Exploratory Data Analysis for Loan Assessment

- By Ankan Roy

Business Objectives

Loan provider aims to identify patterns which indicate if a client has difficulty paying their instalments 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.

Assumptions

Loan provider lends loans to those who are having following

- Applicants history
- Good income
- Good family status
- Good occupations
- Good surroundings
- Client discipline

***Note** : This dataset will be analysed inline based on the assumptions listed above.*

Dataset Relationship and Facts

Application data

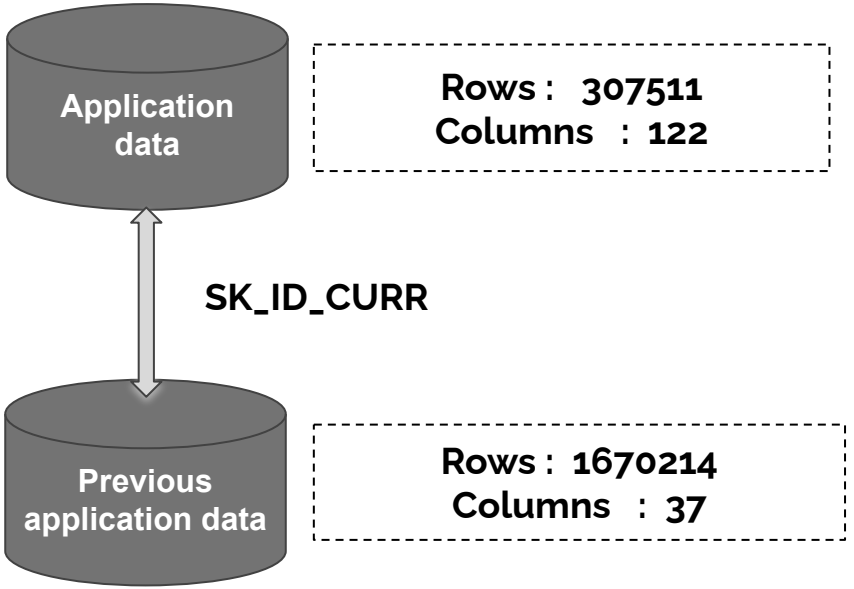
- Contains client's current information at the time of loan application
- No duplicate records
- Too many columns

Previous application data

- Contains information about the client's previous loan application.
- Duplicated rows based on SK_ID_CURR (single applicant has submitted loan application multiple times)

Column application

- Contains description of the columns
- Used for reference not for analysis.



Dataset Alteration

Application data

Columns Merged

FLAG_MOBIL,FLAG_EMP_PHONE,FLAG_WORK_PHONE,FLAG_PHONE and FLAG_EMAIL these columns are clients contact information, hence merged them to one column called CONTACT_INFO

There are 21 columns related to FLAG_DOCUMENT. Assumption made that all the documents are important in loan application. Hence merged to one column called DOCUMENT_PROVIDED

There are 3 columns related to EXT_SOURCE. Assumption made to take sum of all the external source and add in one column EXT_SOURCE, because it will capture the total normalized score of external data.

Columns Dropped

Apartment information related columns, credit bureau, address etc

Columns reduced from 122 to 32.
73% of reduction in columns

Inner join

Previous application data

Considering following columns

- SK_ID_CURR
- NAME_CONTRACT_STATUS
- CODE_REJECT_REASON
- NAME_CLIENT_TYPE

The intention is to get the client history of loan status and then will compare with the current client dataset

Considered 4 columns out of 32 columns.

Merged dataset

Loan_risk_analysis :

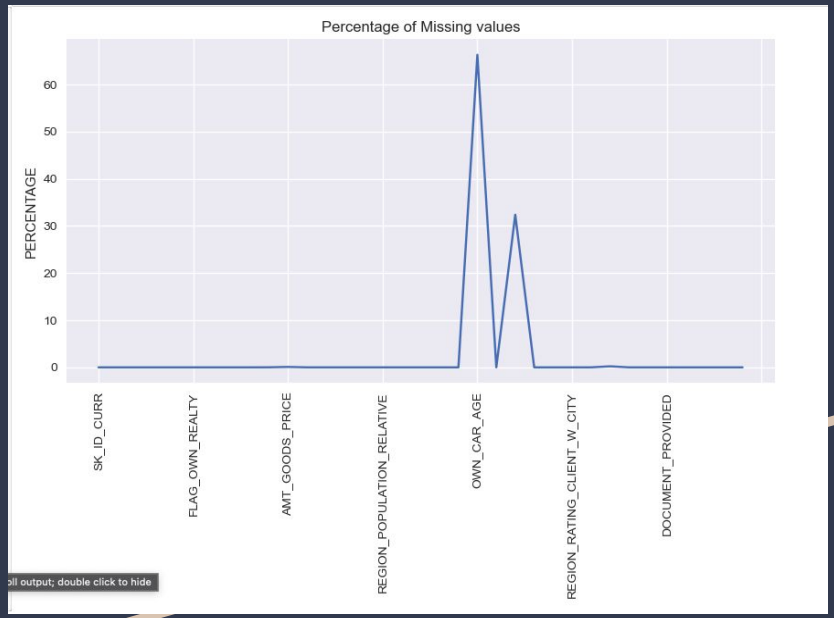
The new dataset will contain information on clients current application and its corresponding previous history.

Row - 1413701

Columns - 36

Further analysis will be performed from the merged dataset.

Missing Value



Missing values are present in

- AMT_ANNUITY - 0.01%
- AMT_GOODS_PRICE - 0.09%
- OWN_CAR_AGE - 66.29%
- OCCUPATION_TYPE - 32.37%

OCCUPATION_TYPE and ORGANIZATION_TYPE

- 32% of Occupation type has missing value in NaN. Will convert it to a different category.
- Organisation type missing value is XNA, will convert it to a different category.
- Drop those values because 31% values are missing.

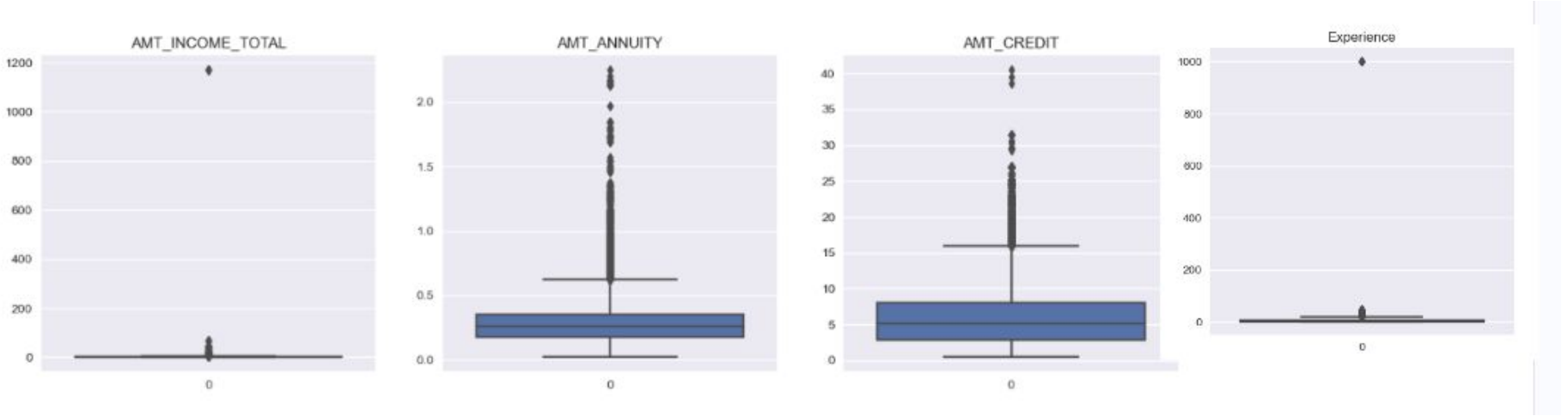
AMT_ANNUITY and AMT_GOODS_PRICE

- Percentage of missing annuity is very less, so we can drop the values

CODE_GENDER

- Dropped XNA because there is nothing called XNA in gender.

Outliers



- Outlier is present in credit amount
- Outlier is present in income amount
- Outlier is not present in annuity amount
- Outlier is not present in in Age
- Outlier is present in experience.

Created bins/range for following columns to treat outliers -

- AMT_INCOME_TOTAL
- AMT_CREDIT
- Experience

Detailed data analysis

- Perform data standardization
- Understand of duplicate values related to SK_ID_CURR
- Univariate analysis
- Bivariate analysis
- Multivariate analysis

Standardize Values

- DAYS_BIRTH & DAYS_EMPLOYED contains negative values.
- Removed negative values and converted days to years.

Duplicate values related to SK_ID_CURR

- Presence of duplicate SK_ID_CURR is legitimate which implies that same applicant and applied for loan multiple times.
- ~78 % of clients have submitted more than one time for 50 % of the data.

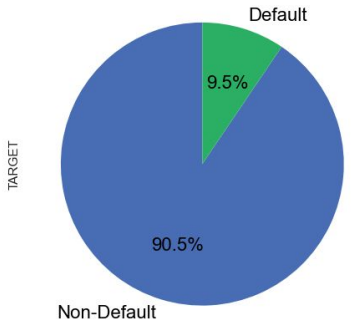
Next slides will be related to univariate,bivariate and multivariate analysis with below assumption. ***(also mentioned in slide #2.)***

Assumptions

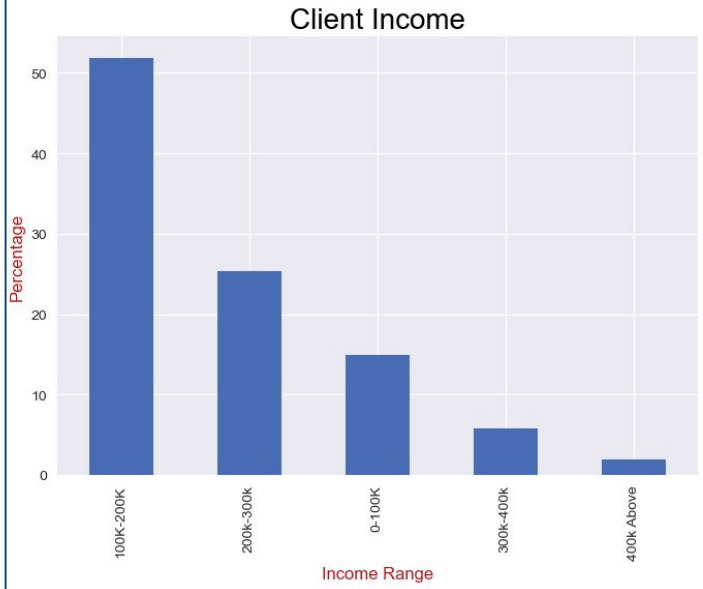
Loan provider lends loans to those who are having following

- Applicants history
- Good income
- Good family status
- Good occupations
- Good surroundings
- Client discipline

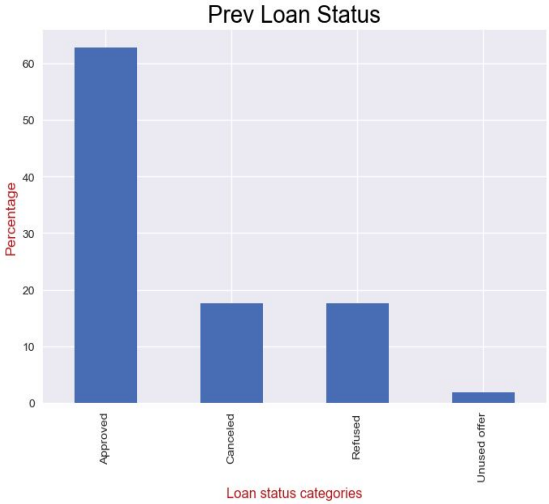
***Note** : This dataset will be analysed inline based on the assumptions listed above.*



- Target has two values - 0 (clients without payment difficulties) and 1 (with payment difficulties)
- 0 has been designated as Non-Default and 1 has been designated as Default
- **Data is highly imbalanced** because the percentage of Non-Default is 90.5 % and Default is 9.5 %

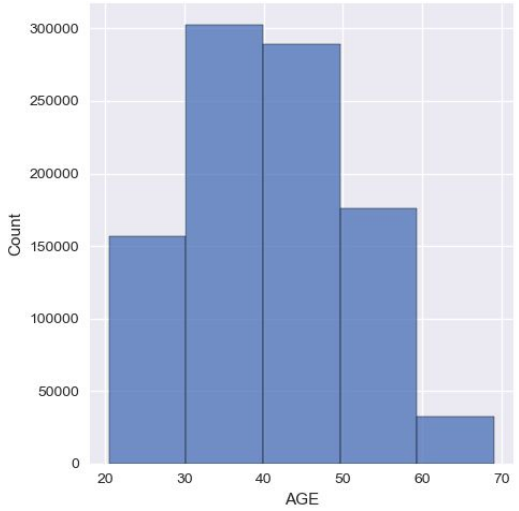


- **52 %** of applicants earn between the range of 100K-200K
- Roughly 2 % of clients are highly paid.
- Mean income - 189533.2



- Approved(Loan approved) - **63%**
- Cancelled (cancelled the application) - 18%
- Refused(Loan rejected) - **18%**
- Unused Offer(Client cancelled the loan during diff stages of the process) - 2 %

Univariate Analysis



- Client between age group between 30-50 is more
- Senior citizen (60-70) - less than 50,000

Assumption

- Rating 1 is tier 1 city
- Rating 2 is tier 2 city
- Rating 3 is tier 3 city

Takeaway

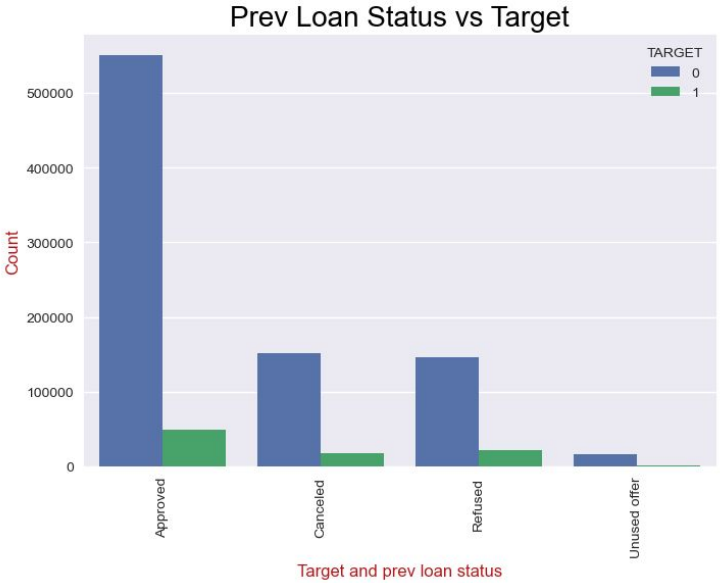
- The proportion of clients seeking for loans from tier 2 cities is approximately **75%.**
- Clients from Tier 1 cities do not apply for loans in large numbers.

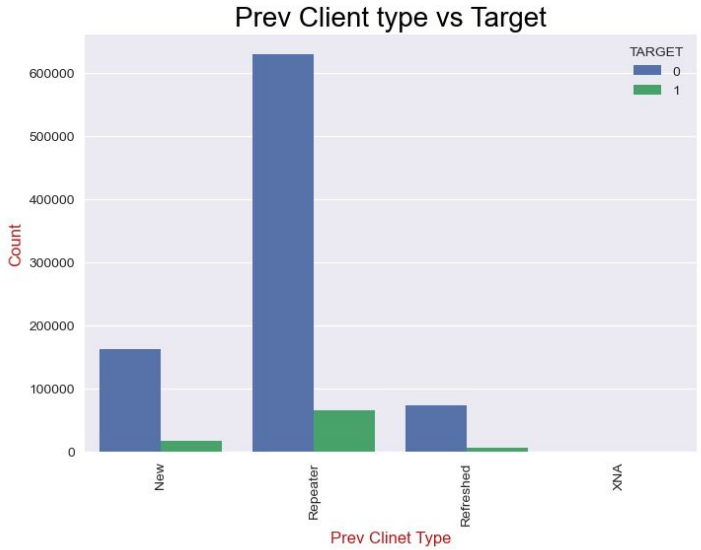
Univariate Analysis

Applicant's History

In order to understand the history of the application. Analysis between following has to be performed -

NAME_CONTRACT_STATUS, **NAME_CLIENT_TYPE** and **TARGET**



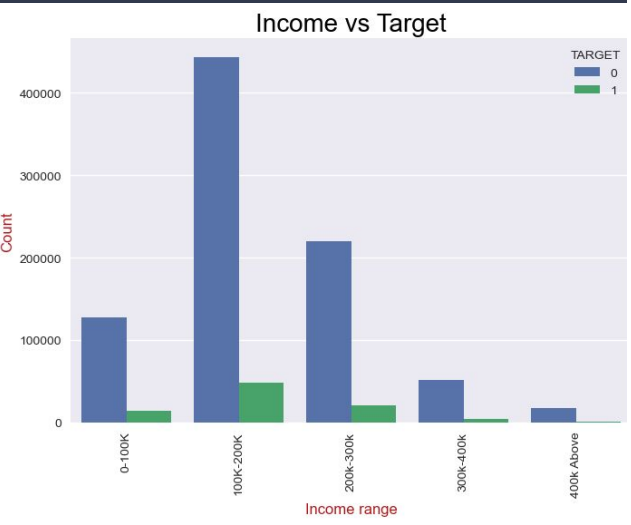


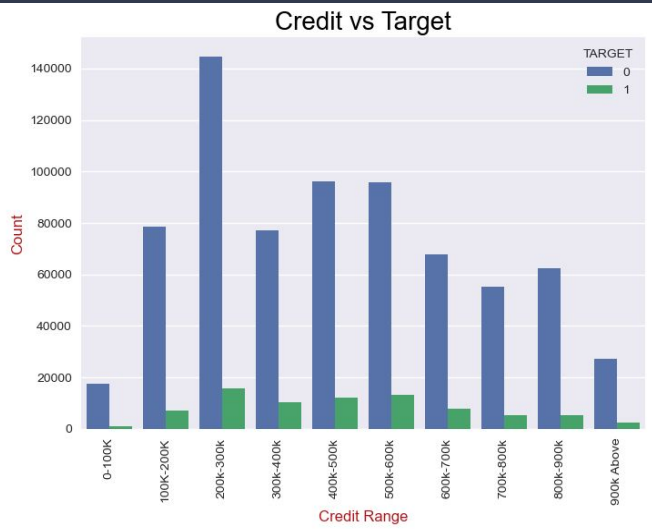
- **~ 87%** of client whose previous loan got rejected are able to pay in the current application.
- Loans got approved for most of the clients who are **repeat customers**. Bank should consider more on the repeat clients.
- Applicants history like **repeat customers and prev approved users** are most likely to get loans approved for current application

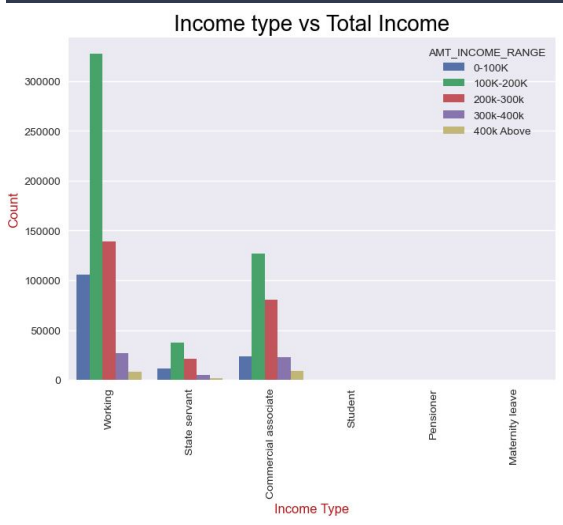
Applicant's Income

In order to understand the income of the application. Analysis between following has to be performed -

AMT_INCOME_RANGE, AMT_CREDIT_RANGE and NAME_INCOME_TYPE



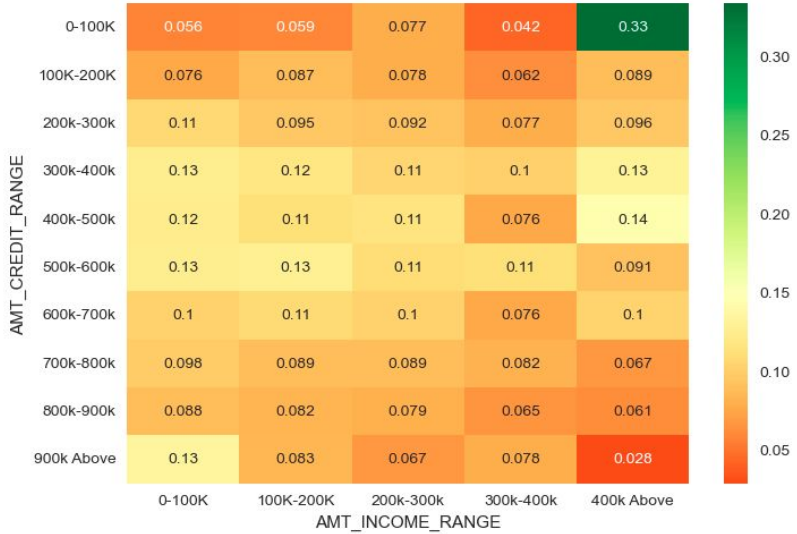


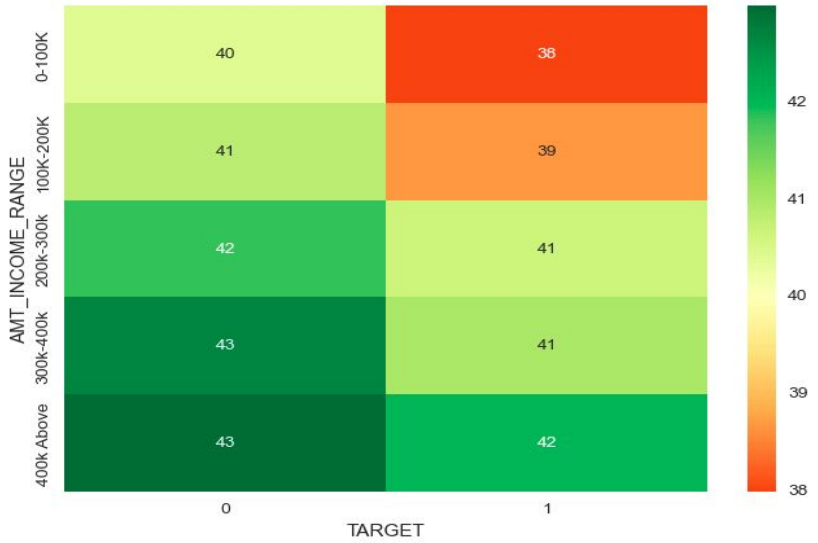


- High income clients are most likely to get loans approved.
- **~92%** of the clients earn in the income range of 100K to 300K.
- For low income clients there are **~10%** chance the loan will be rejected.
- Clients falling under the credit range of 200K to 300K are likely to get the loans approved.
- High chance of loan acceptance for working class client than non working class client due to low income.

Applicant's Income

Multivariate analysis between Credit amount, income range , Target and Age



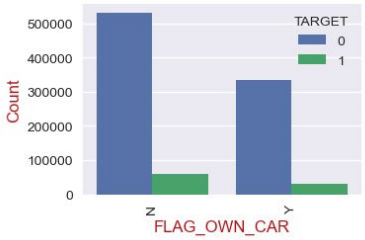


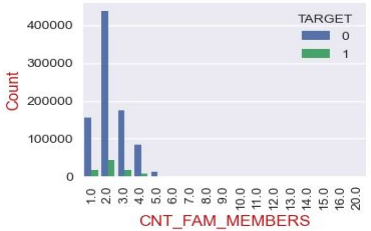
- Mean of age 40 with income range 200K - 300K have tend to no default case.
- **High correlation** between high income range and low credit range.

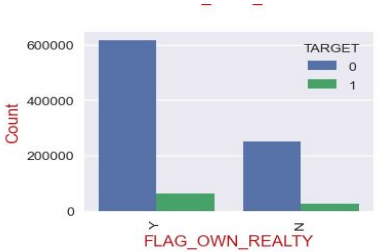
Applicant Status

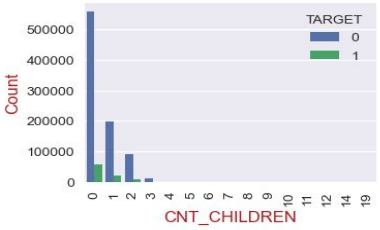
In order to understand the status or about the applicant. Analysis between following has to be performed

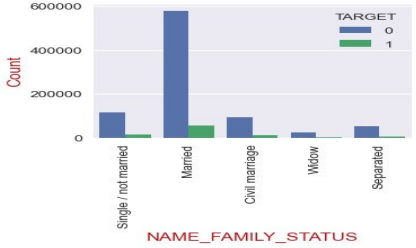
FLAG_OWN_CAR, FLAG_OWN_REALTY, CNT_CHILDREN, CODE_GENDER, NAME_FAMILY_STATUS,NAME_HOUSING_TYPE,AGE,CNT_FAM_MEMBERS

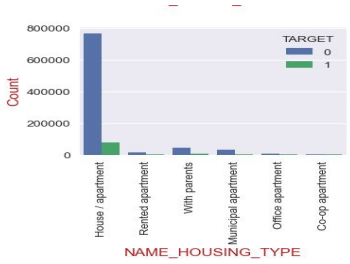










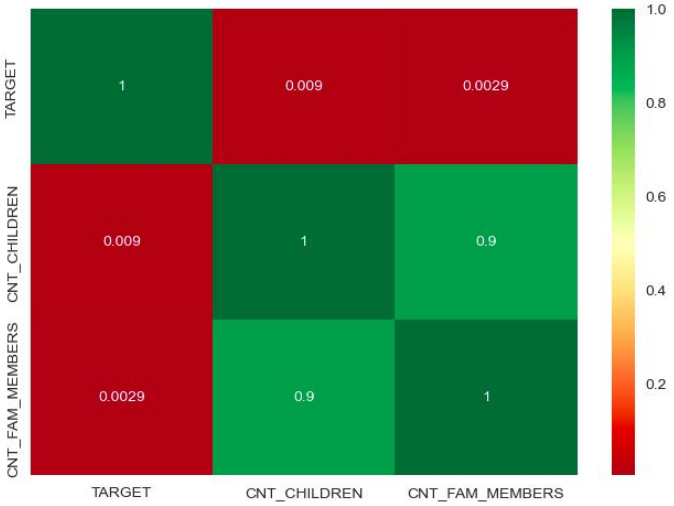


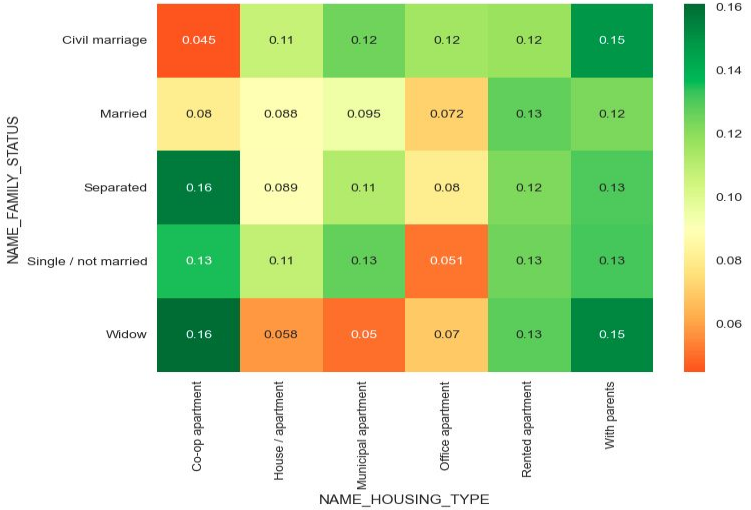
High Chances to get loan for following -

- Female more than Male
- Low children count
- Own a house
- No car
- Moderate family member size

Applicant Status

Multivariate analysis between Credit amount, income range , Target and Age



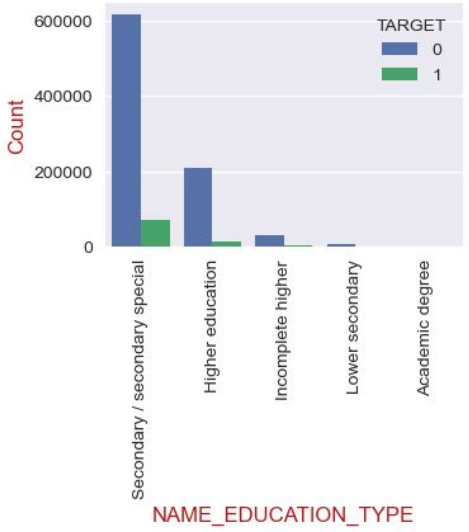


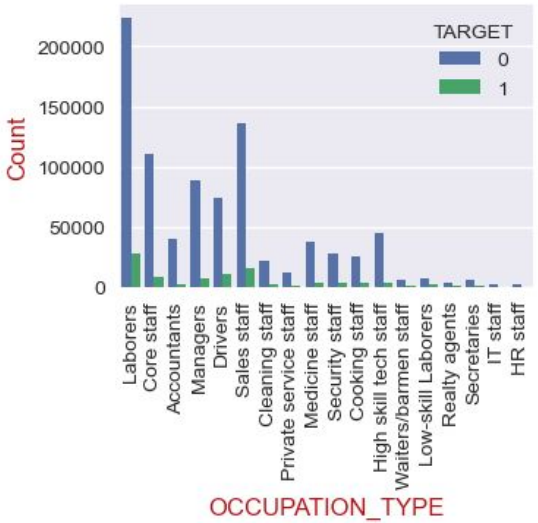
- **Good correlation** between CNT_FAM_MEMBERS and CNT_CHILDREN
- CNT_CHILDREN & CNT_FAM_MEMBERS: The more members there are, the **less likely** they are to acquire a loan.
- **Less chance** to get loan for Single/not married living in office apartment.

Client Occupation

In order to understand the occupation of the applicant. Analysis between following has to be performed

NAME_EDUCATION_TYPE,Experience_Range and OCCUPATION_TYPE



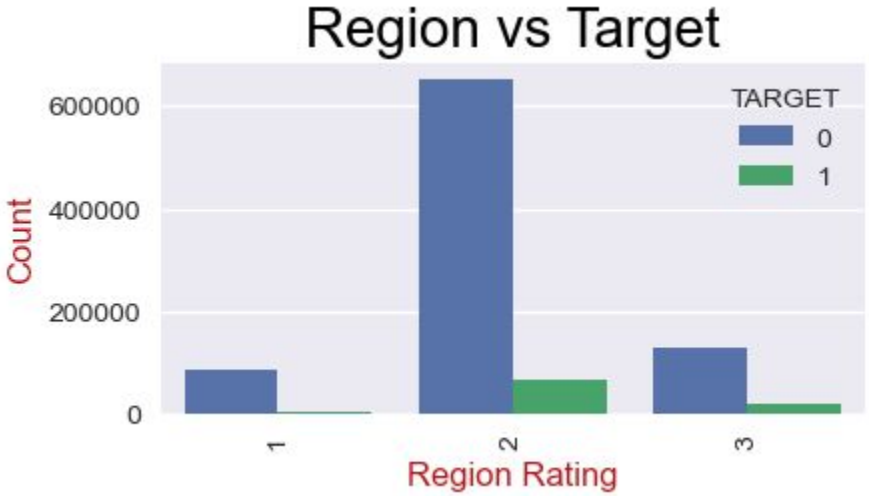


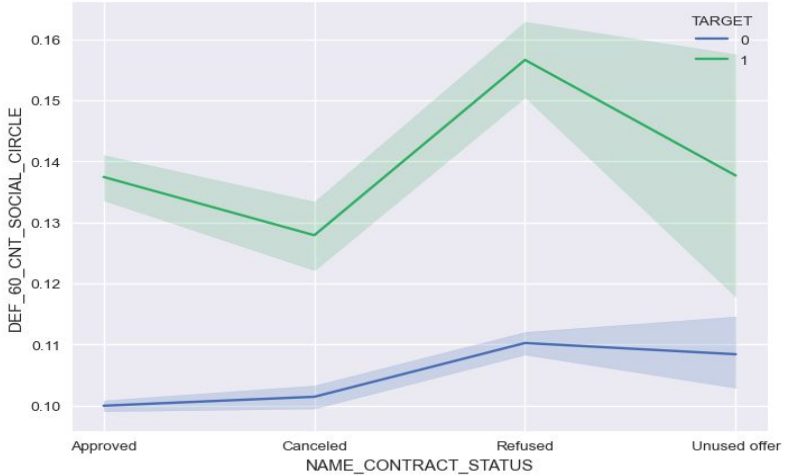
- **Higher** the education good chances to get a loan.
- Experience range 0-20 years will be good chance to get a loan.

Client's Surroundings

In order to understand the surroundings of the applicant. Analysis between following has to be performed

REGION_RATING_CLIENT and DEF_60_CNT_SOCIAL_CIRCLE





- Having a average or higher social environment range of **0.14** reduces your chances of getting a loan.
- Less chance to get the loan having bad social surroundings.
- 2 rating clients are more prone to get be non defaulter.
- Clients with region rating 3 having high chance of loan rejection

Summary

- Data is highly imbalance.
- **NAME_CONTRACT_STATUS and NAME_CONTRACT_STATUS** - Applicant's history, such as repeat customers and previously approved users, are more likely to have loans accepted for their current application.
- **AMT_INCOME** - Loans are most likely to be authorised for high-income clients.
- **AMT_CREDIT** - There is a strong link between a high income and a low credit range.
- **NAME_INCOME_TYPE** - Working-class clients have a higher likelihood of loan acceptance than non-working-class clients due to lower income.
- **CODE_GENDER**: Men have a higher default rate than women.
- **CNT_CHILDREN & CNT_FAM_MEMBERS**: The more members there are, the less likely they are to acquire a loan.
- **NAME_EDUCATION_TYPE** - The higher the education, the better the chances of obtaining a loan.
- **DAYS_EMPLOYED** - If you have 0-20 years of experience, you have a decent probability of getting a loan.
- **DAYS_BIRTH** - Mean of age 40 with income range 200K - 300K have tend to no default case.
- **REGION_RATING_CLIENT**: RATING 2 are the good clien region rating for loan acceptance.
- **DEF_60_CNT_SOCIAL_CIRCLE** - Having a average or higher social environment range of 0.14 reduces your chances of getting a loan.
- If documents are not provided then there is high chance for the client to be defaulter
- Less chance to get the loan having bad social surrounding