CREDIT EDA CASE STUDY

By- RAHUL PATHAK

PROBLEM STATEMENT-I

INTRODUCTION

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

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 to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Data Understanding

The dataset has 3 files as explained below:

1.'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

2.'previous_application.csv' contains information about the client’s previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

3.'columns_description.csv' is data dictionary which describes the meaning of the variables.

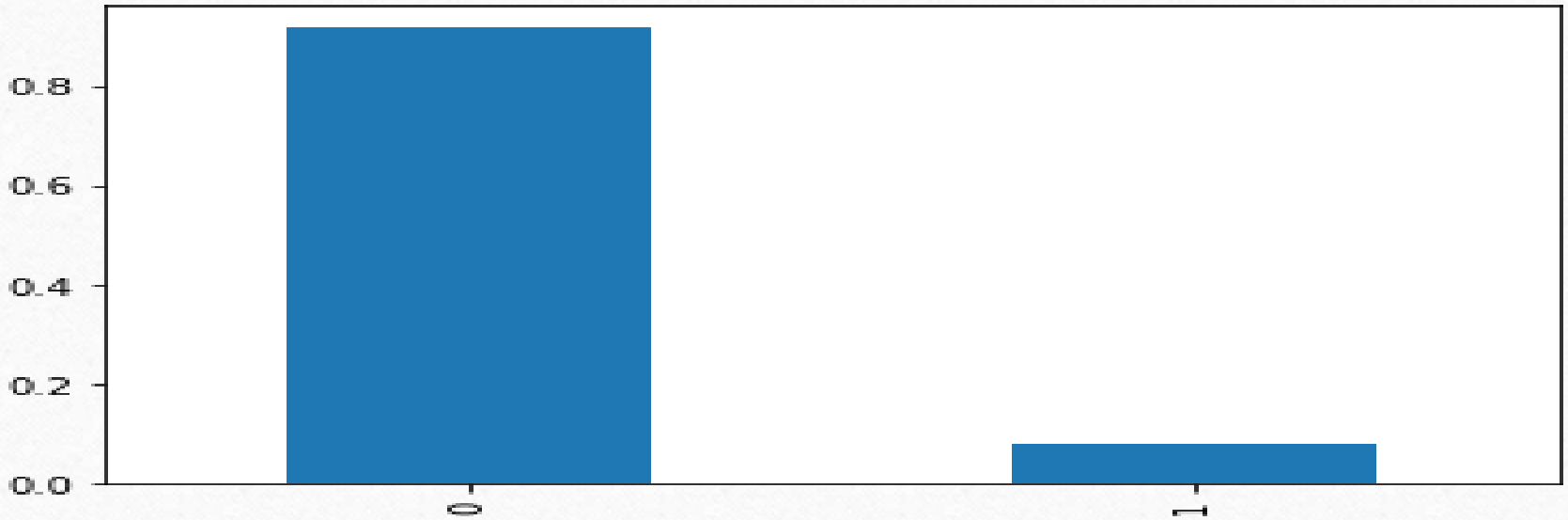
Process Flow:

I have taken our "application_data.csv" dataset into the account first:

1. I have cleaned the "application_data.csv" dataset.
2. I have dropped the columns having null value percentage more than 40%.
3. I have performed univariate, bivariate and segmented univariate analysis on all the important columns and try to find out how much they are influencing our target variable.

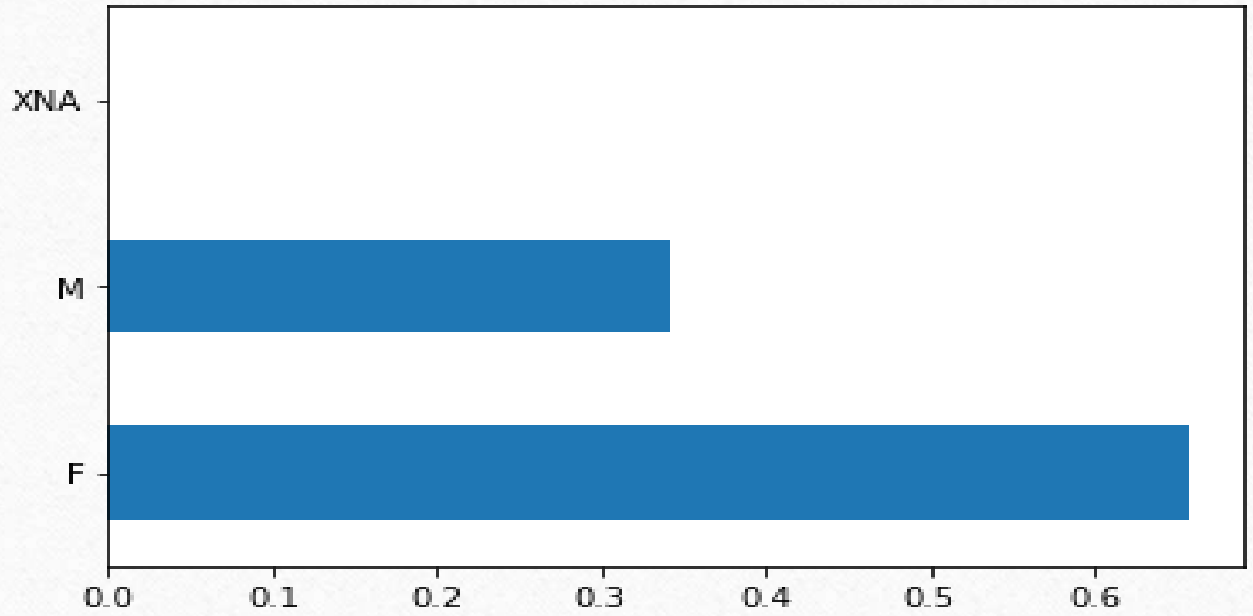
Note: The target variable is our "TARGET" column in the "application_data.csv" dataset which tells if the loans are approved or not. Also, we will make sure Target variable has binary entry only. 1 for Those who are defaulters and 0 for those who are not the defaulters.

Categorical Univariate analysis for TARGET column



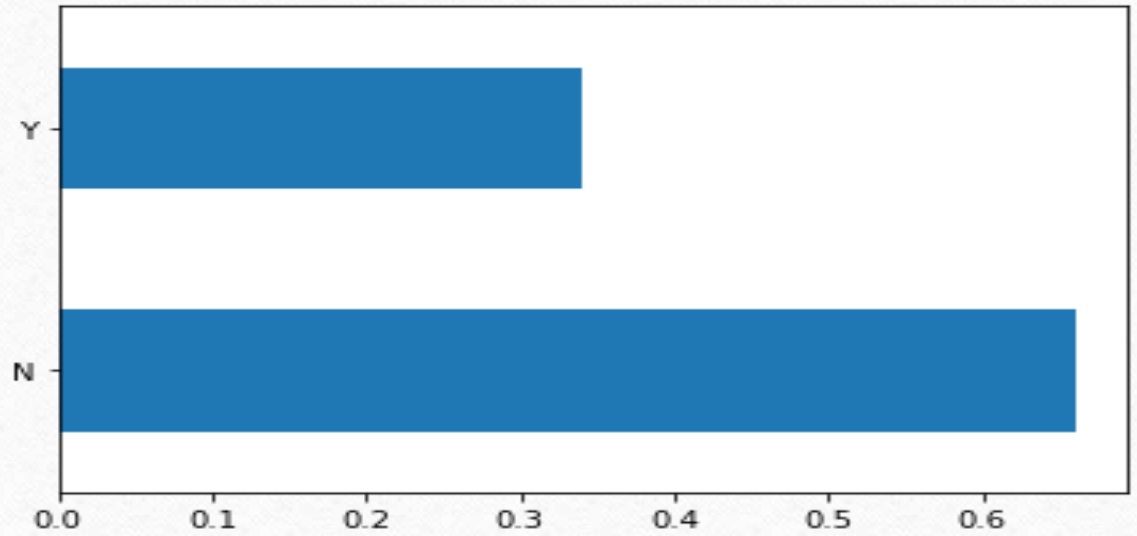
•The data suggests that almost 8% of the people are defaulters amongst all the applicants.

Lets have a look over the bar plots for each gender type who have applied for the loan.



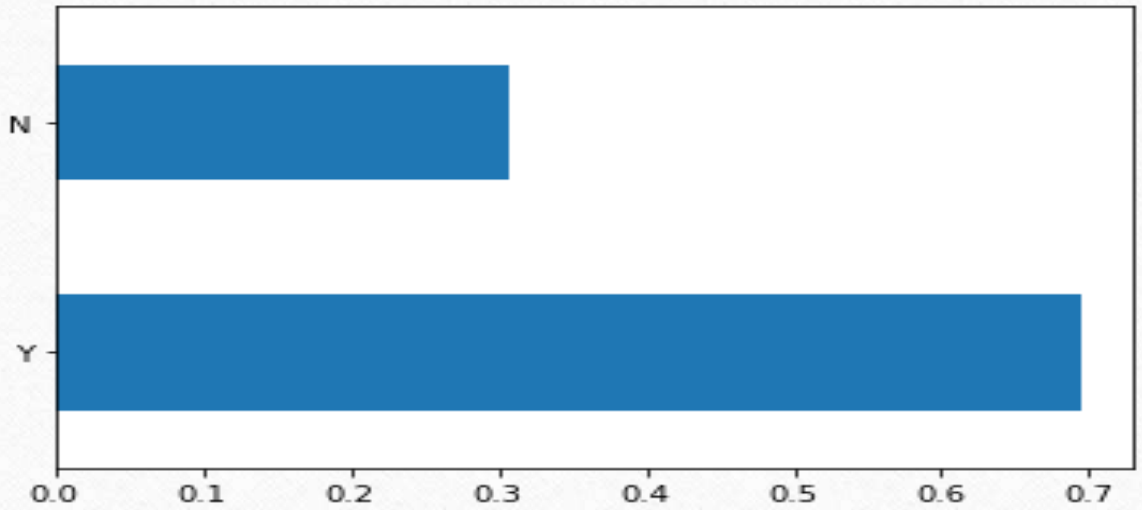
• This concludes female population is significantly high in terms of loan application.

Lets have a look over the bar plots to check if the applicants own a car or not.



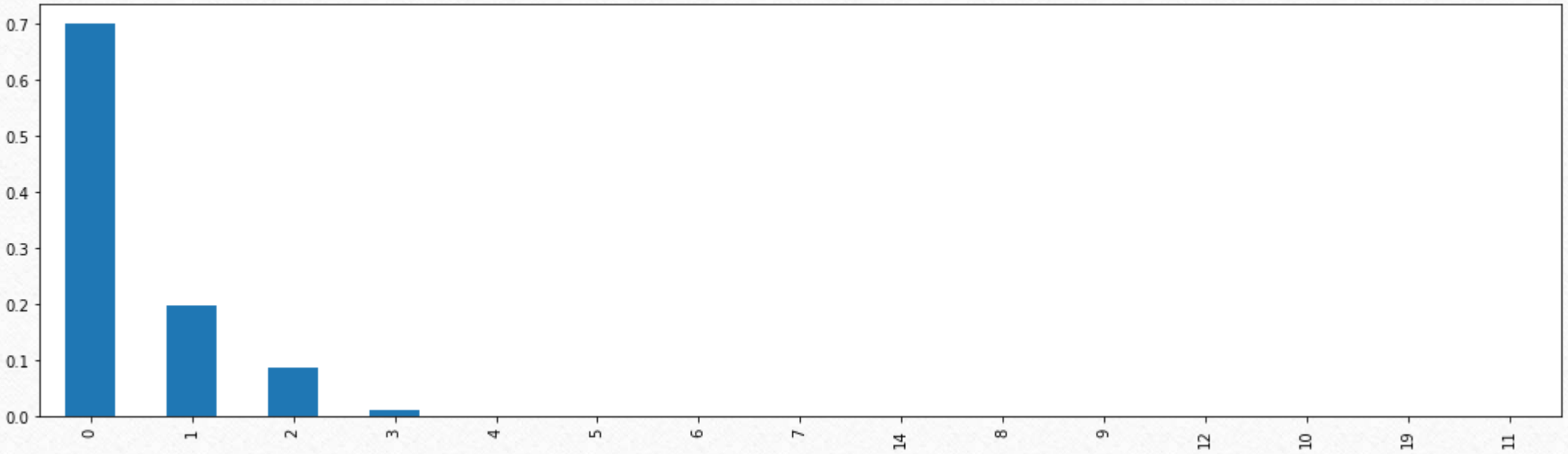
•This concludes the population which doesn't own a car is significantly high in terms of loan application.

Lets have a look over the bar plots to check if the applicants own a real estate property or not.



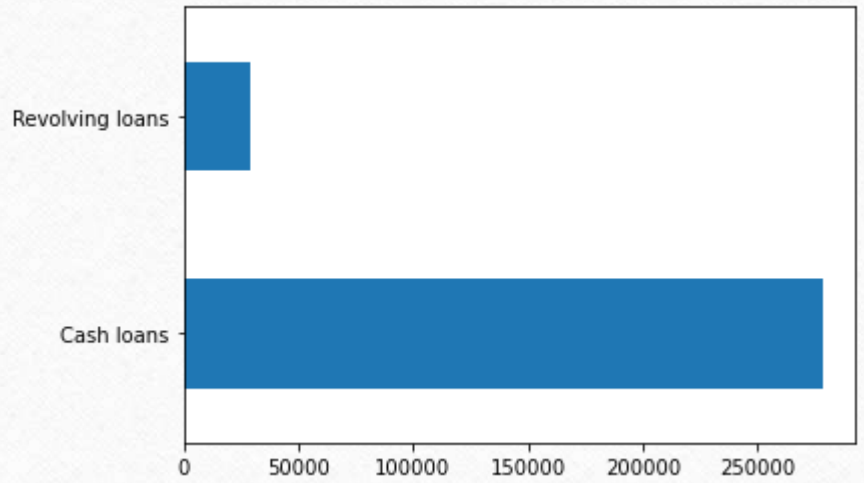
•This concludes the population which owns a estate property is significantly high in terms of loan application.

Lets have a look over the bar plots to check the count of children for the applicants.



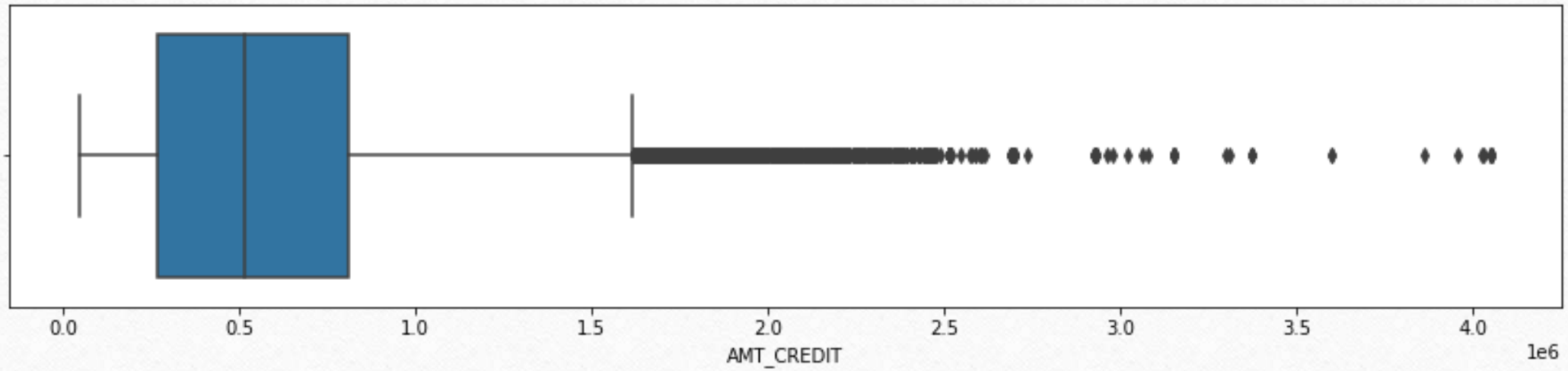
•This concludes: Amongst the applicants, almost 70% of the population doesn't have any child and almost 20% have single child.

Lets have a look over the bar plots to check different means of loans for the applicants.



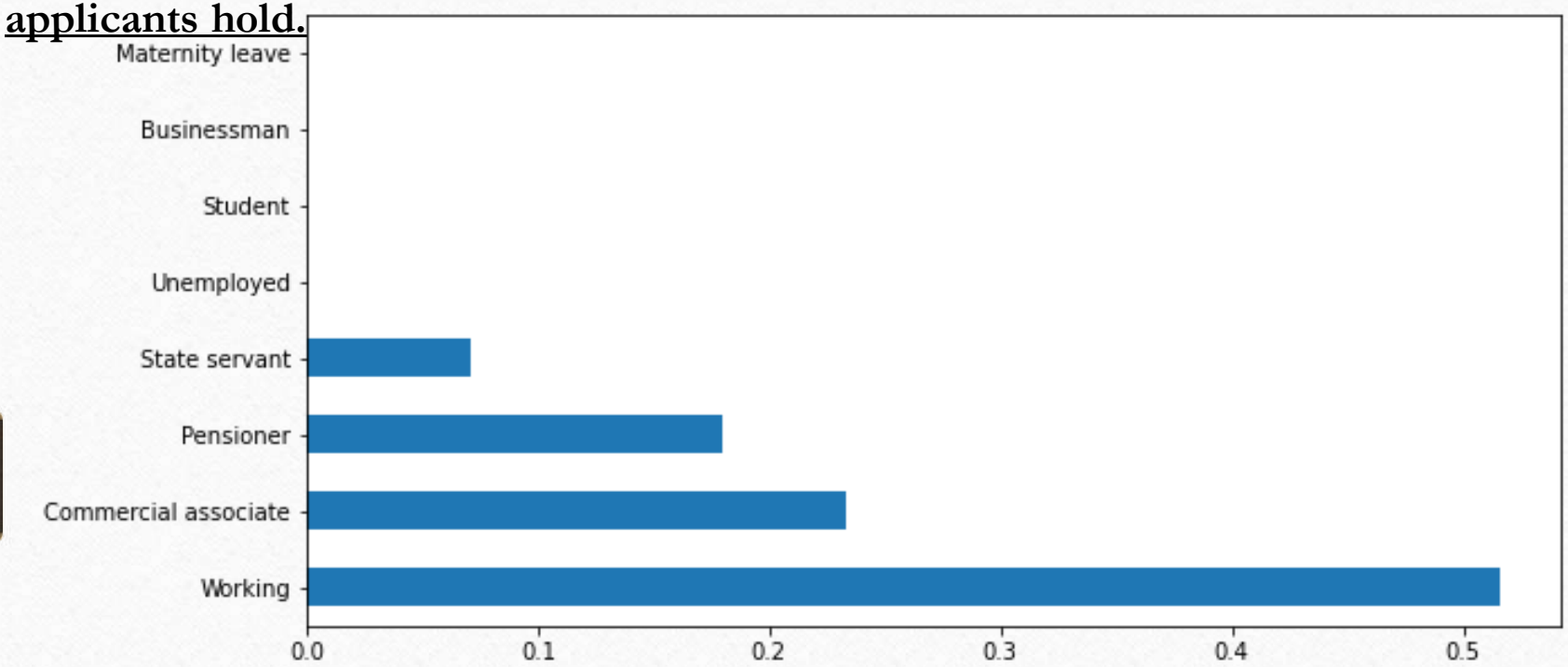
•People taking cash loans seems to apply more as per the plotted bar above.

Lets have a look over the box plot to check the amount of loan applicants are approved for.



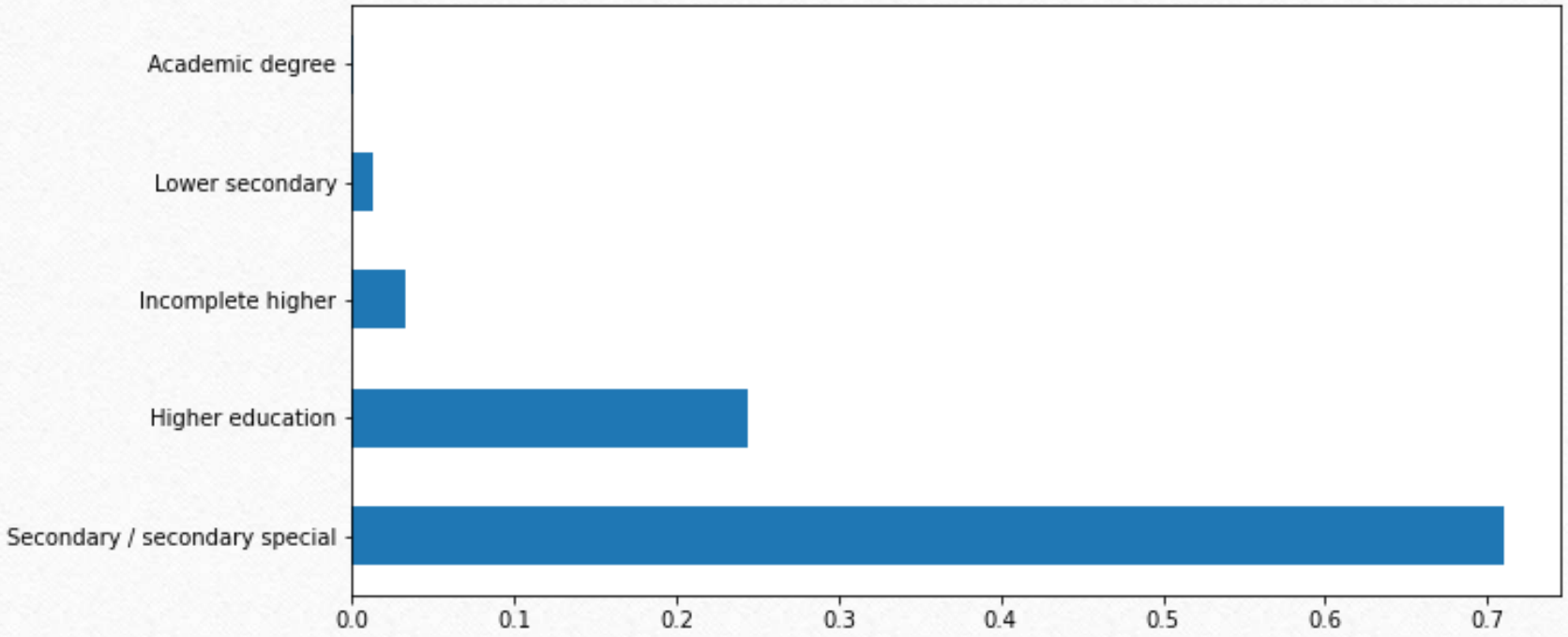
•This reflects, majority of the people or say (50%) of the population who apply for loan are getting approved for the loan amount between 2.5L to 5.0L.

Lets have a look over the bar plots to check the types of income the applicants hold.



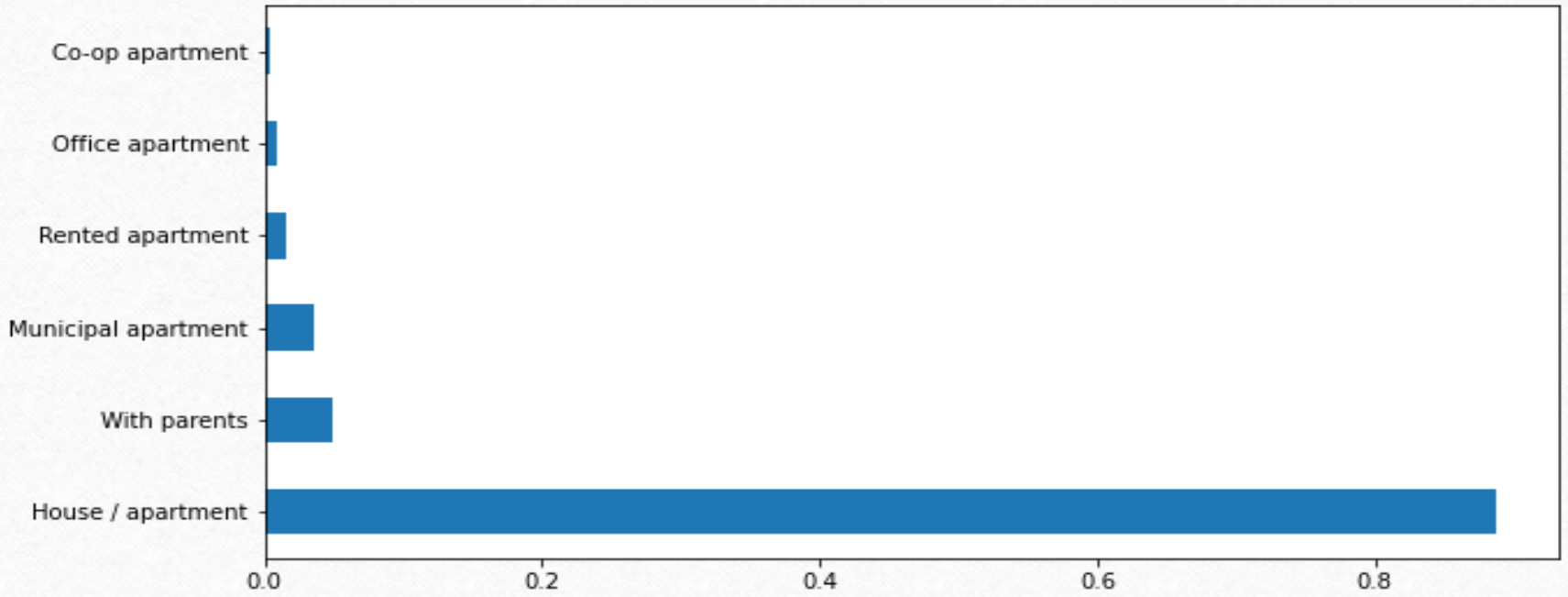
•We can clearly see which income category has the most number of applicants and come to the conclusion that people who have a source of income are most likely to apply for a loan. However, the set of people who probably have the best source of income i.e. Businessmen are least likely to apply for a loan.

Lets have a look over the bar plots to check the types of education the applicants possess.



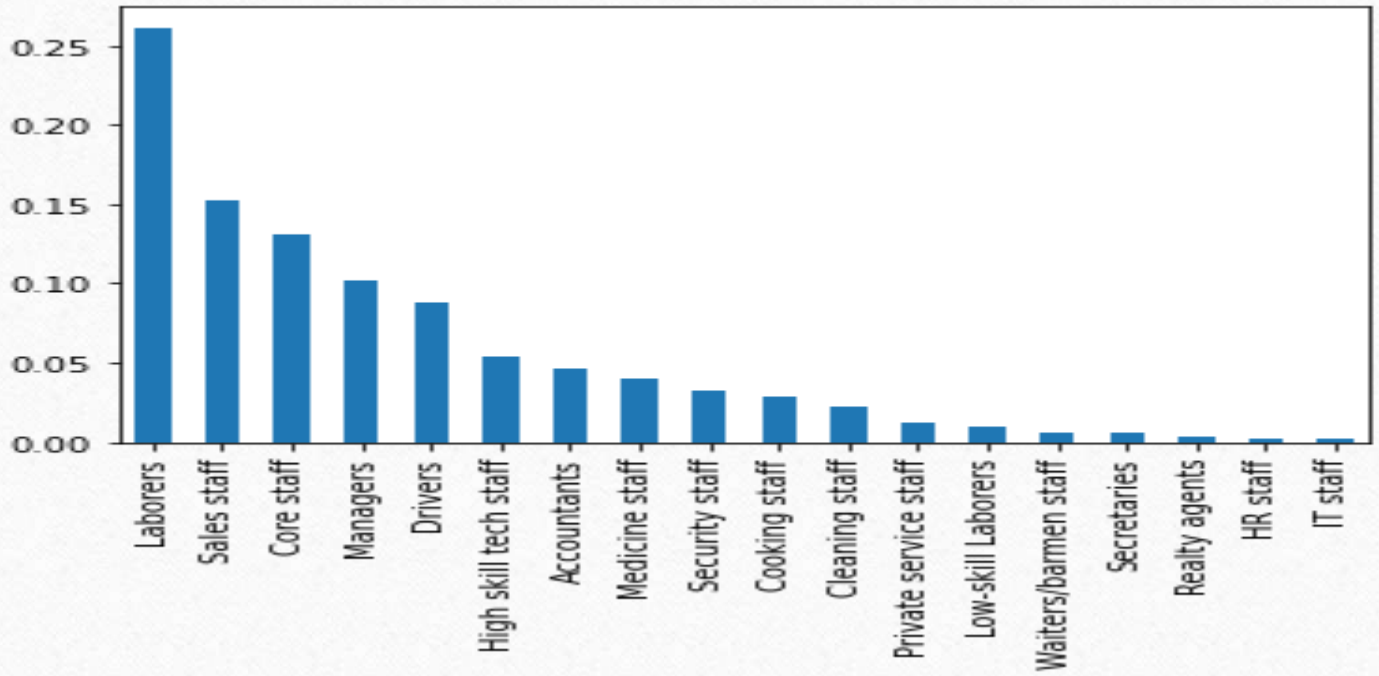
•This concludes: more educated people are most likely to apply for a loan.

Lets have a look over the bar plots to check the housing types for the applicants.



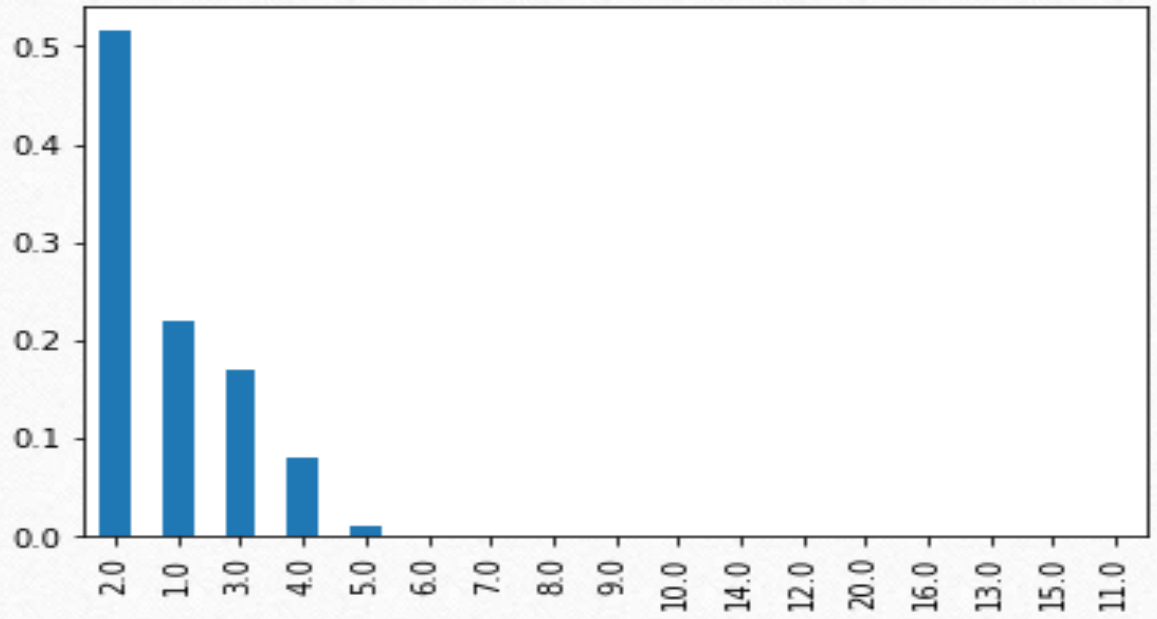
- Here we may conclude that people having their own houses and apartments are most likely to apply for a loan.
- We may also have a look over the categories as well and figure out how likely they are to apply for a loan.

Lets have a look over the bar plots to check the occupation types for the applicants.



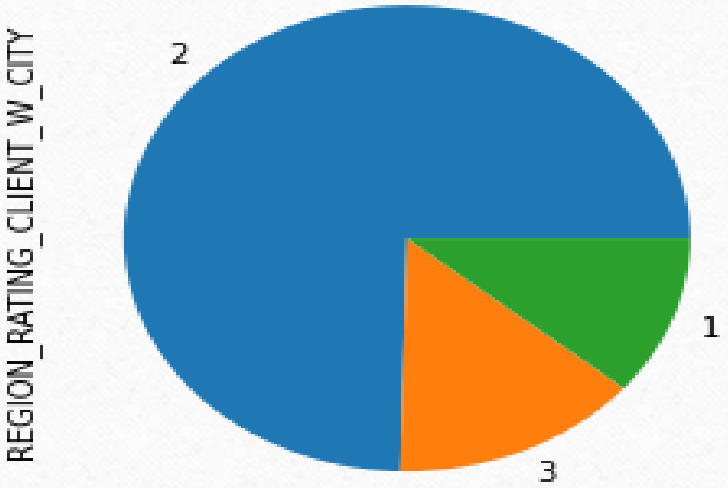
•We can have a detailed look here and figure out that people from which occupation goes for the loan more often.

Lets have a look over the bar plots to check the count of family members for most of the applicants.



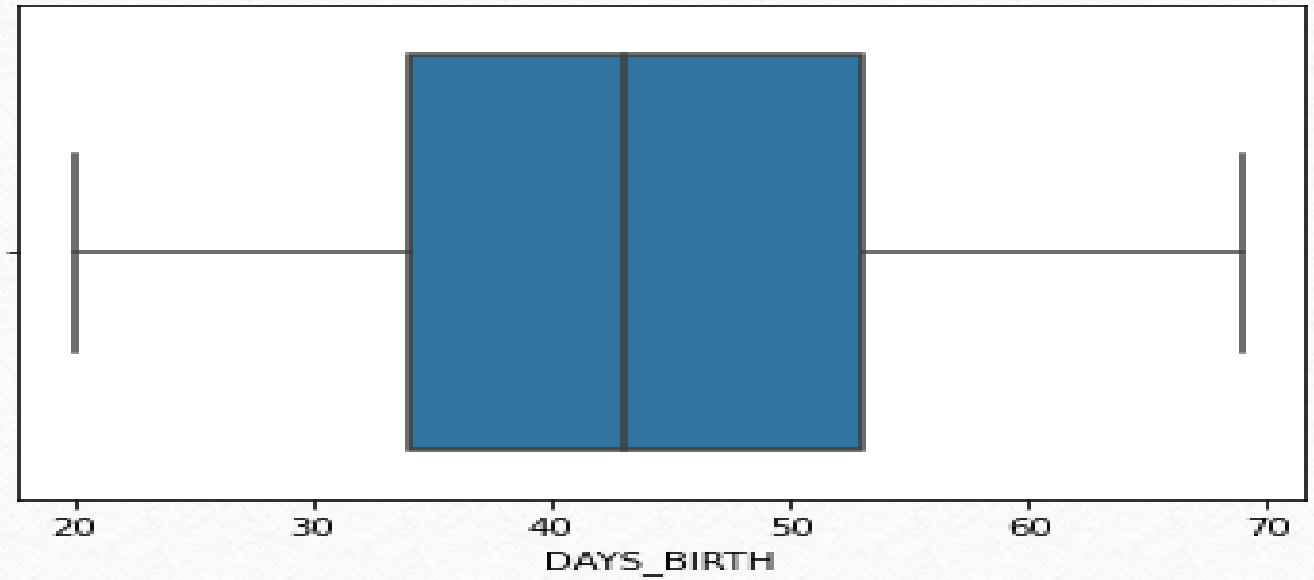
- A family of two is most likely to apply for the loan.
- Also, the trend suggests, higher the count of family members is indirectly proportion to the possibility that people will apply for a loan. Only the first 2 bars are exceptions.

Lets have a look over a pie chart to check the how much applicants belong to which kind of CITY.



- Applicants are most likely from the cities which are rated as "2" say, "Medium" rated.
- Then comes the applicants from "3" rated cities say, "Highest" rated.
- Applicants from lower rated cities are also somewhat low.

Lets have a look over a box plot to check the age of the applicants.



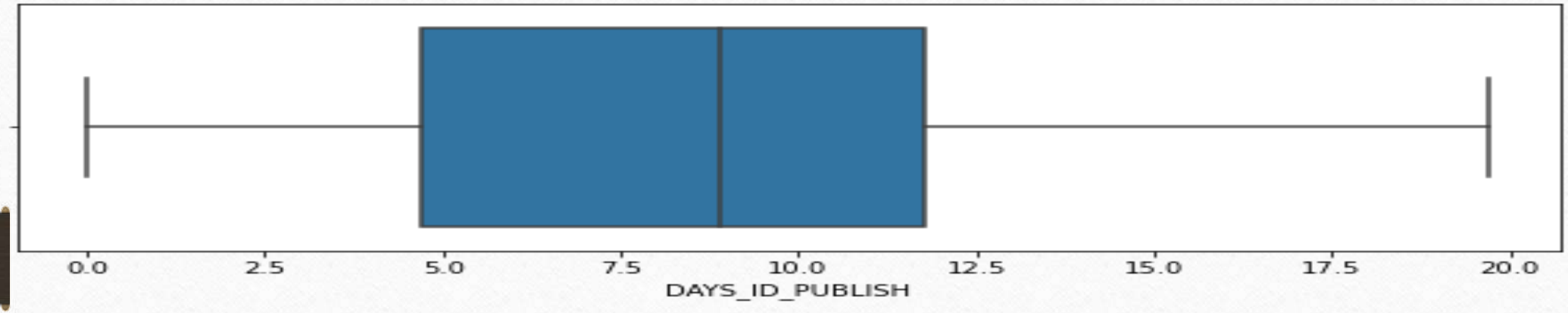
•This concludes 50% of the people's age lie within the range of 34 to 53 yrs. Also, 75% of the people who apply for the loan are below 53 yrs.

Lets have a look over a box plot to check the how many years ago the applicant joined the last employement.

Since it doesn't look much informative as it contains some unreal values in the column. So, to have a better insight lets take few quantile values till 80 percentile of the population.

•Here we can easily conclude that almost 50% of the population within the range who have been working for more than 2.5 years and less than 10 years.
•Also, people working for more than 12 years seem less likely to apply for a loan and same goes with people working for less than an year.

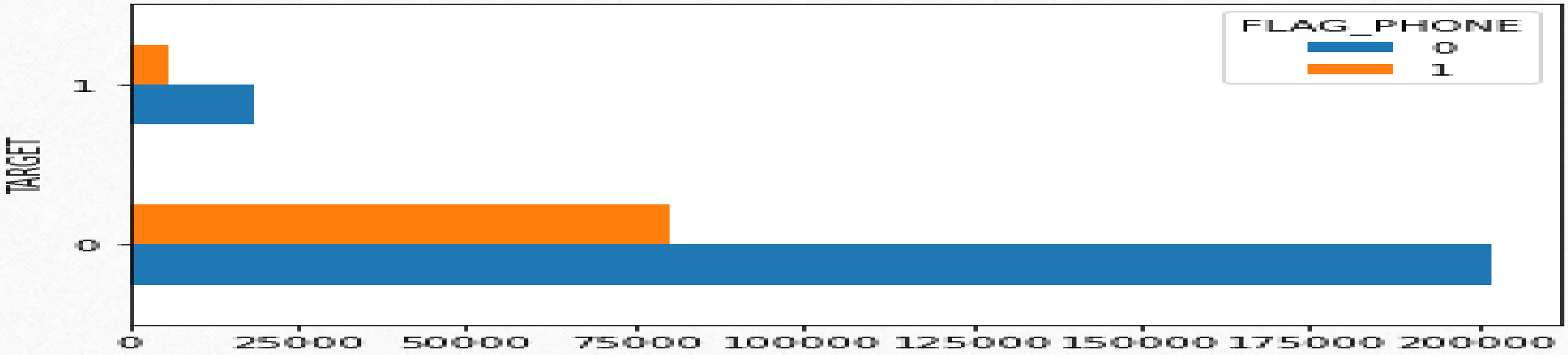
Lets have a look over a box plot to check the how many years ago did applicant got a new id issued before applying for the loan .



•We may conclude that 50% of the population issued a new id to apply for the loan approximately 5 to 12 years ago.
•There is no applicant whose id was issued 20 years ago or beyond.

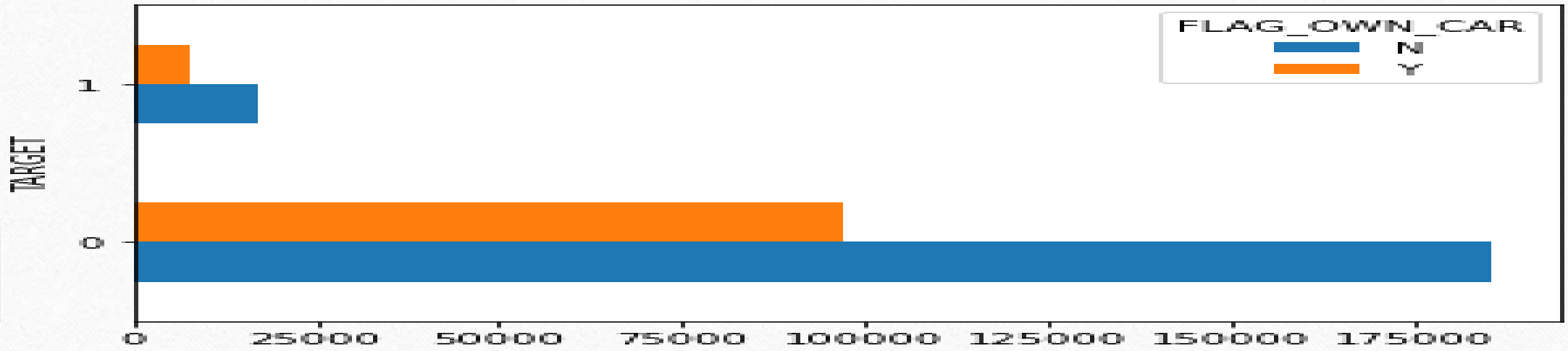
**From here on there will be total of 5 Flag variables which is categorized in two entries "1" and "0" with "1" being "YES" and "0" being "NO". Also, these Flag variables I will plot w.r.t. our TARGET variable.

Lets have a look over a bar plot between our TARGET column and FLAG_PHONE column to observe the trends they are following.



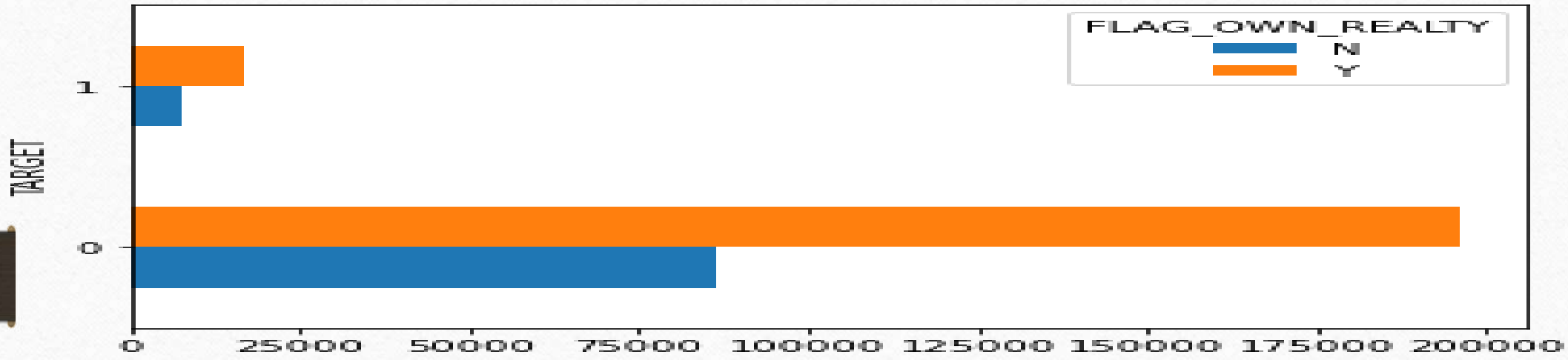
•We may conclude that the applicants generally don’t share their home phone number and in the defaulter’s list, the population of such people is a bit on a higher side.

Lets have a look over a bar plot between our TARGET column and FLAG_OWN_CAR column to observe the trends they are following.



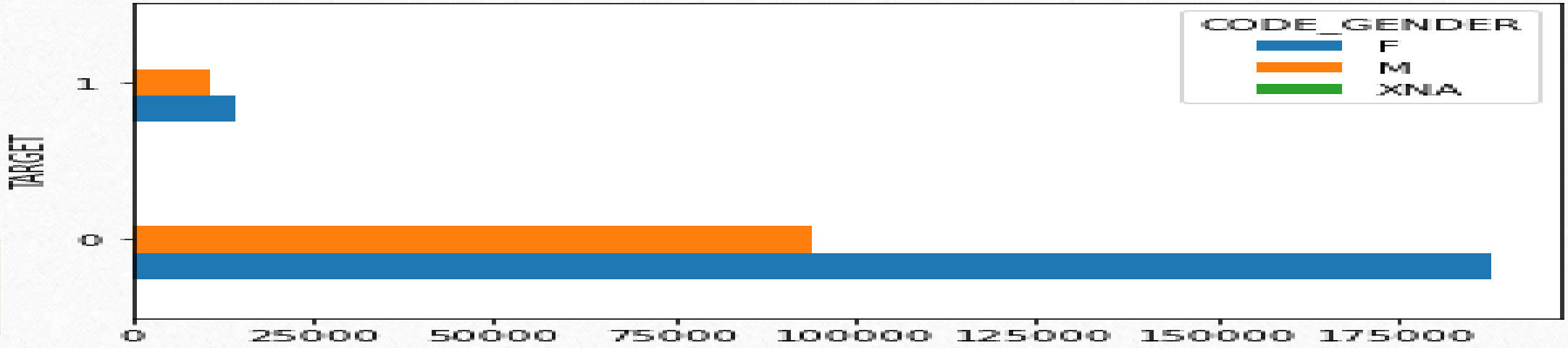
- We may conclude that the applicants generally don't own a car.
- Also, if we minutely observe the plots, we may say, those applicants who own a car are less likely to default.

Lets have a look over a bar plot between our TARGET column and FLAG_OWN_REALTY column to observe the trends they are following.



- We may conclude that the applicants generally don't own a real estate property or a house.
- Also, if we minutely observe the plots, we may say, those applicants who own a house are little more likely to default.

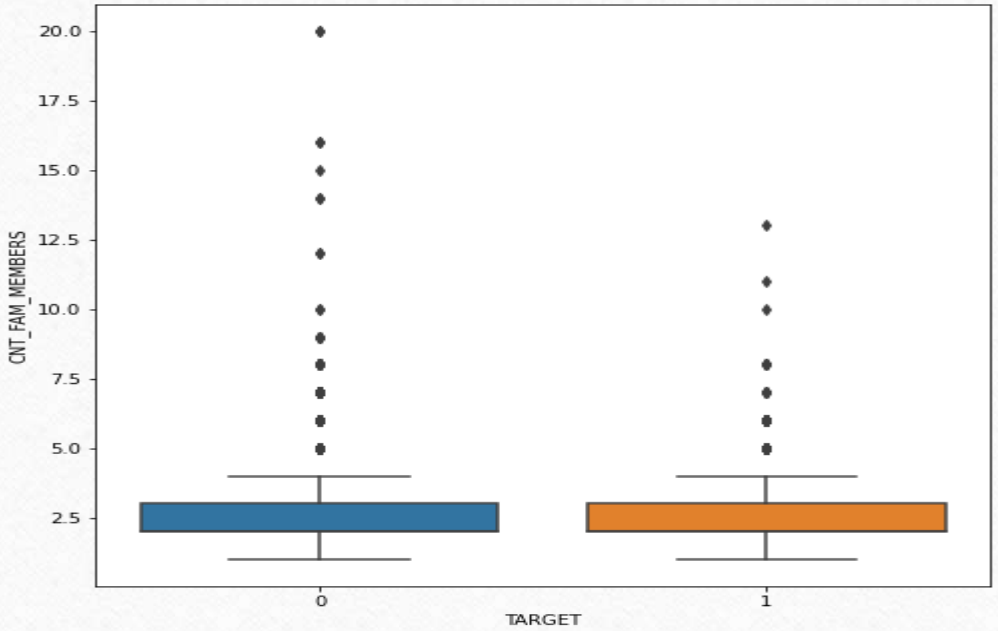
Lets have a look over a bar plot between our TARGET column and CODE_GENDER column to observe what trend they are following.



- We can easily see that female population who apply for a loan is significantly higher than male.
- But on the other hand, male population is much likely to be on the defaulter's list.

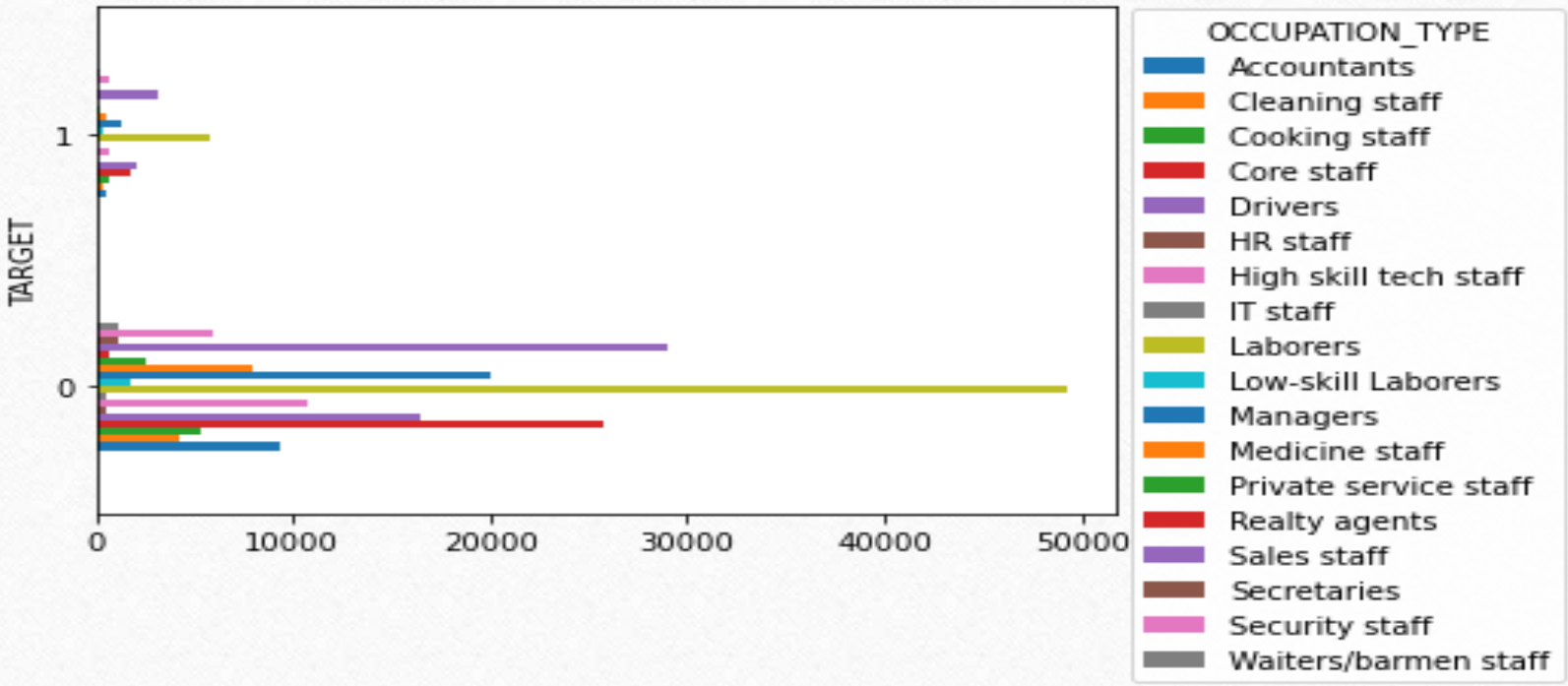
Here on, We will perform few of the interesting bivariate analysis.

Lets have a look over a boxplot between our TARGET column and CNT_FAM_MEMBERS column to observe the trends they are following.



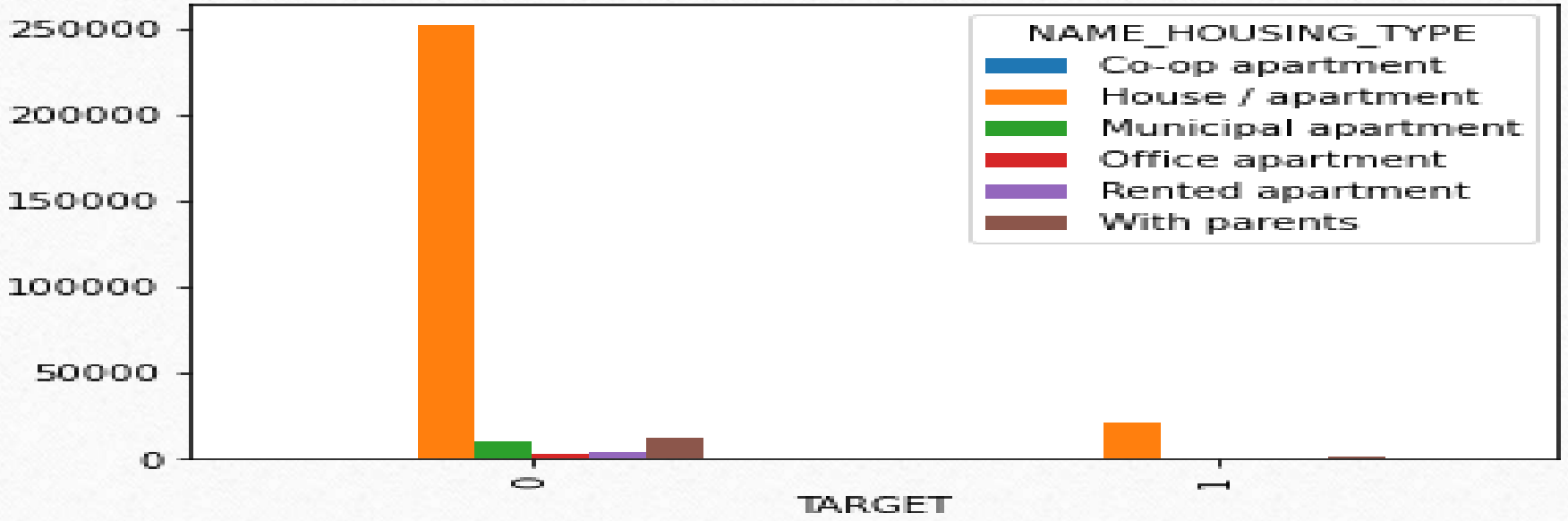
•The data looks almost the same for the same count of family members in both the categories in the TARGET column.

Lets have a look over the bar plots between our TARGET column and OCCUPATION_TYPE column to observe the trends they are following.



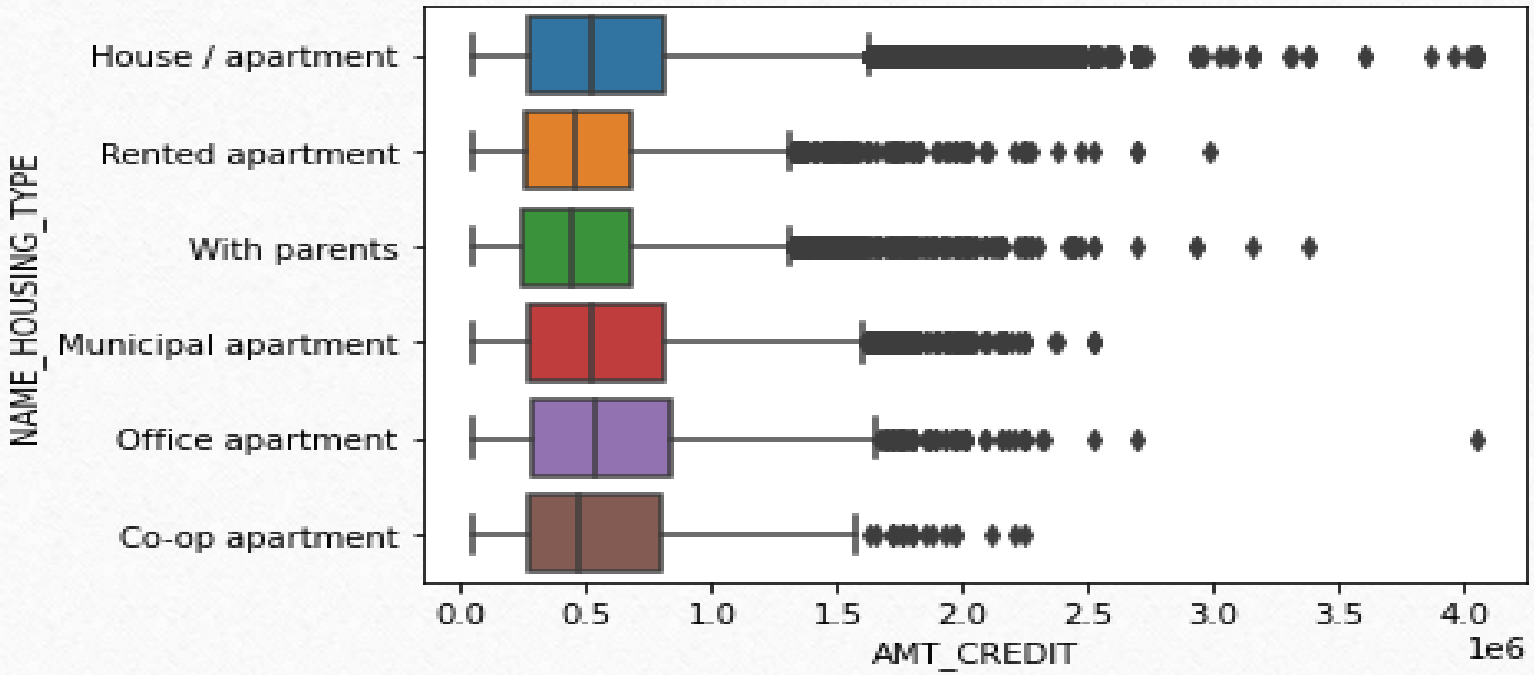
- The graph intends to provide an insight on what trends the applicants from each of the occupation are following.

Lets have look over a bar plot between our TARGET column and NAME_HOUSING_TYPE column to observe the trends they are following.



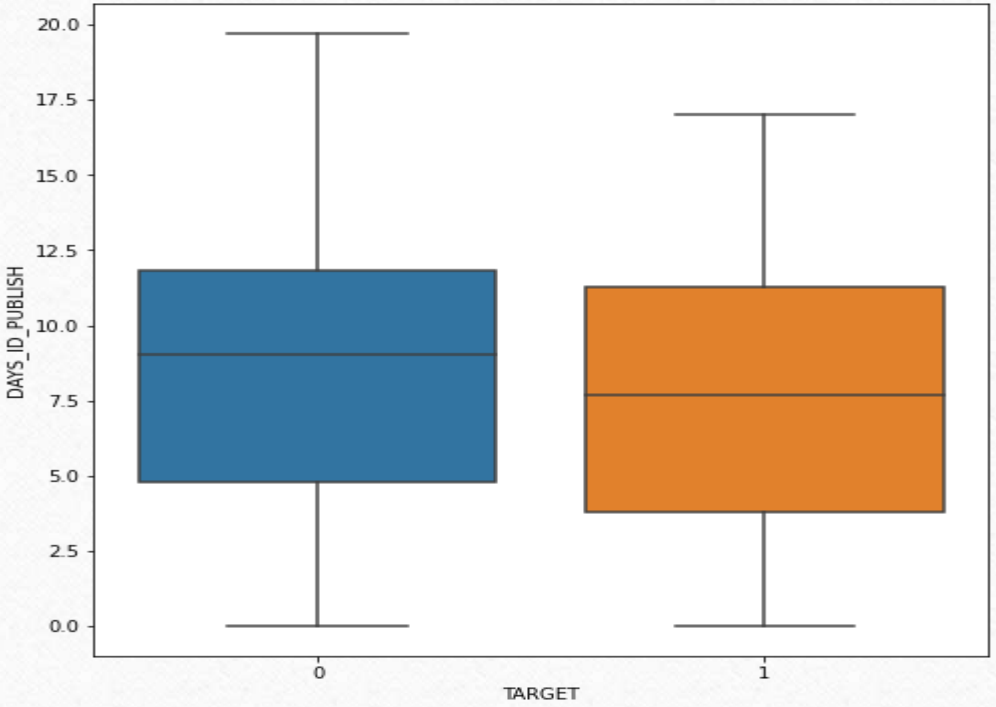
- The graph intends to provide an insight on what trends the applicants from each of the housing type are following.

Lets have a look over a boxplot to check which housing category gets what amount of loan credited.



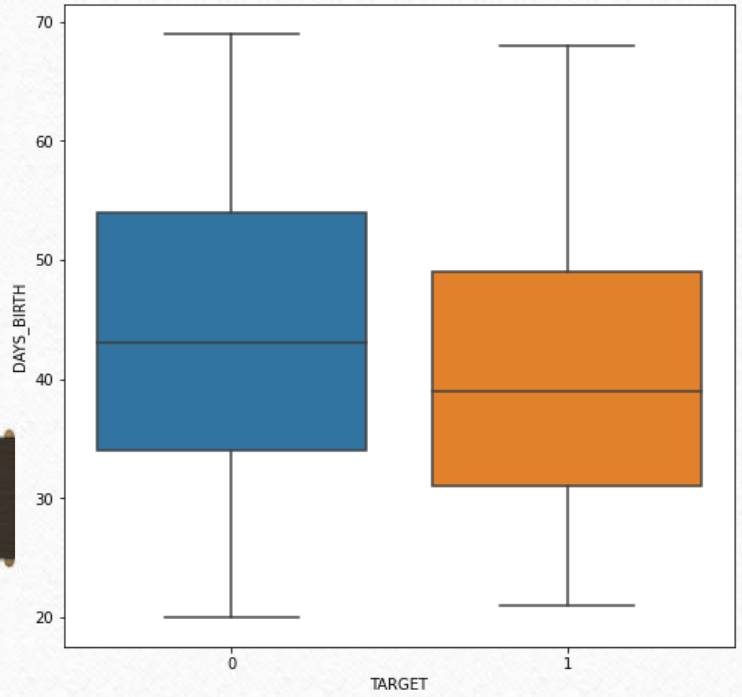
AMT_CREDIT= (“x” ticks)*10^6

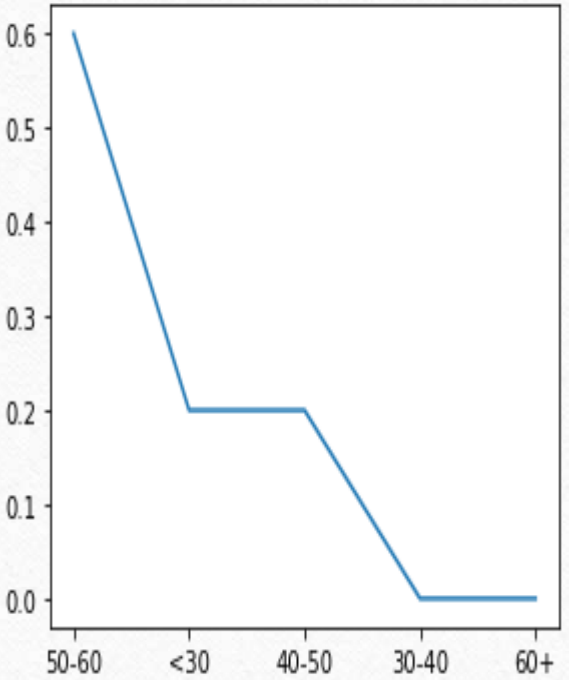
Lets have a look over the TARGET column with respect to the DAYS_EMPLOYED variable now.

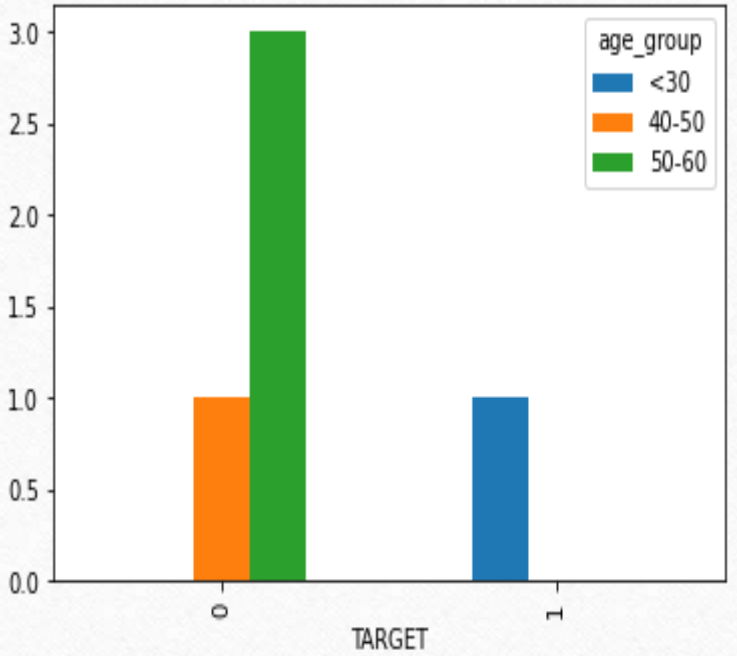


- Here we can say that people who are more likely to default are those who have recently changed the identity document.

Lets have a look over the TARGET column with respect to the age variable now i.e. DAYS_BIRTH.

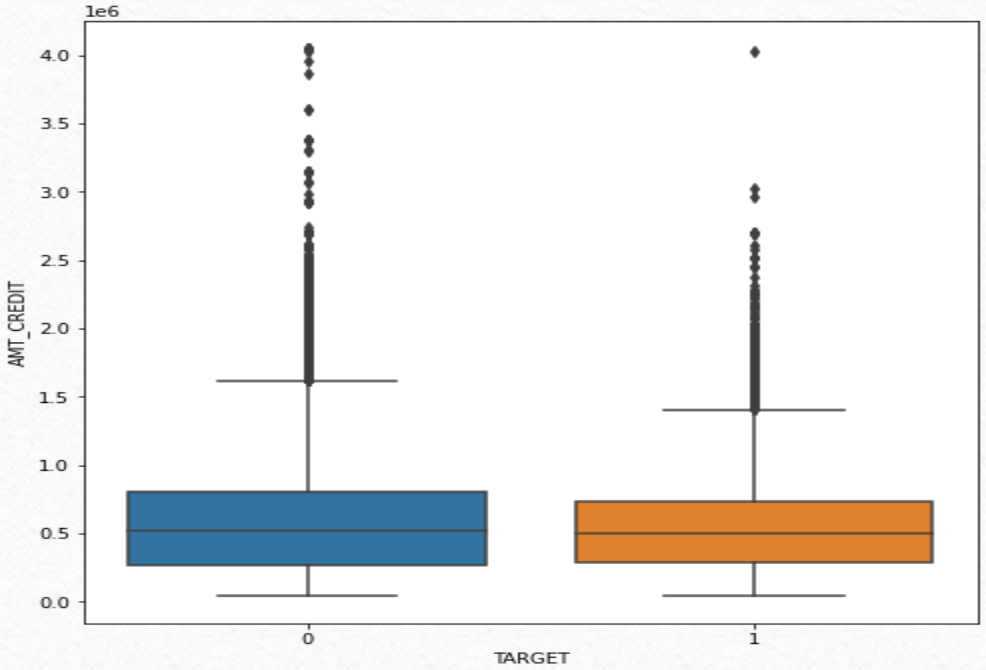






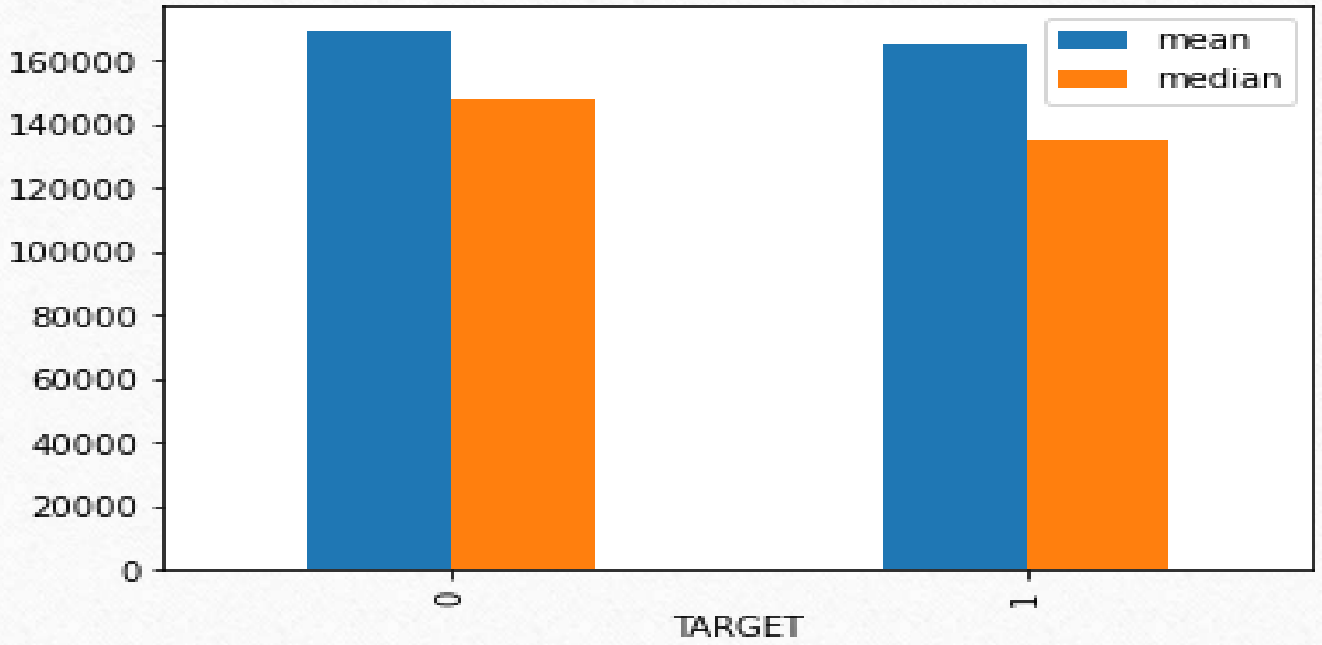
- Here, we can say that people within the least age range and people with higher age range are less likely to default in an overall comparison.
- People under 30 years of age are very likely to default.

Lets have a look over a box plot for the TARGET column with respect to the AMT_CREDIT variable now.



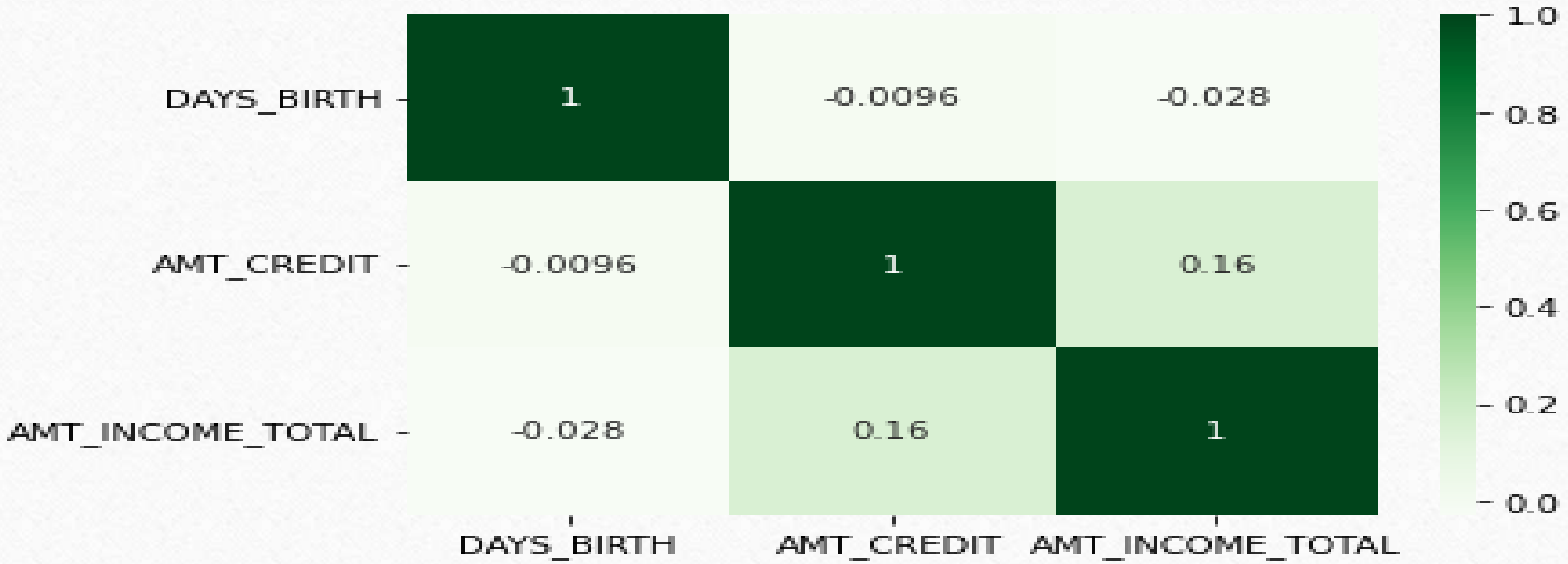
•Here we don't get too much of an insight. But, we can say a customer is less likely to default in case he has applied for a credit within the lowest and the highest range.

Lets have a look over the bar plots of AMT_INCOME_TOTAL's mean and median with TARGET.



•We can't figure out much with this since, there is not much difference between the bars. However, it suggests people with lower income are a bit more likely to default.

Lets have a look if there is any correlation between variables (DAYS_BIRTH, AMT_CREDIT, AMT_INCOME_TOTAL) with the help of heat map.

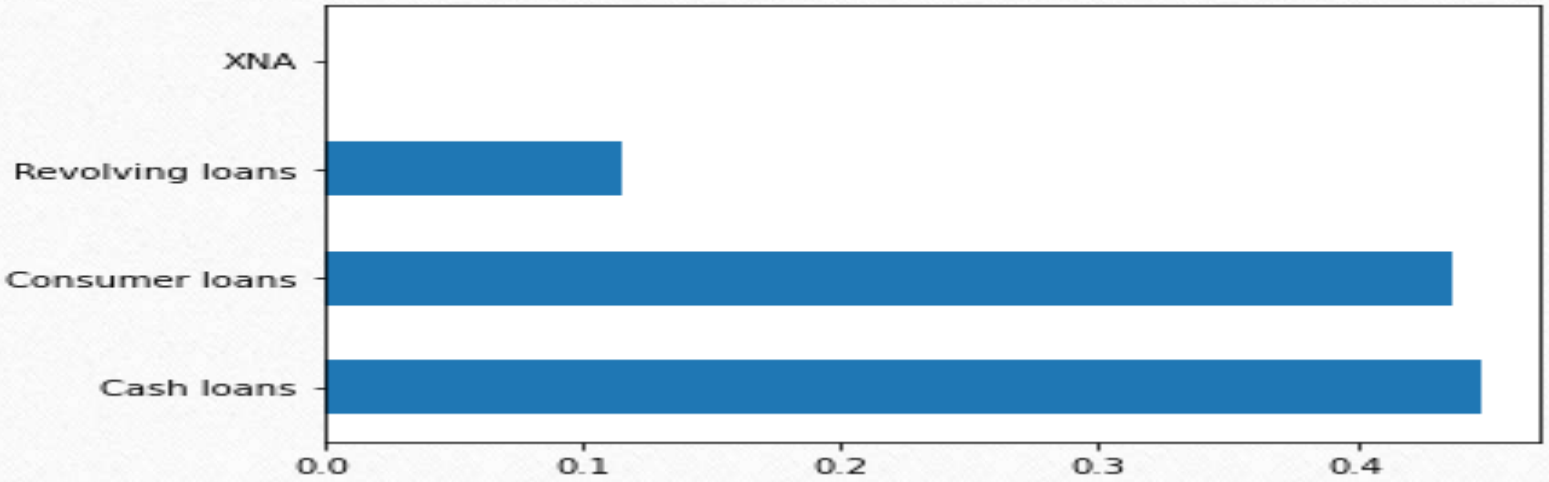


•Here we can see there is no strong correlation among these 3 variables.

Moving on to the next dataset now i.e. PREVIOUS_APPLICATION DATASET

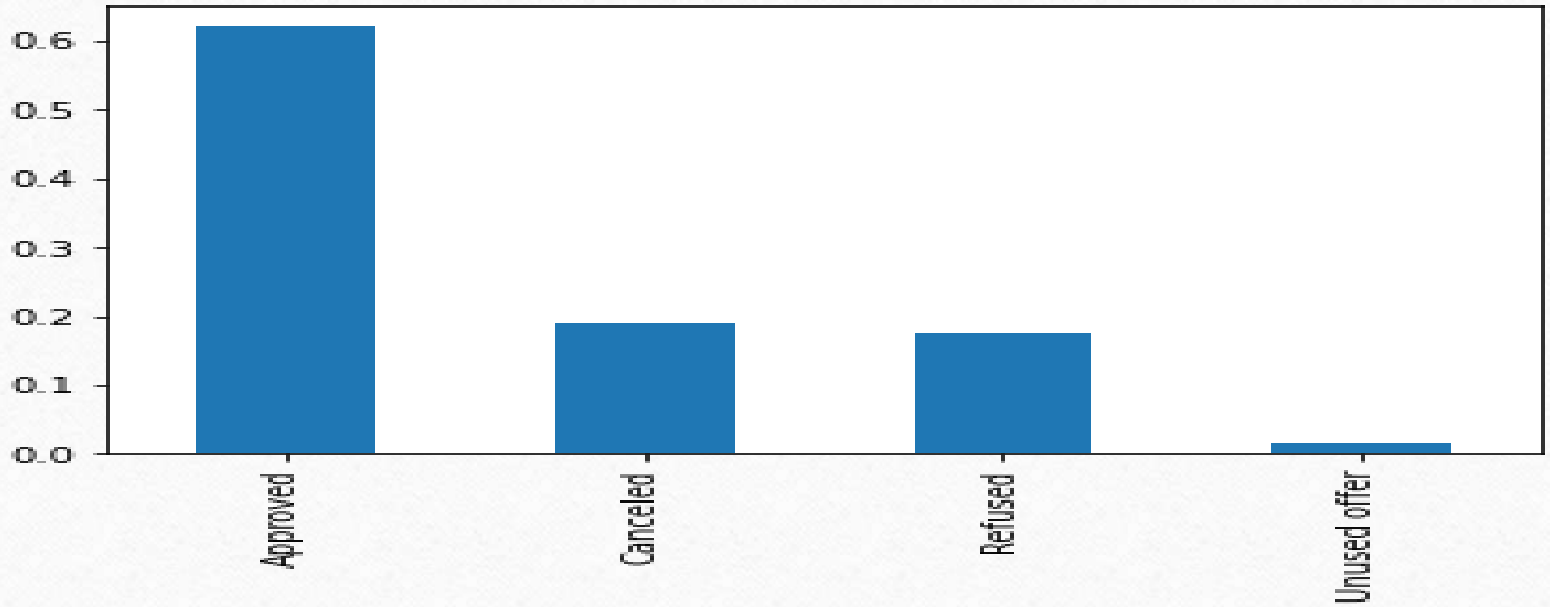
Lets start with some univariate analysis now. I will analyze a set of 4 to 5 important columns as per the best of my understanding.

I will start with our target variable in this dataset which is NAME_CONTRACT_TYPE that gives the category of loans customers has gone for in the previous applications.



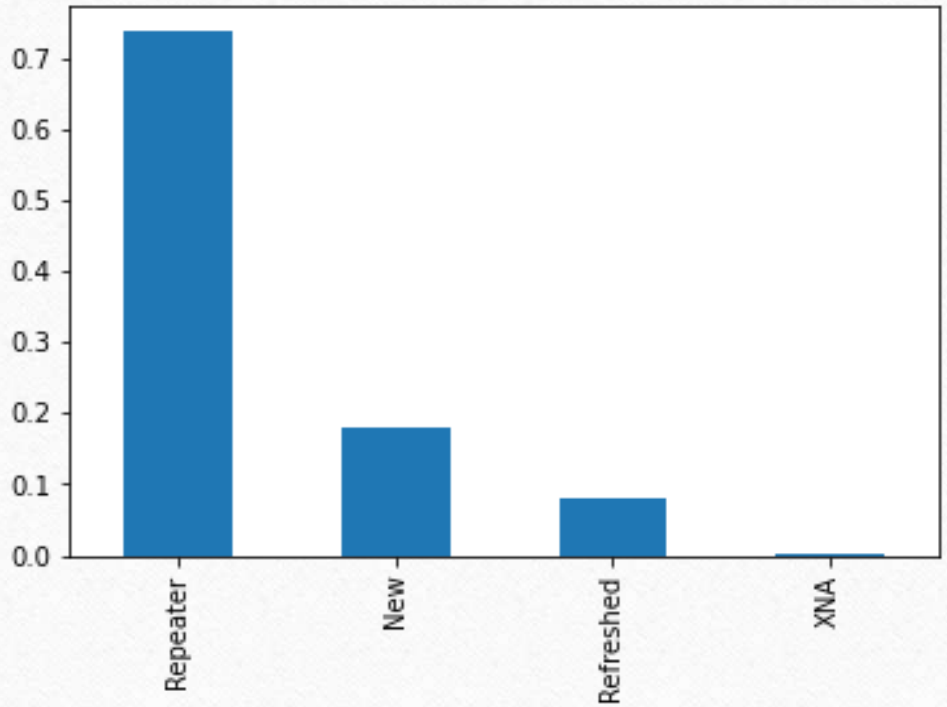
•Cash loan applicants are the highest and consumer loan applicants are also not far behind.

Lets have a look over the univariate analysis on the column NAME_CONTRACT_STATUS



- Almost 62% of the application shows approved amongst the applicants.
- Cancelled and refused applications percentage looks almost similar.

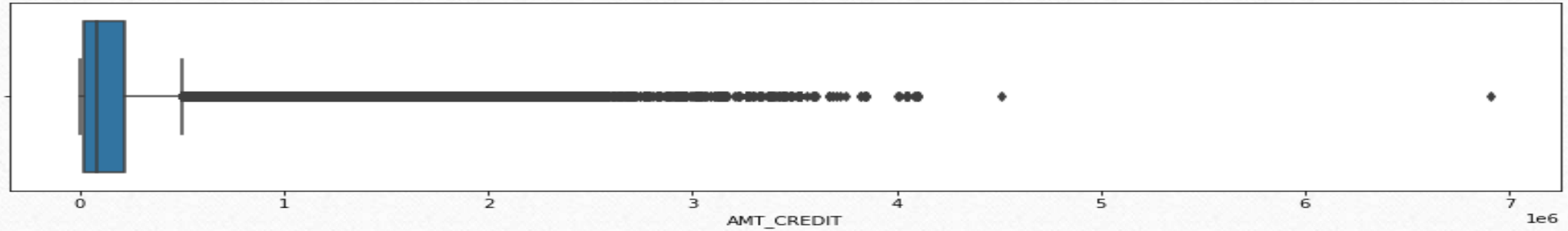
Lets have a look over the univariate analysis on the column NAME_CLIENT_TYPE.



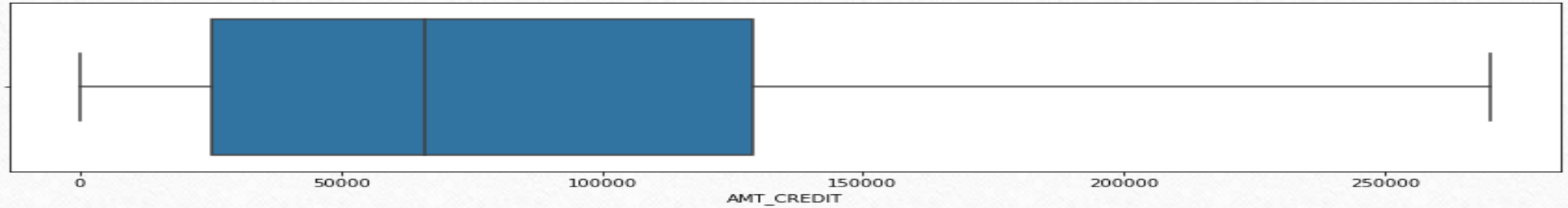
- Here almost 75% of the old clients seems to submit the application again.

Lets have look over the univariate analysis on some numerical columns now.

Lets do a univariate analysis on the column AMT_CREDIT.

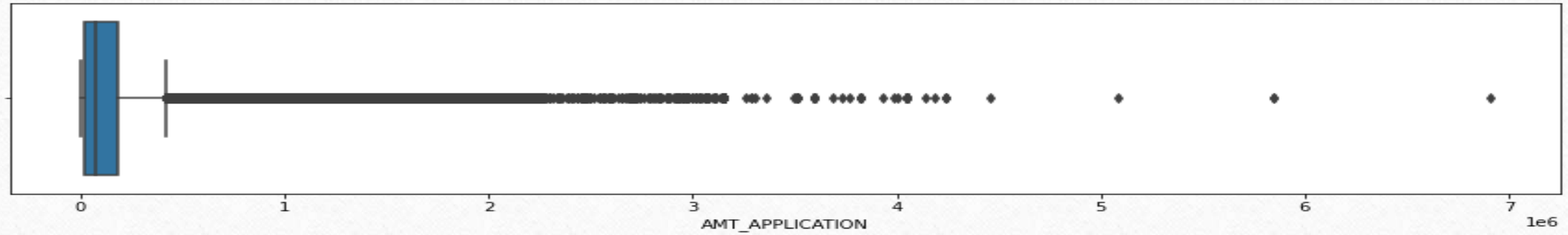


Taking quantile till 0.8 to generate a clear and more meaningful insight.

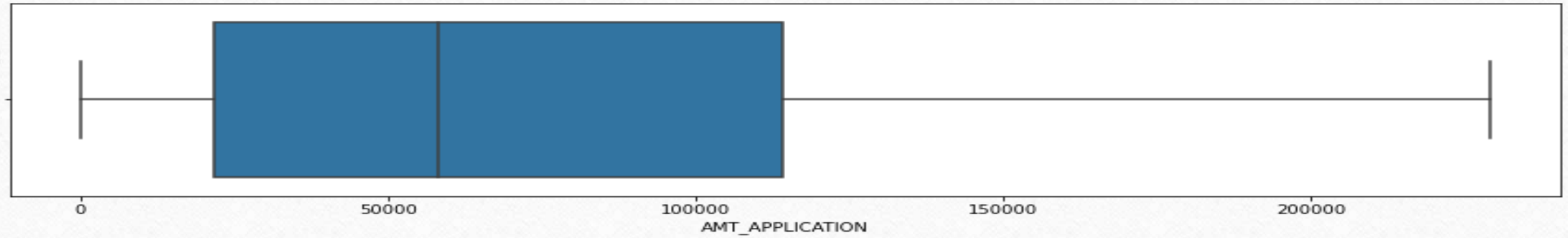


- Almost 80% of the people are within the range of 2.75L loan amount.
- Among these 80% of the population, 50% of the people lie within the range of 25k to 1.25L loan amount.

Lets have a look over the univariate analysis on the column AMT_APPLICATION.



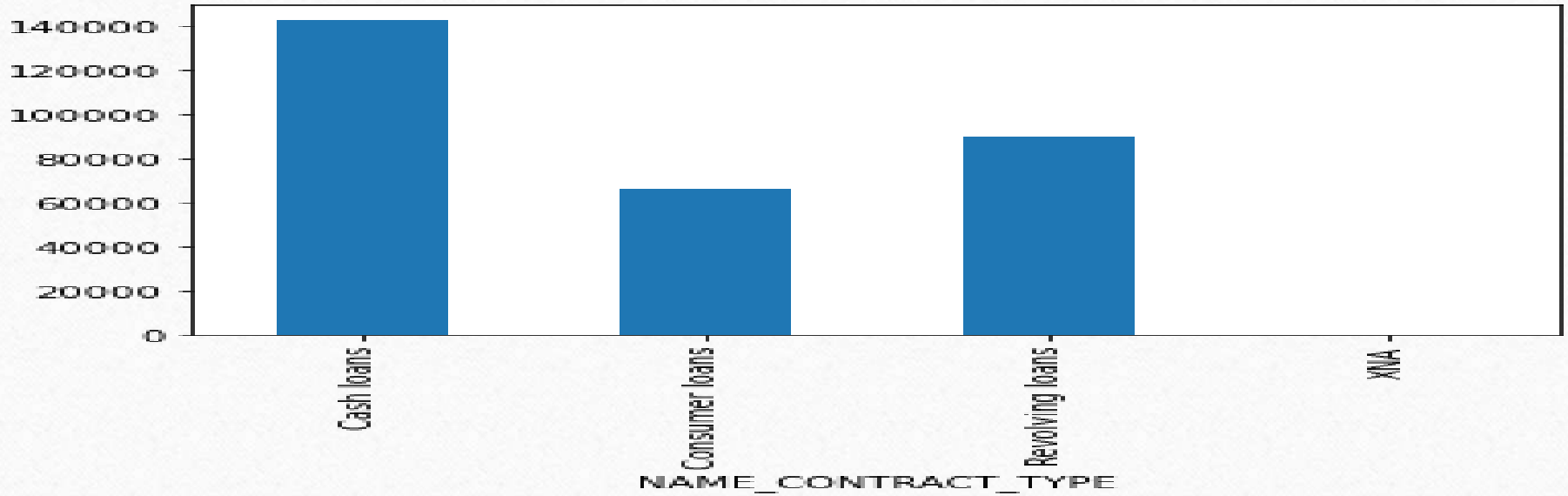
Taking quantile till 0.8 to generate a clear and more meaningful insight.



- Almost 80% of the people have requested for a loan amount lesser than 2.5L.
- Among these 80% of the population, 50% of the people are those who requested for a loan amount within the range of 25k to 1.20L.

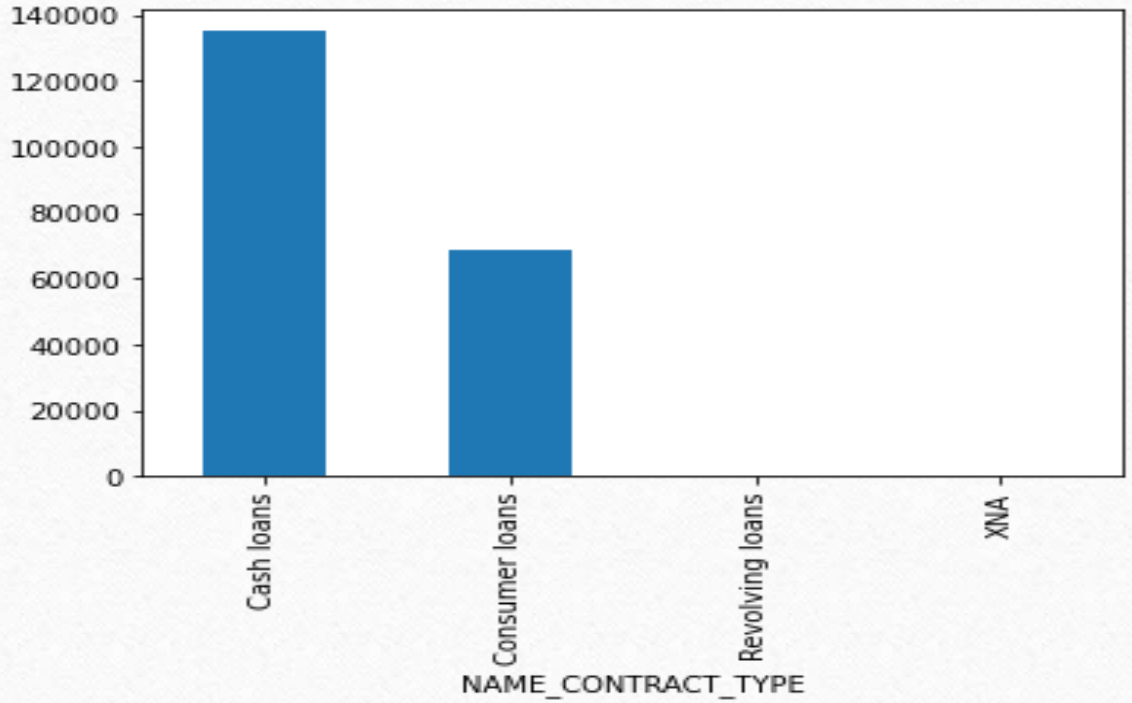
Here on, I will do some of the bivariate analysis based on our target column i.e. NAME_CONTRACT_TYPE and analyze the results on the basis of some plots.

Lets have a look over the bar plot between NAME_CONTRACT_TYPE and AMT_CREDIT.



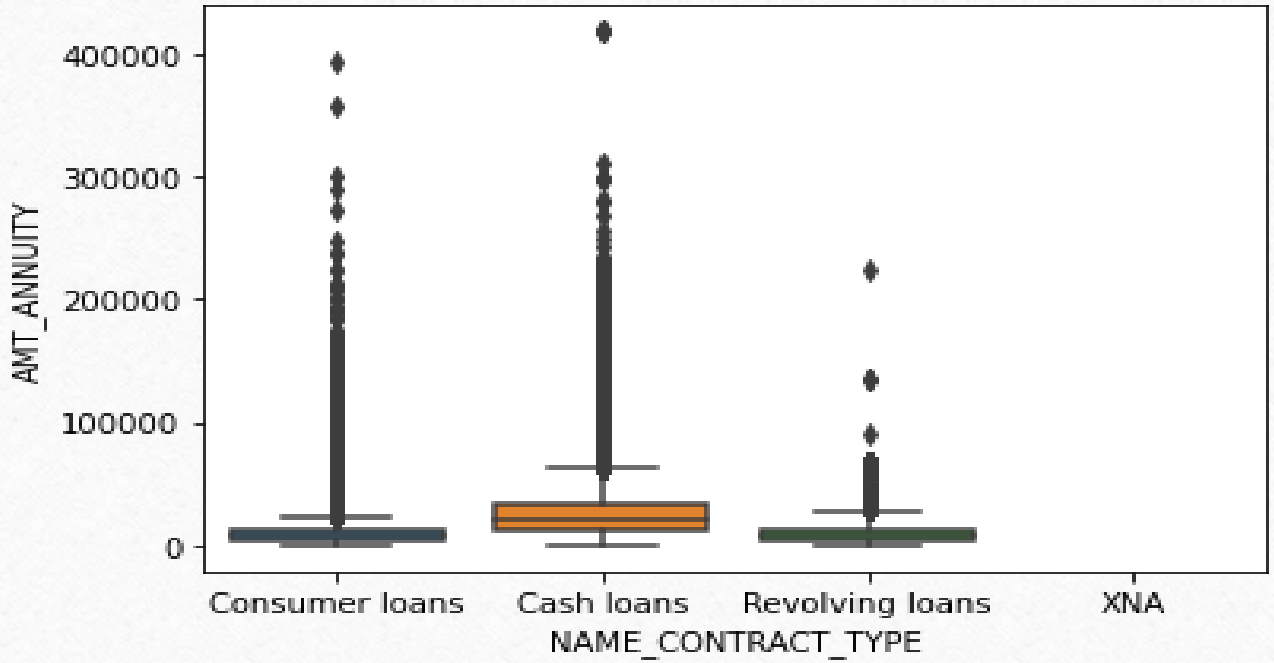
• It reflects, one who has taken cash loans have often been credited with a significantly higher loan amount.

Lets have a look over the bar plots between NAME_CONTRACT_TYPE and AMT_APPLICATION.



- Again, people requesting cash loans tend to request for a significantly higher amount.
- Negligible requests are made for revolving loans.

Lets have a look over the box plot of NAME_CONTRACT_TYPE w.r.t. to AMT_ANNUITY.

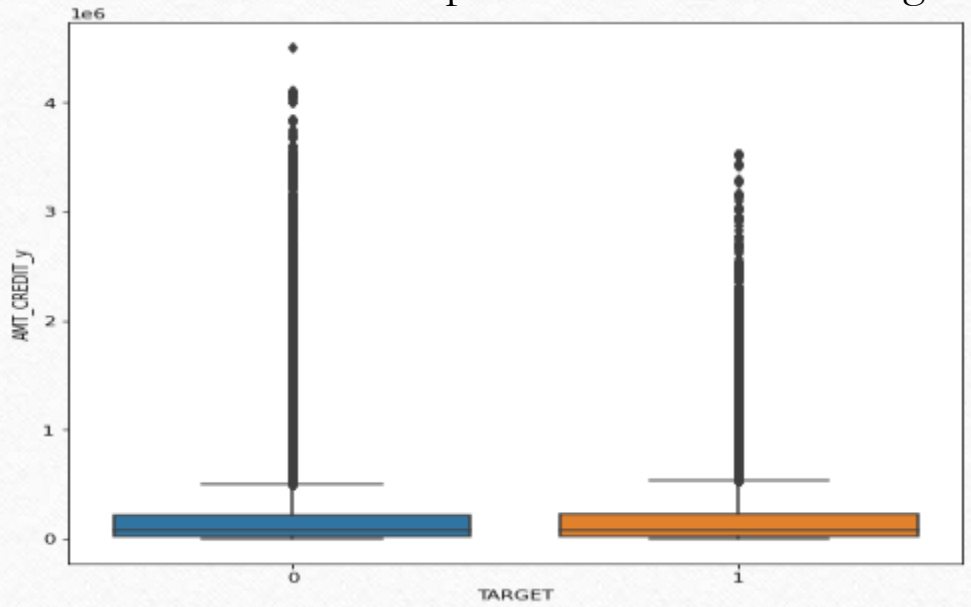


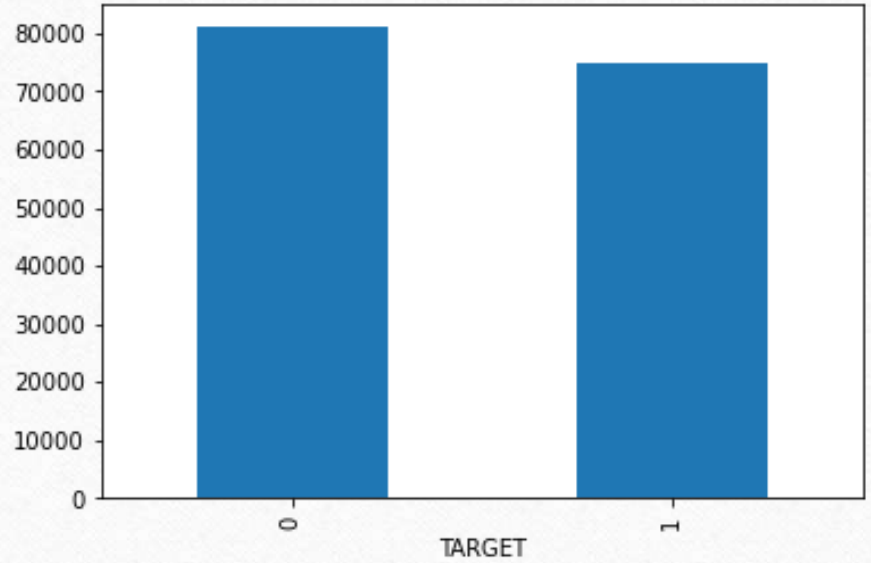
•Again as expected, we can conclude that annuity for cash loan seems to be on significantly higher side.

We will now merge both the table doing inner join based on the column SK_ID_CURR and try to figure out the insights based on original target column i.e. TARGET column from the table application_data.csv

Lets do some bivariate analysis based on our TARGET column in the application_data.csv table.

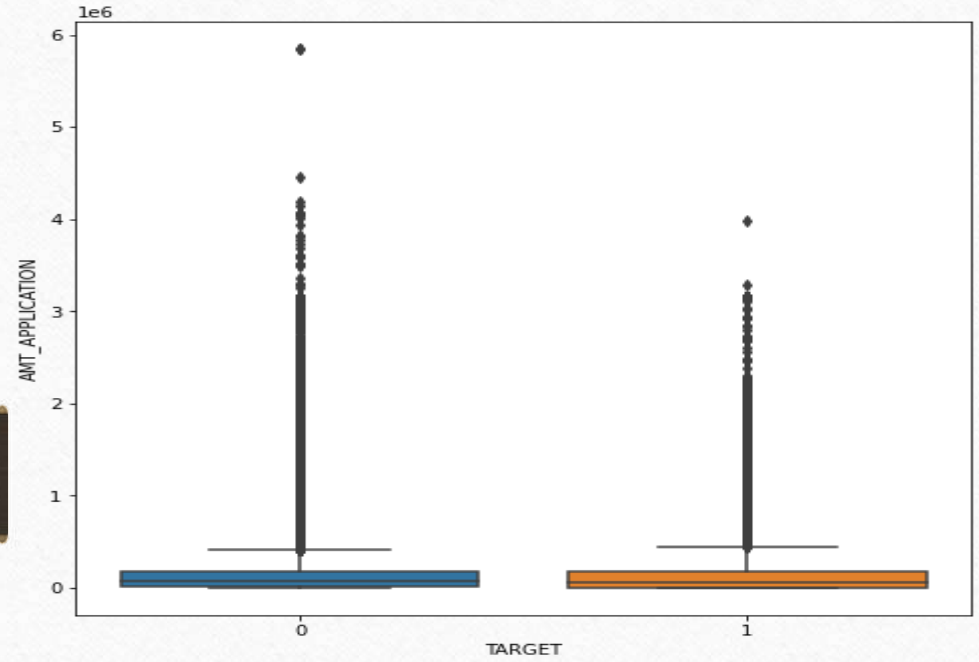
Lets have a look over the box plot of AMT_CREDIT grouped by TARGET column.

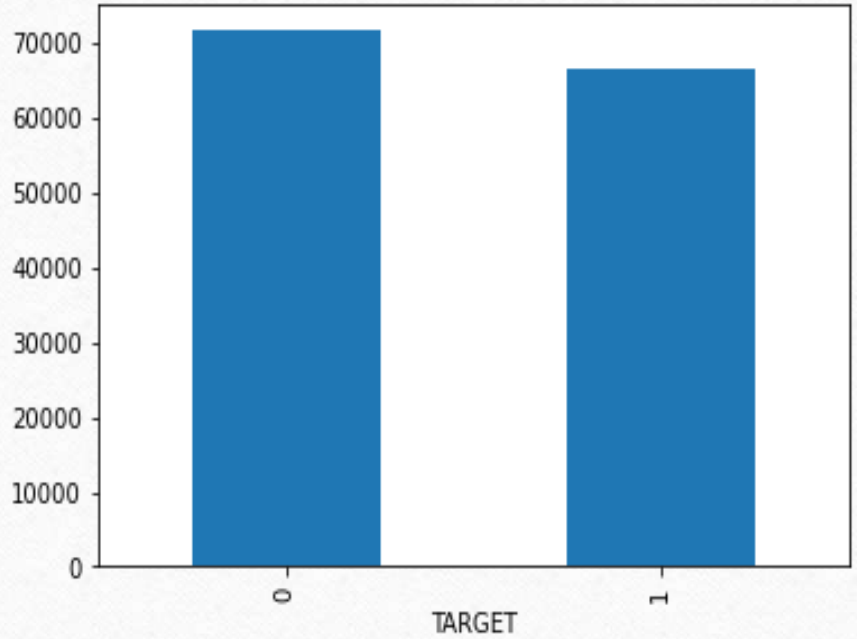




•Not much of an insight to draw from a boxplot but, if we look over the bar plot, we can certainly say an applicant is less likely to default in case he has applied for a credit within the highest range.

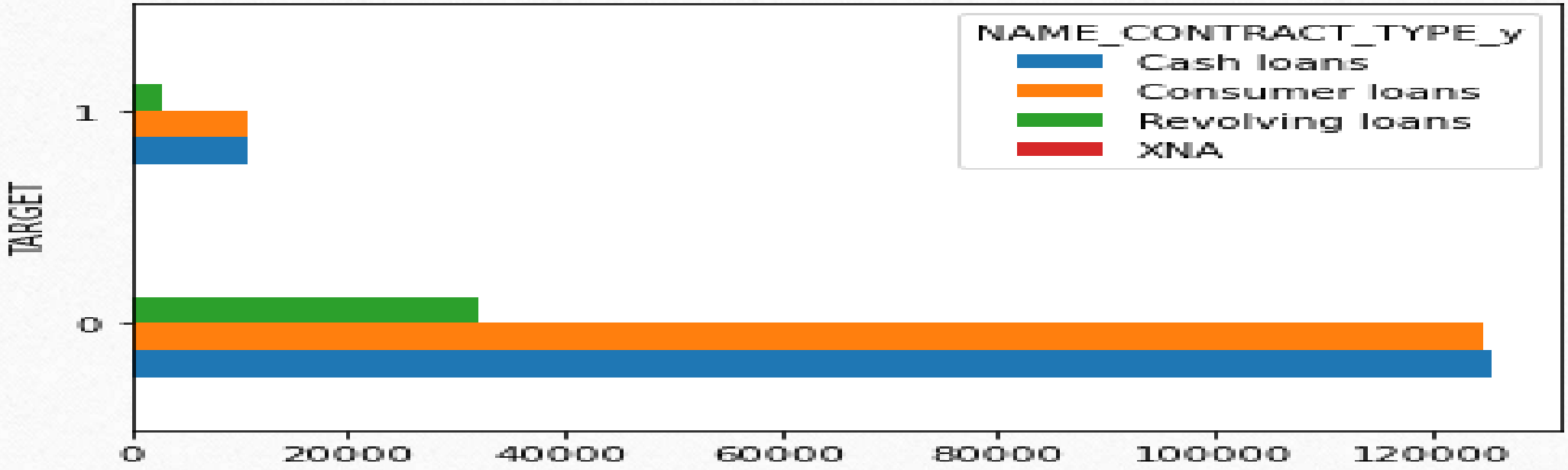
Lets have a look over the box plot of AMT_APPLICATION grouped by TARGET column.





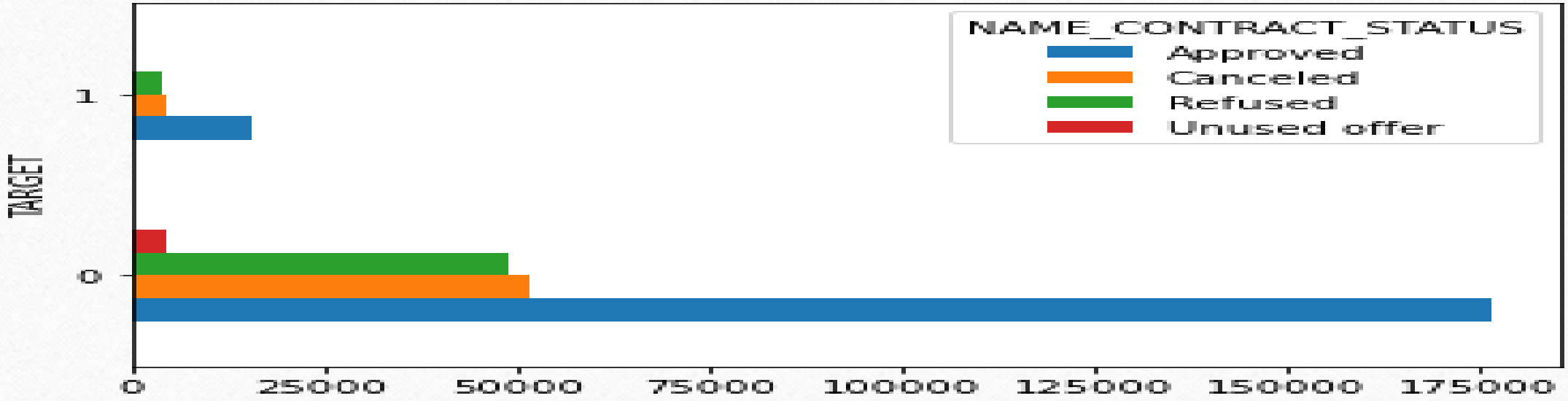
•Not much of an insight to draw from a boxplot but, if we look over the bar plot, we can certainly say an applicant is less likely to default in case he has given an application for the loan amount within the highest range.

Lets have a look over the bar plots between our TARGET column and NAME_EDUCATION_TYPE column from the previous application dataset to observe the trends they have been following.



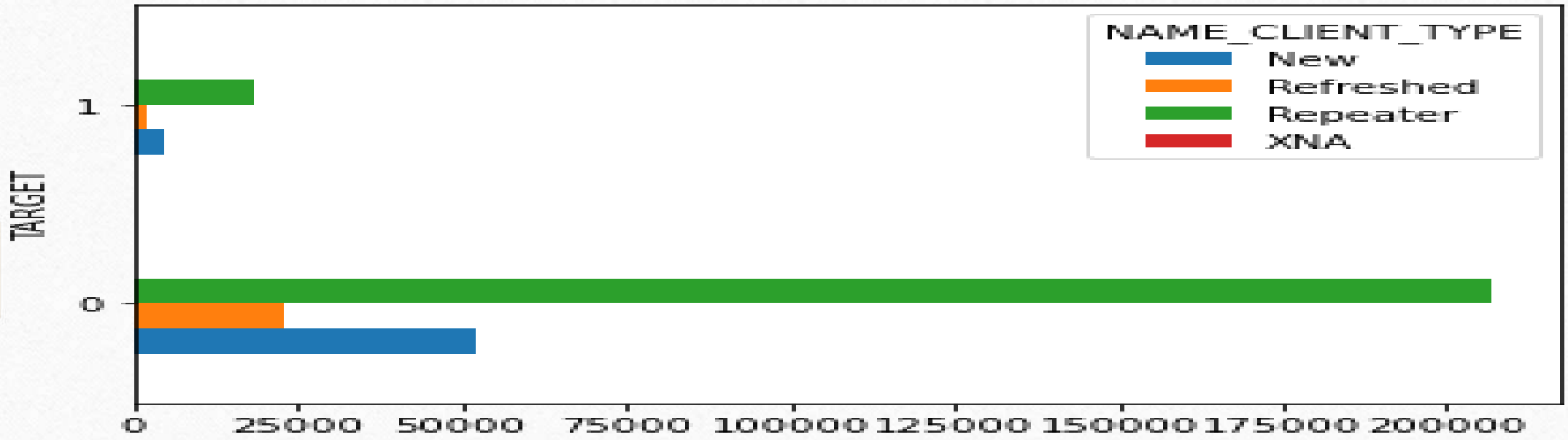
• If we look over the minutest of the details, we can conclude that people opting or going for revolving loan type tend to be in the defaulter's list quite more.

Lets have a look over the bar plots between our TARGET column and NAME_CONTRACT_STATUS column to observe the trends they have been following.



• This bar plot can basically show us an insight on an approximate count of people in defaulter's and non-defaulter's lists based on their contract status in the past.

Lets have a look over the bar plots between our TARGET column and NAME_CLIENT_TYPE column to observe the trends they are following.



• This bar plot can basically show us an insight on an approximate count of people in defaulter's and non-defaulter's lists based on the fact that they have been new customers or already existing customers or refreshed customers based on the past data.

I hope I was able to provide a good understanding and insights on both the datasets individually and also on the merged dataset.

I have done some basic use of Python's "numpy" and "pandas" libraries which came in handy to perform data cleaning on the given raw datasets.

Also, I have shared most of the insights with the basic use of Python's "matplotlib" and "seaborn" libraries to give a graphical insights based on the datasets.

I am sure you will find it a lot more easy to analyze and observe the patterns in the data and will be able to draw some interesting conclusions out of it.

THANK YOU !