CREDIT EDA LOAN RISK ANALYSIS

BY

SHIVAM SHARMA

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of 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. Identification of such applicants using EDA is the aim of this case study.

Problem Statement

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Purpose of Analysis

Our analysis will assist the company in making a loan approval decision for the applicant. This can help the company control its losses and avoid financial losses.

EDA Process Steps

1. **Data Understanding**
   - Data Reading
   - Data Types
   - Data Imbalance

2. **Data Cleaning and Manipulation**
   - Date – Days , Year , Month
   - Missing Values
   - Imputation
   - Outliers

3. **Data Analysis**
   - Univariate & Segmented Univariate Analysis – Bar Plots , Pie Plots
   - Bivariate Analysis – Bar Plots , Scatter Plots
   - Correlation b/w Variables using Heat Map and Correlation Matrix Percentage.

4. **Conclusions & Key Factors**

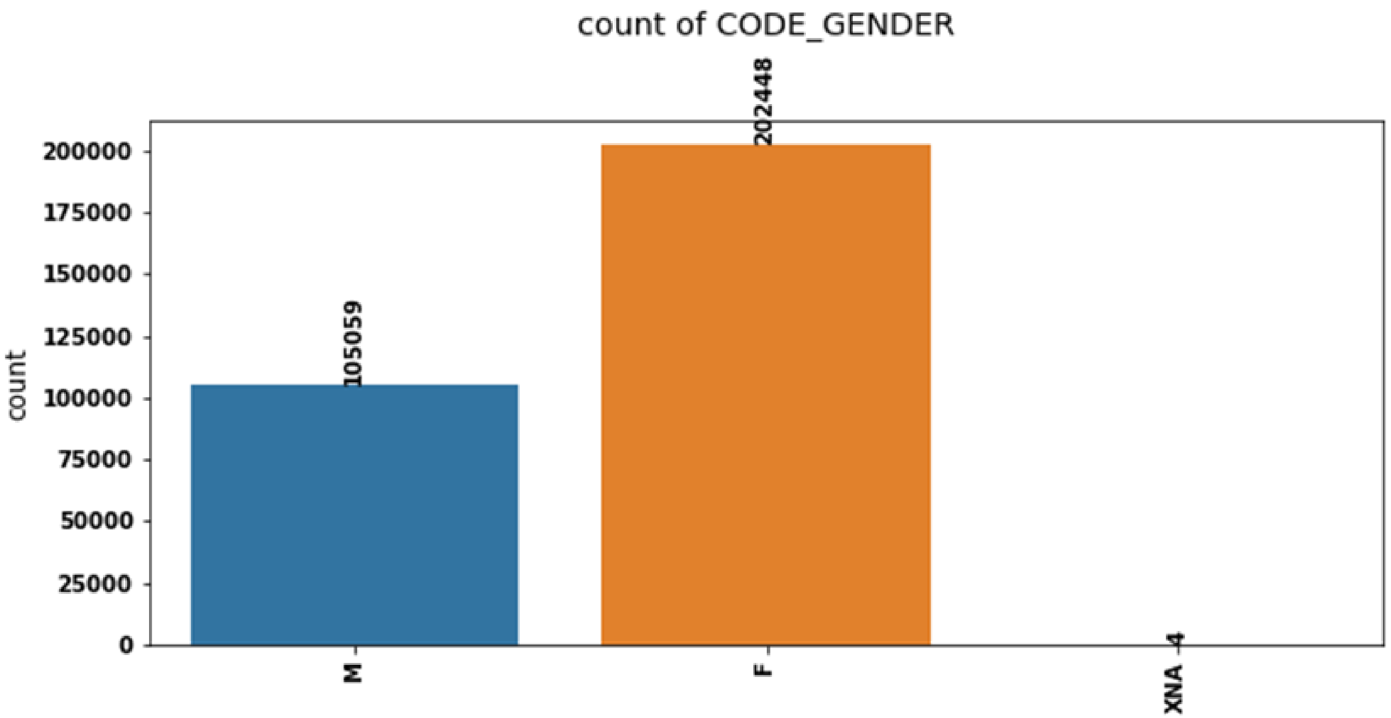
EDA Requirements

1. **Datasets**
   - Current Application
   - Previous Application
   - Columns Description
2. **Python Libraries**
   - Pandas - For Data Reading , Cleaning & Manipulation
   - Numpy - For calculations
   - Matplotlib - For Data Visualization
   - Seaborn - For Data Visualization
3. **Computing Platform**
   - Jupyter Notebook

Univariate Analysis

Observation

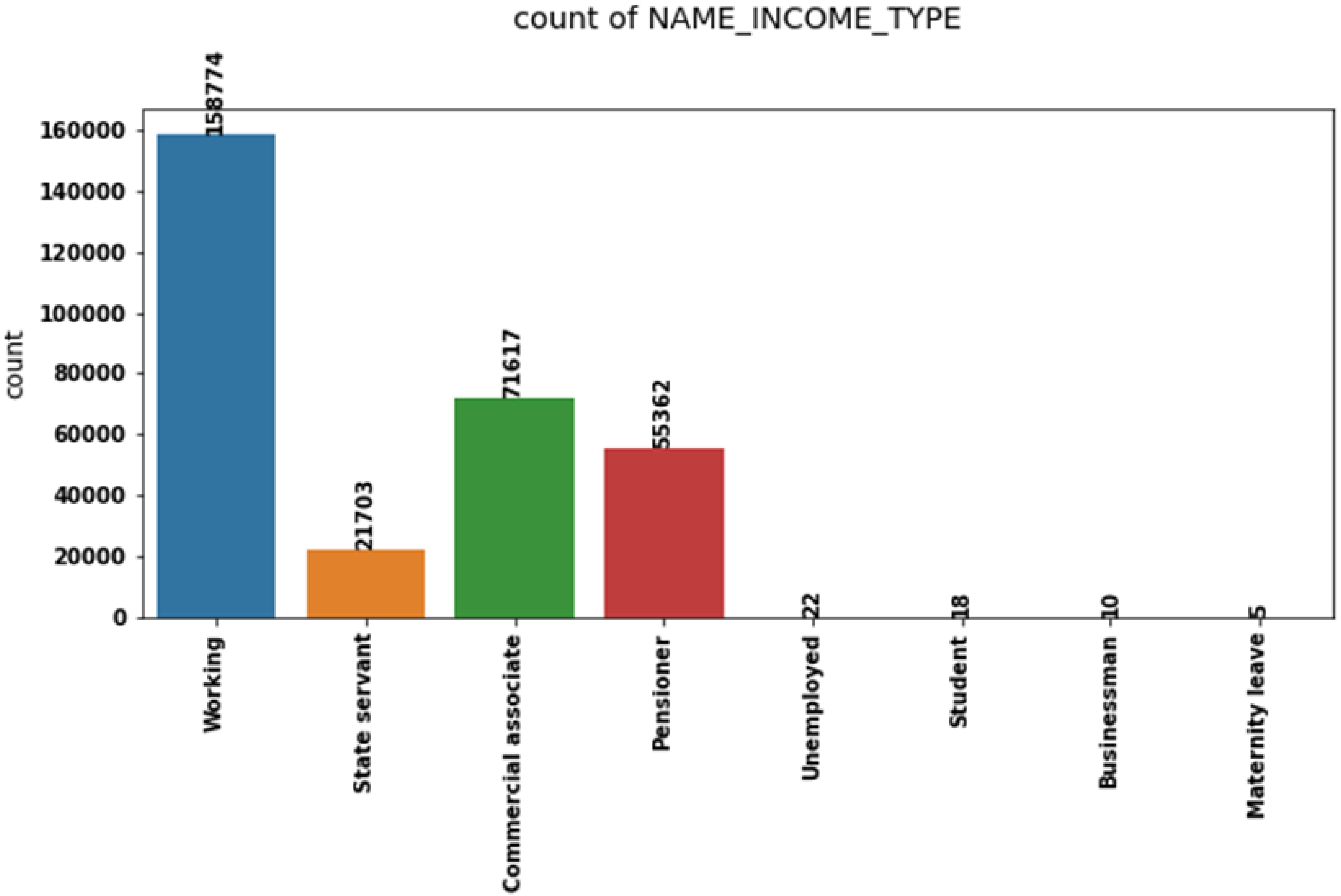
- Females take out more loans.
- Men take out far fewer loans than women.



Univariate Analysis

Observation

- Working, retired, and commercial associates take out the most loans, according to the graph.
- Unemployed, students, business owners, and mothers on maternity leave take out fewer loans.



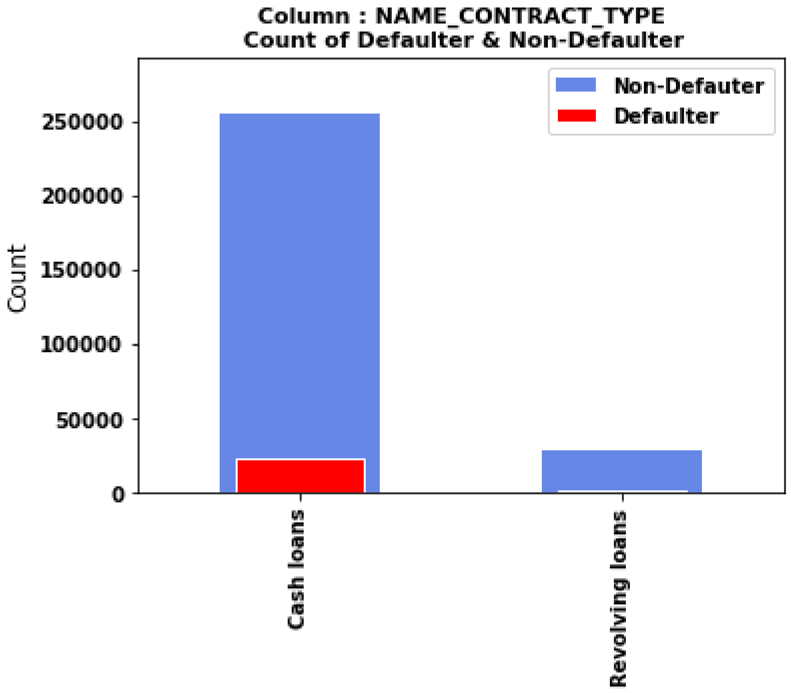
Who are taking more loans ?

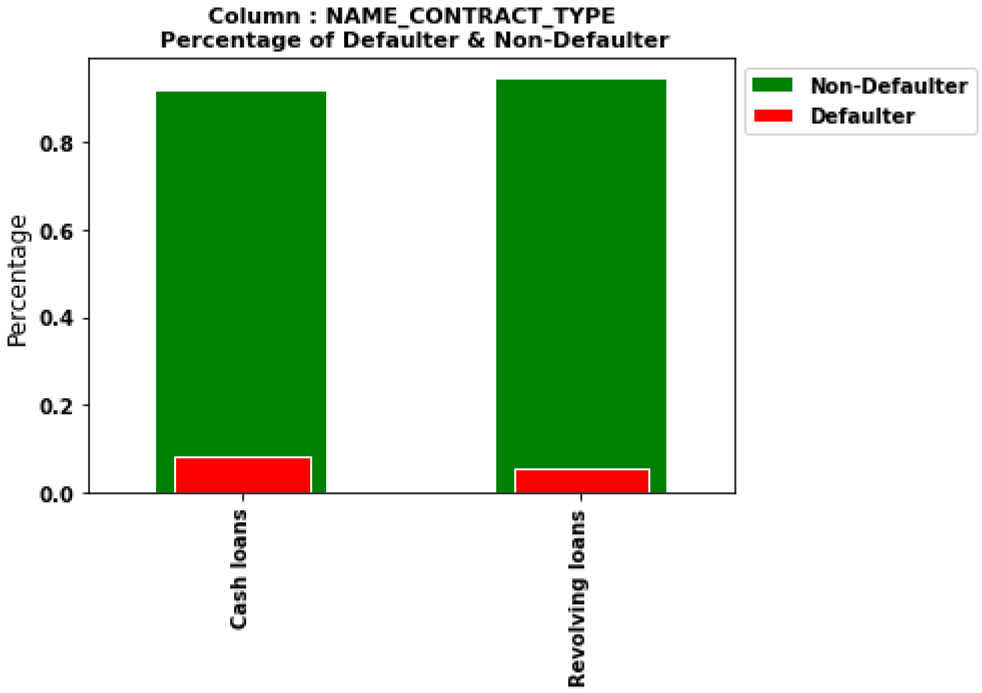
- Females borrow more than men.
- People with a secondary/special education are the most likely to apply for a loan.
- People tend to take out more cash loans, and the default rate on revolving loans is lower.
- People aged 27 to 41 take out the most loans.
- When compared to other categories, married people take out more loans.
- People who own a home or an apartment are more likely to take out loans.
- People who do not own a car are more likely to take out loans.
- People who own real estate tend to take out more loans.

Segmented Univariate Analysis

- ❖ Categorical Segmented Univariate Analysis
- ❖ Numerical Segmented Univariate Analysis

Categorical Univariate Segmented Analysis

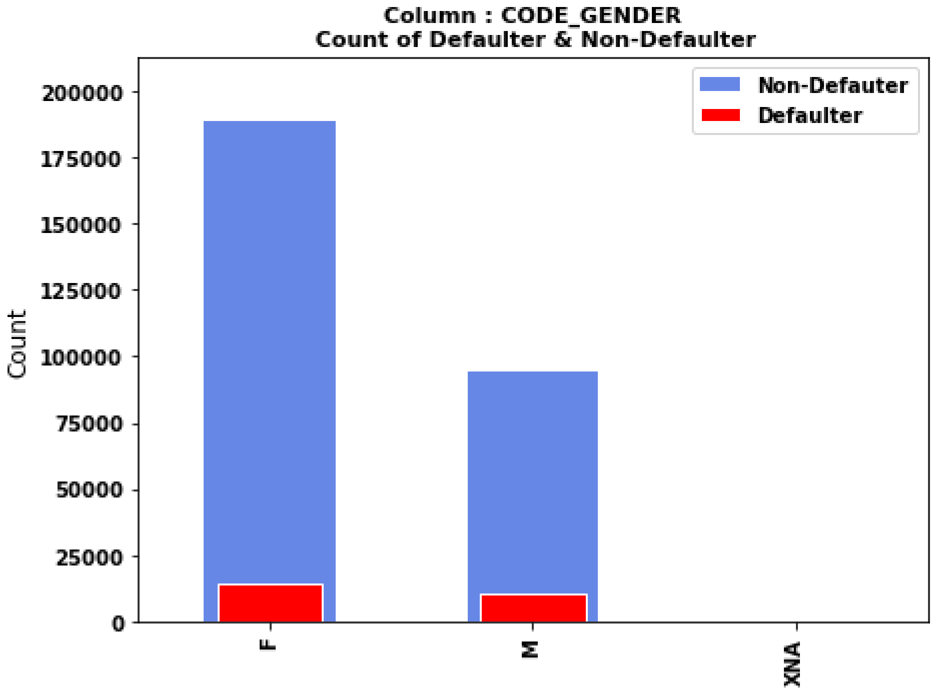


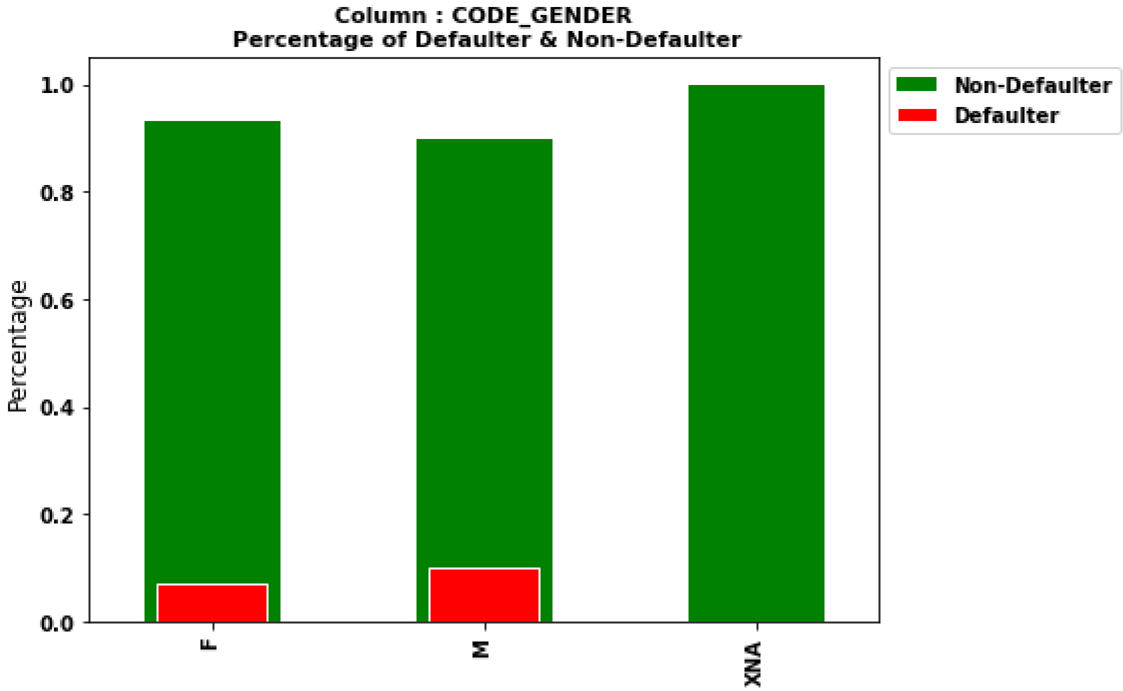


Observation

- People take **Cash** loans more than **Revolving** loans.
- Percentage of defaulters in Cash loans are more than **Revolving** loans. Revolving loans are less risky than cash loans.

Categorical Univariate Segmented Analysis

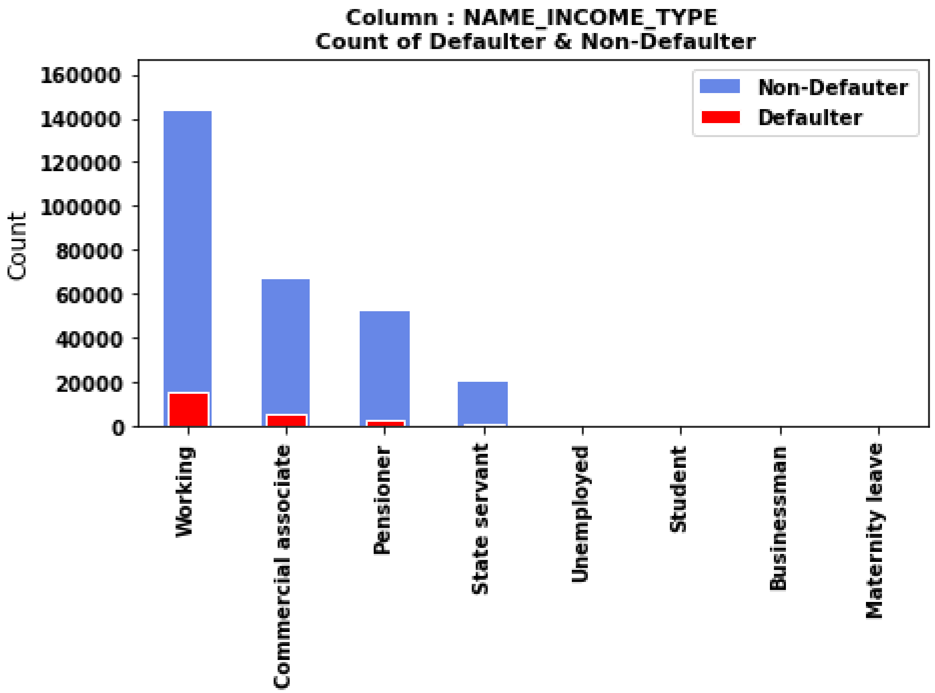




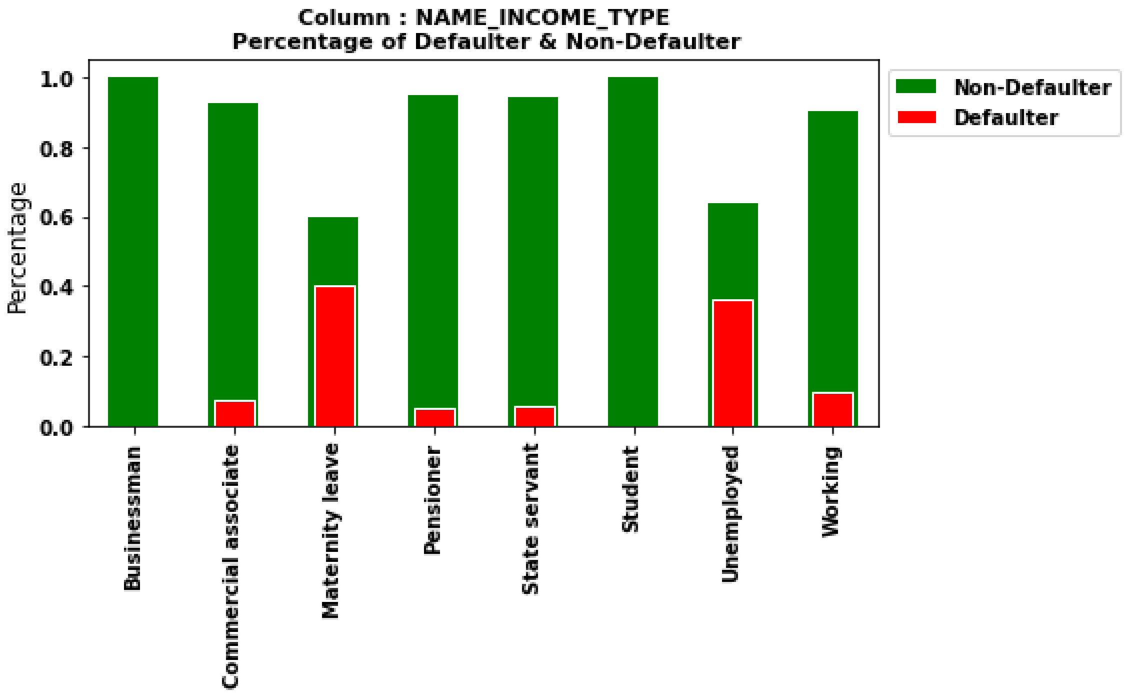
Observation

- Females take more loans than Men.
- Percentage of defaulters in Females are less than the Men. It is less risky to give loans to Females than Men.

Categorical Univariate Segmented Analysis



Column : NAME_INCOME_TYPE
Count of Defaulter & Non-Defaulter

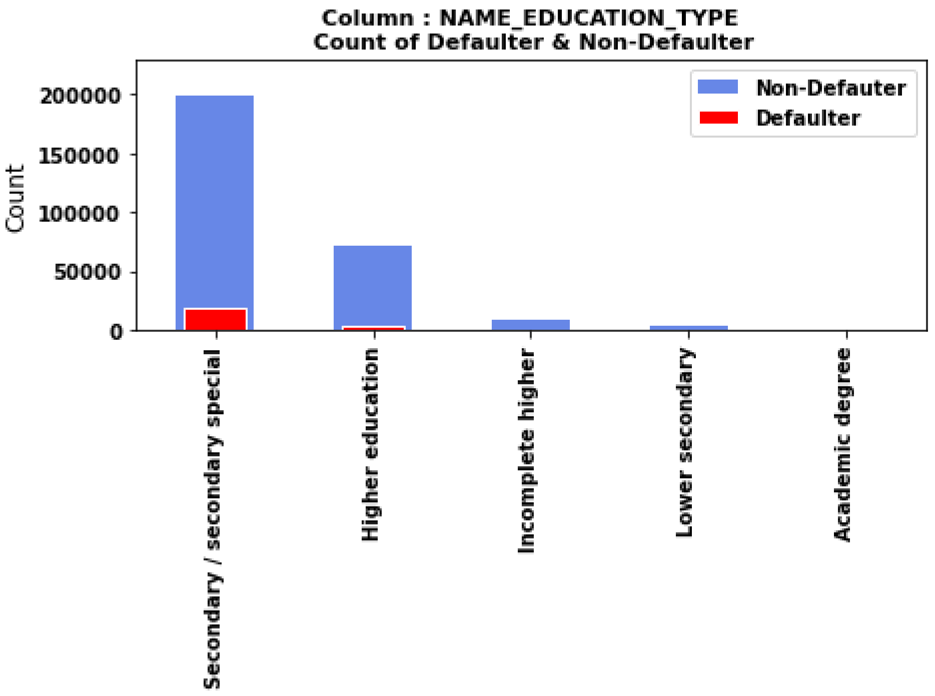


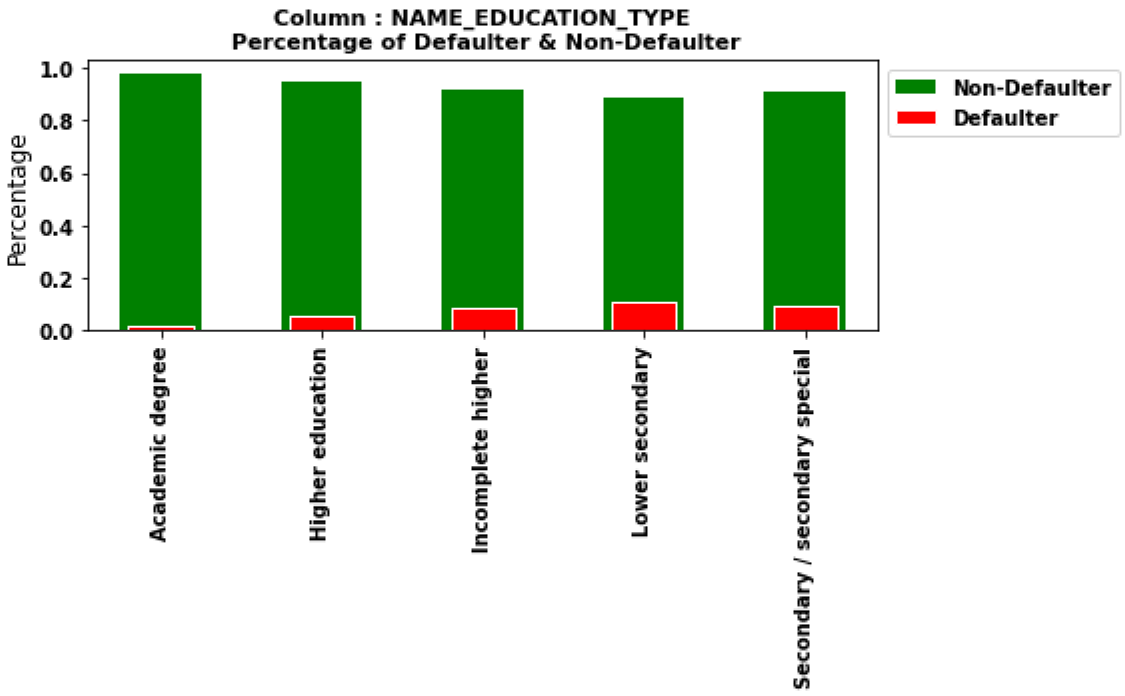
Column : NAME_INCOME_TYPE
Percentage of Defaulter & Non-Defaulter

Observation

- Working, Commercial Associate, Pensioner, State servant take loans more and having less percentage of defaulters.
- Unemployed and maternity leave people take loans less but having very high percentage of defaulter.
- Businessman and Student take loans less and no defaulter found in them according to data.

Categorical Univariate Segmented Analysis

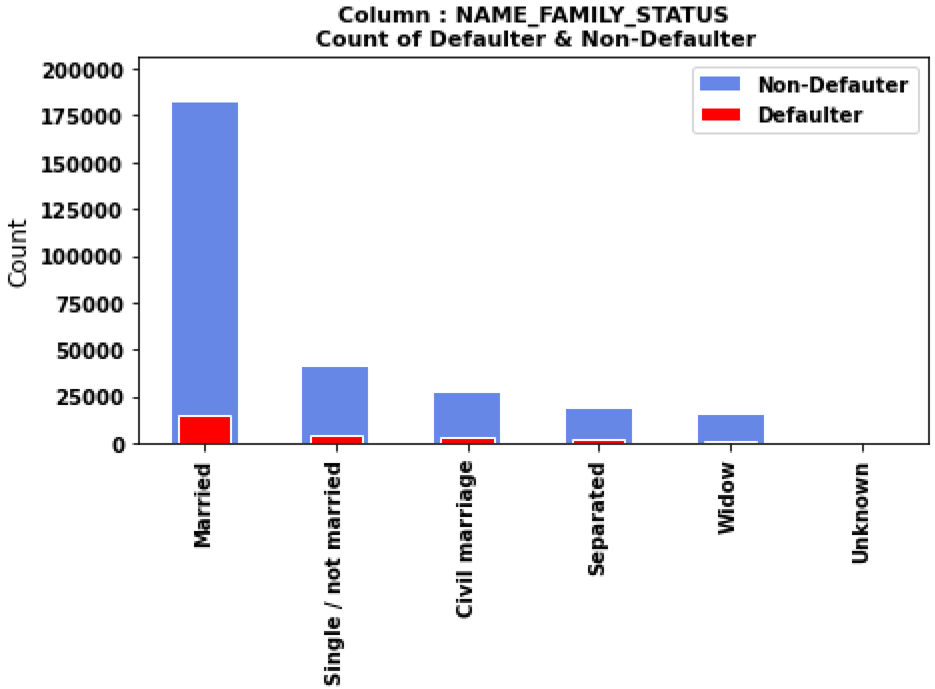


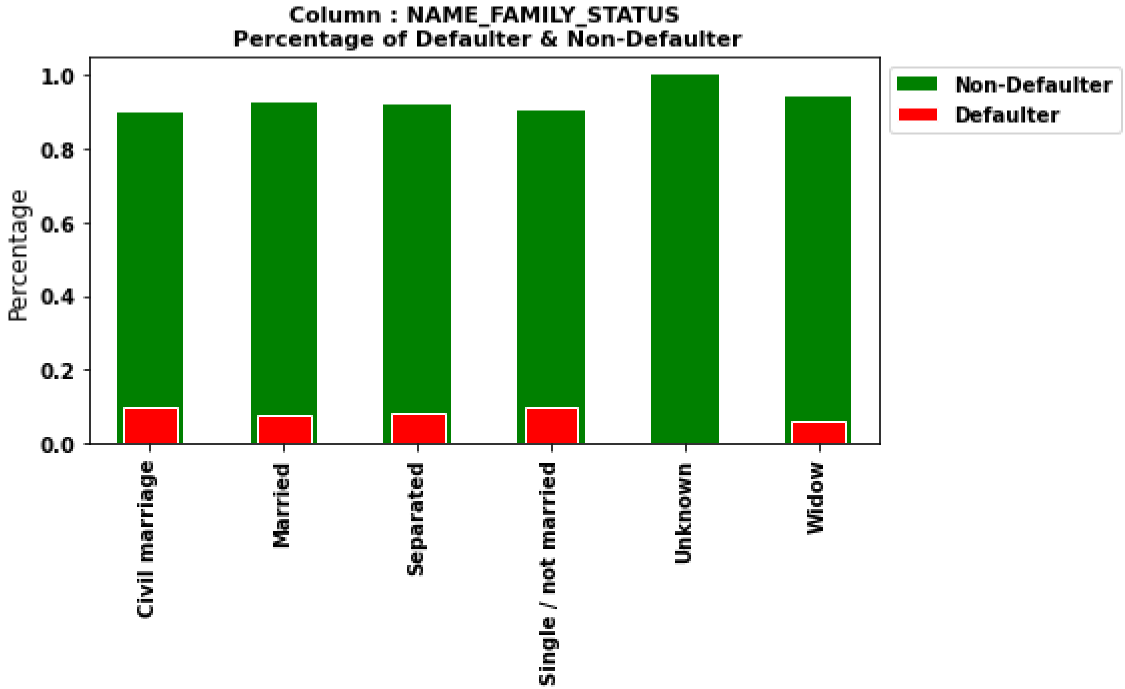


Observation

- Lower and incomplete education people take loans more and having high percentage of defaulter in them.
- Academic or Higher educated people defaulter percentage is less.

Categorical Univariate Segmented Analysis

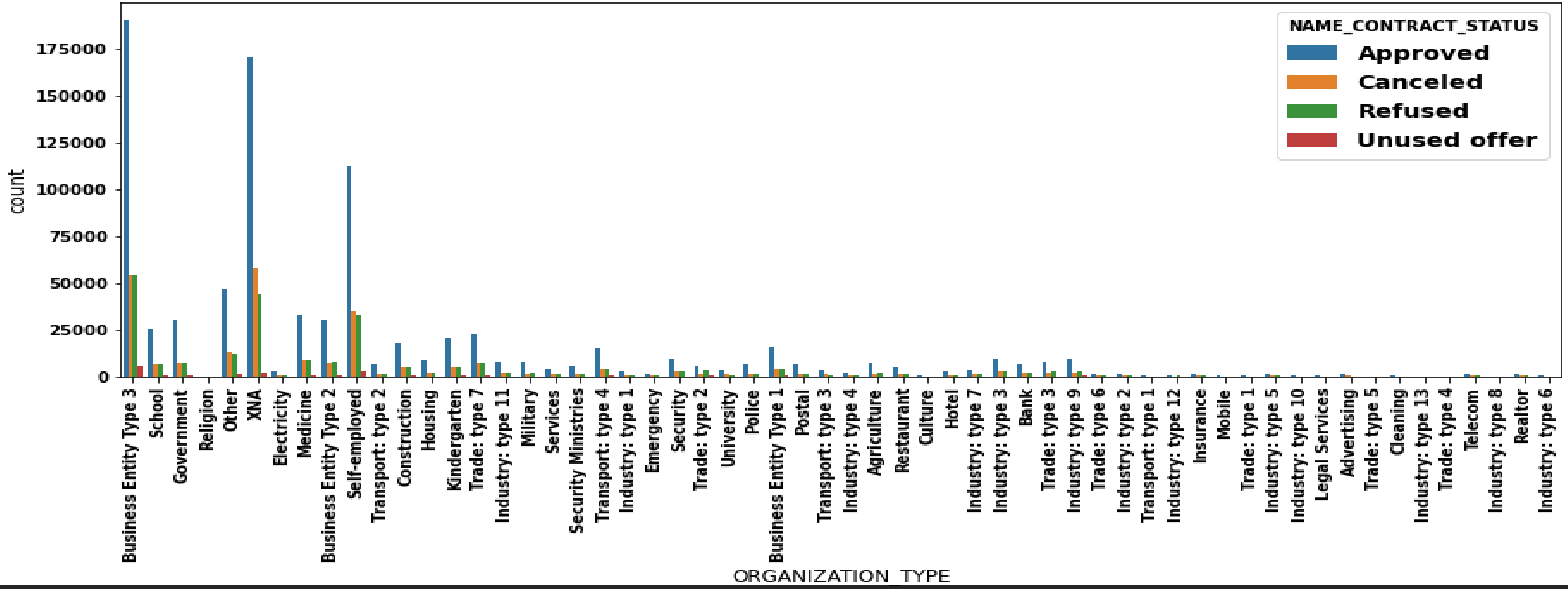




Observation

- Married people take more loans.
- Single and Civil marriage people take less loans but the single & civil marriage people having high percentage of defaulter than married people.

Categorical Univariate Segmented Analysis



Observation

- Loans for Business Entity Type 3, XNA, and Self-employed People were approved at a high rate.

Categorical Univariate Segmented Analysis

Overall Observation

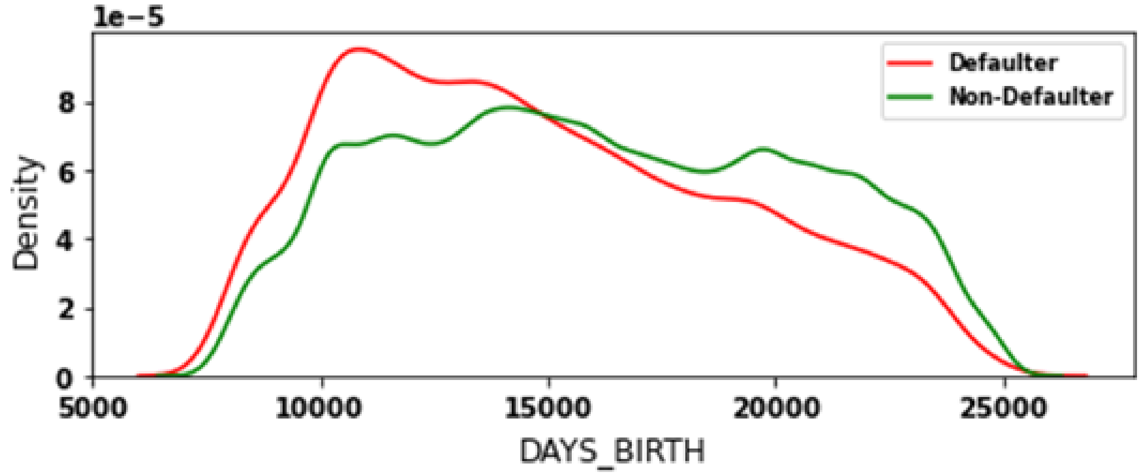
- NAME_CONTRACT_TYPE - People take Cash loans more than Revolving loans. Percentage of defaulters in Cash loans are more than Revolving loans. Revolving loans are less risky than cash loans.
- CODE_GENDER - Females take more loans than Men. Percentage of defaulters in Females are less than the Men. It is less risky to give loans to Females than Men.
- FLAG_OWN_CAR - People who do not own car take loans more than people who do not own car. People who owns car & not own car are having approx. same percentage defaulters in them. But people who owns car are less likely to default .
- FLAG_OWN_REALTY - People having their own realty take loans more than who do not have. Percentage of defaulters & non defaulters are approx. same. But the weightage of the people who owns realty is high, so instead of cancelling their applications, we can set some rules to minimize the risk of default
- NAME_INCOME_TYPE - Working, Commercial Associate, Pensioner, State servant take loans more and having less percentage of defaulters. Whereas Unemployed and maternity leave people take loans less but having very high percentage of defaulter. Businessman and Student take loans less and no defaulter found in them according to data.
- NAME_EDUCATION_TYPE - Lower and incomplete education people take loans more and having high percentage of defaulter in them. Higher educated people defaulter percentage is less.
- NAME_FAMILY_STATUS - Married people take more loans whereas single and civil marriage people take less loans but the single & civil marriage people having high percentage of defaulter than married people.
- NAME_HOUSING_TYPE - People living in their house & apartment take loans more whereas people living in rented apartments and with parents take less loans but the percentage of people living in rented apartment and with parents are having high percentage of defaulting risk than the people living in house or apartment.
- OCCUPATION_TYPE - Labor class, core staff, managers, sales staff, drivers take loans more. According to graph all type people having approx. equal percentage of defaulting risk. Lower skill labor take less loans and having high percentage of defaulters.
- ORGANIZATION_TYPE - Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%), and Restaurant: type 3 (16%) are the organizations with the highest percentage of loans not repaid (less than 12 percent). Self-employed people have a high default rate and should be avoided when applying for a loan or providing a loan with a higher interest rate to mitigate the risk.

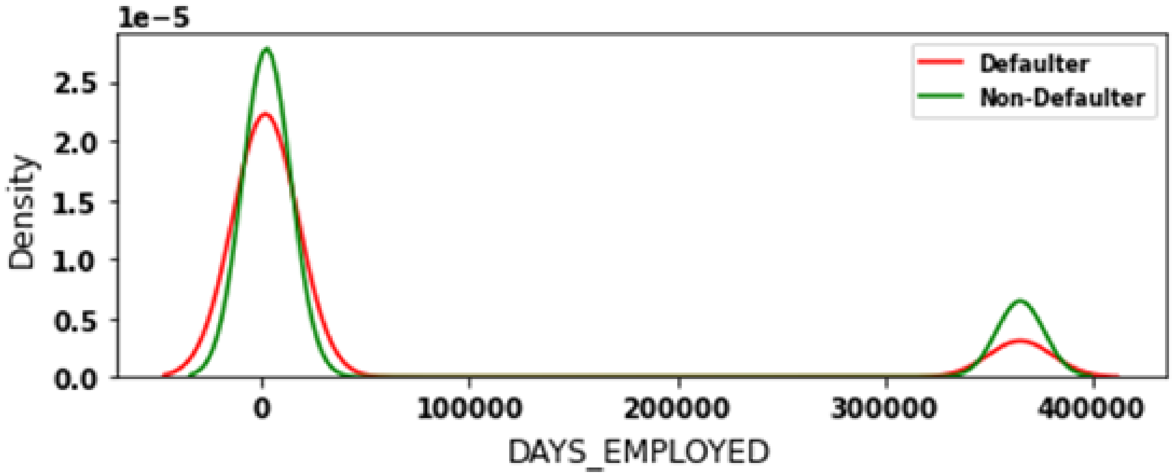
Numerical

Univariate Segmented Analysis

Observation

- DAYS_BIRTH - People aged 10000 days (27 years) to 15000 days (41 years) are more likely to take out loans and have a higher rate of default.
- DAYS_EMPLOYED - People who have recently been employed prior to applying for a loan are more likely to default.



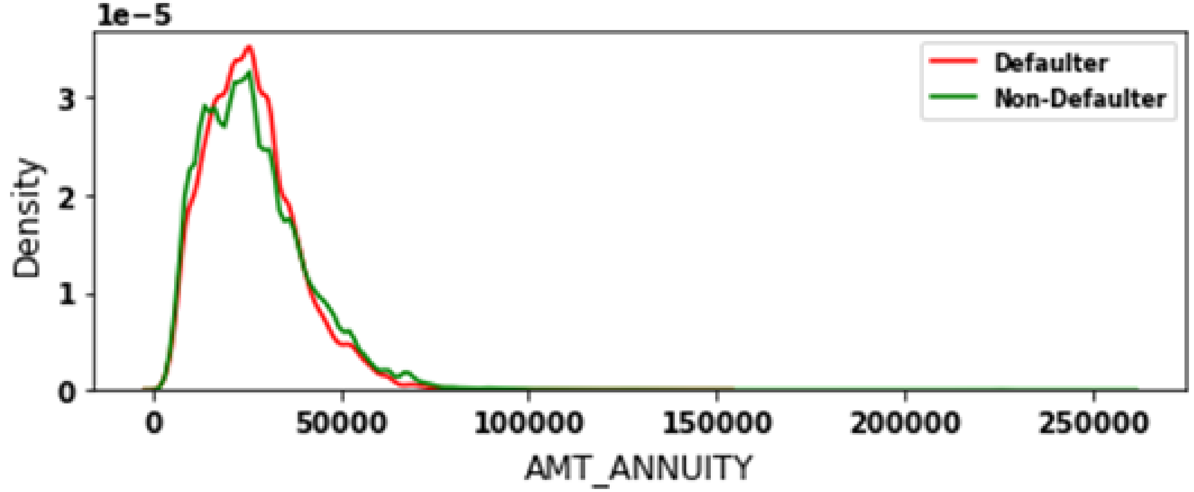


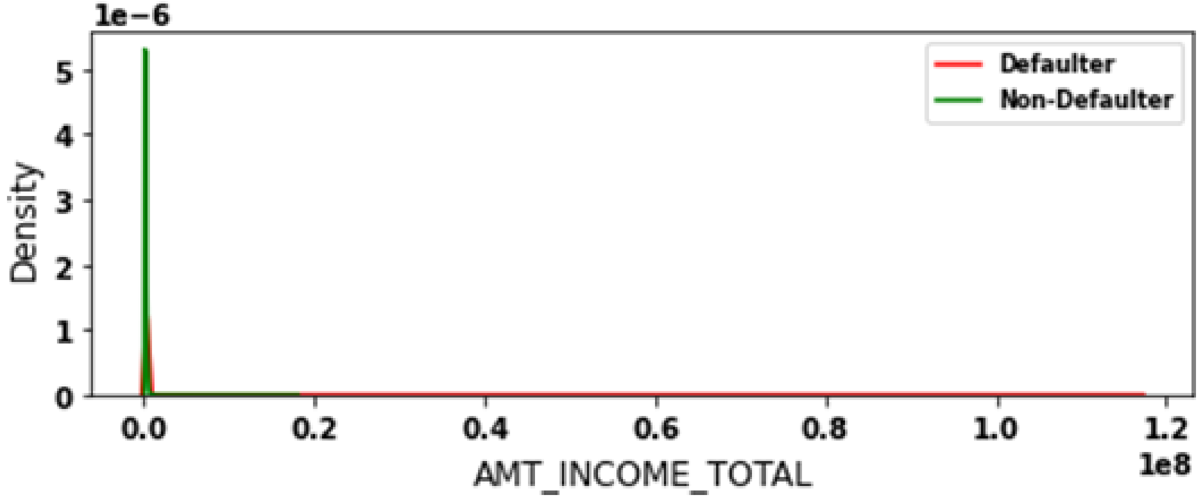
Numerical

Univariate Segmented Analysis

Observation

- AMT_INCOME_TOTAL – Low income individuals are more likely to default.
- AMT_ANNUITY - Annuity with a low payout has a large number of loans and are more likely to default.



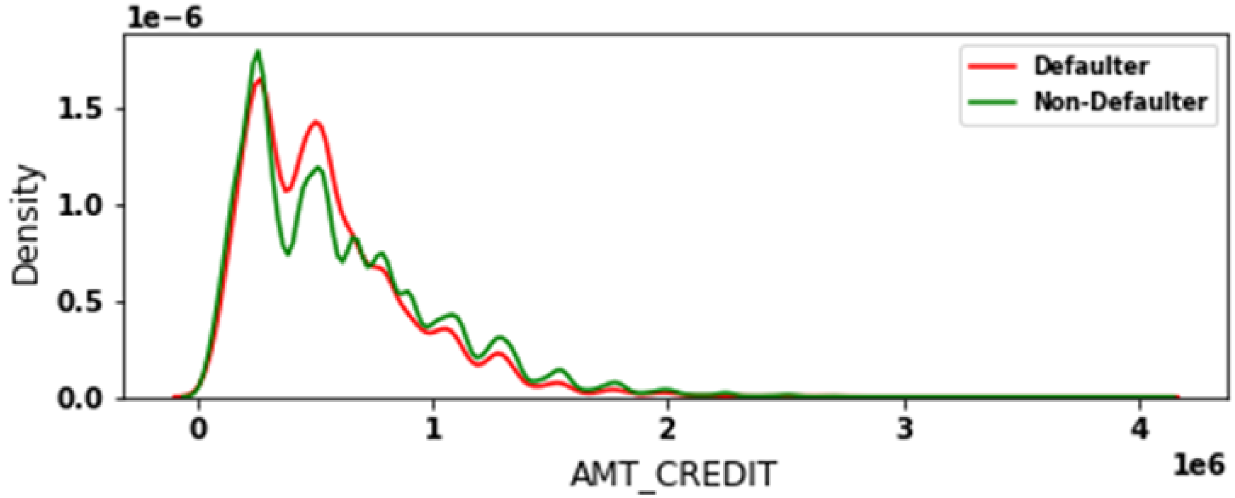


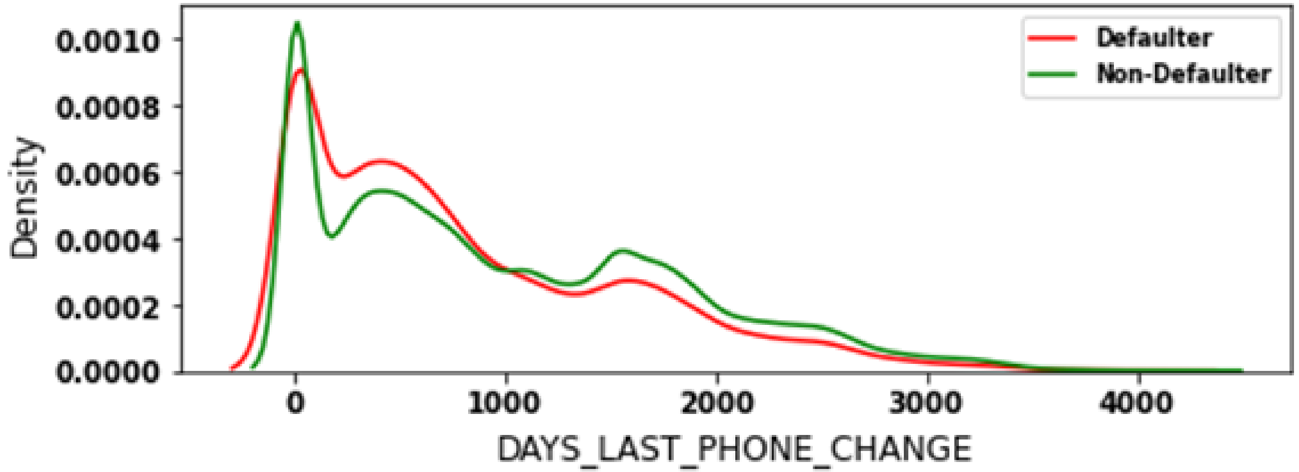
Numerical

Univariate Segmented Analysis

Observation

- AMT_CREDIT - The people who take less amount of loans are more likely to default.
- DAYS_LAST_PHONE_CHANGE - People who have changed their phone number within the last 200 days borrow more and have a higher default rate.



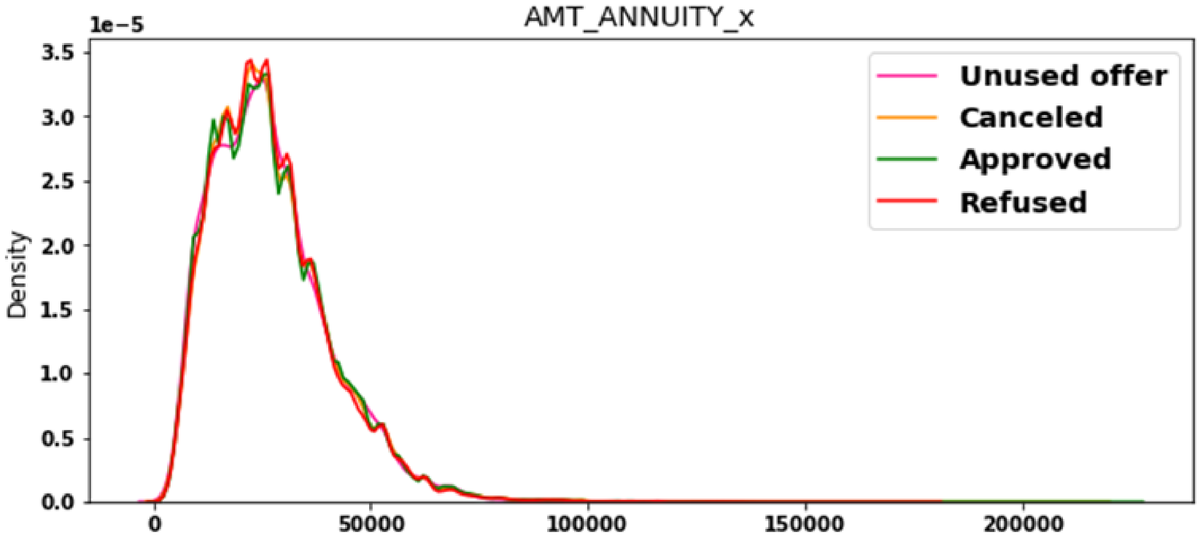


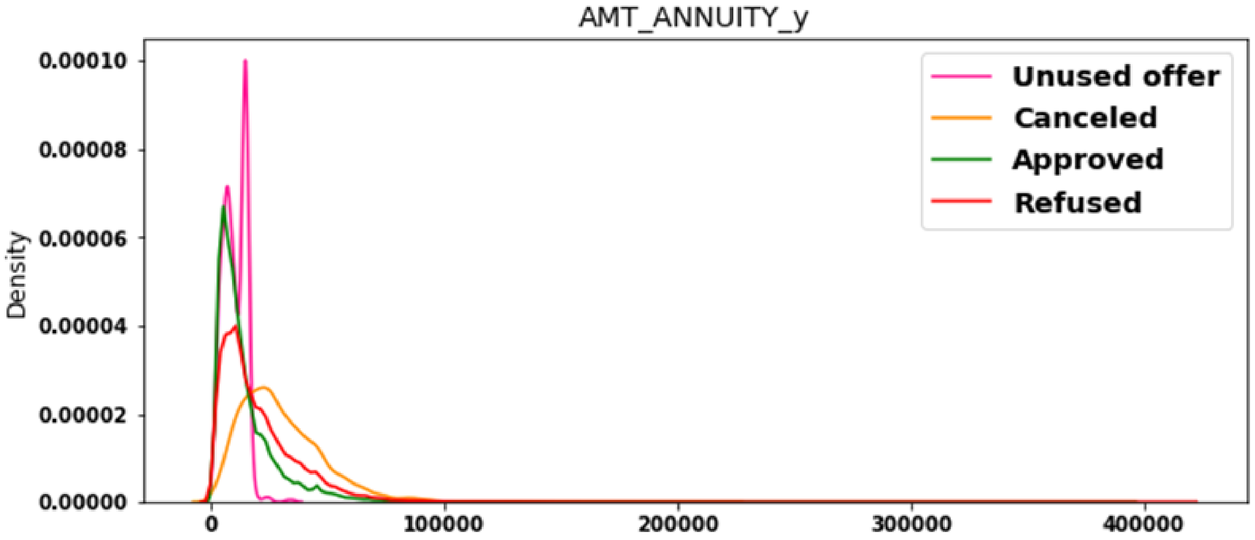
Numerical

Univariate Segmented Analysis

Observation

- AMT_ANNUITY_x - The current number of cancelled/refused offers for AMT ANNUITY is comparable.
- AMT_ANNUITY_y - the bank had a high number of unused offers at low Annuity.





Numerical Univariate Segmented Analysis

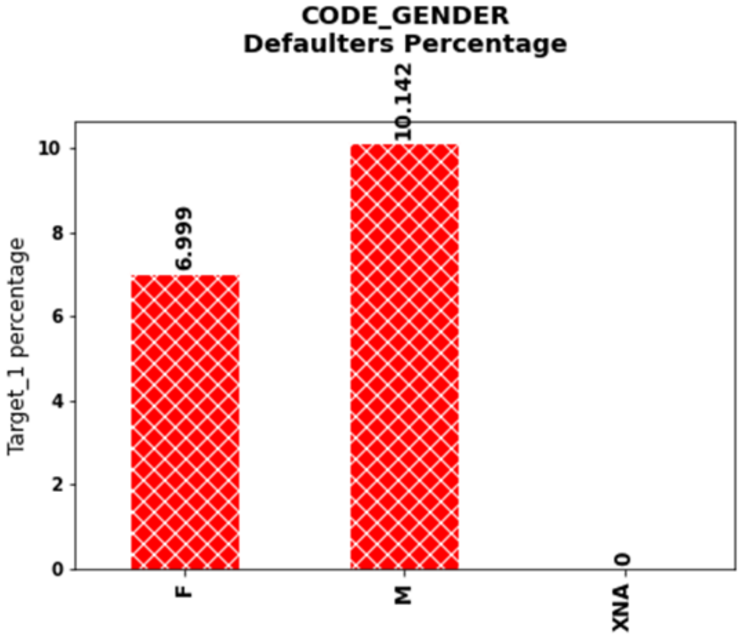
Overall Observation

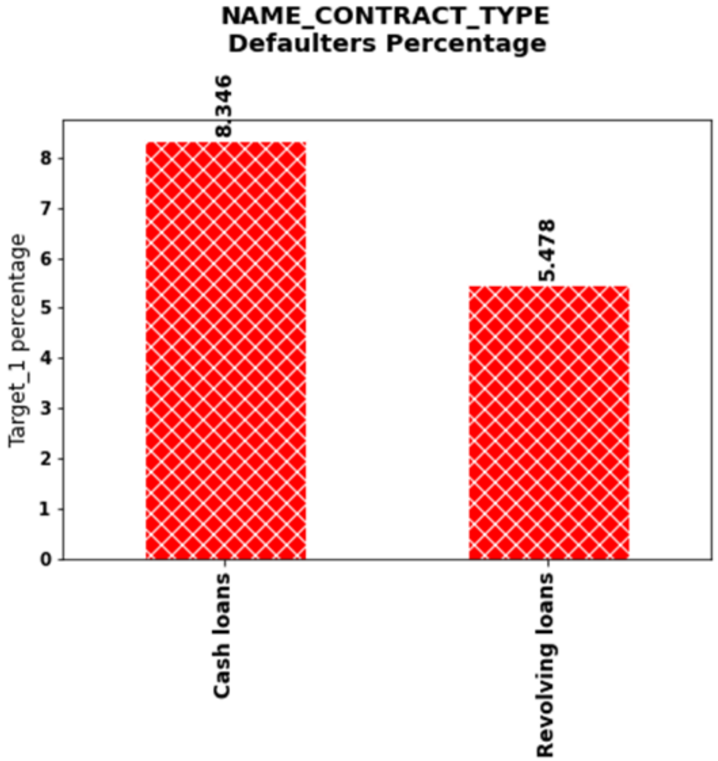
- DAYS_BIRTH - People aged 10000 days (27 years) to 15000 days (41 years) are more likely to take out loans and have a higher rate of default.
- DAYS_EMPLOYED - People who have recently been employed prior to applying for a loan are more likely to default.
- DAYS_ID_PUBLISH - People whose IDs were published between 4000-5000 days take out more loans and have a higher default rate.
- DAYS_LAST_PHONE_CHANGE - People who have changed their phone number within the last 200 days borrow more and have a higher default rate.
- HOURS_APPR_PROCESS_START - Between 10:00 am to 2:00 pm high amount of loans are given and have high defaulter rate.
- CNT_FAM_MEMBERS - People with family members 2-5 are like to take more loans.
- AMT_INCOME_TOTAL - Low-income individuals are more likely to default.
- AMT_ANNUITY - Annuity with a low payout has a large number of loans and are more likely to default.
- AMT_GOODS_PRICE - People Having Low Price of goods default the most.
- AMT_CREDIT - The people who take less amount of loans are more likely to default.

Bivariate Analysis

- ❖ Categorical – Categorical Bivariate Analysis
- ❖ Categorical – Numerical Bivariate Analysis
- ❖ Numerical – Numerical Bivariate Analysis

Categorical - Categorical Bivariate Analysis

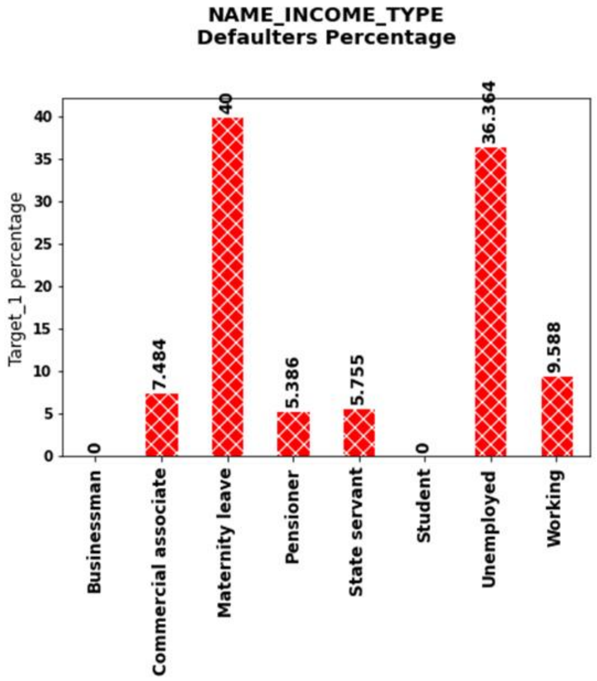

CODE_GENDER
Defaulters Percentage

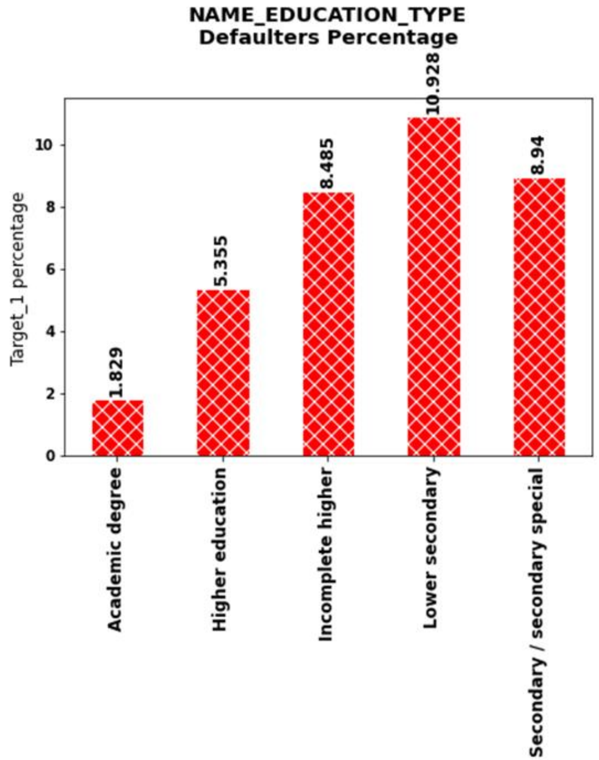

NAME_CONTRACT_TYPE
Defaulters Percentage

Observation

- Default rate is high in Cash loans than revolving loans.
- Default rate is high in Males then Females.

Categorical - Categorical Bivariate Analysis





Observation

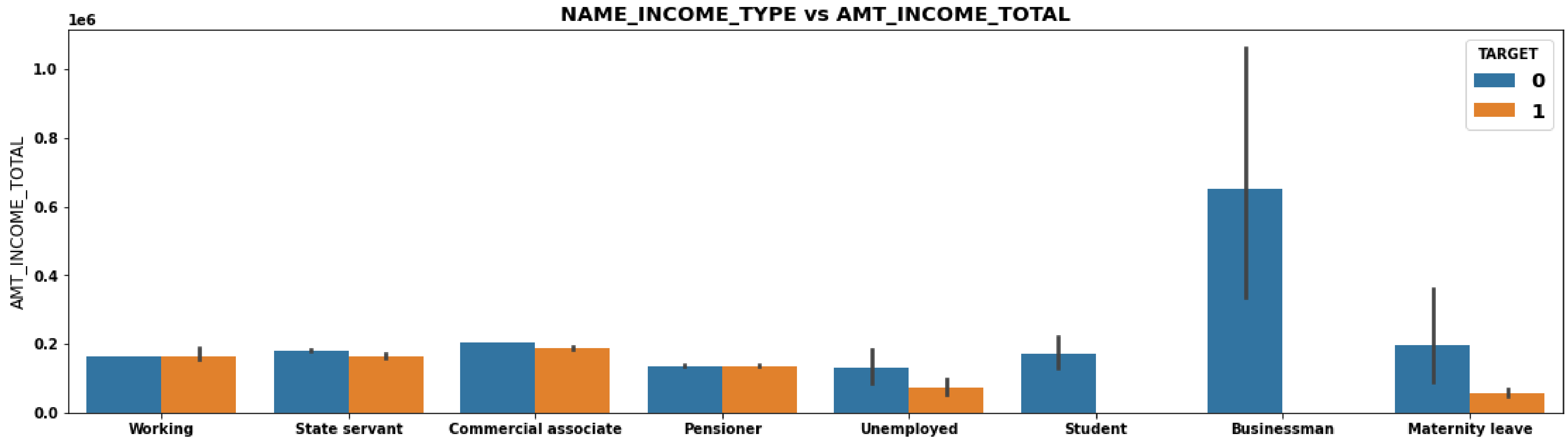
- Unemployed , Maternity leave & Working people have high default rate than commercial associate, pensioner & state servant.
- Lower education people have high default rate than people having higher education.

Categorical - Categorical Bivariate Analysis

Overall Observation

- Default rate is high in Cash loans than revolving loans.
- Default rate is high in Males then Females.
- People who do not own car have high default rate than people who own car.
- People who do not own Realty have high default rate than people who own Realty.
- Unemployed , Maternity leave & Working people have high default rate than commercial associate, pensioner & state servant.
- Lower education people have high default rate than people having higher education.
- Civil marriage and single people have high default rate.

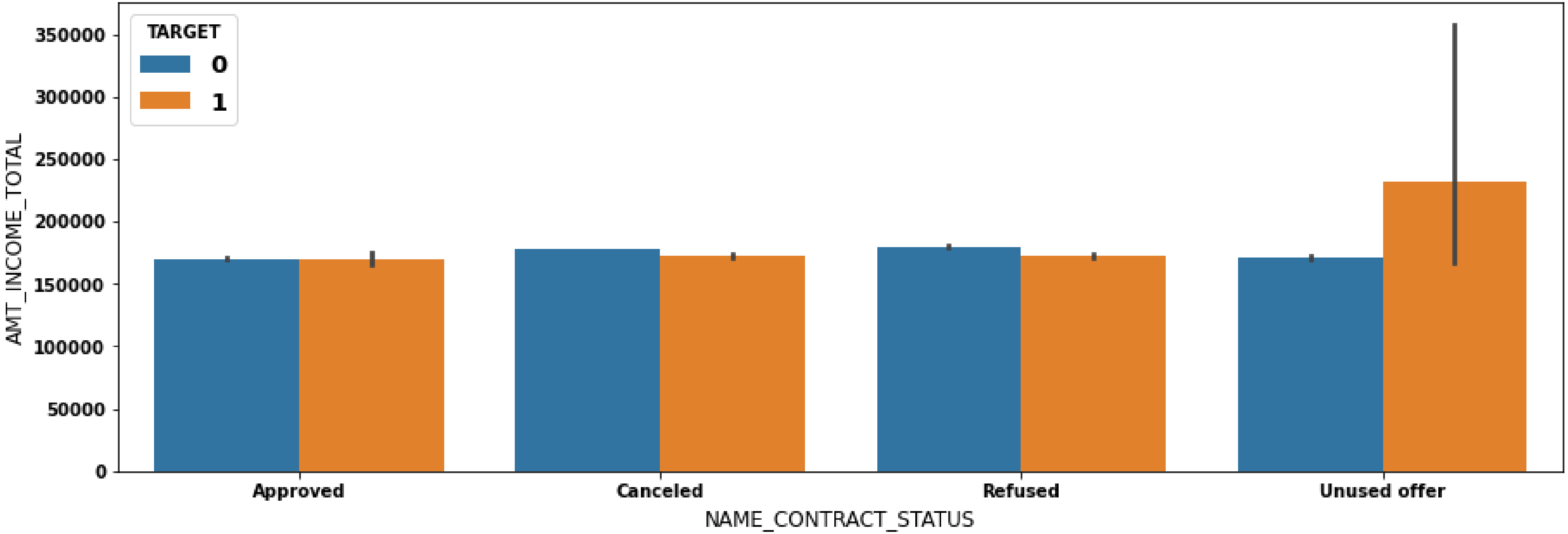
Numerical - Categorical Bivariate Analysis



Observation

- As shown in the graph above, people with high income, such as businessmen, have a 0% default rate.
- People with low income, less than or equal to 2 lakh, can be either defaulters or non-defaulters.

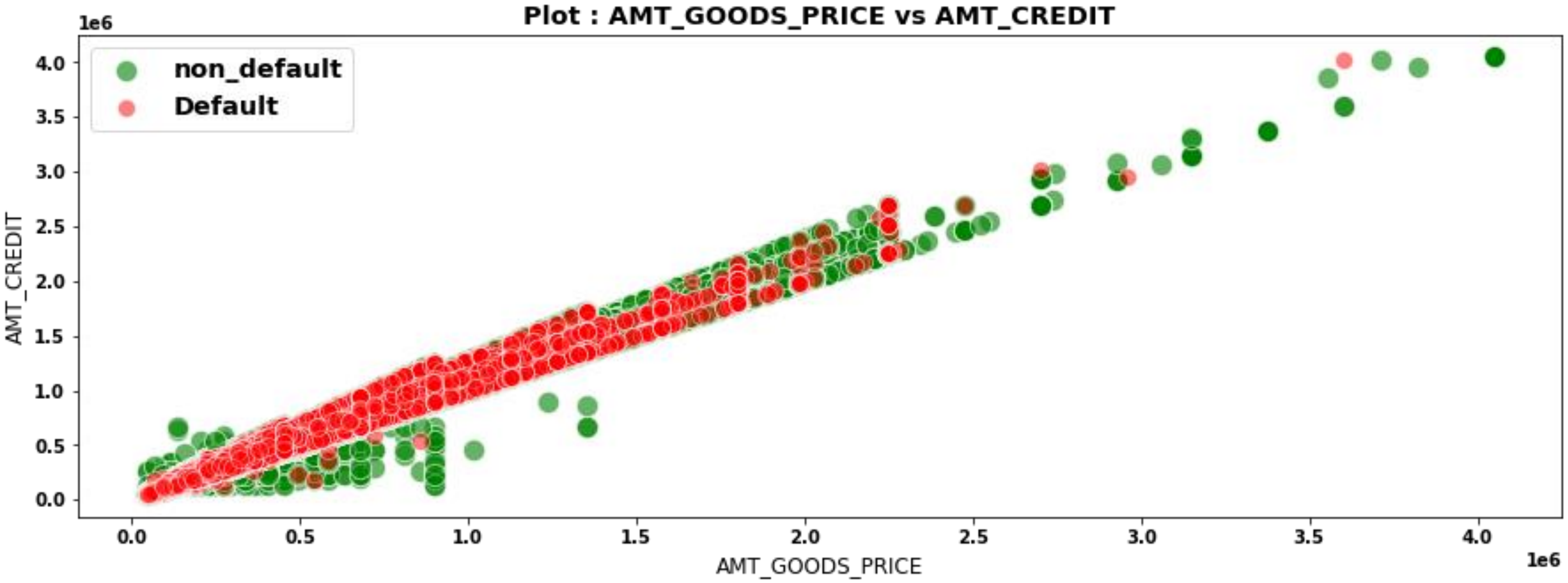
Numerical - Categorical Bivariate Analysis



Observation

- Above graph shows that the people who have not used offer earlier have defaulted more even when there average income is higher than others

Numerical - Numerical Bivariate Analysis

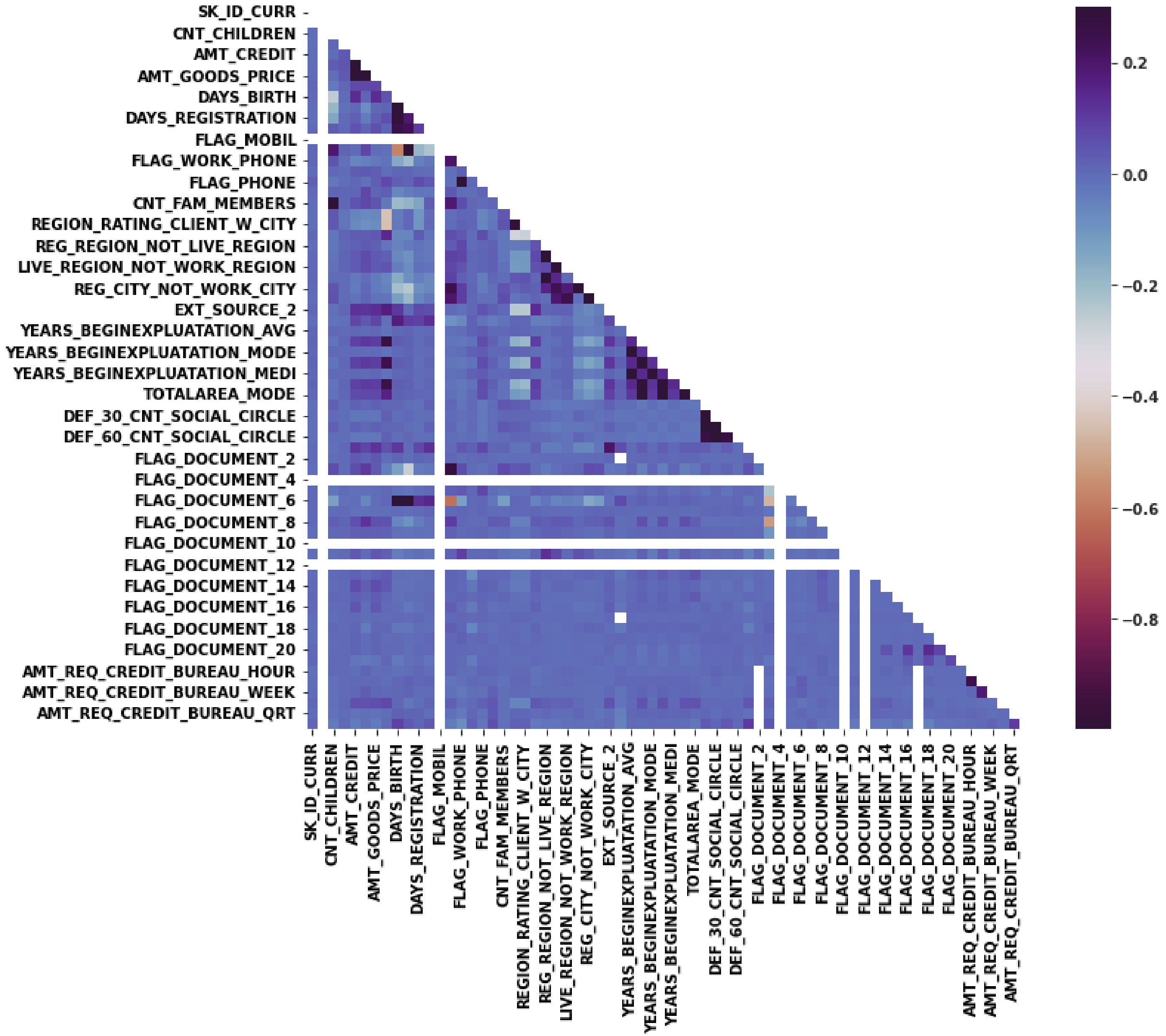


Observation

- People with low AMT_CREDIT and AMT_GOODS_PRICE have a high default rate
- People with high AMT_CREDIT and AMT_GOODS_PRICE have a low default rate.

Top 10 Correlations for Non-Defaulters

1. DAYS_EMPLOYED - FLAG_EMP_PHONE 0.999756
2. OBS_60_CNT_SOCIAL_CIRCLE
   OBS_30_CNT_SOCIAL_CIRCLE
   0.998508
3. AMT_GOODS_PRICE - AMT_CREDIT 0.987250
4. REGION_RATING_CLIENT_W_CITY
   REGION_RATING_CLIENT
   0.950149
5. CNT_CHILDREN - CNT_FAM_MEMBERS 0.878571
6. REG_REGION_NOT_WORK_REGION
   LIVE_REGION_NOT_WORK_REGION
   0.861861
7. LIVE_CITY_NOT_WORK_CITY -
   REG_CITY_NOT_WORK_CITY
   0.830381
8. AMT_ANNUITY - AMT_GOODS_PRICE 0.776686
9. AMT_CREDIT - AMT_ANNUITY 0.771309
10. DAYS_EMPLOYED - DAYS_BIRTH 0.626114



Top 10 Correlations for Defaulters

1. DAYS_EMPLOYED - FLAG_EMP_PHONE 0.999705

2. OBS_60_CNT_SOCIAL_CIRCLE
OBS_30_CNT_SOCIAL_CIRCLE
0.998269

3. AMT_GOODS_PRICE - AMT_CREDIT 0.983103

4. REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
0.956637

5. CNT_CHILDREN - CNT_FAM_MEMBERS 0.885484

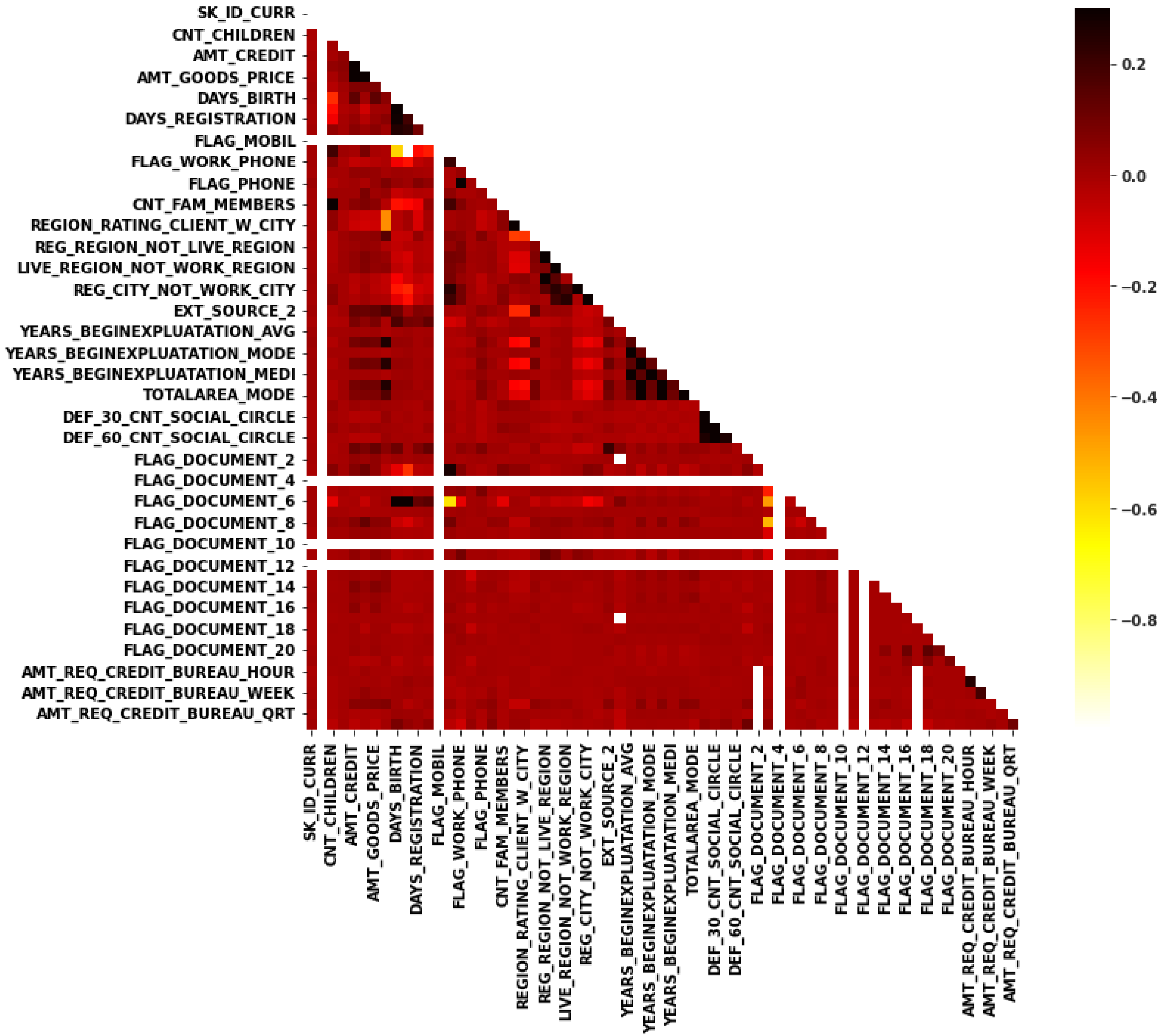
6. LIVE_REGION_NOT_WORK_REGION
REG_REGION_NOT_WORK_REGION
0.847885

7. REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
0.778540

8. AMT_ANNUITY - AMT_GOODS_PRICE 0.752699

9. AMT_CREDIT - AMT_ANNUITY 0.752195

10. DAYS_EMPLOYED - DAYS_BIRTH 0.582185



Conclusions

After analyzing the datasets, there are a few characteristics of a client that the bank can use to determine whether or not they will repay the loan. The analysis is comprised of the following factors and categorization:

- ❖ Factors to determine whether an applicant will be a Re-payer
- ❖ Decisive Factor whether an applicant will be Defaulter

Factors to determine whether an applicant will be a Re-payer

- **CODE_GENDER:** Females default less.
- **NAME_CONTRACT_TYPE:** The default rate on revolving loans is lower.
- **NAME EDUCATION TYPE:** There are fewer defaults for Academic degree.
- **NAME INCOME TYPE:** There are no defaults for students or business owners.
- **ORGANIZATION TYPE:** Less than 3% of clients with Trade Types 4 and 5 and Industry Type 8 have defaulted.
- **DAYS BIRTH:** People over the age of 50 have a low likelihood of defaulting.
- **DAYS EMPLOYED:** Clients with 40+ years of experience have a default rate of less than 1%.
- **AMT INCOME TOTAL:** Applicants earning more than $700,000 are less likely to default.
- **CNT CHILDREN:** People with 0 to 2 children are more likely to repay their loans.

Decisive Factor whether an applicant will be Defaulter

- **CODE GENDER:** Men have a higher default rate.
- **NAME FAMILY STATUS:** People who have civil marriages or are single frequently default.
- **NAME EDUCATION TYPE:** People with a Lower Secondary and Secondary education are more likely to default.
- **NAME INCOME TYPE:** Clients on Maternity Leave OR Unemployed frequently default.
- **OCCUPATION TYPE:** Avoid low-skilled laborers, drivers, waiters/bartenders, security personnel, laborers, and cooks
- because the default rate is extremely high.
- **ORGANIZATION TYPE:** Transport: type 3 , Industry: type 13 , Industry: type 8 , and Restaurant: type 3 are the organizations with the highest percentage of loans not repaid (less than 12 percent). Self-employed people have a high default rate and should be avoided when applying for a loan or providing a loan with a higher interest rate to mitigate the risk.
- **DAYS BIRTH:** Avoid young people between the ages of 20 and 40 because they are more likely to default.
- **DAYS EMPLOYED:** People with less than five years of experience have a high default rate.
- **CNT CHILDREN & CNT FAM MEMBERS:** Clients with children equal to or greater than 9 default 100%, and their applications will be rejected.

Thank You

Submitted By
Shivam Sharma