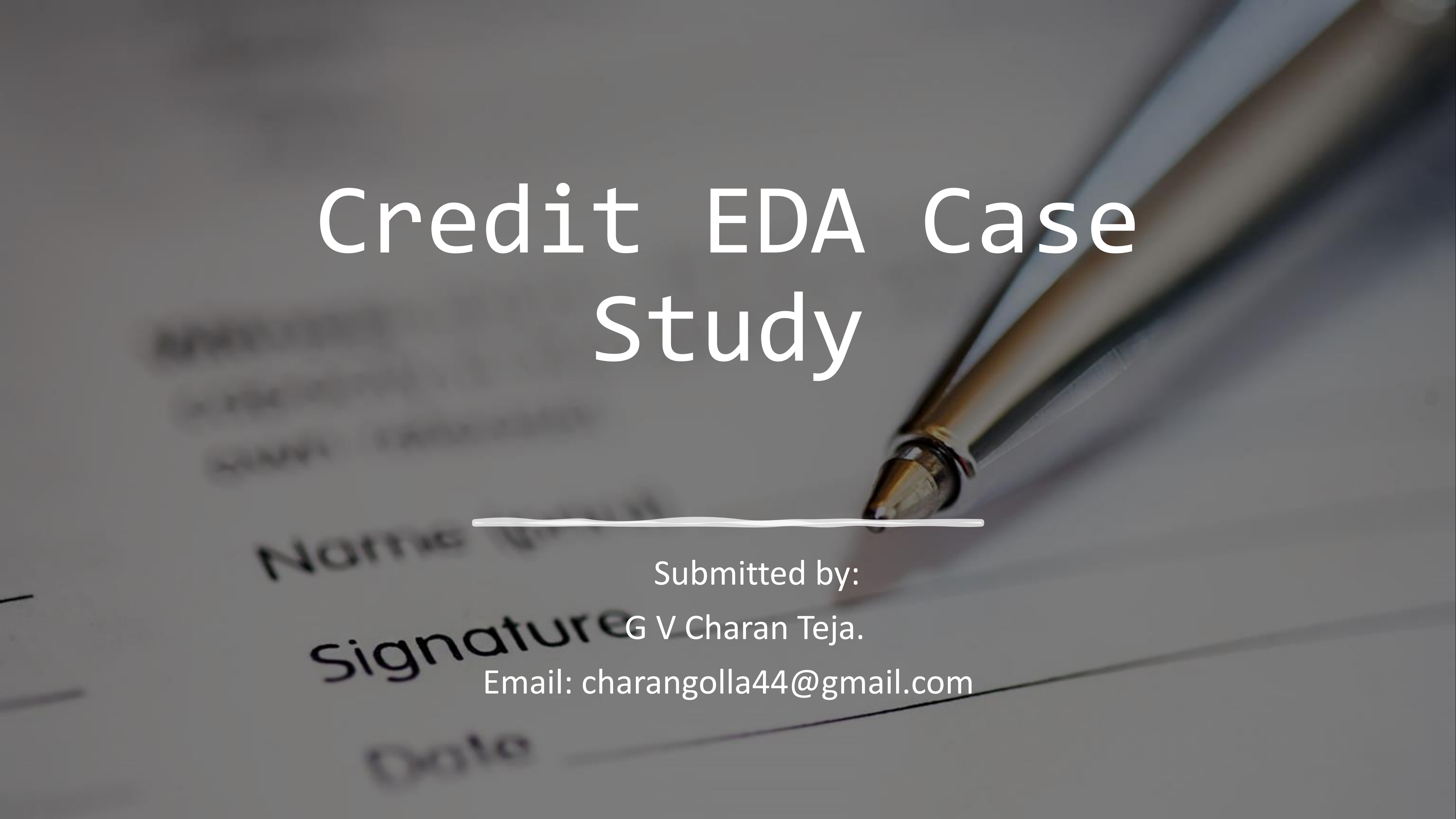


Credit EDA Case Study

Submitted by:

G V Charan Teja.

Email: charangolla44@gmail.com

Objective:

- The main objective of this EDA is to find what are the driving factors which resulting in customer/client defaulting the loan.

Assumptions:

- Customers having less income likely to default more.
- Customers whose age is less than 20 are more likely to default as they don't have income.
- Customers who have less credit scores are more likely default the loan.
- Blue collar workers are more likely to default the loan.

EDA Approach:

- Business Understanding
- Data Understanding
  - Structure of the data.
  - Numerical summary statistics.
  - Checking the data types of the variables.
  - Identifying if there are null values in columns.
  - Checking if there any duplicate rows in the data.

EDA Approach:

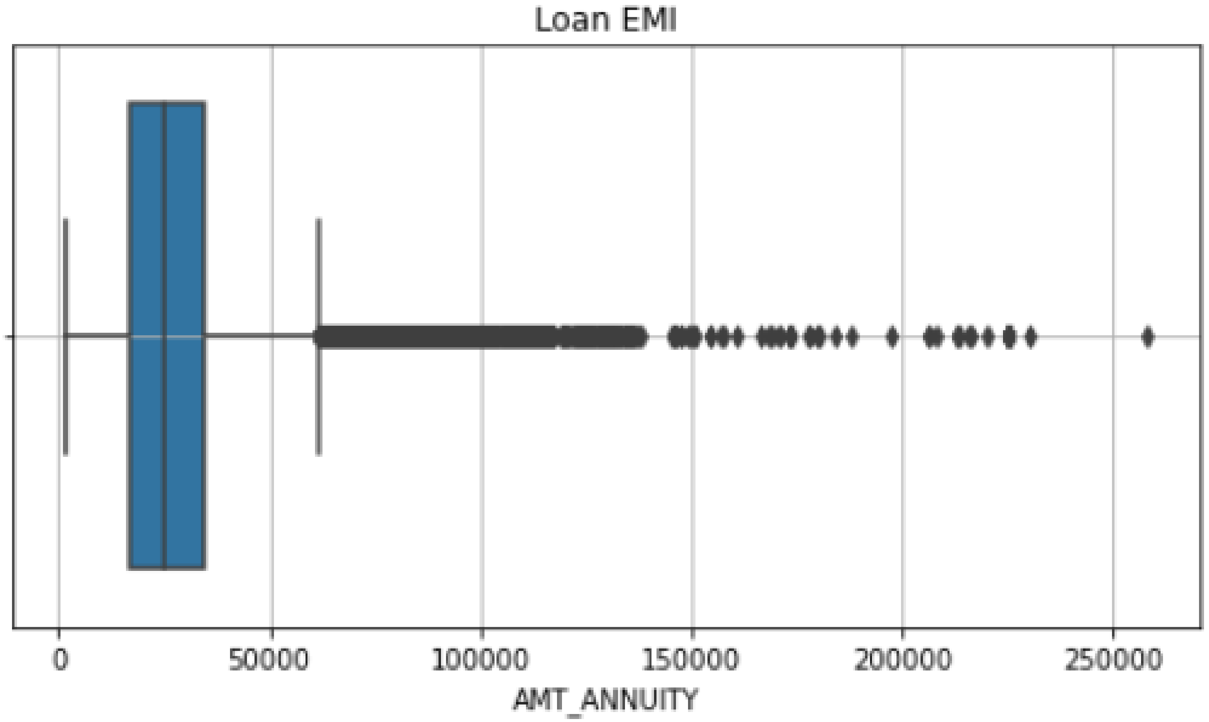
- Data quality check and handling error in data.
- Missing Value Check.
- Outlier Check.
- Binning Of Continuous Variables.
- Data Imbalance Check.
- Univariate Analysis.
- Univariate Segmented Analysis.
- Bivariate Analysis.

Missing Value Treatment:

- Missing values are imputed in different ways for numerical column and categorical column.
- *For numerical column: W*e can impute with "MEAN" if there are no outliers in the column or we can impute with "MEDIAN" if there are outliers in the column
- *For categorical column: W*e can impute with "MODE“ or we can create another category as "Unknown" to avoid inaccurate analysis.

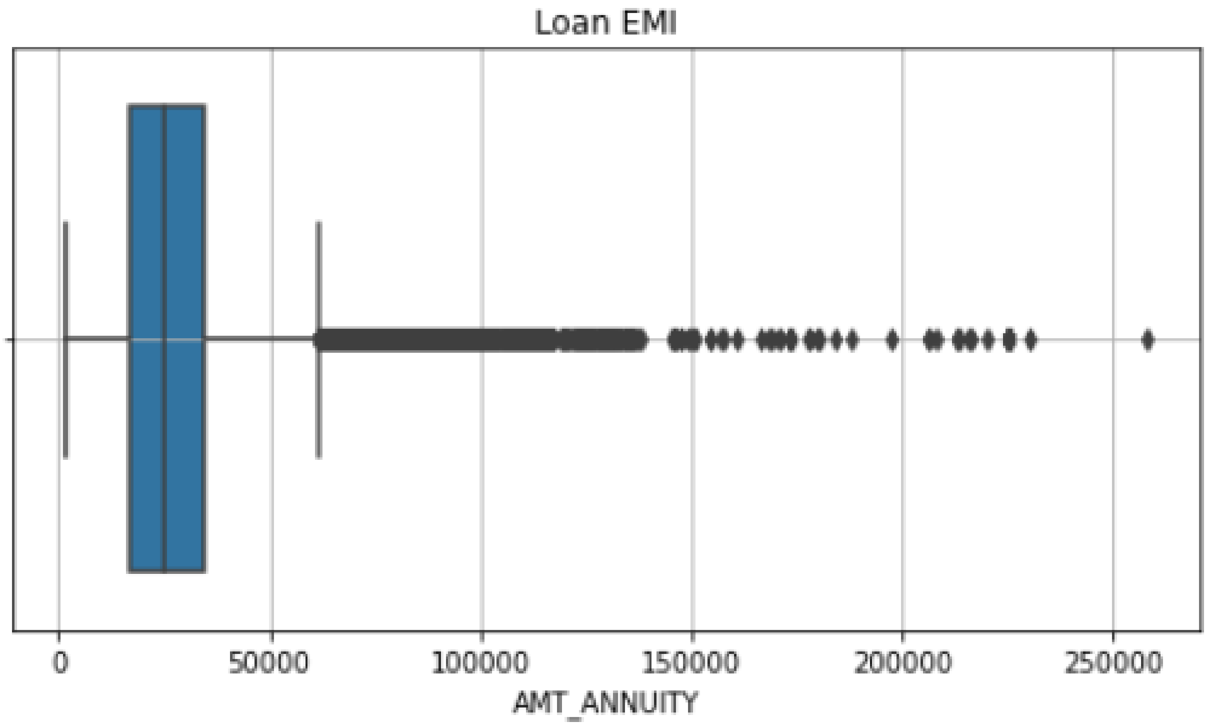
Outlier Check:

- Outlier check for any column can be done in two ways:
  - We can use BOX PLOT as shown in figure to detect outliers in the data.
  - We can use summary statistics to check if there are outliers.
  - If the difference between 99 quartile and max is large, we can say that there are outliers in that column



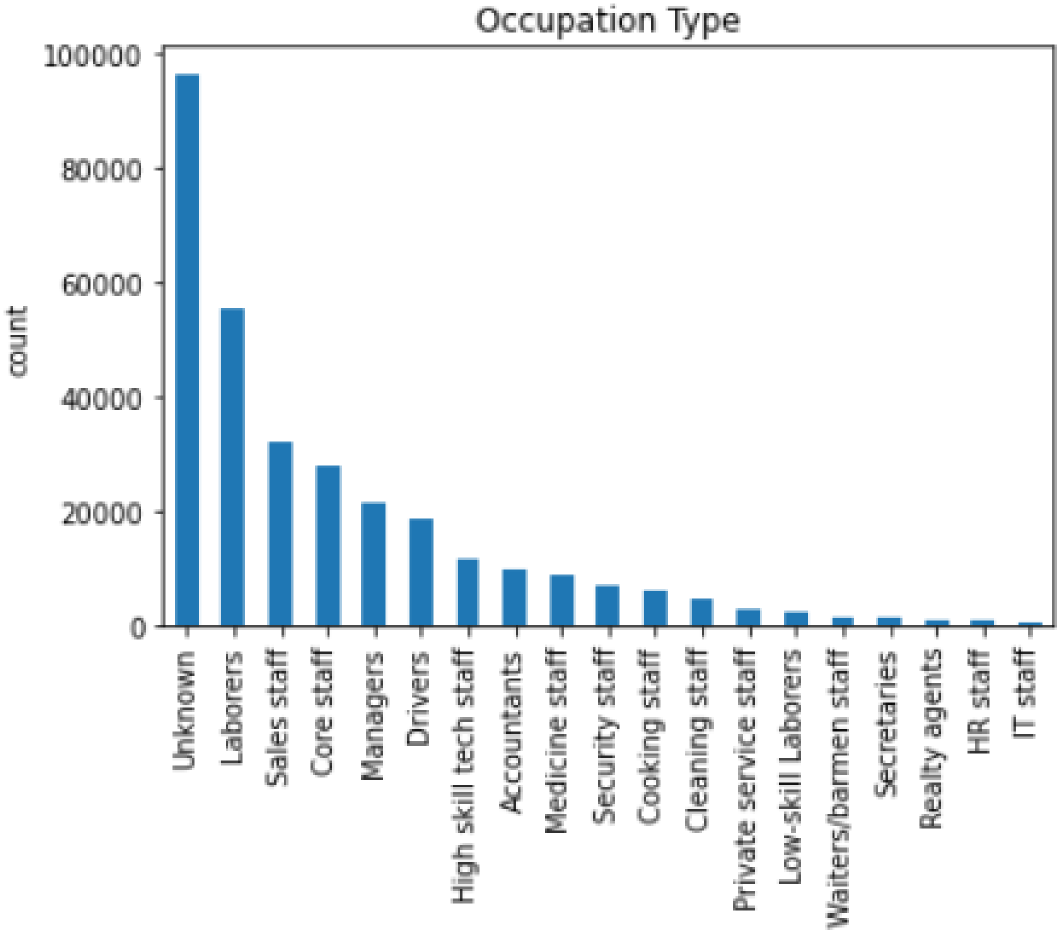
Outlier Treatment:

- Approach to the treat outliers:
  - Deletion of outliers.
  - Binning of values.
  - Capping the Outliers



Occupation Types (Univariate Analysis)

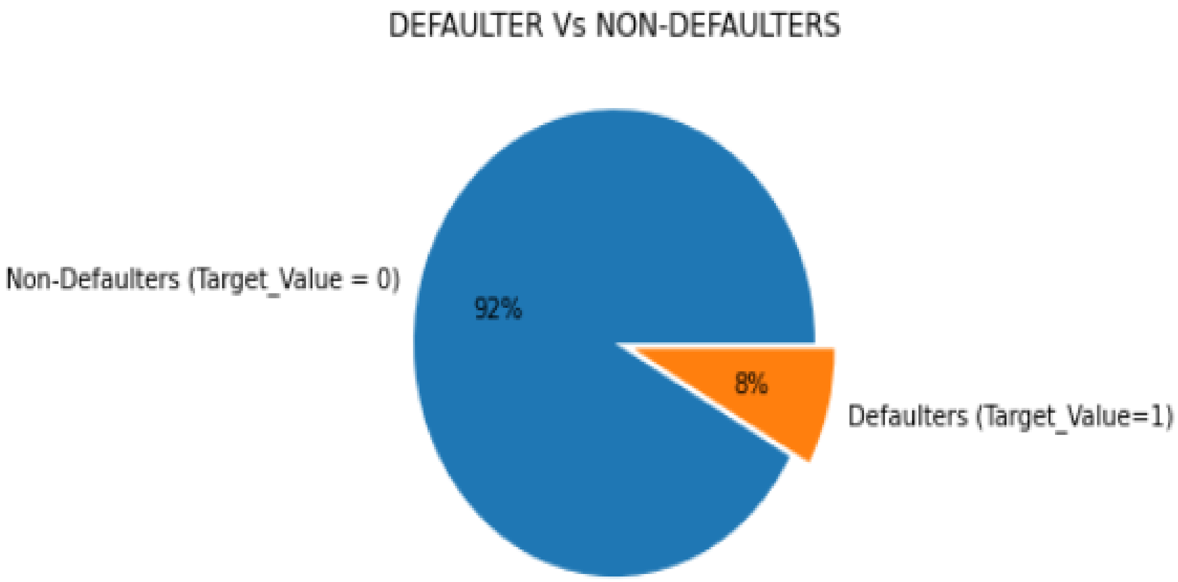
- We can see that there are more laborers in our data set followed by sales staff.
- The Unknown column is imputed in missing values. Hence not considering that category



Data Imbalance Check:

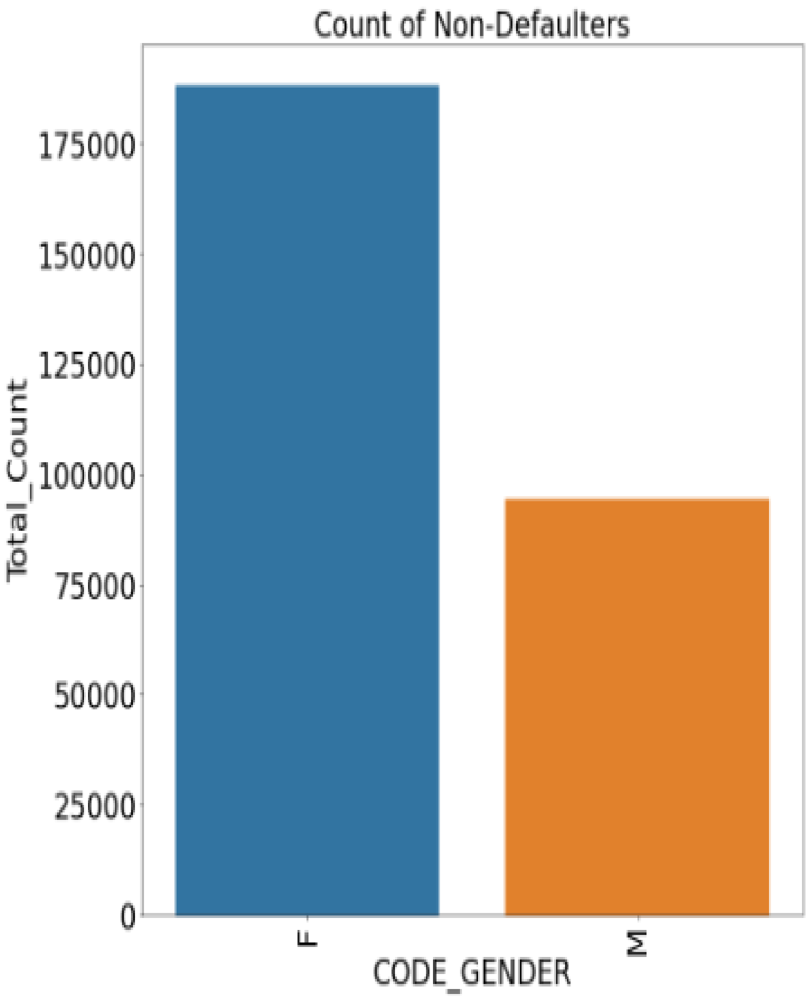
- From pie chart, we can say that data imbalance is high between defaulters and non-defaulters

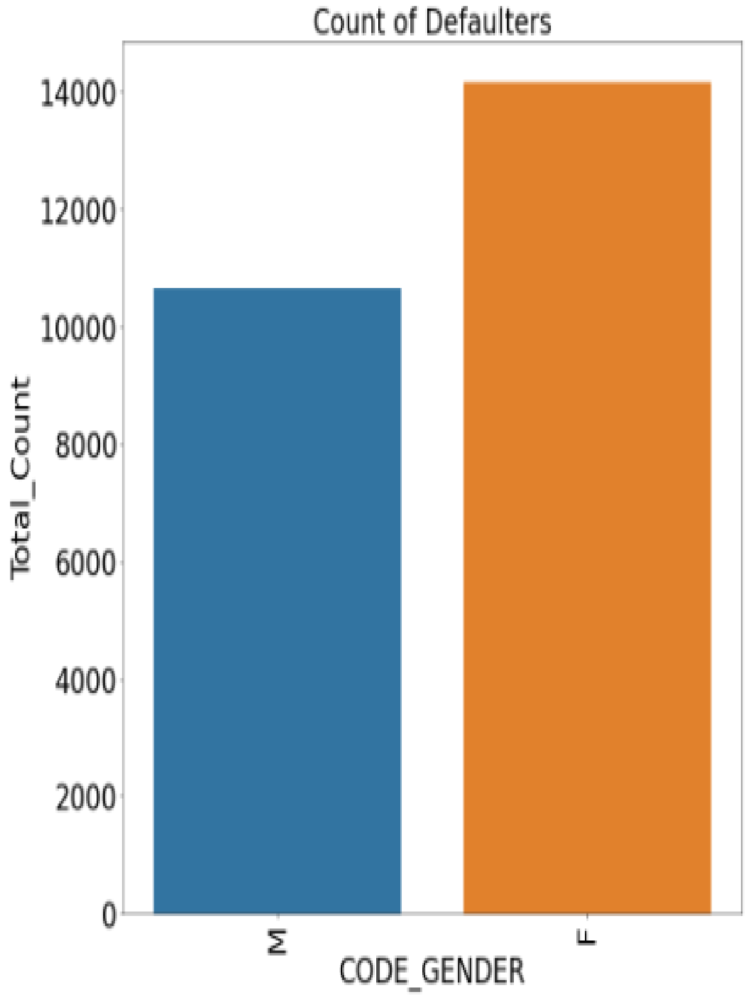
- Also, we can say that 92% of the customers are paying their loans and only 8% of the customers are defaulting their loans.



Analysis Based On Gender (Univariate Analysis) :

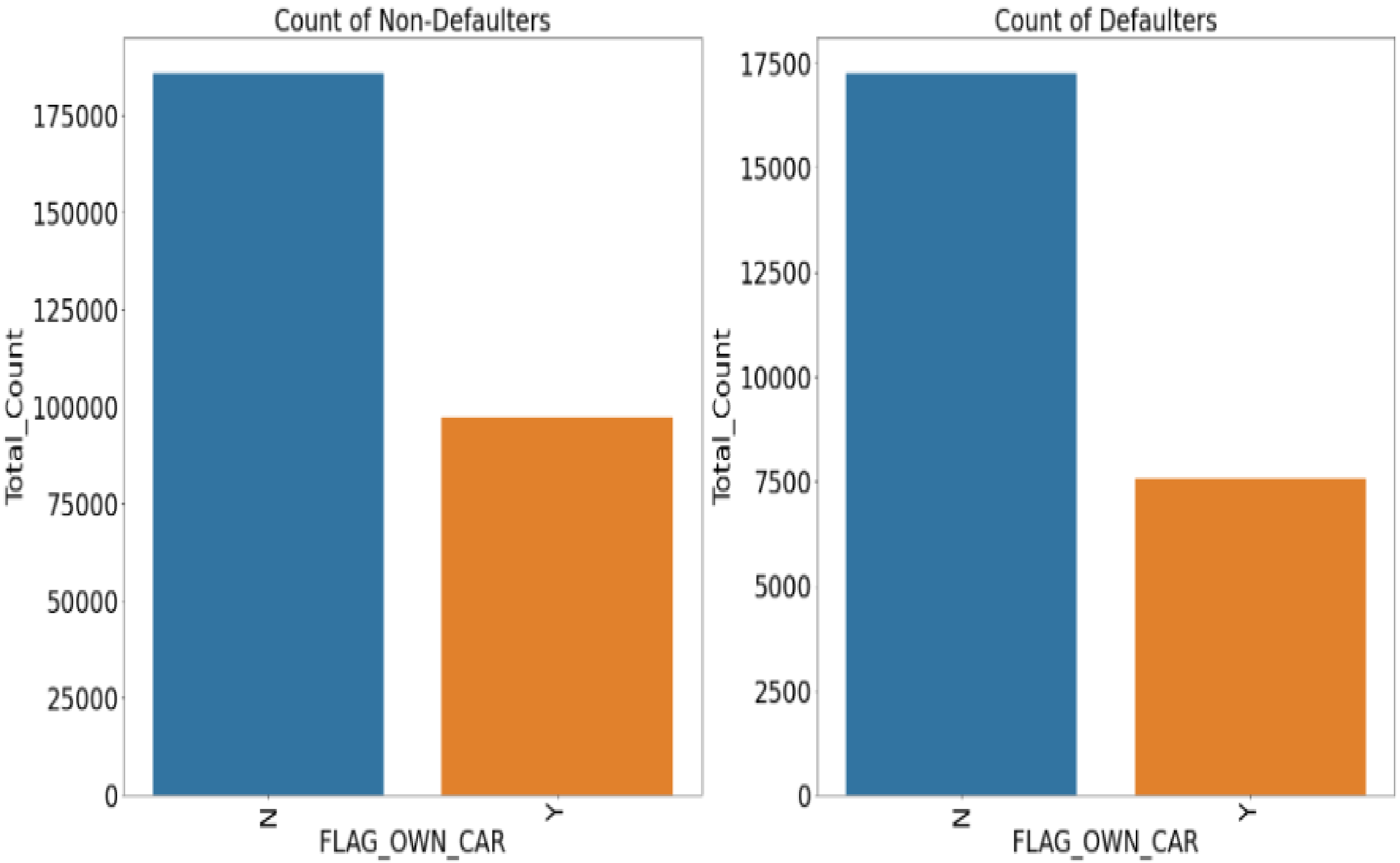
- From the plots we can see that there are more females are applying for the loans.
- Also the female defaulters are high compared to male defaulters this is because of the reason that female applicants are high



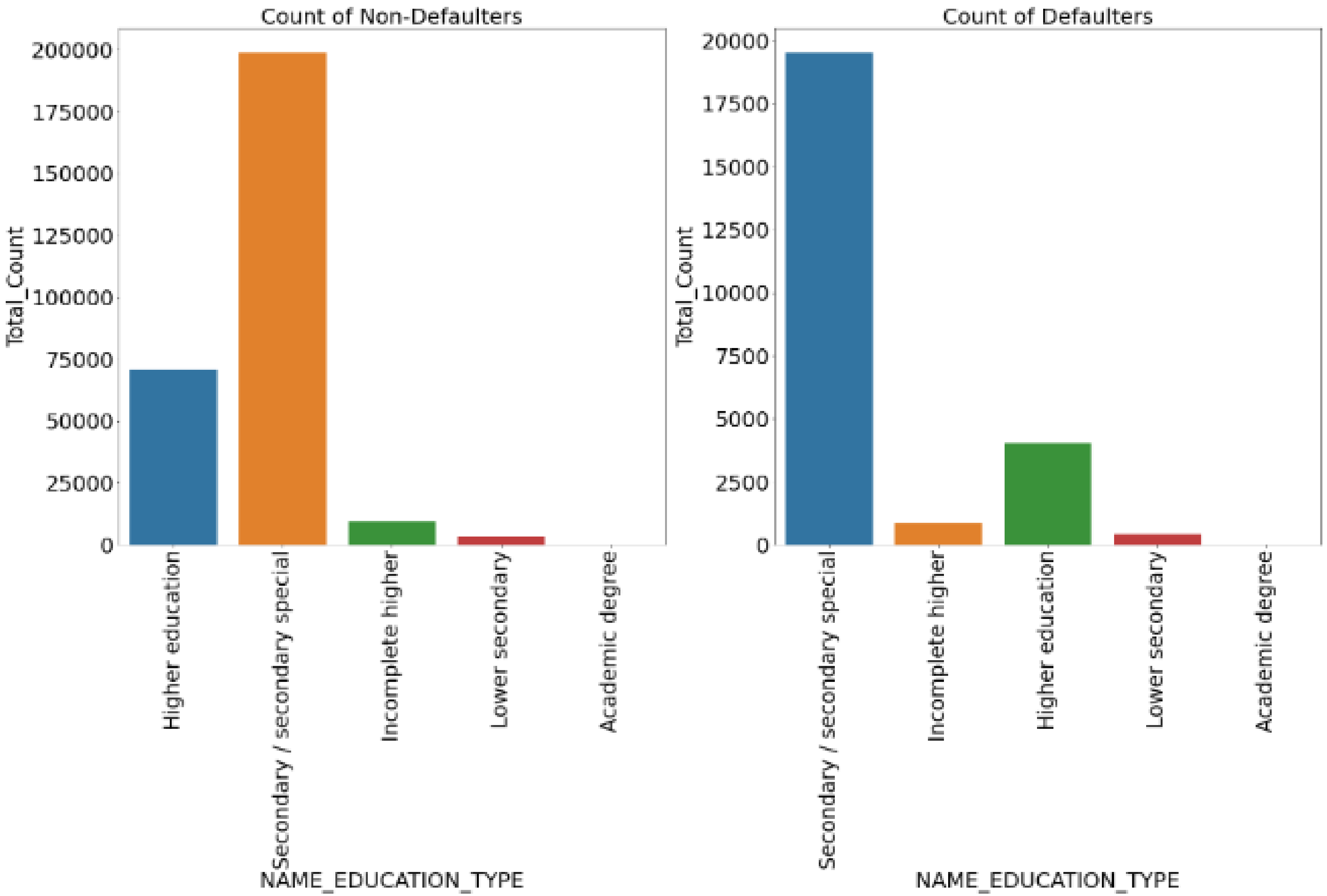


Analysis Based On Whether Customer Own Car (Univariate Analysis) :

- We can see that people who own car are less likely to default compared to the people who doesn't own car

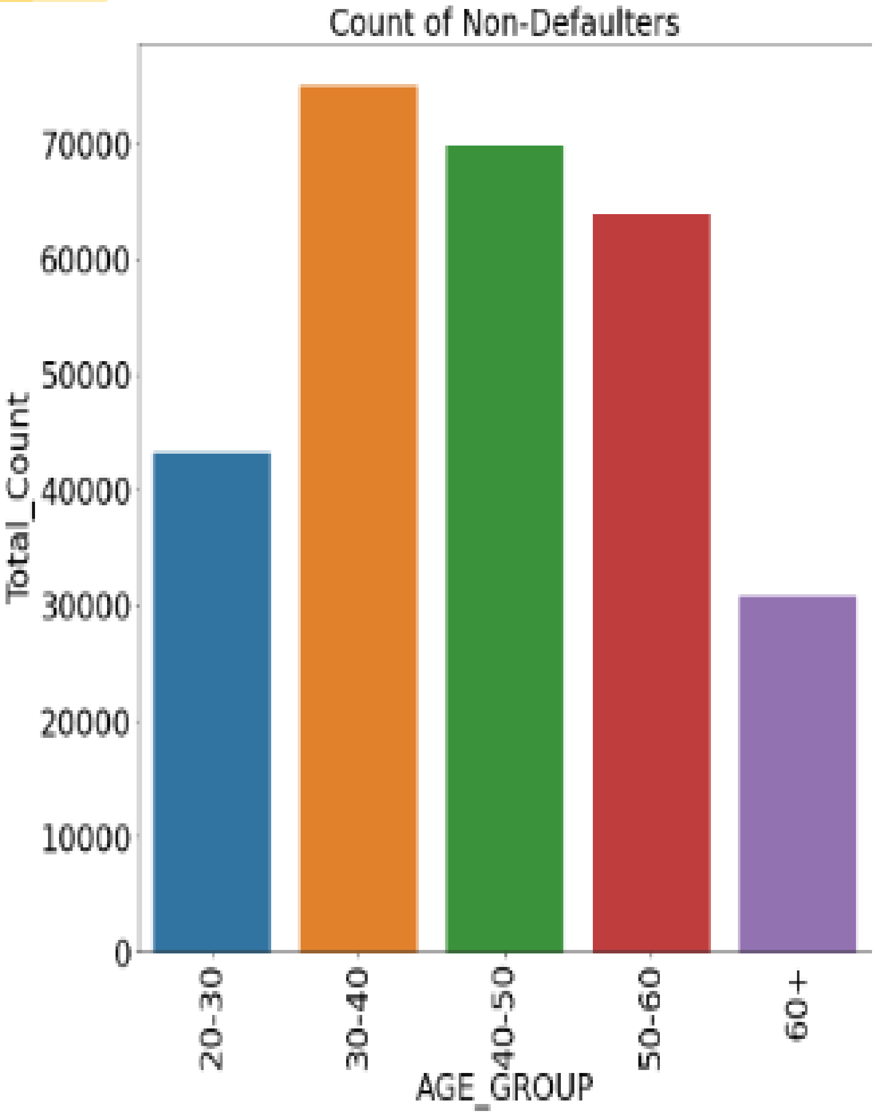


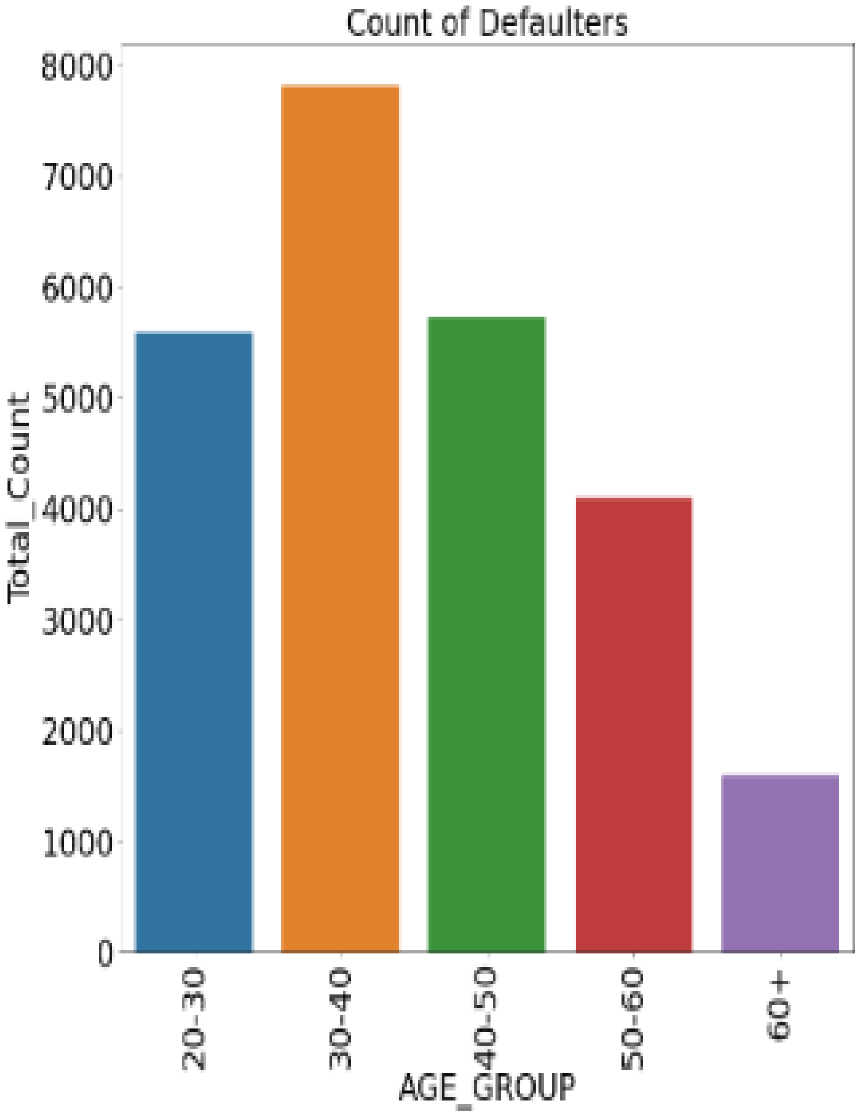
Analysis Based On Customer Education (Univariate Analysis) :



- From the plot we can say that the people whose highest education is secondary/secondary special are defaulting more.

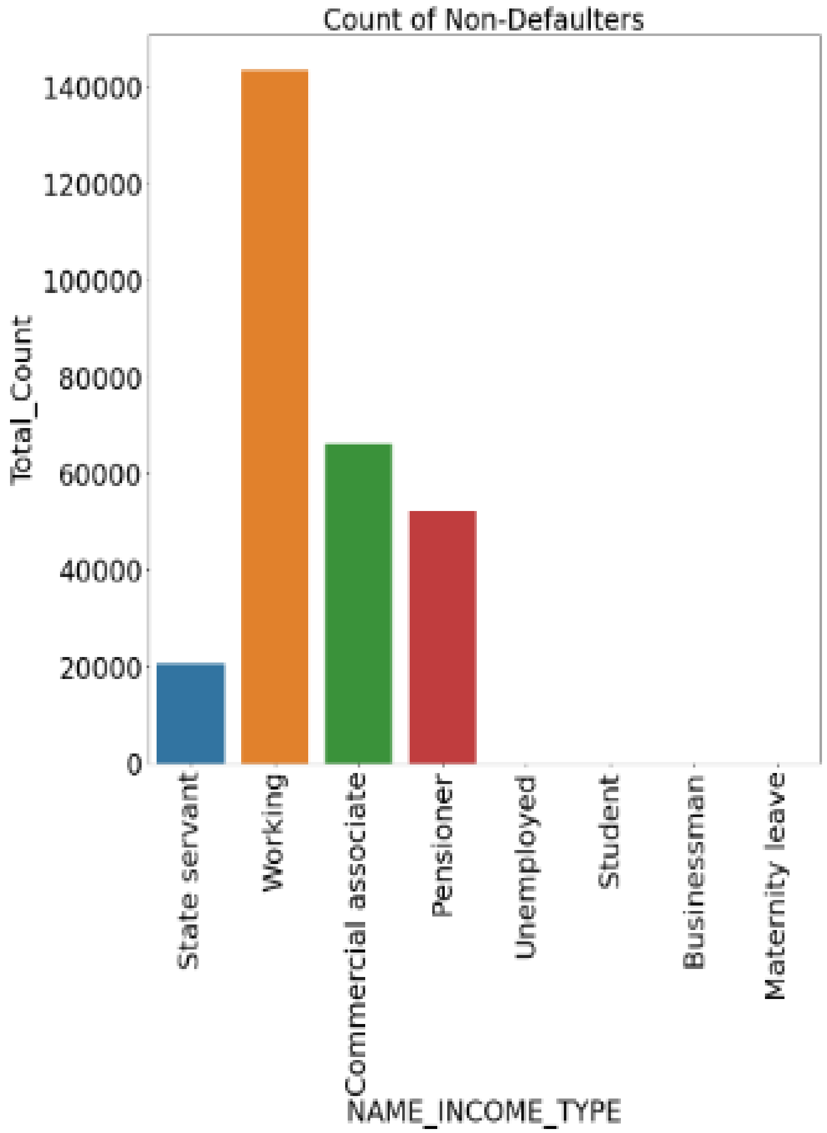
Analysis Based On Customer Age Group (Univariate Analysis):

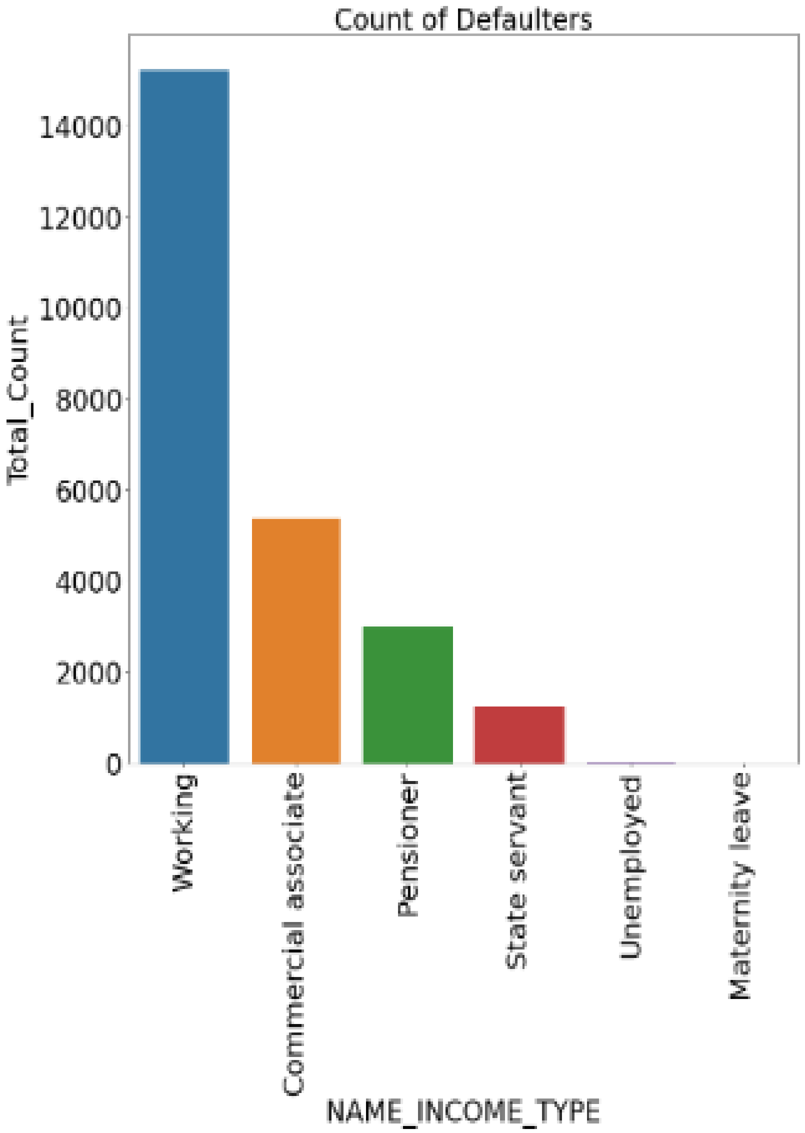




- From the plots, we can say that 30-40 age group are most likely to default.
- Clearly, we see that with increase in age group there is a decrease in defaults

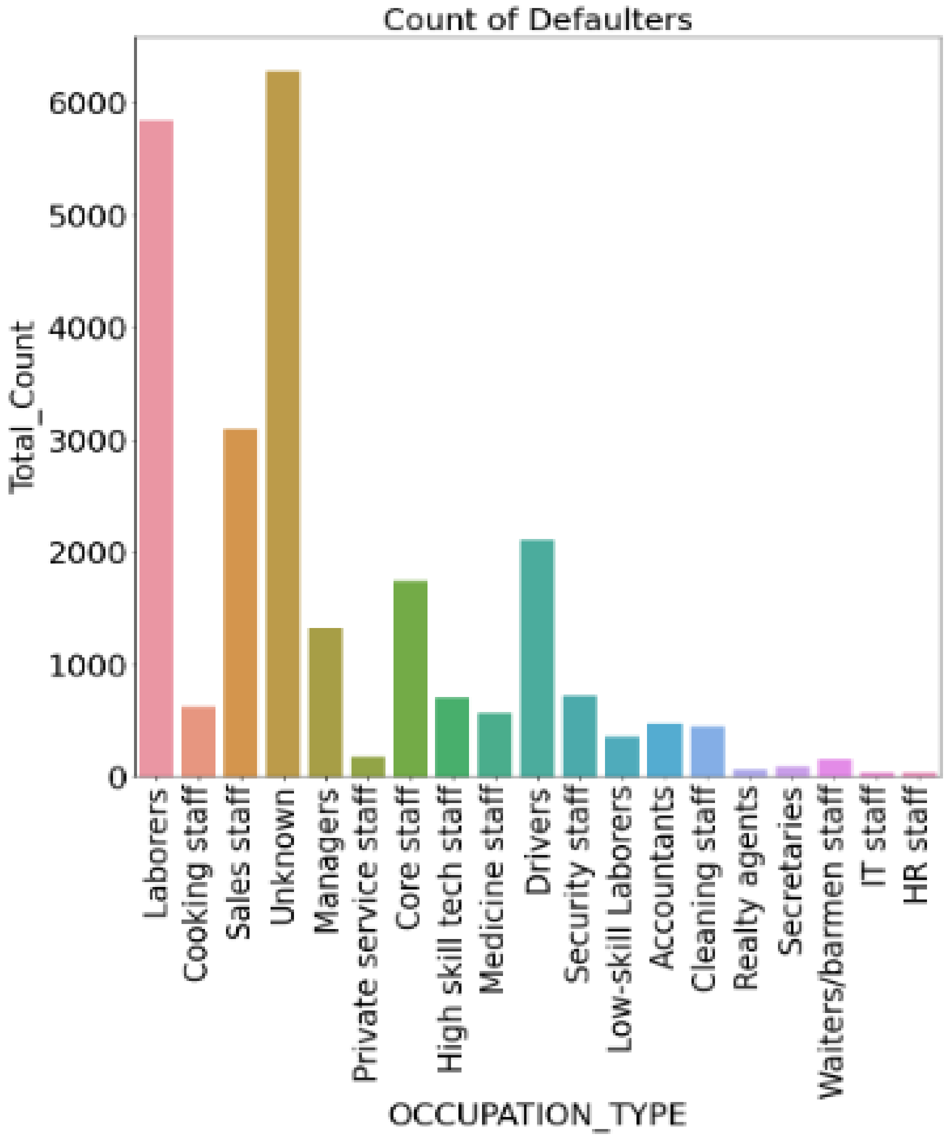
Analysis Based On Income Type (Univariate Analysis) :





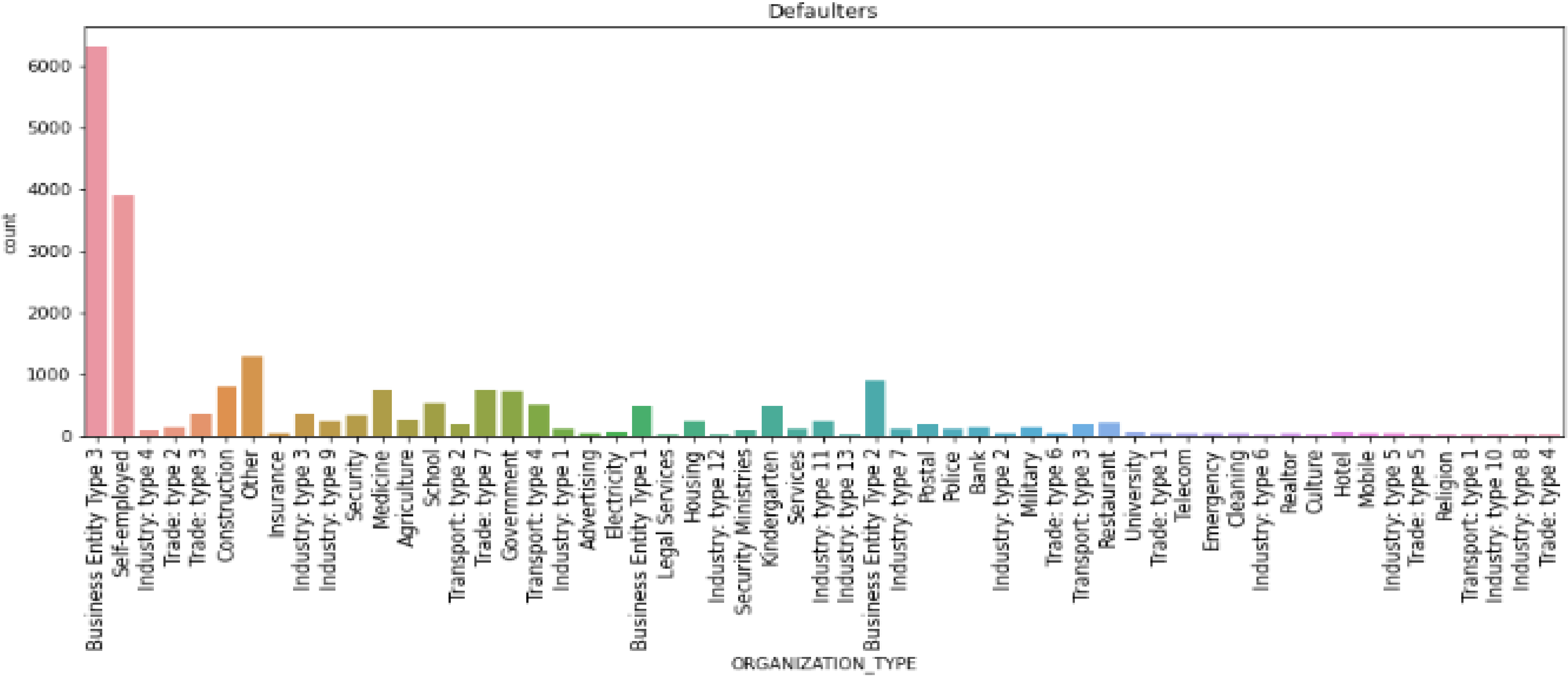
- We can see from plots, that students never default. This could be because banks give less loans to the students
- Clearly we can see that bank provide more loans to working class people
- Also the working class people are more likely default

Analysis Based On Occupation Type (Univariate Analysis) :



- We can see that laborers are more likely to default followed by Sales staff and Drivers
- The Unknown column is imputed in missing values. Hence not considering that category

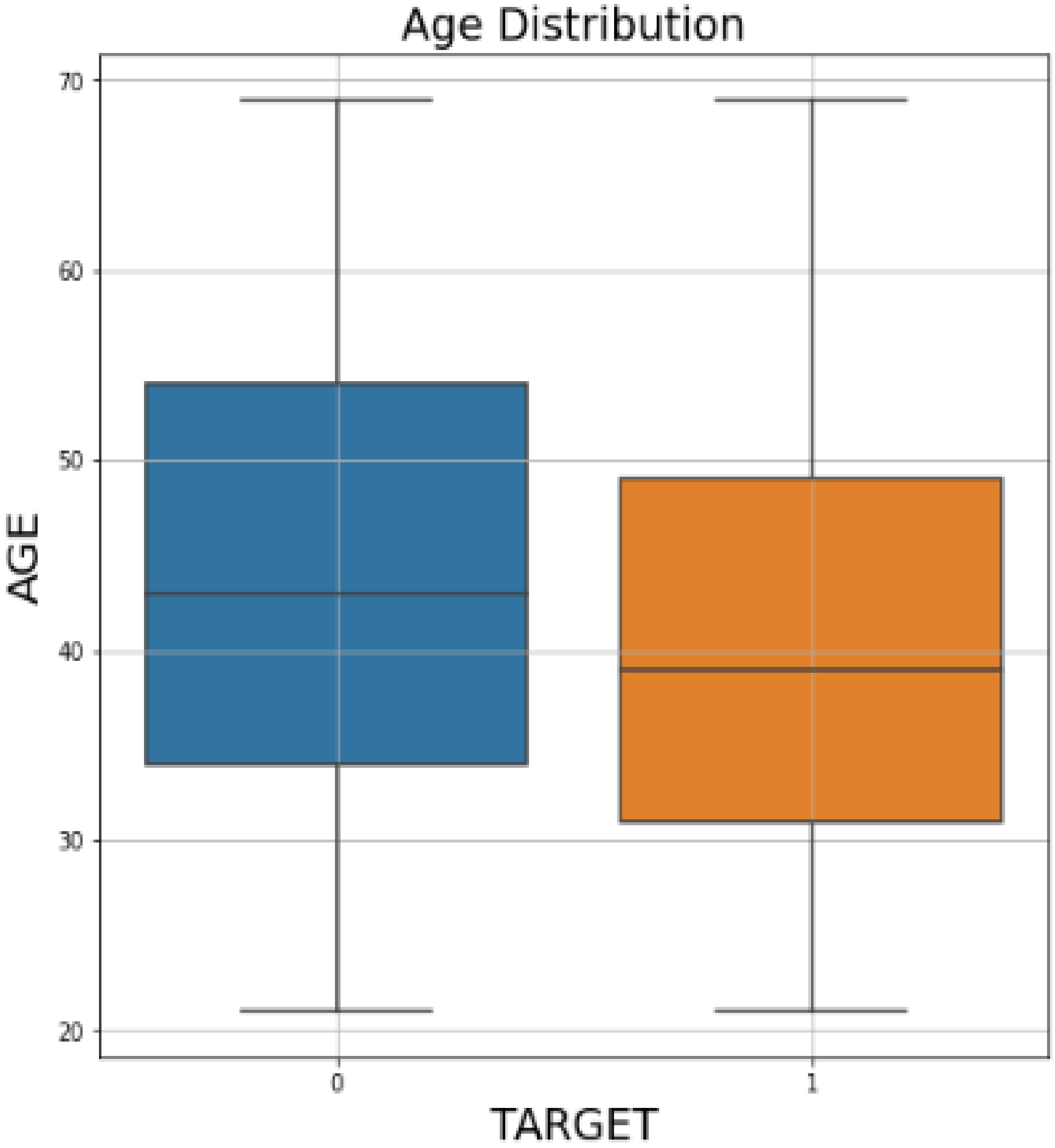
Analysis Based On Organization Type :



- *Business entity type-3 and Self employed people are more likely default the loans*

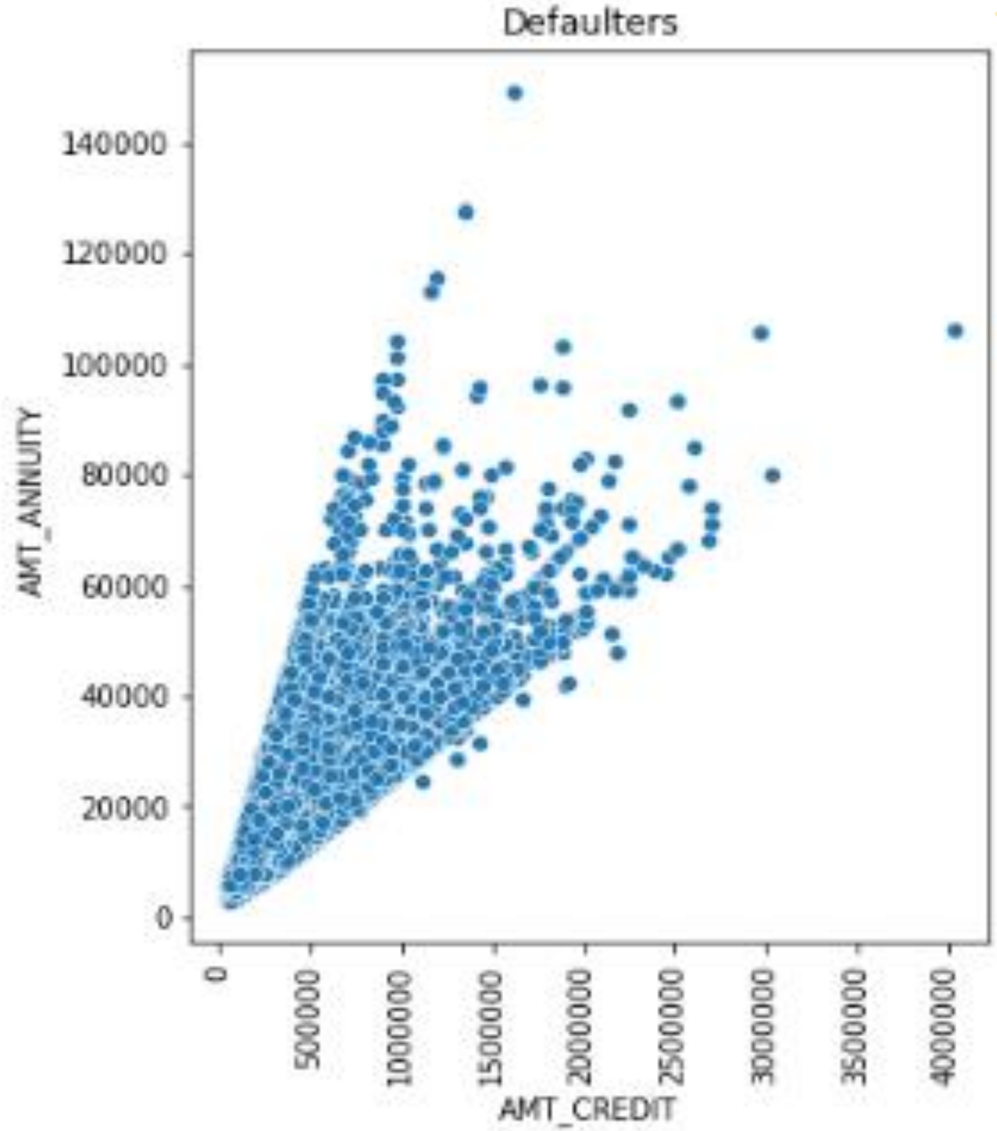
Age Distribution Defaulters Vs Non-Defaulters (Bivariate Analysis) :

- From box plot, we can conclude that, the people in the age between 30-50 are more likely to default
- For defaulters, the median age is around 40
- For non-defaulters, the median age is around 45



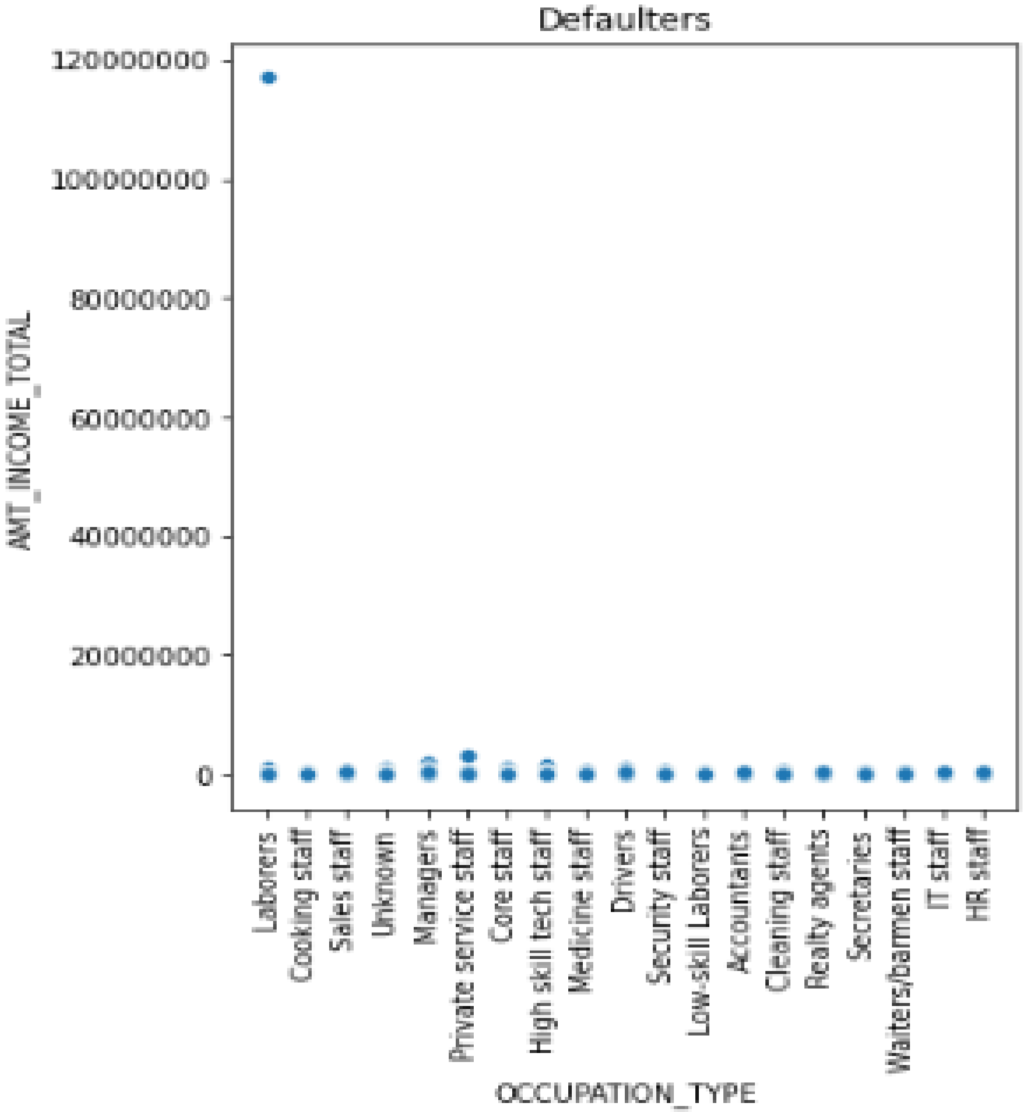
Credit Amount Of the Loan Vs EMI of the Loan(Bivariate Analysis):

- With increase in credit amount the EMI is also increasing which is true as our EMI depends upon the amount of credit



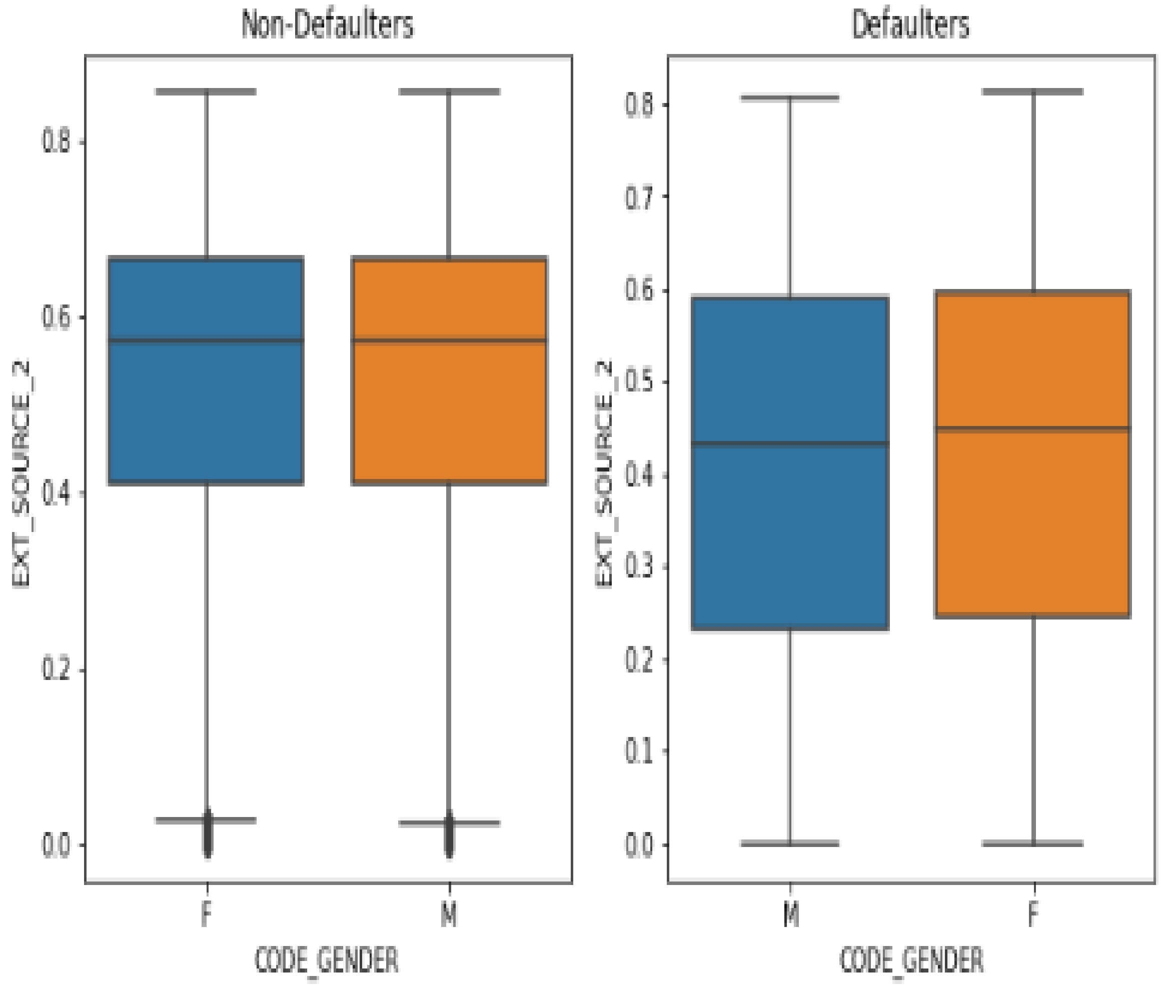
Occupation of the customer Vs Income of the Customer (Bivariate Analysis):

- We can clearly see from plot, the income of defaulters are very low for all occupation_types



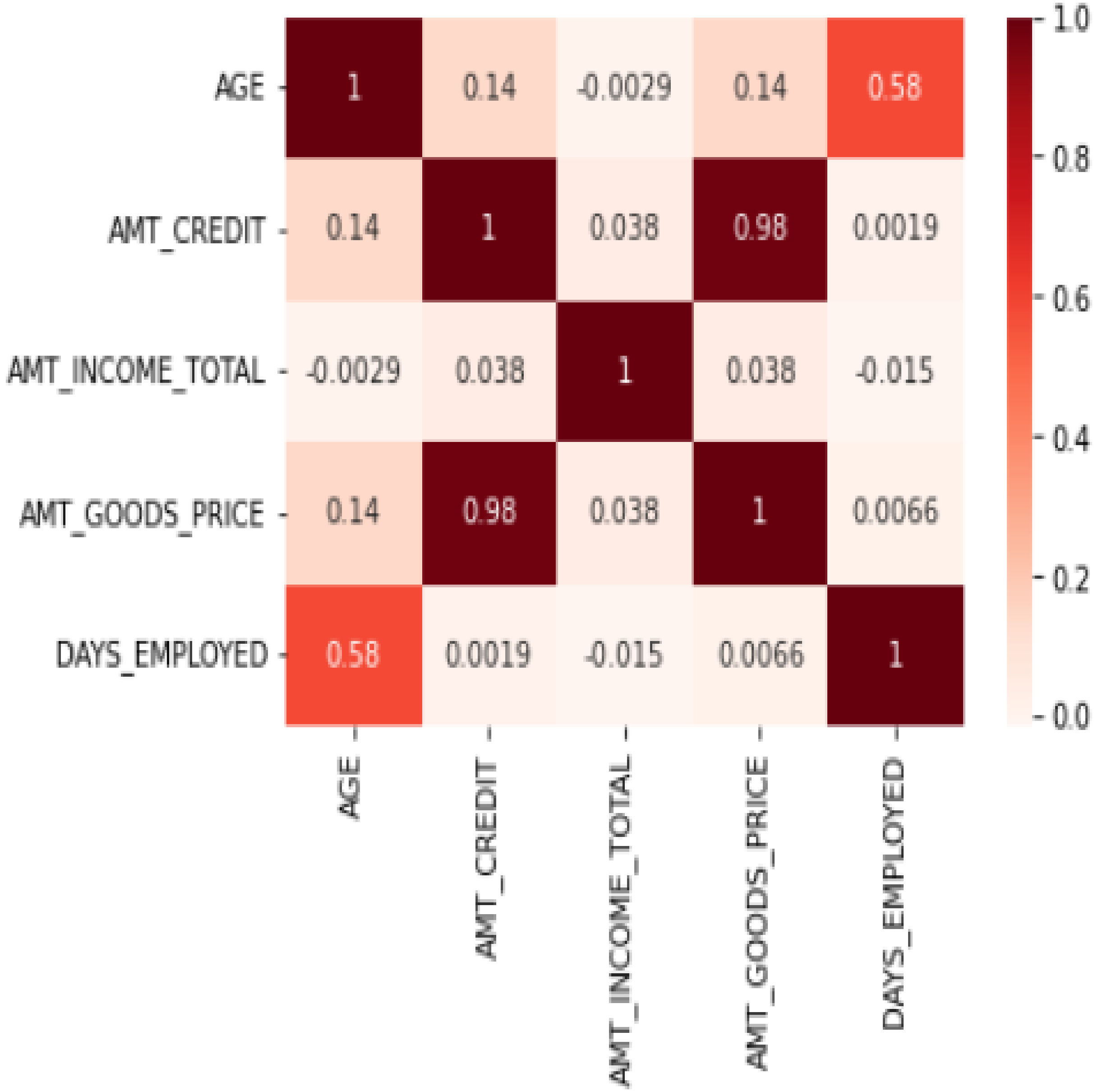
Target Column Vs Credit Score(Bivariate Analysis)

- From the box plot we can say that, the median credit score of both male and female are low for defaulters compared to non-defaulters



Correlation Matrix

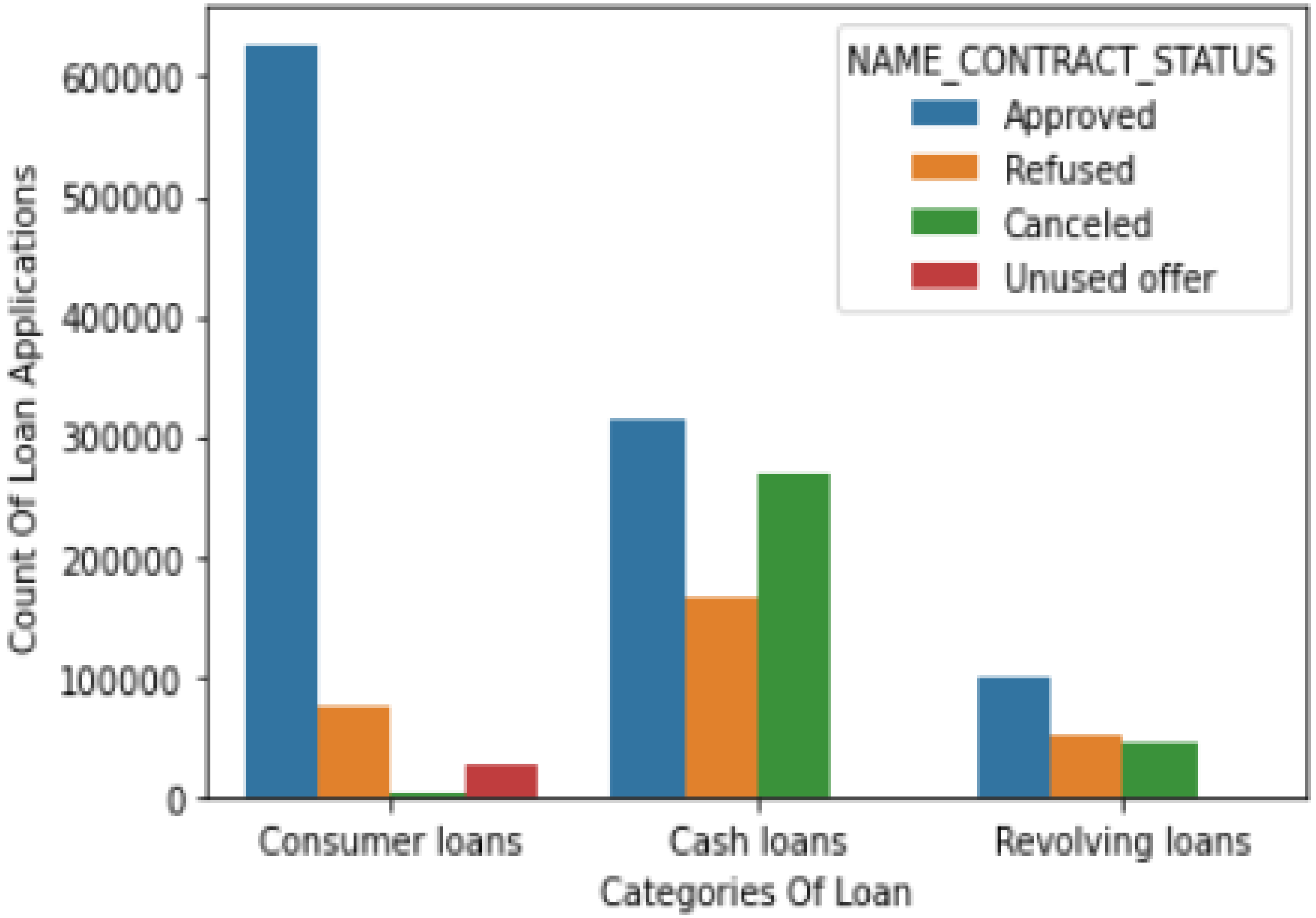
- From heatmap, we can say that price of the goods and credit amount are highly correlated.



	AGE	AMT_CREDIT	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	DAYS_EMPLOYED
AGE	1	0.14	-0.0029	0.14	0.58
AMT_CREDIT	0.14	1	0.038	0.98	0.0019
AMT_INCOME_TOTAL	-0.0029	0.038	1	0.038	-0.015
AMT_GOODS_PRICE	0.14	0.98	0.038	1	0.0066
DAYS_EMPLOYED	0.58	0.0019	-0.015	0.0066	1

Previous Loan Category vs Loan Status:

- We can clearly see that most of the applications are for consumer loans and also the approval rate is also high
- We can also observe that more cash loans are refused compared to consumer and revolving loans



Summary:

- Females are taking more loans compared to the male counterparts and the loan defaults are also high in females.
- Customers who own the car are less likely to default.
- Secondary education people are more likely to default.
- People in the age group of 30-50 are more likely to default.
- As the age increases, the chances of default is becoming less.
- State servants are less likely default.
- Laborers are more likely default followed by sales staff and drivers.
- Self employed people are more likely default.