Credit EDA:
Case study analysis

- A Anurag

Problem Statement and Approach

- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. 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.
- The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:
  - **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
  - **All other cases:** All other cases when the payment is paid on time.
- We aim to analysis the given datasets and tell the bank which client with certain conditions will have payment difficulties and for some who don't have.

- The given two datasets are: *'application_data.csv' and 'previous_application.csv'.*
- After loading the required libraries we load these two datasets in our ipynb file for performing actions on it.
- Then we clean the dataset(Like treating the missing values, deleting the columns which will not be part of our analysis, conversion of the columns to right format, taking care of negative value, checking for outliers etc.
- Univariate, Bivariate and Multivariate analysis on some of the columns will done to trends or insights from them.
- Then after merging both the datasets and analysed for the insights.

Summary of Datasets

Dataset 1: Application data

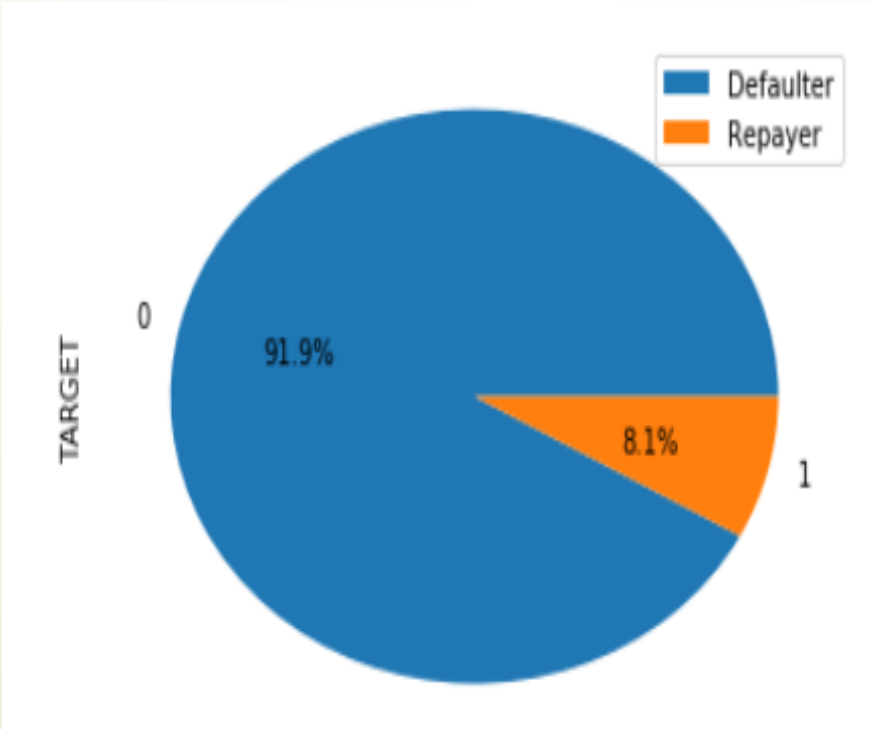
- Initially we had 122 columns, but it was reduced to 45 columns which will be used for analysis.
- Null values percentage greater than 40% values columns were removed.
- We did following things on the dataset
  - Removed unwanted columns for the analysis
  - Treated the null values wherever required.
  - Converted the days to years to its proper format
- Values are also made to positive
- With all these changes this dataset now is ready for the analysis.

Dataset 2: Previous Application data

- Same procedure is followed in this dataset too as in dataset 1.
- There were 1670214 Rows & 37 Columns. We reduced the number of columns to 16 columns.
- Null values percentage greater than 40% values columns were removed.
- We did following things on the dataset
  - Removed unwanted columns for the analysis
  - Treated the null values wherever required.
  - Converted the days to years to its proper format
- Values are also made to positive
- With all these changes this dataset now is ready for the analysis.

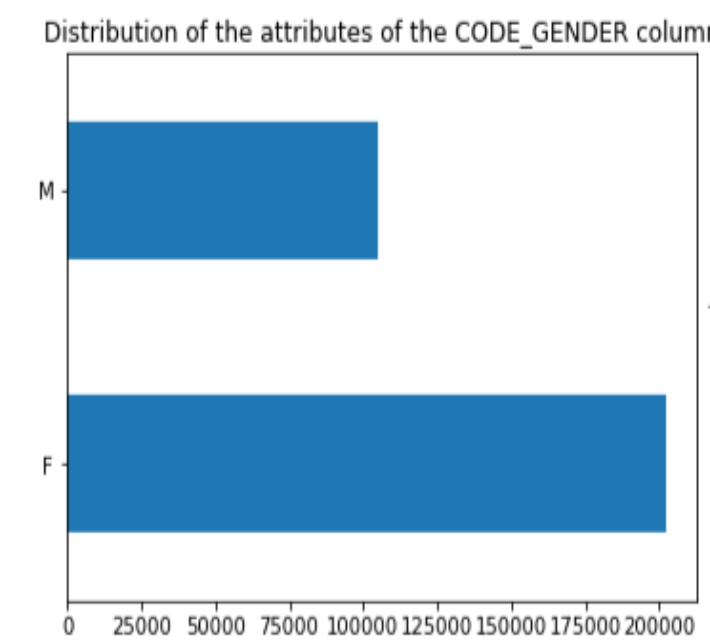
Data Imbalance

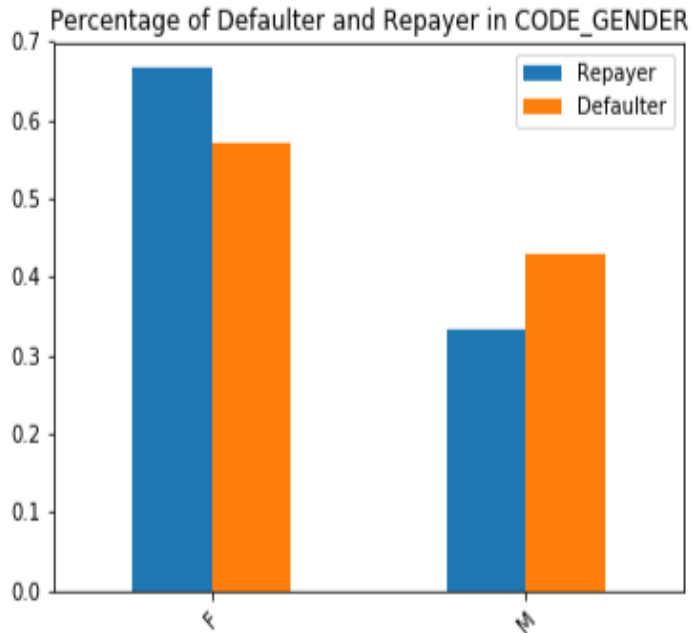
- Defaulter percentage is around 91.9%
- Repayer percentage is around 8.1%
- So, Imbalance ration between can be calculated and it came to 11.39%.
- Shown in pie chart below:



Univariate and Bivariate Analysis

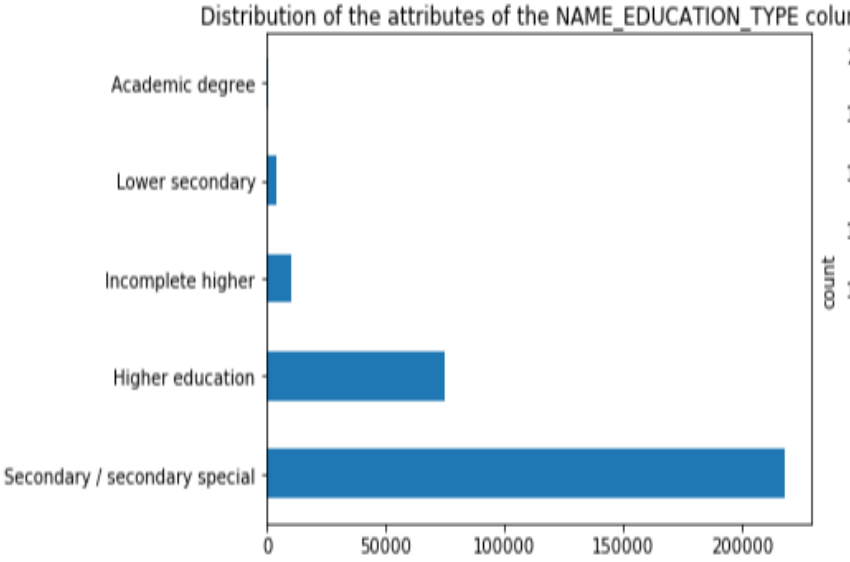
- Analysis based on Gender
- Females take more loans than the males.
- Even though females take more loans but males have high paying difficulties compared to women.

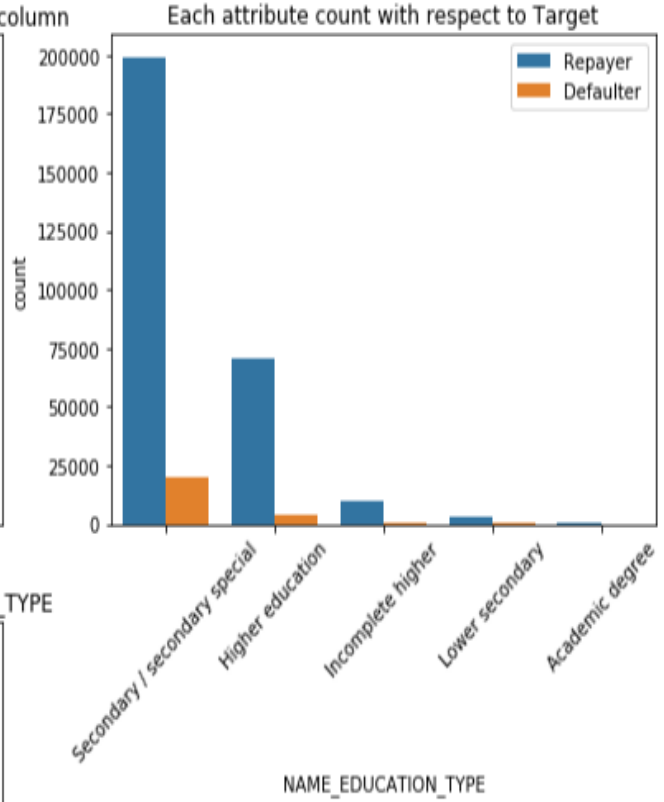


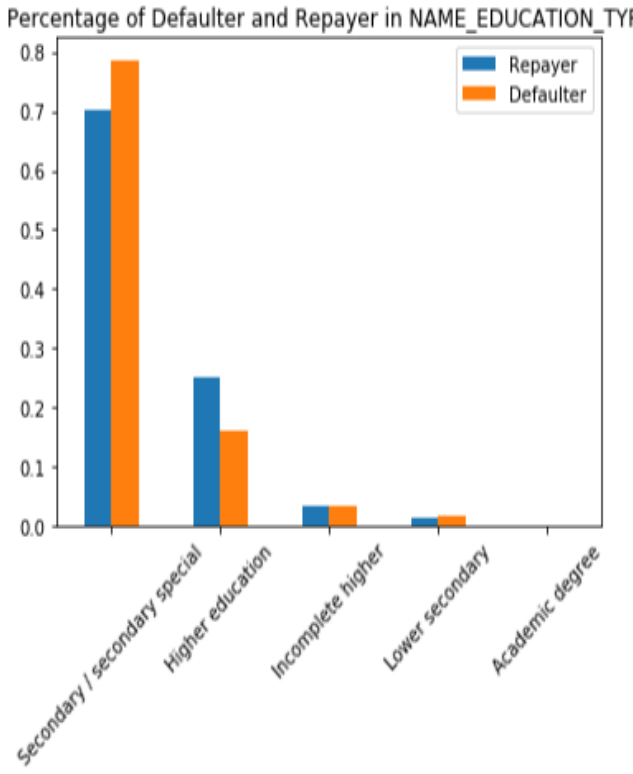


Contd.

- Analysis based on Education
- People having Secondary/ Secondary special are given most loans followed by Higher education clients.
- It is clearly visible that people with Higher Education have less difficulties with respect to the Secondary/Secondary special.
- So, better education means better jobs they get.

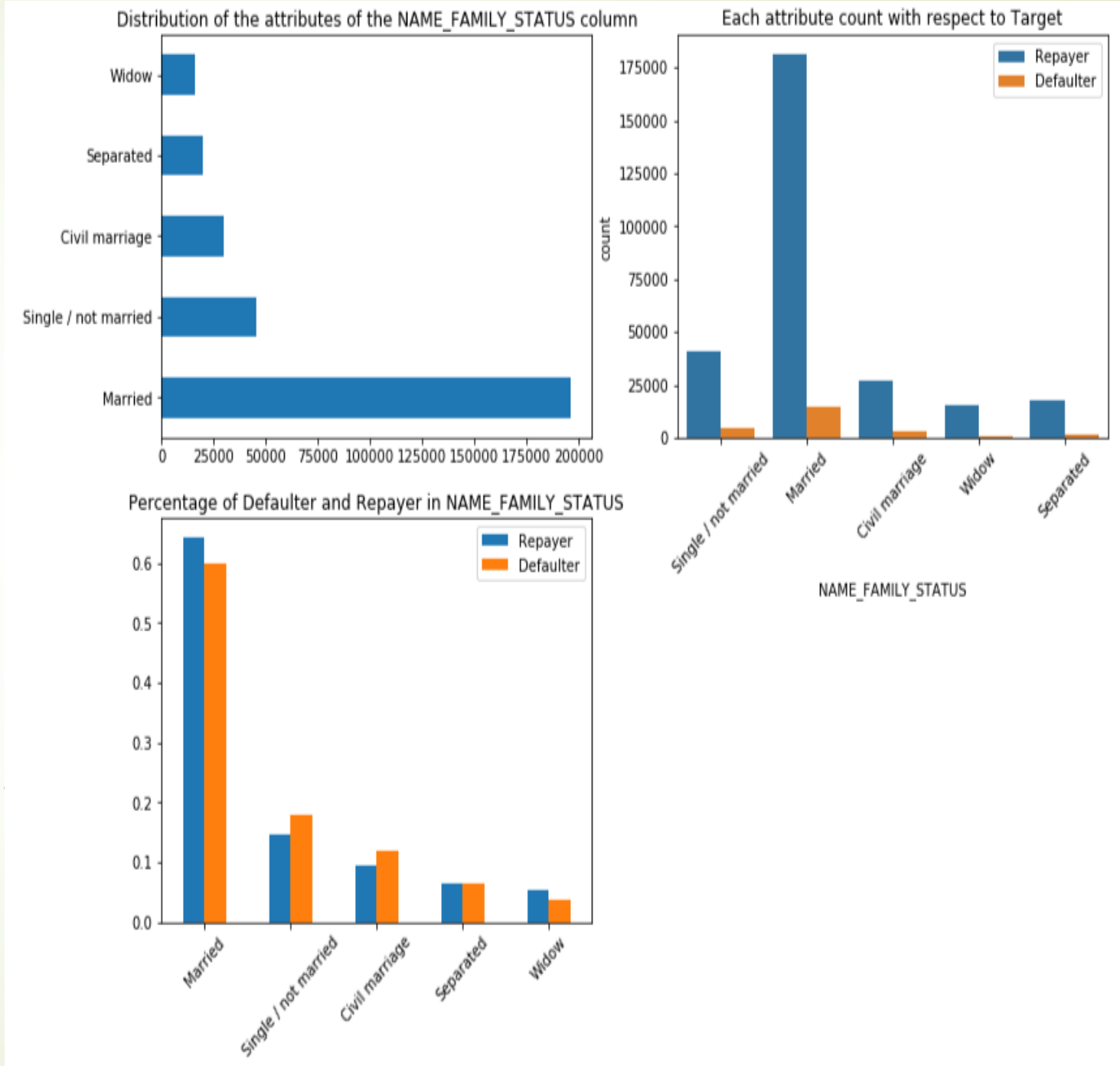






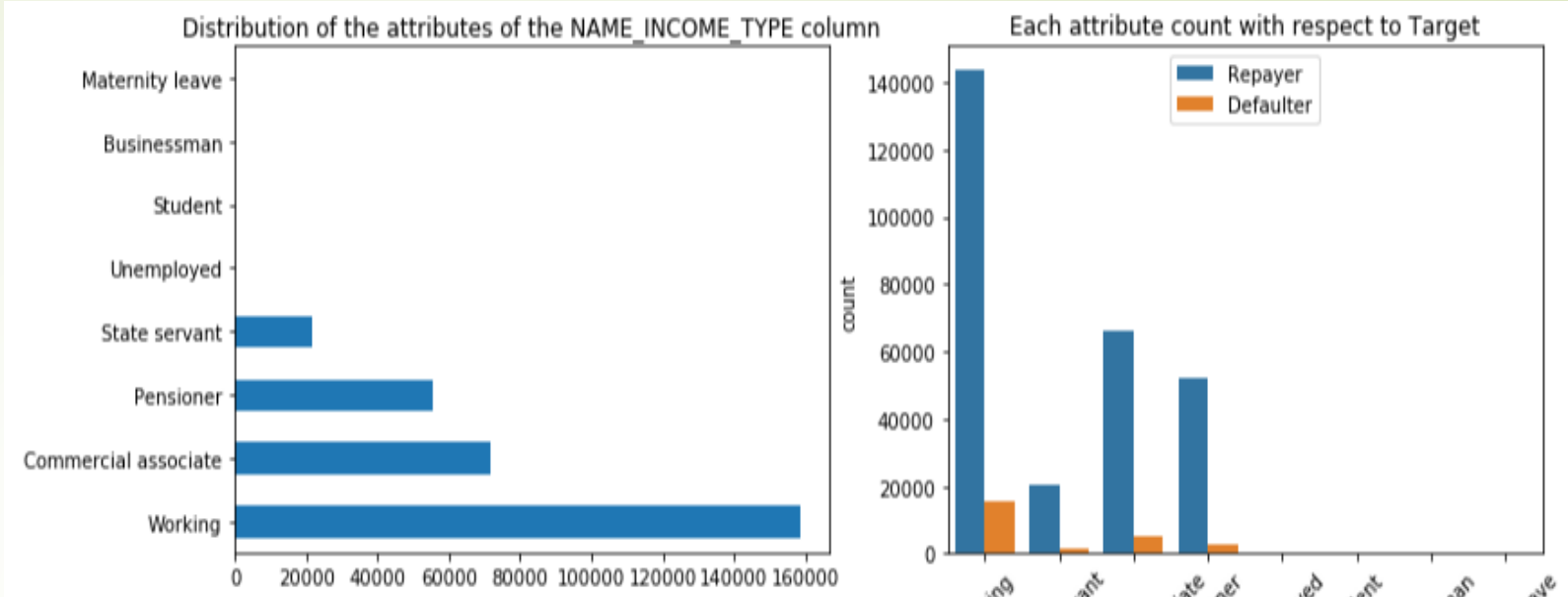
Contd.

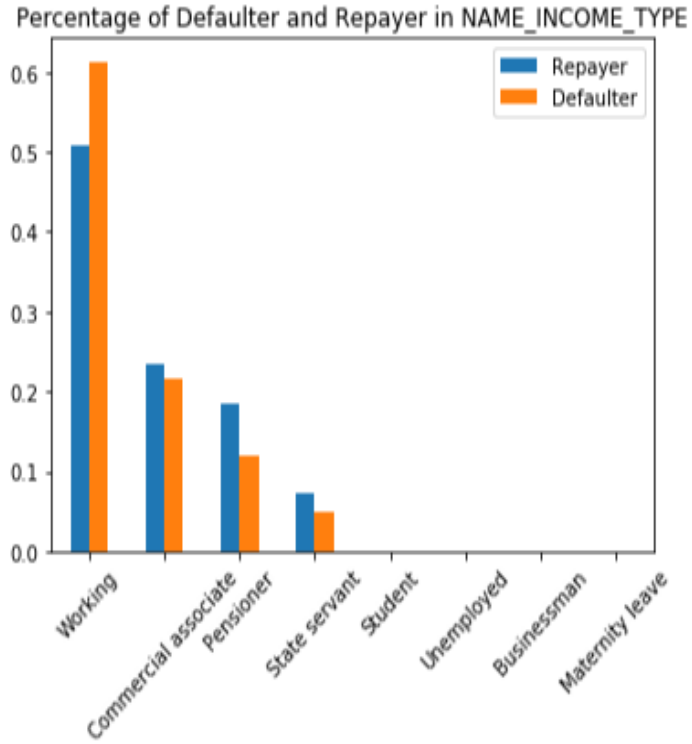
- Analysis based on Family status.
- Most of the loans are given to the married people, more than half of the loans.
- Even though most loans are taken by married people percentage of defaulter in them are less when compared to Single/not married clients.
- This could be because of married clients may have dual source of income.



Contd.

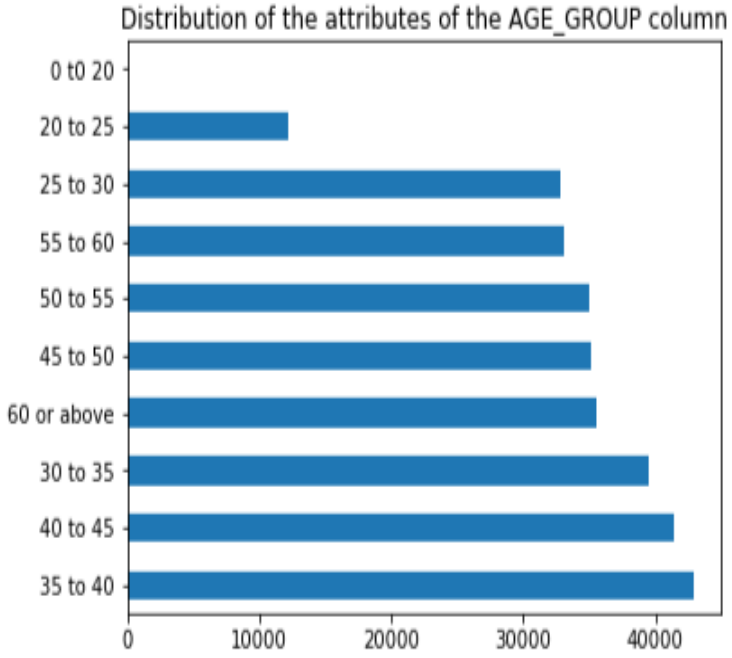
- Analysis based on Income.
- Most of the loans here are taken by the working people, followed by Commercial associate etc.
- It can be seen that clients who are working have more payment difficulties when compared to other income types.

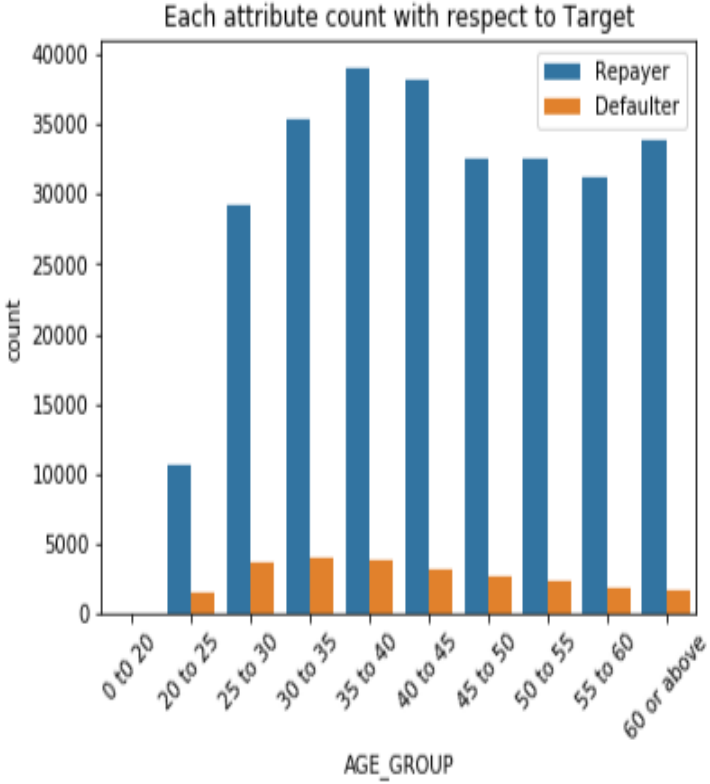


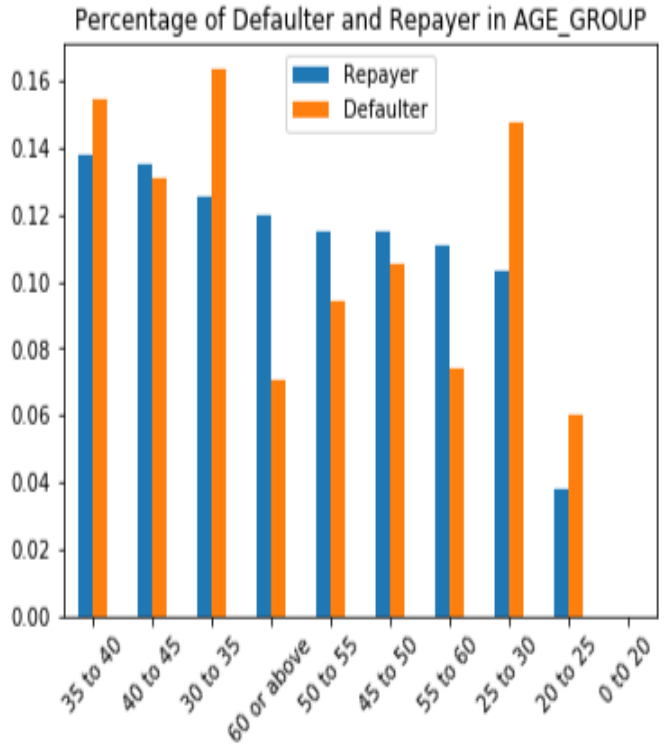


Contd.

- Analysis based on Age.
- Loans taken by middle aged clients that is clients age ranging from 30 to 60 have nearly same percentage of loans being taken.
- Clients who are old are taking less loans as they may be retired and may have limited amount of income.
- Apart from this clients whose age group is between 20 to 30 have more defaulters percentage as this is when they start earning and may not know how to invest and end up spending all.

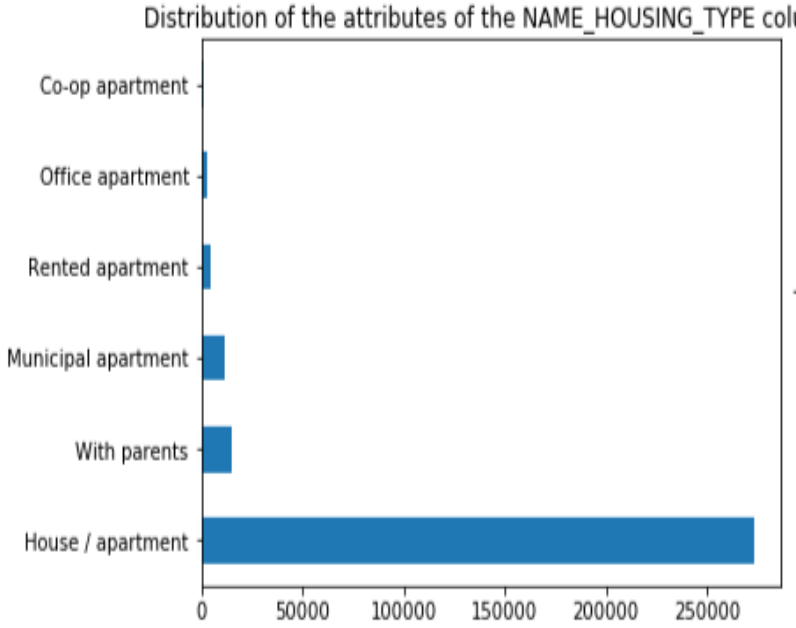


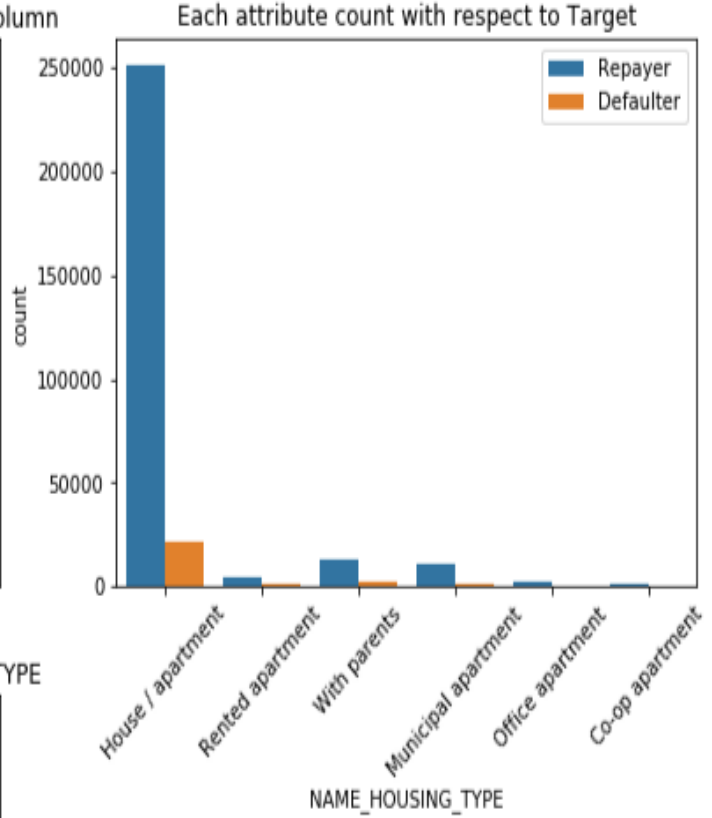


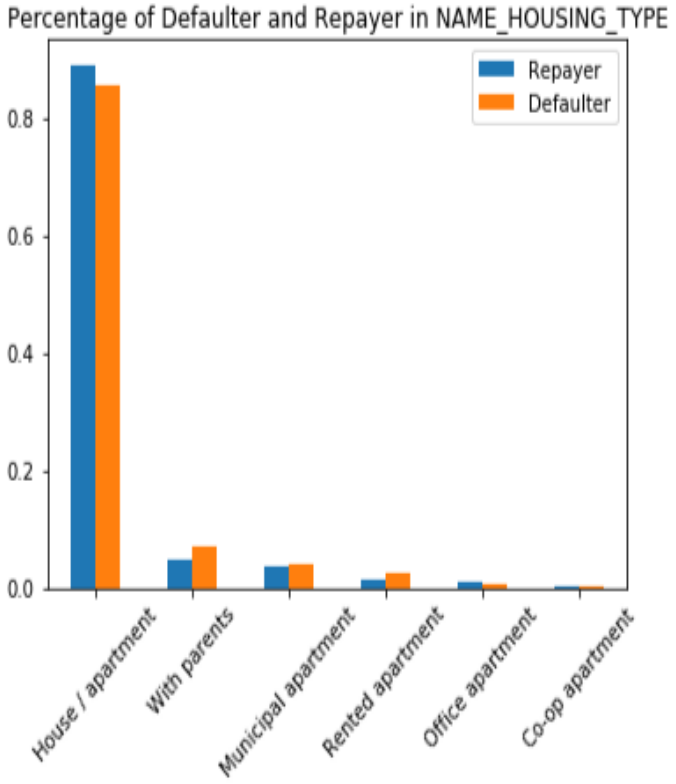


Contd.

- Analysis based on the clients residence
- Clients who stay in Apartments or houses have been taking more loans.
- Clients who are facing payment difficulties are those who stay with their parents, this might be probably because of the increased people to feed, their health etc.

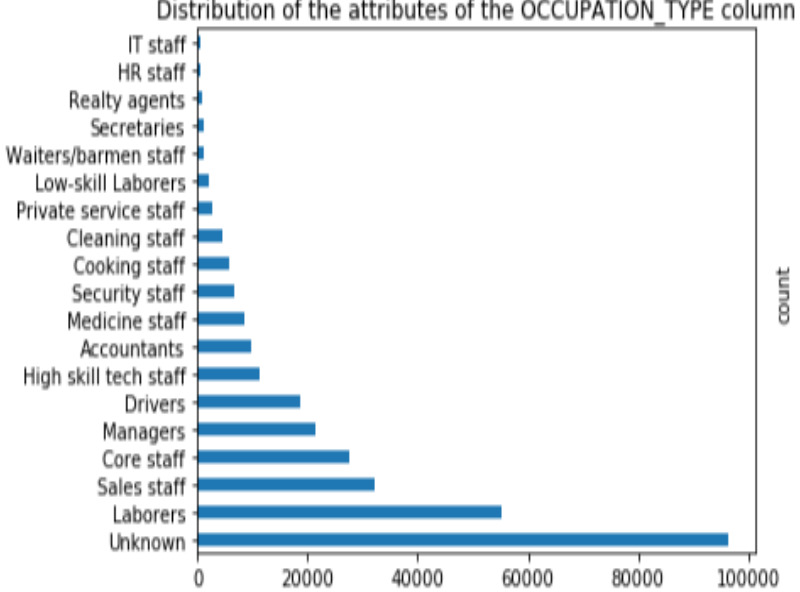


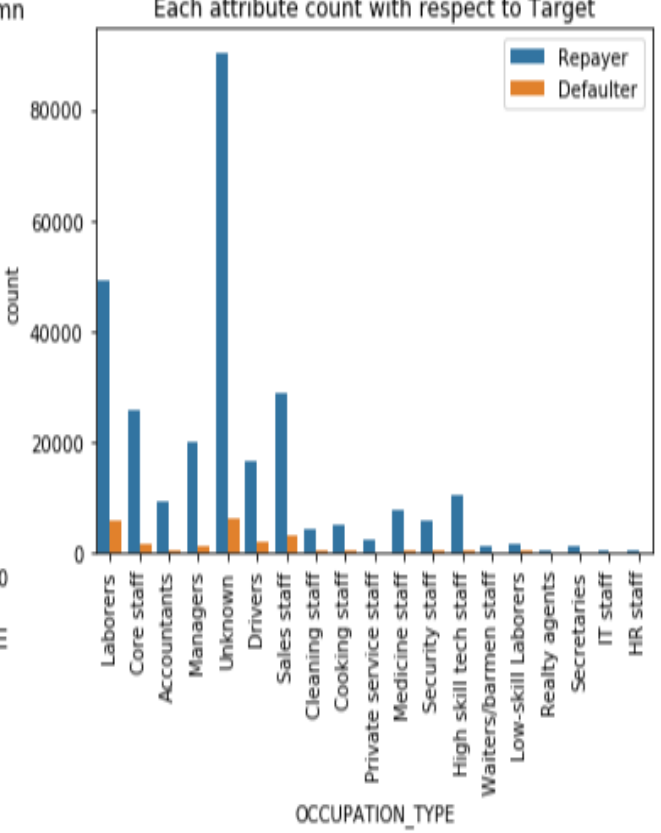


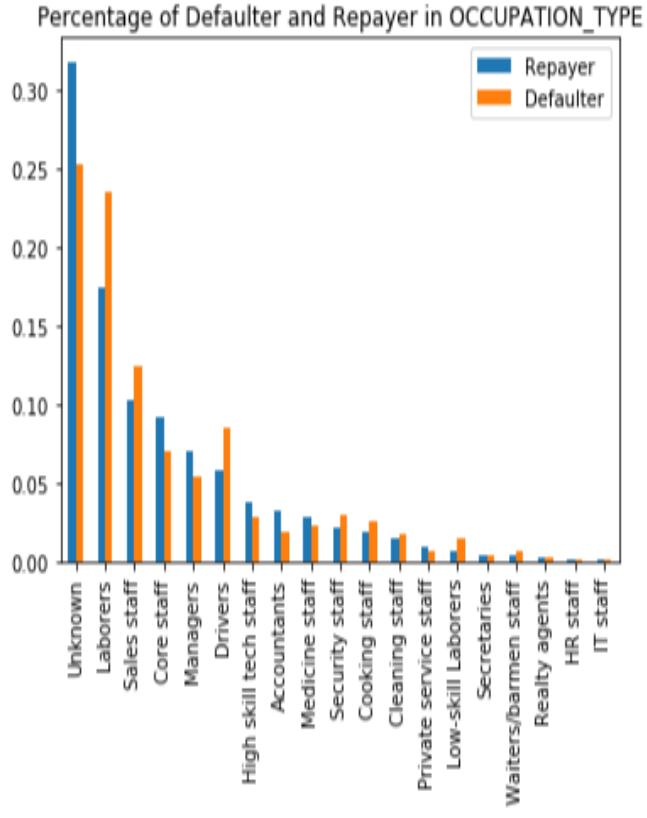


Contd.

- Analysis based on Occupation of the clients.
- Here Low skill level or laborers clients are the ones who are having most payment difficulties this might be due to their lack of education or environment.
- IT staff are very less likely to apply for loan as compared to others.

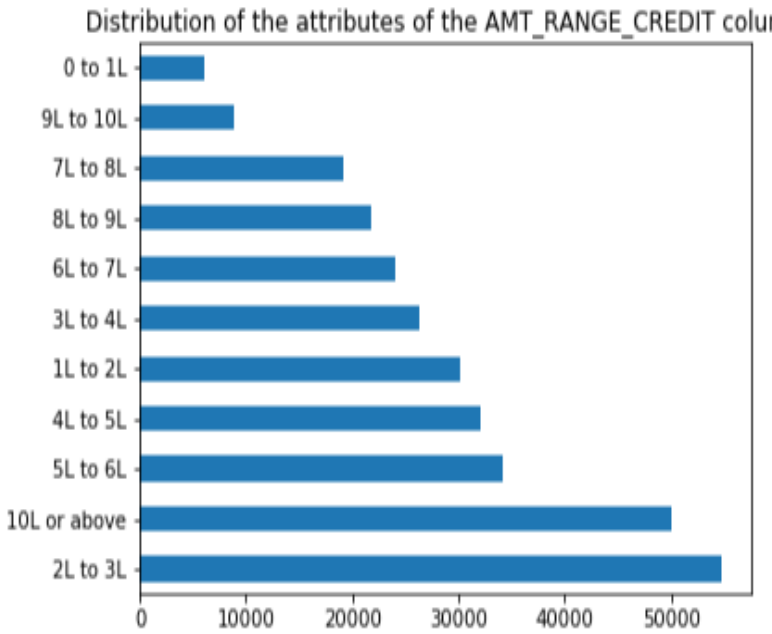


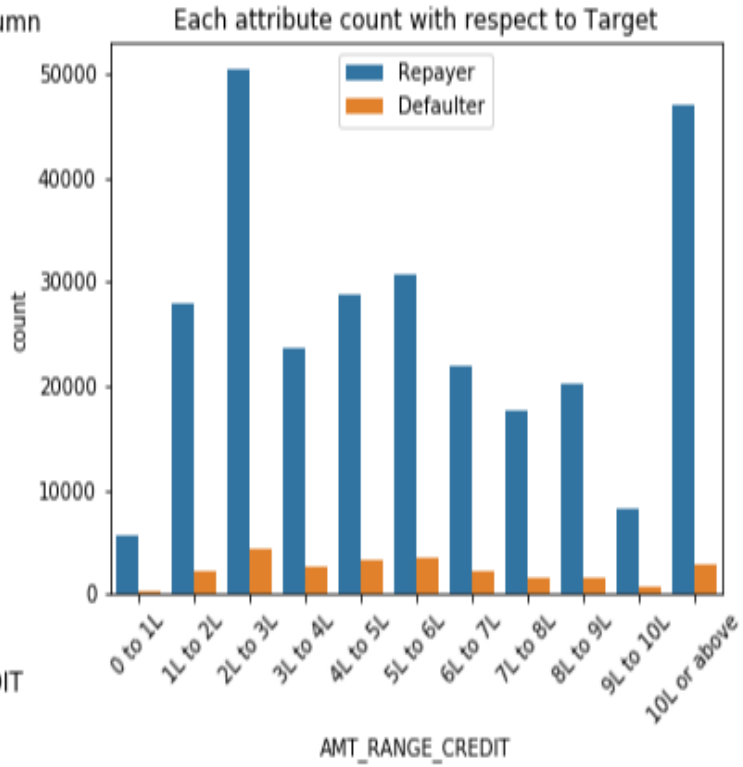


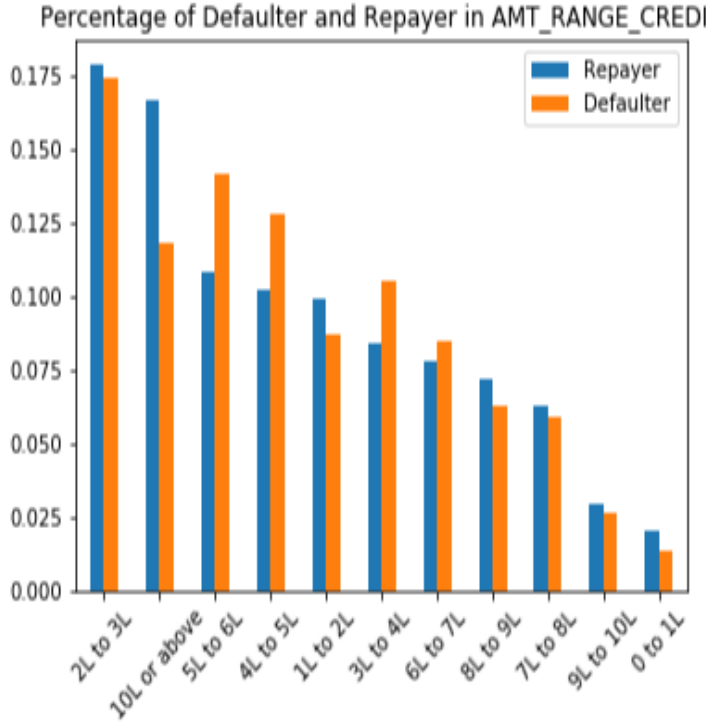


Contd.

- Analysis on Credit amount.
- Most of the loans are taken by the clients whose credit amount is between 2 lakh to 3 lakh, followed by 10 lakh or above.
- Clients with better credit amount are less likely to be defaulter that is 10 lakh or above, whereas clients whose credit amount is up to 5 lakh are having more payment difficulties.

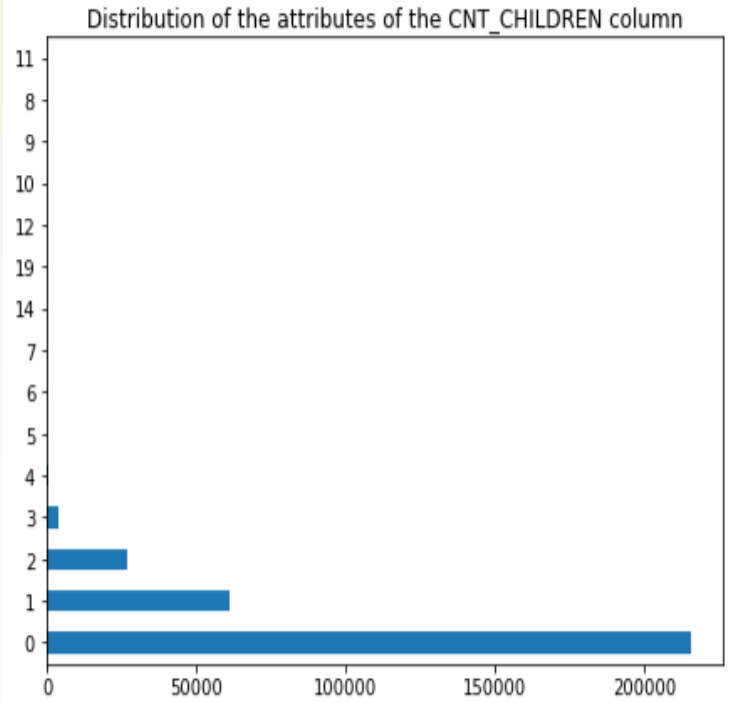


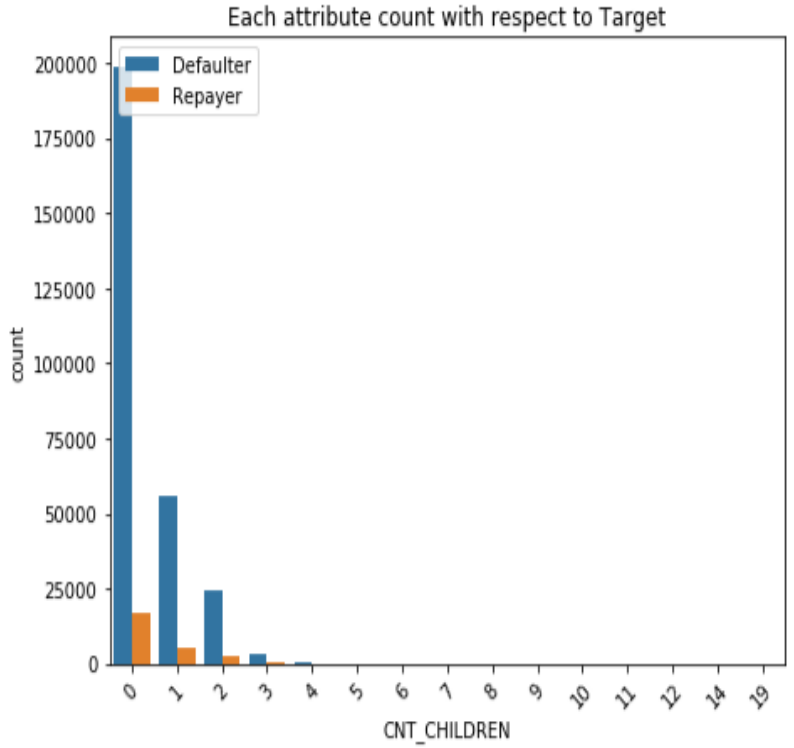


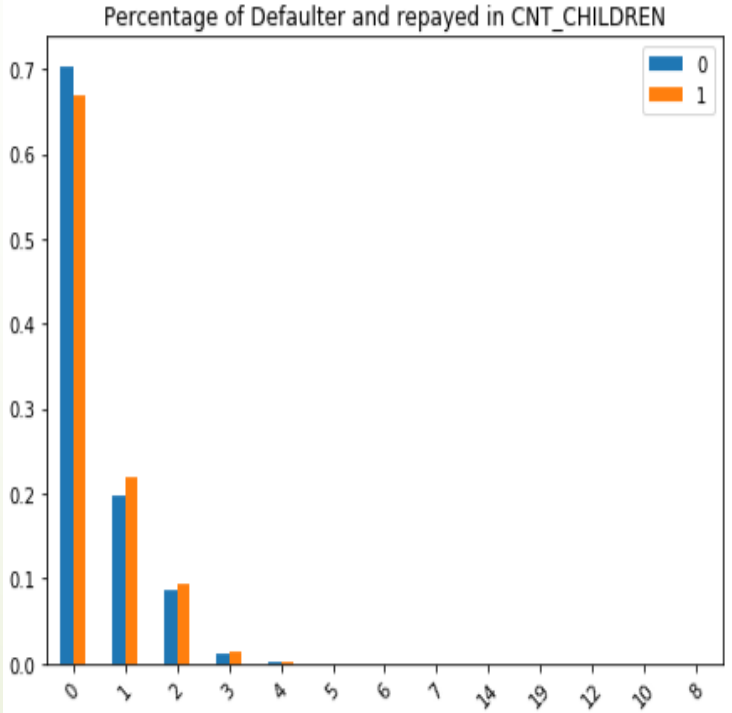


Contd.

- Analysis based on number of children client has.
- Close to 70% of the clients who took loans does not have children.
- As the number of children increase we can see payment difficulties, whereas not much payment difficulties for clients who does not have children.

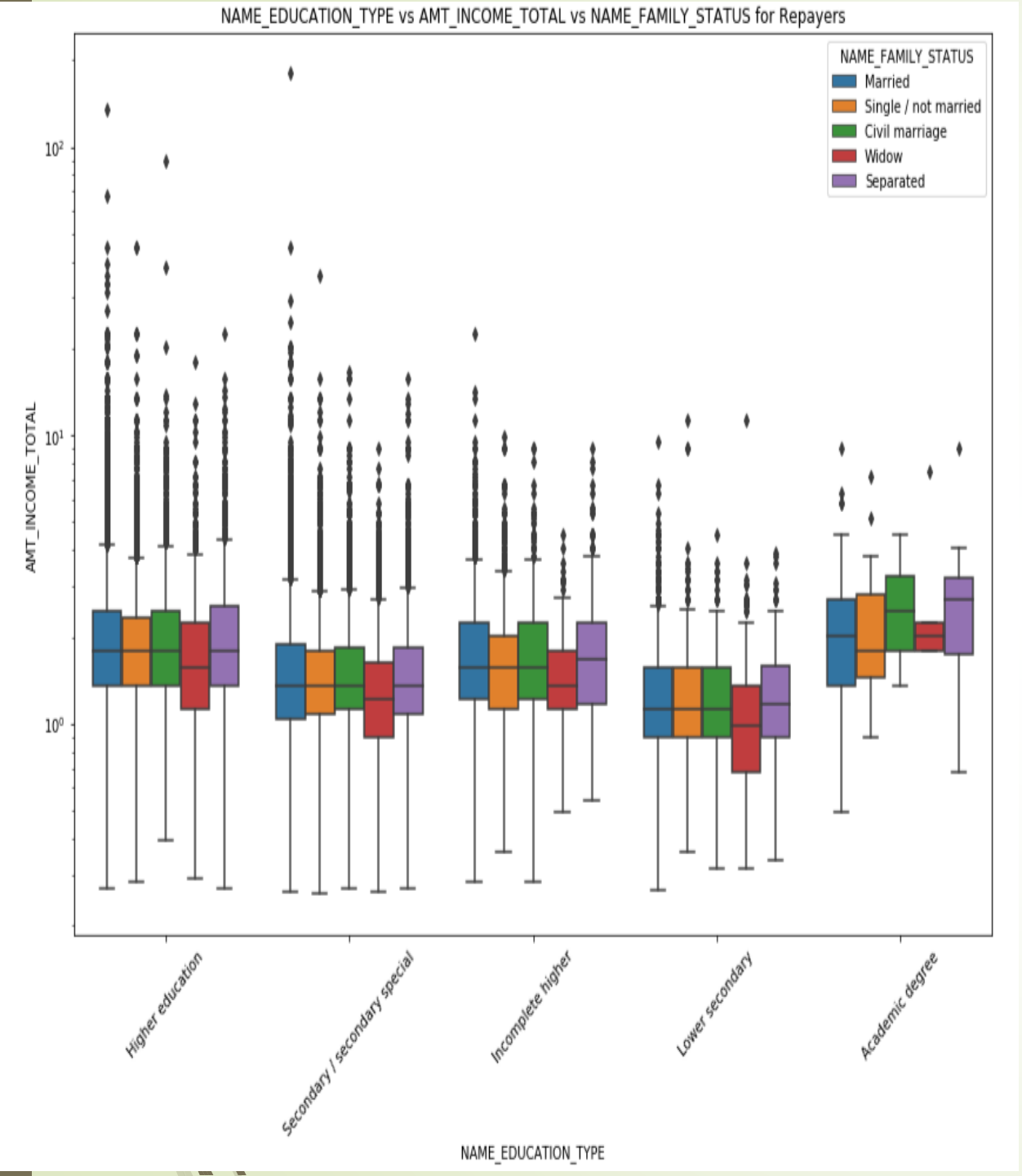


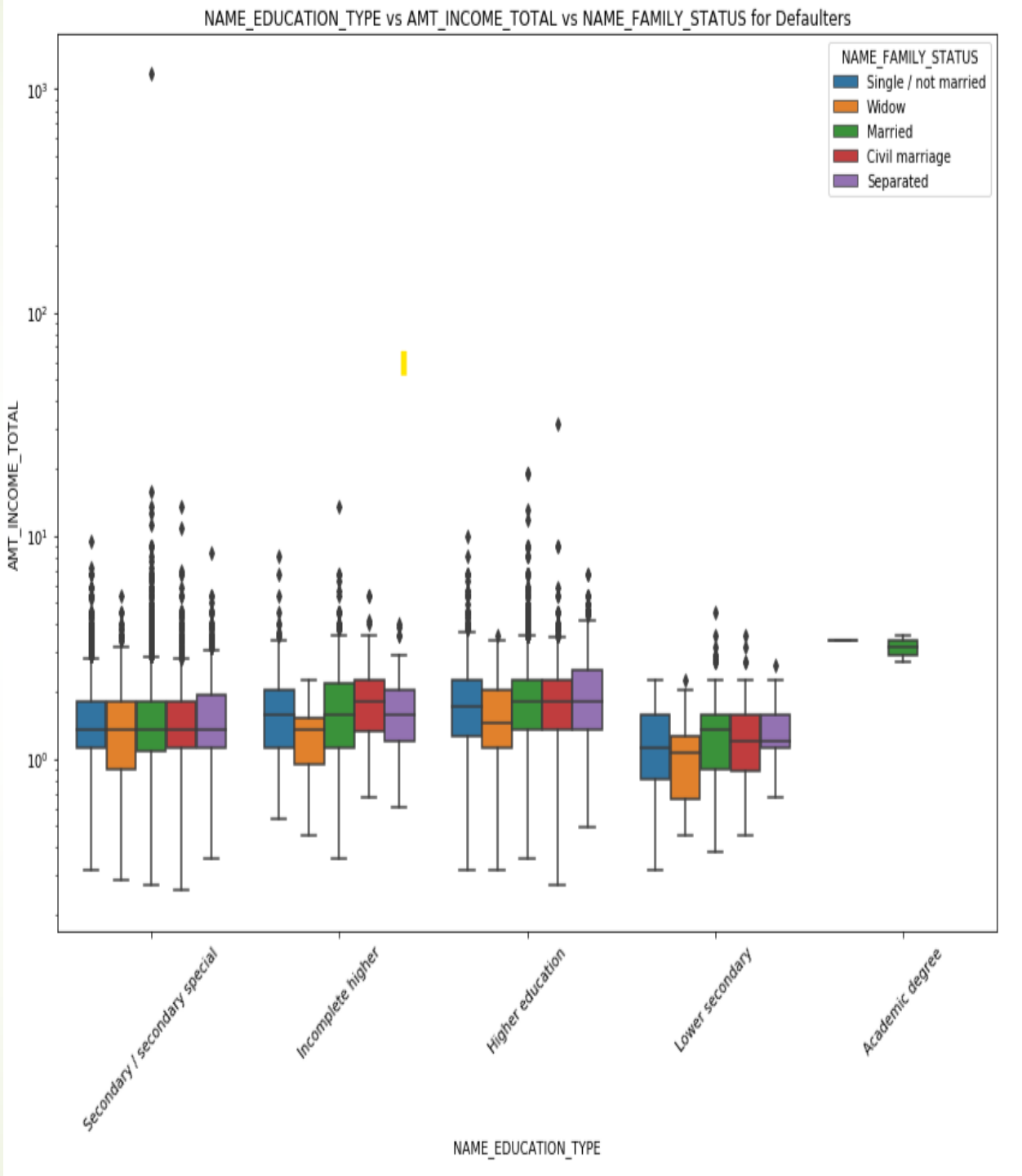




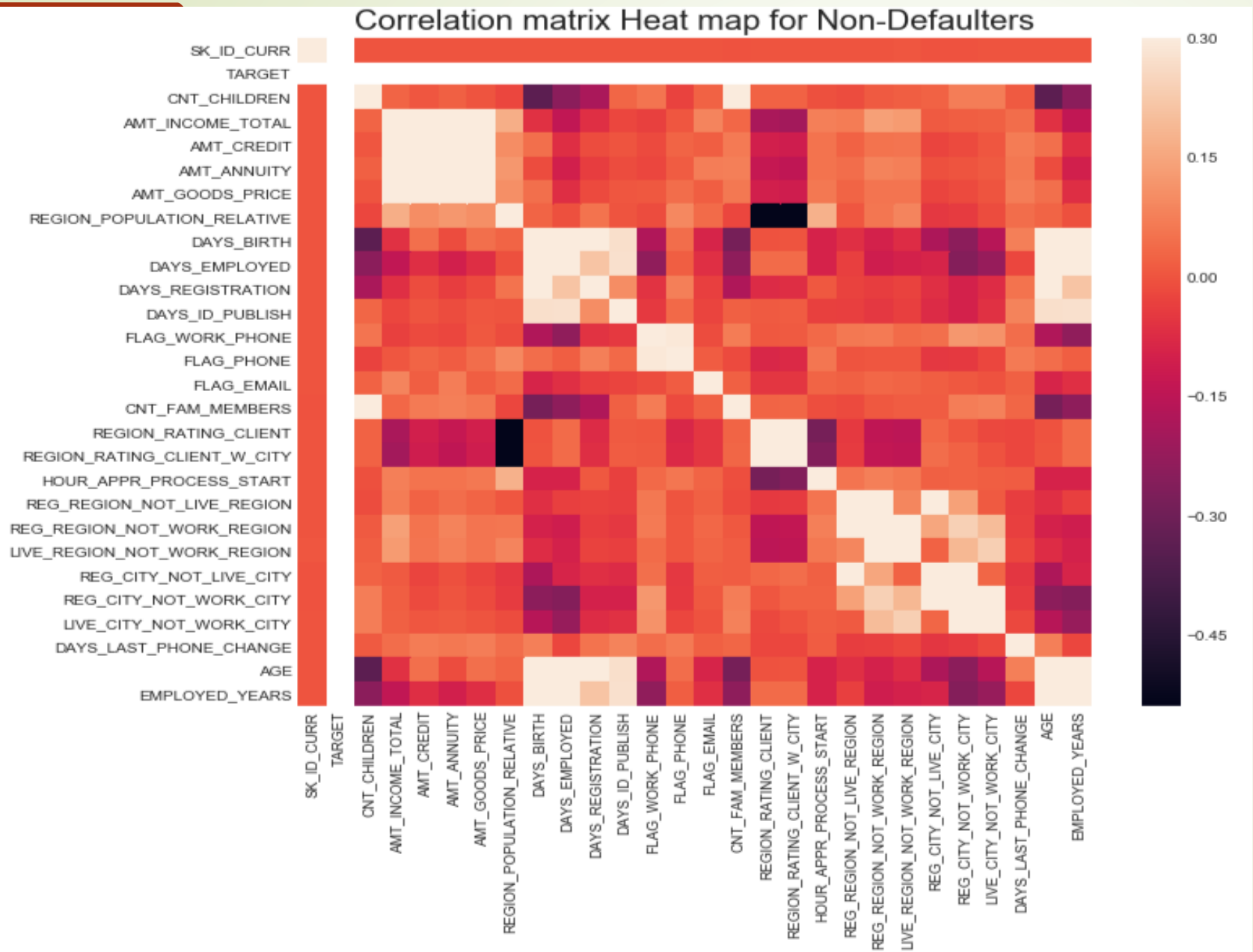
Multivariate Analysis

- *Non-defaulters* : "The income total for families with both higher education and 'Secondary/secondary special' education types is generally similar, and both of these categories exhibit numerous outliers. Conversely, families with academic degree education types tend to have fewer outliers, and their income total appears slightly higher."
- *Defaulters* : The majority of outliers can be attributed to the education types 'Secondary/secondary special,' 'Incomplete higher,' and 'Higher education.' On the other hand, lower secondary and academic education types exhibit only a limited number of outliers. It's worth noting that within the 'Single,' 'Civil marriage,' and 'Separated' family status groups, the income amounts under the 'Secondary/secondary special' status are nearly identical.


NAME_EDUCATION_TYPE vs AMT_INCOME_TOTAL vs NAME_FAMILY_STATUS for Repayers
NAME_FAMILY_STATUS
Married
Single / not married
Civil marriage
Widow
Separated
AMT_INCOME_TOTAL
NAME_EDUCATION_TYPE
Higher education
Secondary / secondary special
Incomplete higher
Lower secondary
Academic degree

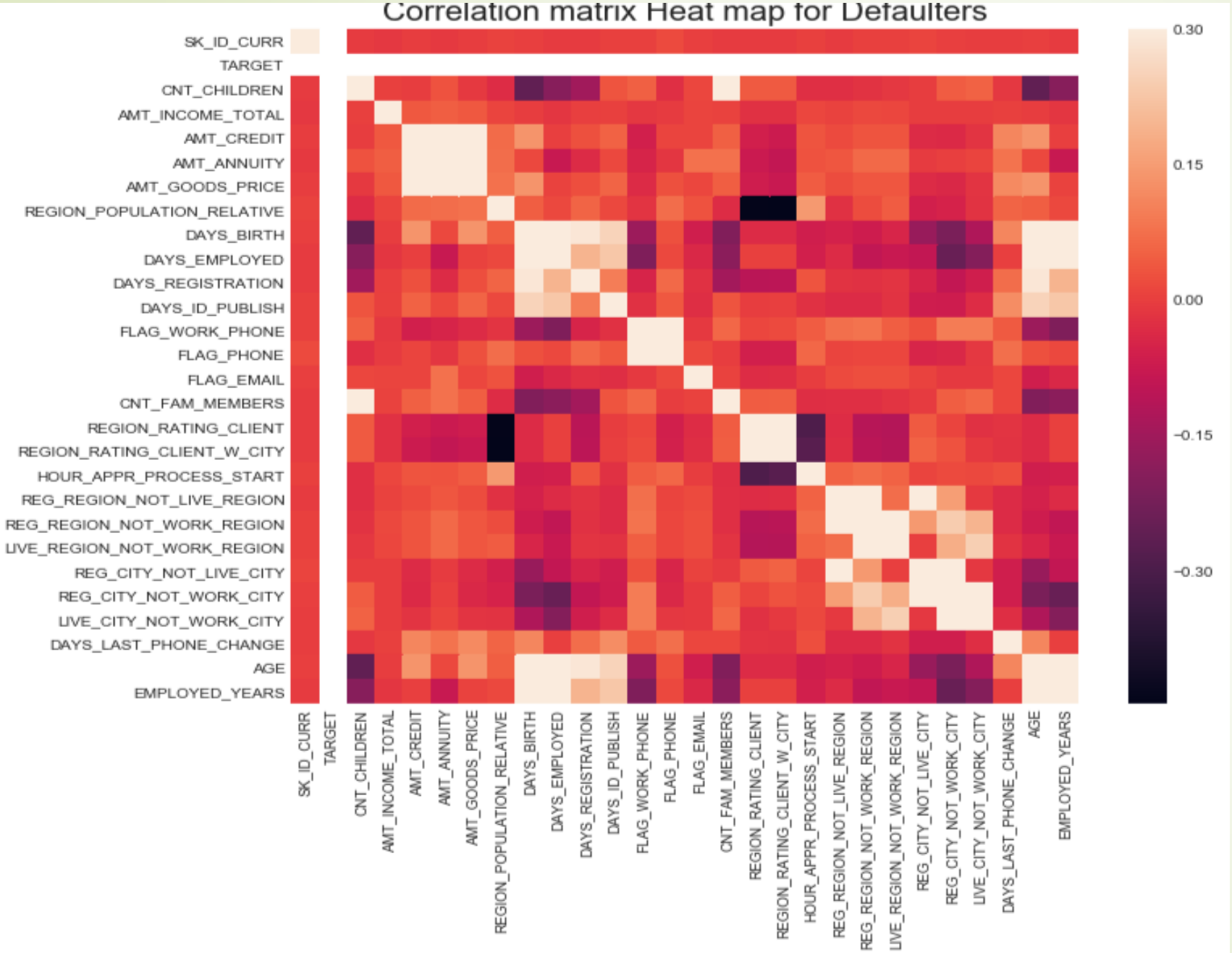

NAME_EDUCATION_TYPE vs AMT_INCOME_TOTAL vs NAME_FAMILY_STATUS for Defaulters
NAME_FAMILY_STATUS
Single / not married
Widow
Married
Civil marriage
Separated
AMT_INCOME_TOTAL
NAME_EDUCATION_TYPE
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Academic degree

Correlation



- Some Observations from the previous slide are:
  - The income amount is inversely related to the number of children a client has, which means that having fewer children leads to higher income, while having more children results in lower income.
  - Client living in densely populated area has less number of children
  - Client living in densely populated area has higher credit amount

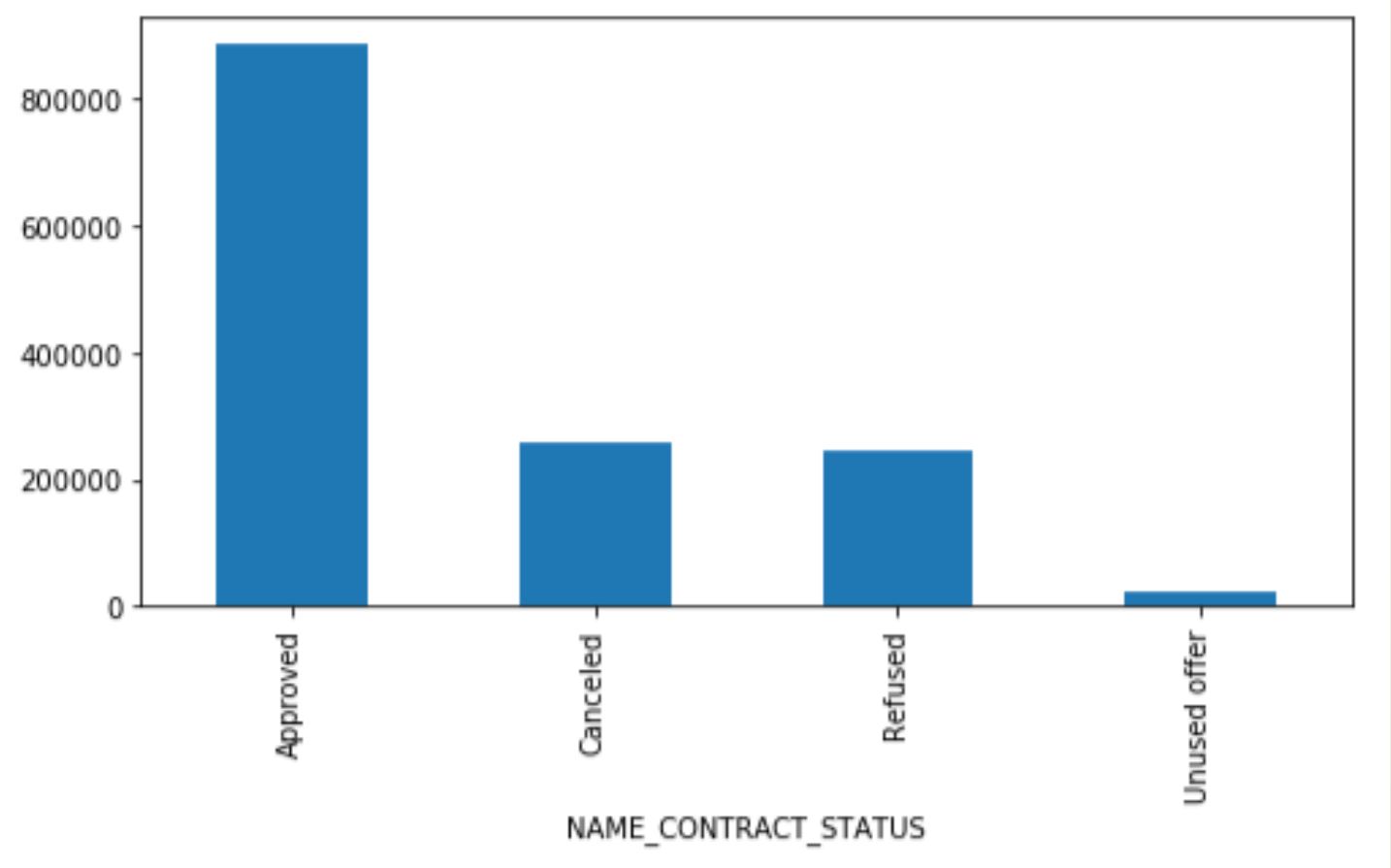
Contd.



- Some Observations are:
  - When the client's permanent address differs from their work address, it is correlated with having fewer children, and conversely so.
  - The client's permanent address not matching the contact address is associated with having fewer children, and the opposite is also true.
  - Similarly to the previous correlation quite a bit both are same.

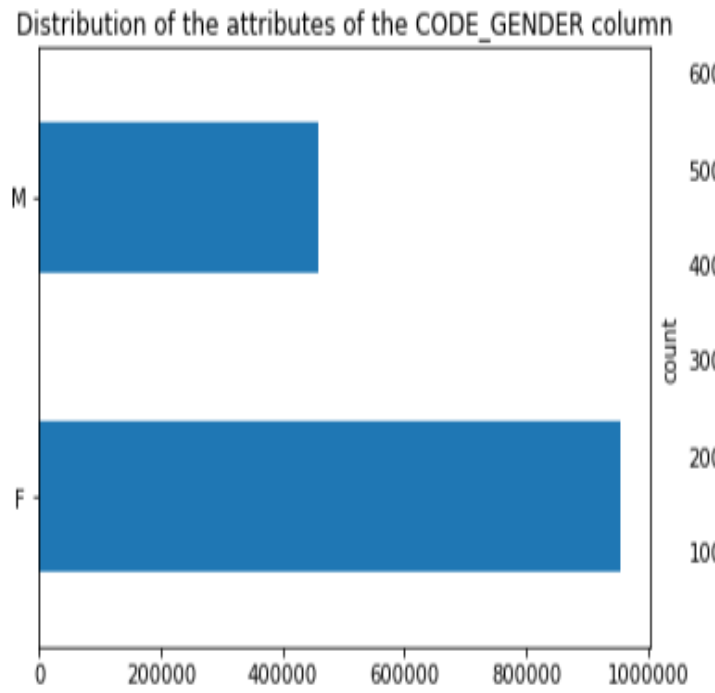
Merging both the datasets

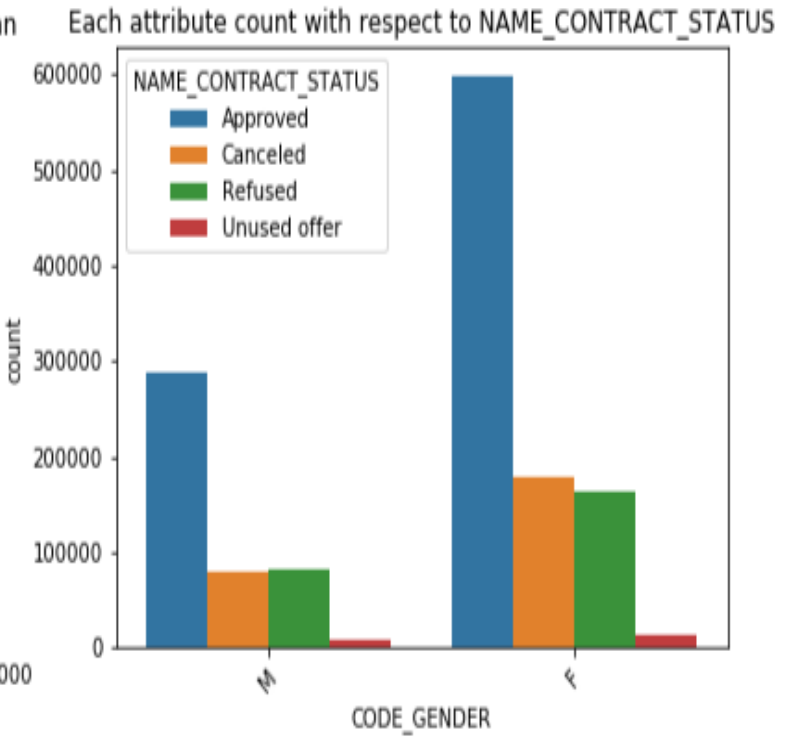
- After merging both the datasets, loans which are approved are more as compared to all other.

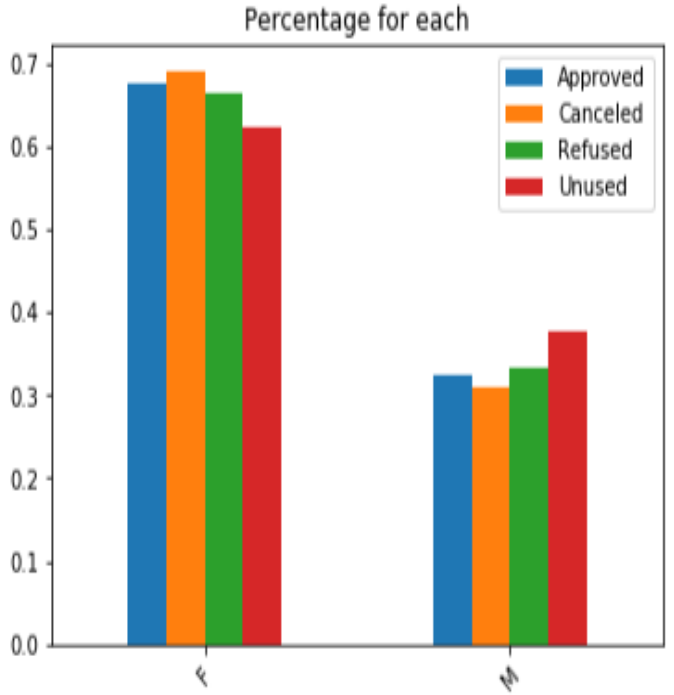


Contd.

- Analysis of Gender for merged dataset.
- Female clients have more number of applications with approved percentage to be high.
- Male clients has comparatively low values for all four types.

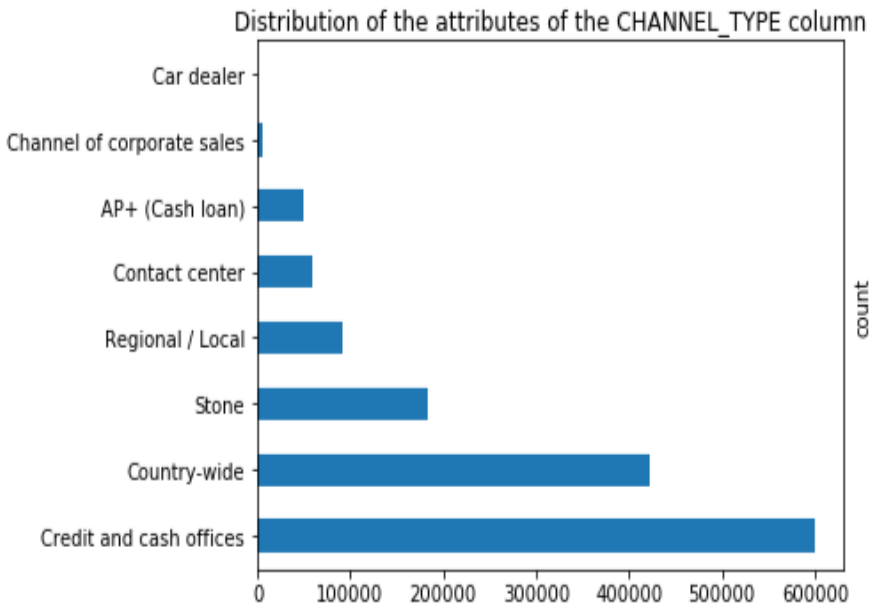


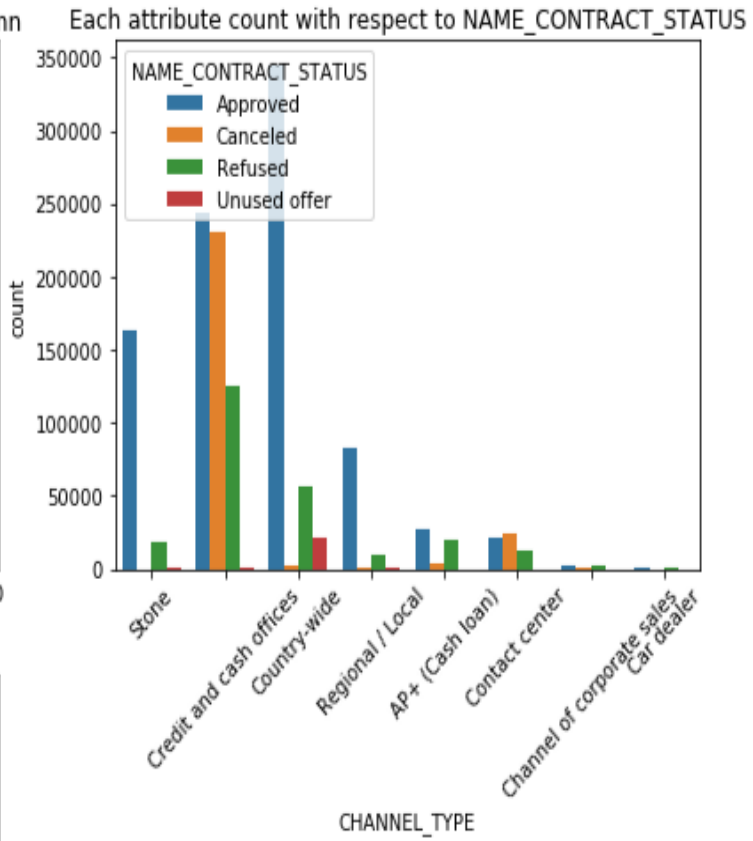


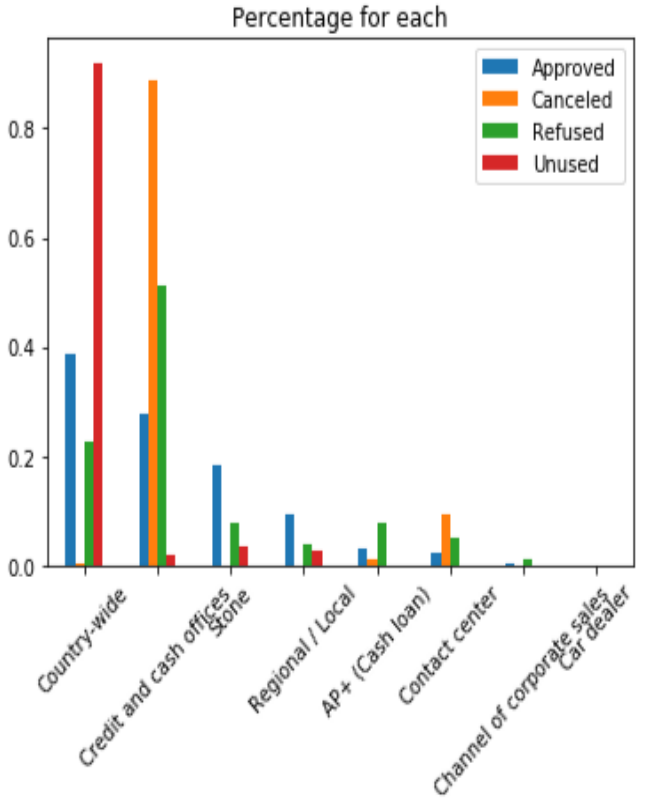


Contd.

- Analysis of Channel type.
- For Country-wide the approved loan percentage is more as to others.
- For Credit and Cash Cancelled loan percentage is very high.







Conclusions

- Banks should prioritize lending to individuals pursuing 'Higher Education' and steer clear of lending to those in Secondary/Secondary Special, Incomplete Higher, or Lower Secondary levels due to the challenges they may encounter with repayment.
- Prioritize clients with income types such as Commercial Associate, Pensioner, and State Servant, as they generally exhibit a more reliable payment history, while exercising caution when considering clients categorized as 'Working,' as they may have a higher incidence of payment challenges.
- Banks should prioritize clients within the age range of 35 to 70, as they are typically financially stable and exhibit fewer payment difficulties.
- Give preference to clients residing in housing types classified as 'House/apartment,' as they tend to experience fewer payment difficulties.
- The bank should concentrate on the 'Country-wide' channel type, which shows a higher volume of approved loans, while the 'Credit and cash offices' channel type tends to have a higher number of canceled and refused loan applications.