EDA of Loan Default Risk

Lujun Lyu & Brooke Stealey

Section 1 — Business & Problem Understanding

Project Overview

Goal: Identify repayment patterns among clients

Data Sources :

- application_data.csv (current application information)
- previous_application.csv (loan history)
- columns_description.xlsx (data dictionary)

Focus:

- Applicant demographics
- Asset ownership & housing
- Loan amount & income distributions
- Credit bureau request behavior
- Previous application history
- Default rate

Section 2 — Data Understanding

Section 2 - Shape

First Steps :

- Loaded the three datasets into the notebook
- Confirmed their number of rows and columns
- Checked variable types (numerical vs categorical)
- Looked at functions .head(), .info(), and .describe() to see the shape/any patterns

Section 2 - application_data.info()

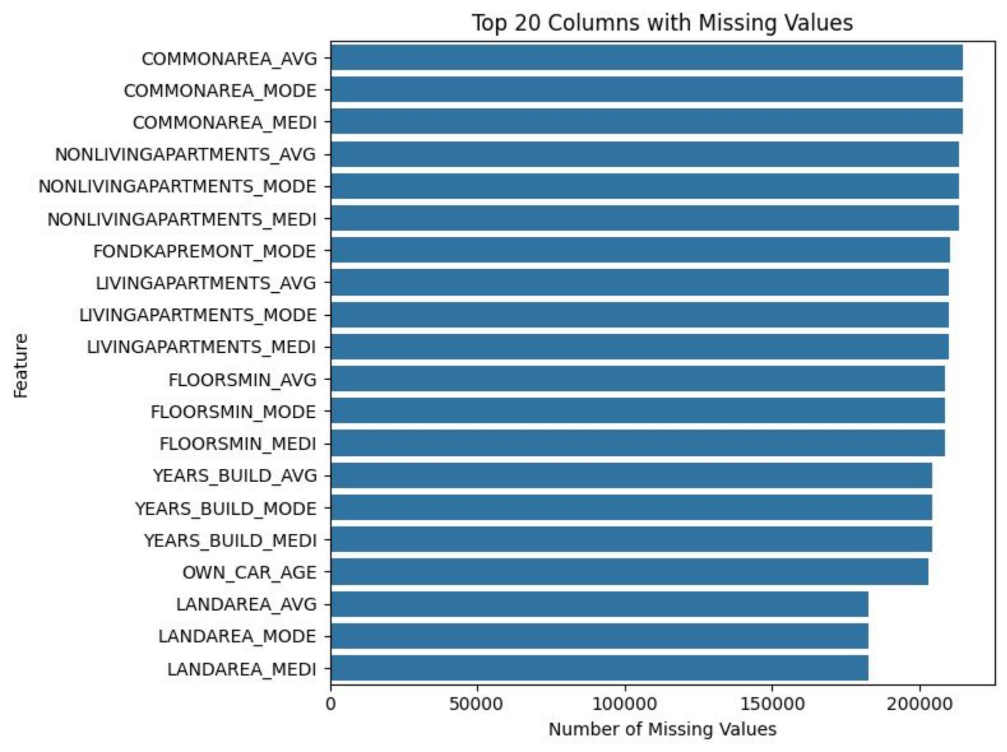
```
application_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

What this text shows :

- The application data set contains 307,511 rows and 122 columns
- Includes a mix of numerical (float/int) and categorical (object) variables
- Several columns have missing values (see next slide)

Section 2 - Missing Values



- Missing values were identified using summary functions to find which of the top 20 columns had the largest gaps

Results :

  - Some features have extensive missing data
  - Most missing columns are tied to housing/apartment data

Section 2 - Preparation Before Analysis

Steps Completed :

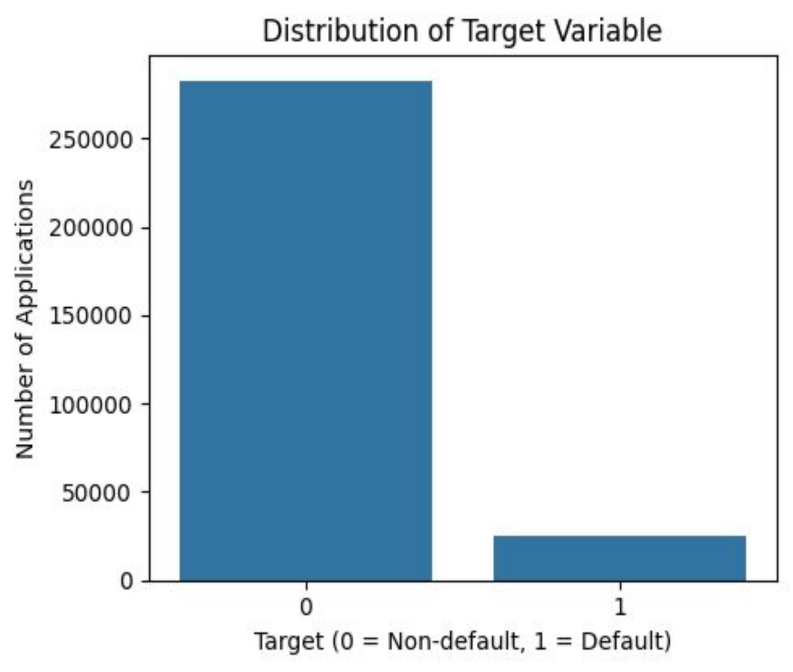
- Abbreviated the data set names for simpler codes
- Verified that SK_ID_CURR can be used to merge current/past records
- Identified financial variables with strong skew
- Reviewed data structure

★ These steps make sure the dataset is structured properly before moving into the main EDA

Section 3 — Univariate Analysis

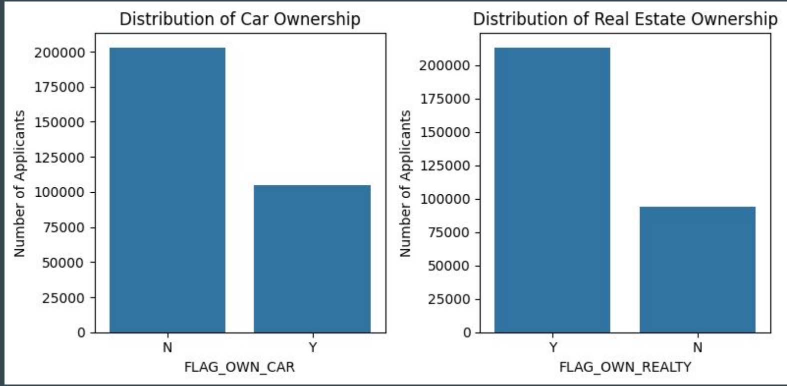
Section 3: Target Variable & Class Imbalance

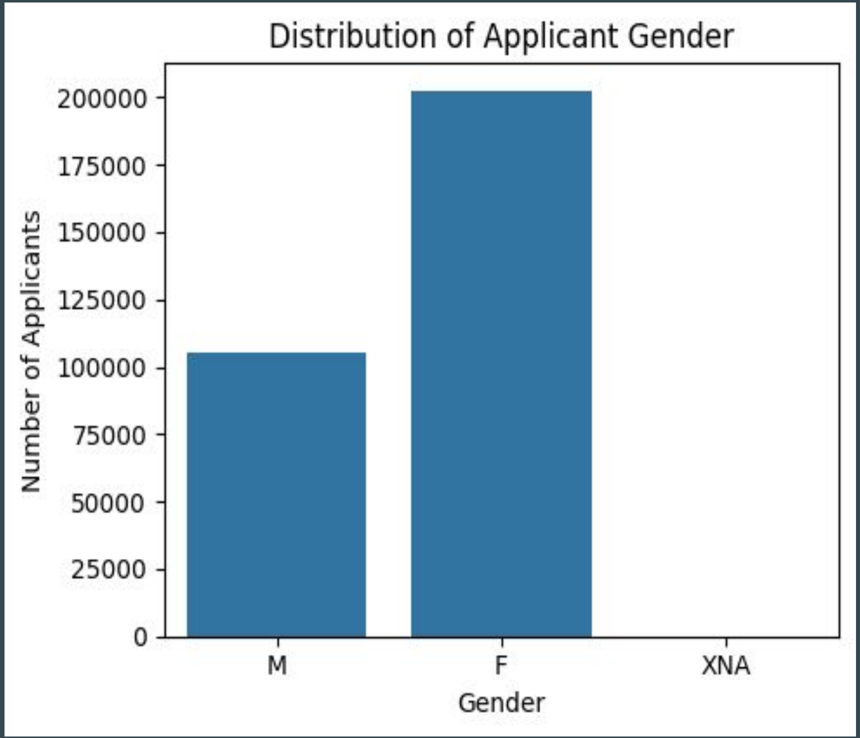
- The target variable indicates whether an applicant experienced payment difficulties.
- About **8%** of applicants defaulted, while **92%** repaid on time.
- This strong class imbalance is typical in credit risk data and provides important context for later analysis.



Section 3: Applicant Profile Demographics & Assets

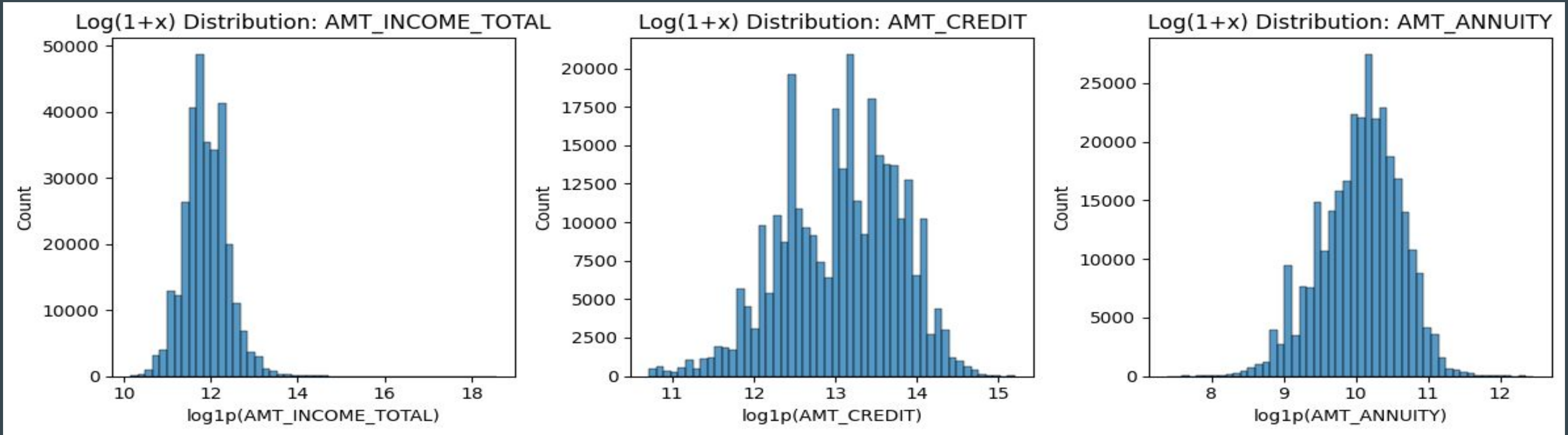
- Applicant population is **gender-skewed** , with more female applicants
- Most applicants report **no children** , indicating smaller households
- **Asset ownership varies** , reflecting heterogeneity in financial stability





Section 3: Financial Characteristics & Distribution Skew

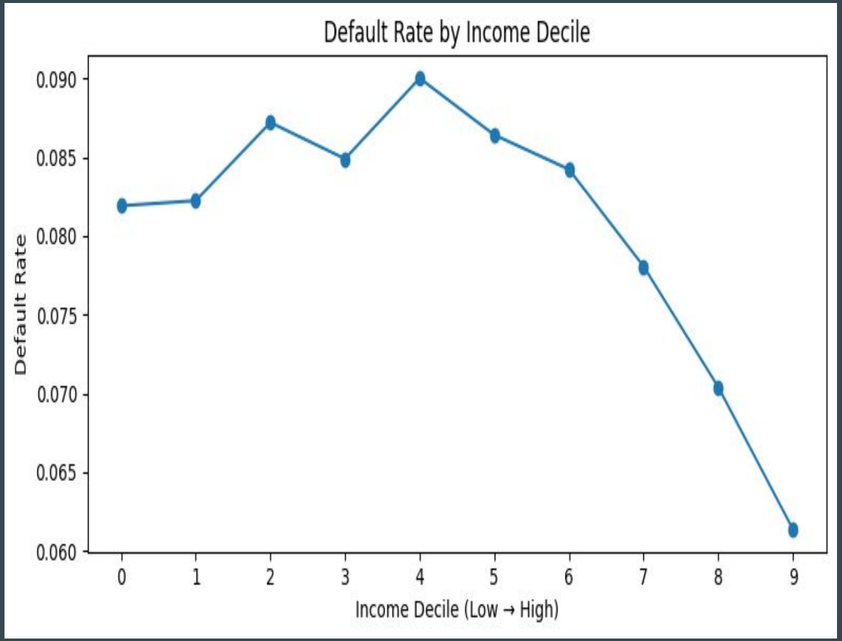
- Income, credit amount, and annuity exhibit strong right skew.
- Log transformation reveals clearer distributional structure.
- This scaling enables more meaningful comparison in later analysis.



Section 4 — Bivariate Analysis

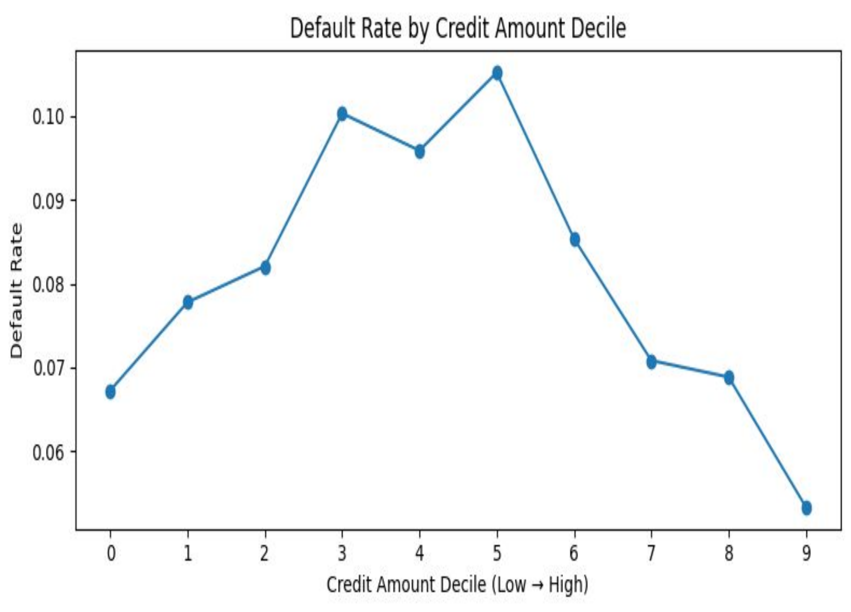
Section 4: Default Risk by Income Level

- Default rates are higher among lower-income applicants.
- Default probability generally decreases as income increases.
- However, income alone does not fully explain default behavior.



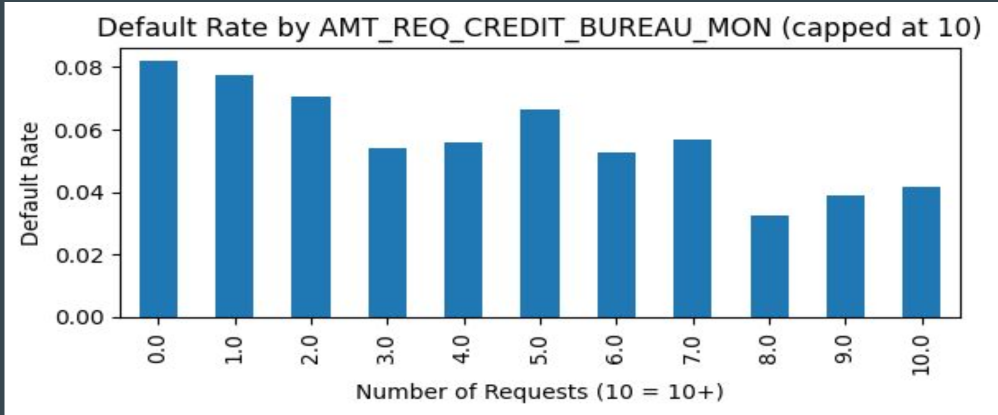
Section 4: Default Risk by Loan Size (Credit Amount)

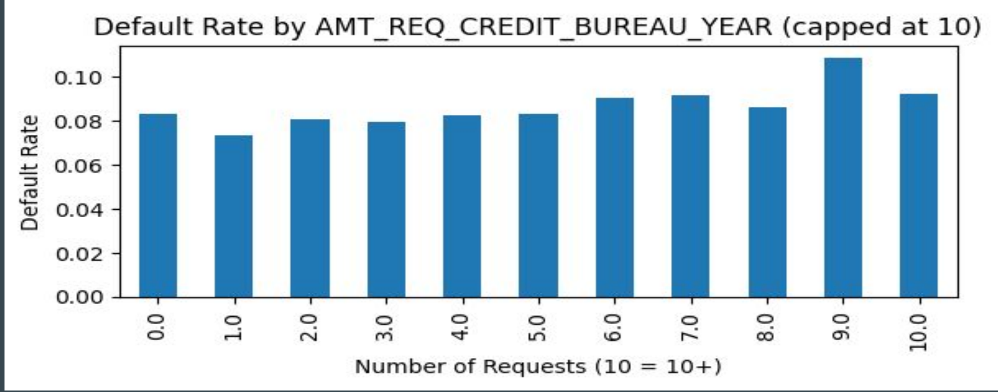
- Default risk varies non-linearly across loan size categories.
- Mid-sized loans display relatively higher default rates.
- Loan size alone is not a sufficient indicator of default risk.



Section 4: Default Risk by Categorical & Behavioral Features

- Default rates differ across **basic applicant characteristics** , such as gender and asset ownership.
- Applicants **without a car or real estate** tend to exhibit higher default rates.
- Credit bureau request behavior shows variation in default risk, suggesting recent credit activity may contain risk signals.

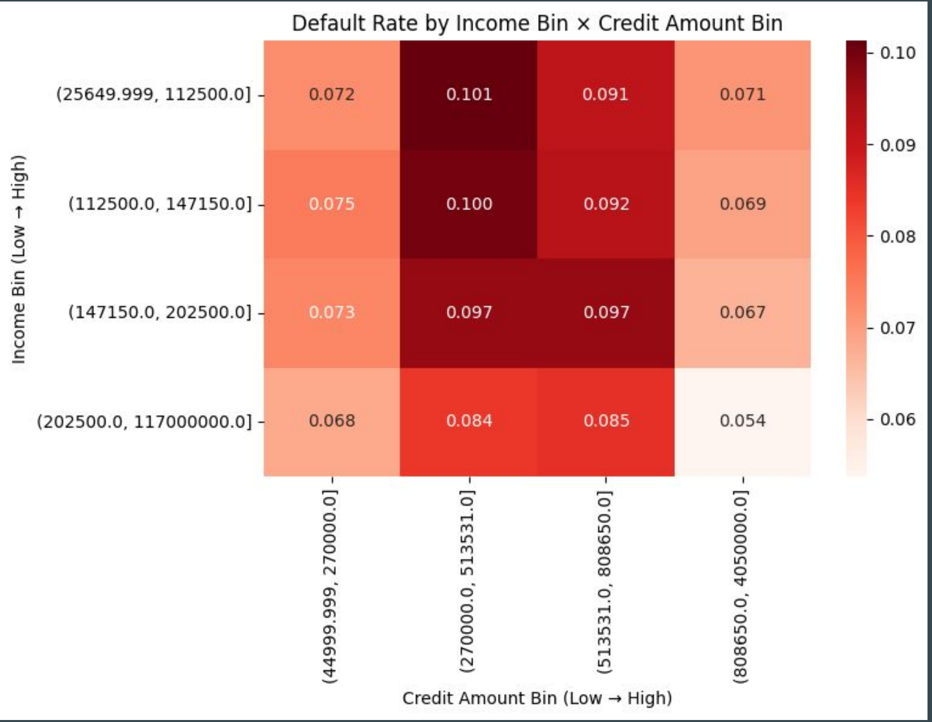




Section 5 — Segment and Interaction Analysis

Section 5: Interaction Analysis: Income × Credit Amount

- Default risk varies substantially across **income–loan size combinations** .
- **Low-income applicants with mid-to-high credit amounts** exhibit the highest default rates.
- High-income segments maintain relatively low default risk even at higher loan levels.

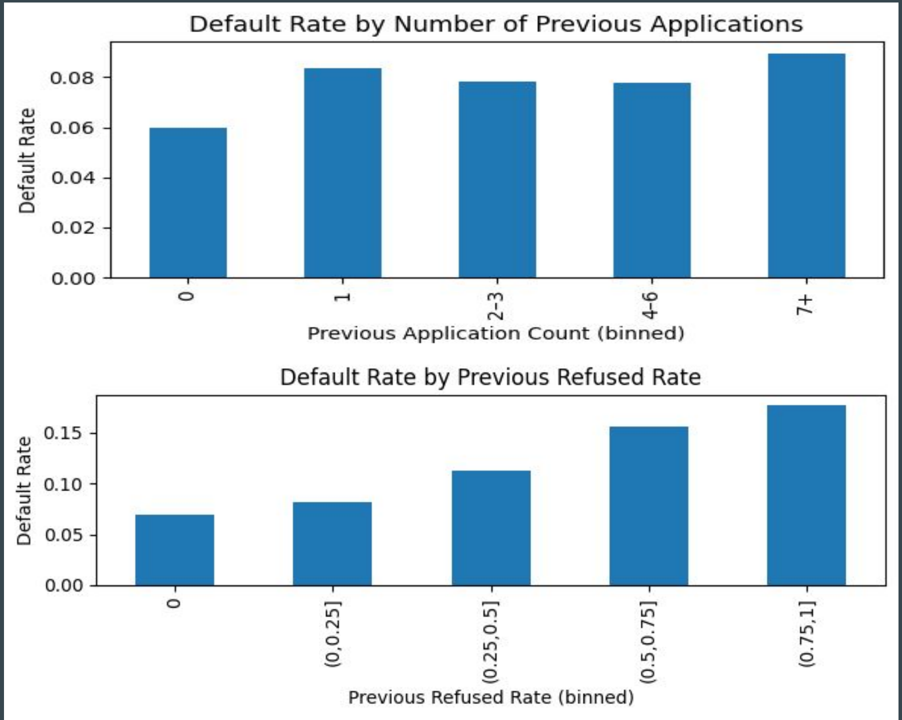


Section 5: Historical Credit Behavior & Default Risk

- Default risk increases with the number of previous loan applications.

- Applicants with higher historical refusal rates exhibit substantially higher default risk.

- Behavioral history provides strong predictive signal beyond current financial characteristics.



Section 6 — Key Insights & Conclusion

Section 6 - Insights & Conclusion

What we found :

- Lower income → higher default rates
- Medium-sized loans →highest risk
- Income & credit amount reveals high-risk groups
    - **Highest risk: low-income + medium credit amount**
    - **Lowest risk: high income + high credit amount**
- Past refusals → strongly linked to higher default
- Asset ownership & credit checks show financial stability of applicant

Conclusion :

EDA helps identify the factors most closely related to default risk, ultimately giving banks and lenders a foundation for stronger models and policies. These insights help drive more reliable, data-driven decisions when reviewing applicants.

Thank You!