Credit EDA Case Study

By

- **Ayush Abhinav**
- **Amber Kaushal**

Batch: IIITB C26 DS Nov 2020
Date of submission: March 01, 2021

Problem Statement

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. The company wants 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.

Case Study Objective

This case study aims to analyse the customer profile data to

- find patterns which indicates if client has any payment issues.
- Highlight the cases where company has denied loan to potential clients thus helps in avoid loss of business opportunity .

Contents

- Application Data Analysis
  - Load Data
  - Choose column of interest
  - Analyse columns for missing value
  - Check for outliers
  - Perform Validation check on data
  - Know the customer we are serving
  - Different Vs Target Relation
  - Univariate Analysis
  - Multivariate Analysis
- Previous Application Data Analysis
  - Load Data, missing value treatment
- Merge Previous Application Data with Current Application
- Multivariate Analysis on Merged Data
- Conclusion/Recommendation.

Application Data Analysis

- Load Data
  - Data was given in .csv file. It was load in Pandas dataframe for analysis
    - df = pd.read_csv('application.csv')
  - Data shape : There are 122 columns and 307511 rows of data

- Choose column of our interest
  - There are 122 columns. we have to identify the columns of our interest and then start our analysis.
  - Column of our interest are : 'SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'OWN_CAR_AGE', 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL', 'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS', 'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY', 'REG_REGION_NOT_LIVE_REGION', 'REG_REGION_NOT_WORK_REGION', 'LIVE_REGION_NOT_WORK_REGION', 'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CITY', 'ORGANIZATION_TYPE', 'EXT_SOURCE_1', 'EXT_SOURCE_2', 'EXT_SOURCE_3'

- Analyse columns for missing value
  - We will analyse the columns which has missing value and establish if it is MCAR, MAR, MNAR. We will list down method for handling such missing values.

  - OCCUPATION_TYPE and OWN_CAR_AGE have considerable share(31 and 66 respectively).Other columns with missing values has less 1% share and should not affect our analysis.

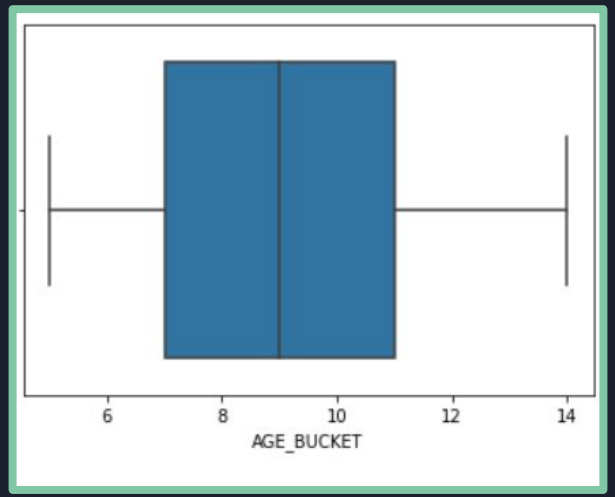
  - We checked if missing in OCCUPATION_TYPE has any relationship with Name_INCOME_TYPE. We found that NAME_INCOME_TYPE is similar for both missing and not missing OCCUPATION_TYPE and has no relation with missingness of OCCUPATION_TYPE. Hence Occupation_TYPE is MCAR. Asit is more than 13% missing value, we can drop this column. NAME_INCOME_TYPE gives the same intuition as OCCUPATION_TYPE

  - Code_Gender has missing value as XNA. Those rows were dropped from dataframe.

- Check for outliers

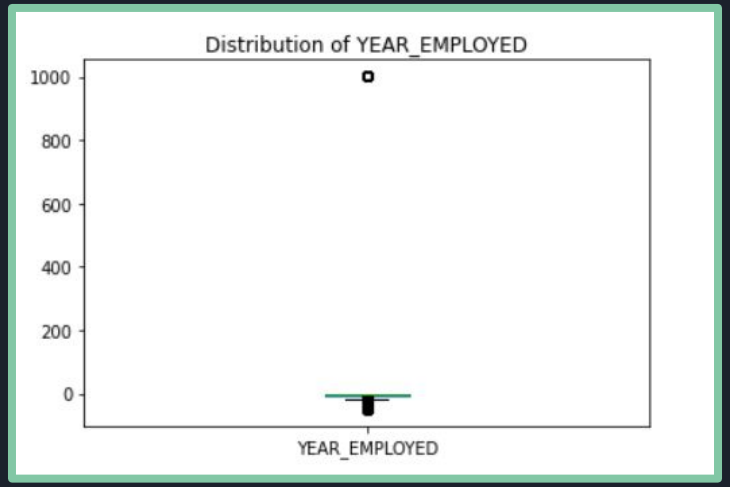
We will check for the outliers in data and establish if they are legitimate data or outlier because of data collection errors. We will also drop the outliers where necessary.

AGE_BUCKET



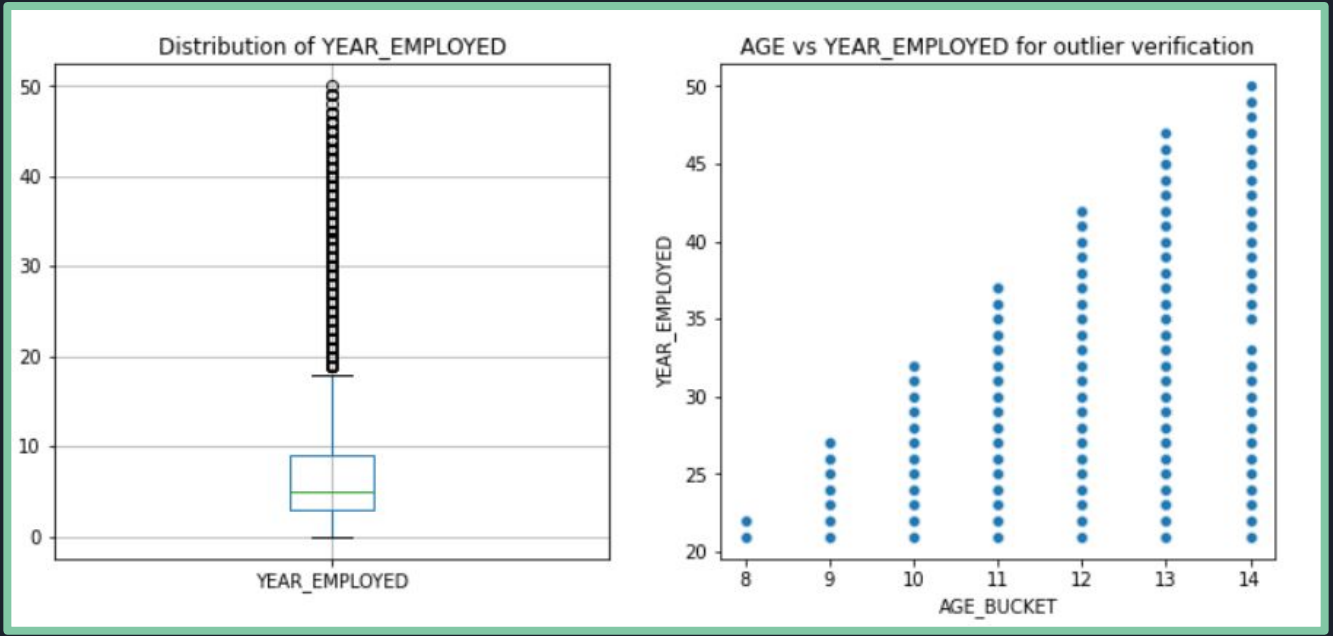
Does Not look to have outlier

Distribution of YEAR_EMPLOYED



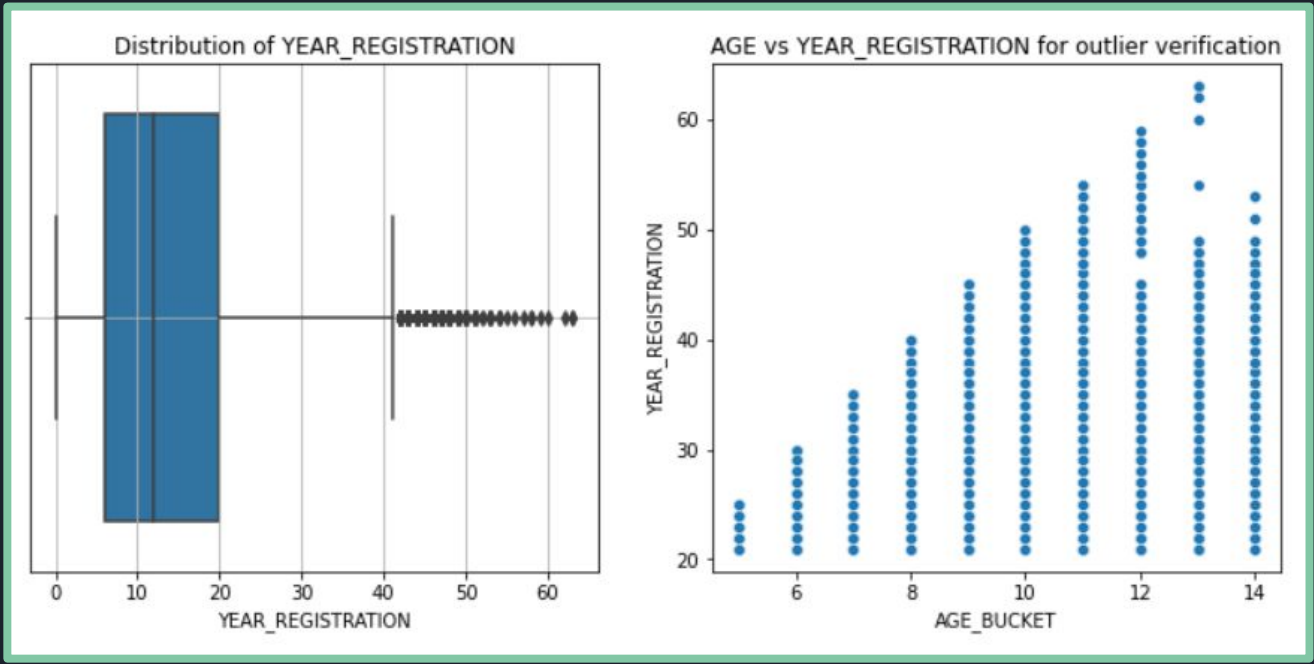
There are outlier. With the value of outlier, it appears to be a data collection issue. Need to drop such records

Distribution of YEAR_EMPLOYED



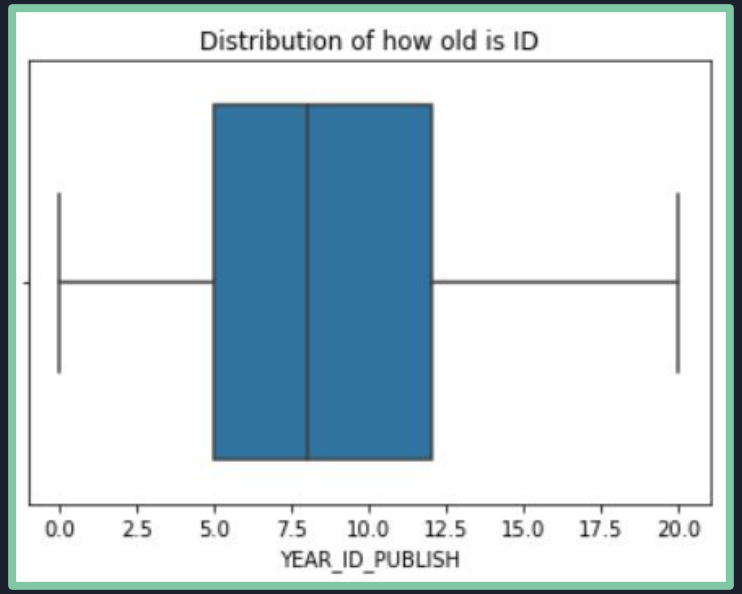
Mass of year employment lies in range 4-9 years. Also the box plot shows outlier values but these values are continuous and looks legitimate as aged person has more year of experience. Hence it should be included in our analysis

Distribution of YEAR_REGISTRATION



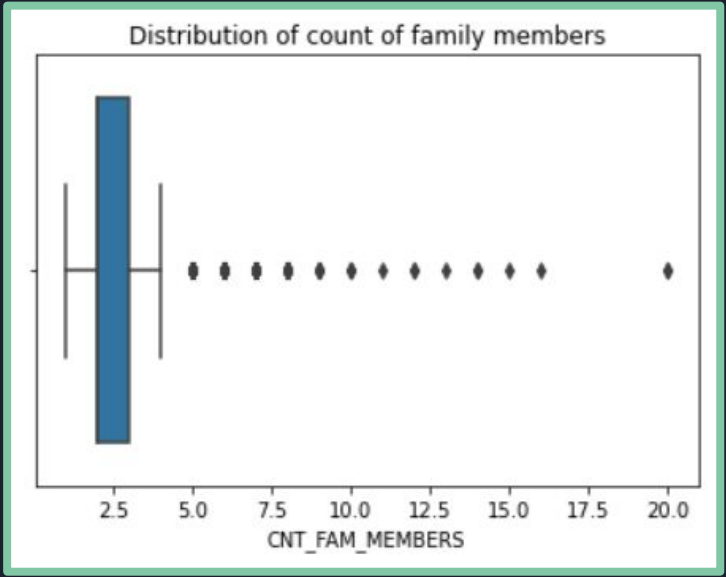
Mass of year registration lies in range 6-20 years. Also the box plot shows outlier values but these values are continuous and looks legitimate as aged person can get his doc registered earlier. Hence it should be included in our analysis

Distribution of YEAR_ID_PUBLISH



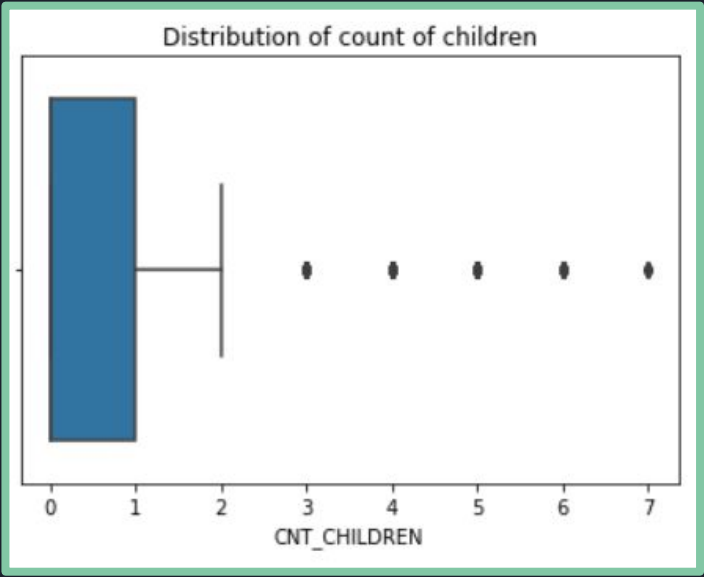
Mass of id published lies in range 5 to 12 year old id. No outliers as such

Distribution Of Family Member



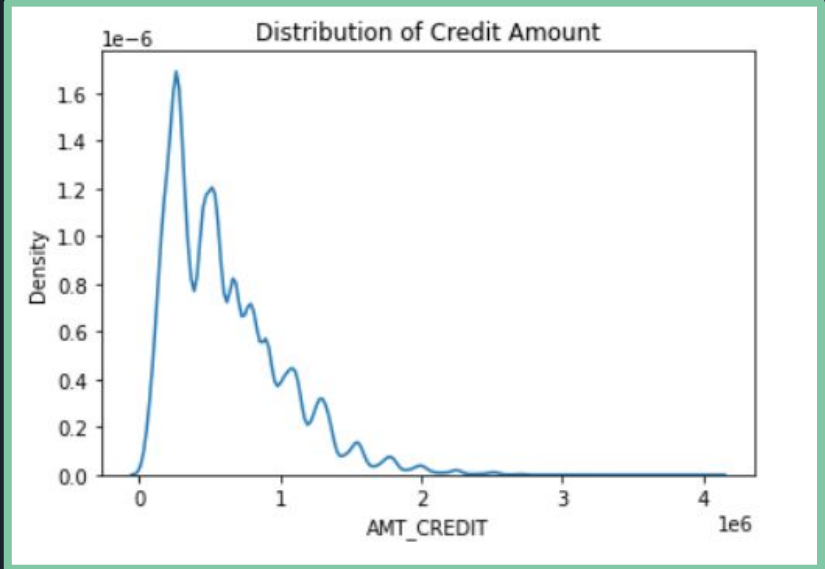
People with 9 family members can be considered as joint family. But above that are outliers and can be dropped from analysis

Distribution Of Children



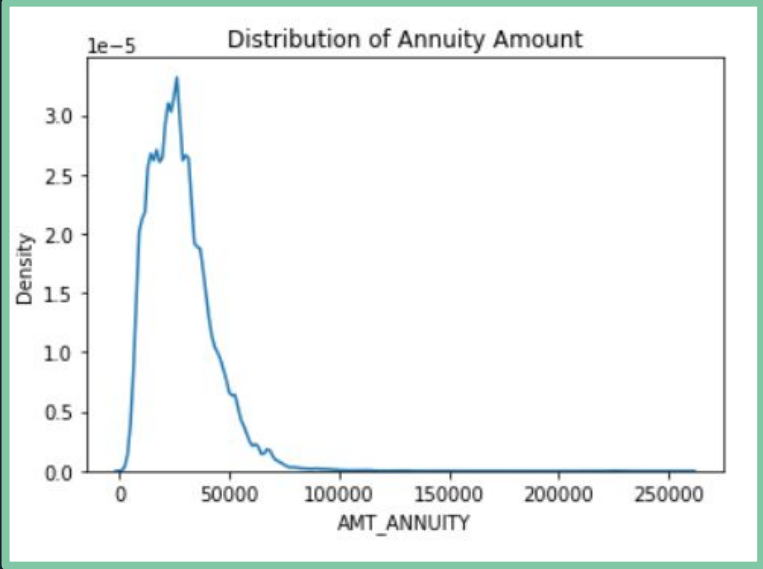
The mass lies for 0 and 1 child. But there are outliers but these values seem to be legitimate

Distribution of Credit Amount



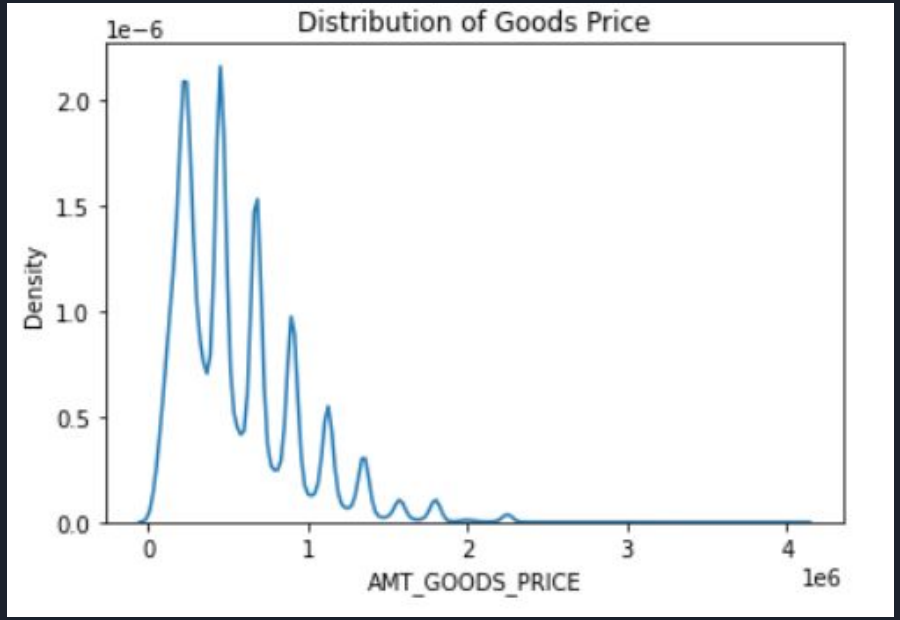
Data is multimodal and right skewed. Maximum people has requested for lower credit amount

Distribution of Annuity Amount



AMT_ANNUITY is right skewed. Mass of distribution is towards the lower value of Annuity

Distribution of Goods Price



This is a multimodal distribution. It seems there are particular goods/good's price for which loans are applied more.Also data is right skewed.

- Perform Validation Check on Data

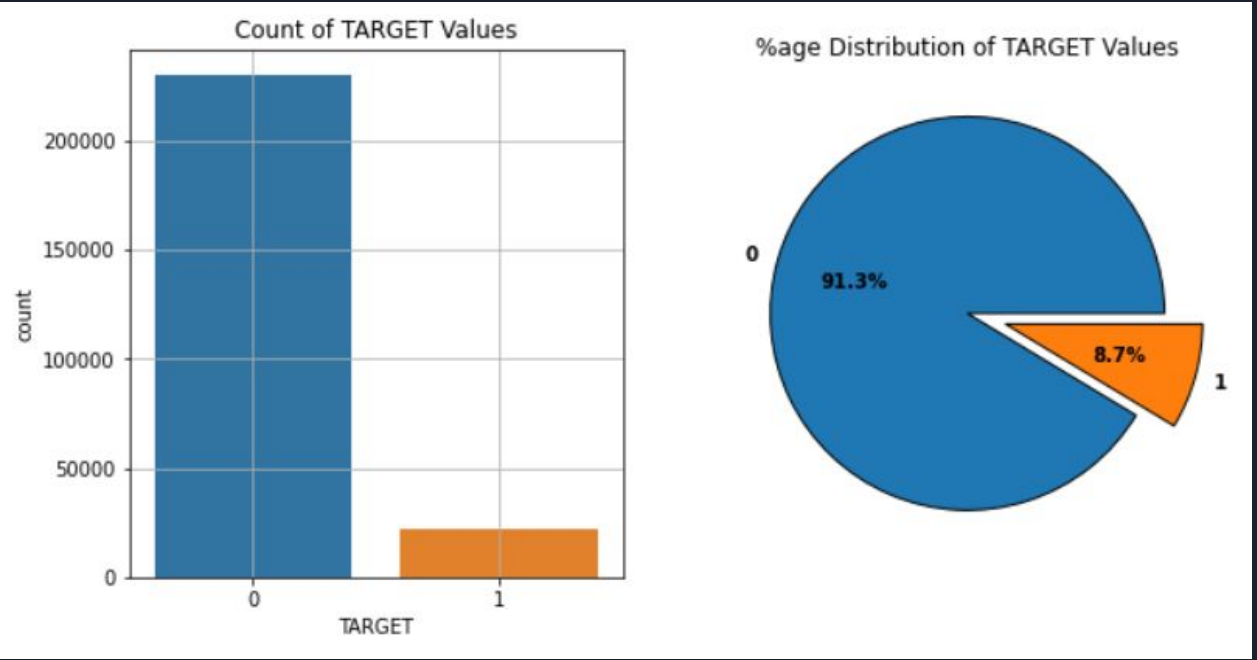
Following two validation checks were performed on data

- Age should be greater than YEAR_EMPLOYED, YEAR_REGISTRATION, YEAR_ID_PUBLISH
- Count of family member should be greater than /equal to count of children

All Data passed above two checks.

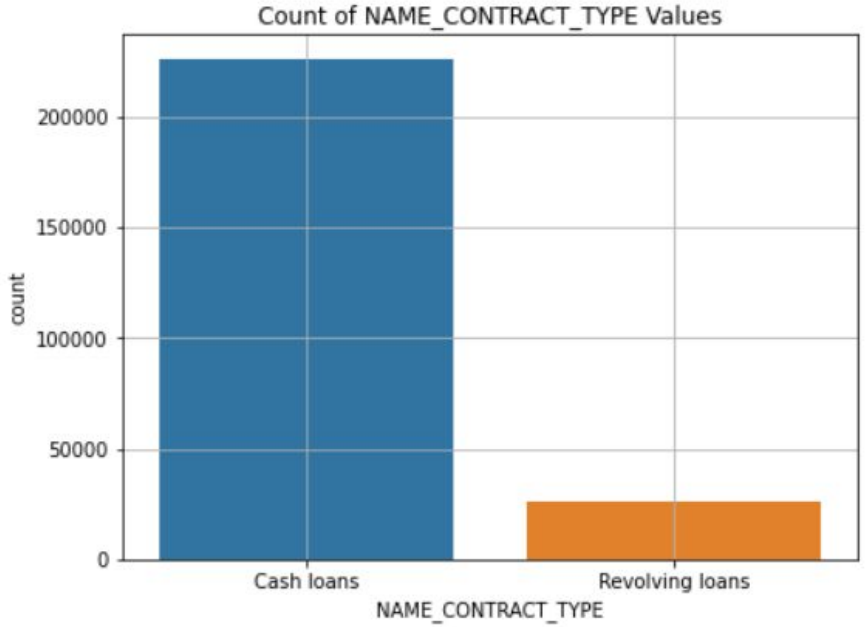
Know the customer we are serving

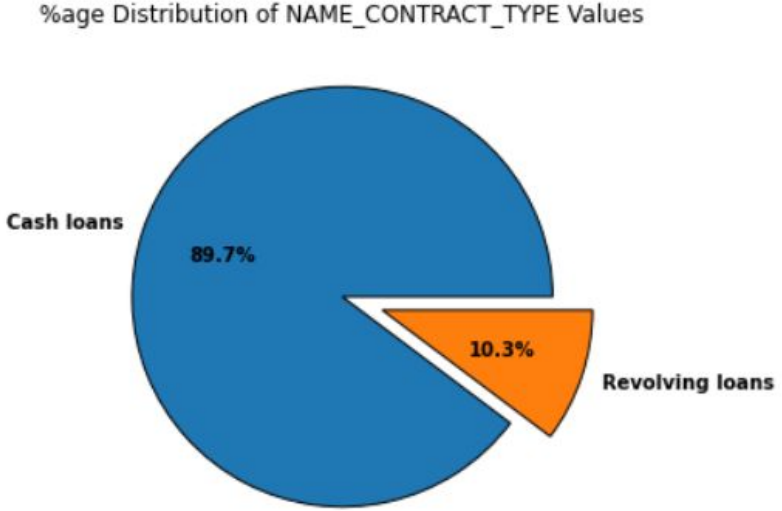
Distribution of TARGET Values



Circa 9% of data is of client with payment difficulties

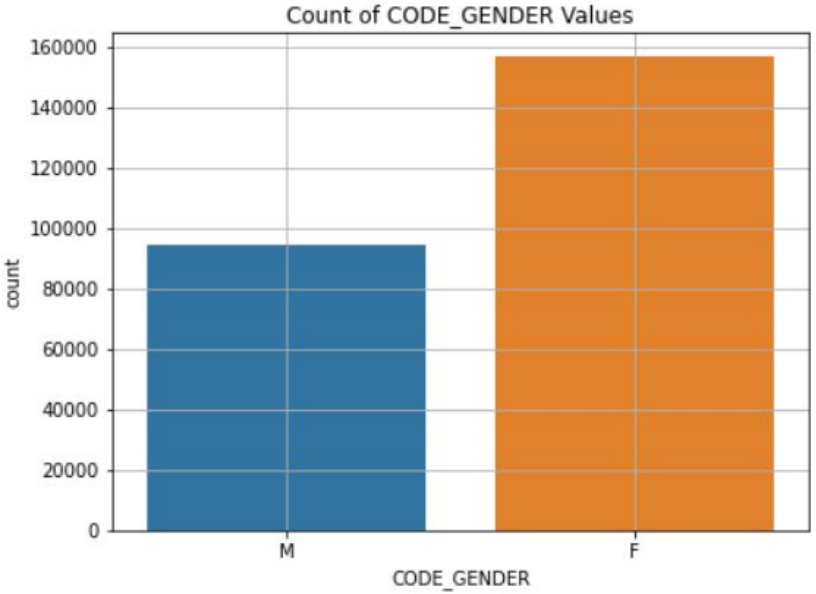
Distribution of NAME_CONTRACT_TYPE values

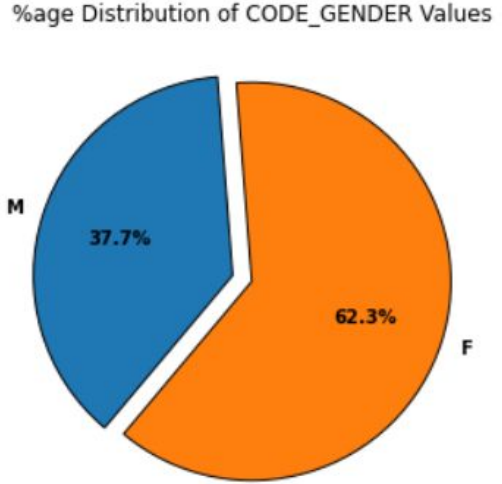




Maximum data is of Cash loans. Revolving loans is only 10.3%

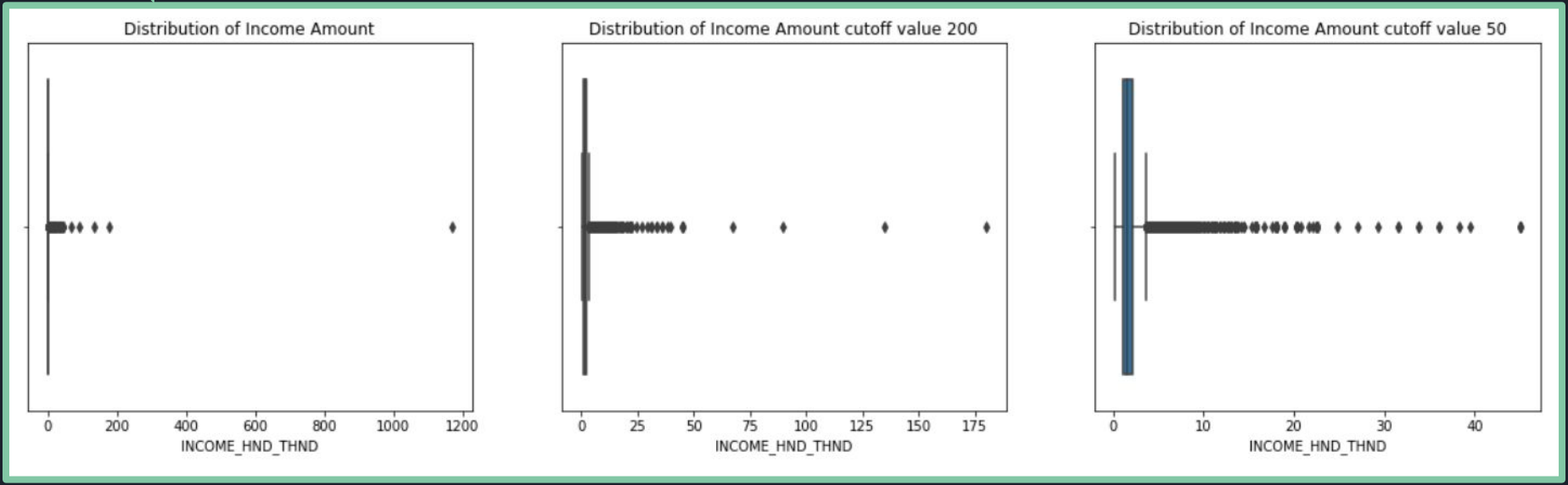
Distribution of CODE_GENDER values





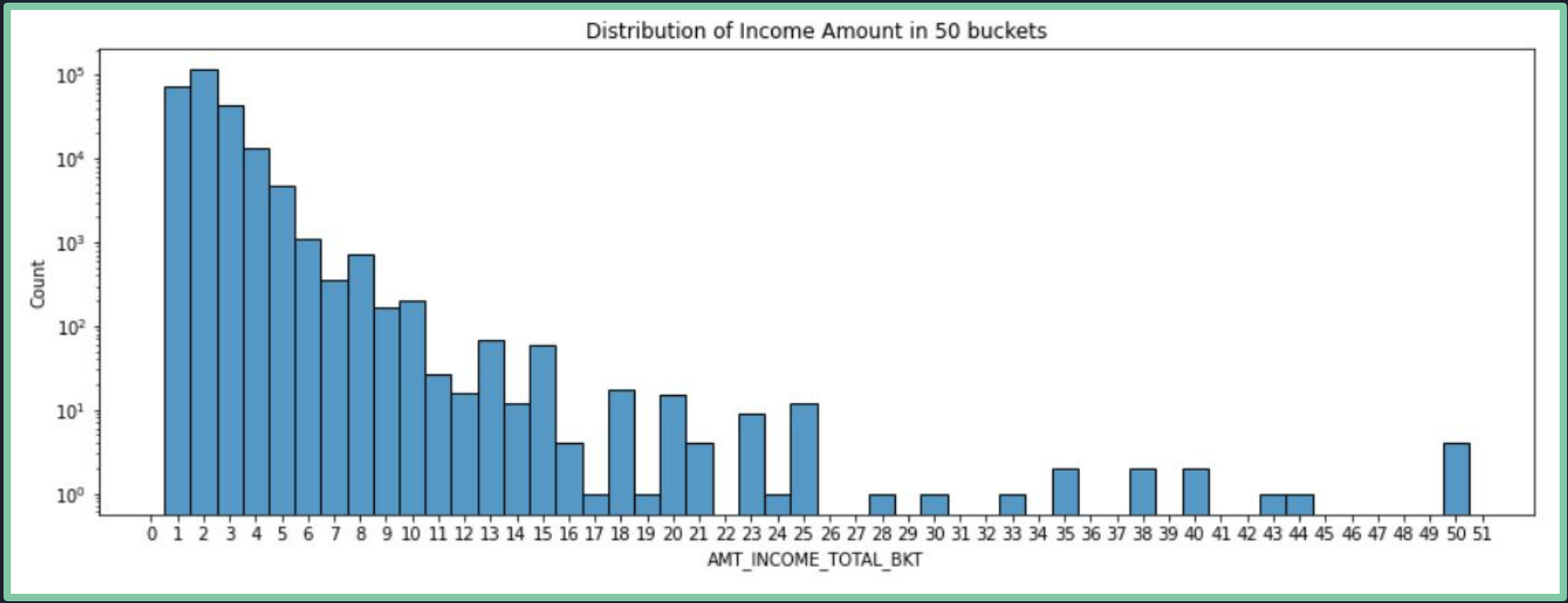
Twice Females as compared to Males

Distribution of AMT_INCOME values



There are outliers in data that need to be removed.

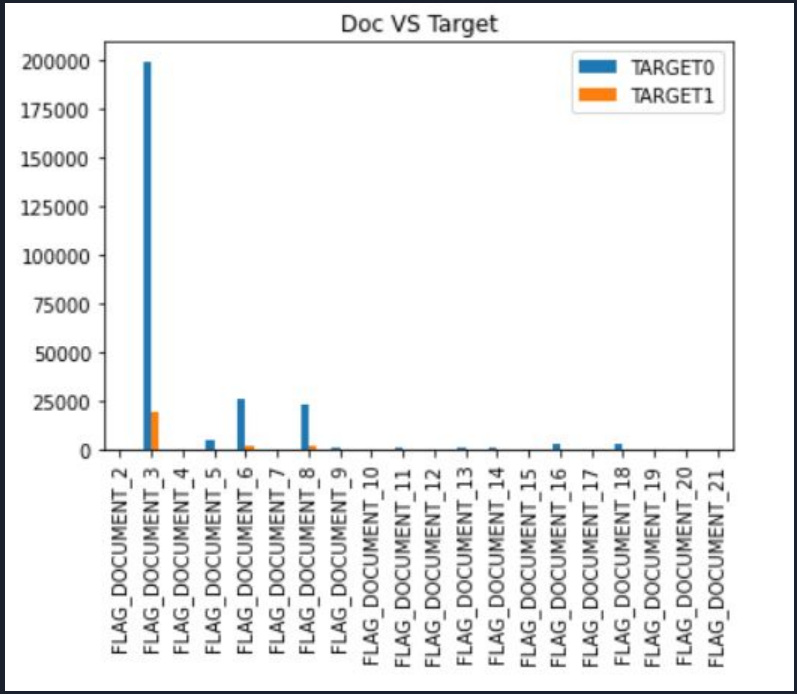
Bucketing the Income into 50 buckets



Some observable points are:-

- 90% of business is with cash loans
- 2/3 of clients are female
- 99% of clients have income less than equal to 25,00,000
- 9% of clients have payment difficulties

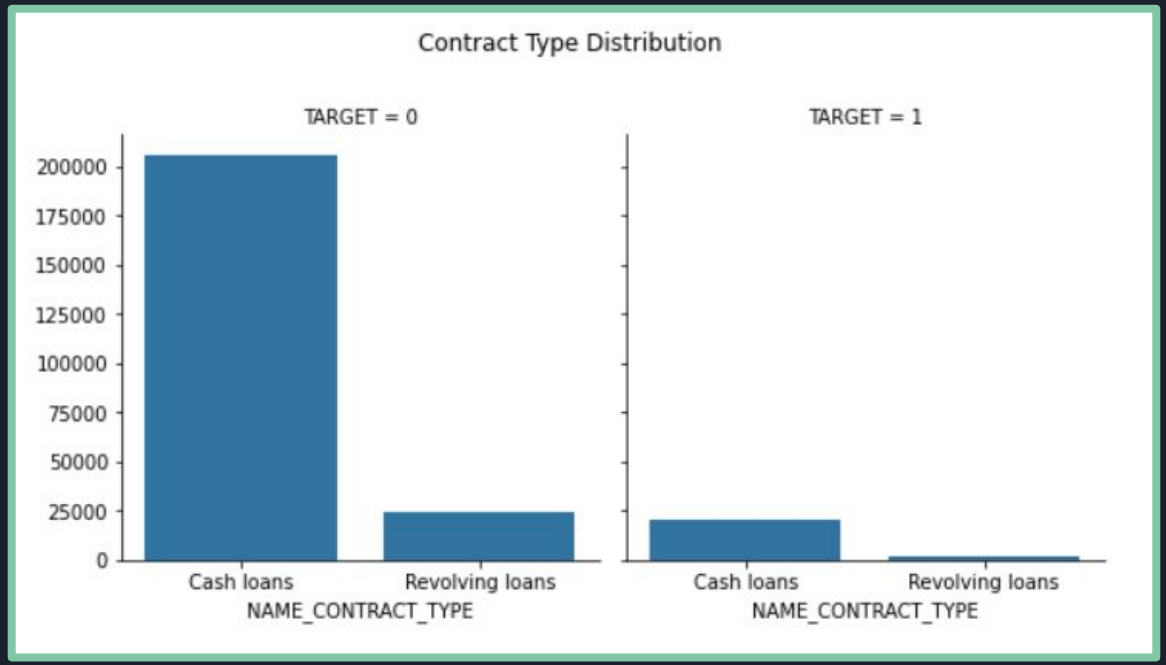
Doc Vs Target relation



Doc 3, 6, 8 are most submitted by loan applicant.

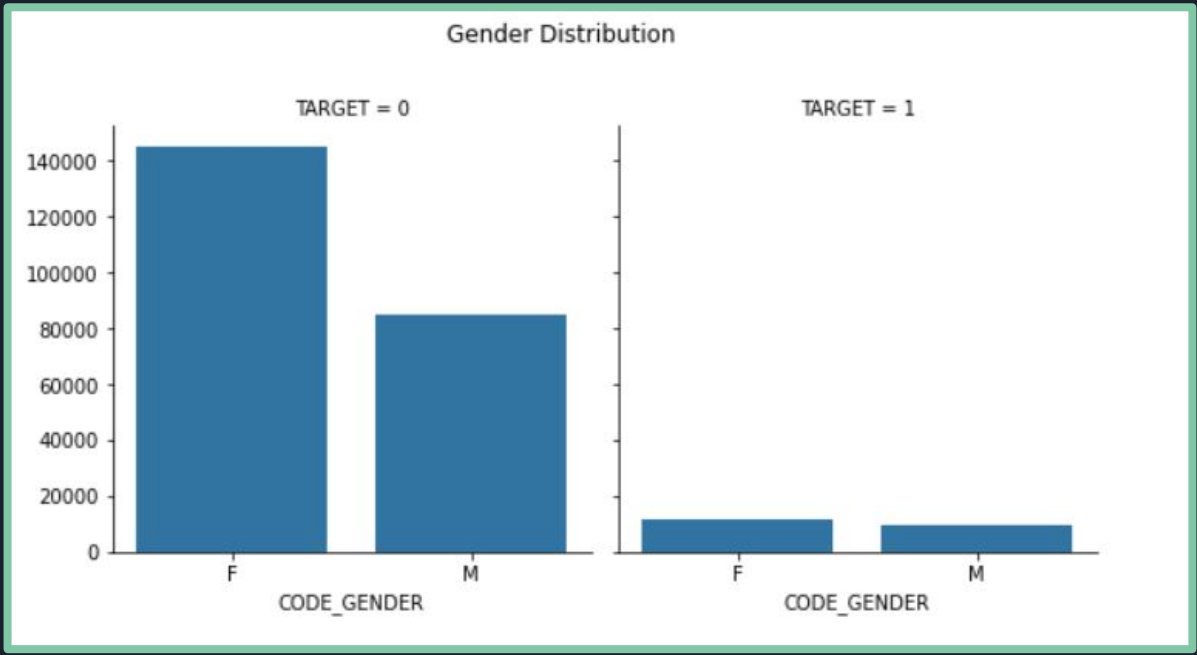
Univariate Analysis

Contract Type Distribution



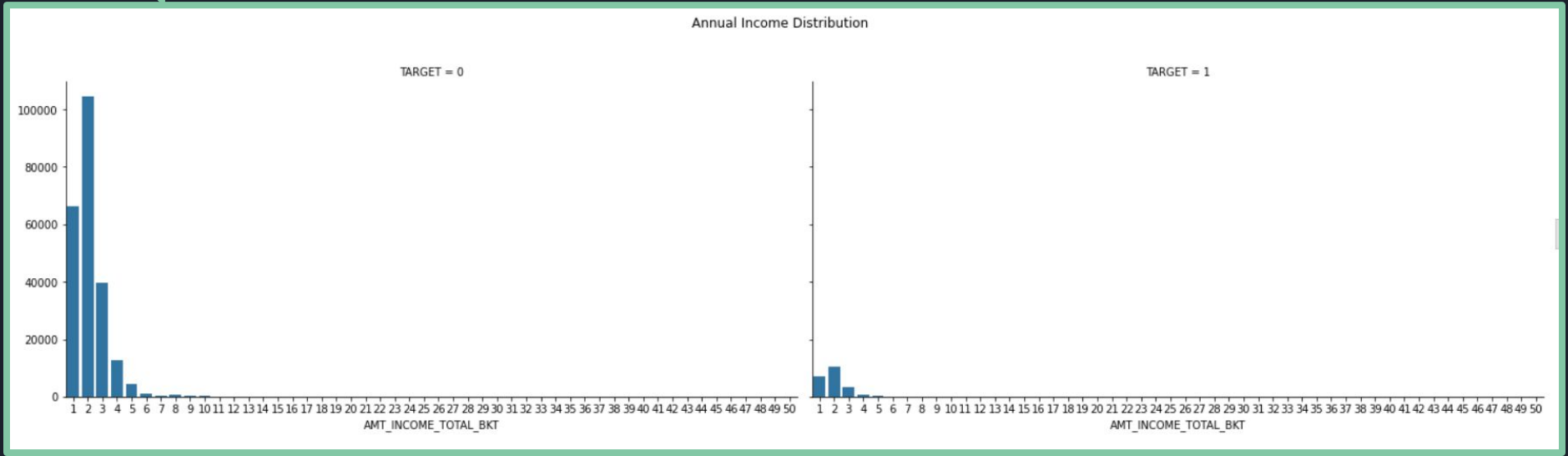
Cash loan has been preferred by customers both with and without payment difficulties

Gender Distribution



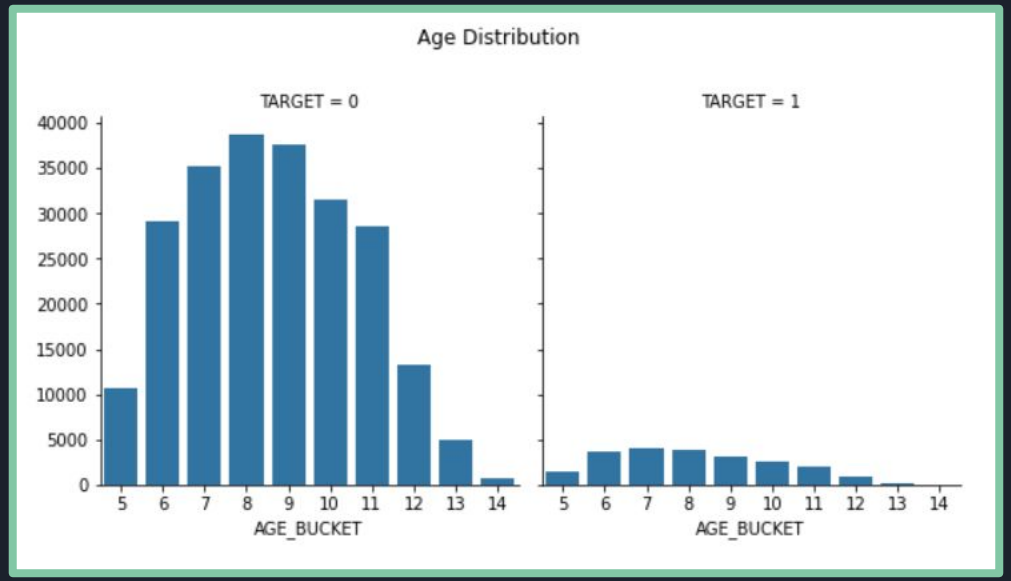
More number of female has take loan. Comparatively more males have payment issues

Annual Income Distribution



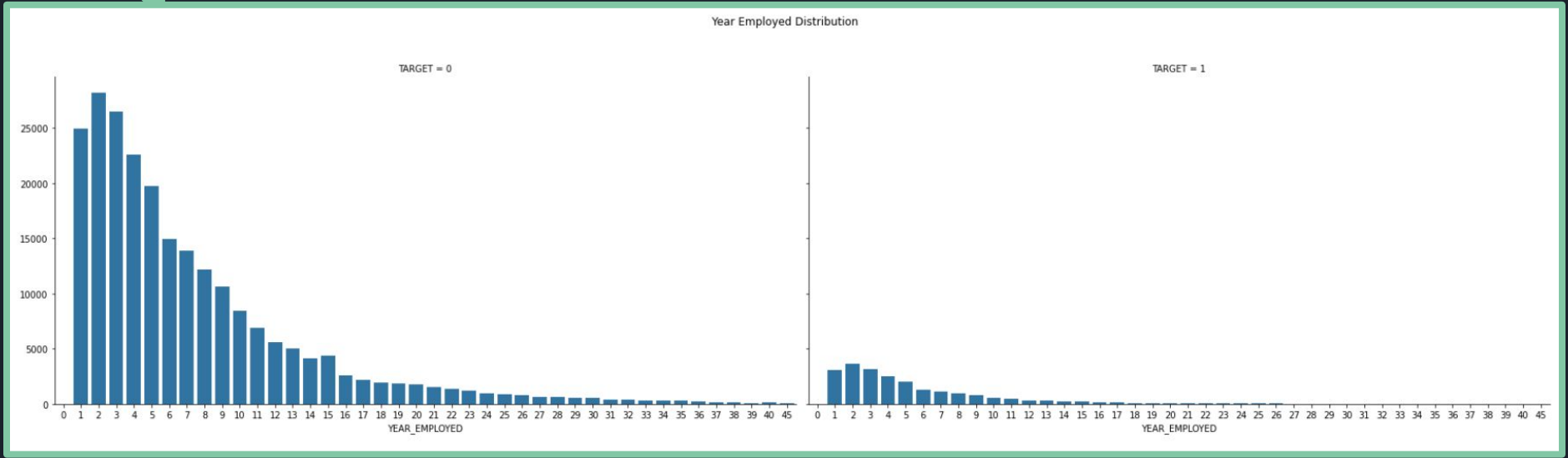
Most customers have income in 2nd bucket. Customers with low income have payment difficulties

Age Distribution



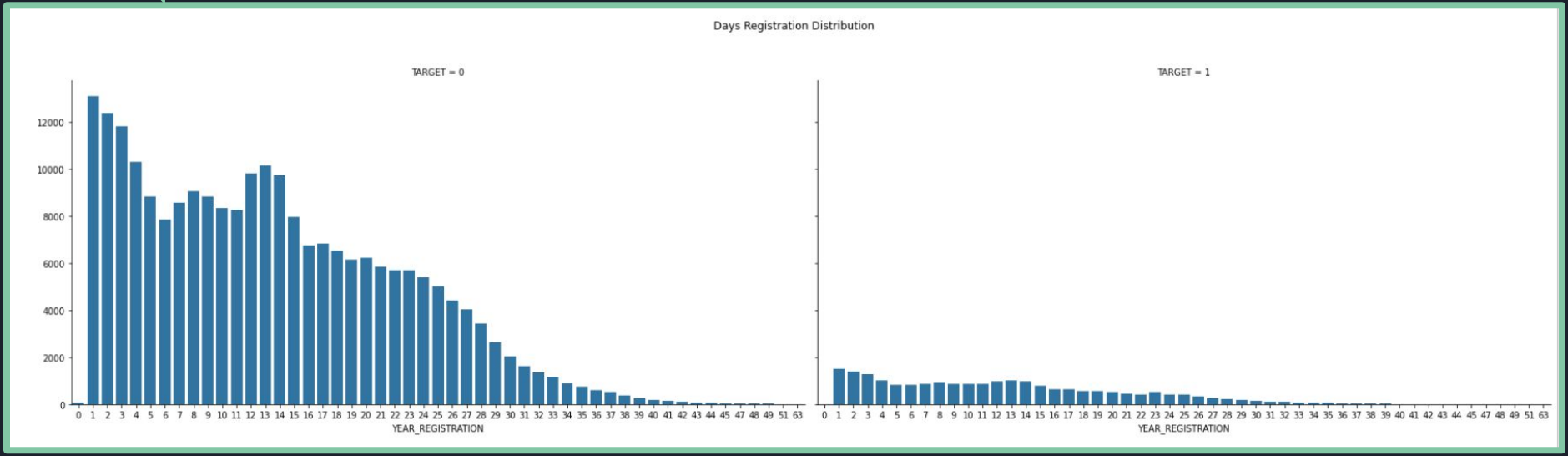
Age column has same pattern for both target 0 and target 1 customer.

Year Employed Distribution



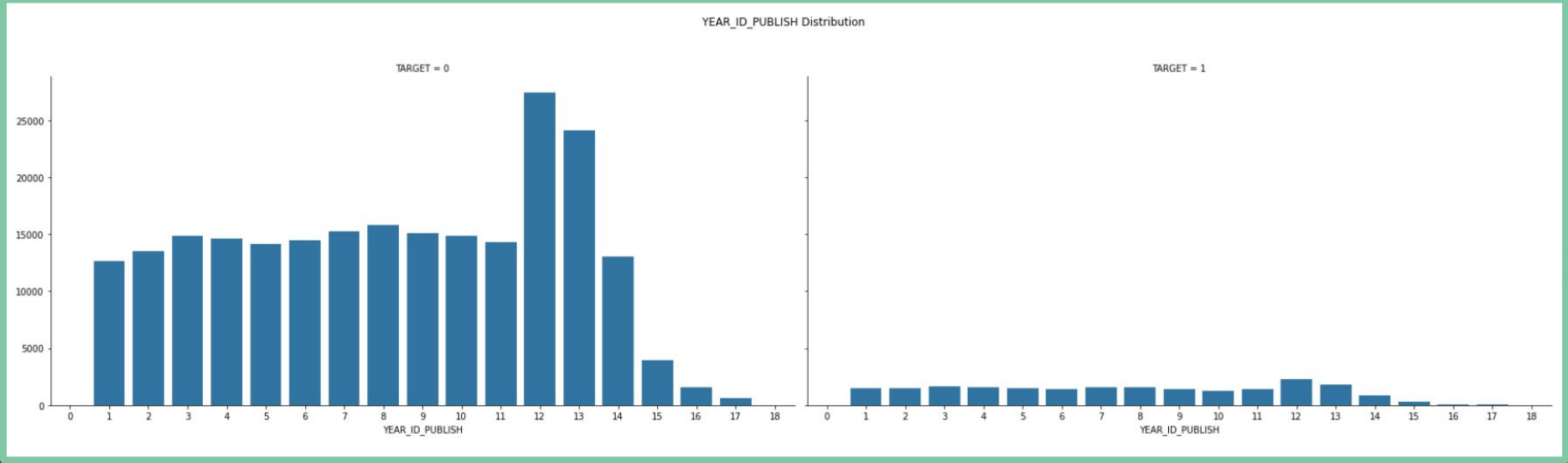
Most customer has less year of employment. Pattern is same for target 0 and target 1.

Days Registration Distribution



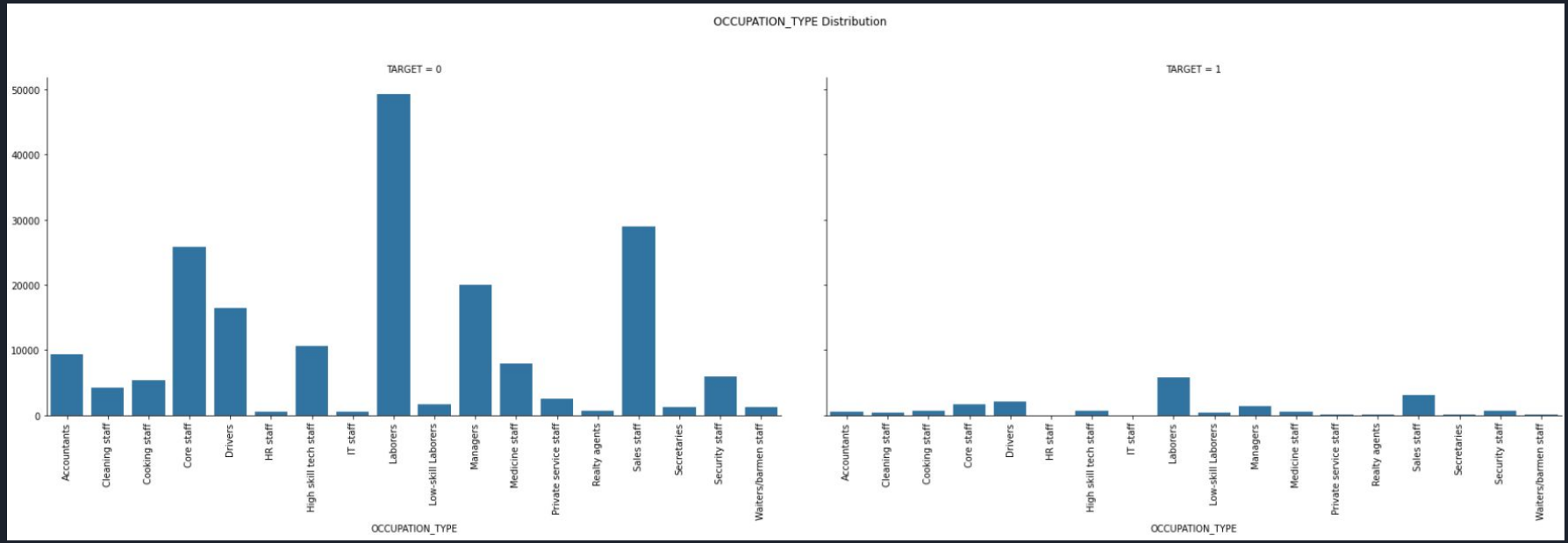
Days registration pattern is same for both target0 and target1 customer.

YEAR_ID_PUBLISH Distribution



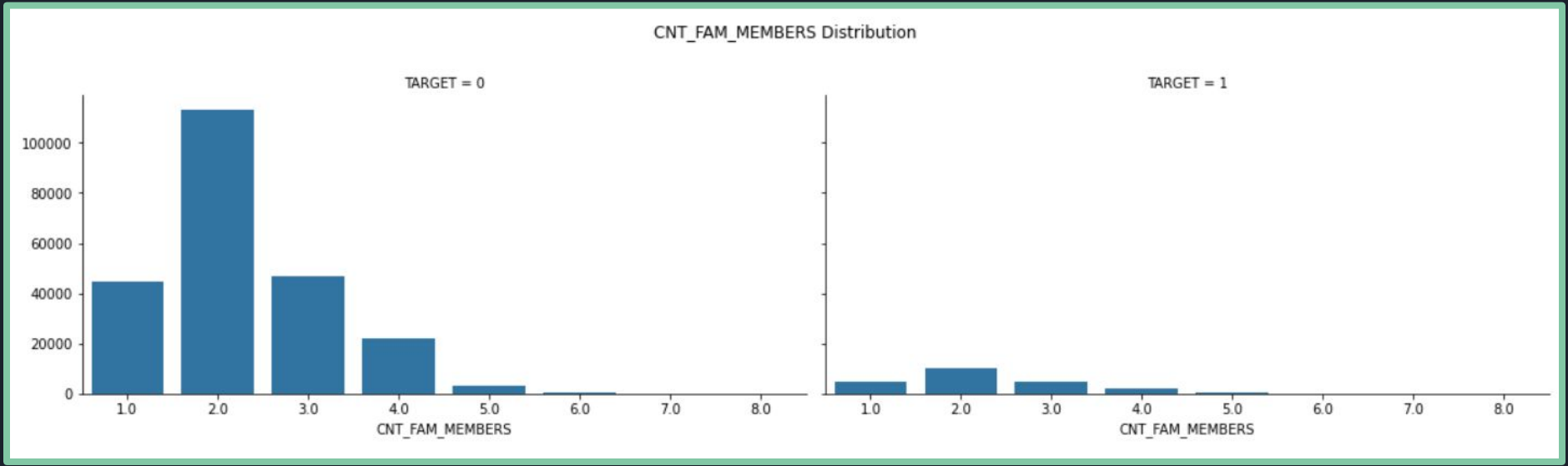
YEAR_ID_PUBLISH pattern is same for both target0 and target1 customer

Occupation type Distribution



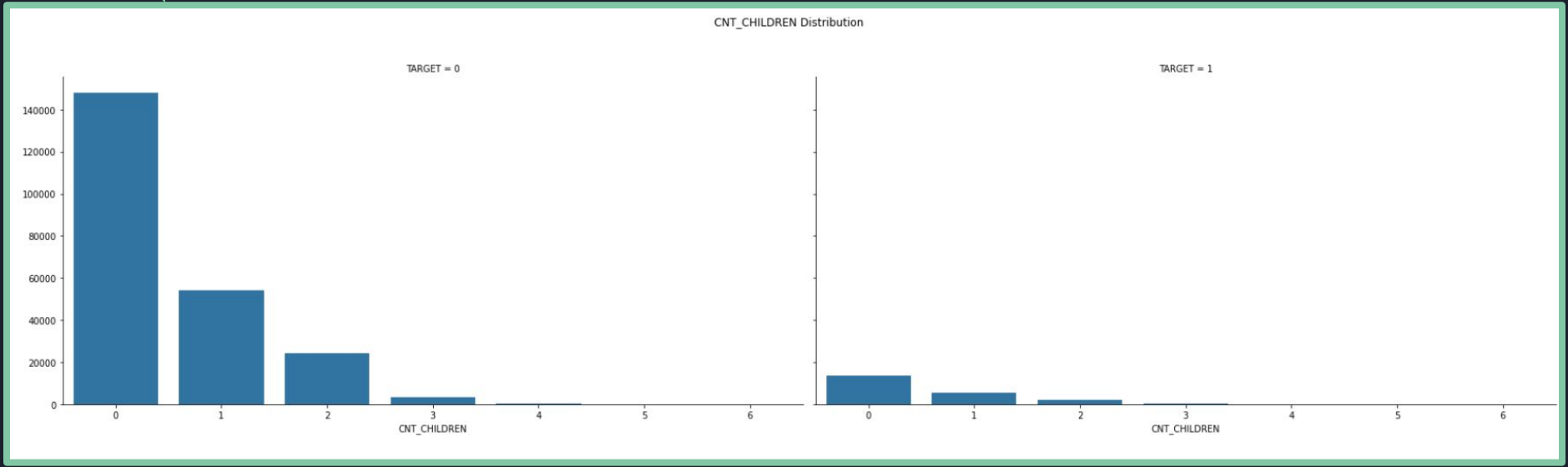
Occupation type pattern is same for both target 0 and target 1

CNT_FAM_MEMBERS Distribution



Family Size Distribution pattern is same for both target 0 and target 1. Very few people has family size 5 or greater. Most loan is taken by couple - Family size 2.

CNT_CHILDREN Distribution

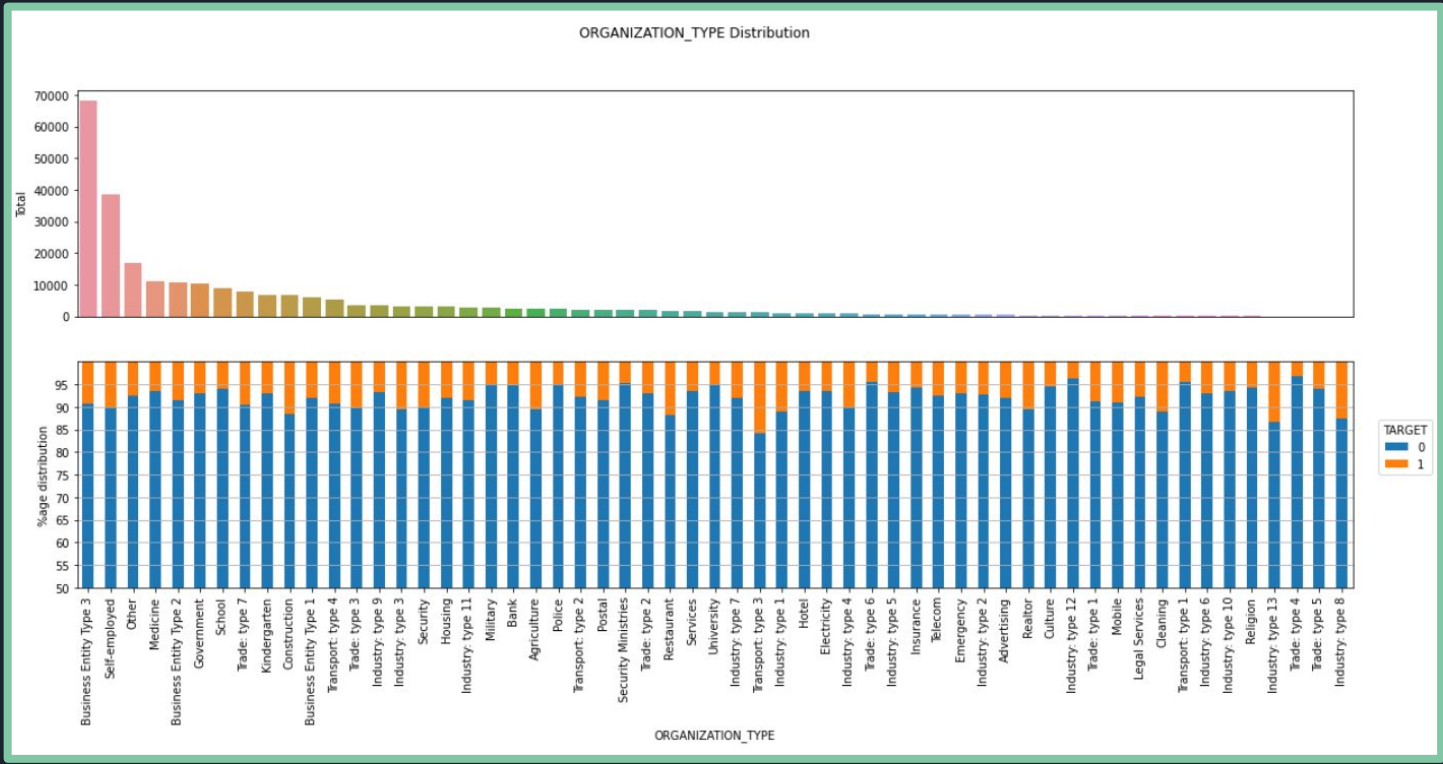


Children count distribution pattern in same for both target 0 and target 1.

REGION_RATING_CLIENT Distribution

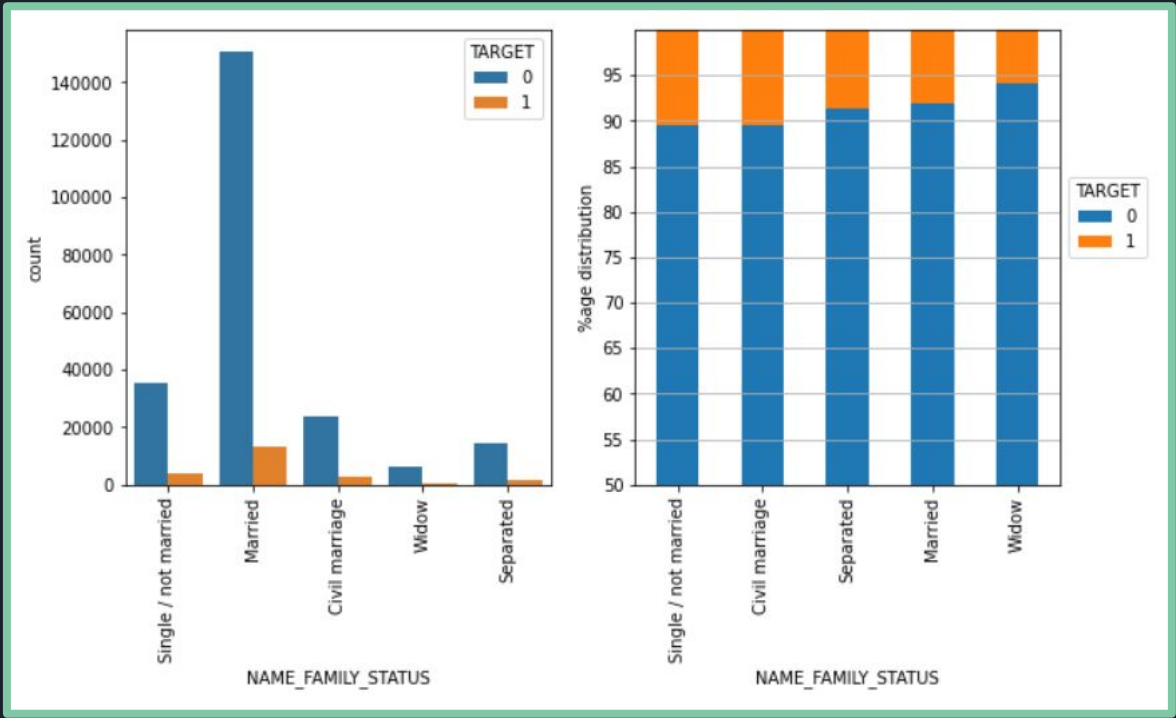
More people from region 2 has taken loan. More Payment issues has been recorded in region 3(~ 12%)

ORGANIZATION_TYPE Distribution



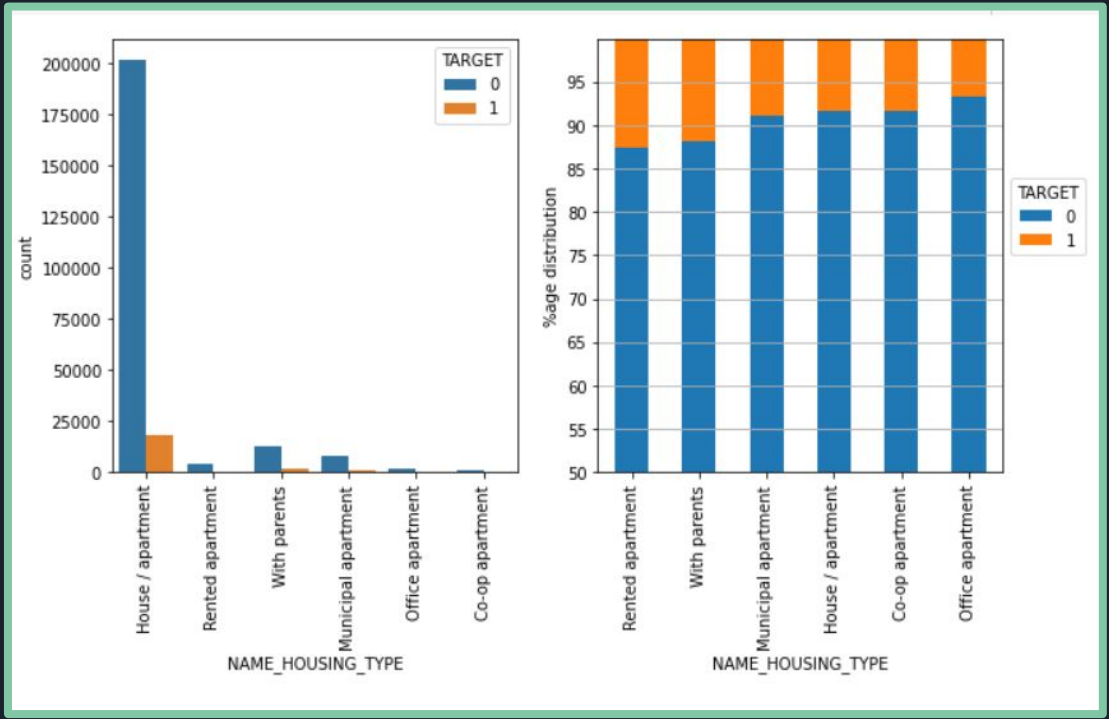
- Top two Organization type from where maximum people has taken loan are Business Entity Type 3 and Self Employed
- Payment issues for above two business type is 10% almost.

NAME_FAMILY_STATUS Distribution



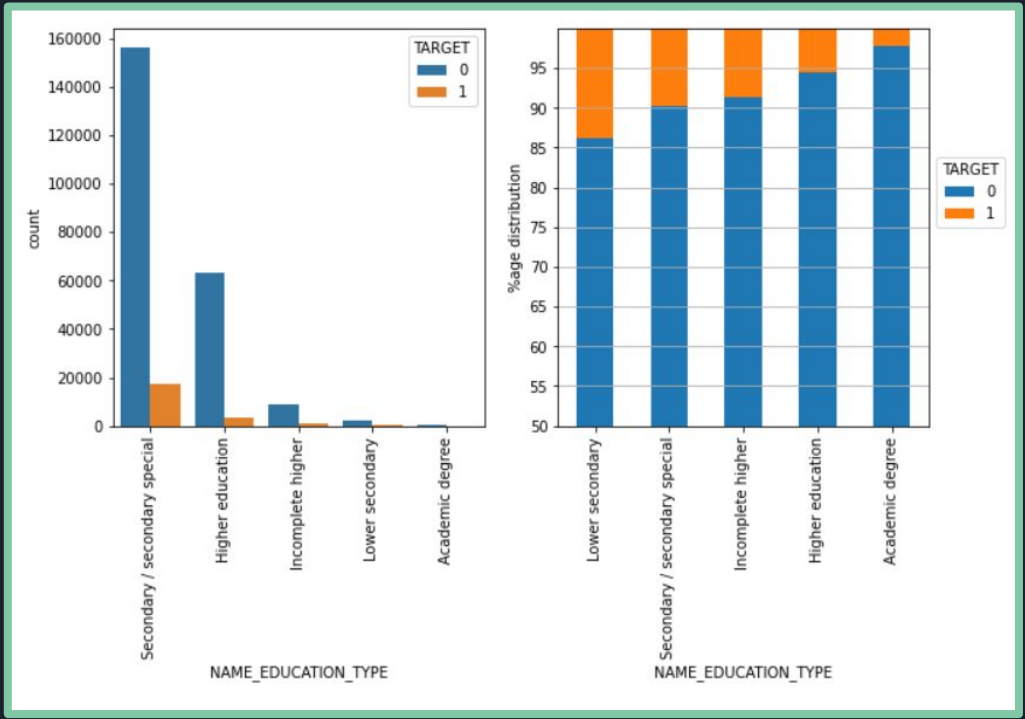
- More number of Married people has taken loan
- Also payment issue with them are less

NAME_HOUSING_TYPE Distribution



- More number of people living in House/apartment has taken loan. Pattern is same for both target 0 and target 1 customer. Less than 10% of them has payment issue.

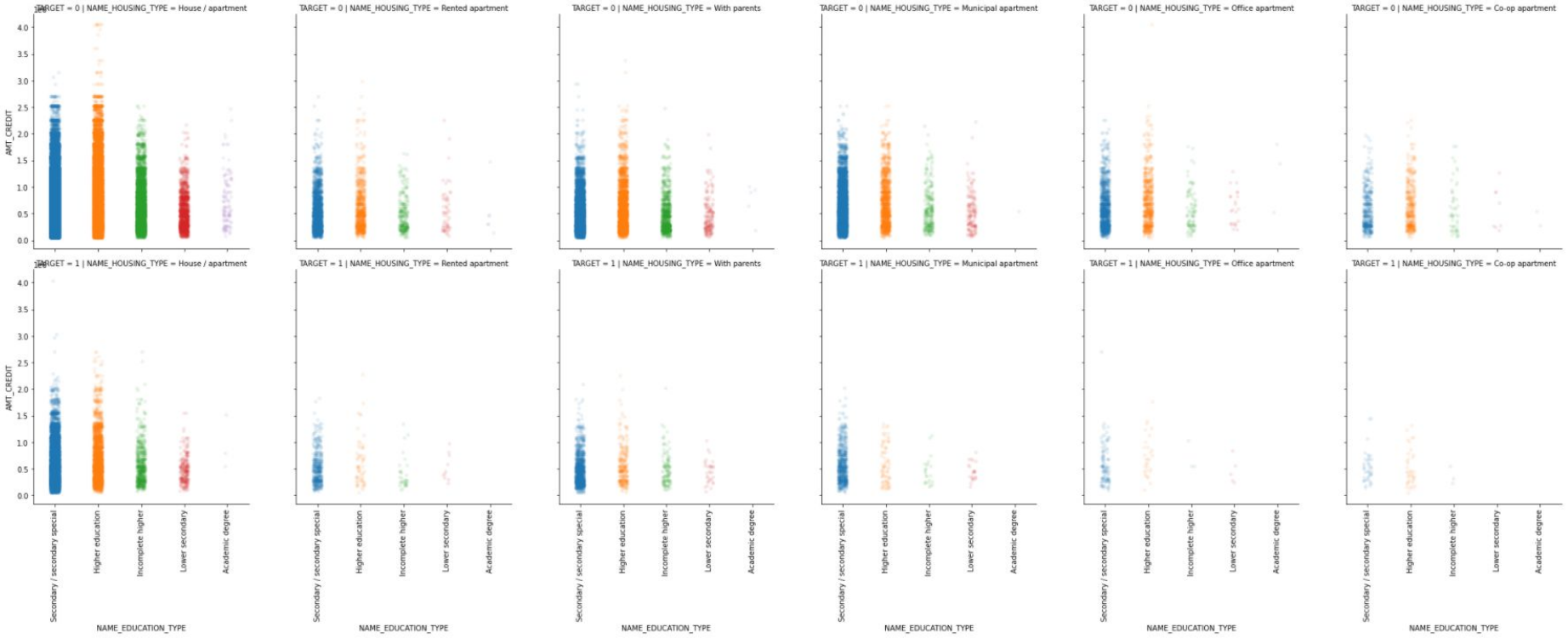
NAME_EDUCATION_TYPE Distribution



- Maximum Secondary education people has taken loan with 10% having payment issues. Higher education has second highest no of loan taker with 6% having payment issues.

Multivariate Analysis

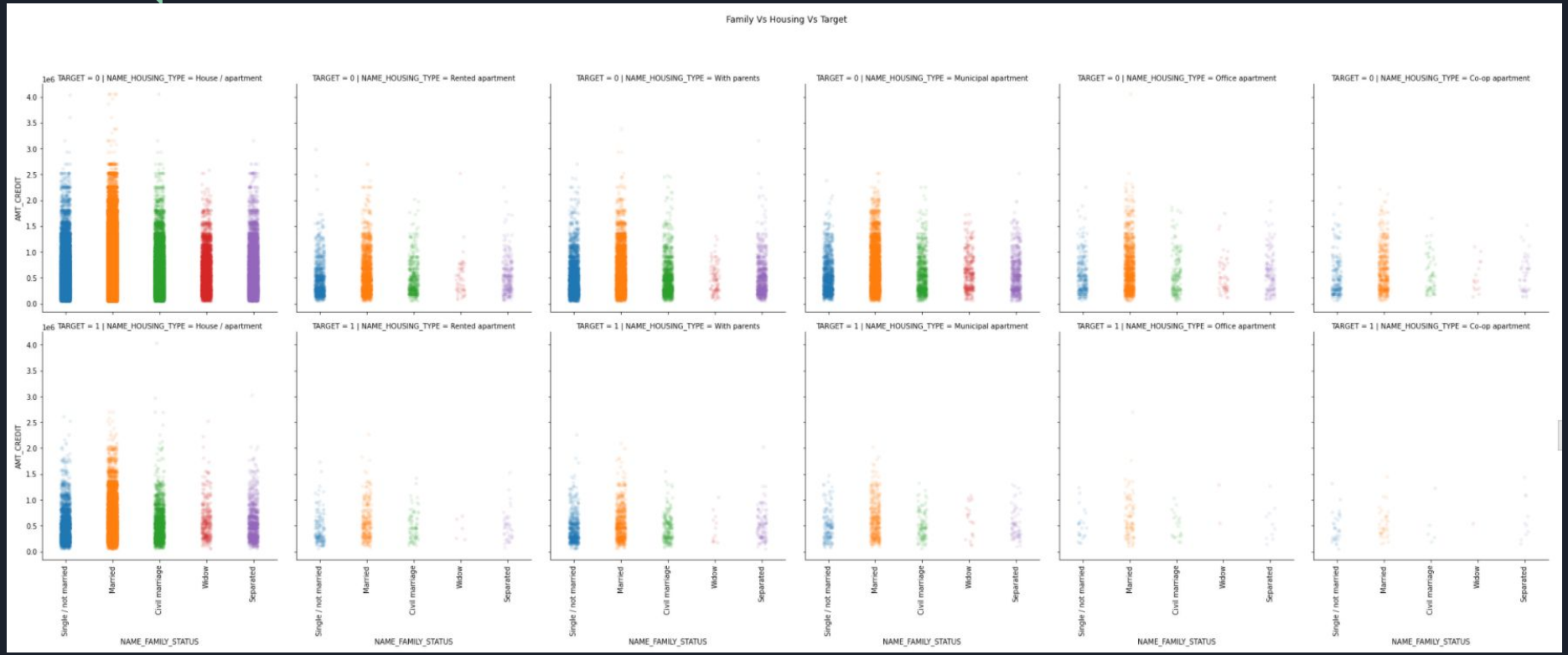
Education vs Housing vs Target



Education vs Housing vs Target

- Amount credit is higher for higher education people followed by Secondary Education.
- Fewer people with education as Academic degree and lower secondary has taken loan.
- Our customer living preferences are House/Apart > With Parents > Municipal Apart > Rented Apart > office Apart > Co-op Apart
- For all segment for combination of Education type and Living preference, people without payment difficulties are more than those with difficulties.So, we are not able to conclude, which segment to avoid based on above segregation.

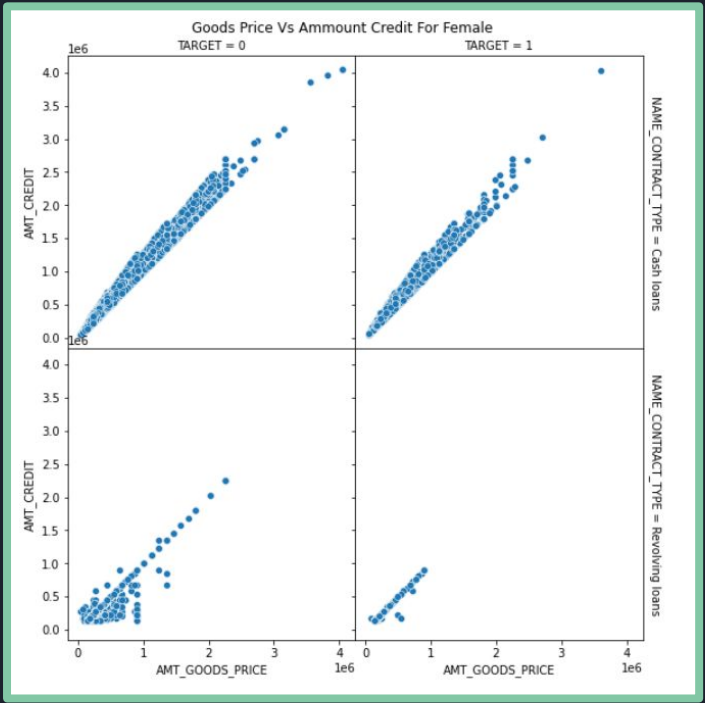
Family vs Housing vs Target



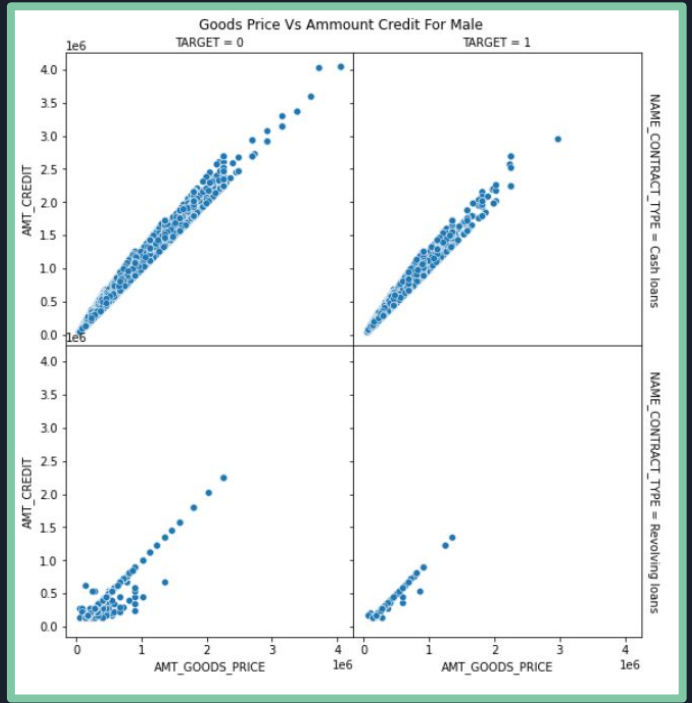
Family vs Housing vs Target

- Our Customers are mostly from House/Aprt living preferences.
- Among them, married people has taken maximum amount credit.
- Over all married people are out major customer in all housing segment.
- The pattern is similar for both target - 0 as well as target-1. On this data,we cannot figure out which segment to avoid.

Goods Price Vs Amount Credit For Female

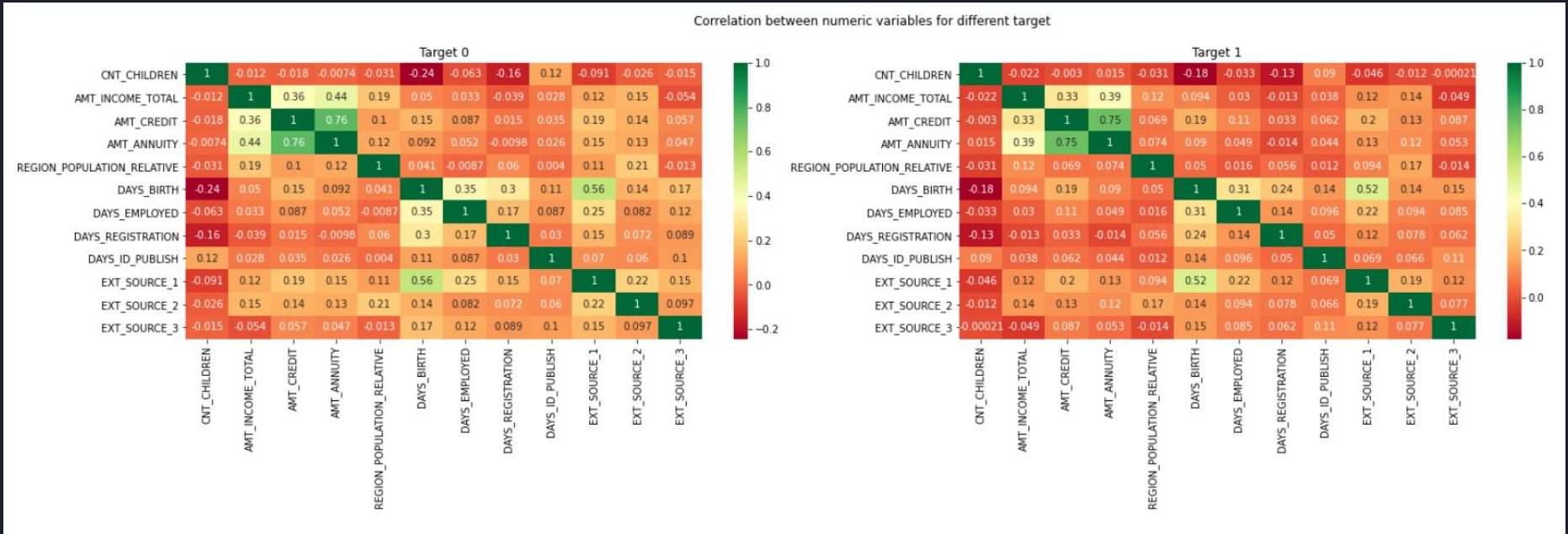


Goods Price Vs Amount Credit For Male



Pattern is same in all 8 boxes. Decision cannot be taken on which segment to avoid

Correlation between numeric variables for different target



Correlation between numeric variables for different target

- Amount Credit and Amount Annuity is positively linearly correlated ->(~0.8)
- Income Total and Amount Annuity is positively linearly correlated -> (~ 0.4)
- Days Birth(Age) has positive correlation Ext_Source1 (~0.5)
- These correlations are same for both target 0 and target 1 customers.

Previous Application Data Analysis

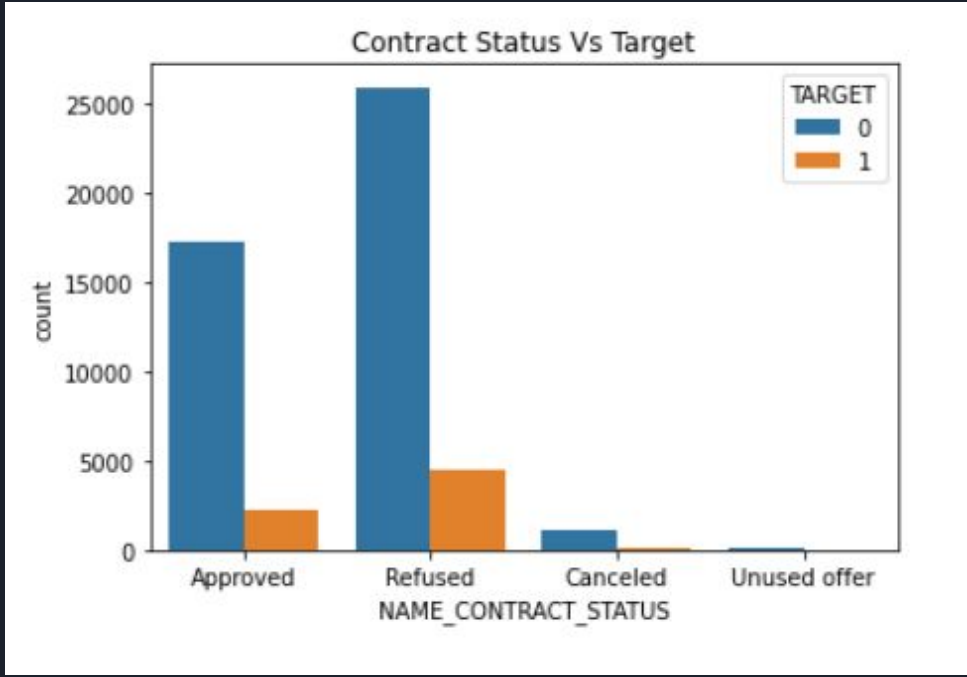
- Load Data, missing value treatment
  - We have loaded the previous application data in panda dataframe
  - We deleted the rows having null value, XNA, XAP.
  - We identified the column of our interest. Such columns are: 'SK_ID_PREV', 'SK_ID_CURR', 'NAME_CONTRACT_TYPE', 'AMT_APPLICATION', 'AMT_CREDIT', 'NAME_CASH_LOAN_PURPOSE', 'NAME_CONTRACT_STATUS', 'NAME_PAYMENT_TYPE', 'CODE_REJECT_REASON', 'NAME_CLIENT_TYPE', 'NAME_GOODS_CATEGORY', 'NAME_PORTFOLIO', 'NAME_PRODUCT_TYPE','CHANNEL_TYPE', 'PRODUCT_COMBINATION'

- Merge Previous Application data with Current Application Data on column SK_ID_CURR

- Multivariate Analysis to find which segment to avoid and which segment to focus

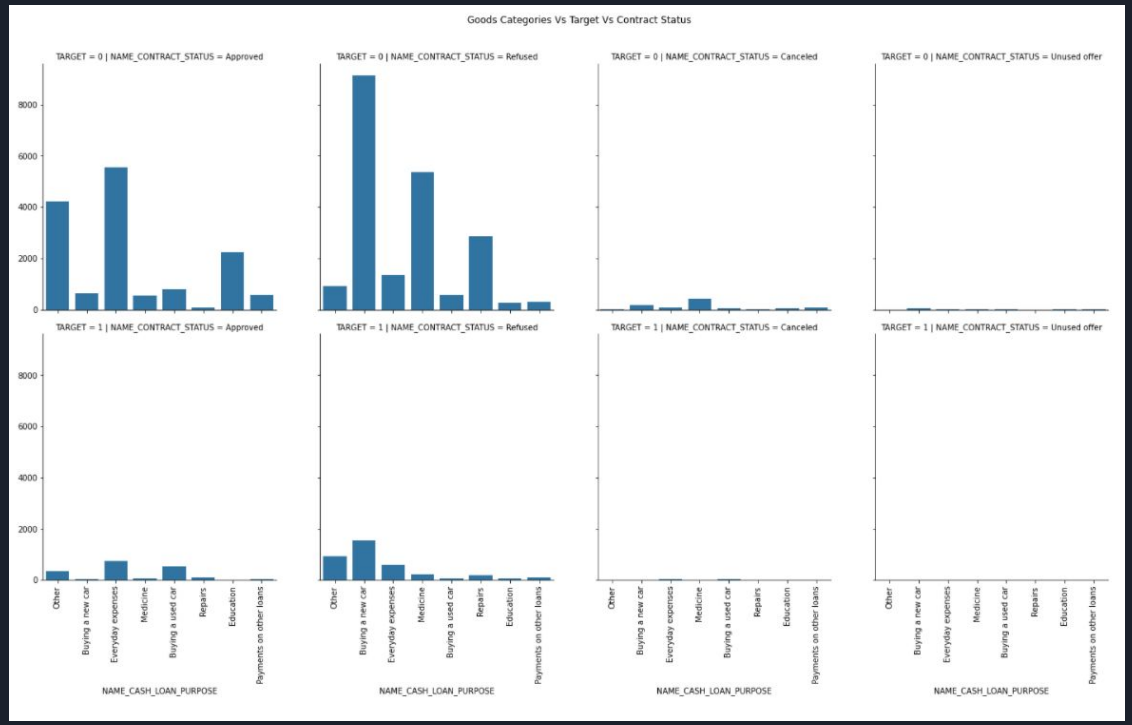
Multivariate Analysis on Merged Data

Contract Status vs Target



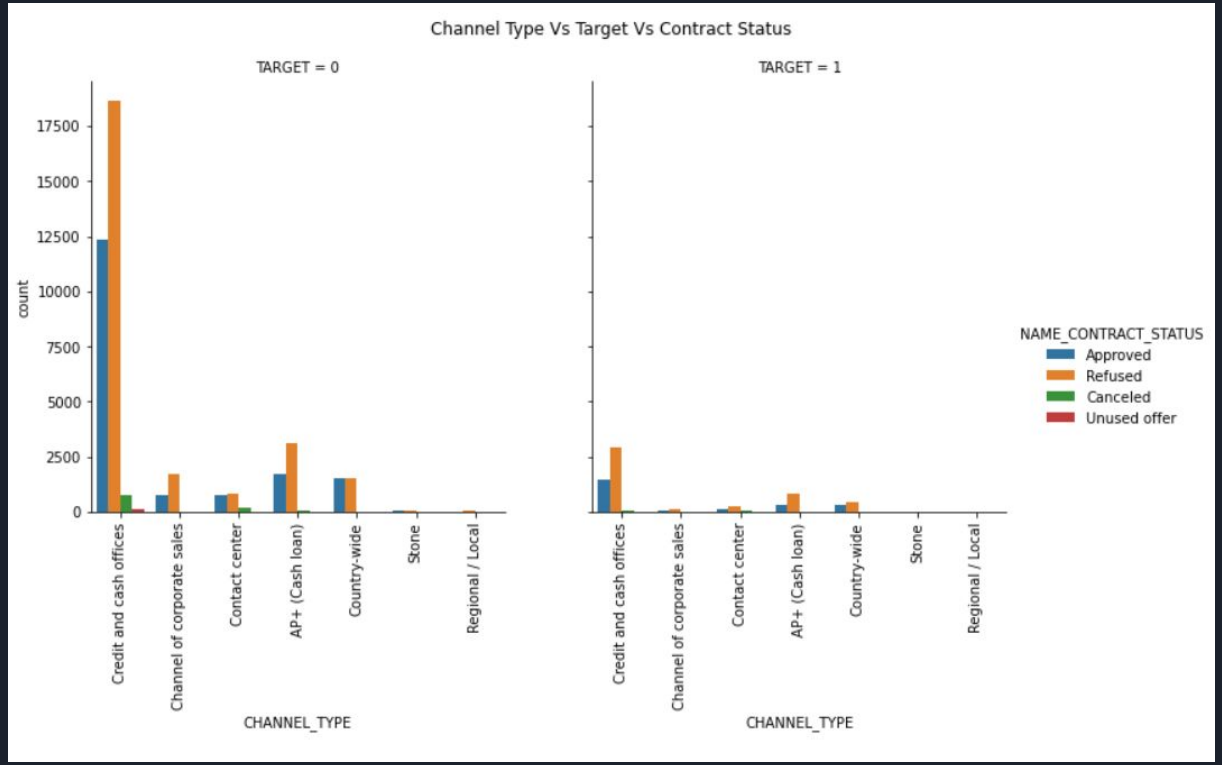
Loan has been approved to people with payment difficulties. Also it has been refused to people who don't have payment difficulties.

Goods Categories vs Target vs Contract Status



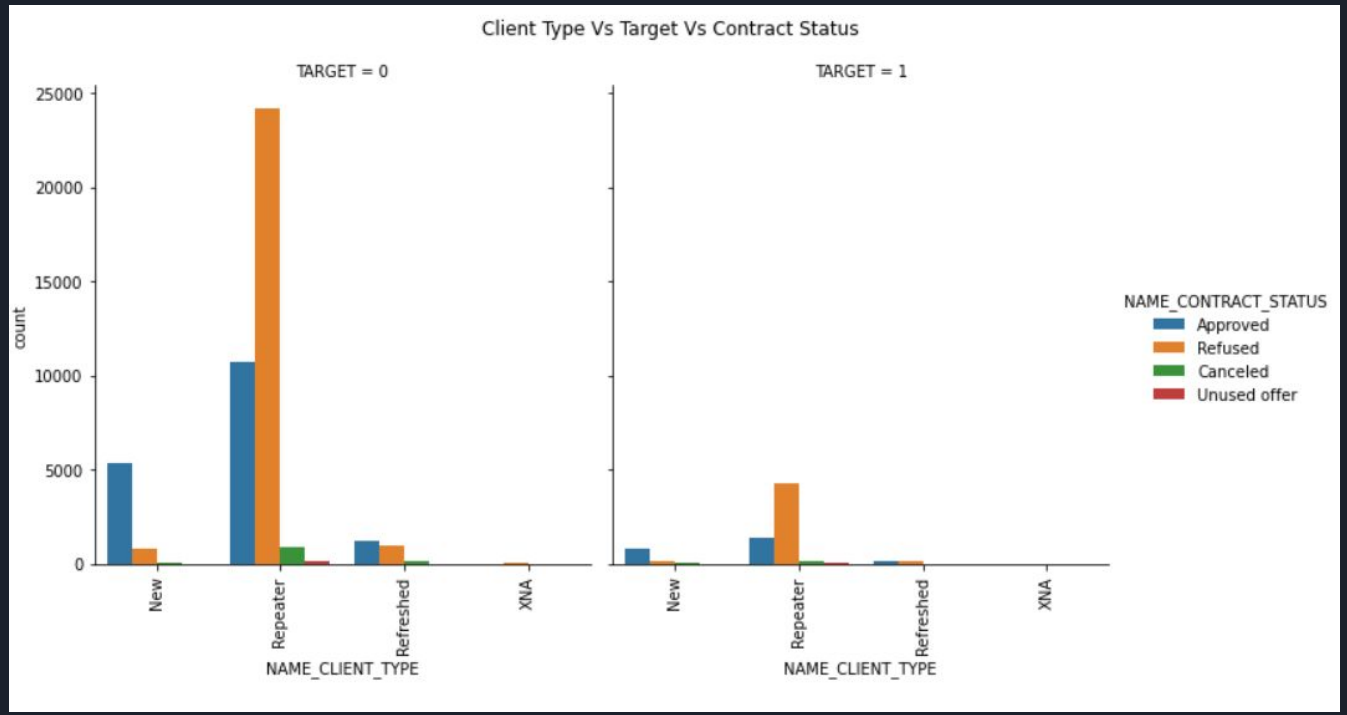
Buying a car, Medicine and Repairs are top categories where loan applications were rejected but these applicants donot have payment difficulties. Loans for these people can be approved.

Channel Type vs Target vs Contract Status



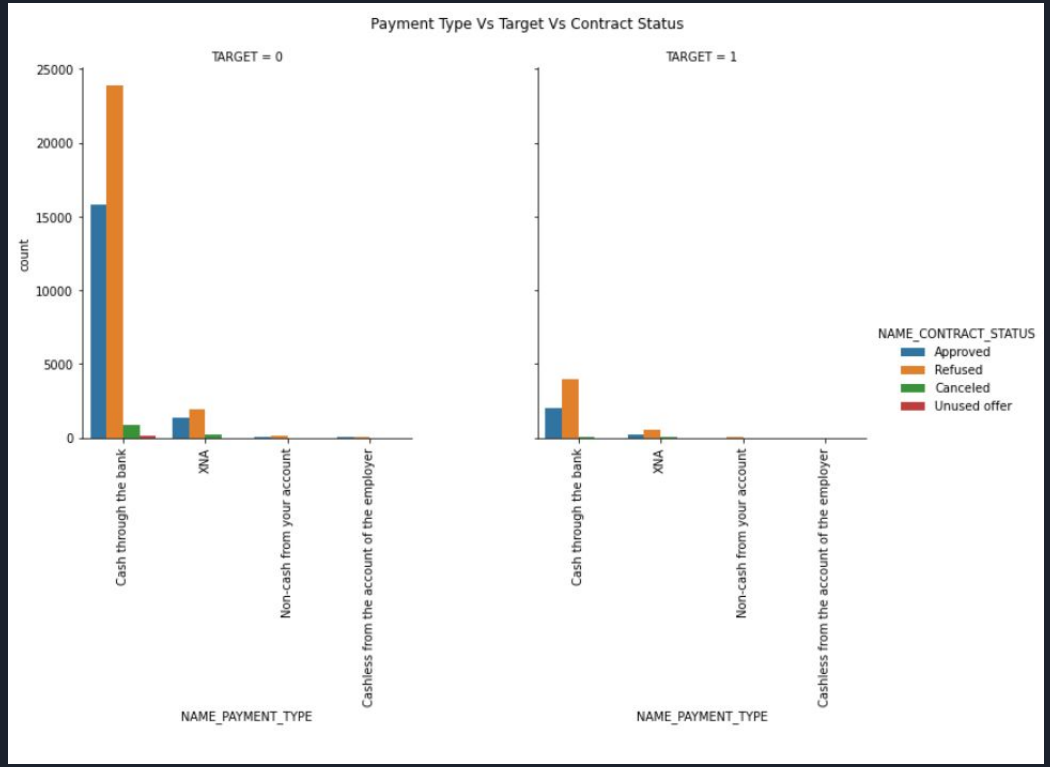
- Target 1 data looks fine. It has more rejection as compared to approval.
- Target 0 data also has more rejection specially for Credit & Cash Offices and AP+(Cash loans). These users do not have payment difficulties. Loan can be approved for them

Client Type vs Target vs Contract Status



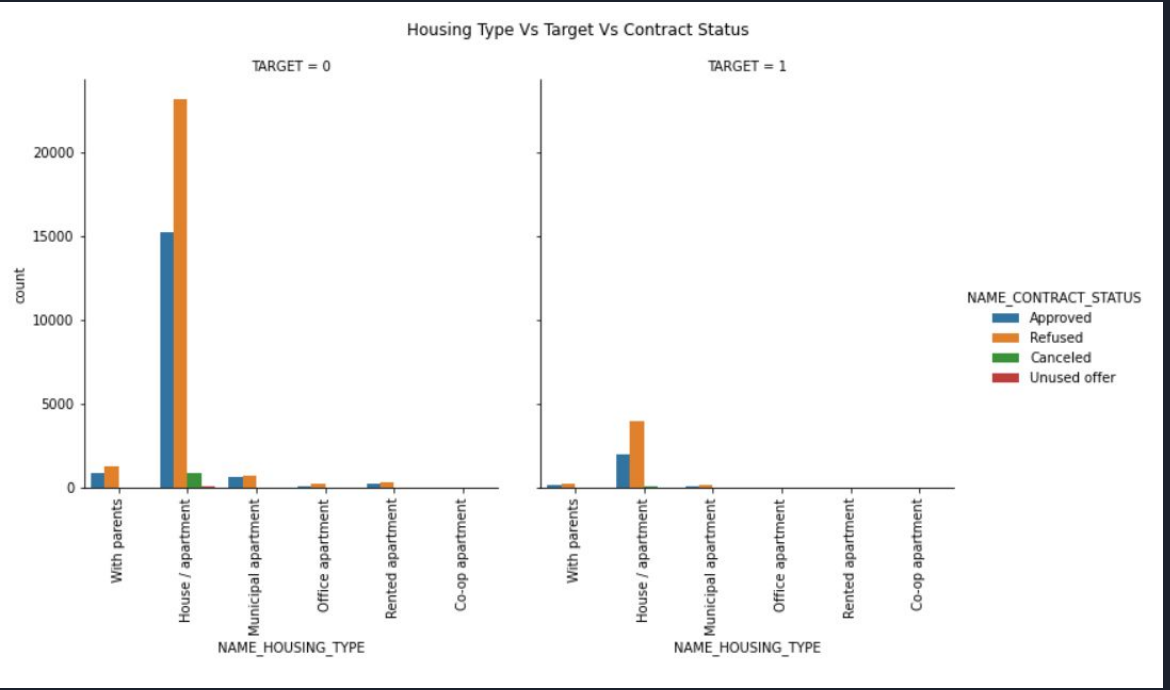
- Organization is good in acquiring new clients but retention is poor.
- There are large no of repeater loan applications who do not have payment difficulties but still their loan application was rejected. Such applications can be approved

Payment Type vs Target vs Contract Status



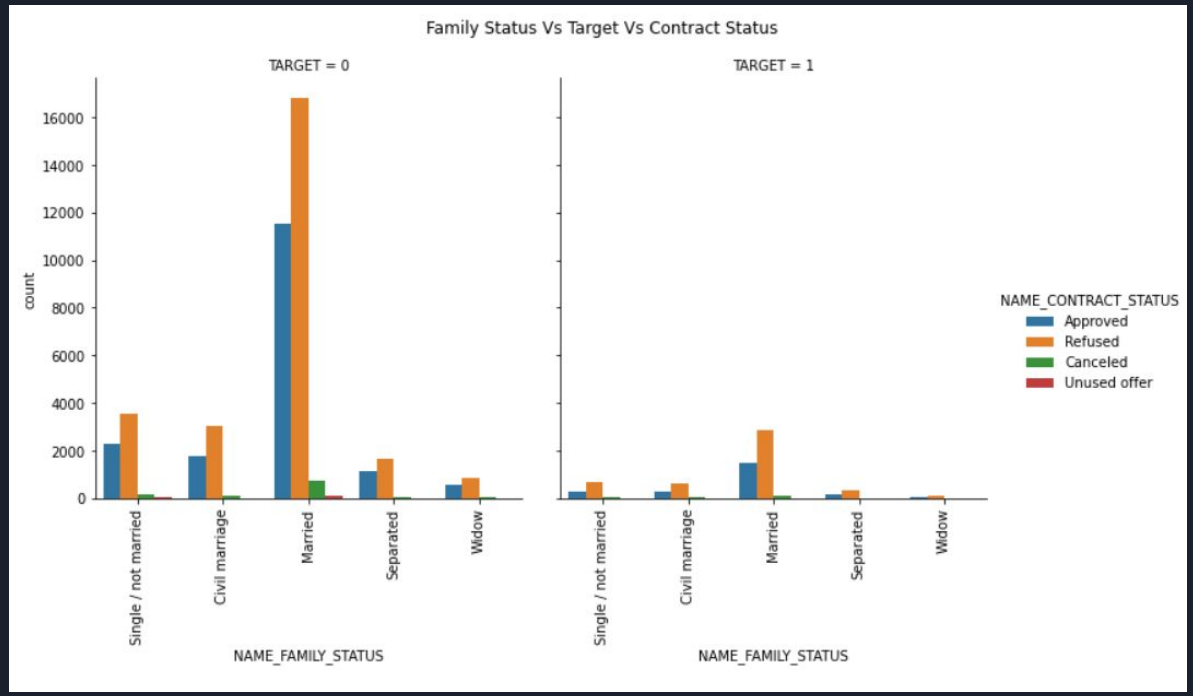
- Large count of loan applications where Cash through the bank was payment method was rejected. They do not have payment difficulties. These applications can be approved.

Housing Type vs Target vs Contract Status



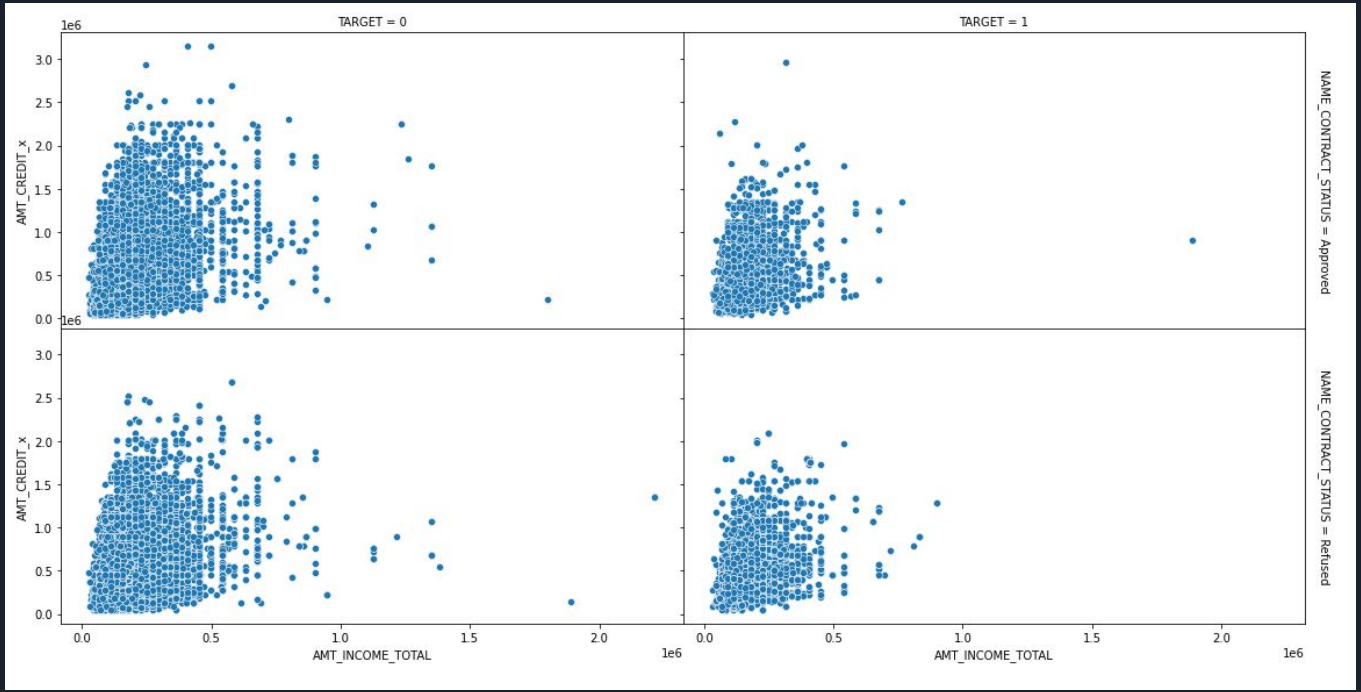
- Large count of loan applications for people living in House/Apartment was rejected. They do not have payment difficulties. These applications can be approved.

Family Status vs Target vs Contract Status



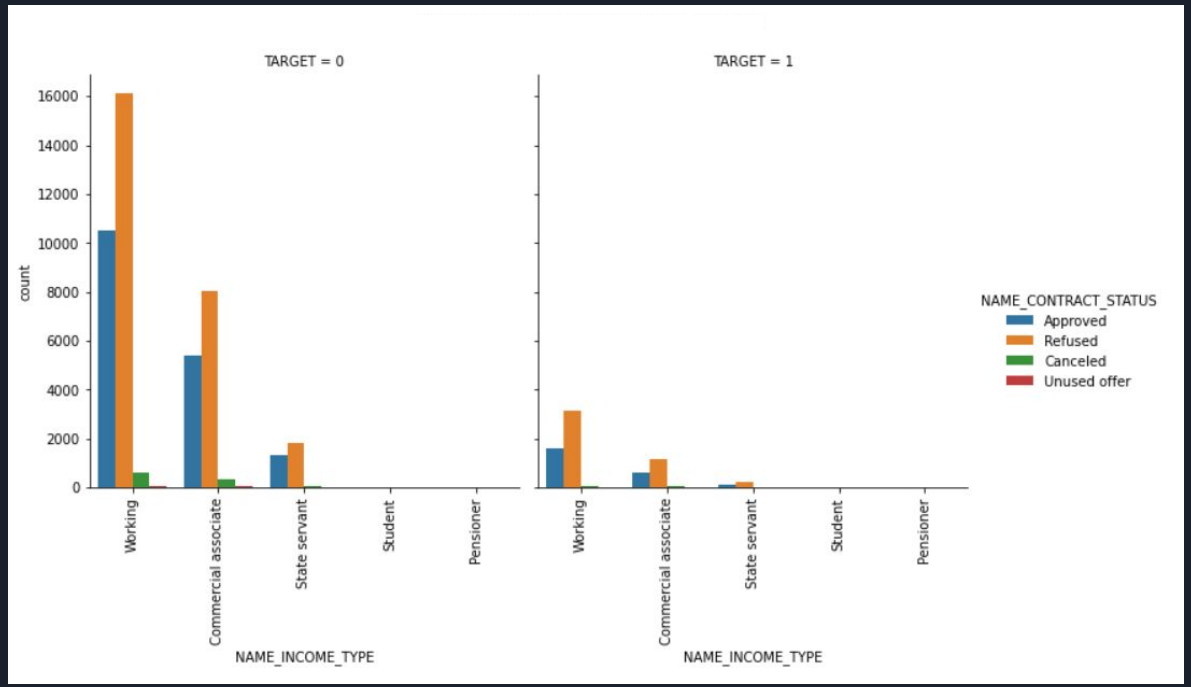
- Loan rejection is high across all family status. There should be other reason for rejection.

Income Vs credit Vs Target Vs Contract Status



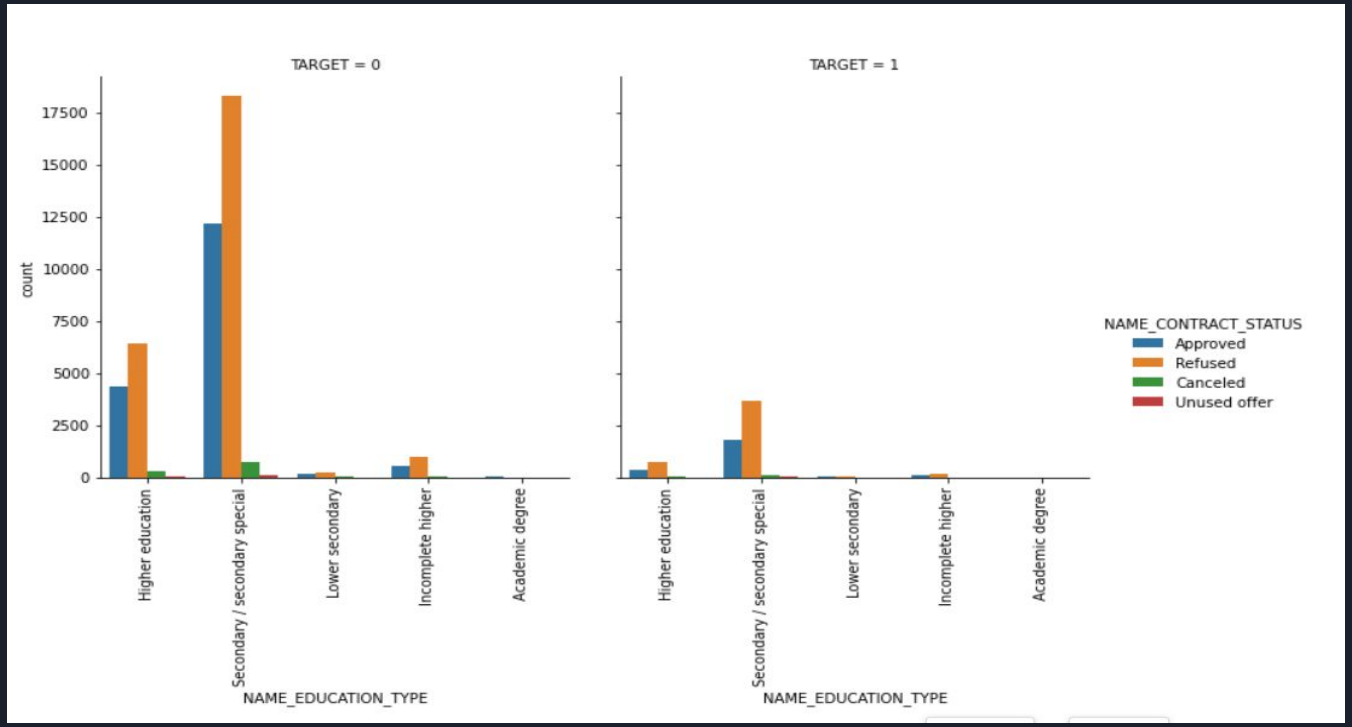
- Distribution pattern is same for all 4 boxes. There should be some other driving factor

Income Type Vs Target Vs Contract Status



Working income type shows higher count in payment issues. Amount of loan for them can be reduced.

Education Type Vs Target Vs Contract Status



Secondary/Secondary Special Education shows higher count in payment issues. Amount of loan for them can be reduced.

Conclusion

Following has been concluded from above analysis

- Approve more loan applications for
  - people living in House/Apartment. They do not have payment difficulties.
  - For Buying a car, Medicine and Repairs are top categories. People requesting loans for these good have less payment issues.
  - For people where payment type is Cash through the bank.
- Restrict loan for
  - Secondary/secondary special Education people. They have more payment issues.
  - Working income type people. They have more payment issues.