



# EXPLORATORY DATA ANALYSIS OF LOAN APPLICATIONS

Name: Ritik Sanjay Patel

Batch: May 2024


Email: [ritikpatel976@gmail.com](mailto:ritikpatel976@gmail.com)



# INTRODUCTION

The primary goal of this project is to analyze loan application data to identify patterns and insights that can help predict loan defaults. By understanding the factors contributing to defaults, financial institutions can improve their risk management strategies and lending policies.

## Datasets Used:

1. Application Data: Contains information about the clients at the time of their loan application, including demographic details, financial status, and loan characteristics.
  2. Previous Application Data: Includes historical loan data for clients, such as previous loan amounts, statuses, and purposes.
- 



# APPROACH

Here is the step by step approach used for this project:

## 1. Data Collection:

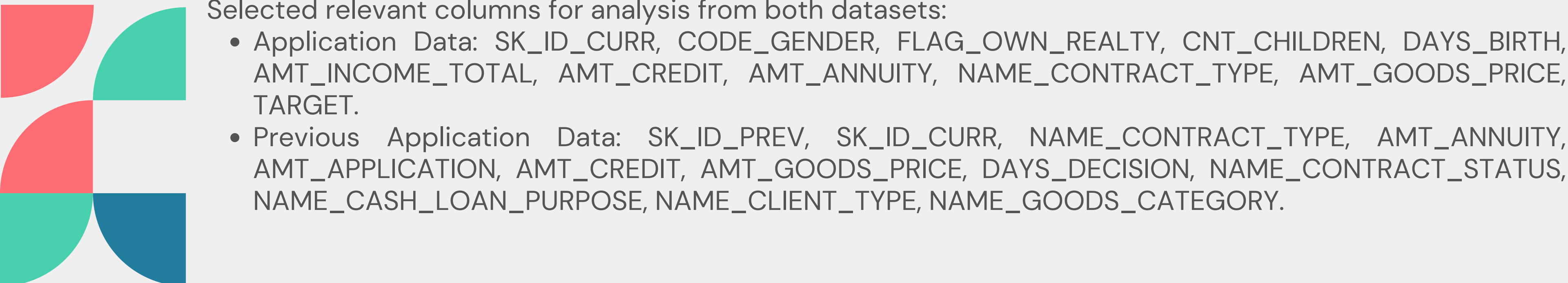
- Gathered data from two primary sources: application\_data.csv and previous\_application.csv.

## 2. Data Cleaning:

- Identified and handled missing values.
- Converted negative days column into positive value
- Detected and addressed outliers using Z-score method.

## 3. Feature Selection:

Selected relevant columns for analysis from both datasets:

- Application Data: SK\_ID\_CURR, CODE\_GENDER, FLAG\_OWN\_REALTY, CNT\_CHILDREN, DAYS\_BIRTH, AMT\_INCOME\_TOTAL, AMT\_CREDIT, AMT\_ANNUITY, NAME\_CONTRACT\_TYPE, AMT\_GOODS\_PRICE, TARGET.
  - Previous Application Data: SK\_ID\_PREV, SK\_ID\_CURR, NAME\_CONTRACT\_TYPE, AMT\_ANNUITY, AMT\_APPLICATION, AMT\_CREDIT, AMT\_GOODS\_PRICE, DAYS\_DECISION, NAME\_CONTRACT\_STATUS, NAME\_CASH\_LOAN\_PURPOSE, NAME\_CLIENT\_TYPE, NAME\_GOODS\_CATEGORY.
- 



# APPROACH

## 4. Merging Dataset:

- Merged `application_data` with `previous_application` on `SK_ID_CURR` to create a comprehensive dataset for analysis.

## 5. Exploratory Data Analysis (EDA):

- Conducted univariate, bivariate, and multivariate analyses to uncover patterns and relationships.
- Visualized data using histograms, scatter plots, box plots, count plots, and heatmaps.



## 6. Generating Insights:

- Analyzed the impact of various features on the target variable (`TARGET`).
- Derived actionable insights to predict loan defaults and improve loan approval processes.



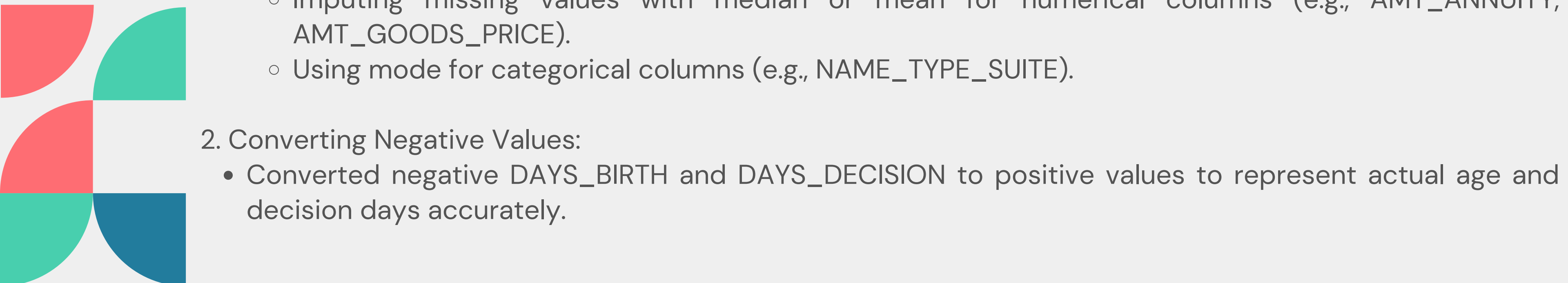
# DATA CLEANING

Data cleaning steps:

## 1. Handling Missing values:

- Analyzed columns with significant missing values.
- Decided on appropriate strategies such as:
  - Dropping columns with extremely high missing values (e.g., RATE\_INTEREST\_PRIVILEGED, RATE\_INTEREST\_PRIMARY).
  - Imputing missing values with median or mean for numerical columns (e.g., AMT\_ANNUITY, AMT\_GOODS\_PRICE).
  - Using mode for categorical columns (e.g., NAME\_TYPE\_SUITE).

## 2. Converting Negative Values:

- Converted negative DAYS\_BIRTH and DAYS\_DECISION to positive values to represent actual age and decision days accurately.
- 



# DATA CLEANING


## 3. Handling Outliers:

- Used Z-score method to detect outliers in key numerical columns (AMT\_ANNUITY, AMT\_APPLICATION, AMT\_CREDIT, AMT\_GOODS\_PRICE, DAYS\_DECISION).
- Addressed outliers by capping extreme values to a threshold (95th percentile) to minimize their impact on analysis.

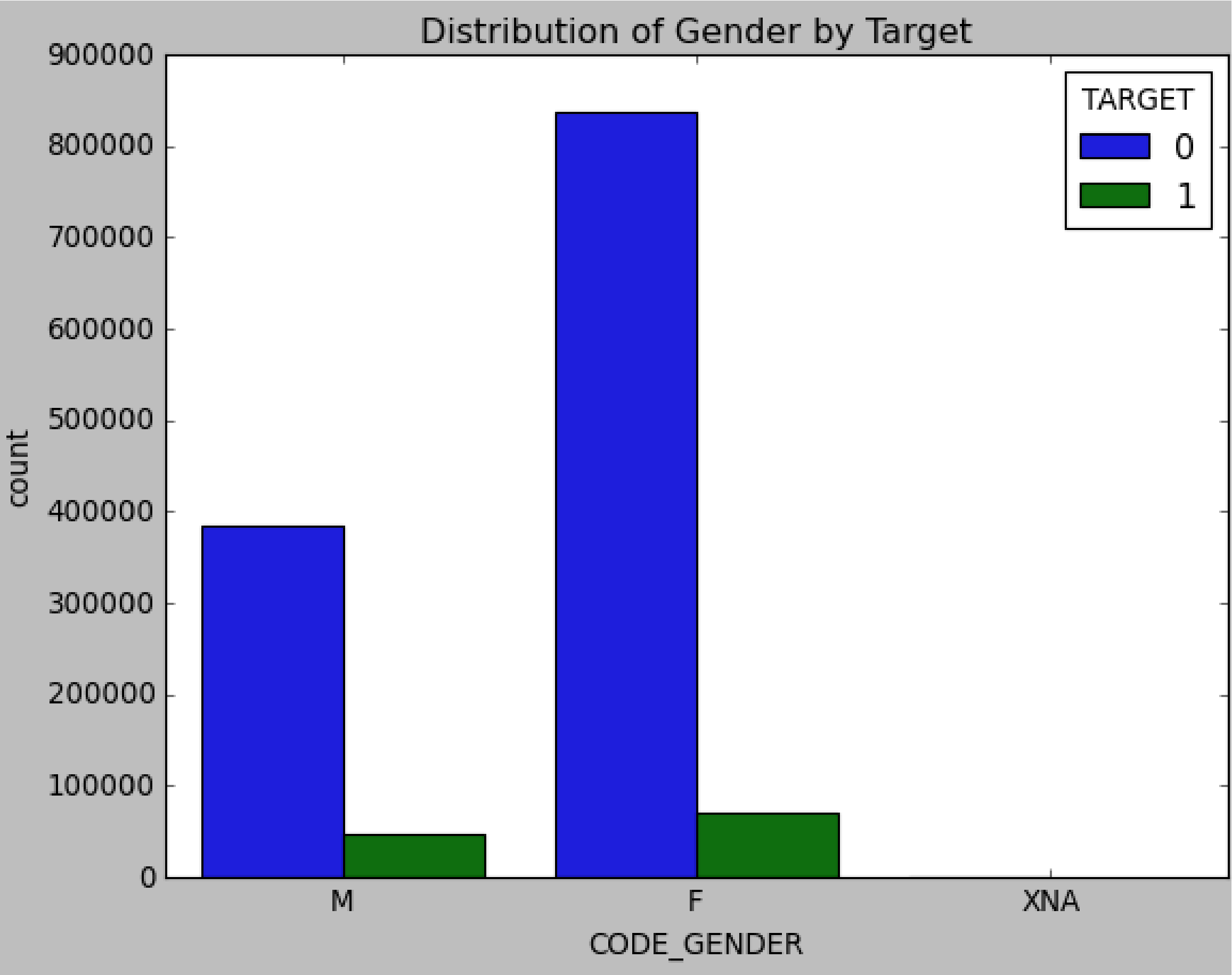
## 4. Feature Engineering:

- Created new columns for better analysis:
  - AGE\_YEARS by converting DAYS\_BIRTH to age in years.
  - Binned AGE\_YEARS into age groups for more granular insights in visualizations.

## 5. Data Merging:

- Merged application\_data with previous\_application on SK\_ID\_CURR to integrate current and previous loan details.
  - Ensured consistency and integrity of merged data for comprehensive analysis.
- 

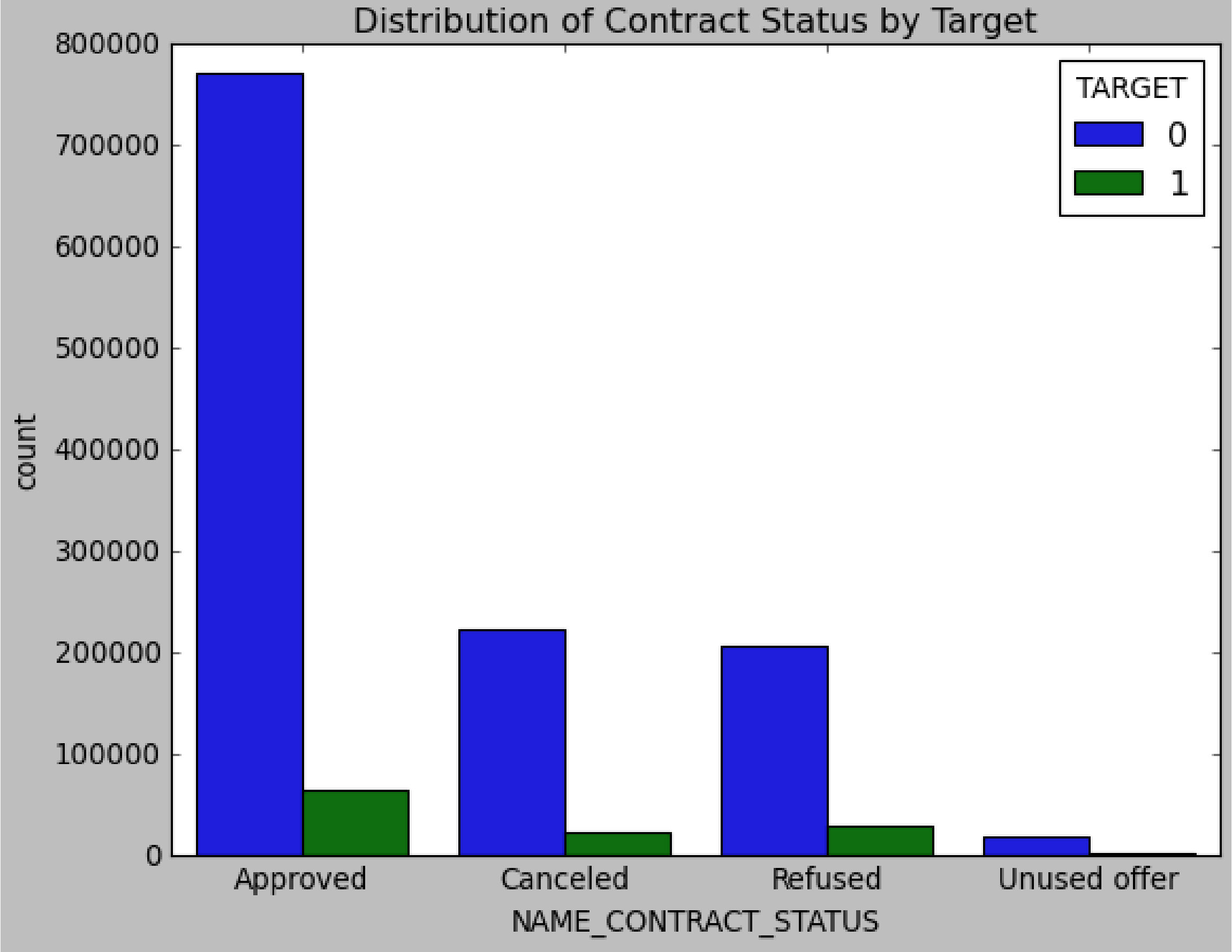
# EDA INSIGHTS



## Insight 1: Gender Distribution

Both male and female applicants have defaults, with a higher number of female applicants overall. The proportion of defaults to non-defaults appears similar for both genders.

# EDA INSIGHTS

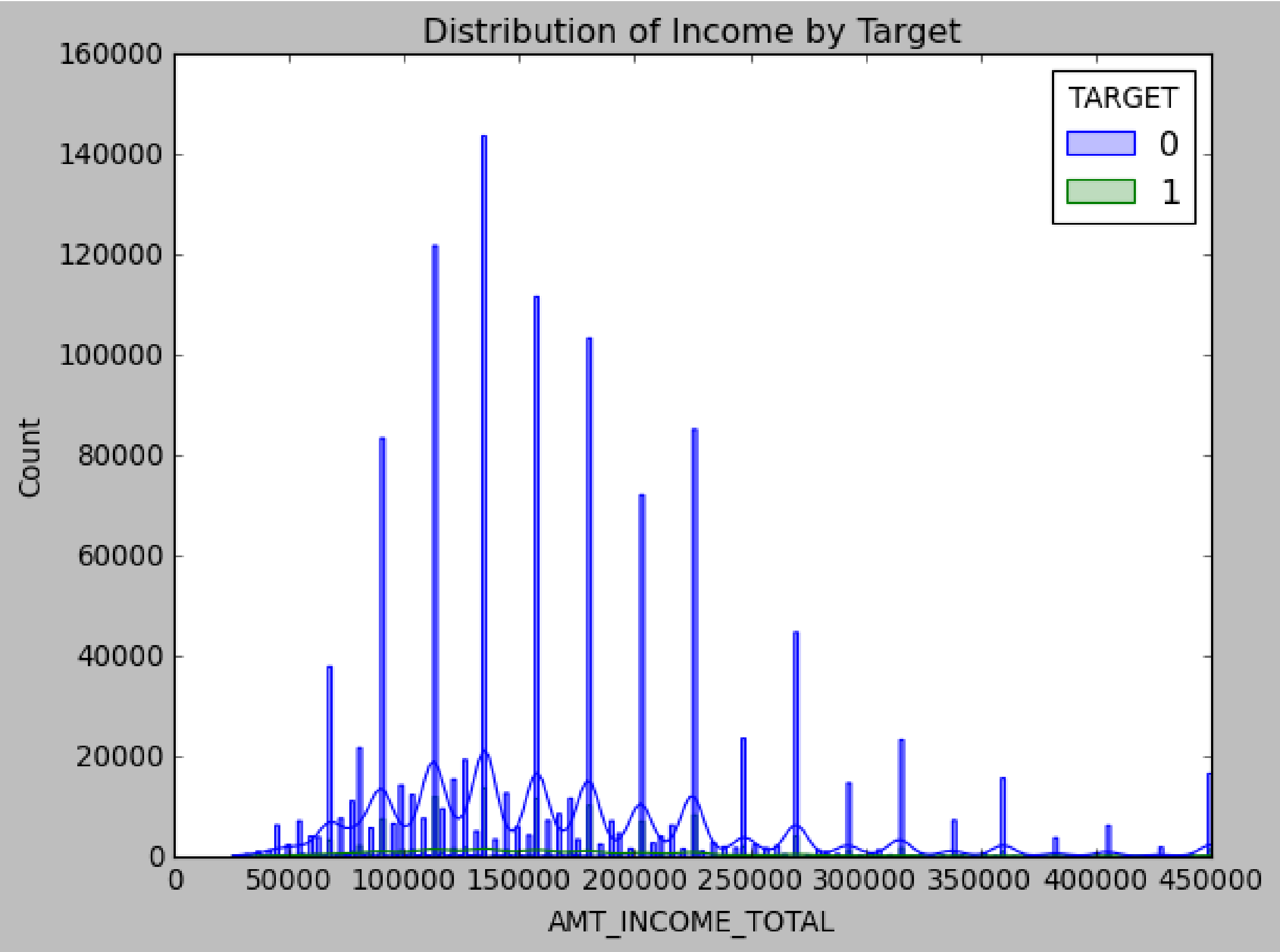


## Insight 2: Contract Status Distribution by Target

- Majority of loans approved, with notable defaults among approved loans
- Refusal process seems effective, with fewer defaults among refused loans



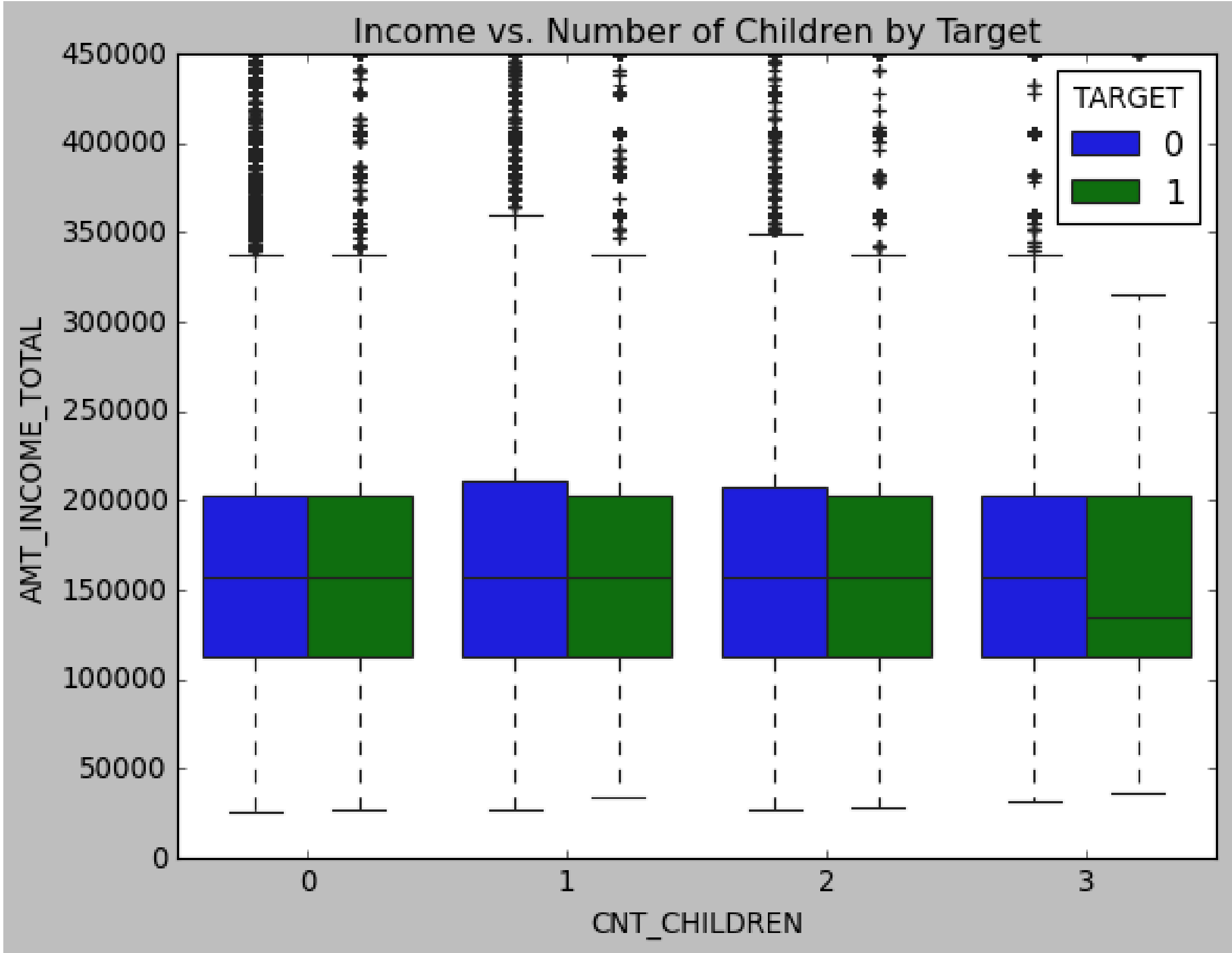
# EDA INSIGHTS



## Insight 3: Income Distribution by Target

- Defaults concentrated in lower income range up to around 200,000
- Higher income levels show fewer defaults, indicating correlation between higher income and lower default risk

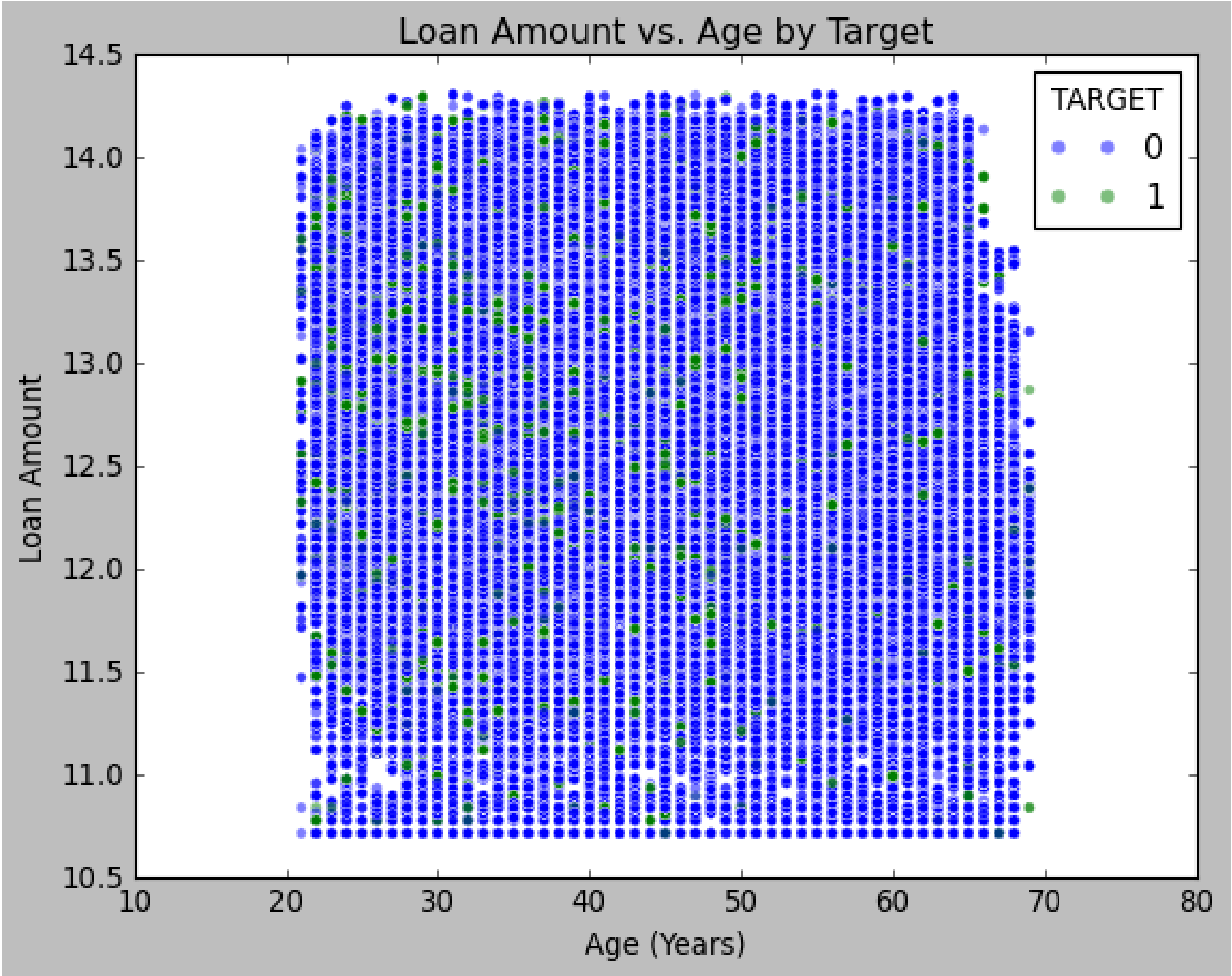
# EDA INSIGHTS



## Insight 4: Income vs. Number of Children by Target

- Median income for non-default clients slightly higher compared to default clients across all categories of CNT\_CHILDREN
- Number of children does not significantly affect income distribution

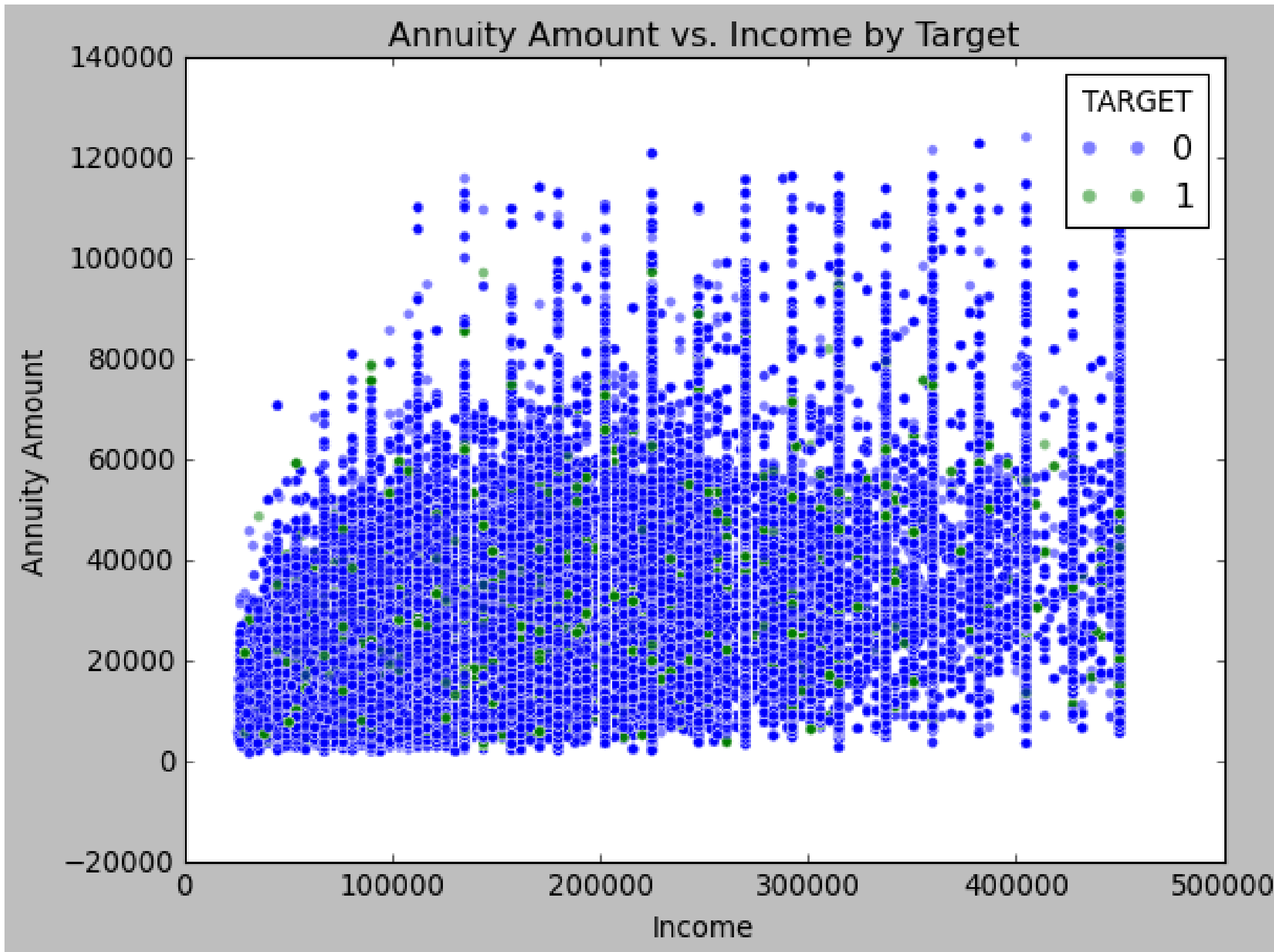
# EDA INSIGHTS



## Insight 5: Loan Amount vs. Age by Target

- Applicants' ages range from approximately 20 to 70 years
- Both default and non-default clients distributed throughout age range and loan amount spectrum

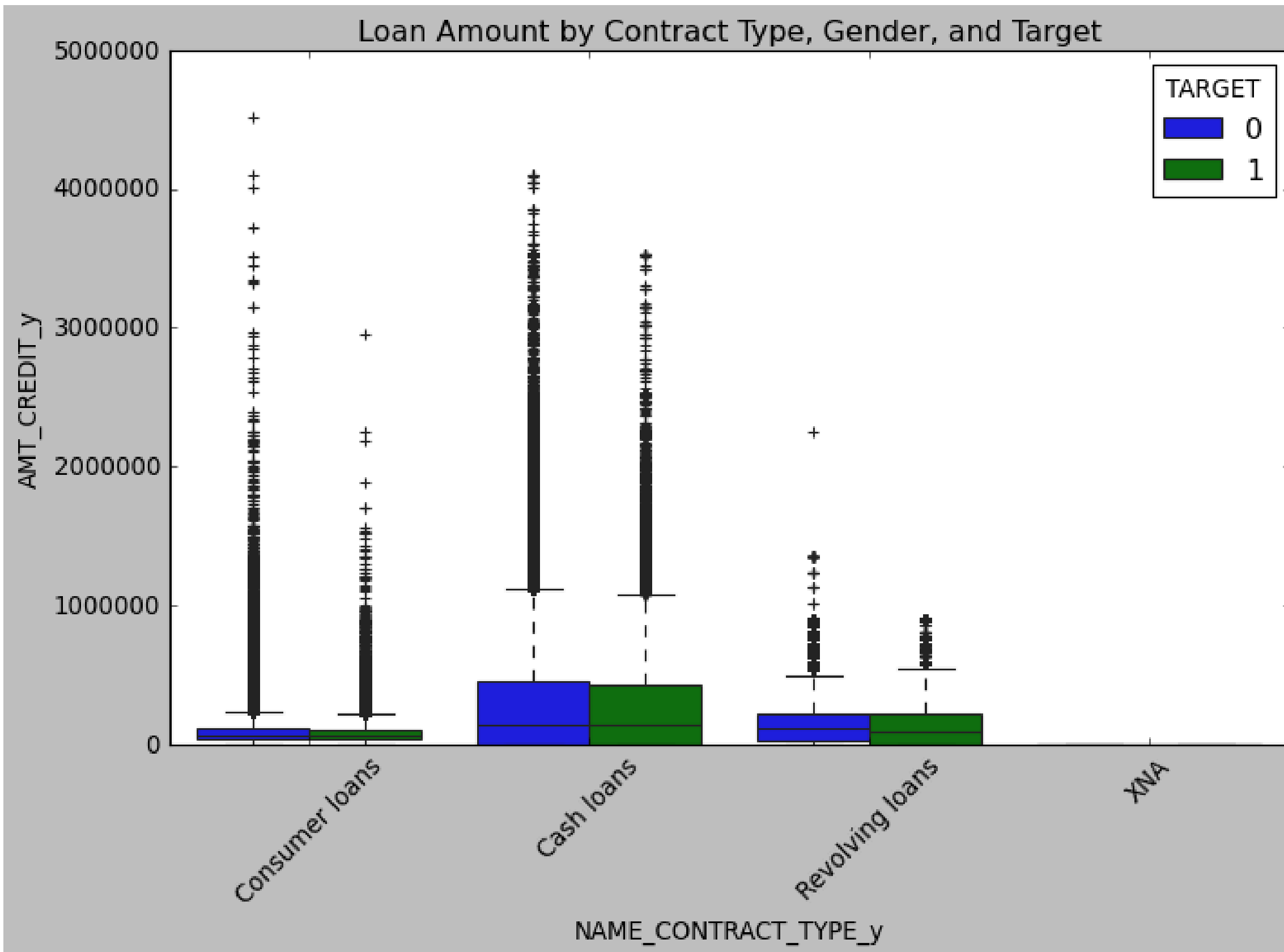
# EDA INSIGHTS



## Insight 6: Annuity Amount vs. Income by Target

- Positive correlation between income and annuity amount
- No distinct separation between defaulters and non-defaulters based on income and annuity amounts alone

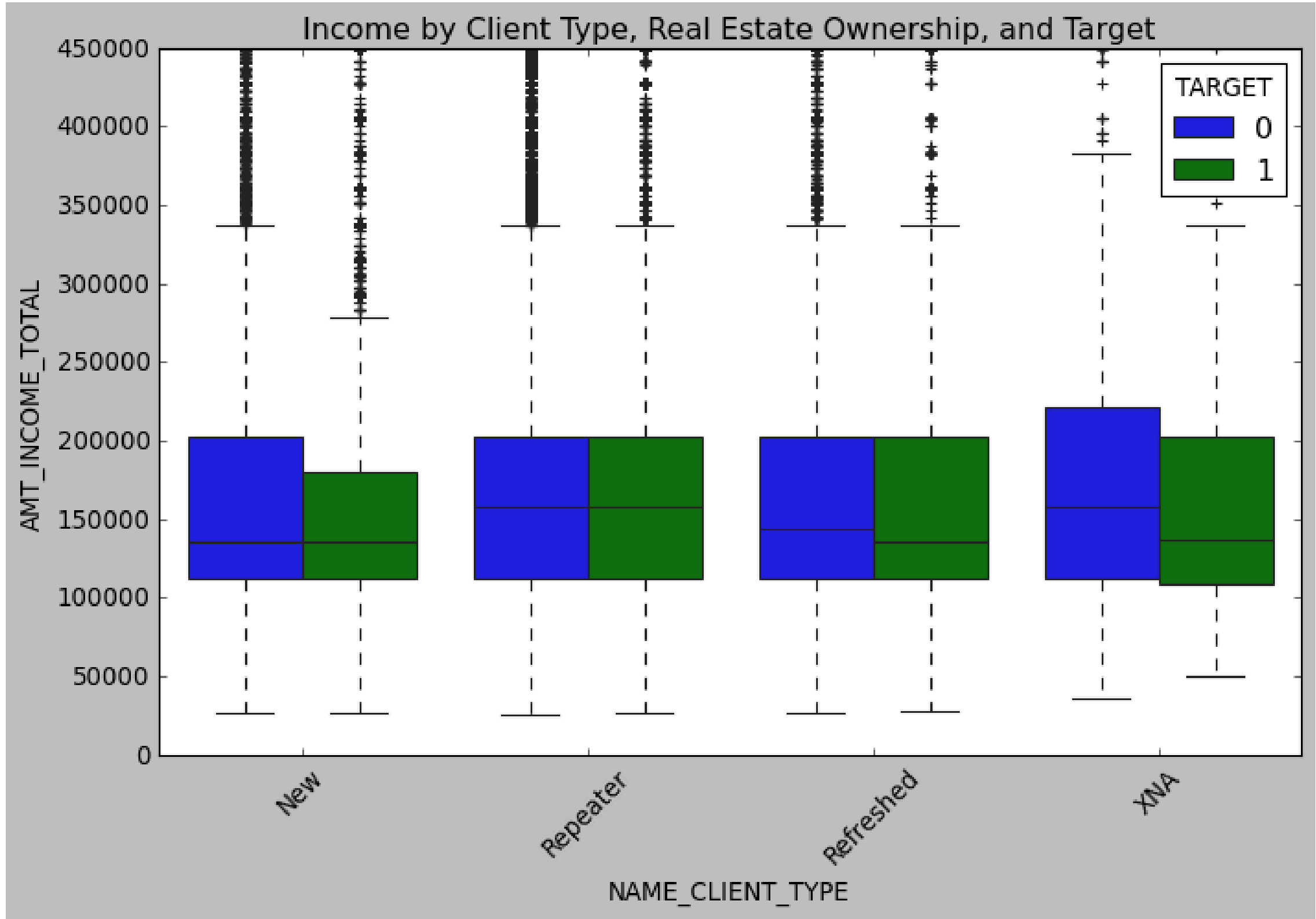
# EDA INSIGHTS



## Insight 7: Loan Amount by Contract Type and Target

- For consumer loans, loan amounts appear similar for both defaulters and non-defaulters
- For cash loans, non-defaulters generally have higher loan amounts than defaulters
- Revolving loans show non-defaulters have slightly higher loan amounts compared to defaulters

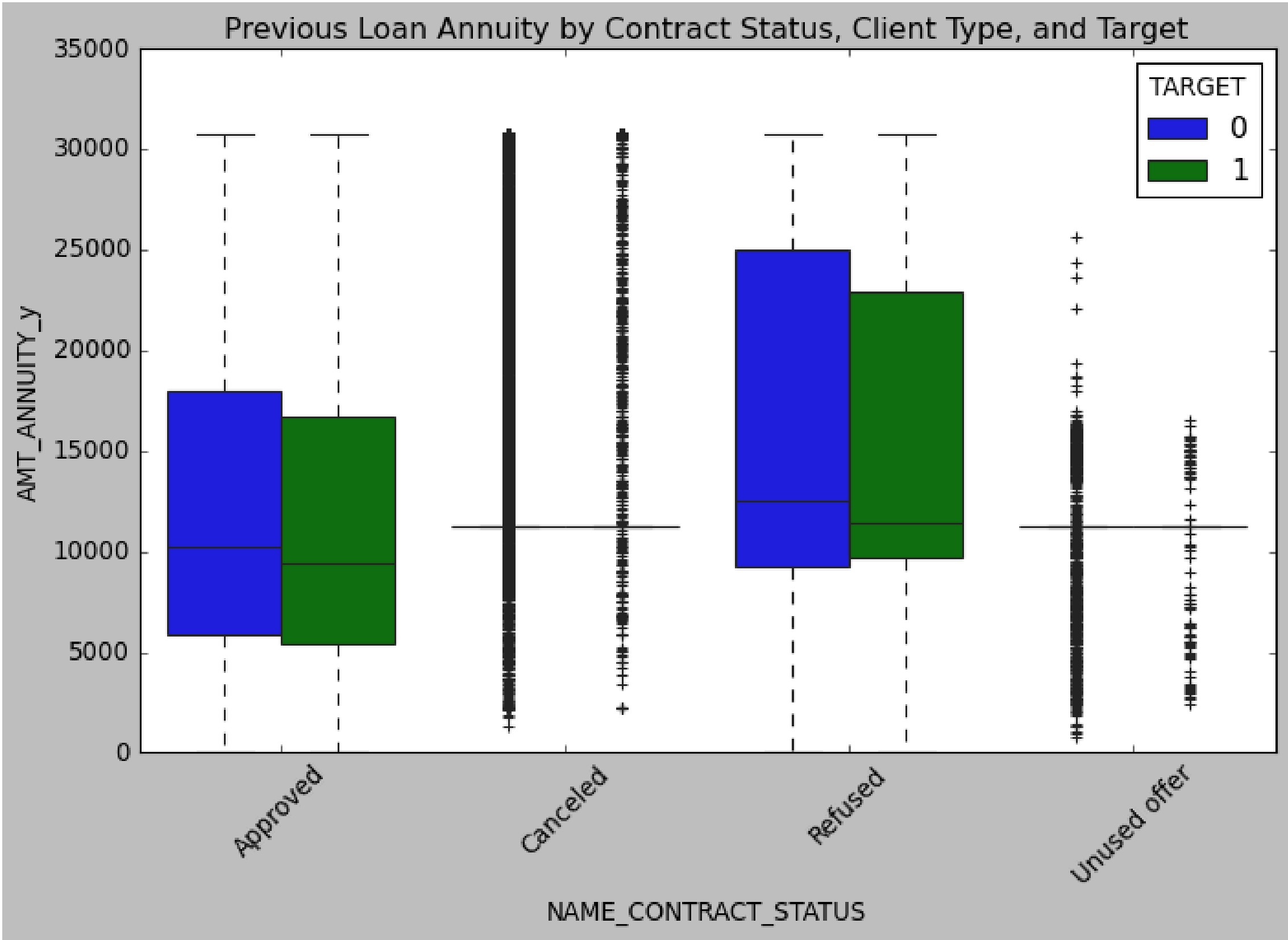
# EDA INSIGHTS



## Insight 8: Income by Client Type and Target

- Median incomes for defaulters and non-defaulters are relatively similar across all client types
- Income alone may not be a strong predictor of default status within these client types

# EDA INSIGHTS

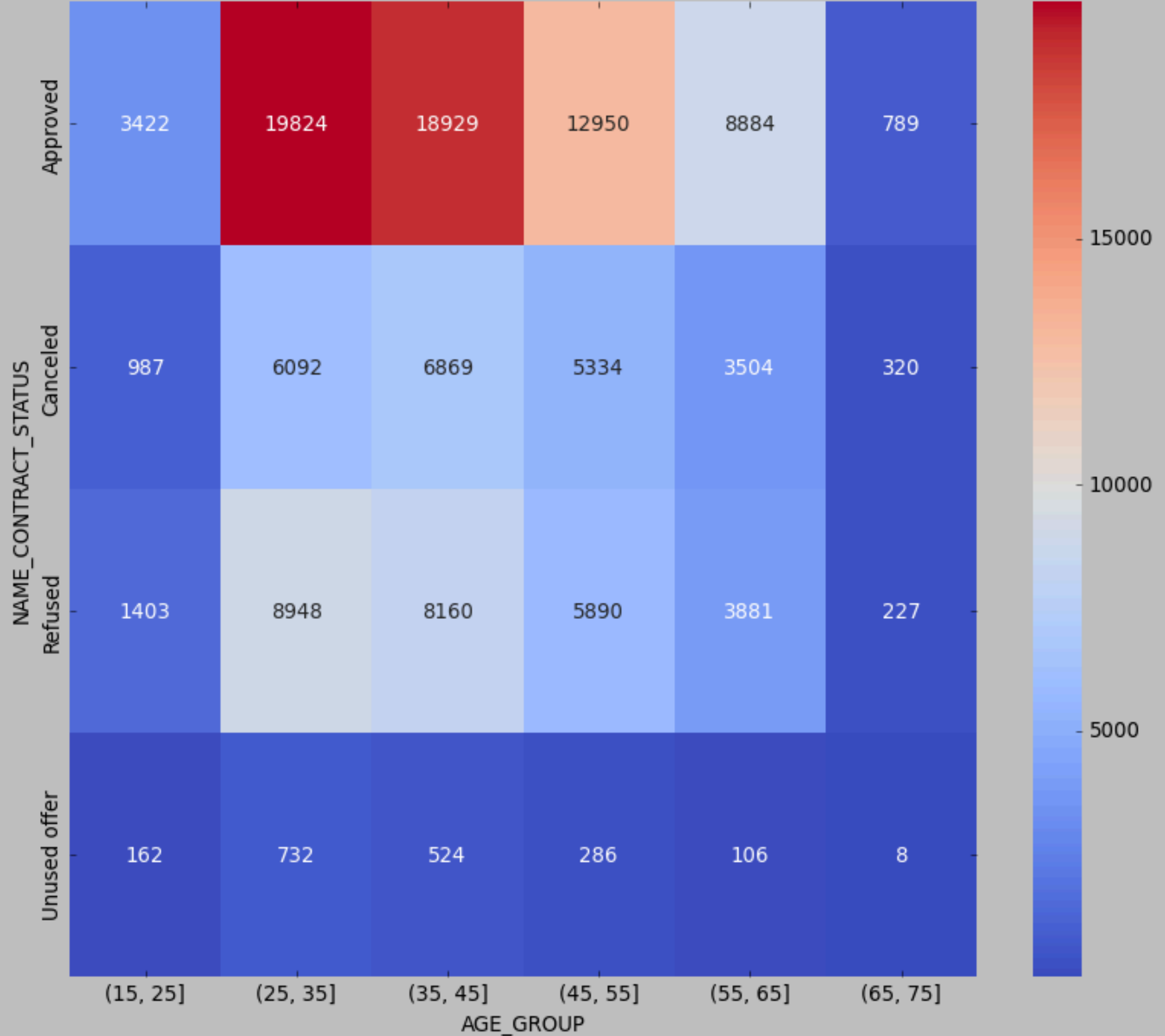


## Insight 9: Previous Loan Annuity by Contract Status and Target

- Approved loans: Annuity amounts for non-defaulters and defaulters are relatively similar
- Cancelled loans: Wide spread in annuity amounts, similar distributions for non-defaulters and defaulters
- Refused loans: Defaulters have slightly lower median annuity amounts compared to non-defaulters
- Unused offers: Both defaulters and non-defaulters show similar distributions

# EDA INSIGHTS

Heatmap of Target by Contract Status and Age Group

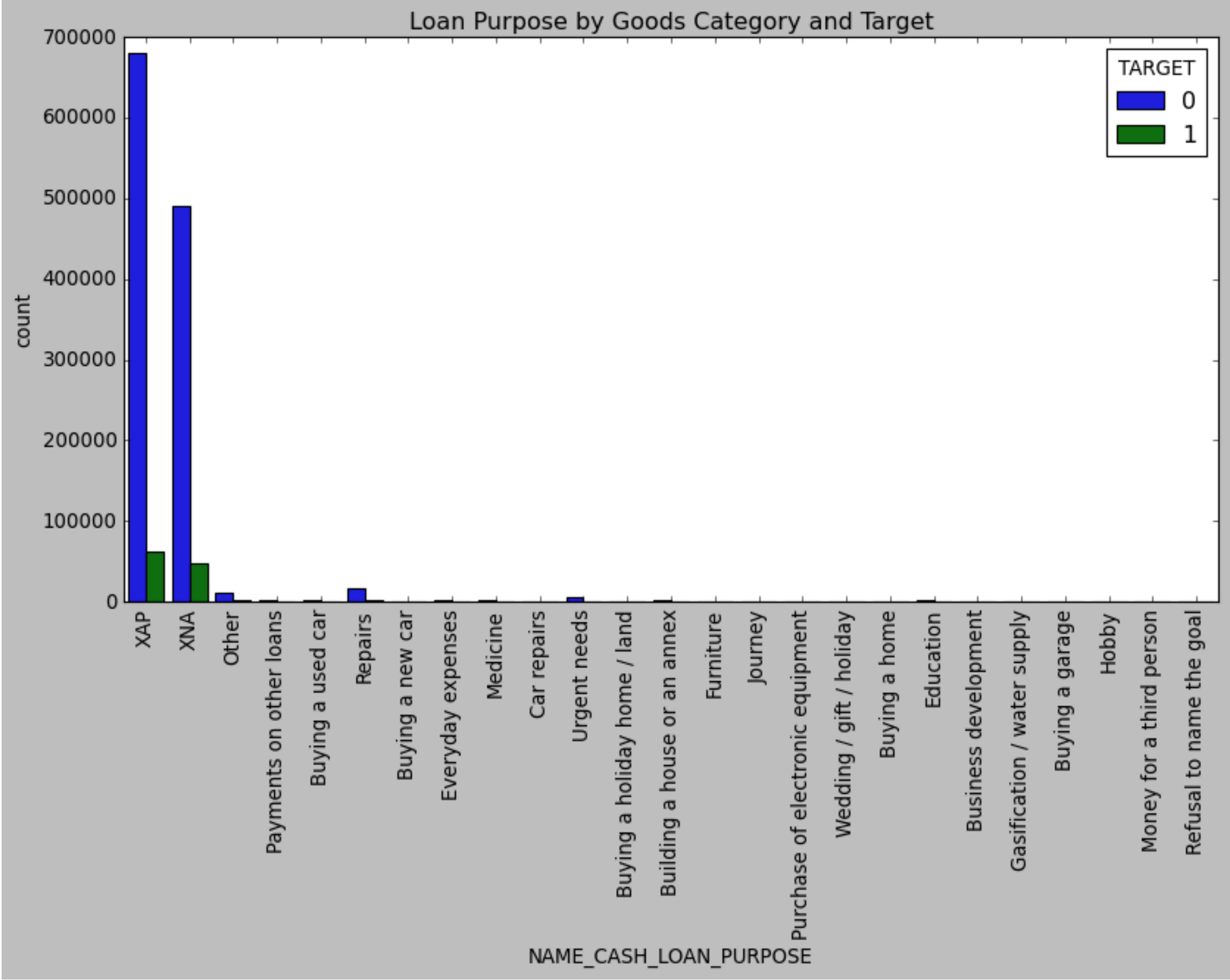


## Insight 10: Target by Contract Status and Age Group

- Higher default rates for approved loans in ages 25–35 and 35–45
- Refused or cancelled loans have defaulted in the current application



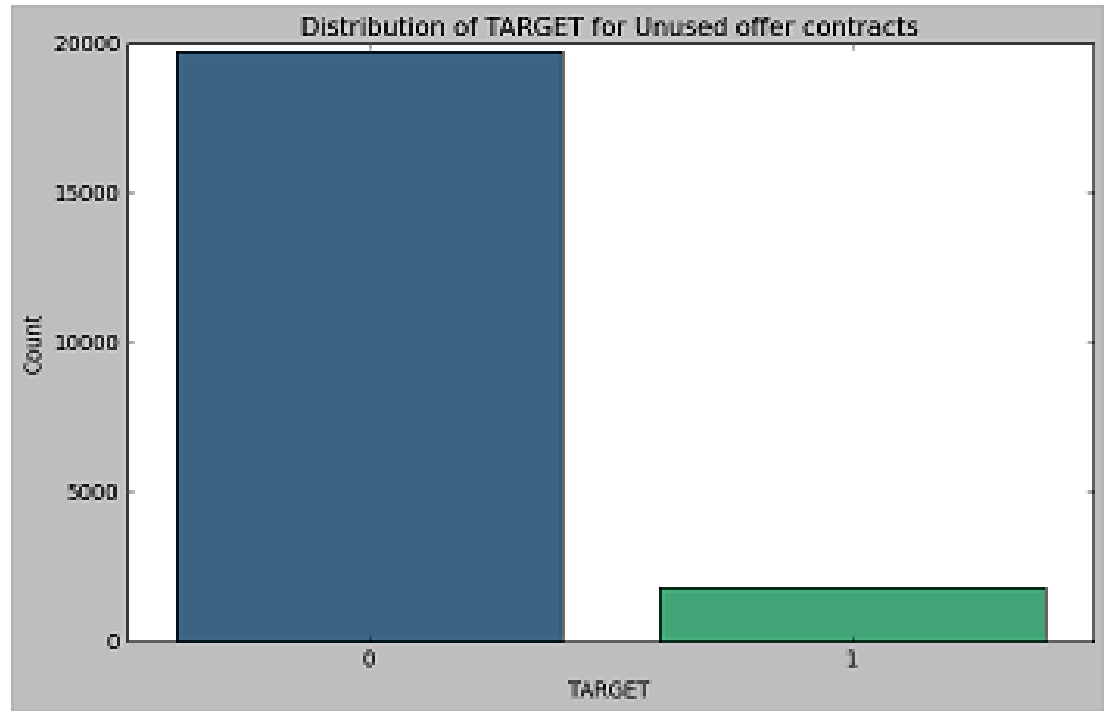
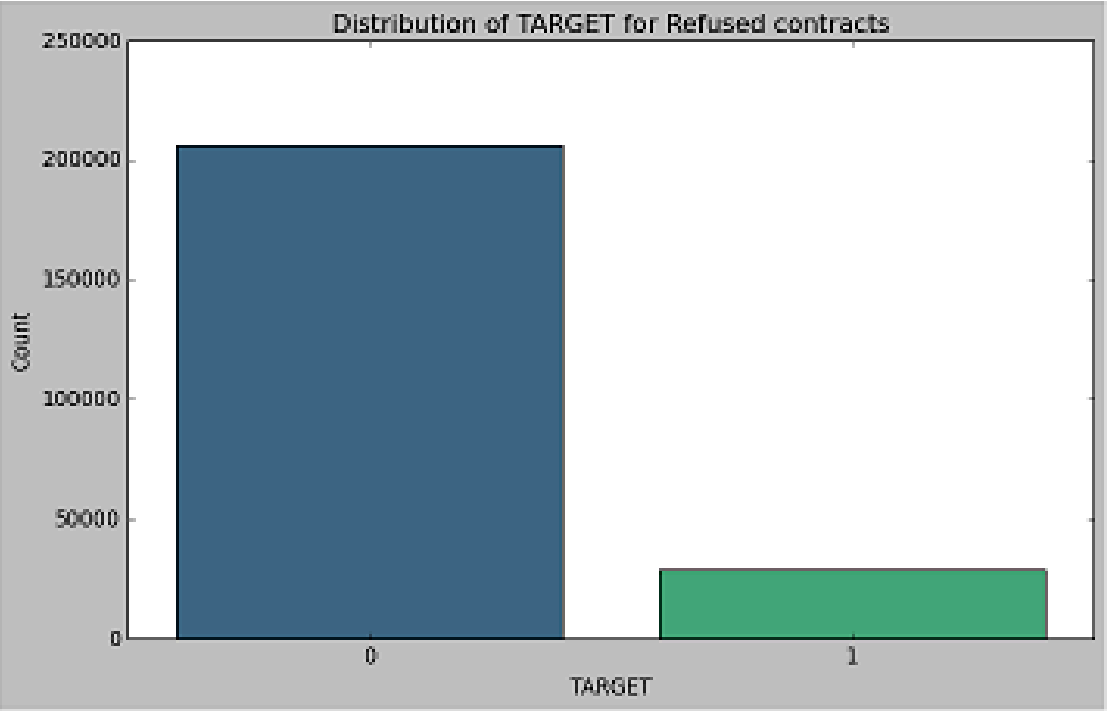
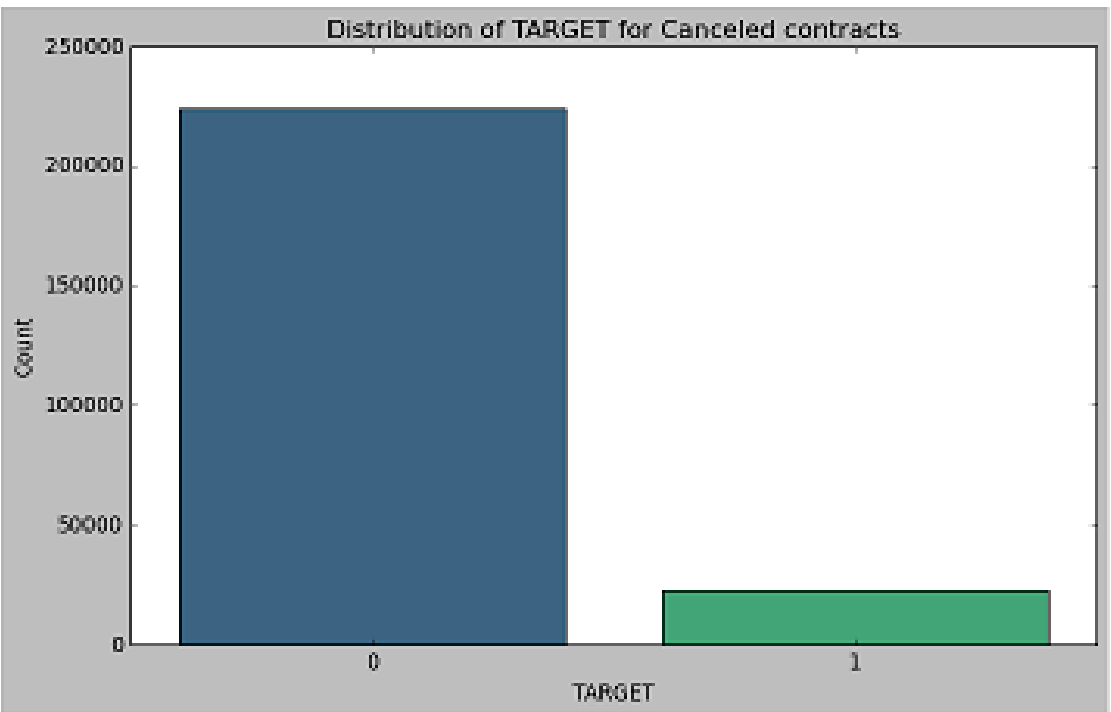
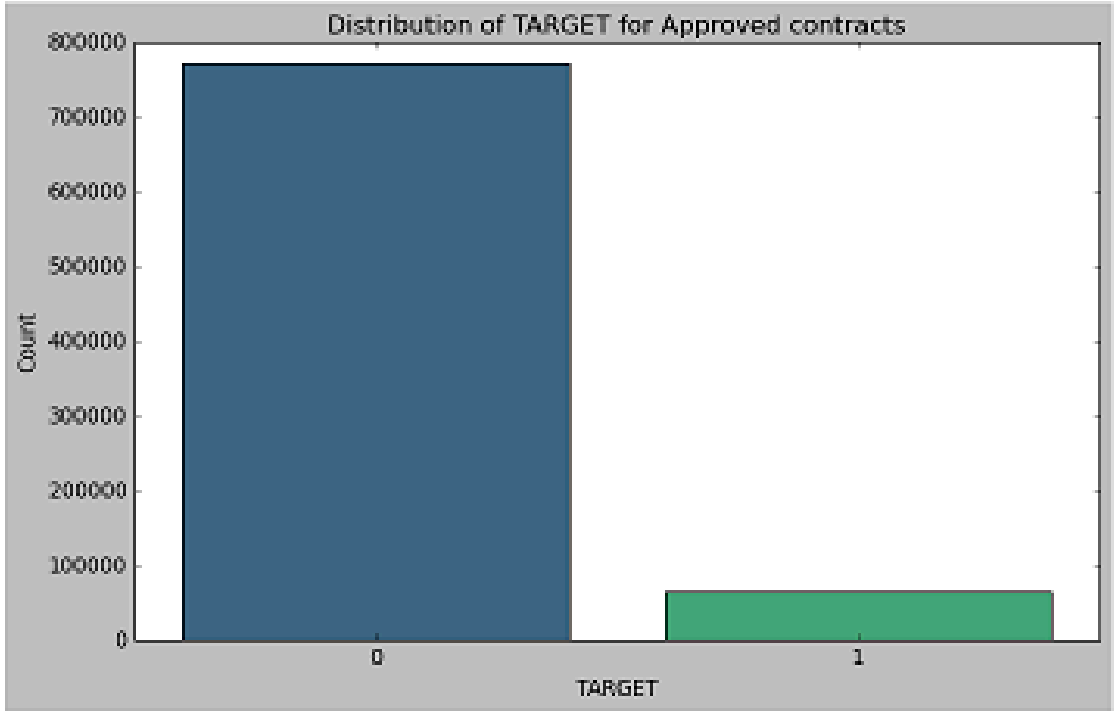
# EDA INSIGHTS



## Insight 11: Loan Purpose by Goods Category and Target

- Majority of loan purposes fall under categories XAP and XNA, showing significant numbers of both non-defaulters and defaulters
- Specified purposes like payments on other loans, buying a used car, repairs, etc., have fewer loans but noticeable proportion of defaulters

# EDA INSIGHTS



## Insight 12: Distribution of Target by Contract Status

- Approved loans: Majority given to clients who do not default
- Cancelled loans: Slightly higher default rate compared to approved contracts
- Refused loans: Highest proportion of defaulters, suggesting some effectiveness in the refusal process
- Unused offers: Slightly higher default rate compared to approved contracts

# TOP 10 CORRELATIONS INSIGHTS

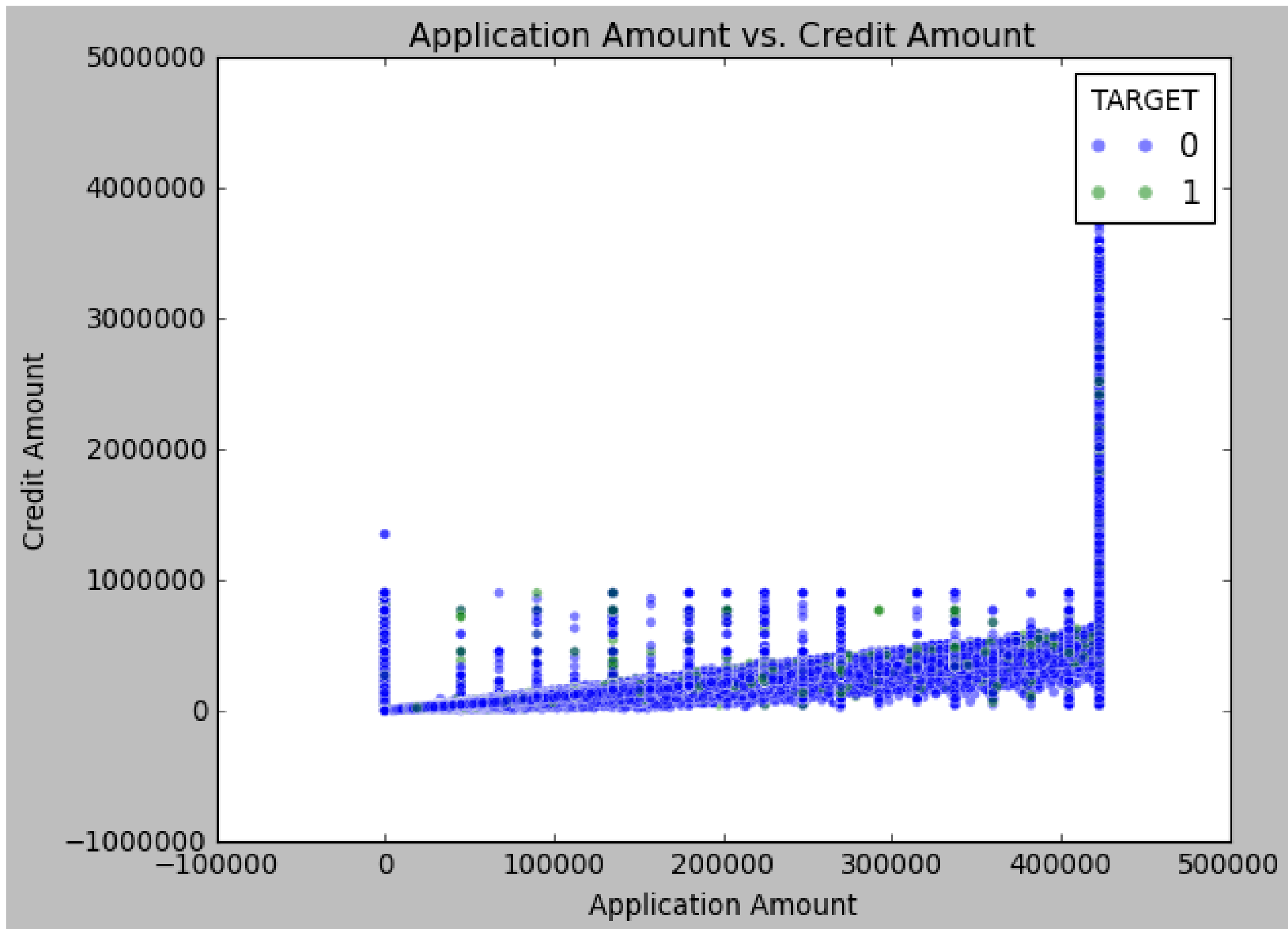


## Scatter Plot Insights:

Insight 1 AMT\_CREDIT\_x vs.  
AMT\_GOODS\_PRICE\_x:

- This scatter plot visualizes the strong positive correlation between the credit amount and the price of goods, highlighting that higher-priced goods typically require larger loans.

# TOP 10 CORRELATIONS INSIGHTS

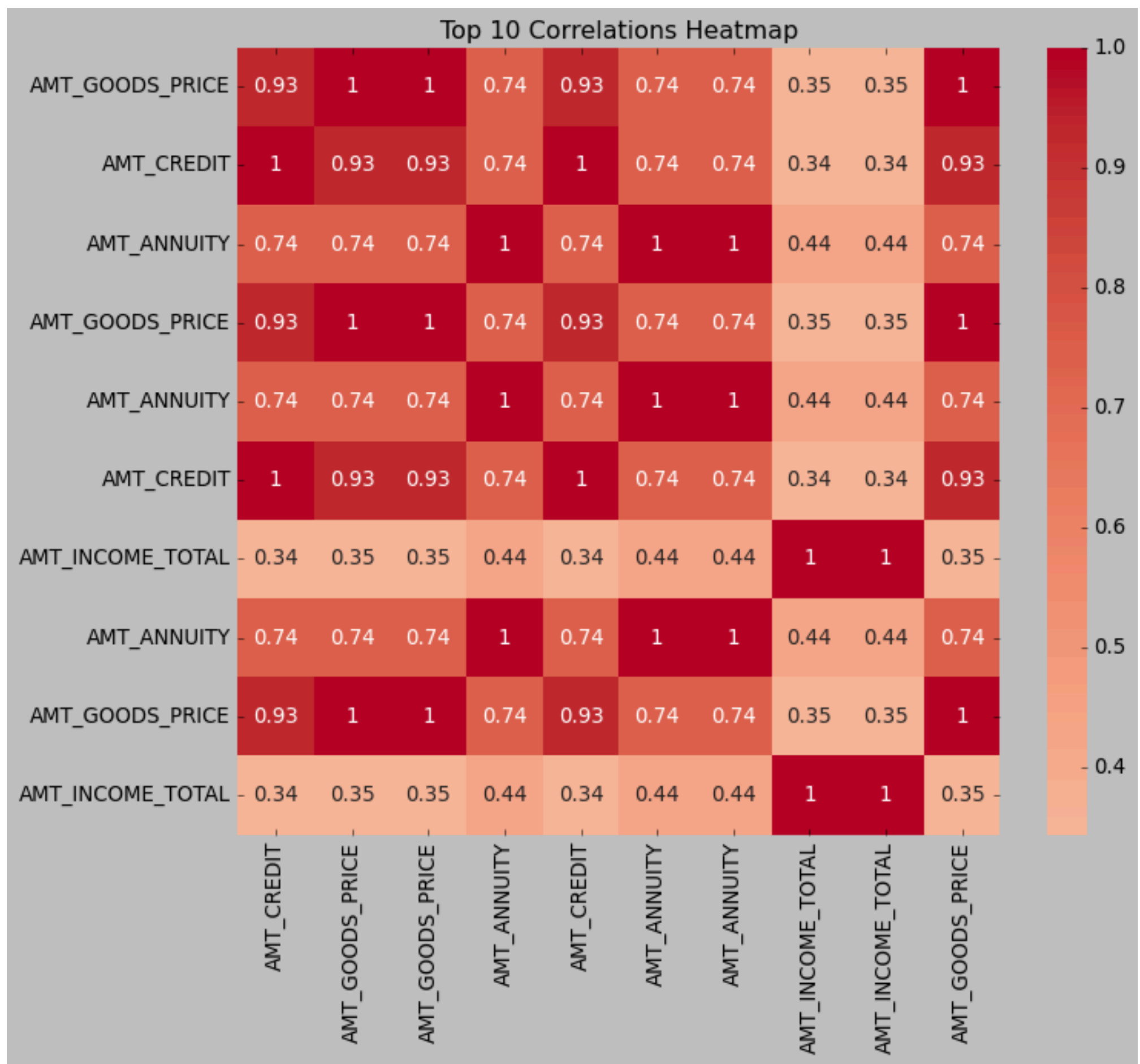


## Scatter Plot Insights:

Insight 2 AMT\_APPLICATION vs.  
AMT\_CREDIT\_y:

- This scatter plot demonstrates the notable correlation between the amount applied for and the credit amount granted, indicating that applicants often receive loan amounts close to what they apply for.

# TOP 10 CORRELATIONS INSIGHTS

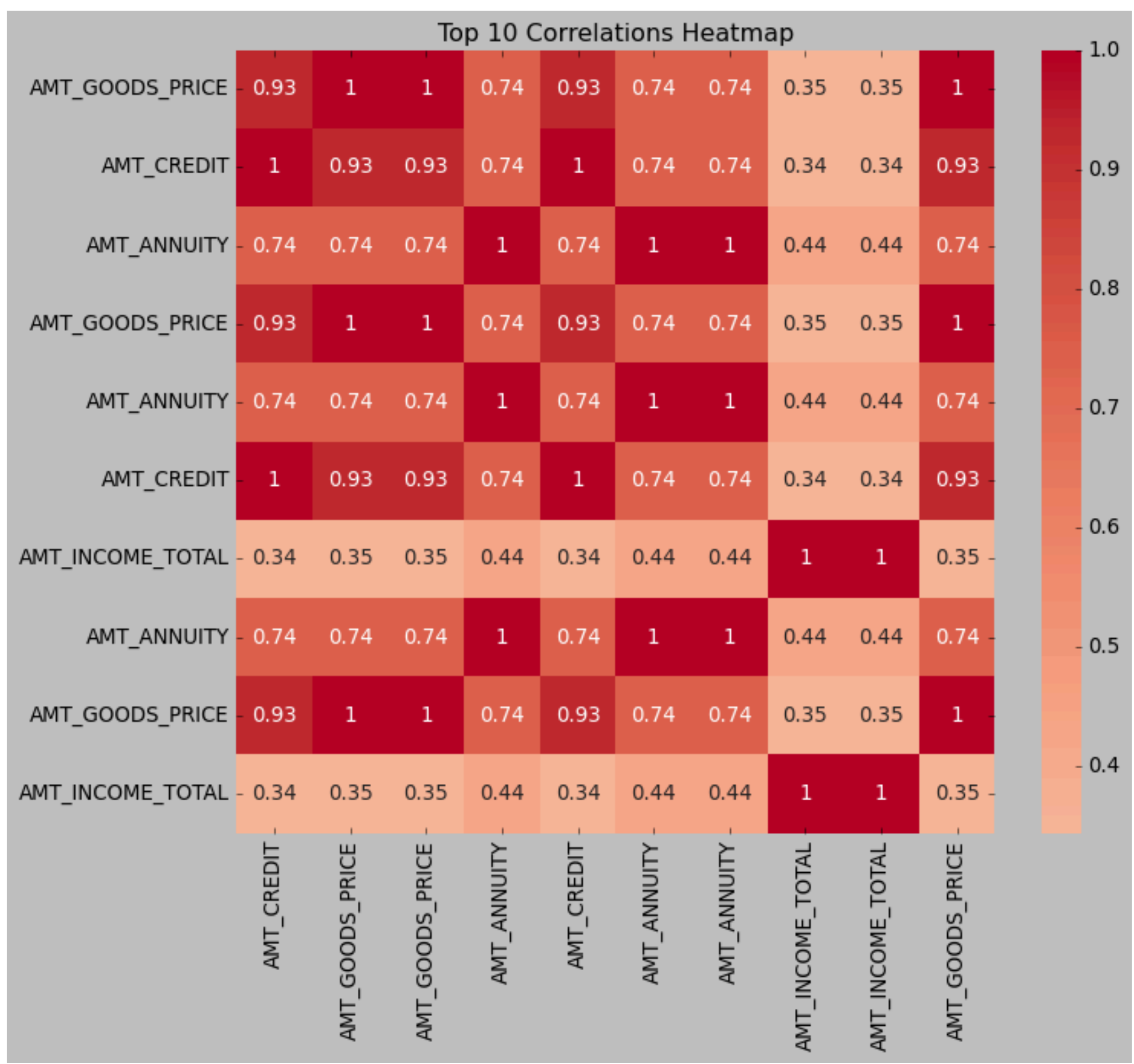


## Heat Map Insights:

The heatmap showing the correlation matrix of the top 10 correlations provides a clear visual representation of the relationships between key financial indicators.

- Credit Amount and Goods Price: Strong positive correlation (0.93) indicating that higher-priced goods require larger loans.
- Credit, Annuity, and Goods Price: High correlations above 0.74 suggest these financial indicators increase together, highlighting their interconnectedness.
- Income Influence: Moderate correlation of income with credit (0.34) and annuity (0.44), indicating income levels influence loan amounts, but not as strongly as other factors.

# TOP 10 CORRELATIONS INSIGHTS



- Variable Redundancy: Near-identical correlations suggest some variables may be redundant, indicating potential for consolidation in models.
- Consistency: Correlations between AMT\_CREDIT and AMT\_GOODS\_PRICE are consistently strong, indicating a reliable relationship for loan assessment.



# FUTURE SCOPE

Future scope of the analysis:


## 1. Advanced Feature Engineering:

- Explore creating more derived variables to capture complex relationships.
- Include interaction terms and polynomial features for better predictive power.

## 2. Incorporating External Data:

- Integrate external economic indicators (e.g., employment rates, GDP growth) to enrich the dataset.
- Use geographic data to analyze regional trends in loan default rates.

## 3. Modeling and Prediction:

- Implement advanced machine learning models (e.g., Random Forest, Gradient Boosting) to predict loan defaults.
  - Perform hyperparameter tuning and cross-validation to improve model performance.
- 

The image features a light gray background with the text "THANK YOU" centered in a bold, blue, sans-serif font. The corners are decorated with abstract geometric patterns. The top-left corner has a series of thin, parallel, light blue diagonal lines. The top-right corner features a cluster of overlapping semi-circles in yellow, red, and teal. The bottom-left corner shows a similar cluster of overlapping semi-circles in red, teal, and blue. The bottom-right corner contains a large, thin, light blue arc with several parallel diagonal lines extending from its base.

**THANK YOU**