EDA ASSIGNMENT

Basic understanding of risk analytics in banking and financial services

by Abhradeep Chandra Paul

Introduction

This assignment aims to give you an idea of applying **EDA** in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develope a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the **driving factors** (or driver variables) behind **loan default**, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop the understanding of the domain, It is advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Why Risk Analysis is In Banking is Important?

1

Financial Stability and Solvency

Effective risk analysis helps banks identify potential risks that could lead to significant financial losses, such as credit defaults, market fluctuations, or operational failures.

2

Regulatory Compliance

Banks operate under stringent regulatory frameworks that require them to manage various risks effectively. Risk analysis helps in complying with these regulations, such as anti-money laundering (AML) laws, Know Your Customer (KYC) requirements, and other financial standards.

3

Operational Efficiency

Risk analysis helps banks identify potential operational risks, such as system failures, fraud, or process inefficiencies. By addressing these risks proactively, banks can enhance their operational efficiency and reduce the likelihood of disruptions.

4

Protecting Stakeholder Interests

By managing risks effectively, banks can protect customers' deposits and investments, fostering trust and loyalty. This is crucial for maintaining a stable customer base and ensuring long-term business success.

Common Bank Loan Challenges

Credit Risk Assessment and Monitoring

One of the most significant problems in loan management is the accurate assessment of borrowers' creditworthiness.

Operational Inefficiencies

Many banks still rely on manual processes for loan origination, processing, and management, which can be time-consuming and prone to errors. Ensuring that all loan documentation is complete, accurate, and compliant with regulatory standards is a significant challenge.

Solutions to Overcome Challenges

Implement advanced AI and machine learning models, automate loan origination and processing, use compliance management software to stay updated with evolving regulations and ensure accurate and timely documentation and reporting.

Risk Analysis Steps

1) At the start, I made sure to import all the important **libraries** needed for the analysis.

After that, I read in the two data sets that were provided - **'application_data.csv'** (categorized as **'app_inp0df'**) and **'previous_application.csv'** (categorized as **'prev_app_inp1df'**).

I then quickly checked the shape of the datasets and the data types of the features using the **.info('all')** method.

Next, I looked into the datasets to identify any columns with missing values.After analyzing i saw the **'app_inp0df'** data sets had 49 columns with over **40%** missing values. I checked correlation "EXT_SOURCE_1","EXT_SOURCE_2","EXT_SOURCE_3".

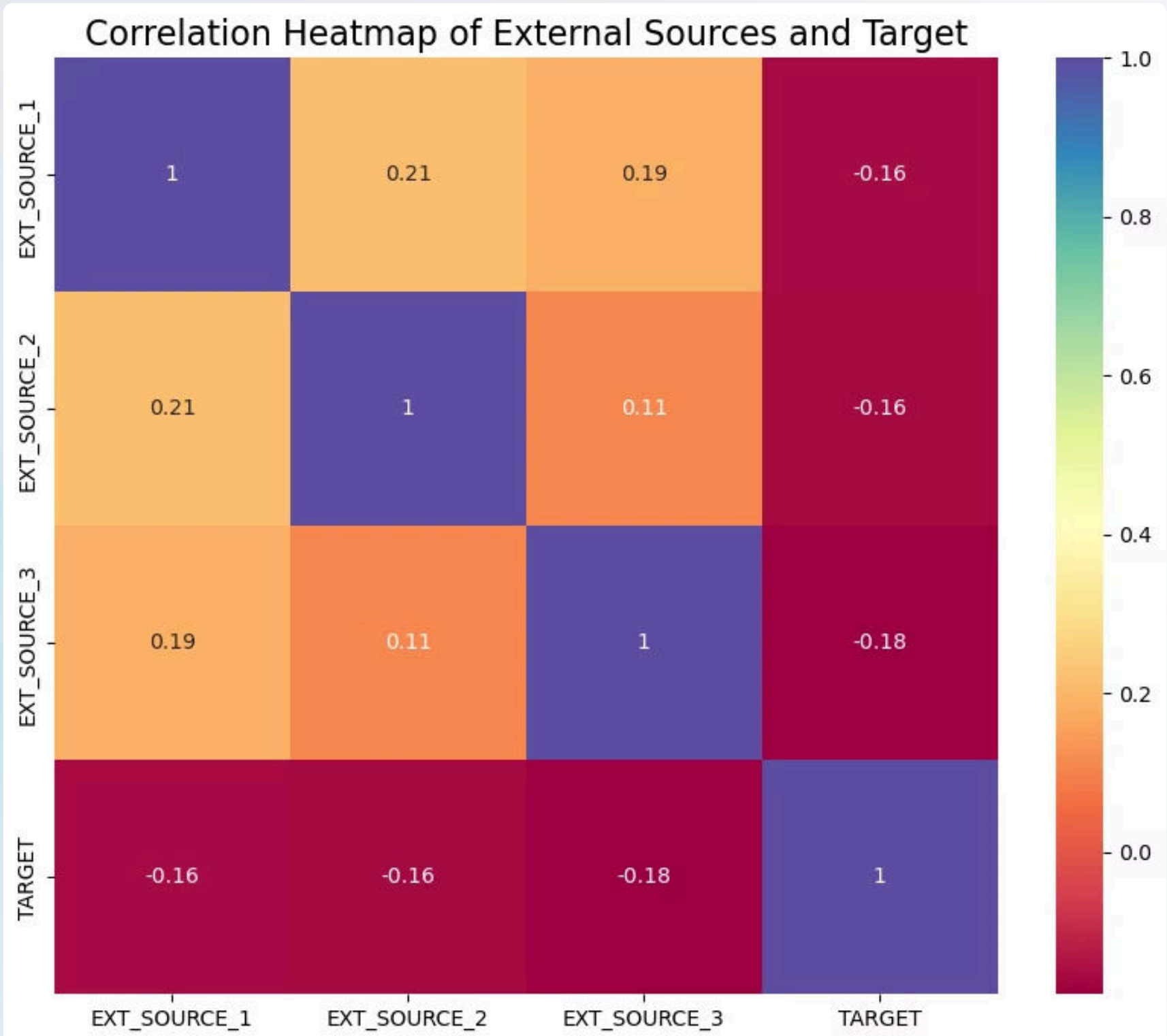


Figure: 1.0

And found out that there is almost no correlation between **EXT_SOURCE** columns and **target** column, thus we can drop these columns. Thus i **dropped** the column that had missing values more than 40%,that are related to different area sizes on apartment owned/rented by the loan applicants.

2) Next, I analyzed the remaining 73 columns 'of app_inp0df' data sets, and created a new list for documents needed for loan called **'col_doc'**. I used a **count plot** on that list to analyze the distribution of document submissions.

Figure: 1.1

After analyzing the count plots, I found that in most loan application cases, clients who applied for loans did not submit 'FLAG_DOCUMENT_X' except for **'FLAG_DOCUMENT_3'**. If the borrower has submitted **'FLAG_DOCUMENT_3'**, then there is a lower chance of them defaulting on the loan.

Based on this insight, except for FLAG_DOCUMENT_3, I decided to drop the rest of the columns in the 'col_doc' list. This will help streamline the dataset and focus the analysis on the most relevant document submission data.

Then I created a new variable called **'contact_column'** and stored the columns 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL','TARGET' and checked if there is any correlation between mobile phone, work phone etc, email, Family members and Region rating.

After analyzing the heat map i found out that there is no correlation between flags of mobile phone, email etc with loan repayment; thus these columns got dropped.

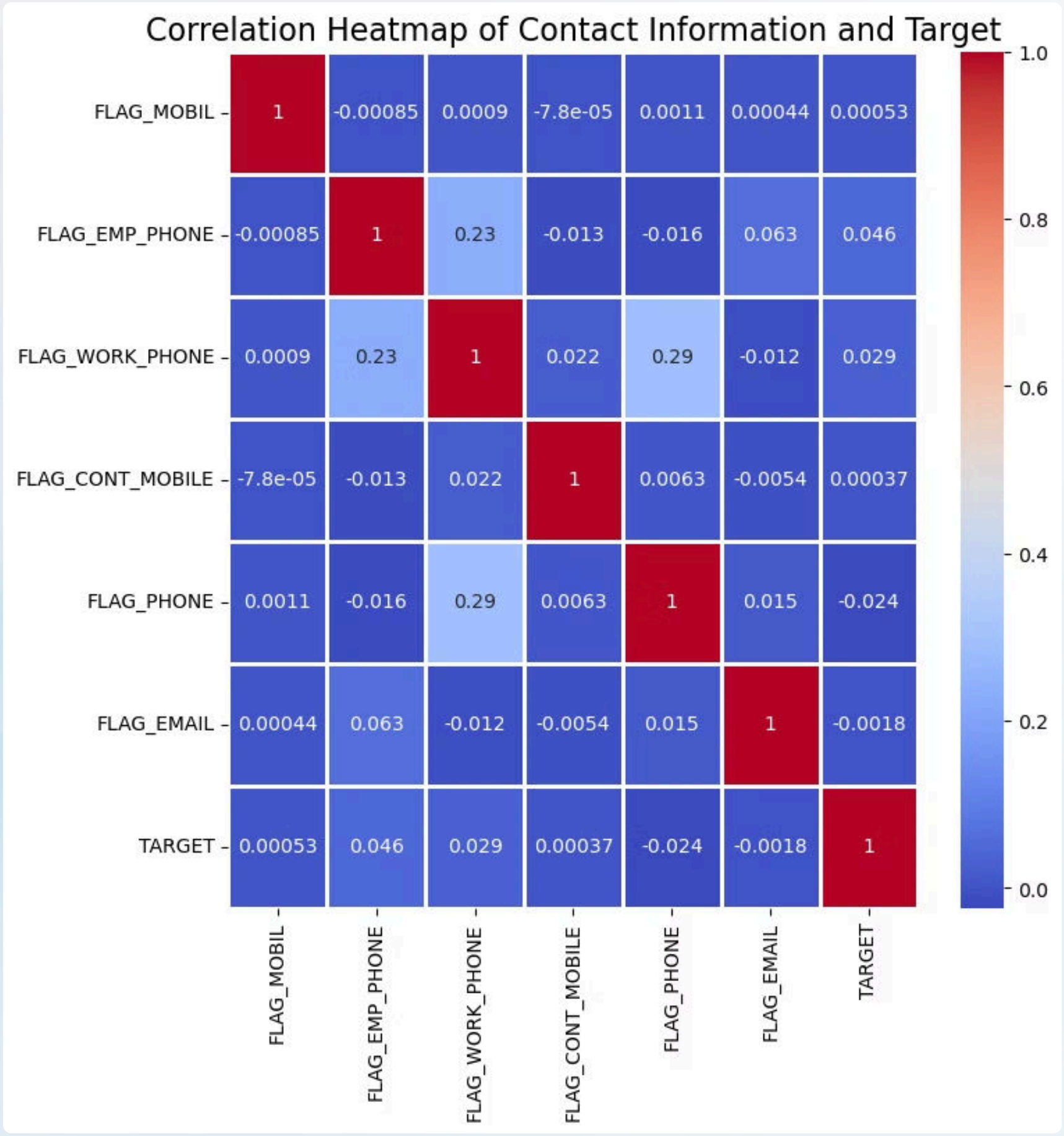
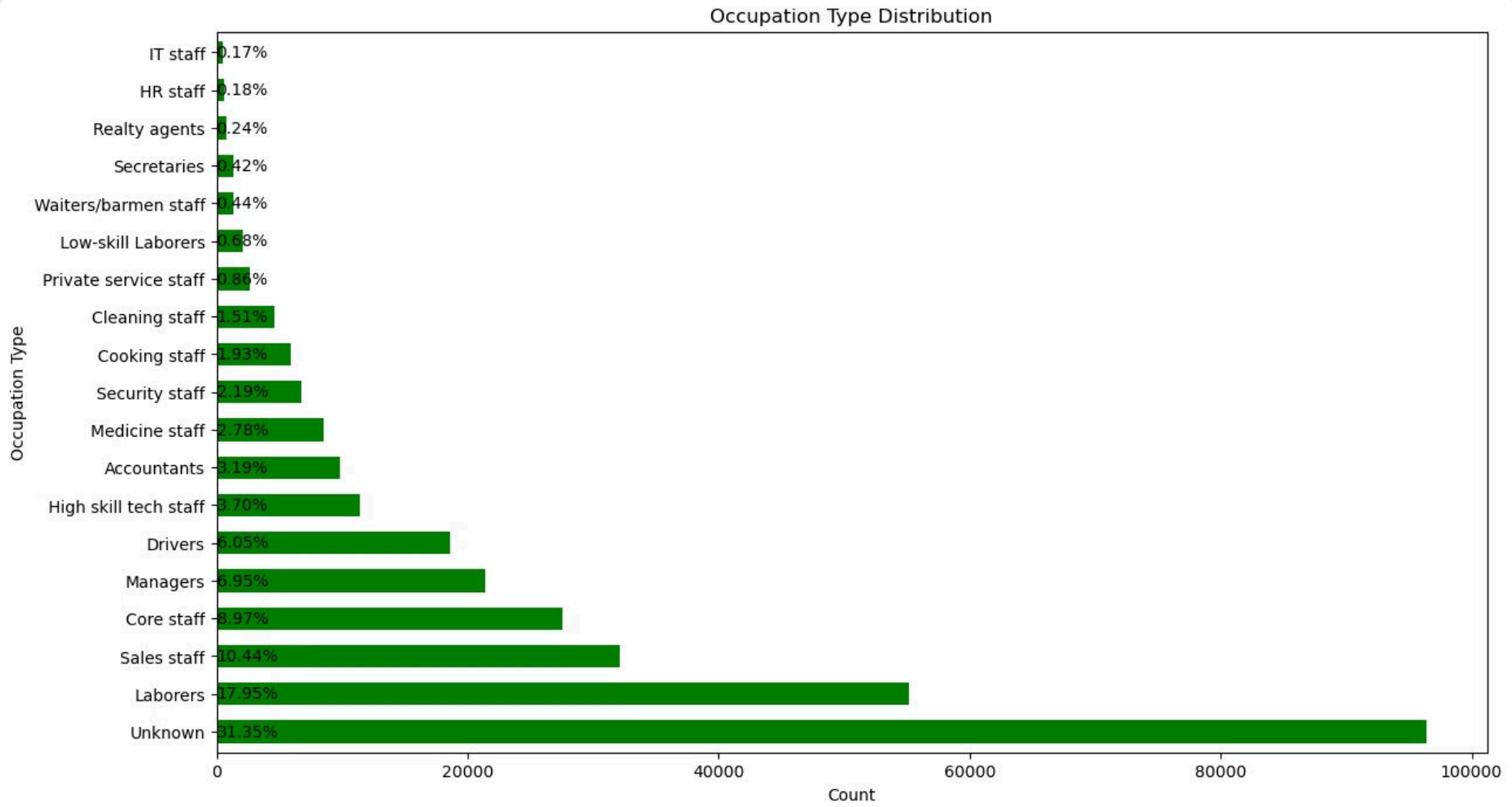


Figure: 1.2

Therefore I created a new variable called **'unwanted_col'** and stored the columns that's not needed in the analysis. Then I dropped those **'unwanted_col'** columns from the dataset.

3)After that, I started imputing missing values. I **imputed** 'Unknown' for the missing values of the 'OCCUPATION_TYPE' column.

Then, I used a **horizontal bar plot** on the **"OCCUPATION_TYPE"** column to analyze it further. This visualization helped me gain more insights into the distribution of occupation types in the dataset.



Figure;2.0

After analyzing the data, I found that the highest percentage of values belong to the **"Unknown"** group at **26.14%**, while the lowest percentage belongs to **"IT staff"** at **0.25%**. I then **imputed** gender 'F' for the 'XNA' values since they were very low and female is the majority.

Then I created a new column called **'numerical_columns'**, to convert columns to numerical values.

Standardize Values for app_inp0df

Next, I **standardized** the values of columns like 'DAYS_BIRTH', 'DAYS_EMPLOYED', etc. which had negative values, (bacause, days can not have negative values) by taking the absolute value using the abs() function.

4) Next, I standardized the values of columns like 'DAYS_BIRTH', 'DAYS_EMPLOYED', etc. which had negative values, by taking the absolute value using the abs() function.

I also created a new column **'Employment_years'** from the 'Days Employed' column and used a bar plot to analyze the distribution of employee tenure. The insights showed a significant majority of employees have a short tenure of 0-5 years, with very few staying beyond 25 years.

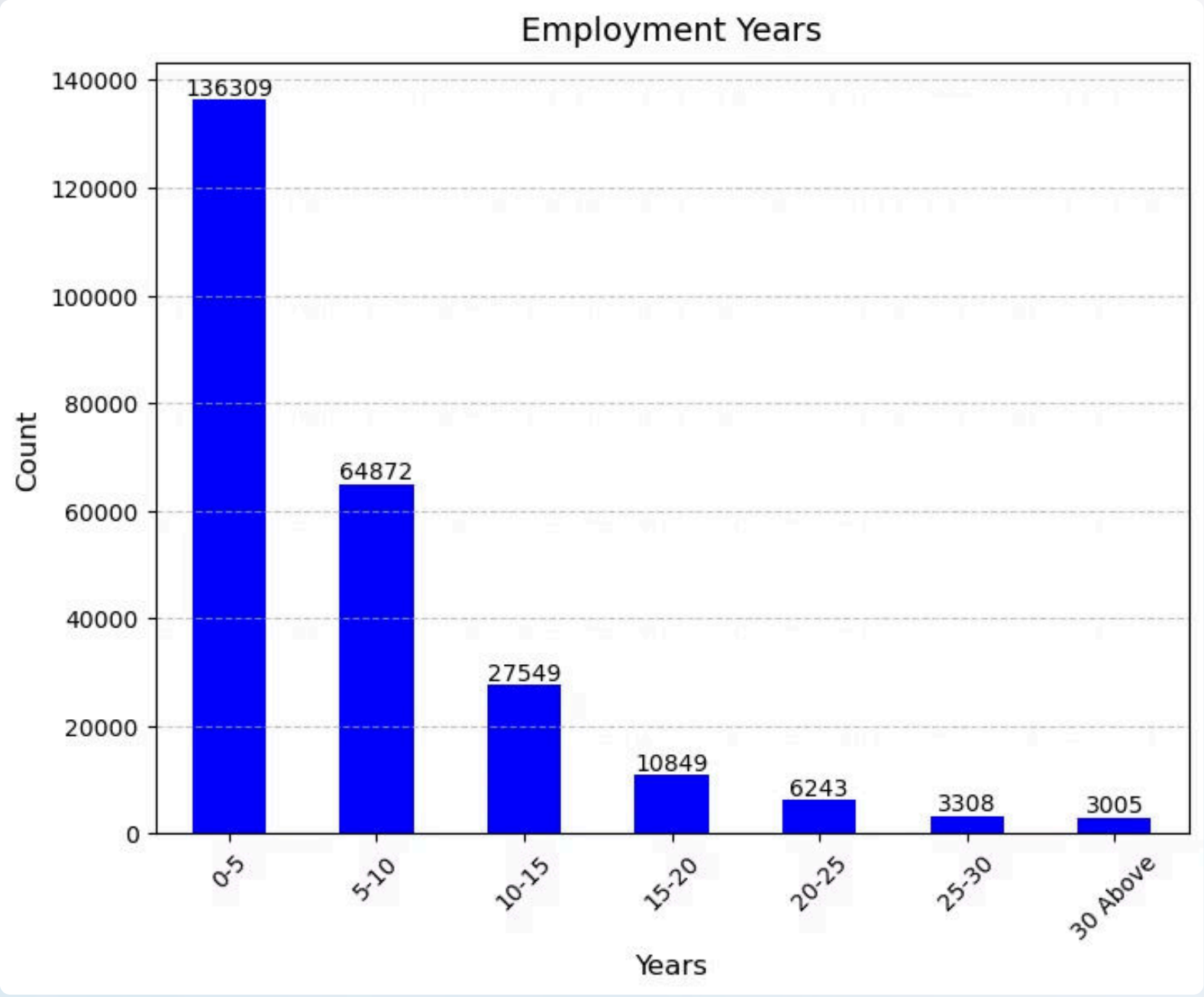


Figure: 3.0

The analysis showed that a significant majority of employees have a short tenure, with about **136,309** individuals having 0-5 years of employment. Very few stay beyond 25 years, with only **3,308** in the 25-30 range and 3,005 in 30+.

I then converted 'DAYS_BIRTH' and 'DAYS_EMPLOYED' to years and binned the values for better understanding.

5) Checking Outliers for app_inp0df data set

Next, I identified potential outliers by checking the columns with high differences between the max and 75th percentile using the describe() function. I created a list of these columns and analyzed them using box plots.

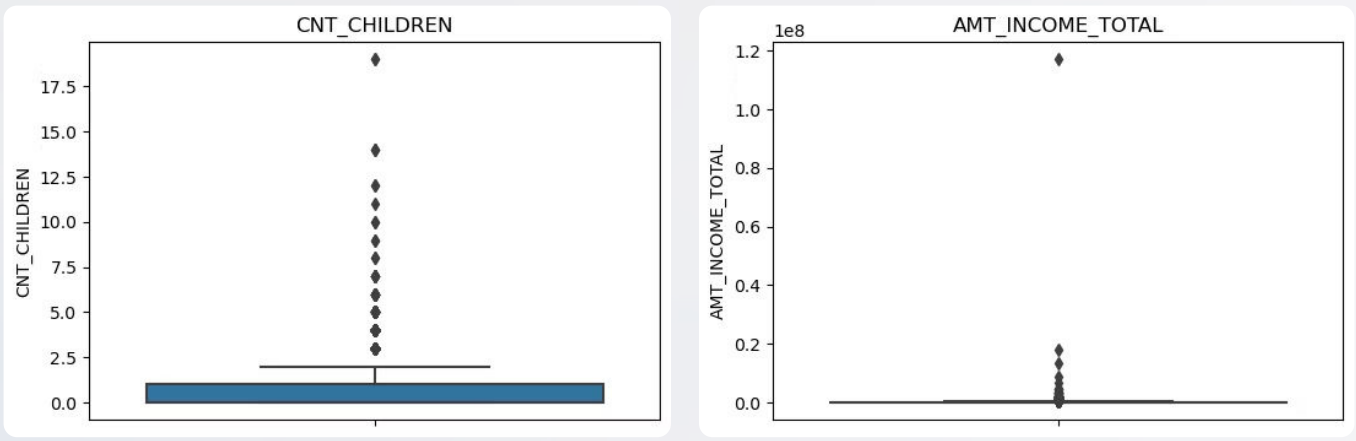


Figure: 4.0 - 4.7

After analyzing the data, I discovered a few key insights. First, I found that the features **'CNT_CHILDREN', 'AMT_ANNUITY', 'AMT_CREDIT',** and **'AMT_GOODS_PRICE'** all contained **outliers** that need further investigation.

Additionally, the 'DAYS_EMPLOYED' feature had some very high values, up to 958 years, which is clearly impossible. This indicates incorrect data that should be addressed.

On a positive note, the 'DAYS_BIRTH' feature appeared to have no outliers, so the date of birth data can be trusted. However, the 'AMT_INCOME_TOTAL' feature had a lot of outliers, meaning some loan applicants have unusually high incomes compared to the rest.

Next, I created a new list called **'cat_col'** to store the columns that need to be converted to categorical variables.

After **imputing** and converting, I was left with 43 columns for further analysis.

6) I started analyzing the **'previous_application.csv'** dataset, also known as 'prev_app_inp1df'. Analysis for missing values showed the that **'prev_app_inp1df'** data set have 4 columns that have missing value over 50% ,then I checked the columns with over **50%** missing values and stored them in a variable called 'missing_val_column_50' and dropped them.

Next, I identified and stored some unwanted columns in **'prev_data_unused_col'**, then dropped them. After this cleanup, the dataset had 1,670,214 rows and 29 columns.

After that I **imputed** 'Unknown' for the missing values to "NAME_TYPE_SUITE" column.I found that columns like 'DAYS_FIRST_DUE', 'DAYS_TERMINATION', 'DAYS_FIRST_DRAWING', 'DAYS_LAST_DUE_1ST_VERSION', and 'DAYS_LAST_DUE' had missing values, which I decided to keep as they represented days.

I also found that 'CNT_PAYMENT' with 0 as the 'NAME_CONTRACT_STATUS' indicated that most of these loans were not started. This provided an interesting insight into the data.

7) I plotted a **kde plot** for "AMT_GOODS_PRICE" to understand the distribution.

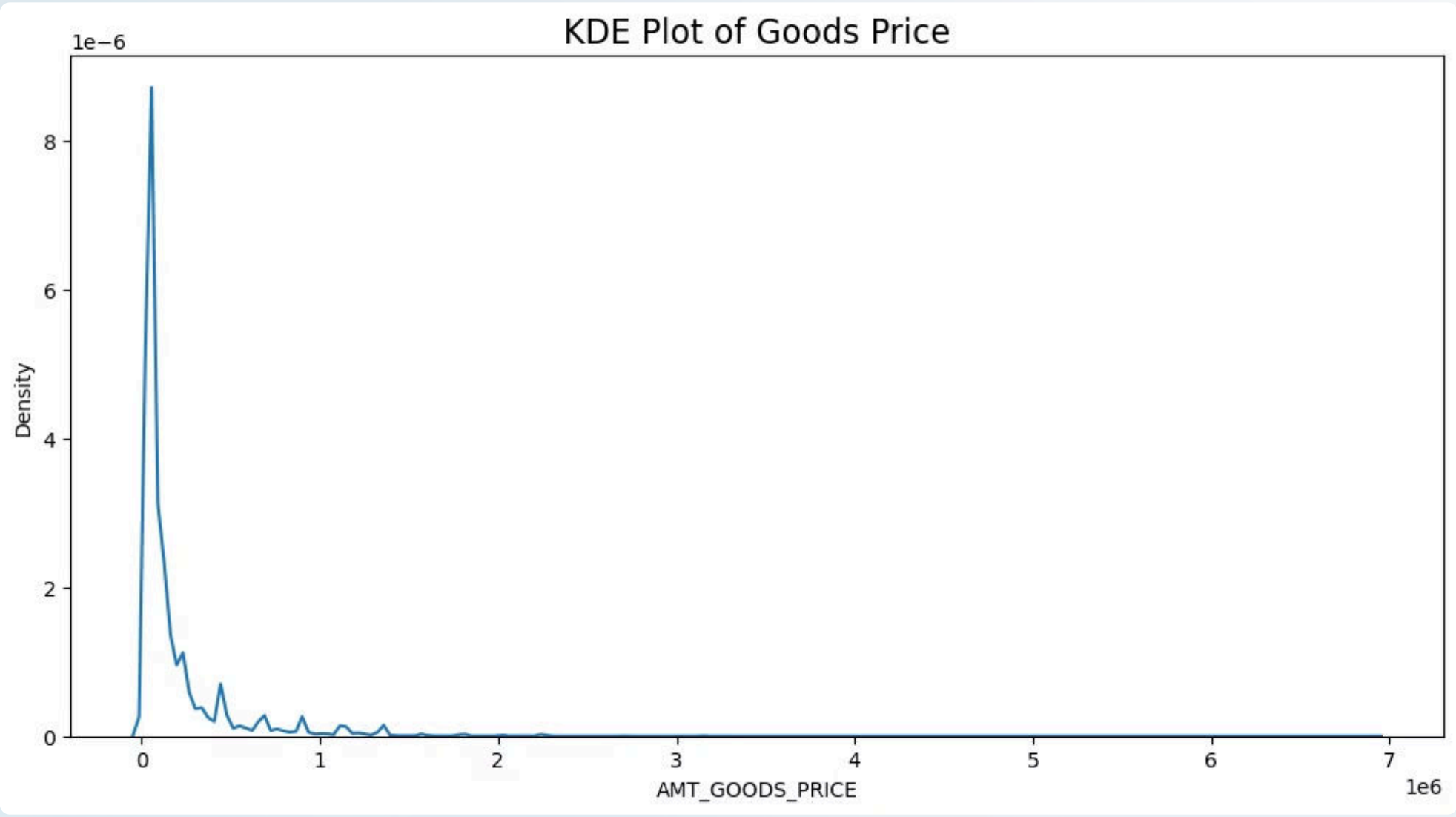


Figure: 5.0

After analyzing the data, I found that there are several peaks along the distribution of the "AMT_GOODS_PRICE" variable. This suggests the data may not have a normal distribution. To better understand the distribution, I will need to impute the mean, median, and mode to see if the distribution is the same across these different central tendency measures.

8) After analyzing the "AMT_GOODS_PRICE" variable, I created a new data frame with columns imputed using the mode, median, and mean values. For this i used **sub plot**. This will allow me to better understand the distribution of this key variable and identify any potential issues or outliers.

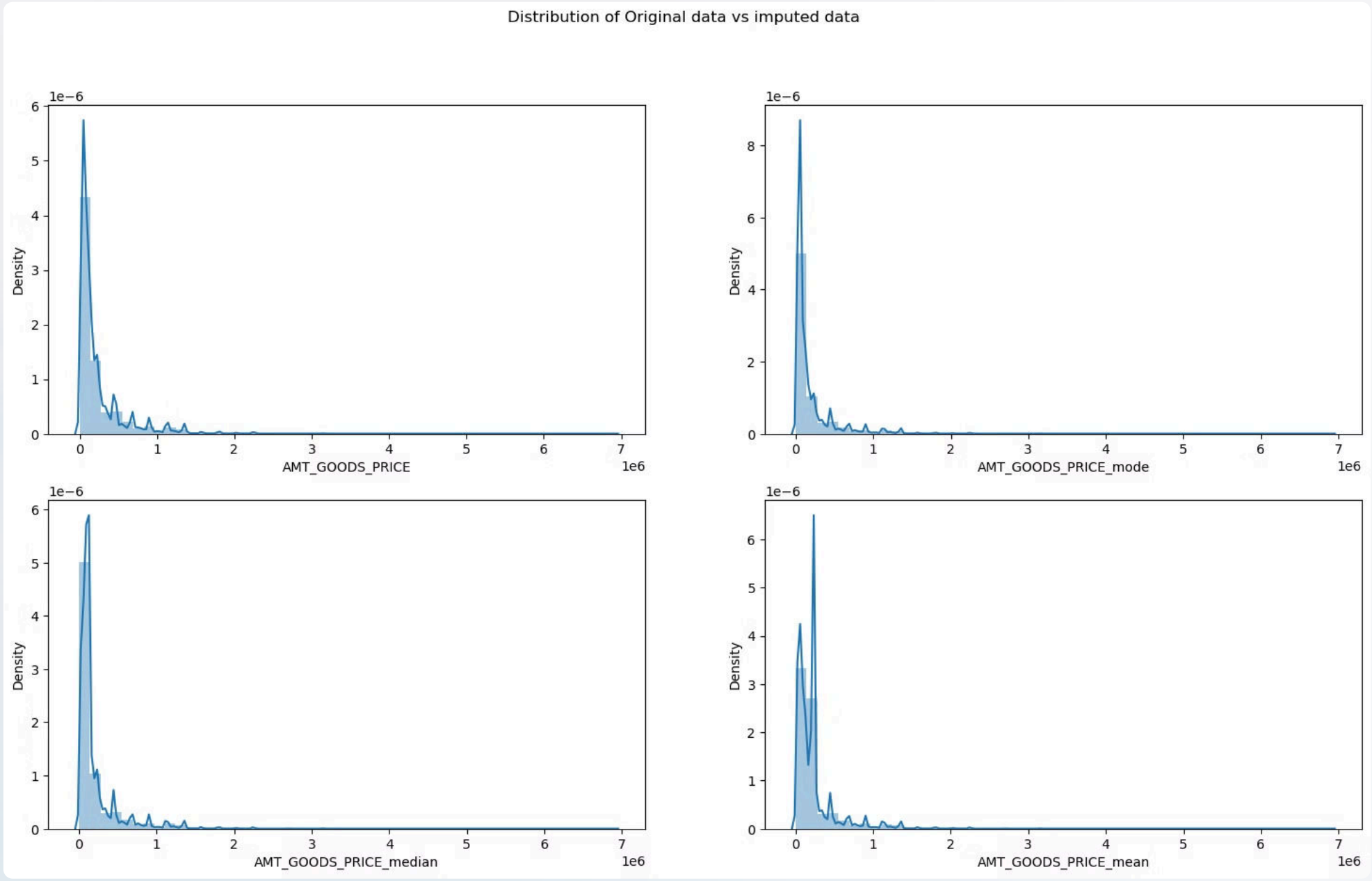


Figure: 5.1-5.4

After analyzing the graph i found out that the original distribution is very close to the distribution of data imputed with mode in this case, So I imputed mode for missing values.

9) Then I plotted a **kdeplot** for "AMT_ANNUITY" to understand the distribution.

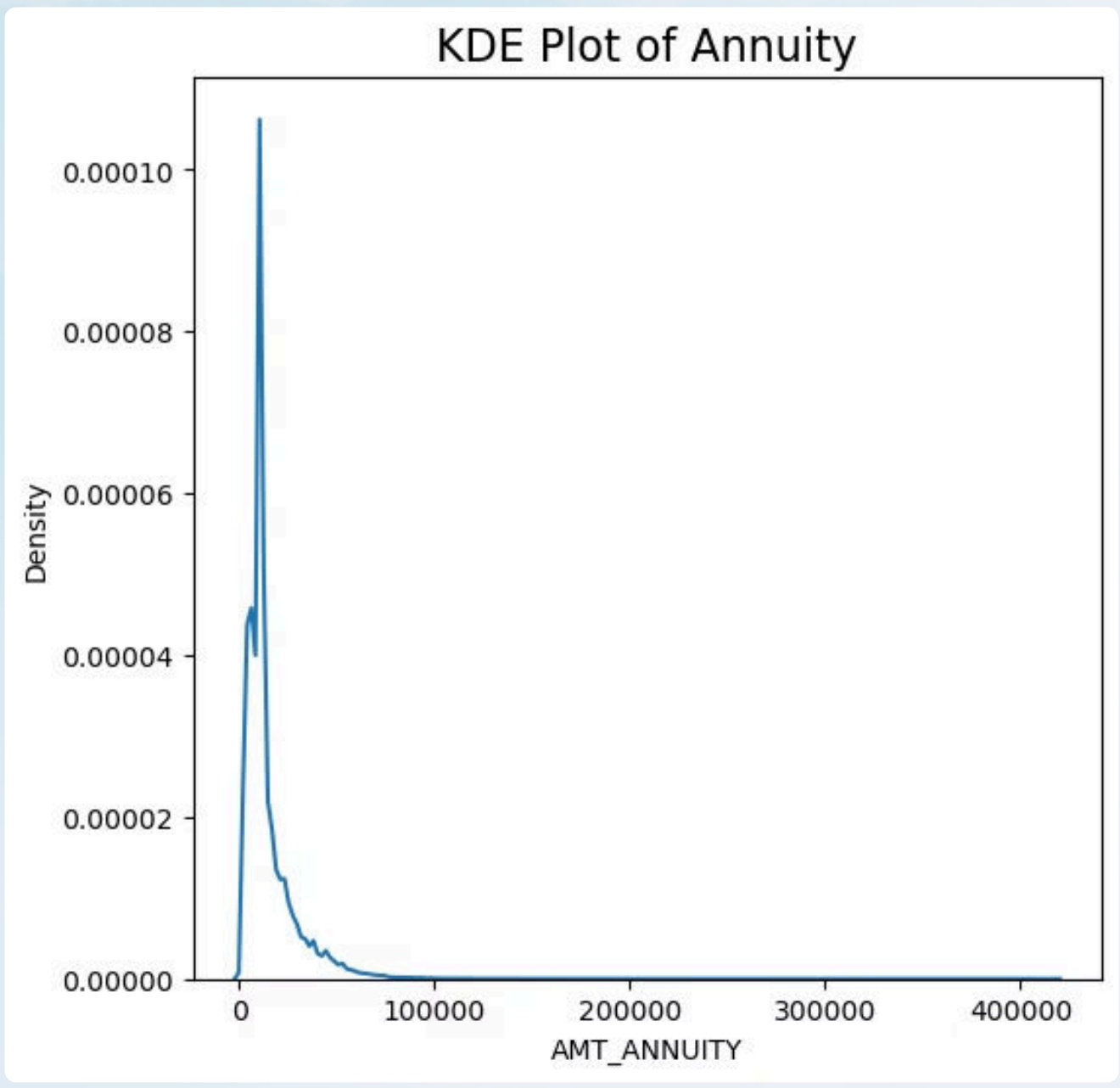


Figure: 5.5

After analyzing the distribution of "AMT_GOODS_PRICE", I can see there is a single peak at the left side of the distribution. This peak indicates the presence of **outliers** in this variable.

Given the presence of outliers, imputing with the median would be more appropriate than using the mean. The mean can be heavily influenced by extreme values, so the median is a more robust measure of central tendency in this case.

10) After **imputing** the missing values for 'AMT_ANNUITY' and 'AMT_GOODS_PRICE' using the median, I noticed that these variables had some outliers. Using the median as the imputation method helped me address the impact of these extreme values on the central tendency measures.

Standardize Values for prev_app_inp1df

Next, I converted the negative values in the days-related columns to positive values using the **abs()** function. This created a new variable called **"prev_data_days_col"** that allowed me to work with the data in a more meaningful way.

Diving deeper into the 'YEARLY_DECISION' and 'DAYS_DECISION' variables, I found some interesting insights. About **35%** of loan applicants had applied for a new loan within the first year of their previous loan decision. Another **23%** had applied within the second year, and 13% within the third year.

To further prepare the data, I converted several columns to categorical variables and stored them in the **'prev_cat_col'** variable.

11) Checking Outliers for 'prev_app_inp1df' data set

After this, I checked for any additional outliers in the **'prev_app_inp1df'** dataset. I used the describe() function to identify columns with a high difference between the 75th percentile and maximum values, and stored those columns in a new variable called **'prev_data_out_col'**.

I used boxplot on 'prev_data_out_col' to check outliers.

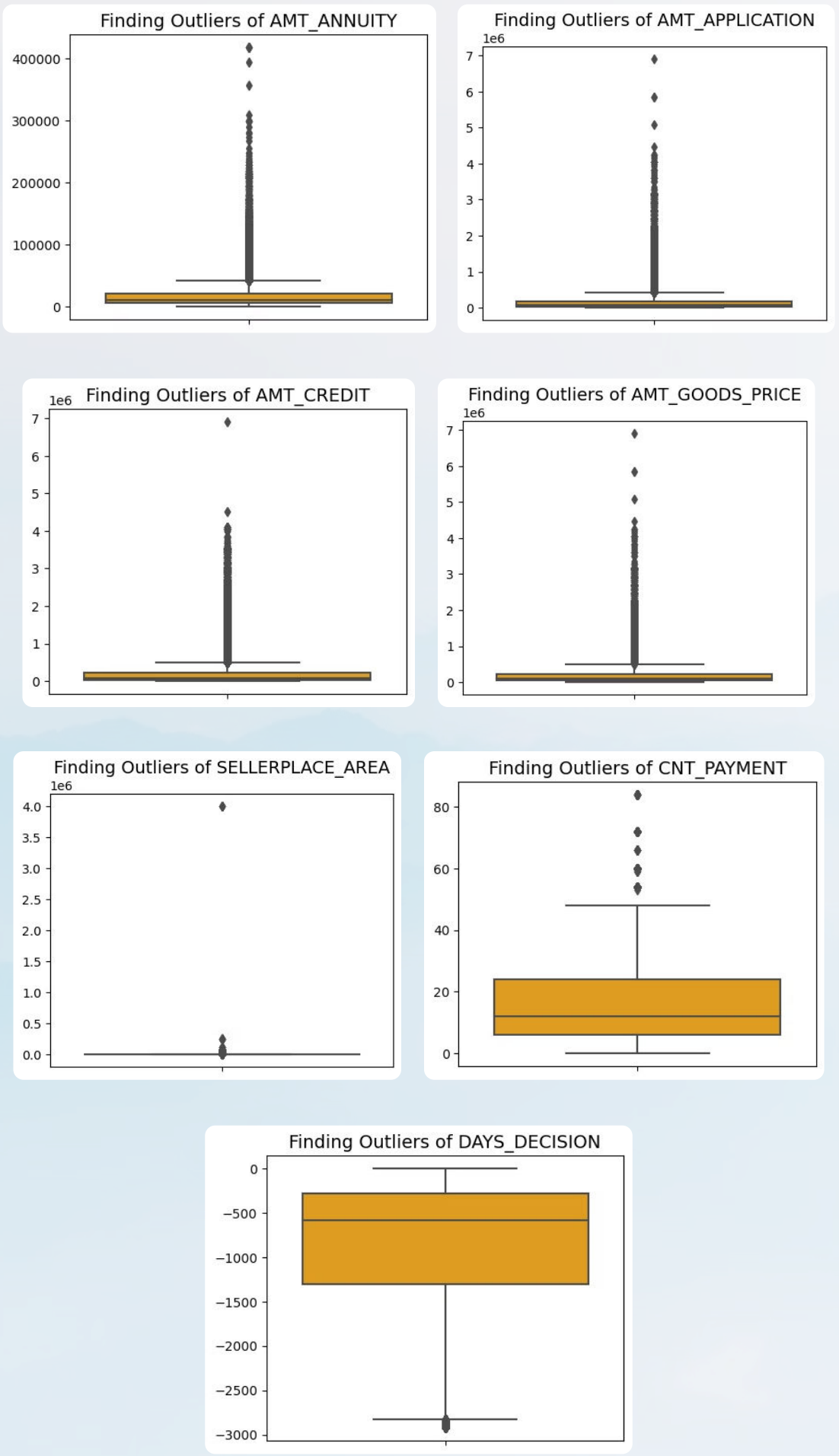


Figure: 6.0-6.6

I noticed that '**AMT_ANNUITY', 'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE',** and '**SELLERPLACE_AREA'** had a **high number of outliers**. The 'CNT_PAYMENT' column also had a few outliers, while 'DAYS_DECISION' had relatively fewer outliers, indicating those previous application decisions were made long ago.

12) Data Analysis

I started with checking Imbalance of data

I used **bar plot** on "Repayers" and "Defaulters" to check **Imbalance plotting** of 'Repayers Vs Defaulters'

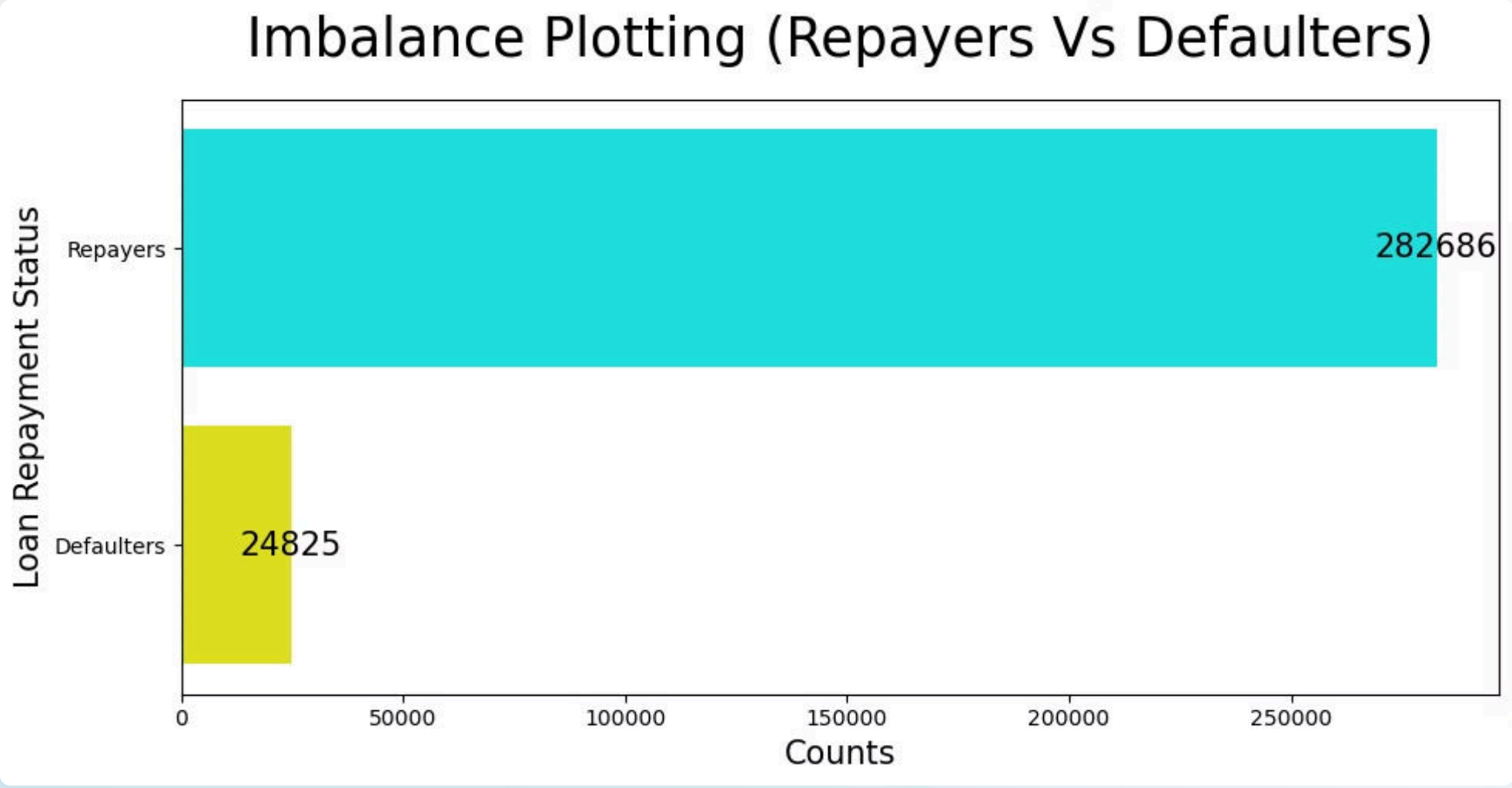


Figure: 7.0

The analysis shows that the number of repayers is much higher than the number of defaulters. Specifically, there are **282,686** repayers and **24,825** defaulters.

This information provides important context for understanding the overall risk profile of the loan applicants. The significantly higher number of repayers compared to defaulters suggests that the majority of applicants are lower-risk borrowers.

However, it's still important to analyze the data further to identify any potential patterns or factors that may contribute to loan defaults.

After that i checked the ratio of imbalance percentage with respect to defaulters and repayers and found out that :

Repayer Percentage is 91.93%
Defaulter Percentage is 8.07%
Imbalance Ratio with respect to Repayers and Defaulters is given: 11.39/1 (approx)

13) Plotting Functions for Univariate Analysis

I created a **'univariate_data'** function that plots a count plot and a plot showing the percentage of defaulters. The **'data_type'** function checks if the data is numerical or categorical. For numerical data, it plots a **histogram**. For categorical data, it creates a **subplot, bar plot, and count plot.** I also created **'cat_column'** and **'num_column'** variables to store the column types..

a) Segmented Univariate Analysis

A) First, I checked the **gender ratio** in terms of clients and defaulters. This provided an initial overview of any potential differences in loan repayment patterns between genders.

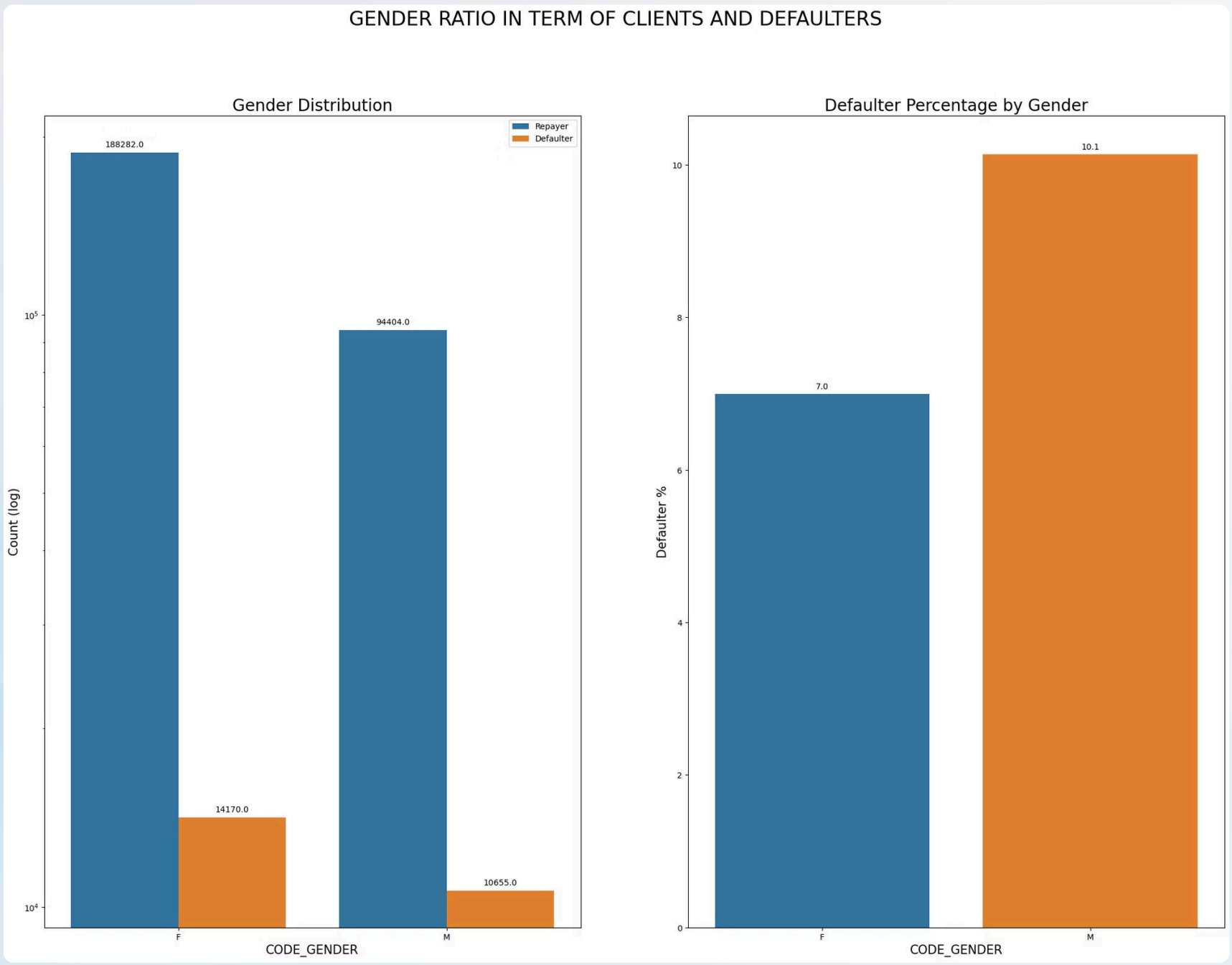


Figure: 8.0

Based on the analysis, it appears that the number of **female** counts is **quite high**. The data on defaulters shows that **males** have a higher chance of being defaulters compared to females. The percentage of male defaulters is around **10%**, while the percentage of female defaulters is around **7%**.

This suggests that gender plays a role in the risk profile of loan applicants. It's important to further investigate the reasons behind this difference and consider it as a factor in the risk assessment process. Understanding the relationship between gender and default rates can help banks make more informed decisions when evaluating loan applications.

B) After analyzing the previous datasets, I wanted to dive deeper into the relationship between **contract type** and loan repayment status. This could provide valuable insights into how the structure of the loan contract impacts a client's ability to repay.

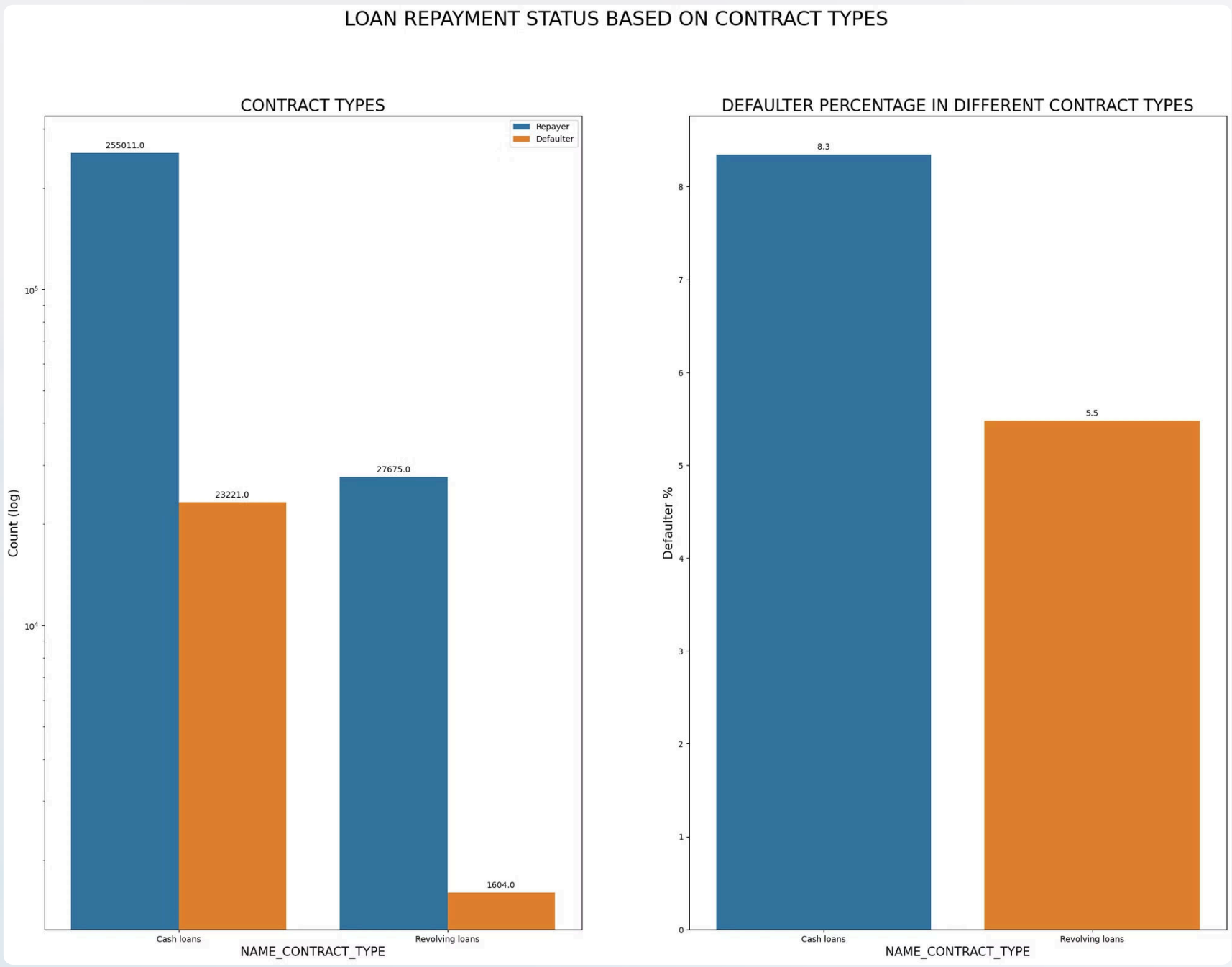


Figure: 8.1

Revolving loans are just a small fraction of total loans. About **8%** of **Cash loan** applicants and **5.5%** of **Revolving loan** applicants are in default.

C) After analyzing the input datasets, I wanted to check if owning **real estate** is related to loan repayment status. This could provide valuable insights into how property ownership impacts loan performance.

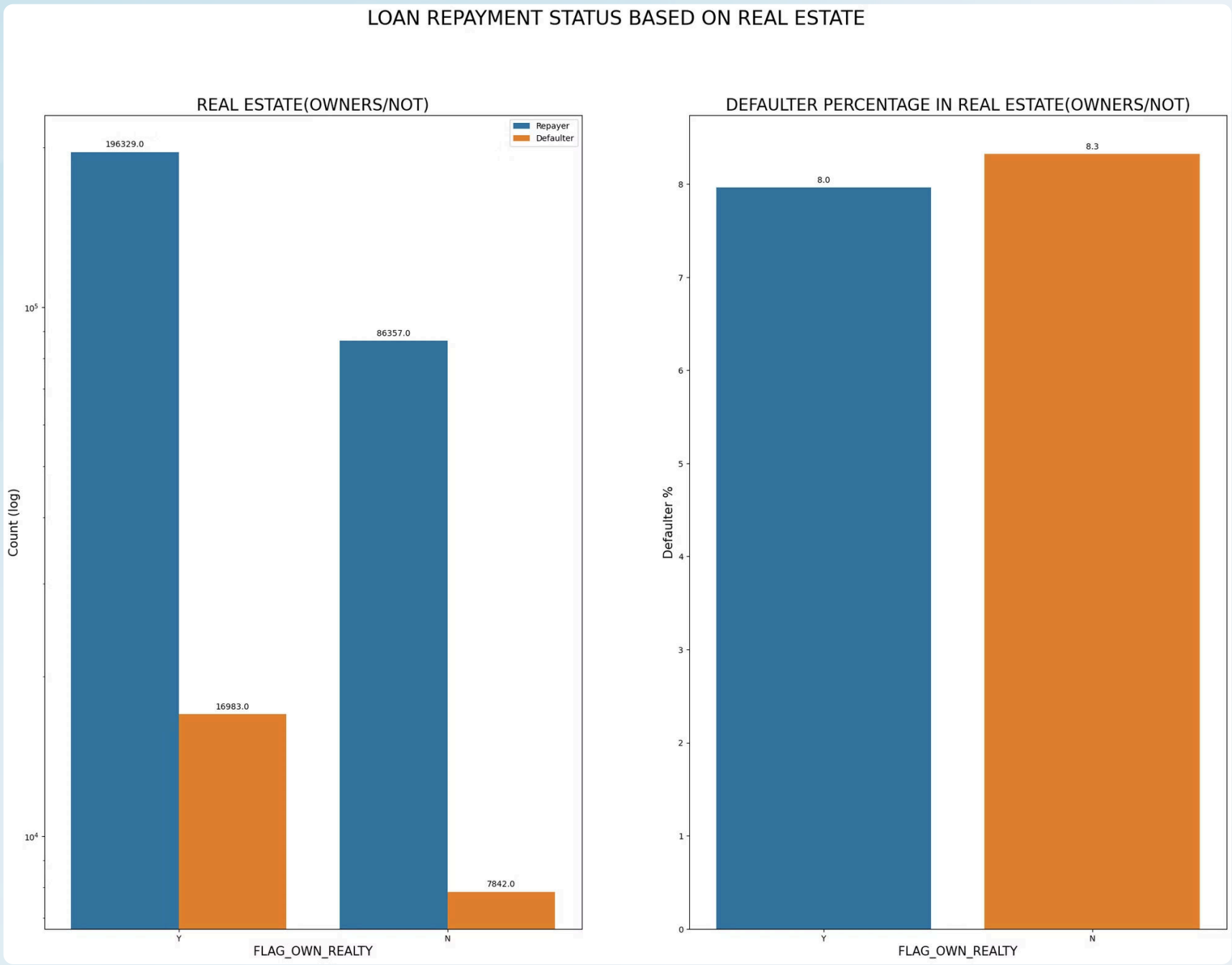


Figure: 8.2

The analysis showed that clients who own real estate are more likely to take loans, but the defaulting rate is similar for both groups at around **8%**. This suggests there is **no correlation** between owning real estate and defaulting on loans.

D) Analyzing the relationship between **family status** and loan repayment status could provide important insights. By understanding how an individual's family situation may impact their ability to repay a loan, we can make more informed underwriting decisions and better manage risk.

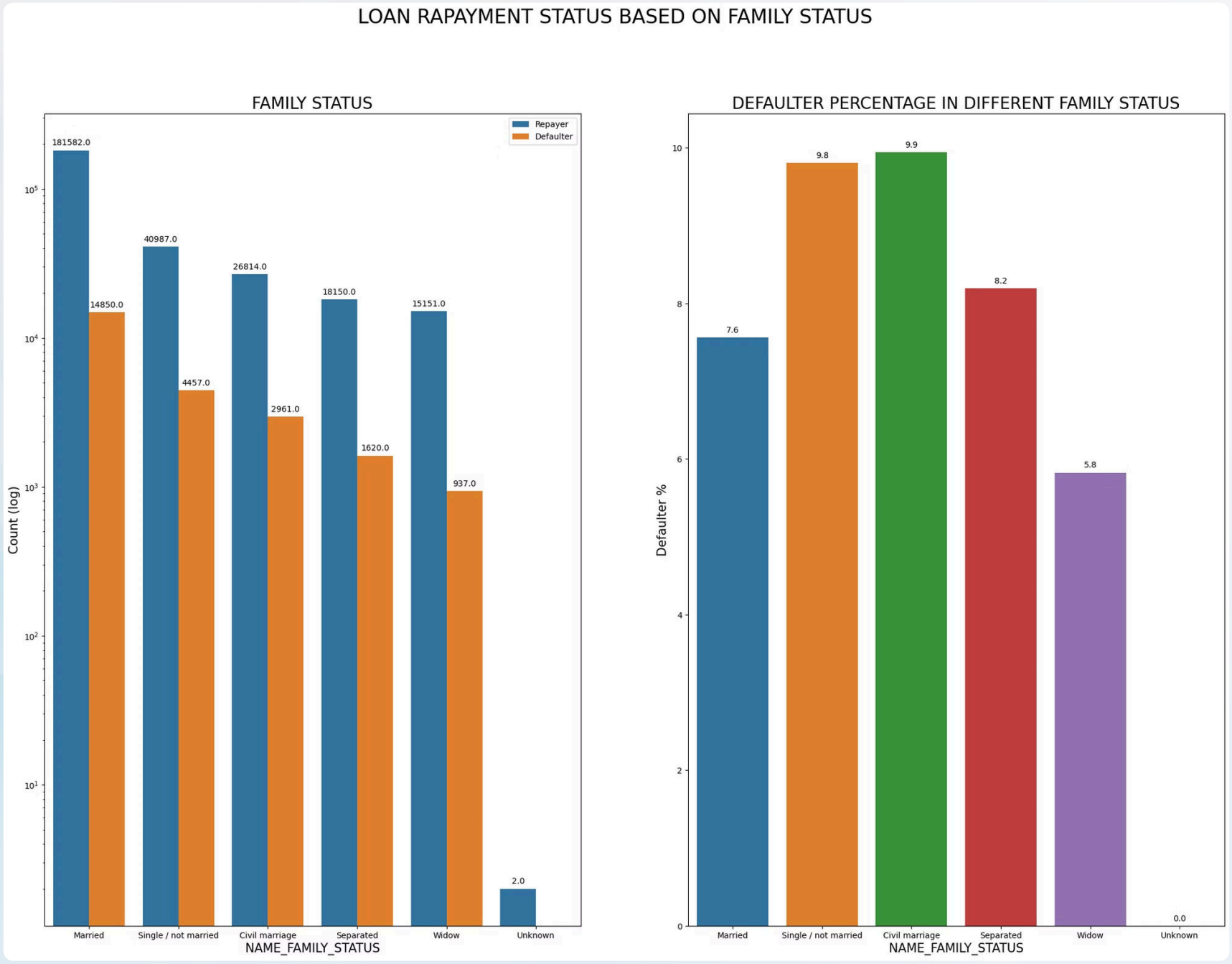


Figure: 8.3

The analysis showed that the highest number of loan takers are married, followed by single/not, civil marriage, separated, and widow. In terms of defaulters, **civil marriage** has the highest percentage at **9.9%**, while **widow** has the lowest at **5.8%** (excluding unknown).

E) After that, I decided to take a closer look at the relationship between **housing type** and loan repayment status. This analysis could provide valuable insights into how different living situations may impact an individual's ability to manage their finances and meet their loan obligations.

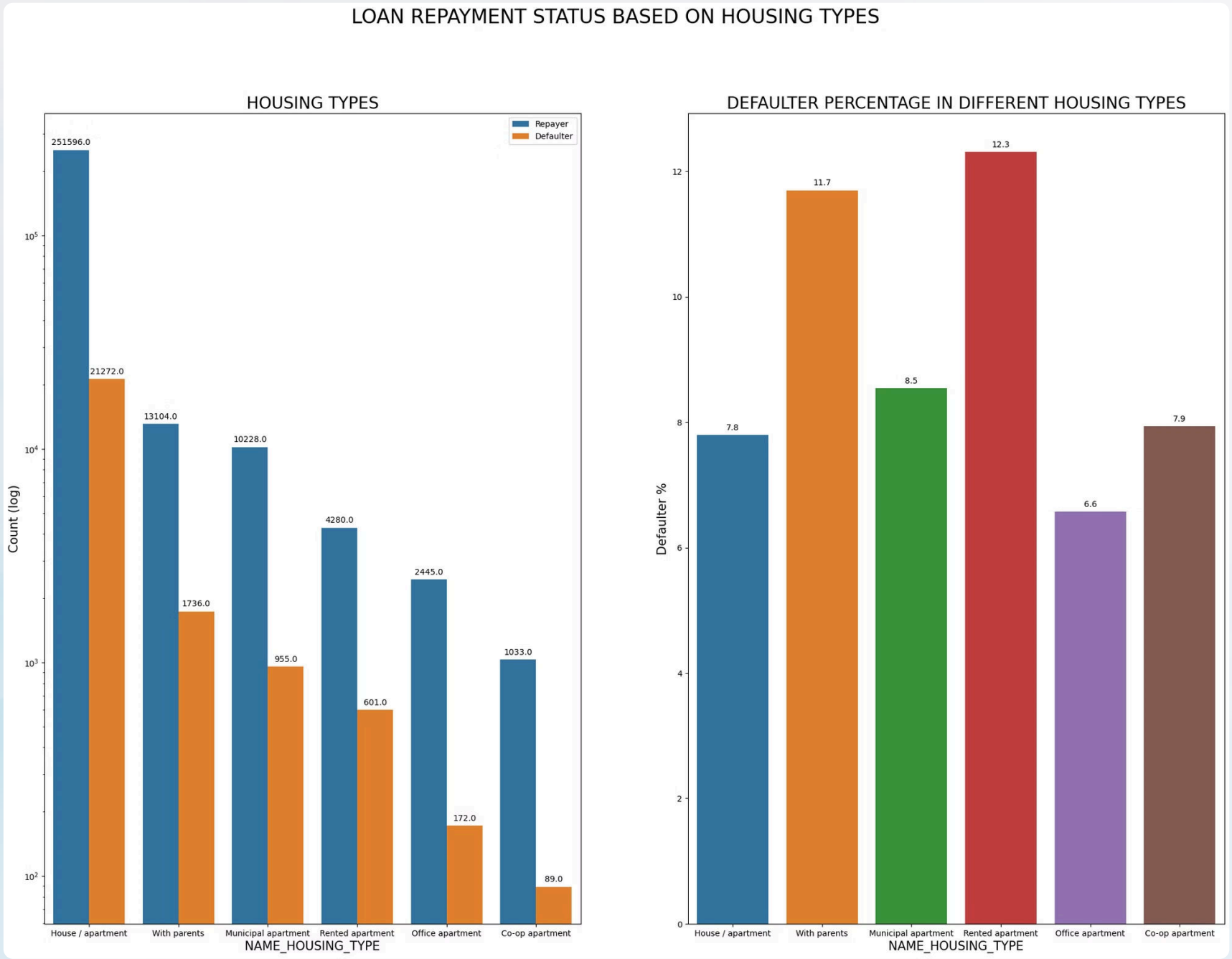


Figure: 8.4

The analyis showed that the most loan borrowers live in houses/apartments, followed by living with parents, municipal apartments, rented apartments, office apartments, and co-op apartments. Those in **office apartments** have the lowest default rate at **6.6%**, while those in **rented apartments** have the highest at **12.3%**. **People living with parents** also have a higher default rate around **11.7%**, making rented apartment residents the **riskiest** borrowers.

F) After that i analyzed **Education Type** based on loan repayment status.The data on loan repayment status provides an interesting window into how an individual's living situation can impact their financial well-being. After analyzing the numbers, a clear pattern emerges regarding the relationship between housing type and loan repayment rates.

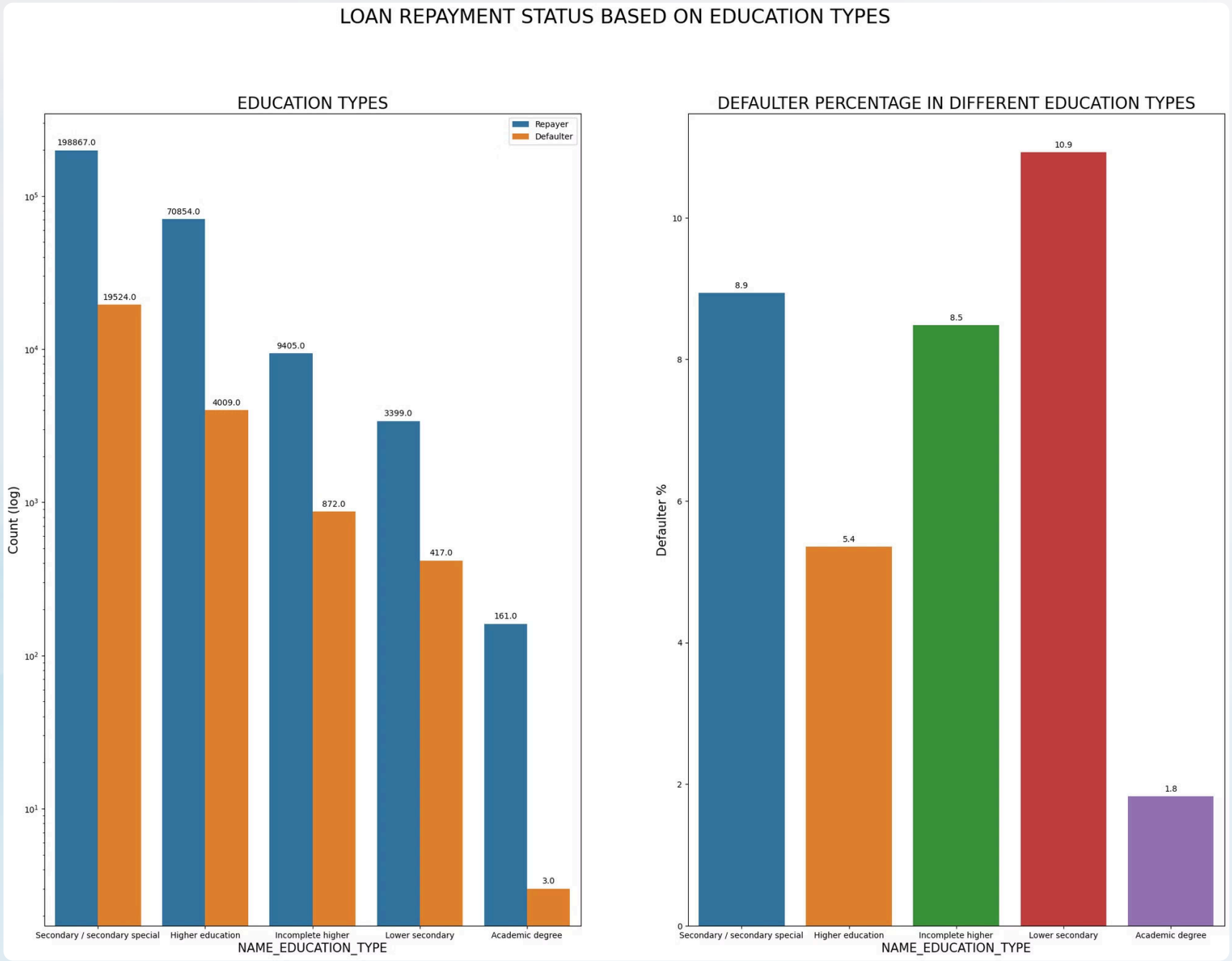


Figure: 8.5

The data shows that borrowers with lower **secondary education** have the highest default rate at **10.9%**, while those with **academic degrees** have the lowest at just **1.8%**. This suggests a strong correlation between education level and loan repayment ability.

G) After examining the relationship between housing type and loan repayment status, I wanted to dig deeper into the data to see if there were any patterns based on the applicant's **occupation type**.

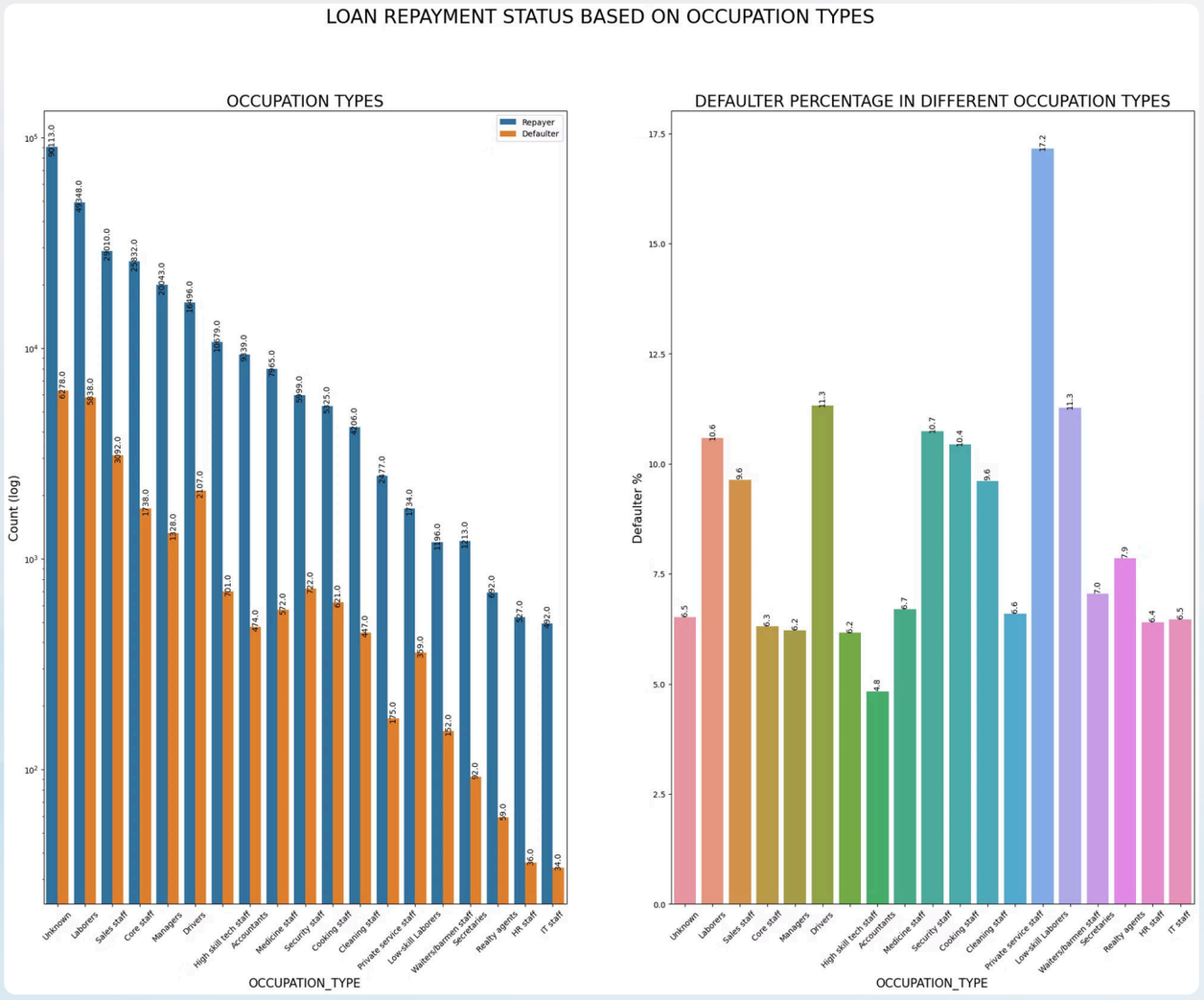


Figure: 8.6

The data shows occupational differences in loan repayment. Borrowers in more **stable, higher-earning jobs** like accounting have lower default rates, around **5%**. In contrast, those in **service or temporary roles** like laborers, drivers, and waiters face much higher default rates, up to **17%**. This suggests financial stability, whether through employment or housing, is key for successful loan repayment.

H) After analyzing the relationship between housing type and loan repayment status, I decided to take a closer look at **income type** as well. This could provide additional insights into how an individual's financial situation impacts their ability to manage their loan obligations.

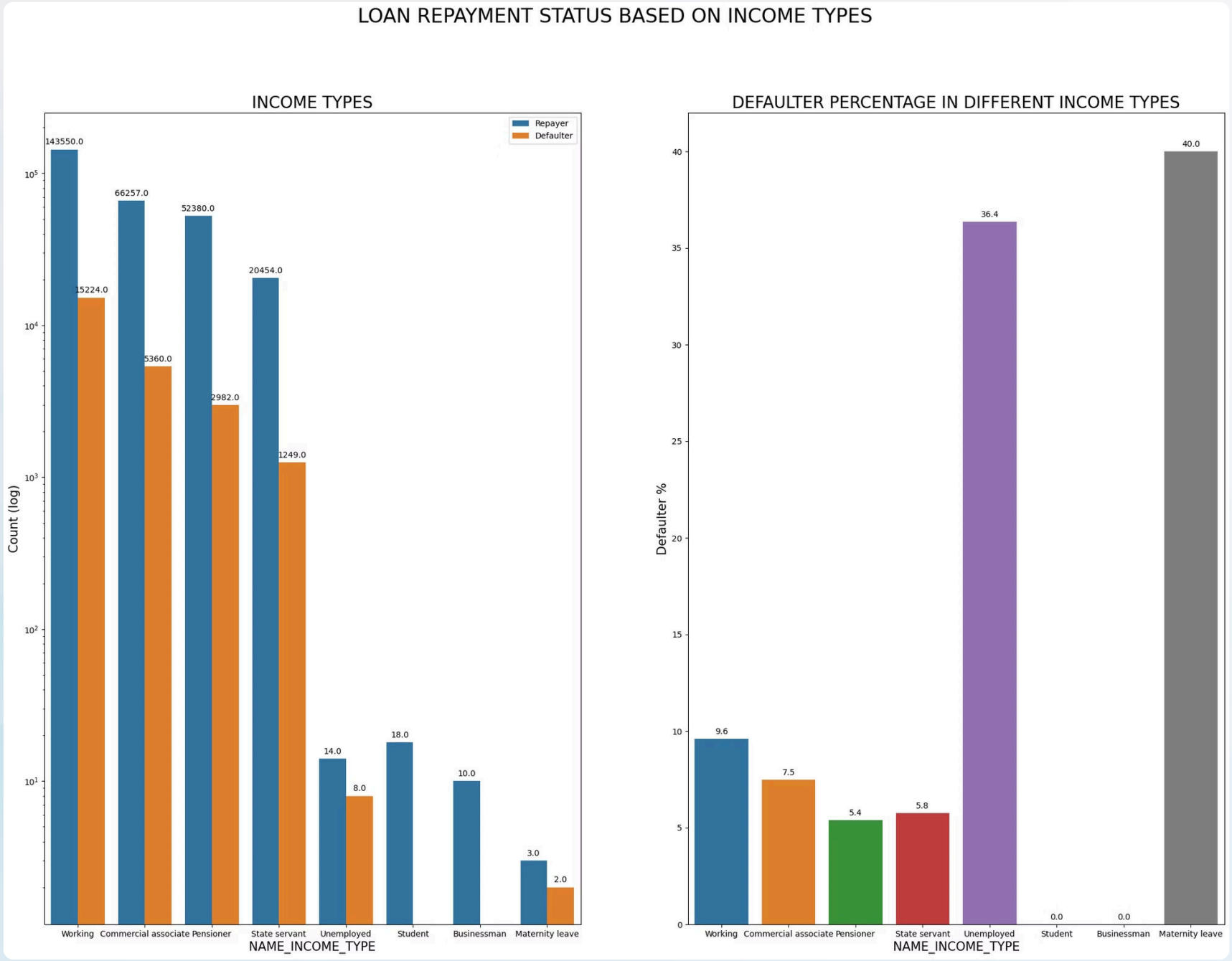
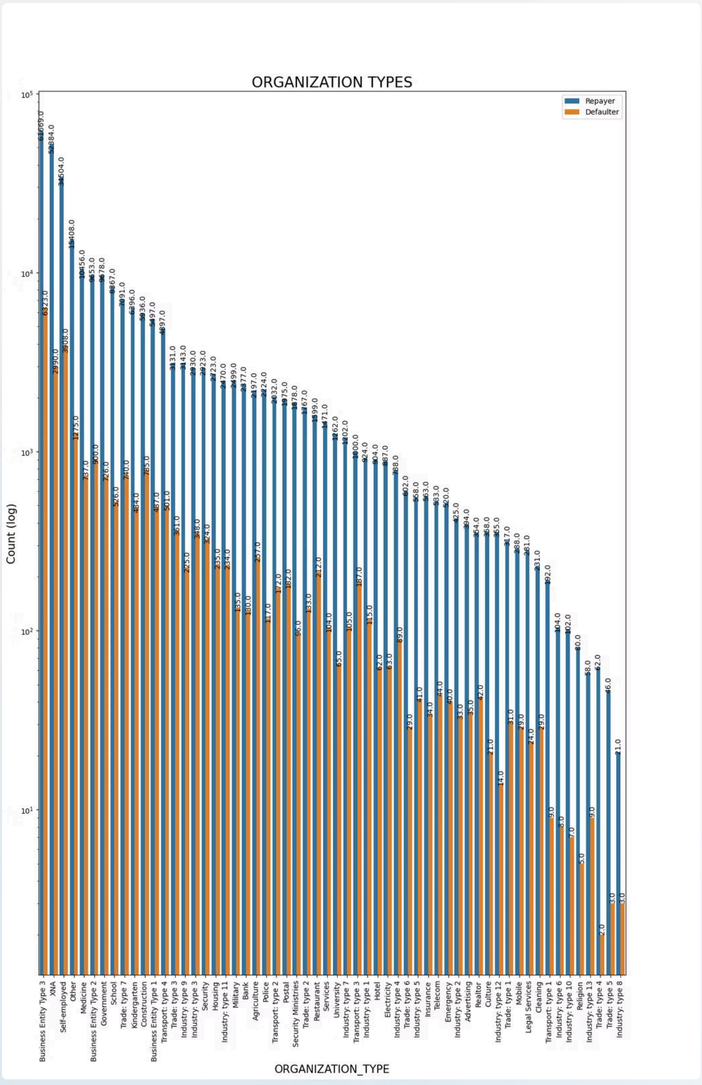


Figure: 8.7

The data shows that borrowers with the riskiest income types, such as those on **maternity leave** (**40%** default rate) and the **unemployed (36.4%** default rate), have much higher default rates compared to more stable income sources like salaries (around **10%** default). Students and business owners, though less likely to take loans, have no recorded defaults. This suggests **financial stability**, whether through employment or other reliable income, is key for successful loan repayment.

I) After analyzing the relationship between housing type and loan repayment status, it made sense to take a closer look at the impact of **organization type** as well. This could uncover additional patterns and factors that influence an individual's ability to manage their loan obligations.

LOAN REPAYMENT STATUS BASED ON ORGANIZATION TYPES



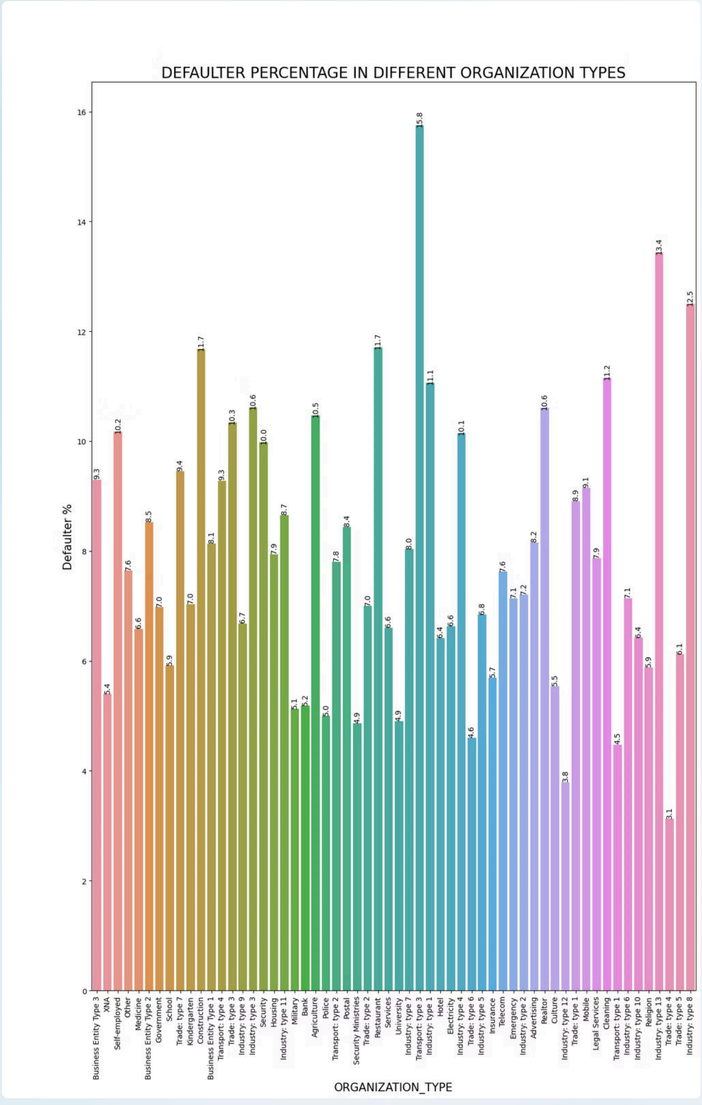


Figure:8.8

Most loan applications are from Business Entity **Type 3**. Many applications lack organization type info. Highest default rates are in Transport (**15.8%**), Industry (**13.4%, 12.5%**), and Restaurants (**11.7%**). Self-employed borrowers have a **10.2%** default rate - consider higher interest to mitigate risk. T**rade Type 4/5** and **Industry 8** have **lower** default rates, so are safer for lending.

J) The data shows that analyzing loan repayment status by **age group** can yield some interesting insights.

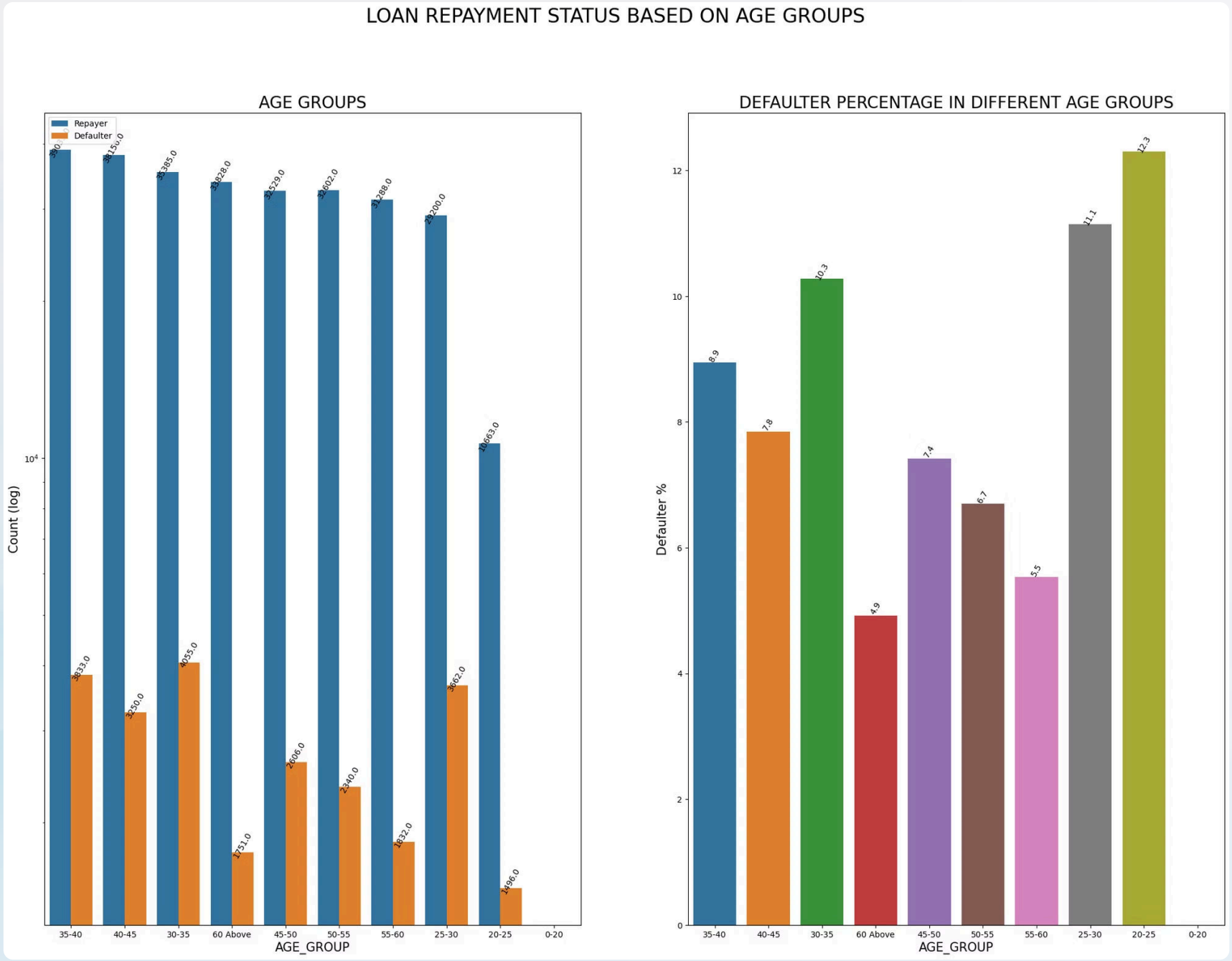


Figure: 8.9

The data shows that younger borrowers, particularly those aged between **20-35**, tend to have **higher** default rates compared to older age groups. This could be due to less financial stability and experience managing debt early in one's career. Conversely, those **aged 60** and above have the **lowest** default rates, likely benefiting from more established careers and financial resources.

K) Now I've analyzed the **'Employment_Year'** data in relation to loan repayment status. This can provide some interesting insights into how an individual's stage in their career may impact their ability to manage their financial obligations.

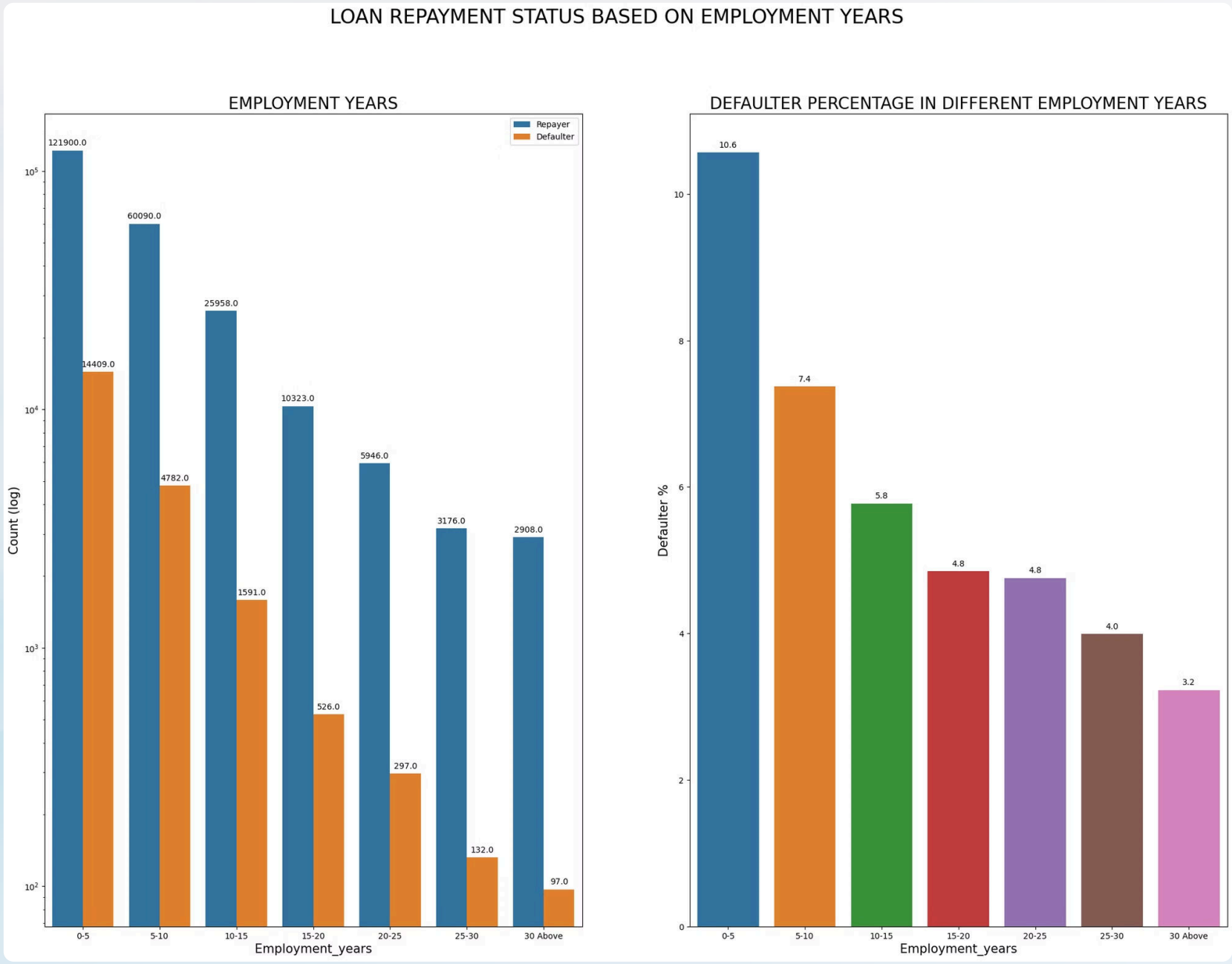


Figure: 8.10

The data shows that newer employees with **0-5 years** of experience have the highest application volume but also the highest default rate at **10.6%**. Meanwhile, those with **30+ years** of experience have the lowest default rate at **3.2%**, though lower application numbers. **Tightening eligibility criteria** for the 0-5 years group, such as **requiring more documentation** or **higher credit scores**, could help **decrease** their default rate. Overall, default rates tend to decrease as years of experience increase.

L) I wanted to take a closer look at how the number of **children** an individual has may impact their loan repayment status. This could be an important factor to consider, as the financial obligations and responsibilities of raising a family may create unique challenges when it comes to managing student debt and other loans.

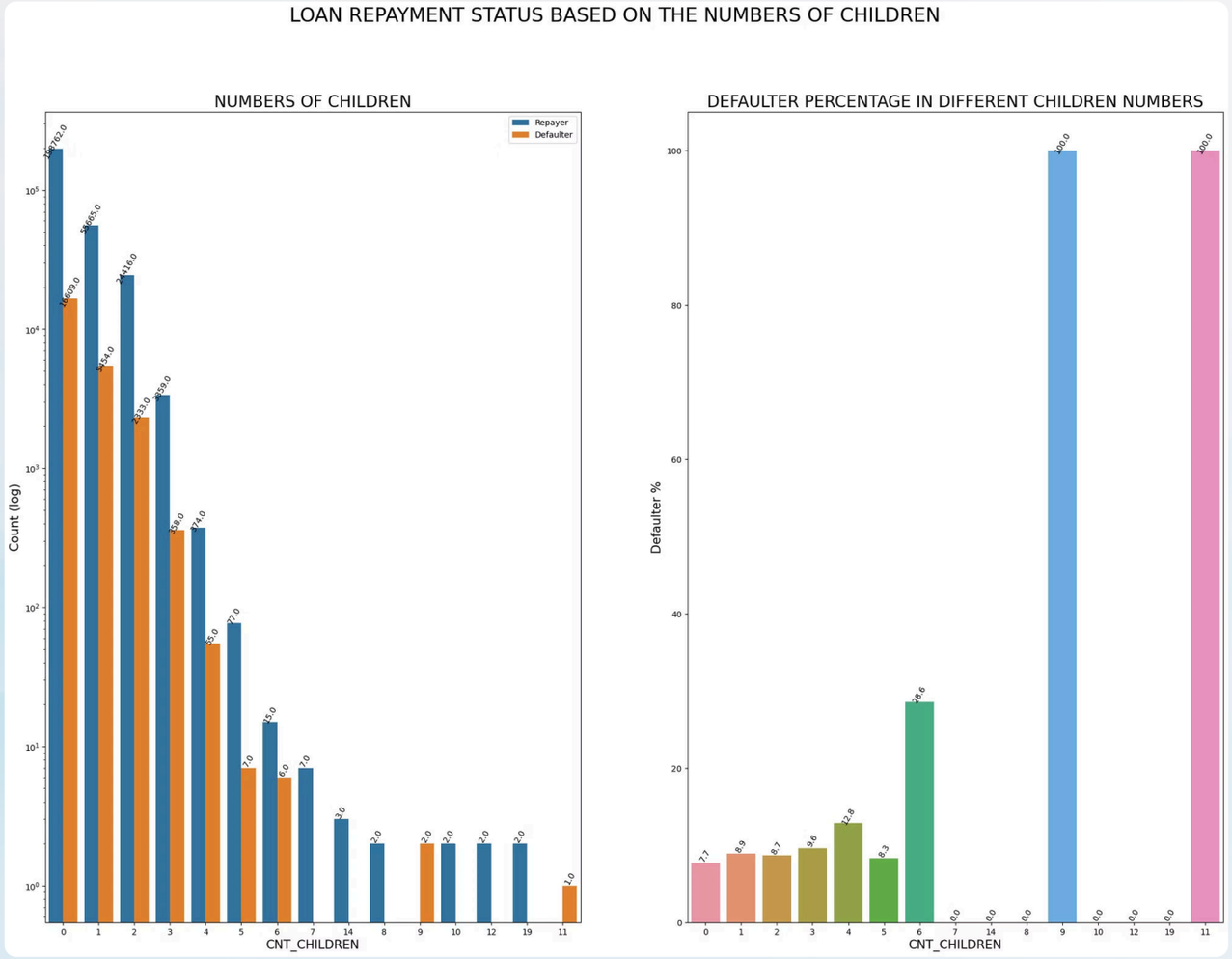


Figure: 8.11

The data shows that borrowers with **larger** families tend to have higher default rates. Those with **6 children** have a **28.6%** default rate, while those with **9** or **11** children have a **100%** default rate. This suggests that family size is an important factor to consider when evaluating loan risk.

b) Numerical Univariate Analysis

This card provides a deeper look at the **numerical columns** related to loan amounts. I **plotted** the distribution, so I can gain insights into the density and spread of these important financial variables.Selecting the numerical columns related to amount.

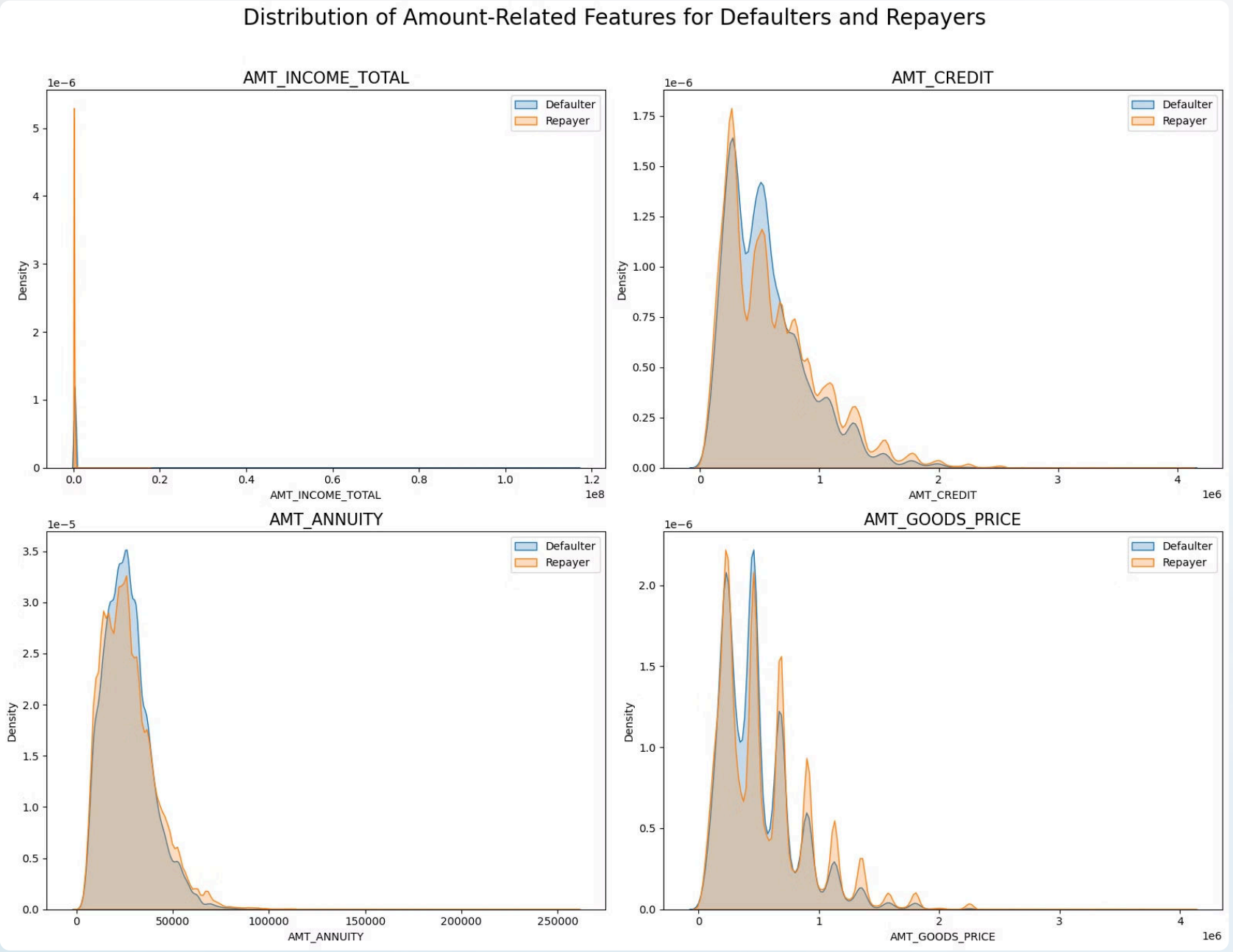


Figure: 9.0

Insights we got from this graphs the credit amount distributions show **defaulters** and **repayers** have similar **right-skewed patterns**, with the highest densities at **lower credit amounts**. Subtle differences exist, like repayers having slightly higher density at the lowest credits, while defaulters have more in the mid-range. The annuity amount distributions are also right-skewed, with **defaulters** having slightly higher densities in the **30,000 to 50,000** range. Goods prices show multi-modal distributions, with defaulters and repayers having very similar patterns but minor differences in density peaks.

14) I plotted **pairplot** between amount variable to draw reference against loan repayment status.Selected the relevant columns from the DataFrame and filter out rows with null values in 'AMT_GOODS_PRICE' and 'AMT_ANNUITY'.

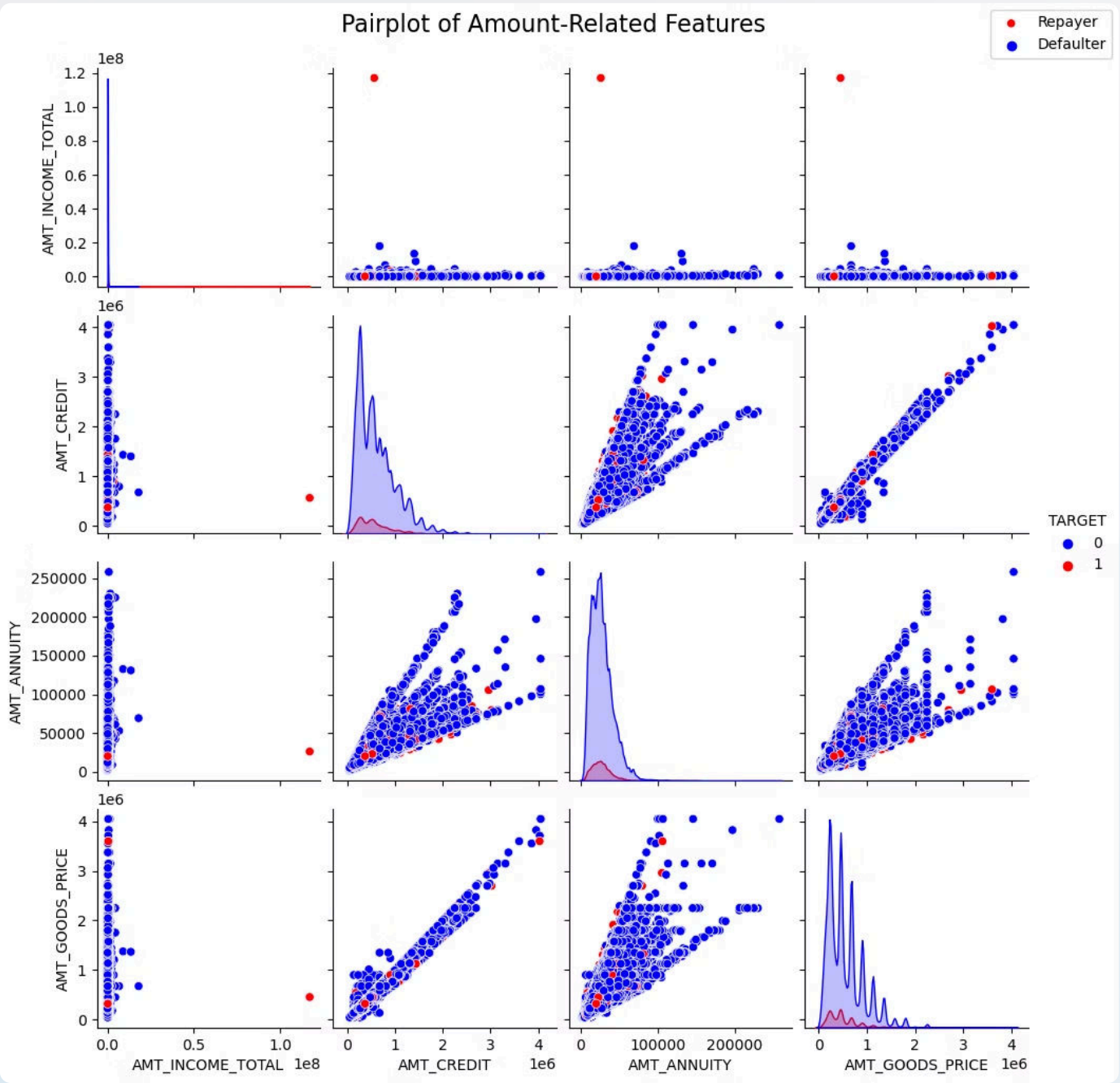


Figure: 10.0

Analysis shows that the defaulters are less likely when annuity is over **15K** and goods price is **over 20 Lakhs**. There are fewer defaulters with credit **over 20 Lakhs**. Most data points for both groups are at lower income levels, with some high-income outliers among repayers. Credit amount shows positive correlation with annuity and goods price. AMT_CREDIT and AMT_ANNUITY are positively correlated with other features. Defaulters have slightly higher annuity at the higher end. There is a **strong positive correlation** between AMT_CREDIT and AMT_GOODS_PRICE.

15) Bivariate Analysis

A) bivariate categorical analysis

I created a function called **'bivariate_categorical'** for bivariate analysis on categorical columns.

I analyzed Income type vs Income Amount Range on a **Seaborn Barplot**:

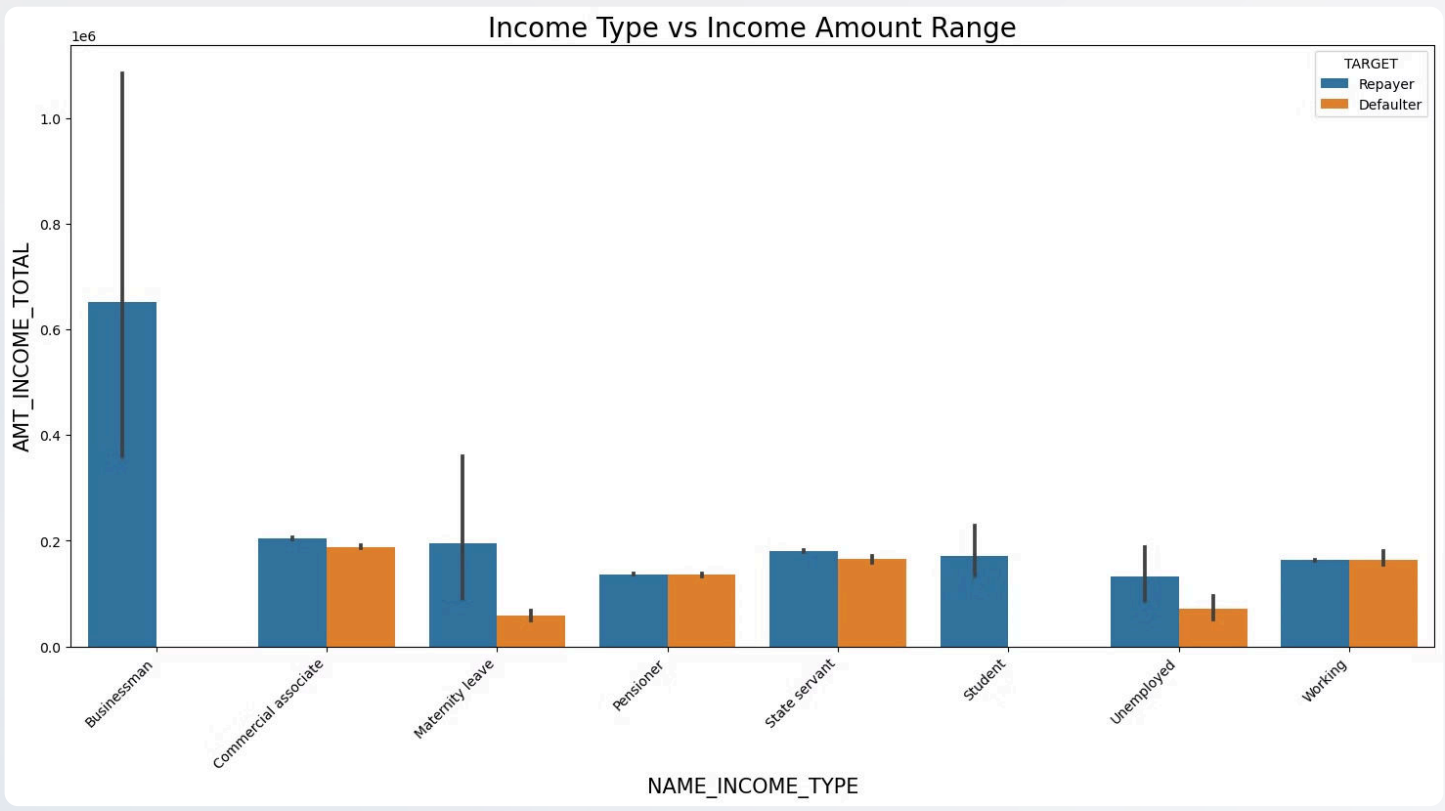


Figure: 11.0

Businessmen have the **highest average income**, but high variation suggests inconsistent earnings. **Commercial Associates** earn less but have **lower default** rates, likely due to more stable finances. **Unemployed** individuals have very low incomes and high default rates. **Maternity leave** leads to **lower incomes** and **higher defaults**. **State servants** have moderate incomes but low default rates, likely due to job security.

B) Bivariate Numerical Analysis

I created a function called **'bivariate_numerical'** for bivariate analysis on numerical columns.

I analyzed the relationship between 'AMT_GOODS_PRICE' and 'AMT_CREDIT' and comparing with loan repayment status.

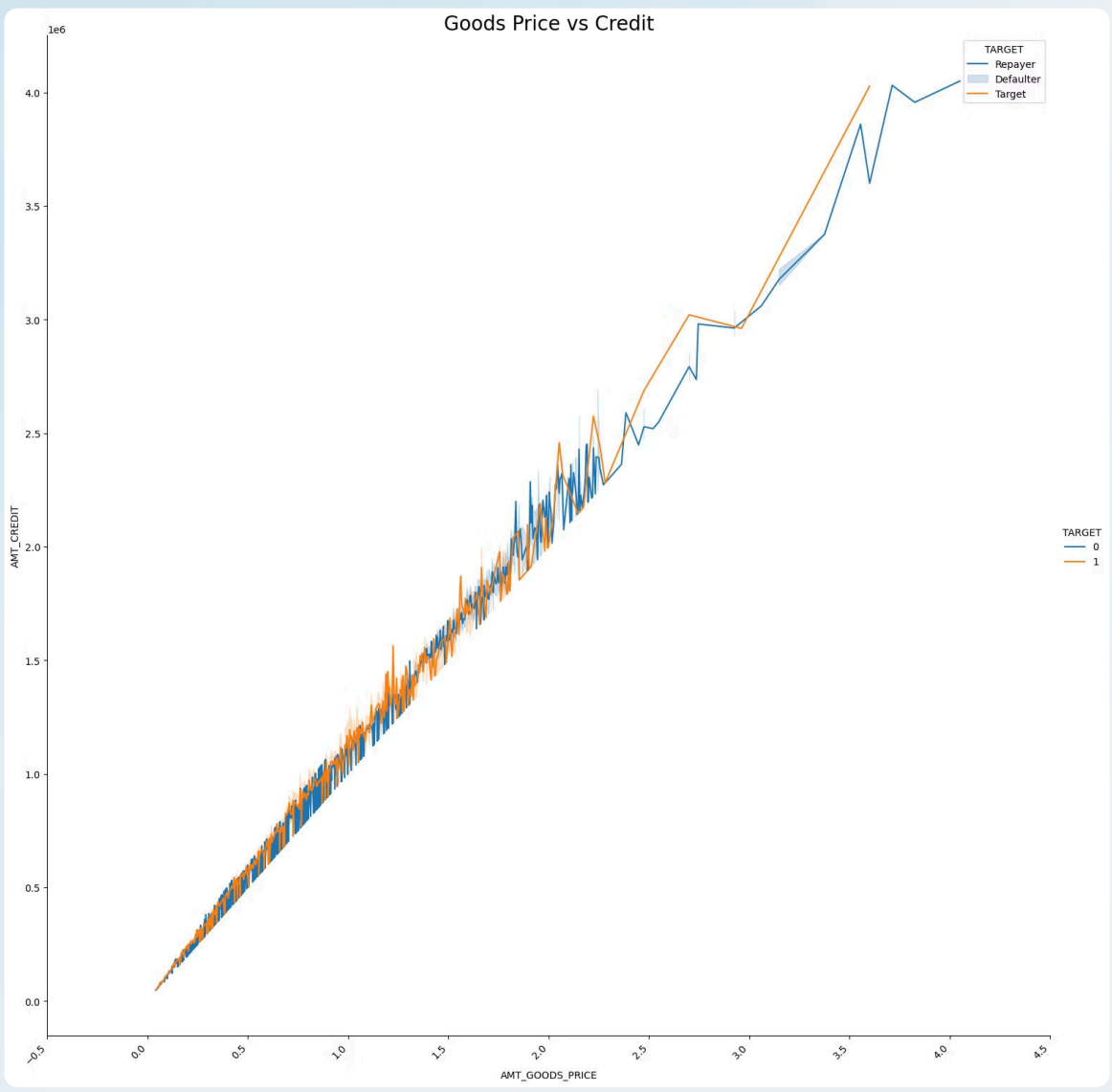


Figure: 11.1

Insights we got when credit **exceeds 30 Lakhs**, defaulters increase. Goods price and credit amount show a linear relationship - higher prices mean higher credit, for both repayers and defaulters.

18) Correlation between numeric variable

A) As we've explored the various factors that can impact a borrower's ability to successfully manage their loan obligations, I felt it would be valuable to take a closer look at the relationship between housing type and repayment status. This analysis could provide key insights to help guide the development of more targeted support programs and policies.

To dive deeper into this area, I've plotted a heatmap to visualize any **correlations** that may exist among different **Repayers** profiles.

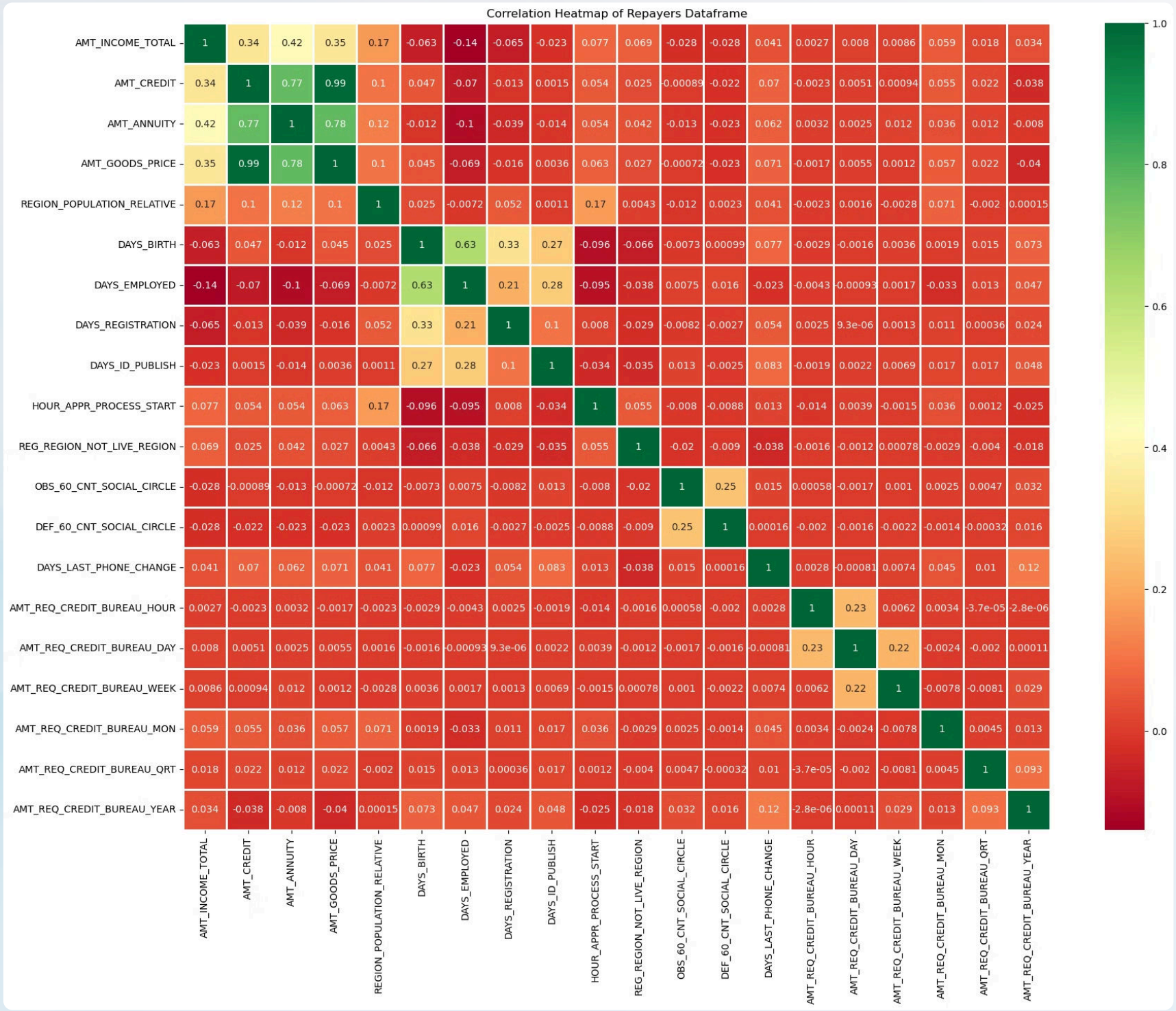


Figure: 12.0

The heatmap analysis revealed several key correlations: - **Higher credit amounts** are associated with **higher annuities** (**0.77**) - **Credit amount** is closely tied to **goods price** (**0.99**) - **Higher annuities** correspond to **higher goods prices** (**0.78**) - **Higher incomes** tend to have **higher annuities** (**0.42**) - **Higher incomes** are somewhat related to **higher credit** (**0.34**) - **Older applicants** are employed **longer** (**0.33**) - **Longer employment** is linked to **longer registration periods** (**0.32**) - **Longer employment** may be slightly associated with **lower credit** (**-0.14**) - **Older individuals** tend to have **lower incomes** (**-0.063**) Most other features show weak or no significant correlations.

B) I felt it would be valuable to take a closer look at the relationship between housing type and repayment status. This analysis could provide key insights to help guide the development of more targeted support programs and policies.

To dive deeper into this area, I've created a heatmap visualization to uncover any **correlations** that may exist among different **Defaulters** profiles.

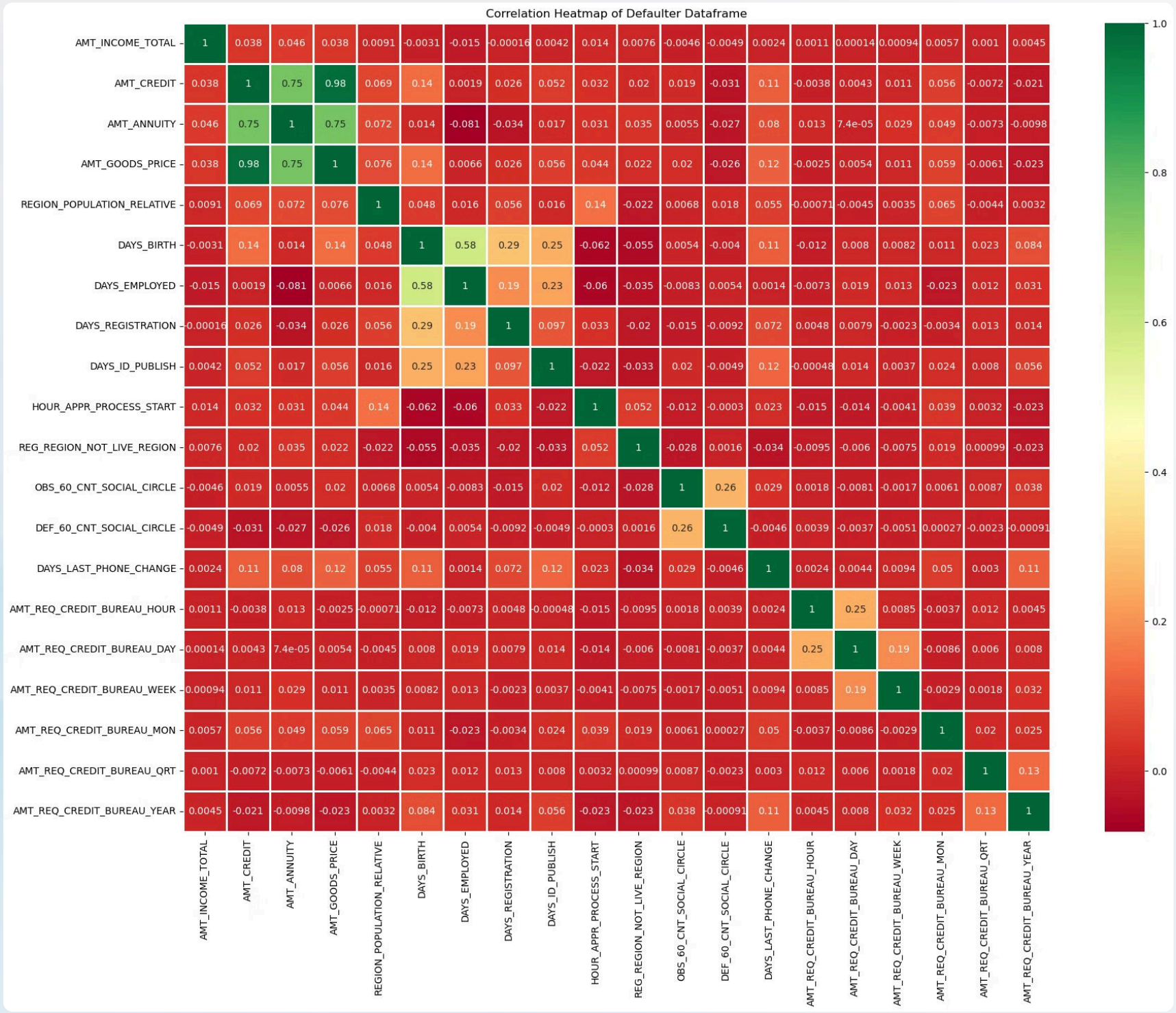


Figure: 12.1

The heatmap analysis revealed several key correlations: - **Higher credit** amounts are associated with **higher goods prices** (**0.77**) - **Higher incomes** tend to have **higher annuities** (**0.42**) - **Older applicants** are employed for **fewer days** (**−0.74**) - **Older individuals** registered for **longer periods** (**−0.71**) However, other features like application approval time and days since last phone change show weak or no significant correlations.

19) Merged data frame

After that I **merged** the '**app_inp0df**' and '**prev_app_inp1**' data frames for more analysis.The new data frame called **'loan_inp3'** had around 1413701 rows and 76 columns.

I created a new function called **'univariate_categorical_merged'** for univariate analysis of the merged data frame. Then I bisected the new dataframe **'loan_inp3'** into "**L0**"(Repayers) and "**L1**"(Defaulters).

A) Then I plotted **Contract status vs Purpose of the loan**. This analysis provides an interesting perspective on how the purpose of a loan may relate to whether the borrower is able to repay it successfully.

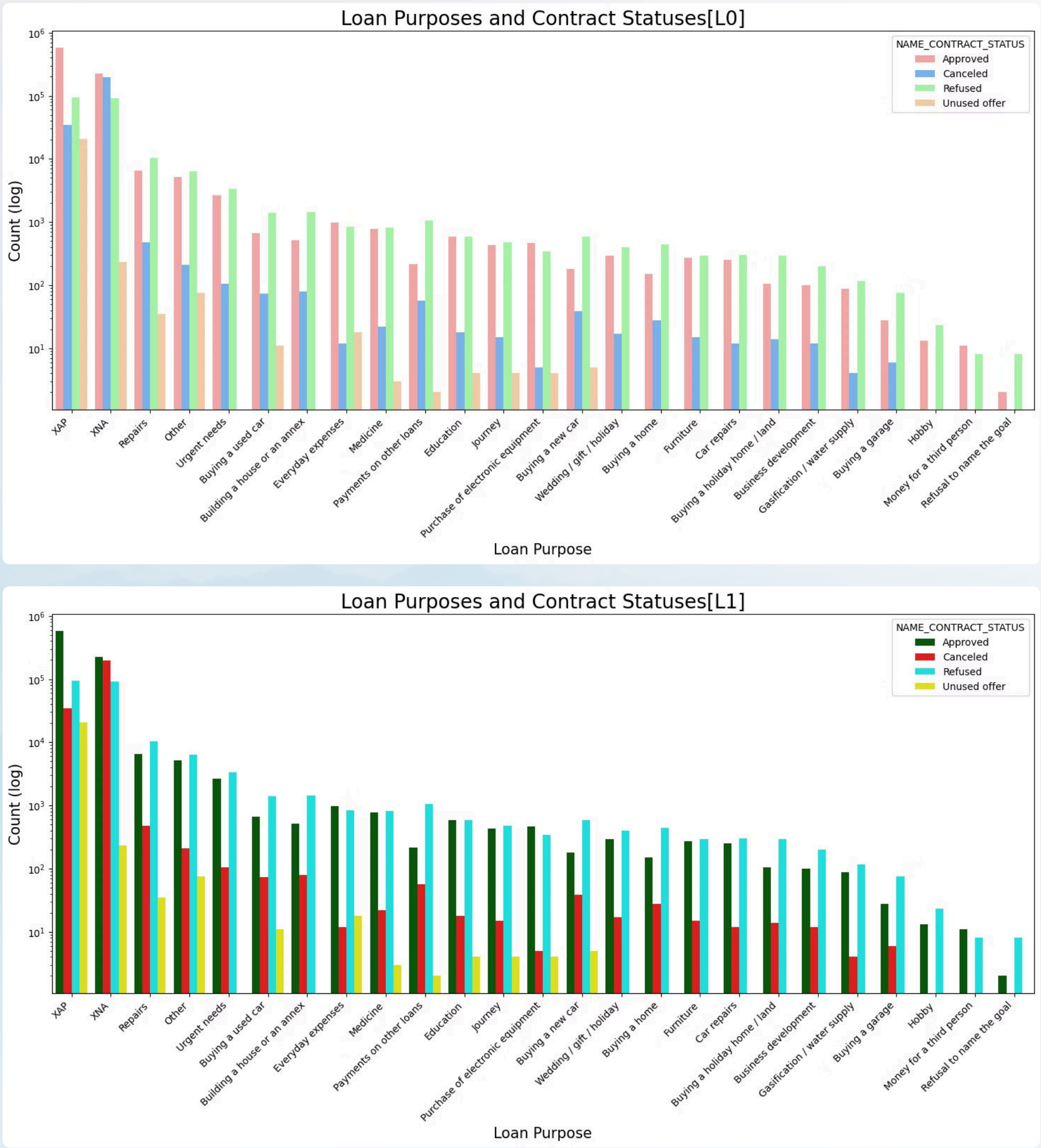


Figure: 13.0-13.1

Insigts I got from this was the most **common loan purposes** are **'XAP'**, **'XNA'**, **'Repairs'**, and **'Other'**. Most purposes have more **'Approved'** statuses than other statuses. **'Refused'** is often the second most common status. **'Canceled'** and **'Unused offer'** have fewer counts, but **'Canceled'** is notable for 'Everyday expenses', 'Buying a new car', and 'Education'. Some niche categories like 'Gasification' have lower overall counts but still more approvals. **'Refusal to name the goal'** has a **high rate** of **'Refused' loans.**

B) After analyzing the relationship between housing type and loan repayment status, I decided to take a closer look at another interesting dynamic - **the connection between a client's social circle and their loan contact status.**

By examining whether people who defaulted on their loans in **the last 60 days are part of the client's social network**, we may be able to uncover patterns that could inform credit decisions. This analysis could shed light on the potential influence of social relationships on financial behavior and loan repayment.

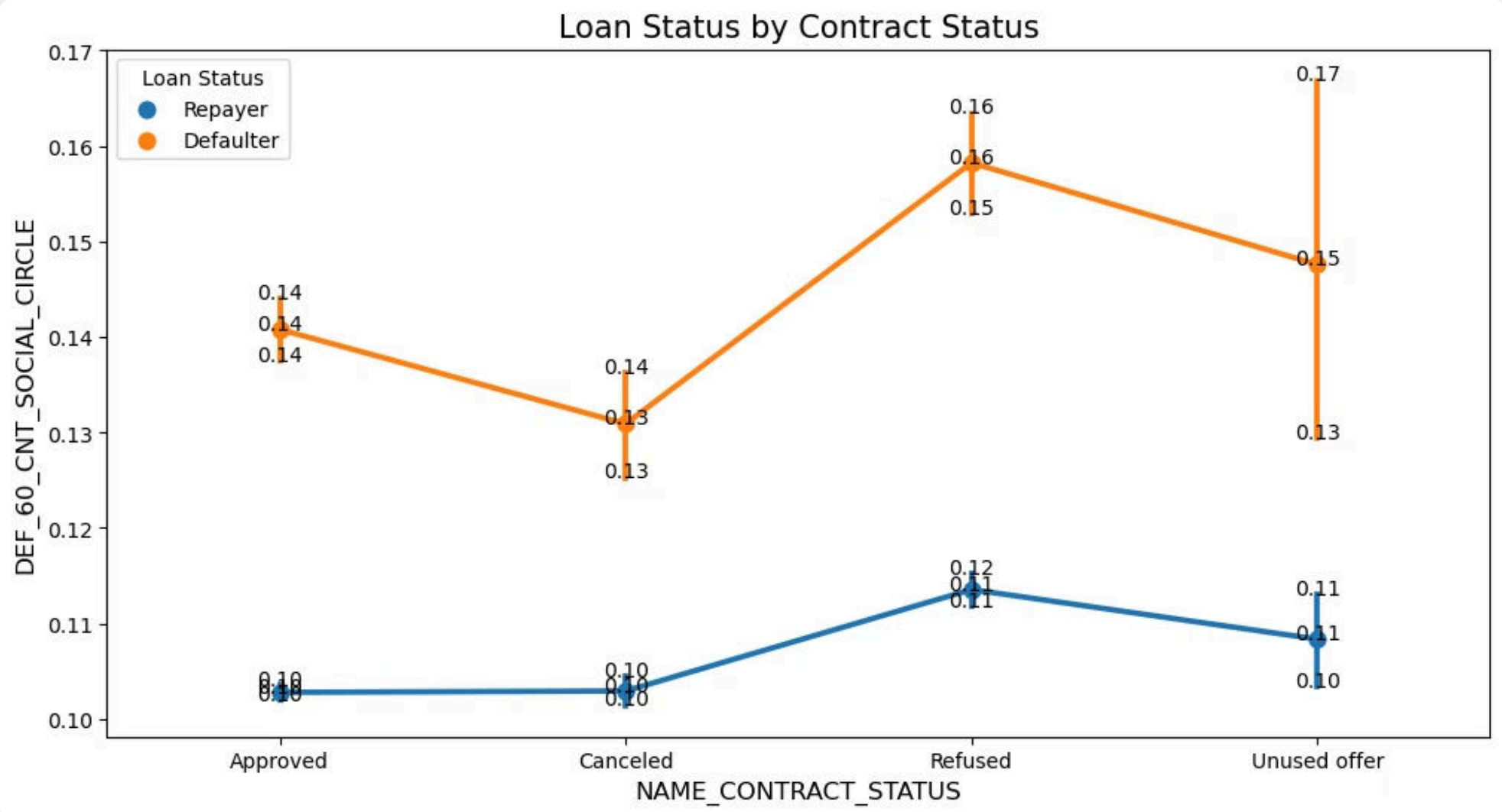


Figure: 13.2

Analysis shows that clients who have **average of 0.13 or higher,** their DEF_60_CNT _SOCIAL_CIRCLE score has **higher** default rate and so analysing client's social circle could help in disbursment of the loan.

C) The data shows an interesting **relationship between a client's total income and their loan contract status**. By plotting this information, I can start to uncover patterns that may inform the lending institution's decision-making process and risk management strategies.

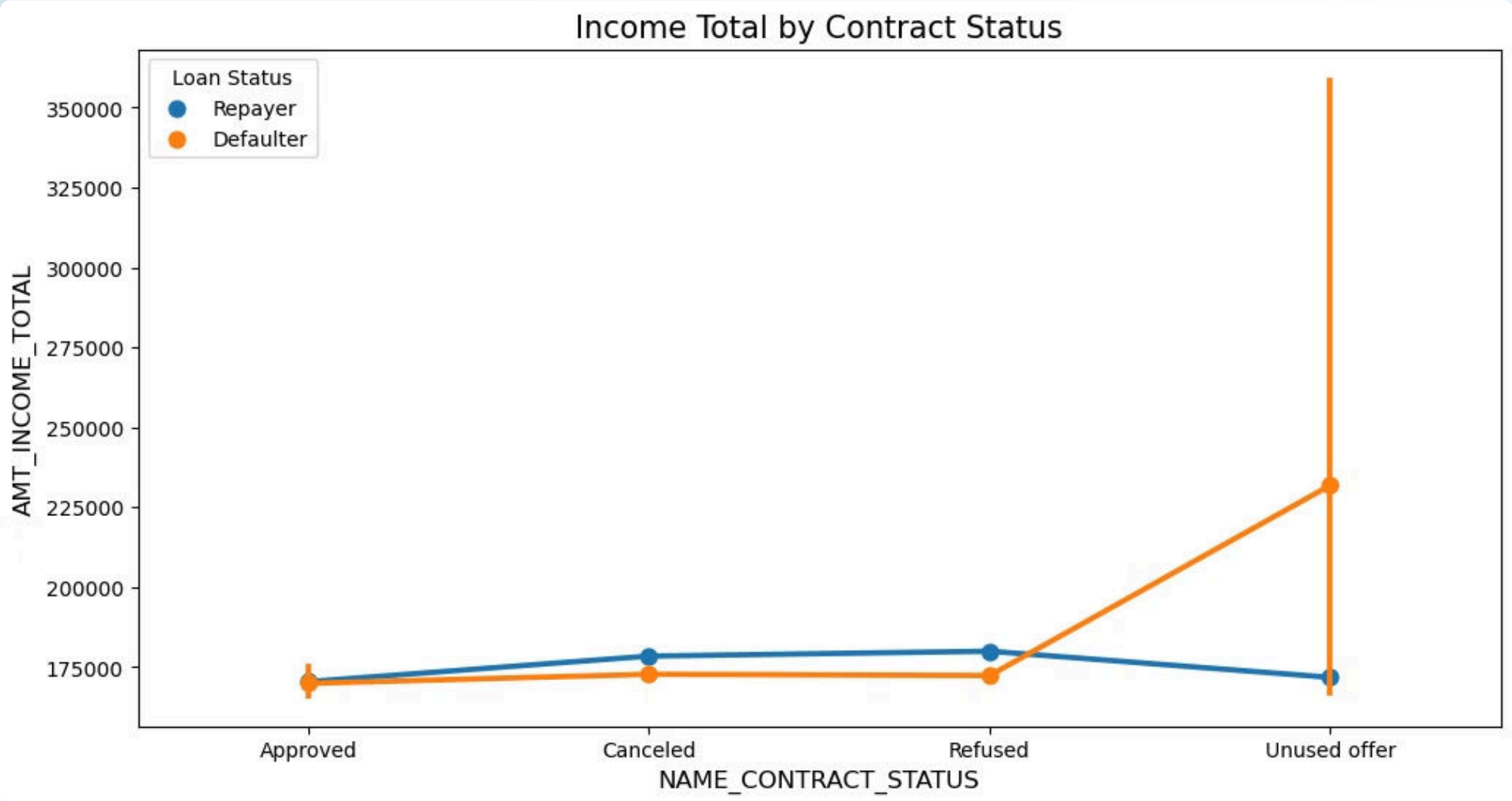


Figure: 13.3

The point plot shows that the people who have not **used offer** earlier have **defaulted** even when there average income is **higher** than other people.

D) After reviewing the previous sections on time management strategies and challenges, I decided to shift the focus of my analysis to the relationship between a client's **housing situation and their loan repayment status**. This is an important factor to consider, as an individual's living arrangements can significantly impact their financial stability and ability to meet loan obligations.

By examining **the contract status based on loan repayment data**, I hope to uncover any potential connections between housing type and financial outcomes.

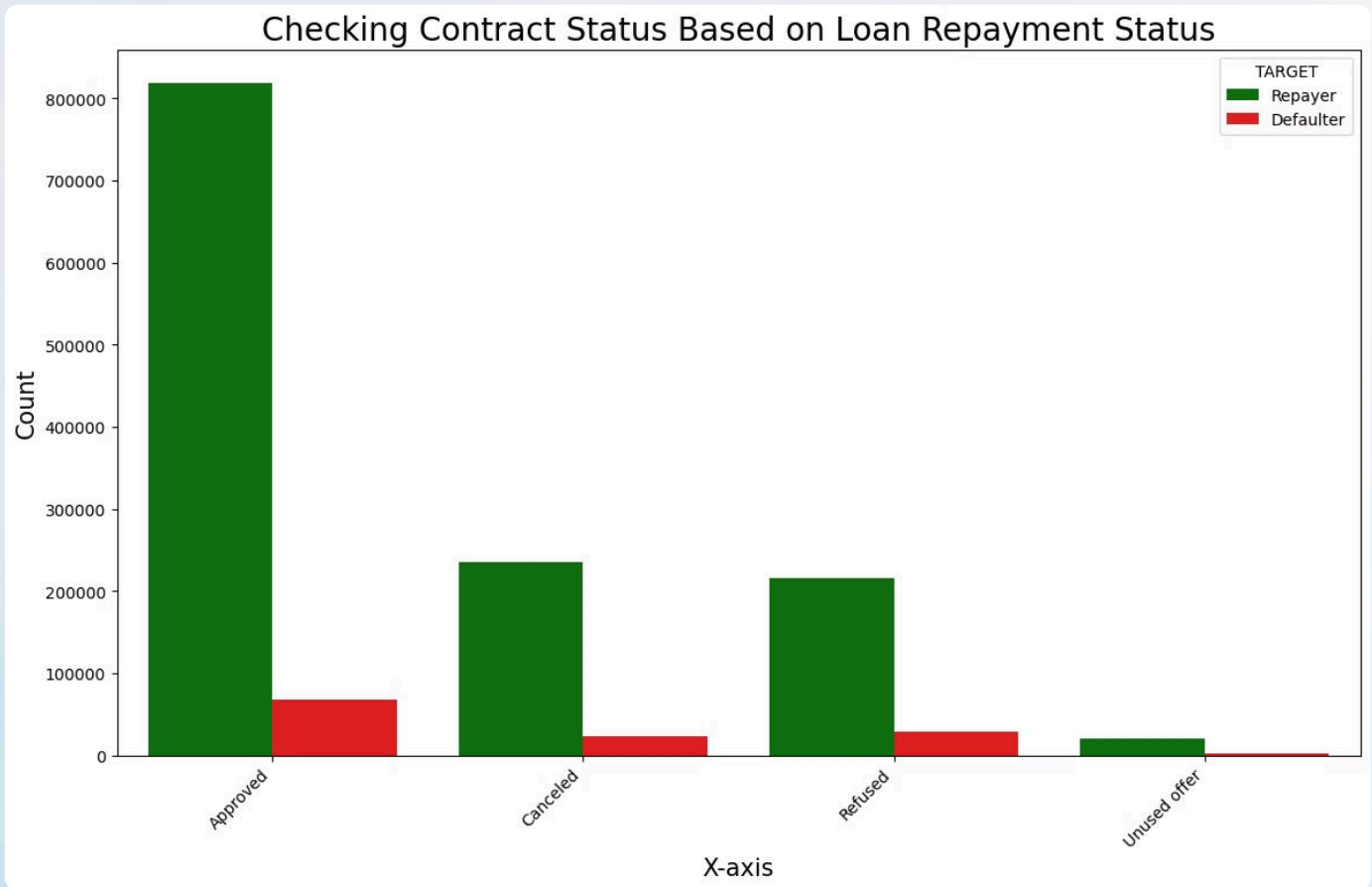


Figure: 14.0

The data reveals a clear dominance of **Repayers** across all contract statuses. **92.41%** of **approved contracts** are held by **Repayers**, indicating their **strong creditworthiness**. In contrast, **Defaulters** account for only **7.59%** of **approved contracts**.

Interestingly, the proportion of **canceled contracts** is also **higher** among **Repayers** at **90.83%**, suggesting that cancellations are more common in this group. However, the **refused contracts** show a different trend, with **12.0%** held by **Defaulters** compared to **88.0%** by **Repayers**.

The distribution of **unused offers** follows a similar pattern, with **Repayers** holding **91.75%** and **Defaulters** accounting for **8.25%**. This consistent trend across contract statuses highlights the **Repayer dominance** and the **stringency of the approval process** in limiting Defaulters.

20) Conclusions:

After analysing the datasets **app_inp0df and prev_app_inp1 and loan_inp3**, there are few attributes of a client with which the bank would be able to identify if they have chances to became defaulters or not. The analysis is thoroughly checked, and some important points are in below with the contributing factors and categorization:

Factors Contributing to Repayment:

- **Educational Attainment:** Applicants with academic degrees exhibit lower default rates, indicating higher repayment reliability.
- **Income Sources:** Students and businessmen show no defaults, establishing them as dependable borrowers.
- **Industry and Business Type:** Trade Type 4, Trade Type 5, and Industry Type 8 show default rates below 3%, highlighting their reliability.
- **Age and Experience:** Applicants above 50 years old and clients with over 40 years of experience demonstrate a lower probability of defaulting.

Factors Contributing to Default:

- **Gender:** Men tend to default more frequently compared to women.
- **Marital Status:** Civil marriage or single status correlates with higher default rates.
- **Education Level:** Lower Secondary and Secondary education levels are associated with higher default probabilities.
- **Income Source:** Maternity leave and Unemployed statuses show higher default rates.

Mitigating Default Risks:

- **Offer higher interest loans** for applicants living in rented apartments or with parents to offset potential losses from defaults.
- Implement **higher interest rates** for loans in the 300,000 to 600,000 credit range, which tend to have higher default rates.
- Apply **higher interest rates** to loans for applicants with a total income less than 300,000, as they exhibit higher default probabilities.
- Offer **higher interest loans** or reject applications from clients with 4 to 8 children, who show a very high default rate.
- Maintain **rejection or higher interest** for loans taken for repairs, as they have the highest default rates.

Optimizing the Loan Portfolio:

- **Revisit interest rates or loan terms** for clients with previous loan refusals or cancellations to capitalize on their improved repayment behavior.
- Clients with a history of loan refusals now demonstrate a **high repayment rate (88%)**, indicating potential for revised risk assessments.
- Leverage insights from past data to **convert historically high-risk clients into reliable borrowers**, enhancing business opportunities and minimizing risk.
- Reassess risk models based on these insights to **further optimize loan approval processes** and improve overall portfolio performance.