Credit EDA Case Study

Submitted by:
Sawal Malhotra
&
Sanu Sinha

Introduction:

- Bank and financial institution lend a sum of money i.e. Loan to the Customer who have requested for a given amount.
- However, there are risks involved in lending money to the customers.
- Risks being:
  - 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.
- This report is formed in order to help in decision making after analysing various factors so that the bank neither looses a good business nor incurs a financial loss.

Problem Statement:

- The aim of Case study is to identify the Driving factors for a Customer to Default and thus , help Banks/Financial institution take a better decision to reduce the risk.

- Based on the applicants data and historical data(if present), financial institutions can either deny the loan, reduce the amount of loan, lend(to risky applicants) at a higher interest rate.

Analysis approach:

Following is the approach followed in sequential manner to analyse the given case study-

- **Data Sourcing**: Data Sourcing is the process of getting the data from organizations(i.e. Private data) or public platforms(i.e. Public data).

  As part of the step, we are provided with the Financial institution data.

- **Data Cleaning**: Data Cleaning is the process of removing any irregularities from the Data set , which can be in form of missing value, incorrect format, spelling mistake and outliers.
  - As part of this step, we have performed Data cleaning on the data set.
  - In the cases where outliers are present, they are being mentioned in the report.

- **Univariate Analysis**: It is process of analysing and visualising a single variable from the Data set.

  As part of the step, we have performed analysis on both Continuous and Categorical variables.

- **Bivariate and Multivariate Analysis**: It is the process of analysing and visualising of two variables from the Data set.
  - Continuous - Continuous
  - Categorical - Categorical
  - Categorical - Continuous

Data Reading

- Initially there were two datasets provided.
- inp_1 = “application_data.csv” - current applicants
- inp_2 = “previous_application.csv” - previous data of the applicants

- inp_0 = inp_1 merged with inp_2 after data cleaning

Data Cleaning: Identify missing values

- Missing values were Identified in data set inp_0, inp_1 and inp_2 data set .

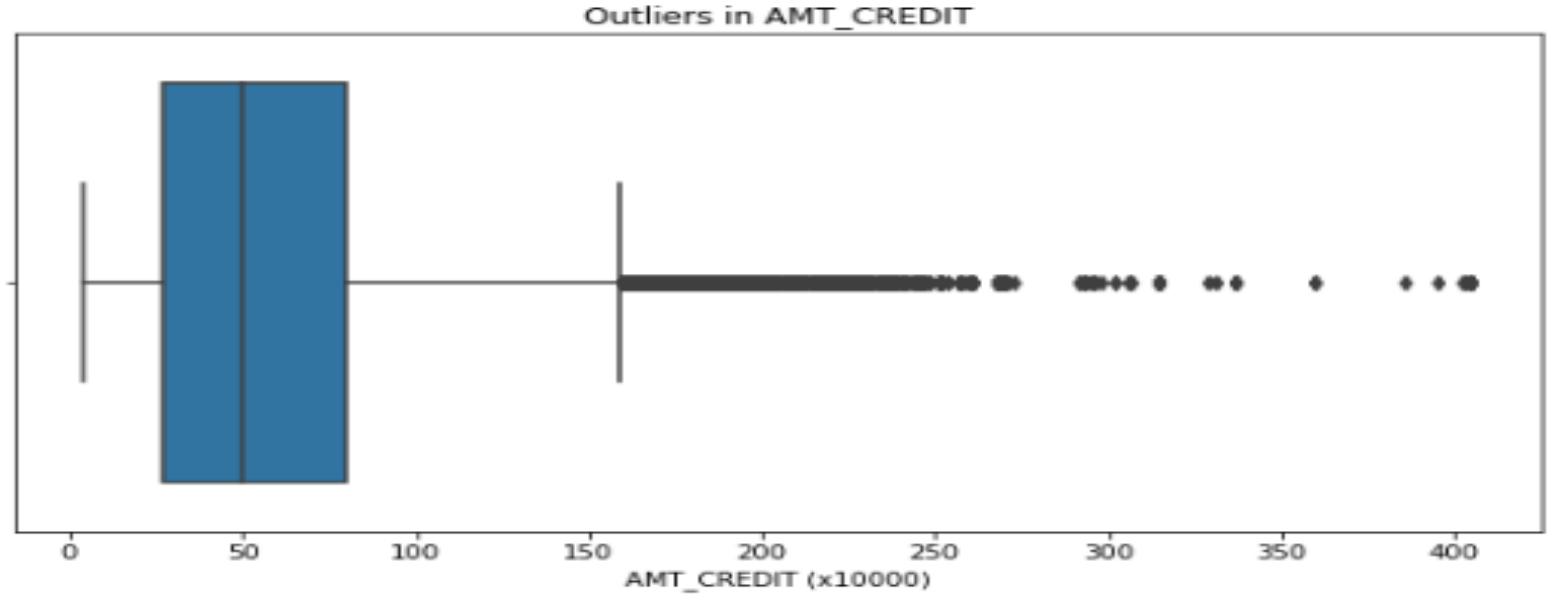
- Based the percentage of missing values- Rows were either removed and in other case it was replaced by relevant attribute like mean, median or mode in case of numerical variables.

- For one of the categorical variables, since there was no data provided for more than 30% of the population, we replaced the Null value with “Unknown” rather than dropping the rows.

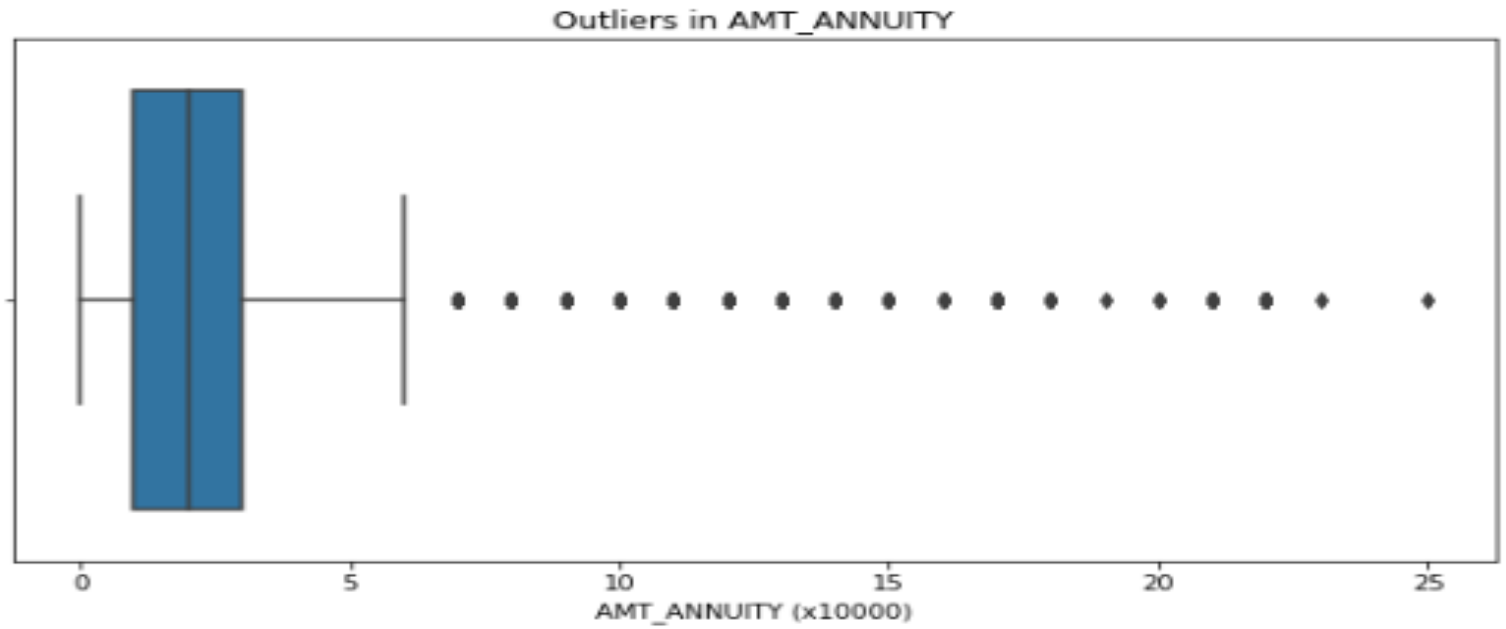
Data Cleaning: Identify Outliers

- In the Case Study, Outliers were identified by using statistics function and visually by using Box Plot.

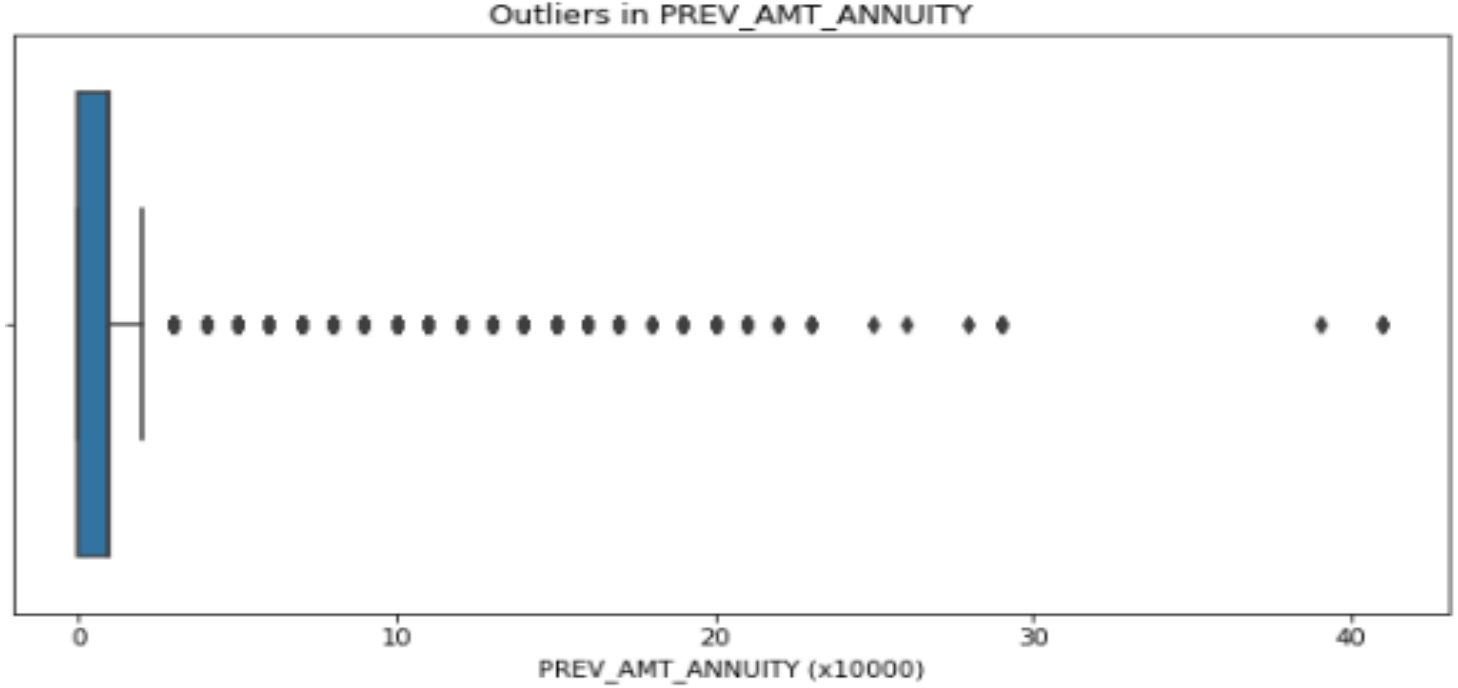
- The presence of outliers are mentioned in the report.



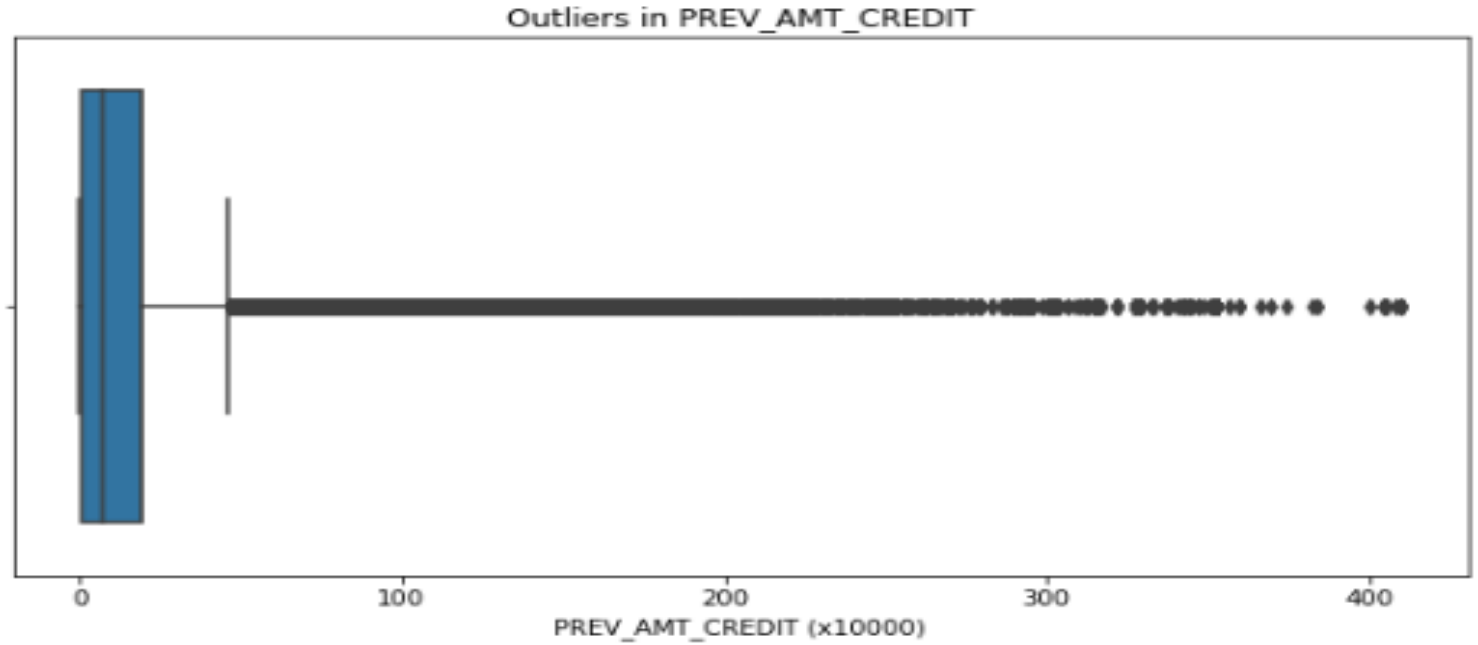
- Continuous values that lie outside the whiskers are not outlier.
- Except for value beyond 4000000 which can be treated as an outlier.



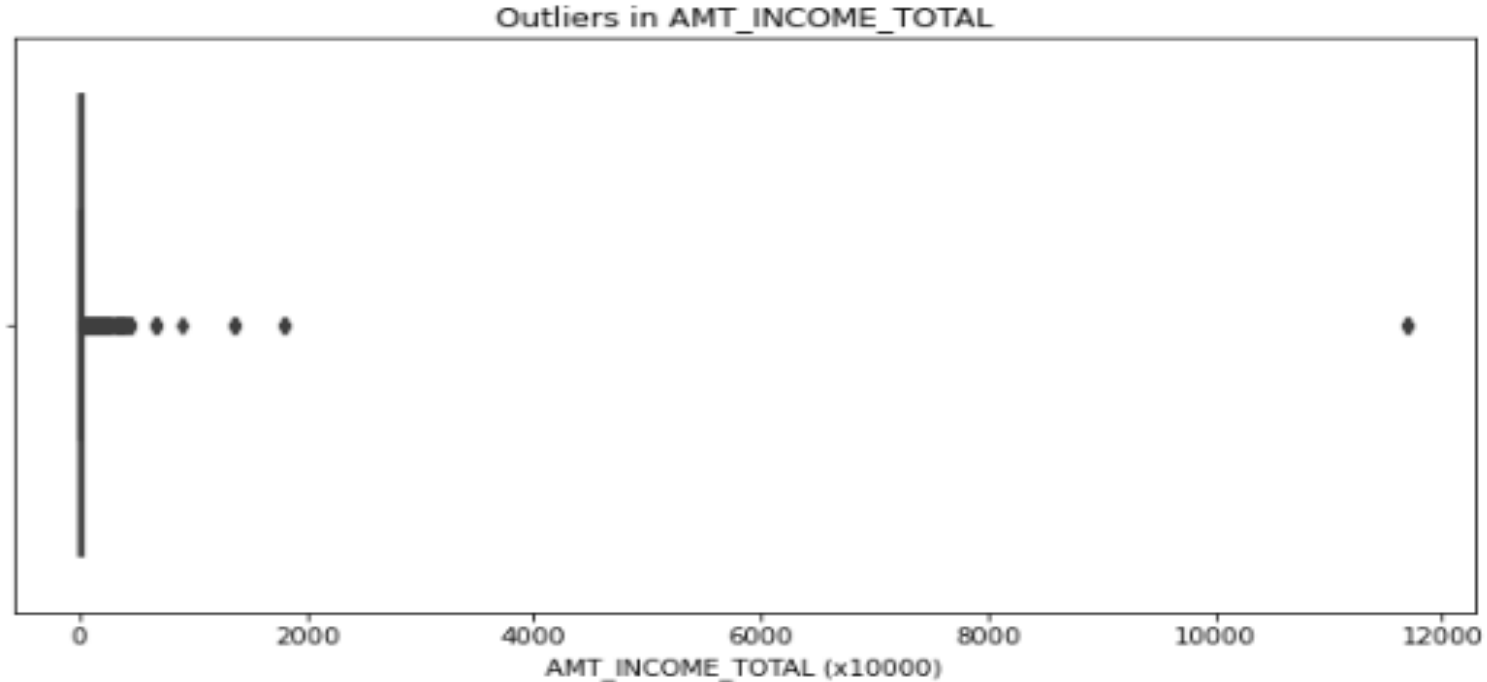
- Continuous values that lie outside the whiskers are not outlier.
- Except for value near 250000 which can be treated as an outlier.



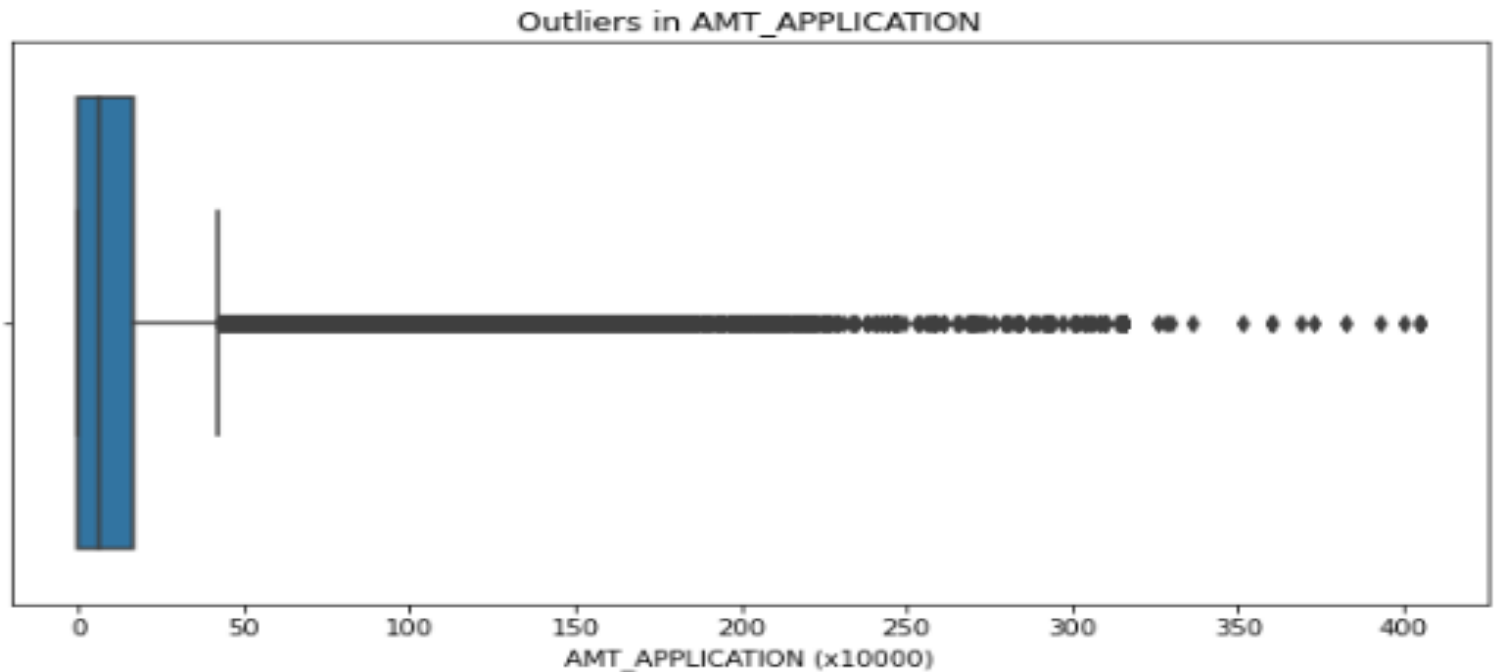
- For Previous Historical data, we observe that there are a few outliers near 400000 for PREV_AMT_ANNUITY value.



- There are continuous set of values , so they all cannot be treated as Outlier.



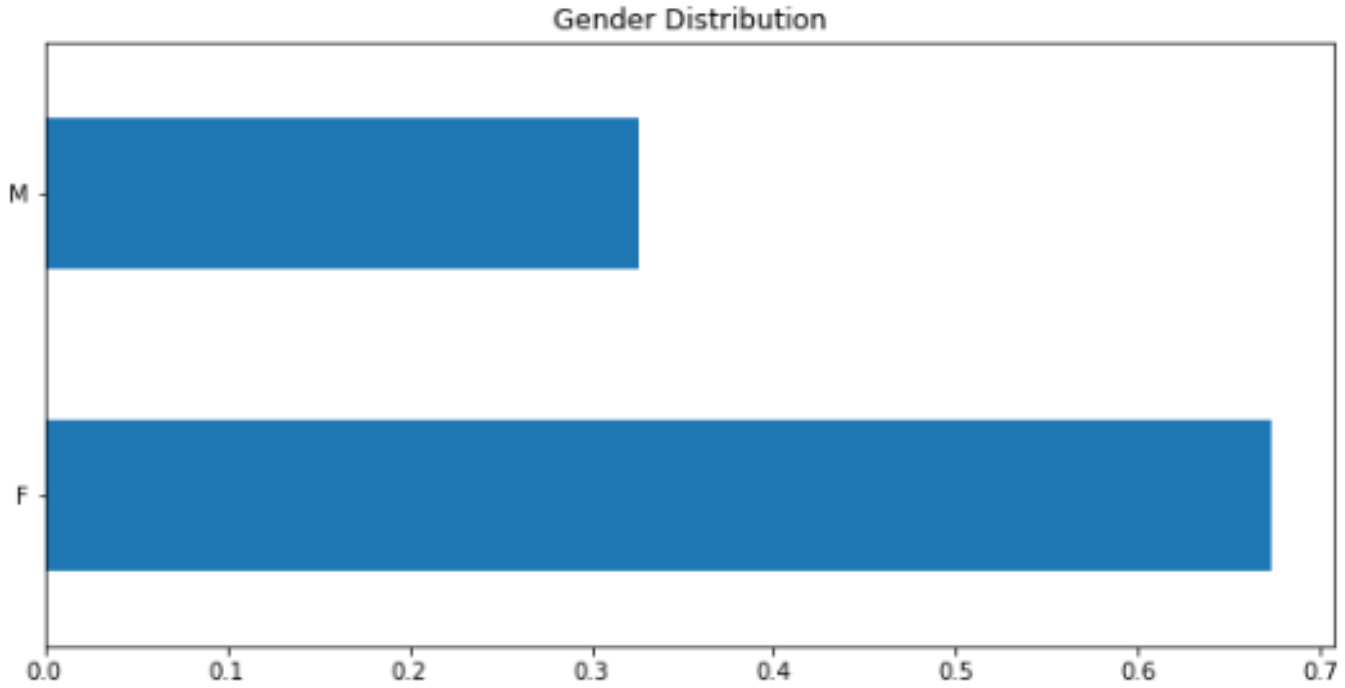
- There are continuous set of values , so they all cannot be treated as Outlier.
- There's is a Outlier present after a huge gap from the continuous distribution near 12000(X10000)



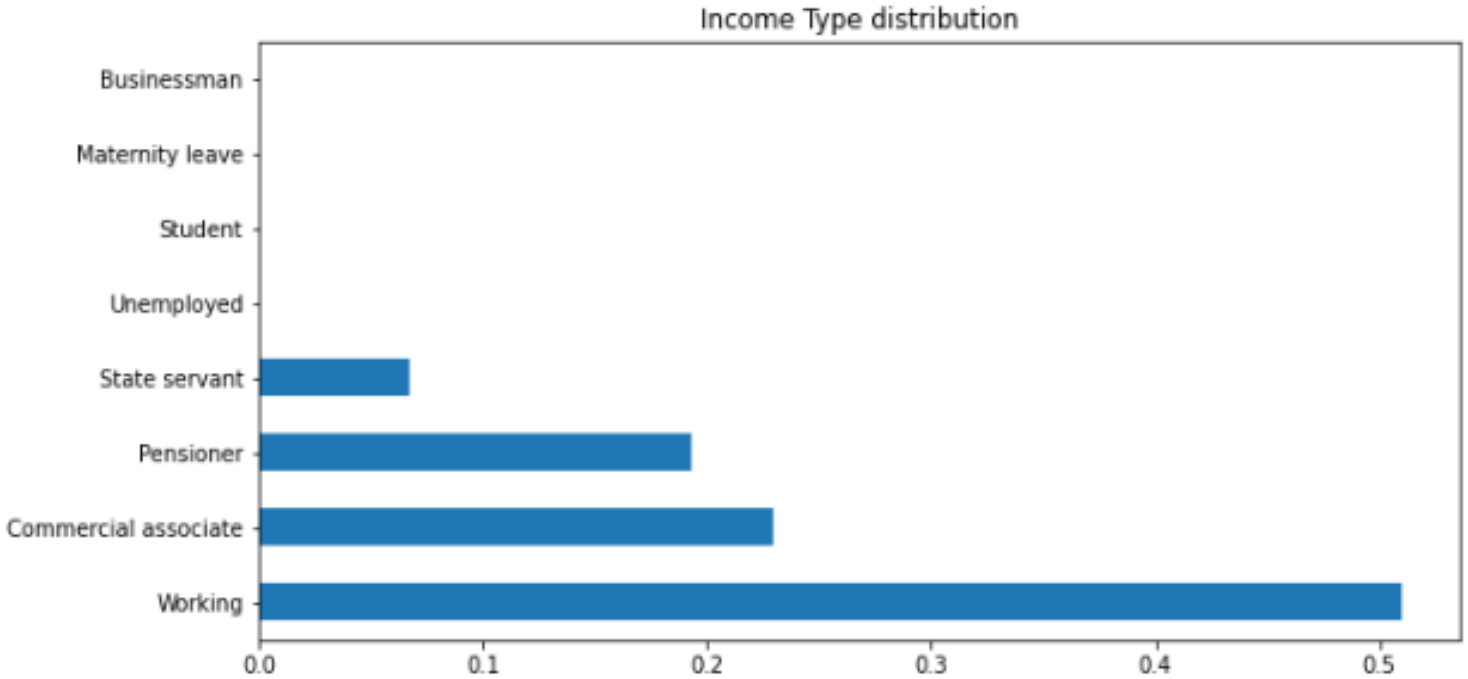
- There are continuous set of values , so they all cannot be treated as Outlier.

Data cleaning: Identify Data imbalance

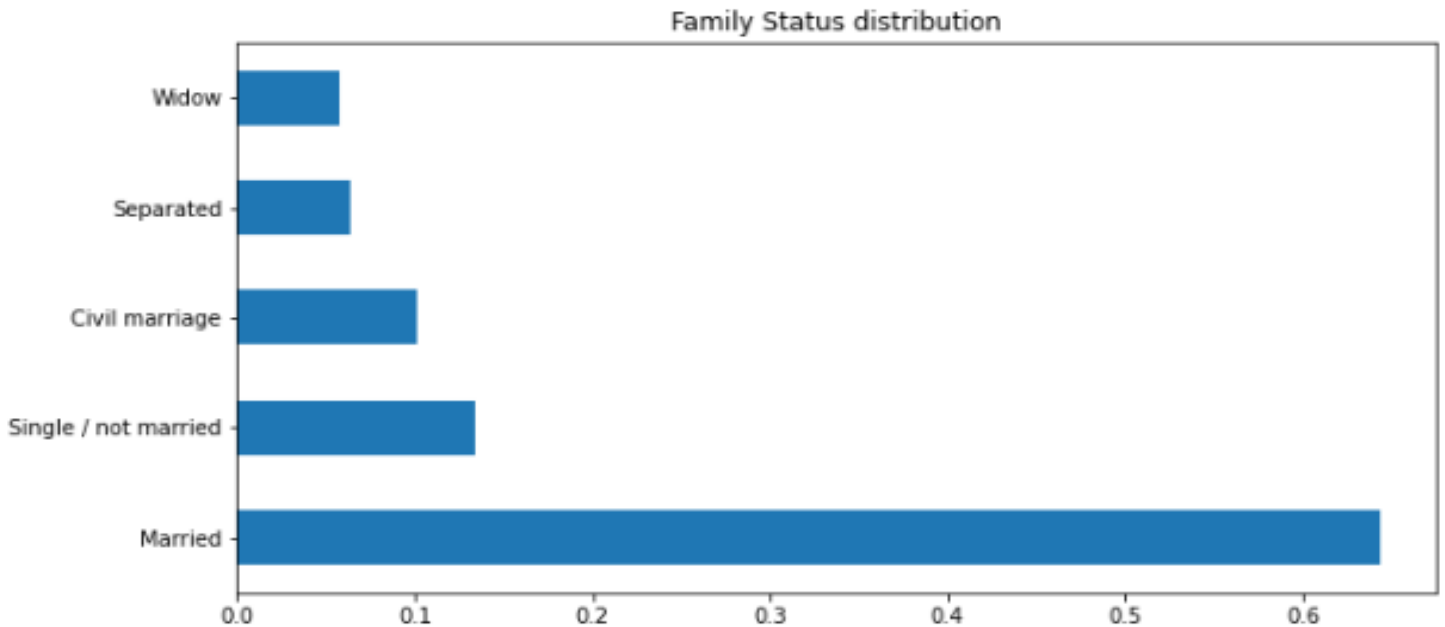
- We found unequal distribution of data in columns and which have been identified graphically using Bar Plot for both Default and non-Default applicants.



- More than 60% of the Total applicants are Females.



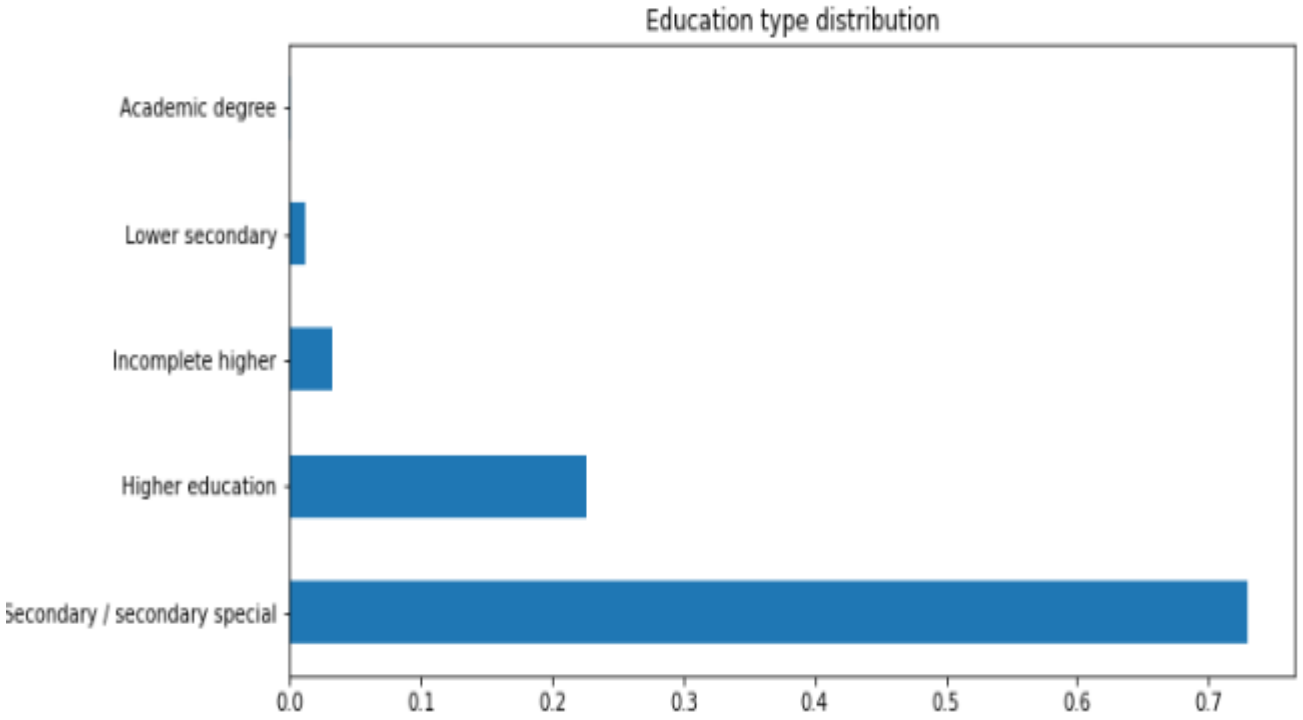
- Applicant having Income type- Working have taken maximum no of Loan.
- Applicants with Income type as Students, Unemployed, Maternity leave and Businessman have taken the least no of loans.



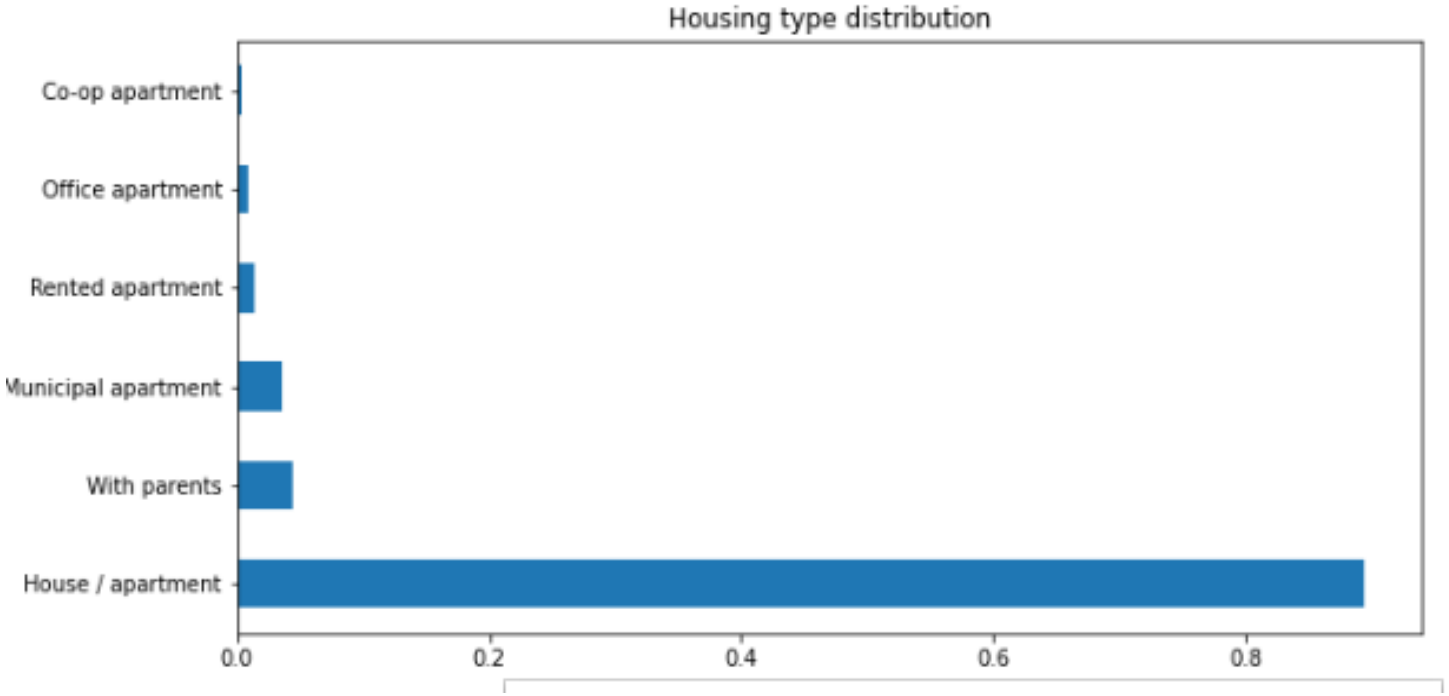
- Among all applicants, married ones are highest in number whereas Widowers and Separated are the least in terms of applicants applied for the loan.

Univariate analysis before segmenting the data frame

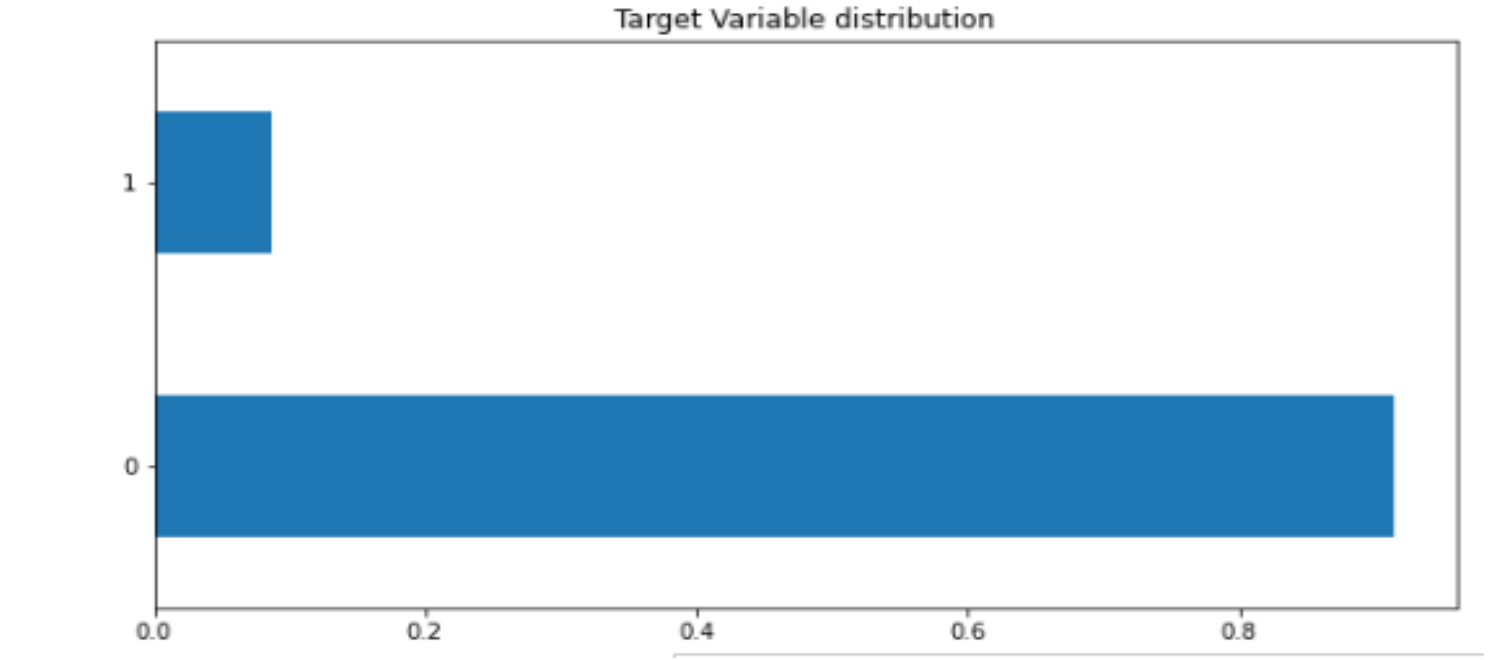
- It covers analysis of single categorical variable . We used dist plot, bar plots and count plots for Univariate analysis.



- Education does play an important role for Loan applicants.
- Most educated people with Education type Secondary/secondary special have shown more interest in applying for a Loan based on the no of applications received.
- Applicants with Highest qualification as Academic degree have taken least no of loans



- Most loan applicants were people living in Housing/apartments type while least no of applicants were residing in Co-op apartments and Office apartments.

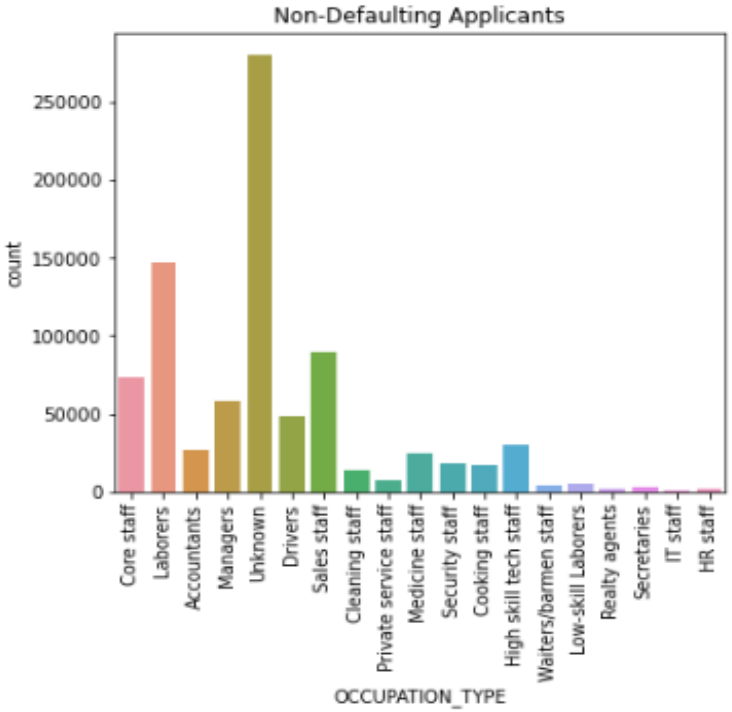


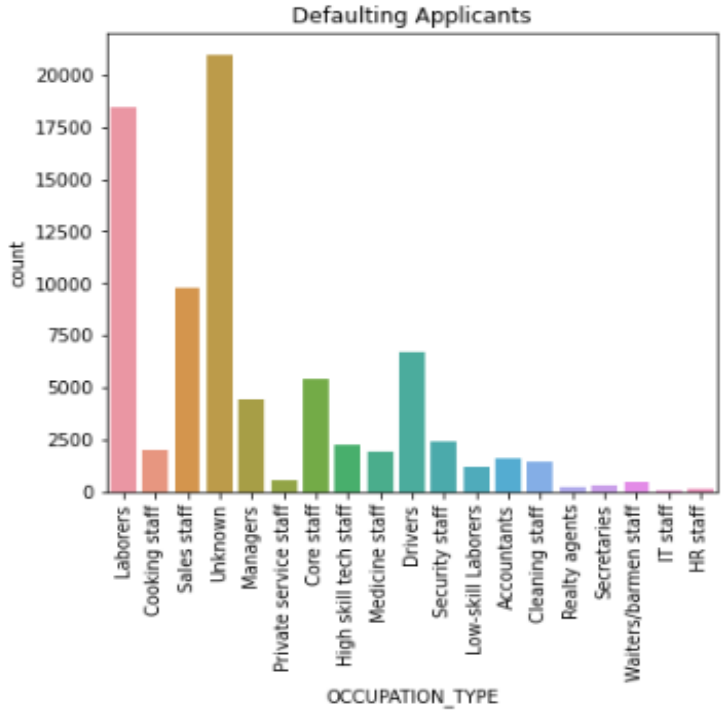
- Around 90% of Loan Takers have not Defaulted i.e. Target variable=0 ,while the remaining Around 10% of the Loan takers are Default i.e. Target Variable=1

Univariate analysis after segmentation:

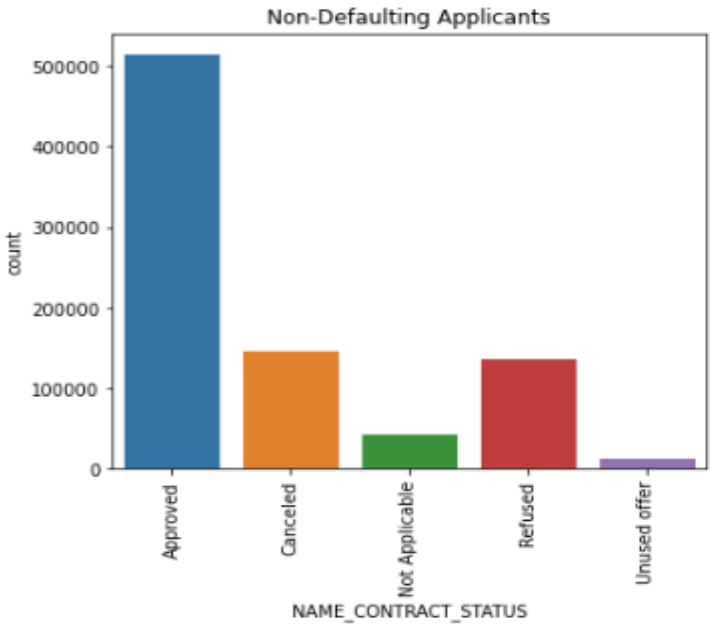
- Based on the Target variable, two segmented data frames (yd and nd data frames) were created representing default and non-default applicants.

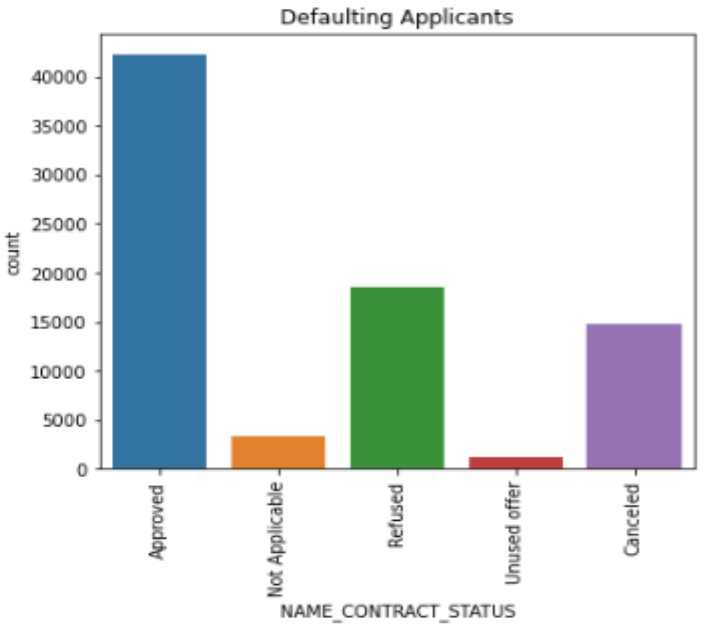
- Categorical values were plotted and identified for visual patterns using Bar chart





- Laborers tends to default the most followed by Sales Staffs and Drivers amongst the known Occupation type.
- HR Staff and IT Staff are the least seeker of loans regardless of their Target type





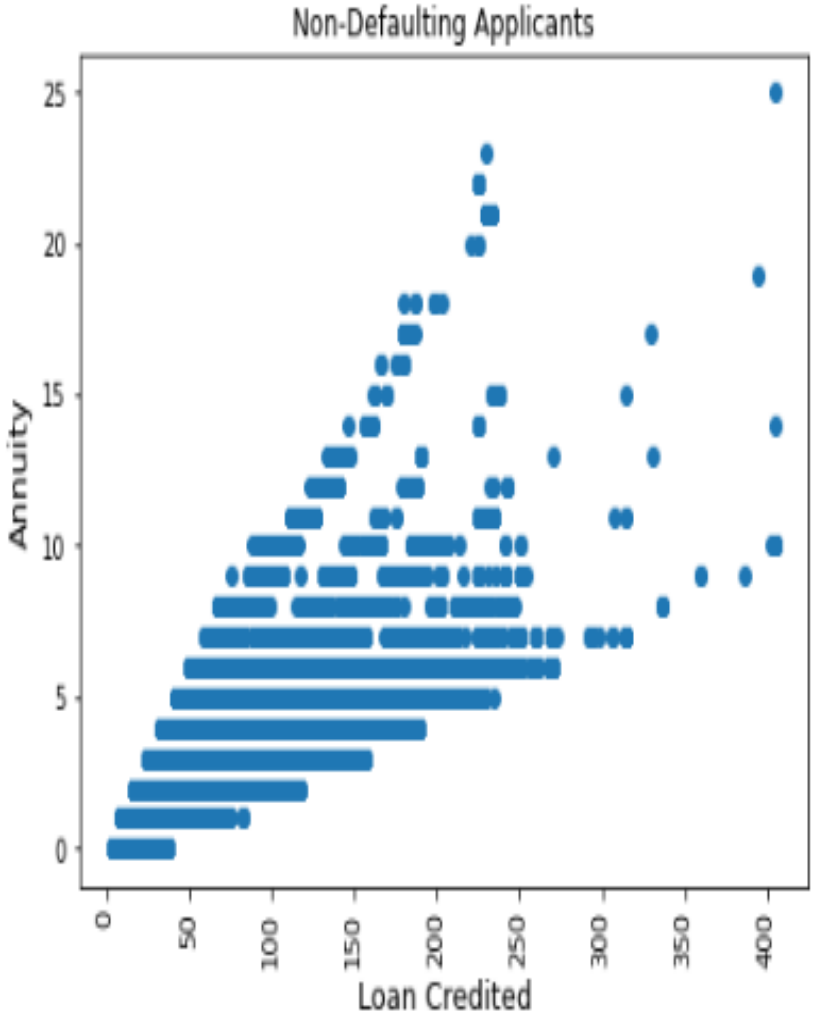
- In case of Non-Defaulters, Cancelled applications are more than Refused ones.
- In case of Non-Defaulters, Refused applications are more than Cancelled ones.

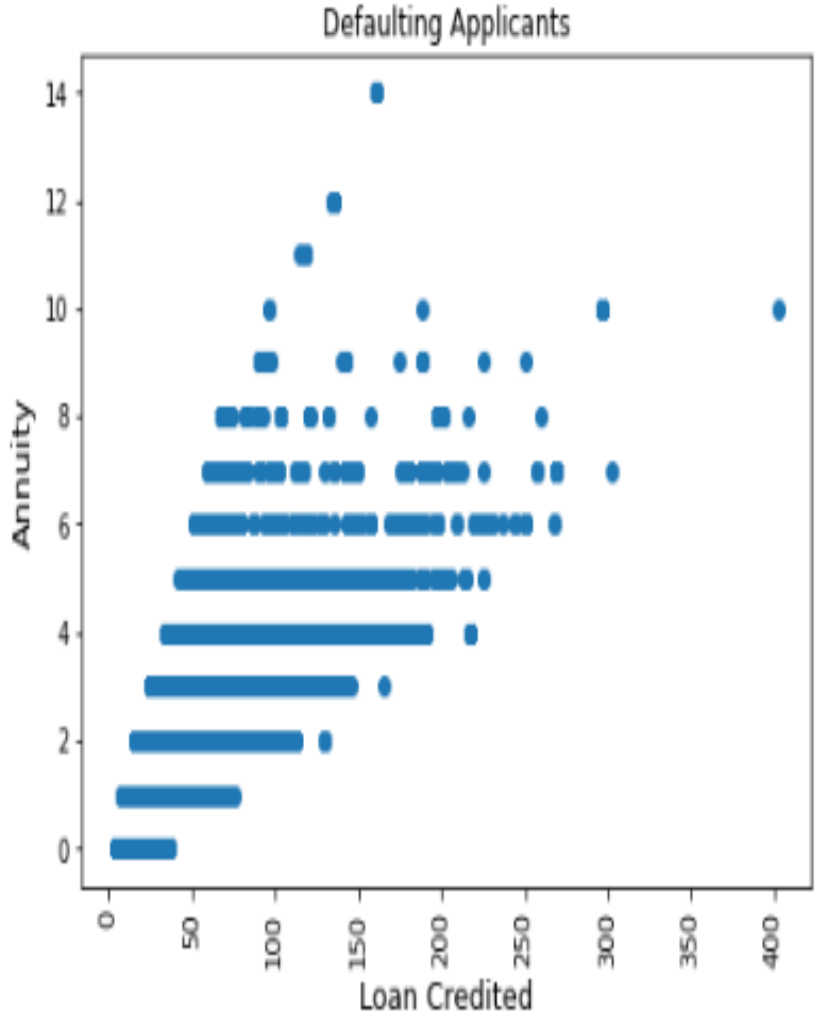
Bivariate analysis

- In Bivariate analysis, we use two separate variables other than the target variable to identify any pattern or relationship between them.

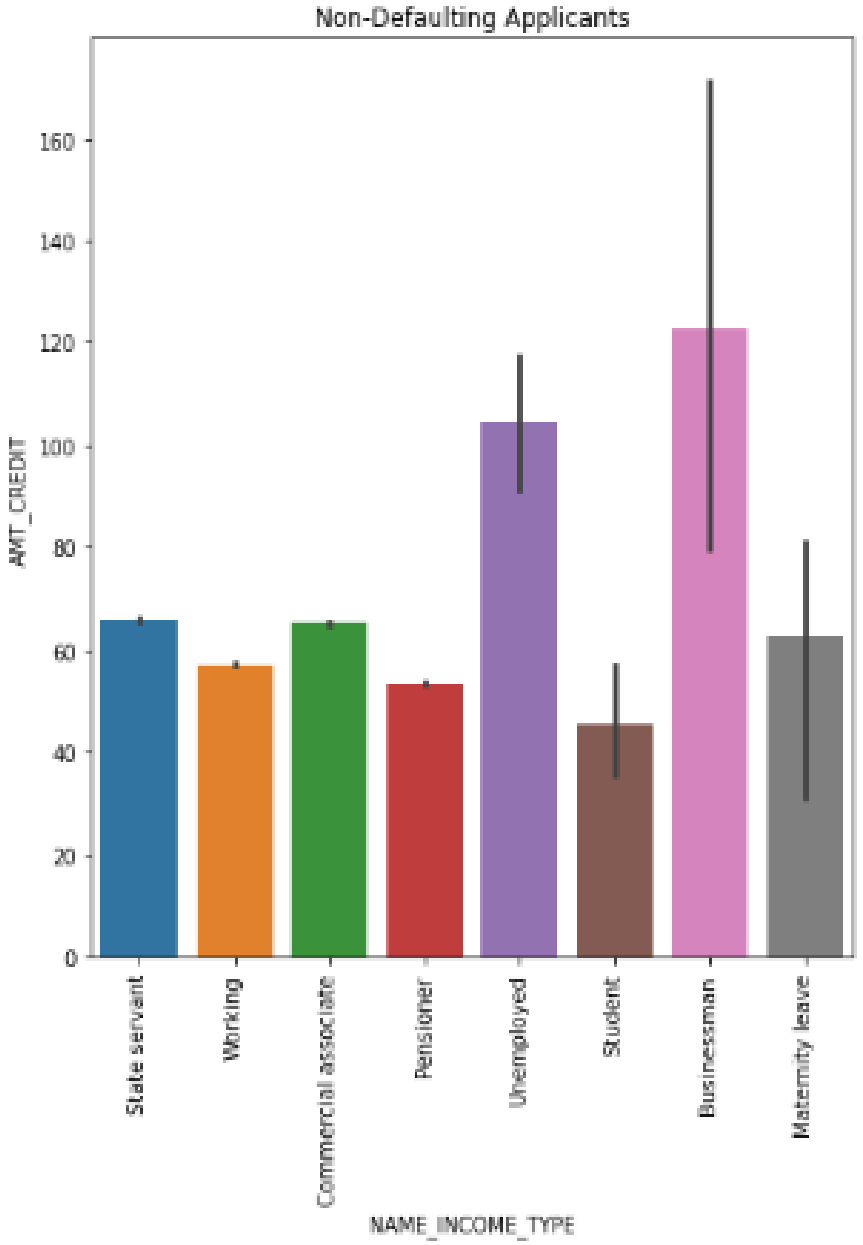
- Scatter Plot, Pair Plot and Heat map are mostly common graphs used for finding and representing the Correlation between the variables.

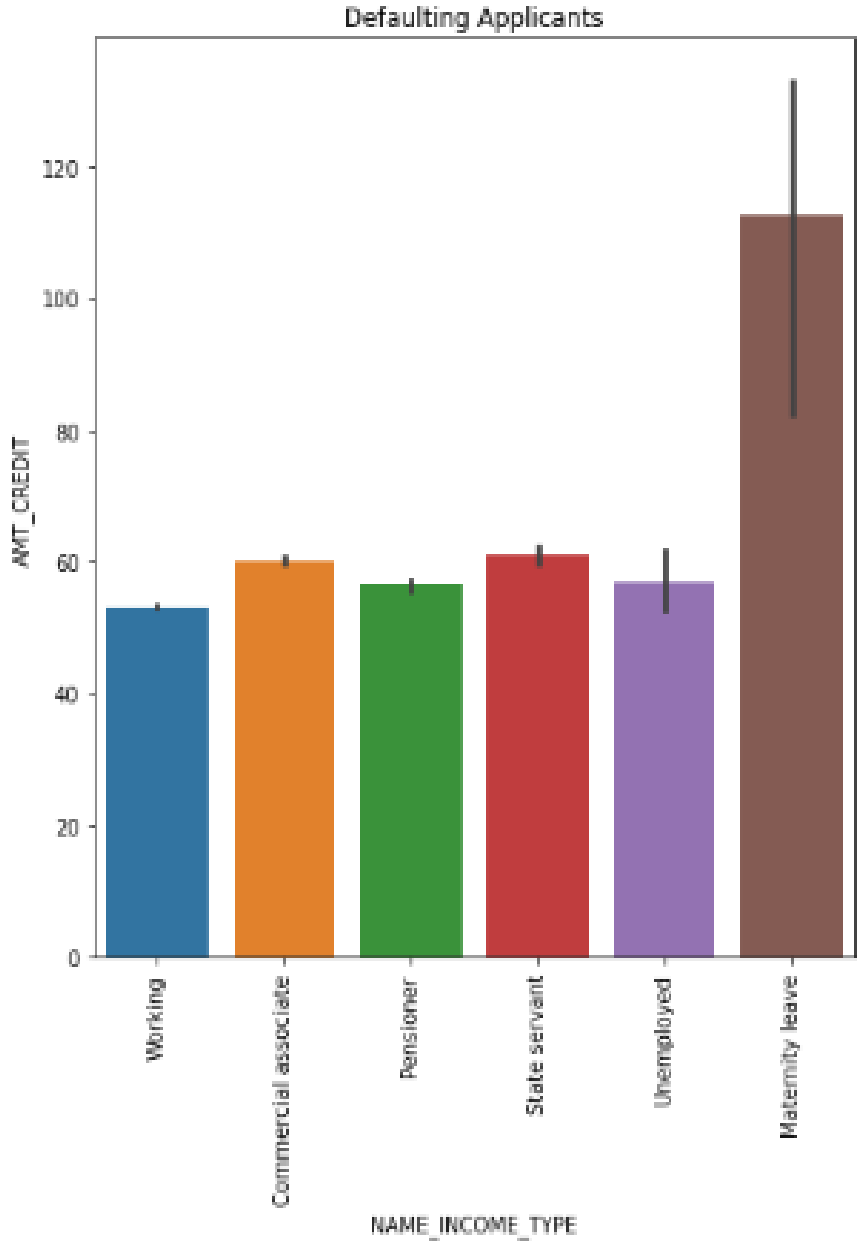
- In Correlation Matrix Heat map, it can be positive correlation i.e indicating a String correlation b/w them and negative correlation i.e indicating weak relation b/w the variables.



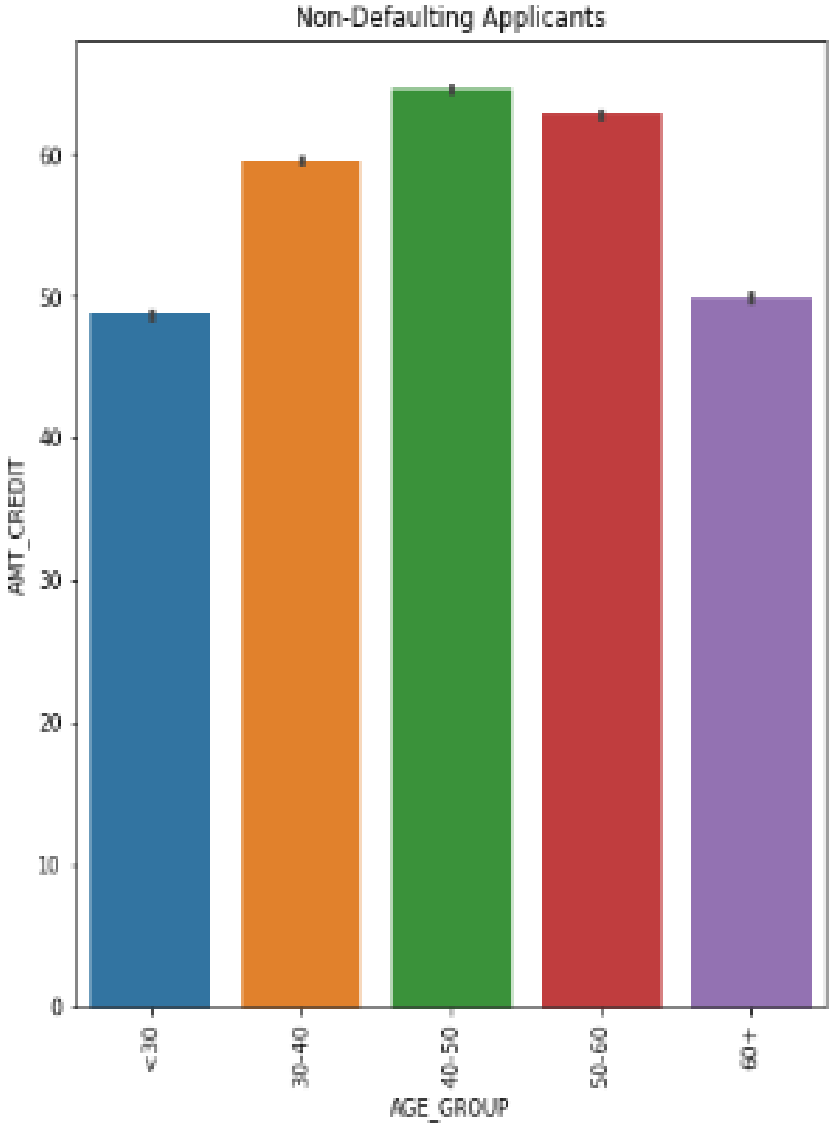


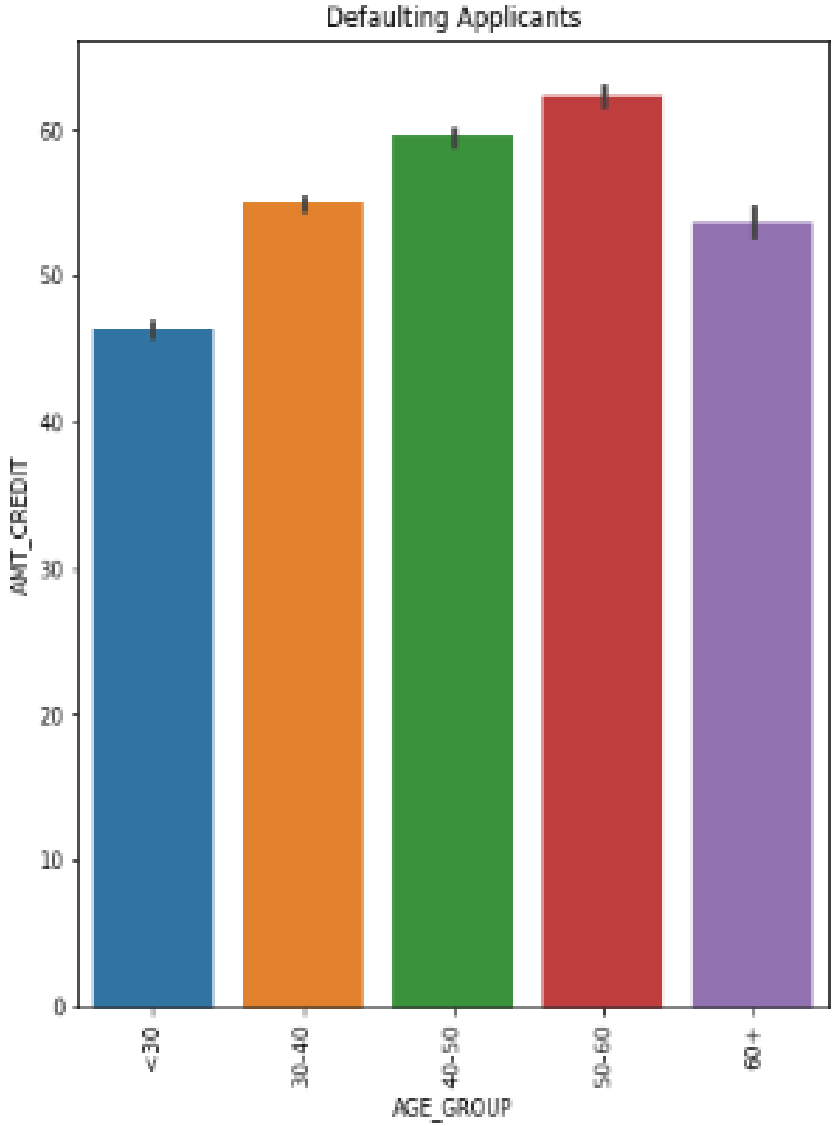
- Defaulters tend to stick to the lower annuity amount when compared to non-defaulters.
- The plot shows that Annuity vs Loan Credited for defaulters has a sharper incline when compared to non-defaulters.



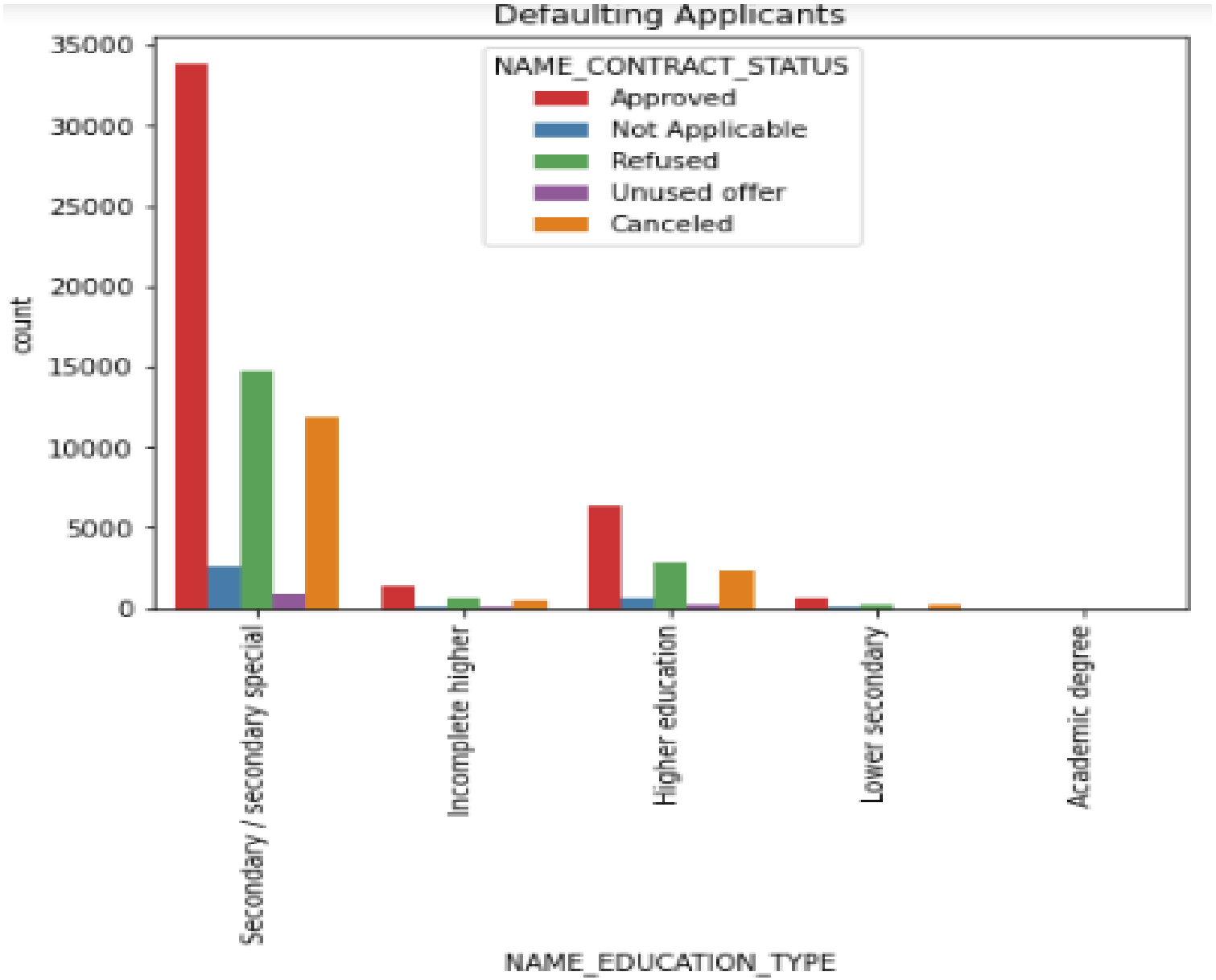


- Women applying for loan during their maternity leave are more likely to default.
- Applicants with income type Business are the highest seeker of loan who pay on time, followed by unemployed applicants.
- Businessmen and Students loan applicants are not likely to default.





- Amount of credit among the defaulters is most in the age-group 50-60 and least in the age group within 30.
- Amount of credit among the non-defaulters is most in the age group 40-50.



- Default applicants with Highest education as 'Secondary/secondary special' have had maximum Approved loans, as well as highest Refused and Cancelled Loans in comparison to Applicants who are less educated.

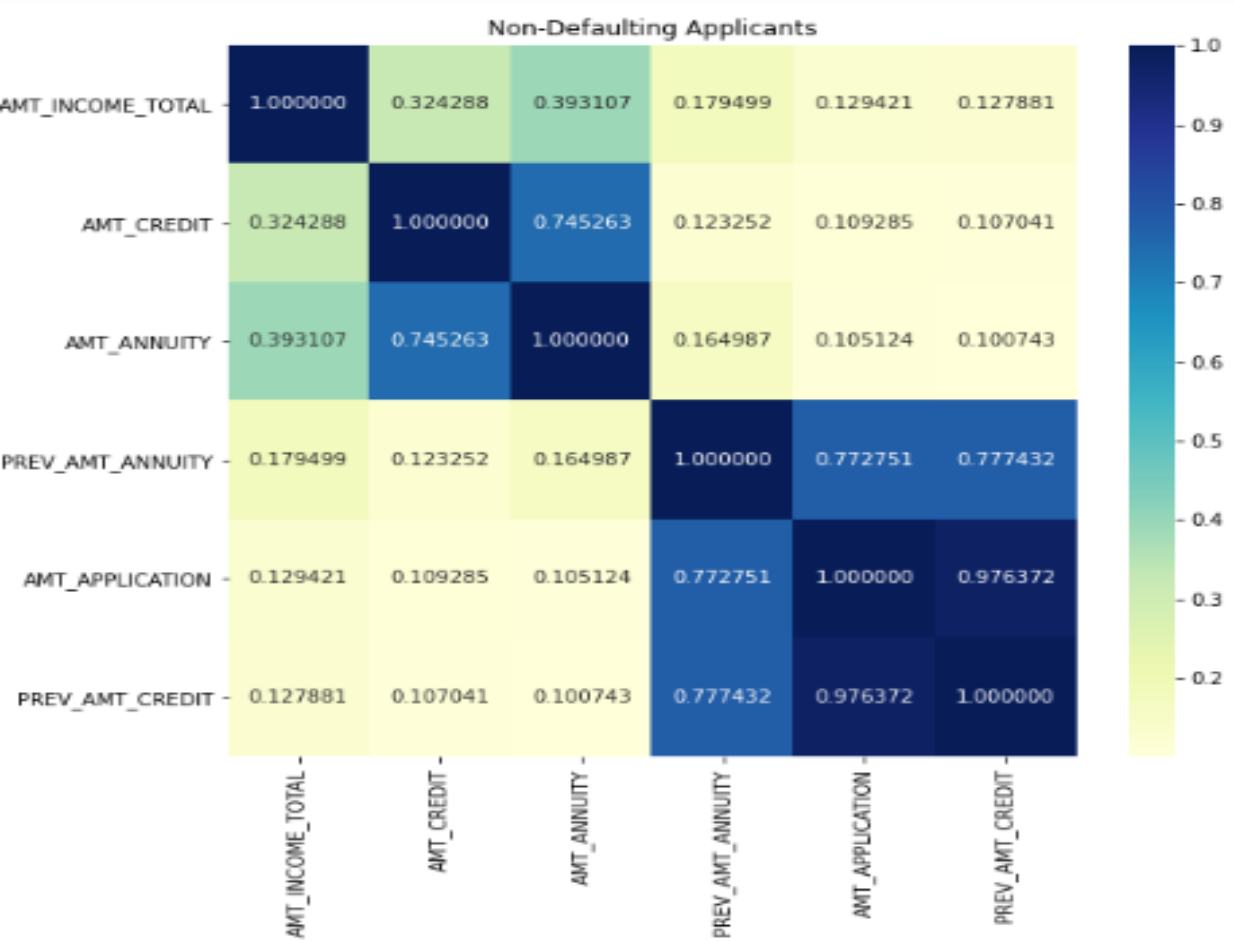
Correlations

To identify Correlations, we use following types of plots to visualize:

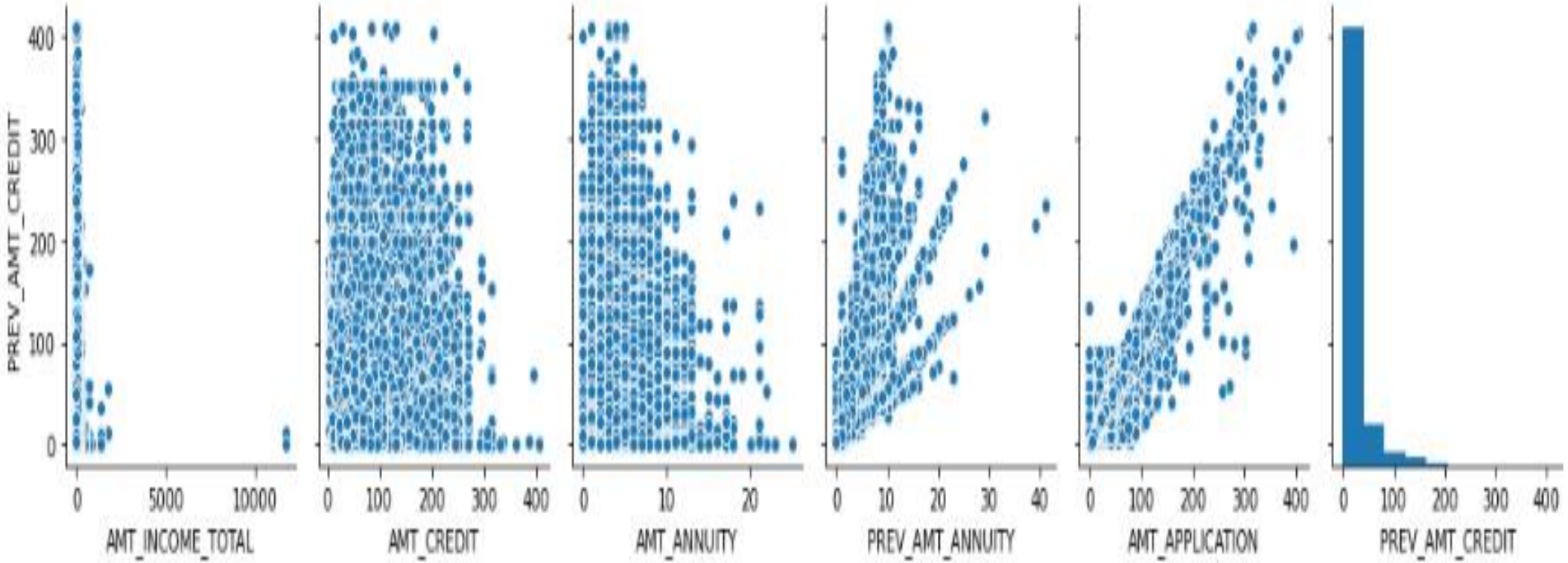
- Scatter plot
- Pair plot
- Correlation matrix
- Heat map

Correlation identified as part of analysis:

- For Target Variable(1) i.e. Default customers
- For Target variable(2)i.e. non-Default customers.



- Annuity amount is very low for the defaulters when seen against their Total Income .
- Amount credited to the defaulters are considerably less with respect to Total Income when compared from Non-Defaulter.
- Previous Loan Annuity amount is also very low for the defaulters when seen against their Total Income .
- When comparing Annuity with Amount Credited, there is not much difference between Defaulters and Non-Defaulters.
- The correlation of Credit asked from the applicant in previous application and the amount credited for the previous application is higher regardless of the applicant being defaulter or non-defaulter.



- As observed in the Pair plot, there is a correlation between PREV_AMT_CREDIT, PREV_AMT_ANNUITY & AMT_APPLICATION.

Overall, there is a strong, positive correlation between:

Pair plot:

- PREV_AMT_CREDIT increases with increase in the PREV_AMT_ANNUITY variable.

- PREV_AMT_CREDIT increases with increase in the AMT_APPLICATION variable.

Heatmap:

- AMT_CREDIT have a positive correlation & increases with increase in the AMT_ANNUITY

- PREV_AMT_ANNUITY have a positive correlation & increases with increase in the AMT_APPLICATION

- AMT_ANNUITY and AMT_INCOME_TOTAL have a correlation of around 0.39 .

- For Default applicants, correlation value between AMT_CREDIT and PREV_AMT_CREDIT is higher as compared to the non-Defaulters.

Thank You