CREDIT EDA CASE STUDY

BY CHETNA SAHU & SWETA MISHRA

Problem Statement

To analyze Bank Data on loans and find patterns in the data that are predictors of loan defaults. This will ensure that future loan decisions are made more logically and reduce possible defaults.

There are two types of risks associated with any loan request:

- If the applicant is likely to repay the loan, then not approving the loan results in 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 of the company.

Analysis of the data set has been done in Python on a Jupyter Notebook.

Types of Analysis Done

Step:

1. Check missing values; which to handle, how to handle.
2. Check outliers; Check data imbalance, ratio of imbalance.
3. Top 10 correlation for the client with payment difficulties other variables within Application DF & Previous App DF.
4. Which correlation is most relevant.

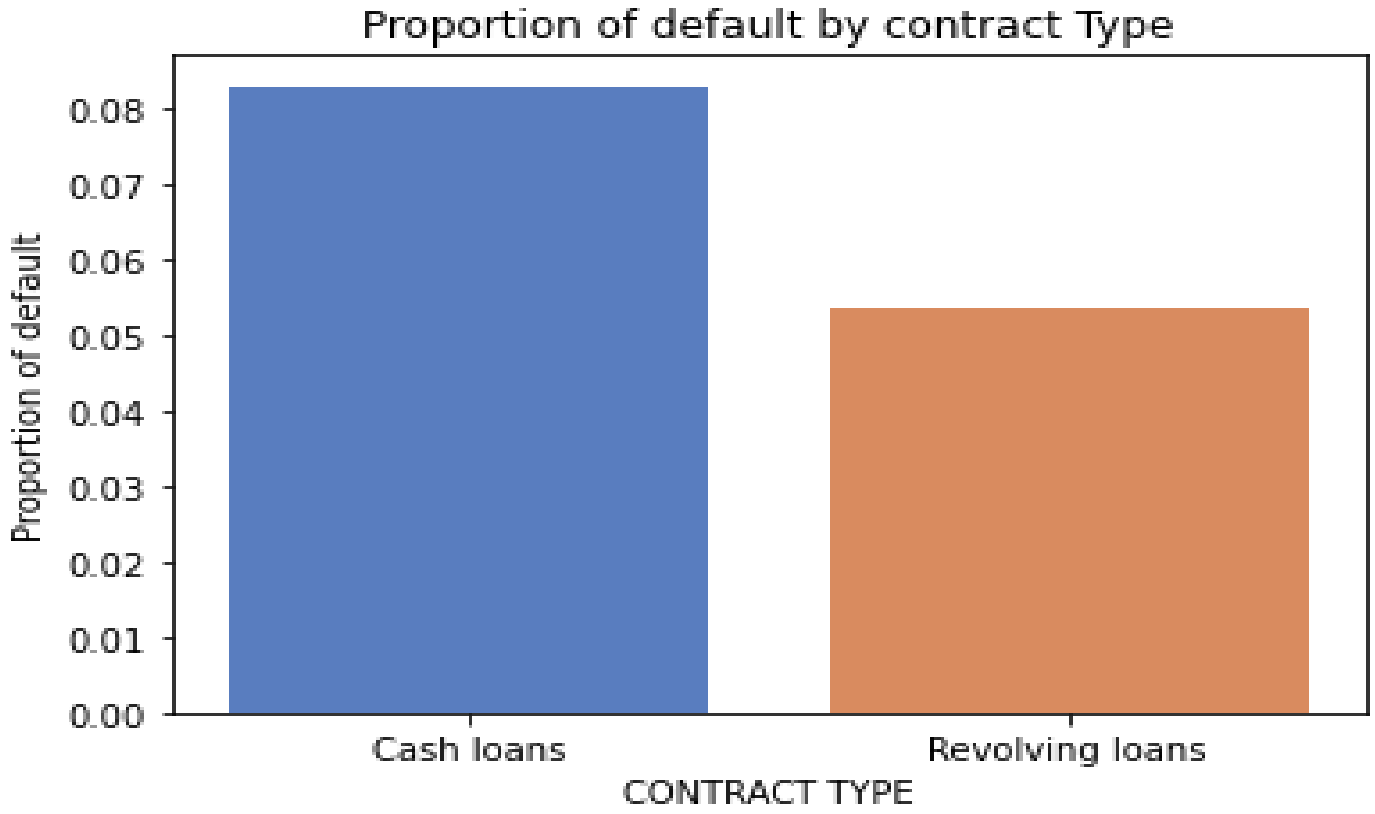
Divided into following tasks:

- Task 1: Reading the data and finding the information about the data.
- Task 2: Inspecting Data for Data Cleaning: Null Values, Which columns to drop and which to impute.
- Task 3: Imputing Values(Categorical- Mode; Numeric- depending on type of distribution.
- Task 4: Checking datatypes of columns.
- Task 5: Checking for outliers & handling.
- Task 6: Some more data cleaning operations.
- Task 7: Classifying data into bins.
- Task 8: Checking the data after data cleaning operations.
- Task 9: Check data imbalance.
- Task 10: Finding co-relations in the data.
- Task 11: Bivariate Analysis of numerical variables.
- Task 12: Analysis of Previous application v/s defaults and finding co relations.

Univariate analysis on categorical variables in Application data

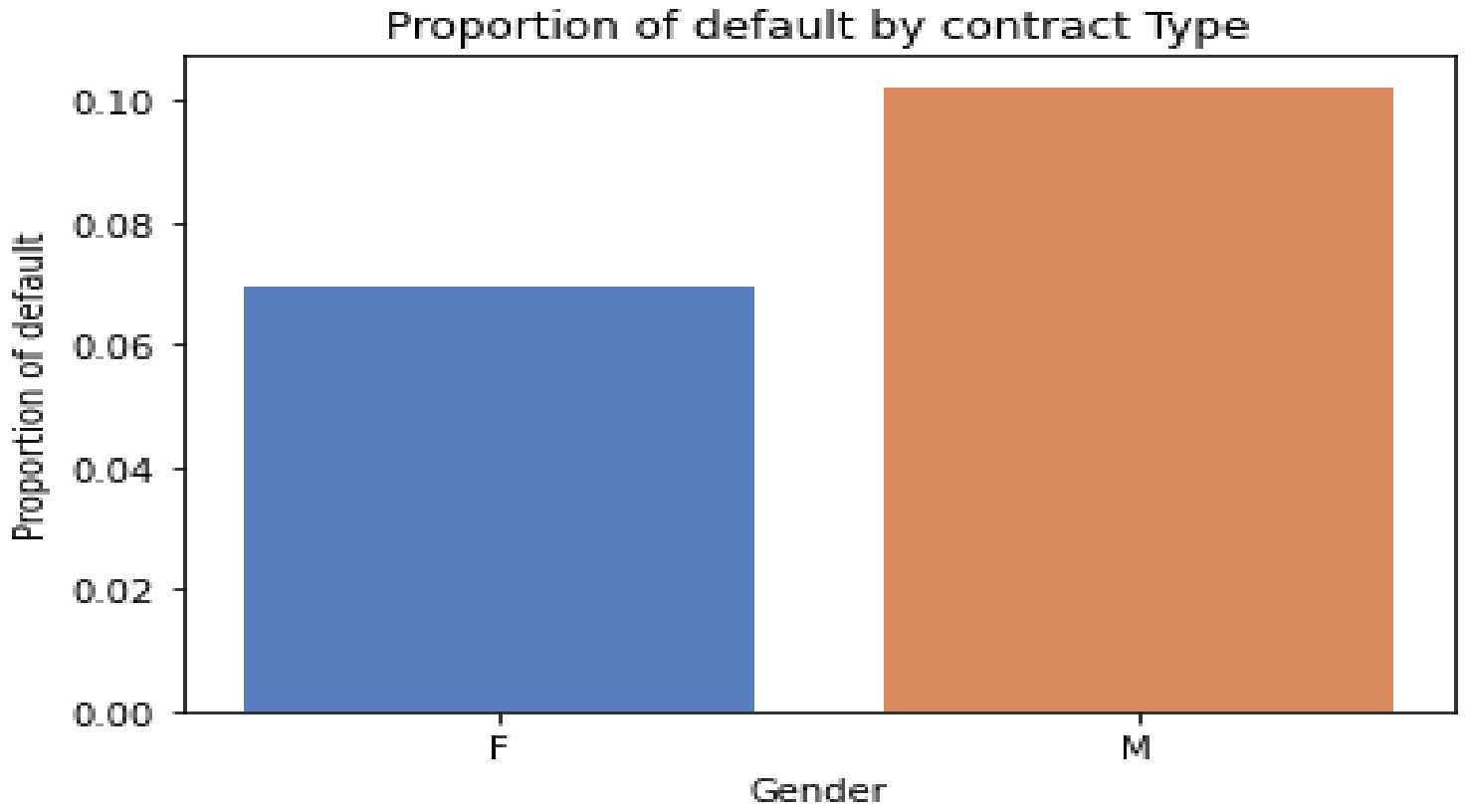
Contract Type

The below fig shows that the proportion of default is higher in cash loan category compared to the revolving loans.



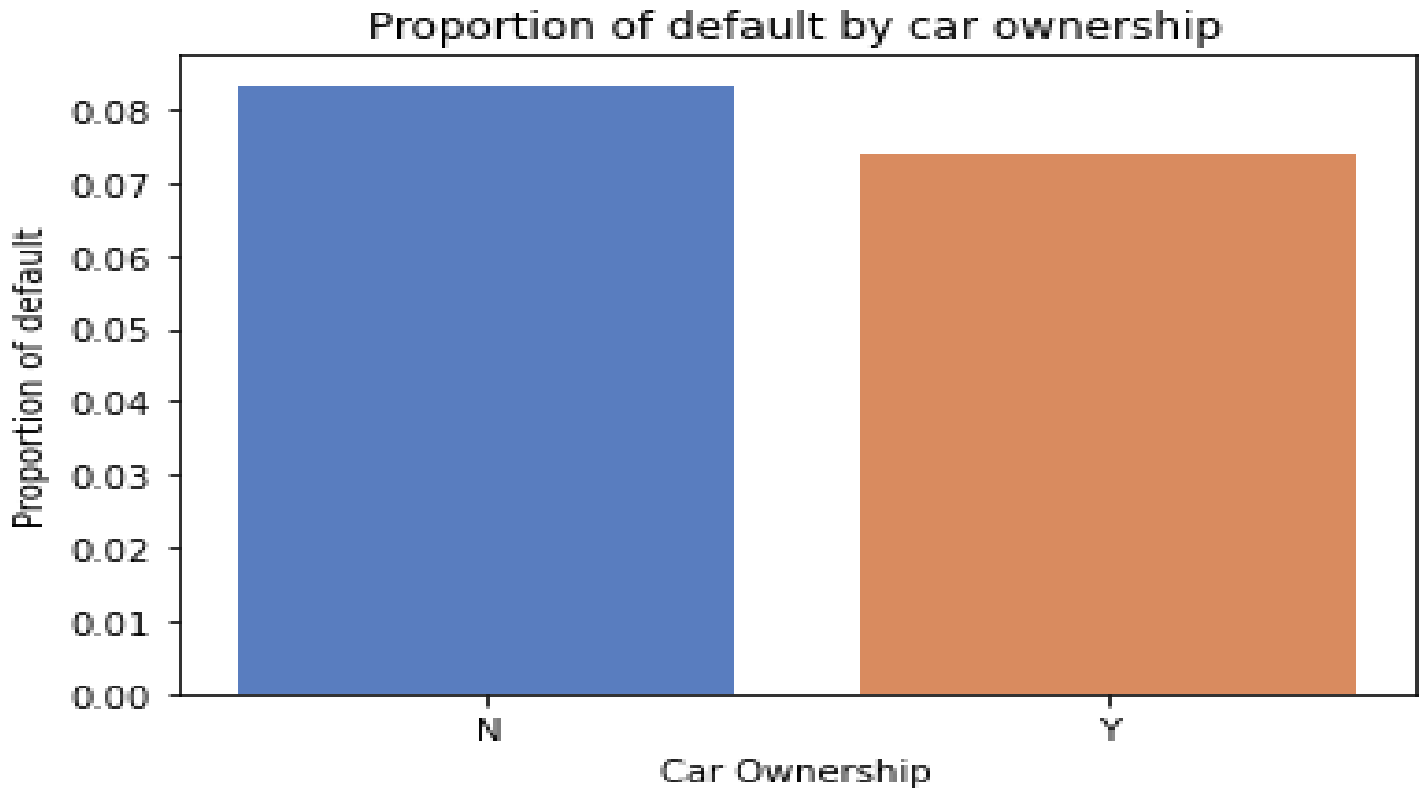
Gender

Number of Female clients are more in both the cases default/non-default. And the proportion of default is lower among female applicants than that of the male.



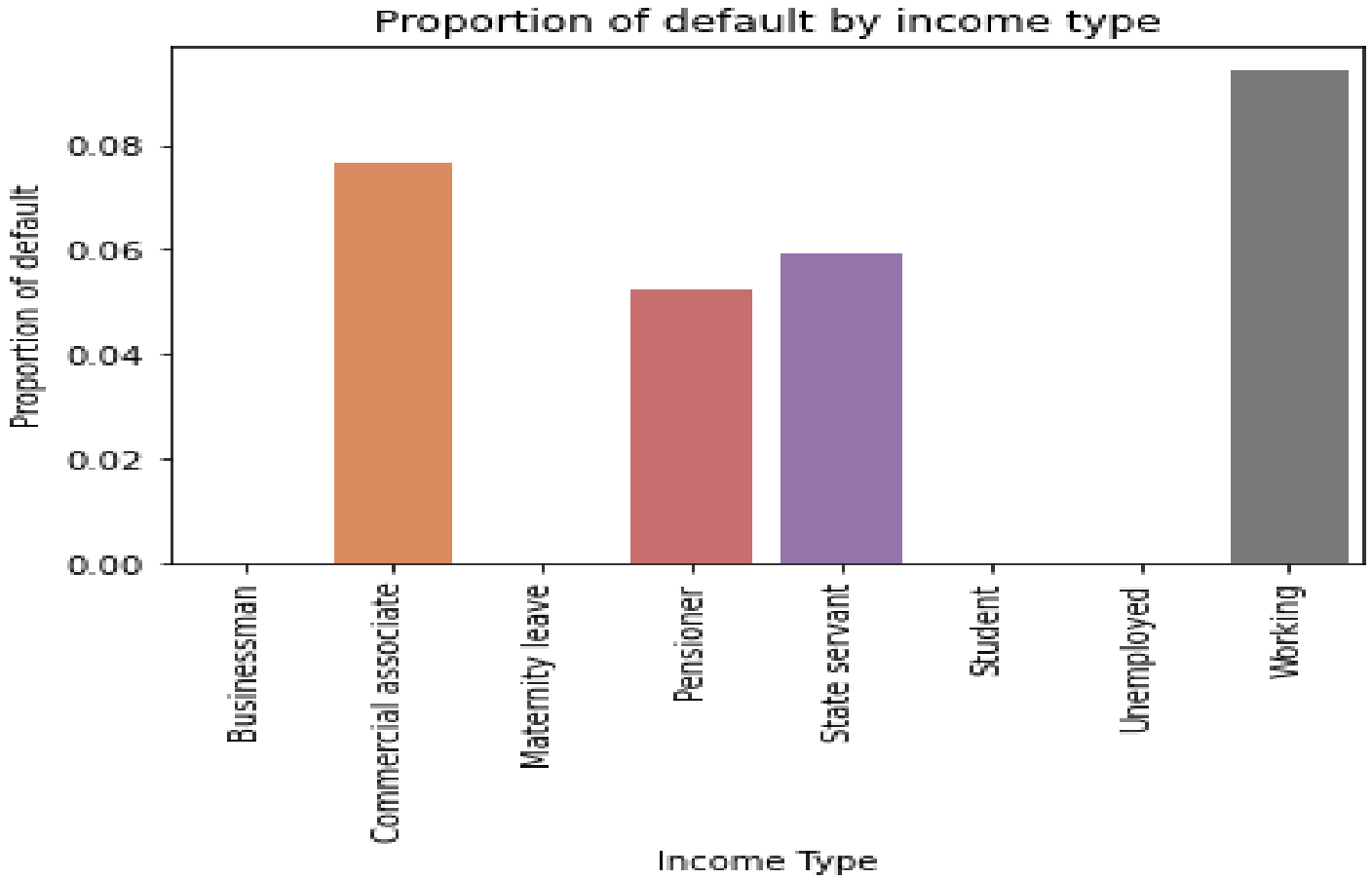
Car Ownership Status

We can see that in both Default/Non-default cases, the count of client who do not own car is higher than who owns a car. The proportion of default is higher for non-car owners relative to the car owners.



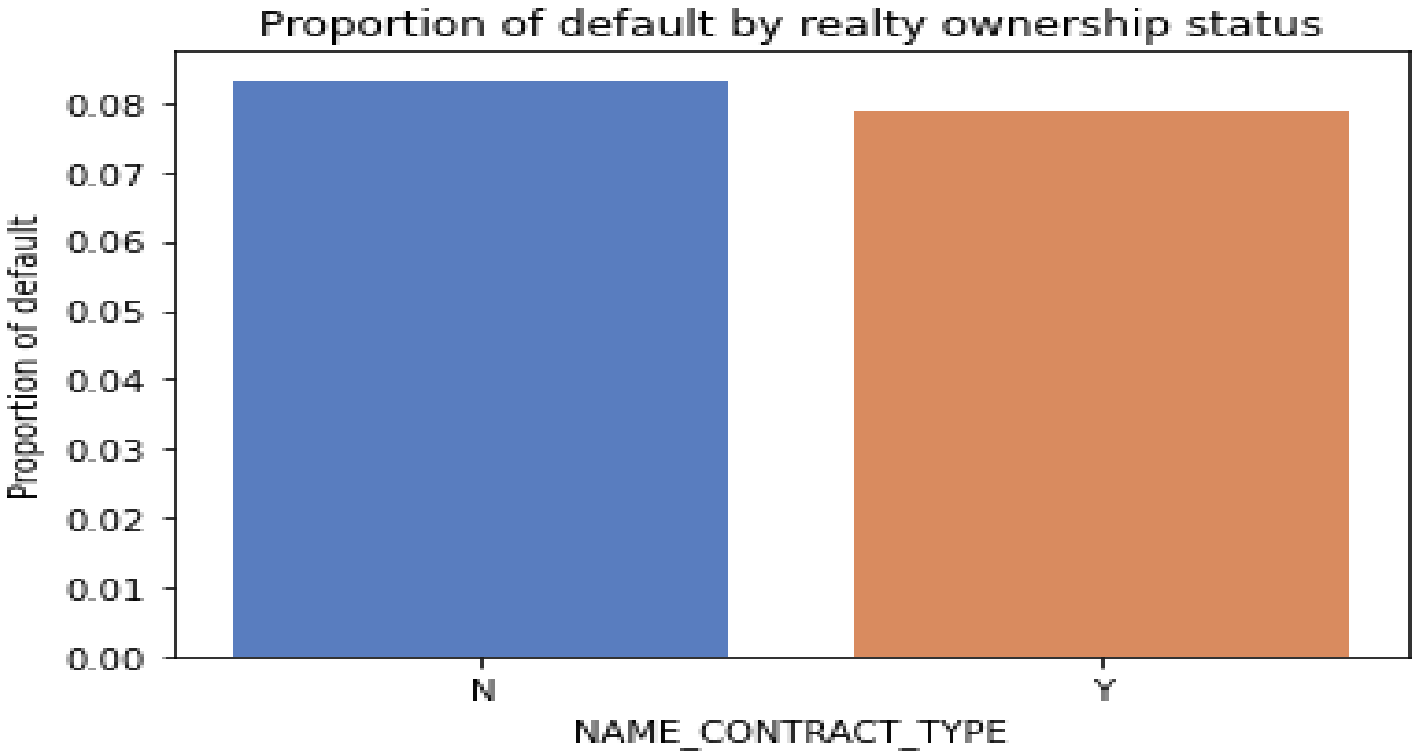
Income Type

Majority of the applicants are from working, commercial associate, pensioners and state servants. The remaining categories of income types are very small. The proportion of default is high among the working and the commercial associates. It is relatively lower for the pensioner and state servant.



Realty Ownership Status

Default/Non-default both the cases, the count of client who do own real estate is higher than who has not. Applicants with no realty ownership has a higher propensity to default than the clients who own real estate



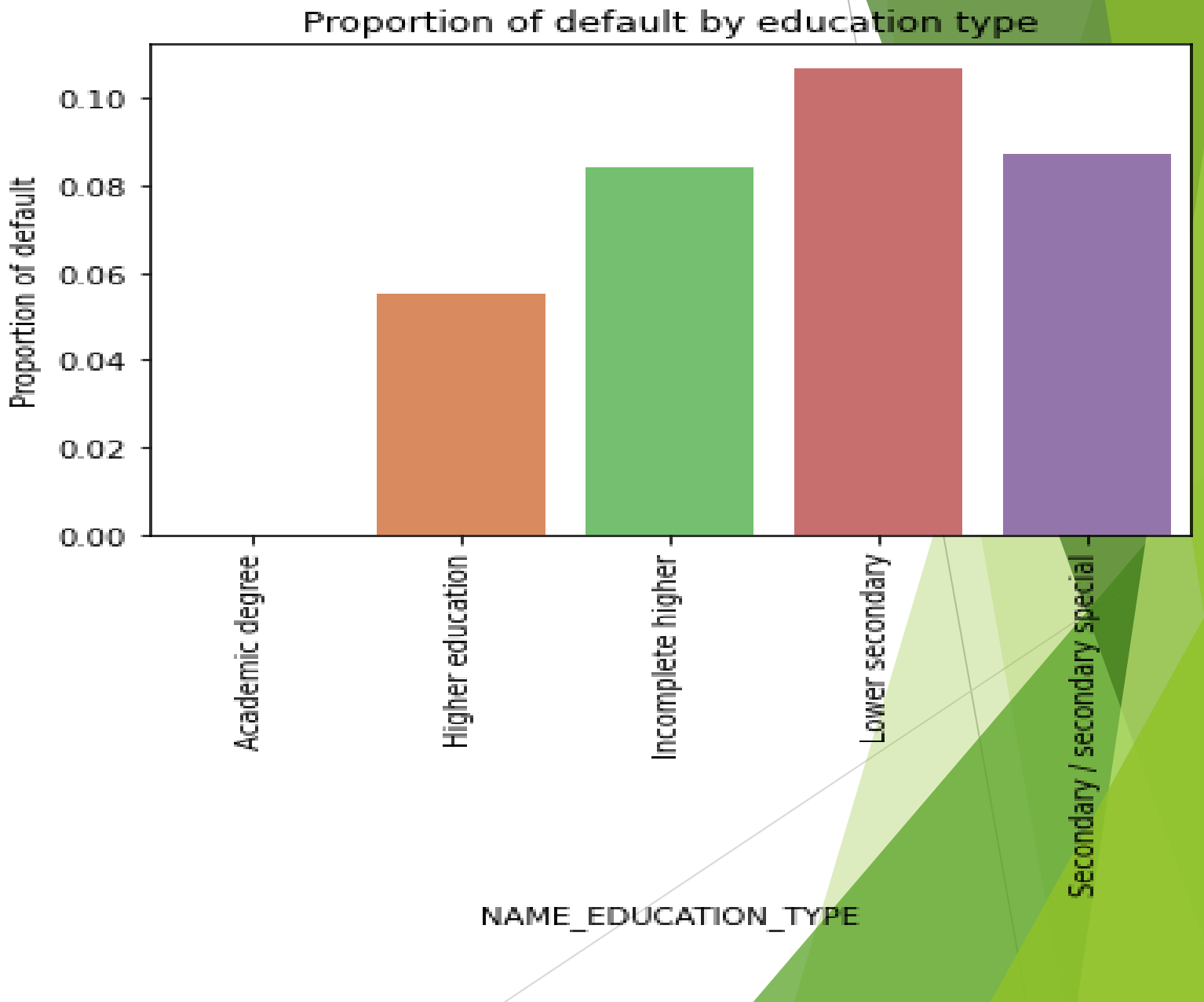
Education Type

1.Applicants with secondary and higher secondary education are among the highest defaulters as well as not defaulters.

2.Whereas, applicants with academic degrees are the smallest group of applicants that have applied for the loan and applicants from this background has no recorded of default.

3.From the above figure, we see that a distinct pattern emerges.

4.The chances of default is lower as the education level of the applicants increases.

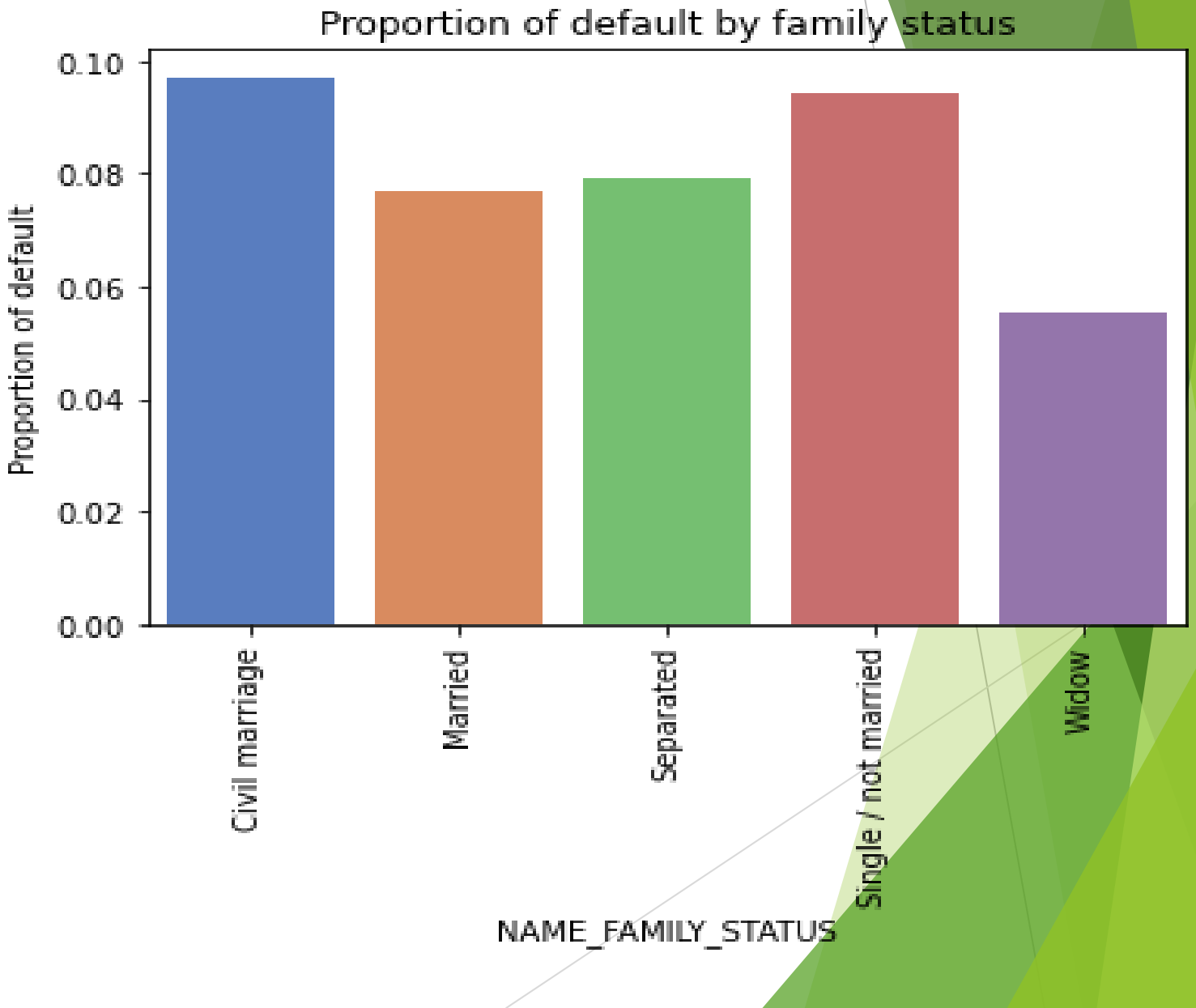


Family Status

1.Applicants who are married are among the highest number of defaulters and non-defaulters.

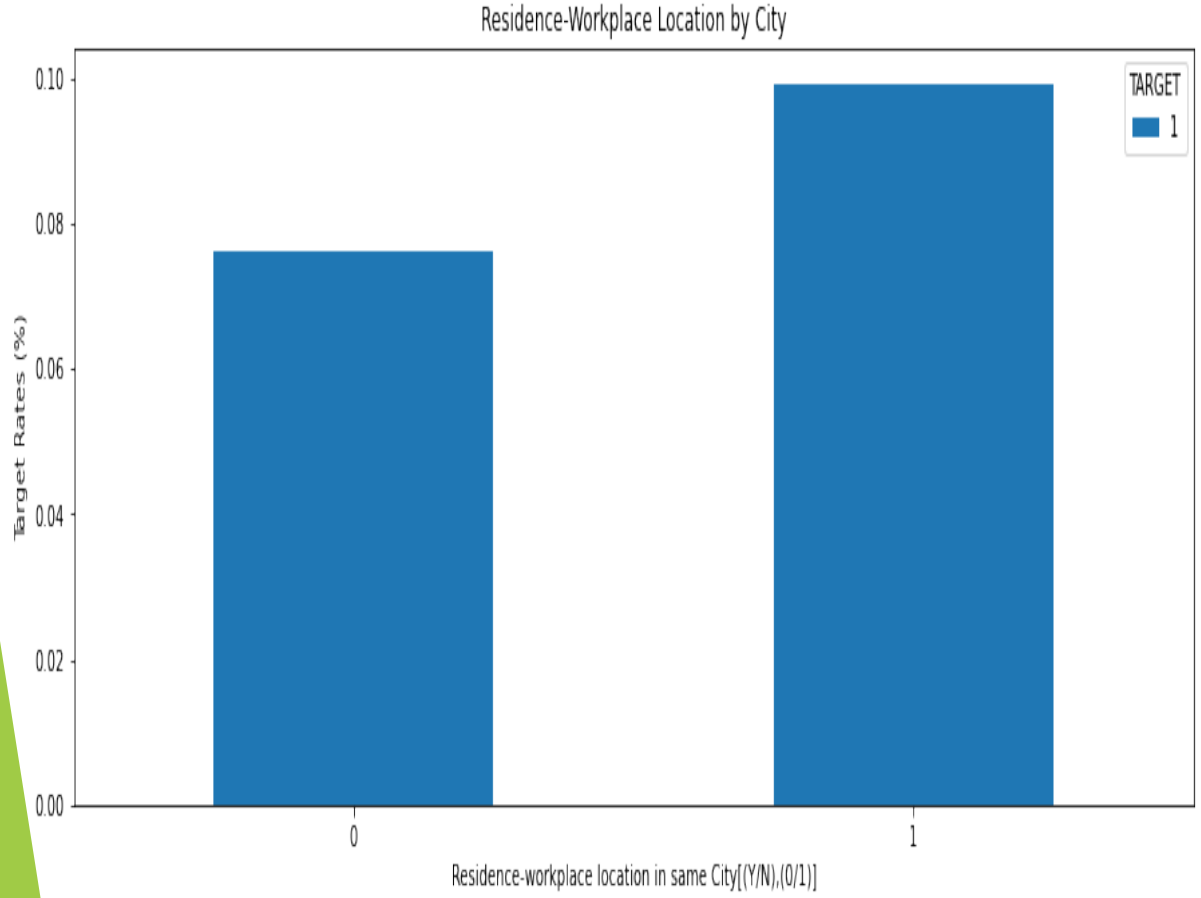
2.Whereas, widows are the lowest number of defaulters and non-defaulters.

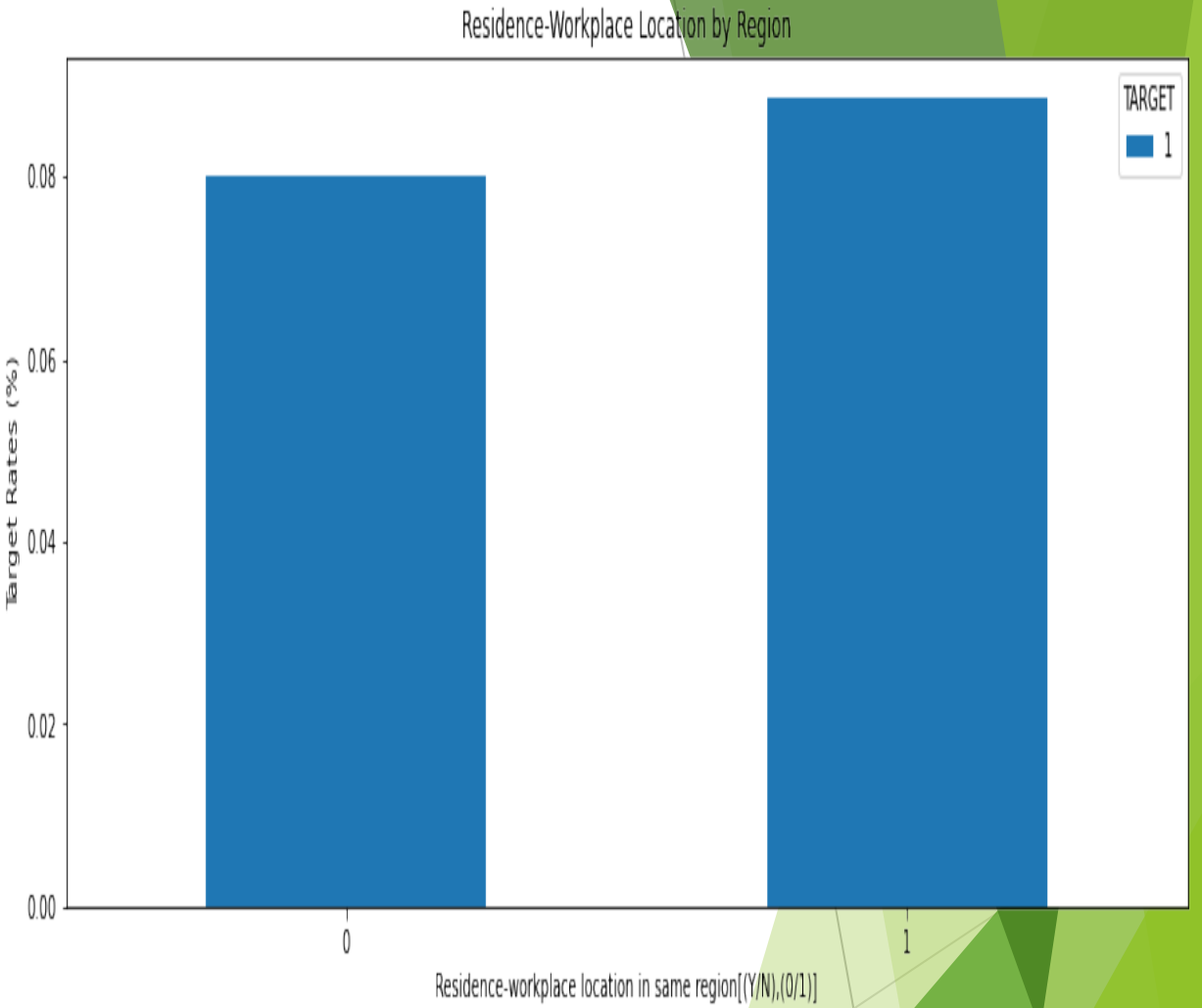
3.The proportion of default is the highest among the applicants who are in civil marriage category followed by applicants who are single.



Spatial Effect

Most of the applicants live and work in the same city/region. And applicants who doesn't live and work in the same city/region has the higher chances of default.





Univariate analysis on continuous columns in Application Data

DAYS_BIRTH

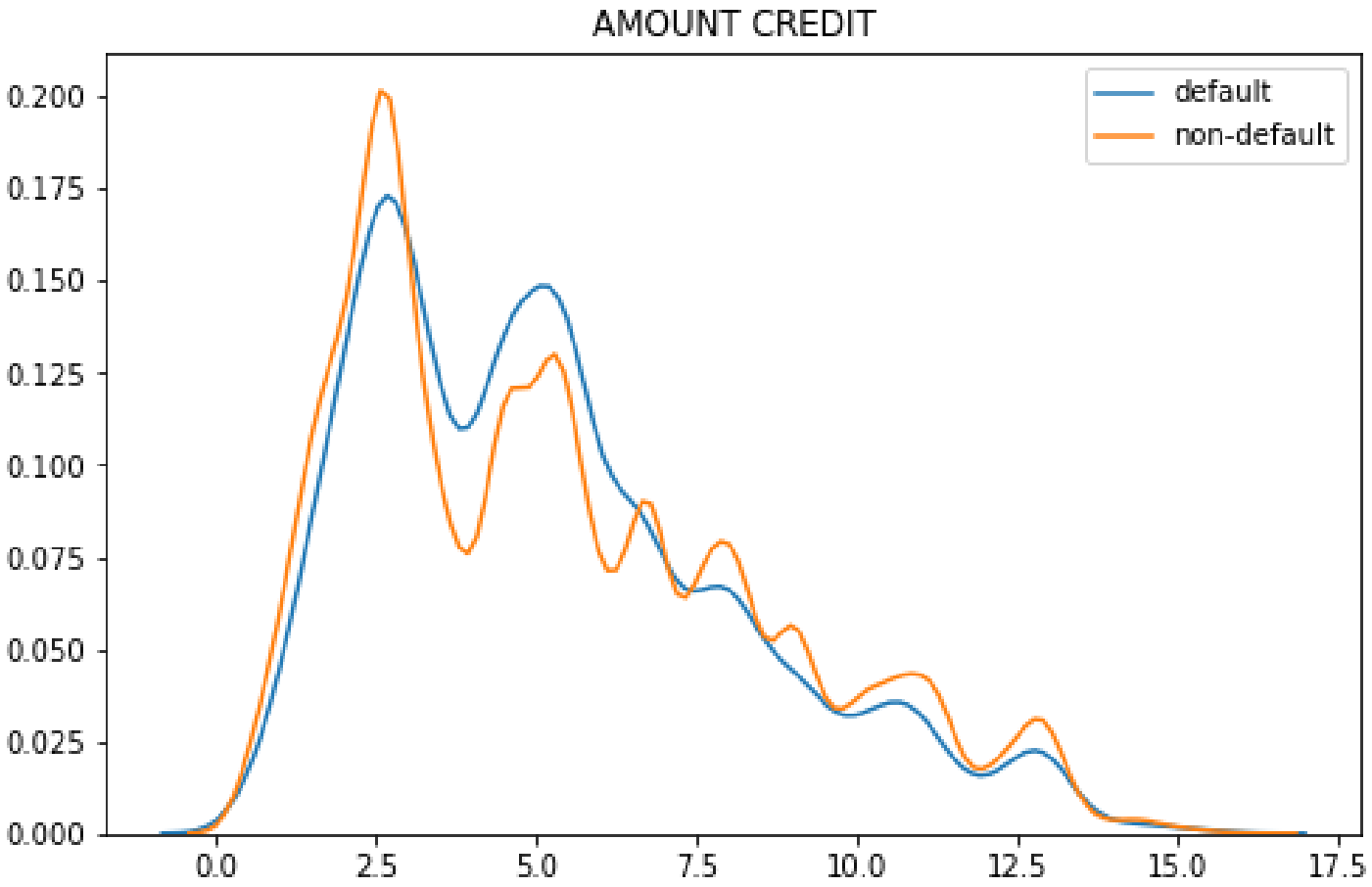
Around 29 years to 40 years people are more defaulters. There is high chance to be defaulted of the young people. Non-defaulted people are almost equally distributed.

AMT_INCOME_TOTAL

There are interesting patterns in both the default and non-default people with respect to Income Total. From 75000 to 200000 income has some spikes and then higher the income, the lesser spike we can see for default people. However, the pattern is same for both the default and non-default. The frequency range of non-default people is larger than default.

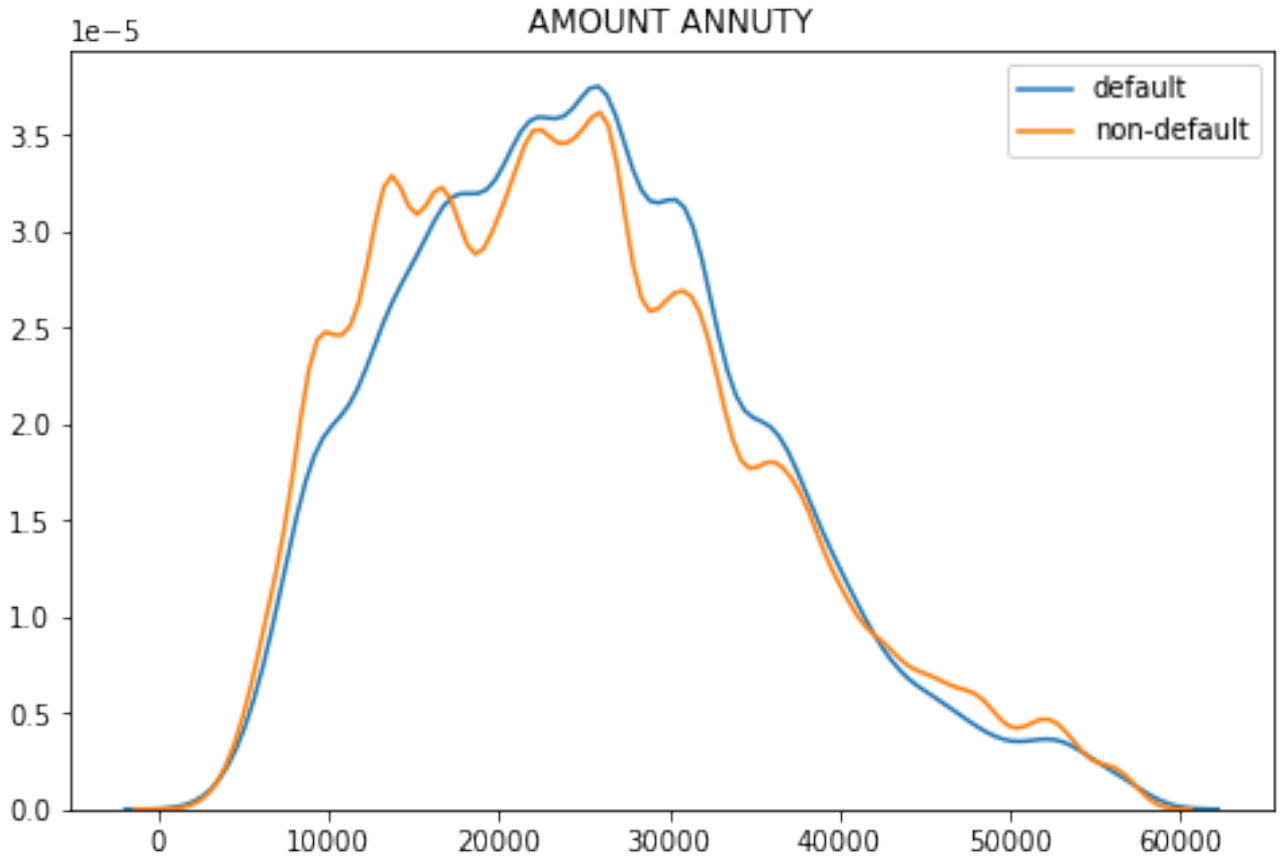
AMT_CREDIT

Here, we can see that the lesser loan credit amount, the higher the default chances. We can do bivariate analysis with Occupation Type to find out more insight.



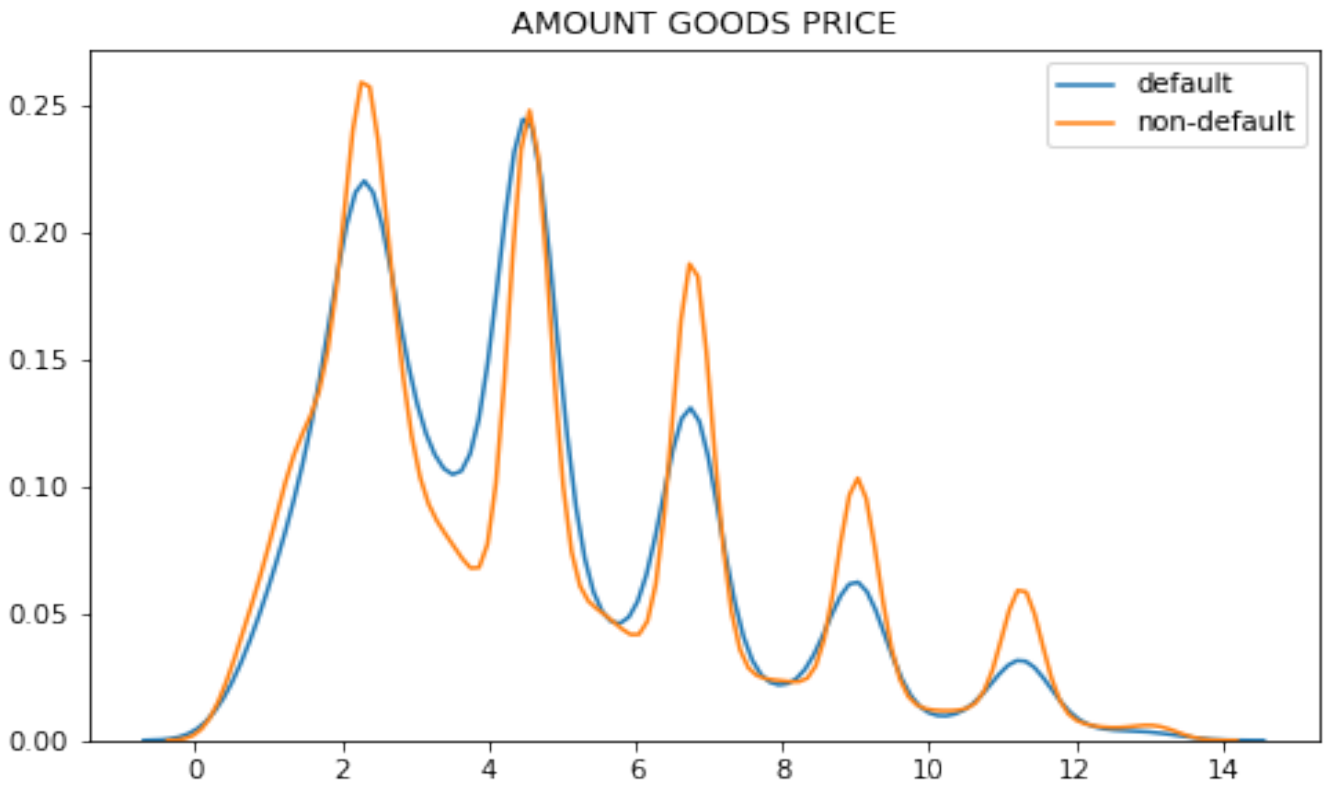
AMT_ANNUITY

Here also we can see the same pattern in both the default and non-default. The loan annuty is mostly concentrated within 10000 to 40000 range in both the cases.



AMT_GOODS_PRICE

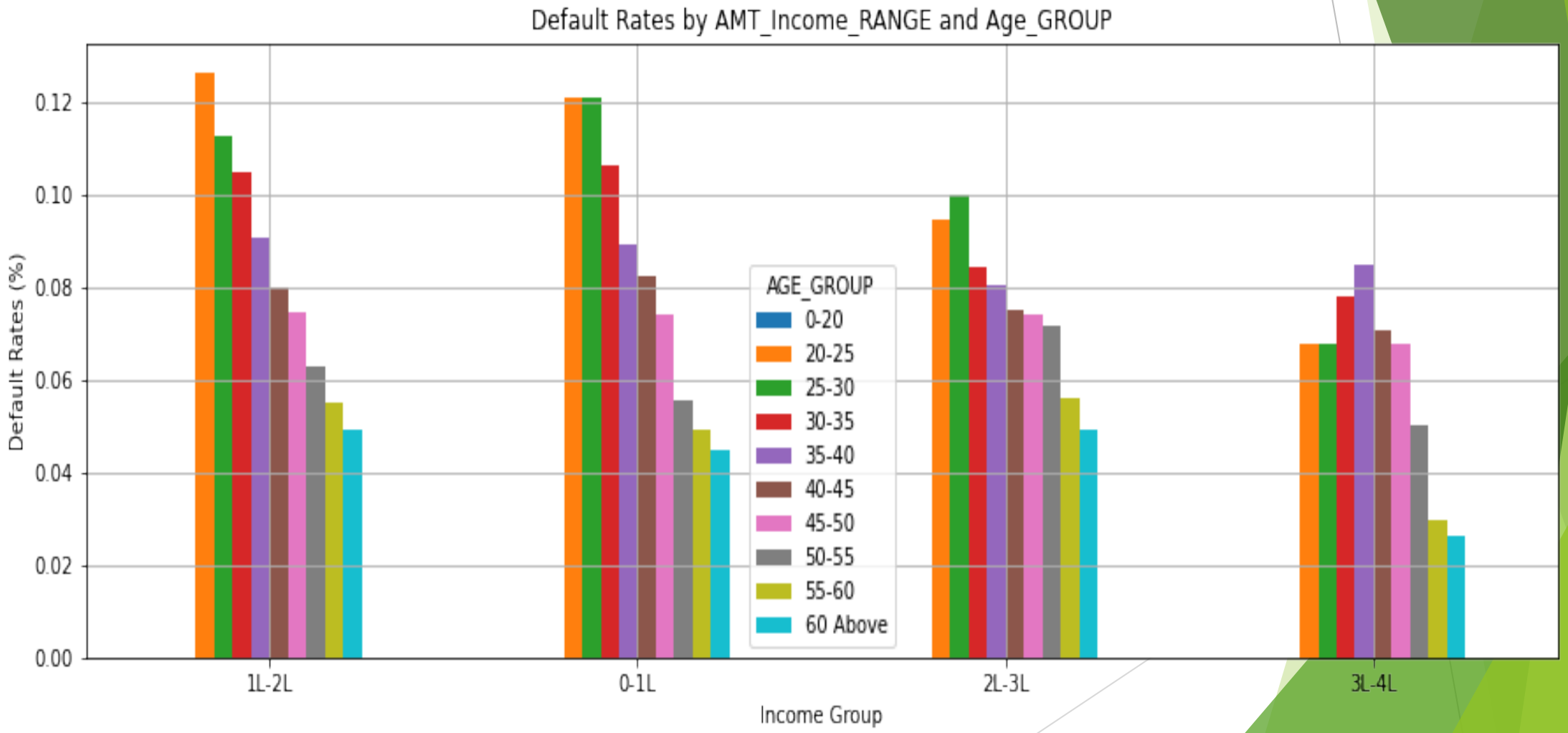
Here also we can see an interesting pattern. Both the curves are following the similar frequency distribution. We can see some spikes from 150000 to 220000, then around 500000 price. At this range people are more defaulted and higher the goods price, people are becoming the less defaulted. We can infer that, rich people are buying costly product and thus they are becoming less defaulted.



Segmented univariate analysis of Application Data

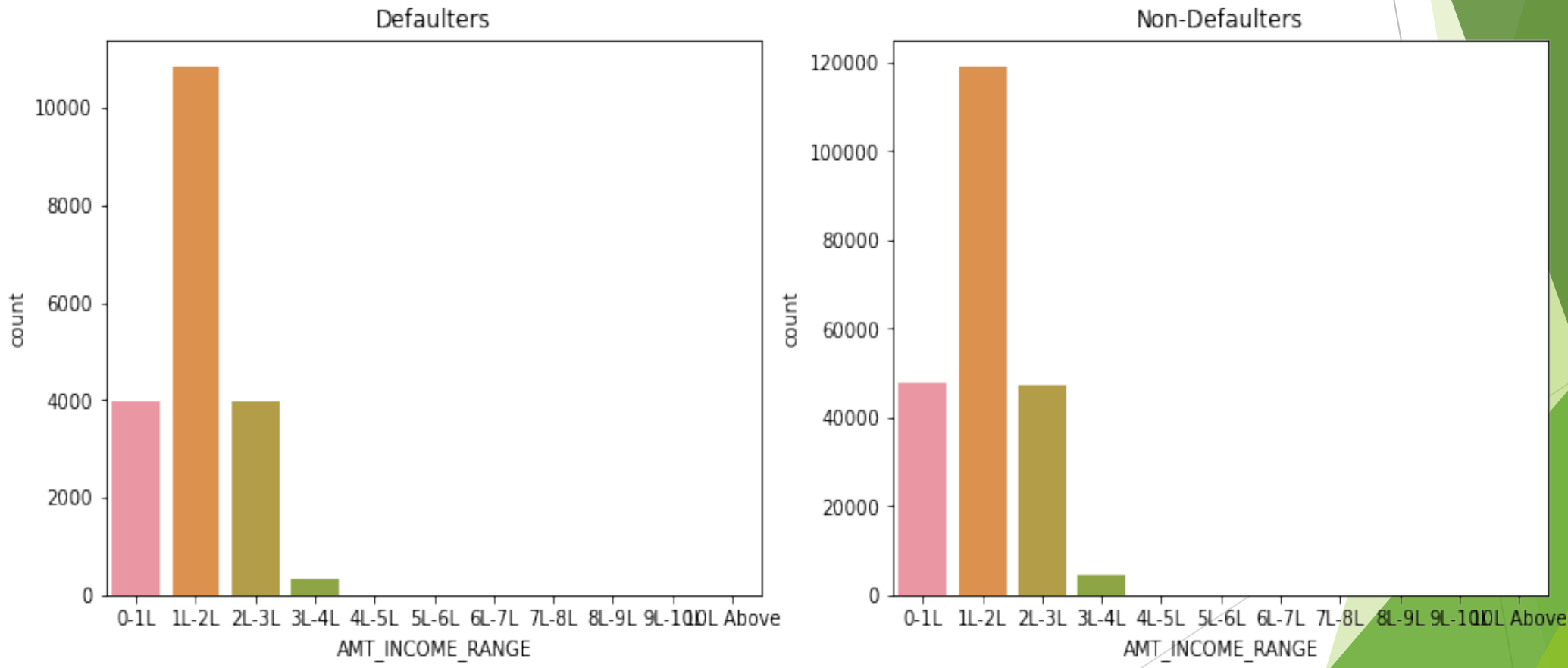
Age and Income Category segmented two variables

We see from the below diagram that irrespective of the income groups, the chances of default decreases as the age of the applicants increases.



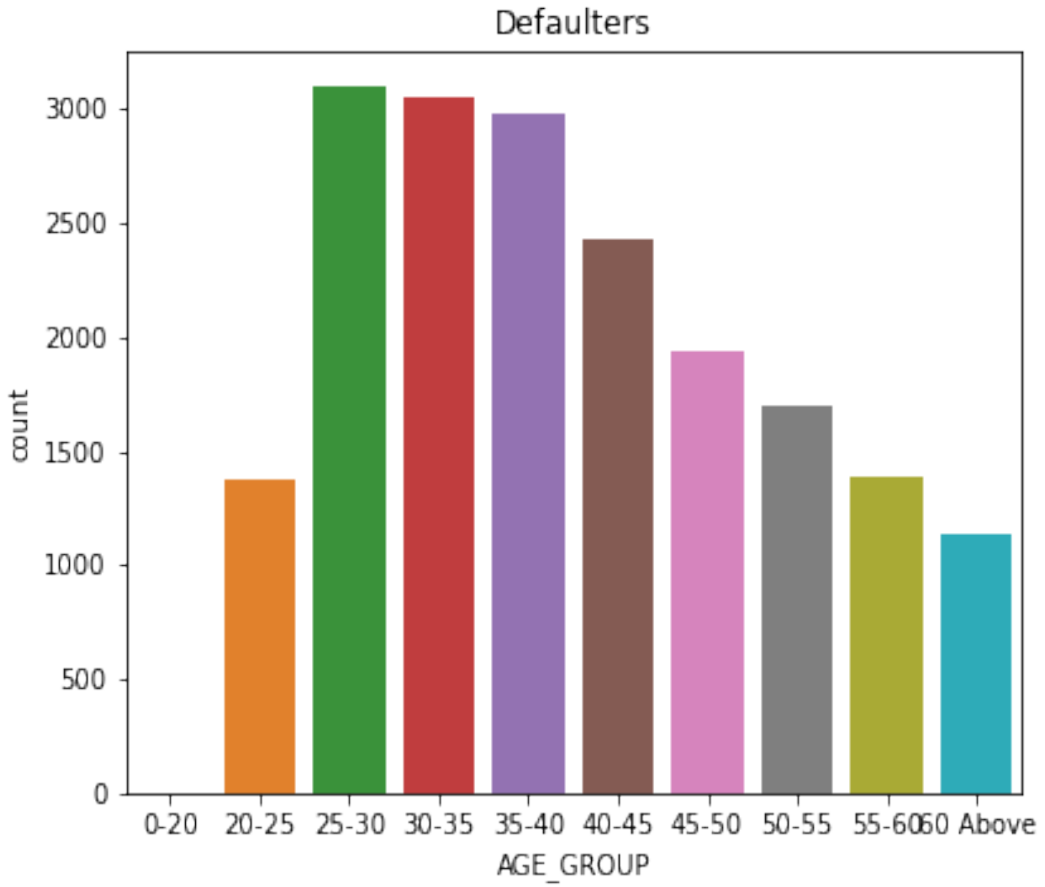
Income category segments for both the default and non-default

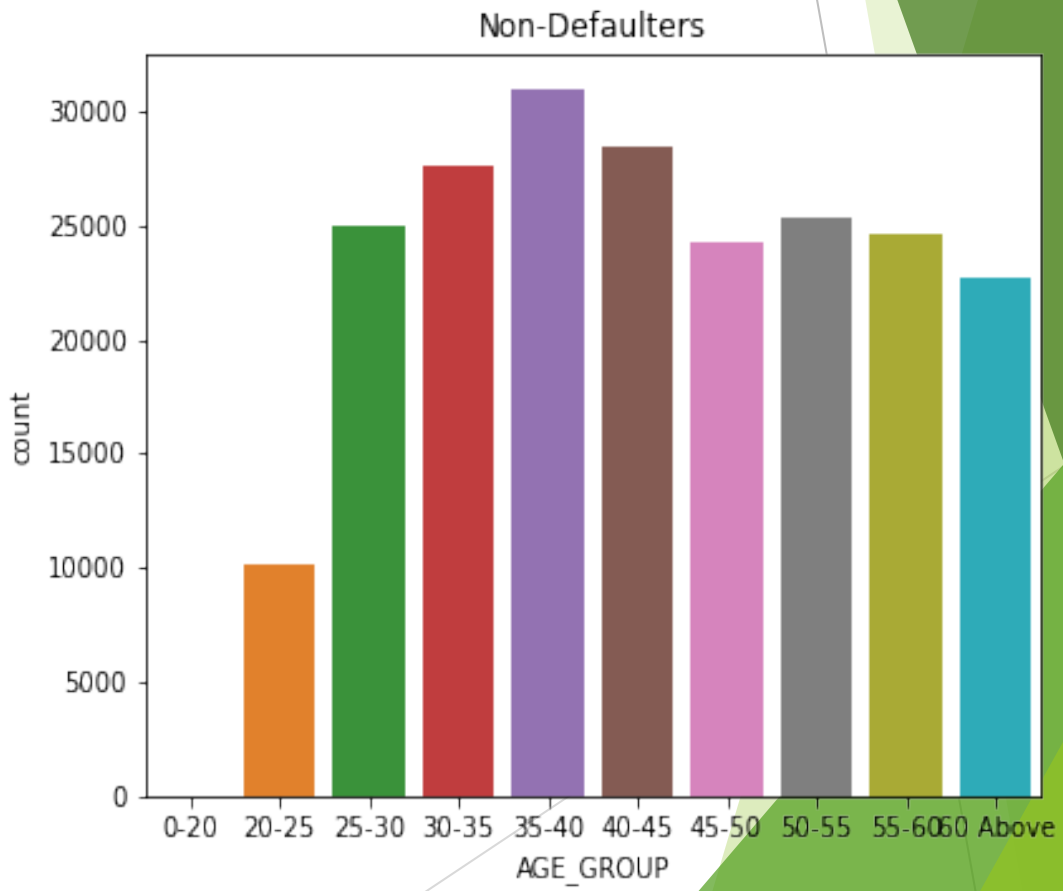
Low income group has more defaulter followed by high income group.



Age category segments for both the default and non-default

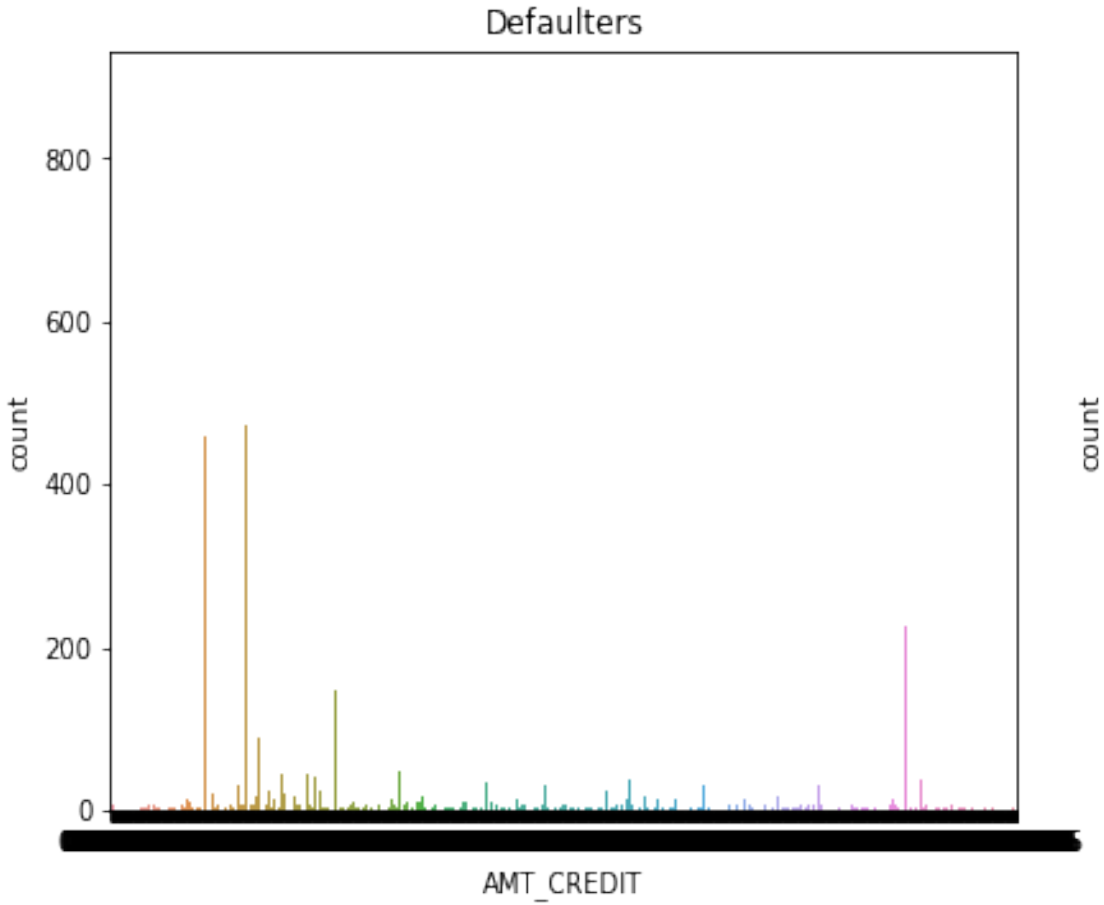
Mid age (35-55) age group of people are more likely to be defaulted followed by the young people.

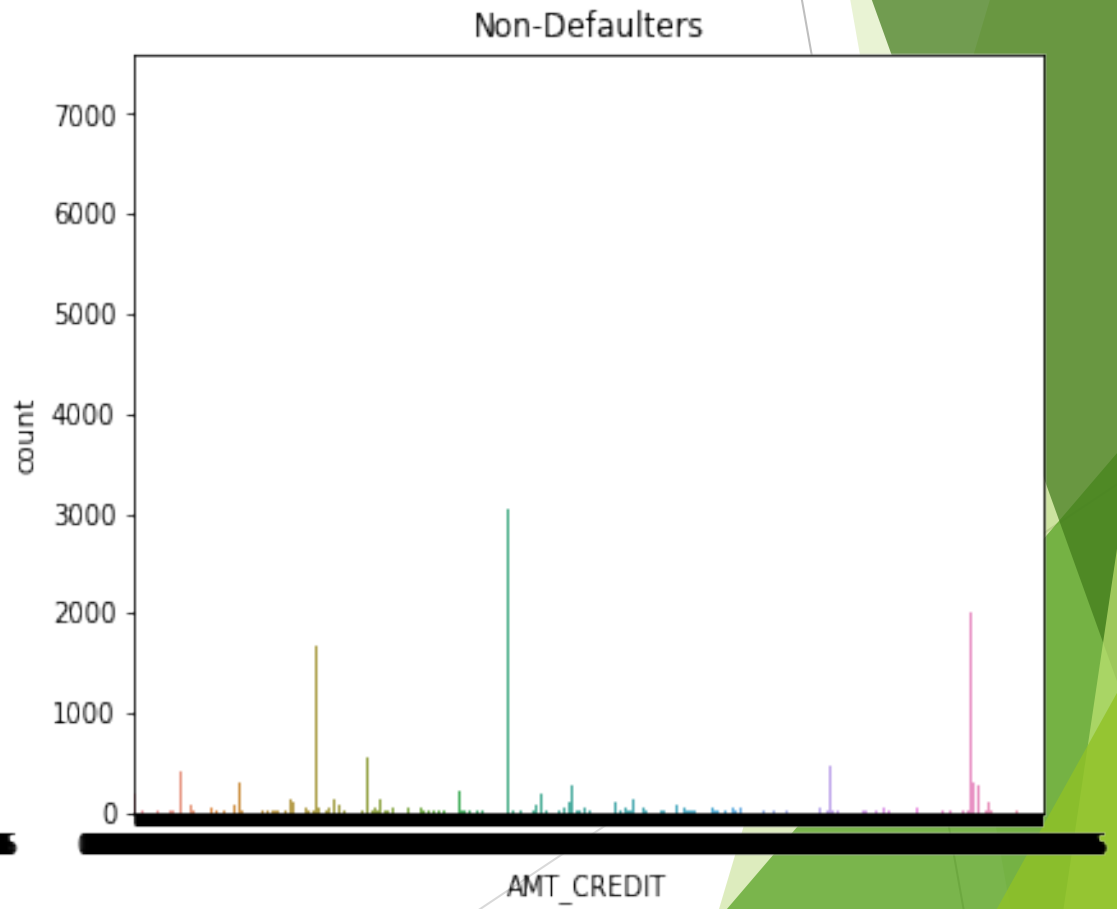




Amt credit segments for both the default and non-default

Low category of loan amount credited people are more likely to be defaulted than high amount loan credit.

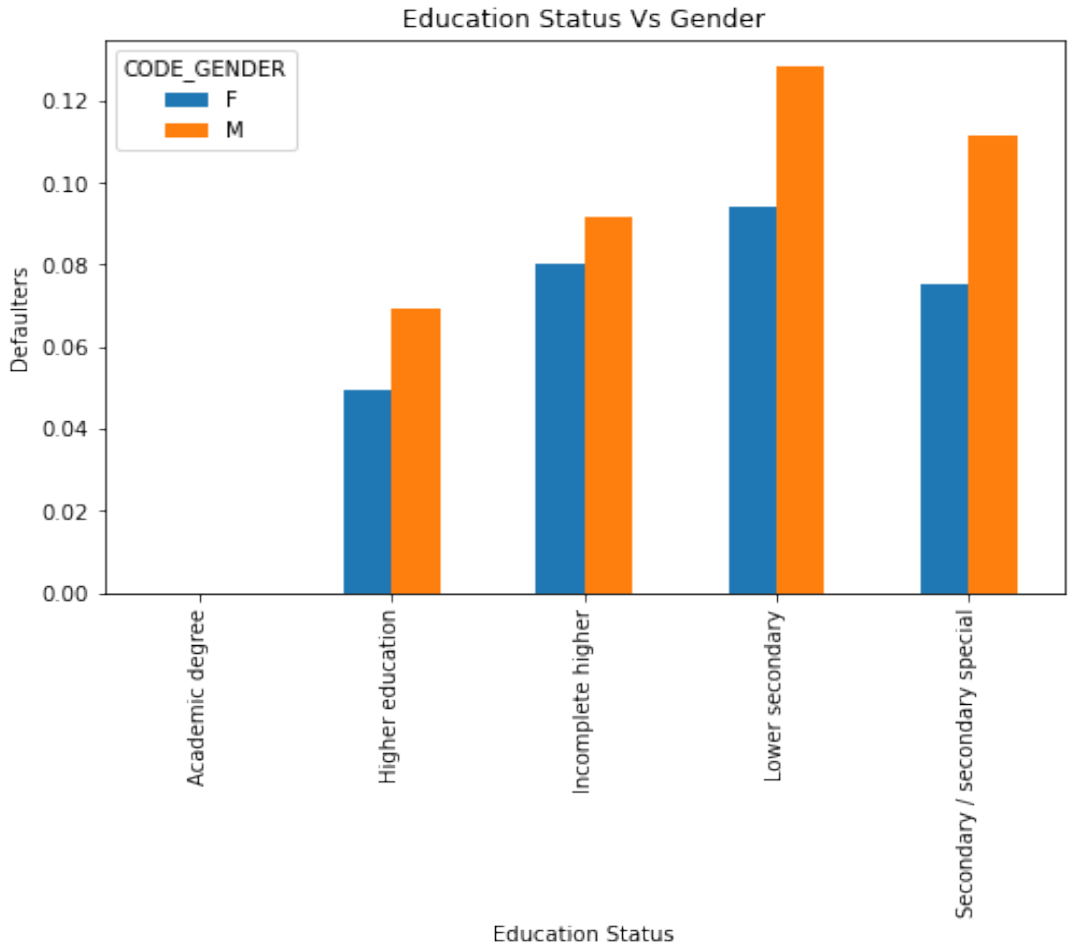


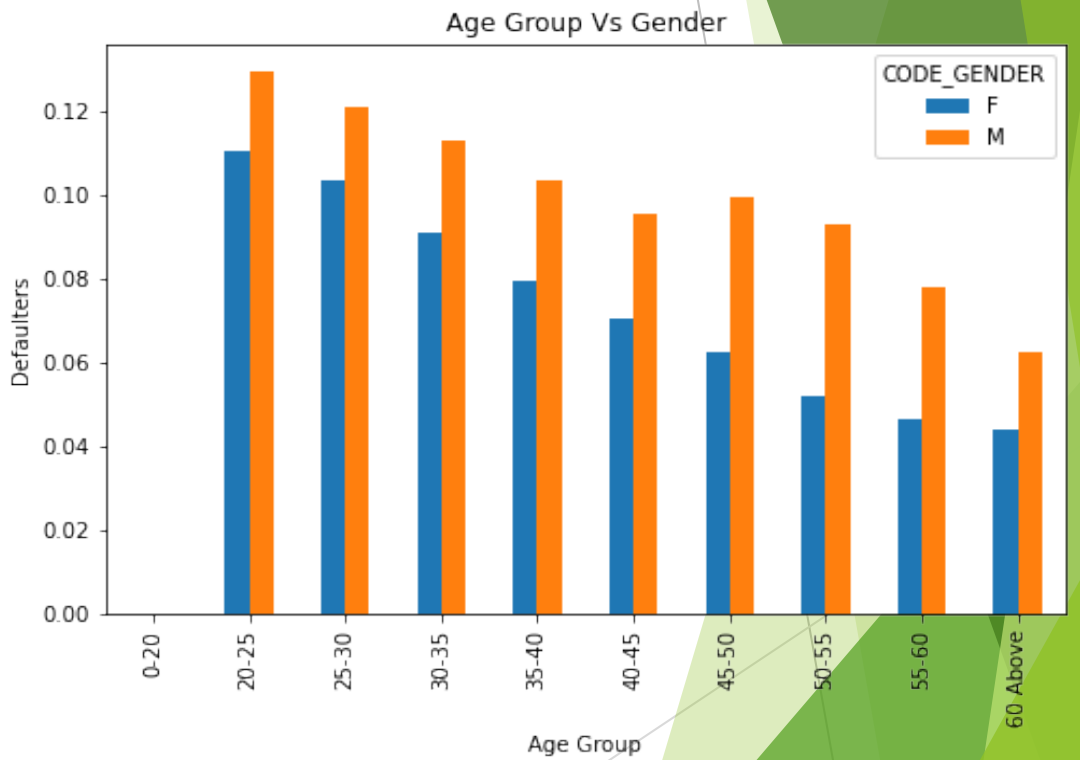


Education Status Vs Gender wise defaulters & Age Group Vs Gender

1.Male with lower secondary education are more defaulted followed by Secondary/secondary special education.

2.Young male clients are more in number to be defaulted.



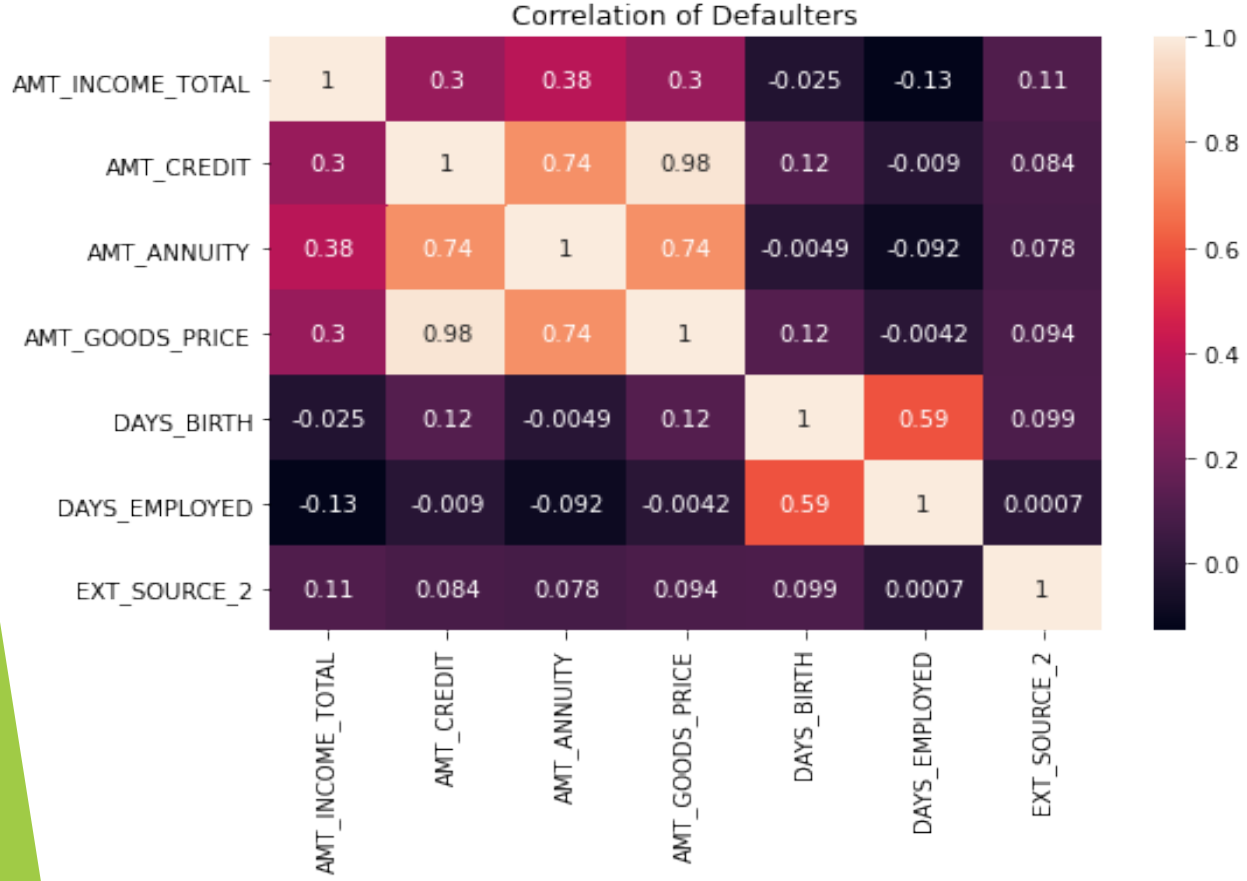


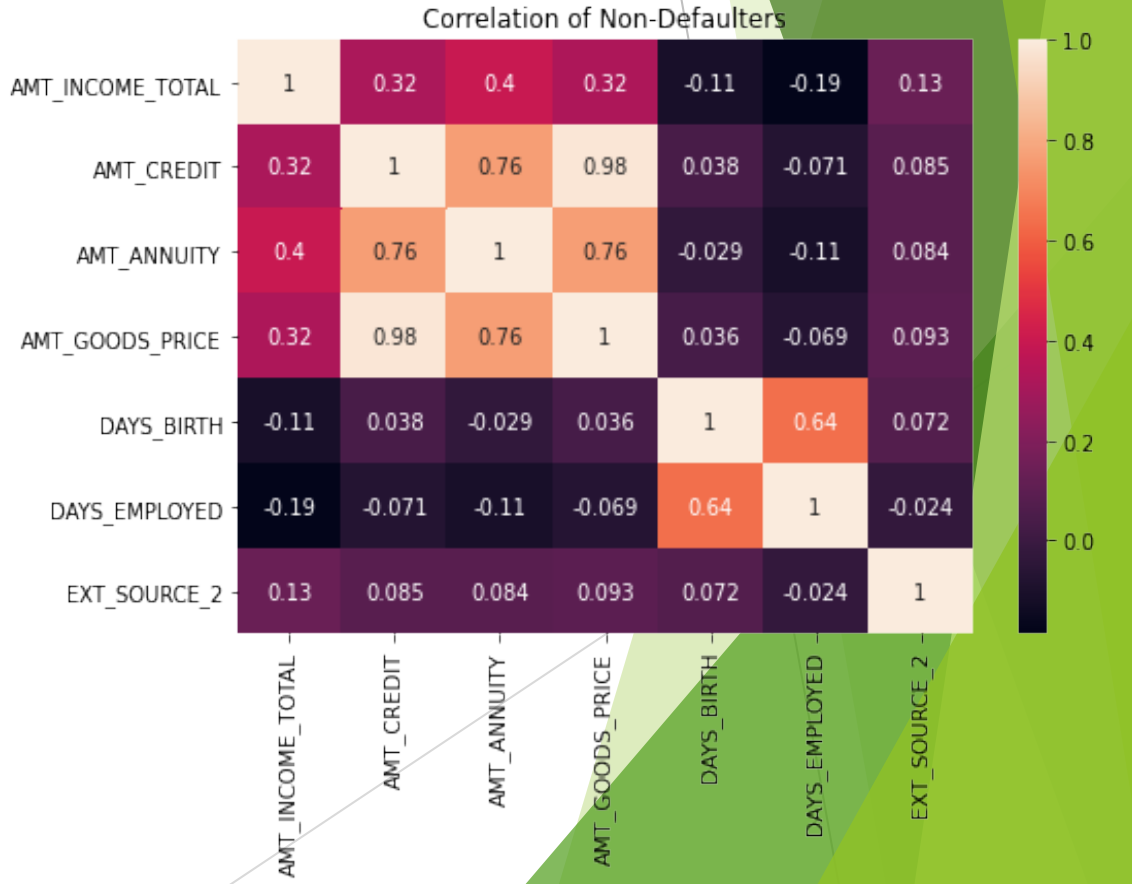
Bivariate analysis of Categorical columns

Correlation of the continuous variables of Defaulter data frame & Correlation of the continuous variables of Non-Defaulter data frame

We can see that GOODS_PRICE and AMT_CREDIT, AMT_ANNUTY and AMT_AMT_CREDIT are highly correlated. External Rating is highly correlated with all DAYS_BIRTH(Age), GOODS_PRICE, AMT_CREDIT.

We can see that GOODS_PRICE and AMT_CREDIT, AMT_ANNUTY and AMT_AMT_CREDIT are moderately correlated with each other. External Rating is highly correlated with all DAYS_BIRTH(Age), GOODS_PRICE, AMT_CREDIT.

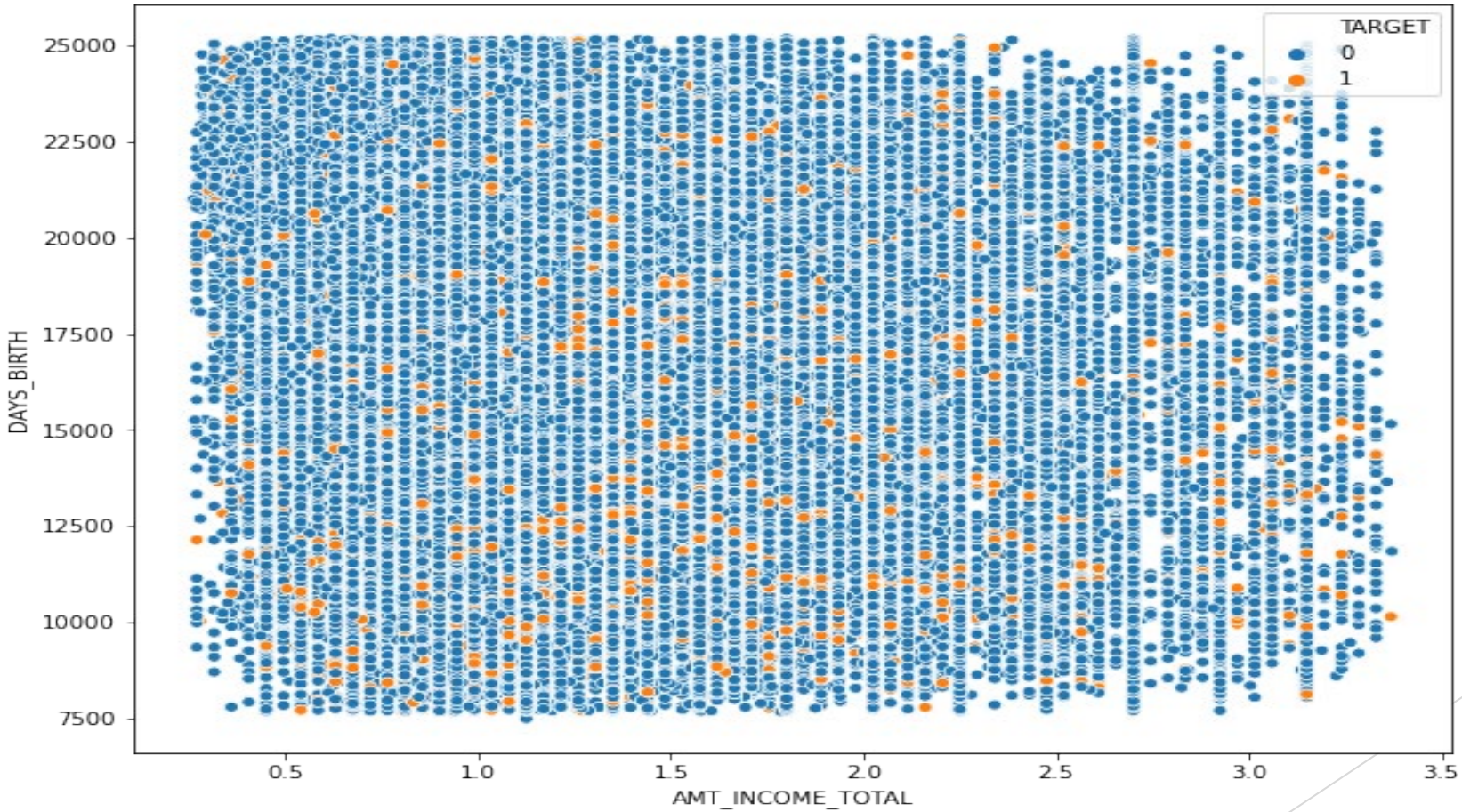




Bivariate analysis on continuous columns

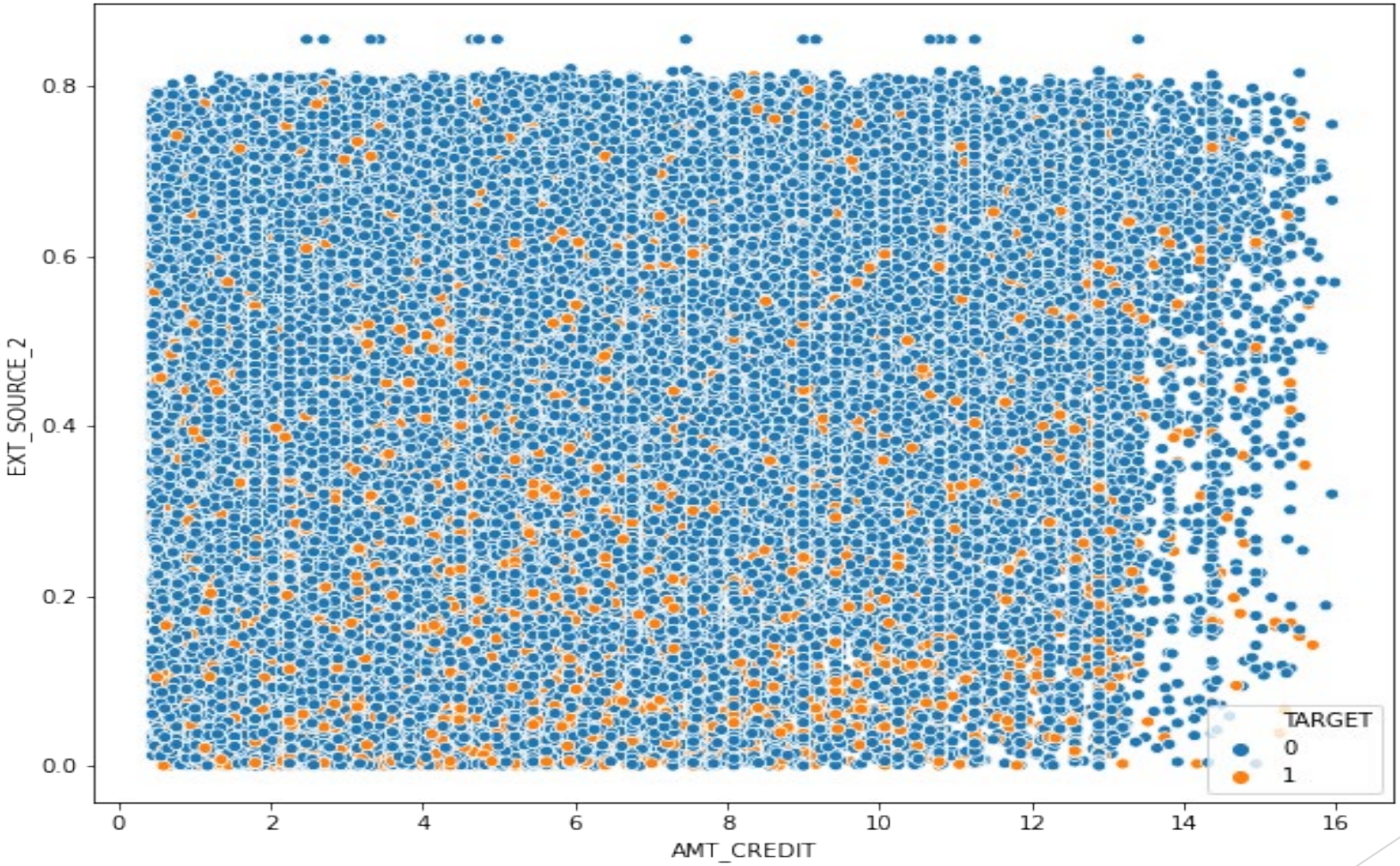
Age and Income

From the above figure, we see that the number of default applications are concentrated more when the days of birth i.e. age is lower, irrespective of the income.



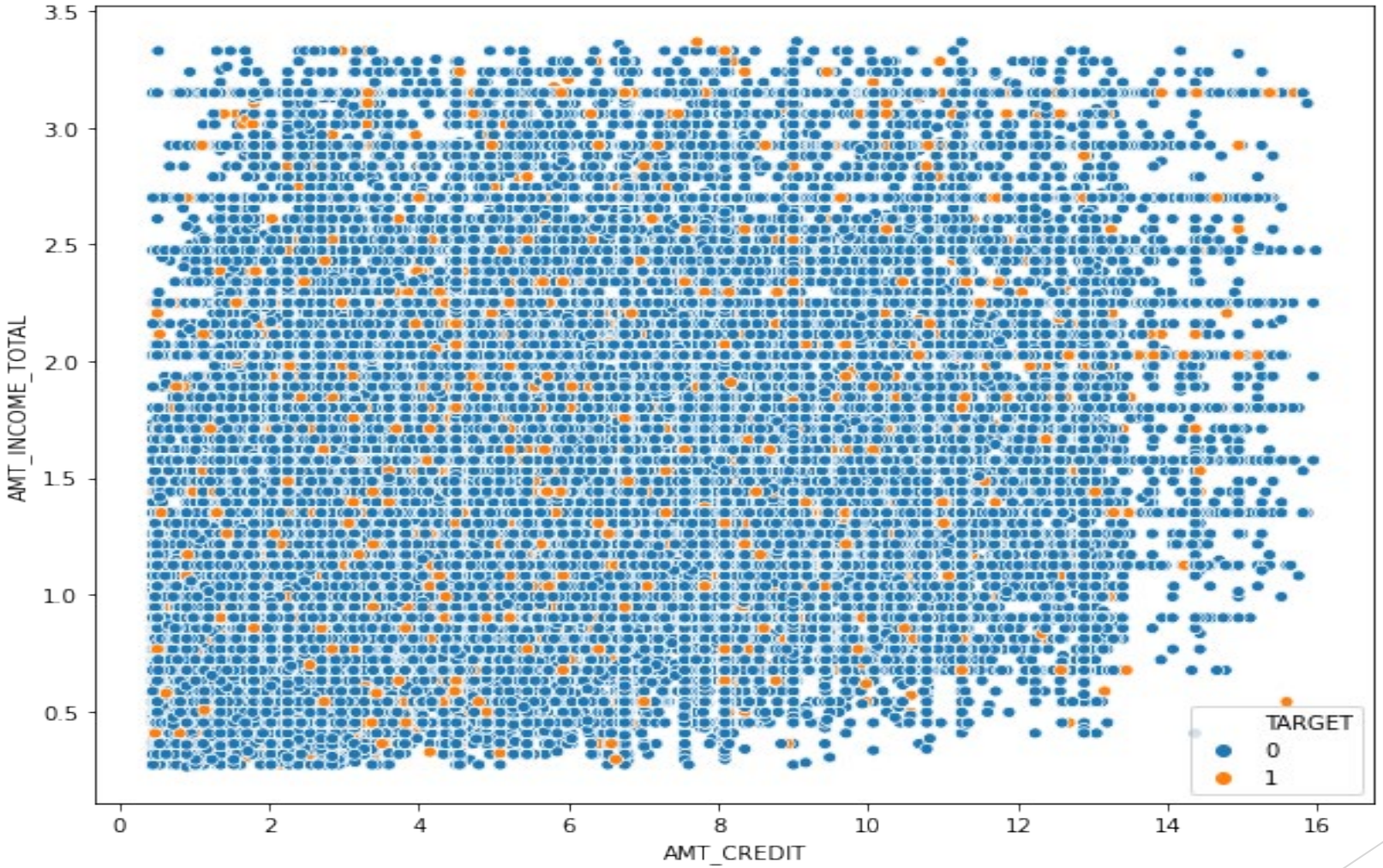
Loan Credit amount and Rating

From the above plot, we cannot get much insight as the data is scattered across the plot. However, we can see some concentration of defaulters near the low rating region between 0.0 to 3.0.



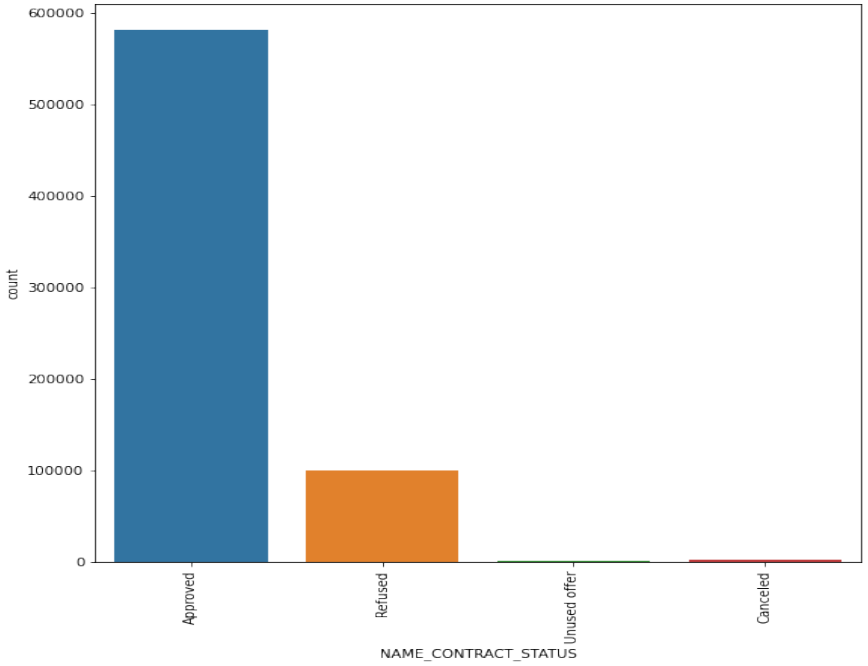
Loan Credit amount and Total Income

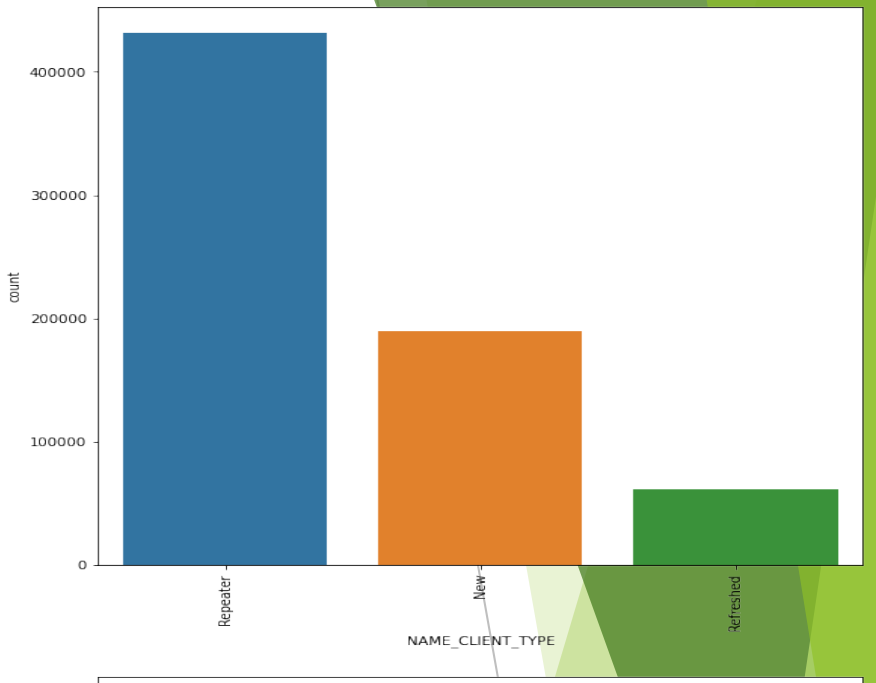
Here also, we can hardly find any interesting pattern.

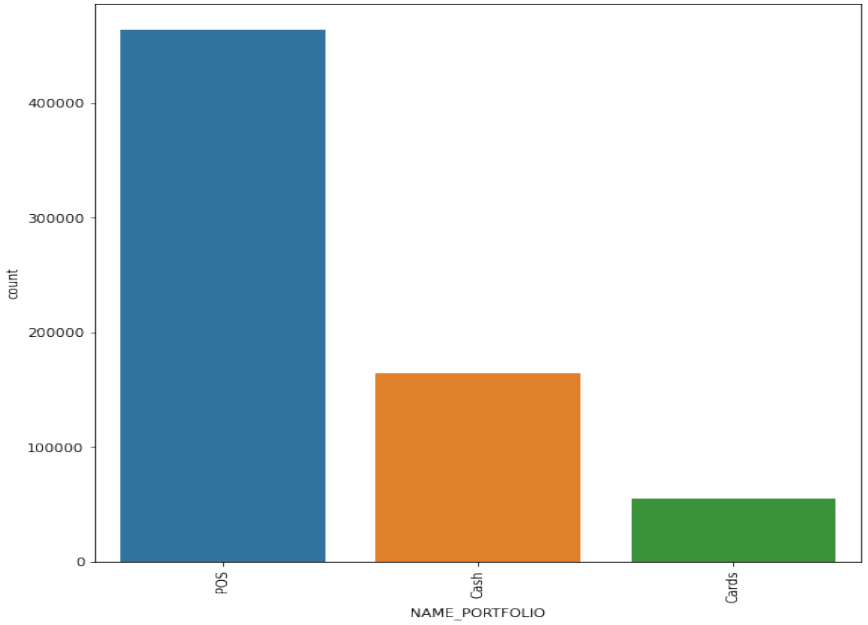


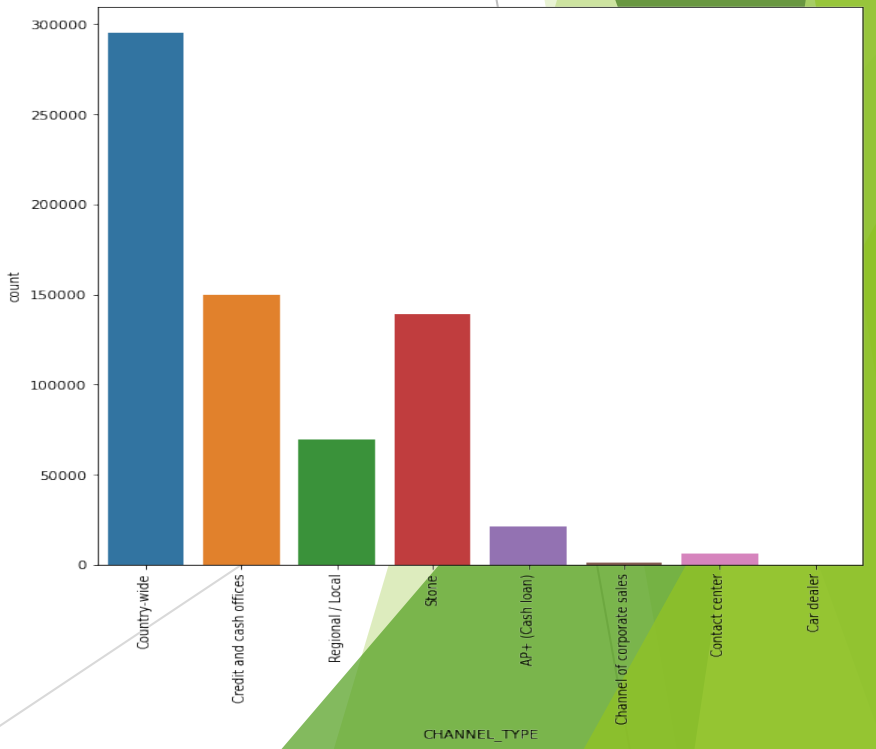
Univariate Analysis on categorical columns in Previous Application data

1.Approved loan status is huge than rejected or canceled.
2.Repeater clients are highest in number than new client.
3.POS loans are highest rather than cash loans.
4.Country-wide channel type is the most used channel followed by Credit and cash offers.









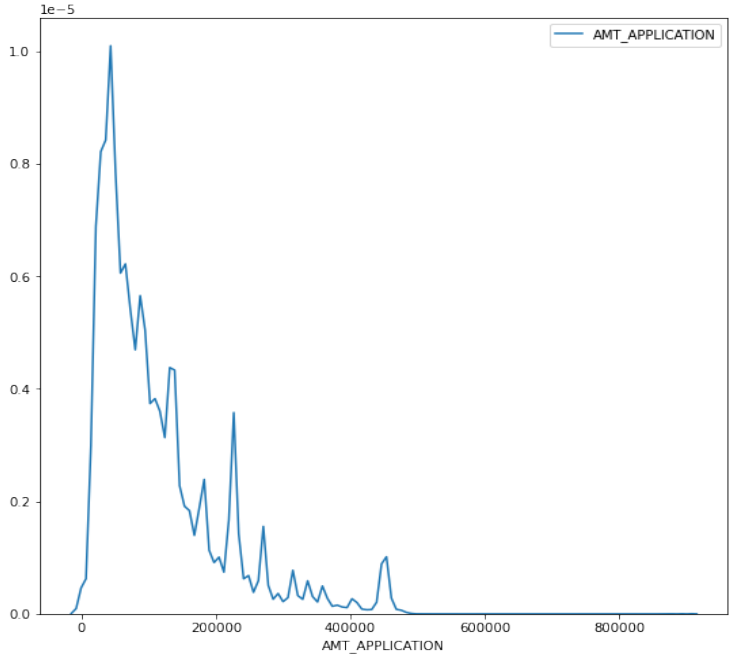
Univariate analysis on continuous columns

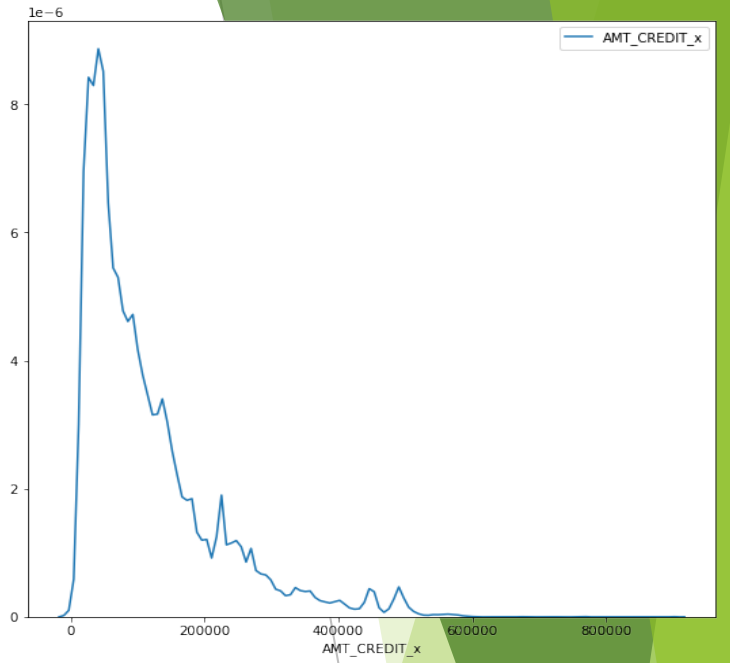
1.Most of the loan application amount were below 500000, we can see a huge spike around 100000 amount.

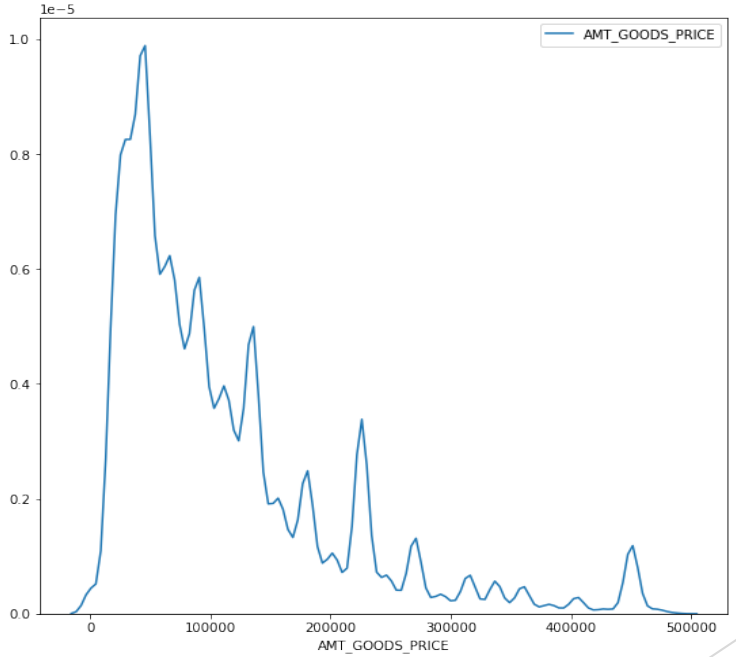
2.Amount credited, is also following the pattern of loan application. We already saw that most of the application was approved in previous plots.

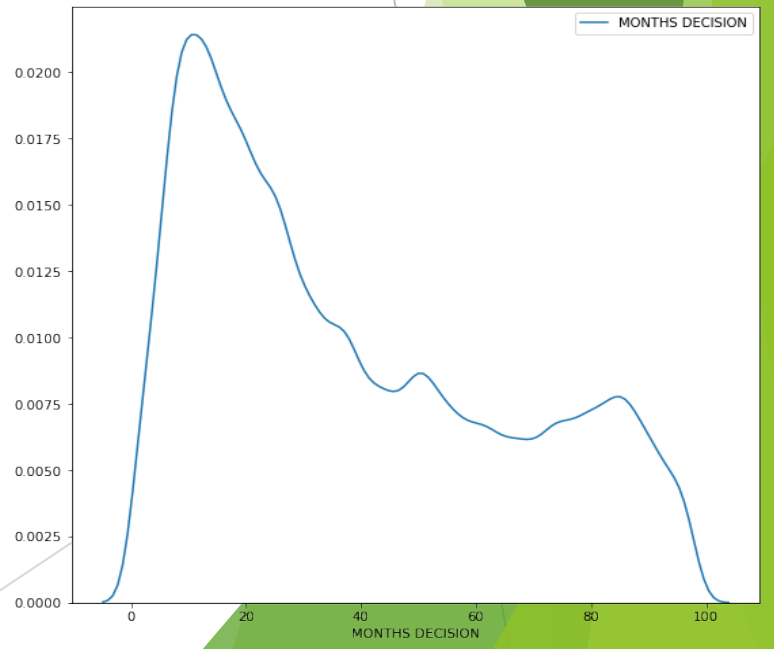
3.Amount of the goods price is also following the same distribution like application amount and amount credited. Because, based on the price of the goods, the loan was approved and amount was credited.

4.Most of the applications decision took around 10 to 30 months.



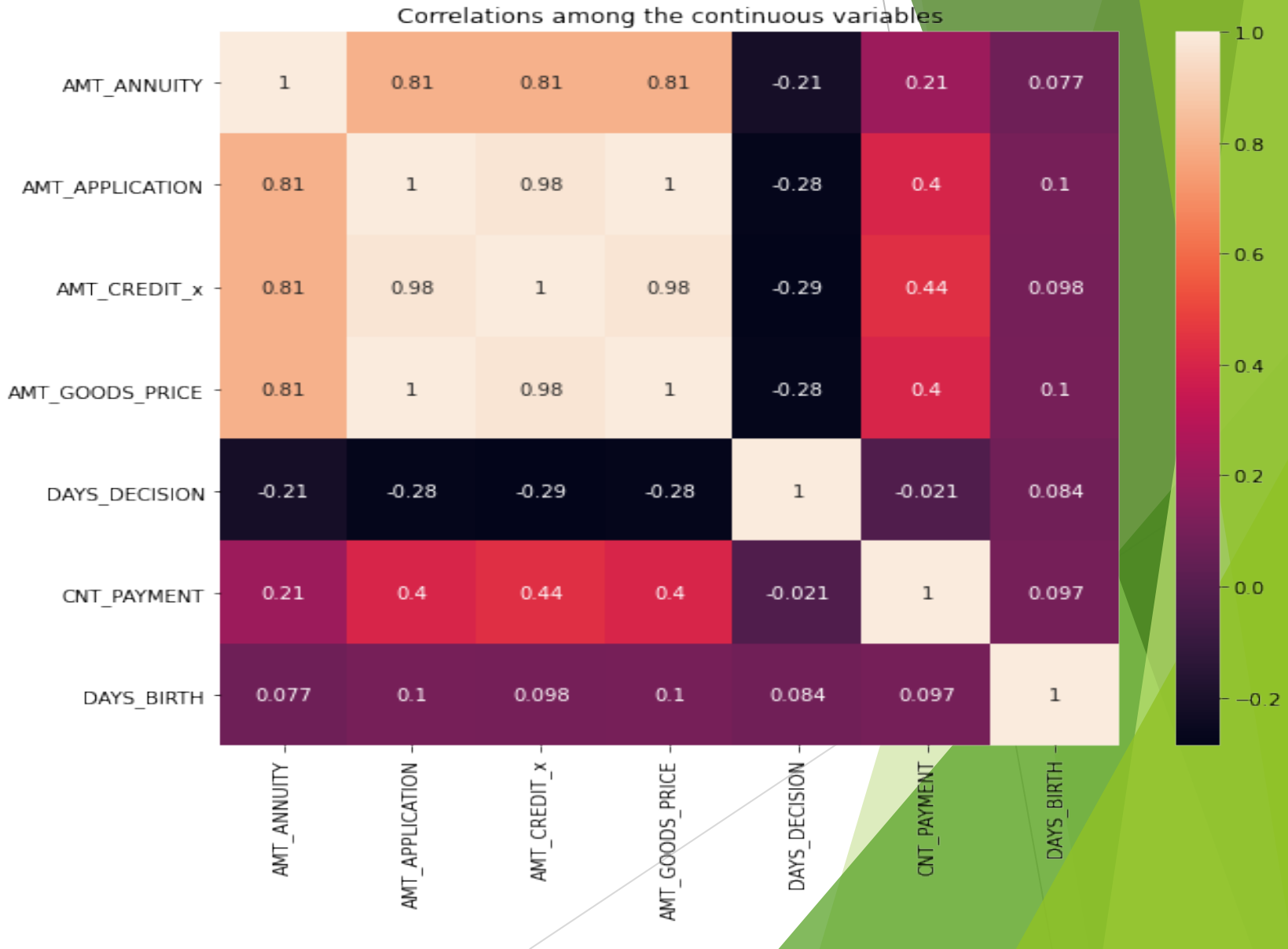






Bivariate analysis on Categorical variables.

There are strong correlations between below variables DAYS_BIRTH(AGE) is correlated with all the variables AMT_APPLICATION is correlated with AMT_ANNUTY, AMT_AMT_CREDIT, AMT_GOODS_PRICE.



Bivariate analysis on continuous columns

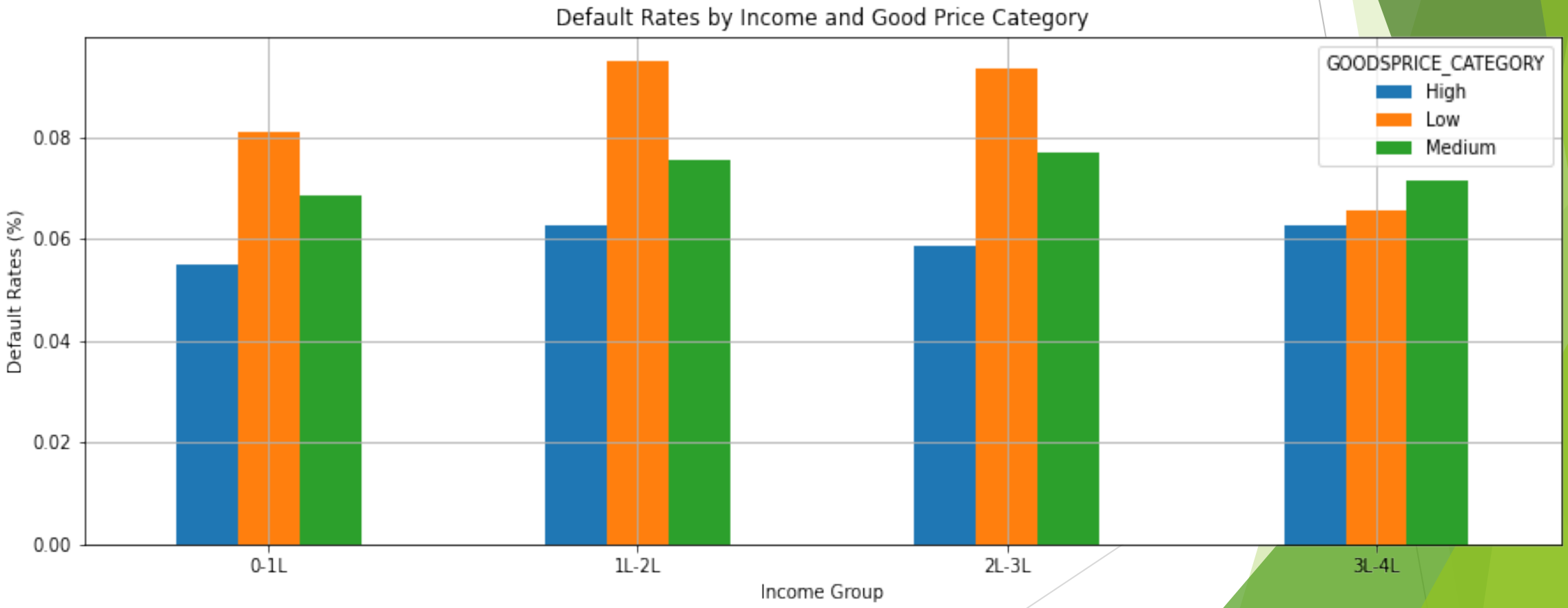
AMT_GOODS_PRICE and AMT_CREDIT are positively correlated and mostly concentrated near the lower region. High AMT_CREDIT loans are most likely to be refused.

Credit amount and the application is highly correlated.

Segmented univariate analysis

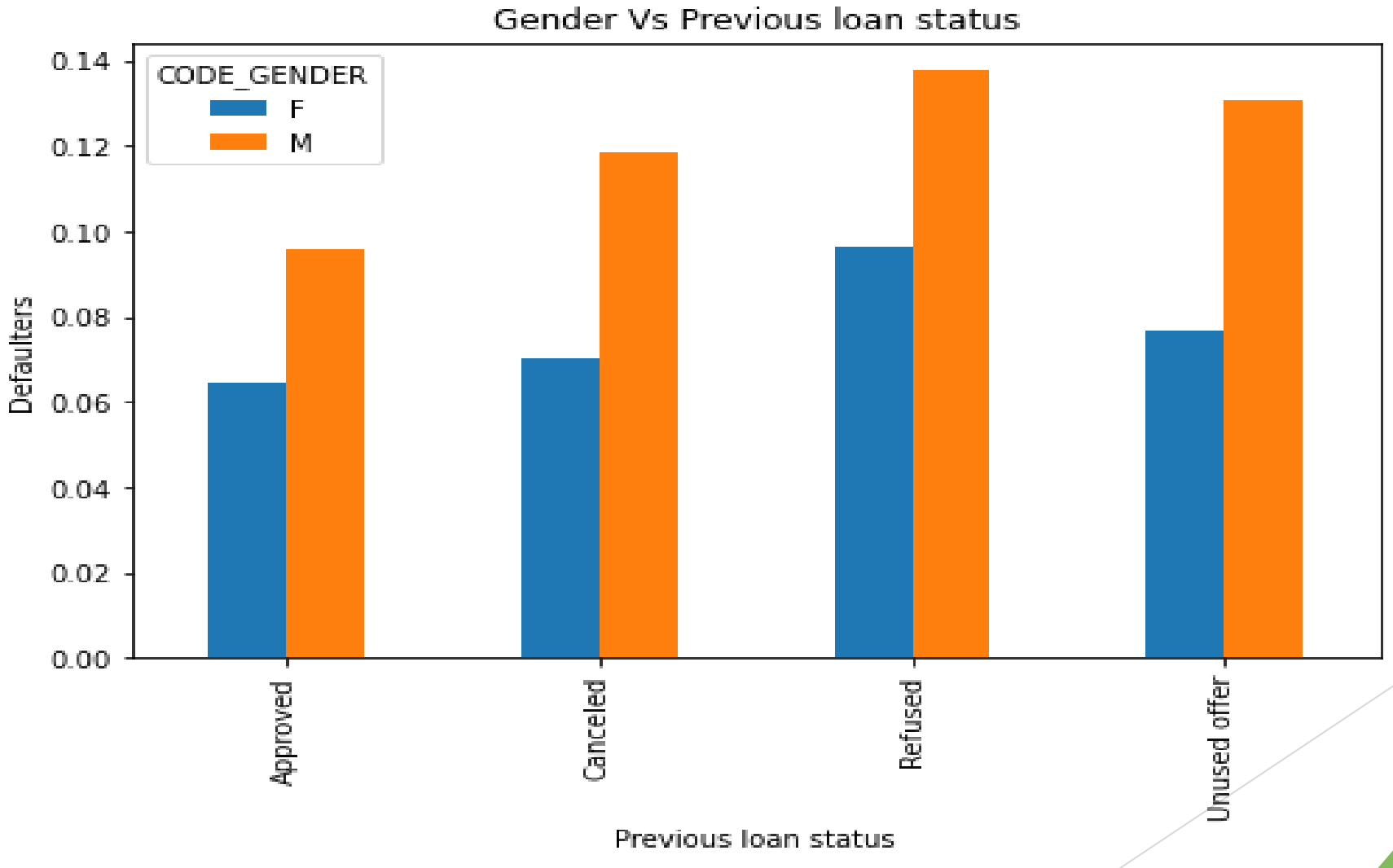
AMT_INCOME_RANGE & GOODSPRICE_CATEGORY

From the above analysis, we find that irrespective of the income groups, the lowest price of the good has the highest chances of default . Interestingly, the highest price category of goods has the lowest probability of default for all the income groups.



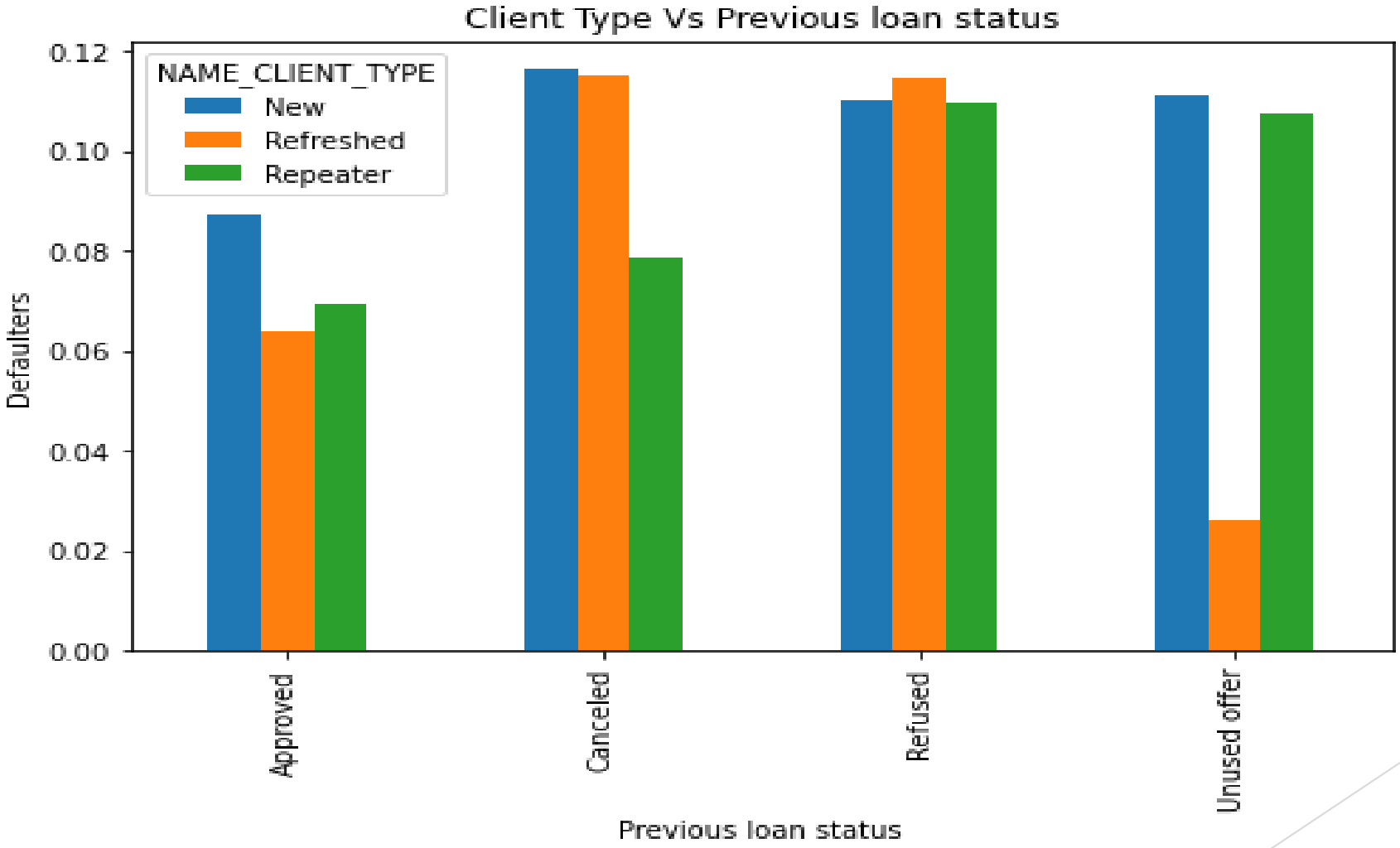
Previous loan Status Vs Current Defaulters Plot

Male clients are more defaulted than female client. Also, previously refused customer are more defaulted in current application.



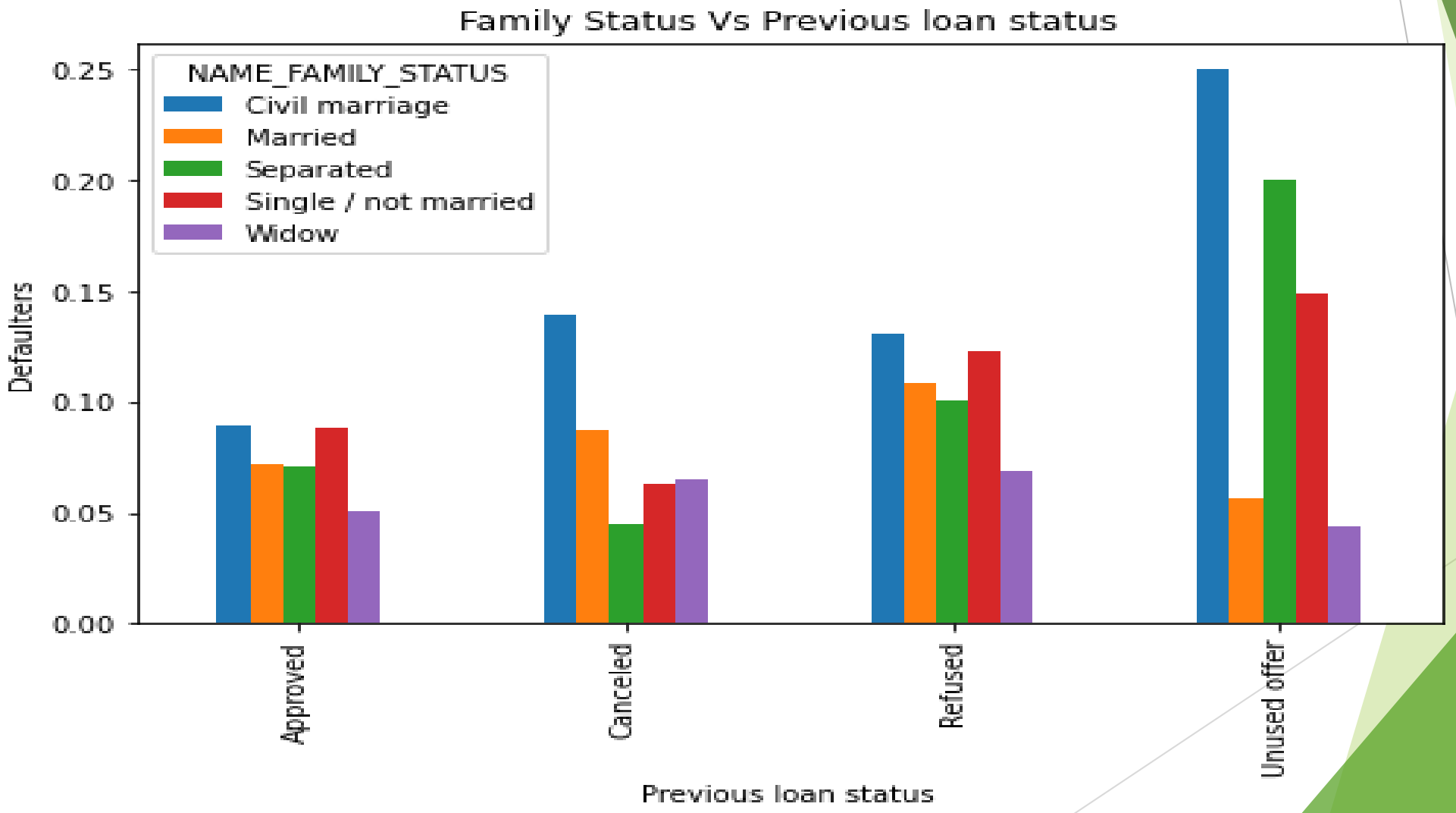
Client Type Vs Previous loan status plot

Previously cancelled New and Refreshed clients are more defaulted than repeater clients



Family Status Vs Previous loan status

Client who did civil marriage with previously unused loan offers are more defaulted currently.



Education status Vs Previous loan status

Previously refused people with lower secondary education are more defaulted in current application.