EXPLORATORY DATA ANALYSIS

CREDIT AND RISK ANALYSIS

BY: SUMANTH.ASHOK METIMATH

Agenda

- Business Objective
- Dataset
- Approach and Methodology
- Analysis
- Conclusion

Business Objective

- The finance company is seeking key attributes within an applicant's profile to aid in the decision-making process for approving or declining a loan application.

- The company wants to understand the driving factors behind loan default, i.e. the variables which are strong indicators of default for risk assessment.

Dataset

This dataset consists of three files:

1.**'application_data.csv':** It has information about clients when they applied for a loan, specifically whether they have trouble making payments.

2.**'previous_application.csv':** This file contains details about clients' past loan applications, revealing whether those applications were approved, cancelled, refused, or resulted in an unused offer.

3.**'columns_description.csv':** This file acts as a simple guide, explaining the meanings of the different variables in the dataset.

Approach and methodology

- Understanding the Data
- Data Cleaning and Handling
- Univariate and Bivariate Analysis
- Visual Representation

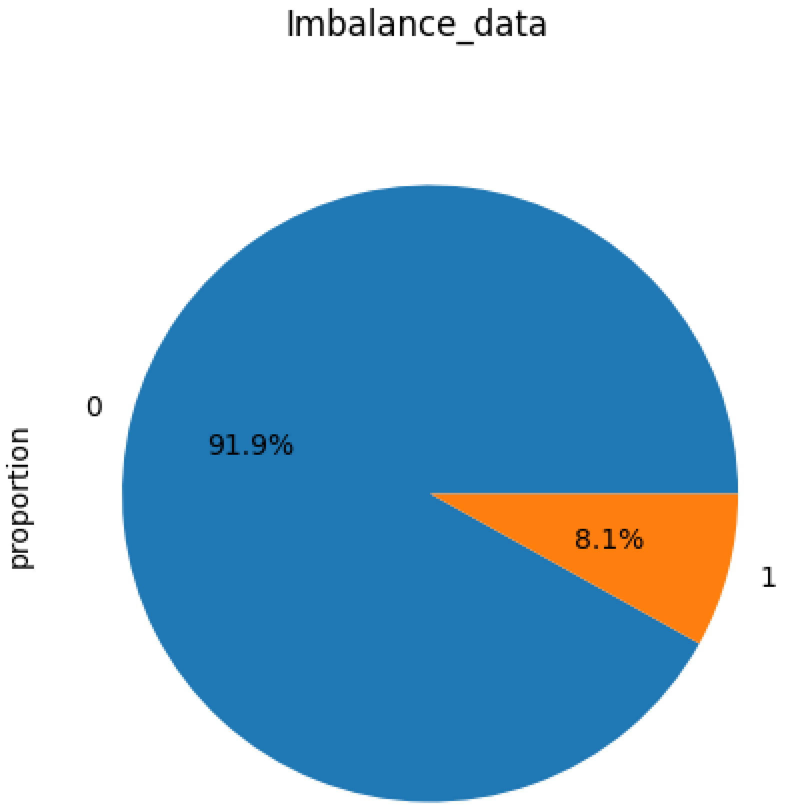
Dataset Cleaning And Handling

- Understanding shape of the data
- Identifying the percentage of null values present in each column
- Removing the columns which has more than 50% of its data as null values in application_data.csv and 30% and more in previous_data.csv
- Removing additional columns on careful consideration
- Further in the previous_data.csv, we had to specifically remove the “XNA”,”XPA”, just understand more about the purpose of the loan
- Handling outliers by creating bins.

ANALYSIS

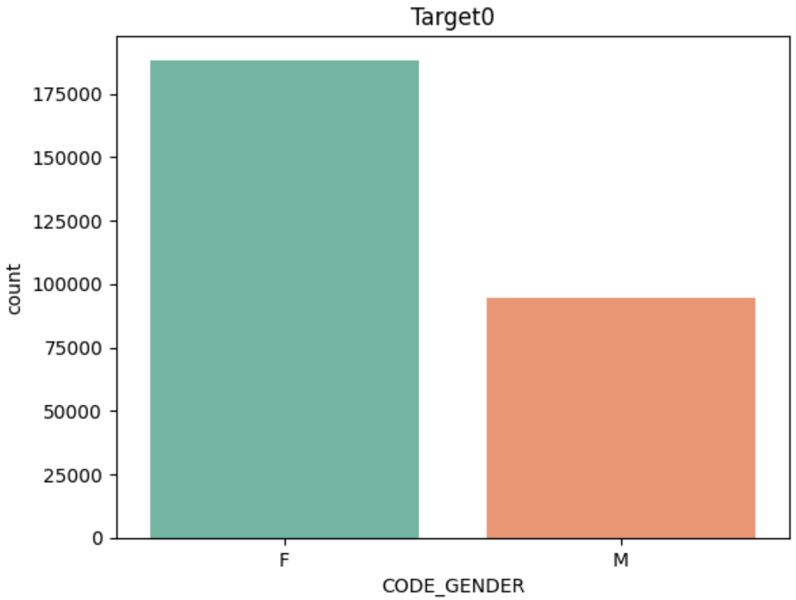
Analysis Based On Target Column

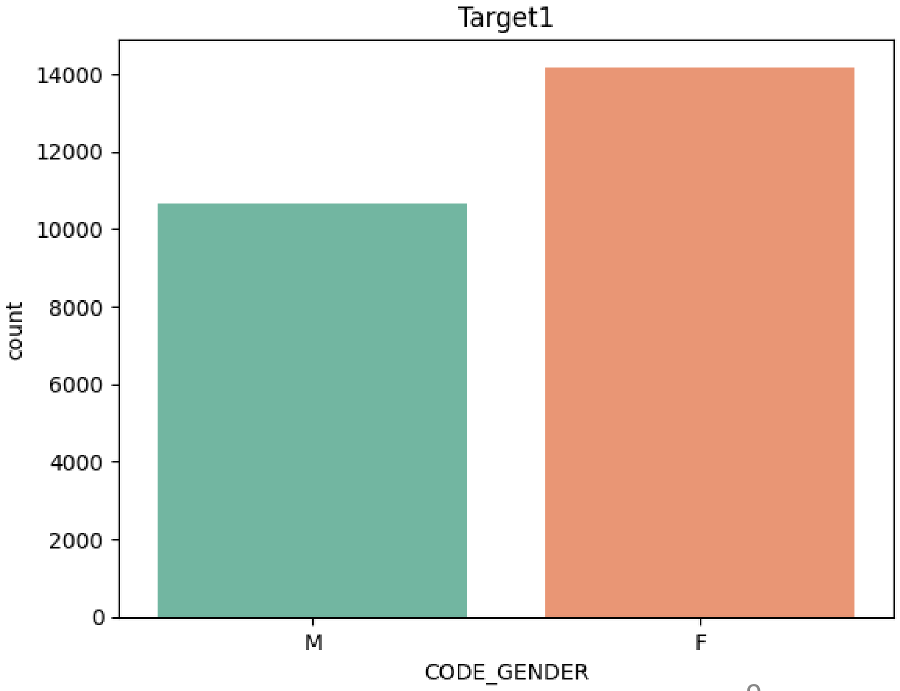
- Target Column is the column, which tells about the clients who has been paying the loan on time and who has become defaulters.
- In the column , 0 represents people who pay loan on time and 1 represents defaulters
- From the graph we can see 92% of the people pay loan on time and around 8% who are facing difficulties.
- As there is imbalance in data , we will do analysis separately on the each of the unique value i.e is 0 and 1.



Univariate analysis

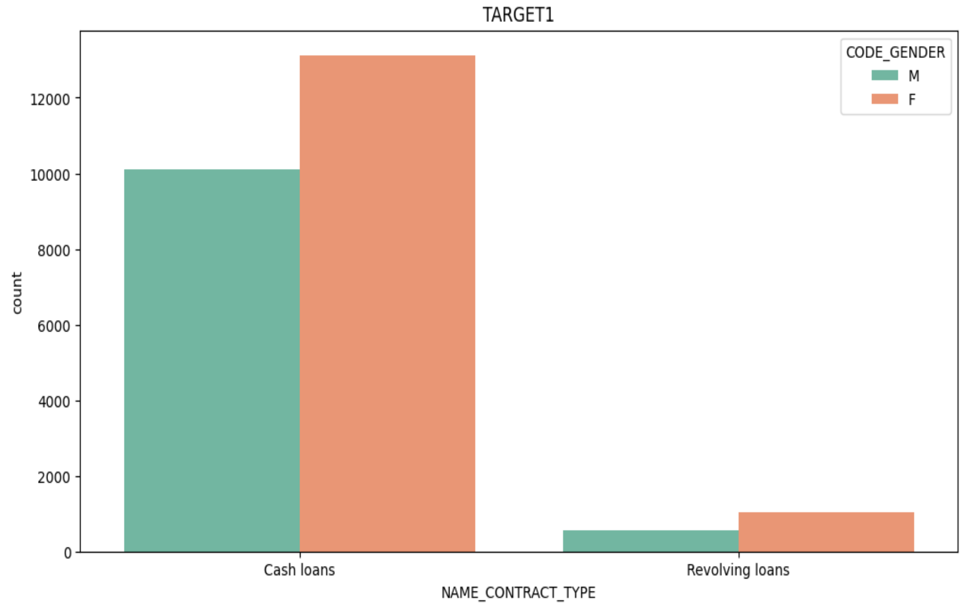
- We tried to understand the gender ratio in the non-defaulters and the defaulters list.
- From the graph , we can see that female are the most who apply for the loan.
- In defaulter we can see comparatively female are higher than male.
- Around 14,000 female are defaulters and around 10500 male are defaulters.

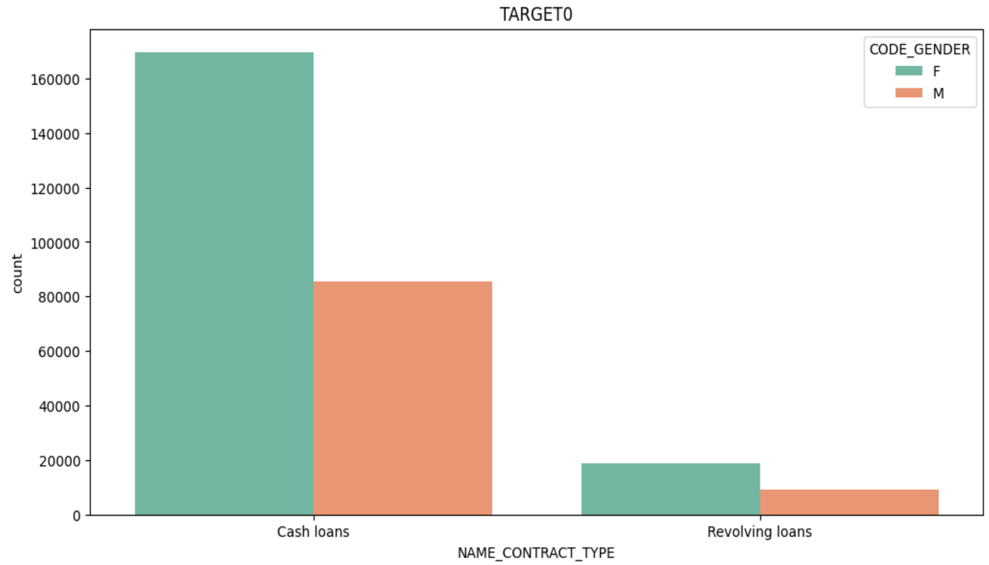




Univariate analysis

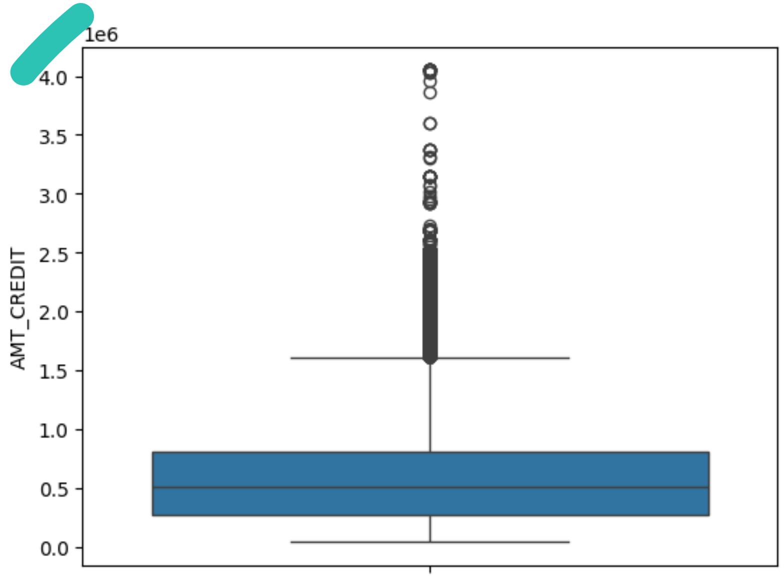
- Cash loan is the most preferred type of contract while opting a loan.
- From the graph , we can see that female are the most who apply for the loan.
- In defaulter we can see comparatively female are higher than male.

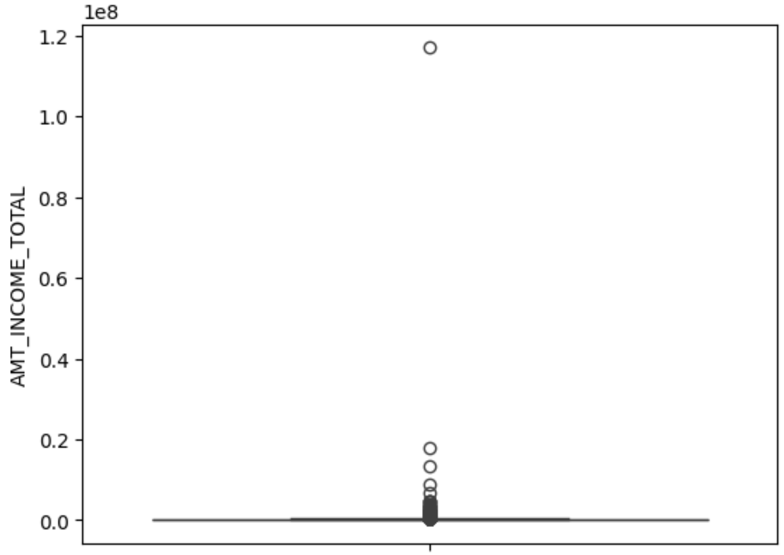




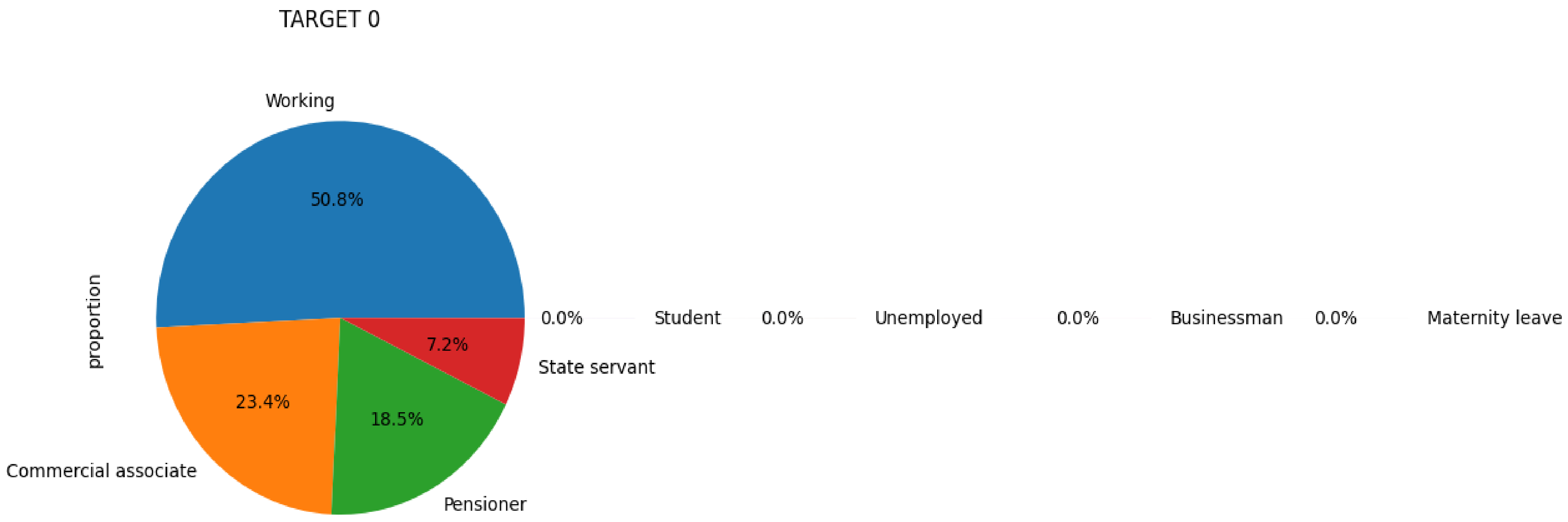
Univariate analysis

- We can see from the box plots that the credit amount and Income amount have lot of outliers
- These can make our analysis and study go wrong.
- Hence, we will be using the binning method
- Binning is a method in which , the whole data is grouped into bins.



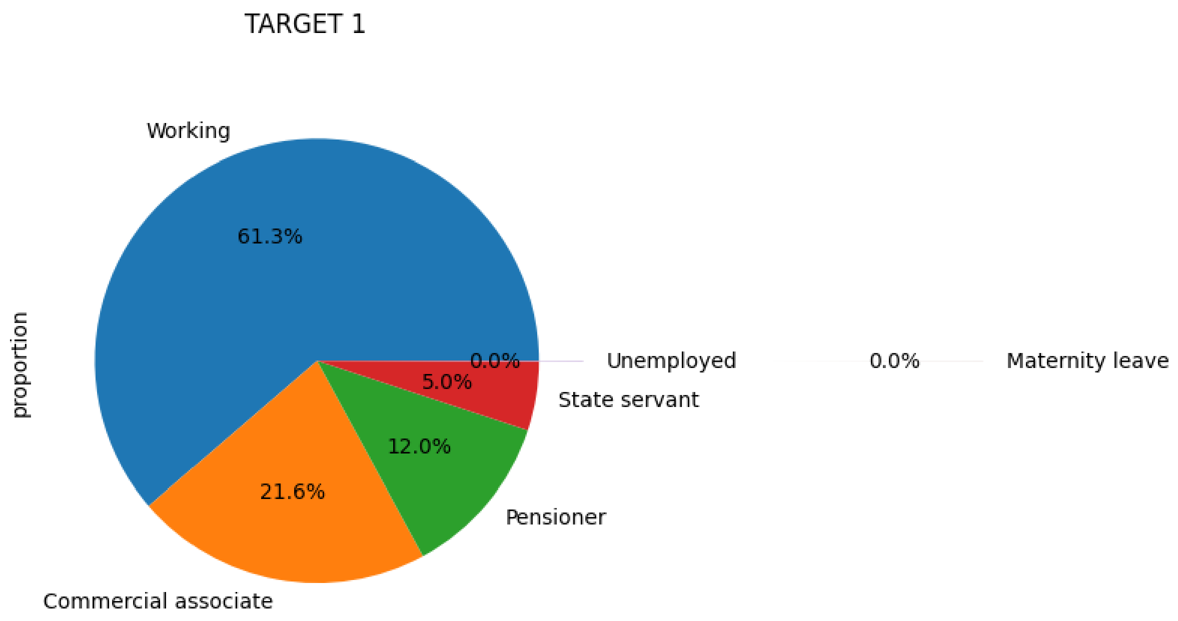


INFLUENCE OF INCOME TYPE ON LOAN



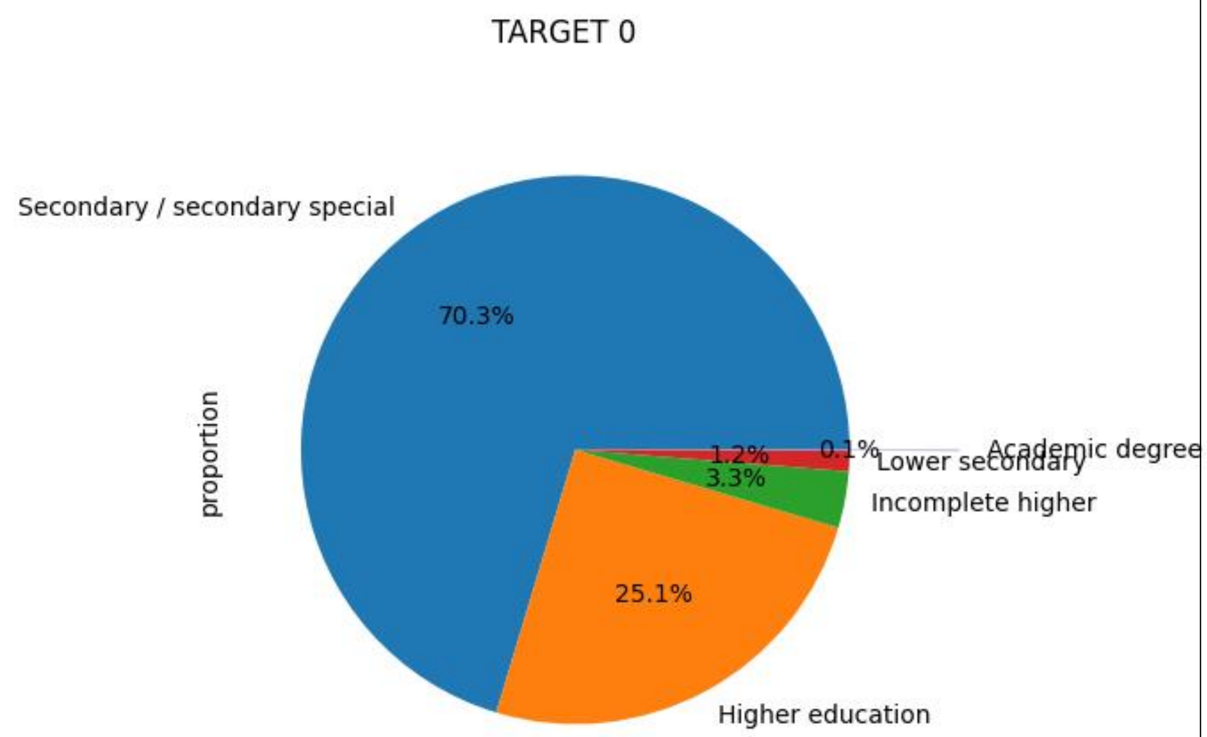
- From this we can say, majority loan appliers are from the working class followed by commercial associate and pensioner
- Even though there are application from student ,unemployed and businessman, in the huge data set , the count remains between 90 – 100 out of 100000, due to which while representation it has considered it as 0

INFLUENCE OF INCOME TYPE ON LOAN

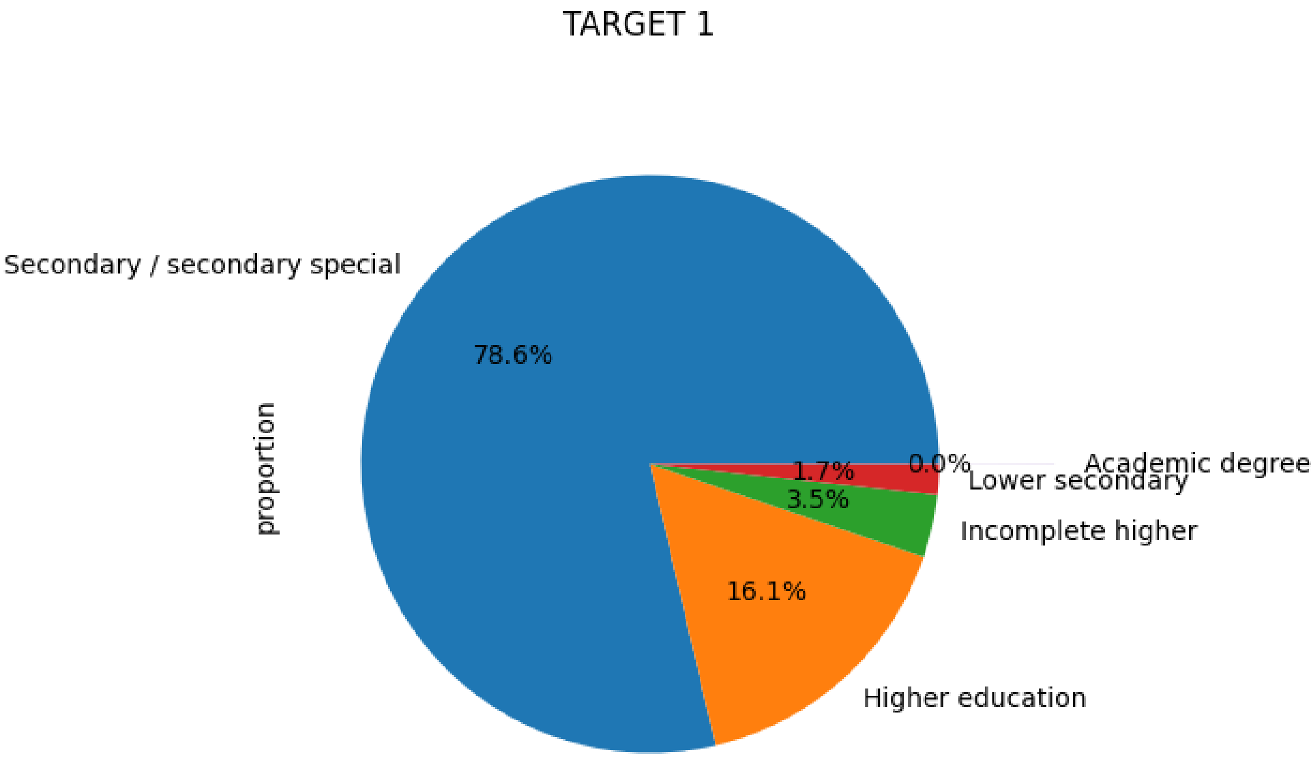


- From this we can say, majority loan Defaulters are from the working class followed by commercial associate and pensioner
- As we know , in whole dataset there is around 8% of clients who become defaulters, in which majority contribution is from the working profession , who fail to repay the loan on time.

INFLUENCE OF EDUCATION ON LOAN



Customers who have paid the loan



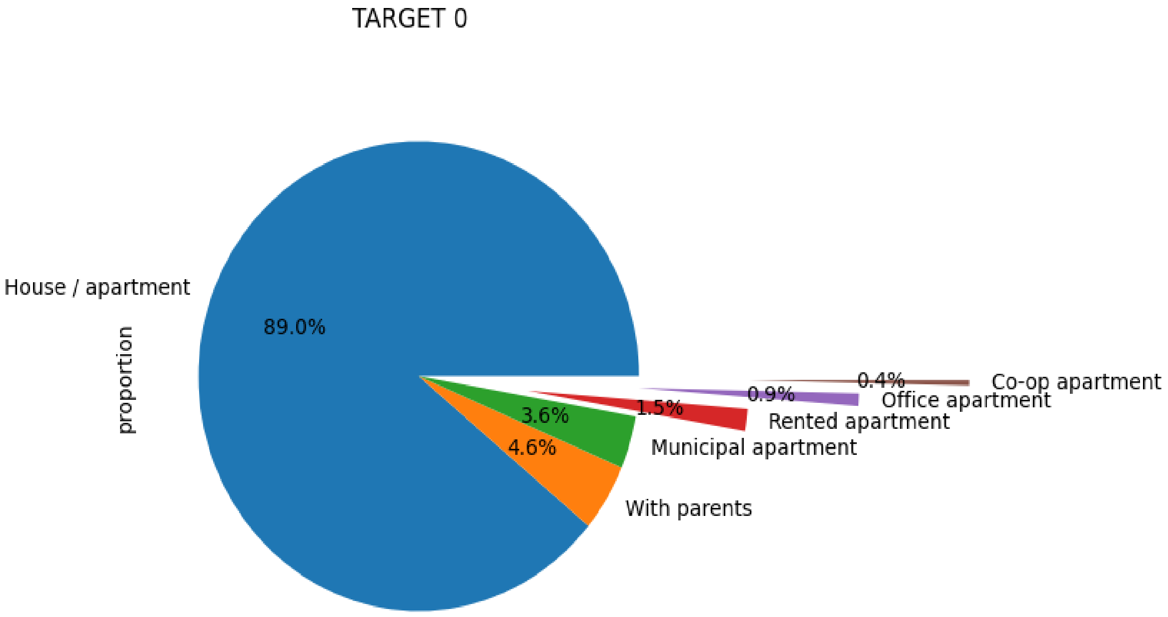
Customers facing difficulties paying the loan

INFLUENCE OF EDUCATION ON LOAN

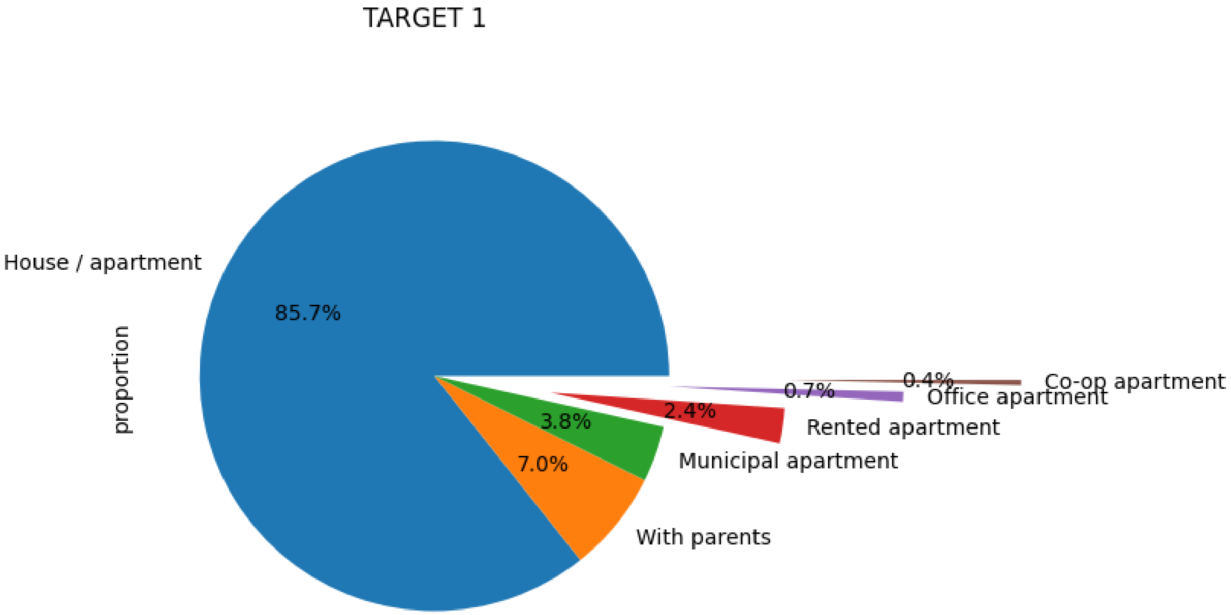
Education plays an indirect and important role in risk analysis.

- Customers with a secondary education background are more likely to defaulters
- On the contrary, customers with higher education tend to be more punctual in repayment of their loans.
- In conclusion, there is an inverse relationship between education level and risk – higher education correlates with lower risk, and vice versa.

INFLUENCE OF HOUSING ON LOAN



Customers who have paid the loan



Customers facing difficulties paying the loan

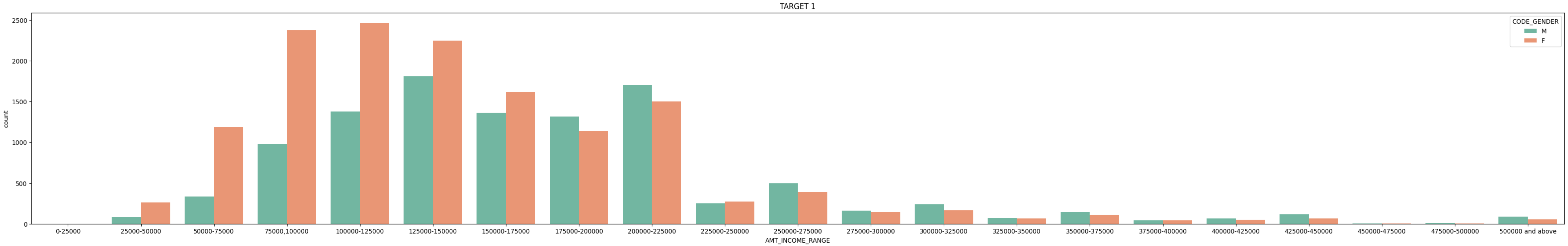
INFLUENCE OF HOUSING ON LOAN

Housing plays in reducing the risk in the loan by making it as collateral, hence understanding it is a must.

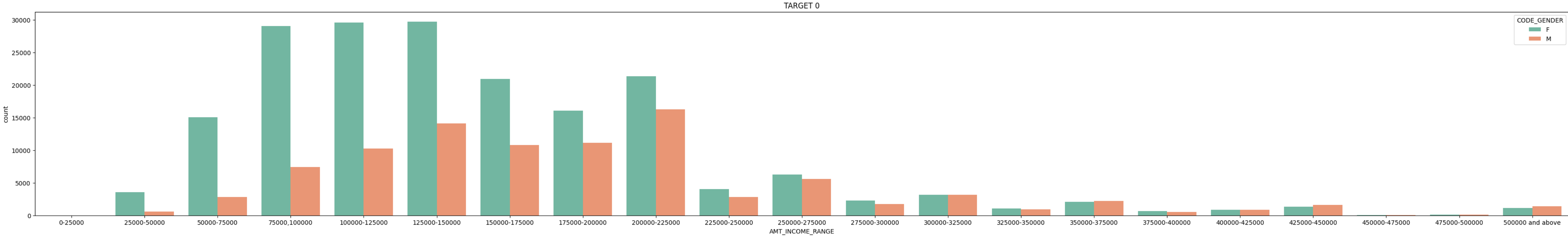
- Clients living with parents are more likely to be defaulters, as we can see the ratio between them in the graph

i.e only 5% of them pay their loan on time , but 7% of them fail to pay the amount

THE INCOME & THE LOAN



Customers facing difficulties paying the loan

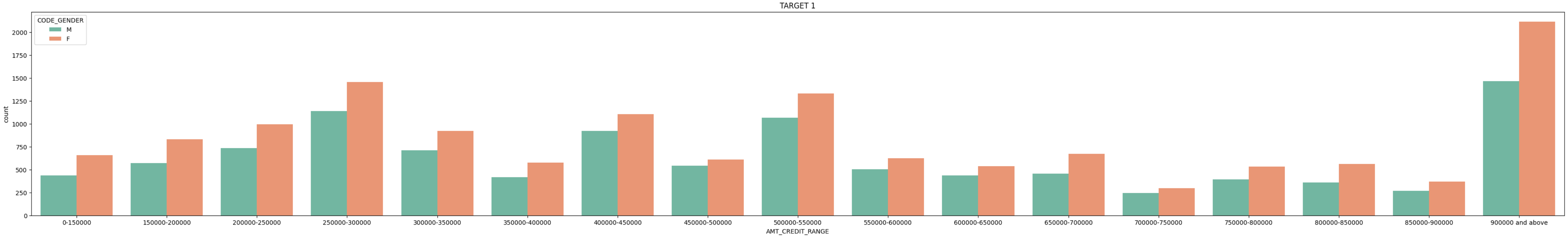


Customers who have paid the loan

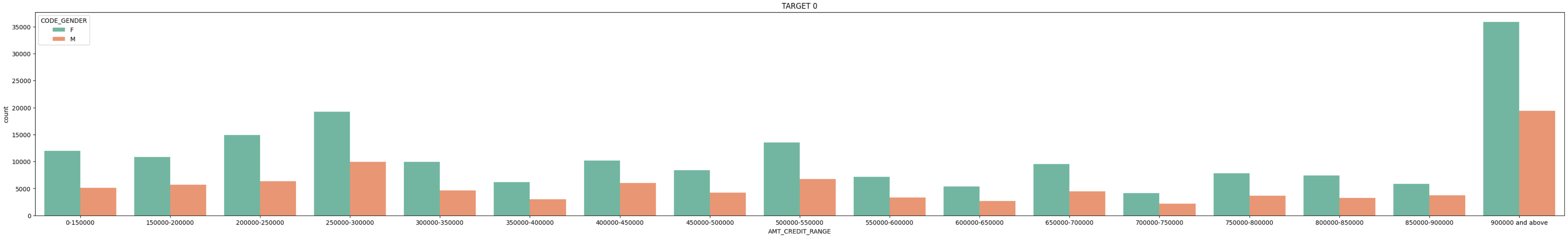
THE INCOME & THE LOAN

- People with income of 0 to 2 lakhs are the majority clients who apply for the loan
- People with income of 0 to 2 lakhs face most difficulties in paying back the loan

Credit Amount of the loan



Customers facing difficulties paying the loan

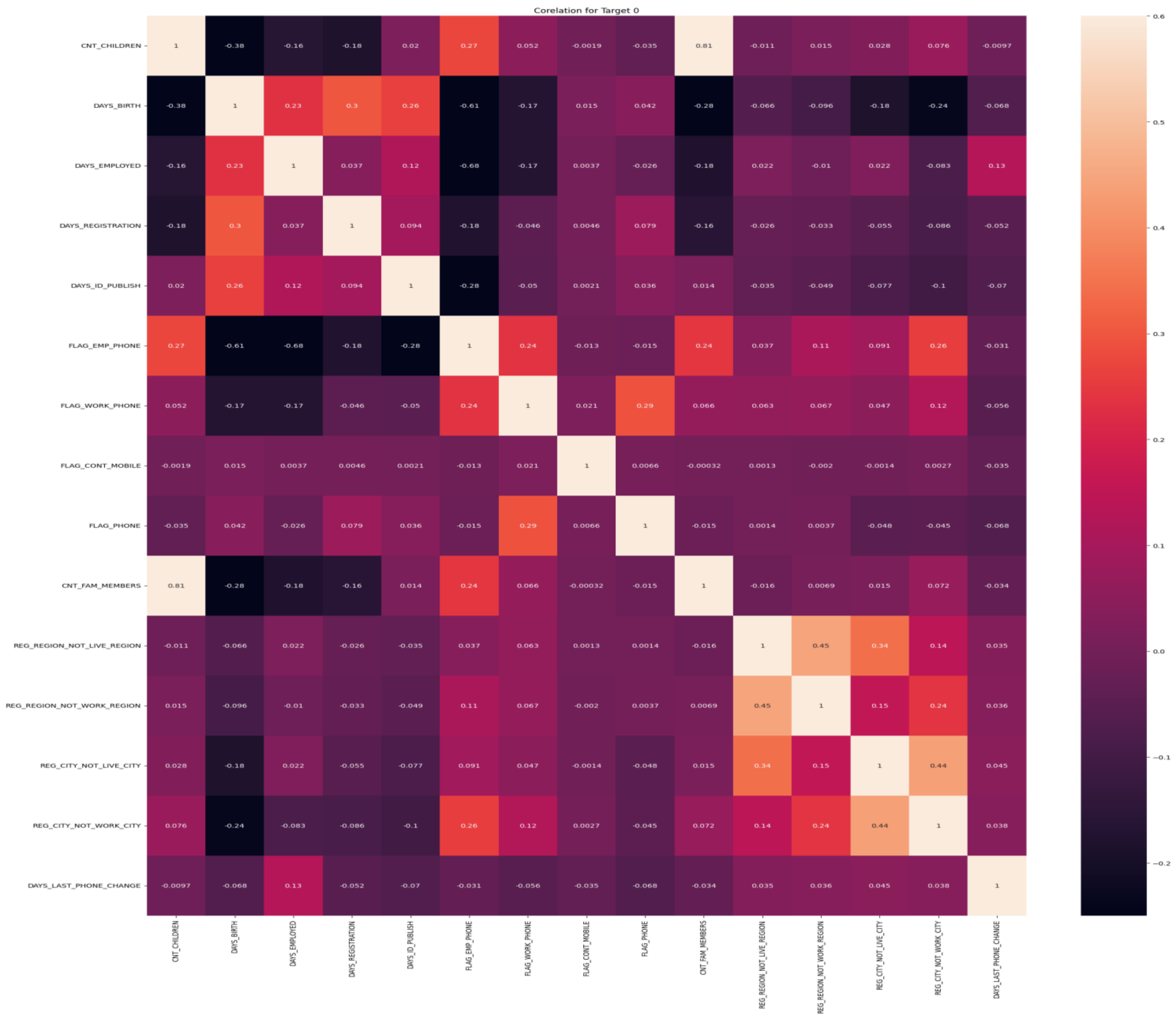


Customers who have paid the loan

THE INCOME & THE LOAN

- People usually tend to take , Loan amount more than Rs.900000 and above
- The outliers who had taken considerably more loan all are considered in this bin hence majority clients are in this range.

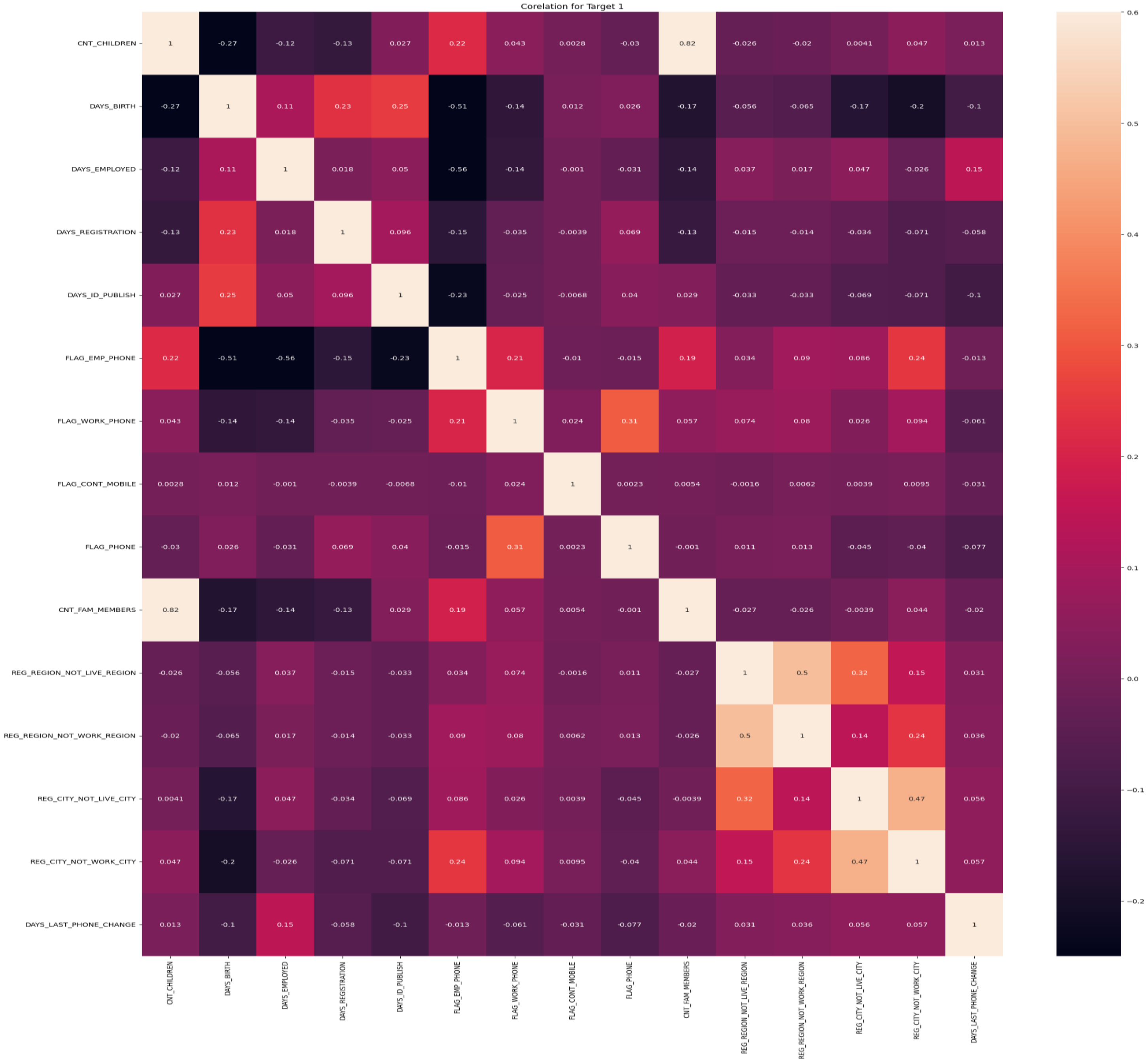
HEATMAP TARGET 0



	CNT_CHILDREN	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	FLAG_EMP_PHONE	FLAG_WORK_PHONE	FLAG_CONT_MOBILE	FLAG_PHONE	CNT_FAM_MEMBERS	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	DAYS_LAST_PHONE_CHANGE
CNT_CHILDREN	1	-0.38	-0.16	-0.18	0.02	0.27	0.052	-0.0019	-0.035	0.81	-0.011	0.015	0.028	0.076	-0.0097
DAYS_BIRTH	-0.38	1	0.23	0.3	0.26	-0.61	-0.17	0.015	0.042	-0.28	-0.066	-0.096	-0.18	-0.24	-0.068
DAYS_EMPLOYED	-0.16	0.23	1	0.037	0.12	-0.68	-0.17	0.0037	-0.026	-0.18	0.022	-0.01	0.022	-0.083	0.13
DAYS_REGISTRATION	-0.18	0.3	0.037	1	0.094	-0.18	-0.046	0.0046	0.079	-0.16	-0.026	-0.033	-0.055	-0.086	-0.052
DAYS_ID_PUBLISH	0.02	0.26	0.12	0.094	1	-0.28	-0.05	0.0021	0.036	0.014	-0.035	-0.049	-0.077	-0.1	-0.07
FLAG_EMP_PHONE	0.27	-0.61	-0.68	-0.18	-0.28	1	0.24	-0.013	-0.015	0.24	0.037	0.11	0.091	0.26	-0.031
FLAG_WORK_PHONE	0.052	-0.17	-0.17	-0.046	-0.05	0.24	1	0.021	0.29	0.066	0.063	0.067	0.047	0.12	-0.056
FLAG_CONT_MOBILE	-0.0019	0.015	0.0037	0.0046	0.0021	-0.013	0.021	1	0.0066	-0.00032	0.0013	-0.002	-0.0014	0.0027	-0.035
FLAG_PHONE	-0.035	0.042	-0.026	0.079	0.036	-0.015	0.29	0.0066	1	-0.015	0.0014	0.0037	-0.048	-0.045	-0.068
CNT_FAM_MEMBERS	0.81	-0.28	-0.18	-0.16	0.014	0.24	0.066	-0.00032	-0.015	1	-0.016	0.0069	0.015	0.072	-0.034
REG_REGION_NOT_LIVE_REGION	-0.011	-0.066	0.022	-0.026	-0.035	0.037	0.063	0.0013	0.0014	-0.016	1	0.45	0.34	0.14	0.035
REG_REGION_NOT_WORK_REGION	0.015	-0.096	-0.01	-0.033	-0.049	0.11	0.067	-0.002	0.0037	0.0069	0.45	1	0.15	0.24	0.036
REG_CITY_NOT_LIVE_CITY	0.028	-0.18	0.022	-0.055	-0.077	0.091	0.047	-0.0014	-0.048	0.015	0.34	0.15	1	0.44	0.045
REG_CITY_NOT_WORK_CITY	0.076	-0.24	-0.083	-0.086	-0.1	0.26	0.12	0.0027	-0.045	0.072	0.14	0.24	0.44	1	0.038
DAYS_LAST_PHONE_CHANGE	-0.0097	-0.068	0.13	-0.052	-0.07	-0.031	-0.056	-0.035	-0.068	-0.034	0.035	0.036	0.045	0.038	1

CORRELATION AND CO-EFFICIENT

- The correlation is indicating a very weak positive relationship. This means that as the number of family members increases, the amount of credit tends to increase slightly.
- The correlation is indicating a moderate negative relationship. This means that as the number of children increases, the amount of credit decreases.



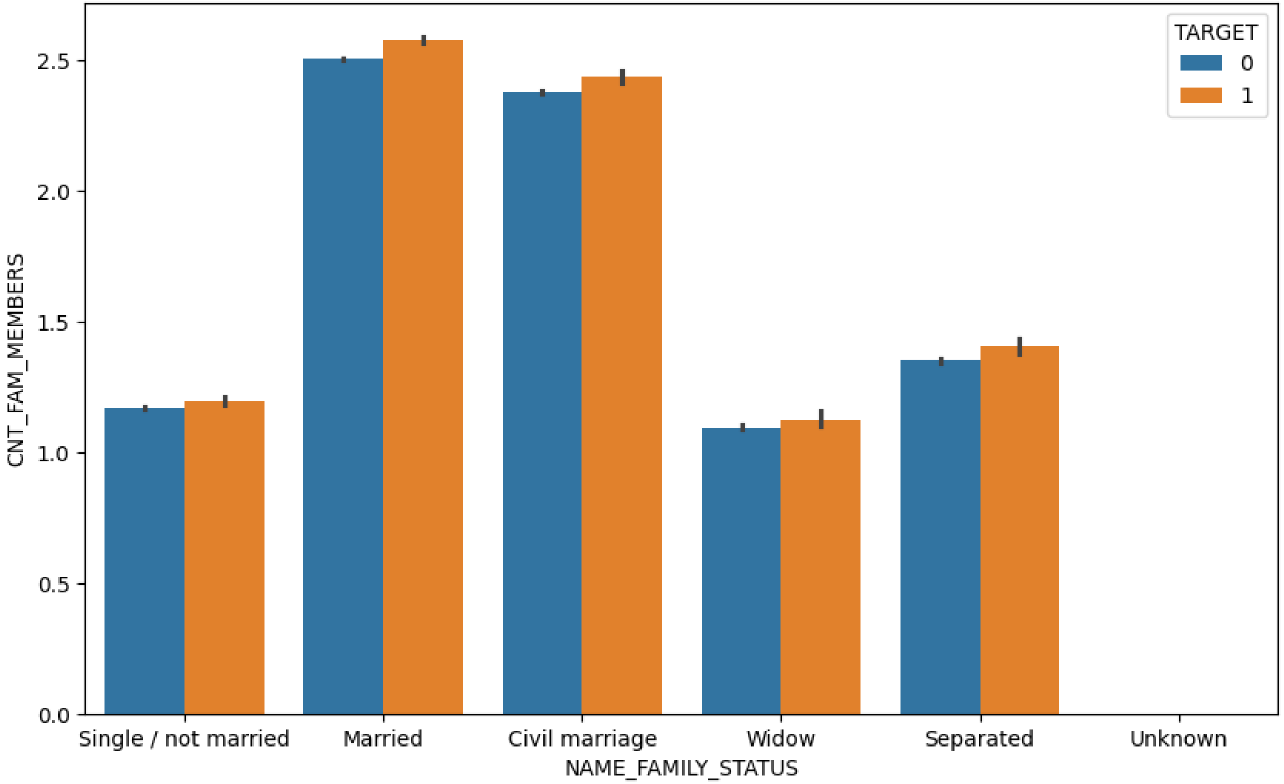
HEATMAP TARGET 1

CORRELATION AND CO-EFFICIENT

- This heat map for Target 1 is also having quite a same observation just like Target 0.
- There is a moderate positive correlation between credit amount and days employed. This suggests that clients who have been employed for a longer time are more likely to have higher credit amounts.

BIVARIATE ANALYSIS

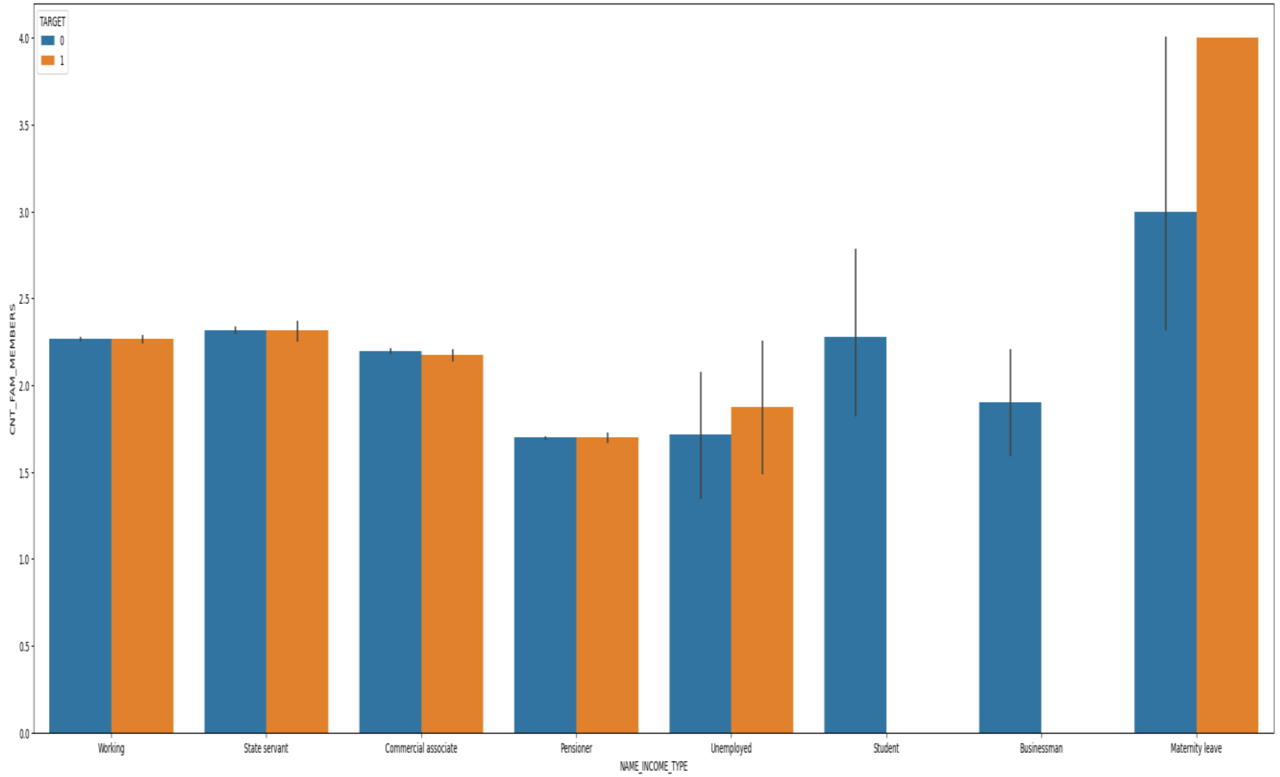
Family Status vs Count Of Family Members



- If someone is married and has five or more children, there's a higher chance they might struggle to repay loans. This could be because supporting a bigger family might create financial challenges for them.

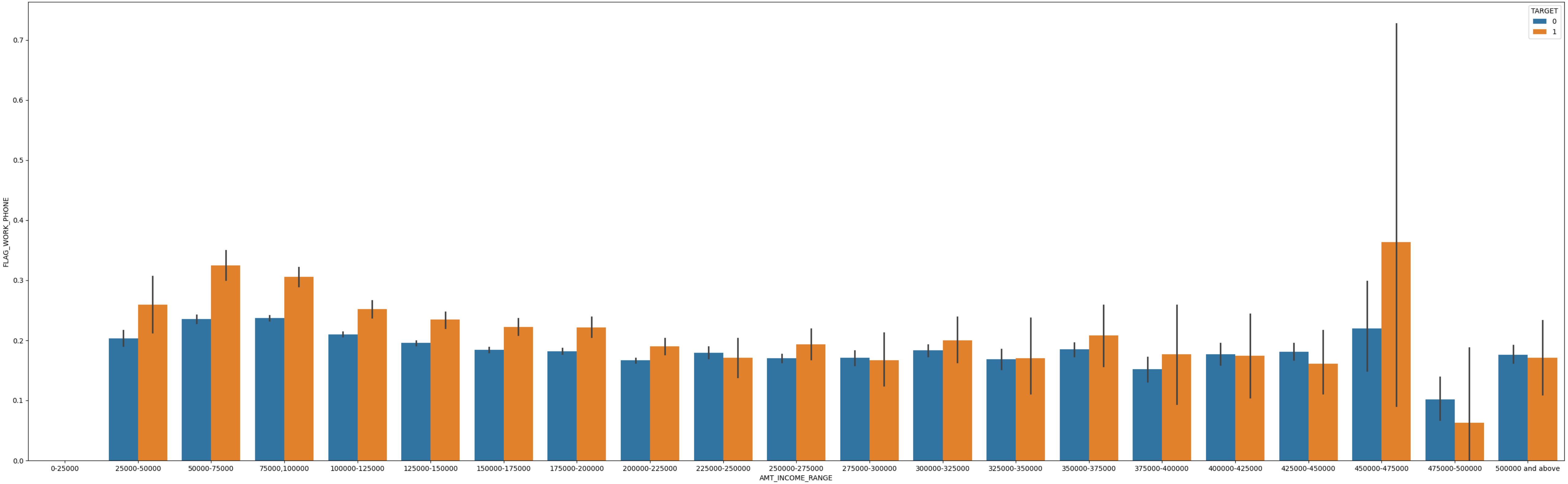
BIVARIATE ANALYSIS

Income type vs Count Of Family Members



- Those who are in maternity leave , tend to become defaulters, because the family member is gradually increasing

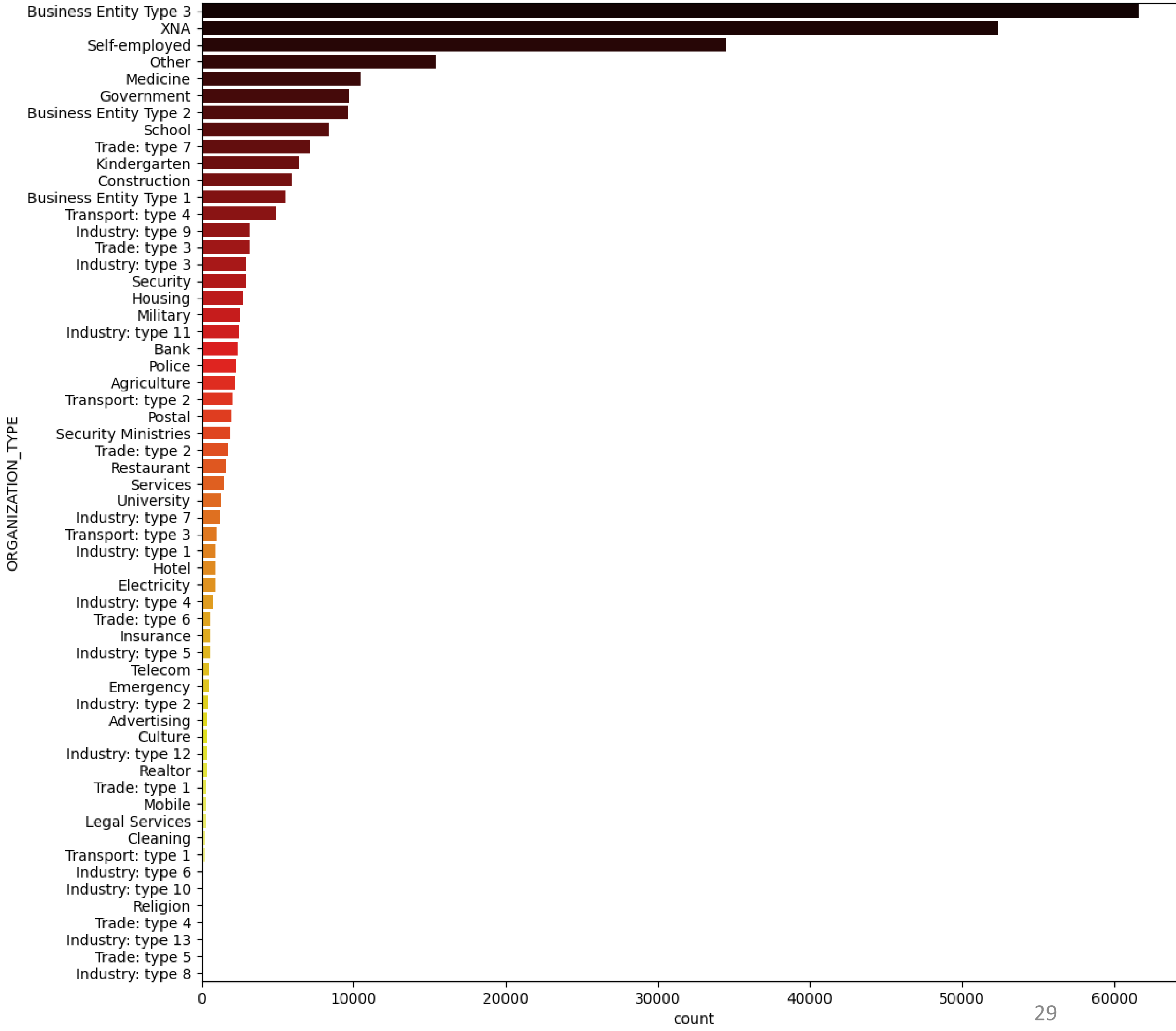
Income Range vs Flag Work phone



- People whose income range is between 4.5 to 5 lakh tend be defaulters

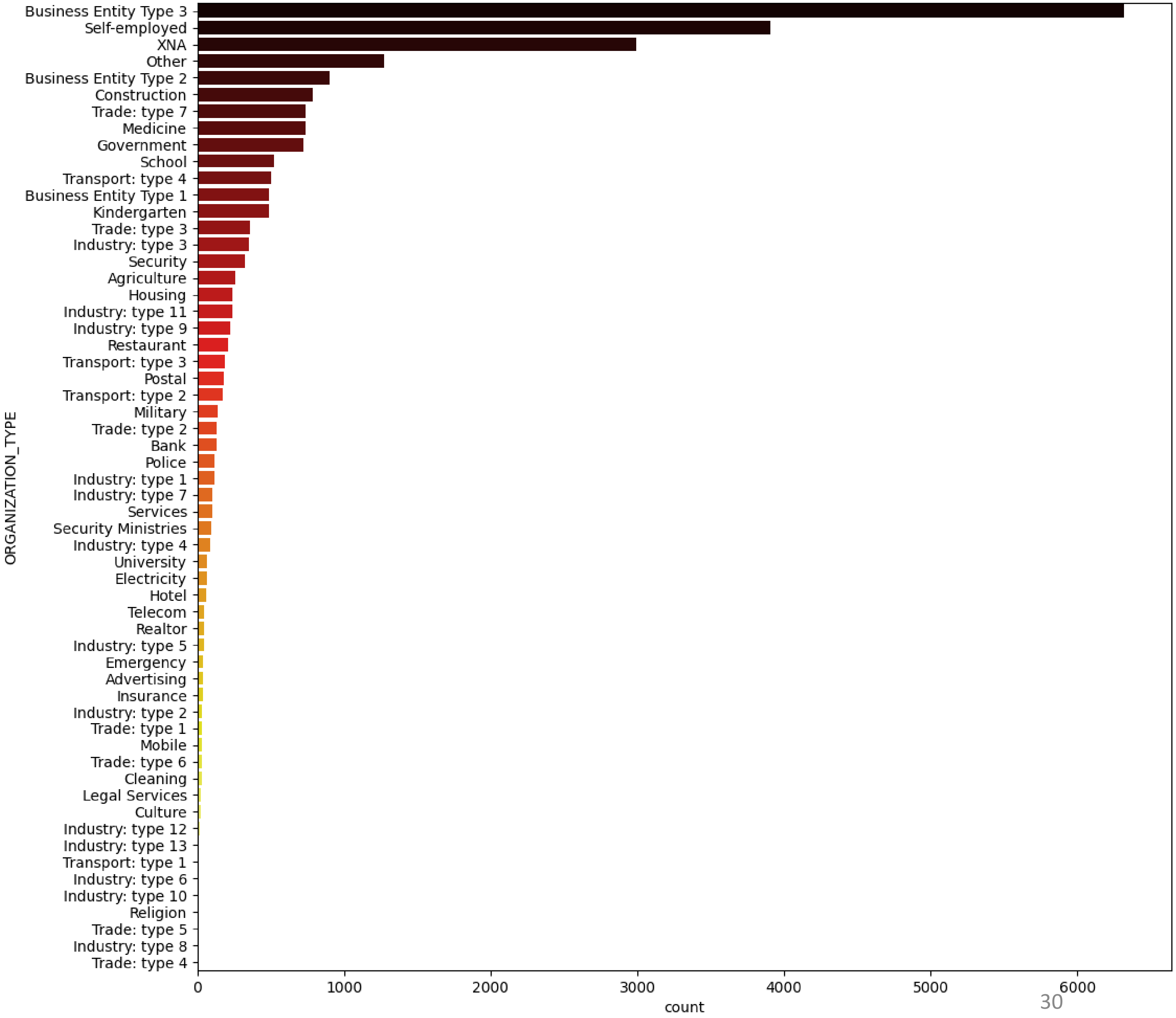
Distribution of organization type

TARGET - 0



Distribution of organization type

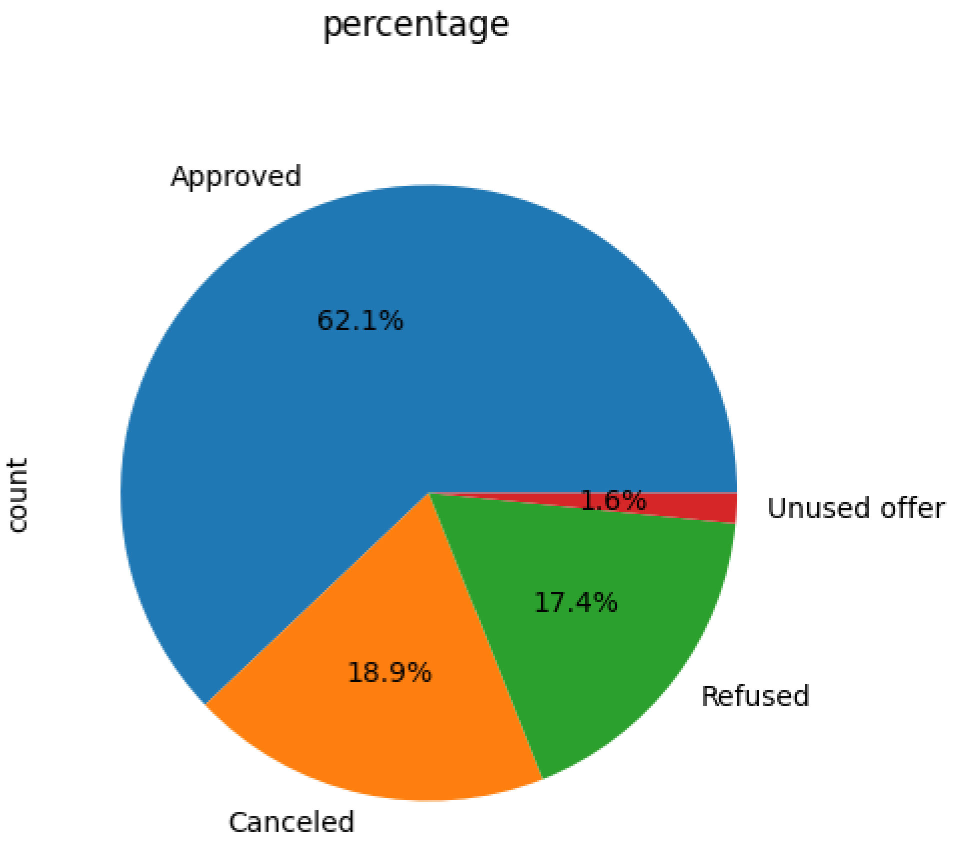
TARGET - 1



ANALYSIS ON SECOND DATASET

(PREVIOUS_DATA.CSV)

Analysis on Loan Status



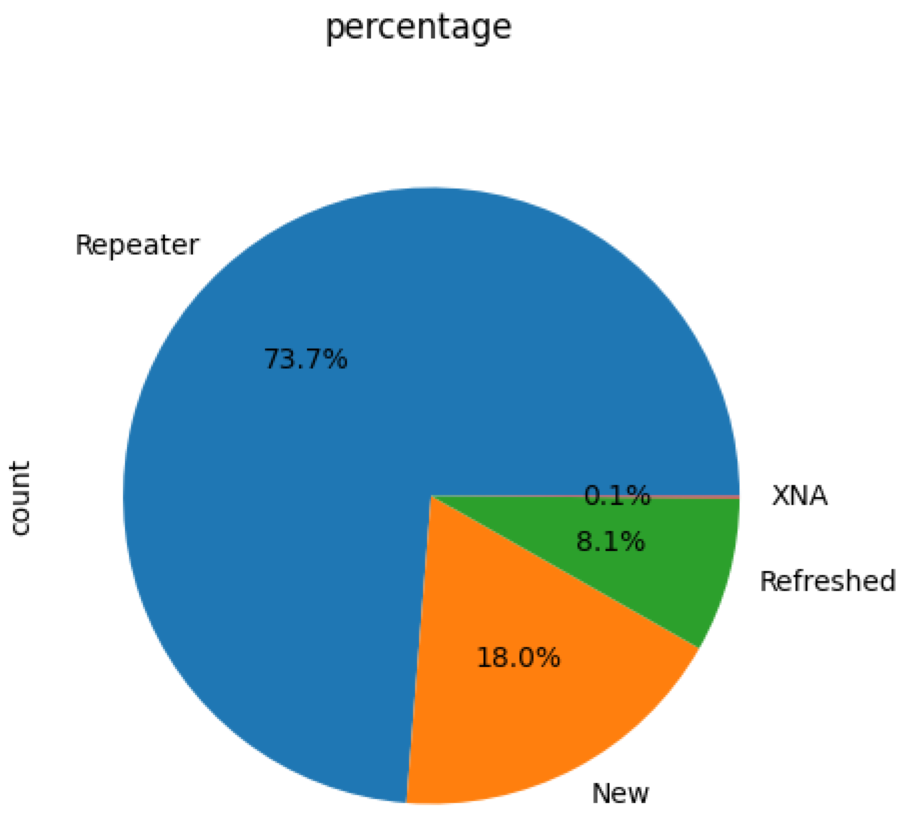
- From we can see that , around 62% of the loan application were approved

- Around 19% of the Loan application were canceled

- 17.5 % of them were refused.

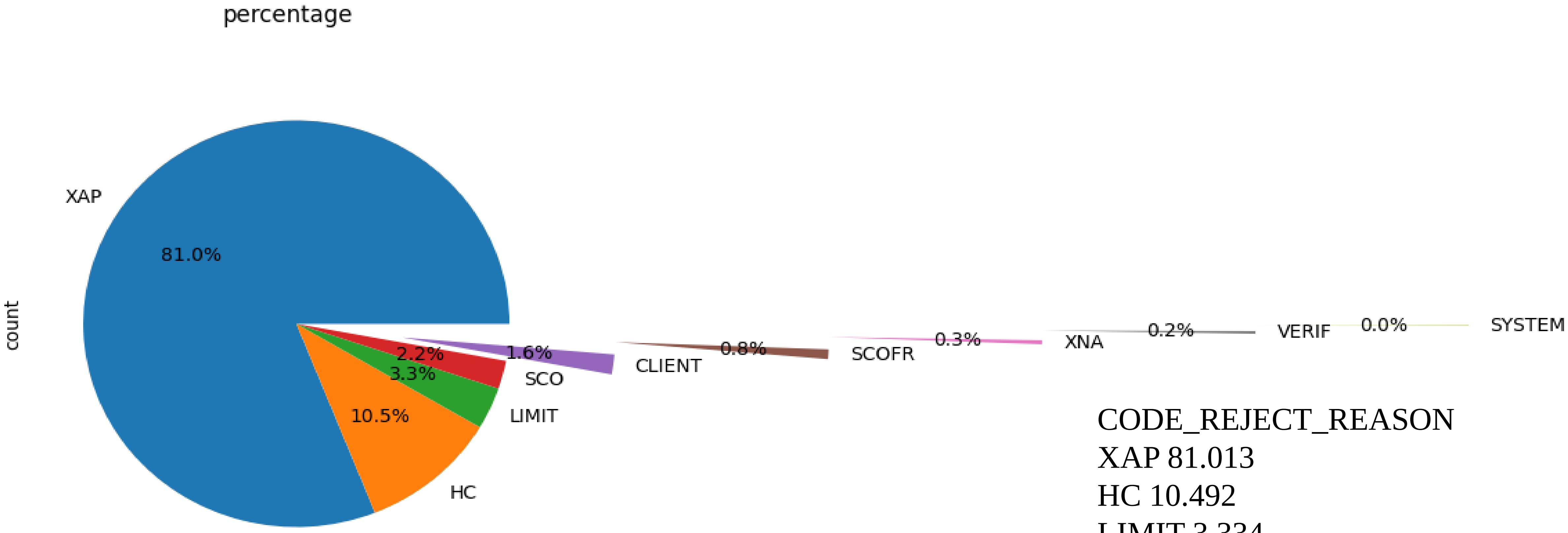
- Rest were unused offers.

Analysis on Type of Clients



- Around 74% of the Clients are repeaters who already take a loan
- Around 18% of the Loan appliers are new ,who will seeing more info about the loan
- 8.1% are refreshed.
- Rest of client details are unknown.

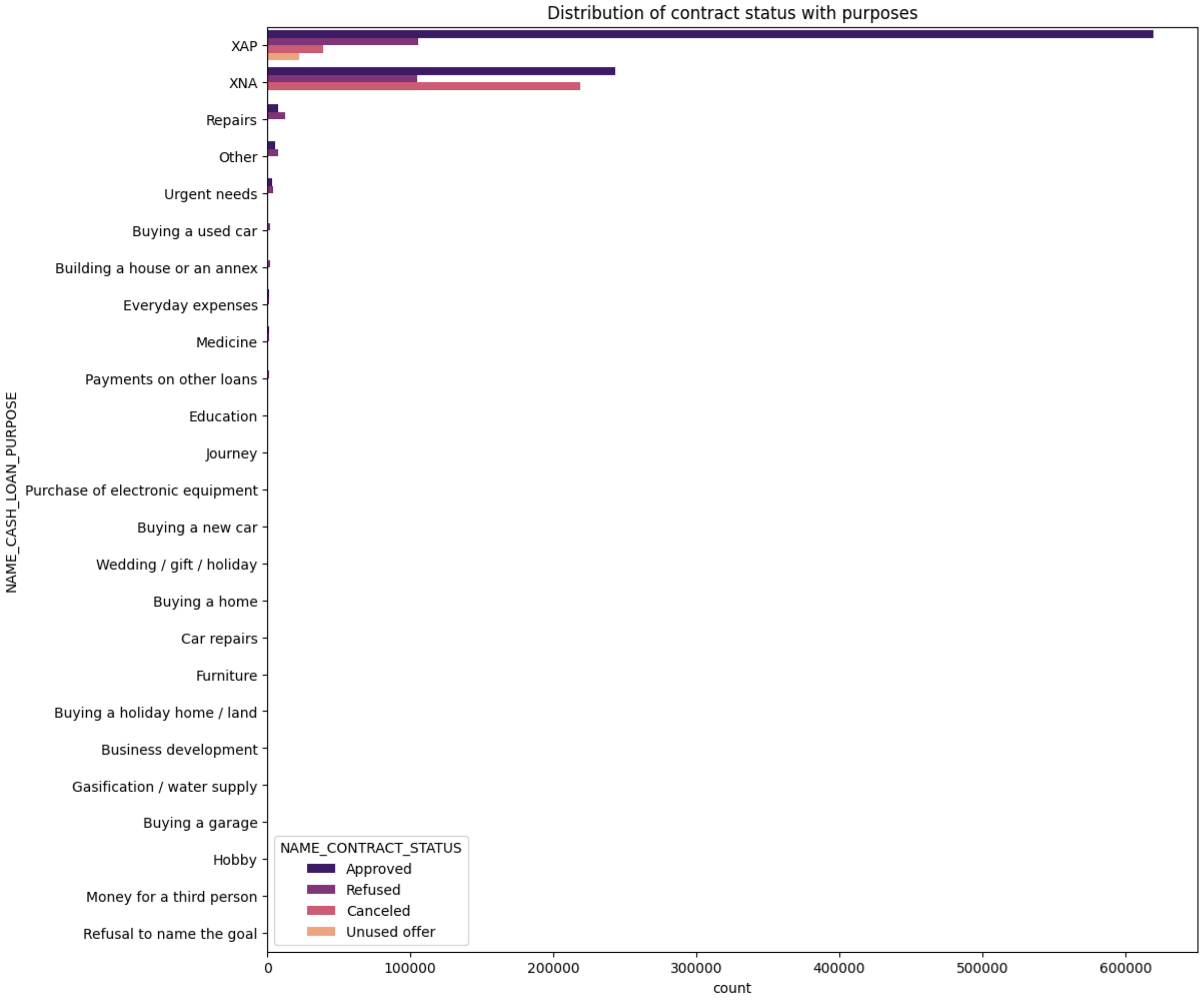
Analysis on Rejection of loan



CODE_REJECT_REASON	
XAP	81.013
HC	10.492
LIMIT	3.334
SCO	2.243
CLIENT	1.583
SCOFR	0.767
XNA	0.314
VERIF	0.212
SYSTEM	0.043

ANALYSIS ON MERGED DATASET

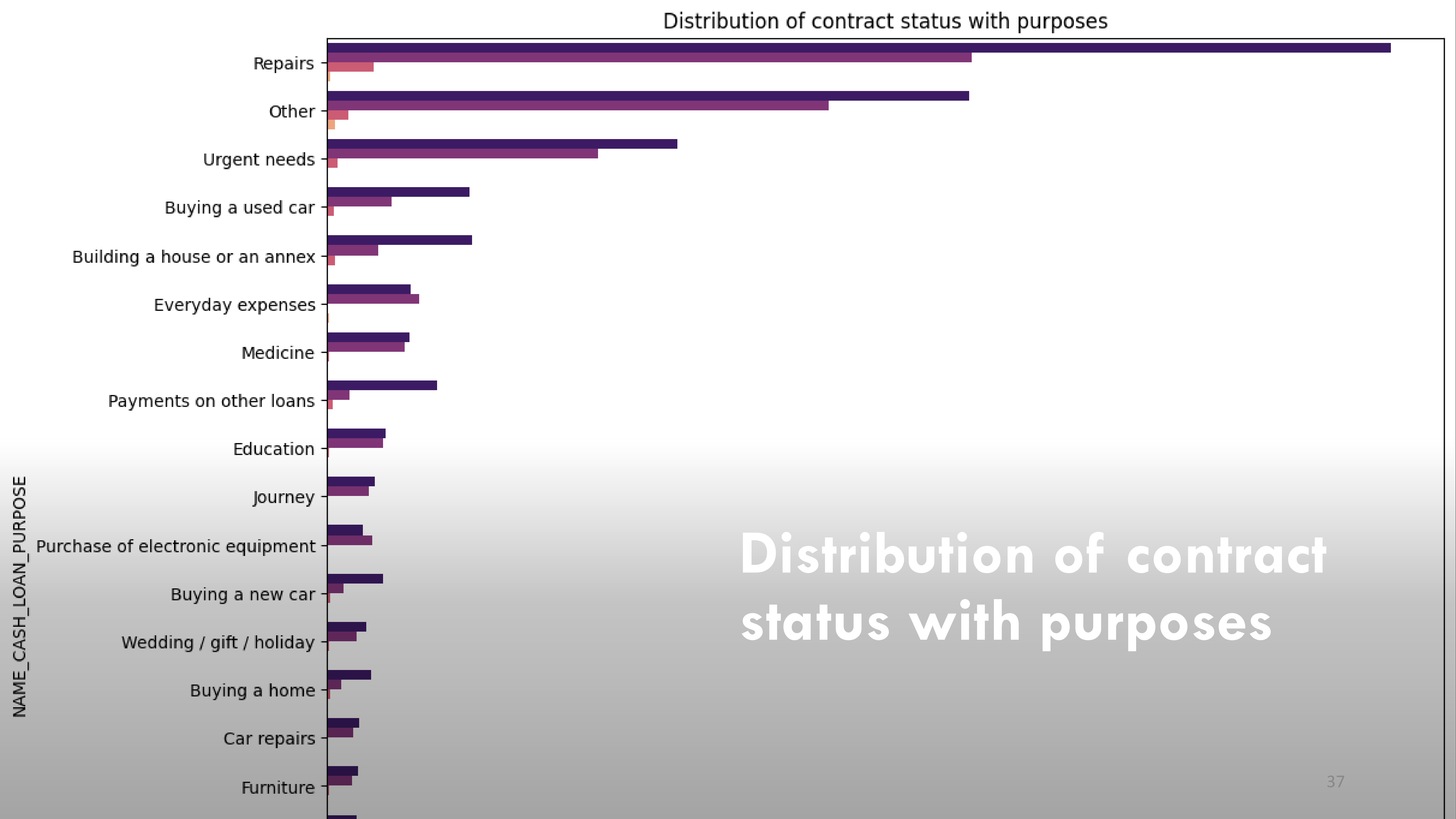
Distribution of contract status with purposes



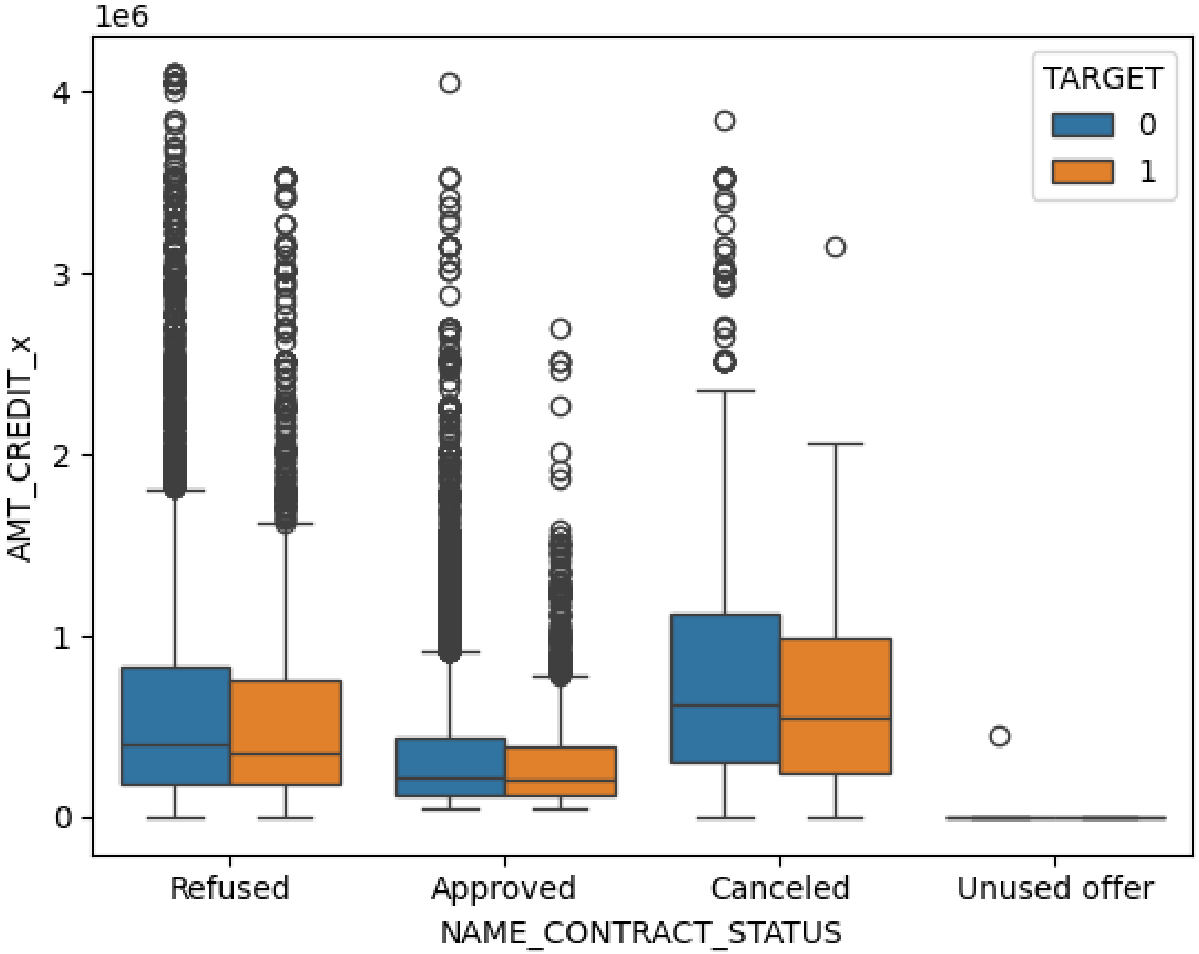
- As, we can see from the graph that majority of the reasons for loan is unknown(xna and xap)

- The total count of applications with XAP is 786905 XNA is 567329

- We will , demonstrate another graph, with xap and xna removed.

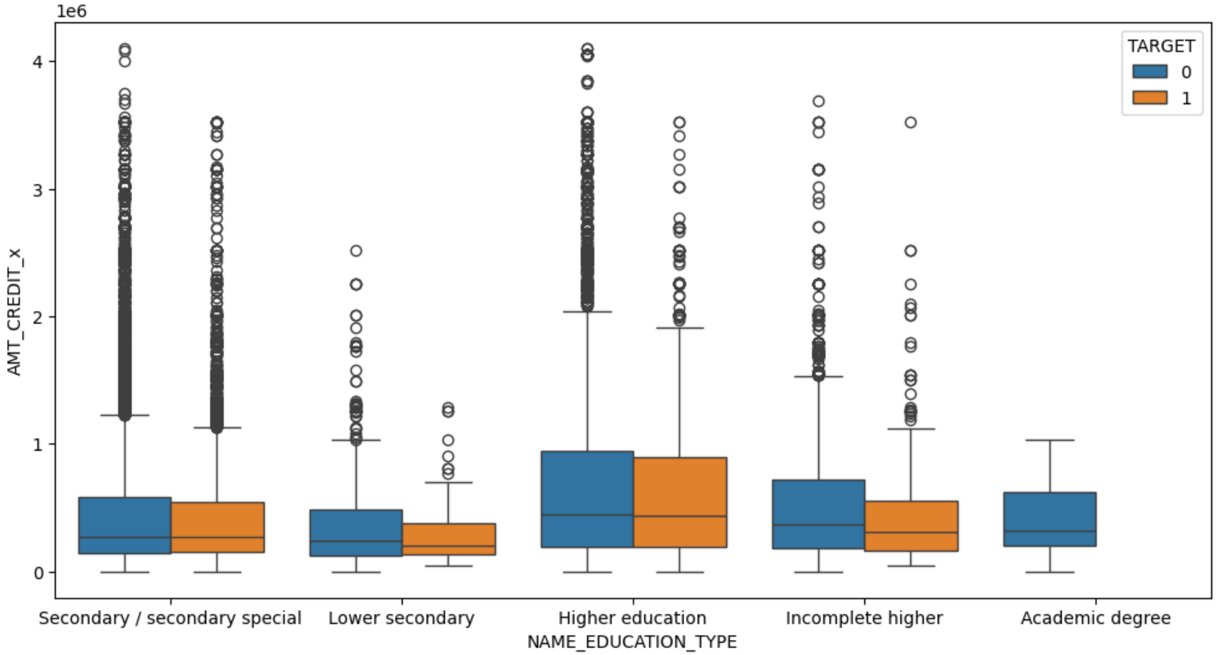


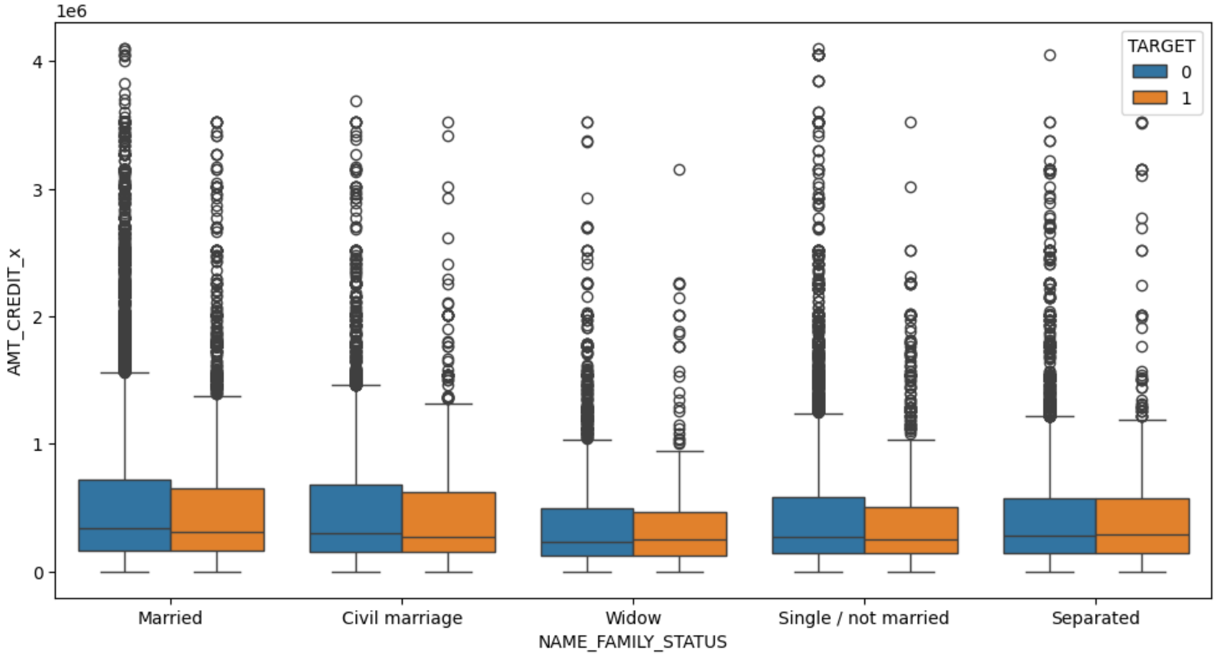
Distribution of contract status with purposes



Analysis on Rejection of loan

Analysis on Rejection of loan





Analysis on Rejection of loan

Points From The Analysis

- The data was highly imbalanced , around 92% of the clients pay the loan on time, only 8% of them face difficulties in paying the loan back.
- Females most often apply for the loan and the cash loans are highest loans taken so far.
- Working profession people most often become defaulters ,followed by commercial associate
- Education ,when considered we got to know ,clients with high education to pay the loan amount on time when compared to secondary .
- Next was the income range , we got to know that majority clients who apply for the loan lie between 2-4 lakh.
- The clients who have taken more than Rs.900000 of credit amount tend to become defaulters.

Points From The Analysis

- Then we saw that , clients who are married and have a greater number of children tend to become defaulters.
- Later ,we got to know students and businessmen tend to pay loan on time when rest up others are almost in equal ratio , but clients in maternity leave tend to become more defaulters when compared to others.
- At last, we saw the count of various applicants from different organization, where business type 3 were highest applicant of loans.
- Next up , we understand previous loan applicant’s details.
- We saw details like how many of the of the loan applications were approved and then the reason for rejection and type of the loan applicants whether the person is a repeater , new.

Points From The Analysis

- Then, Once the two data were merged, we tried understood what was the purpose of the loan and status of the loan.
- The majority were xna or xpa which in terms of data is unknown, so we removed those variables and got to know majority of the loan was for the repairs followed by others.
- Previously the rejection of the loan was done when applicants requested for a high amount of credit.
- The married clients usually tend to take high loan amount and when education is considered people with higher education tend to take high loan amount.

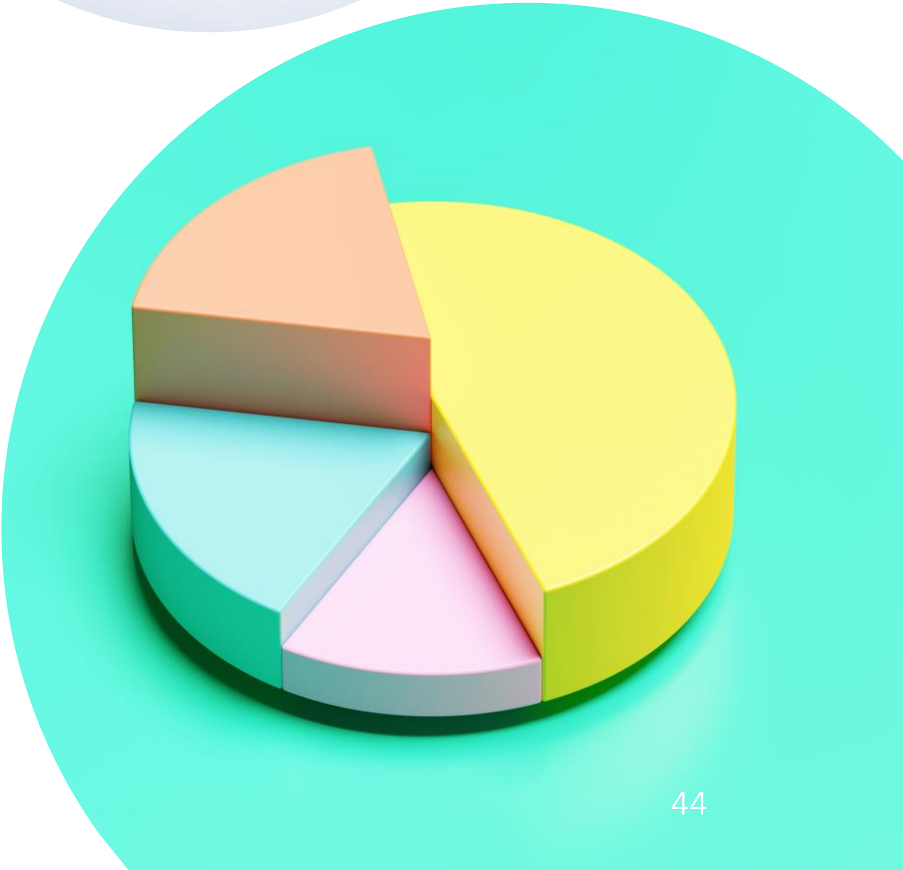
Summary

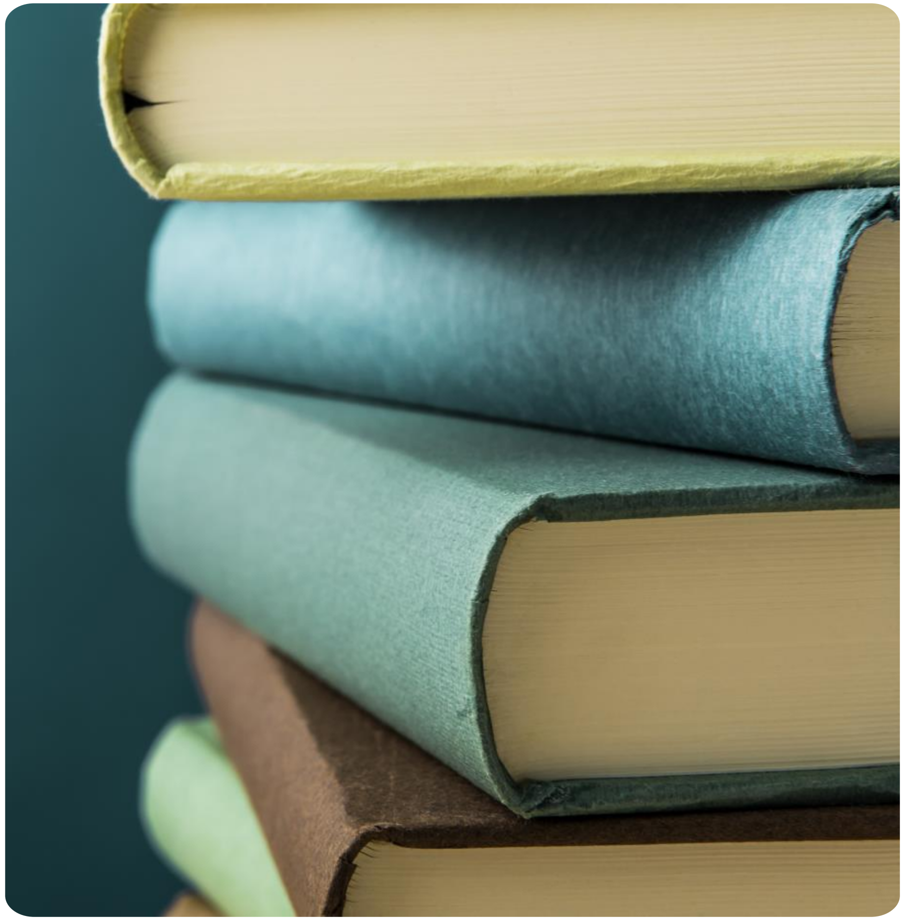
We've gained valuable insights from this data, allowing us to assess risk effectively.

By analyzing various parameters such as income groups, gender, and occupation, we've identified patterns that distinguish high-risk from low-risk individuals.

For more complex analysis we can consider many other aspects like the rating of the are and other factors which will increase the complexity As of now, these insights offer a solid foundation for understanding and addressing risk in the given context.

With access to more extensive time-series multiple data, we could enhance our models for even better risk mitigation.





Thank you

Sumanth.Ashok Metimath
sumanthmeti331@gmail.com