CREDIT RISK ASSESSMENT

PYTHON EDA PROJECT

PRESENTED BY,
SWETHA KIZHAVANA JOSEPH
BUSINESS ANALYST INTERN

ABOUT ME

Name: Swetha Kizhavana Joseph

Education:

- MSc Data Analytics
- MSc Statistics
- BSc Mathematics

Programming & Data Analysis: Python, C, R, SAS, SQL, Advanced Excel

Visualisation: Tableau & Power BI

Areas of Interest:

- Data Analytics
- Business Intelligence
- Data Visualization
- Predictive Modelling

INTRODUCTION

- Evaluating credit risk is crucial in the modern financial environment to reduce potential losses and maintain the stability of lending institutions while ensuring that credit is extended to deserving applicants.

- This project which was conducted using Python (Jupyter Notebook) in Visual studio Code focuses on Loan Credit Risk Assessment, where the goal is to identify patterns and factors that contribute to loan defaults. By leveraging Python libraries like NumPy, Pandas, Matplotlib, and Seaborn to perform Exploratory Data Analysis (EDA), we aim to gain insights into client behaviour, credit history, and other key variables that influence the likelihood of default.

- The analysis includes cleaning and preparing two datasets: one containing current loan application data and the other detailing previous loan applications. A comprehensive Exploratory Data Analysis (EDA) was performed on both datasets individually, as well as on the merged dataset, to uncover trends and correlations that may indicate higher credit risk.

- The insights gained from this project aim to support more informed lending decisions, which could lower the rate of loan defaults and enhance the overall quality of the loan portfolio.

PROBLEM STATEMENT

- The goal of this project is to identify patterns that indicate a client's difficulty in repaying loans. By analysing these patterns, we can make informed decisions such as denying loans, adjusting loan amounts, or offering higher interest rates to high-risk clients. This helps ensure that loans are granted to clients who are likely to repay them, thus minimizing financial losses.

- Objectives: Identify key factors that signal loan default and use these insights to improve portfolio and risk assessment.

- Business Impact: Reduce losses by targeting risky applicants and optimise lending strategies to maintain a healthy loan portfolio.

DATA OVERVIEW

01

application_data.csv

- **Current Loan Applications**: This dataset includes detailed information on current loan applications, such as client demographics, income, loan amount, and other financial indicators. Each row represents a unique loan application.
- **Rows**: 307,511 **Columns:** 122

02

previous_application.csv

- **Previous Loan Applications:** This dataset contains records of previous loan applications made by clients, providing insights into their borrowing history and past interactions with the lending institution.
- **Rows:** 1,670,214, **Columns**: 37

DATA CLEANING & PREPARATION

- **Initial Data Exploration** such as statistical summary, displaying basic information etc was conducted.

- **Handling Missing Values:** Calculated the percentage of missing values in each column. Columns with more than 50% missing data were dropped to avoid introducing bias. For columns with fewer missing values, imputation techniques such as Mean or median imputation for Numerical Data, depending on the distribution and Mode imputation for Categorical Data are used.

- **Dropping Unnecessary Columns**: Analysed and Removed columns that were not relevant to the analysis, such as those with redundant information organised the dataset and focused on key variables.

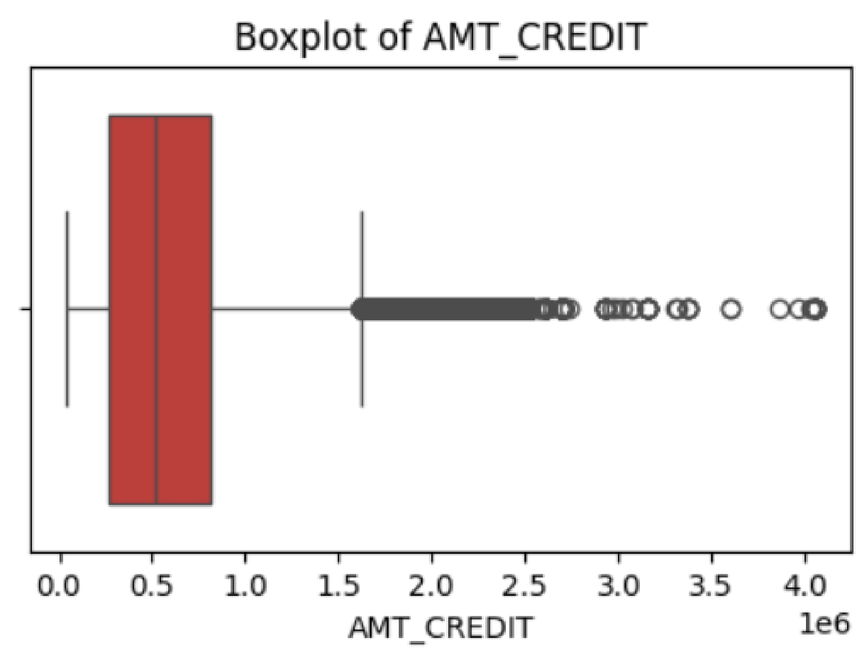
- **Filtering and Enhancing Data**: Applied filters to focus on relevant records and added new columns derived from existing data to capture additional insights, such as categorised variables.

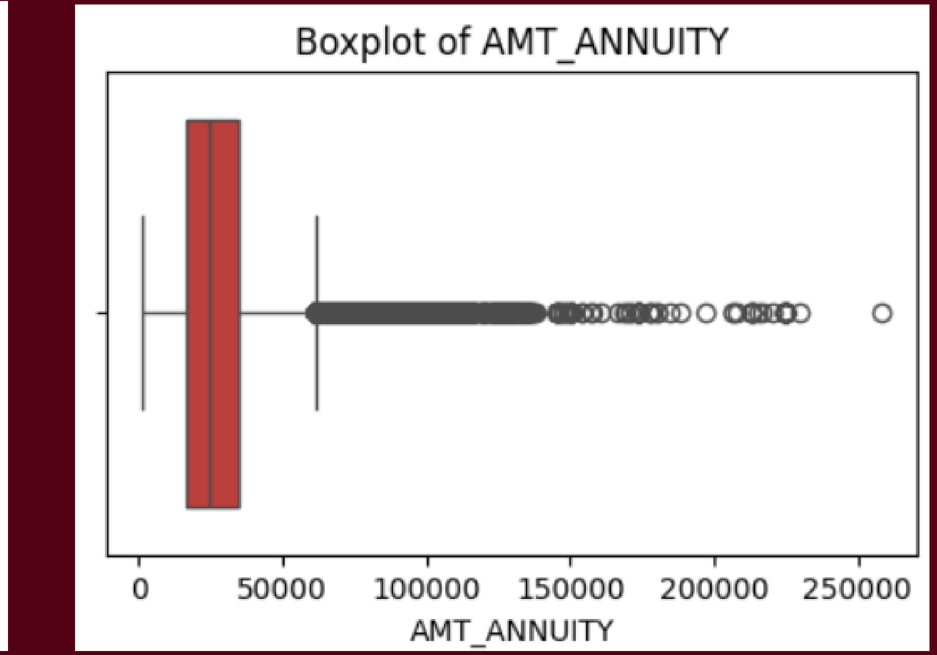
- **Duplicates**: Checked for duplicates and no duplicate records were found which ensured the integrity of the dataset.

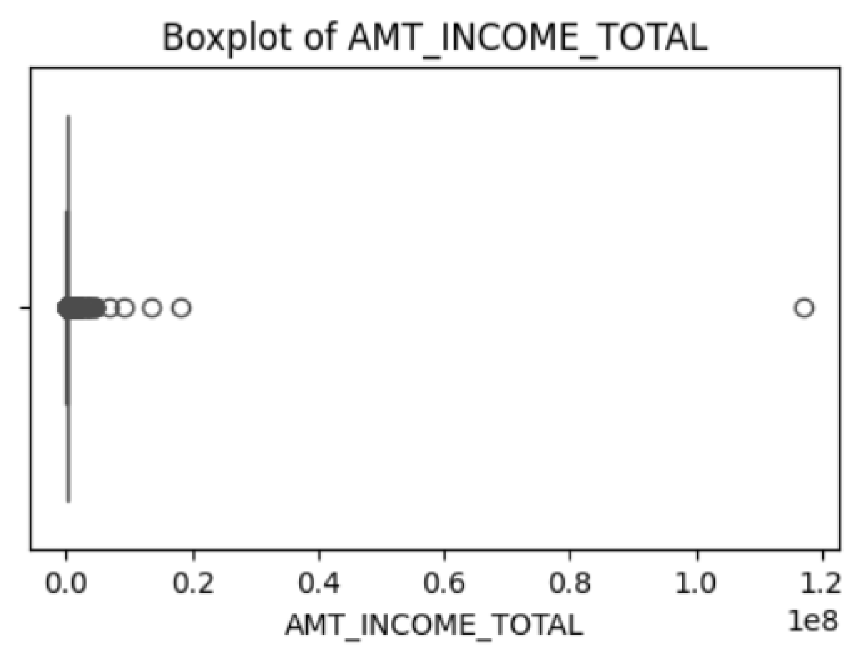
- **Correct Data Types**: Transformed data to suitable types and ensured numerical columns were accurately typed for efficient processing and storage.

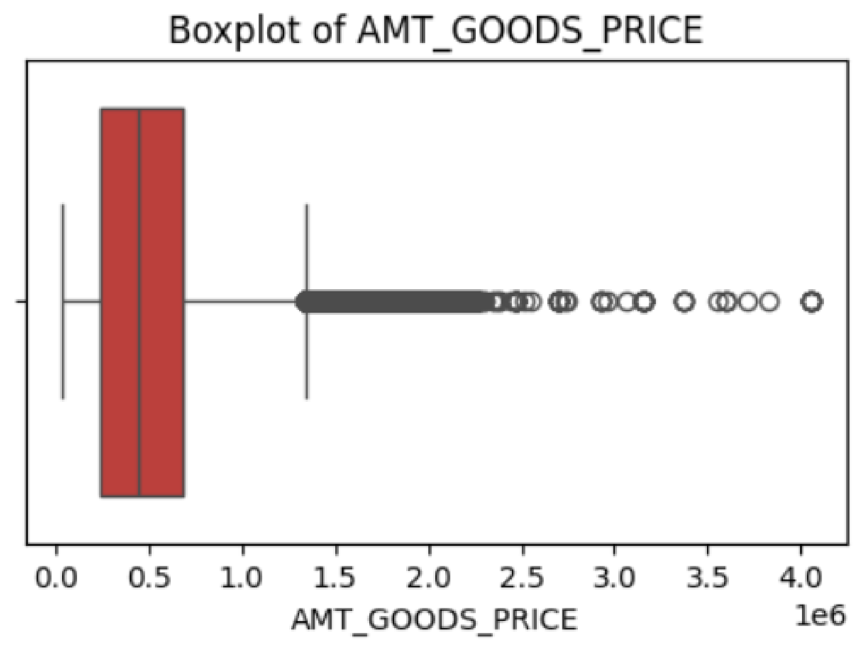
OUTLIER DETECTION USING BOXPLOTS (application_data)

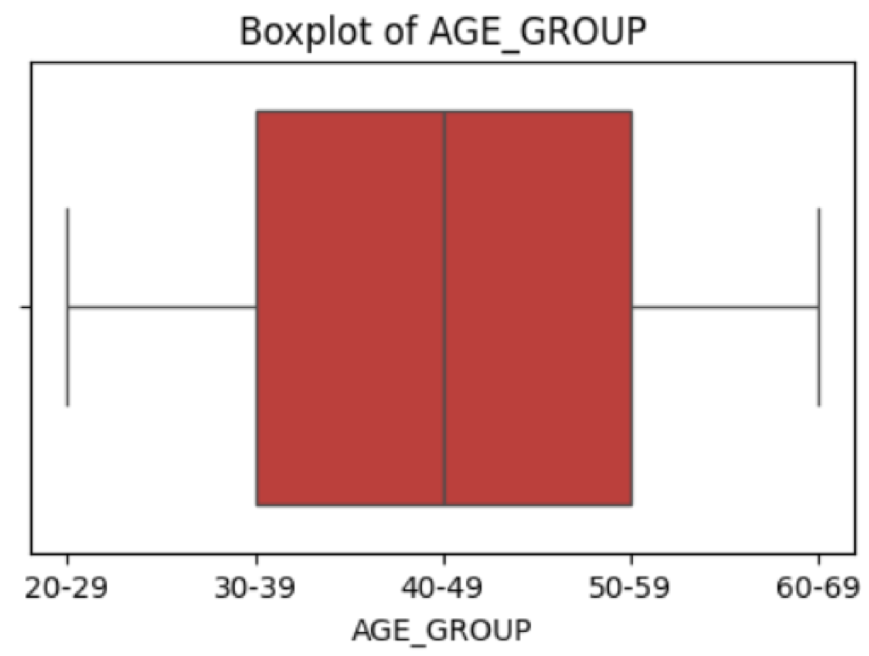
The boxplots show outliers in key financial variables like AMT_CREDIT and AMT_ANNUITY, while AGE_GROUP displays a consistent distribution with no outliers.

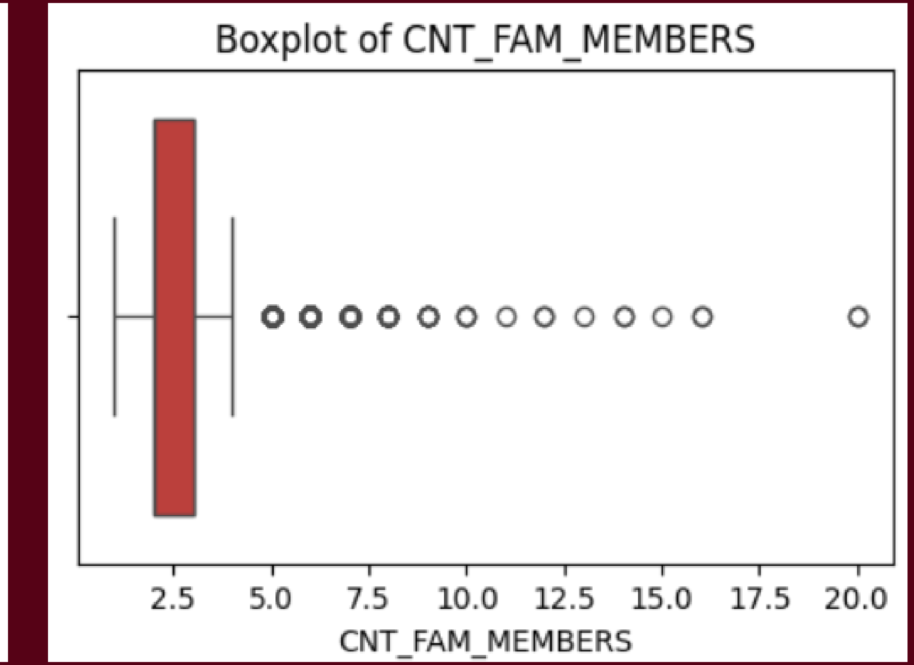


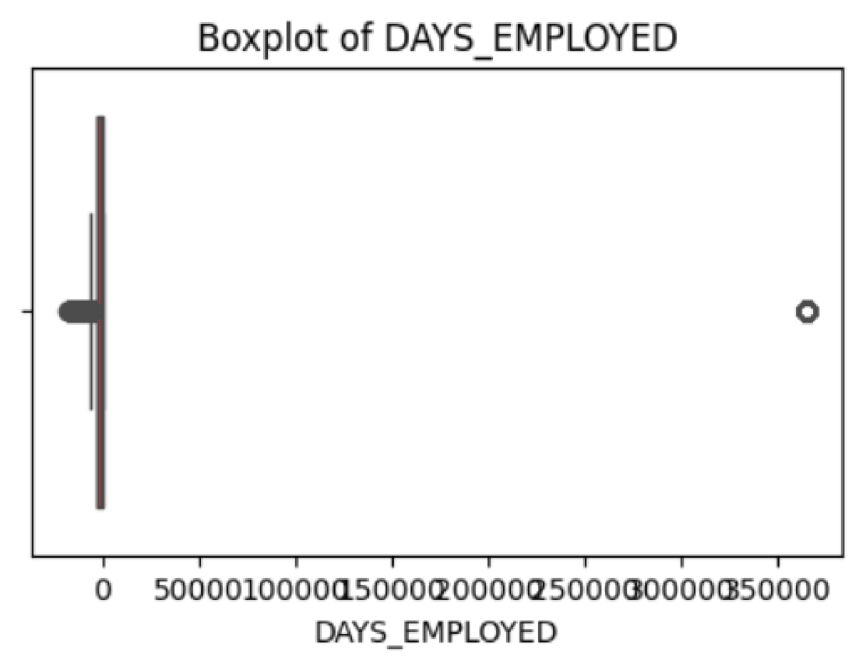


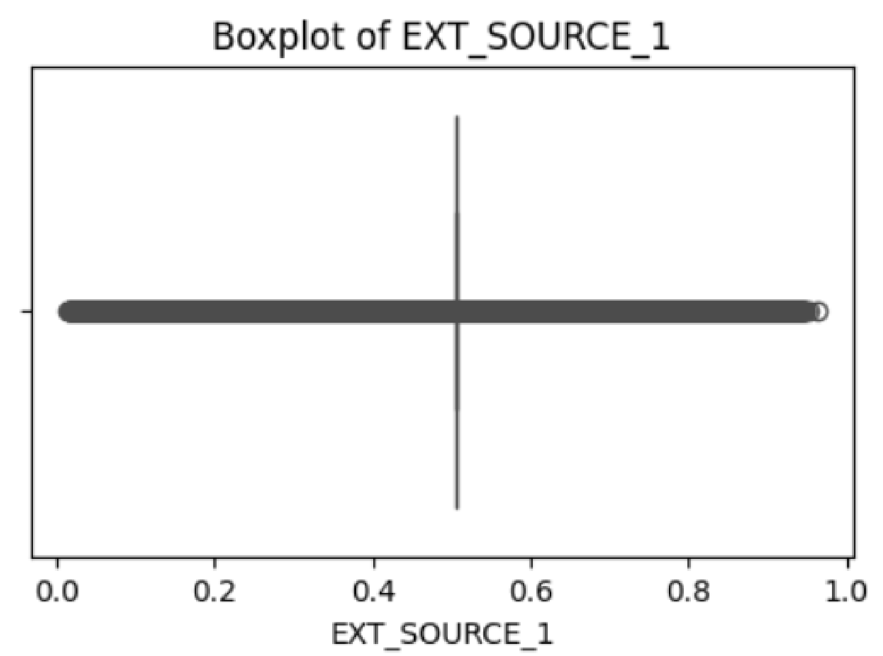






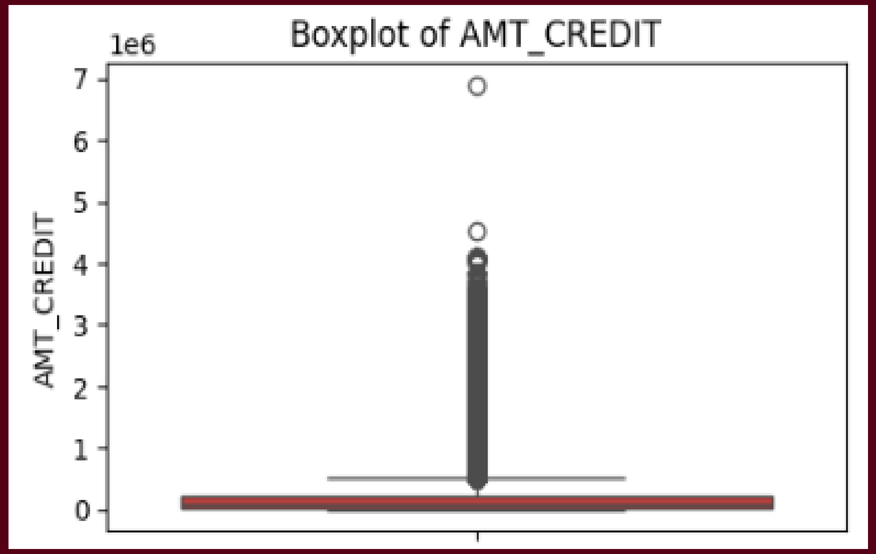


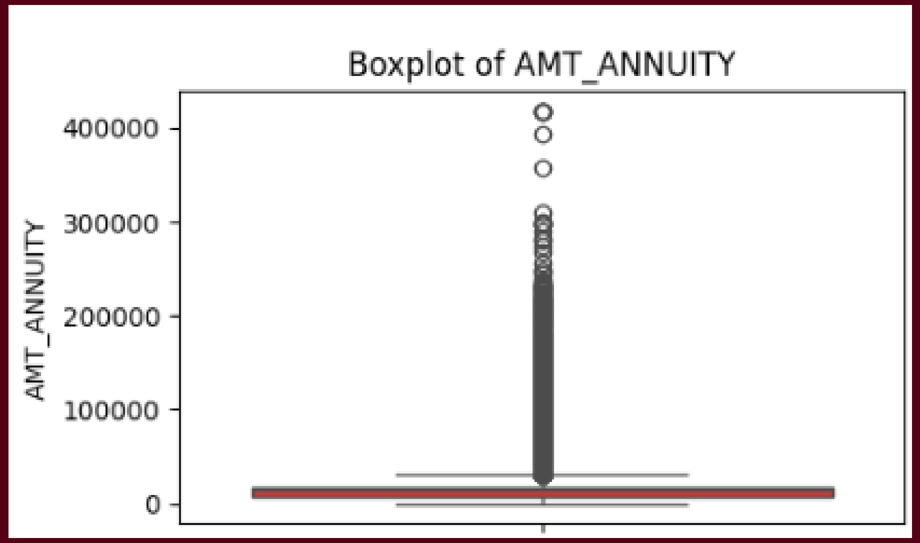


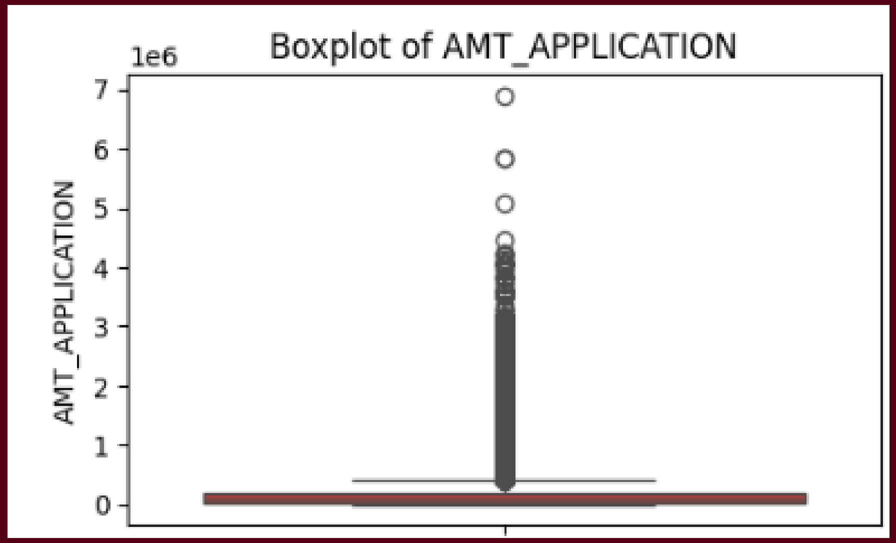


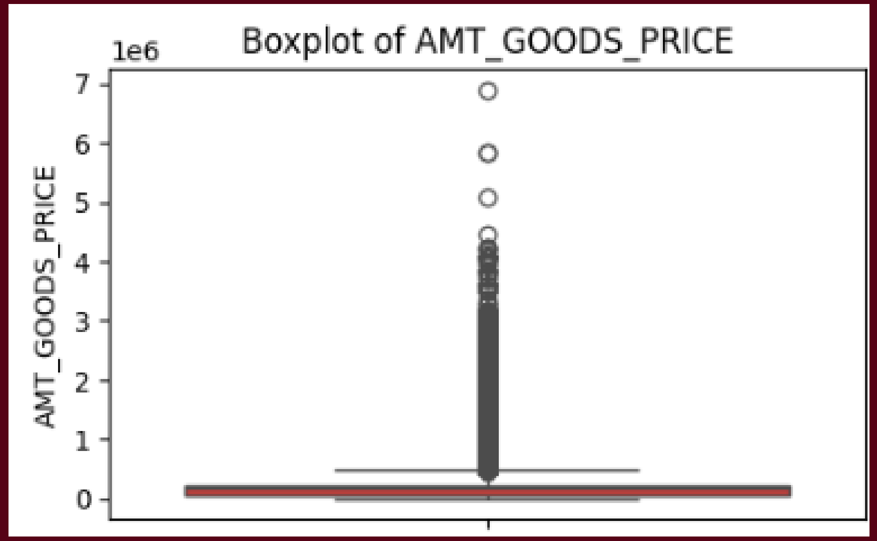
OUTLIER DETECTION USING BOXPLOTS (previous_application)

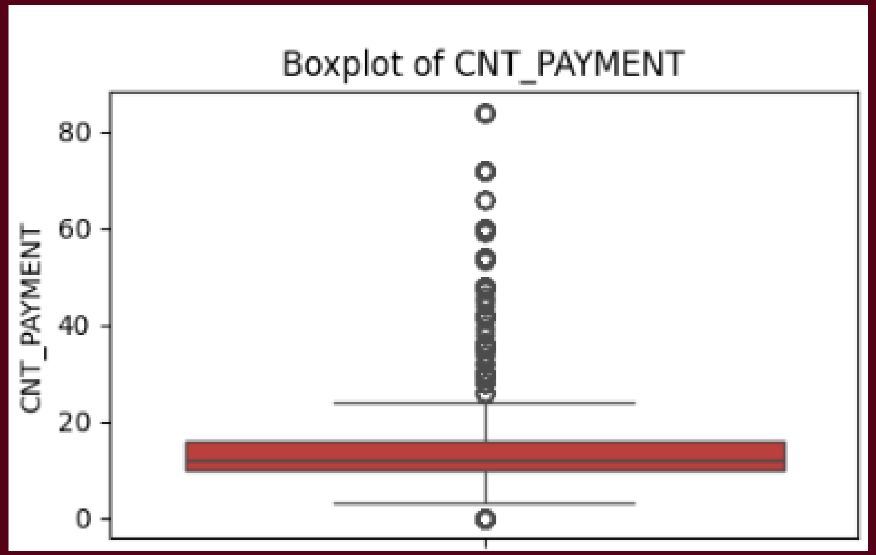
All the features in the boxplots—AMT_CREDIT, AMT_ANNUITY, AMT_APPLICATION, AMT_GOODS_PRICE, CNT_PAYMENT, DAYS_DECISION, and DAYS_TERMINATION show significant outliers.

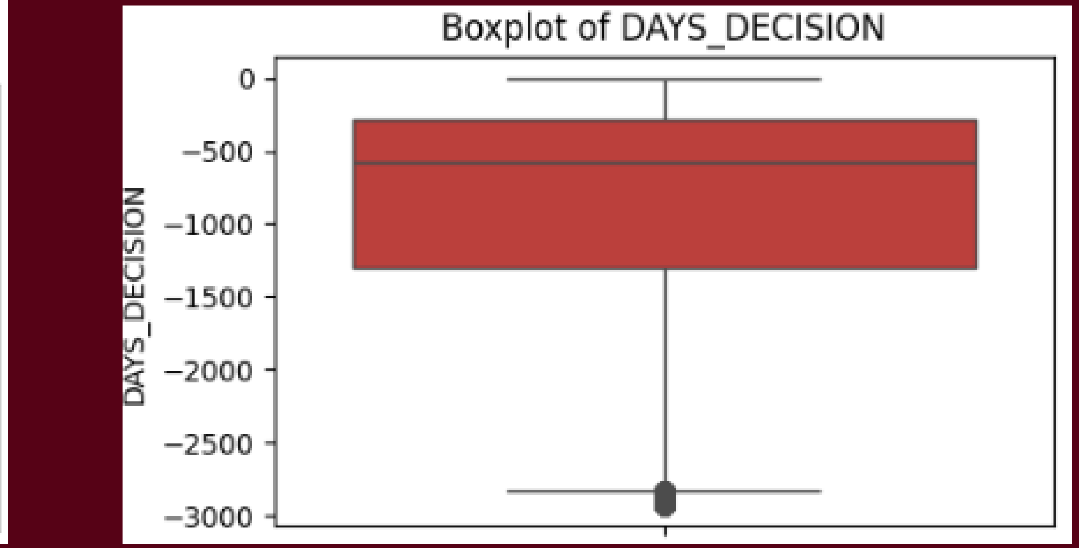


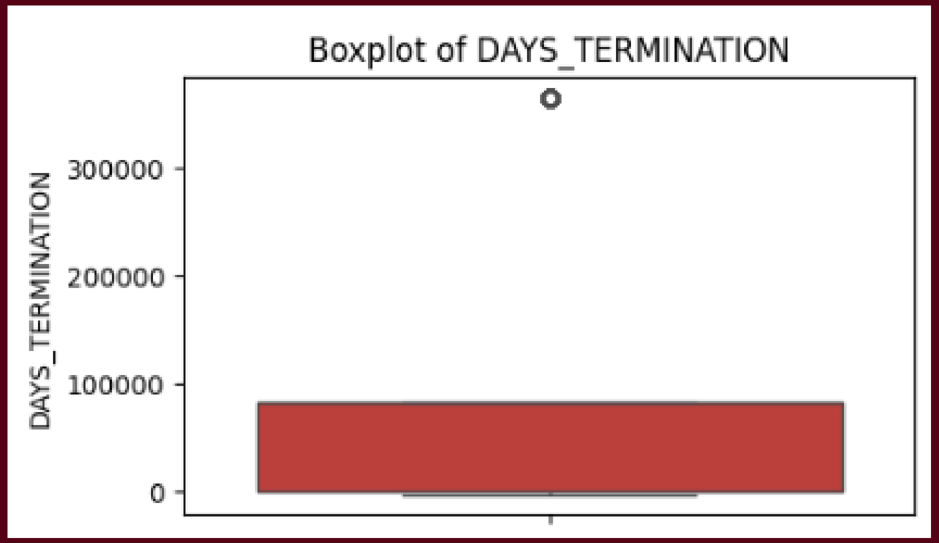












OUTLIER DETECTION USING IQR METHOD

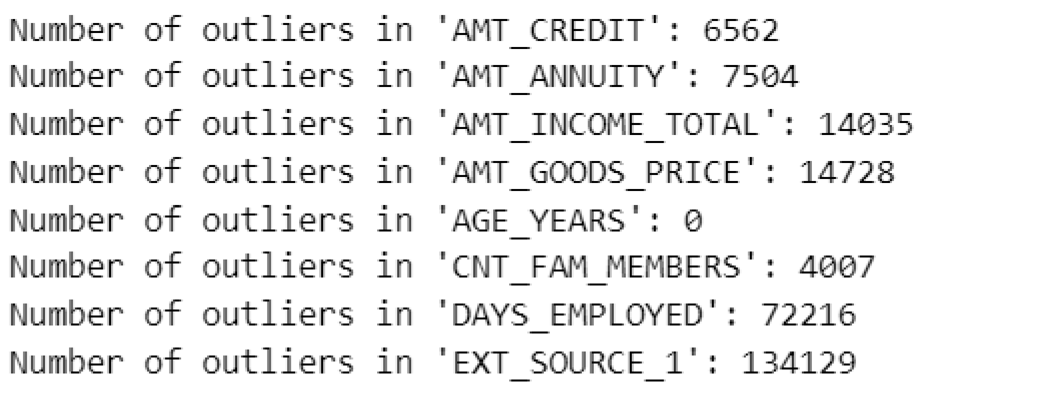
Interquartile Range (IQR) Method is used to identify and quantify outliers in key financial and demographic variables to understand data variability and potential anomalies better.

Interquartile Range (IQR) Method:

- Calculate each numerical variable's first quartile (Q1) and third quartile (Q3)
- IQR = Q3 - Q1
- Lower bound = Q1 – (1.5 x IQR)
- Upper bound = Q3 + (1.5 x IQR)
- Outliers are data points falling outside these bounds.

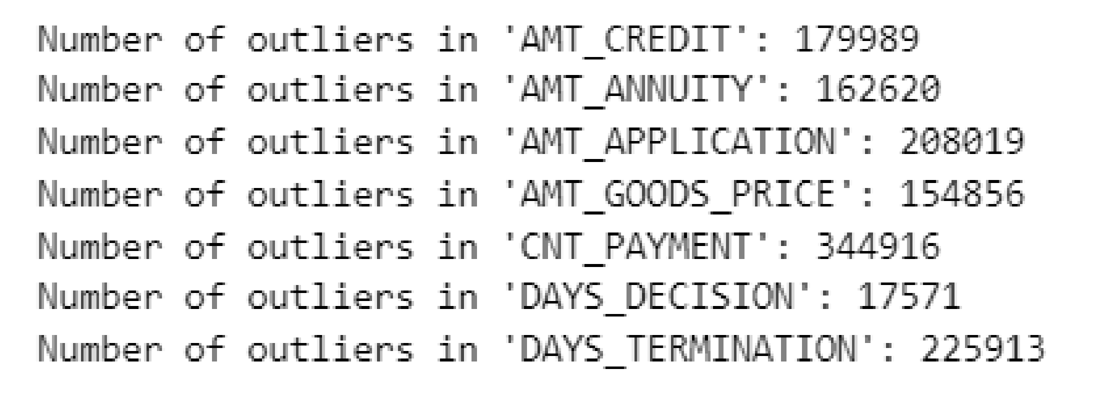
- **Results :**

application_data.csv



```
Number of outliers in 'AMT_CREDIT': 6562
Number of outliers in 'AMT_ANNUITY': 7504
Number of outliers in 'AMT_INCOME_TOTAL': 14035
Number of outliers in 'AMT_GOODS_PRICE': 14728
Number of outliers in 'AGE_YEARS': 0
Number of outliers in 'CNT_FAM_MEMBERS': 4007
Number of outliers in 'DAYS_EMPLOYED': 72216
Number of outliers in 'EXT_SOURCE_1': 134129
```

previous_application.csv

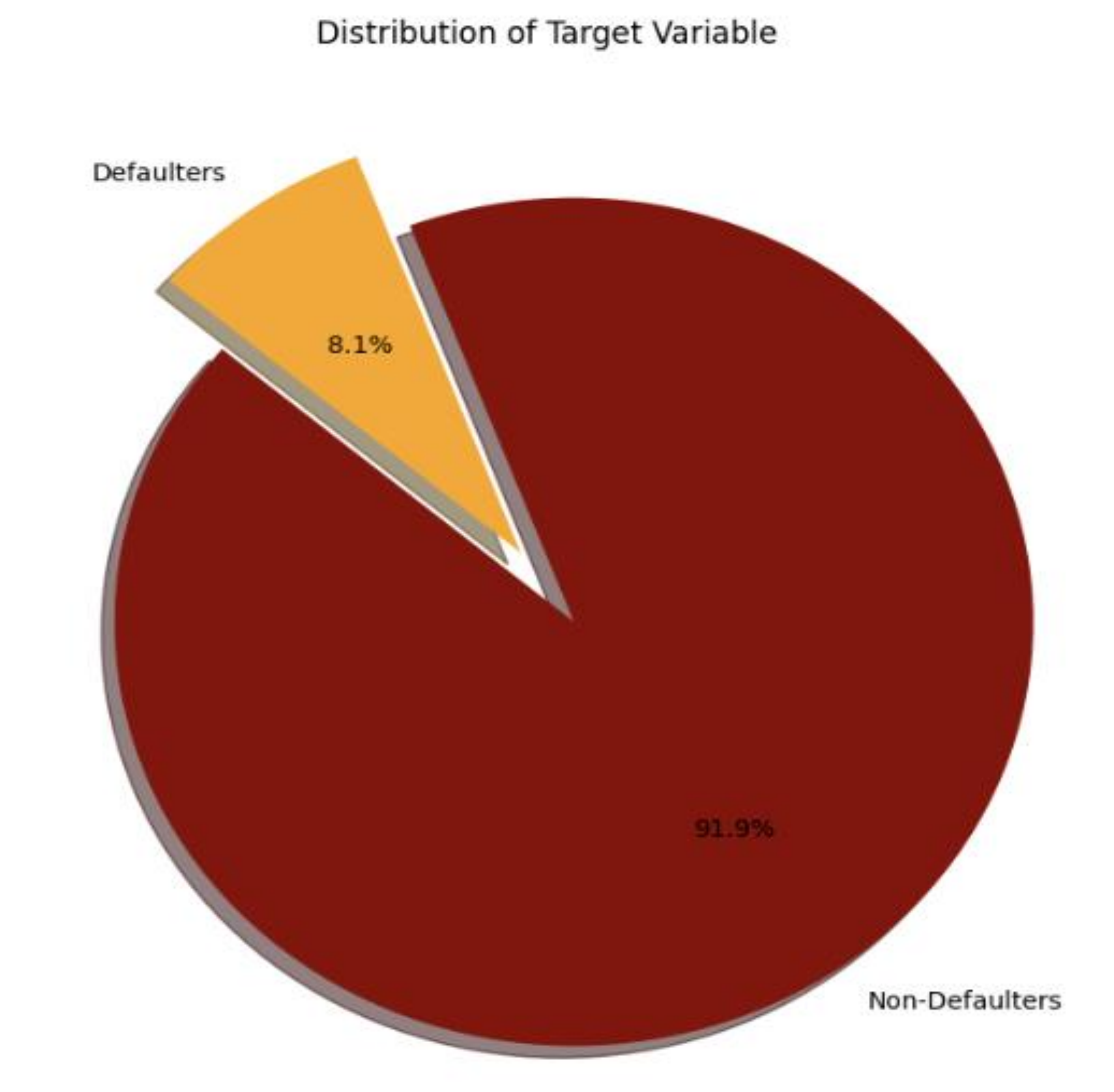


```
Number of outliers in 'AMT_CREDIT': 179989
Number of outliers in 'AMT_ANNUITY': 162620
Number of outliers in 'AMT_APPLICATION': 208019
Number of outliers in 'AMT_GOODS_PRICE': 154856
Number of outliers in 'CNT_PAYMENT': 344916
Number of outliers in 'DAYS_DECISION': 17571
Number of outliers in 'DAYS_TERMINATION': 225913
```

The two datasets contain a significant number of outliers across multiple features, indicating potential extreme values that could heavily influence analysis outcomes, requiring transformation or removal to ensure the accuracy of the analysis.

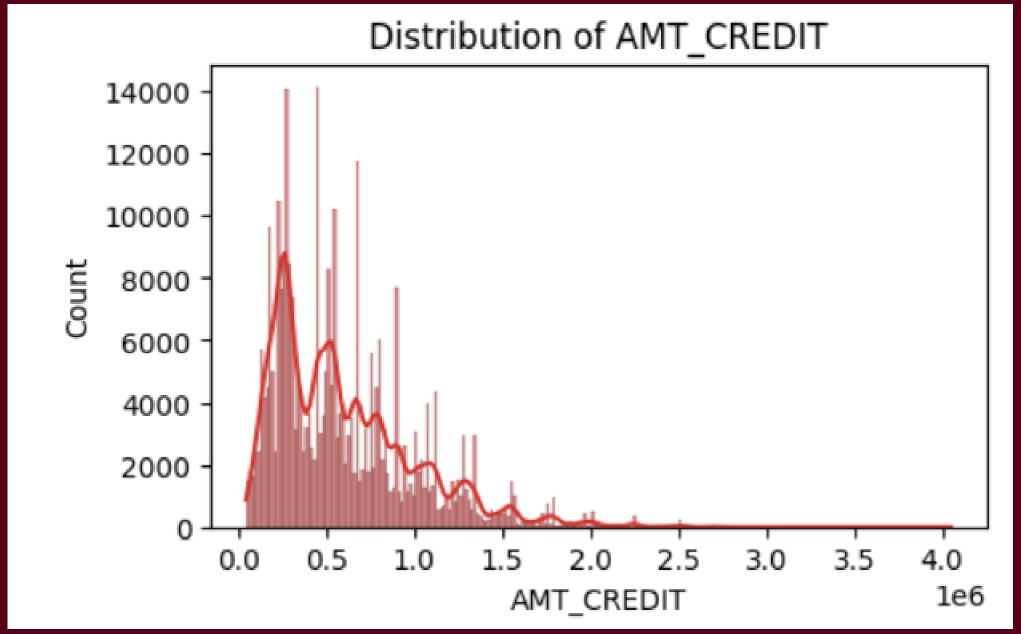
EXPLORATORY DATA ANALYSIS (application_data)

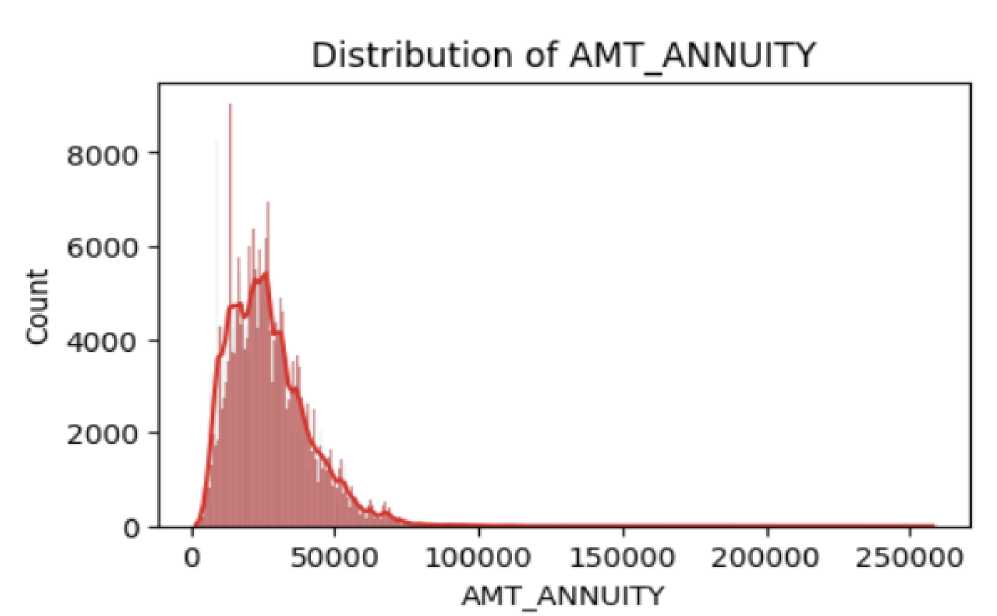
To identify key differences and risk factors affecting the Loan default, split data into Defaulters (TARGET = 1) indicating clients who have defaulted on their loans. and Non-Defaulters (TARGET = 0) indicating clients who have not defaulted for comparative analysis.

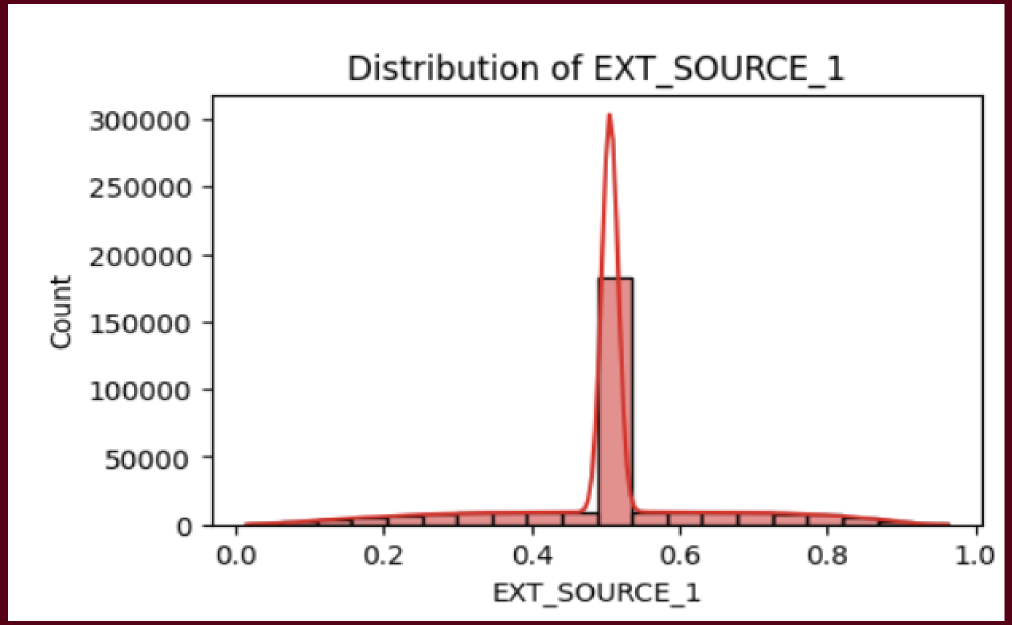


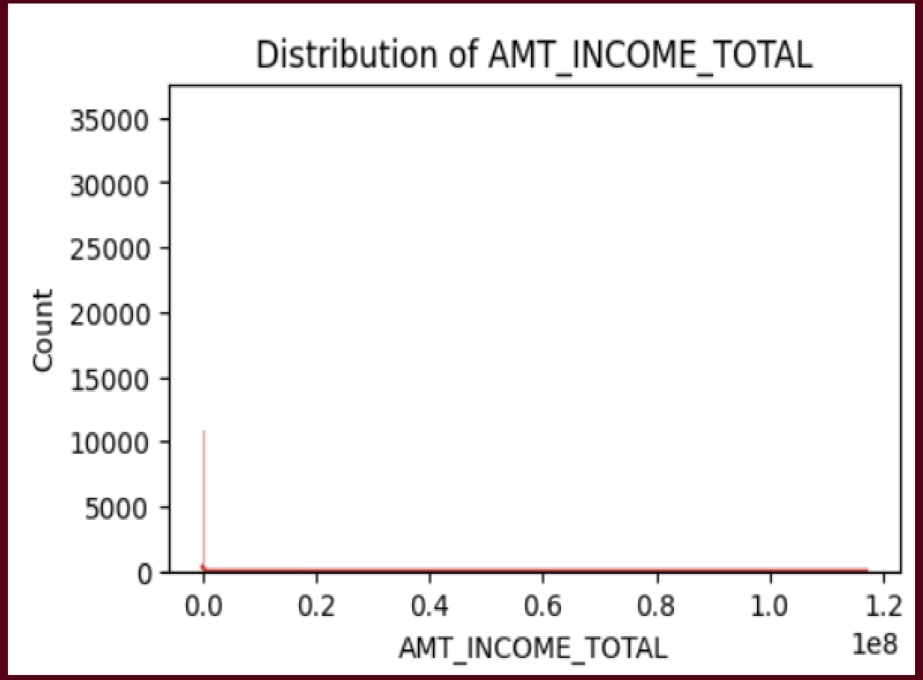
The pie chart shows the distribution of the **TARGET** variable, with far more non-defaulters (91.9%) than defaulters (8.1%), indicating that risk factors for loan defaults might be less common and require careful analysis to identify.

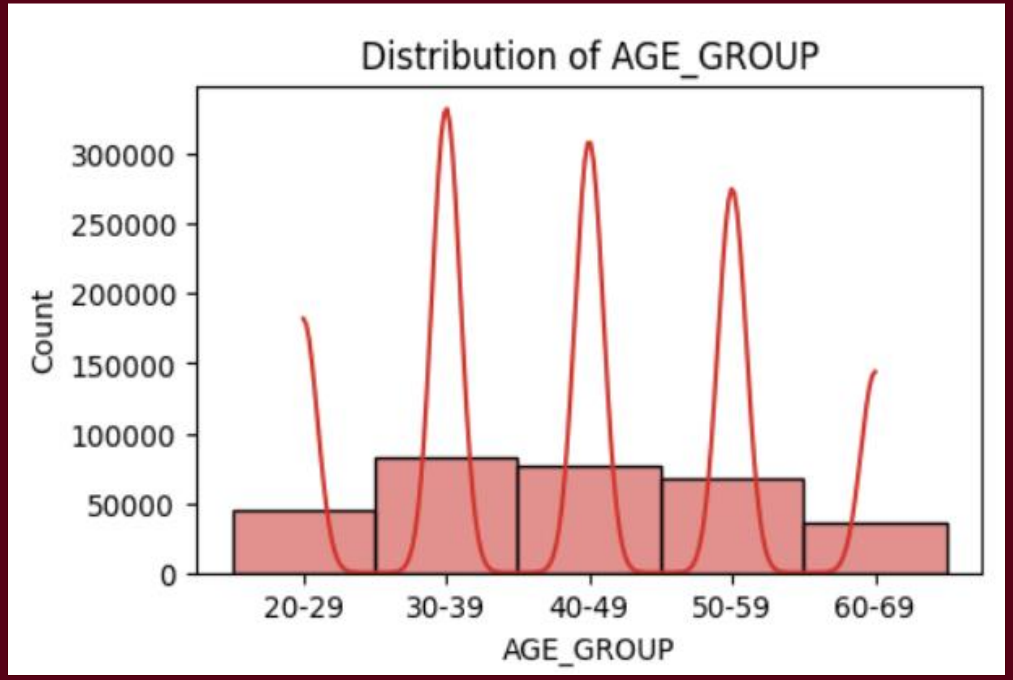
RISK ANALYSIS: NUMERICAL FEATURES (application_data)











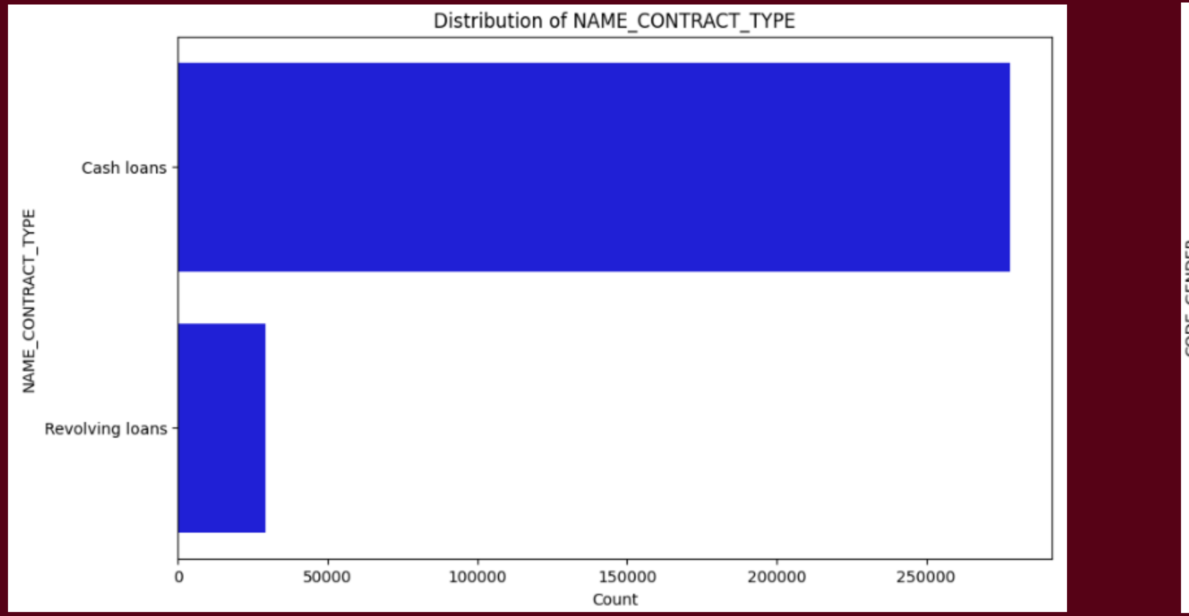
- The right-skewed distributions of AMT_CREDIT, and AMT_ANNUITY, show most clients have lower amounts, with a few high-value cases that may represent higher risk.

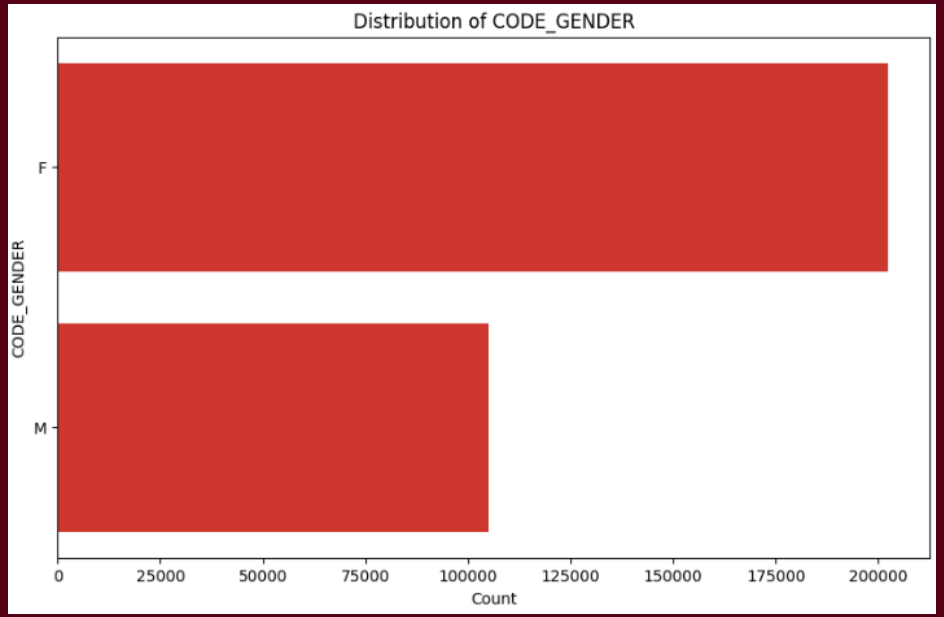
- The distribution of loan applicants is fairly even across the 30-59 age range, with fewer applicants in the 20-29 and 60-69 age groups.

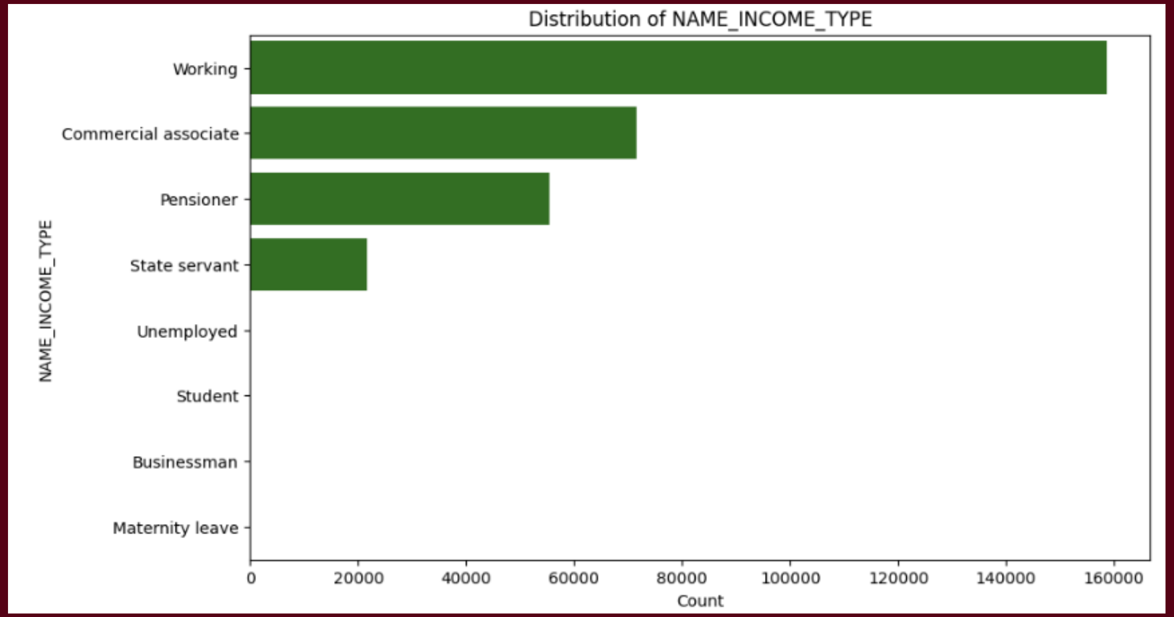
- The extreme right-skew in AMT_INCOME_TOTAL, with a few clients having exceptionally high incomes, may suggest overstated incomes or unusually high loan-to-income ratios, increasing default risk.

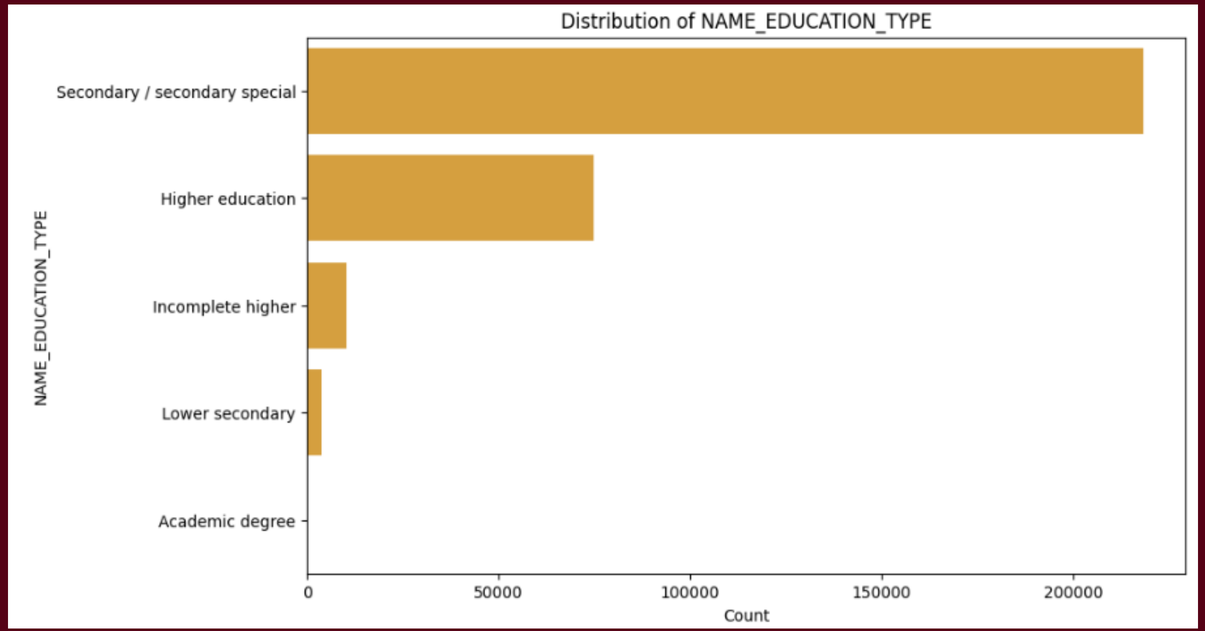
- The sharp, narrow distribution of EXT_SOURCE_1 indicates most clients have similar external credit scores, limiting its usefulness in distinguishing risk.

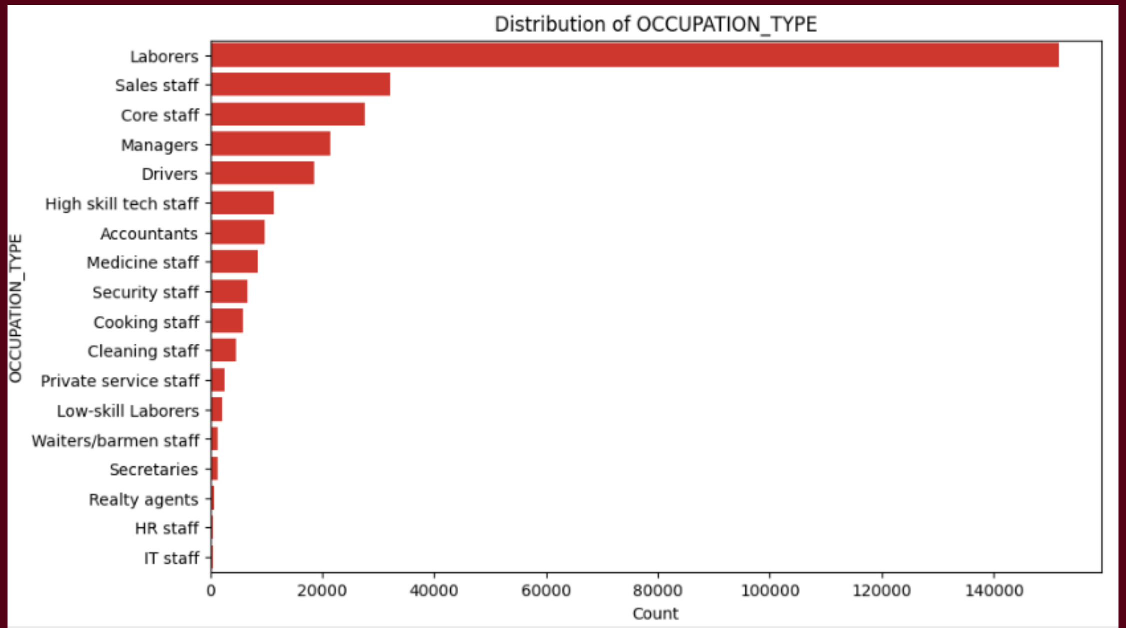
RISK ANALYSIS: CATEGORICAL FEATURES (application_data)





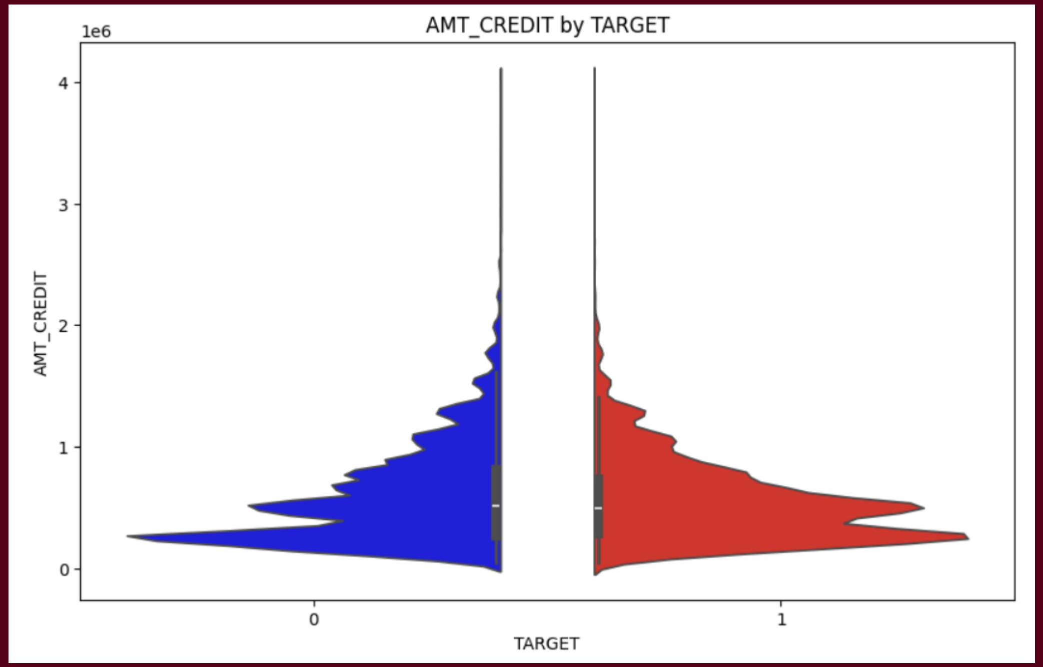


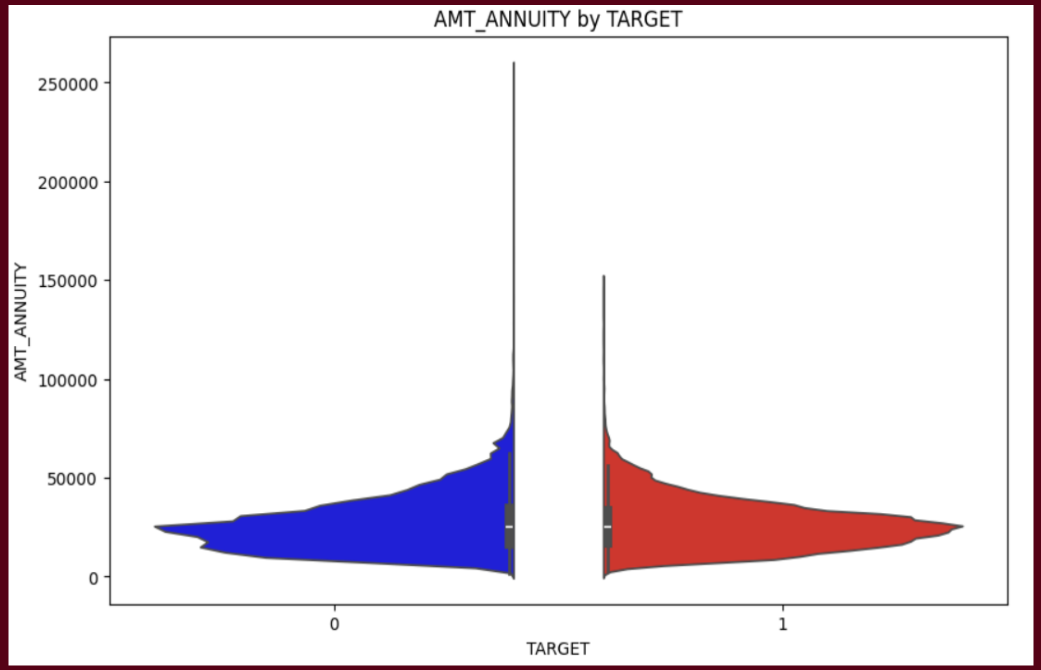


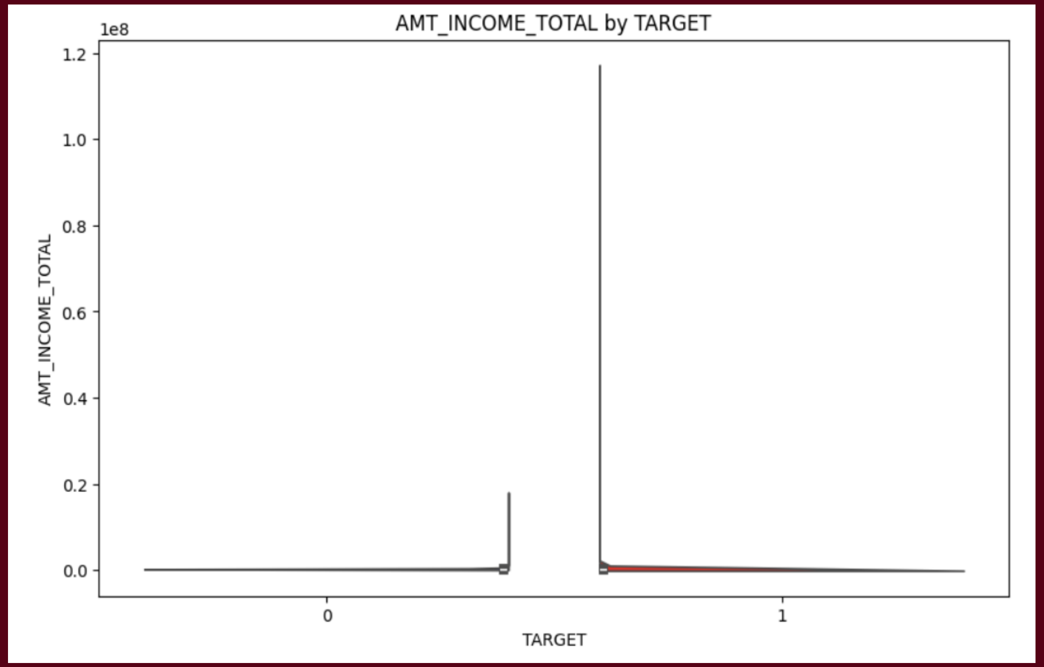


The distribution plots show that most loans are cash loans, with a higher proportion of female applicants and "Working" income types dominating. Key risk factors to consider include loan type, gender, income source, education level, and occupation, as they might influence the likelihood of loan default.

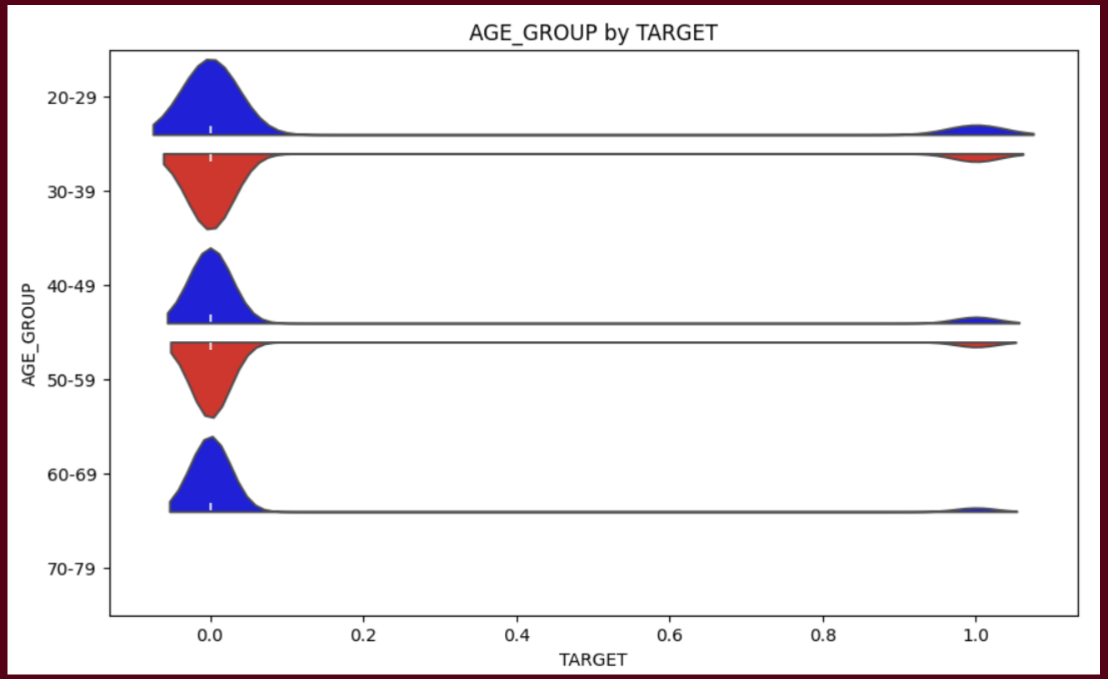
IMPACT OF NUMERICAL FEATURES ON LOAN DEFAULT (application_data)

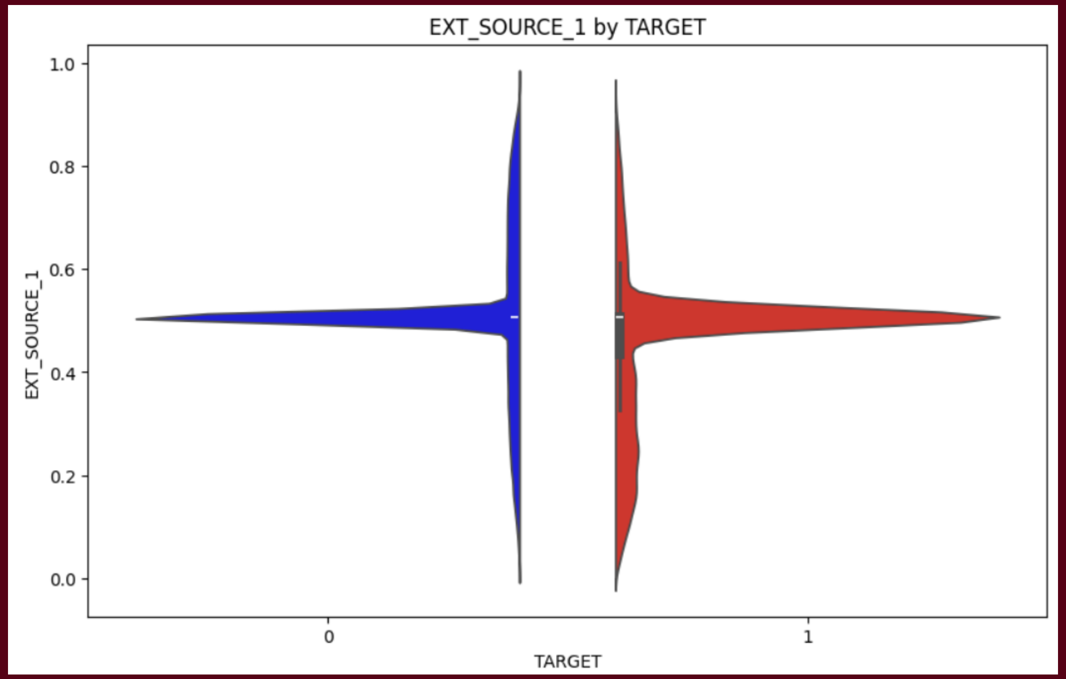




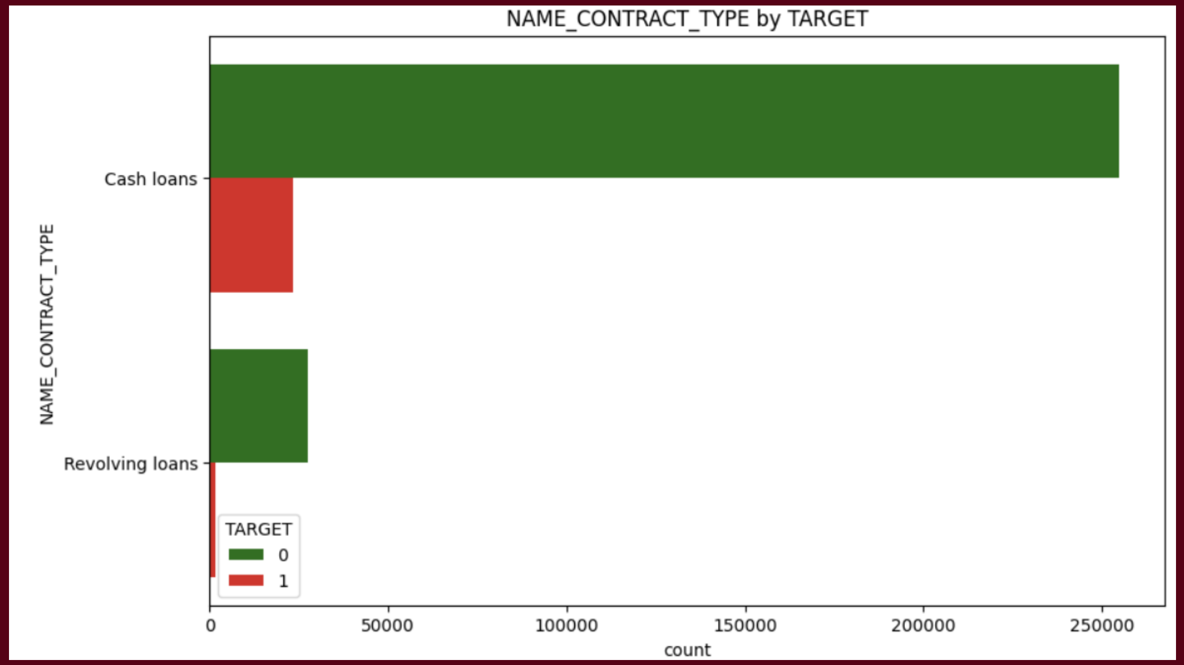


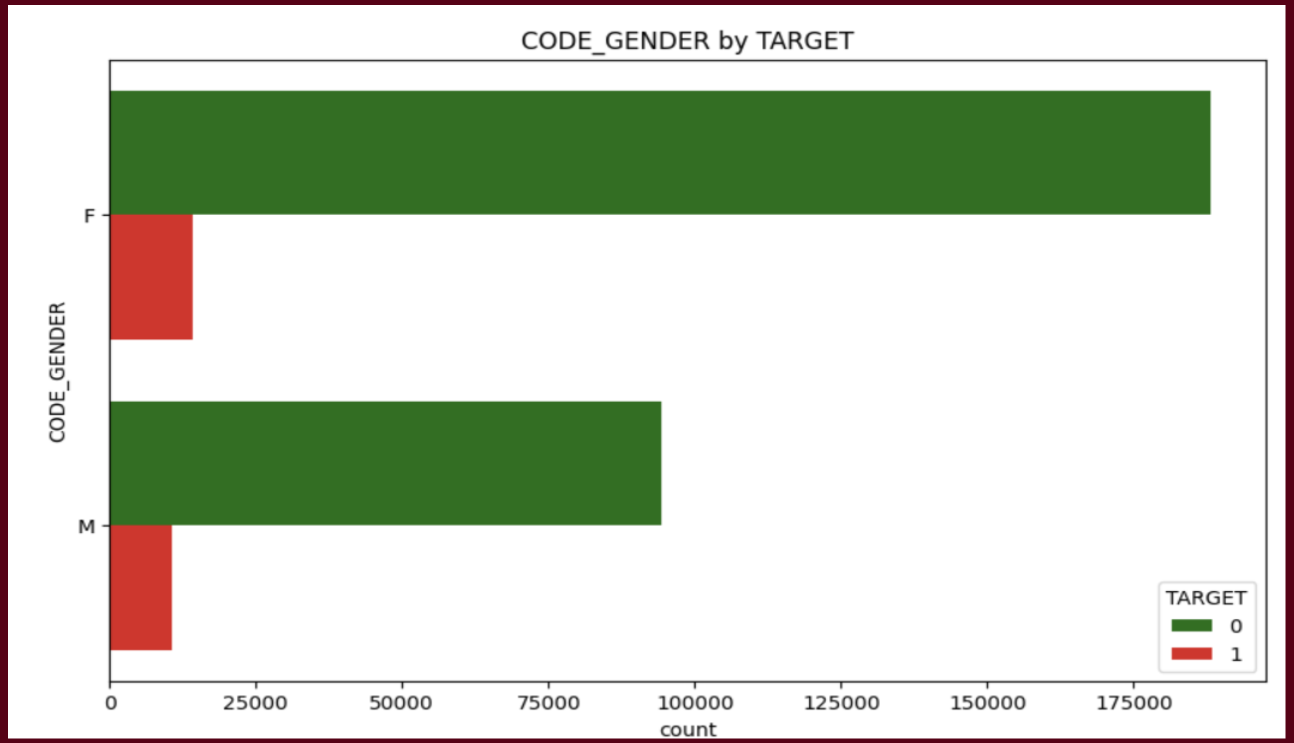
- AMT_CREDIT and AMT_ANNUITY distributions show similar ranges for defaulters and non-defaulters, with defaulters slightly leaning towards lower amounts.
- AMT_INCOME_TOTAL indicates high-income outliers in both groups, but the majority have lower incomes, suggesting income may not be a strong differentiator.
- AGE_GROUP reveals that younger age groups (20-39) have a higher proportion of defaulters, implying age may influence default risk.
- EXT_SOURCE_1 shows defaulters generally have slightly lower external credit scores, indicating this score could be crucial in risk assessment.





IMPACT OF CATEGORICAL FEATURES ON LOAN DEFAULT (application_data)

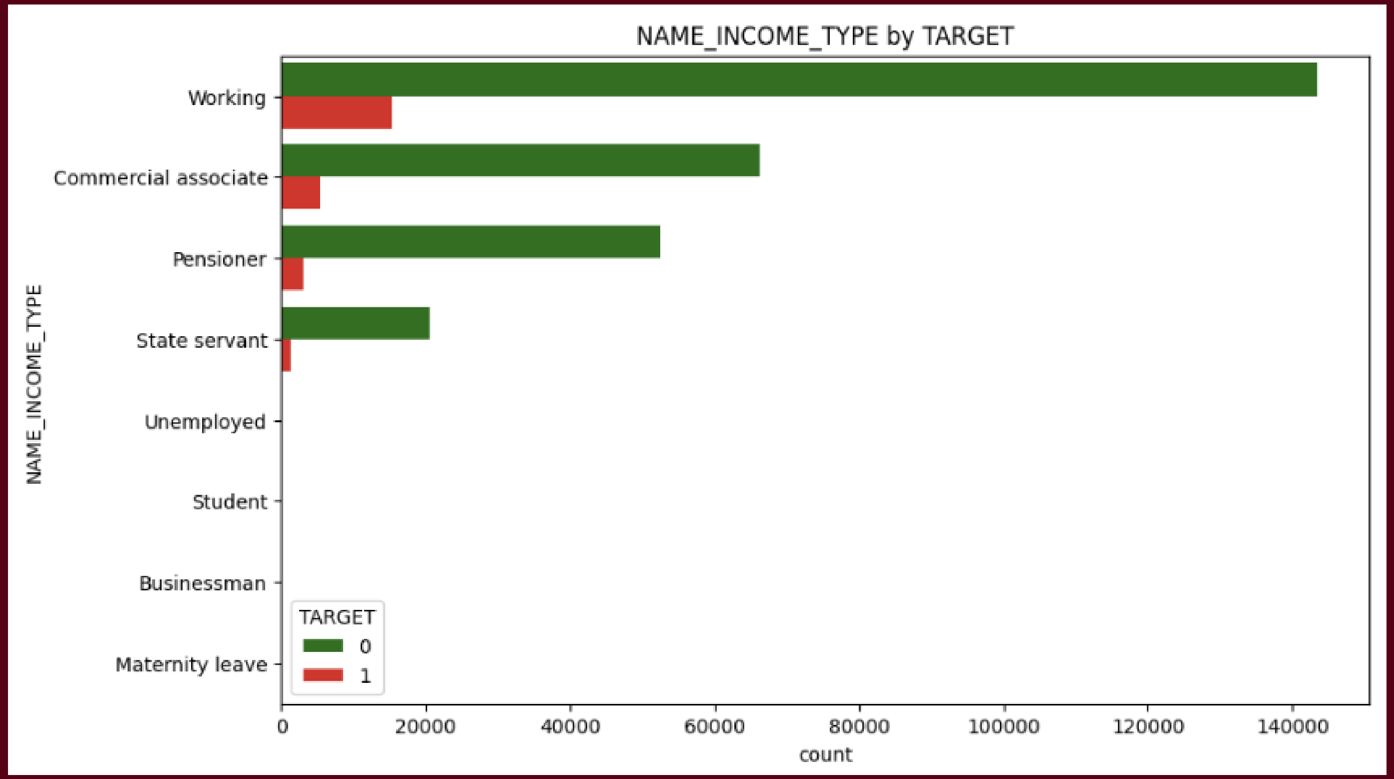




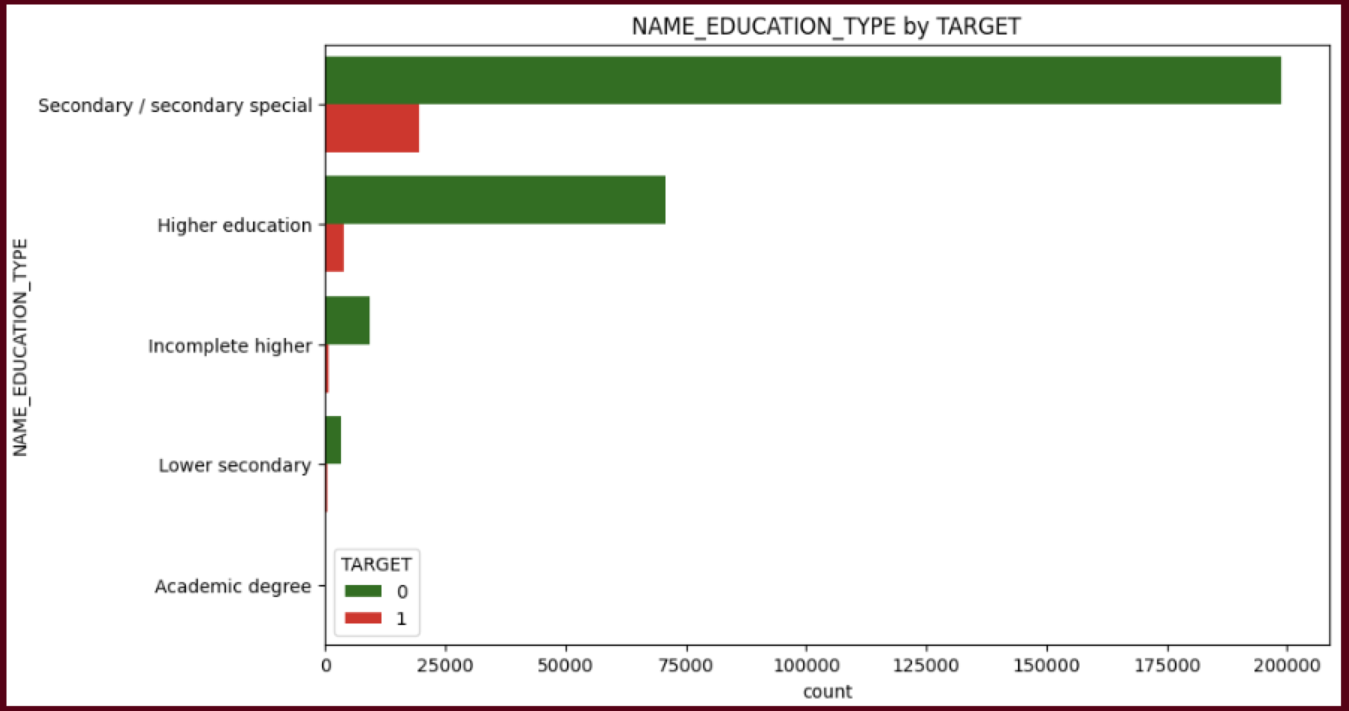
- NAME_CONTRACT_TYPE: Cash loans are far more common than revolving loans, and the majority of cash loans are repaid without default, though there is a notable number of defaults in this category.

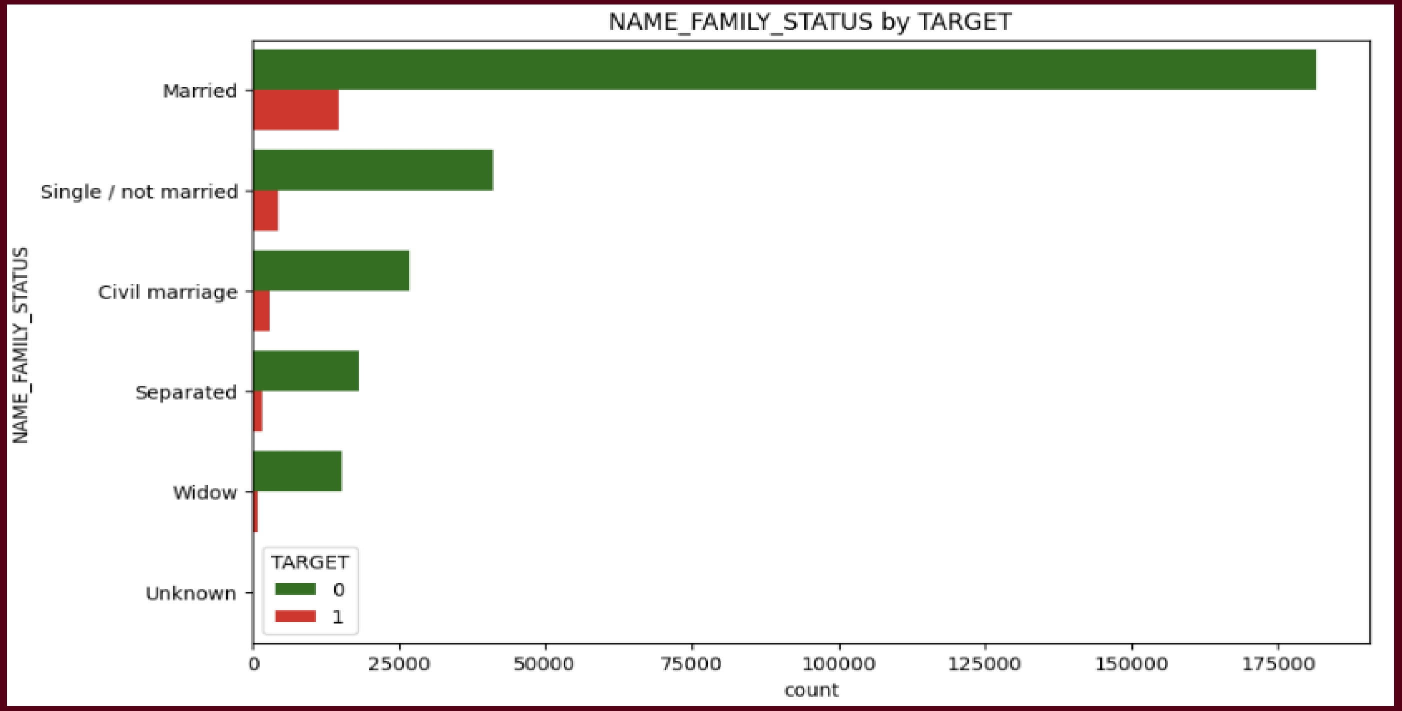
- CODE_GENDER: Females appear to have a higher count of non-defaults compared to males, but both genders have similar proportions of defaults, indicating gender may not be a significant differentiator in repayment difficulties.

- NAME_INCOME_TYPE: Working individuals and Commercial associate show a higher count of non-defaulters, implying stable income sources contribute to easier loan repayment.



IMPACT OF CATEGORICAL FEATURES ON LOAN DEFAULT (application_data)

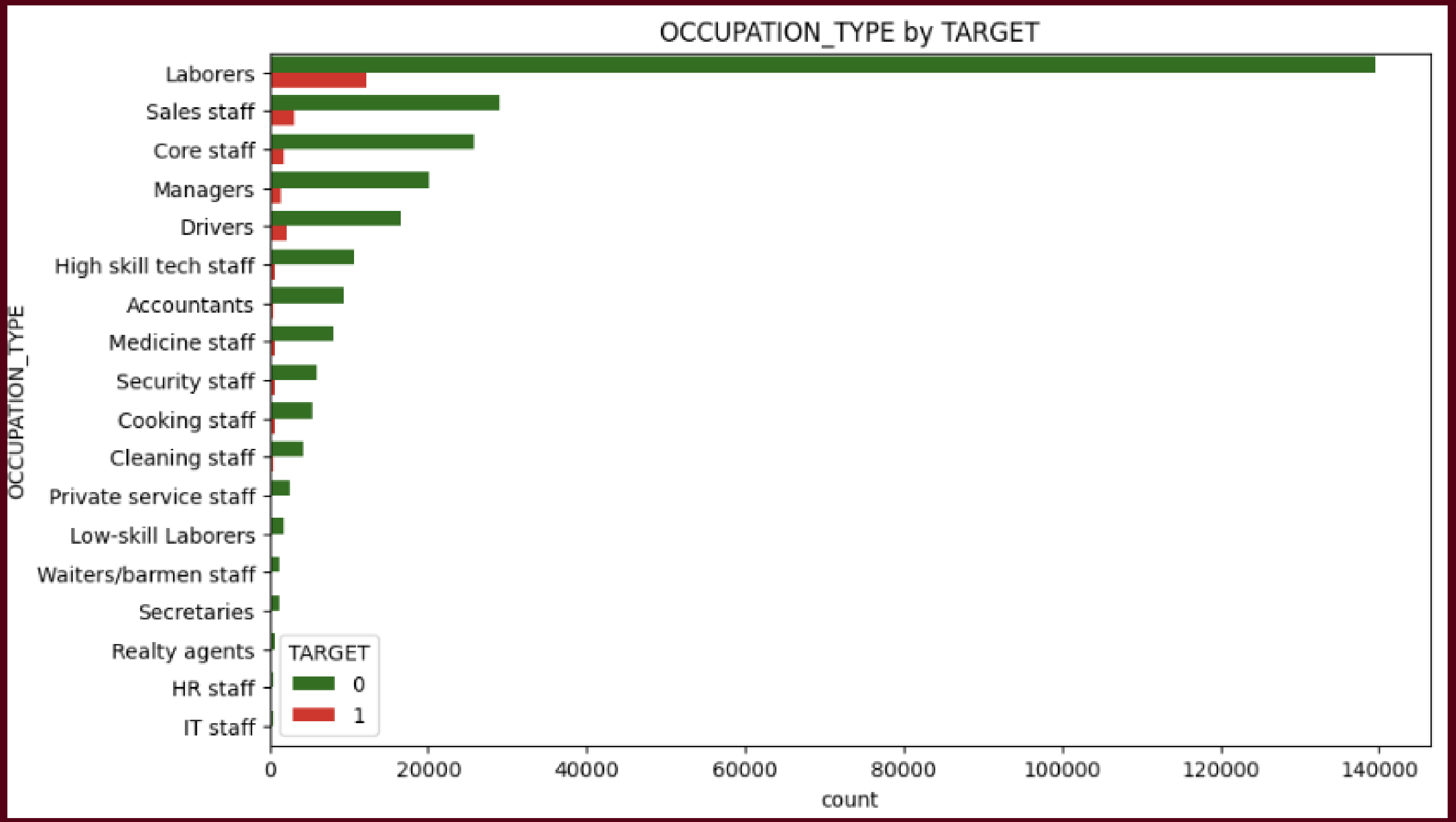




- NAME_EDUCATION_TYPE: Individuals with higher education or secondary special education tend to have fewer repayment difficulties, possibly due to better job opportunities and financial literacy.

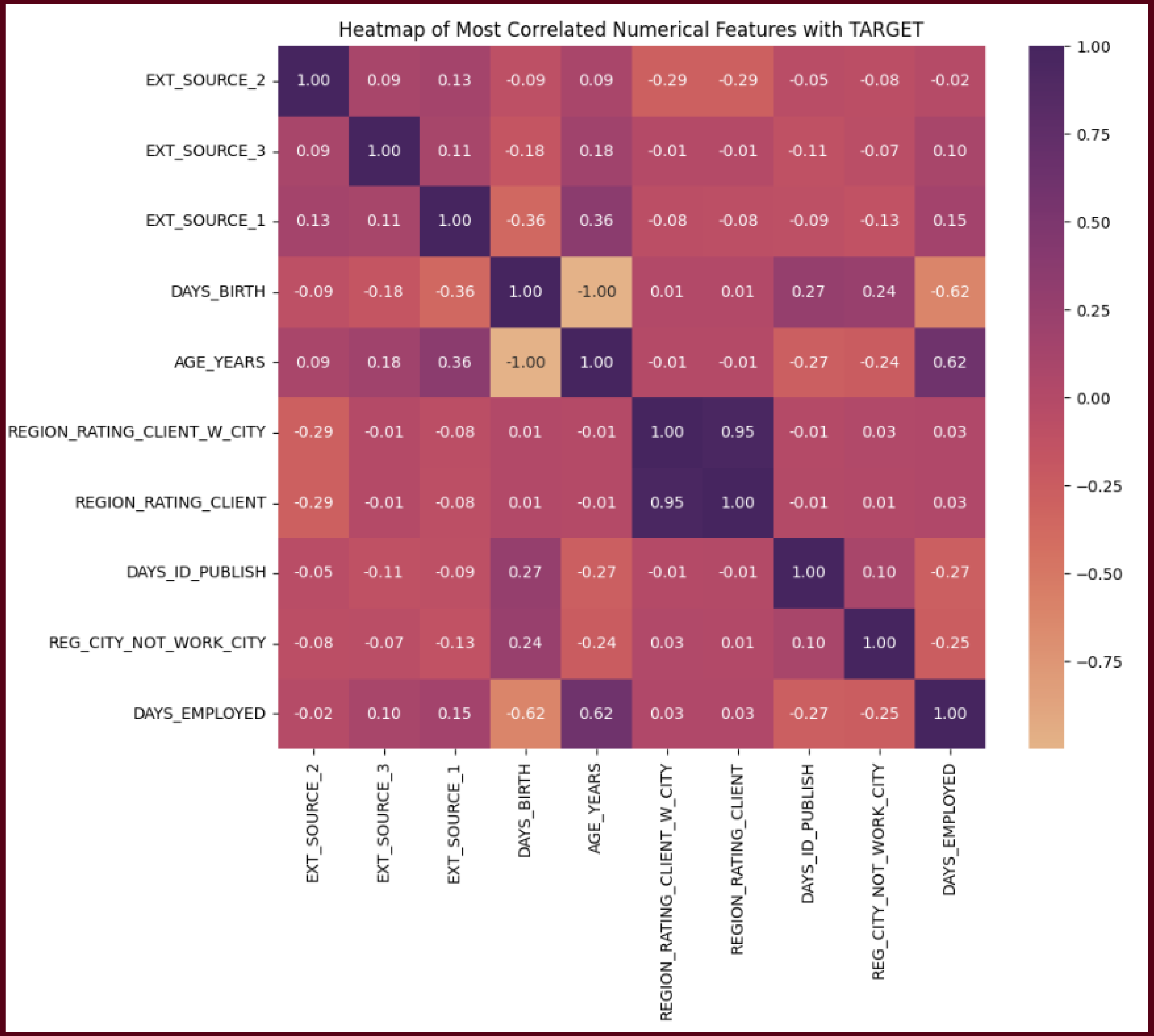
- NAME_FAMILY_STATUS: Married individuals show a higher count of non-defaulters, indicating that marital stability might contribute to better financial management and loan repayment.

- OCCUPATION_TYPE: Laborers and low-skill laborers show a higher count of Non-defaulters although they show defaults.



CORRELATION ANALYSIS

- The heatmap displays the most correlated numerical features with the TARGET with correlation coefficients, which range from -1 to 1.
- The values 0 to 1 indicate Positive correlation and values -1 to 0 indicate negative correlation.
- Darker colours indicate stronger correlations; lighter colours suggest weaker ones.
- The heatmap shows that higher external credit scores (EXT_SOURCE_1, 2, 3) and older age (DAYS_BIRTH, AGE_YEARS) are associated with a lower likelihood of loan default, while shorter employment duration (DAYS_EMPLOYED) is slightly linked to higher default risk. Other variables show weak correlations with default.



PAIR PLOT ANALYSIS

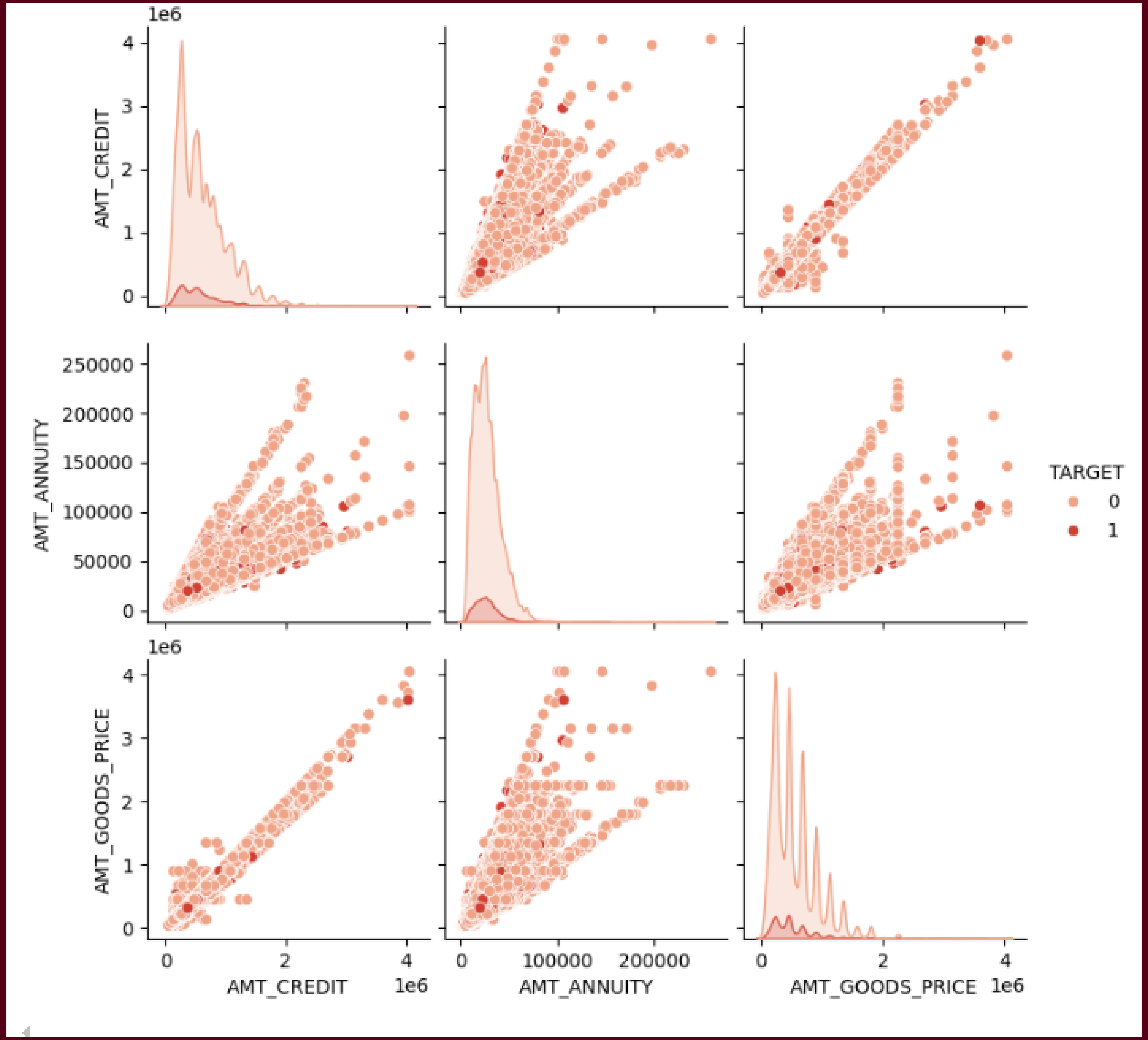
- The pair plot shows the relationships between AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE across defaulters (TARGET = 1) and non-defaulters (TARGET = 0).

- The diagonal plots indicate that these features have a right-skewed distribution.

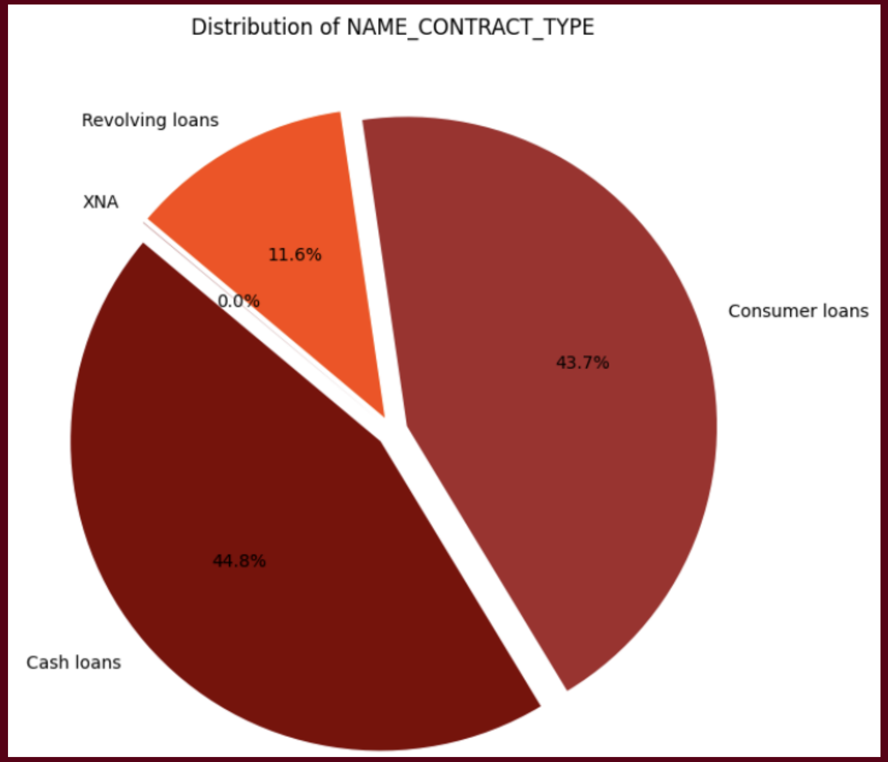
- The off-diagonal scatter plots show a positive correlation between these variables, suggesting that as one increases, the others tend to increase as well.

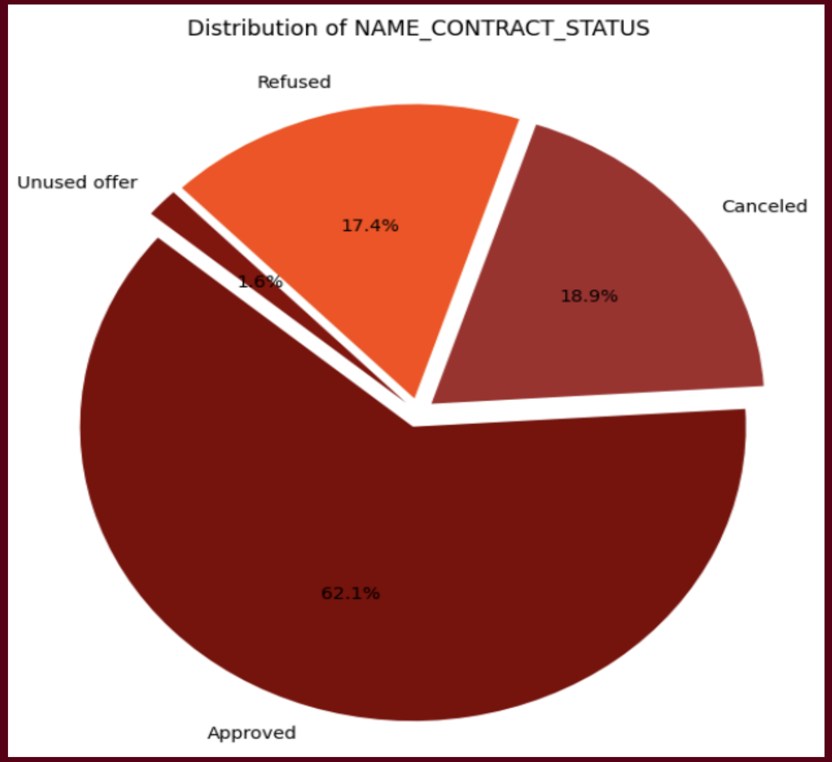
- There is a strong positive correlation between AMT_CREDIT and AMT_GOODS_PRICE, as well as between AMT_CREDIT and AMT_ANNUITY, indicating that higher credit amounts are associated with higher goods prices and annuities.

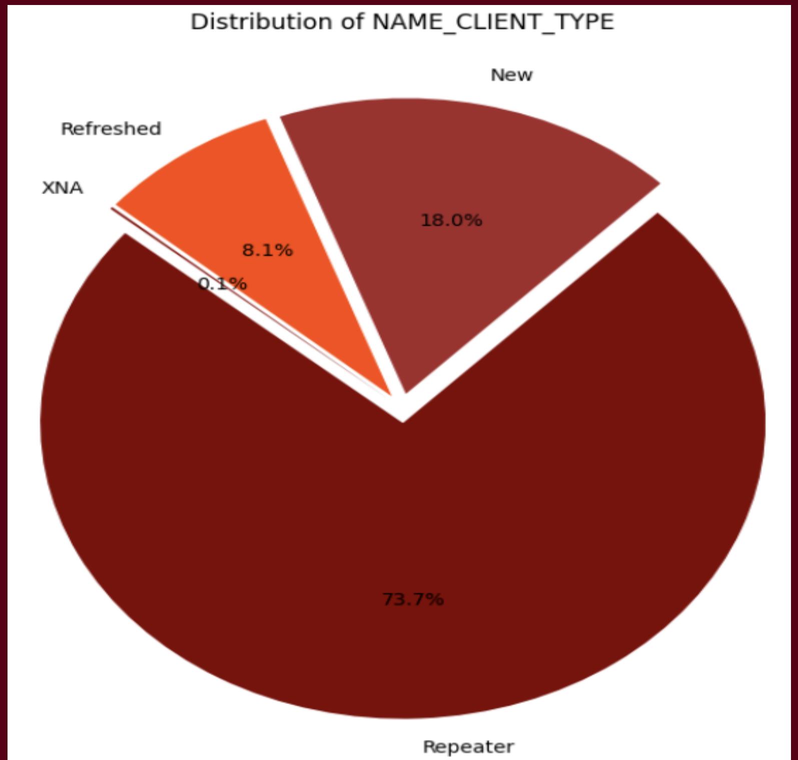
- The distribution shows that defaulters (TARGET=1) are spread across similar ranges but may be slightly more concentrated in certain value ranges, suggesting that loan amount and goods price could be related to default risk.



EXPLORATORY DATA ANALYSIS (previous_application)



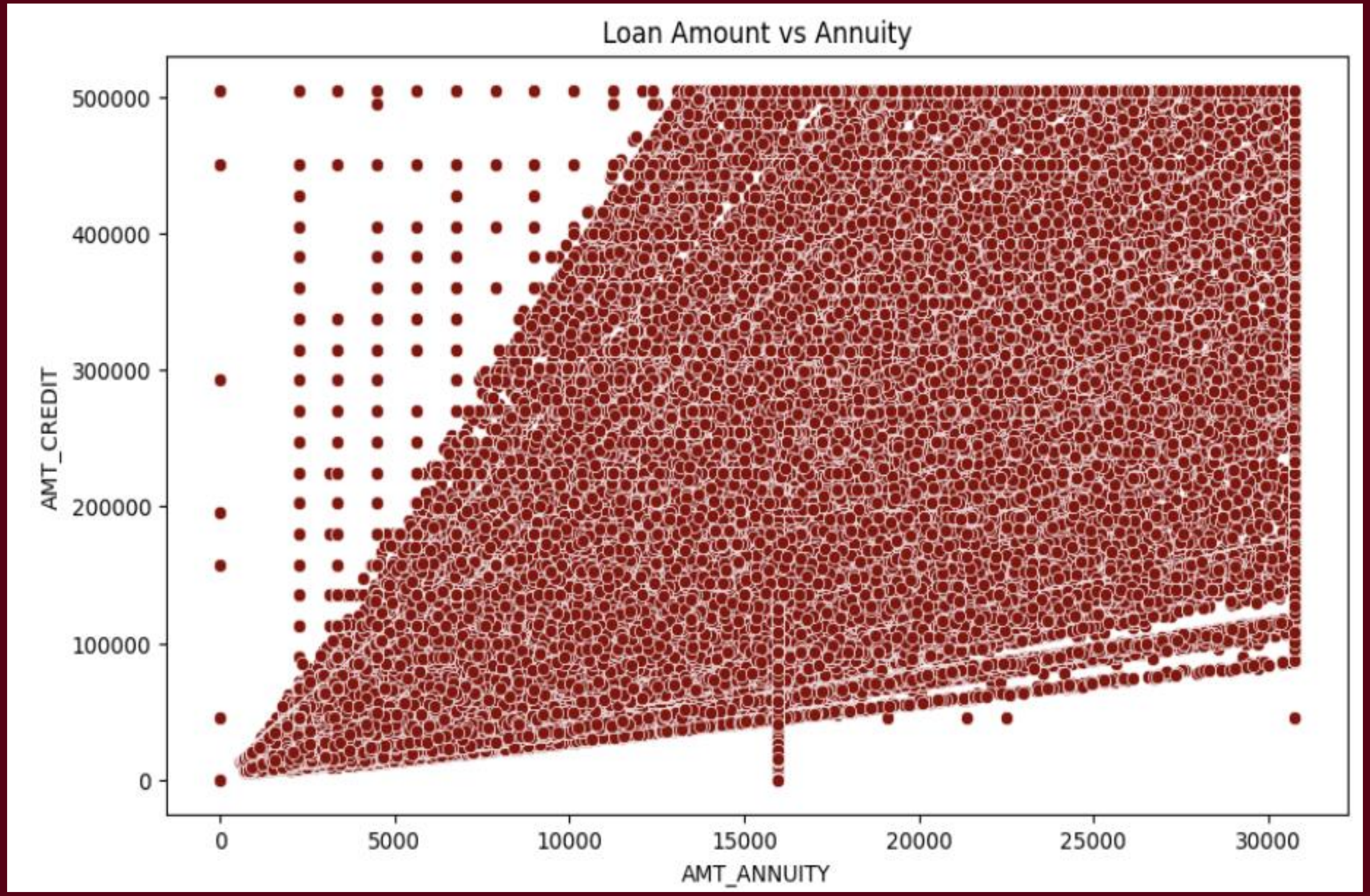


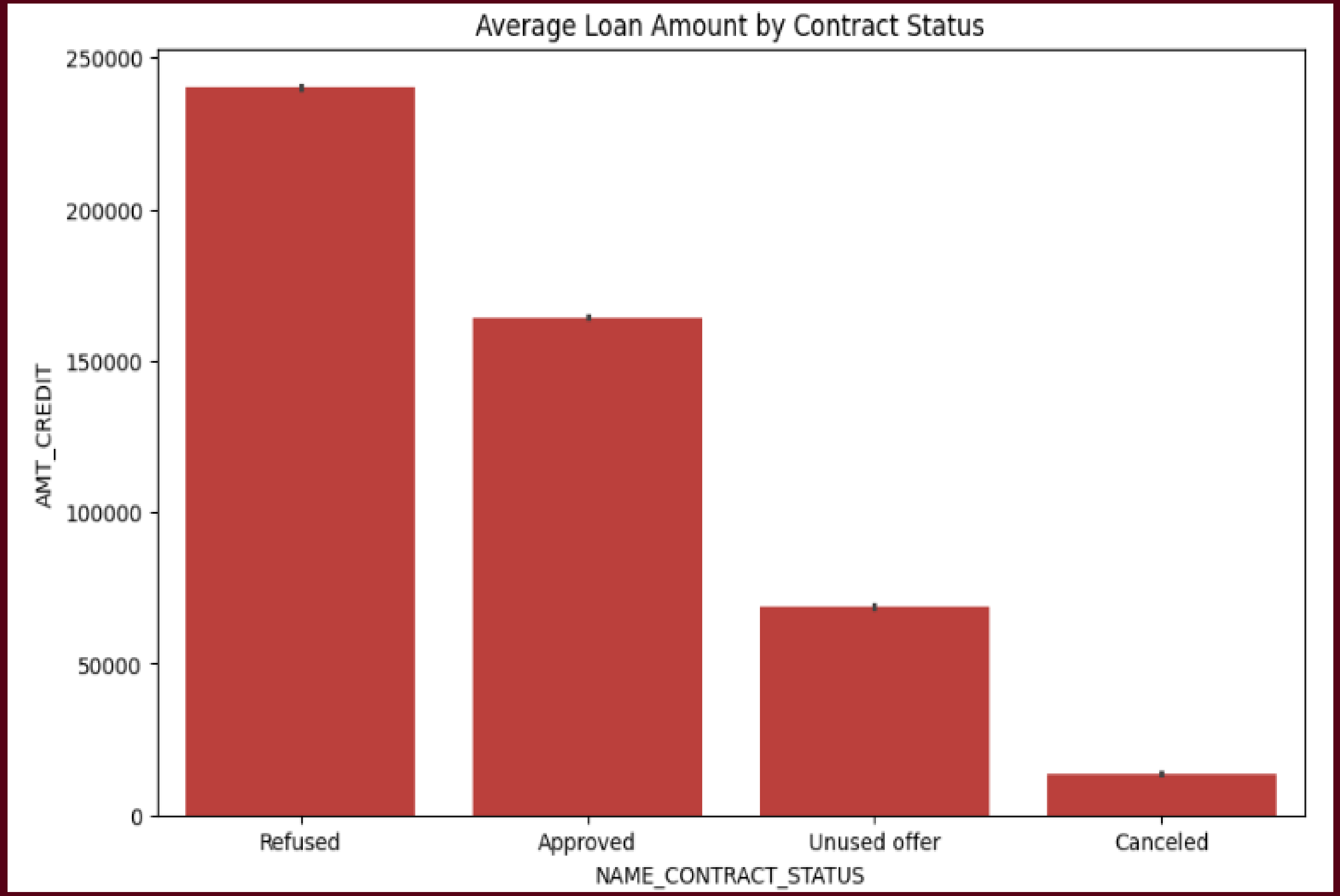


DISTRIBUTION OF CATEGORICAL FEATURES:

- The pie charts indicate that cash loans and consumer loans make up the majority of contract types, with cash loans (44.8%) being slightly more prevalent.
- Most contract statuses (62.1%) are approved, followed by cancelled and refused offers.
- Among client types, repeat clients (73.7%) dominate, with a significant portion of new clients as well, while refreshed clients and unknown categories are minimal.
- This suggests a concentration on certain loan types and client profiles, which could inform risk assessments and targeting strategies.

EXPLORATORY DATA ANALYSIS (previous_application)

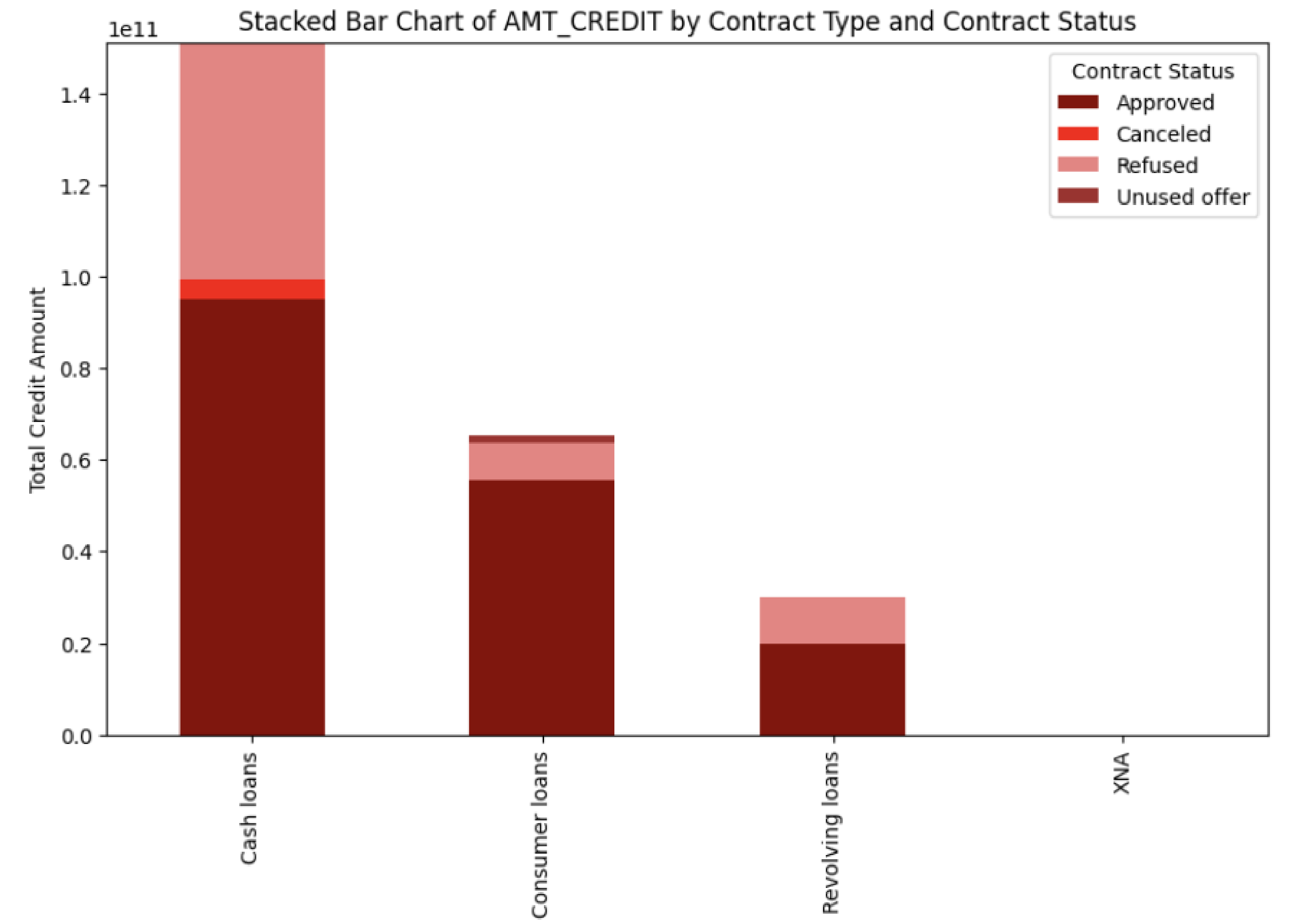


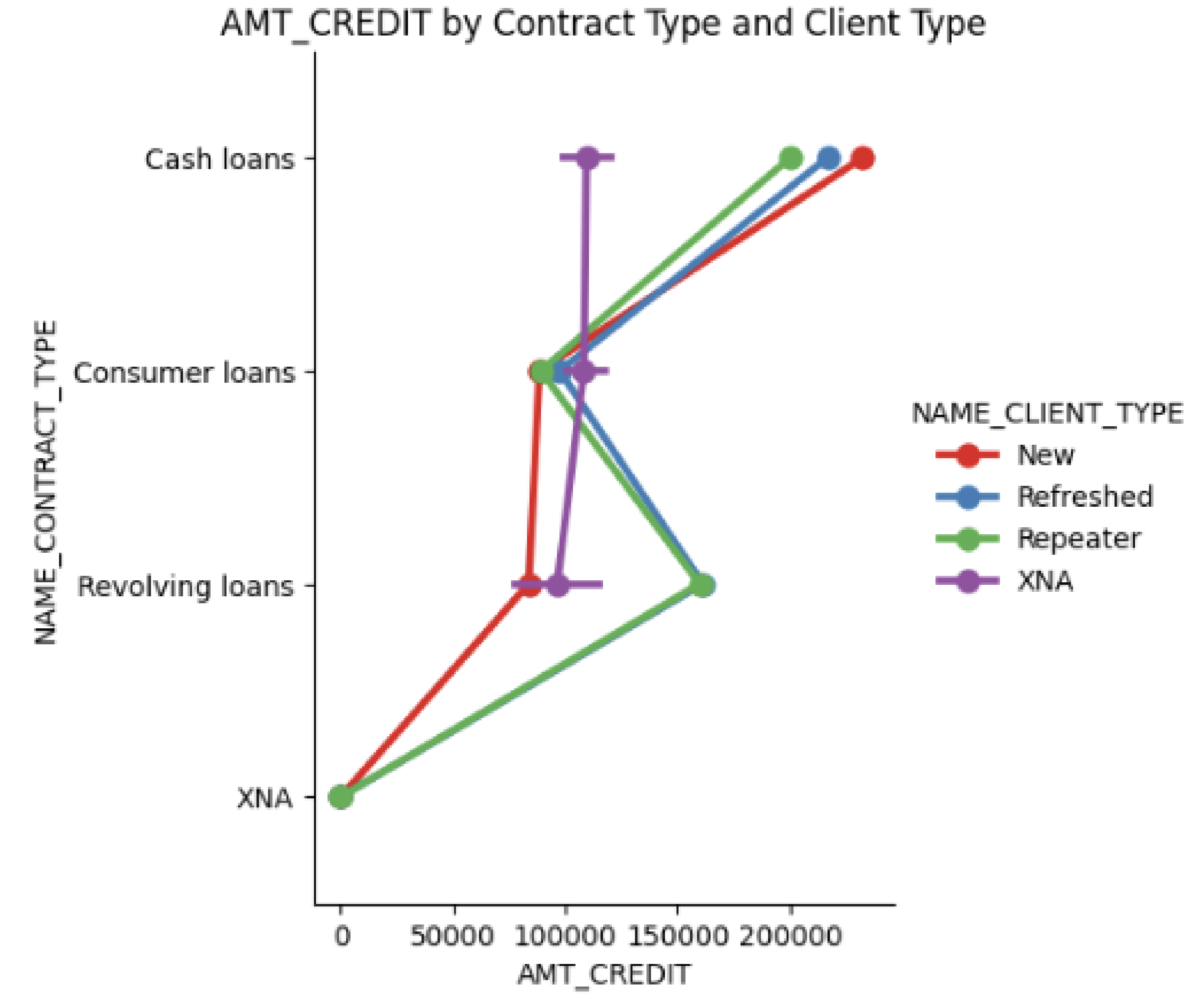


- The scatter plot shows a positive correlation between loan amount (**AMT_CREDIT**) and annuity amount (**AMT_ANNUITY**), meaning as the loan amount increases, the annuity tends to increase as well. The distribution is dense and suggests structured loan and annuity packages, but some scattered points indicate exceptions or outliers.

- The bar plot shows that Refused loans have the highest average amounts, indicating larger loans face more evaluation and are often declined. Approved loans are moderate in size, while smaller loans are more likely to be cancelled or unused.

EXPLORATORY DATA ANALYSIS (previous_application)

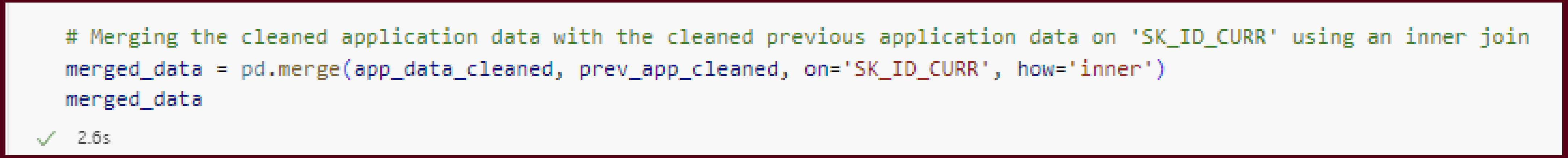




- The Stacked Bar chart shows that Cash loans have the highest credit amount, mostly approved, while Consumer and Revolving loans have smaller amounts with more cancellations and refusals.

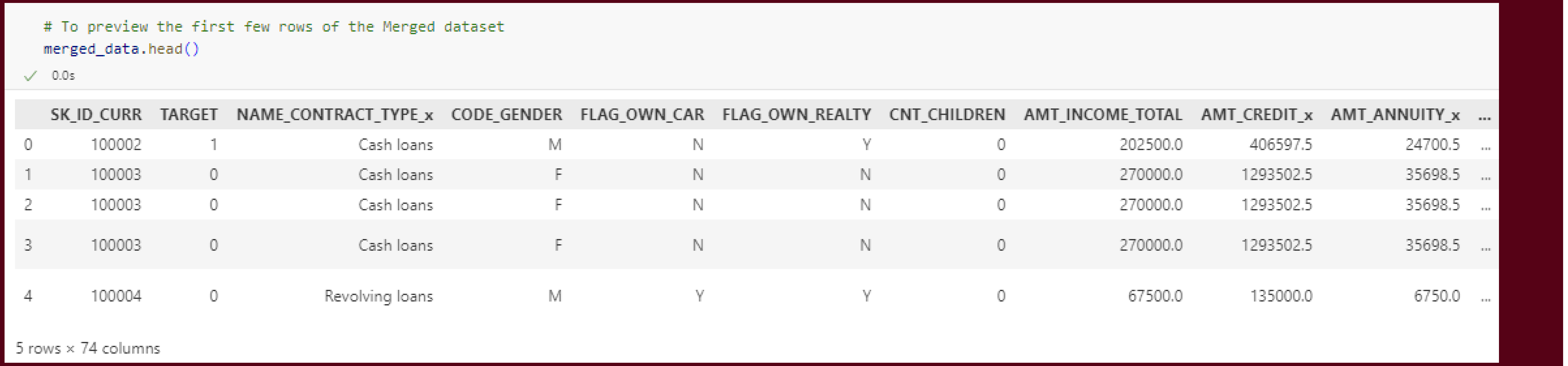
- This Line plot shows that Cash loans generally have the highest credit amounts across all client types, while Revolving loans have lower amounts. Repeat clients tend to have higher loan amounts across all contract types, indicating they might be trusted more by lenders.

MERGED DATASET



```python
# Merging the cleaned application data with the cleaned previous application data on 'SK_ID_CURR' using an inner join
merged_data = pd.merge(app_data_cleaned, prev_app_cleaned, on='SK_ID_CURR', how='inner')
merged_data
```

✓ 2.6s

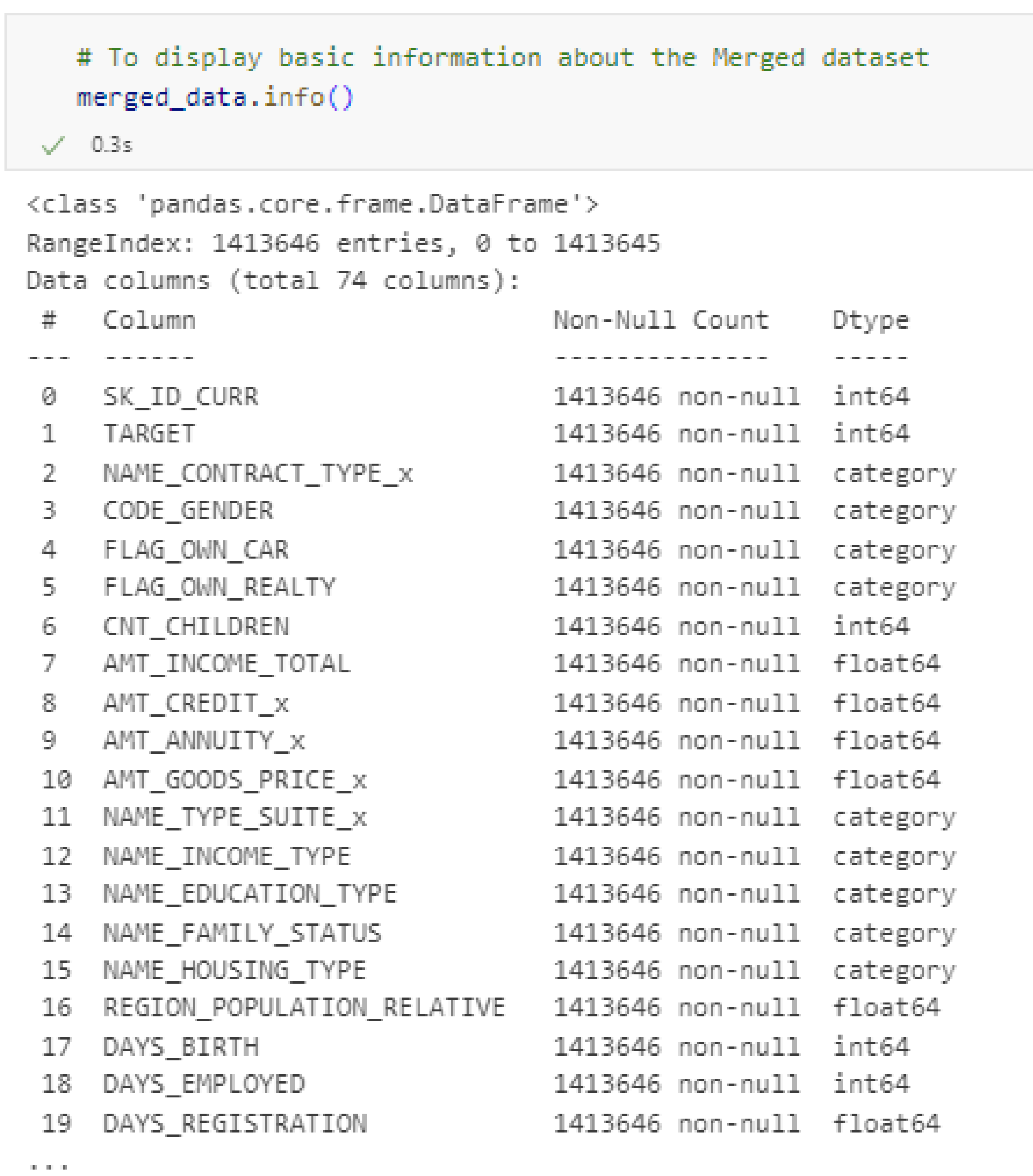


```python
# To preview the first few rows of the Merged dataset
merged_data.head()
```

✓ 0.0s

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT_x	AMT_ANNUITY_x	...
0	100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	...
1	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	...
2	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	...
3	100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	...
4	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	...

5 rows × 74 columns



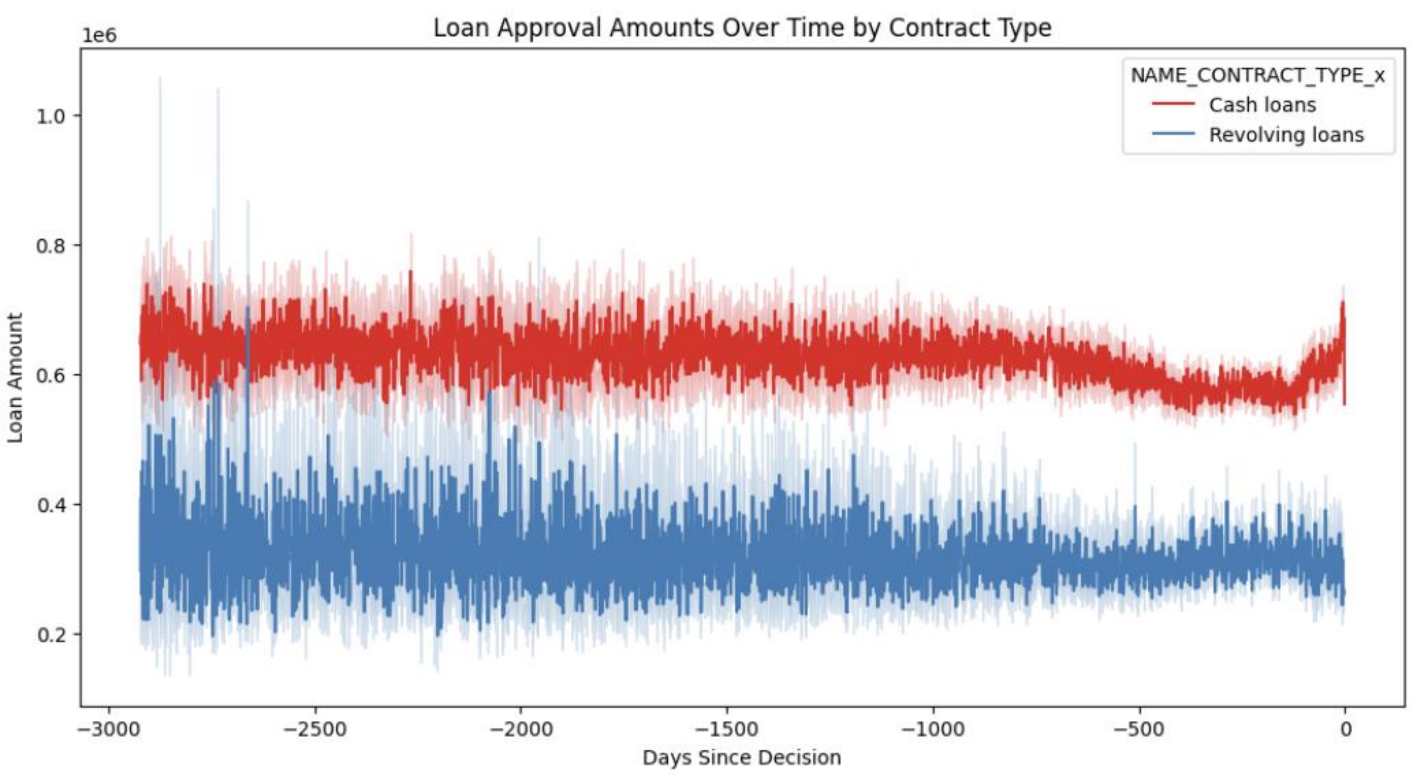
```python
# To display basic information about the Merged dataset
merged_data.info()
```

✓ 0.3s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1413646 entries, 0 to 1413645
Data columns (total 74 columns):
 #   Column                      Non-Null Count    Dtype
---  ------                      --------------    -----
 0   SK_ID_CURR                  1413646 non-null  int64
 1   TARGET                      1413646 non-null  int64
 2   NAME_CONTRACT_TYPE_x        1413646 non-null  category
 3   CODE_GENDER                 1413646 non-null  category
 4   FLAG_OWN_CAR                1413646 non-null  category
 5   FLAG_OWN_REALTY             1413646 non-null  category
 6   CNT_CHILDREN                1413646 non-null  int64
 7   AMT_INCOME_TOTAL            1413646 non-null  float64
 8   AMT_CREDIT_x                1413646 non-null  float64
 9   AMT_ANNUITY_x               1413646 non-null  float64
 10  AMT_GOODS_PRICE_x           1413646 non-null  float64
 11  NAME_TYPE_SUITE_x           1413646 non-null  category
 12  NAME_INCOME_TYPE            1413646 non-null  category
 13  NAME_EDUCATION_TYPE         1413646 non-null  category
 14  NAME_FAMILY_STATUS          1413646 non-null  category
 15  NAME_HOUSING_TYPE           1413646 non-null  category
 16  REGION_POPULATION_RELATIVE  1413646 non-null  float64
 17  DAYS_BIRTH                  1413646 non-null  int64
 18  DAYS_EMPLOYED               1413646 non-null  int64
 19  DAYS_REGISTRATION           1413646 non-null  float64
...
```

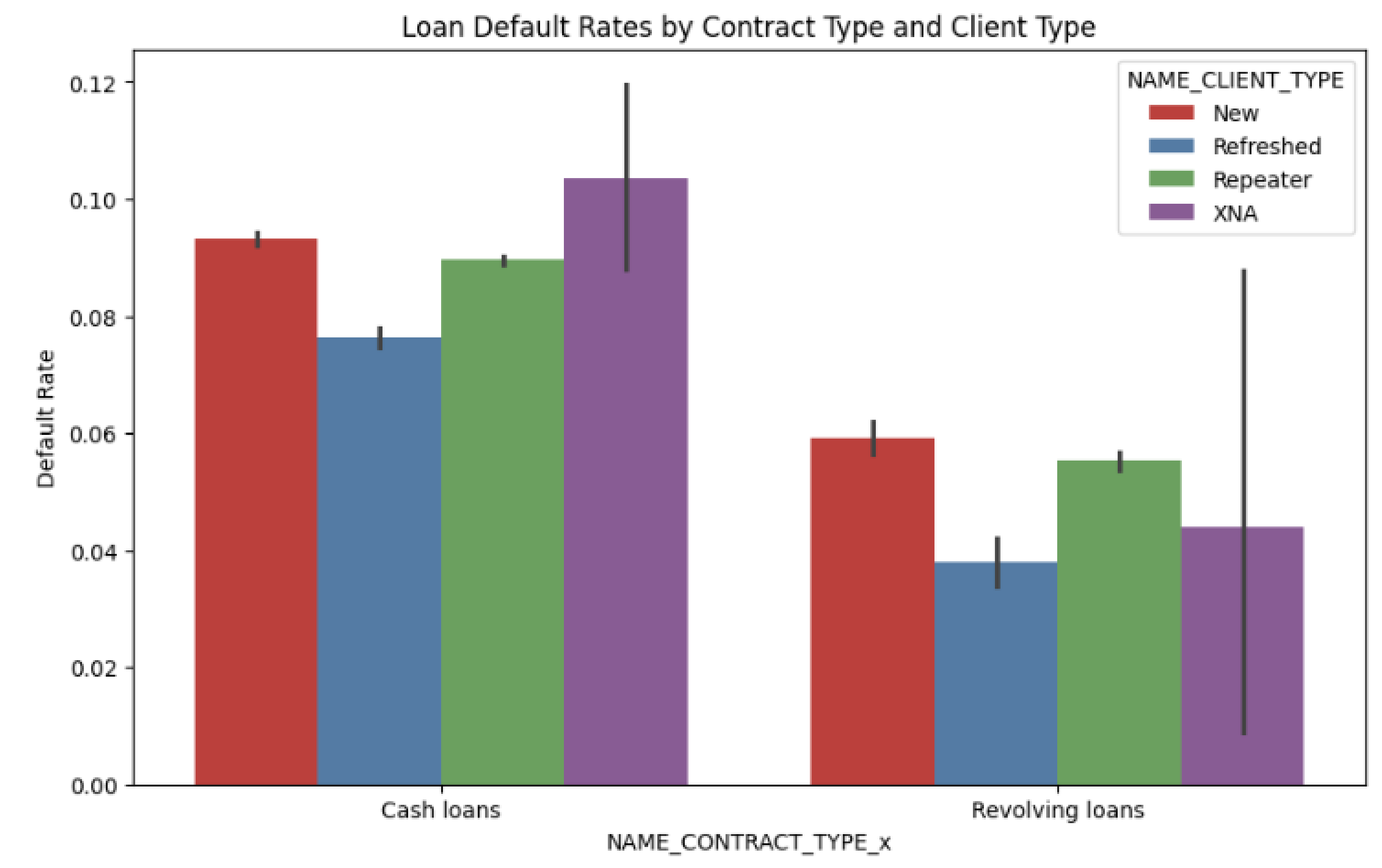
- The merged_data was created by performing an inner join on SK_ID_CURR between app_data_cleaned and prev_app_cleaned to consolidate relevant information from both datasets for comprehensive analysis.

EXPLORATORY DATA ANALYSIS (MERGED DATA)

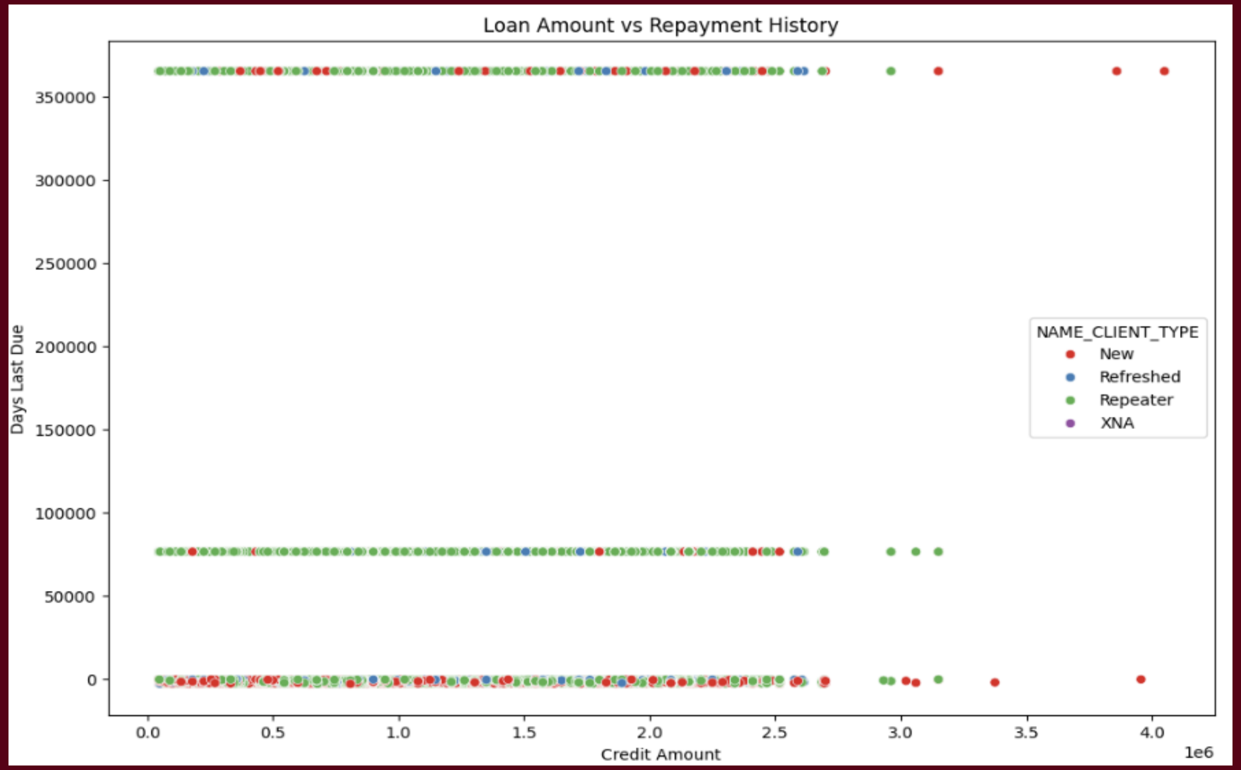


APPLICATION APPROVAL TRENDS:
The line plot shows that loan approval amounts for both cash loans and revolving loans have generally remained stable over time, with cash loans consistently having higher amounts than revolving loans. However, there is a noticeable increase in loan amounts as the decision date approaches.

COMPARISON OF LOAN DEFAULT RATES BY CONTRACT & CLIENT HISTORY:
The bar plot shows that Cash loans have similar default rates across client types, with "XNA" showing the highest. For revolving loans, "Repeaters" tend to default more, while "New" and "Refreshed" clients have lower rates. Client history and loan type both influence default risk.

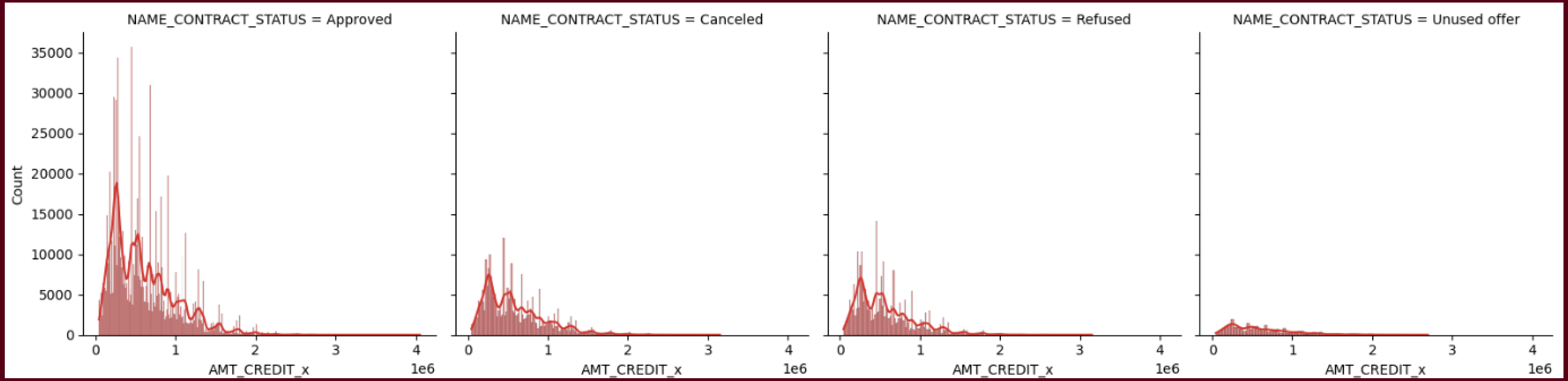


EXPLORATORY DATA ANALYSIS (MERGED DATA)



LOAN AMOUNT VS REPAYMENT HISTORY:
The scatter plot shows that loan amounts are consistent across different client types, with repayment days segmented into distinct groups (0, ~100,000, ~350,000). There is no clear relationship between the amount of credit and the days past due, suggesting that loan size does not strongly influence repayment behaviour across different client types.

DISTRIBUTION OF CREDIT AMOUNTS BY CONTRACT STATUS:
This FacetGrid plot shows the distribution of loan amounts (AMT_CREDIT_x) across different contract statuses. This suggests that higher loan amounts are more likely to be approved, while smaller loans are often refused or left unused.



CONCLUSION & RECOMMENDATIONS

- The data shows a significant imbalance with 91.9% non-defaulters and 8.1% defaulters, indicating that defaults are less common but critical to identify for financial stability.
- Younger age groups (20-39 years) show a higher proportion of defaults, indicating that age may influence repayment behaviour. However, gender and income type are less definitive in predicting defaults.
- Variables like age, external credit scores, and contract types are significant predictors of loan default.
- Cash loans dominate, with varying default rates depending on client history.
- There are strong correlations between loan amounts and goods prices, as well as between external credit scores and defaults.
- A client's previous loan history, particularly the type and status of those loans, plays a significant role in determining their risk of default on new loans.
- Consider prioritizing loan approval for clients with longer employment histories, as this correlates with lower default risk and can be beneficial in loan repayment.
- Focus on high-risk clients, particularly younger individuals and those with low credit scores, by enhancing scoring models with additional variables like occupation and contract types.
- Diversify the loan portfolio by focusing on products with lower default rates.
- Enhanced Risk Assessment Models can be developed to improve the prediction of defaults.

Thank You

✉ swethakjoseph16@gmail.com

🌐 https://www.linkedin.com/in/swetha-kizhavana-joseph-04b68721b/