EDA CASE STUDY :Bank Loan Risk Analysis

Bank loan default risk analysis is a critical process for financial institutions to assess the likelihood of loan default by borrowers. It involves evaluating various factors to determine the potential risk associated with lending money.

~ Prithibi Mondal

Problem Statement

Enhancing Loan Decision-making Through EDA"

- Objective: Minimize financial risk and optimize lending decisions.
- Challenge: Limited credit history leading to potential defaults.
- EDA Goals:
    - Risk Minimization
    - Pattern Identification
    - Decision Factors Analysis
    - Portfolio Enhancement
- Empowerment: Informed decisions, improved risk assessment, and enhanced portfolio management.

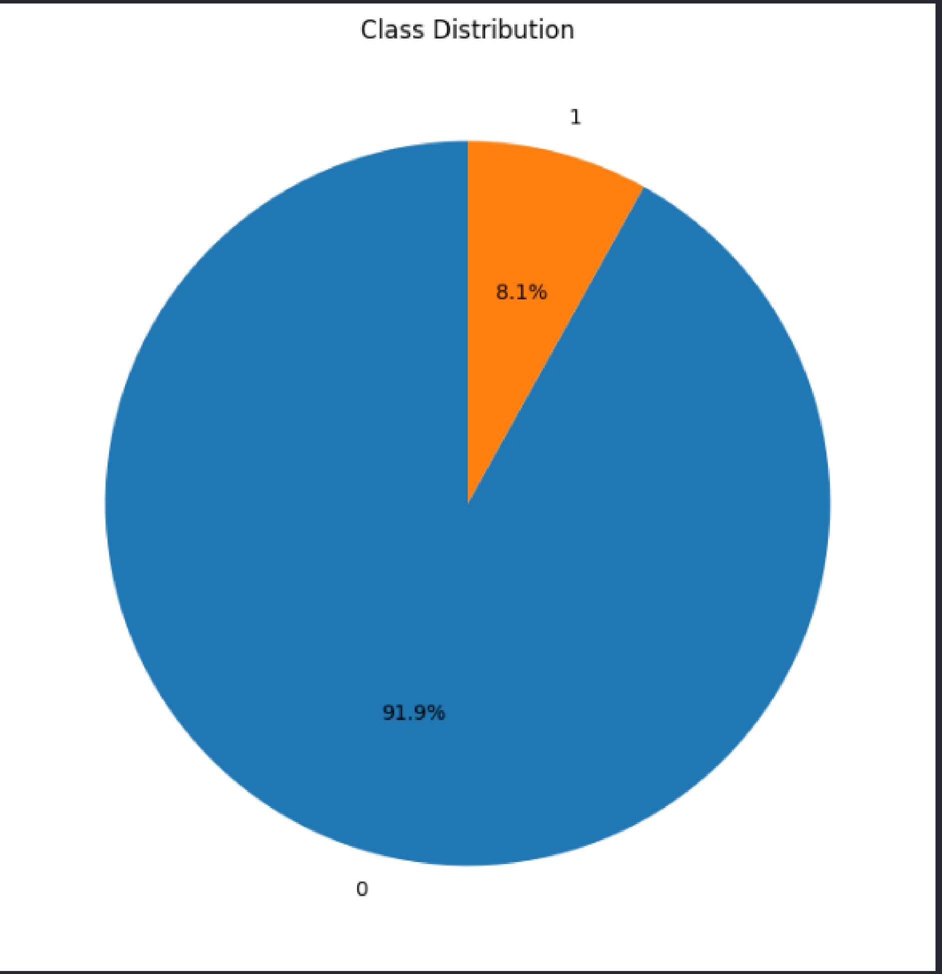
Analysis of the data set has been done in Python on Google Colab

Approach and Methodology

1. **Data Preparation:**
    - Collect comprehensive loan application data, addressing missing values and outliers.
    - Ensure dataset cleanliness for accurate analysis.
2. **Exploratory Visualization:**
    - Utilize univariate, bivariate, and multivariate analyses to identify patterns and relationships.
    - Gain insights into consumer attributes and loan characteristics.
3. **Insightful Feature Discovery:**
    - Identify key factors influencing loan defaults through data-driven insights.

Data Imbalance

Data imbalance can significantly impact the accuracy of models, potentially leading to biased outcomes. Addressing this issue is crucial for the reliability of loan default risk analysis.



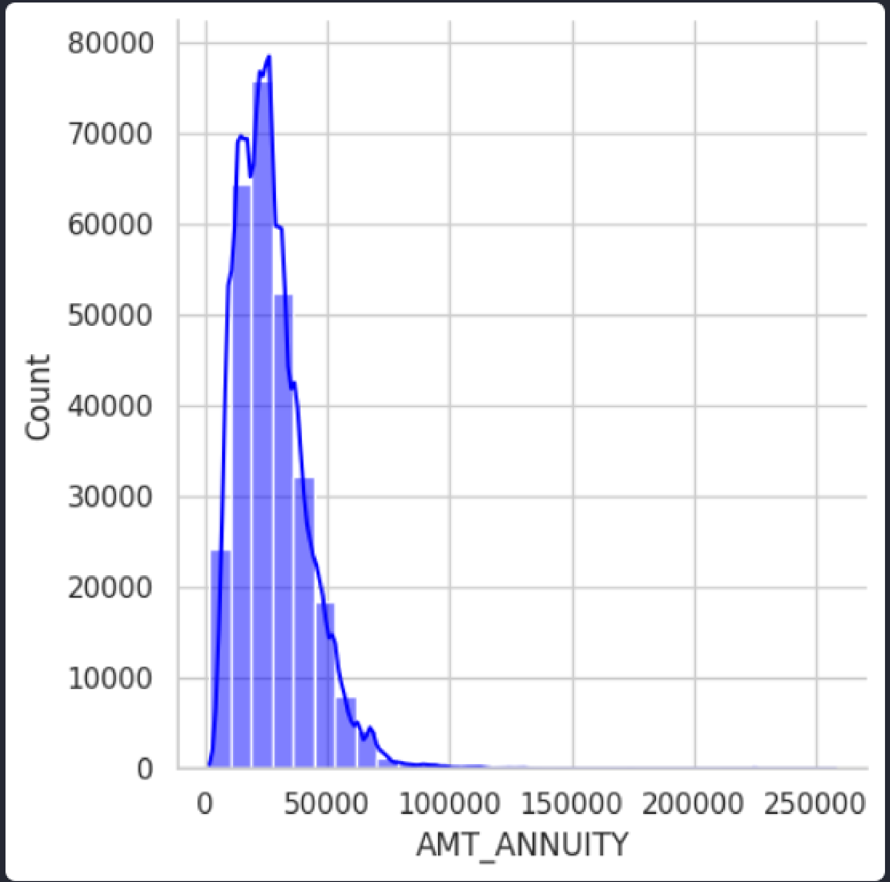
The application data is highly imbalanced

Defaulters are 8.1% of total

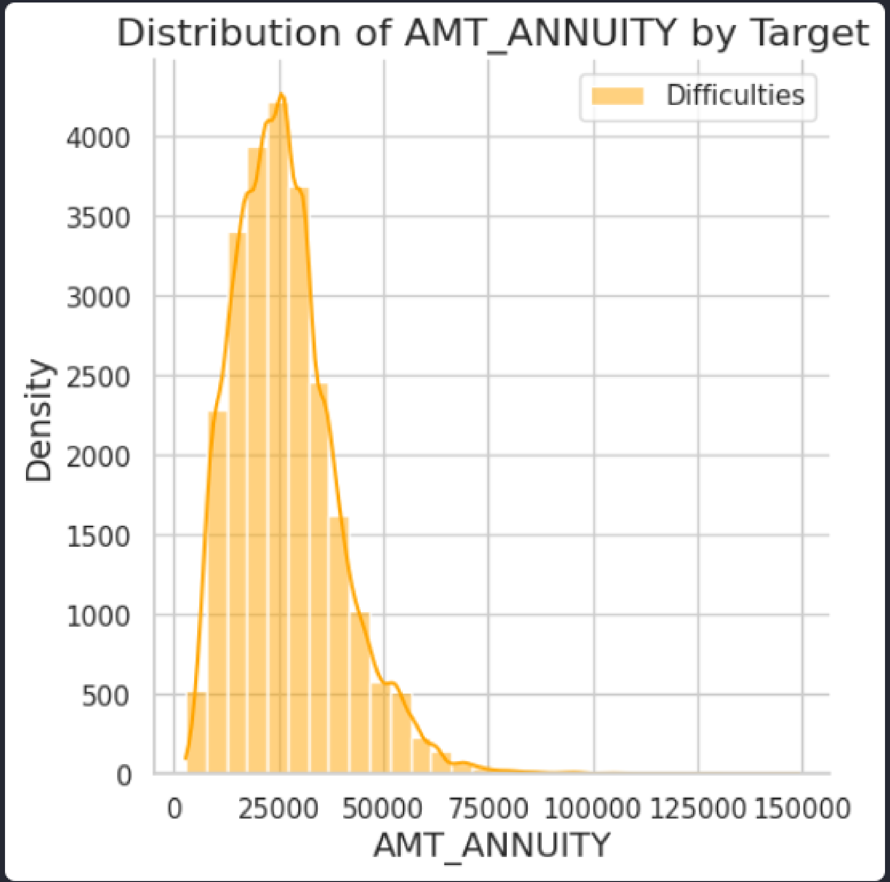
91.9% are non-defaulters

Imbalance Ratio: 11.387150

Proportion of Defaulters by Annuity AMT



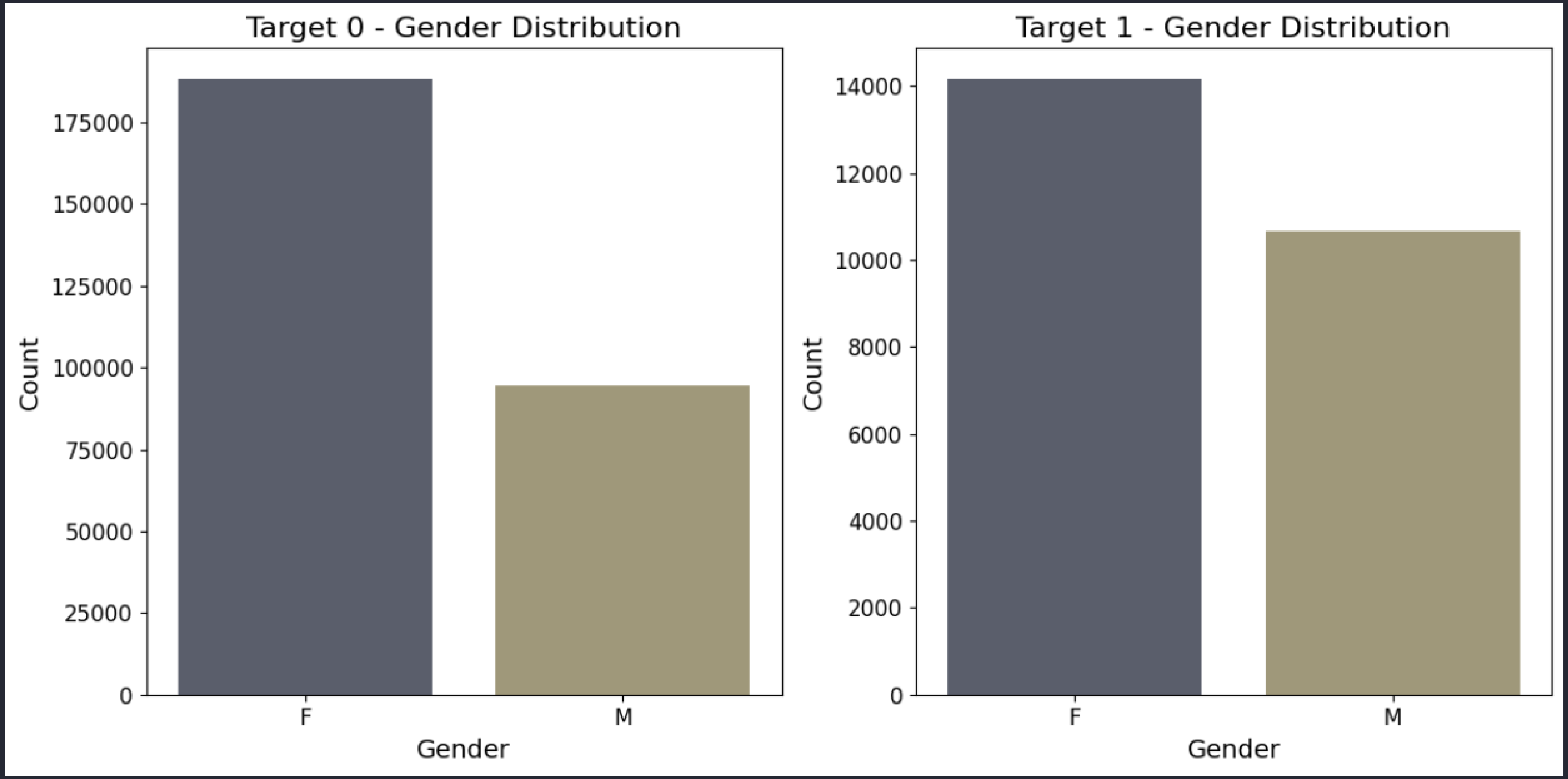
Target-0



Target-1

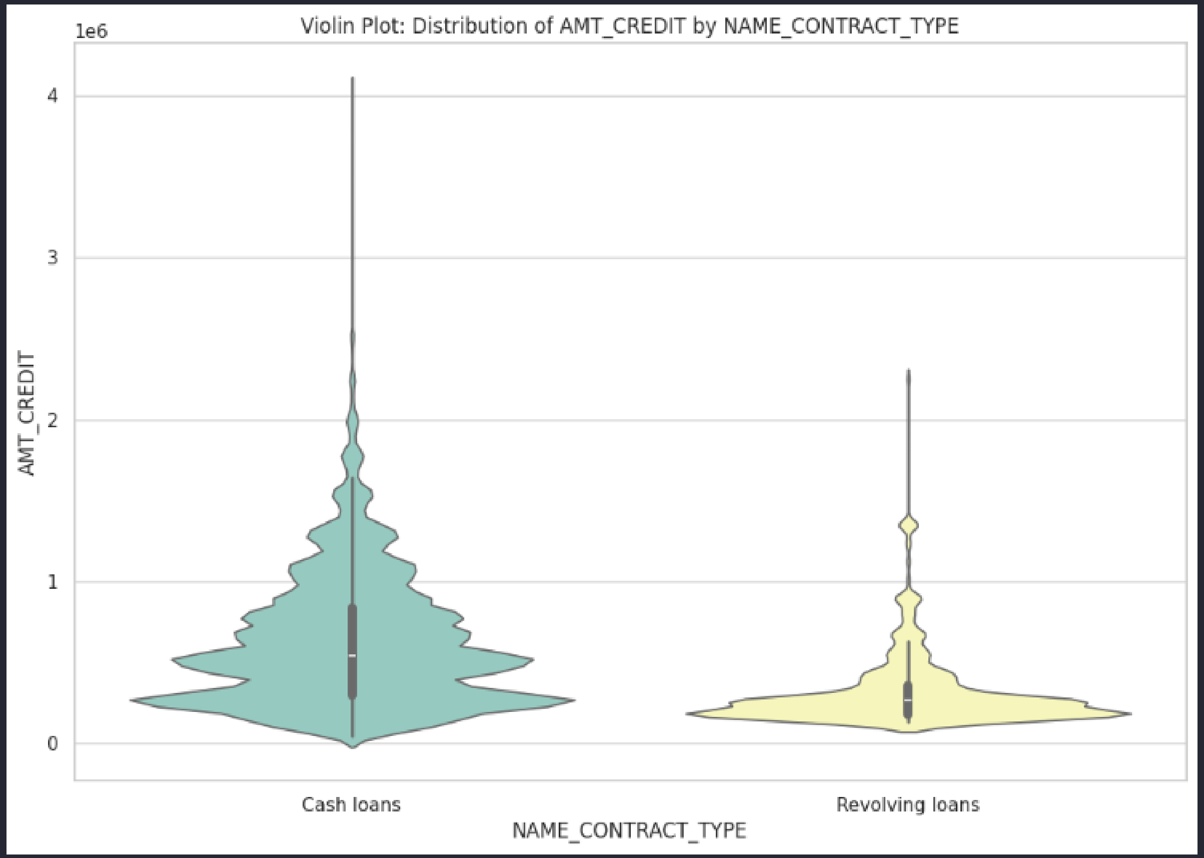
- The annuity shapes are clearly identical for Target 0 and Target 1, as the graphic makes evident. Additionally, the plots show those who struggle to repay loans in relation to their annuities.

Gender Distribution



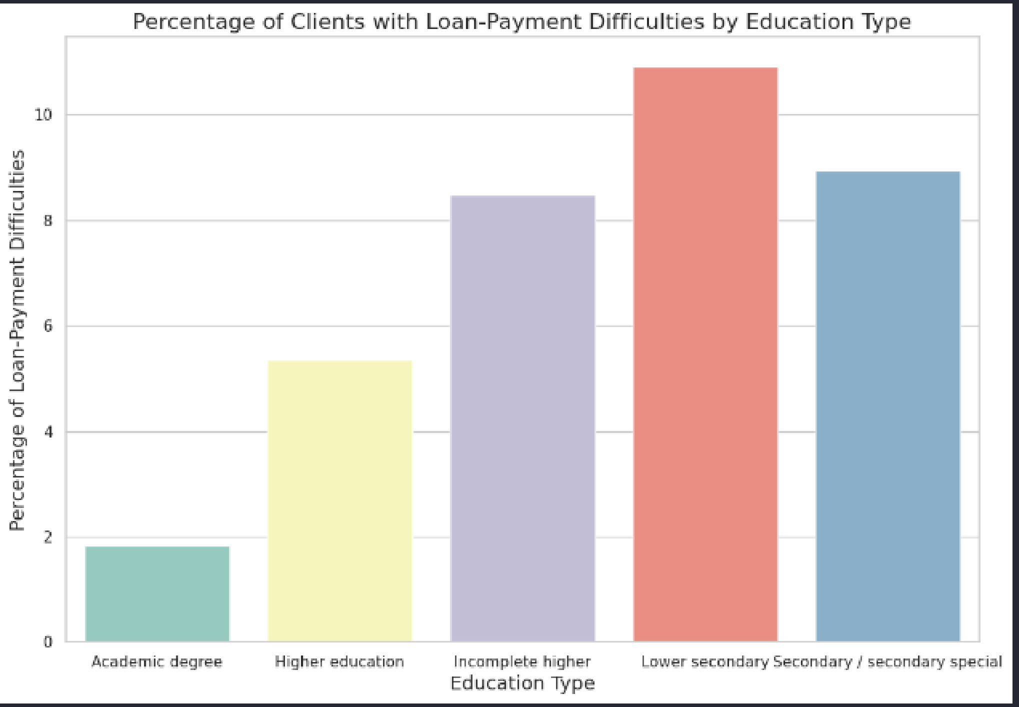
- It appears that more female clients than male consumers applied for loans. 33.4% of male clients and 66.6% of female clients are non-defaulters. 42% of male clients default, compared to 57% of female clients.

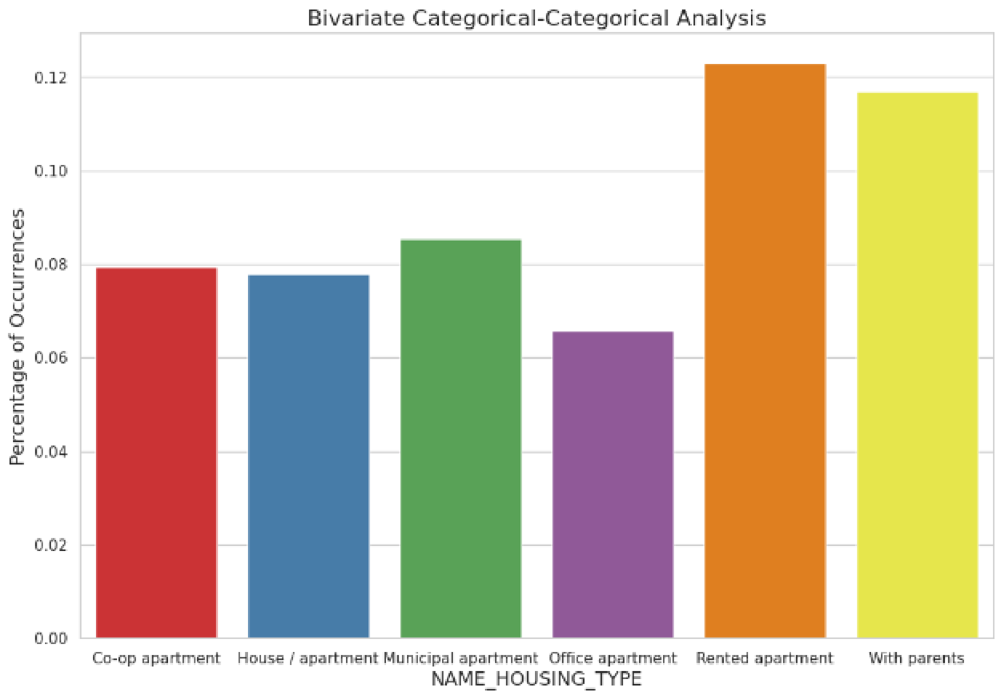
Distribution of Credit AMT by NAME_CONTRACT_TYPE



- This is the violin plot distribution of credit amount by contract type. While a relatively small percentage of clients have asked for revolving loans for both defaulters and non-defaulters, the majority of clients have requested for cash loans.

Loan Payments Difficulties



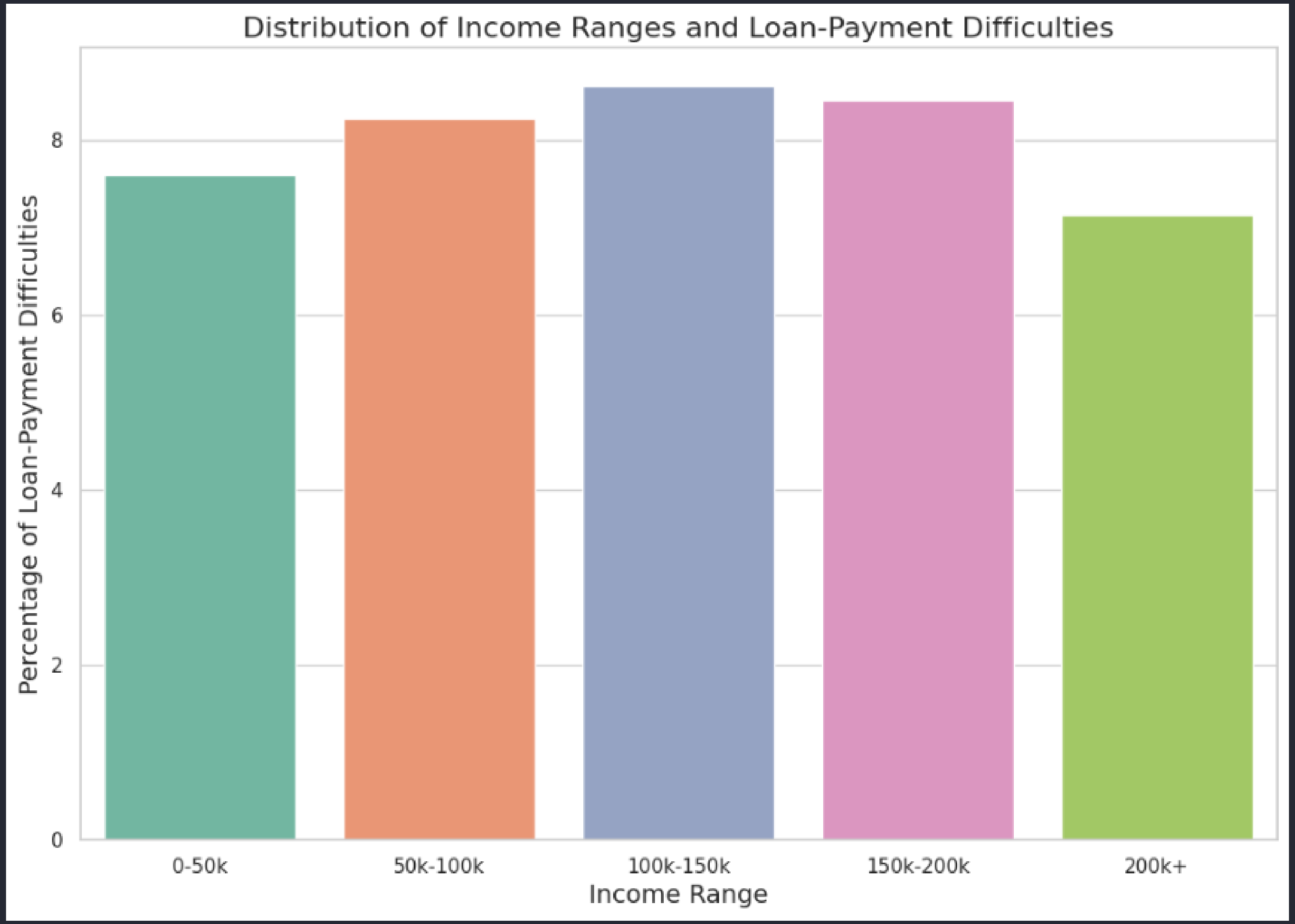


- Education Type with Maximum Loan-Payment Difficulties: Lower secondary

  Loan-Payment Difficulties percentage: 10.93% .
- Housing Type with Maximum Loan-Payment Difficulties: Rented apartment

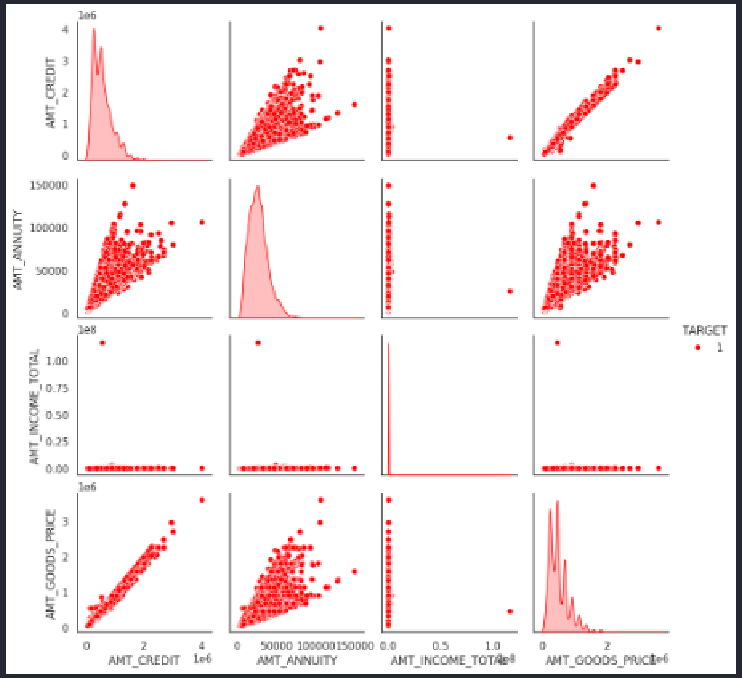
  Loan-Payment Difficulties percentage: 12.31%

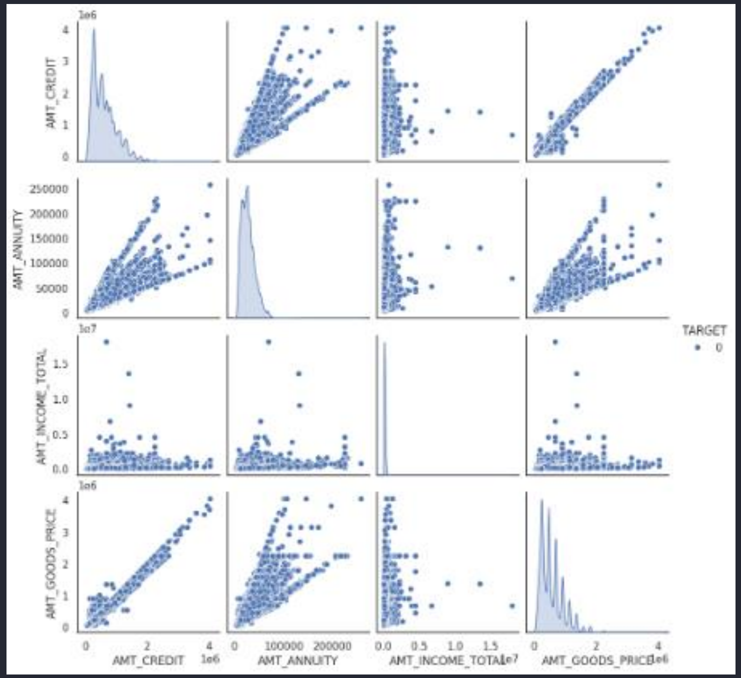
Distribuition of Loan Payment Difficulties by Income Range



- Income Range with Maximum Loan-Payment Difficulties: 100k-150k
  Loan-Payment Difficulties percentage: 8.62%
- Applicant with Income more than 200k are less likely to default than others

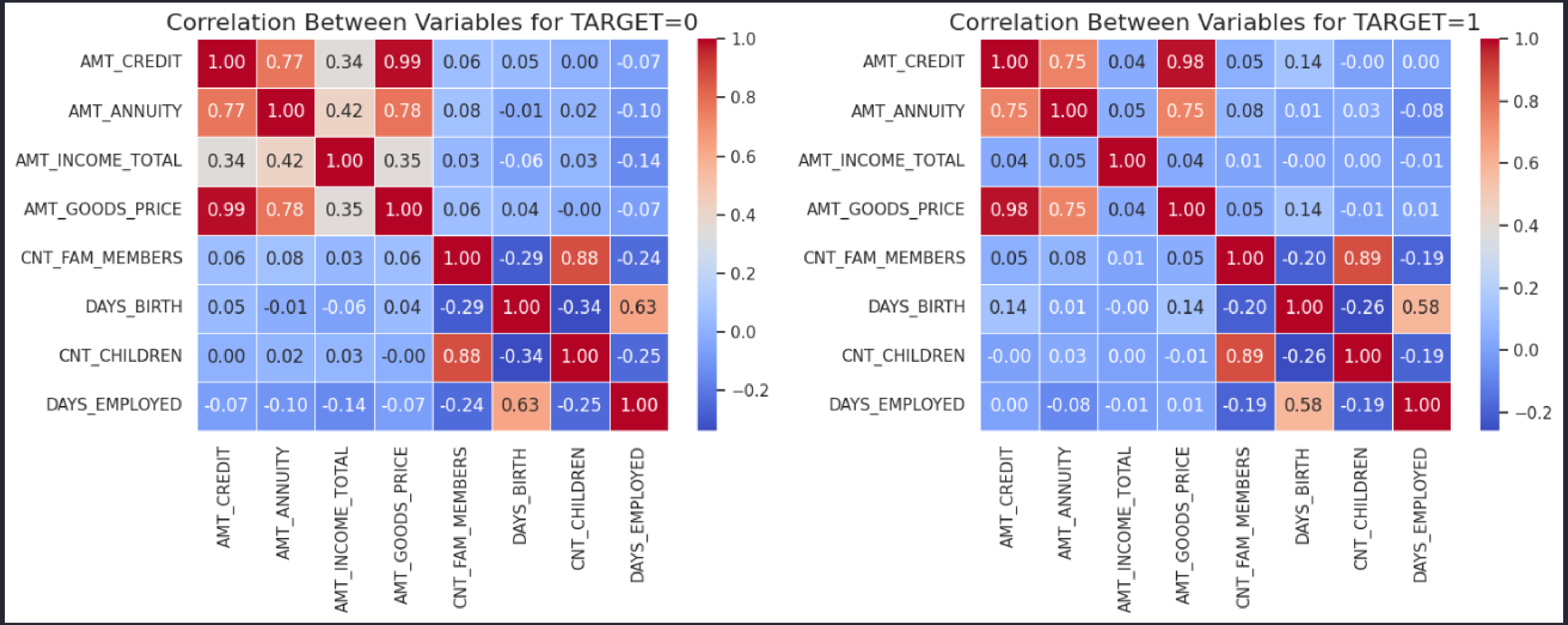
Correlations-Pairplot





- The variables AMT_CREDIT and AMT_GOODS_PRICE exhibit a strong correlation with both defaulters and non-defaulters. Thus, the loan amount rises in tandem with the price of the residence. For both defaulters and non-defaulters, there is a strong correlation between the variables AMT_CREDIT and AMT_ANNUITY (EMI). Thus, it makes sense that when the price of a home rises, so does the EMI amount. The strong correlation between the three variables—AMT_CREDIT, AMT_GOODS_PRICE, and AMT_ANNUITY—for both defaulters and non-defaulters suggests that they may not be useful for identifying defaulters.

Correlations-Heatmap



- AMT_CREDIT is inversely correlated with DAYS_BIRTH; individuals in the lower age group tend to take out larger credit amounts, and vice versa. Because AMT_CREDIT and CNT_CHILDREN are negatively correlated, clients with fewer children have larger credit amounts, and vice versa. Since AMT_INCOME_TOTAL and CNT_CHILDREN are inversely correlated, clients with fewer children will have higher incomes, and vice versa.

Conclusion

Following dataset analysis, specific client attributes crucial for predicting loan repayment have been identified. These contributing factors are categorized, offering the bank key insights for informed lending decisions and enhanced risk assessment

1. AMT_INCOME_TYPE: Customers in the middle of the pay scale are more likely to apply for a loan, whether they are non-defaulters or defaulters. Low- and middle-class clients are more likely to default.
2. AMT_CREDIT_TYPE: For both defaulters and non-defaulters, the majority of clients asked for loans with a medium credit amount. There is a significant default risk for clients applying for both high and low credit
3. AMT_INCOME_TOTAL: Applicant with Income more than 200,000 are less likely to default
4. CNT_CHILDREN: People with zero to two children tend to repay the loans.
5. NAME_EDUCATION_TYPE: It is more likely that clients with Secondary or Secondary Special education will apply for the loan. Clients with Secondary or Secondary Special education are more likely to default. Other forms of schooling carry less danger
6. OCCUPATION_TYPE: The majority of loan applications, both from defaulters and non-defaulters, have come from pensioners. The largest group, pensioners, are most likely to default, followed by workers.