Credit Default Risk Analysis: Report

Prepared By:

Manu E Thomas

Table of Contents

Chapter 1

Introduction

This report presents a detailed summary of an exploratory data analysis (EDA) conducted on a loan application dataset by Home Credit Group. The primary focus of this analysis is to identify key factors that contribute to loan defaults, which is essential for developing effective risk mitigation strategies. The dataset, which consists of 122 columns, includes a variety of information about loan applicants, their loan applications, and their repayment behaviour. The analysis encompasses several critical steps, including data pre-processing, feature selection, feature engineering, and both univariate and bivariate analysis of key features. The target variable in the dataset indicates whether a loan was repaid (0) or defaulted (1). By examining these factors, the analysis aims to provide actionable insights for optimizing lending practices and reducing potential financial losses.

Chapter 2

Objectives

The main objectives of this exploratory data analysis are

- **Data Understanding:** To gain a comprehensive understanding of the structure, content, and quality of the loan application dataset, including the identification of missing values, outliers, and data distributions.
- **Risk Factor Identification:** To determine the key demographic, financial, and application-related factors that are significantly associated with loan defaults.
- **Univariate Analysis:** To examine the distribution of individual variables, especially the relationship between categorical features and the target variable, as well as the distributions of numerical features for both defaulters and repayers.
- **Bivariate Analysis:** To investigate the relationships between pairs of variables, with a particular focus on how these relationships correlate to loan defaults.
- **Feature Selection and Engineering:** To identify and select relevant features from the dataset, and create new features that may enhance the predictive power of any downstream models, while removing less informative or redundant data.
- **Actionable Insights:** To provide actionable insights and recommendations for the lender to improve risk assessment, target specific demographics, and refine loan offerings to optimize the lending portfolio.
- **Further Analysis:** To identify areas that require more in-depth investigation

Chapter 3

Approach

The main objectives of this exploratory data analysis are

- Data Preprocessing
- Univariate Analysis (categorical)
- Correlation Analysis
- Univariate Analysis (numerical)
- Bivariate Analysis
- Futher Analysis
- Key Takeaways and Recommendations

Chapter 4

Data Preprocessing

4.1 Importing Libraries & Loading Data

Several key libraries were imported to facilitate data analysis and visualization. The pandas library is used for data manipulation and handling structured data, while numpy provides support for numerical operations. For data visualization, matplotlib and seaborn are employed to create a wide range of plots and graphs, enhancing the presentation and interpretation of the data.

The CSV file **application_data.csv** is loaded using pandas.

4.2 Missing Value Handing

During data preprocessing, columns with over 40% missing values were dropped. Missing values in categorical columns, specifically *CNT_FAM_MEMBERS*, *OCCUPATION_TYPE*, and *NAME_TYPE_SUIT*, were filled using the mode. Numerical columns with missing values, *AMT_ANNUITY* and *AMT_GOODS_PRICE*, were imputed using the median. The missing values in *AMT_REQ_CREDIT_BUREAU* features were also imputed using the mode.

4.3 Outlier Treatment

For the time being, I have decided to retain the outliers and proceed with the exploratory data analysis (EDA). This approach will help identify any anomalies in the data and ensure that no unusual patterns, which warrant significant attention, are overlooked. By keeping the outliers, we can gain a comprehensive understanding of the dataset and make more informed decisions.

Chapter 5

Exploratory Data Analysis

5.1 Univariate Analysis (Categorical Variables)

Univariate analysis involves examining individual variables in a dataset to understand their distributions and characteristics. In the context of loan default risk, this type of analysis is crucial for identifying patterns and trends within single features that may correlate with the likelihood of a loan default. For categorical variables, this includes understanding the distribution of each category and its relation to the target variable. For example, the analysis looks at the different categories within features like NAME_CONTRACT_TYPE, CODE_GENDER, NAME_INCOME_TYPE, and others to see which categories have a higher default rate. For numerical features, the analysis focuses on understanding the central tendencies, such as the median, and the spread, such as the interquartile range, as well as identifying outliers, and how these characteristics vary between defaulters and non-defaulters.

5.1.1 NAME_CONTRACT_TYPE

Cash loans constitute about 90% of the total loans and have a default rate of 7.7 compared to revolving loans with a default rate of 5.

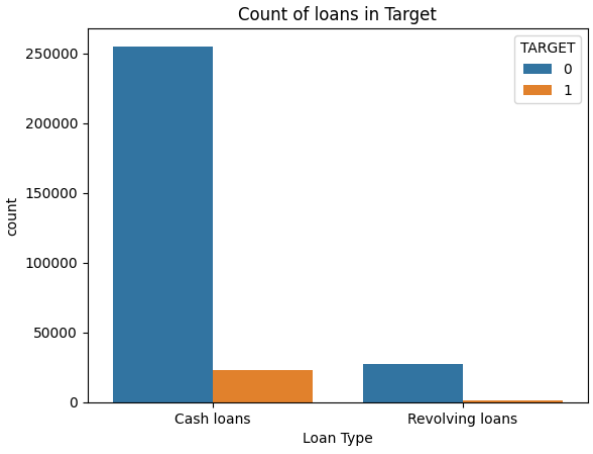
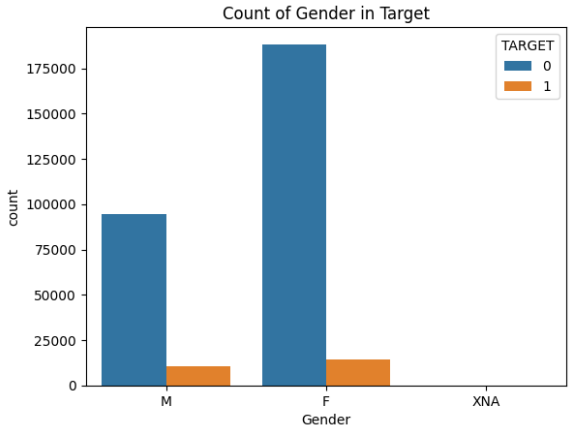


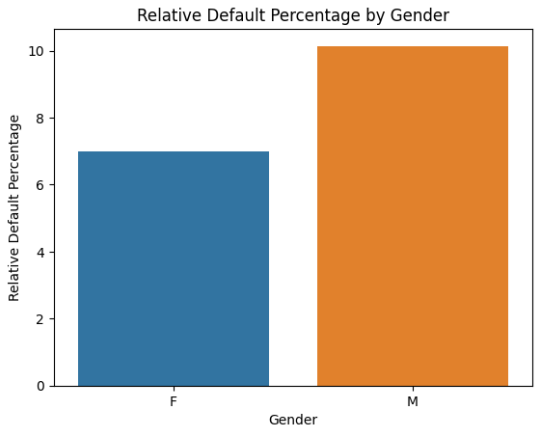
Figure 5.1: Count of Loan Types by Target

5.1.2 CODE_GENDER

65% of loans are taken by females, who are less likely to default than males.



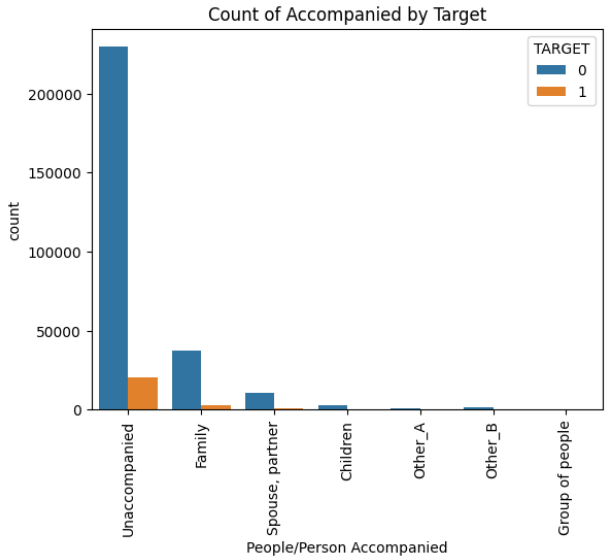
(a) Count of Gender by Target



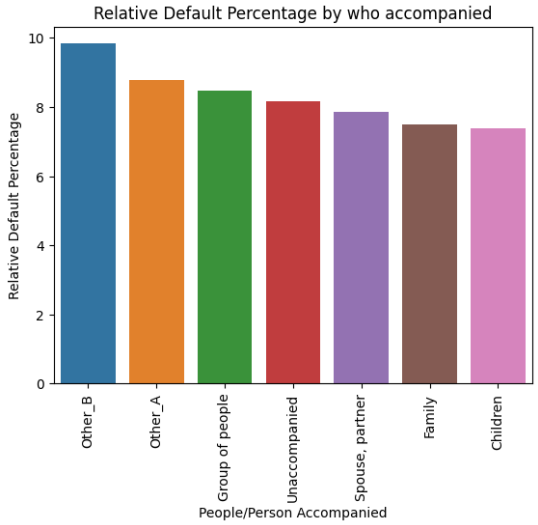
(b) Default Percentage by Gender

5.1.3 NAME_TYPE_SUITE

Most applicants apply individually, but those accompanied by ”Other_B” are more likely to default, while those with children are the safest. (The dataset does not give any information about ‘Other_B’ category).



(a) Count of Client Accompanied



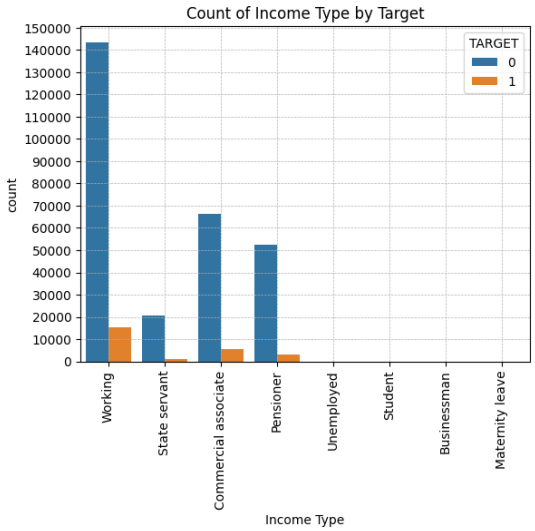
(b) Relative Default Percentage by Accompanied

5.1.4 NAME_INCOME_TYPE

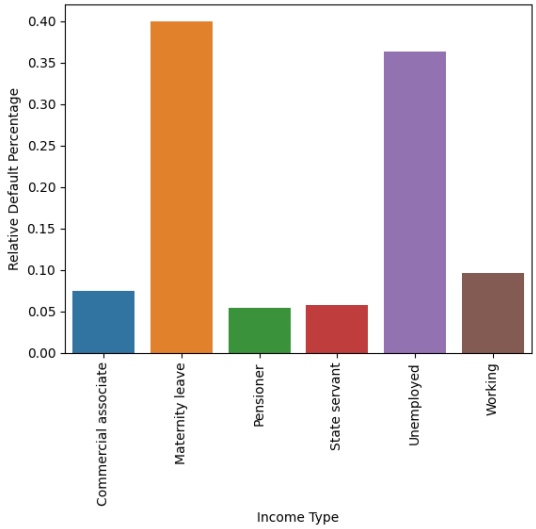
Most applicants are Working professionals and commercial associates. They are the most promising target groups due to their already significant representation and reliable repayment records.This makes them attractive for credit offerings with minimal resource allocation.

Mothers on maternity leave exhibit high default rates, likely due to financial strains such as increased healthcare costs and reduced income. Addressing these issues through enhanced credit mechanisms or social security measures could provide necessary support.

Loans to unemployed individuals require stringent scrutiny. Implementing targeted schemes and benefits to upskill the unemployed can empower them to secure employment and repay their debts, ultimately supporting their financial stability and contributing to economic growth.



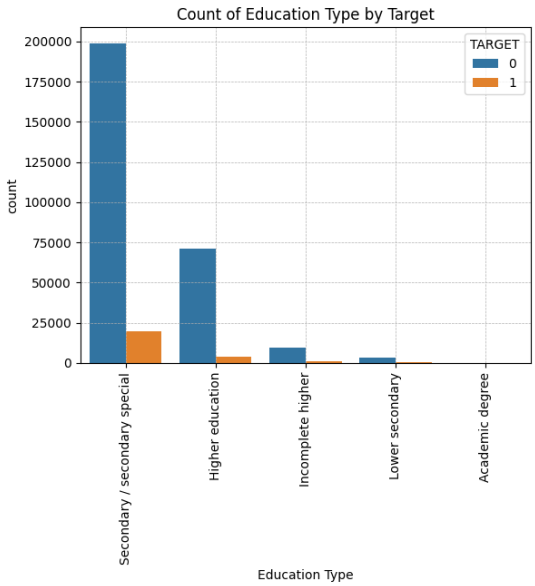
(a) Count of Income Type by Target



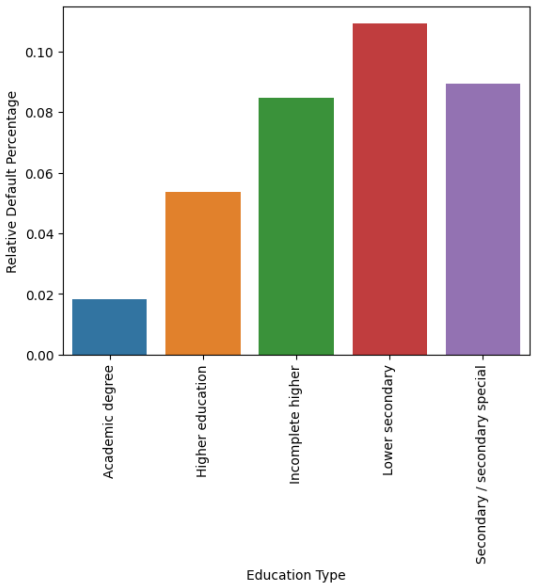
(b) Relative Default Percentage by Income Type

5.1.5 NAME_EDUCATION_TYPE

A significant share of loans is taken by applicants with secondary/secondary special education 71%, but those with lower secondary education exhibit the highest relative default rate. Applicants with higher education and academic degrees appear more promising.



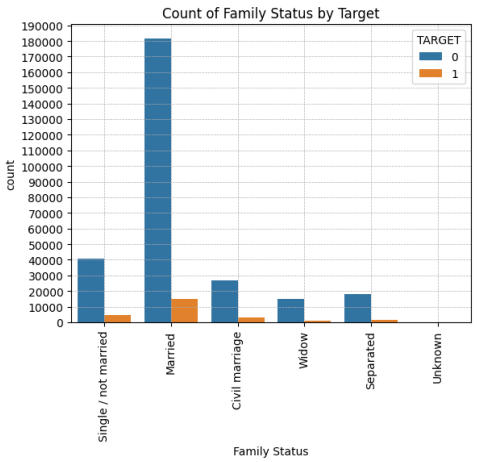
(a) Education Count by Target



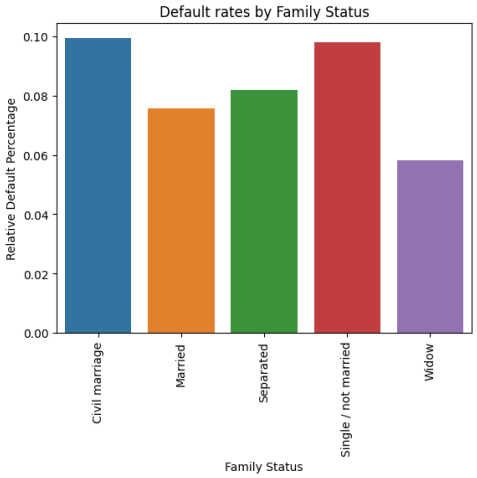
(b) Relative Default Percentage by Education Type

5.1.6 NAME_FAMILY_STATUS

The majority of loan applicants, constituting 64%, are married. This is followed by individuals who are either single or in a civil marriage. Although applicants in civil marriages and those who are single represent a smaller portion of the total applicants, they exhibit the highest default rates, nearing 10%. Married applicants, while not demonstrating the safest profile, have a relatively lower default rate compared to other categories, with the exception of widowed applicants. Notably, widowed applicants exhibit the lowest default rate among all groups.



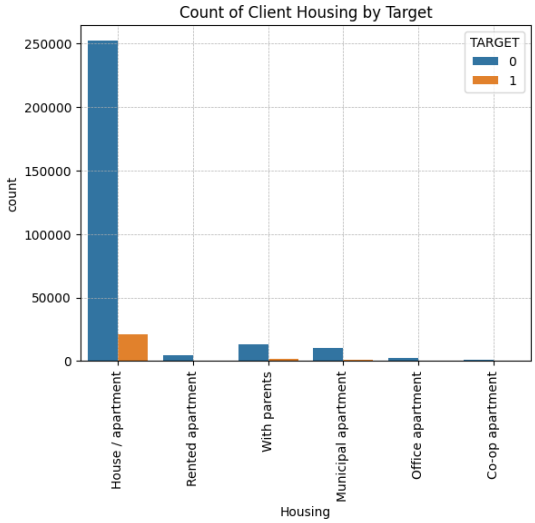
(a) Family Status count by Target



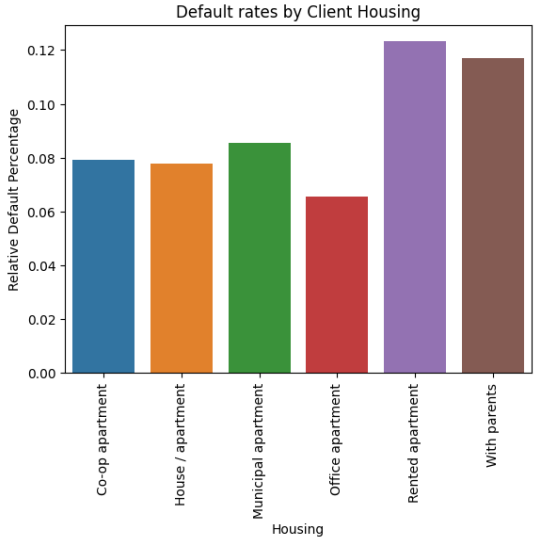
(b) Relative Default Percentage by Family Status

5.1.7 NAME_HOUSING_TYPE

Nearly 89% of applicants are homeowners, followed by those living with parents (5%), in municipal apartments (4%), and in rented apartments (1.2%). Applicants residing in rented apartments and those living with parents demonstrate a higher propensity for default. Among consumer clients, homeowners represent the most secure category. Commercial clients utilizing office apartments constitute a small share of the total applicants. Despite their lower representation, they exhibit the lowest default rates. Focusing on them could significantly enhance future lending strategies. These clients can benefit from more affordable loans and flexible repayment structures. Given the often long-standing relationships between these clients and the bank, this approach can secure a steady revenue stream



(a) Client Housing Count by Target

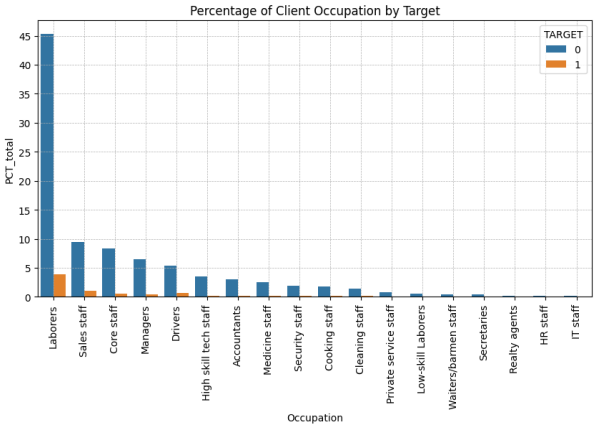


(b) Relative Default Percentage by Client Housing

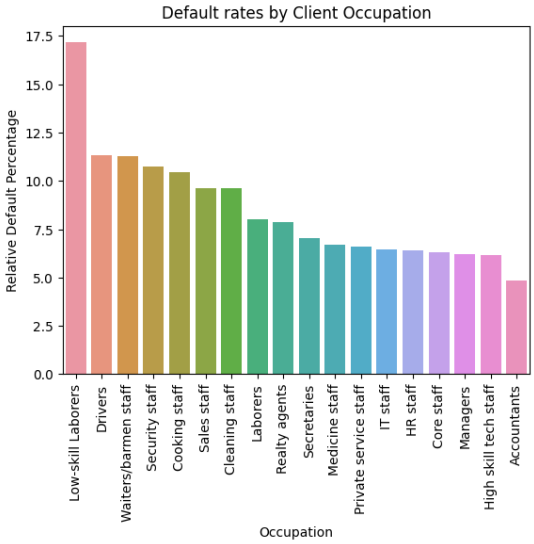
5.1.8 OCCUPATION_TYPE

Nearly 50% of the applicants are laborers, followed by sales staff (10%) and core staff (9%). Although Low-Skill Labourers represent less than 1% of the applicants, this group is more likely to default. Other occupations, with the exception of accountants, exhibit default rates ranging from 6.5% to 11%. IT staff, HR staff, core staff, managers, and high-skill tech staff are less likely to default, likely due to their higher income levels compared to other occupations. Accountants stand out as the safest target group, with the lowest default rates, constituting just 3% of total loans.

Overall the data suggests that occupations associated with higher incomes tend to have lower default rates.



(a) Percentage of client Occupation by Target



(b) Default Rates by Client Occupation

5.1.9 ORGANIZATION_TYPE

Business Entity Type 3, XNA, and Others constitute the largest share of applicants. Transport type 3, Industry type 3, and Industry type 8 are serious defaulters, while Trade Type 3 has the least default rate.

Not much information was obtained from this column. A more effective approach would be to classify these categories into subcategories such as Industry, Transport, Business, Medicine, etc. A sectoral analysis could provide more meaningful trends within each sector, offering better guidance for strategic decision-making.

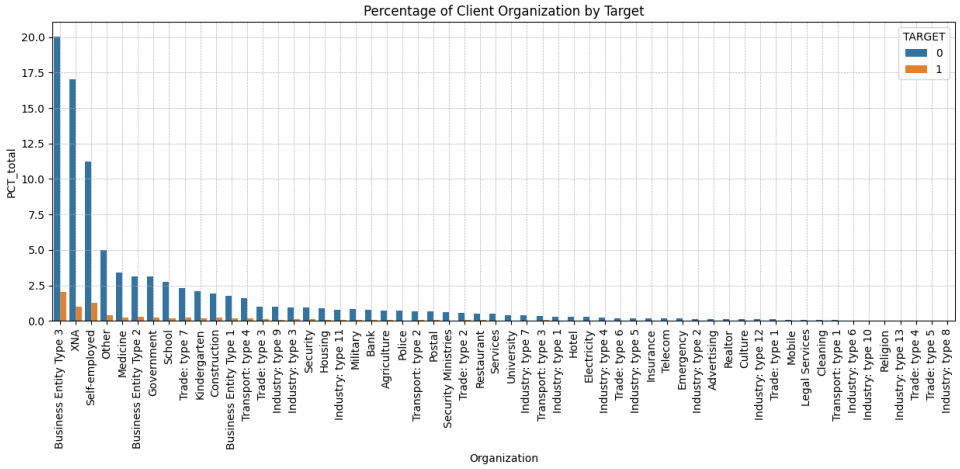


Figure 5.9: Percentage of Client Organization by Target

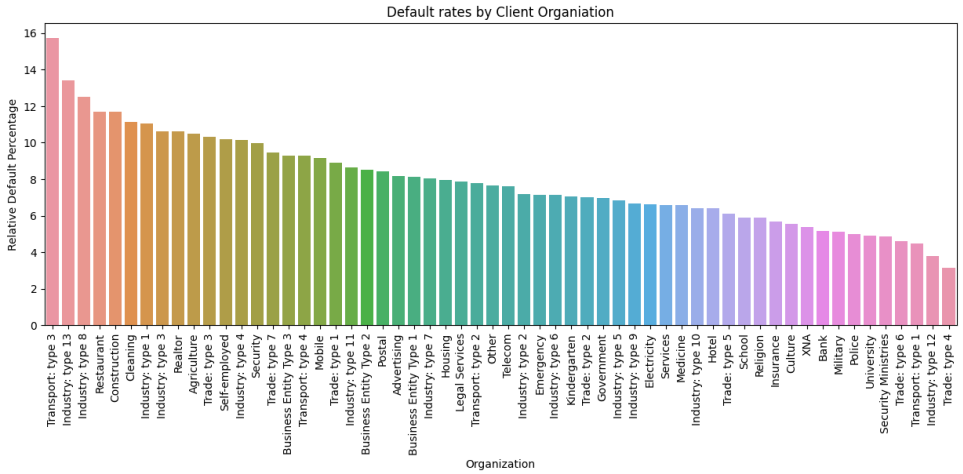


Figure 5.10: Relative Default Percentage of Client Organization by Target

5.2 Correlation Between Numberical Variables

Upon inspecting the dataset, it is evident that there is a significant **class imbalance**, with repayers comprising 92% of the data and defaulters making up only 8%. This imbalance can potentially skew the analysis.

To conduct a more meaningful and robust analysis, we should address this imbalance. One effective approach is to split the dataframe into two separate subsets: defaulters and non-defaulters. This stratification will enable us to analyze each group independently, understand their unique characteristics, and develop strategies tailored to each segment.

5.2.1 Correlation Analysis

Correlation analysis on numerical variables involves examining the statistical relationships between pairs of numerical features in a dataset. The goal is to identify how these variables change in relation to each other. This is different from univariate analysis, which focuses on the characteristics of single variables. In the context of loan default risk, understanding correlations between numerical features can reveal patterns that may not be apparent when looking at each variable in isolation.

Correlation was measured using Pearson's method. The analysis identified several pairs of numerical variables with high correlations. It is interesting to note that both the defaulters and non-defaulters showed similar correlation results

- **OBS_60_CNT_SOCIAL_CIRCLE** and **OBS_30_CNT_SOCIAL_CIRCLE** show a very high positive correlation. These variables count the number of observations of a client's social circle with 60 and 30 days past due, respectively.

- **AMT_GOODS_PRICE** and **AMT_CREDIT** are highly positively correlated. This indicates that the price of the goods and the amount of credit extended are very closely related.

- **REGION_RATING_CLIENT_W_CITY** and **REGION_RATING_CLIENT** also show a strong positive correlation. This indicates that the client's region rating with and without city are closely related

- **CNT_FAM_MEMBERS** and **CNT_CHILDREN** are also highly correlated. This indicates that the number of family members and the number of children are related as would be expected.

5.2.2 Flag Columns

The columns starting with *"FLAG_"* were identified as binary indicators. Analysis revealed that few documents other than *FLAG_DOCUMENT_3* were commonly submitted. A correlation analysis on a subset of these flag columns showed very low correlation values among each other. Consequently, all *FLAG* columns, along with *EXT_SOURCE_2* and *EXT_SOURCE_3*, can be dropped due to limited predictive power.

5.3 Univariate Analysis (Numerical Variables)

For numerical features, the analysis focuses on understanding the central tendencies, such as the median, and the spread, such as the interquartile range, as well as identifying outliers, and how these characteristics vary between defaulters and non-defaulters.

5.3.1 AMT_CREDIT

- The **median credit** amount for both defaulters and non-defaulters is approximately **0.5 million**, indicating a comparable central tendency in credit amounts across both groups.

- The **Interquartile Range** (IQR), representing the middle 50% of the data, is also similar for both categories, suggesting a comparable spread of credit amounts within the middle range.

- Both categories exhibit a significant number of **outliers**, with credit amounts extending up to approximately 4.05 million. This indicates the presence of individuals in both groups with exceptionally high credit amounts.

- The smaller range of the **lower whisker** suggests that the lower range of credit values is relatively small, while the larger range of the upper whisker indicates a higher spread due to the presence of outliers.

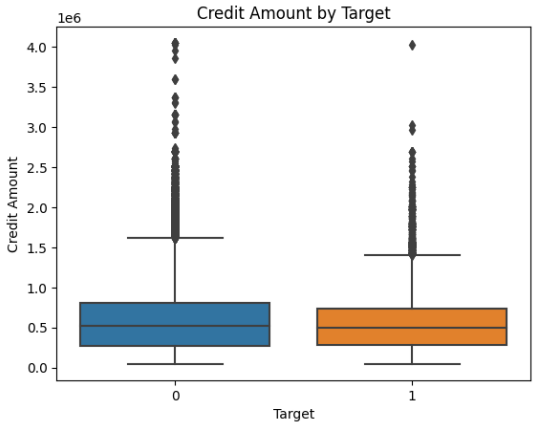


Figure 5.11: Distribution of Credit Amount by Target

Overall, the distribution of credit amounts is quite similar between those who repaid their loans on time and those who defaulted. This suggests that credit amount alone may not be a strong differentiator between these two groups.

5.3.2 AMT_INCOME_TOTAL

The violin plot analysis indicates that median income for both defaulters and non-defaulters is comparable, approximating to 150,000. The majority of clients have incomes in the range of 0.1 million to 0.25 million.

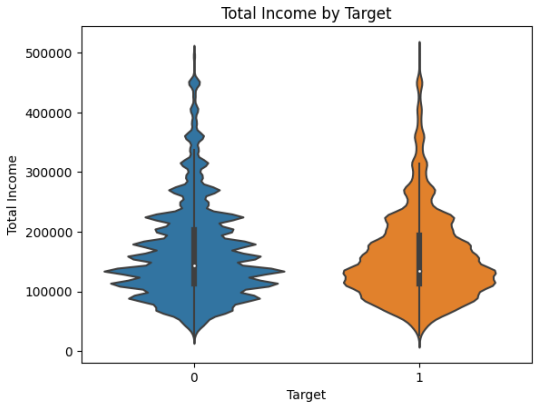


Figure 5.12: Distribution of Total Income by Target

5.3.3 AMT_ANNUITY

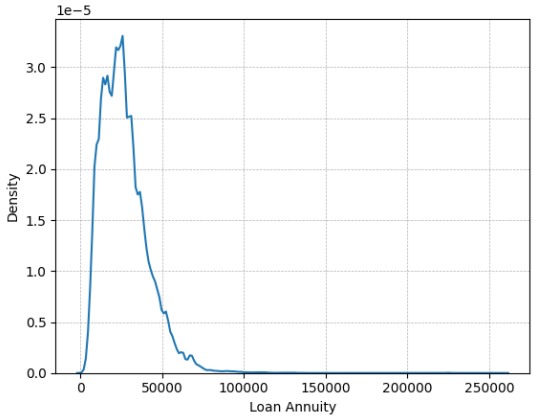
The KDE plot analysis reveals

- The majority of annuity amounts are concentrated between **0 and 100,000, with a peak around 30,000**.
- The **strong right skewness** indicates that most data points are lower annuity amounts, with fewer instances of higher annuity amounts.

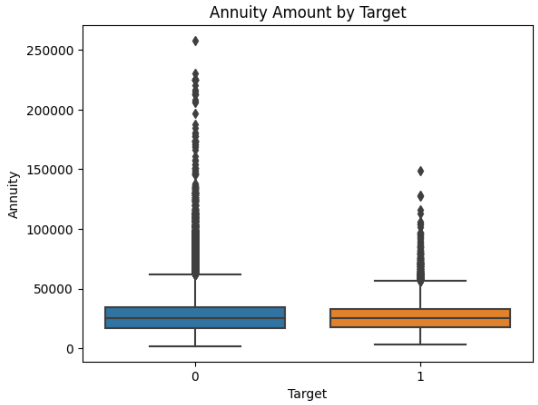
- Beyond 50,000, the density sharply decreases and continues to taper off, suggesting that **higher loan annuity amounts are less common**.

- Descriptive statistics reveal that the **mean and median values are similar**.

The box plot analysis indicates that the spread of outliers is larger for those who repaid their loans on time compared to defaulters.



(a) Distribution of Annuity Amount



(b) Distribution of Annuity Amount by Target

5.3.4 AMT_GOODS_PRICE

- Most asset valuations range between **0.25 million to 0.7 million**.

- The **median asset value** is around **0.5 million** for both defaulters and non-defaulters.

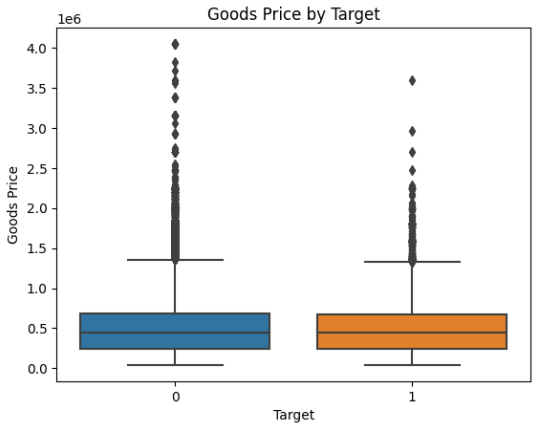


Figure 5.14: Distribution of Asset Valuations by Target

5.4 Bivariate Analysis

5.4.1 AMT_CREDIT and AMT_GOODS_PRICE

Both features are highly positively correlated. Most defaulters have credit amounts less than 1.5 million, while for non-defaulters, the credit amounts are typically under 2.5 million.

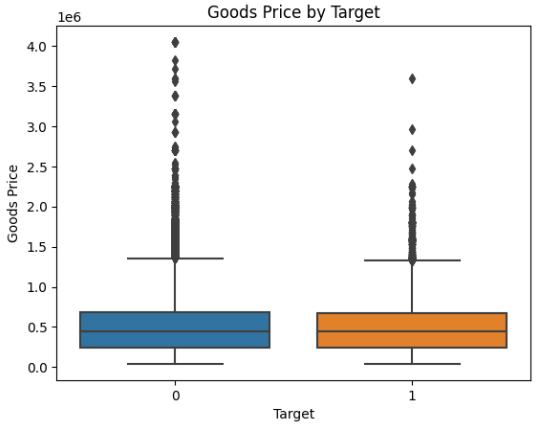
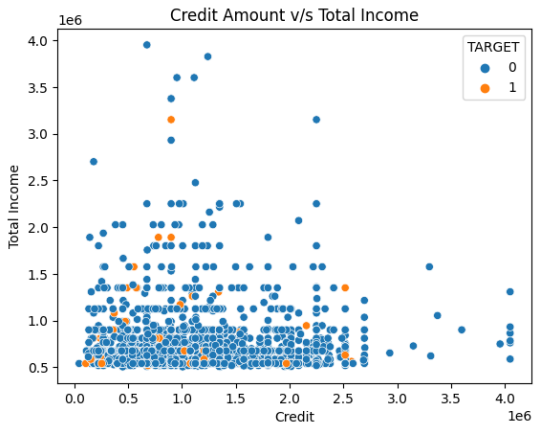


Figure 5.15: Credit v.s Goods Price

5.4.2 AMT_CREDIT v/s AMT_INCOME_TOTAL

- There is no clear relationship between total income and credit values.
- **Most defaults** occur for loan amounts under **1.5 million**.
- It appears **safer to target** customers with a **total income greater than 150,000 and credit amounts between 2 million and 2.5 million**.

(a) Credit Amount v.s Total Income



(b) Credit Amount v.s Total Income (outliers)

While a direct relationship between total income and credit values cannot be clearly established, focusing on customers with higher total incomes and specific credit ranges may help mitigate default risks and improve loan portfolio quality.

5.4.3 AMT_CREDIT v/s CNT_CHILDREN

- The majority of applicants have no children (70%), followed by those with one child (20%) and two children (9%).
- The variability in credit values for the top three groups (0, 1, and 2 children) is minimal, with average credit values around 600,000
- There is negligible difference in loan repayment or default rates among the top three groups

So we can conclude the number of children does not significantly impact the variability in credit values or the likelihood of loan repayment versus default for the majority of applicants

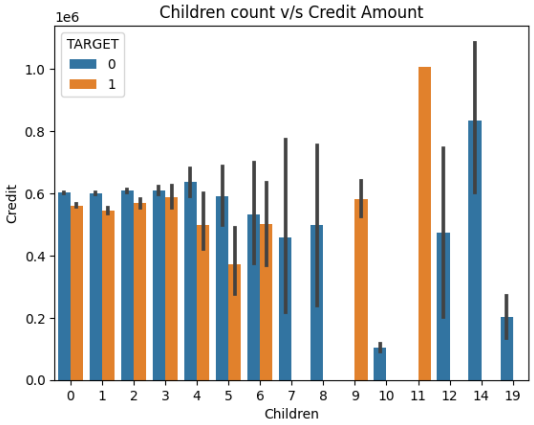


Figure 5.17: Credit v.s Count of Children

5.5 Scope for Further Analysis

- A detailed analysis of the 'ORGANIZATION_TYPE' column canb be done by classifying categories into subcategories such as industry, transport, business, and medicine to identify trends within each sector.
- Despite low correlations, *FLAG_* columns may offer valuable insights when combined with other features.

Chapter 6

Key Takeaways and Recommendations

In this analysis I have identified specific demographics and financial factors that significantly impact loan repayment, enabling the bank to make more informed decisions and tailor the offerings.

6.1 Targeting

- **Low-Risk Groups:** Widows, married individuals, those with higher education, accountants, commercial clients with office apartments, and clients who own a house or apartment.
- **High-Risk Groups:** Unemployed individuals, mothers on maternity leave, those with lower secondary education, people in civil marriages, single people, people who rent, low-skill laborers, clients who live with parents, clients from transport, industry types 3 and 8.

6.1.1 Recommendation:

- Offer smaller, flexible loans for high-risk groups and more lucrative offers for low-risk groups, enhancing customer satisfaction and maximizing profit.
- Provide more affordable loans and flexible repayment structures to commercial clients to secure a steady revenue stream and build loyalty
- Offer schemes and benefits to unemployed individuals aimed at upskilling them for employment and debt repayment

6.2 Policy Adjustments for Risk Mitigation

It is necessary to conduct a thorough review of existing policies regarding clients on maternity leave, addressing financial strains through enhanced credit mechanisms or social security measures to reduce default rates

6.3 Opportunities for Growth

- The current portfolio reveals a low share of commercial clients in the bank's total loan allocations. The data shows that commercial clients are reliable and present a promising target group. Diversifying revenue streams by increasing focus on commercial clients will mitigate the risk of over-reliance on consumer clients, ensuring long-term stability and consistent returns.
- Among consumer clients, those owning houses/apartments represent the safest category. Their significant presence in the business facilitates targeted outreach with minimal resource allocation

6.4 Safe Credit Range

With the current data that we have, It appears safer to target customers with a total **income greater than 150,000 and credit amounts between 2,000,000 and 2,500,000**.

By adopting these recommendations, we can significantly enhance our lending strategies, minimize default risks, and ensure sustainable business growth. This approach will also enhance customer relationships through tailored solutions, optimizing resource allocation and strengthening our position in the market. These measures ensure that the bank not only increases its profitability but also minimizes its losses due to loan defaults.