CREDIT EDA CASE STUDY

By Rajat Sachan

Introduction

This assignment focuses on applying EDA (Exploratory Data Analysis) in a practical business context. It provides insight into risk analytics within the banking and financial services sector, showcasing how data analysis is used to mitigate financial lending risks. Through EDA techniques, you'll gain valuable skills while understanding how data plays a crucial role in minimizing potential losses when lending to customers.

Business Understanding

- In the context of a consumer finance company, we aim to analyze loan application data to identify patterns in loan defaults. The company faces two key risks: approving loans to likely repayers (avoiding loss of business) and to likely defaulters (resulting in financial loss).
- The data distinguishes clients with payment difficulties and those who pay on time. There are four possible outcomes: approved, canceled, refused, and unused offers. Our goal is to use EDA to understand how client and loan attributes influence loan default tendencies, aiding the company in better lending decisions.

Business Objectives

- This case study's objective is to detect patterns indicating clients' payment difficulties, enabling actions like loan denial, reduced loan amounts, or higher interest rates for risky applicants. The goal is to ensure deserving loan applicants are not rejected.
- In simpler terms, the company aims to identify key factors (driver variables) leading to loan defaults—strong indicators of default. This knowledge aids portfolio management and risk assessment. To better grasp the domain, it's advisable to conduct some independent research on risk analytics, focusing on variable types and their importance.

Workflow Walkthrough

- Data Understanding
- Data Cleaning and Manipulation
- Handling Outlier
- Addressing Data Imbalance
- Analyzing Data: Univariate, Segmented, Bivariate, and Correlation Analysis
- Analyzing Previous Applications with Similar Processing
- Merged Both Application and Previous Application
- Integrated Data Analysis: Univariate, Bivariate, and Correlation Analysis

Initial Data Exploration

- Imported essential libraries: Pandas, Matplotlib, Numpy and Seaborn.
- Created two data frames from CSV files: “app_df" and “pre_app_df”.
- The "app_data" dataframe has a dimension of 307,511 rows and 122 columns, while the "pre_app_df" consists of 1,670,214 rows and 37 columns.
- Both dataframes consist of varying data types, including int64, float64, and object.
- Examine summary statistics for numerical columns in the dataframes.

Exploring Missing Data And Performing Data Cleaning And Manipulation

- Examining null values in each column and displaying percentage of null values.
- I assume that "OCCUPATION_TYPE" significantly impacts loan approval by reflecting the applicant's financial stability and profession, justifying a 40% threshold for column removal.
- Removing columns with null values exceeding 40%.
- Some columns contain varying percentages of null values, ranging from 30% to 19%, or even as low as 0.33%. I decided to impute these null values using either the mean or mode, as appropriate.
- I identified certain columns containing NAN values or XNA entries. To tackle this, we removed the corresponding rows or replaced NAN or XNA values with the mode.

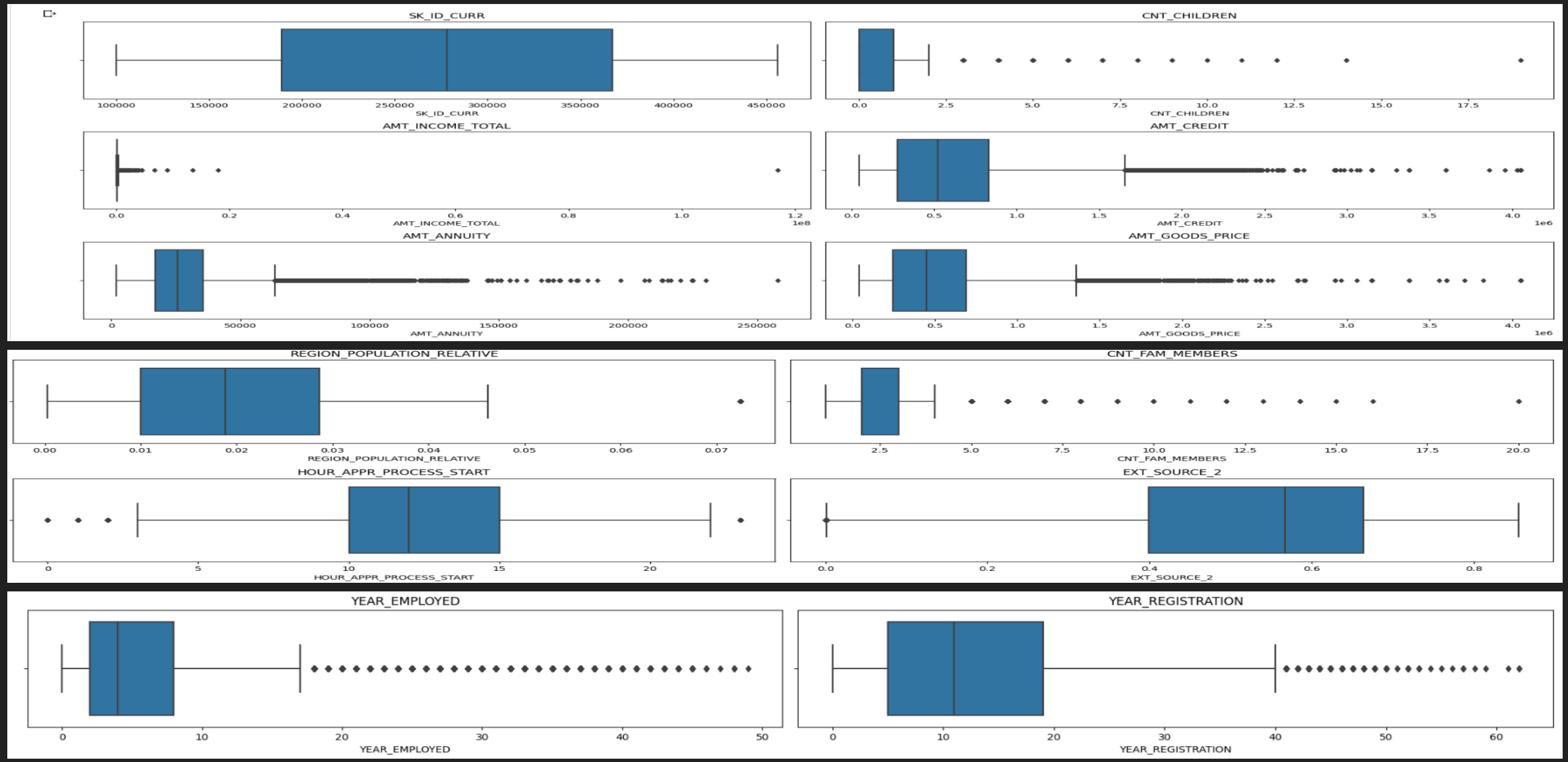
```
missing_percentage[missing_percentage > 30]
```

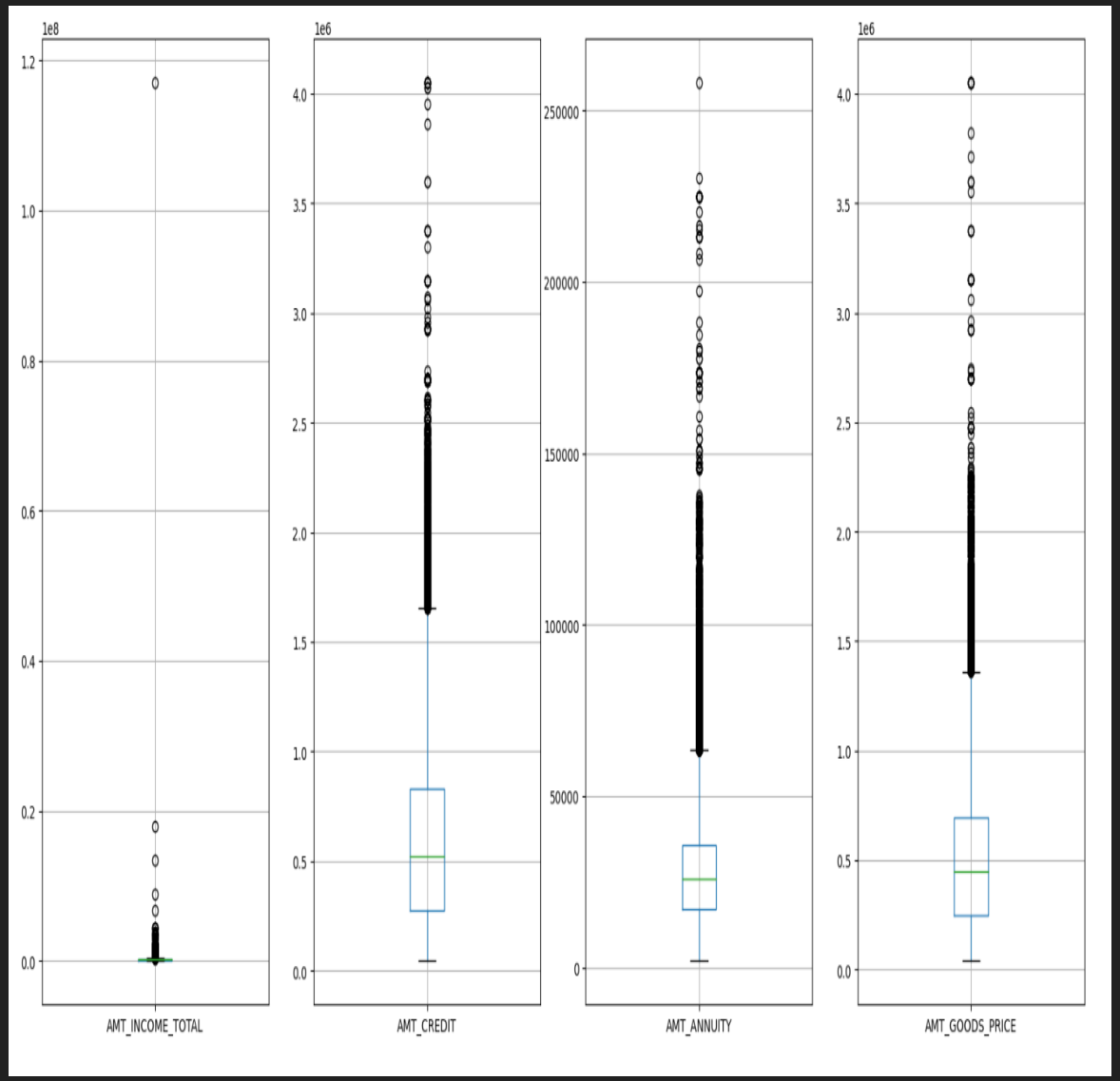
```
OWN_CAR_AGE                    65.99
OCCUPATION_TYPE                31.35
EXT_SOURCE_1                   56.38
APARTMENTS_AVG                 50.75
BASEMENTAREA_AVG               58.52
YEARS_BEGINEXPLUATATION_AVG    48.78
YEARS_BUILD_AVG                66.50
COMMONAREA_AVG                 69.87
ELEVATORS_AVG                  53.30
ENTRANCES_AVG                  50.35
FLOORSMAX_AVG                  49.76
FLOORSMIN_AVG                  67.85
LANDAREA_AVG                   59.38
LIVINGAPARTMENTS_AVG           68.35
LIVINGAREA_AVG                 50.19
NONLIVINGAPARTMENTS_AVG        69.43
NONLIVINGAREA_AVG              55.18
APARTMENTS_MODE                50.75
BASEMENTAREA_MODE              58.52
YEARS_BEGINEXPLUATATION_MODE   48.78
YEARS_BUILD_MODE               66.50
COMMONAREA_MODE                69.87
ELEVATORS_MODE                 53.30
ENTRANCES_MODE                 50.35
FLOORSMAX_MODE                 49.76
FLOORSMIN_MODE                 67.85
LANDAREA_MODE                  59.38
LIVINGAPARTMENTS_MODE          68.35
LIVINGAREA_MODE                50.19
NONLIVINGAPARTMENTS_MODE       69.43
NONLIVINGAREA_MODE             55.18
APARTMENTS_MEDI                50.75
BASEMENTAREA_MEDI              58.52
YEARS_BEGINEXPLUATATION_MEDI   48.78
YEARS_BUILD_MEDI               66.50
COMMONAREA_MEDI                69.87
ELEVATORS_MEDI                 53.30
ENTRANCES_MEDI                 50.35
FLOORSMAX_MEDI                 49.76
FLOORSMIN_MEDI                 67.85
```

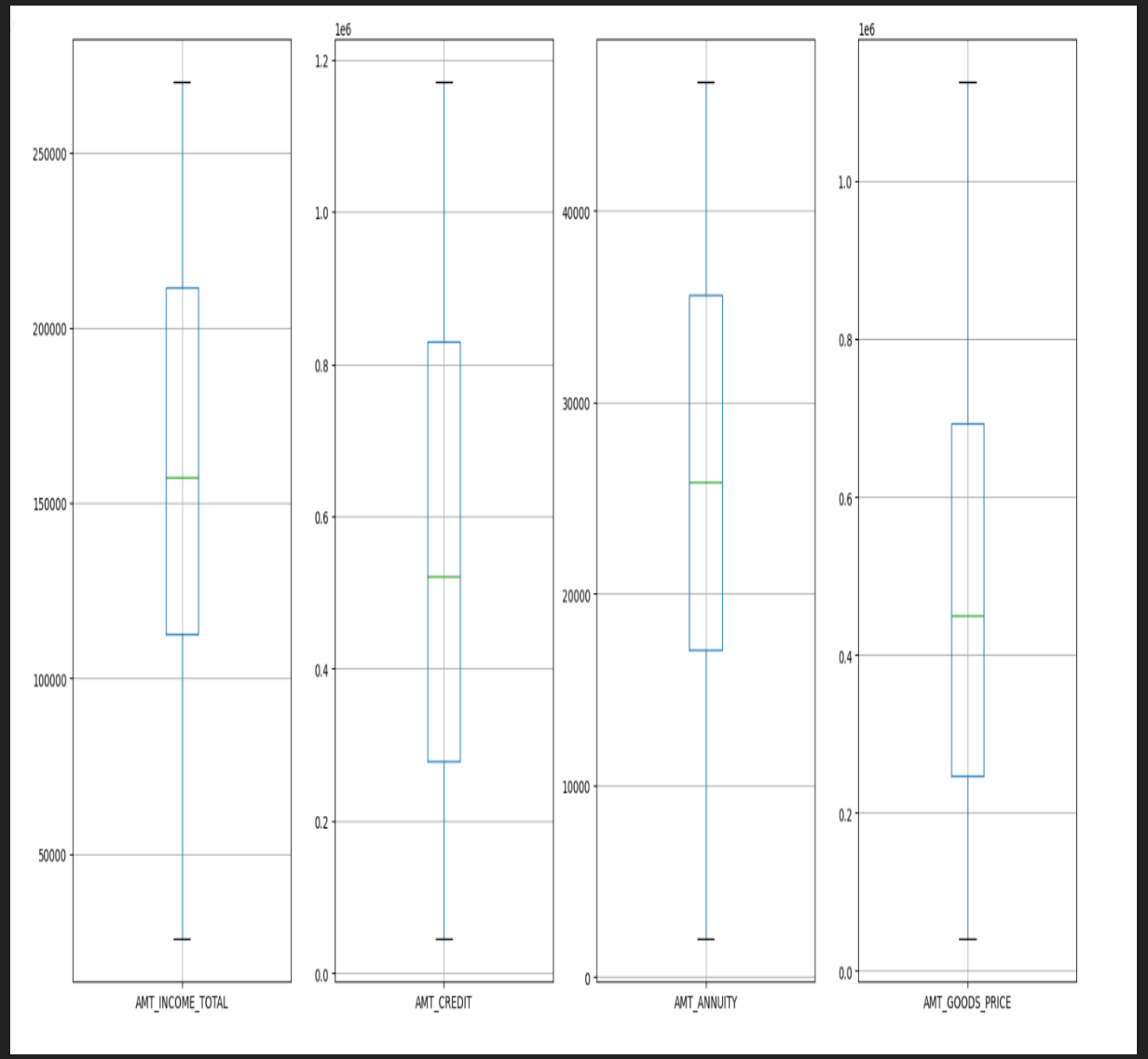
Outlier Detection and Mitigation: Capping and Winsorization Techniques

- Initial observation: We detected the presence of outliers in several numerical columns, which could adversely affect our data analysis.
- Exploratory visualization: To identify these outliers, we created box plots for the numerical columns, enabling us to visually pinpoint the extreme values.
- Mitigation strategy: Following the identification of outliers, we implemented two techniques to handle them effectively: capping and winsorization.
- Capping: For extreme values exceeding a predefined threshold, we applied capping, replacing them with the threshold value to limit their impact on subsequent analyses.
- Winsorization: As an alternative approach, we employed winsorization, which involved replacing extreme outliers with the nearest non-outlier values, preserving the overall data distribution while reducing the influence of outliers.
- We included box plots below, which clearly visualize the presence of outliers in the data, reaffirming the importance of addressing these data points during our analysis.
- Negative Data Values**:** Some columns, including "DAYS_BIRTH," "DAYS_EMPLOYED," "DAYS_REGISTRATION," and "DAYS_ID_PUBLISH," contain negative data values.
- Conversion to Absolute Values**:** In order to rectify this issue, we transformed these columns by converting the negative values into their absolute counterparts.

- Box Plot Analysis: Detecting Outliers in Data







- These financial fields exhibit far-reaching outliers, signifying extreme values that warrant special attention and outlier management techniques.
- Outlier Display: The left side displays outliers.
- Post-Outlier Handling: On the right side, you can see the data after outlier handling.

Univariate Analysis of Application Data

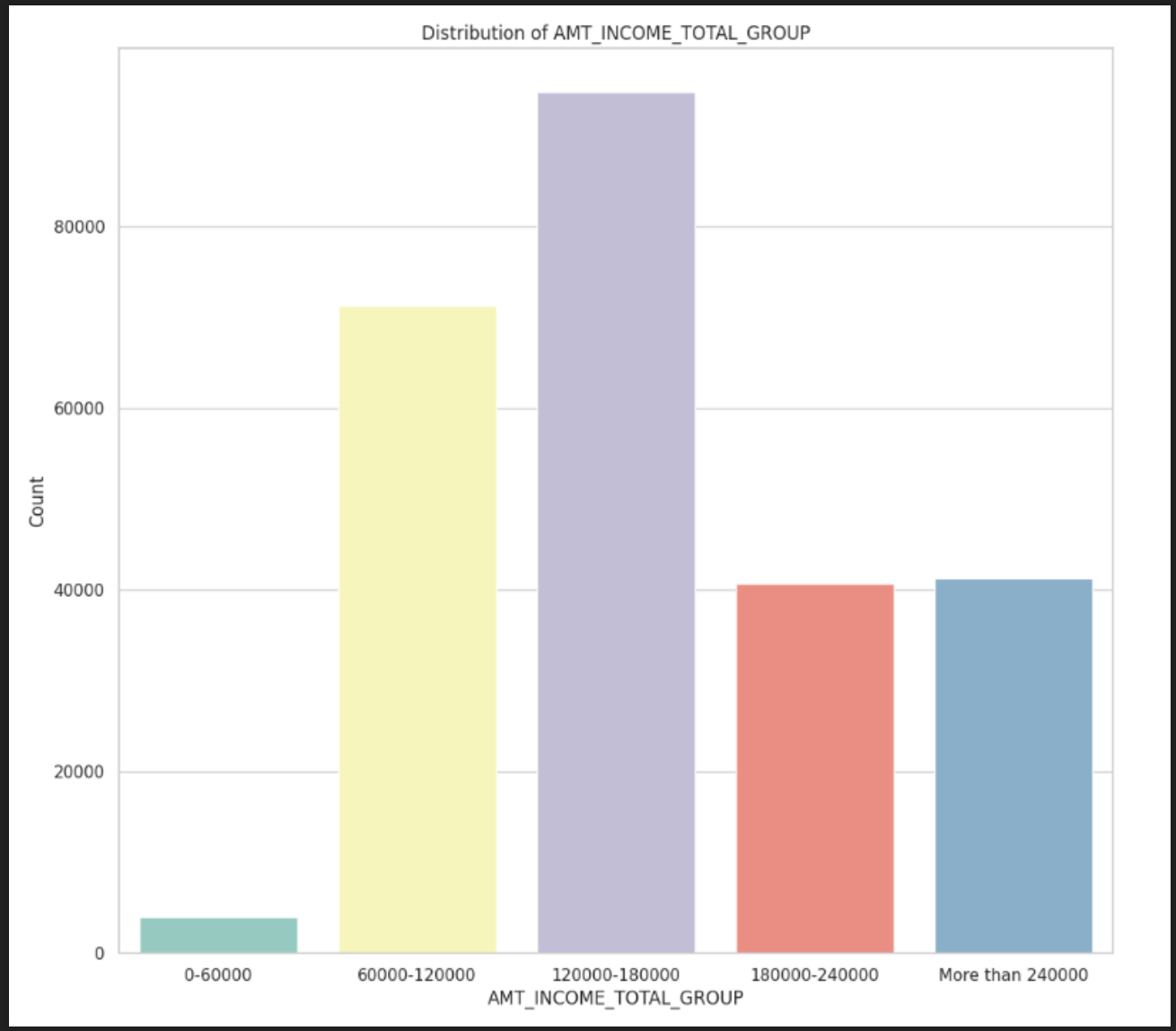
- Univariate Analysis of Numerical Columns: We conducted a univariate analysis focusing on our numerical columns.
- Binning Data: To facilitate this analysis, we initially divided the data into bins.
- Bar Graph Visualization: Subsequently, we created bar graphs to visually represent the data distribution for specific columns.
- Columns Included: Some of the columns analyzed in this manner included "AMT_GOODS_PRICE," "AMT_INCOME_TOTAL," "AMT_CREDIT," and "AMT_ANNUITY," among others. There are bar graphs displayed beneath this slide.
- Also created count plot and histogram visualizations for fields with object data types.
- In our data analysis, we've used count plots and histograms to visualize various dataset fields, gaining valuable insights for further analysis.

Key Observations from AMT_INCOME_TOTAL Analysis

1. **Income Group Analysis:** When we grouped the income data into AMT_INCOME_TOTAL_GROUP bins, a prominent trend emerged.

2. **Dominant Income Range:** The analysis clearly indicates that the largest number of applicants belong to the income range of 120,000 to 180,000 within the AMT_INCOME_TOTAL_GROUP.

3. **Lowest Applicant Count:** In contrast, the smallest number of applicants can be observed in the income bracket of 0 to 60,000, as revealed by the graph.

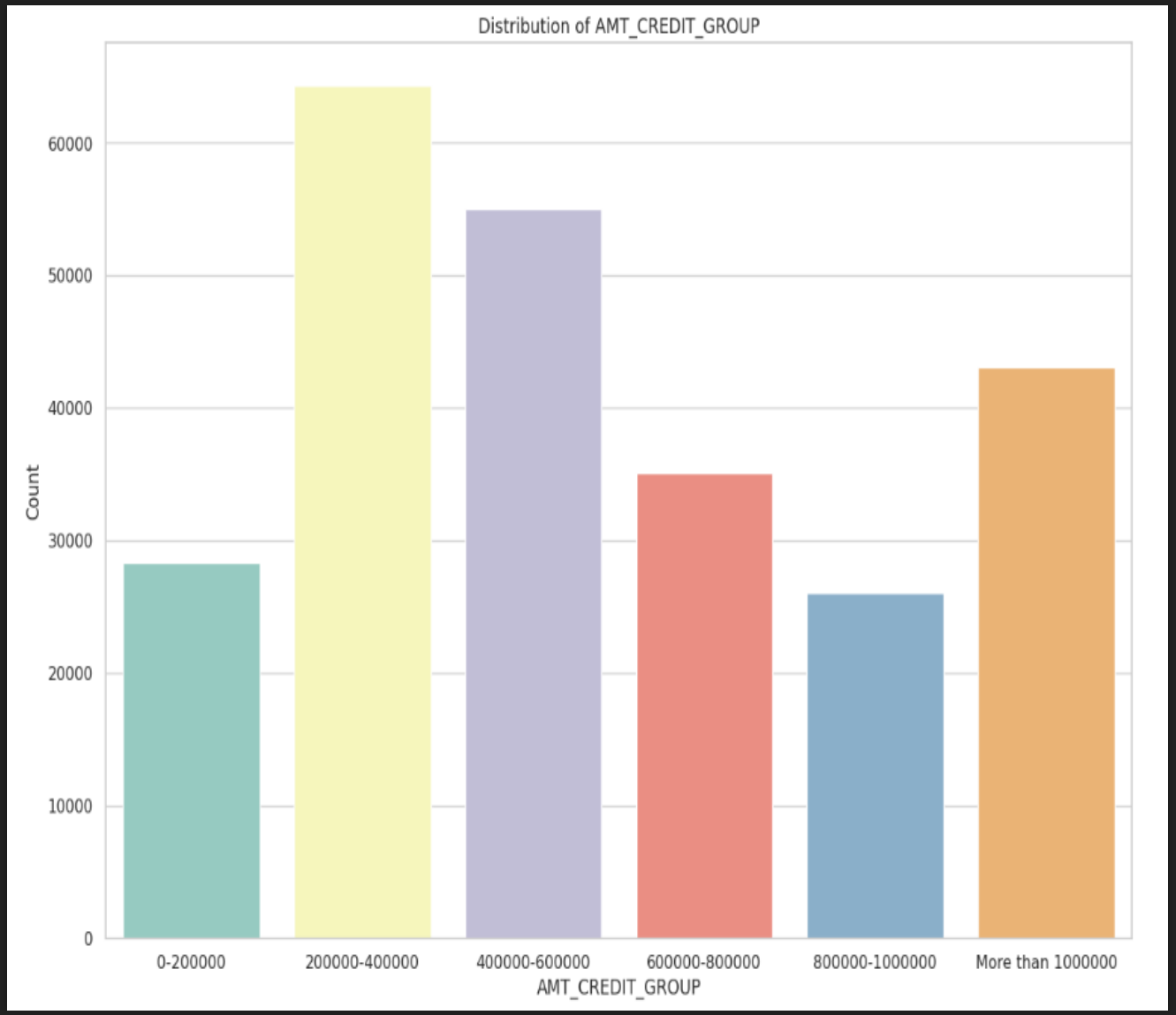


Key Insights from AMT_CREDIT Analysis

1. **AMT_CREDIT Grouping:** After categorizing AMT_CREDIT into AMT_CREDIT_TOTAL_GROUP bins, a distinct pattern emerges from the analysis.

2. **Dominant Credit Range:** The analysis highlights that the largest number of applicants falls within the credit amount range of 200,000 to 400,000 within the AMT_CREDIT_TOTAL_GROUP.

3. **Lowest Applicant Count:** Conversely, the graph illustrates that the smallest number of applicants can be found in the credit amount bracket of 800,000 to 1,000,000.

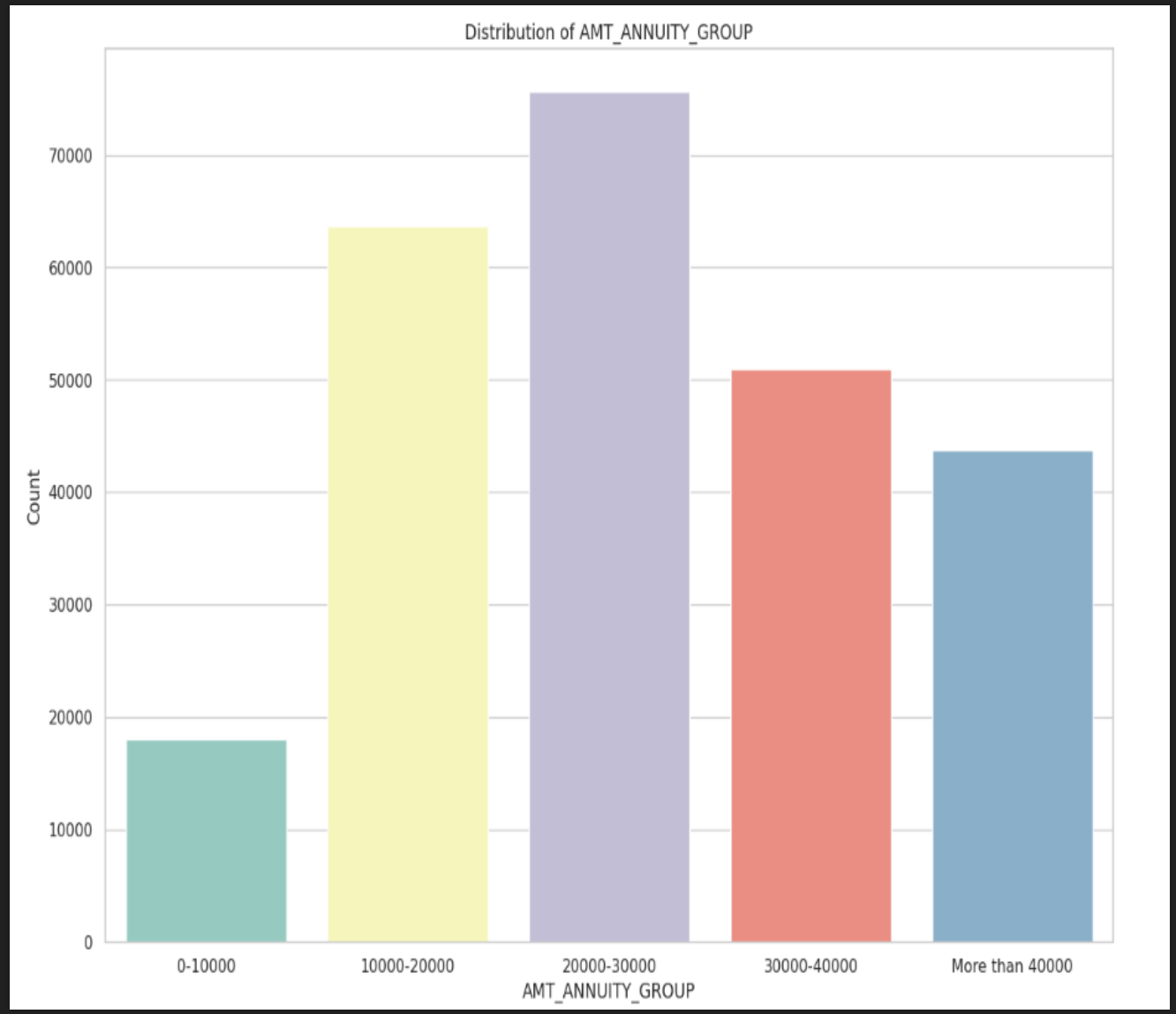


Key Observations from AMT_ANNUITY Analysis

1. **AMT_ANNUITY Grouping:** After segmenting AMT_ANNUITY into AMT_ANNUITY_GROUP categories, a distinct pattern becomes evident in our analysis.

2. **Dominant Annuity Range:** The graph clearly indicates that the highest number of applicants falls within the loan annuity range of 20,000 to 300,000 within the AMT_ANNUITY_GROUP.

3. **Lowest Applicant Count:** In contrast, the graph illustrates that the smallest count of applicants can be identified within the loan annuity range of 0 to 10,000.

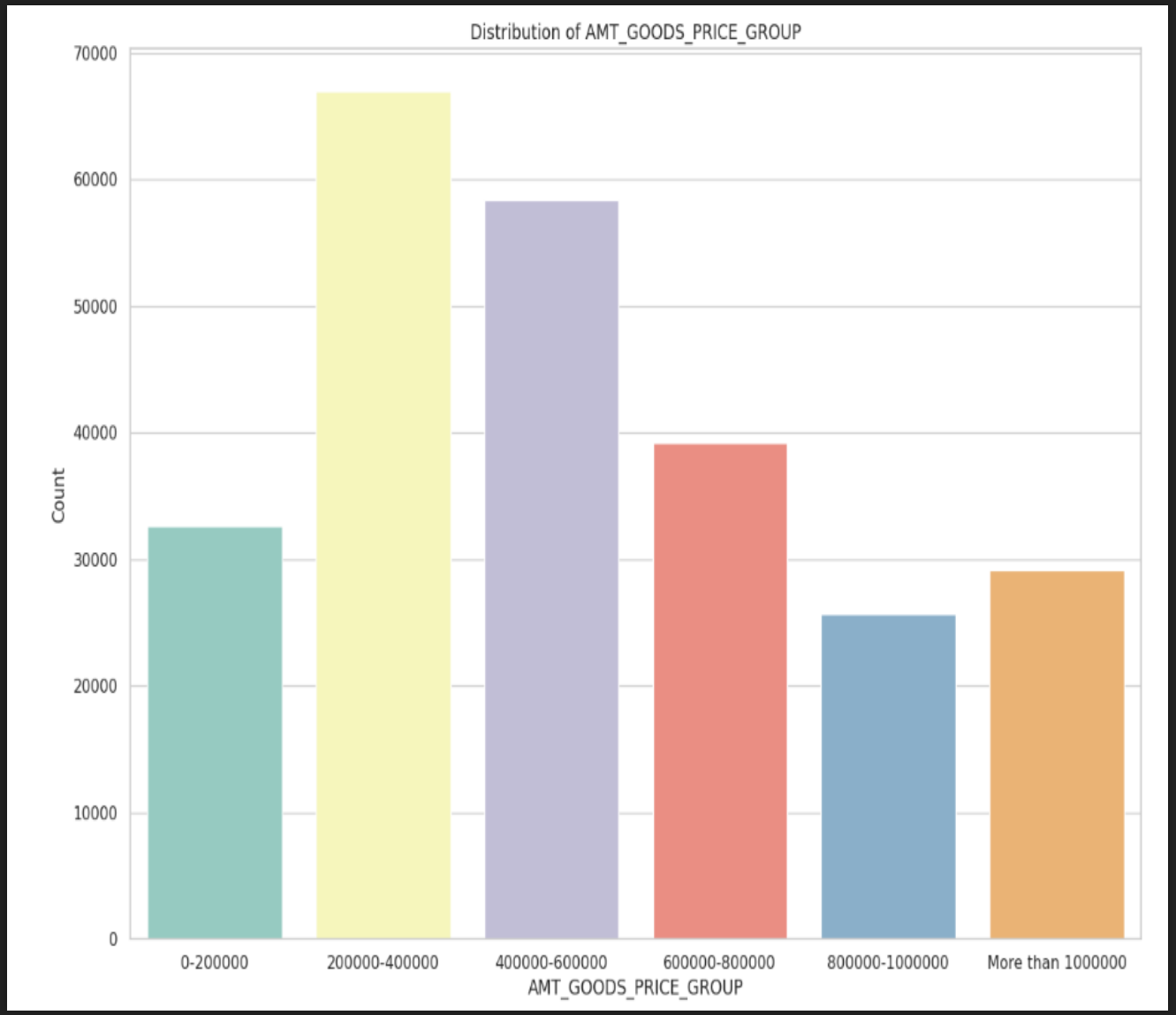


Key Insights from AMT_GOODS_PRICE Analysis

1. **AMT_GOODS_PRICE Grouping:** After categorizing AMT_GOODS_PRICE into AMT_GOODS_PRICE_GROUP categories, our analysis brings to light distinct trends.

2. **Dominant Goods Price Range:** The graph prominently showcases that the highest number of applicants can be found within the price range of 200,000 to 400,000 for goods in the AMT_GOODS_PRICE_GROUP.

3. **Lowest Applicant Count:** Conversely, the graph demonstrates that the smallest applicant count falls within the price range of 800,000 to 1,000,000 for goods within this grouping.

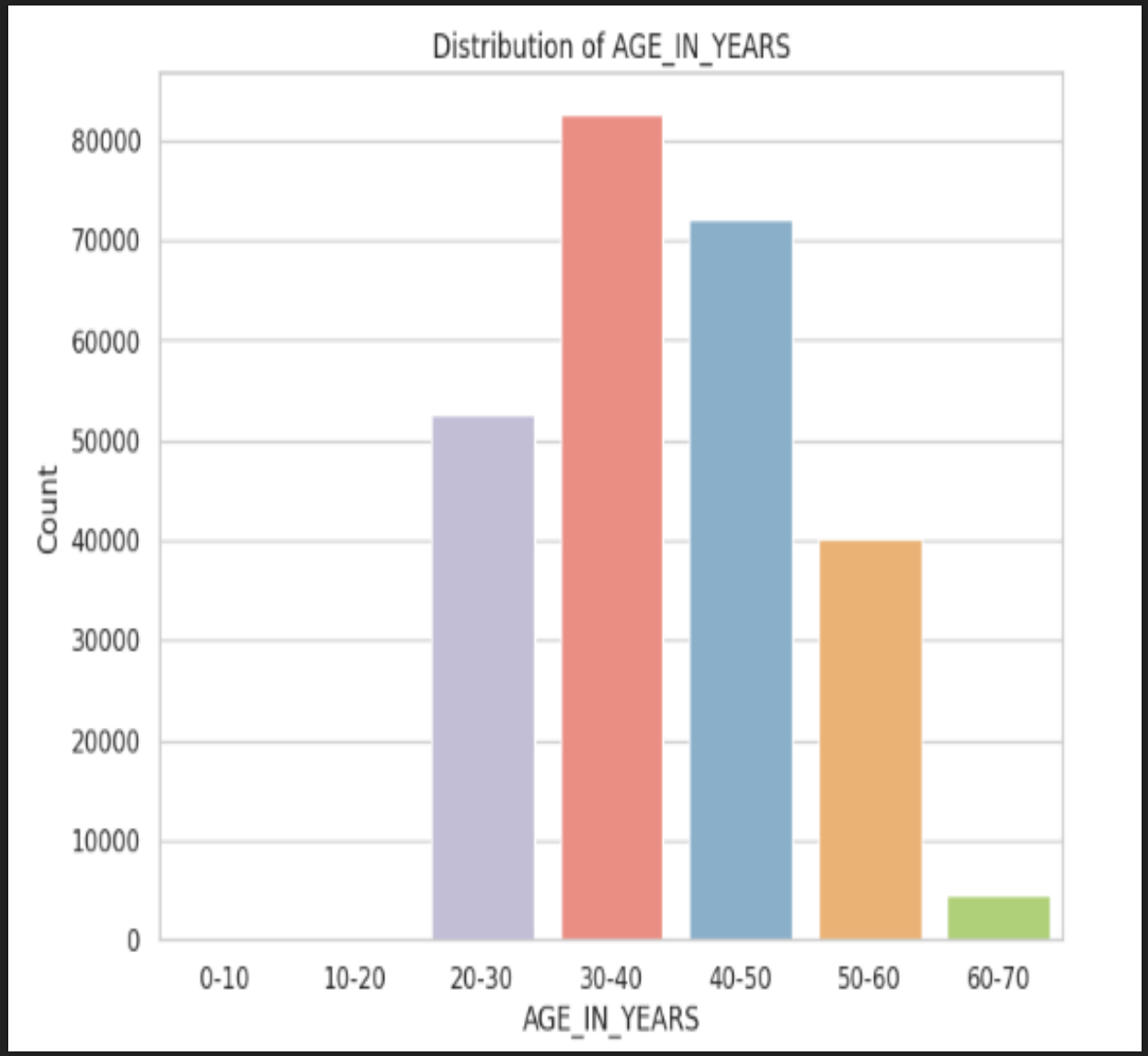


Insights from Age Data Distribution Analysis

1. **Age Distribution Analysis:** When examining the count graph for the "Age_in_years" category, a notable pattern emerges.

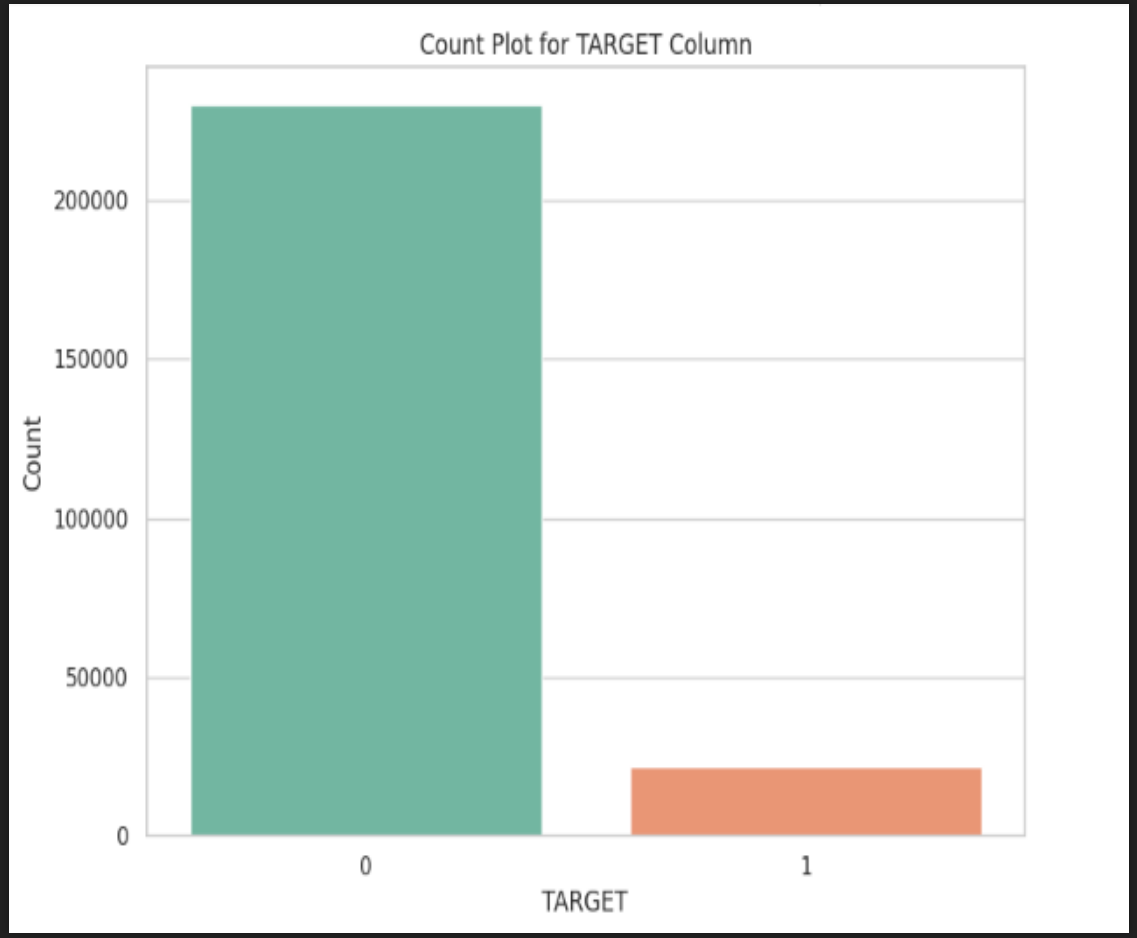
2. **No Applicants Below 20:** Interestingly, the graph reveals that there are no applicants aged between 0 and 20 years.

3. **Dominant Age Group:** The most significant observation is that the majority of applicants belong to the age category of 30 to 40 years, reflecting the highest number of applicants within this age range.

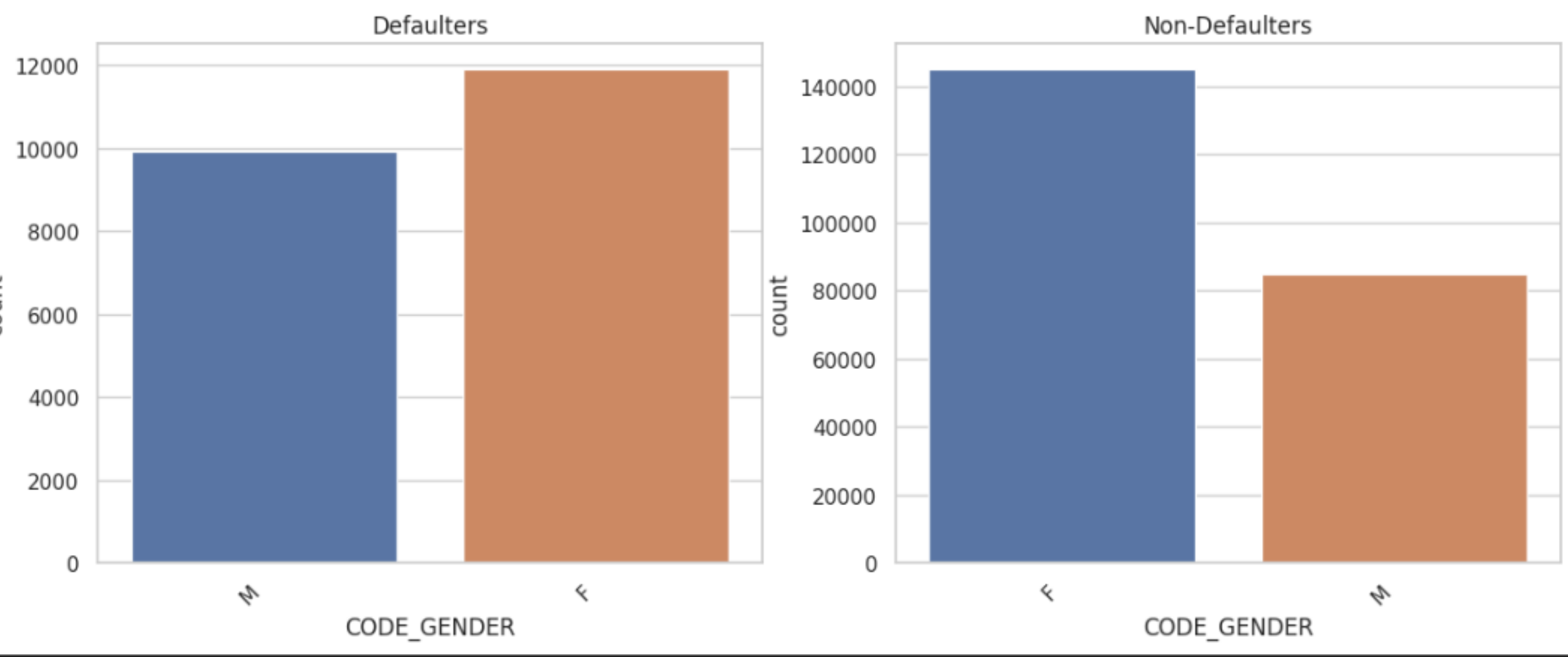


Data Imbalance

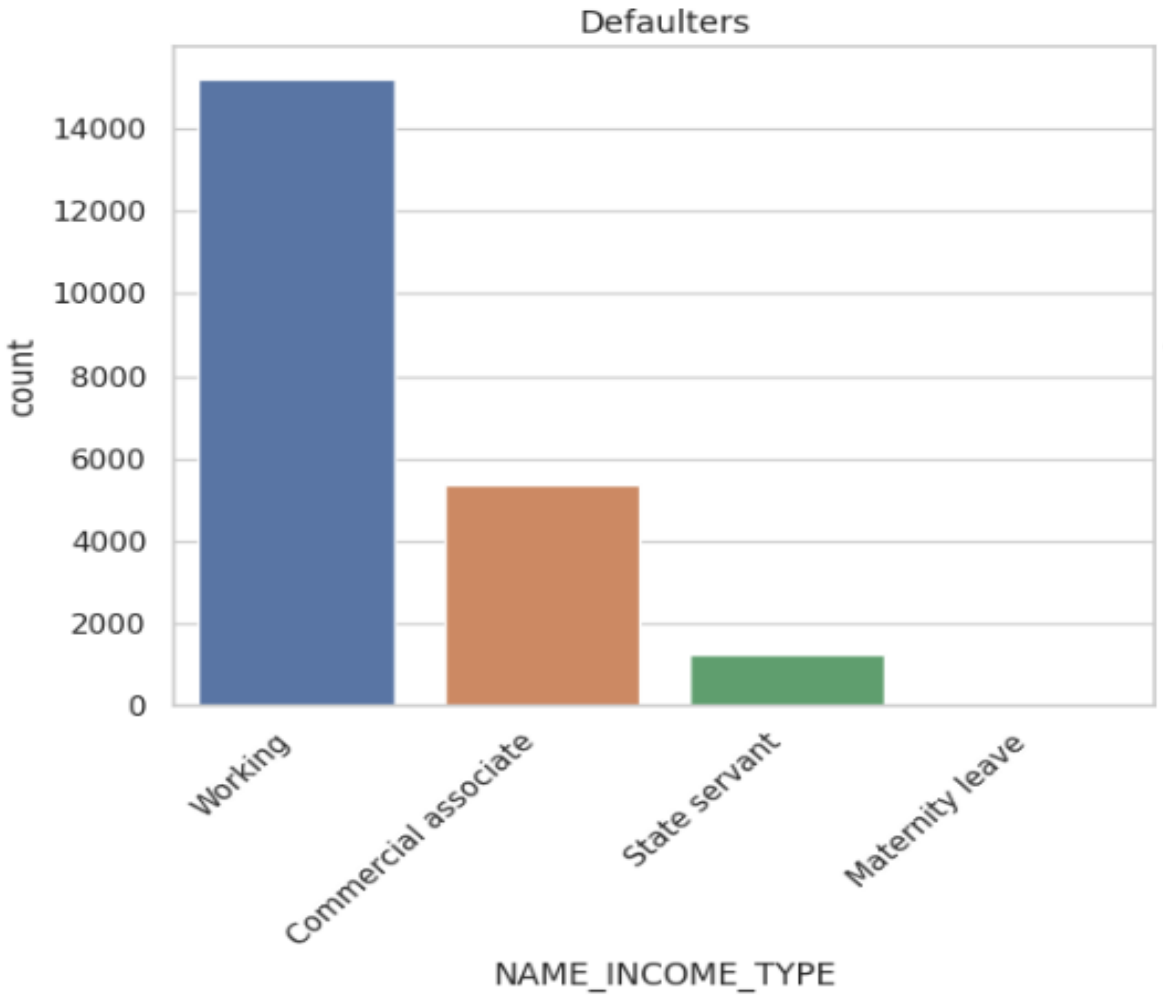
- **Understanding Data Imbalance:** Data imbalance occurs when there's a significant disparity in the number of instances representing different categories within a dataset.
- **Identification of Data Imbalance:** The analysis of the target column indicates the presence of data imbalance, where the target variable takes on values of 0 or 1.
- **Data Split:** To address this, we divided the data frame into two separate subsets: one for Target 1 and another for Target 0.
- **Target Variable 1:** Target variable 1 represents clients with payment difficulties. These individuals had late payments for at least one of the first Y installments of the loan, with a delay of more than X days, as defined in our sample.
- **Target Variable 0:** Target variable 0 encompasses all other cases, signifying clients without payment difficulties.
- **Data Imbalance in the Target Variable:** The target variable highlights a data imbalance, where a value of 1 corresponds to clients who encountered payment difficulties, characterized by late payments on certain early installments. Meanwhile, a value of 0 encompasses all other cases.
- **Imbalance Ratio Calculation:** After performing calculations, it was determined that the data imbalance ratio stands at 10.55.
- **Separate Analysis:** We conducted separate data analyses for these two datasets, involving univariate and bivariate analyses as well as correlation graph plotting, to gain a comprehensive understanding of each dataset independently.

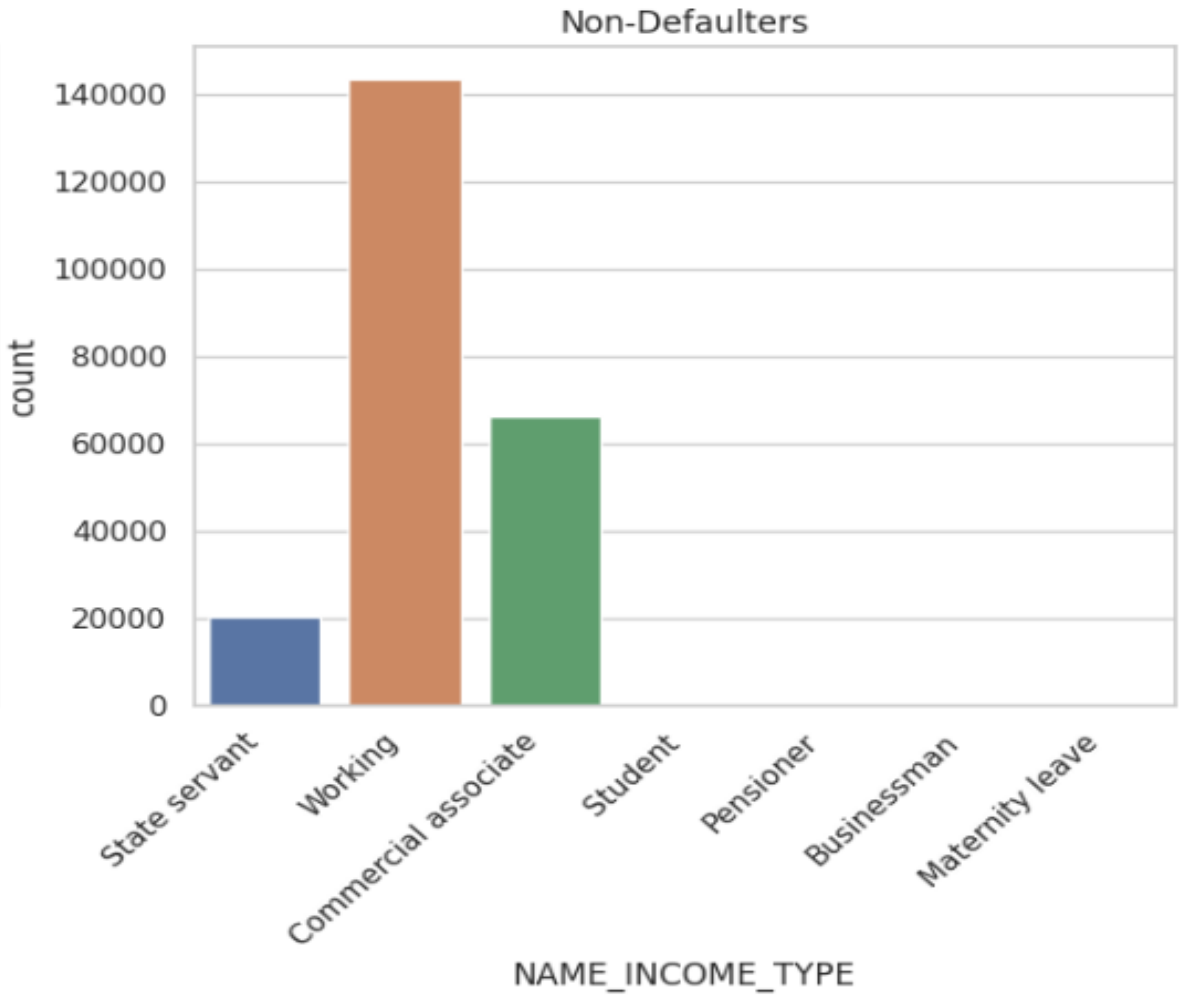


Segmented Univariate analysis and Bivariate analysis

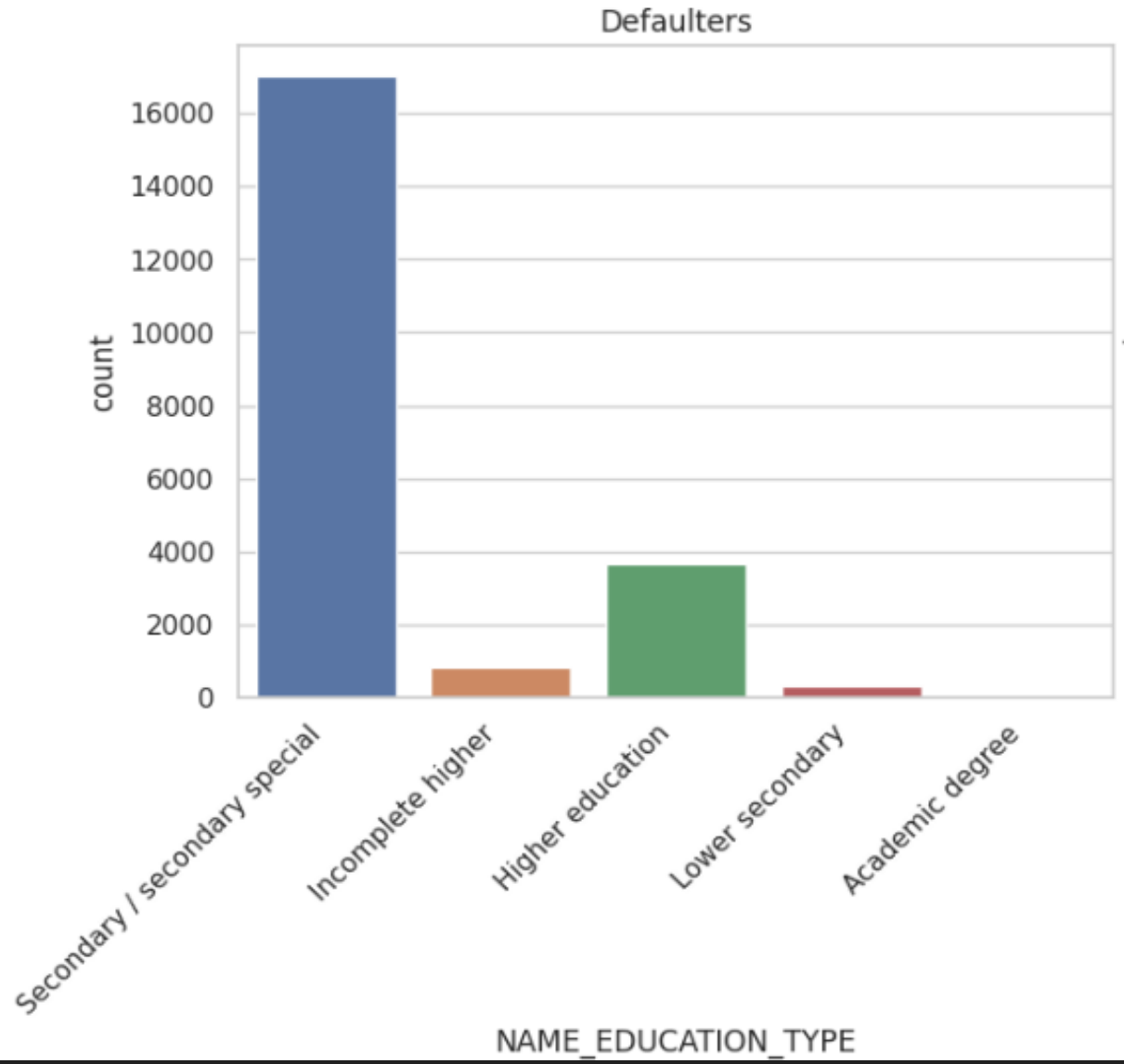


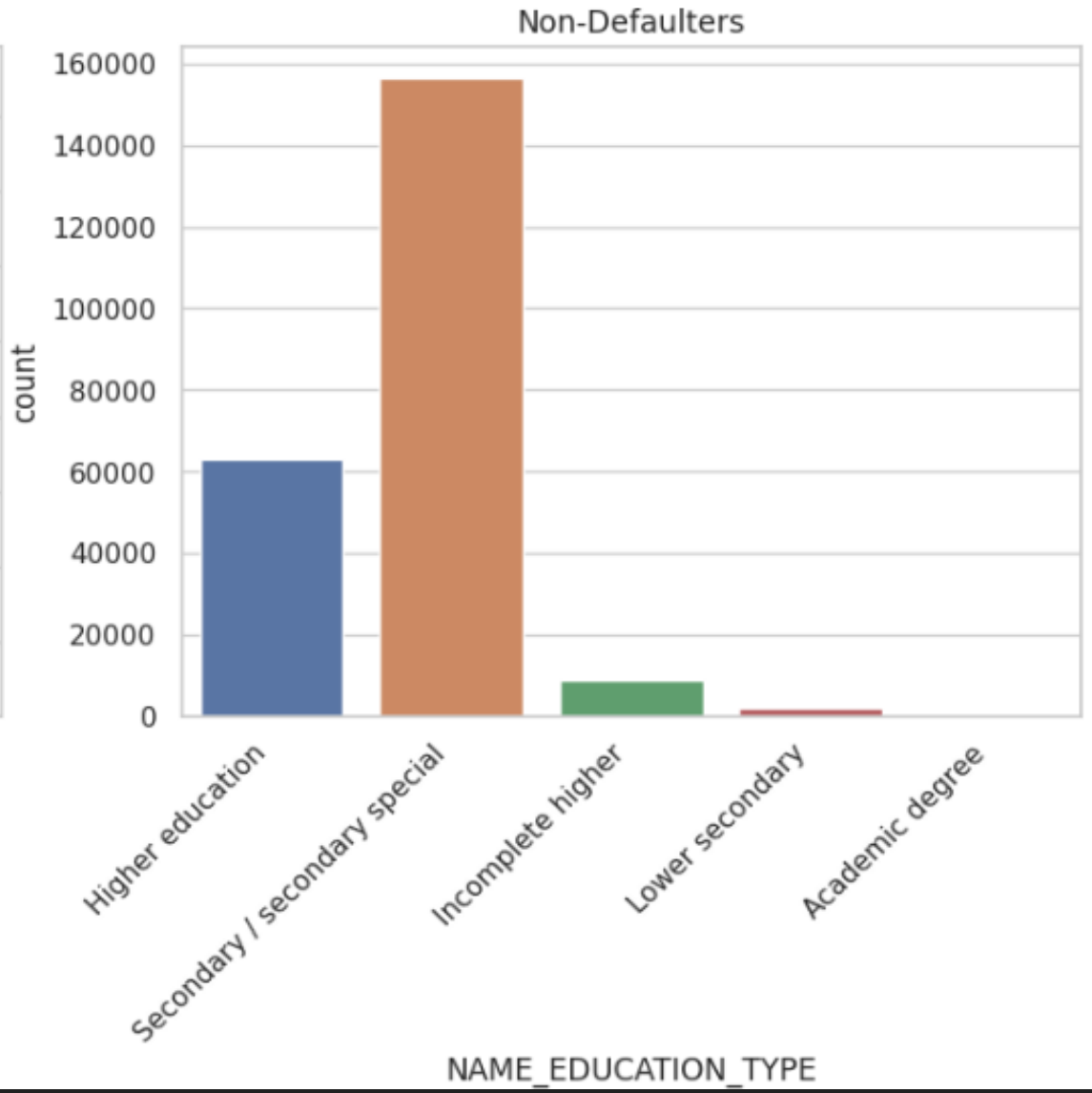
- The plot clearly illustrates that females have a higher ratio in both Target0 and Target1 groups.
- However, within Target0 (Non-Defaulters), females have a higher ratio than in Target1 (Defaulters).
- Consequently, it is advisable for the bank to prioritize granting loans to females, as they exhibit a higher likelihood of successful payments.
- In contrast, males have a higher ratio in Target1 compared to Target0, but their representation among Non-Defaulters is lower than among Defaulters.



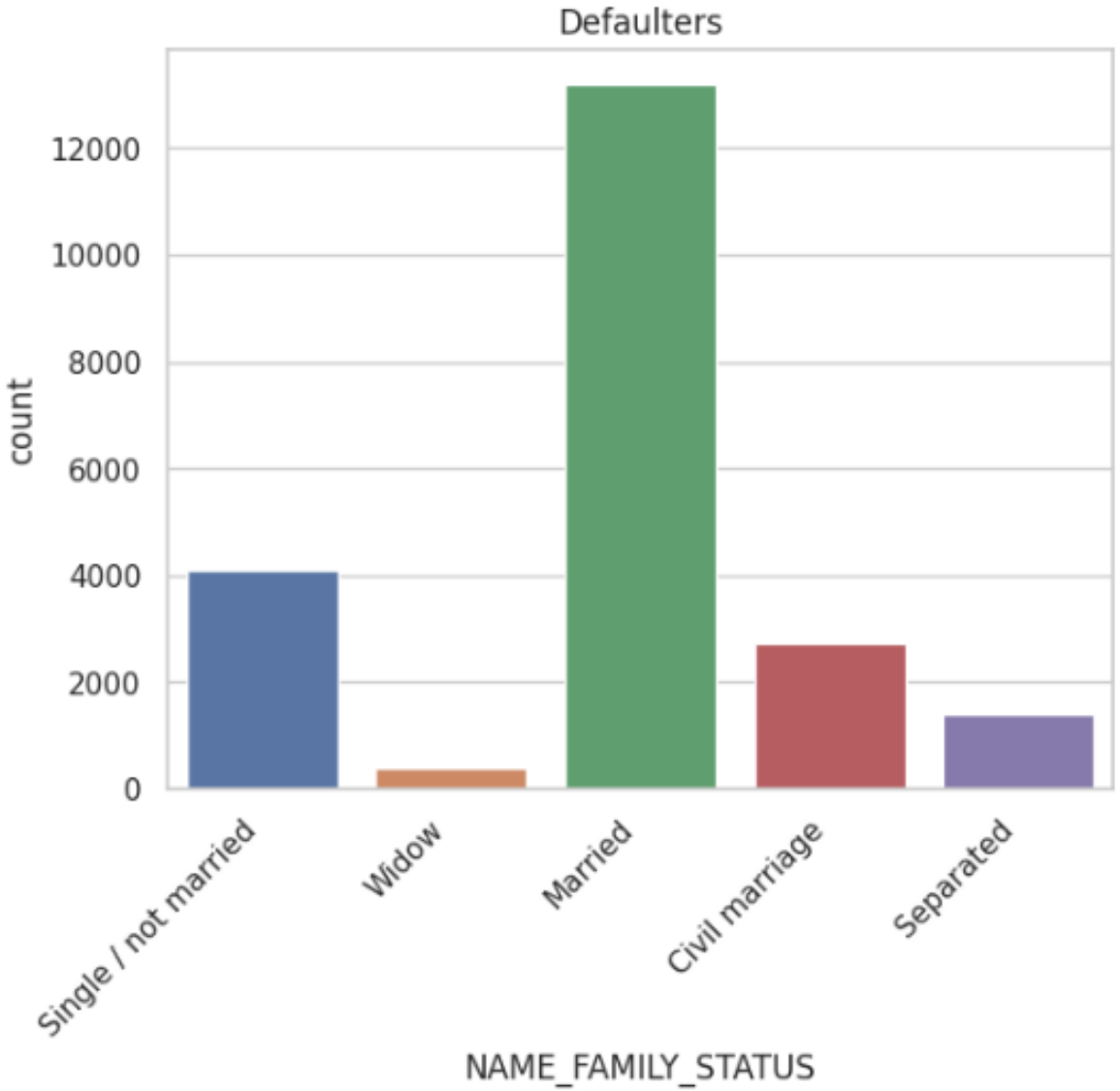


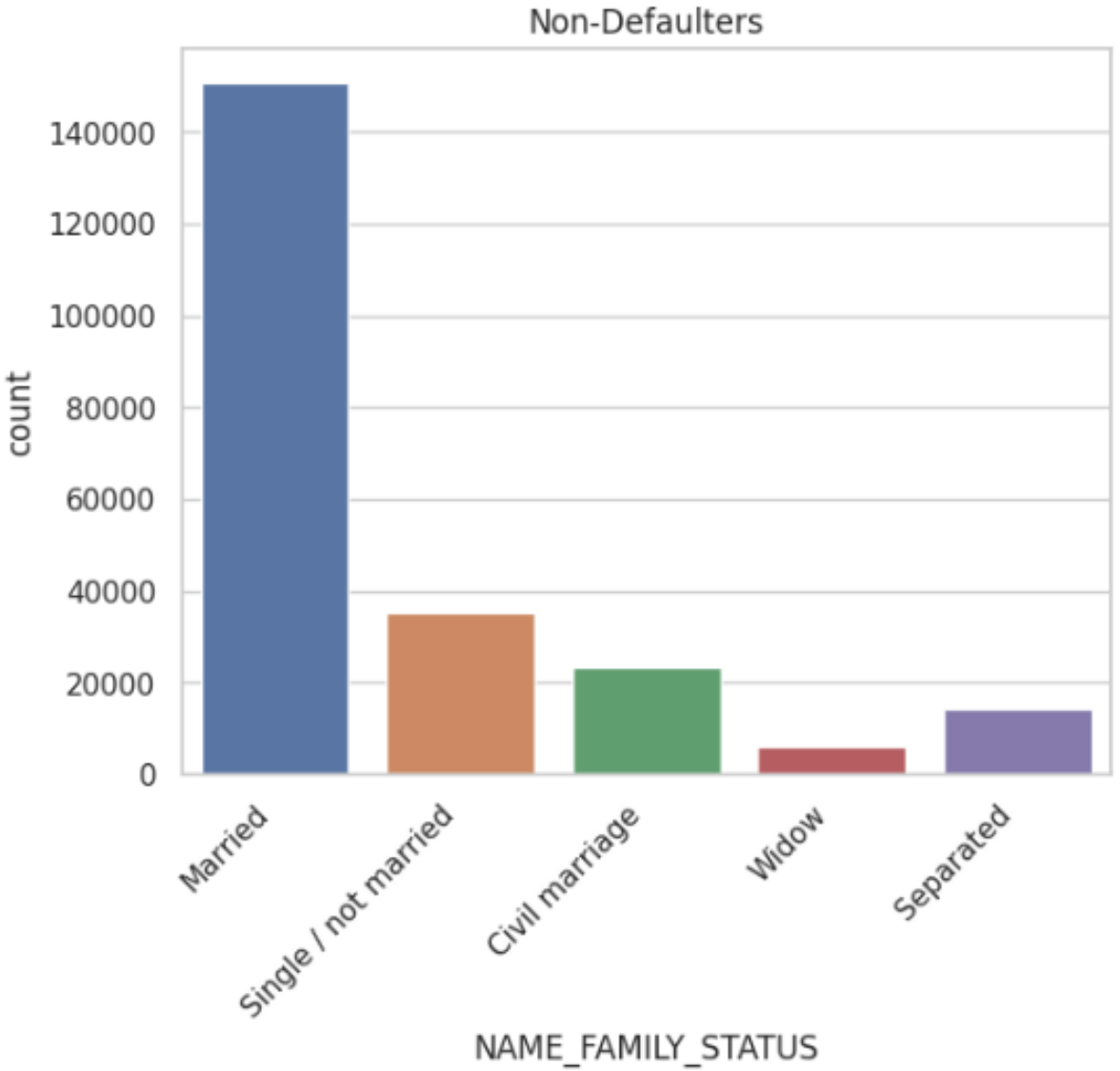
- **Income Types in Both Target Groups:** In both Target 0 (Non-Defaulters) and Target 1 (Defaulters), the prevalence of "working" income type is notably higher.
- **Low Frequency of State Servants:** It's worth noting that the frequency of state servants is comparatively low. Therefore, the bank should carefully evaluate loan applications from this category.
- **Loan Consideration for Working Class:** Considering the working class with "income type" as "working," the bank may need to offer loans with higher interest rates as this group seems to have a higher representation among both defaulters and non-defaulters.



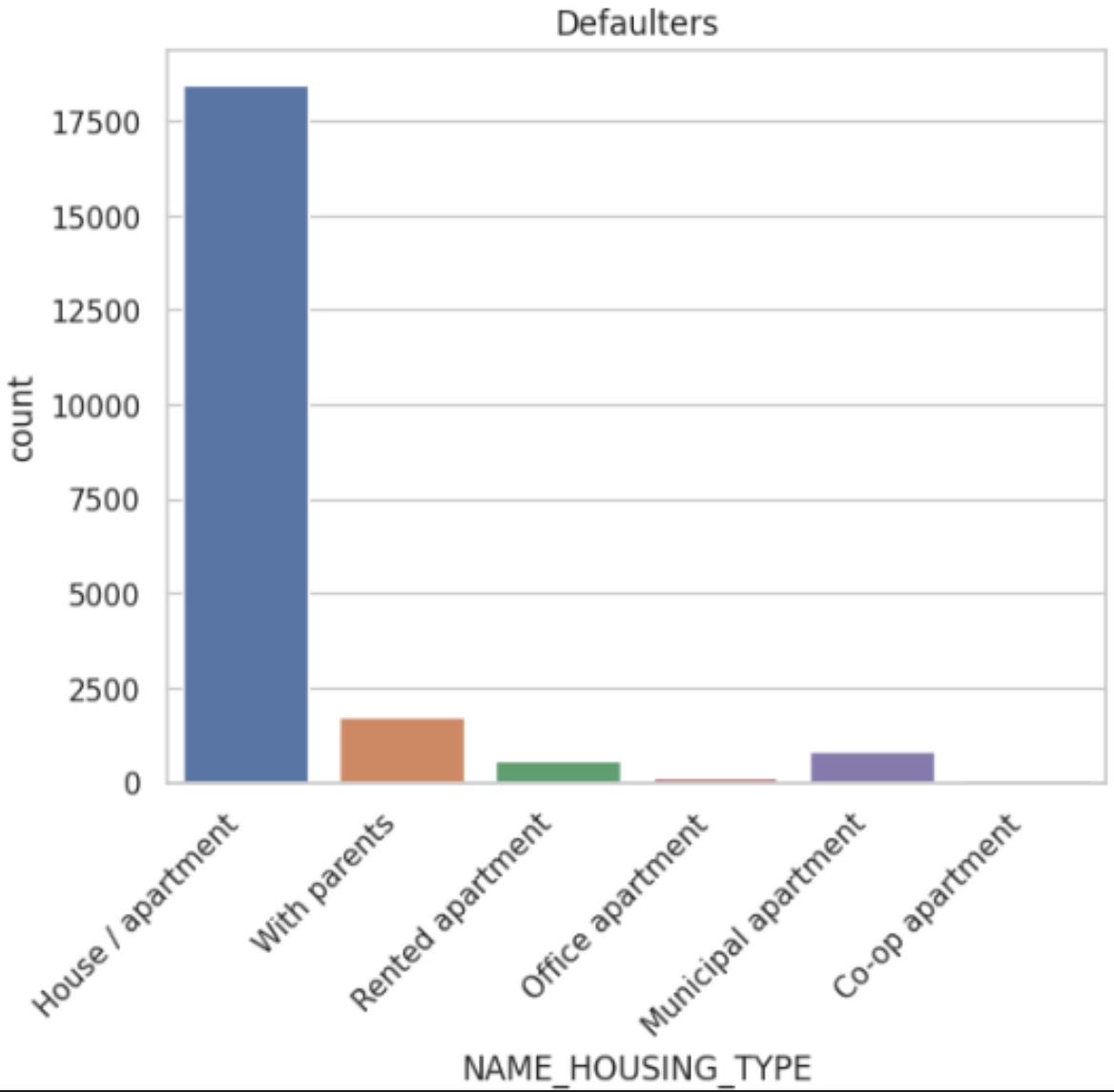


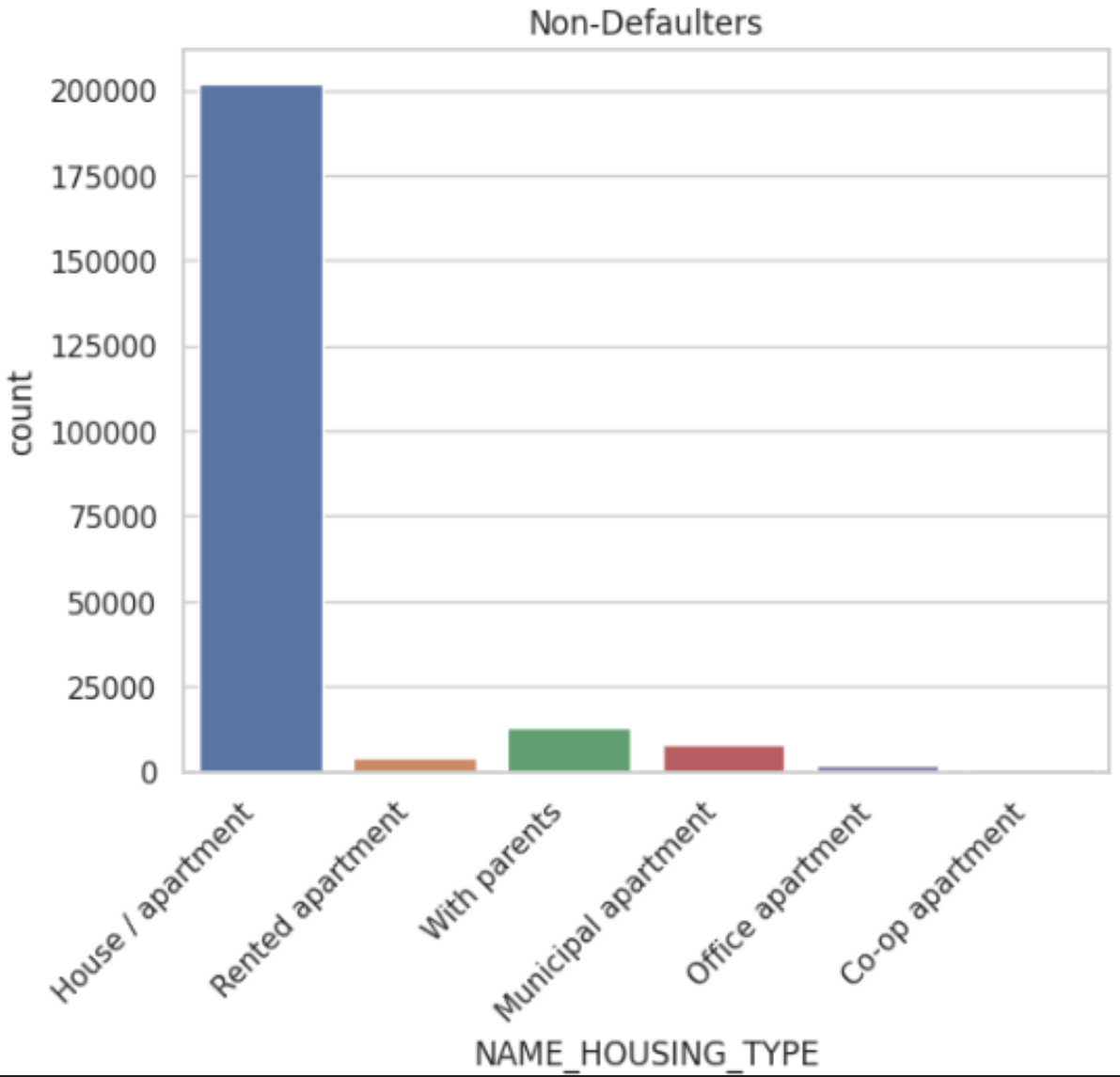
- **Education Type Insights:** Secondary/secondary special education types show higher representation in both Target 0 (Non-Defaulters) and Target 1 (Defaulters). However, it's essential to note that Target 0 has a significantly larger count than Target 1 in this category.
- **Consideration for Loans:** While secondary/secondary special education types may be considered, other factors should also be taken into account when evaluating loan applications.
- **Lowest Academic Degree Representation:** Academic degree education types have the lowest representation in both Target 0 (Non-Defaulters) and Target 1 (Defaulters).



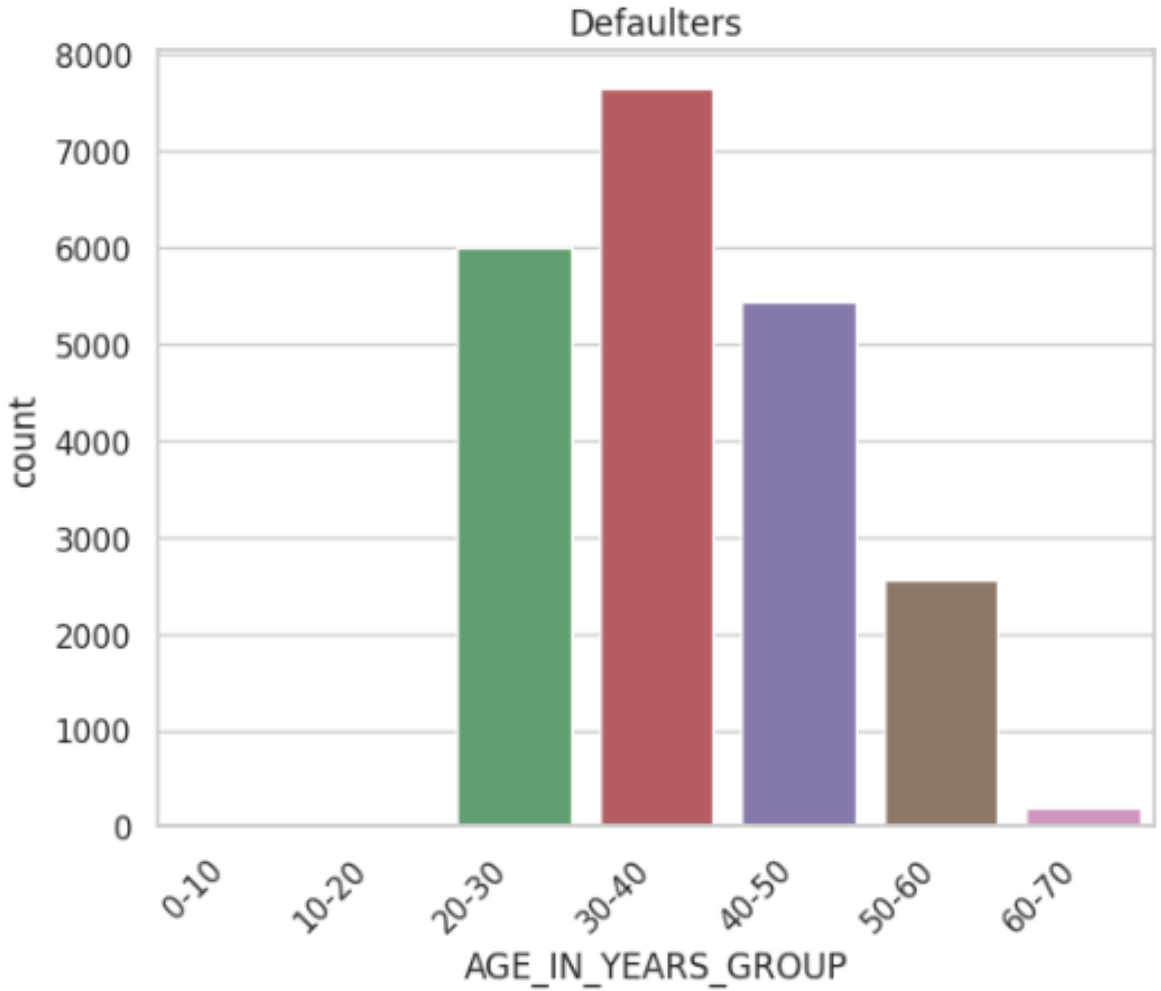


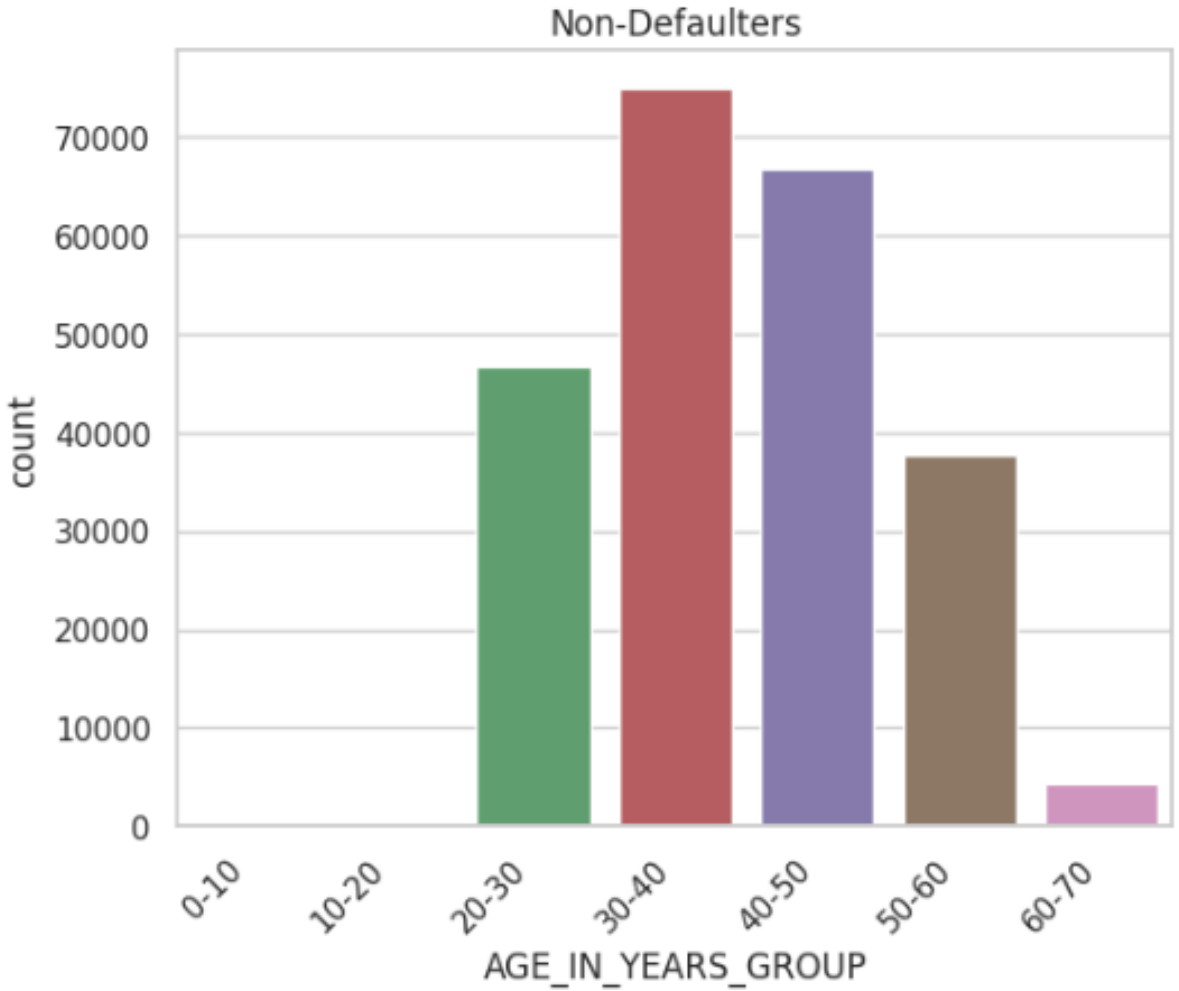
- **Family Status Insights:** Marriage family status is more prevalent in both Target 0 (Non-Defaulters) and Target 1 (Defaulters). However, it's important to note that Target 0 has a significantly larger count than Target 1 in this category.
- **Loan Evaluation:** While considering applicants with marriage family status may be relevant, it's advisable for the bank to factor in other criteria as well.
- **Lowest Representation:** Widow family status have the lowest representation in both Target 0 (Non-Defaulters) and Target 1 (Defaulters).



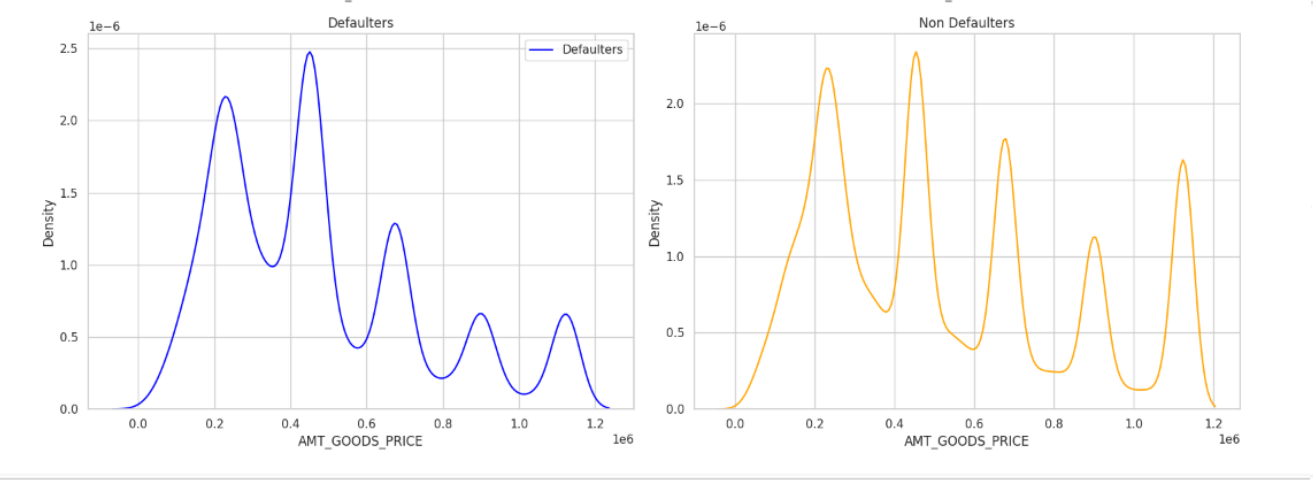


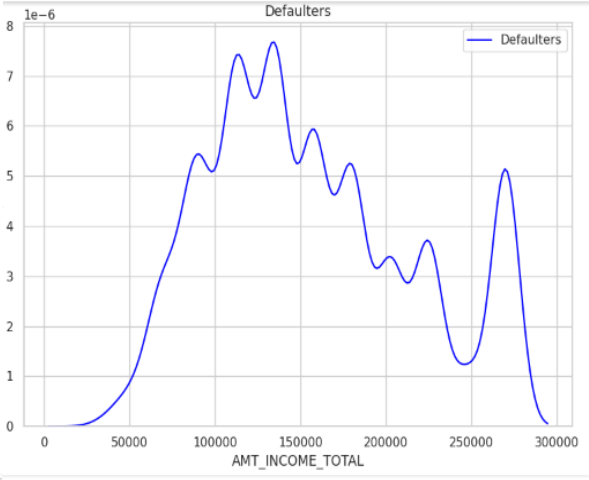
- **House Type Insights:** House/apartment house types show a higher presence in both Target 0 (Non-Defaulters) and Target 1 (Defaulters). However, it's crucial to note that Target 0 significantly outnumbers Target 1 in this category.
- **Loan Evaluation:** While applicants with house/apartment house types may be considered, it's advisable for the bank to take into account other relevant factors as well.
- **Lowest Representation:** Co-op apartment house types have the lowest representation in both Target 0 (Non-Defaulters) and Target 1 (Defaulters).

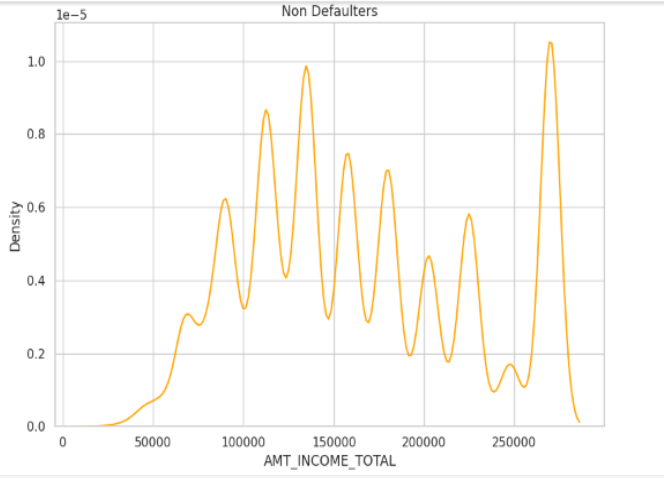


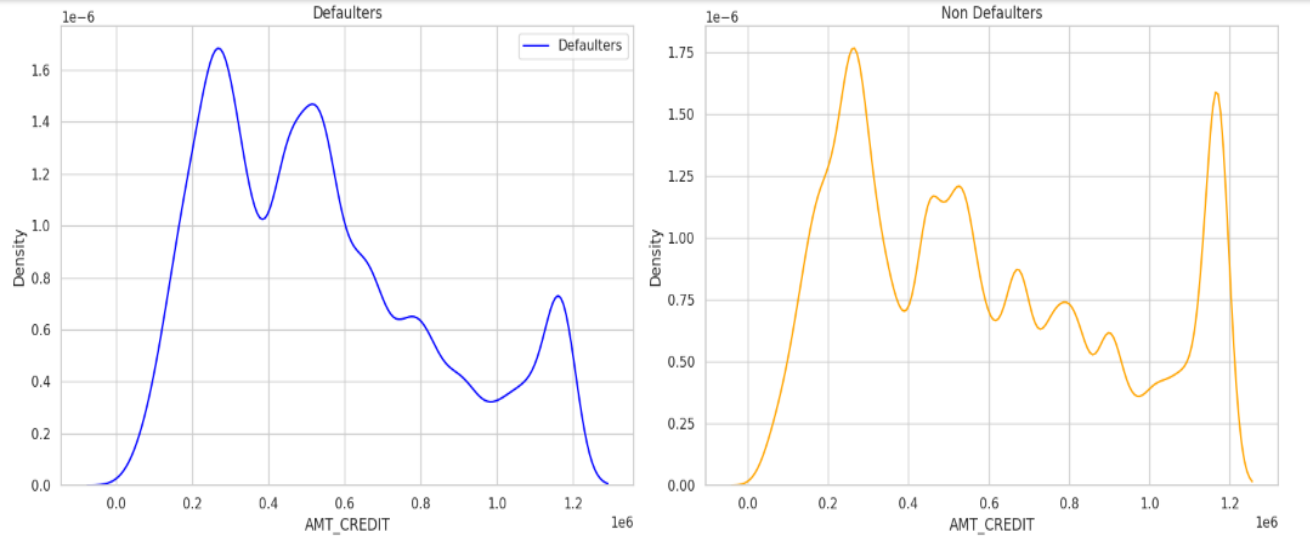


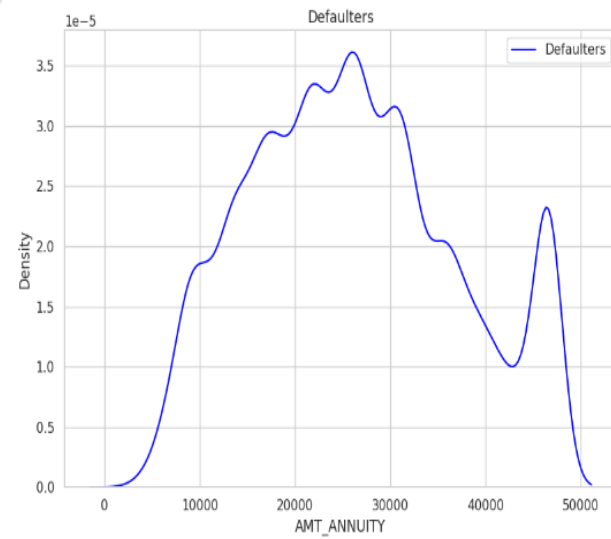
- **Age Group Insights:** Notably, there are no applications within the 0-20 age range. The 30-40 age group exhibits higher representation in both Target 0 (Non-Defaulters) and Target 1 (Defaulters). However, it's important to acknowledge that the count of Target 0 significantly surpasses that of Target 1 within this age bracket.
- **Loan Evaluation:** While considering applicants from the 30-40 age group may be pertinent, the bank should also take into consideration other relevant factors.
- **Lowest Representation:** The 60-70 age group has the lowest representation in both Target 0 (Non-Defaulters) and Target 1 (Defaulters).

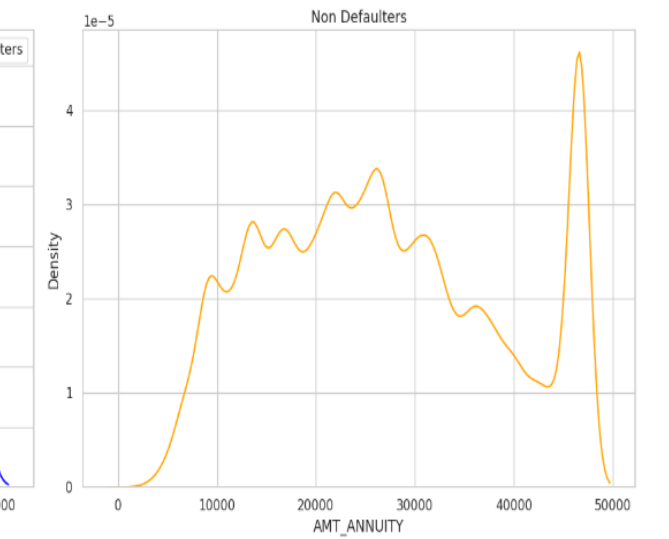




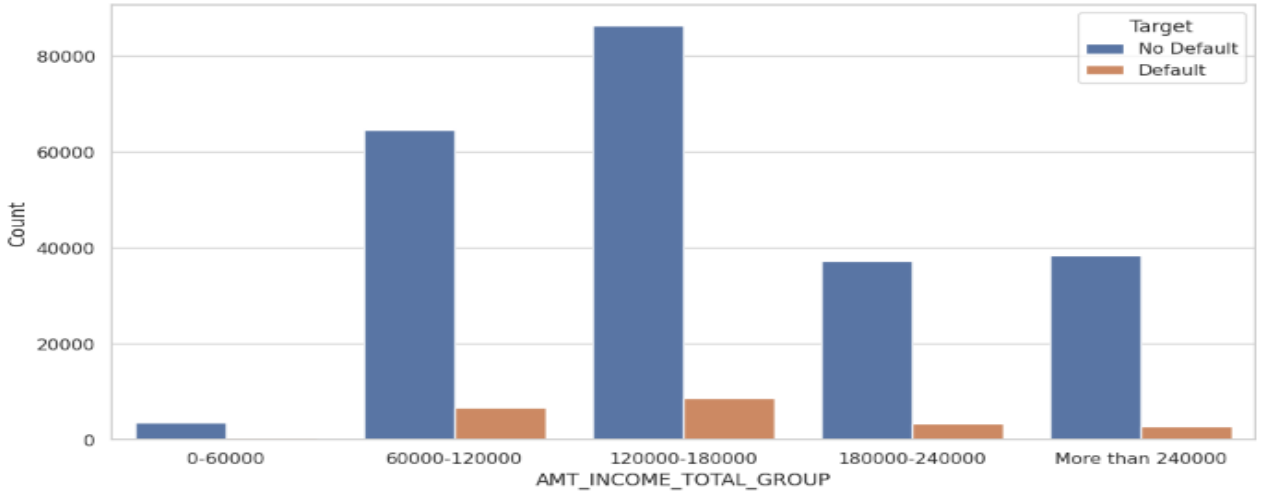


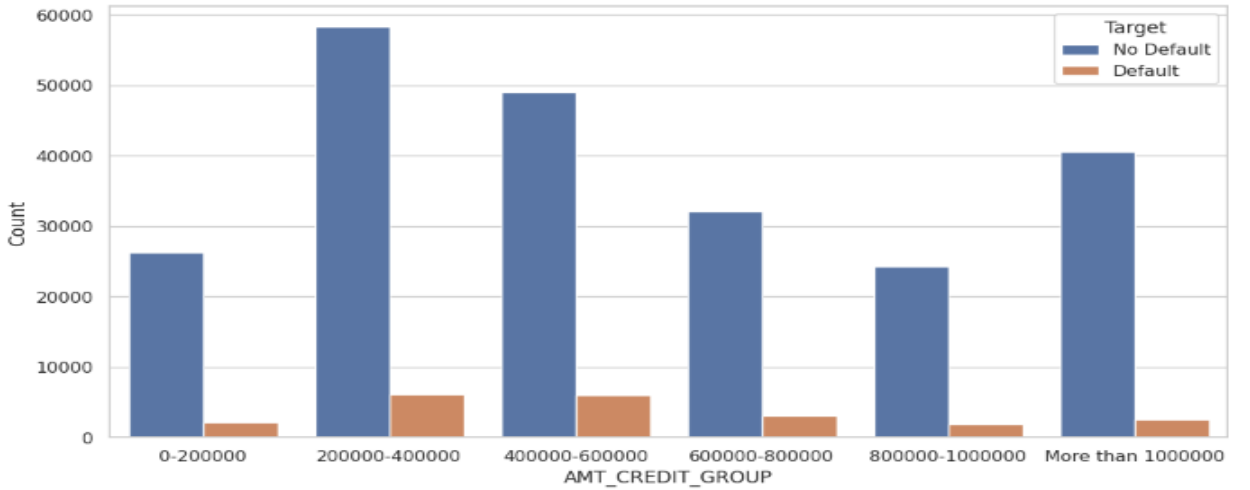


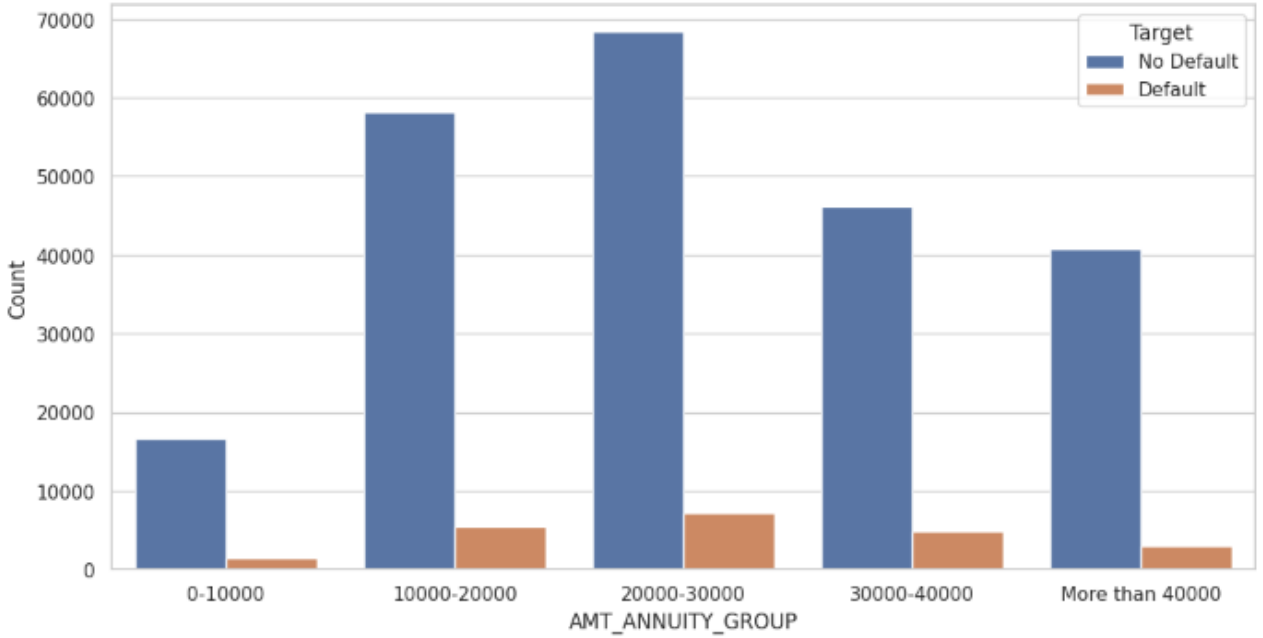


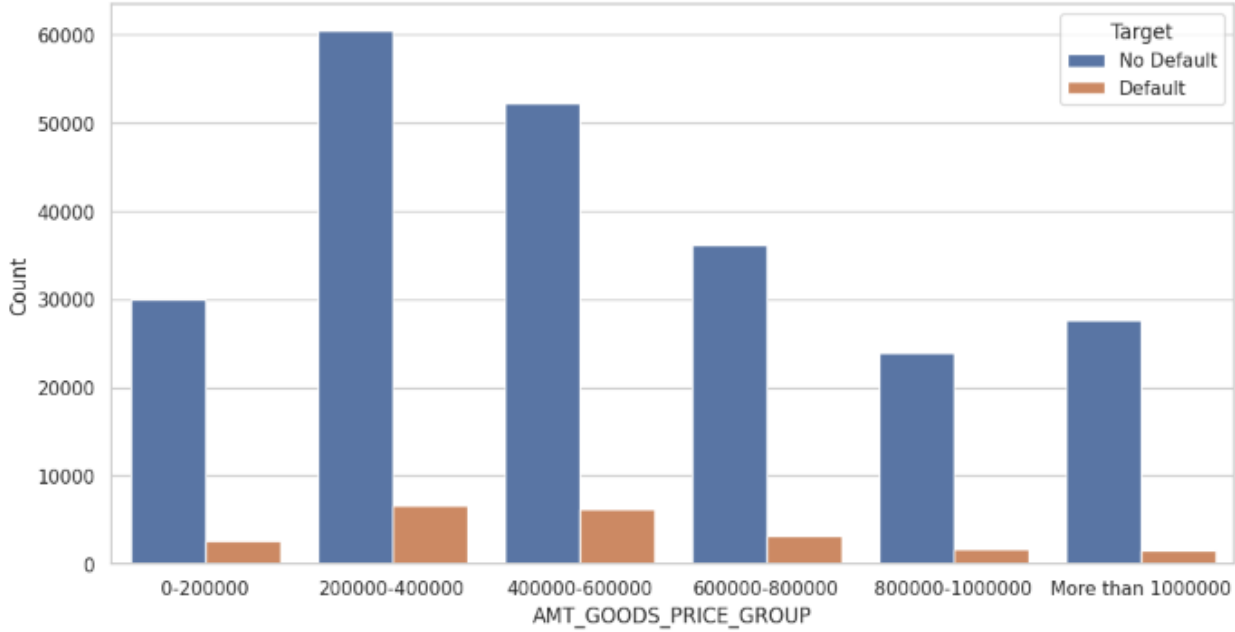


- **Financial Group Representation:** Certain financial groups are represented in both Target 0 (Non-Defaulters) and Target 1 (Defaulters).
- **Density Plot Analysis:** Utilizing density plots for both Target 0 (Non-Defaulters) and Target 1 (Defaulters), we can perform a comprehensive analysis of these groups.









- **Common Financial Group Representation:** Certain financial groups are shared between Target 0 (Non-Defaulters) and Target 1 (Defaulters).
- **Count Plot Analysis:** We've utilized count plots for both target groups to streamline data examination, enabling us to assess differences in the number of defaulters.

Plotted several bivariate graphs to conduct our analyses

Default Rates by Education and Gender:

- The highest default rates are observed among males with lower secondary education, followed by those with secondary/secondary special education.
- Default rates are notably higher among males compared to females, particularly within the age range of 20 to 30 years.

Total Income vs. Gender:

- Looking at the 'AMT_INCOME_TOTAL_GROUP' vs. Gender graph, males exhibit a higher default rate compared to females, especially in the income range of 0 to 60,000.

Credit Amount vs. Gender:

- Analyzing the 'AMT_CREDIT_GROUP' vs. Gender graph, a similar trend emerges with higher default rates among males, particularly in the credit amount range of 400,000 to 600,000.

Loan Annuity vs. Gender:

- In the 'AMT_ANNUITY_GROUP' vs. Gender graph, once again, males show a higher propensity for default, particularly within the loan annuity range of 10,000 to 40,000.

Goods Price vs. Gender:

- Shifting focus to the 'AMT_GOODS_PRICE_GROUP' vs. Gender graph, males consistently exhibit higher default rates, especially in the goods price range of 200,000 to 600,000.

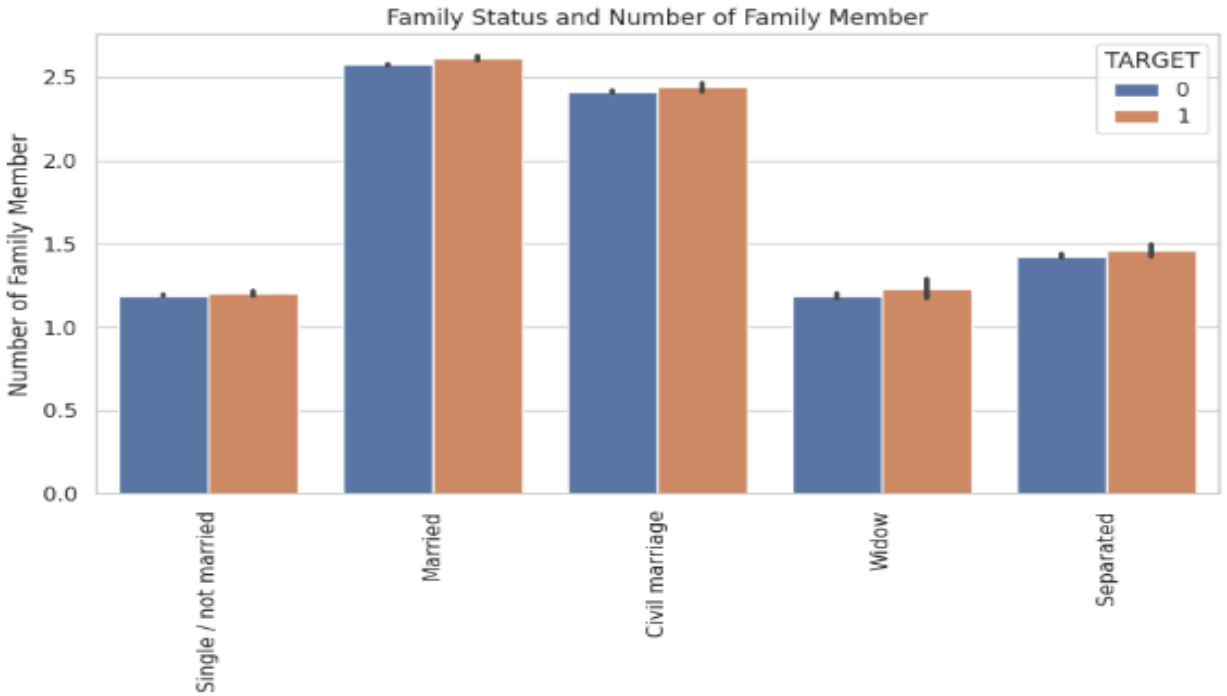
Family Size and Default Risk:

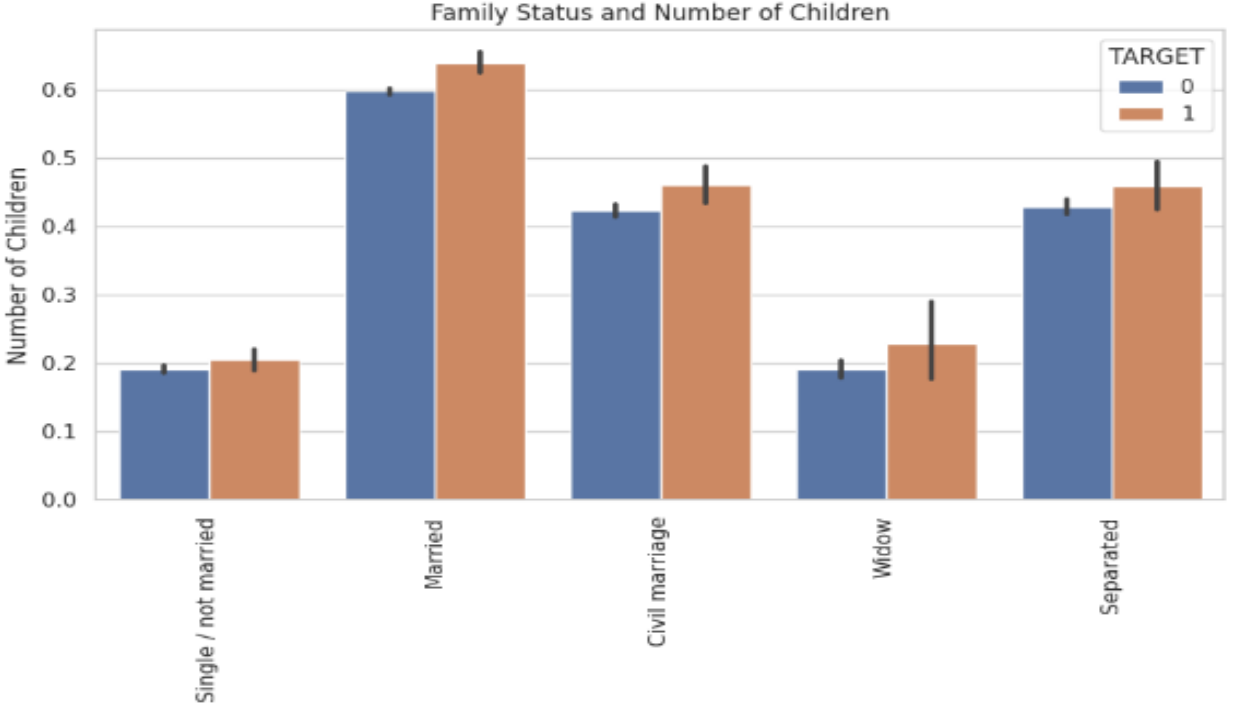
- Married clients with five or more children face an elevated risk of default, likely due to economic strains associated with supporting multiple dependents.
- Married clients with larger family sizes exhibit increased risks in both defaulting and non-defaulting categories, possibly due to the impact of additional family members on their economic situation.

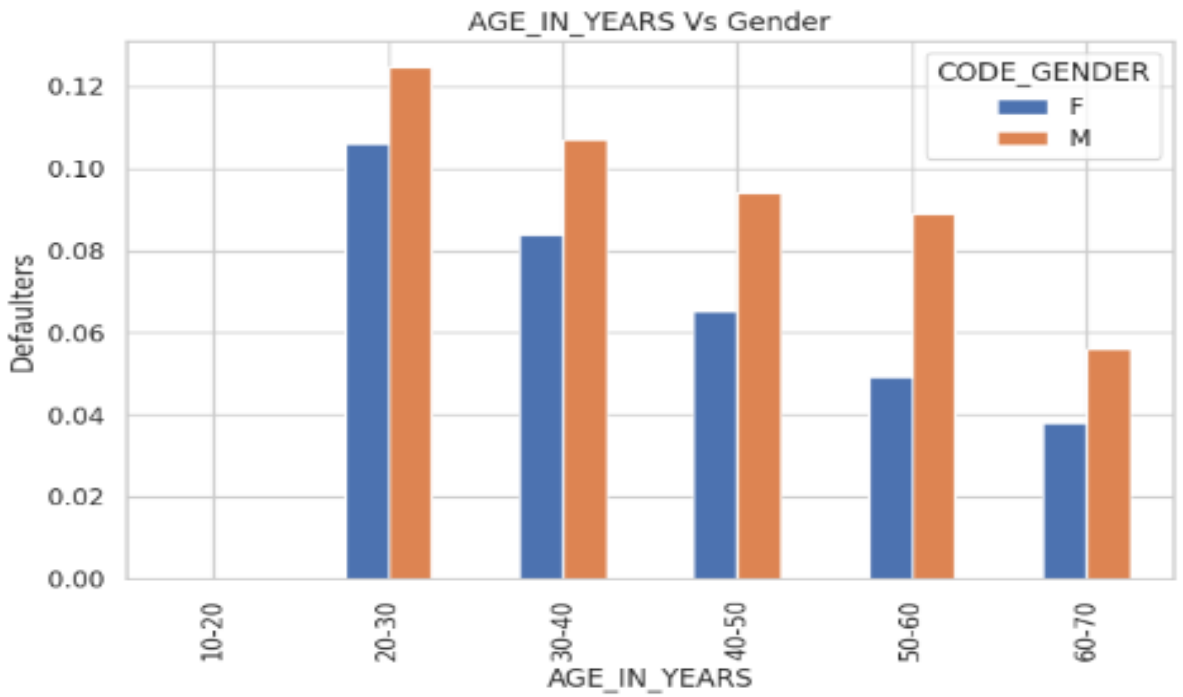
Income Source and Family Size:

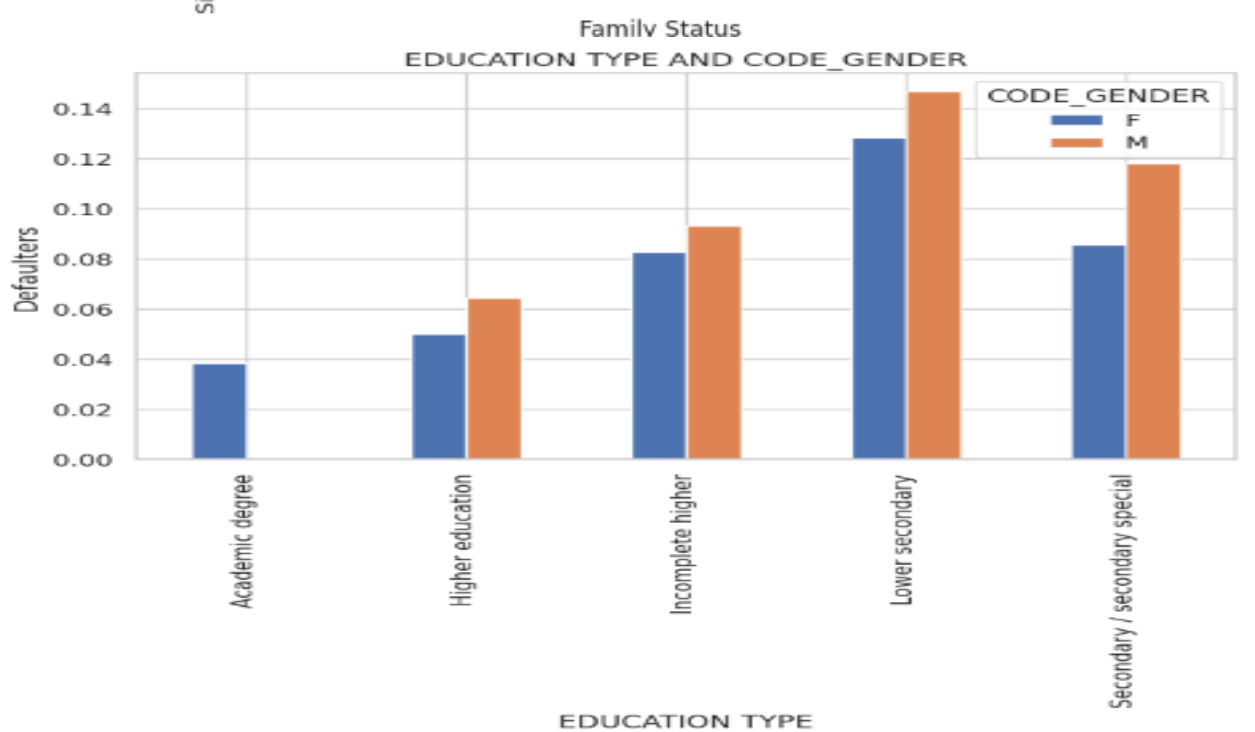
- Individuals receiving income through Maternity Leave tend to have a higher likelihood of default when they have a greater number of family members.
- Similarly, individuals receiving income via Maternity Leave have a higher tendency to default when they have a larger number of children.

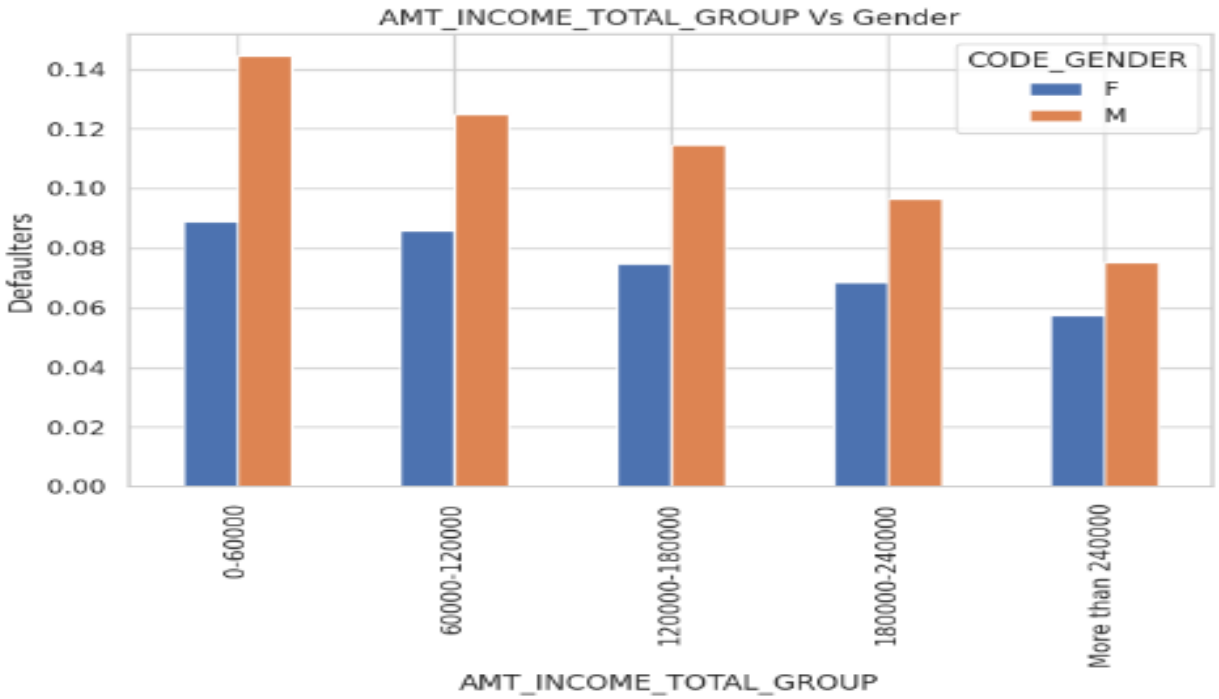
Conclusion: Across various financial aspects such as income, credit, loan annuity, and goods price, males consistently exhibit higher default rates compared to females. This trend holds true across different income and expense ranges. Additionally, family size, especially in cases of marriage and larger families, is associated with an increased risk of default.

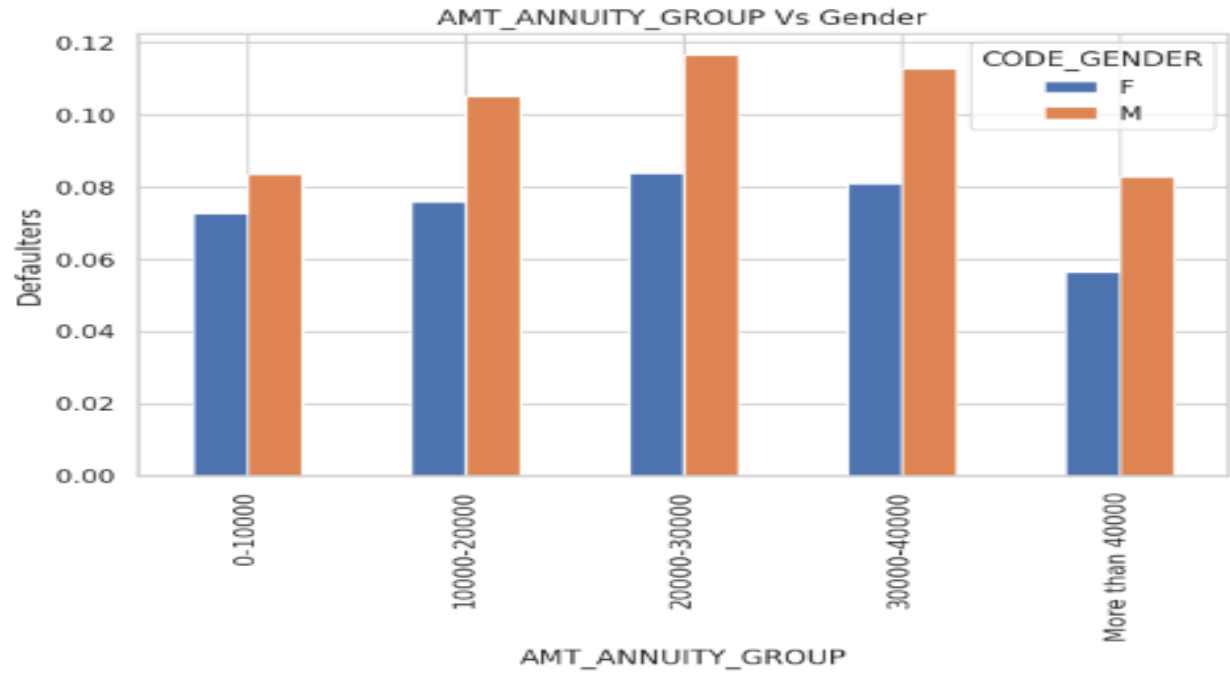

Family Status and Number of Family Member
Number of Family Member
TARGET
0
1
0.0
0.5
1.0
1.5
2.0
2.5
Single / not married
Married
Civil marriage
Widow
Separated
Family Status


Family Status and Number of Children
Number of Children
TARGET
0
1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Single / not married
Married
Civil marriage
Widow
Separated

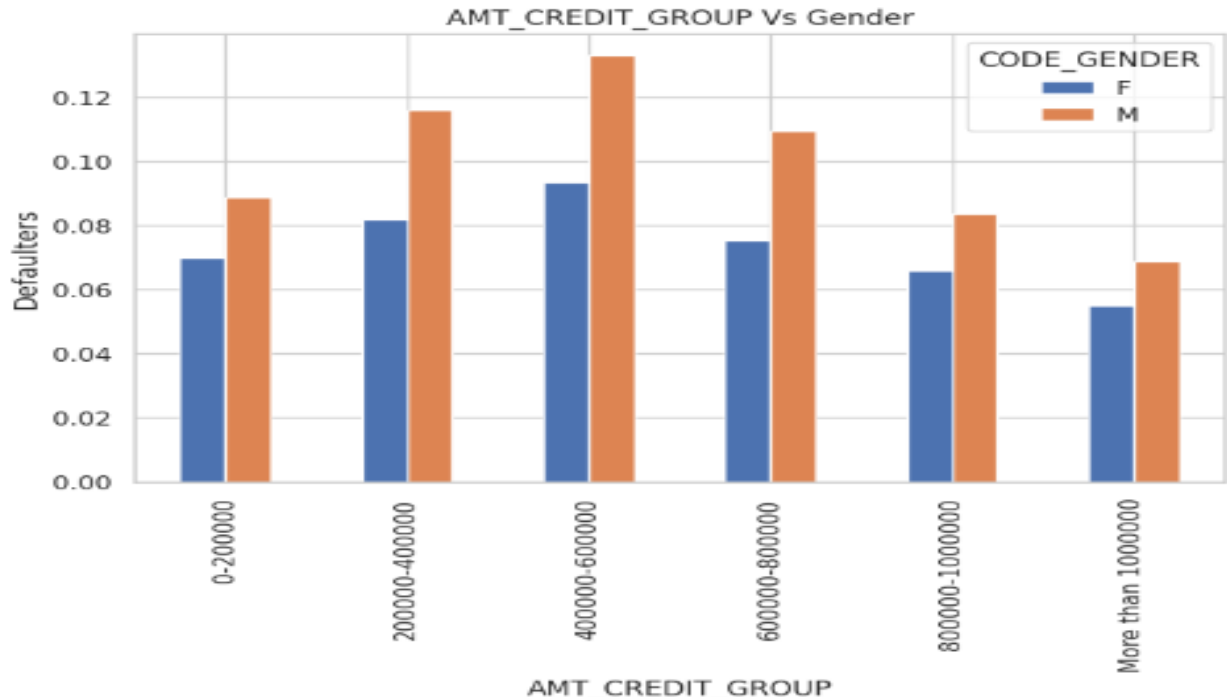

AGE_IN_YEARS Vs Gender
Defaulters
CODE_GENDER
F
M
0.00
0.02
0.04
0.06
0.08
0.10
0.12
10-20
20-30
30-40
40-50
50-60
60-70
AGE_IN_YEARS


Family Status
EDUCATION TYPE AND CODE_GENDER
Defaulters
CODE_GENDER
F
M
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
EDUCATION TYPE


AMT_INCOME_TOTAL_GROUP Vs Gender
CODE_GENDER
F
M
Defaulters
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0-60000
60000-120000
120000-180000
180000-240000
More than 240000
AMT_INCOME_TOTAL_GROUP


AMT_ANNUITY_GROUP Vs Gender
CODE_GENDER
F
M
Defaulters
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0-10000
10000-20000
20000-30000
30000-40000
More than 40000
AMT_ANNUITY_GROUP


AMT_GOODS_PRICE_GROUP Vs Gender
CODE_GENDER
F
M
Defaulters
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0-200000
200000-400000
400000-600000
600000-800000
800000-1000000
More than 1000000
AMT_GOODS_PRICE_GROUP


AMT_CREDIT_GROUP Vs Gender
CODE_GENDER
F
M
Defaulters
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0-200000
200000-400000
400000-600000
600000-800000
800000-1000000
More than 1000000
AMT_CREDIT_GROUP

Correlation of Non Defaulters

Highly Correlated Variables:

- There's a strong correlation (0.98) between AMT_CREDIT and AMT_GOODS_PRICE.
- REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT are highly correlated (0.95).
- CNT_FAM_MEMBERS and CNT_CHILDREN show a notable correlation (0.85).
- AMT_ANNUITY and AMT_CREDIT exhibit a strong correlation (0.82).

Relationships:

- Higher AMT_CREDIT is associated with clients who have fewer children, while lower AMT_CREDIT is associated with those having more children.
- Increased income is correlated with clients having fewer children, while lower income is linked to clients with more children.
- Higher income tends to be associated with clients having fewer family members, whereas lower income is related to clients with larger families.
- In densely populated areas, clients tend to have fewer children, as CNT_CHILDREN is inversely proportional to REGION_POPULATION_RELATIVE.
- In densely populated areas, AMT_CREDIT tends to be higher, indicating a direct relationship between AMT_CREDIT and REGION_POPULATION_RELATIVE.
- AMT_INCOME_TOTAL is typically higher in densely populated areas, suggesting an inverse relationship between income and REGION_POPULATION_RELATIVE.


Correlation for Target 0
CNT_CHILDREN 1 -0.022 -0.024 -0.011 -0.027 -0.031 0.0019 -0.012 0.0026 -0.034 0.01 0.85 0.036 0.034 -0.023 -0.015 -0.0055 0.0024 0.0075 0.013 -0.027 -0.012 -0.27 -0.03 -0.16 0.12
AMT_INCOME_TOTAL -0.022 1 0.4 0.47 0.4 0.1 -0.0037 -0.12 -0.0061 -0.0021 0.086 -0.012 -0.21 -0.24 0.077 0.16 0.15 -0.00084 -0.013 -0.0042 0.18 -0.055 0.054 0.059 -0.039 0.037
AMT_CREDIT -0.024 0.4 1 0.82 0.98 0.057 -0.0027 -0.054 0.033 0.023 0.0069 0.046 -0.09 -0.099 0.015 0.041 0.044 -0.041 -0.037 -0.011 0.14 0.041 0.17 0.1 0.017 0.038
AMT_ANNUITY -0.011 0.47 0.82 1 0.82 0.061 -0.0021 -0.062 0.032 0.0099 0.057 0.053 -0.12 -0.13 0.033 0.07 0.069 -0.02 -0.024 -0.0079 0.13 0.033 0.1 0.074 -0.01 0.027
AMT_GOODS_PRICE -0.027 0.4 0.98 0.82 1 0.066 -0.0028 -0.032 0.029 0.039 0.0066 0.045 -0.093 -0.1 0.016 0.042 0.045 -0.042 -0.039 -0.013 0.15 0.046 0.17 0.11 0.016 0.04
REGION_POPULATION_RELATIVE -0.031 0.1 0.057 0.061 0.066 1 0.0029 -0.014 -0.0072 0.058 0.026 -0.022 -0.42 -0.43 -0.027 0.029 0.053 -0.049 -0.032 -0.005 0.19 -0.0077 0.042 0.00046 0.04 0.011
FLAG_EMP_PHONE 0.0019 -0.0037 -0.0027 -0.0021 -0.0028 0.0029 1 -0.00052 -0.00032 -0.0041 -0.00076 1.7e-05 0.0017 0.0027 0.00091 -0.00084 -0.0013 0.0021 0.0014 0.00057 -0.0031 -0.0018 -0.0024 -0.003 0.0024 -0.00054
FLAG_WORK_PHONE -0.012 -0.12 -0.054 -0.062 -0.032 -0.014 -0.00052 1 0.026 0.34 -0.029 0.0085 0.016 0.023 0.058 0.042 0.018 0.027 0.066 0.062 -0.024 -0.031 -0.034 0.014 -0.0064 0.019
FLAG_CONT_MOBILE 0.0026 -0.0061 0.033 0.032 0.029 -0.0072 -0.00032 0.026 1 0.0068 -0.004 0.0044 0.015 0.016 0.0032 -0.00061 -0.0024 0.00014 0.0064 0.0069 -0.0053 0.0093 0.0093 0.0072 0.0022 -0.002
FLAG_PHONE -0.034 -0.0021 0.023 0.0099 0.039 0.058 -0.0041 0.34 0.0068 1 0.017 -0.017 -0.087 -0.082 0.00074 0.0059 0.0071 -0.05 -0.047 -0.023 0.073 -0.00018 0.045 0.055 0.077 0.034
FLAG_EMAIL 0.01 0.086 0.0069 0.057 0.0066 0.026 -0.00076 -0.029 -0.004 0.017 1 0.0034 -0.055 -0.054 0.018 0.026 0.022 0.0099 -0.012 -0.019 0.022 -0.045 -0.067 -0.029 -0.024 -0.01
CNT_FAM_MEMBERS 0.85 -0.012 0.046 0.053 0.045 -0.022 1.7e-05 0.0085 0.0044 -0.017 0.0034 1 0.04 0.041 -0.028 -0.021 -0.0089 -0.0078 0.011 0.027 -0.0089 0.0054 -0.17 0.016 -0.14 0.11
REGION_RATING_CLIENT 0.036 -0.21 -0.09 -0.12 -0.093 -0.42 0.0017 0.016 0.015 -0.087 -0.055 0.04 1 0.95 -0.044 -0.15 -0.16 0.039 0.016 -0.015 -0.33 -0.0051 -0.035 0.016 -0.089 -0.0034
REGION_RATING_CLIENT_W_CITY 0.034 -0.24 -0.099 -0.13 -0.1 -0.43 0.0027 0.023 0.016 -0.082 -0.054 0.041 0.95 1 -0.04 -0.15 -0.15 0.051 0.04 0.0047 -0.32 -0.003 -0.035 0.014 -0.082 -0.00092
REG_REGION_NOT_LIVE_REGION -0.023 0.077 0.015 0.033 0.016 -0.027 0.00091 0.058 0.0032 0.00074 0.018 -0.028 -0.044 -0.04 1 0.46 0.09 0.34 0.14 0.0034 0.023 -0.041 -0.059 -0.065 -0.018 -0.027
REG_REGION_NOT_WORK_REGION -0.015 0.16 0.041 0.07 0.042 0.029 -0.00084 0.042 -0.00061 0.0059 0.026 -0.021 -0.15 -0.15 0.46 1 0.86 0.15 0.22 0.18 0.045 -0.036 -0.038 -0.087 -0.015 -0.021
LIVE_REGION_NOT_WORK_REGION -0.0055 0.15 0.044 0.069 0.045 0.053 -0.0013 0.018 -0.0024 0.0071 0.022 -0.0089 -0.16 -0.15 0.09 0.86 1 0.015 0.17 0.22 0.044 -0.019 -0.013 -0.063 -0.0077 -0.0086
REG_CITY_NOT_LIVE_CITY 0.0024 -0.00084 -0.041 -0.02 -0.042 -0.049 0.0021 0.027 0.00014 -0.05 0.0099 -0.0078 0.039 0.051 0.34 0.15 0.015 1 0.44 0.012 -0.044 -0.062 -0.17 -0.12 -0.039 -0.055
REG_CITY_NOT_WORK_CITY 0.0075 -0.013 -0.037 -0.024 -0.039 -0.032 0.0014 0.066 0.0064 -0.047 -0.012 0.011 0.016 0.04 0.14 0.22 0.17 0.44 1 0.82 -0.093 -0.046 -0.11 -0.13 -0.047 -0.034
LIVE_CITY_NOT_WORK_CITY 0.013 -0.0042 -0.011 -0.0079 -0.013 -0.005 0.00057 0.062 0.0069 -0.023 -0.019 0.027 -0.015 0.0047 0.0034 0.18 0.22 0.012 0.82 1 -0.075 -0.015 -0.029 -0.07 -0.027 -0.0017
EXT_SOURCE_2 -0.027 0.18 0.14 0.13 0.15 0.19 -0.0031 -0.024 -0.0053 0.073 0.022 -0.0089 -0.33 -0.32 0.023 0.045 0.044 -0.044 -0.093 -0.075 1 0.086 0.15 0.099 0.073 0.06
EXT_SOURCE_3 -0.012 -0.055 0.041 0.033 0.046 -0.0077 -0.0018 -0.031 0.0093 -0.00018 -0.045 0.0054 -0.0051 -0.003 -0.041 -0.036 -0.019 -0.062 -0.046 -0.015 0.086 1 0.16 0.12 0.078 0.094
AGE_IN_YEARS -0.27 0.054 0.17 0.1 0.17 0.042 -0.0024 -0.034 0.0093 0.045 -0.067 -0.17 -0.035 -0.035 -0.059 -0.038 -0.013 -0.17 -0.11 -0.029 0.15 0.16 1 0.31 0.27 0.082
YEAR_EMPLOYED -0.03 0.059 0.1 0.074 0.11 0.00046 -0.003 0.014 0.0072 0.055 -0.029 0.016 0.016 0.014 -0.065 -0.087 -0.063 -0.12 -0.13 -0.07 0.099 0.12 0.31 1 0.13 0.11
YEAR_REGISTRATION -0.16 -0.039 0.017 -0.01 0.016 0.04 0.0024 -0.0064 0.0022 0.077 -0.024 -0.14 -0.089 -0.082 -0.018 -0.015 -0.0077 -0.039 -0.047 -0.027 0.073 0.078 0.27 0.13 1 0.036
YEAR_ID_PUBLISH 0.12 0.037 0.038 0.027 0.04 0.011 -0.00054 0.019 -0.002 0.034 -0.01 0.11 -0.0034 -0.00092 -0.027 -0.021 -0.0086 -0.055 -0.034 -0.0017 0.06 0.094 0.082 0.11 0.036 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
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
EXT_SOURCE_2
EXT_SOURCE_3
AGE_IN_YEARS
YEAR_EMPLOYED
YEAR_REGISTRATION
YEAR_ID_PUBLISH
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

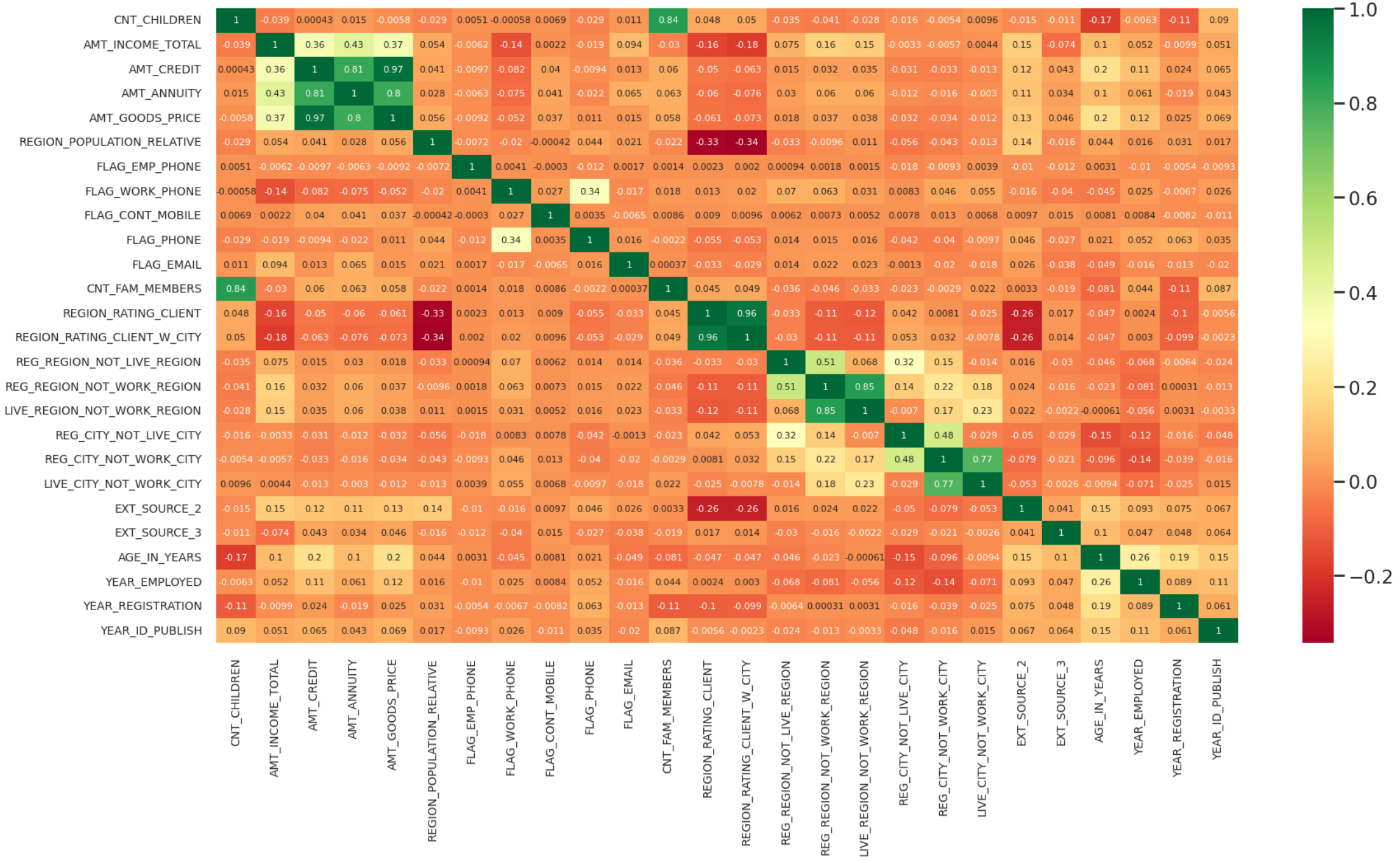
Correlation of Defaulters

Highly Correlated Variables:

- AMT_CREDIT and AMT_GOODS_PRICE display a strong correlation of 0.97.
- REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT are highly correlated, with a coefficient of 0.96.
- CNT_FAM_MEMBERS and CNT_CHILDREN show a notable correlation, registering at 0.84.
- AMT_ANNUITY and AMT_CREDIT exhibit a strong correlation of 0.81.

Relationship Insights:

- Clients with higher AMT_CREDIT tend to have fewer children, whereas those with lower AMT_CREDIT tend to have more children.
- Increased income correlates with clients having fewer children, while lower income correlates with clients having more children.
- Higher income is linked to clients having fewer family members, while lower income is associated with clients having more family members.
- In densely populated areas, clients tend to have fewer children, as CNT_CHILDREN inversely correlates with REGION_POPULATION_RELATIVE.
- Densely populated areas show a tendency for higher AMT_CREDIT, suggesting a direct relationship between AMT_CREDIT and REGION_POPULATION_RELATIVE.
- AMT_INCOME_TOTAL tends to be higher in densely populated areas, indicating an inverse relationship between income and REGION_POPULATION_RELATIVE.


Correlation for Target 1
CNT_CHILDREN 1 -0.039 0.00043 0.015 -0.0058 -0.029 0.0051 -0.00058 0.0069 -0.029 0.011 0.84 0.048 0.05 -0.035 -0.041 -0.028 -0.016 -0.0054 0.0096 -0.015 -0.011 -0.17 -0.0063 -0.11 0.09
AMT_INCOME_TOTAL -0.039 1 0.36 0.43 0.37 0.054 -0.0062 -0.14 0.0022 -0.019 0.094 -0.03 -0.16 -0.18 0.075 0.16 0.15 -0.0033 -0.0057 0.0044 0.15 -0.074 0.1 0.052 -0.0099 0.051
AMT_CREDIT 0.00043 0.36 1 0.81 0.97 0.041 -0.0097 -0.082 0.04 -0.0094 0.013 0.06 -0.05 -0.063 0.015 0.032 0.035 -0.031 -0.033 -0.013 0.12 0.043 0.2 0.11 0.024 0.065
AMT_ANNUITY 0.015 0.43 0.81 1 0.8 0.028 -0.0063 -0.075 0.041 -0.022 0.065 0.063 -0.06 -0.076 0.03 0.06 0.06 -0.012 -0.016 -0.003 0.11 0.034 0.1 0.061 -0.019 0.043
AMT_GOODS_PRICE -0.0058 0.37 0.97 0.8 1 0.056 -0.0092 -0.052 0.037 0.011 0.015 0.058 -0.061 -0.073 0.018 0.037 0.038 -0.032 -0.034 -0.012 0.13 0.046 0.2 0.12 0.025 0.069
REGION_POPULATION_RELATIVE -0.029 0.054 0.041 0.028 0.056 1 -0.0072 -0.02 -0.00042 0.044 0.021 -0.022 -0.33 -0.34 -0.033 -0.0096 0.011 -0.056 -0.043 -0.013 0.14 -0.016 0.044 0.016 0.031 0.017
FLAG_EMP_PHONE 0.0051 -0.0062 -0.0097 -0.0063 -0.0092 -0.0072 1 0.0041 -0.0003 -0.012 0.0017 0.0014 0.0023 0.002 0.00094 0.0018 0.0015 -0.018 -0.0093 0.0039 -0.01 -0.012 0.0031 -0.01 -0.0054 -0.0093
FLAG_WORK_PHONE -0.00058 -0.14 -0.082 -0.075 -0.052 -0.02 0.0041 1 0.027 0.34 -0.017 0.018 0.013 0.02 0.07 0.063 0.031 0.0083 0.046 0.055 -0.016 -0.04 -0.045 0.025 -0.0067 0.026
FLAG_CONT_MOBILE 0.0069 0.0022 0.04 0.041 0.037 -0.00042 -0.0003 0.027 1 0.0035 -0.0065 0.0086 0.009 0.0096 0.0062 0.0073 0.0052 0.0078 0.013 0.0068 0.0097 0.015 0.0081 0.0084 -0.0082 -0.011
FLAG_PHONE -0.029 -0.019 -0.0094 -0.022 0.011 0.044 -0.012 0.34 0.0035 1 0.016 -0.0022 -0.055 -0.053 0.014 0.015 0.016 -0.042 -0.04 -0.0097 0.046 -0.027 0.021 0.052 0.063 0.035
FLAG_EMAIL 0.011 0.094 0.013 0.065 0.015 0.021 0.0017 -0.017 -0.0065 0.016 1 0.00037 -0.033 -0.029 0.014 0.022 0.023 -0.0013 -0.02 -0.018 0.026 -0.038 -0.049 -0.016 -0.013 -0.02
CNT_FAM_MEMBERS 0.84 -0.03 0.06 0.063 0.058 -0.022 0.0014 0.018 0.0086 -0.0022 0.00037 1 0.045 0.049 -0.036 -0.046 -0.033 -0.023 -0.0029 0.022 0.0033 -0.019 -0.081 0.044 -0.11 0.087
REGION_RATING_CLIENT 0.048 -0.16 -0.05 -0.06 -0.061 -0.33 0.0023 0.013 0.009 -0.055 -0.033 0.045 1 0.96 -0.033 -0.11 -0.12 0.042 0.0081 -0.025 -0.26 0.017 -0.047 0.0024 -0.1 -0.0056
REGION_RATING_CLIENT_W_CITY 0.05 -0.18 -0.063 -0.076 -0.073 -0.34 0.002 0.02 0.0096 -0.053 -0.029 0.049 0.96 1 -0.03 -0.11 -0.11 0.053 0.032 -0.0078 -0.26 0.014 -0.047 0.003 -0.099 -0.0023
REG_REGION_NOT_LIVE_REGION -0.035 0.075 0.015 0.03 0.018 -0.033 0.00094 0.07 0.0062 0.014 0.014 -0.036 -0.033 -0.03 1 0.51 0.068 0.32 0.15 -0.014 0.016 -0.03 -0.046 -0.068 -0.0064 -0.024
REG_REGION_NOT_WORK_REGION -0.041 0.16 0.032 0.06 0.037 -0.0096 0.0018 0.063 0.0073 0.015 0.022 -0.046 -0.11 -0.11 0.51 1 0.85 0.14 0.22 0.18 0.024 -0.016 -0.023 -0.081 0.00031 -0.013
LIVE_REGION_NOT_WORK_REGION -0.028 0.15 0.035 0.06 0.038 0.011 0.0015 0.031 0.0052 0.016 0.023 -0.033 -0.12 -0.11 0.068 0.85 1 -0.007 0.17 0.23 0.022 -0.0022 -0.00061 -0.056 0.0031 -0.0033
REG_CITY_NOT_LIVE_CITY -0.016 -0.0033 -0.031 -0.012 -0.032 -0.056 -0.018 0.0083 0.0078 -0.042 -0.0013 -0.023 0.042 0.053 0.32 0.14 -0.007 1 0.48 -0.029 -0.05 -0.029 -0.15 -0.12 -0.016 -0.048
REG_CITY_NOT_WORK_CITY -0.0054 -0.0057 -0.033 -0.016 -0.034 -0.043 -0.0093 0.046 0.013 -0.04 -0.02 -0.0029 0.0081 0.032 0.15 0.22 0.17 0.48 1 0.77 -0.079 -0.021 -0.096 -0.14 -0.039 -0.016
LIVE_CITY_NOT_WORK_CITY 0.0096 0.0044 -0.013 -0.003 -0.012 -0.013 0.0039 0.055 0.0068 -0.0097 -0.018 0.022 -0.025 -0.0078 -0.014 0.18 0.23 -0.029 0.77 1 -0.053 -0.0026 -0.0094 -0.071 -0.025 0.015
EXT_SOURCE_2 -0.015 0.15 0.12 0.11 0.13 0.14 -0.01 -0.016 0.0097 0.046 0.026 0.0033 -0.26 -0.26 0.016 0.024 0.022 -0.05 -0.079 -0.053 1 0.041 0.15 0.093 0.075 0.067
EXT_SOURCE_3 -0.011 -0.074 0.043 0.034 0.046 -0.016 -0.012 -0.04 0.015 -0.027 -0.038 -0.019 0.017 0.014 -0.03 -0.016 -0.0022 -0.029 -0.021 -0.0026 0.041 1 0.1 0.047 0.048 0.064
AGE_IN_YEARS -0.17 0.1 0.2 0.1 0.2 0.044 0.0031 -0.045 0.0081 0.021 -0.049 -0.081 -0.047 -0.047 -0.046 -0.023 -0.00061 -0.15 -0.096 -0.0094 0.15 0.1 1 0.26 0.19 0.15
YEAR_EMPLOYED -0.0063 0.052 0.11 0.061 0.12 0.016 -0.01 0.025 0.0084 0.052 -0.016 0.044 0.0024 0.003 -0.068 -0.081 -0.056 -0.12 -0.14 -0.071 0.093 0.047 0.26 1 0.089 0.11
YEAR_REGISTRATION -0.11 -0.0099 0.024 -0.019 0.025 0.031 -0.0054 -0.0067 -0.0082 0.063 -0.013 -0.11 -0.1 -0.099 -0.0064 0.00031 0.0031 -0.016 -0.039 -0.025 0.075 0.048 0.19 0.089 1 0.061
YEAR_ID_PUBLISH 0.09 0.051 0.065 0.043 0.069 0.017 -0.0093 0.026 -0.011 0.035 -0.02 0.087 -0.0056 -0.0023 -0.024 -0.013 -0.0033 -0.048 -0.016 0.015 0.067 0.064 0.15 0.11 0.061 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
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
EXT_SOURCE_2
EXT_SOURCE_3
AGE_IN_YEARS
YEAR_EMPLOYED
YEAR_REGISTRATION
YEAR_ID_PUBLISH
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

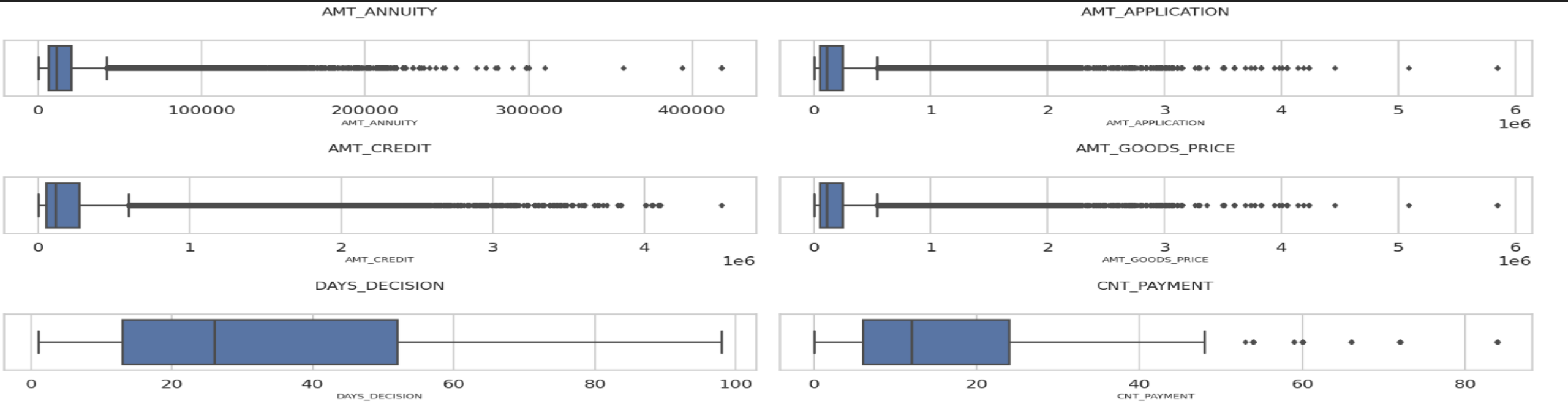
Previous application

Exploring Missing Data and Data Cleaning:

- We began by examining null values in each column and calculated the percentage of missing data.
- Columns with null values exceeding 30% were removed, while others with varying percentages of null values were assessed, ranging from 23% to 0.33%.
- Columns lacking significance were removed, and for the remaining null values, we opted for imputation using either the mean or mode as appropriate.
- Inspecting and addressing individual columns with missing data in a sequential manner.
- Additionally, we identified columns containing NAN or XNA entries and addressed them by either removing corresponding rows or replacing NAN, XNA, or XAP values with the mode.

Outlier Detection and Mitigation Using Capping and Winsorization:

- Initial observation revealed the presence of outliers in numerous numerical columns, prompting concern for their impact on data analysis.
- To identify these outliers, we created box plots for numerical columns, providing a visual representation of extreme values.
- To mitigate the influence of outliers, we implemented two techniques: capping and winsorization.
- Capping involved replacing extreme values exceeding a predefined threshold with the threshold value, limiting their effect on subsequent analyses.
- Winsorization, as an alternative approach, replaced extreme outliers with the nearest non-outlier values, preserving t overall data distribution.
- Box plots were included to visualize outlier presence, emphasizing the necessity of addressing these data points.
- Furthermore, we resolved negative values in specific columns by converting them into their absolute counterparts to rectify this issue.

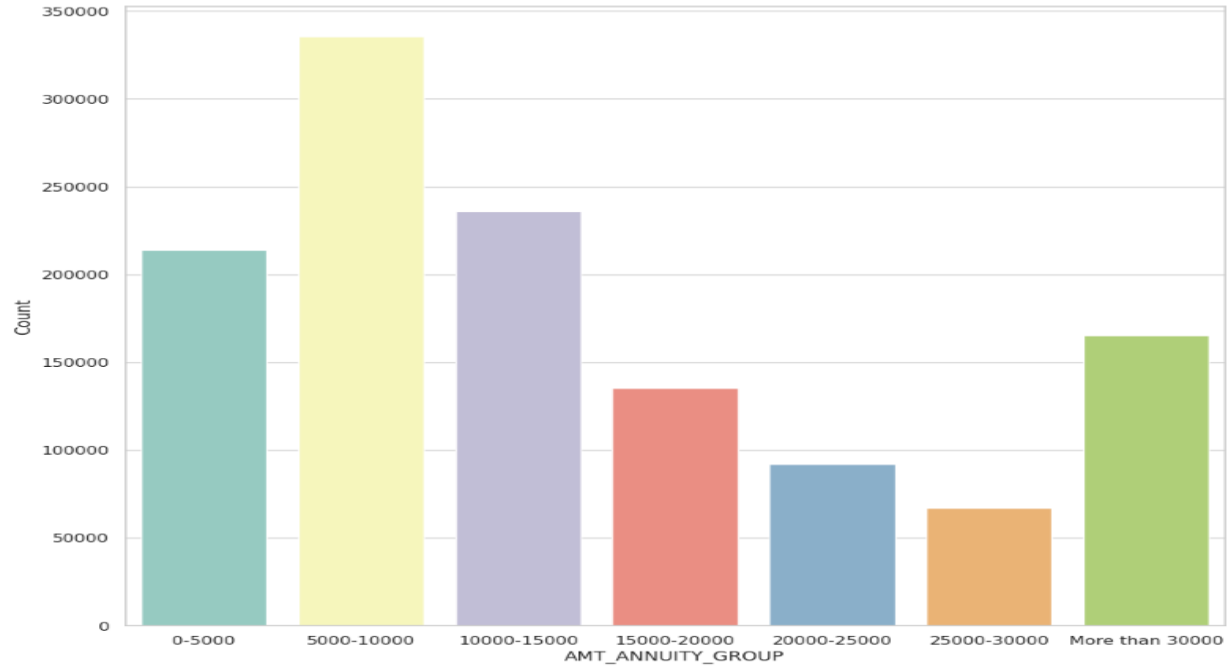


Exploratory Data Analysis Insights for Previous Application Data

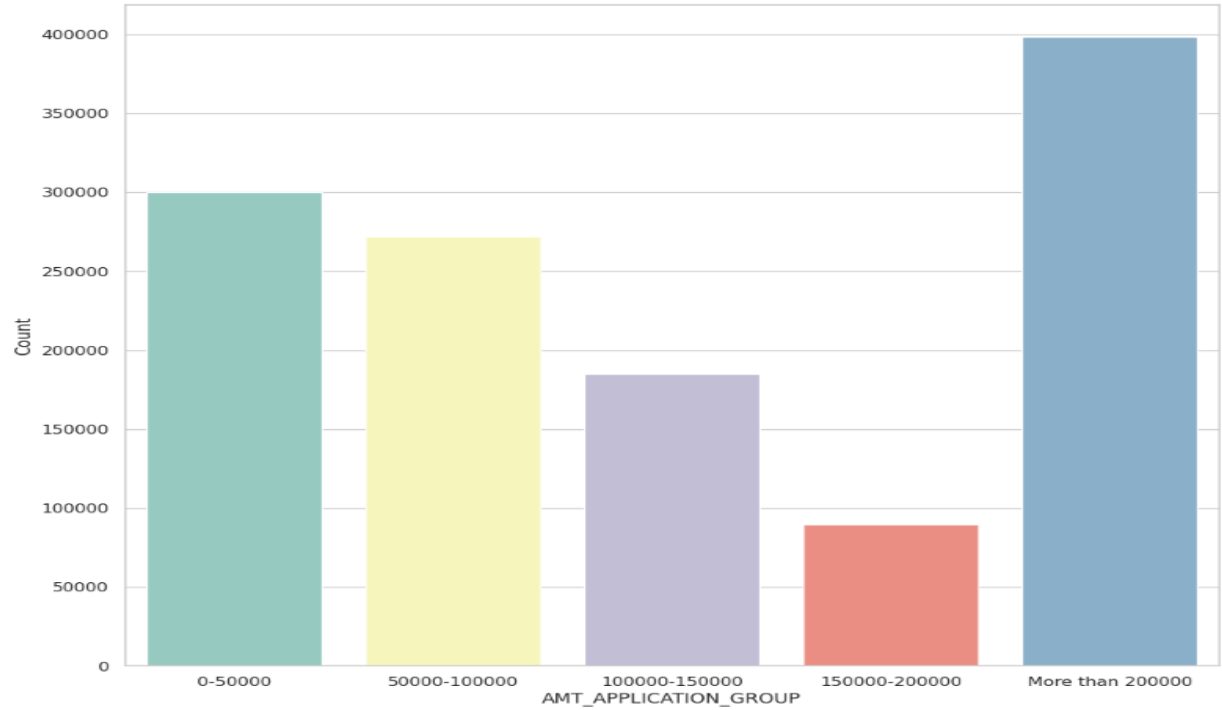
These observations were based on exploratory data analysis using various types of graphs for the previous application data.

- The graph revealed that 'DAYS_DECISION' exhibited a negative correlation ratio with other variables.
- Assumption: Renaming the column from 'DAYS_DECISION' to 'MONTHS_DECISION' was suggested.
- Utilized Various Graph Types for Analyzing Previous Application Data, Presented Below:

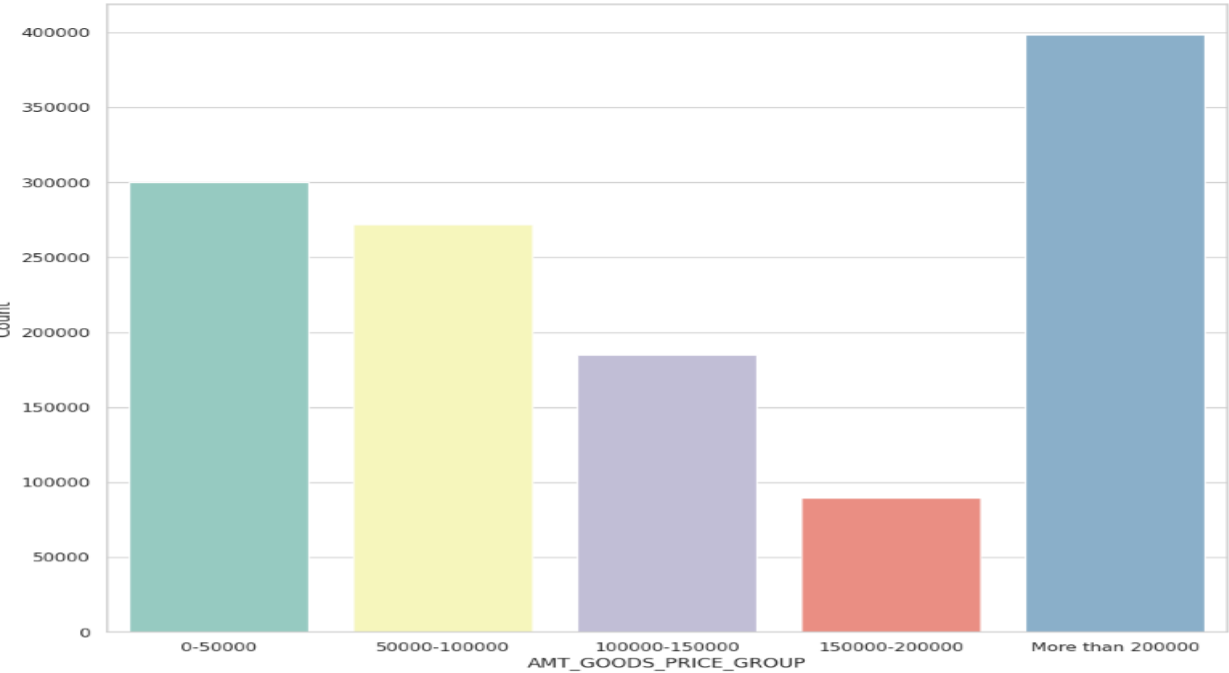
Count graph for AMT_ANNUITY_GROUP



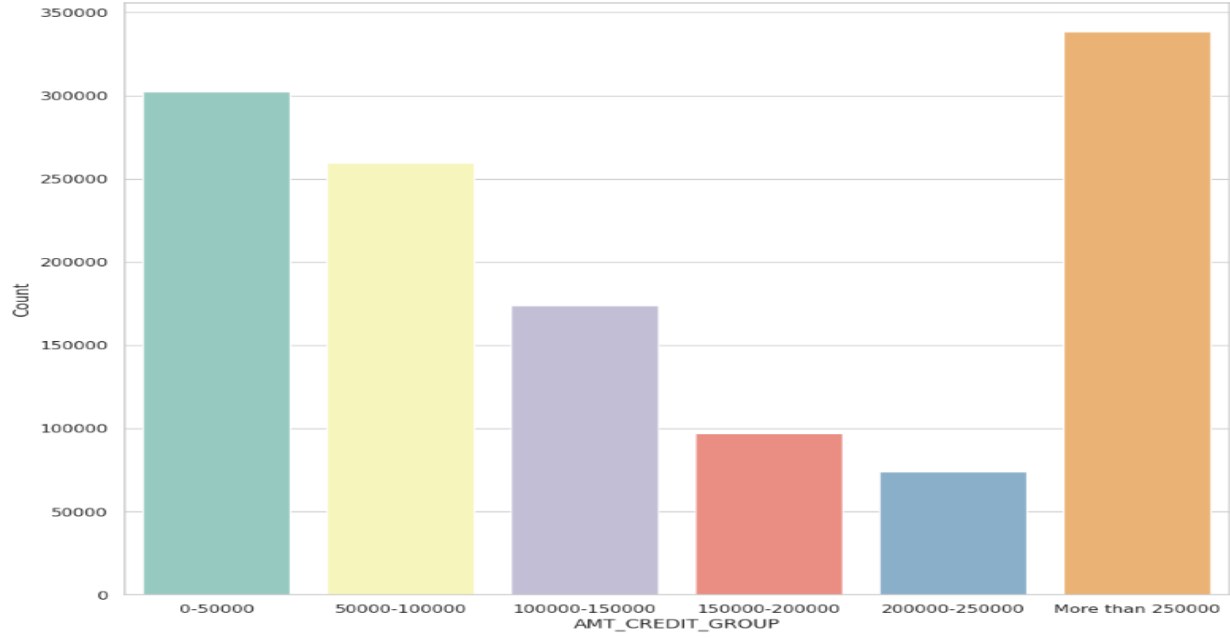
Count graph for AMT_APPLICATION_GROUP



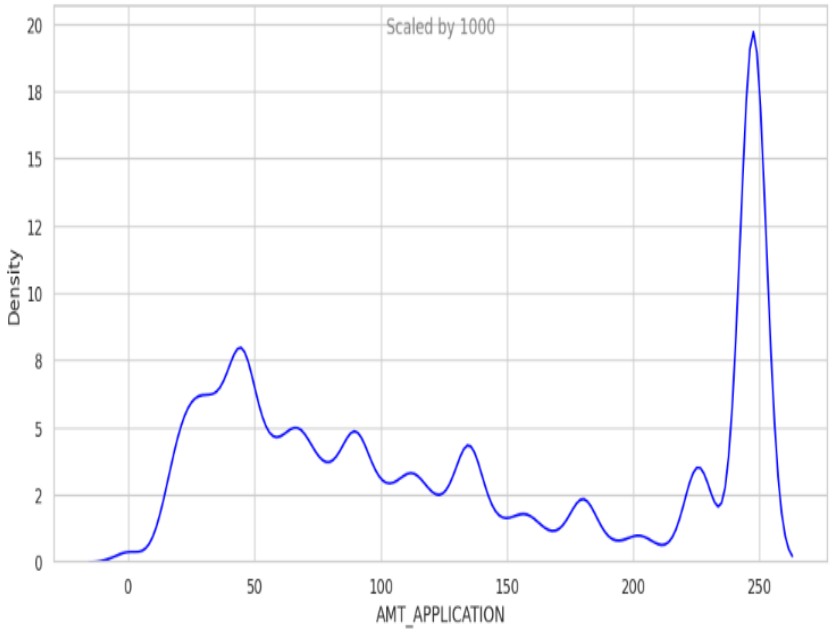
Count graph for AMT_GOODS_PRICE_GROUP



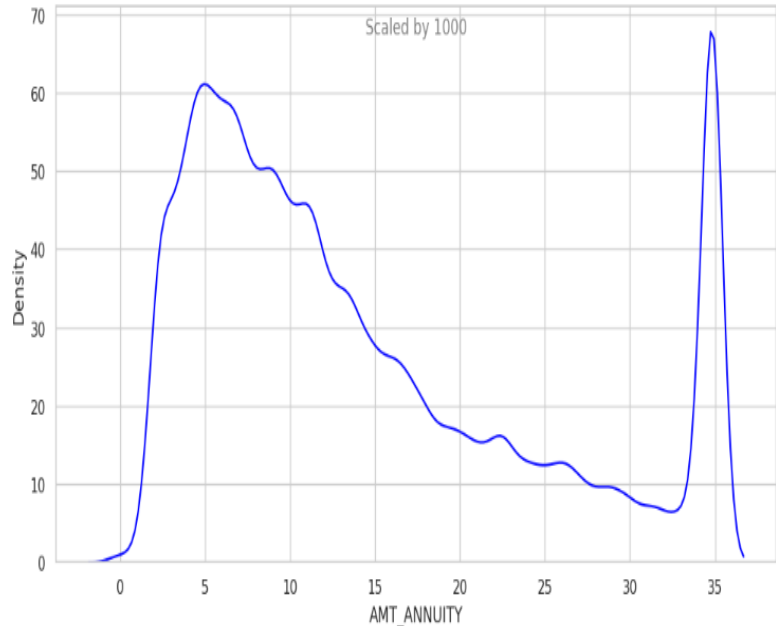
Count graph for AMT_CREDIT_GROUP



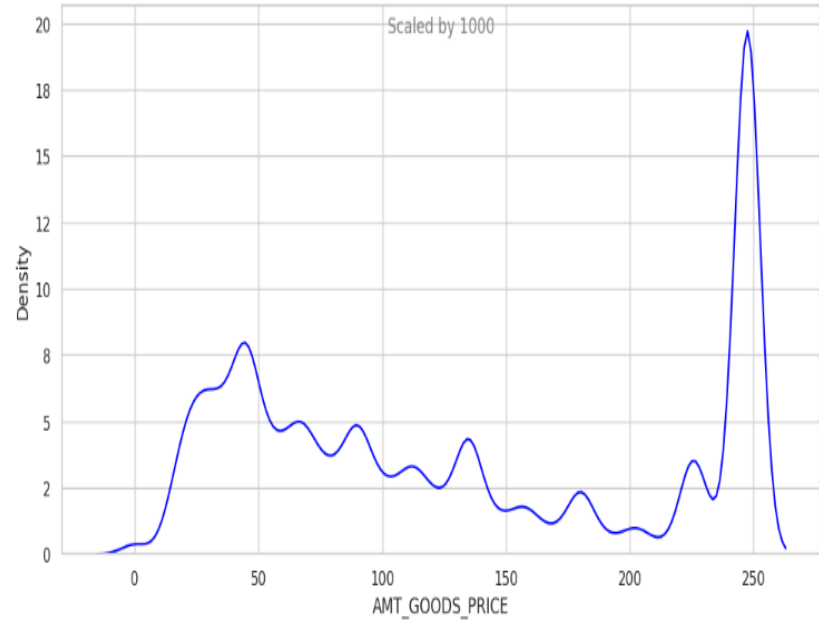
Density Plot for AMT_APPLICATION



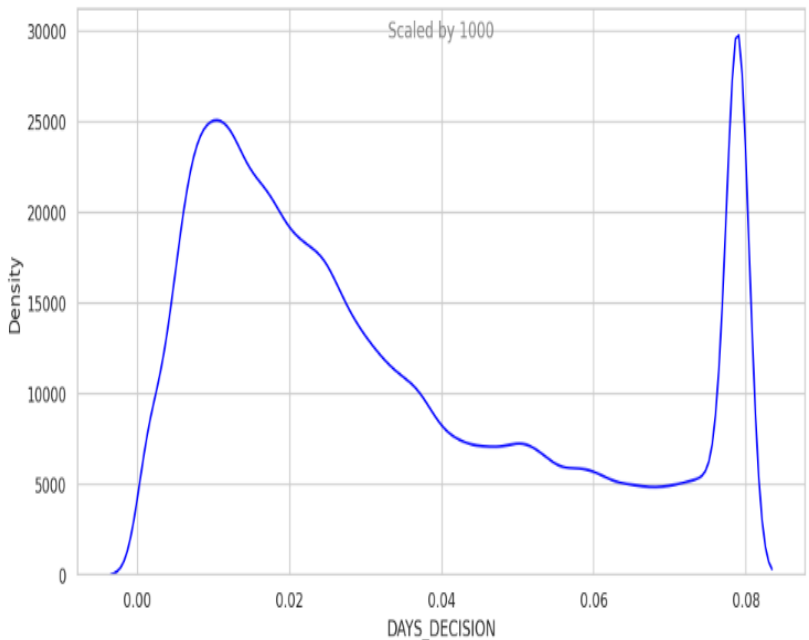
Density Plot for AMT_ANNUITY



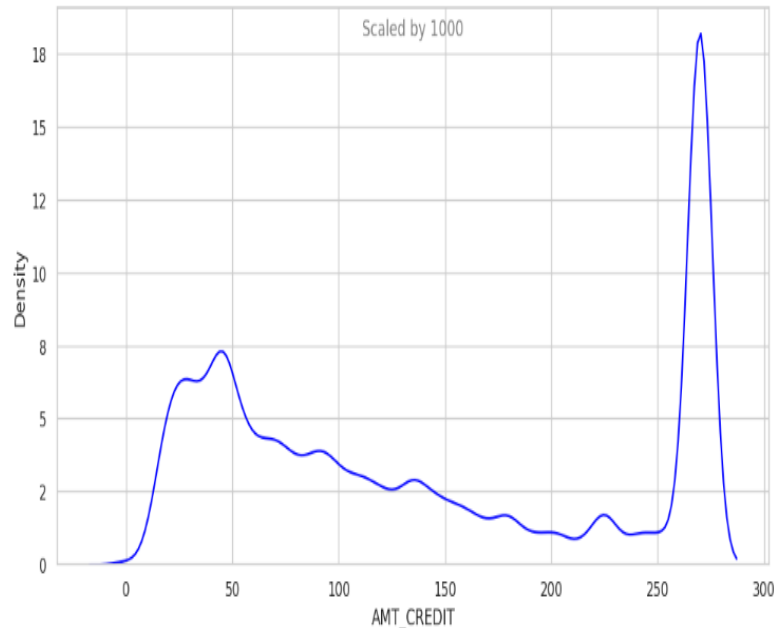
Density Plot for AMT_GOODS_PRICE



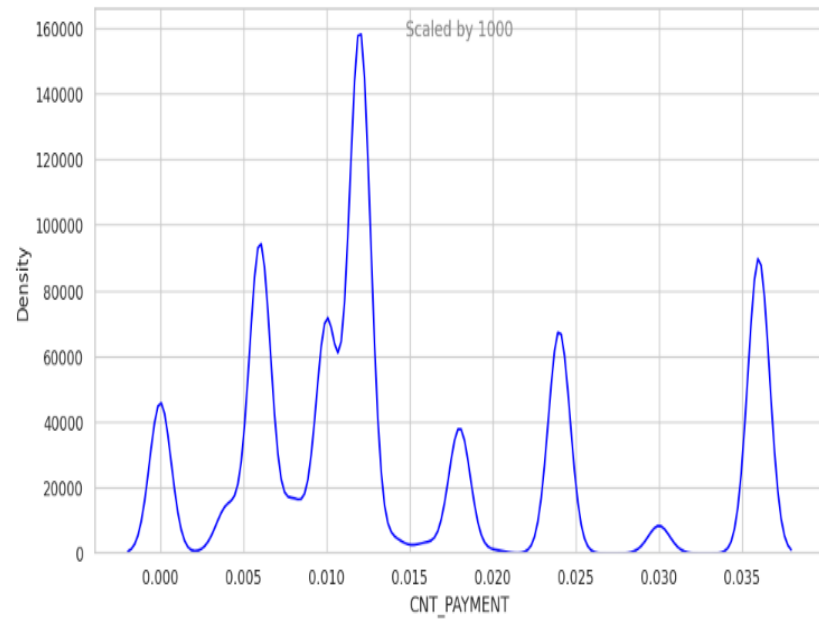
Density Plot for DAYS_DECISION

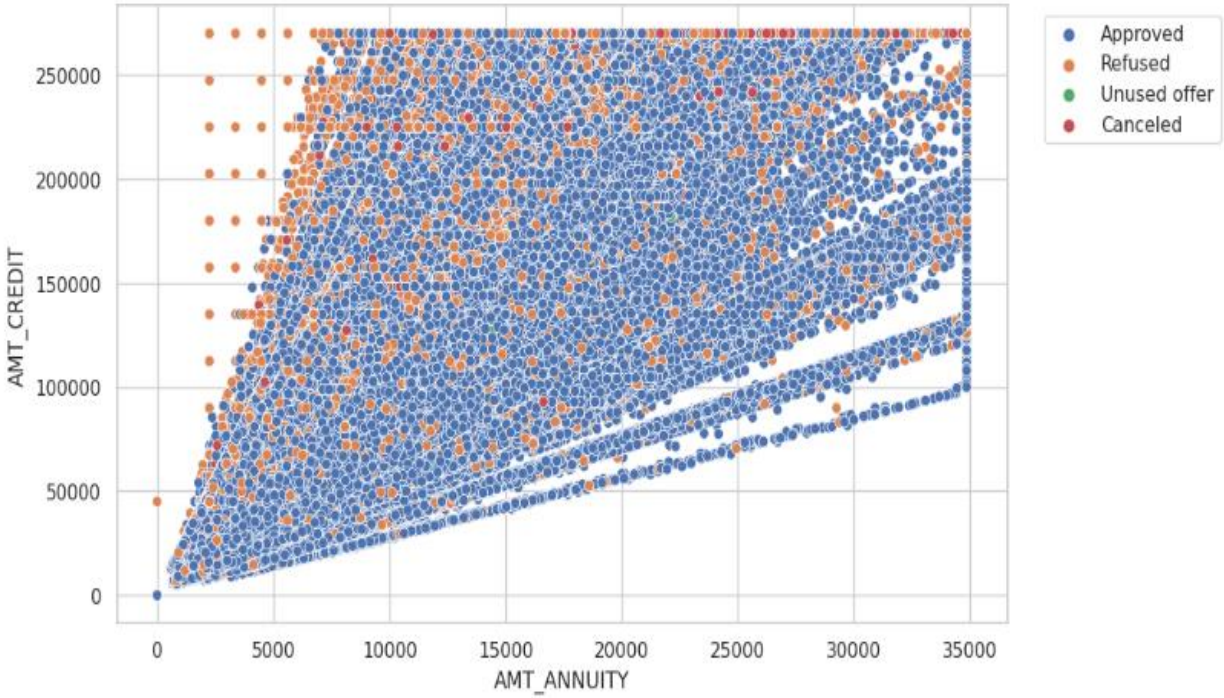


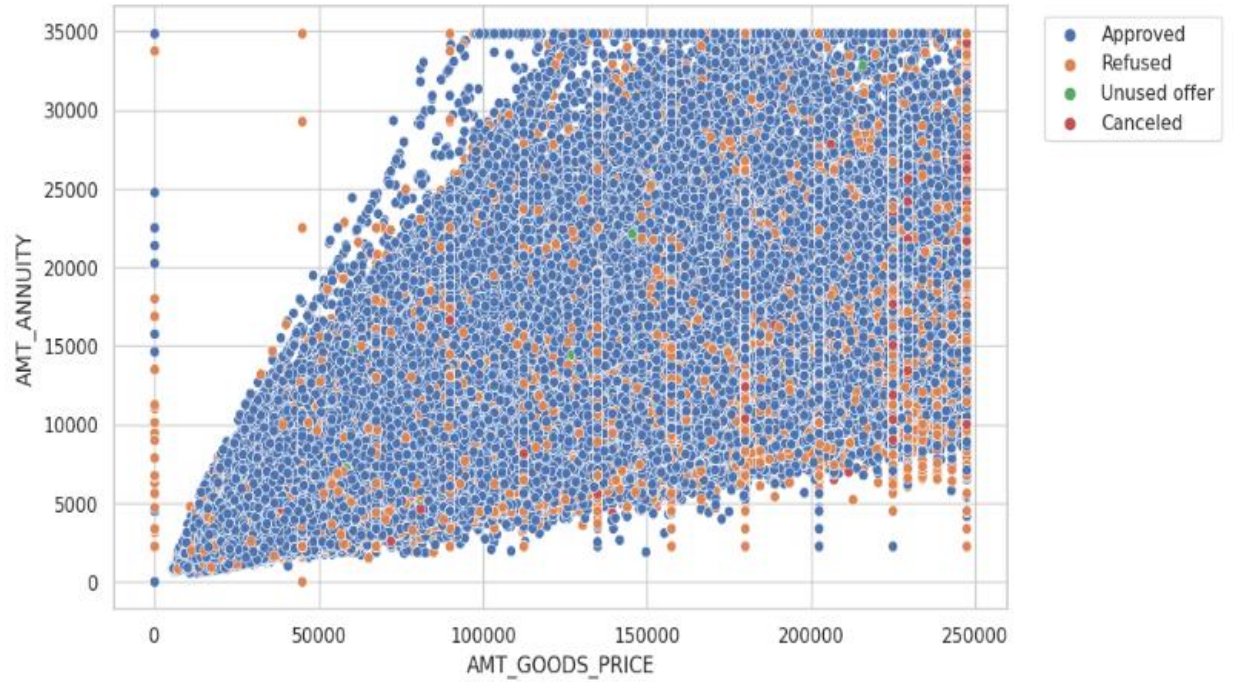
Density Plot for AMT_CREDIT

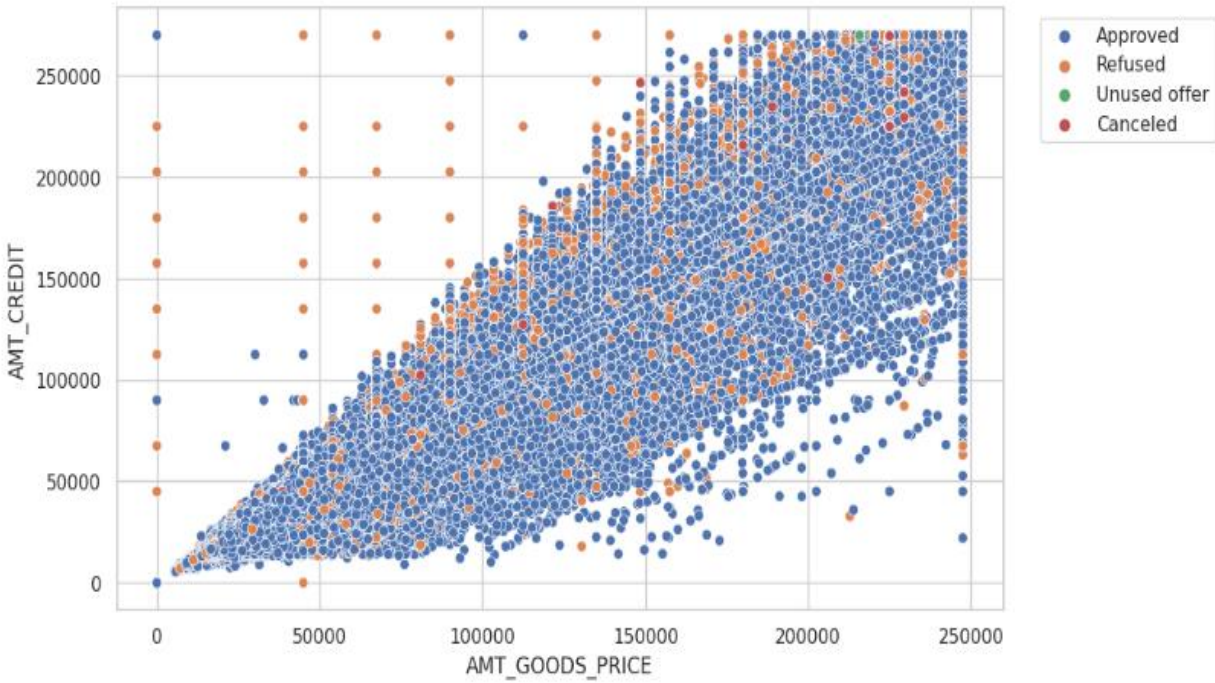


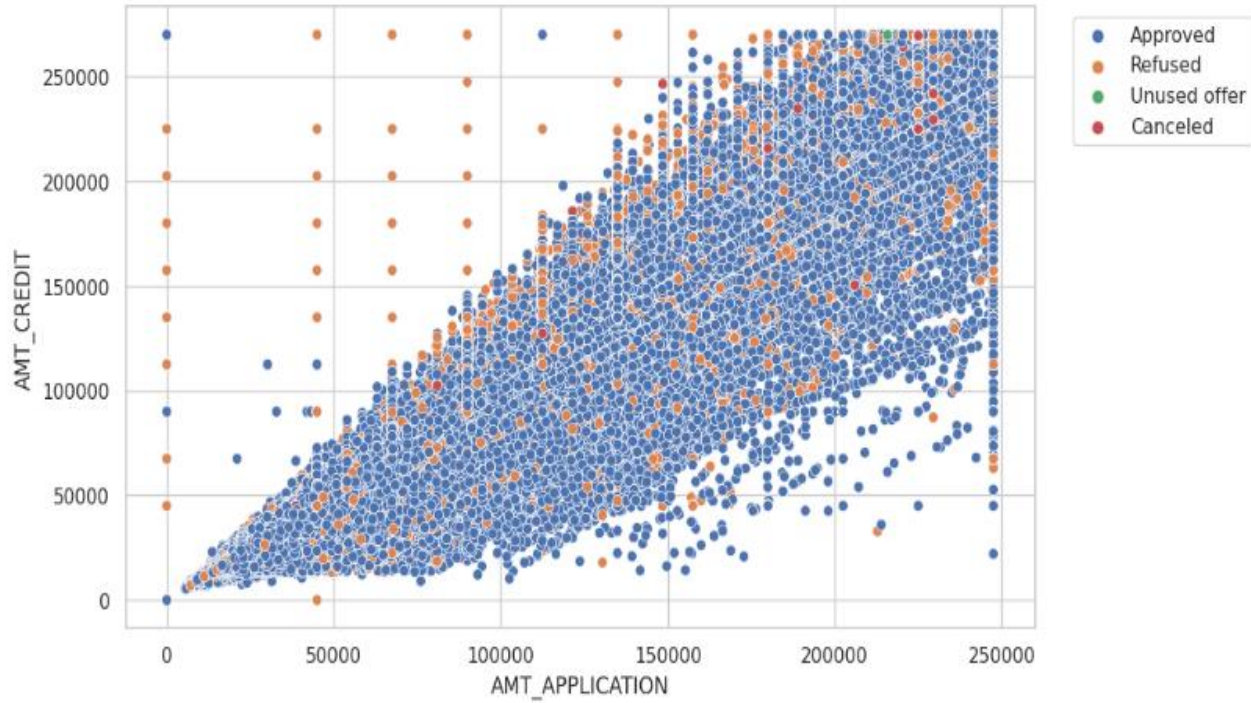
Density Plot for CNT_PAYMENT

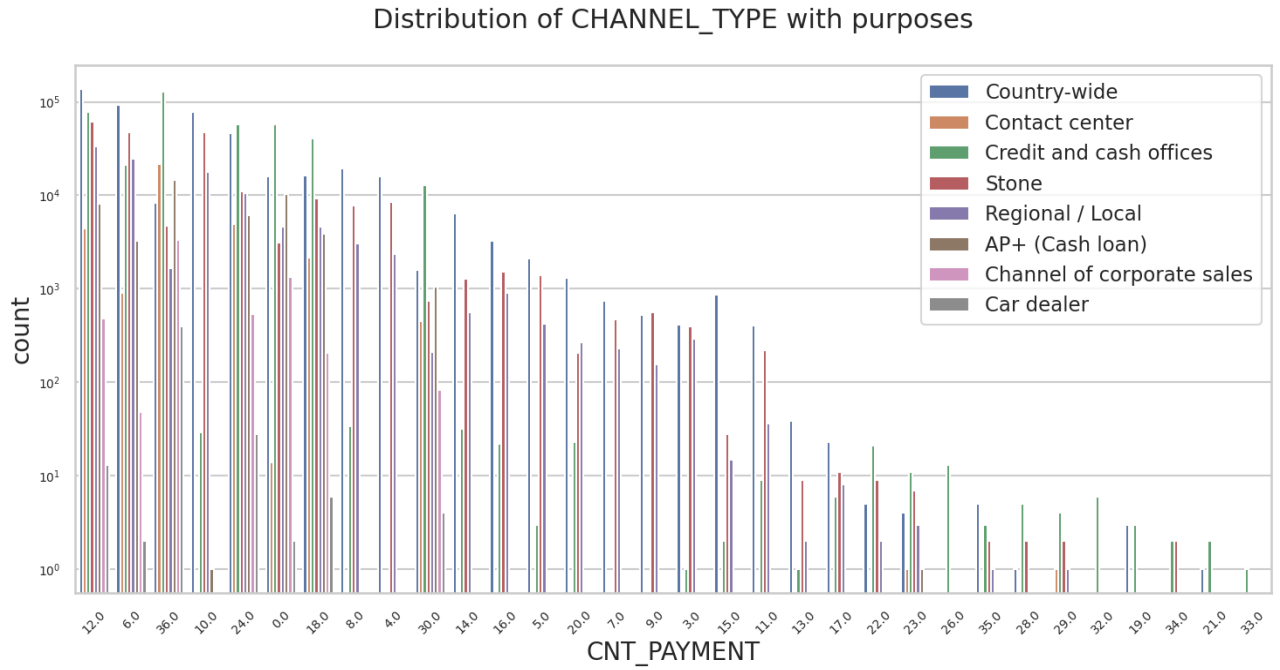


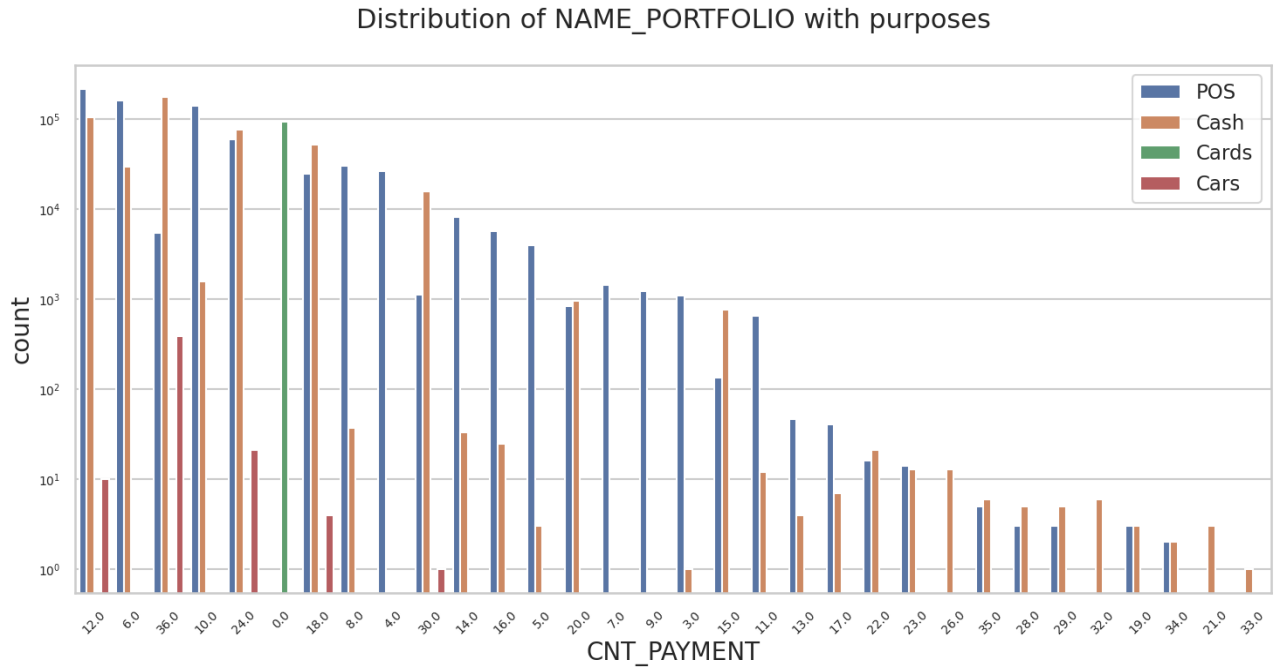

AMT_CREDIT
AMT_ANNUITY
0
50000
100000
150000
200000
250000
0
5000
10000
15000
20000
25000
30000
35000
Approved
Refused
Unused offer
Canceled

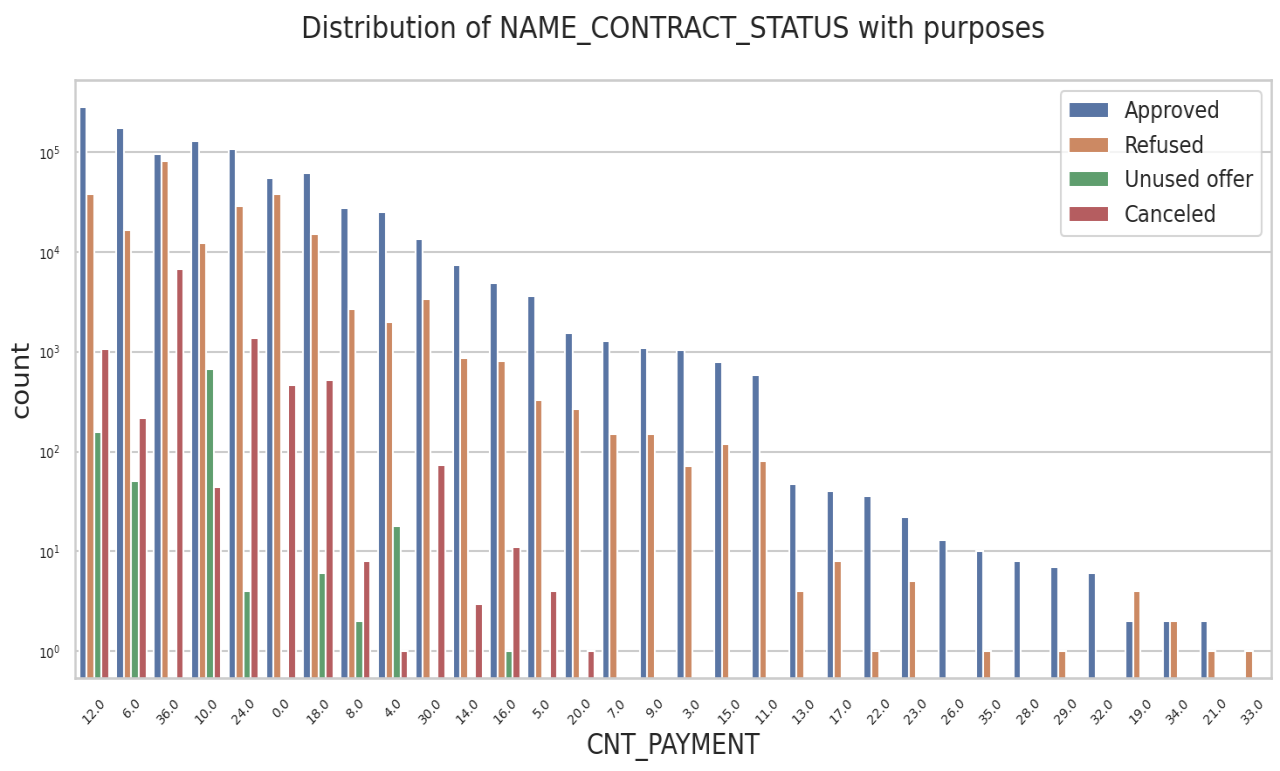

AMT_ANNUITY
AMT_GOODS_PRICE
0
5000
10000
15000
20000
25000
30000
35000
0
50000
100000
150000
200000
250000
Approved
Refused
Unused offer
Canceled

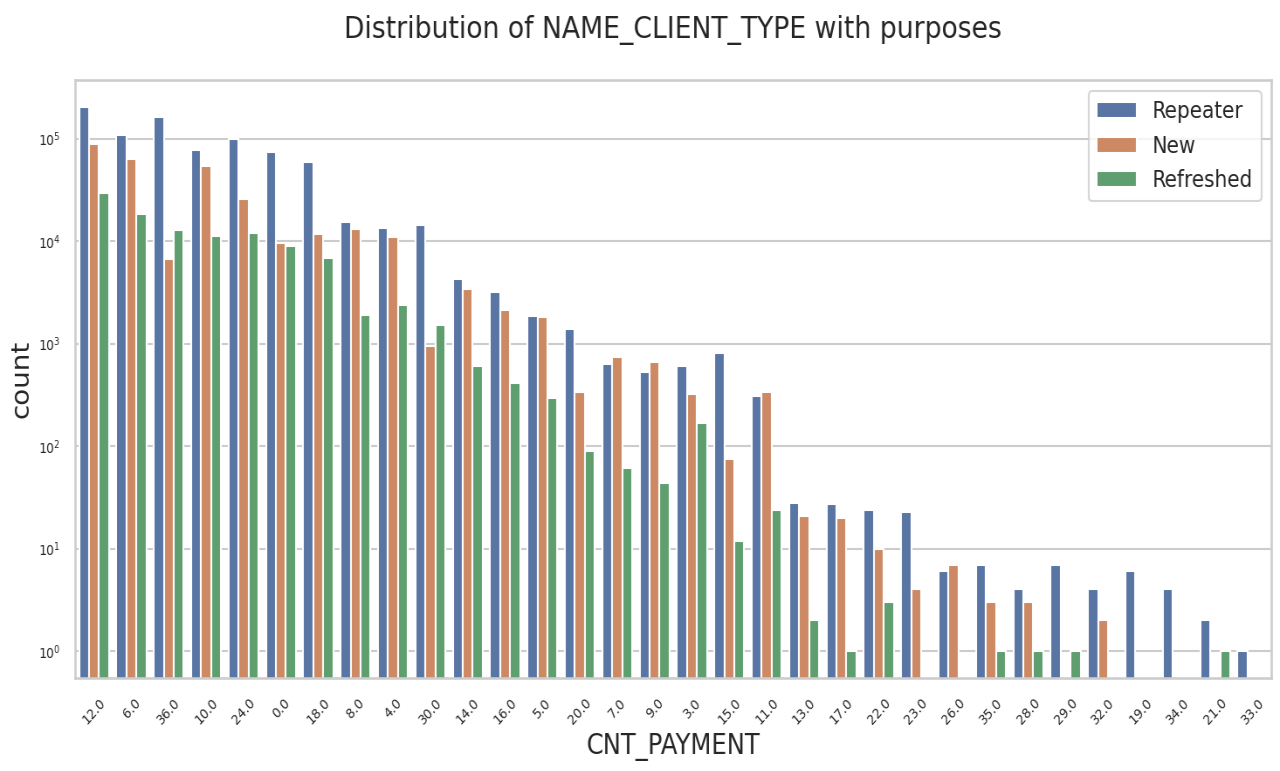

AMT_CREDIT
AMT_GOODS_PRICE
0
50000
100000
150000
200000
250000
Approved
Refused
Unused offer
Canceled


AMT_CREDIT
AMT_APPLICATION
0
50000
100000
150000
200000
250000
Approved
Refused
Unused offer
Canceled

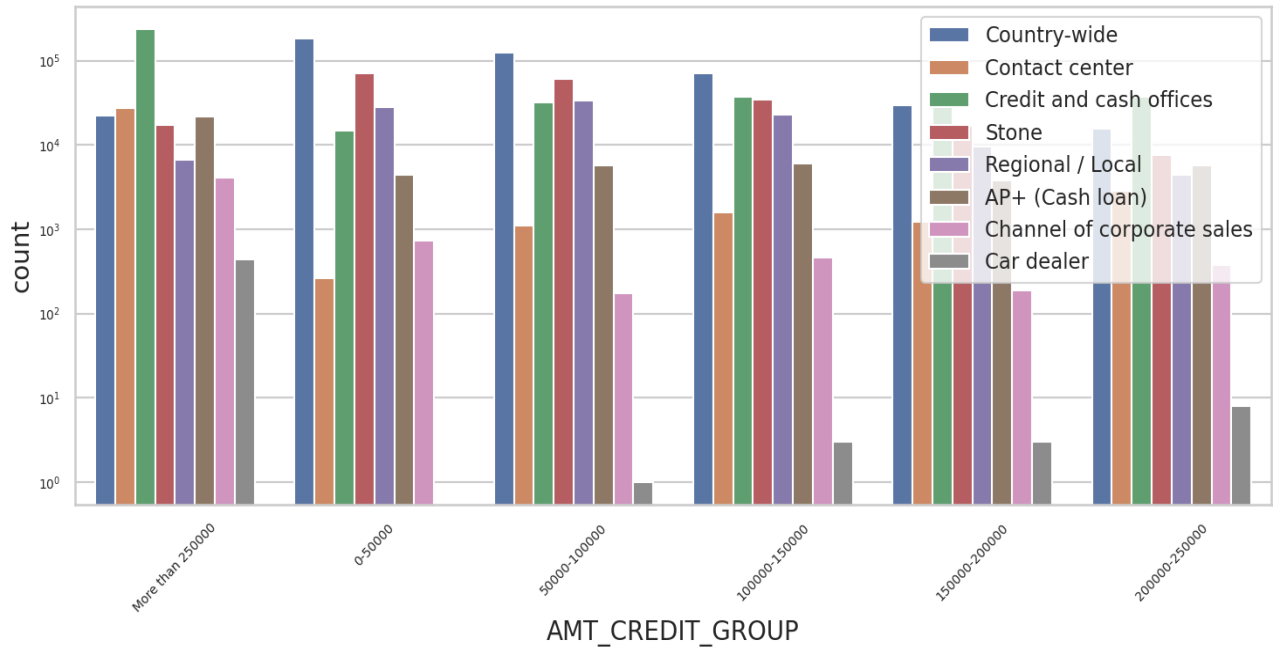

Distribution of CHANNEL_TYPE with purposes
Country-wide
Contact center
Credit and cash offices
Stone
Regional / Local
AP+ (Cash loan)
Channel of corporate sales
Car dealer
count
CNT_PAYMENT


Distribution of NAME_PORTFOLIO with purposes
POS
Cash
Cards
Cars
count
CNT_PAYMENT

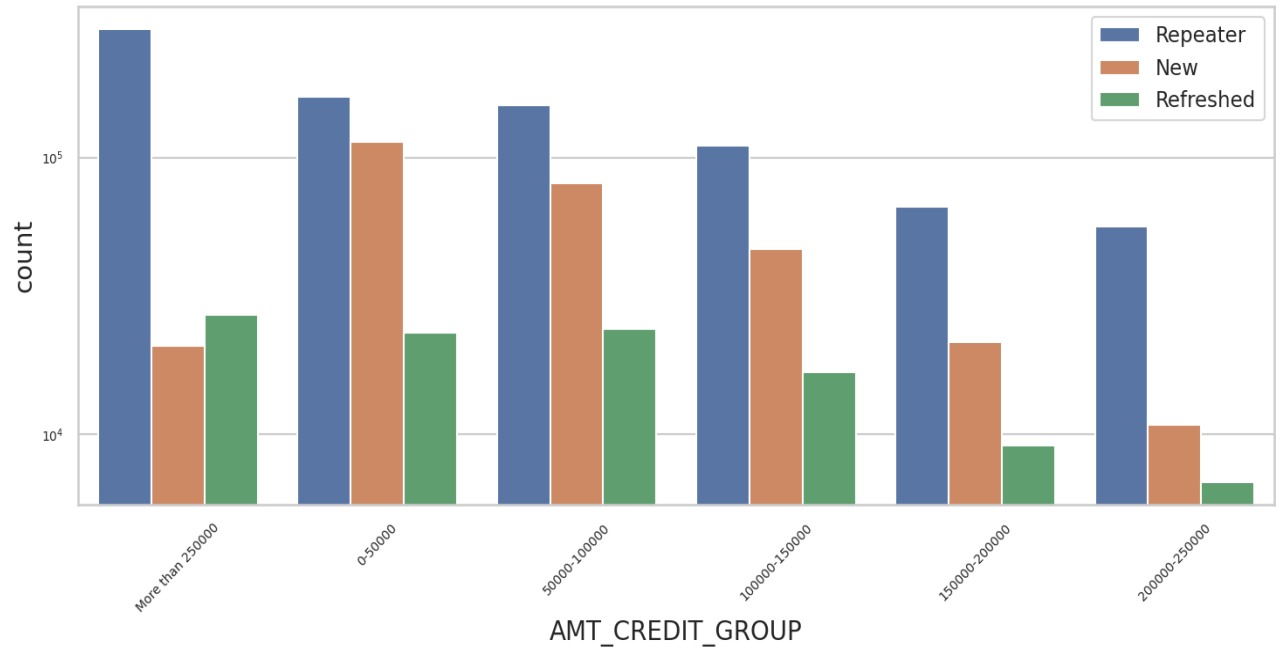

Distribution of NAME_CONTRACT_STATUS with purposes
Approved
Refused
Unused offer
Canceled
count
CNT_PAYMENT


Distribution of NAME_CLIENT_TYPE with purposes
Repeater
New
Refreshed
count
CNT_PAYMENT

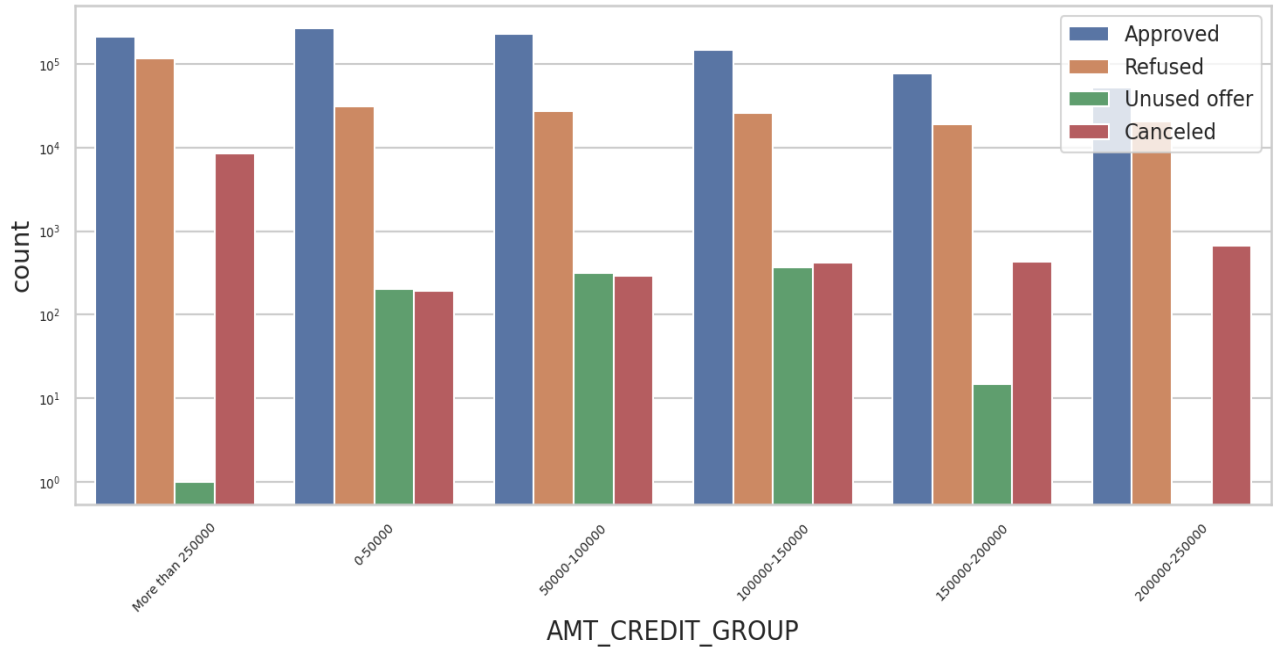
Distribution of CHANNEL_TYPE with purposes



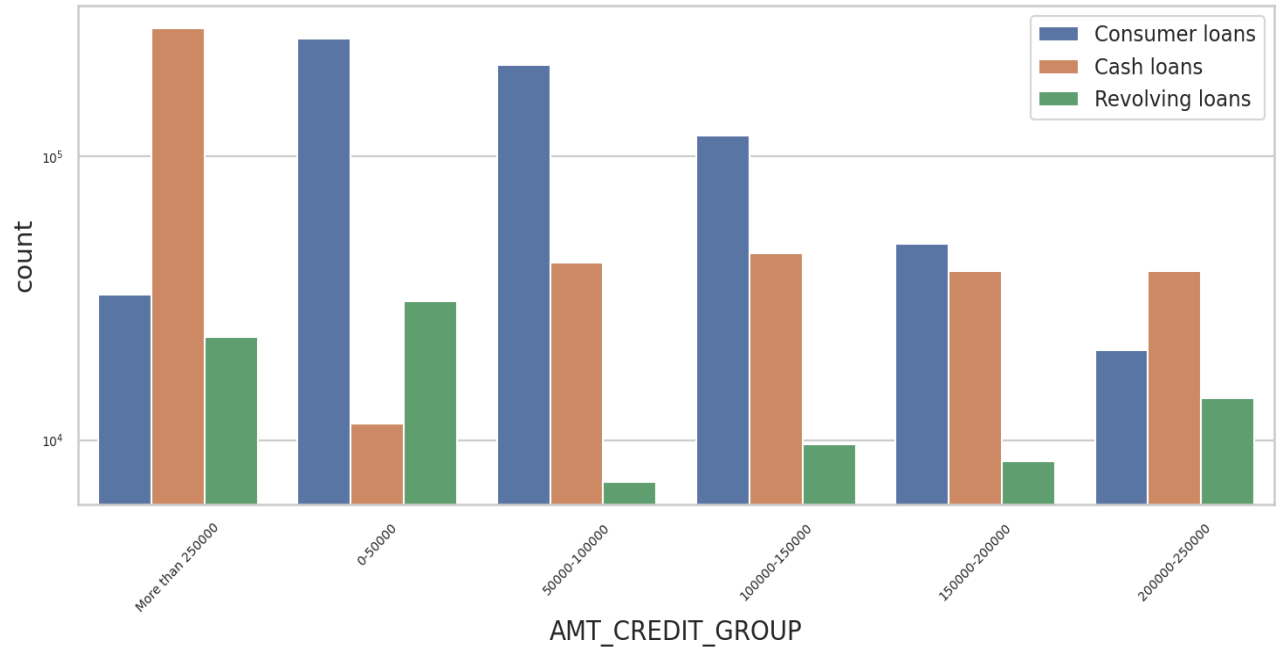
Distribution of NAME_CLIENT_TYPE with purposes



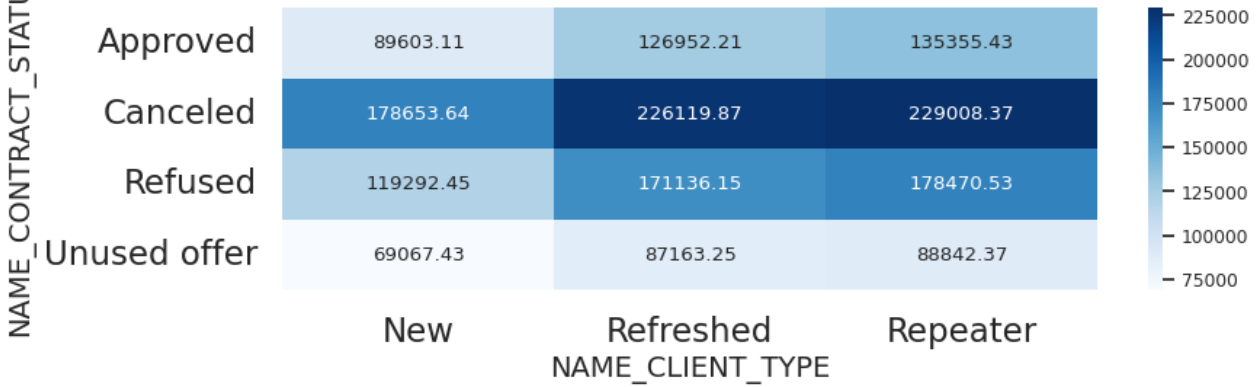
Distribution of NAME_CONTRACT_STATUS with purposes



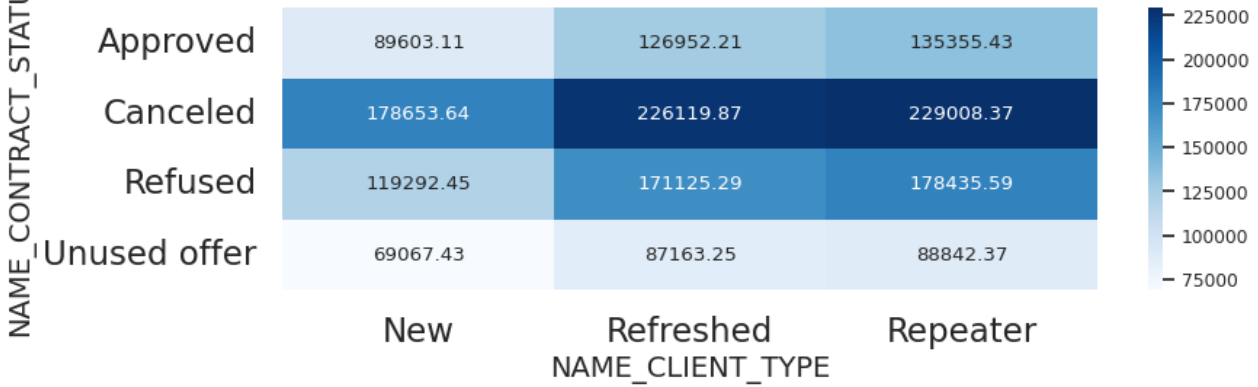
Distribution of NAME_CONTRACT_TYPE with purposes



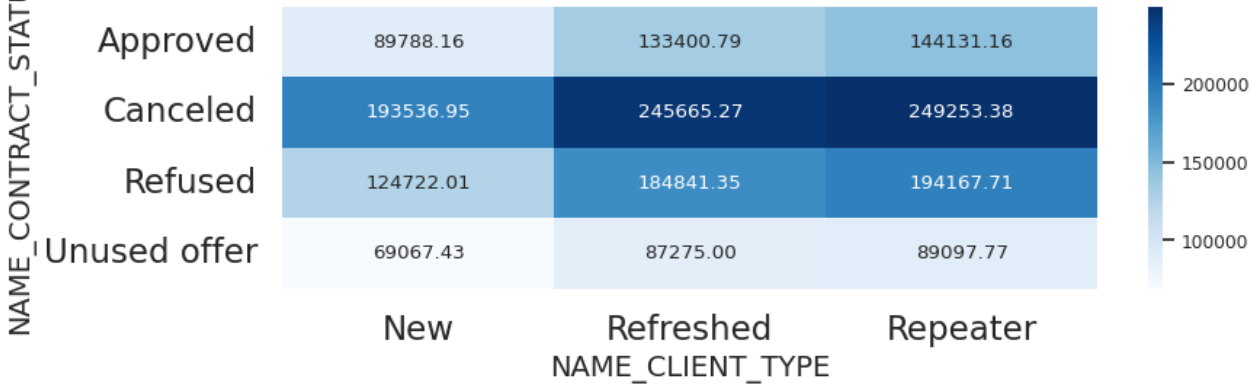
Heatmap of AMT_GOODS_PRICE by Contract Status and Client Type



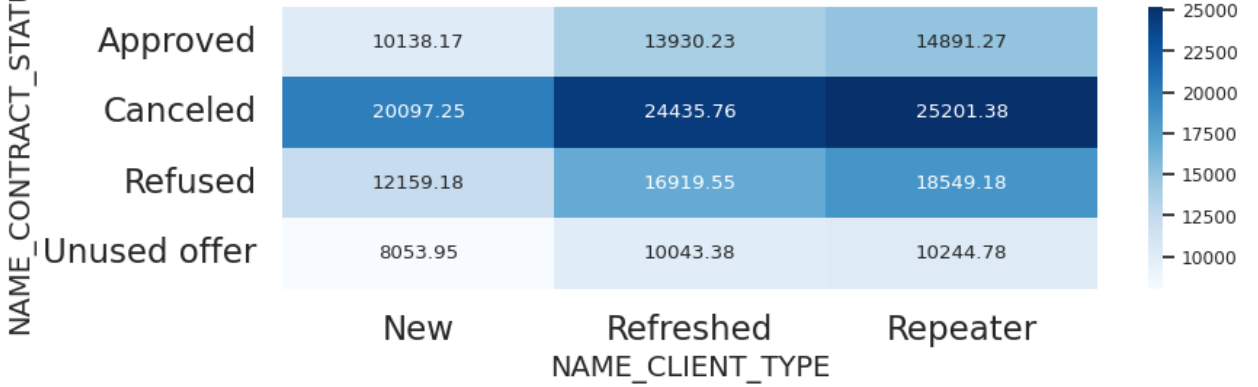
Heatmap of AMT_APPLICATION by Contract Status and Client Type



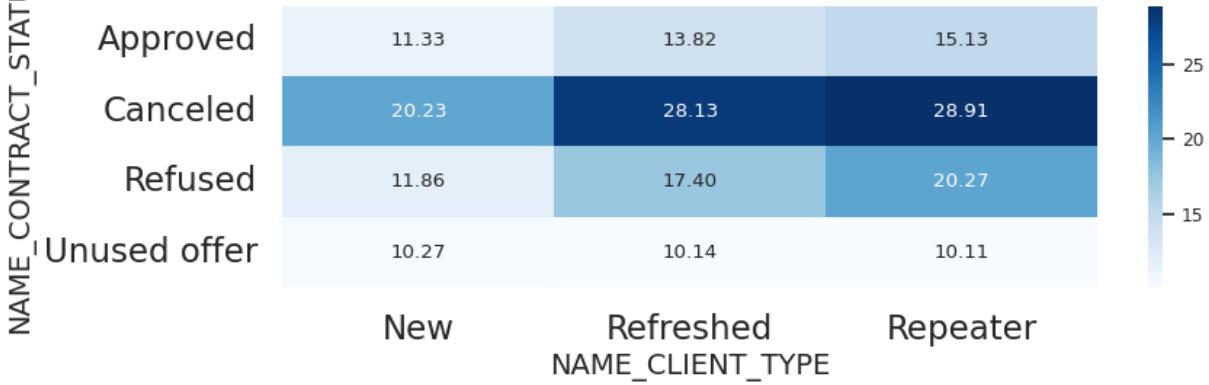
Heatmap of AMT_CREDIT by Contract Status and Client Type

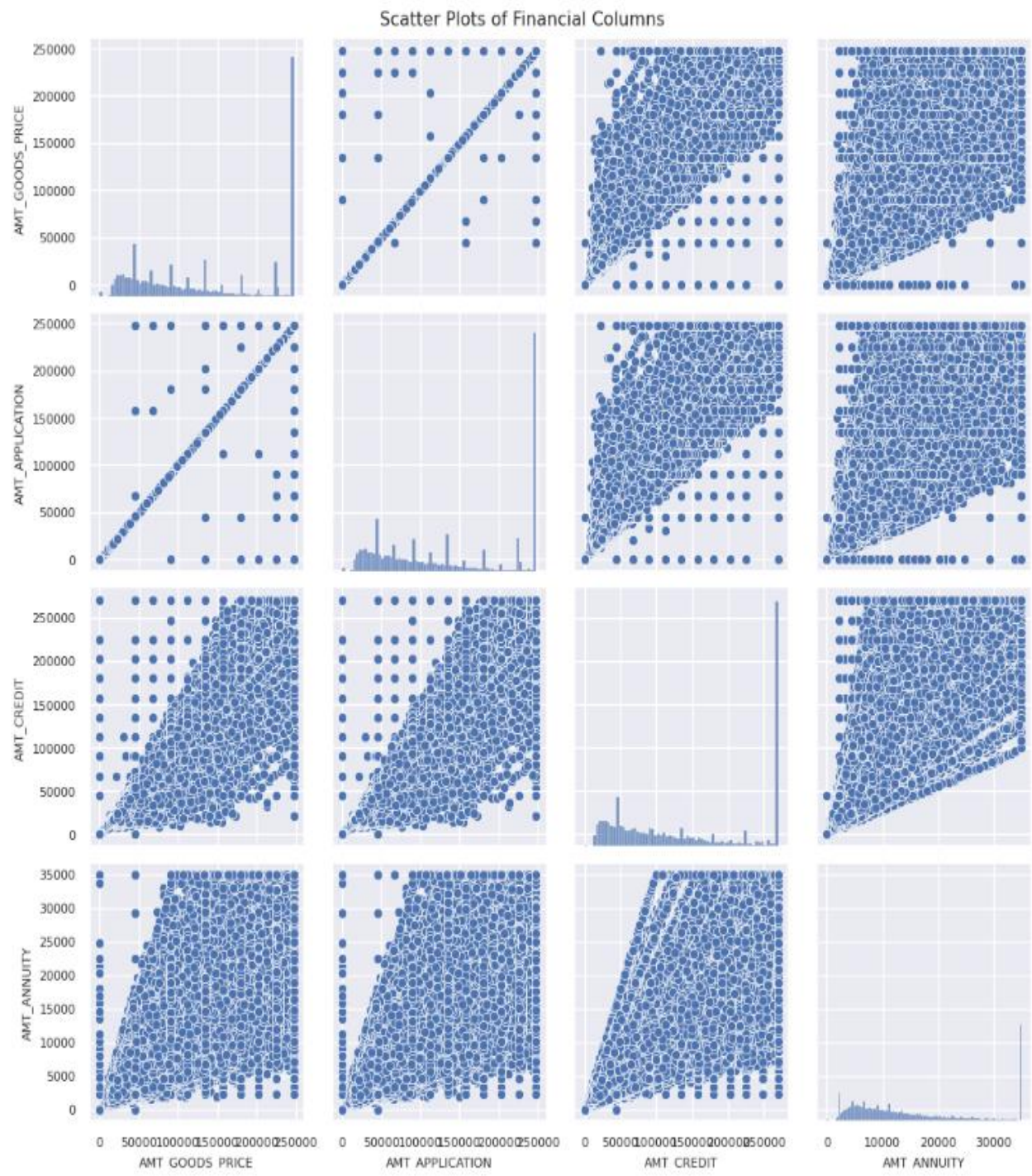


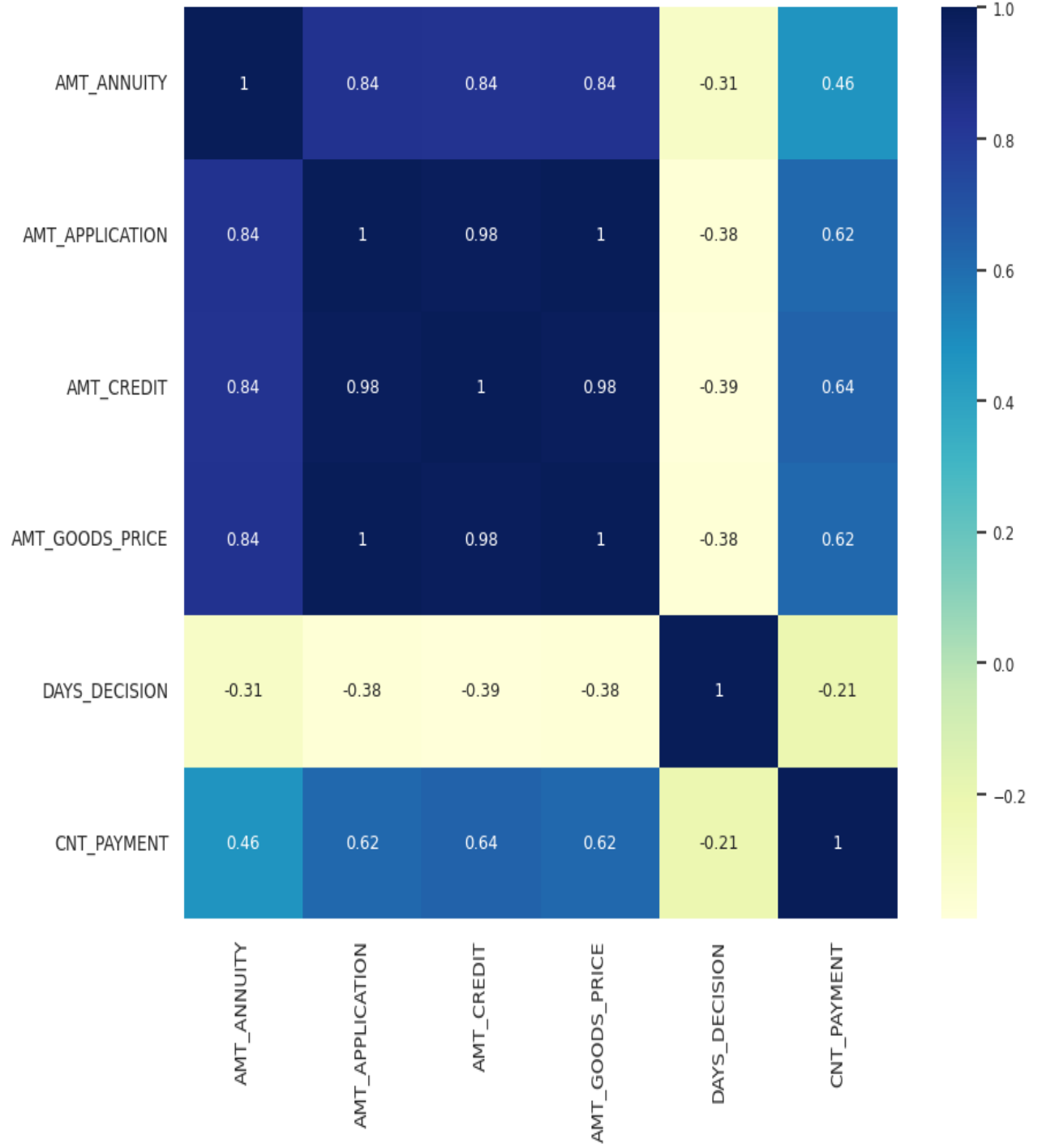
Heatmap of AMT_ANNUITY by Contract Status and Client Type



Heatmap of CNT_PAYMENT by Contract Status and Client Type




Scatter Plots of Financial Columns
AMT_GOODS_PRICE
AMT_APPLICATION
AMT_CREDIT
AMT_ANNUITY
AMT GOODS PRICE
AMT APPLICATION
AMT CREDIT
AMT ANNUITY


AMT_ANNUITY 1 0.84 0.84 0.84 -0.31 0.46
AMT_APPLICATION 0.84 1 0.98 1 -0.38 0.62
AMT_CREDIT 0.84 0.98 1 0.98 -0.39 0.64
AMT_GOODS_PRICE 0.84 1 0.98 1 -0.38 0.62
DAYS_DECISION -0.31 -0.38 -0.39 -0.38 1 -0.21
CNT_PAYMENT 0.46 0.62 0.64 0.62 -0.21 1
AMT_ANNUITY
AMT_APPLICATION
AMT_CREDIT
AMT_GOODS_PRICE
DAYS_DECISION
CNT_PAYMENT

Clarifications on the types of graphs used for the previous application data:

Histogram Analysis:

- Histograms were created for numerical columns in the dataset, revealing distinct density distributions.

Bar Plot Analysis (Bivariate):

- In bar graphs comparing "AMT_CREDIT" and "AMT_GOODS_PRICE_GROUP" across various categorical columns, a consistent trend was observed. Moving right on the graphs indicated an increase in represented values.

Scatter Plot Analysis (Bivariate):

- Scatter plots between numerical columns and the "NAME_CONTRACT_STATUS" field showed a linear relationship, suggesting a consistent correlation pattern. The majority of data points were associated with approved loans.

Count Graph Analysis (Bivariate):

- Count graphs illustrated relationships between numerical values and different categorical values.

Heat Map Analysis:

- Heat maps revealed notable points in various financial aspects. For instance, in aspects like 'AMT_GOODS_PRICE', 'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_ANNUITY', and 'CNT_PAYMENT,' there were exceptionally low unused offers for new client types. Conversely, cancelled application amounts were high for Repeater client types, indicating more favorable policies or interest rates for repeat applicants.

Pair Plot Analysis:

- The Pair Plot analysis highlighted strong correlations between 'AMT_GOODS_PRICE,' 'AMT_ANNUITY,' 'AMT_APPLICATION,' 'AMT_CREDIT,' and 'CNT_PAYMENT.' These correlations were expected because higher item values were associated with larger loan amounts and annual installments. Additionally, a notable correlation existed between 'AMT_CREDIT' and 'AMT_GOODS_PRICE' due to the inherent link between the credit granted and the price of financed goods.

Merged Both DataSets

Default Rate Among Approved Loans:

Among approved loans, there is an 8.2% default rate.

Concerning Trends in Previous Applications:

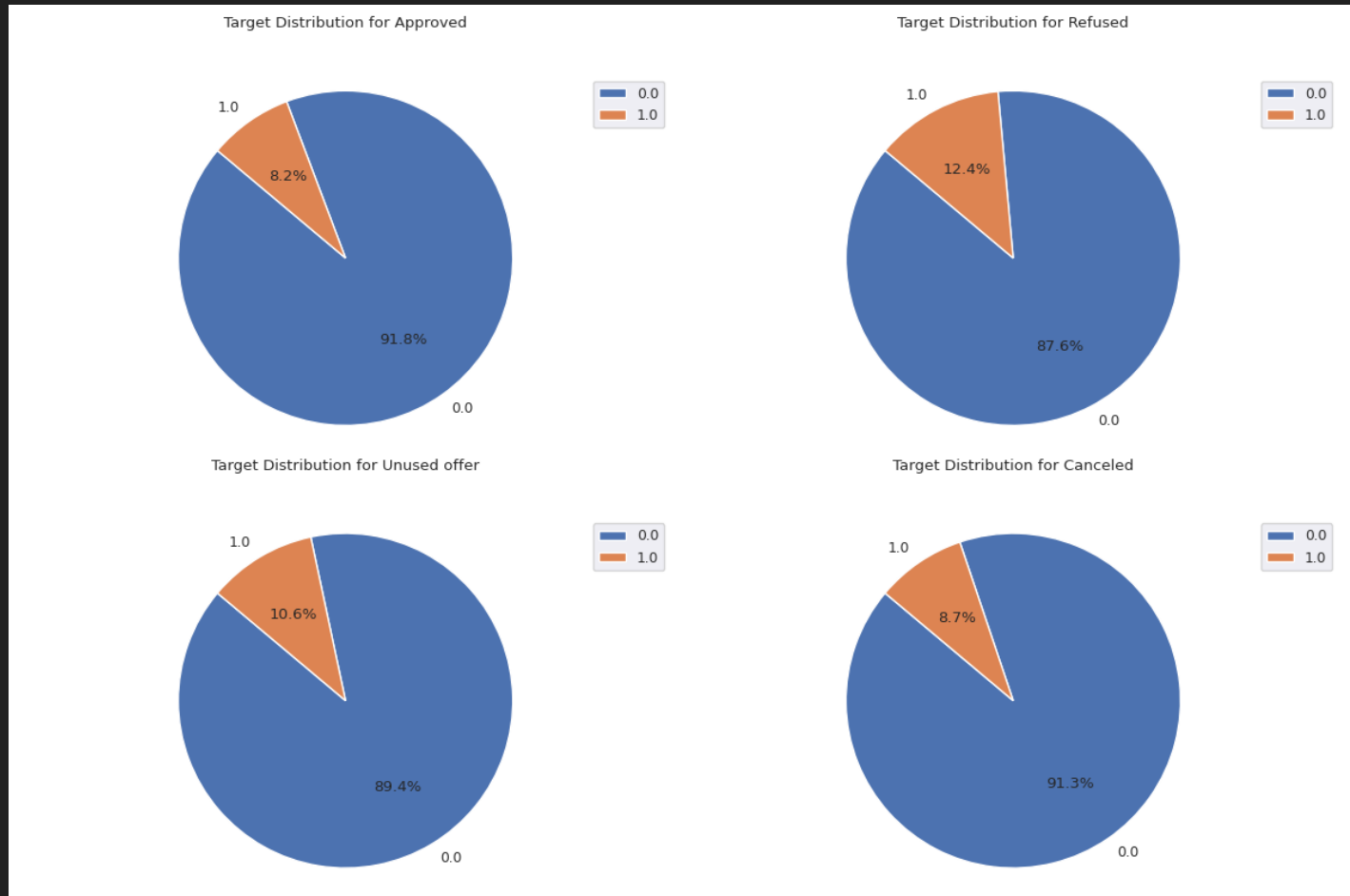
Alarming trends emerge when defaults are observed in previous applications that were either refused, cancelled, or unused.

Financial Company Actions:

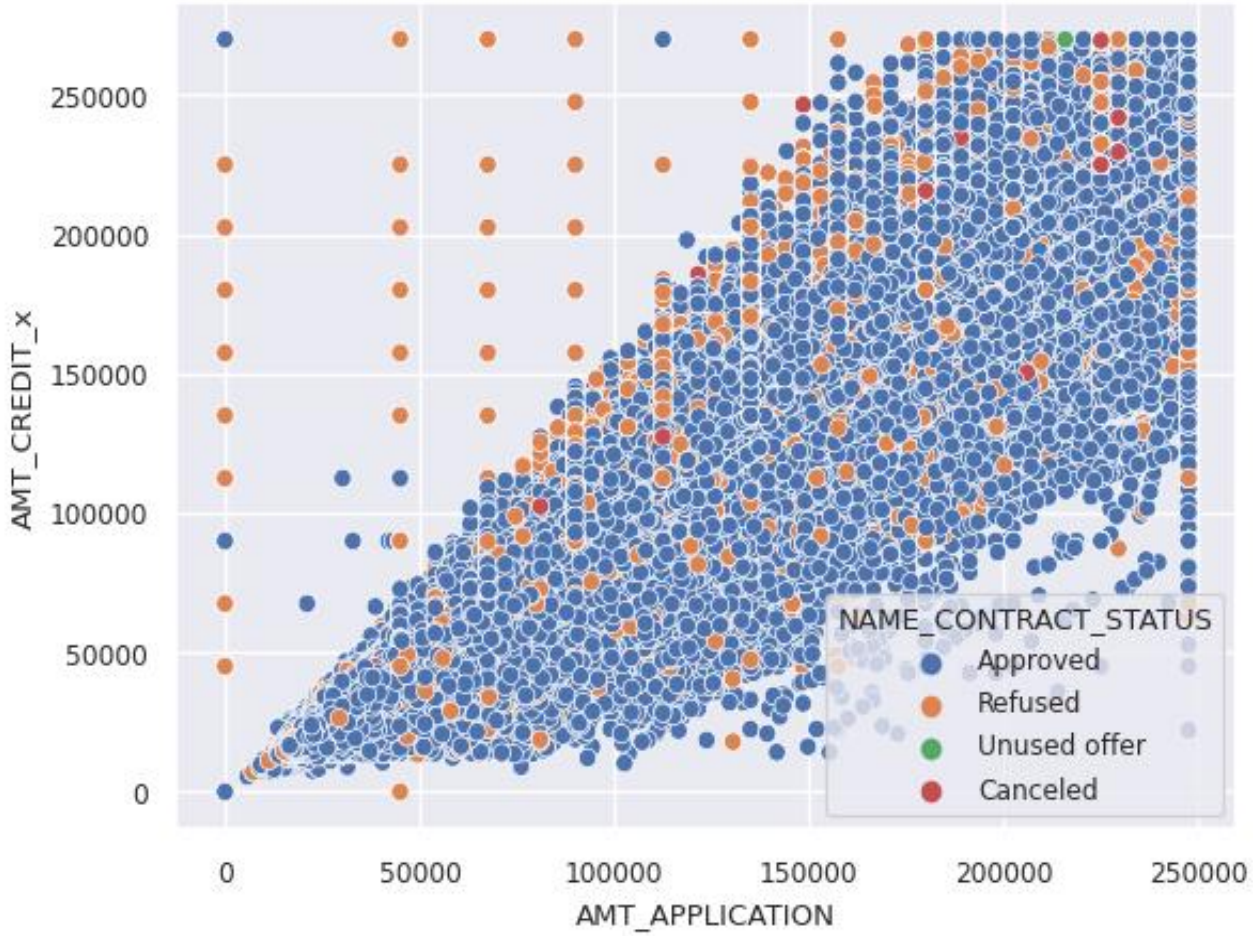
These observations suggest that despite the financial company's refusal or cancellation of previous applications, the current approved loans are still experiencing defaults.

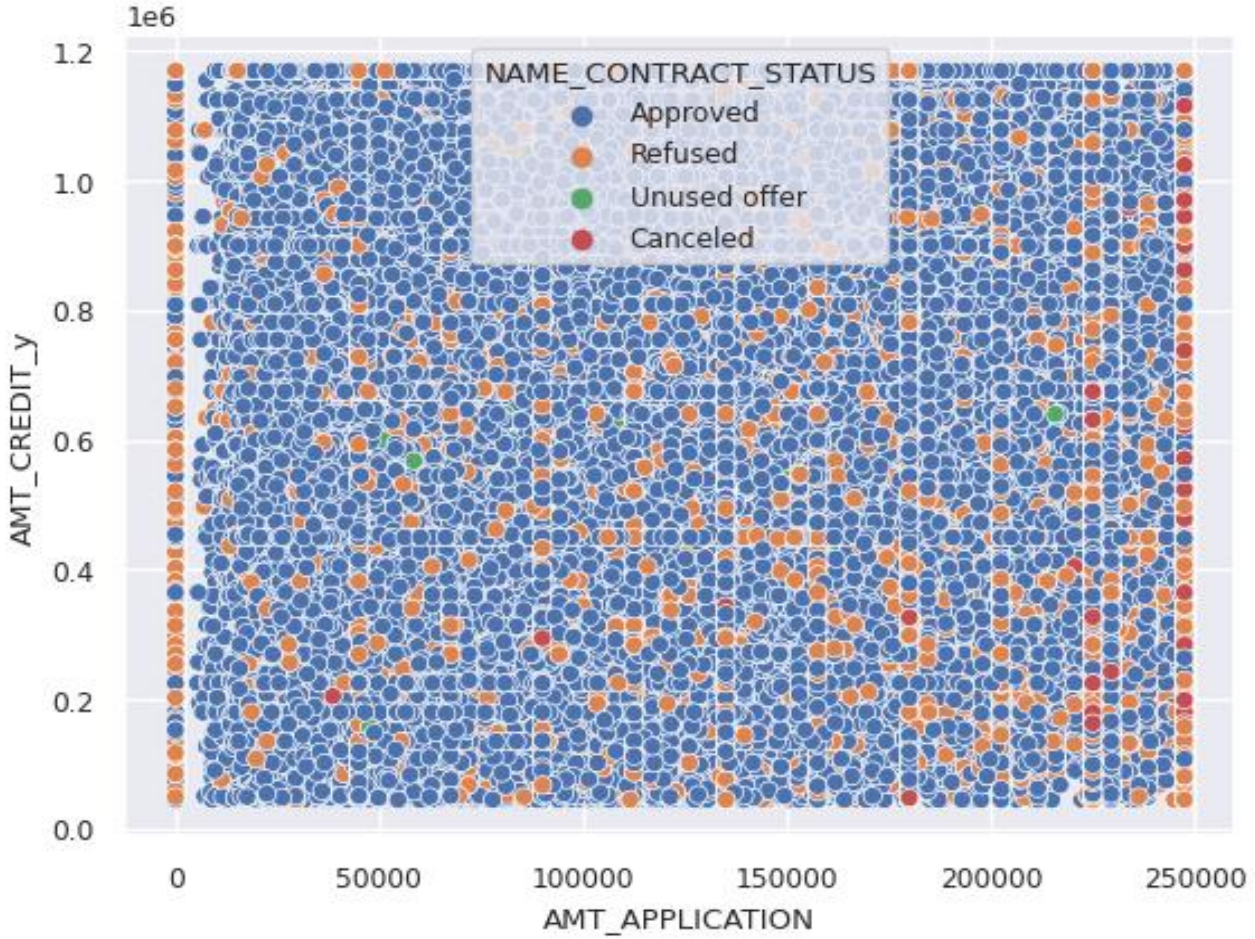
Persistent Default Risk:

The presence of defaults in such cases indicates a persistent risk of default, warranting further investigation and risk assessment.

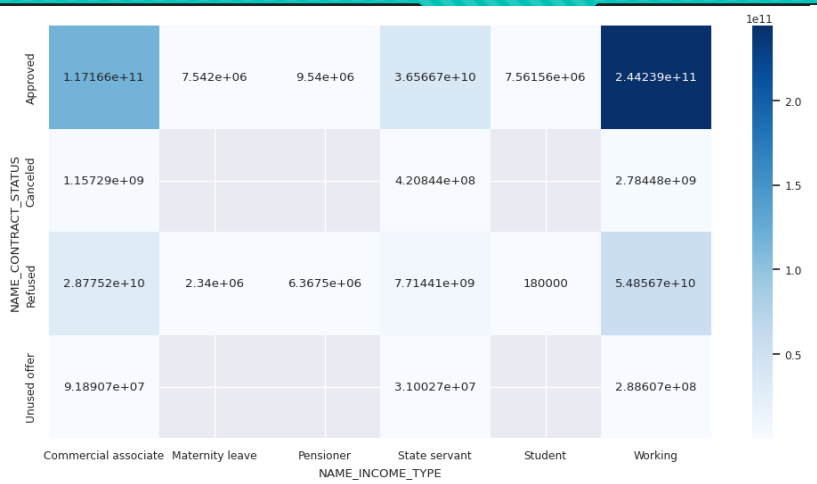


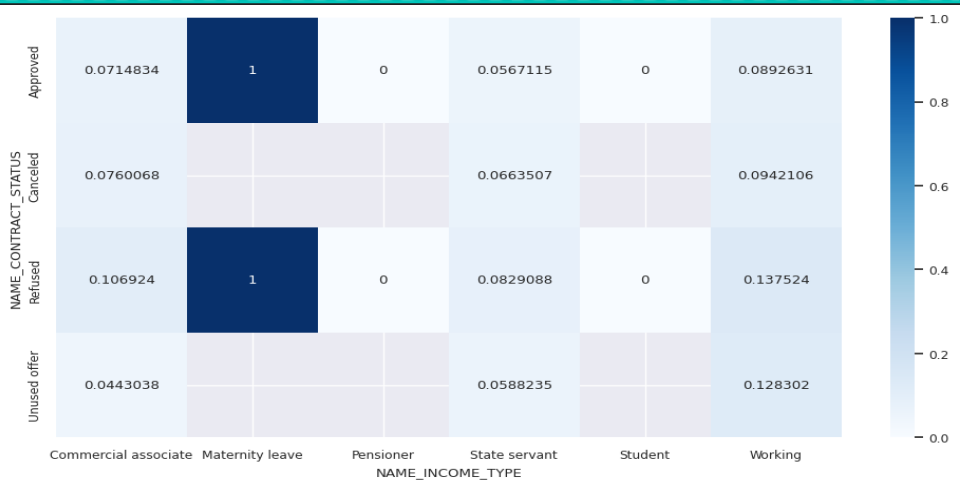
Exploring Relationships with Scatterplots

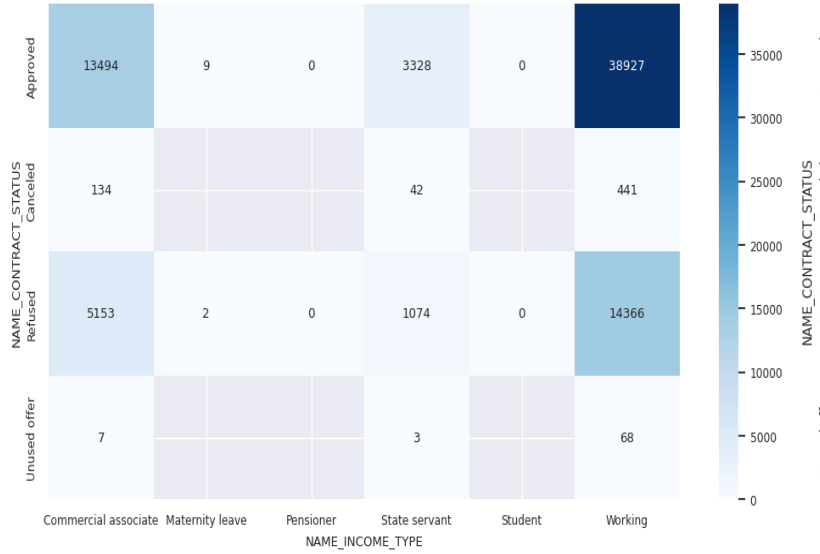




Key Insights from Analysis and Correlation Graphs







Here are several key points derived from the graph analysis above:

Heatmap Analysis for Default Correlation:

- A higher value in the upper matrix of the heatmap indicates a stronger correlation with default (Target 1).

Income Types and Default Trends:

- Working applicants with an approved status have the highest number of defaults among various income types.

Concerning Trends in Previous Applications:

- Notably, previous applications with refused, cancelled, or unused loans also show defaults, indicating persistent default risks despite prior rejections or cancellations.
- Within the working class applicants, 14,366 individuals who were previously refused have defaulted on their current loans.

Age Groups and Default Rates:

- The age group 30-40 exhibits the highest number of defaulters among approved loans, followed by age groups 20-30 and 40-50, which also display significant default rates.
- Previous applications with refused and cancelled loans are correlated with defaults in current applications.

Credit Offerings and Applicant Categories:

- A significant observation is the higher credit offerings to individuals in the working and commercial associate categories.
- Refused applicants, especially students, tend to receive smaller credit amounts.

Default Analysis by Applicant Categories:

- Among approved loans, the POS associate category has the highest count of defaulters.
- Approved loans in the car portfolio show only 14 instances of default.

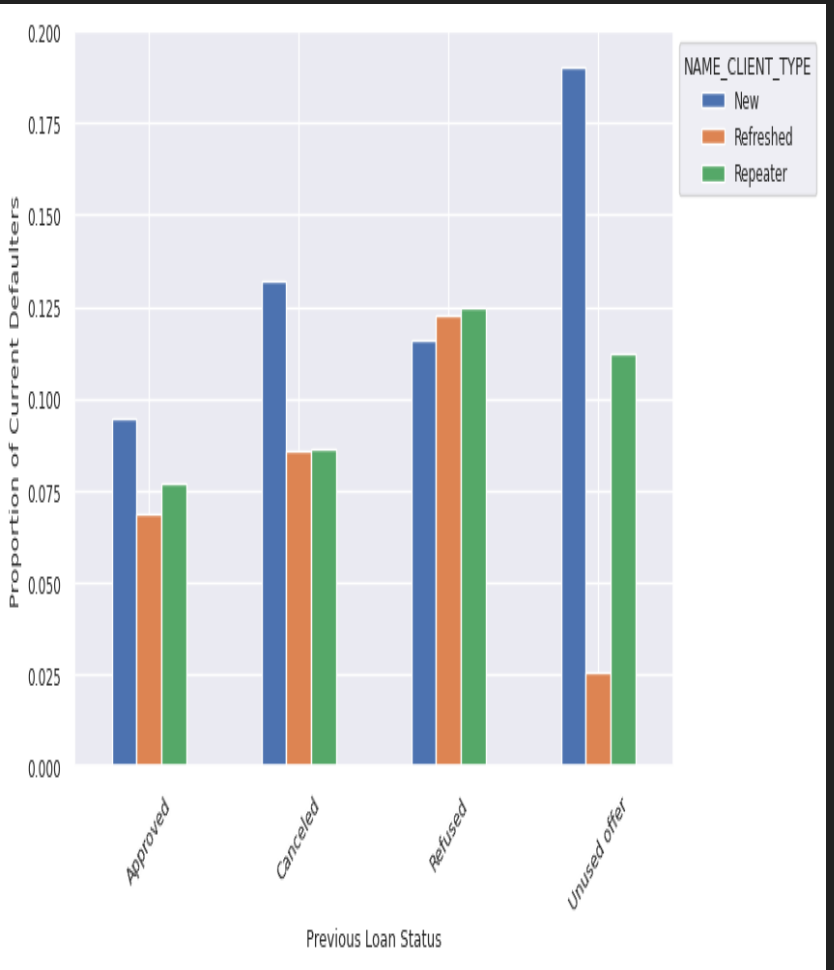
Default Cases by Country-wide and Car Dealer Categories:

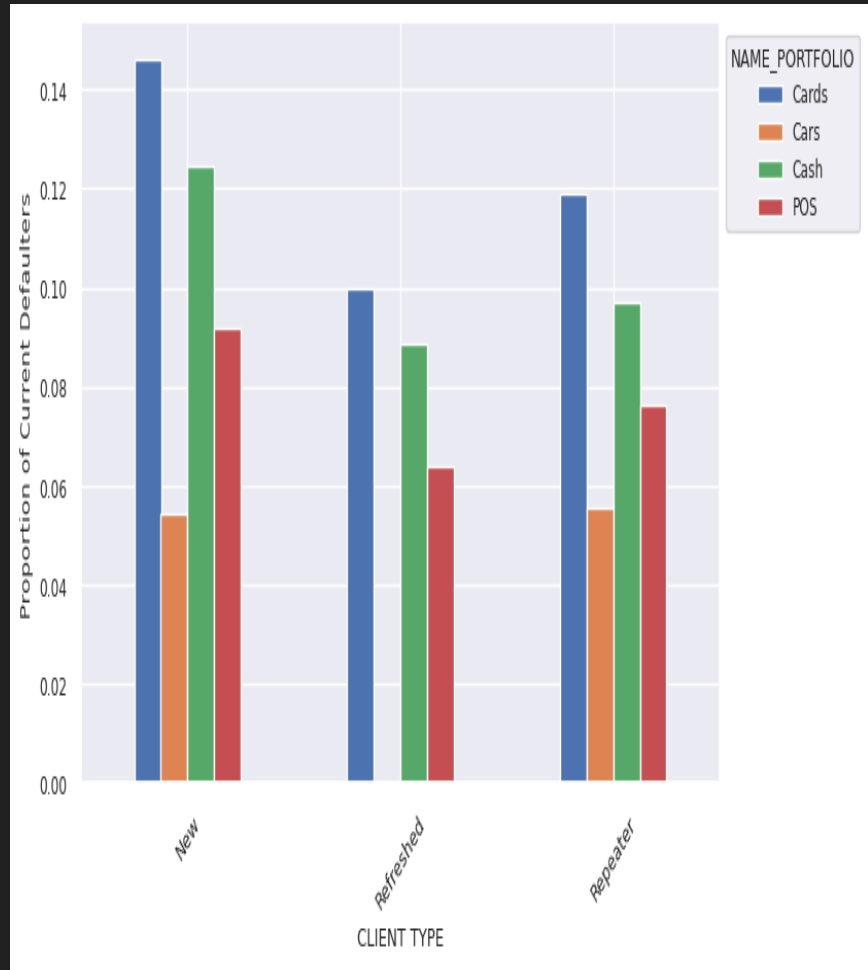
- Among approved loans, the country-wide associate category has the highest number of default cases.
- In the car dealer category of approved loans, there are only 14 instances of defaults.

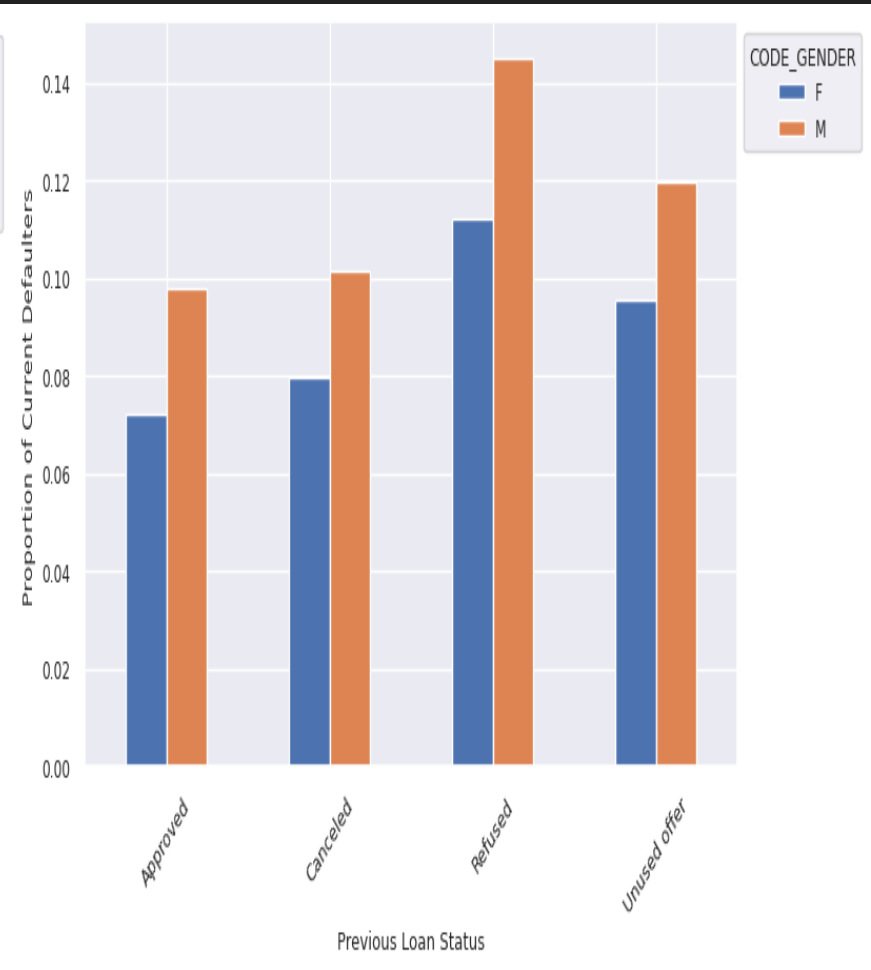
Maternity Leave and Other Categories:

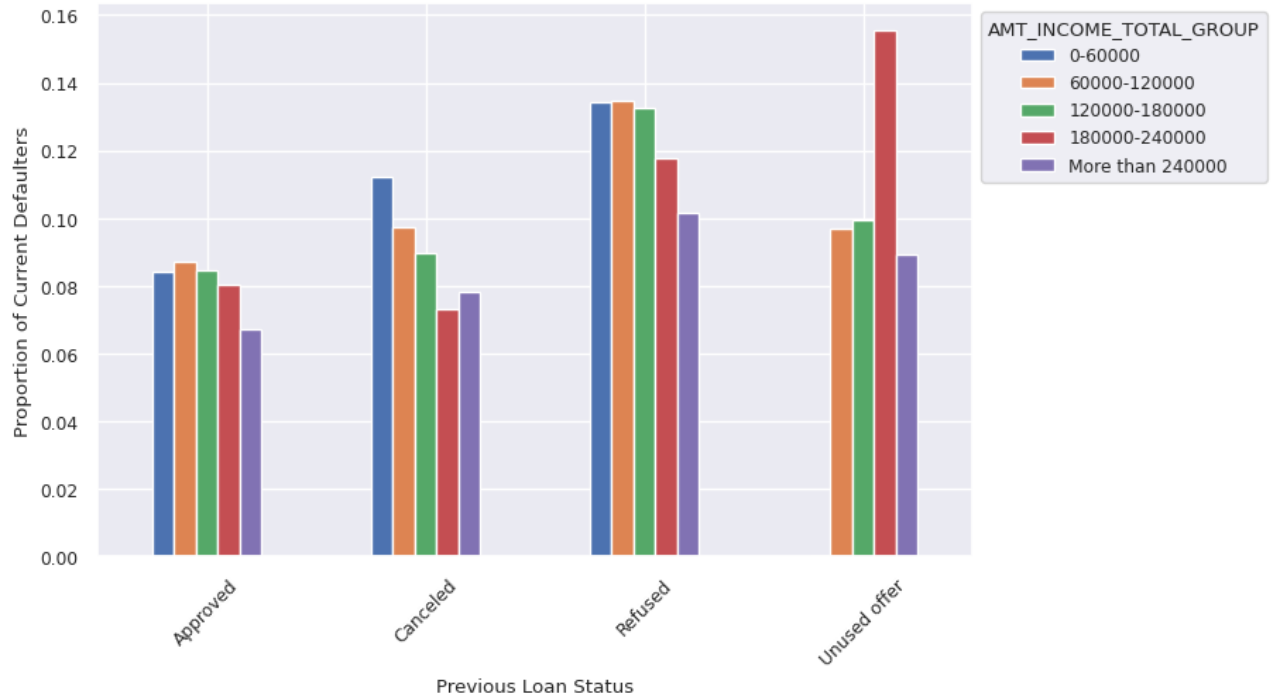
- When considering both approved and refused loans, the Maternity leave associate category stands out with the highest count of default cases.
- No instances of defaults are observed in the pensioner and student categories for both approved and refused loans.

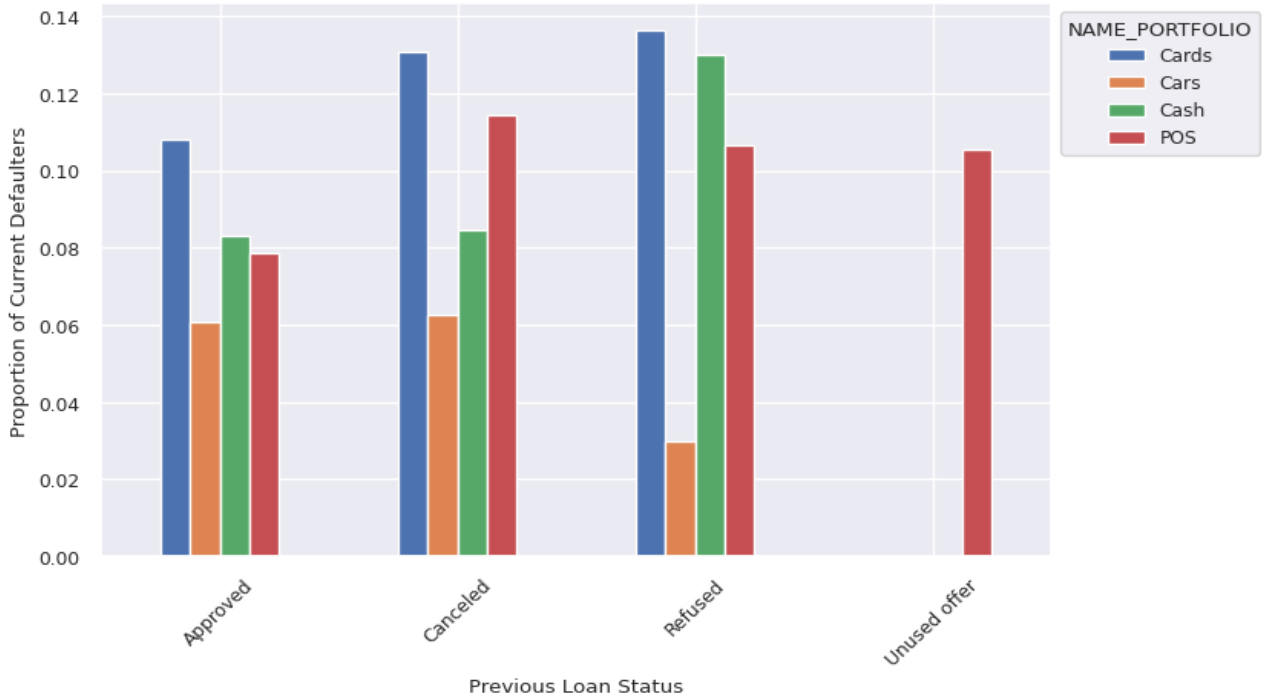
Exploring Column Relationships through Bivariate Bar Graphs

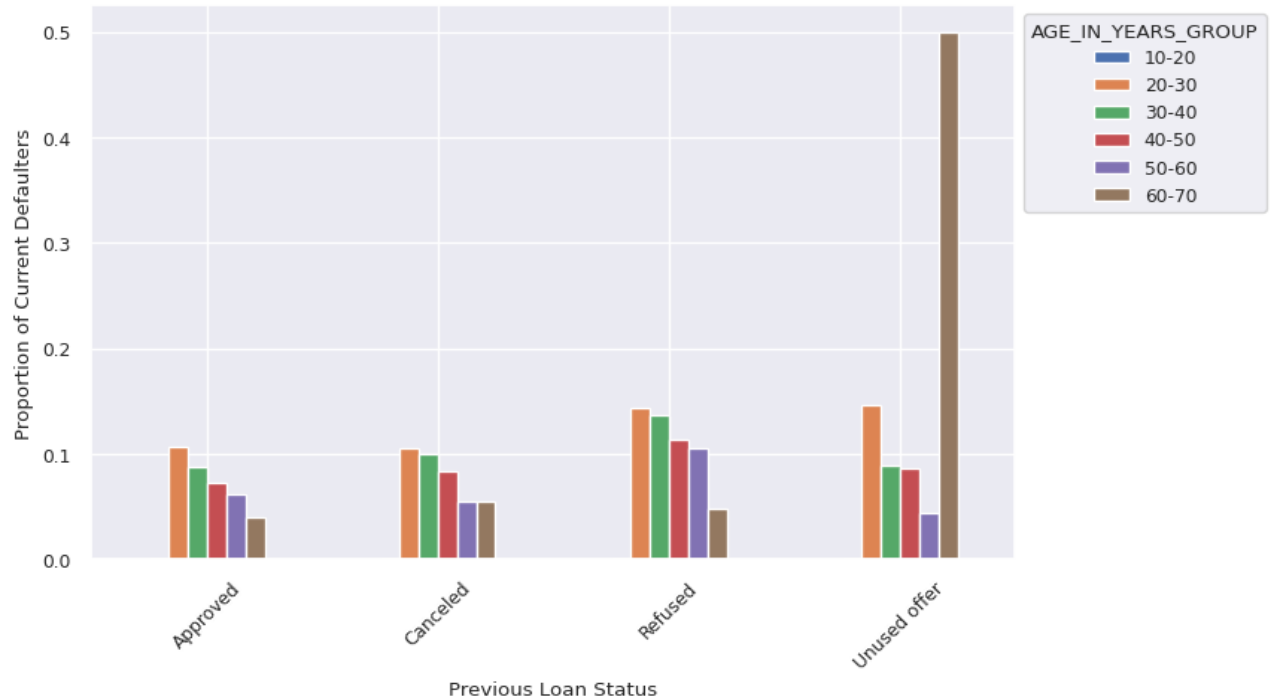


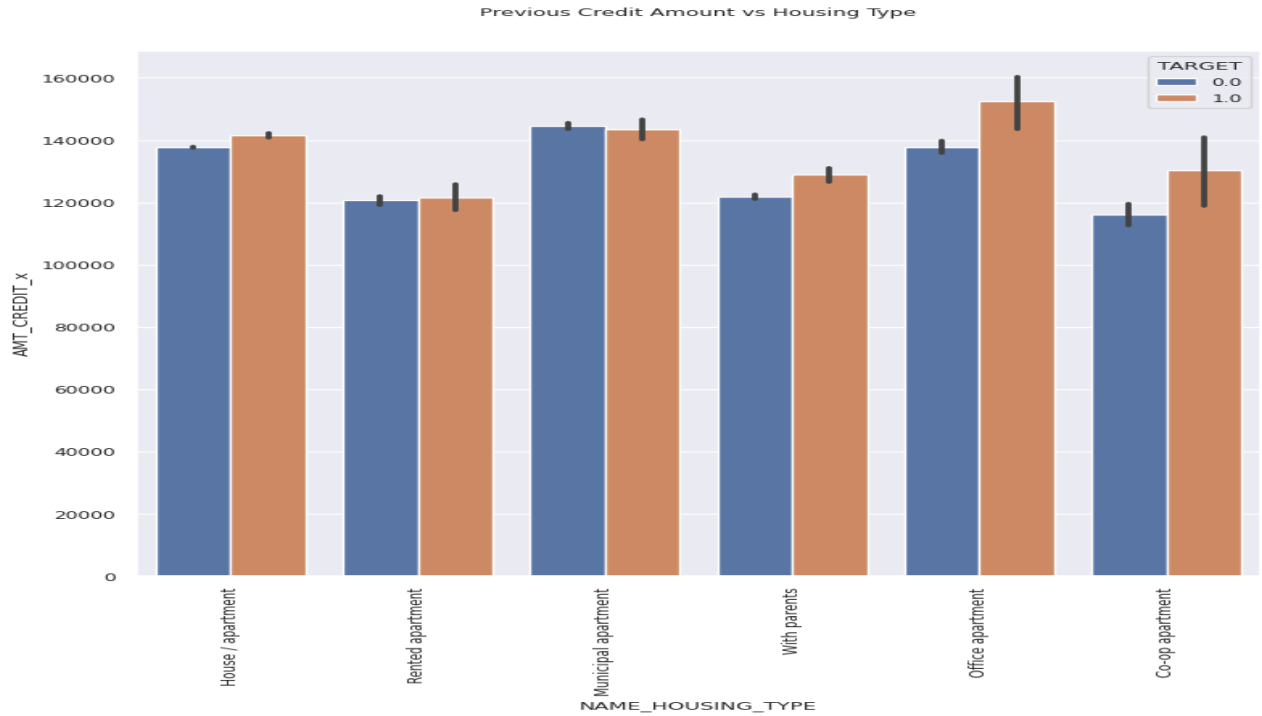


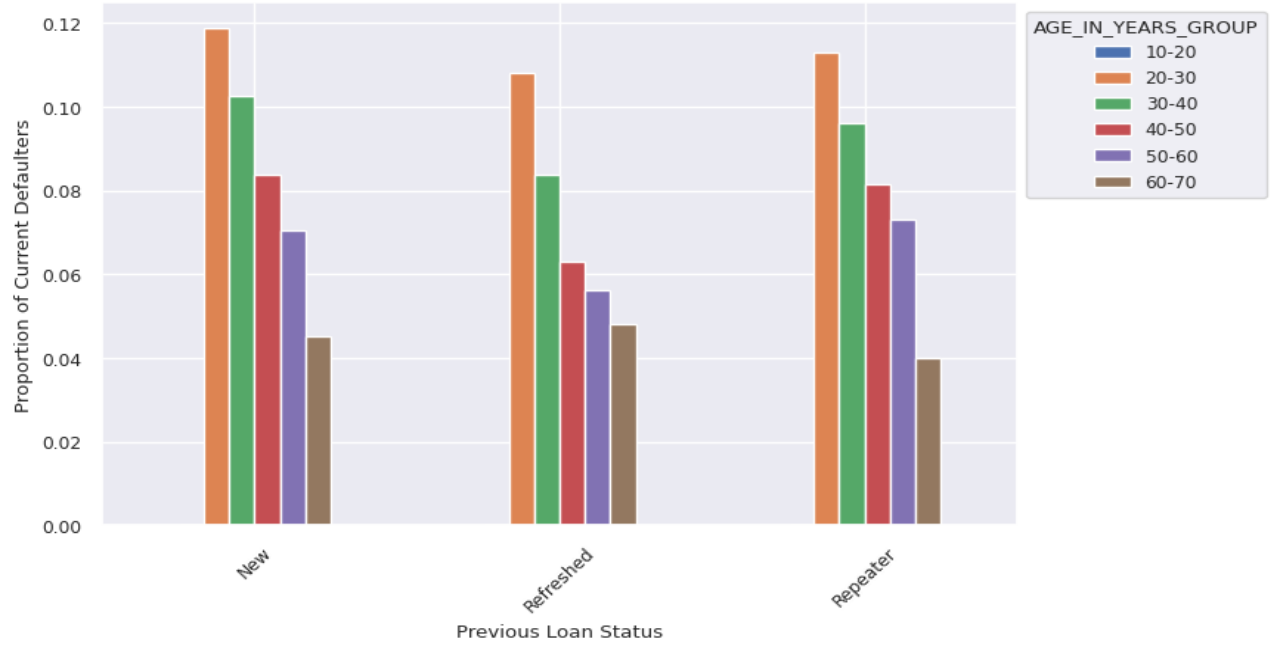


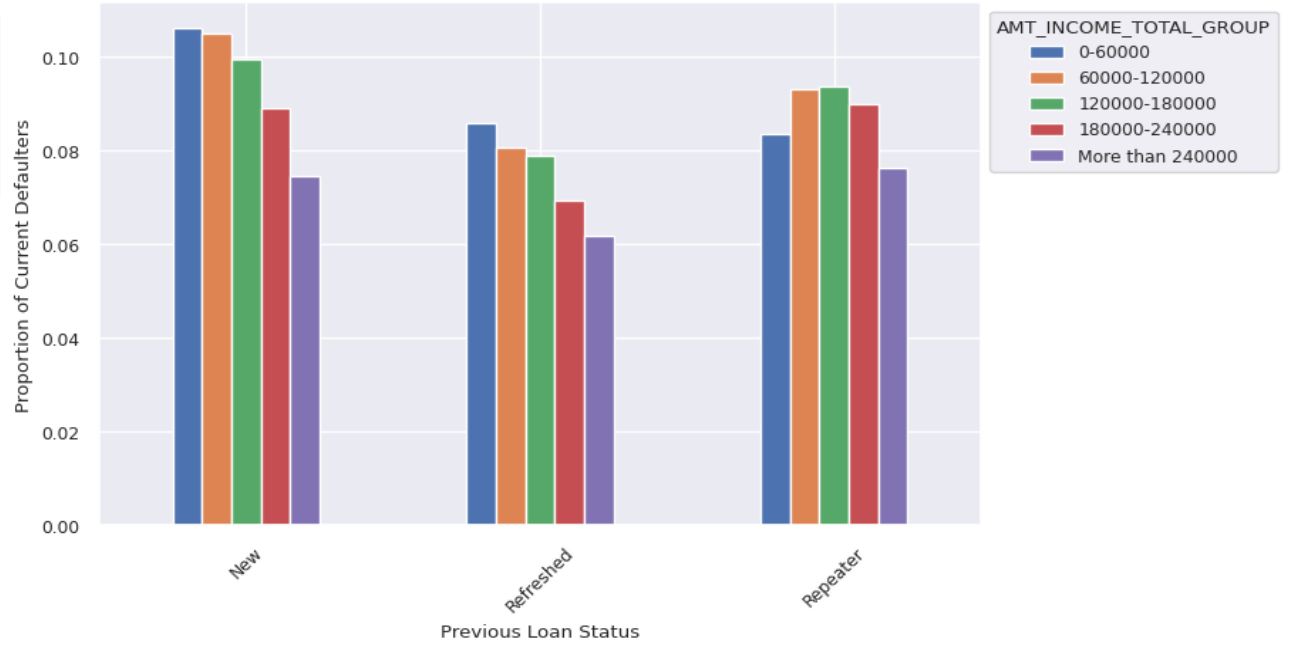

AMT_INCOME_TOTAL_GROUP
0-60000
60000-120000
120000-180000
180000-240000
More than 240000
Proportion of Current Defaulters
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Approved
Canceled
Refused
Unused offer
Previous Loan Status

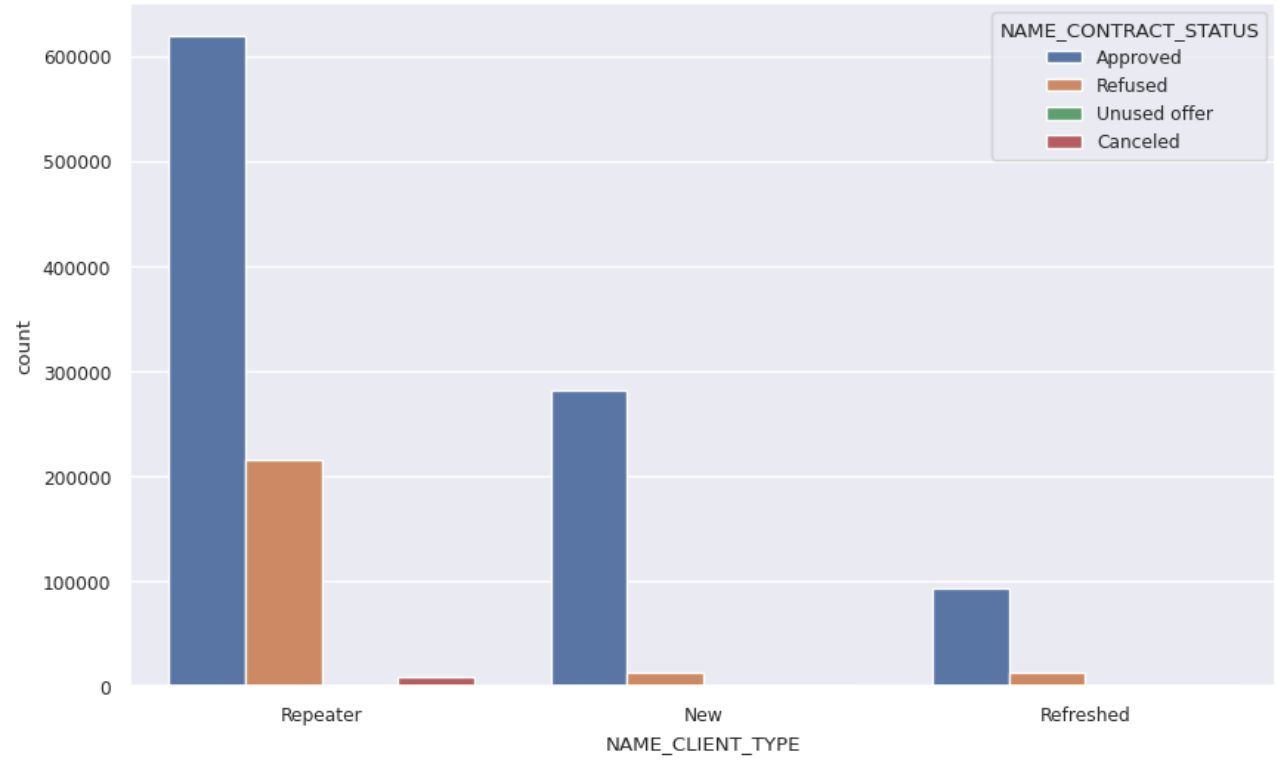

NAME_PORTFOLIO
Cards
Cars
Cash
POS
Proportion of Current Defaulters
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Approved
Canceled
Refused
Unused offer
Previous Loan Status

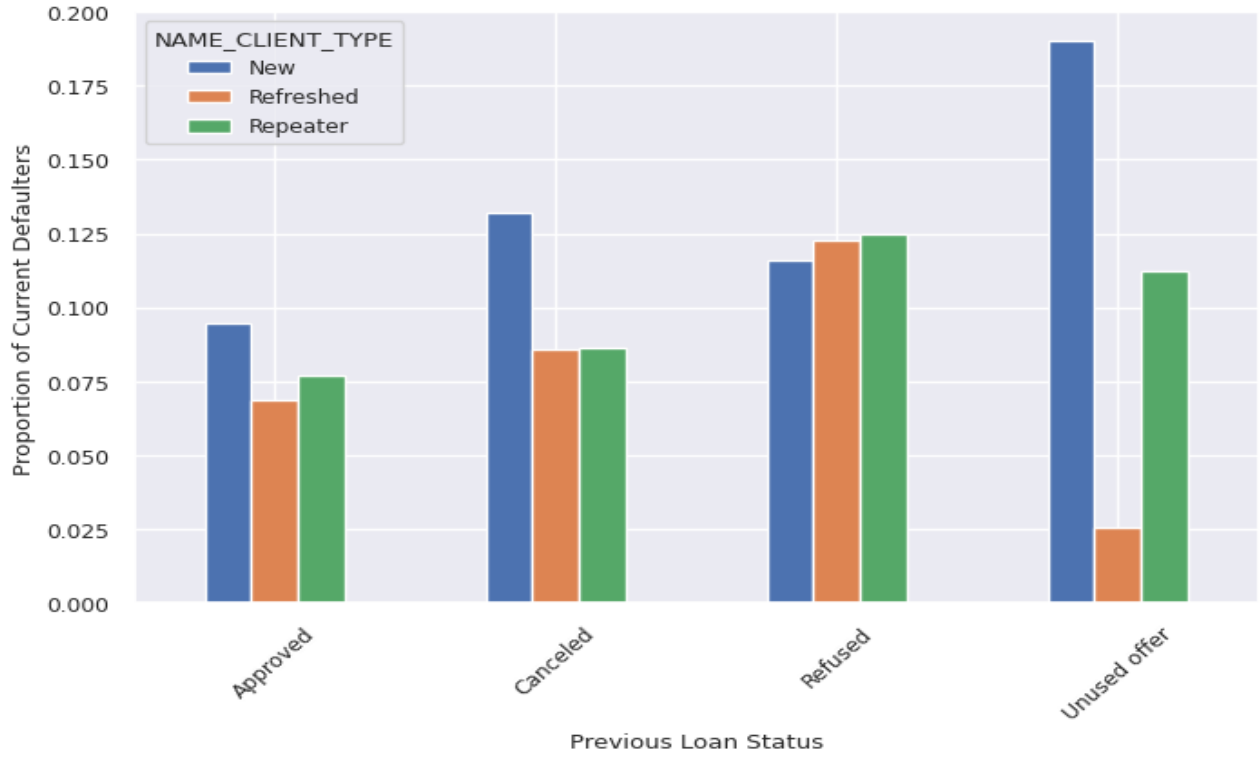

AGE_IN_YEARS_GROUP
10-20
20-30
30-40
40-50
50-60
60-70
Proportion of Current Defaulters
0.5
0.4
0.3
0.2
0.1
0.0
Approved
Canceled
Refused
Unused offer
Previous Loan Status


Previous Credit Amount vs Housing Type
TARGET
0.0
1.0
AMT_CREDIT_x
160000
140000
120000
100000
80000
60000
40000
20000
0
House / apartment
Rented apartment
Municipal apartment
With parents
Office apartment
Co-op apartment
NAME_HOUSING_TYPE


AGE_IN_YEARS_GROUP
10-20
20-30
30-40
40-50
50-60
60-70
Proportion of Current Defaulters
0.12
0.10
0.08
0.06
0.04
0.02
0.00
New
Refreshed
Repeater
Previous Loan Status


AMT_INCOME_TOTAL_GROUP
0-60000
60000-120000
120000-180000
180000-240000
More than 240000
Proportion of Current Defaulters
0.10
0.08
0.06
0.04
0.02
0.00
New
Refreshed
Repeater
Previous Loan Status


NAME_CONTRACT_STATUS
Approved
Refused
Unused offer
Canceled
count
600000
500000
400000
300000
200000
100000
0
Repeater
New
Refreshed
NAME_CLIENT_TYPE


NAME_CLIENT_TYPE
New
Refreshed
Repeater
Proportion of Current Defaulters
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Approved
Canceled
Refused
Unused offer
Previous Loan Status

Here are several key points derived from the graph analysis above:

Default Patterns by Previous Loan Status and Gender:
- Refused clients have a higher default rate compared to approved clients.
- Across all loan statuses, males have a higher default rate than females.

Default Patterns by Previous Loan Status and Client Type:
- Defaulters are more prevalent among clients with a previous loan status of "Unused offers," especially among "New" clients.
- For previous "Approved" statuses, new clients show a higher default rate, followed by repeater clients.
- Among applicants with a previous status of "Refused," defaulters are more common among repeater clients.
- Among applicants with a previous status of "Canceled," new clients have a higher default rate.

Default Patterns by Age Group for Approved Clients:
- For approved clients, younger applicants tend to have a higher default rate.
- Senior applicants among approved clients display a lower default rate.

Default Patterns by Income Range for Unused Offers:
- Among unused offers, applicants with an income range of 180,000-240,000 have a higher default rate, while incomes exceeding 240,000 show the least defaults.
- Across all loan statuses, default rates are relatively consistent across various income groups.

Default Patterns by Loan Status and Portfolio Type:

- Clients who refused applied loans for cards exhibit a higher default rate.
- For approved loan statuses, clients applying for cars have a lower default rate.
- Clients who refused applied loans for cars also show a lower default rate.

Housing Type and Default Rates:

- Office apartments have a higher credit default rate (Target 1), while municipal apartments show a higher credit repayment rate (Target 0). This suggests caution when granting loans for office apartments.

Client Types and Approved Loans:

- Repeater clients have a higher count of approved loans compared to New and Refreshed clients.

Default Patterns by Previous Loan Status and Client Type (Continued):

- Among clients previously categorized as "Unused offers," there is a higher occurrence of defaulters, particularly among "New" clients.
- For clients with a previous loan status of "Approved," "New" clients have a higher count of defaults, followed by "Repeater" clients.
- Among applicants with a previous loan status of "Refused," instances of defaulters are higher among "Refreshed" clients.
- Among those with a previous loan status of "Canceled," the count of defaulters is higher among "New" clients.

Age Groups and Default Rates (Continued):

- Across all previous loan statuses, younger applicants exhibit a higher frequency of defaults.
- Senior applicants show a lower incidence of defaults compared to other age groups, regardless of the previous loan status.

Income Ranges and Default Rates:

- New applicants with an income range of 0-60,000 tend to have a higher rate of defaults.
- Among repeater applicants, those with an income range of 0-60,000 show a lower default rate compared to other client types.
- Repeater applicants with an income exceeding 240,000 exhibit a higher default rate compared to other income ranges.

Portfolio Types and Default Rates:

- Among the cards portfolio, new applicants display the highest default rate.
- Within the cars portfolio, refreshed applicants have no instances of default.
- Among different client types in the cards portfolio, refreshed applicants exhibit the lowest default rate compared to others.

Key Insights for Risk Assessment and Lending Strategies

Leading Factors in Defaults:

- Medium income levels, age groups of 20-30 and 30-40, being male, and occupation types like business type 3 and self-employment are associated with higher default rates.

Family and Education Factors:

- Married clients with five or more children are at an elevated risk of defaulting.
- Default rates are highest among males with lower secondary education.
- Maternity leave applicants tend to have a higher likelihood of default when they have a larger number of children.

Unused Applications and Age:

- Unused applications are more prevalent in the higher age group, potentially explaining their underutilization.

Gender and Default Rates:

- Female applicants have lower default rates, suggesting they could be given extra consideration.

Working Applicants and Caution:

- While more defaulters are working applicants, this does not imply outright refusal. Further examination of other parameters is necessary.

Educational Background Consideration:

- Banks should prioritize applicants with a background of 'Higher education' while exercising caution with those who have completed 'Secondary/secondary special,' as they tend to face difficulties in making payments.

Income Type and Risk:

- It is advisable for banks to be cautious with clients with an income type of 'Working' due to their higher likelihood of payment difficulties. Instead, they should focus on clients classified as 'Commercial associate,' 'Pensioner,' and 'State servant.'

Housing Type and Payment Issues:

- To reduce payment difficulties, banks should pay more attention to clients residing in 'House/apartment' housing types, as they have fewer payment issues.

Channel Types and Loan Outcomes:

- Banks should note that the 'Country-wide' channel type results in more approved loans, whereas the 'Credit and cash offices' channel type is associated with a higher number of canceled and refused loans.