EDA
ASSIGNMENT

UNDERSTANDING CONSUMER AND LOAN ATTRIBUTES TO MINIMIZE RISK IN LENDING

By- Vaibhavi B Raj

CONTENT

PROBLEM STATEMENT

Objective:

- Learn the main causes of loan defaults.
- Reduce the risk from high-risk applicants and make sure qualified borrowers are not turned away.

Key Challenges:

- Striking a balance between accepting high-risk applications and turning away qualified borrowers.
- Addressing credit histories that are incomplete or nonexistent.

Scenarios Analyzed:

- Loan applications that have been approved, denied, cancelled, and unused.
- distinction between those who have defaulted and those who have not.

Goal:

- Utilise EDA to find trends that affect repayment issues.

DATASETS:

This dataset has 3 files as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

2. 'previous_application.csv' contains information about the client’s previous loan data. It contains data on whether the previous application was Approved, Cancelled, Refused, or Unused.

3. 'columns_description.csv' is a data dictionary describing the variables' meaning.

APPROACH AND METHODOLOGY

Data Cleaning:

- Handling Missing Data:
  - Imputation methods for missing values were used:
    a) Variables with numbers: In order to ensure reliable estimates against outliers, the median/mean was substituted.
    b) Using categorical variables preserves logical coherence.
  - Columns having a high percentage of missing values (>19%) were dropped.

- Negative Value Treatment: To make interpretation simpler, all negative values in columns such as DAYS_BIRTH were changed to positive ones.

- Categorical_columns Cleanup: 'XNA' and other unusual values were changed to the most common logical value as part of the categorical cleanup process.

APPROACH AND METHODOLOGY

Exploratory Data Analysis (EDA)

- Univariate analysis: Individual variable distribution.

- Bivariate analysis: examines the connections between variables and the intended outcome.

- Segmentation: Distinguished analysis between non-defaulters and defaulters.
    a) Non-Defaulters: Applicants who have successfully paid back their loans (TARGET = 0).
    b) Defaulters (TARGET = 1): Repayment-challenged applicants.

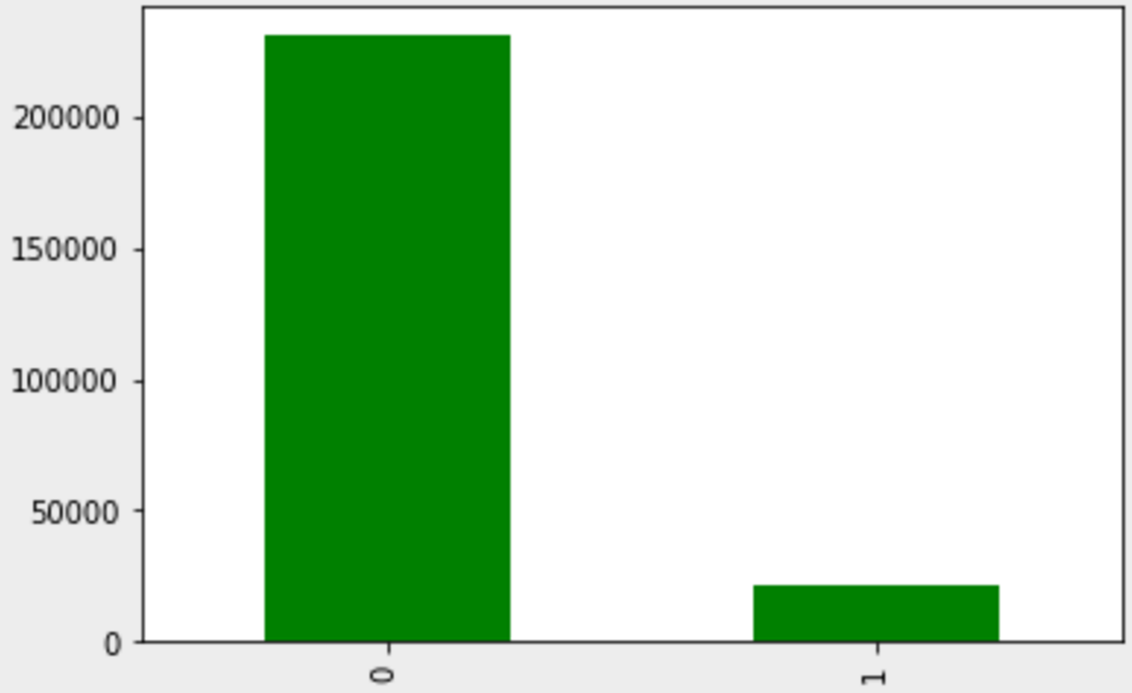
- Correlation Analysis: Determine the most important impacting factors using correlation analysis.

- Outlier Detection: Box plots were used to find anomalies in variables such as CNT_CHILDREN and AMT_CREDIT.Although they were noted, outliers were kept for a thorough examination.

GRAPHS AND INSIGHTS From *application_data.csv* as *data0*

Target Variable Distribution

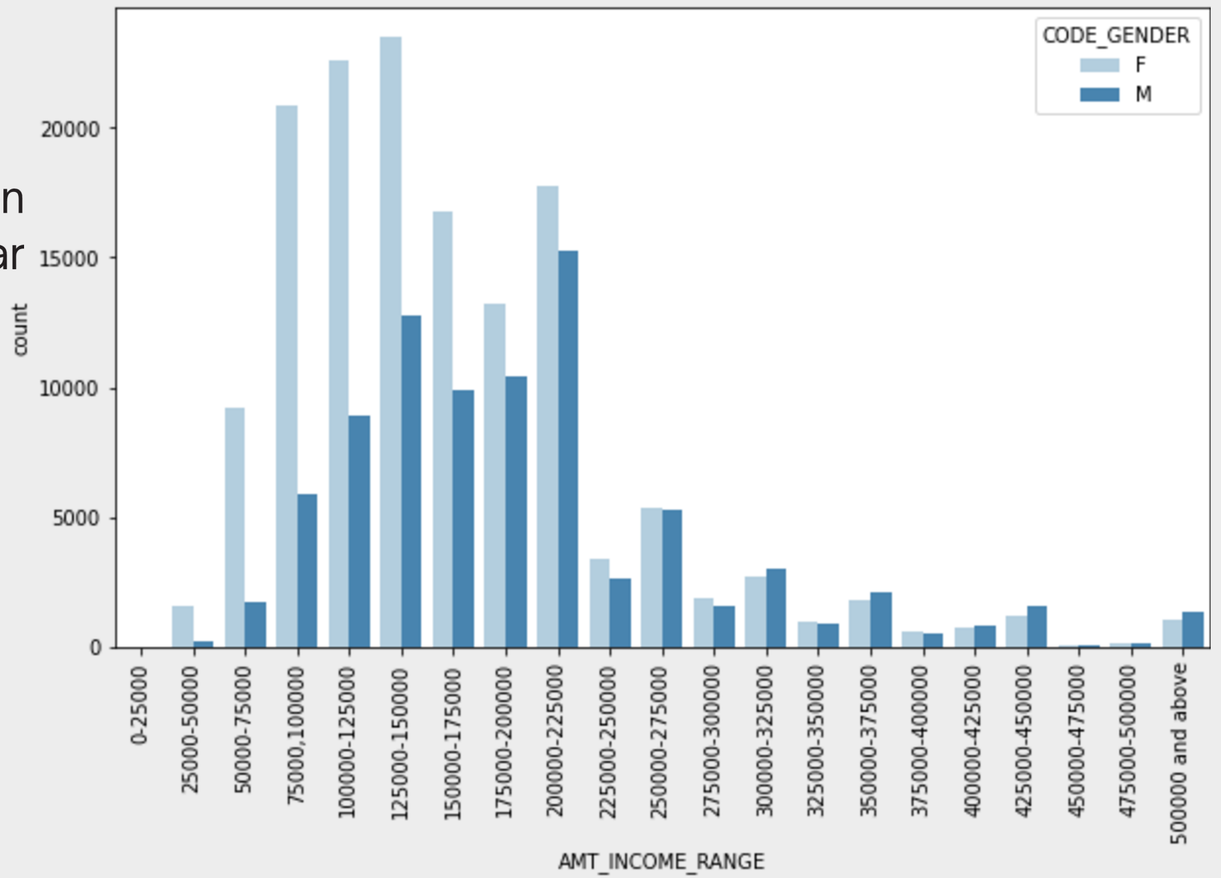
- The distribution of the TARGET variable is shown in the bar chart.
    - With a ratio of almost 10:1, the dataset shows a class imbalance.
    - The majority of clients successfully return their loans, as indicated by this imbalance.



GRAPHS AND INSIGHTS

Income Distribution and Default Risk

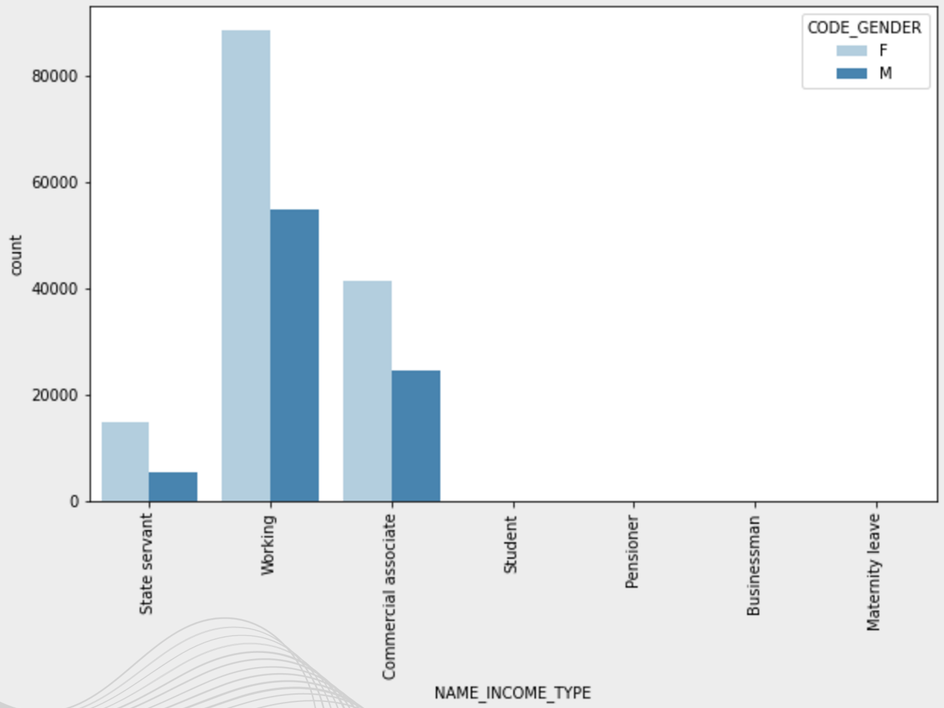
- The analysis of the income distribution between defaulters and non-defaulters is shown in the bar chart.
    - Default rates are higher for applicants with lower incomes.
    - Stricter checks or different lending amounts for low-income groups can be implemented with insight.



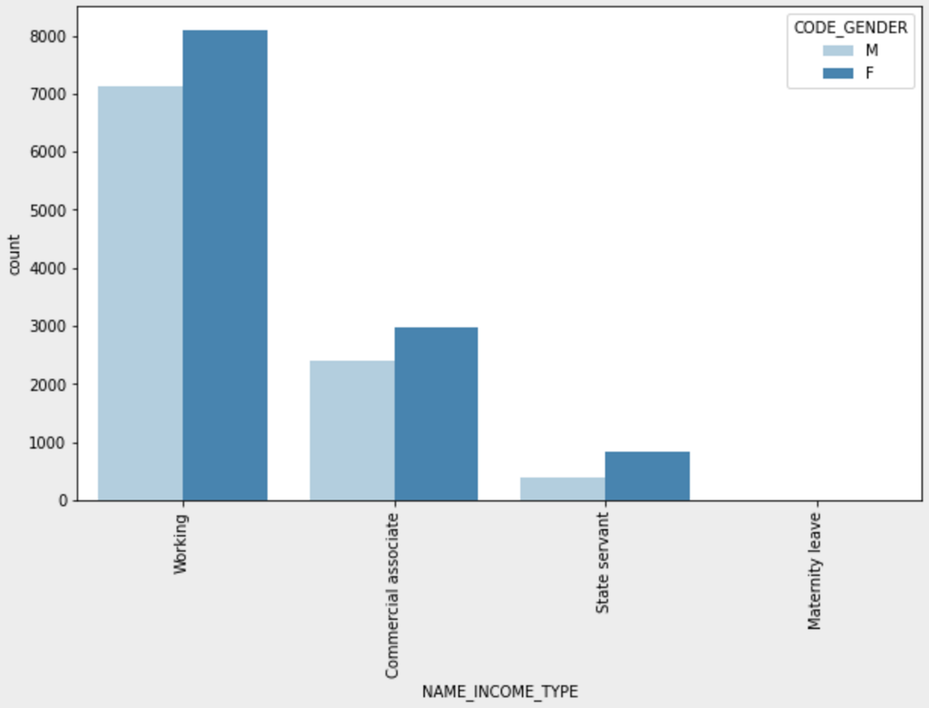
GRAPHS AND INSIGHTS

- The distribution of applicants' income kinds (such as working and pensioner) is shown in the bar chart.
- Although self-employed and lower-income groups have higher default inclinations, the majority of applicants are employed.

Non-Defaulters

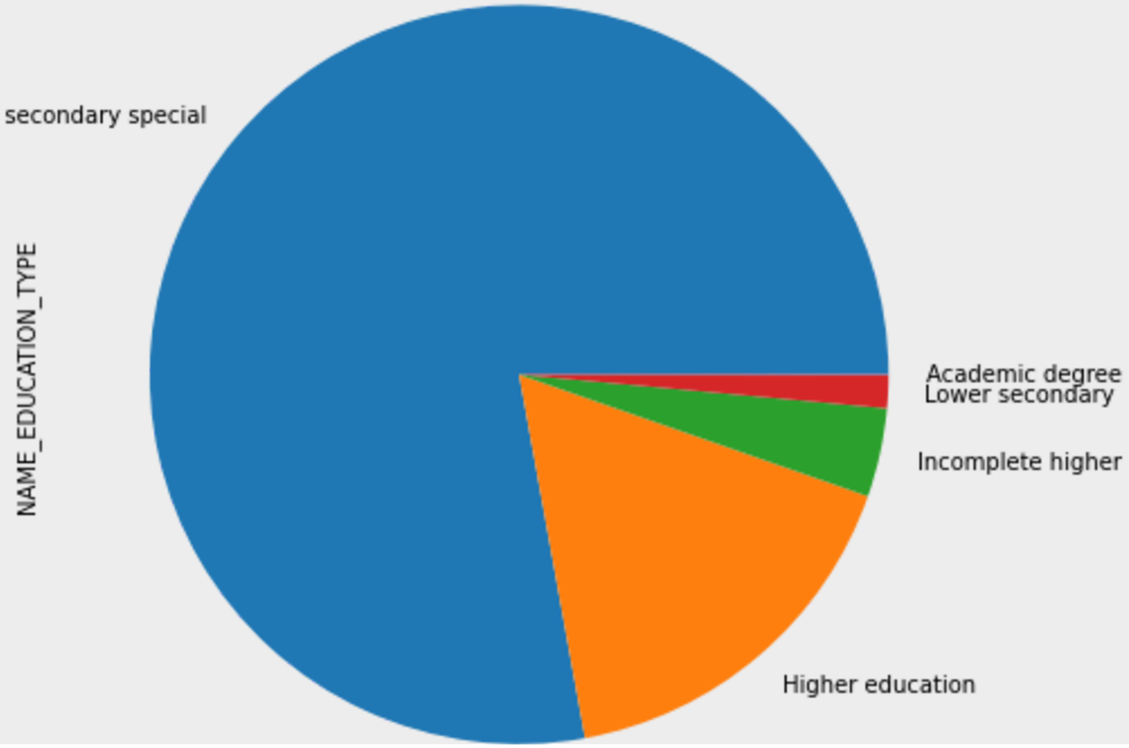


Defaulters



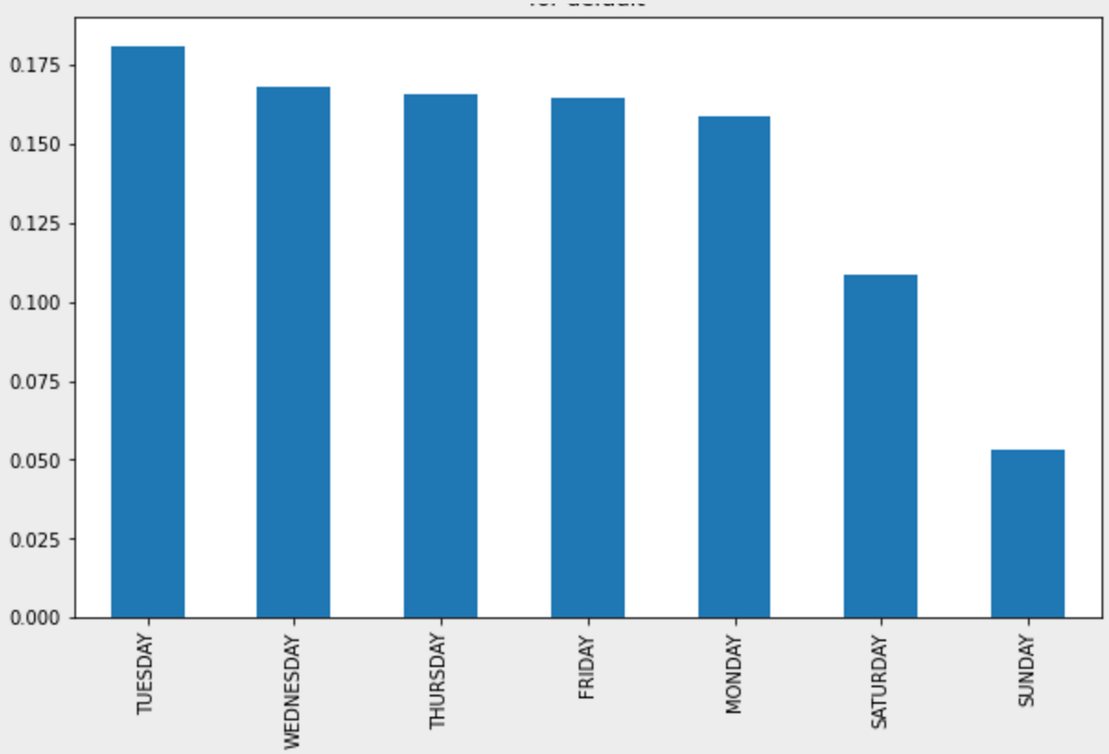
GRAPHS AND INSIGHTS

- The percentage of loan applications with varying educational backgrounds.
  - A pie chart that shows candidates' educational backgrounds (e.g., Secondary, Higher, Graduate).
  - Particularly among defaulters, secondary education predominates, indicating a link between education and default risk.



GRAPHS AND INSIGHTS

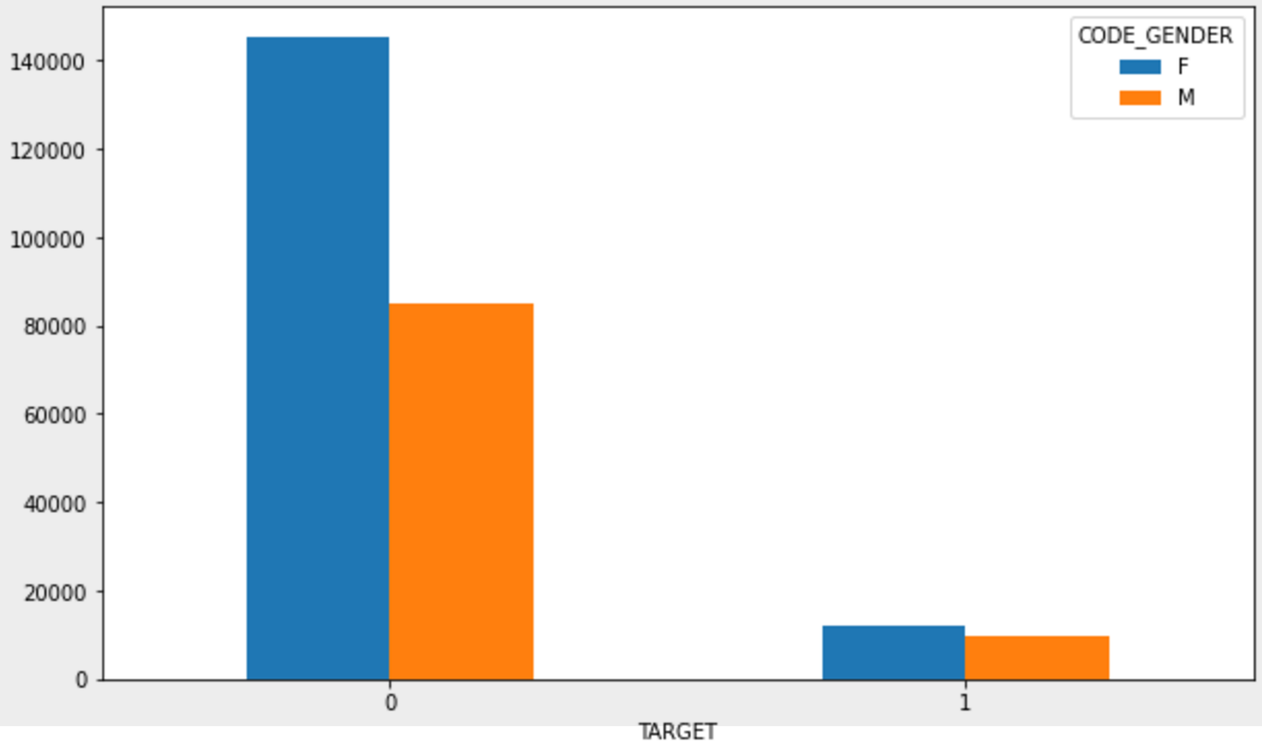
- Application Timing Bar Chart (WEEKDAY_APPR_PROCESS_START). Determine the patterns in loan applications by day of the week.
  - A bar graph that displays the number of loan applications for every day of the week.
  - Weekdays—especially Tuesdays—have the biggest number of applications, while weekends see a decline.



GRAPHS AND INSIGHTS

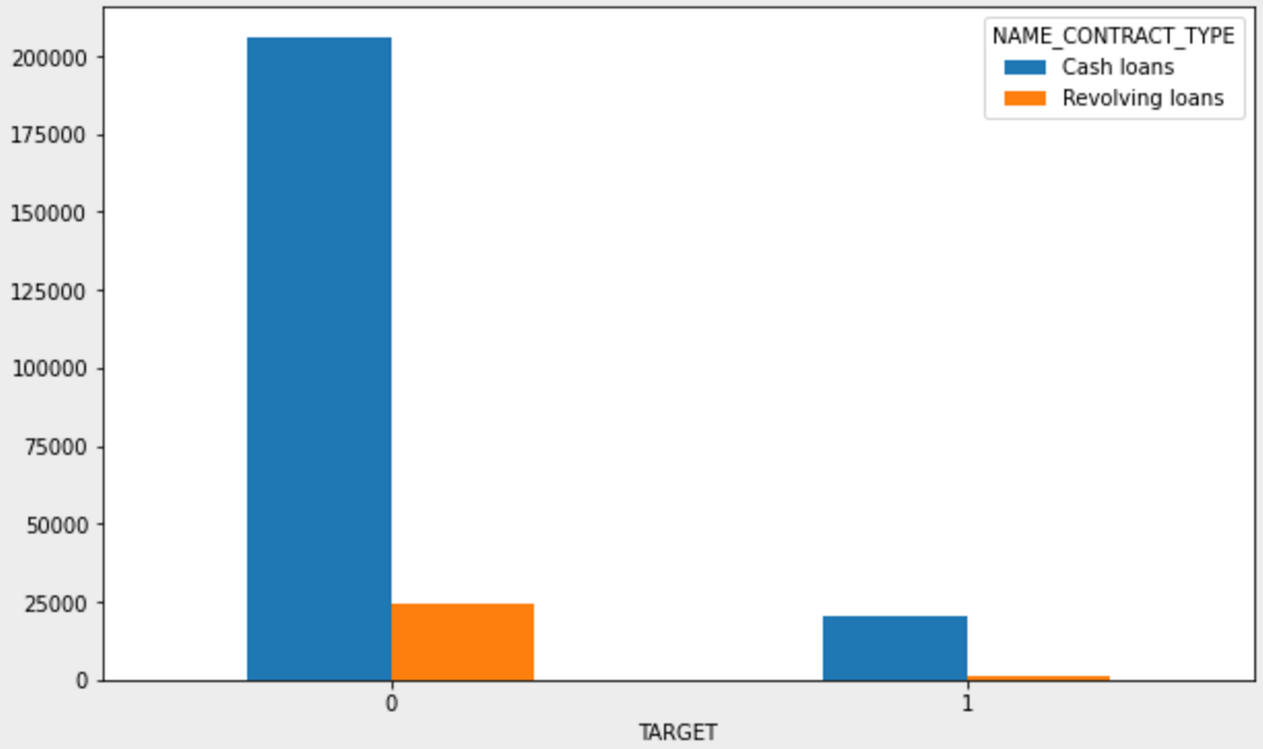
Loan Type and Status

- Gender Distribution Bar Chart (CODE_GENDER). Analyse the loan distribution by gender.
  - The number of male and female applications for defaulters and non-defaulters is displayed in a bar chart.
  - Although defaulters are more evenly divided between the sexes, females are more likely to be approved for loans.



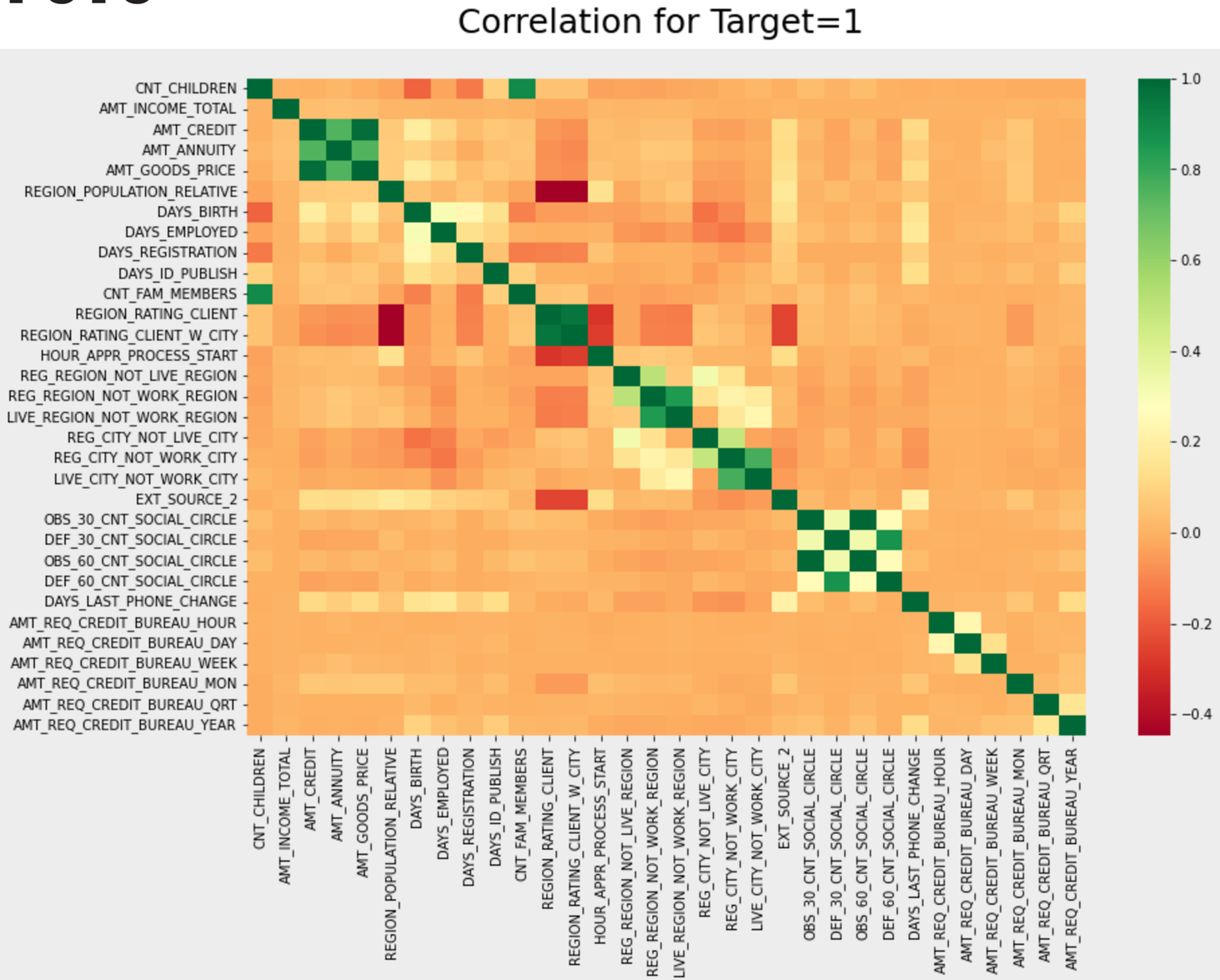
GRAPHS AND INSIGHTS

- NAME_CONTRACT_TYPE Count Plot by TARGET. Percentage of loan types between non-defaulters and defaulters (cash versus revolving).
  - Cash loans are more popular, but revolving loans show relatively lower default rates.
  - This highlights cash loans as a potential risk area requiring better risk assessment.



CORRELATION ANALYSIS

- Correlation Heatmap of TARGET-segmented relationships between numerical variables.
  - Negative correlation with DAYS_EMPLOYED (recently employed are higher risk) is one of the strong connections for defaulters.
  - positive association with AMT_CREDIT (the chance of default rises with loan amounts).

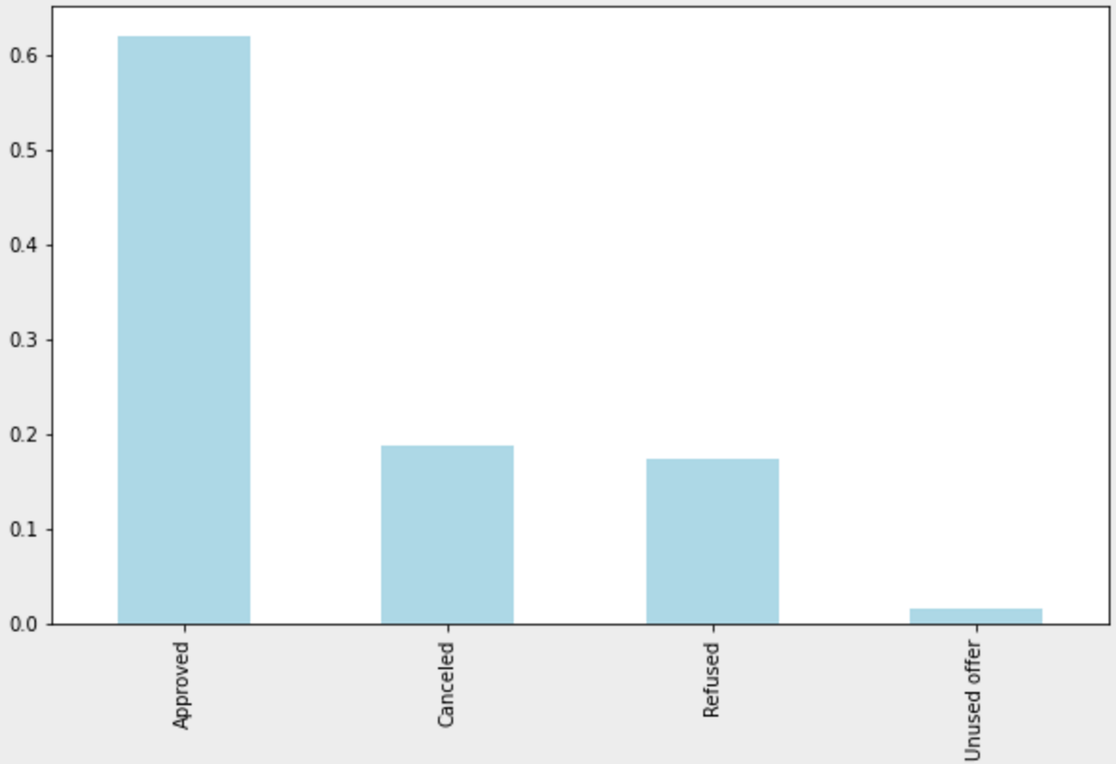


GRAPHS AND INSIGHTS

From *previous_application.csv* as *data1*

Loan Status Distribution

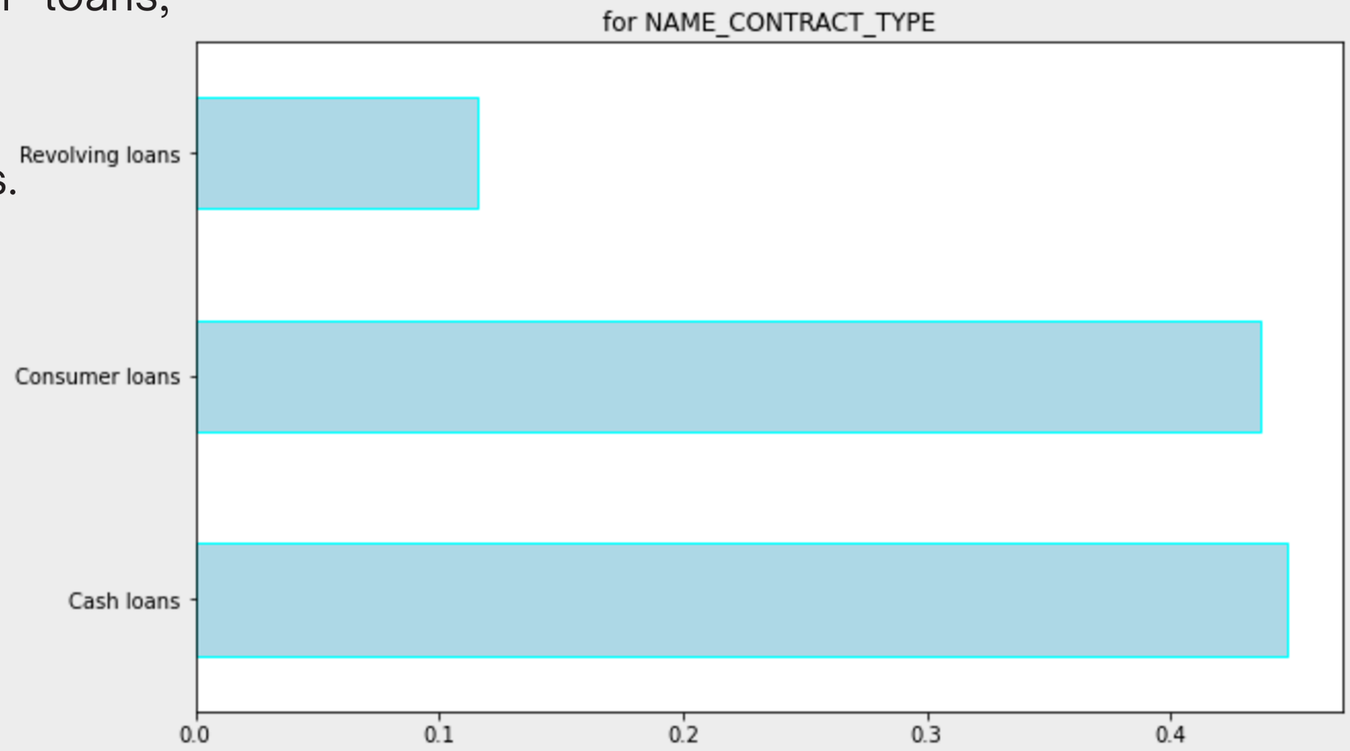
- NAME_CONTRACT_STATUS ChartLoans are distributed according to their status: approved, cancelled, refused, or unused.
    - While the majority of loans are granted, a sizable portion are denied.
    - Refusals reveal the company's risk management strategies.
    - Loans that have been approved require additional risk factor analysis.



GRAPHS AND INSIGHTS

Loan Purpose Analysis

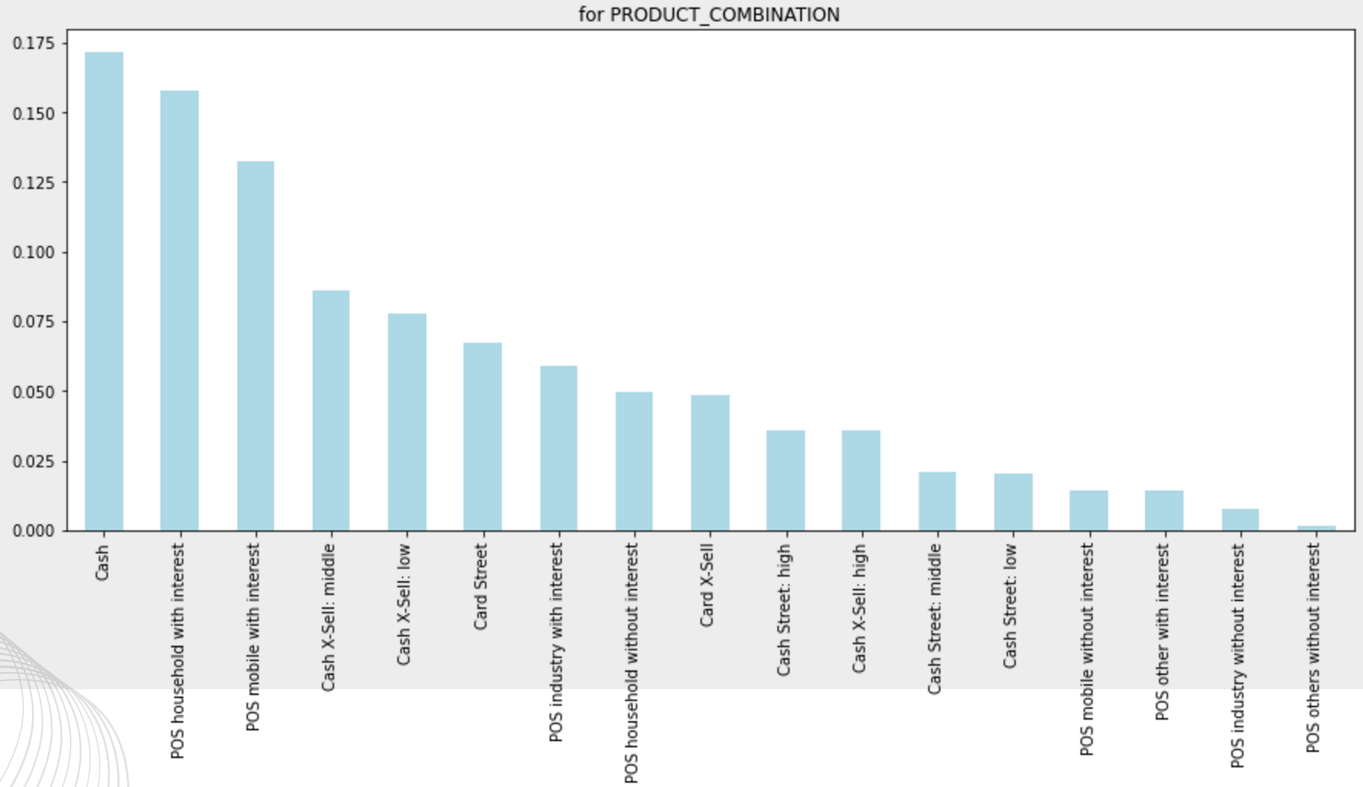
- Bar Chart for NAME_CONTRACT_TYPE Horizontally the type loan (e.g., cash loans, consumer loans etc.).
    - The majority of loans are cash loans.
    - These goals could be a sign of financial distress and increased default risk.



GRAPHS AND INSIGHTS

Product Combination Analysis

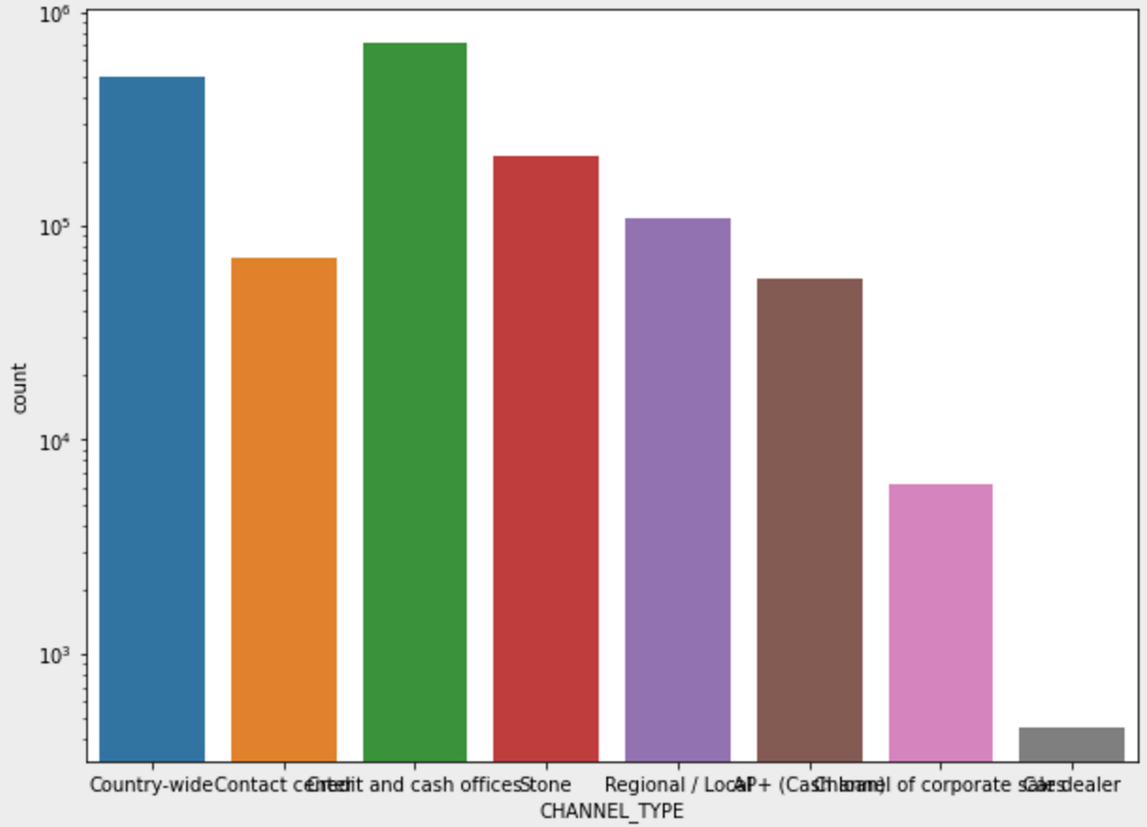
- Product_Combination Bar Chart Distribution of product combinations (e.g., cash loan with items, standalone cash loan) for loans that have been accepted.
- The majority of loans are simple cash loans, with reasonable approval rates for loans paired with products.
- Stricter checks may be necessary for standalone cash loans.



GRAPHS AND INSIGHTS

Channel Analysis

- Bar Chart of CHANNEL_TYPE Distribution of loan processing channels (e.g., POS channels, credit and cash offices).
    - A significant number of loans are handled by credit and cash offices.
    - Relatively fewer loans are displayed by cash dealers, which may indicate a lesser risk.



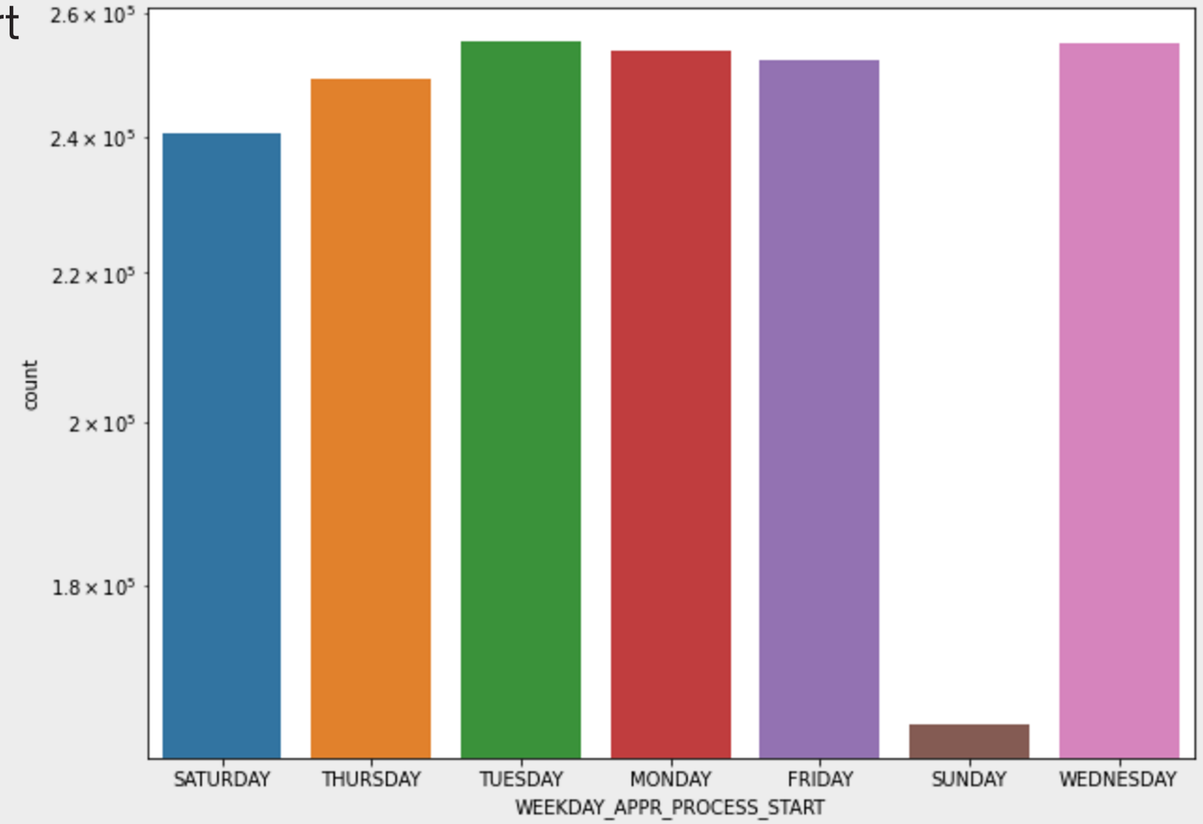
GRAPHS AND INSIGHTS

Loan Applications by Day of the Week

- WEEKDAY_APPR_PROCESS_START bar chart showing loan applications by day of the week.

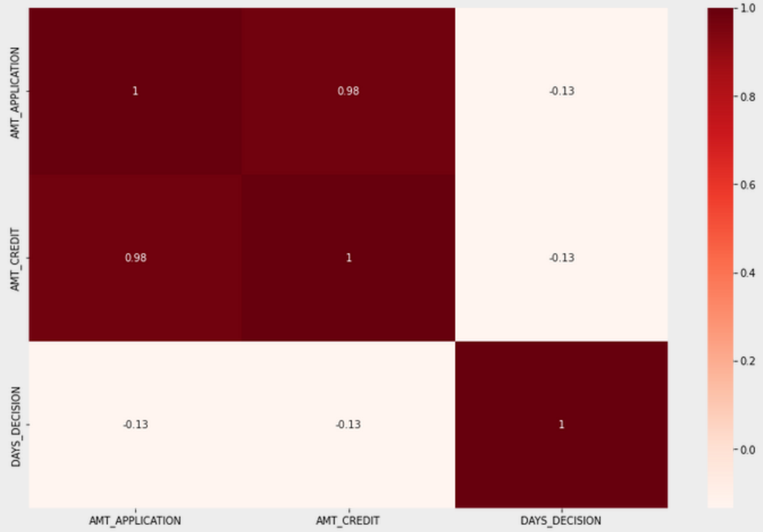
  - Weekdays, particularly the middle of the week, saw a rise in loan applications.

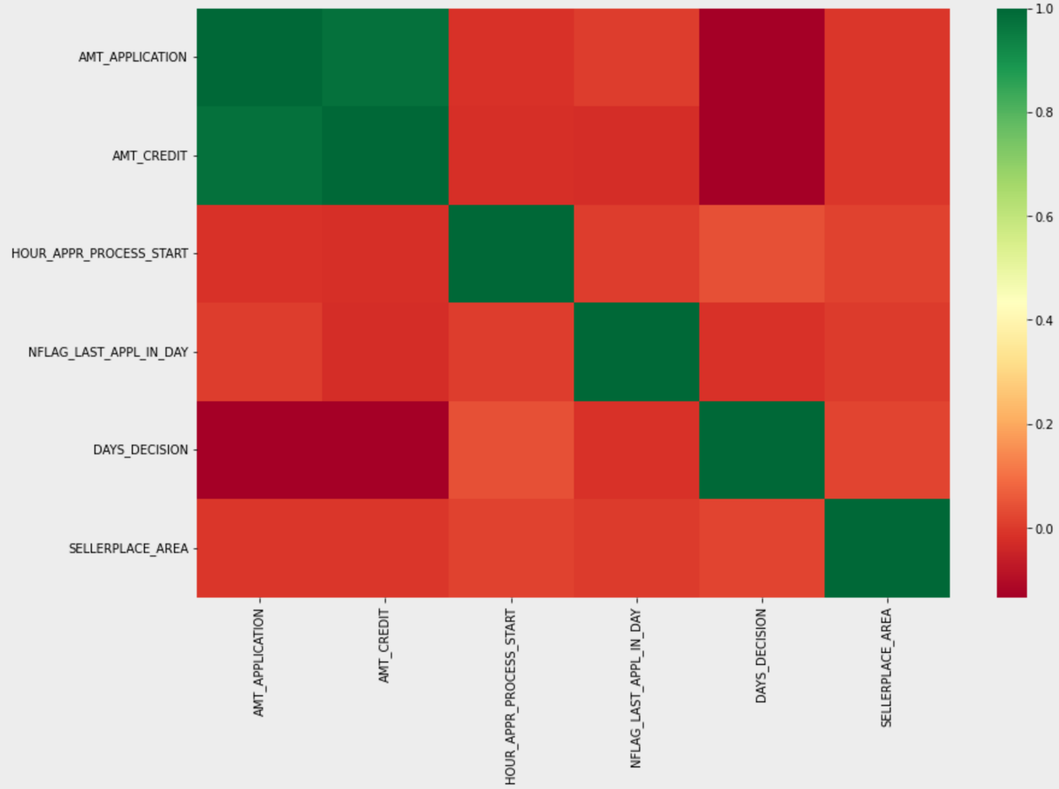
  - Weekends (Saturday and Sunday) saw a sharp decline in applications, indicating lower operational activity.



CORRELATION ANALYSIS

- Correlations between numerical variables such as AMT_CREDIT, DAYS_DECISION, and AMT_APPLICATION are depicted in a heatmap of data1.
  - AMT_APPLICATION and AMT_CREDIT showed strong connections.
  - Weaker correlations indicate that DAYS_DECISION (time since application) has less sway.





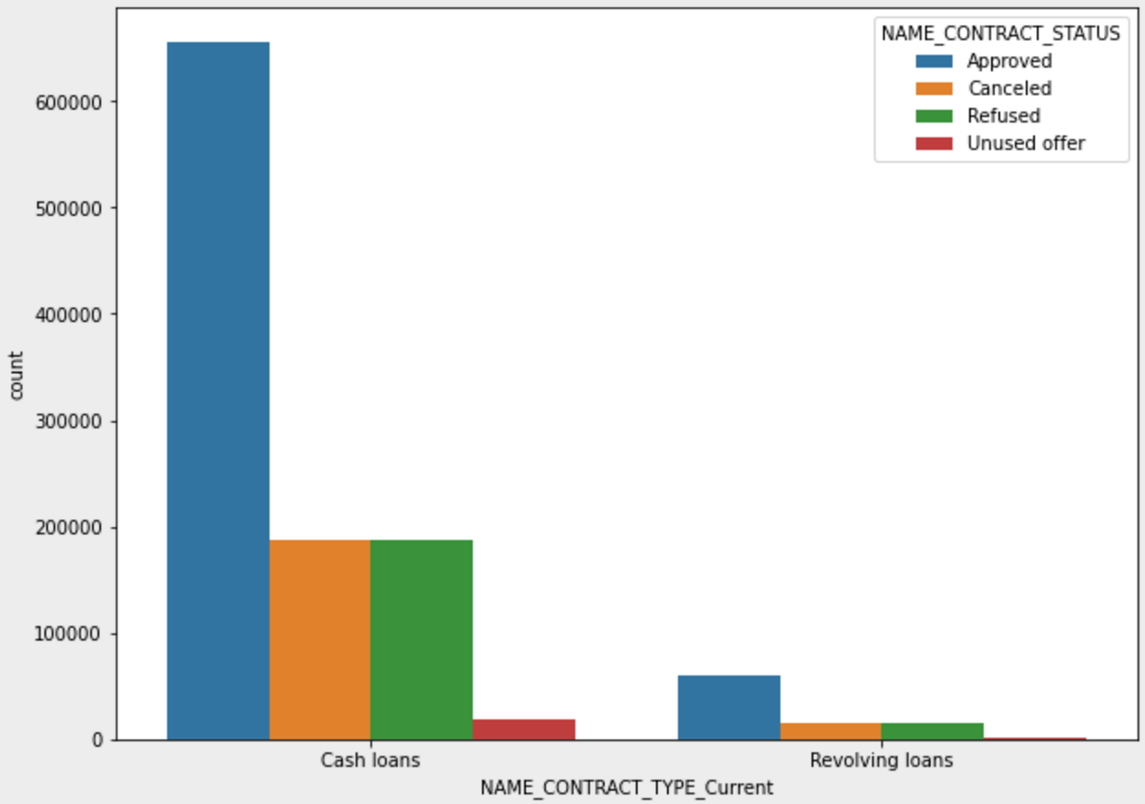
GRAPHS AND INSIGHTS From *merged_df* (data1+data0)

Loan Status and Current Loan Type

- Distribution of loan statuses (approved, denied, etc.) among the many loan types that are now available (cash, revolving).

- Approved applications are dominated by cash loans.

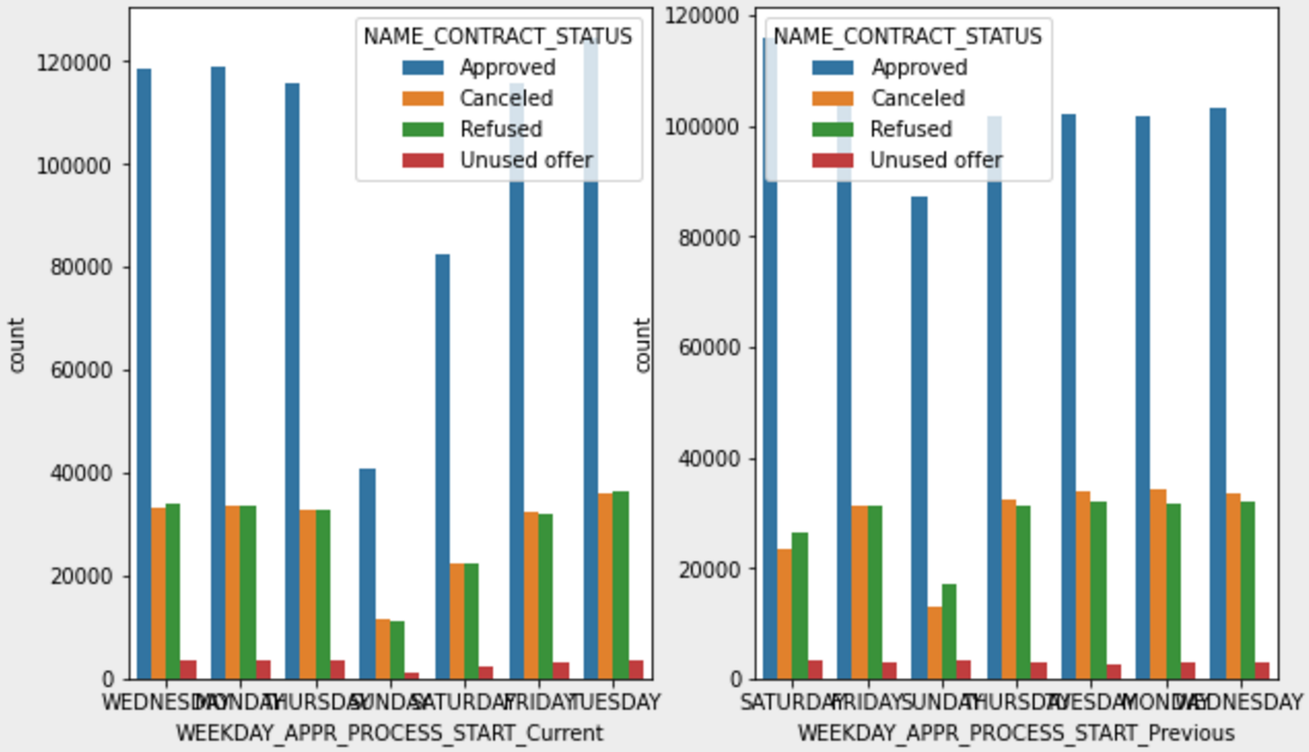
- Revolving loans are more likely to be refused, which suggests that these loans are subject to more stringent scrutiny.



GRAPHS AND INSIGHTS

Day of Application Processing

- Weekday_Appr_Procprocess_Start_Current vs. Weekday_Appr_Procprocess_Start_Previous Count Plot
- Weekdays saw more applications for both existing and past loans than weekends.
- Operational modifications could maximize activity during the busiest weekdays.

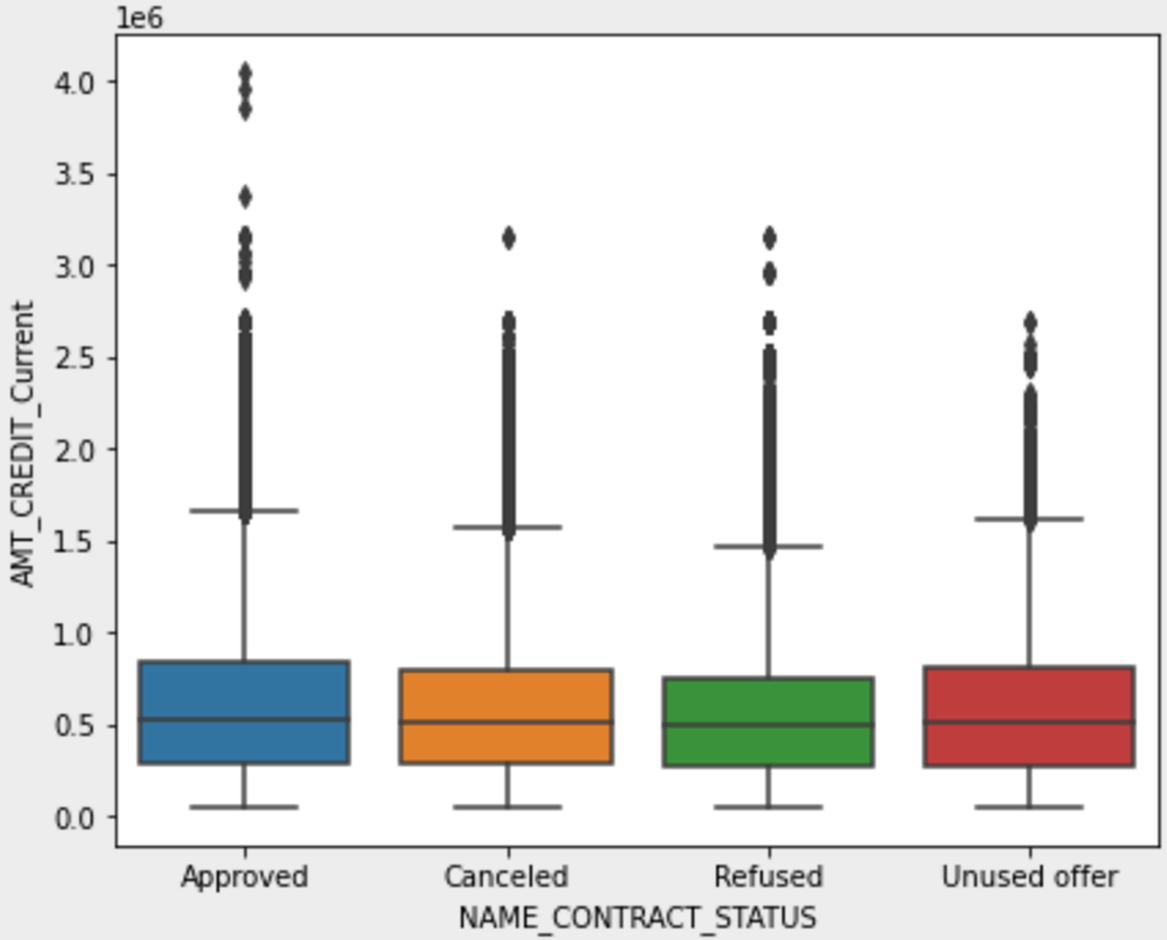


GRAPHS AND INSIGHTS

Loan Amounts in Current vs. Previous Applications

- AMT_CREDIT_Current vs. NAME_CONTRACT_STATUS box plot
- Credit amounts for various debt statuses are distributed.

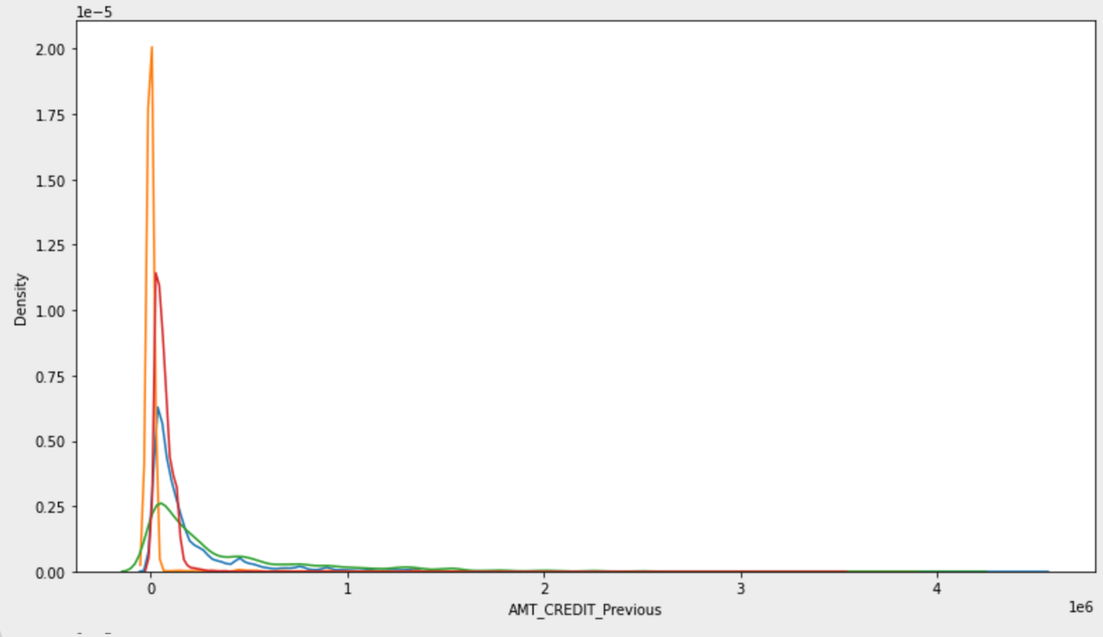
- In general, approved loans have larger credit amounts than those that are denied.
- The range of unused and cancelled offers is comparable, perhaps as a result of client-initiated cancellations.

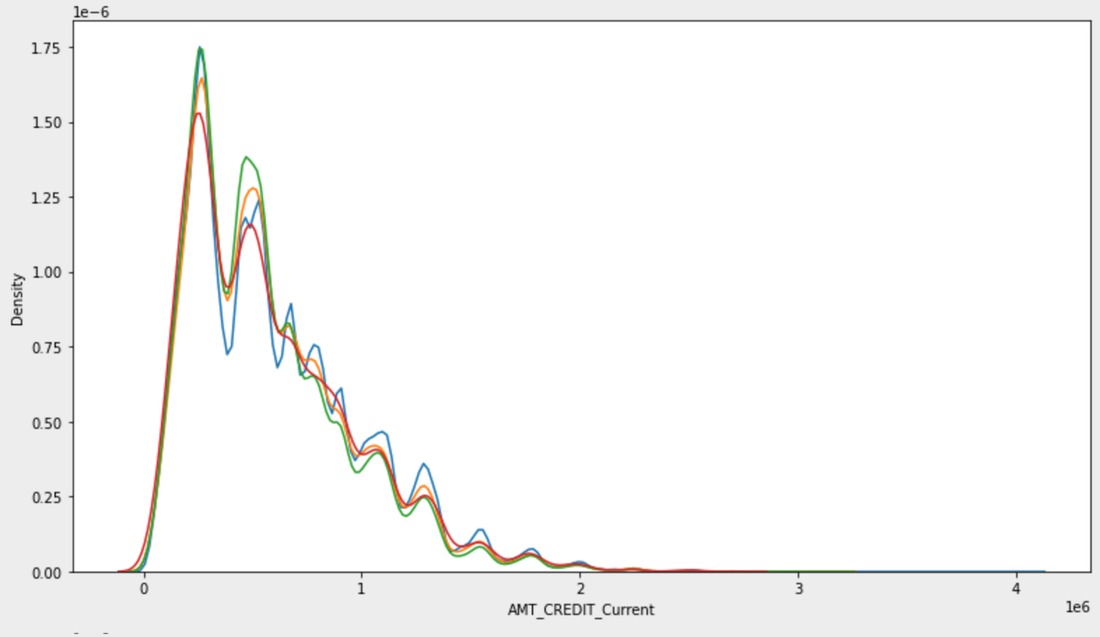


GRAPHS AND INSIGHTS

Loan Amounts in Current vs. Previous Applications

- The density distribution of credit amounts in the current and prior applications is shown in the AMT_CREDIT_Current vs. AMT_CREDIT_Previous KDE plot.
- While many customers keep their credit ranges same from one application to the next, some apply for far larger sums in later loans.

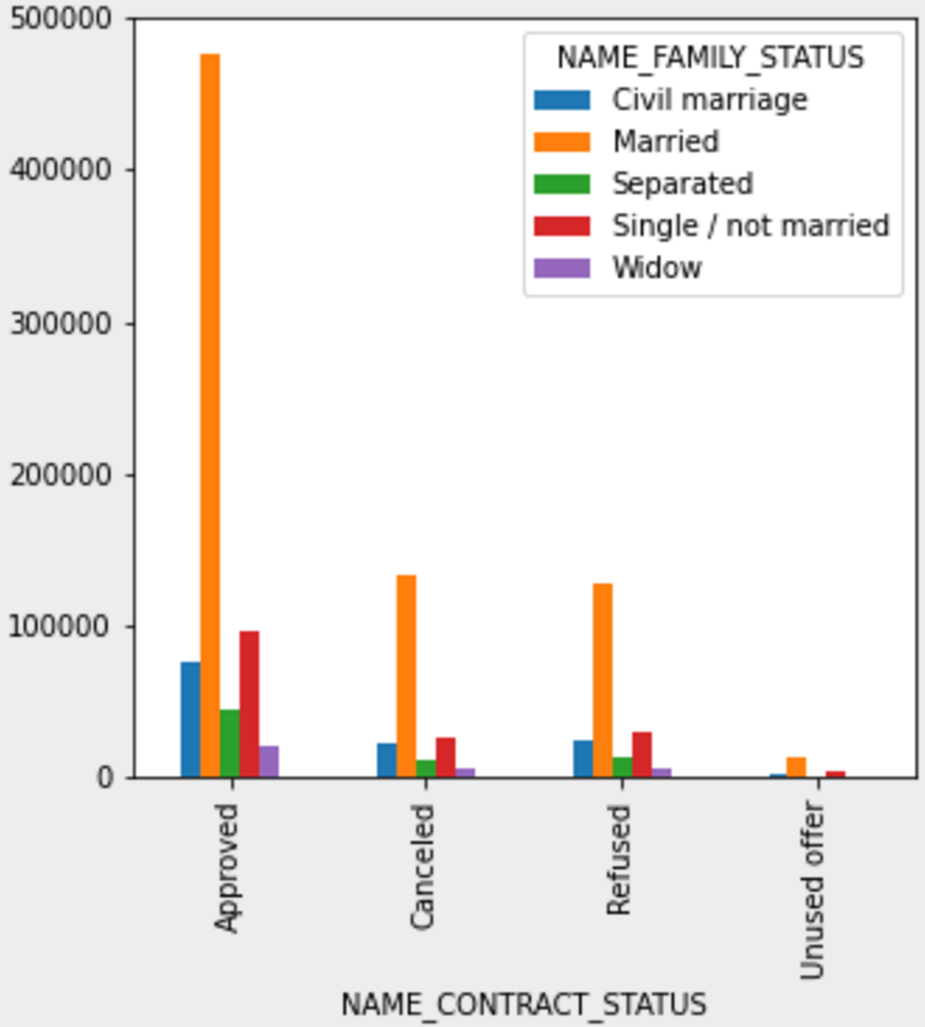




GRAPHS AND INSIGHTS

Family Status and Loan Status

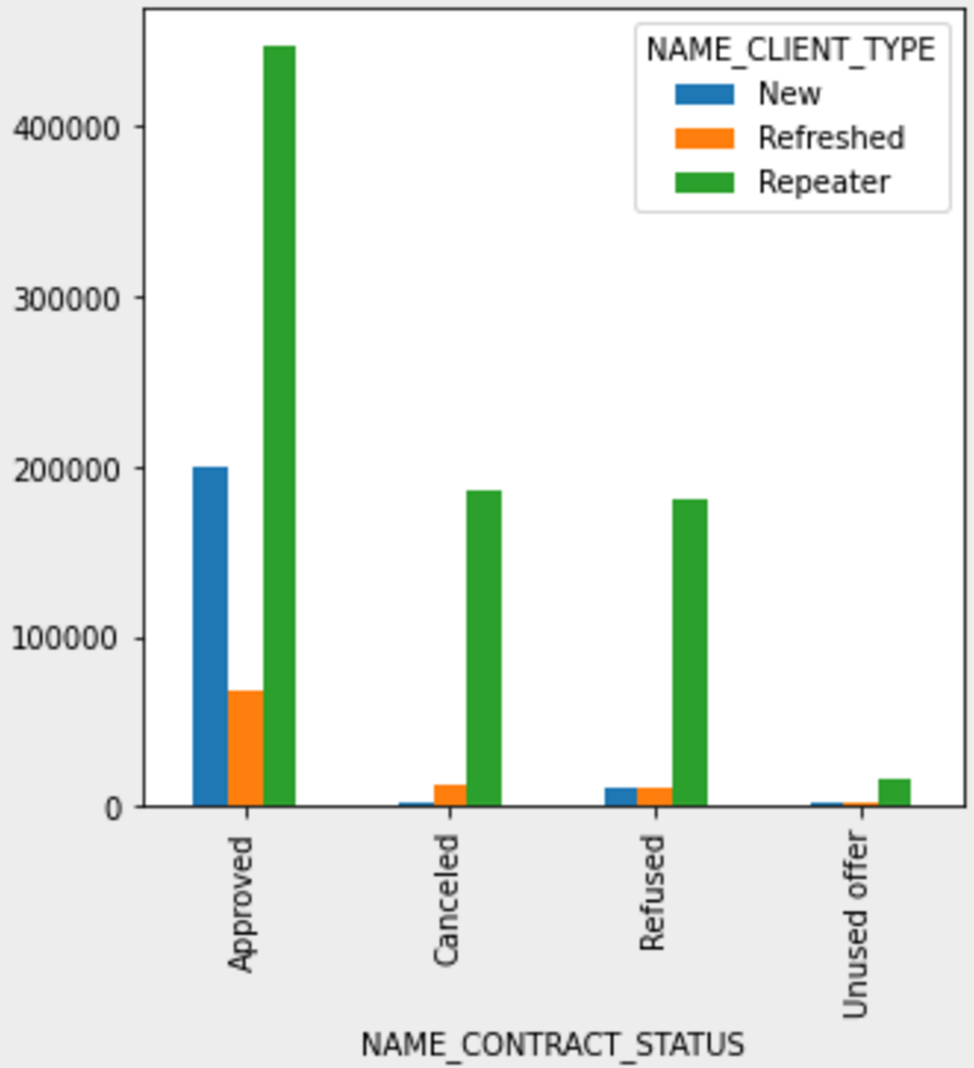
- The NAME_FAMILY_STATUS vs. NAME_CONTRACT_STATUS bar chart shows the applicants' family status for each loan status.
- While single applicants have a greater denial rate, married applicants make up the majority of loans that are accepted.
- The stability of the family may have an impact on loan acceptance.



GRAPHS AND INSIGHTS

Client Type and Loan Status

- NAME_CLIENT_TYPE versus NAME_CONTRACT_STATUS Outplot Distribution of loan statuses by client category (New, Repeater, etc.).
- Compared to new clients, repeat customers are more likely to get approved.
- Refusals are more likely to occur for new clients, underscoring the significance of thorough risk assessment for new applications.



CONCLUSION:

1. Default vs. Non-Default: Across all income and demographic groups, non-defaulters far outnumber defaulters.

2. Income and Credit: People with lesser incomes are more likely to default, whereas those who are not in default typically have greater income levels and larger credit amounts.

3. Demographics: Women apply for loans more often than males, and those who are married and have completed high school are more likely to do so.
Despite being the most accepted loan type, cash loans have a higher default rate than revolving loans.

5. Application Timing: Weekdays, particularly Tuesdays, see the highest volume of applications, while weekends see the lowest.

6. Correlation & Outliers: There are important relationships between employment, credit, and income; outliers are displayed by features like `AMT_CREDIT` and `CNT_CHILDREN`.

THANK YOU