Credit EDA Case Study

SUBMITTED BY :
SHANKAR KOKADWAR
&
SHUBHANGI TRIVEDI

Problem Statement

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.

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default

Approach to Analysis

- Importing file and Check structure – Application_data.csv
- Data Quality Check and handling missing values
  - If the columns are categorical, imputed it with mode value
  - For numerical columns, if it contain outliers, imputed the null values with median otherwise with mean value
- Categorizing columns and Binning of Continuous Variable
- Outlier Analysis
- Bifurcating data frame on basis of payment difficulty
- Analysis – For analysis used below approach
  - Univariate Analysis
  - Correlation Analysis
  - Bivariate Analysis
- Importing file and Check structure – Previous_Application.csv
- Data Merging- Merging Application_data and Previous_application
- Analysis of merged data set
  - Univariate Analysis
  - Bivariate Analysis
  - Multivariate Analysis

Data Quality Check and Missing Values

Importing file and Check structure – Application_data.csv

Data Quality check :

- We identified columns that have more than 50% values as null and hence removed 41 such columns.

Null Value Handling:

- AMT_ANNUITY : Since there was minimum difference between max and mean value, this suggested presence of outliers (confirmed using box plot) and hence null values need to imputed by median instead of mean.
- AMT_GOOD_PRICE : Since 99 percentile value and max value had huge difference, it confirmed the presence of outliers and hence null values need to be imputed by median.
- NAME_TYPE_SUITE : Since this is a categorical column, we imputed the null value with Mode , and after imputation confirmed that no major change has occurred in percentage of categories.
- CNT_FAM_MEMBERS : Again a categorical column, hence used mode value
- EXT_SOURCE_2: On plotting , a right skewed plot is observed that confirms outliers and hence median is well suited for imputing null values
- OBS_30_CNT_SOCIAL_CIRCLE : Here also outliers observed hence null value imputed with median

Categorizing columns and Binning continuous Variable

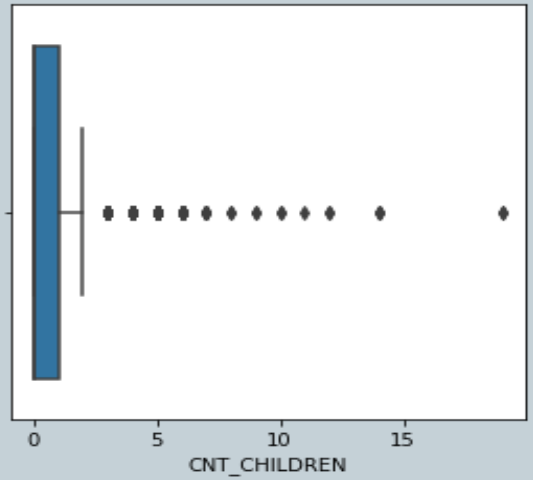
Categorizing Columns:

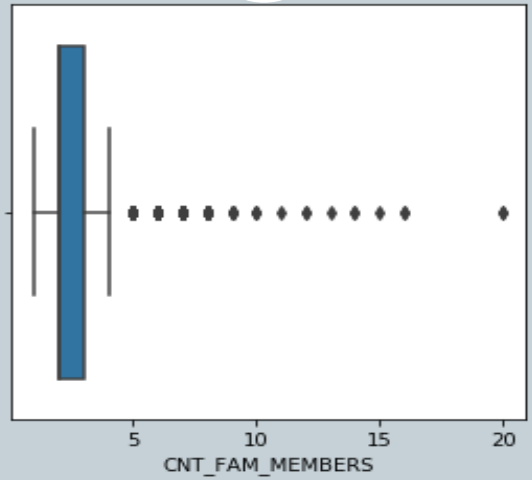
- To identify if a column is numerical or categorical, for each we identified distinct values and any column having count of distinct values more than 8, we considered it as categorical column and other as numerical/Continuous column.
- Few exceptions: ***ORGANIZATION_TYPE and OCCUPATION_TYPE ,*** though identified as numerical, we moved them to categorical columns

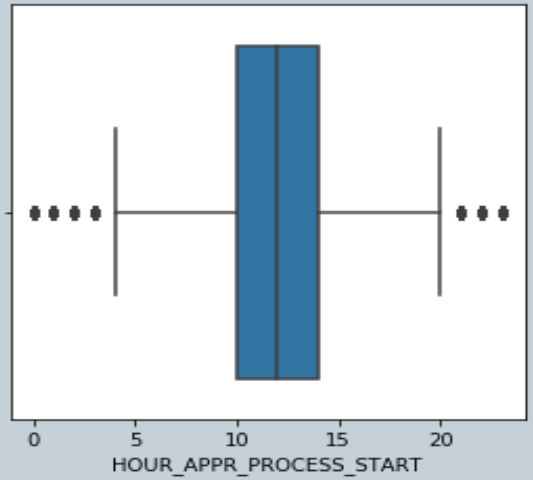
Binning Continuous Variable:

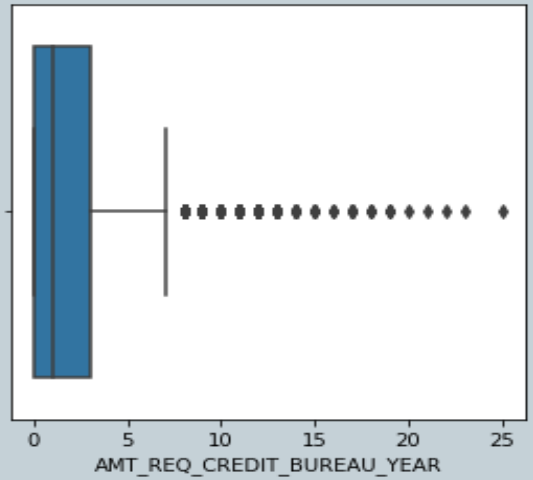
- We have categorised the CREDIT_AMT into categories such as *Low_exposure, moderate_exposure, medium_exposure, High_exposure and very_high_exposure* based on credit amount given to the customers
- Similarly we have created bins for income i.e. AMT_INCOME_TOTAL column, creating categories such as *0-1 lakh, 1-3 lakh, 3-5 lakh, 5-10 lakh, 10-15 lakh, 15-20 lakh and 20+*

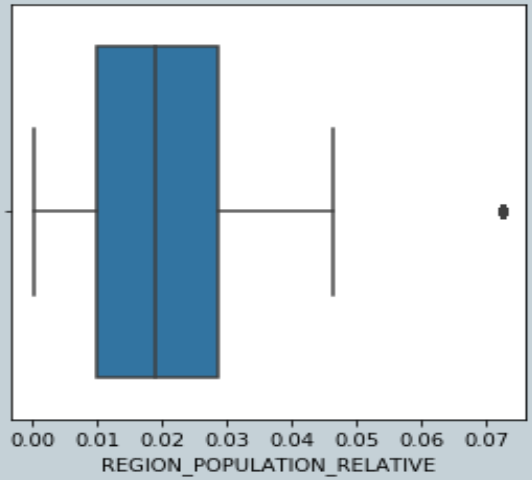
Outlier Analysis

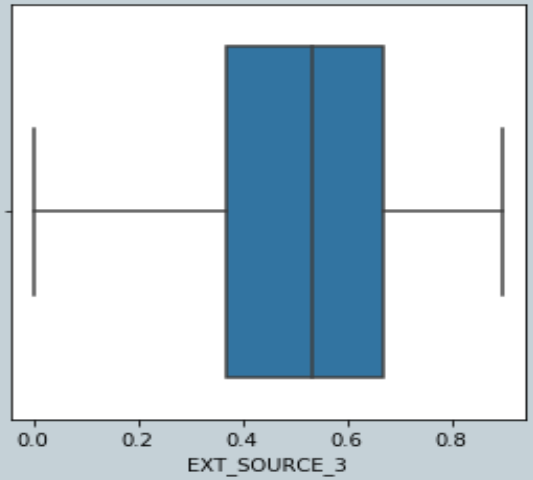












Outlier Analysis

POINTS CONCLUDED:

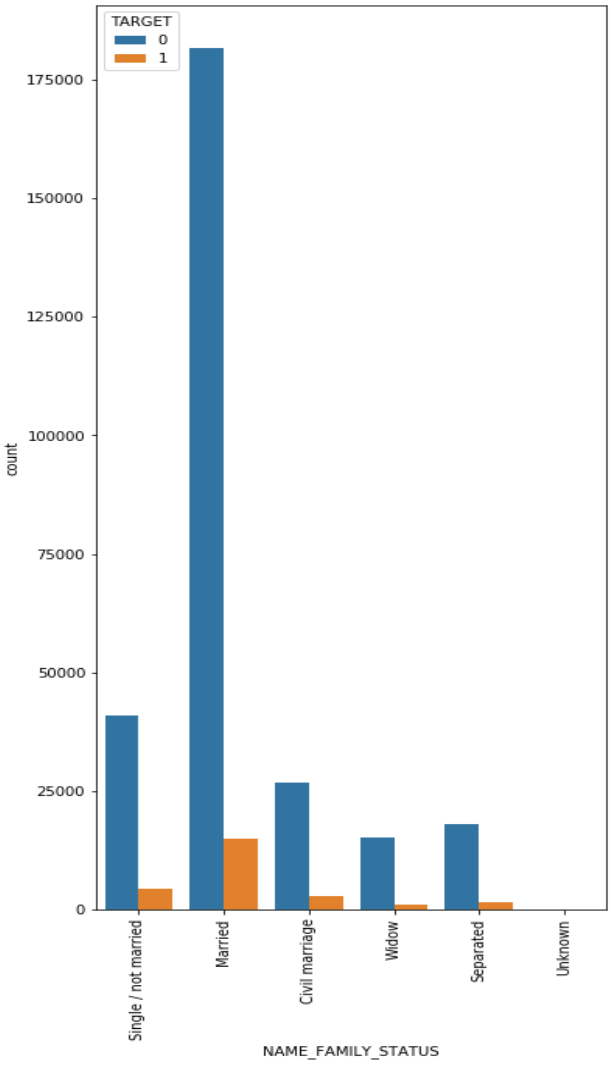
- We can see some outliers in the CNT_CHILDREN column. Most of the customers are having children in the range of 0 to 2 but having children more than 5 is rare but certainly possible hence we may bin them in case of analysis.
- We can see some outliers in CNT_FAM_MEMBERS column. Most people have 2 to 3 family members. But some people have as large as 20 which is a possibility in Indian context where joint families exist.
- We can see some outliers in case of HOUR_APPR_PROCESS_START. We can see that most of the customers have applied between 10 AM to 2 PM which can be considered as office time. But it may be possible that bank has online application for loan and hence other timings may very well be possible. Or else they can be considered as outliers
- We can see some outliers in AMT_REQ_CREDIT_BUREAU_YEAR. We can see that for most customers no of enquiries to credit bureau are in the range 0 to 3 which is a good sign. Generally if credit bureau enquiries are more than 6 customer is not considered credit worthy. Since the customer have many options to take credit more than 10 credit enquiries is very much possible and in reality the values are possible.
- There is only a single outlier in REGION_POPULATION_RELATIVE who is living in a very populated area.
- There is no outlier in EXT_SOURCE_3 column.

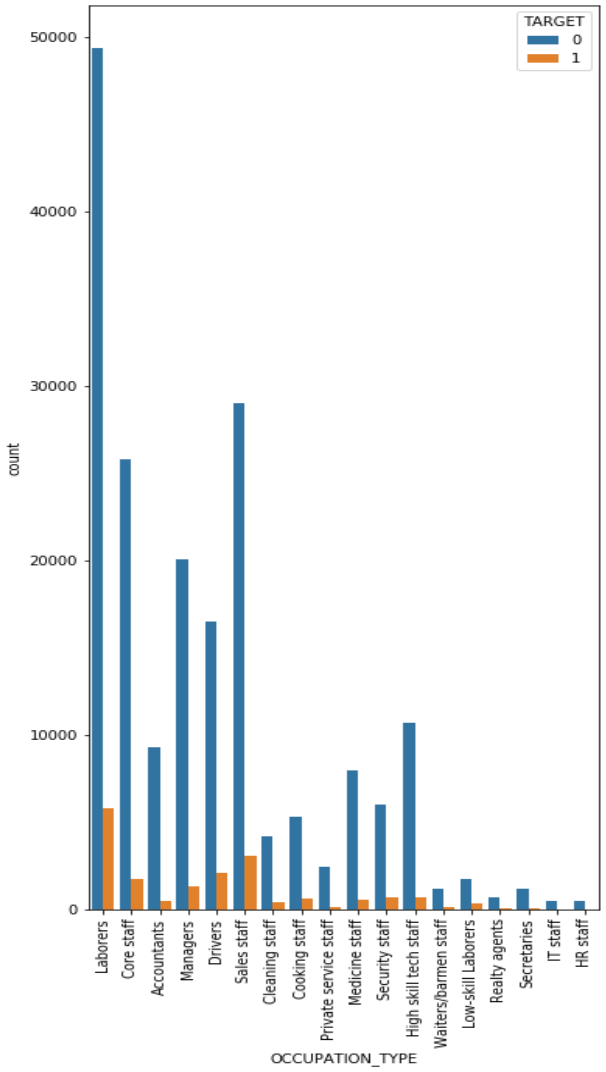
Univariate Analysis

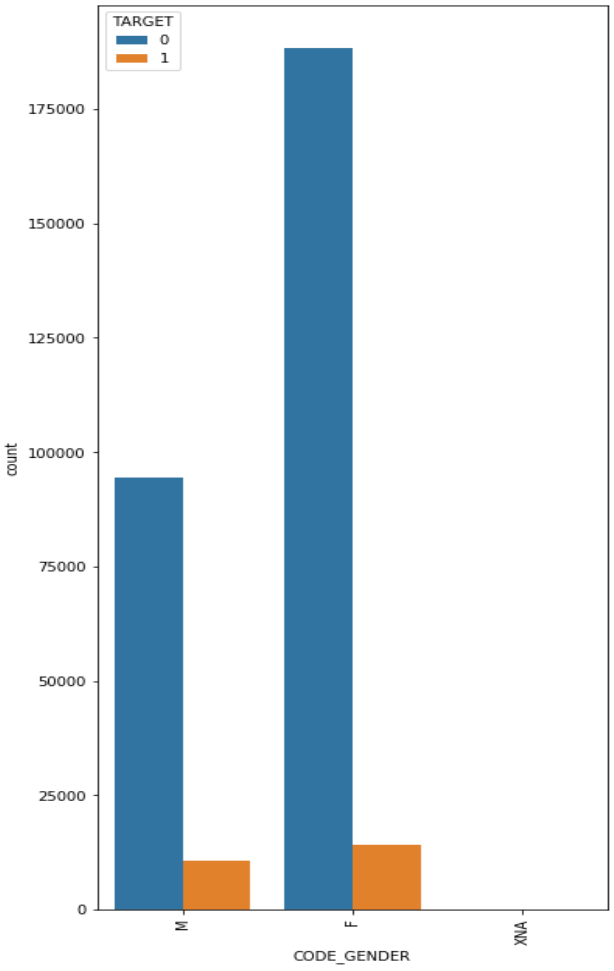
WE HAVE BIFURCATED THE DATAFRAME FOR APPLICATION DATA BASED ON PAYMENT DIFFICULTY OVER TARGET COLUMN VALUE

FOR UNIVARIATE ANALYSIS WE HAVE USED LOOP GRAPHING TECHNIQUE

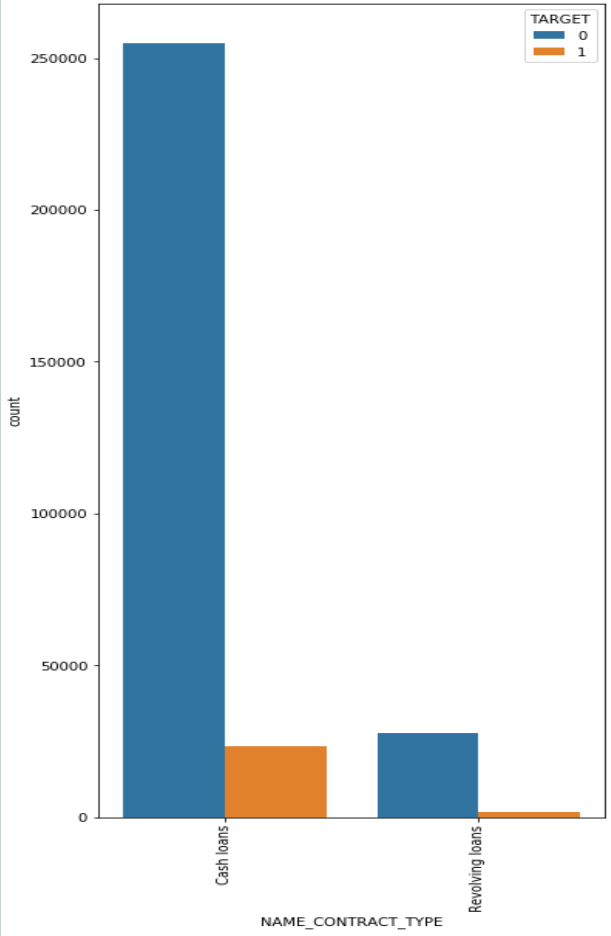
Univariate Analysis

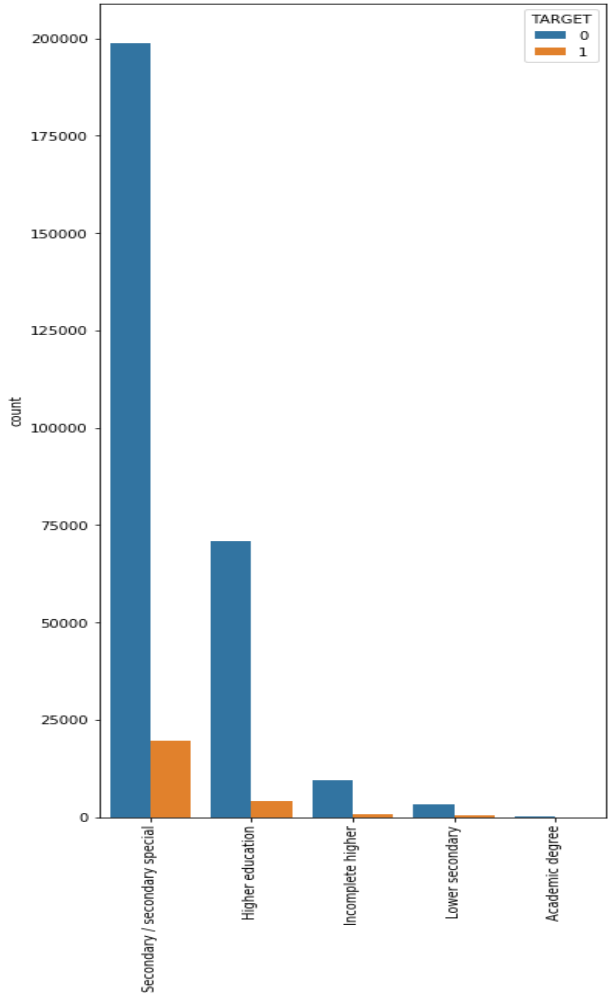


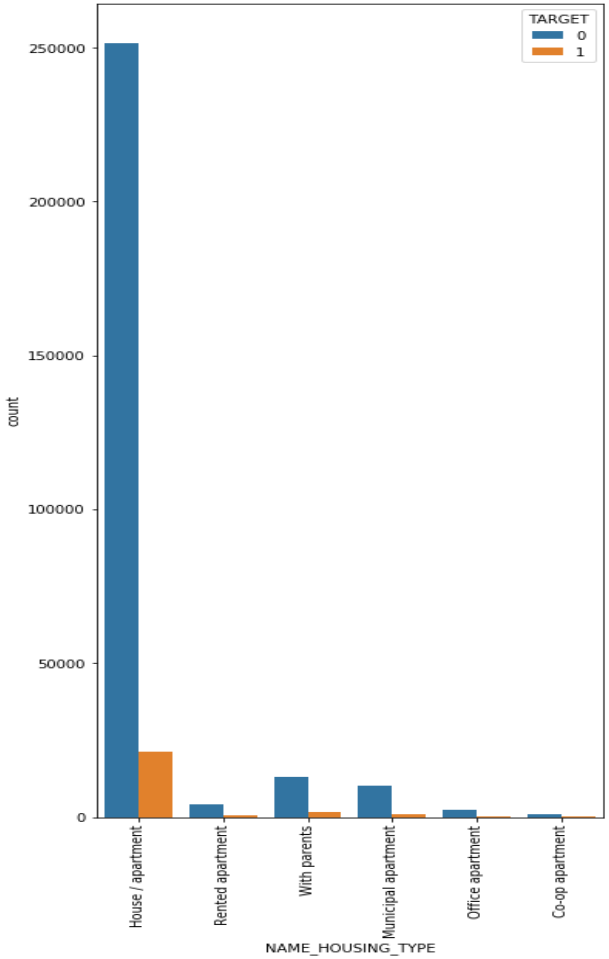




Univariate Analysis







Univariate Analysis

POINT CONCLUDED:

- We can see from the above that most of the customers of the bank are Married (married and civil marriage category). The second most customers are in the category single.
- Most of the customers of the bank are either labourers or sales staff.
- Most of the customers of the bank are females and compared to female borrowers there is comparatively more payment difficulty in case of male borrowers.
- There are more cash loans as compared to revolving loans.
- Most of the customers with secondary/secondary special education or with higher education. However, there are very few customers with academic degree.
- Most of the customers are with Housing Type as House/Apartment

Correlation : Target 0 & 1

POINT CONCLUDED : PAYMENT DIFFICULTY

- There is a strong correlation between AMT_CREDIT and AMT_GOOD_PRICE , this shows that loans is granted based on good price.
- There is a strong correlation between AMT_ANNUITY and AMT_CREDIT, this shows that instalments are decided on credited amount.
- There is correlation between DAYS_BIRTH and DAY_EMPLOYED, that means a person with more age will have more work experience.
- There is a very weak correlation between DAY_EMPLOYED and AMT_INCOME_TOTAL , this shows that experience have little impact on salary intake.
- There is weak correlation between DAYS_BIRTH and AMT_ANNUITY, this shows that credit amount decrease with increase in age.

Correlation : Target 0 & 1

POINT CONCLUDED : WITHOUT PAYMENT DIFFICULTY

- There is a strong correlation between AMT_CREDIT and AMT_GOOD_PRICE , this shows that loans is granted based on good price.
- There is a strong correlation between AMT_ANNUITY and AMT_CREDIT, this shows that instalments are decided on credited amount.
- There is correlation between DAYS_BIRTH and DAY_EMPLOYED, that means a person with more age will have more work experience.
- There is a very weak correlation between DAY_EMPLOYED and AMT_INCOME_TOTAL , this shows that experience have little impact on salary intake.
- There is weak correlation between DAYS_BIRTH and AMT_ANNUITY, this shows that credit amount decrease with increase in age

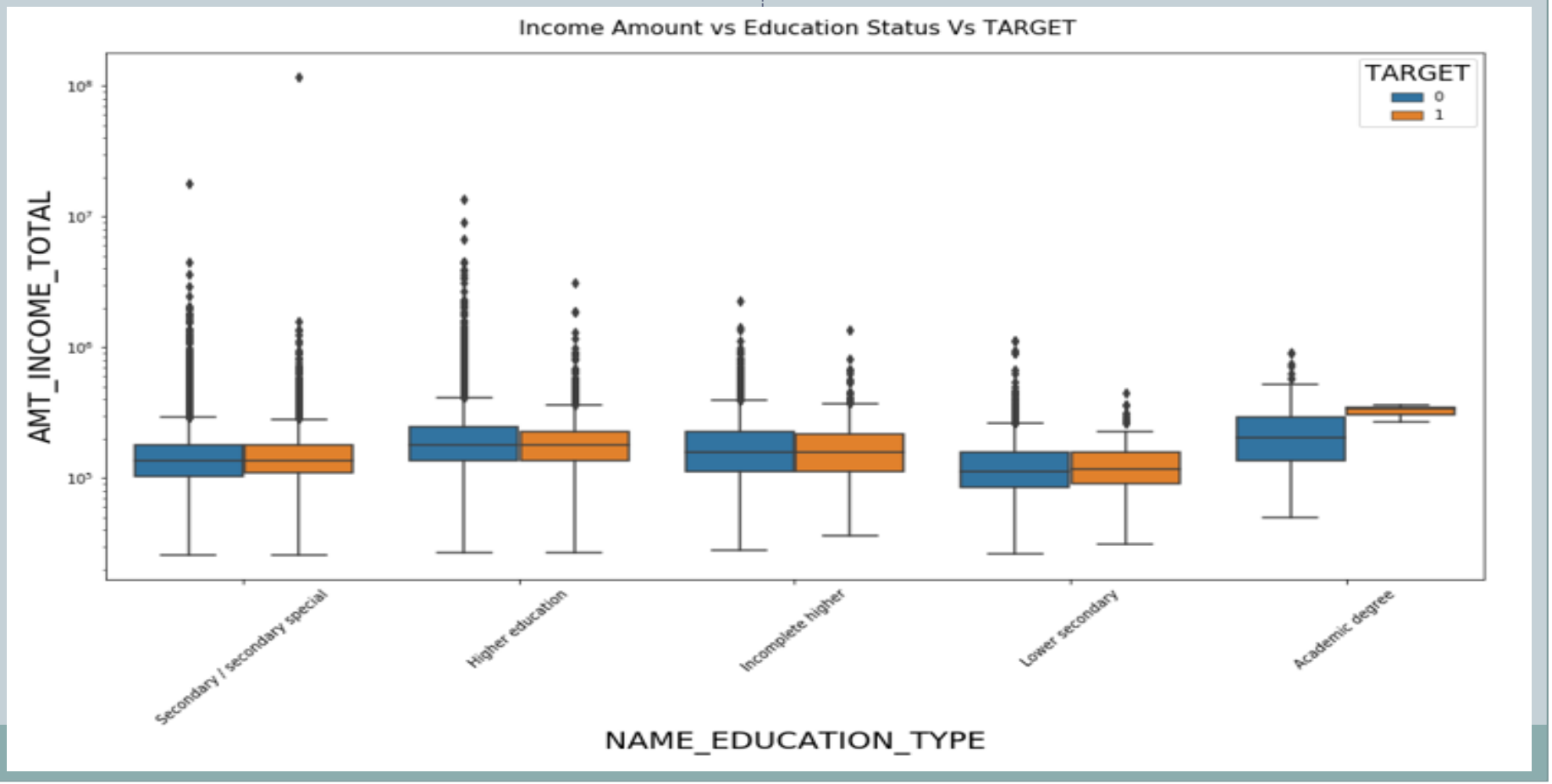
CORRELATION FINAL CONCLUSION

Highest correlation factor are same irrespective it is Target 0 or Target 1

Bivariate Analysis - Distribution of Income and Education with Payment Difficulty

POINTS CONCLUDED:

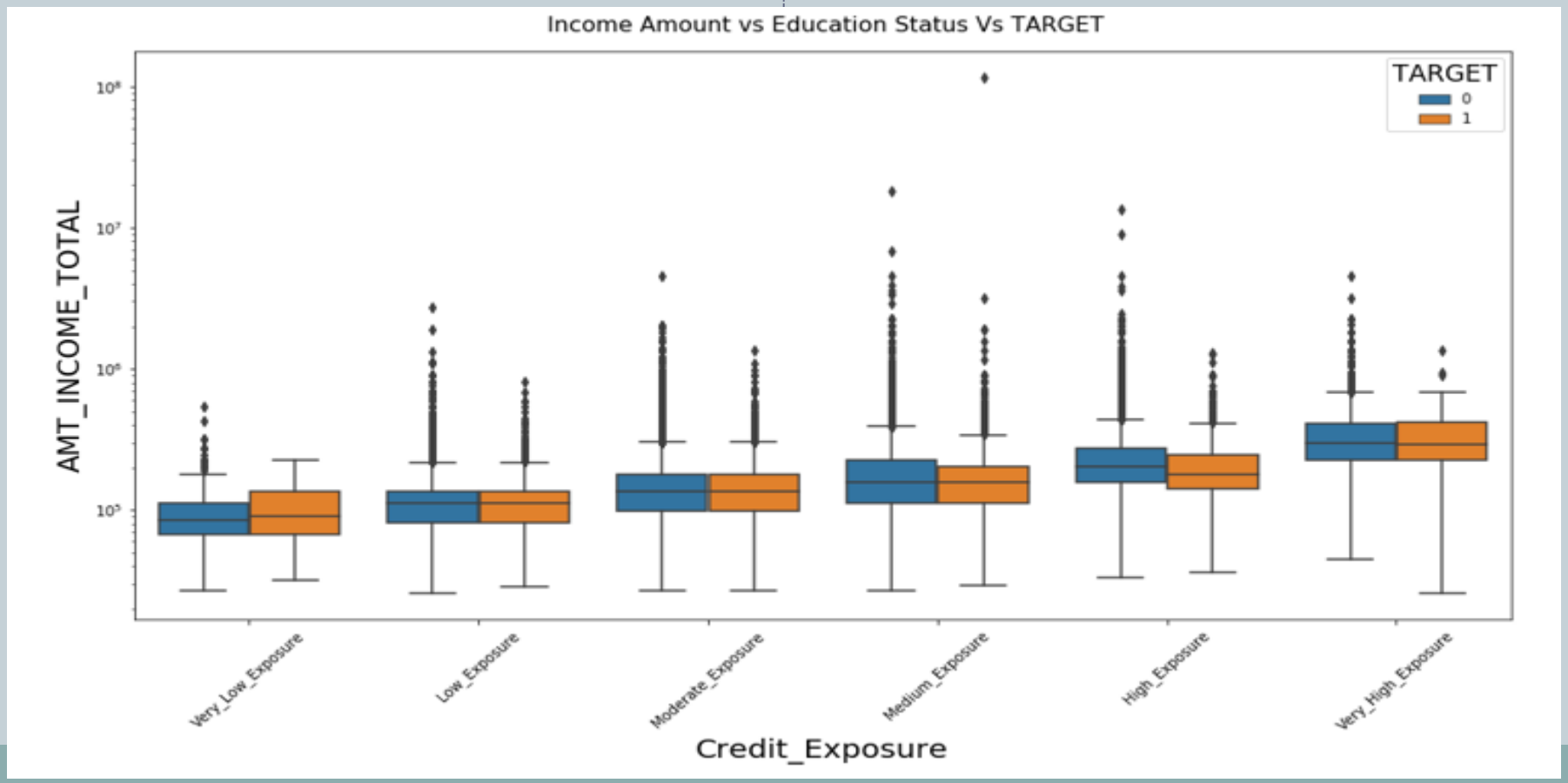
- Organisation should try to focus more on people with academic degree
- People with incomplete education and higher education has lesser payment difficulty.
- There are lot of outliers in case of higher education.



Bivariate Analysis - Distribution of Income and Credit with Payment Difficulty

POINTS CONCLUDED:

- Organization is taking exposure to customers based on their income level. As with increase in exposure we can observe the increase in income as well.
- There is more payment difficulty in cases where there is low exposure i.e. exposure of less than 50,000/-



Univariate Analysis after Merging Previous Data

PREREQUISITE :

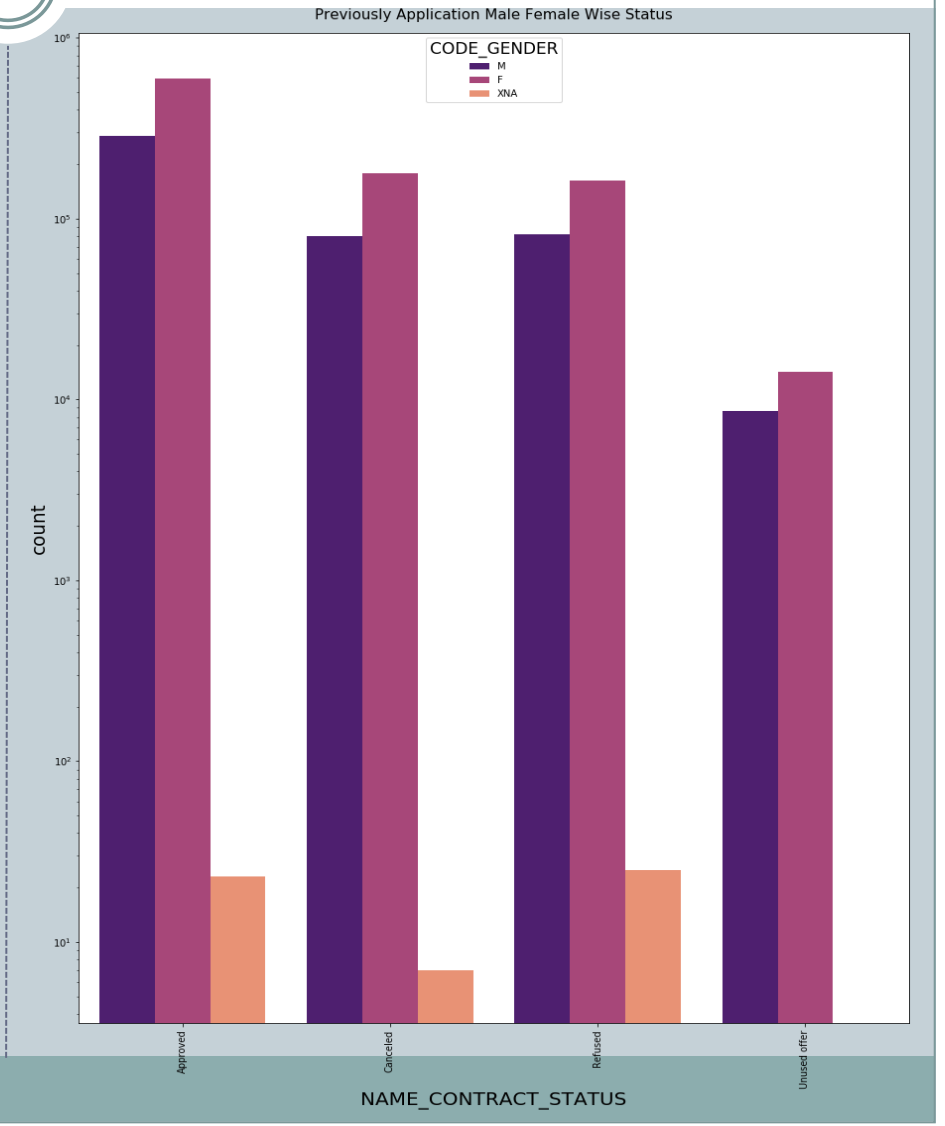
HANDLING PREVIOUS APPLICATION DATA SET

- IMPORTING FILE AND CHECK STRUCTURE – PREVIOUS_APPLICATION.CSV
- DATA QUALITY CHECK AND HANDLING MISSING VALUES
- REMOVING UNNECESSARY COLUMNS
- MERGING PREVIOUS APPLICATION DATA WITH APPLICATION DATA

Distribution of Contract Status with Gender

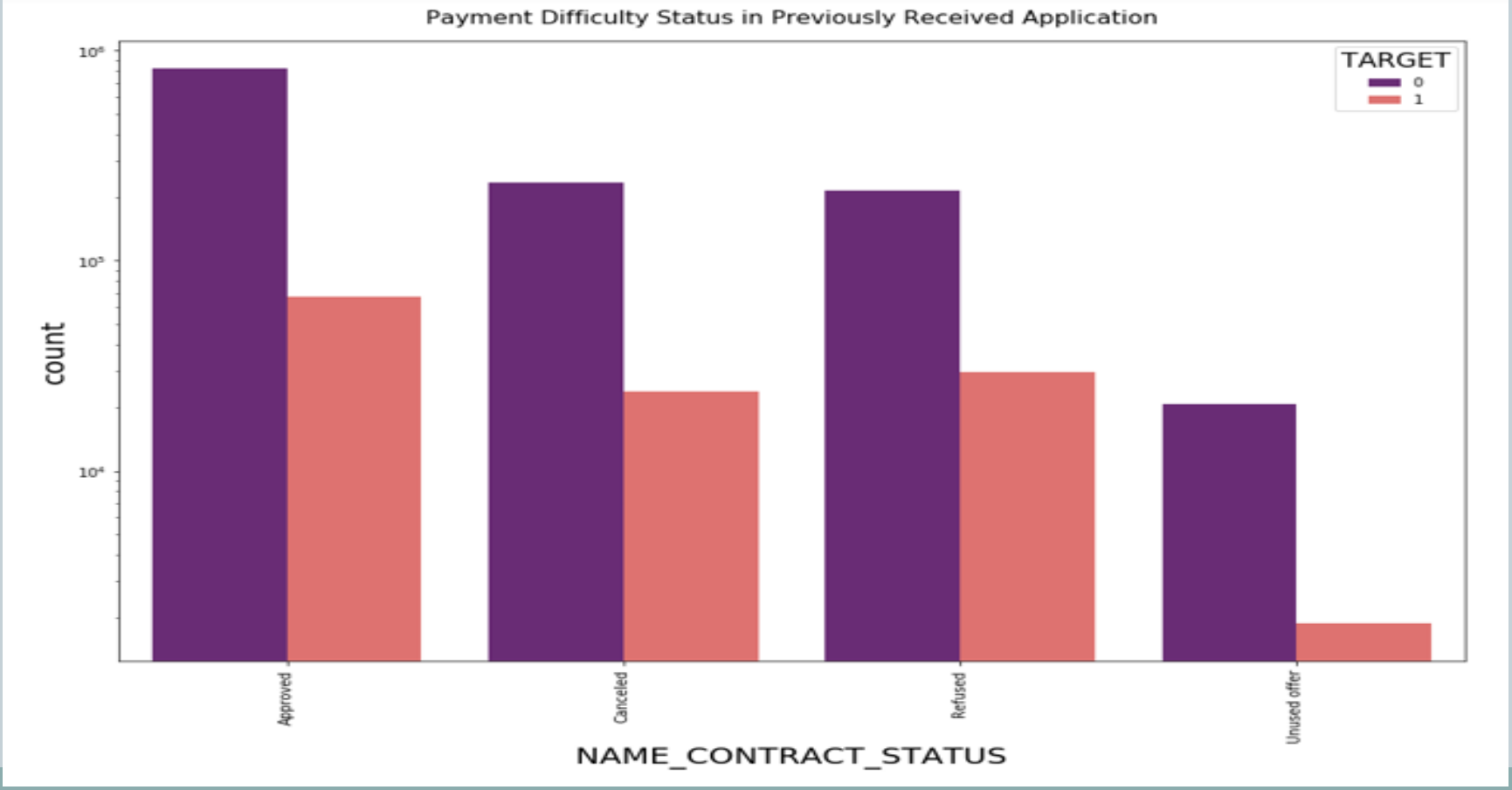
POINTS CONCLUDED:

- We can see that there is not much difference in the application approval/rejection based on Gender. So we can say that gender has not a major role in approval or rejection of loan.



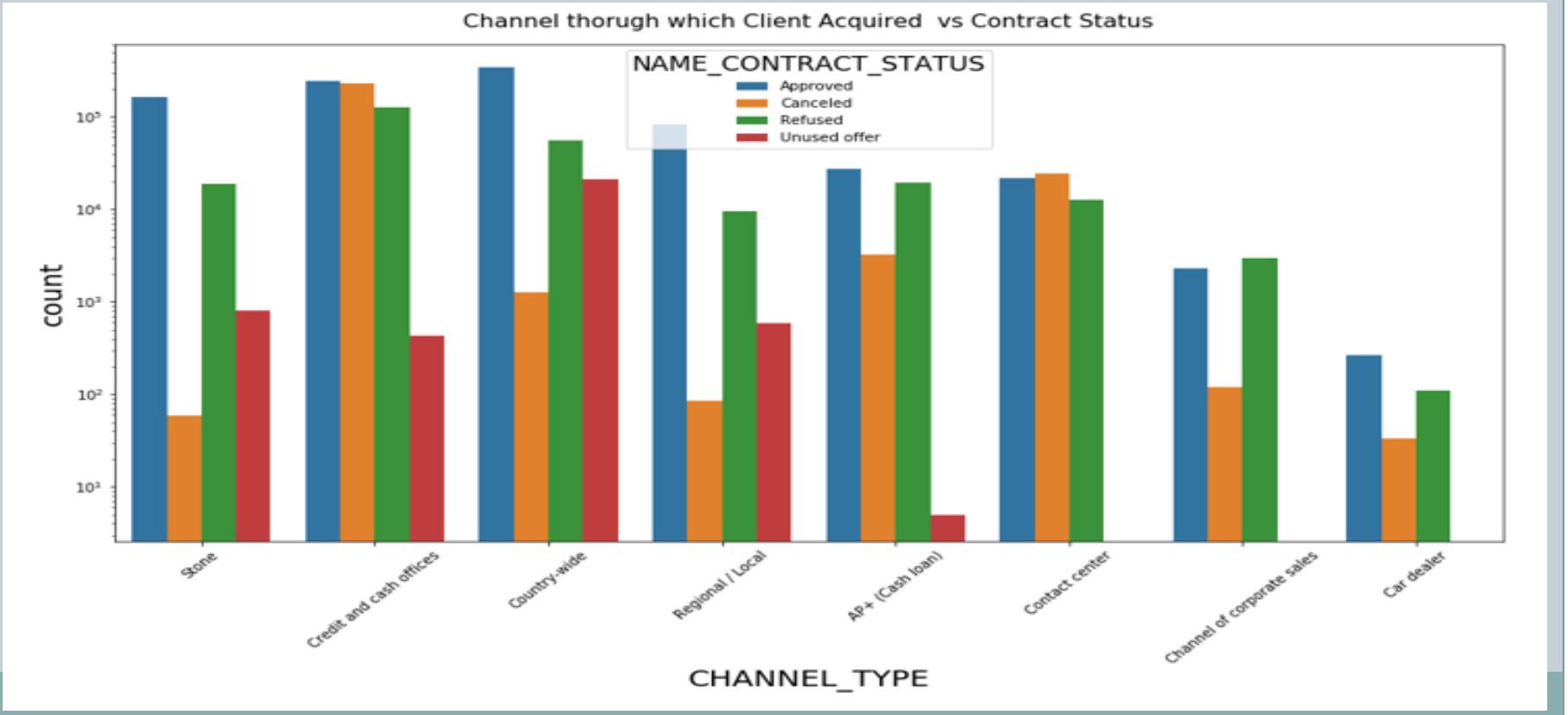
Distribution of Contract Status with Payment Difficulty

1. We can see from the above that in many of the previously rejected cases there is no payment difficulty of payment difficulty percentage is similar to previously approved cases. This can be considered as a loss to the company.
2. In case of previously unused offers the difficulty in payment is relatively very less.



Distribution of Channel with Contract Status

1. In case of customers acquired through contact centre there is very high percentage of cancellation or refusal by the bank.
2. Highest no of application cancelled in between are in case of customers acquired via credit and cash offices.
3. In case the channel is corporate sales there is high percentage of cancellation compared to total application received.



Distribution of Amount Credit with Contract Status

1. The bank needs to improve its rejection mechanism as many of the clients who have earlier refused loans have not faced payment difficulty.
2. There is payment difficulty observed even in case the client is a repeater or refreshed.

Final Conclusion and Recommendations

1. Bank should focus on diversifying its customers base as most of the customers presently as having income less than 3 lakhs.

2. Most of the customers of the bank are either labourers or sales staff. Bank should plan a loan campaign for attracting customers from other professions like IT staff, HR staff, Secretaries as well. This will help the bank to deconcentrate their risk profile.

3. There is less payment difficulty in case of female borrowers.

4. Bank should take step to understand the credit need of customers with academic as presently there are very few such customers with bank.

5. Bank should give preference to customers with less # of credit bureau enquiries. Alternatively, bank can devise a scheme of Risk Premium where in Risk Premium is added to interest rate based on no of credit bureau enquiries.

6. Bank should give preference to customers with more external credit rating source 3.

7. Bank should give preference to customers who have employed for more days and have a sufficient amount of employment left so as to repay the loan.

8. Bank should exercise greater caution in case if low exposure i.e. exposure of less than 50,000/- loans. Bank may consider taking collateral security or having a regular monitoring mechanism for such customers.

9. In case of Customers with Secondary / secondary special education, customer having car or realty should be preferred.

10. Banks credit appraisal system is Gender neutral which bank can highlight to promote their bank as approval and rejection rate is not affected by gender.

11. In case of revolving loans more no. of customers have either cancelled the application or bank has refused them. So bank may rethink of this product.

12. In many of the previously rejected cases there is no payment difficulty or payment difficulty percentage is similar to previously approved cases. This can be considered as a loss to the company.

13. In case of previously unused offers the difficulty in payment is relatively very less. Bank may have a mechanism of regular follow-up with the customers who have not yet used their offer.

14. In case the channel is corporate sales there is high percentage of cancellation compared to total application received. The bank may want to relook at this channel.