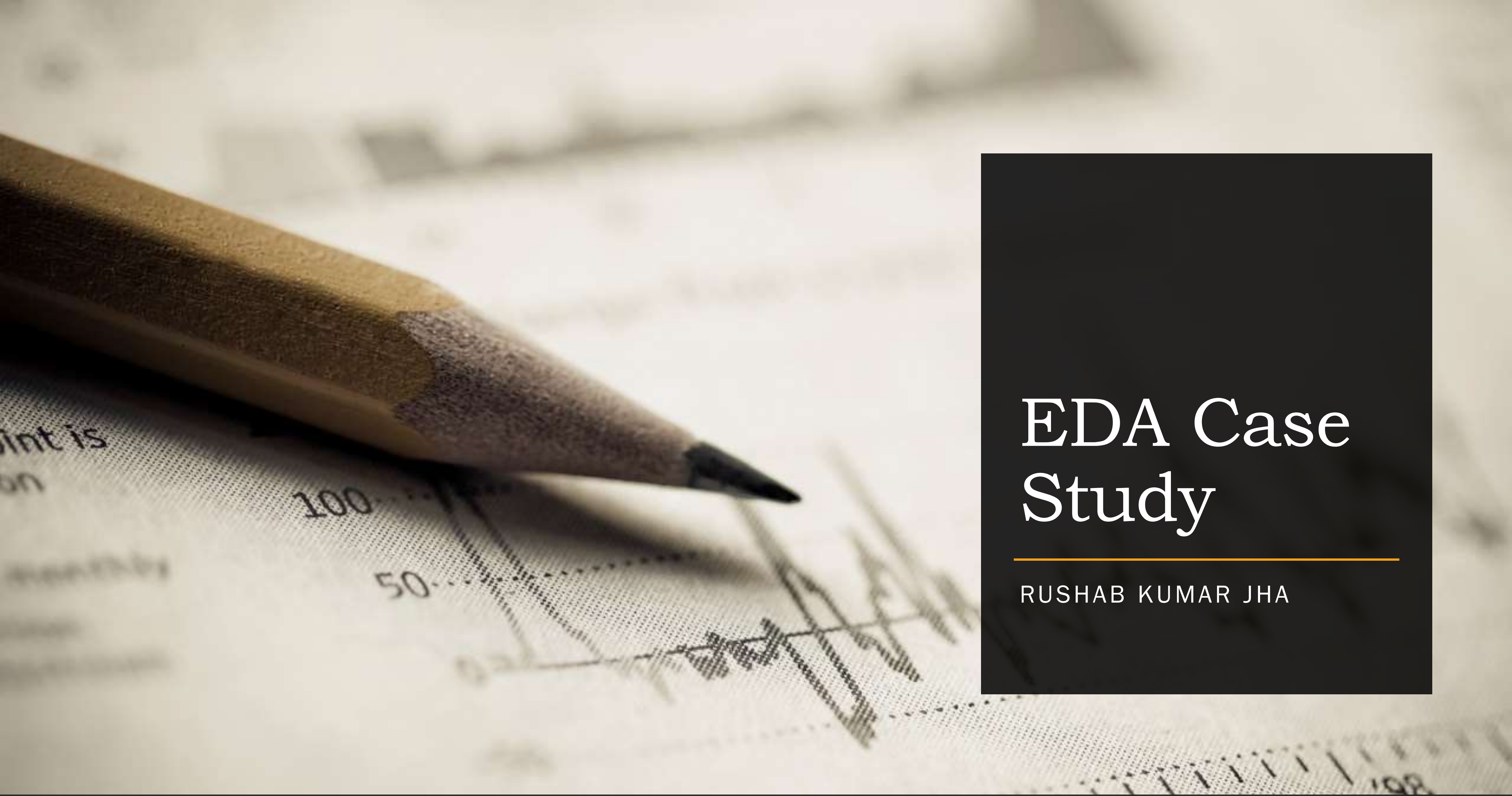


EDA Case Study

RUSHAB KUMAR JHA

Problem Statement

The Case Study aims to identify the indicating factors that results in clients having difficulty to pay for there installments which can be used for taking decisions such as denying loans, reducing the amount of loans, lending (to risky applicants) at a higher interest rate, etc. Hence ensuring that the customers capable of repaying the loans are not rejected.

Identification of such applicants using the EDA is the aim of the provided Case Study.

Steps Involved in Analysis.

Application Data

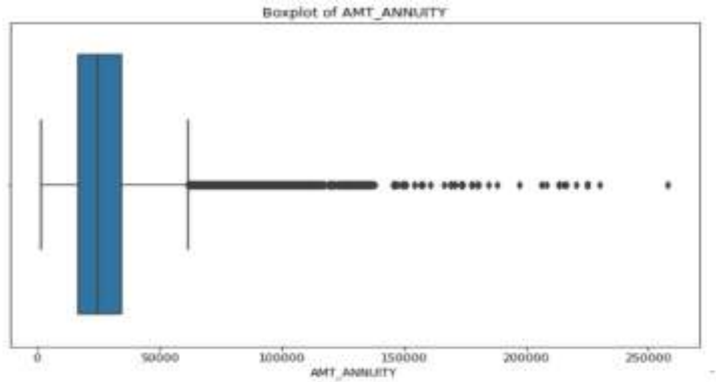
- Importing Libraries
- Inspecting the dataframe.
- Cleaning the Data.
- Analyzing the Outliers.
- Univariate Analysis.
- Univariate Analysis for Repayers and Defaulters.
- Bivariate and Multivariate Analysis.

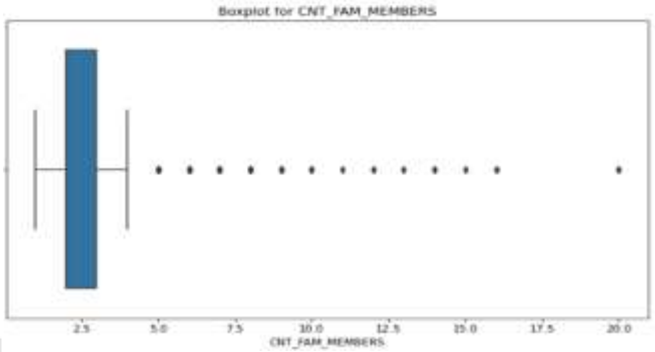
Previous Application

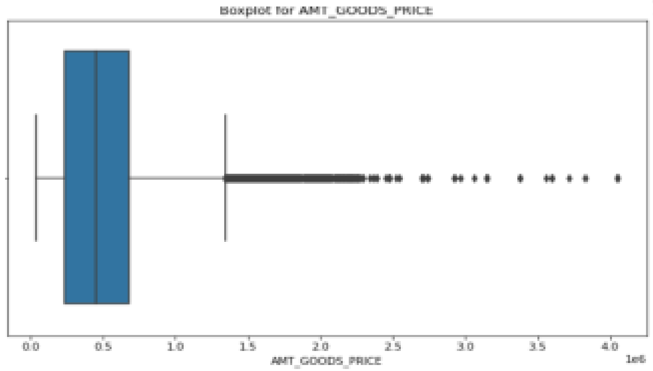
- Importing Libraries
- Inspecting the Dataframe.
- Cleaning the Data.
- Analyzing the Outliers.
- Univariate Analysis.
- Merging Application and Previous Application dataframe.
- Univariate Analysis for Repayers and Defaulters.
- Bivariate / Multivariate Analysis.

Application Data

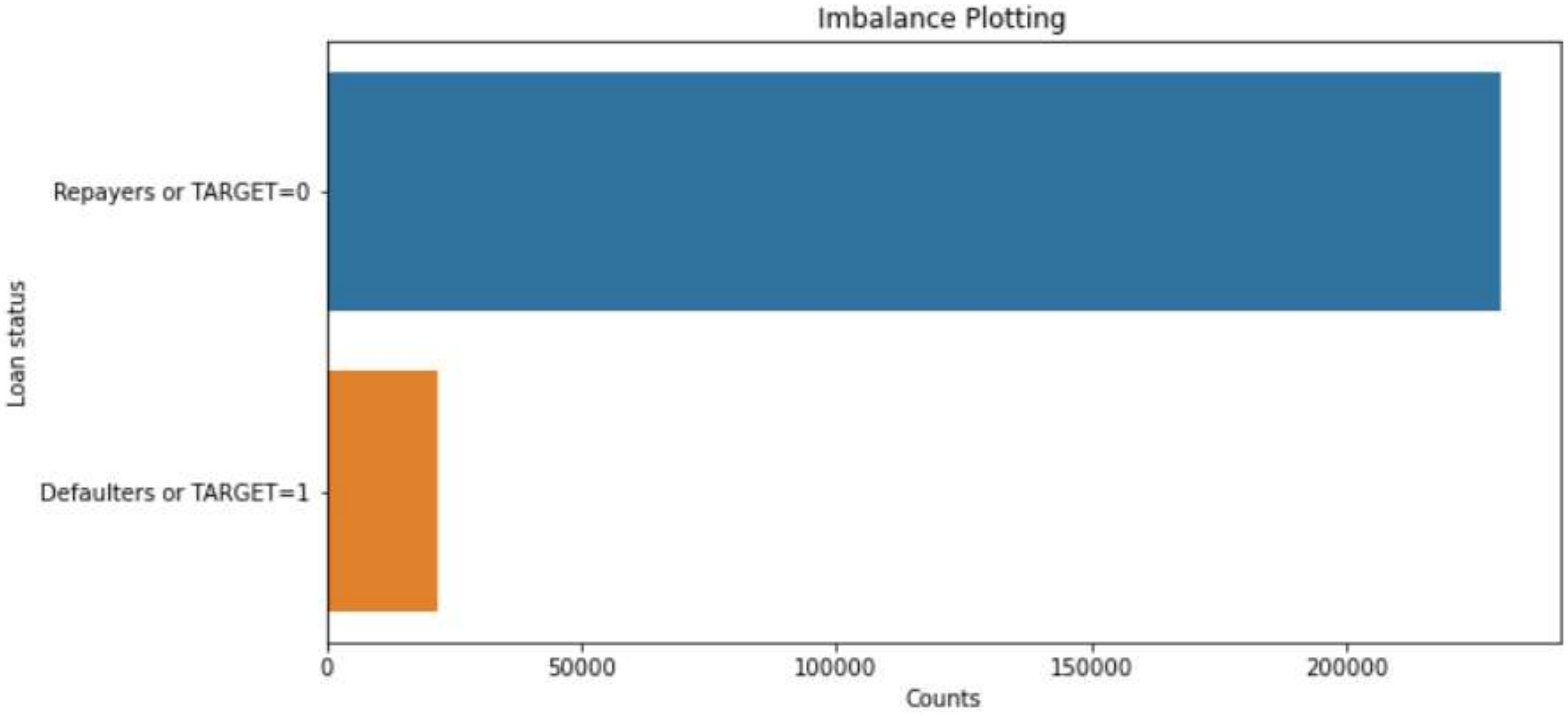
- Importing the necessary libraries.
- Then we inspect the dataframe and understand all the attributes.
- Finding the missing data from the dataframe and then find the percentage of null values,
- Then taking the data drop percentage(30%) and removing columns accordingly from the dataset.
- We then clean the data left and replace the left-over null values by mean value of the respective column as implemented on “AMT_ANNUITY”, “CNT_FAM_MEMBERS” and ‘AMT_GOODS_PRICE’ column.
- Also applied absolute and binning on certain columns like DAYS_BIRTH.







Imbalance percentage



Upon analyzing the Target=0 (Repayer) and Target=1 (Defaulter) Count we found that the imbalance ratio for clients is 10.55.

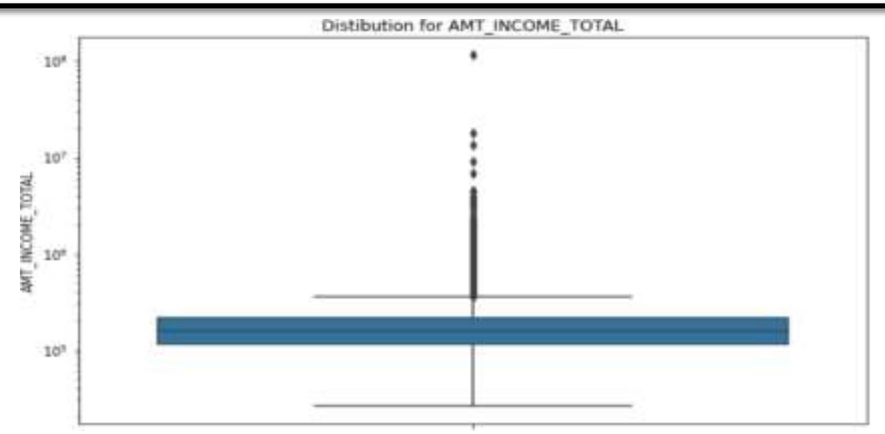
Analyzing the Outliers

- Analyzing AMT_INCOME_TOTAL we can observe that it has outliners above the 10^6 range which is almost above the upper whisker.

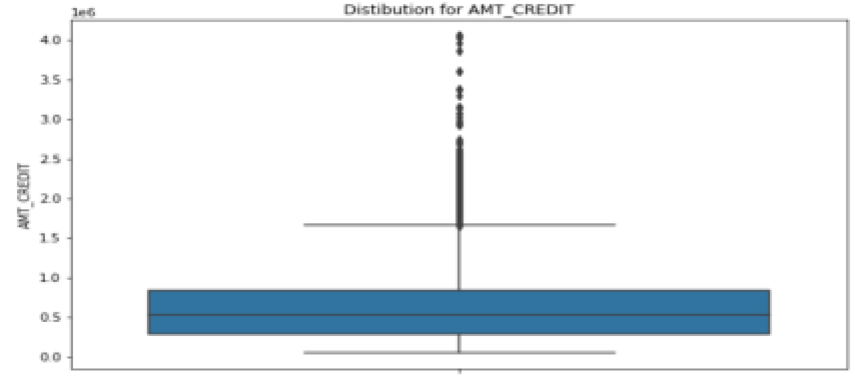
- Analyzing AMT_CREDIT we can observe that it has outliners above the 2.0x10^6 range which is almost above the upper whisker.

- Analyzing AMT_ANNUITY we can observe that it has outliners above the 100000 range which is almost above the upper whisker.

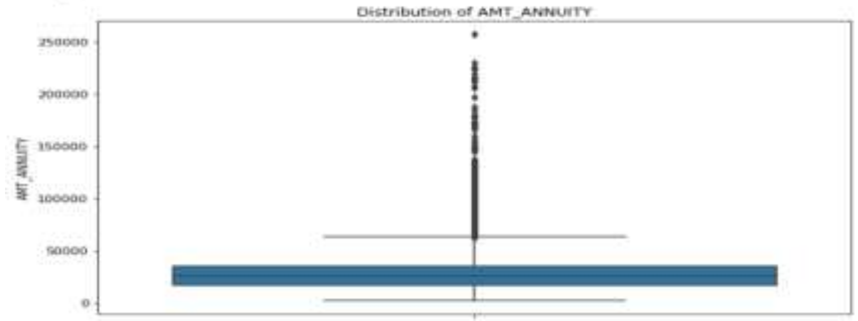
* In the DAYS_REGISTRATION we can see outliner yet in the DAYS_BIRTH analysis we observed that there are no outliners because we had applied absolute function on to it.(kindly refer the code for more details.)



Distibution for AMT_INCOME_TOTAL

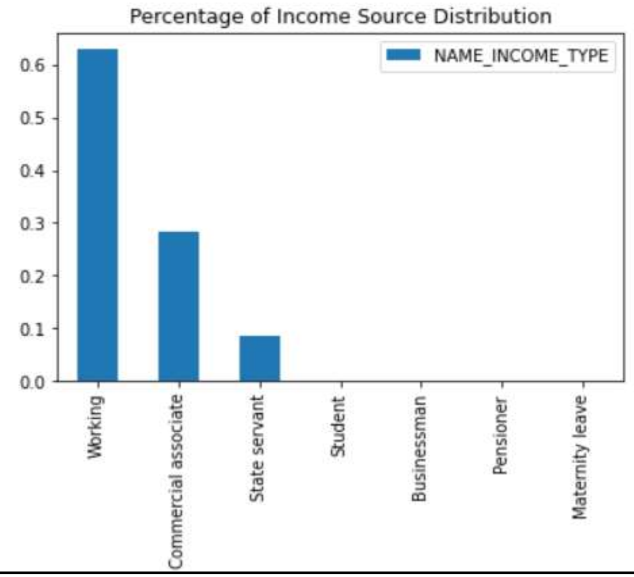


Distibution for AMT_CREDIT

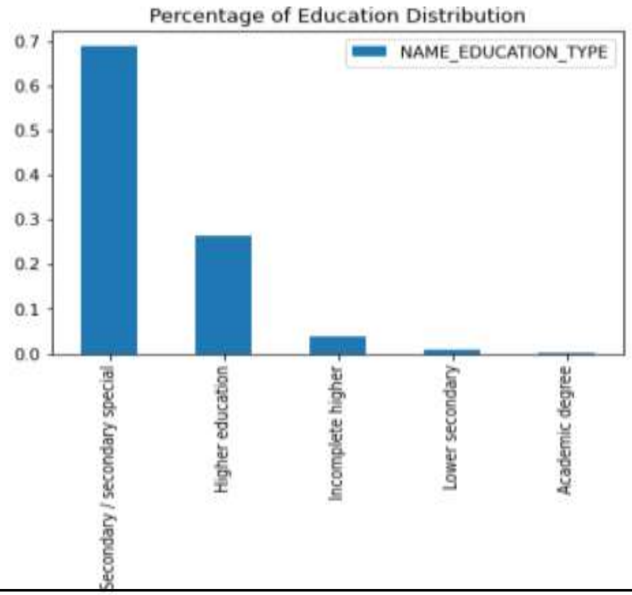


Distribution of AMT_ANNUITY

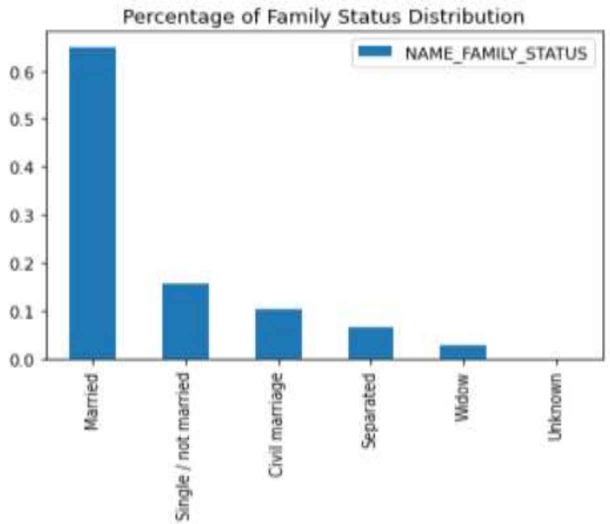
Univariate Analysis



- In the above analysis of Income Type we can see that Working category has the highest percentage of Loan Applicants



- In the above analysis of Education Distribution, we can see that Secondary/ secondary special have the highest percentage of Loan applicants.

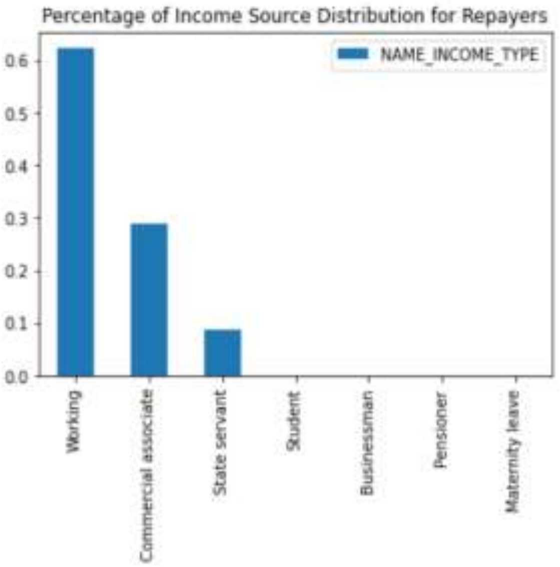


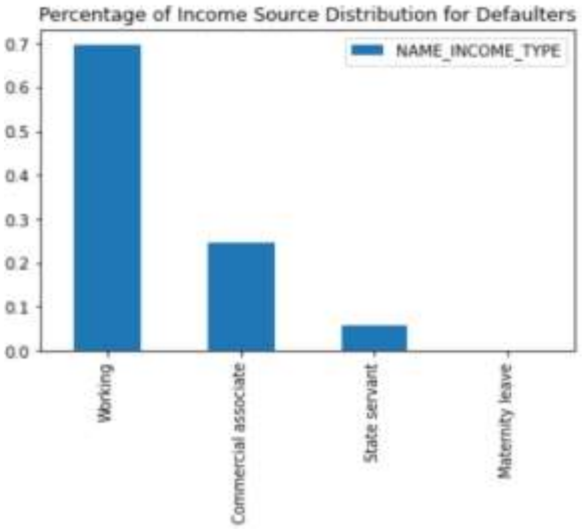
- In the above analysis of Family Type, we can see that Married clients have the highest percentage of Loan applicants.

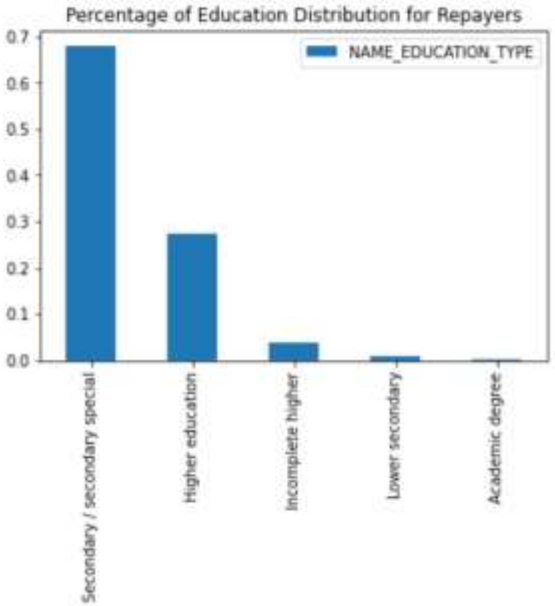
Note:-

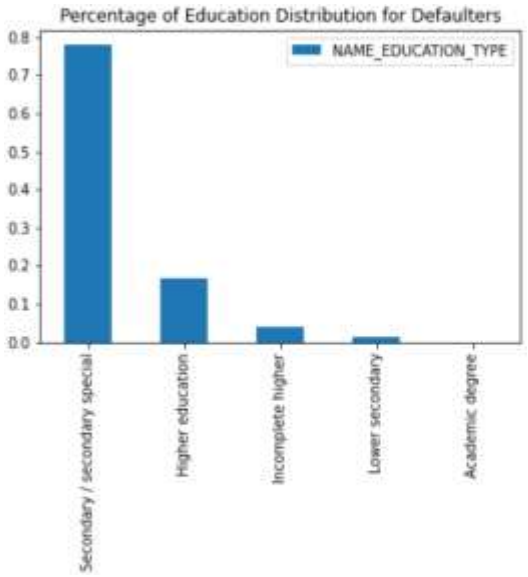
* Similar to the above observation’s, univariate analysis on some other attributes like Gender and Housing Type is also done which provides interesting view on the clients who are applying for loans.

Univariate Analysis for Repayer and Defaulter.


Percentage of Income Source Distribution for Repayers


Percentage of Income Source Distribution for Defaulters


Percentage of Education Distribution for Repayers


Percentage of Education Distribution for Defaulters

Income Type

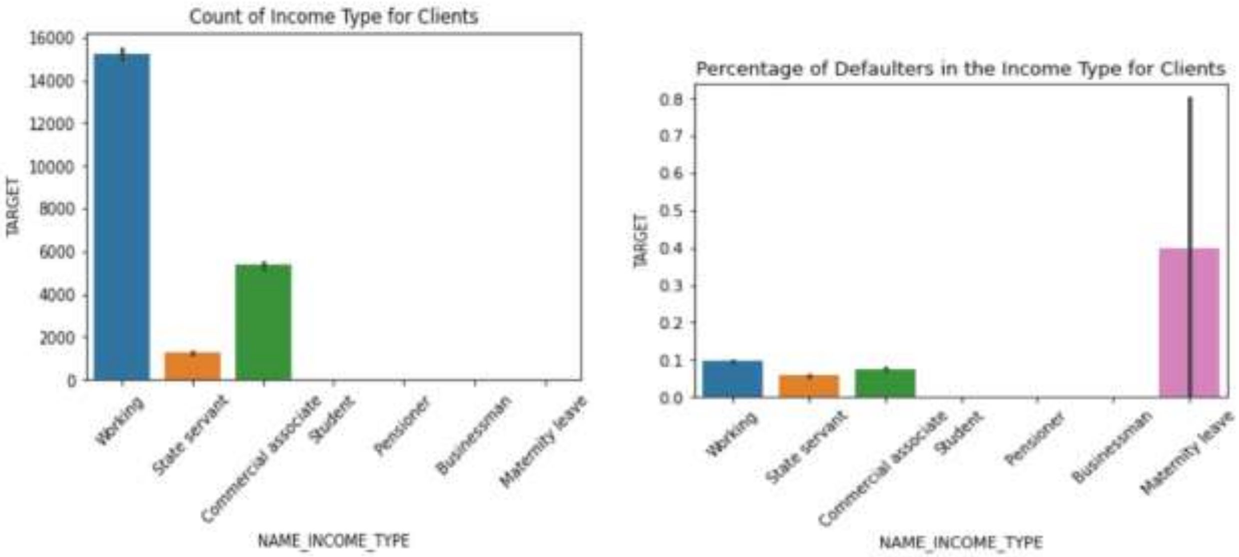
- Through the above analysis in NAME_INCOME_TYPE for Repayer and Defaulter, we can say that there is an increase in the Loan Applicants from Working Type.

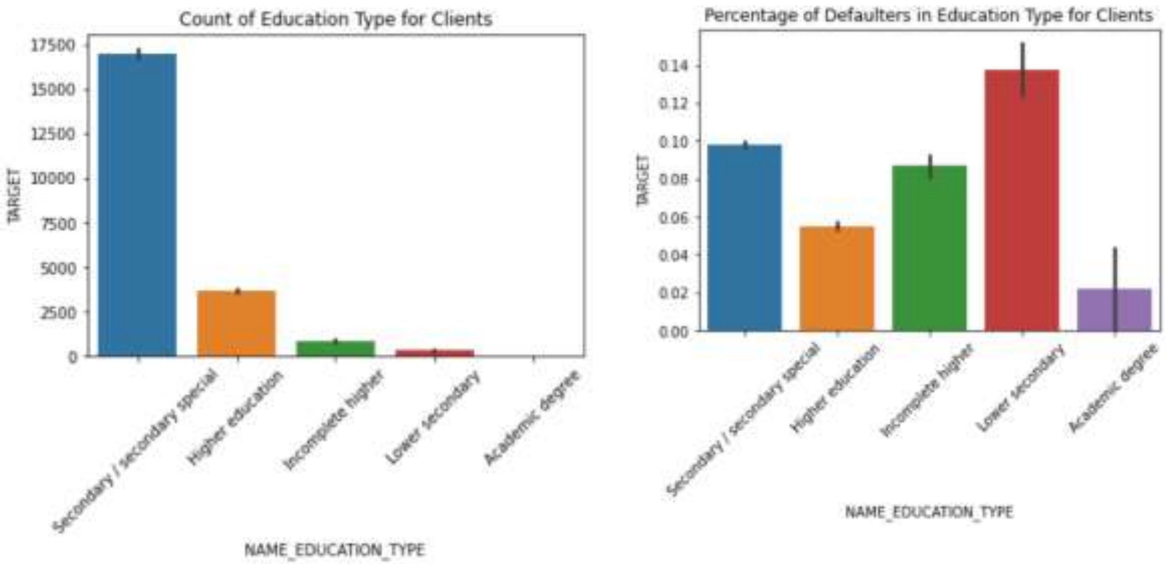
Education Status

- On analyzing the data, we find that there is an increase in percentage of Loan Payment Difficulties whose educational qualifications are secondary/secondary special.
- We also find an interesting observation of a decrease in the percentage of Loan Payment Difficulties who have completed higher education when compared with the percentages of Loan Payment Difficulties and Loan Non-Payment Difficulties

* In Loan Type there is an increase in an increase in the cash loan percentage for both Repayer and Defaulter (To know more about the other analysis kindly go through the code).

Bivariate Analysis





Income Type

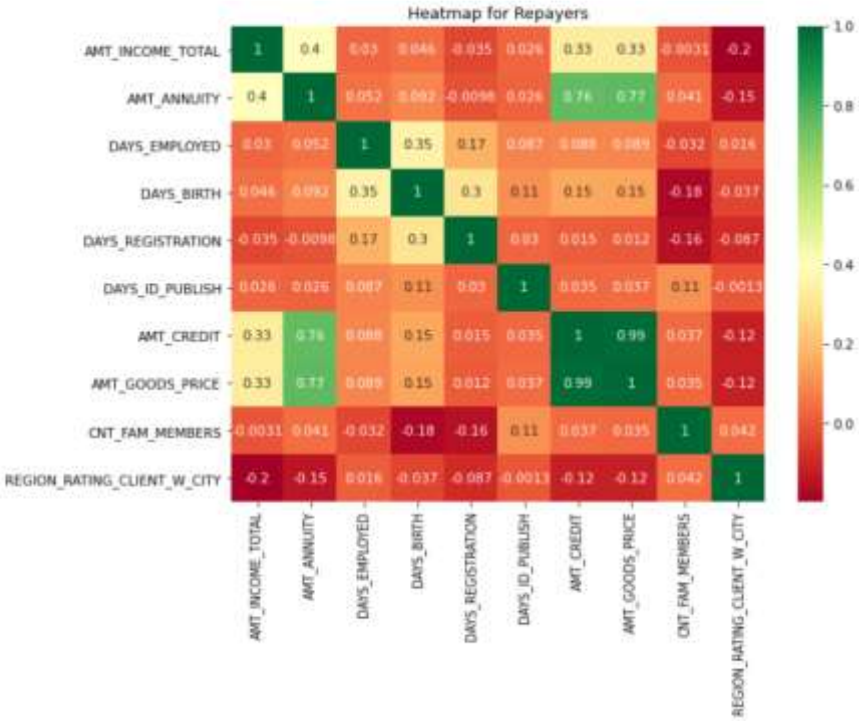
- On observing the above bar plot we can see that there is a huge increase in the Defaulter percentage in the clients with "Maternity Leave" Income Type.
-

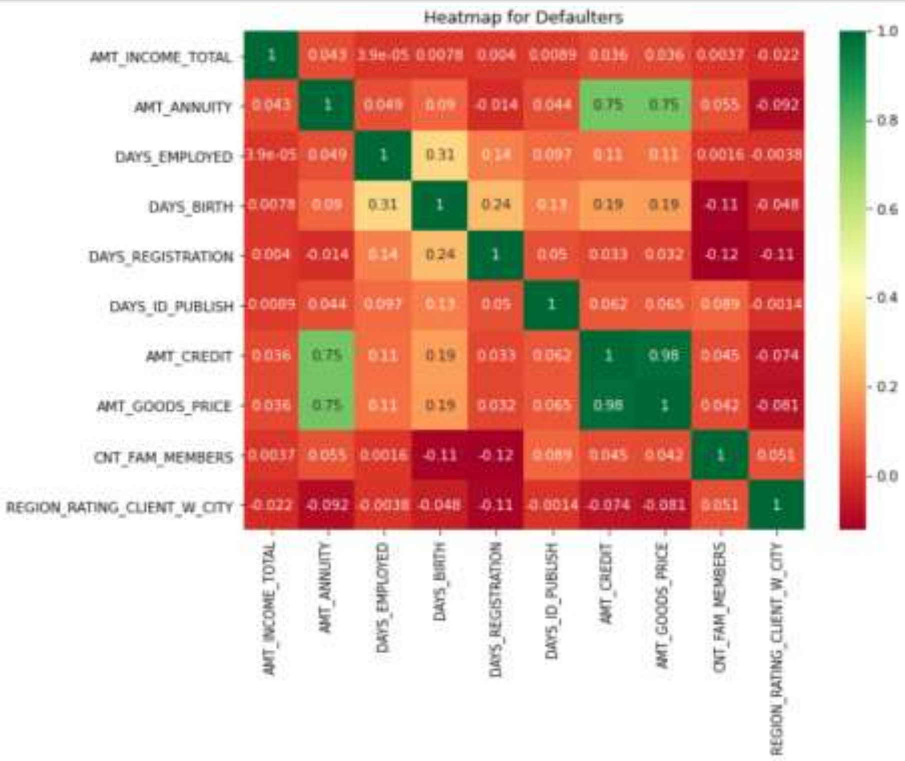
Education Type

- Through the above analysis we can see that the clients with the Lower secondary Education has the highest Defaulter percentage.

* In Loan Type there is an increase in an increase in the cash loan percentage for both Repayer and Defaulter (To know more about the other analysis kindly go through the code).

Multivariate Analysis





Heatmap of Repayer and Defaulter

- We observe that there is a high correlation between Credit amount and Goods Price.
- We can also find that there appears to be some deviancies in the correlation of Loan-Payment Difficulties(Defaulter) and Loan- Non Payment Difficulties(Repayer) such as Credit amount v/s Income Total.

* To check the Top 10 Correlation for Repayer and Defaulter client, kindly go through the program.

Previous Application

- Importing the necessary libraries.

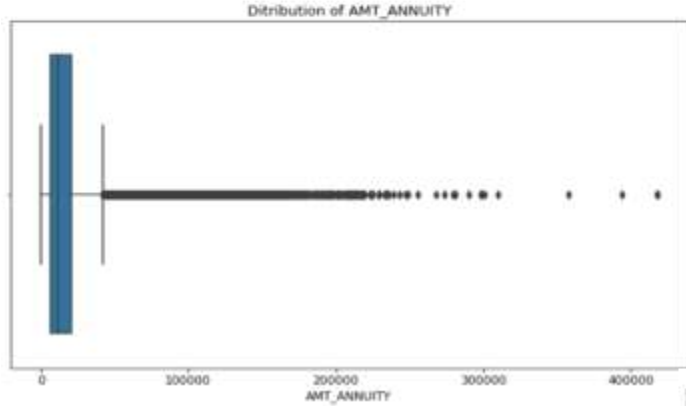
- Then we inspect the dataframe and understand all the attributes.

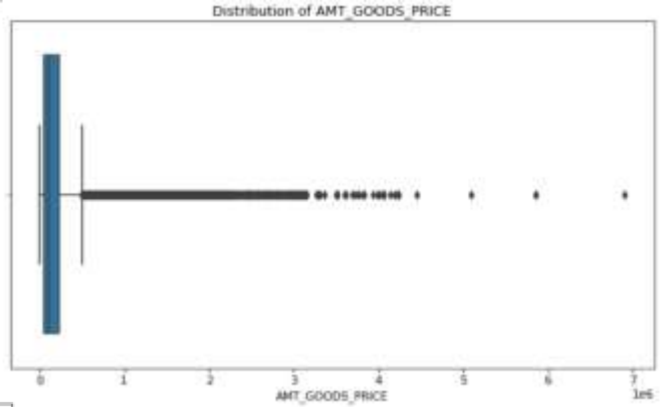
- Finding the missing data from the dataframe and then find the percentage of null values,

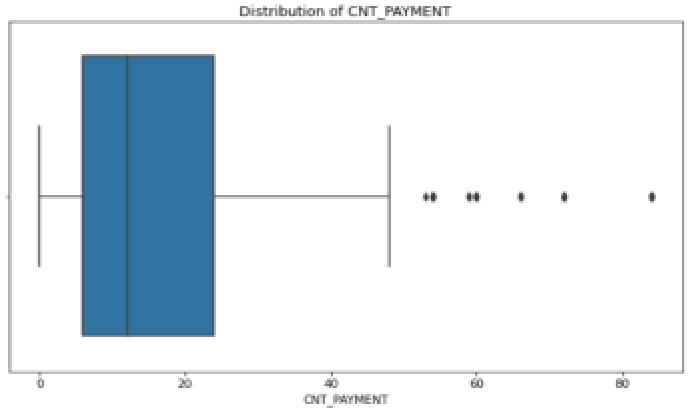
- Then taking the data drop percentage(30%) and removing columns accordingly from the dataset.

- We then clean the data left and replace the left-over null values by mean value of the respective column as implemented on “AMT_ANNUITY”, “CNT_PAYMENT” and ‘AMT_GOODS_PRICE’ column.

- In later stage of analysis we also do merge application and previous application dataframe in terms of specified columns.





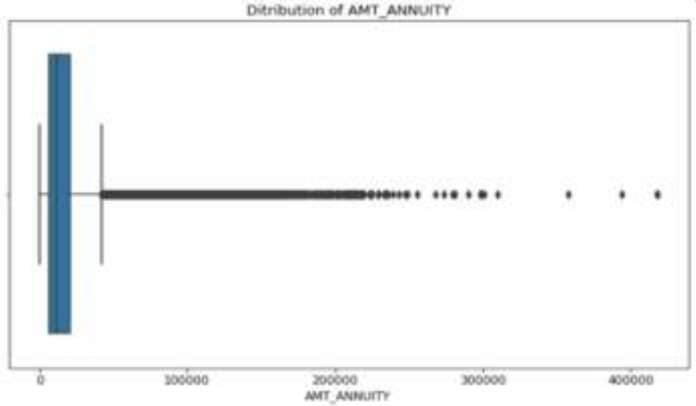


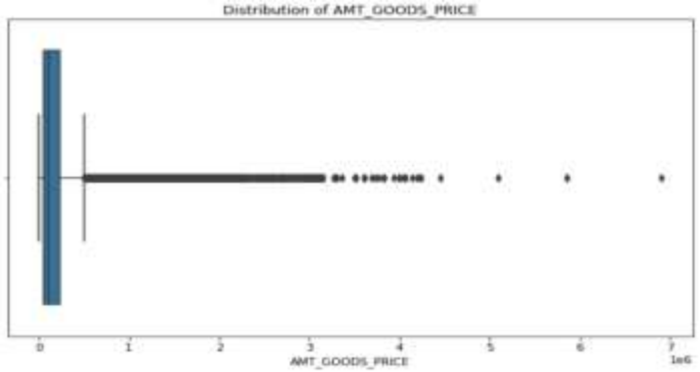
Analyzing the Outliers

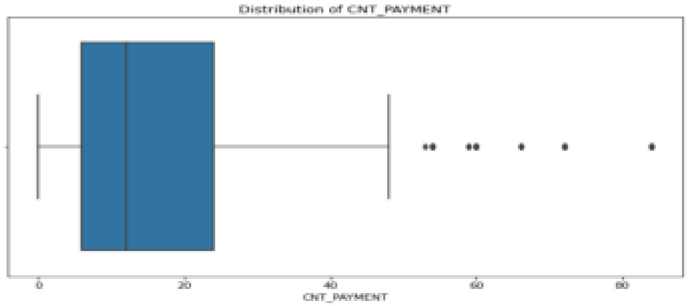
- Analyzing AMT_ANNUITY we can observe that it has outliners above the 100000 range which is almost above the upper whisker.

- Analyzing AMT_GOODS_PRICE we can observe that it has outliners above the 1.0x10^6 range which is almost above the upper whisker.

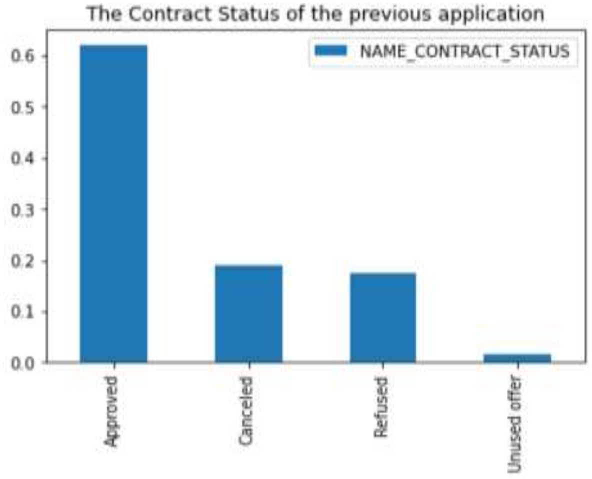
- Analyzing CNT_PAYMENT we can observe that it has outliners almost above the range from 60 which is almost above the upper whisker.



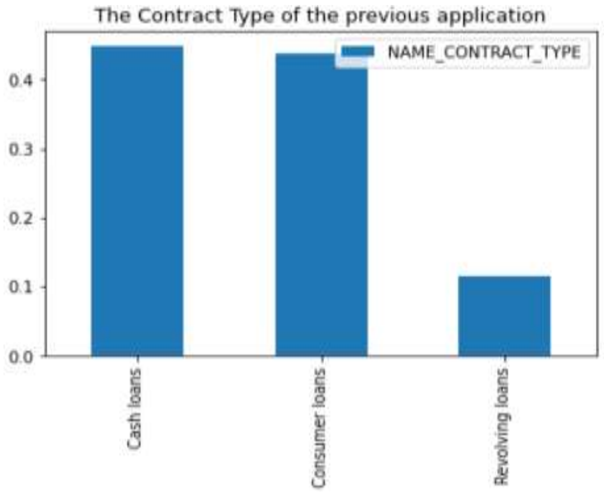




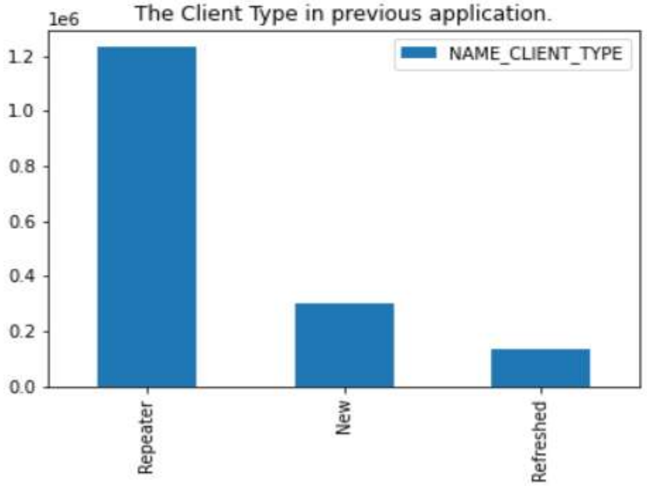
Univariate Analysis



- We can see observe that majority percentage of the loans are approved.
- We can also see quite less percentage of loans are being refused.
- The lest percentage of loans are the ones which are in the Unused offer category.



- We can see observe that majority percentage of the loans were Cash loans.
- The lest percentage of loans are the ones which are in the Revolving loans category.

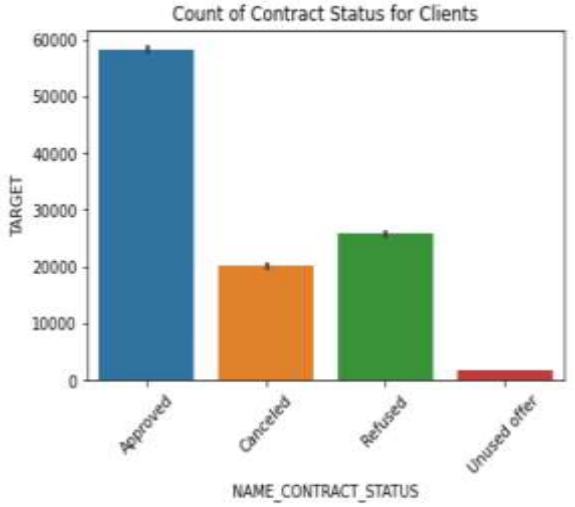


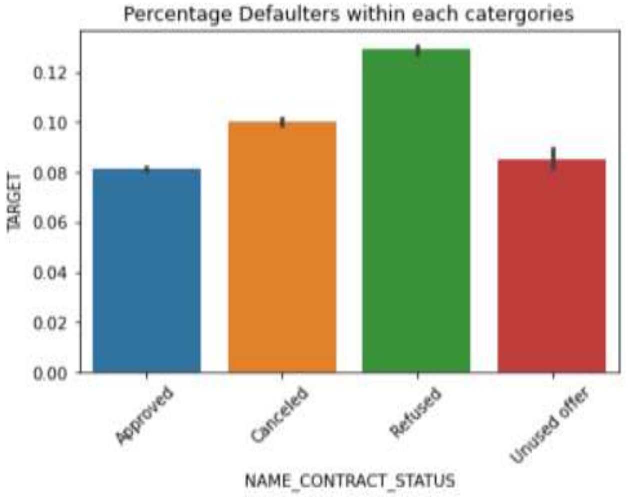
- On observing the above plot we can see that most of the clients are Repeater.

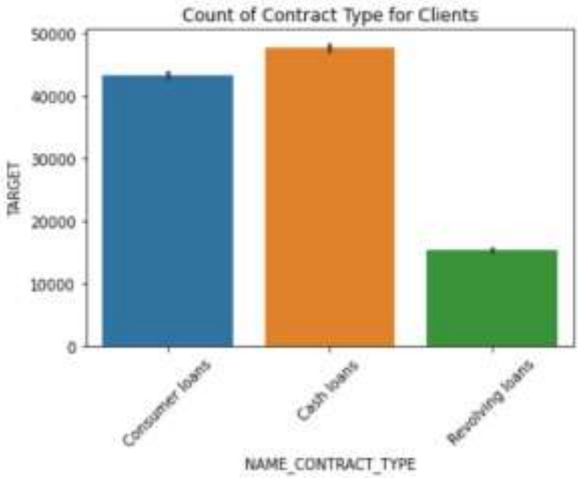
Note:-

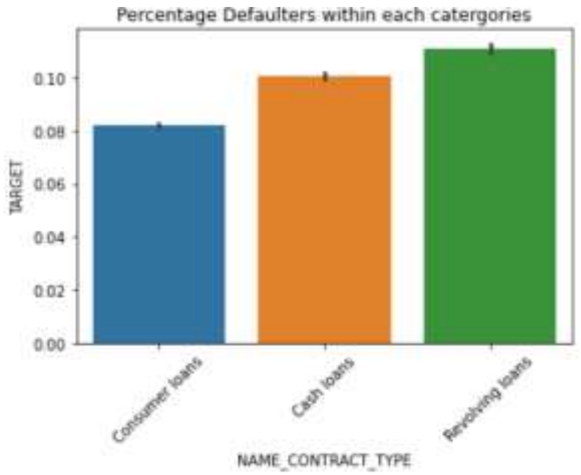
* Similar to the above analysis, univariate analysis on some other attributes like Payment type , Goods Category etc is also done which provides interesting view on the clients who had applied for loans.

Univariate Analysis for Repayer and Defaulter.









Contract Status

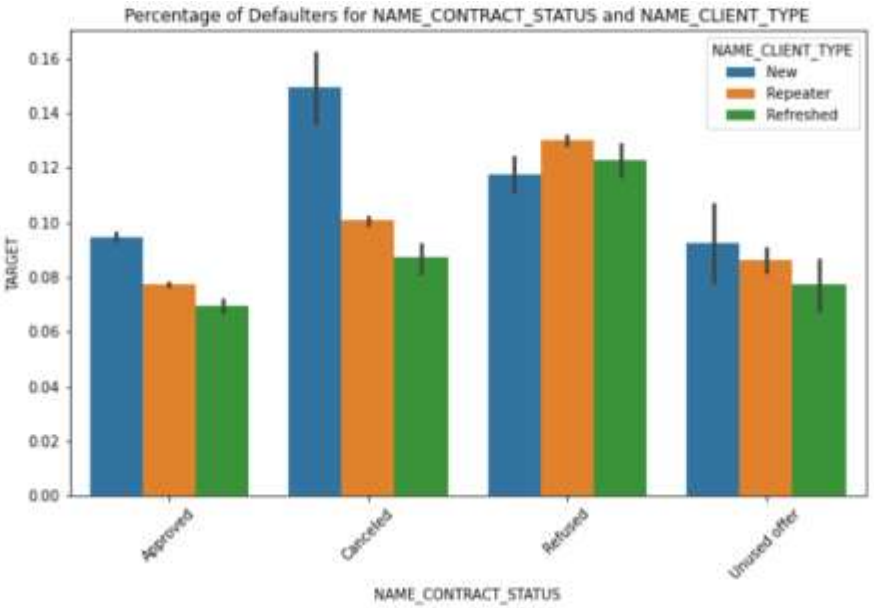
- On observing the first graph it can be seen that most of the contracts from previous application have been Approved.
- We analyze the following from the second graph.
  - In the previous application 'Refused' contracts are the ones who have maximum percentage of Defaulters from the current application.
  - In the previous application 'Approved' contracts are the ones who have minimum percentage of Defaulters from the current application.

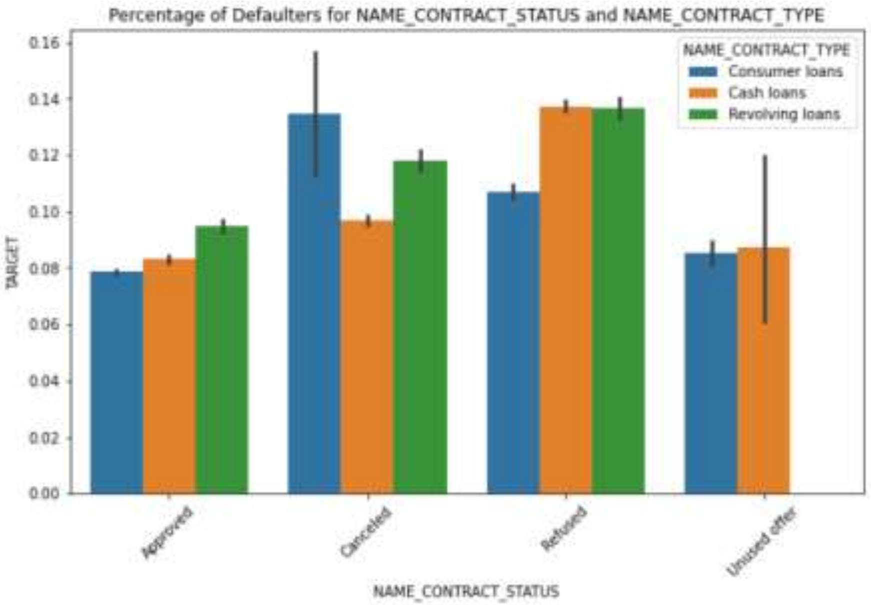
Contract Type

- On observing the first graph it can be seen that most of the contracts from previous application have been Approved.
- We analyze the following from the second graph.
  - In the previous application 'Revolving loans' contracts are the ones who have maximum percentage of Defaulters from the current application.
  - In the previous application 'Consumer loans' contracts are the ones who have minimum percentage of Defaulters from the current application.

* In Client Type and other attributes as well, we have made some analysis (kindly go through the code file to get in-depth information on the analysis).

Bivariate / Multivariate Analysis





Contract Status

- We can observe from the above graph that Client who where 'New' and had 'Cancelled' previous application tend to have more percentage of Loan-Payment Difficulties in current application.

Contract Type

- We can observe from the above graph that Clients with 'Revolving loans' and with 'Refused' previous application tend to have more % of Loan-Payment Difficulties in current application.

* In Client Type as well we have made some analysis (kindly go through the code file to get in-depth information on the analysis).

Conclusion

Application Data

- In the 'NAME_INCOME_TYPE' we can see that the count of 'Maternity Leave' is comparatively very less, and it also displayed that it has maximum percentage of payment difficulties. Therefore, we can say that the income type of 'Maternity leave' are the driving factors for Loan Defaulters.

- In the 'NAME_INCOME_TYPE' we can see that the count of 'Working' clients is the highest, yet it has the least percentage of Loan Defaulter clients.

- In the 'NAME_EDUCATION_TYPE' we can see that the count of 'Lower Secondary' is comparatively very less, and it also displayed that it has maximum percentage of payment difficulties. Therefore, we can say that the education type of 'Lower Secondary' are the driving factors for Loan Defaulters.

- In the 'NAME_EDUCATION_TYPE' we can also see that the count of 'Secondary/ secondary special' clients are quite high, yet it has comparatively less percentage of Loan Defaulter then that seen in 'Lower Secondary'.

- In the 'NAME_CONTRACT_TYPE' we can see that both count and percentage of defaulters are having Loan type as Cash loans. Therefore, we can say that the loan type of 'Cash Loans' are the driving factors for Loan Defaulter clients.

Previous Application

- In the 'NAME_CONTRACT_STATUS' we can see that the count of 'Refused' is comparatively less and it also has the maximum percentage of payment difficulties. Therefore, clients who are having the contract status as 'Refused' in previous application are the driving factors for Loan Defaulters.

- In the 'NAME_CONTRACT_TYPE' we can see that the count of 'Revolving Loans' is comparatively very less, and it also has the maximum percentage of payment difficulties. Therefore, clients who are having the contract type as 'Revolving loans' in previous application are the driving factors for Loan Defaulters.

- The observation from the above presented graph is that Clients with 'Revolving loans' and 'Refused' previous application tend to have more percentage of payment difficulties in current application.

- The above written observations can be proven correct because 'Revolving loans' and 'Refused' is comparatively less, clients with ‘Revolving Loans’ and ‘Refused’ previous application are driving factors for Loan Defaulters.

* In Client Type as well we have made some analysis (kindly go through the code file to get in-depth information on the analysis).

THANK YOU