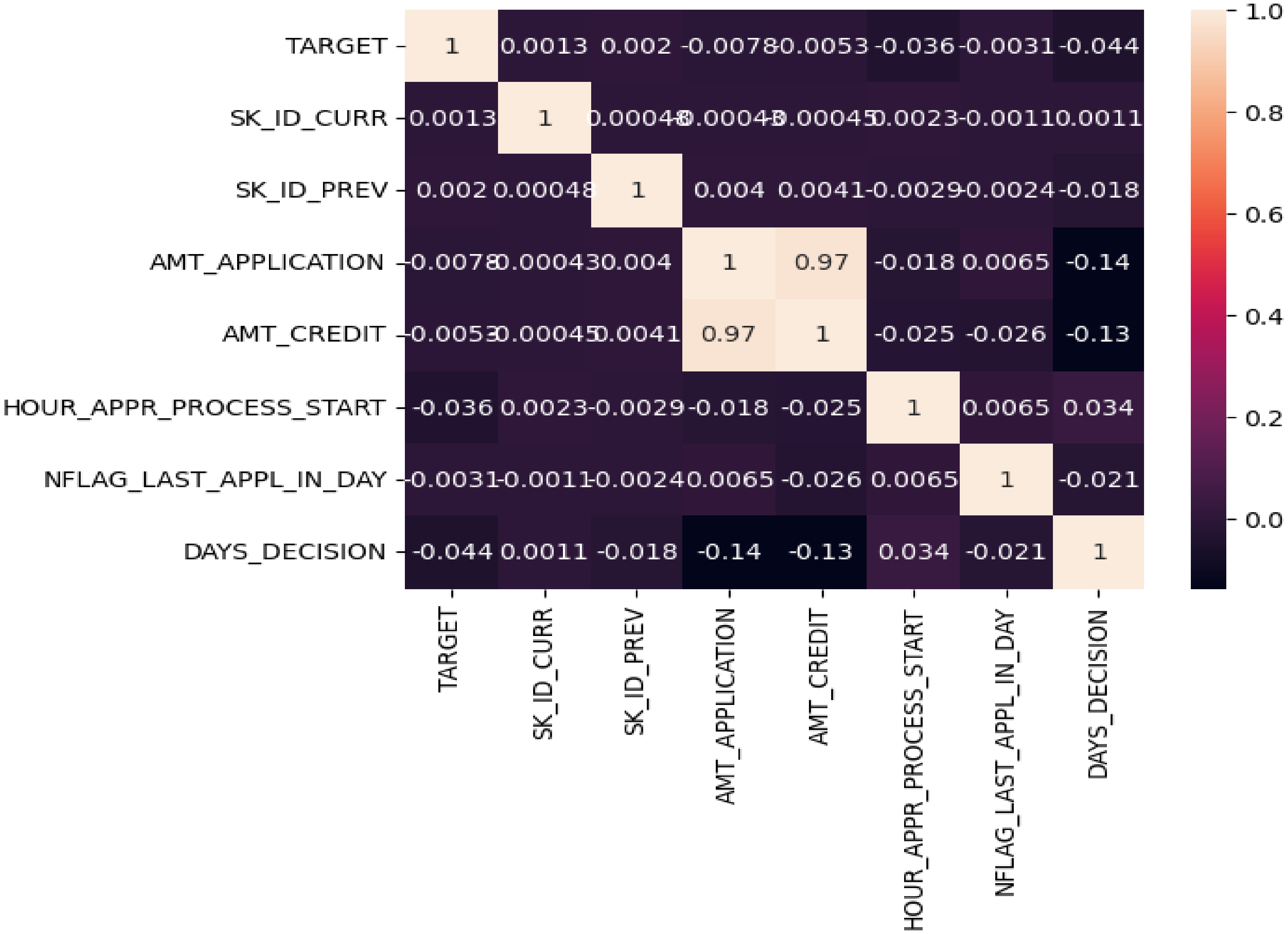
# Correlation Plot of Amount Variables



## Insight

- Loan amount and annuity amount are highly positively correlated.
- Clients Income amount is also positively related to loan amount.
- Some slight correlation seen between clients who have enquired about loan a day before and clients who have enquired about loan a hour before application.

# Correlation Plot  of Previous data of Clients



Insight

- Final credit amount on the last application is highly positively correlated with amount of credit asked by client on previous application.

# Analysis

Exploratory Data Analysis showcased some interesting patterns and trends which differentiated clients with default payment history and al other cases. Our data showed following inferences:

- 92% of clients are good applicants while 8% of clients are faulty applicants.
- In Gender, we have 7% of increased risk in Men.
- In Contract type, The Cash Loan seems to have 3% more risk of having defaulters.
- in Education, people with Secondary education as highest education has 9% increased risk.
- Single people seem to have 3% increased risk of being a defaulter.
- Laborers also seems to have 5% increased risk.

- Majority of the clients are actually of the age group 30-50.
- It is also noticed that the people with less age tend to represent more in TARGET 1 group.
- Similar to age(and highly correlated), clients with less experience tend to represent more in TARGET 1 group.
- If the Phone is changed within last couple of years there's an increased chance of being a TARGET 1 group client as 6% increased risk seen if phone changed less than a year.
- Previously refused loan seems to have 8% more increased risk.
- If the EXT_SOURCE_2 and and EXT_SOURCE_3 scores are lower, client tend to be riskier.

- Owning a home or property seem to reduce the risk by 1%,
- Majority of the client has salary between 100k and 250k Rupees.
- Most of the loan amounts are less than 800k Rupees, though we have outliers till 4M Rupees.
- Final credit amount on the last application is highly positively correlated with amount of credit asked by client on previous application.
- Loan amount and annuity amount are highly positively correlated also clients Income amount is also positively related to loan amount.
- Rating of Clients registered city is highly correlated with clients living region.

# Conclusion

- Low income, less experienced young adult should have more risk due to lack of knowledge on importance of credit scores
- Low income causing unexpected credits, which later people tend to struggle to pay back.
- Single people most likely to be young adults, who also likely take immature decision on personal finance. The same applies to directly if age is low or experience is less.
- Men seems to be responsible for family income and responsibilities, Also men tend to be in majority of the low income jobs. We were expecting to have more defaults due to the low income challenges.
- More kids and family member, might cause increased risk of being defaulter, but on the other hands if the family members are adults and earning income it should reduce the risk.