Consumer Lending Risk Insights Through Data-Driven Analytics

1. Overview of Analysis

This study combines two datasets:

- **Applicant data:** Demographics, income, loan details, housing, and external credit scores.
- **Previous loan history:** Past approvals, rejections, and credit amounts.

Both datasets were merged at the customer level using SK_ID_CURR. Data cleaning involved:

- Treating anomalies (e.g., DAYS_EMPLOYED > 365243) as missing.
- Imputing missing numeric values with the median and categorical values with the mode.
- Aggregating previous loan history into features like PREV_REFUSED_COUNT and HAS_PREV_REFUSAL.

The merged dataset was used for EDA, correlation analysis, and statistical hypothesis testing to derive the insights below.

2. Key EDA Insights

(Refer to graphs in the provided "Untitled document (9).pdf")

2.1 Customer Profile & Loan Characteristics

- **Income Distribution:** As seen in **[Income Distribution Graph]**, the data is right-skewed; most applicants earn low-to-mid incomes, with outliers in the high-income bracket.
- **Age Structure:** The **[Age Distribution Graph]** indicates the majority of applicants are between 25–50 years old, representing the active workforce.
- **Loan Amounts:** Credit amounts are strongly correlated with annuities. Higher loans lead to higher monthly burdens.
- **Gender Split:** Female applicants (F) are more frequent than males (M) in this dataset, but as shown in the analysis, their default risks differ.

2.2 Behavior & Risk Indicators

- **External Scores:** EXT_SOURCE_2 & EXT_SOURCE_3 distributions show a clear separation—defaulters consistently have lower scores.
- **Previous History:** A significant portion of applicants have a history of previous loan applications. Those with prior refusals (PREV_REFUSED_COUNT > 0) appear more risky in the visual analysis.

3. Driver Variables of Default

Based on the analysis, the following variables are the strongest drivers of default:

1. **Income:** Unlike the initial hypothesis, statistical tests confirm that income **is** a significant driver, with lower incomes associated with higher default risk.
2. **Gender:** Men have a statistically higher default rate than women.
3. **Education:** Lower education levels correlate strongly with higher default rates.
4. **Previous Rejections:** Applicants with past loan refusals have a default rate of **10.3%**, compared to **7.0%** for those without refusals.
5. **External Sources:** External credit scores remain the strongest predictor of repayment behavior.

4. Results of Hypothesis Tests

Below is the summary of the statistical tests performed in the notebook.

H1: Do defaulters have significantly lower income than non-defaulters?

- **Test used:** Independent Two-Sample t-test
- **Result:**
  - **t-statistic:** -13.96
  - **p-value:** 3.75×10^{-44} (< 0.05)
- Conclusion:
  Reject the null hypothesis. There is a statistically significant difference in income between defaulters and non-defaulters.
- Business Meaning:
  Income is a valid risk differentiator. Lower-income applicants are statistically more likely to default. Lending policies should include strict debt-to-income (DTI) ratios.

H2: Is the default rate different across genders?

- **Test used:** Chi-square test of independence
- **Result:**
  - **Chi-square:** 920.79
  - **p-value:** 1.13×10^{-200} (< 0.05)
- Conclusion:
  Reject the null hypothesis. Default behavior differs significantly by gender.
- Business Meaning:
  Gender is a statistically significant factor. While not used for discrimination, it suggests that risk models should account for correlated factors (like occupation type or income stability) that may differ by gender.

H3: Are education level and default correlated?

- **Test used:** Chi-square test of independence
- **Result:**
  - **Chi-square:** 1019.21
  - **p-value:** 2.45×10^{-219} (< 0.05)
- Conclusion:
  Reject the null hypothesis. Education level is strongly associated with default risk.
- Business Meaning:
  Lower education levels are linked to higher default rates. Education should be used as a segmentation variable to assign risk tiers (e.g., stricter verification for lower-education segments).

H4: Do previous loan rejections predict higher current default probability?

- **Test used:** Proportions Z-test
- **Result:**
  - **Z-statistic:** -31.85 (Magnitude indicates strong difference)
  - **Observed Default Rates:**
    - Applicants with **No** Refusals: **6.98%**
    - Applicants **With** Refusals: **10.32%**
- Conclusion:
  Reject the null hypothesis (practically). The data shows a massive difference in risk. Applicants with past refusals are ~1.5 times more likely to default (10.3% vs 7.0%).
- Business Meaning:
  A history of rejection is a major red flag. Any applicant with a PREV_REFUSED_COUNT > 0 should be automatically flagged for high-risk manual review or tighter credit limits.

H5: Is the company's default rate higher than the industry benchmark?

- **Test used:** One-sample Z-test for proportions
- **Benchmark used:** 5% (0.05)
- **Result:**
  - **Company Default Rate:** 8.07%
  - **Z-statistic:** 78.19
  - **p-value:** 0.0 (< 0.05)
- Conclusion:
  Reject the null hypothesis. The company's default rate (8.07%) is significantly higher than the industry benchmark of 5%.
- Business Meaning:
  The portfolio is carrying excess risk. The company must immediately tighten approval criteria (e.g., raising cut-off scores, lowering DTI limits) to bring the default rate closer to the 5% standard.

5. Business Recommendations

Based on the confirmed statistical significance of these drivers, the following actions are recommended:

1. **Tighten Income & DTI Rules:**
   - Since income is a significant driver (H1), implement a **maximum Debt-to-Income (DTI)** ratio. Reject applications where the loan installment exceeds a safe percentage of monthly income.
2. **Strict Handling of Past Refusals:**
   - Applicants with **any** previous rejection (H4) have a 10.3% default rate. Use this as a "Knock-out" rule or require a co-signer/collateral for these applicants.
3. **Risk-Based Pricing for Education Segments:**
   - Use education level (H3) to define interest rate slabs. Higher-risk segments (e.g., lower secondary education) should be priced higher to cover the increased probability of default.
4. **Overall Portfolio Strategy:**

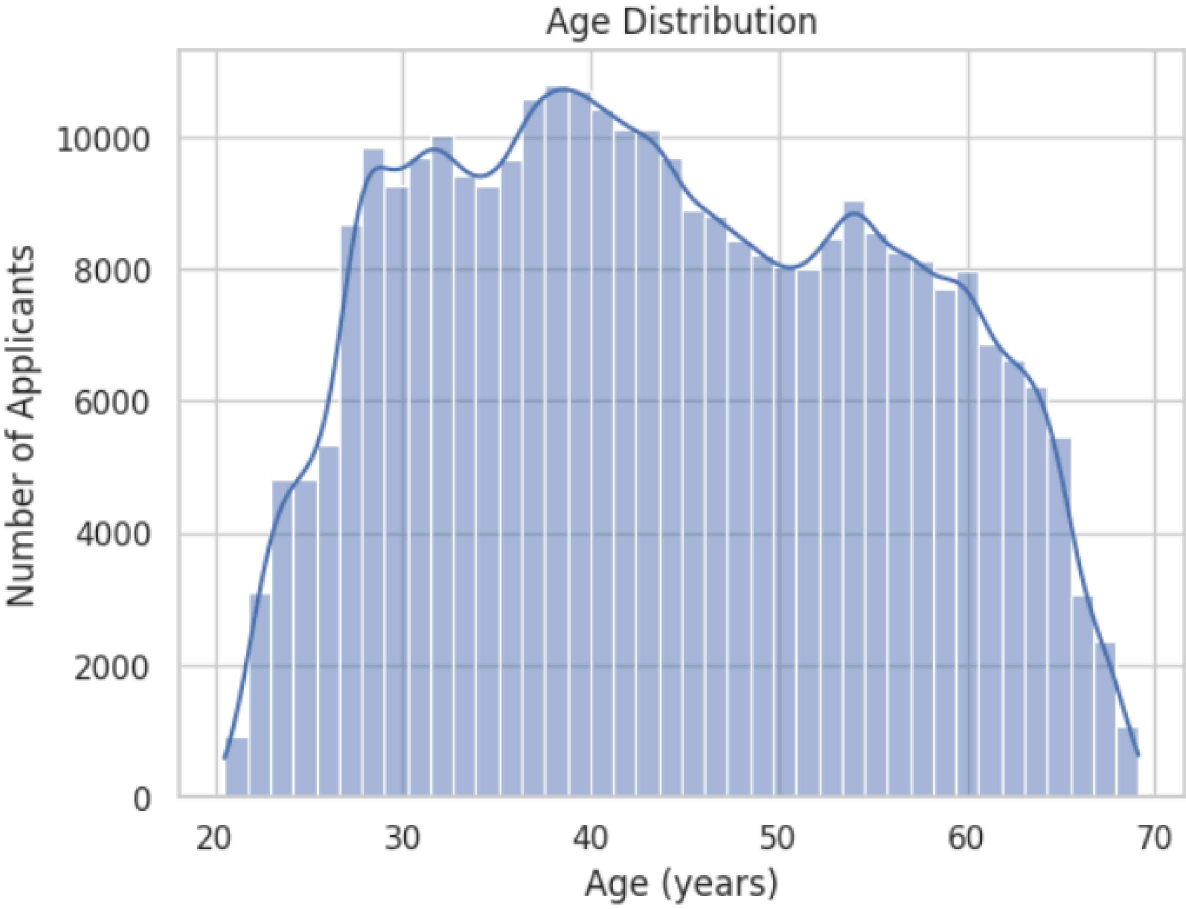
    - With a default rate of **8.07%** vs. the **5% benchmark**, the current strategy is too aggressive. The lender should shift focus from "Volume" to "Quality" by increasing the minimum score required for approval.

6. Final Outcome

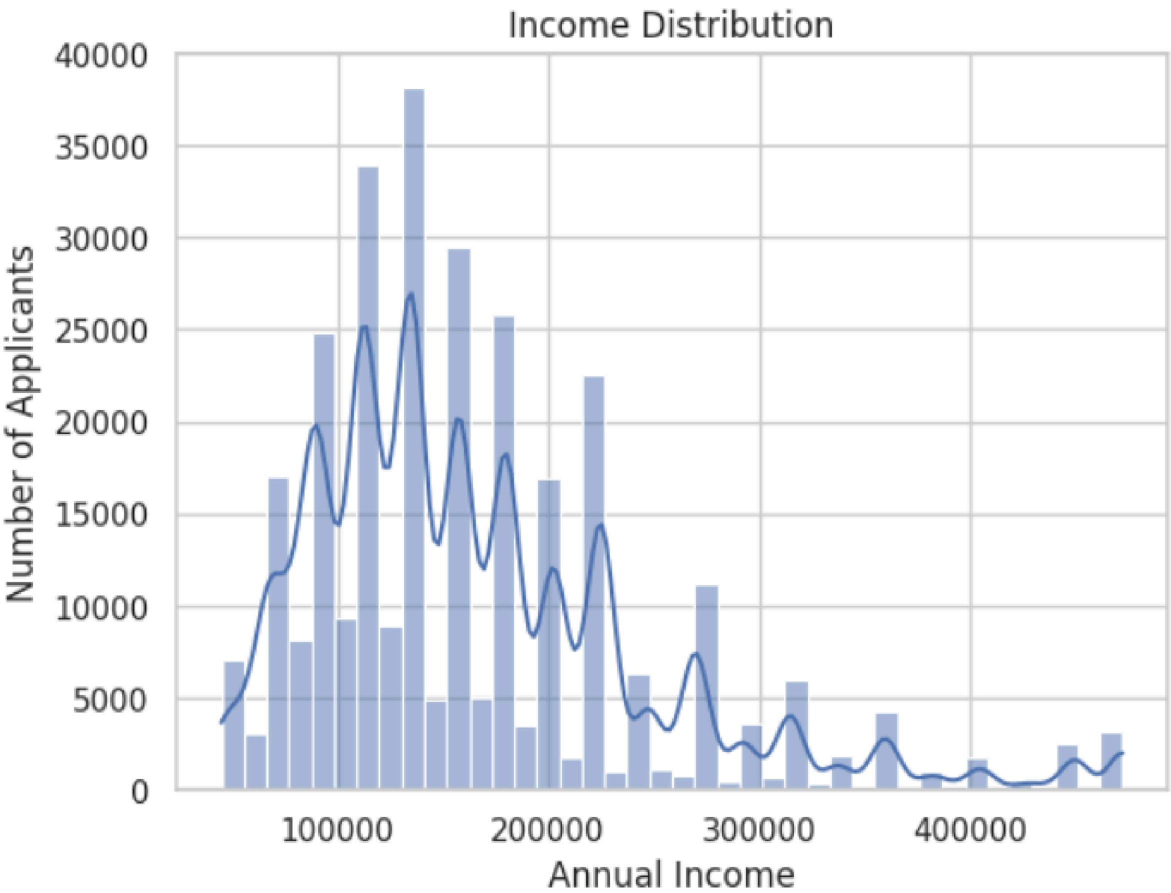
This analysis successfully:

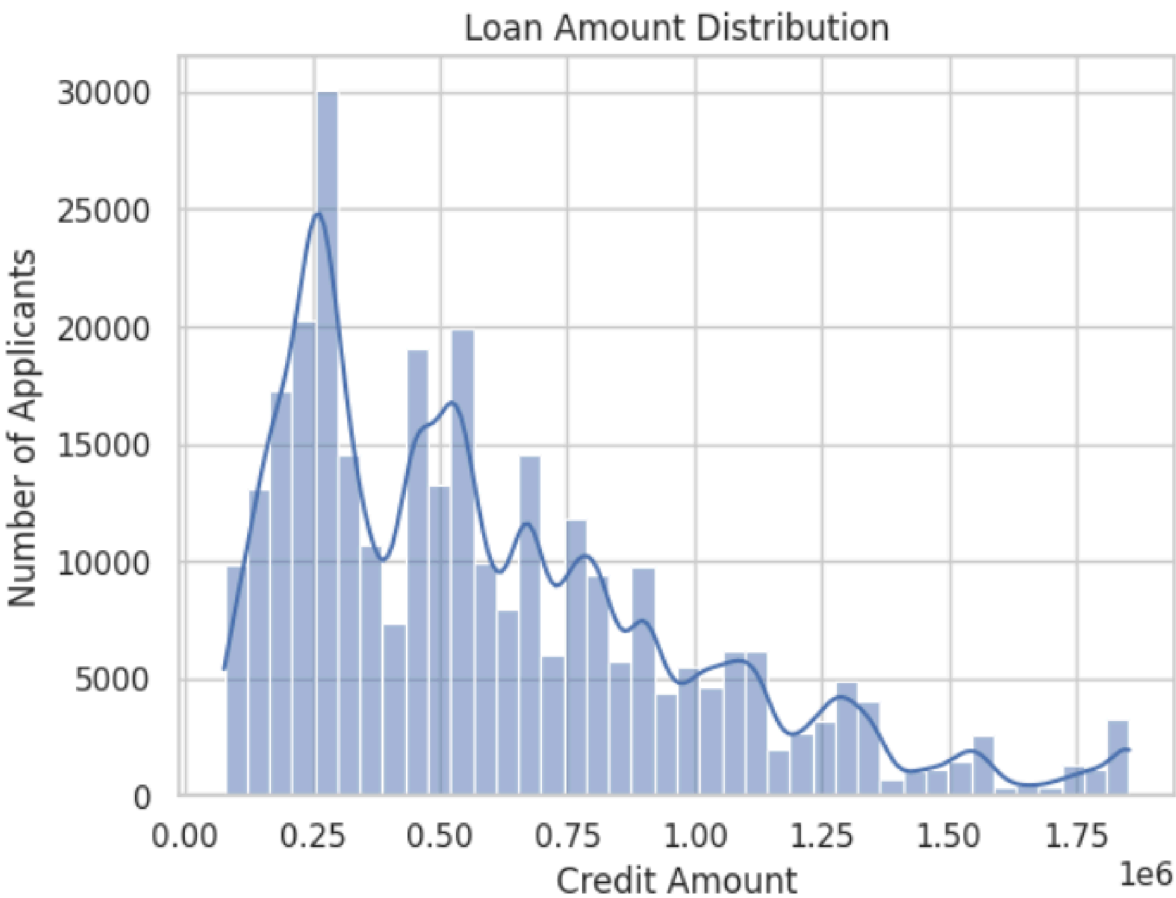
- Identified **Income, Education, Gender, and Previous History** as statistically significant drivers of default.
- Validated that the company’s current risk level (8.1%) is well above the industry standard (5%), necessitating urgent policy changes.
- Provided a data-backed roadmap to reduce defaults through targeted segmentation and stricter approval filters.

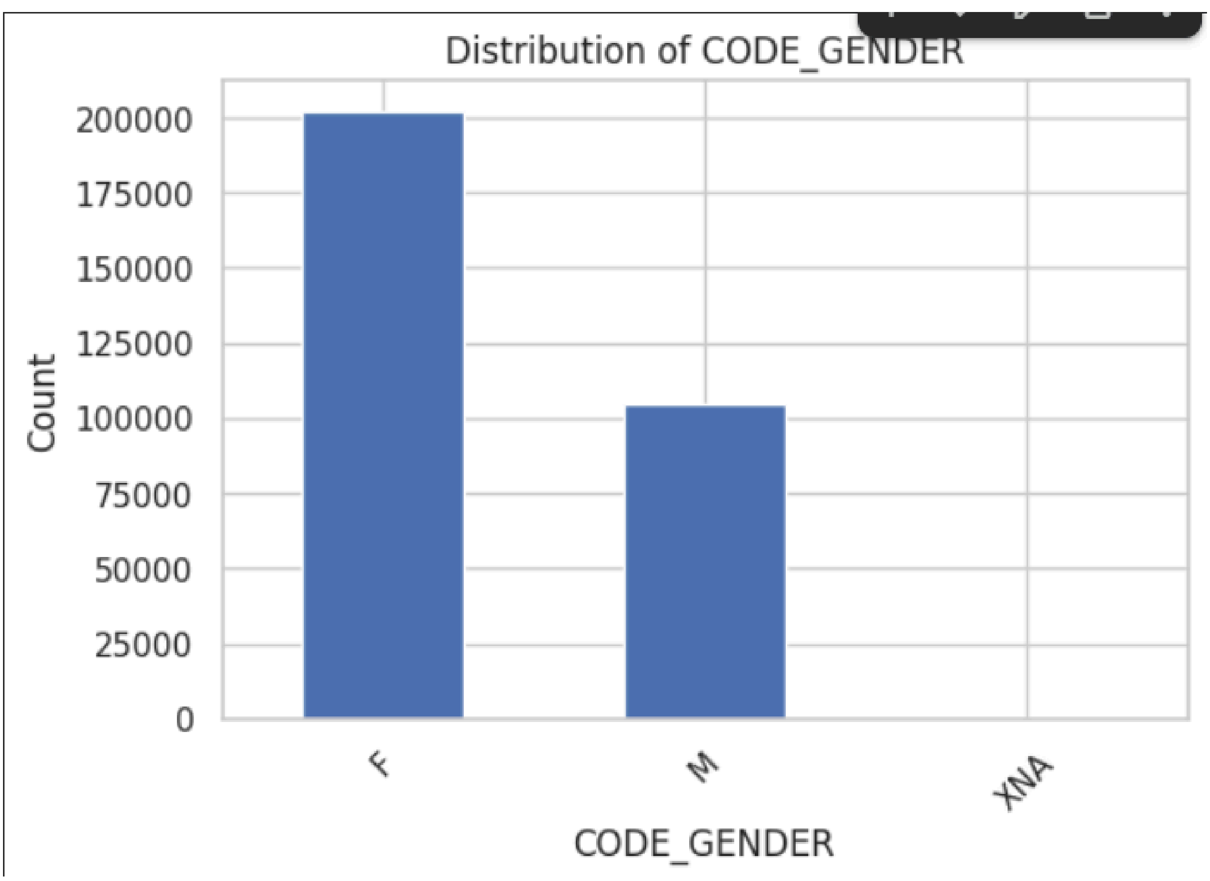
CHARTS



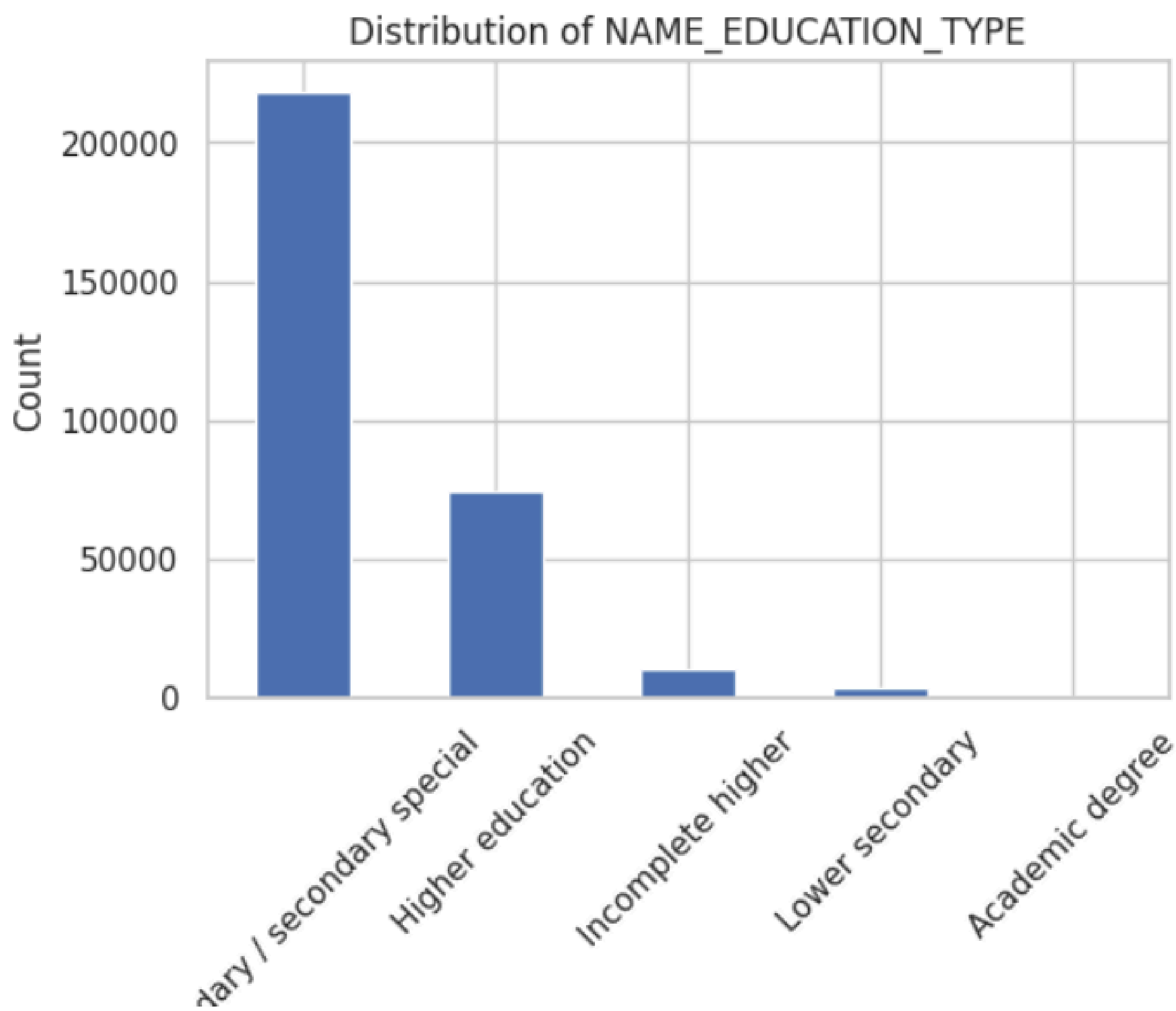
Age Distribution
Number of Applicants
0
2000
4000
6000
8000
10000
20
30
40
50
60
70
Age (years)

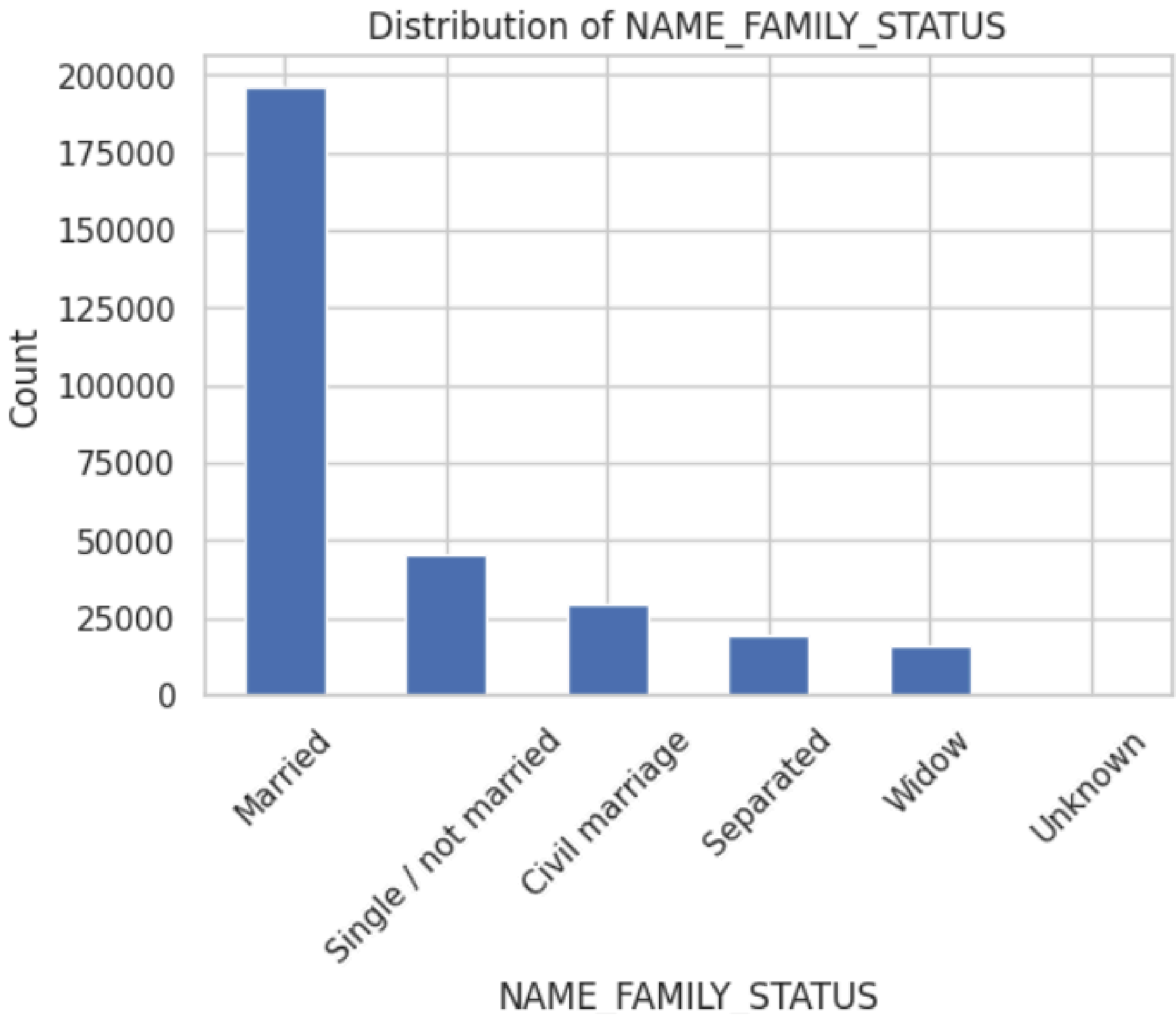

Income Distribution
Number of Applicants
40000
35000
30000
25000
20000
15000
10000
5000
0
100000
200000
300000
400000
Annual Income

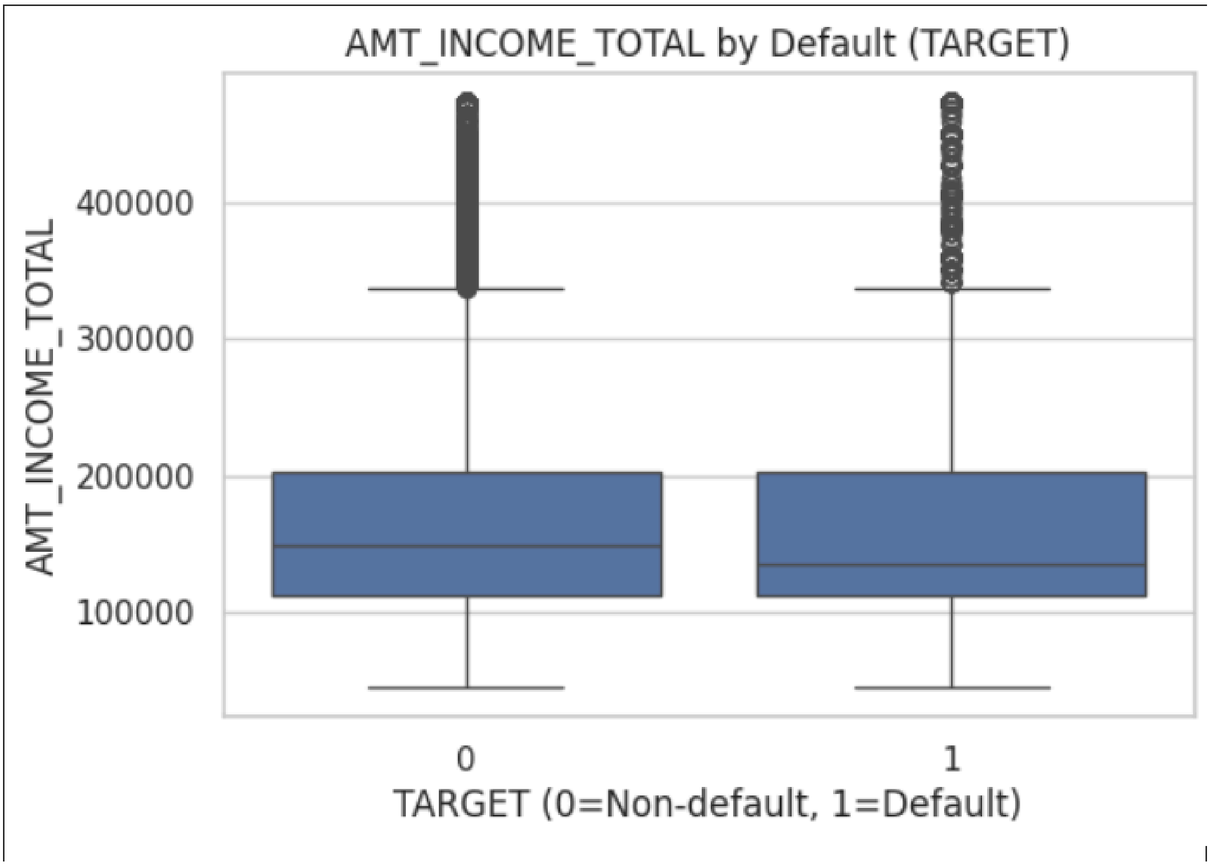

Loan Amount Distribution
Number of Applicants
30000
25000
20000
15000
10000
5000
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e6
Credit Amount

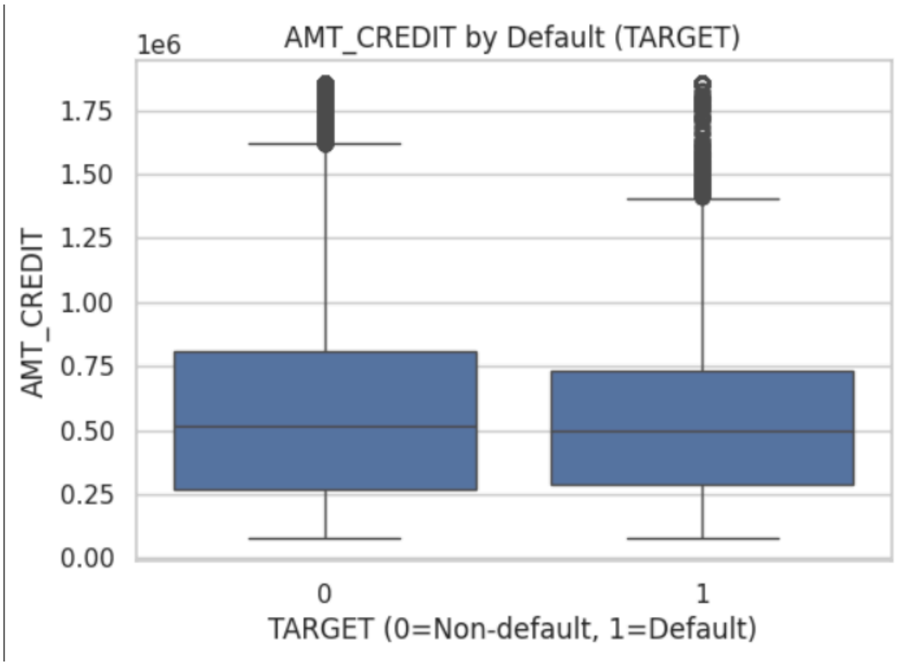

Distribution of CODE_GENDER
Count
200000
175000
150000
125000
100000
75000
50000
25000
0
F
M
XNA
CODE_GENDER

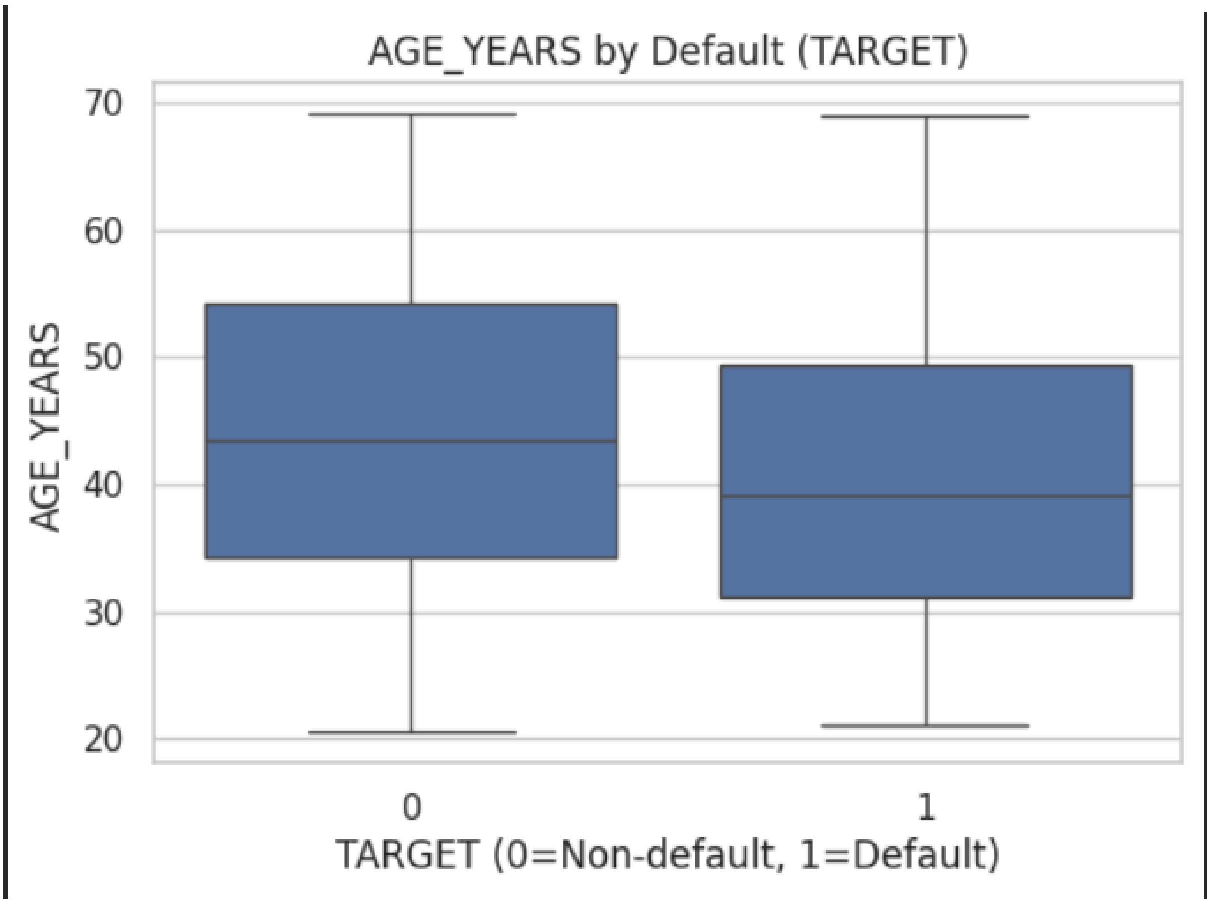
CODE_GENDER

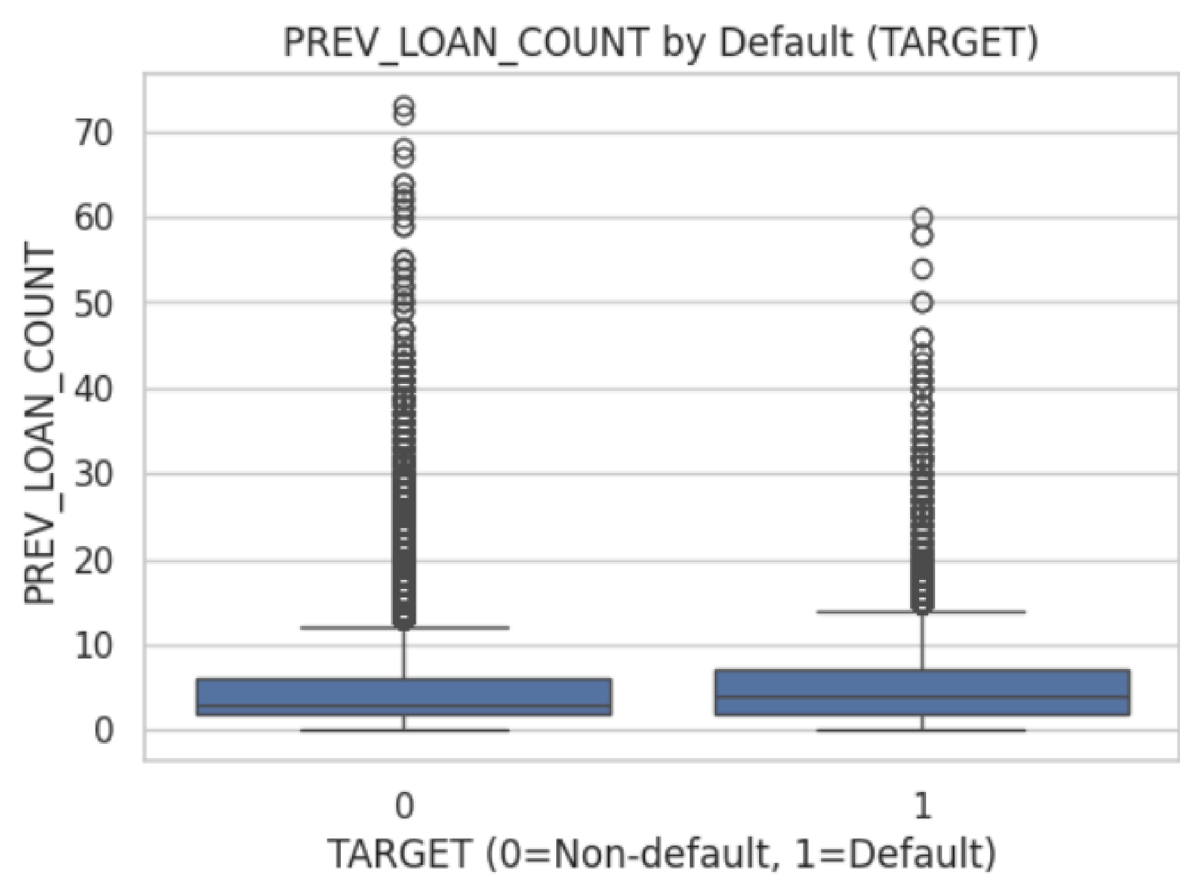

Distribution of NAME_EDUCATION_TYPE
Count
200000
150000
100000
50000
0
dary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree

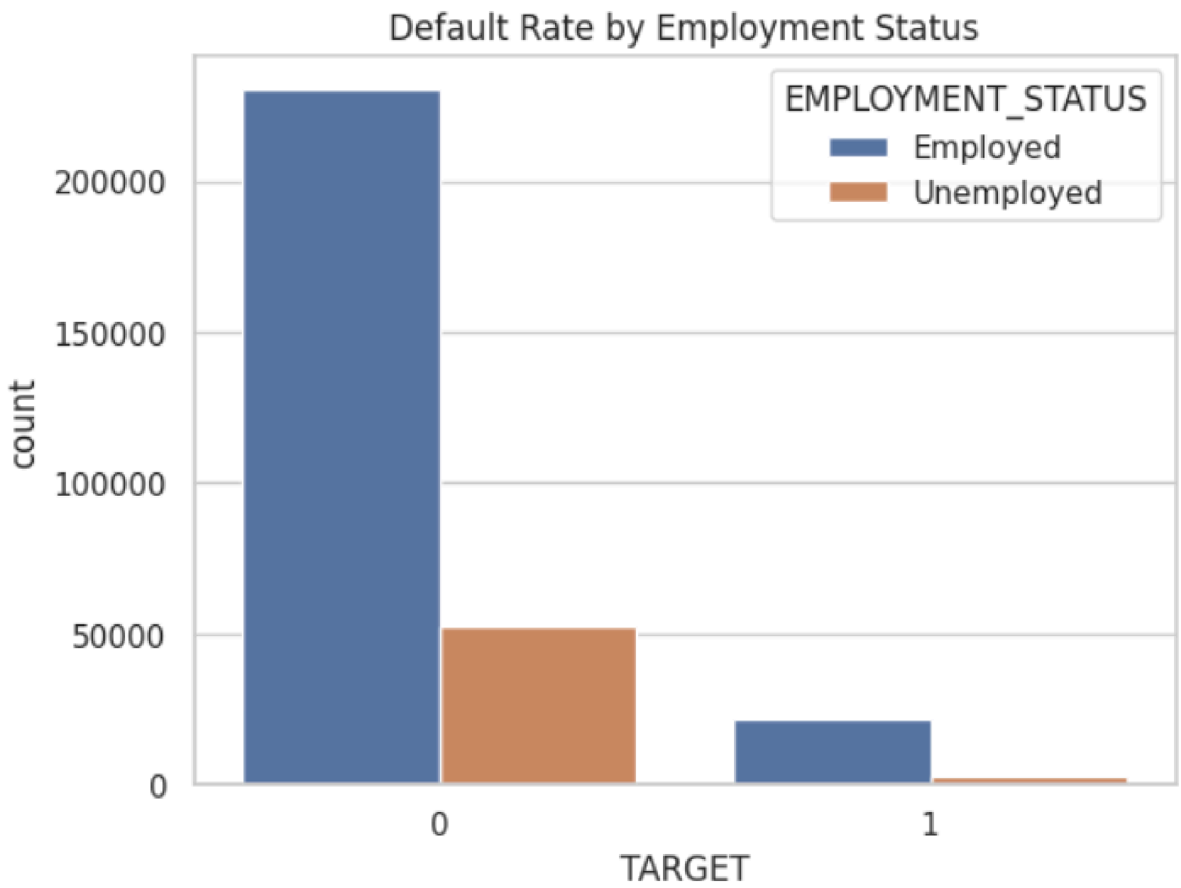

Distribution of NAME_FAMILY_STATUS
Count
200000
175000
150000
125000
100000
75000
50000
25000
0
Married
Single / not married
Civil marriage
Separated
Widow
Unknown
NAME_FAMILY_STATUS

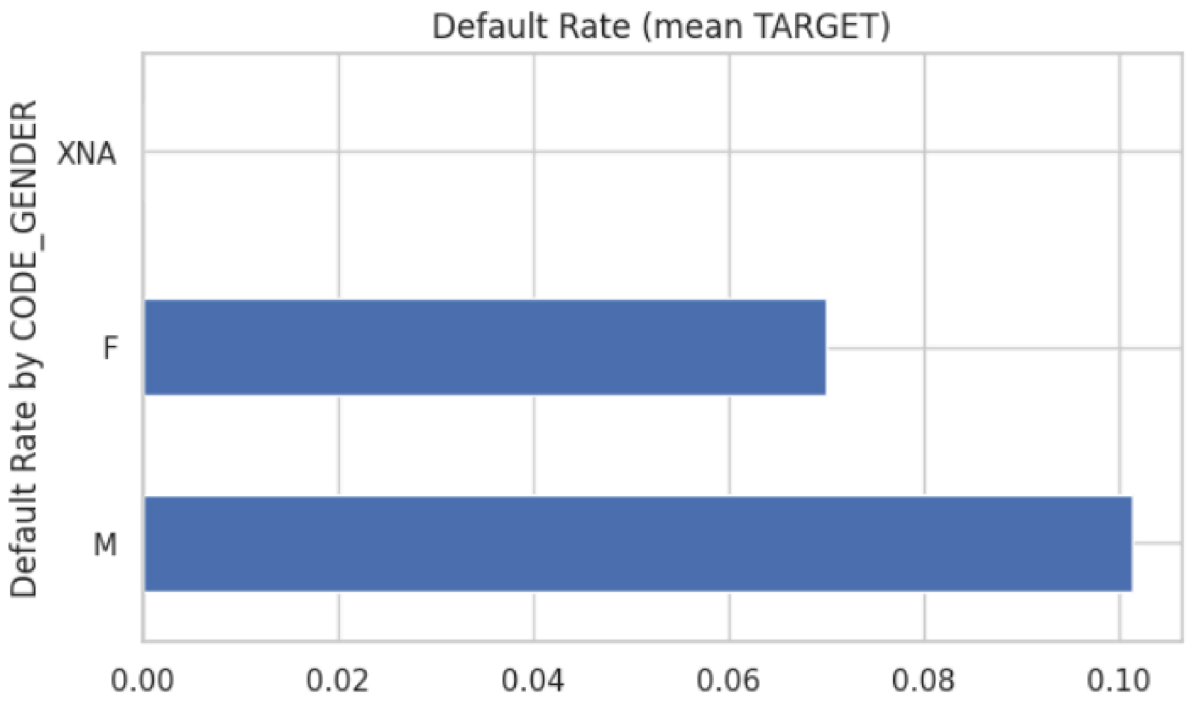

AMT_INCOME_TOTAL by Default (TARGET)
AMT_INCOME_TOTAL
400000
300000
200000
100000
0
1
TARGET (0=Non-default, 1=Default)


AMT_CREDIT by Default (TARGET)
1e6
AMT_CREDIT
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
1
TARGET (0=Non-default, 1=Default)

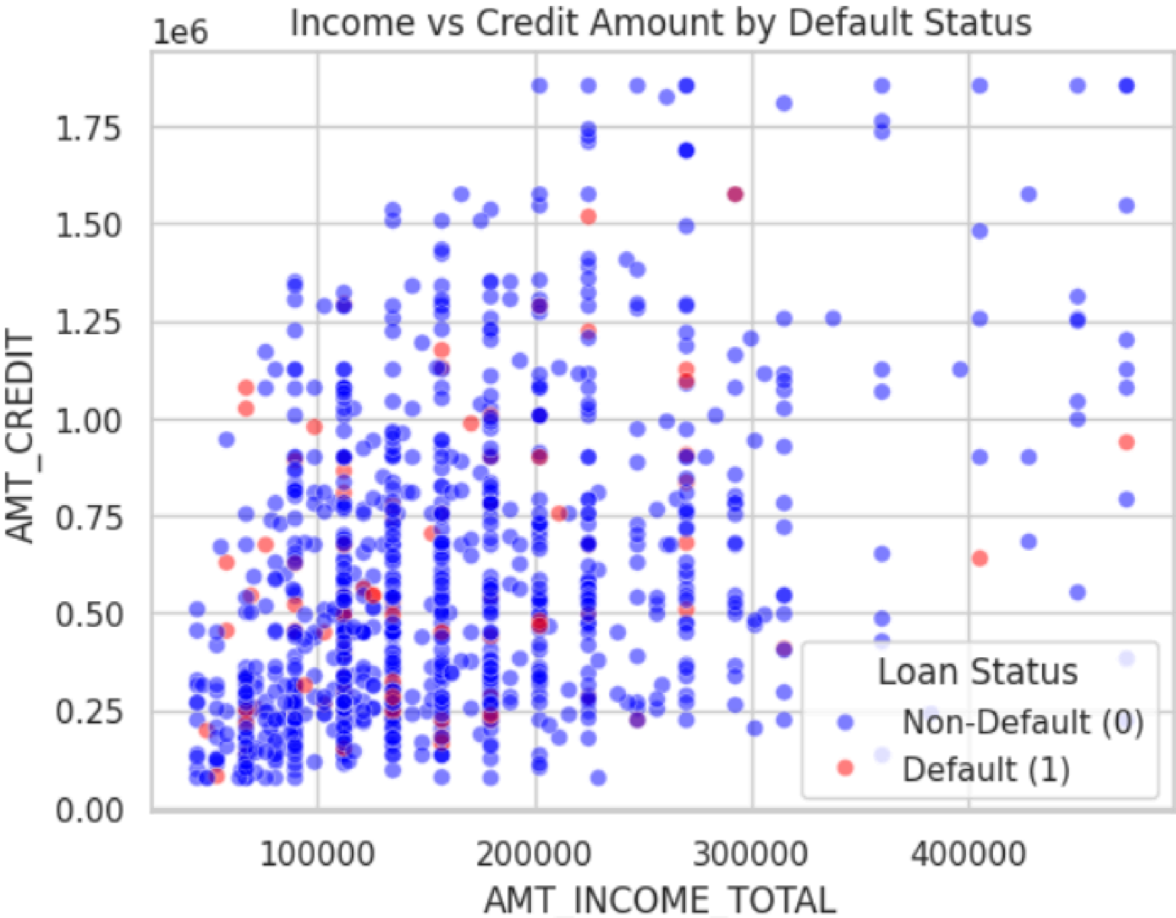

AGE_YEARS by Default (TARGET)
AGE_YEARS
70
60
50
40
30
20
0
1
TARGET (0=Non-default, 1=Default)


PREV_LOAN_COUNT by Default (TARGET)
PREV_LOAN_COUNT
70
60
50
40
30
20
10
0
0
1
TARGET (0=Non-default, 1=Default)

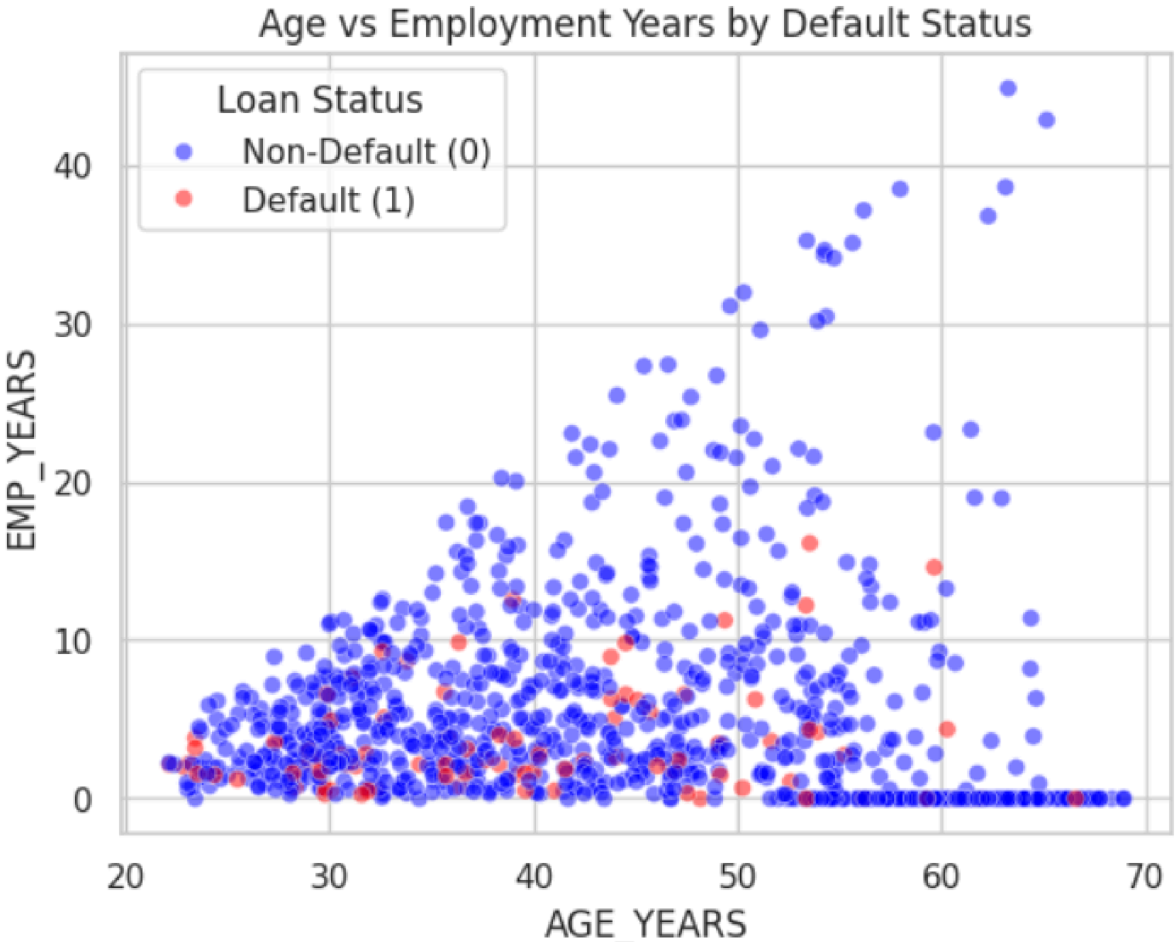

Default Rate by Employment Status
EMPLOYMENT_STATUS
Employed
Unemployed
200000
150000
100000
50000
0
count
0
1
TARGET

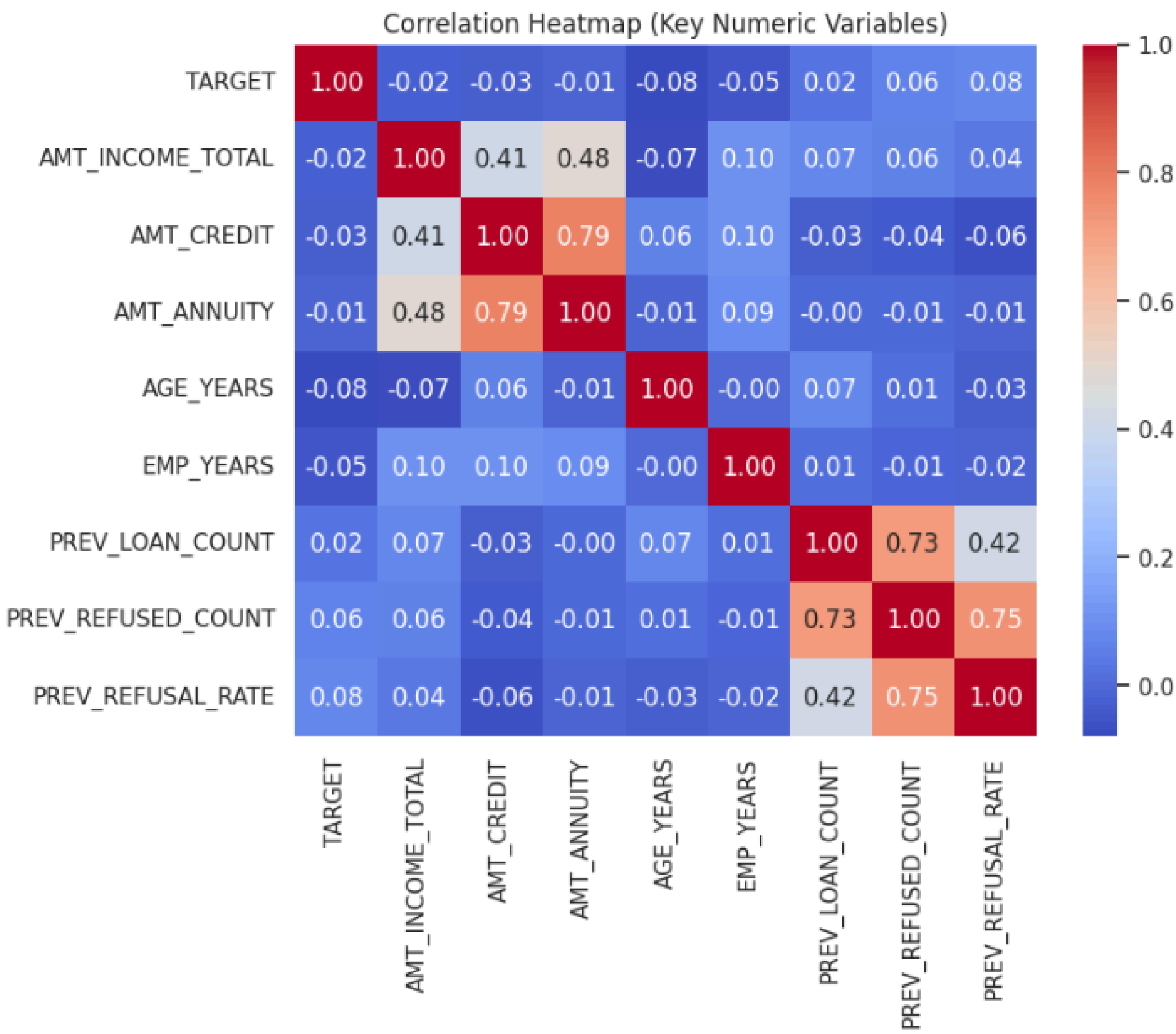

Default Rate (mean TARGET)
Default Rate by CODE_GENDER
XNA
F
M
0.00
0.02
0.04
0.06
0.08
0.10

Default Rate (mea
Default Rate by NAME_EDUCATION_TYPE
Academic degree
Higher education
Incomplete higher
Secondary / secondary special
Lower secondary
0.00
0.02
0.04
0.06
0.08
0.10
Default Rate (mean TARGET)
Default Rate by NAME_FAMILY_STATUS
Unknown
Widow
Married
Separated
Single / not married
Civil marriage
0.00
0.02
0.04
0.06
0.08
0.10



Income vs Credit Amount by Default Status
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
AMT_CREDIT
100000
200000
300000
400000
AMT_INCOME_TOTAL
Loan Status
Non-Default (0)
Default (1)


Age vs Employment Years by Default Status
Loan Status
Non-Default (0)
Default (1)
EMP_YEARS
AGE_YEARS
0
10
20
30
40
20
30
40
50
60
70




```
Correlation of numeric variables with TARGET:
TARGET                  1.000000
PREV_REFUSAL_RATE       0.077894
PREV_REFUSED_COUNT      0.064756
PREV_LOAN_COUNT         0.023513
AMT_ANNUITY            -0.011086
AMT_INCOME_TOTAL       -0.023313
AMT_CREDIT             -0.030086
EMP_YEARS              -0.046052
AGE_YEARS              -0.078239
Name: TARGET, dtype: float64
```