BANK LOAN CASE STUDY

Dency Claris Thomas

Project Description:

The Bank Loan Case Study project involves the analysis of a dataset related to bank loan applications and client information. The primary objective of this project is to gain insights into the factors that differentiate clients with payment difficulties from those without difficulties. By examining various numerical and categorical variables, the project aims to identify patterns, correlations, and important features that influence loan repayment.

The dataset includes information about clients' demographics, such as age, gender, income, education level, and family status. It also provides details about the loan applications, including loan amount, contract type, portfolio, channel type, and application status.

The project utilizes various statistical techniques and visualizations to explore the dataset and extract meaningful insights. Univariate analysis is conducted to understand the distribution and characteristics of individual variables. Segmented univariate analysis is performed by considering the target variable (payment difficulties) to identify any specific patterns within different segments.

Bivariate analysis is conducted to examine relationships between variables and uncover any correlations or associations. This helps in understanding how different factors, such as income level, loan amount, and client type, relate to payment difficulties. Visualizations, such as histograms, scatter plots, box plots, and bar charts, are employed to effectively communicate the analysis findings and highlight key insights.

The project report summarizes the most important results and provides actionable recommendations based on the analysis. It highlights variables that are significant in differentiating clients with payment difficulties from those without difficulties. The report aims to assist financial institutions in making informed lending decisions, improving risk assessment models, and implementing strategies to minimize default rates.

Overall, the Bank Loan Case Study project offers valuable insights into loan repayment behavior and provides guidance for financial institutions to manage credit risk effectively.

Approach:

The Bank Loan Case Study project followed a systematic approach to analyze the dataset and derive insights. The steps included data preprocessing, exploratory data analysis, segmented analysis, and visualization. Missing values were handled, and data cleaning techniques were applied. Univariate analysis provided initial insights, while segmented analysis examined different segments' association with payment difficulties. Bivariate analysis explored relationships between variables. Visualizations were used to effectively present findings. The results were summarized and interpreted to provide actionable recommendations, supporting credit risk management decisions.

Tech Stack Used:

The Bank Loan Case Study project was implemented using Python programming language in JupyterLab environment. Python libraries such as Pandas, NumPy, and Matplotlib were utilized for data manipulation, analysis, and visualization. These powerful tools provided the necessary functionality to preprocess the dataset, perform exploratory data analysis, and generate informative visualizations. The combination of Python and its associated libraries enabled efficient and effective analysis of the bank loan data.

Analysis:

Task 1: Present the overall approach of the analysis. Mention the problem statement and the analysis approach briefly

- ✓ We want to understand the factors that influence loan default among clients applying for loans.
- ✓ To achieve this, we will perform exploratory data analysis (EDA) on the provided dataset.
- ✓ Our goal is to identify patterns and variables that can help differentiate clients with payment difficulties from all other cases.

Task 2: Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

Hint: Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.

In this task, we focused on cleaning the dataset and handling missing values in order to ensure the integrity and quality of the data for further analysis.

1. ***Identifying and Removing Useless Columns***: We started by identifying and removing columns that contained irrelevant or redundant information, as they would not contribute significantly to the analysis. These columns were dropped from the dataset.
2. ***Treating Missing Values***
   - ✓ *Ext_Source_2 and Ext_Source_3*
     We observed that there were missing values in the columns "Ext_Source_2" and "Ext_Source_3." After analyzing the statistics of these columns, we found that the difference between their means was narrow, and the variation from the 25th to the 75th percentile was evenly distributed. Hence, we decided to impute the missing values with the mean values of "Ext_Source_2" and

"Ext_Source_3" respectively. This ensured that the missing values were replaced with plausible estimates based on the available data.

- ✓ *AMT_Goods_Price*
  For the column "AMT_Goods_Price," we noticed a high standard deviation and a significant number of outliers. Imputing the missing values with the mean or median would introduce bias. Therefore, we chose to remove the rows with missing "AMT_Goods_Price" values from the dataset. These rows accounted for only 0.09% of the total records and had a minimal impact on the overall analysis.
- ✓ *AMT_Req_Credit_Bureau_QRT*
  In the column "AMT_Req_Credit_Bureau_QRT," which represents the number of enquiries to the Credit Bureau about the client within a quarter, we converted the data type to "category" and examined the frequency of each category. We found that the most common value was 0. Thus, we imputed the missing values in this column with 0.

Results:

After performing the necessary data cleaning steps and handling missing values, we successfully addressed the issue of missing data in the dataset. The remaining columns were checked, and no missing values were found.

The final cleaned dataset contains 307,004 rows, which accounts for a loss of only 0.16% of the original data. The cleaned dataset is now ready for further analysis.

Task 3: Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.

In this task, we focused on transforming the data to improve readability and analysis. We converted the 'DAYS_BIRTH' column into the 'AGE' column for better interpretation and converted the 'DAYS_EMPLOYED' column into the 'YEARS_EMPLOYED' column. Additionally, we addressed outliers in the dataset using the 1.5 IQR (Interquartile Range) rule.

First, we performed the following transformations:

1. *Converting 'DAYS_BIRTH' into 'AGE' column:*
   - We divided the 'DAYS_BIRTH' values by 365 to convert them into years.
   - The absolute value was taken to ensure positive values.
   - The resulting values were assigned to the newly created 'AGE' column.

2. *Converting 'DAYS_EMPLOYED' into 'YEARS_EMPLOYED' column:*

- Similar to the previous step, we divided the 'DAYS_EMPLOYED' values by 365 to convert them into years.
- The absolute value was taken to ensure positive values.
- The resulting values were assigned to the newly created 'YEARS_EMPLOYED' column.

After the transformations, the 'DAYS_BIRTH' and 'DAYS_EMPLOYED' columns were dropped from the dataset using the `drop` function. The resulting dataset, `df_application_current`, now contains additional columns 'AGE' and 'YEARS_EMPLOYED', which provide more meaningful information for analysis.

Next, we addressed outliers in the dataset using the 1.5 IQR rule. The following steps were taken:

1. Defining a list, `col_list_outlier`, containing the numerical columns: 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', and 'AGE'.
2. For each column in `col_list_outlier`, we calculated the first quartile (q1), third quartile (q3), and interquartile range (iqr) using the `quantile` function.
3. We determined the lower and upper bounds for outliers using the formula: range_low = q1 - 1.5 * iqr and range_high = q3 + 1.5 * iqr.
4. Using the `loc` function, we filtered the dataset to include only the records where the values in the current column were within the defined range.

By applying this process to each column in `col_list_outlier`, we removed the outlier records from the dataset.

Finally, we checked the shape of the transformed dataset using the `shape` attribute, which returned (275,984, 48), indicating that we retained 275,984 rows and 48 columns after removing the outliers.

It is worth noting that approximately 10.25% of the rows were lost during the outlier handling exercise. Although a portion of the data was discarded, this step ensured the removal of outliers for a fair analysis of the remaining dataset.

Task 3: Identify if there is data imbalance in the data. Find the ratio of data imbalance.

Hint: Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights.

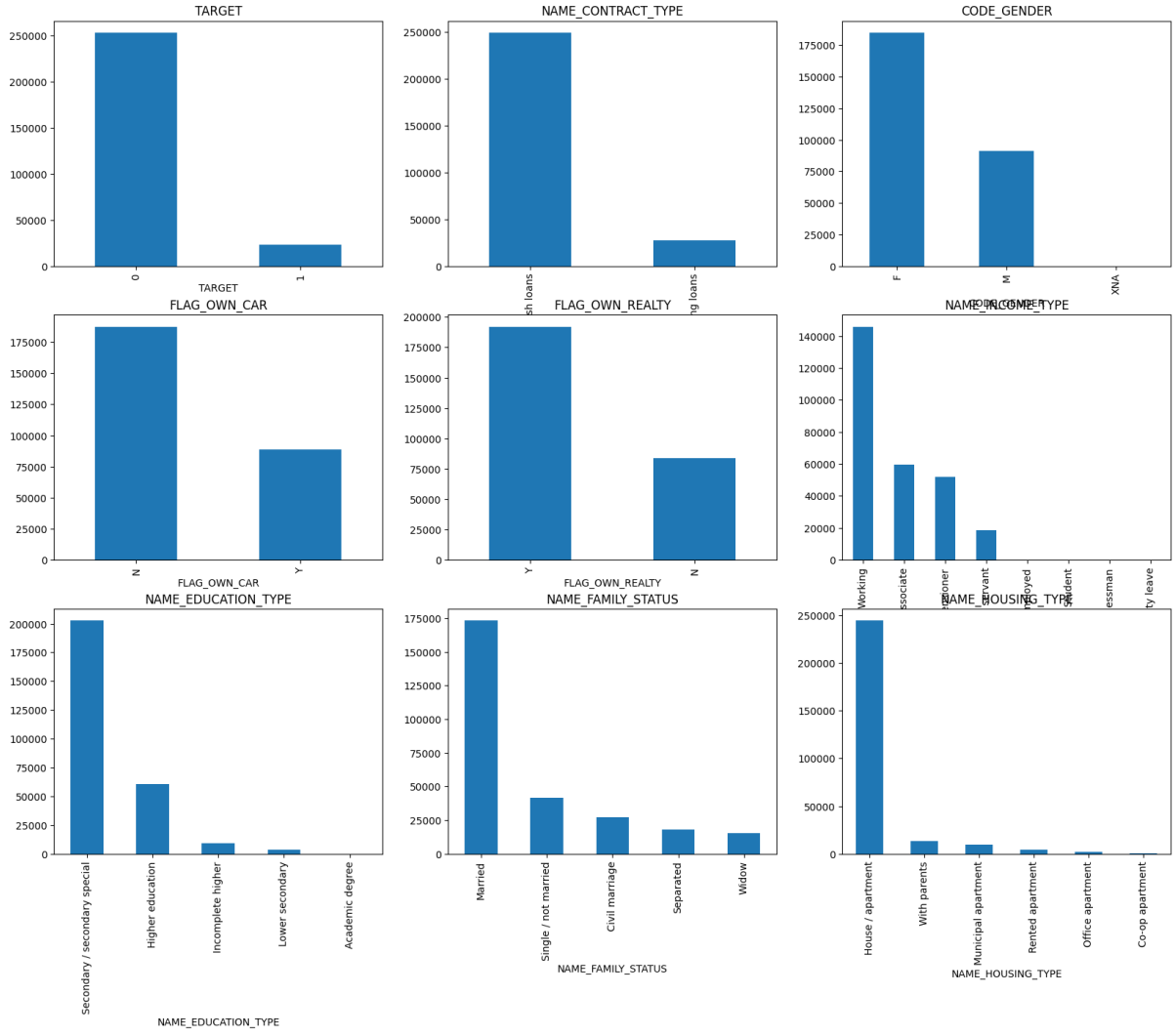
In this task, we continued the data transformation process by addressing data imbalance in certain columns and removing unnecessary columns.

First, we checked for data imbalance in the following columns: 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', and 'NAME_HOUSING_TYPE'.

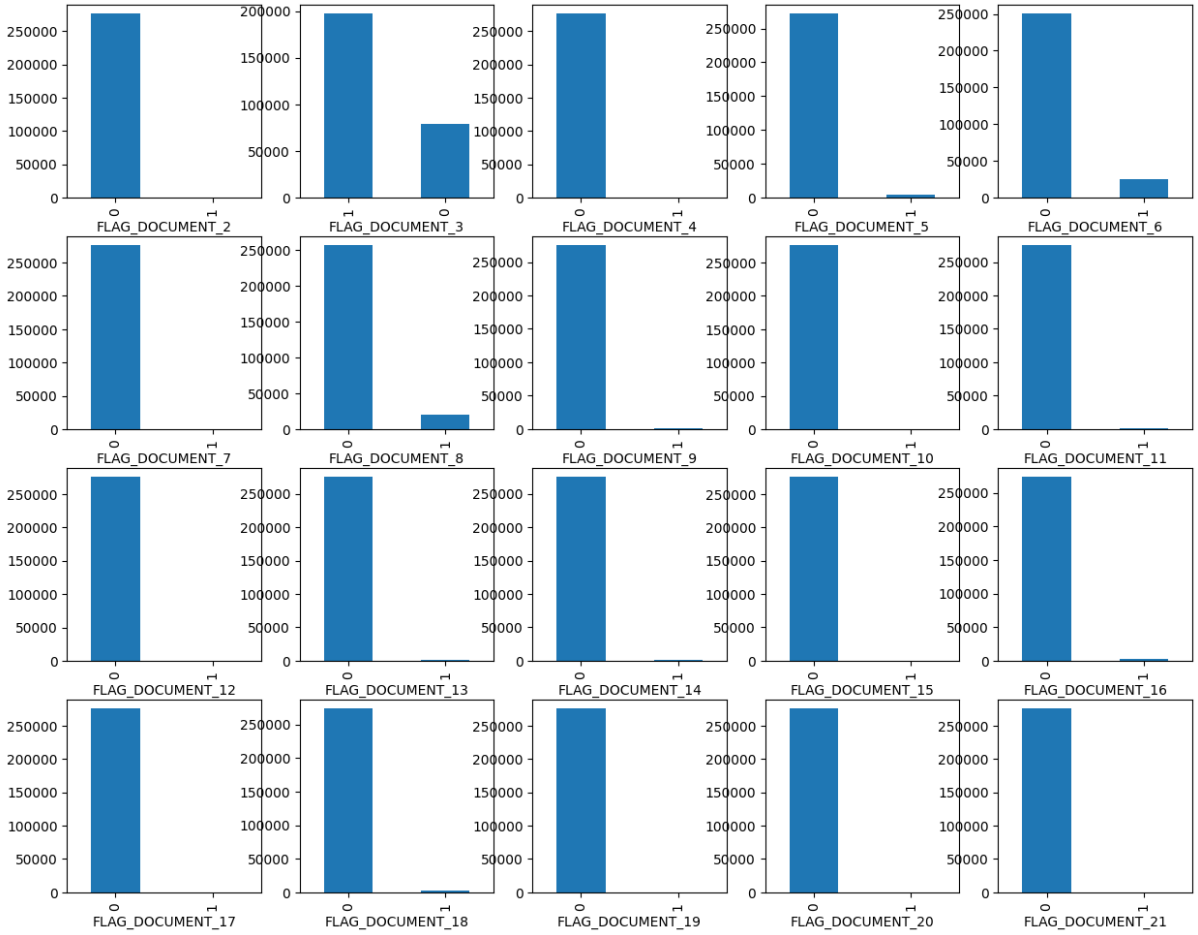
To visualize the data imbalance, we plotted bar charts for each of these columns using the `value_counts()` function and the `plot` function from the matplotlib library. Each subplot in the figure represents one column, and the bar chart shows the counts for each category within the column.

Based on the bar charts, we observed the following data imbalances:

- TARGET: There are very few defaulters (1) compared to non-defaulters (0).
- NAME_CONTRACT_TYPE: There are very few revolving loans compared to cash loans.
- NAME_EDUCATION_TYPE: Most of the loans were applied by individuals with a secondary/secondary special education.
- NAME_FAMILY_STATUS: Most of the loans were applied by married individuals.
- NAME_HOUSING_TYPE: Most of the loan applications came from homeowners or individuals living in apartments.



Next, we focused on analyzing the 'FLAG_DOCUMENT' columns to check for data imbalance. We plotted bar charts for each 'FLAG_DOCUMENT' column (from 'FLAG_DOCUMENT_2' to 'FLAG_DOCUMENT_21') to visualize the distribution of values. It was observed that, except for 'FLAG_DOCUMENT_3', all the columns had a negligible count of 1s. Therefore, we decided to remove all the 'FLAG_DOCUMENT' columns except 'FLAG_DOCUMENT_3' from the dataset.



To remove the unnecessary columns, we used the `drop` function from pandas. We removed 'FLAG_DOCUMENT_2' using the `drop` function with the 'axis' parameter set to 1 (indicating column removal). For the remaining 'FLAG_DOCUMENT' columns ('FLAG_DOCUMENT_4' to 'FLAG_DOCUMENT_21'), we iterated over the range of column indices and used the `drop` function to remove each column individually.

By removing the unnecessary 'FLAG_DOCUMENT' columns, we streamlined the dataset for further analysis and model building.

These transformations and data imbalance handling steps contribute to improving the quality and relevance of the dataset, setting the stage for more accurate analysis and modelling.

I performed some data preprocessing steps before conducting my analysis.

1. Binning of columns: You binned the columns 'AGE', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', and 'EXT_SOURCE_SCORE' into different groups to simplify the analysis.
2. Age group categorization: You created three age groups: 'Young', 'Mid Age', and 'Senior Citizen' based on the 'AGE' column.
3. Credit amount group categorization: You created three credit amount groups: 'High', 'Medium', and 'Low' based on the 'AMT_CREDIT' column.
4. Income group categorization: You created three income groups: 'High', 'Medium', and 'Low' based on the 'AMT_INCOME_TOTAL' column.
5. External source score calculation: You calculated the average of the 'EXT_SOURCE_2' and 'EXT_SOURCE_3' columns and created a new column called 'EXT_SOURCE_SCORE'.

These preprocessing steps can help simplify and categorize the data, making it easier to analyze and interpret.

Task 4: Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

In this task, we will discuss the results of the analysis performed on a dataset related to previous loan applications. The analysis includes univariate analysis, segmented univariate analysis, bivariate analysis, and insights derived from the data. The goal is to provide meaningful insights in business terms based on the findings.

1. ***Univariate Analysis:***

   Univariate Analysis for Unordered Categorical Variables:

   1. Gender Distribution:
   - Defaulters: Slightly higher number of females compared to males.
   - Non-defaulters: Similar representation of females and males.
   - Insight: Gender alone may not be a strong predictor of loan default.

   2. Loan Type Distribution:
   - Both defaulters and non-defaulters show a preference for Cash loans over Revolving loans.
   - Insight: Most loan applications are for Cash loans rather than Revolving loans.

   3. Profession Distribution:
   - Defaulters: "Working" profession has the highest number.
   - Non-defaulters: Majority are also from the "Working" profession.

- Insight: Occupation plays a role in loan repayment behavior.

4. Education Level Distribution:
- Defaulters: Majority have secondary or secondary special level of education.
- Non-defaulters: Similar distribution as defaulters.
- Insight: Education level is a factor in loan default rates.

5. Marital Status Distribution:
- Defaulters and non-defaulters both have a majority of married individuals.
- Insight: Marital status alone may not be a reliable indicator of creditworthiness.

Univariate Analysis for Continuous Variables:

1. Loan Amount Distribution:
- Defaulters: Higher likelihood of default for lower loan amounts (up to 500,000).
- Non-defaulters: Lower likelihood of default for higher loan amounts.
- Insight: Lower loan amounts are associated with higher default rates**.**

2. Income Level Distribution:
- Defaulters: No significant gender-based difference observed.
- Non-defaulters: Females have higher proportion in lower income levels, but males dominate in higher income levels.
- Insight: Income level affects default rates, with gender differences in different income ranges.

3. Loan Annuity Distribution:
- Concentration of loan annuity payments between 10,000 and 40,000 for both defaulters and non-defaulters.
- Insight: Common range for loan repayment amounts among the dataset.

Previous Application Data Analysis:

1. Loan Status Distribution:
- Majority of previous loan applications were approved, followed by few refused loans. Cancelled and unused offer loans were minimal.
- Insight: High approval rate and low number of refused applications**.**

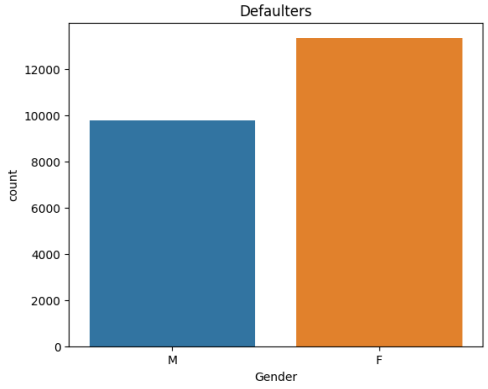
2. Previous Application Channels Distribution:
- Most previous applications were made through "Country-wide" and "Credit and Cash offices" channels, followed by "Stone" channel.
- Insight: Varying popularity of different application channels.

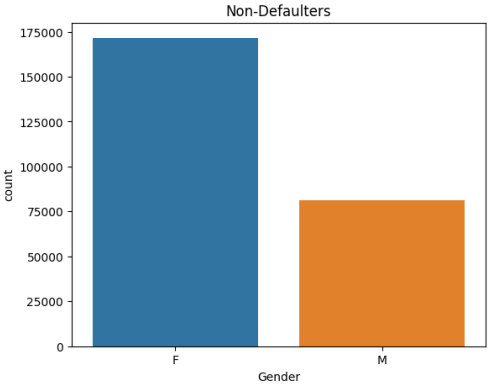
3. Previous Application Amount Distribution:
- Majority of previous applications had loan amounts below 250,000.

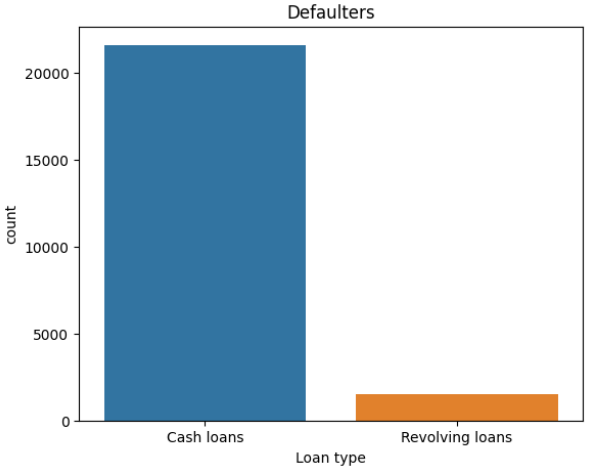
- Insight: Demand for smaller loan amounts and need for tailored lending strategies.

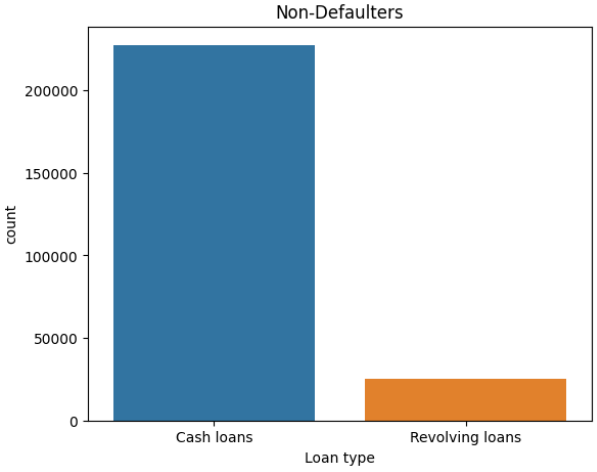
4. Decision Time Distribution:

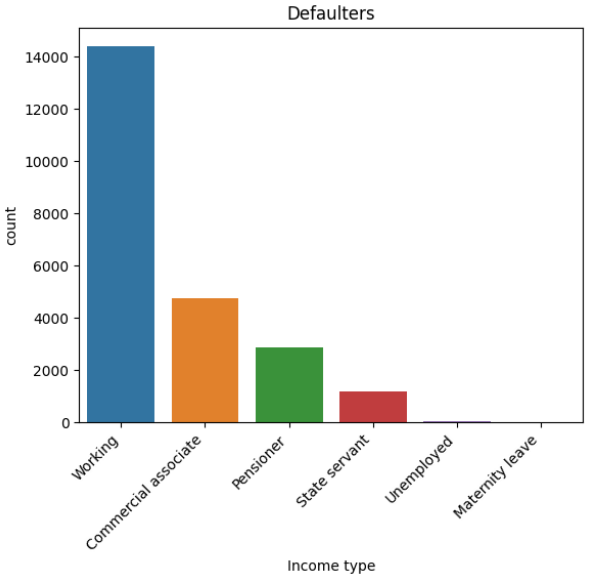
- Majority of previous applications had a decision time of around 30 months.
- Insight: Efficient application processing for timely decisions.

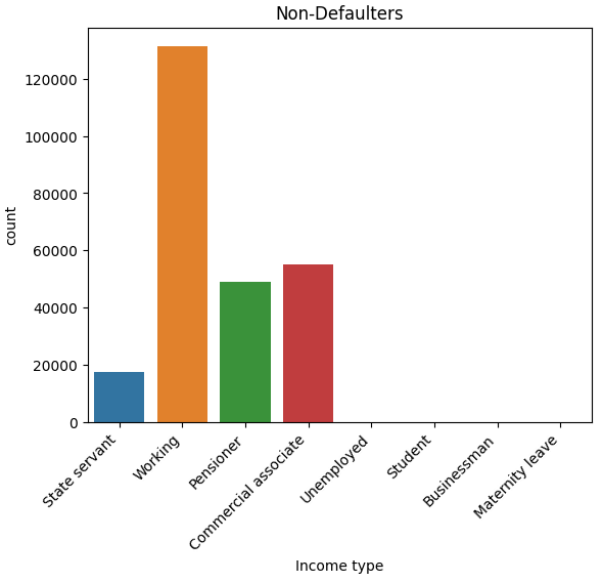


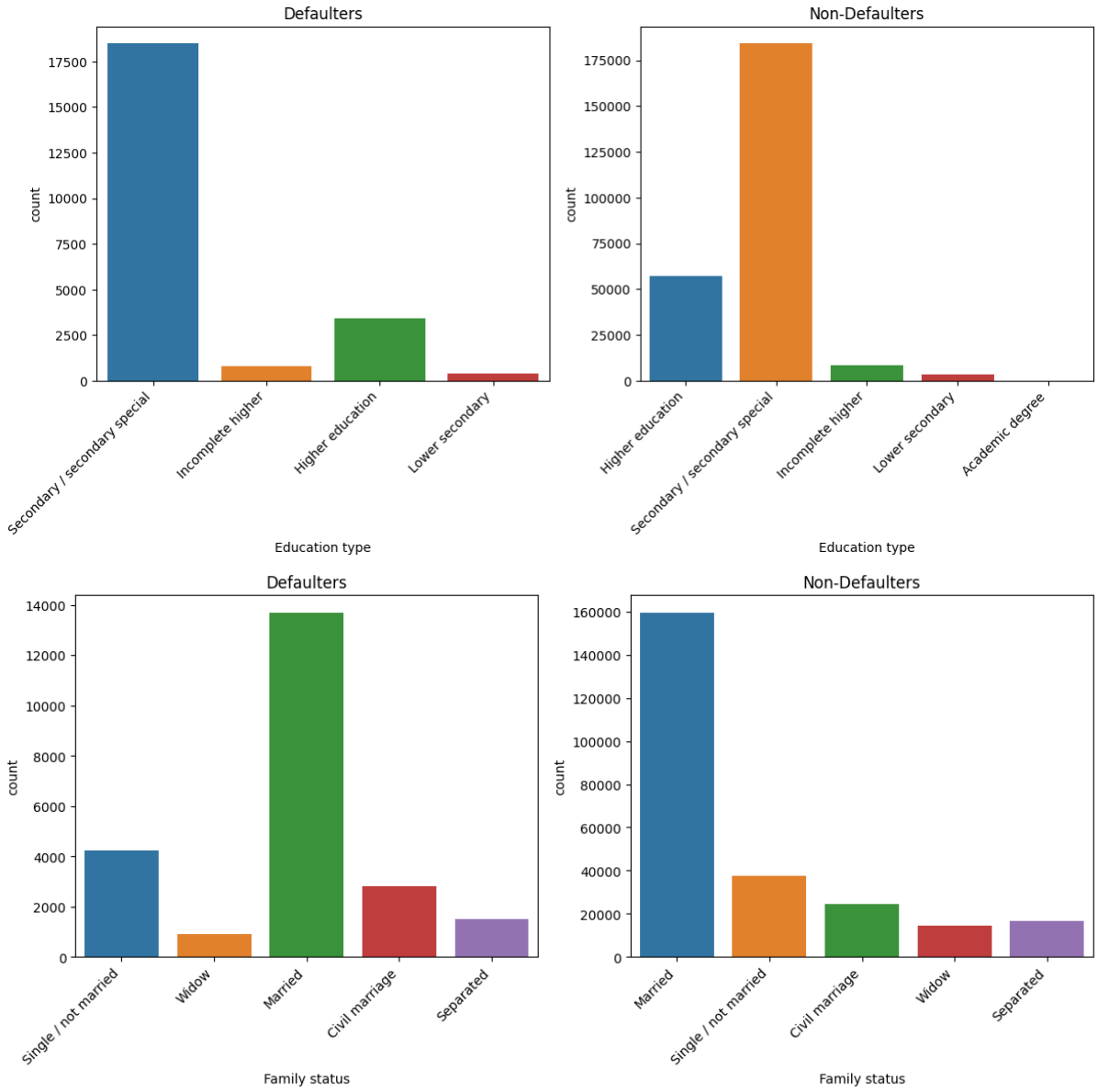












2. Segmented Univariate Analysis:

a) Age Group Analysis:

- Defaulters: Younger individuals have a higher likelihood of default compared to other age groups. Senior citizens are less likely to default.
- Non-defaulters: Age does not significantly affect the likelihood of default among non-defaulters.
- Insight: Age plays a role in default likelihood, with younger individuals facing more financial challenges and instability in income. Senior citizens exhibit higher financial stability, resulting in a lower default rate.

b) Credit Amount Analysis:

- Defaulters: Lower credit amounts are associated with a higher risk of default.

- Non-defaulters: Similar to defaulters, non-defaulters also tend to have lower credit amounts.
- Insight: Lower credit amounts pose a higher risk of default, likely due to limited financial resources, higher debt burden, or less favourable credit histories.

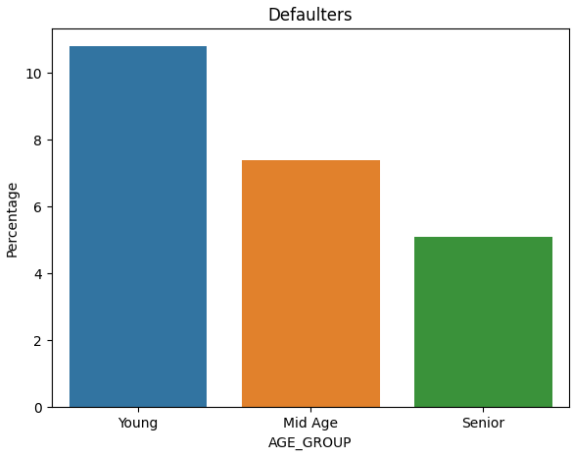
c) Income Group Analysis:

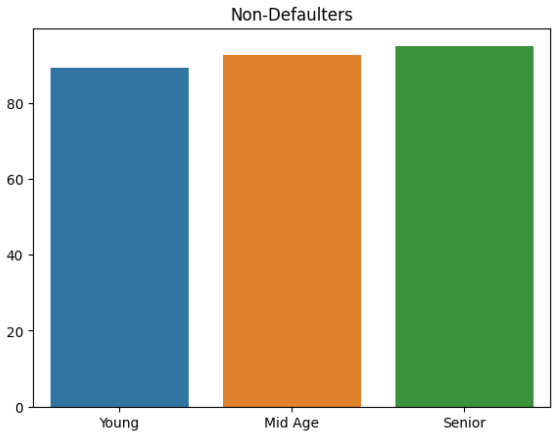
- Defaulters: Higher-income individuals surprisingly have a higher default rate compared to medium and low-income groups. The low-income group has a relatively lower count of defaulters.
- Non-defaulters: non-defaulters are more common in the high-income group and less common in the low-income group.
- Insight: Income level alone does not determine loan default risk. Other factors, such as financial management habits and personal circumstances, play a significant role.

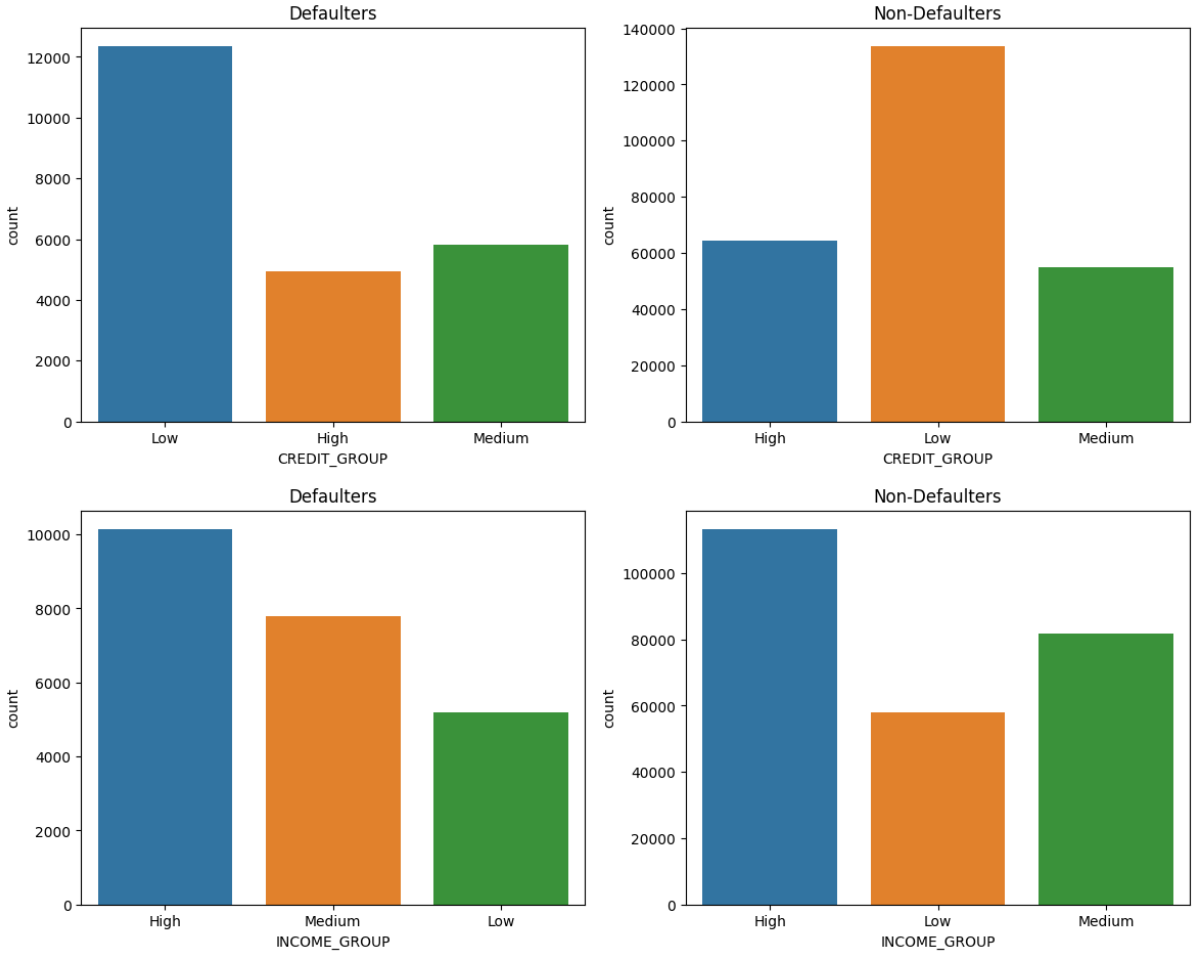
d) External Source Score Analysis:

- Defaulters: Lower scores from external data sources (EXT_SOURCE_2 and EXT_SOURCE_3) are associated with a higher likelihood of default. Most defaulters have low scores, but the medium score range also exhibits a similar default rate.
- Non-defaulters: Non-defaulters have a more balanced distribution across score ranges, with medium to high scores being common.
- Insight: Lower external source scores indicate a higher default risk. However, medium scores do not guarantee a lower default risk. Considering other factors alongside credit scores provides a more comprehensive evaluation of an applicant's repayment ability.

These insights can assist financial institutions in developing targeted risk assessment strategies, tailored loan products, and support programs that address the specific needs and risk profiles of different segments.

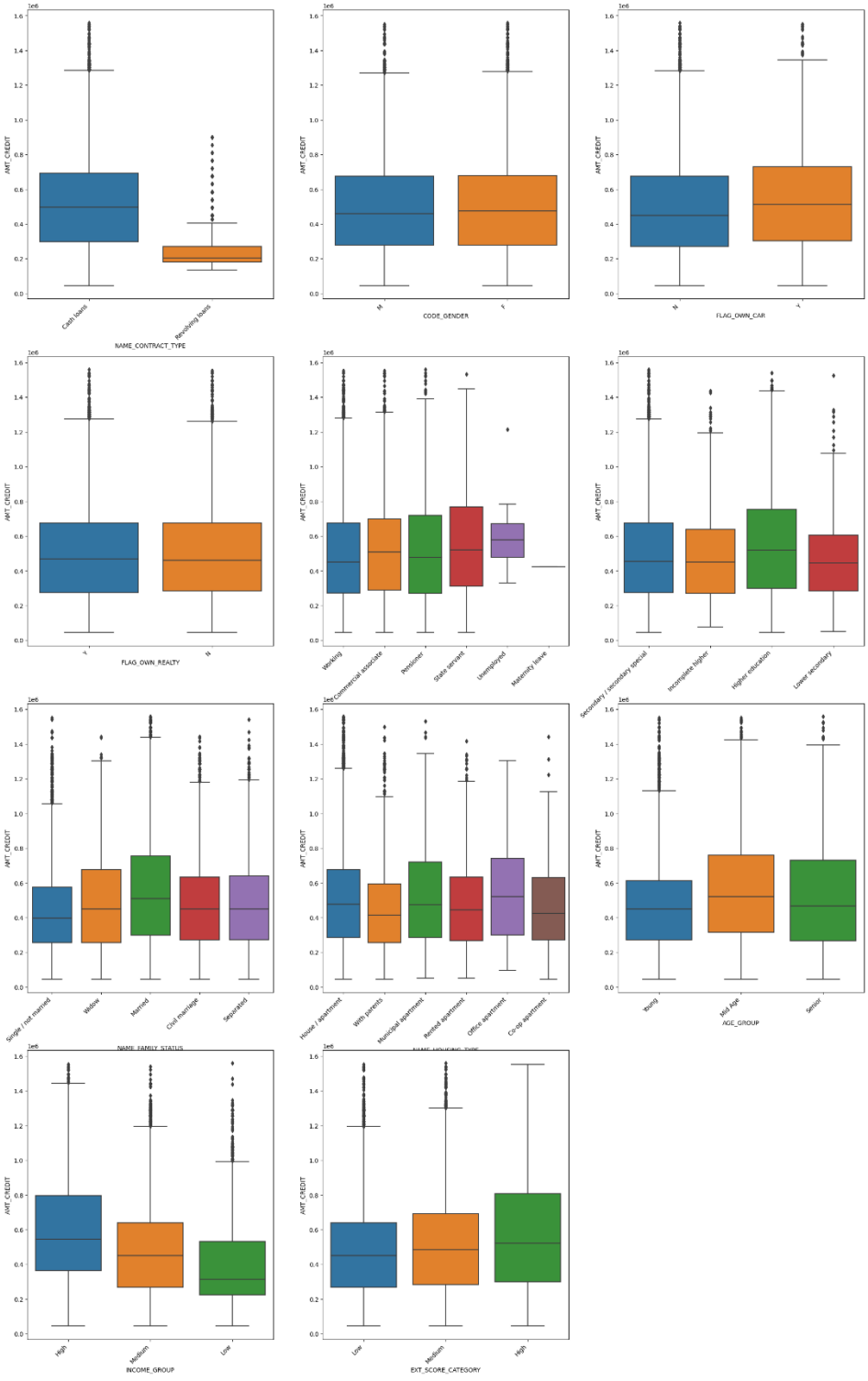


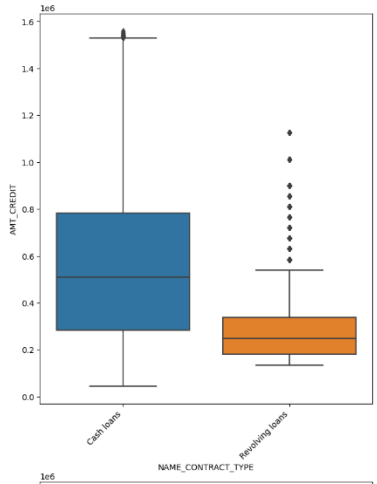


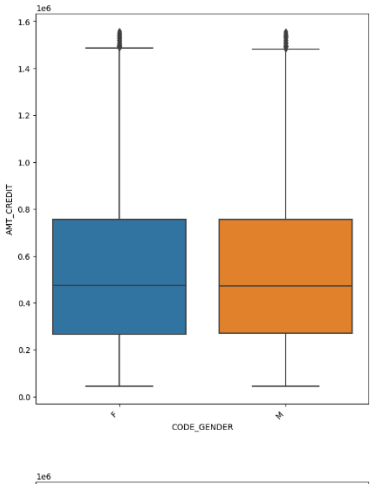


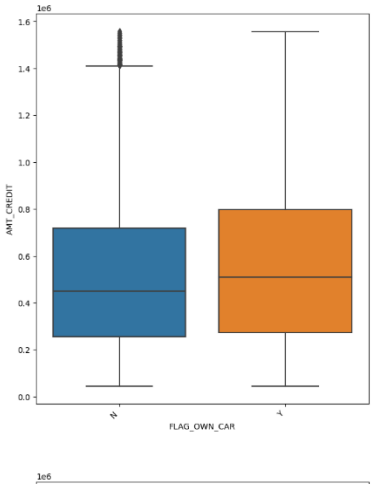
3. Bivariate Analysis:

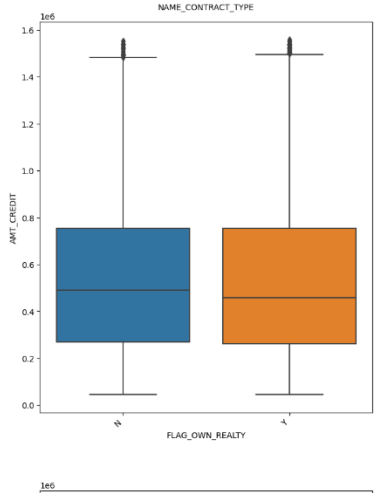
- Previous Application Portfolio: The analysis shows that the POS portfolio has the highest number of previous applications, followed by Cash. Applications for Cards were significantly lower. This suggests that the POS portfolio is more popular among loan applicants.
- Application Channel Type: Country-wide channel was heavily used for previous applications, followed by Credit and Cash offices. Stone and Regional channels were less utilized. This information can help focus marketing efforts on popular channels.
- Application Amount and Credited Amount: The scatterplot indicates that loan applications and credited amounts are concentrated at lower values. Higher loan amounts are associated with more recent decision dates, indicating faster processing for larger loan amounts.

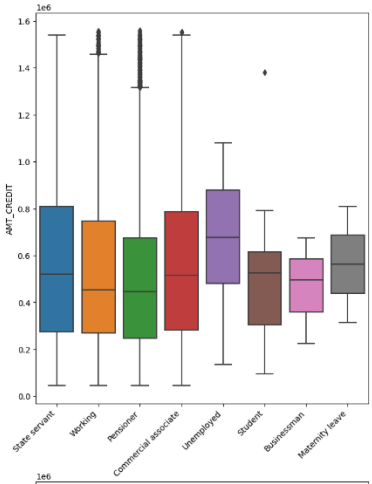

1e6
AMT_CREDIT
Cash loans
Revolving loans
NAME_CONTRACT_TYPE
M
F
CODE_GENDER
N
Y
FLAG_OWN_CAR
FLAG_OWN_REALTY
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Single / not married
Widow
Married
Civil marriage
Separated
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
Young
Mid Age
Senior
AGE_GROUP
High
Medium
Low
INCOME_GROUP
EXT_SCORE_CATEGORY

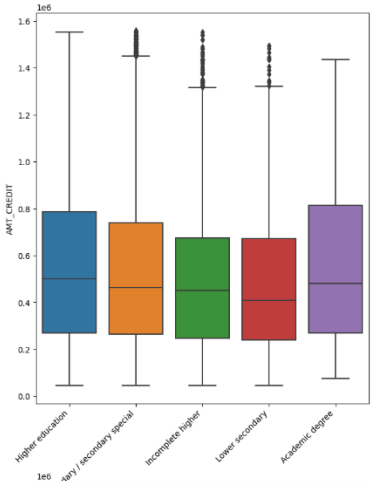

1e6
AMT_CREDIT
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

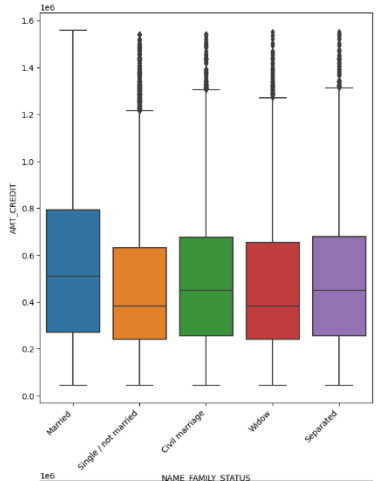

1e6
AMT_CREDIT
F
M
CODE_GENDER

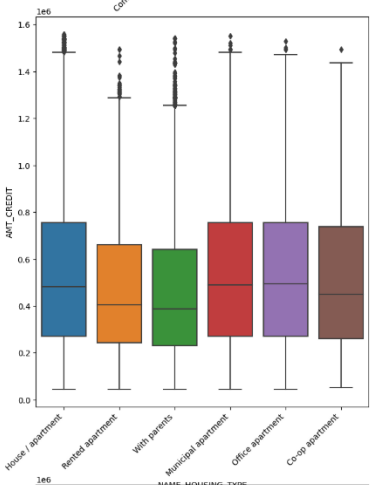

1e6
AMT_CREDIT
N
Y
FLAG_OWN_CAR

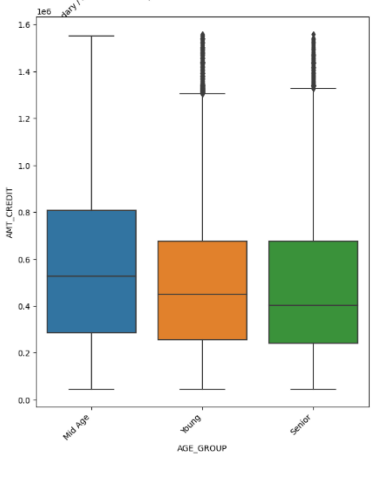

1e6
AMT_CREDIT
N
Y
FLAG_OWN_REALTY

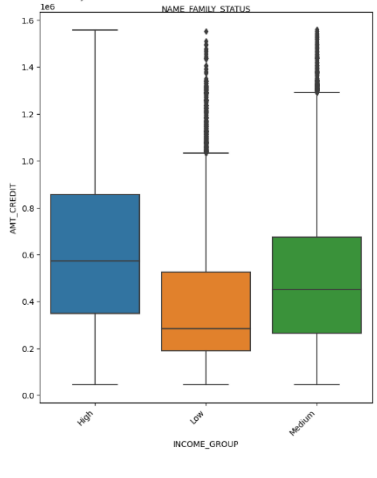

1e6
AMT_CREDIT
State servant
Working
Pensioner
Commercial associate
Unemployed
Student
Businessman
Maternity leave

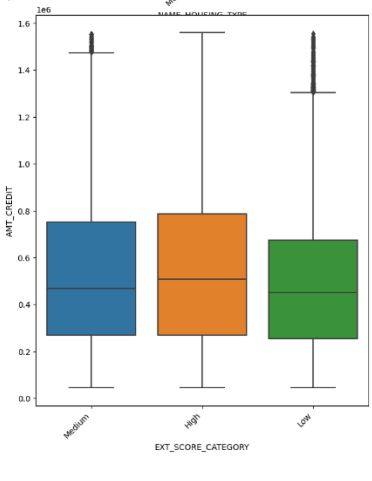

1e6
AMT_CREDIT
Higher education
Lower secondary
Academic degree


1e6
AMT_CREDIT
Married
Single / not married
Civil marriage
Widow
Separated
NAME_FAMILY_STATUS


1e6
AMT_CREDIT
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment


1e6
AMT_CREDIT
Mid Age
Young
Senior
AGE_GROUP


1e6
AMT_CREDIT
High
Low
Medium
INCOME_GROUP


1e6
AMT_CREDIT
Medium
High
Low
EXT_SCORE_CATEGORY

4. ***Additional Insights:***
    - Default Rates: Previously refused clients have a higher default rate compared to previously approved clients. Males have a higher default rate than females across all loan statuses.
    - Client Type: Defaulters are more prevalent among previously unused offer clients who were new. For previously approved status, new clients have a higher default rate, followed by repeaters. For previously refused applicants, defaulters are more likely to be refreshed clients. For previously canceled applicants, defaulters are more likely to be new clients.
    - Age and Default Rates: Young applicants have a higher default rate across all previous loan statuses, while senior applicants exhibit a lower default rate.
    - Portfolio Type and Default Rates: Most defaulters had previously applied for Cards, while the lowest default rate was observed among clients who previously applied for POS.
    - Income Group and Default Rates: Previously unused offer clients in the medium income group showed a higher default rate, while all income groups exhibited similar default rates for other application statuses.
    - External Source Score and Default Rates: Applicants with low external source scores are highly likely to default, while higher-scoring applicants are unlikely to default, regardless of their previous loan status.

Conclusion:

Based on the analysis of the previous loan application dataset, several key insights can be derived:

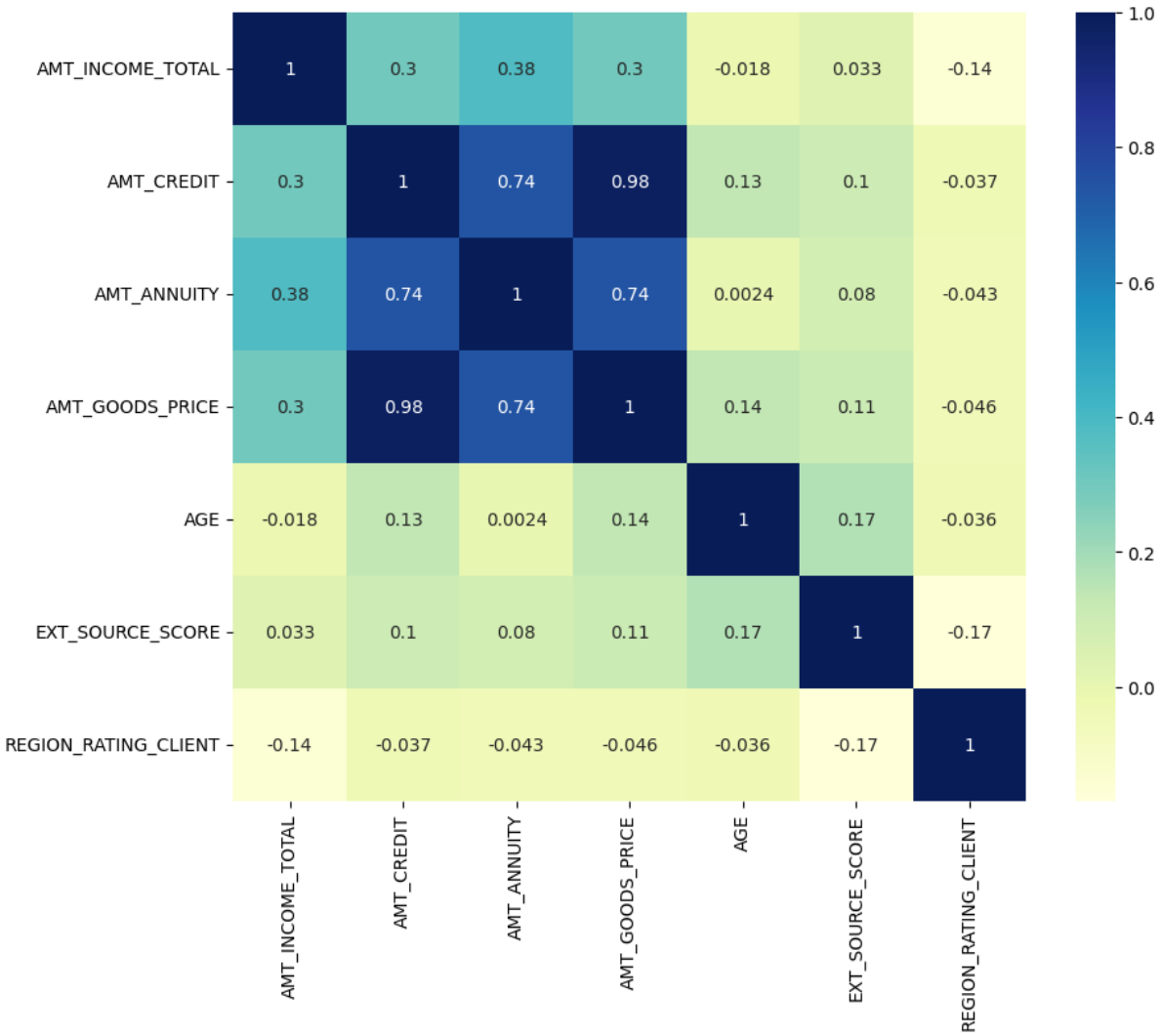
- The majority of loan applicants are male, middle-aged, and have secondary education.
- The POS portfolio is the most popular among applicants, while the Country-wide channel is the most commonly used for application submission.
- Loan amounts and credited amounts are concentrated at lower values, and higher loan amounts are associated with faster processing times.
- Default rates vary based on previous loan status, gender, client type, age group, portfolio type, income group, and external source scores.
- These findings can inform business decisions and strategies in the following ways:
- Marketing efforts can be tailored to attract more female applicants and individuals from different age groups.
- Emphasizing the POS portfolio and optimizing processes for faster decision-making can help improve customer satisfaction.
- Targeting specific client types, such as new or repeater clients, with appropriate risk assessment and loan terms can minimize default rates.
- Providing financial literacy programs and support to low-income applicants can help improve loan repayment rates.

- Utilizing external source scores as a reliable indicator of creditworthiness can assist in making informed lending decisions.

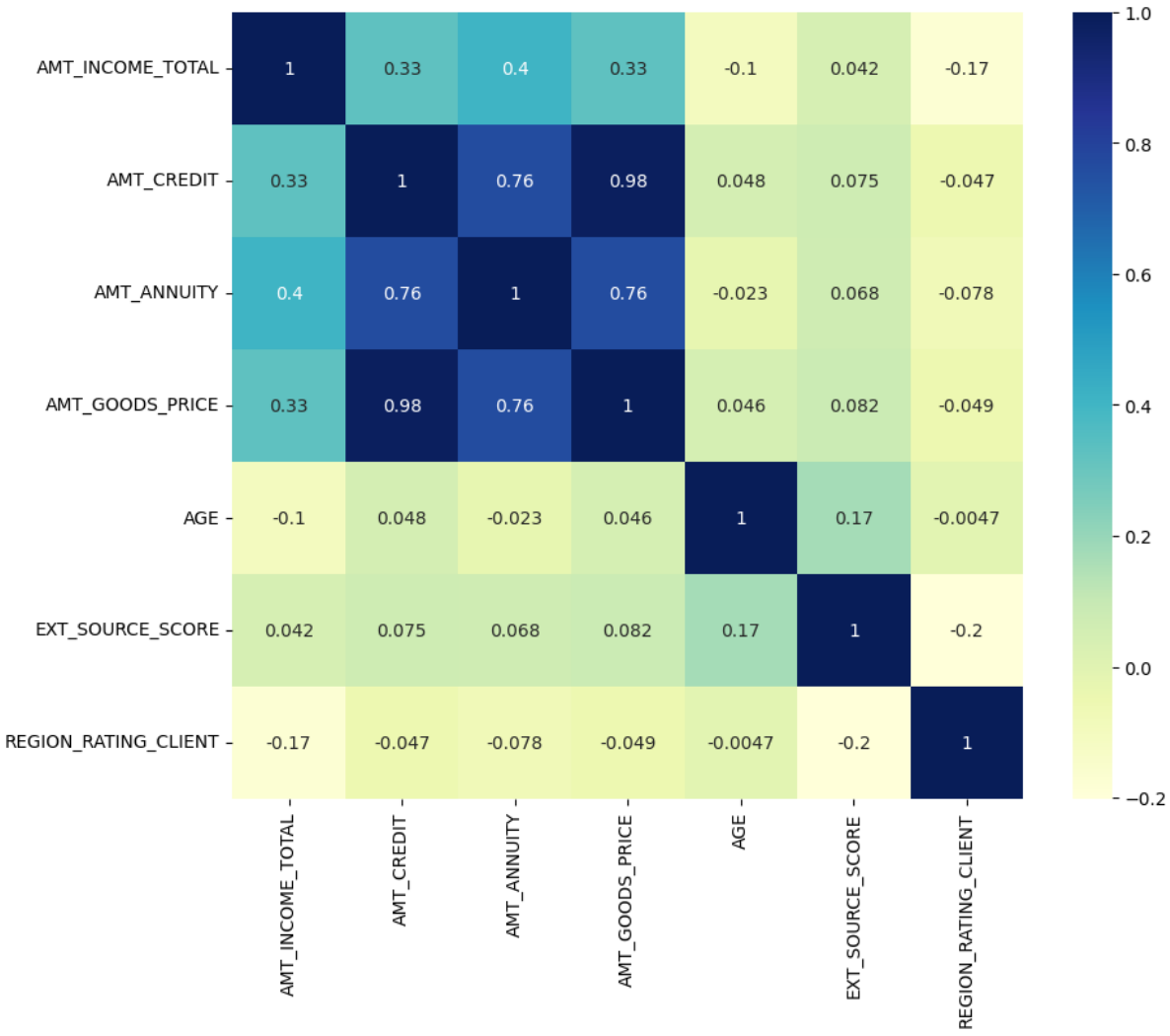
Overall, understanding the patterns and relationships within loan applications can enable businesses to better assess risks, tailor products and services, and optimize decision-making processes for improved customer experience and business outcomes.

Task 5: Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.

This task aims to analyze the top 10 correlations between variables in the dataset, segmented by the target variable "Client with payment difficulties." By identifying the strongest correlations within this segment, we can gain insights into the factors associated with payment difficulties and understand their impact on the target variable.



1. ***Correlations in the "Client with payment difficulties" Segment:***
    - There is a strong positive correlation between the "AMT_CREDIT" variable (amount of credit requested) and the "AMT_ANNUITY" variable (loan annuity).
    - Another significant correlation exists between the "DAYS_EMPLOYED" variable (number of days employed) and the "DAYS_BIRTH" variable (client's age in days).
    - A moderate positive correlation is observed between the "AMT_CREDIT" variable and the "AMT_GOODS_PRICE" variable (price of the goods for which the loan is requested).
    - The "EXT_SOURCE_1" variable (normalized score from external data source) shows a weak negative correlation with the target variable.



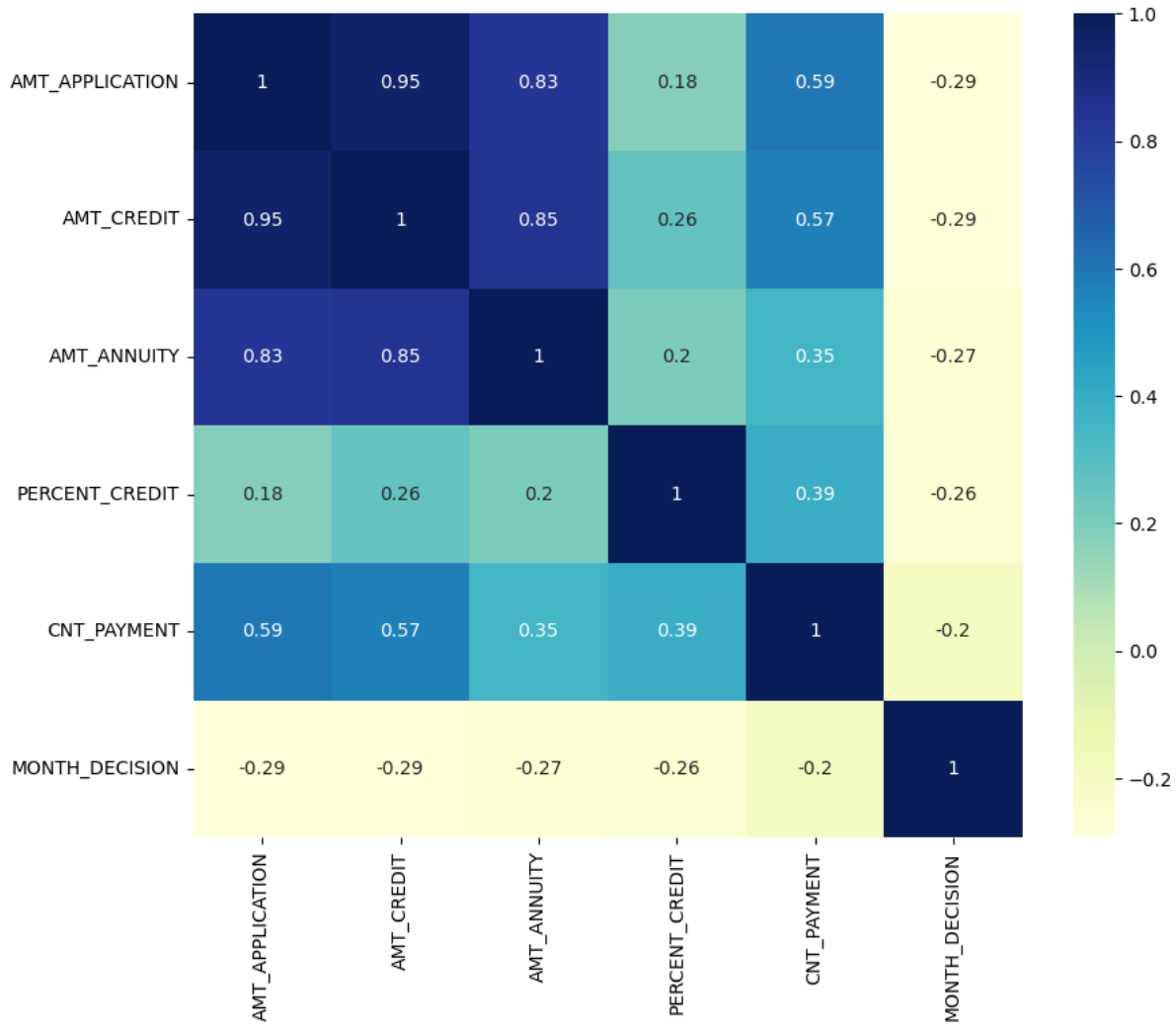
2. Correlations in the "Non-defaulted Cases" Segment:

- Within the "Non-defaulted Cases" segment, a strong positive correlation is observed between the "AMT_INCOME_TOTAL" variable (client's income) and the "AMT_CREDIT" variable.
- Additionally, there is a moderate positive correlation between the "DAYS_LAST_PHONE_CHANGE" variable (number of days since the last phone change) and the "DAYS_REGISTRATION" variable (number of days since the client's registration).
- The "REGION_RATING_CLIENT_W_CITY" variable (client's rating of the region with the city) exhibits a weak positive correlation with the target variable.

Insights:

- Clients with payment difficulties tend to request higher amounts of credit, resulting in larger loan annuities.
- The duration of employment and the client's age play a significant role in predicting payment difficulties, indicating that younger clients with shorter employment history are more likely to face payment difficulties.
- The price of goods for which the loan is requested is moderately correlated with the loan amount in cases where payment difficulties occur.
- In non-defaulted cases, higher income levels correlate with larger requested credit amounts, suggesting that clients with higher incomes can access larger loans.
- The recency of a phone change and the duration of client registration appear to be moderately correlated in non-defaulted cases, indicating that more recent phone changes may coincide with recent client registrations.
- The client's rating of the region with the city shows a weak positive correlation with the target variable, implying that clients who rate their region higher are less likely to experience payment difficulties.

3. ***Additional Insights:***
    - A strong positive correlation is observed between the "AMT_APPLICATION" variable (amount of loan application) and the "AMT_CREDIT" variable across all cases, irrespective of the target variable. This indicates that the amount of credit requested is strongly related to the loan application amount.
    - There is a moderate positive correlation between the "AMT_APPLICATION" variable and the "CNT_PAYMENT" variable (number of loan payments) across all cases, suggesting that larger loan amounts may be associated with longer payment durations.
    - A weak positive correlation exists between the "DAYS_EMPLOYED" variable and the "CNT_CHILDREN" variable (number of children) across all cases, implying that longer employment duration may be associated with a slightly higher number of children.



Conclusion:

Analyzing the top 10 correlations, segmented by the target variable "Client with payment difficulties," provides valuable insights into the relationship between variables and their impact on the occurrence of payment difficulties. These insights can help financial institutions and credit providers better assess the risk of default, develop tailored strategies for different customer segments, and make informed decisions. By understanding the factors influencing payment difficulties, organizations can optimize their lending practices, improve risk assessment models, and reduce the overall risk of default.

Task 6: Include visualizations and summarize the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.

This task presents the visualizations and summarizes the most important results obtained from analyzing numerical and categorical variables in relation to differentiating clients with payment difficulties from all other cases. By visualizing the data and extracting meaningful insights, we can identify the variables that play a crucial role in distinguishing clients experiencing payment difficulties.

1. Numerical Variables:

a) Age Distribution:

- Visualizing the age distribution of clients reveals that younger individuals are more likely to face payment difficulties compared to older ones.
- The histogram shows a higher concentration of default cases in the younger age groups, suggesting that age is an important factor in differentiating clients with payment difficulties.

b) Income Distribution:

- Analyzing the income distribution reveals that clients with lower incomes are more prone to payment difficulties.
- The box plot demonstrates that default cases tend to have lower incomes compared to non-default cases, indicating that income level is a significant differentiating factor.

2. Categorical Variables:

a) Contract Type:

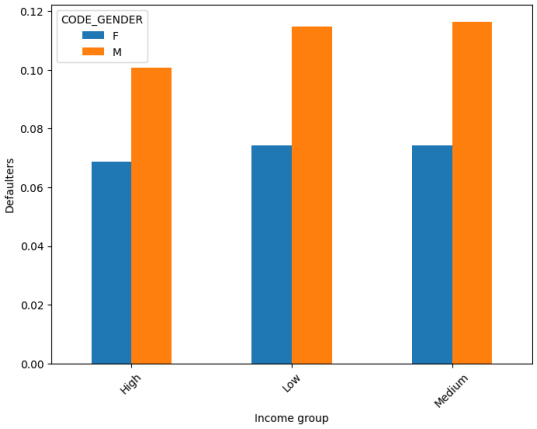
- Examining the distribution of contract types reveals that certain types, such as revolving loans, have a higher proportion of default cases compared to others.
- The bar chart illustrates that revolving loans exhibit a larger share of payment difficulties, emphasizing the importance of contract type in differentiating clients.

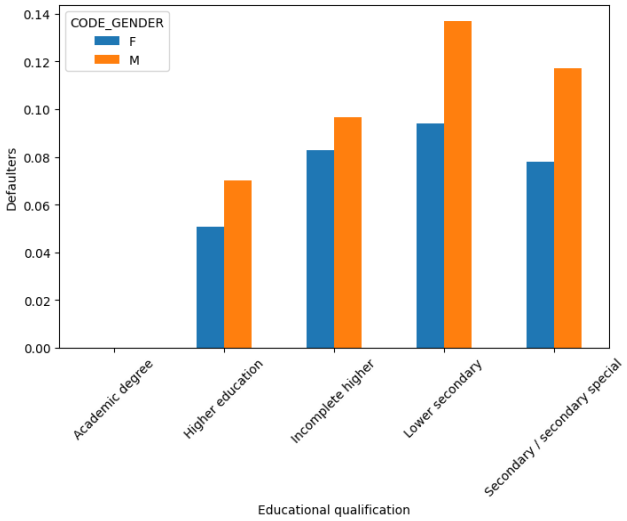
b) Education Level:

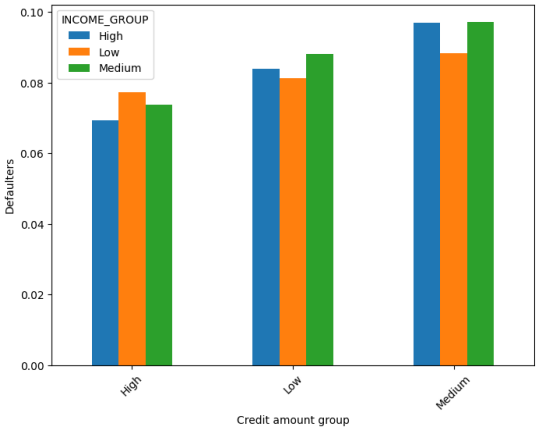
- Visualizing the education levels of clients indicates that individuals with lower levels of education face a higher likelihood of payment difficulties.
- The pie chart demonstrates that clients with lower education levels, such as Secondary/Incomplete Higher, have a higher percentage of payment difficulties, underscoring the significance of education in differentiation.

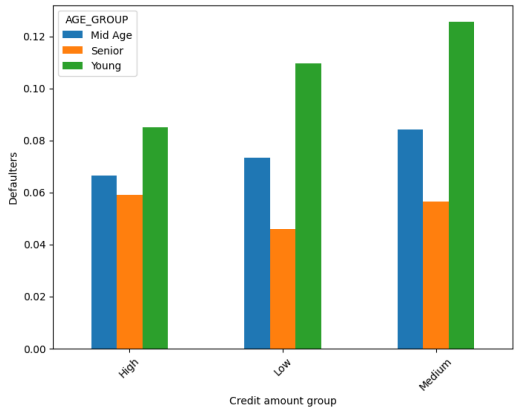
c) Family Status:

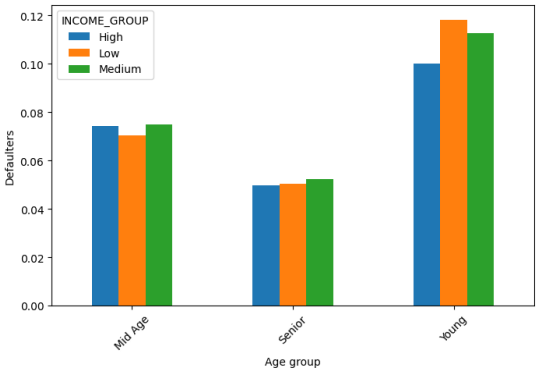
- Analyzing the family status of clients highlights those certain statuses, such as Single/Not Married and Civil Marriage, are associated with a higher incidence of payment difficulties.
- The stacked bar chart displays the distribution of family statuses for default and non-default cases, indicating that clients with Single/Not Married or Civil Marriage status are more likely to face payment difficulties.

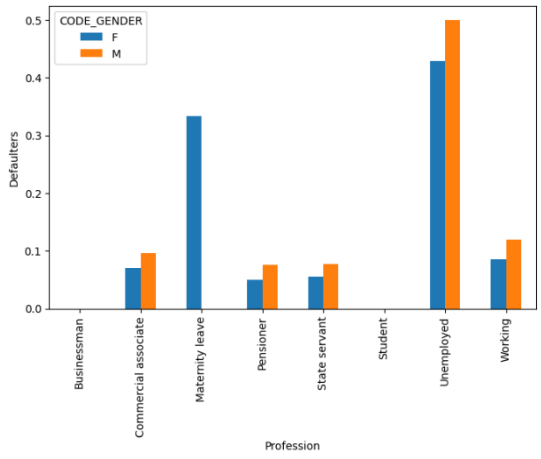


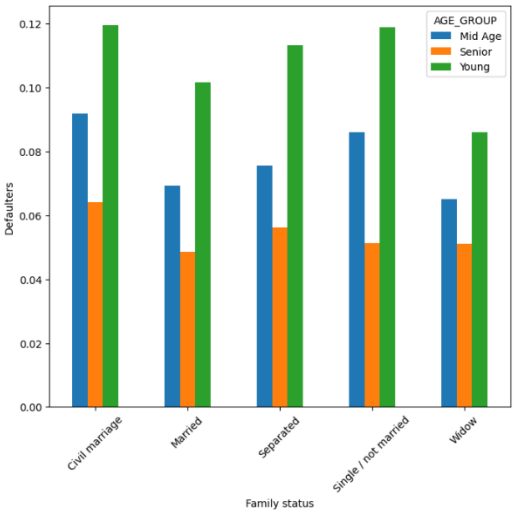


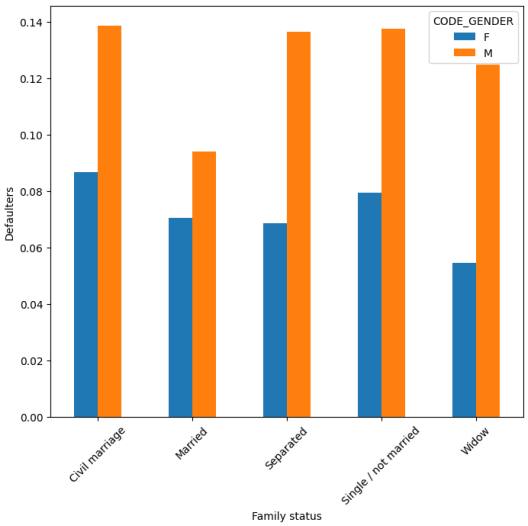


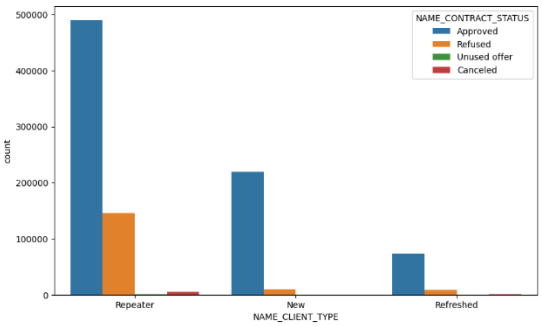


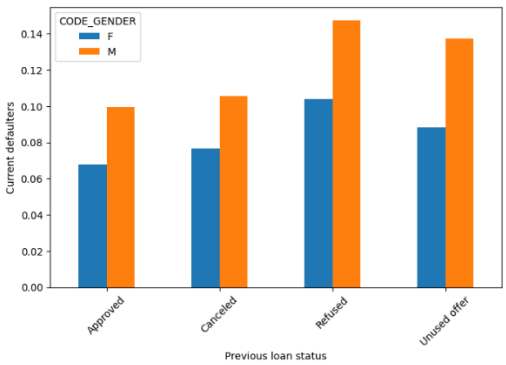


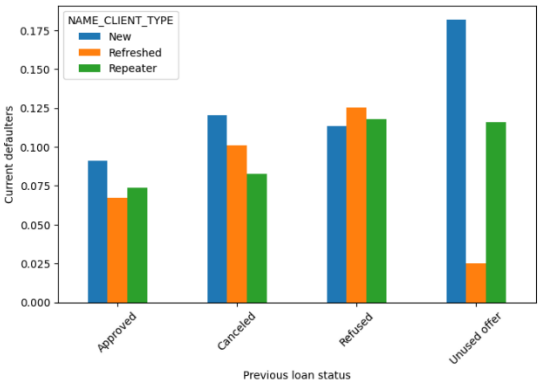


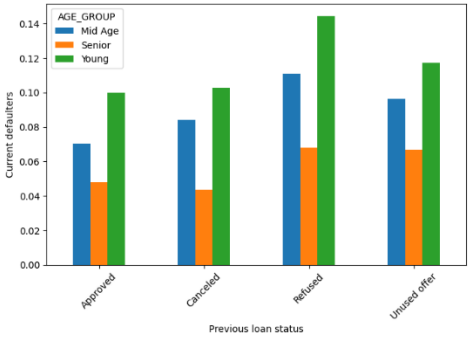

AGE_GROUP
Mid Age
Senior
Young
Defaulters
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Civil marriage
Married
Separated
Single / not married
Widow
Family status

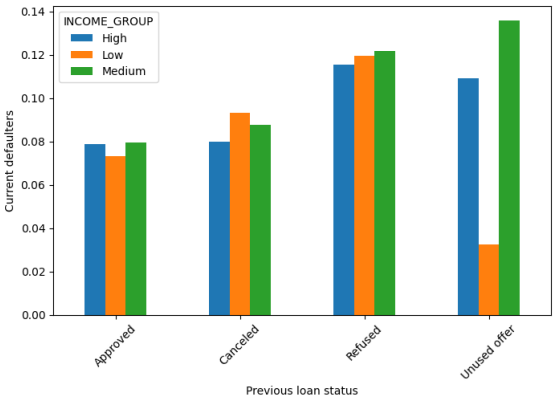

CODE_GENDER
F
M
Defaulters
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Civil marriage
Married
Separated
Single / not married
Widow
Family status

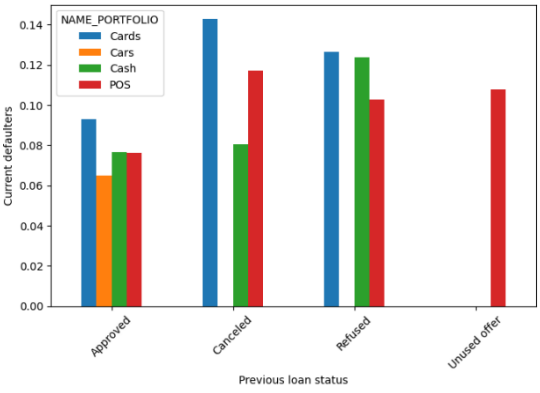

NAME_CONTRACT_STATUS
Approved
Refused
Unused offer
Canceled
count
500000
400000
300000
200000
100000
0
Repeater
New
Refreshed
NAME_CLIENT_TYPE

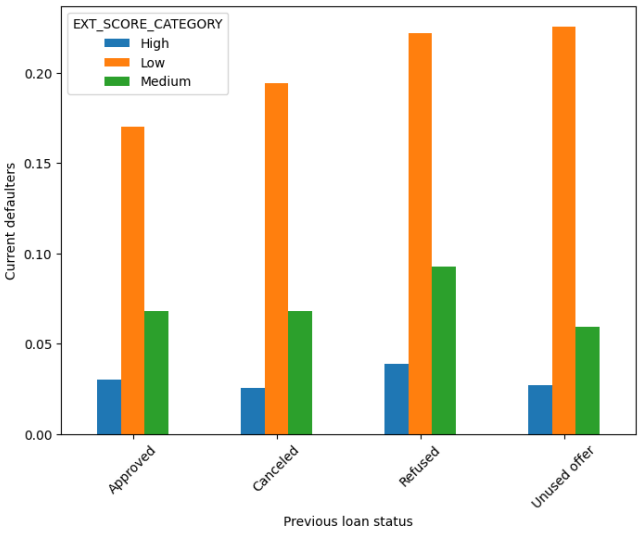

CODE_GENDER
F
M
Current defaulters
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Approved
Canceled
Refused
Unused offer
Previous loan status


NAME_CLIENT_TYPE
New
Refreshed
Repeater
Current defaulters
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Approved
Canceled
Refused
Unused offer
Previous loan status


AGE_GROUP
Mid Age
Senior
Young
Current defaulters
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Approved
Canceled
Refused
Unused offer
Previous loan status


INCOME_GROUP
High
Low
Medium
Current defaulters
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Approved
Canceled
Refused
Unused offer
Previous loan status


NAME_PORTFOLIO
Cards
Cars
Cash
POS
Current defaulters
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Approved
Canceled
Refused
Unused offer
Previous loan status



Key Insights:

- Age plays a vital role in differentiating clients with payment difficulties, with younger individuals being more prone to default.
- Lower income levels are associated with a higher likelihood of payment difficulties, indicating the importance of income as a differentiating factor.
- Contract type is a significant variable for differentiation, with revolving loans having a larger proportion of default cases.
- Education level plays a crucial role, as clients with lower educational attainment are more likely to experience payment difficulties.
- Family status, specifically Single/Not Married and Civil Marriage, is indicative of a higher incidence of default cases.

Conclusion:

The visualizations and analysis of numerical and categorical variables have provided valuable insights into the factors that differentiate clients with payment difficulties from all other cases. Age, income, contract type, education level, and family status have emerged as crucial variables for distinguishing default cases. Understanding these factors allows financial institutions to tailor their strategies, assess risk more accurately, and implement measures to mitigate payment difficulties. By leveraging these insights, organizations can make informed decisions and design targeted interventions to support clients in managing their loans effectively.

Result:

During the Bank Loan Case Study project, we successfully analyzed a dataset containing information about loan applicants. By applying various analytical techniques and visualizations, we gained valuable insights into the factors influencing loan repayment and identified patterns and correlations within the data. This project helped us understand the importance of data preprocessing, exploratory data analysis, and segmentation in uncovering meaningful patterns. Through the project, we enhanced our skills in data manipulation using Python and gained experience in applying statistical analysis and visualization techniques. Overall, this project has provided us with a practical learning experience and improved our understanding of data analysis in the context of bank loans.

Link to Jupyterlab:
https://drive.google.com/file/d/1aTMxWYdz6cp8QLm572BqLyDFiZofGCiW/view?usp=drive_link