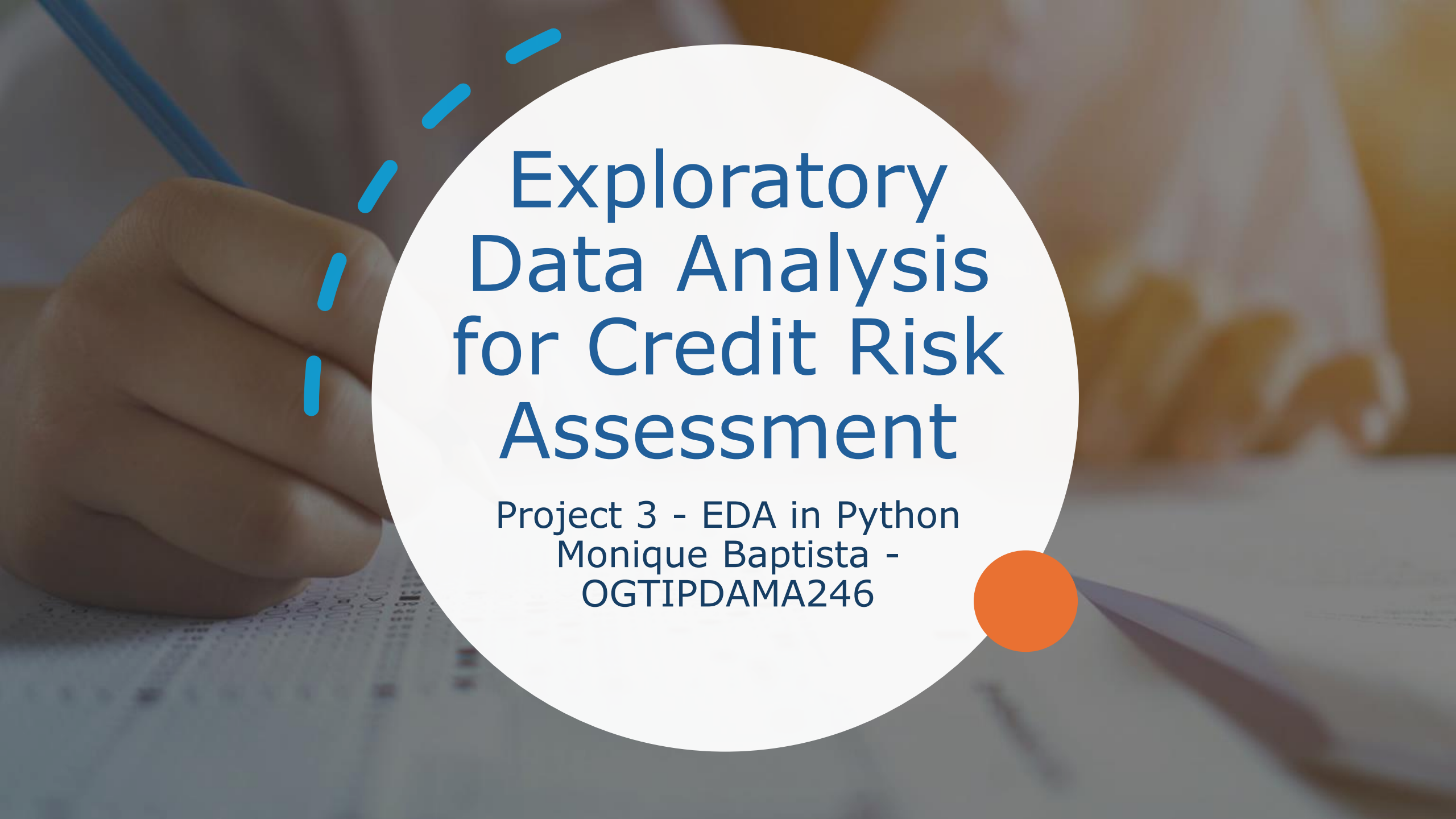


Exploratory Data Analysis for Credit Risk Assessment

Project 3 - EDA in Python

Monique Baptista - OGTIPDAMA246

Agenda

1. Introduction
2. Target Univariate Analysis
3. Types of Contract Vs Target - Bivariate Analysis
4. Gender and Age - Univariate Analysis
5. Gender vs Age - Bivariate Analysis
6. Gender Vs Age per Target group - Multivariate Analysis
7. Income Type and Income Amount - Univariate Analysis
8. Income Type and Amount for Target - Multivariate Analysis
10. Correlation Matrix: Income, Credit, Annuity and Price of the Good - Multivariate Analysis
11. Pair Plot: Income, Credit, Annuity and Price of the Good correlation Multivariate Analysis
12. Summary

Introduction

Goal: Conduct an Exploratory Data Analysis (EDA) to identify key patterns and variables indicating potential difficulties in loan repayment. This analysis aims to assist the company in risk assessment and enable informed decisions regarding loan approvals, amounts, and interest rates.

Tasks: 1. Data Collection: Load the current and previous datasets and merge them; 2. Data Cleaning: Duplicate the original dataset, drop duplicated rows, handle missing values by dropping or imputing values, transform data types and create new features for enhanced analytical insights; 3. Data Analysis and Visualisation: Perform univariate, bivariate, and multivariate analyses, uncovering patterns and complex relationships in a more accessible and insightful way.

Data: Two datasets in CSV format were provided: Application data and Previous application. The Application data is the current dataset with 307,511 unique ID numbers and 124 features. The columns include client demographic information such as gender, educational level, and marital status; financial information such as the client's income, credit amount received, and loan annuity; and more specific information such as the number of inquiries to the Credit Bureau and flags indicating if the client provided various documents. The second dataset contains previous financial information, including product type, financial aspects like annuity and credit amounts, timing of applications, flags indicating the status of the previous application, rates and ratios related to down payment and interest, and client-specific information.

Target
Univariate Analysis

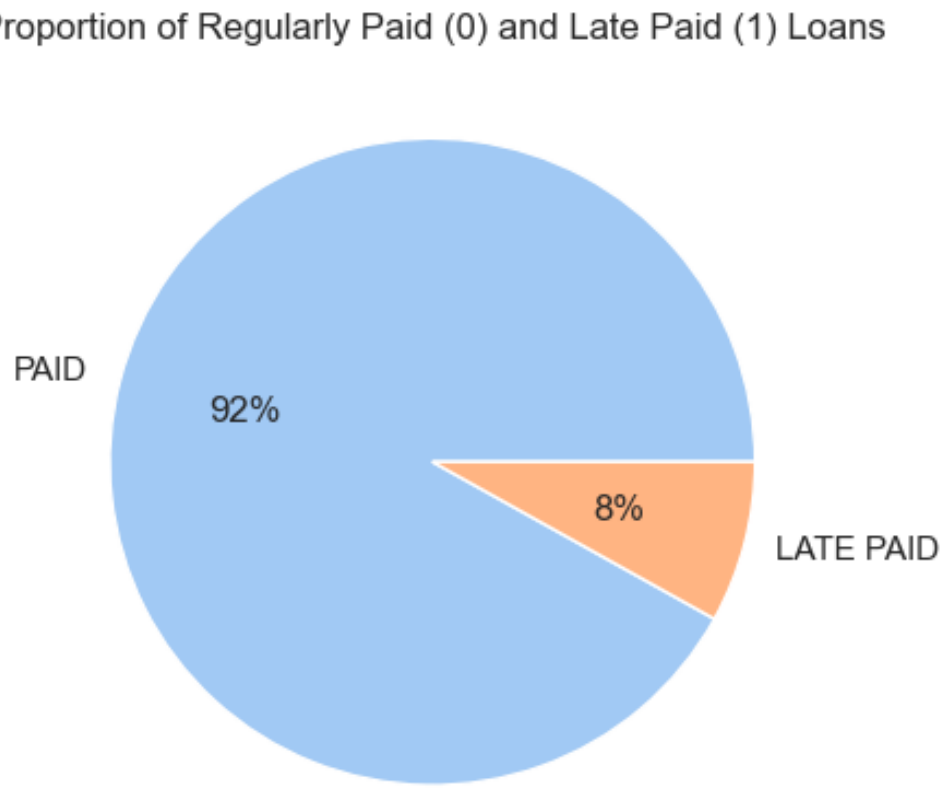


Figure 1

- The target variable is highly imbalanced, with Target 1 (Late Paid Loan) representing only 8% and Target 0 (Regularly Paid Loan) dominating at 92% in our dataset.
- Given the focus on understanding credit risk factors, the group of interest is Target Group 1, which consistently demonstrates difficulties in repaying loans.

Types of Contract Vs Target
Bivariate Analysis

Cash Loans are prevalent in both Target groups, constituting 94% in Target 1 and 90% in Target 0.

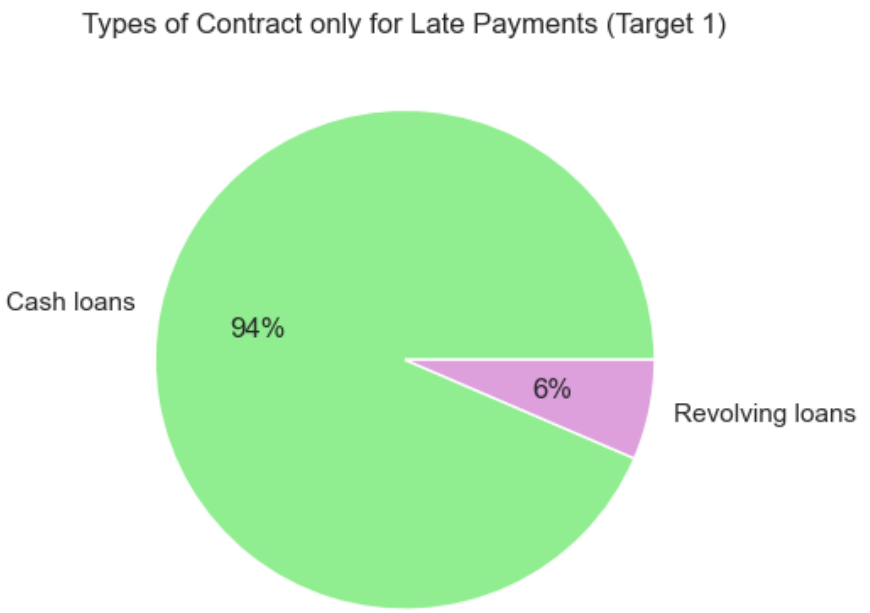


Figure 2

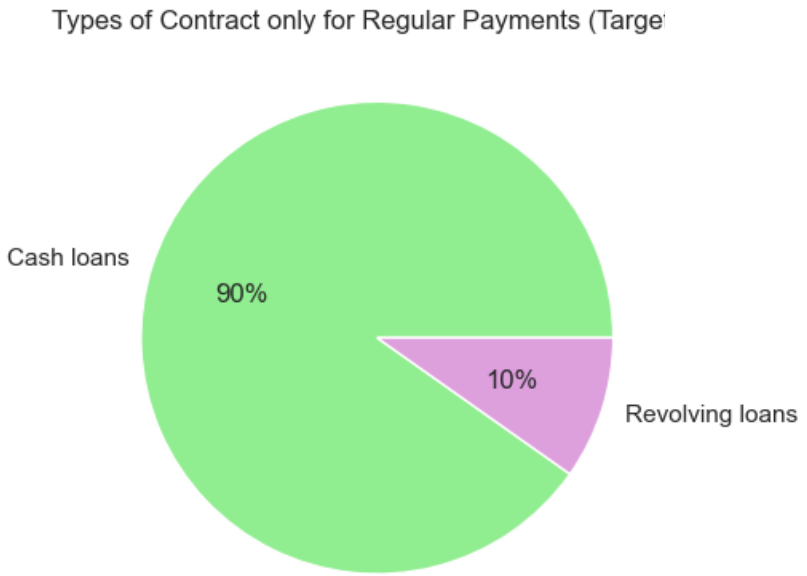


Figure 3

Gender and Age Univariate Analysis

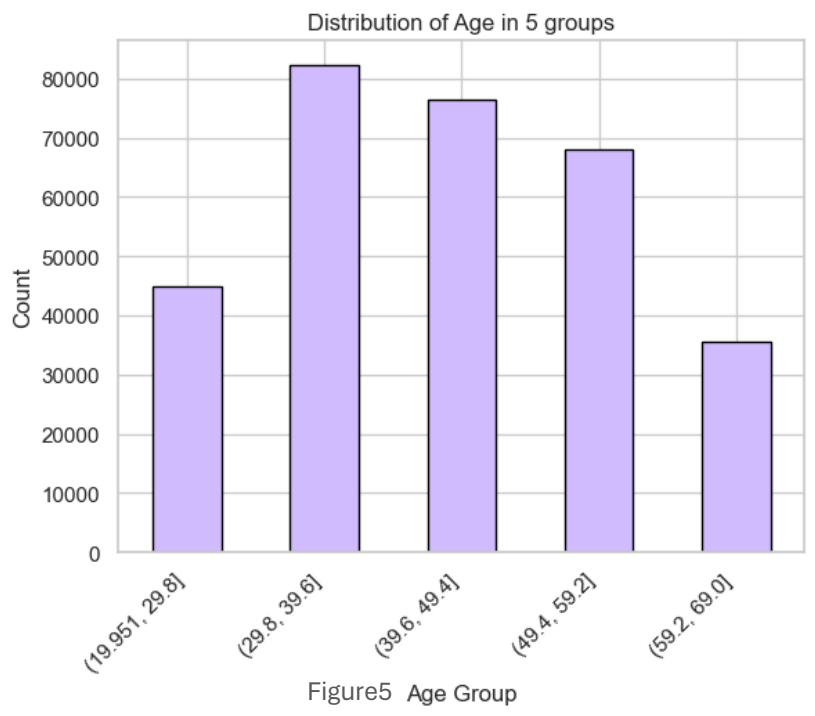


Figure5

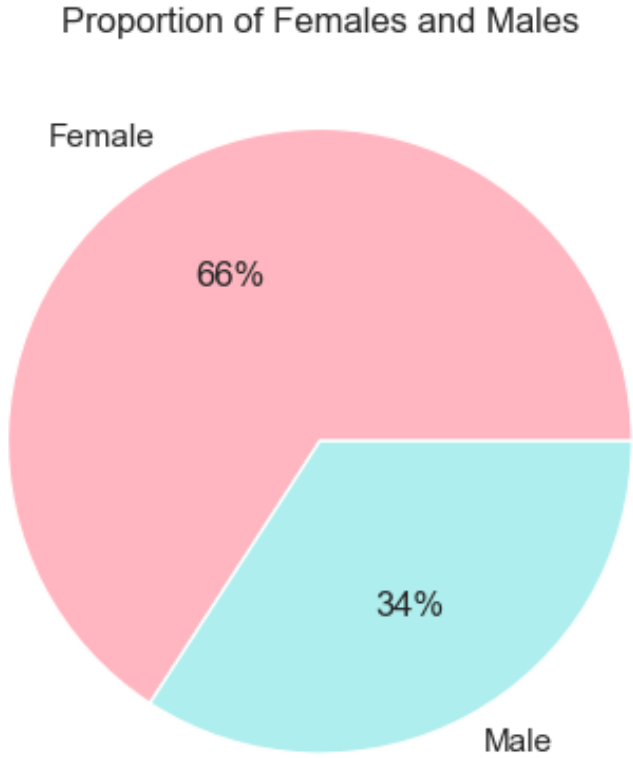


Figure 4

- Fig4- Females represent 66%, twice the percentage of males (34%) in our dataset.
- Fig5- The most frequent age range is 30 to 40, constituting 27% of the data, followed by the range 40 to 50 years old at 25%.

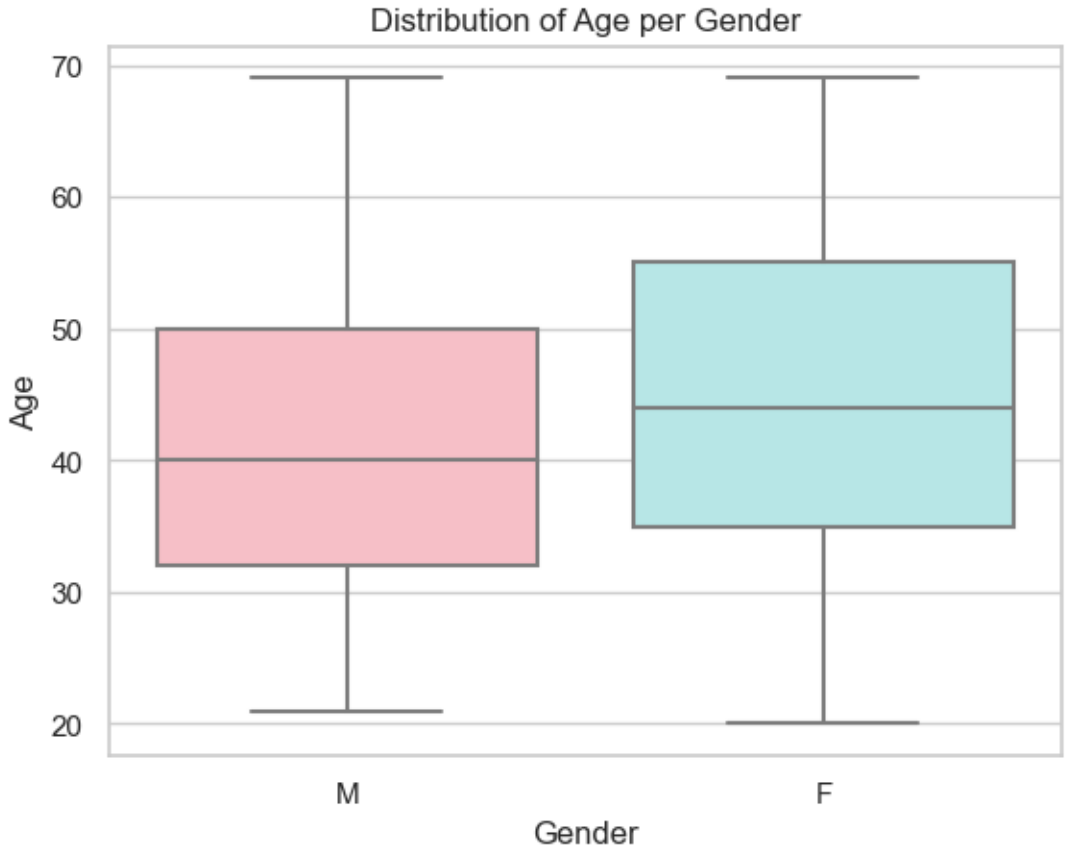


Figure 6

Gender Vs Age - Bivariate Analysis

- The minimum and maximum age values are around 20 and 70 years old, respectively, for both groups.
- The female median is around 45 years old, while the male median is 40 years old.
- No outliers were found.

Gender Vs Age per Target group - Multivariate Analysis

- Fig7 - Target group 0 has an Interquartile Range Age for males around 30 to 50 with 40 as the median, while the female group is much older, ranging from 45 to 55 with a median of 45 years old.
- Target group 1 has a much younger Interquartile Range Age, with males ranging from 30 to 45 and females from slightly more than 30 to 50, with a median of 40—5 years less than Target group 0.
- Fig8 - Females have a wider interquartile range, reaching 50 years old for the 75th percentile, whereas males have 45 for the same mark.

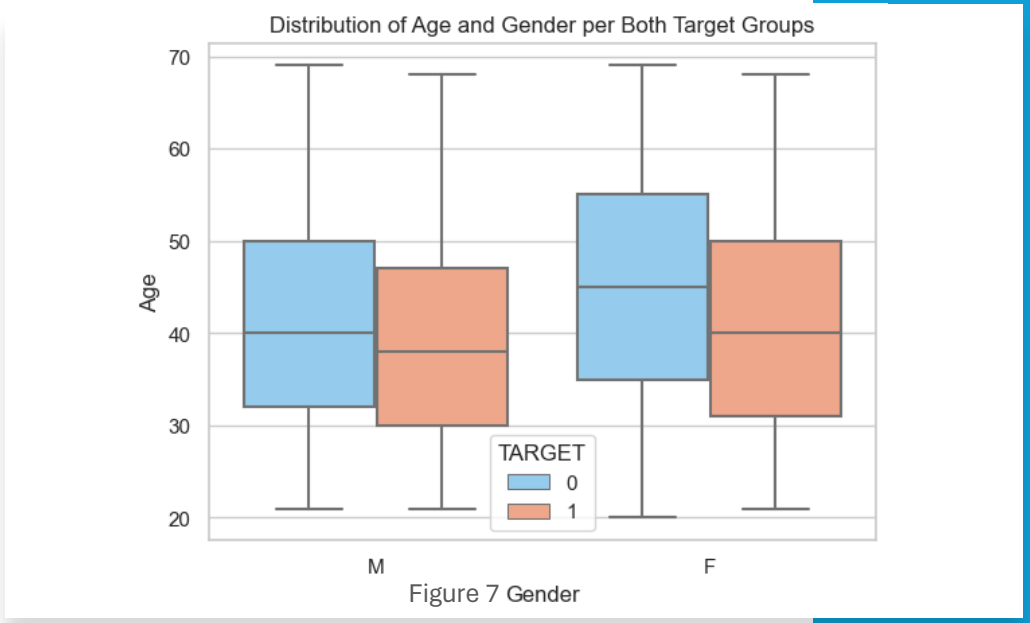


Figure 7 Gender

Figure 8

Income Type and Income Amount - Univariate Analysis

- Fig9 - Working group leads with more than half of Income types (52%), followed by Commerce Associates at almost a quarter (23%), then Pensioners (18%), and State servants (7%) in the fourth position.
- Fig10 - Notable outliers can be observed, with the client's income having a maximum value of 117,000,000.

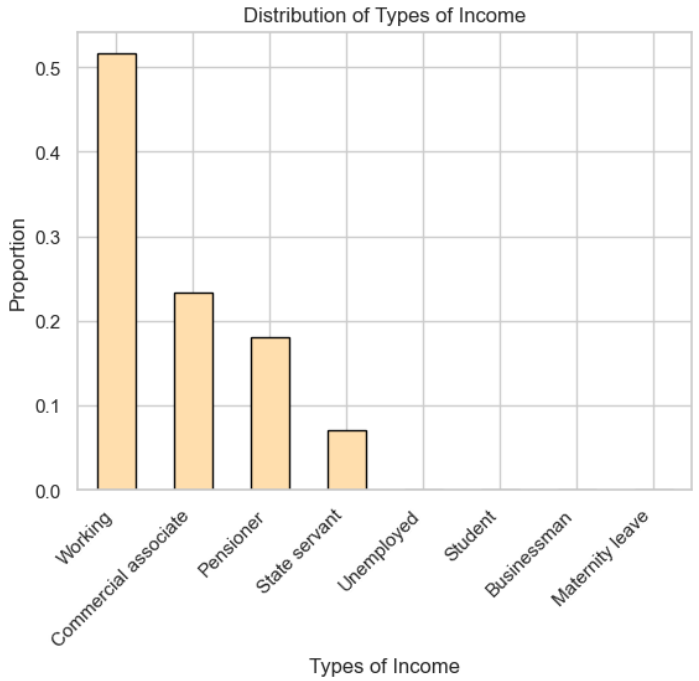


Figure 9

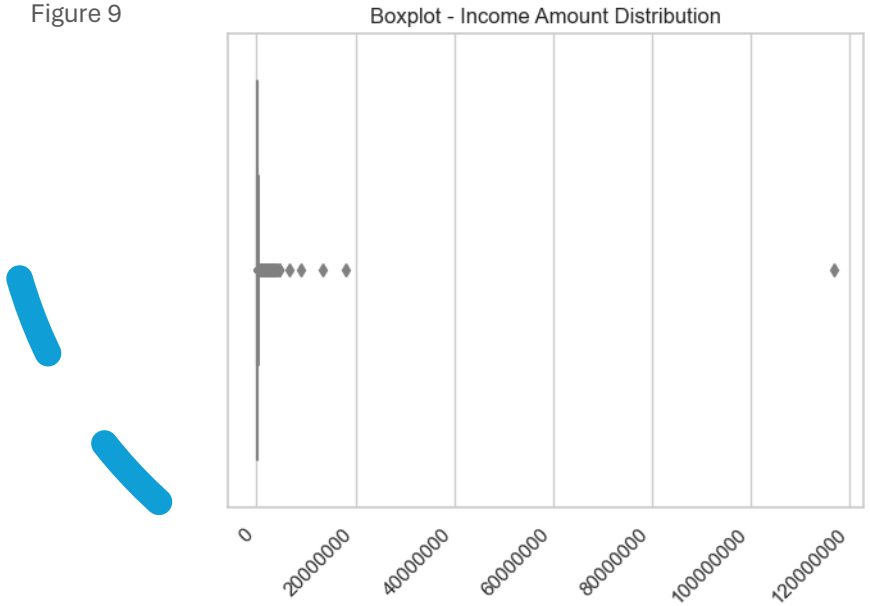


Figure 10

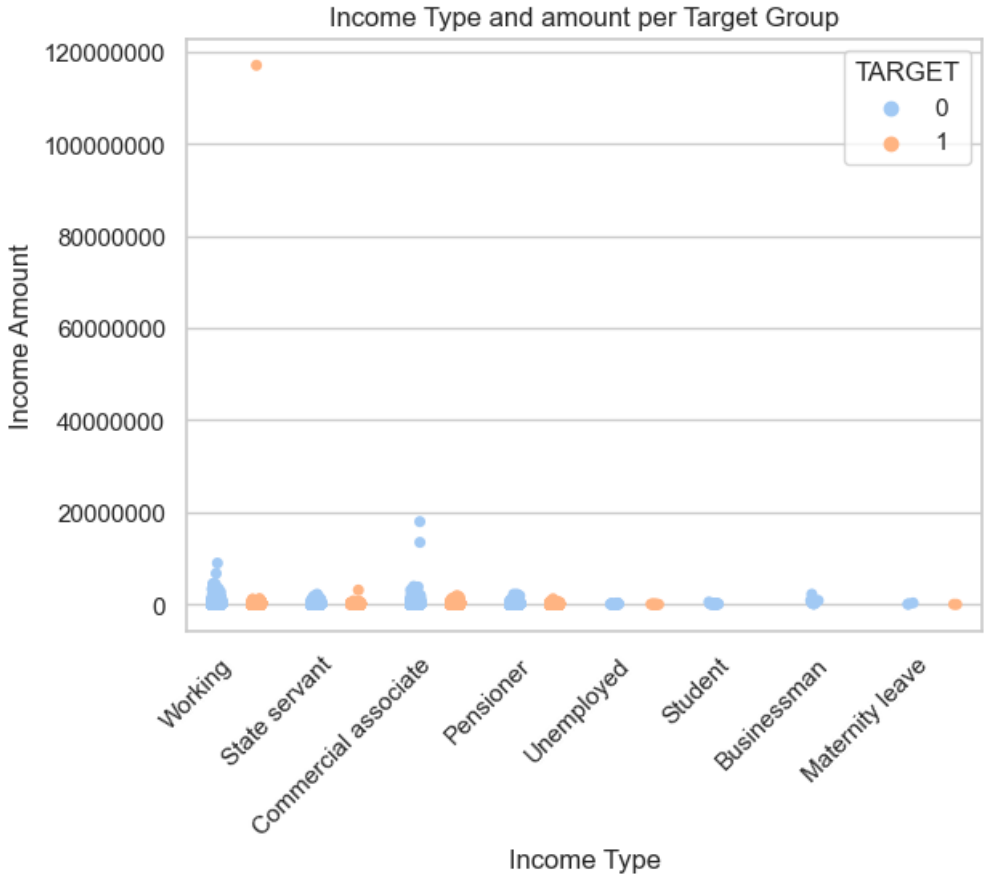


Figure 11

Income Type and amount for Target - Multivariate Analysis

- A Late Paid outlier is observed near the mark of 120,000,000 of Income Amount.
- Student and Businessman Income categories not appeared in Target Group 1.

Correlation Matrix: Income, Credit, Annuity and Price of the Good - Multivariate Analysis

- A strong positive correlation exists between Credit Amount and Price of Goods granted with a loan features.
- There is a good positive correlation between Credit Amount and Loan Annuity.
- Income presents a weak correlation with other features.

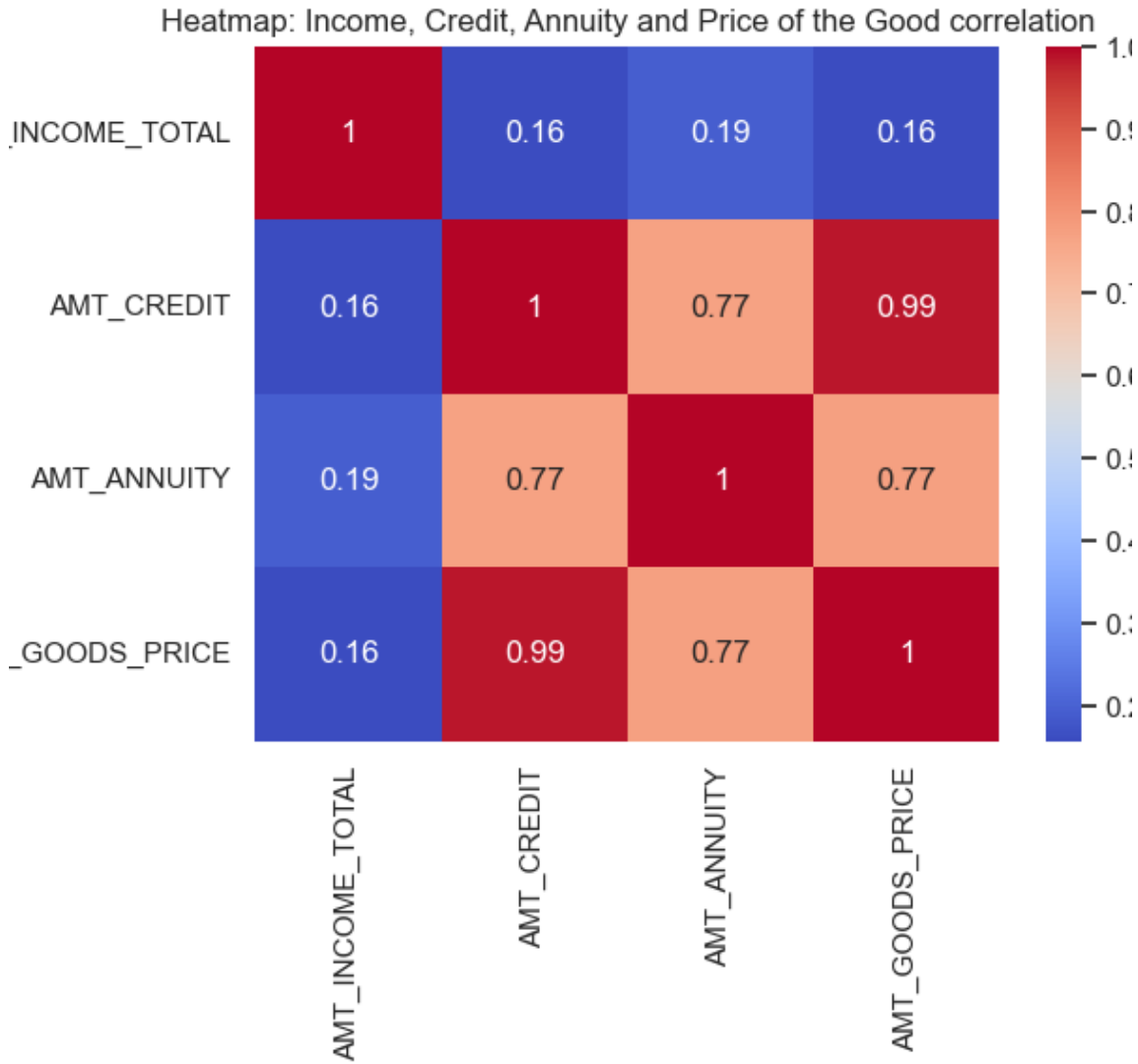


Figure 12

Pair Plot: Income, Credit, Annuity and Price of the Good correlation Multivariate Analysis

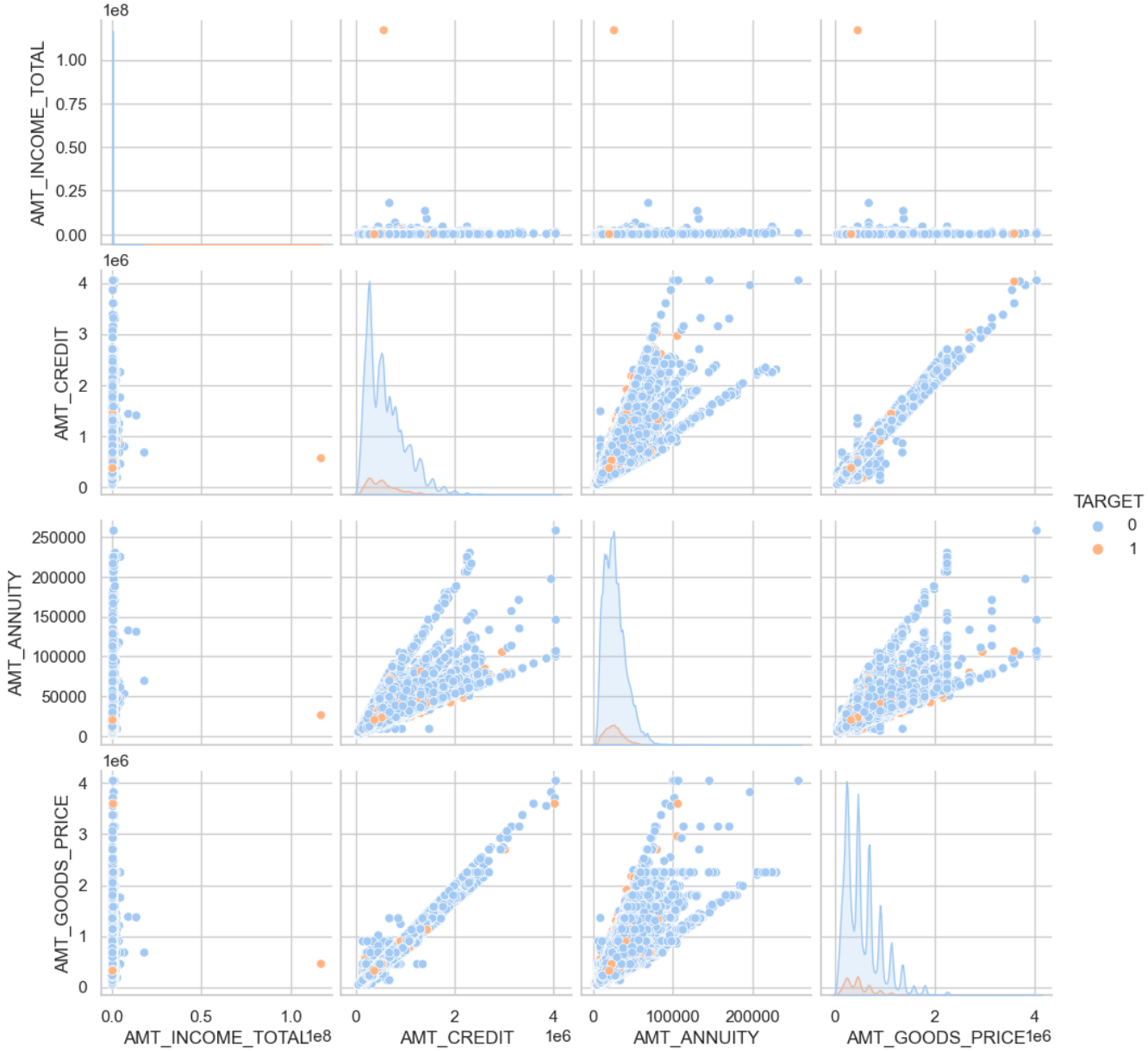
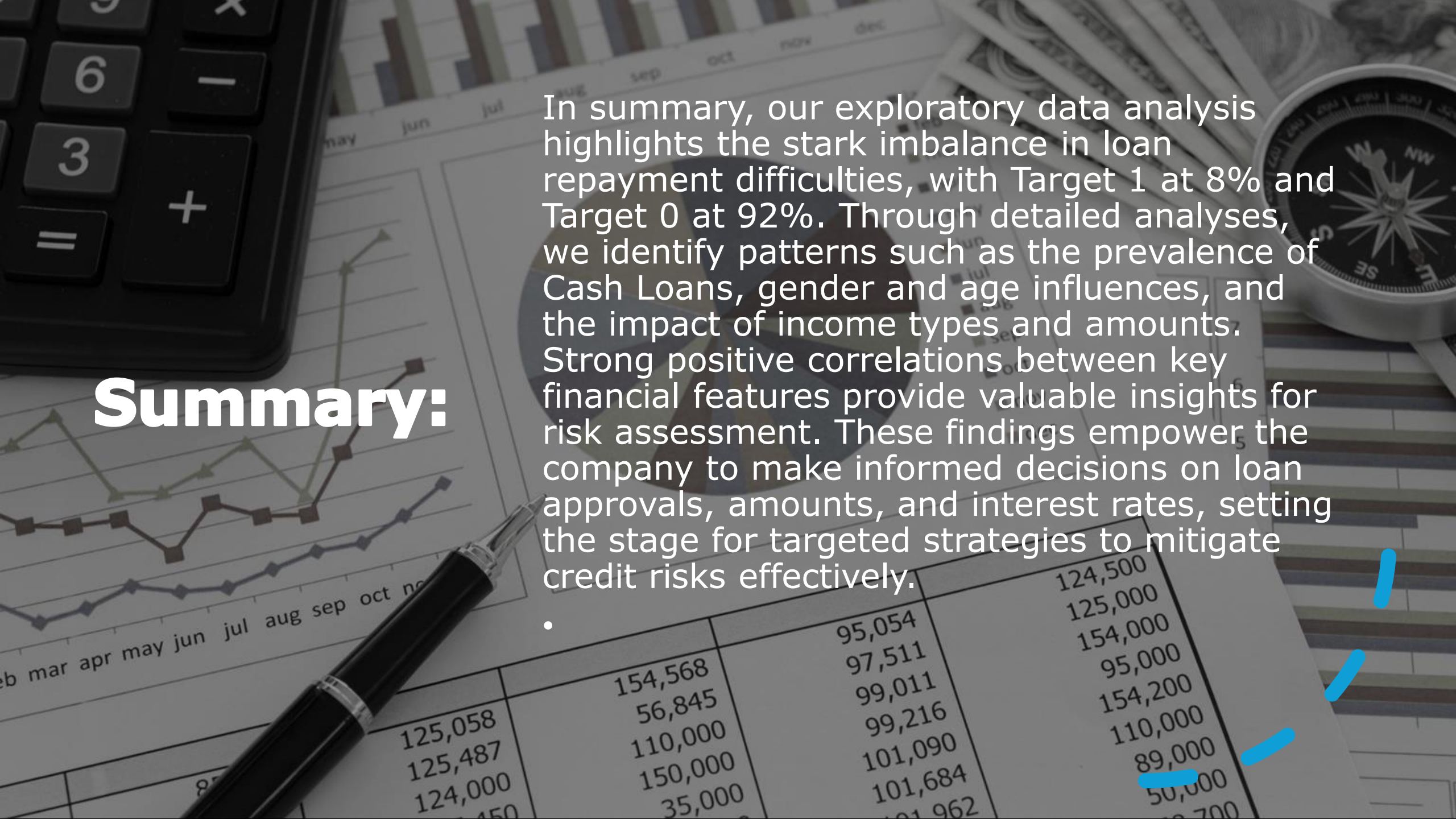


Figure 13



Summary:

In summary, our exploratory data analysis highlights the stark imbalance in loan repayment difficulties, with Target 1 at 8% and Target 0 at 92%. Through detailed analyses, we identify patterns such as the prevalence of Cash Loans, gender and age influences, and the impact of income types and amounts. Strong positive correlations between key financial features provide valuable insights for risk assessment. These findings empower the company to make informed decisions on loan approvals, amounts, and interest rates, setting the stage for targeted strategies to mitigate credit risks effectively.

.

List of Figures: