Case Study on Loan Risk Analysis EDA

KRUNAL TANNA

ALAKNANDA AGARWAL

Introduction

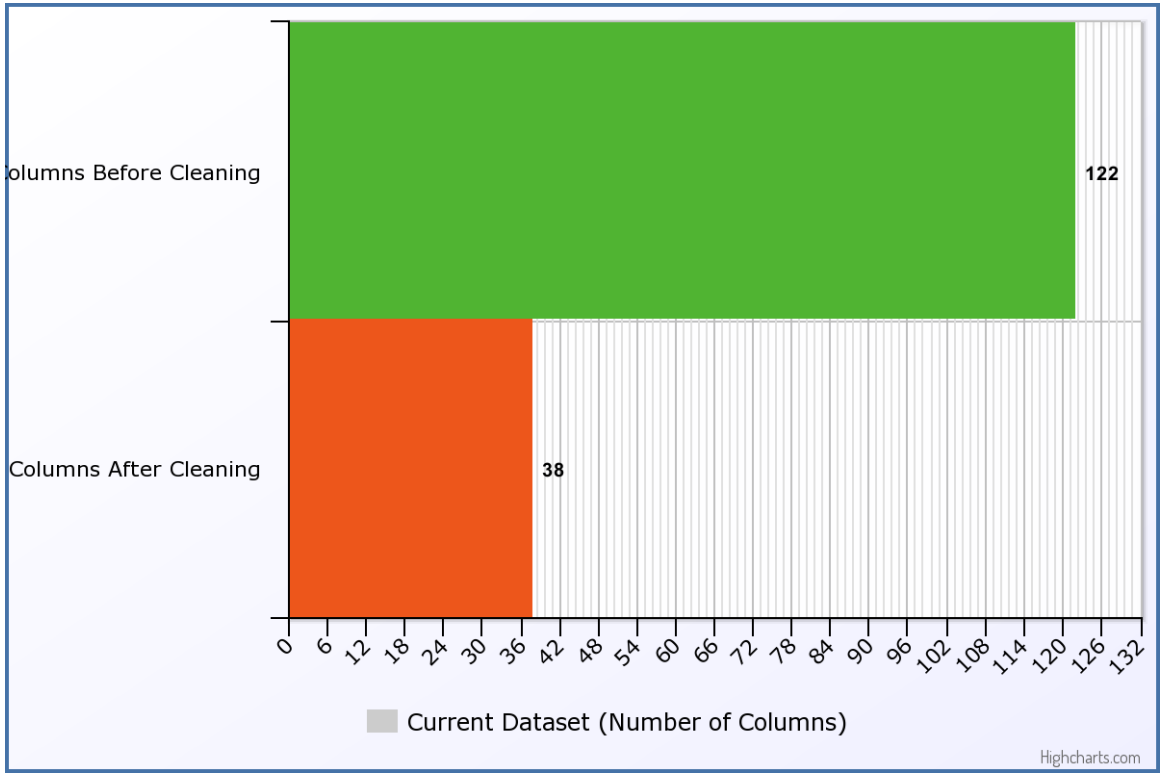
- Credit Risk Analysis in banks for Loans sanction process is an essential and important part that needs to be implemented in a more accurate and comprehensive way
- We have painted a picture detailing the most important aspects related to the loans and performed **EDA (Exploratory Data Analysis)** to find our driving variables which plays important role in analyzing risk associated with loans
- A Current Loan Dataset is used having important attributes such as Age, Housing Type, Annual Income, Credited Amount, Organization Type, Defaulters and many more
- Along with this, A Previous Dataset is used which have some extra columns like Loan status
- The purpose of this EDA is to find insights, important attributes and patterns by doing various steps like Data cleaning, Data preparation, Data Analyzing and Data visualization.

Data Cleaning

- Removed columns having more than 50% NULL values, Columns found irrelevant have been dropped directly
- For columns having lesser null values, found ways in which it can be imputed with appropriate values using mean/median
- Checked for outliers, if any
- Changed some datatypes and derived new columns using existing columns
- Created bins for continuous variables like '**Amount Credited'** and '**Annual Income'**
- Nulls in column **AMT_GOODS_PRICE** can be imputed with the median of the column.
- Columns **AMT_CREDIT_RANGE, AMT_INCOME_RANGE, AMT_ANNUITY_RANGE, DAYS_BIRTH** have been binned.
- Datatype conversion done in columns **FLAG_OWN_CAR, FLAG_OWN_REALTY** from Object to Boolean.

Data Cleaning

- Before Cleaning: Loan shape - (307511,122)
- After Cleaning: Loan shape - (307511,38)



Data Analysis

Have split the application loan data into two datasets- **default** with Target 0, and **non_default** with Target 1.

There is a huge imbalance in the data, the percentage of default rows is **8.07%** and non_default is **91.93%.**

Have plotted various plots for univariate and bivariate analysis of variables with respect to the Target column.

1. Univariate Analysis (categorical variables)

In Fig. 1, It is clearly observed that people with Cash loans have more chances of getting default than people with Revolving Loans

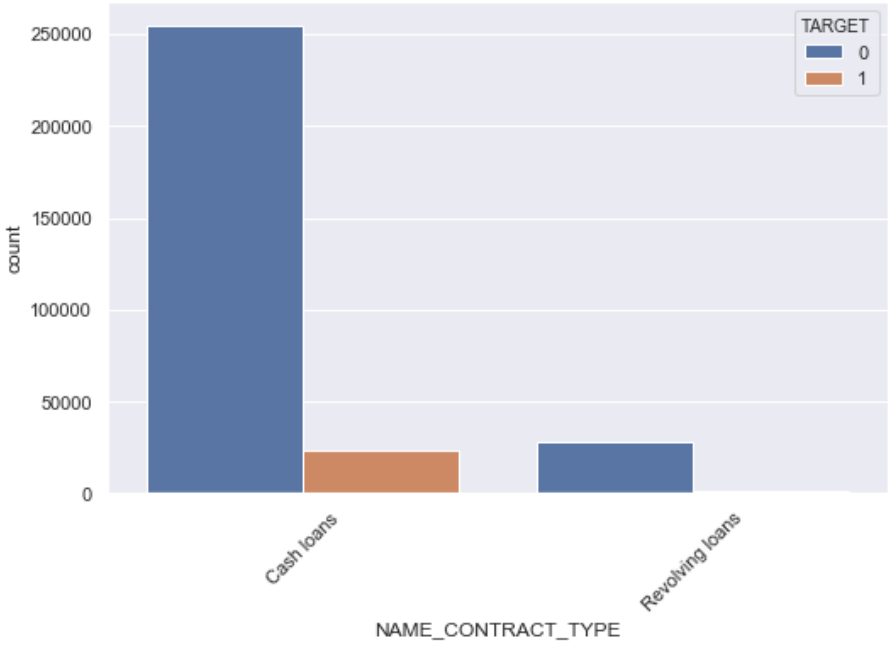


Fig. 1

In Fig. 2, It is observed that applicants having own house/apartment have higher chances of getting defaulted as well as non defaulted

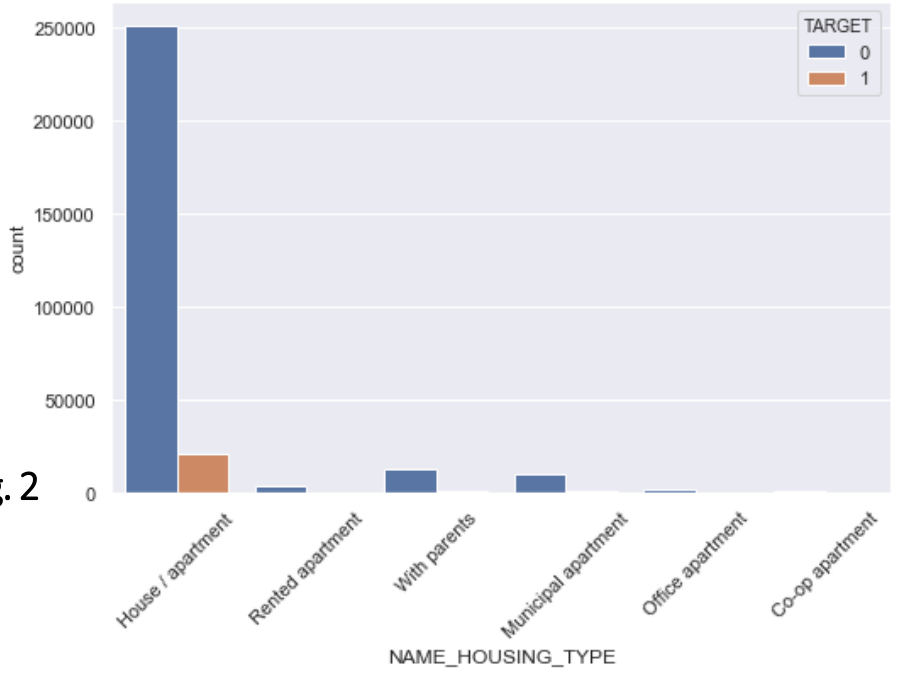


Fig. 2

2. Univariate Analysis (categorical variables)

Fig. 3 shows the differences for v/s unaccompanied. It is visible that chances of getting non-defaulters can be greater than applicants living alone

Fig. 4 shows a sharp spike in working population having higher chances of non-defaulters in compare to State servants

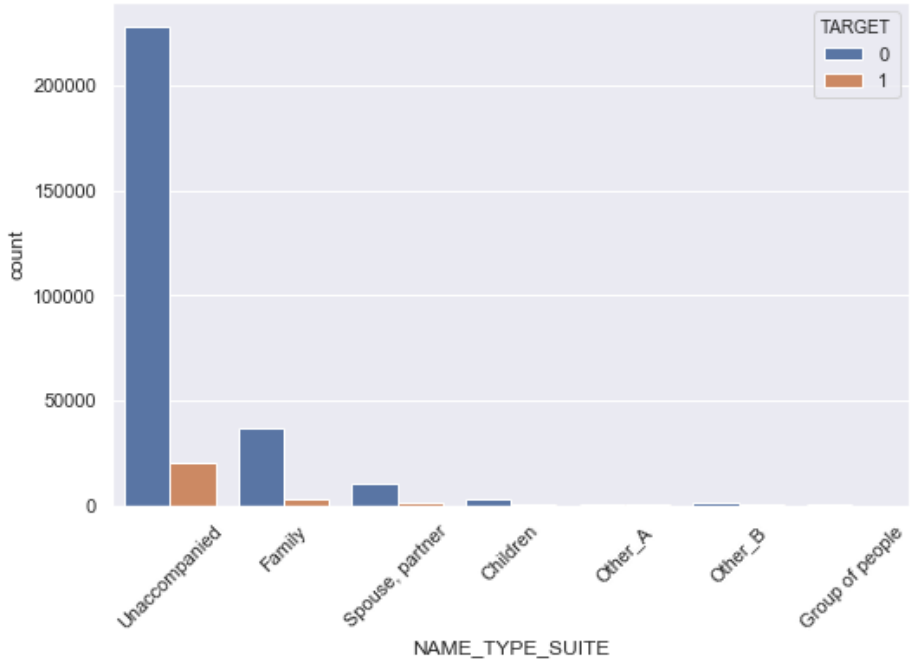


Fig. 3

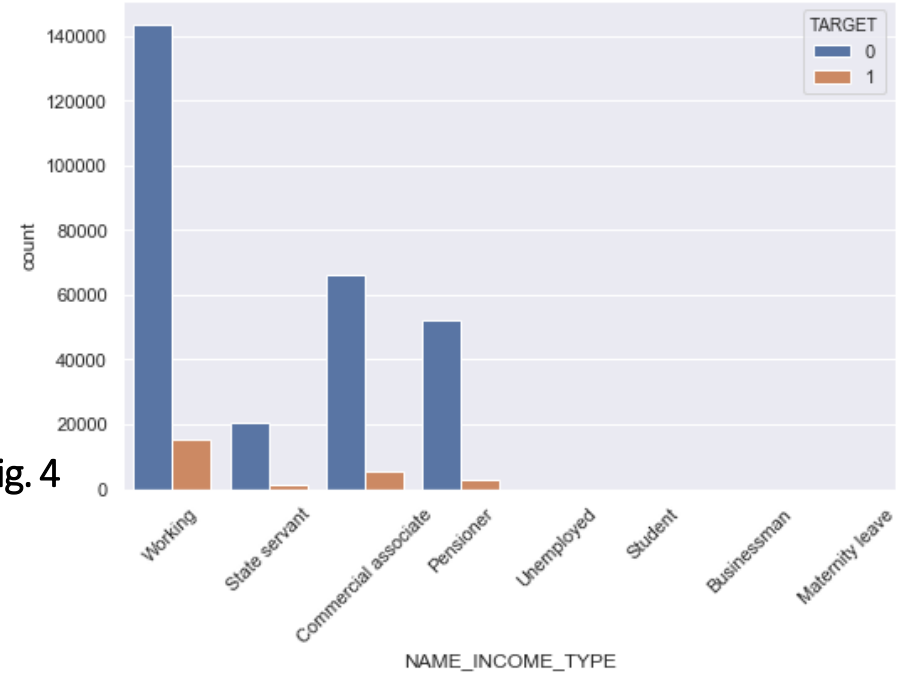
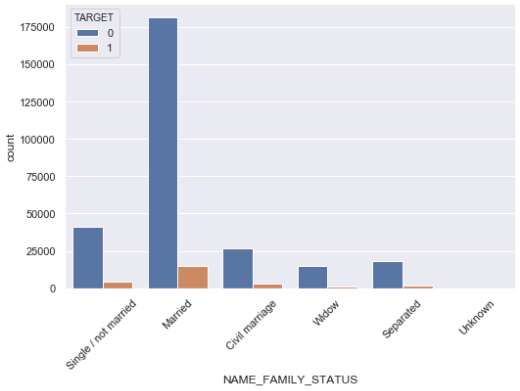


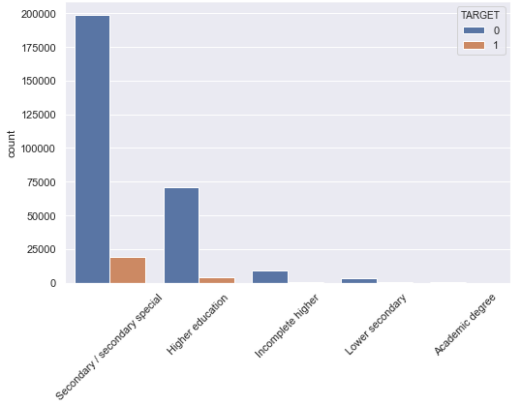
Fig. 4

Family Status



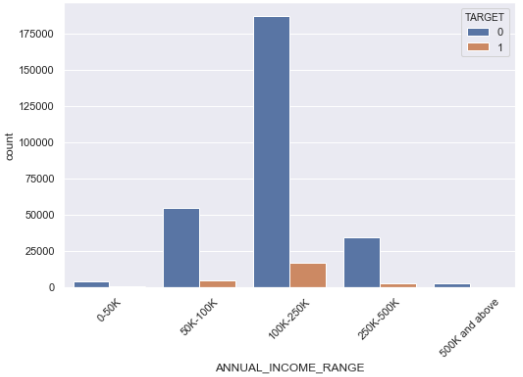
It is clearly seen that applicants who are unmarried have less non-defaulters than married.

Education



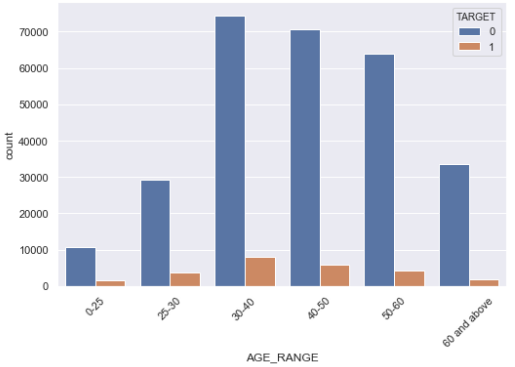
For Education type, it is noticed that as the level of education increases, there are less chances of people getting defaulted.

Annual Income Range



It is observed that applicants having Annual income of 100-250K have more chances of getting non-defaulted.

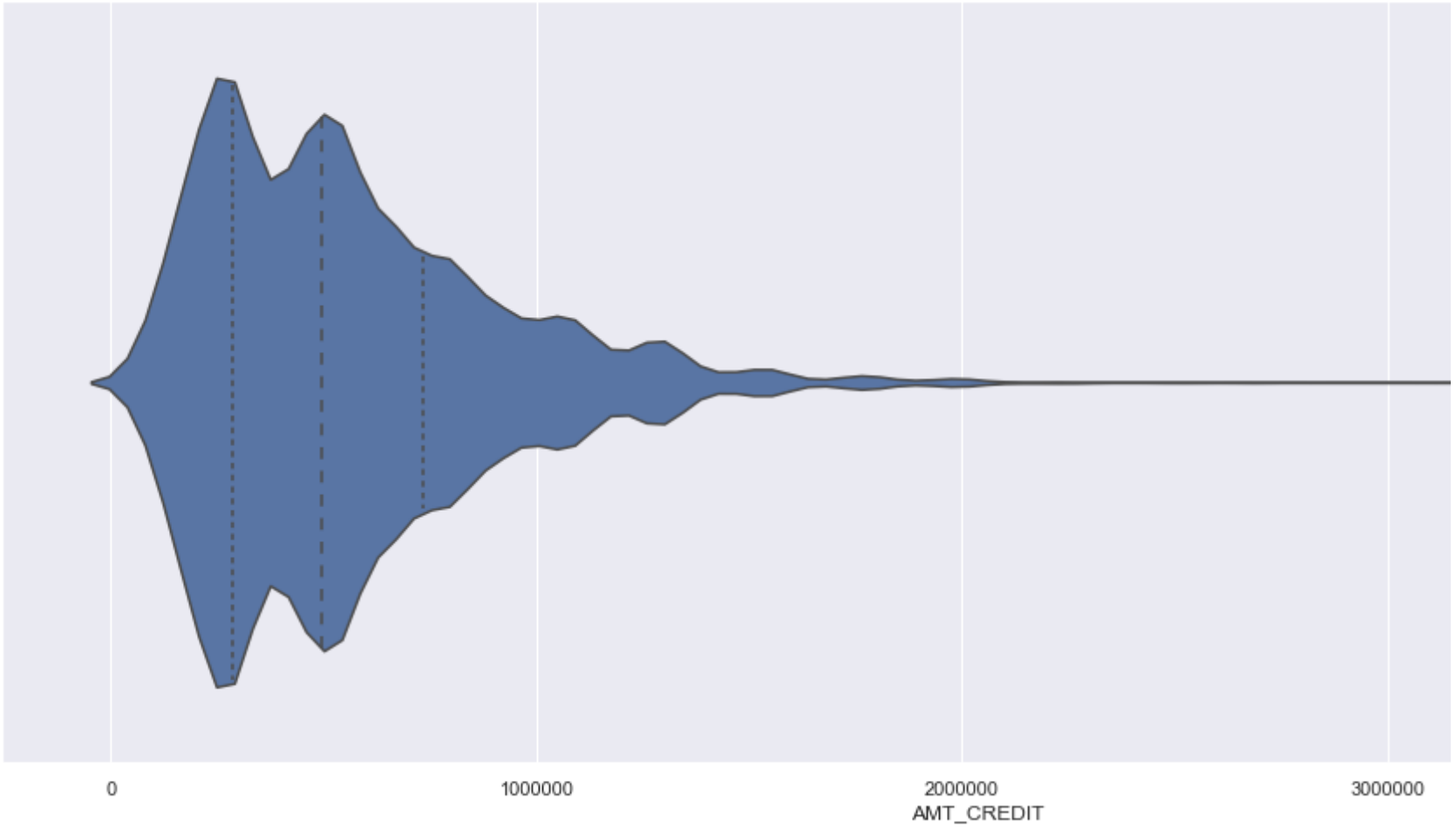
Age Range

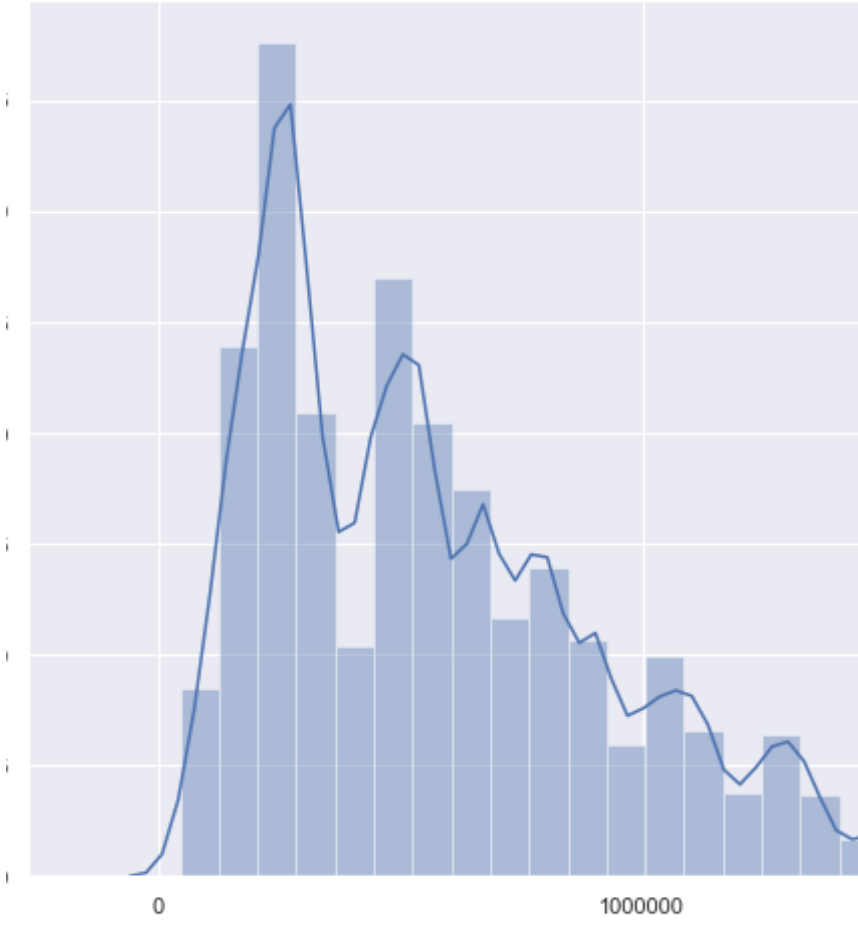


Age Range plot resembles a normal distribution, having range of 30-50 years with maximum Non-defaulters and defaulters.

3. Univariate Analysis (categorical & continuous binned variables)

CONTINUOUS VARIABLES ARE BINNED TO FORM A CATEGORICAL VIEW FOR BETTER ANALYSIS

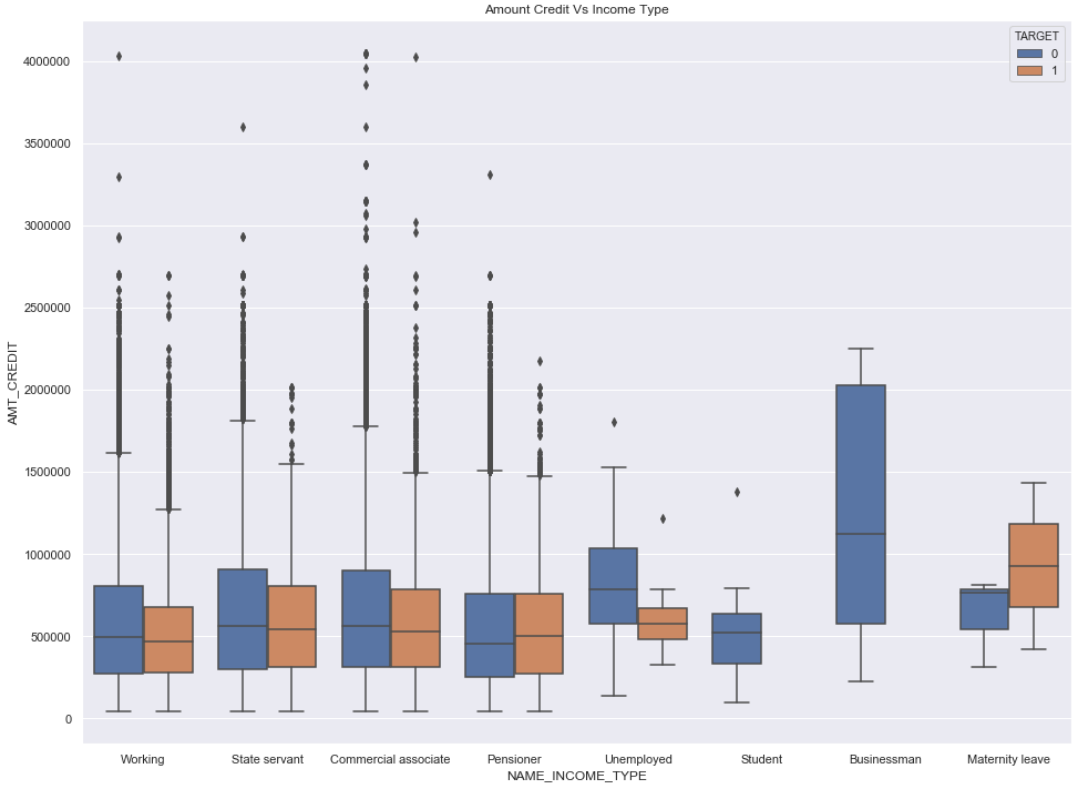


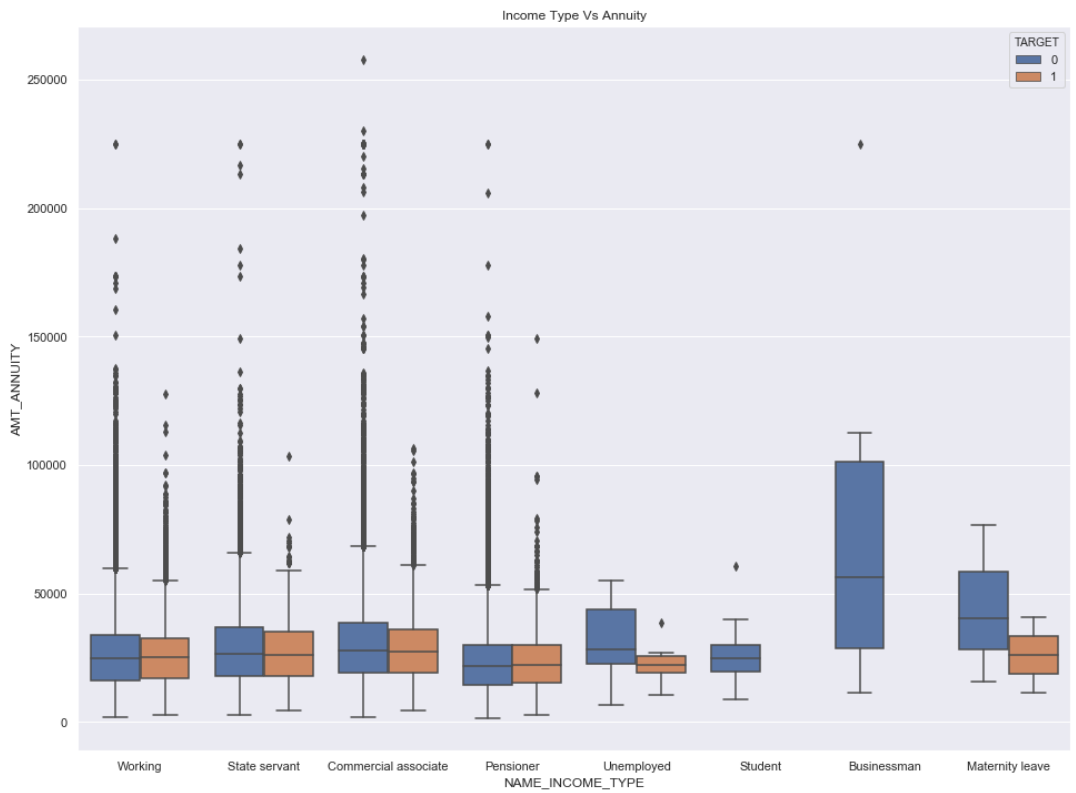


4. Univariate Analysis (continuous variables)

A **VIOLIN PLOT** HAS BEEN USED WHERE IT IS CLEARLY OBSERVED THAT MOST OF THE CREDITED LOAN AMOUNT ARE DISTRIBUTED BETWEEN **400K-750K$** FOR **DEFAULTERS**

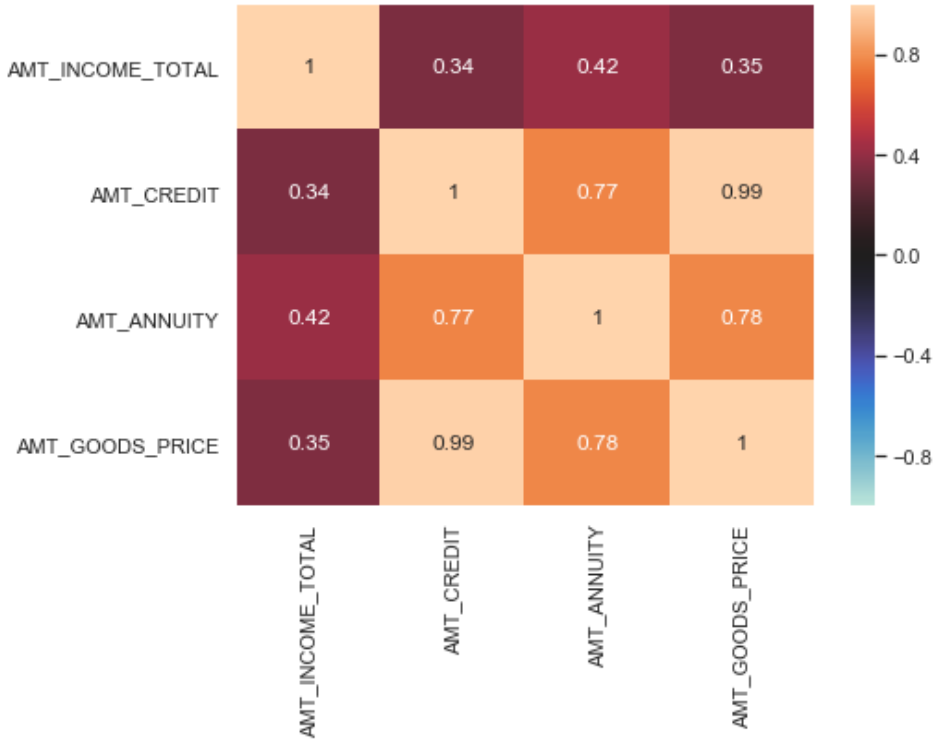
A **DIST PLOT** IS USED FOR **NON-DEFAULTERS** AND FOUND THAT MOST OF THE CREDITED LOAN AMOUNT ARE DISTRIBUTED BETWEEN **450K-800K$**.



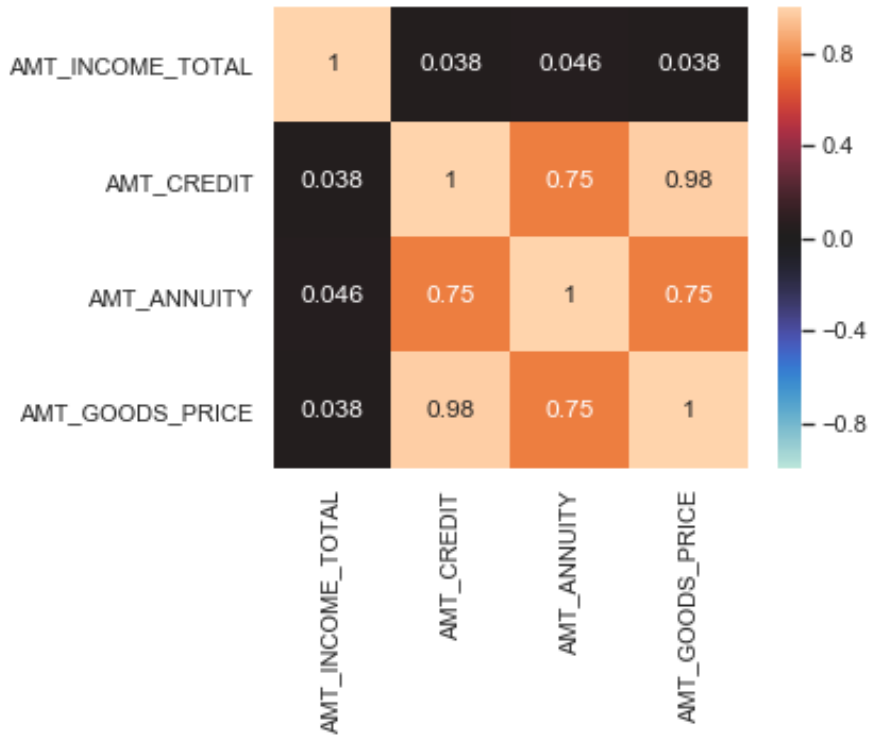


5. Univariate Analysis (continuous v/s categorical)

A **BOX PLOT** IS USED TO VISUALIZE **INCOME TYPE** OF APPLICANT WITH **AMOUNT CREDITED** AND **ANNUITY** FOR **DEFAULTERS** AND **NON DEFAULTERS**. FOR **MATERNITY LEAVE**, IT IS SEEN THAT THE LESSER THE AMOUNT OF **ANNUITY**, MORE ARE THE CHANCES OF GETTING **DEFAULT** AND **STUDENTS** AND **BUSINESSMAN** ARE VERY UNLIKELY TO GET **DEFAULTED**.

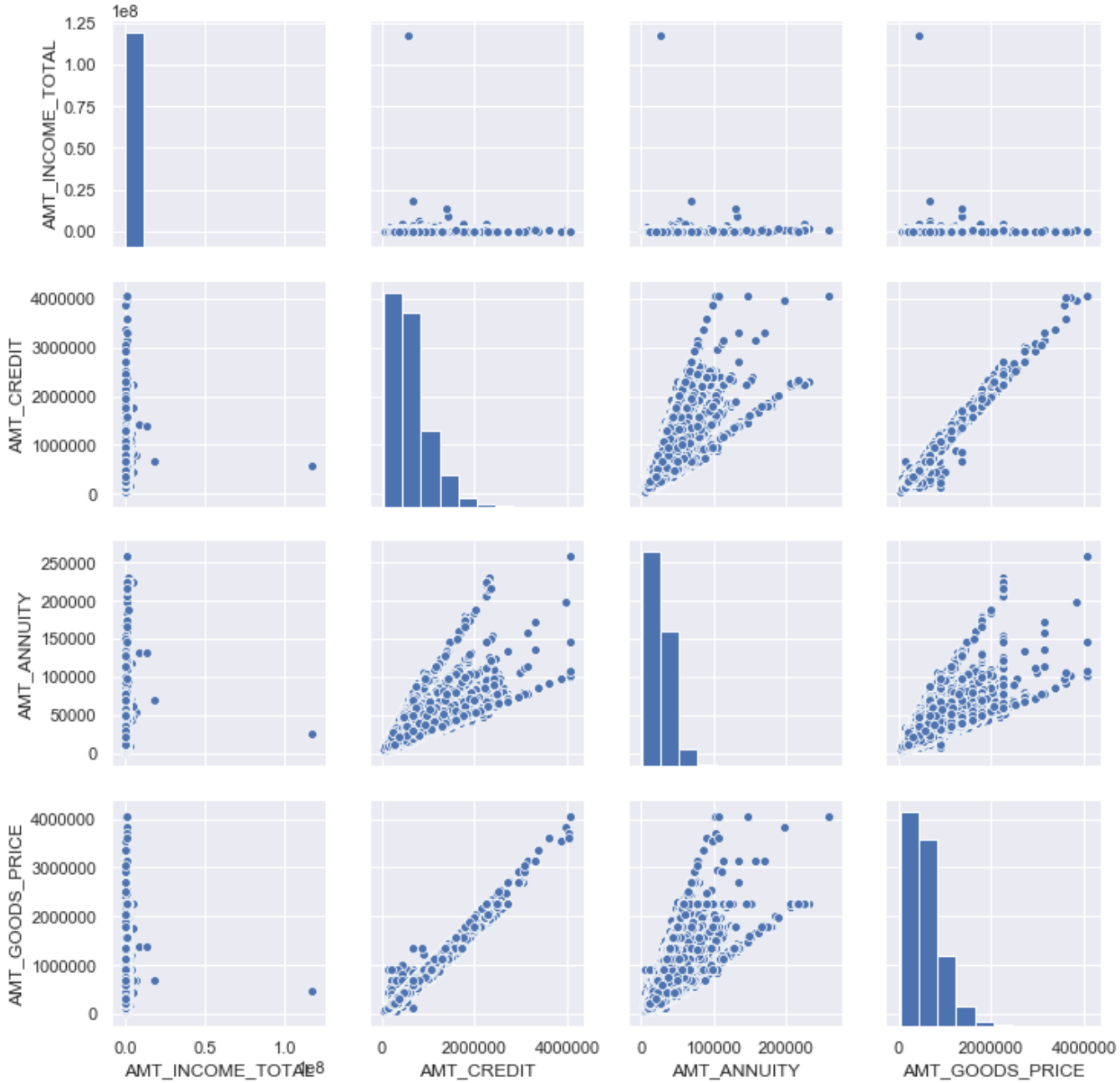


For Non-Defaulters, it is again observed that Amounts Good price has the highest level of correlation with Amount Credit whereas Amount Income total has moderate level of correlation with other 3 variables.



For Defaulters, it is observed that Amounts Good price has the highest level of correlation with Amount Credit whereas Amount Income total has lowest level of correlation with every other variables.

1. Multivariate Analysis (correlation) – Using Heat Map

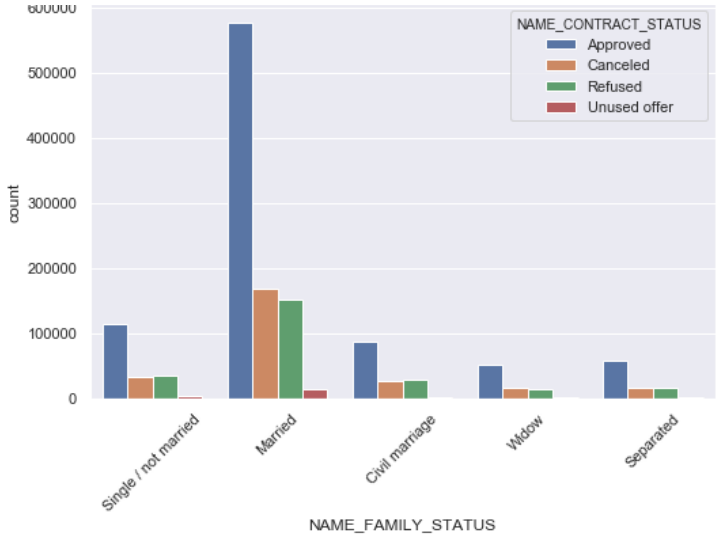


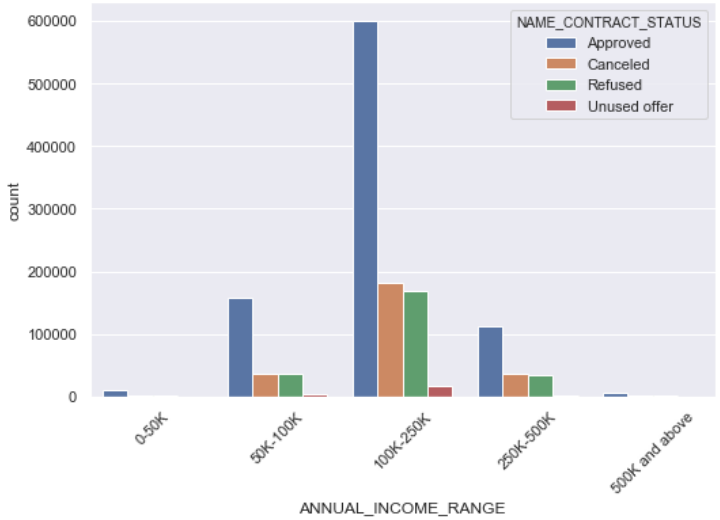
2. Multivariate Analysis (correlation)

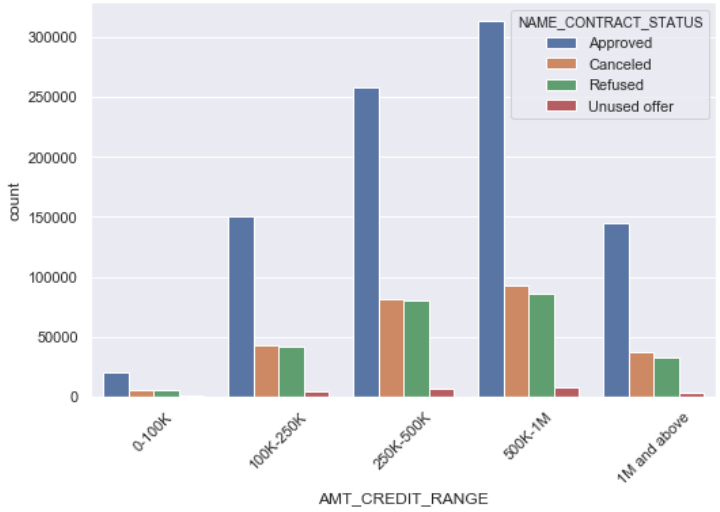
A **PAIR PLOT** IS USED TO FIND A CORRELATION BETWEEN MULTIPLE CONTINUOUS VARIABLES LIKE **ANNUAL INCOME, AMOUNT CREDITED, ANNUITY AND GOODS PRICES**.

IT IS EVIDENT **FROM THE PLOTTING THAT AMOUNT CREDIT, AMOUNT ANNUITY AND GOOD PRICES** ARE **LEFT SKEWED** AND NOT NORMALLY DISTRIBUTED

AMOUNT CREDIT VS **GOODS PRICES** FOLLOWS A LINEAR CURVE



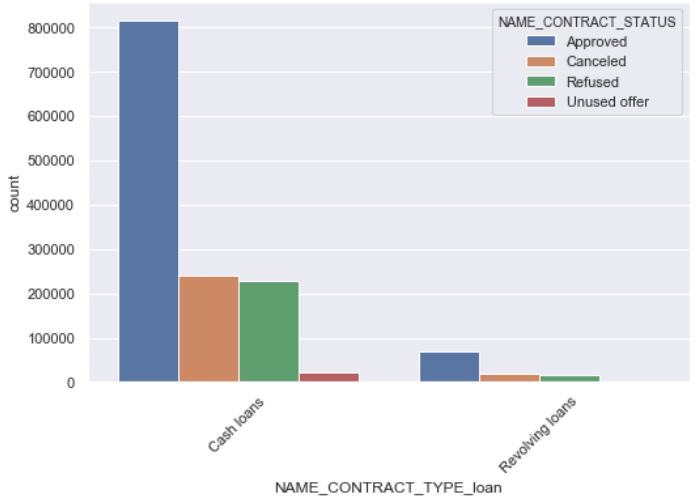


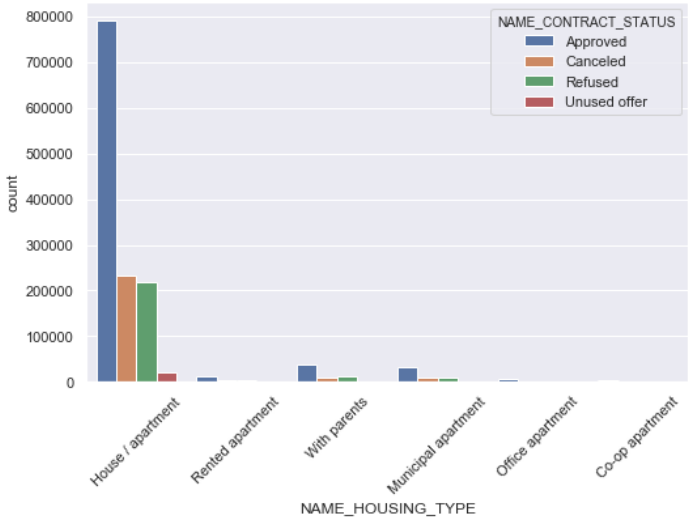


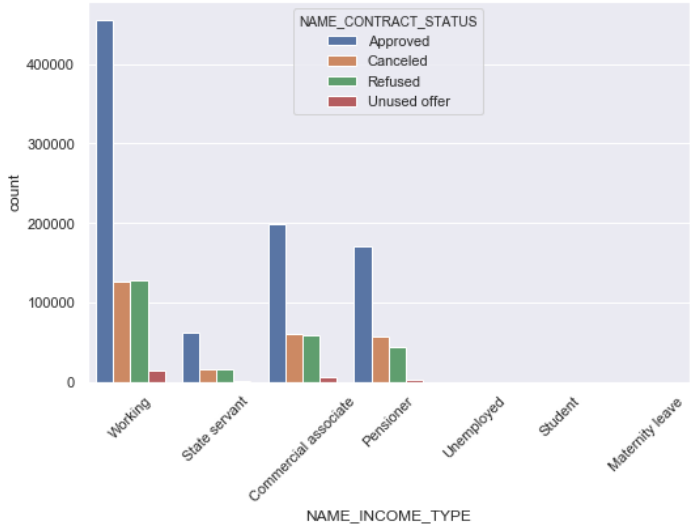
Merging Previous Datasets

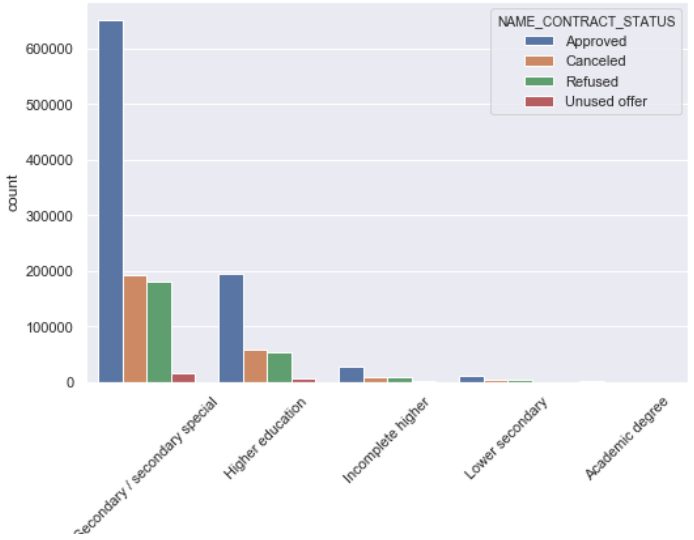
Loan dataset is merged with **Previous application dataset**. This was tricky as there were multiple entries for a single **SK_CURR_ID** in the previous application dataset.

After merging, analysis is performed with respect to **NAME_CONTRACT_STATUS** variable.









Few Analysis on Merged Dataset

- It is observed that number of Canceled and Refused loan applications are almost same in case of Cash loans, and chances of loans getting approved is higher in case of Cash Loans.
- In case of Family Status, chances of Loan getting approved is higher if the user is married
- For Age Range, it is observed that chances of loans approving are higher if a user is in range of 30-40 years and lowest if applicant is under 25.
- For Annual Income, it is seen that, for Applicant having salary of 100-250K, chances of loans getting Approved is higher

Conclusion

Variables to look out for:

Age Range - Chances of loans approving are higher if a user is in range of 30-40 years and chances of loan being approved is lowest if age is under 25 and above 50 years.

Income Type - It was seen that all loans can be approved for Student and Businessman Income Types, as there are 0 defaulters. For Unemployed type, lower the credit amount, more the defaults, so bank should be careful while approving their loans.

Amount Credit - Chances of loan getting approved are higher if Amount Credit for the applicant is low.

Family Status - Chances of loan getting approved is higher if the user is Married.

Annual Income - For Applicant having salary of 100-250K, chances of loans getting Approved is higher.