CREDIT
EDA CASE STUDY

JITHIN PRAKASH K

ABHISHEK BHATTACHARYA

Introduction

This case study is performed in accordance with the data provided for current and previous loan application status for customers. The case study will help to analyze the factors involved for the customers to pay or not to pay the loan. There are numerous parameters which came into play during this analysis and with the help of different statistical techniques we have tried to get a good amount of insight on the Loan applicant tendency of paying or defaulting a loan.

Since the dataset provided for case study is pretty large , the objective is to consider those variable which has a direct impact on the application of the customer also consider the variables with a correct amount of data.

The whole study has been done on the google colab IDE using Python coding utilizing the numpy, pandas, matplotlib, seaborn libraries.

Approach

The case study has been done by keeping several things in mind which includes several considerations like analyzing the missing values, imputation method with reasons to replace the missing values (but for this EDA the missing data has been kept as it is).

The analysis has been done on the columns/variable which may not be impacting the analysis and can be removed for the sake of improving efficiency of data retrieval. Identifying the variables as Categorical and Numerical and then deducing their correlation is also a part of this study. The analysis also includes several statistical methods like univariate and bivariate analysis along with data computation which includes binning.

Key Steps Involved

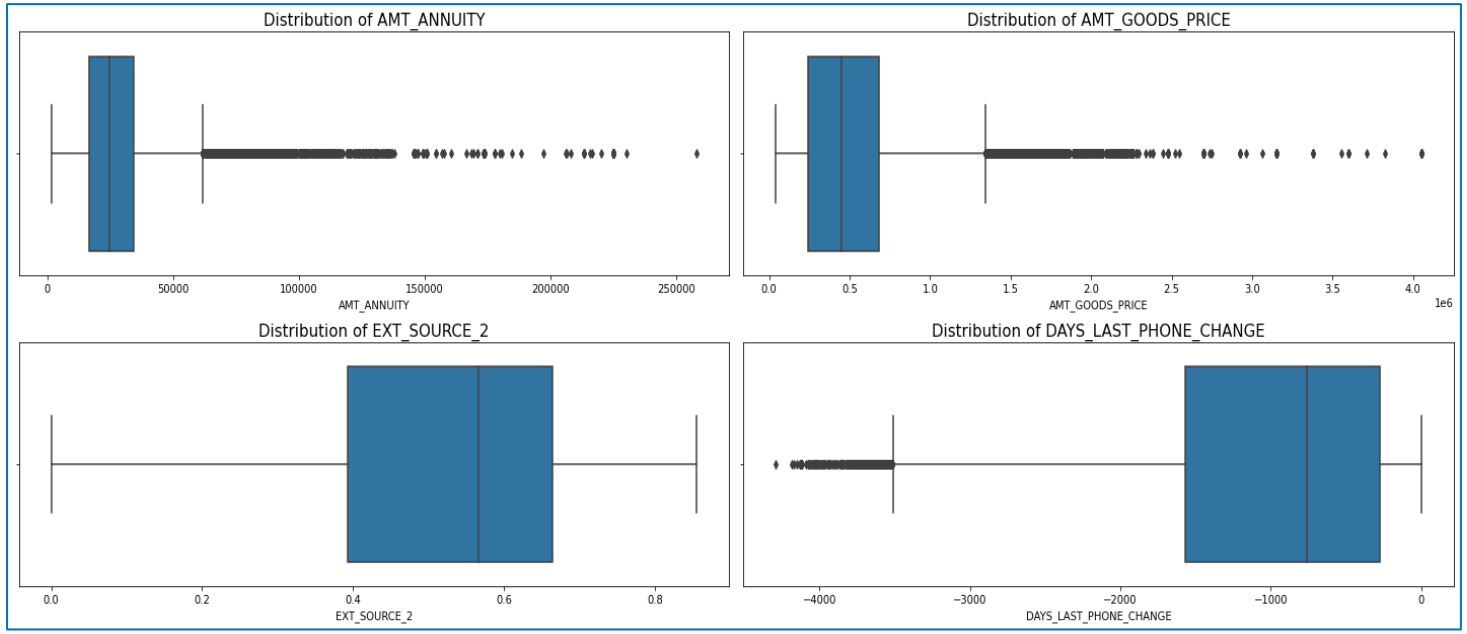
- Reading and loading the data from source file in python pandas data frame.
- Analysing the data-frame to get the information of the columns, datatypes, their statistical values like mean, median, mode and percentiles.
- To remove the columns from analysis having large amount of null values (more than 50%)
- Finding the missing values and providing suggestions to impute the same.
- Data transformation to convert some negative values in positive like no. of days, prices etc.
- Finding out the category and numerical variables.
- Analysing the numeric variables to detect the outliers.
- Performing binning operation on continuous variables
- Getting the correlation between the Target and other variables
- Performing univariate , bivariate and multivariate analysis for different variables to understand the relationship between them.
- Finding the correlation between different variables and their effects on the Target Variable.
- Plotting different charts for better insights and easy understanding.
- Concluding the insights.

Current Application

Distribution of Numerical variables having null values < 15%

Inferences drawn from graphs present

- Loan Annuity, Price of Goods have outliers present
- No. of days before applying the client change the phone have negative values and also have outliers
- External Source columns does not seem to have outliers present – this is normalized data.



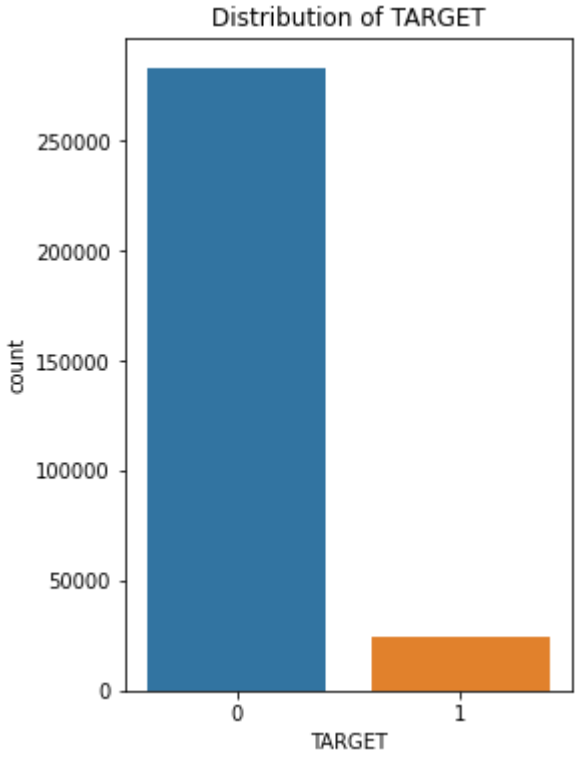
Imputation techniques suggested

- Numeric data - For Null values and outliers can be imputed using median
- Numeric data - For Null values without outliers can be imputed using mean
- Categorical data - The missing values can be imputed using mode categories.

Analysing Imbalance for Target variable

Data Imbalance check is crucial, as this will depend the father prediction methodologies and data insights.

Checking the data imbalance for the TARGET variable.



Inferences

From the graph, it is clear that there are

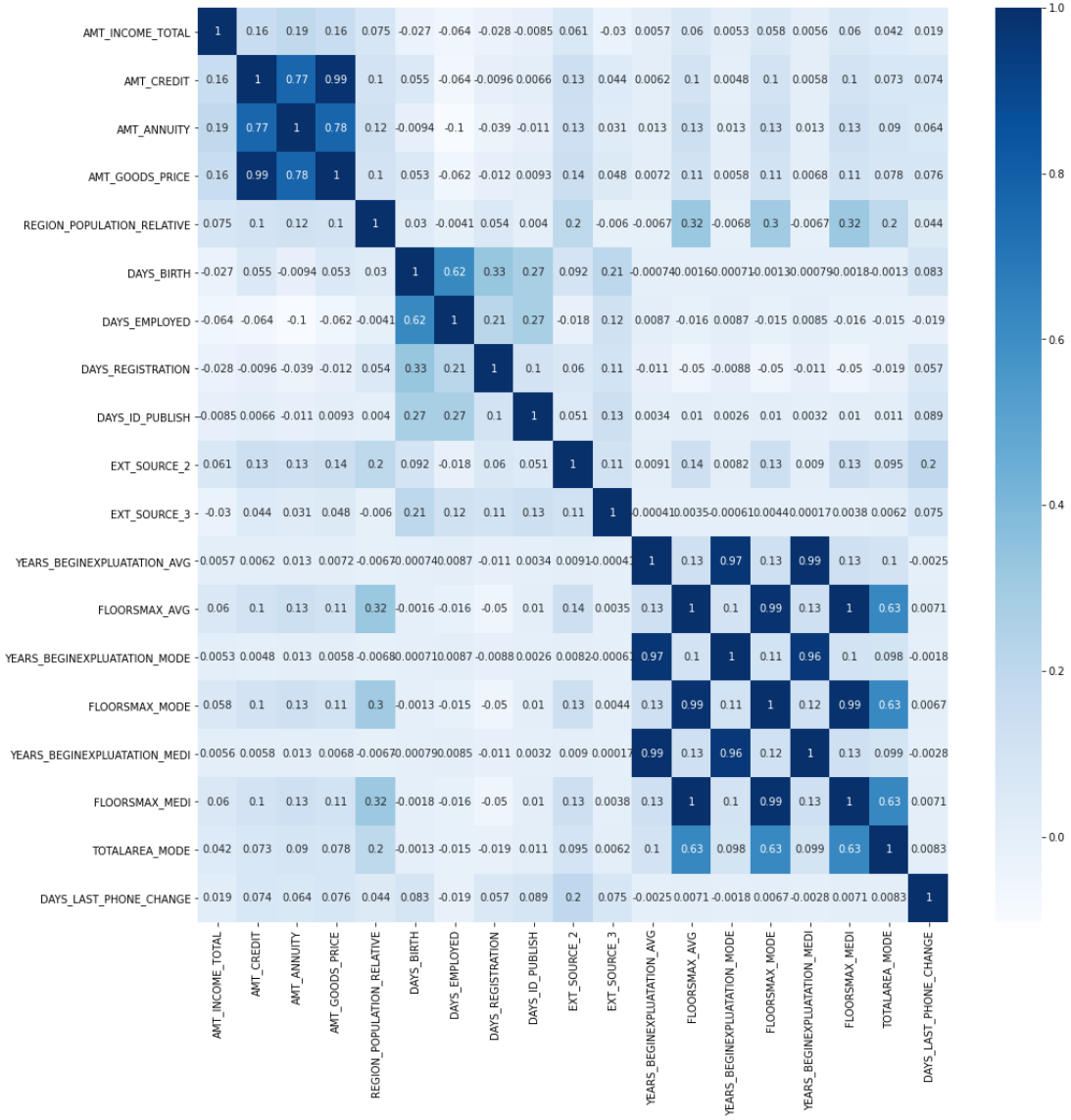
- approximately **91.9%** of values of TARGET as 0
- only **8.1%** as 1

i.e. 91.9% clients with No payment difficulties and there are 8.1% clients with payment difficulties

The difference in the Target variable is huge and hence there is a clear imbalance in the data.

There are far more number of loans repaid on time, than those are defaulted.

Top 10 correlated variables



Inference

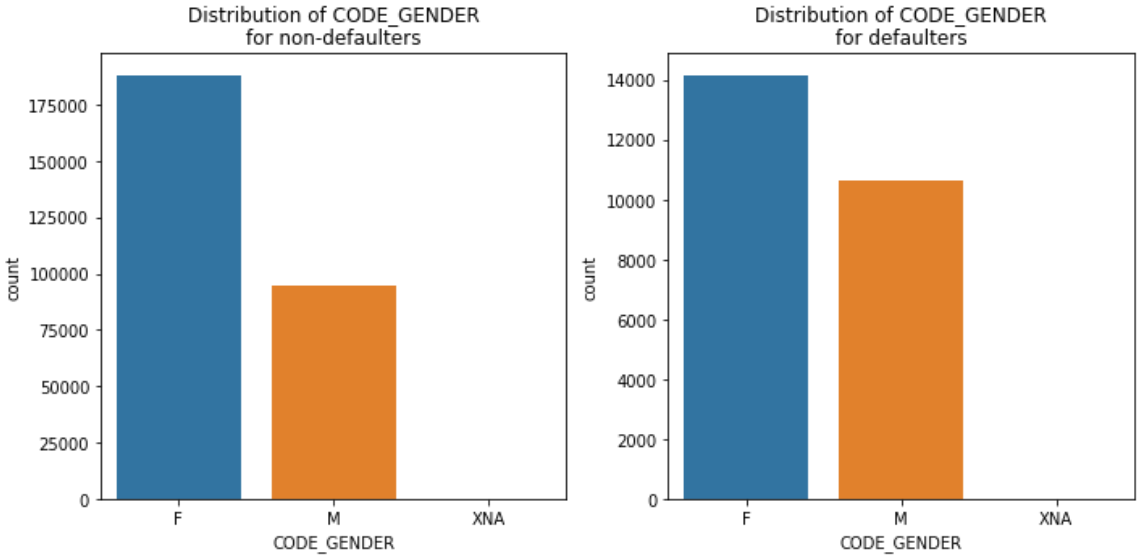
Same analysis is carried out for the defaulter and non-defaulter dataset and found to be similar for top correlations

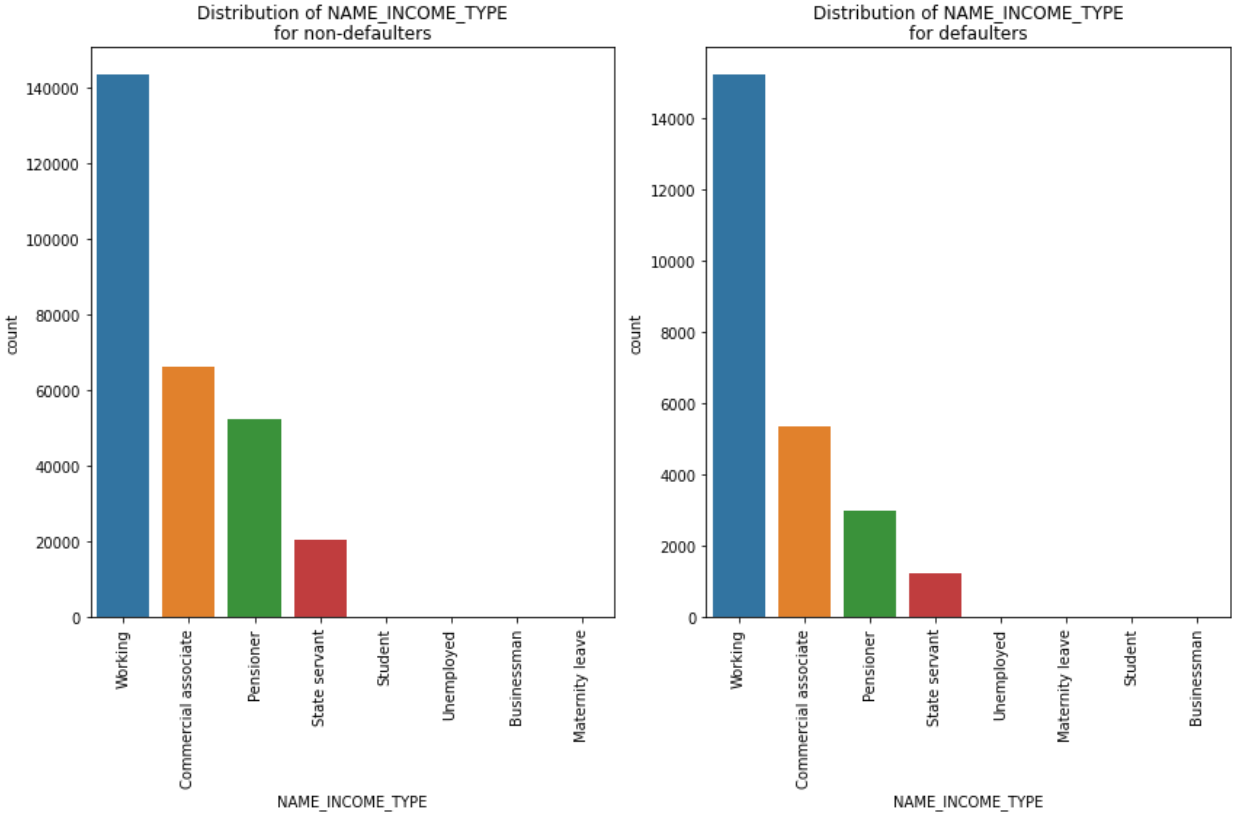
From the heat map it is evident that there is a strong relation exists between the following variables:

ATTRIBUTE_1	ATTRIBUTE_2	CORRELATION
FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997034
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.993825
FLOORSMAX_MEDI	FLOORSMAX_MODE	0.988237
AMT_GOODS_PRICE	AMT_CREDIT	0.986968
FLOORSMAX_MODE	FLOORSMAX_AVG	0.985689
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.971893
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.963539
AMT_GOODS_PRICE	AMT_ANNUITY	0.775109
AMT_ANNUITY	AMT_CREDIT	0.770138
TOTALAREA_MODE	FLOORSMAX_AVG	0.632595

Univariate analysis of categorical variable

Gender , Income Type





Inferences

Females have taken higher number of loans compared to males for both defaulters and non-defaulters

The ration of graph for males are higher for defaulters, i.e. ***Male have higher chance to become defaulters.***

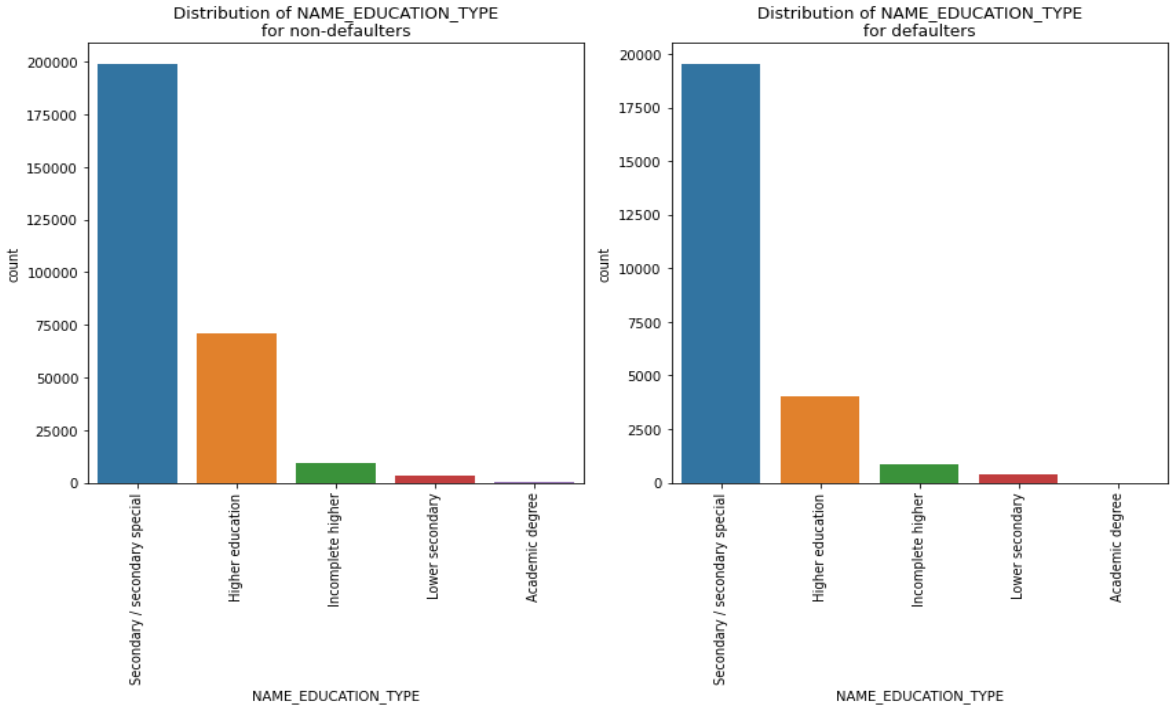
Inferences

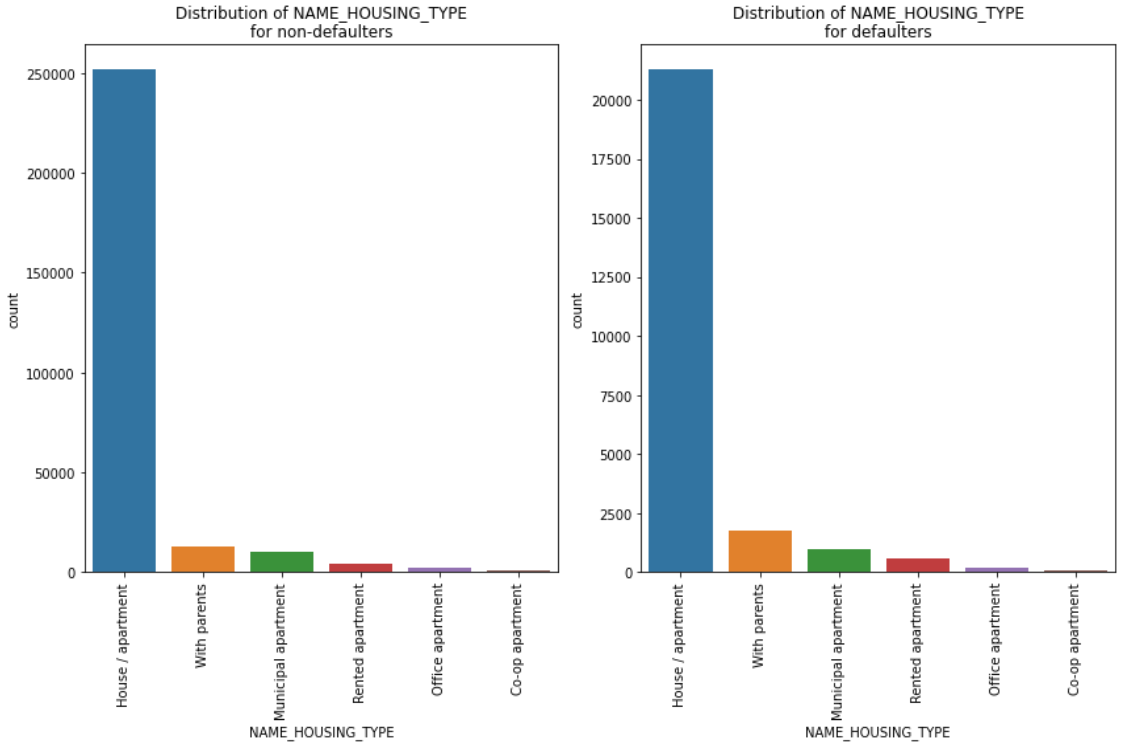
Working type clients availed higher number of loans across defaulters and non-defaulters.

Commercial Associates and pensioners have higher ratio for non-defaulters, i.e. Bank should focus more on Commercial associates and Pensioners

Univariate analysis of categorical variable

Education Type, Housing Type





Inferences

Most number of loans are taken by clients who have completed secondary(special) education.

Loans taken by clients with Higher education shows a high ratio for non-defaulter. So bank should concentrate more on such clients as well.

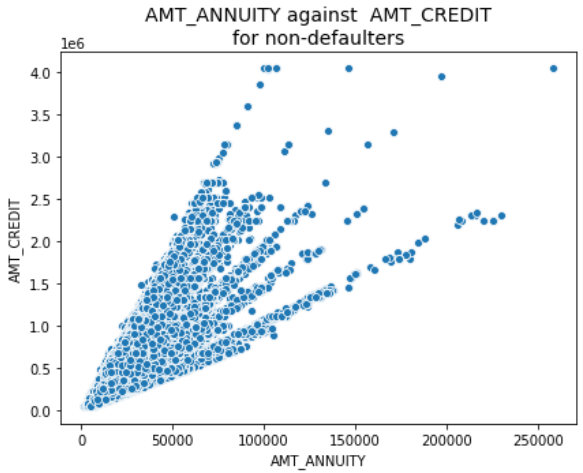
Inferences

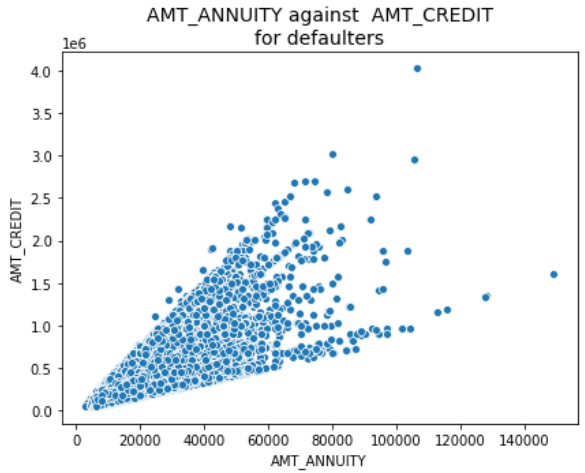
Most clients who availed the loan own a House/Apartment.

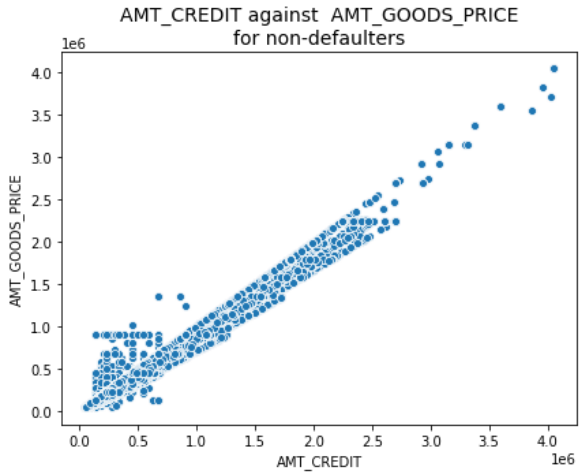
Ratio of People who live ***With Parents*** is ***more for defaulter*** than non-defaulters. Those who live in ***Rented Apartments*** show similar results, so both the categories tend to have higher chance of payment difficulties

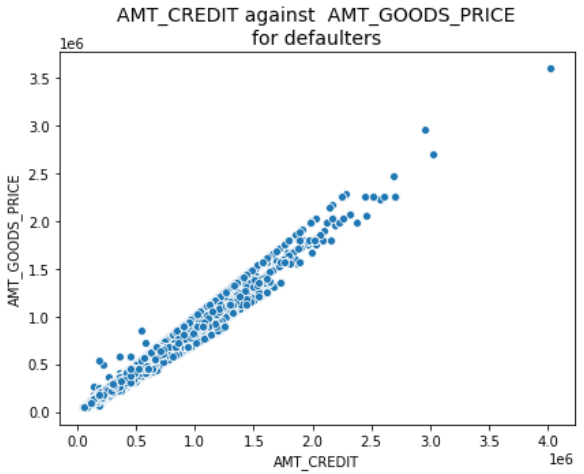
Bivariate analysis – Continuous - Continuous variables

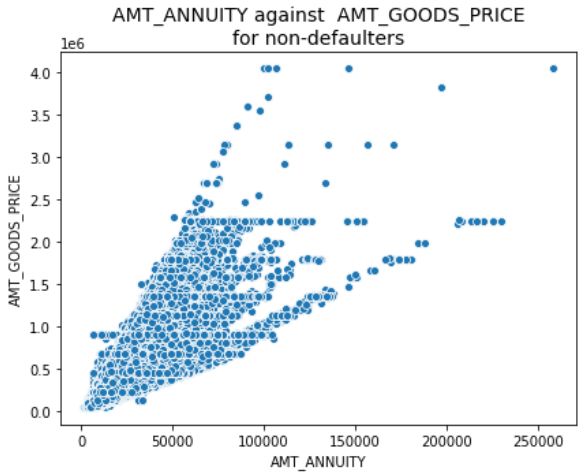
Credit, Annuity, Goods price amounts

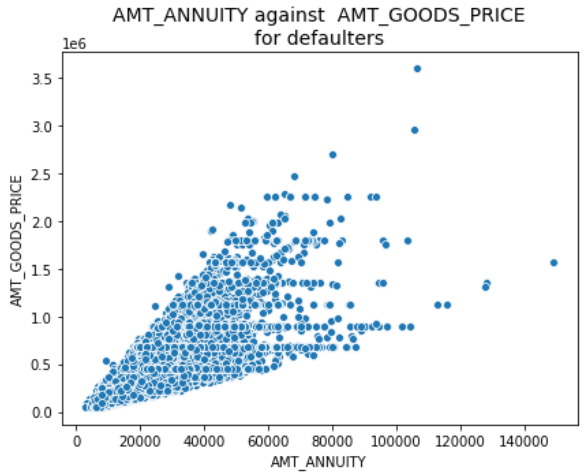










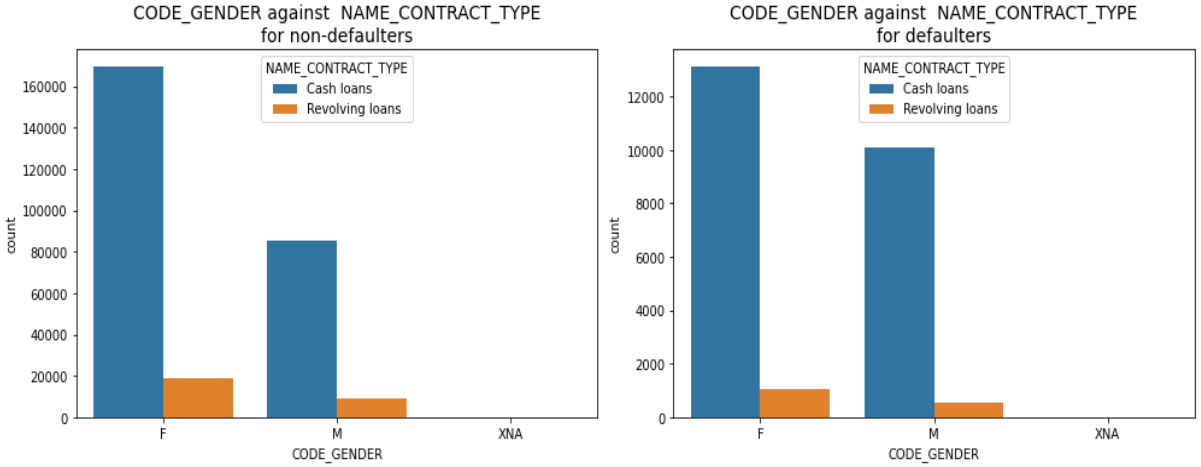


Inferences

- AMT_ANNUITY vs AMT_CREDIT
- AMT_ANNUITY vs AMT_GOODS_PRICE
- AMT_CREDIT vs AMT_GOODS_PRICE

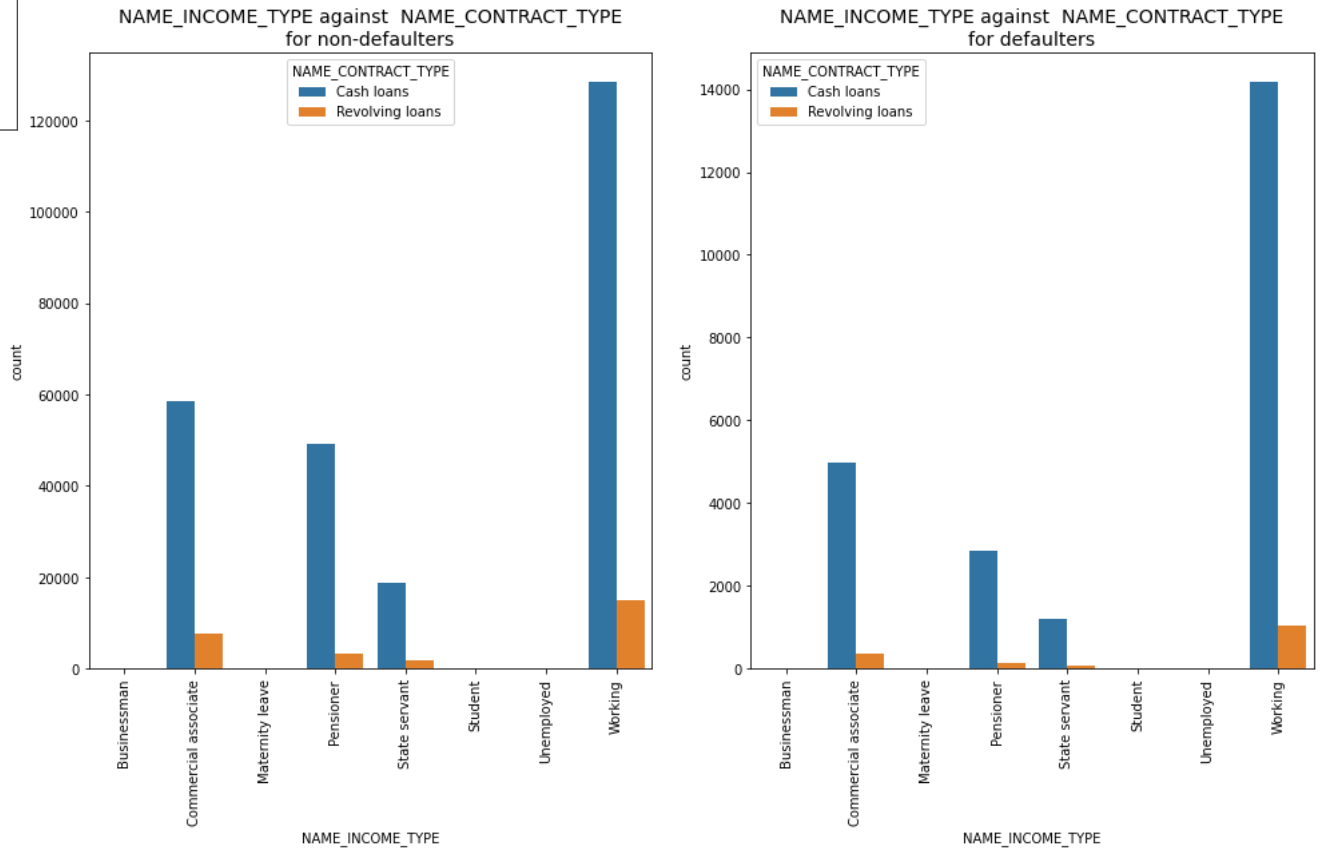
All the three continuous variables shows a high positive linear correlation, i.e. as x variables go higher y variable also is higher. For defaulters the correlation is bit low compared to non-defaulters.

Bivariate analysis – Categorical - Categorical variables



Inferences

For both genders, cash loans are of higher number for defaulters and non-defaulters. Women who avail revolving loans have higher non-defaulters count. ***Ratio of cash loan defaulters are higher in Males.***
Men who avail cash loan have a higher chance of payment difficulties

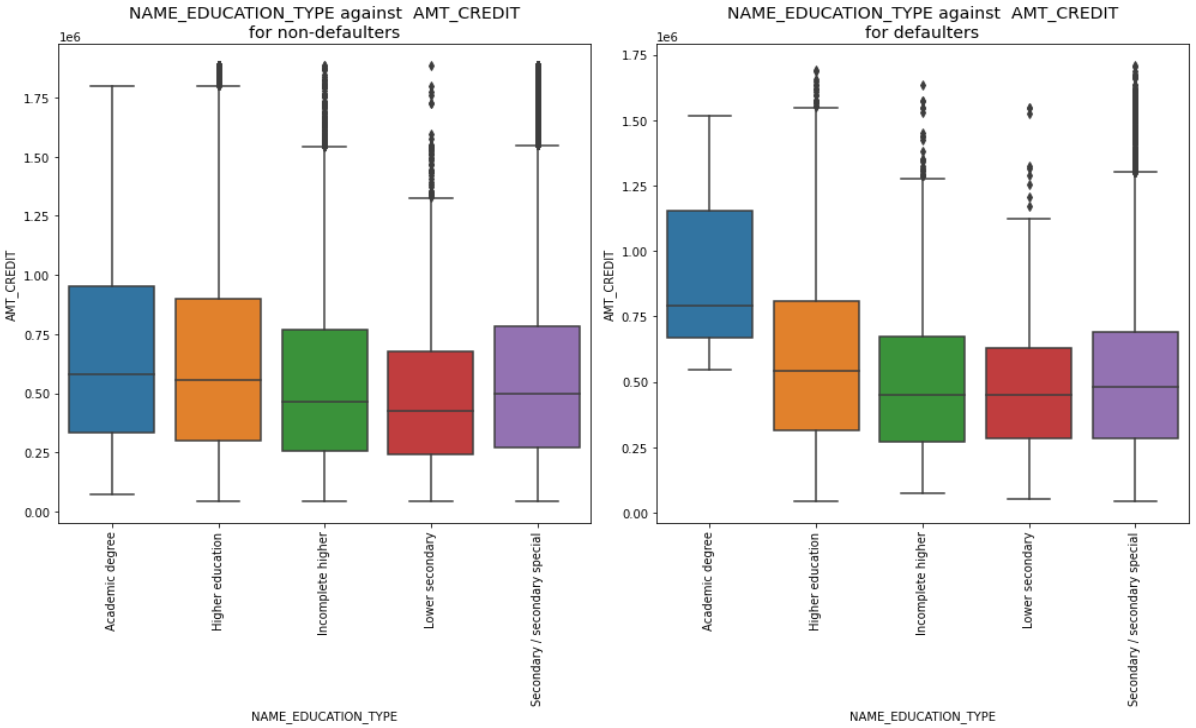


Inferences

The Ratio of loan defaulter is low for both cash and revolving loans for commercial associate, Pensioners and state servants, but for working people it shows same trend for both defaulters and non-defaulters.

Working people have a higher default rate for cash loans, where as pensioners and commercial associates are largely non-defaulters for cash loans and revolving loans

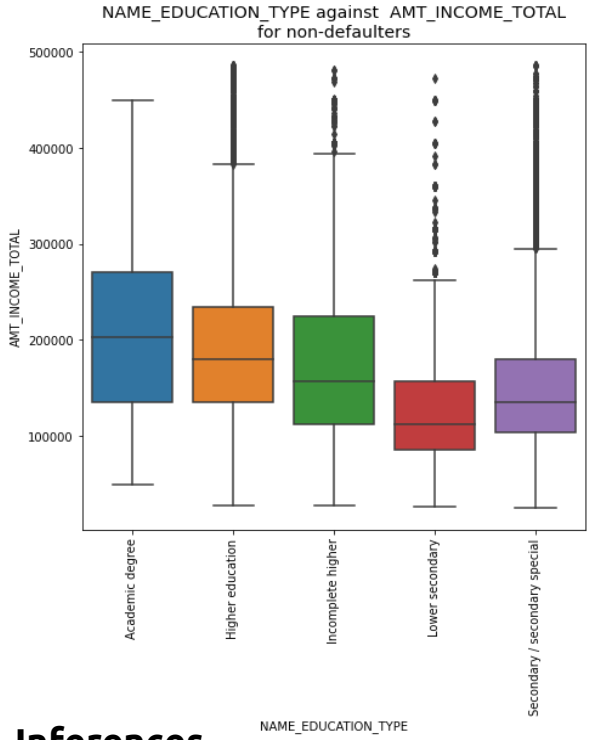
Bivariate analysis – Continuous - Categorical variables…1/2

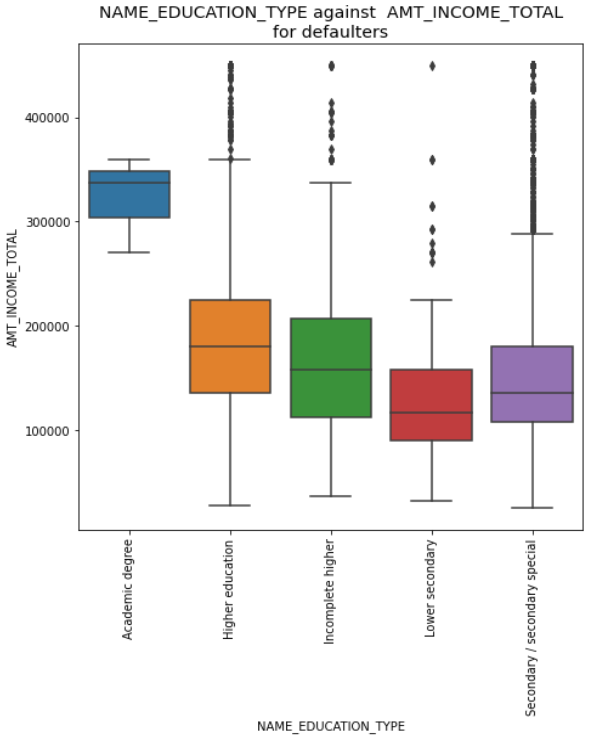


Inferences

All the Education types except Academic degree shows a similar trend of Credit amount for both defaulters and non-defaulters. ***For Academic degree, spread of credit amount is more for non-defaulters and normal from median which is around 60000,*** for defaulters the spread is short and skewed, the median is higher at ~800000 approx.

Clients with academic degree have defaulters when the credit amount is higher.



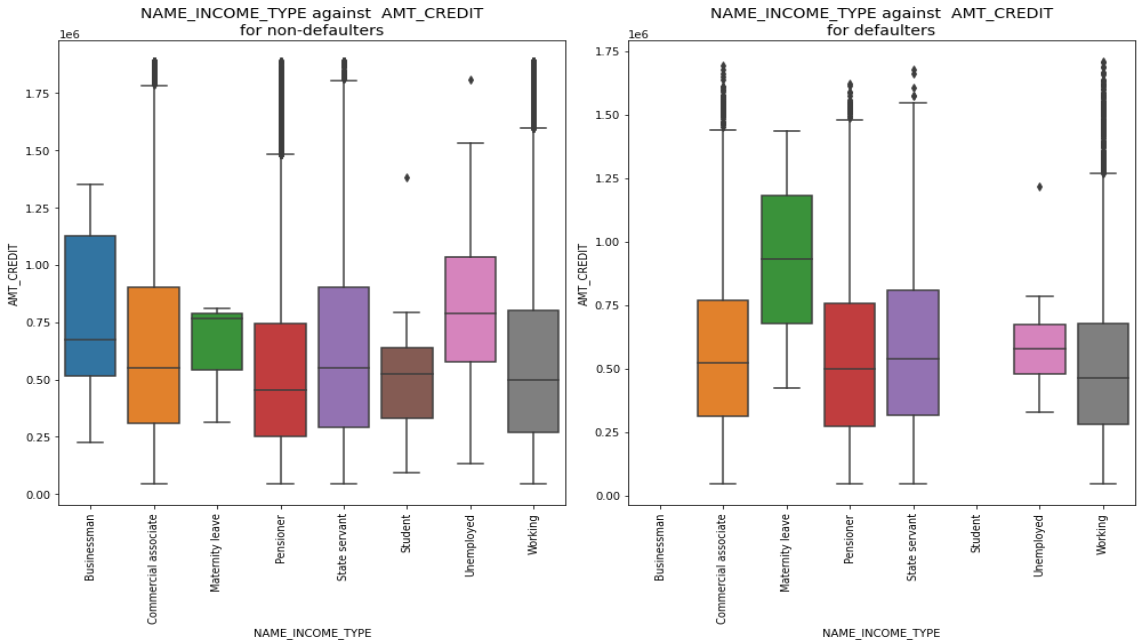


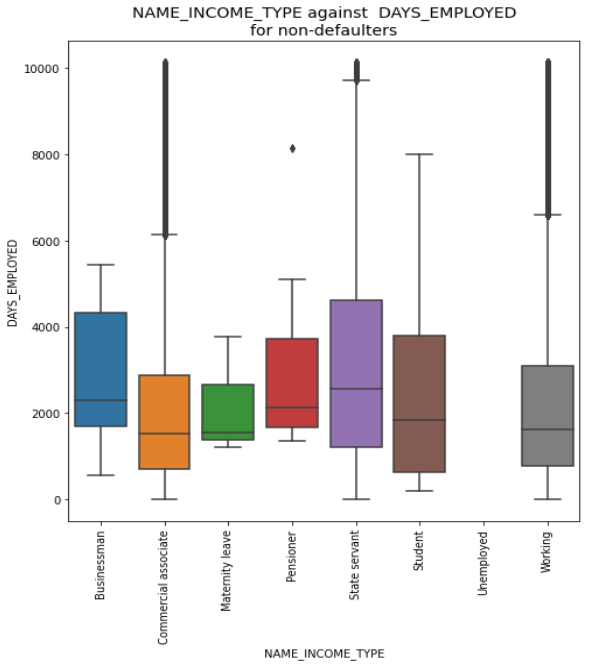
Inferences

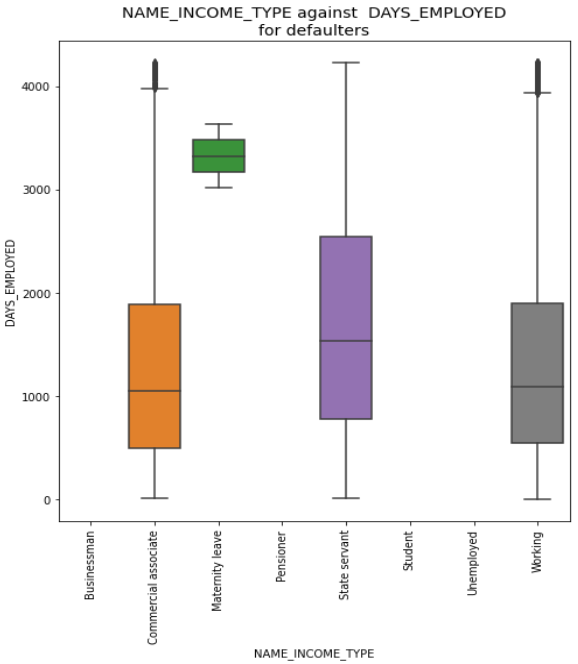
All the Education types except Academic degree shows a similar trend of income for both defaulters and non-defaulters. The median values for these education types are a bit lower for the defaulters. For Academic degree, spread of income is more for non-defaulters and normal from median which is around 200000, for defaulters the spread is short and the median lies much higher at ~340000 approx.

Clients with academic degree have defaulters where the income is higher, however the spread of non-defaulters are higher, so bank has to concentrate on such groups .

Bivariate analysis – Continuous - Categorical variables…2/2







Inferences

For **Businessman and students**, the credit amount is not available in the defaulter dataset. ***Bank has to concentrate more on such groups.***

Maternity leave shows a uniform normal distribution from median for defaulters whereas it is skewed for non-defaulters. Also with a higher Credit Amount, Maternity leave types tend to have higher defaulters. Median lies much higher for maternity leaves in case of defaulters.

Inferences

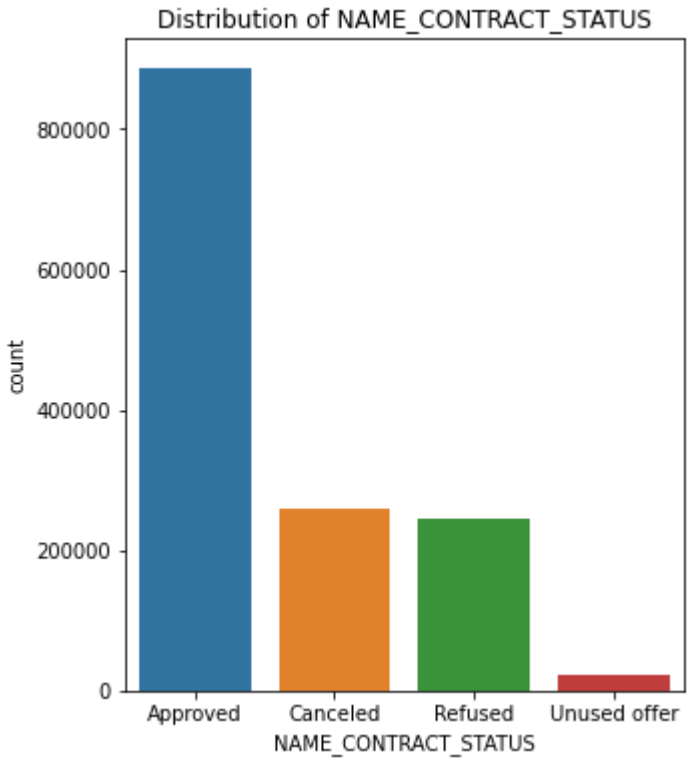
For the Businessman, Pensioner and Students there are no defaulter values for days employed

For Maternity leave, the days employed is higher for defaulters and the distribution is normal from median. Median lies much higher at above 3000days.

For other categories of income type, defaulter median lie lower than that of non-defaulters, i.e. median employed date are lower for defaulters i.e. at the start of their careers

Previous Application

Outcome of previous applications is NAME CONTRACT STATUS



Inferences drawn from graphs present

From the above analysis, it is clear that

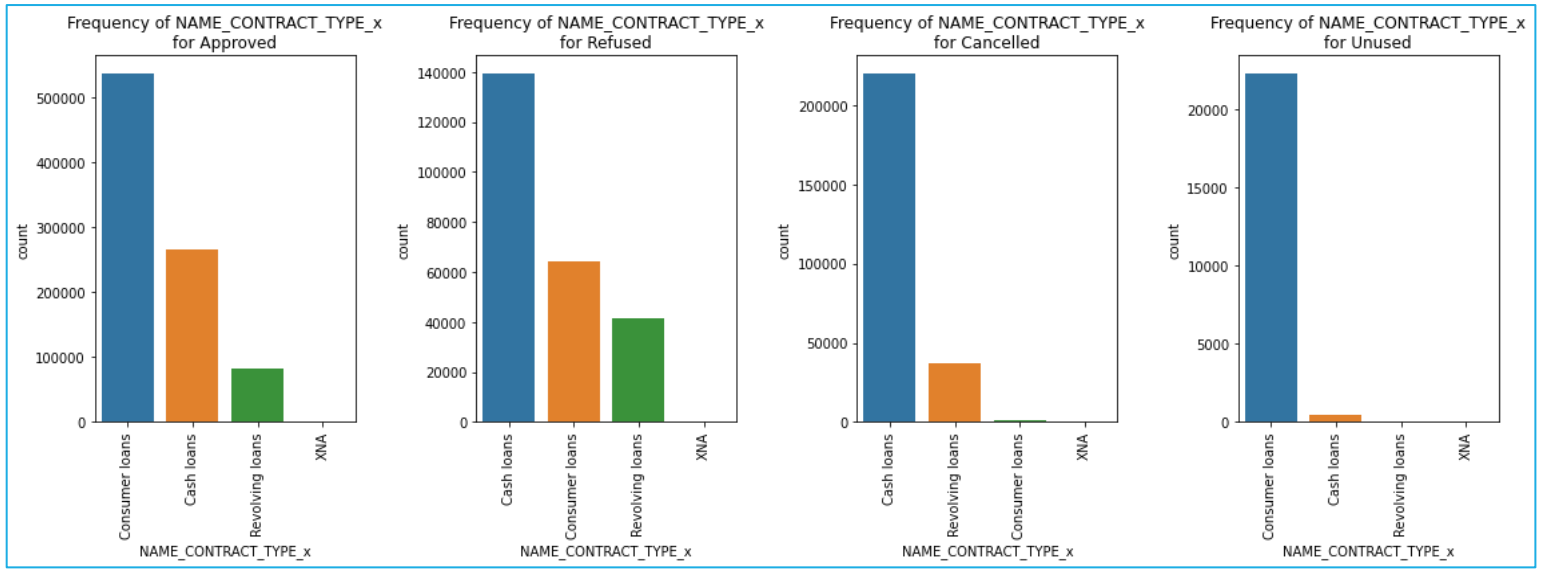
- 62.7% loans are Approved
- 18.4% loans are Cancelled
- 17.4% loans are Refused and
- 1.6% loans are Unused

The loan approval rate looks good as per the graph for previous applications made by the clients

The amount of loans cancelled and refused are almost present in equal ratio

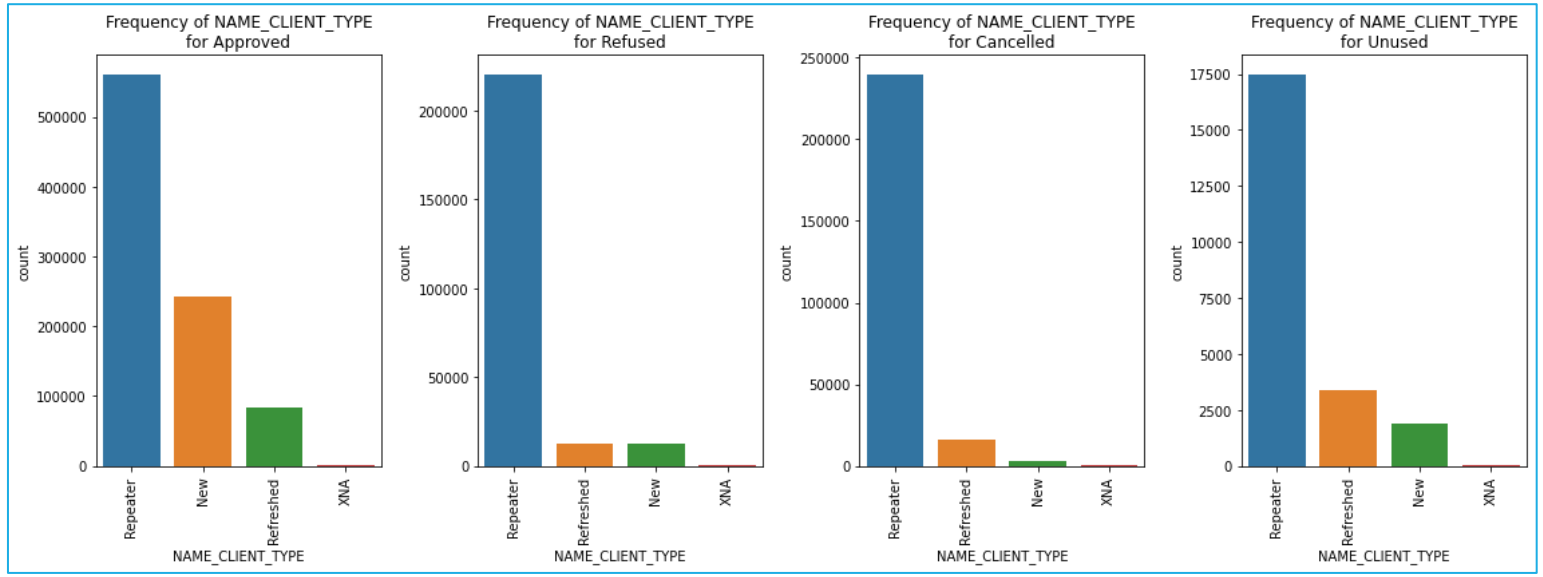
There is a very less percentage of loans which went unused.

Univariate analysis of categorical variable



Inferences

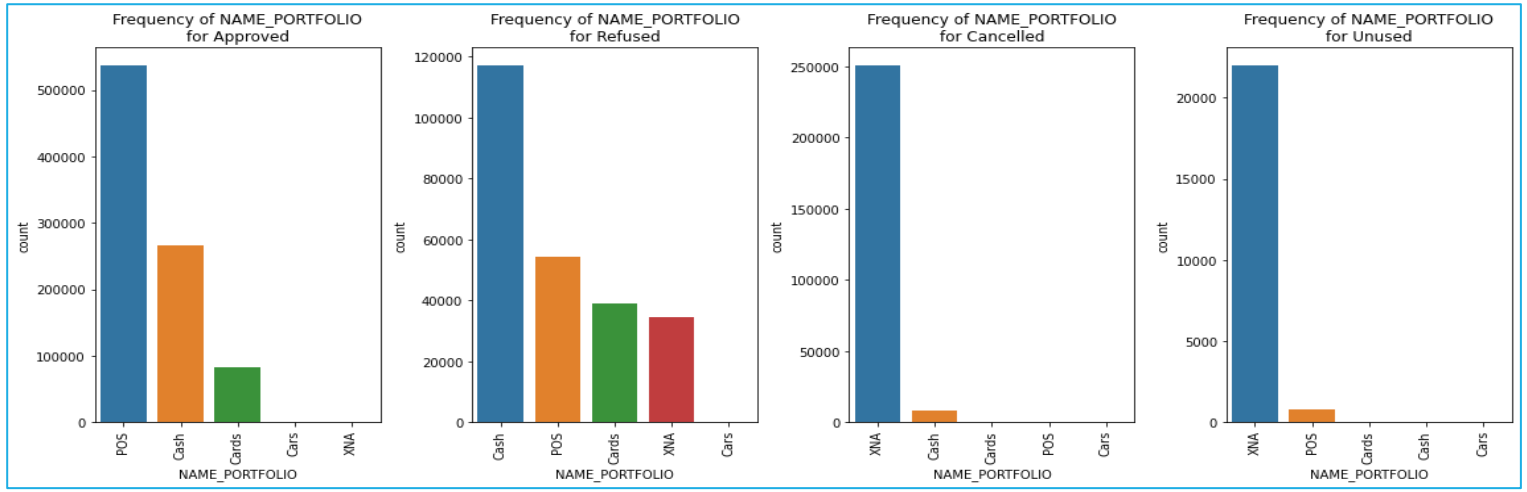
- Consumer loans are higher in number for Approved and Unused loans.
- Cash loans are more for cancelled and refused loan status types.
- Consumer loans are very less in number for cancelled loans compared to other 3 categories.



Inferences

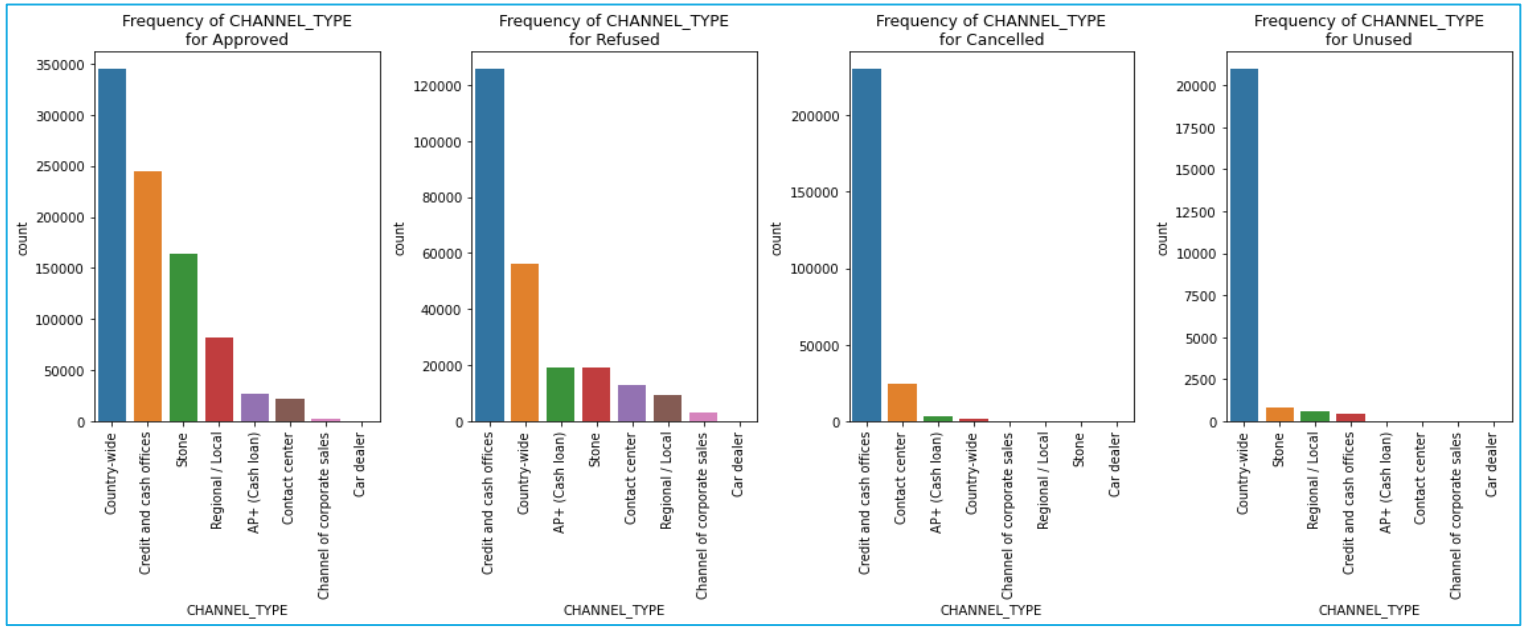
- In all the loan status types, the highest number of loans are taken by client who is a repeater, i.e. the clients do reapply for loans most of the time.
- New clients have a higher chance of loans approved as per the previous data.
- Refreshed loans are higher compared to new loans in case of cancelled loans and unused loans.
- For refused loans refreshed loans and new loans are approximately equal in number

Univariate analysis of categorical variable



Inferences

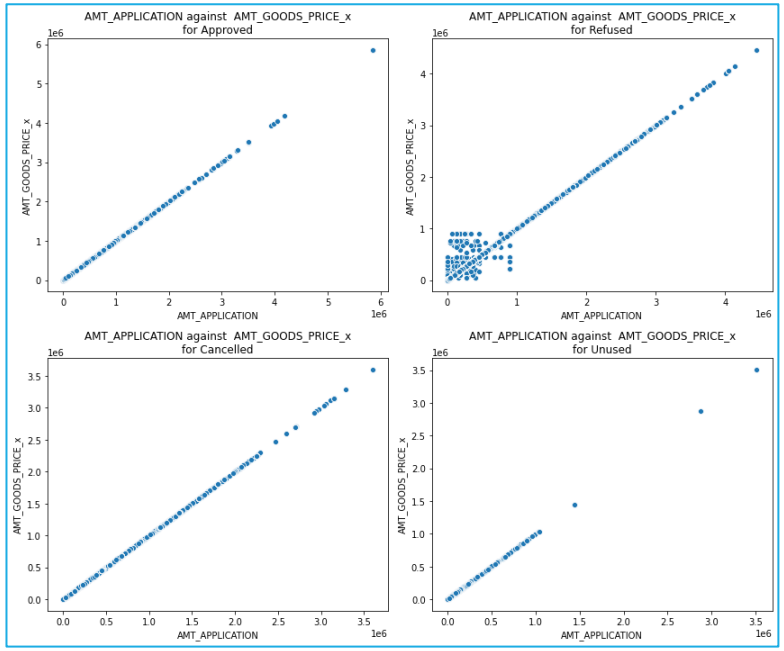
- Number of loans approved are higher in case of POS.
- More number of refused loans are Cash portfolio type.
- All the types show very low or nil quantity for 'cars' portfolio
- For Cancelled and Unused, Unknown (XNA) portfolios have highest count.

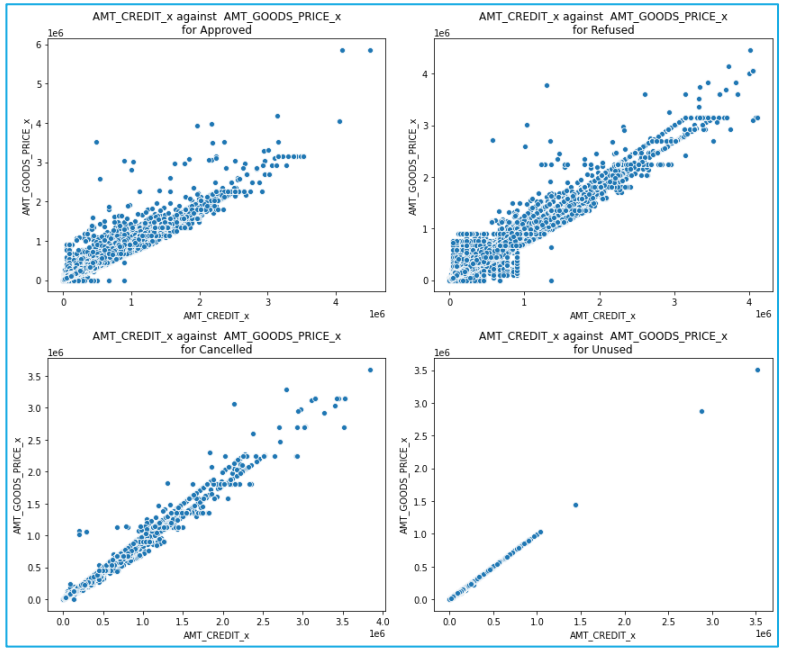


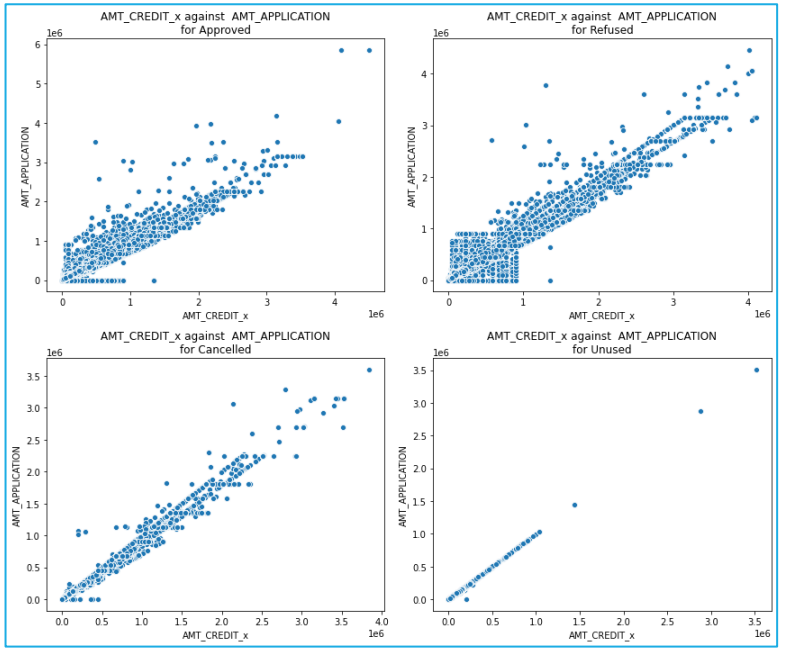
Inferences

- Most number of loans are from Country-wide channel for Approved and unused loan types.
- Most number of cancelled and refused loans are in the credit and cash offices

Bivariate analysis – Continuous - Continuous variables







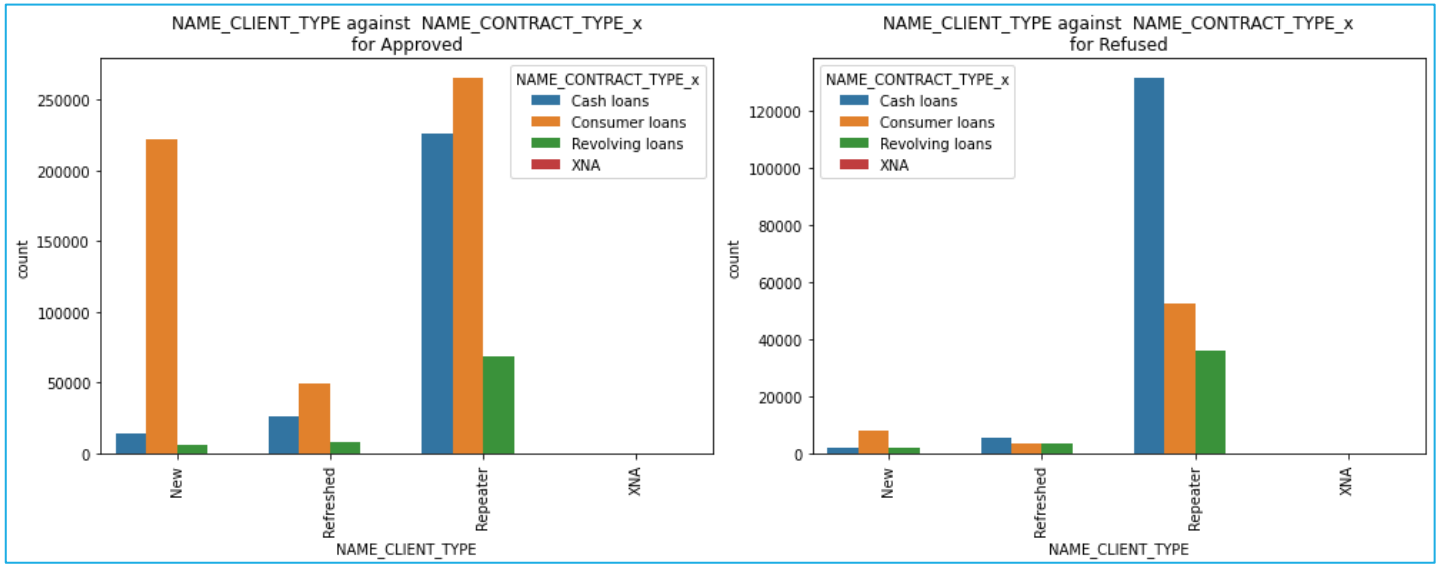
Inferences

AMT_APPLICATION and AMT_GOODS_PRICE shows a very high positive linear correlation for all the loan types. For refused loan types, there is a spread for the lower values of the AMT_APPLICATION and AMT_GOODS_PRICE

AMT_CREDIT and AMT_GOODS_PRICE shows a very high positive linear correlation for all the loan types. The amount for CREDIT and GOODS PRICE is higher for refused types compared to the approved loan types.

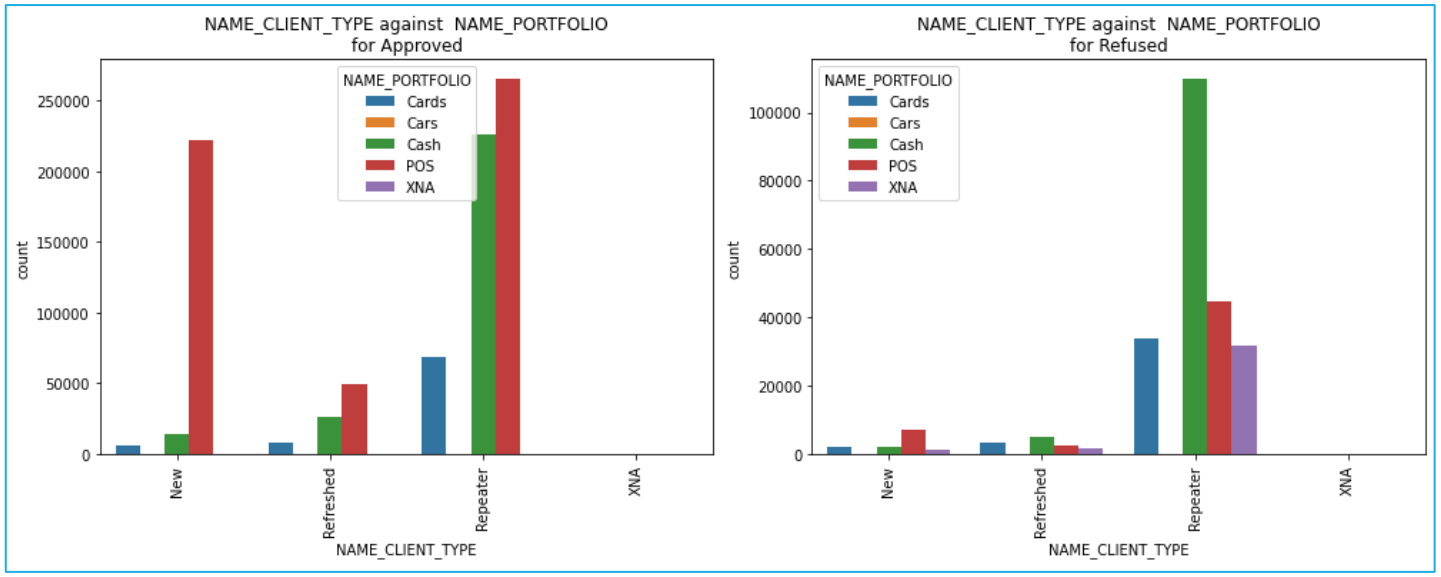
AMT_CREDIT and AMT_APPLICATION also shows a very high positive linear correlation for all the loan status types. The amount for CREDIT and APPLICATION is higher for refused types compared to the approved loan types.

Bivariate analysis – Categorical - Categorical variables …1/2



Inferences

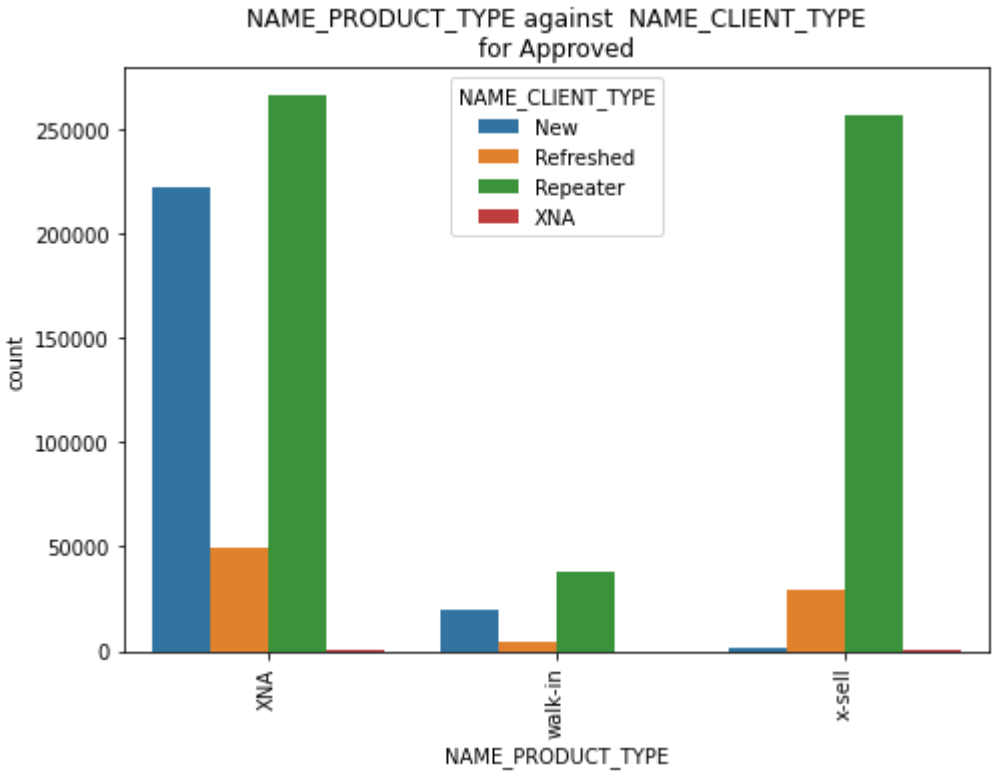
- Approved : New loans are much higher for consumer loans. For refreshed and repeated also consumer loans have higher numbers compared to cash and revolving loans.
- Refused : Cash loans are much higher for repeaters compared to other contract types

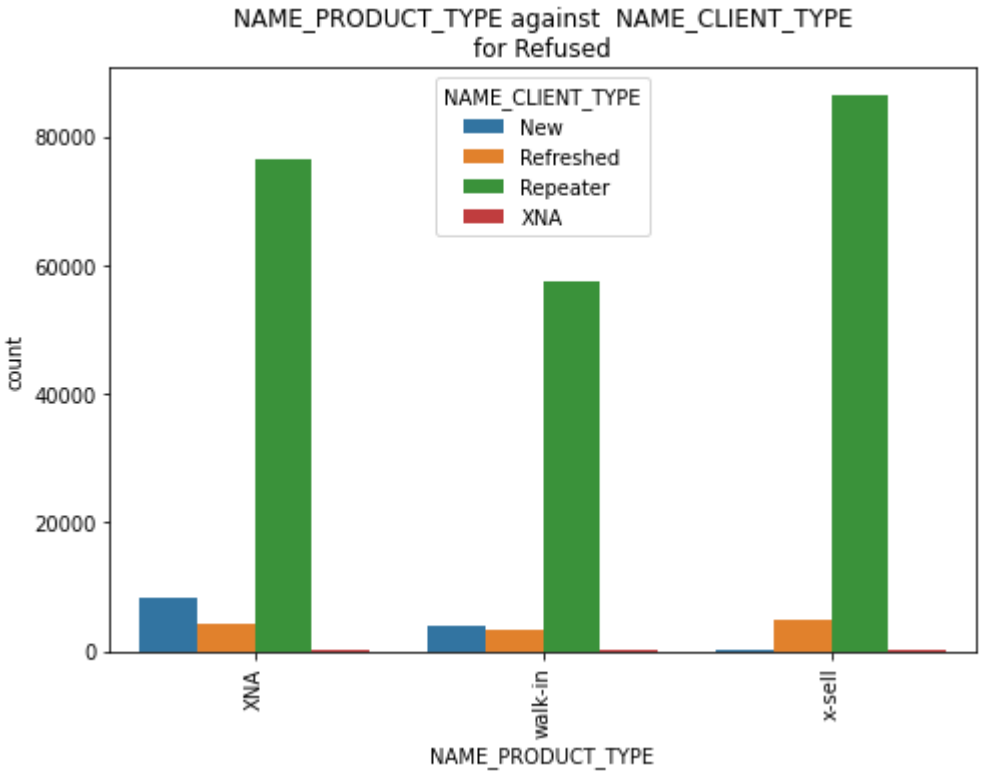


Inferences

- For Approved loan type, New loans are much higher for POS portfolio. For refreshed and repeated also POS portfolio has higher numbers compared to cash and cards portfolio.
- For Refused loan type, Cash portfolio are much higher for repeaters compared to other portfolios.

Bivariate analysis – Categorical - Categorical variables…2/2

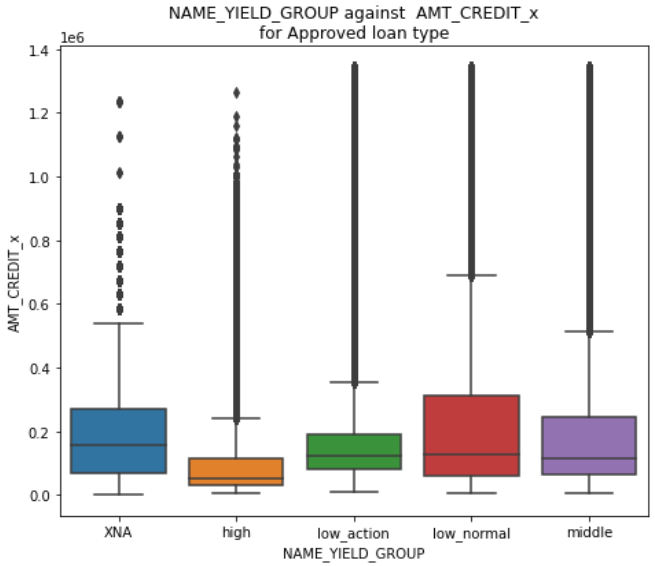


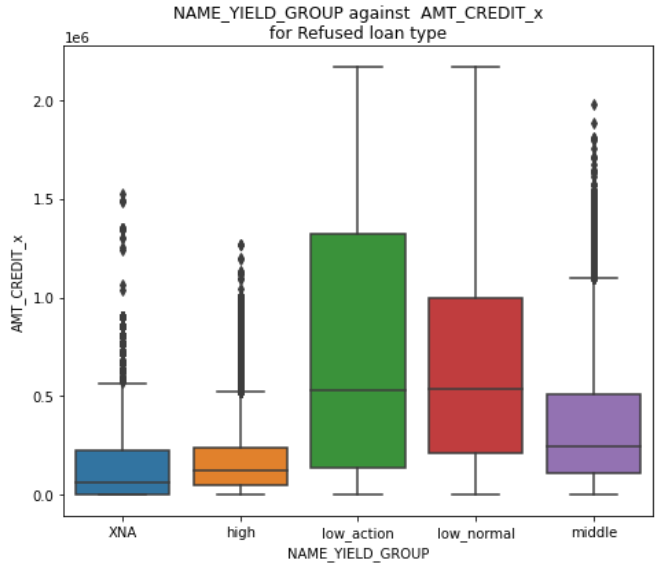


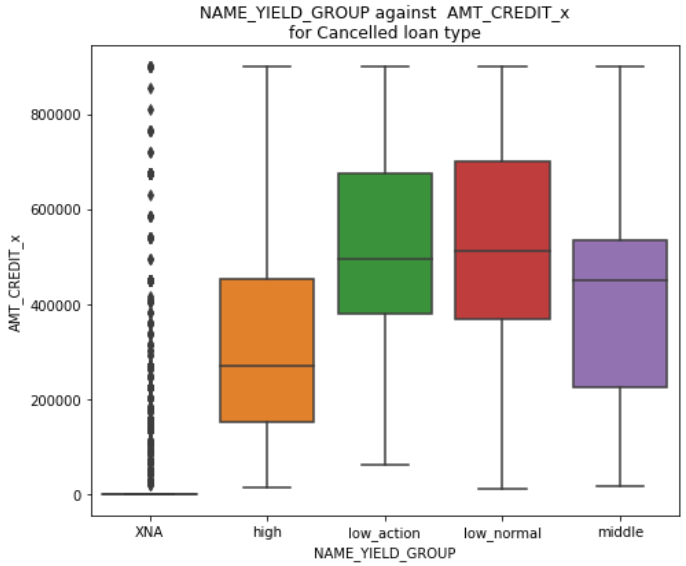
Inferences

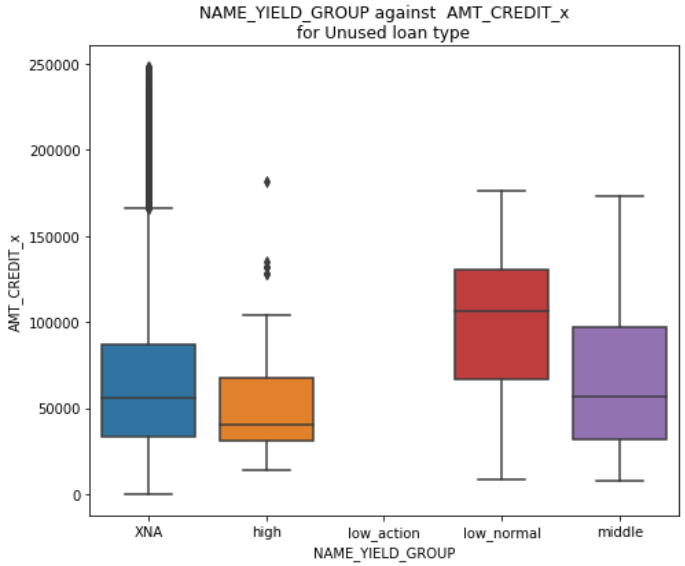
- For Approved loan type, Repeater client type has higher number of loans for Unknown (XNA - Null) and x-sell Product types.
- For Refused loan type, Repeater client type has higher number of loans for x-sell Product types, also Repeater client type has higher loans for other product types as well.
- Walk-in - repeaters have a high tendency of loans getting refused and New - Unknowns have a higher chance of approved loans..

Bivariate analysis – Continuous - Categorical variables









Inferences

- For Approved loan type, low-normal yield group has wider spread for the credit amount. All the yield groups have outliers towards the higher side. The median for high is lower at around 40k. median for Unknown (XNA) is highest at 150k.

- For Refused loan type, low-action yield group has wider spread for the credit amount. Unknown(XNA), high and middle yield groups have outliers towards the higher side. The median for Unknown (XNA) is lower at around 80k. Median for low-action and low-normal yield group is highest at 500k.

- For Cancelled loan type, Unknown (XNA) yield groups shows lowest spread for the credit amount. Unknown (XNA) has more data at 0 and outliers for rest of the amount towards the higher side. The median for Unknown (XNA) is lower at around 0. Median for low-normal yield groups is highest at 480k.

- For Unused loan type, low-action and middle yield groups shows wider spread for the credit amount, however low-normal has a highest value for the median at around 110k. High yield group has the lowest spread and lowest median at around 45k. low-action do not have any data for unused loan type. Unknown (XNA) and high yield groups have outliers towards the higher application amount value.

Conclusion

- Current application data is imbalanced - There are far more number (92%) of loans repaid on time, than those (8%) are defaulted.

Bank should try to limit

- Ratio of males are higher for defaulters, especially Men who avail cash loan have a higher chance of payment difficulties
- Working people have a higher default rate for cash loans
- Clients living with parents and in Rented Apartments tend to have higher chance of payment difficulties

defaulters are more in such groups

Bank should focus more on

- Businessman and students.
- Commercial associates and Pensioners.
- Clients with higher education.
- Clients with academic degree have defaulters when the credit amount and income is higher, bank should concentrate more on clients with academic degree with nominal credit amount, when amount goes higher chances are more for defaulting.

defaulters are less in such groups

Thank you