CREDIT EDA ASSIGNMENT

Pratik P Punyawant

PROBLEM STATEMENT

- ✓ **Loan providers struggle to lend to individuals with insufficient or non-existent credit histories, leading some consumers to default.**
- ✓ **Working for a consumer finance company specializing in various urban loans, your task is to use EDA to analyze data patterns.**
- ✓ **This analysis will help ensure that capable applicants are approved, reducing the risk of default and improving loan approval accuracy.**

ASSUMPTIONS MADE

- **Focus on columns with less than 50% null values: Prioritize columns with higher data availability.**
- **Drop columns with over 50% null values: Remove columns with significant missing data.**
- **Ensure a cleaner dataset: Enhance data quality by reducing noise and inconsistencies.**
- **Improve model predictions: Increase the reliability of models in identifying potential loan defaulters.**

METHODOLOGY

Analysing the Given Data. → Assumptions for the data. → Cleaning the data. → Treating the missing values.

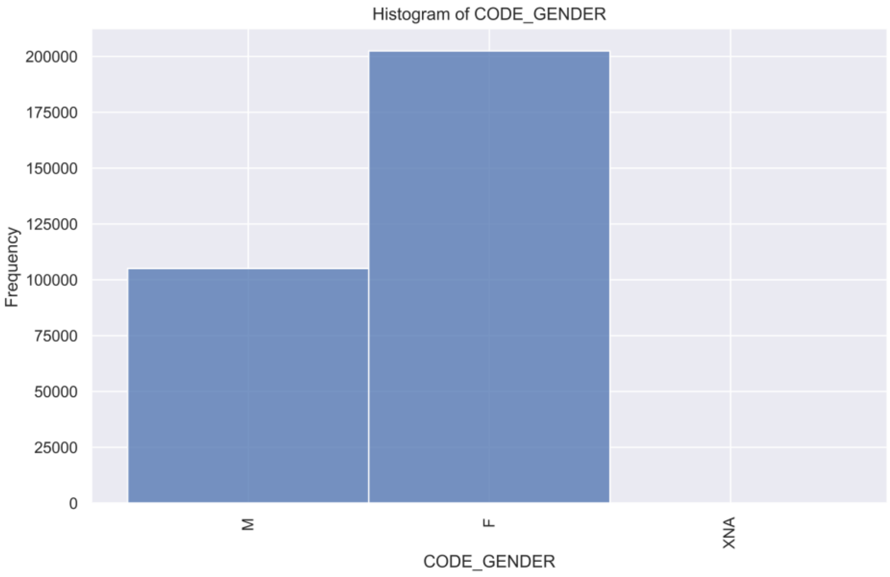
↓

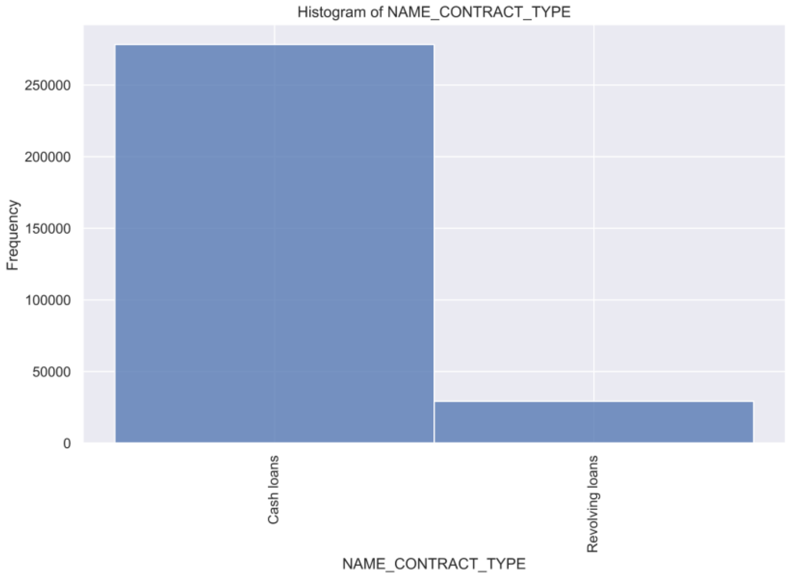
Bivariate Analysis ← Merging the data. ← Finding out the outliers. ← Univariate Analysis.

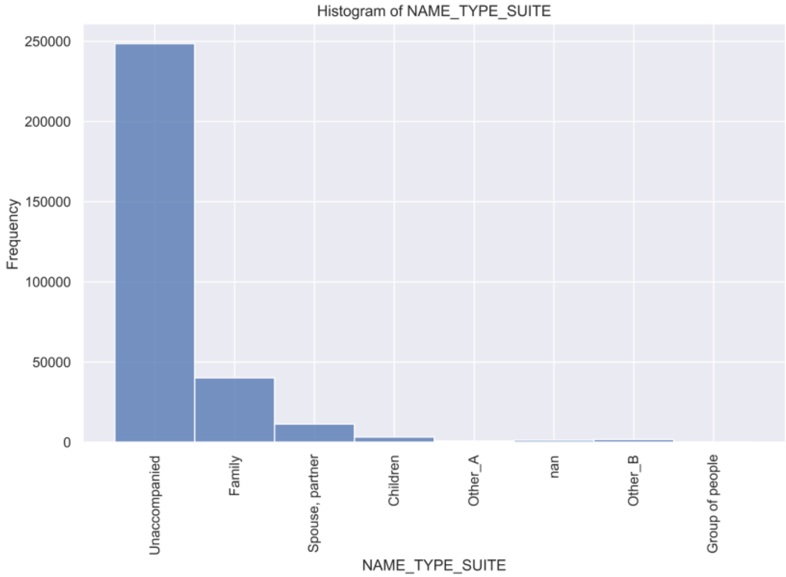
APPROACH

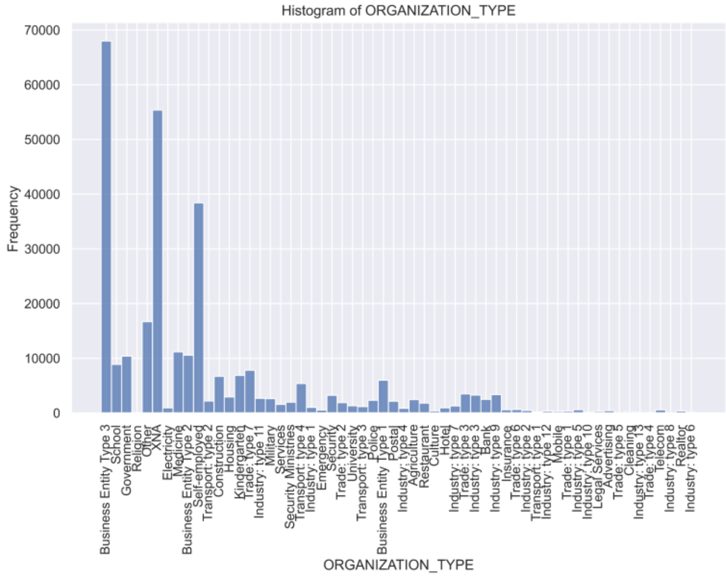
- Importing all the necessary datasets like NumPy, Pandas, etc.
- Dropped the columns as per the assumption.
- Clean the data for each columns by filling mean, median, mode.
- Done the Univariate analysis for both categorical and numerical data.
- Find the correlation and defaulter correlation in the data
- Finding the outliers.
- Done the Bivariate analysis.

ANALYZING CATEGORICAL COLUMNS THROUGH GRAPHS



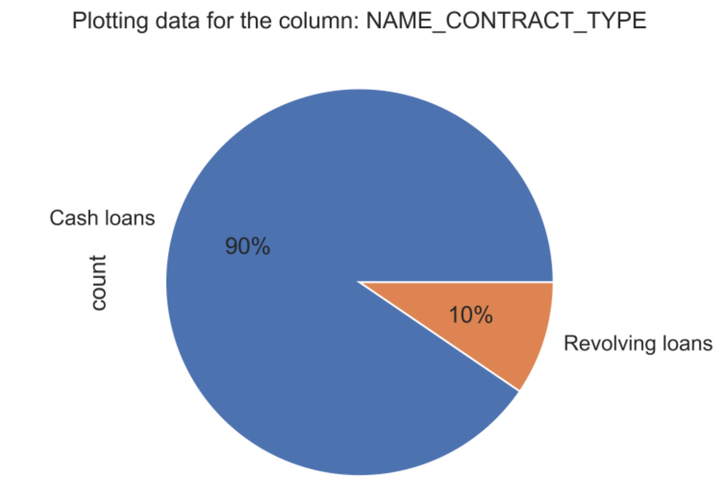


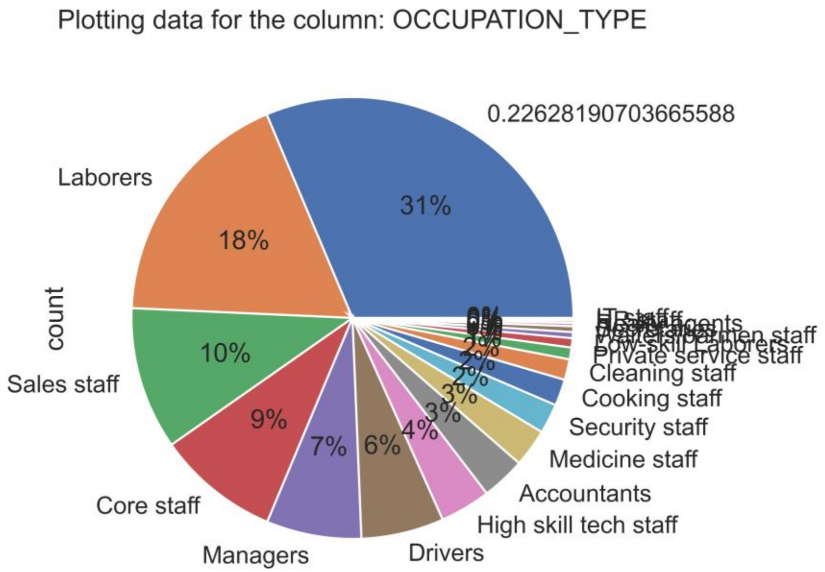


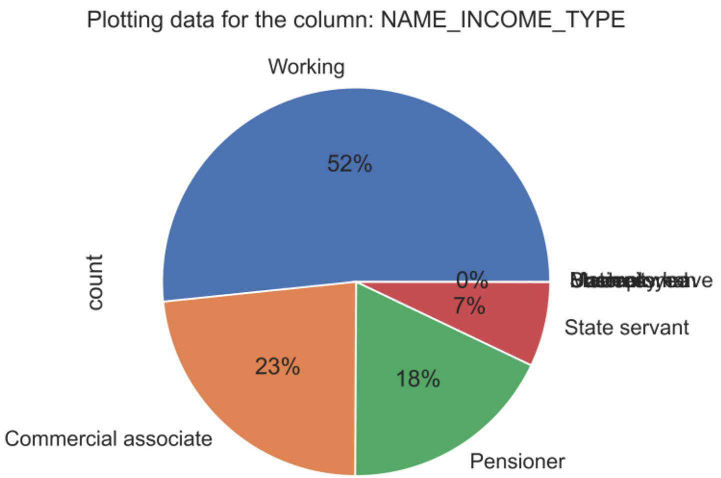


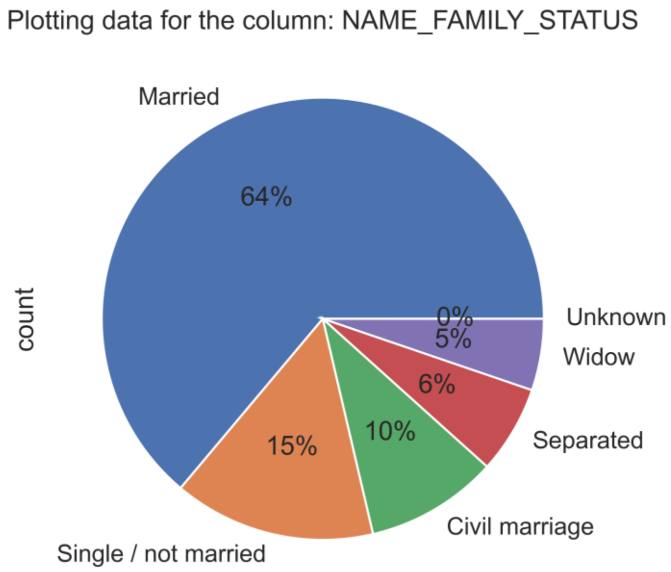
- We can observe that the cash loans are higher than the revolving loans in ‘NAME_CONTRACT_TYPE’.
- Similarly, the unaccompanied is higher, compared to other in column ‘NAME_TYPE_SUITE’.
- In the same way frequency of females is higher than the male in the ‘CODE_GENDER’ column.

CATEGORICAL UNIVARIATE ANALYSIS



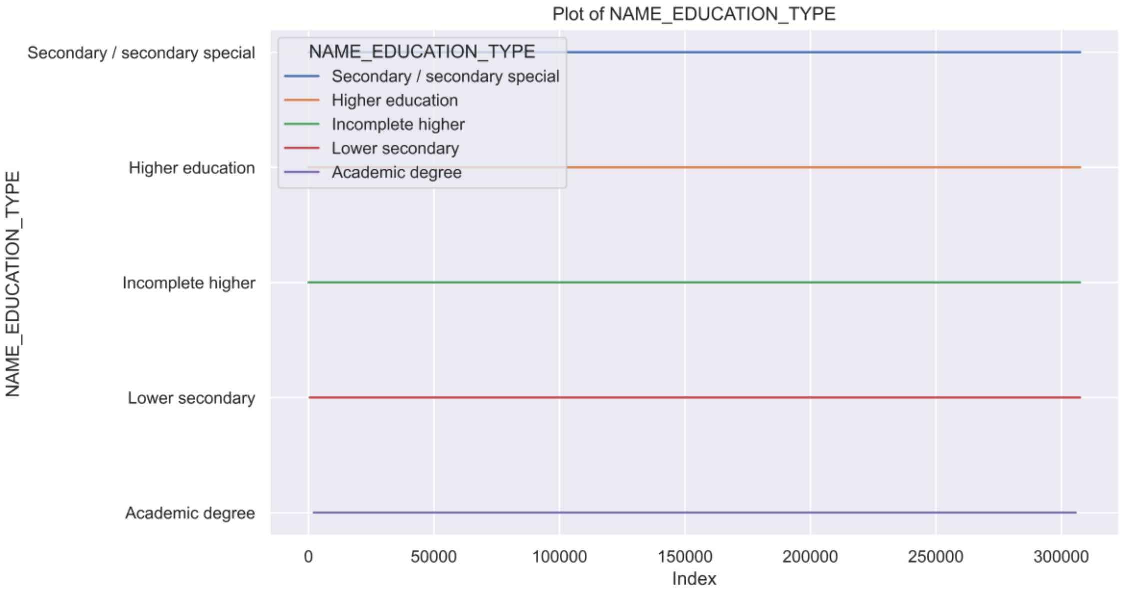


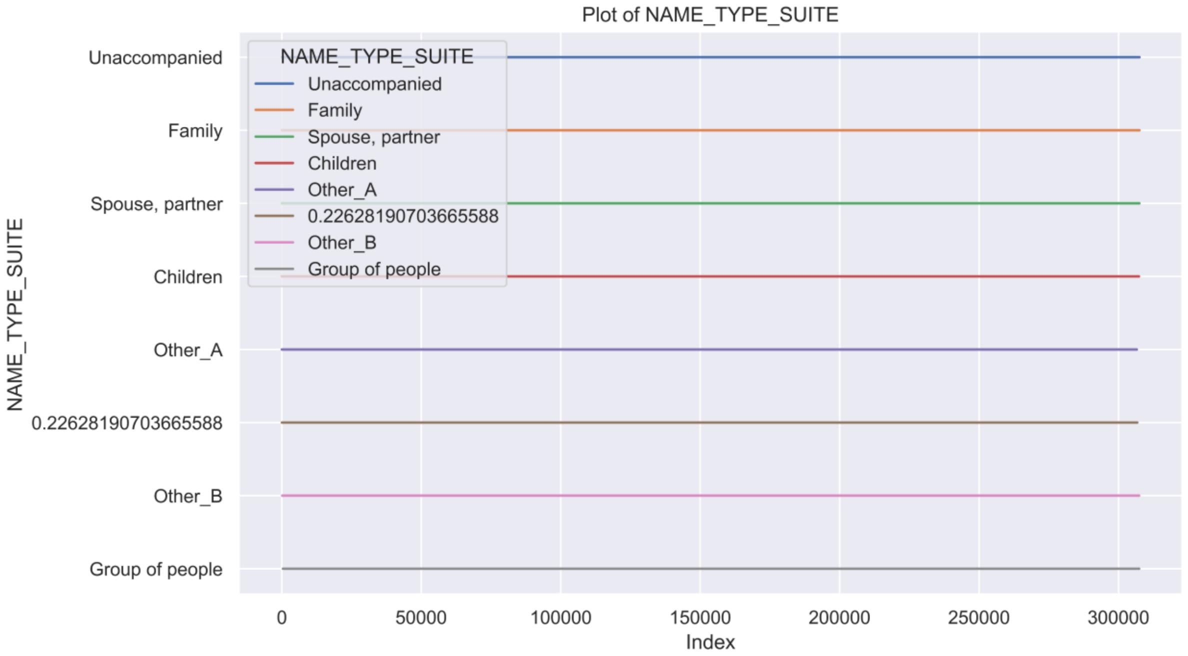


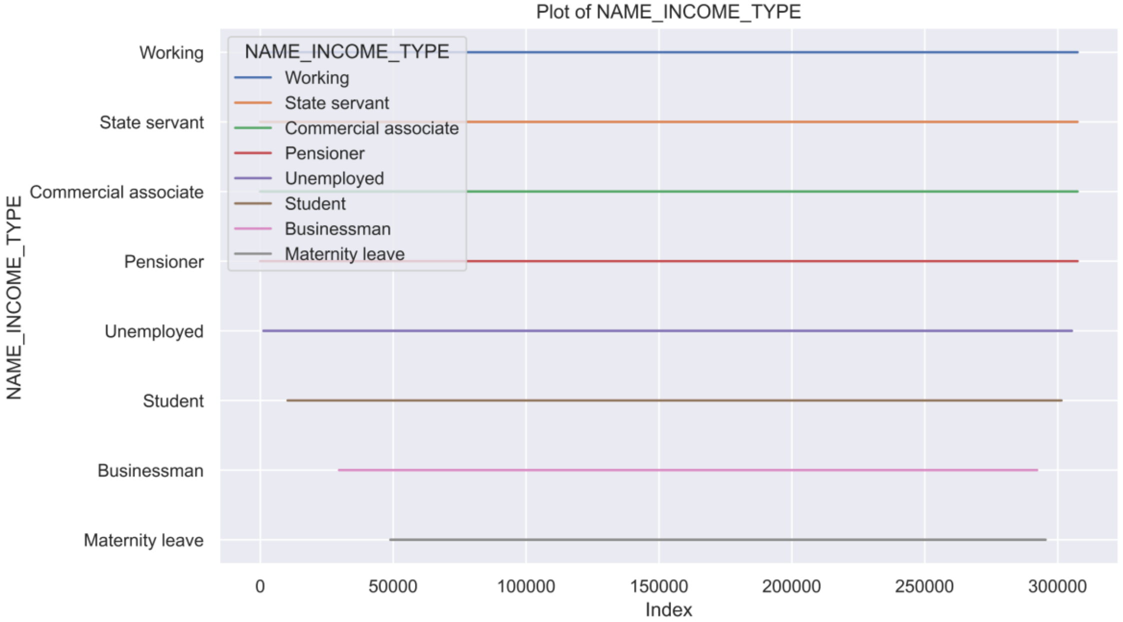


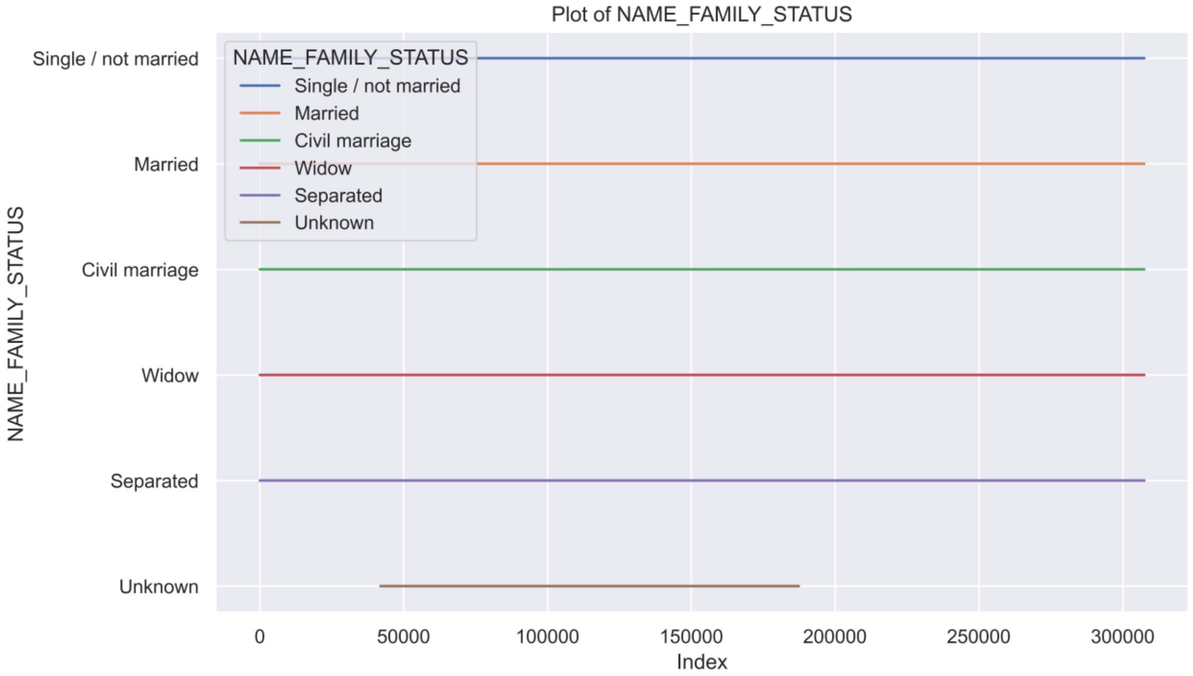
- 10% of the revolving loans out of 100% is occupied, where as 90% of that is cash loans.
- We can observe that, with 18% of the occupation is covered with laborers which is highest of all other occupations.
- More than 50% is filled with working professions compared to others.

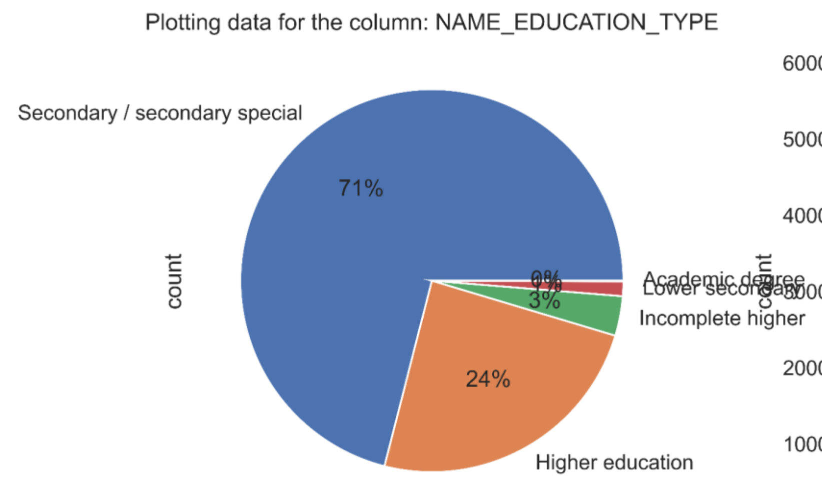
OUTLIERS

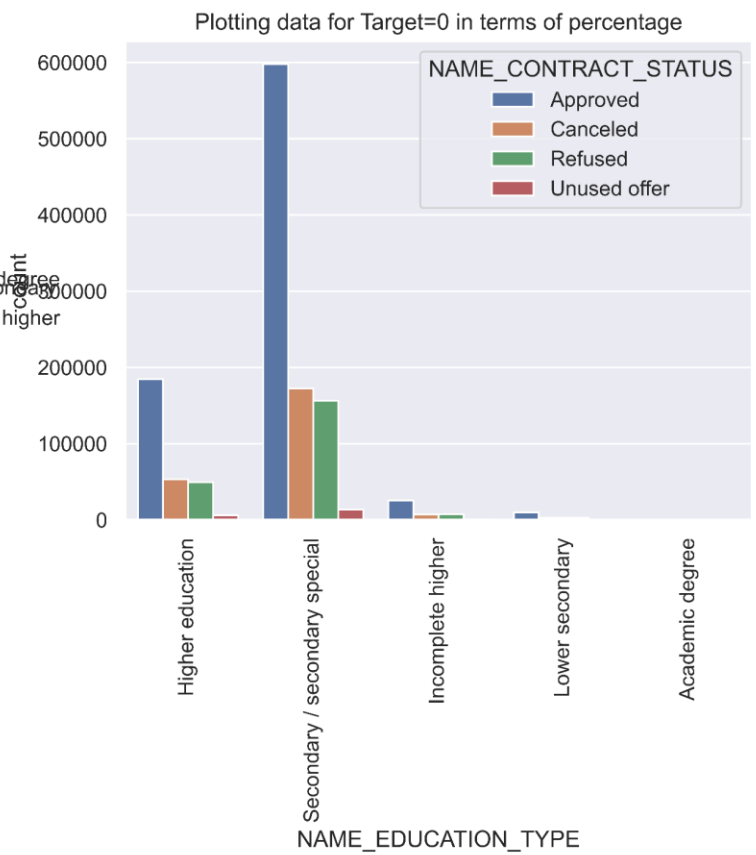

Plot of NAME_EDUCATION_TYPE
NAME_EDUCATION_TYPE
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
0
50000
100000
150000
200000
250000
300000
Index

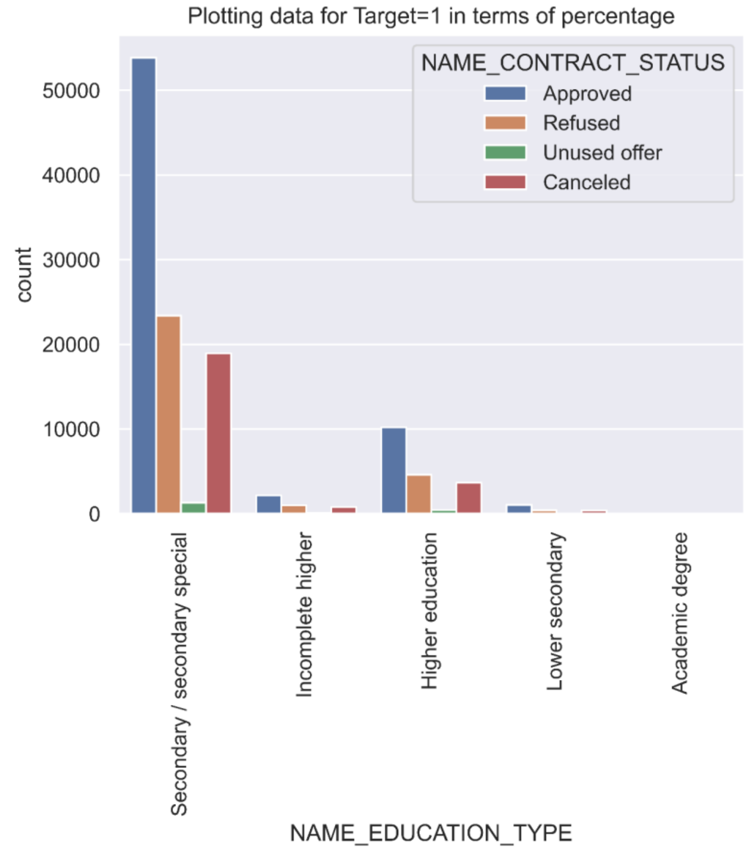

Plot of NAME_TYPE_SUITE
NAME_TYPE_SUITE
Unaccompanied
Family
Spouse, partner
Children
Other_A
0.22628190703665588
Other_B
Group of people
0
50000
100000
150000
200000
250000
300000
Index


Plot of NAME_INCOME_TYPE
NAME_INCOME_TYPE
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
0
50000
100000
150000
200000
250000
300000
Index


Plot of NAME_FAMILY_STATUS
NAME_FAMILY_STATUS
Single / not married
Married
Civil marriage
Widow
Separated
Unknown
0
50000
100000
150000
200000
250000
300000
Index







BIVARIATE ANALYSIS

- This is the one of the example for bivariate analysis.
- It is observed that secondary/secondary special is the highest in approved loans.
- Similarly the lower secondary is the least approved loans.

CONCLUSION

- Example for defaulters (medium pay, age 25-35 then 35-45, male, jobless).
- People working in workers, sales, driving jobs of the business type 3, and not buying houses usually have small loan amounts.
- Additionally, consideration should be given to female candidates, as they historically have lower default rates, than their male counterparts.
- However, many of the past applications listed as denied, withdrawn, or inactive often exhibit well-timed payments, which suggests previous decisions were flawed.
- Previously, dubious applications with rejected, cancelled, or unveiled loans were also involved in defaults, and thus initiated decision making.

THANK YOU