CREDIT EDA CASE STUDY

ARUSHI SHREE

ABHISHEK PORWAL

Objective:

To perform EDA analysis over the data collected by a bank over its customers and identified whether a customer associated with the bank who had taken a loan have fallen into the list of defaulters or not.

In the data, bank have collected information based on various parameters which could help in analysing that if a customer will tend to fall under the defaulter and furthermore the modelling based on this analysis will help the bank company to predict the likelihood that customer will be a defaulter.

The EDA is carried out by following the classic procedural steps as below:

- Data understanding
- Data cleaning
- Data manipulation
- Analysis of Data (univariate/bivariate/multivariate)
- Governing factors.

Assumptions/Approach for Data cleaning.

- Identifying the null values in data set.
- Imputation of Null values :

  - Columns having null values greater than 50% are dropped.
  - Columns having null values percentage between 0< <13% are taken and best values are identified which can be imputed in place of null values.
  - Null values in categorical columns are treated by taking MODE of column values.
  - Null values in a numerical(continuous) type column are handled by two ways depending upon presence of outliers in them. i.e. If outliers are present then MEDIAN value is taken else MEAN value.

Analysis overview:

Univariate analysis

- Distribution plots

- Count plots

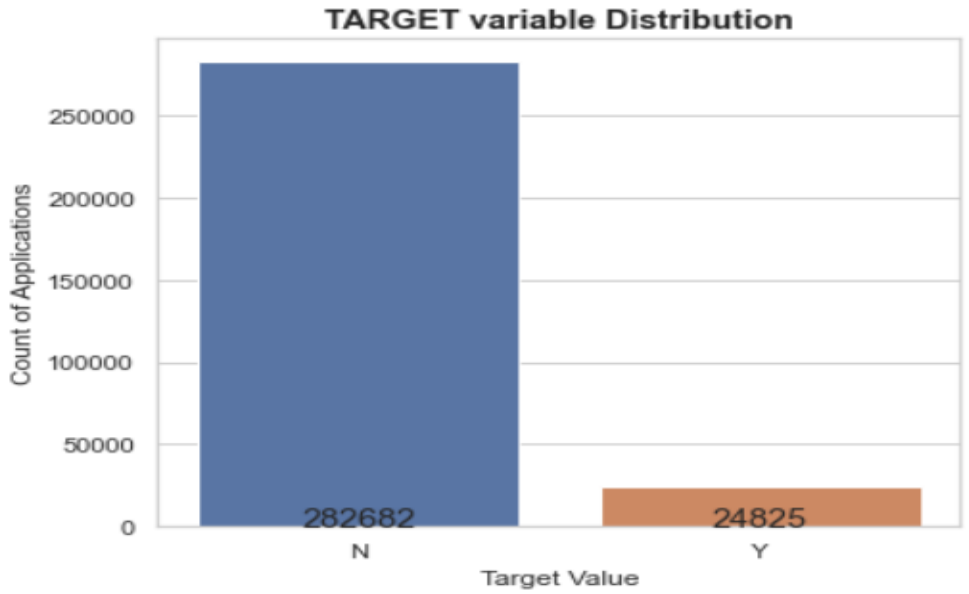
Bivariate analysis

- Categorical vs Categorical

- Categorical vs Numerical

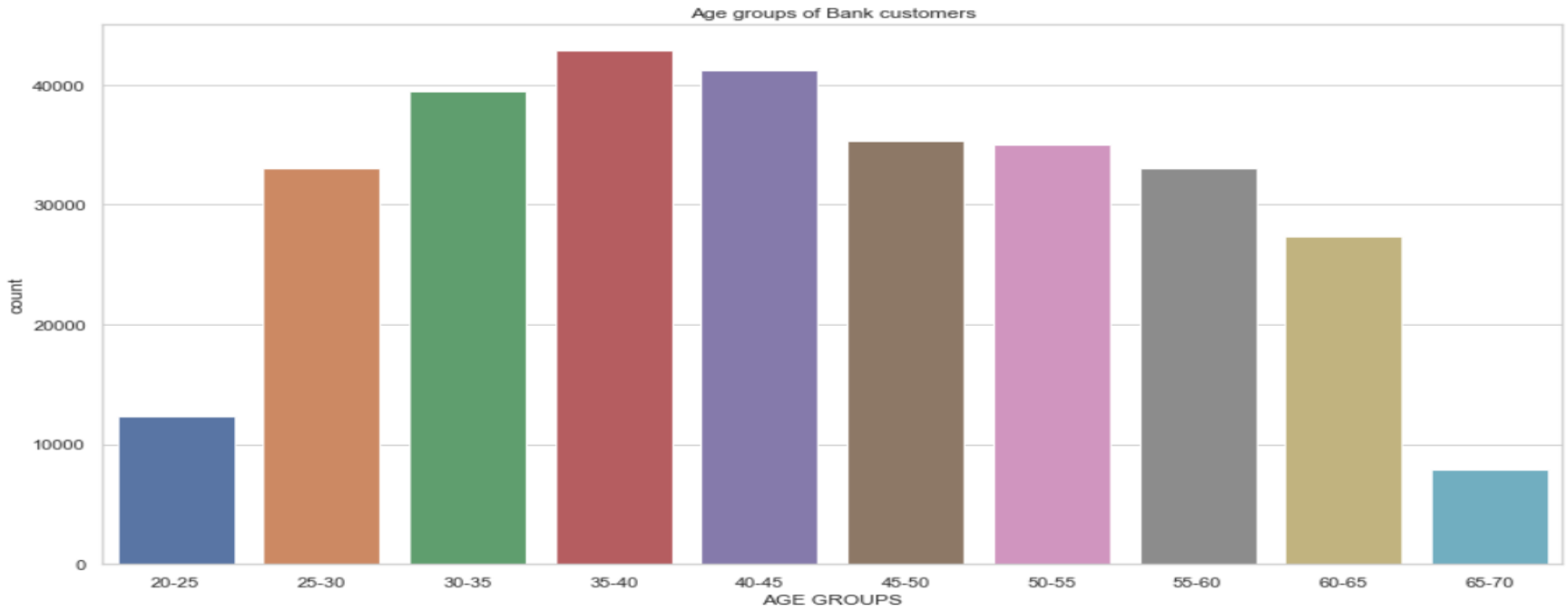
- Numerical vs Numerical

Distribution of target variables



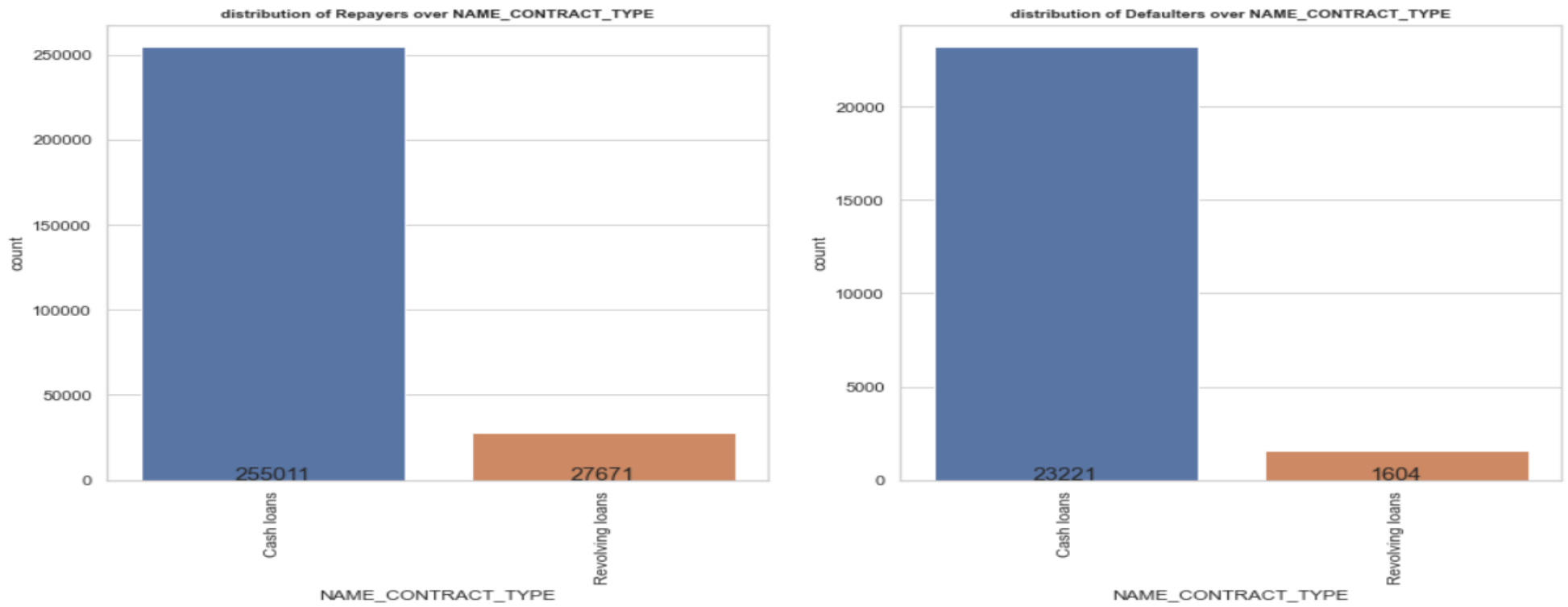
We can infer that 282686 applicants are without payment difficulties and only around 24825 are with payment difficulties.

Distribution of client across age groups



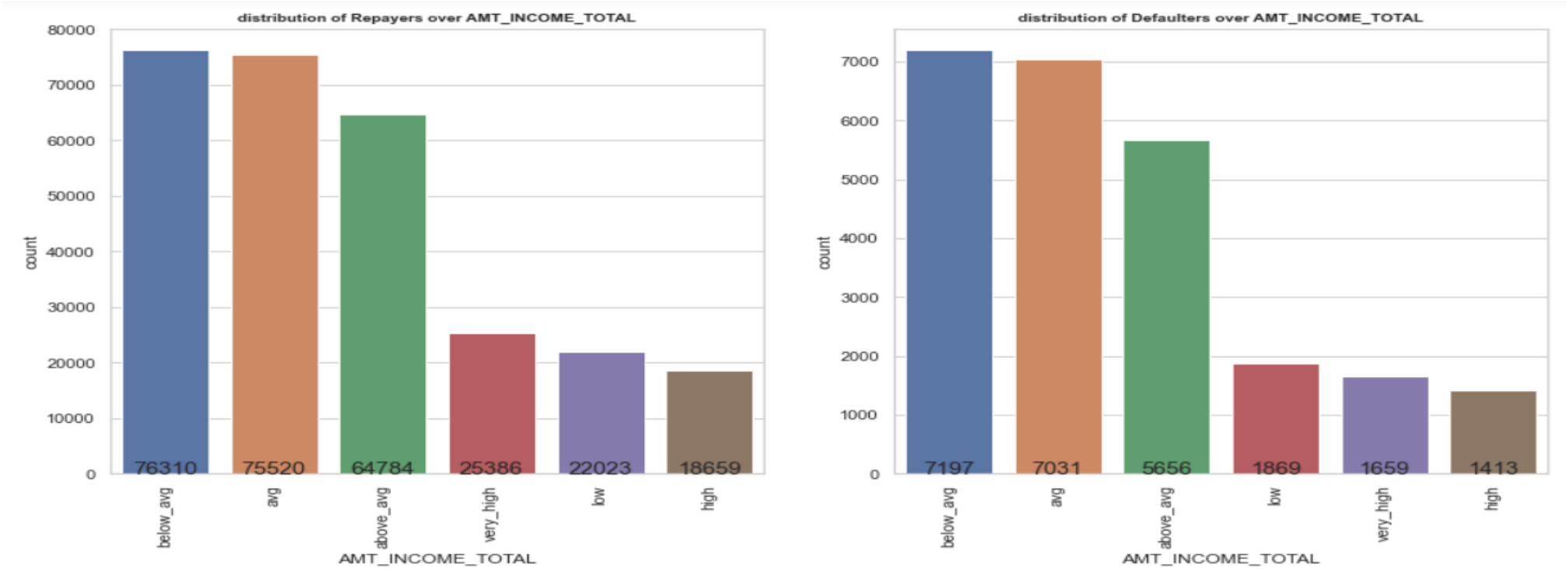
- We can see that majority of the clients are in range of 25 to 65.
- The very young (<25) and very old (>65) clients count is very less compared to other age groups.

Distribution of client over Name contract type



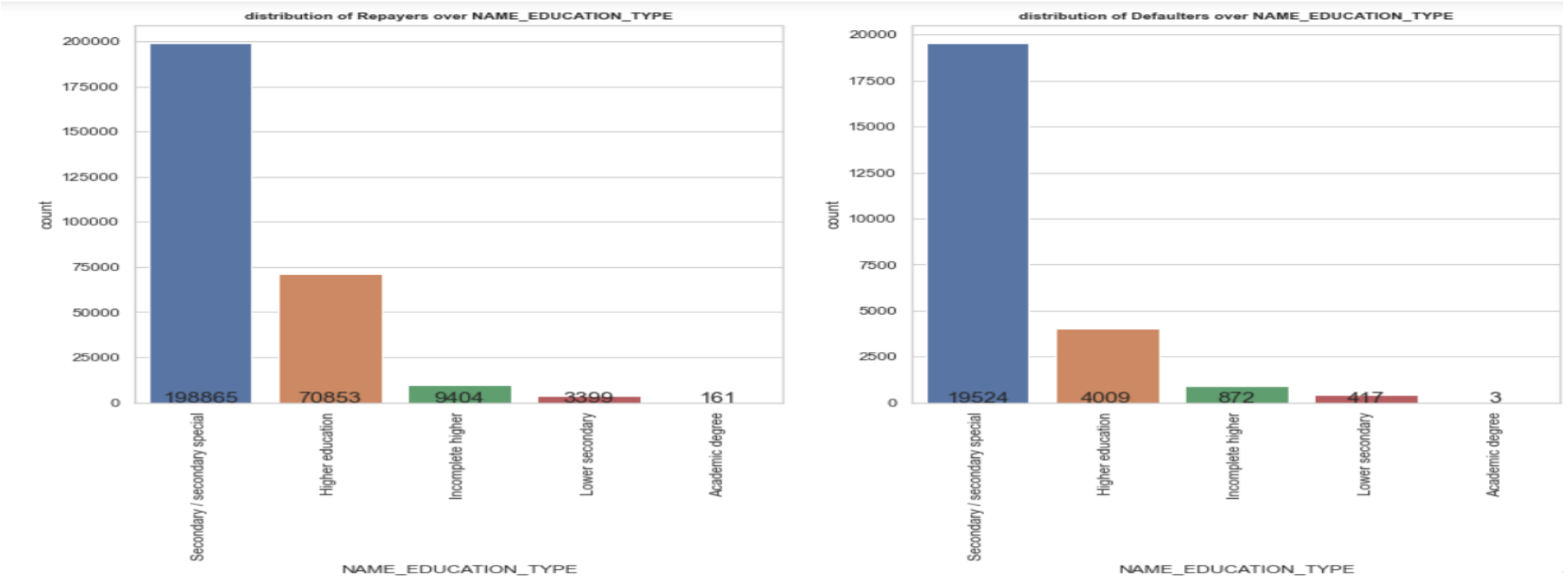
- Most of the clients opt for cash loan type instead on revolving loan type in both Repayer as well as defaulter case .

Distribution of client across income groups



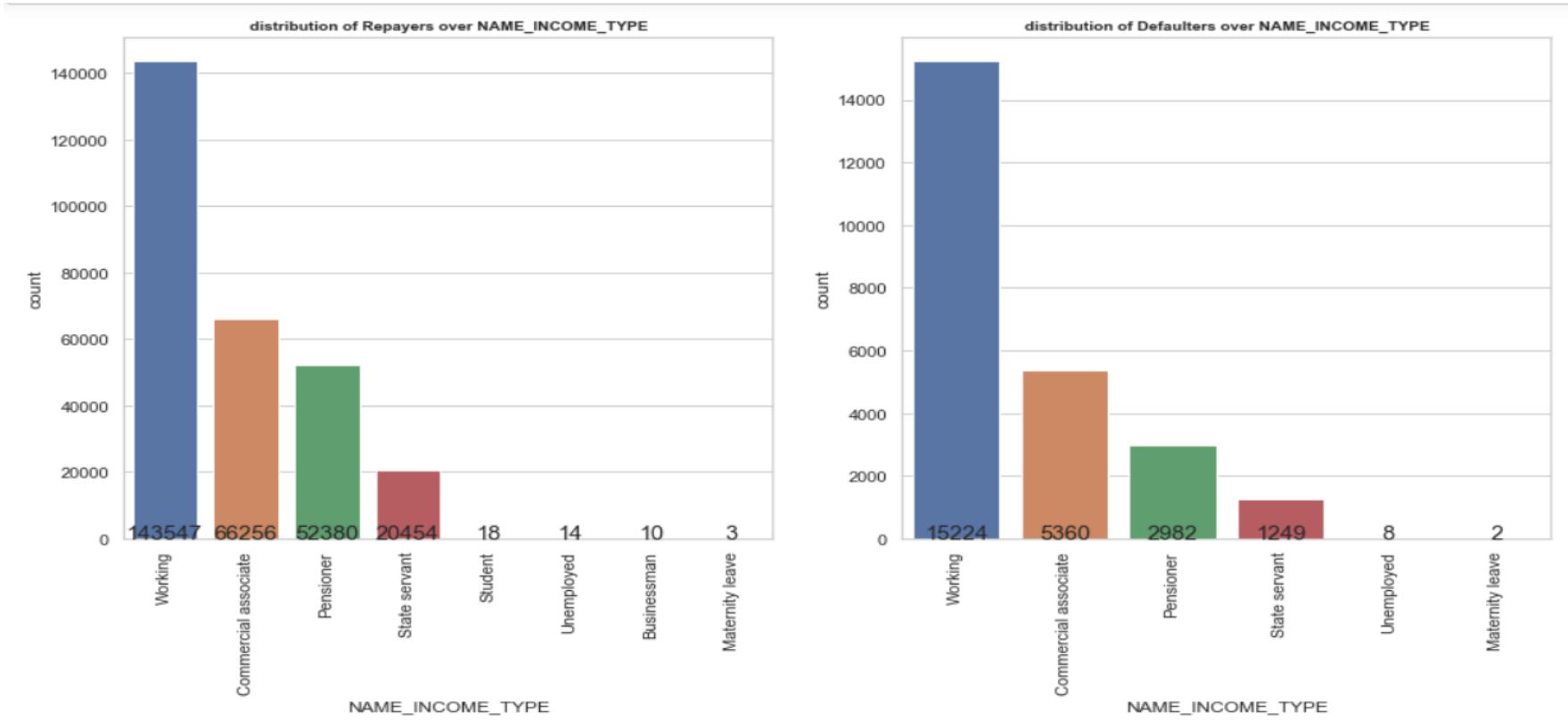
- We can infer that majority of the clients fall within the income category below_avg,
- While minority of clients fall within the income category high.

Distribution of client across Education type



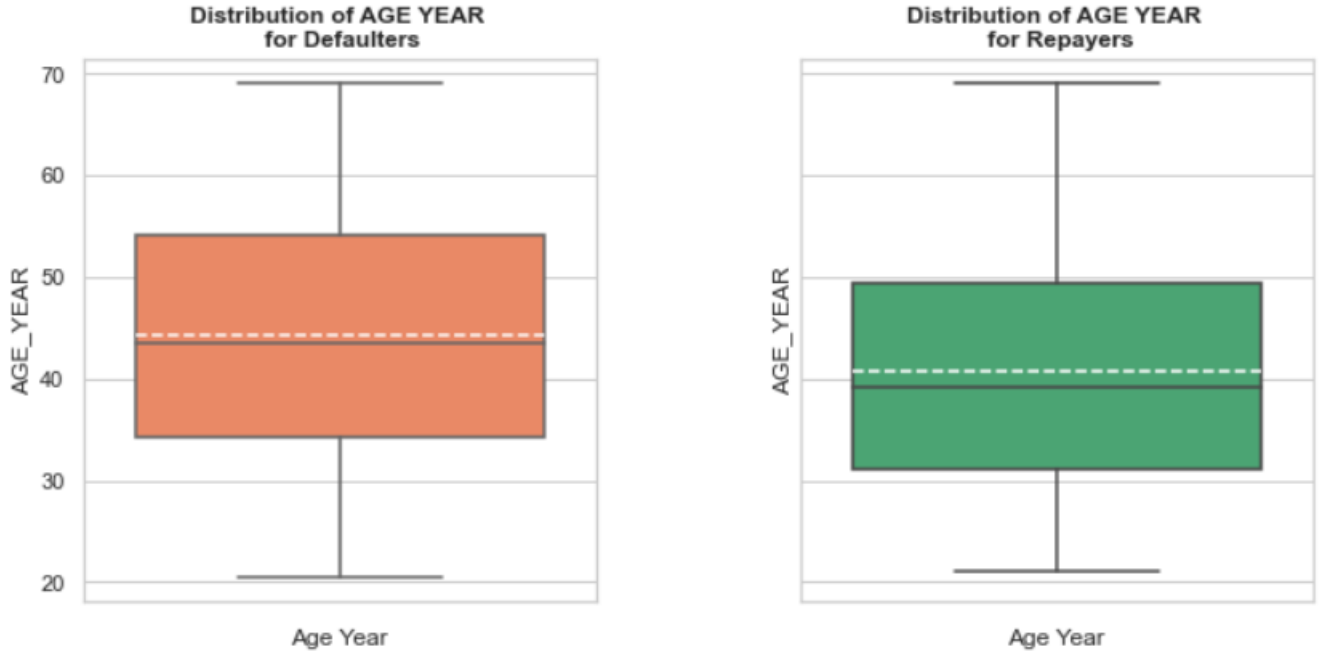
- We can infer that majority of the clients fall within the Education type Secondary/secondary special,
- While minority of clients fall within the Education type Academic degree.

Distribution of client across income type



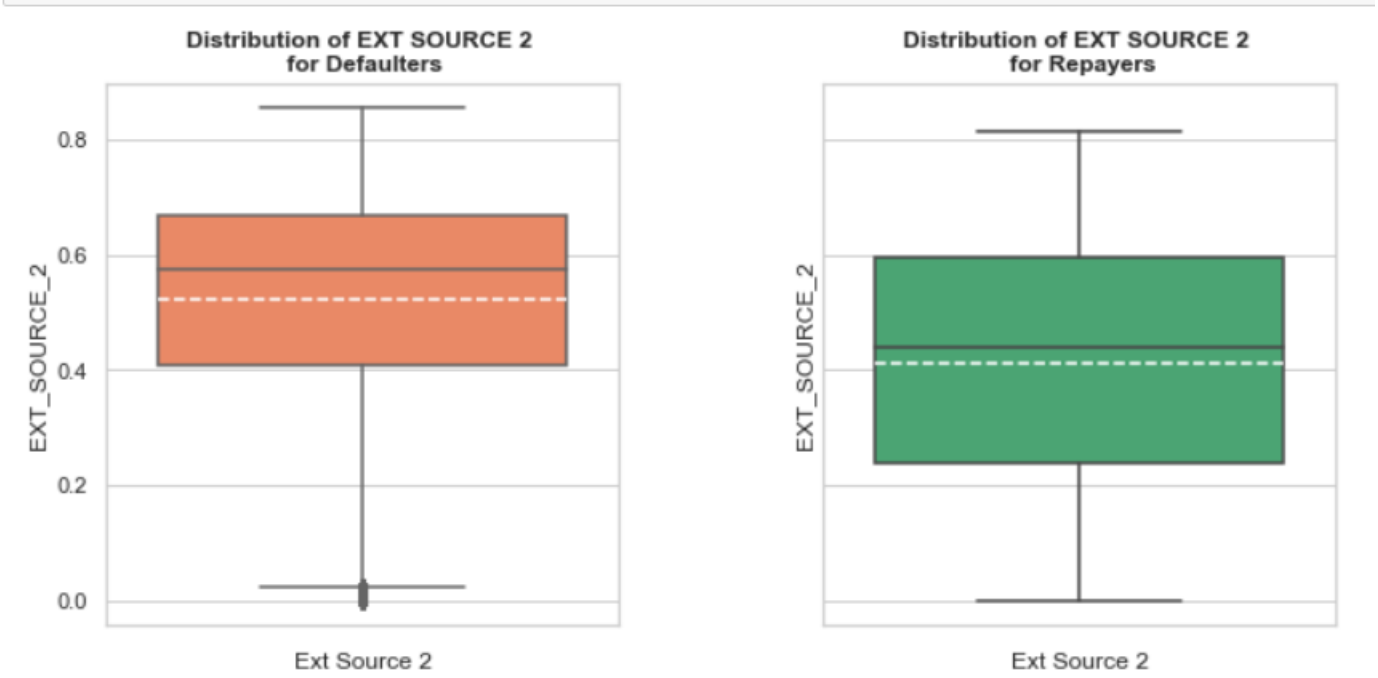
- We can infer that none of businessmen nor students are present in defaulters data

Distribution of age groups across target



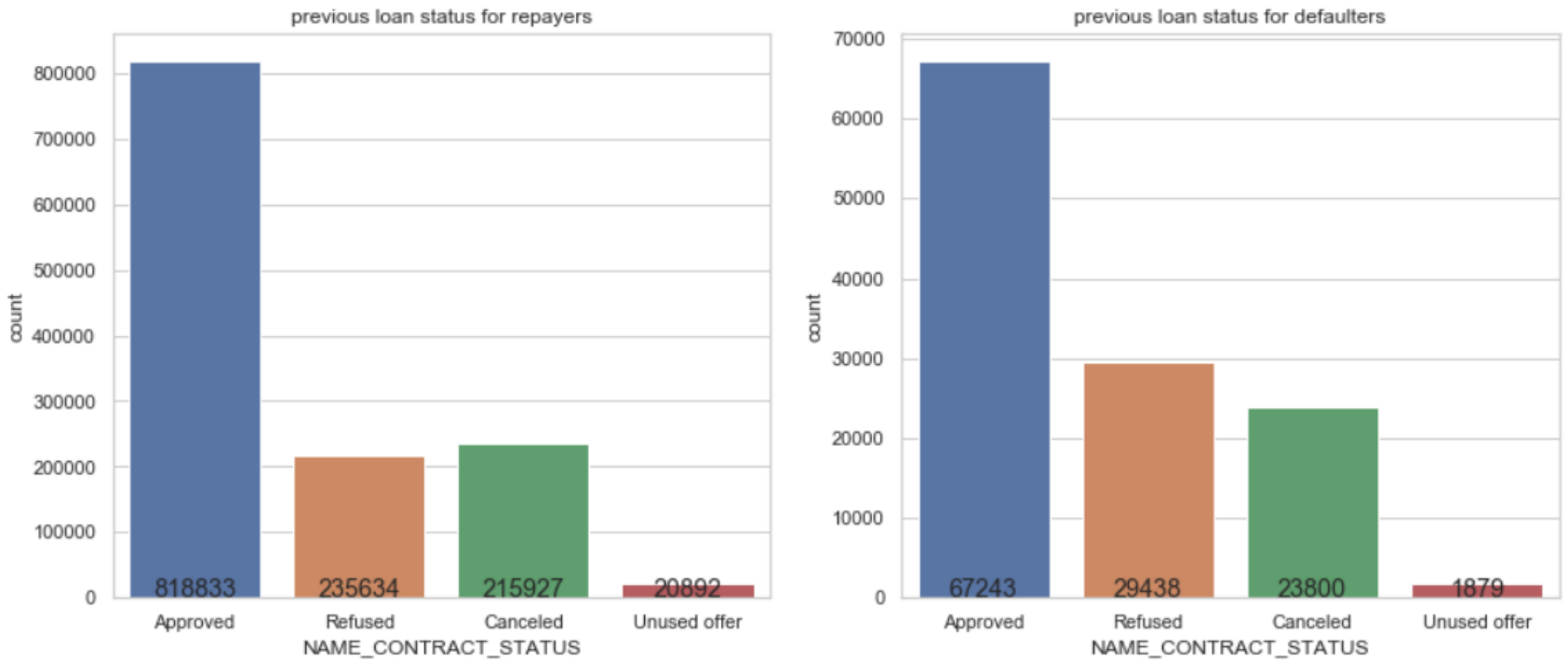
- Average age of repayers is slightly less than non-defaulters.
- Also for defaulters majority of age is distributed approx. between 35-55.

External source 2 vs. Target



- Average value for external source:2 is higher in Repayer case as compared to defaulters case.
- Majority of Ext. source 2 value falls between 0.2 to 0.6 for defaulters, while for Repayer it falls 0.4 to 0.7

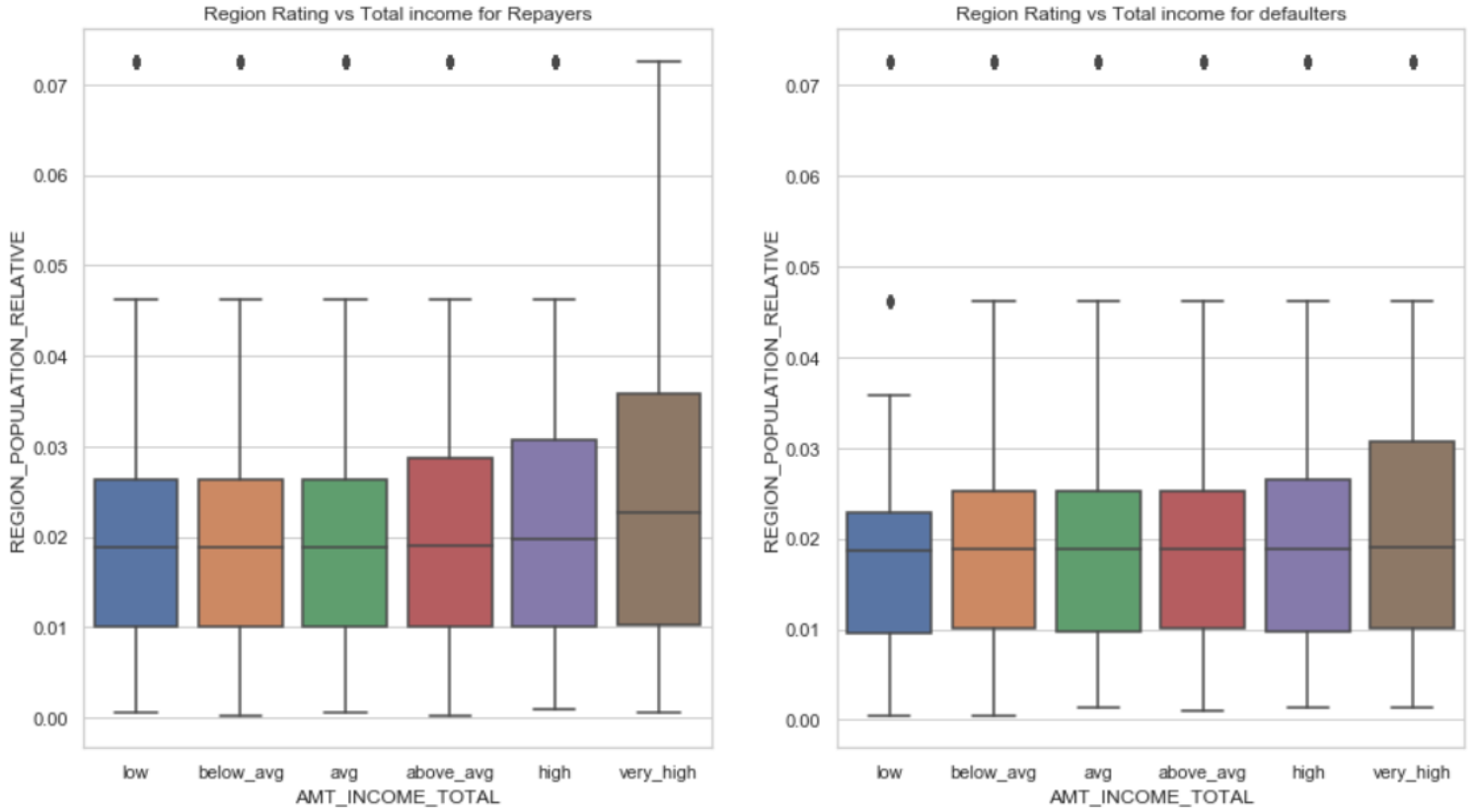
Analysis based on previous loan status:



An interesting insight can be drawn from above plot is that:

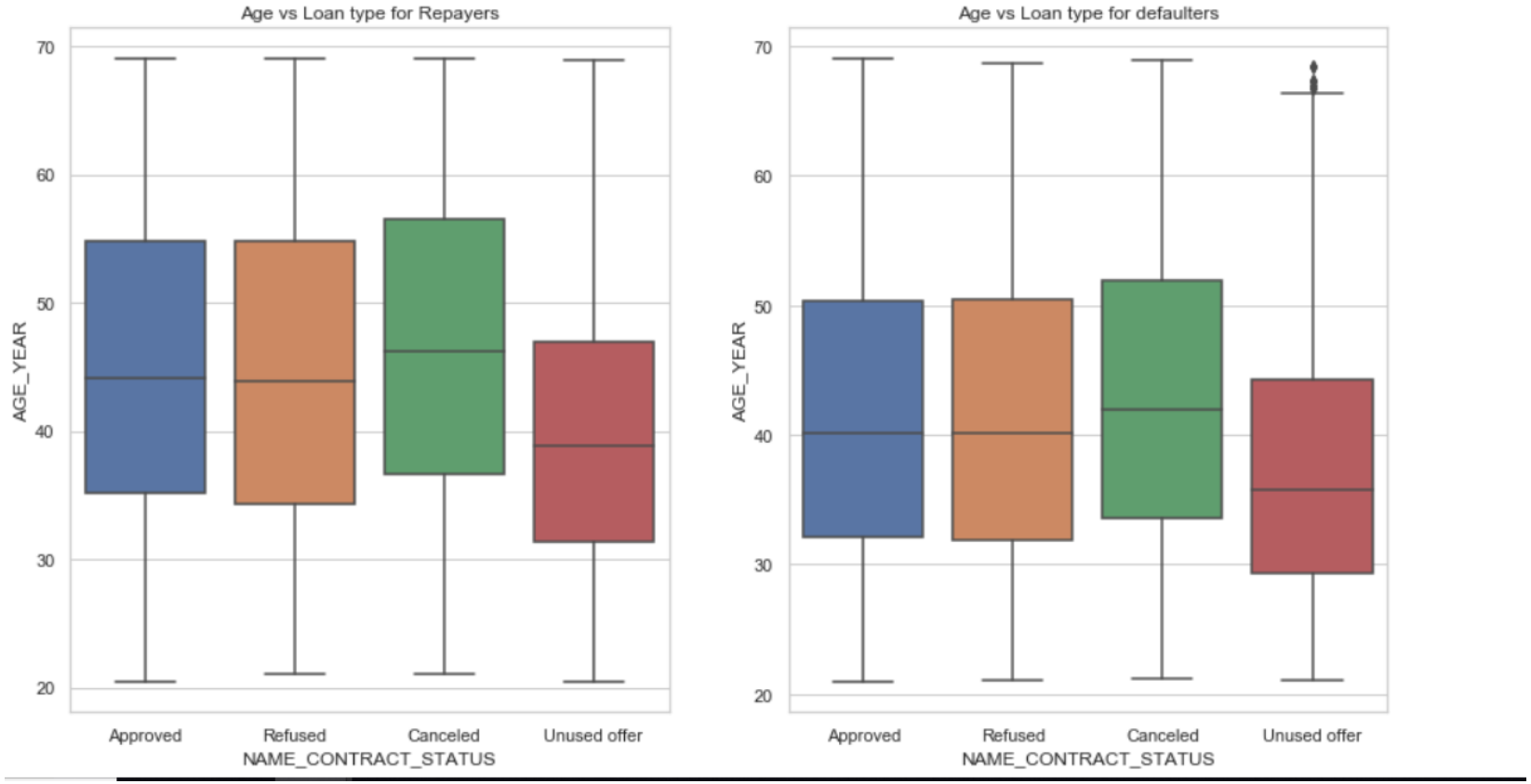
- count of approved application for current repayers is more as compared to defaulters.
- previous application was comparetively highly refused for current repayers than current defaulters.

Region rating vs Total income under both target segments:



From above distribution we can se that although most customers fall under high region rating but distribution is more vastly distributed for Repayers rather than defaulters.

Age vs Contract Status for both target Segments:



Customers who repaid the loan and have their previous application approved are more distributed in the upper age group range wherin defaulters having their loan contract approved in the lower age -group.

Inferences

Variable which are strong indictors of default :

- AMT_INCOME_TOTAL
- AMT_ANNUITY & AMT_CREDIT
- **DAYS_BIRTH (YEARS_BIRTH) –**

  Selecting clients with middle age group will reduce the risk of default
- **OCCUPATION_TYPE –**

  Selecting clients with higher occupation type will reduce the risk of default
- **REGION_RATING_CLIENT** –

  Selecting clients with highest region rating provided by bank itself will reduce risk of default

Thank You-