EDA Assignment on a Bank Data

AGENDA

Problem statement.

Exploratory data analysis (EDA) approach.

Univariate analysis results.

Bivariate analysis result.

Summary and recommendation.

PROBLEM STATEMENT

To identify patterns which indicate if a client is capable for a loan or not.

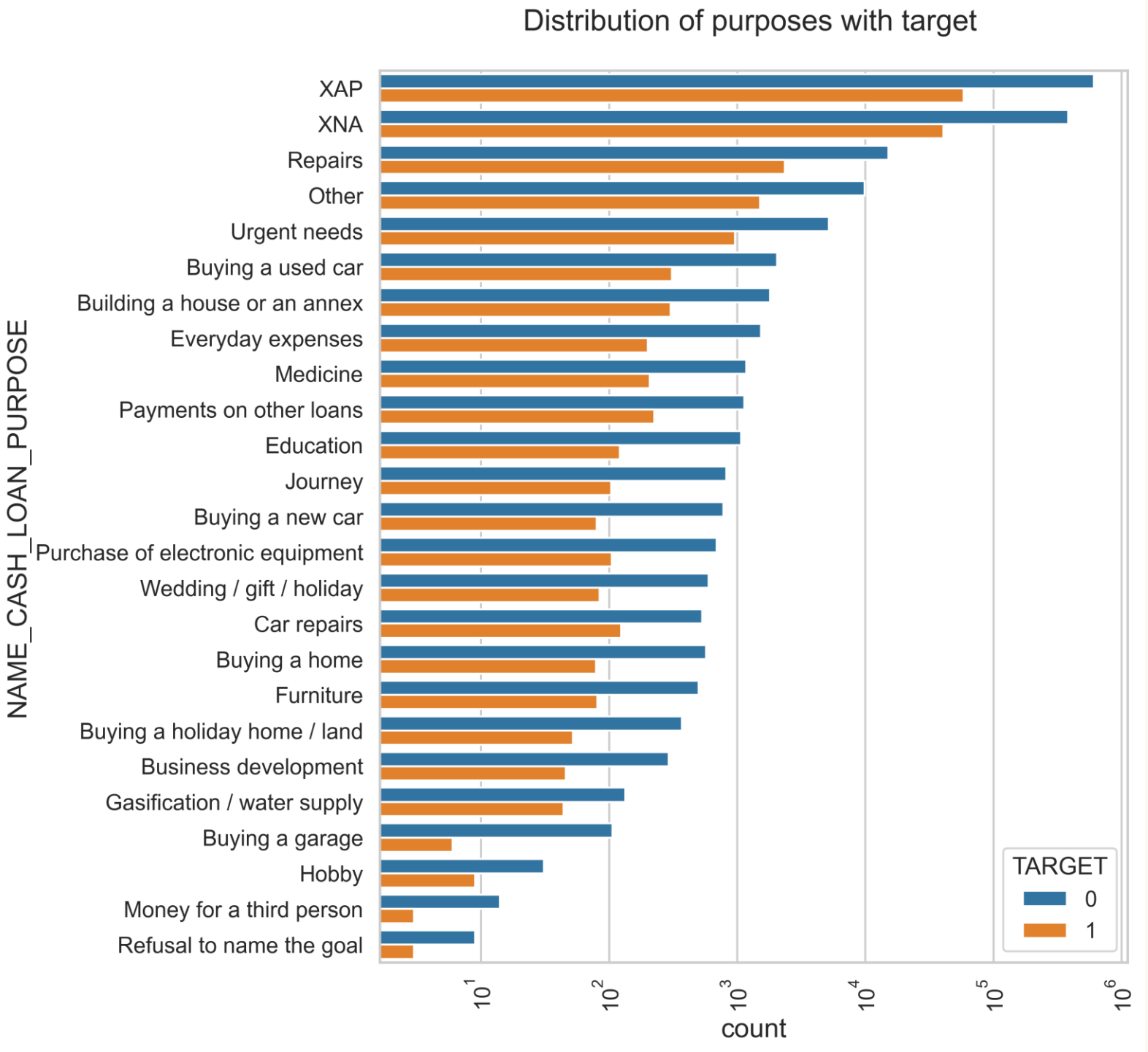
EXPLORATORY DATA ANALYSIS (EDA) APPROACH

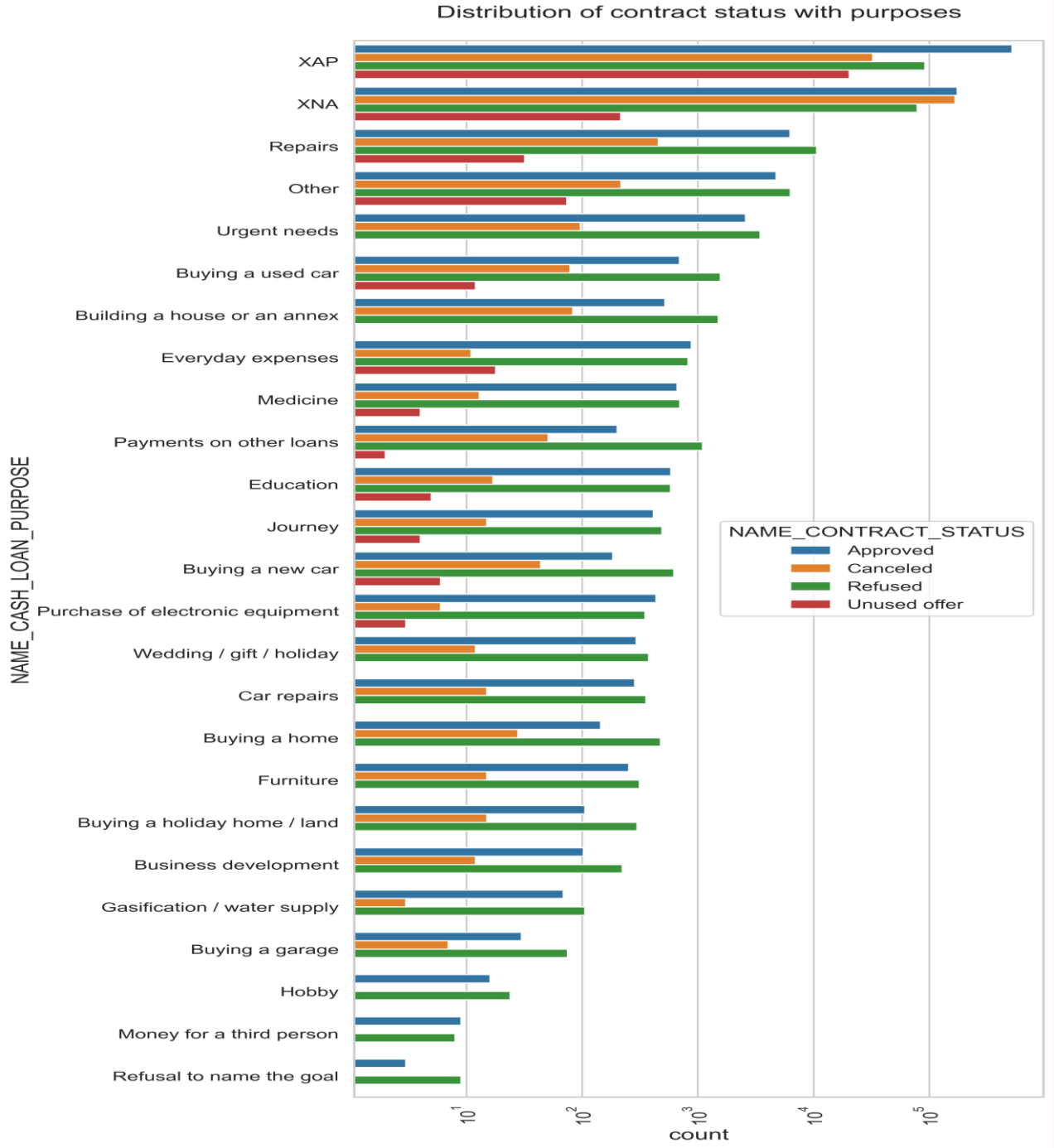
- Very basic thing to start with importing the csv file and checking for the shape, size , info and describe function to get a knowledge of the data given.
- Once I got the insights of the data I checked the missing values in the columns to get a overview converted it to a present value for better understanding.
- Before I can drop the column converted the missing value to a data frame then selected all the columns and converted it to a list.
- Most of the data was missing from the column it is of no use as it can hamper the analysis for the best result dropped the data
- After dropping the columns we can still check there are still 73 columns left. Now I can see the flag columns which represent the document provided by the client or not.

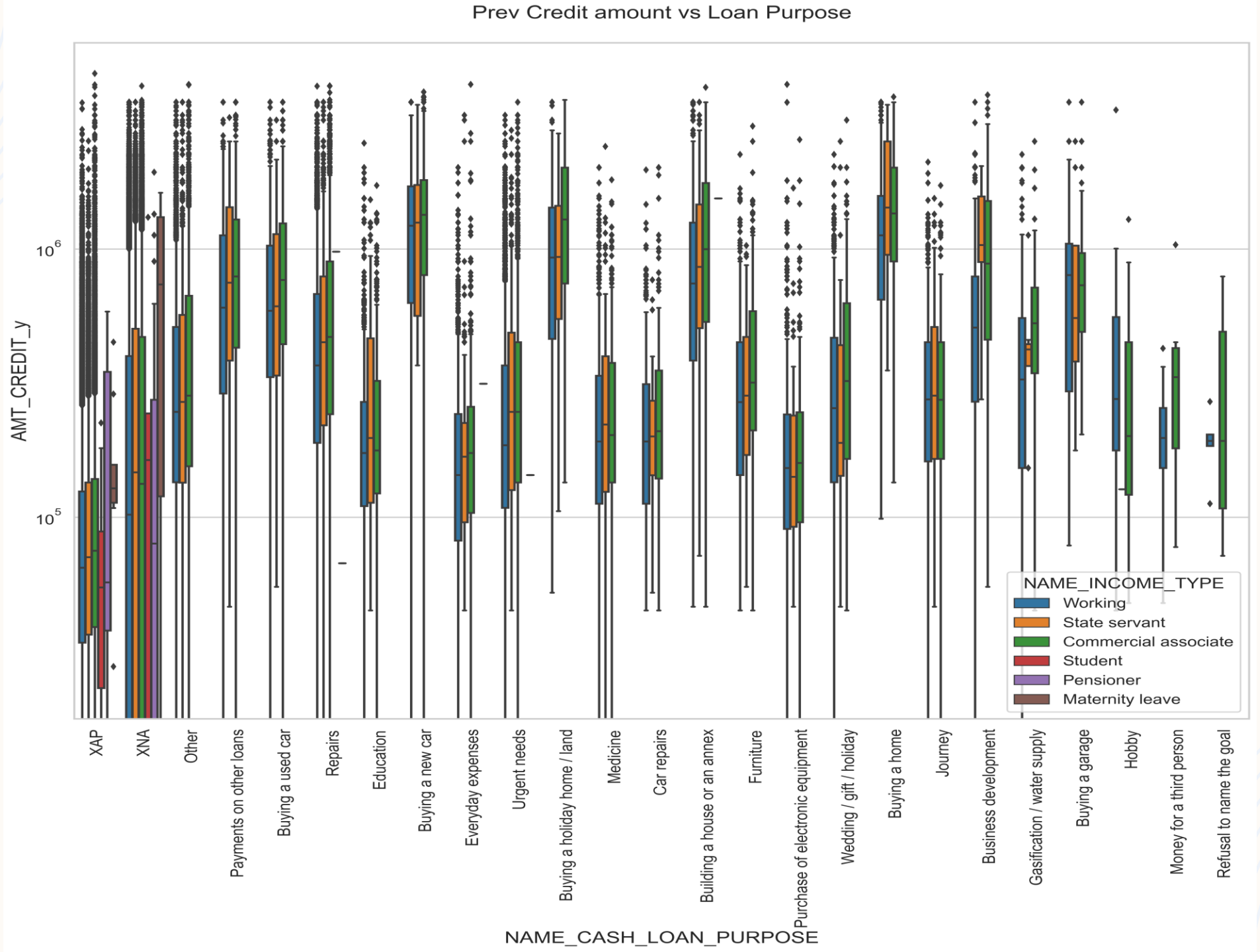
- Checked the all the Flag_documents columns with the Target columns.
- Using subplot and count plot check the relation with the Target between them.
- Once checked some of the Flag columns were still left out so for that I just made a separate data to store them and check the relation between them as well.
- To check the Correlation Coefficients between them I used a heat map
- As there Correlation Coefficients values was less tha +0.8 and -0.8 so item was of no use hence dropped the columns.
- After removing the normalized columns as plotted a heat map and fine the same hence dropped both the columns
- Checked the very first columns for the gender for the male and female find a missing value for XNA however the missing value was less than 1% of the data hence drop the rows of the same.

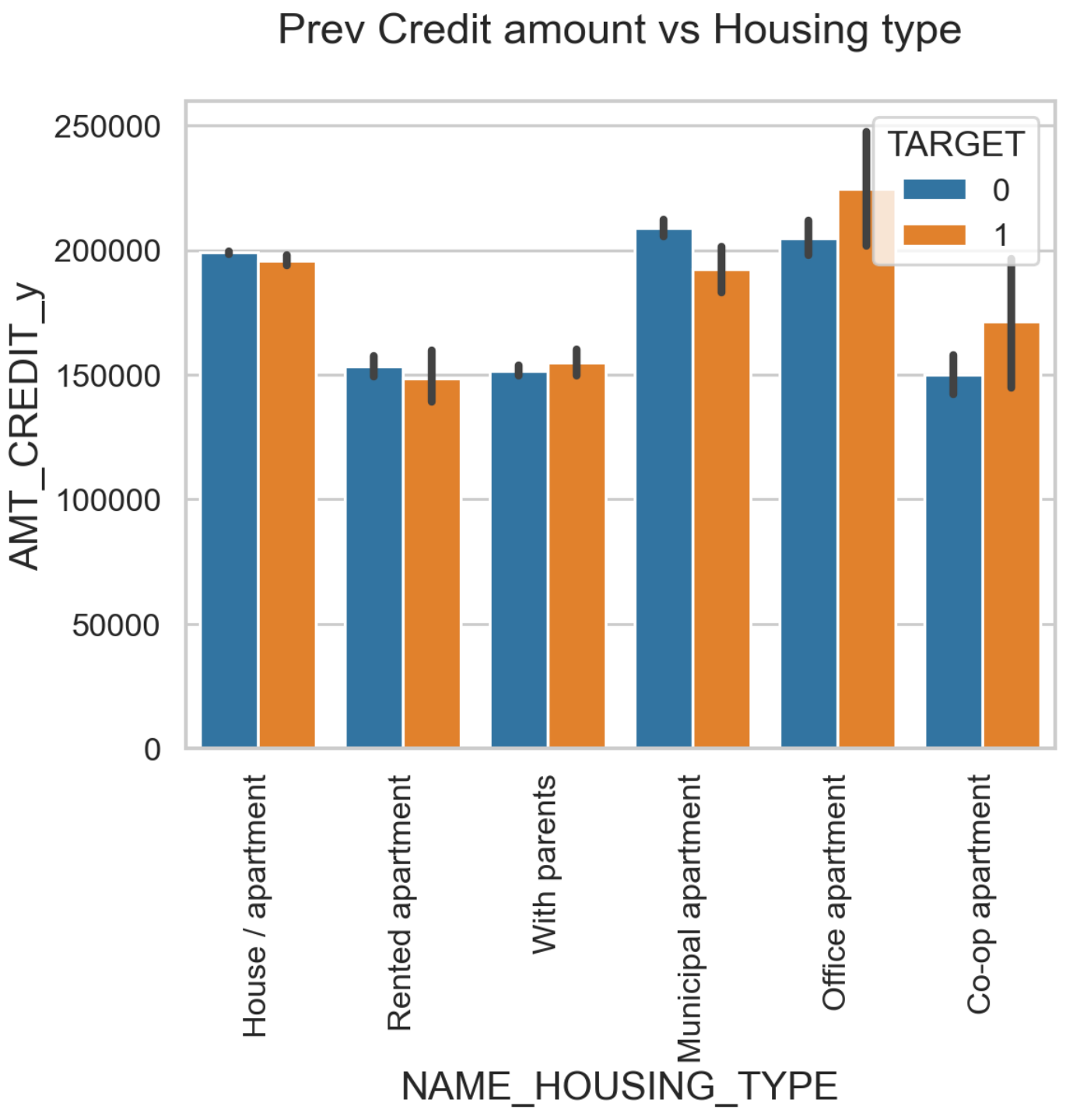
- Once it was done we can check some of the numeric values were in a negative integer . Converted the same value to the positive.
- From some of the numeric columns used a box plot to check if the columns have a outliers . Once plotted the same we can identify that the numeric columns have outliers.
- Created bin for the continues variable categorical columns.
- Once all unwanted and missing columns were removed, we then proceed to one of the important factor of the data set Target.
- From now the analysis was started and divided the Target value in 2 part as 0's and 1's.
- Once it was done I started Categorical Univariate Analysis in logarithmic scale for target=0 .
- Categorical univariate analysis is value scale to check to count the value.

- Once completed the above step, then performed bivariate analysis for the numeric values for Target 0 and same for Target 1.
- Done box plot for the both Target 0 and Target 1.
- Checked for the correlation for both Target 0 and Target 1.
- Now imported the prev_application and check the basic info, shape and describe .
- Checked the missing value and converted them into percentage value to get a better understand.
- Dropped the missing value to more than 50%
- Merged both the data together for a better analysis
- Once merged performed univariate analysis and bivariate analysis for the data to check whether the client falls in the data of defaulter and non- defaulter.


Distribution of purposes with target
NAME_CASH_LOAN_PURPOSE
XAP
XNA
Repairs
Other
Urgent needs
Buying a used car
Building a house or an annex
Everyday expenses
Medicine
Payments on other loans
Education
Journey
Buying a new car
Purchase of electronic equipment
Wedding / gift / holiday
Car repairs
Buying a home
Furniture
Buying a holiday home / land
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
TARGET
0
1
10^1
10^2
10^3
10^4
10^5
10^6
count


Distribution of contract status with purposes
NAME_CASH_LOAN_PURPOSE
XAP
XNA
Repairs
Other
Urgent needs
Buying a used car
Building a house or an annex
Everyday expenses
Medicine
Payments on other loans
Education
Journey
Buying a new car
Purchase of electronic equipment
Wedding / gift / holiday
Car repairs
Buying a home
Furniture
Buying a holiday home / land
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
count


Prev Credit amount vs Loan Purpose
AMT_CREDIT_y
10^6
10^5
NAME_INCOME_TYPE
Working
State servant
Commercial associate
Student
Pensioner
Maternity leave
XAP
XNA
Other
Payments on other loans
Buying a used car
Repairs
Education
Buying a new car
Everyday expenses
Urgent needs
Buying a holiday home / land
Medicine
Car repairs
Building a house or an annex
Furniture
Purchase of electronic equipment
Wedding / gift / holiday
Buying a home
Journey
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
NAME_CASH_LOAN_PURPOSE


Prev Credit amount vs Housing type
AMT_CREDIT_y
250000
200000
150000
100000
50000
0
TARGET
0
1
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE

SUMMARY & RECOMMENDATION

- Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.

- Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.

- Bank can focus mostly on housing type ‘with parents’ , ‘House/apartment’ and ‘municipal apartment’ for successful payments.