CREDIT EDA CASE STUDY

Financial Risk Assessment – Loan Application

PRESENTER – ISHAN SRIVASTAVA

UP GRAD & IIITB-DATA SCIENCE PROGRAM-SEPTEMBER 2024 SA

Content Index

Introduction

1. **Objective**:
    a. Conduct an Exploratory Data Analysis (EDA) to identify patterns and key factors influencing loan defaults.
    b. Provide actionable insights to financial institutions for effective risk assessment and decision-making.
2. **Significance**:
    a. High default rates can lead to significant financial losses for lenders.
    b. By identifying high-risk applicants, lenders can mitigate losses and optimize loan portfolios.
3. **Methodology**:
    a. Analyze applicant demographic, financial, and loan-specific attributes.
    b. Perform data cleaning, univariate, and bivariate analysis to derive actionable insights.

Dataset Overview

1. **Source**: Loan application data from a consumer finance company via UpGrad.
2. **Data Set Attributes**: columns_description.csv
3. **Demographics**: Gender, Family Status, Housing Type, Education , Occupation.
4. **Loan Details**: Loan Type, Contract Status, Payment Difficulties.
5. **Financials**: Income Type, Annuity Amount, Credit Amount,etc.
6. **Dataset :** application_data.csv , previous_application.csv
7. **Target Variable**: *Defaulters (Payment Difficulties)* vs. *Non-Defaulters (Timely Payments)*.
8. **Outliers**: Identified and analyzed for key numerical columns.
9. **Data Imbalance**: Non-defaulters significantly outnumber defaulters.
10. **Challenges :** Missing /Error Values, Wrong Inputs in columns.

Problem Statement

1. **Objective**:
    a. Develop a deeper understanding of loan defaults by analyzing applicant attributes.
    b. Identify key factors (e.g., demographics, financial metrics) that are strong indicators of risk.
2. **Challenges for Lenders**:
    a. Rejected Good Applicants: Loss of business due to over-cautious risk assessment.
    b. Approved Risky Applicants: Financial losses due to defaults.
3. **Business Needs**:
    a. Ensure reliable applicants are not rejected.
    b. Minimize losses by identifying and managing risky applicants.
4. **Expected Outcome**:
    a. Insights into driver variables for defaults.
    b. Strategies for reducing default risk while optimizing loan approval rates.

Data Cleaning and Preprocessing - 1

1. **Objective**: Prepare the dataset for analysis by addressing quality issues.
2. **Steps Taken**:

   A) Handling Missing Values:

   i. Columns like OCCUPATION_TYPE had significant missing values. So we can't drop this column because this plays a vital role in analysis ,so created a new category in this column called "Missing".

   ii. Applied imputation strategies such as mode replacement or adding a new category for categorical variables.

   B) Removing Redundant Data:

   i. Dropped columns with excessive missing values (e.g., >40% , except Occupation Type) and finally we get best 38 columns out of 122 for EDA.

   ii. Some columns were also removed by own understanding that we don't need these column for better understanding.

   iii. Missing values were replaced by Mean() / Mean() in some Continuous Numerical and Categorical columns like - EXT_SOURCE_3, EXT_SOURCE_2 , AMT_GOODS_PRICE , ATM_ANNUITY , AMT_REQ_CREDIT_BUREAU.

   iv. There were also some correction in Gender columns , XNA replace with Femal as the number of female gender is more so we replace it Female.

   v. After doing some observation in data frame we can see that there is some negative values in 'DAYS_BIRTH','DAYS_EMPLOYED','DAYS_ID_PUBLISH' ,Days Can't be negative we have to deal with it by changing it to absolute value.

Data Cleaning and Preprocessing - 2

C) Outlier Detection:

i. Outliers were retained to understand their impact on loan defaults.

ii. Here we are managing outliers by capping and flooring method

iii. Identified outliers in numerical attributes like AMT_INCOME_TOTAL and AMT_CREDIT using boxplots and IQR method.

D) Data Transformation:

i. Encoded categorical variables (e.g., NAME_INCOME_TYPE) for better interpretability.

ii. Converted dates (days / 365) to extract useful features like age.

E) Data Imbalance:

i. Identified imbalance between defaulters and non-defaulters (e.g., defaulters < 10%).

Exploratory Data Analysis (EDA)

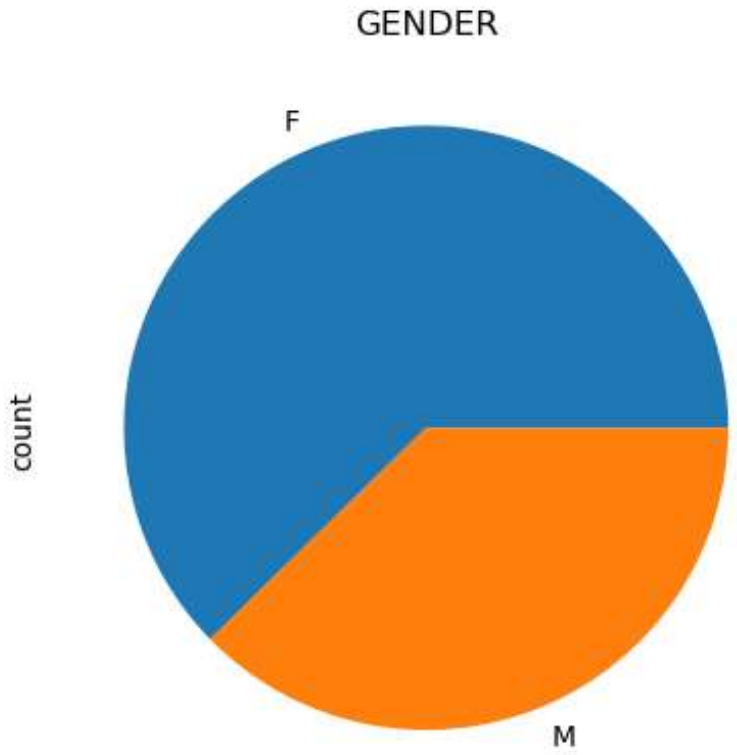
1. **Univariate Analysis:**
    a. Objective: Analyze individual attributes to identify distributions and trends.
    b. Examples : *Working* and *Commercial Associate* dominate loan applications , Loan Contract Types: Majority are *Cash Loans* , Gender: Higher loan applications from *Females*.
2. **Bivariate Analysis:**
    a. Objective: Explore relationships between features and the target variable.
    b. Examples: Credit Amount vs. Family Status: Married applicants tend to have higher loan amounts , Income Type vs. Default Rate: Applicants with *Pensioner* status have lower default rates , Housing Type vs. Target: *Co-op Apartments* have a higher proportion of defaults.
3. **Data Imbalance Observations:**
    a. Majority of records are from non-defaulters.
    b. Visualized using bar charts and pie charts , line plot , KDE and Heatmaps for correlations and bivariate relationships.

Univariate Analysis

Univariate Analysis: Gender Distribution

Observations:

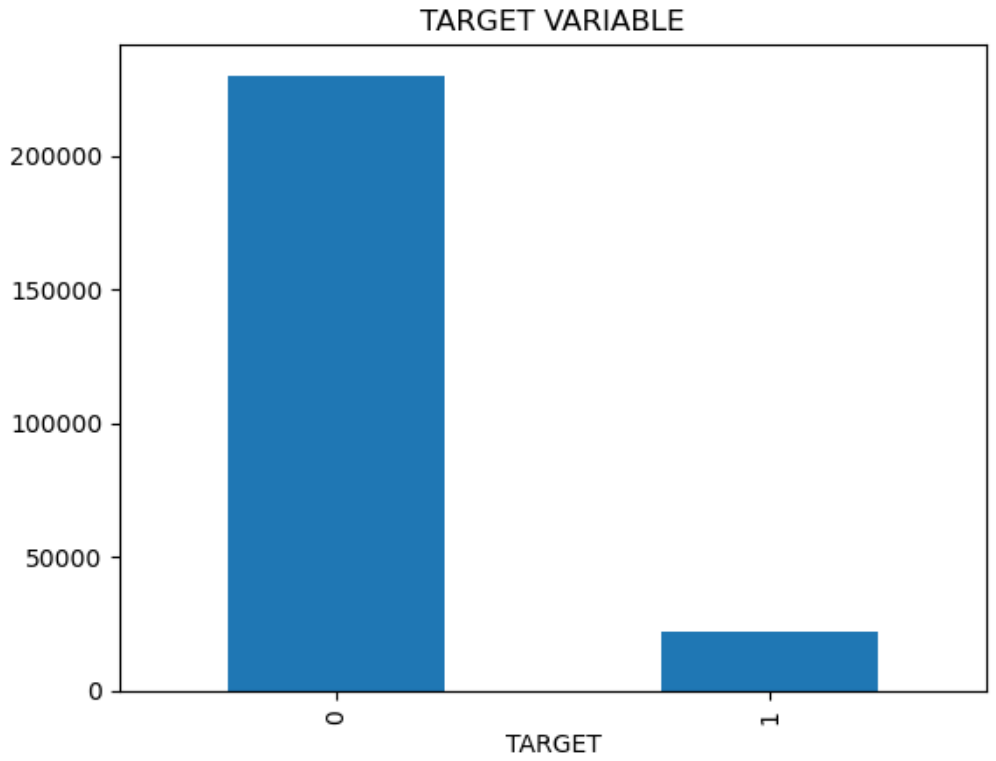
1. **Female Dominance:** Female applicants significantly outnumber male applicants.
2. **Potential Factors:** Societal changes, economic empowerment, or product-specific trends might be influencing this gender disparity.



Univariate Analysis: Target Variable

Observations:

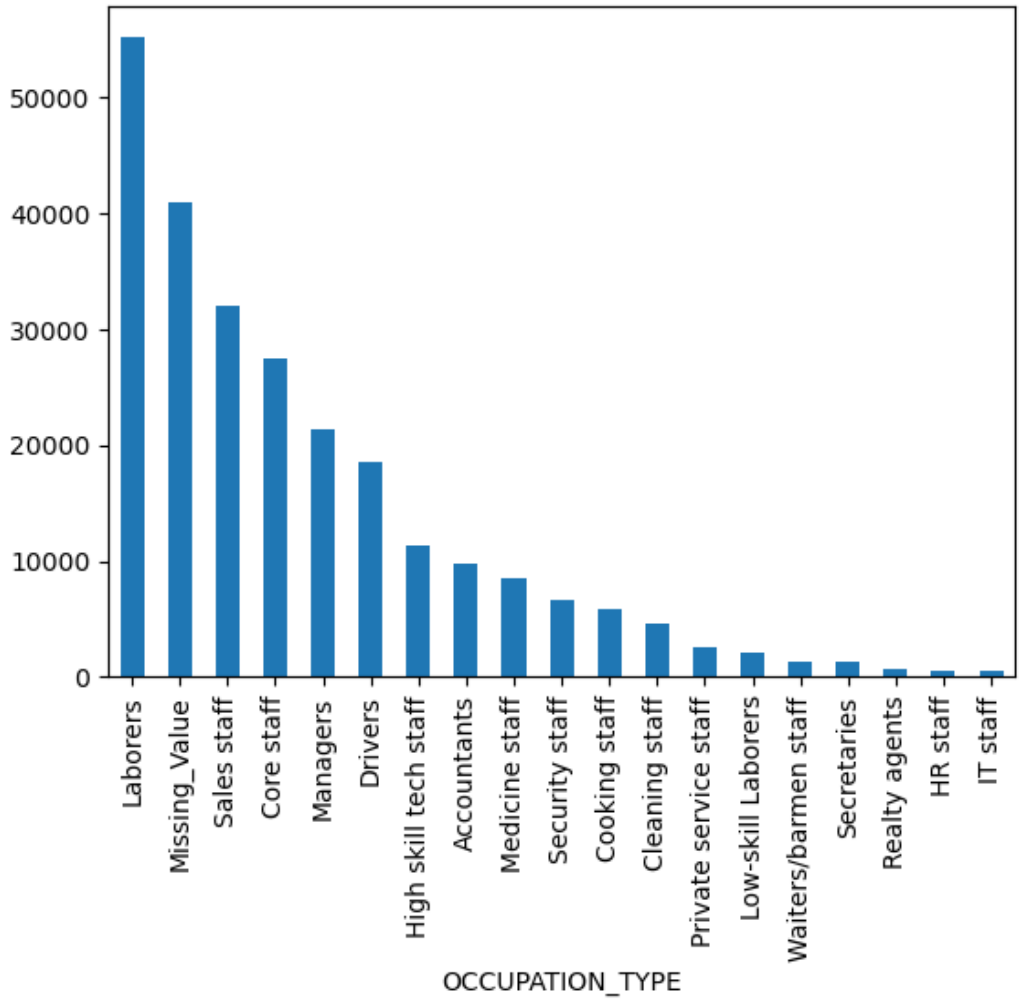
1. **Low Default Rate:** Less than 10% of applicants are classified as defaulters.
2. **Majority Creditworthy:** Most loan applications are approved to creditworthy individuals.
3. **Risk Management:** Despite the low default rate, understanding the factors contributing to default risk is essential for effective risk management.



Univariate Analysis: Occupation Type

Observations:

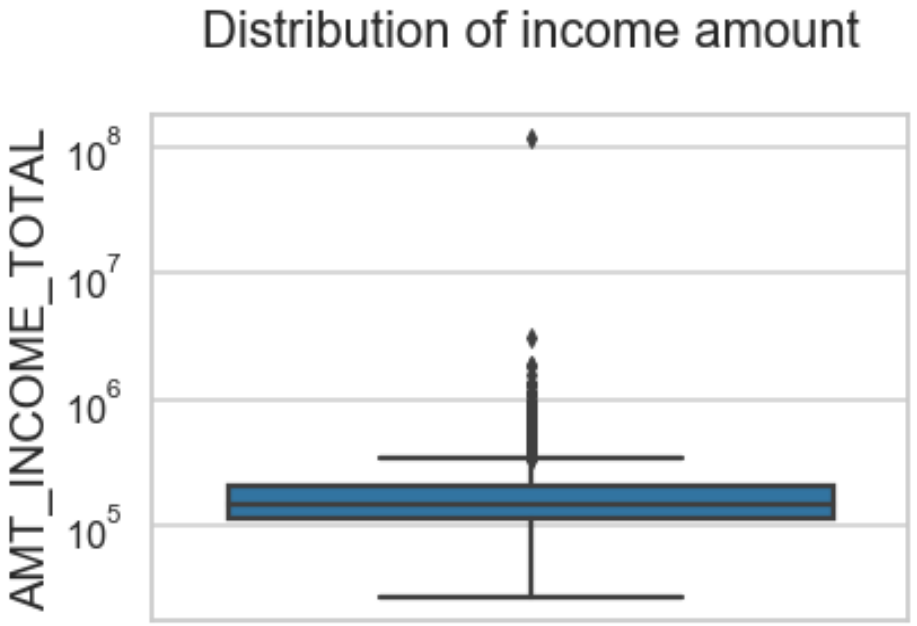
1. **Dominant Occupations:** Laborers, Sales staff, and Core staff are the most prevalent occupations among applicants.
2. **Underrepresented High-Skill Jobs:** High-skill occupations like IT staff, Medicine staff, and Accountants are less common.
3. **Missing Occupation Data:** A significant portion of data lacks occupation information, which might impact analysis and modeling.
4. **Low-Skill Dominance:** Low-skill jobs like Laborers, Drivers, and Cleaning staff constitute a substantial portion of the applicant pool.



Univariate Analysis: Income Amount

Observations:

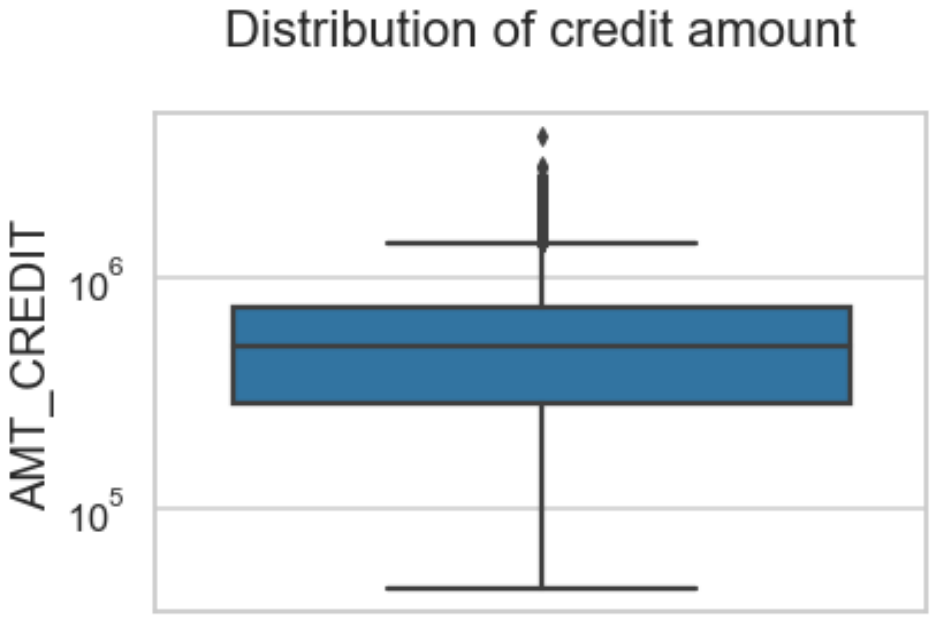
1. Presence of outliers suggests some high-income individuals.
2. The third quartiles is very slim for income amount.
3. income distribution is indicating a significant number of low-income individuals.
4. Majority of individuals have lower incomes.



Univariate Analysis: Credit amount

Observations:

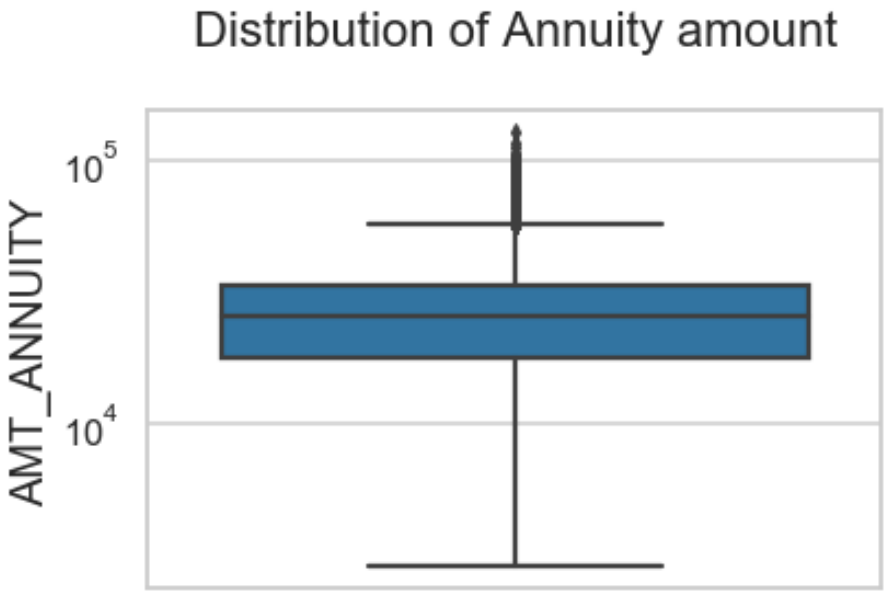
1. The credit amount distribution is indicating a significant number of smaller loans.
2. The presence of outliers suggests some very large loan amounts.
3. The majority of loans are concentrated in the lower range, with fewer larger loans.

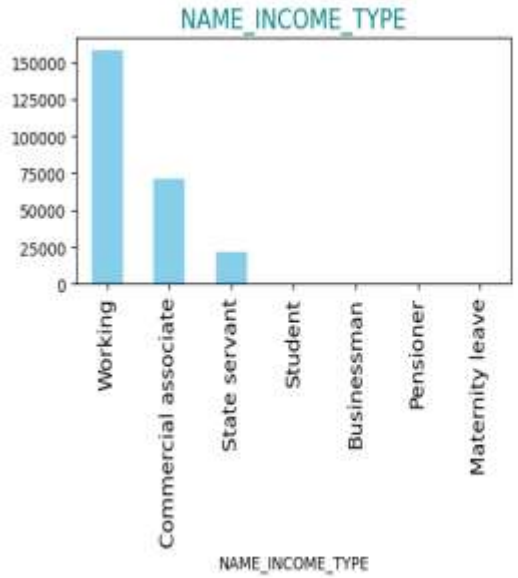


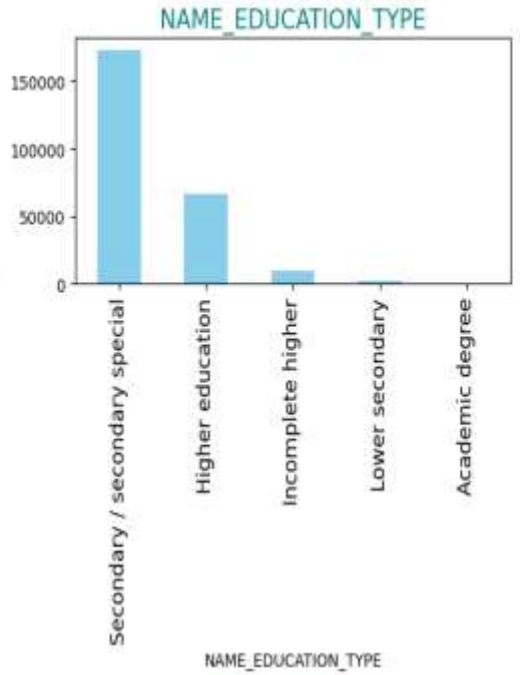
Univariate Analysis: Annuity amount

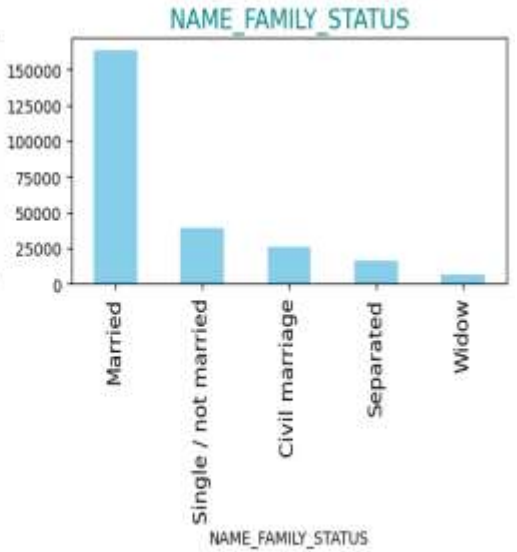
Observations:

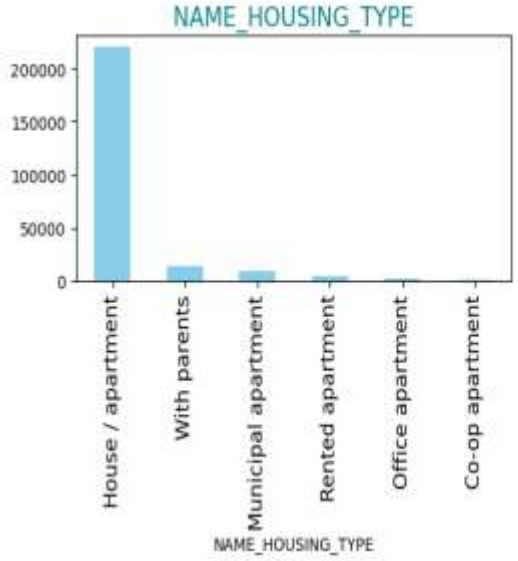
1. Some outliers are noticed in annuity amount, Indicating a significant number of lower annuity amounts.
2. The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile. The majority of clients have lower annuity amounts.

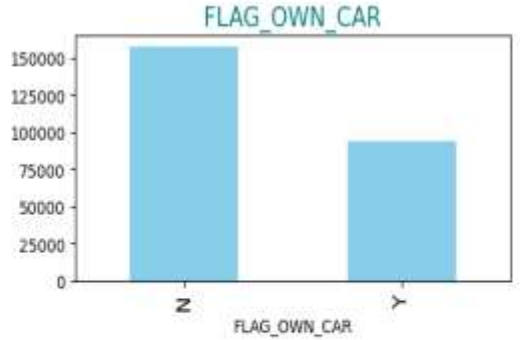


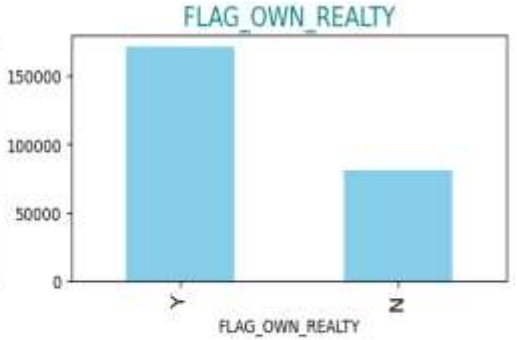












Univariate Analysis: Categorical Features

Univariate Analysis: Categorical Features –1

Observations:

1. **Income Type:** Most applicants are 'Working', followed by 'Commercial associate'.
2. **Education Type:** A significant portion of applicants have 'Secondary / secondary special' education.
3. **Family Status:** 'Married' is the most common family status among applicants.
4. **Housing Type:** Most applicants live in 'House / apartment'.
5. **Car Ownership:** A majority of applicants do not own a car.
6. **Real Estate Ownership:** A majority of applicants own real estate.
7. **Contract Type:** 'Cash loans' are the most common type of loan applied for.

Univariate Analysis: Categorical Features –2

Insights:

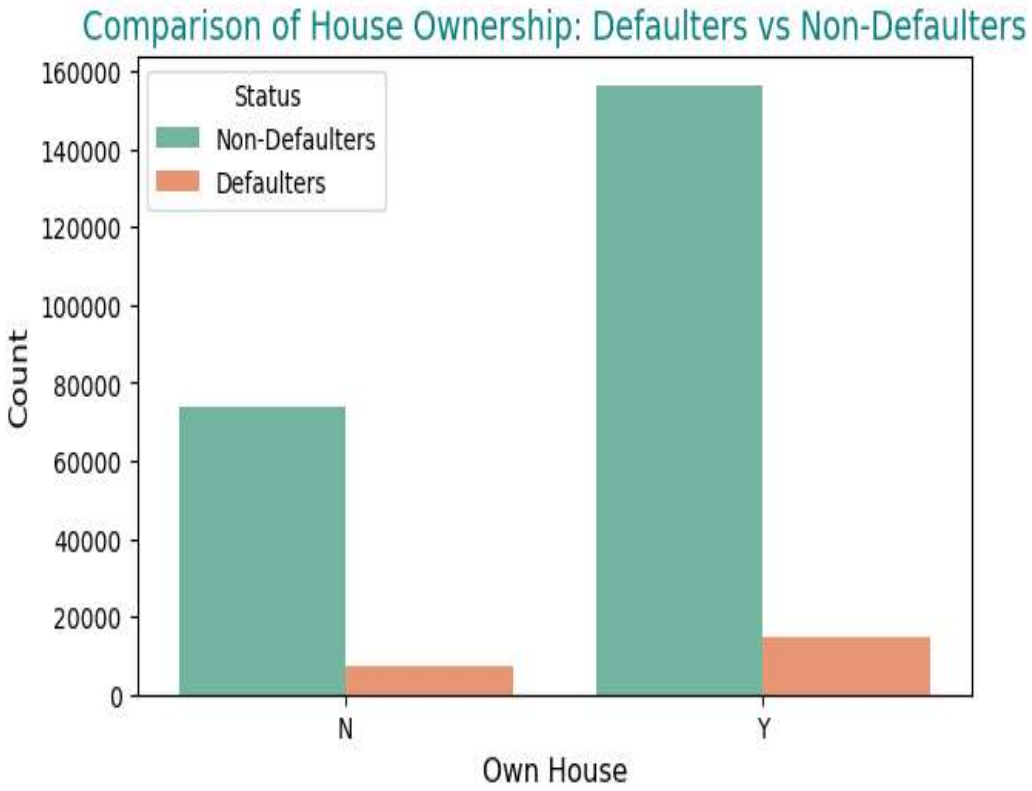
1. **Target Audience:** The analysis suggests that the majority of applicants are working individuals with secondary education and are married.
2. **Product Focus:** Cash loans seem to be the most popular product.
3. **Marketing Strategy:** Targeting individuals with specific housing types and car ownership can be effective.

Bivariate Analysis

Bivariate Analysis: House Ownership

Key Observations

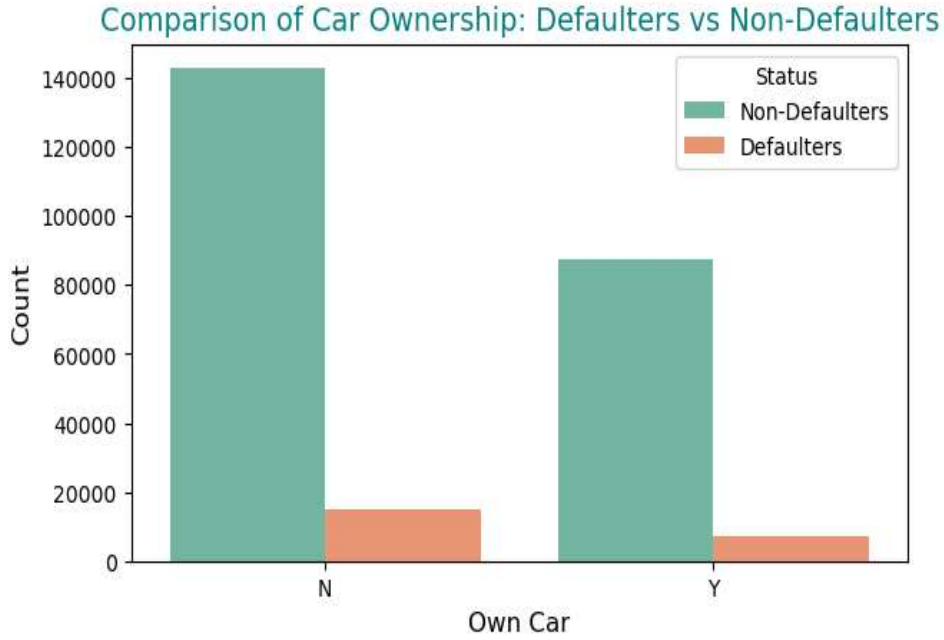
1. **Majority Own Property**: Most defaulters and non-defaulters own property.
2. **Similar Distribution**: Ownership patterns are alike for both groups.
3. **Implications**
4. Owning a house alone is not a strong predictor of default; factors like income and debt-to-income ratio are likely more impactful.



Bivariate Analysis: Car Ownership

Key Observations

1. **Majority Doesn't Own a Car**: Most defaulters and non-defaulters lack car ownership.
2. **Similar Distribution**: Car ownership patterns are consistent across both groups.
3. **Implications :** Car ownership is not a strong predictor of default; factors like income and debt-to-income ratio may have more influence.



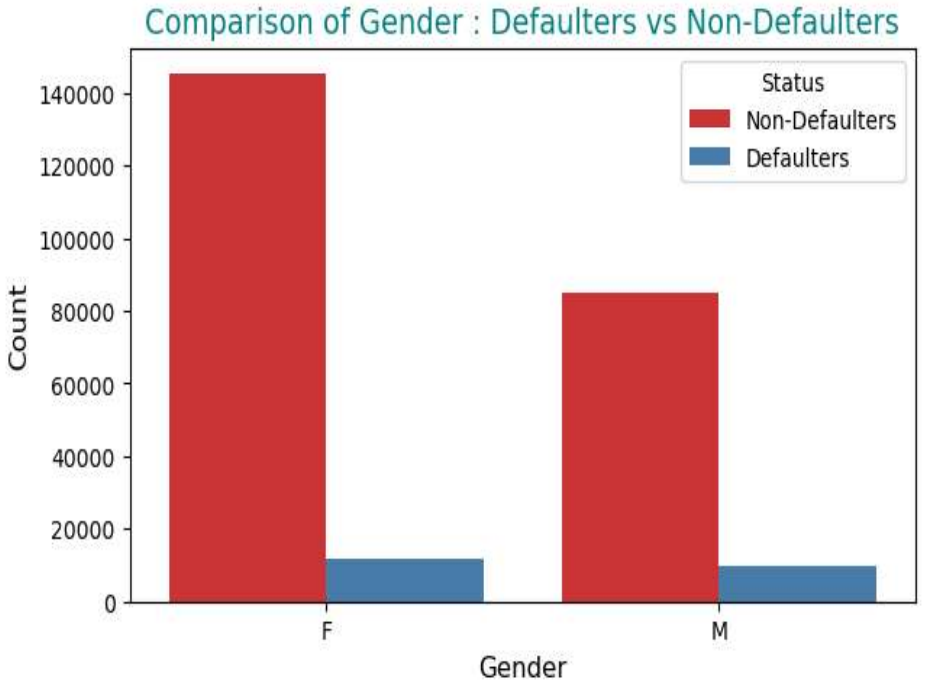
Bivariate Analysis: Gender

Key Observations

1. **Female Dominance**: Most loan applications come from females.
2. **Higher Female Default Rate**: Females show a higher default proportion despite being the majority.

Implications

1. Gender influences credit risk but must be assessed alongside other factors like income and employment.
2. Tailored strategies may help address challenges faced by female borrowers.



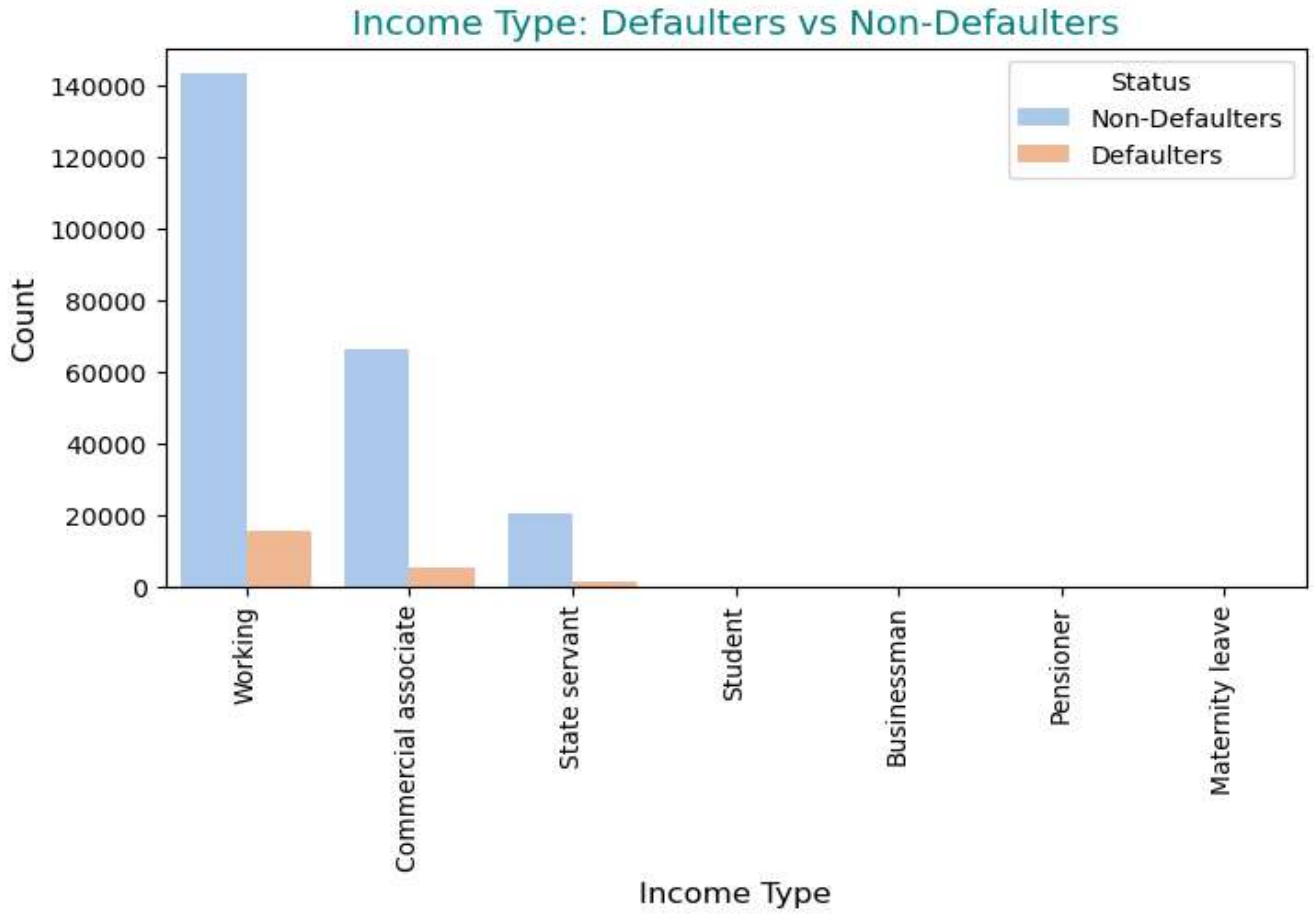
Bivariate Analysis: Income Type

Key Observations

1. **Dominant Group**: "Working" is the most common income type among defaulters and non-defaulters.
2. **Commercial Associates**: A significant portion belongs to this group.
3. **Other Types**: Groups like "State servant," "Student," and "Pensioner" are less common.

Implications

1. Assess financial stability of "Working" individuals and Commercial Associates.
2. Address unique challenges of less common income types.



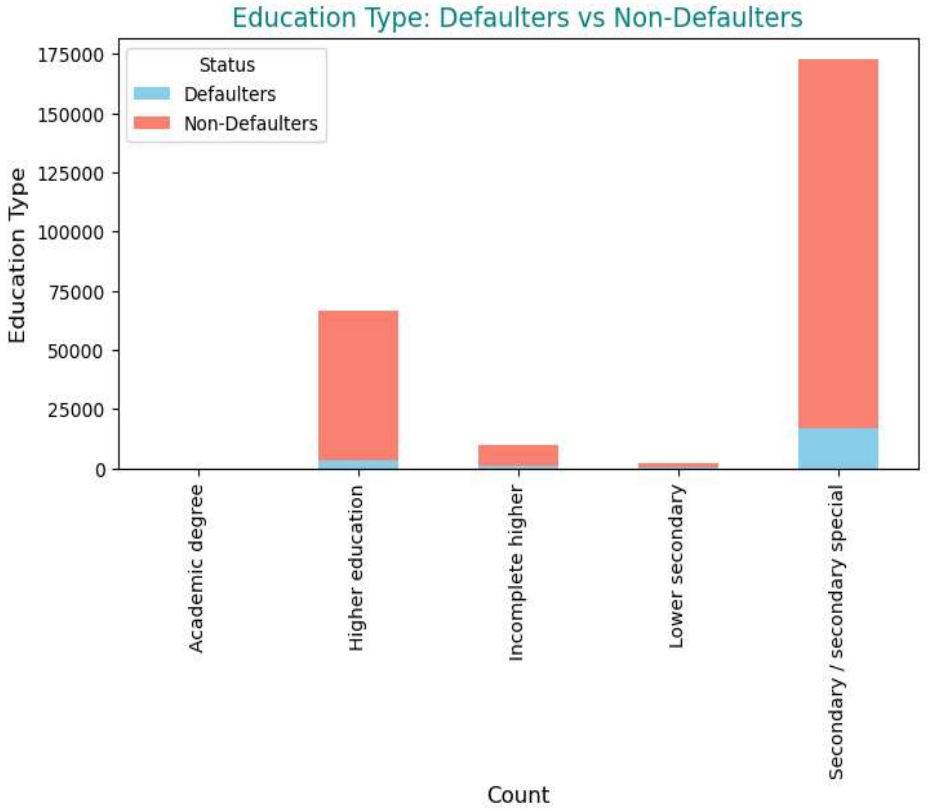
Bivariate Analysis: Education Type

key Observation:

1. **Secondary/Secondary Special Education:** A significant portion of both defaulters and non-defaulters have this level of education. However, a higher proportion of defaulters fall into this category.

Implications:

1. **Risk Assessment:** Individuals with Secondary/Secondary Special education might pose a higher risk of default.
2. **Targeted Financial Literacy:** Implementing financial literacy programs for this group could help reduce default rates.
3. **Product Tailoring:** Customized loan products with flexible repayment terms and lower interest rates could be considered for this segment.



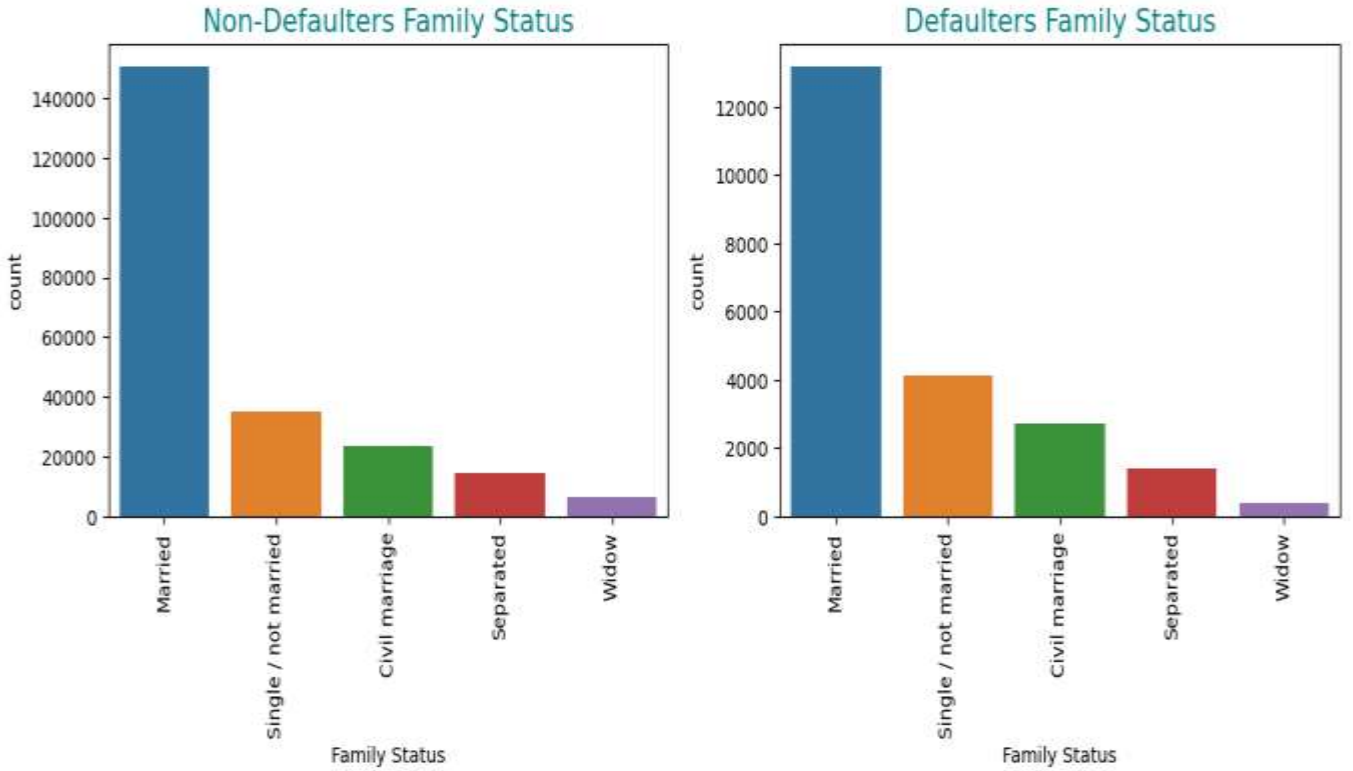
Bivariate Analysis: Family Status

Key Observations:

1. **Married Dominance:** Married individuals form the largest group among both defaulters and non-defaulters.
2. **Other Statuses:** Other family statuses like "Single / not married," "Civil marriage," "Separated," and "Widow" are less common.

Implications:

1. **Marital Status and Risk:** While marriage might be a positive indicator, it's not a definitive factor in predicting default risk.
2. **Other Factors:** Factors like income, employment, and loan amount are likely more influential.



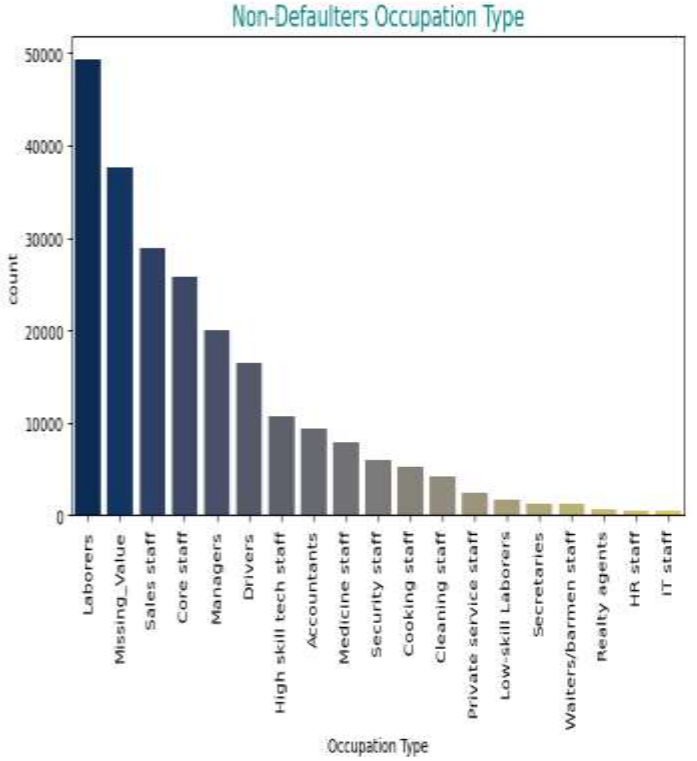
Bivariate Analysis: Occupation Type and Default Status

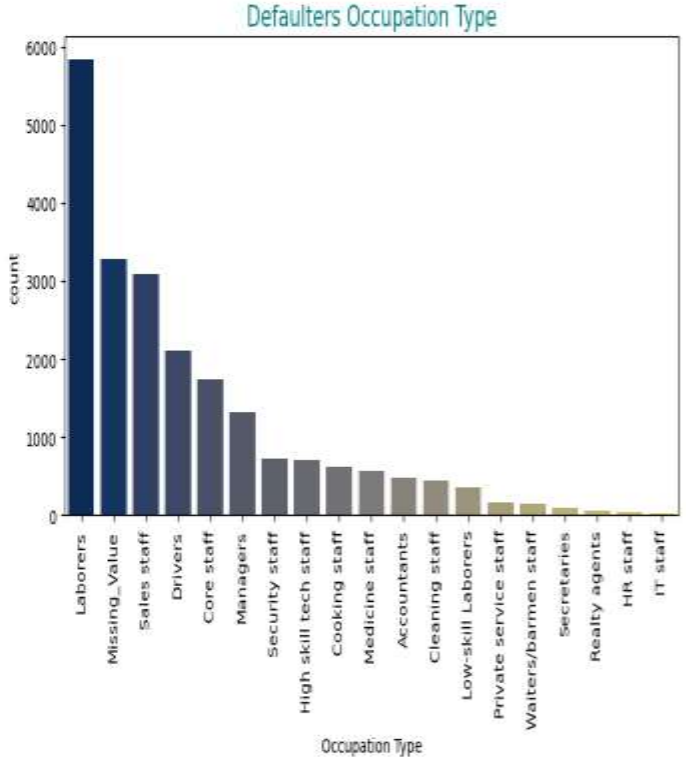
Key Observation:

1. Low-Skill Dominance: A significant proportion of defaulters belong to low-skill occupations like Laborers, Drivers, and Cleaning staff.
2. High-Skill Underrepresentation: High-skill occupations have a lower representation among defaulters.

Implications:

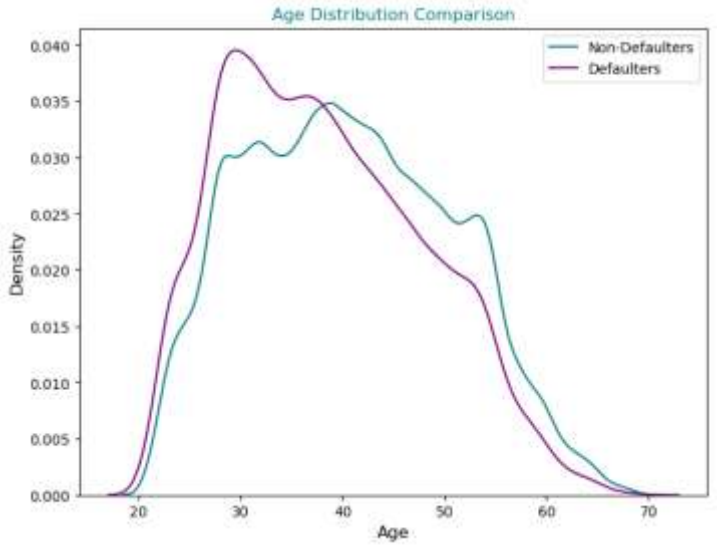
1. Targeted Risk Assessment: Lenders should consider occupation as a factor in assessing creditworthiness, especially for low-skill jobs.
2. Financial Literacy: Targeted financial literacy programs for low-income individuals can help improve their financial management skills and reduce default risk.





Bivariate Analysis: Age and Default Risk

1. **Middle-aged Defaulters:** A significant portion of loan defaulters fall within the age group of 25-45 years.
2. **Age Distribution Similarity:** Both defaulters and non-defaulters exhibit similar age distribution patterns.



Multivariate Analysis

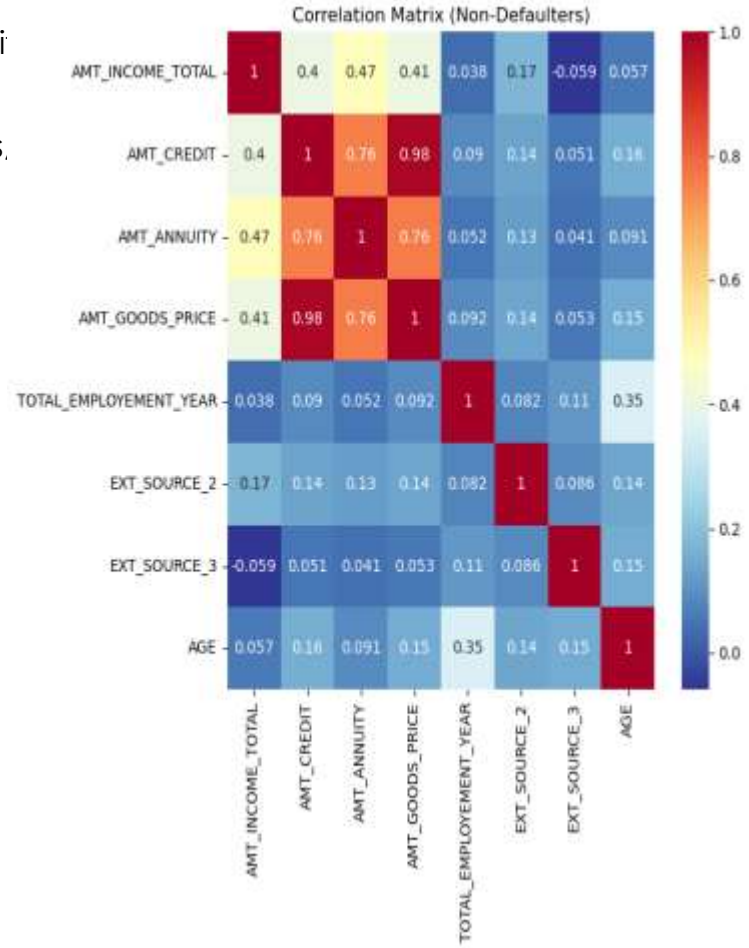
Multivariate Analysis: Correlation Matrix

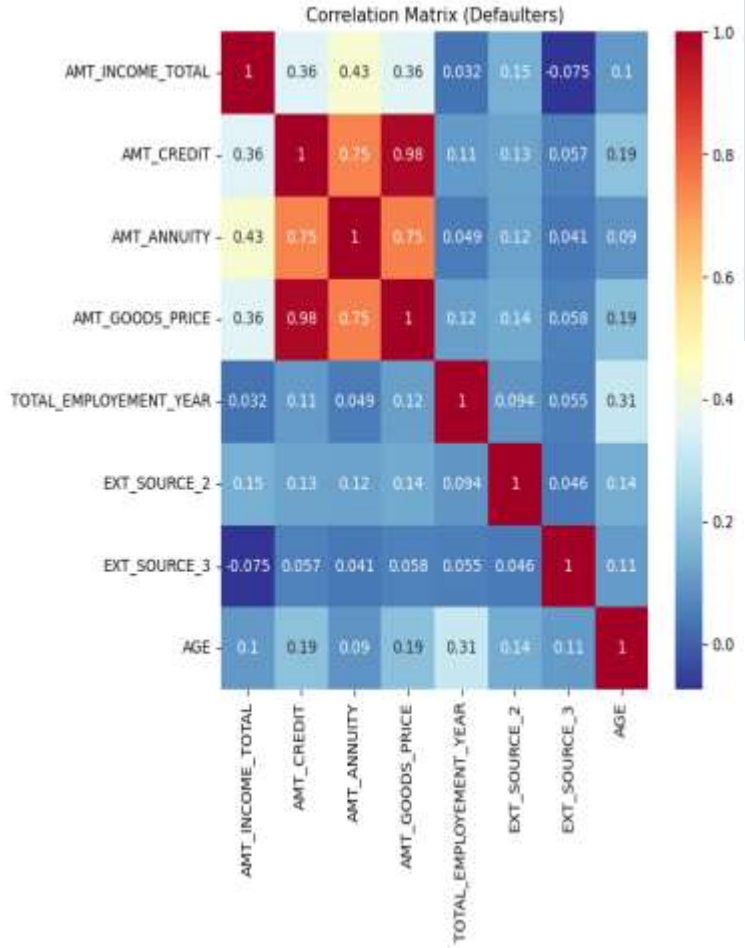
Key Observations:

1. **Strong Correlations:** Both defaulters and non-defaulters exhibi strong positive correlations between AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE. This indicates a strong relationship between loan amount, monthly payments, and the cost of goods.
2. **Age and Employment:** A moderate positive correlation between Age and TOTAL_EMPLOYMENT_YEAR suggests that older individuals tend to have longer employment histories.
3. **External Sources:** The EXT_SOURCE variables show moderate correlations with other features, indicating their potential influence on creditworthiness.

Implications:

1. **Feature Importance:** These highly correlated features can be important predictors of default risk.
2. **Multicollinearity:** High correlation between features can lead to multicollinearity issues in modeling, which might impact model performance.
3. **Feature Engineering:** Creating new features or combining existing ones can help mitigate multicollinearity and improve model performance.





Pair Plot Analysis: Key Observations

1. Income and Credit:

A. **Positive Correlation:** There's a strong positive correlation between AMT_INCOME_TOTAL and AMT_CREDIT. This indicates that individuals with higher incomes tend to receive larger loans.

B. **Outliers:** A few outliers are visible, suggesting some individuals with high incomes but relatively low loan amounts or vice versa.

2. Credit and Annuity:

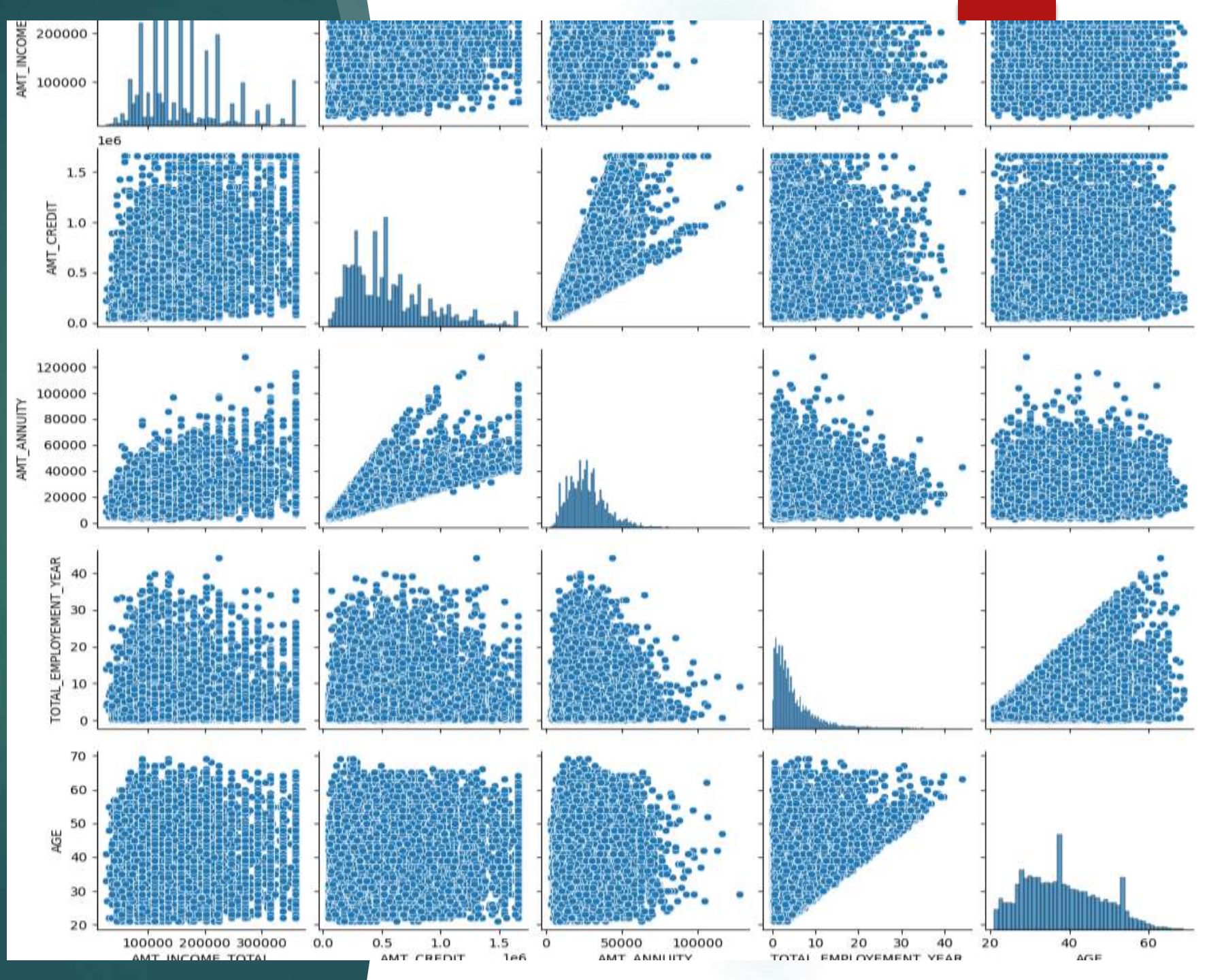
A. **Strong Positive Correlation:** A strong positive correlation exists between AMT_CREDIT and AMT_ANNUITY. This suggests that larger loans are associated with higher monthly payments.

3. Age and Employment:

A. **Positive Correlation:** There's a positive correlation between AGE and TOTAL_EMPLOYMENT_YEAR, indicating that older individuals tend to have longer employment histories.

4. Variable Distributions: Several variables, such as AMT_INCOME_TOTAL, AMT_CREDIT, and AMT_ANNUITY, exhibit right-skewed distributions, suggesting a larger number of lower-value observations.

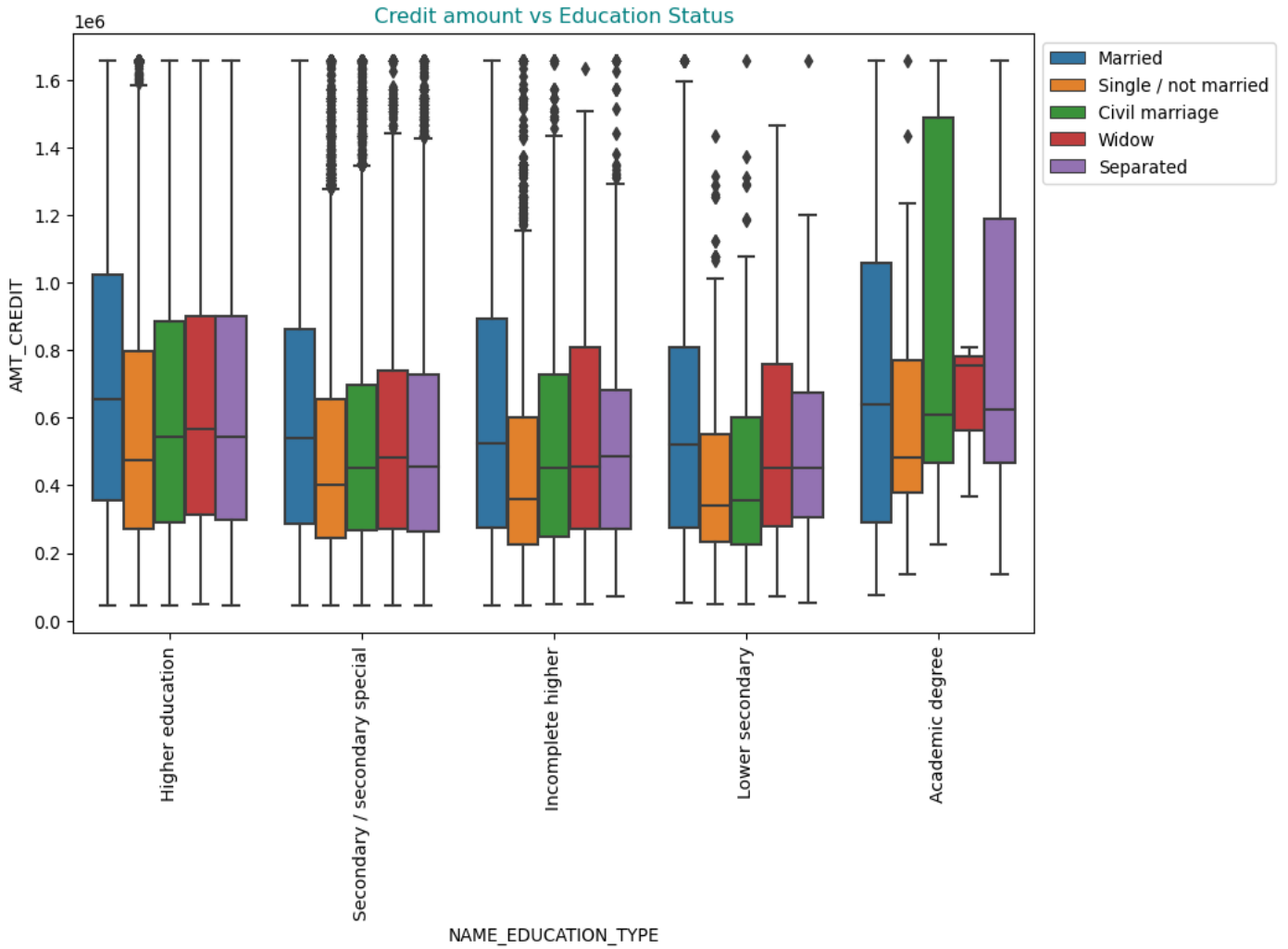
Pair Plot Analysis: Key Observations



Multivariate Analysis: Education and Family Status vs Loan Amount

Key Observations:

1. **Education and Loan Amount:** Individuals with higher education levels (Academic degree, Higher education) tend to receive larger loans compared to those with lower education levels.
2. **Family Status and Loan Amount:** Married individuals generally have higher loan amounts, followed by those in civil marriages. Single/not married individuals have relatively lower loan amounts.
3. **Outliers:** Some individuals, particularly in higher education groups, have significantly higher loan amounts, which might be due to factors like higher income, specific loan purposes, or lower risk profiles.



Multivariate Analysis: Credit Amount vs Real Estate and Car Ownership

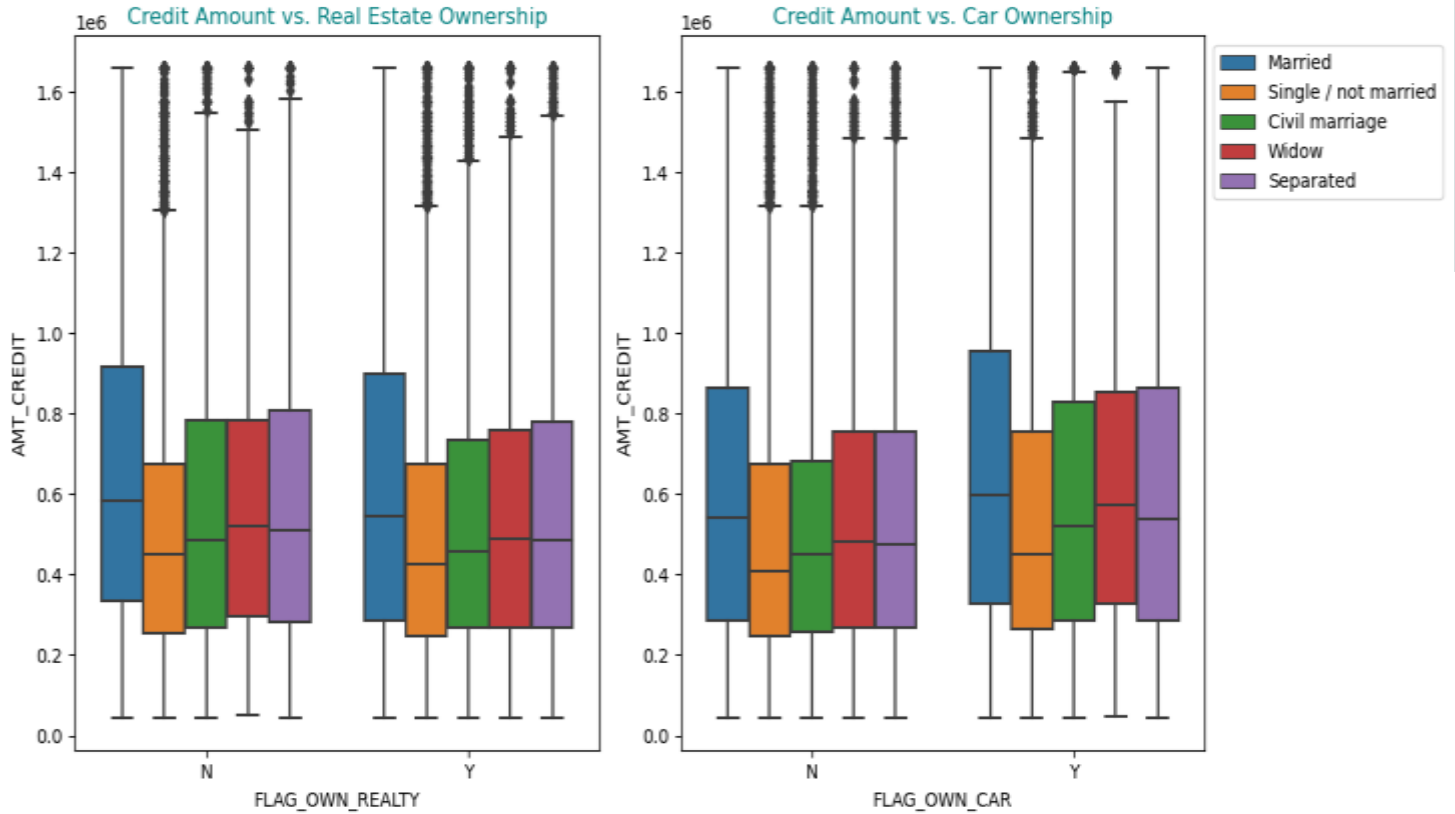
Key Observations:

Real Estate Ownership:

1. **Higher Median Loan Amount:** Individuals owning real estate tend to have slightly higher median loan amounts.
2. **Similar Distribution:** The overall distribution of loan amounts is similar for both real estate owners and non-owners.

Car Ownership:

1. **Less Significant Impact:** Car ownership seems to have a less pronounced impact on loan amounts compared to real estate ownership.
2. **Similar Distribution:** The distribution of loan amounts is relatively similar for car owners and non-owners.



Previous Application Dataset Analysis

Analyzing Categorical Features

(Plot is on next slide.)

Key Observations: Previous Appplications Summary

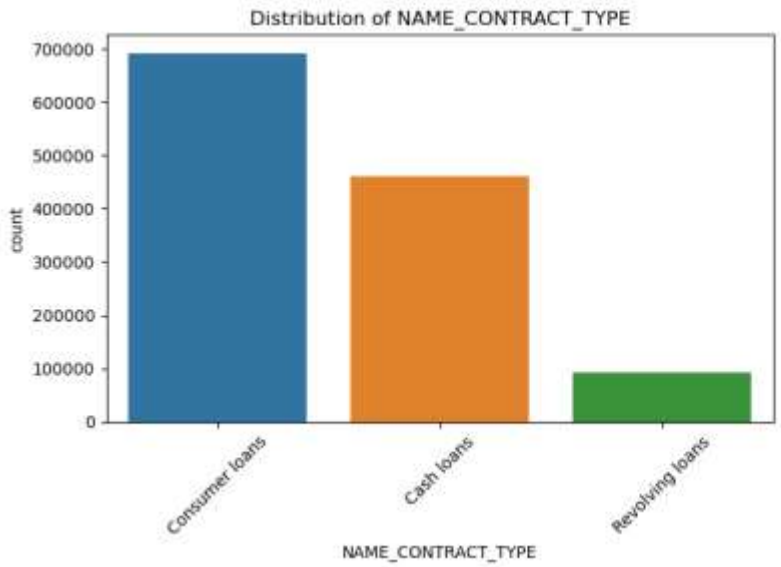
Loan Types: Consumer loans are the most prevalent, followed by cash loans and revolving loans.

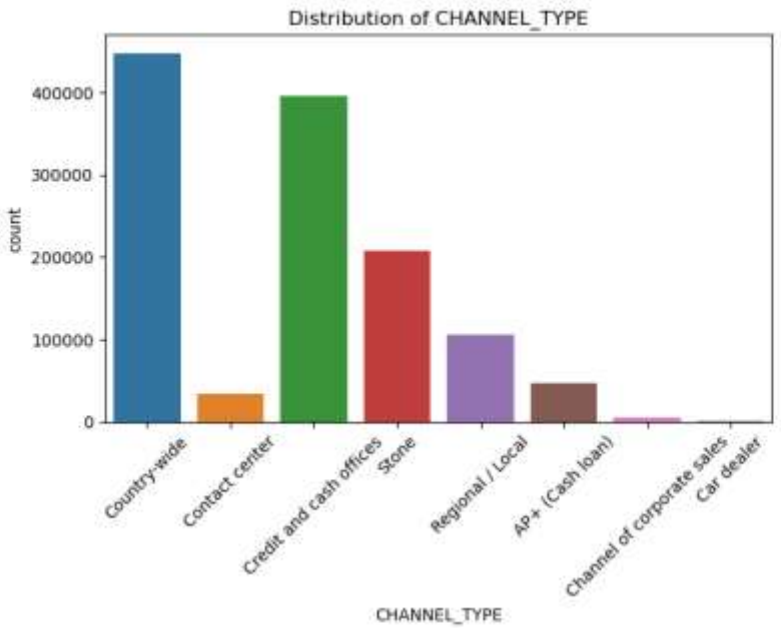
Application Channels: Most applications are received through contact centers, AP+Cash loans, and car dealers.

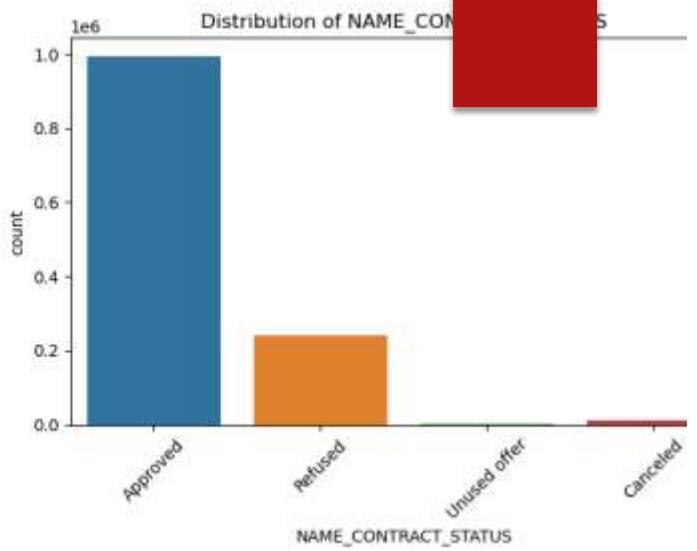
Loan Status: A majority of loan applications are approved, with a smaller proportion being refused or canceled.

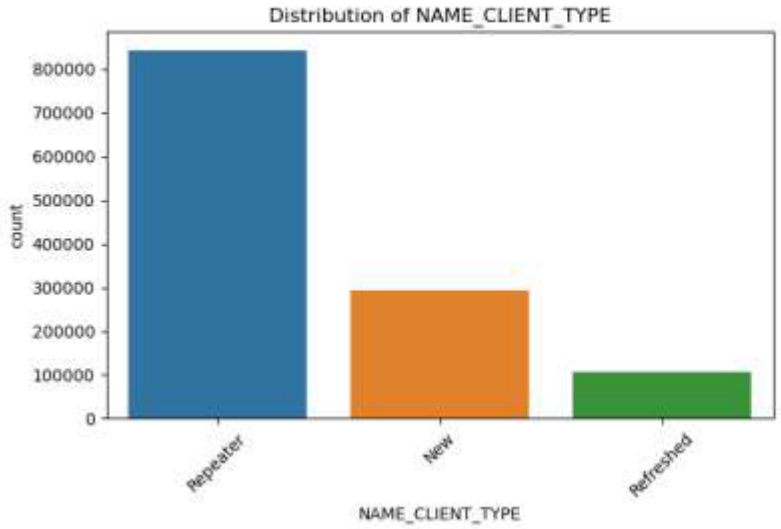
Client Type: Repeat customers form a significant portion of the applicant base.

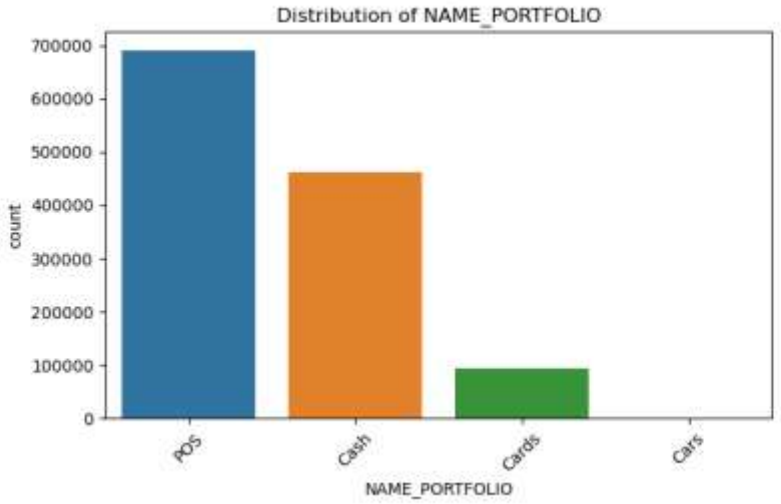
Loan Portfolio: POS loans dominate, followed by cash loans and credit cards.

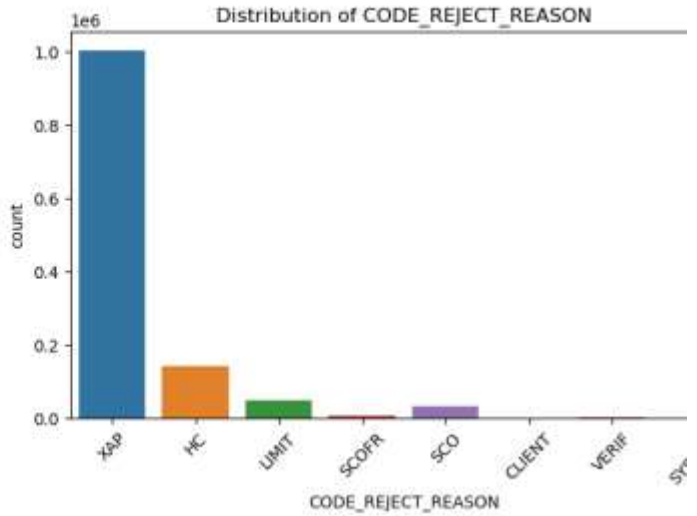

Distribution of NAME_CONTRACT_TYPE
count
700000
600000
500000
400000
300000
200000
100000
0
Consumer loans
Cash loans
Revolving loans
NAME_CONTRACT_TYPE


Distribution of CHANNEL_TYPE
count
400000
300000
200000
100000
0
Country-wide
Contact center
Credit and cash offices
Stone
Regional / Local
AP+ (Cash loan)
Channel of corporate sales
Car dealer
CHANNEL_TYPE


1e6
Distribution of NAME_CON
S
count
1.0
0.8
0.6
0.4
0.2
0.0
Approved
Refused
Unused offer
Canceled
NAME_CONTRACT_STATUS


Distribution of NAME_CLIENT_TYPE
count
800000
700000
600000
500000
400000
300000
200000
100000
0
Repeater
New
Refreshed
NAME_CLIENT_TYPE


Distribution of NAME_PORTFOLIO
count
700000
600000
500000
400000
300000
200000
100000
0
POS
Cash
Cards
Cars
NAME_PORTFOLIO


1e6
Distribution of CODE_REJECT_REASON
count
1.0
0.8
0.6
0.4
0.2
0.0
XAP
HC
LIMIT
SCOFR
SCO
CLIENT
VERIF
CODE_REJECT_REASON

Univariate Analysis: Categorical Features on Merged dataset

(Plot is on next slide.)

Education Type

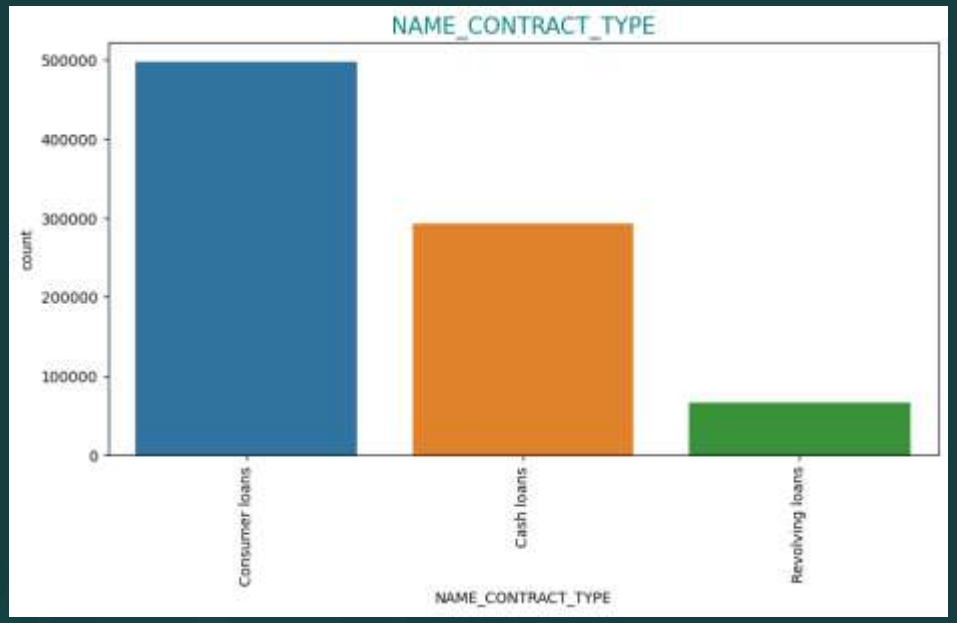
1. **Dominant Level:** "Secondary/Secondary Special" is the most common education level among applicants.
2. **Lower Levels:** "Higher education" and "Academic degree" are less common.
3. **Implications:** Understanding the financial capabilities and risk profiles of different education levels is crucial for effective lending decisions.

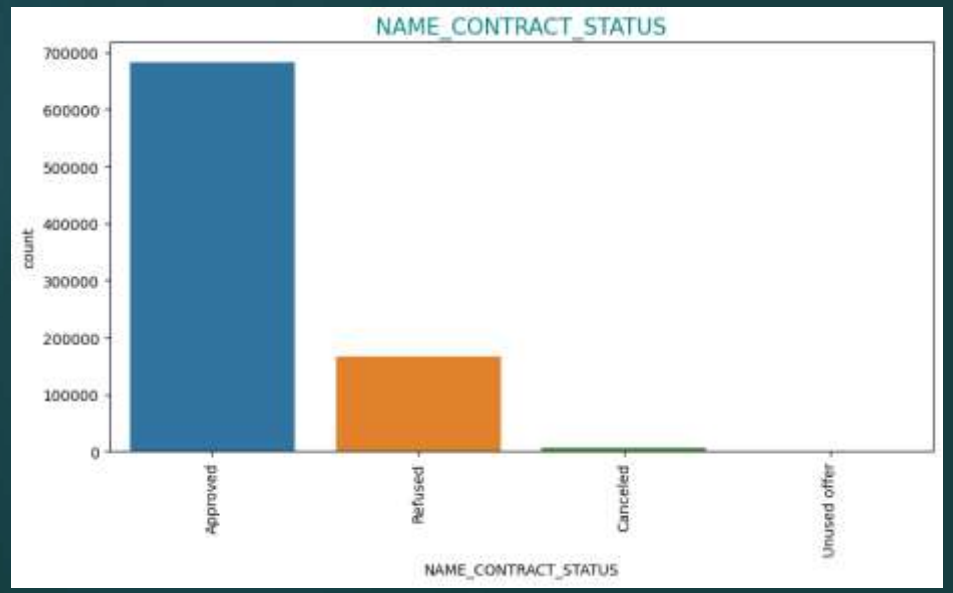
Contract Status

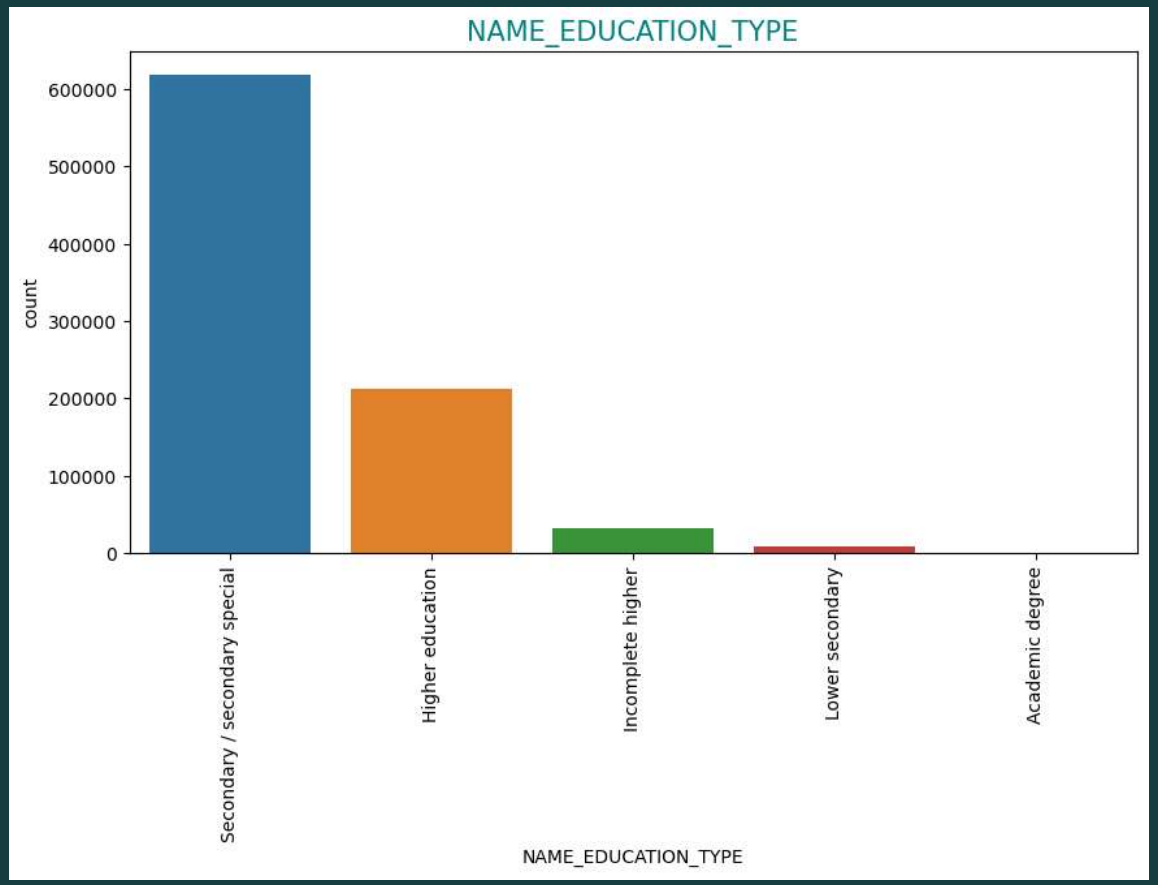
1. **Approved Loans:** A significant majority of loan applications are approved.
2. **Refused and Canceled Loans:** A smaller portion of applications are refused or canceled.
3. **Unused Offers:** A very small number of offers remain unused.

Contract Type

1. **Consumer Loans:** Consumer loans are the most prevalent type of loan applied for.
2. **Cash Loans and Revolving Loans:** These two types of loans are less common compared to consumer loans.


NAME_CONTRACT_TYPE
500000
400000
300000
200000
100000
0
count
Consumer loans
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

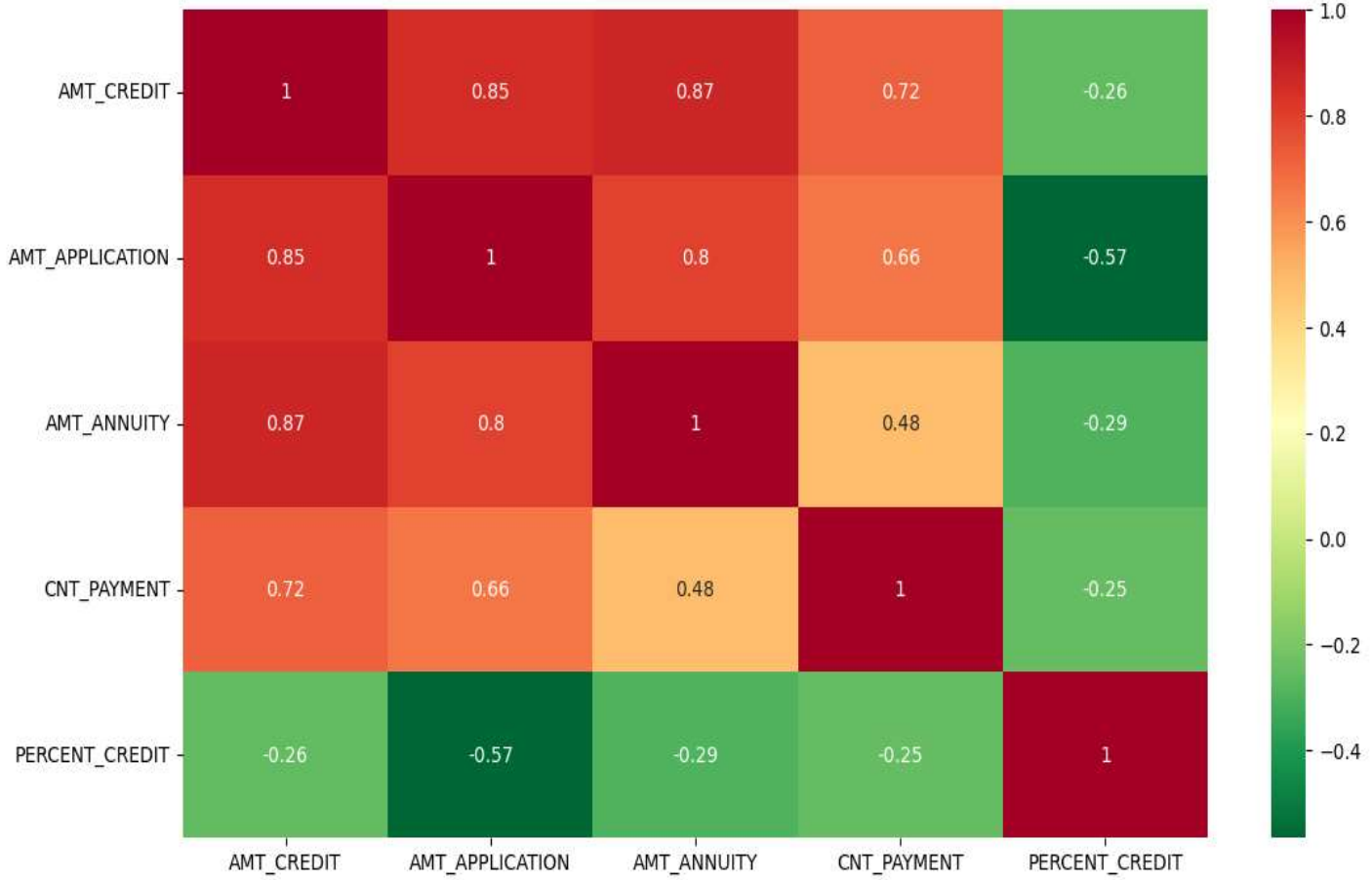

NAME_CONTRACT_STATUS
700000
600000
500000
400000
300000
200000
100000
0
count
Approved
Refused
Canceled
Unused offer
NAME_CONTRACT_STATUS


NAME_EDUCATION_TYPE
600000
500000
400000
300000
200000
100000
0
count
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

Multivariate Analysis: Correlation Heatmap

Key Observations:

1. **Strong Positive Correlations:**
    1. AMT_CREDIT, AMT_APPLICATION, and AMT_ANNUITY are highly correlated, indicating that larger loan amounts are often associated with higher monthly payments and application amounts.
    2. CNT_PAYMENT (number of payments) is also positively correlated with these variables, suggesting that larger loans are typically repaid over longer periods.
2. **Negative Correlation with PERCENT_CREDIT:**
    1. PERCENT_CREDIT (credit amount as a percentage of application amount) is negatively correlated with other variables. This might indicate that larger loans tend to have lower percentages of the total application amount.



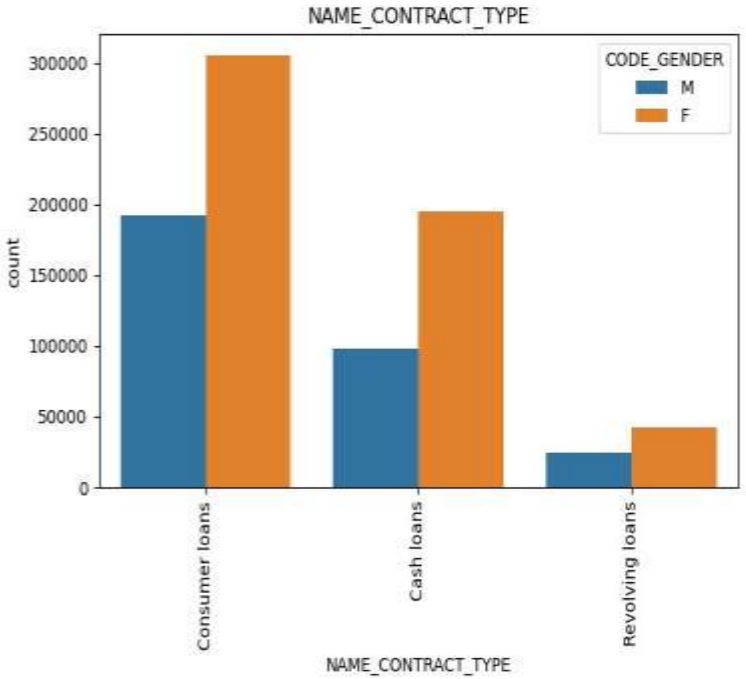
Multivariate Analysis: Gender and Categorical Features

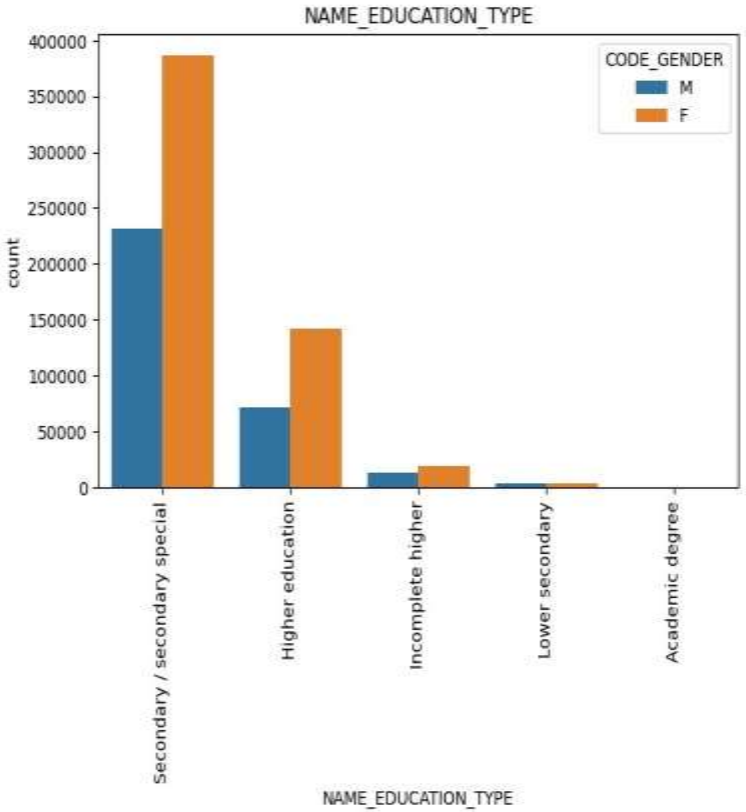
Key Observations:

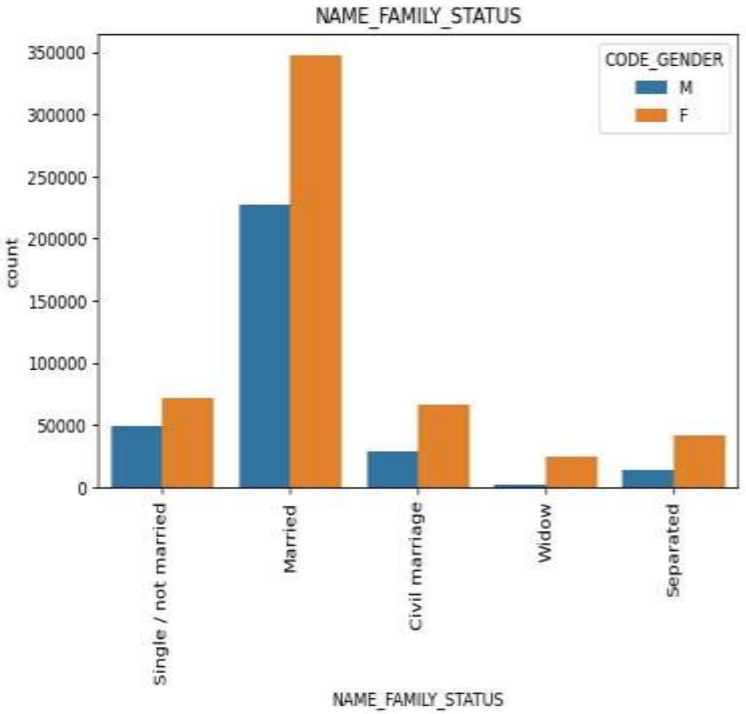
1. **Contract Type:** Females tend to opt for consumer loans more frequently than males.
2. **Education Level:** A higher proportion of females have a secondary/secondary special education level compared to males.
3. **Family Status:** The distribution of family statuses is relatively similar between males and females, with married individuals being the most common.

Implications:

1. **Gender-Based Product Preferences:** Understanding gender-specific preferences can help tailor product offerings and marketing strategies.
2. **Risk Assessment:** Analyzing default rates for different gender groups can help identify potential differences in risk profiles.
3. **Financial Literacy:** Targeted financial literacy programs can be designed to address specific needs and challenges faced by different gender groups.


NAME_CONTRACT_TYPE
CODE_GENDER
M
F
count
300000
250000
200000
150000
100000
50000
0
Consumer loans
Cash loans
Revolving loans
NAME_CONTRACT_TYPE


NAME_EDUCATION_TYPE
CODE_GENDER
M
F
count
400000
350000
300000
250000
200000
150000
100000
50000
0
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE


NAME_FAMILY_STATUS
CODE_GENDER
M
F
count
350000
300000
250000
200000
150000
100000
50000
0
Single / not married
Married
Civil marriage
Widow
Separated
NAME_FAMILY_STATUS

Conclusion

Key Findings:

1. **Demographic Factors:** Female applicants, especially those with lower education levels, and married individuals, particularly females, exhibit higher default rates.
2. **Loan Characteristics:** Consumer loans, particularly for less educated individuals, pose higher default risks.
3. **Applicant Experience:** Repeat applicants, despite a higher chance of default, also have a higher chance of non-default compared to new applicants.
4. **Occupation and Default:** Low-skilled occupations are associated with higher default rates.
5. **Age and Default:** Middle-aged individuals (25-45) are more prone to default.
6. **Income and Default:** Higher income levels are generally associated with lower default rates.
7. **Asset Ownership:** While owning a house or car can be a positive indicator, it's not a definitive factor in predicting default risk.
8. **Loan Type and Status:** Consumer loans are the most prevalent, with a majority of applications being approved.

Recommendations for Lenders

1. **Targeted Lending:** Focus on specific demographic segments with lower default risk, such as individuals with higher education levels, stable income sources, and favorable family statuses.
2. **Robust Risk Assessment**: Implement robust creditworthiness assessment models that consider a combination of factors, including income, employment, education, family status, and loan purpose.
3. **Product Diversification:** Offer a diverse range of loan products to cater to different customer needs and risk profiles.
4. **Financial Literacy Programs**: Provide financial literacy programs to help borrowers make informed decisions and manage their finances effectively.
5. **Data-Driven Decision Making:** Utilize data analytics to identify high-risk segments, optimize underwriting processes, and improve customer retention.
6. **Continuous Monitoring and Adaptation:** Regularly monitor performance metrics and adjust lending strategies to adapt to changing market conditions and customer behavior.
7. By carefully considering these insights and implementing appropriate strategies, lenders can mitigate credit risk, enhance customer satisfaction, and achieve sustainable growth.

Thankyou