FINANCIAL
RISK
ANALYSIS
IN BANKING
LOAN RSISK ANALYSIS

About Me

- Proficient in **Python**, **SQL**, **Tableau**, and **Power BI**
- Pursuing a **Diploma in Data Analysis** from **Hyper Island** (2023-2025)
- Experienced in teaching statistics
- Skilled in **data visualization** and **predictive modeling**
- Passionate about leveraging data-driven insights to make informed decisions and tackle complex business challenges

Contact Information:

- LinkedIn

Project objective

- Identify key patterns that indicate loan repayment risk among applicants.
- Assist the bank in making better decisions:
  Deny or reduce loan amounts for high-risk customers.
  Offer loans to reliable applicants at lower interest rates.
- Reduce loan default rates by analyzing repayment patterns.

Database Overwiew

Databases Used in Credit Risk Analysis:

Application_Data

- Contains details about loan applicants, including demographic information, loan conditions, and financial history.
- **Total Records**: 307,511
- **Columns**: 122
- **Memory Usage**: 286.2 MB

Previous_Application

- Includes historical data on prior loan requests made by applicants.
- **Total Records**: 1,670,214
- **Columns**: 37

Key Metrics

- **TARGET**: Indicates if a customer had difficulty repaying a loan (1) or not (0).
- **AMT_INCOME_TOTAL**: Total income of the applicant.
- **AMT_CREDIT**: Total amount of the loan requested.
- **AMT_ANNUITY**: Yearly or monthly loan installment amount.
- **DAYS_EMPLOYED**: Number of days the applicant has been employed (can also be converted to years).
- **CNT_CHILDREN** and **CNT_FAM_MEMBERS**: Number of dependents and family members of the applicant.
- **Other Categorical Features**: Such as CODE_GENDER, NAME_EDUCATION_TYPE, OCCUPATION_TYPE, and NAME_HOUSING_TYPE.

Data Cleaning

Steps Taken:

- Identified Missing Values
- Dropped Columns with >60% Missing Values
- Filled Missing Values in Numerical Columns with Mean
- Converted Financial Columns to Integer Format
- Converted Negative Values Columns to Positive Values
- Processed Missing Values in Categorical Columns by Filling with "Unknown"

```python
# Data cleaning
# checking for missing value
missing_value_pre = pre_data.isnull().sum()

# Calculate percentage of missing values for each column
miss_value_pre_percentage = (missing_value_pre / len(pre_data)) * 100
print('miss_value_pre_percentage:')
print(miss_value_pre_percentage)
```

```python
# filling valous for columns that have missing values for app_data_clean

#selecting numerical colmns
app_select_numerical_with_nun= app_data_clean.select_dtypes(include=['int64','float64']).columns

# filling numerical data
for columns in app_select_numerical_with_nun:
    if app_data_clean[columns].isnull().sum()>0:
        app_data_clean[columns].fillna(app_data_clean[columns].mean(),inplace=True)
```

```python
# Convert `AMT_*` columns to integer
amt_columns_previous = ['AMT_APPLICATION', 'AMT_CREDIT', 'AMT_DOWN_PAYMENT', 'AMT_GOODS_PRICE']
for col in amt_columns_previous:
    pre_data_clean[col] = pre_data_clean[col].astype('int')

# List of `DAYS_*` columns to convert to absolute values
days_columns_previous = ['DAYS_FIRST_DUE', 'DAYS_LAST_DUE_1ST_VERSION', 'DAYS_LAST_DUE',
                         'DAYS_TERMINATION']
for col in days_columns_previous:
    pre_data_clean[col] = pre_data_clean[col].abs()
```

```python
# Find the names of columns that have more than 60% missing values
columns_app_to_drop=miss_value_app_percentage[miss_value_app_percentage>60].index

# drop these columns from the original DataFrame
app_data_clean= app_data.drop(columns=columns_app_to_drop)

# Display the remaining columns
print('remaining columns:',app_data_clean.columns)
```

```python
# List of day-related columns to make positive
days_columns = ['DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH']

# Converting each column to its absolute value
for col in days_columns:
    app_data_clean[col] = app_data_clean[col].abs()
```

```python
#selecting string columns with missing values for app_data_clean
app_select_string_with_nun = app_data_clean.select_dtypes(include=['object']).columns

# fillinh with unknon
for col in app_select_string_with_nun:
    if app_data_clean[col].isnull().sum():
        app_data_clean[col].fillna('unknown', inplace=True)
```

Analyzing Data

The dataset was divided into two categories:

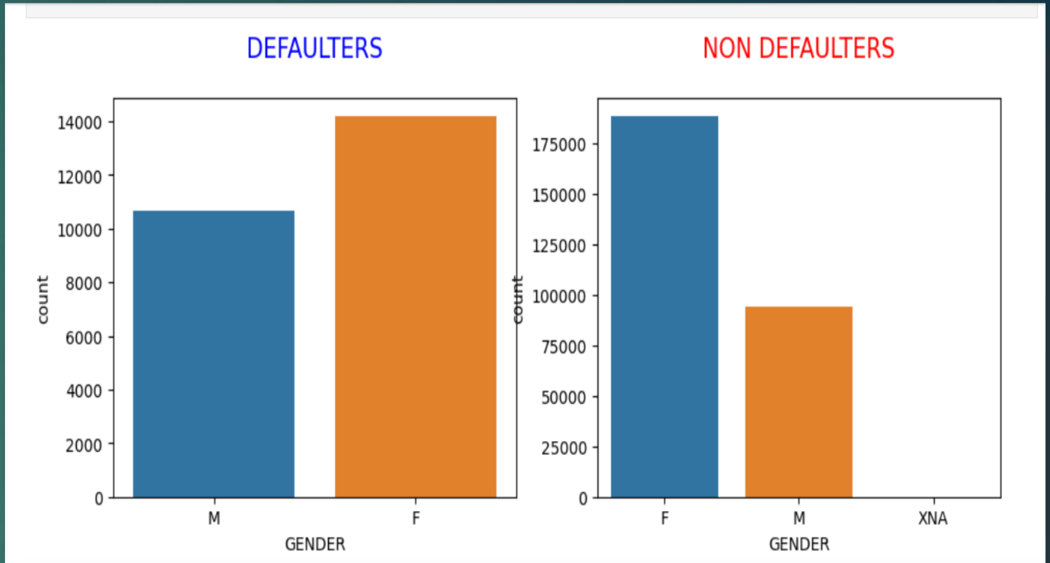
- **Defaulters**: Clients who had difficulty repaying their loans.
- **Non-Defaulters**: Clients with no repayment issues.

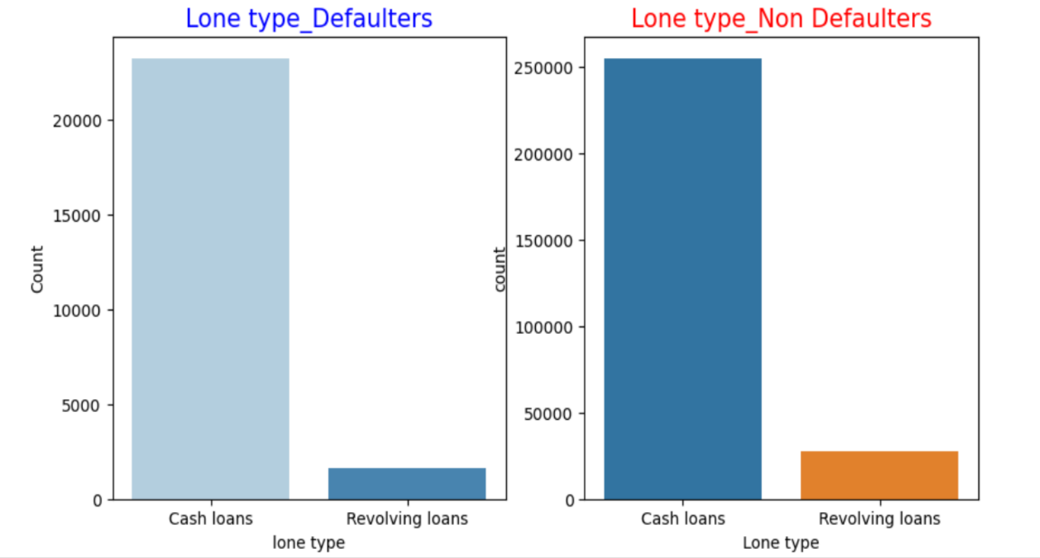
Gender Distribution:

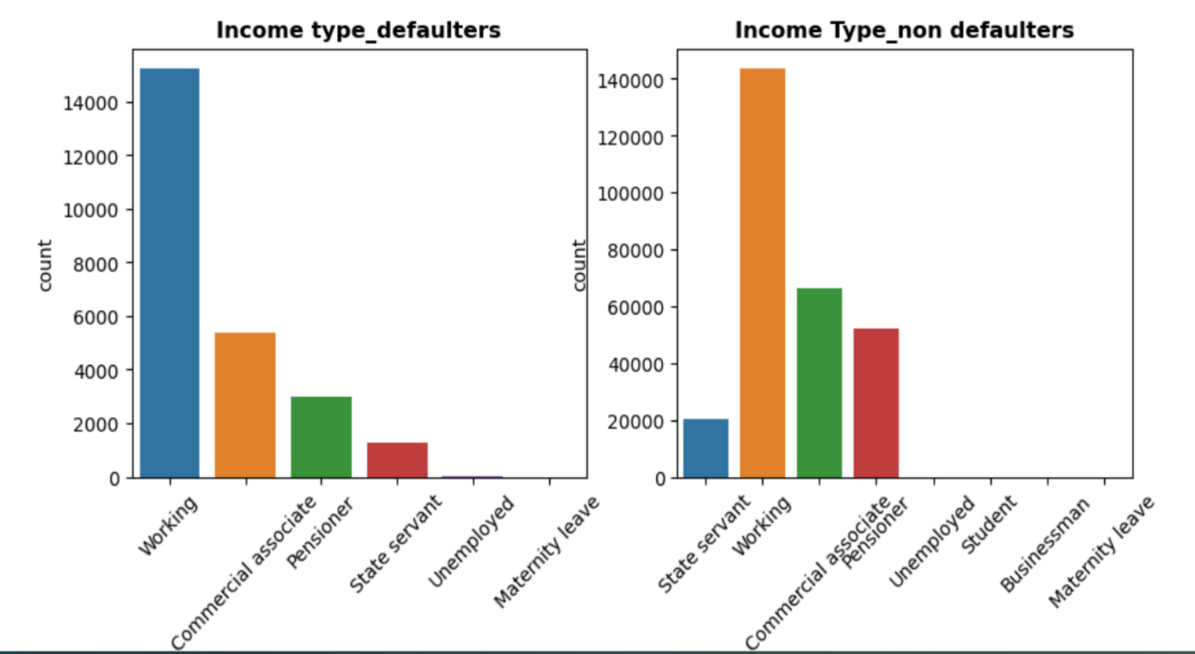
- In **Defaulters**: Higher proportion of females compared to males.
- In **Non-Defaulters**: Females are still a majority, but the distribution is more balanced.

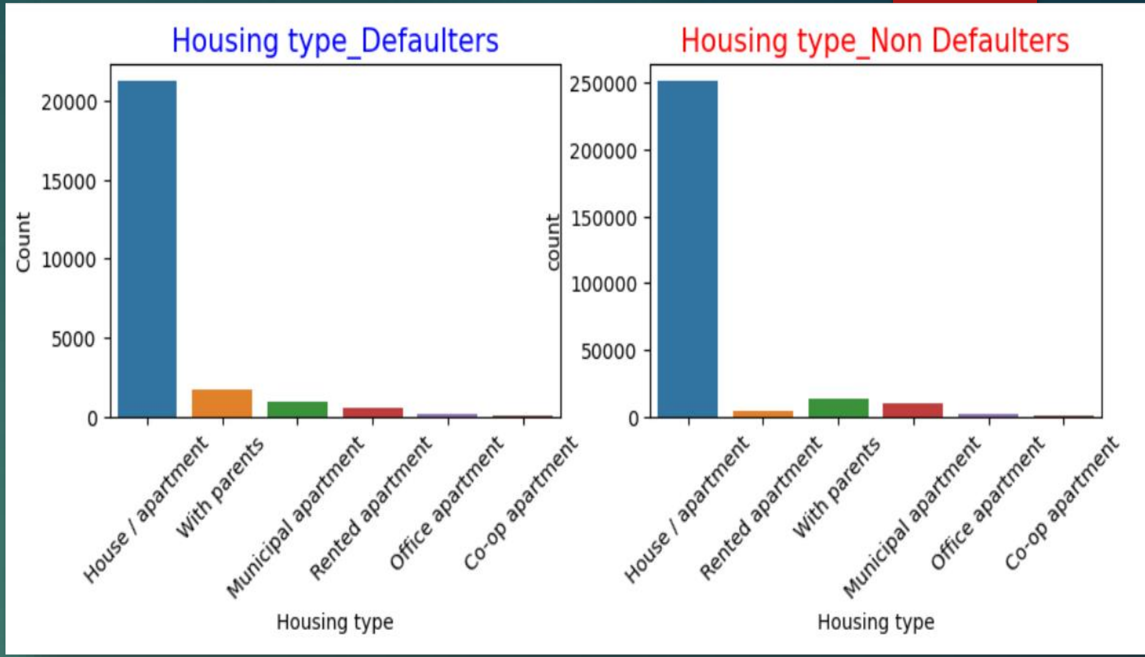
Loan Type:

- **Cash Loans** are the predominant loan type for both defaulters and non-defaulters.
- **Revolving Loans**: Slightly higher proportion among defaulters, though still a small percentage overall.









Income Type Distribution

•**Defaulters**:

- The majority are in the "Working" category, followed by "Commercial associate" and "Pensioner."
- Smaller representation from "State servant," "Unemployed," and "Maternity leave" categories.

•**Non-Defaulters**:

- A higher proportion in "State servant" and "Pensioner" categories compared to defaulters.
- The "Working" and "Commercial associate" categories remain prominent but at a larger scale than defaulters.

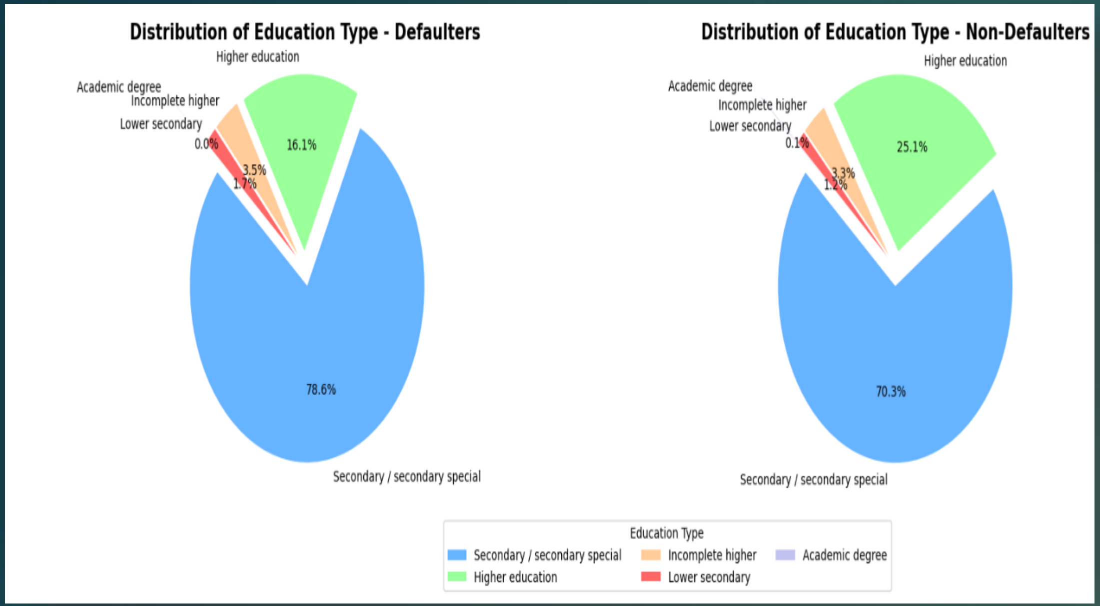
Housing Type Distribution

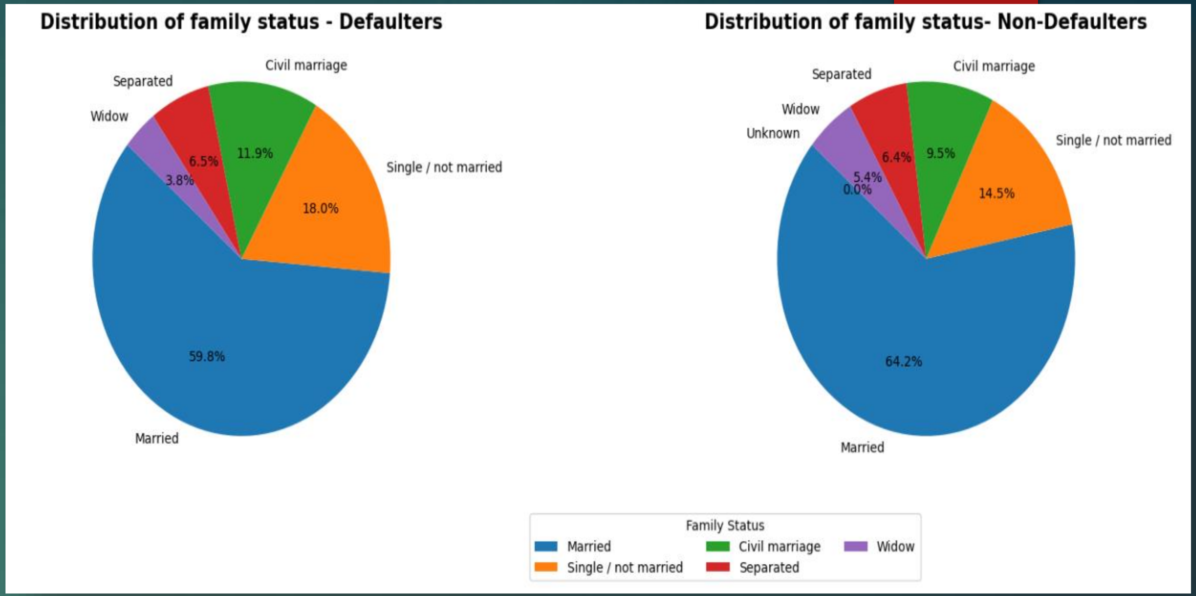
•**Defaulters**:

- Most live in "House / apartment," with smaller proportions living "With parents" or in "Rented apartment.«

•**Non-Defaulters**:

- Similar trend where the majority live in "House / apartment," but a slightly higher proportion also live "With parents."





Distribution of Education Type

•**Defaulters**:

- The majority have "Secondary / secondary special" education (78.6%).
- A smaller percentage have "Higher education" (16.1%), with even fewer in "Incomplete higher" or "Academic degree" categories.

•**Non-Defaulters**:

- Similarly, most have "Secondary / secondary special" education (70.3%), though a slightly higher percentage have "Higher education" (25.1%).

Distribution of Family Status

•**Defaulters**:

•A significant portion are "Married" (59.8%), followed by "Single / not married" (18%).

•Other statuses like "Civil marriage" and "Separated" have smaller representations.

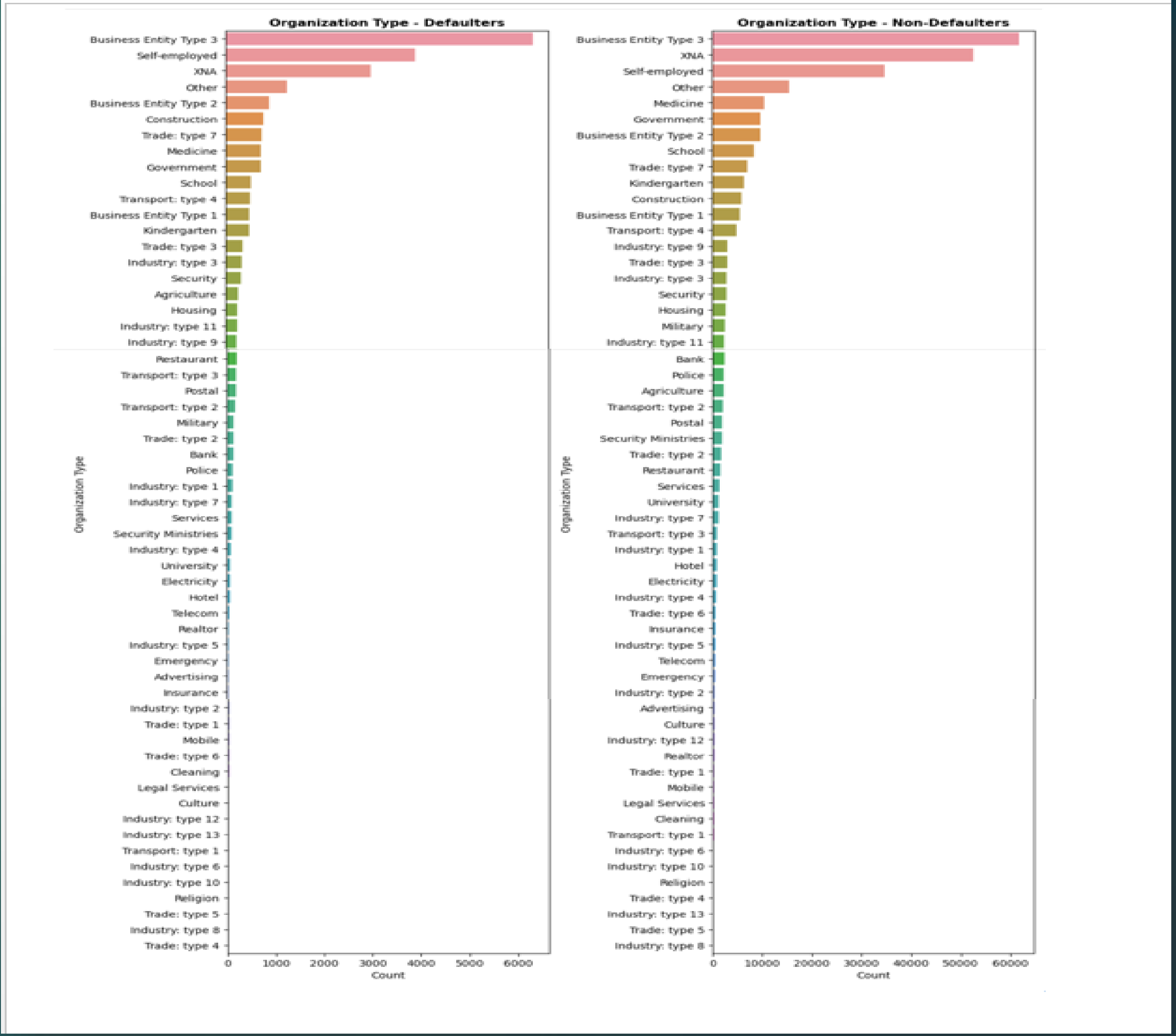
•**Non-Defaulters**:

•"Married" status dominates more (64.2%), and the "Single / not married" category is also significant (14.5%).

•Similar patterns are observed for "Civil marriage" and other categories, with slight variations.

Organization Type distribution:

- **Defaulters**:
  - Highest in "Business Entity Type 3"
  - High count in "Self-employed" and "XNA"
  - Noticeable numbers in "Construction," "Medicine," and "Government"
- **Non-Defaulters**:
  - Highest in "Business Entity Type 3"
  - Significant in "Self-employed" and "XNA"
  - Higher representation in stable sectors: "Government," "School," and "Medicine"



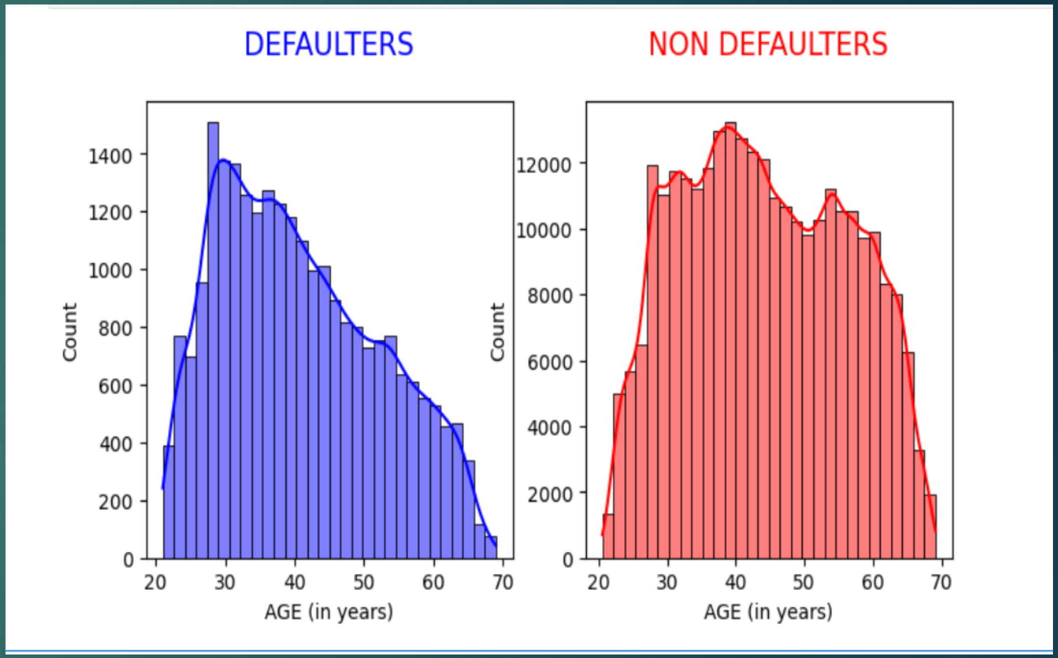
Age distribution for defaulters and non-defaulters

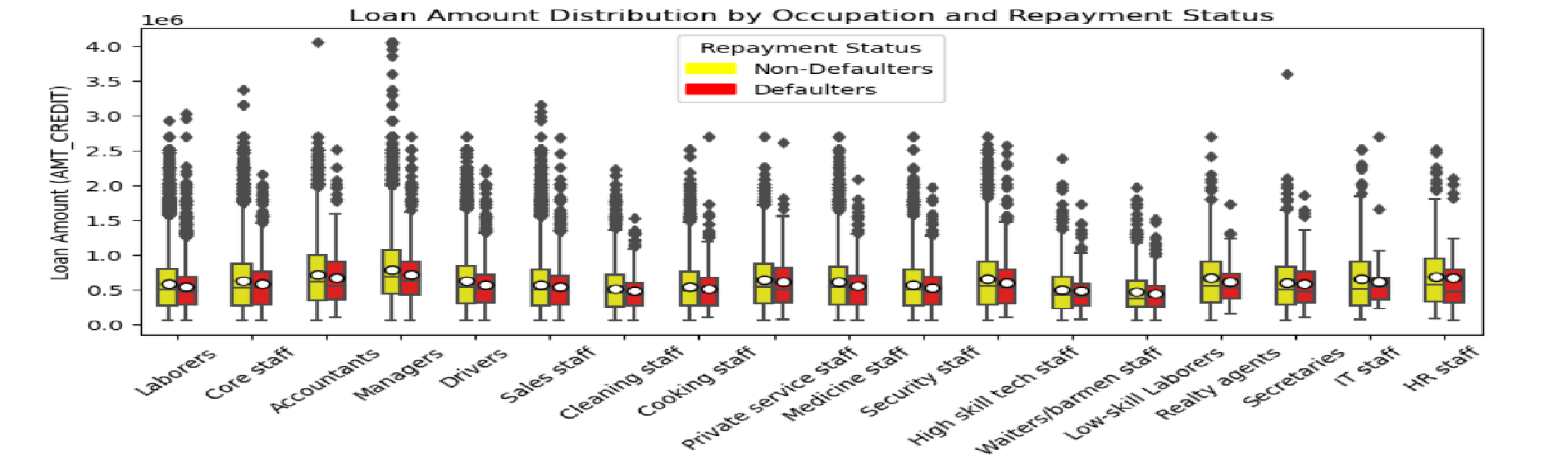
Defaulters:

- The age distribution peaks around 30-35 years, indicating that younger individuals are more likely to default on loans.
- The count steadily declines as age increases, showing fewer older individuals among defaulters.

Non_Defaulters:

- The distribution is more evenly spread across ages, with a noticeable peak around the ages of 40-50.
- This suggests that non-defaulters are often older and have more financial stability, which might contribute to a lower default rate.





- The chart show loan amount across different job occupation Defaulters and Non_Defaulters:

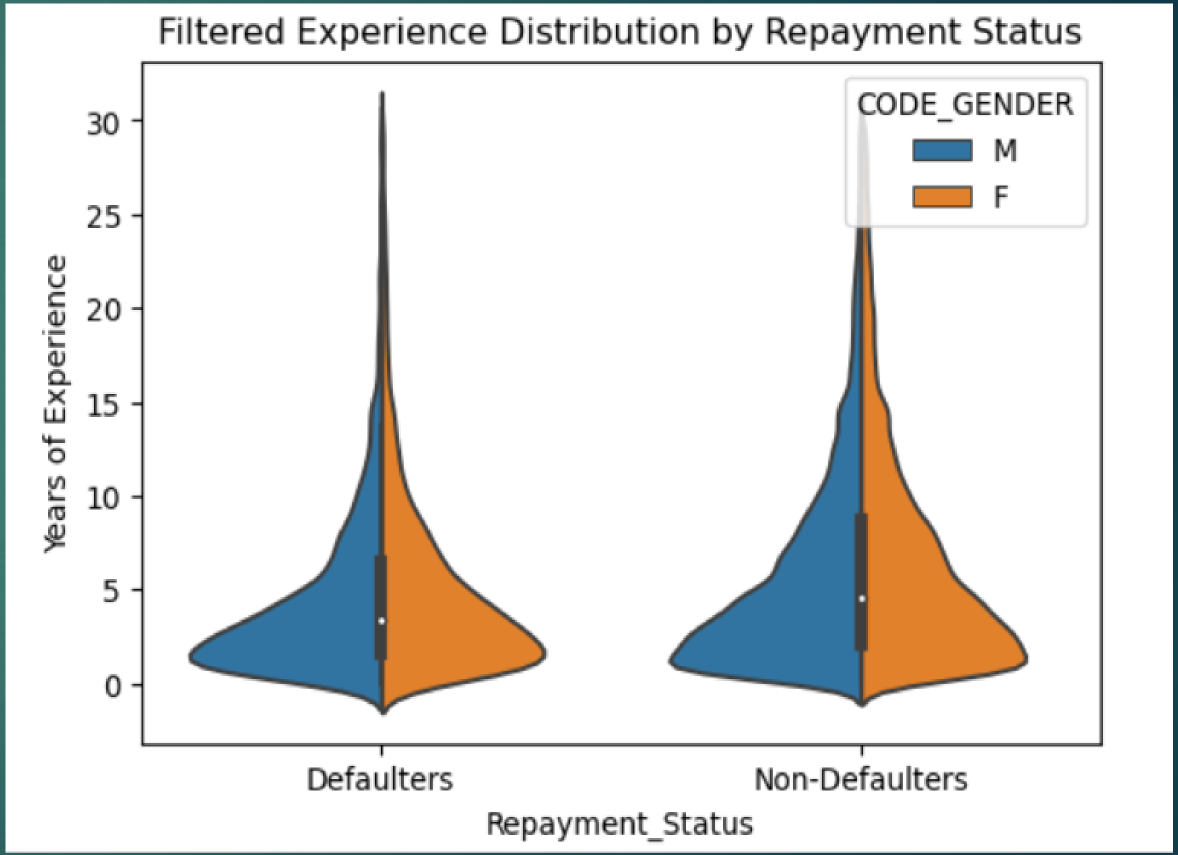
- **Higher Loan Amounts**: "Managers" and "Core Staff" often have higher loans, likely due to better financial standing.
- **Lower Loan Amounts**: Roles like "Cleaning Staff" and "Laborers" usually have smaller loans, reflecting lower income levels.
- **Wider Loan Spread for Defaulters**: Defaulters, especially among "Core Staff" and "Managers," show a broader range of loan amounts, indicating more variability in financial stability.
- **High Outliers Among Non-Defaulter Accountants**: Some non-defaulter accountants have very high loans, suggesting strong financial stability for certain individuals.

Years of Experience by Repayment Status

Defaulters tend to have slightly less experience than non-defaulters, but the distributions are quite similar overall.

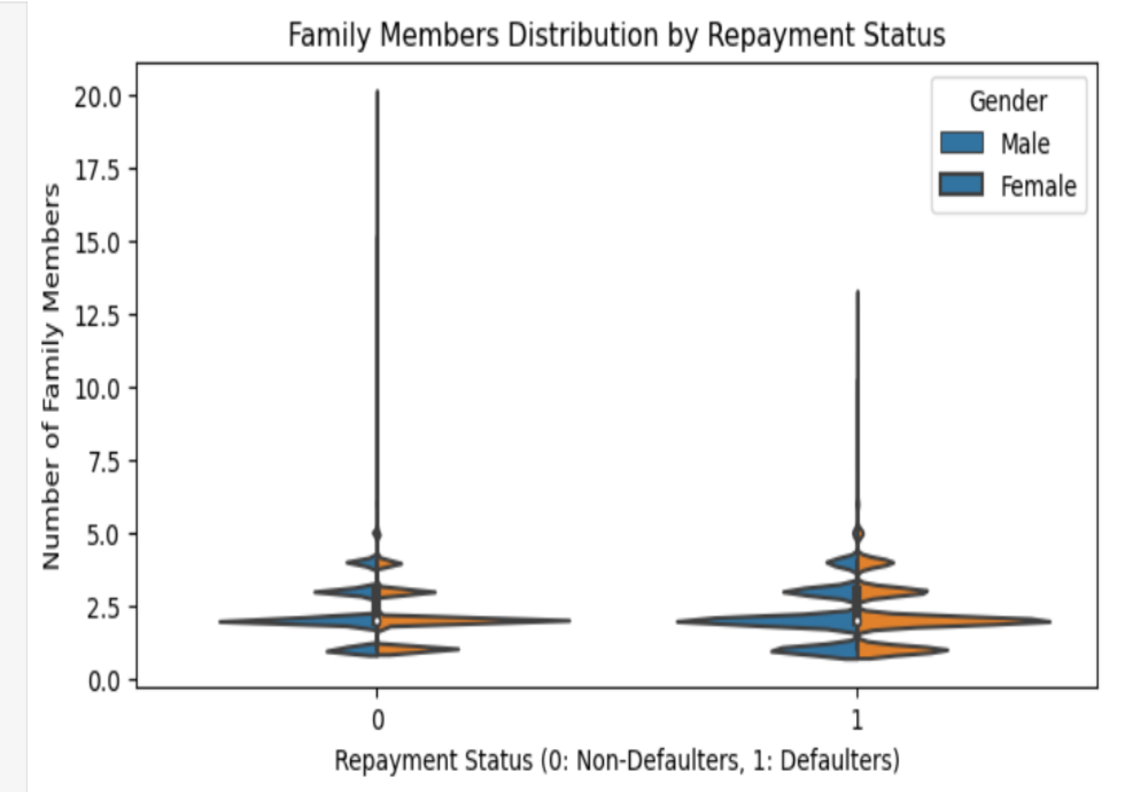
Gender Differences: The experience distribution is fairly balanced between males and females in both defaulters and non-defaulters, with a slight increase in years of experience for males in both categories.

Range: Both groups have a long tail, indicating a few individuals with significantly higher experience levels.



Family Members Distribution by Repayment Status

- **Median Family Size**: Both defaulters and non-defaulters mostly have small family sizes, typically around 2-3 members.
- **Outliers**: There are extreme outliers in both groups with family sizes larger than 10, though they are very rare.
- **Gender Proportions**: The distribution across family sizes appears similar for both males and females within each repayment status, with minor differences.

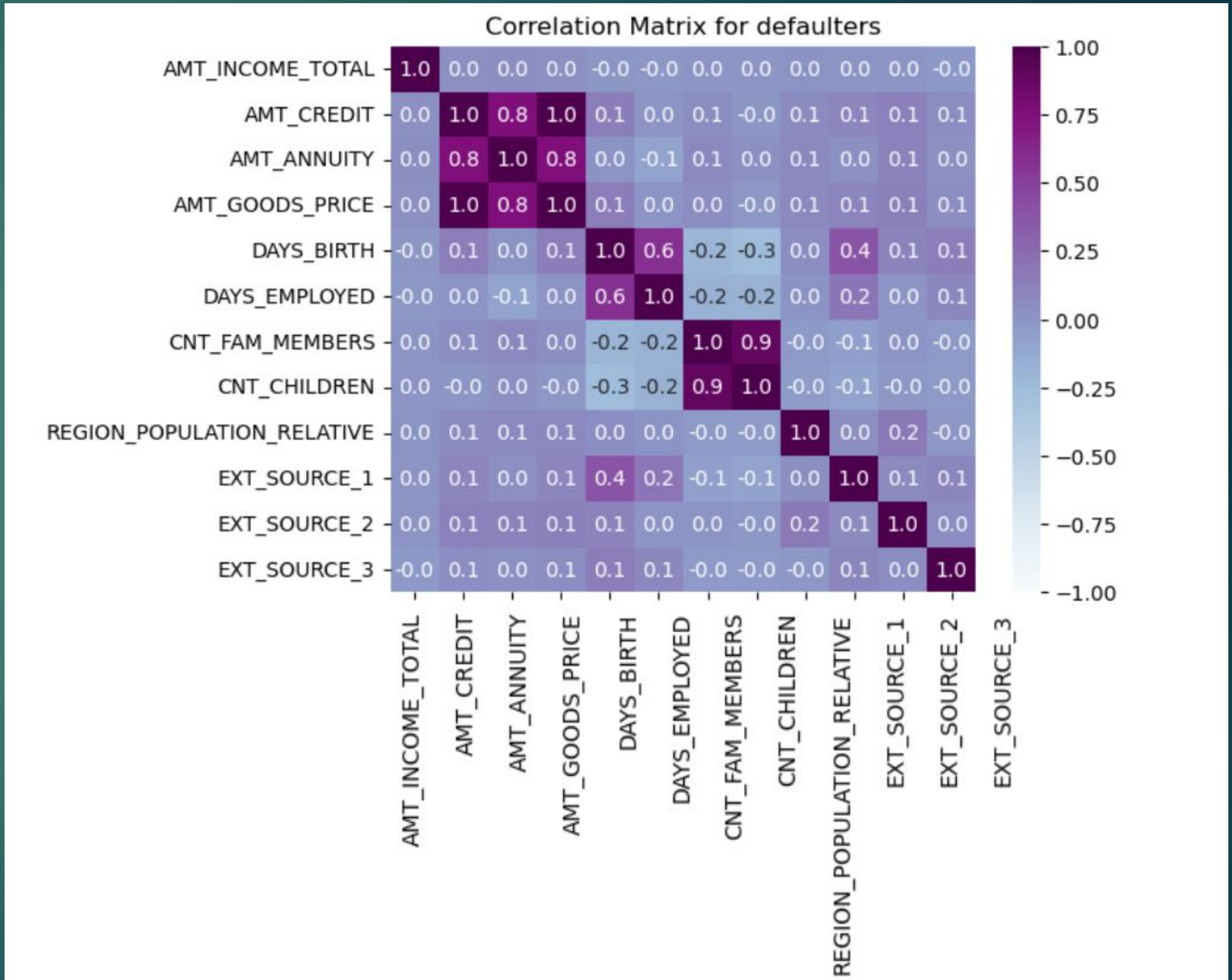


The correlation matrix for defaulters:

- **Loan & Goods Price**: Strong correlation between AMT_CREDIT and AMT_GOODS_PRICE, linking higher loans with higher-priced goods.

- **Age & Risk Scores**: Negative correlation between DAYS_BIRTH and EXT_SOURCE_3, indicating older applicants often have better risk scores.

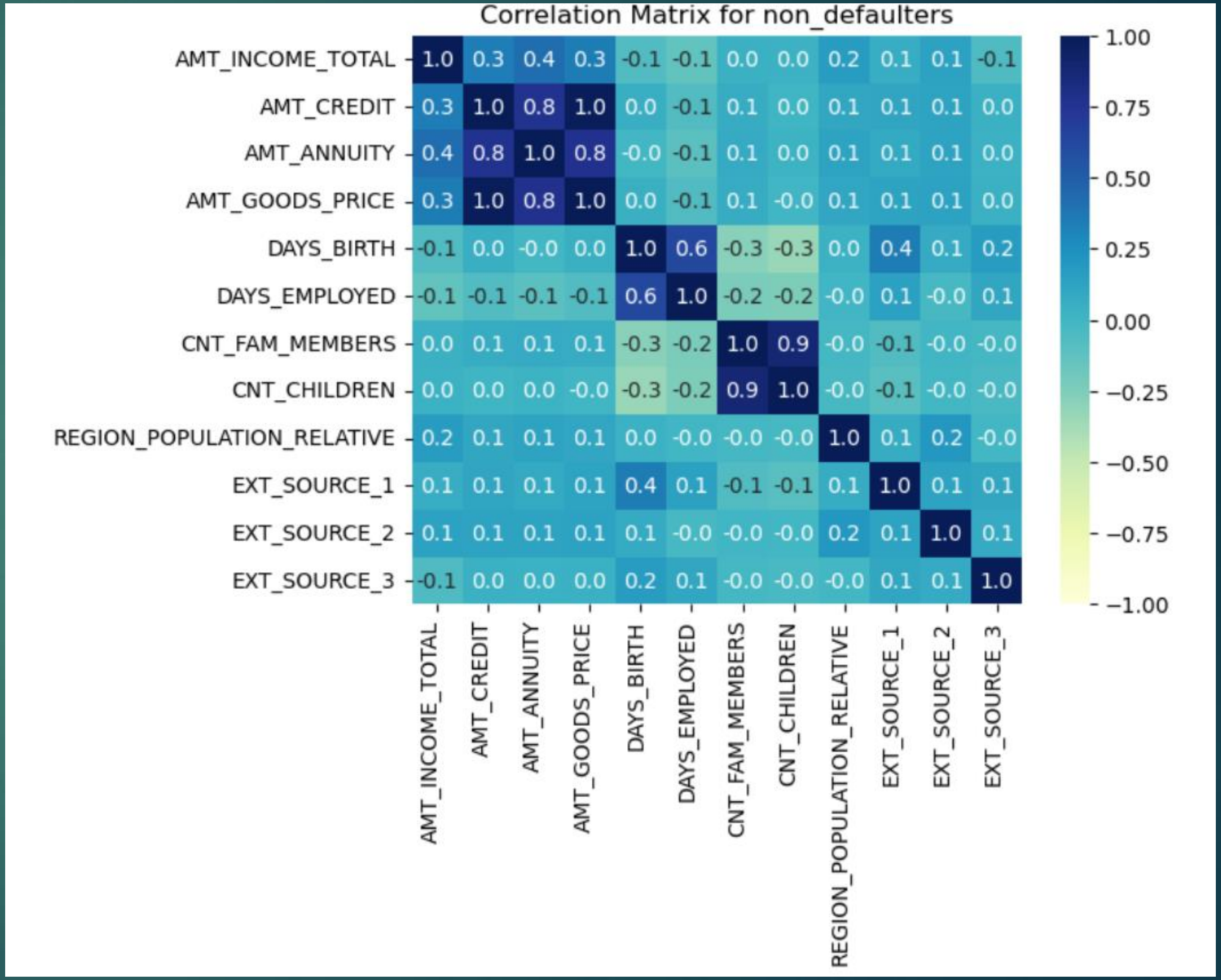
- **Employment & External Scores**: Weak correlation between DAYS_EMPLOYED and EXT_SOURCE values, suggesting job tenure has minimal impact on risk scores.



- **High Correlation**: Strong positive correlation between AMT_GOODS_PRICE, AMT_CREDIT, and AMT_ANNUITY, suggesting larger loans are associated with higher payments.

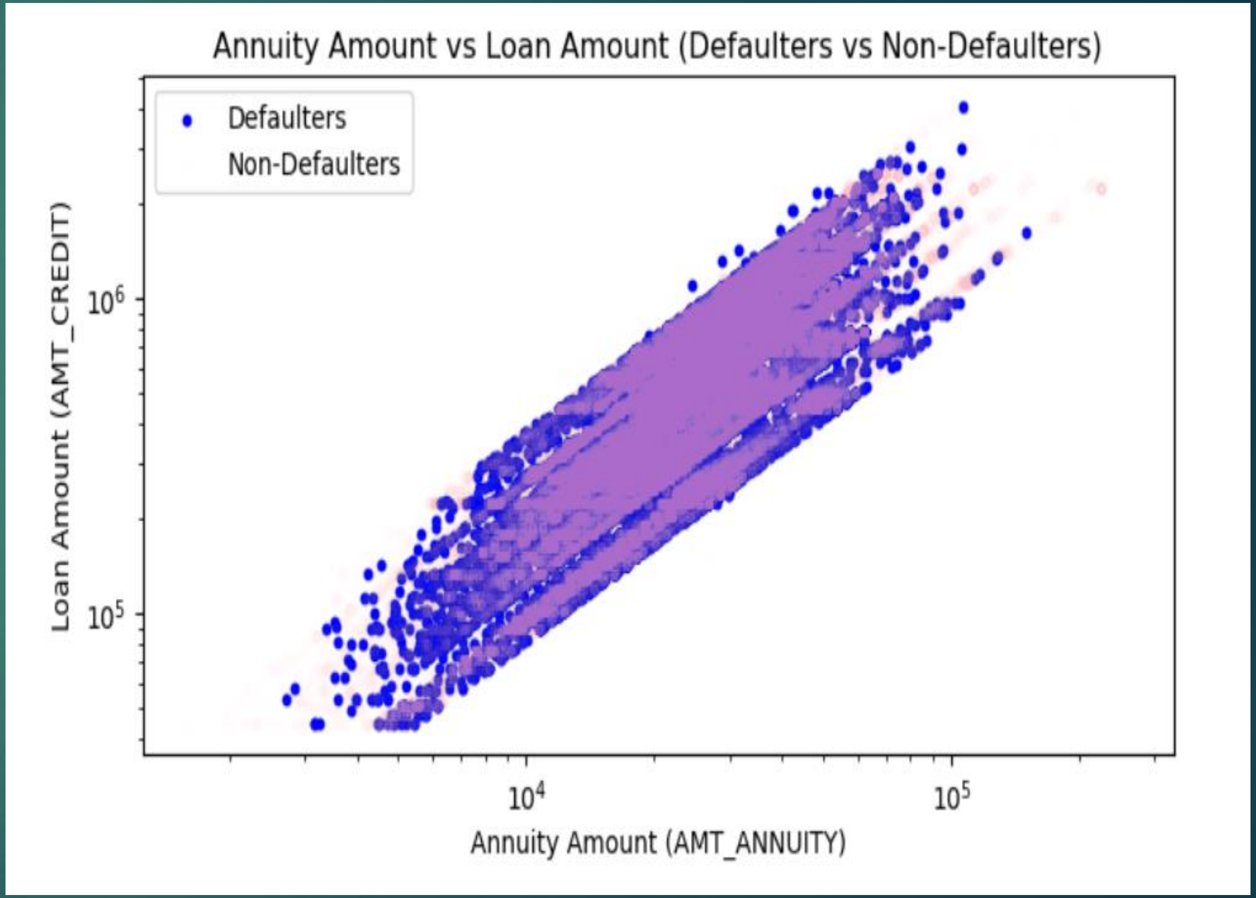
- **Age Correlation**: DAYS_BIRTH has a moderate positive correlation with DAYS_EMPLOYED, indicating older clients tend to have longer employment histories.

- **External Sources**: EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3 have low to moderate positive correlations, possibly representing independent risk factors.



Relationship between Annuity Amount and Loan Amount for Defaulters vs. Non-Defaulters:

- **Strong Correlation**: Loan amount and annuity show a clear positive correlation for both groups.
- **Overlap**: Significant overlap between defaulters and non-defaulters; these features alone don't clearly separate the two groups.
- **Range**: Non-defaulters tend to have slightly higher loan and annuity amounts compared to defaulters.

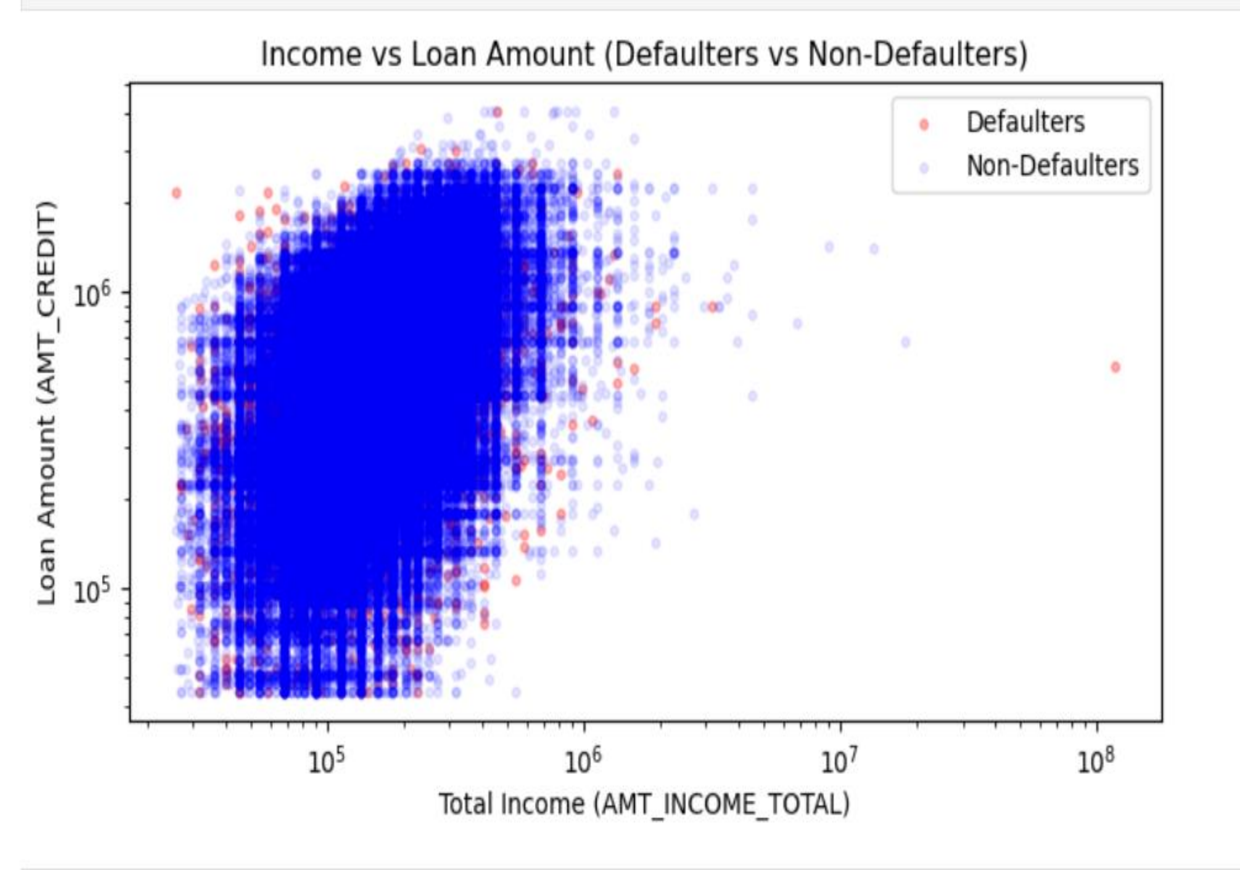


Income VS Loan Amount

•**Significant Overlap**: Hard to distinguish defaulters from non-defaulters based on income and loan amount.

•**Broad Income Range**: Both groups span a wide income range.

•**Dense Cluster**: Most applicants are in lower income and loan ranges.



Recommendations

- **Tighten Criteria** for younger, high-risk applicants.
- **Higher Interest** for low-income or high loan applicants.
- **Incentivize Low-Risk** profiles (stable jobs, higher education) with better rates.
- **Use External Scores** to refine risk segmentation.
- **Investigate High Loans** among non-defaulter accountants.
- **Monitor Loan-Goods Price** correlation to ensure affordability.


THANK YOU!