Credit EDA Assignment

Problem Statement:

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

We have Two data sets namely 'application_data.csv' and 'previous_application.csv'. I will analyze the data to find the driving factors which are the indicators of becoming a default.

Application Data:

Observations:

- There are 122 columns and 307511 rows available in the data set.
- There are Many null values in the data.
- And there are some columns starts with "DAYS" are showing negative values which are incorrect.
- I will drop some columns which are not very useful for analysis.
- I will also impute the appropriate value in place of Null values.
- I will drop the columns which are having more then 40% of missing data.

Assumptions and Approach:

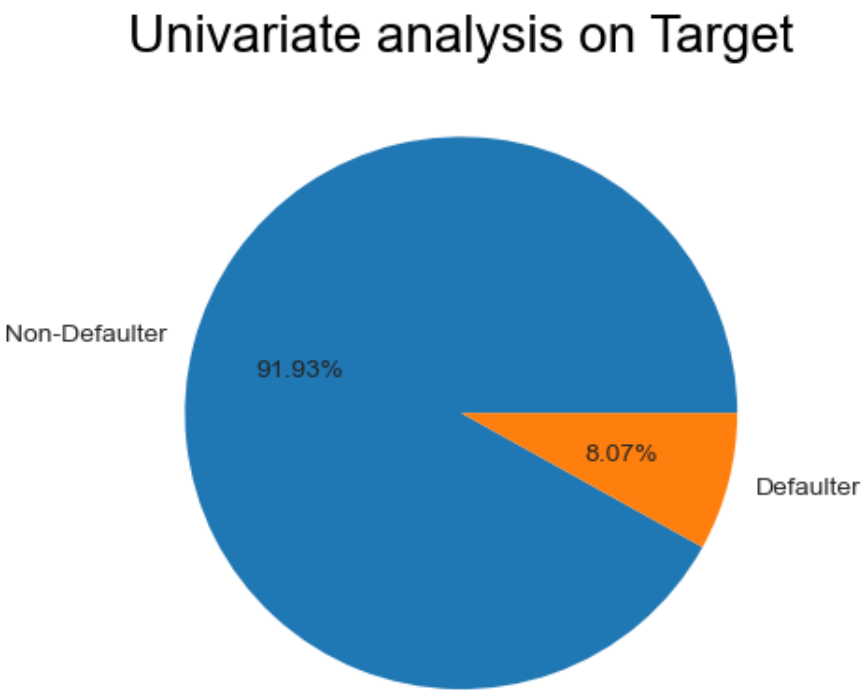
- By observing the missing data under the OCCUPATION_TYPE column, there are different income type applicants available. In this case we cannot impute the missing value with "MODE". The mode value here in this case is "Laborers".
- I am assuming that there are mixed income type of the missing data under OCCUPATION_TYPE, so I will mark this as a separate variable as "Others".
- EXT_SOURCE_3 Data is missing here is not at random. I am imputing the median value for missing data for better analysis.
- NAME_TYPE_SUITE is a categorical data type, imputing the missing values with "mode" is the best way for analysis.

Continued....

- Missing data in AMT_ANNUITY column are having AMT_CREDIT details and almost all the applicants NAME_TYPE_SUITE is "Unaccompanied".
- As we can observe that each one having the AMT_CREDIT details, there is high chances of AMT_ANNUITY. So I will impute the missing value with "median".
- There are 2 null values in CNT_FAM_MEMBERS. NAME_FAMILY_STATUS for the both the null values is Unknown and NAME_TYPE_SUITE is mentioned as Unaccompanied.
- We can assume that applicant might be single but we will try to observe the mean, median, mode under this CNT_FAM_MEMBERS,. Mean, median and mode value is same which is "2" so we will impute median value in place of missing value.
- There is only one null value in DAYS_LAST_PHONE_CHANGE column. This null value may be missing at random. I will impute median value to analyze the data.
- Converting 'DAYS_BIRTH‘ and 'DAYS_EMPLOYED' to years and creating new columns "Age" and "Experience". This is done for better understanding and analysis.

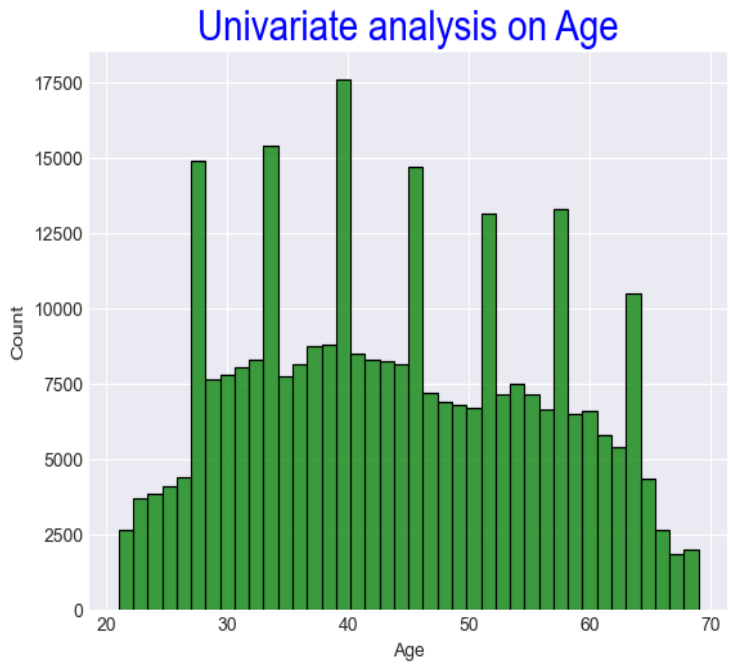
Data Insights

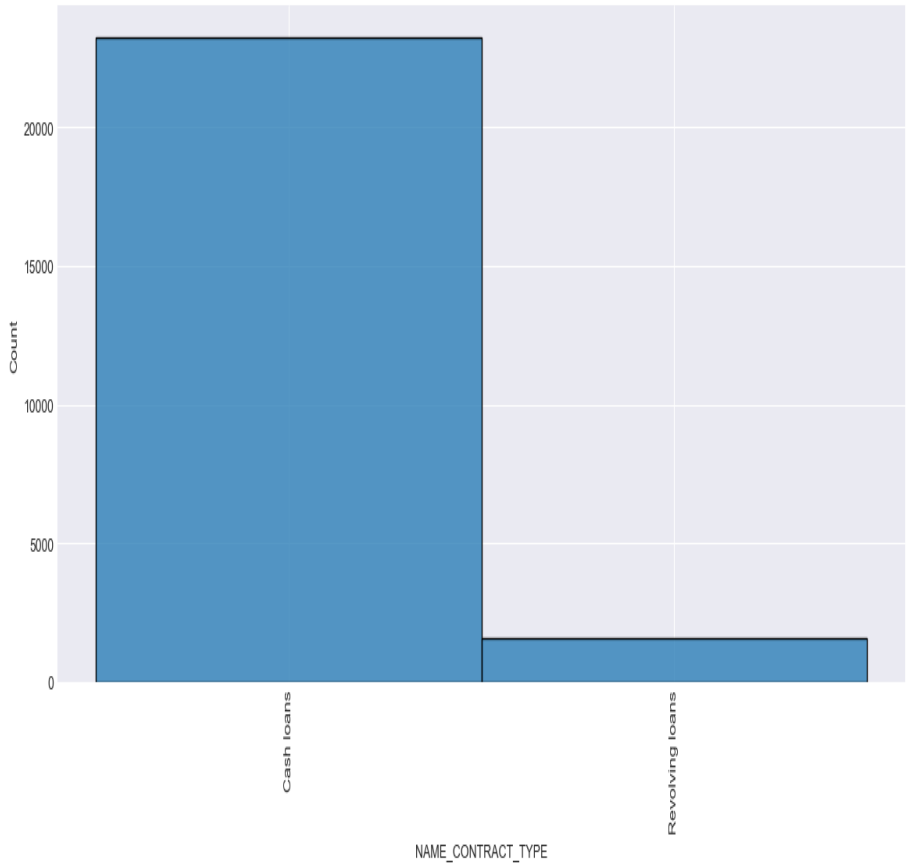
- There is data imbalance in the given data. We can observe that defaulters percentage is 8.07% and non defaulters are 91.93%. We can observe that defaulters percentage is 8.07% and non defaulters are 91.93%.



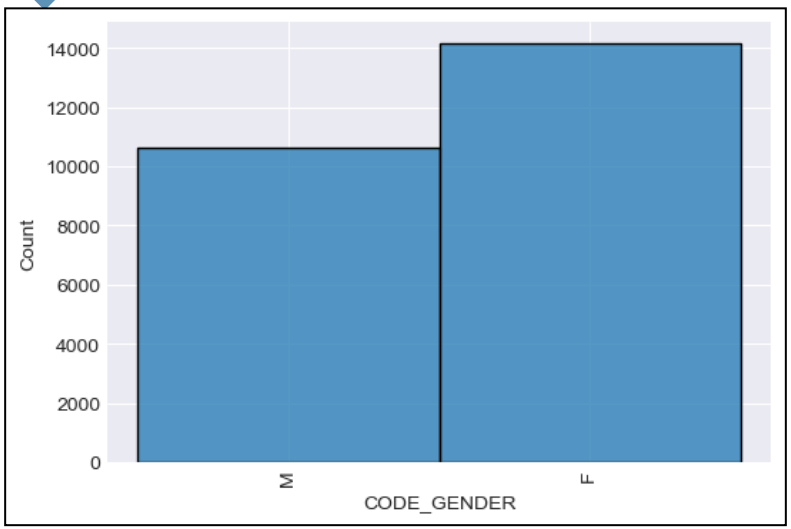
Age 40 are applied the most for the loan.

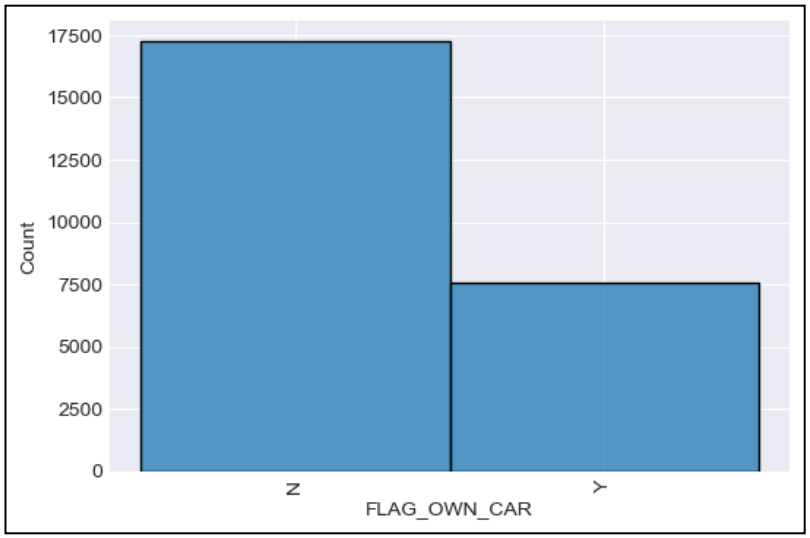
There are high number applicants who are applied for cash loans have payment difficulties.

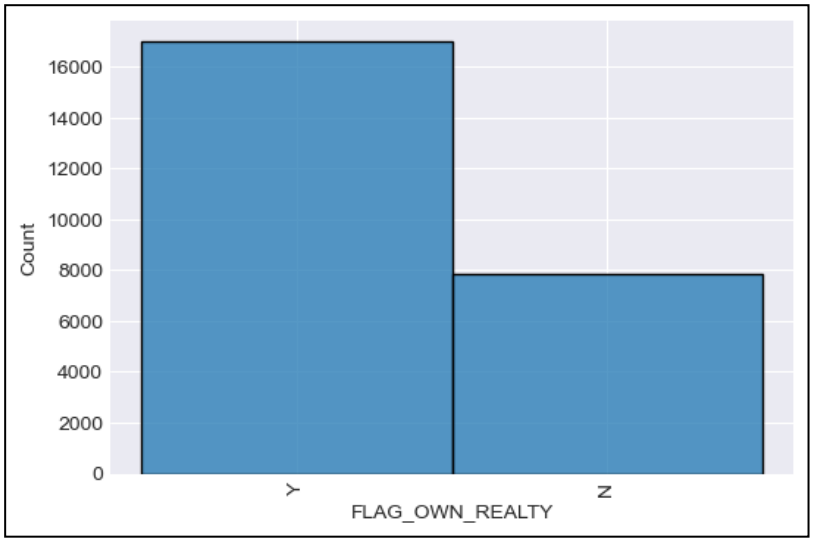


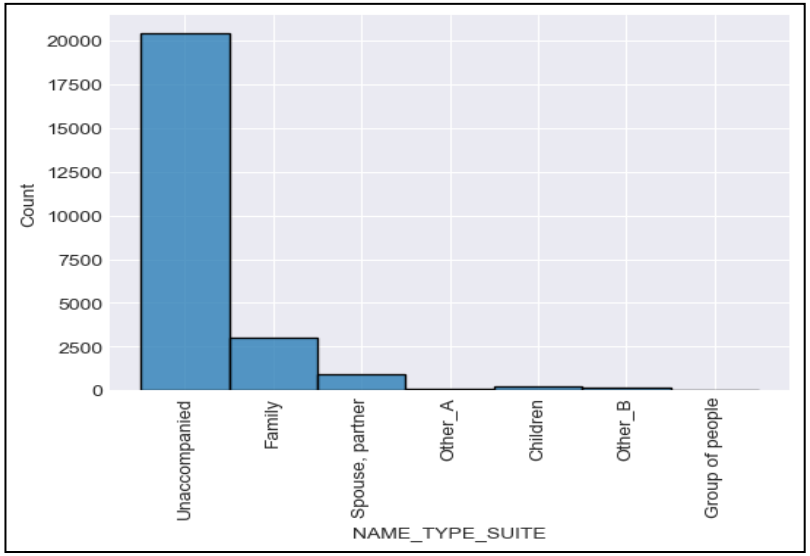


- Female applicants are comparatively more in number in terms for payment difficulties.
- People who are not having the own reality and car are having more payment difficulties then people who are having the same.
- Applicants who are Unaccompanied while applying the loan are more high in number and they are having payment difficulties.



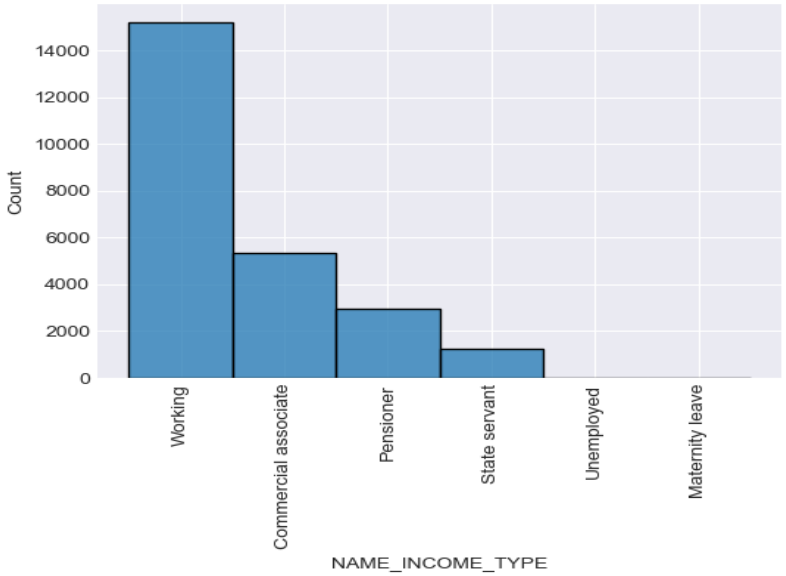


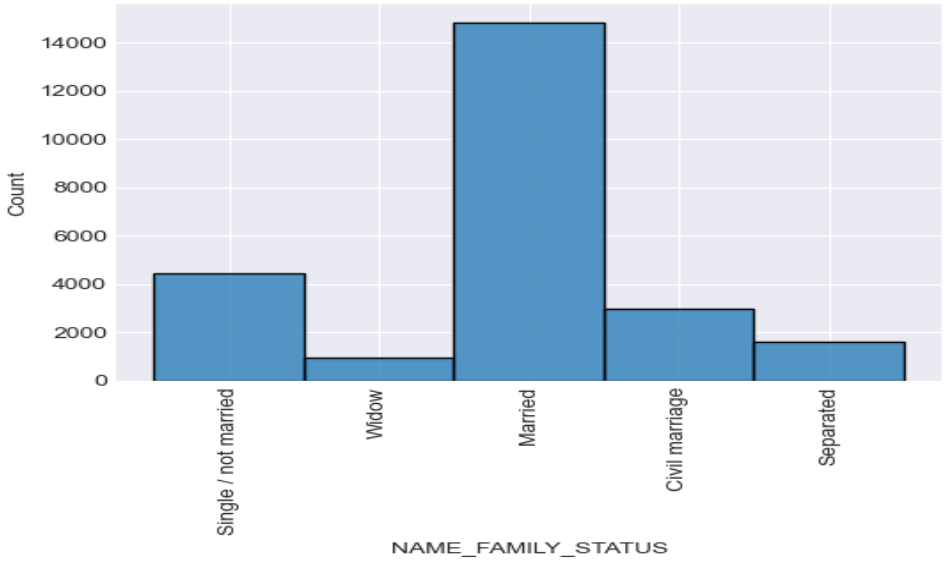


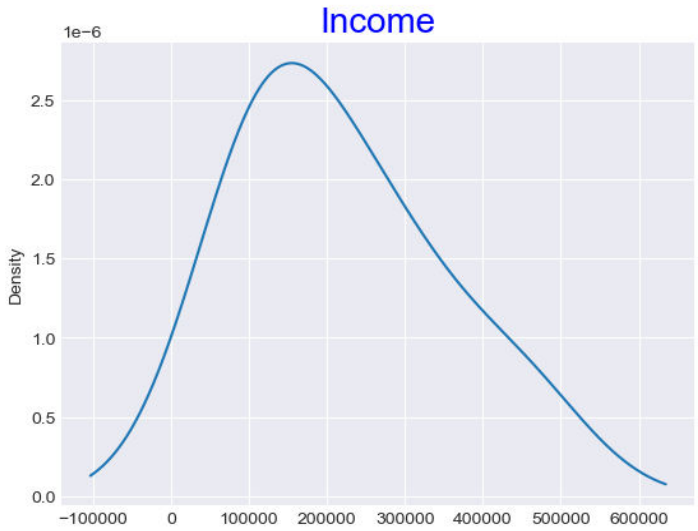


Working professional, people with secondary education and married people are having more payment difficulties.

Applicants whose income between 1,00,000 to 2,00,000 are highly facing payment difficulties.

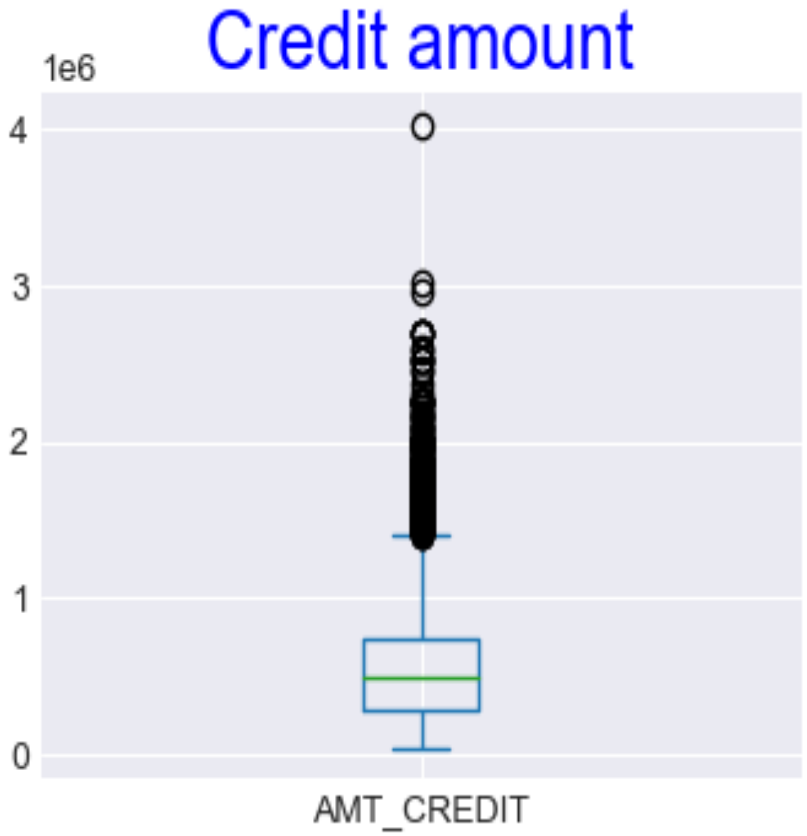


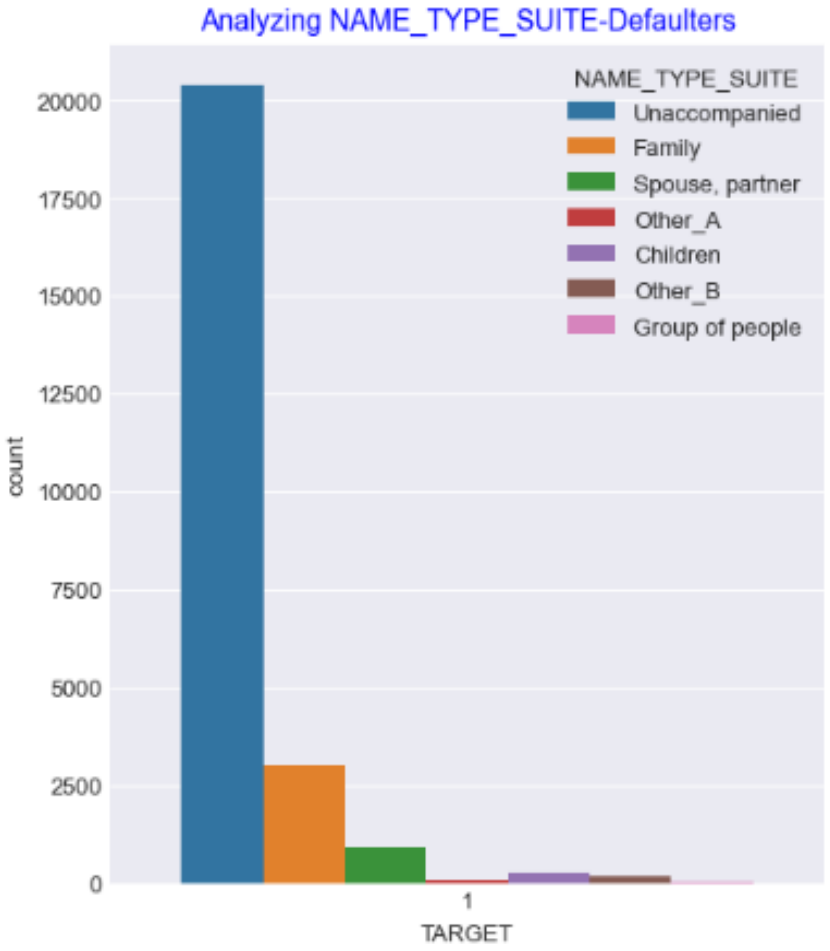




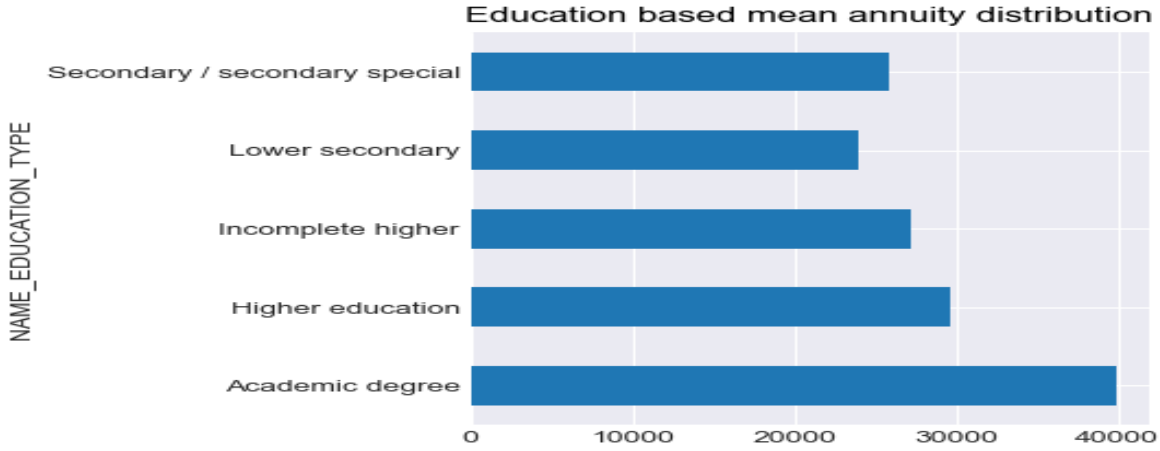
Credit amount have some outliers.

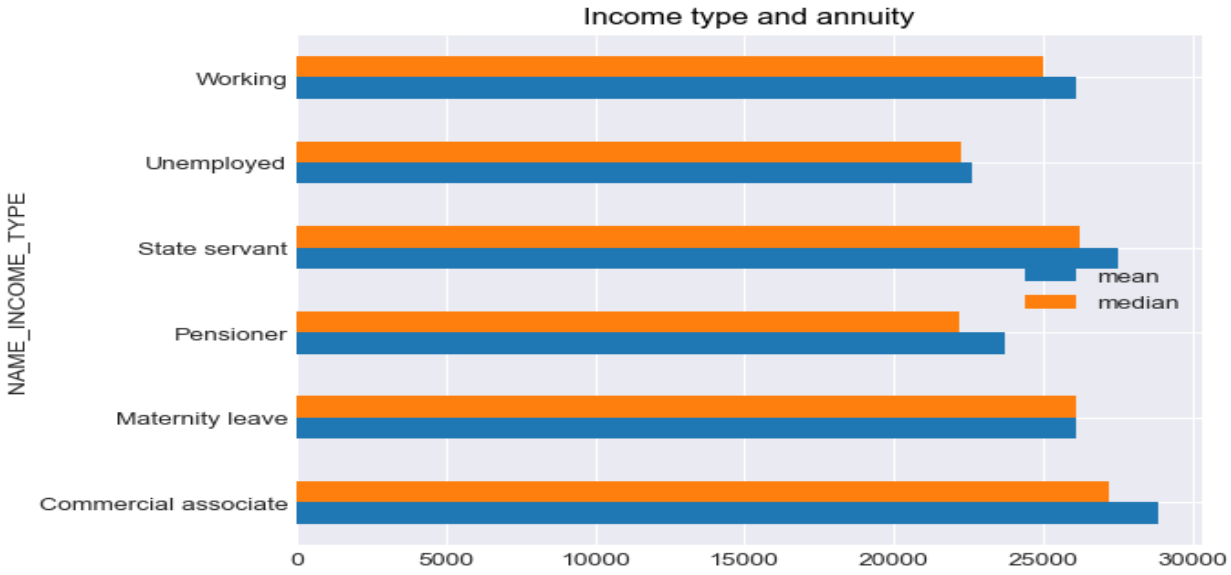
Clients who are unaccompanied while applying the loan are more in number and having payment difficulties.





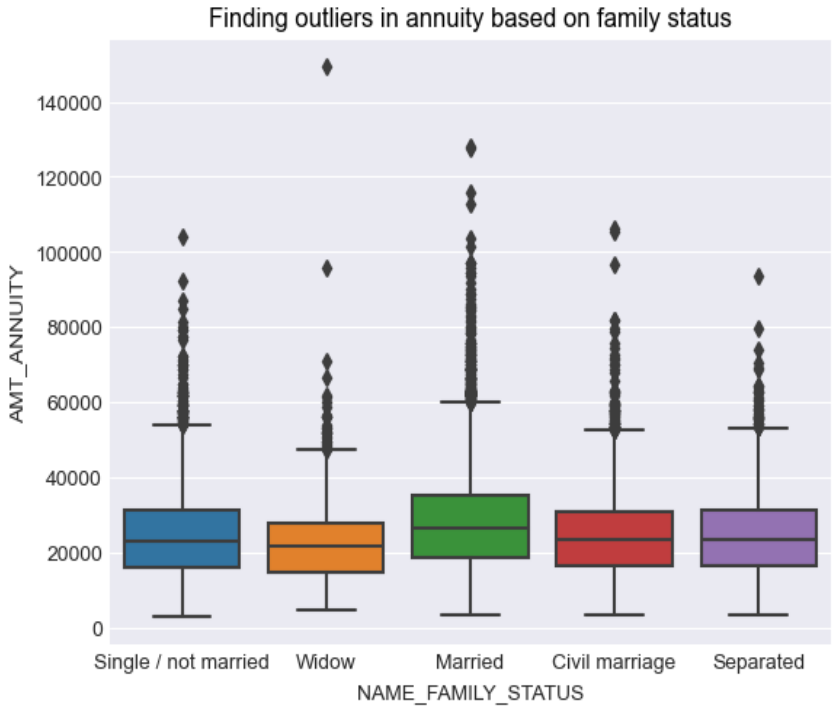
- Education based mean annuity distribution
- Income type and annuity

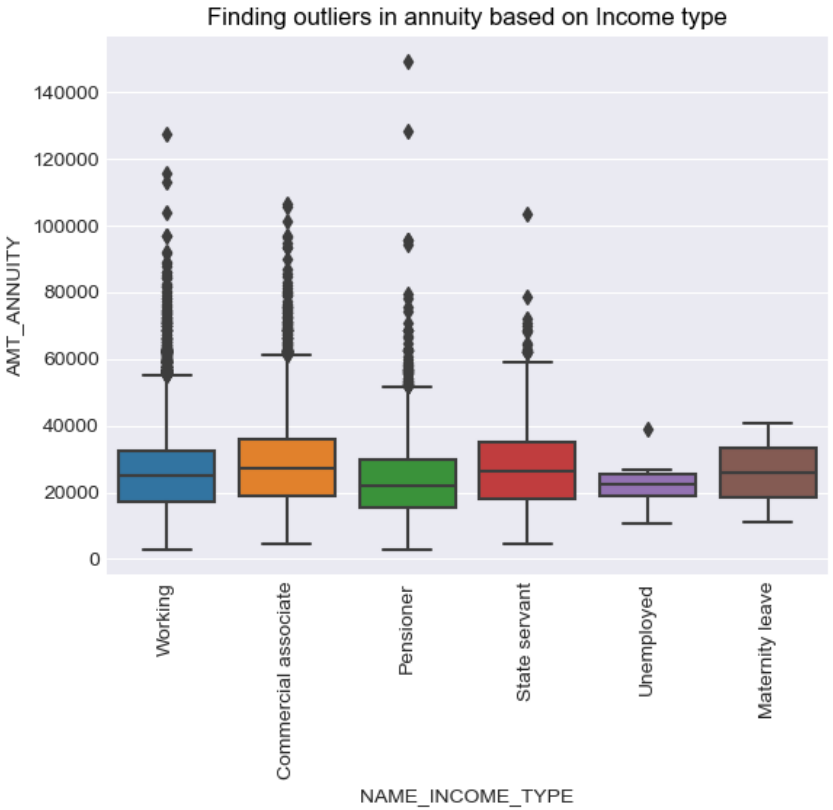




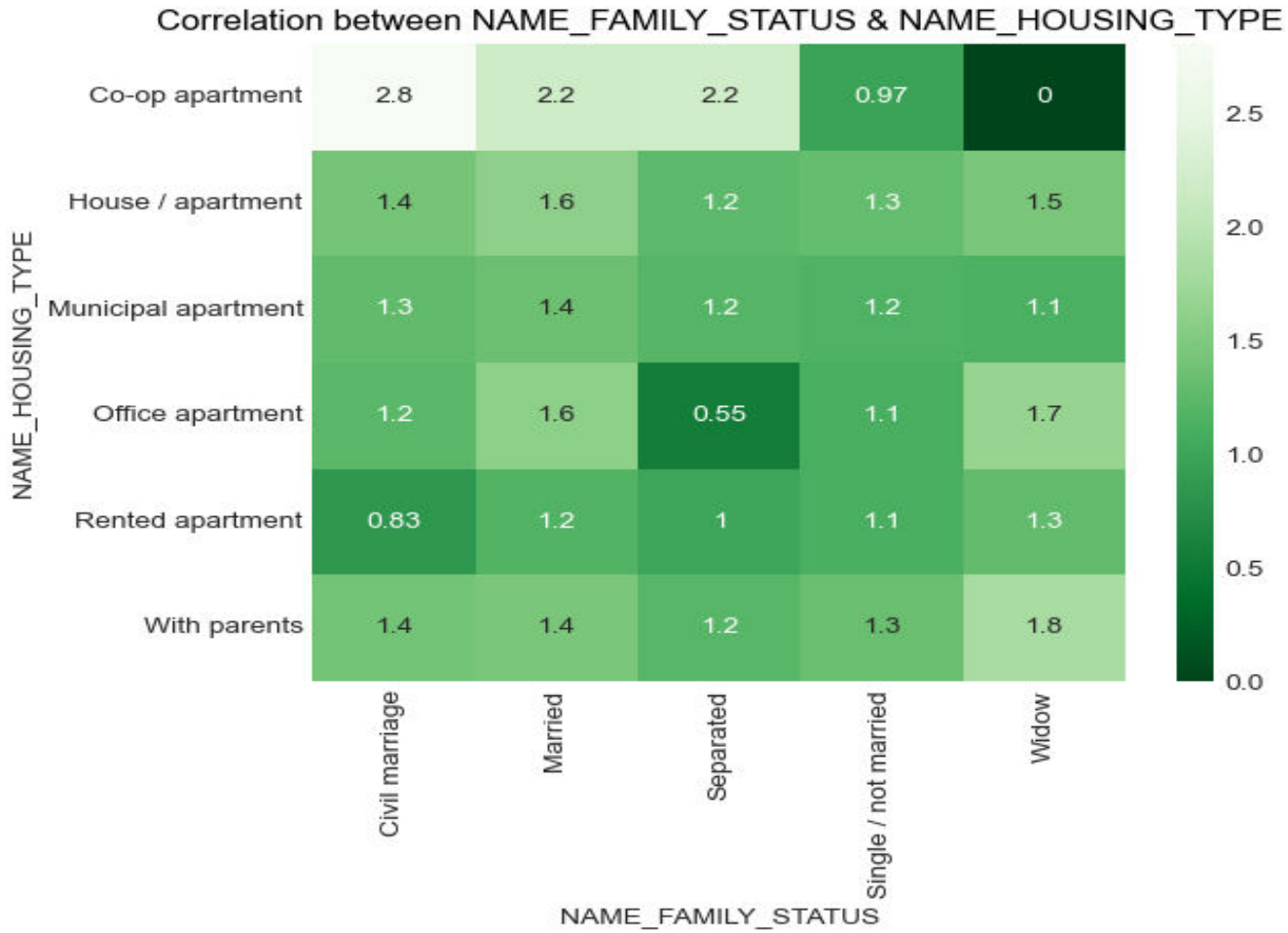
Outliers in Annuity based on Family status

Outliers in annuity based on income type





Correlation between NAME_FAMILY_STATUS & NAME_HOUSING_TYPE



FINDINGS

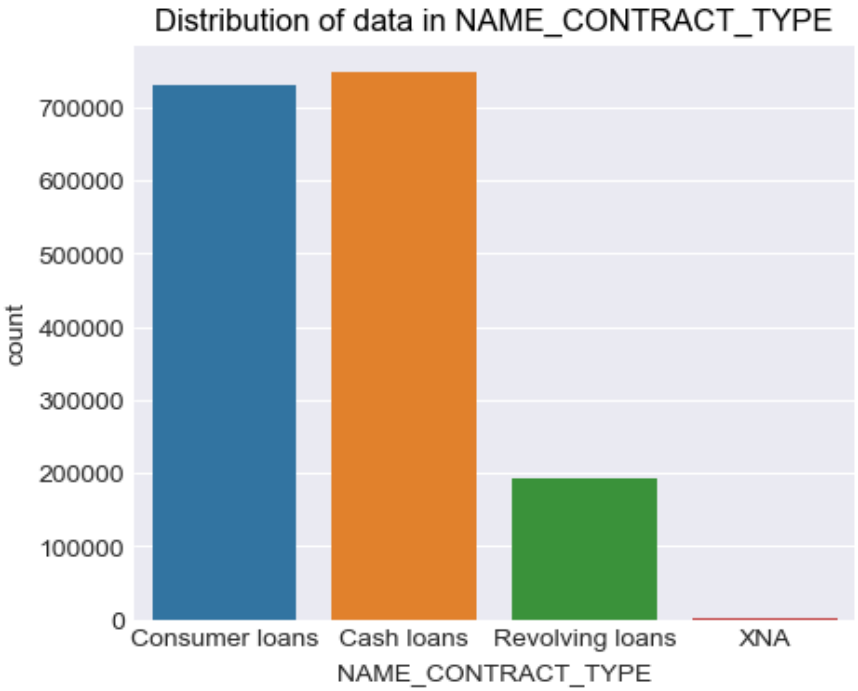
- We have data imbalance in the given data set, We can observe that defaulters percentage is 8.07% and non defaulters are 91.93%.
- We can observe that people with age 40 are applied the most for the loan.
- There are high number applicants who are applied for cash loans have payment difficulties.
- Female applicants are comparatively more in number in terms for payment difficulties.
- People who are not having the own reality and car are having more payment difficulties then people who are having the same.
- Applicants who are Unaccompanied while applying the loan are more high in number and they are having payment difficulties.
- Working professional, people with secondary education and married people are having more payment difficulties.
- Applicants whose income between 1,00,000 to 2,00,000 are highly facing payment difficulties.
- Applicants with 0 to 1 year experience are facing more payment difficulties.
- Commercial associates and working people mean income is more then the others.
- People with academic degree are having high income.
- Mean total of annuity is high for cash loans.
- Married applicants have high annuity.
- People live in Co-op apartment and civil married are having high 60 DPD (days past due) default.

Previous Application Data: Observations And Assumptions:

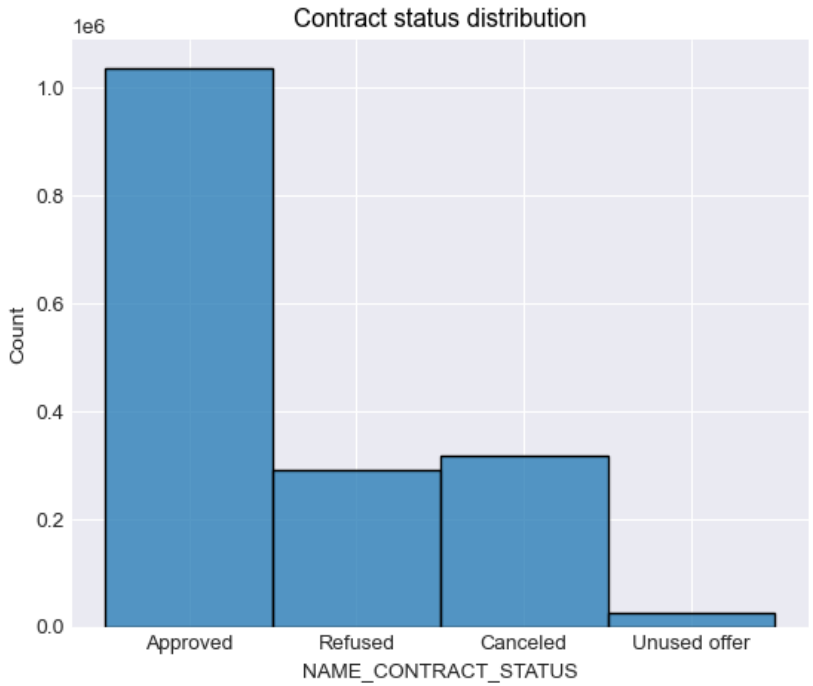
- There are 1670214 rows and 37 columns in the previous application data.
- There are huge null values in the previous application data set.
- There are some columns with more then 40% of missing values. I will drop the columns with missing values more than 40%. These missing values will affect our analysis.
- DAYS_DECISION column have some negative values which may occurred while entering the data or it can be system error. I will convert these values to absolute values for better analysis.
- There are many null values in annuity and goods price columns. There are many null values in annuity and goods price columns.
- There are many null values in annuity and goods price columns. There is a possibility that applicant is applying loan for the first time. Imputing the missing value with mean or median may affect the analysis hence I am imputing the 0("Zero") value.

Analyzing the Previous application data

- Distribution of contract type in the previous application data.

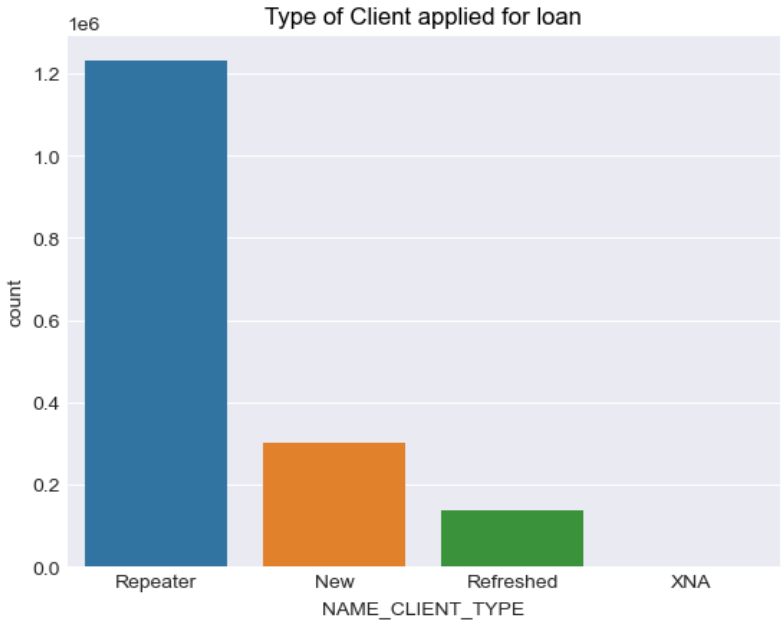


- Distribution of contract status of the application.



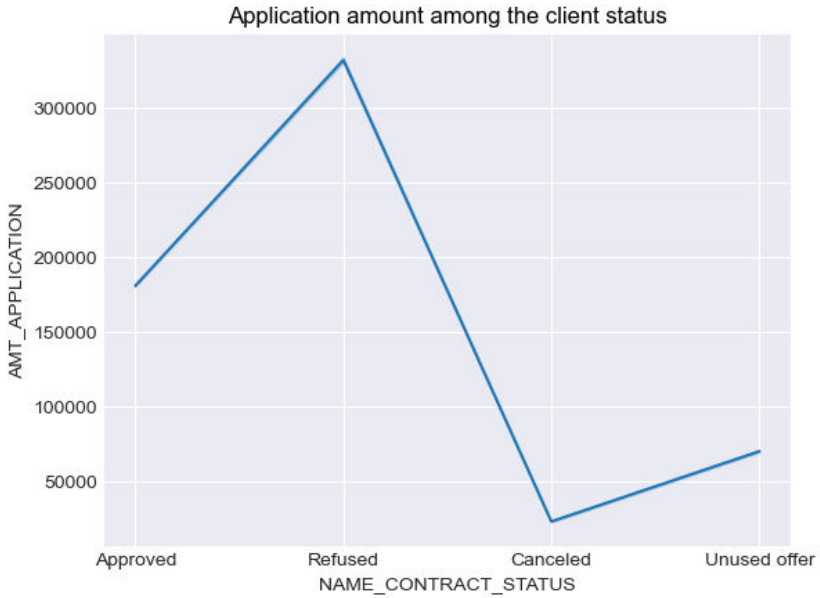
Distribution of Type of Clients applied for loan

- We can observe that Repeated clients are applying for loan.

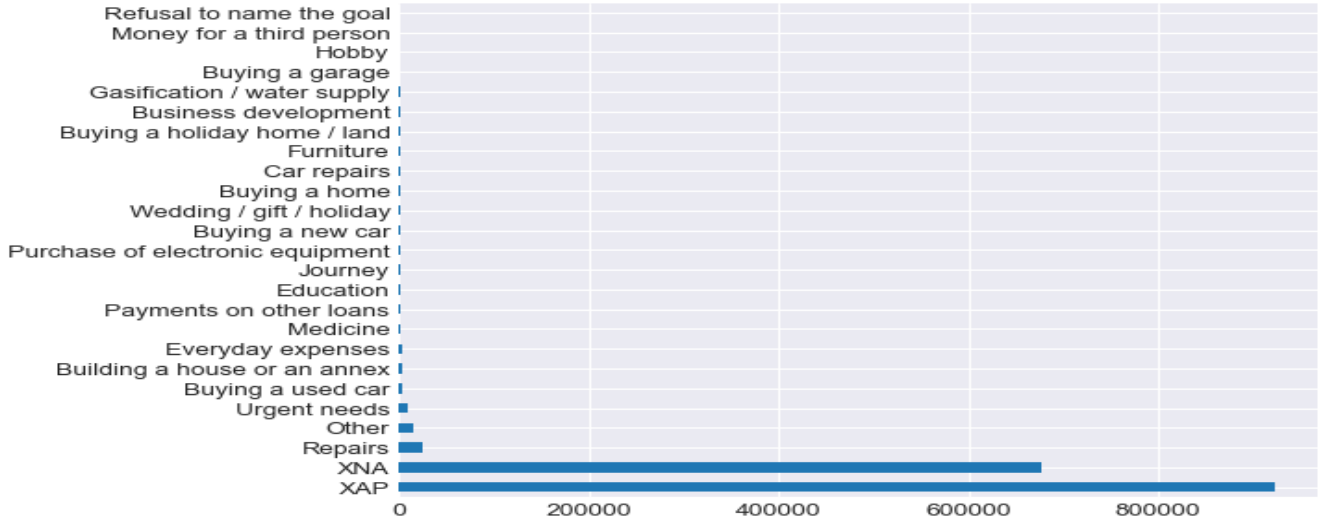


Application amount among the clients status

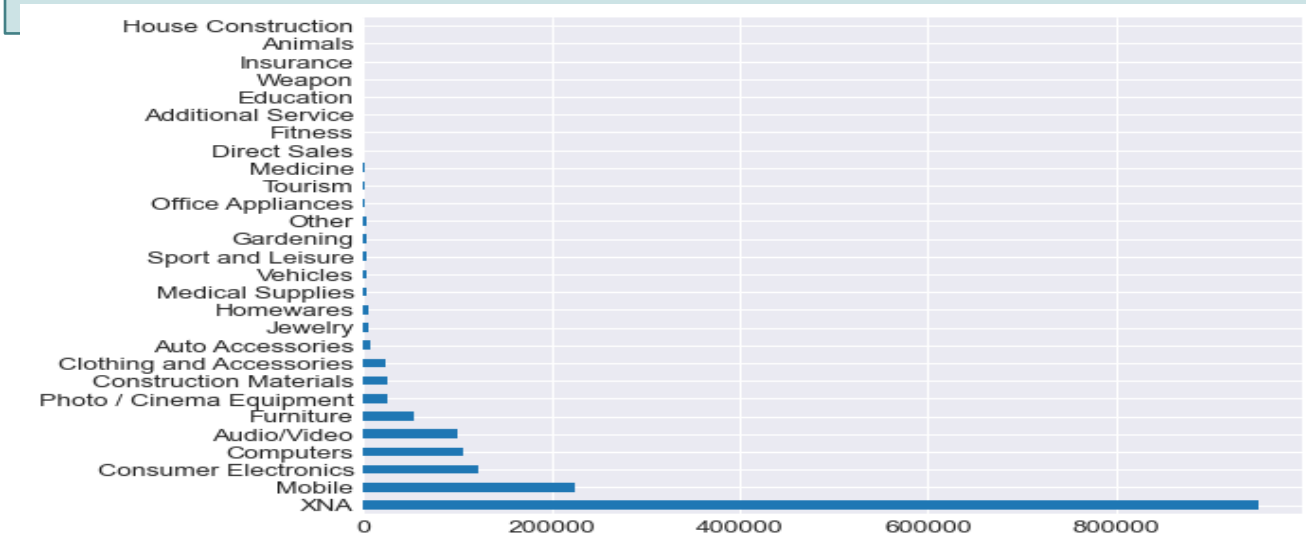
- By observing the data the highest application amount is above 3lakh which is under refused status. Highest approved amount lies between 1.5lakh to 2lakh.



Cash loan purpose distribution



Goods category values distribution

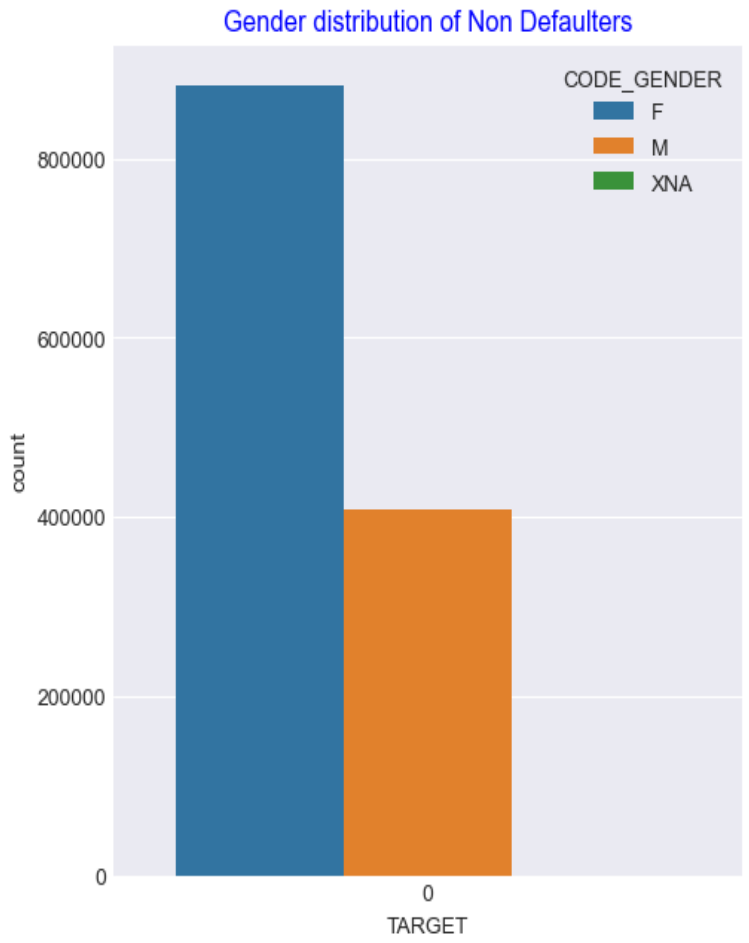


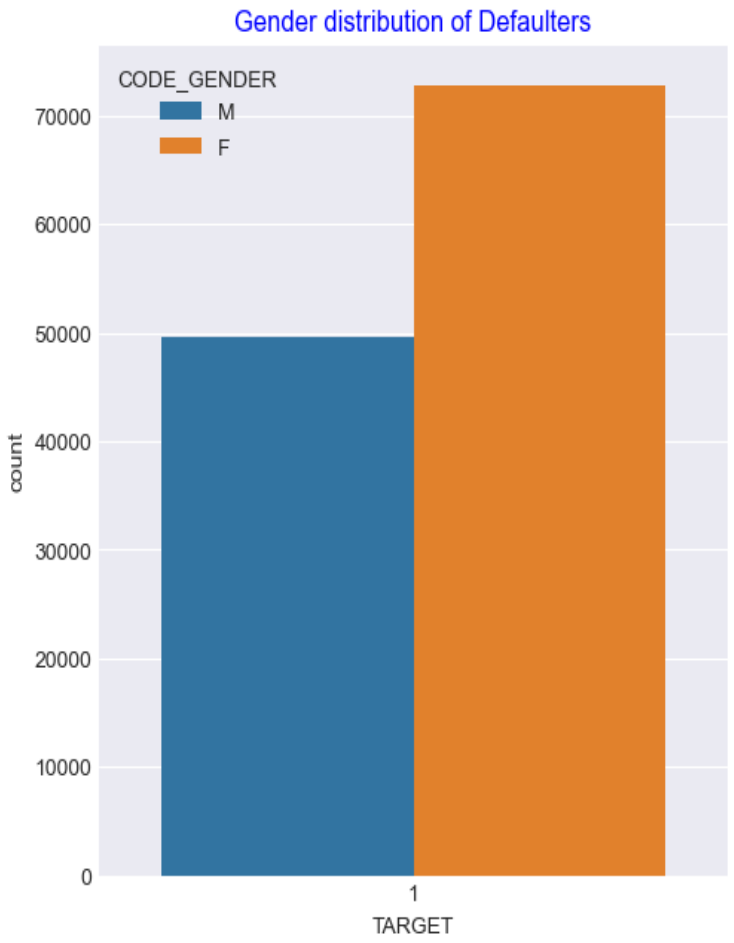
FINDINGS

- Previous application data contains 1670214 rows and 37 columns.
- There are very high number of missing values in the data.
- I have dropped columns which are having 40% and more missing values.
- Cash loans and consumer loans are high in number compared to revolving loans.
- Approved loans are high in number.
- clients who are applied for loan are mostly repeated clients.
- Most of the loan purpose is in XAP and XNA.
- Mobiles comes in second place for loan application.

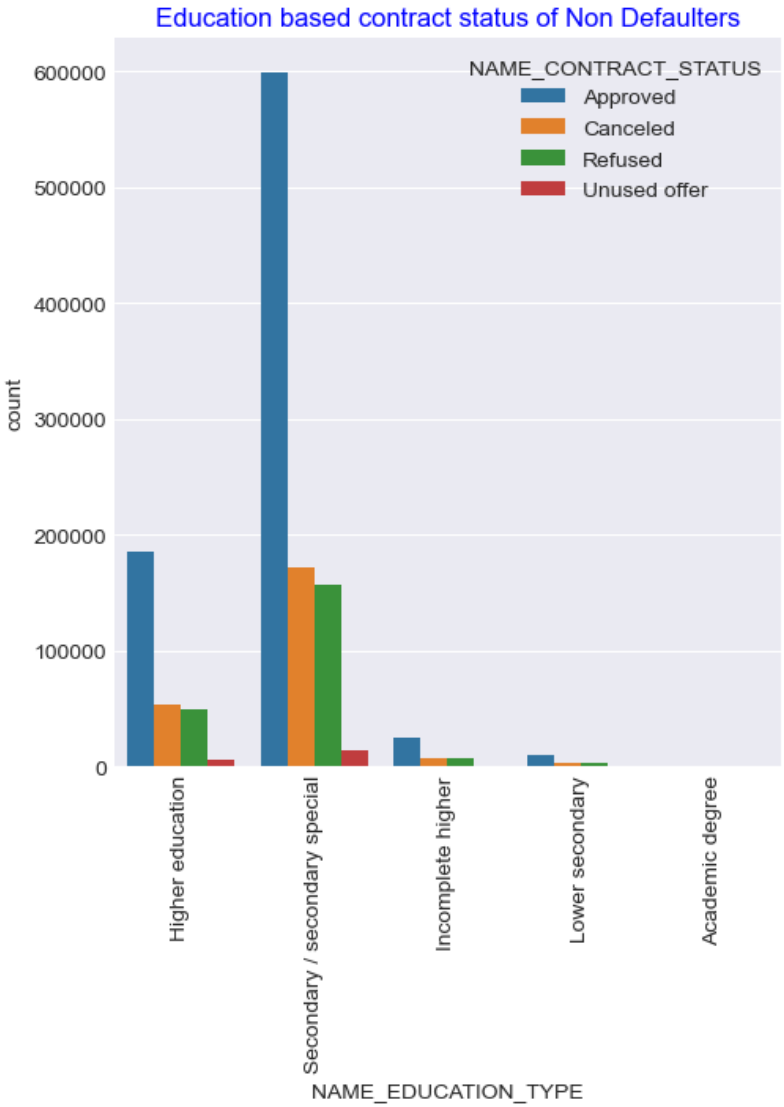
Analyzing Merged Data

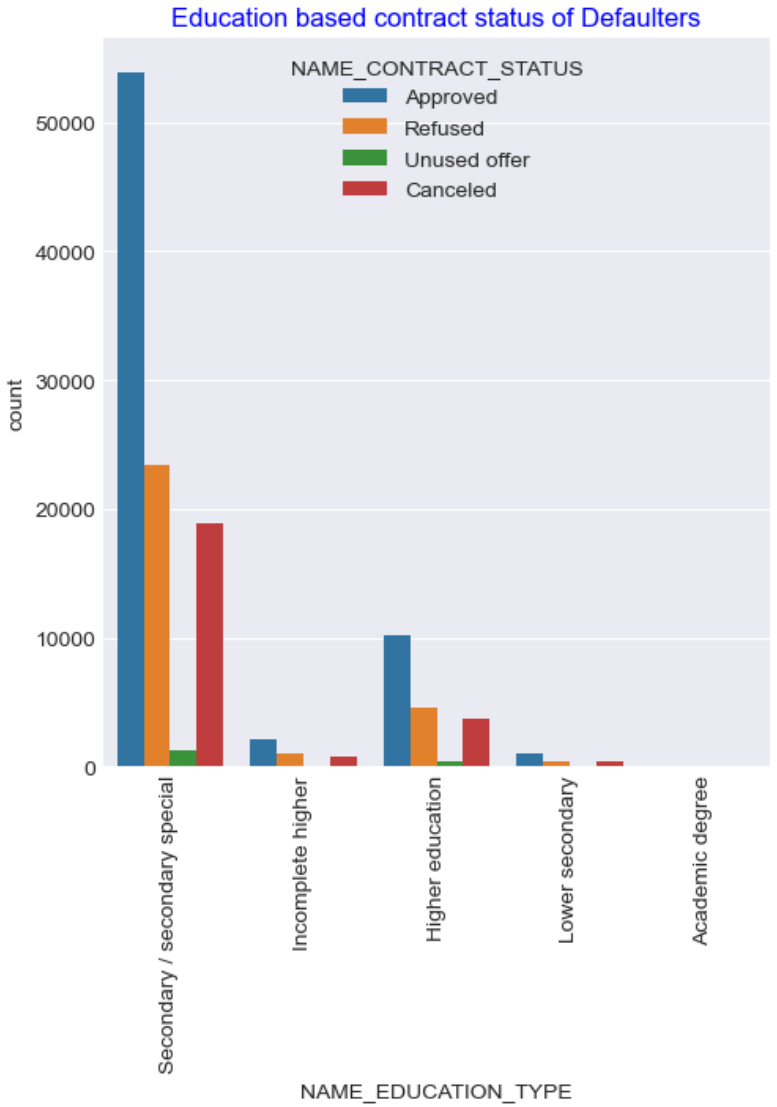
- Gender distribution of the given data.



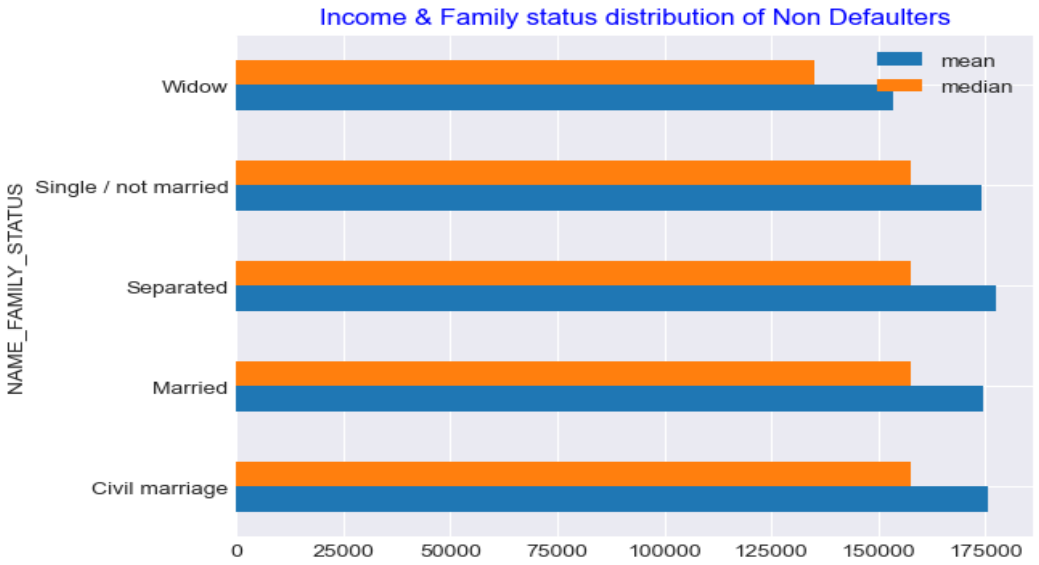


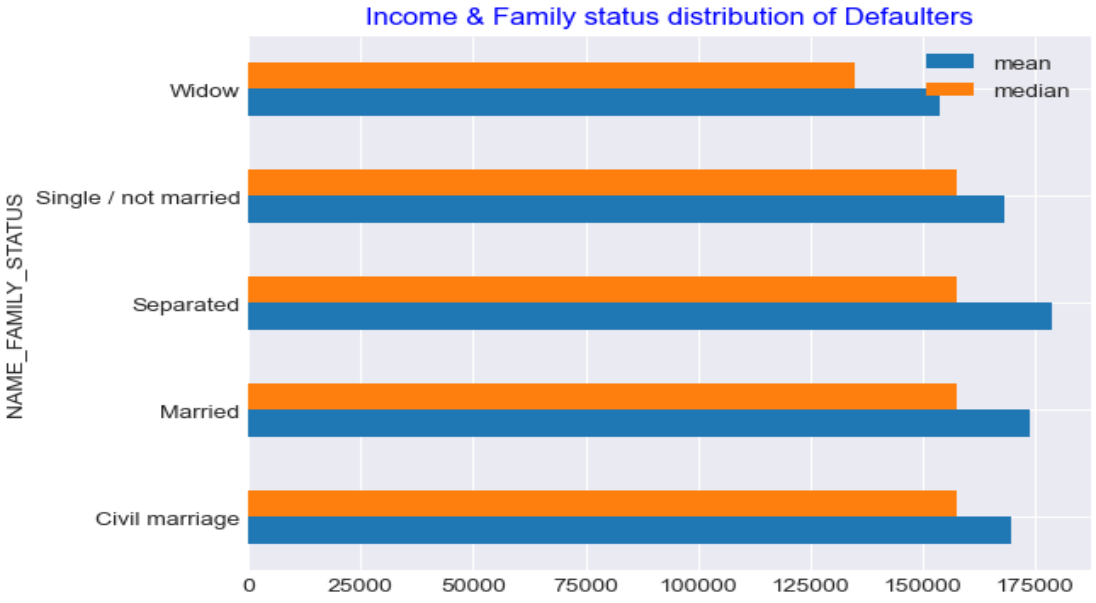
Education based contract status





Income & Family status distribution





FINDINGS

- After merging the data, there are 1413701 rows and 69 columns available.
- There 67243 approved loans who are having payment difficulty.
- Female clients have more payment difficulties compared to males.
- Clients with secondary education is having the highest approval count.

CONCLUSION

- Clients who are applied for cash loans have more payment difficulties.
- Working professional, people with secondary education and married people are having more payment difficulties.
- Female applicants are comparatively more in number in terms for payment difficulties.
- People who are not having the own reality and car are having more payment difficulties then people who are having the same.
- Applicants whose income between 1,00,000 to 2,00,000 are highly facing payment difficulties.
- People live in Co-op apartment and civil married are having high 60 DPD (days past due) default.
- Married clients whose education is secondary and income between 1,00,000 to 2,00,000 are having more payment difficulties.`