Exploratory Data Analysis in Banking and Risk Analytics

PREPARED BY
DIGANTH RAI

Introduction

A credit risk assessment is a process of assessing the likelihood that a borrower will be unable to meet its financial obligations, such as payments or reaching a loan agreement This assessment examines a borrower’s financial stability, credit history , and other relevant factors to make sound and supportive lending or investment decisions.

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of the loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected.

Data Preparation:

Handling Missing Value:

- Identify missing values, Begin by identifying where data is missing in the dataset.
- Options for handling missing values:
  - Removal: We can remove rows or columns with a significant number of missing values, being cautious about potential data loss.
  - Imputation: Impute missing values by filling them in with reasonable estimates. Common imputation methods include mean, median ,mode or predictive techniques.

```
# Calculate the percentage of null values in each column
((df3.isnull().sum() / len(df3)) * 100).sort_values(ascending=False)
```

```
COMMONAREA_MEDI                69.872297
COMMONAREA_AVG                 69.872297
COMMONAREA_MODE                69.872297
NONLIVINGAPARTMENTS_MODE       69.432963
NONLIVINGAPARTMENTS_AVG        69.432963
NONLIVINGAPARTMENTS_MEDI       69.432963
FONDKAPREMONT_MODE             68.386172
LIVINGAPARTMENTS_MODE          68.354953
LIVINGAPARTMENTS_AVG           68.354953
LIVINGAPARTMENTS_MEDI          68.354953
FLOORSMIN_AVG                  67.848630
FLOORSMIN_MODE                 67.848630
FLOORSMIN_MEDI                 67.848630
YEARS_BUILD_MEDI               66.497784
YEARS_BUILD_MODE               66.497784
YEARS_BUILD_AVG                66.497784
OWN_CAR_AGE                    65.990810
LANDAREA_MEDI                  59.376738
LANDAREA_MODE                  59.376738
```

The above SS shows the presence of null value in the datasets that was given us . To handle this ,

We initially identified and addressed missing data. Most instances with more than 40% missing values were removed to maintain data integrity.

For the remaining missing values, we applied suitable statistical methods. Numeric data was imputed using mean, mode, or predictive techniques. For object-type data, we used the mode value or the label "missing" to maintain data consistency.

These steps ensured that our dataset is now reliable and ready for meaningful analysis and modeling, minimizing the potential impact of missing values.

```
In [15]: df3.head()
         #Go through the data for further cleaning,also notice the negative values presnt the days section
```

Out[15]:

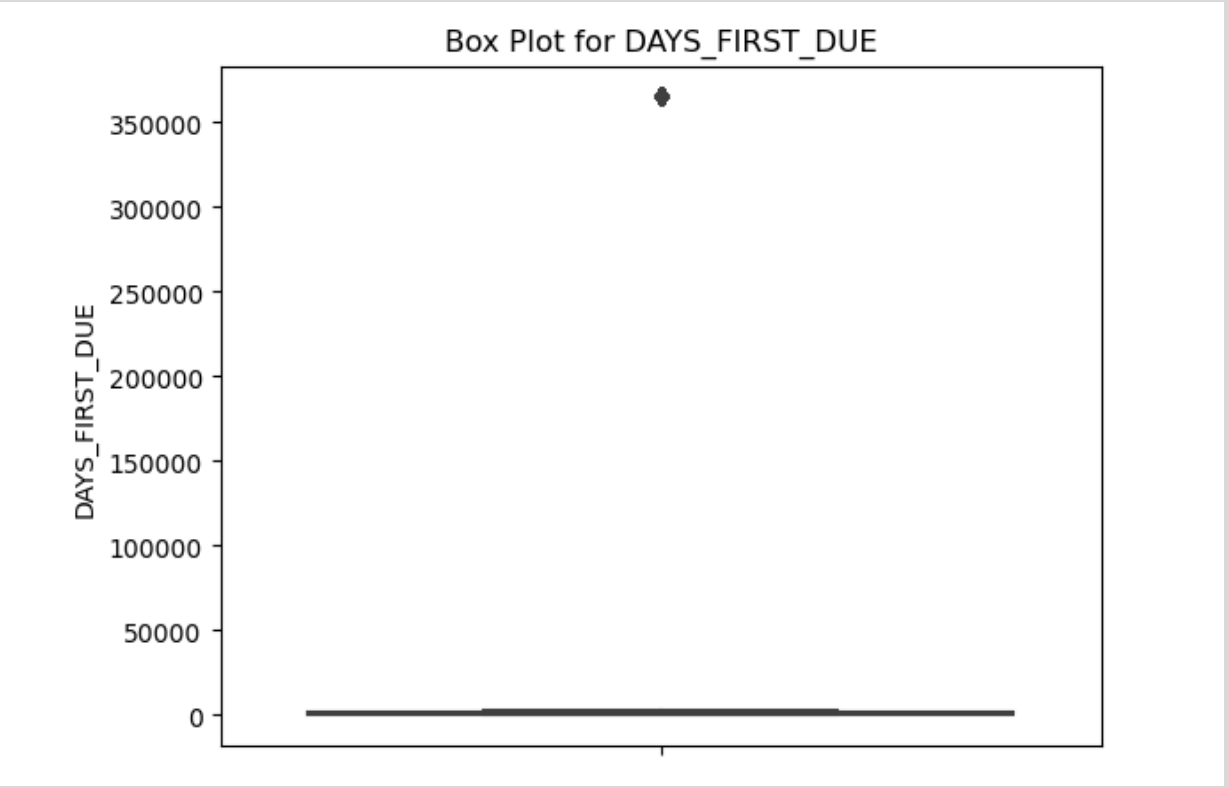
_ATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	FLAG_MOBIL	FLAG_EMP_PHONE
0.018801	-9461	-637	-3648.0	-2120	1	1
0.003541	-16765	-1188	-1186.0	-291	1	1
0.010032	-19046	-225	-4260.0	-2531	1	1
0.008019	-19005	-3039	-9833.0	-2437	1	1
0.028663	-19932	-3038	-4311.0	-3458	1	1

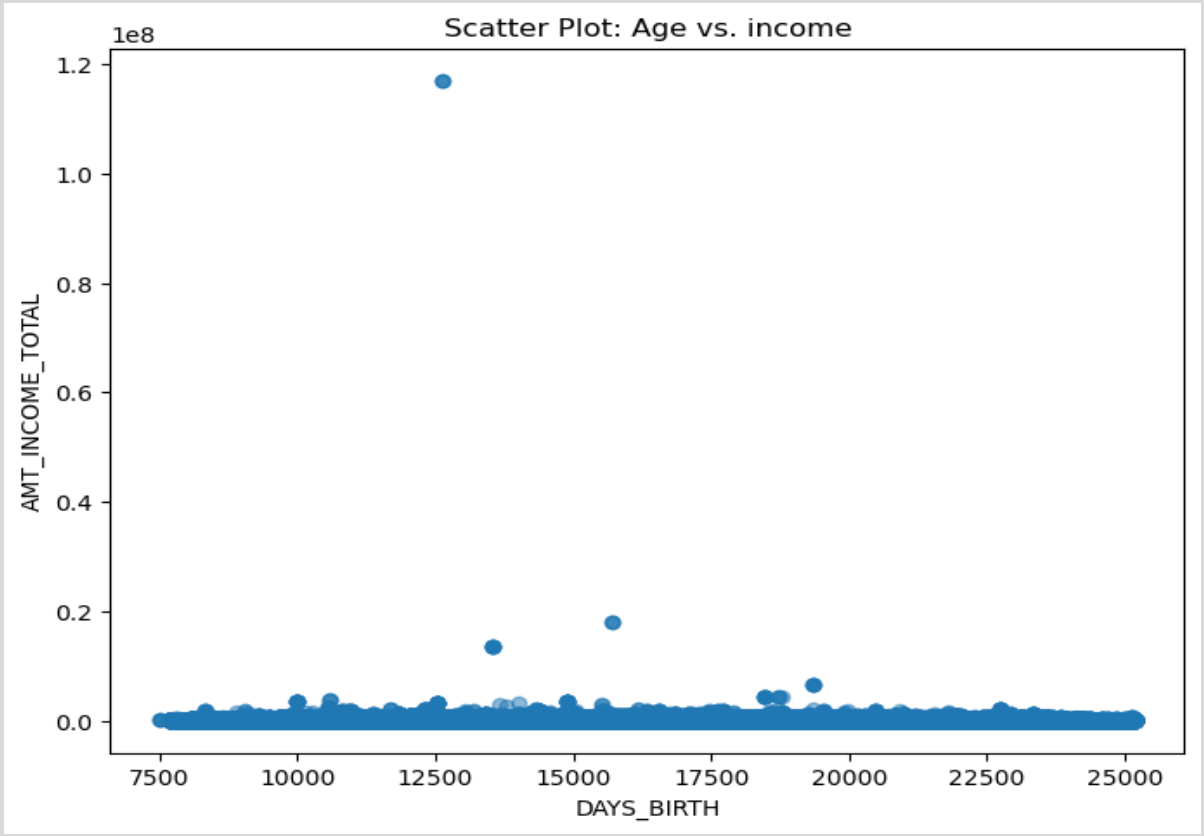
Identified negative values in columns related to days, which need to be adjusted. Negative values in day-related columns might be an error or represent a specific condition, and converting them to positive values can facilitate further analysis. Used abs() function for the same.

Data Preparation:

Outlier Detection and Handling:

- Identify outliers: Outliers are data points that significantly deviate from the majority of the data. Utilize statistical methods or data visualization to identify outliers.
- Options for handling outliers
  - Removal: We can remove rows or columns with the outlier, being cautious about potential data loss.The outlier can be stored in sperate dataset for future reference.
  - Imputation: Replace outlier with reasonable estimates.





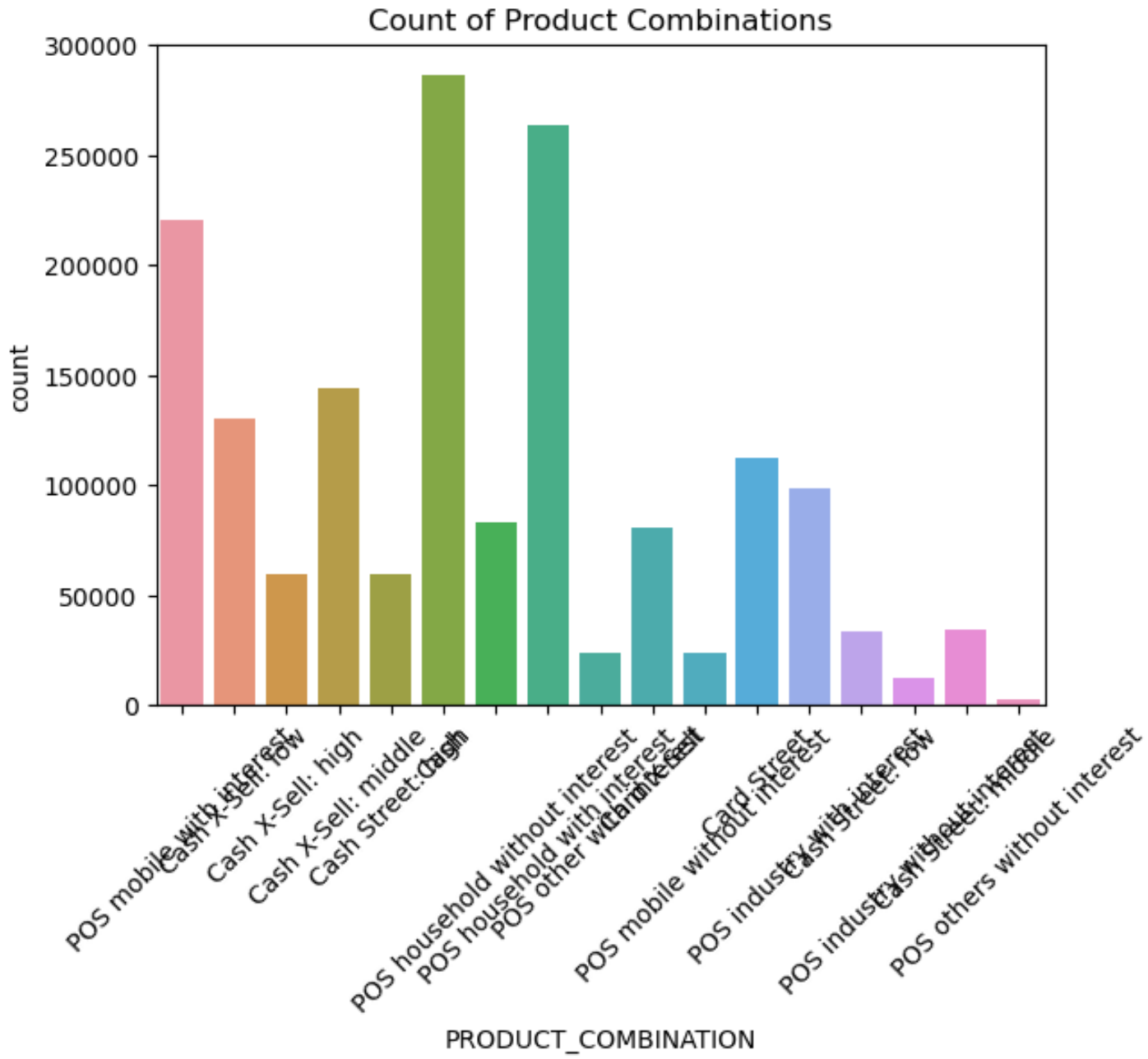
The above two plots show outliers. The first plot shows a number of days of first due nearly 35,000, while the others are within a range of 100. There are few more outlier that is shown in analysis

The second plot shows an outlier in annual income that is much higher than the others in the column.

Univariate and Bivariate Analysis:

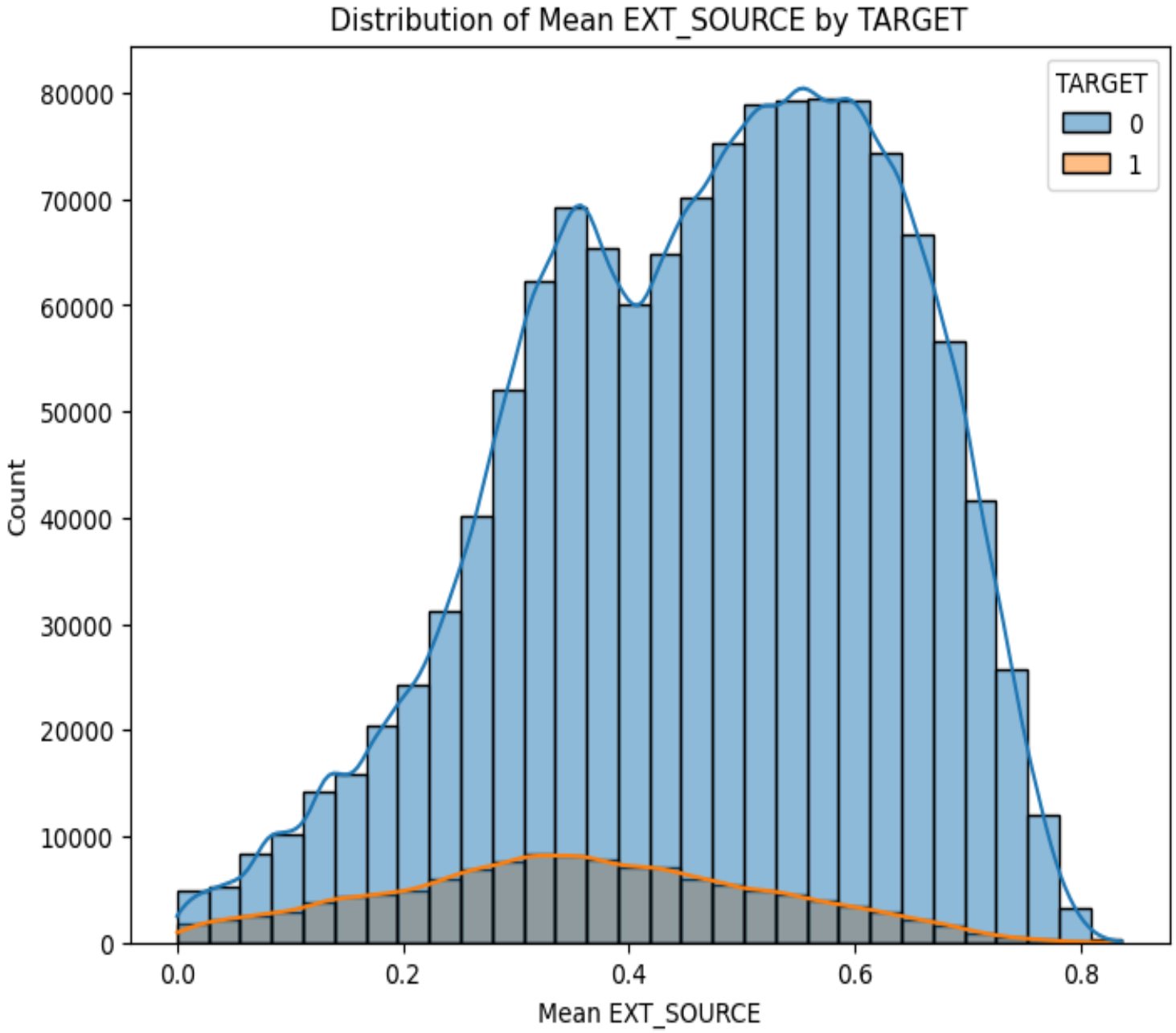
- Univariate analysis is a statistical method used in data analysis to examine and describe the characteristics and properties of a single variable in isolation
- It is used for for understanding and summarizing individual variables, making data-driven decisions, and preparing for more comprehensive analyses.
- Bivariate analysis is a statistical method used in data analysis to explore and analyze the relationship between two variables.
- Understanding the relationship between two variables is valuable for making predictions, identifying factors that influence outcomes, and informing decision-making processes.

A count plot can be used to decide on the values that are to be imputed with. For example, if you have a dataset with missing values in the PRODUCT_COMBINATION column, you could use a count plot to identify the most common product combinations. You could then impute the missing values with the most common product combinations that is mode().



In the shown (bivariate) hist plot of mean of normalized score from external data source, we observe that the number of non defaulters increased as the score increased till 0.6.

In other words, individuals with higher scores from the external data source are less likely to default on some financial obligation or criteria you are studying. This could be indicative of a potential relationship between the external data source's score and the creditworthiness or reliability of individuals in fulfilling their financial obligations.



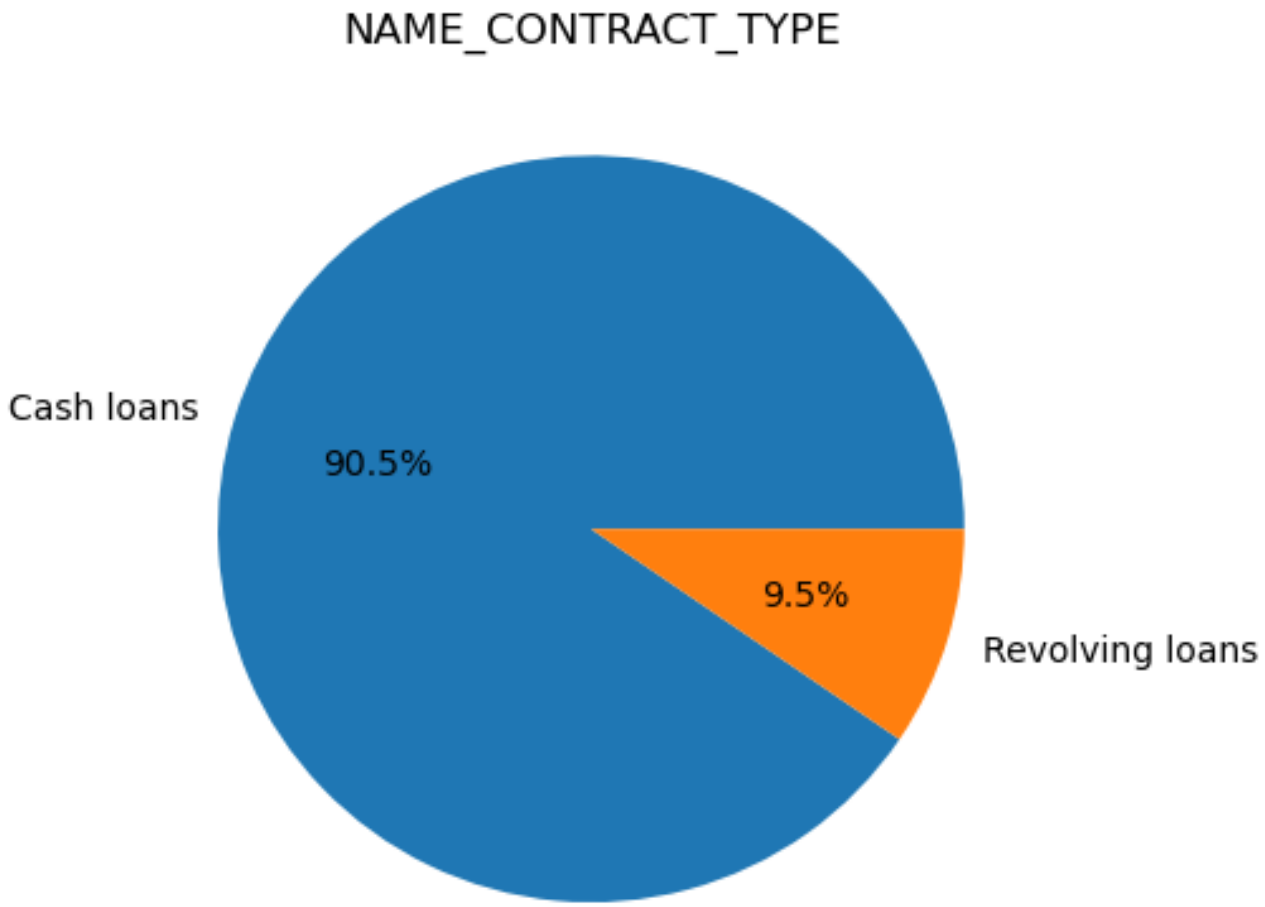
Visualizations:

Data visualizations are graphical representations of data that make complex information more accessible, understandable, and actionable. They play a crucial role in data analysis and communication by helping to reveal patterns, trends, and insights that might be difficult to discern from raw data.

1. Line plots
2. Bar chart
3. Histogram
4. Box Plot
5. Scatter Plot
6. Pair Plot
7. HeatMap

Insight

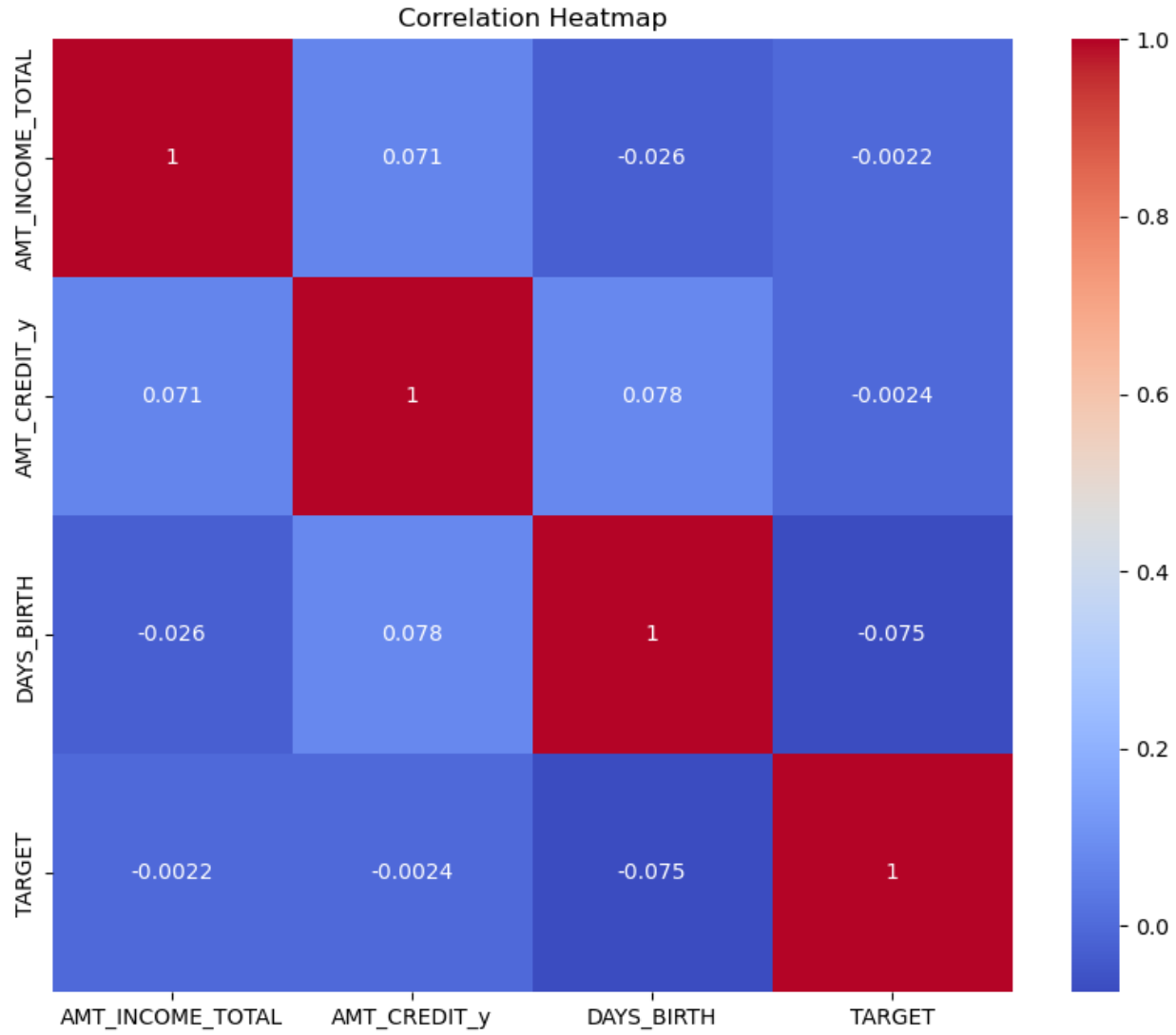
Among the options shown in the pie chart, credit is the most popular with customers. These conclusions are based on data and graphics presented in the chart, which shows that a larger portion of the pie is allocated to the "cash-type loan" category, which means it has a larger share of consumer preferences as are compared with other types with loans included in the scheme



Insight

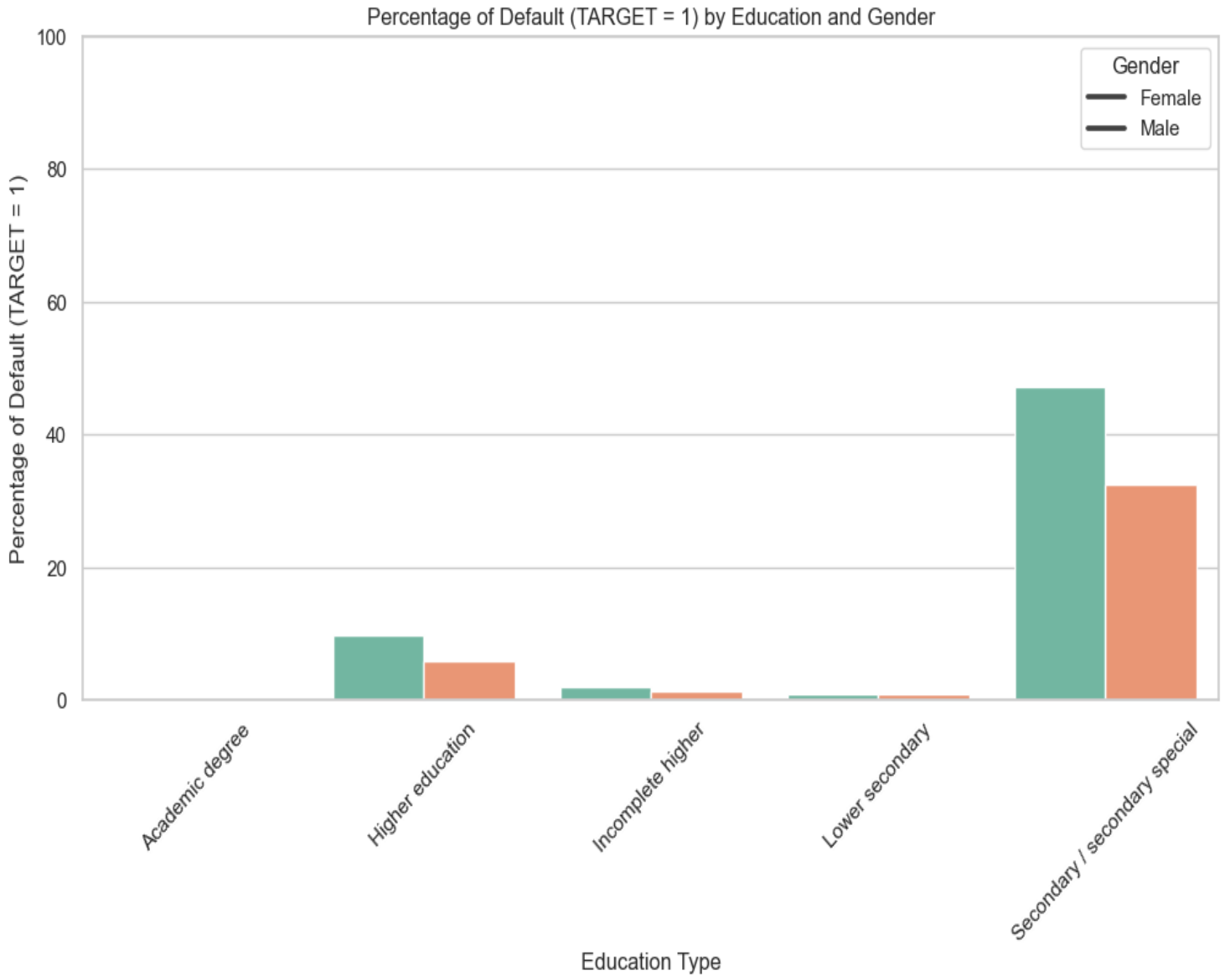
There is no strong trends ,but the negative correlation between AMT income and AMT credit can be used by creditors to assess a borrower's credit risk.

Borrowers with higher AMT income and lower AMT credit are generally considered to be more risky borrowers..



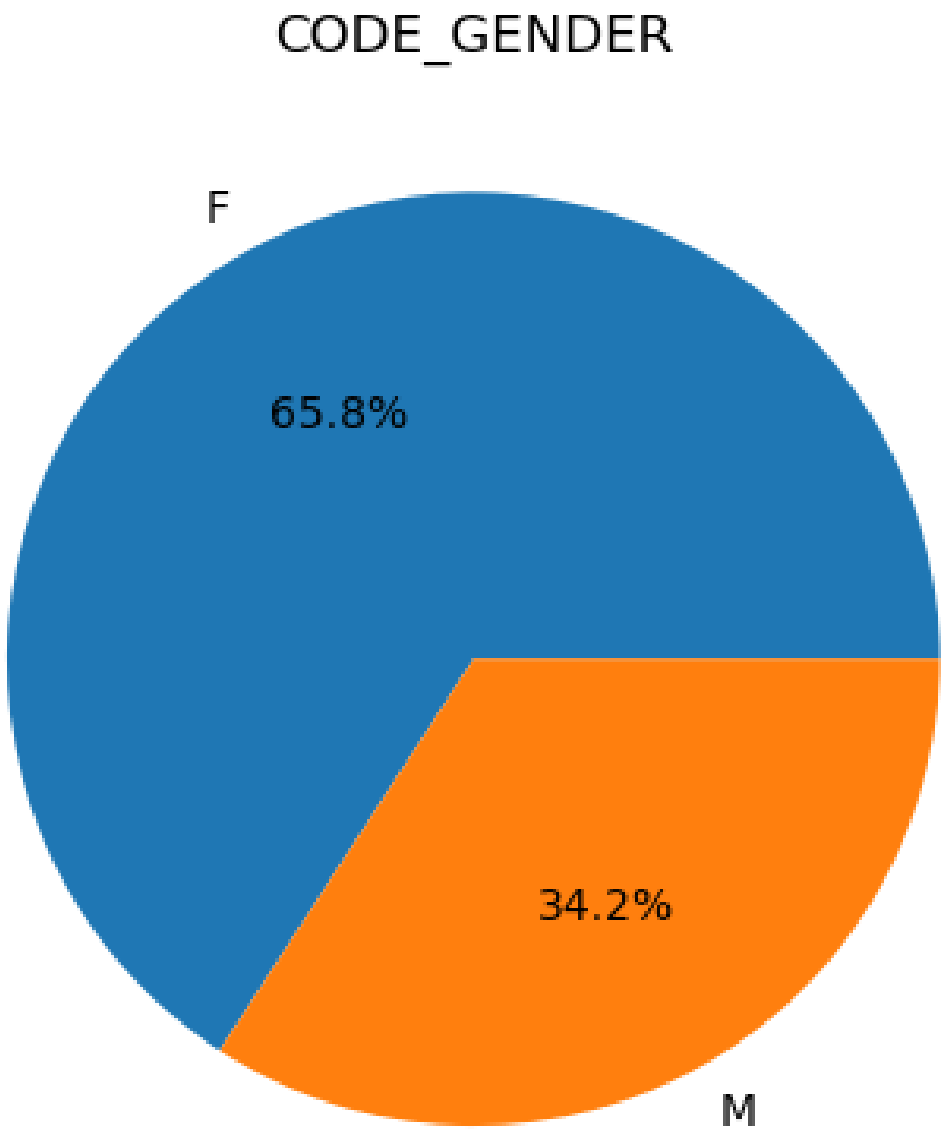
Insight:

Educated females more likely to borrow and not repay loans than educated males, possibly due to lower income and more responsibilities.



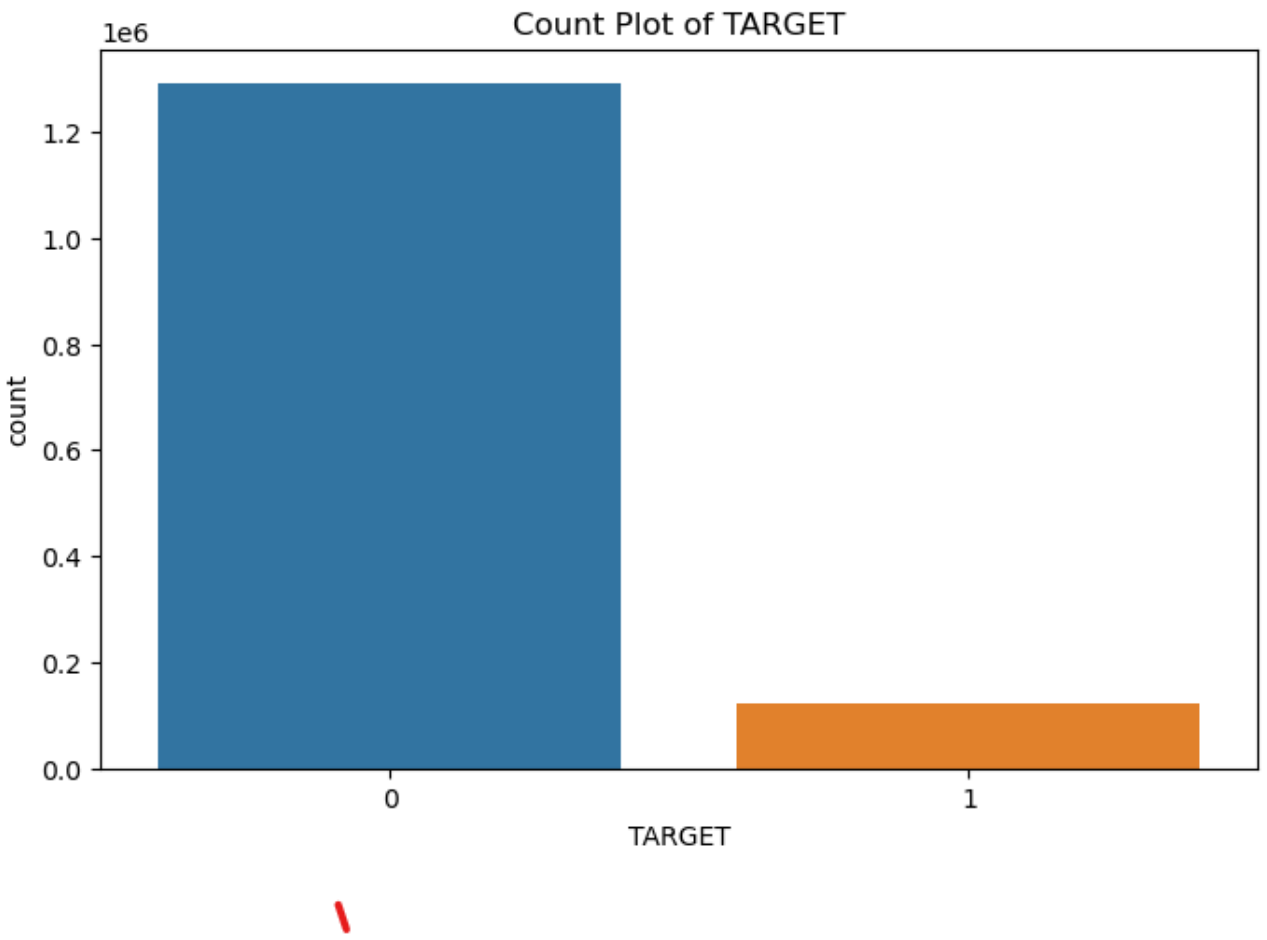
Insight

Majority of the costumers applying for loans are female (i.e 65.8%).It is also an imbalanced data.



Insight:

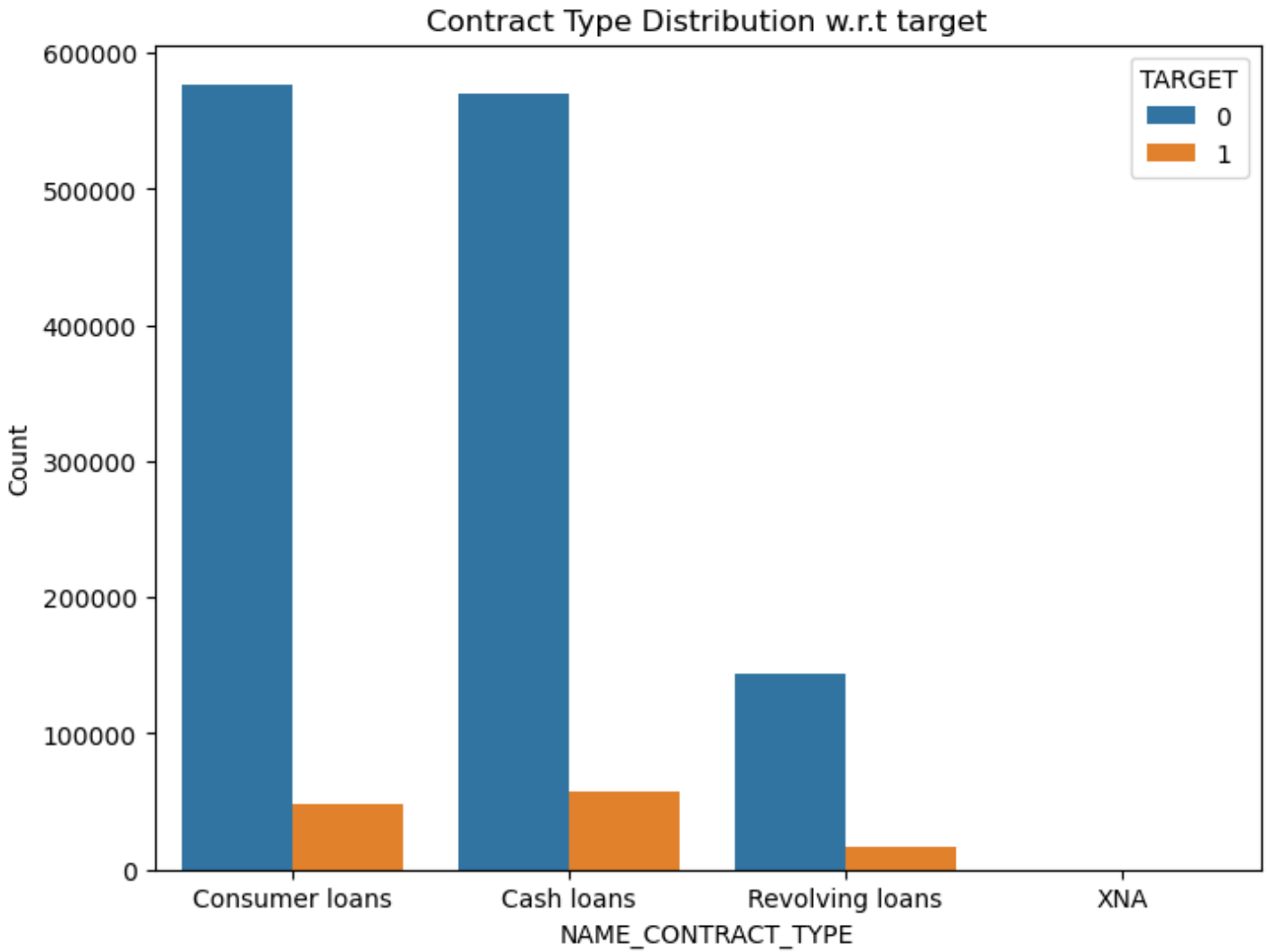
The target class in credit risk analysis dataset is imbalanced, meaning that there are many more clients without payment difficulties than clients with payment difficulties. This can be a problem for machine learning algorithms, so it is important to take steps to address it.



Insight

The majority of loans are cash loans. This means that the borrowers used the loan money for general purposes

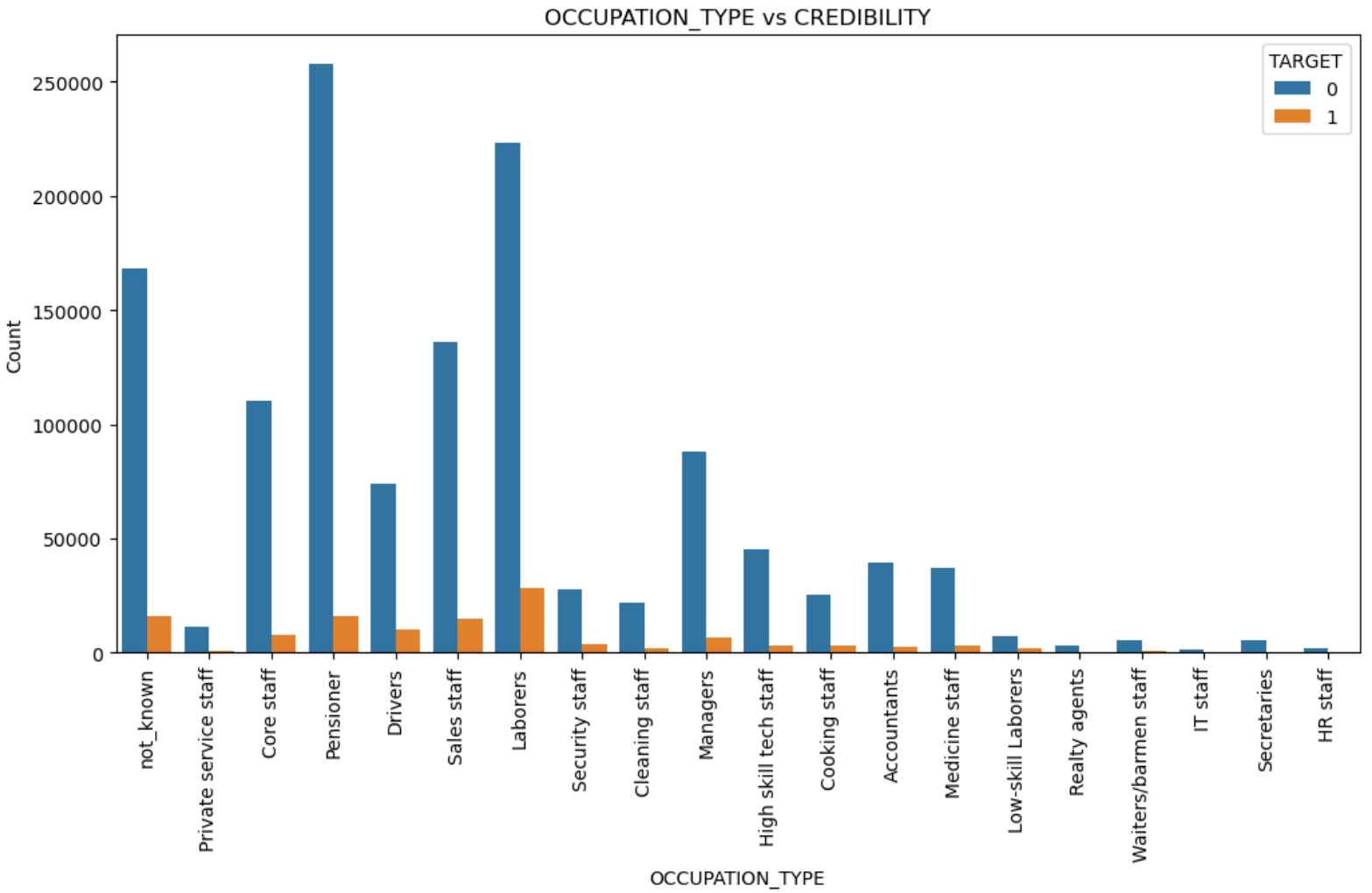
The default rate for cash loans is higher than the default rate for other types of loans. This suggests that borrowers who take out cash loans may be more likely to default on their loans.



Insight

The plot provided shows the default rate for different types of borrowers, based on their occupation.

The plot shows that laborers have the highest default rate



Insights

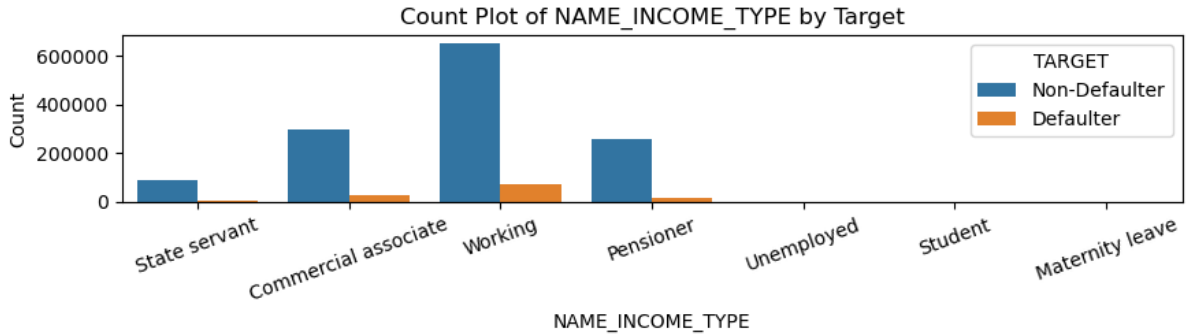
In the above bar chart,

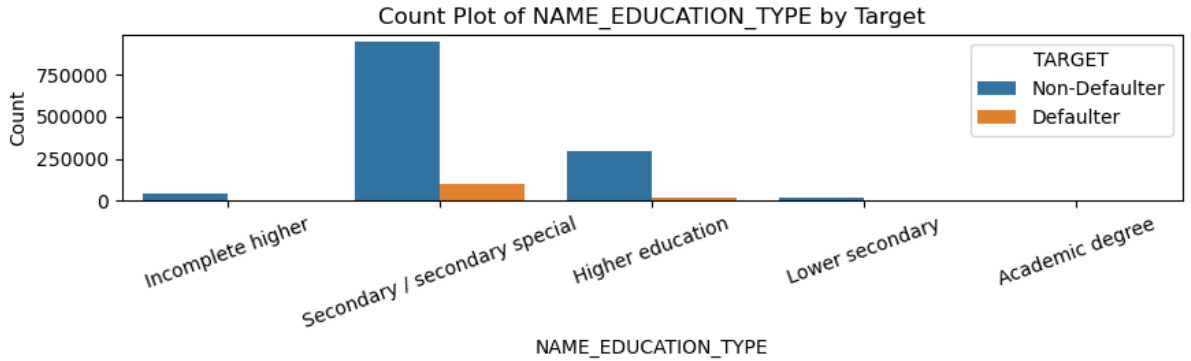
Clients who are employed in Working, Commercial associate, State servent and Pensioner roles are more likely to apply for loans, with the highest application rate among the Working class. Conversely students, and unemployed individuals are less likely to apply for loans. Among these groups, the Working category has a higher risk of default, while State Servants exhibit a minimal risk of default.

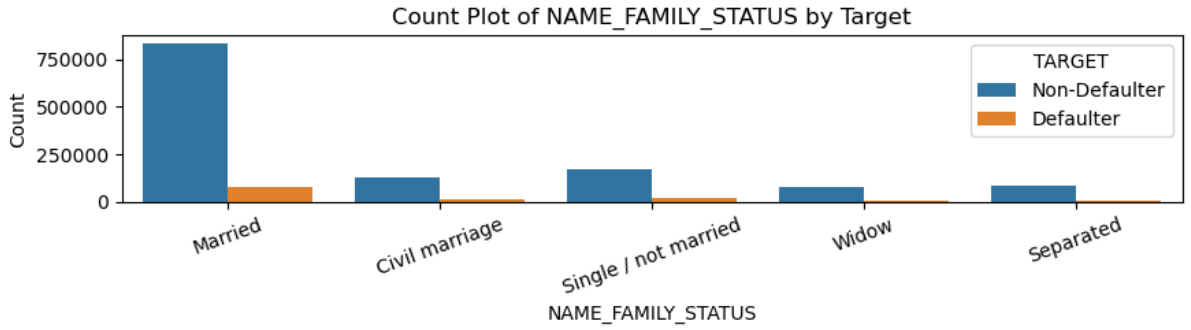
Costumer with secondary education are the majority of the applicants

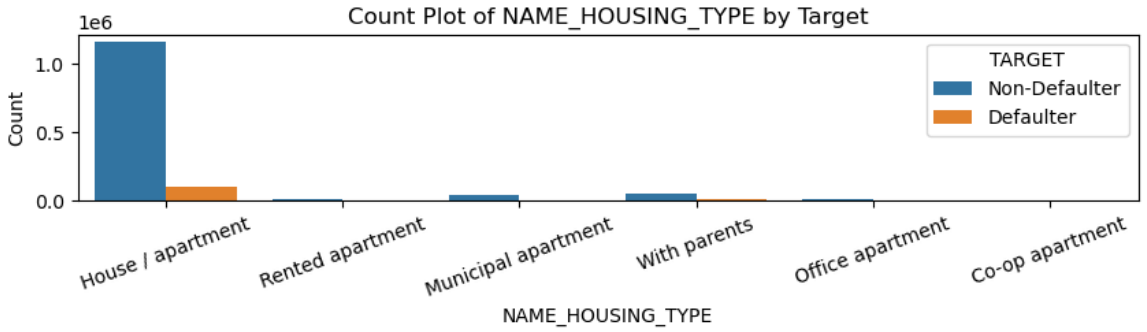
Married clients are the most common applicants for loans, regardless of whether they are defaulters or non-defaulters.Among defaulters, clients with a single relationship status appear to have a lower risk of default compared to other marital statuses.Defaulting clients who are widowed show the least risk of defaulting, indicating that widows have a more stable repayment history.

4)Majority of clients, whether they are defaulters or non-defaulters, reside in either houses or apartments. This suggests that owning a house or living in an apartment is the most common housing situation among both groups.









Thank you!!!