CREDIT EDA CASE-STUDY

By FIZA & AYUSH

PROBLEM STATEMENT

- This case study aims to identify patterns which indicate the driving factors (or driver variables) behind loan default

- In other words 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

TWO TYPES OF RISKS FOR THE BANK

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
- Hence this case-study is to mitigate the risks for the bank

UNDERSTANDING RAW DATA

- Analysis is done on two data set -
  - Loan Application Dataset
  - Previous Application Dataset

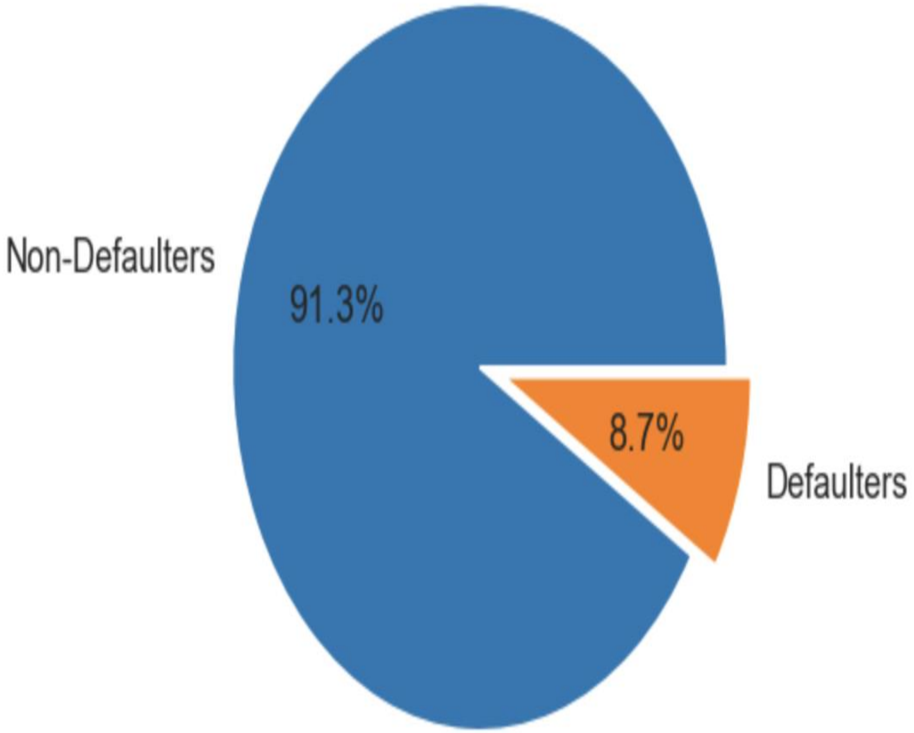
- **Loan Application Dataset** - Contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties.**

- **Previous Application Dataset -** contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

ANALYSIS APPROACH

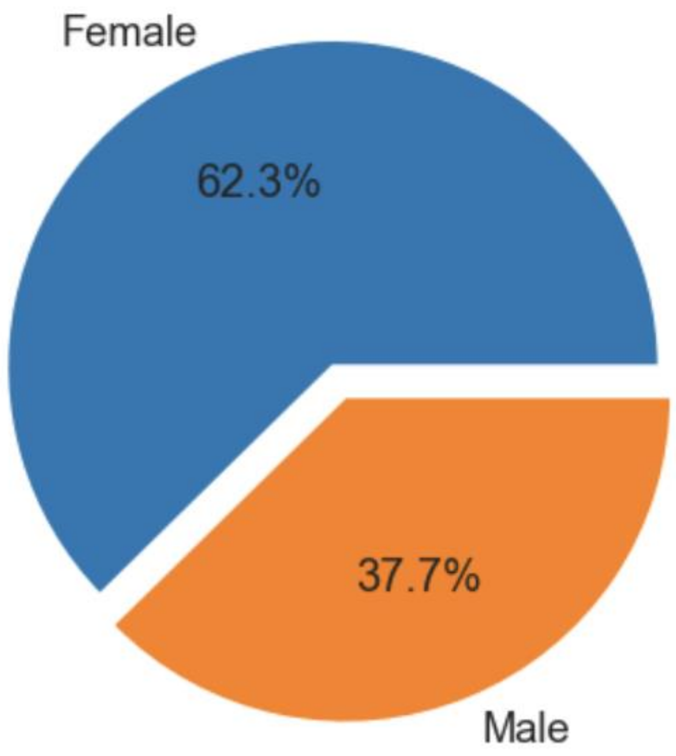
- Identifying the missing data and use of appropriate method to deal with it.
- Identifying the outliners in the data.
- Finding Imbalance and depicting the ratio of Imbalance
- Inferencing top 10 correlation for the **Client with payment difficulties** and **all other cases**
- Inclusion of Plots and visualization approach for better understanding of the Analysis
- Summary and conclusion depicting the further steps to be taken

DEFAULTERS AND NON-DEFAULTERS DATA IMBALANCE

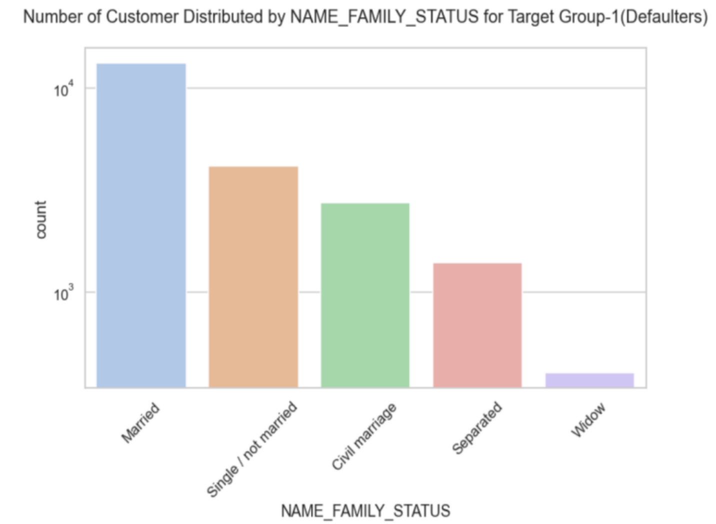


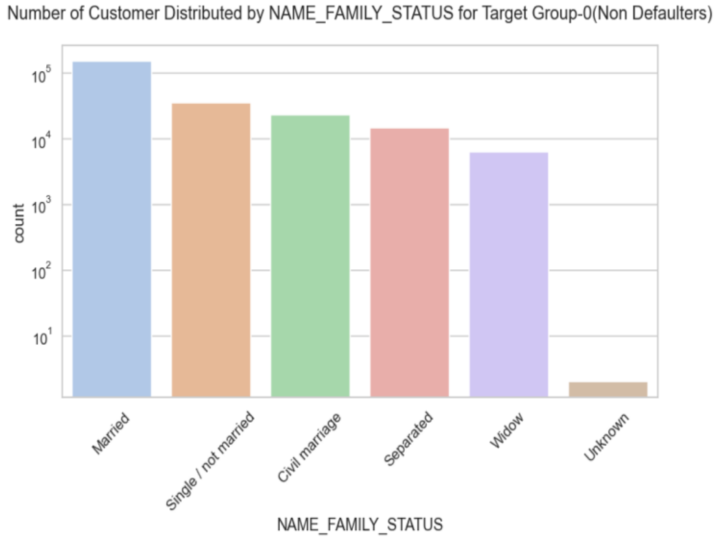
- Here after calculating Imbalance percentage for target column, we can clearly infer that Target_1 i.e., Defaulter are **10.55** times less the Non-Defaulters
- Defaulter are **8.7 %** of the total data and Non Defulters are **91.3%**
- Hence **Ratio** of Data Imbalance is **10.55**

GENDER IMBALANCE IN THE DATA



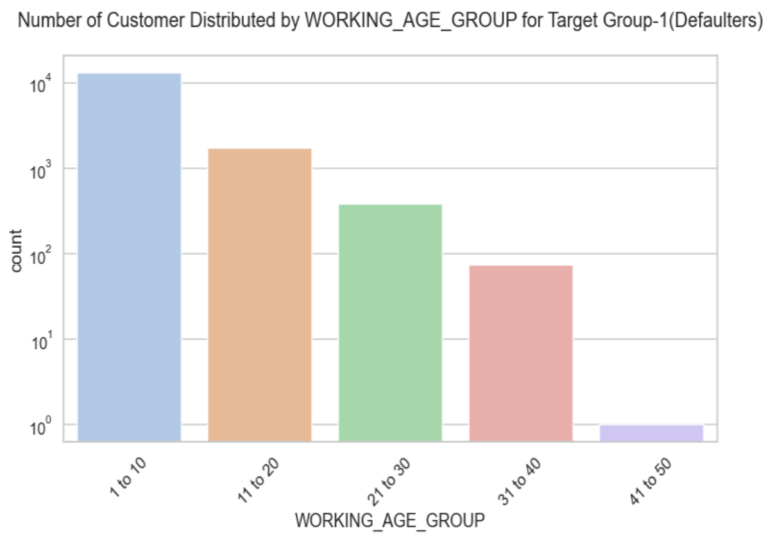
- Calculating Gender Imbalance we can determine that Female are 62.3% and Males are 37.7%
- Clearly more Loan customers are female.

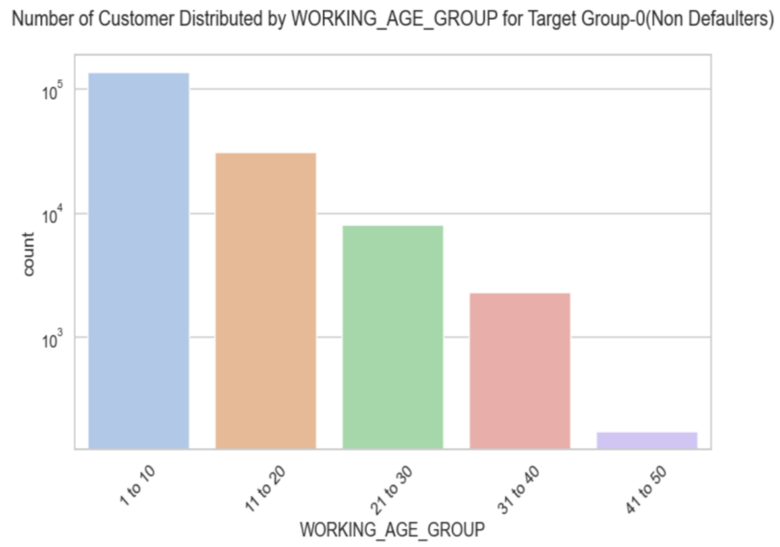




FAMILY STATUS DISTRIBUTION

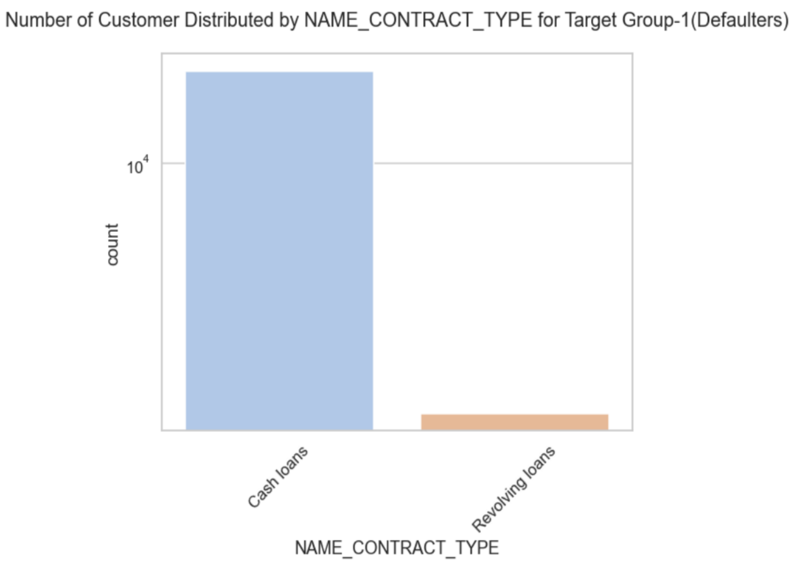
- This is Family Status data for Defaulter and Non-Defaulters
- We Can infer that Except married all the Family statuses are less in the defaulters.
- Hence, we can loosely say that People who are single, civil married, separated, widow are **less chances of defaulting the loan**

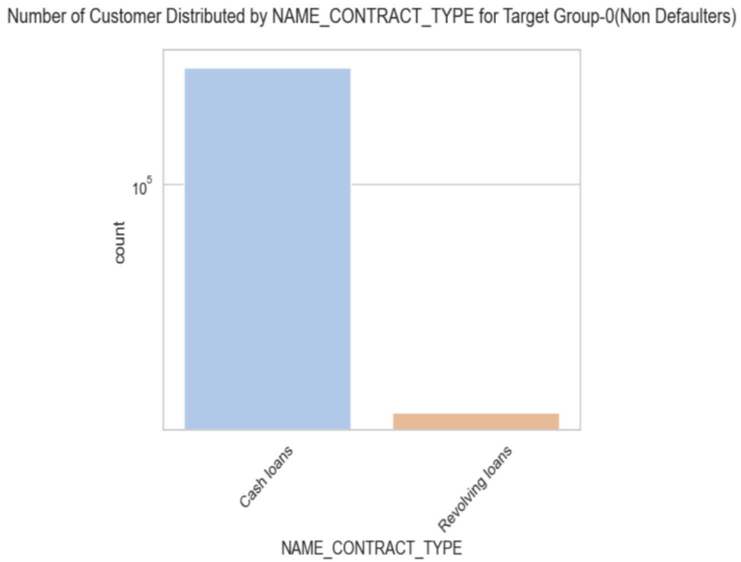




WORKING SINCE DISTRIBUTION

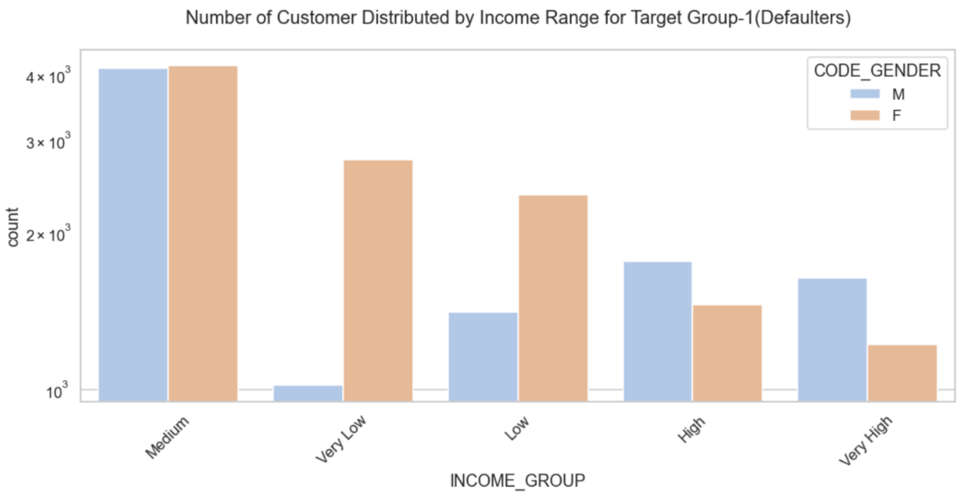
- This is graph shows the distribution of **How many days before the application the person started current employment**
- Plots of both Defaulter and Non-Defaulters are quite similar
- We can clearly see that most of the defaulters as well as Non-Defaulters are Working since 1 to 10 years

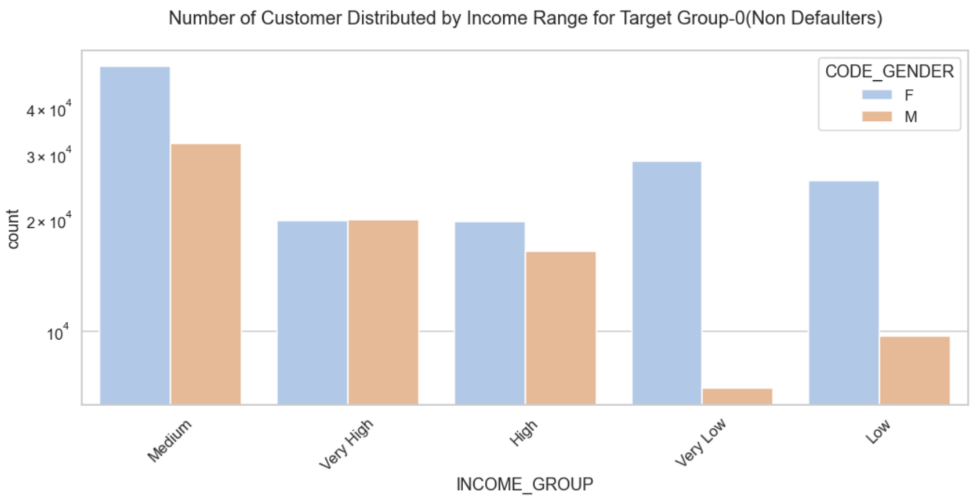




CONTRACT TYPE DISTRIBUTION

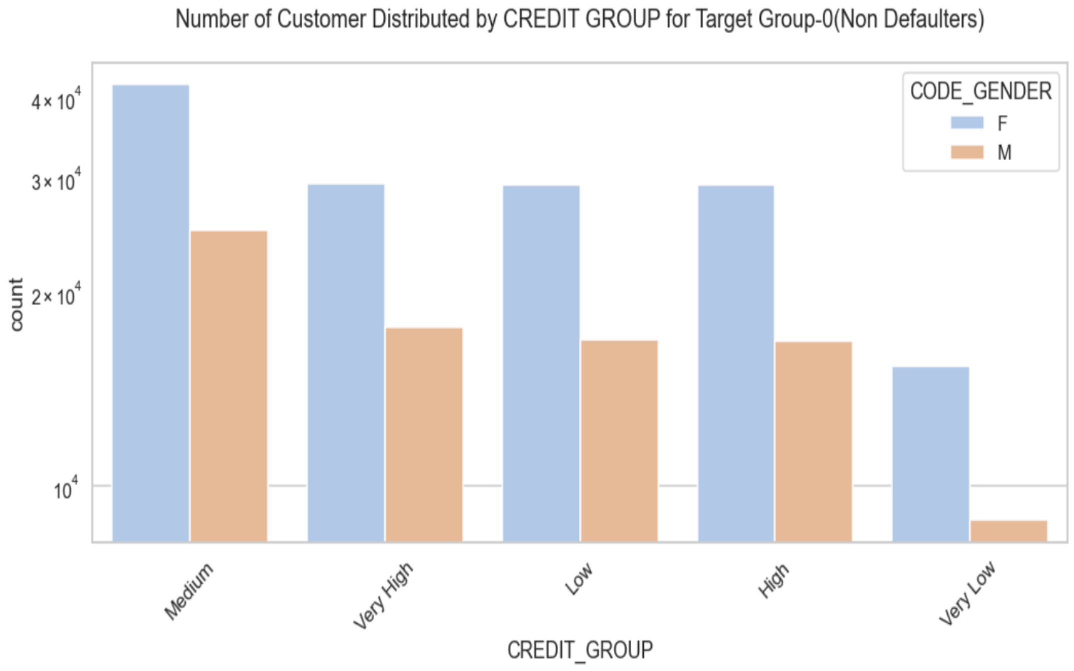
- This is graph shows the distribution **Contract product types**
- Plots of both Defaulter and Non-Defaulters are quite similar
- We can clearly see that most of the defaulters as well as Non-Defaulters take cash-loans
- Hence, we can clearly say that banks should focus more on cash-loans

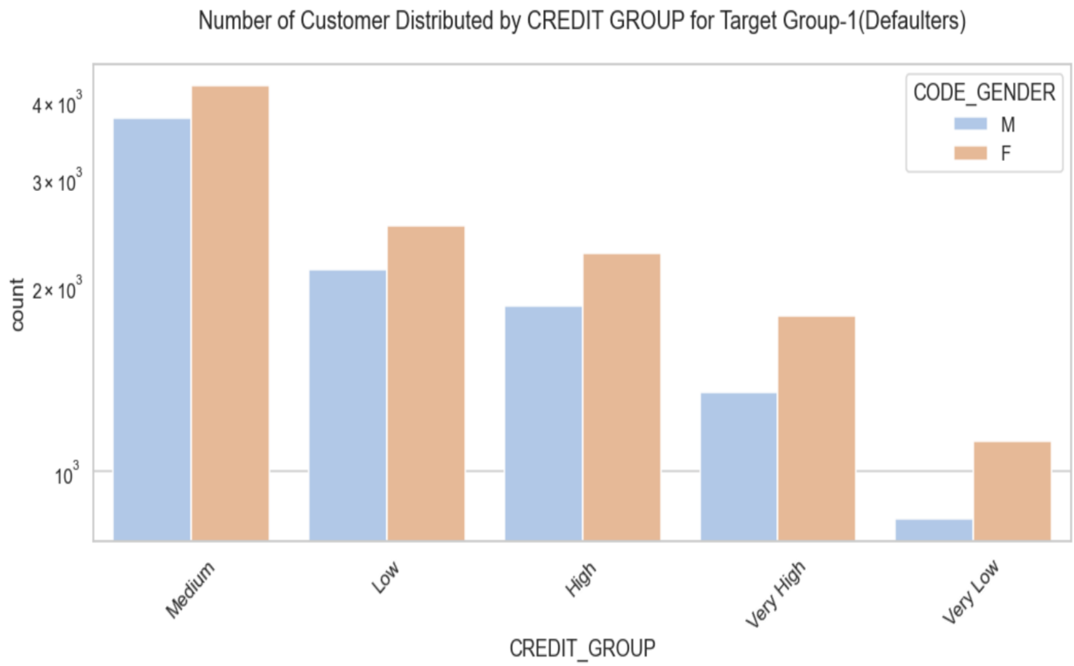




INCOME WITH GENDER DISTRIBUTION

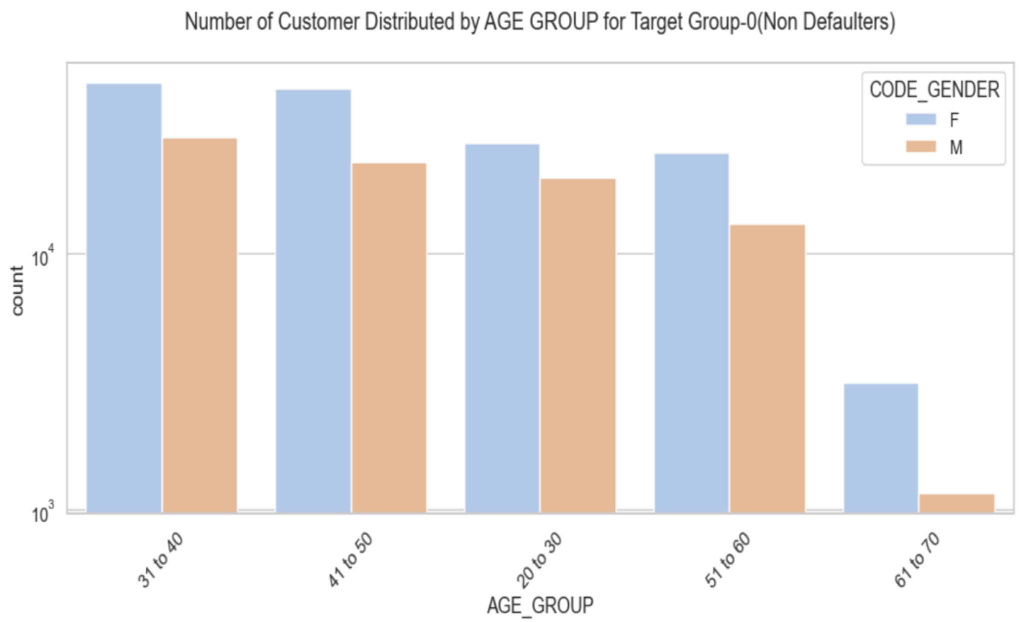
- This is graph shows the **Income types along with Gender distribution.**
- In Defaulters for **Medium** income-group we can see that **Males are equal to Female**
- Except that it is evident that **Females** are **more loan getters** as well as **more defaulters**
- We can also Infer that **most** of the **Application for Loan** ares from **Medium Income-Group**

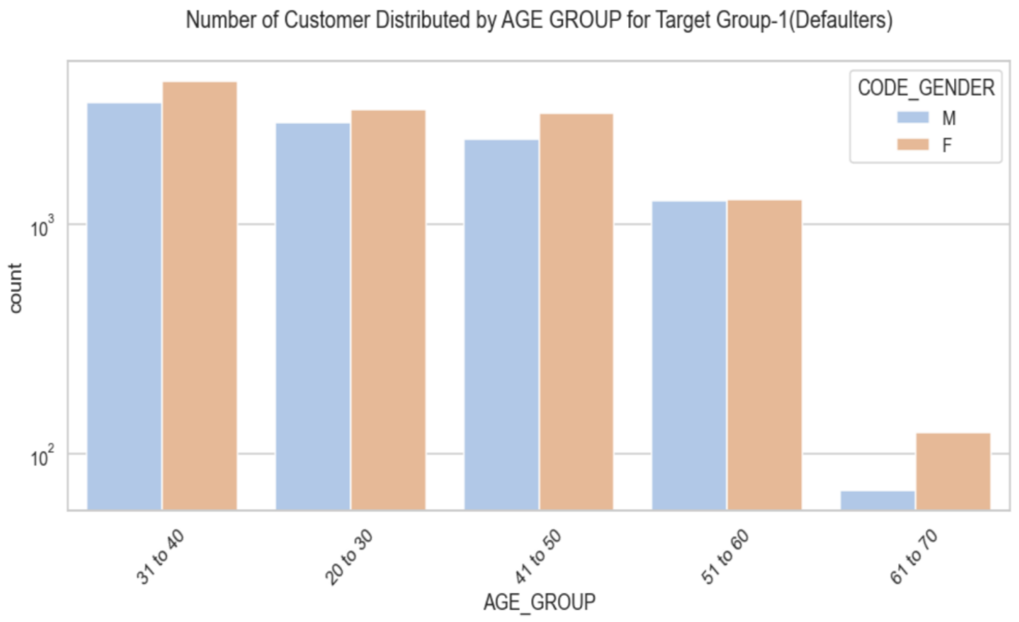




LOAN CREDIT WITH GENDER DISTRIBUTION

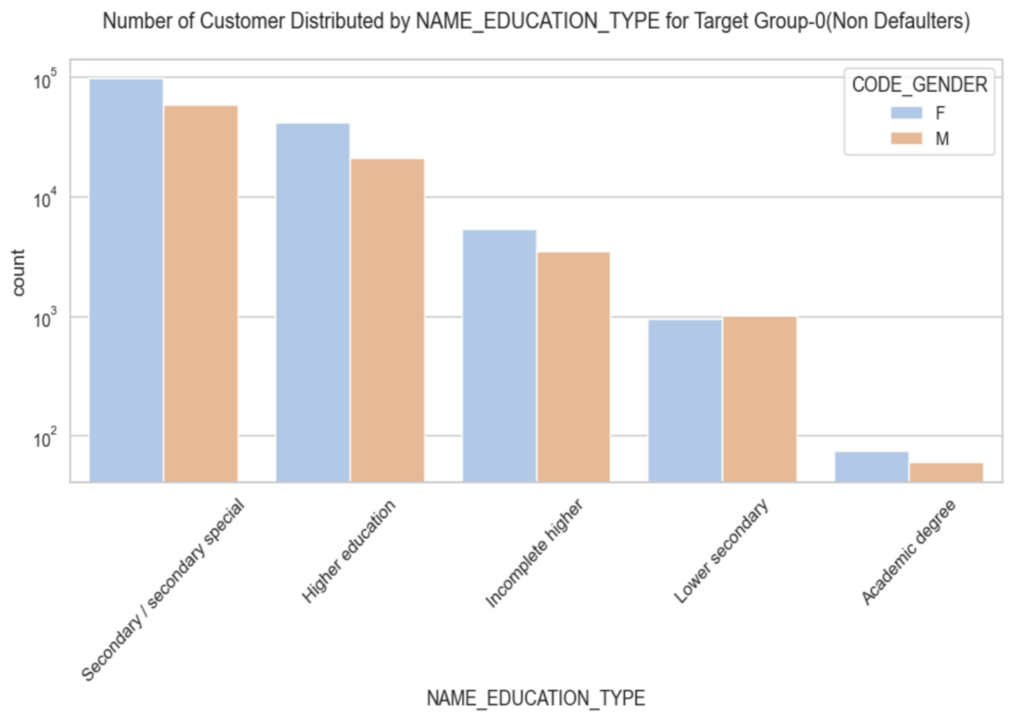
- This is graph shows the **Loan Credit along with Gender distribution.**
- In Defaulters for **Medium** Credit-group we can see that **Males are Almost equal to female.**
- Except that it is evident that **Females** are **more credit getters** as well as **more defaulters**
- We can also Infer that **the greatest** number of applications are for **Medium** Credit group for both Defaulters and Non-Defaulters

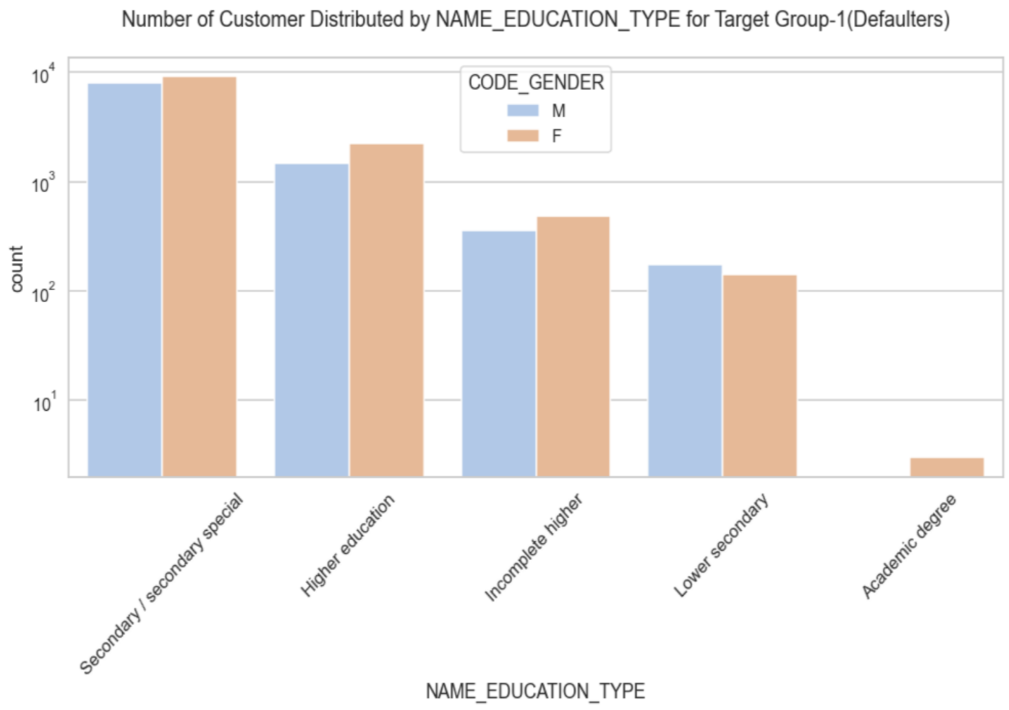




AGE GROUP WITH GENDER DISTRIBUTION

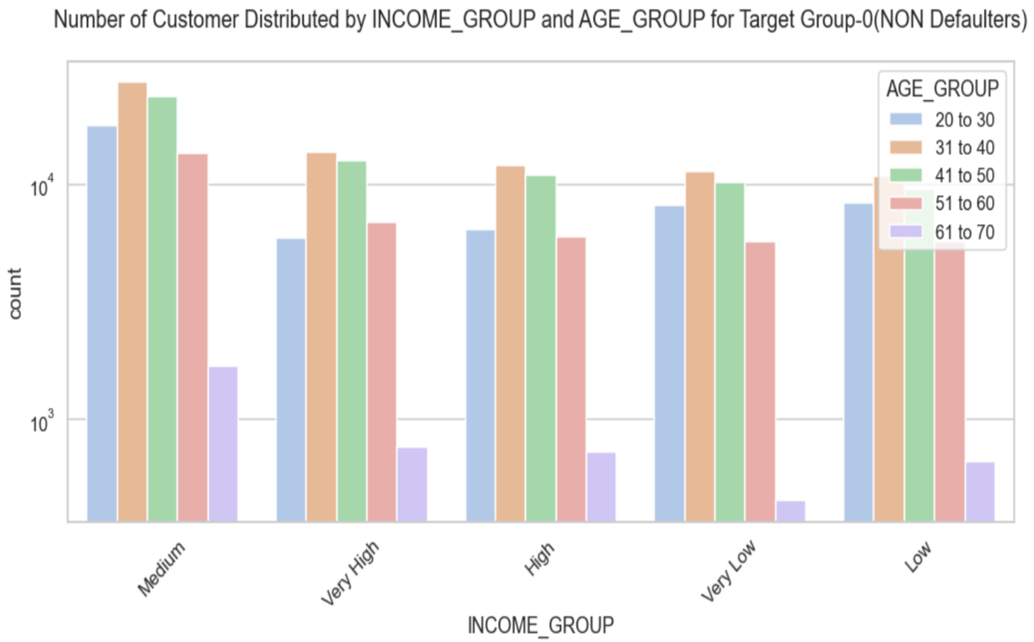
- This is graph shows the **Age Group along with Gender Distribution.**
- Most of the applications are from Age group **31-40 age group** for both Defaulters and Non-
- **Females** are more in every Age-Group for both defaulter and Non-Defaulters
- **Except for** Age-Group **51-60 in Defaulters** here number of application for **Males and Females** are almost **equal.**





EDUCATION TYPE WITH GENDER DISTRIBUTION

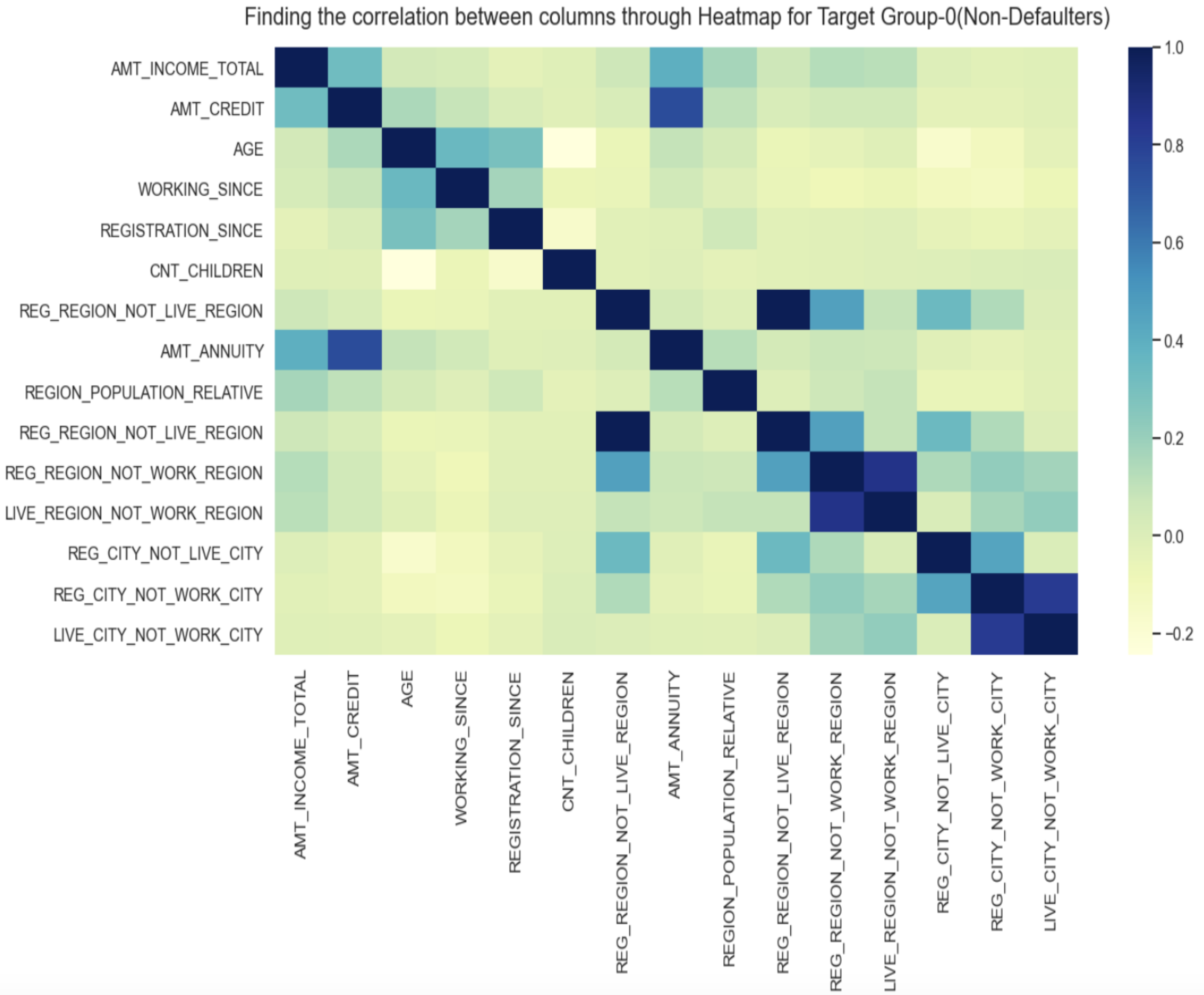
- This is graph shows the **Education Type along with Gender Distribution.**
- Most of the applications are from Education type – **Secondary/Secondary Special group** for both Defaulters and Non-Defaulters
- **Females** are more in every Education type for both defaulter and Non-Defaulters
- **Except for** Education type **Lower Secondary** in Non-Defaulters**.** here number of application for **Males and Females** are almost **equal.**



INCOME GROUP & AGE GROUP DISTRIBUTION

- This is graph shows the **Income group and Age Group Distribution.**
- Most of the **Non-Defaulter** customers are from **Medium** salary with AGE_GROUP as **31-40 and 41-50**
- Most of the **Defaulter** customers are from **Medium** salary with AGE_GROUP as **31-40 and 20-30**
- Hence Giving loan to Age-Group **20-to 30** of **Medium Income** is **RISK**

CORRELATIONS



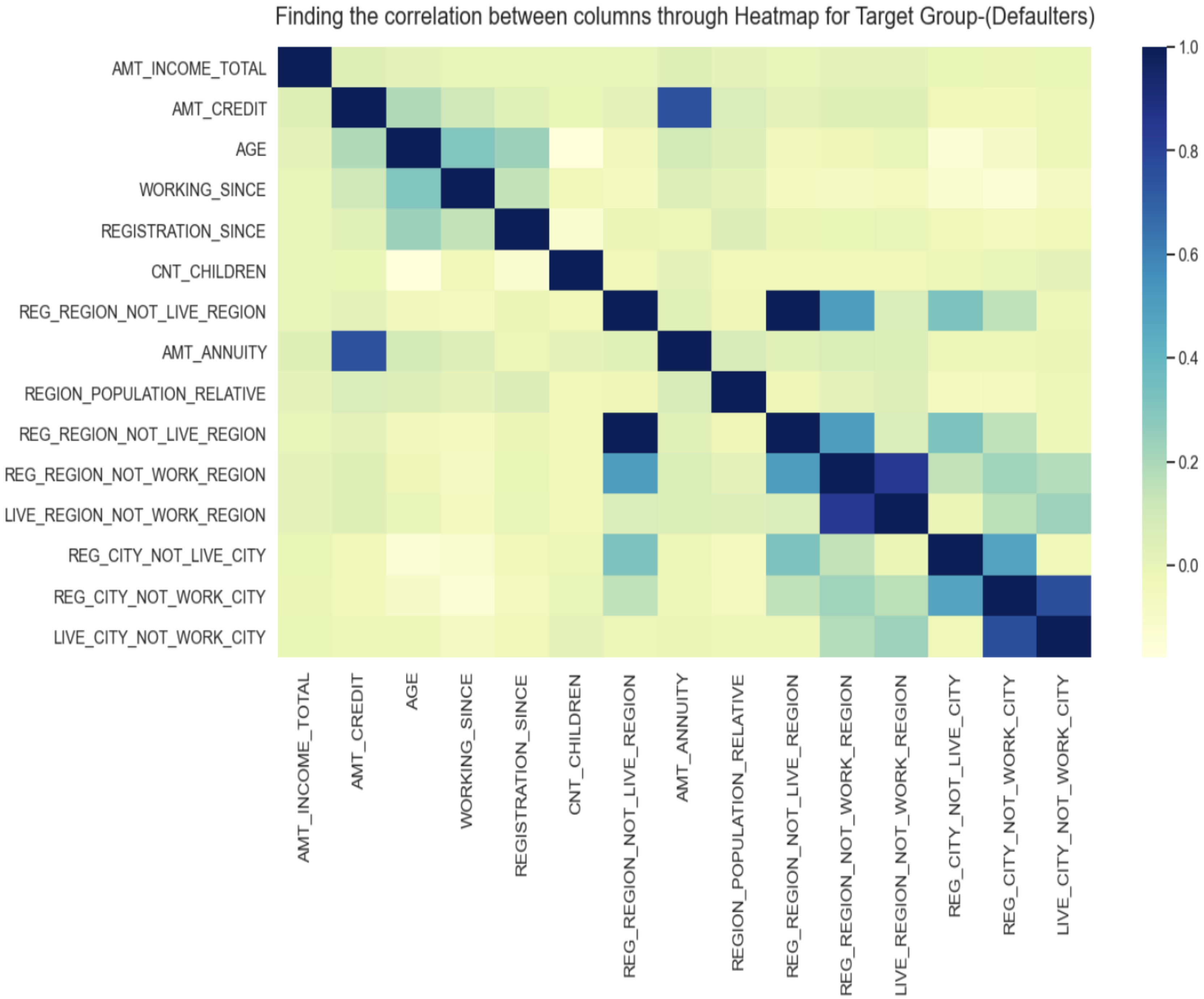
CORRELATIONS BETWEEN VARIABLES FOR NON-DEFAULTERS

From the Heatmap above, we can find the following correlations between columns for the **NON-DEFAULTERS:**

* Between the column **LIVE_REGION_NOT_WORK_REGION** and **REG_REGION_NOT_WORK_REGION**, there are positively correlated.

* Followed with **REG_CITY_NOT_WORK_CITY** and **LIVE_CITY_NOT_WORK_CITY, AMT_ANNUITY** and **AMT_CREDIT.**

* Negatively correlated columns such as **CNT_CHILDREN** and **AGE.**



CORRELATIONS BETWEEN VARIABLES FOR DEFAULTERS

From the Heatmap above, we can find the following correlations between columns for DEFAULTERS:

* Based from the column **LIVE_REGION_NOT_WORK_REGION** and **REG_REGION_NOT_WORK_REGION**, we can conclude that both of the columns are positively correlated.

* This positively correlated column followed with and LIVE_CITY_NOT_WORK_**REG_CITY_NOT_WORK_CITY CITY, AMT_ANNUITY** and **AMT_CREDIT.**

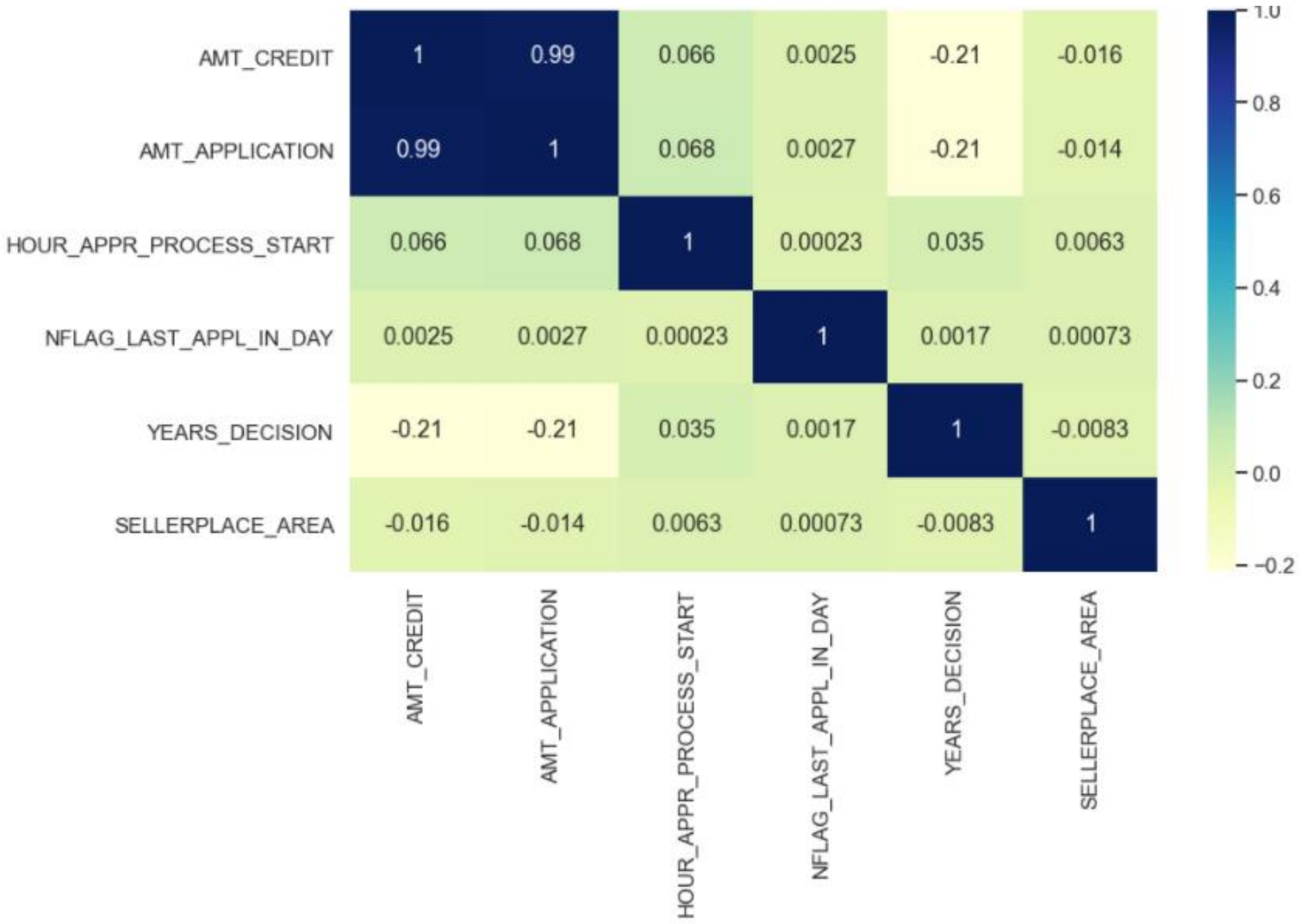
* There are also columns which are negatively correlated such as **AGE and CNT_CHILDREN, REG_CITY_NOT_LIVE_CITY** and **AGE** so on.

CONCLUSION AND SUMMARY FOR APPLICATION DATASET

- Proportion of Defaulter to Non-Defaulters is 8.7%
- Females are More Loan Getter with Greater risk of being loan Defaulters
- Females get higher amount of credit then males
- People who are single, civil married, separated, widow are **less chances of defaulting the loan**
- Giving loan to Age-Group **20-to 30** of **Medium Income** is **RISK**
- Less defaults when applicants have Longer employment and Longer Registration days
- Banks should focus more on cash-loans as they are more revenue generating along with more defaulters are in the Cash-loans

MERGING THE DATASET

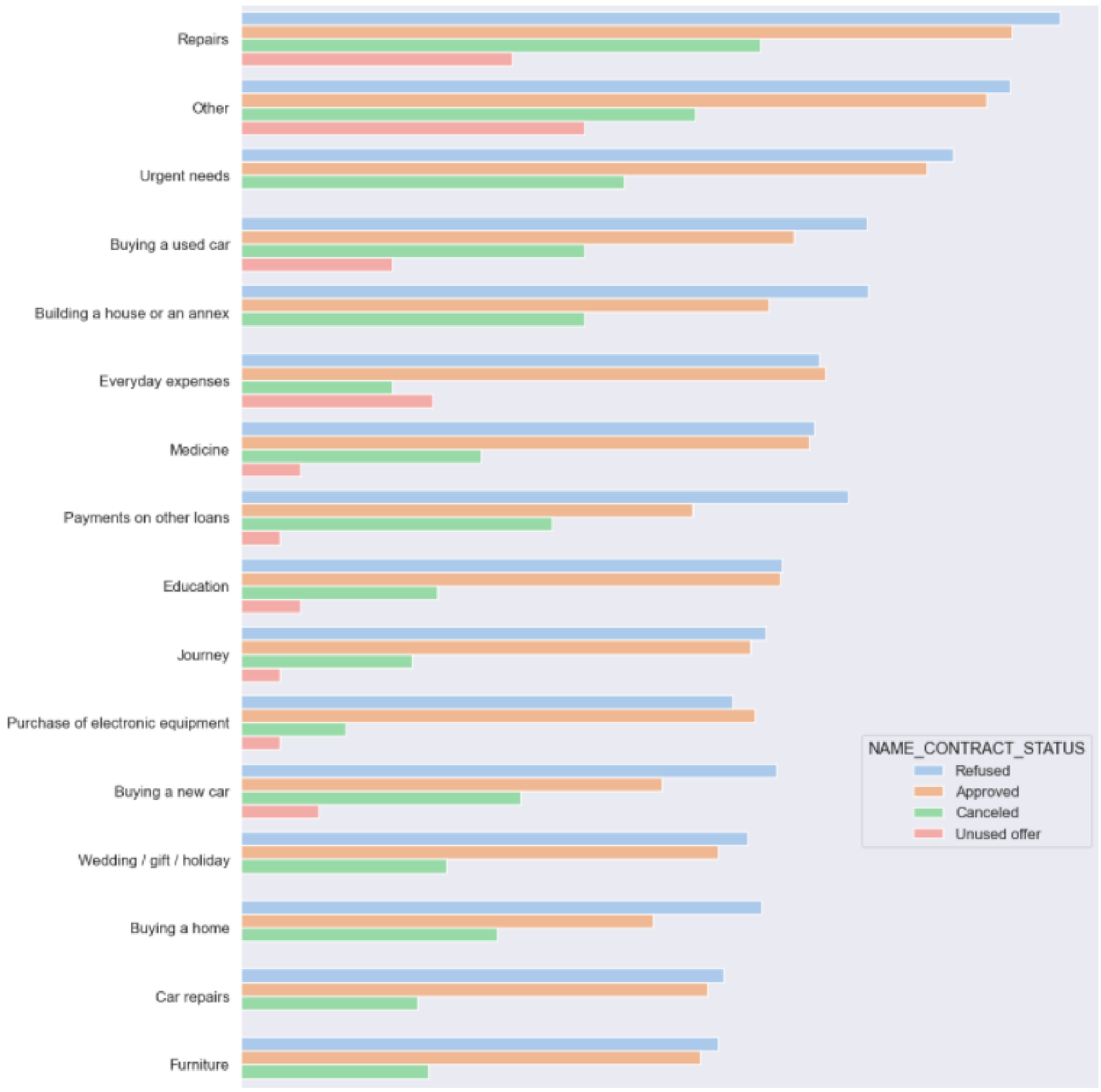
CORRELATIONS OF THE VARIABLES IN THE DATASET

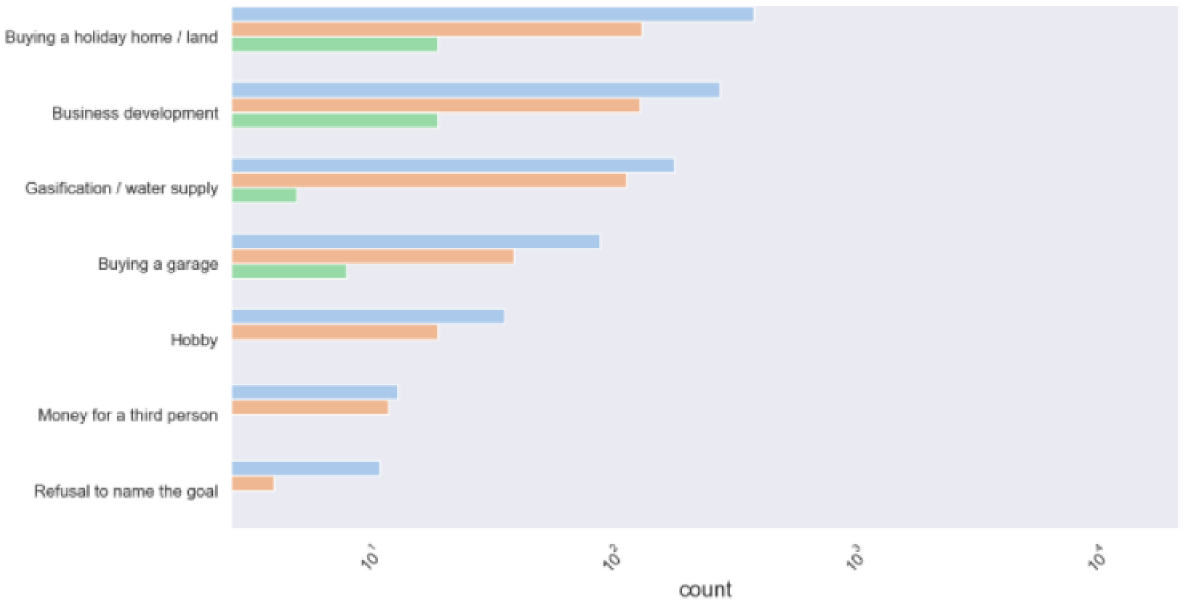


Inferences:

- The correlations that we can conclude from the heatmap will be as follows;
  - AMT_APPLICATION and AMT_CREDIT
    - Positively Correlated – **99%**
  - HOUR_APPR_PROCESS_START and AMT_CREDIT
    - Positively Correlated -- **6.6%**
  - HOURS_APPR_PROCES_START and DECISION
    - Positively Correlated – **3.5%**
  - SELLERPLACE_AREA and AMT_CREDIT
    - Negatively Correlated – **-1.6%**
  - SELLERPLACE_AREA and AMT_APPLICATION
    - Negatively Correlated – **-1.4%**

DISTRIBUTION OF CONTRACT STATUS WITH THE LOAN PURPOSE

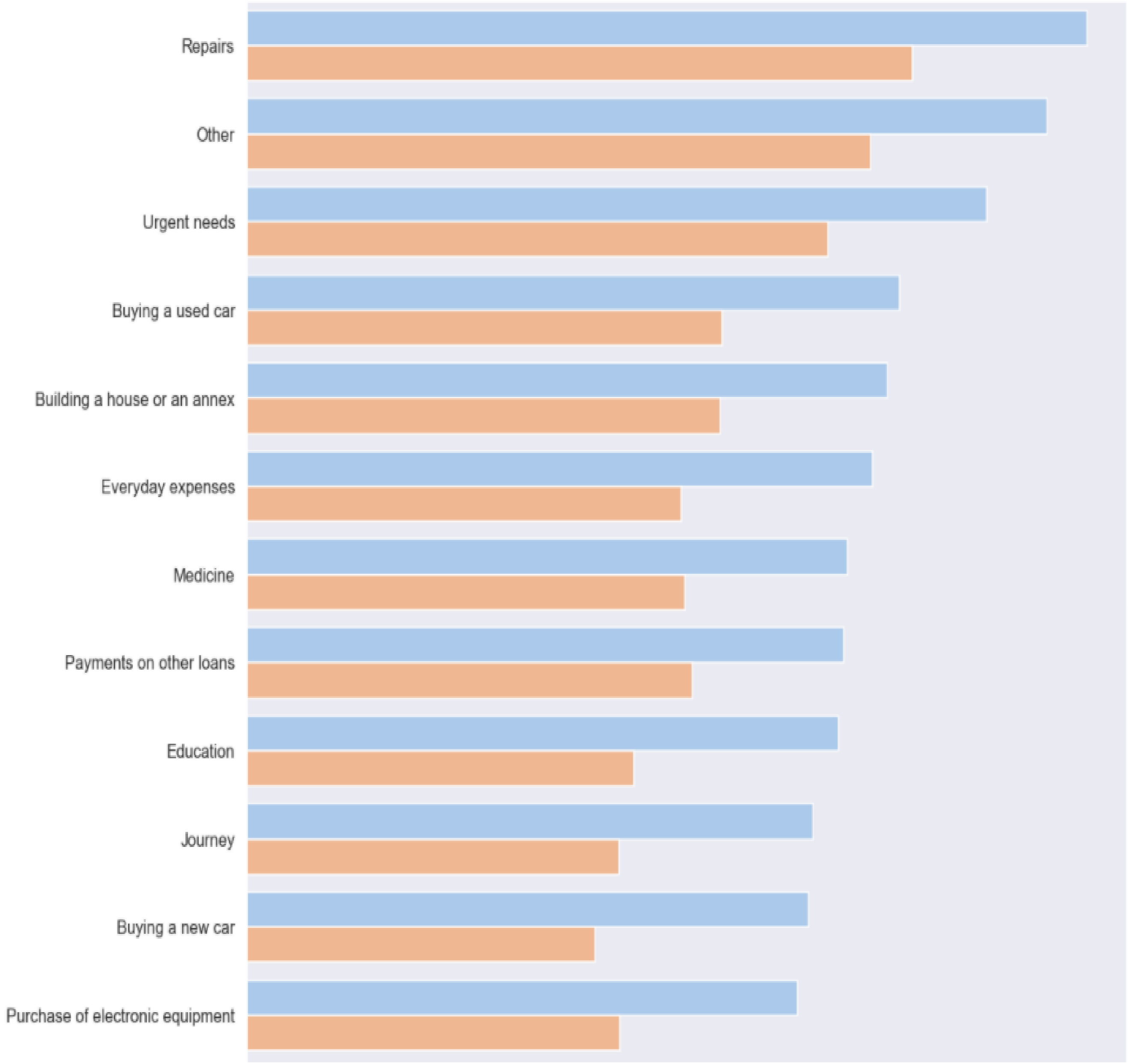


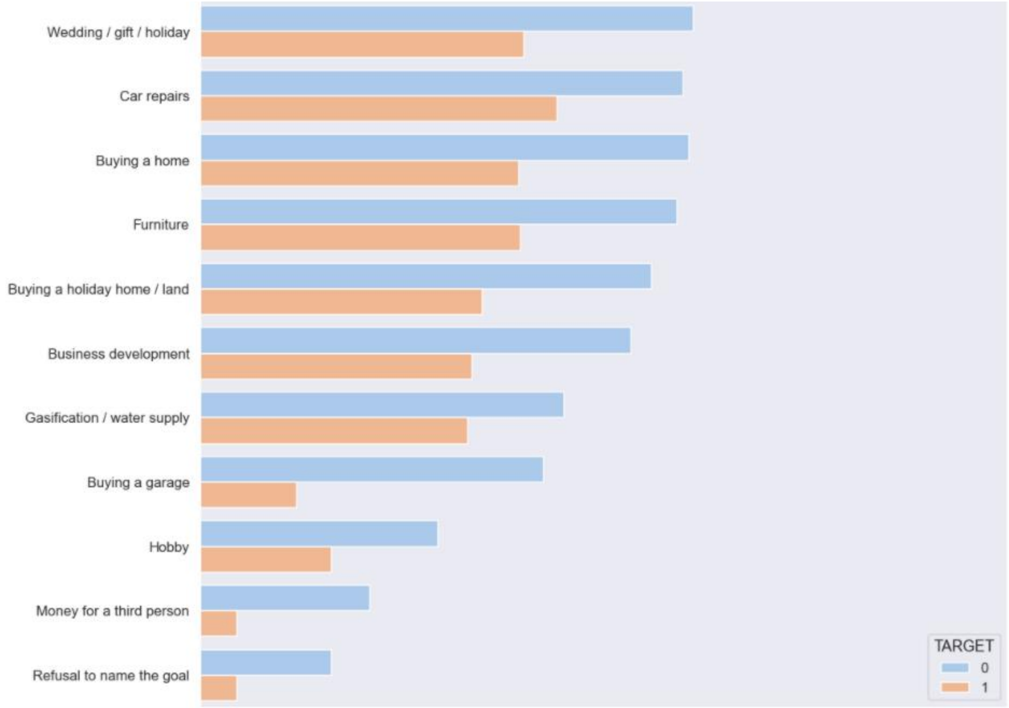


Inferences:

- Application of Loans for the purposes of 'Repairs' carried most Rejections
- Significant Rejections outcome can be seen for 'Other' purposes as well, followed closely by applying loan for the purposes of 'Urgent Need'
- There are also Applications where the applicant refused to name the goal of the application process, this carried the least Approval and Rejection cases
- In a summary, we can see that the number of Rejections exceeded the number of Approved loans.

DISTRIBUTION OF CONTRACT STATUS WITH TARGET

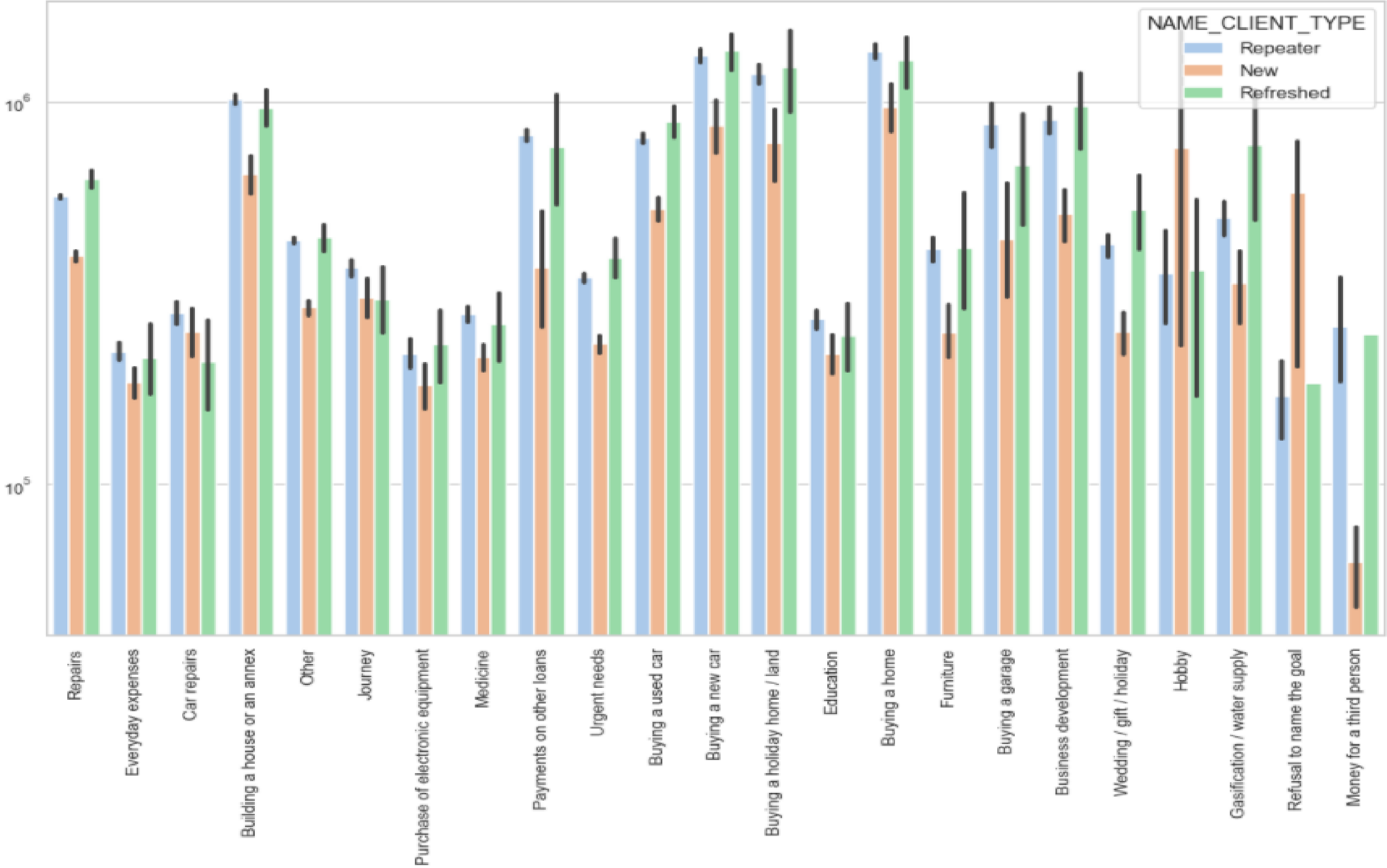




Inferences:

- Application for the purpose of Repairs are facing more difficulties in payment on time.
- We can see significance differences in regards with the loan payment than the difficulties of payments such as for the purpose of 'Buying a land', 'Buying a garage',' Buying a new car' and so on.
- In a summary, we can focus on the purposes which lead to very less of payment difficulties and highly repayment for the bank.

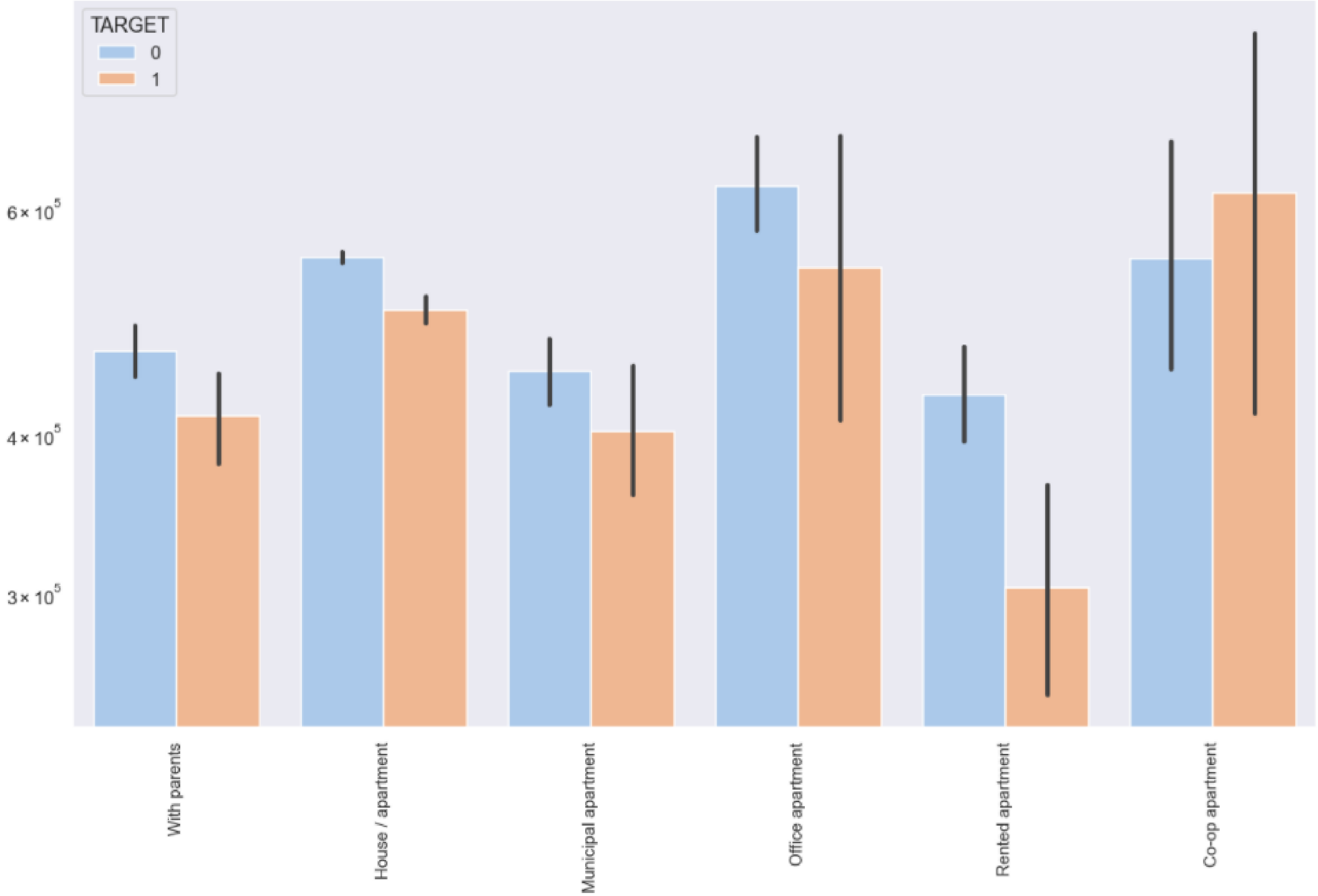
CREDIT AMOUNT VS LOAN PURPOSES



Inferences:

- The Amount Credit for the purpose of Buying a new car, buying a home and buying a land are higher for the repeated and refreshed clients.
- Refreshed clients have a significant amount of credit applied.
- Money for a third person have less credits applied for.

CREDIT AMOUNT VS HOUSING TYPE



Inferences:

- For the Office Apartment are higher Amount of Credit especially for the non-Defaulters as compared with the Defaulters.
- Here we can conclude that the bank should avoid applying loans for Co-op apartment as they have the greater chances of not making payment.
- In a nutshell, the **focus for the bank should be for the housing type of House/Apartment, or With parents** as they have greater values of making payments.

CONCLUSION AND SUMMARY

- Bank should approve more loans for the Housing Type of House/Apartment, Office Apartment, or With parents as there are having less payment difficulties.
- Bank can focus more on the 'Working' Females as they applied most of the applications, less focus for the Pensioner, Age range within 61-70.
- Also, Bank should provide more loans to 'Business entity Type 3' and 'Self Employed'.

THANK YOU