EDA ASSIGNMENT

"Decoding Loan Risk with EDA: Enhancing Lending Strategies"

BY

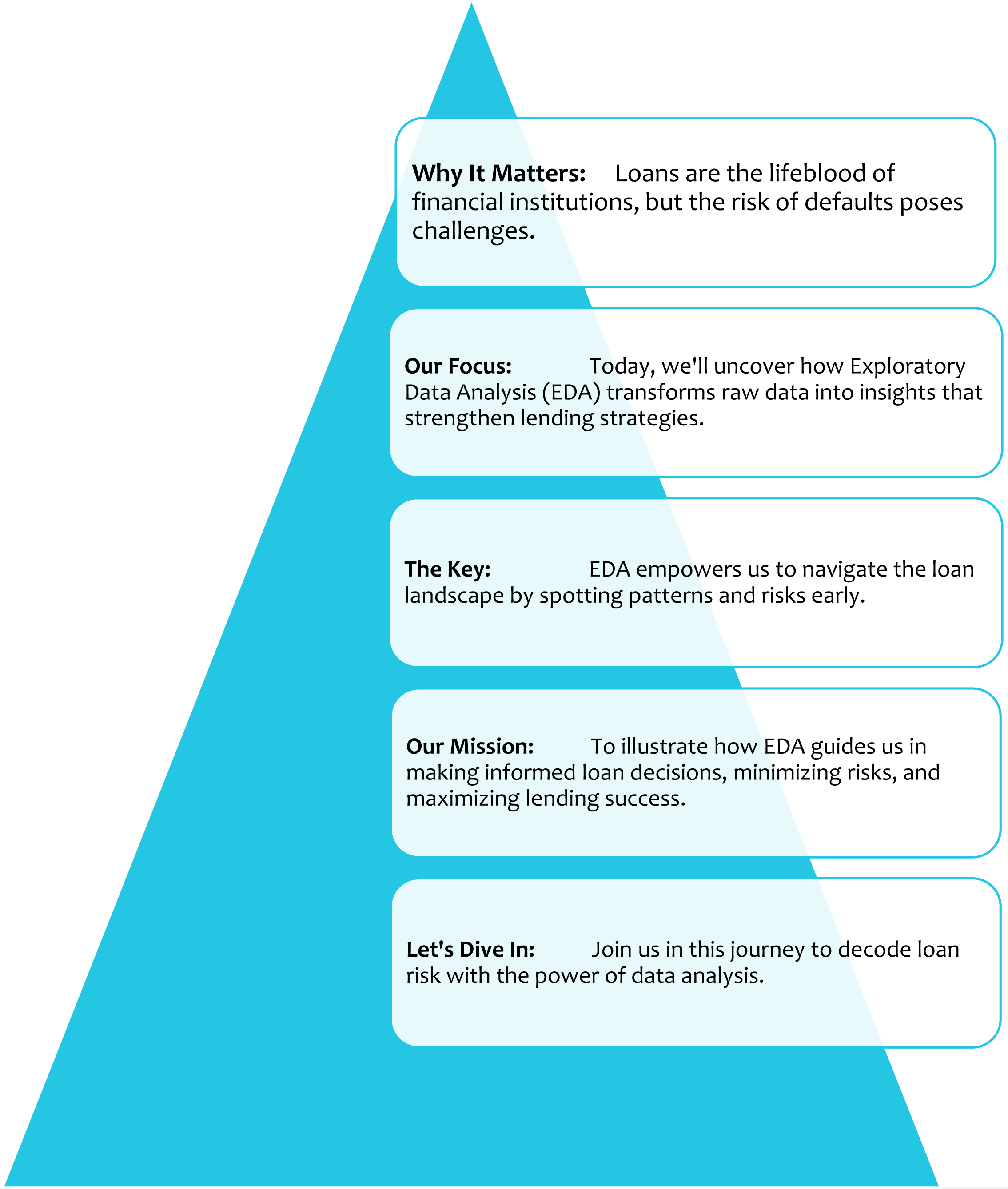
SONMOY JANA

Index - Mapping Our Presentation Journey

- Introduction
- Business Understanding
- Dataset Overview
- Data Inspection and Cleaning
- Data Standardization
- Identifying Outliers for Robust Analysis
- Univariate Analysis
- Bivariate and Multivariate Analysis
- Analyzing Combined Datasets
- Conclusion

Introduction

Empowering Lending Decisions Through Data Analysis



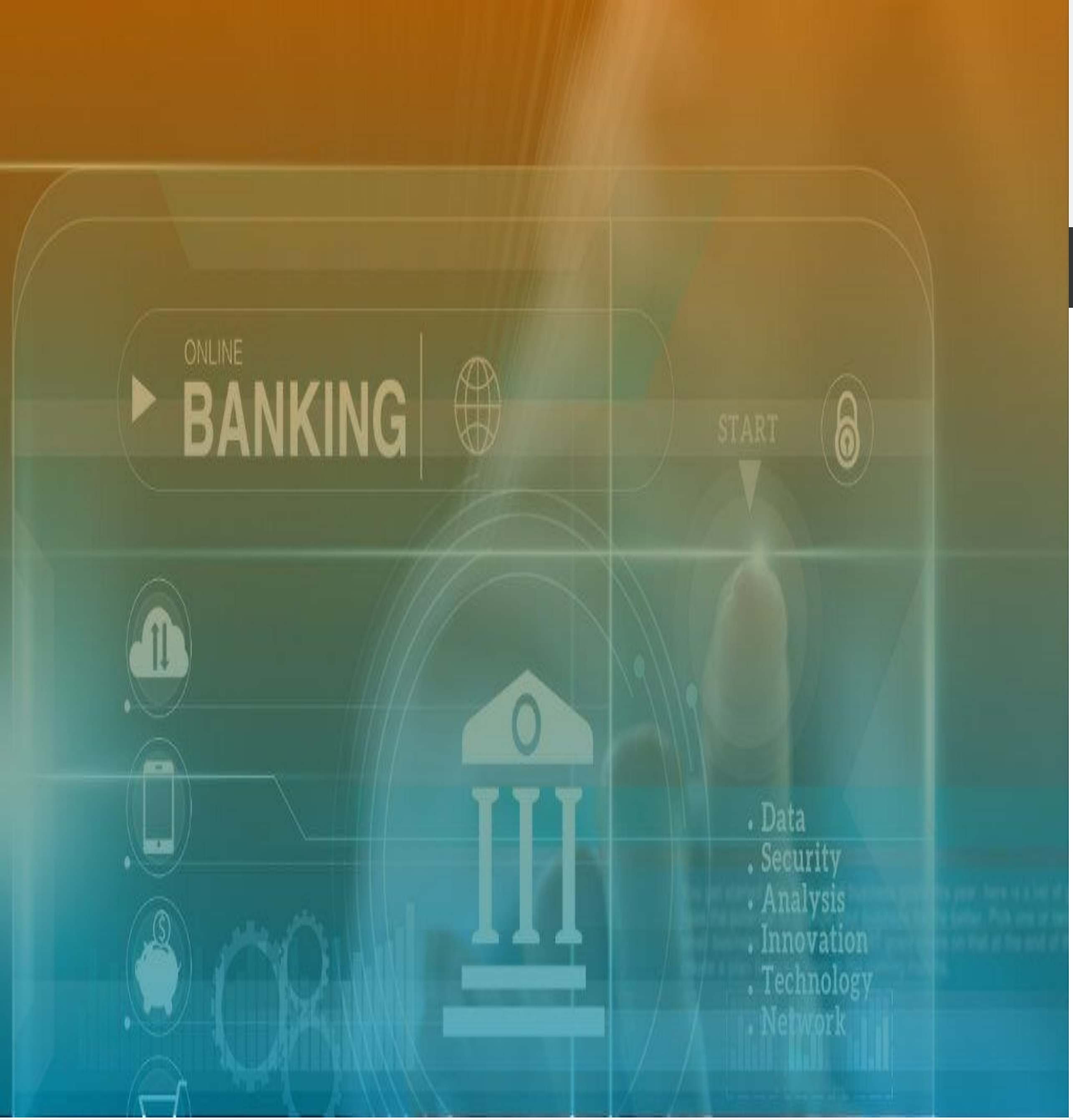
Why It Matters: Loans are the lifeblood of financial institutions, but the risk of defaults poses challenges.

Our Focus: Today, we'll uncover how Exploratory Data Analysis (EDA) transforms raw data into insights that strengthen lending strategies.

The Key: EDA empowers us to navigate the loan landscape by spotting patterns and risks early.

Our Mission: To illustrate how EDA guides us in making informed loan decisions, minimizing risks, and maximizing lending success.

Let's Dive In: Join us in this journey to decode loan risk with the power of data analysis.



Business Understanding

Unveiling Insights from Analysis

- **Lending Challenges:** Lenders often struggle to decide who gets loans due to limited credit history knowledge.
- **Finding the Right Customers:** In a vast lending landscape, identifying the most suitable customer segment to target becomes a strategic puzzle.
- **Balancing Loan Range and Interest Rates:** Ensuring loan ranges and interest rates are enticing to customers, while minimizing defaults and benefiting the company, requires careful calibration.
- **Making It Fair:** EDA helps us spot patterns, making sure we don't miss out on deserving borrowers.

Dataset Overview

Exploring Core Data: Our presentation's foundation rests on two pivotal datasets: application data and previous application.

Main Variable: The focal point of our analysis is the TARGET variable, categorizing loan applicants as defaulters or non-defaulters.

Categorical Understanding:

Contract & Gender: NAME_CONTRACT_TYPE and CODE_GENDER define loan types and applicant genders, influencing lending strategies.

Suite & Income Type: NAME_TYPE_SUITE and NAME_INCOME_TYPE provide insights into social circles and income sources, enriching customer profiling.

Education & Family: NAME_EDUCATION_TYPE and NAME_FAMILY_STATUS reflect education levels and family statuses, impacting loan eligibility.

Housing & Occupation: NAME_HOUSING_TYPE and OCCUPATION_TYPE shed light on housing preferences and job types, guiding market targeting.

Weekday & Organization: WEEKDAY_APPR_PROCESS_START and ORGANIZATION_TYPE uncover application timings and applicant industries, enhancing strategic decisions.

Dataset Overview

Numerical Insights:

Income Insights: AMT_INCOME_TOTAL offers a numerical glimpse into applicant incomes, guiding risk assessment and lending limits.

Credit & Annuity: AMT_CREDIT and AMT_ANNUITY define loan amounts and payment structures, crucial for designing sustainable loans.

Goods Price: AMT_GOODS_PRICE reflects the cost of financed goods, shaping loan structures and applicant affordability.

Strategic Approach: Our analytical journey revolves around assessing variables against the TARGET variable, enabling us to formulate ethical lending strategies that balance profitability and risk management.

Targeted Insights: By dissecting relationships between variables and the variable, we aim to extract actionable insights that drive prudent risk assessment, customer profiling, and strategic lending decisions.

Data Inspection and Cleaning

- **Inspecting and Understanding Data:**

  We initiated our analysis by examining key attributes using .shape, .info(), and .describe() to gain an overview of the dataset.

- **Data Cleaning Process:**

  To ensure data quality, we addressed missing values by calculating the percentage of null values within the dataset.

  Columns with null values exceeding 45% were removed to streamline our analysis.

- **Imputation and Transformation:**

  For columns with empty values, mode imputation was employed for all categorical attributes. In the case of numerical columns, most of the missing values were imputed using the median. However, after conducting an in-depth analysis, we opted to impute 0 for two specific columns within the numeric category. We encountered columns with negative or mixed values, such as 'DAYS_BIRTH,' 'DAYS_EMPLOYED,’ 'DAYS_REGISTRATION,' 'DAYS_ID_PUBLISH,' and 'DAYS_LAST_PHONE_CHANGE.' These were transformed to absolute values for analysis.

Data Inspection and Cleaning

- **Data Type and Value Handling:**

During our exploration, the field 'CODE_GENDER' had 'XNA' entries, indicating unavailability. We replaced these with the most common value, 'F' (female).

Similarly, 'ORGANIZATION_TYPE' and 'OCCUPATION_TYPE' had 'XNA' and null values respectively, which we replaced with 'Pensioner' based on associations with 'NAME_INCOME_TYPE'.

Data Standardization:

- **Transitioning to Years:**

  Focusing on columns representing time spans ('DAYS_BIRTH,' 'DAYS_EMPLOYED,' 'DAYS_REGISTRATION,' 'DAYS_ID_PUBLISH,' and 'DAYS_LAST_PHONE_CHANGE'), I transformed the values from days to years.

  This conversion yielded a more relatable and interpretable scale, contributing to the clarity of our analysis.

- **Binning for Analysis:**

  We've fine-tuned attribute ranges to match their characteristics for precise insights. Our approach involved thorough analyses on vital attributes like 'YEARS_AGE,' 'YEARS_EMPLOYED,' 'AMT_INCOME_TOTAL,' 'AMT_CREDIT,' 'AMT_ANNUITY,' and 'AMT_GOODS_PRICE.' By crafting purposeful ranges, we've boosted our analytical depth, unveiling valuable insights essential for informed lending strategies.

Identifying Outliers for Robust Analysis

- **Detecting Outliers in Key Attributes:**

  We concentrated on identifying potential outliers within significant attributes crucial for our analysis.

  Using box plots, we examined attributes such as 'CNT_CHILDREN,' 'AMT_INCOME_TOTAL,' 'AMT_CREDIT,' 'AMT_ANNUITY,' 'AMT_GOODS_PRICE,' and more.

- **Structured Visualization:**

  Our exploratory analysis is showcased through a structured arrangement of box plots.

  With each plot representing a specific attribute, we visually assessed the data distribution for potential outliers.

- **Key Observations:**

  EXT_SOURCE_2, YEARS_AGE, and YEARS_ID_PUBLISH: These attributes exhibit a consistent absence of outlier values, contributing to the robustness of our analysis.

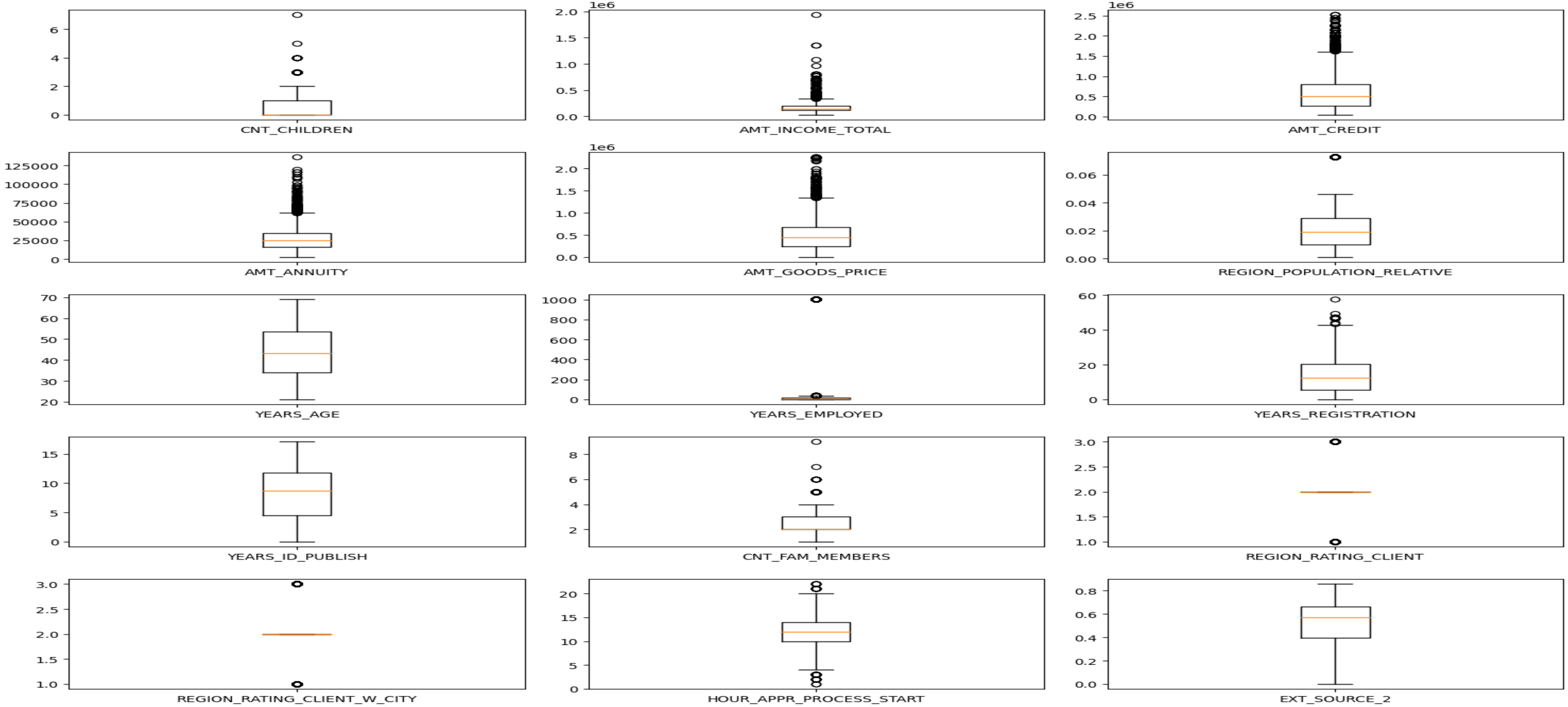
  REGION_POPULATION_RELATIVE: With only one outlier value, this attribute maintains a relatively stable distribution within the dataset.

  Remaining Attributes: A significant number of attributes showcase multiple outlier values, potentially indicating noteworthy data points that warrant careful consideration.

- **Data Integrity Assurance:**

  By meticulously pinpointing outliers, we ensure data integrity and set the stage for robust and informed decision-making in our lending practices.

Outliers

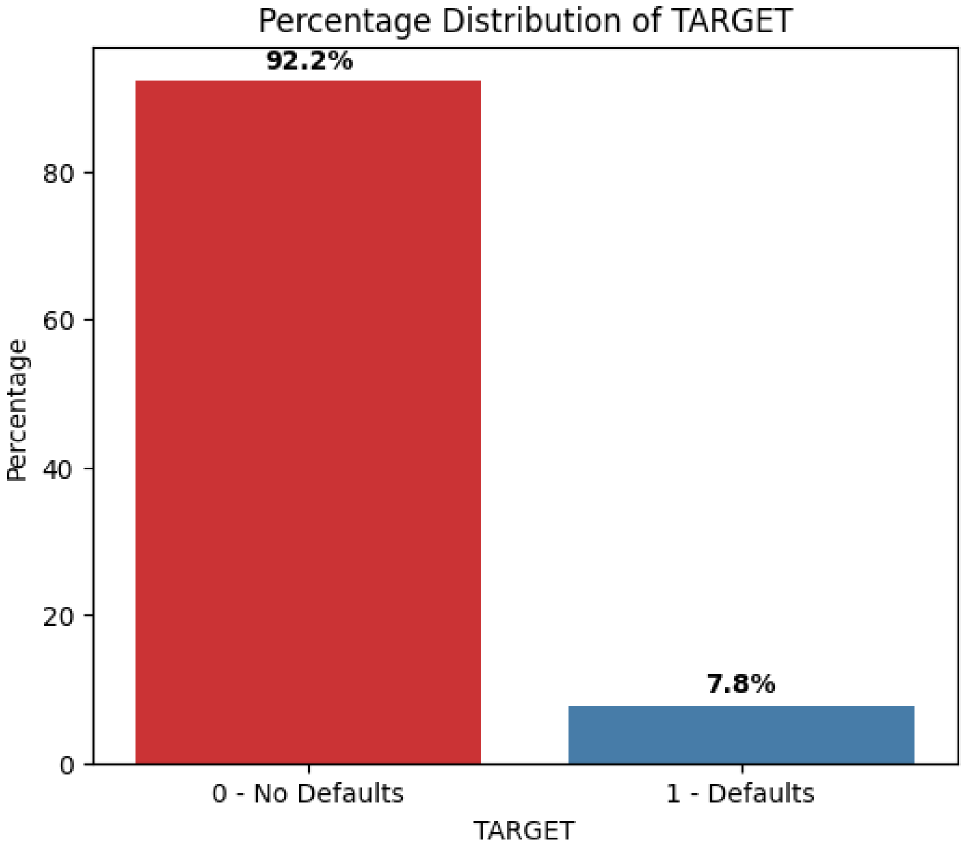


0
2
4
6
CNT_CHILDREN
1e6
0.0
0.5
1.0
1.5
2.0
AMT_INCOME_TOTAL
1e6
0.0
0.5
1.0
1.5
2.0
2.5
AMT_CREDIT
0
25000
50000
75000
100000
125000
AMT_ANNUITY
1e6
0.0
0.5
1.0
1.5
2.0
AMT_GOODS_PRICE
0.00
0.02
0.04
0.06
REGION_POPULATION_RELATIVE
20
30
40
50
60
70
YEARS_AGE
0
200
400
600
800
1000
YEARS_EMPLOYED
0
20
40
60
YEARS_REGISTRATION
0
5
10
15
YEARS_ID_PUBLISH
2
4
6
8
CNT_FAM_MEMBERS
1.0
1.5
2.0
2.5
3.0
REGION_RATING_CLIENT
1.0
1.5
2.0
2.5
3.0
REGION_RATING_CLIENT_W_CITY
0
5
10
15
20
HOUR_APPR_PROCESS_START
0.0
0.2
0.4
0.6
0.8
EXT_SOURCE_2

Univariate Analysis

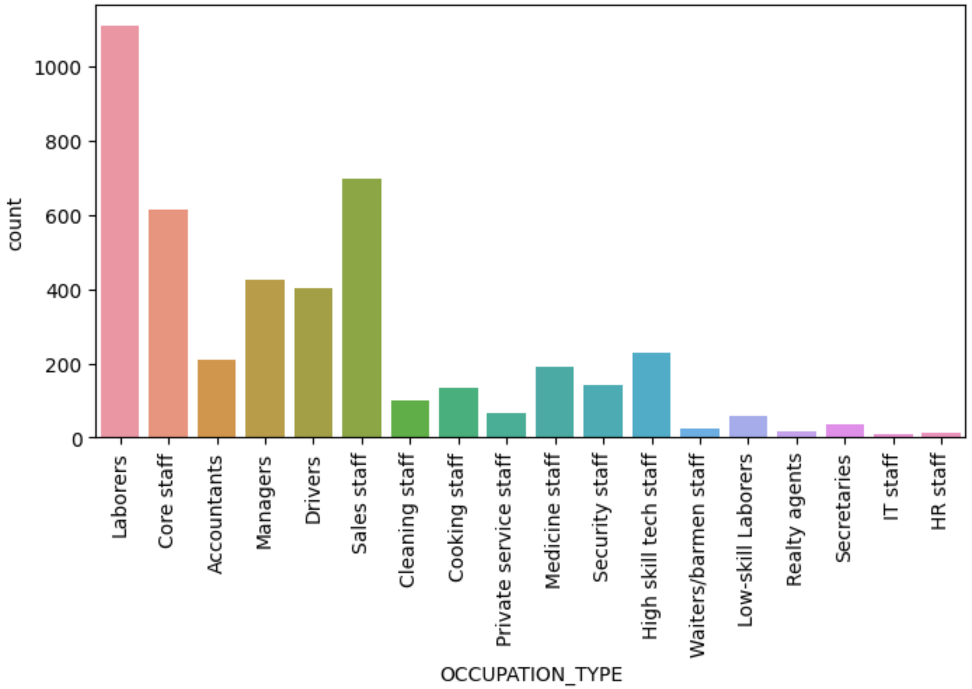
Exploring Individual Attributes

Target Variable



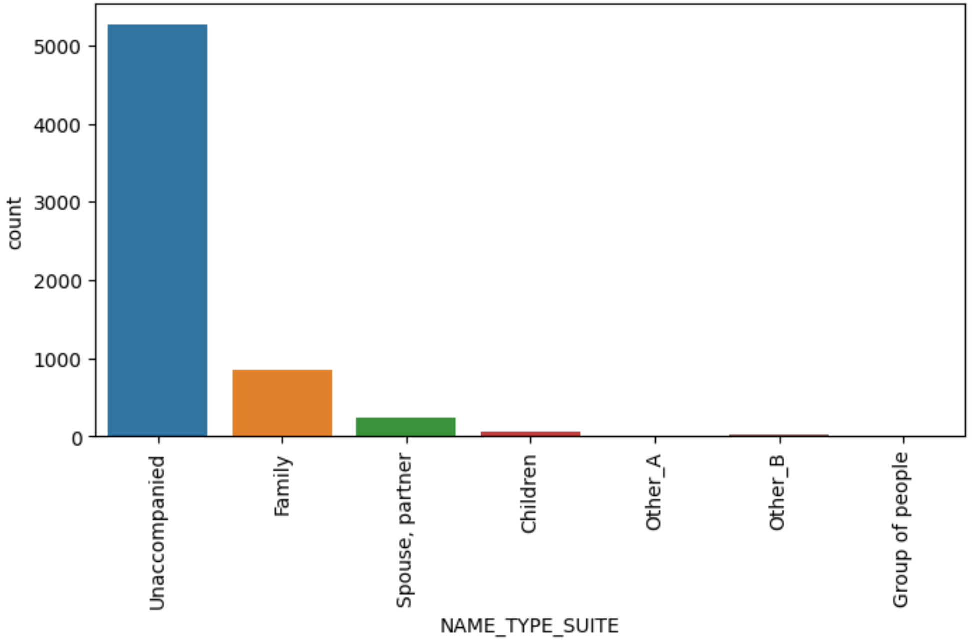
- **7.8**% of clients experience payment difficulties, as observed in the target variable's distribution during analysis.

Occupation Type



- ❖ Laborers emerge as the predominant occupation type in the dataset, showcasing their higher representation among applicants.

Name Type Suite



- ❖ Unaccompanied individuals primarily make up the "NAME_TYPE_SUITE" category, highlighting their prevalence among applicants.

Analyzing Categories & Default Risk: Insights into Categorical Variables' Impact on Loan Default Probability

NAME_CONTRACT_TYPE :

- Most of the customers has taken cash loan & persons who taken cash loan are less defaulter.

CODE_GENDER :

- Most of the loan taken by female & has approx. 7% default rate which is safer than male.

NAME_TYPE_SUITE :

- Most of the loans taken by Unaccompanied people & default rate is approx. 8%, which is safer.

Continue

Analyzing Categories & Default Risk: Insights into Categorical Variables' Impact on Loan Default Probability

NAME_INCOME_TYPE :

- Working, commercial associate & pensioners are safest segment to target.

NAME_EDUCATION_TYPE :

- Higher education is safest segment to target & default rate is less than 5%.

NAME_FAMILY_STATUS :

- Married people are safe to target & default rate is less than 8%.

Continue

Analyzing Categories & Default Risk: Insights into Categorical Variables' Impact on Loan Default Probability

NAME_HOUSING_TYPE :

- Customer who owns house/apartment are safest to target, defaults rate is approx. 8%.

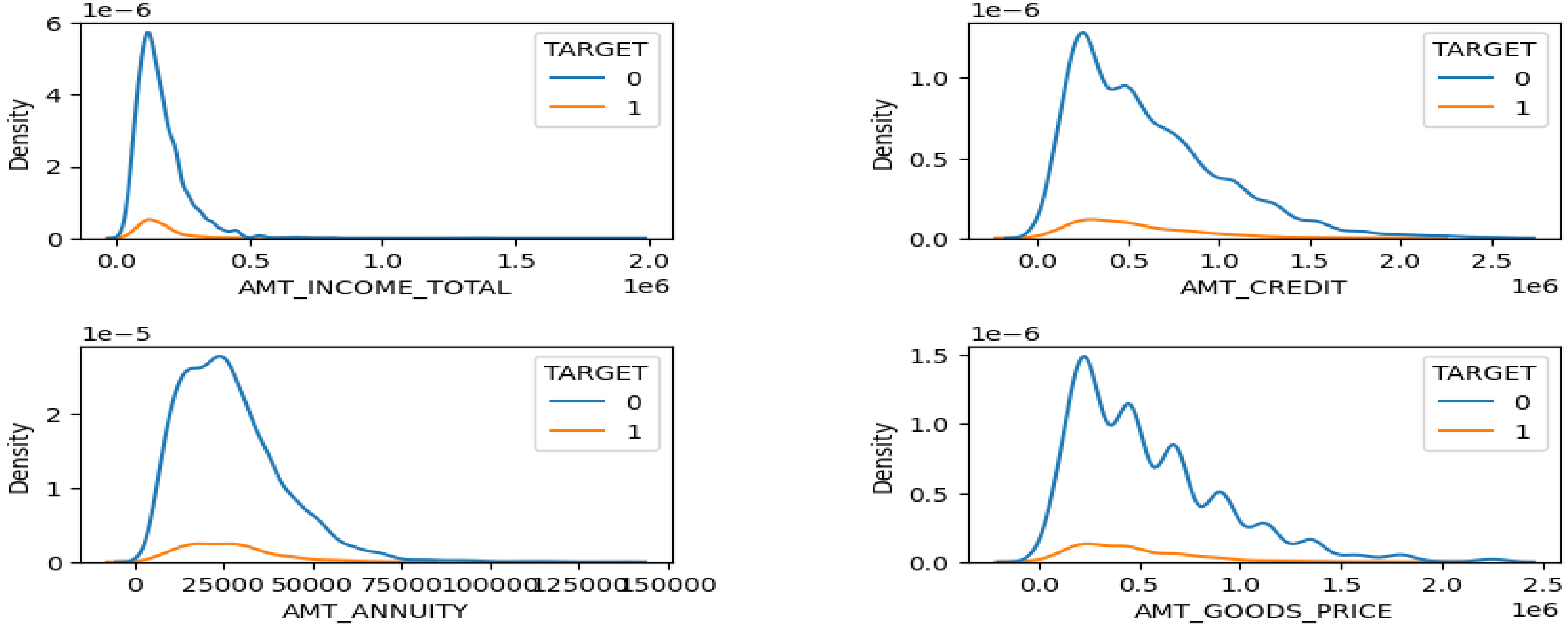
OCCUPATION_TYPE :

- laborers, accountants, core stuff, managers are less defaulters.
- low skill labors & drivers are higher default rate.

ORGANIZATION_TYPE :

- Transport type 3 are highest default rate.
- others, Business entity 3 are safest target, default rate is less than 10%.

Exploring Numerical Variables: Evaluating Density Distributions with Respect to Loan Default Probability



Exploring Numerical Variables: Evaluating Density Distributions with Respect to Loan Default Probability

AMT_INCOME_TOTAL :

- Most of the customer income between 0-10 Lakh.

AMT_CREDIT :

- Most of the credit amount between 0-10 Lakh.

AMT_ANNUITY :

- Most of the customer pay annuity between 0-50 Thousand.

Continue

Exploring Numerical Variables: Evaluating Density Distributions with Respect to Loan Default Probability

AMT_GOODS_PRICE :

- Most of the goods price between 0-10 Lakh

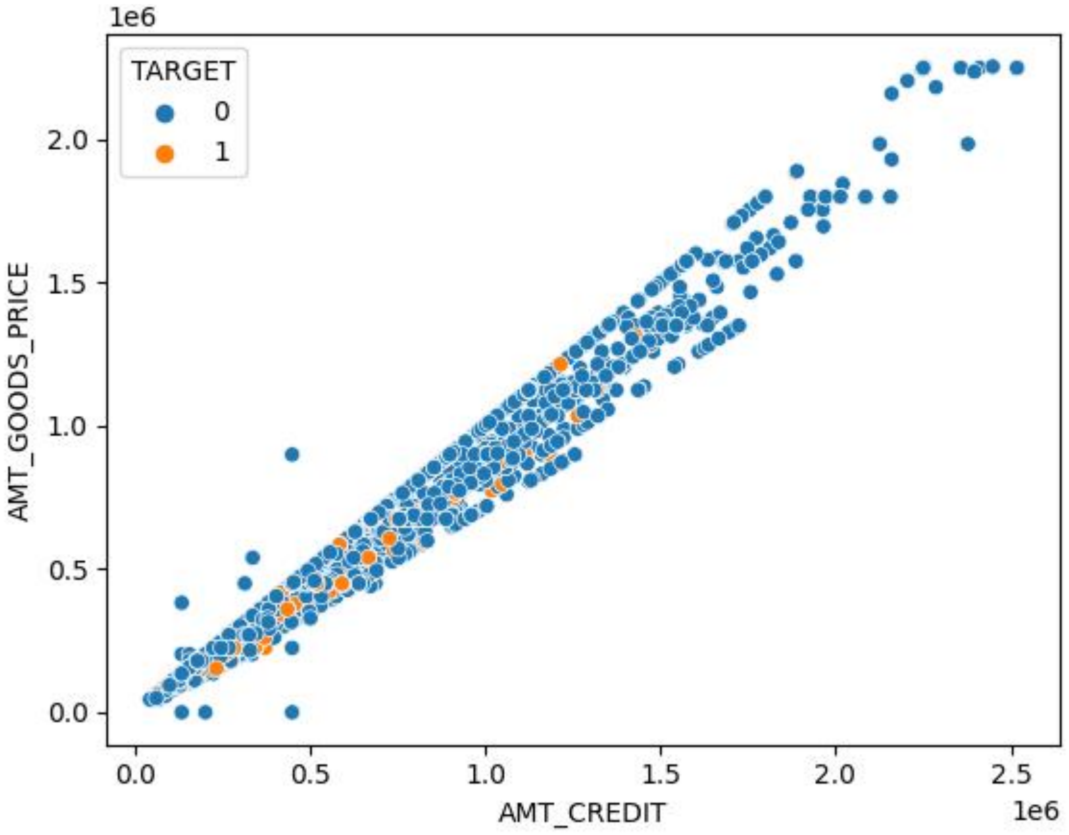
Bivariate and Multivariate Analysis

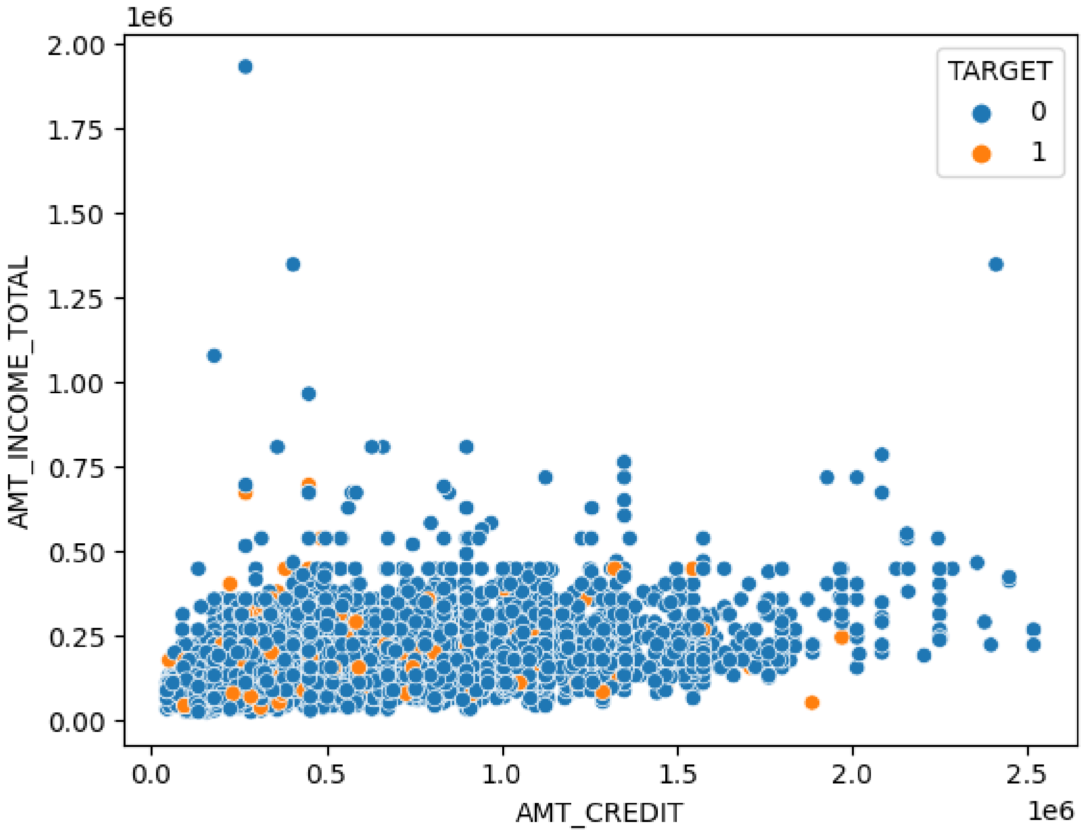
Revealing Relationships and Insights

Scatter Plot

- ❖ Amount credit and Goods price are co-related. If the amount credit increases, default rate decreases.

- ❖ Those customers whose income below 10 Lakh and taking loan above 15 Lakh are less defaulters.



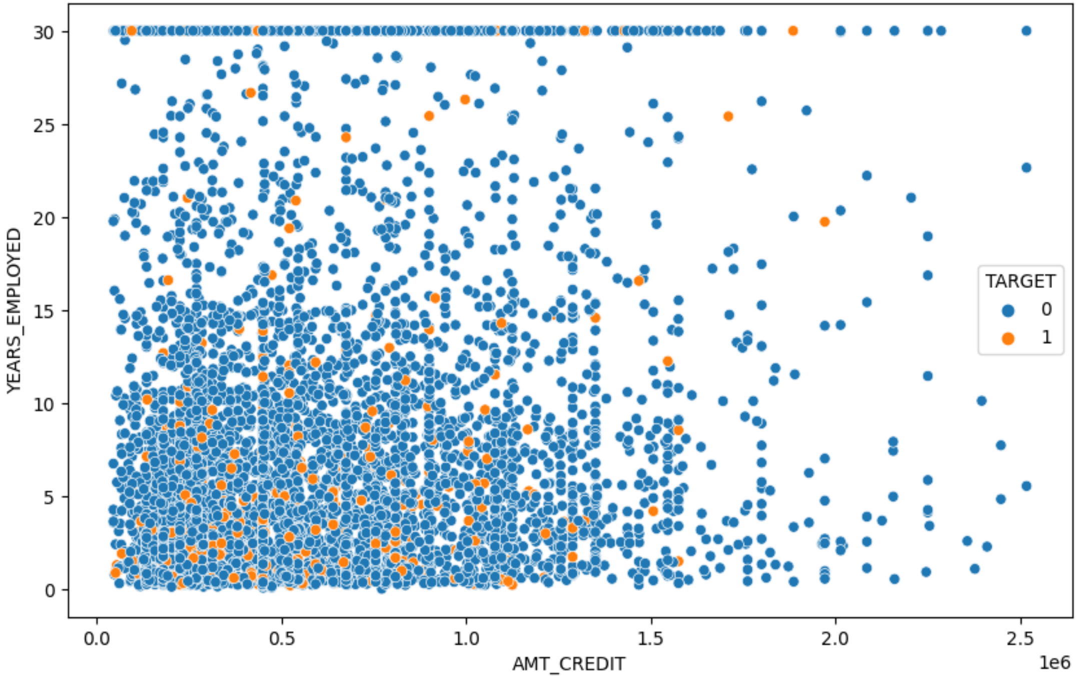


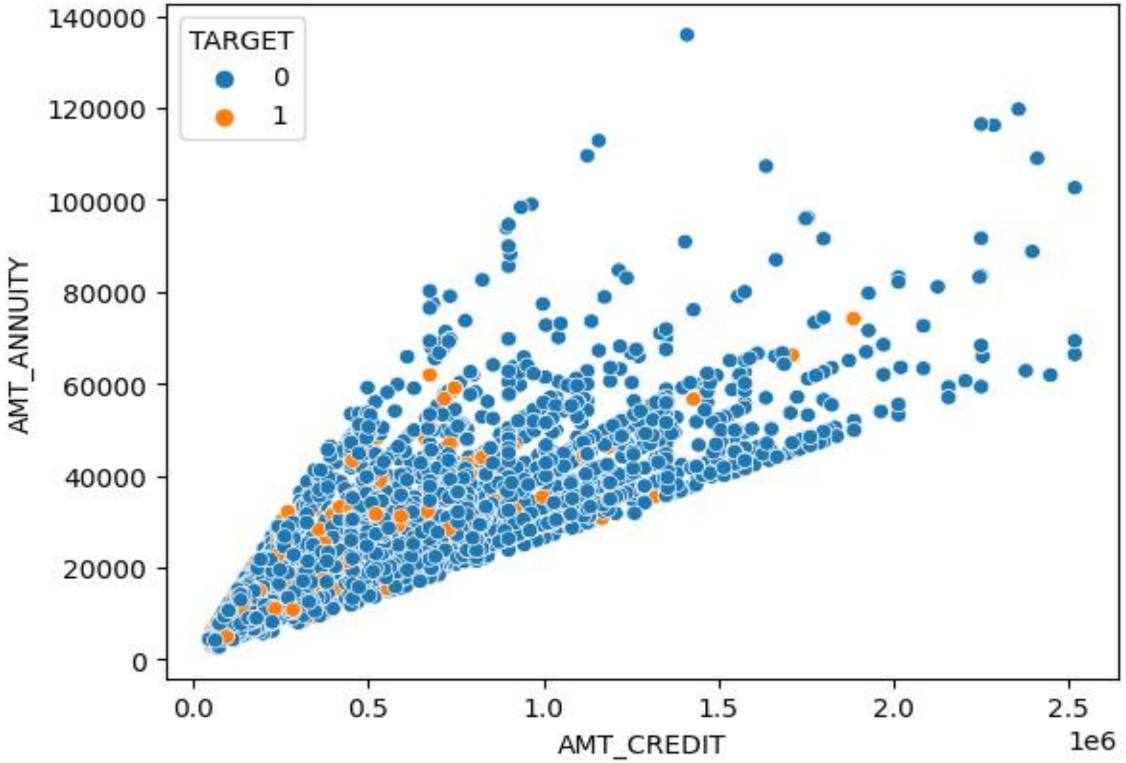
Continue

Scatter Plot

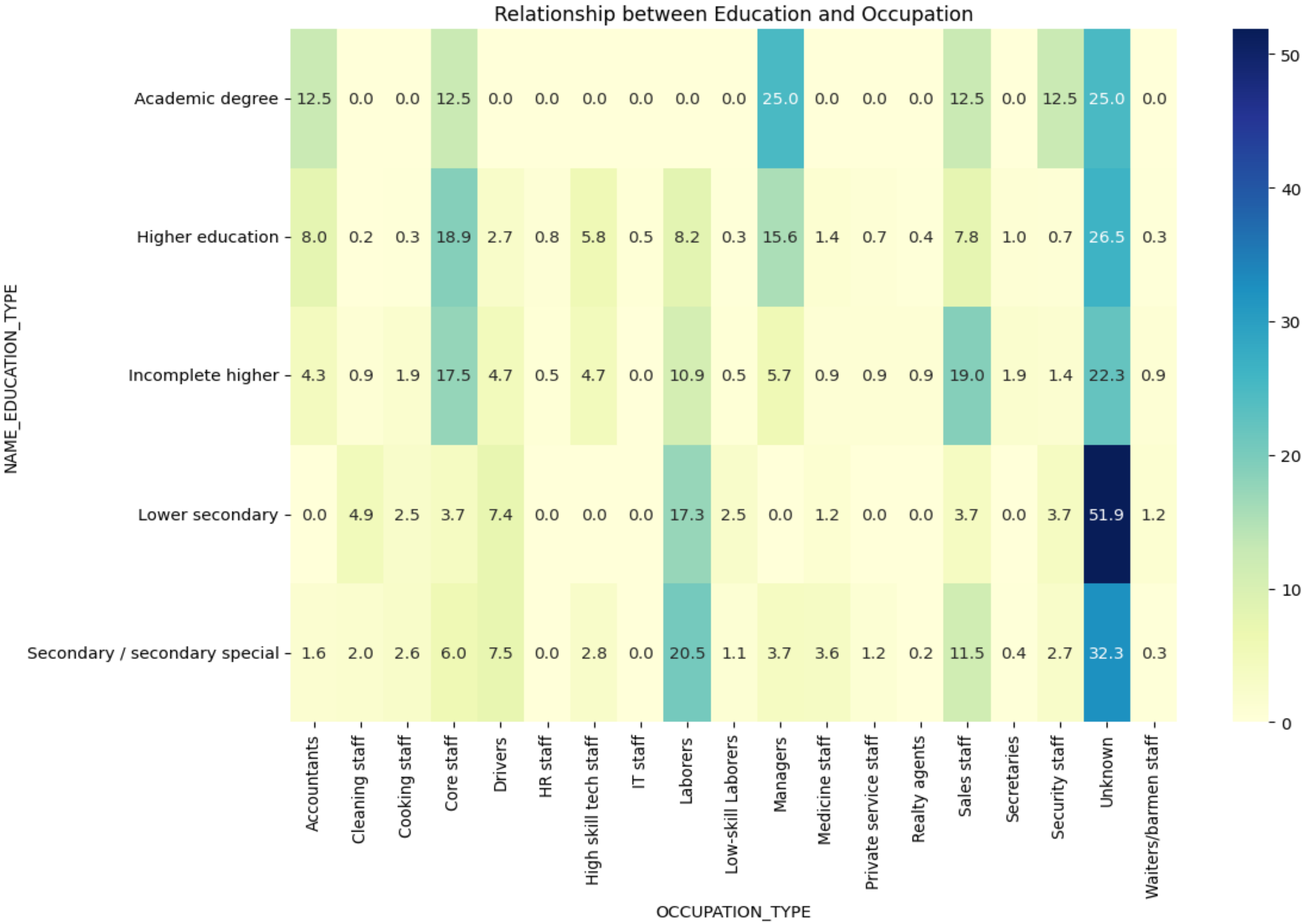
- People having more than 15 years of experience and taking loan above 15 lakh are less defaulters.

- People who are paying more than 40 thousand annuity and taking loan more tahn 15 lakh to 25 lakh are less defaulters and safer target.



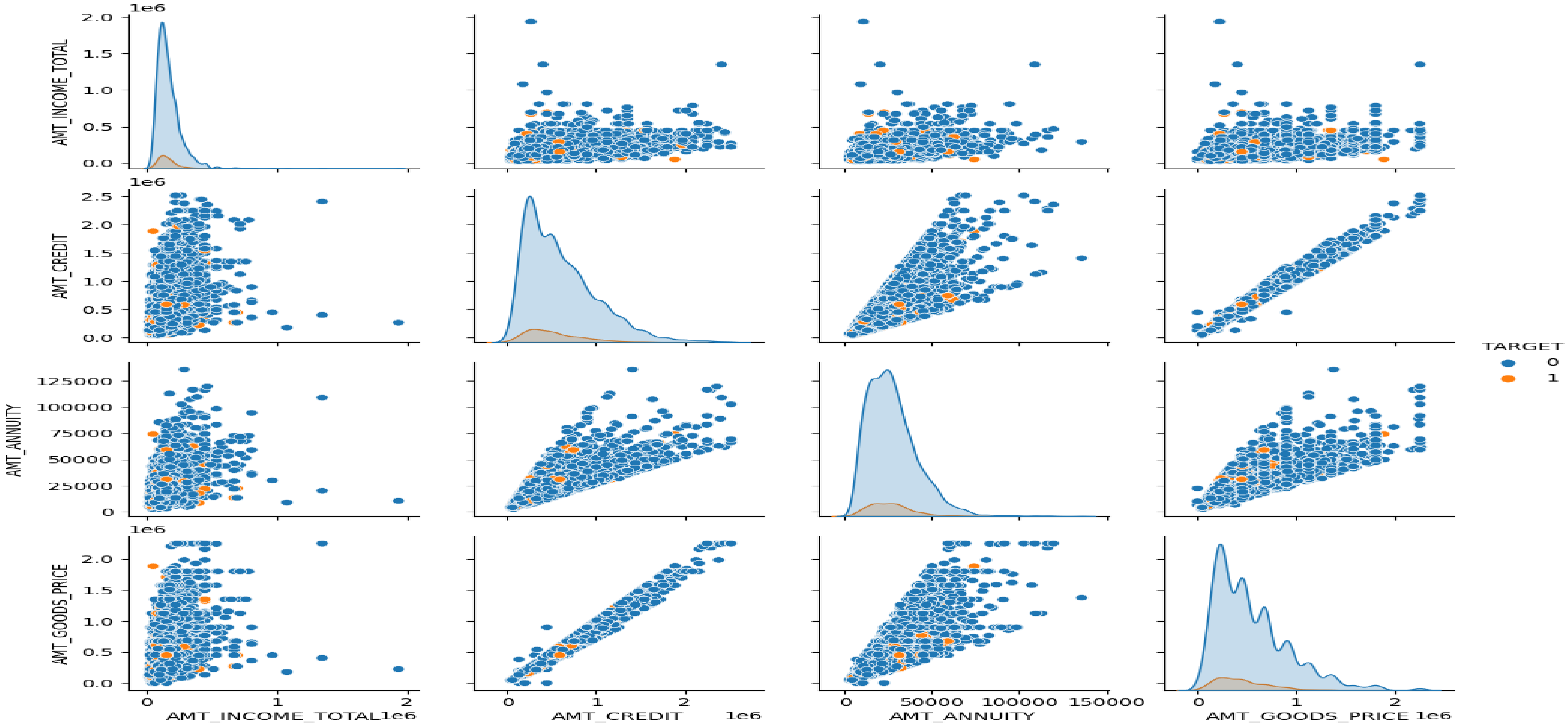


Heatmap



- ❖ Managers are predominantly associated with higher education and academic degrees.
- ❖ Core staffs are primarily linked to higher education and incomplete higher education levels.
- ❖ Laborers are predominantly associated with secondary and lower secondary education levels.

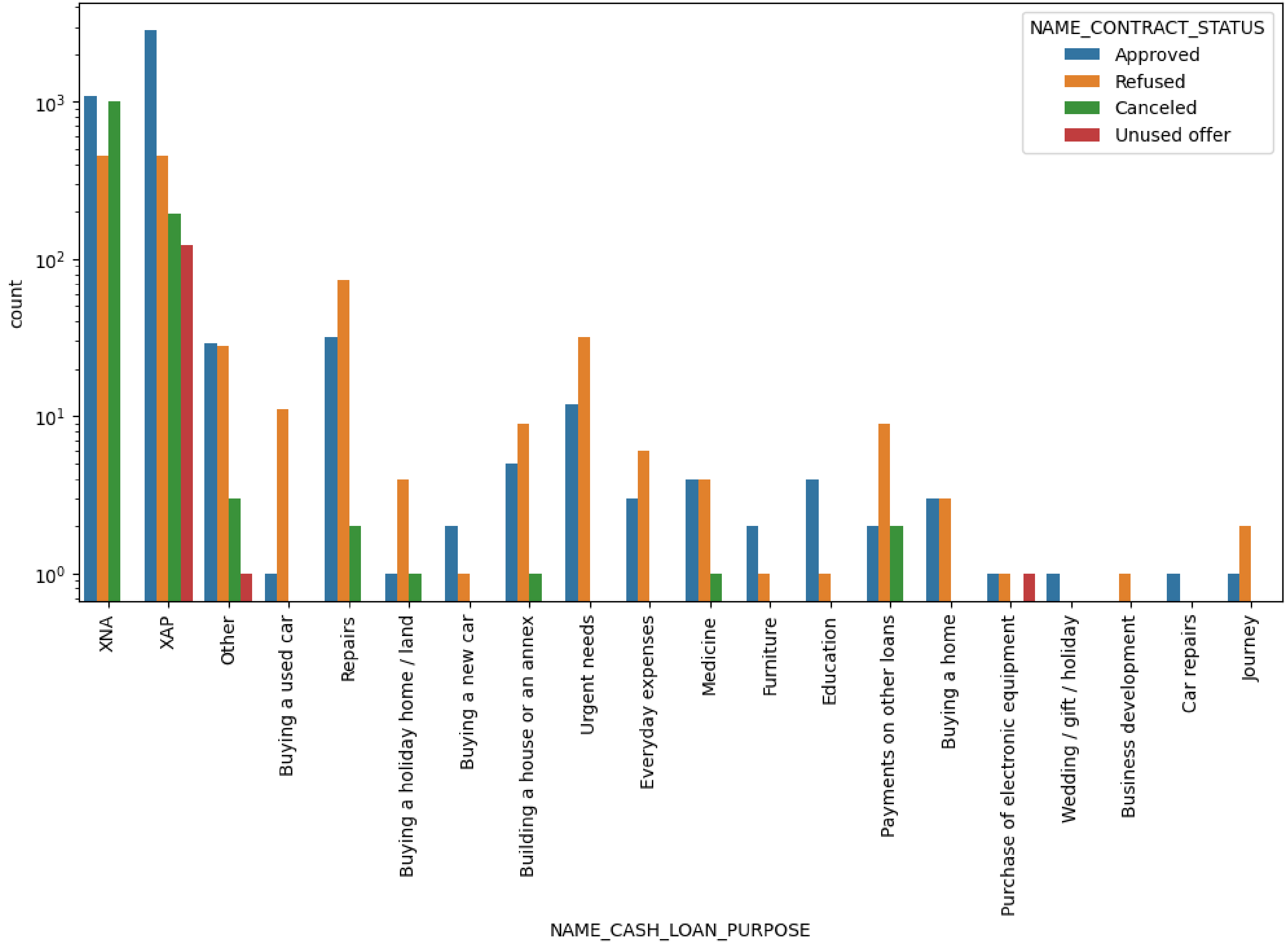
Exploring Relationships: Pair Plot Analysis



Continue

Analyzing Combined Datasets

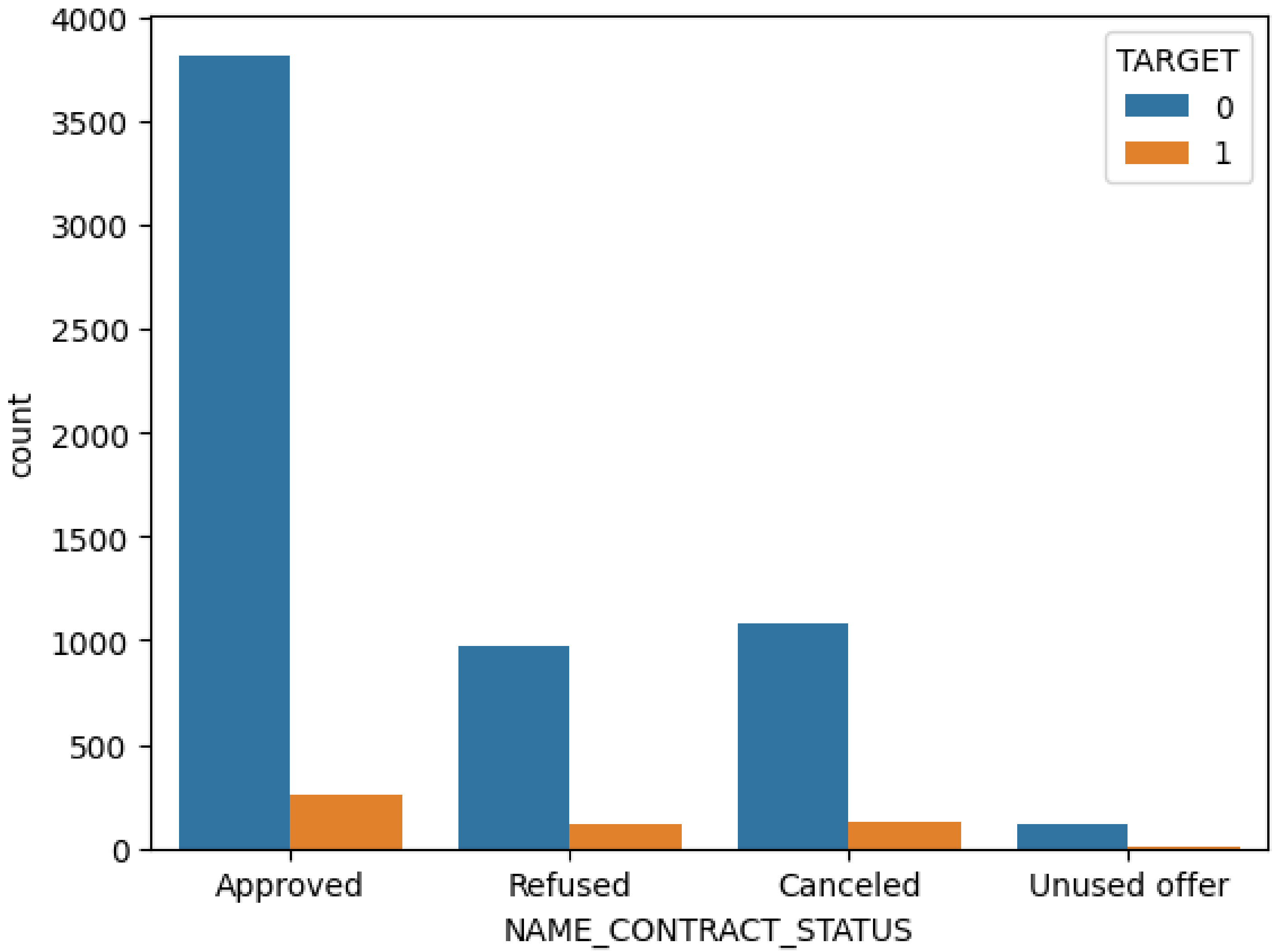
Count Plot : 'NAME_CASH_LOAN_PURPOSE' vs 'NAME_CONTRACT_STATUS'



- Repairs, Others & Urgent Needs are having most approved and refused cases respectively.

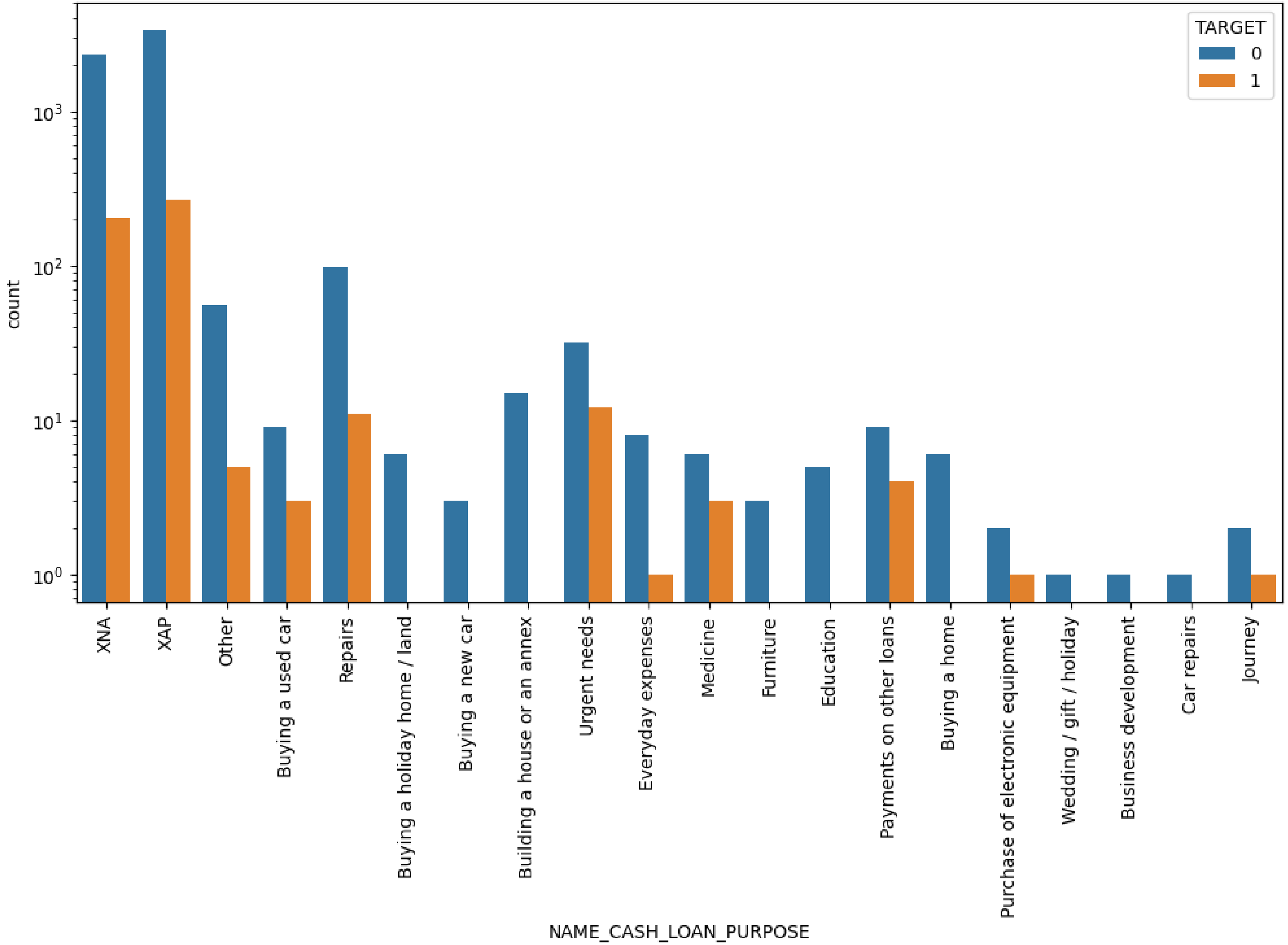
Continue

Count Plot : 'NAME_CONTRACT_STATUS' vs 'TARGET'



- ❖ Most of the application which are previously cancelled or refused, approx. 90 % of them are consider in current data set.

Count Plot : 'NAME_CASH_LOAN_PURPOSE' vs 'TARGET'



- Buying a garage category of previously loan purpose are less defaulters.
- Hobby category of previously loan purpose are highest percentage of defaulters.
- Repairs, Others and Urgent need category of previously loan purpose are above 85 % of success paying rate.

Conclusion: Drawing Insights and Guiding Strategies

Recommendations and Precautions

Enhancing Lending Strategies:

Recommendation:

Bank should target the customers

- Customers having income below 10 Lakh
- Customers working with Other, Business entity 3, self employed and pensioners
- Customers working as accountants, Core stuff, Managers and laborers
- Customers having own house/apartment
- Customers having higher education
- Customers paying annuity up to 50 thousands
- Customers are married
- Unaccompanied Customers
- Female Customers
- Credit amount should be up to 10 lakh
- Bank should analyze and consider previously cancelled and refused customers

Continue

Precautions:

Bank should avoid the customers

- Customers are transport type 3 from organization type
- Customers are Low skill Laborers and Drivers from occupation type

Thank You

Sonmoy Jana

9123329073

Sonmoy.jana123@gmail.com

IIIT-Bangalore Alumini

TREY research