Credit EDA Case Study

Umashankar Somasekar
Saksham Kapur

DS C33 upGrad & IITB - Data Science - 30th June 2021

Credit EDA Case Study

This case study is designed on the lines of Risk Analytics, the aim is to understand what are the major risks and recommendations that can be derived from the insights of the data.

The Business Problem

The Credit Risk analysis will help the bank to understand the approval of loans to clients through their own profiles, to identify problem areas and optimise business profits.

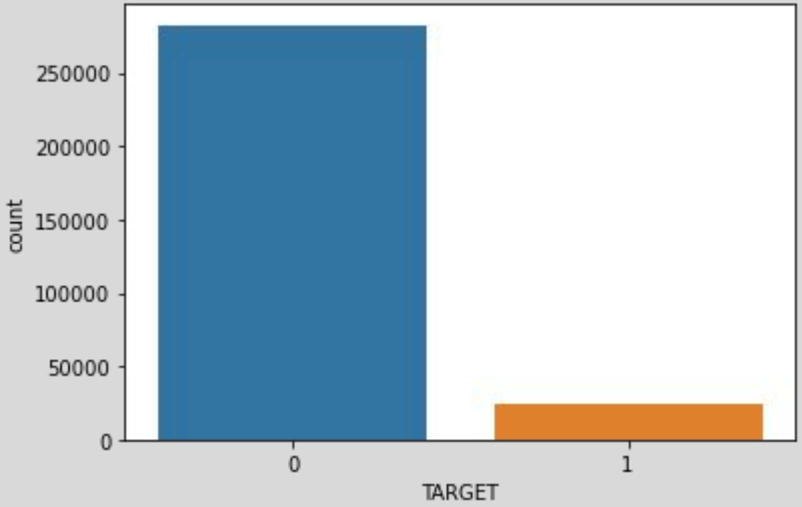
We will utilise EDA as a tool to achieve this.

The Steps Followed

There is no defined structure to EDA but in a nutshell here are the steps that were followed.

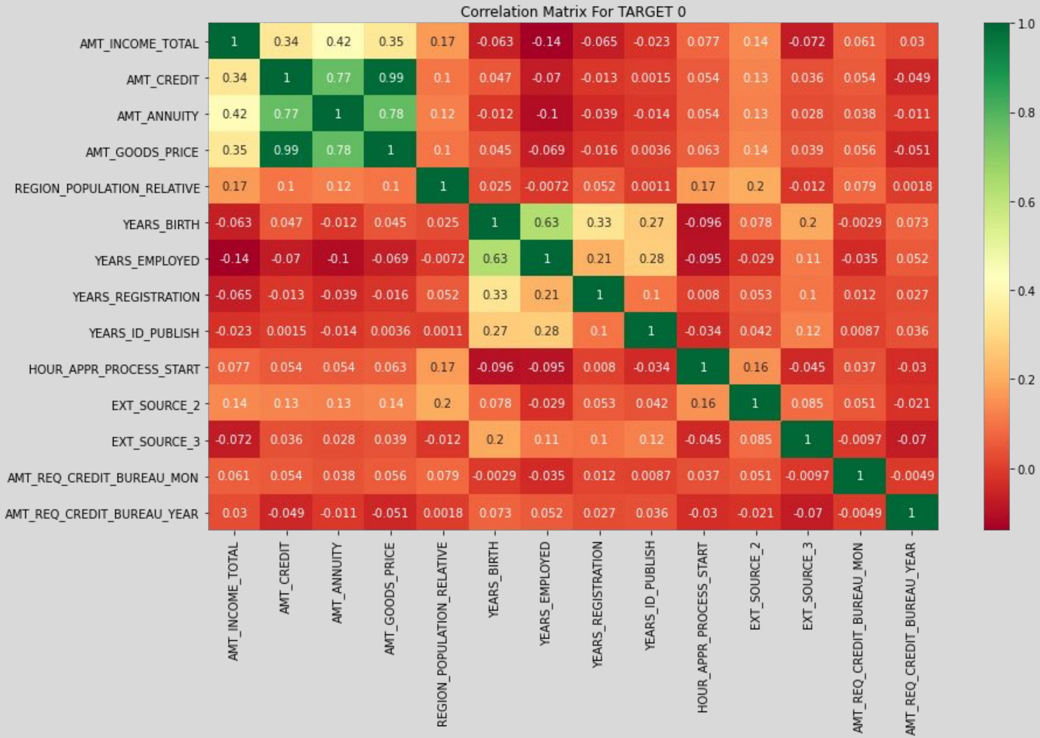
- Step 1: Data Sourcing and Understanding.
- Step2: Check the Data for Quality issues, and binning.
- Step3: Check for Data imbalance, identifying outliers, perform analysis, by Univariate, Bi-variate, and find the correlations.
- Step4: Appropriately Segment the Data into two Target variables, and found out all possible relationships between Variables.
- Step 5: Merged the Data from Previous Application and New Applications.
- Step6: Perform Multi_variate analysis and drawing insights.
- Final Conclusions or Takeaways.

DATA IMBALANCE



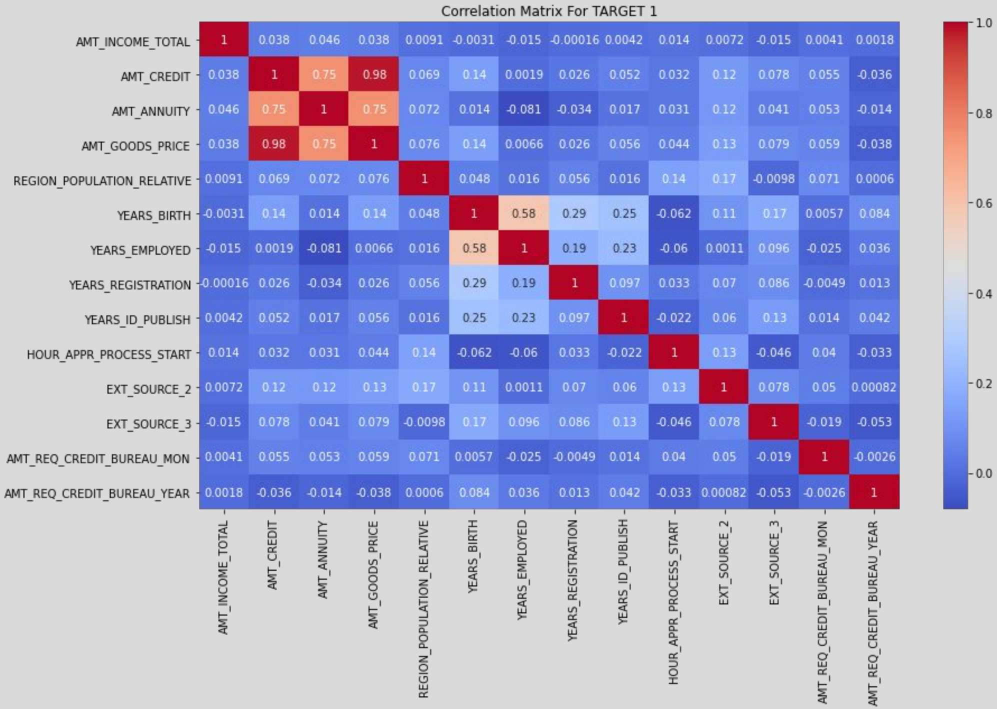
INFERENCE: The Data is highly imbalanced, the Ratio for TARGET 0/ TARGET 1 Came out to be **11.38**

CORRELATION FOR TARGET 0(Clients Without Payment Difficulties)



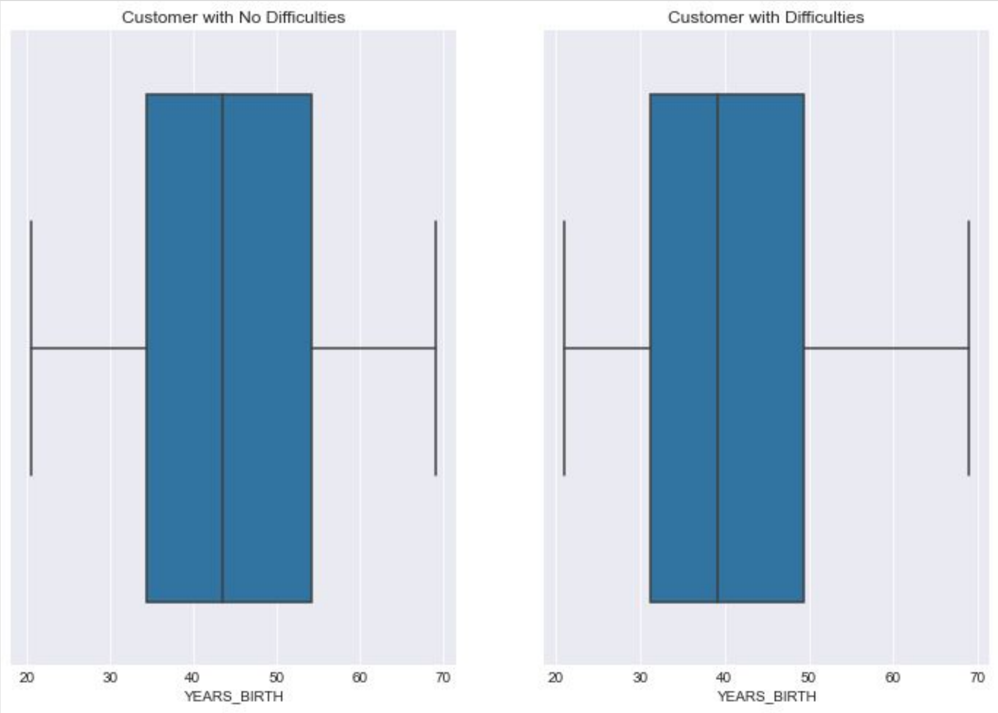
INFERENCE: AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE and AMT_INCOME_TOTAL were highly correlated to Each other.

CORRELATION FOR TARGET 1(Clients With Payment Difficulties.)



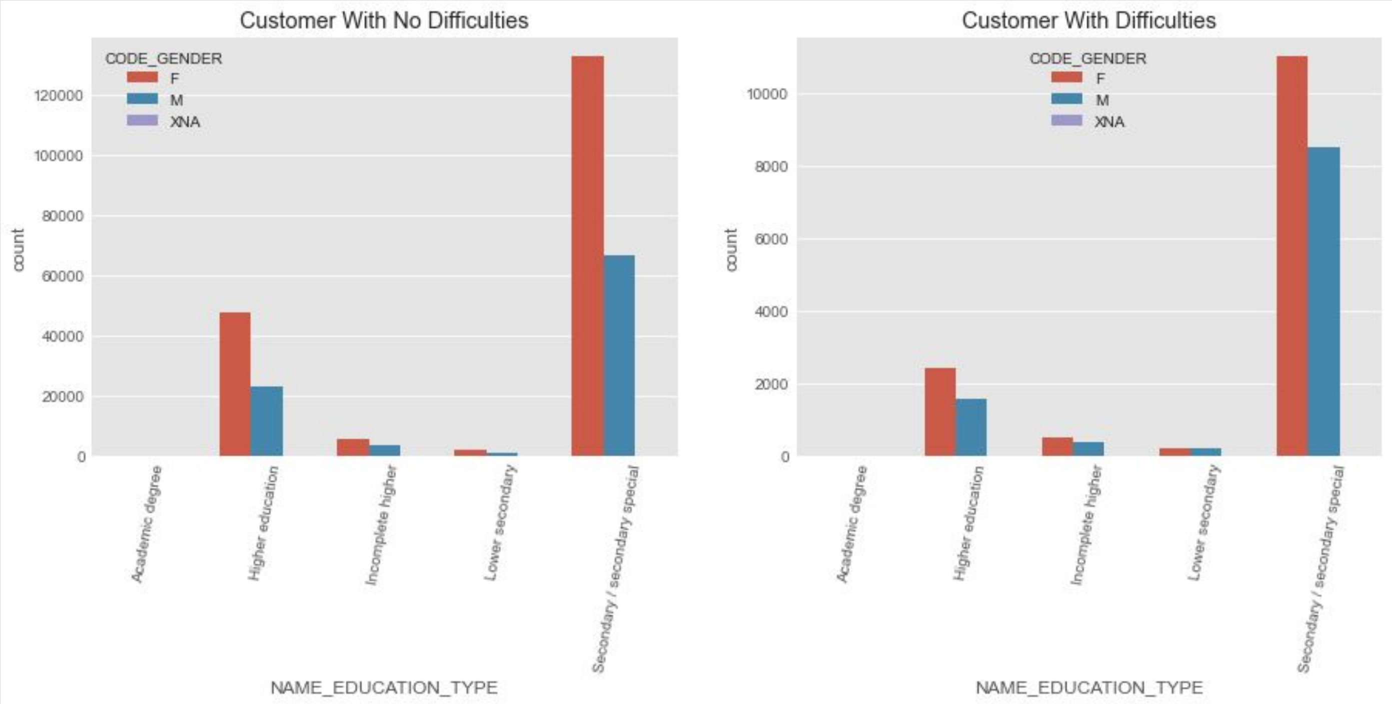
INFERENCE: AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE and AMT_INCOME_TOTAL were highly correlated to Each other.

Univariate - YEARS_BIRTH



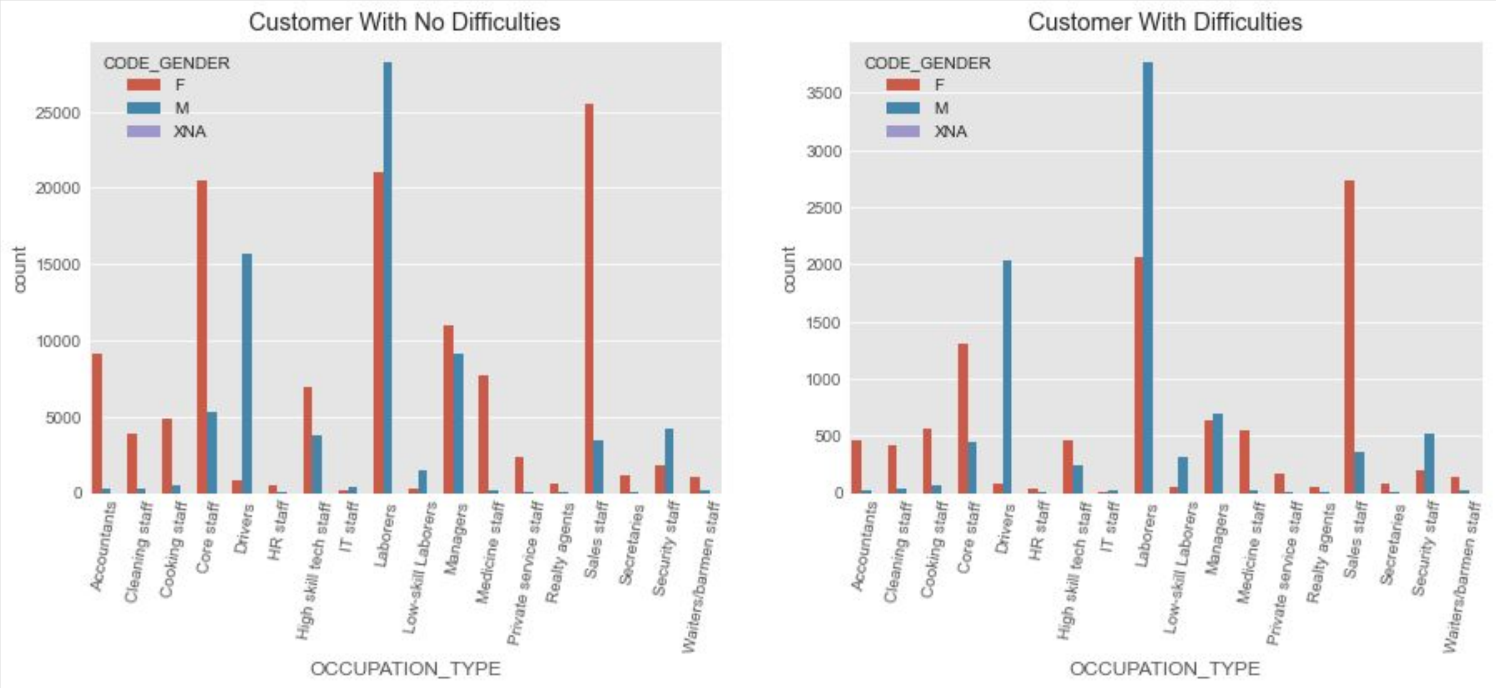
INFERENCE: The Customers with No difficulties, have higher number in the 50+ side, also, younger clients seem to be defaulting more.

Univariate - EDUCATION TYPE



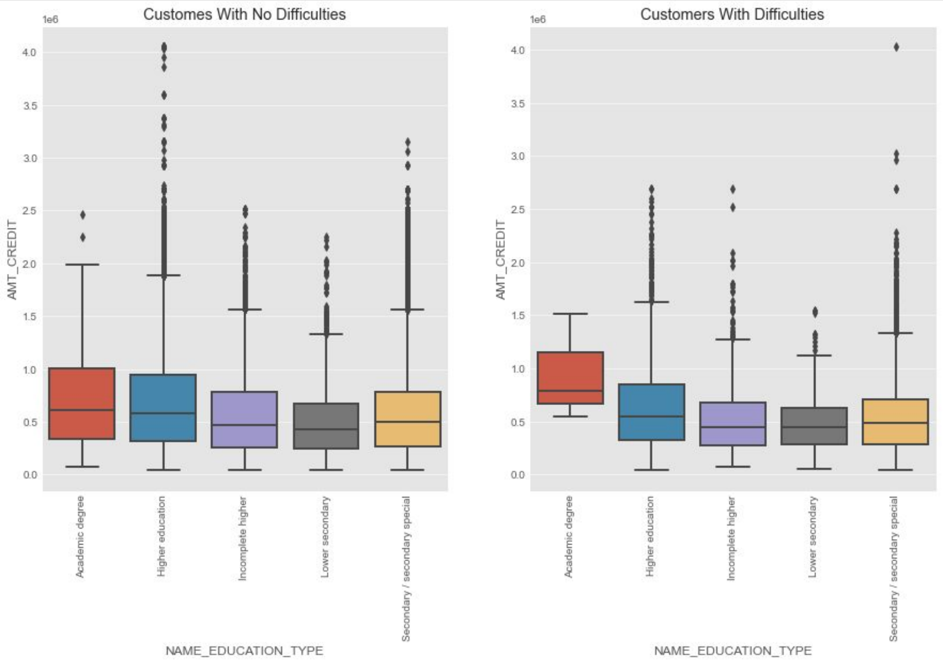
INFERENCE: Females are lending more than Males, and Education point of view, the Customers With Difficulties are more in Secondary Education, and Males who have secondary education have a high number of clients with Payment Difficulties.

Univariate - OCCUPATION TYPE



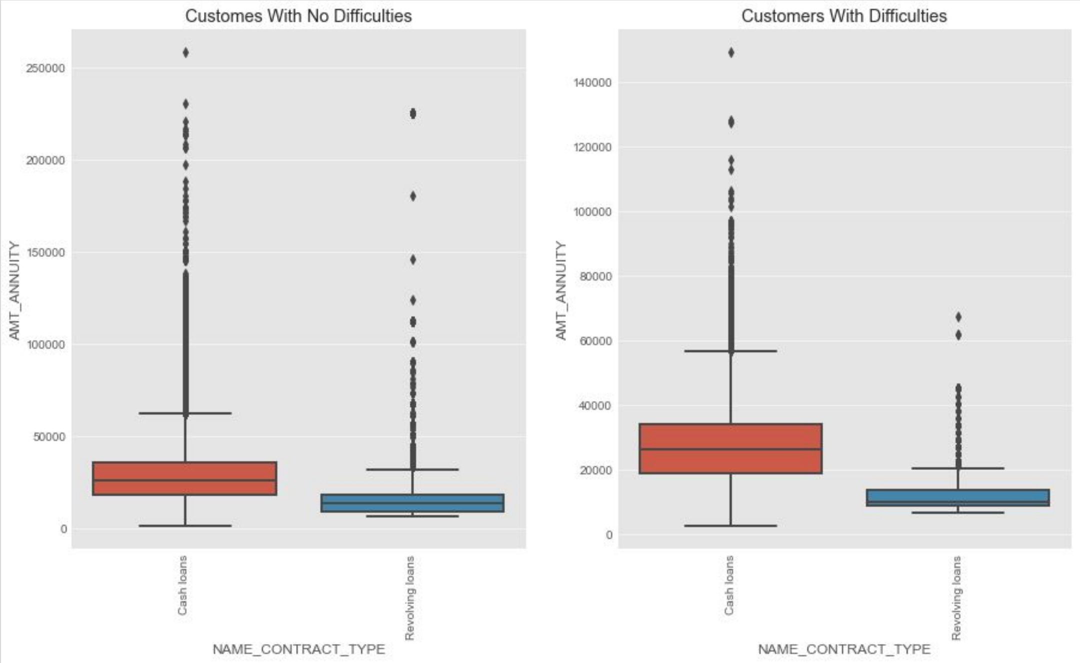
INFERENCE: Laborers are lending the most followed by Sales Staff. So, maybe there are more areas of opportunities.

Bi-Variate - AMT_CREDIT vs Education Type



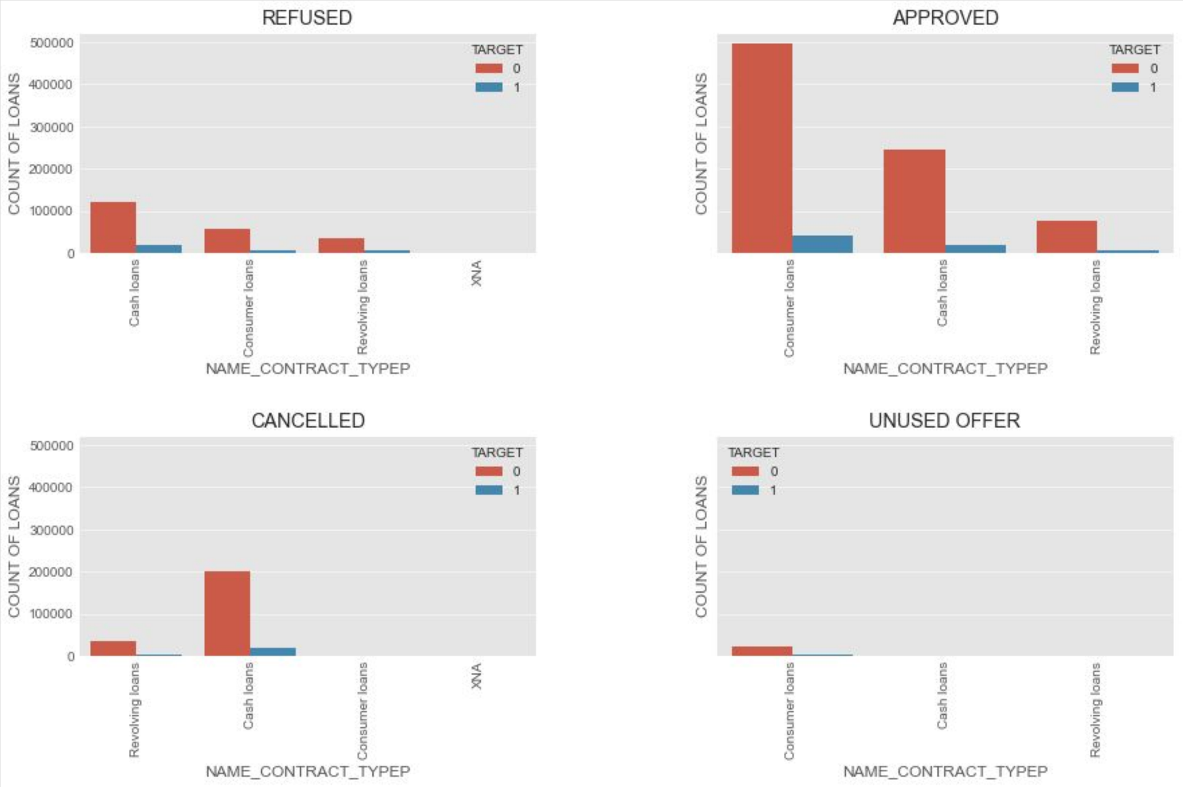
INFERENCE: Higher Education seems to be drawing more Credit than others, Where as Lower Secondary seems to be drawing the Lesser side of Credit.

Bi-Variate - AMT_ANNUITY vs NAME_CONTRACT_TYPE



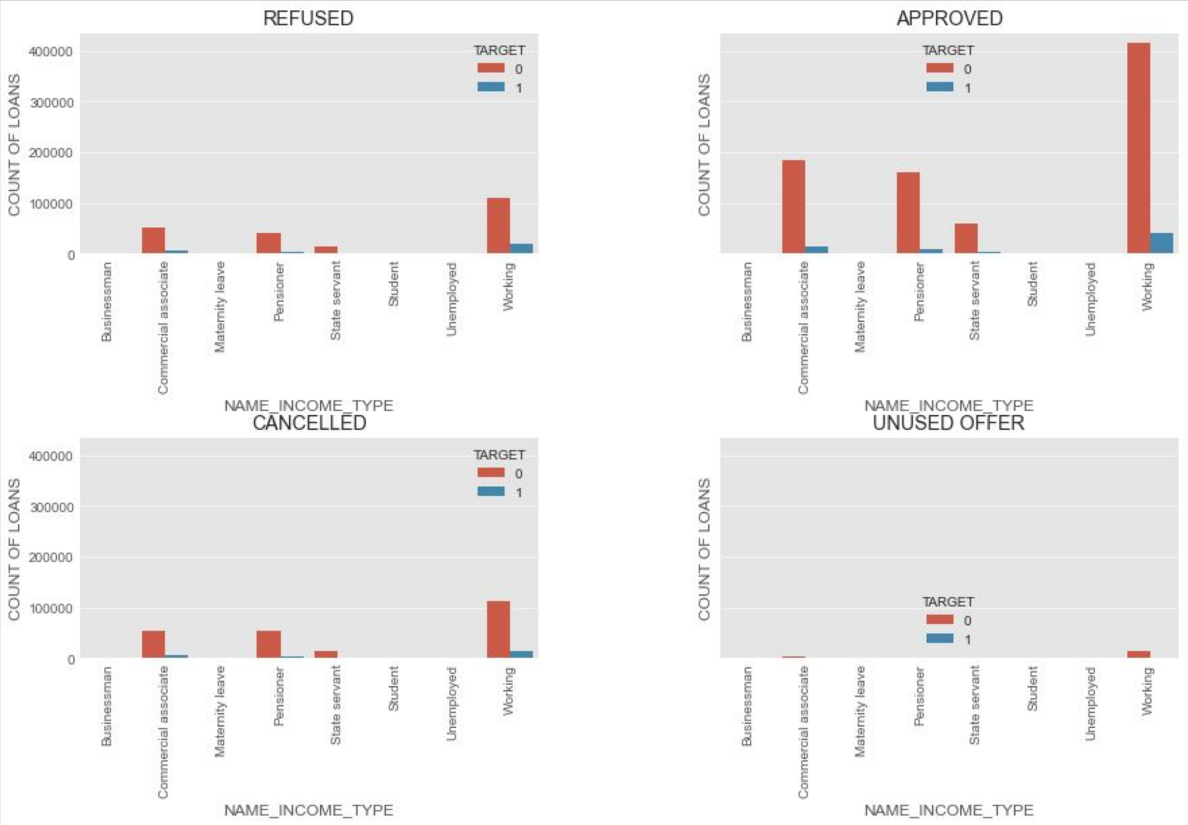
INFERENCE: Cash Loans attract Higher Annuity, and later on we will see that Cash loans have the Highest Cancellations, probably due to higher annuity it attracts.

Final Dataset - Count of Loans Vs NAME_CONTRACT_TYPE



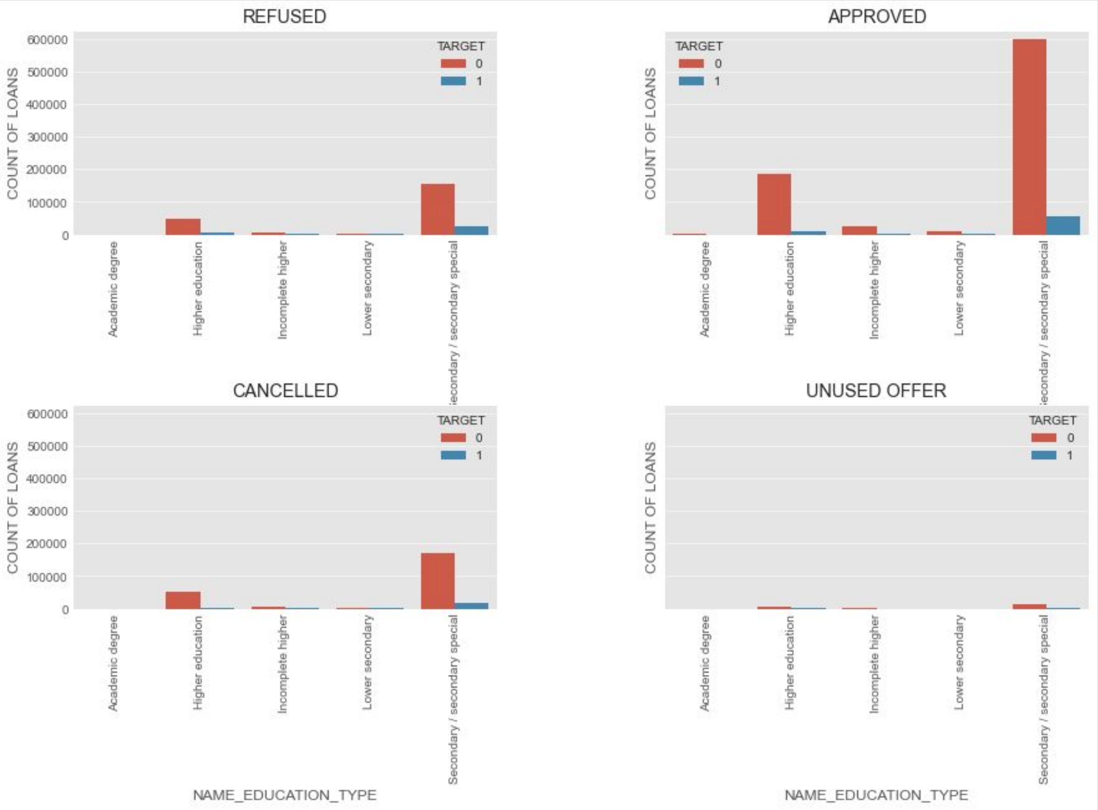
INFERENCE: 1) Clients Prefer Cash Loans and Consumer Loans Over other types. 2) People With Approved Consumer Loans have the highest Difficulty in paying the loan. 3) Cash loans have the highest Cancellation rate. As we saw earlier due to to Higher Annuity they attract.

Final Dataset - Count of Loans Vs NAME_INCOME_TYPE



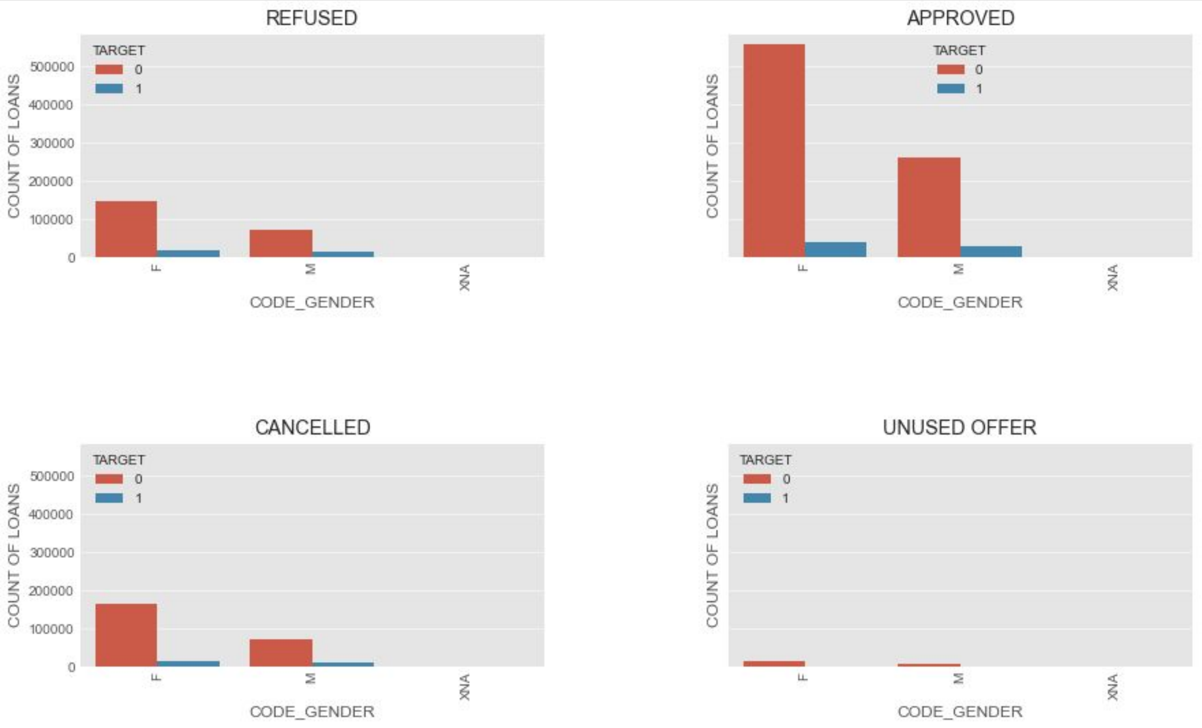
INFERENCE: Working Professionals have the highest number of Applications, along with Highest Number of Clients With Payment Difficulties. Hence More Focus should be put on other professionals. 2) State Servants have a very low default chance, hence more focus should be put here.

Final Dataset - Count of Loans Vs EDUCATION_TYPE



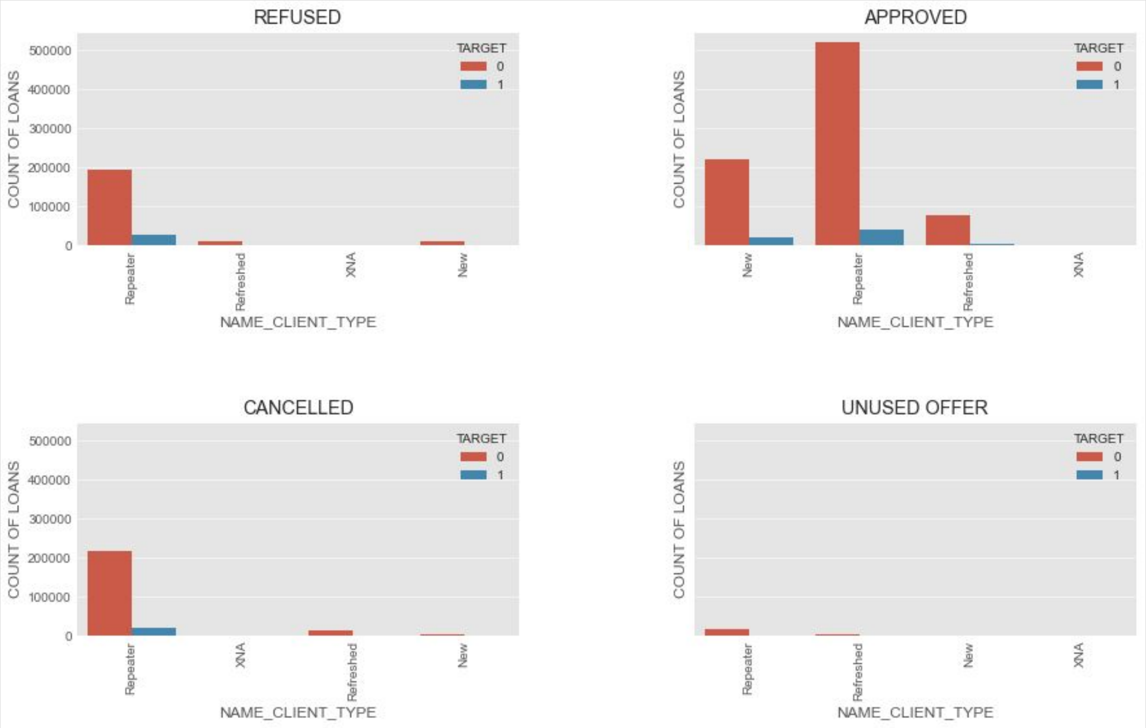
INFERENCE: 1) Secondary have the highest number of Applications, and for the Approved loans the Maximum number of Clients with Difficulties. 2) Higher Education Category has a low number of Clients with Difficulties, and a high refusal rate for Clients without Difficulties, Maybe focusing on this category will help bring more business

Final Dataset - Count of Loans Vs Gender



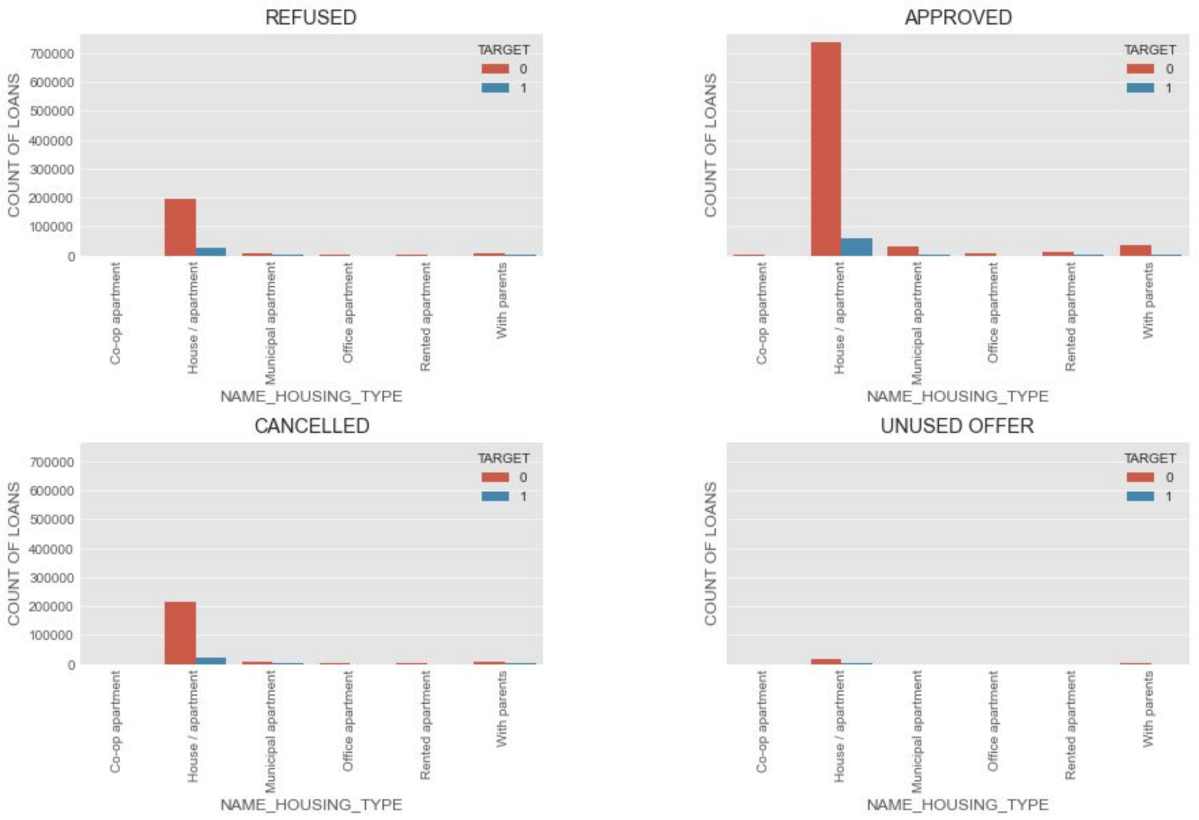
INFERENCE: Females, have higher lending applications and the lower default rates compared to males, hence Females, tend to get higher approvals as well.

Final Dataset - Count of Loans Vs Client_type



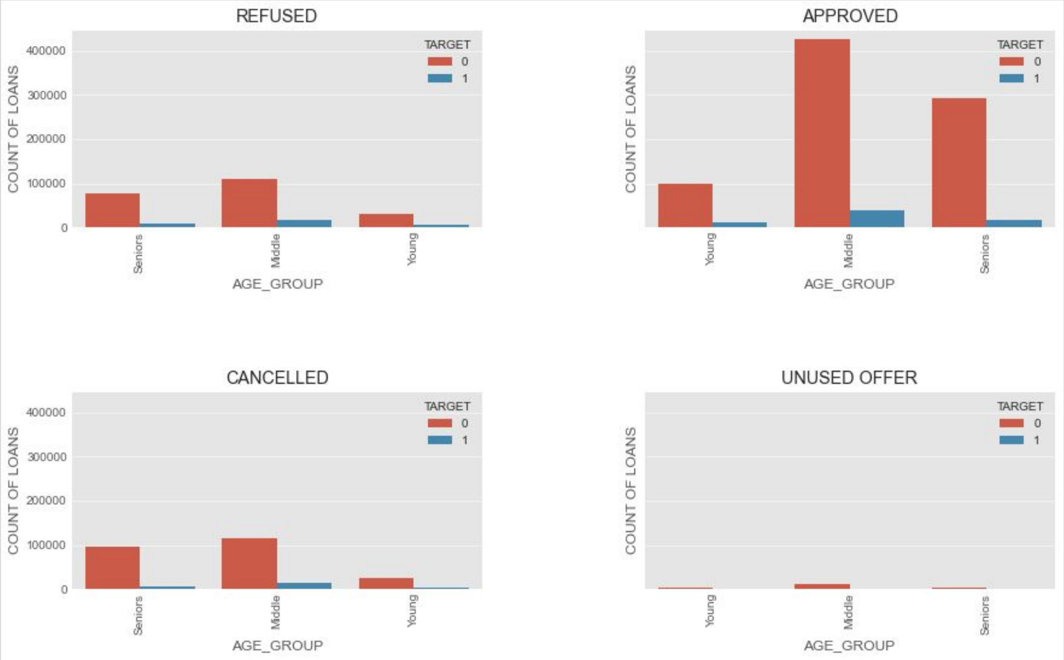
INFERENCE: Repeaters have higher chances of Approval, also the banks have refused the loan to repeaters who had trouble paying in the past.

Final Dataset - Count of Loans Vs Housing_Type



INFERENCE: Clients with House/Apartments have the highest Approval numbers, also Clients Staying with Parents have the least chance of Default.

Final Dataset - Count of Loans Vs AGE_GROUP



#Pct of Default in
#Young : 11.91%
#Middle: 9.42%
#Seniors : 6.47%

INFERENCE: Senior Citizens have the lowest Default Rates, and hence are suitable to get loans. Where as the Middle ages clients have the high number of Applications but at the same time **Have the highest Default Rates.**

Final Dataset - TARGET

INFERENCE: It is a good indicator that bank is approving more loans for Customers without payment difficulties, But at the same time a huge number of customers without payment difficulties were Refused, This might lead to loss of Business.

Who All Can Get A Loan (Less Chance to Default)

- ***Clients Working As State Servants***
- ***Senior Citizens have low default number***
- ***Females with Higher Education***
- ***Refreshed Clients with ‘Unused Offer’ Status.***
- ***Clients from High Income Group***
- ***Those Who did not Have any default previously.***
- ***Clients living with Parents are defaulting less.***

Who All are risky (High Chance of Default)

- ***Lower Secondary and Secondary Special have high default numbers***
- ***Working Professionals also have high Default Rates.***
- ***Males seek lesser number of Applications and yet have high chances of Defaulting.***
- ***Those who were refused the Loan Previously, will have a higher chance of default if Approved.***

Thank You!