CREDIT EDA ASSIGNMENT

By
MAYURI PARAG KHONDEKAR.

INTRODUCTION

- This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding - 1

- • 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 the data. This will ensure that the applicants capable of repaying the loan are not rejected. • 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.

Business Understanding - 2

- 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. When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
- 1. Approved: The Company has approved loan Application
- 2. Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
- 3. Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
- 4. Unused offer: Loan has been cancelled by the client but on different stages of the process. In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Business Objectives

- This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e., the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Data Understanding

- This dataset has 3 files as explained below:
- 1. 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
- 2. 'previous_application.csv' contains information about the client’s previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- 3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Reading the dataset.

- Initially two datasets are given. And we have to import necessary libraries like numpy, pandas, matplotlib and seaborn etc.
- Now we have to read the path of application dataset, and previous dataset as well.

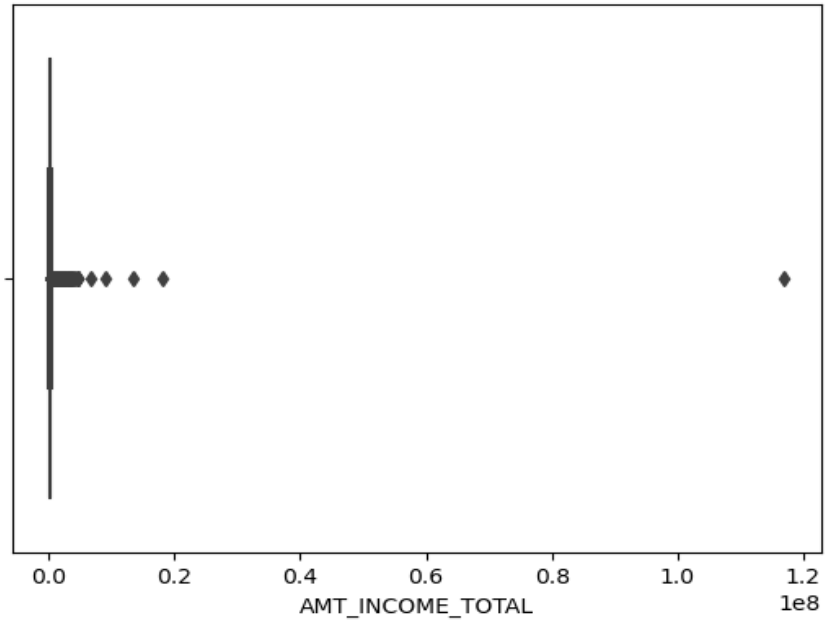
To clean the dataset.

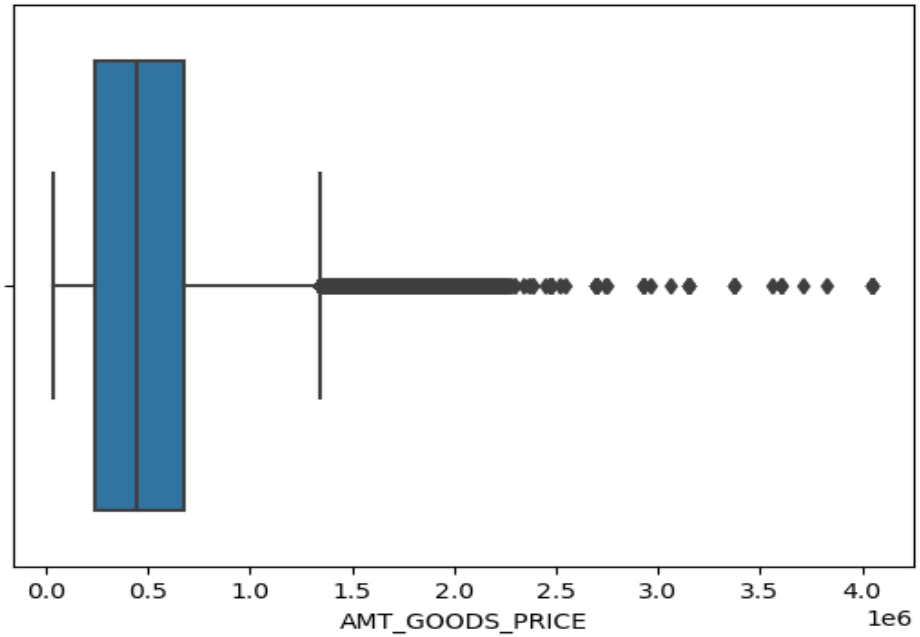
- Before we analyse the data and make insights form the data, we need to observe and understand the data after that need to clean the data with various steps.
- Like finding null values and outliers and fixing them with appropriate strategy.
- Also if there are any anomalies then we should fix them.
- by following these steps before analysis process will make the analysis easier and finding insights will go in a right direction which would be benificial for our respective organization.

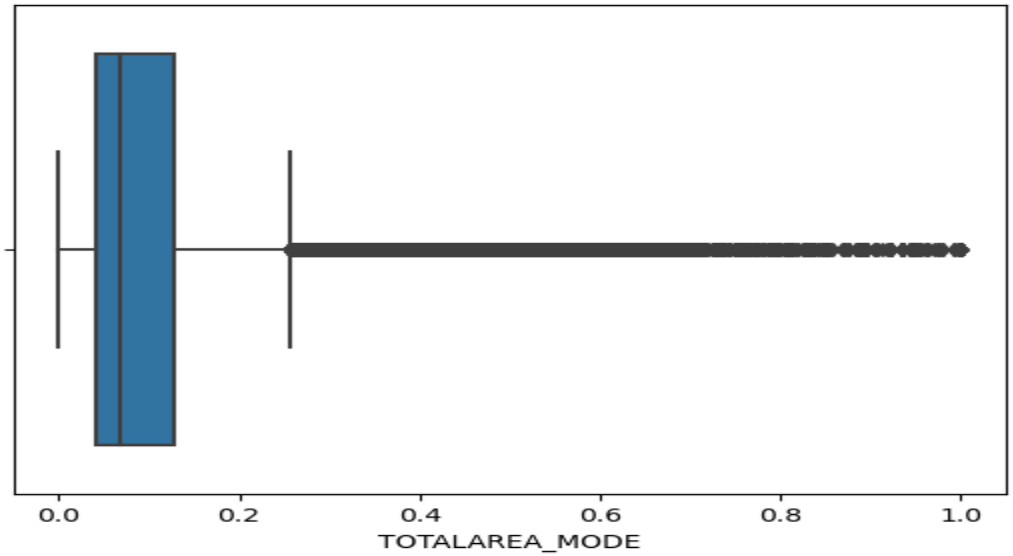
Identification and treatment of null values in the dataset

- To identify the null values in the dataset we can use isnull() function. And find the percentages of the null values.
- After that we should drop the null values which are above 50%. With the help of the drop() function.

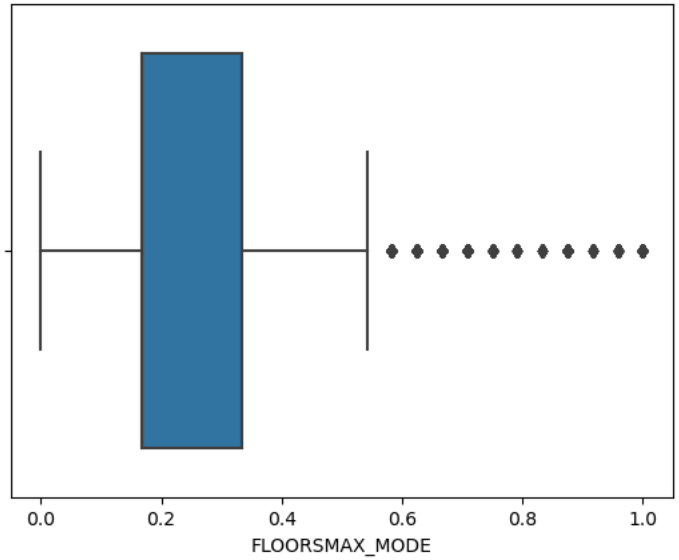
Identification of outliers in the dataset using boxplot

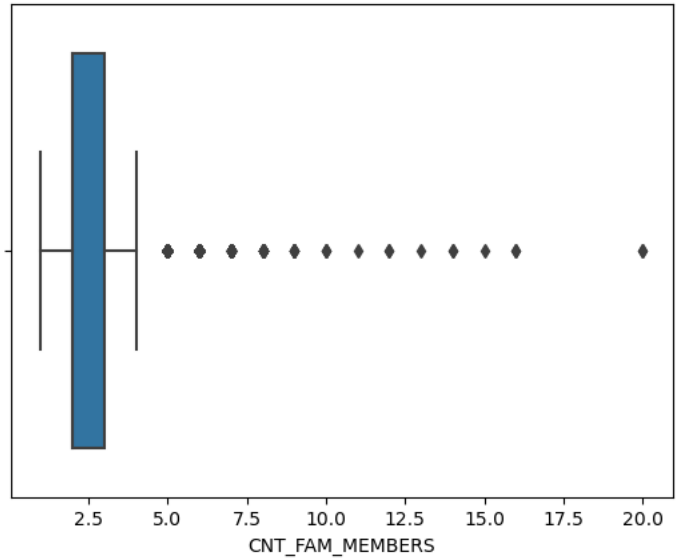


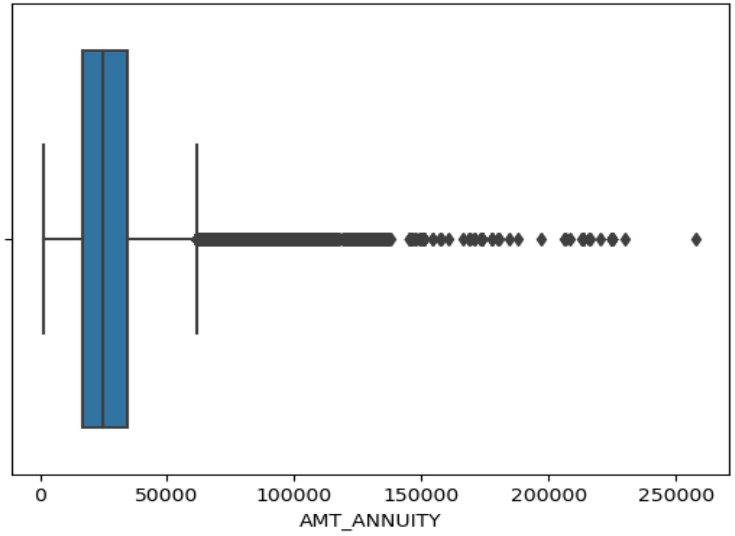




Identification of outliers using boxplot





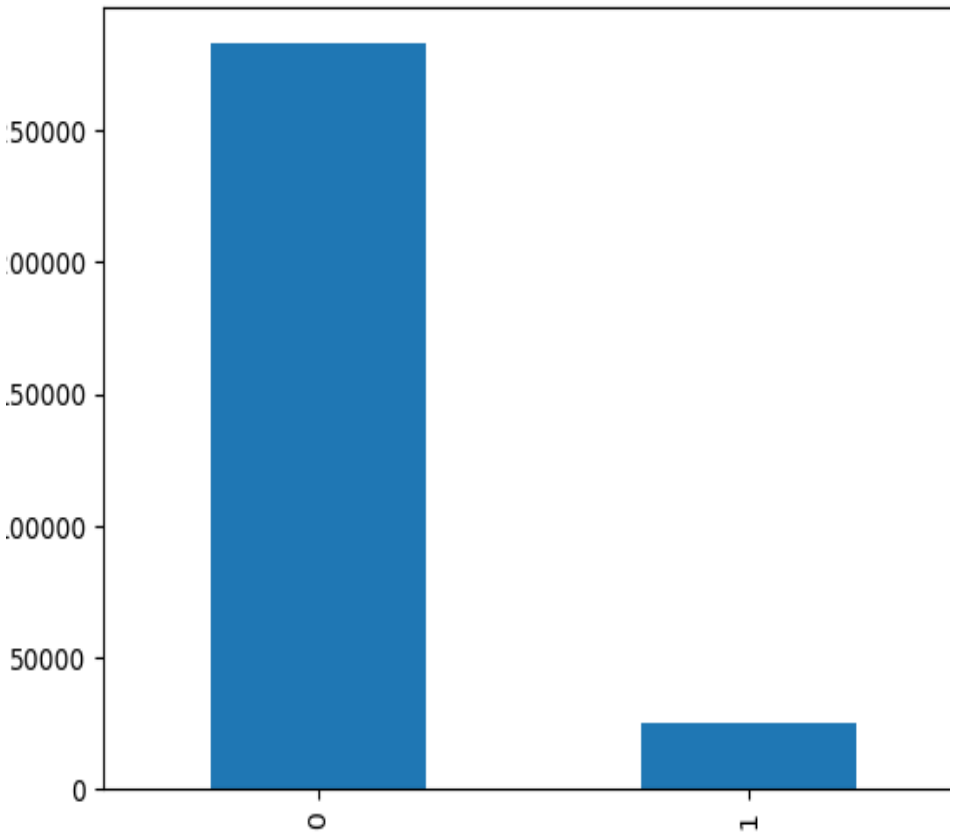


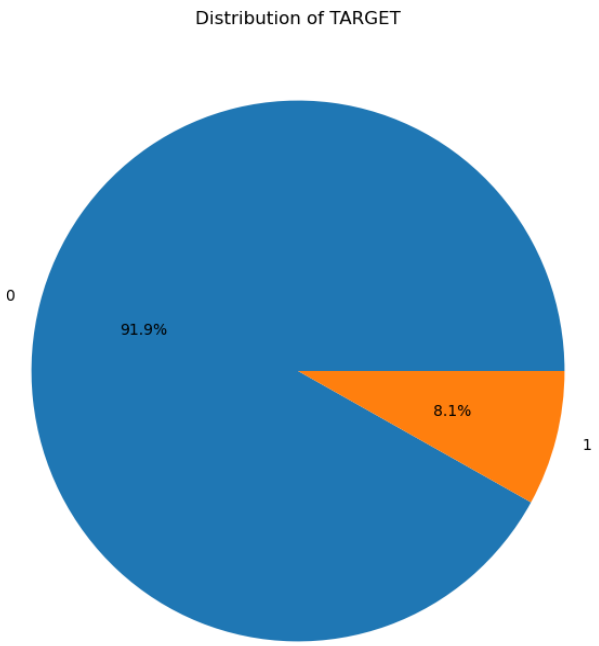
Treatment of outliers in the dataset

- We can treat the outliers by three method
- 1. Imputation method
- 2. Capping method or IQR method
- 3. Remove the outliers.
- -Imputation method is not preferred because it will affect the other values in the column so that insights would be wrong.
- -Removing the outliers is not a good practice as it may leads to loss of the important data.
- -Capping or IQR method is mostly preferred because it very easy to perform and it will not affect other values in the column,
- so that insight of the data, observation and decisions taken by data analyst will go in a right direction.
-
- -In these ways we can take care of an outliers in each columns from the dataset.

ANALYSIS

- Definition :
- Data analysis is the process of inspecting, cleaning, transforming, and interpreting data to discover useful information, draw conclusions, and support decision-making.
- Data Collection.
- Data Cleaning.
- Data Exploration
- Data Visualization.
- Reporting and Presentation.
- These are few steps that should be followed in data analysis process.





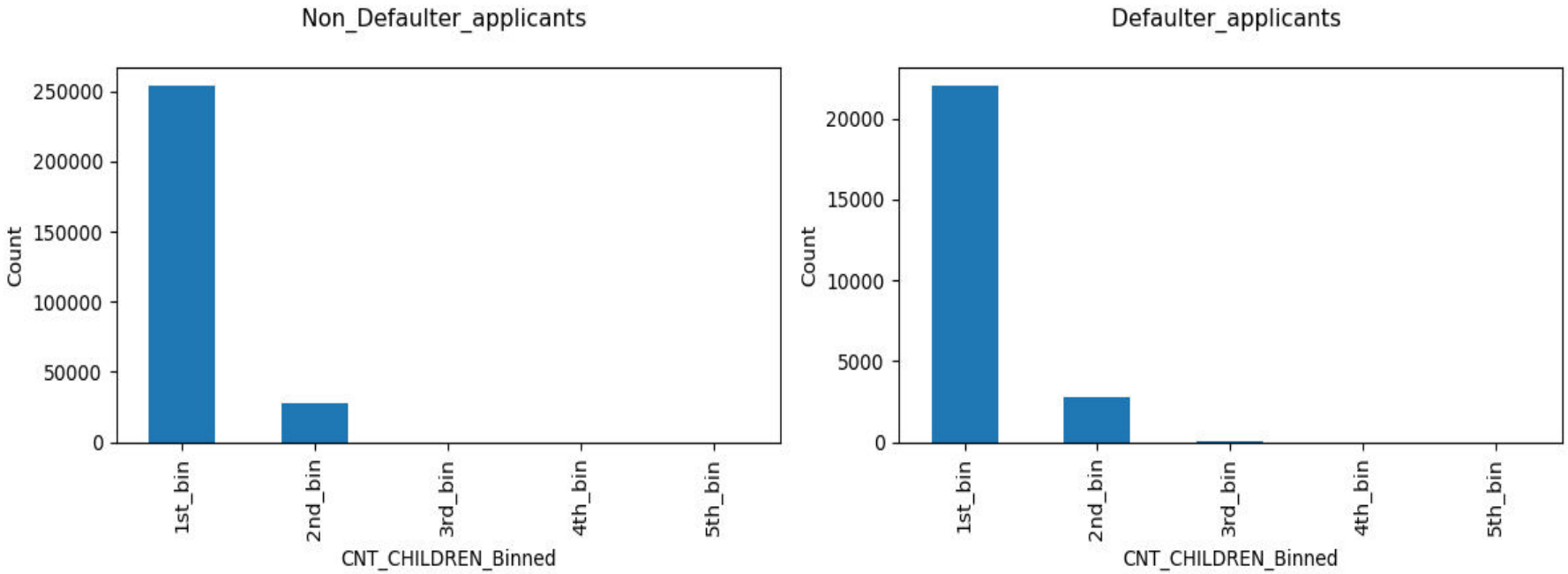
To check percentage of imbalance of TARGET variable.

Conclusion : It means that 91.92% applicants are non_defaulter while only 8.07% applicants are defaulter.

Strategy for univariate and bivariate/multivariate analysis.

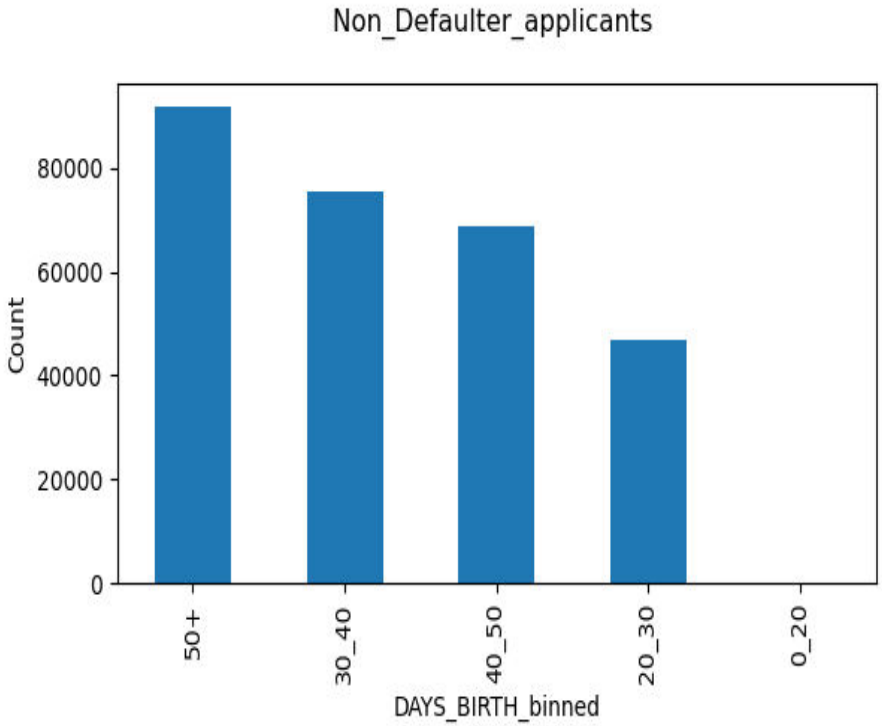
- If we want to visualise a categorical column then we can visualise it with the help of pie chart, barplot and stacked bar charts as well.
- If we want to visualise a numerical column then we can visualise it with the help of histogram, boxplot, Violin Plots and meny more plots .

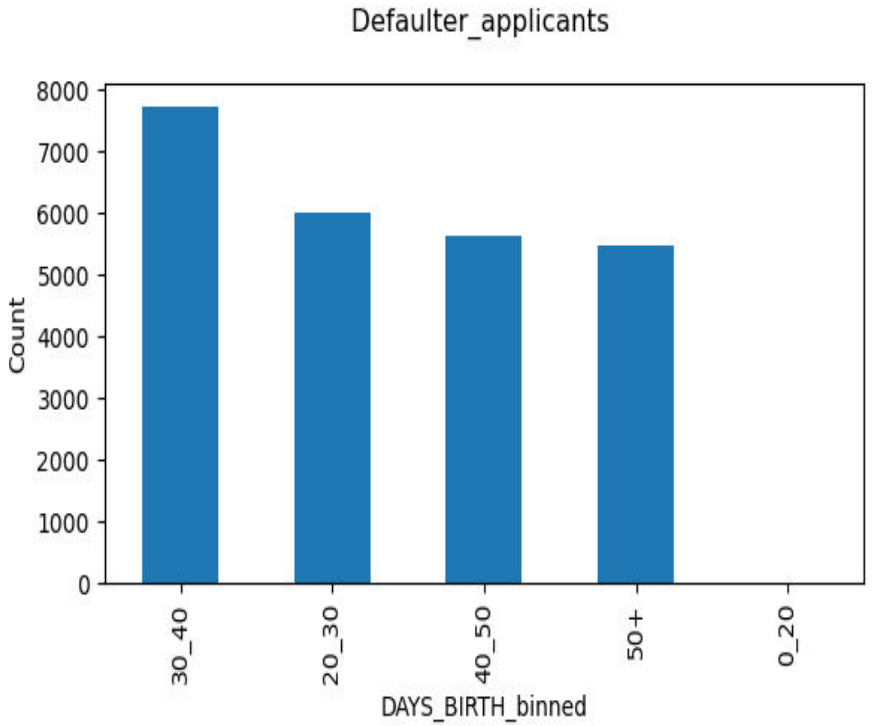
Univeriate analysis of CNT_CHILDREN_Binned column



- from these subplots we can say that the count of 1st_bin with count of children ranges between 0 to 1 (applicants are having no chidren or only 1 child) are comparetively high.i.e. they are mostly non_defaulter applicants.
- whereas count of 1st_bin with children ranges between 0 to 1 (applicants are having no chidren or only 1 child) are comparetively low.i.e.(applicants are having 2 or more children) they are mostly defaulter applicants.

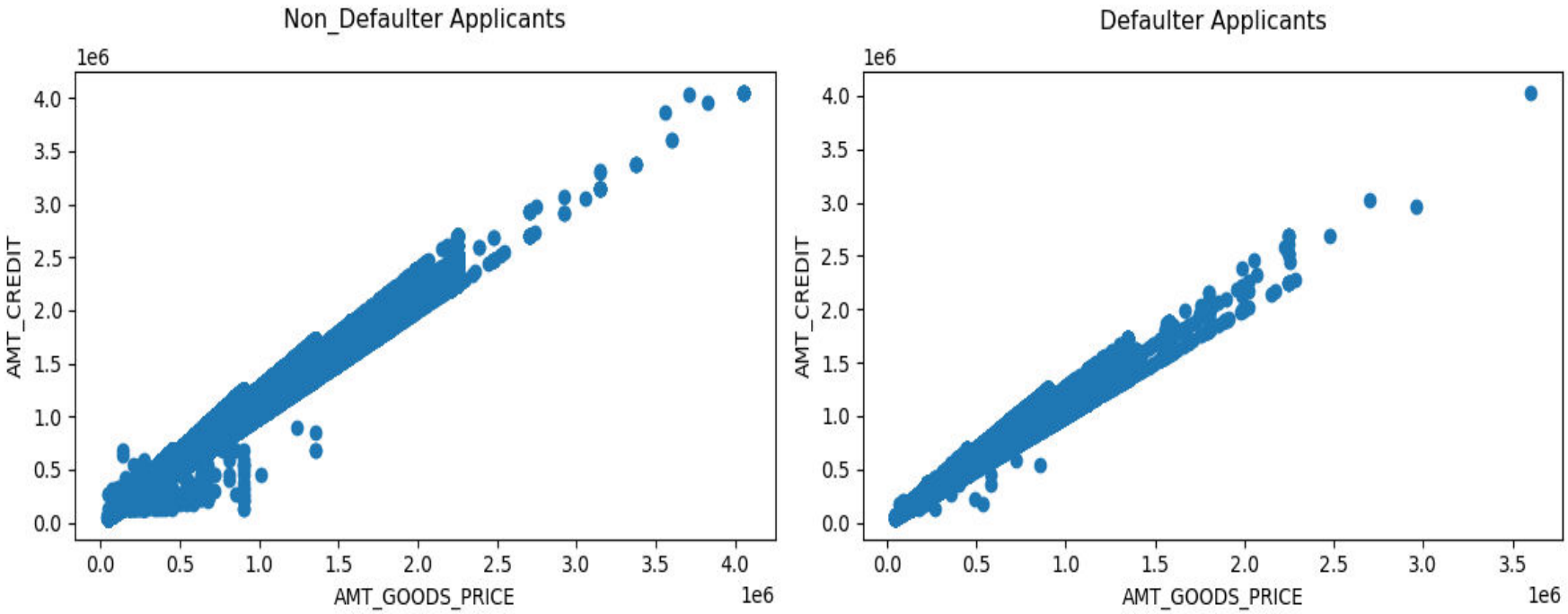
Univeriate analysis of DAYS_BIRTH_binned column





- from these subplots we can say that the old_People with age ranges above 50 are mostly non_defaulter applicants.
- whereas Adult people age ranges between 30 to 40 are mostly belongs to defaulter applicants.

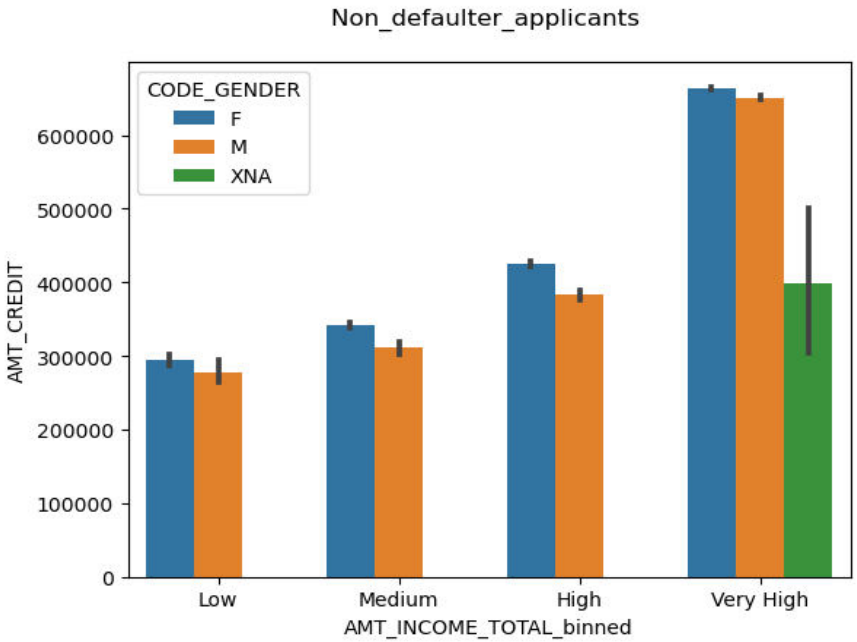
Bivariate analysis / multiveriate analysis of AMT_GOODS_PRICE & AMT_CREDIT



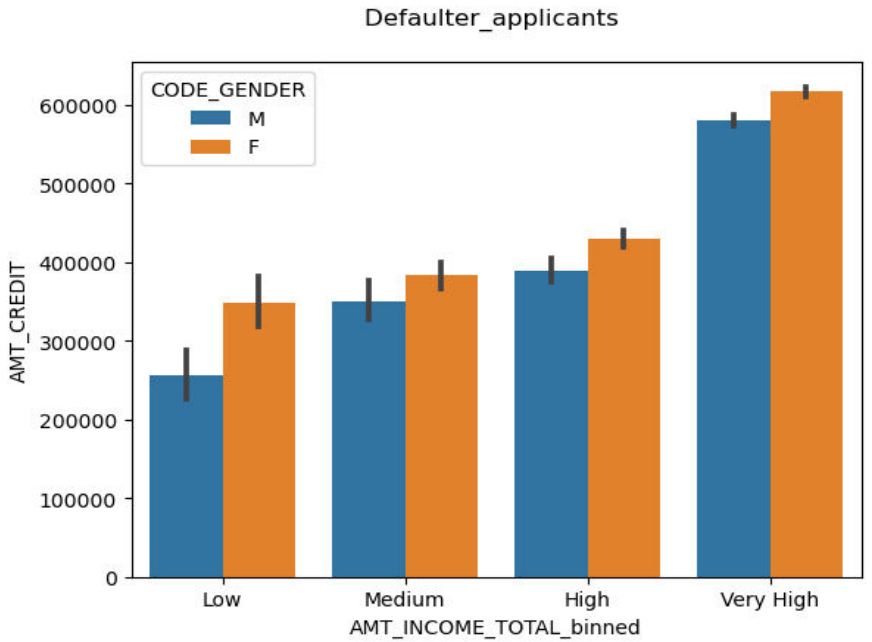
- AMT_GOODS_PRICE and AMT_CREDIT have strong positive correlation between them. From these two plots we can say that as AMT_GOODS_PRICE increases, the Credit_Amount also increases.

Bivariate analysis of AMT_INCOME_TOTAL_binned and AMT_CREDIT with CODE_GENDER as hue.

TARGET_0_APPLICANTS



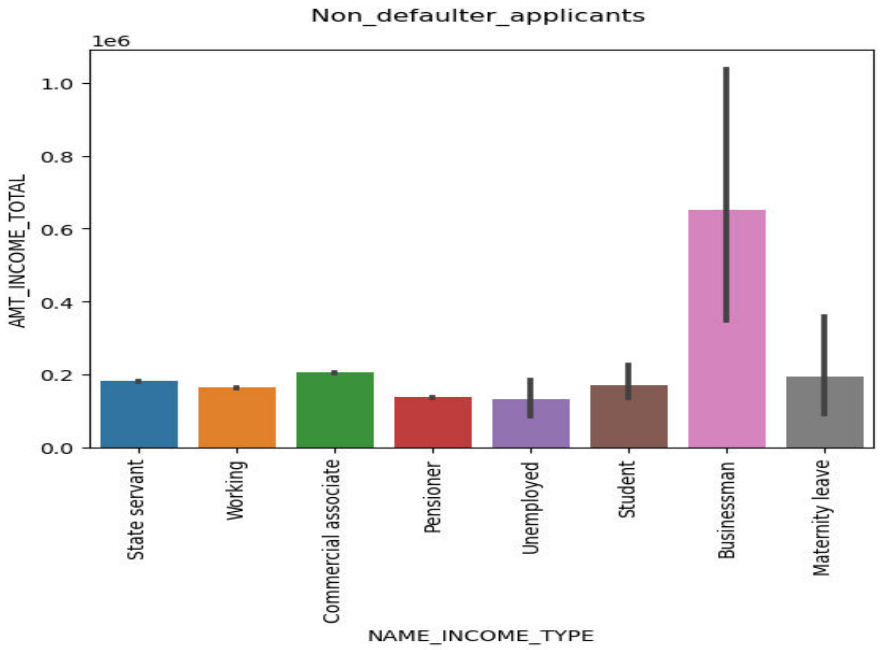
TARGET_1_APPLICANTS



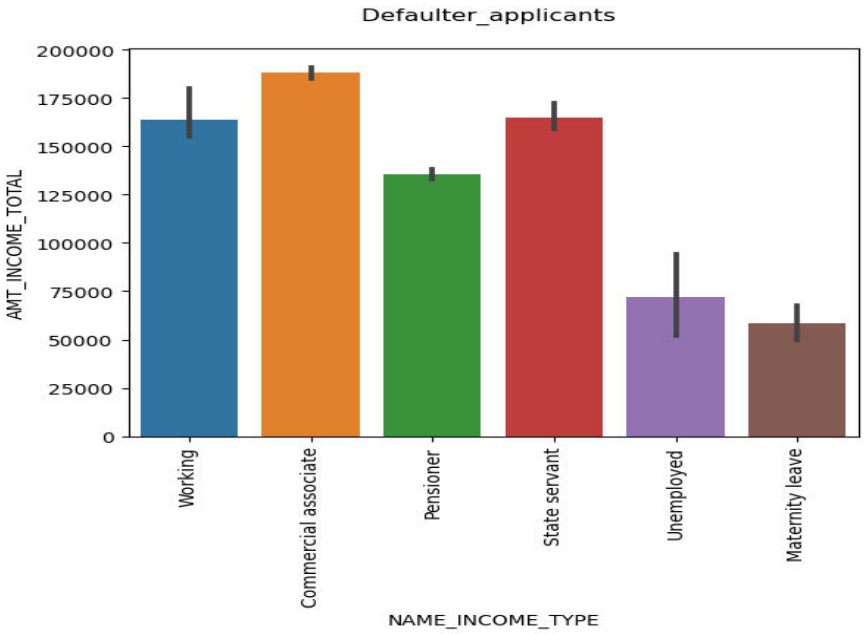
- Male with very high income amount get highest credit amount than female comes in non defaulter category.
- while male with very high income get low credit amount than female comes in defaulter category.

Bivariate analysis of NAME_INCOME_TYPE and AMT_INCOME_TOTAL without hue.

TARGET_0_APPLICANTS

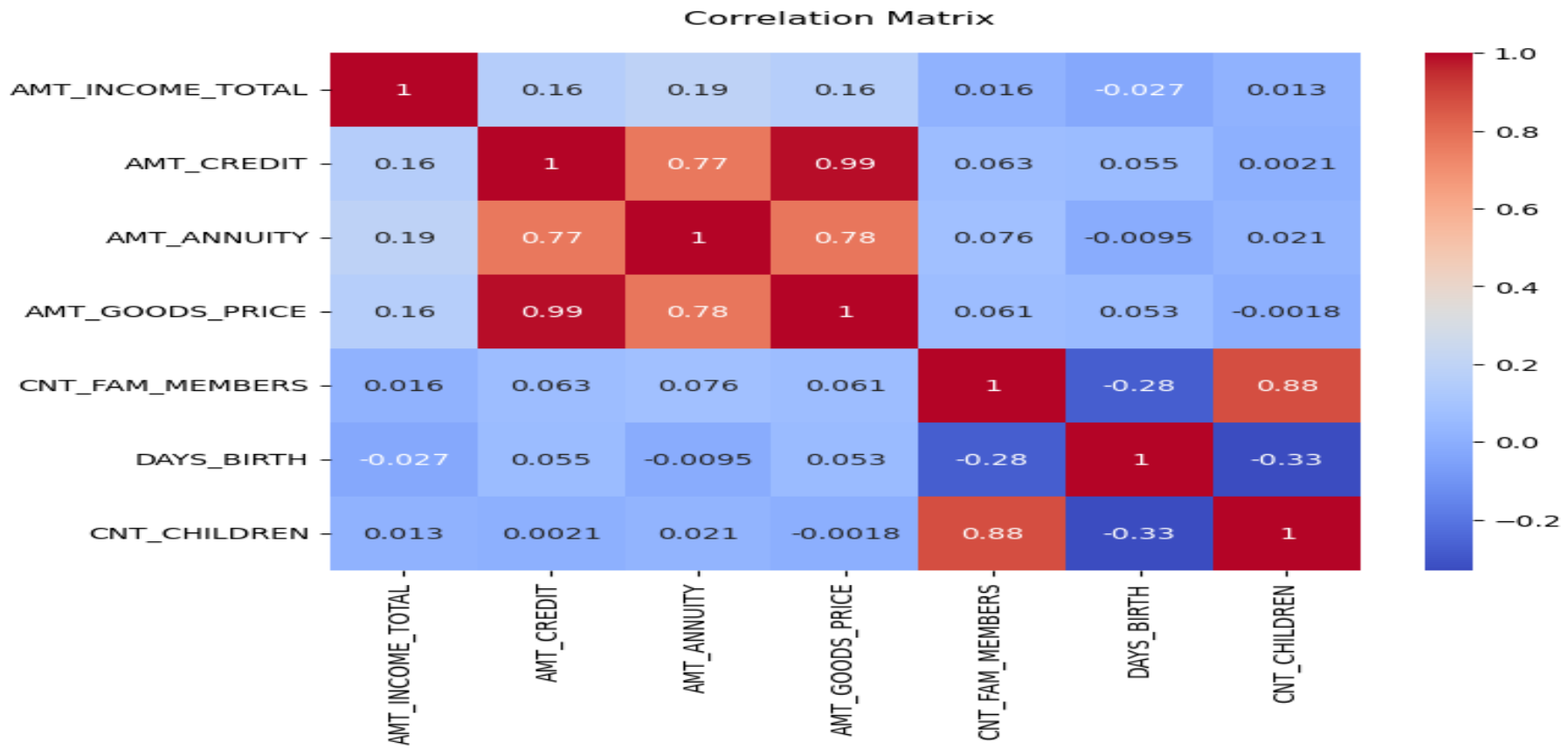


TARGET_1_APPLICANTS



- From the above barplot we can say that unknown males are having highest income range are mostly non defaulter applicants.
- On the other hand Married males are mostly Defaulter applicants.
- use of Hue in visualization plots can make easy to find insights from the data with respect to hue column.

Correlation Matrix

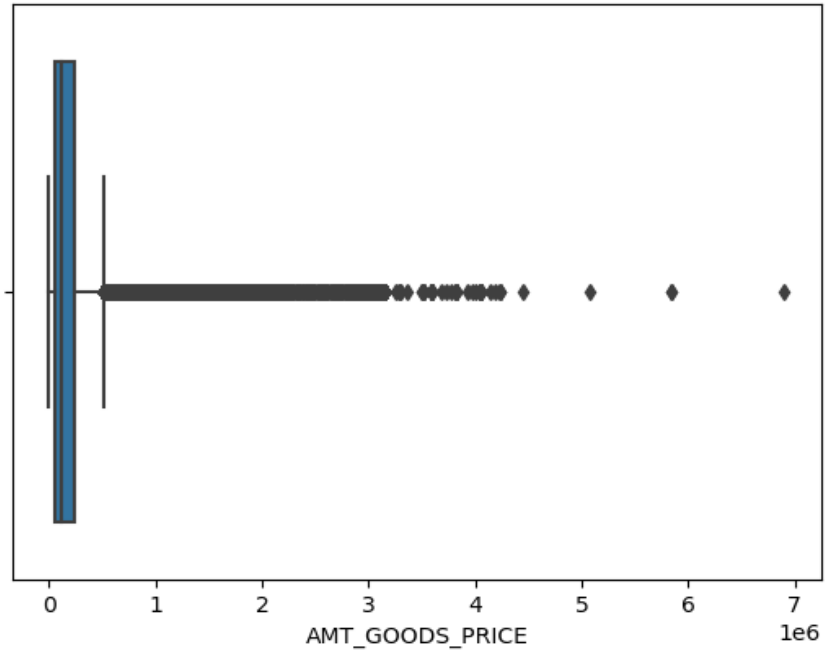


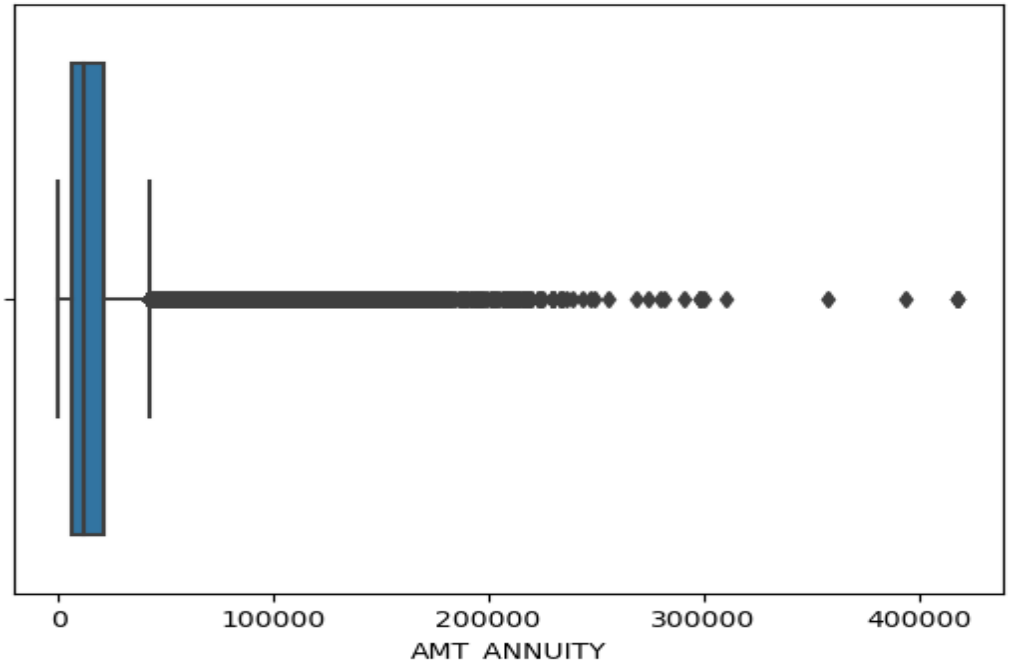
- AMT_GOODS_PRICE and AMT_CREDIT have very strong positive correlation between them.
- From these two plots we can say that as AMT_GOODS_PRICE increases,the Credit Amount also increases.
- CNT_FAM_MEMBERS are less strongly correlated CNT_CHILDREN the count of children increases so the count of family members also increases.

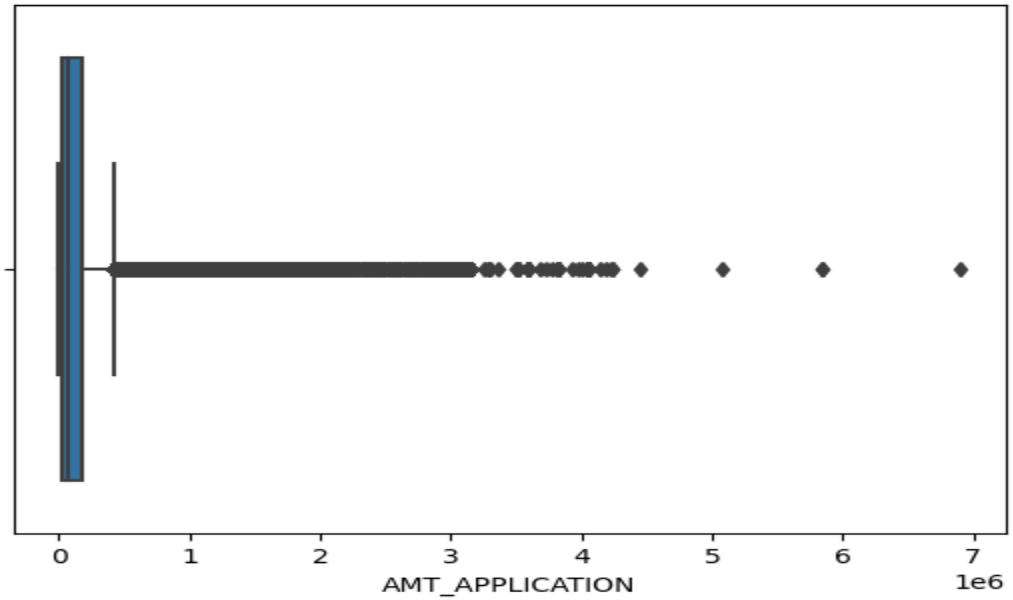
Reading the previous_application dataset

- Before we start analysis of previous dataset we again need to follow all the steps to clean the data. Like,
- Like finding null values and outliers and fixing them with appropriate strategy as mentioned in previous slides.
- Also if there are any anomalies then we should fix them.
- Now it we be easier to do analysis.

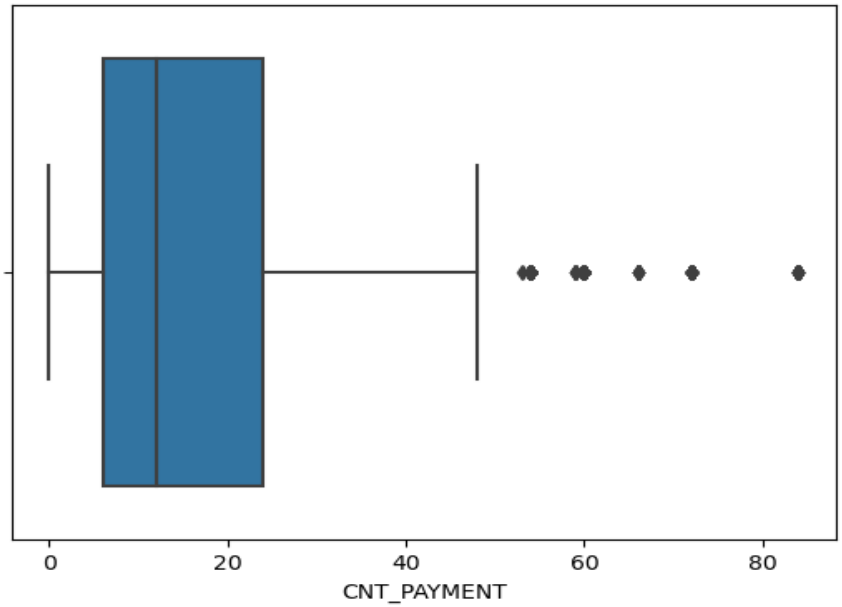
Identification of outliers using boxplot

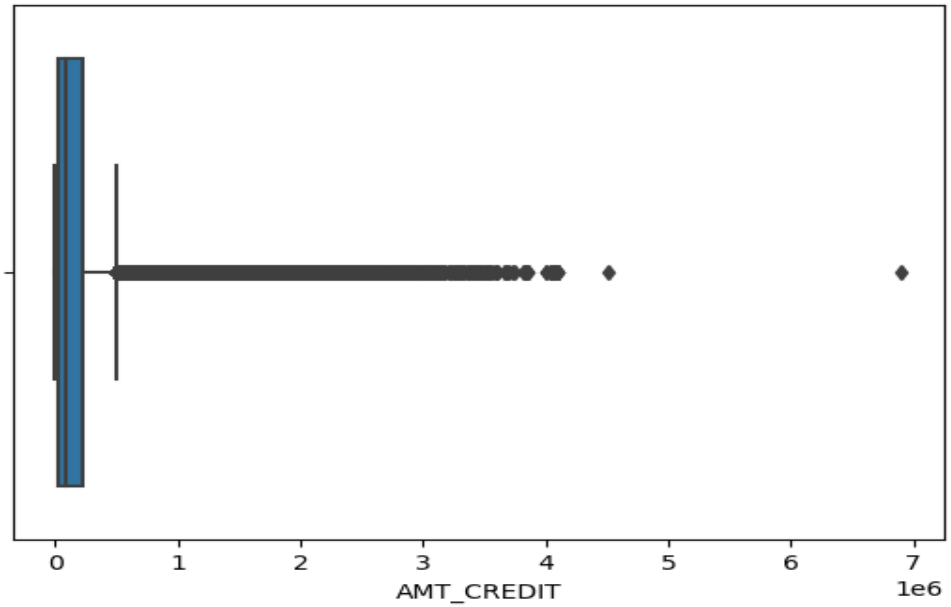


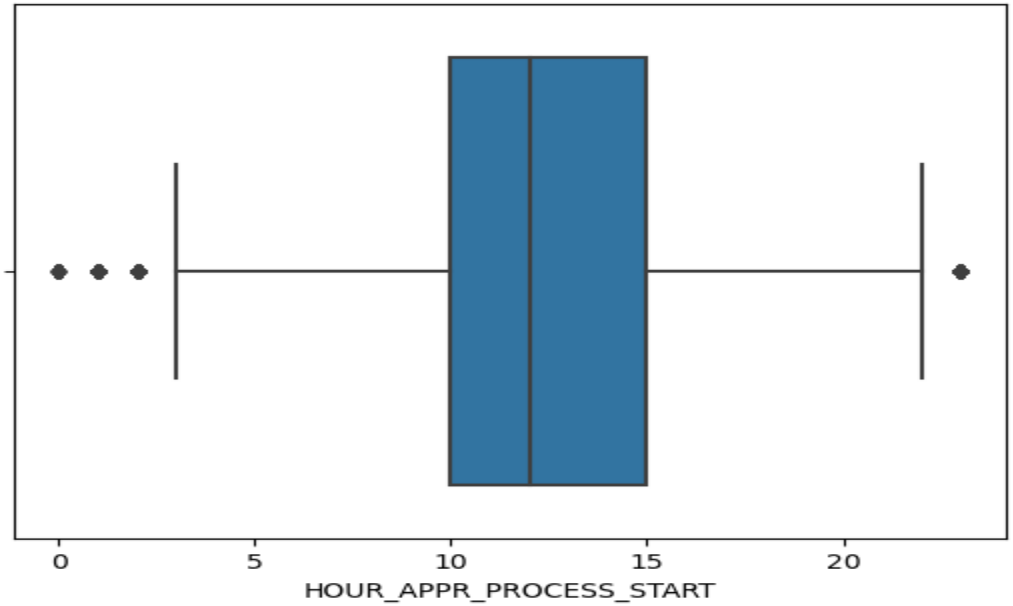




Identification of outliers using boxplot







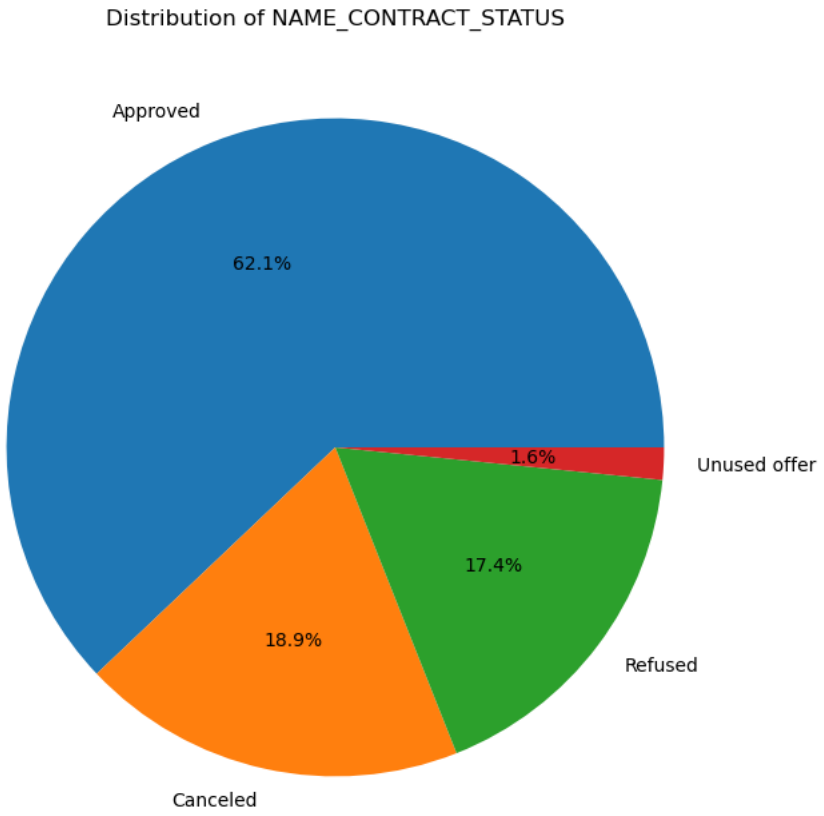
Treatment of outliers in previous_ application dataset.

- We can treat the outliers by three method

- # Imputation method
- # Capping method or IQR method
- # Remove the outliers.
- -Imputation method is not preferred because it will affect the other values in the column so that insights would be wrong.
- -Removing the outliers is not a good practice as it may leads to loss of the important data. -Capping or IQR method is mostly preferred because it very easy to perform and it will not affect other values in the column, so that insight of the data, observation and decisions taken by data analyst will go in a right direction.
- -In these ways we can take care of an outliers in each columns from the dataset.

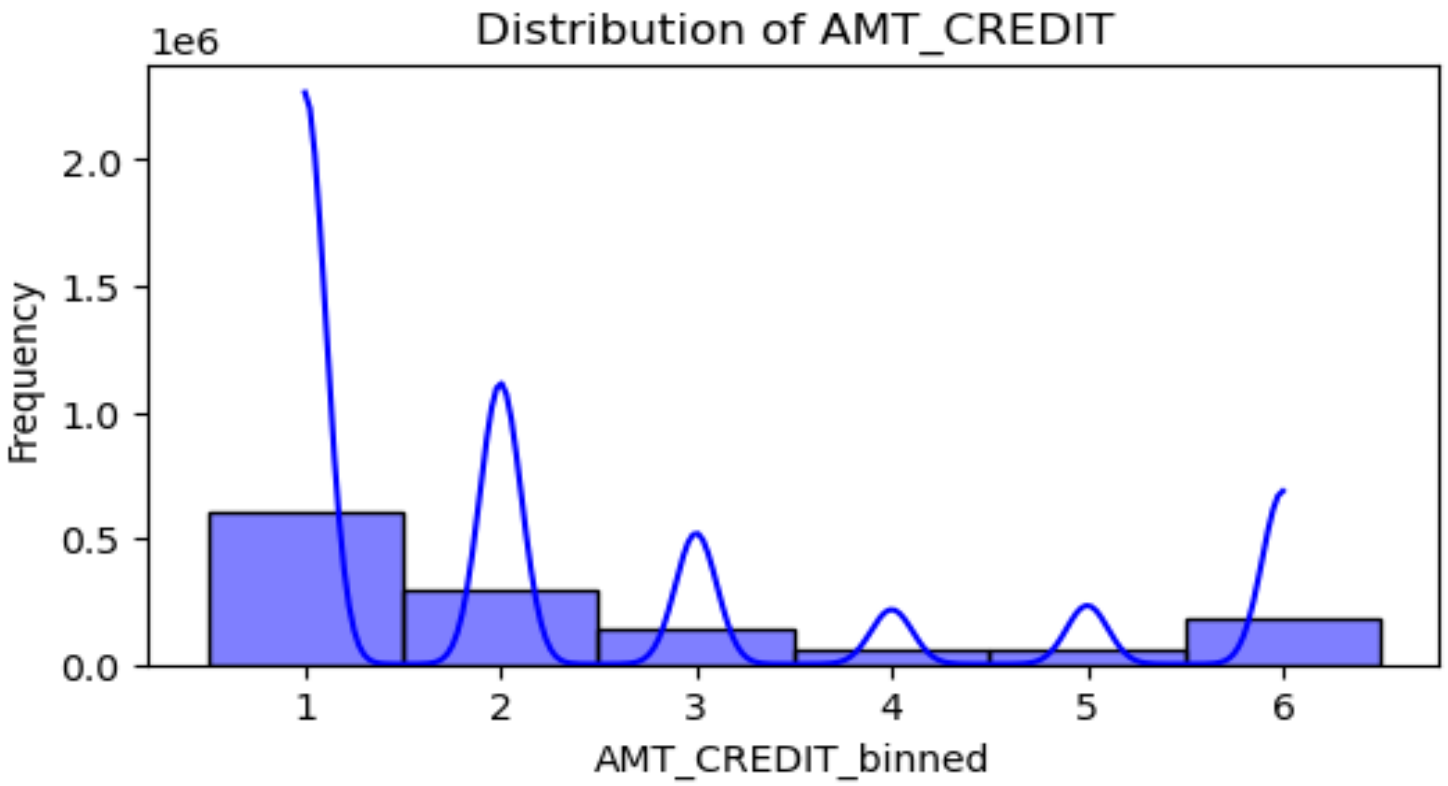
univariate analysis of previous dataset

univariate analysis of categorical column.



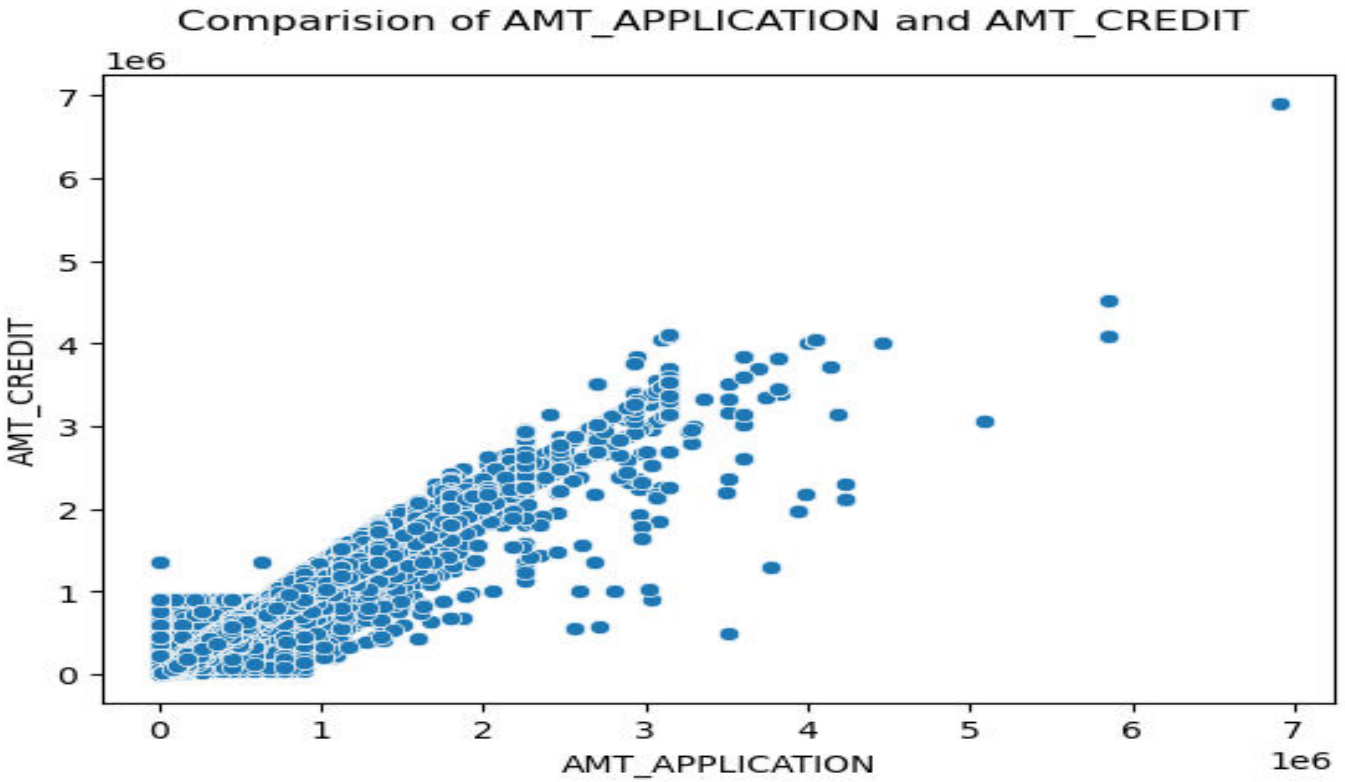
- Approved loan status is the highest among all loan applications
- Canceled loan status is the second highest among all loan applications

univariate analysis of categorical column



- By this plot we can assume that, most people received the loan amount that they applied for.

Bivariate/ multivariate analysis of numerical columns

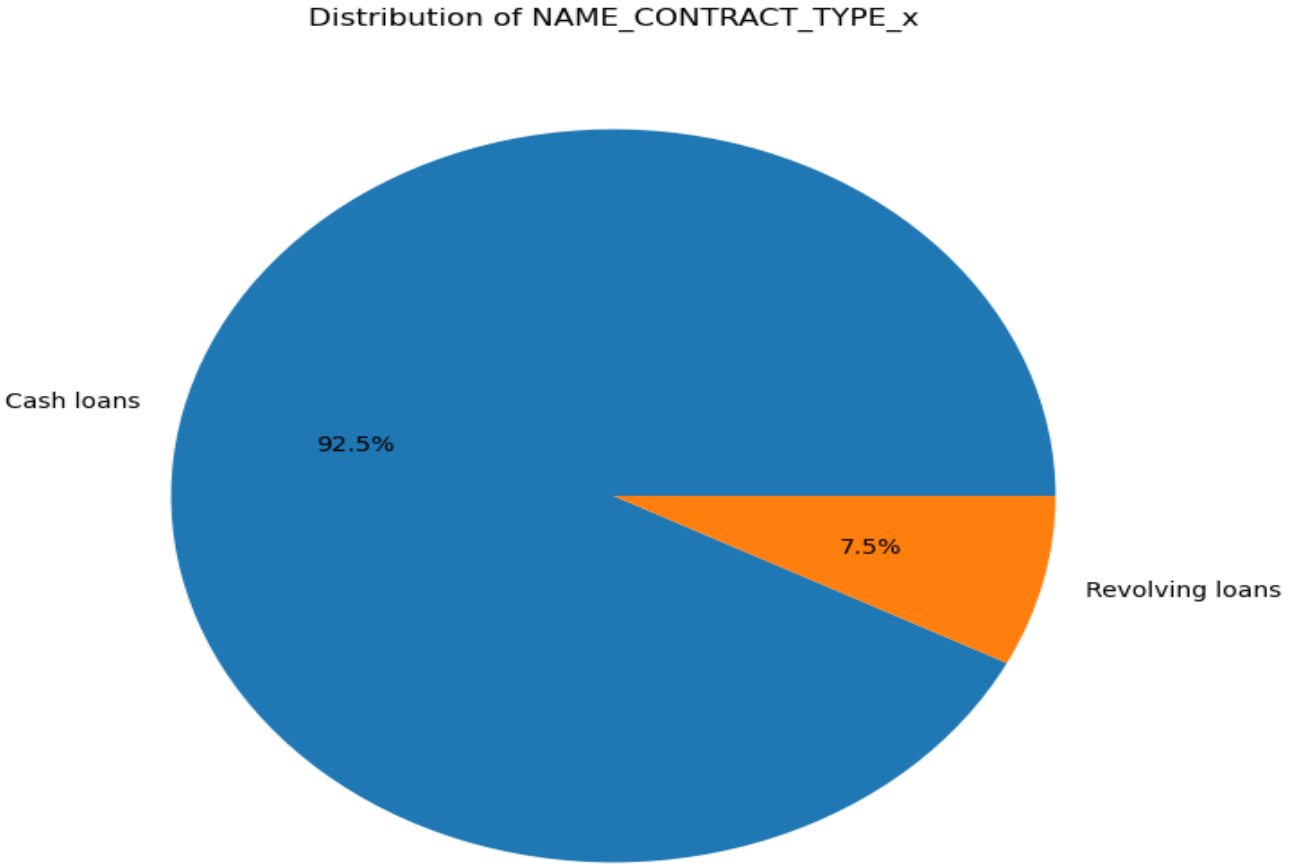


- Both the columns AMT_APPLICATION and AMT_CREDIT are strongly correlated with each other.
- We can say that, as count of AMT_APPLICATION increases so the count of AMT_CREDIT also increases.

Now we have to merge both the datasets and try to analyse it

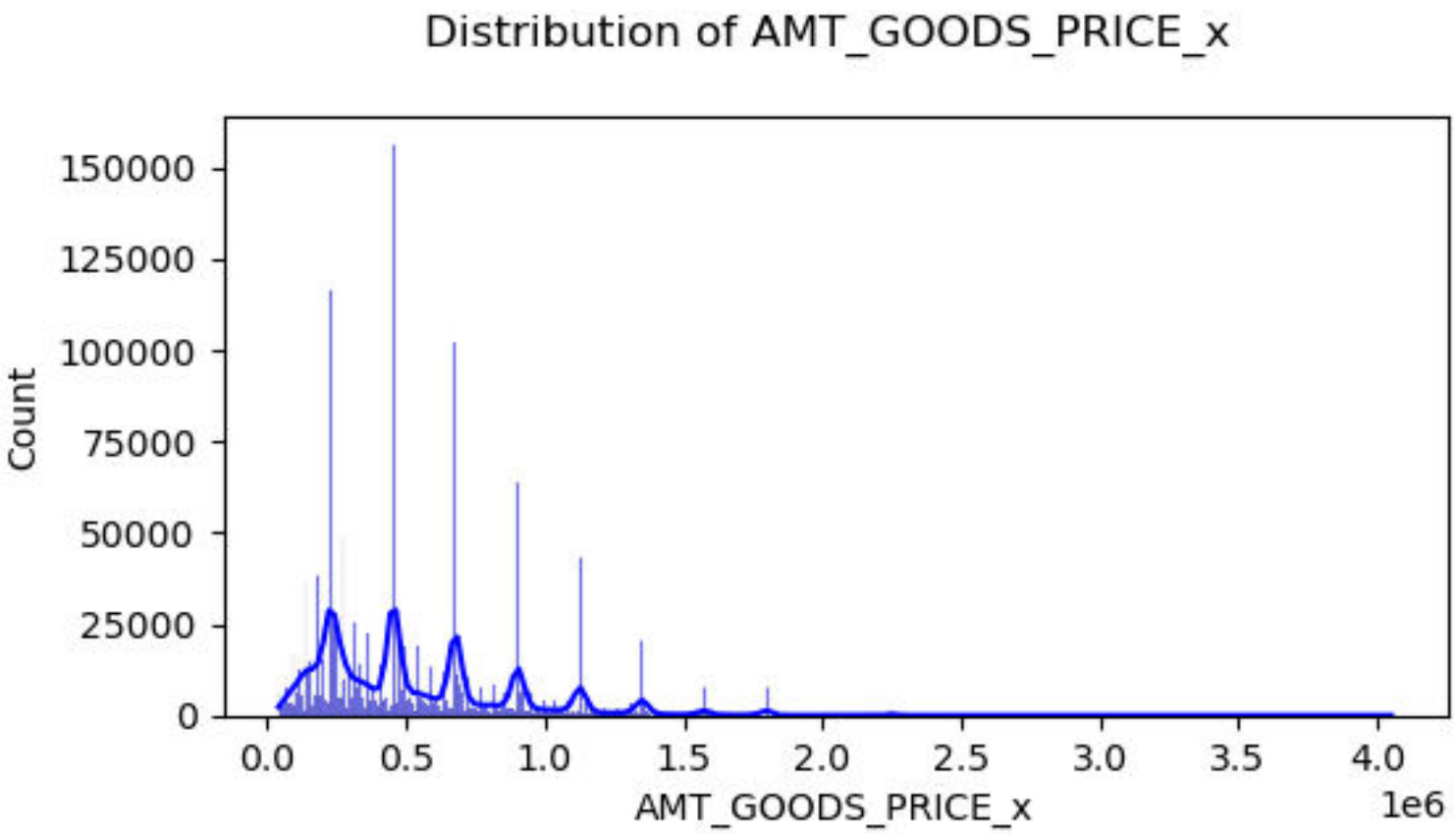
Univariate analysis of merged dataset

univariate analysis of categorical column.



- From the piechart we must say that 90.5% clients are having their contract type as cash loans.
- while on the other hand only 9.5% clients are having their contract type as revolving type.

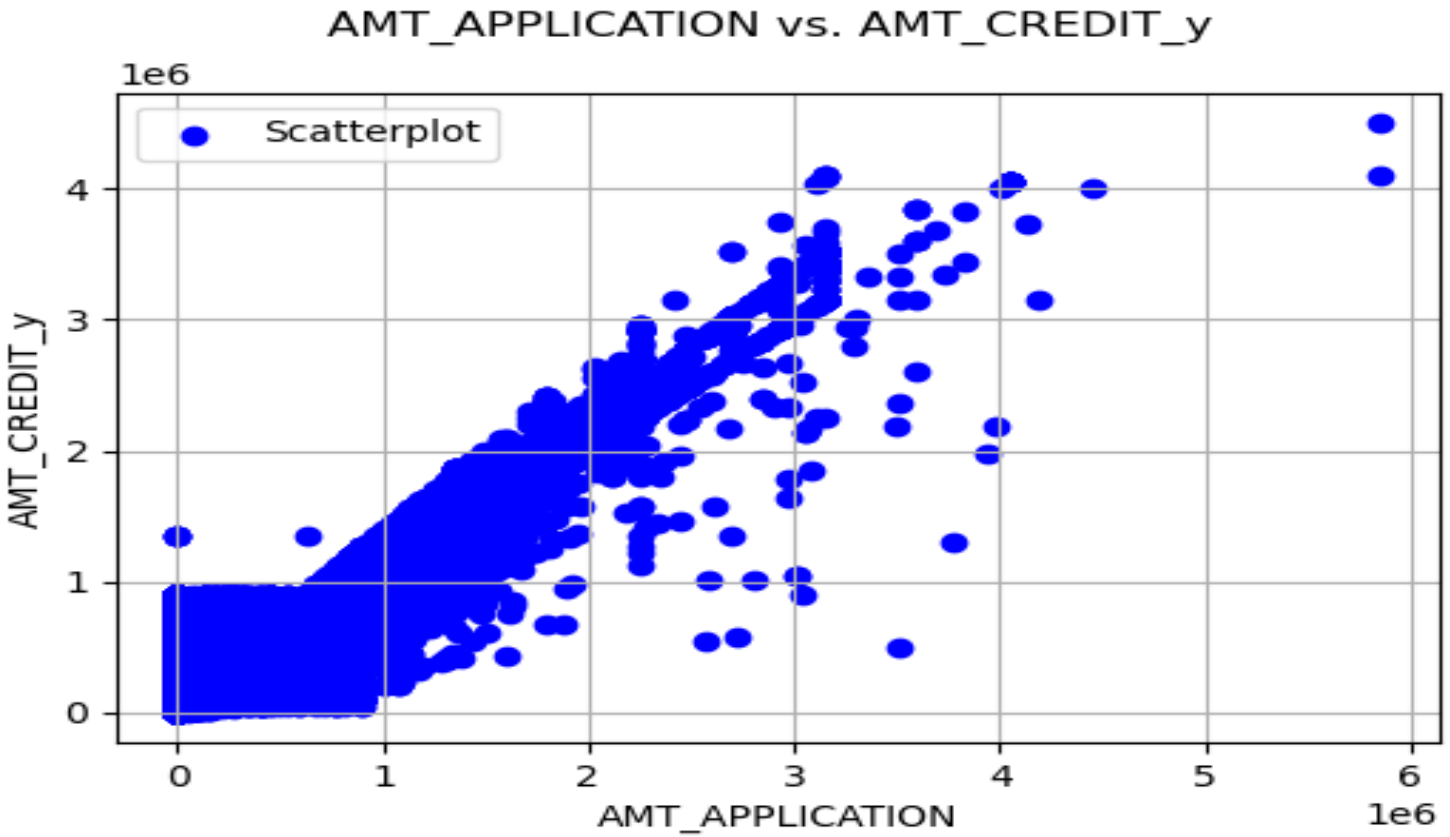
univariate analysis of numerical column



- Most of the goods price asked by clients in previous application is less than 100000.

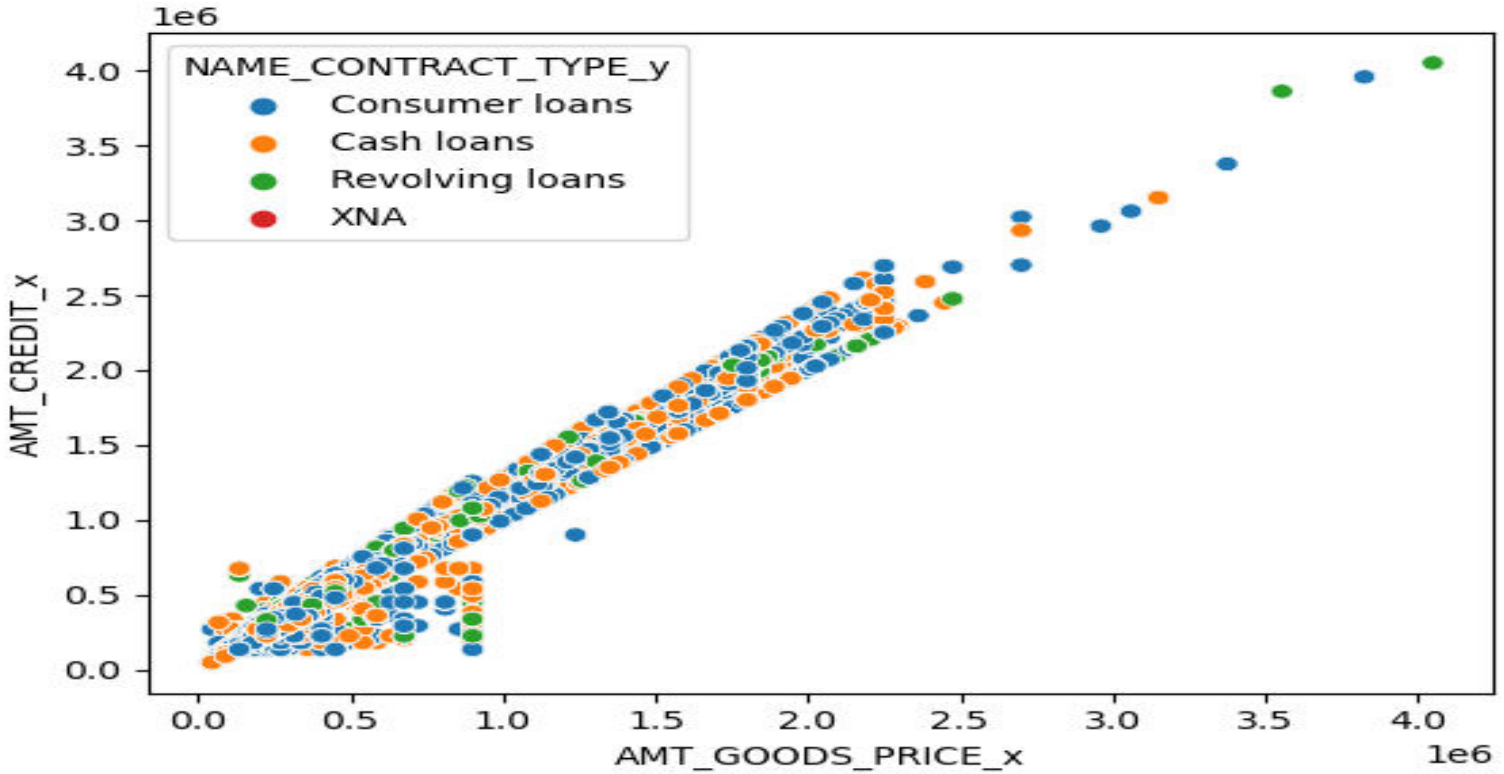
Biveriate / multiveriate analysis of merged dataset

Bivariate analysis of numerical columns



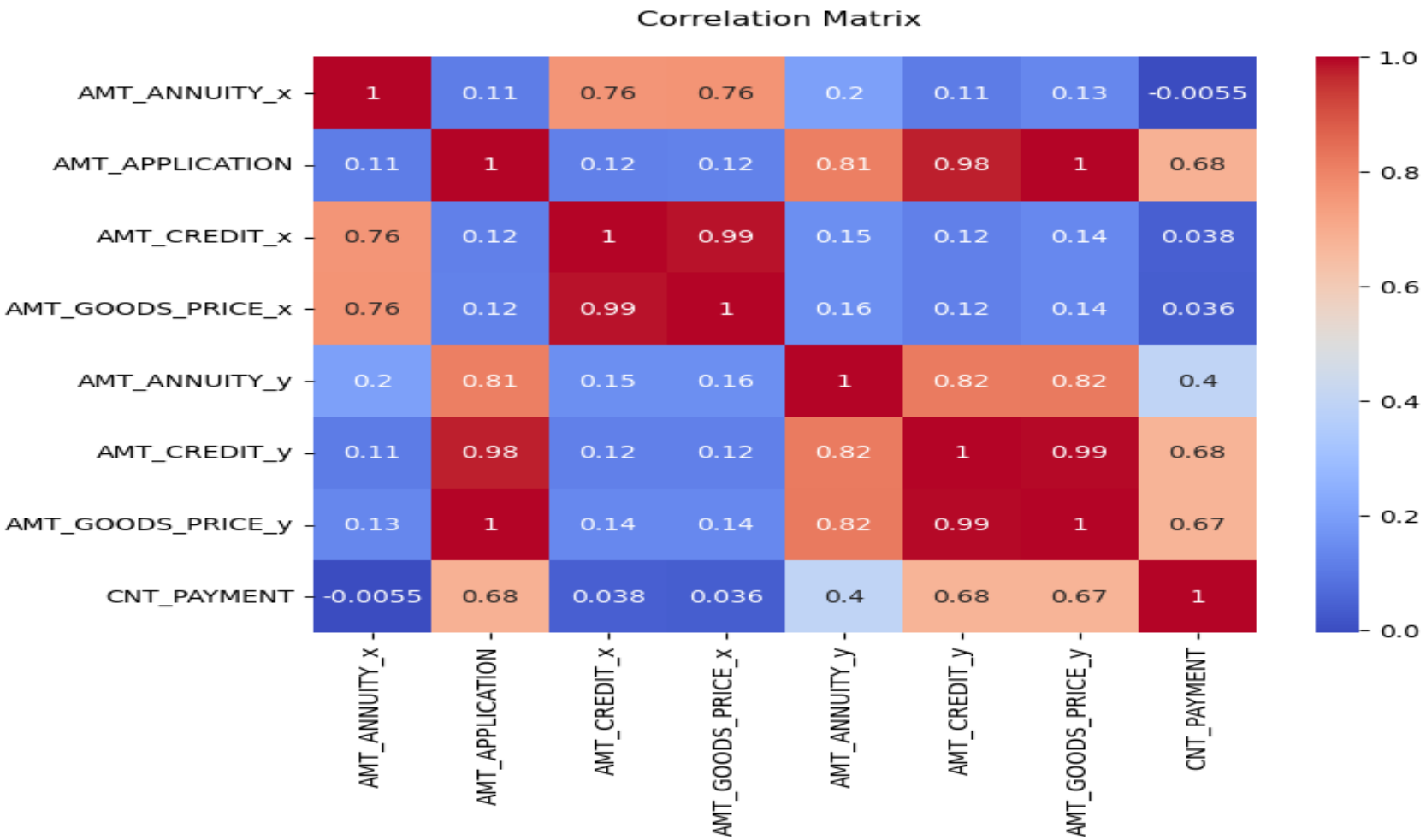
- Application amount has strong positive correlation with Credit amount
- It can be say that application amount increases so the credit amount will also increases.

scatterplot to show the count study of contract types of loans with goods price and credit amount columns.



- from this scatterplot we can find that,
- count of consumer loan is highest in this plot.
- credit amount has strong positive correlation with Goods price amount.
- count of cash loans type contract is very high. whereas count of revolving loans type contract is very low.

Correlation Matrix in merged dataset



- AMT_APPLICATION has a high correlation with AMT_ANNUITY_y, AMT_CREDIT_y.
- AMT_CREDIT_x has a high correlation with AMT_GOODS_PRICE_x.
- AMT_ANNUITY_x has a lowest correlation with CNT_PAYMENT.

CONCLUSION

- Client categories should be targeted to provide loan
- Clients who are employed for more than 19 years
- Clients in the age range 30-40 and 40-50
- Clients who are Married
- Male clients with Academic degree
- Students and Businessman
- Repeater clients.

THANK YOU