Credit EDA Assignment

By Osheen Bindroo

Introduction

Objective: Perform Exploratory Data Analysis (EDA) on two datasets, 'application_data.csv' and 'previous_application.csv,' to gain insights into credit risk factors and loan approval trends in consumer finance. Create a presentation summarizing the analysis approach and key findings.

Background: Consumer finance companies often struggle with assessing creditworthiness and minimizing loan default risks. This analysis aims to utilize EDA to identify factors influencing loan defaults and improve lending decisions.

Four potential loan decisions:

- **Approved:** Loan application accepted.
- **Cancelled:** Client cancels, often due to changing terms.
- **Refused:** Loan rejected for various reasons.
- **Unused Offer:** Client cancels at different stages.

Two primary risks are associated with lending decisions:

- **Loss of Business:** Rejecting creditworthy applicants results in missed business opportunities.
- **Default Risk:** Approving loans to high-risk applicants can lead to financial losses.

Introduction

Datasets:

- 'application_data.csv' - Contains client information at the time of application, including data on payment difficulties.
- 'previous_application.csv' - Provides information on the client's previous loan data, categorized as Approved, Cancelled, Refused, or Unused offer.

Analysis Approach:

- **Handling Missing Data:** Identify and address missing data in both datasets, deciding whether to remove columns or replace missing values based on the data's nature.
- **Outlier Detection:** Detect outliers in the datasets and provide explanations for their significance. No data points will be removed at this stage.
- **Data Imbalance Assessment:** Determine if data imbalance exists, especially regarding the 'Target variable' (clients with payment difficulties and all other cases). Calculate the data imbalance ratio and use various plots to analyze it.
- **Univariate and Bivariate Analysis:** Conduct univariate, segmented univariate, and bivariate analyses to explore relationships and insights within the datasets.
- **Top Correlations:** Find the top 10 correlations for each segment of the data (Client with payment difficulties and All other cases) and discuss any meaningful insights.

Data Preprocessing

Identifying Missing and Incorrect Data:

- Imported the necessary libraries and read the dataset.
- Checked the data types and general information about the dataset using data2.info() and data2.describe().
- Identified columns with negative and positive values, particularly those related to days.

Handling Missing Values:

- Created a function null_values to calculate the percentage of missing values in each column.
- Checked for columns with more than 50% missing values and dropped them.
- Imputed missing values in the "NAME_TYPE_SUITE" column with the label "Unknown."

Transforming Negative Days to Positive Days:

- Converted columns related to days (e.g., "DAYS_DECISION," "DAYS_FIRST_DRAWING," etc.) to absolute values since negative days may indicate data entry errors.

Data Preprocessing

Creating a New Feature:

- Created a new feature called "YEARLY_DECISION" by categorizing the "DAYS_DECISION" column into bins representing different time periods.

Data Exploration:

- Explored the distribution of the "YEARLY_DECISION" feature, showing the percentage of loan applicants who submitted fresh loan applications within different time periods.

Summary Statistics:

- Checked the number of unique values in each column using data2.nunique().

Final Check for Missing Values:

- Checked for any remaining missing values in the dataset

Data Imbalance Analysis

Loading the Data:

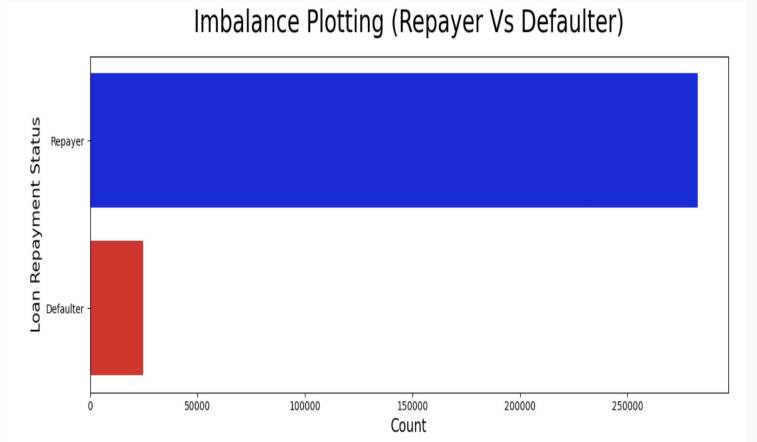
The code starts by reading the "application_data.csv" file into a Pandas DataFrame named data1.

Plotting Data Imbalance:

- It creates a bar plot to visualize the data imbalance between two classes: "Repayer" and "Defaulter." The data imbalance is based on the "TARGET" column in the dataset.
- The plot shows the count of each class on the y-axis and the counts of "Repayer" and "Defaulter" on the x-axis, using different colors (blue and red) for each class.
- The plot includes labels, titles, and appropriate formatting for better readability.

Calculating Imbalance Percentages:

- The code calculates and prints the percentage of "Repayer" and "Defaulter" in the dataset.
- It also calculates and prints the imbalance ratio between "Repayers" and "Defaulters."

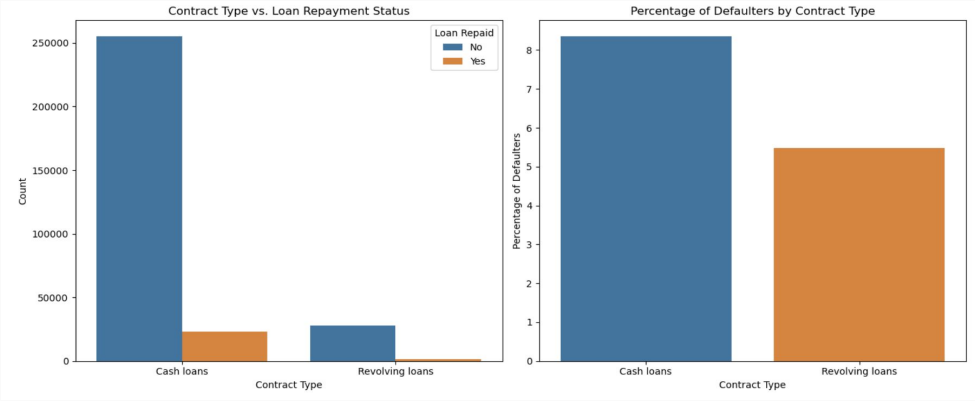


Key Takeaways

- Repayer Percentage is 91.93%
- Defaulter Percentage is 8.07%
- The imbalance ratio between Repayers and Defaulters is approximately 11.39/1.

Exploratory Data Analysis

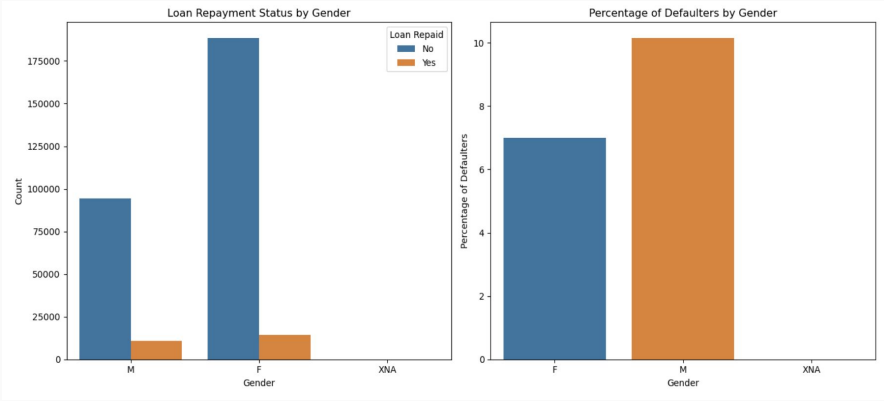
Contract Type vs. Loan Repayment Status



Key Takeaways

- Only 10% of all loans are revolving loans, which is a very tiny percentage.
- Approximately 8–9% of people who apply for cash loans and 5%–6% of people who apply for revolving loans are defaulters.

Gender vs. Loan Repayment Status

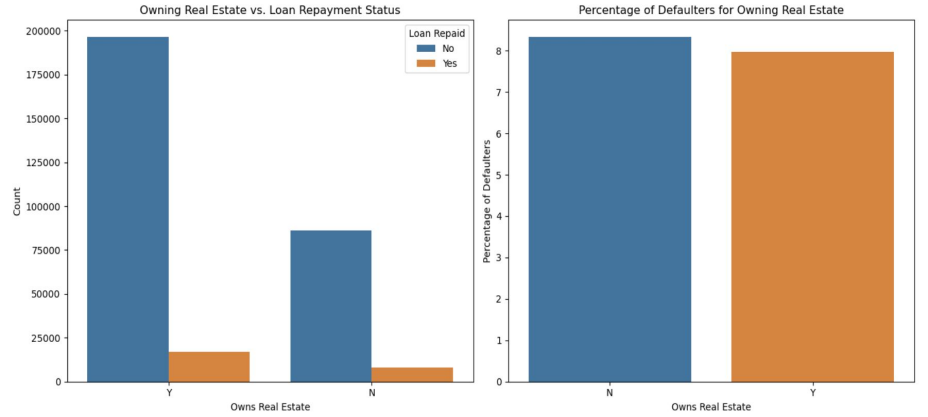


Key Takeaway

- The proportion of female customers is almost two times that of male customers.
- According to the rate of defaulted credits, men have a 10% higher probability than women do of not repaying their loans.

Exploratory Data Analysis

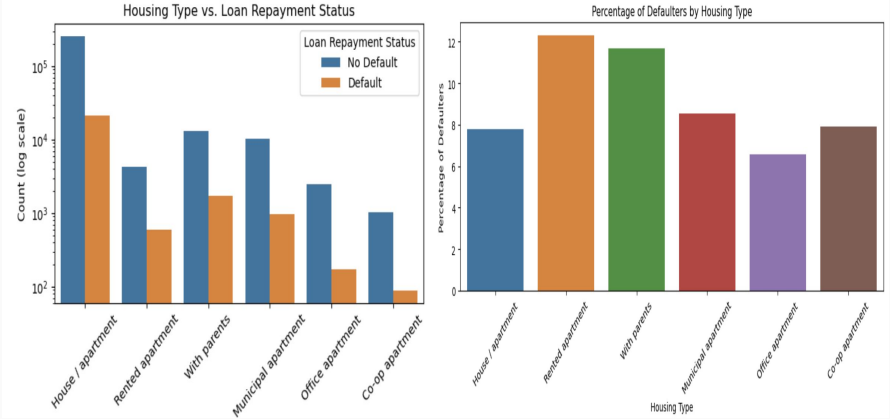
Ownership of Real Estate vs. Loan Repayment Status



Key Takeaways

- More than twice as many clients own real estate as those who do not.
- Both categories have around the same (8%) defaulter rates. Therefore, it follows that there is no corelation between owning a real estate and making a debt default.

Housing Type vs. Loan Repayment Status

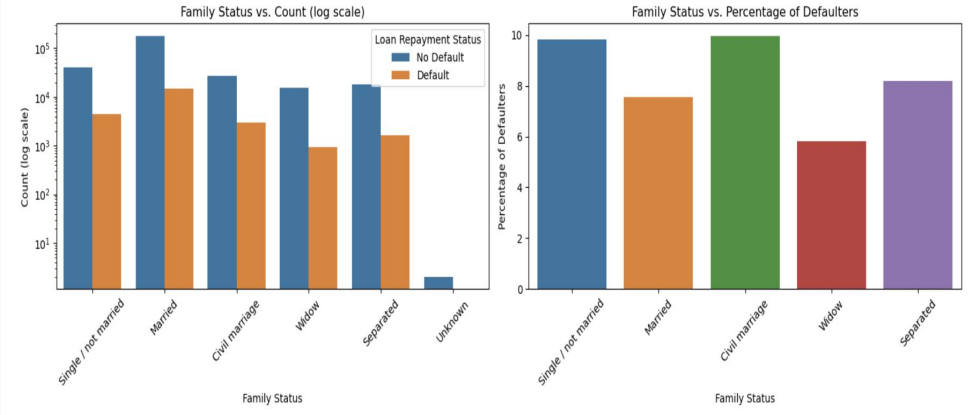


Key Takeaway

- Most individuals reside in homes or apartments.
- The lowest rate of default is among those who live in office apartments.
- People who live with their parents (11.5%) and in rented flats (>12%) are more likely to defaulting.

Exploratory Data Analysis

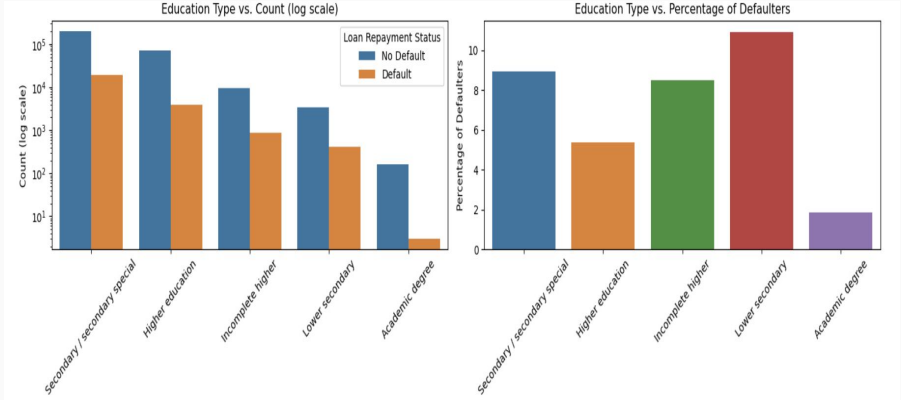
Family Status vs. Loan Repayment Status



Key Takeaways

- The majority of those who have taken out loans are married, followed by single/unmarried and civil marriage.
- Civil marriage has the highest percentage of defaulters (about 10%), while widows have the lowest (approximately 6%) (exception being Unknown).

Education Type vs. Loan Repayment Status

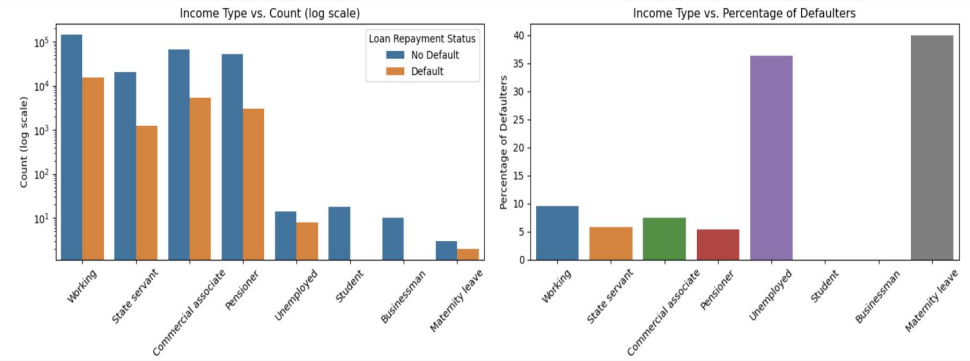


Key Takeaway

- Those with secondary or secondary special education make up the majority of clients, followed by those with higher education.
- Academic degrees are very rare among clients
- A high rate of default occurs in the lower secondary category, around 11%.
- Defaulter rates are lowest among people with academic degrees.

Exploratory Data Analysis

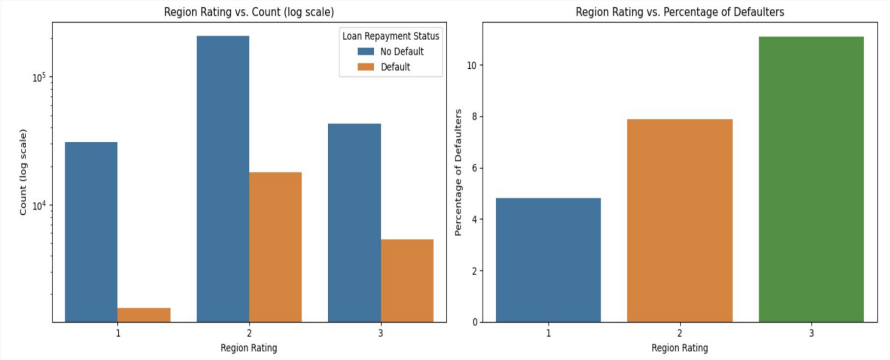
Income Type vs. Loan Repayment Status



Key Takeaways

- In terms of income type, the majority of applicants for loans have an income type of Working, followed by Commercial Associate, Pensioner, and State Servant.
- Maternity leave applicants have the highest defaulting rate of 40%, followed by unemployed applicants (37%). There are around 10% defaultees under the rest.
- Though they are fewer in number, students and businessmen do not have default records. Providing loans in these two categories is the safest option.

Region Rating vs. Loan Repayment Status

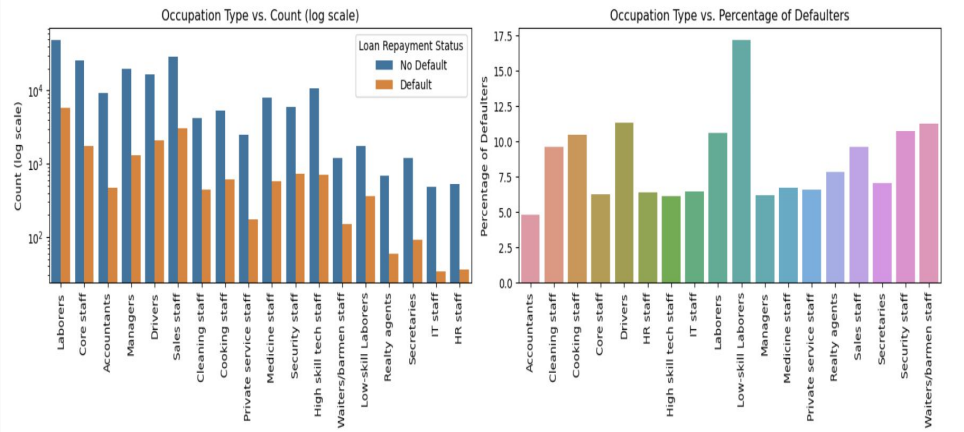


Key Takeaway

- The most common place of residence for applicants is a Rating 2 region. The default rate in Region Rating 3 is the highest (11%). An applicant residing in Region_Rating 1 has the lowest likelihood of default, making loan approval safer.

Exploratory Data Analysis

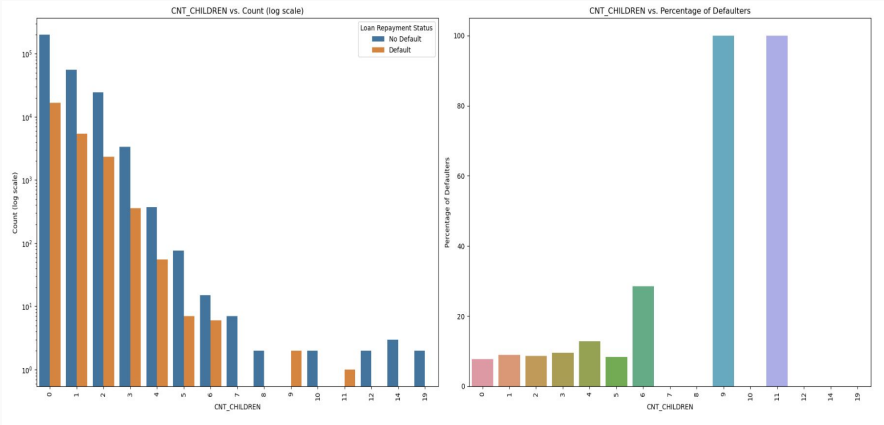
Occupation Type vs. Loan Repayment Status



Key Takeaways

- Laborers take out the most loans, followed by Sales staff. Less IT staff will apply for loans. Low-skill laborers had the highest percentage of fraud (over 17%), followed by drivers, waiters/bartenders, laborers, Security staff, and Cooking staff.

Children Count vs. Loan Repayment Status:

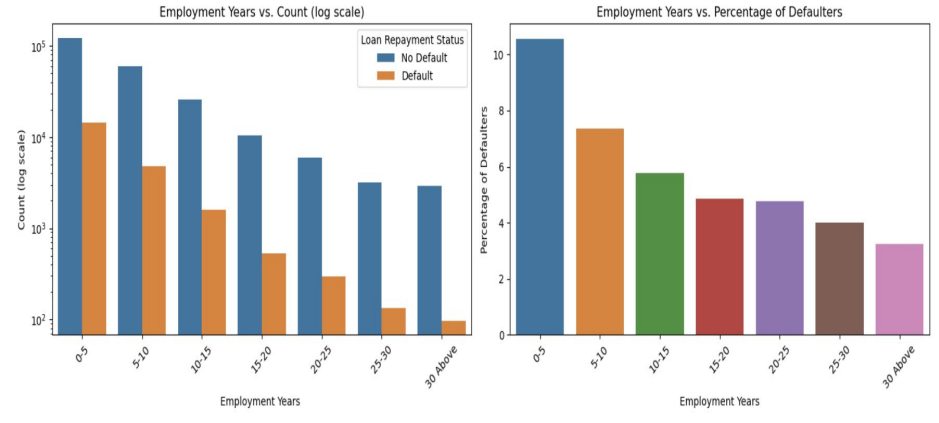


Key Takeaways

- The majority of applicants do not have any children.
- A small percentage of clients have more than three children.
- Clients with more than four children exhibit a notably high default rate, particularly those with child counts of nine and eleven, which both show a 100% default rate.

Exploratory Data Analysis

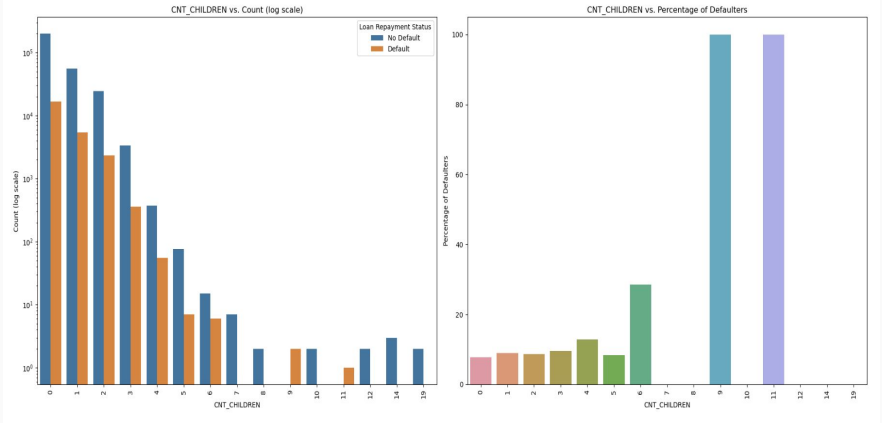
Employment Years vs. Loan Repayment Status



Key Takeaways

- The majority of candidates with 0 to 5 years of job experience are defaulters. This category also has the greatest default rate, which is roughly 10%. Defaulting rate is continuously falling as employment years grow. with employees having 40+ years of experience, the default rate is less than 1%.

Family Members Count vs. Loan Repayment Status

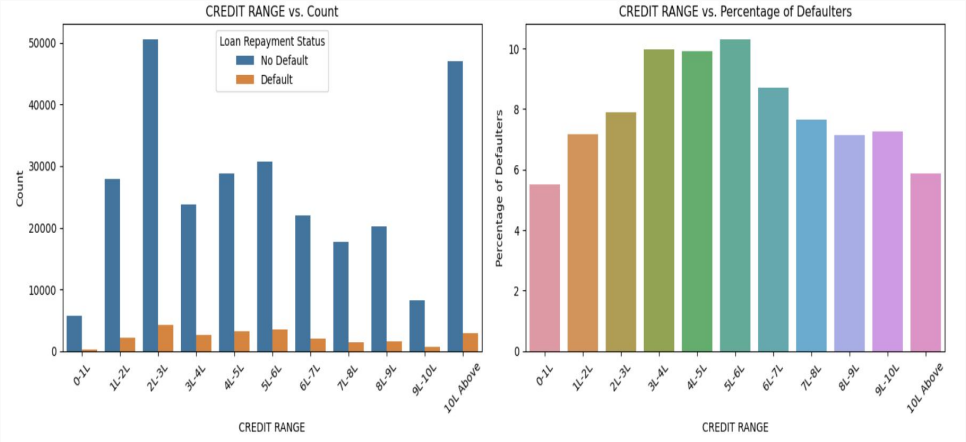


Key Takeaway

- The trend observed with family members is consistent with that of children, as having a higher number of family members also correlates with an increased risk of defaulting.

Exploratory Data Analysis

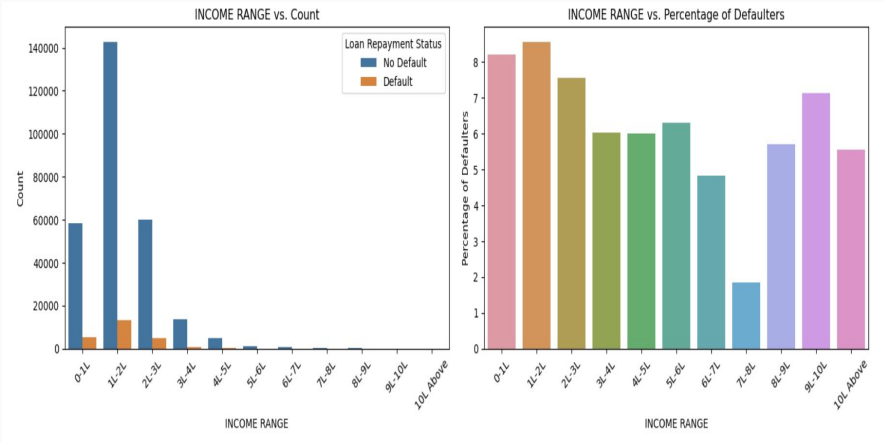
Credit Range vs. Loan Repayment Status



Key Takeaways

- A large proportion of applicants have loans between two and three lakhs, with 10 lakhs being the next highest amount. More people who receive loans between 3-6 lakhs default than those than other loan range.

Income Range vs. Loan Repayment Status

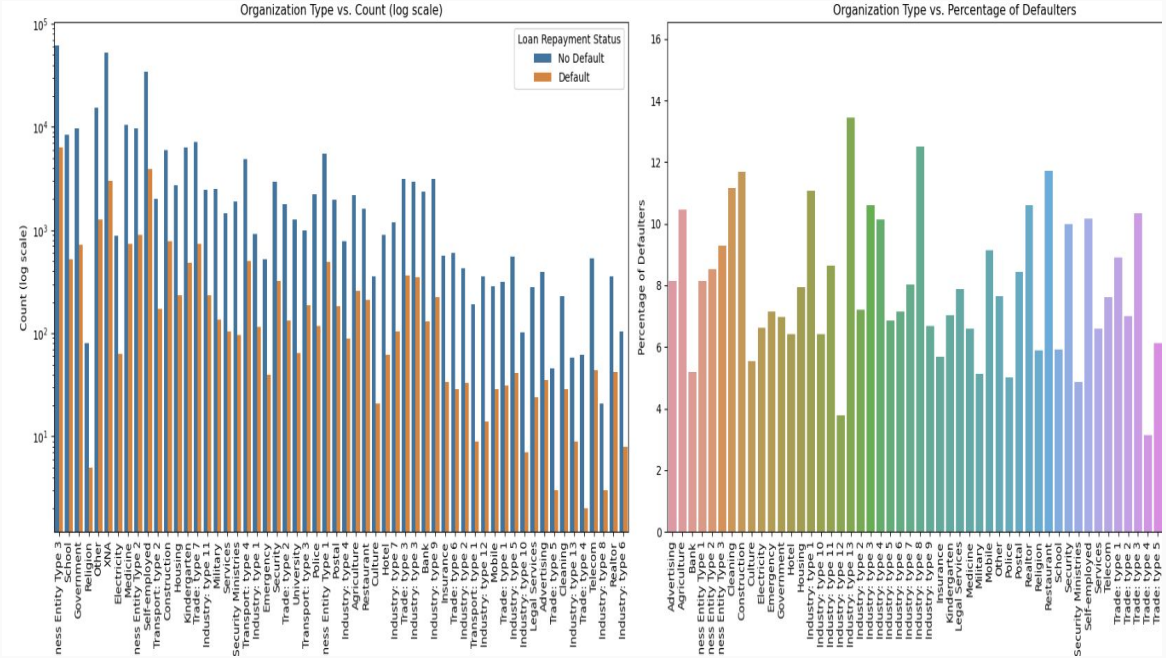


Key Takeaway

- The proportion of female customers is almost two times that of male customers.
- According to the rate of defaulted credits, men have a 10% higher probability than women do of not repaying their loans.

Exploratory Data Analysis

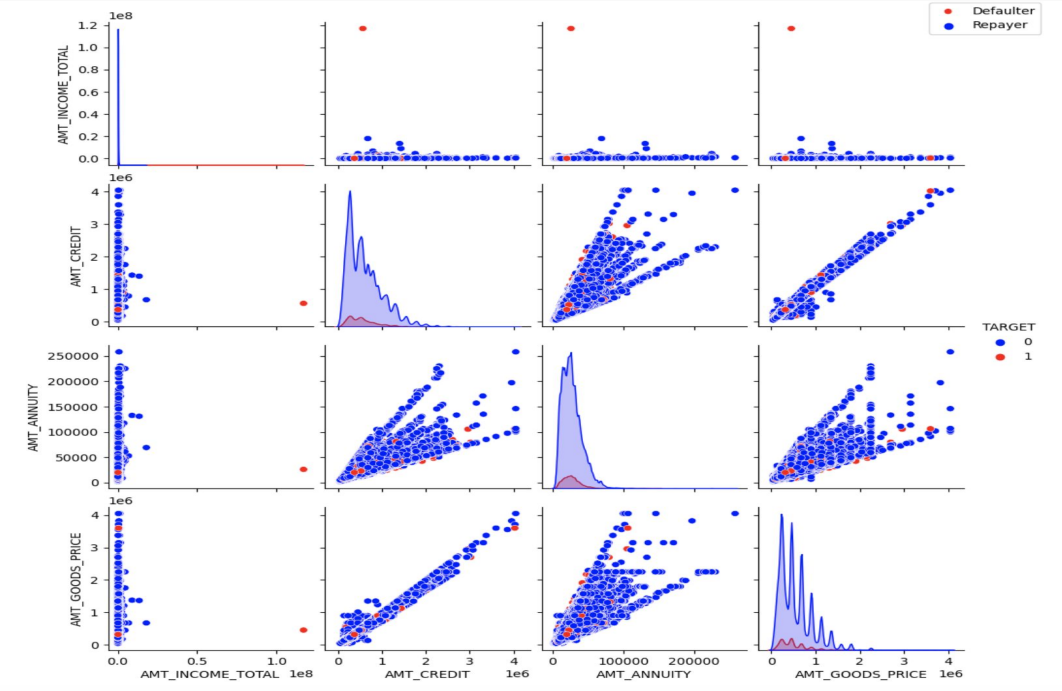
Organization Type vs. Loan Repayment Status



Key Takeaway

- Organizations with the Highest Default Rates:
- Among the various organizations, those with the highest percentage of defaulters are as follows: Transport: Type 3 (16%) Industry: Type 13 (13.5%) Industry: Type 8 (12.5%) Restaurant (less than 12%)
- Self-employed individuals exhibit a relatively higher rate of defaulting on loans.
- The majority of loan applicants are affiliated with Business Entity Type 3.
- Applications with Unavailable Organization Type: It's worth noting that a significant number of loan applications lack organization type information, marked as 'XNA' in the dataset.
- Safer Organization Types for Loan Disbursement: Certain organization types tend to have lower default rates, making them safer options for providing loans. These include Trade Type 4 and 5, as well as Industry Type 8.

Univariate and Bivariate Analysis



Key Takeaways

Annuity Amount and Good Price Amount:

When the annuity amount exceeds 15,000 and the good price amount is greater than 20 lakhs, the likelihood of defaulters is reduced.

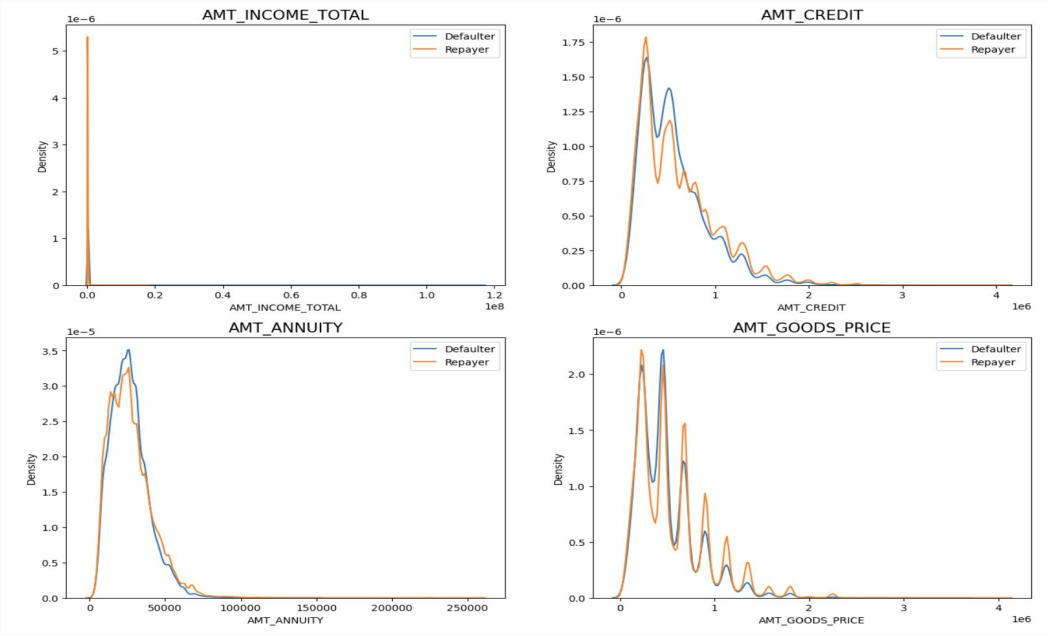
Loan Amount (AMT_CREDIT) and Goods Price (AMT_GOODS_PRICE):

- There is a strong positive correlation between loan amount (AMT_CREDIT) and goods price (AMT_GOODS_PRICE).
- The scatterplot reveals a concentration of data points forming a line, indicating a high correlation.

Loan Amount > 20 Lakhs:

There are notably fewer defaulters when the loan amount (AMT_CREDIT) exceeds 20 lakhs.

Univariate and Bivariate Analysis



Key Takeaways

Loan Amount for Goods: The majority of loans are granted for goods priced below 10 lakhs.

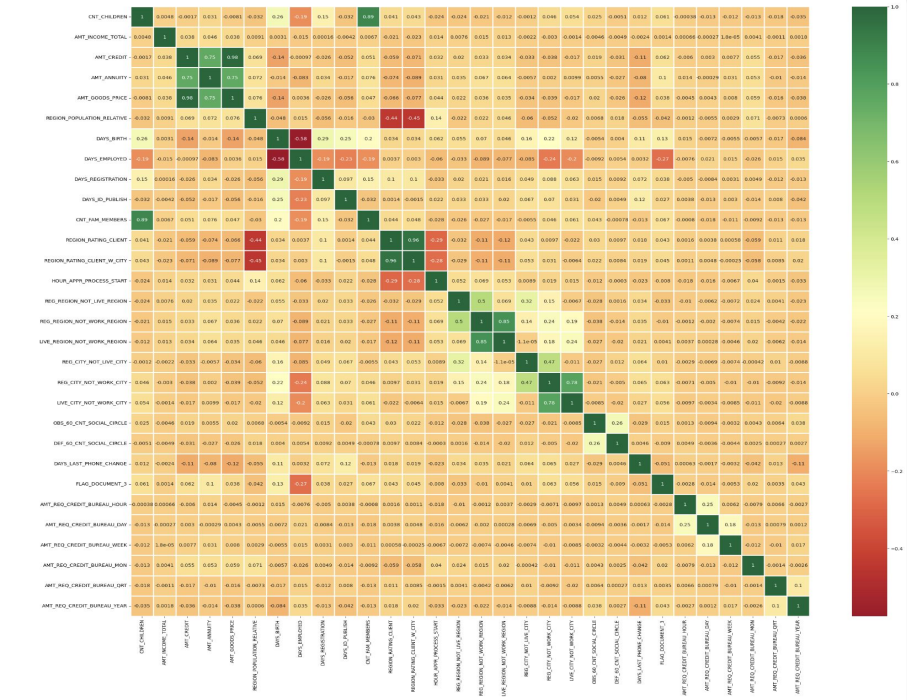
Annuity Payment: Most borrowers make annuity payments below 50,000 for their credit loans.

Credit Loan Amount: The credit loan amount typically falls below 10 lakhs for most borrowers.

Repayers and Defaulters:

- The distribution of repayers and defaulters overlaps in all the plots.
- Consequently, using any of these variables in isolation may not be sufficient to make a definitive decision.

Correlation Analysis



Key Takeaways

Credit Amount and Good Price Amount:

- Both are highly correlated, indicating a strong relationship.
- This correlation is consistent for repayers.

Loan Annuity and Credit Amount:

- In repayers, there is a strong correlation of 0.77.
- In defaulters, the correlation slightly reduces to 0.75.

Number of Days Employed:

- Among repayers, there is a relatively high correlation of 0.62.
- Among defaulters, the correlation is slightly lower at 0.58.

Total Income and Credit Amount:

- A notable difference is observed between repayers and defaulters:
- Repayers have a high correlation of 0.342.
- Defaulters show a significantly lower correlation of 0.038.

Days Birth and Number of Children:

- Among repayers, there is a correlation of 0.337.
- Among defaulters, the correlation reduces to 0.259.

Social Circle and Defaulted to Observed Count:

- For both repayers and defaulters:
- The correlation with defaulted to observed count is relatively low.
- Defaulters show a slightly higher correlation of 0.264 compared to repayers' 0.254.

Conclusion

A. Factors Indicating Repayment Potential:

- Education: Applicants with academic degrees tend to default less.
- Income Type:Students and businessmen have a low default rate.
- Region Rating: Rating 1 regions appear safer.
- Organization Type: Clients in trade types 4 and 5 and industry type 8 have defaulted less than 3%.
- Age (DAYS_BIRTH): People above 50 have a lower probability of defaulting.
- Employment Length (DAYS_EMPLOYED): Clients with 40+ years of experience have less than a 1% default rate.
- Income Total (AMT_INCOME_TOTAL): Applicants with incomes above 700,000 are less likely to default.
- Loan Purpose (NAME_CASH_LOAN_PURPOSE): Loans for hobbies and buying garages are most likely to be repaid.
- Children (CNT_CHILDREN): Borrowers with zero to two children are more likely to repay their loans.

B. Factors Indicating Default Risk:

- Gender (CODE_GENDER): Men have a relatively higher default rate.
- Family Status (NAME_FAMILY_STATUS): Civilly married or single individuals default more.
- Education (NAME_EDUCATION_TYPE): Lower secondary and secondary education levels have a higher default rate.
- Income Type: Maternity leave or unemployed clients tend to default.
- Region Rating: Regions with a rating of 3 have the highest defaults.
- Occupation Type: Avoid low-skill laborers, drivers, waiters/barmen staff, security staff, laborers, and cooking staff due to high default rates.
- Organization Type: Organizations with the highest loan default rates include Transport: type 3, Industry: type 13, Industry: type 8, and restaurants.
- Age (DAYS_BIRTH): Young people aged 20-40 have a higher probability of defaulting.
- Employment Length (DAYS_EMPLOYED): Clients with less than 5 years of employment have a high default rate.
- Children (CNT_CHILDREN) & Family Members (CNT_FAM_MEMBERS): Clients with 9 or more children default 100%.

Conclusion

C. Factors for High-Interest Loans to Mitigate Default Risk:

- Housing Type (NAME_HOUSING_TYPE): Many loan applicants live in rented apartments or with parents, so offering loans with higher interest rates to this group could mitigate potential losses.
- Loan Amount (AMT_CREDIT): Loans between 3-6 lakhs tend to default more, so higher interest rates for this credit range may be advisable.
- Income (AMT_INCOME_TOTAL): Since 90% of applicants have incomes less than 3 lakhs and a high probability of default, offering them loans with higher interest rates could mitigate risk.
- Children (CNT_CHILDREN) & Family Members (CNT_FAM_MEMBERS): Clients with 4 to 8 children have a very high default rate, justifying higher interest rates.
- Loan Purpose (NAME_CASH_LOAN_PURPOSE): Loans for "Repairs" have the highest default rate. A significant number of applications with purposes like "Repair" or "Other" were previously rejected, indicating a high-risk perception.

Suggestions:

Previously Canceled Clients: Record the reasons for cancellation to negotiate terms with these repaying customers in the future and increase business opportunities.

Previously Refused Loan Applicants: Document the reasons for loan refusals to mitigate business losses, and consider offering loans to these clients who have now turned into reliable borrowers.