OPTIMIZING LOAN APPROVALS

UNVEILING PATTERNS THROUGH EDA IN RISK ANALYSIS FOR CONSUMER FINANCE

DATE – 21/12/2023

Agenda

1. Business Understanding
2. EDA Process
    - Data Understanding
    - Data Cleaning
    - Data Analysis
3. Recommendation

Meet the presenter

SAURADIP PRADHAN

BUSINESS UNDERSTANDING

In the realm of consumer finance, our company faces the challenge of optimizing loan approval decisions. The absence of robust credit histories often results in difficulties distinguishing creditworthy applicants from potential defaulters. How can we leverage Exploratory Data Analysis (EDA) to uncover patterns, ensuring we approve loans to applicants capable of repayment? This presentation explores Risk Analytics in Banking and Financial Services, aiming to identify patterns of payment difficulties and understand key factors driving loan defaults . Our goal is to empower informed decision-making, minimizing the risk of financial loss due to defaulted loans.

EDA PROCESS
DATA UNDERSTANDING

IMPORT FUNTIONS AND UPLOAD THREE FILES

- Import pandas, NumPy, matplotlib and seaborn.
- ‘application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
- ‘previous_application.csv' contains information about the client’s previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- 'columns_description.csv' is data dictionary which describes the meaning of the variables

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
```

```
In [2]: # upload the files
        ad =  pd.read_csv('application_data.csv')
        pa = pd.read_csv("previous_application.csv")
        cd = pd.read_csv('columns_description.csv',encoding="ISO-8859-1")
```

IMPORT FUNCIONS AND UPLOAD FILES

Name of ‘Application_data ‘ is load as ‘ad’

Name of ‘Previous_data’ is load as ‘pa’

Name of ‘Columns_description’ as ‘cd’

DATA UNDERSTANDING

- **HERE WE EXPLORE THE SHAPE OF THE CSV FILES**

  CODE :

  ad.shape, pa.shape, cd.shape

  Shape of 'ad' is (307511,122)
  Shape of 'pa' is (1670214 , 37)
  Shape of 'cd' is (160 ,5)

- **EXPLORE THE TOP 5 ROWS OF THE COLUMN**

  CODE :

  ad.head() , pa.head() , cd.head()

- **EXPLORE THE COLUMNS OF THE CSV FILES**

  CODE:

  ad.columns , pa.columns, cd.columns

DATA UNDERSTANDING

- **MERGE THE TWO FILES BY A COMMON COLUMN:**

CODE:

merged_df = pd.merge(ad,pa, on='SK_ID_CURR')

- **CHECKING THE DUPLICATE COLUMNS IN THE MERGED COLUMN:**

CODE:

duplicate_columns = merged_df.columns[merged_df.columns.duplicated()]

print(duplicate_columns)

- **CHECKING THE DATA TYPES OF THE MERGED COLUMN**

CODE:

merged_df.dtypes( )

DATA CLEANING

- **CHECKING THE PERCENTAGE OF THE NULL VALUES IN THE MERGED TABLE :**

CODE :

```
merged_df.isnull().sum()/shape_ad[0]*100
```

- **IDENTIFY THE COLUMNS WHICH HAVE MORE THAN 30% NULL VALUES:**

CODE:

```
columns_more_than_30 = isnull[isnull > 30].index
print(columns_more_than_30)
```

- **DELETE THE COLUMNS WHICH HAVE MORE THAN 30% NULL VALUES**

CODE:

```
merged_df = merged_df.drop(columns=columns_more_than_30)
```

DATA CLEANING

- **EXPLORE THE SHAPE OF THE MERGED FILE:**

CODE:

```
merged_df.shape
```

Shape = (1413701 , 87)

- **FILL THE MISSING VALUES:**

CODE:

```
merged_df['AMT_GOODS_PRICE_x'] =
merged_df['AMT_GOODS_PRICE_x'].fillna(mean)
```

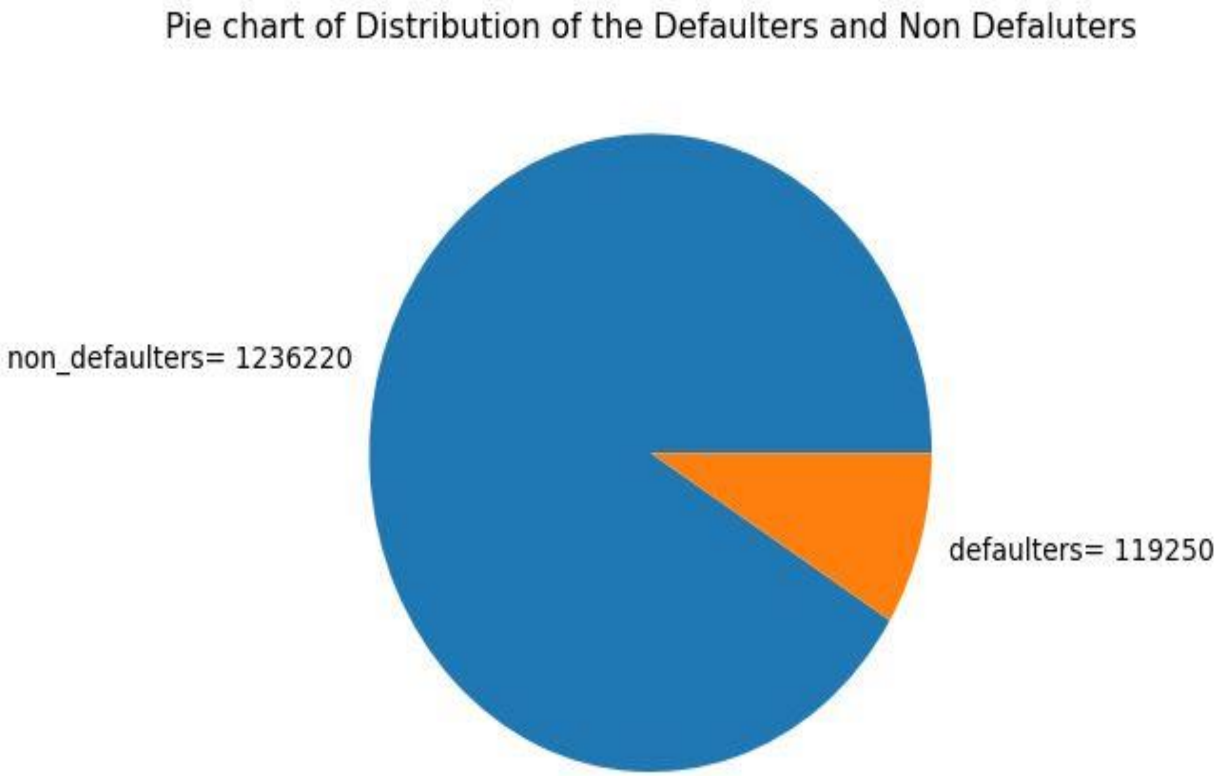
- **DROP THE COLUMNS WHICH ARE NOT NECESSARY:**

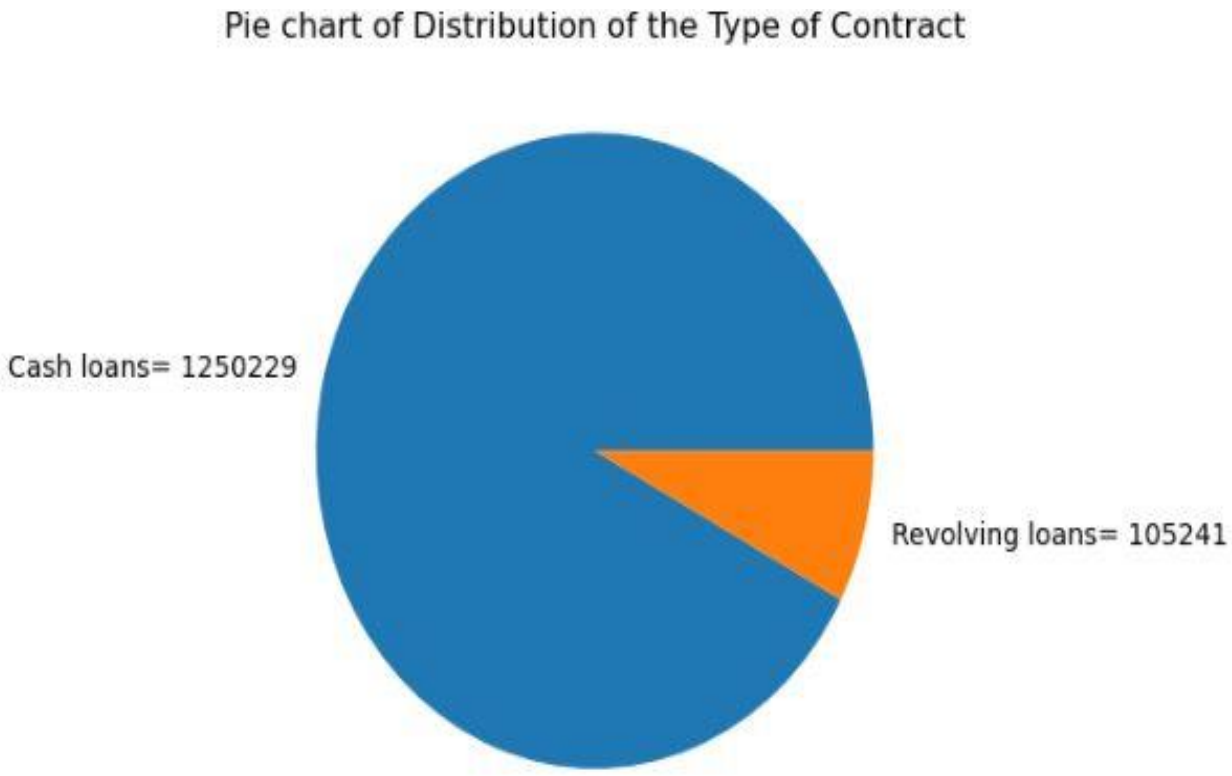
CODE:

```
merged_df = merged_df.drop(merged_df.columns[40:60], axis=1)
```

DATA ANALYSIS

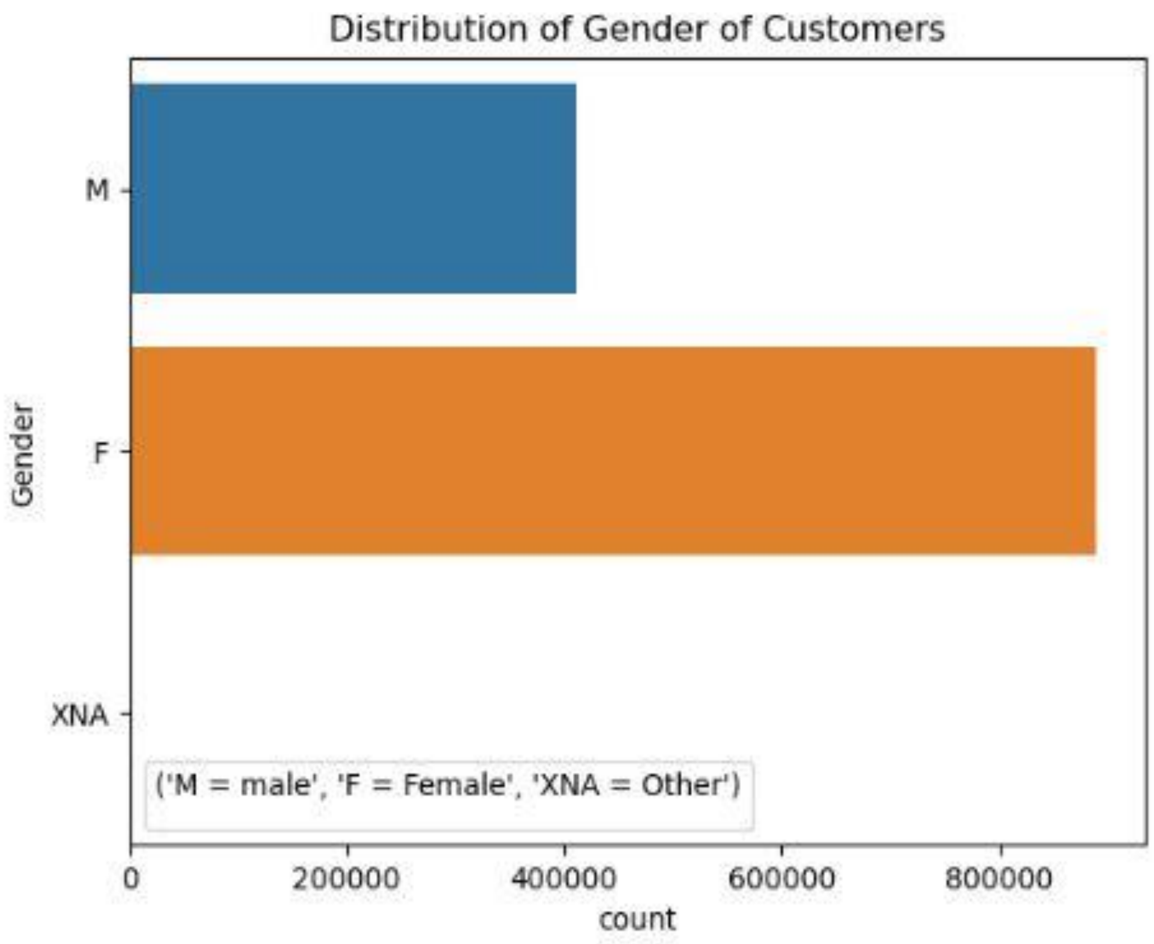
UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS

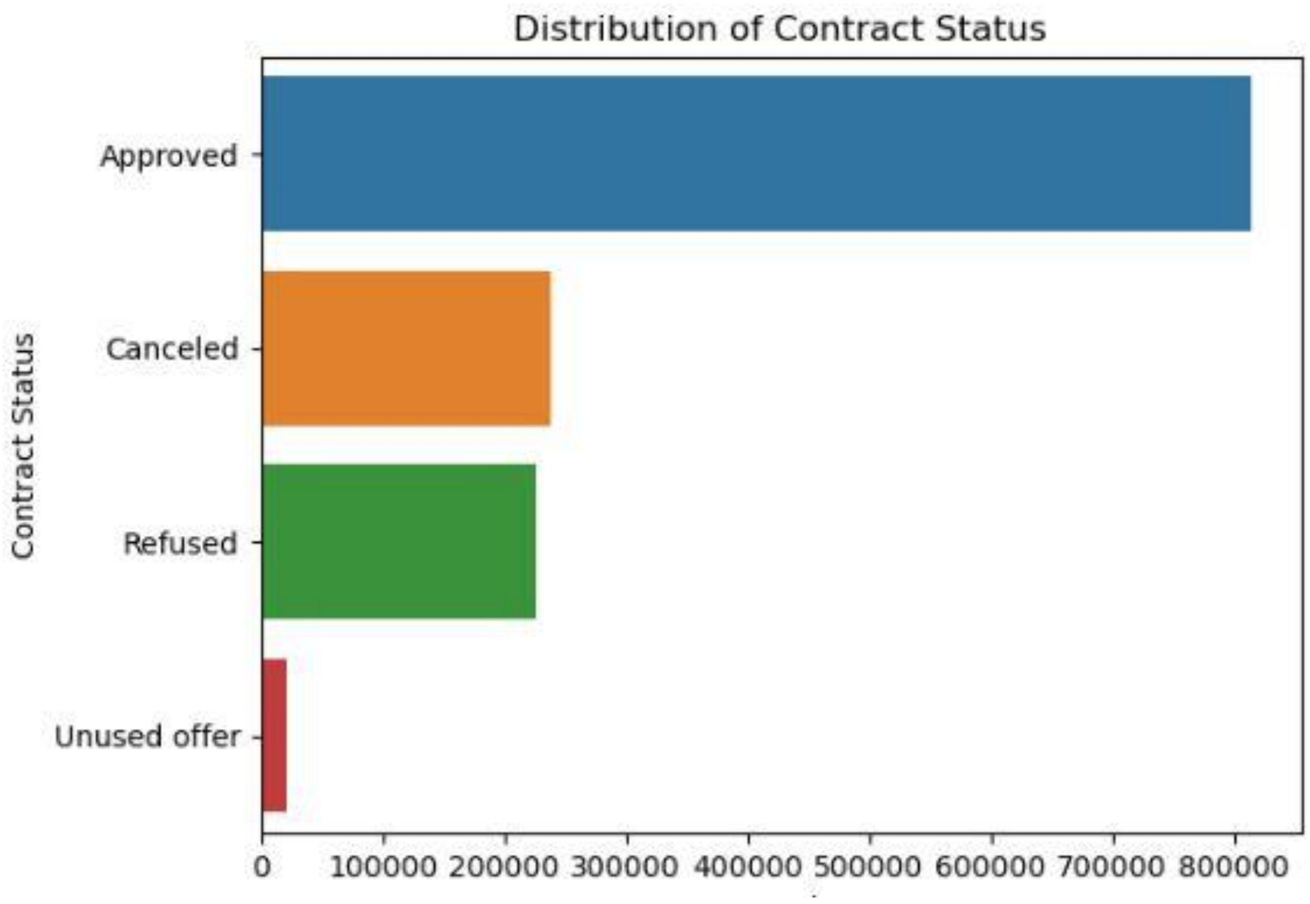




DATA ANALYSIS

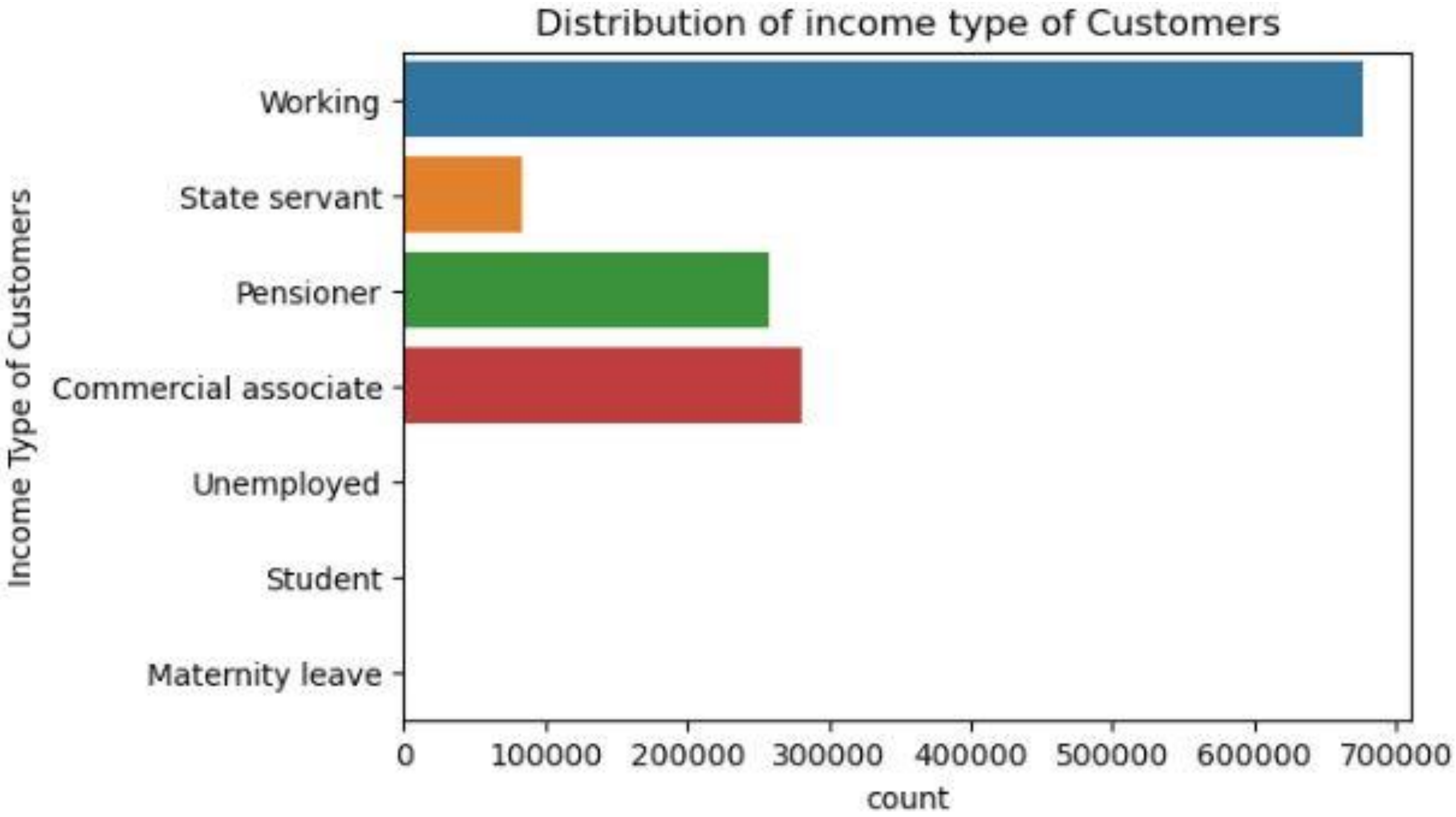
UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS





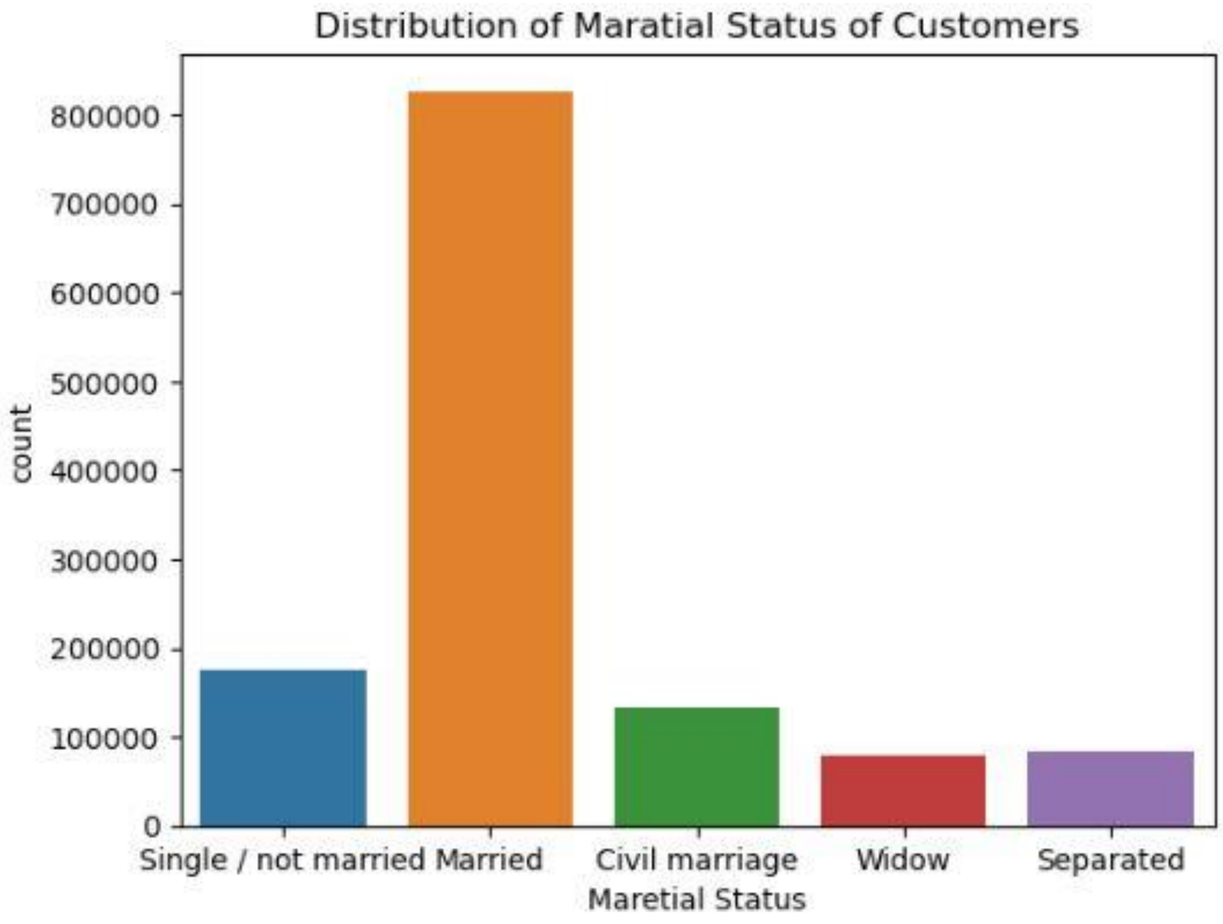
DATA ANALYSIS

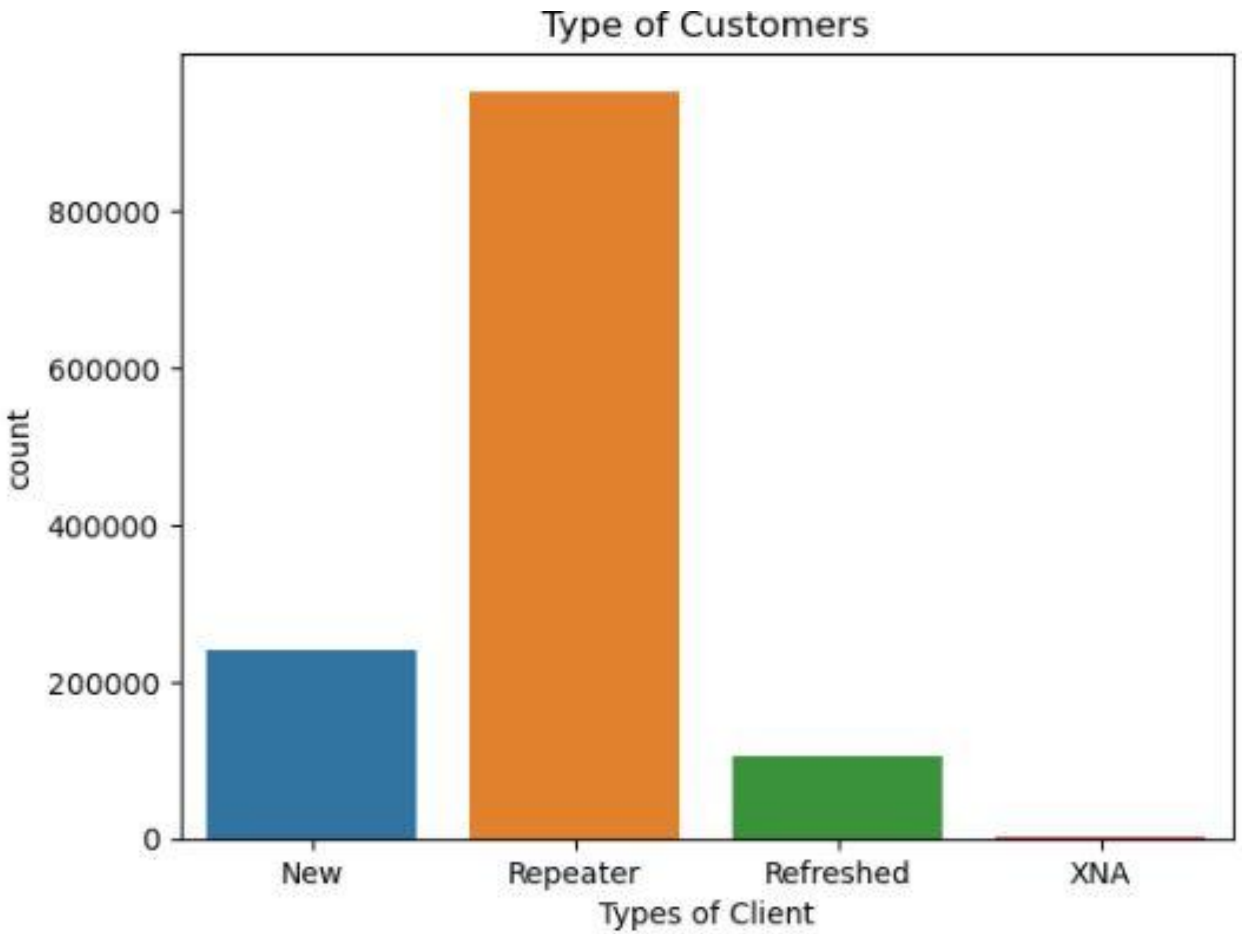
UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS



DATA ANALYSIS

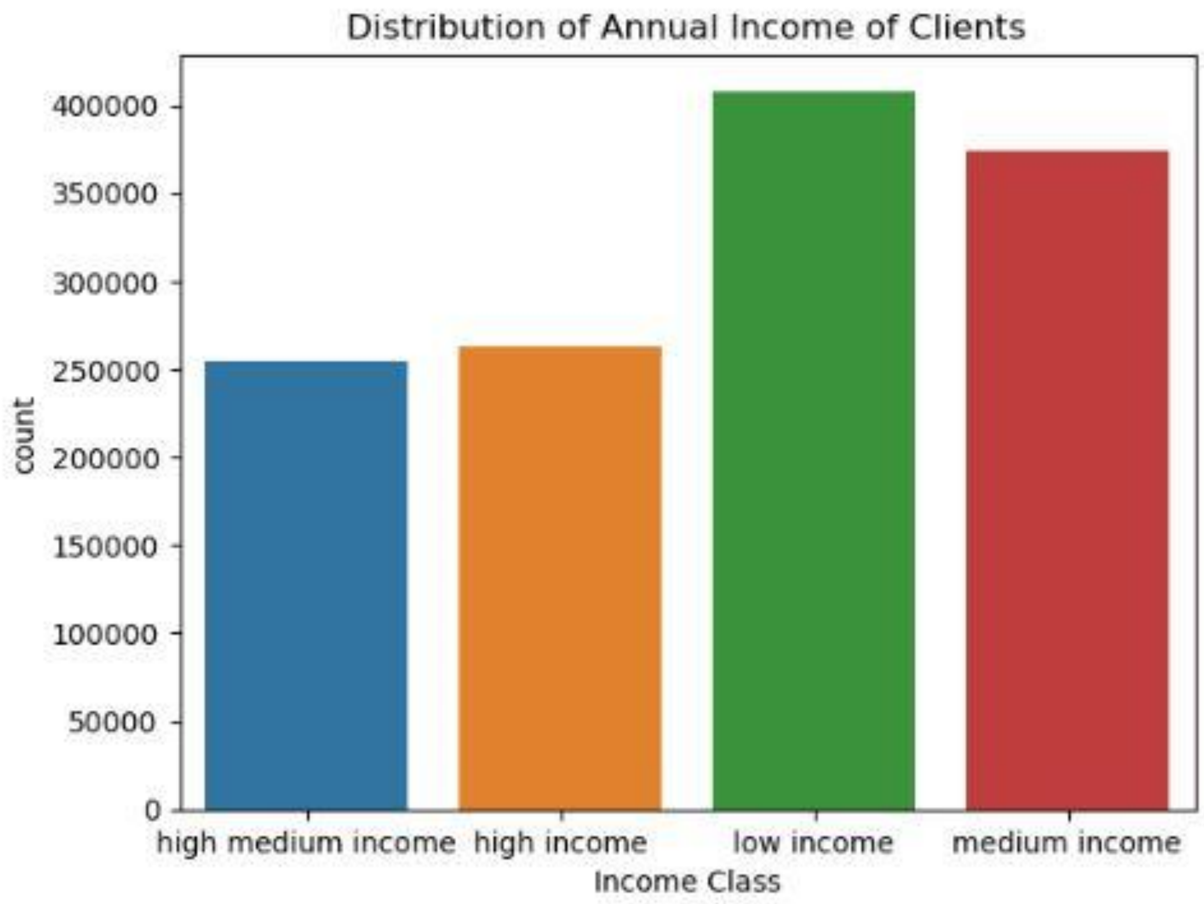
UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS

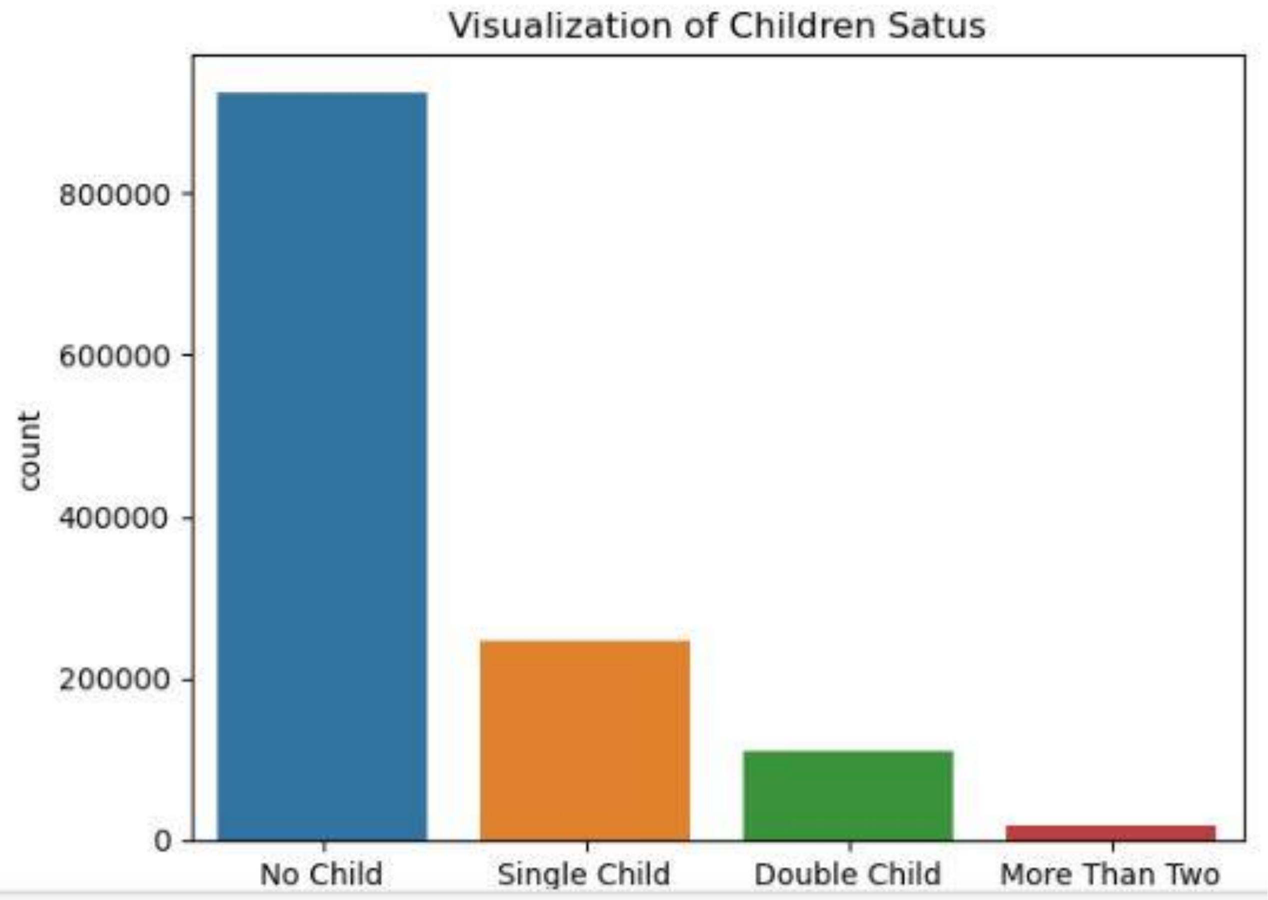




DATA ANALYSIS

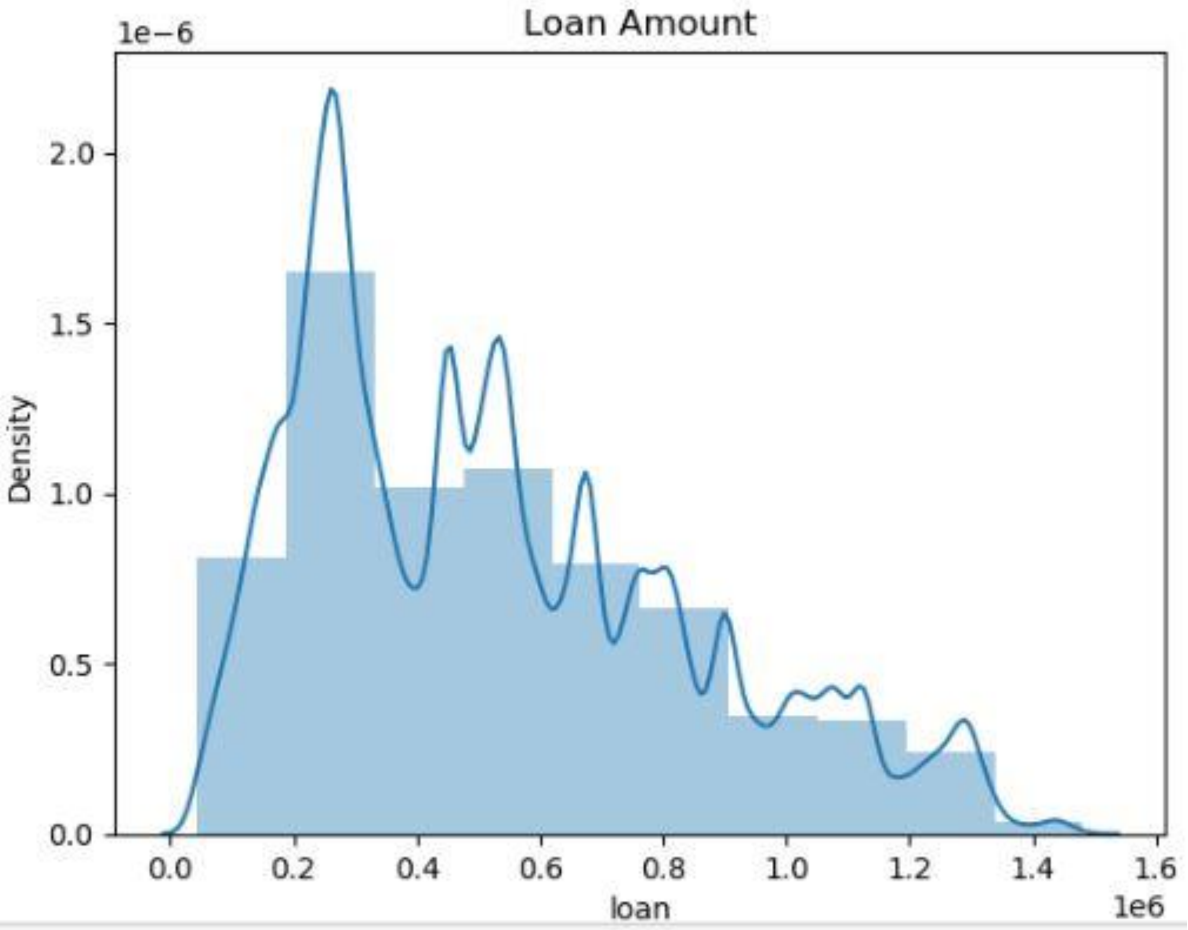
UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS

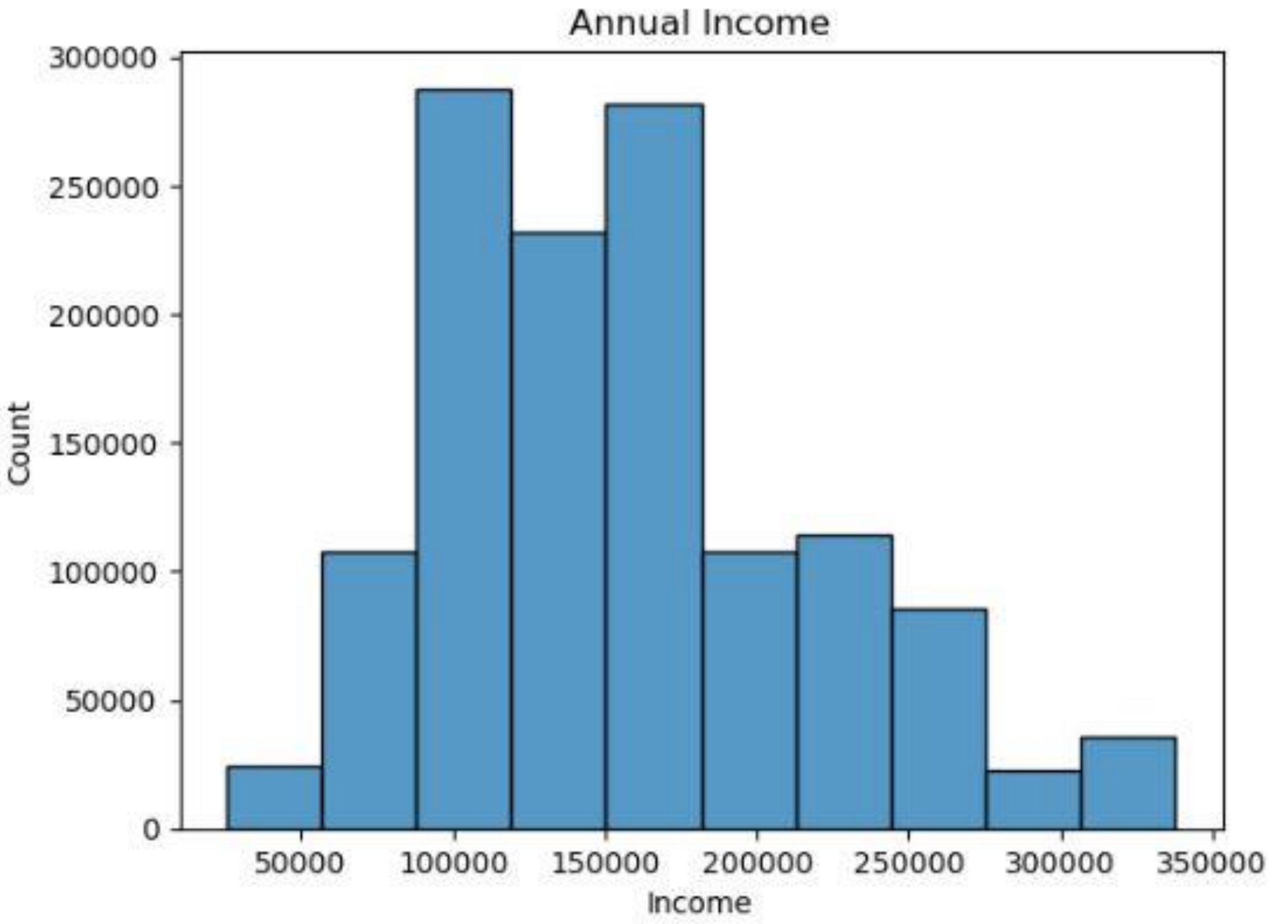




DATA ANALYSIS

UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS

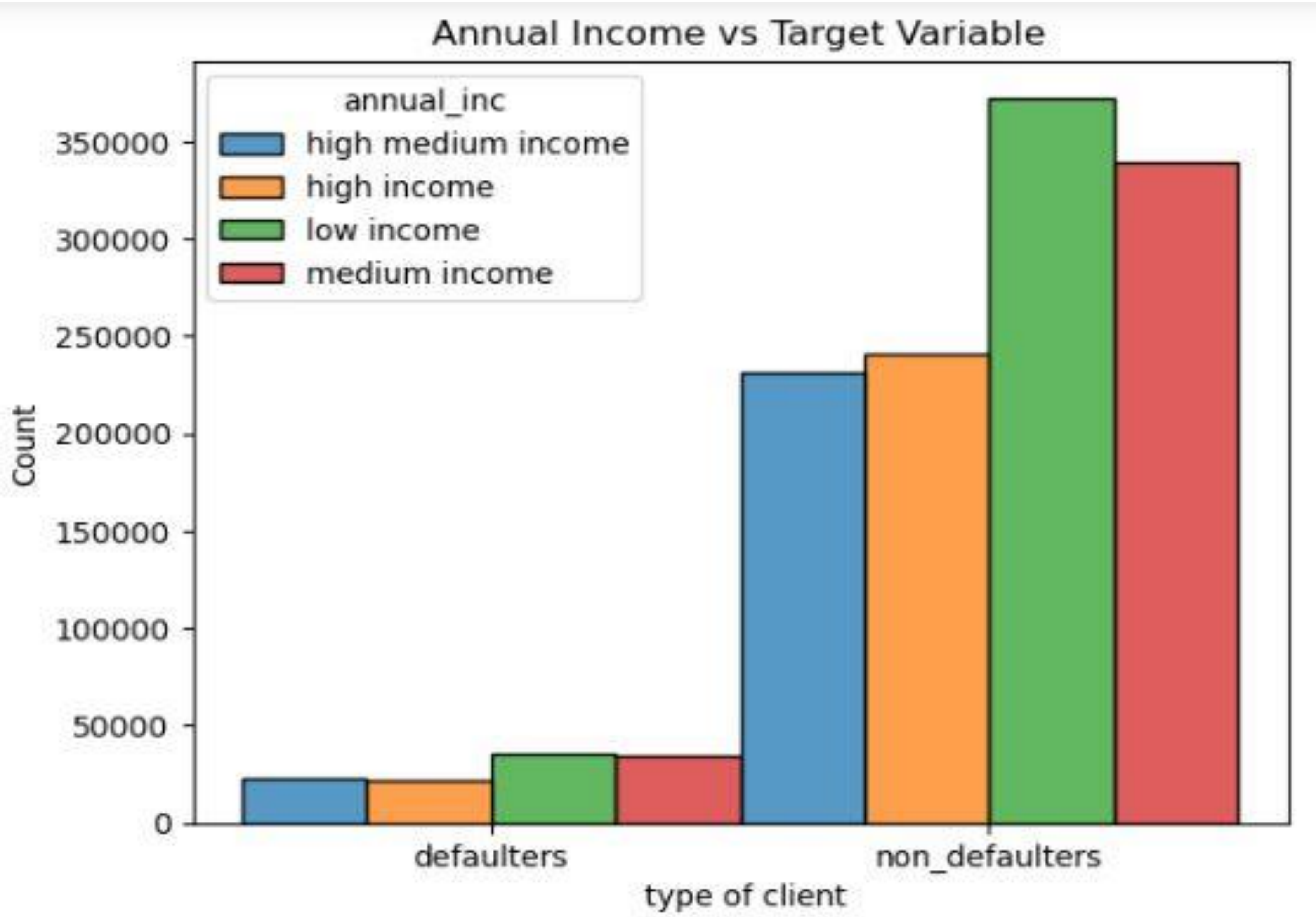




DATA ANALYSIS

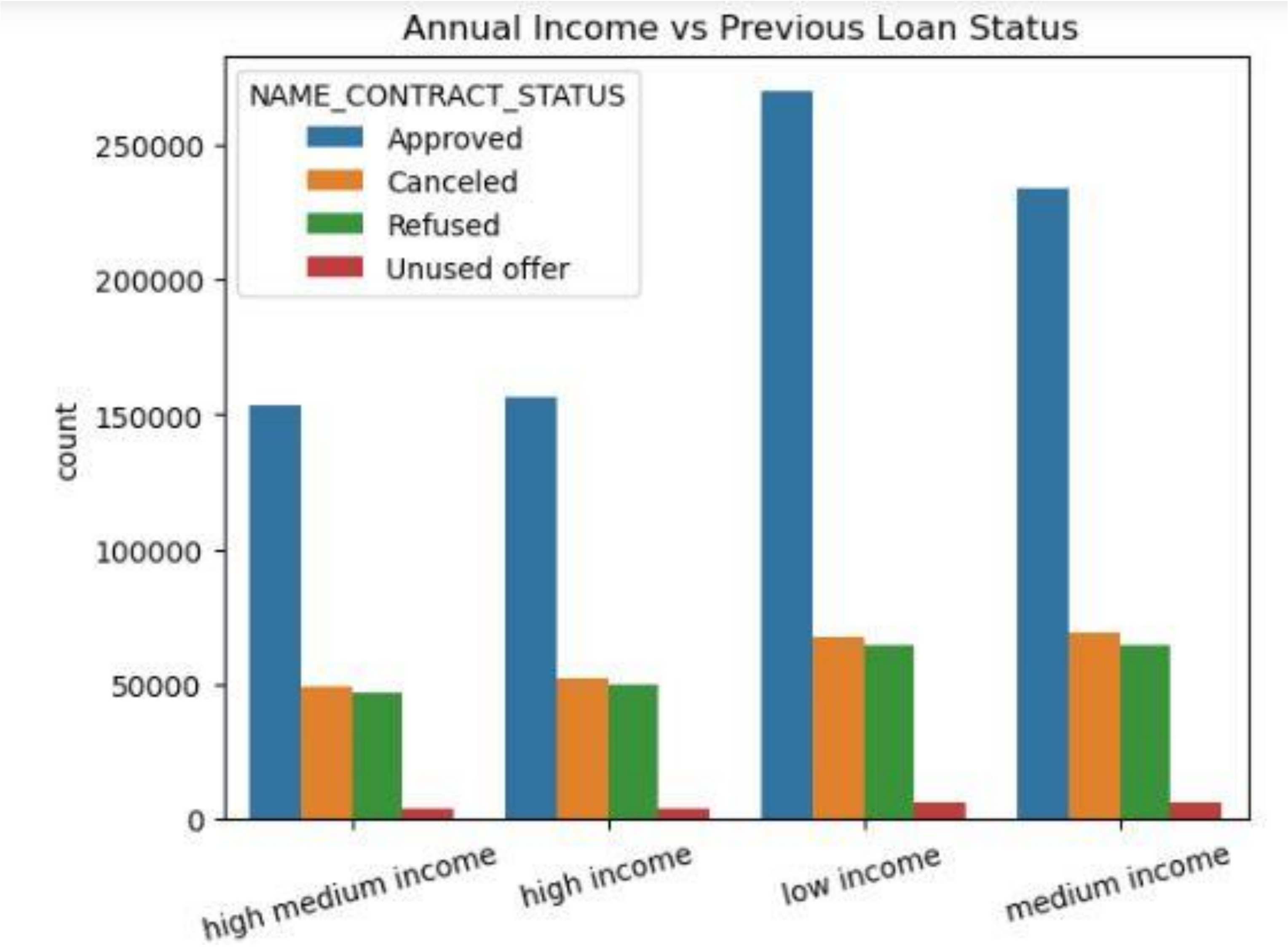
BIVARIATE ANALYSIS OF IMPORTANT COLUMNS

FIVE IMPORTANT DRIVER VARIABLES



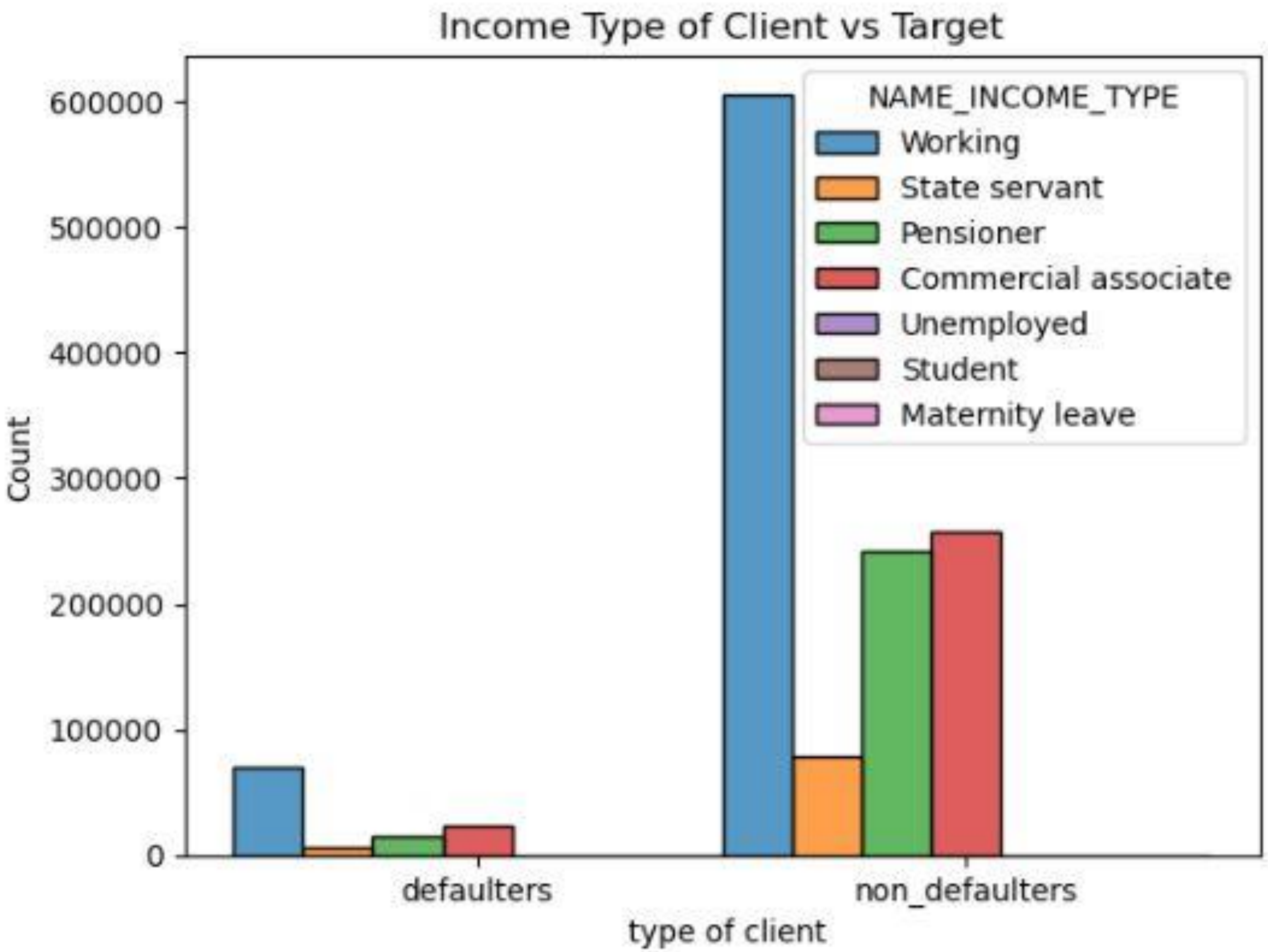
DATA ANALYSIS

BIVARIATE ANALYSIS OF IMPORTANT COLUMNS



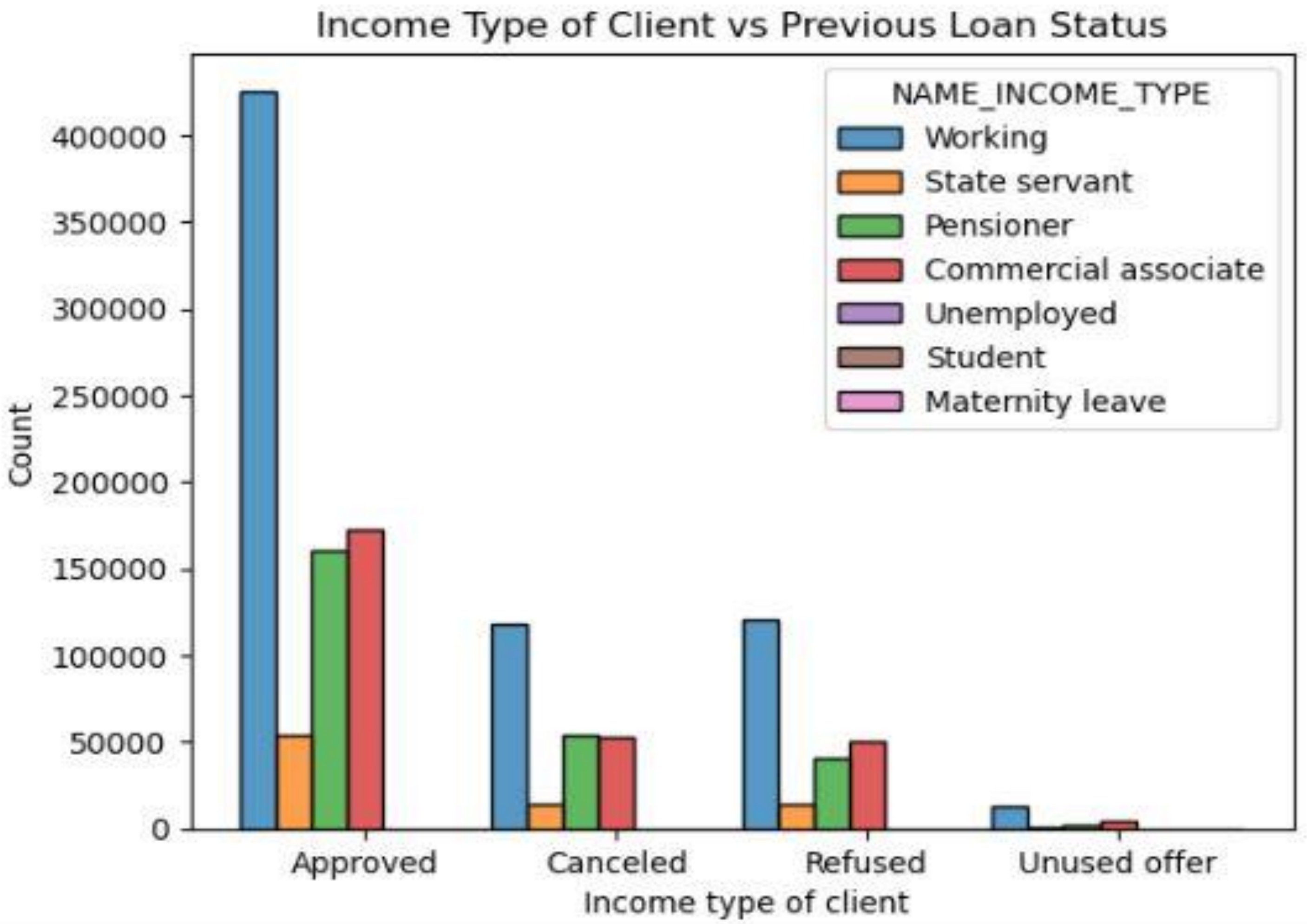
DATA ANALYSIS

UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS



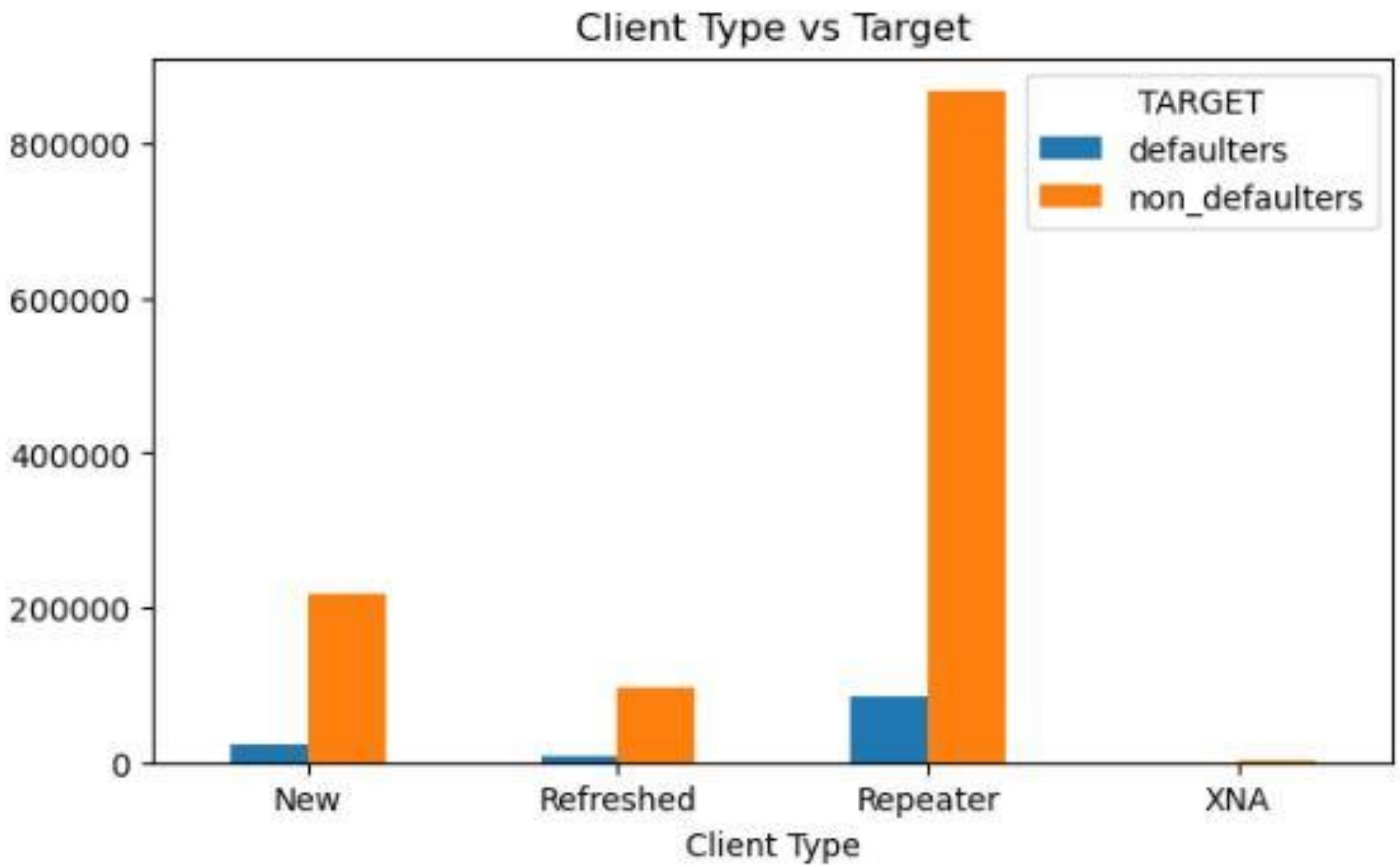
DATA ANALYSIS

UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS



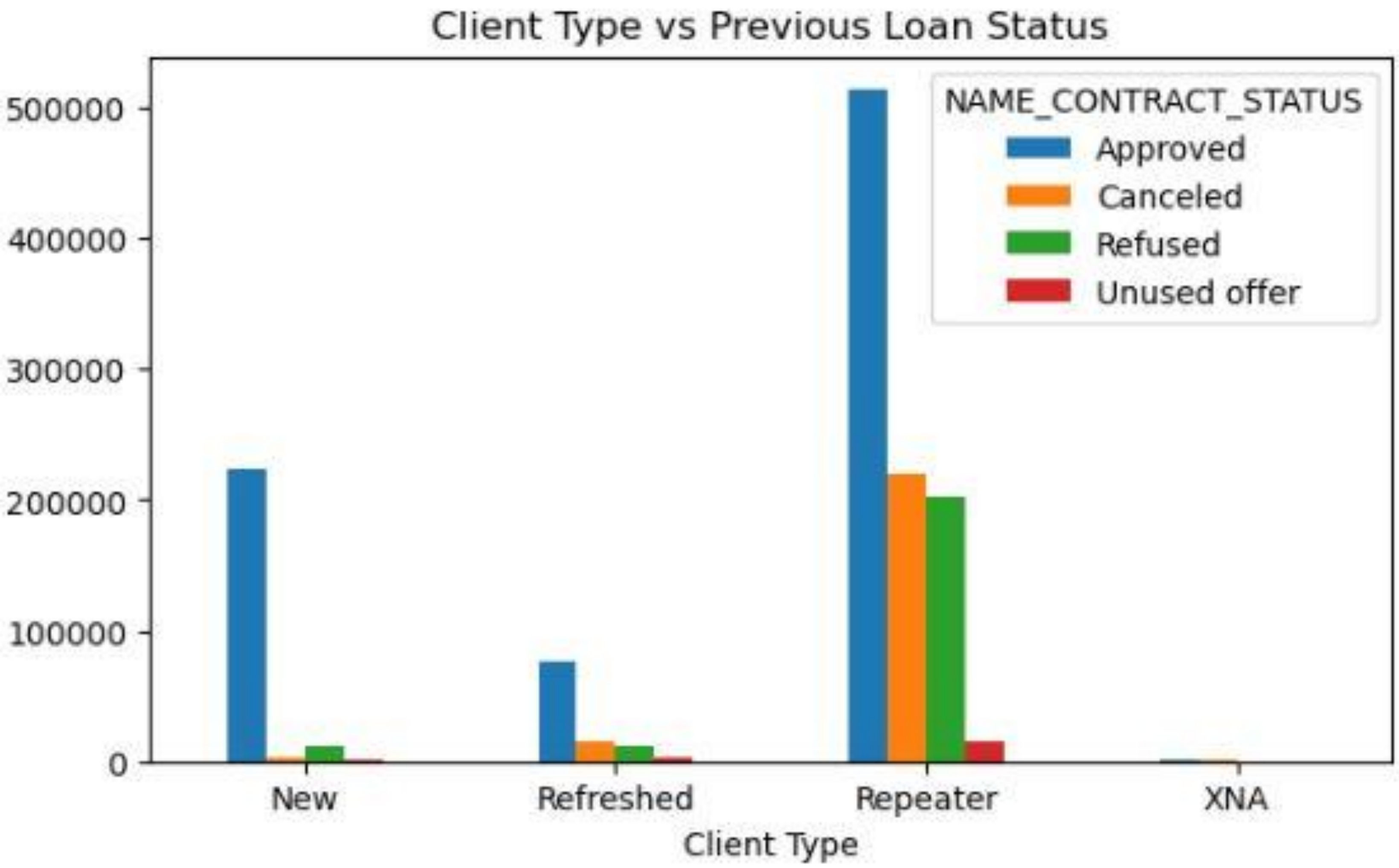
DATA ANALYSIS

UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS



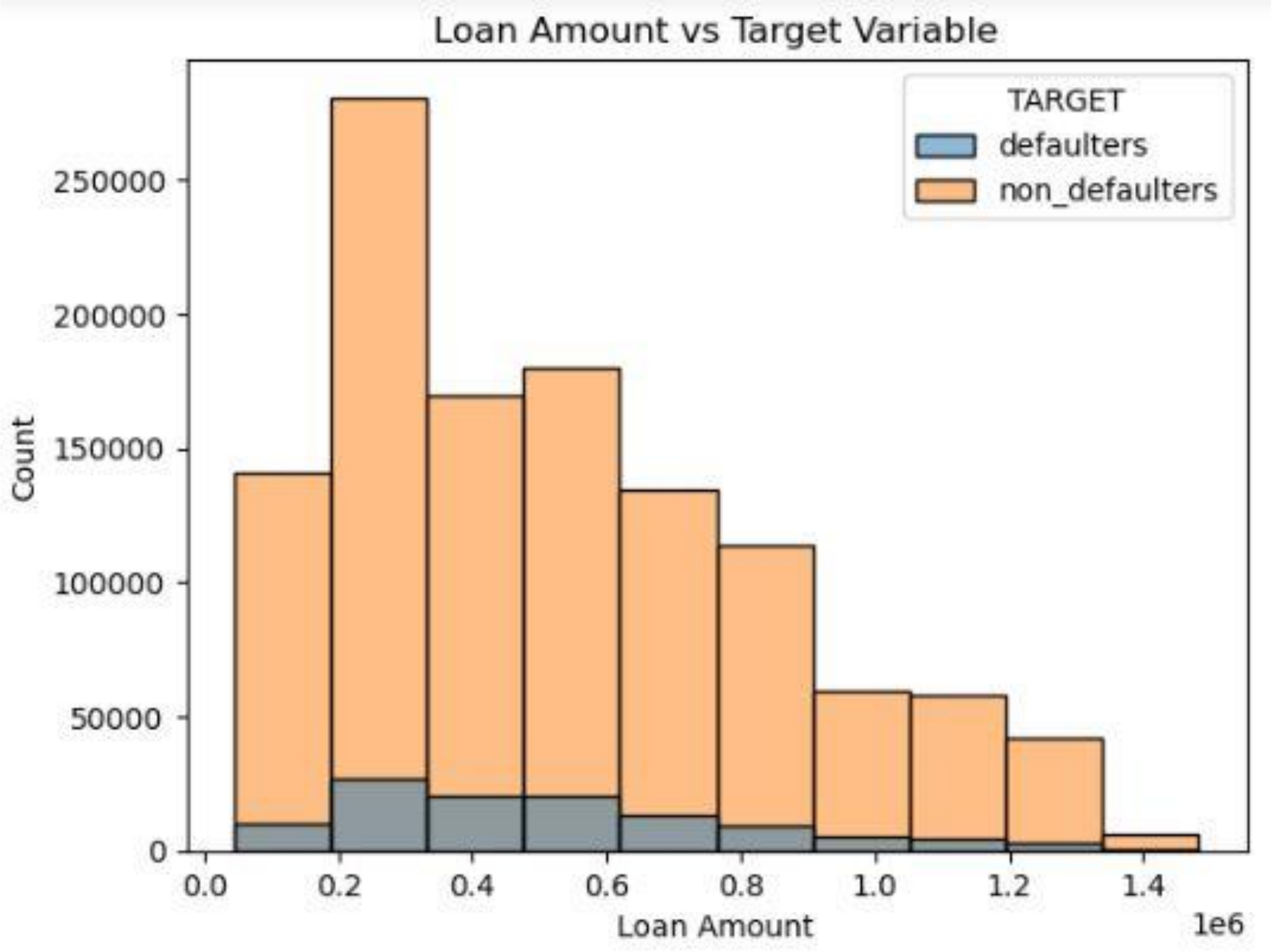
DATA ANALYSIS

UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS



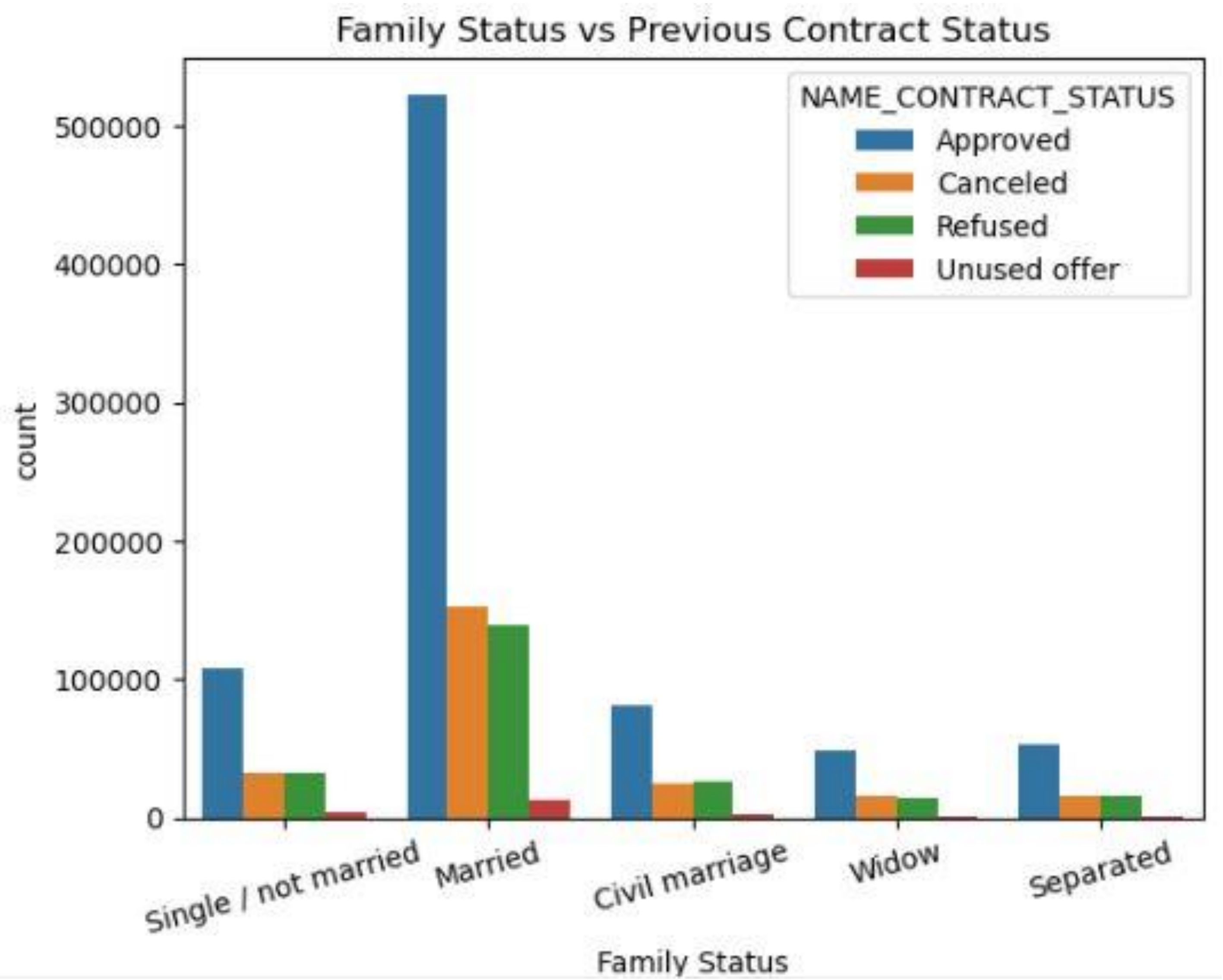
DATA ANALYSIS

UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS



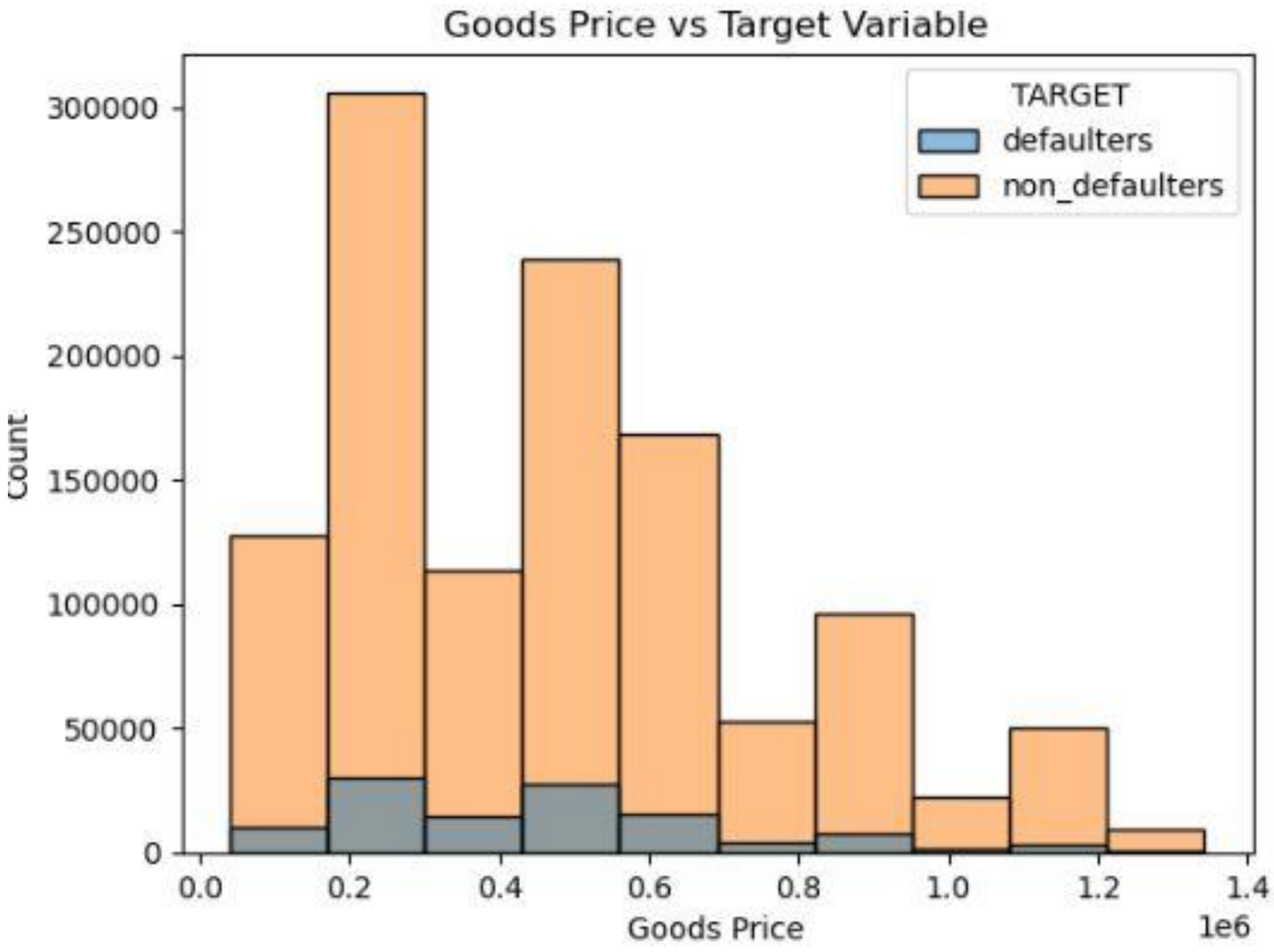
DATA ANALYSIS

UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS



DATA ANALYSIS

UNIVARIATE ANALYSIS OF IMPORTANT COLUMNS



RECOMMENDATION

1. **Defaulters vs. Non-Defaulters:**
    1. The number of defaulters is notably low compared to non-defaulters.
2. **Gender Distribution:**
    1. A majority of clients are female.
3. **Income Type:**
    1. The predominant income type among clients is "working."
4. **Previous Loan Approval:**
    1. Most clients have had their previous loans approved. The number of clients canceling or refusing loans is minimal and comparable, both being low compared to approved loans.
5. **Marital Status:**
    1. The majority of clients are married.
6. **Client History:**
    1. A significant portion of clients consists of repeat customers.
7. **Annual Income Distribution:**
    1. Clients with low to medium annual income levels are the majority.
8. **Income Level and Defaulters:**
    1. Non-defaulters are prominently found among clients with low annual income, with the second-highest among those with medium income.

RECOMMENDATION

1.Income Type and Previous Loan Approval:

1. Clients with the income type 'working' show a higher likelihood of being non-defaulters, and this group also has a higher rate of previous loan approvals. Commercial associates are the second-highest in this regard.

2.Repeat Clients and Loan Approval:

1. Repeat clients are predominantly non-defaulters, with a history of previous loan approvals.

3.Marital Status and Loan Approval:

1. Married clients have a higher likelihood of previous loan approvals, with single clients in the second position.

4.Loan Amount for Non-Defaulters:

1. Most non-defaulters have loan amounts ranging from 0.2 to 0.6 million.

5.Goods Price and Defaulters:

1. Defaulters tend to have goods prices ranging from 0.4 to 0.6 million, while non-defaulters predominantly fall within the 0.2 to 0.3 million range.

These insights provide a comprehensive understanding of the risk factors associated with the dataset, offering valuable information for decision-making in the context of loans and client profiles.