EDA CREDIT CASE STUDY ASSIGNMENT

SUBMISSION

By:

Shireen Dash

Ketki Kale

Business Understanding

Consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision :

- **Interest loss**: If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- **Credit loss**: 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.

Business Objectives

Driving Factors or Driver Variables behind loan default, i.e. the variables which are strong indicators of default.

Problem Solving Methodology

Domain Understanding:

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

Data Understanding

Data Sets provided are:

- **Application_data.csv**: Information of the client at the time of application. It contains whether the client has difficulties in payment.
- **Previous_application.csv**: Information of the client about previous loan application. It contains whether the client has been approved, cancelled, refused and unused offer.
- **Columns_description.csv**: It is data dictionary which describes the meaning of the variables.

Shape of Data Sets:

- Application_data.csv has 307511 rows and 122 columns.
- Previous_application.csv has 1670214 rows and 37 columns.

To obtain the attributes that influence the tendency to default, we will be using TARGET variable:

- **Target=1:** It indicates the client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample
- **Target= 0:** all other cases.

Data Cleaning

Data Cleaning Part I:

1. **For Application_data.csv:**
   - Finding the percentage of missing values in the columns and removing the columns with more than 20% missing values.
   - Imputing missing values for those columns which have less than 13% missing values in the columns.
   - Correcting the datatypes of all columns.
   - Checking out outliers for the numerical columns and treating them.
   - Binning of continuous variables for segmented analysis.
2. **For Previous_application.csv:**
   - Finding the percentage of missing values in the columns and removing the columns with more than 40% missing values.

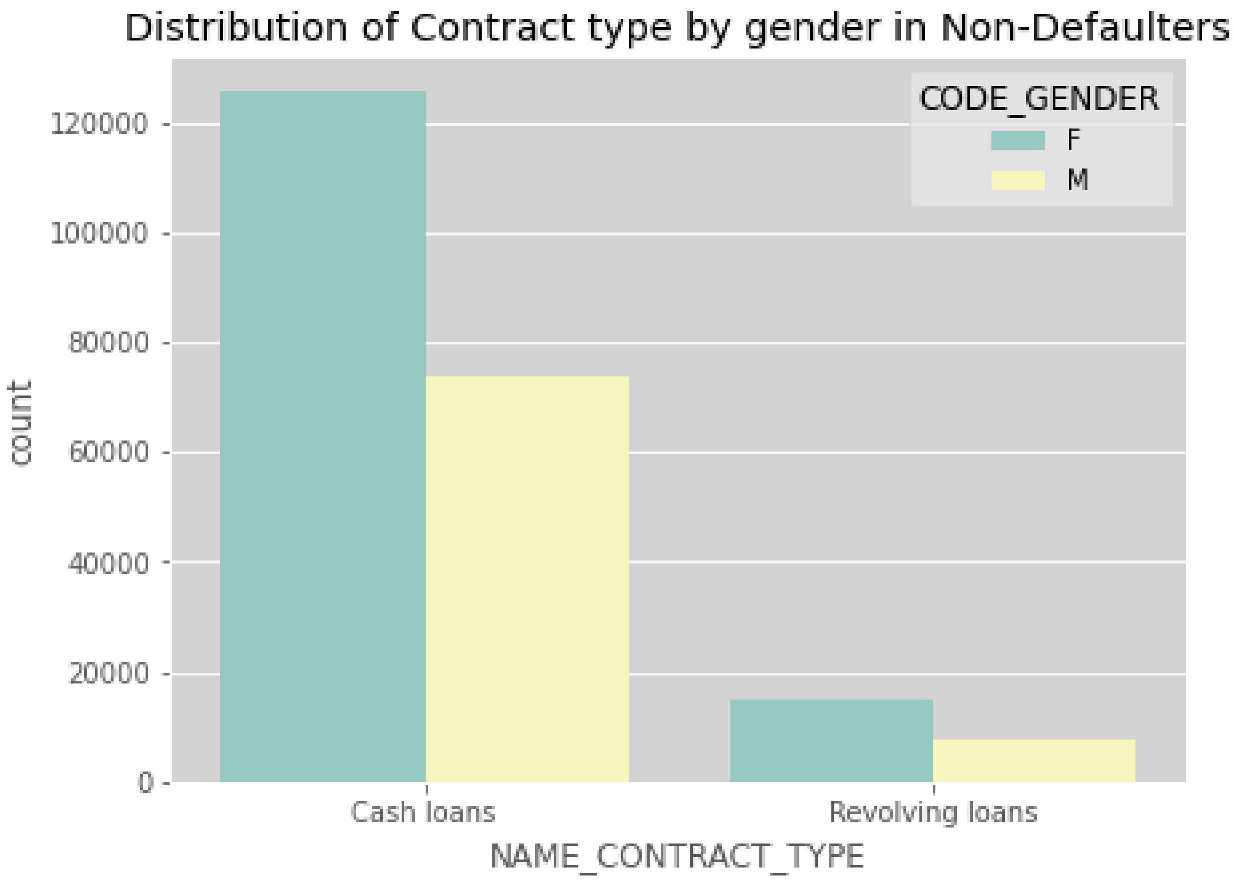
Data Cleaning Part II:

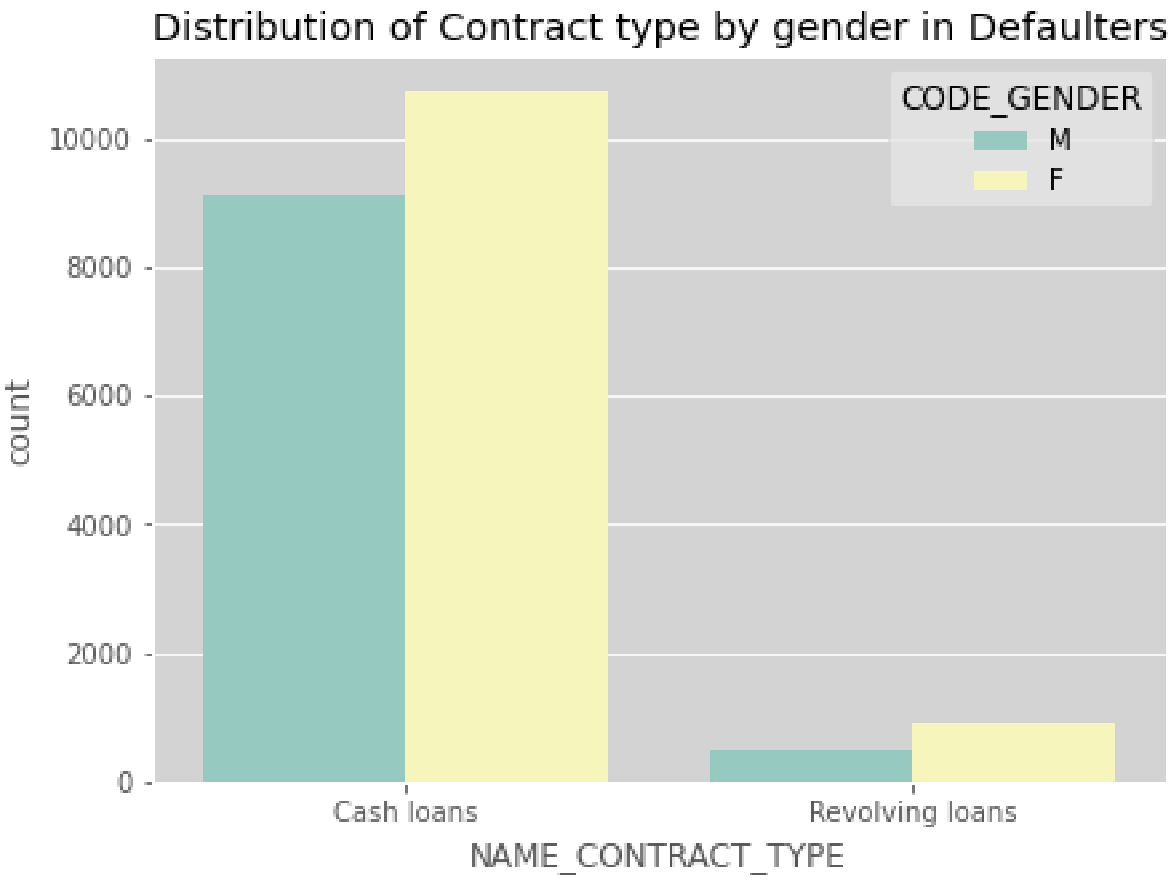
- Merging of Application_data.csv and Previous_application.csv.
- Dropping missing values in the merged data set.
- Shape was found to be 314226 rows and 64 columns.

I. Analysis of application_data.csv data set:

1. Obtaining the data imbalance ratio by using TARGET variable. The Data imbalance ratio was found to be 10.45(approx.).
2. Later, we segregated the data set into 2 sets: Defaulters(TARGET=1) and Non-defaulters(TARGET=1).
3. Performed univariate analysis on categorical variables for both defaulters and non-defaulters and compared them.
4. Checked correlation for numerical columns for both defaulters and non-defaulters. It was found to be the same for both defaulters and non-defaulters.
5. Performed univariate analysis on numerical variables for both defaulters and non-defaulters and compared them.
6. Performed bivariate analysis on numerical variables for both defaulters and non-defaulters and compared them.

Univariate Analysis on Contract type:

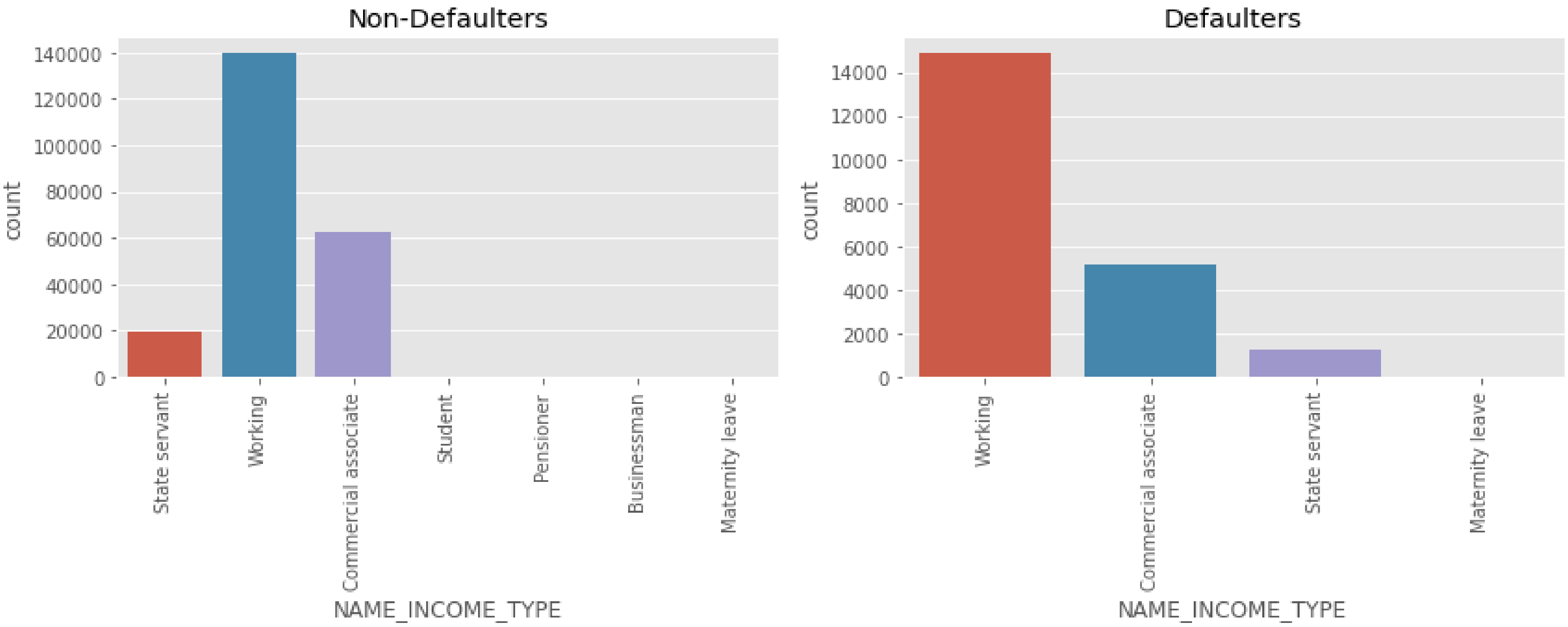




Observation:

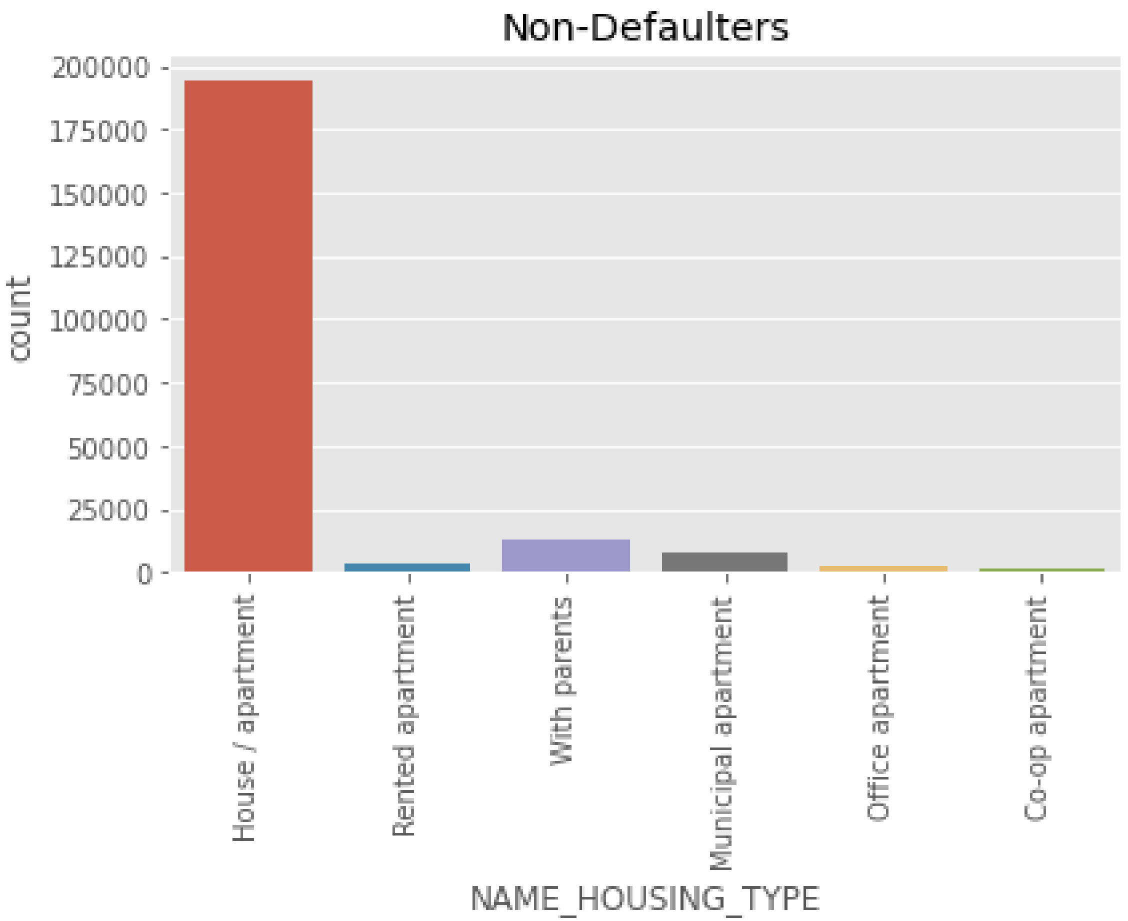
1. Contract-type ‘cash loans’ are applied more than ‘Revolving loans’.
2. In both the cases, Females have applied for more loans than males.

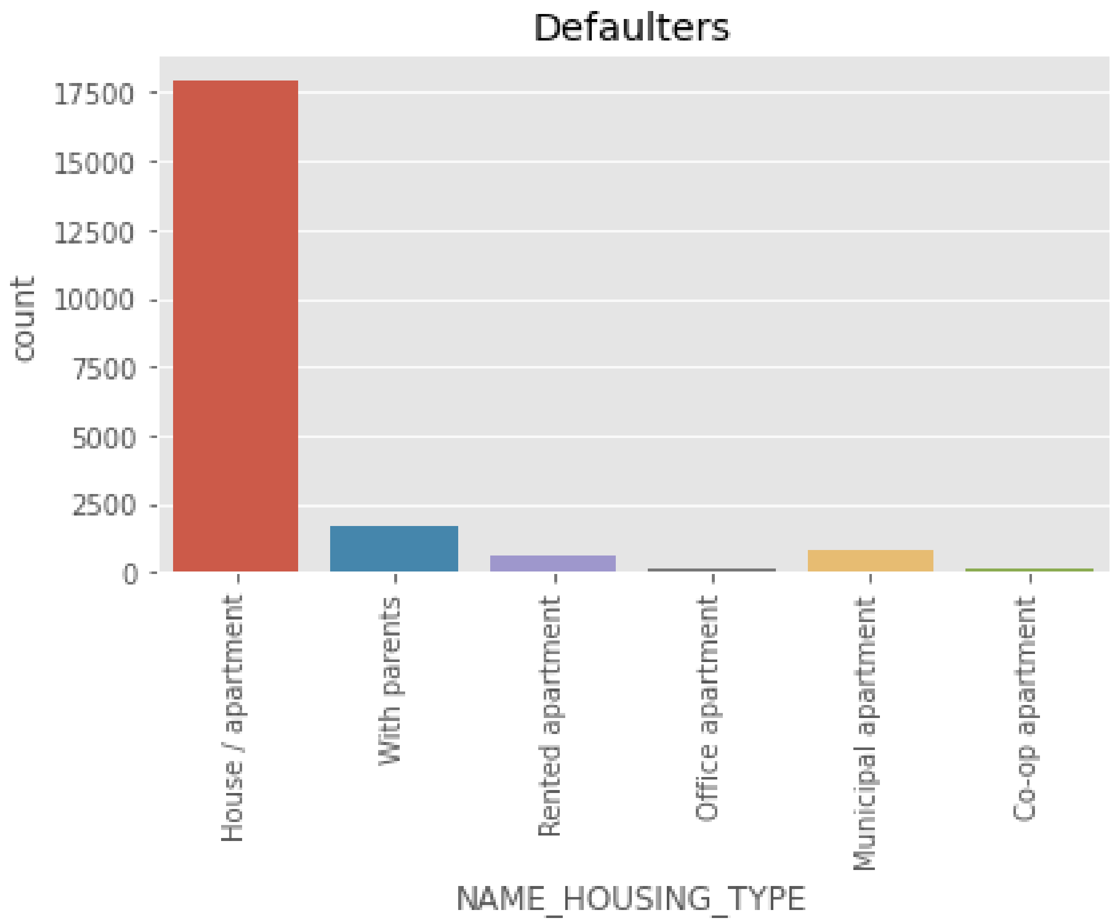
Univariate Analysis on Income type:



Observation: State servant, Commercial Associate and Working professionals have higher non-defaulters.

Univariate Analysis on Housing type:

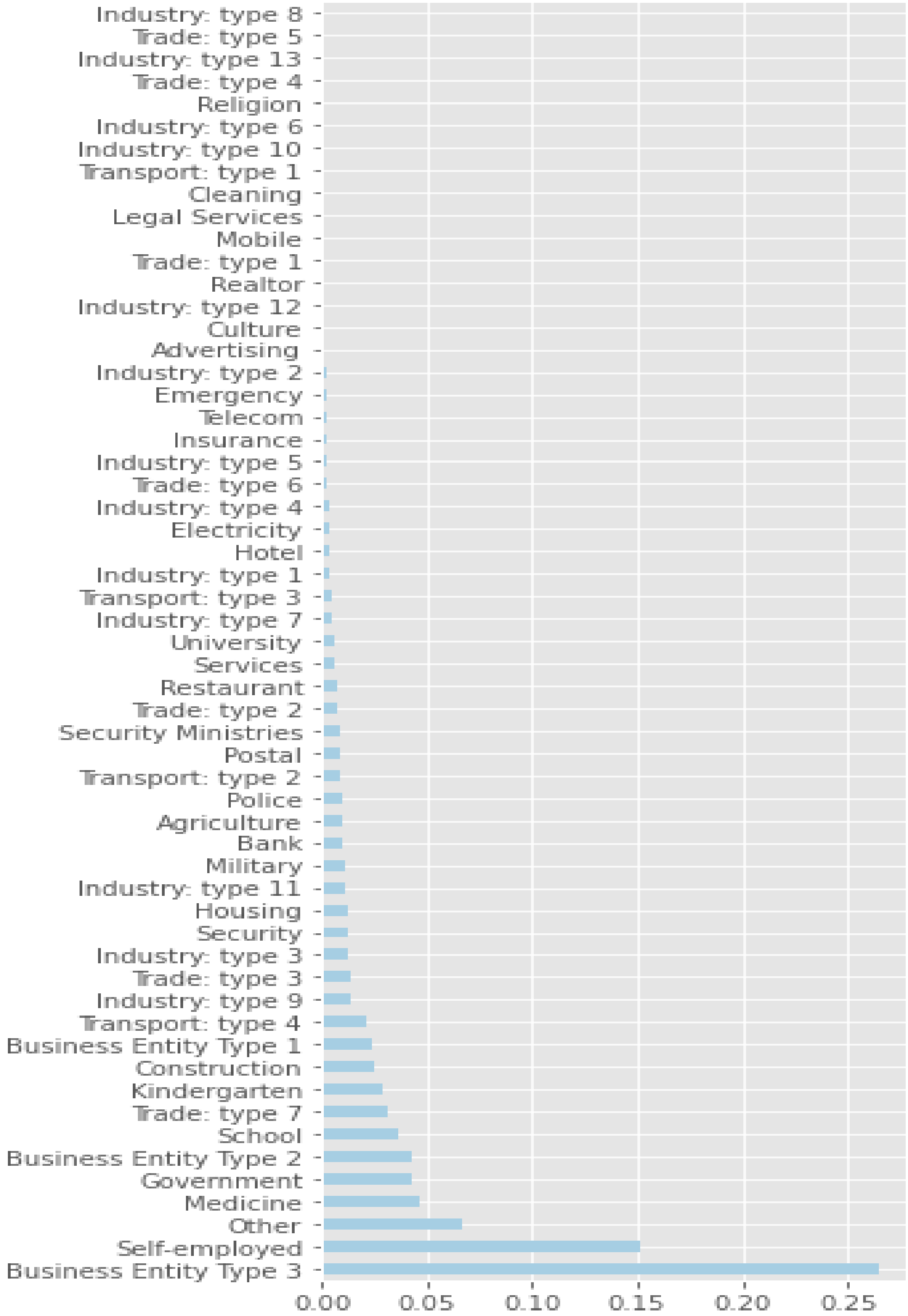


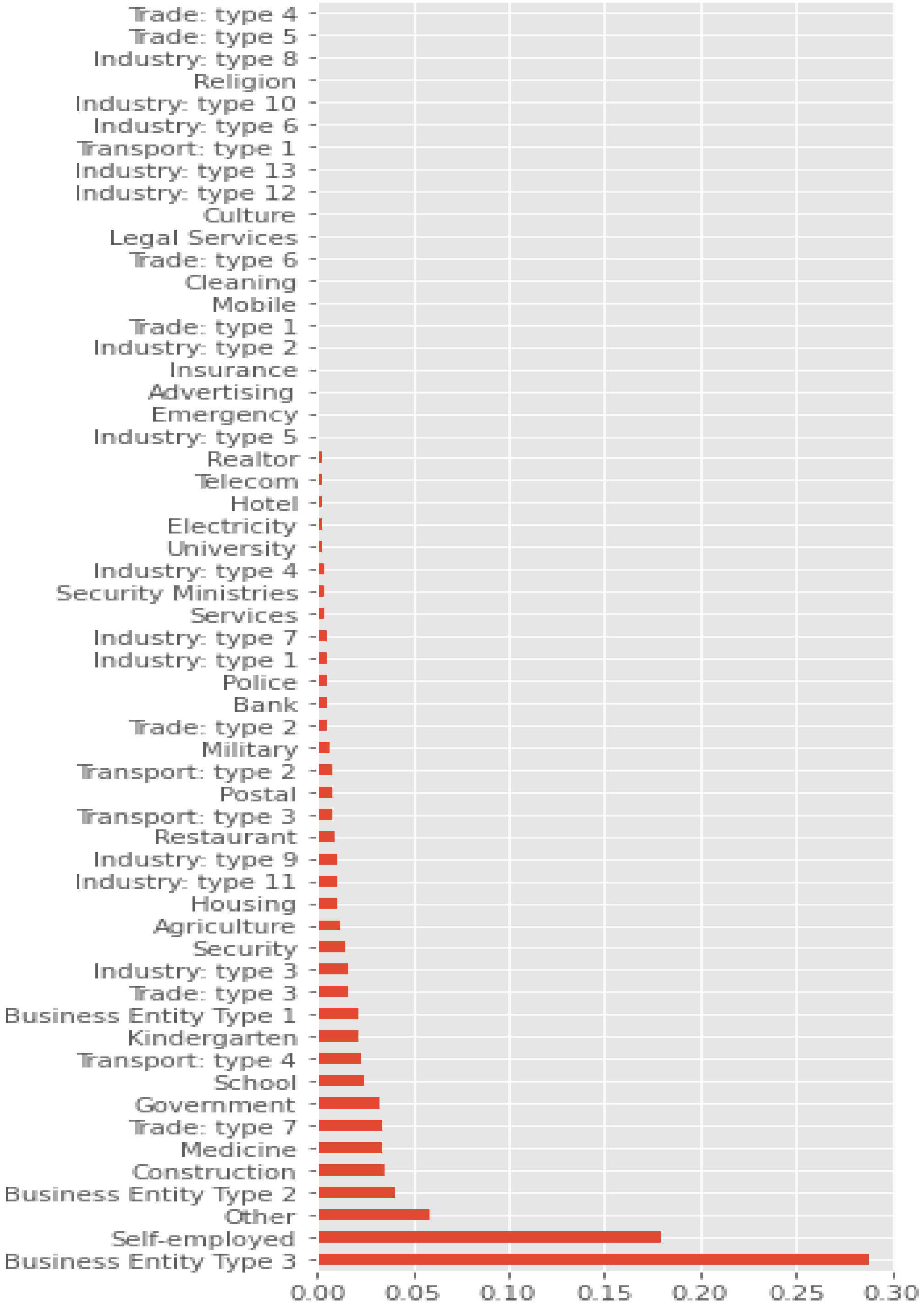


Observation: People with own house/apartment or living 'with parents' are the category of people who have applied for maximum loans.

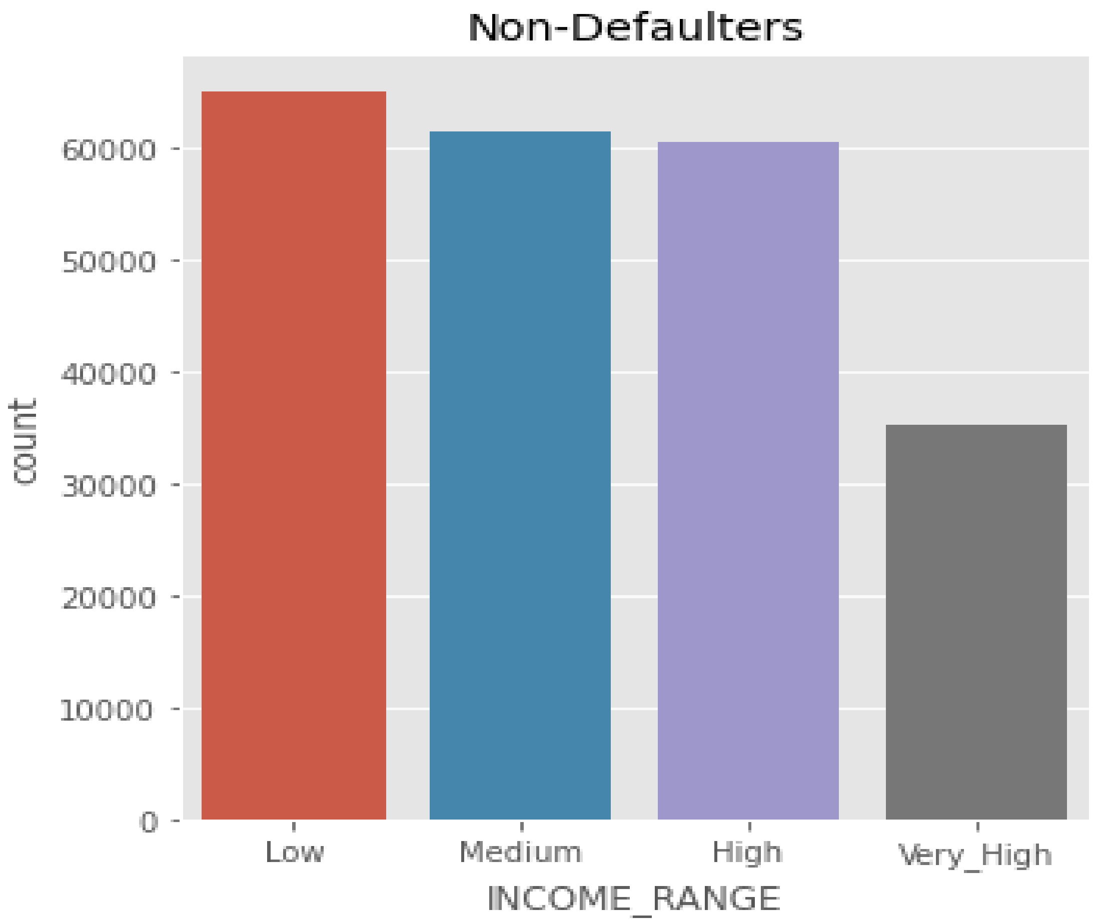
Univariate Analysis on Organisation type:

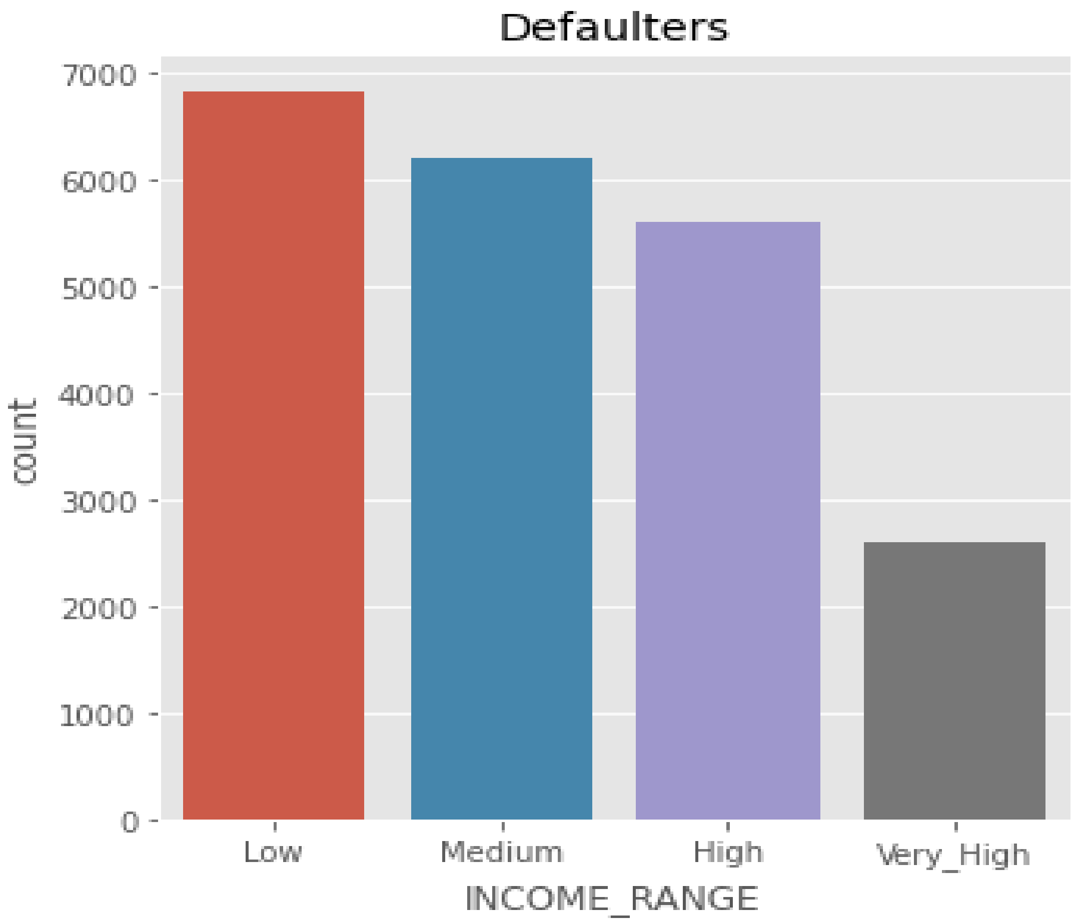
Observation: Most of the loans have been applied from the organization types - 'Business Type 3', 'Self-employed', 'Other'.





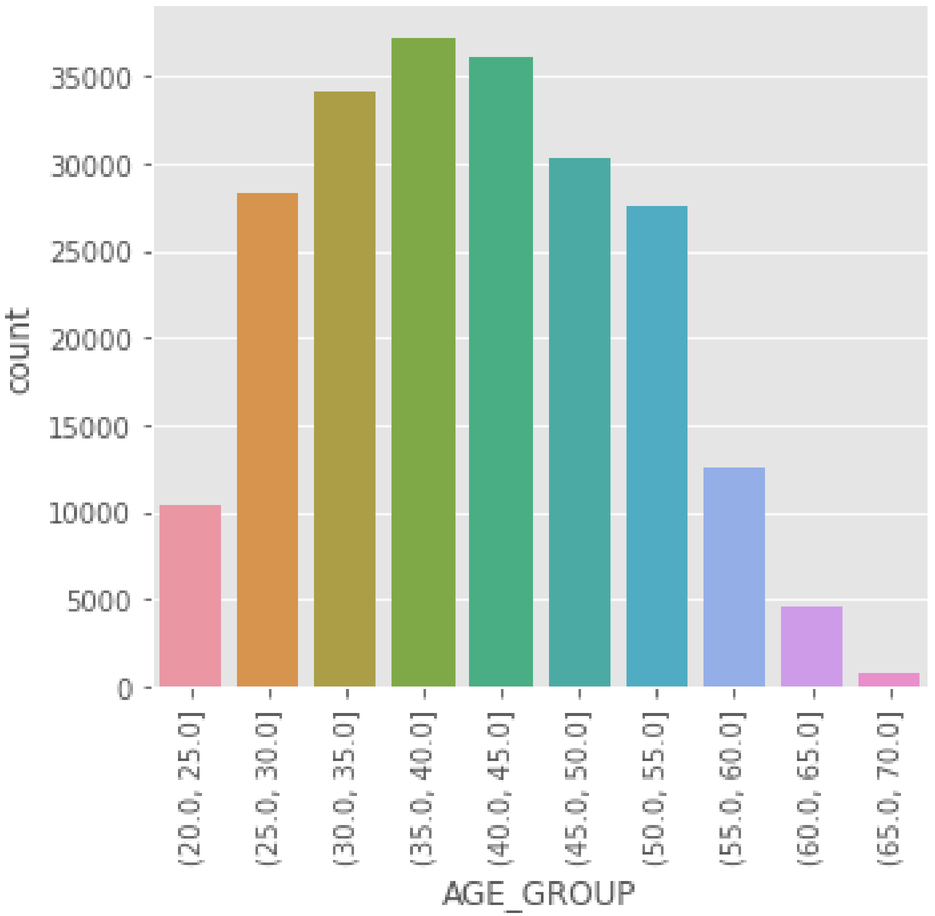
Segmented Univariate Analysis on Income Range:

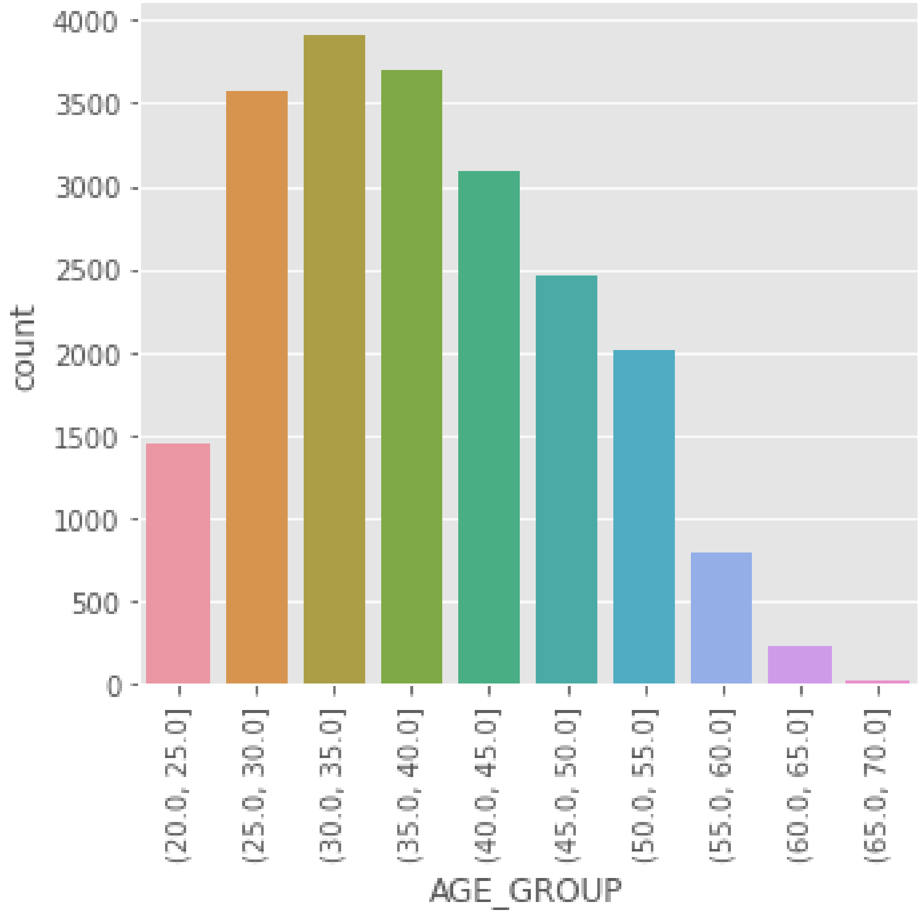




Observation: Most of the applied loans are from the category - 'Low' which lies in the range 25k to 120k

Segmented Univariate Analysis on AGE_GROUP:

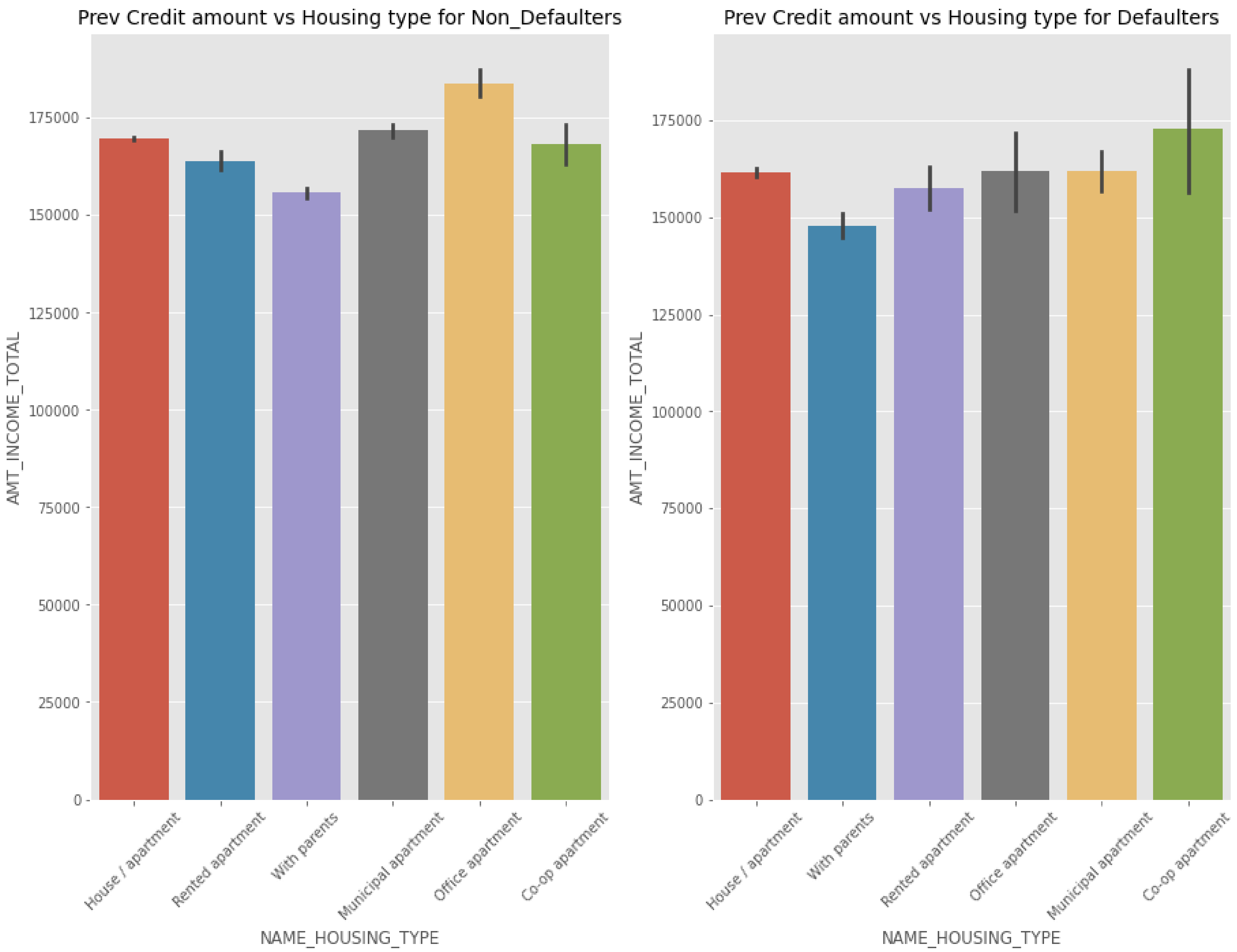




Observation: Most of the loans have been applied by people from the age 25 to 45.

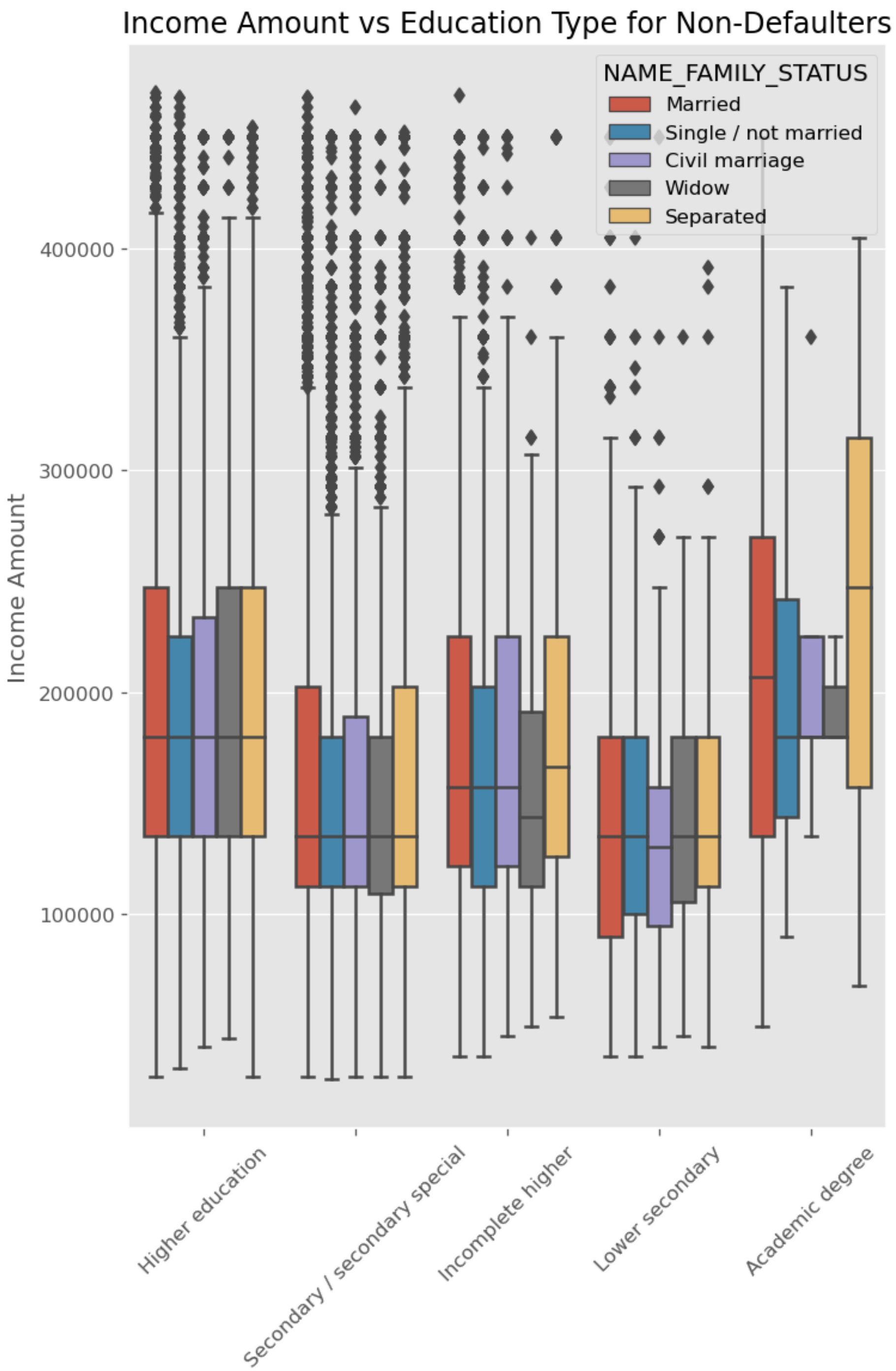
Bivariate Analysis on Housing Type Vs Income:

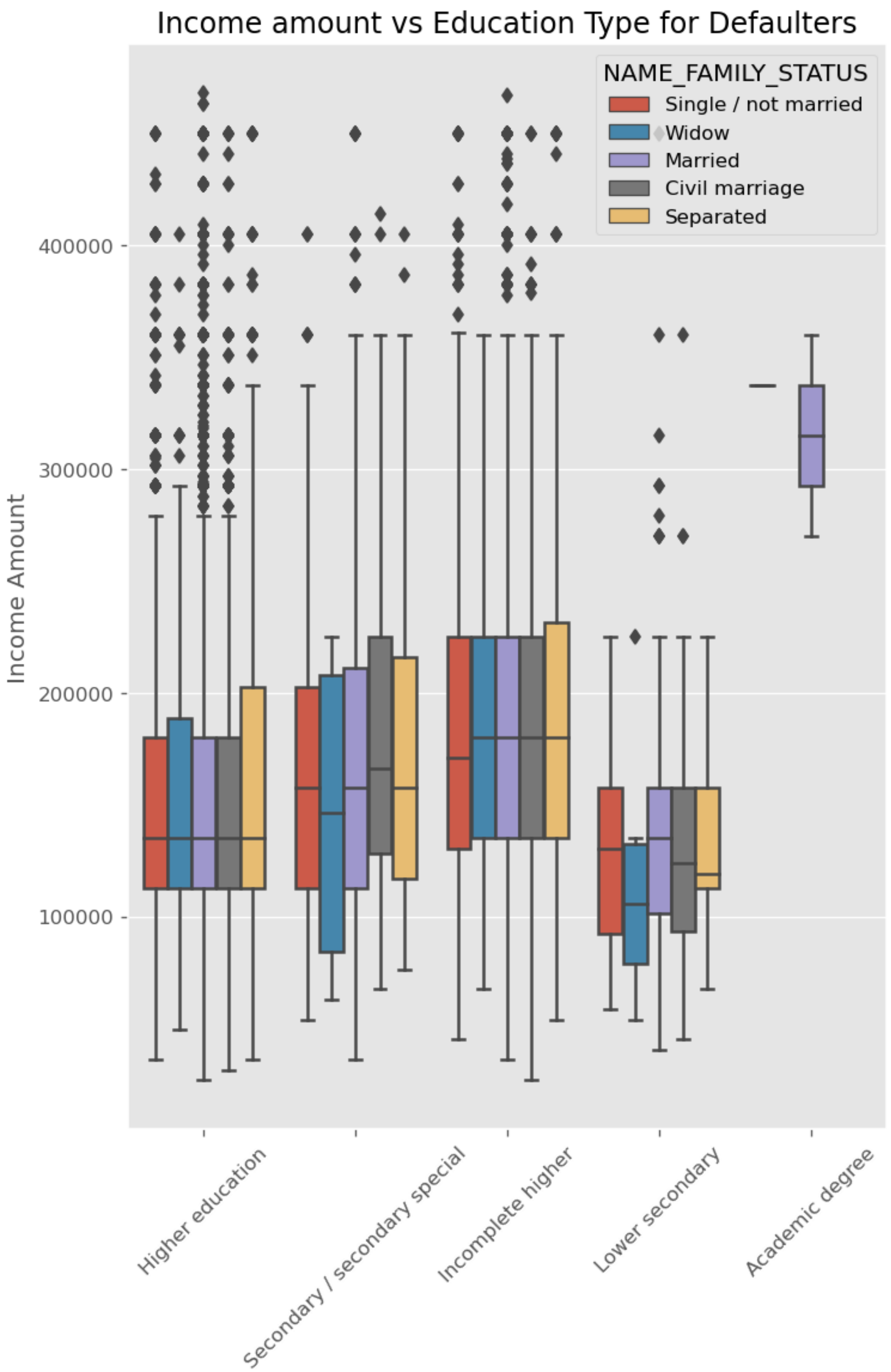
Observation: A significant difference can be seen in housing-types - 'with parents' and 'office apartments'. These categories have higher non-default rate as compared to the default rate, therefore to target these customers can be beneficial.



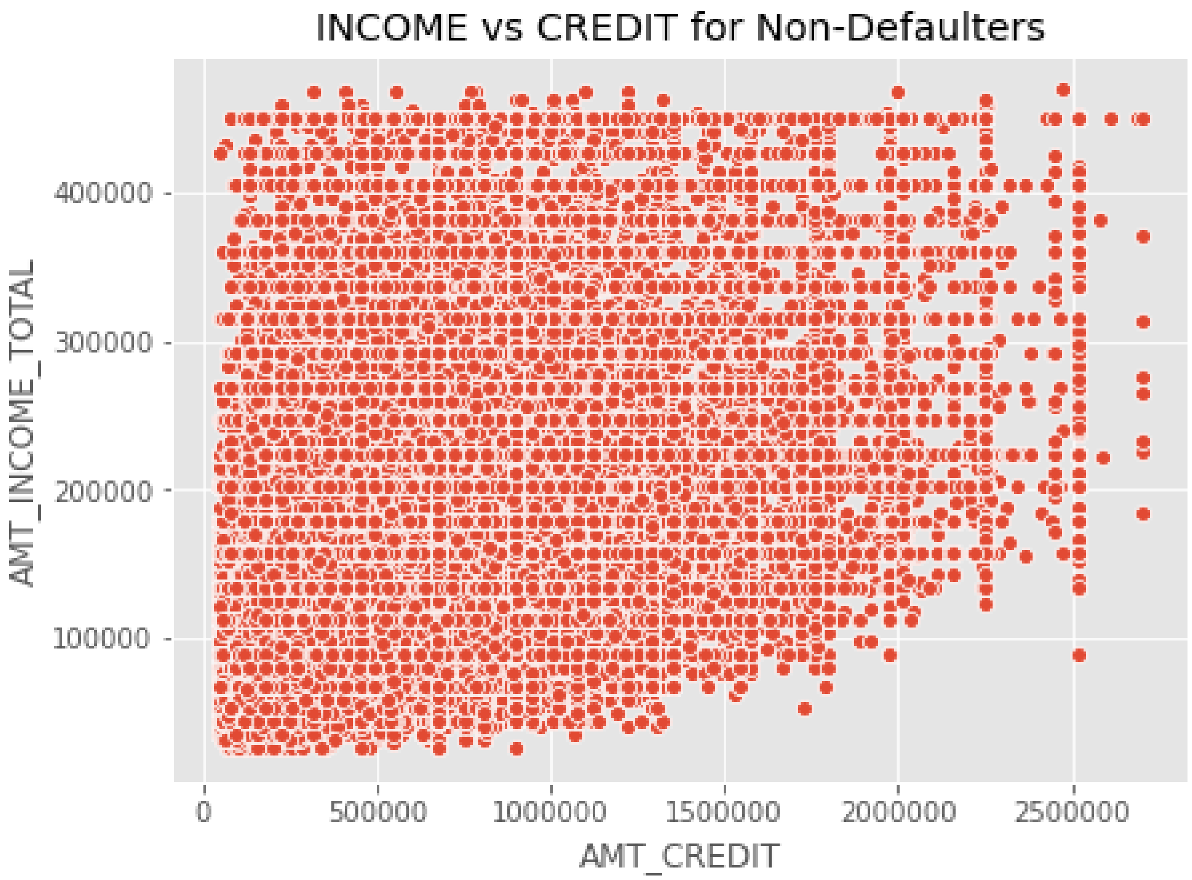
Bivariate Analysis on Education Type Vs Income:

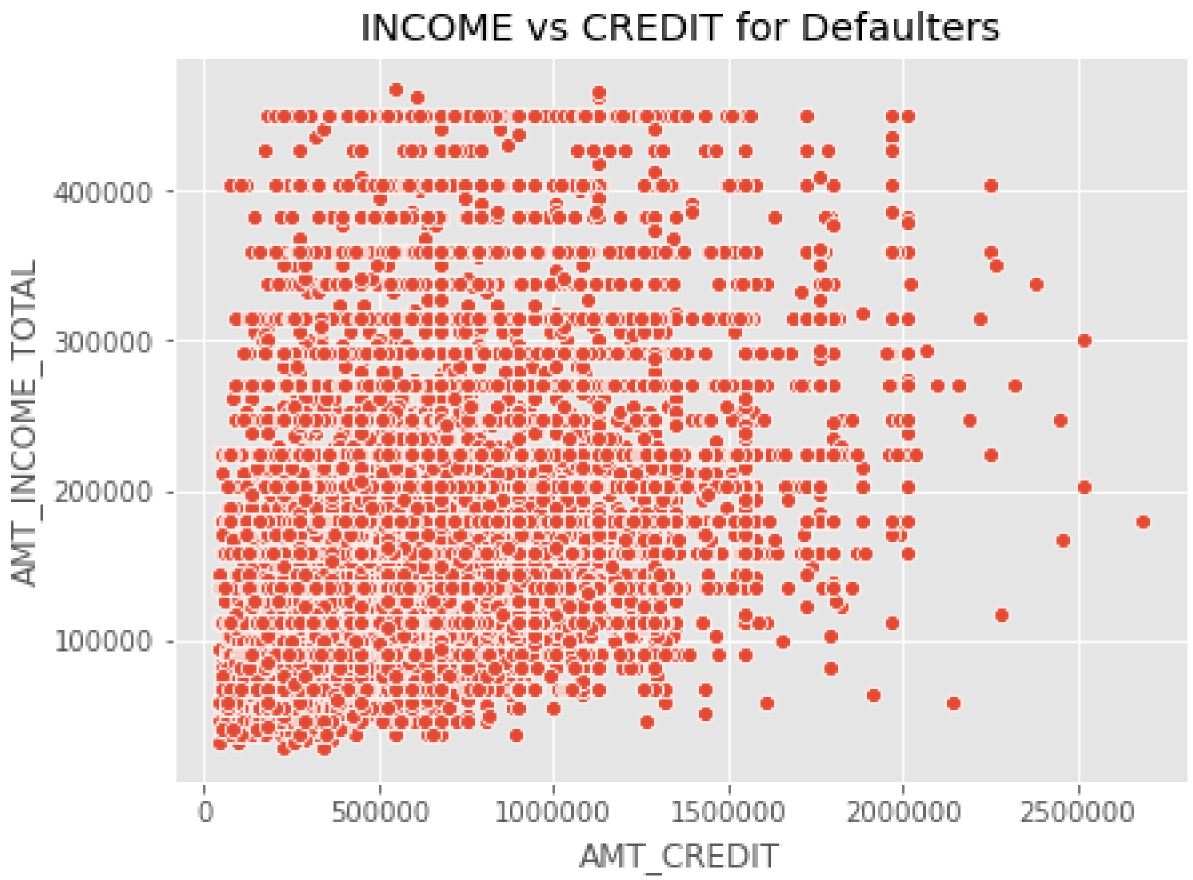
Observation: From the defaulters plot, 'Academic-degree' category has the highest default rate between income range of approx 2.5 lacs to 4.5 lacs and has family status as 'Married'.





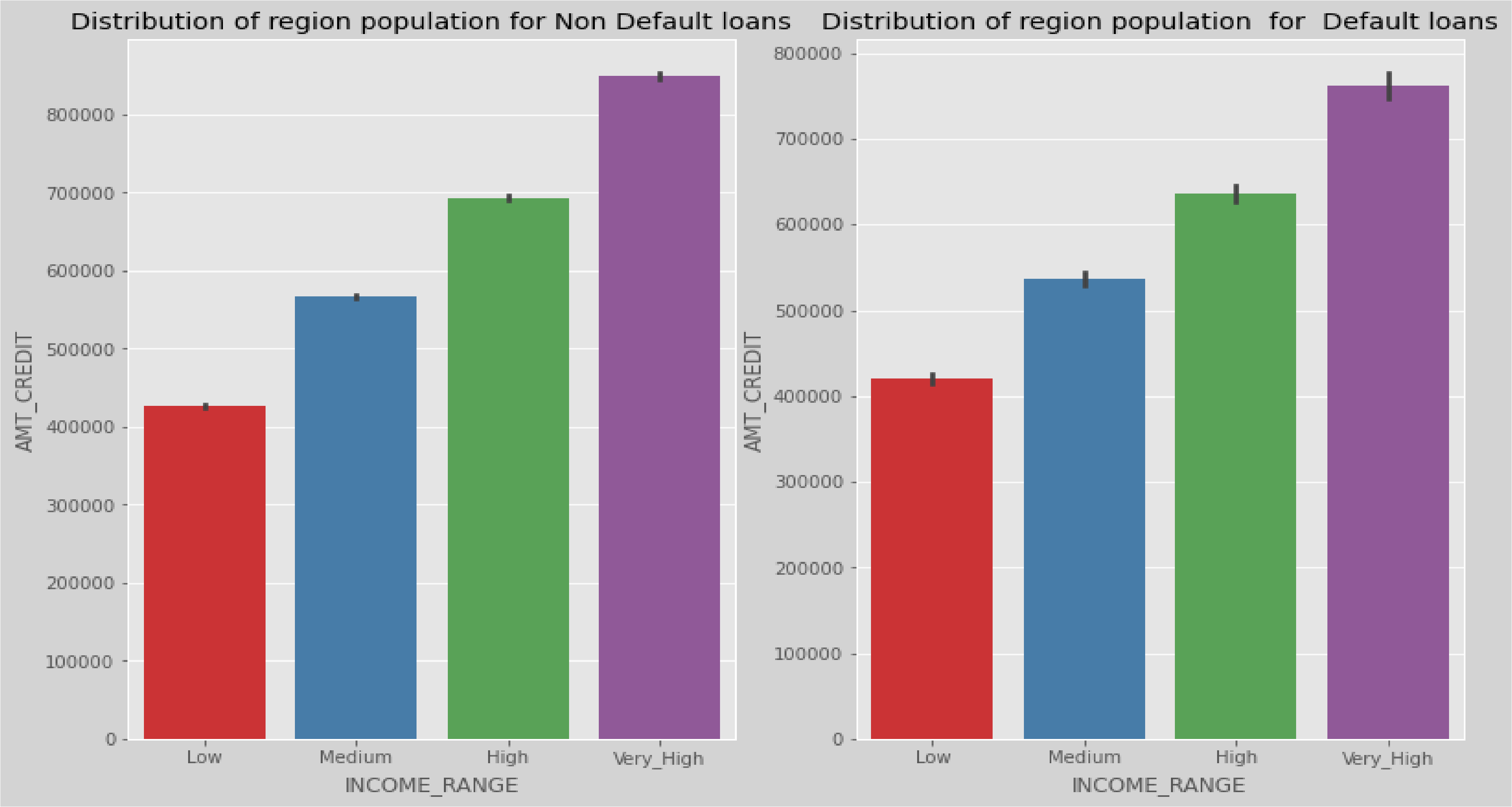
Bivariate Analysis on Income Vs AMT_CREDIT:





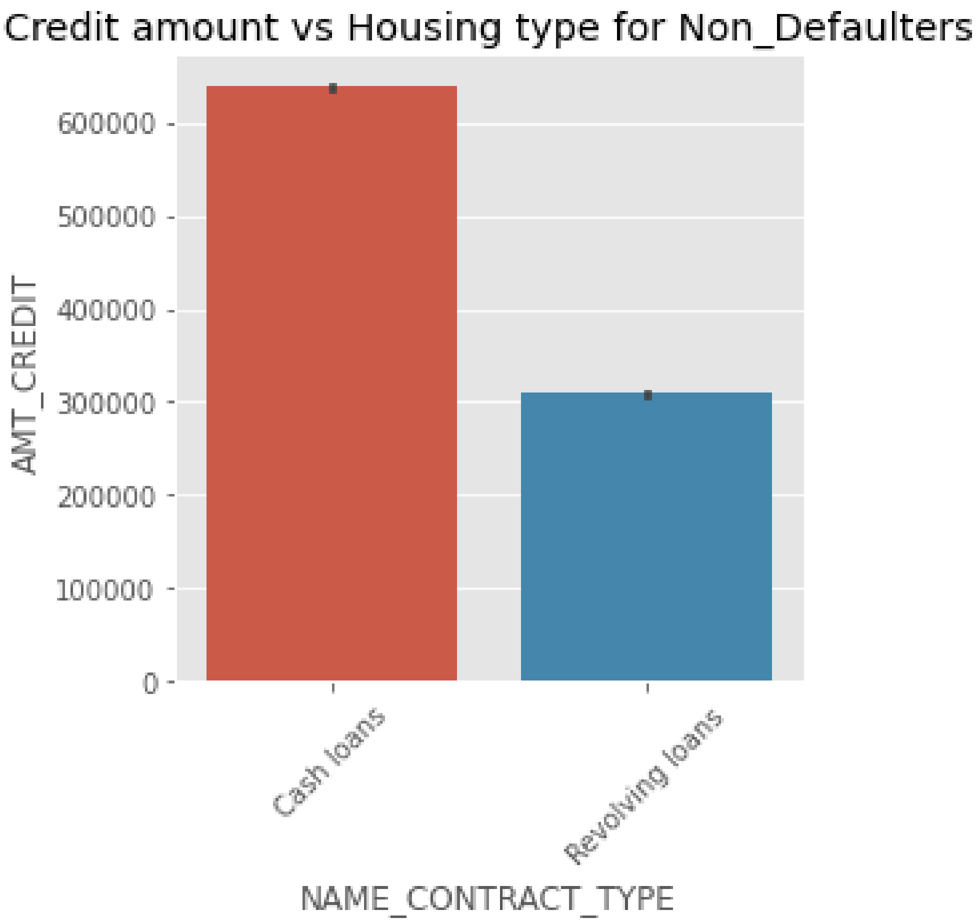
Observation: Lower density of defaults can be seen where income is higher than 3 lacs or credit is greater than 15 lacs.

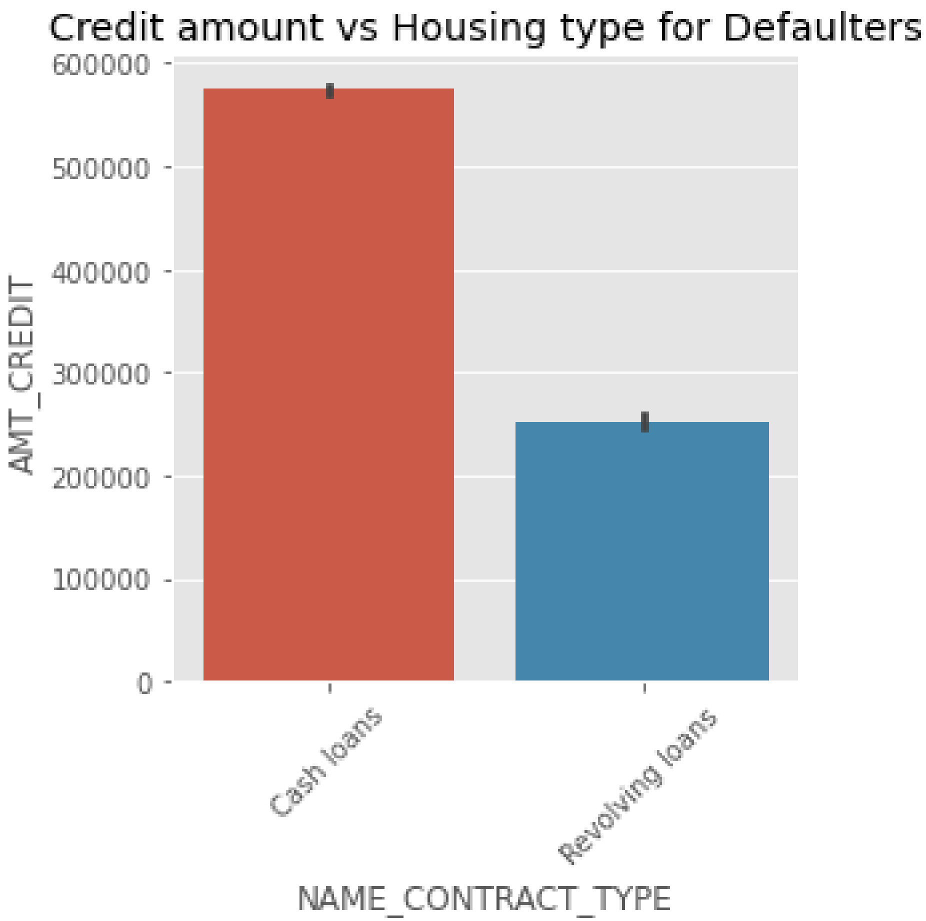
Bivariate Analysis on Income Range Vs AMT_CREDIT:



Observation: People with 'High' and 'Very-high' income range are least likely to default as they have higher non-default rate than default rate.

Bivariate Analysis on Contract Type Vs AMT_CREDIT:



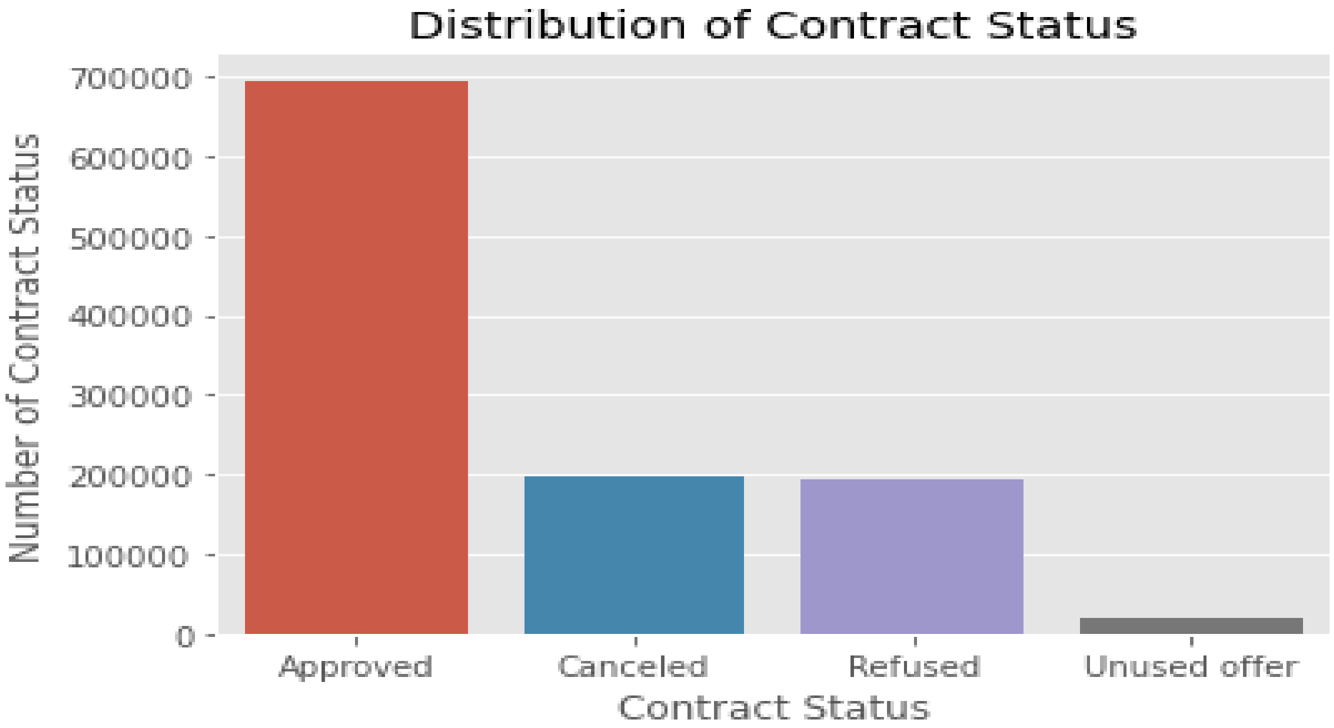


Observation: People with high credit amounts tend to have applied for more cash loans than revolving loans.

II. Analysis of merged data set:

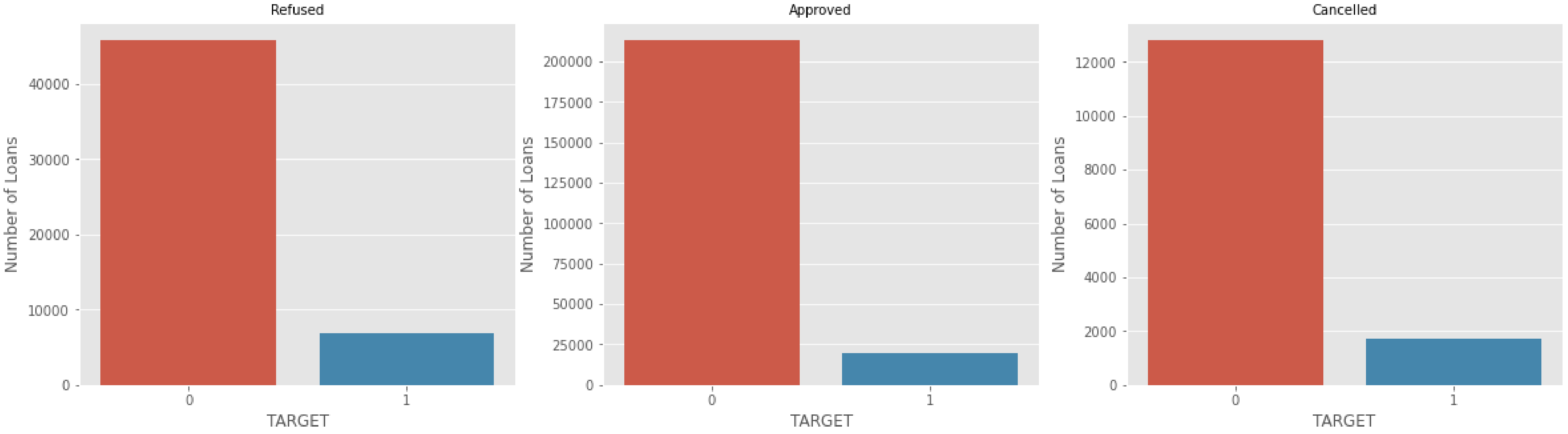
We merged the **Application_data.csv** and **Previous_application.csv** dataset with respect to SK_ID_CURR (ID of loan in our sample) and performed univariate and bivariate analysis on segregated contract statuses (approved, cancelled, refused and unused offer) based on Target=1 are the defaulters and Target=0 are the non-defaulters

.



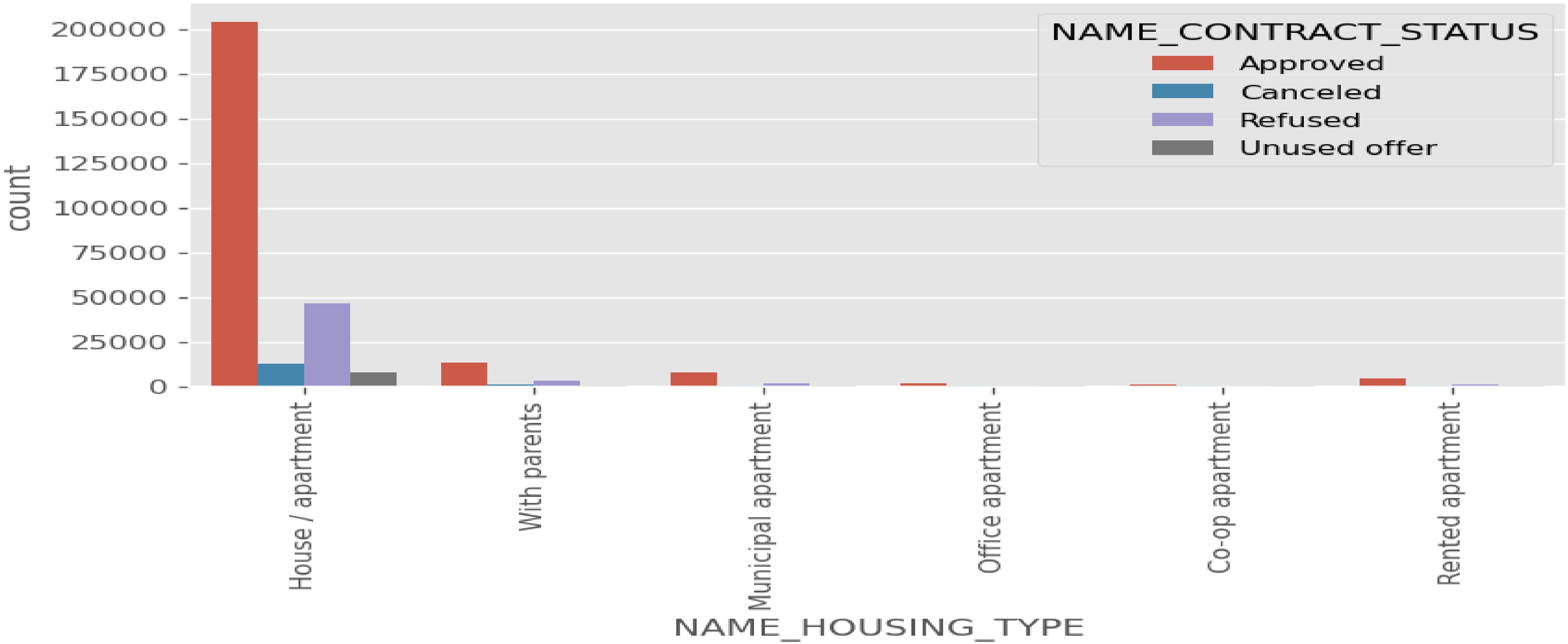
Number of loan applications based on whether the Contract Status is approved, cancelled, refused and unused offer.

Univariate Analysis on Contract Status :



Observation: Loans which have been refused and cancelled before have more chances to default as compared to the approved ones.

Univariate Analysis on Housing type:

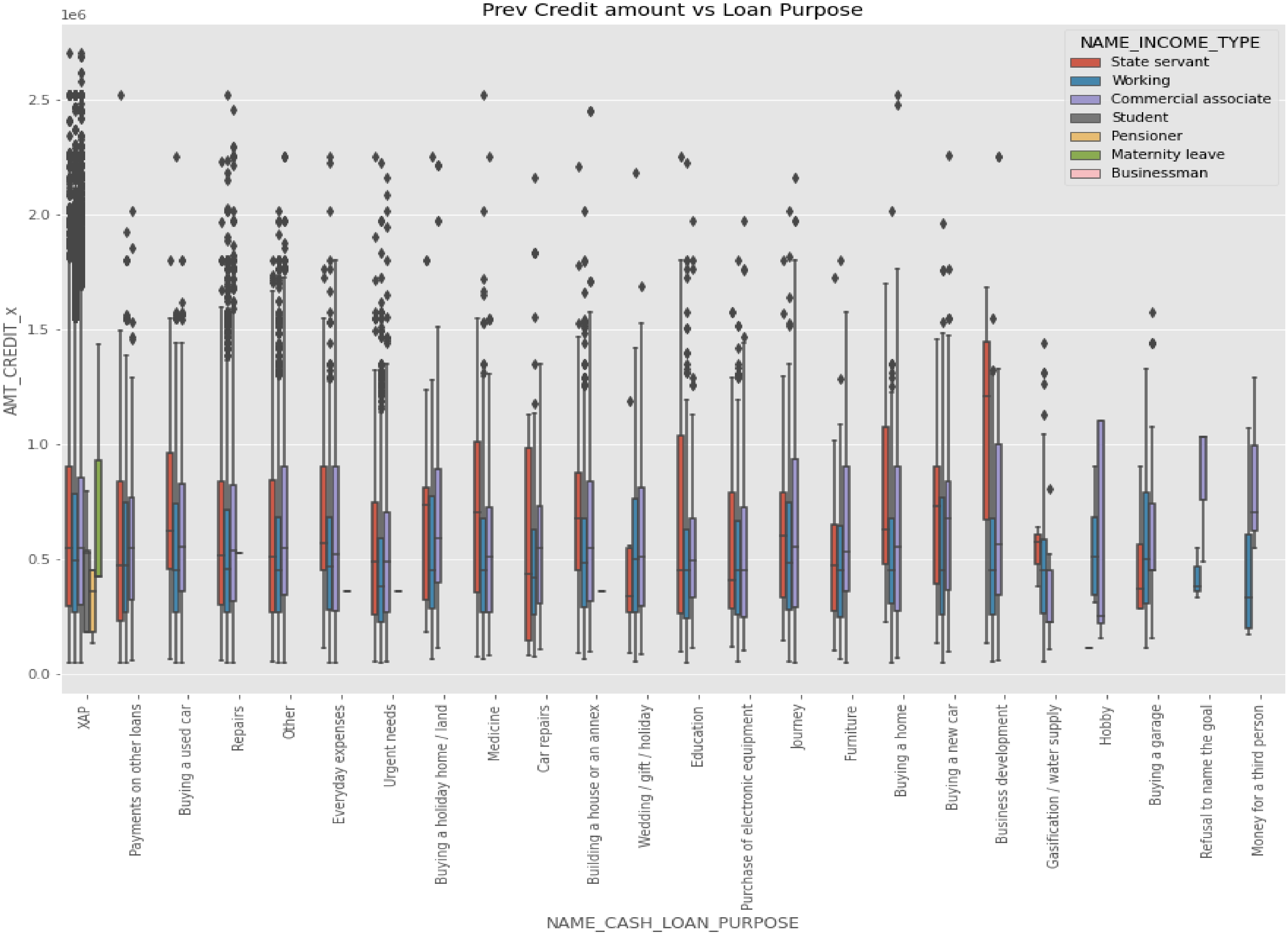


Observation: Highest approved loans are from the housing type 'house/apartment'.

Bivariate Analysis on Previous credit Vs Loan Purpose:

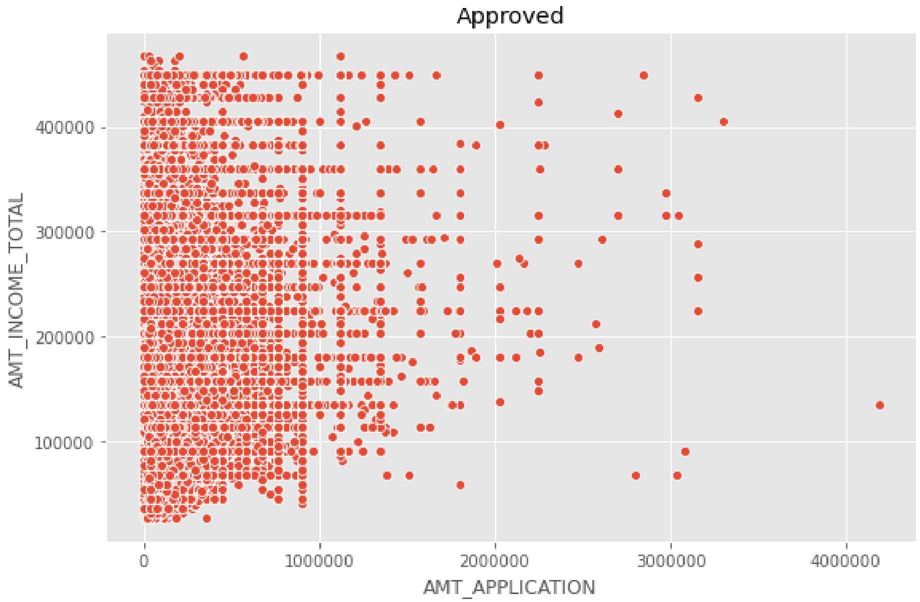
Observation:

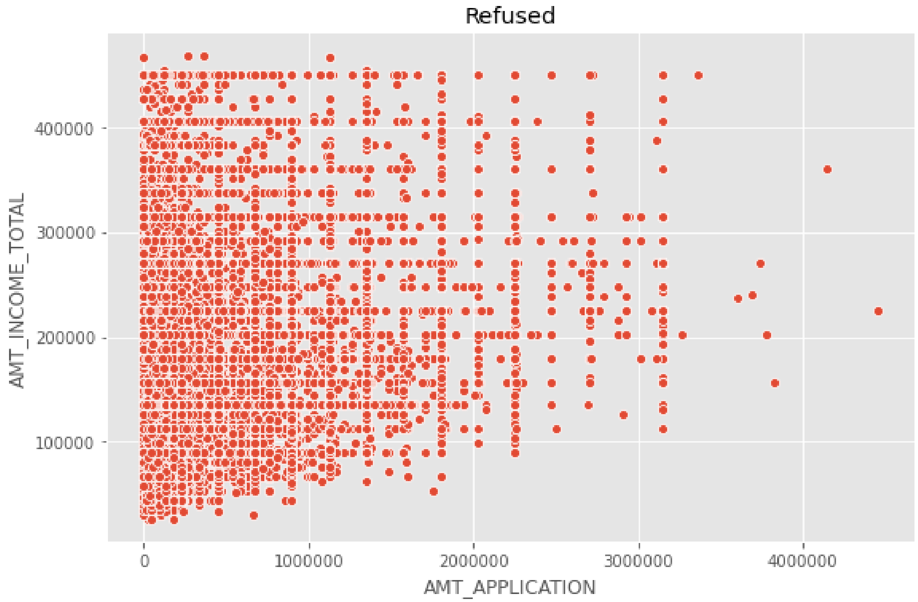
1. Credit amount for loan purposes like education, buying a home, business development is higher.
2. Credit amount of commercial associate for loan purposes like everyday expenses and journey is very high.
3. State Servant income type has significant amount of credit for many loan purposes.

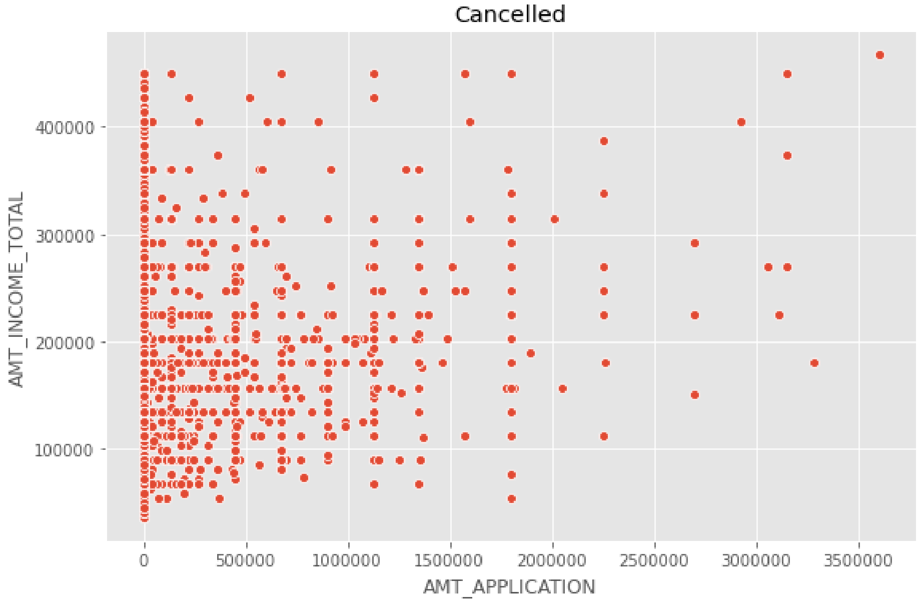


Bivariate Analysis on AMT_Application Vs Income:

Observation:
('AMT_APPLICATION' is - For how much credit did client ask in the previous application).
We can clearly observe - Highly dense cancelled loans fall under the income range of 2 lacs. Therefore, people with income lesser than 2 lacs have more chances of cancelling their loans.

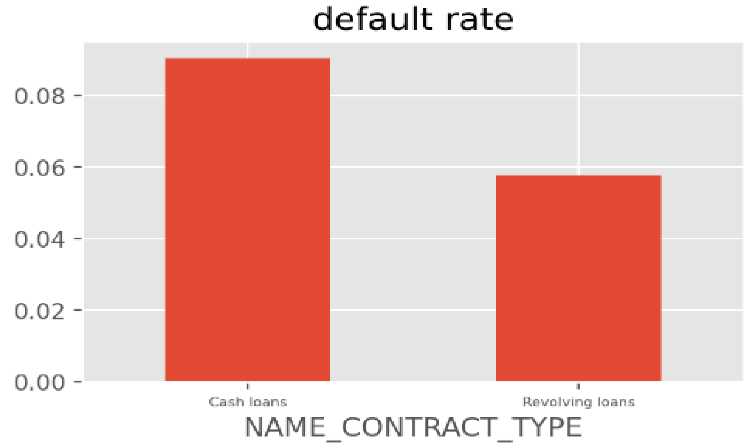


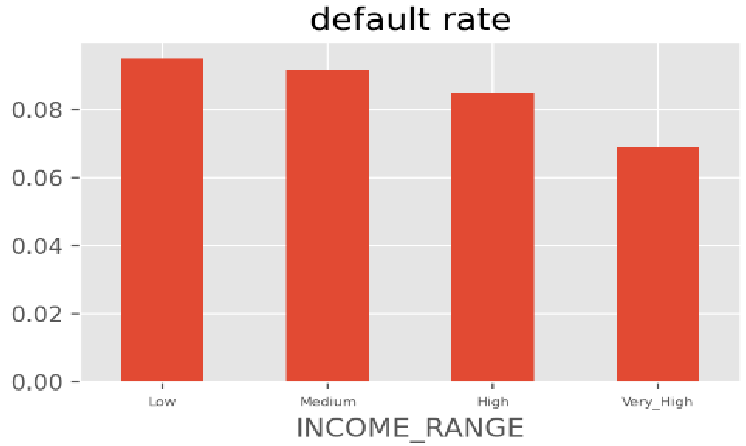


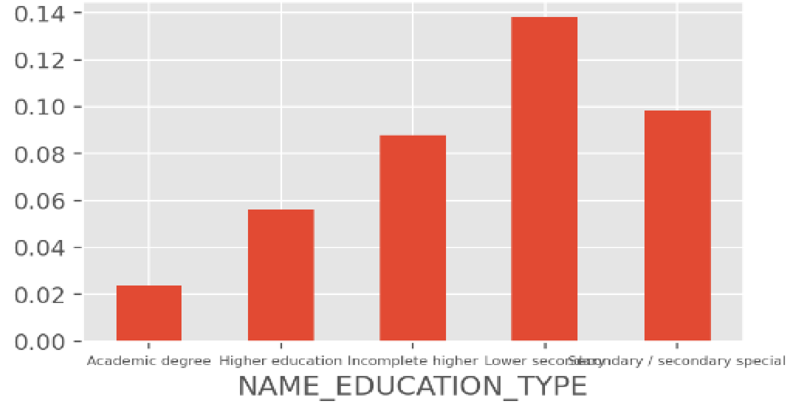


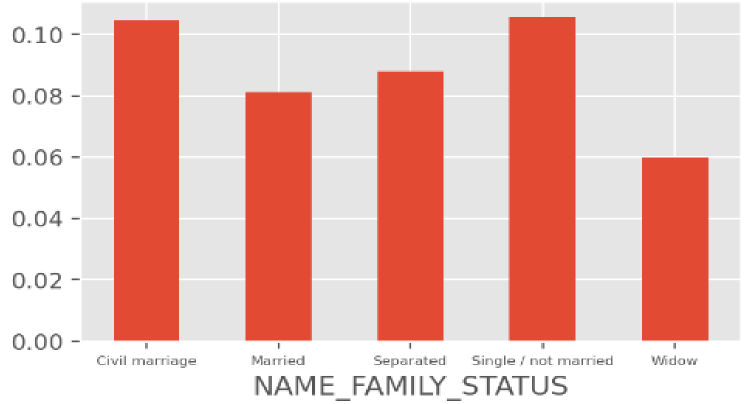
Default Rate Analysis:

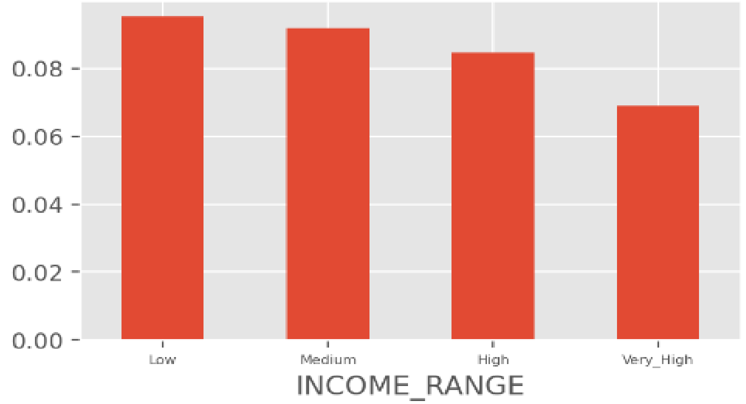
Default rate is a ratio of Number of defaulted / Total Number of Loans in the segment that we are observing. Default rate tells us which category from a particular segment has had the highest to least number of defaults.











RECOMMENDATIONS:

- People with 'Low' income range have higher chances of defaulting, therefore we should focus on other income ranges over this.
- People with 'Lower secondary' education and 'Single' status have the highest default rate. Therefore, we should be very careful while providing them loans. An authenticated guarantor's presence should be considered mandatory.
- Among both genders, even though females are higher applicants than male, it is still observed that females are lesser defaulters than males. Therefore, providing loans to females over males can be a plus point.
- People with 'Rented apartments' as their housing type are the highest defaulters. Therefore, we should check the security assets as well as the income of the applicant thoroughly.
- Age group (20-25) are the highest defaulters. Whereas, income stability is better in the age groups from 30 to 60 and they are less likely to default. Therefore, we should offer more loans to (30-60) age groups.
- People with housing types - 'office apartments' and 'with parents' are least likely to default as compared to other categories. Therefore, we should focus more on providing loans to these applicants.
- Banks should focus less on income type ‘Working’ as they have most number of unsuccessful payments.