Credit EDA Assignment

By,
Amal A A

Introduction

This assignment aims to give you a sense of how to use EDA in a genuine business setting. In addition to using the skills you learned in the EDA module, you will gain a fundamental grasp of risk analytics in banking and financial services and learn how data is used in this assignment to reduce the risk of losing money when lending to customers.

Business Understanding-1

- Due to their weak or nonexistent credit histories, loan providers find it challenging to grant loans to individuals. Because of this, some customers take advantage of it by defaulting. Imagine you work for a consumer finance business that provides urban customers with different kinds of loans. To analyse the trends found in the data, you must use EDA. By doing this, it will be ensured that only those applicants who can repay the debt will be accepted.
- When a loan application is received, the company must determine whether to approve the loan based on the applicant's profile. The bank's choice is subject to two different kinds of risks:
    - If the borrower is likely to repay the loan, refusing to grant it results in the firm losing business.
    - If the borrower is not expected to pay back the loan or is expected to default, then authorising the loan may result in a loss of revenue for the business.

Business Understanding -2

- Four decisions could be made by a client or company in response to a loan application :
  - **Approved:** The loan application has been accepted by the company
  - **Cancelled:** During the approval process, the client cancelled the registration. Either the client changed his/her mind about the loan, or in some instances because the client was a higher risk, he/she received unfavorable pricing.
  - **Refused:** The loan has been rejected by the company(because the client does not meet their requirements etc.).
  - **Unused offer:** The client has cancelled the loan, but the procedure is still in progress.
- You will use this case study to apply EDA to your understanding of how loan and customer characteristics affect default risk.

Data Understanding

Three files in the given dataset are:

1. All of the client's information from the moment of application is contained in the file "application_data.csv". The information relates to a client's ability to make payments.

2. 'The client's prior credit data is contained in the file "previous_application.csv". It includes information about whether the previous application was accepted, rejected, canceled, or not used.

3. A data dictionary named *'columns_description.csv'* describes the meaning of the variables.

Data Cleaning Approach

- Removed columns with missing values of more than 50% from both datasets. And we have also dropped the columns that seemed irrelevant to the future analysis of the data.

- In the other numerical columns with less than 50% missing values, we imputed the null values with the mean or median.

- In the categorical columns, we have replaced the null values with the highest occurring category in most cases but in the case of categorical variables like NAME_PRODUCT_TYPE, NAME_GOODS_CATEGORY etc, we observed a very large count of missing values so in such cases we left them as it is.

Missing Data

Application Dataset

DAYS_LAST_PHONE_CHANGE	0.00
CNT_CHILDREN	0.00
FLAG_DOCUMENT_8	0.00
NAME_CONTRACT_TYPE	0.00
CODE_GENDER	0.00
FLAG_OWN_CAR	0.00
FLAG_DOCUMENT_2	0.00
FLAG_DOCUMENT_3	0.00
FLAG_DOCUMENT_4	0.00
FLAG_DOCUMENT_5	0.00
FLAG_DOCUMENT_6	0.00
FLAG_DOCUMENT_7	0.00
FLAG_DOCUMENT_9	0.00
FLAG_DOCUMENT_21	0.00
FLAG_DOCUMENT_10	0.00
FLAG_DOCUMENT_11	0.00
FLAG_OWN_REALTY	0.00
FLAG_DOCUMENT_13	0.00
FLAG_DOCUMENT_14	0.00
FLAG_DOCUMENT_15	0.00
FLAG_DOCUMENT_16	0.00
FLAG_DOCUMENT_17	0.00
FLAG_DOCUMENT_18	0.00
FLAG_DOCUMENT_19	0.00
FLAG_DOCUMENT_20	0.00
FLAG_DOCUMENT_12	0.00
AMT_CREDIT	0.00
AMT_INCOME_TOTAL	0.00
FLAG_PHONE	0.00
LIVE_CITY_NOT_WORK_CITY	0.00
REG_CITY_NOT_WORK_CITY	0.00
TARGET	0.00
REG_CITY_NOT_LIVE_CITY	0.00
LIVE_REGION_NOT_WORK_REGION	0.00
REG_REGION_NOT_WORK_REGION	0.00
REG_REGION_NOT_LIVE_REGION	0.00
HOUR_APPR_PROCESS_START	0.00
WEEKDAY_APPR_PROCESS_START	0.00
REGION_RATING_CLIENT_W_CITY	0.00
REGION_RATING_CLIENT	0.00
FLAG_EMAIL	0.00
FLAG_CONT_MOBILE	0.00
ORGANIZATION_TYPE	0.00
FLAG_WORK_PHONE	0.00
FLAG_EMP_PHONE	0.00
FLAG_MOBIL	0.00
DAYS_ID_PUBLISH	0.00
DAYS_REGISTRATION	0.00
DAYS_EMPLOYED	0.00
DAYS_BIRTH	0.00
REGION_POPULATION_RELATIVE	0.00
NAME_HOUSING_TYPE	0.00
NAME_FAMILY_STATUS	0.00
NAME_EDUCATION_TYPE	0.00
NAME_INCOME_TYPE	0.00
SK_ID_CURR	0.00

COMMONAREA_MEDI	69.87
COMMONAREA_AVG	69.87
COMMONAREA_MODE	69.87
NONLIVINGAPARTMENTS_MODE	69.43
NONLIVINGAPARTMENTS_AVG	69.43
NONLIVINGAPARTMENTS_MEDI	69.43
FONDKAPREMONT_MODE	68.39
LIVINGAPARTMENTS_MODE	68.35
LIVINGAPARTMENTS_AVG	68.35
LIVINGAPARTMENTS_MEDI	68.35
FLOORSMIN_AVG	67.85
FLOORSMIN_MODE	67.85
FLOORSMIN_MEDI	67.85
YEARS_BUILD_MEDI	66.50
YEARS_BUILD_MODE	66.50
YEARS_BUILD_AVG	66.50
OWN_CAR_AGE	65.99
LANDAREA_MEDI	59.38
LANDAREA_MODE	59.38
LANDAREA_AVG	59.38
BASEMENTAREA_MEDI	58.52
BASEMENTAREA_AVG	58.52
BASEMENTAREA_MODE	58.52
EXT_SOURCE_1	56.38
NONLIVINGAREA_MODE	55.18
NONLIVINGAREA_AVG	55.18
NONLIVINGAREA_MEDI	55.18
ELEVATORS_MEDI	53.30
ELEVATORS_AVG	53.30
ELEVATORS_MODE	53.30
WALLSMATERIAL_MODE	50.84
APARTMENTS_MEDI	50.75
APARTMENTS_AVG	50.75
APARTMENTS_MODE	50.75
ENTRANCES_MEDI	50.35
ENTRANCES_AVG	50.35
ENTRANCES_MODE	50.35
LIVINGAREA_AVG	50.19
LIVINGAREA_MODE	50.19
LIVINGAREA_MEDI	50.19
HOUSETYPE_MODE	50.18
FLOORSMAX_MODE	49.76
FLOORSMAX_MEDI	49.76
FLOORSMAX_AVG	49.76
YEARS_BEGINEXPLUATATION_MODE	48.78
YEARS_BEGINEXPLUATATION_MEDI	48.78
YEARS_BEGINEXPLUATATION_AVG	48.78
TOTALAREA_MODE	48.27
EMERGENCYSTATE_MODE	47.40
OCCUPATION_TYPE	31.35
EXT_SOURCE_3	19.83
AMT_REQ_CREDIT_BUREAU_HOUR	13.50
AMT_REQ_CREDIT_BUREAU_DAY	13.50
AMT_REQ_CREDIT_BUREAU_WEEK	13.50
AMT_REQ_CREDIT_BUREAU_MON	13.50
AMT_REQ_CREDIT_BUREAU_QRT	13.50
AMT_REQ_CREDIT_BUREAU_YEAR	13.50
NAME_TYPE_SUITE	0.42
OBS_30_CNT_SOCIAL_CIRCLE	0.33
DEF_30_CNT_SOCIAL_CIRCLE	0.33
OBS_60_CNT_SOCIAL_CIRCLE	0.33
DEF_60_CNT_SOCIAL_CIRCLE	0.33
EXT_SOURCE_2	0.21
AMT_GOODS_PRICE	0.09
AMT_ANNUITY	0.00
CNT_FAM_MEMBERS	0.00

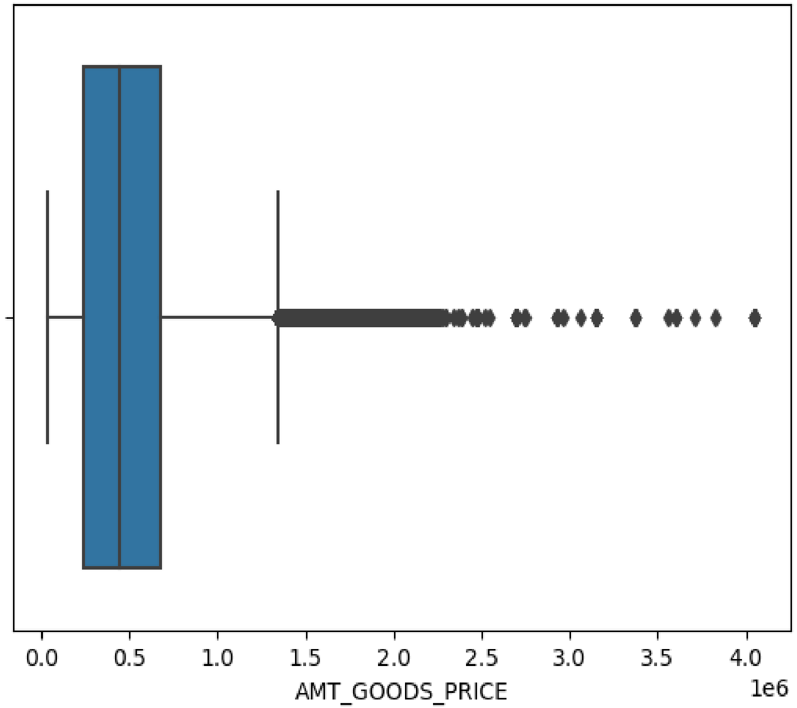
Previous Application Dataset

COMMONAREA_MEDI	69.87
COMMONAREA_AVG	69.87
COMMONAREA_MODE	69.87
NONLIVINGAPARTMENTS_MODE	69.43
NONLIVINGAPARTMENTS_AVG	69.43
NONLIVINGAPARTMENTS_MEDI	69.43
FONDKAPREMONT_MODE	68.39
LIVINGAPARTMENTS_MODE	68.35
LIVINGAPARTMENTS_AVG	68.35
LIVINGAPARTMENTS_MEDI	68.35
FLOORSMIN_AVG	67.85
FLOORSMIN_MODE	67.85
FLOORSMIN_MEDI	67.85
YEARS_BUILD_MEDI	66.50
YEARS_BUILD_MODE	66.50
YEARS_BUILD_AVG	66.50
OWN_CAR_AGE	65.99
LANDAREA_MEDI	59.38
LANDAREA_MODE	59.38
LANDAREA_AVG	59.38
BASEMENTAREA_MEDI	58.52
BASEMENTAREA_AVG	58.52
BASEMENTAREA_MODE	58.52
EXT_SOURCE_1	56.38
NONLIVINGAREA_MODE	55.18
NONLIVINGAREA_AVG	55.18
NONLIVINGAREA_MEDI	55.18
ELEVATORS_MEDI	53.30
ELEVATORS_AVG	53.30
ELEVATORS_MODE	53.30
WALLSMATERIAL_MODE	50.84
APARTMENTS_MEDI	50.75
APARTMENTS_AVG	50.75
APARTMENTS_MODE	50.75
ENTRANCES_MEDI	50.35
ENTRANCES_AVG	50.35
ENTRANCES_MODE	50.35
LIVINGAREA_AVG	50.19
LIVINGAREA_MODE	50.19
LIVINGAREA_MEDI	50.19
HOUSETYPE_MODE	50.18
FLOORSMAX_MODE	49.76
FLOORSMAX_MEDI	49.76
FLOORSMAX_AVG	49.76
YEARS_BEGINEXPLUATATION_MODE	48.78
YEARS_BEGINEXPLUATATION_MEDI	48.78
YEARS_BEGINEXPLUATATION_AVG	48.78
TOTALAREA_MODE	48.27
EMERGENCYSTATE_MODE	47.40
OCCUPATION_TYPE	31.35
EXT_SOURCE_3	19.83
AMT_REQ_CREDIT_BUREAU_HOUR	13.50
AMT_REQ_CREDIT_BUREAU_DAY	13.50
AMT_REQ_CREDIT_BUREAU_WEEK	13.50
AMT_REQ_CREDIT_BUREAU_MON	13.50
AMT_REQ_CREDIT_BUREAU_QRT	13.50
AMT_REQ_CREDIT_BUREAU_YEAR	13.50
NAME_TYPE_SUITE	0.42
OBS_30_CNT_SOCIAL_CIRCLE	0.33
DEF_30_CNT_SOCIAL_CIRCLE	0.33
OBS_60_CNT_SOCIAL_CIRCLE	0.33
DEF_60_CNT_SOCIAL_CIRCLE	0.33
EXT_SOURCE_2	0.21
AMT_GOODS_PRICE	0.09
AMT_ANNUITY	0.00
CNT_FAM_MEMBERS	0.00

Outlier Analysis

Analysis of AMT_GOODS_PRICE

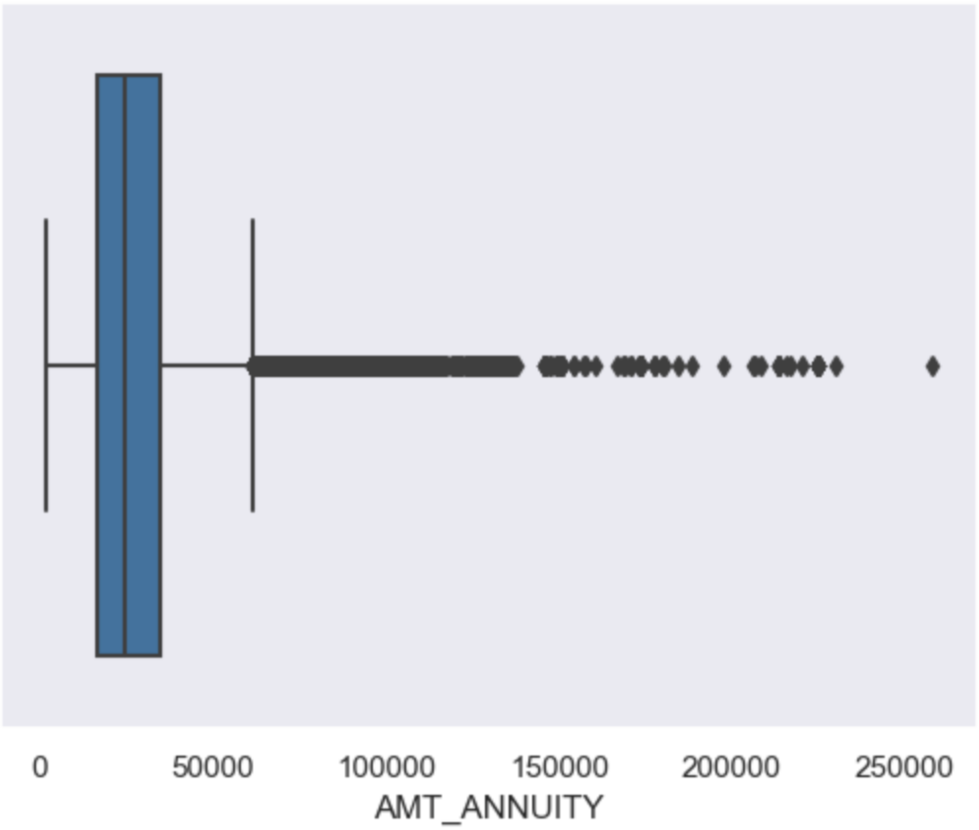
- Values above 900000.0 are outliers but as per the problem statement, we need not make any changes.




```
count     307233.00
mean      538396.21
std       369446.46
min        40500.00
25%       238500.00
50%       450000.00
75%       679500.00
max      4050000.00
Name: AMT_GOODS_PRICE, dtype: fl
```

Analysis of AMT_ANNUITY

- We can see that there are outliers present above 43632 in the column but as per the problem statement, we need not make any changes.

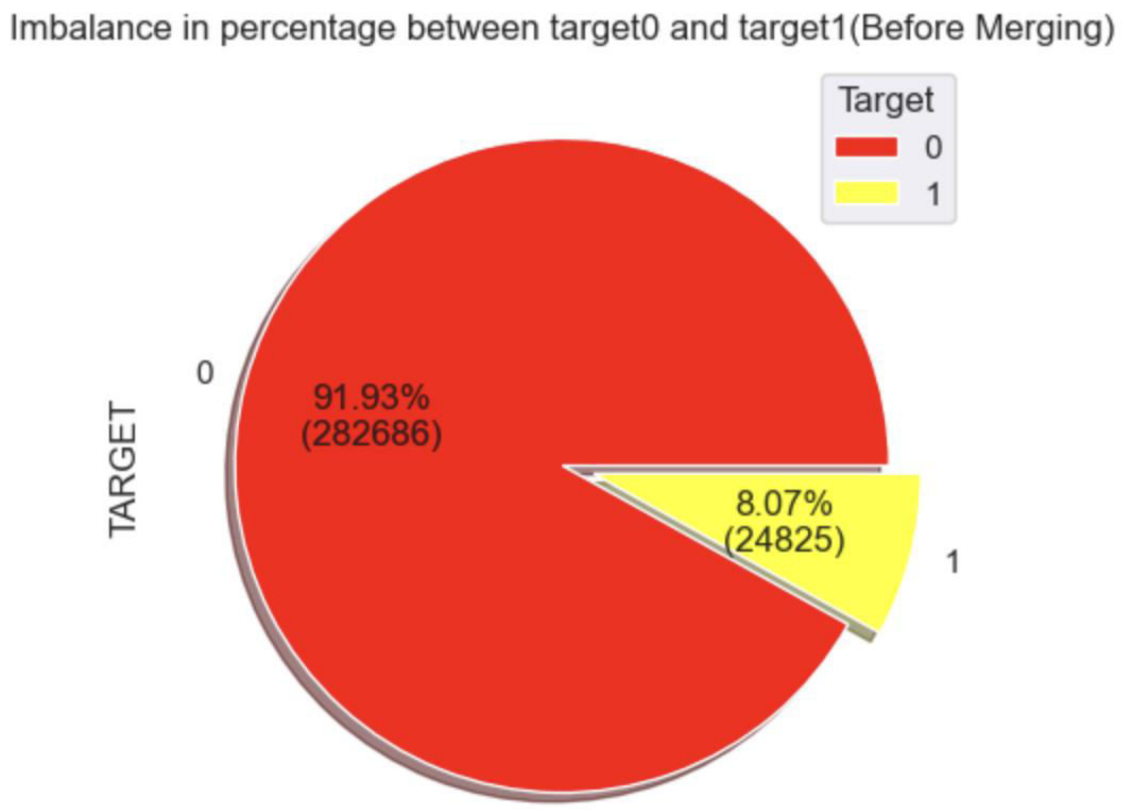



```
count    307499.00
mean      27108.57
std       14493.74
min        1615.50
25%       16524.00
50%       24903.00
75%       34596.00
max      258025.50
Name: AMT_ANNUITY, dtype: float64
```

Methodology-1

- While calculating imbalance ratio using the target variable we found there was an imbalance of 11.38% in the data

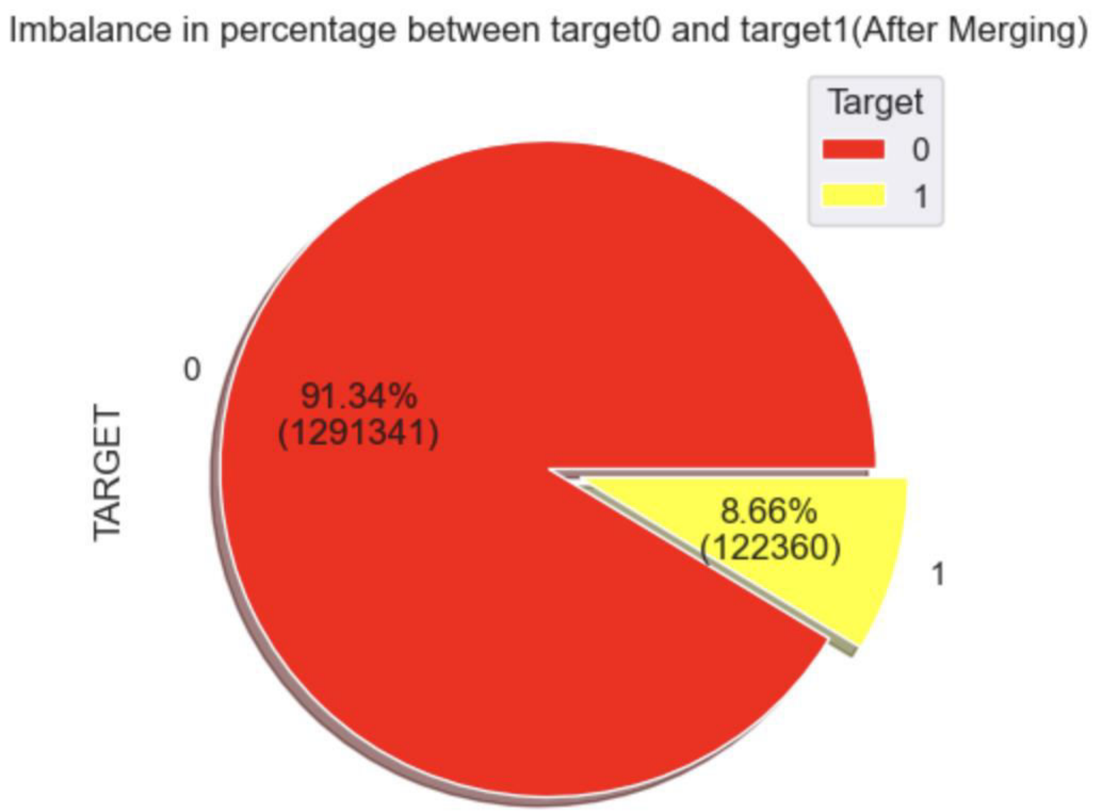
- So during the data analysis, we primarily divided the application data into 2 datasets(target0 and target1).



Methodology-2

- After that, we merged both the previous application and application data into a combined dataset for further bivariate and multivariate analysis.

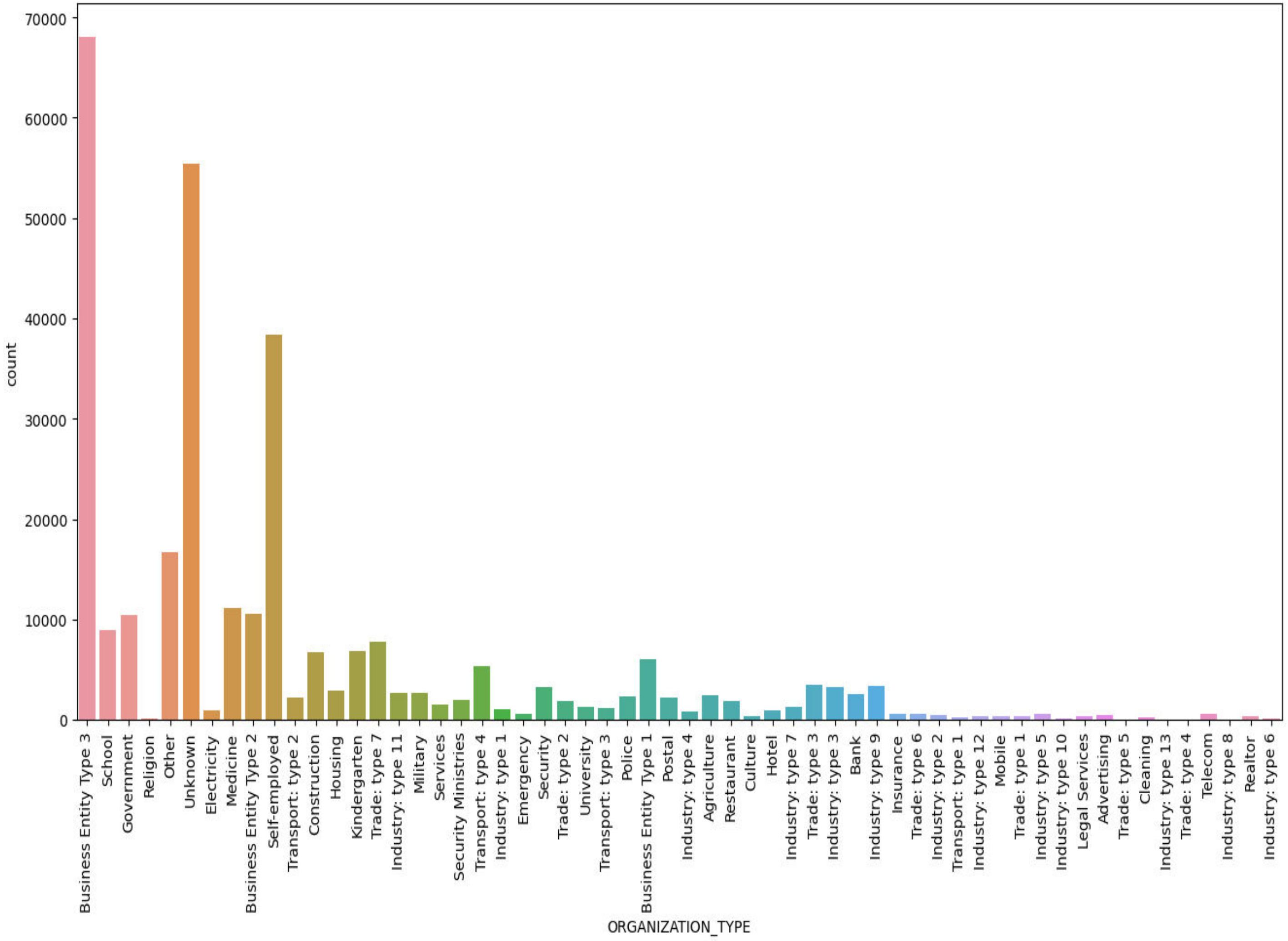
- All the analyses were done using pie plots, count plots and heatmaps.



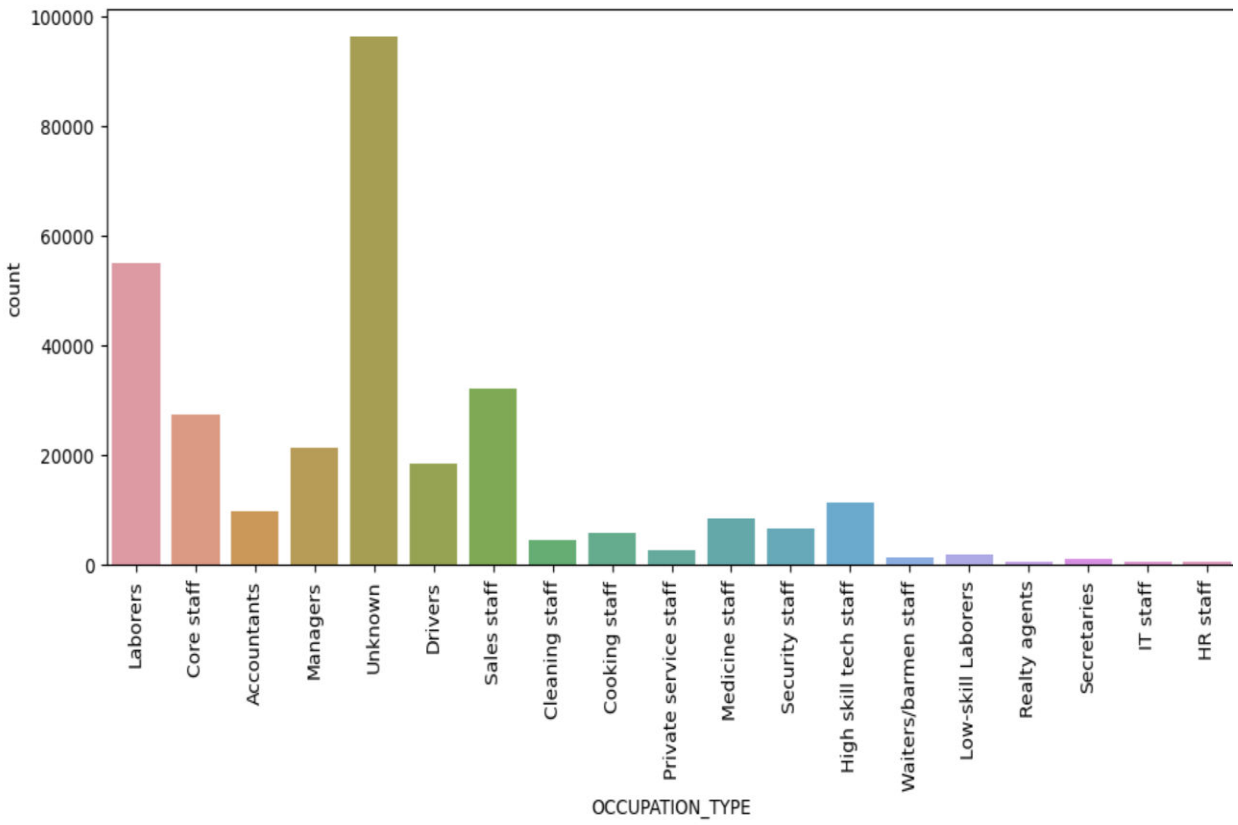
Univariate Analysis

Analysis of ORGANIZATION_TYPE

The majority of clients work in Business Entity Type 3 organizations and least in Industry: type 8.



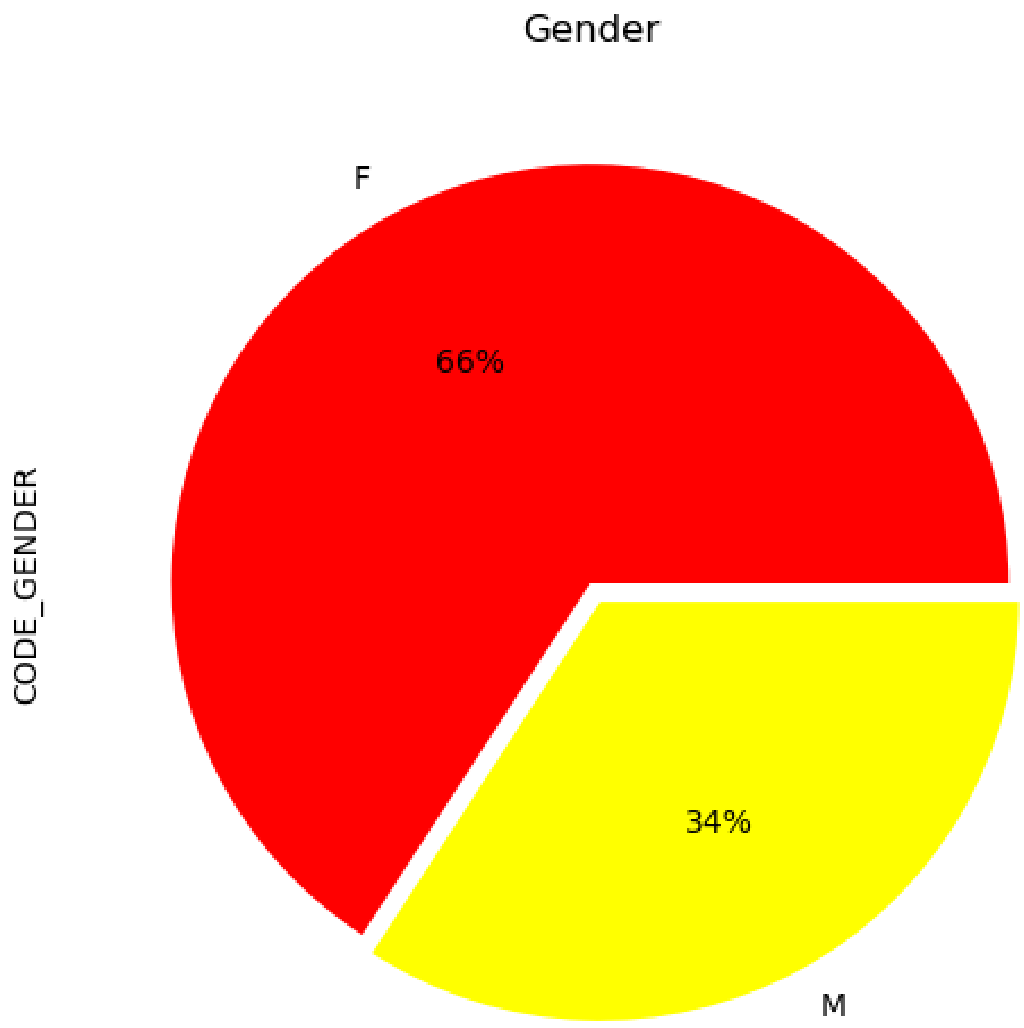
Analysis of OCCUPATION_TYPE



- From the above plot, it's obvious that the majority of clients are Labourers and the minority is from Realty staff.

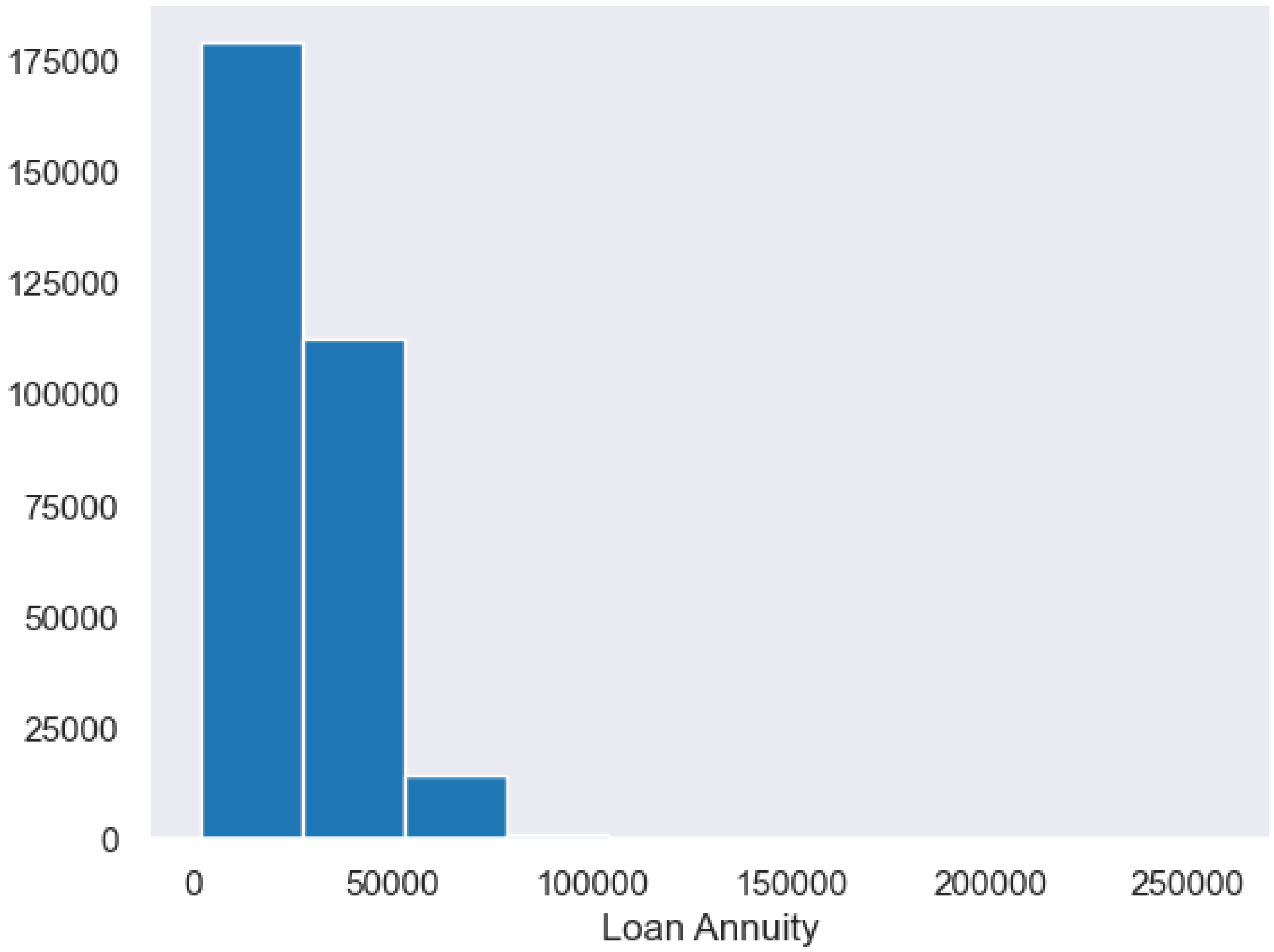
Analysis of CODE_GENDER

CODE_GENDER data consists of 66% of females and 34% of males i.e. majority of the clients are females



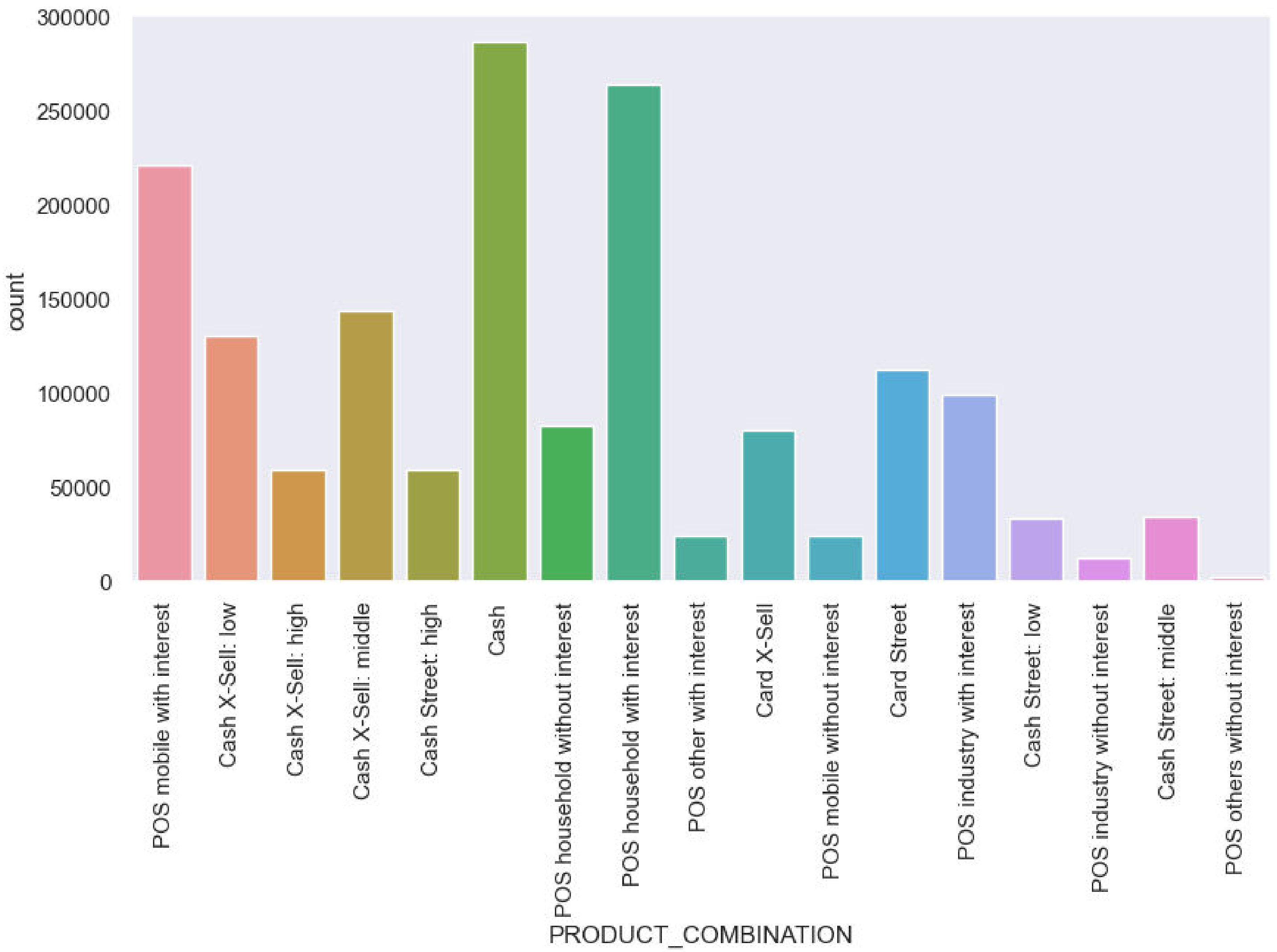
Analysis of AMT_ANNUITY

- Most of the values lie in range of 0 – 50000.



Analysis of PRODUCT_COMBINATION

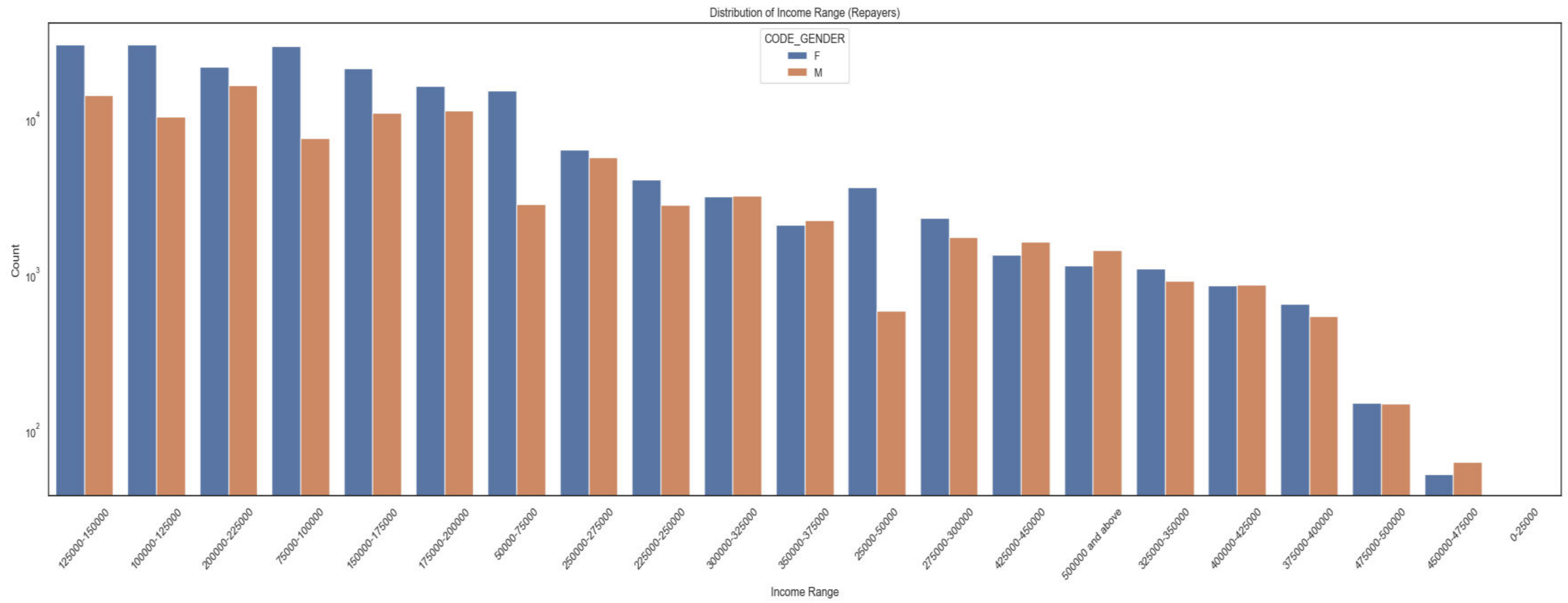
- The category Cash has the highest count.
- POS others without interest have the lowest count.

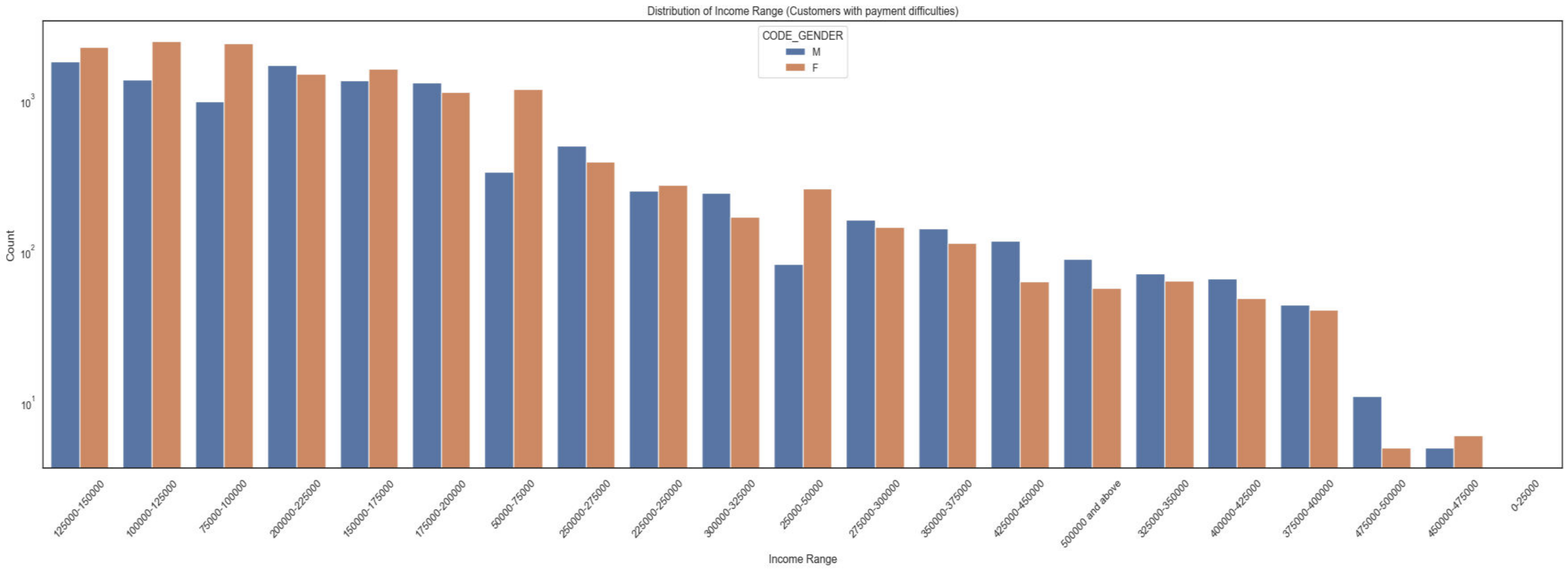


Bivariate/Multivariate Analysis

Analysis of AMT_INCOME_RANGE Vs CODE_GENDER

- Females are better repayers of loan than males.

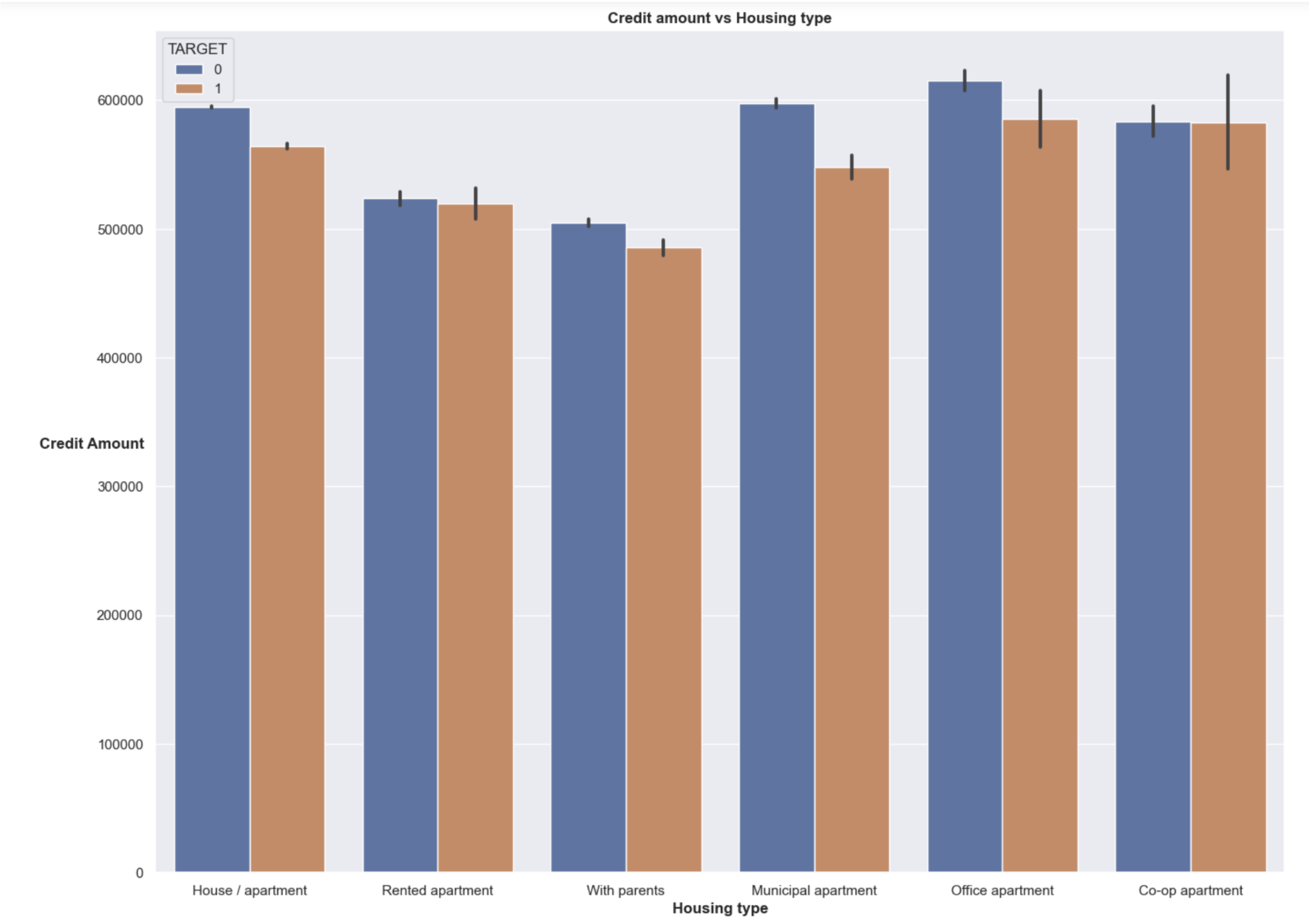




Analysis of NAME_HOUSING_TYPE Vs AMT_CREDIT Vs TARGET

- Clients with office apartment, house/apartment, municipal aparments have the highest repayers.

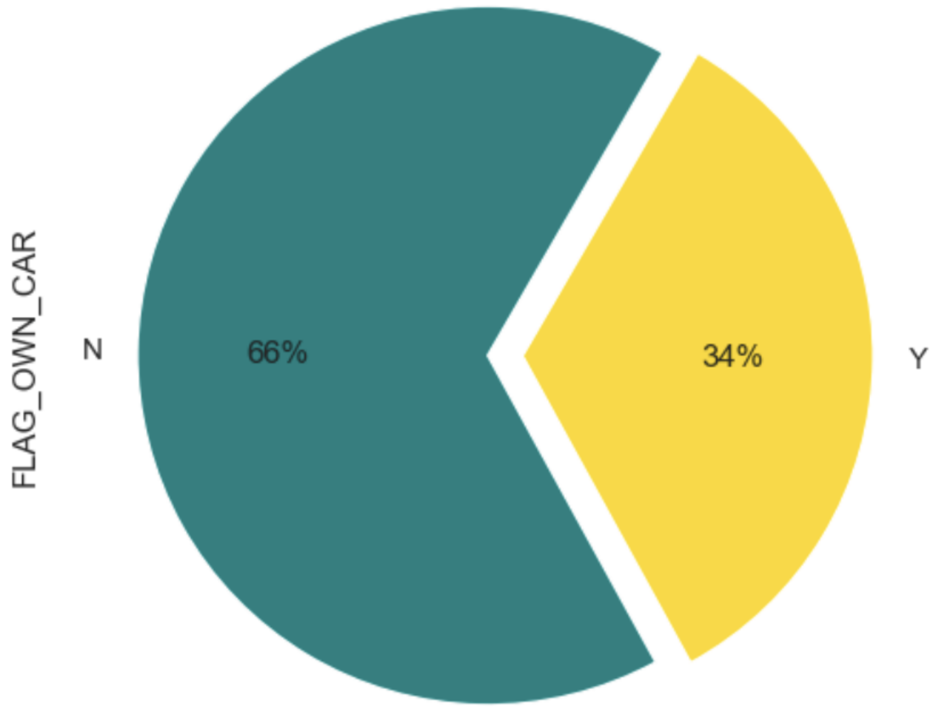
- Clients living with parents or in a parent's aparment have the least amount of repayers and defaulters.



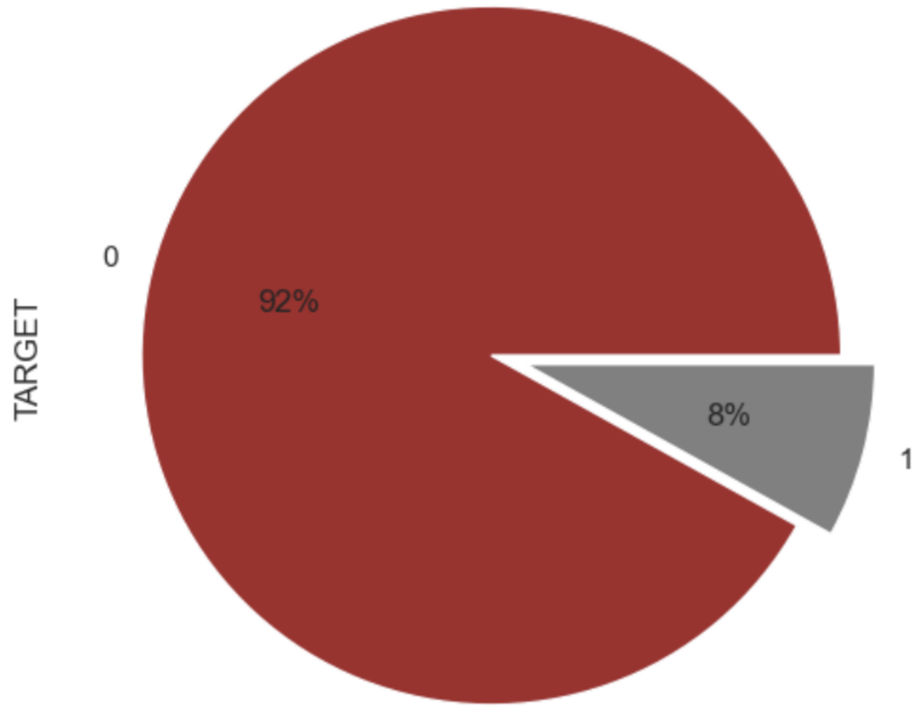
Analysis of FLAG_OWN_CAR Vs Target

- Clients who own a car are better repayers of loans.

Distribution of client by car ownership

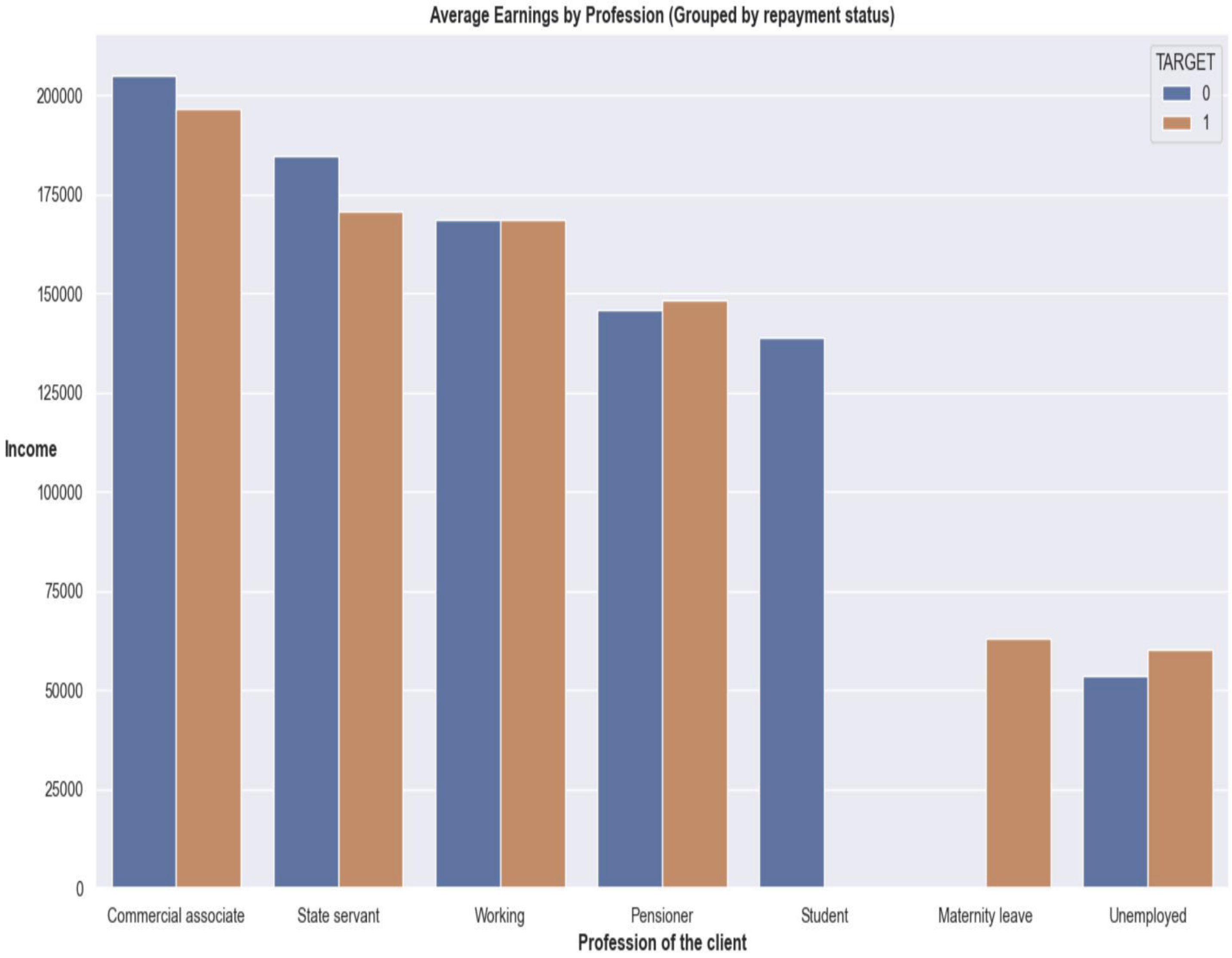


Distribution of client by car ownership based on repayment status



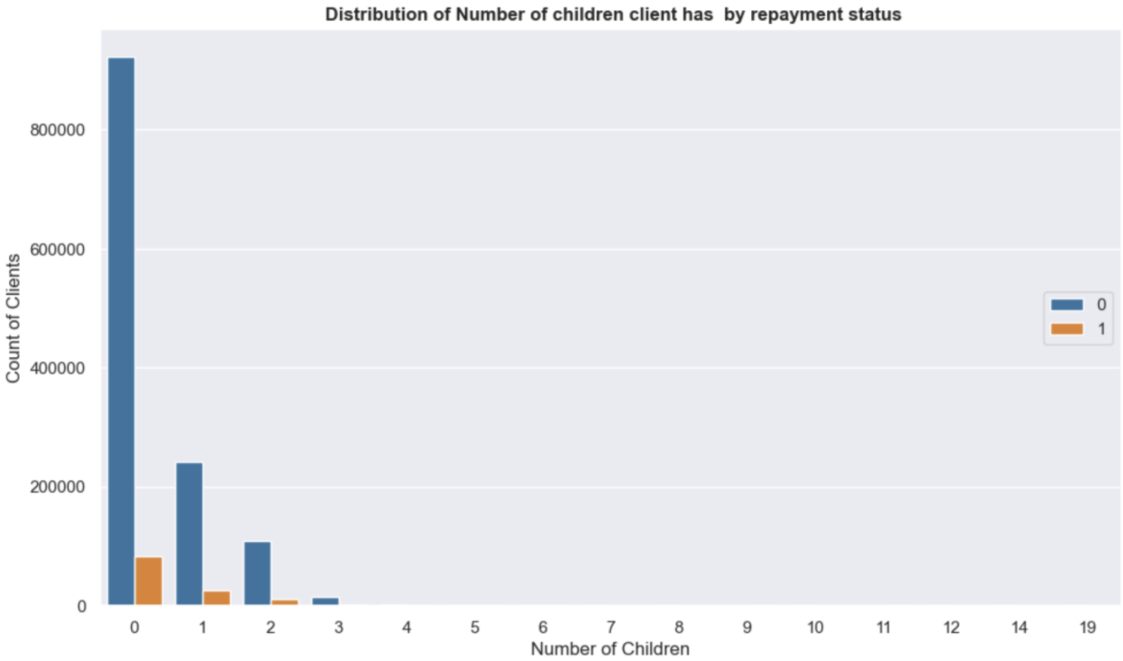
Analysis of NAME_INCOME_TYPE Vs AMT_INCOME_TOTAL Vs Target

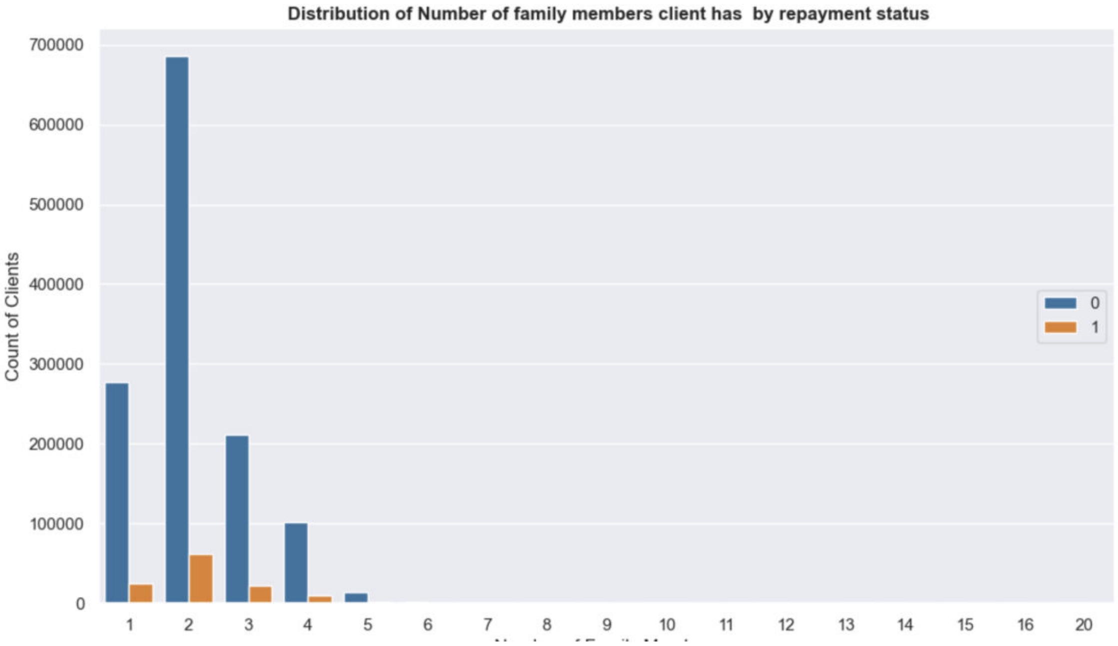
- In both cases of repayment status, commercial associate clients are the highest earners.
- Clients who are on maternity leave and unemployed have difficulty in making payments
- There are almost an equal number of clients under the working category who repay and default.



Analysis of CNT_FAM_MEMBERS Vs TARGET

- In the majority of cases, people with children are finding it difficult to repay the loan.
- People with family sizes less than or equal to 2 are the better repayers of loan.

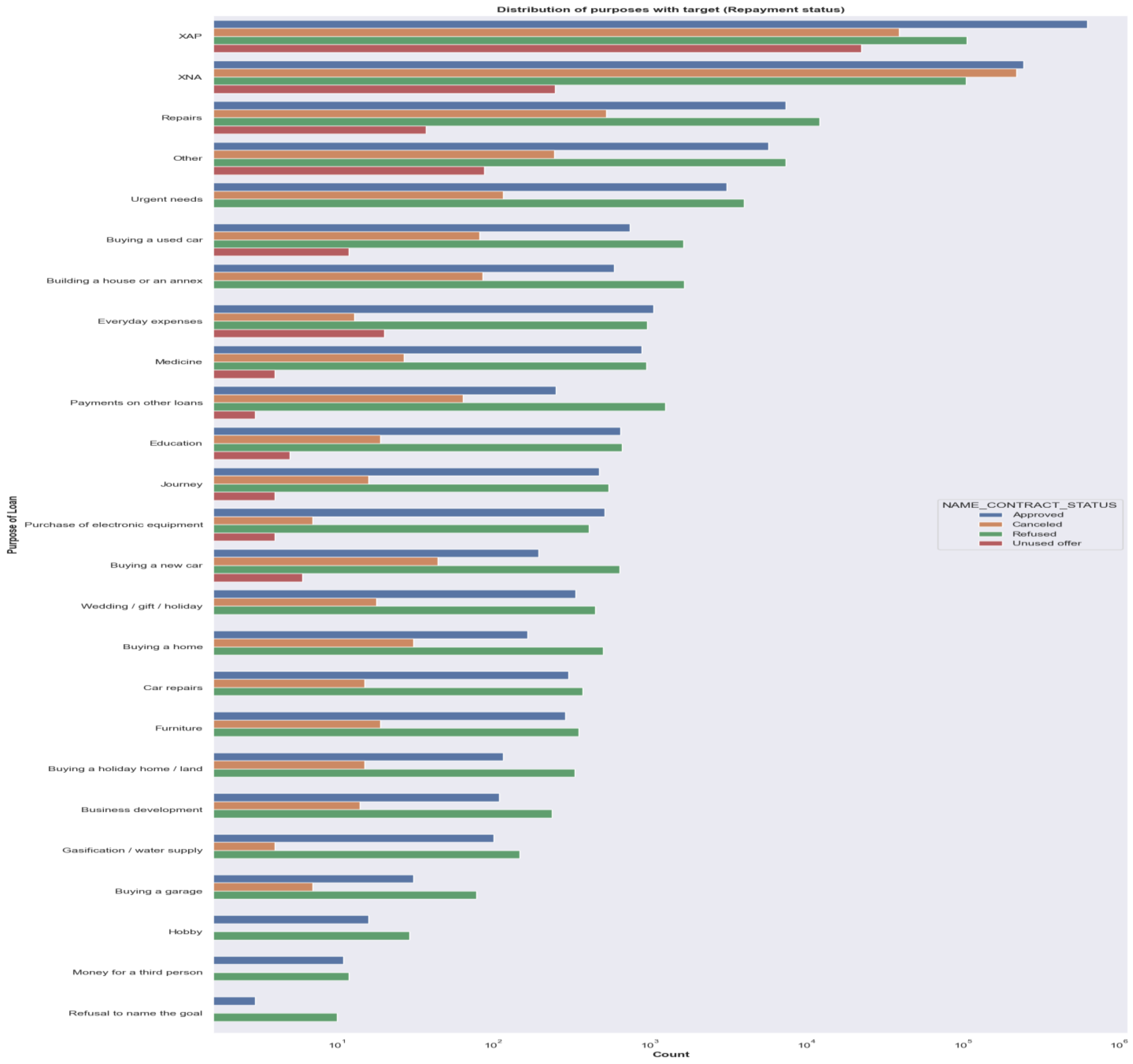




Analysis of NAME_CASH_LOAN_PURPOSE Vs NAME_CONTRACT_STATUS

- Most rejection and approval of loans is when the purpose of the client is based on Repairs.

- For education purposes, we have the equal number of approvals and refusals.



Top 10 Correlations for Merged Data

Non Defaulters

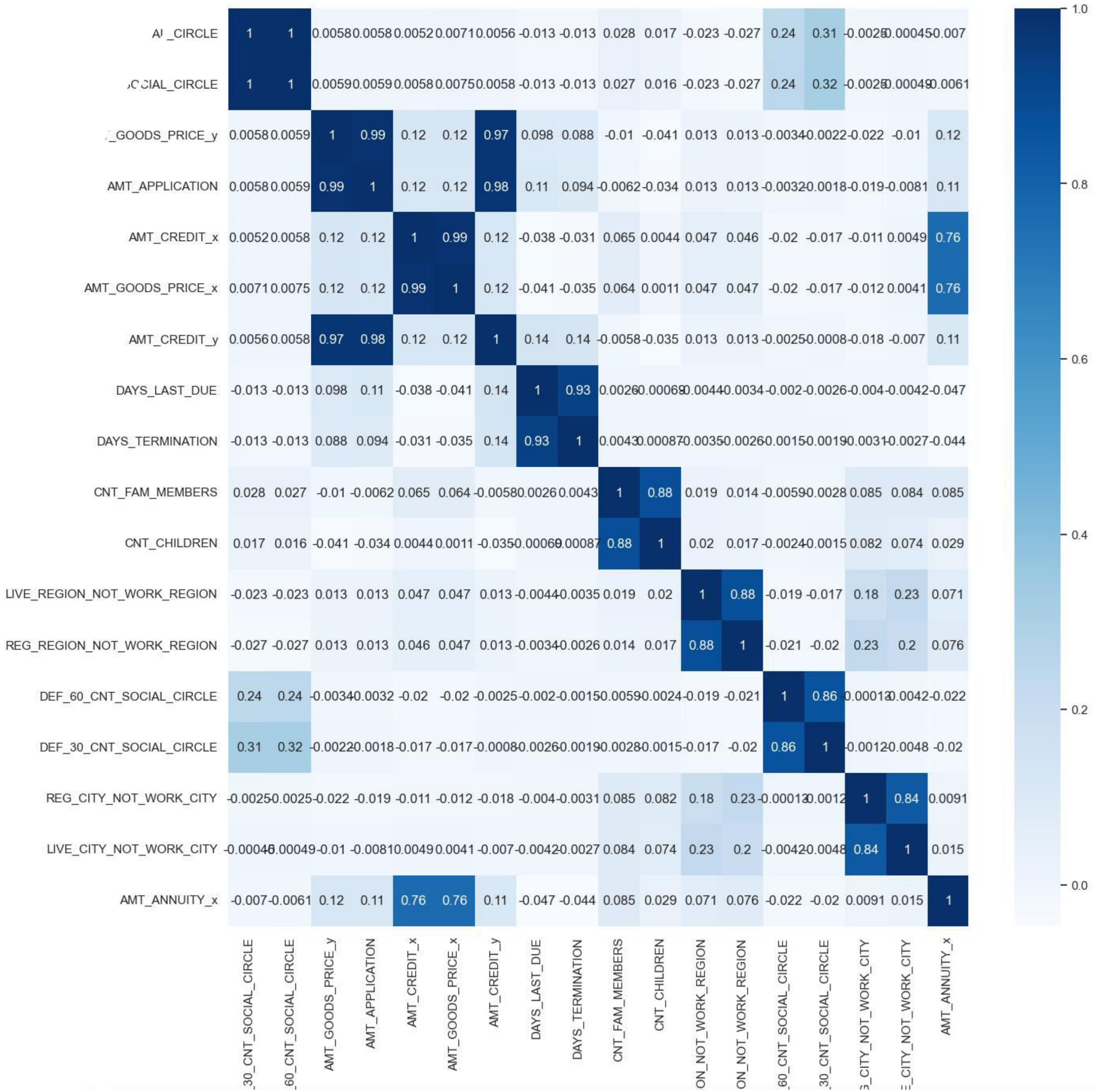
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	1.00
AMT_GOODS_PRICE_y	AMT_APPLICATION	0.99
AMT_CREDIT_x	AMT_GOODS_PRICE_x	0.99
AMT_CREDIT_y	AMT_APPLICATION	0.98
DAYS_LAST_DUE	DAYS_TERMINATION	0.93
CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.88
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.84
AMT_GOODS_PRICE_x	AMT_ANNUITY_x	0.76

Defaulters

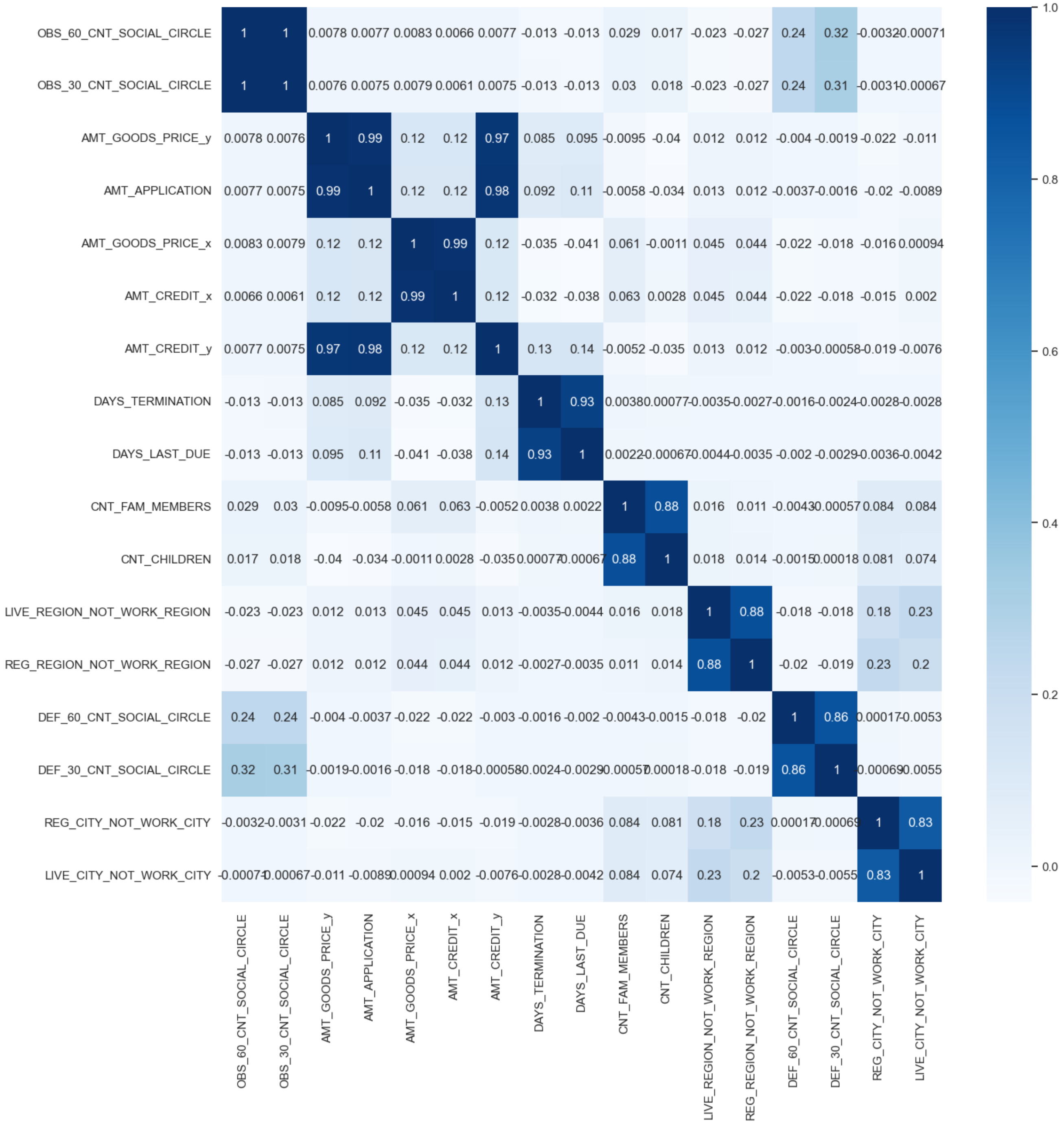
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
AMT_GOODS_PRICE_y	AMT_APPLICATION	0.99
AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.98
AMT_APPLICATION	AMT_CREDIT_y	0.98
AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.97
DAYS_TERMINATION	DAYS_LAST_DUE	0.95
CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.87
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.79

Heatmap for Repayers Data

- AMT_APPLICATION and AMT_GOODS_PRICE have a high correlation i.e goods price of goods that the client asked for (if applicable) on the previous application is directly proportional to the credit the client ask on the previous application.
- If the client's contact address does not match the work address, then there's a high chance that the client's permanent address also does not match the work address.
- DAYS_TERMINATION is highly correlated to DAYS_LAST_DUE i.e application date of the current application when the expected termination of the previous application is highly correlated to the application date of the current application when was the last due date of the previous application.
- A client with children is highly likely to have family members as well because CNT_FAM_MEMBERS is directly proportional to CNT_CHILDREN.



Heatmap for Defaulter Data



- AMT_GOODS_PRICE and AMT_APPLICATION have a higher correlation.
- If the client's contact address does not match the work address, then there's a high chance that the client's permanent address also does not match the work address.
- Higher the goods price, the higher the credit by the client.
- CNT_CHILDREN and CNT_FAM_MEMBERS are highly correlated which means a client with children is highly likely to have family members as well.

Inferences

- In majority of cases people with children are finding it difficult to repay the loan. So they shouldn't be targeted by the bank.
- Female clients should be targeted more as they are better in repaying loans than males.
- Clients with office apartment, house/apartment, municipal aparments should be prioritized.
- Clients who are on maternity leaves shouldn't be targeted as they find it difficult to repay the loan.
- Most rejection and approval of loans is when the purpose of the client is based on Repairs.
- Clients who own cars should be targeted as they repay the loan than those who doesn't own a car.
- Most rejection and approval of loans is when the purpose of the client is based on Repairs. So low risk client catogories should be prioritized.
- Clients with property types such as office apartment, house/apartment, municipal aparments have the highest repayers. So they should be targeted.

Thank You