Banking Risk Analytics

By: Sudhamsha Sagar

25 May, 2025

Problem Statement

This case study focuses on identifying patterns that indicate whether a client is likely to face difficulties in repaying their loan instalments. The objective is to leverage Exploratory Data Analysis (EDA) to uncover key factors that influence loan default. These insights can help a bank make informed decisions, such as rejecting high-risk applications, adjusting loan amounts, or offering loans at higher interest rates to risk-prone applicants.

By understanding the variables most strongly associated with default, the bank can improve its credit risk assessment, minimize financial losses, and ensure that creditworthy clients are not unfairly denied access to loans. Ultimately, the goal is to support data-driven decision-making in loan approval and risk management processes.

Steps Involved

1. Loading the Dataset
2. Initial Explorations
3. Data Cleaning
4. Handling Missing Values and Outliers
5. Data Visualisation
6. Summary

Tools Used

1. **Python**
2. **Numpy**
3. **Pandas**
4. **Matplotlib**
5. **Seaborn**

Data Overview

Shape of Application.csv

307511 Rows

122 Columns

Shape of Previous_Data.csv

1670214 Rows

37 Columns

Major Data Type

float 64, int 64

```
# Checking the shape of the dataset
df.shape
```

```
(307511, 122)
```

Data contains 307511 rows and 122 columns

```
# Checking the overview of the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

Two Data's were available

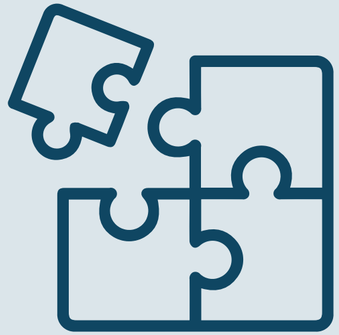
1. Previous_Data.csv
2. Application_Data.csv

Both are handled separately and merged by the end for the final analysis

Data Cleaning and Pre Processing

Handling Missing Values

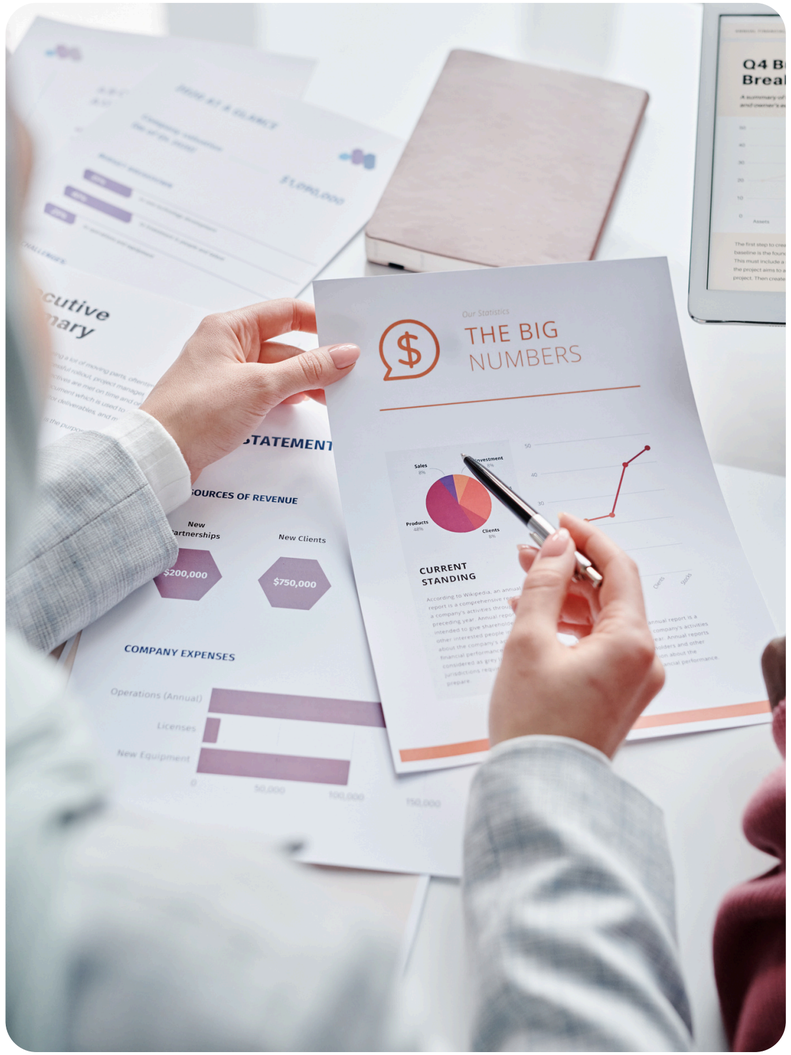
- Removed columns with >40% missing values.
- Imputed missing values using median/mean depending on distribution.



Handling Outliers

- Checked the number of outliers in all the columns and selected the required ones
- Handled outliers by clipping to 1st and 99th percentiles.

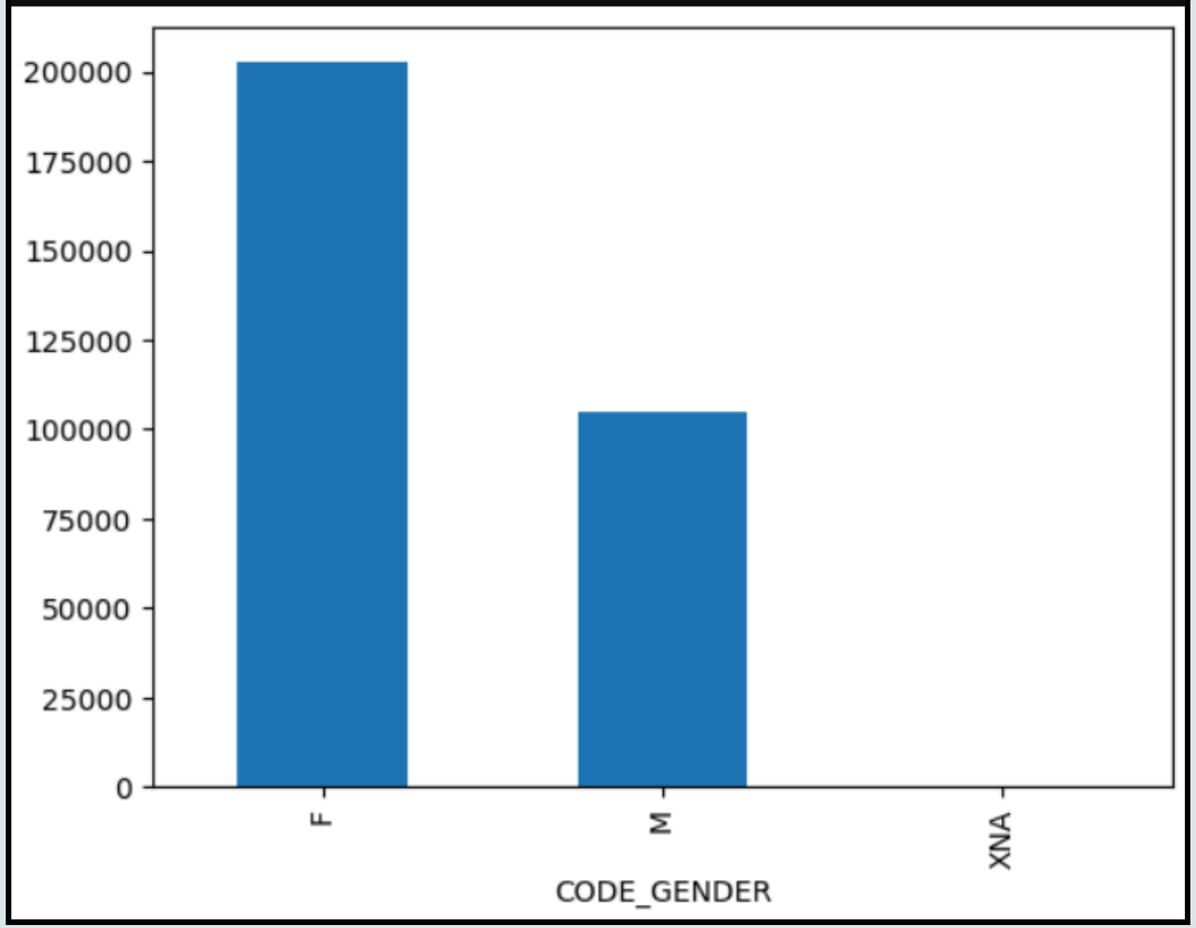
Univariate Analysis



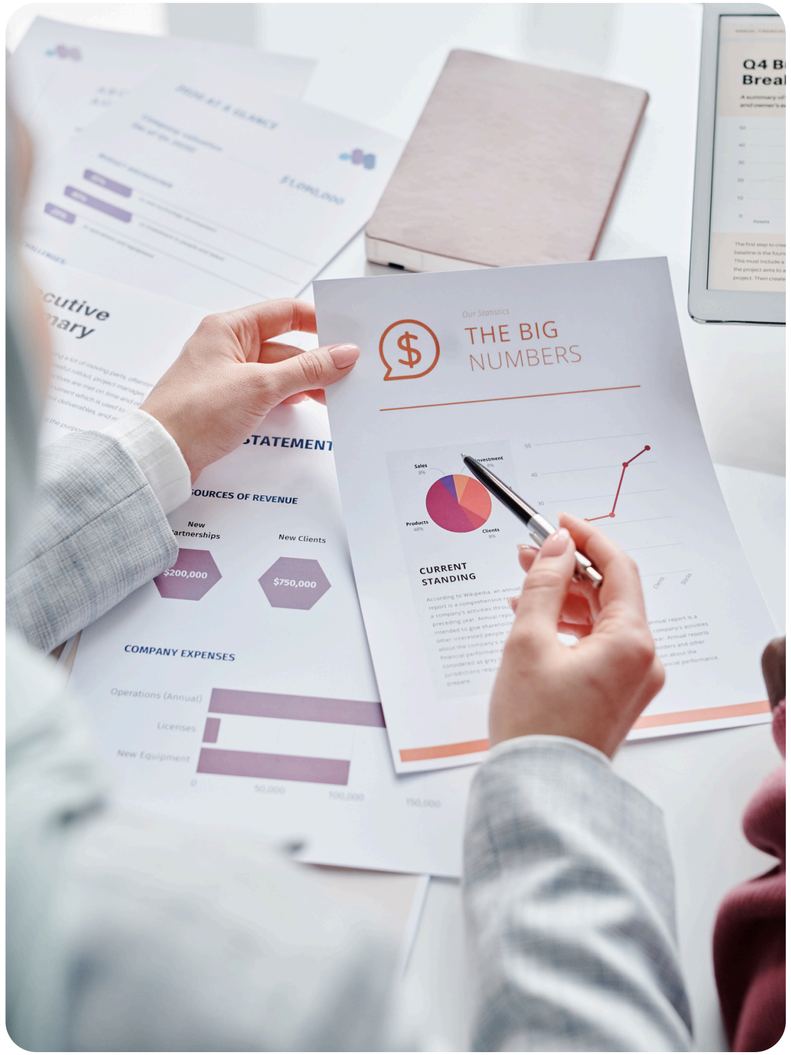
Gender Analysis

- Majority of applicants are female.

```
df.CODE_GENDER.value_counts().plot.bar()
plt.show()
```



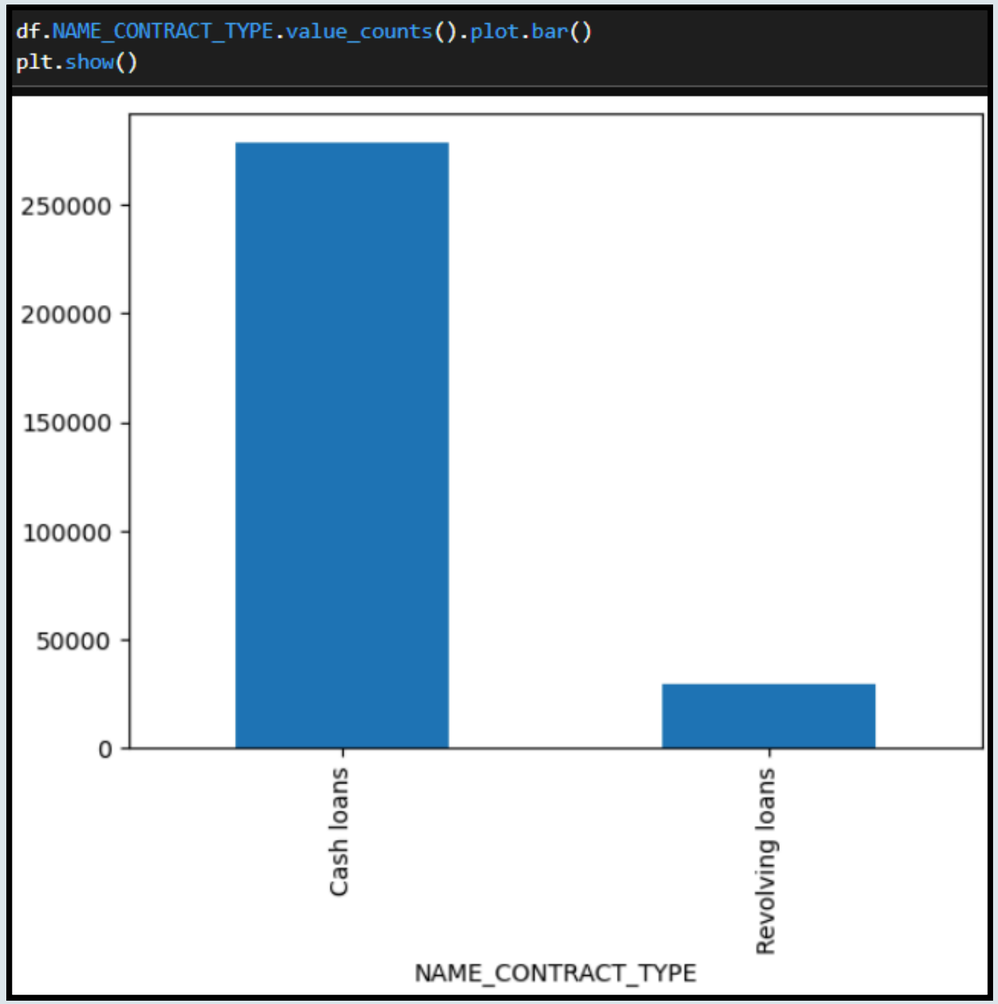
Univariate Analysis



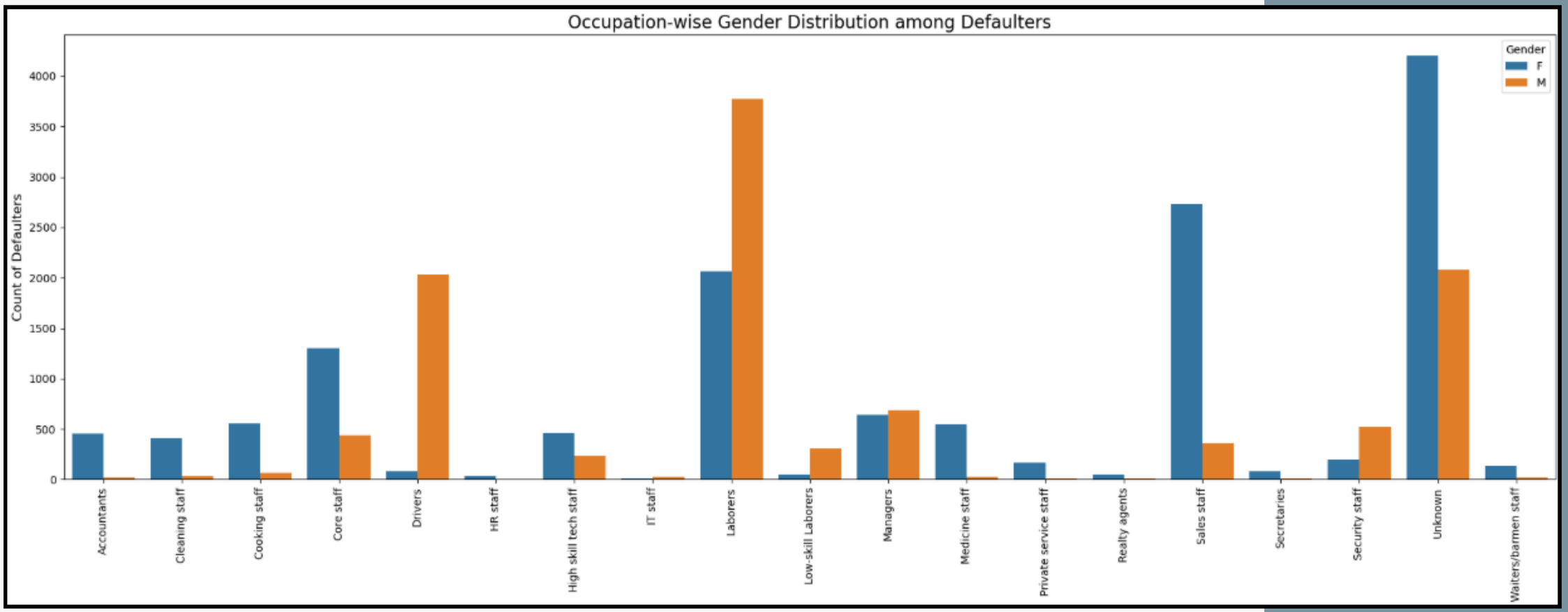
Loan Type Analysis

- Majority of loan type is of Cash Loans

```
df.NAME_CONTRACT_TYPE.value_counts().plot.bar()
plt.show()
```

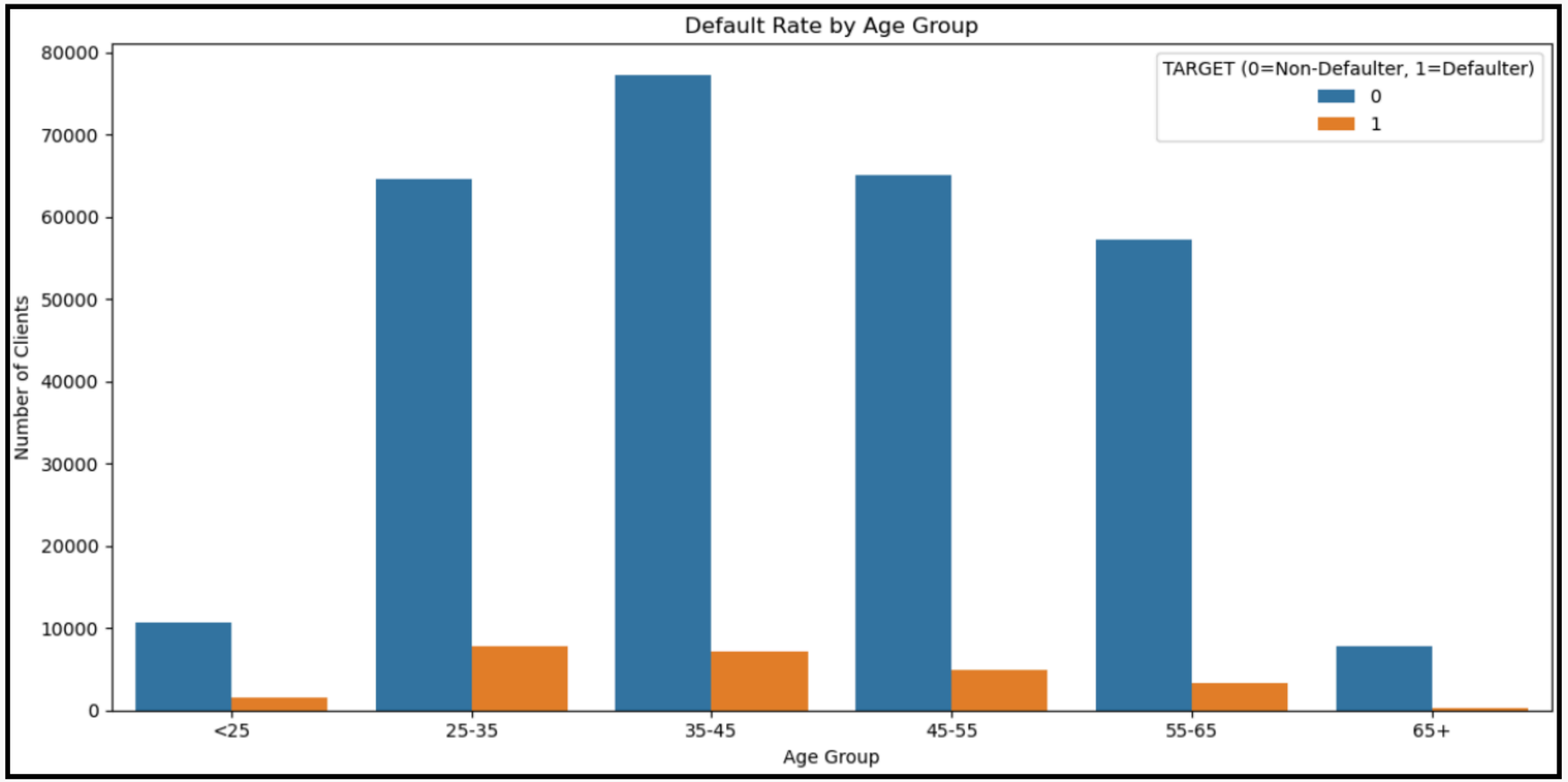


Gender wise Occupation V/s Target



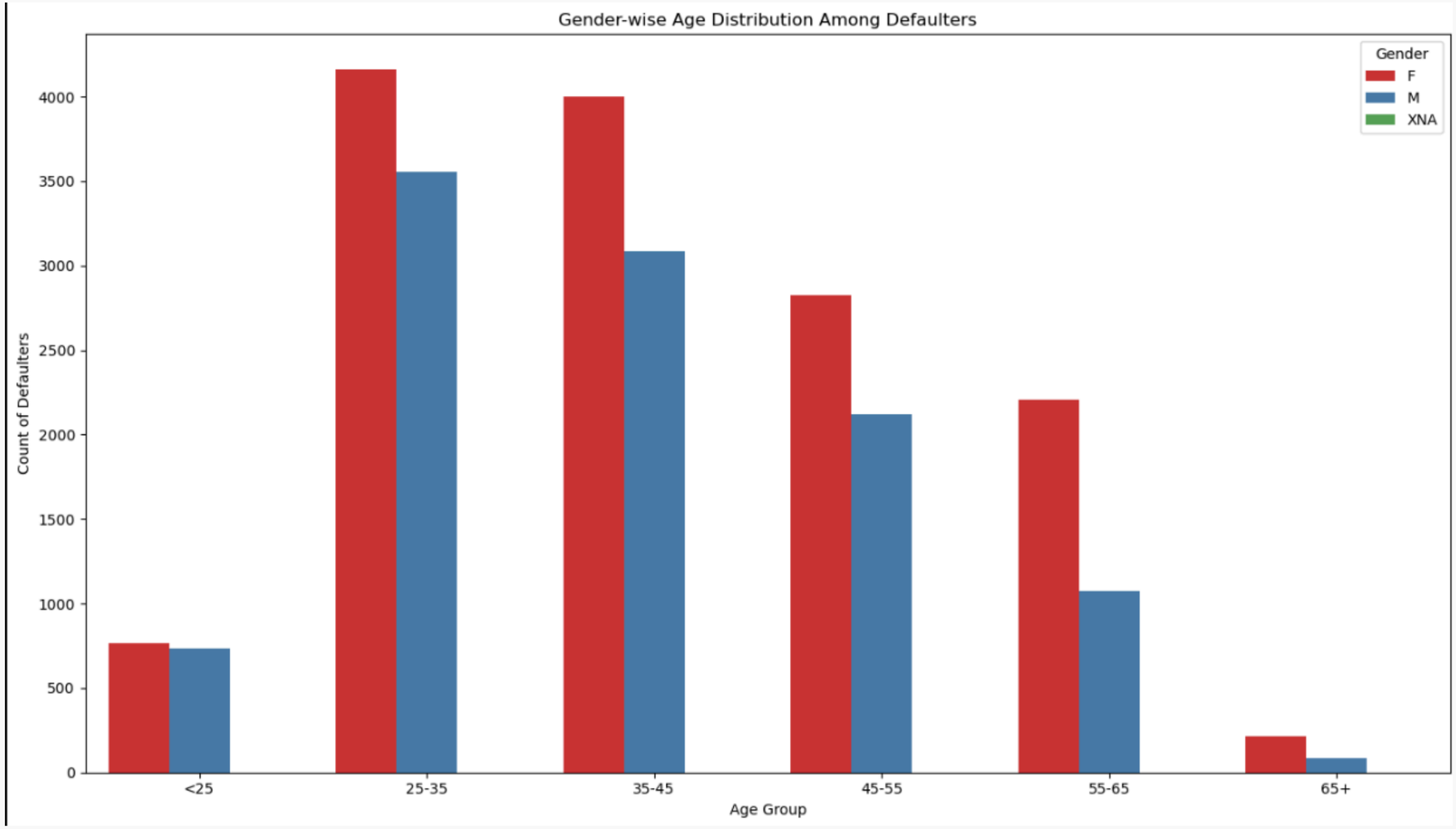
1. **Male who are labourers are more defaulters**
2. **Females who's occupation is not known are more defaulters**
3. **Female Sales Stafs's are more defaulters**

Age Group V/s Target



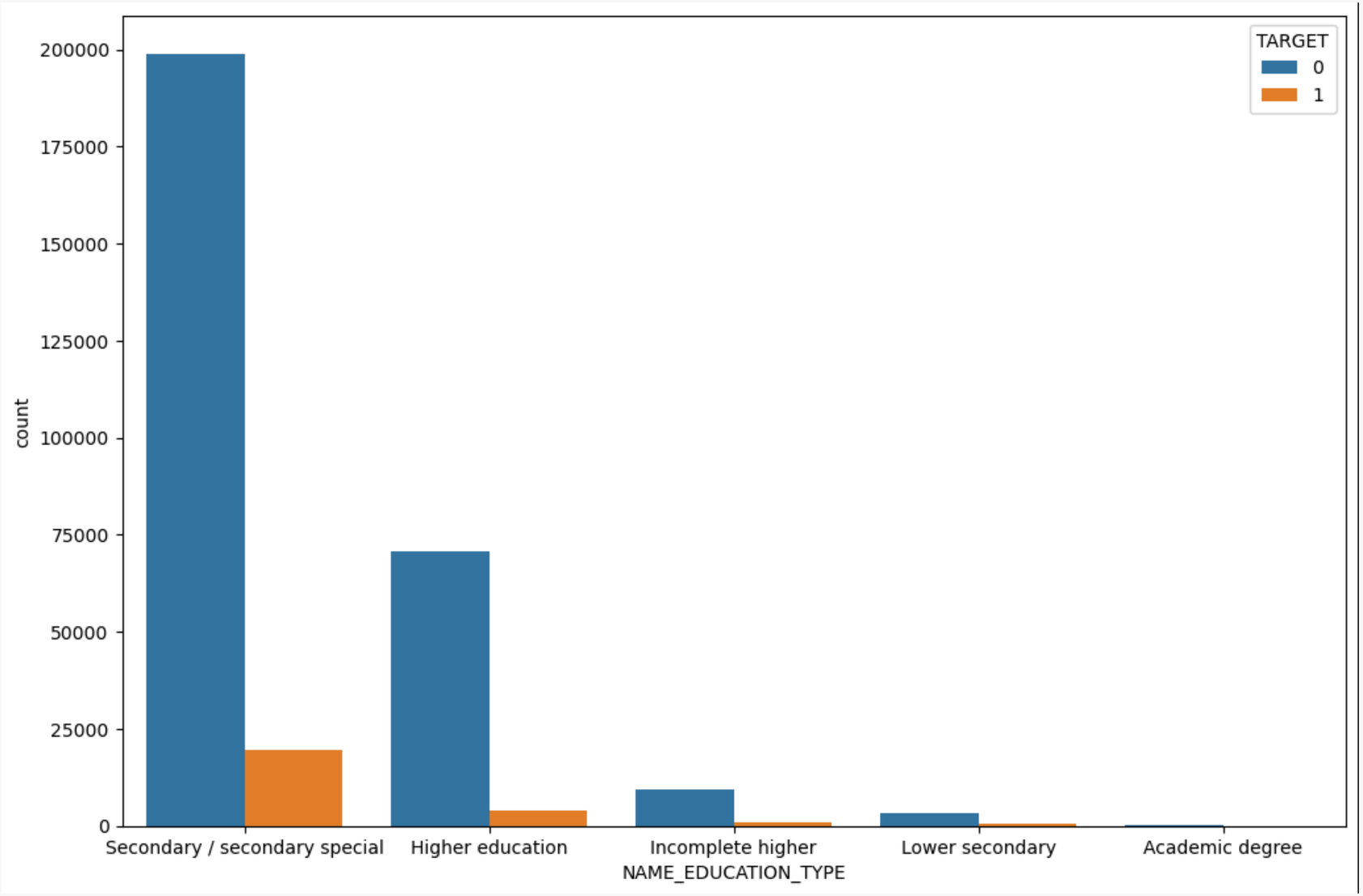
1. **25-35 Age group people are more defaulters**
2. **65+ Age group people are less defaulters**

Gender Wise Age Group V/s Target



1. Females are more defaulters in every age group, more specifically in 25-35 age group

Education V/s Target



1. Most defaulters have secondary education.
2. Applicants with an academic degree are least likely to default.

Conclusion

- Majority of the loan applicants are females.
- Cash loans are more common and have higher default rates.
- Female applicants are more likely to default than males.
- Unknown occupation and laborers show higher default tendencies.
- Females with unknown occupation or in sales roles have higher default rates.
- People without cars and those owning an apartment are more likely to default.
- Age group 25–35 shows the highest default rate.
- Females aged 25–35 are more likely to default compared to males.
- Applicants with secondary education default the most; those with academic degrees default the least.
- Income group ₹50k–₹100k has the highest number of defaulters, followed by <₹50k income group.
- Clear patterns emerged between demographics, financials, and default behavior.

Thank you