Exploratory Data Analysis (EDA) for Client Payment Difficulties

A Data-Driven Approach to Understanding Payment Difficulties in Clients

Vy Pham

Table of contents

01

Initial Data Understanding.

02

Data Cleaning and Handling Missing Values.

03

Exploratory Data Analysis (EDA)

04

Insights and Recommendations

01

Data Understanding

Data Understanding

- **Summary of the dataset:**
  - Number of rows and columns.
  - Description of the target variable (clients with payment difficulties).
  - Key variables (e.g., demographic, transactional data).
- **Identification of missing data:**
  - Overview of missing value patterns using heatmaps or tables.
  - Mention variables with significant missing values.

- **Key Highlights:**
  - **Imbalance in Target Variable:**
    1. **Non-Defaulters (0):** ~92% (282,686 rows).
    2. **Defaulters (1):** ~8% (24,825 rows).
- **Missing Values:**
  - Application Data: Several columns have significant missing values (e.g., OWN_CAR_AGE - 66.2%, EXT_SOURCE_1 - 56.4%).
  - Previous Application: Missing data in columns like NAME_CONTRACT_STATUS.

Dataset	Rows	Columns	Key Variables	Target Variable
Application Data	307,511	122	Demographic (AGE, CNT_FAM_MEMBERS), Credit (AMT_CREDIT, AMT_ANNUITY), Behavioral (DAYS_EMPLOYED)	TARGET
			Derived (TOTAL_DOCS, TOTAL_CREDIT_BUREAU_ENQ)	Binary (1: Payment difficulties, 0: No difficulties)
Previous Application	167,021	37	Application history (DAYS_DECISION, NAME_CONTRACT_STATUS), Credit Terms (AMT_APPLICATION, AMT_CREDIT)	Not applicable

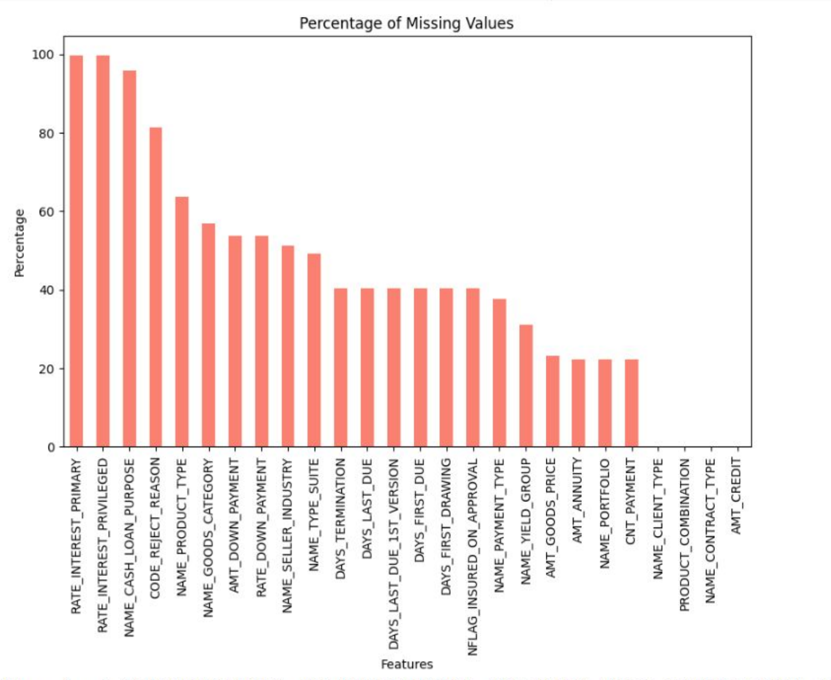
02

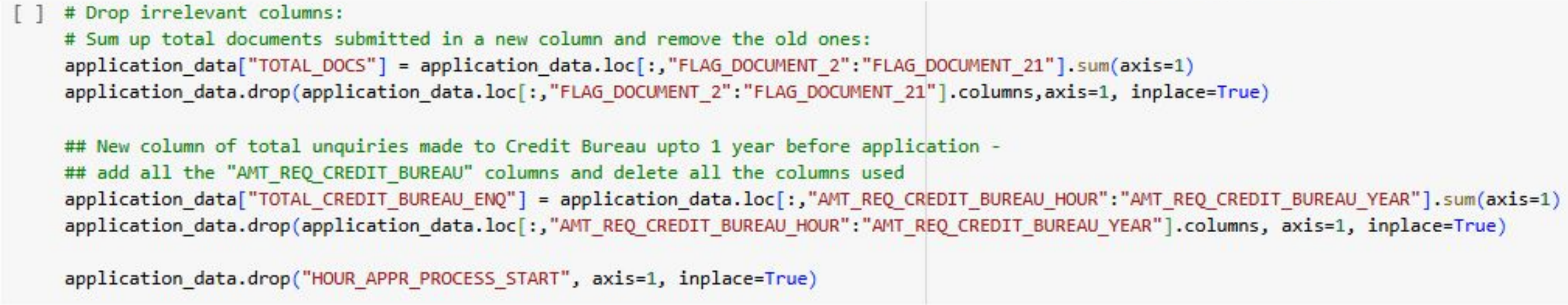
Handling Missing Values

Handling Missing Data

Methodology

1. **Column Removal**:
   - **Criteria**: Columns with >40% missing values dropped to reduce noise.
   - **Dropped Columns**:
     - FLAG_DOCUMENT_2, FLAG_DOCUMENT_21, AMT_REQ_CREDIT_BUREAU_HOUR, HOUR_APPR_PROCESS_START etc
   - **Justification**:
     - Excessive missing values reduce model reliability.
2. **Imputation Techniques**:
   - **Numerical Variables**:
     - Imputed with median values (less sensitive to skewness).
     - Example: AMT_ANNUITY, AMT_CREDIT.
   - **Categorical Variables**:
     - Imputed with mode (most frequent value).
     - Example: OCCUPATION_TYPE, GENDER.





```
# Drop irrelevant columns:
# Sum up total documents submitted in a new column and remove the old ones:
application_data["TOTAL_DOCS"] = application_data.loc[:,"FLAG_DOCUMENT_2":"FLAG_DOCUMENT_21"].sum(axis=1)
application_data.drop(application_data.loc[:,"FLAG_DOCUMENT_2":"FLAG_DOCUMENT_21"].columns,axis=1, inplace=True)

## New column of total unquiries made to Credit Bureau upto 1 year before application -
## add all the "AMT_REQ_CREDIT_BUREAU" columns and delete all the columns used
application_data["TOTAL_CREDIT_BUREAU_ENQ"] = application_data.loc[:,"AMT_REQ_CREDIT_BUREAU_HOUR":"AMT_REQ_CREDIT_BUREAU_YEAR"].sum(axis=1)
application_data.drop(application_data.loc[:,"AMT_REQ_CREDIT_BUREAU_HOUR":"AMT_REQ_CREDIT_BUREAU_YEAR"].columns, axis=1, inplace=True)

application_data.drop("HOUR_APPR_PROCESS_START", axis=1, inplace=True)
```

Handling Missing Data

Handling Special Cases

- **XNA/XAP Values**:
  - Replaced with NaN for accurate identification of missingness.
- **Outlier in DAYS_EMPLOYED**:
  - Value 365,243 replaced with NaN (represents erroneous input).

Justification for Approach

- Dropping high-missing-value columns ensures better model reliability and interpretability.
- Imputation balances retaining data and managing missingness effectively without introducing significant bias.

Visualization of the Process

1. **Columns Dropped**:
   - Table of dropped columns with percentages of missing values.
2. **Imputation Example**:
   - Before/After histogram of a key variable (AMT_CREDIT) showing the impact of imputation.
3. **Remaining Missing Values**:
   - Summary statistics: Percentage of remaining missing data.

03

Exploratory Data Analysis (EDA)

Problem Statement

Problem Overview:

- The objective is to understand the factors influencing clients' payment difficulties.
- By analyzing historical data, we aim to identify key patterns, correlations, and insights that can help predict future payment behaviors.

Business Objective:

- **Target Variable**: Clients with payment difficulties versus all other cases (binary classification).
- **Goal**: To identify patterns, trends, and factors contributing to payment difficulties.

Approach Overview

Exploratory Data Analysis (EDA):

- Uncover hidden insights and patterns.
- Visualize relationships between variables and the target.
- Identify missing data, outliers, and any data imbalances.
- Segment data to find key correlations and business insights.

Steps:

1. Data Cleaning: Handle missing values and outliers.
2. Univariate and Bivariate Analysis: Analyze individual and paired variables.
3. Data Imbalance: Check if there's an imbalance in the target variable and analyze it.
4. Correlation Analysis: Find the top 10 correlations with the target variable.

Target Variable (Data imbalance) Analysis

Key Steps:

- Examined the distribution of the **TARGET** variable:
  - 0: Normal payments.
  - 1: Payment difficulties.

Findings:

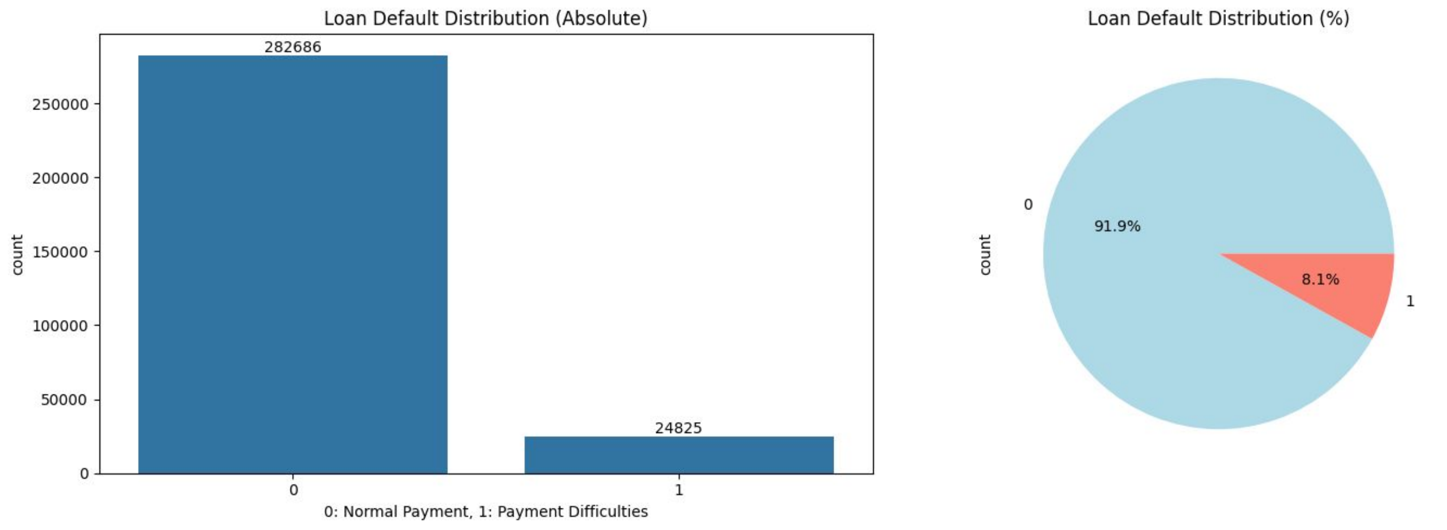
- Class Imbalance:
  - **Non-default cases**: 91.93%+
  - **Default cases**: ~8.1%
  - **Imbalance Ratio**: 11.39:1

Imbalance:

The dataset exhibits a significant imbalance, with non-default cases being over 11 times more frequent than default cases.
Imbalance Ratio: 11.39:1

Implications:

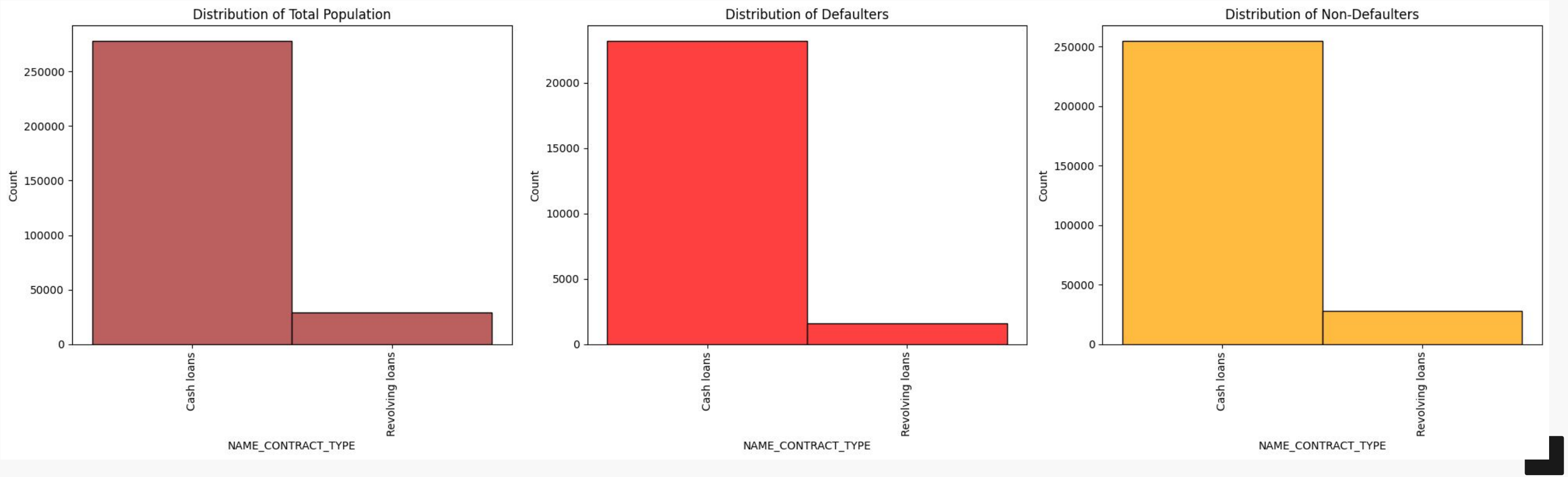
The skewed class distribution highlights the need for techniques such as oversampling, undersampling, or balanced metrics during predictive modeling to address bias.



Univariate Analysis - Categorical Feature Analysis

NAME_CONTRACT_TYPE

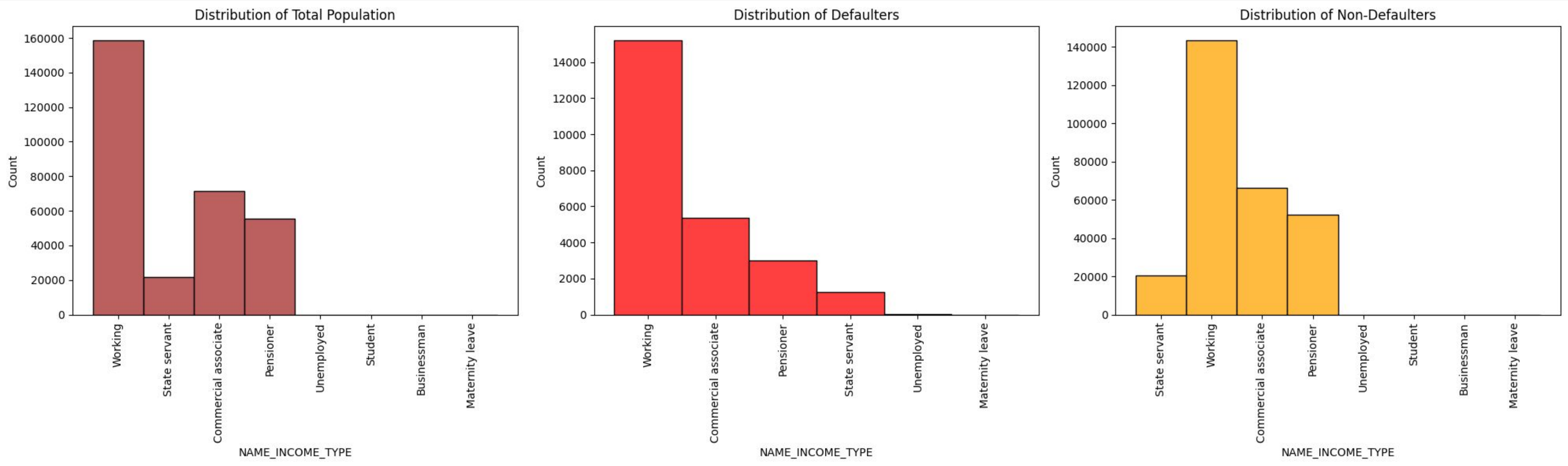
- **Distribution**:
  - The majority of the population opted for **Cash loans**, while a small percentage chose **Revolving loans**.
  - Similar trends are observed among both defaulters and non-defaulters.
- **Insights:**
  - The default rate is slightly higher for **Cash loans** (8.35%) compared to **Revolving loans** (5.47%).
  - Cash loans might indicate a higher risk category, requiring further monitoring.



Univariate Analysis - Categorical Feature Analysis

NAME_INCOME_TYPE

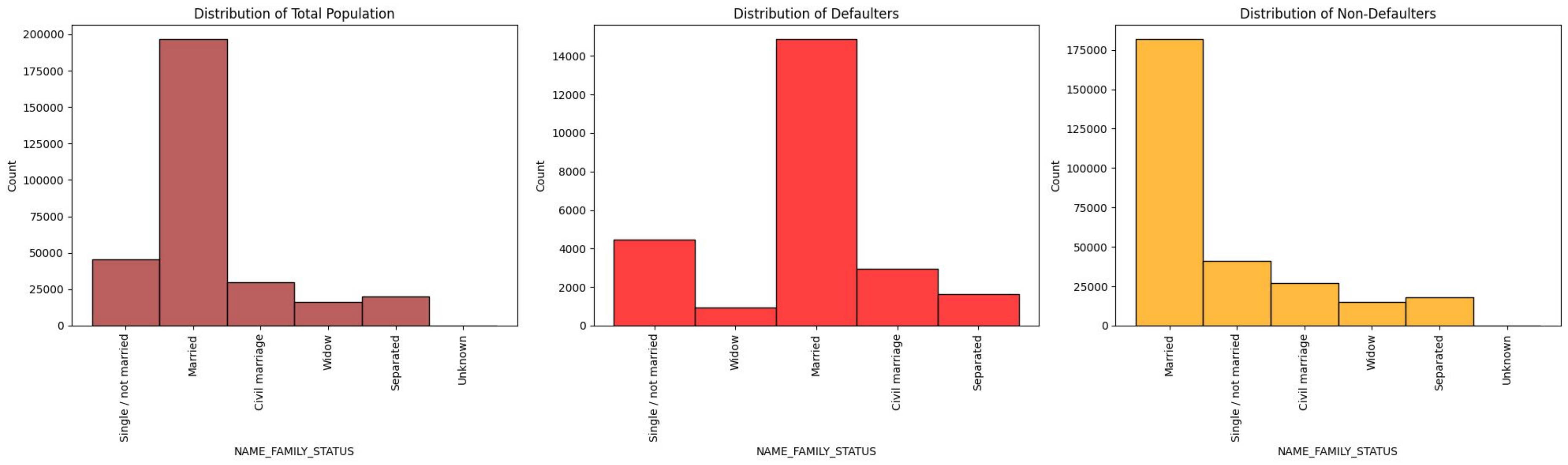
- **Distribution**:
  - Most clients fall under the **Working** or **Commercial Associate** income types.
  - Defaulters are predominantly from the **Working** category.
- **Insights:**
  - The highest default rate is observed among the **Unemployed** (36.36%) and those on **Maternity leave** (33.33%).
  - Clients in unstable or limited income categories are at significantly higher risk of default.



Univariate Analysis - Categorical Feature Analysis

NAME_FAMILY_STATUS

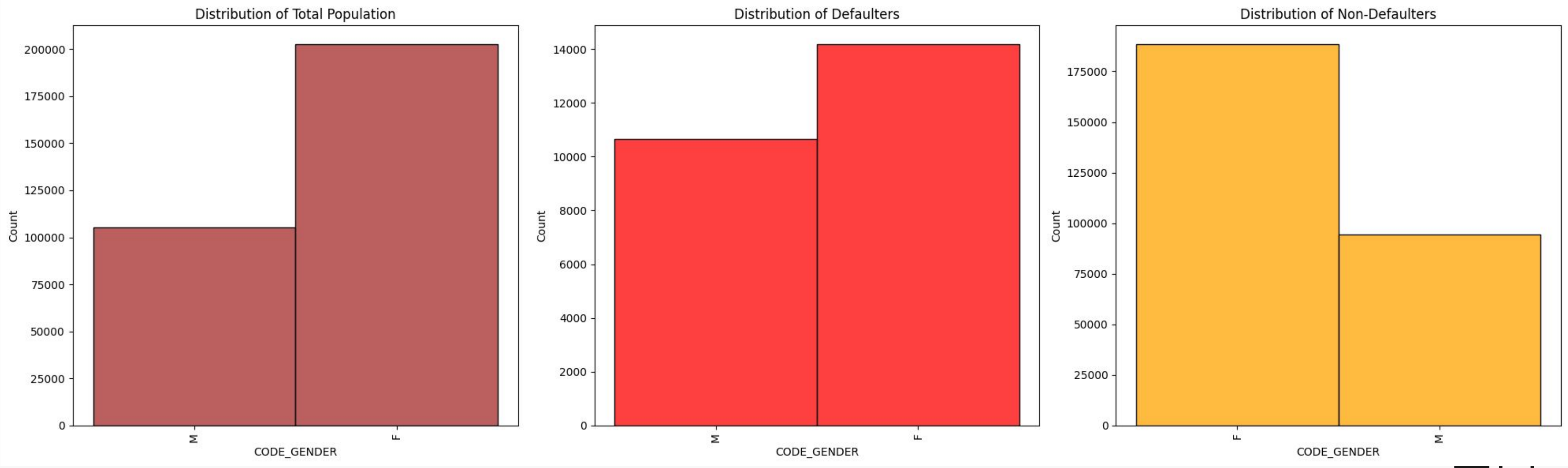
- **Distribution**:
  - Married individuals form the largest group in both defaulters and non-defaulters.
  - Single individuals also constitute a significant proportion.
- **Insights:**
  - The default rate is highest for **Civil marriage** (9.94%), followed by **Separated** (9.18%).
  - Family dynamics and financial responsibilities might impact the likelihood of default.



Univariate Analysis - Categorical Feature Analysis

CODE_GENDER

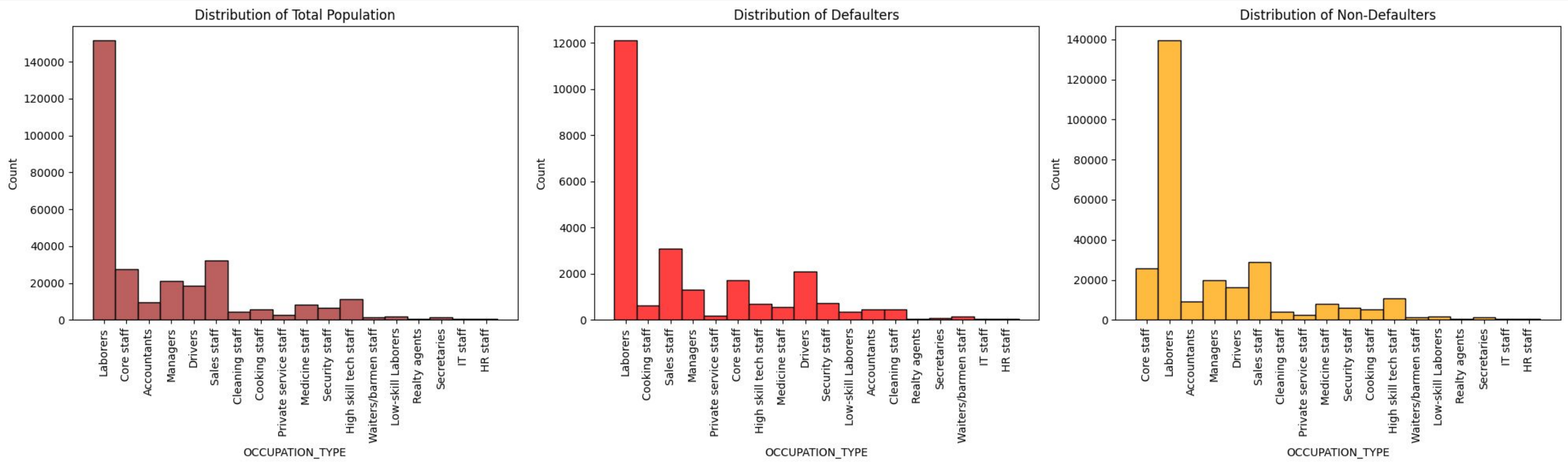
- **Distribution**:
  - Females dominate the population, but defaulters are more evenly split between males and females
- **Insights:**
  - Males exhibit a higher default rate (10.14%) compared to females (6.99%).
  - This could indicate gender-based financial behavior differences.



Univariate Analysis - Categorical Feature Analysis

OCCUPATION_TYPE

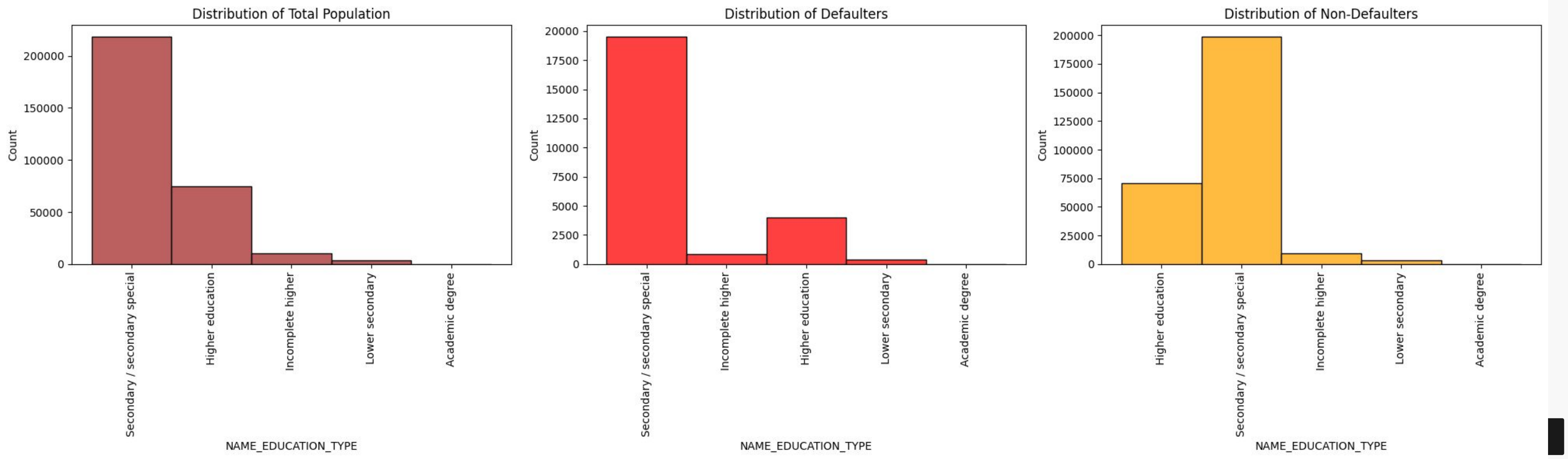
- **Distribution**:
  - The largest group consists of clients in **Laborers**, followed by **Sales staff** and **Core staff**.
  - Defaulters are heavily concentrated in labor-intensive roles.
- **Insights:**
  - Default rates are highest among **Low-skill laborers** (37.55%), followed by **Drivers** and **Cleaning staff**.
  - Occupation types tied to lower-income and unstable jobs are significantly more prone to default.



Univariate Analysis - Categorical Feature Analysis

NAME_EDUCATION_TYPE

- **Distribution**:
  - The majority of the population has a **Secondary/Secondary Special** education level.
  - Defaulters are primarily from this education group as well.
- **Insights:**
  - Default rates decrease as education levels increase.
  - **Secondary/Secondary Special** shows the highest default rate (8.93%), while those with **Academic Degrees** have the lowest (1.83%).



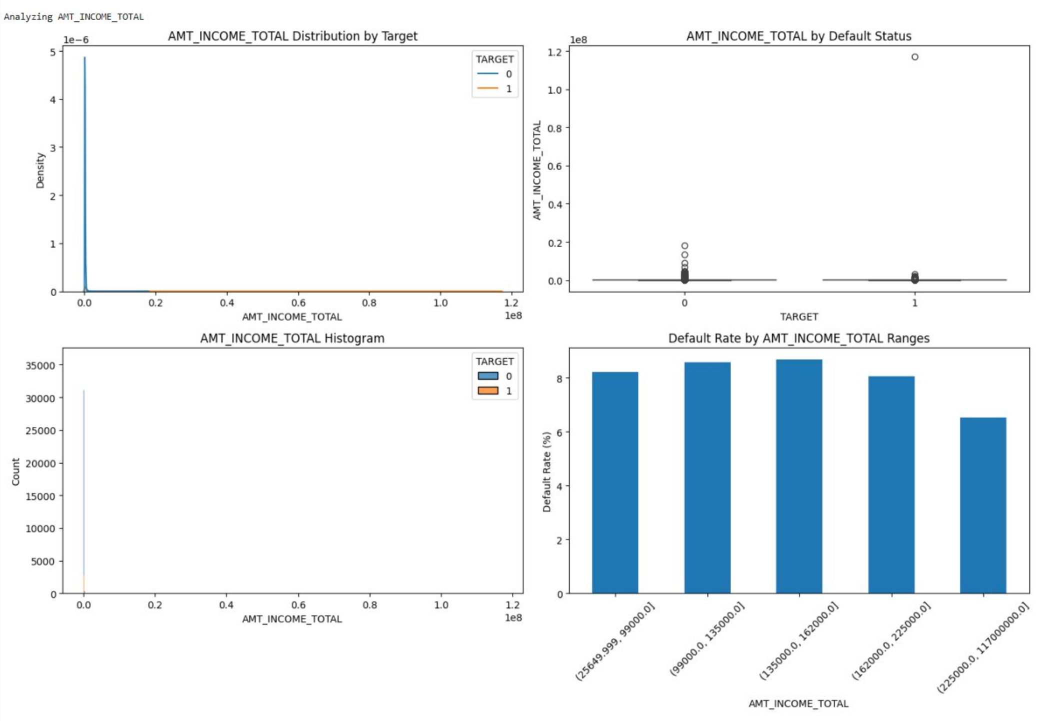
Univariate Analysis - Categorical Feature Analysis

General Observations

1. **Income stability** and **education level** play critical roles in differentiating defaulters from non-defaulters.
2. Certain **occupations** and **family statuses** indicate higher financial vulnerability.
3. **Cash loans** are associated with higher risk, while gender-based trends in defaults need further investigation.

These insights can guide targeted risk mitigation strategies, such as stricter evaluation criteria for high-risk groups or providing financial support programs for vulnerable demographics.

Univariate Analysis - Numerical Feature Analysis



```
Analyzing AMT_INCOME_TOTAL
```

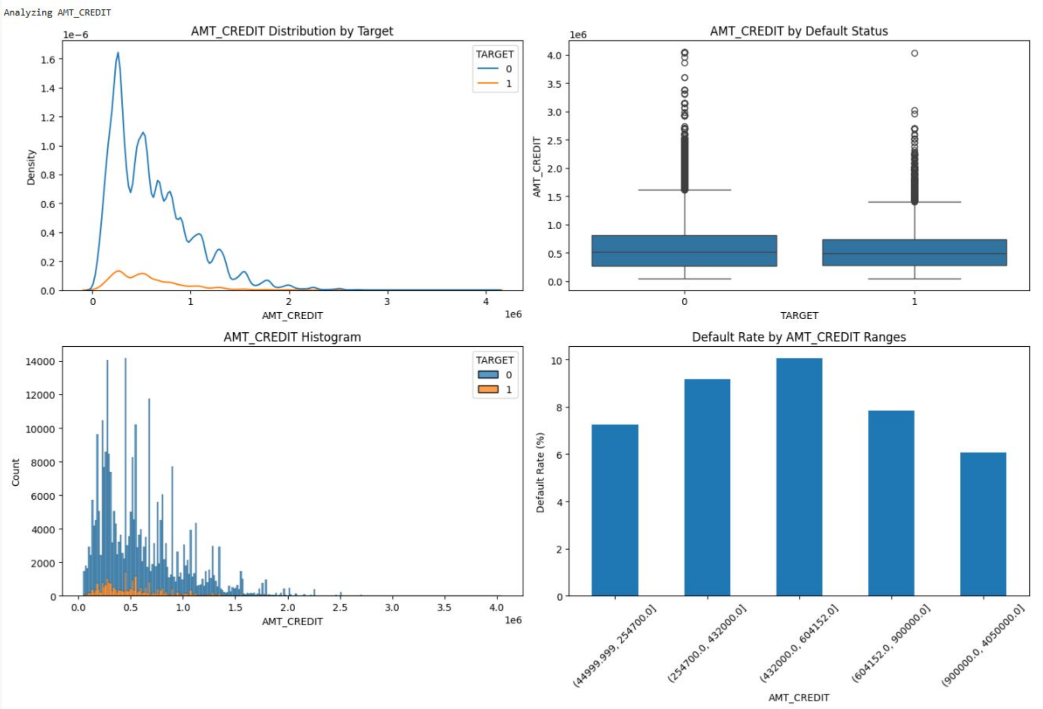
Key Insights: AMT_INCOME_TOTAL

- **Median Income:**
  - **Defaulters:** $135,000
  - **Non-Defaulters:** $148,500
  - Defaulters generally have slightly lower median incomes.
- **Outliers:**
  - **Count:** 14,035 rows (4.56% of data).
  - **Effect:** Skew the data towards higher income values.
- **Default Rates by Income Range:**
  - **Highest Default Rate:** $99,000–$135,000 (8.58%).
  - **Lowest Default Rate:** $225,000+ (6.51%).
- **Income Distribution Percentiles:**
  - Majority of clients earn below $270,000.
  - Higher income groups (> $225,000) show reduced default tendencies.

Recommendations:

1. **Risk Management:**
   - Focus on middle-income groups ($99,000–$135,000), which show the highest default rates.
2. **Outlier Handling:**
   - Consider strategies to address outliers, such as capping or transformation, to improve model performance.
3. **Credit Policy:**
   - Tailor credit policies to incentivize higher-income borrowers (> $225,000) who exhibit lower default risks.

Univariate Analysis - Numerical Feature Analysis



```
Analyzing AMT_CREDIT
```

Key Insights: AMT_CREDIT

- **Summary Statistics:**
    - **Mean Credit Amount:**
        - **Non-Defaulters:** $602,648
        - **Defaulters:** $557,778
    - **Median Credit Amount:**
        - **Non-Defaulters:** $517,788
        - **Defaulters:** $497,520
    - Non-defaulters generally have higher mean and median credit amounts.
- **Outliers:**
    - **Count**: 6,562 rows (2.13% of data).
    - **Effect**: Outliers are concentrated in higher credit ranges (> $1,350,000).
- **Default Rates by Credit Range:**
    - **Highest Default Rate:** $432,000–$604,152 (10.05%).
    - **Lowest Default Rate:** $900,000+ (6.08%).

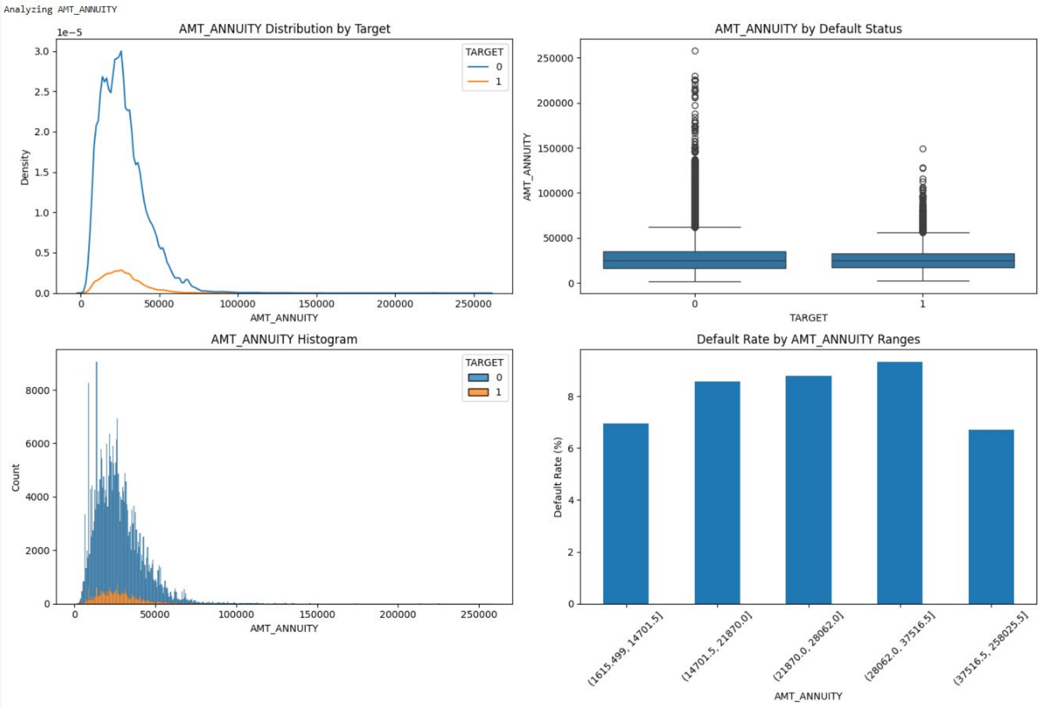
Percentile Analysis:

- **25th Percentile:** $270,000
- **75th Percentile:** $808,650
- **Outlier Threshold:** Values above $1,350,000 (95th percentile).

Recommendations:

1. **Risk Mitigation:**
    - Prioritize risk assessment for clients in the $432,000–$604,152 credit range due to their high default rate.
2. **Outlier Treatment:**
    - Address outliers (e.g., capping or normalization) to improve model robustness and fairness.
3. **Policy Adjustment:**
    - Offer tailored credit policies for high-credit clients (> $900,000), who exhibit the lowest default tendencies.

Univariate Analysis - Numerical Feature Analysis



```
Analyzing AMT_ANNUITY
```

AMT_ANNUITY Distribution by Target

AMT_ANNUITY by Default Status

AMT_ANNUITY Histogram

Default Rate by AMT_ANNUITY Ranges

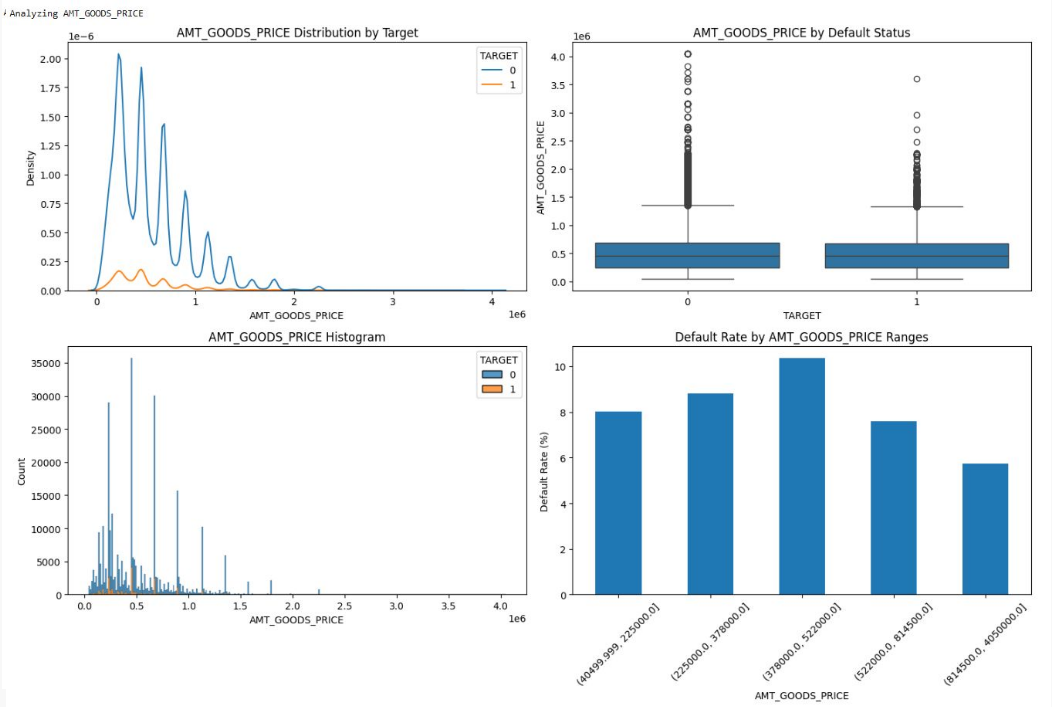
Key Insights: AMT_ANNUITY

- **Summary Statistics:**
  - **Median Annuity Amount:**
    - **Non-Defaulters:** $24,876
    - **Defaulters:** $25,263
  - Slightly higher median annuities observed among defaulters.
- **Outliers:**
  - **Count:** 7,504 rows (2.44% of data).
  - **Effect:** Outliers primarily involve high annuity amounts (> $70,006, 99th percentile).
- **Default Rates by Annuity Range:**
  - **Highest Default Rate:** $28,062–$37,516 (9.34%).
  - **Lowest Default Rate:** $37,516+ (6.70% and below).

Recommendations:

1. **High-Risk Monitoring:**
   - Focus on clients with annuities in the **$28,062–$37,516** range due to elevated default rates.
2. **Outlier Management:**
   - Address high-value outliers through scaling or capping techniques to enhance analytical accuracy.
3. **Policy Refinement:**
   - Introduce flexible repayment options for clients within the high-risk annuity segment.

Univariate Analysis - Numerical Feature Analysis



Analyzing AMT_GOODS_PRICE

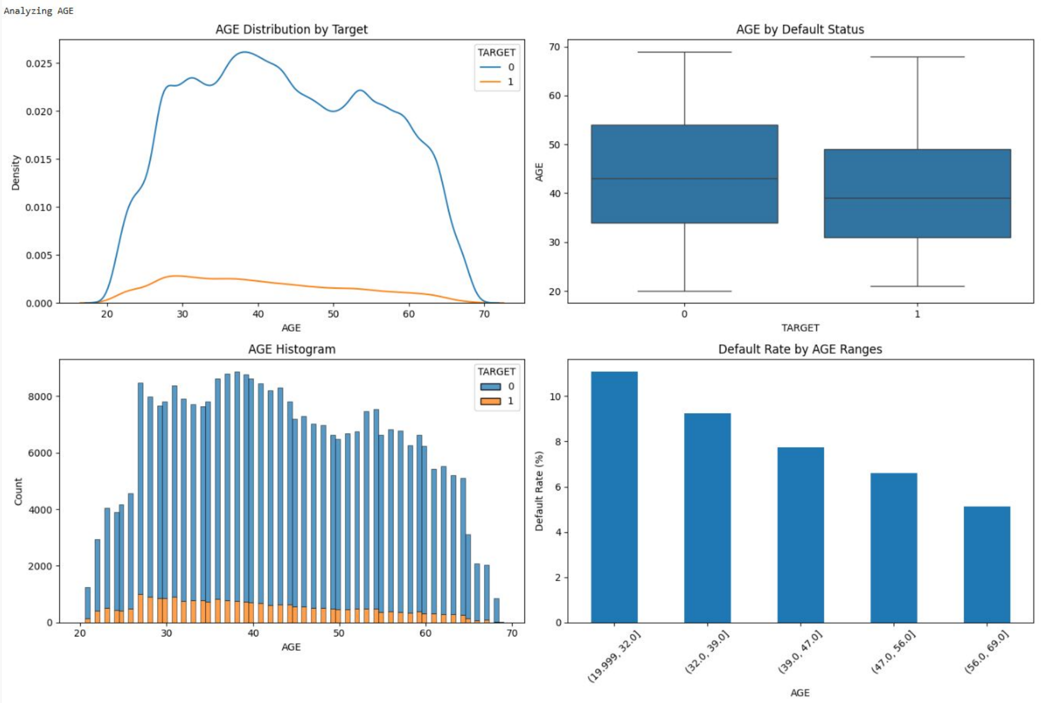
Key Insights: AMT_GOODS_PRICE

- **Summary Statistics:**
  - **Median Goods Price:** $450,000 (same for defaulters and non-defaulters).
  - High concentration of values around the median price.
- **Outliers:**
  - **Count:** 14,728 rows (4.79% of data).
  - **Effect:** Outliers mainly occur at high price ranges above $1,093,500 (95th percentile).
- **Default Rates by Price Range:**
  - **Highest Default Rate:** $378,000–$522,000 (10.35%).
  - **Lowest Default Rate:** $814,500+ (5.73% and below).

Recommendations:

1. **High-Risk Monitoring:**
   - Focus on goods priced in the **$378,000–$522,000** range, as these exhibit the highest default rates.
2. **Outlier Management:**
   - Investigate high-value goods price outliers (> $1,093,500) for potential data irregularities or distinct client behaviors.
3. **Policy Refinement:**
   - Tailor risk-based policies for mid-range goods prices while maintaining flexible terms for lower-risk high-value goods.

Univariate Analysis - Numerical Feature Analysis



```
Analyzing AGE
```

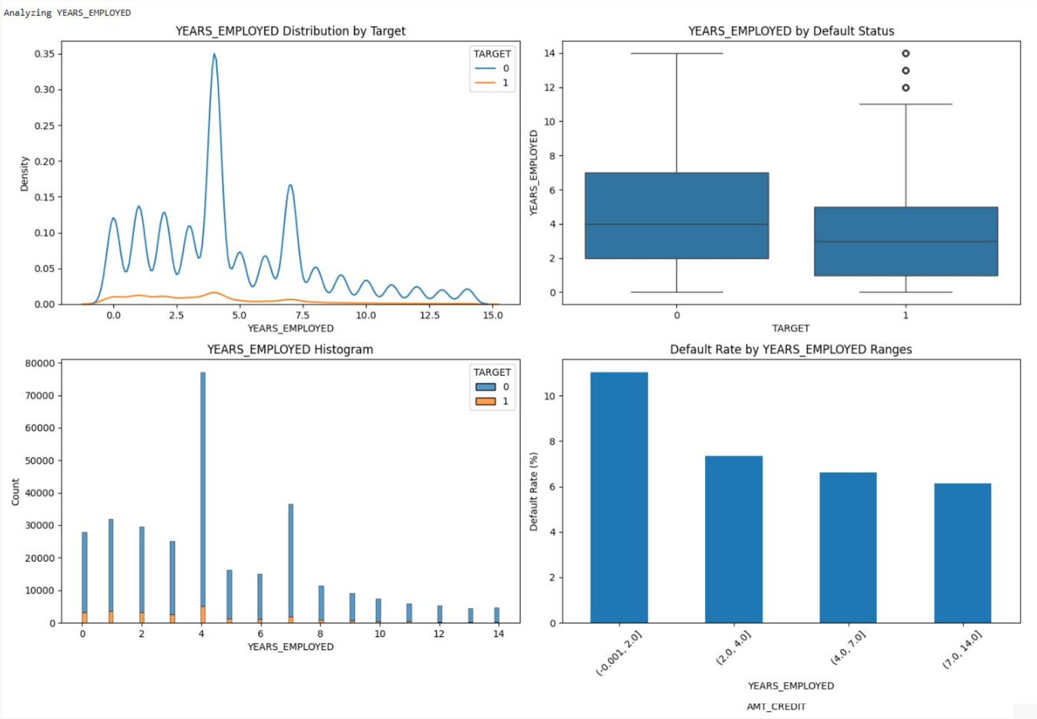
Key Insights: AGE

- **Summary Statistics:**
  - **Mean Age:**
    - **Defaulters:** 40.77 years
    - **Non-Defaulters:** 43.71 years
  - Defaulters are notably younger on average.
- **Default Rates by Age Group:**
  - **Highest Default Rate:** 20–32 years (11.09%).
  - **Lowest Default Rate:** 50+ years (5.94%).
- **Outliers:**
  - No significant outliers detected for AGE.

Recommendations:

1. **High-Risk Monitoring:**
   - Prioritize younger clients (20–32 years) for enhanced risk management strategies.
2. **Policy Design:**
   - Develop tailored financial literacy or risk mitigation programs targeting younger borrowers.
3. **Predictive Modeling:**
   - Use AGE as a key feature for predictive models, as it strongly correlates with default behavior.

Univariate Analysis - Numerical Feature Analysis



```
Analyzing YEARS_EMPLOYED
```

```
AMT_CREDIT
```

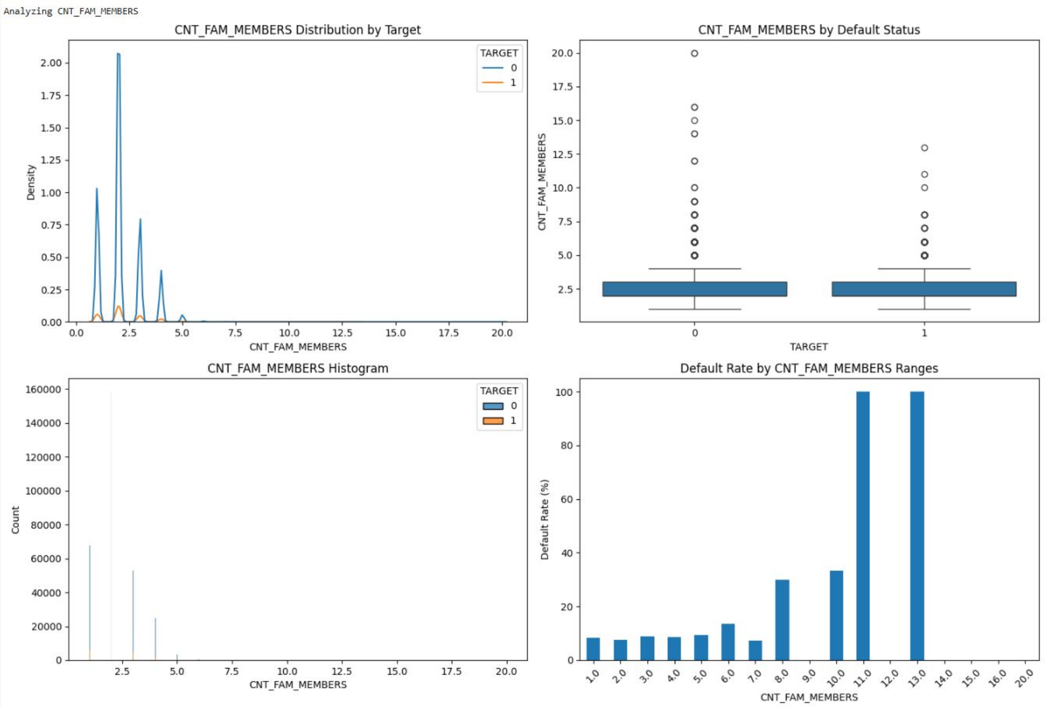
Key Insights: YEARS_EMPLOYED

- **Summary Statistics:**
  - **Mean Employment Duration:**
    - **Defaulters:** 3.83 years
    - **Non-Defaulters:** 4.61 years
  - Shorter employment durations are associated with higher default risk.
- **Default Rates by Employment Duration:**
  - **Highest Default Rate:** Less than 2 years (11.04%).
  - **Lowest Default Rate:** Over 10 years (5.91%).
- **Outliers:**
  - No significant outliers detected for YEARS_EMPLOYED.

Recommendations:

1. **Risk Mitigation:**
   - Focus on clients with employment durations under 2 years, as they exhibit the highest default risk.
2. **Predictive Modeling:**
   - Use YEARS_EMPLOYED as a critical feature for assessing financial stability.
3. **Policy Design:**
   - Develop targeted credit policies or support programs for clients with limited employment histories.

Univariate Analysis - Numerical Feature Analysis



```
Analyzing CNT_FAM_MEMBERS
```

CNT_FAM_MEMBERS Distribution by Target

CNT_FAM_MEMBERS by Default Status

CNT_FAM_MEMBERS Histogram

Default Rate by CNT_FAM_MEMBERS Ranges

Key Insights: CNT_FAM_MEMBERS

- **Summary Statistics:**
  - **Mean Family Members:**
    - **Defaulters:** 2.18
    - **Non-Defaulters:** 2.15
  - Defaulters have slightly larger family sizes on average.
- **Default Rates by Family Size:**
  - **Highest Default Rate:** Families with 11+ members (100%), though these cases are extremely rare.
  - **Moderate Risk:** Families with 7–10 members show elevated default risks.
- **Outliers:**
  - **Count:** 4,007 rows (1.3% of data).
  - **Effect:** Outliers primarily occur in families with 7+ members.

Recommendations:

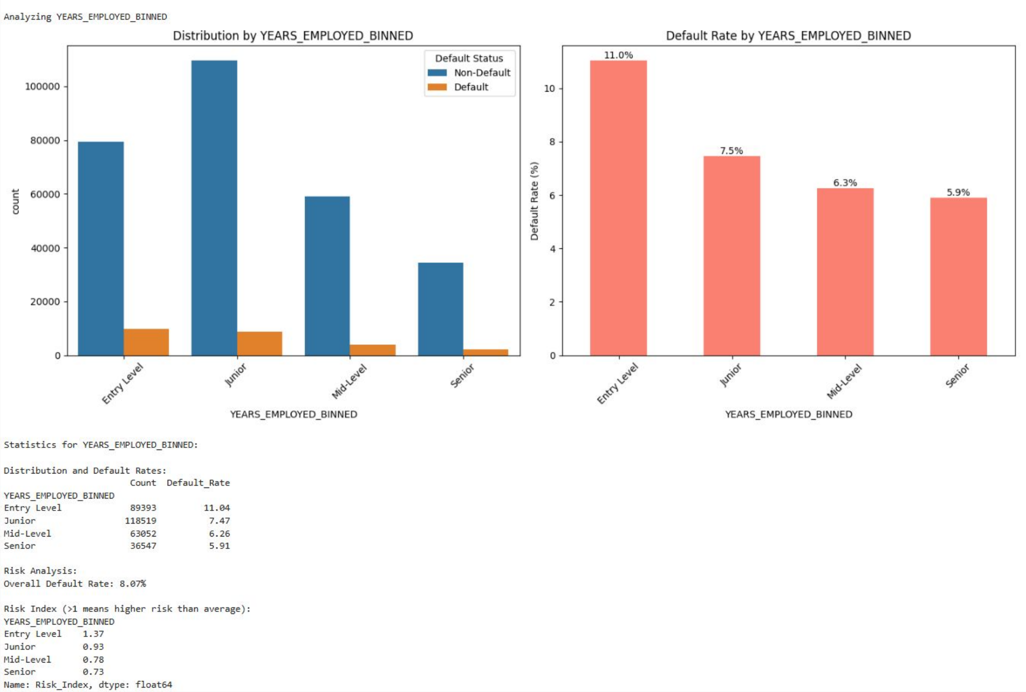
1. **High-Risk Monitoring:**
   - Closely monitor larger families (7+ members) due to their increased default risks.
2. **Policy Considerations:**
   - Incorporate family size into risk assessment models as a factor influencing repayment ability.
3. **Further Analysis:**
   - Investigate if larger families are correlated with other risk factors like lower income or higher credit utilization.

Univariate Analysis - Numerical Feature Analysis

General Observations

1. **AMT_INCOME_TOTAL**:
   - **Observation**: Defaulters tend to have slightly lower incomes. Middle-income groups ($99,000–$135,000) have the highest default rates (8.58%).
2. **AMT_CREDIT**:
   - **Observation**: Default rates peak for credit ranges $432,000–$604,152 (10.05%). High credit values ($900,000+) have lower default risks (6.08%).
3. **AMT_ANNUITY**:
   - **Observation**: Defaulters show elevated risks in the annuity range $28,062–$37,516 (9.34%). High annuity values account for most outliers.
4. **AMT_GOODS_PRICE**:
   - **Observation**: Default rates are highest for goods priced $378,000–$522,000 (10.35%). Higher price ranges ($814,500+) show reduced risks.
5. **AGE**:
   - **Observation**: Younger clients (20–32 years) have the highest default rates (11.09%). Default risk decreases with age.
6. **YEARS_EMPLOYED**:
   - **Observation**: Clients with employment less than 2 years have the highest default rate (11.04%). Longer employment durations reduce risk.
7. **CNT_FAM_MEMBERS**:
   - **Observation**: Larger families (7+ members) exhibit elevated risks. Families with 11+ members have a 100% default rate, though very rare.

Univariate Analysis - Binned Features Analysis



```
Analyzing YEARS_EMPLOYED_BINNED
```

```
Statistics for YEARS_EMPLOYED_BINNED:

Distribution and Default Rates:
                        Count  Default_Rate
YEARS_EMPLOYED_BINNED
Entry Level             89393         11.04
Junior                 118519          7.47
Mid-Level               63052          6.26
Senior                  36547          5.91

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
YEARS_EMPLOYED_BINNED
Entry Level    1.37
Junior         0.93
Mid-Level      0.78
Senior         0.73
Name: Risk_Index, dtype: float64
```

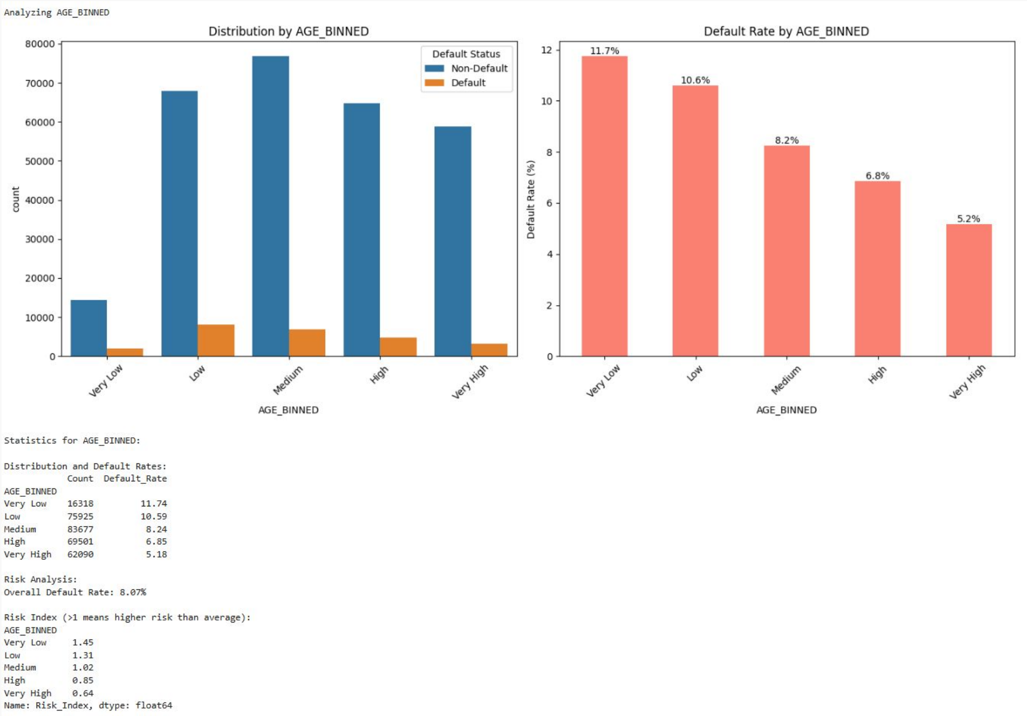
Key Insights: YEARS_EMPLOYED_BINNED

- **Default Rates by Employment Level:**
  - **Highest Risk:**
    - **Entry Level (0–2 years):** 11.0%.
  - **Lowest Risk:**
    - **Senior (8+ years):** 5.9%.
  - Default rates decrease as employment duration increases.
- **Risk Index:**
  - **Entry Level:** 1.37 (above average risk).
  - **Senior:** 0.73 (below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - Focus on **Entry Level** employees for targeted risk management strategies.
2. **Policy Design:**
   - Offer customized repayment terms or credit education for newer employees to mitigate risks.
3. **Predictive Modeling:**
   - Use YEARS_EMPLOYED_BINNED as a key feature to enhance predictive accuracy in default modeling.

Univariate Analysis - Binned Features Analysis



```
Analyzing AGE_BINNED
```

```
Statistics for AGE_BINNED:

Distribution and Default Rates:
            Count  Default_Rate
AGE_BINNED
Very Low    16318         11.74
Low         75925         10.59
Medium      83677          8.24
High        69501          6.85
Very High   62090          5.18

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
AGE_BINNED
Very Low     1.45
Low          1.31
Medium       1.02
High         0.85
Very High    0.64
Name: Risk_Index, dtype: float64
```

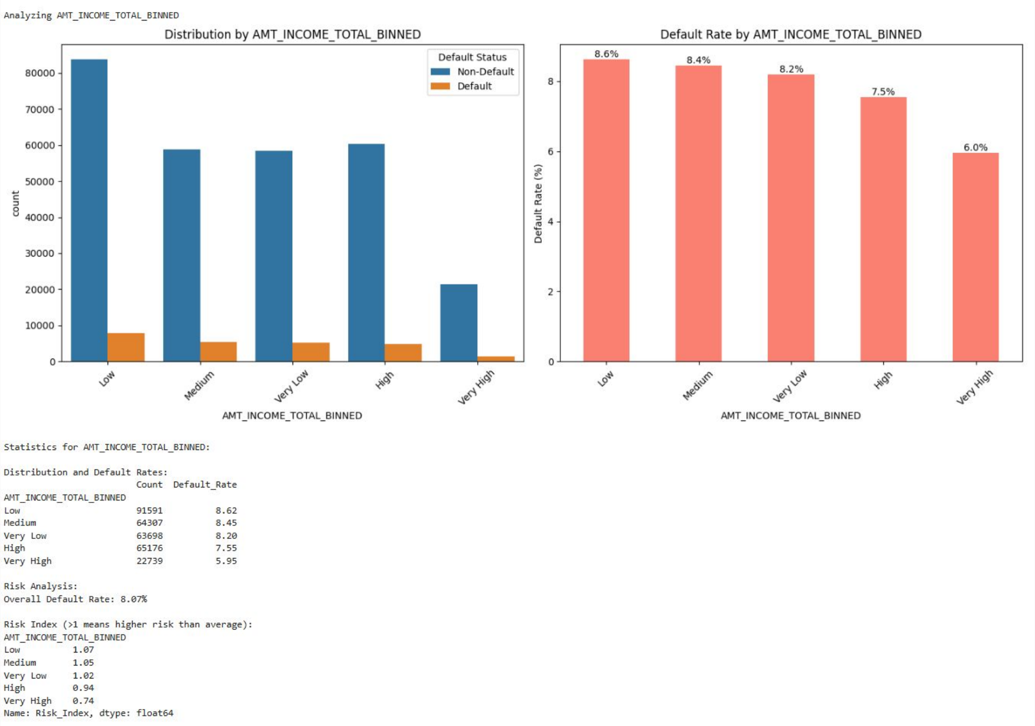
Key Insights: AGE_BINNED

- **Default Rates by Age Group:**
  - **Highest Risk:**
    - **Very Low (20–30 years):** 11.7%.
  - **Lowest Risk:**
    - **Very High (50+ years):** 5.2%.
  - Default rates steadily decrease with age.
- **Risk Index:**
  - **Very Low:** 1.45 (significantly above average risk).
  - **Very High:** 0.64 (significantly below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - Focus on younger clients in the **Very Low** and **Low** age groups for targeted risk management strategies.
2. **Policy Design:**
   - Offer financial literacy programs or stricter loan terms for younger borrowers to mitigate risk.
3. **Credit Optimization:**
   - Prioritize loan approvals for older clients (**High** and **Very High** groups) who demonstrate lower default risks

Univariate Analysis - Binned Features Analysis



```
Analyzing AMT_INCOME_TOTAL_BINNED
```

```
Statistics for AMT_INCOME_TOTAL_BINNED:

Distribution and Default Rates:
                         Count  Default_Rate
AMT_INCOME_TOTAL_BINNED
Low                      91591          8.62
Medium                   64307          8.45
Very Low                 63698          8.20
High                     65176          7.55
Very High                22739          5.95

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
AMT_INCOME_TOTAL_BINNED
Low          1.07
Medium       1.05
Very Low     1.02
High         0.94
Very High    0.74
Name: Risk_Index, dtype: float64
```

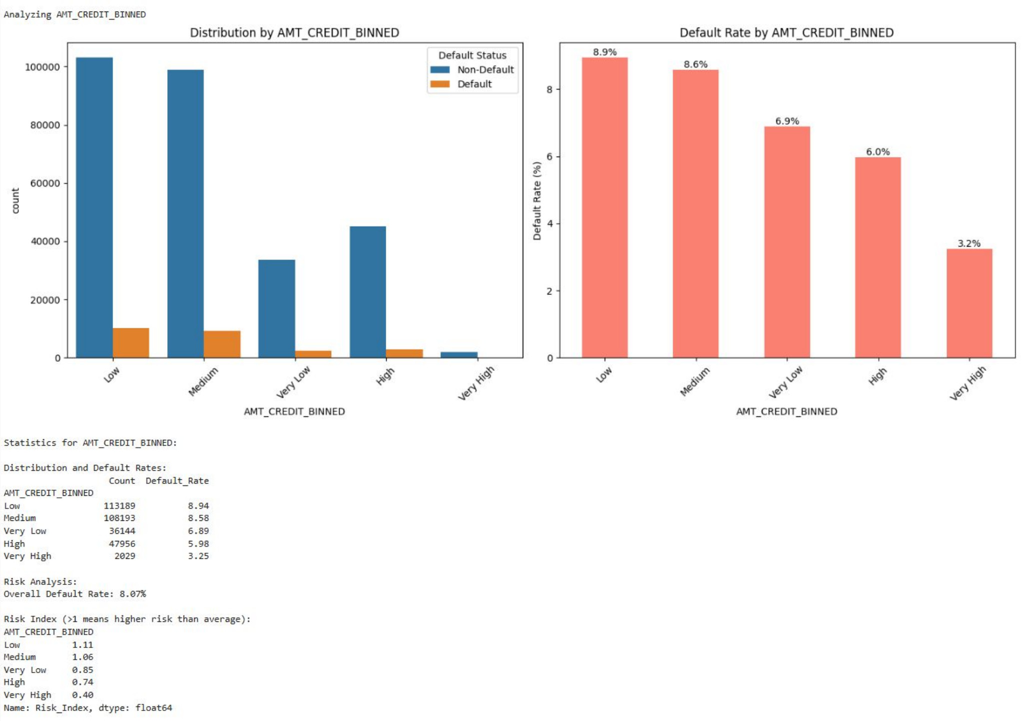
Key Insights: AMT_INCOME_TOTAL_BINNED

- **Default Rates by Income Group:**
  - **Highest Risk:**
    - **Low Income ($0–$135,000):** 8.6%.
  - **Lowest Risk:**
    - **Very High Income ($225,000+):** 6.0%.
  - Default rates decrease as income levels increase.
- **Risk Index:**
  - **Low Income:** 1.07 (above average risk).
  - **Very High Income:** 0.74 (below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - Focus on clients in the **Low Income** group for enhanced risk mitigation strategies.
2. **Policy Design:**
   - Tailor credit policies to incentivize higher-income groups with favorable loan terms.
3. **Predictive Modeling:**
   - Use AMT_INCOME_TOTAL_BINNED as a key feature to capture the impact of income on default likelihood.

Univariate Analysis - Binned Features Analysis



```
Analyzing AMT_CREDIT_BINNED
```

```
Statistics for AMT_CREDIT_BINNED:

Distribution and Default Rates:
                    Count  Default_Rate
AMT_CREDIT_BINNED
Low                113189          8.94
Medium             108193          8.58
Very Low            36144          6.89
High                47956          5.98
Very High            2029          3.25

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
AMT_CREDIT_BINNED
Low          1.11
Medium       1.06
Very Low     0.85
High         0.74
Very High    0.40
Name: Risk_Index, dtype: float64
```

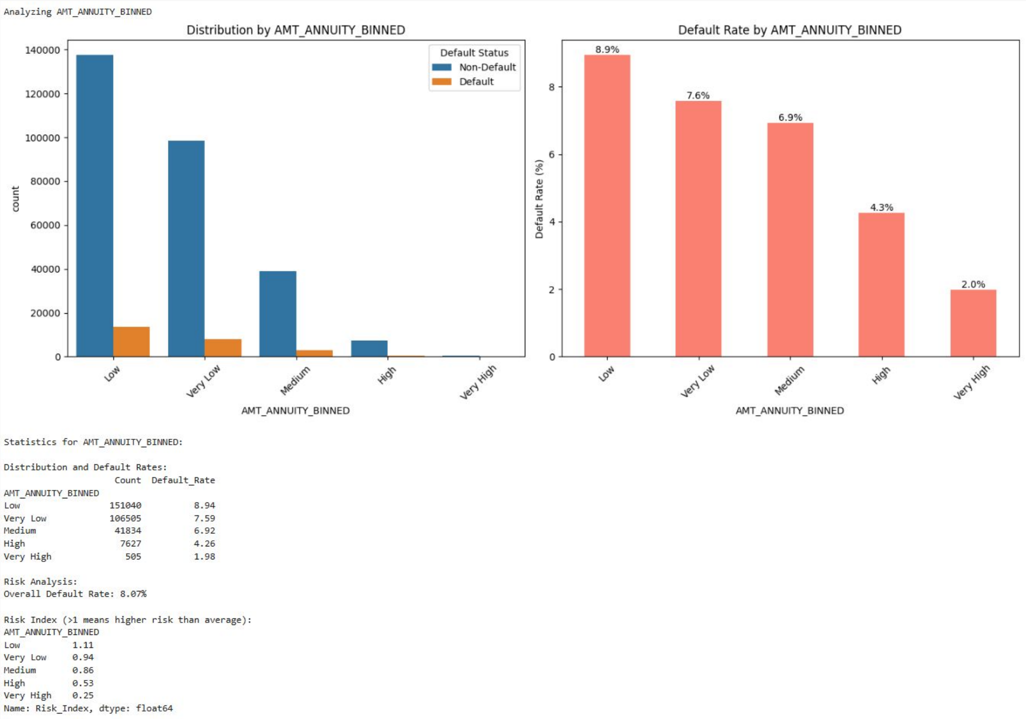
Key Insights: AMT_CREDIT_BINNED

- **Default Rates by Credit Group:**
  - **Highest Risk:**
    - **Low Credit ($0–$500,000):** 8.9%.
  - **Lowest Risk:**
    - **Very High Credit ($1,000,000+):** 3.2%.
  - Default rates slightly decrease with increasing credit amounts.
- **Risk Index:**
  - **Low Credit:** 1.11 (above average risk).
  - **Very High Credit:** 0.40 (significantly below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - Focus on clients with **Low** and **Medium** credit amounts for targeted risk management.
2. **Policy Refinement:**
   - Provide tailored loan options for higher credit clients (**Very High Credit**) who pose lower default risks.
3. **Predictive Modeling:**
   - Use AMT_CREDIT_BINNED as a feature to segment risk effectively across different credit levels.

Univariate Analysis - Binned Features Analysis



```
Analyzing AMT_ANNUITY_BINNED
```

```
Statistics for AMT_ANNUITY_BINNED:

Distribution and Default Rates:
                    Count  Default_Rate
AMT_ANNUITY_BINNED
Low                151040          8.94
Very Low           106505          7.59
Medium              41834          6.92
High                 7627          4.26
Very High             505          1.98

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
AMT_ANNUITY_BINNED
Low          1.11
Very Low     0.94
Medium       0.86
High         0.53
Very High    0.25
Name: Risk_Index, dtype: float64
```

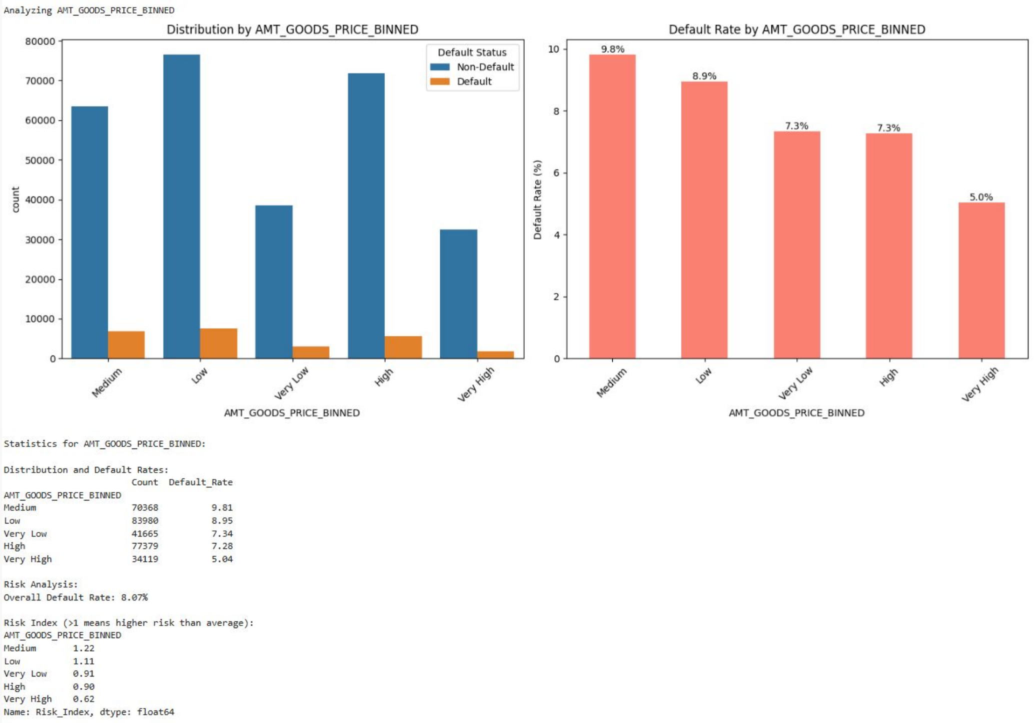
Key Insights: AMT_ANNUITY_BINNED

- **Default Rates by Annuity Group:**
  - **Highest Risk:**
    - **Low Annuity ($0–$30,000):** 8.9%.
  - **Lowest Risk:**
    - **Very High Annuity ($90,000+):** 2.0%.
  - Default rates decrease with increasing annuity amounts.
- **Risk Index:**
  - **Low Annuity:** 1.11 (above average risk).
  - **Very High Annuity:** 0.25 (significantly below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - Focus on clients with **Low** and **Very Low** annuity amounts for targeted risk mitigation.
2. **Policy Refinement:**
   - Offer flexible repayment terms for clients in high-risk annuity groups.
3. **Predictive Modeling:**
   - Use AMT_ANNUITY_BINNED to segment borrowers effectively by annuity-based risk levels.

Univariate Analysis - Binned Features Analysis



```
Analyzing AMT_GOODS_PRICE_BINNED
```

```
Statistics for AMT_GOODS_PRICE_BINNED:

Distribution and Default Rates:
                        Count  Default_Rate
AMT_GOODS_PRICE_BINNED
Medium                  70368          9.81
Low                     83980          8.95
Very Low                41665          7.34
High                    77379          7.28
Very High               34119          5.04

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
AMT_GOODS_PRICE_BINNED
Medium       1.22
Low          1.11
Very Low     0.91
High         0.90
Very High    0.62
Name: Risk_Index, dtype: float64
```

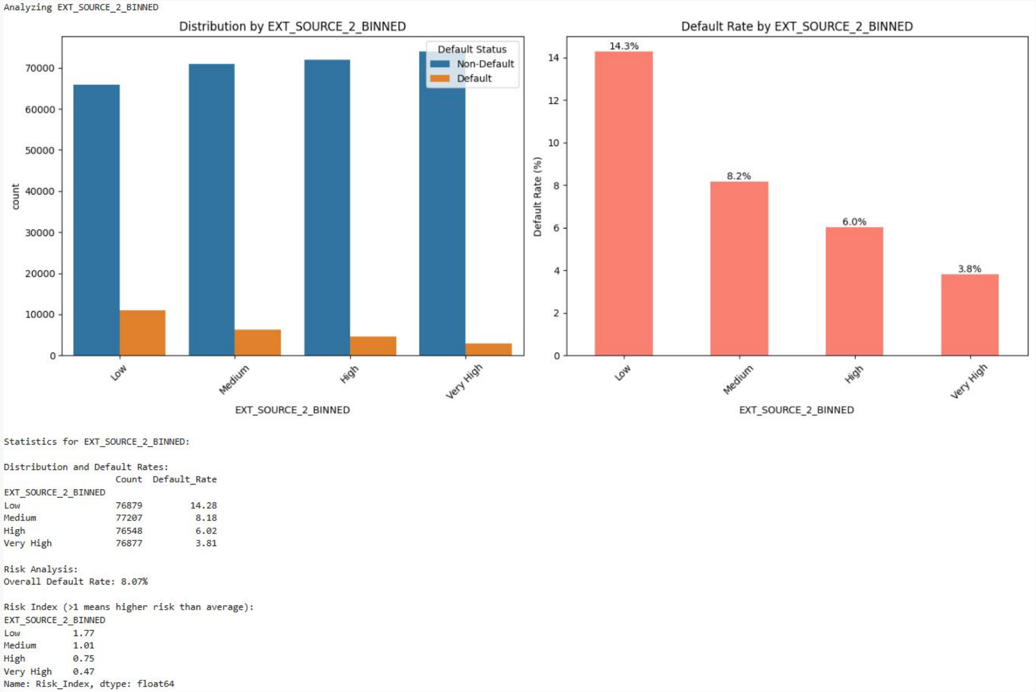
Key Insights: AMT_GOODS_PRICE_BINNED

- **Default Rates by Goods Price Group:**
  - ○ **Highest Risk:**
    - ■ **Medium Price ($400,000–$600,000):** 9.8%.
  - ○ **Lowest Risk:**
    - ■ **Very High Price ($1,000,000+):** 5.0%.
  - ○ Default rates decrease as goods prices increase.
- **Risk Index:**
  - ○ **Medium Price:** 1.22 (highest risk).
  - ○ **Very High Price:** 0.62 (significantly below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - ○ Focus on loans with **Medium** and **Low** goods prices, which show elevated default rates.
2. **Policy Refinement:**
   - ○ Offer favorable terms or additional scrutiny for goods priced in the **High** and **Very High** ranges, as they correlate with lower risk.
3. **Predictive Modeling:**
   - ○ Use AMT_GOODS_PRICE_BINNED to effectively segment borrowers based on goods price-related risk levels.

Univariate Analysis - Binned Features Analysis



```
Statistics for EXT_SOURCE_2_BINNED:

Distribution and Default Rates:
                     Count  Default_Rate
EXT_SOURCE_2_BINNED
Low                  76879         14.28
Medium               77207          8.18
High                 76548          6.02
Very High            76877          3.81

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
EXT_SOURCE_2_BINNED
Low          1.77
Medium       1.01
High         0.75
Very High    0.47
Name: Risk_Index, dtype: float64
```

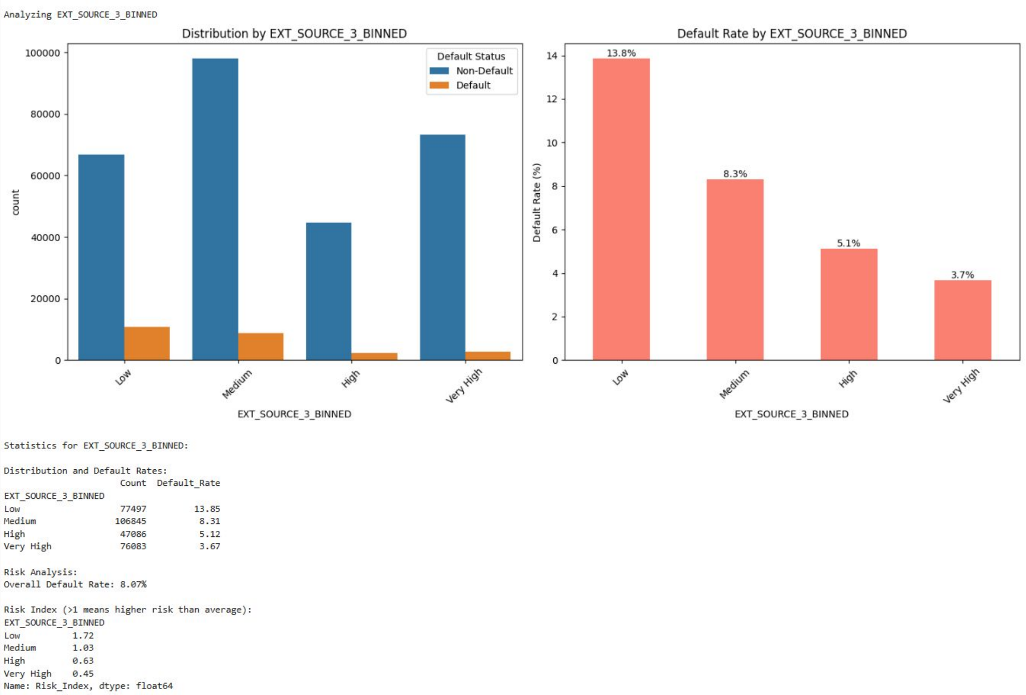
Key Insights: EXT_SOURCE_2_BINNED

- **Default Rates by EXT_SOURCE_2 Group:**
  - **Highest Risk:**
    - **Low (0.0–0.25):** 14.3%.
  - **Lowest Risk:**
    - **Very High (0.75–1.0):** 3.8%.
  - Default rates decrease as external credit scores increase.
- **Risk Index:**
  - **Low:** 1.77 (significantly above average risk).
  - **Very High:** 0.47 (significantly below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - Focus on borrowers in the **Low** bin for additional risk mitigation strategies.
2. **Credit Policy Design:**
   - Provide favorable terms or quicker approval processes for borrowers in the **Very High** bin, as they exhibit the lowest risk.
3. **Predictive Modeling:**
   - Use EXT_SOURCE_2_BINNED as a key feature in risk models, given its strong correlation with default likelihood.

Univariate Analysis - Binned Features Analysis



```
Analyzing EXT_SOURCE_3_BINNED
```

```
Statistics for EXT_SOURCE_3_BINNED:

Distribution and Default Rates:
                    Count  Default_Rate
EXT_SOURCE_3_BINNED
Low                 77497         13.85
Medium             106845          8.31
High                47086          5.12
Very High           76083          3.67

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
EXT_SOURCE_3_BINNED
Low          1.72
Medium       1.03
High         0.63
Very High    0.45
Name: Risk_Index, dtype: float64
```

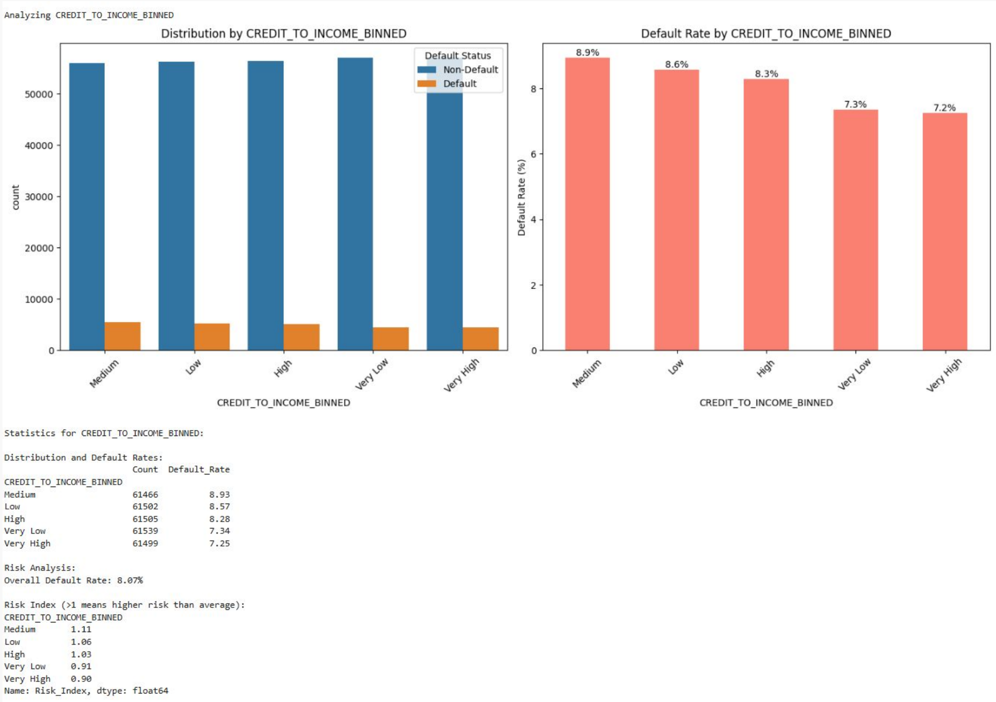
Key Insights: EXT_SOURCE_3_BINNED

- **Default Rates by EXT_SOURCE_3 Group:**
  - **Highest Risk:**
    - **Low (0.0–0.25):** 13.8%.
  - **Lowest Risk:**
    - **Very High (0.75–1.0):** 3.7%.
  - Default rates decrease as the external credit score increases.
- **Risk Index:**
  - **Low:** 1.72 (significantly above average risk).
  - **Very High:** 0.45 (significantly below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - Closely monitor borrowers in the **Low** bin for targeted risk management strategies.
2. **Credit Policy Design:**
   - Incentivize borrowers in the **Very High** bin with favorable loan terms or expedited processing, as they exhibit the lowest default risk.
3. **Predictive Modeling:**
   - Use EXT_SOURCE_3_BINNED as a critical feature to enhance the accuracy of predictive models.

Univariate Analysis - Binned Features Analysis



```
Analyzing CREDIT_TO_INCOME_BINNED
```

```
Statistics for CREDIT_TO_INCOME_BINNED:

Distribution and Default Rates:
                         Count  Default_Rate
CREDIT_TO_INCOME_BINNED
Medium                   61466          8.93
Low                      61502          8.57
High                     61505          8.28
Very Low                 61539          7.34
Very High                61499          7.25

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
CREDIT_TO_INCOME_BINNED
Medium       1.11
Low          1.06
High         1.03
Very Low     0.91
Very High    0.90
Name: Risk_Index, dtype: float64
```

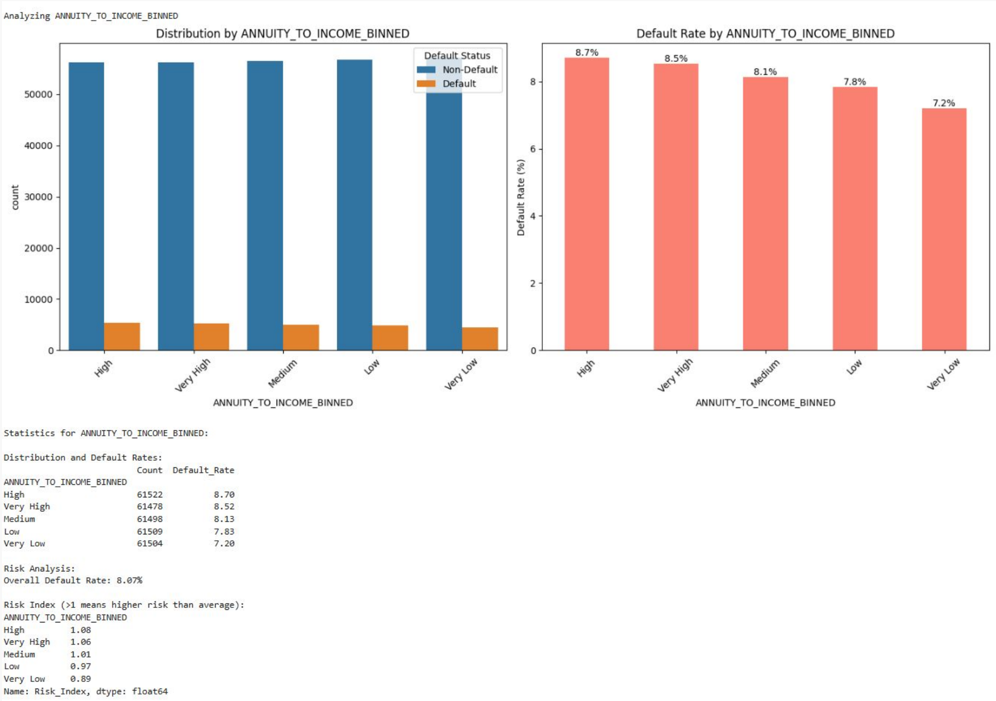
Key Insights: CREDIT_TO_INCOME_BINNED

- **Default Rates by Credit-to-Income Group:**
  - **Highest Risk:**
    - **Medium (0.3–0.5):** 8.93%.
  - **Lowest Risk:**
    - **Very High (> 0.7):** 7.25%.
  - Default rates are moderately higher in the **Medium** bin compared to other groups.
- **Risk Index:**
  - **Medium:** 1.11 (above average risk).
  - **Very High:** 0.90 (below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - Focus on borrowers in the **Medium** and **Low** bins for risk management strategies, as they exhibit higher default rates.
2. **Policy Refinement:**
   - Offer flexible loan repayment terms for borrowers in safer bins (**Very High** and **Very Low**) to encourage reliable borrowers.
3. **Predictive Modeling:**
   - Incorporate CREDIT_TO_INCOME_BINNED to enhance risk segmentation in credit models.

Univariate Analysis - Binned Features Analysis



```
Analyzing ANNUITY_TO_INCOME_BINNED
```

```
Statistics for ANNUITY_TO_INCOME_BINNED:

Distribution and Default Rates:
                          Count  Default_Rate
ANNUITY_TO_INCOME_BINNED
High                      61522          8.70
Very High                 61478          8.52
Medium                    61498          8.13
Low                       61509          7.83
Very Low                  61504          7.20

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
ANNUITY_TO_INCOME_BINNED
High         1.08
Very High    1.06
Medium       1.01
Low          0.97
Very Low     0.89
Name: Risk_Index, dtype: float64
```

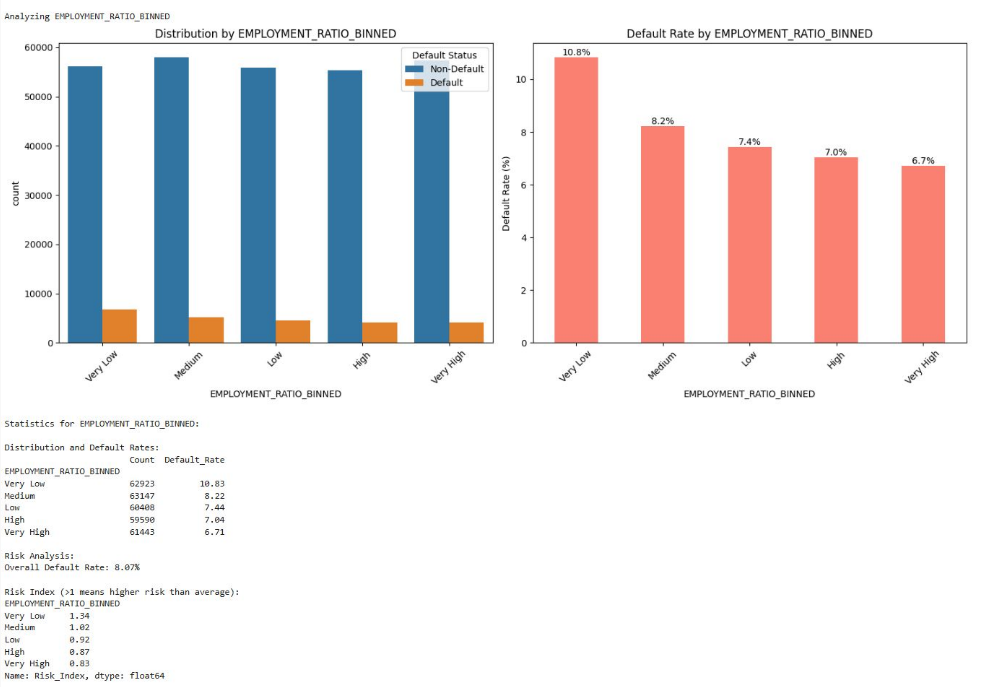
Key Insights: ANNUITY_TO_INCOME_BINNED

- **Default Rates by Annuity-to-Income Group:**
  - **Highest Risk:**
    - **High (0.2–0.3):** 8.70%.
  - **Lowest Risk:**
    - **Very Low (< 0.1):** 7.20%.
  - Default rates increase with higher annuity-to-income ratios.
- **Risk Index:**
  - **High:** 1.08 (above average risk).
  - **Very Low:** 0.89 (below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - Focus on borrowers in the **High** bin for additional assessment and risk mitigation strategies.
2. **Credit Policy Design:**
   - Prioritize loans to borrowers in the **Very Low** bin, as they represent the lowest default risk.
3. **Predictive Modeling:**
   - Utilize ANNUITY_TO_INCOME_BINNED as a key variable for effective segmentation in credit scoring models.

Univariate Analysis - Binned Features Analysis



```
Analyzing EMPLOYMENT_RATIO_BINNED
```

```
Statistics for EMPLOYMENT_RATIO_BINNED:

Distribution and Default Rates:
                        Count  Default_Rate
EMPLOYMENT_RATIO_BINNED
Very Low                62923         10.83
Medium                  63147          8.22
Low                     60408          7.44
High                    59590          7.04
Very High               61443          6.71

Risk Analysis:
Overall Default Rate: 8.07%

Risk Index (>1 means higher risk than average):
EMPLOYMENT_RATIO_BINNED
Very Low     1.34
Medium       1.02
Low          0.92
High         0.87
Very High    0.83
Name: Risk_Index, dtype: float64
```

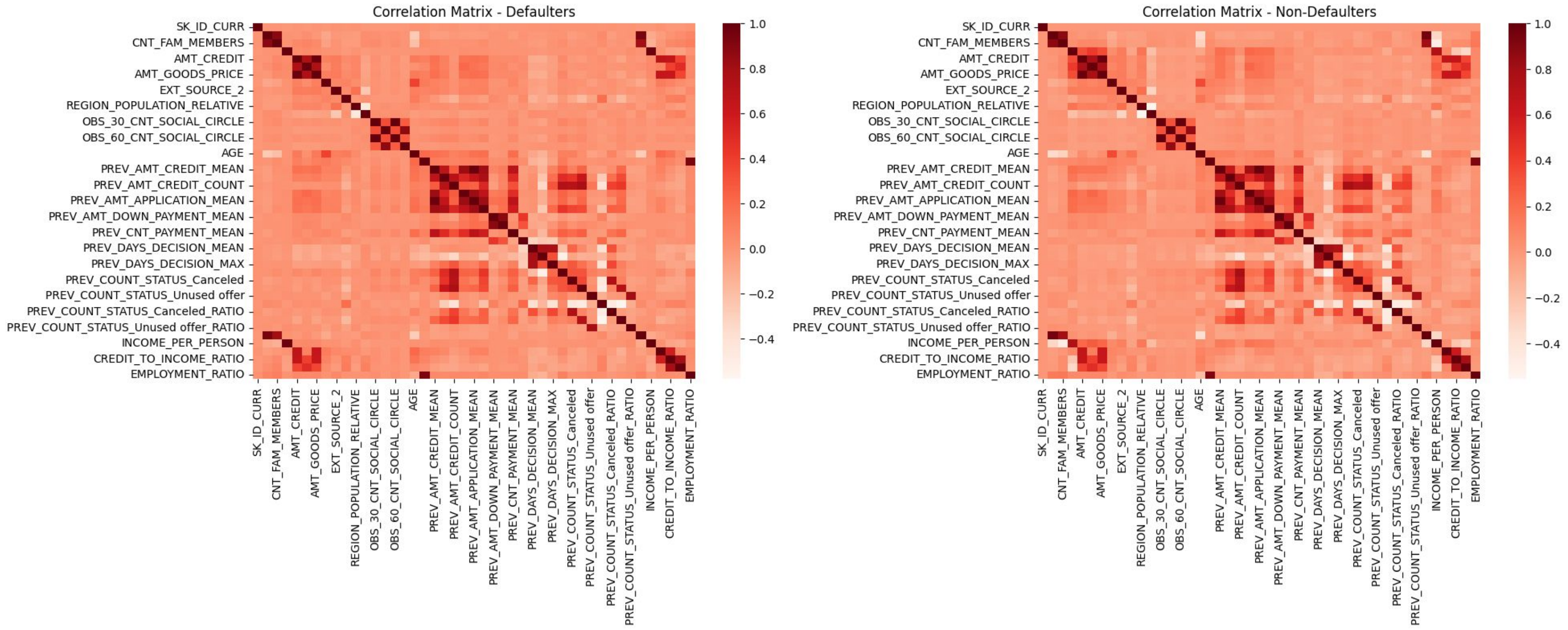
Key Insights: EMPLOYMENT_RATIO_BINNED

- **Default Rates by Employment Ratio Group:**
  - **Highest Risk:**
    - **Very Low (< 0.1):** 10.8%.
  - **Lowest Risk:**
    - **Very High (> 0.7):** 6.7%.
  - Default rates decrease as the employment ratio increases.
- **Risk Index:**
  - **Very Low:** 1.34 (above average risk).
  - **Very High:** 0.83 (below average risk).

Recommendations:

1. **High-Risk Monitoring:**
   - Closely monitor borrowers in the **Very Low** bin as they exhibit the highest default risk.
2. **Credit Policy Refinement:**
   - Prioritize loan approvals for borrowers in the **Very High** bin due to their low risk.
3. **Predictive Modeling:**
   - Use EMPLOYMENT_RATIO_BINNED to enhance the segmentation of risk profiles in credit scoring models.

Correlation Analysis



Correlation Analysis

Key Insights:

Defaulters:

- **Top Correlations:**
  - **OBS_30_CNT_SOCIAL_CIRCLE & OBS_60_CNT_SOCIAL_CIRCLE**: 0.998
  - **AMT_CREDIT & AMT_GOODS_PRICE**: 0.982
  - **AMT_INCOME_TOTAL & INCOME_PER_PERSON**: 0.976
  - **PREV_AMT_CREDIT_MEAN & PREV_AMT_APPLICATION_MEAN**: 0.975
  - **CNT_CHILDREN & CHILDREN_RATIO**: 0.934
  - **YEARS_EMPLOYED & EMPLOYMENT_RATIO**: 0.904
  - **CNT_CHILDREN & CNT_FAM_MEMBERS**: 0.884
  - **PREV_AMT_DOWN_PAYMENT_MEAN & PREV_AMT_DOWN_PAYMENT_MAX**: 0.825
  - **DEF_60_CNT_SOCIAL_CIRCLE & DEF_30_CNT_SOCIAL_CIRCLE**: 0.822
  - **PREV_AMT_CREDIT_MEAN & PREV_AMT_ANNUITY_MEAN**: 0.829
- **Patterns:**
  - Strong correlations between credit-related variables and social circle observations indicate potential predictors for payment difficulties.

Non-Defaulters:

- **Top Correlations:**
  - **OBS_30_CNT_SOCIAL_CIRCLE & OBS_60_CNT_SOCIAL_CIRCLE**: 0.998
  - **AMT_CREDIT & AMT_GOODS_PRICE**: 0.987
  - **PREV_AMT_CREDIT_MEAN & PREV_AMT_APPLICATION_MEAN**: 0.977
  - **YEARS_EMPLOYED & EMPLOYMENT_RATIO**: 0.952
  - **CNT_CHILDREN & CHILDREN_RATIO**: 0.952
  - **CNT_CHILDREN & CNT_FAM_MEMBERS**: 0.878
  - **DEF_30_CNT_SOCIAL_CIRCLE & DEF_60_CNT_SOCIAL_CIRCLE**: 0.896
  - **PREV_AMT_ANNUITY_MEAN & PREV_AMT_APPLICATION_MEAN**: 0.811
  - **PREV_AMT_CREDIT_MEAN & PREV_AMT_APPLICATION_MAX**: 0.810
  - **PREV_AMT_CREDIT_MEAN & PREV_AMT_APPLICATION_MEAN**: 0.807
- **Patterns:**
  - Similar trends to defaulters, with additional emphasis on employment-related correlations.

Recommendations:

1. **Dimensionality Reduction:**
   - Remove or combine highly correlated variables (e.g., OBS_30_CNT_SOCIAL_CIRCLE & OBS_60_CNT_SOCIAL_CIRCLE) to improve model efficiency and avoid redundancy.
2. **Feature Prioritization:**
   - Credit (AMT_CREDIT, AMT_GOODS_PRICE) and social circle variables should be prioritized in risk prediction models due to consistent high correlations.
3. **Segment-Specific Modeling:**
   - Explore differences in correlation patterns between defaulters and non-defaulters to design segment-specific predictive models.
4. **Employment-Related Insights:**
   - Utilize employment duration and ratio features as robust indicators for creditworthiness.

04

Insights and Recommendations

Conclusion: Insights from All Analyses

Credit and Income-Related Features:

- High correlations between AMT_CREDIT, AMT_GOODS_PRICE, and AMT_INCOME_TOTAL suggest redundancy in these features.
- Defaulters generally show higher default rates in lower credit and income bins, indicating financial stress.

Employment and Annuity Trends:

- YEARS_EMPLOYED and ANNUITY_TO_INCOME ratios show a consistent trend: lower bins (less experience or higher annuity burdens) are associated with higher risks.
- Employment stability (EMPLOYMENT_RATIO) shows a significant relationship with default rates, with "Very Low" stability being most risky.

Age and Family Structure:

- Younger individuals (lower AGE bins) and those with more dependents (CNT_FAM_MEMBERS) exhibit higher default risks.
- Borrowers in the "Very High" age bins (older individuals) and smaller families are lower risk.

External Sources (EXT_SOURCE_2 & EXT_SOURCE_3):

- These scores are strong predictors of risk. Lower external source scores are associated with much higher default rates, reflecting poor financial health or external risk evaluation.

Credit Ratios:

- Ratios such as CREDIT_TO_INCOME and ANNUITY_TO_INCOME are effective at segmenting risk:
  - Higher ratios (indicating more financial burden) show higher default rates.
  - These ratios should be prioritized for predictive models.

Behavioral Patterns (Social Circles):

- Features like OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE have strong correlations and can provide insights into default behavior. Larger social circles may indicate potential financial dependencies.

Key Risk Segments Identified:

- **High-Risk Groups:**
  - Borrowers with low income, young age, limited employment, or higher credit/annuity burdens.
  - Low external source scores (EXT_SOURCE_2 & EXT_SOURCE_3).
- **Low-Risk Groups:**
  - Borrowers in higher income or credit bins, older age, and stable employment.

Business Insights

Insights in Business Terms

- **Key Drivers of Default:**
  - Lower income, higher credit amount, and larger family sizes are strongly associated with payment difficulties.
  - Younger individuals showed a higher likelihood of defaulting.
- **Impact of Imbalance:**
  - Models or policies need to address the bias caused by data imbalance, ensuring fairness in predictions.

Recommendations:

1. **Intervention Strategies:**
   - Offer financial counseling or restructuring options to clients identified with high default risk.
2. **Data Collection Improvements:**
   - Reduce missing data by ensuring better recording practices.
3. **Balanced Metrics for Modeling:**
   - Use techniques like SMOTE for balanced training datasets.

Business Insights

Customer Segmentation for Risk Management:

- **High-Risk Customers**:
  - Borrowers with low income (AMT_INCOME_TOTAL) and high credit burden (CREDIT_TO_INCOME_RATIO, ANNUITY_TO_INCOME_RATIO).
  - Younger individuals (AGE) and less stable employment (EMPLOYMENT_RATIO).
  - Lower external credit scores (EXT_SOURCE_2, EXT_SOURCE_3).
- **Low-Risk Customers**:
  - Older individuals with higher income and less financial burden.
  - Borrowers with higher external credit scores and stable employment history.

Actionable Insight: Create segmented loan offerings based on risk profiles to optimize risk-adjusted returns. For instance:

- High-risk customers can be offered loans with higher interest rates, tighter terms, or additional collateral requirements.
- Low-risk customers can be incentivized with better loan terms to build long-term relationships.

Business Insights

Improve Loan Approval and Pricing Strategies:

- Borrowers in lower AGE, AMT_INCOME_TOTAL, and CREDIT_TO_INCOME bins are more likely to default. This group represents an opportunity to apply stricter underwriting criteria.
- Use external scores (EXT_SOURCE_2 and EXT_SOURCE_3) as strong predictors to refine creditworthiness assessments.

Actionable Insight: Develop dynamic pricing models by incorporating predictors like income ratios and external scores. Reward customers with high scores and stable income with lower interest rates to encourage borrowing while mitigating risk.

Improve Loan Approval and Pricing Strategies:

- Borrowers in lower AGE, AMT_INCOME_TOTAL, and CREDIT_TO_INCOME bins are more likely to default. This group represents an opportunity to apply stricter underwriting criteria.
- Use external scores (EXT_SOURCE_2 and EXT_SOURCE_3) as strong predictors to refine creditworthiness assessments.

Actionable Insight: Develop dynamic pricing models by incorporating predictors like income ratios and external scores. Reward customers with high scores and stable income with lower interest rates to encourage borrowing while mitigating risk.

Actionable Recommendations:

1. **Focus on Key Predictors:**
   - Include features like external source scores, credit ratios, age, and employment stability in risk modeling.
   - Eliminate redundant variables (e.g., highly correlated credit-related features).
2. **Enhance Risk-Based Segmentation:**
   - Use binned analysis to develop risk categories for targeted interventions or differentiated loan offerings.
   - Monitor high-risk groups (e.g., low external scores, low income) for early signs of financial distress.
3. **Dimensionality Reduction:**
   - Leverage correlation matrices to reduce feature redundancy, improving model efficiency.
4. **Behavior-Based Insights:**
   - Social circle size and observed defaults in similar groups can offer additional behavioral risk metrics.

Thanks!

“Talent wins games, but teamwork and intelligence win championships.”

- Michael Jordan