



Credit EDA Case Study

Risk analytics in banking and financial services

Presentation By: Saish Hawa



Problem Statement:

To find and analyse the patterns present in the data using EDA to ensure that the clients able to repay the loan are not rejected.

Two types of risks are associated with the bank's decision:

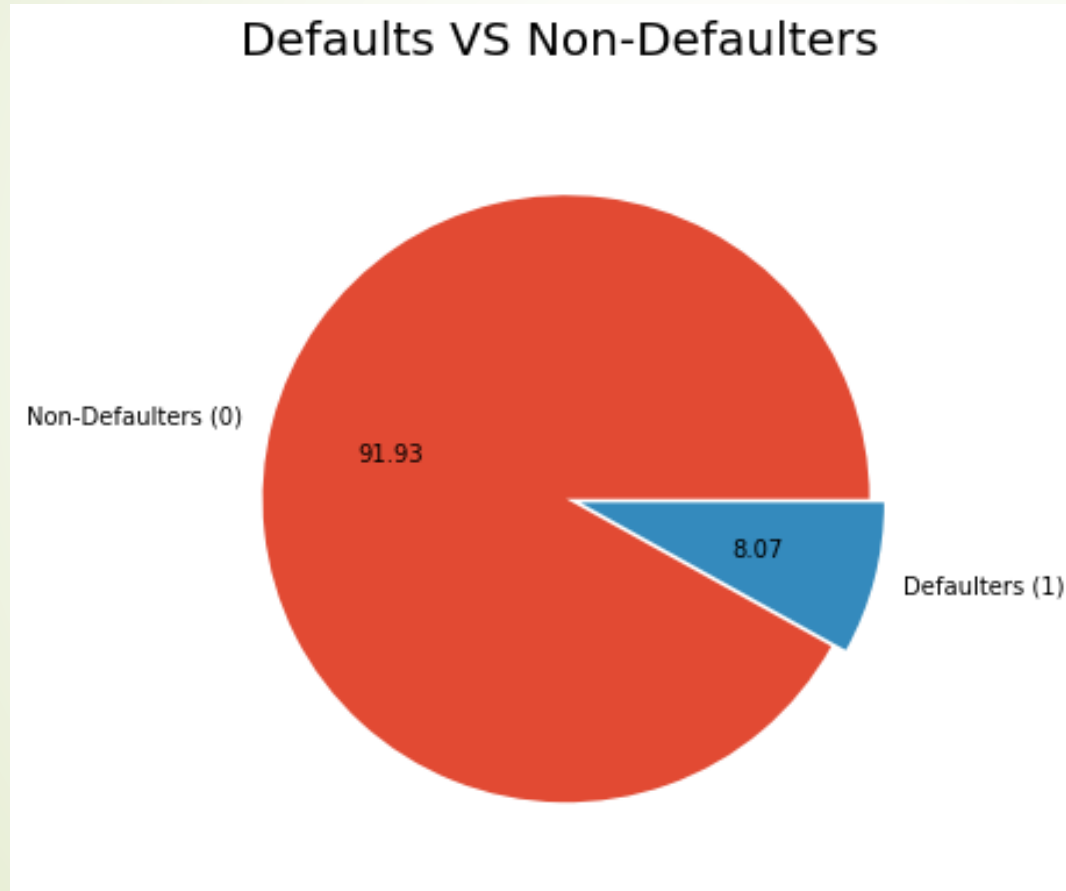
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.



Steps preformed

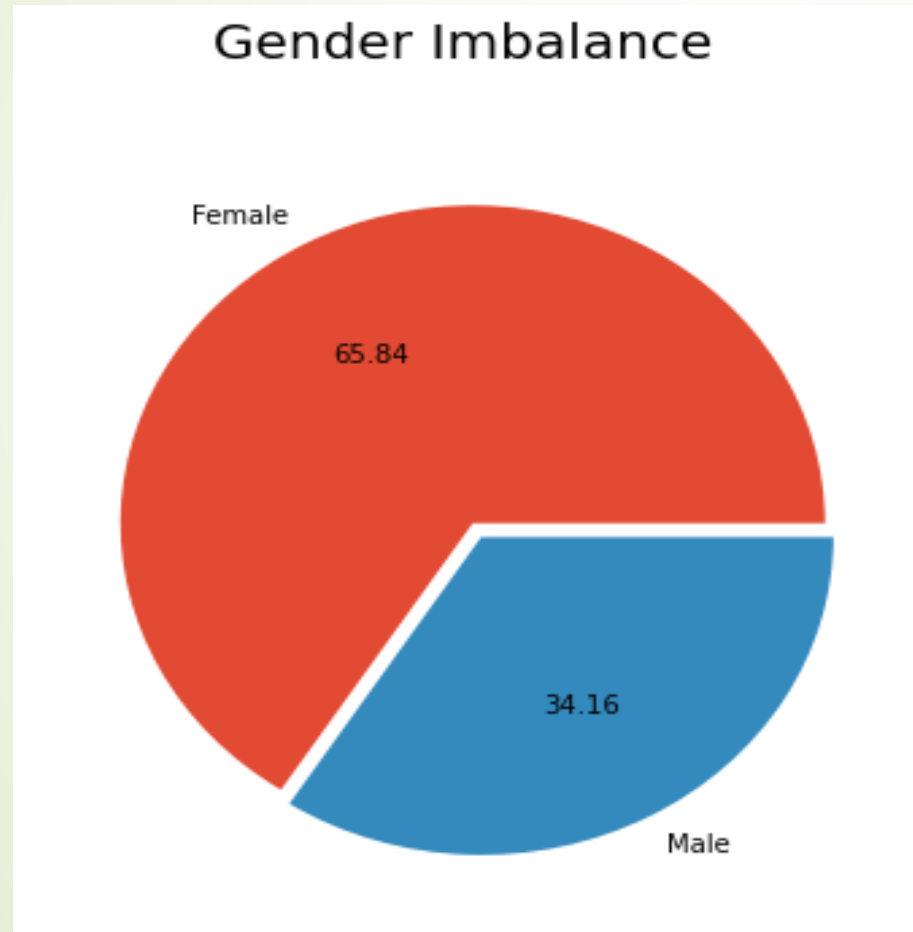
- Two data sets are provided, `application_data` and `previous_application`.
- Basic checks are done on both the data sets like head, shape, info, description and data types.
- Checked the null value percentages in all the columns.
- The columns having greater than 40% null values are dropped.
- The columns having less than 13% null values are selected for the value imputations.
- Columns are checked for the outliers using boxplot and based on the result, mean/median value is suggested for the imputations.

Data Imbalance: Defaulters and Non-Defaulters



Around 92% clients repaid their loans on time i.e. Non-Defaulters and around 8% clients did not repay their loans on time i.e. Defaulters.

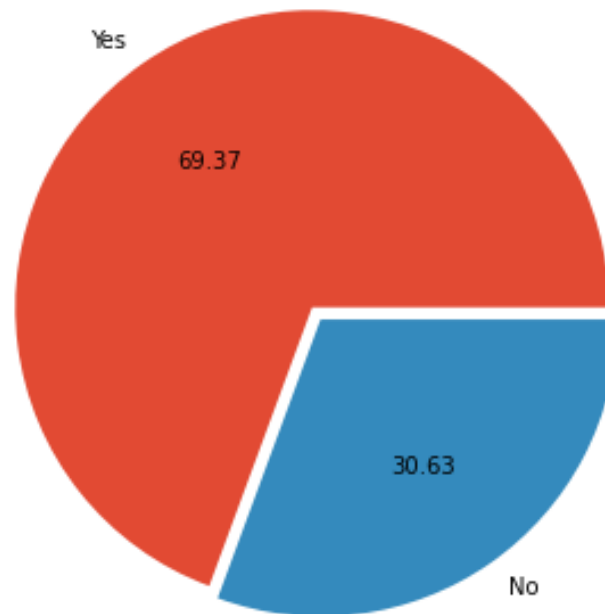
Data Imbalance: Gender



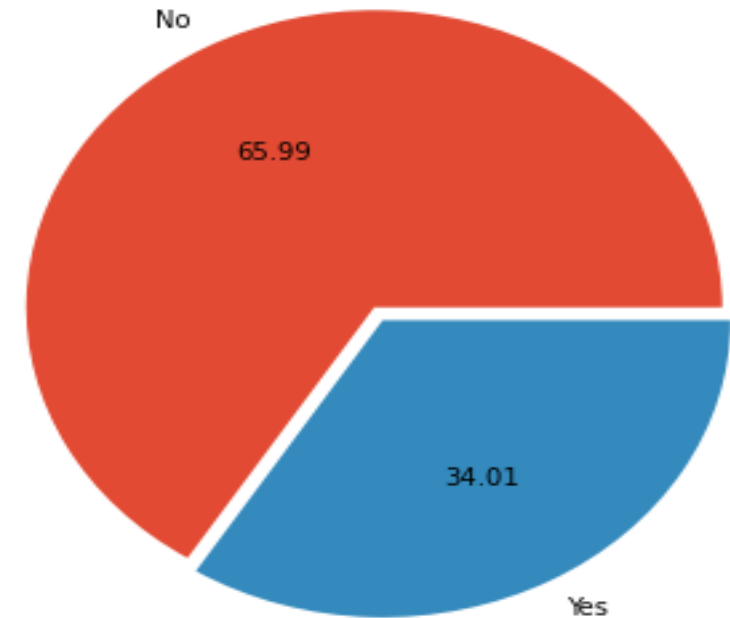
Out of the total loan applications, 65.84% are from Female clients and 34.16% are from Male clients.

Data Imbalance: House and Car Ownership

Imbalance in House / Flat Ownership



Imbalance in Car Ownership

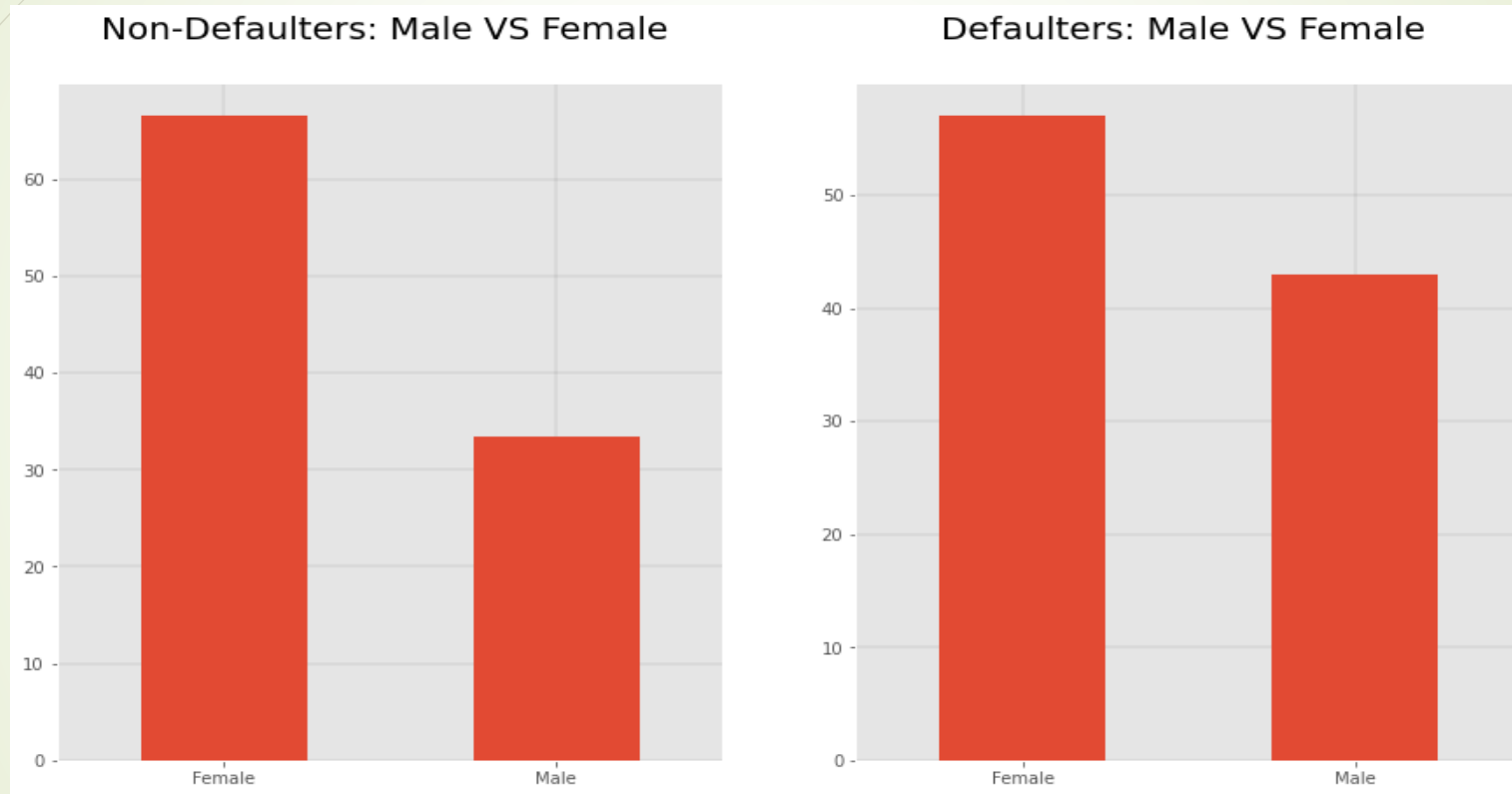




Creating new data frames

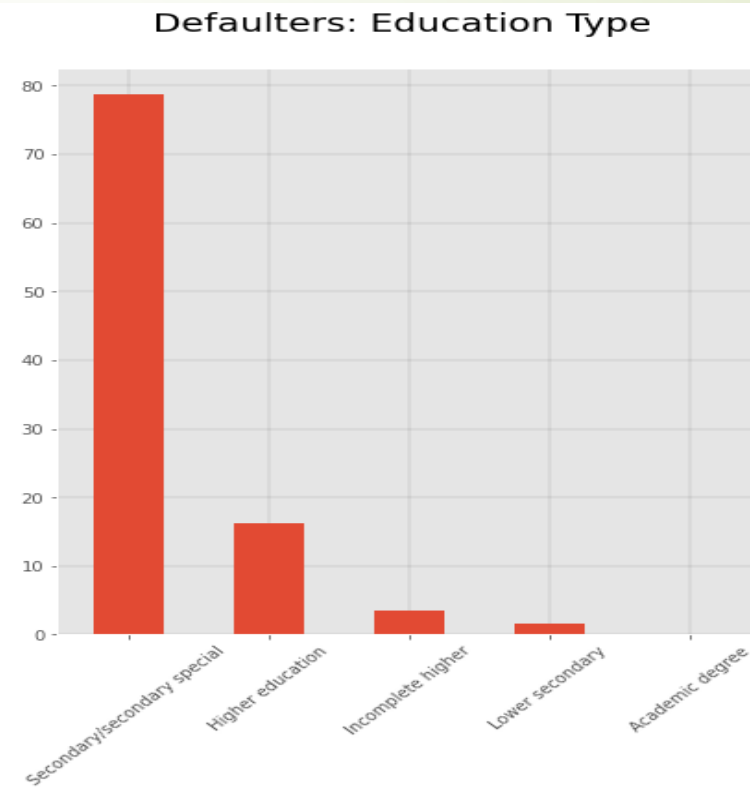
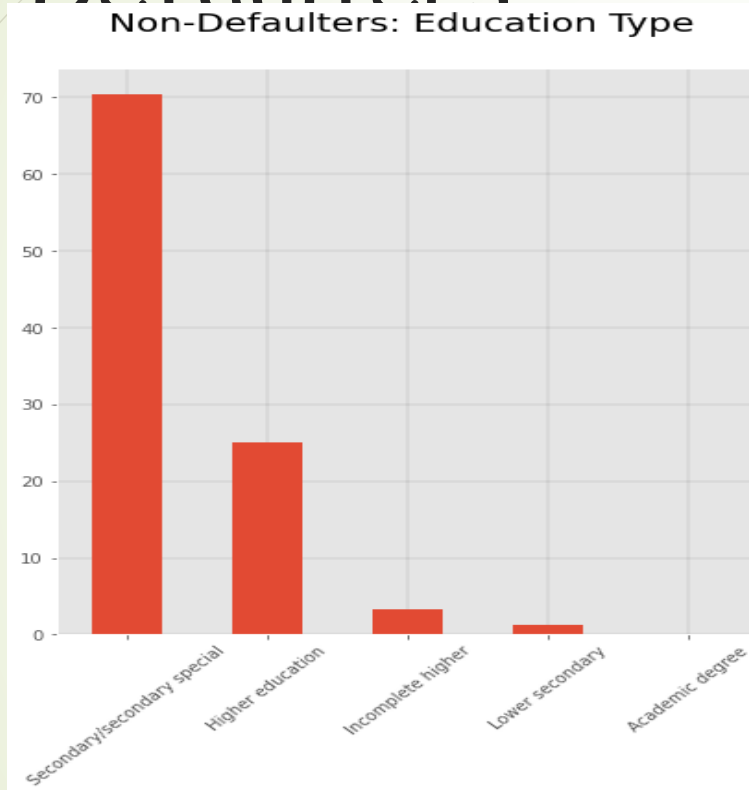
- Two new data frames are created from the application_data data set based on the Target value.
- One data frame with target value 0 contains the data for all the non-defaulters. Clients repaying their loans on time.
- Other data frame with Target value 1 contains the data for all the defaulters. Clients not repaying their loans on time.
- Using these two data frames we will proceed with the EDA.

Gender: Defaulters and Non-Defaulters



- High number of females are both non-defaulters and defaulters but, number of non-defaulters is higher.
- There are male defaulters than non-defaulters.

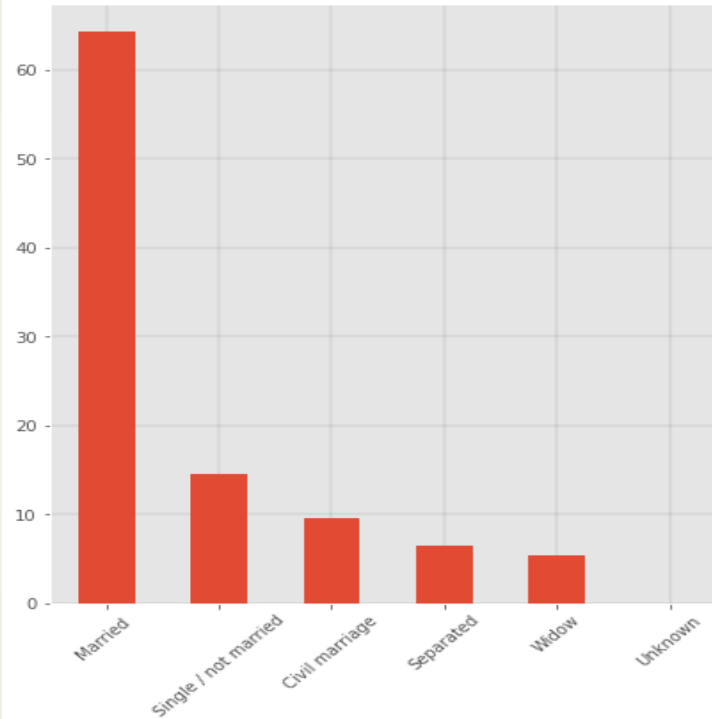
Education Type: Defaulters and Non-Defaulters



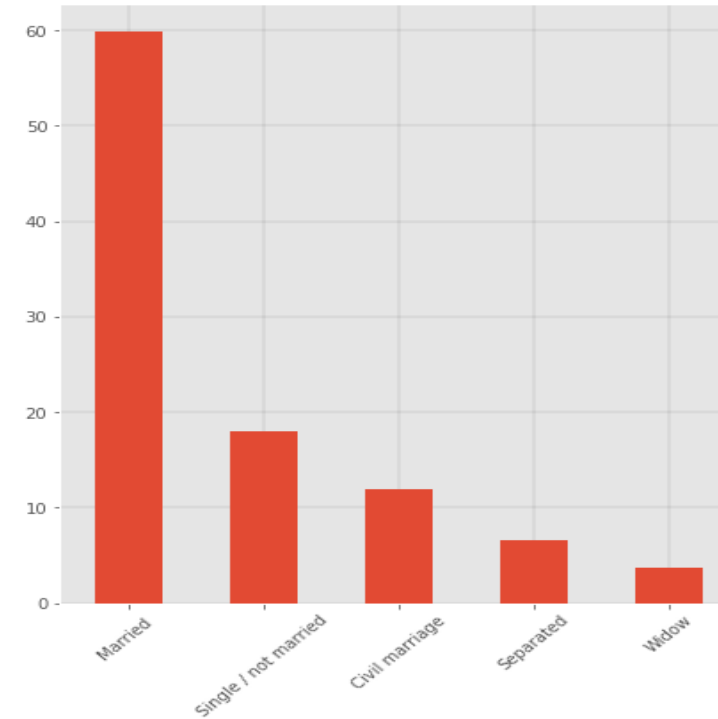
Clients with education type Secondary/secondary special seems to default more than repaying on time.

Family Status: Defaulters and Non-Defaulters

Non-Defaulters: Family Status

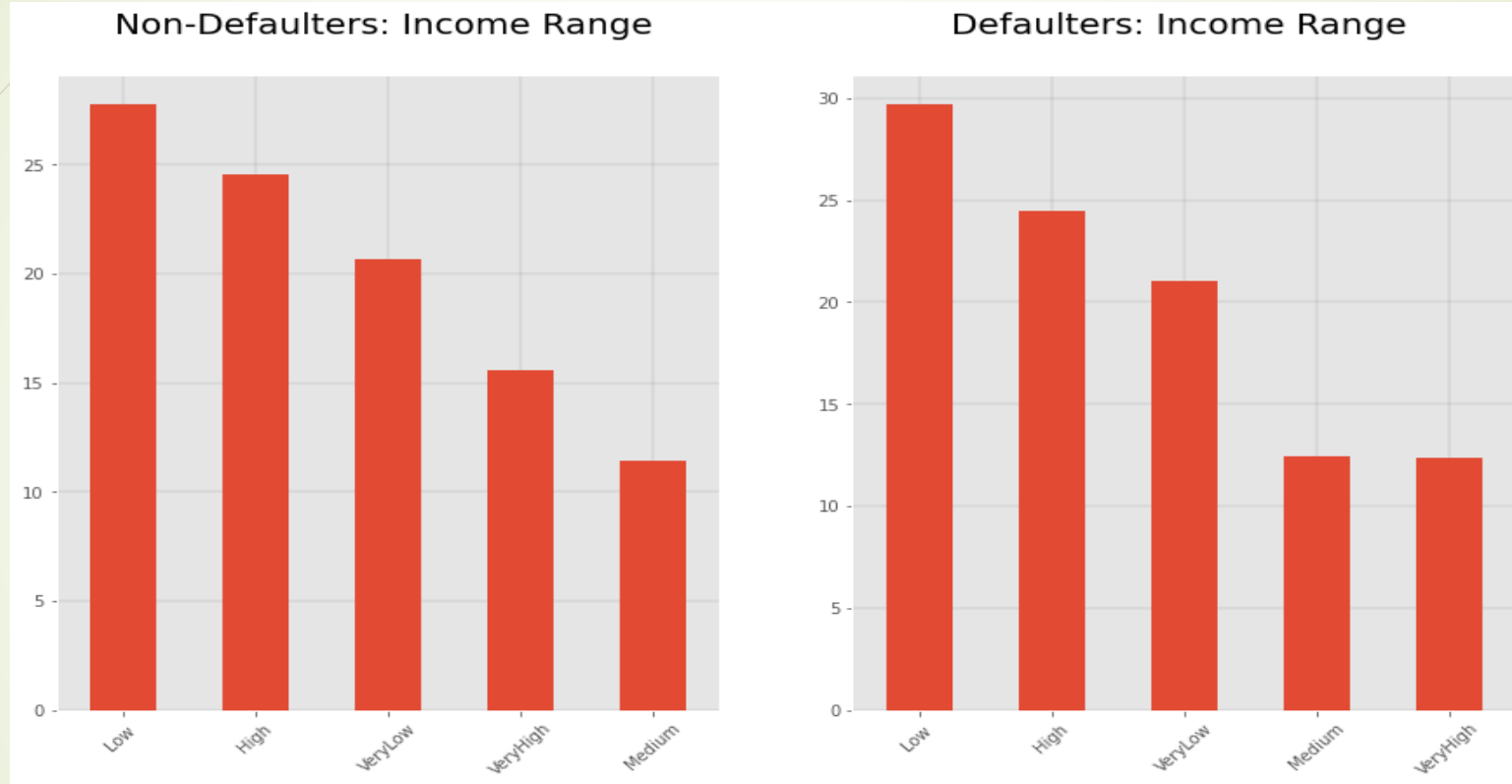


Defaulters: Family Status



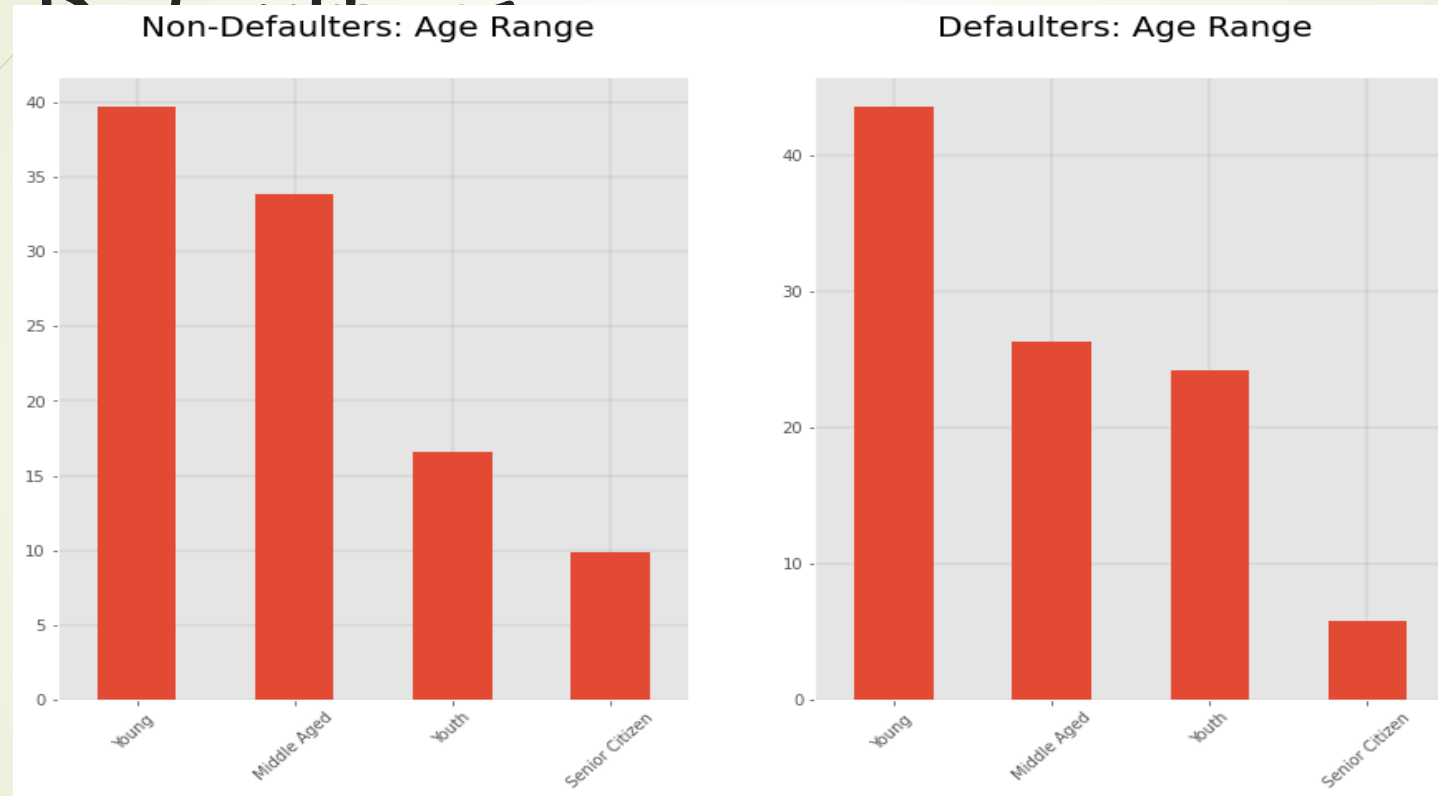
- Married people tends to apply for loan more than other people.
- Single/not married and Civil marriage status clients tends to default their loans more than repaying on time.

Income Range: Defaulters and Non-



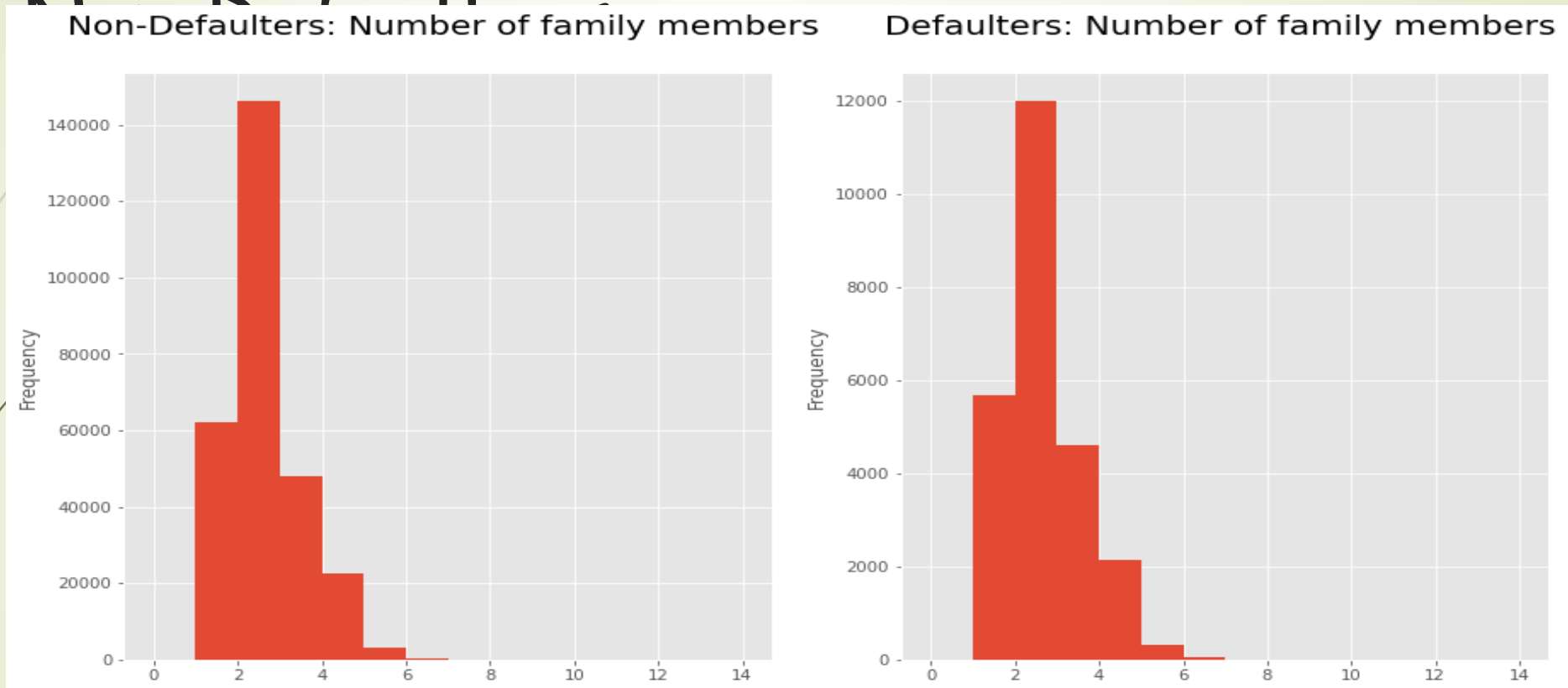
- Clients in Very High income range tends to repay their loans on time more than defaulting the loans.
- Clients in Low income range tends to default their loans more than repaying on time.
- Clients in High, Medium and Very low income range have a very equal distribution in both defaulting and non-defaulting so, they have a 50-50% chances of defaulting or non-defaulting the loan.

Age Group: Defaulters and Non-



- Youth (20-30 yrs) and Young (30-45 yrs) clients tends to default more than repaying on time.
- Middle Aged (45-60 yrs) and Senior Citizen (60-70 yrs) clients tends to repay their loans on time more than defaulting.

No. of family members: Defaulters and



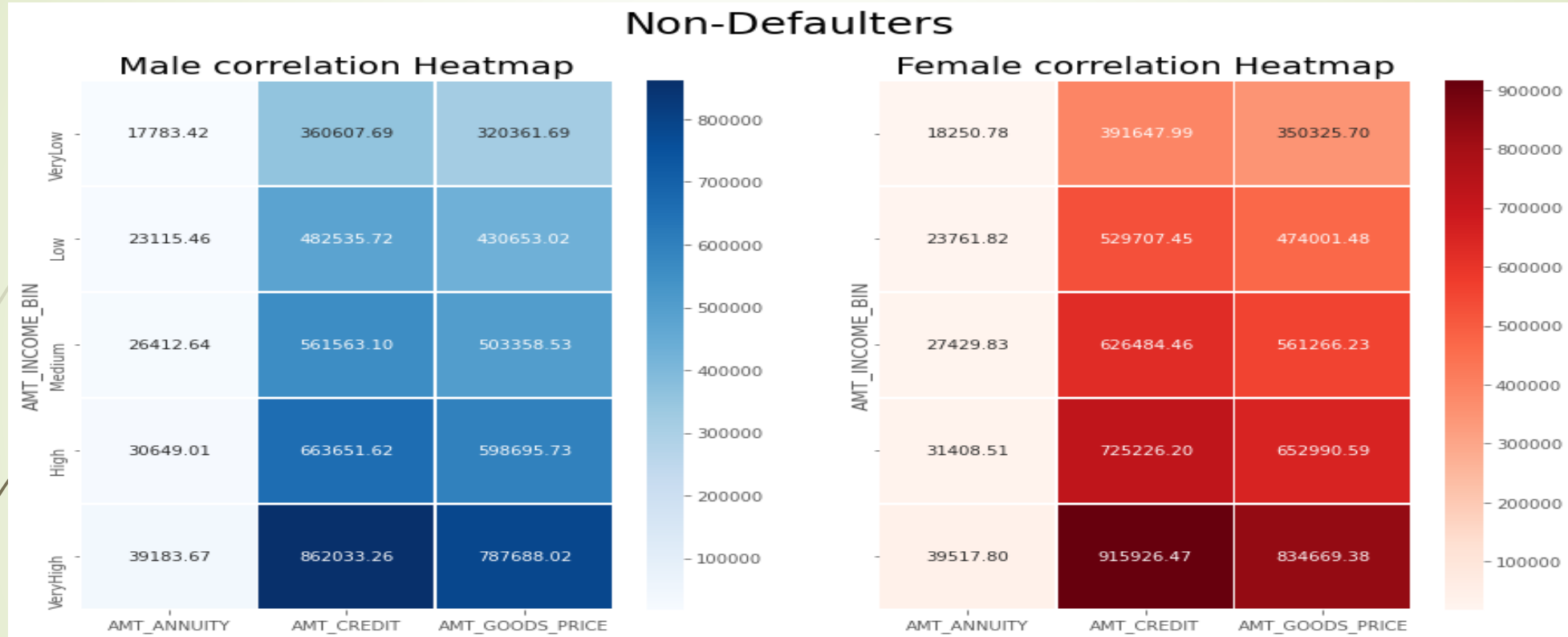
- Families of 1 and 2 members repay the loan on time more than defaulting.
- Majority of the loan applications are from the families of 2 members.



Correlation based on the Target variable

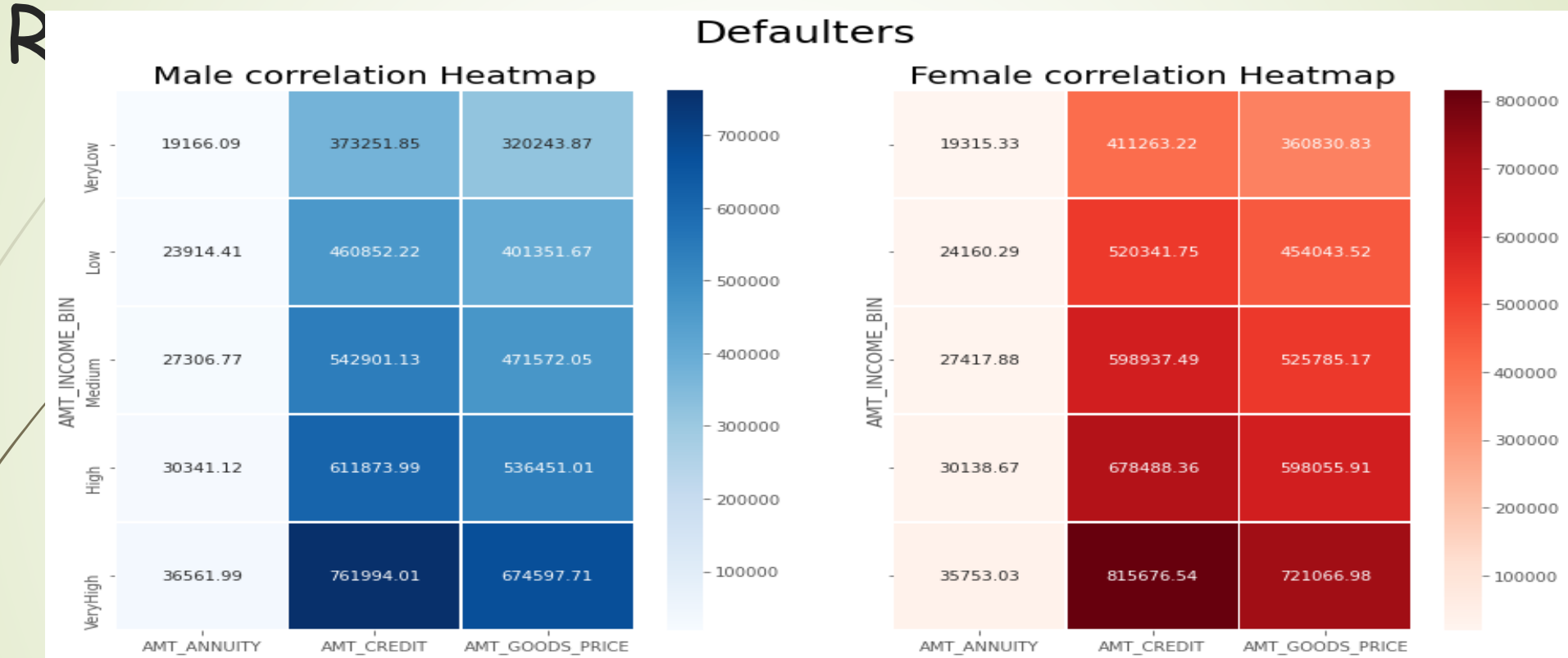
- Creating two new data frames for male data and female data from the previously created Target 0 and Target 1 variable.
- New data frames created are:
 1. Non-Defaulters Male
 2. Non-Defaulters Female
 3. Defaulters Male
 4. Defaulters Female

Non-Defaulters: Correlation between Annuity, Credit and Goods Price based on



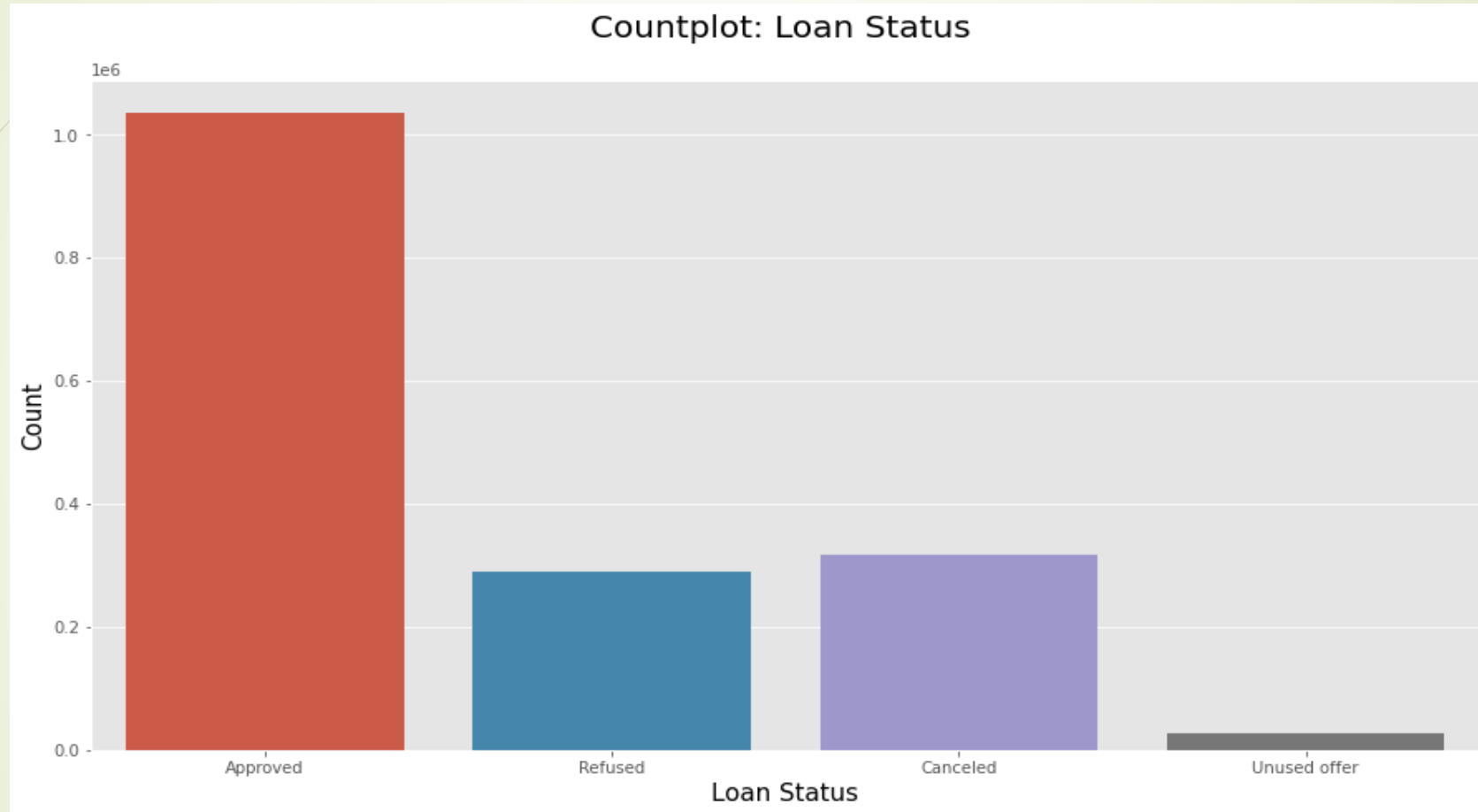
- Both male and females under the Very High category of income have highest Annuity amounts, tends to take huge loan amounts and have highest goods price amount compared to the other income category clients.
- Both male and females under the Very Low category of income have lowest Annuity amounts, tends to take smaller loan amounts and have lowest goods price amount compared to the other income category clients.
- There is a direct proportion between the income and loan amount, the higher the income higher the loan amount for both male and female.

Defaulters: Correlation between Annuity, Credit and Goods Price based on Income



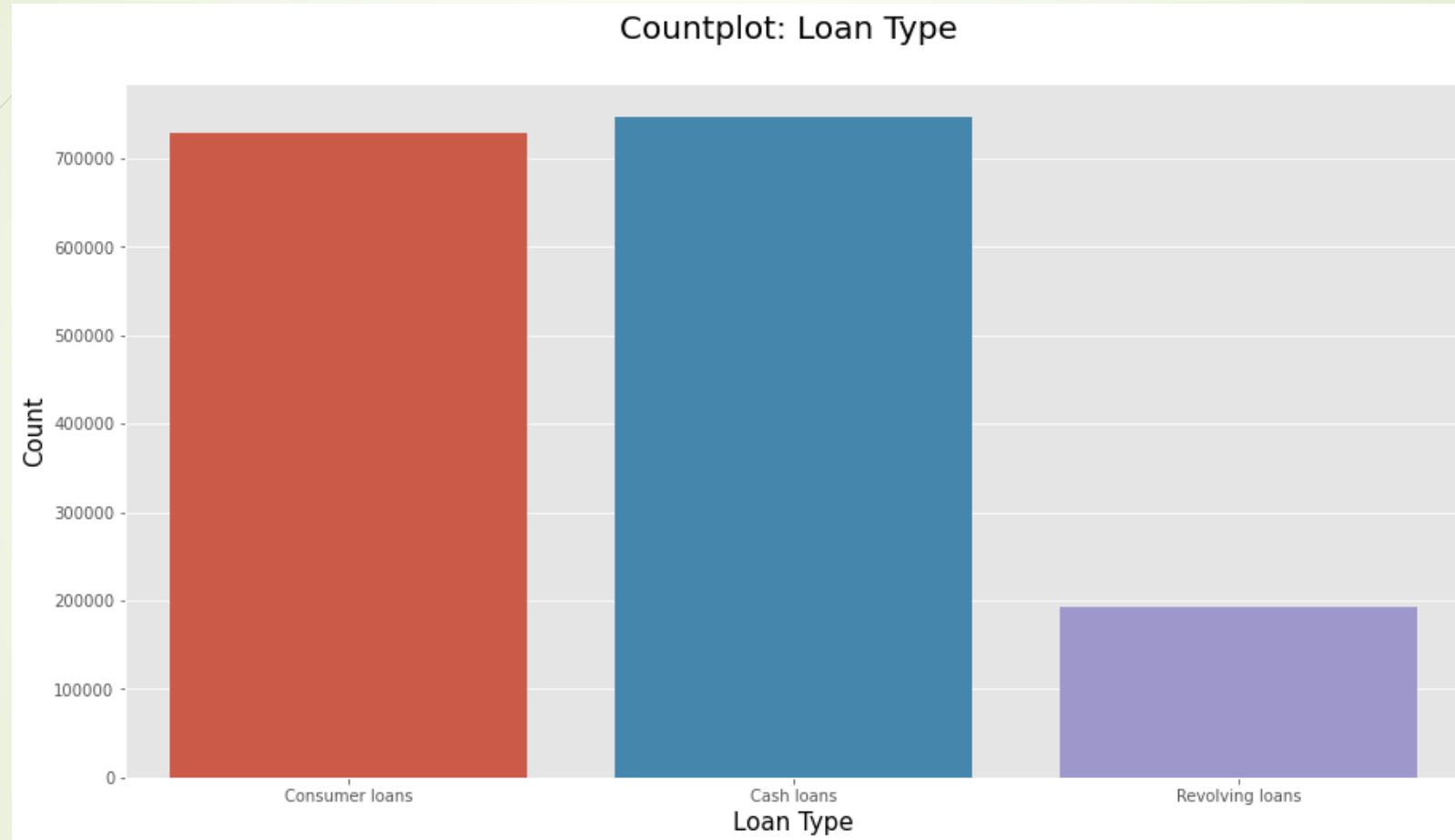
- Both male and females under the Very High category of income have highest Annuity amounts, tends to take huge loan amounts and have highest goods price amount compared to the other income category clients.
- Both male and females under the Very Low category of income have lowest Annuity amounts, tends to take smaller loan amounts and have lowest goods price amount compared to the other income category clients.
- There is a direct proportion between the income and loan amount, the higher the income higher the loan amount for both male and female.

Analysis from previous_application dataset:



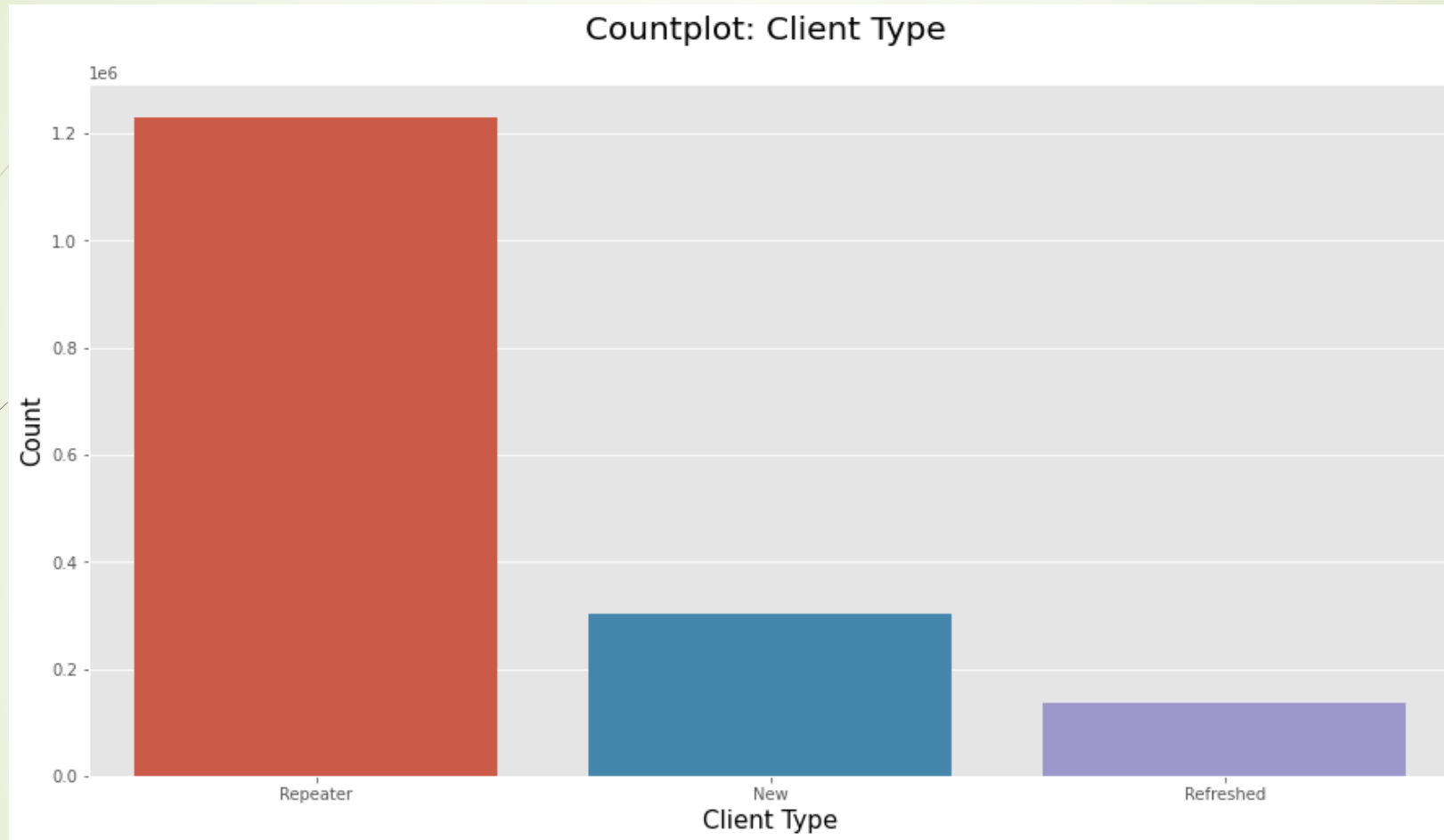
- Most of the loan applications are Approved.
- There are very few unused offers.

Loan Type distribution



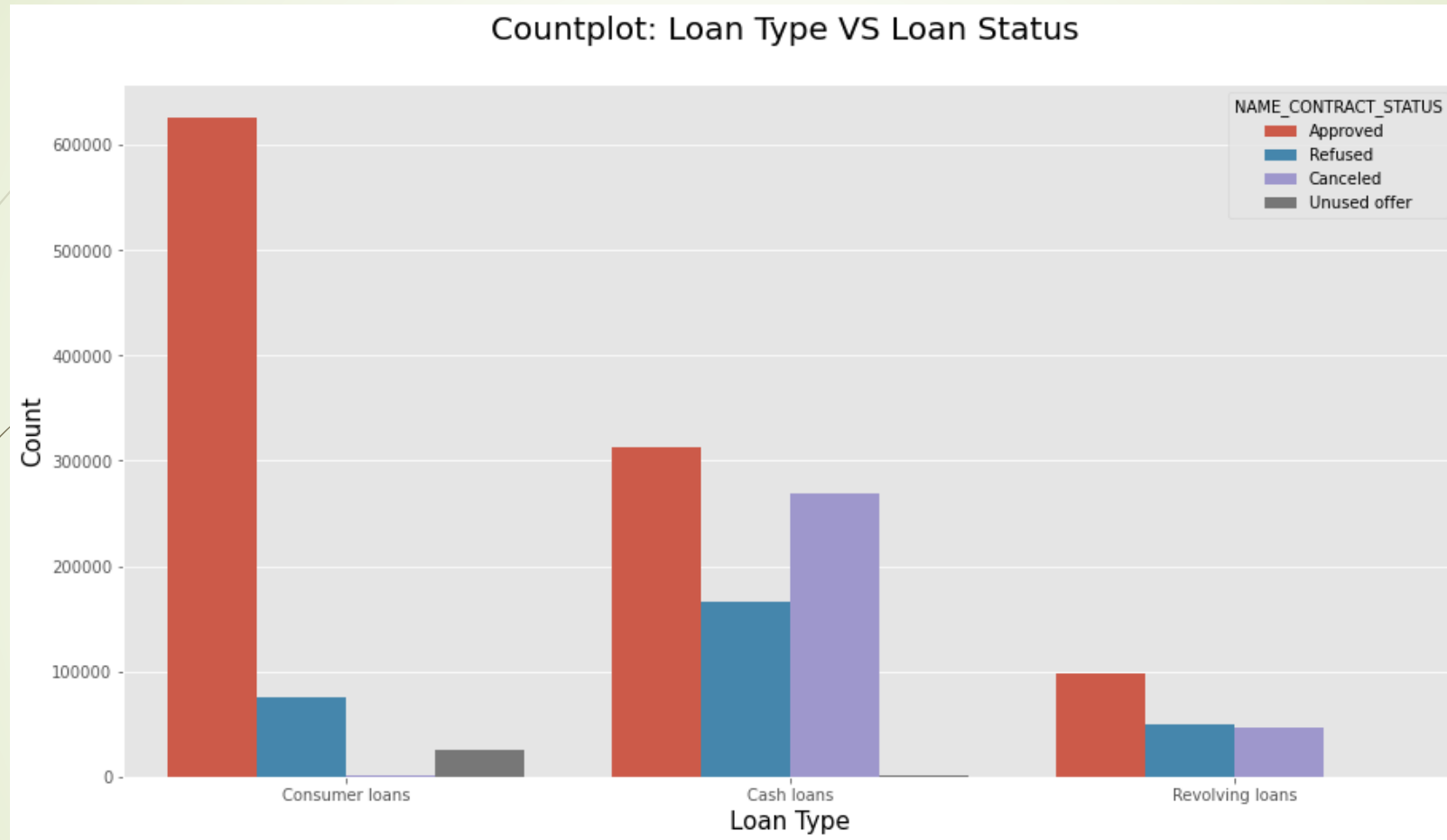
Most of the clients prefer Consumer loans and Cash loans over the Revolving loans.

Client Type distribution



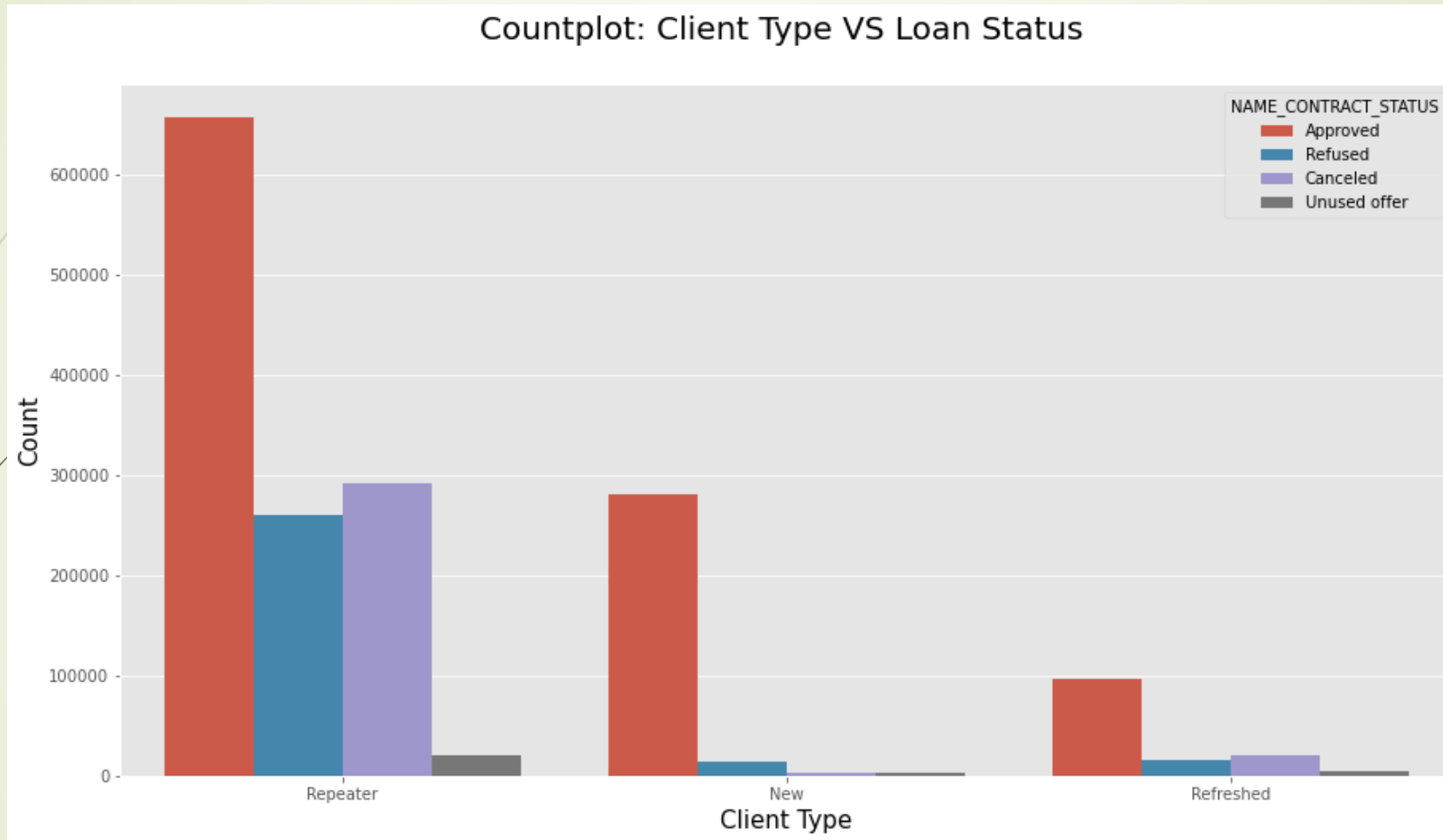
- Most of the loan applications are from the Repeater clients followed by the New clients.
- There are very few Refreshed clients.

Loan Type VS Loan Status distribution



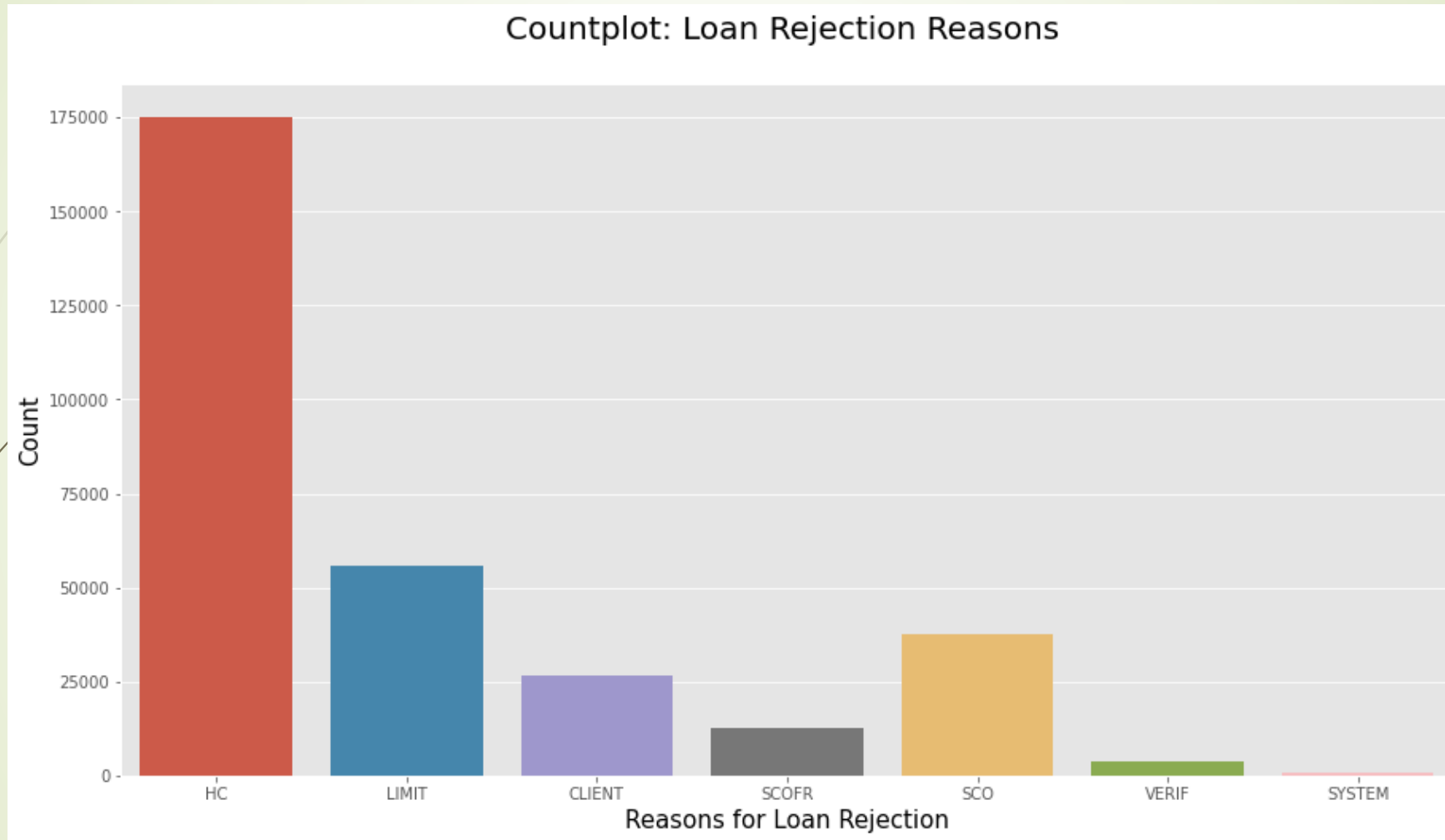
- Most of the applications for Consumer loans are approved and there are few unused offers as well.
- Majority of the applications for Cash loans are approved but, there are many applications which are cancelled and refused.

Client Type VS Loan Status distribution



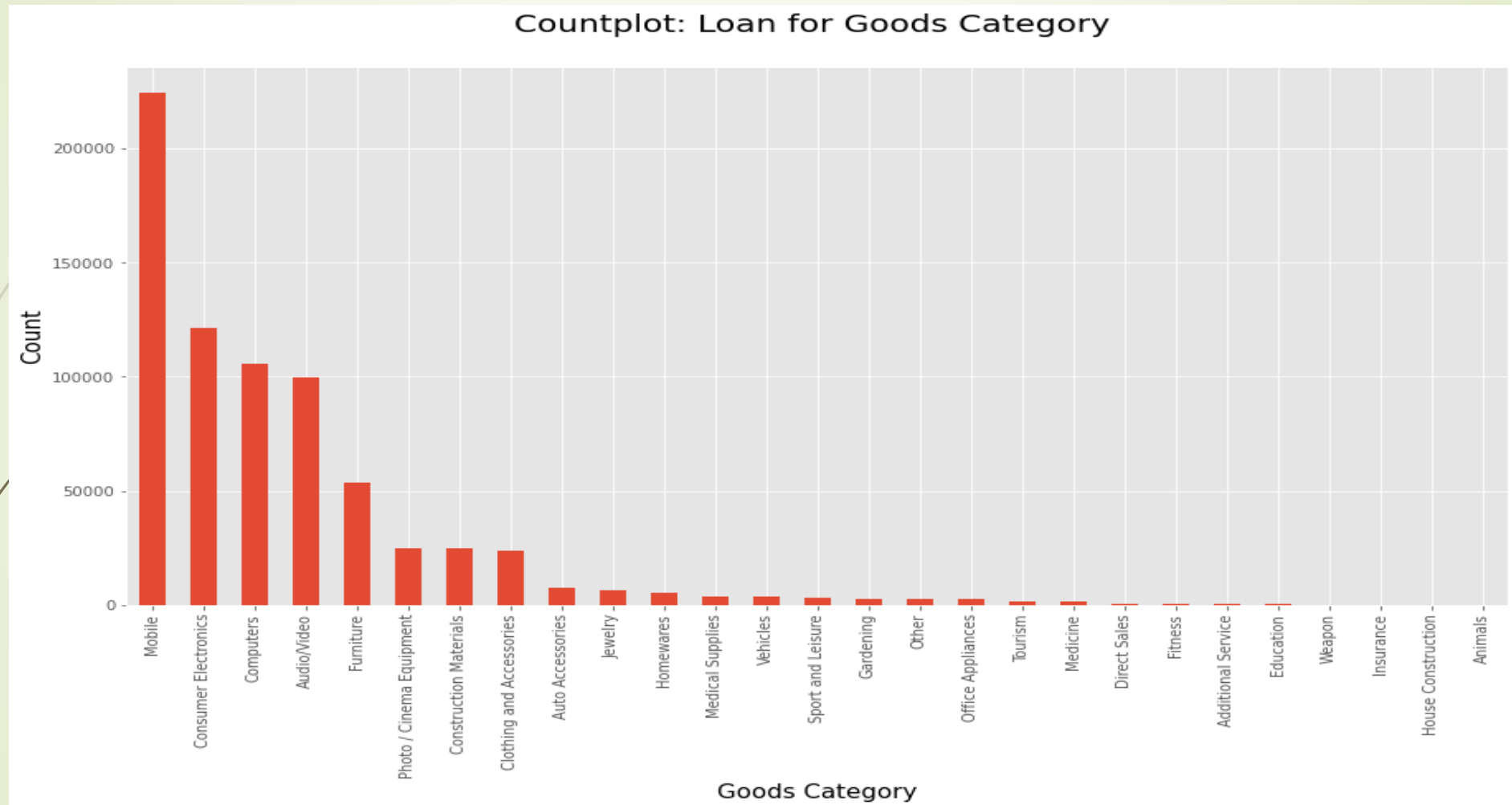
- Majority of the applications from Repeater clients are approved but, there are also quite a few applications which are cancelled and refused.
- Most of the applications by the New clients are approved and very few are refused, cancelled or unused.
- Overall we can say that irrespective of the client type, majority of the applications are approved.

Loan Rejection Reasons distribution



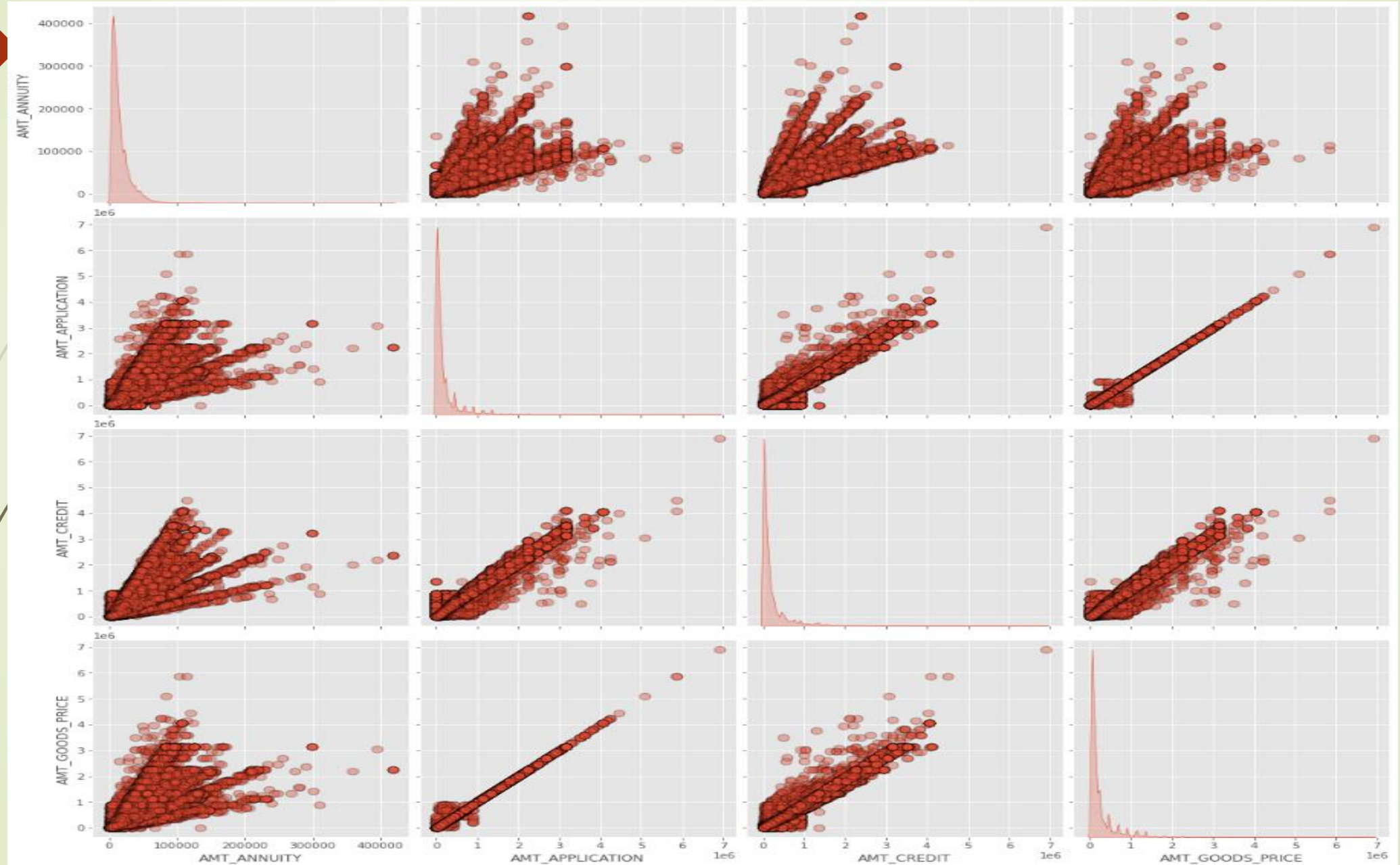
Most number of loans are rejected because of HC followed by Limit.

Types of Goods distribution



Majority of the loans are applied for Mobile followed by Consumer Electronics and Computers.

Bivariate analysis on numerical columns

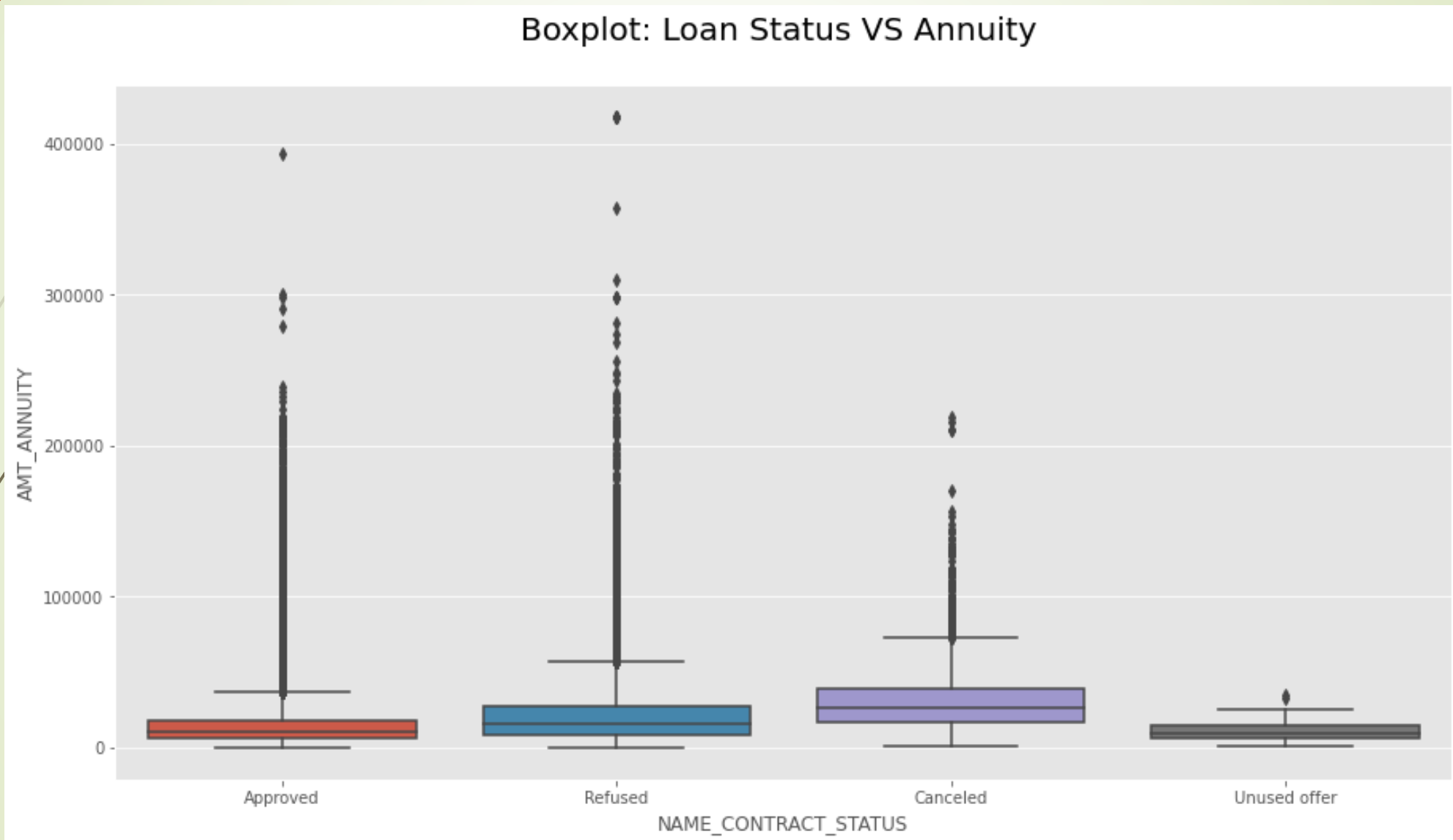


Bivariate analysis on numerical columns

From the plot we can see that:

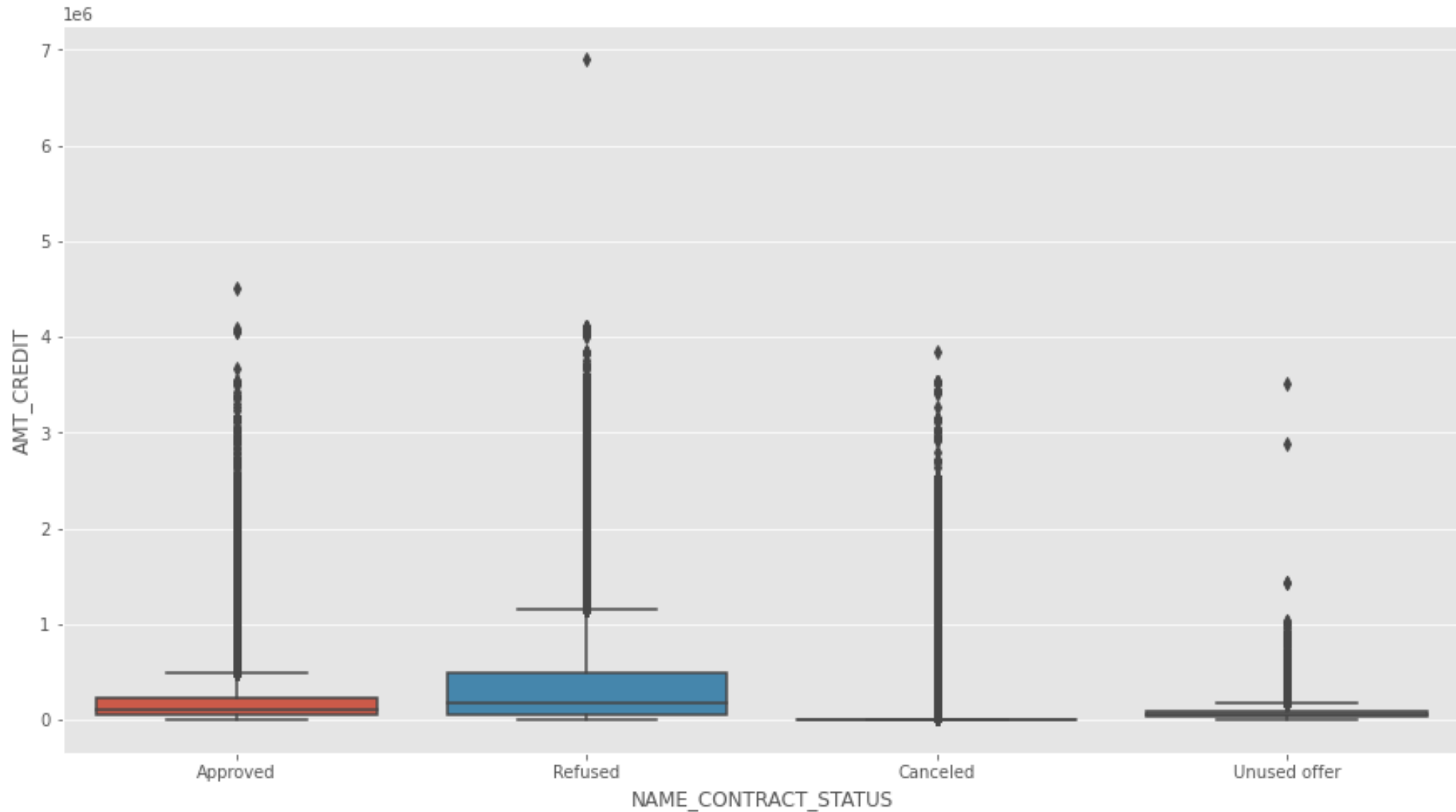
- Credit amount asked by the clients on the previous application is highly related to the Goods price of good that client has asked for on the previous application.
- Final credit amount given to the clients previously is related to the application amount and the goods price of good that client asked for on the previous application.

Loan Status VS Annuity distribution




Loan Status VS Credit Amount

Boxplot: Loan Status VS Credit Amount

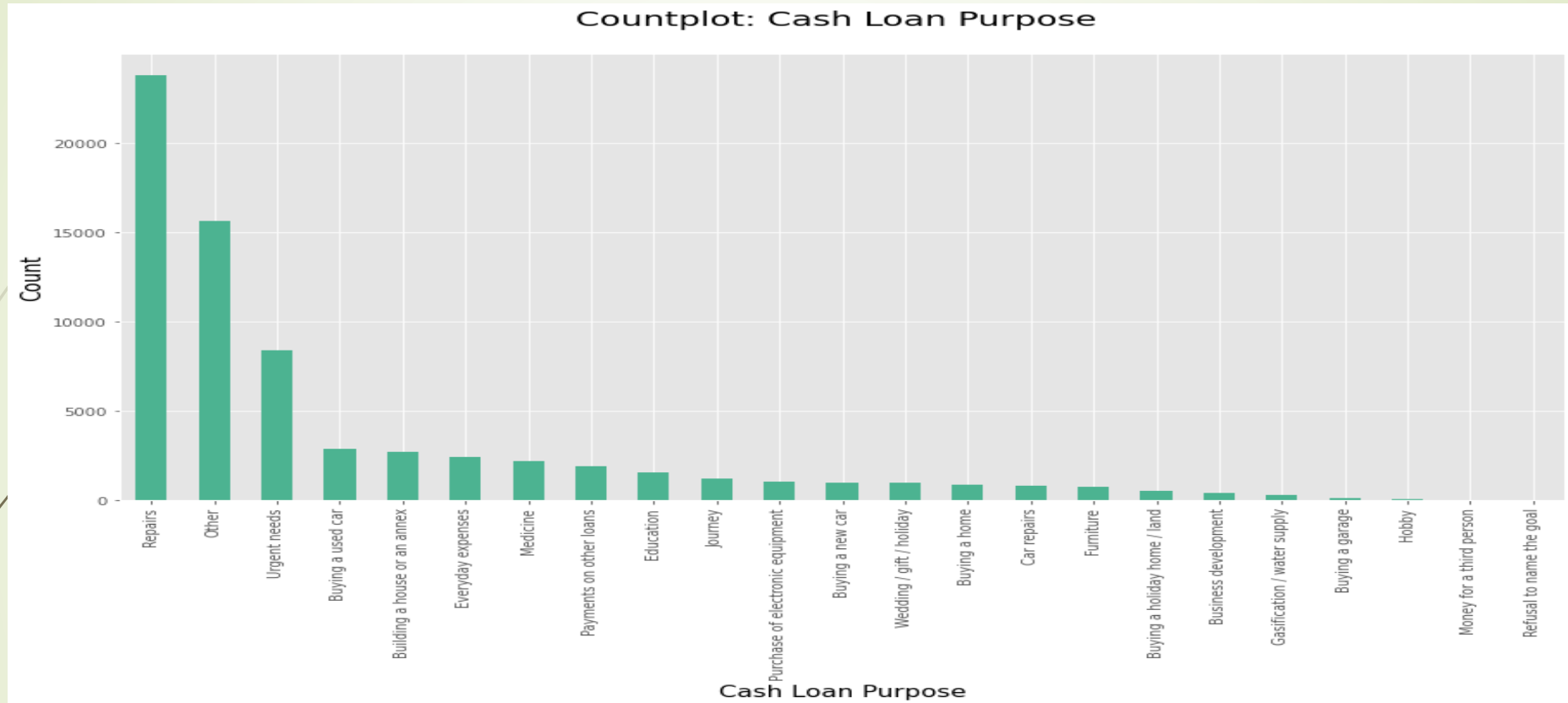




Creating a new data frame

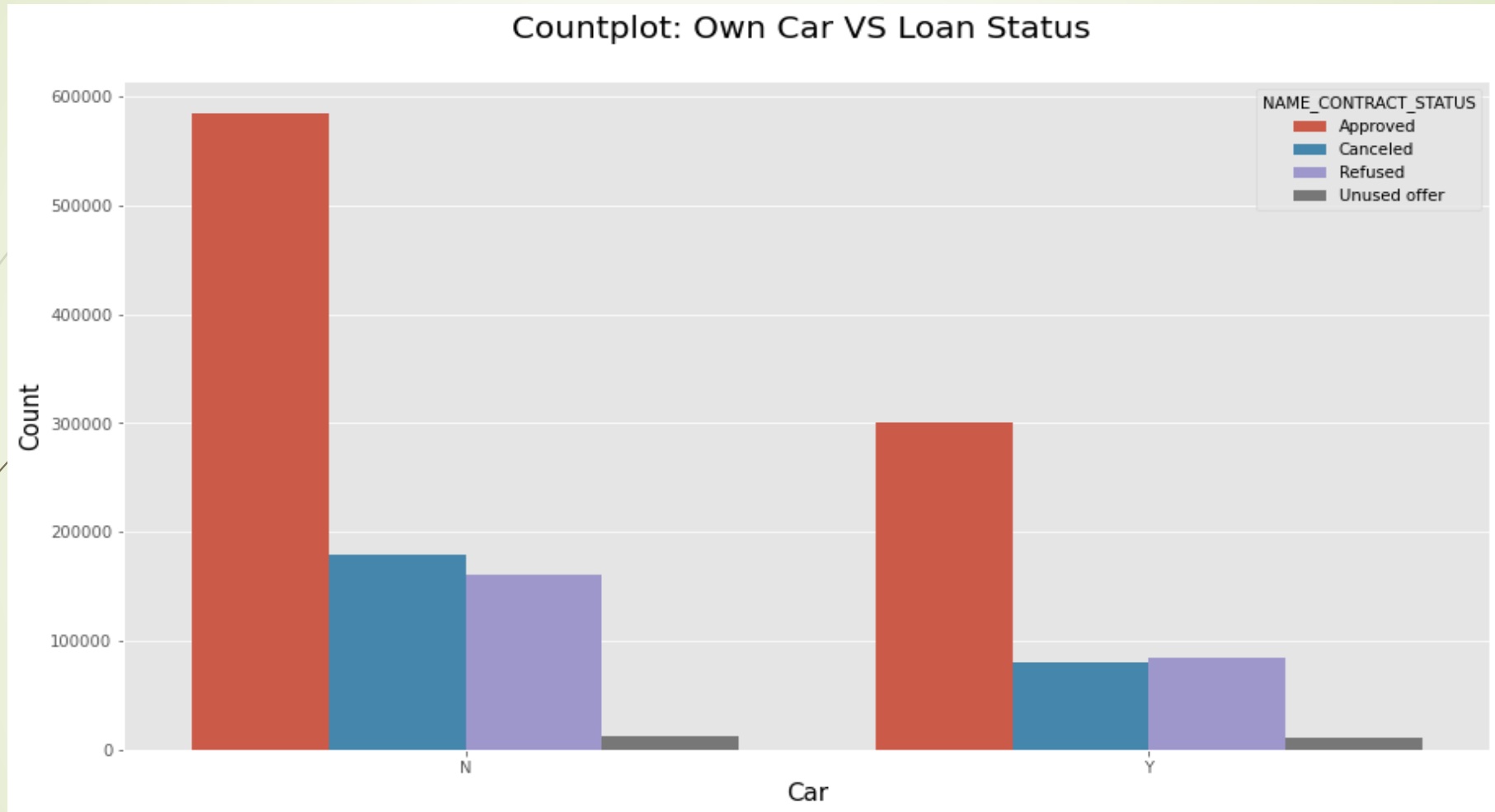
- Creating a new data frame by merging the application_data and previous_application data sets.
 - Merging the two data sets on SK_ID_CURR using left join on application_data data set so that, all the columns from application_data data set are retained.
 - This new data set is used for further analysis.
- 

Cash Loan Purpose distribution



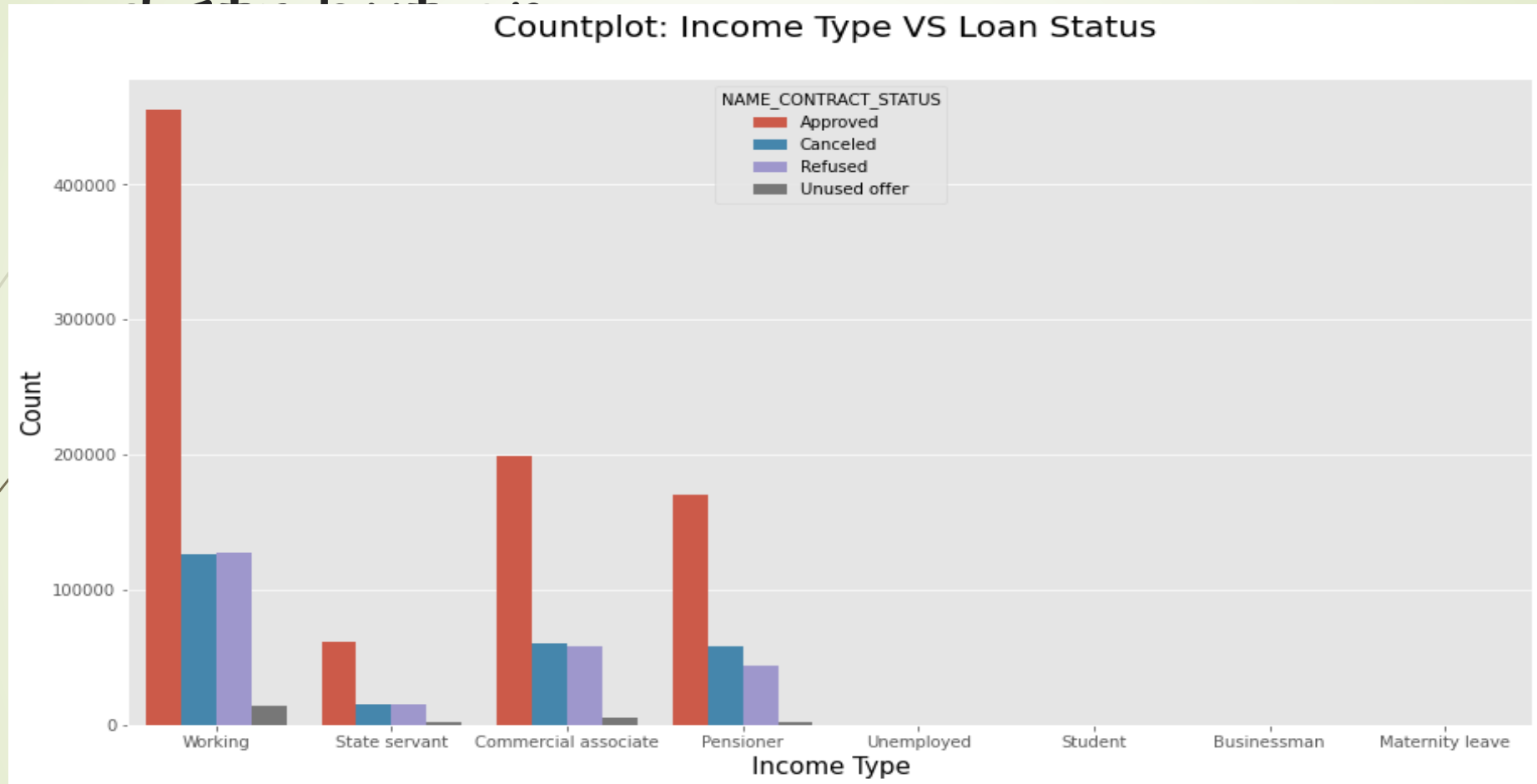
- Most of the cash loan applications are for Repairs purpose followed by Others and Urgent needs.
- In very few cash loan applications, clients have refused to disclose the purpose of taking loan.
- There are least number of applications for 'Money for a third person' and 'Hobby' purposes.

Own Car VS Loan Status distribution



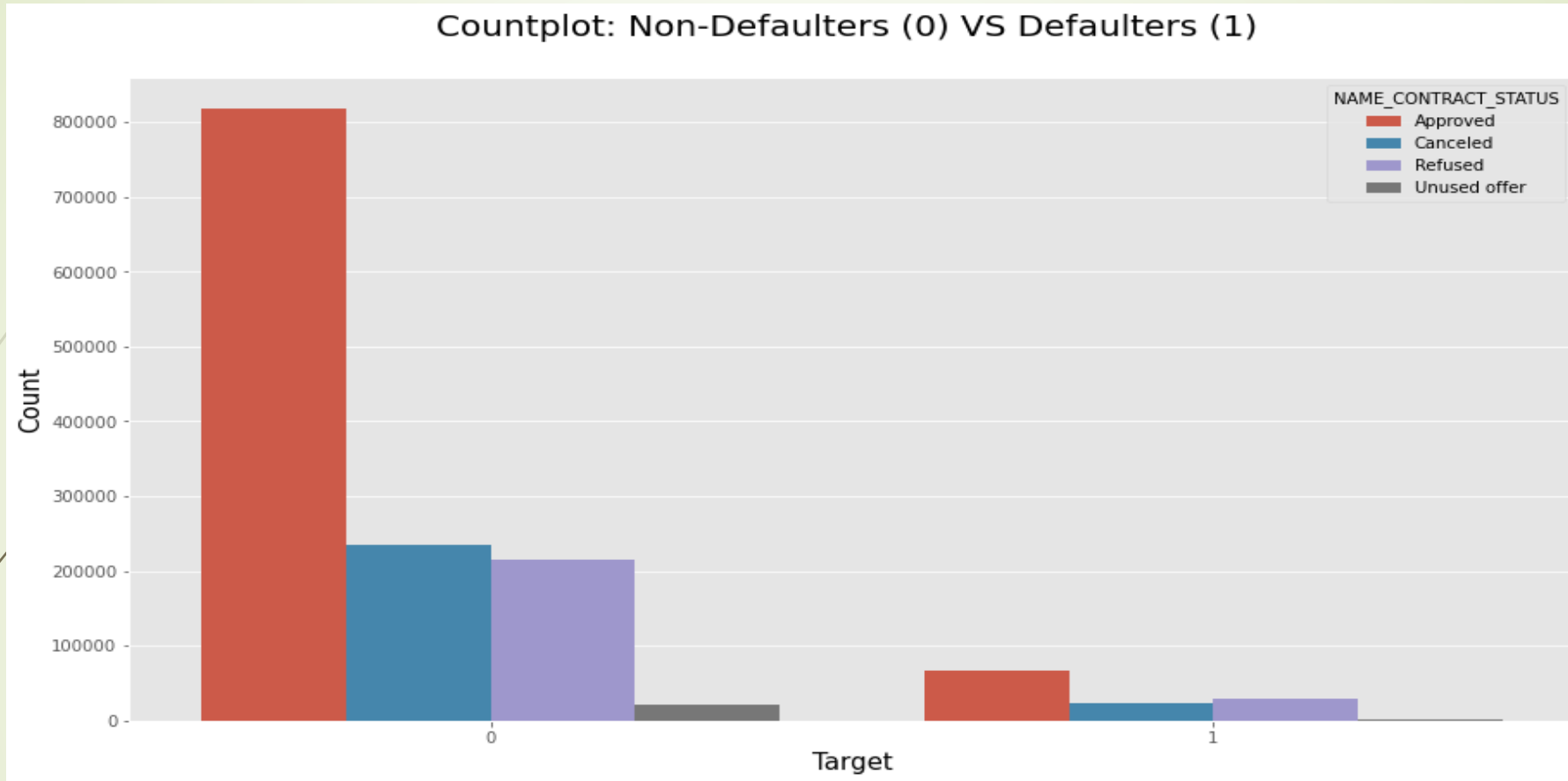
- There is not much difference in loan status if client owns a car or do not own a car.
- People not owning a car seems to apply for loans more than people owning a car.

Income Type VS Loan Status




- People from Working class seems to apply for loan the most followed by Commercial associate and Pensioner.
- There are very few loan applications from the people in Unemployed, Student, Businessman and Maternity leave class.

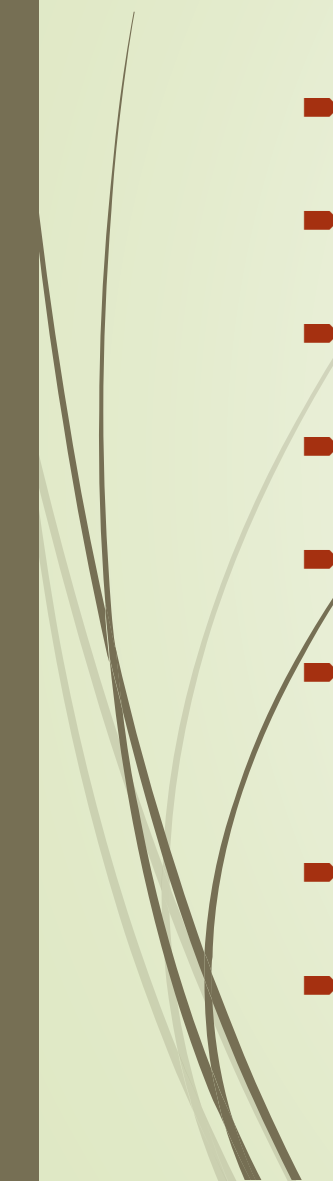
Target VS Loan Status distribution



- There are less number of defaulters.
- There are quite a few loan applications from defaulters which are approved resulting in loss for the company.
- There are also high number of loan applications from non-defaulters which are refused and cancelled resulting in loss of business for the company.



Conclusions from the EDA and Suggestions

- There are 8.07% of defaulters.
 - There are more loan applications from female clients.
 - Clients with Secondary/secondary special education type are more likely to default.
 - Single and not married clients are more likely to default.
 - Married clients are safe to give the loan as they are more likely to repay on time.
 - Younger clients (< 45 years) are more likely to default. Clients greater than 45 years of age are more likely to repay on time.
 - Clients who's previous loan was rejected/cancelled are more likely to default.
 - There are also high number of loan applications from non-defaulters which are refused and cancelled resulting in loss of business for the company.
- 



Thank You...