EDA CASE STUDY

Case Study Business Problem:

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 as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Major problems for the scenario:

There are majorly two problems observed:

1) If a person is capable of paying loan and application is rejected, bank will be ending with Interest loss
2) If a person is not capable of paying loan and application is approved, bank will be ending with credit loss.

For the business problems mentioned above slide can be insighted from analyst percepective and reduce the loss by giving some recommendations for bank.

This process can be achieved by EDA (Exploratory Data Analysis).

EDA Process:

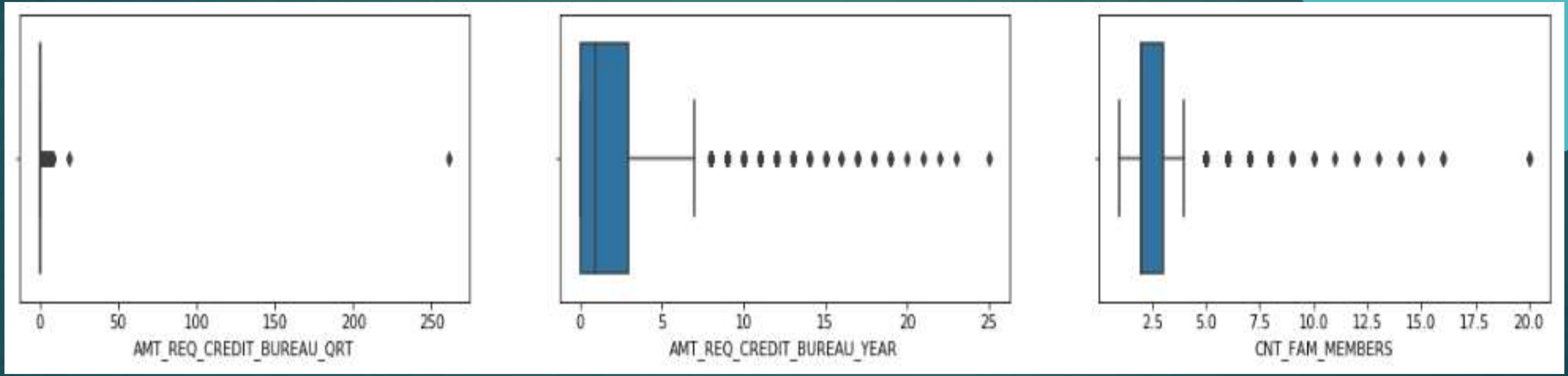
1) Data Cleaning
2) Univariate Analysis
3) Bivariate Analysis.

Data Cleaning

- ❑ Load Application CSV File.
- ❑ Look for few Insights using Python commands(Describe, Info, Shape, data types).
- ❑ Start Looking for Null values as they affect the Quality of Data Analysis on the Dataset.
- ❑ Drop Columns Which have More than 50% of null values in the respective columns as they do not help or show impact on our Analysis.
- ❑ Do not drop other columns which have Less %(Around 13) of Null values as they can be imputed based on the type of Columns Choosing Mean, Median and Mode.
- ❑ There are 41 Columns which are having more than 50% of null values which are Dropped.

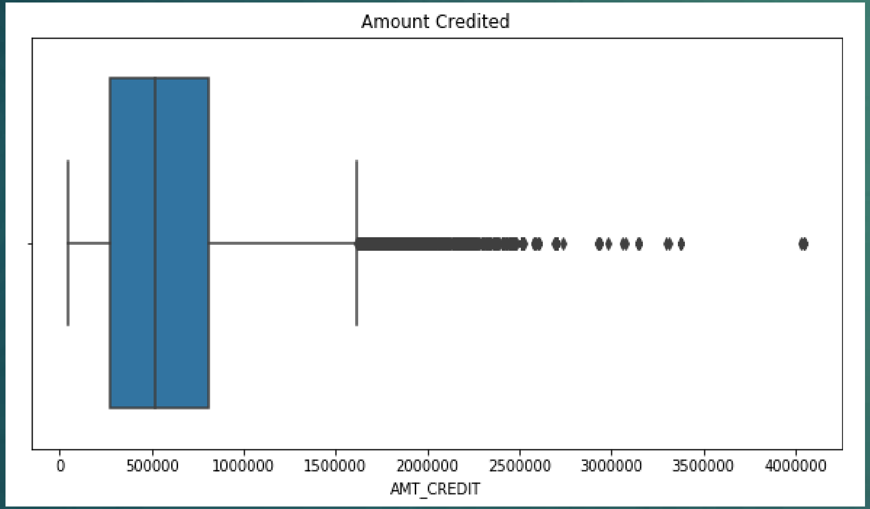
Handling Outliers for continuous Numeric variables

- **'AMT_REQ_CREDIT_BUREAU_QRT’,'AMT_REQ_CREDIT_BUREAU_YEAR','CNT_FAM_MEM BERS'** are continuous numeric variables chosen form the Application CSV file that are having outliers present in the respective columns.
- As per the Box Plot below ‘MEAN’ can have a impact on the actual data if imputed. Hence it is recommended to use MEDIAN for these columns.

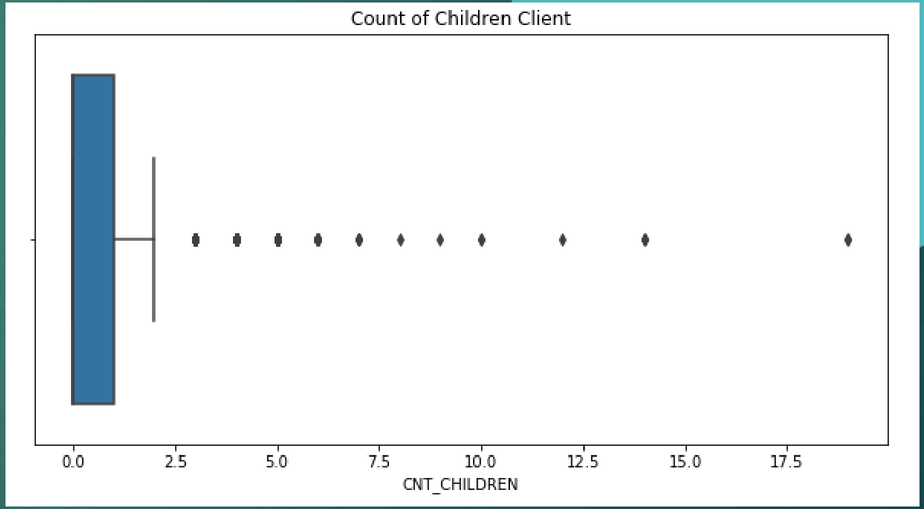


Identifying Outliers for continuous Numeric variables

- For ‘AMT_CREDIT’ column the mean is 599076.2 where as the max amount Is 4050000.



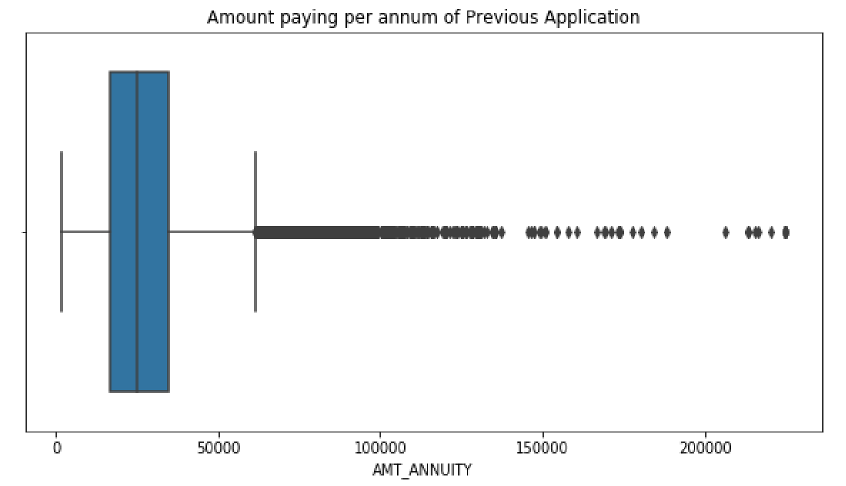
- For ‘CNT_CHILDREN’ column the mean is 0.41where as the max amount Is 19.

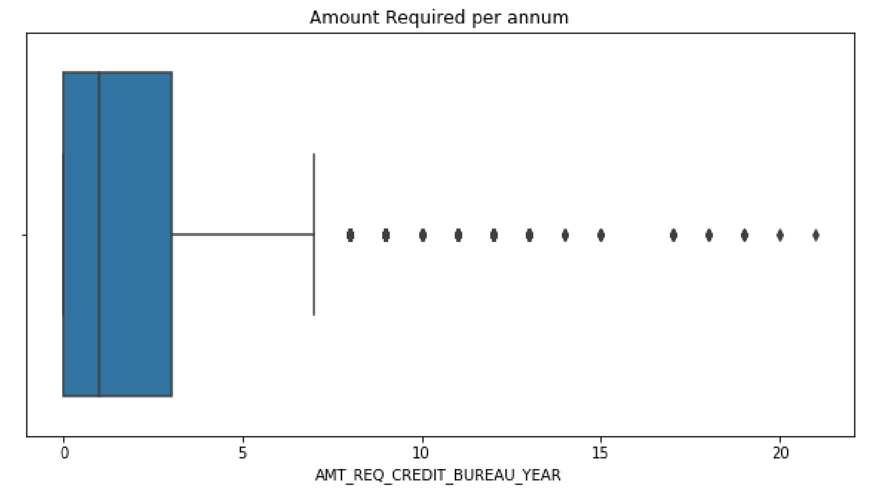


Identifying Outliers for continuous Numeric variables

- For 'AMT_ANNUITY' column the mean is 27117 where as the max amount Is 225000.

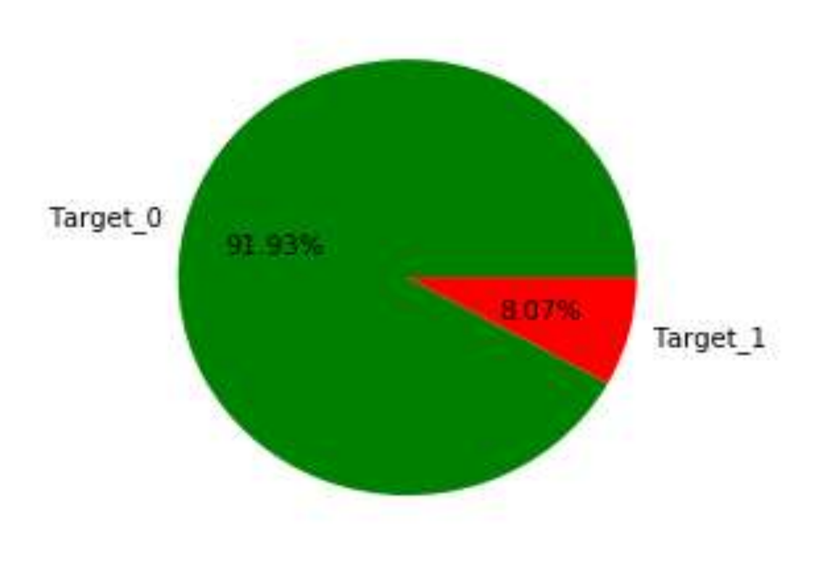
- For 'AMT_REQ_CREDIT_BUREAU_YEAR' column the mean is 1.9where as the max amount Is 21.



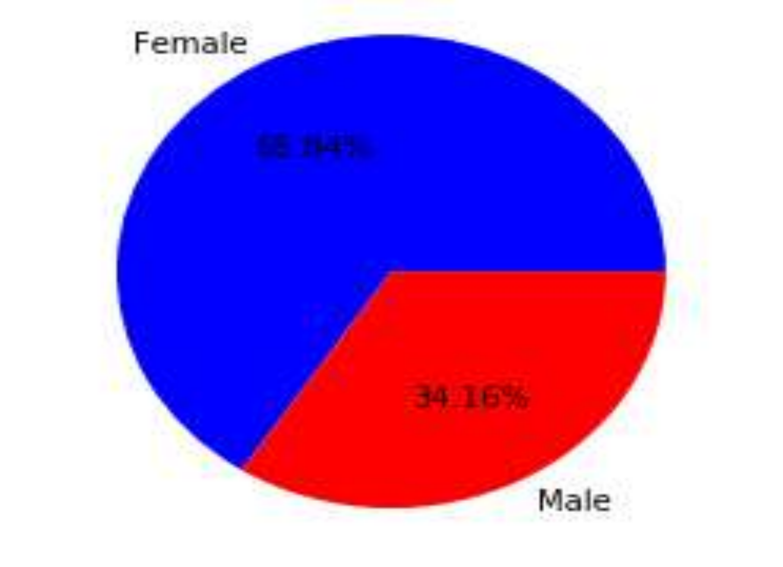


Imbalance Check:

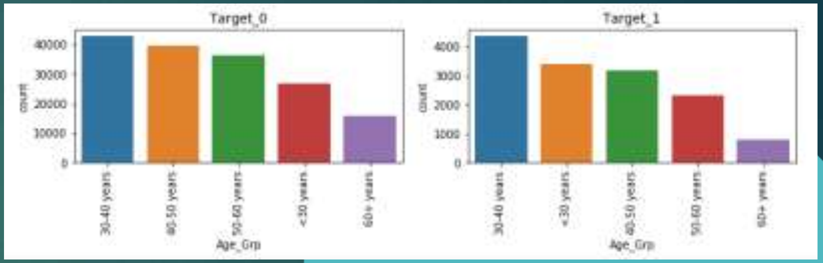
- Based On Target Value 0 and 1
- As you can see in the graph below the data is Imbalance where Target 0 is of 91.93% and Target 1 is of 8.07%



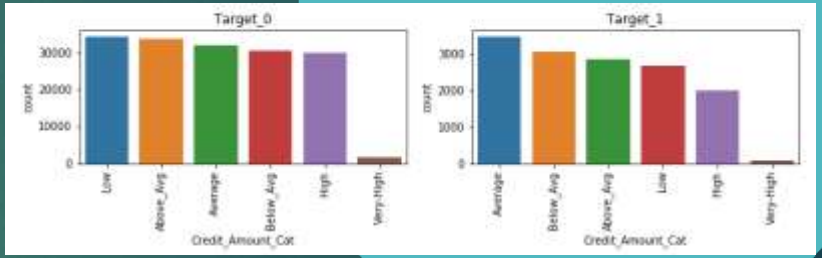
- Based on Male & Female Gender
- As you can see in the graph below the data is Imbalance where Female are of 65.84% and Male are of 34.16%



Univariate Analysis For Categorical Variables of Target 0 and Target 1

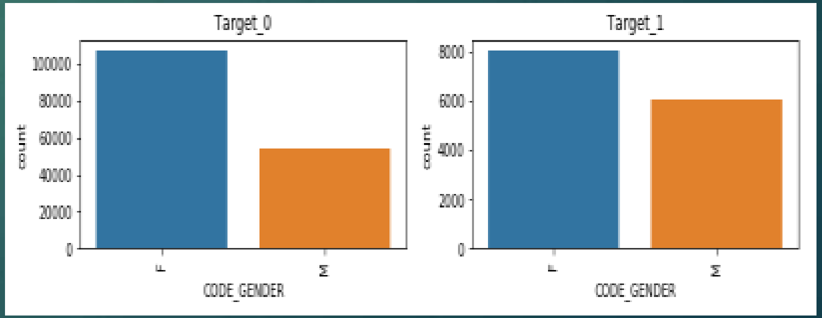


❑ Considering Age group Variable in target 0 and target 1 the chances of a non-defaulter and likely to default for age group of 30-40 is high in both the cases.



❑ For the Credit amount category the low zone is having higher re-payments in target 0 and in target 1 the difficulties is for average credit amount taken holders.

❑ No Significant difference in Housing Type, Family Status, Education Type.



❑ Also for Code Gender category around 125000 Females are more in Number as non-Defaulters than being likely to default and we can see the male category is high in likely to be default than Non-Defaulters.

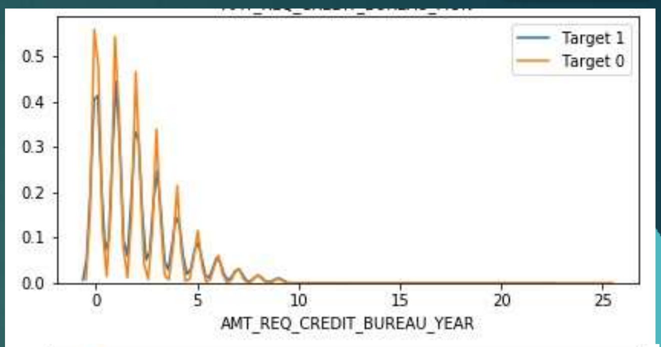
Note: Only limited graphs are shown in the presentation. More illustrations and graph can be found in attached python file

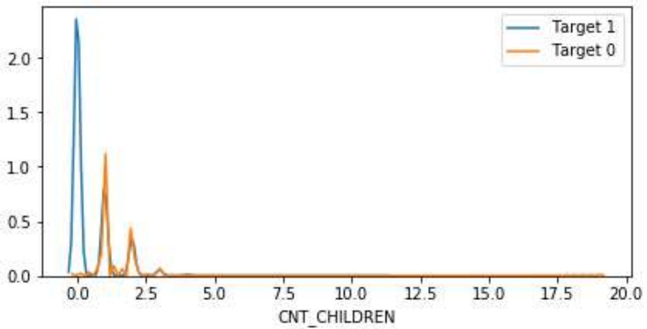
Correlation of numerical Columns of Target ‘0’ and Target ‘1’

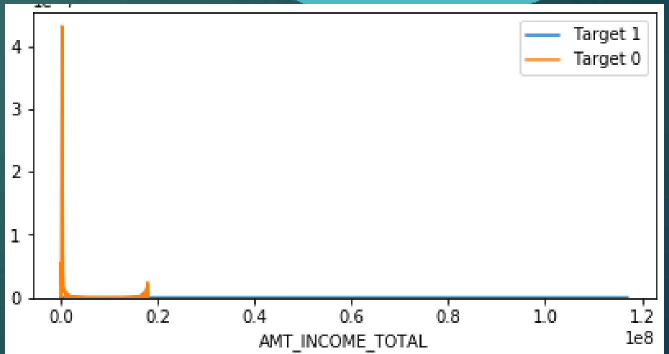
- ❑ Highest correlation exist between AMT_CREDIT and AMT_GOODS_PRICE there is something related to this column of Target ‘0’
- ❑ There is negative correlation existing between CNT_CHILDREN and DAYS_BIRTH in Target ‘0’
- ❑ Least correlation exist between the DAYS_EMPLOYED and DAYS_ID_PUBLISH in Target ‘0’
- ❑ Highest correlation exist between AMT_CREDIT and AMT_GOODS_PRICE there is something related to this column of Target ‘1’
- ❑ There is negative correlation existing between CNT_CHILDREN and DAYS_BIRTH of Target ‘1’ is Comparatively less with 'Target 0'
- ❑ Least correlation exist between the DAYS_EMPLOYED and DAYS_ID_PUBLISH of Target ‘1’
- ❑ For the above correlation we could see almost the values which are correlated with one another is same as in target 0 and 1 data frames. Also by above imputation we could AMT_GOODS_PRICE is highly correlated with AMT_CREDIT and also with reasonably less correlated with AMT_ANNUTITY

Univariate Analysis For Numerical Variables of Target 0 and Target 1

- The client reaching credit Bureau yearly enquiries are more non-defaulters than likely to default.
- The client with no children are high in number of defaulters than not Defaulters.
- If the Client's income is high then the client may not default the bank.
- The applicant For consumer loans is likely to be non-defaulter than being defaulter.
- If the clients days employed is above 350000 than it has more number of Non-Defaulters than likely to Default



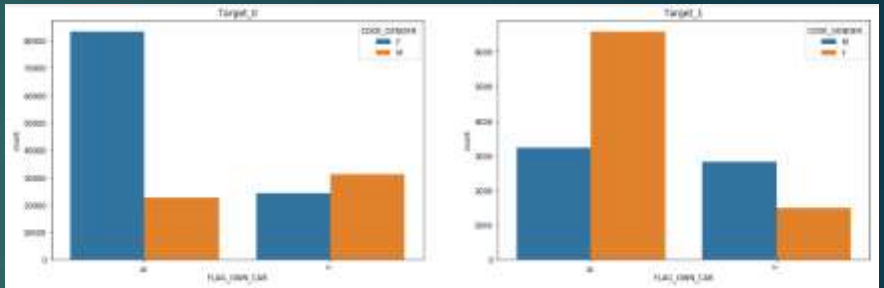


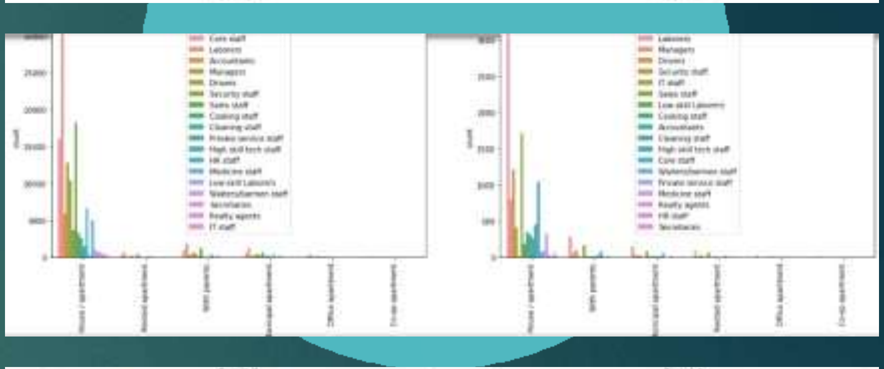


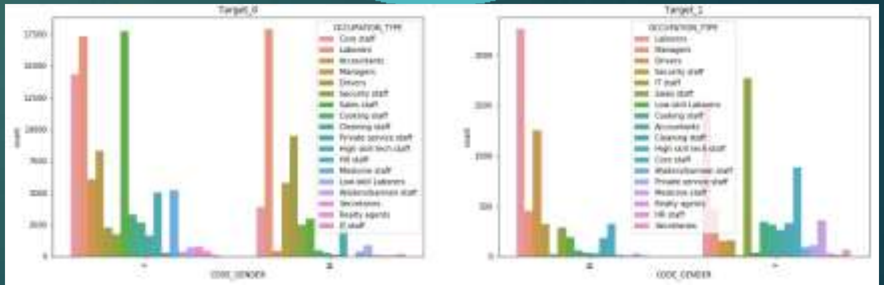
Note: Only limited graphs are shown in the presentation. More illustrations and graph can be found in attached python file

Bivariate Analysis on Categorical columns of 'Target 0' and 'Target 1'

- ❑ The female gender who doesn't own a car are more non-defaulters and its the same the same for likely to default.
- ❑ The male gender who owns a House / apartment are more likely to default than non-defaulters
- ❑ The core staff occupation type who owns of House/apartment are more in number of non-Defaulters whereas the occupation type Labourers who owns of House/apartment are high in number to likely to default
- ❑ The Male Gender Labourers doesn't have significant difference between Non-Defaulting and likely to default and the female sales staff members are more in number of Non-Defaulters than likely to default.



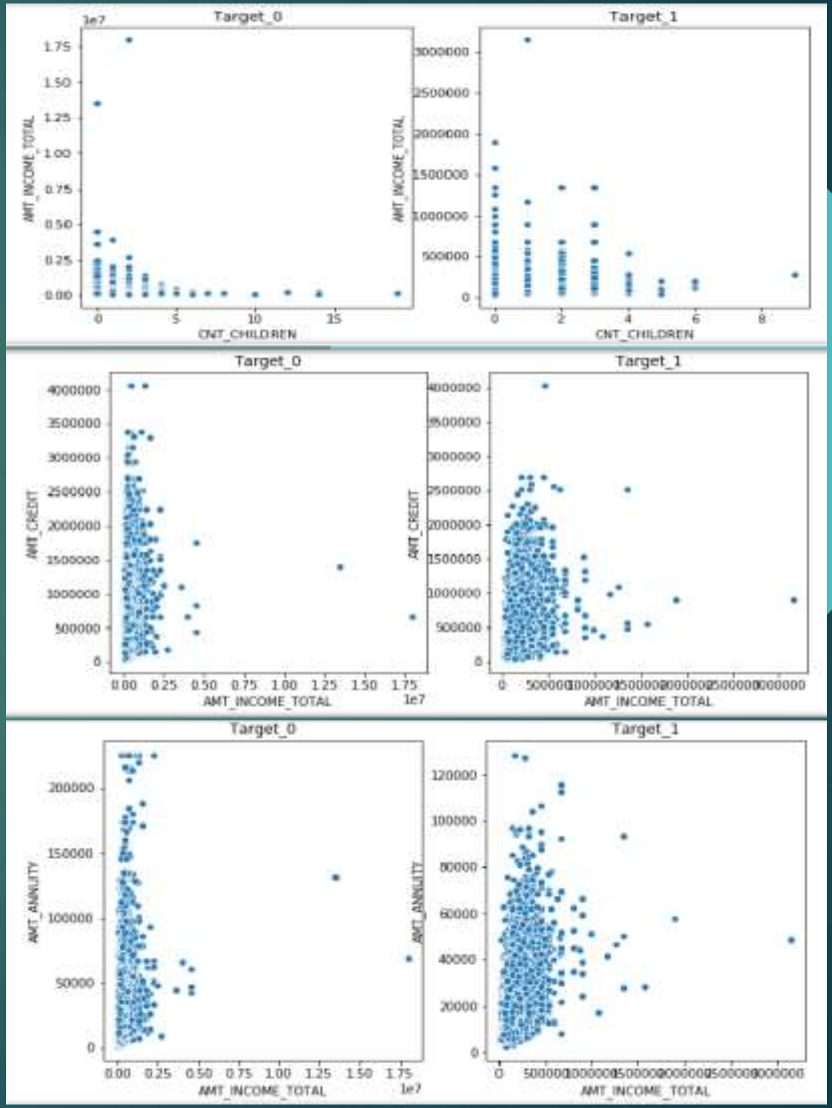




Note: Only limited graphs are shown in the presentation. More illustrations and graph can be found in attached python file

Insights Bivariate Analysis of Continuous and categorical variables and Continuous to continuous

- Applicants who have high income and with no children are more likely to default
- Providing a loan amount of Range 500000-2500000 to the total income of Less 500000 are more likely to default than non-default
- There are more people who haven't paid back their loans on time with a total income of less than 500000 and are more likely to default.
- The variables AMT_ANNUITY and AMT_CREDIT for both non-defaulters and Defaulters has a strong correlation and also has similar pattern between them



Note: Only limited graphs are shown in the presentation. More illustrations and graph can be found in attached python file

Inferences drawn performing Univariate Analysis and Bivariate analysis of Combined data frame of Categorical Variables and Continuous Variables.

- ❑ There are around 70000-80000 whose loans are approved who are likely to default and also over 200000 applicant's loan is refused who are less likely to default this would incur loss to the bank.
- ❑ There is few applicant's loan with secondary / secondary special who face difficulties to pay loan on time than who are likely to pay on time
- ❑ Female Gender are more likely to not face payment difficulties then the male and hence it is recommended to approve more loans of Female Gender than the male gender at the same Female are High in number than who face difficulties than males
- ❑ Labourers are high in number of occupation type list who are likely to default or payment difficulties
- ❑ The Repeater applicant has High chance of non-Defaulting and also has high chance of defaulting when compared to new applicants
- ❑ No millionaire is likely to default so should not refused a application of millionaire's application for loan and Lower Middle class people are high in number to repay the loans

THANK YOU

-Harshad Surya Chandolu