# Credit EDA Case Study

To identify driver variables behind loan defaults

**Sheetal Atre**
**Neeta Talekar**

# Business Objective

## The objective :

- To *identify* patterns which indicate if a customer who borrows a loan is likely to default.
- To highlight *patterns/trends* in customer loan repaying behaviour by analysing customer attributes and his current and previous loan attributes for banks to target such customers.

This information may be used by a banks or finance companies to take actions such as denying the loan, reducing the amount of loan, lending at higher interest rate for risky applicants.

# Approach to EDA Analysis

***Main steps*** for analysis on both current and previous datasets include :

- Data loading and basic data checks
- Data cleaning and creating derived columns
- Perform Univariate analysis for categorical, numerical and continuous numerical columns
- Perform Bivariate for combinations of categorical and numerical columns
- Perform Multivariate Analysis to identify trends or patterns
- From the observations determine driver variables
- Provide recommendations

# Initial Datasets

- ***Initial*** application dataset used for analysis contained:
  - 307511 rows,  122 columns in current data
  - 1670214 rows, 37 columns in previous data
  - Many columns are repeated in past and current data.

# Data cleaning

- ***Steps*** for data cleaning:
  - Remove columns with more than 50% null values
  - Impute columns:
    - Handling ***categorical*** missing values :
      - Introduce and replace with a new value "Unknown“ for columns like occupation type, since they have very large count of missing values which cannot be inferred.
    - Handling ***numerical*** missing values :
      - Impute missing values with median for columns which are already normalized and are related to building/flat where customer lives
      - Replace missing values with sample mean calculated using Central Limit Theorem for 1000 randomly selected samples each of size 500.
      - Remove/replace outliers

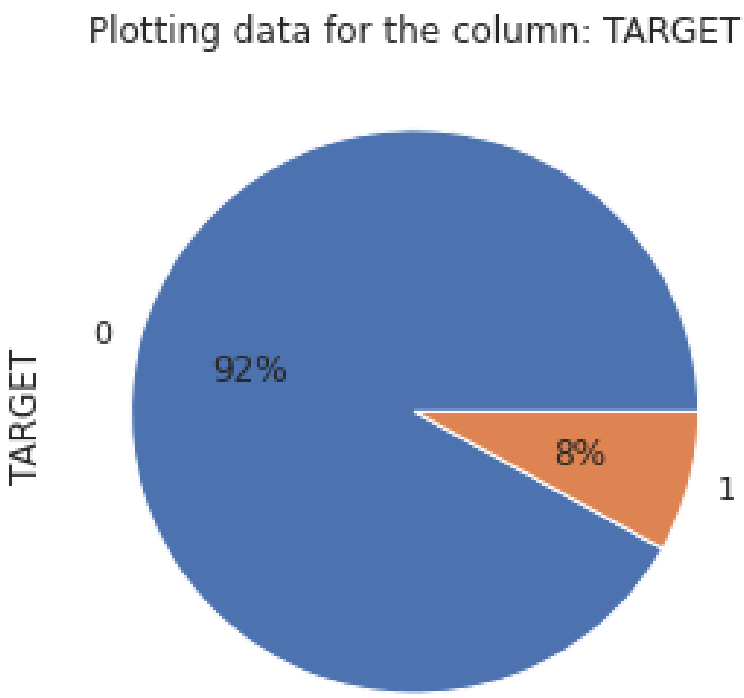
# Deriving columns

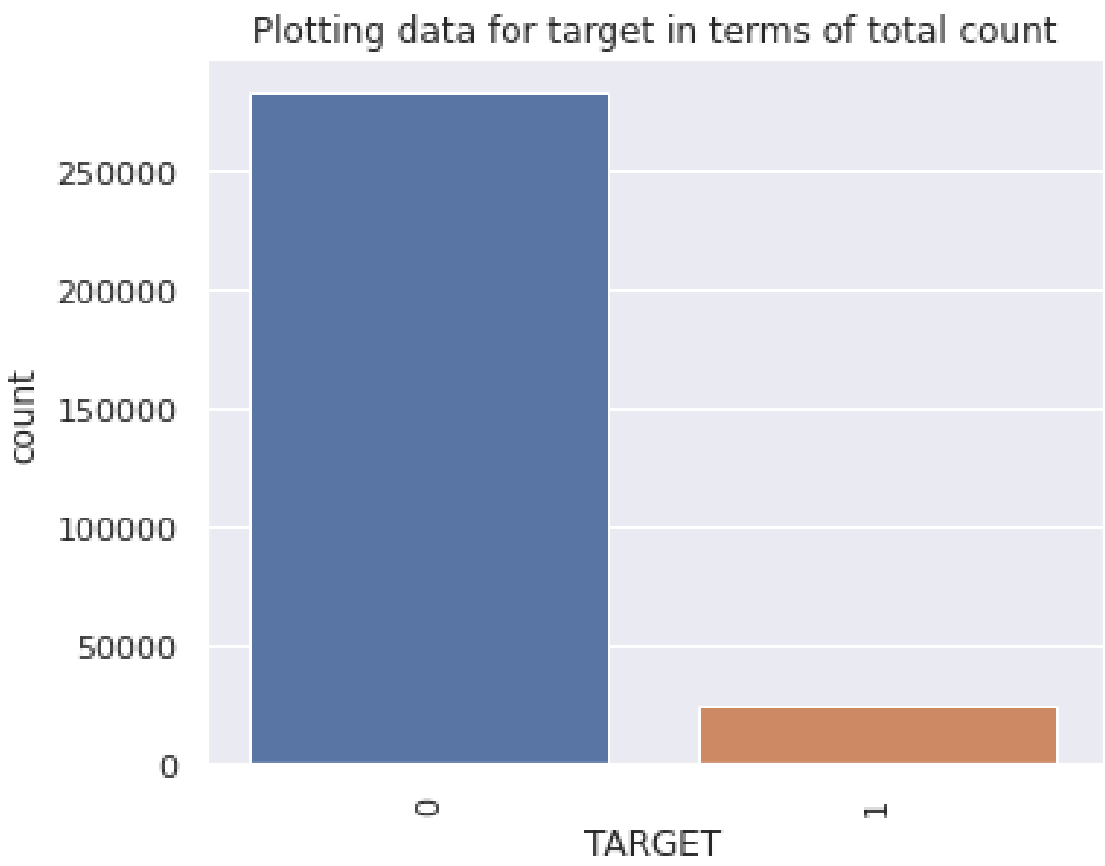
- Create ***new derived*** columns for large values:

  - Calculate “Age” from “Days of Birth” column
  - Convert No. of days of employment to No. of years of employment
  - Divide total income amount column by 100000(1 lakh) for ease of inspection
  - Binning of amount values like Income, Credit Amount, Primary interest rates

# Univariate Analysis of Categorical/Non-continuous numerical data - 1

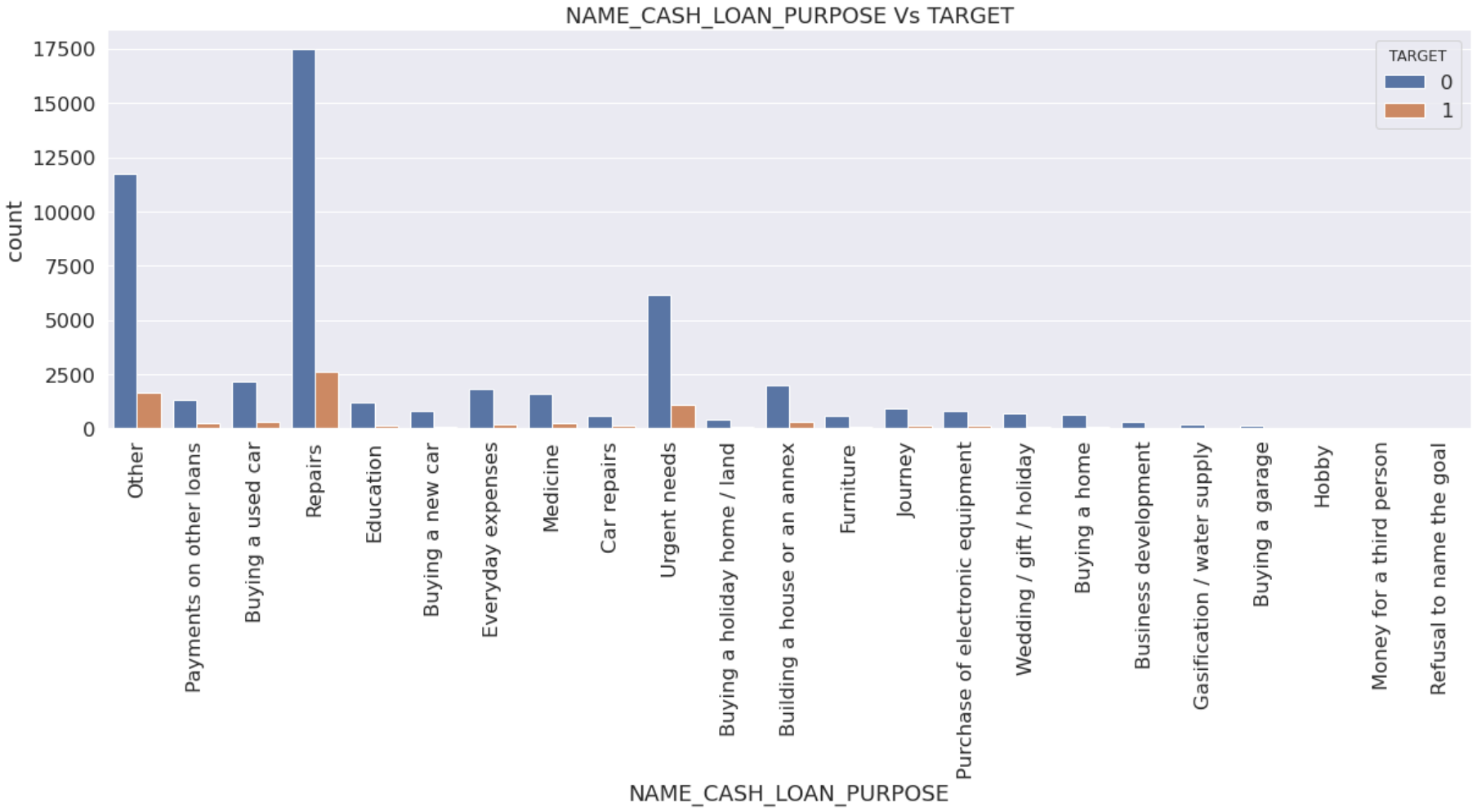
**Target Variable(Defaulters/Non-defaulters)** - 92% loans are fully paid.



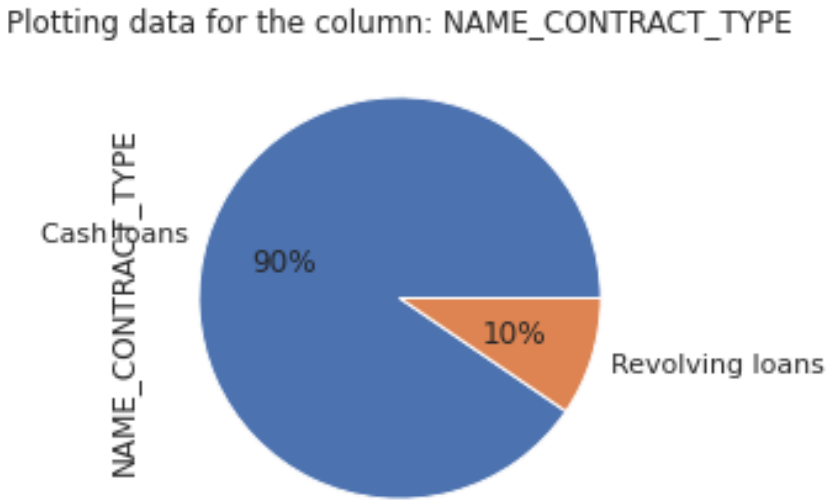
# Univariate Analysis of Categorical/Non-continuous numerical data - 2

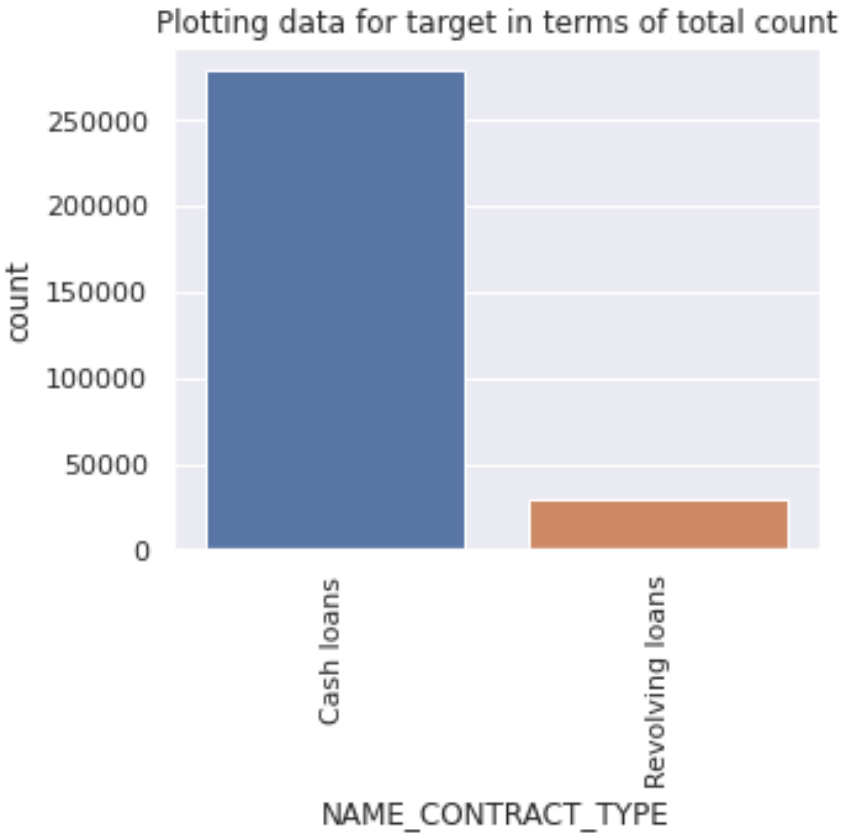
Loan purpose – In past data, about 34% clients applied for Repairs, 22% for Other, 12% for Urgent needs, 4% for buying a used car and 3.8% for building a house/annex



# Univariate Analysis of Categorical/Non-continuous numerical data - 3

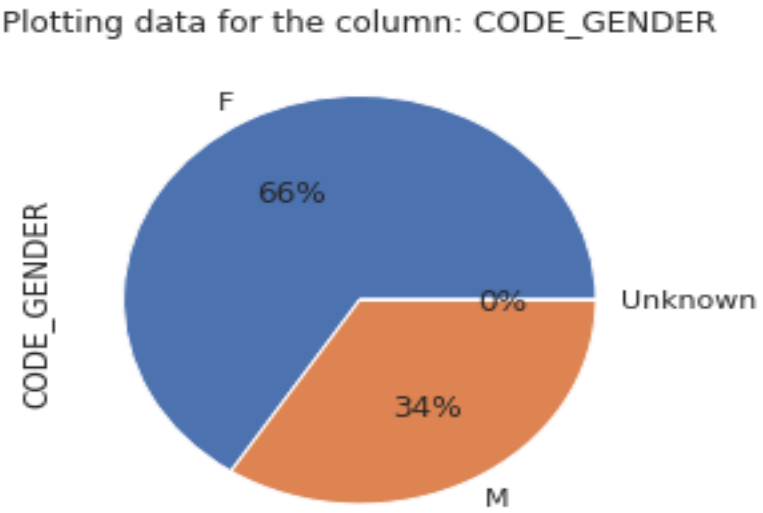
Contract type – 90% applications are for “cash” loans, rest for “revolving” loans

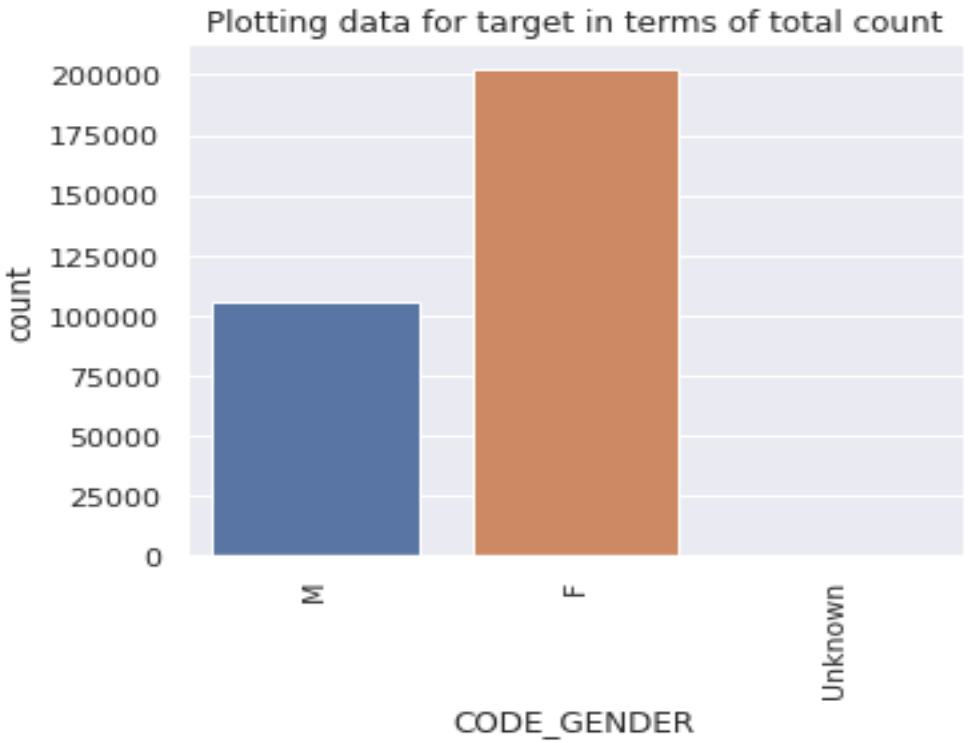


# Univariate Analysis of Categorical/Non-continuous numerical data - 4

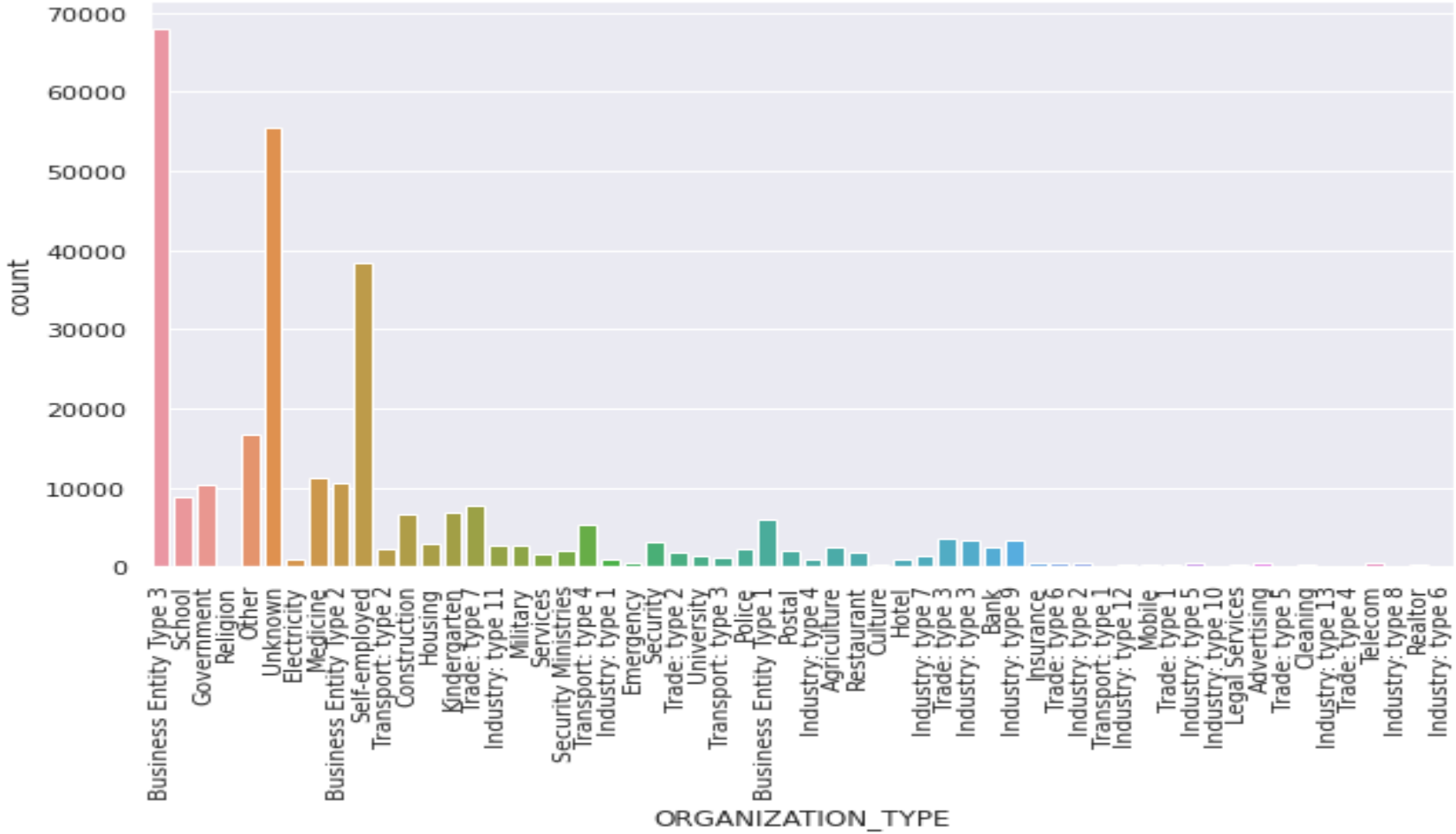
Client Gender - Almost 66% applicants are females



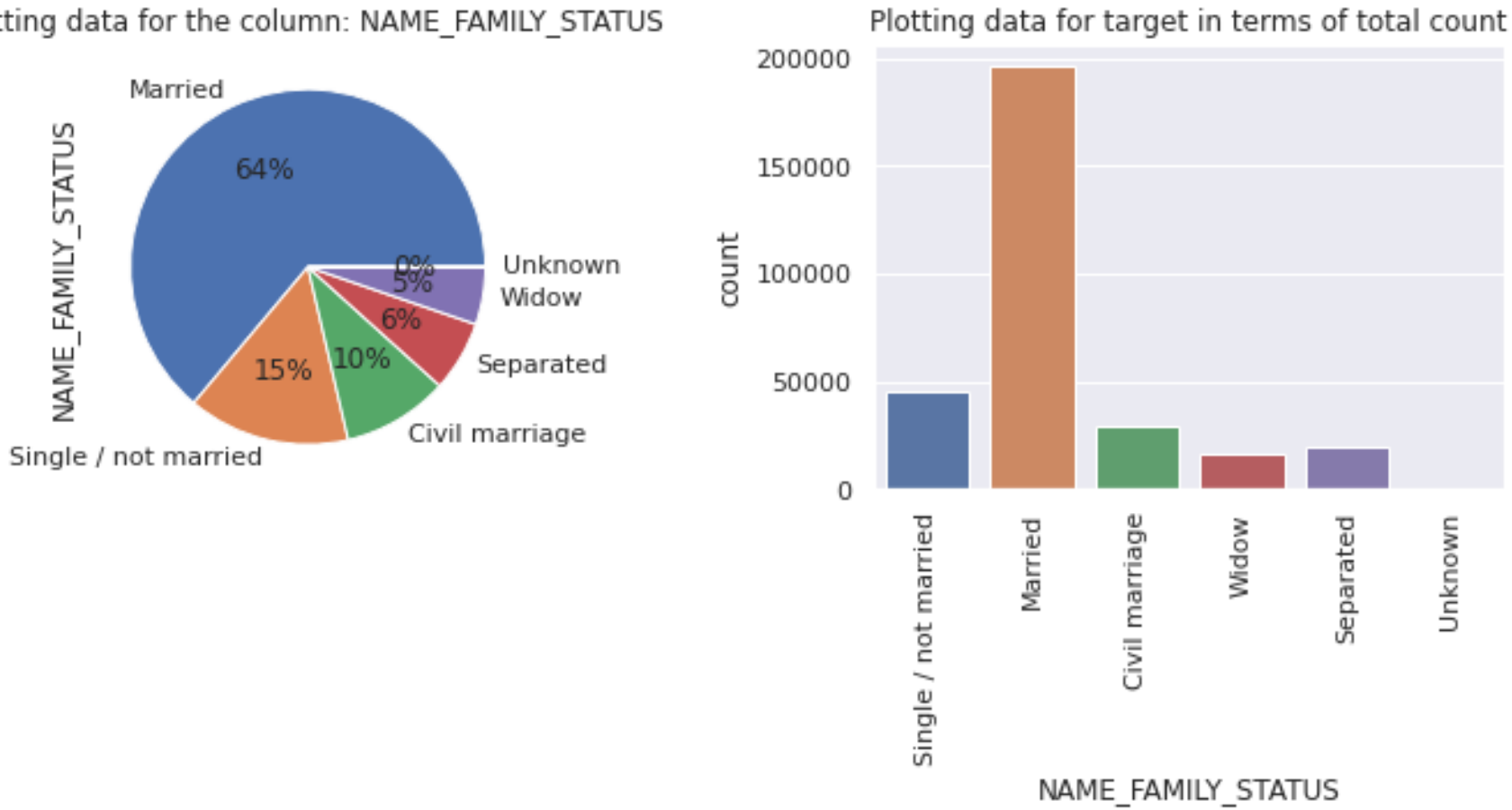
# Univariate Analysis of Categorical/Non-continuous numerical data - 5

Organization where client works - About  22% work in Corporations, LLCs, and Liability organizations, followed by 12% who are self employed. Large records have missing values for this column.



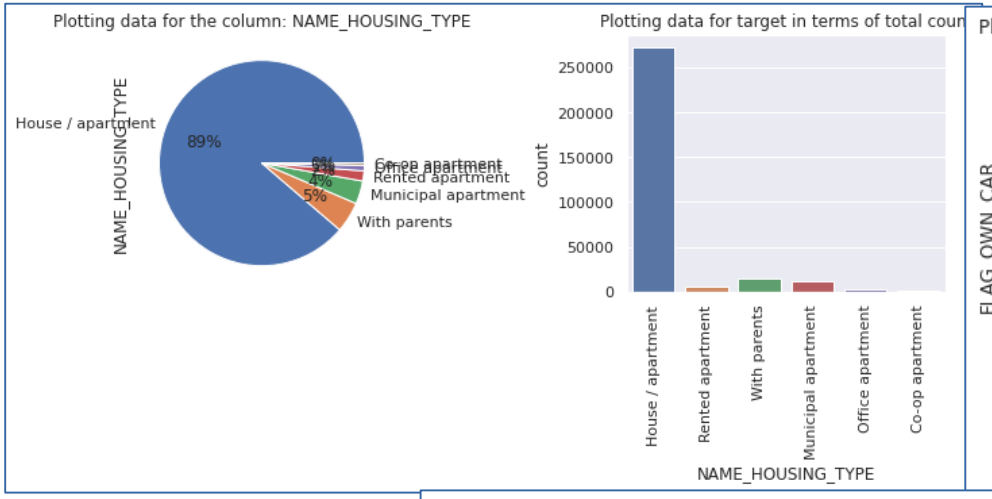
# Univariate Analysis of Categorical/Non-continuous numerical data - 6

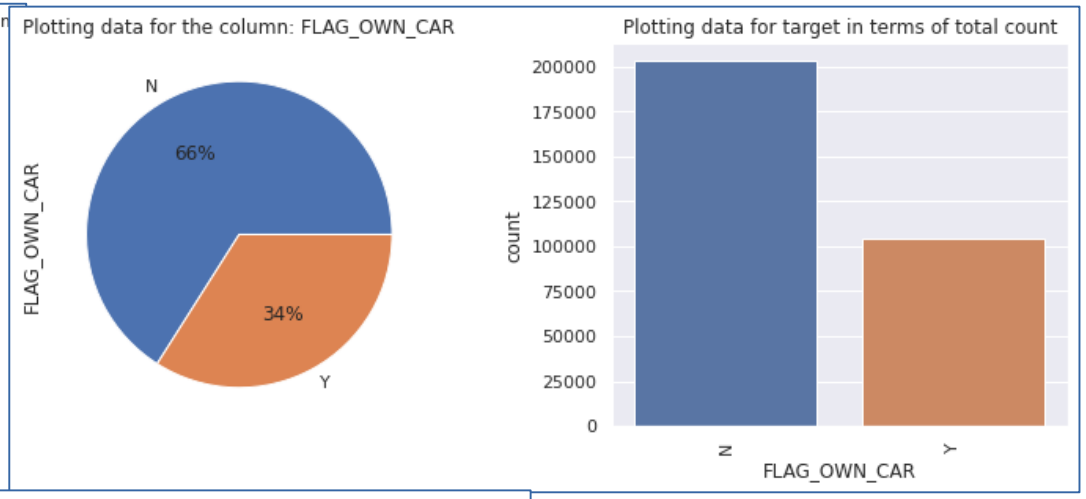
Family status of the client - 64% married people apply for loans

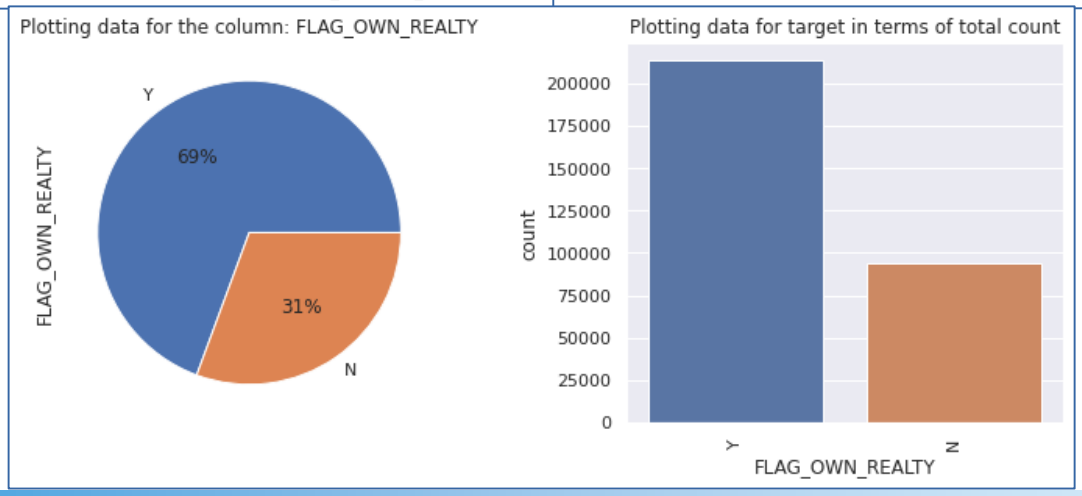


# Univariate Analysis of Categorical/Non-continuous numerical data - 7

Housing situation, Flat/House ownership, Car ownership flags - About 66% do not own car, 69% own a house or flat and 89% live in their own house rather than with parents or municipal apts.

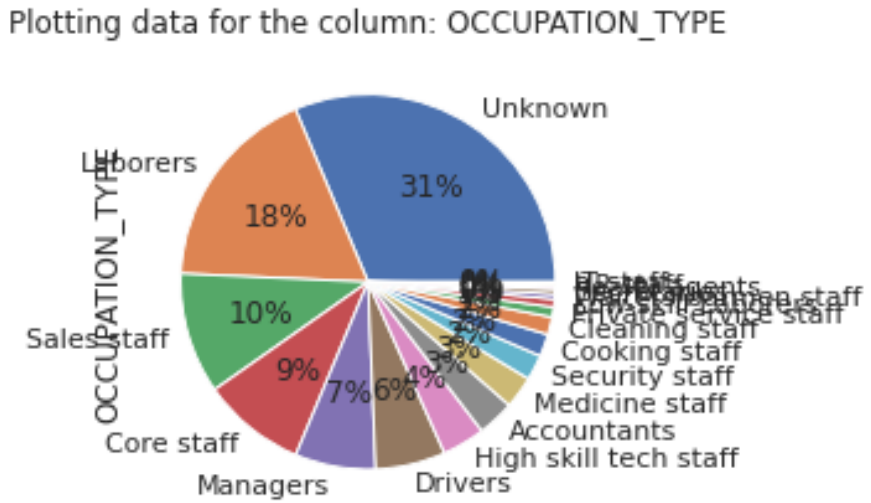


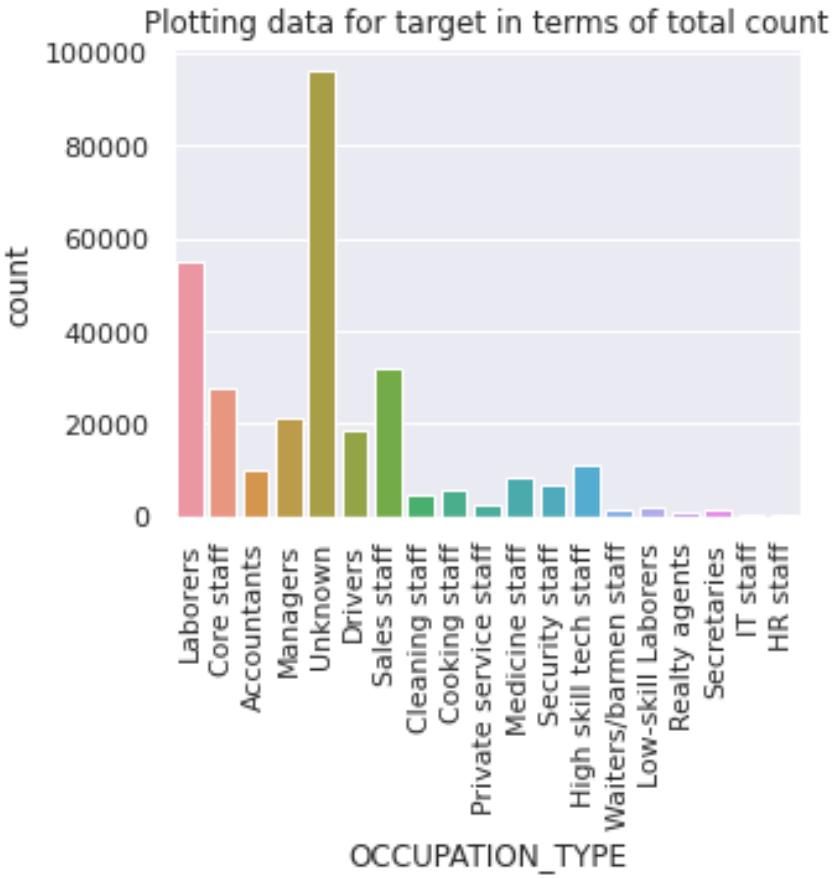



# Univariate Analysis of Categorical/Non-continuous numerical data - 8

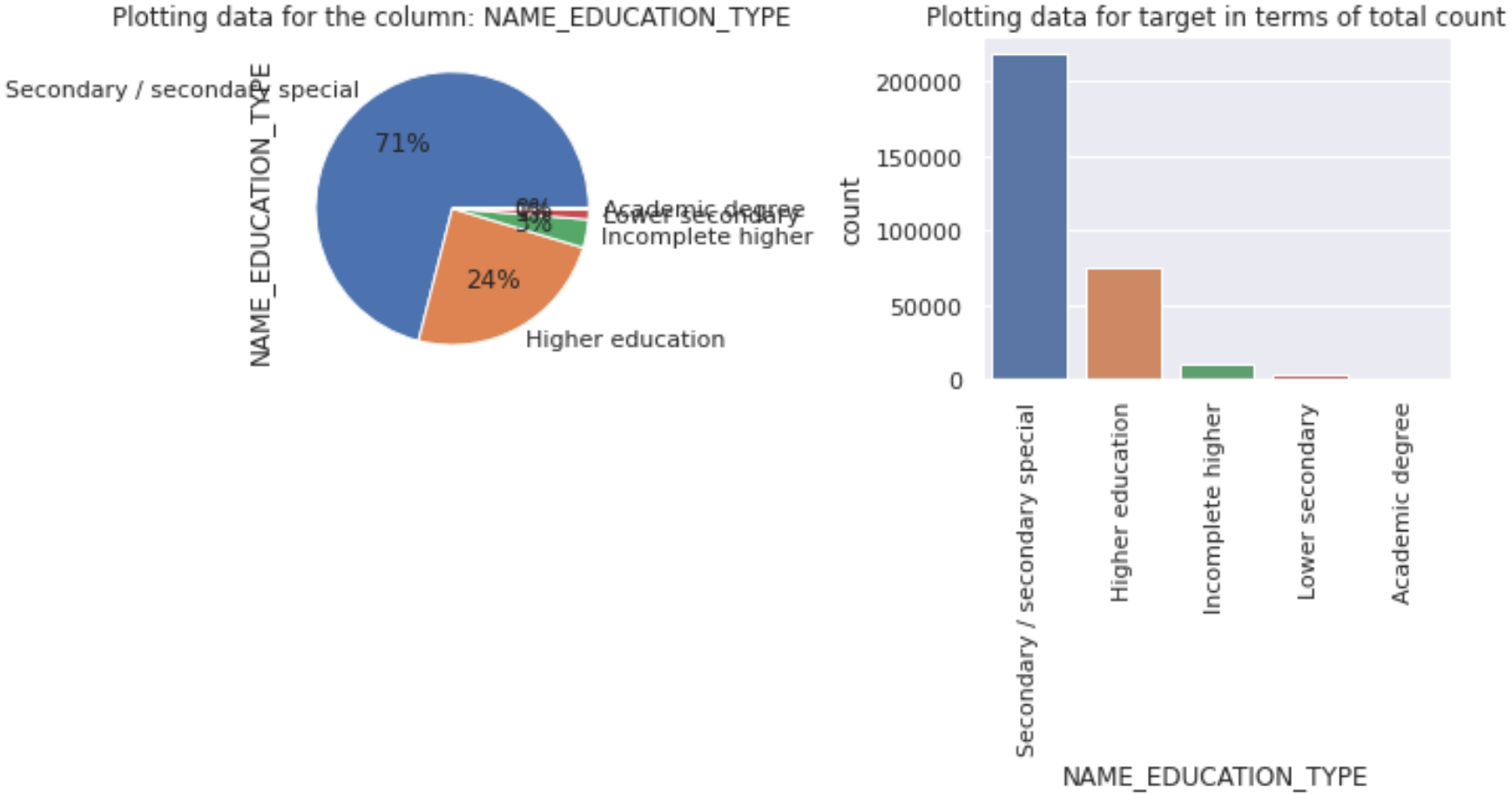
Client occupation – About 18% clients are Laborers, 10% are Sales staff and 9% are Core staff



# Univariate Analysis of Categorical/Non-continuous numerical data - 9

Client highest education - 71% applicants have secondary/special and 24% have completed higher education.



# Univariate Analysis of continuous numerical Data - 1

**Handling *Outliers*** :

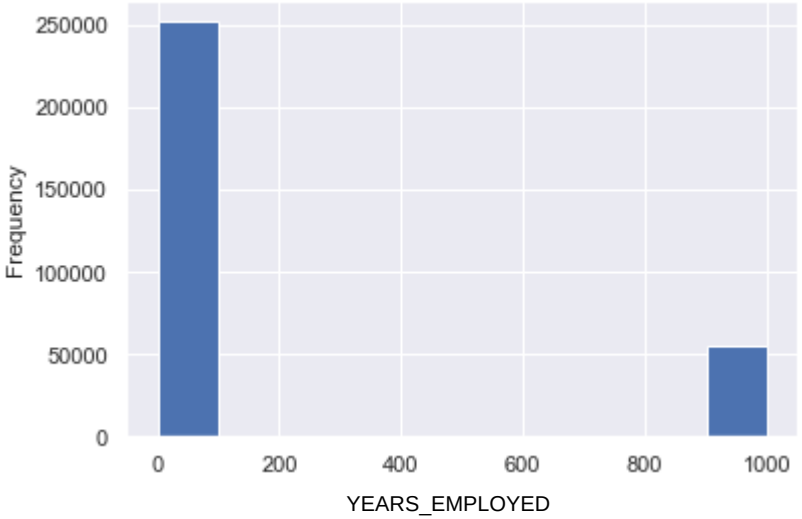
- Describe the data and check the difference between mean and max value for numerical columns.
- Some important variables in ***current*** data:

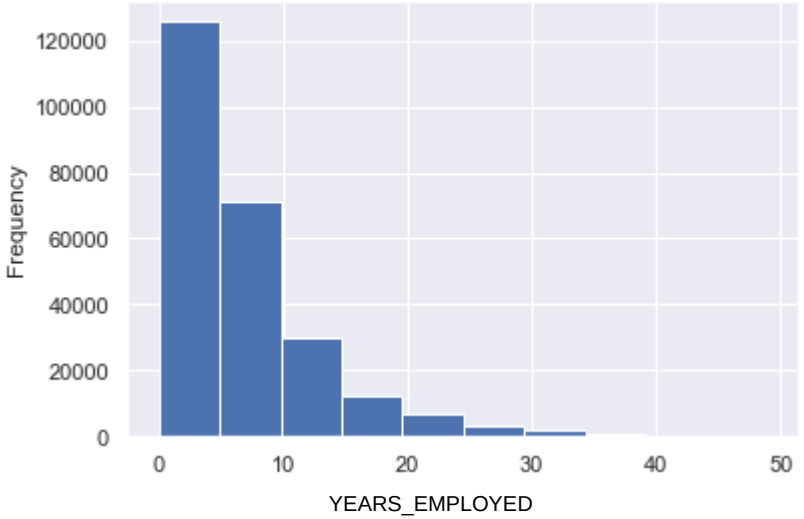
Days(Years) of employment since application - Replace outliers with more than 60 years(1000 days)  of days of employment with 0

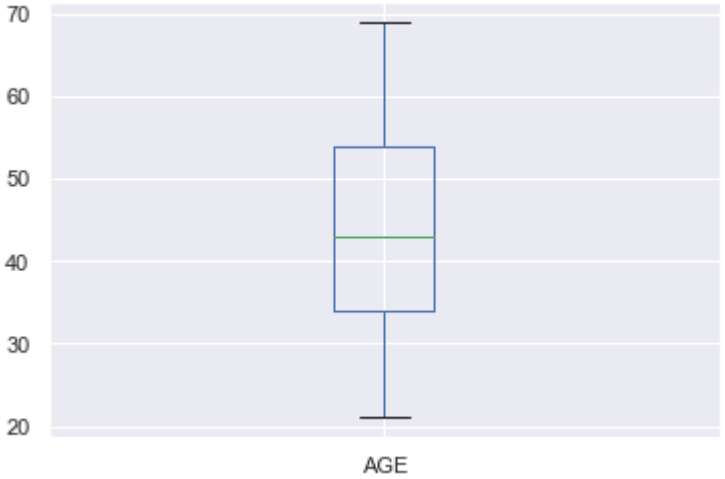
No of family members - Replace outliers > 10 members with -1

Income of client– Remove the records having more than 50.0 lakh income by comparing it with the average income for that occupation type.

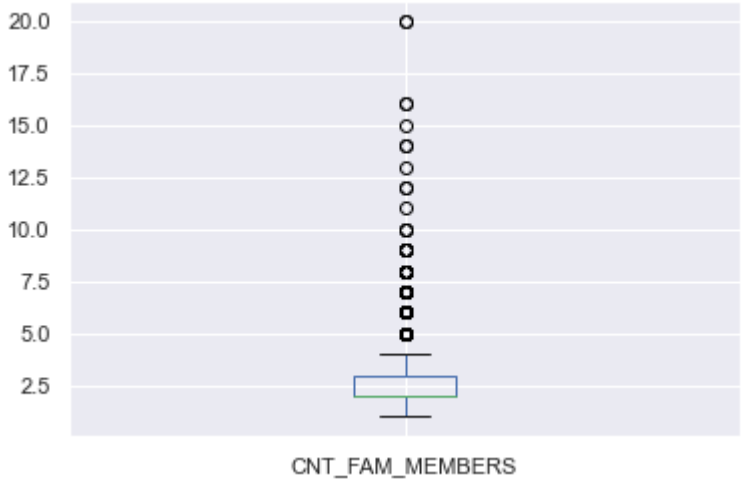
# Univariate Analysis of continuous numerical Data - 2

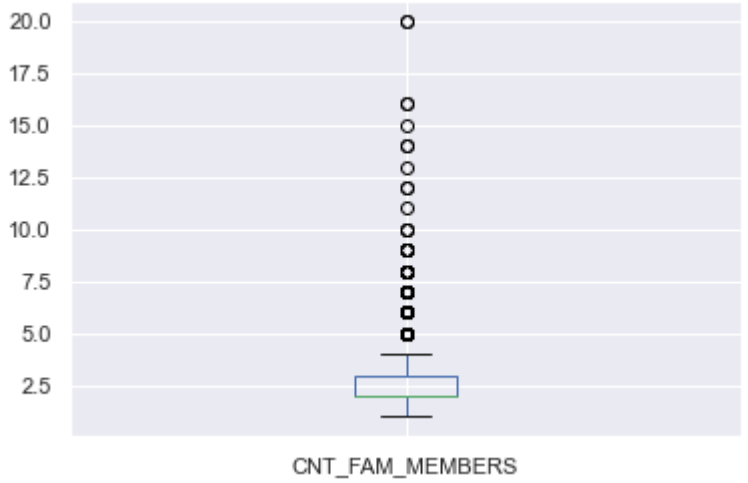


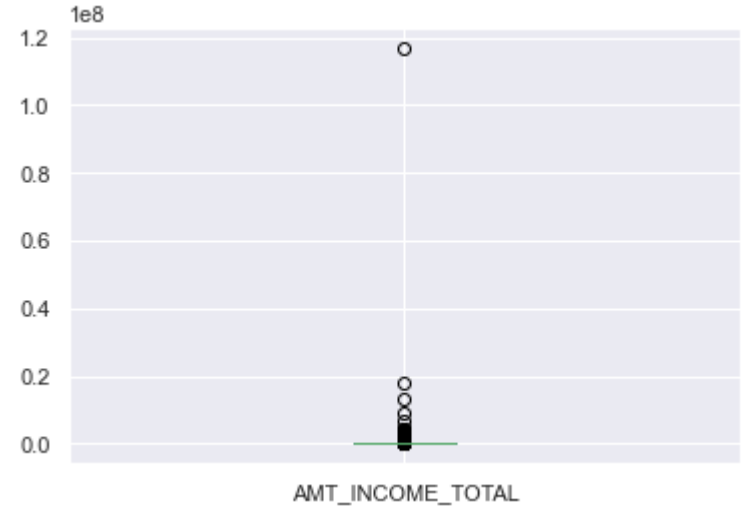



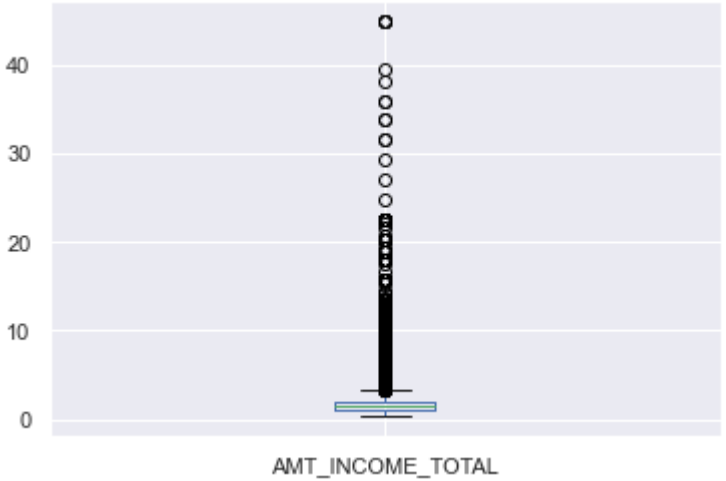



# Univariate Analysis of continuous numerical Data - 3



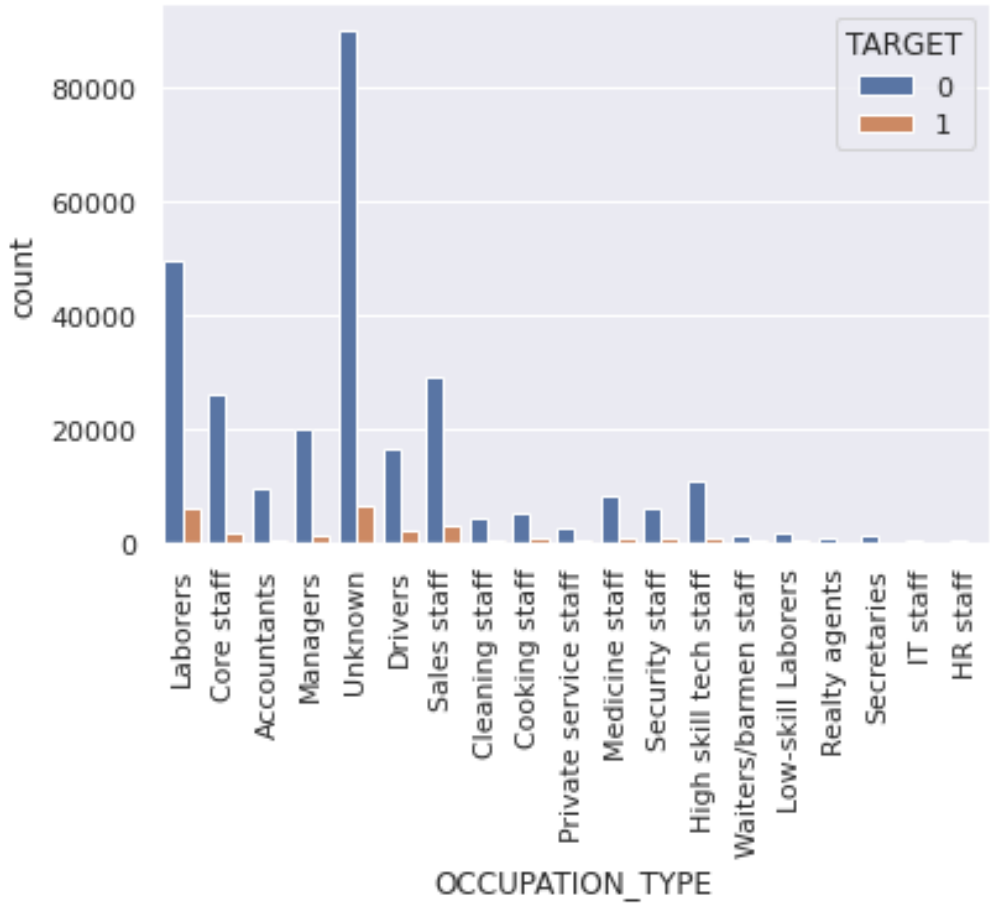



# Univariate segmented analysis of categorical columns - 1

Dataset ***Imbalance ratio***:

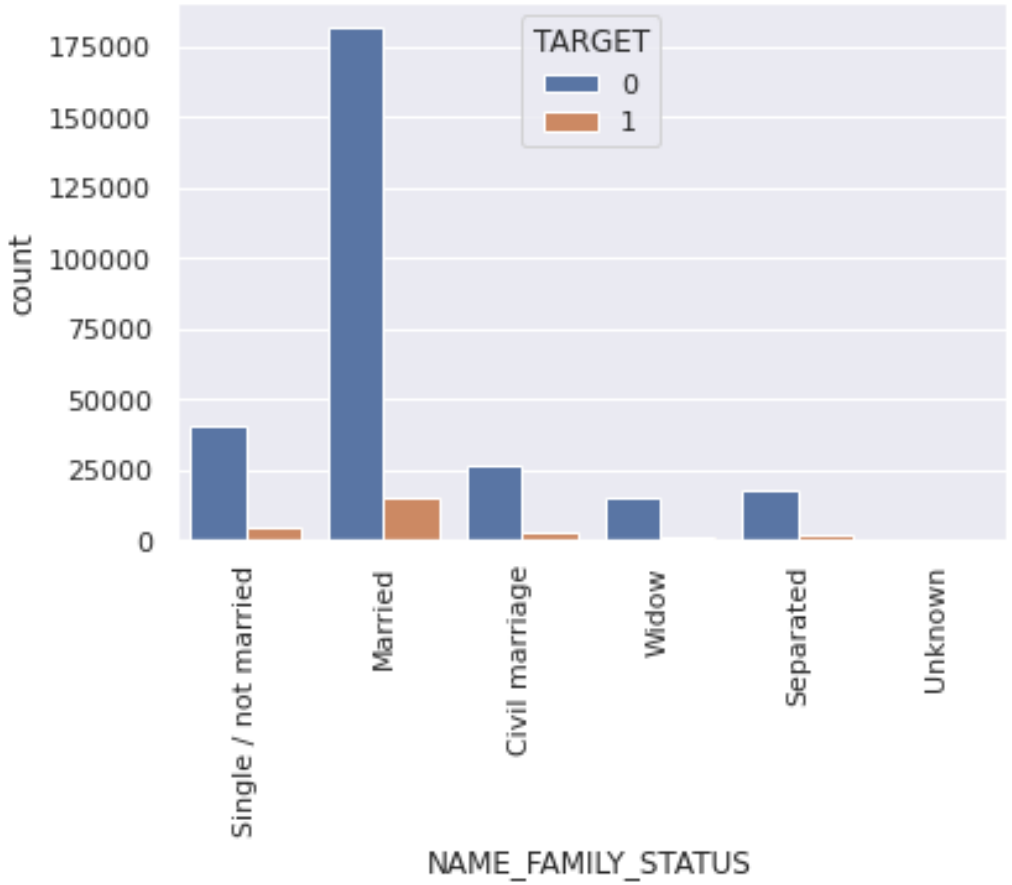
- The ratio of clients paying on time(TARGET == 0) and clients having difficulty to repay(TARGET == 1) is "**11.39**" indicating dataset is imbalanced.
- Due to large imbalance, we have separated analysis of the target variables to study defaulter/non-defaulter behavior.

Client occupation - Sales staff face difficulties than Core staff to pay in time. Laborers are able to pay loans in time, may be for lower loan amounts.

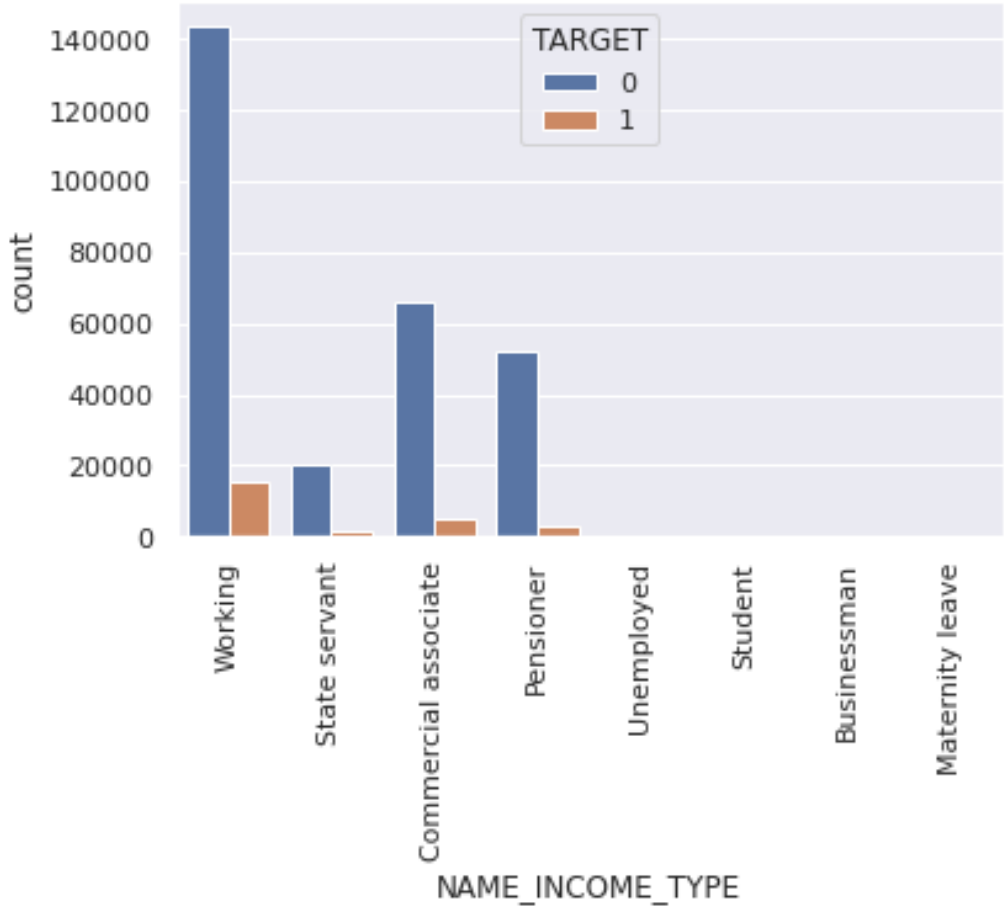


Client Gender - Female customers pay loan amount on time. This may be since banks target more female customers for lending loan.

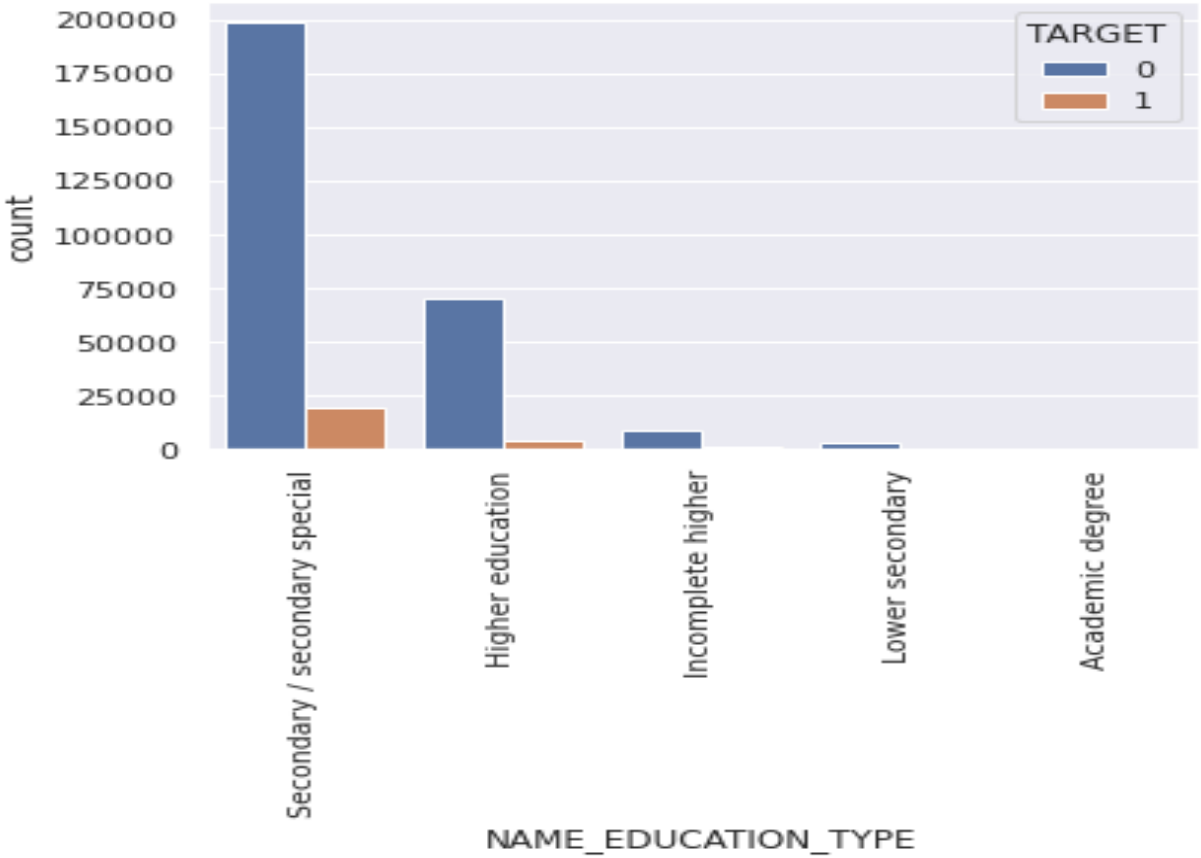
Family status - Widowers/Separated category don't have difficulty to repay loan amount in time.



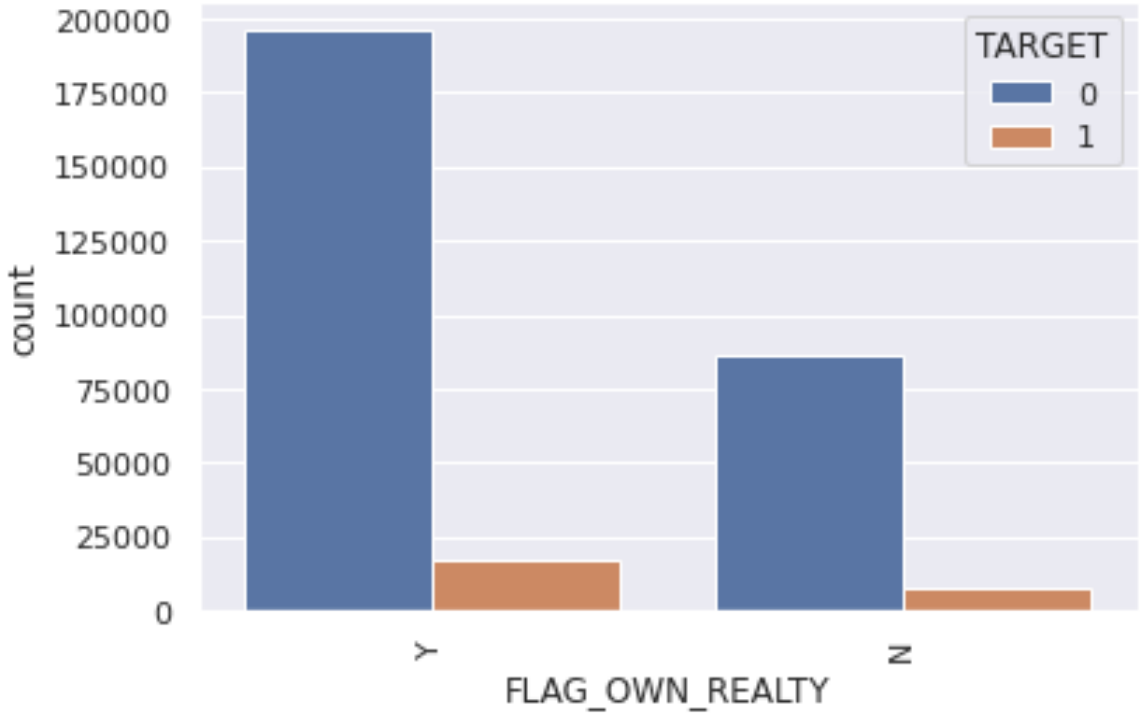
Income type of client - Working customers and pensioners tend to repay loans on time compared to businessman



- Client highest education - Customers with higher and higher/secondary education are most likely to take loans and make payments in time when compared to customers with academic degree.



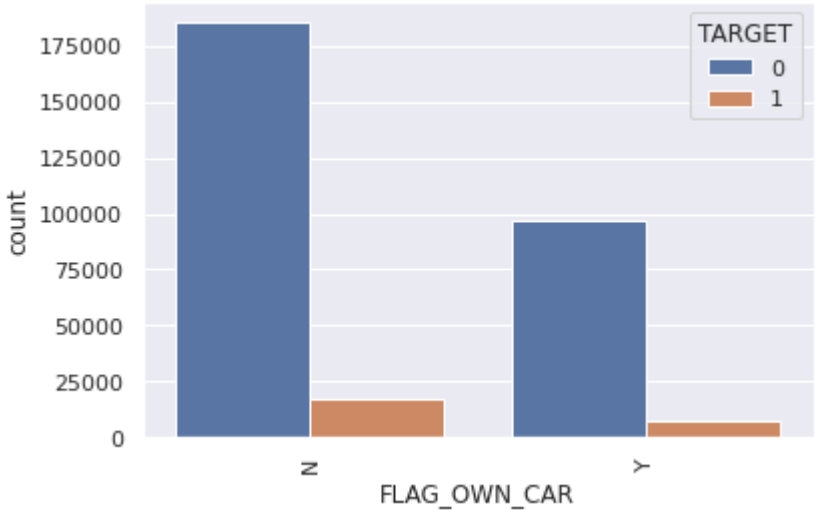
House ownership flag - Customers owning House/flat are most likely to make payments on time compared to those who don’t a house.

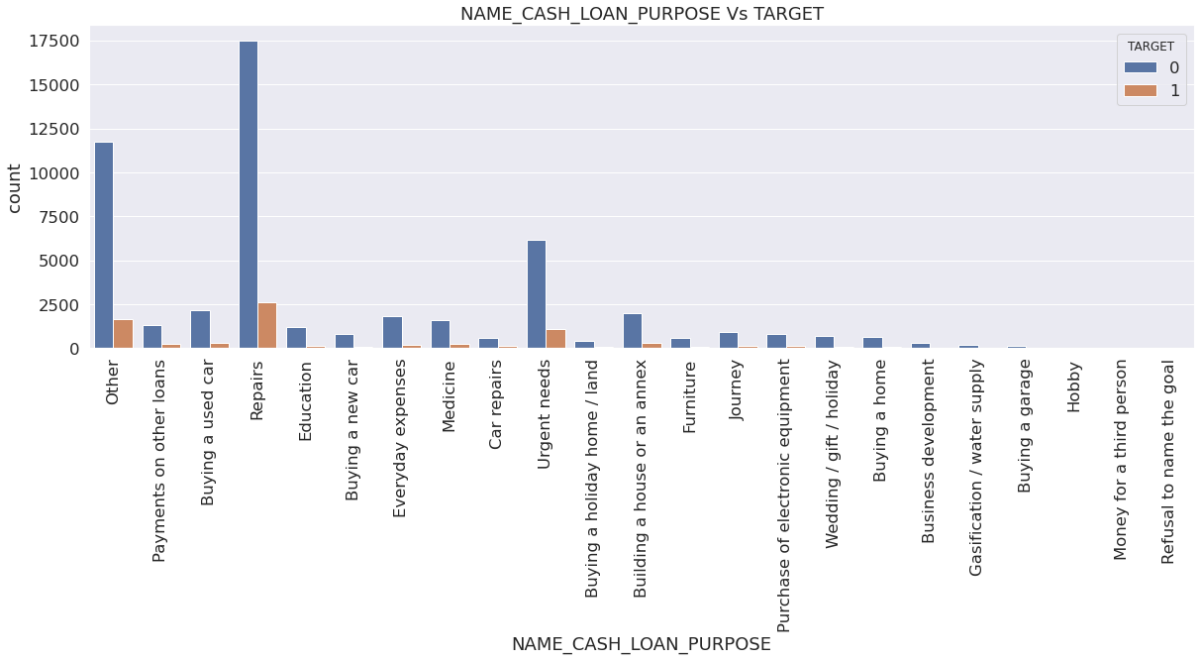


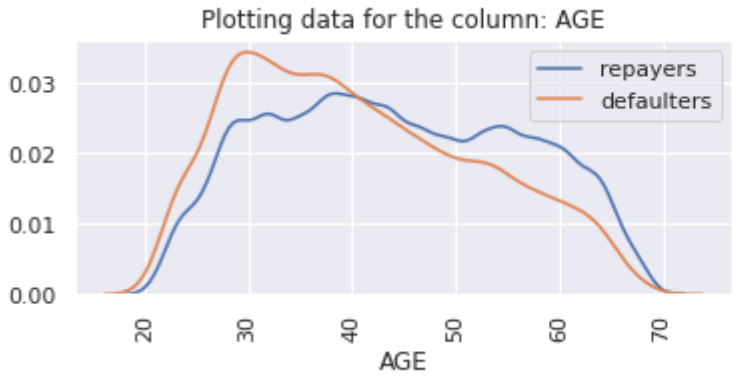
# Univariate segmented analysis of categorical columns - 2

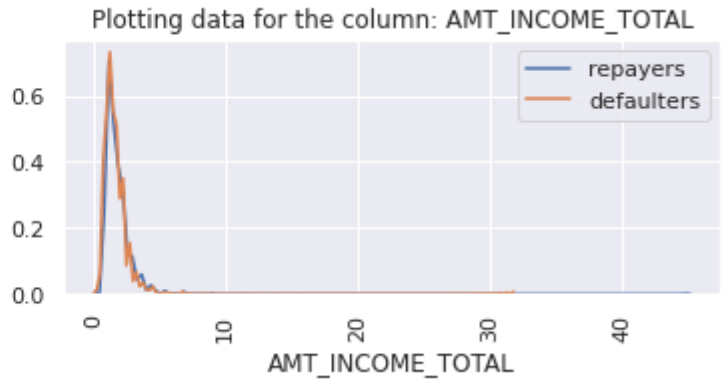
1. Car ownership flag - People not owning cars are able to repay in time.
2. Client Age(Days of birth) - Customers within the Age Group 20-40 shows higher number of Defaulters.
3. Client income- The defaulters count is high for people having income 0-5 Lacs
4. Loan Purpose – There is high count of defaulters for Repairs compared to Others and Urgent needs.

# Univariate segmented analysis of categorical columns - 3

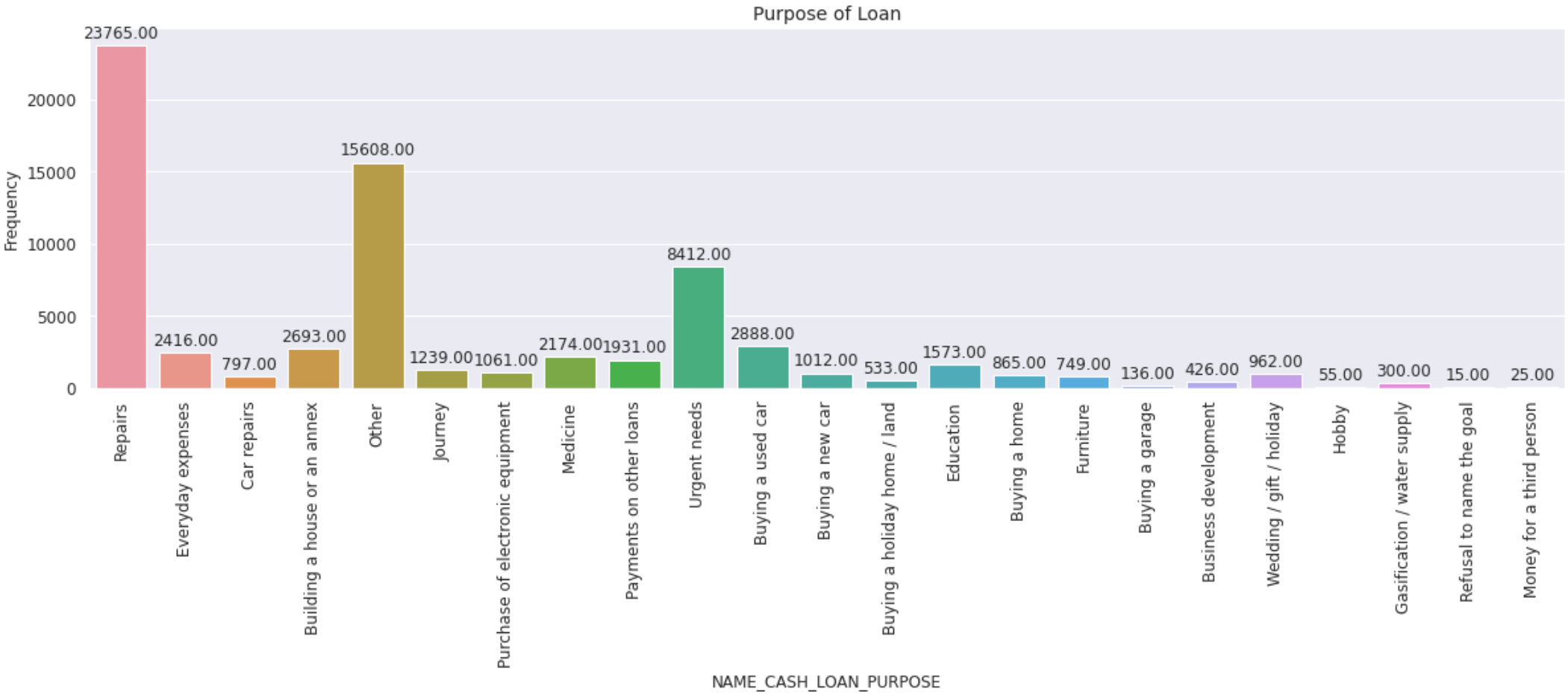


# Univariate segmented analysis of categorical columns - 4



# Bivariate/Multivariate Analysis - 1

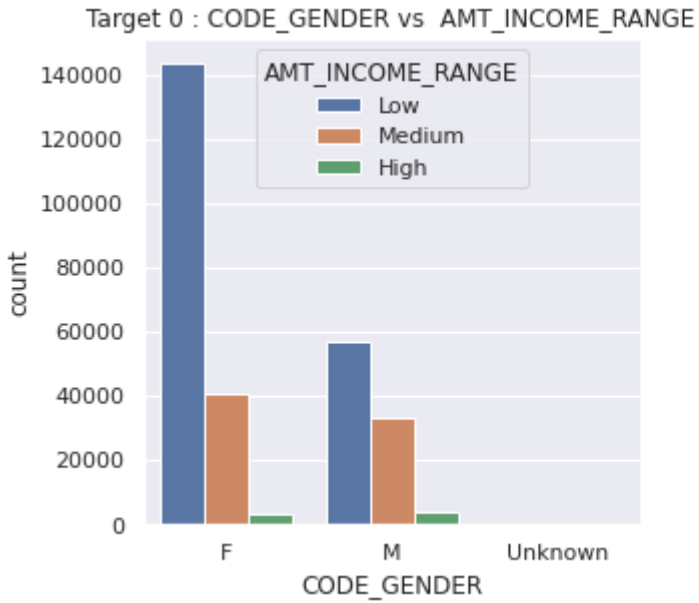
This analysis includes:

- Categorical to Categorical
- Categorical to Numerical and
- Numerical to Numerical column analysis

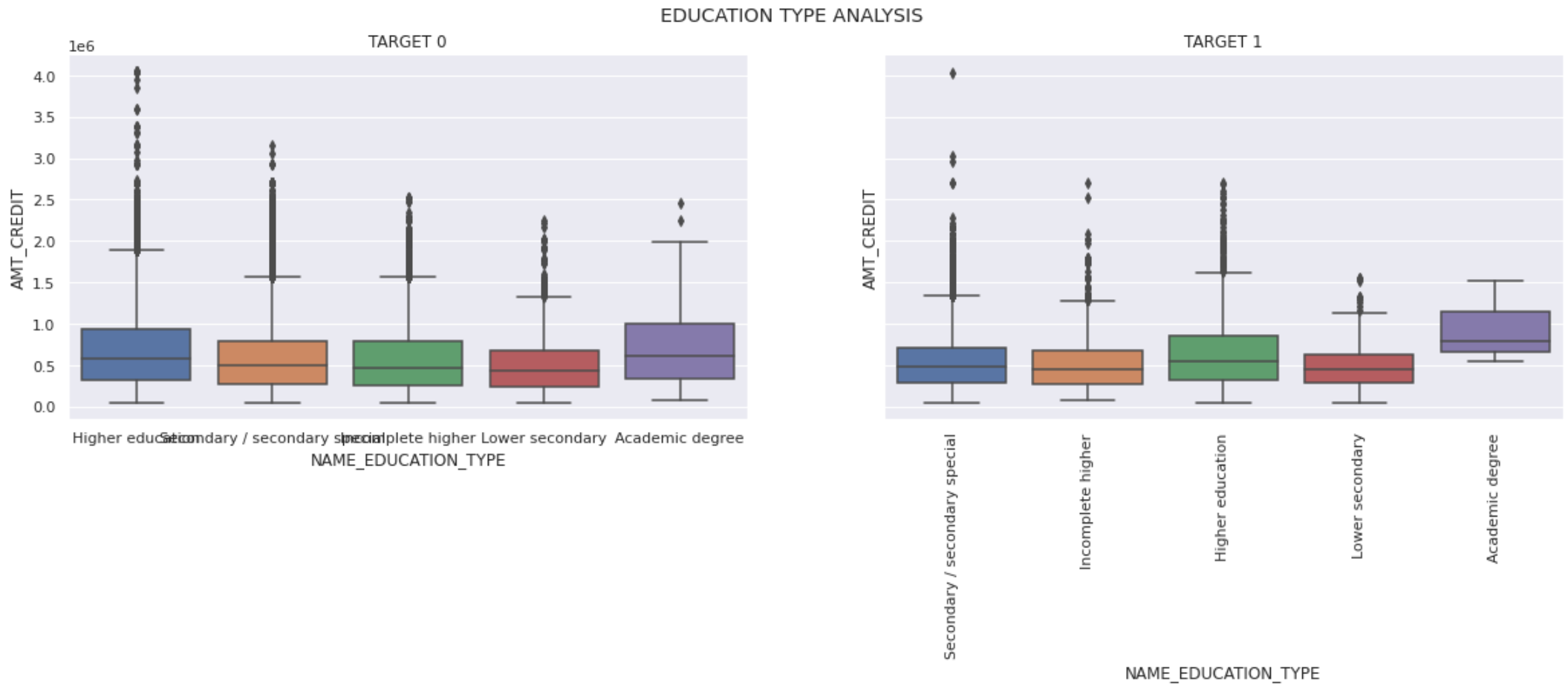
Analysis for :

- Target vs Income vs Credit Amount - Those who have repaid the Loan amount on time have higher Credit Value than defaulters.
- Target vs Client Gender vs Income for Defaulters and non-defaulters - Females with Lower income group are able to repay the loan in time
- Target vs Client Education vs, Credit Amount - Clients having Academic Degree Education take higher Credit than clients with higher secondary/special and secondary education

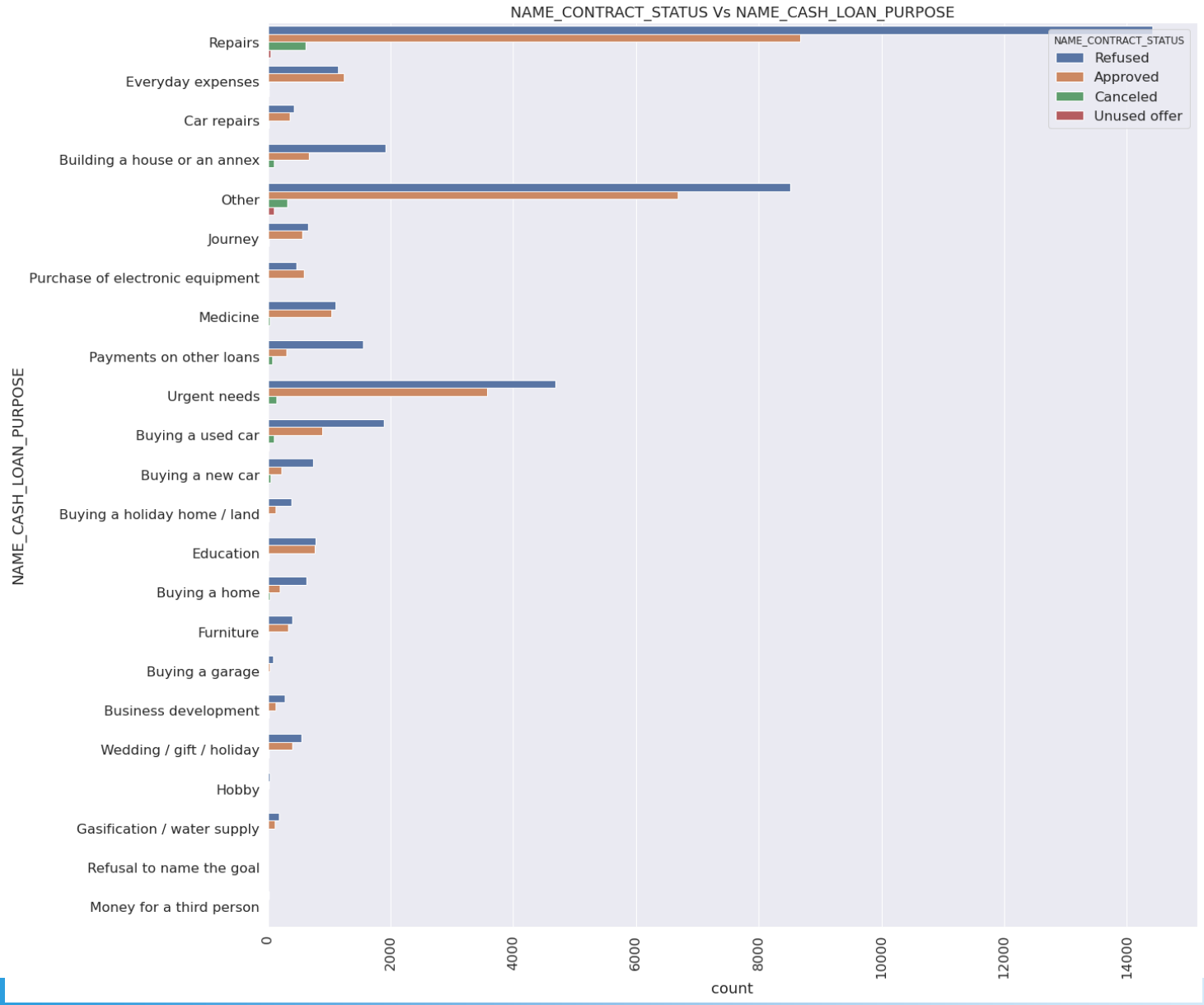
# Bivariate/Multivariate Analysis – 1.1




# Bivariate/Multivariate Analysis – 1.2



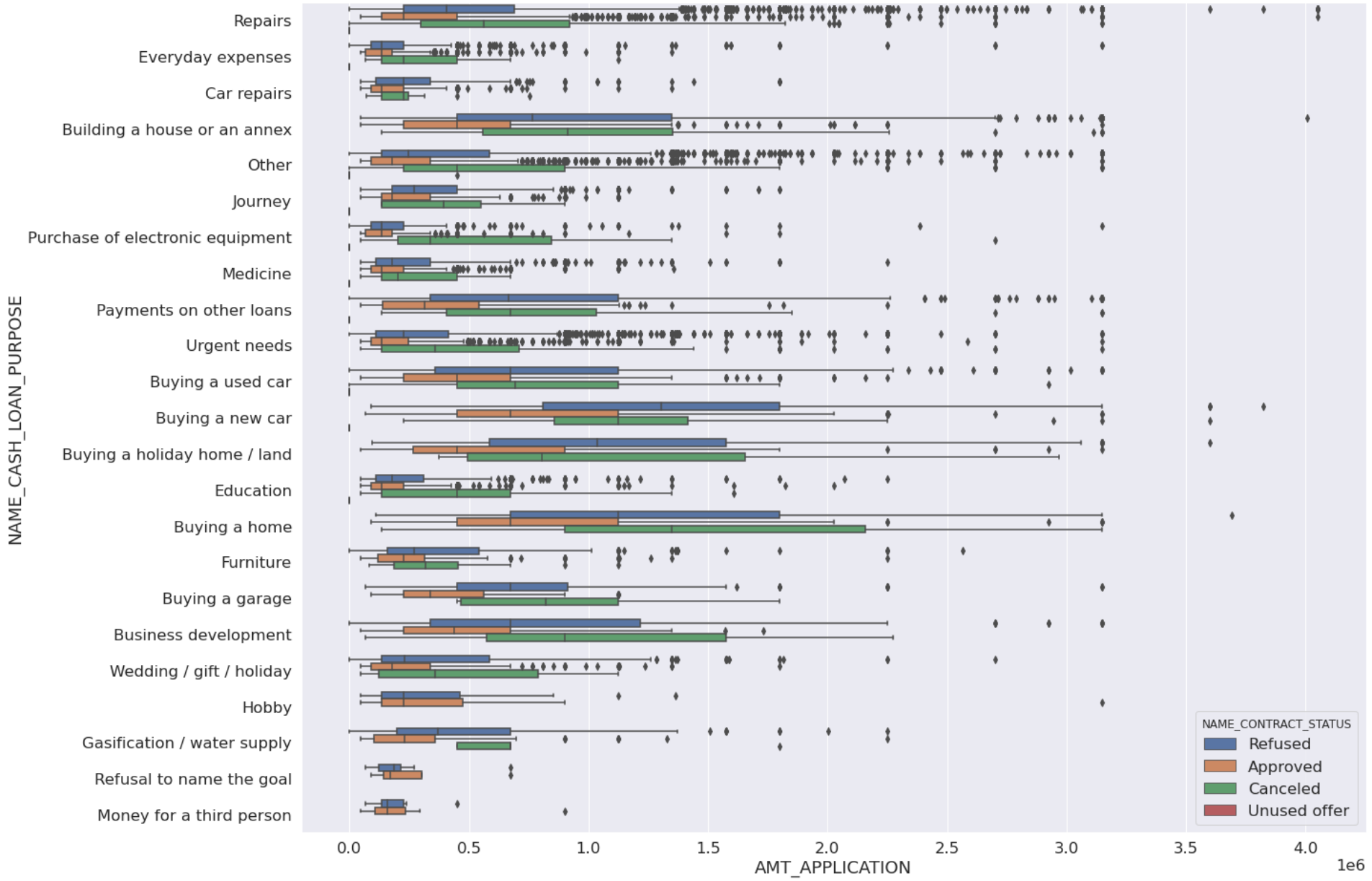
# Bivariate/Multivariate Analysis – 1.3



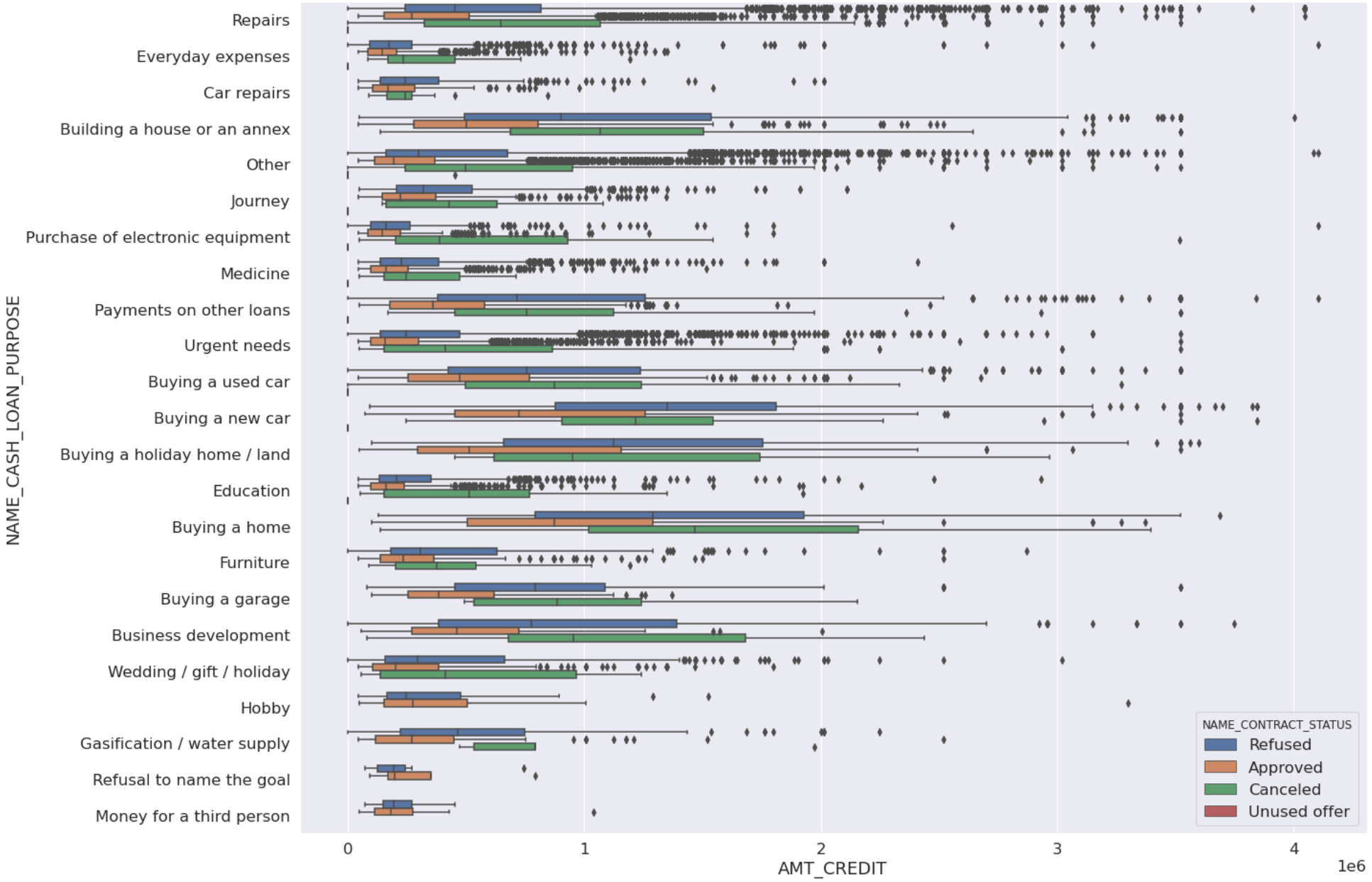
# Bivariate/Multivariate Analysis - 2

- Loan Amt applied by client in past vs Loan purpose vs Contract status – For eg – Highest loan amounts applications was for Repairs and most of the Refused loans were for HC, LIMIT, SCO status in the past
- Credit Amt Vs Loan purpose Vs Contract status – High Loan amounts were approved for buying a new car, a home or for buying holiday home/annex. There was high percentage of loan cancellation for purchasing electronic equipment and Business Development purpose.

# Bivariate/Multivariate Analysis - 4



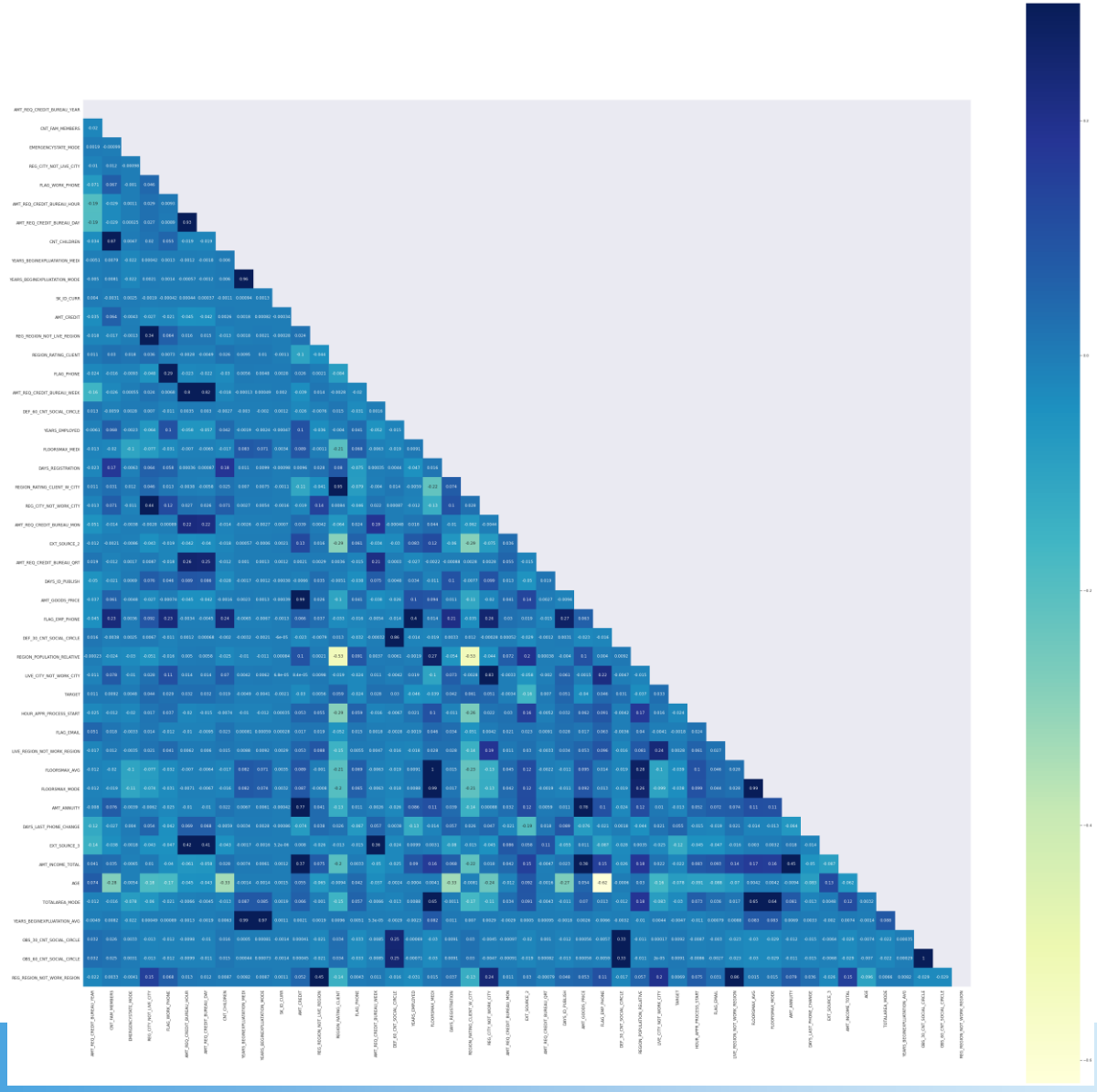
# Bivariate/Multivariate Analysis - 5



# Multivariate Analysis

- Top 10 correlations :

# Multivariate Analysis - 2



# Inferences

| Driver Variables and combinations for Defaulter derived from the analysis are: | | |
|---|---|---|
| Loan purpose | Contract type - cash loans, revolving loans | Client Gender |
| Organization where client works | Family status of the client, No of family members | Housing situation, Flat/House ownership, Car ownership flags |
| Client occupation | Client highest education | Days(Years) of employment since application |
| Income of client | Client Age(Days of birth) | |

| Combination of variables: | | |
|---|---|---|
| Target vs Income vs Credit Amount | Target vs Client Gender vs Income | Target vs Client Education vs, Credit Amount |
| Loan Amt applied by client in past vs Loan purpose vs Contract status | Credit Amt Vs Loan purpose Vs Contract status | |

# Recommendations

Below recommendations are suggested to bank/finance company for portfolio and risk assessment:

- Loans for Repairs, Urgent needs purpose can be allocated with higher interest rate, or detailed investigation of client.
- Bank should target clients having own flat, and no previous car loans
- It is safer to provide discounts or loans to females even having lower income range.
- Labourers Sales staff faces more difficulties to repay loans, hence these categories can be considered for small loan amounts or long term of loans.
- It is safer to target customers between age group 40-60 years
- It is safer to provide loan to customers with income more than 5 lakhs.
- Previous non-defaulters can be safely provided higher loan amounts or with lower interest as they are low-risk.
- Highly educated clients with higher credit in the past are more likely to repay loans in time, hence they can be targetted for reasons like buying Home, buying car, etc
- There was high percentage of loan cancellation for purchasing electronic equipment and Business Development purpose. Bank should investigate this further.

Thank You