Credit EDA Case Study

Two data sets were provided as part of this case study

- Application Data
- Previous application data

In order to perform data quality checks and handling missing values, **we have considered Application data set and performed necessary actions as explained below**.

To find out the missing values and handle it :

- More than 40% or more null row columns are removed from application data.

Imputation on Columns:

- Replaced the XNA values with MODE for CODE_GENDER and for ORGANIZATION_TYPE as the data has with 18% of records replaced with NaN

New Columns Derivation:

- FAMILY_SIZE from CNT_FAM_MEMBERS
- AGE_IN_YEARS from DAYS_BIRTH
- EXPERIENCE_IN_YEARS from DAYS_EMPLOYED
- POPULATION_DENSITY from REGION_POPULATION_RELATIVE
- AGE_RANGE from AGE_IN_YEARS
- EXPERIENCE_RANGE from EXPERIENCE_IN_YEARS
- REGION_RATING from 'REGION_RATING_CLIENT','REGION_RATING_CLIENT_W_CITY'.

UniVarient Analysis :
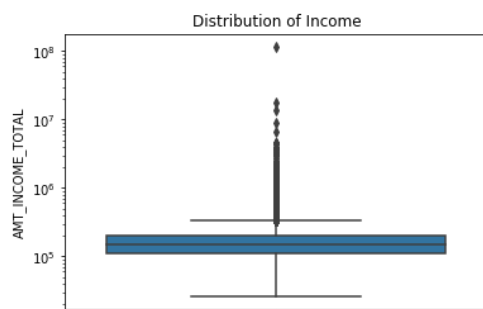
- Upper whisker is very slim from Income compared to Lower whisker states that there are less clients with high income amounts.
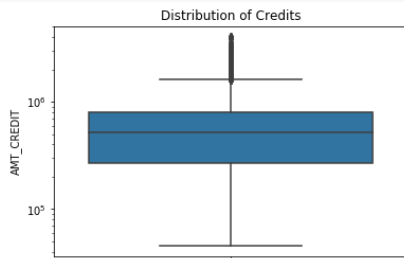
- Some outliers noticed and will be binned based on 95, 99 percentile/quantiles.

- Bins created based on above quantile range and labeled accordingly, outilers are handled by keeping them in a seperate bin

bin=[0,80000,100000,120000,140000,160000,180000,220000,280000,340000,480000,1170000 00]
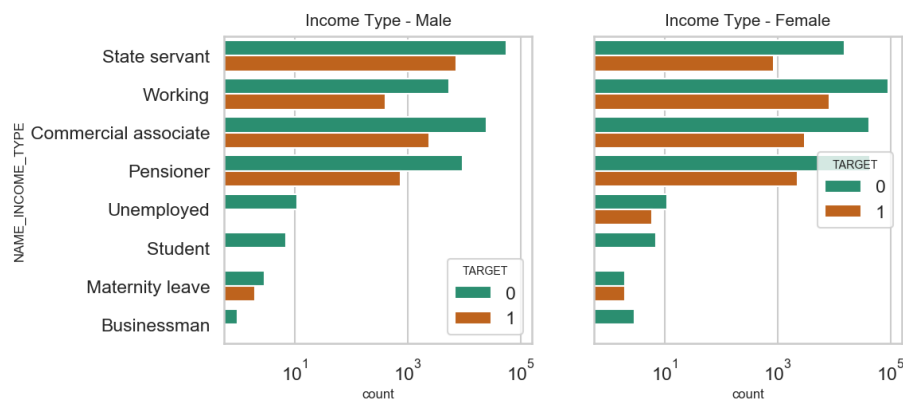
- Upper whisker is very slim from Income compared to Lower whisker states that there are less clients applied for high credits.
- Some outliers noticed and will be binned based on 95, 99 percentile/quantiles
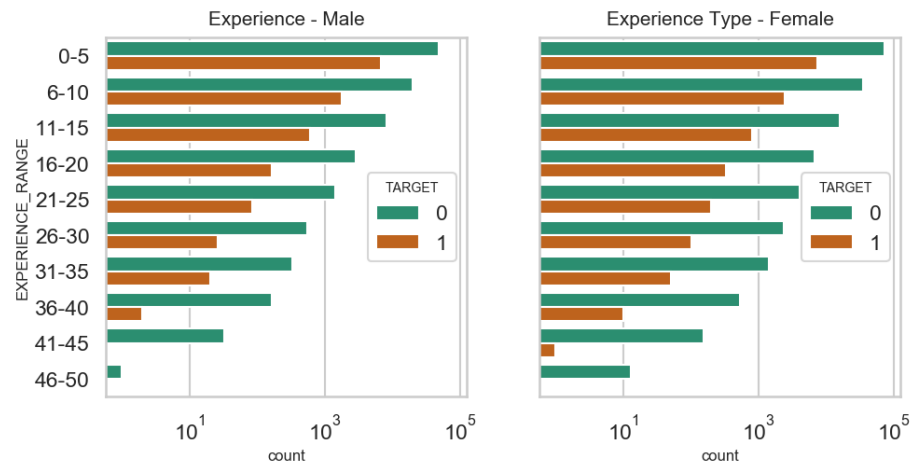
Bivarient Analysis:

- Working,Pensioner Unemployed women are having highest difficulty in repayment of loans
- State servant, Commercial Associate Men are having payment difficulty.
- Repaying capacity of Business women is high compared to men.
- No defaults for Student Clients.

- Clients with Experience 41-50 has no payment difficulty.
- For this both Male and Female almost same but though Female is leading for loans in range of 0-5 work ex.

Previous Application Dataset

- Merge the datasets Application and previous on SK_Current_ID (inner join)
- Took sample out of the output data
- Started visualization
- Target0 - O – Others , Target1- PD – Payment Difficulties.

AGE – CONTRACT STATUS:
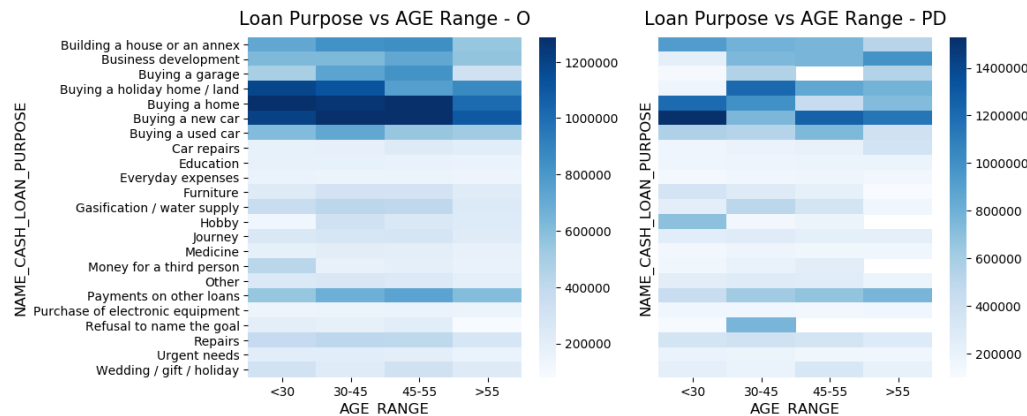
Age vs Contract_Status - O

Age vs Contract_Status - PD
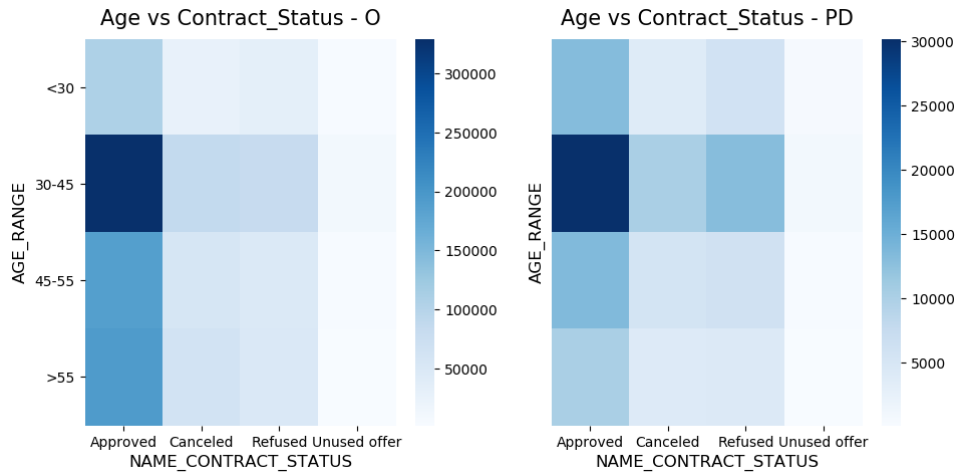
- Approved Applicants for the Age Range 30-45 has high payment difficulty and repay capacity.

EDUCATION TYPE — CONTRACT STATUS:

Education vs Contract_Status - O

Education vs Contract_Status - PD
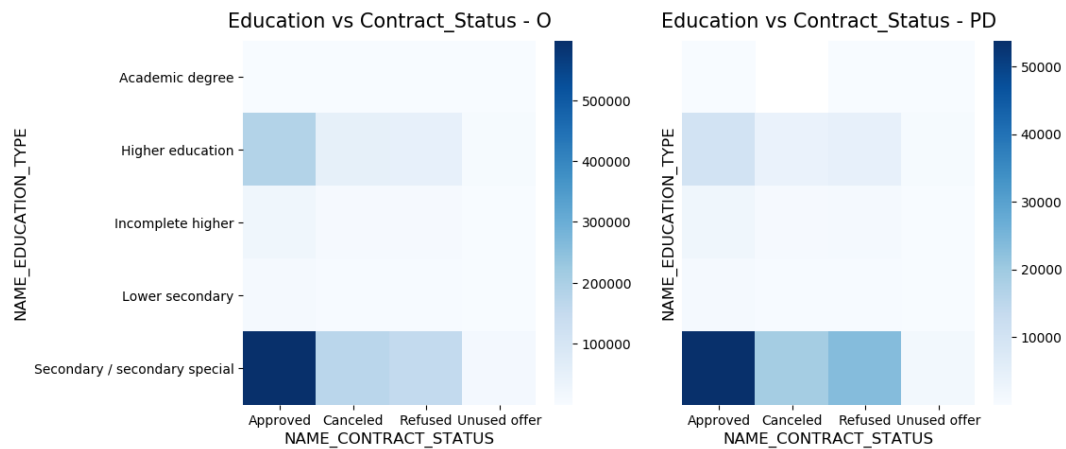
- Secondary Education type has higher Approvals and Refused Applications.

LOAN PURPOSE — AGE:

Loan Purpose vs AGE Range - O   Loan Purpose vs AGE Range - PD

- Approving loans for age 30-45 for buying a car than <30 will be beneficial for bank.
- Approving a Loan for buying a garage for the age 45-55 has high repay capacity

LOAN PURPOSE – POPULATION DENSITY:

Loan Purpose vs Population Density - O   Loan Purpose vs Population Density - PD
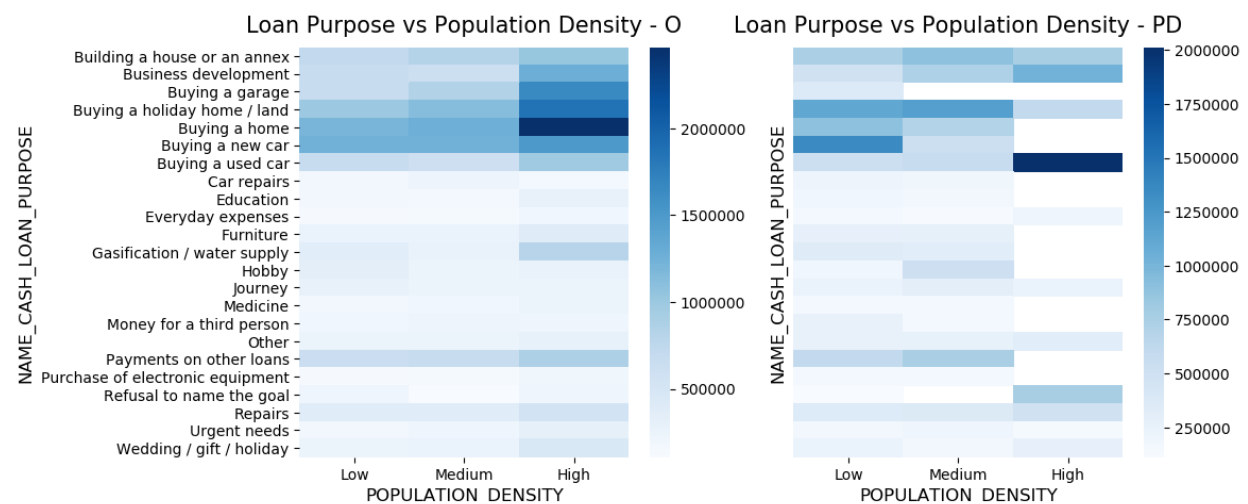
- Client from high populated area Buying a home should be giving high priority.
- Default rate is high for buying a used car from high Populated area

LOAN PURPOSE – INCOME TYPE:

Loan Purpose vs Income Type - O / Loan Purpose vs Income Type - PD
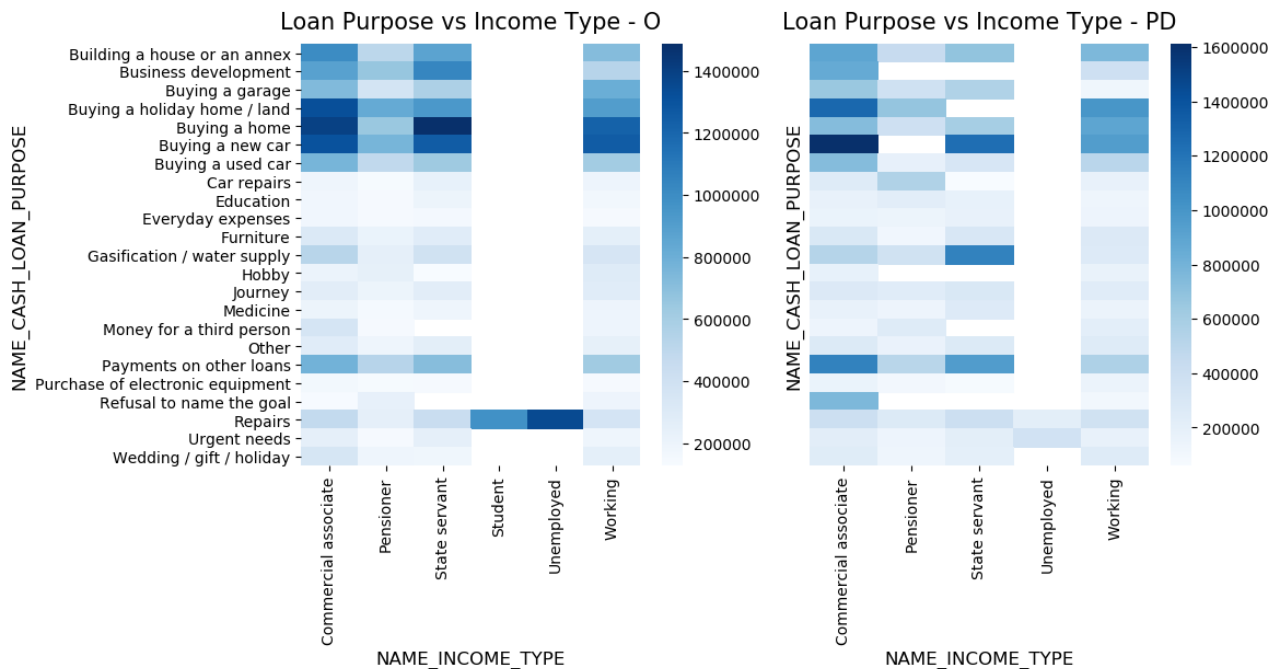
- Higher repay capacity for Commercial Associates with purpose buying a holiday home/land and a house
- State servents has high repay ability for buying a house.
- Unemployed has high repay ability for repairs.
- High default rate for Commercial Associates with purpose buying a car

LOAN PURPOSE – REGION RATING:

Loan Purpose vs Region Rating - O / Loan Purpose vs Region Rating - PD
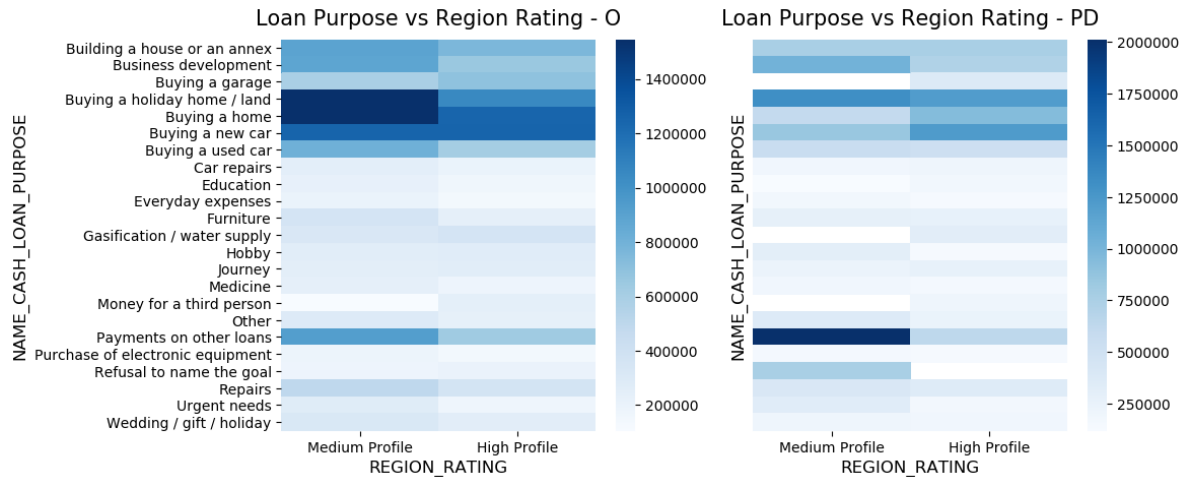
- Clients from Medium Profile Region has high default rate for the loan purpose of Payment of other loans.
- Clients from Medium Profile Region has high repay capacity for the loan purpose of home and land.

CLIENT TYPE – GENDER:

Client Type vs Status - O / Client Type vs Status - PD
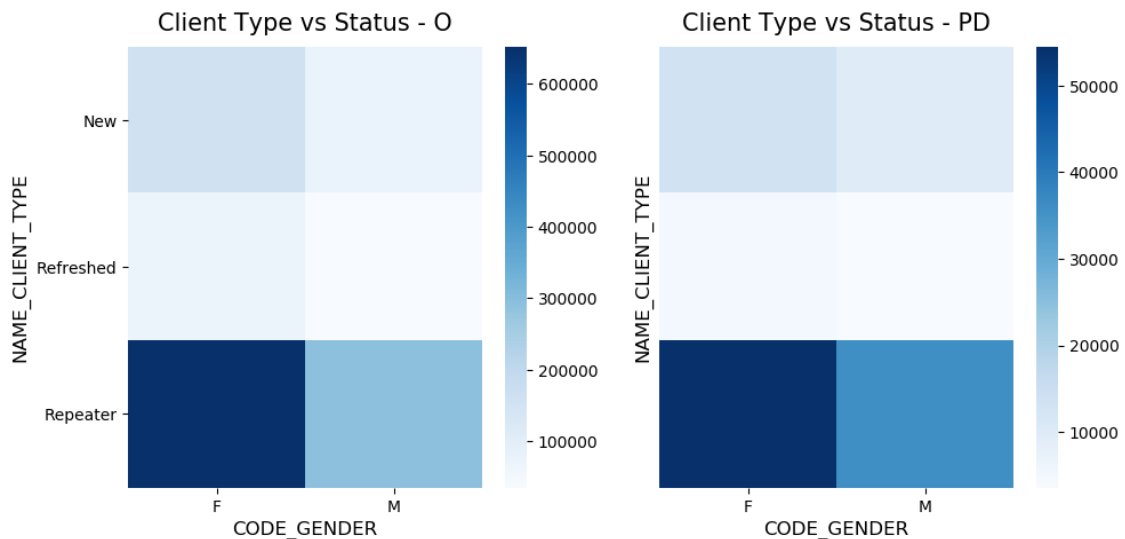
- Male repeated clients has higher default rate.

HOUSING TYPE:
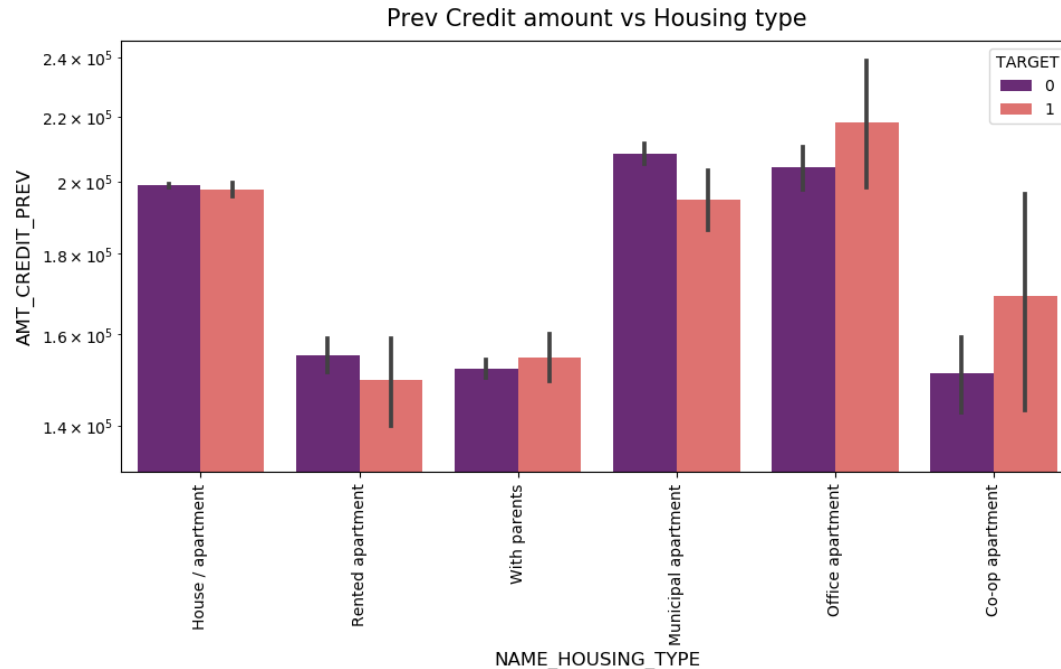
Prev Credit amount vs Housing type

- Loans for Municipal apartment is having higher repay capacity and high payment difficulties for Office apartments and Co-operative apartments

Conclusions:

- Credit Range 500000- 600000 has high payment difficulty with Secondary Education type, Small Family size for Loan type Cash.
- Clients with Academic degree and Student as Education type has no difficulties
- Income range between 1.2L - 1.4L has high payment difficulties with the age range between 30-45.
- Clients with Experience 41-50 has no payment difficulty.
- Payment Difficulty is high for buying a used car from high Populated Density and buying a home should be giving high priority.
- Office apartments and Co-operative apartments has high payment difficulties.
- Higher repay capacity for Commercial Associates with loan purpose buying a holiday home/land and a house and high payment difficulty for buying a car
- Loans for Municipal apartment is having higher repay capacity and high payment difficulties for Office apartments and Co-operative apartments
- Approving loans for age 30-45 for buying a car than <30 will be beneficial for bank.
- Approving a Loan for buying a garage for the age 45-55 has high repayment of loans
- Clients from Medium Profile Region has high repayment for the loan purpose of home and land and high default for Payment of other loans.
- Client Type - Repeater Male has payment difficulty than Women.
- Occupation Type, State servants has high repayment for loan purpose buying a house and Unemployed has high repayment for loan purpose repairs.