CREDIT EDA CASE STUDY

By :

KAUSTUBH BUTTAN

SHWETA PATIL

Business Objectives

- To use EDA to understand how consumer attributes and loan attributes influence the tendency of default.
- To understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment

Data Exploration

The data is made available in the form of below mentioned excel files:

- *'application_data.csv'* contains all the information of the client at the time of application.
  The data is about whether a **client has payment difficulties**
- *'previous_application.csv'* 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.**

Application Data Analysis

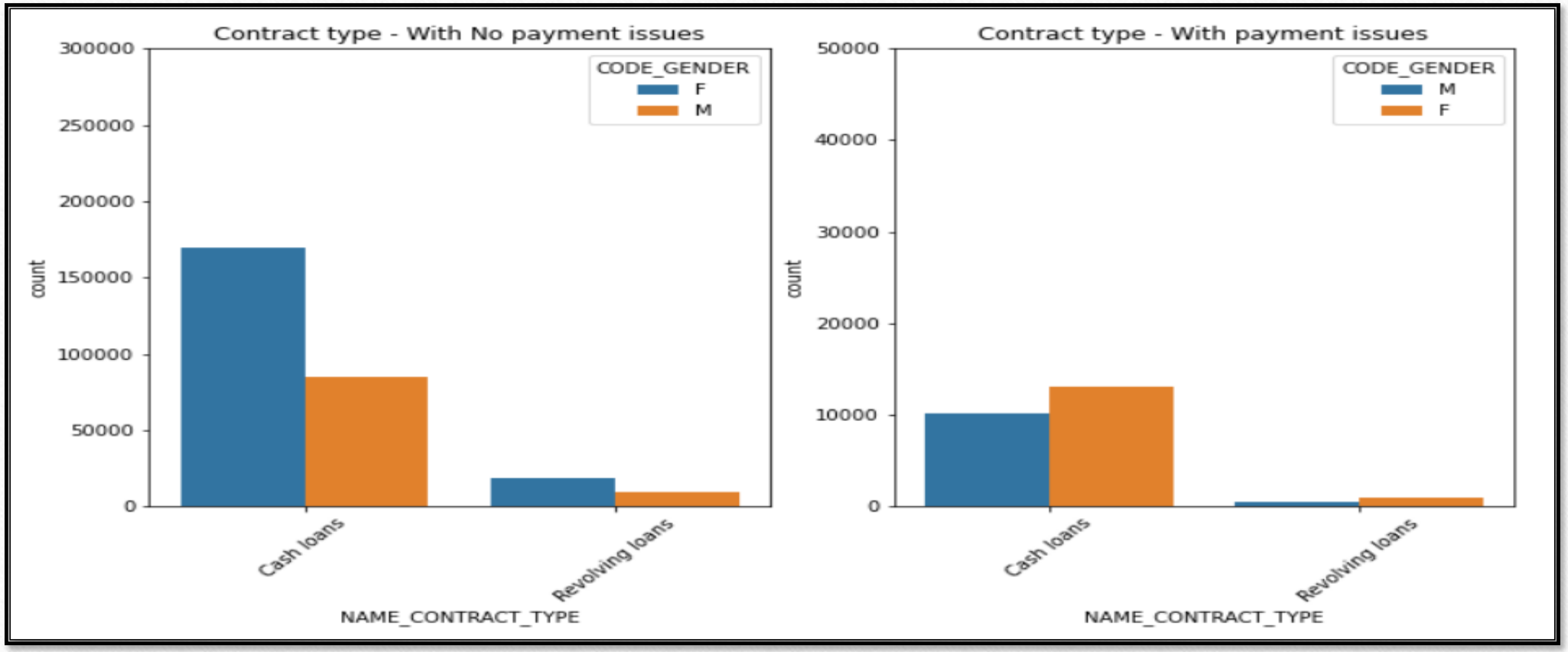
Data Cleaning and Manipulation

Dealing with missing values :

- Columns with more than 40% of values as null values are dropped.
- Unwanted columns are removed.
- The null values of certain columns are replaced with either median or mode.

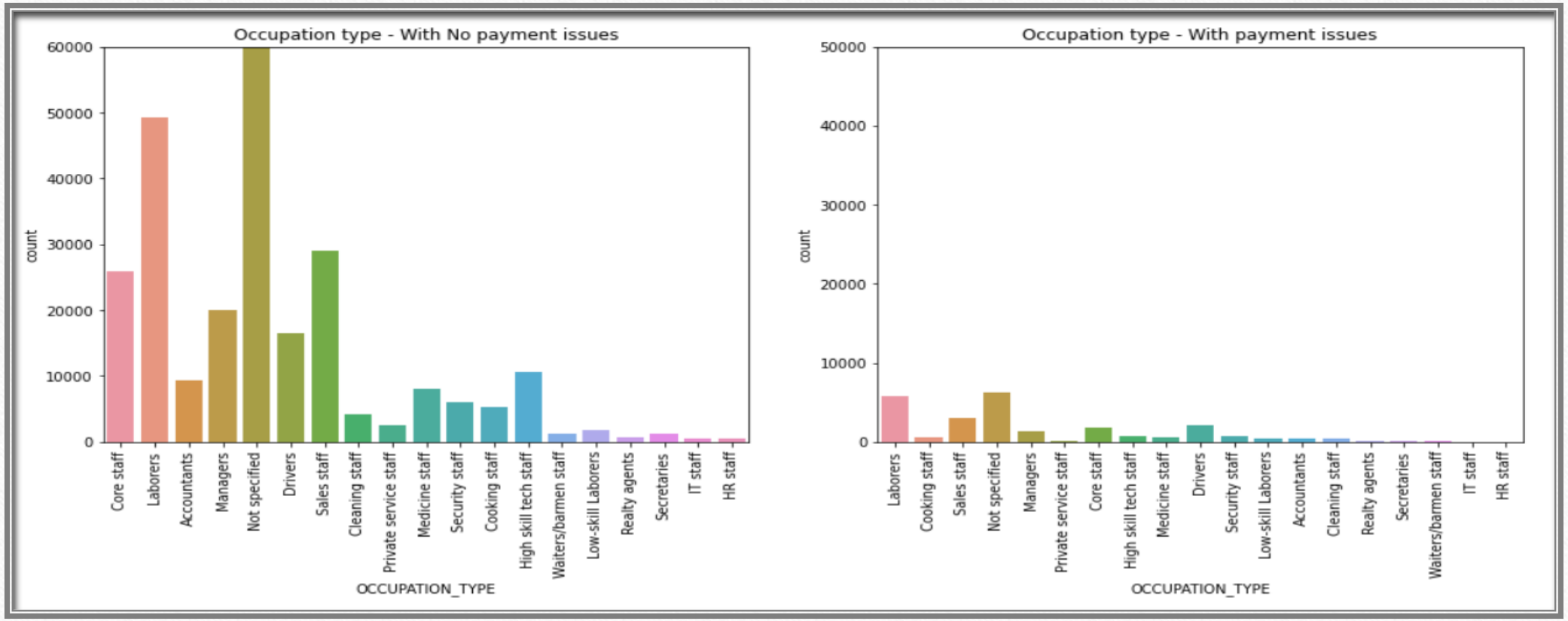
Others :

- Columns having unique values more than 40 can be considered as continuous columns where as columns with unique values less than 40 can be considered as categorical columns for analysis. Also, delete the columns with unique value as '1' as it is insignificant for analysis
- The negative values of certain columns are changed into positives.
- Outliers are treated using techniques like Imputation, Deletion of outliers, Binning of values, Cap the outlier



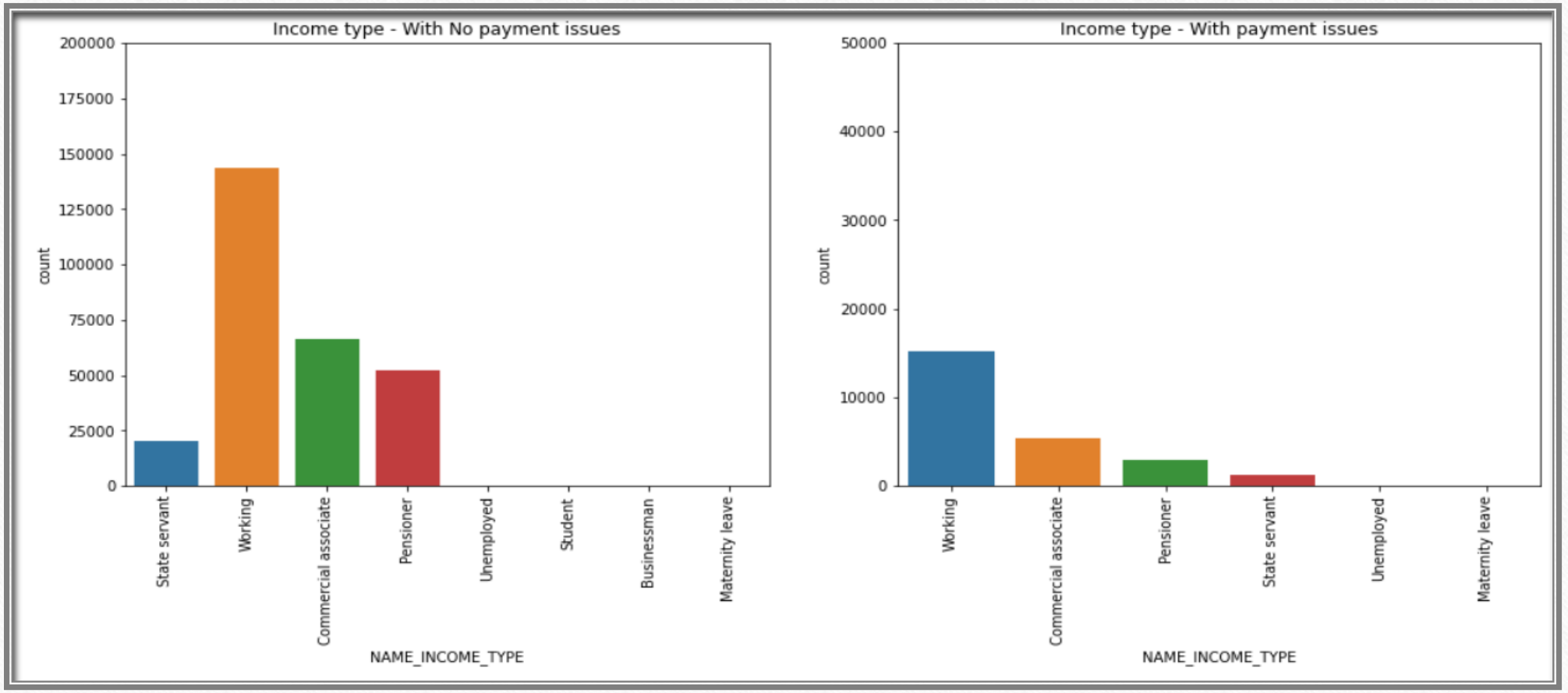
Categorical variable analysis for given CONTRACT TYPE

- It can be observed that people with Cash Loan are facing more trouble than Revolving Loan Customers.



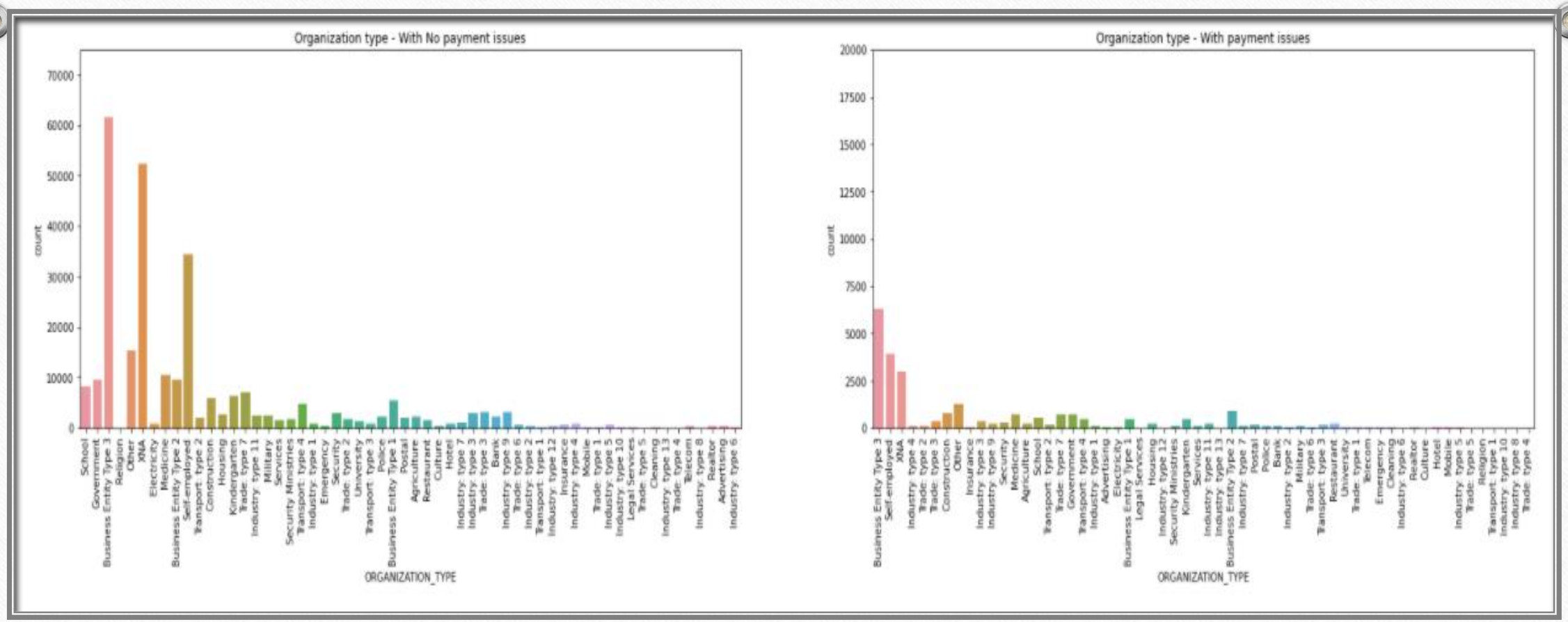
Categorical variable analysis for given OCCUPATION TYPE

- Ignoring the 'not specified' category as these people have not mentioned their occupation type. Leaving apart 'not specified' category, we can see that maximum loan applications received are from laborers.



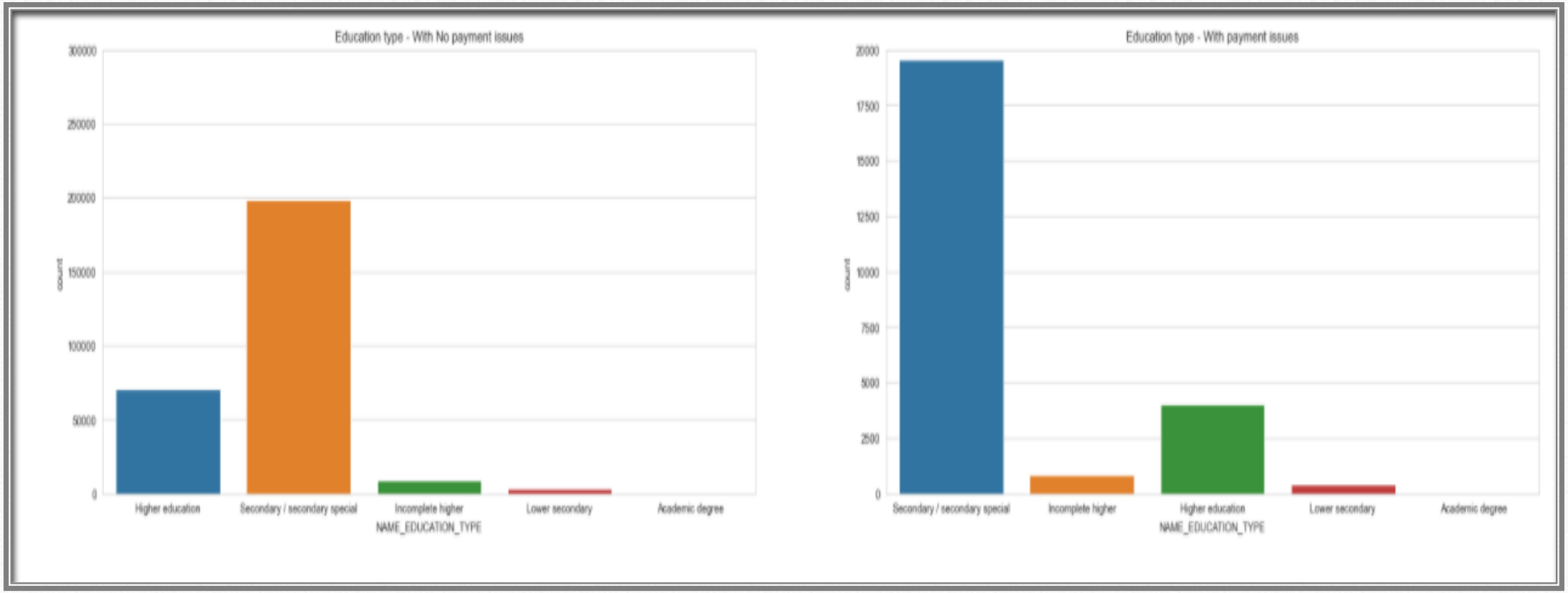
Categorical variable analysis for the given INCOME TYPE

- It can be concluded from the above figure that the Maximum number of application are received by Working class people



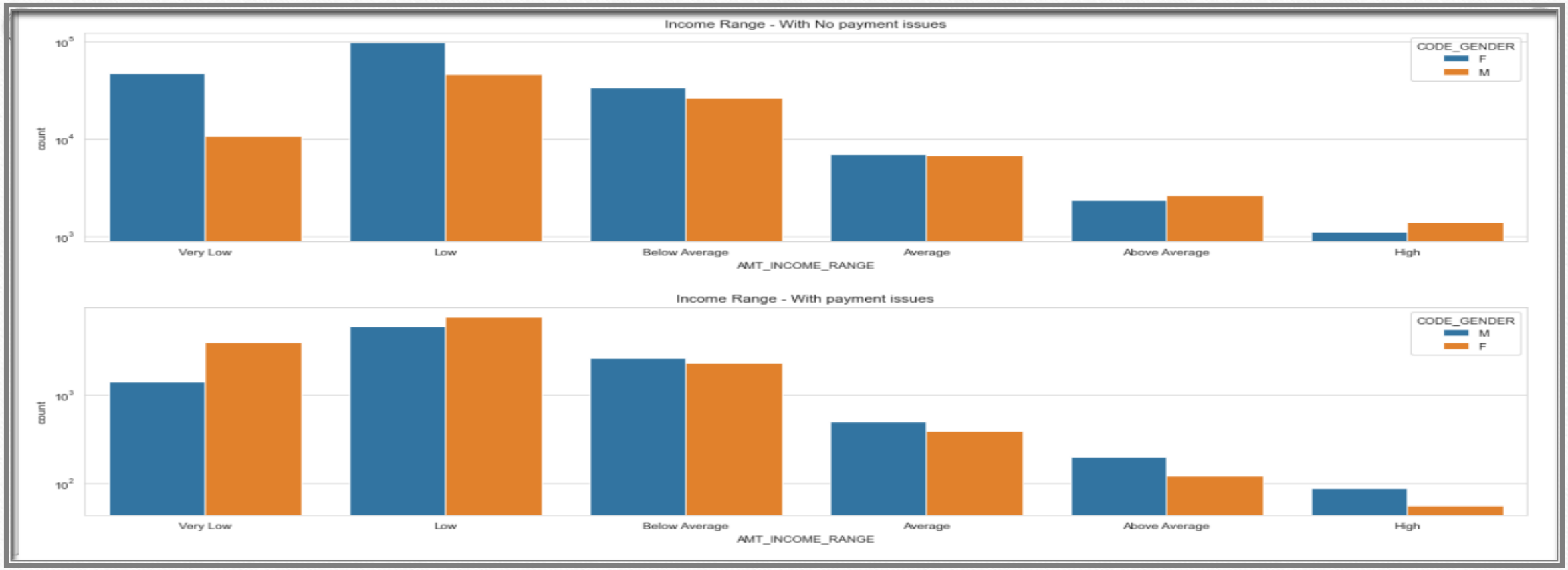
Categorical variable analysis for given ORGANIZATION TYPE

- We can infer from the graph that maximum number of application are received by people from organization - Business entity type 3



Categorical variable analysis for given EDUCATION TYPE

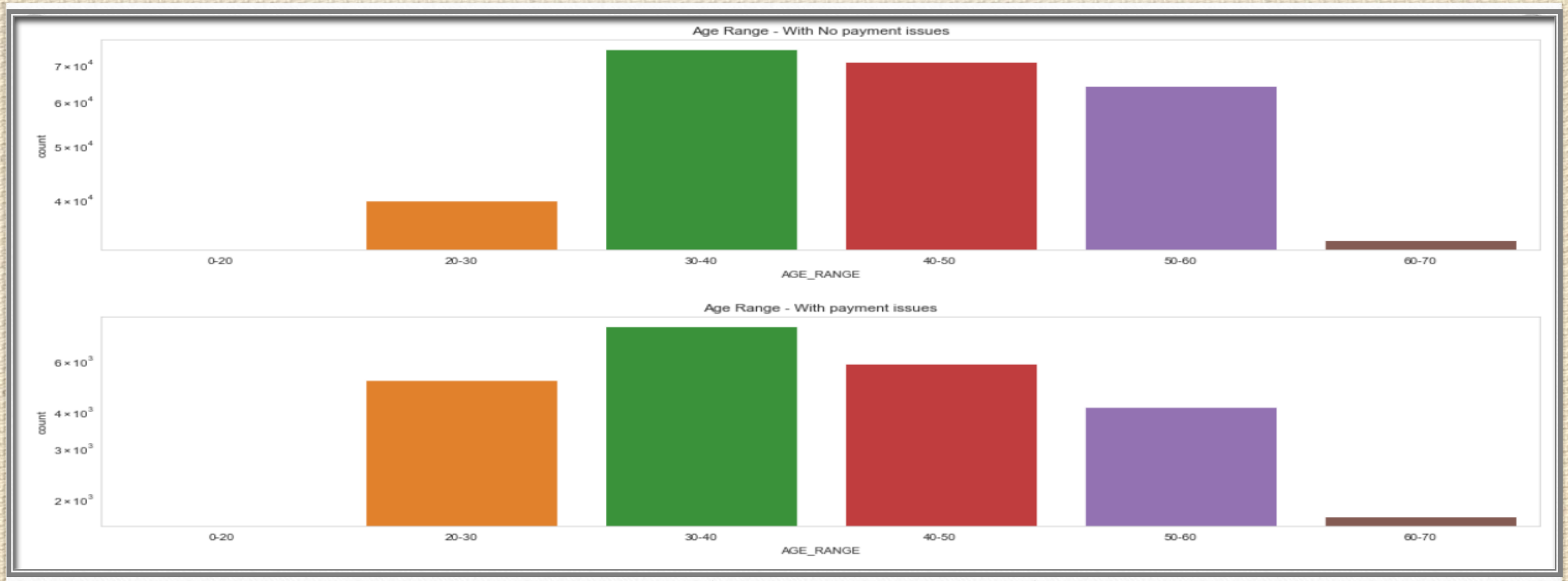
- We can infer from the graph that People with Secondary / Secondary special education are majorly applying for loans



Univariate analysis of target0 and target1 group around Income range

It can be inferred clearly that :

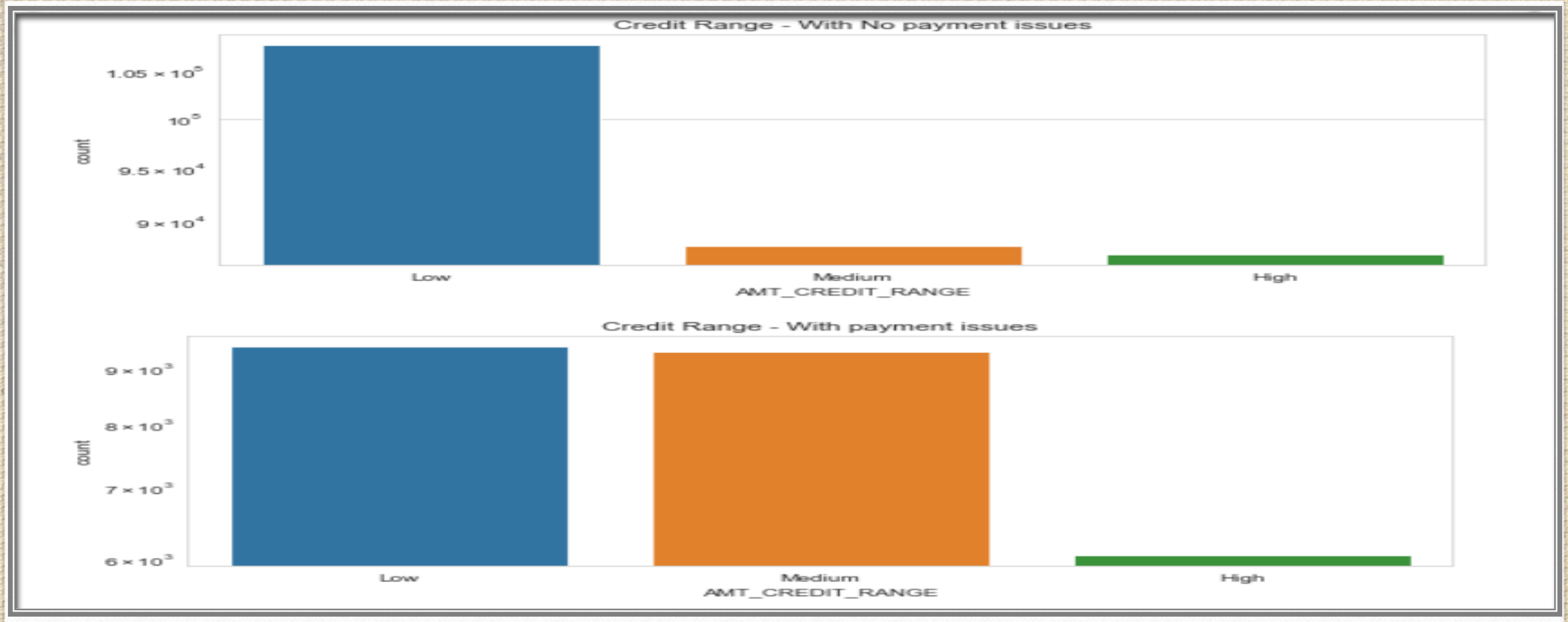
- Female count is higher than Males
- Maximum people seeking for loan lies in the Very Low to Average income slabs.
- Very few applications are received from people in Higher income slab



Analysis of target0 and target1 group around Age range

It can be inferred from the figure :

- People from the age group of 30-60 are mostly having no issues in loan repayments
- It should be observed that more number of people in the age group 20-30 are facing issues with loan repayments

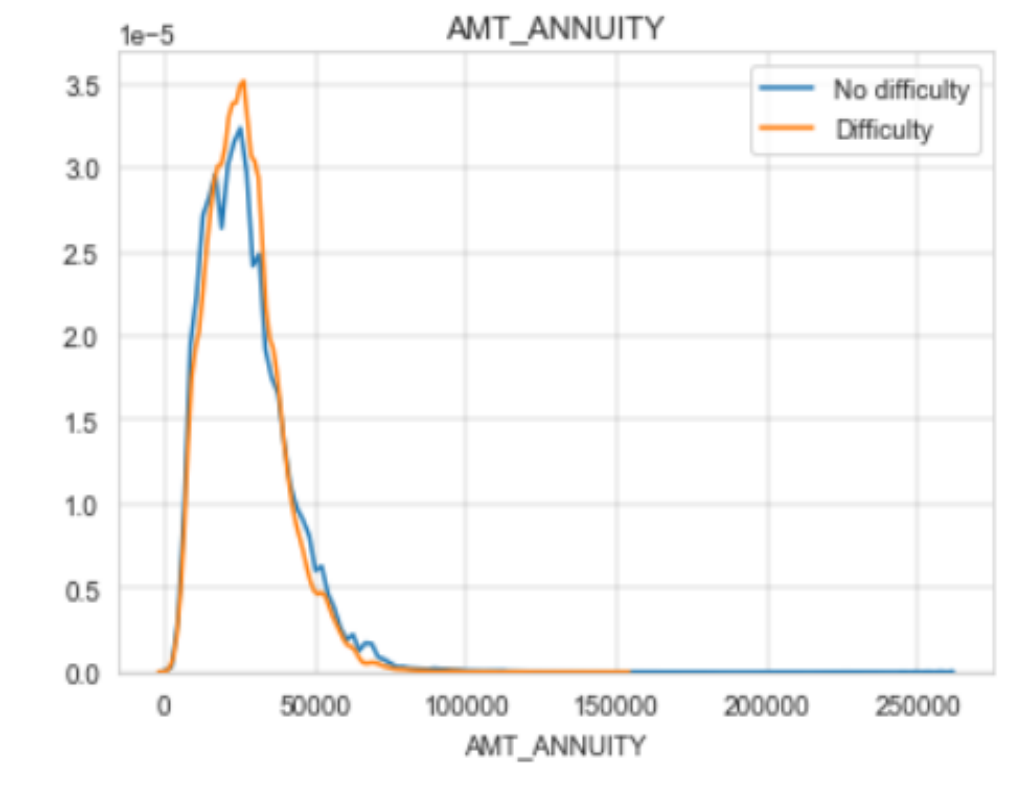


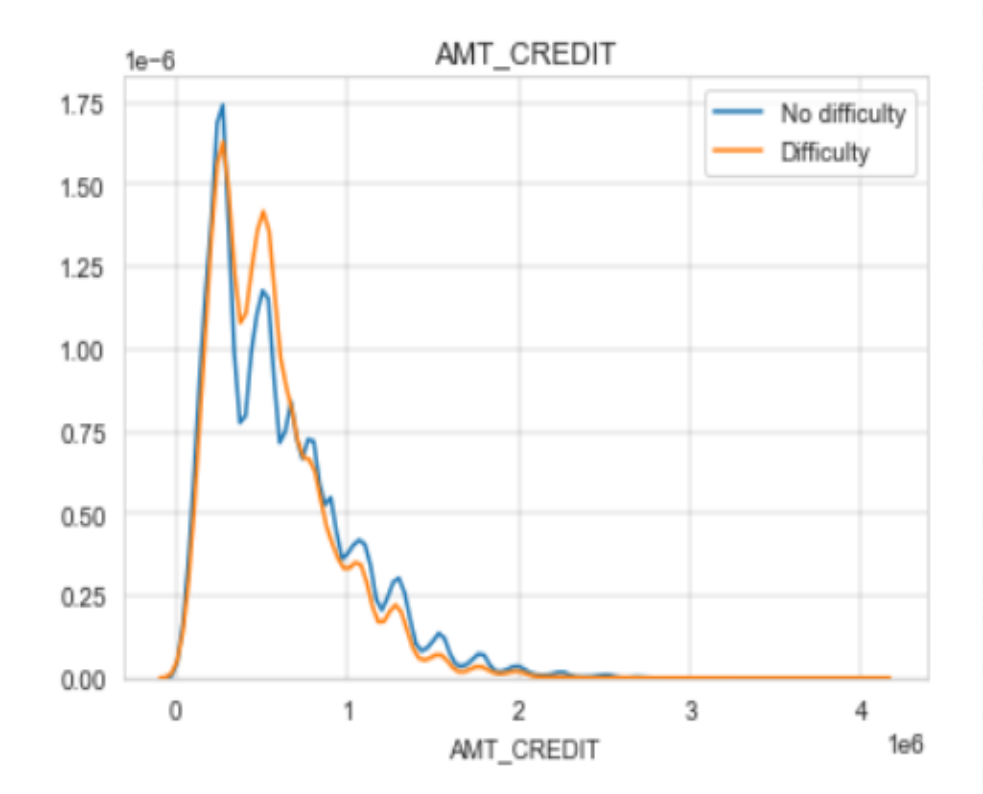
Analysis of target0 and target1 group around Credit range

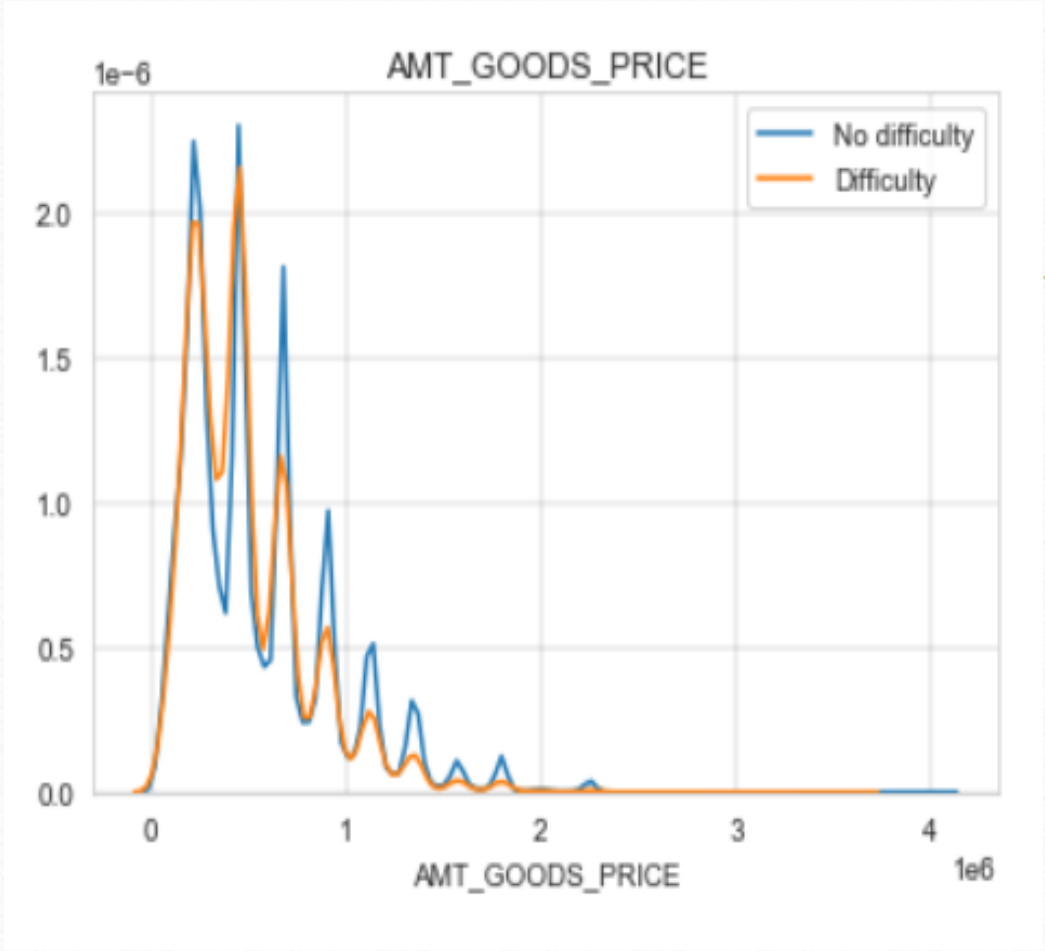
The inferences drawn are :

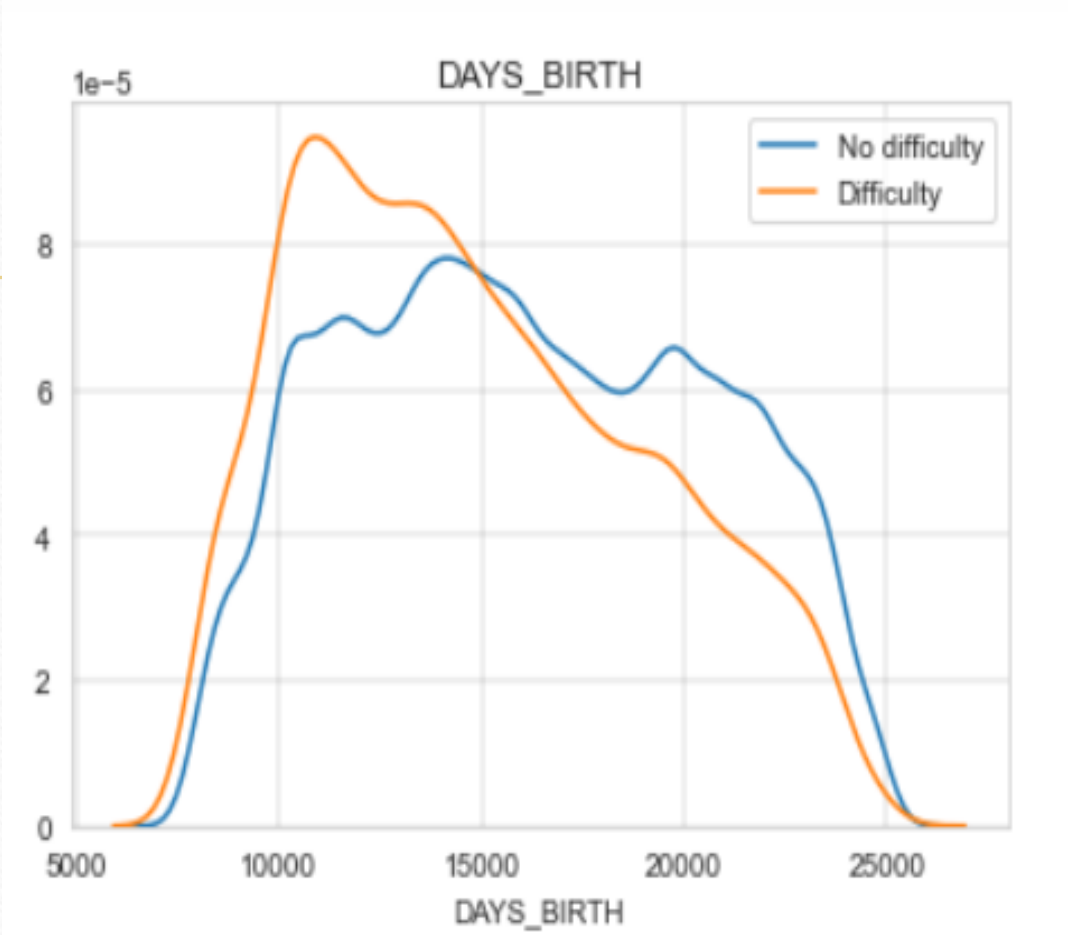
- Majority of the loans given are Low credit loans
- People with low credit loans and medium credit loans are struggling more for repayments

Analyzing continuous columns
(AMT_ANNUITY','AMT_CREDIT','AMT_GOODS_PRICE','DAYS_BIRTH','DAYS_EMPLOYED)






AMT_GOODS_PRICE
1e−6
No difficulty
Difficulty
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
1e6
AMT_GOODS_PRICE

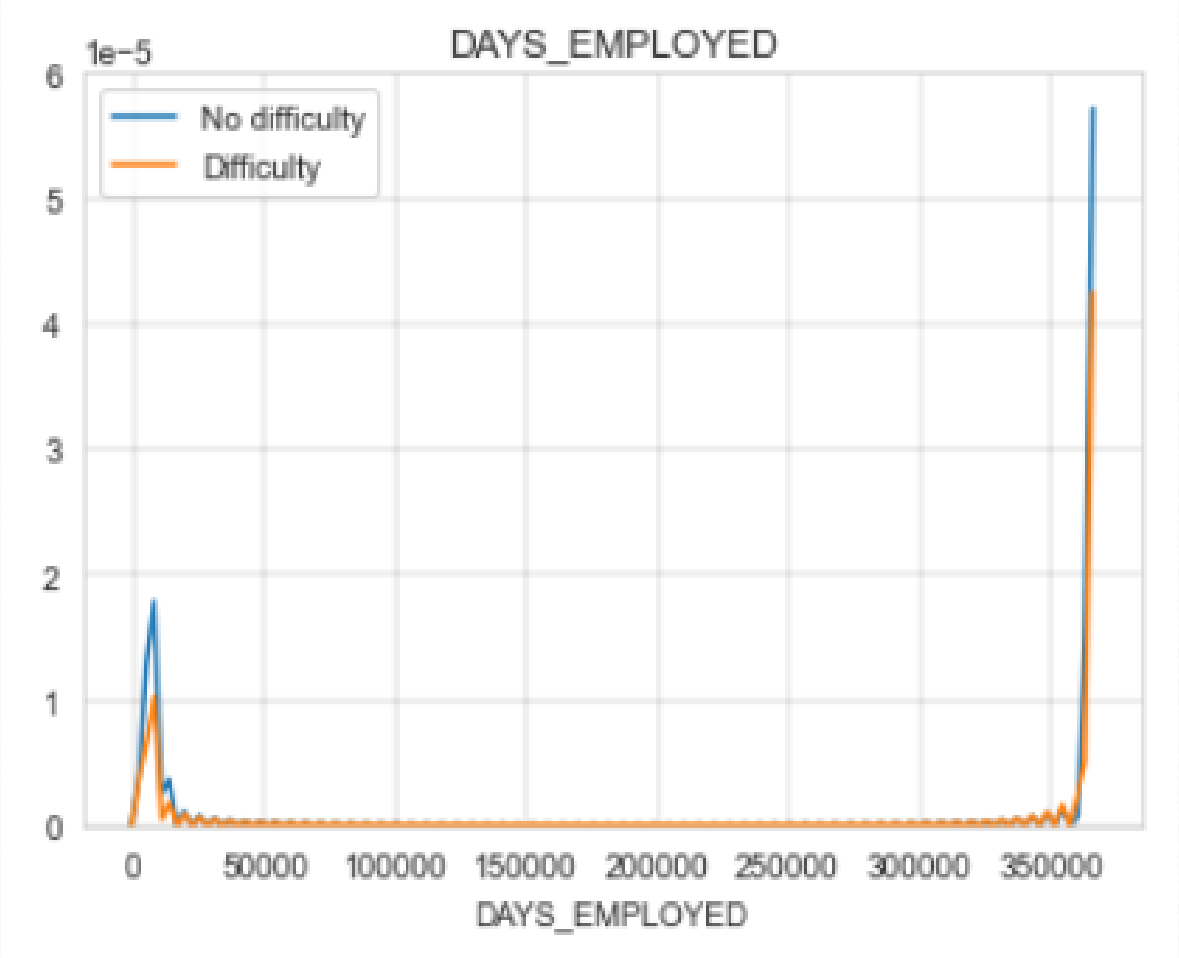

DAYS_BIRTH
1e−5
No difficulty
Difficulty
8
6
4
2
0
5000
10000
15000
20000
25000
DAYS_BIRTH

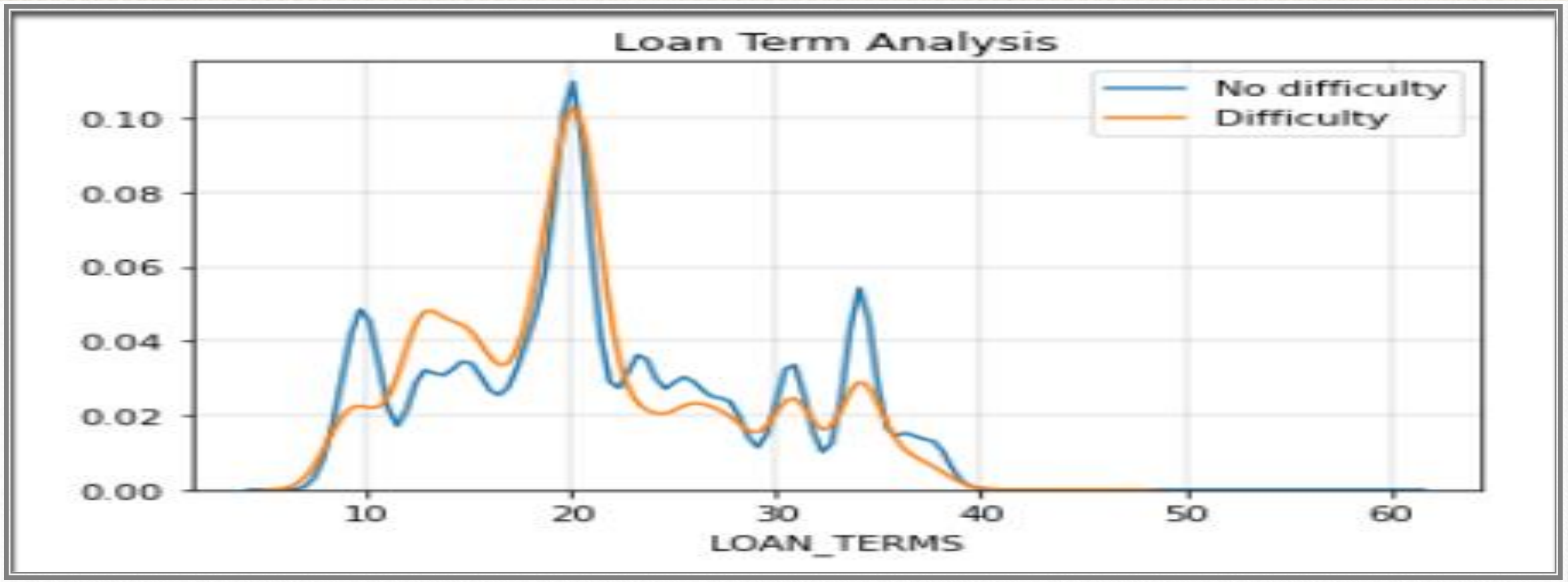
Observation:

- People having difficulty in loan payments, have lower annuity amounts
- Many Lower amount loans are given, which might have been given to less well off people and hence they are struggling with their repayments

---

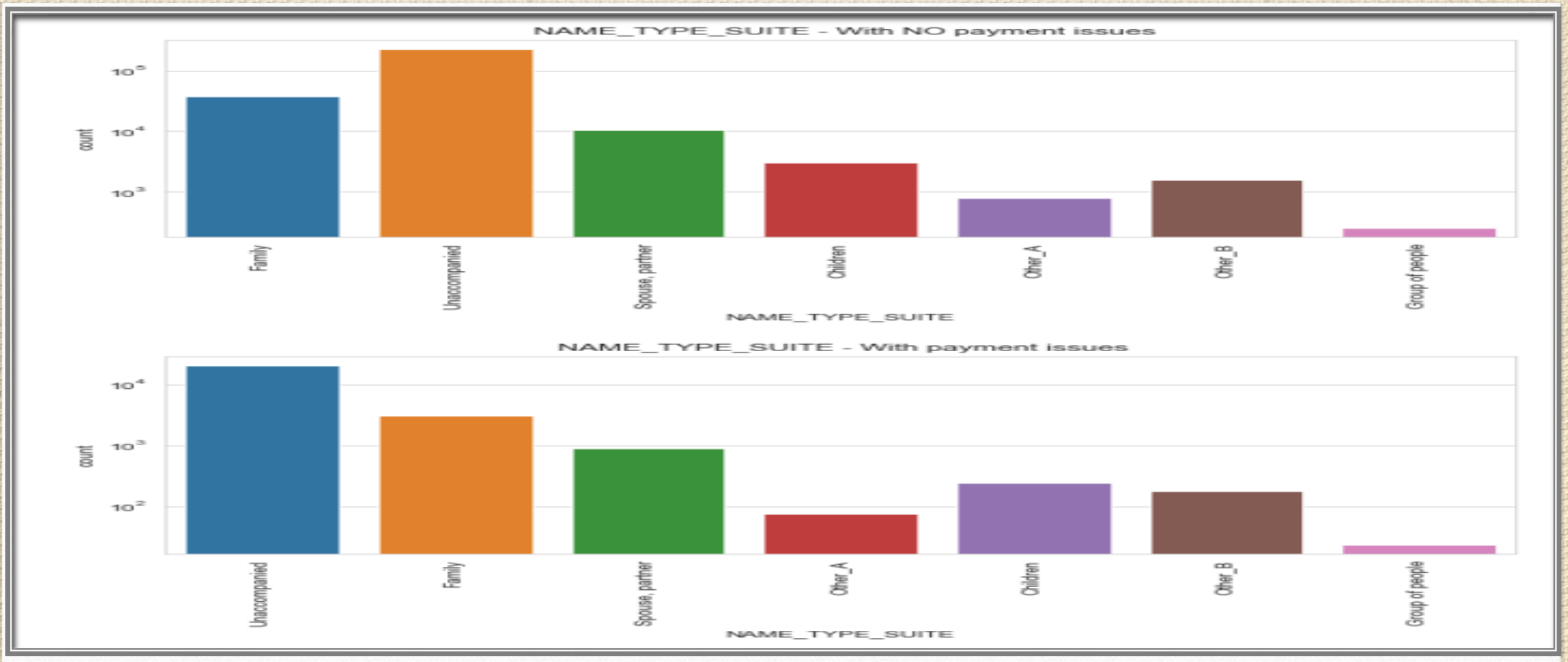
- Younger people are struggling more for loan repayment which is obvious as financial stability comes with more professional experience and age
- When people are employed for less time, they seem to be struggling for loan repayment





Loan Term Analysis

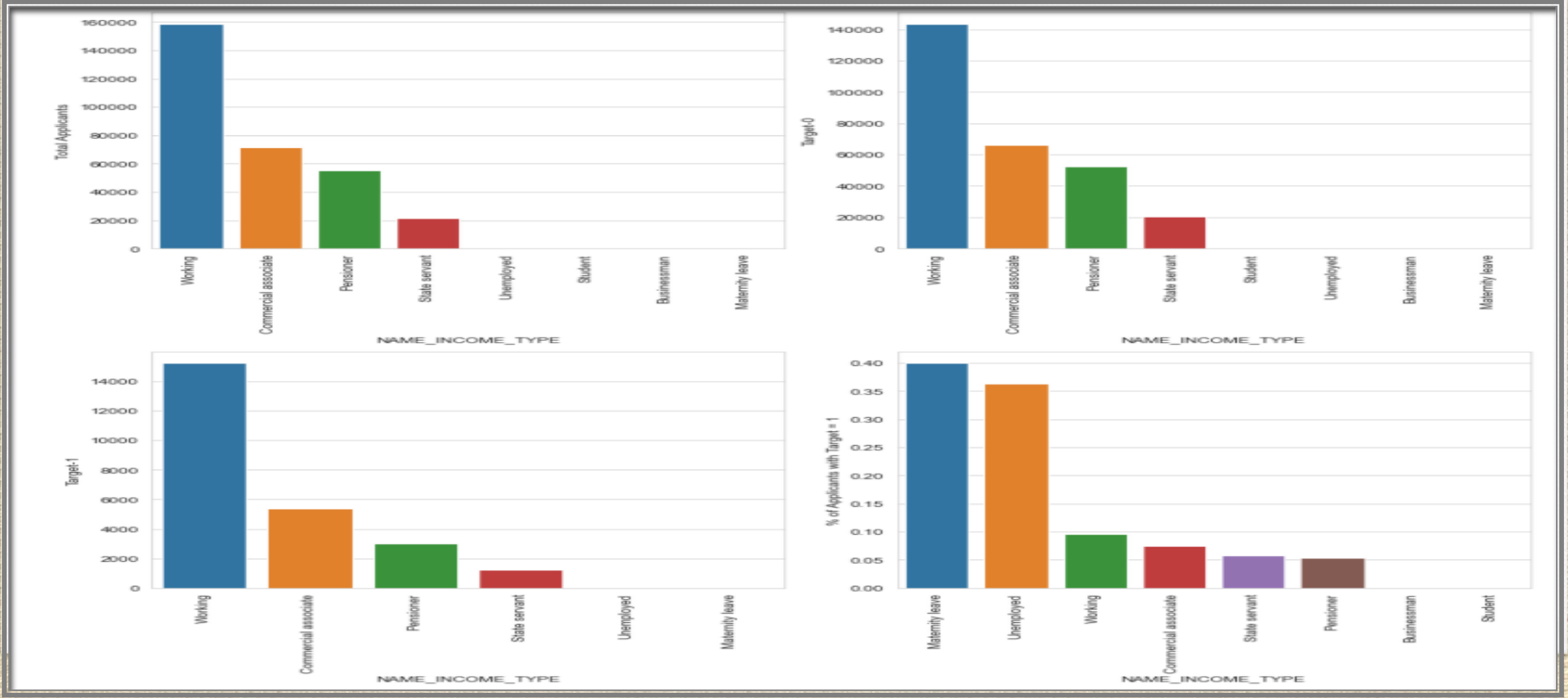
- People with Loan Terms between 19-23 seem to struggle in their loan repayments. Banks can avoid this risk by lending loans for "more terms" or "less terms with more rate of interest.

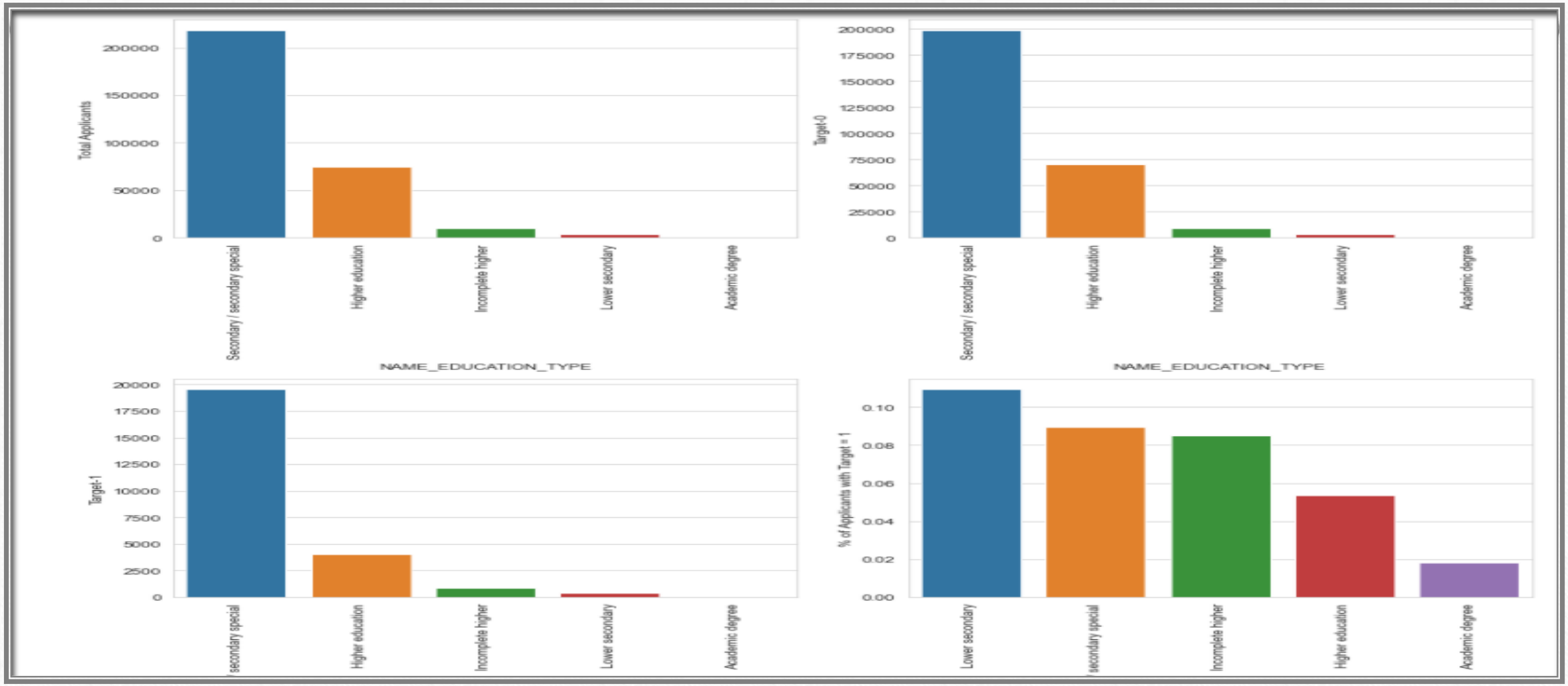


Payment issues with respect to different type of people.

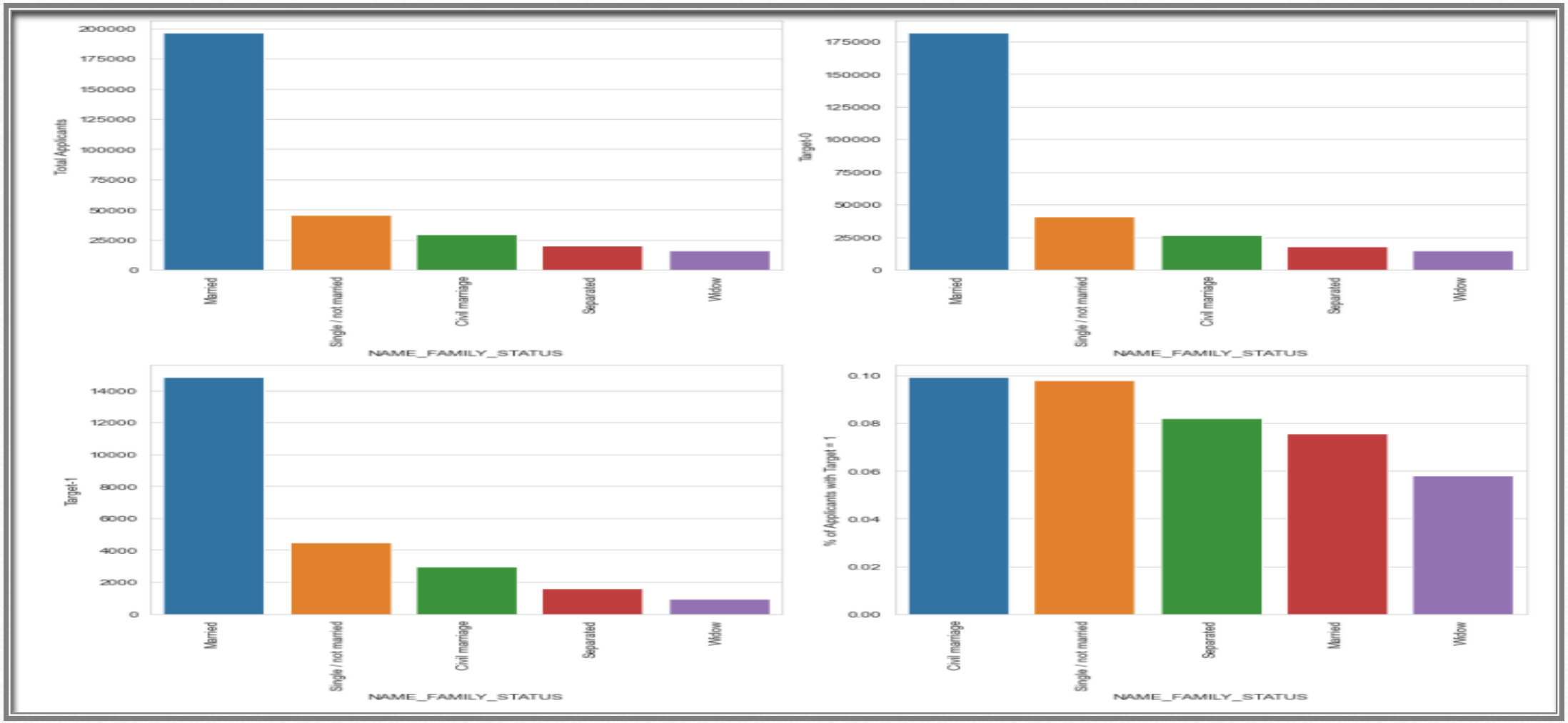
- We can infer from the given graph that Unaccompanied people have the most issues in loan repayment.

Taking into consideration target-0 and target-1analysis ,we can infer that there are very few loan applications from "Unemployed" and "Maternity leave" category or may be banks do not lend much loan to people belonging to this category because as we can see, the probability of people facing issues in loan repayment is very high (visible from 4th graph). Bank should examine closely before lending loans to people from "Unemployed" and "Maternity leave" category.



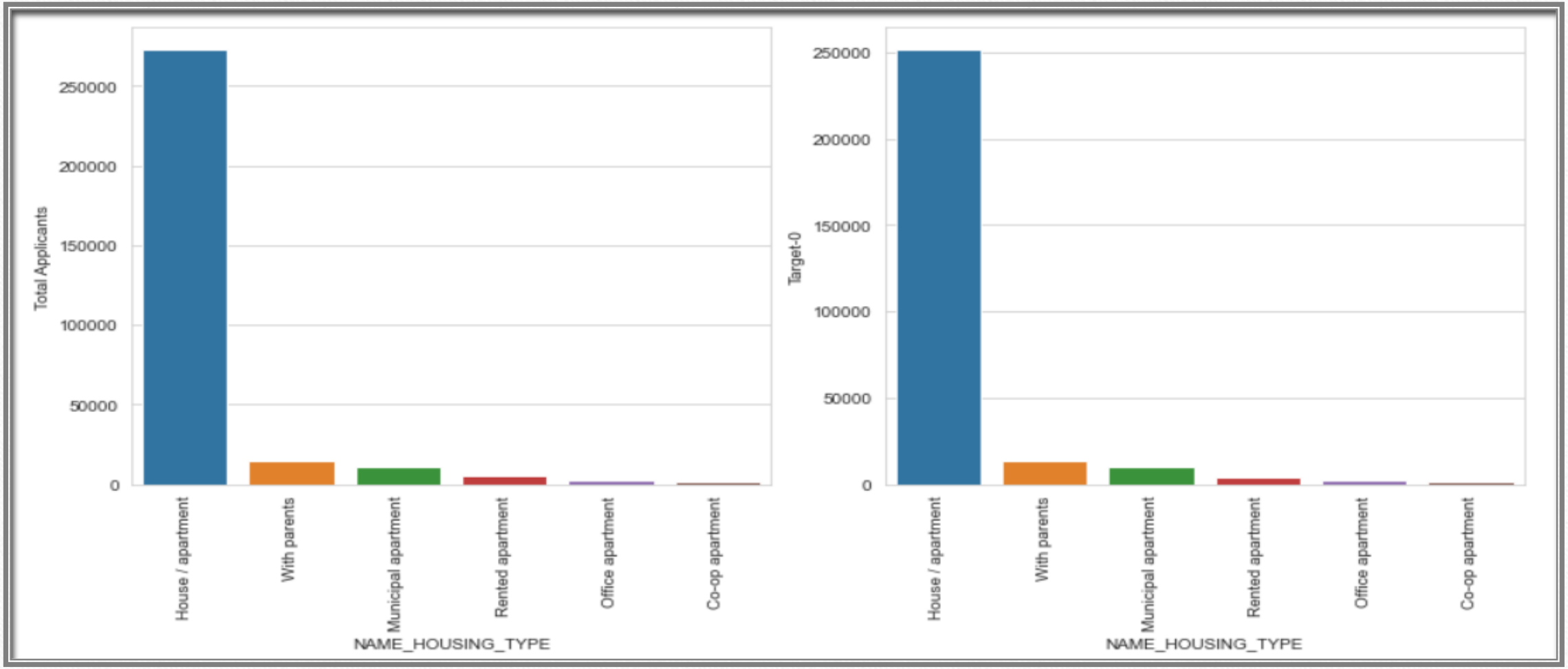


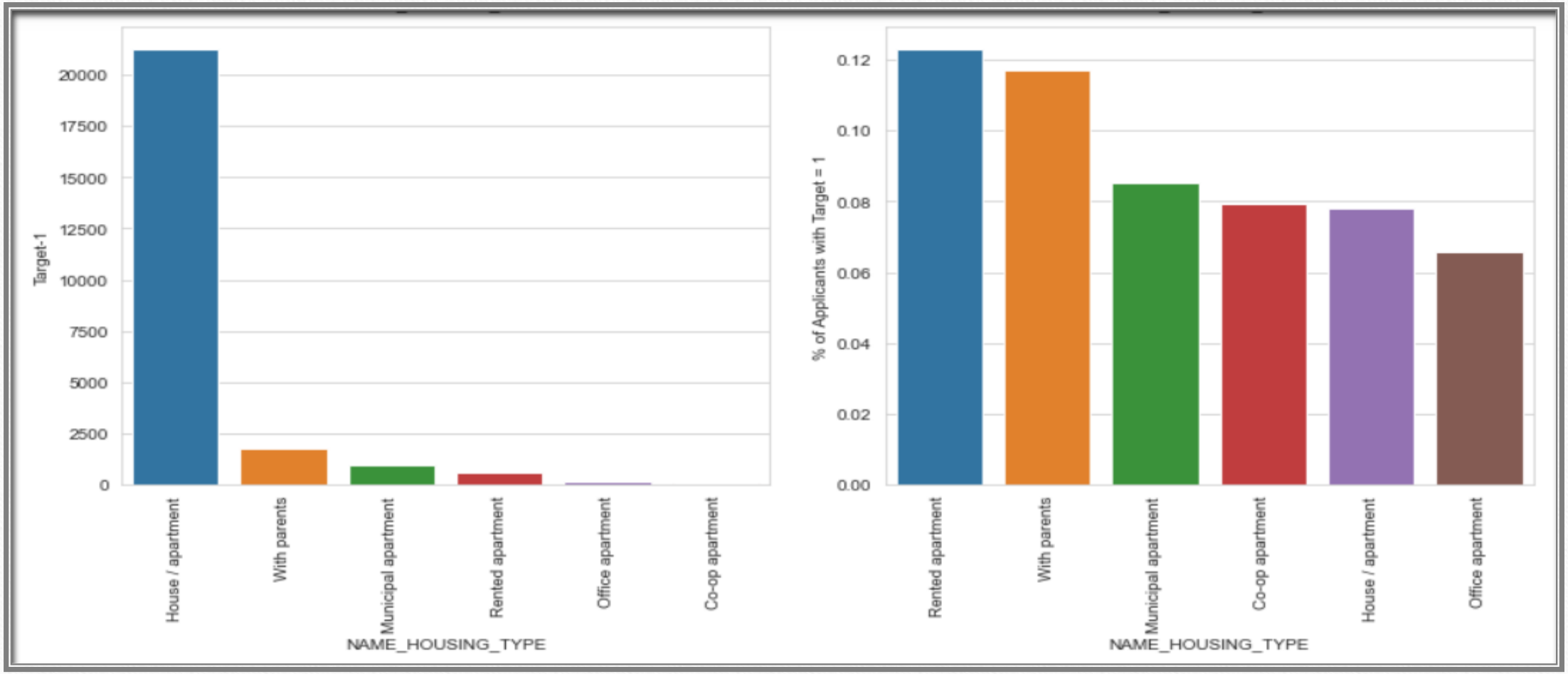
- Similar is the case for different education types where People with "Lower secondary" and "Secondary / secondary special" category are struggling for loan repayment. This seems obvious because mostly people with less educational qualification tend to receive lower income and thus might not be able to manage loan terms.



Now taking into consideration people according to the **Family Status ,** people from "Civil marriage" and "single / not married" category seem to face issues in loan repayment. Bank should scrutinize these people before lending loans

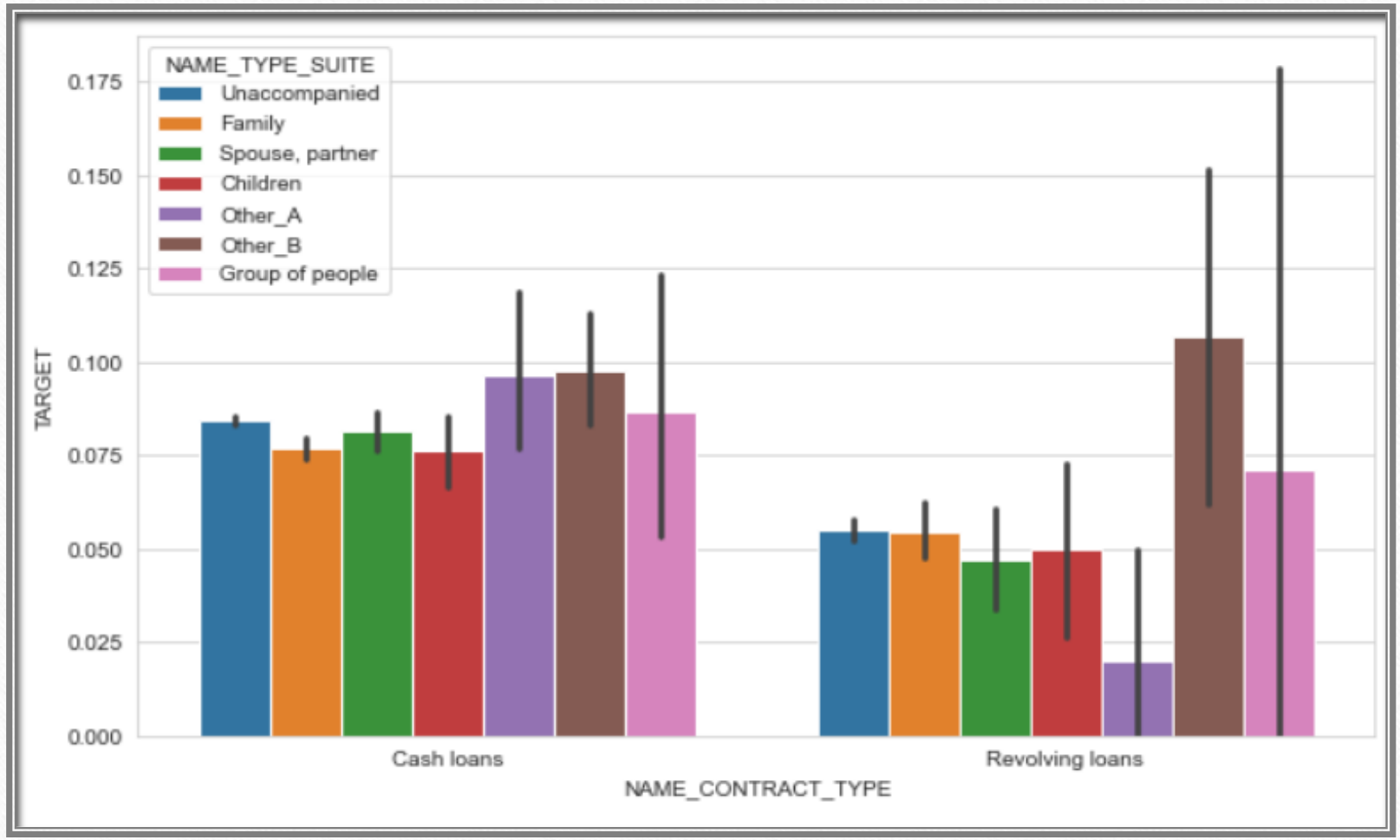
Analysis based on the **Housing Type** the applicants belong to :





- Based on the plots above and in the last slide we can infer that Applicants from "Rented apartment" and those from "With parents" category seems to default. This might be because person has to pay the rent of his house and loan term at the end of every month and this might be difficult to manage. On the other hand those who live with parents have might be having more household expenses as they have more family members and thus they might be facing issues while repaying loans

Bivariate Analysis

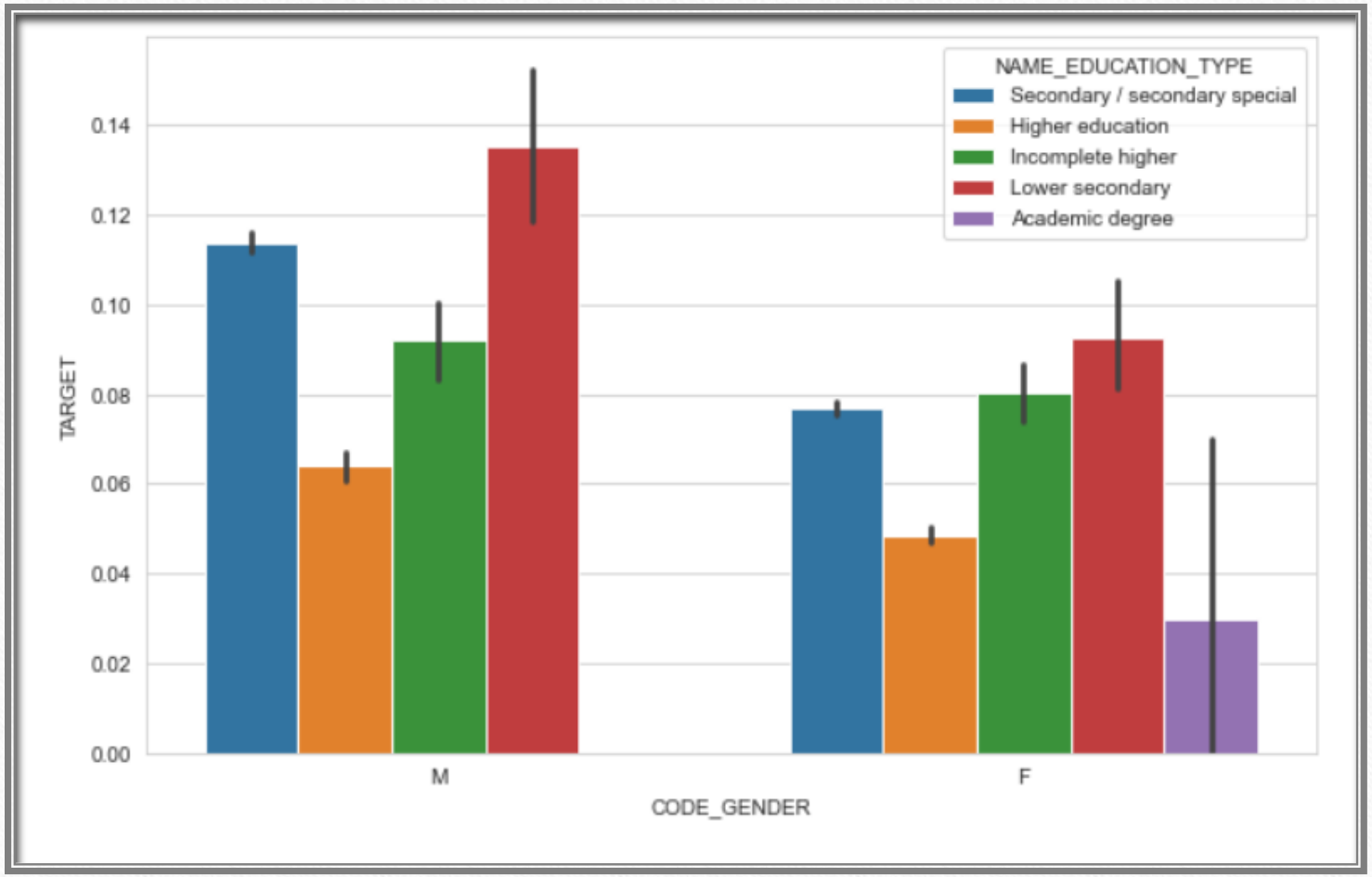


Analysis on the basis of given contract type

Observation:

1.Cash loans have more defaulters than revolving loans

2.Interestingly Other_B category in revolving loans has higher default value than Other_B category in cash loans



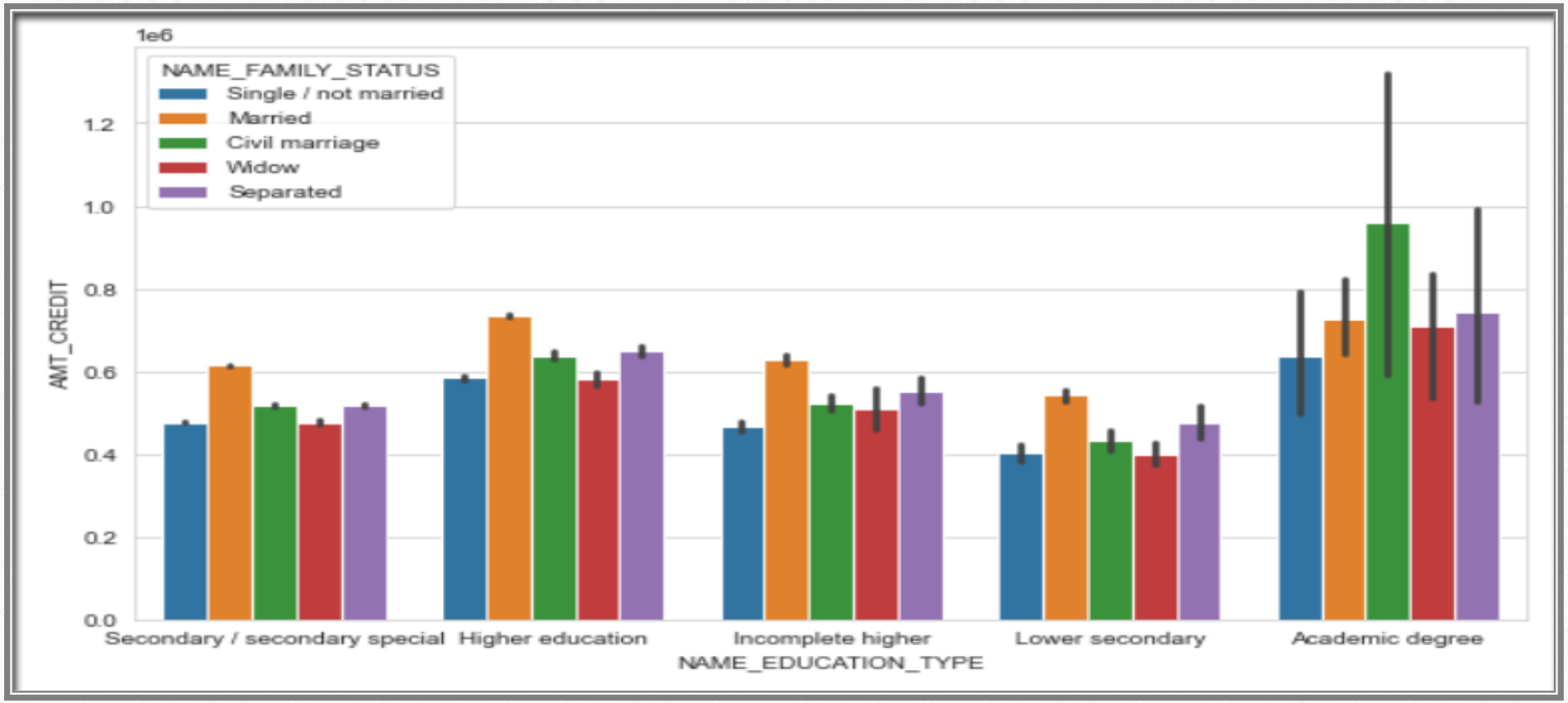
Analysis on the basis of Education Type over Gender category.

Observation:

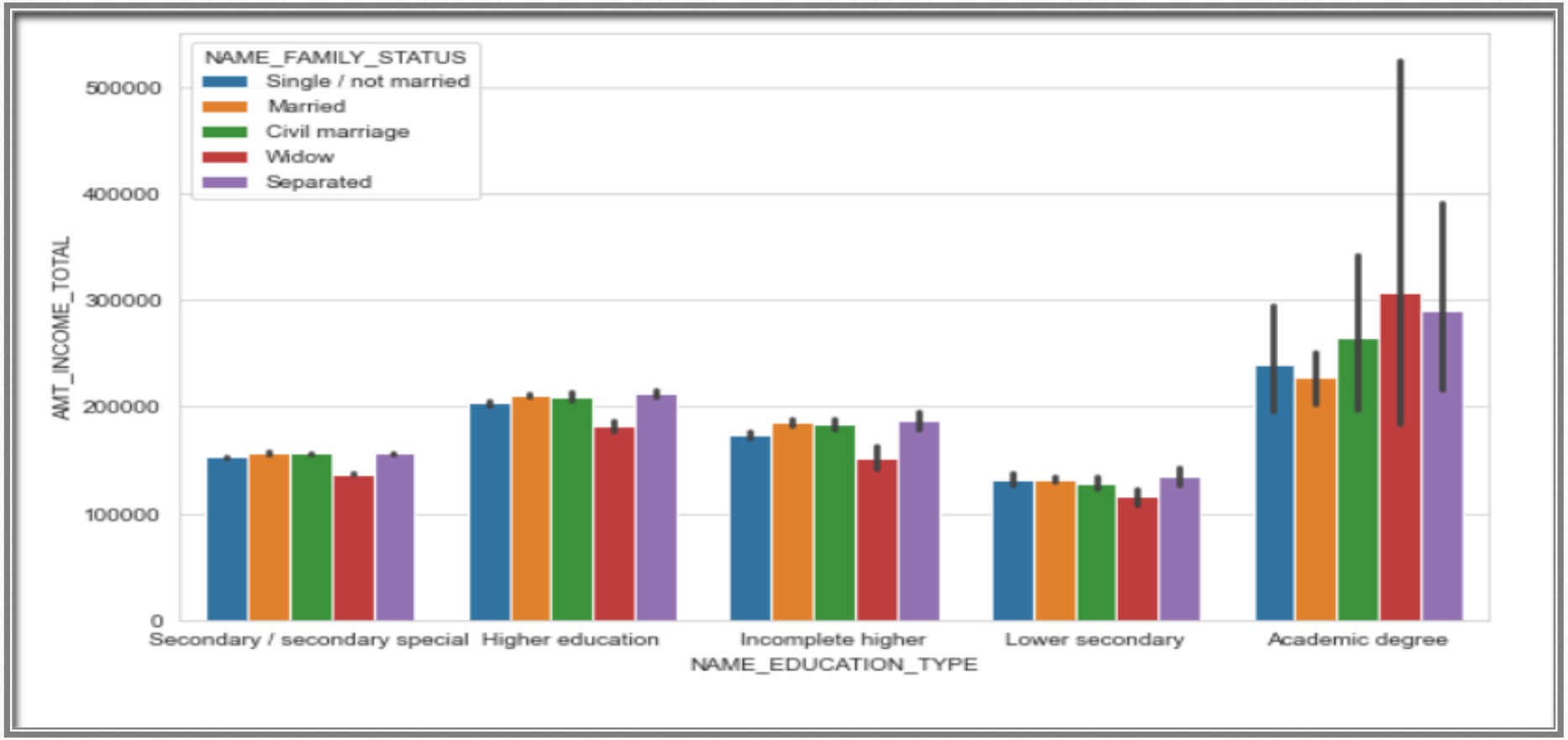
1.It confirms that Lower secondary and secondary/secondary special are more likely to default

2.Females with academic degree tend to default. This insight was not visible earlier and is important one to note here

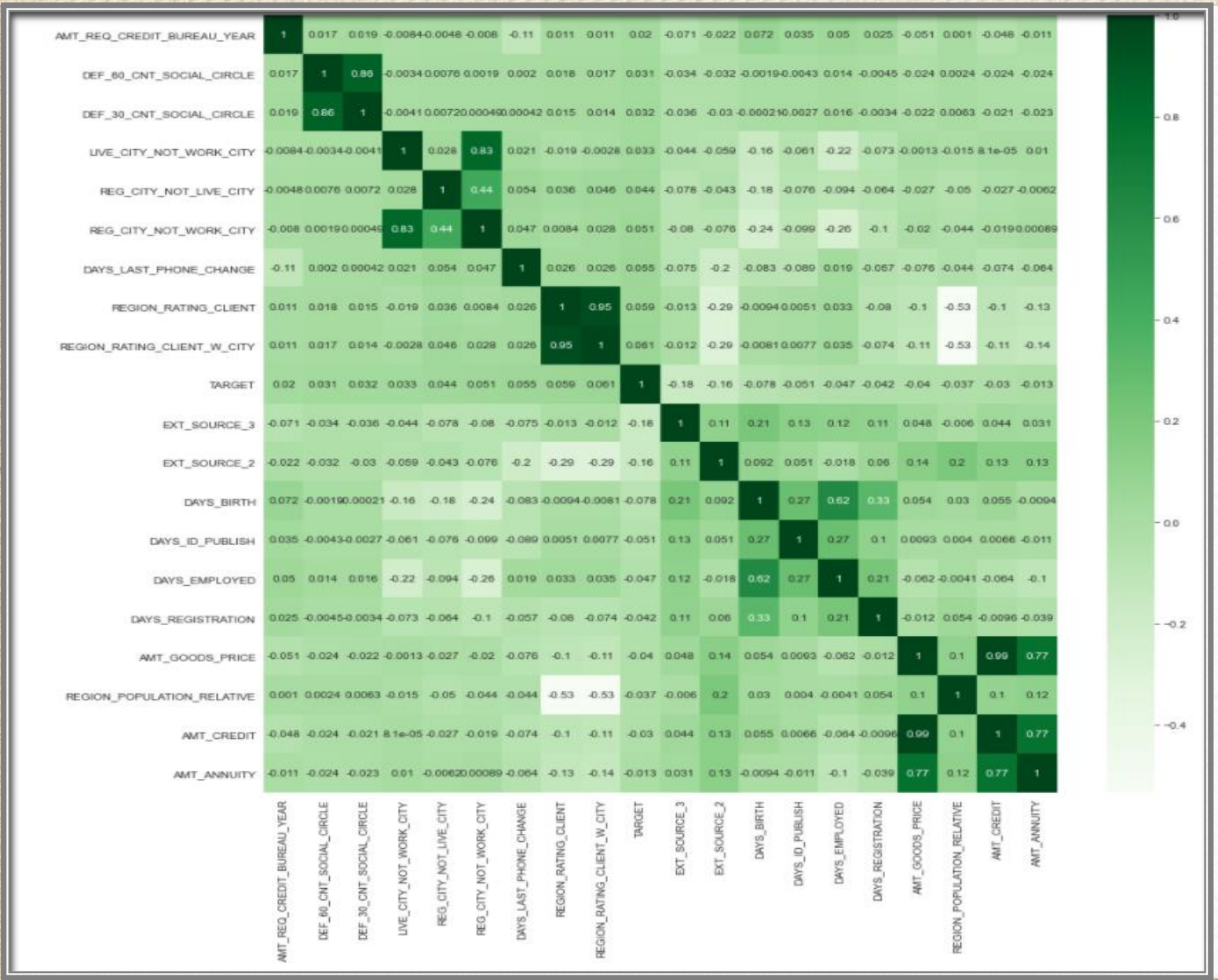
In [451]:



We can infer from the above graph that People with Academic degree tend to have higher amount credits irrespective of family status.



- It can be observed from the graph above that People with Academic degree tend to have higher income irrespective of family status



Generating a correlation heat-map from the Data frame

OBSERVATION :

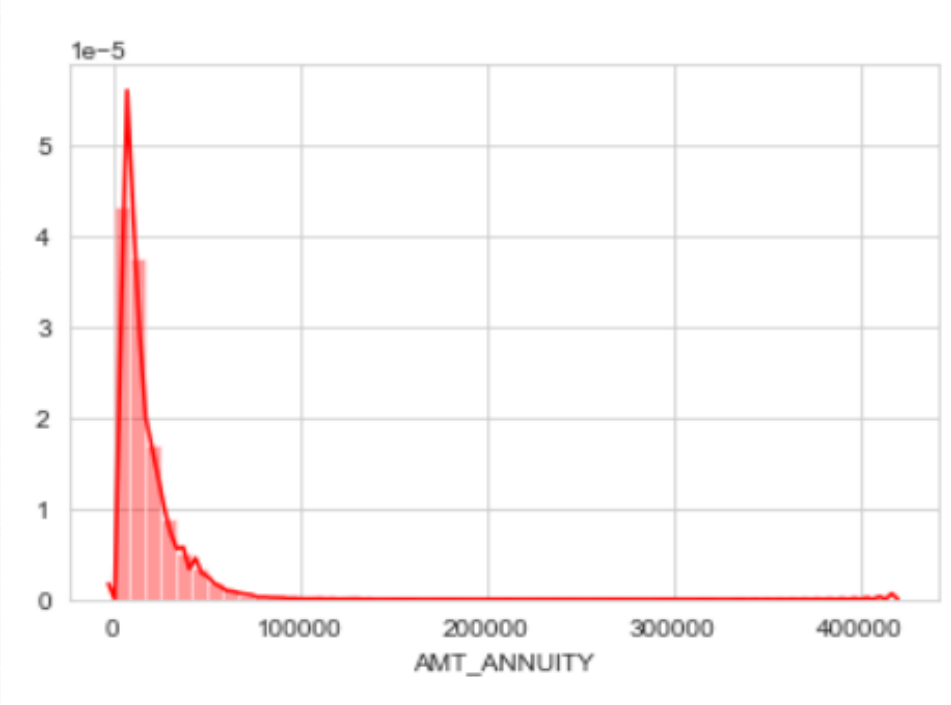
- DAYS_BIRTH and DAYS_EMPLOYED are highly co-related, which is true as people with higher age will have more professional experience
- High correlation between DEF_60_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE .
- AMT_GOODS_PRICE and AMT_CREDIT seem to have higher correlation
- REGION_POPULATION_RELATIVE and REGION_RATING_CLIENT have negative correlation

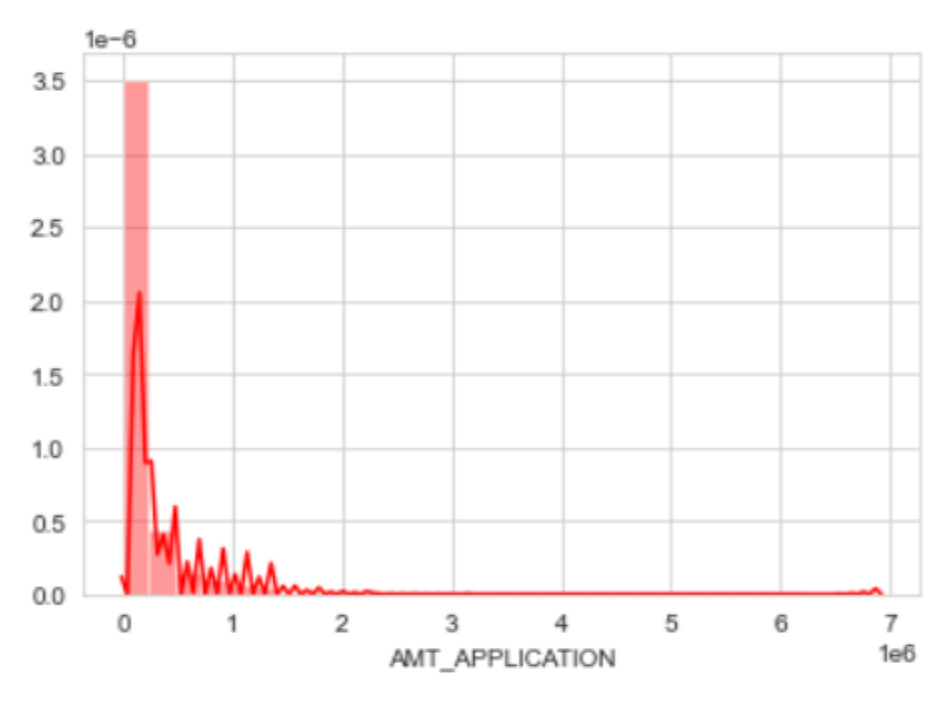
Previous Application Data Analysis

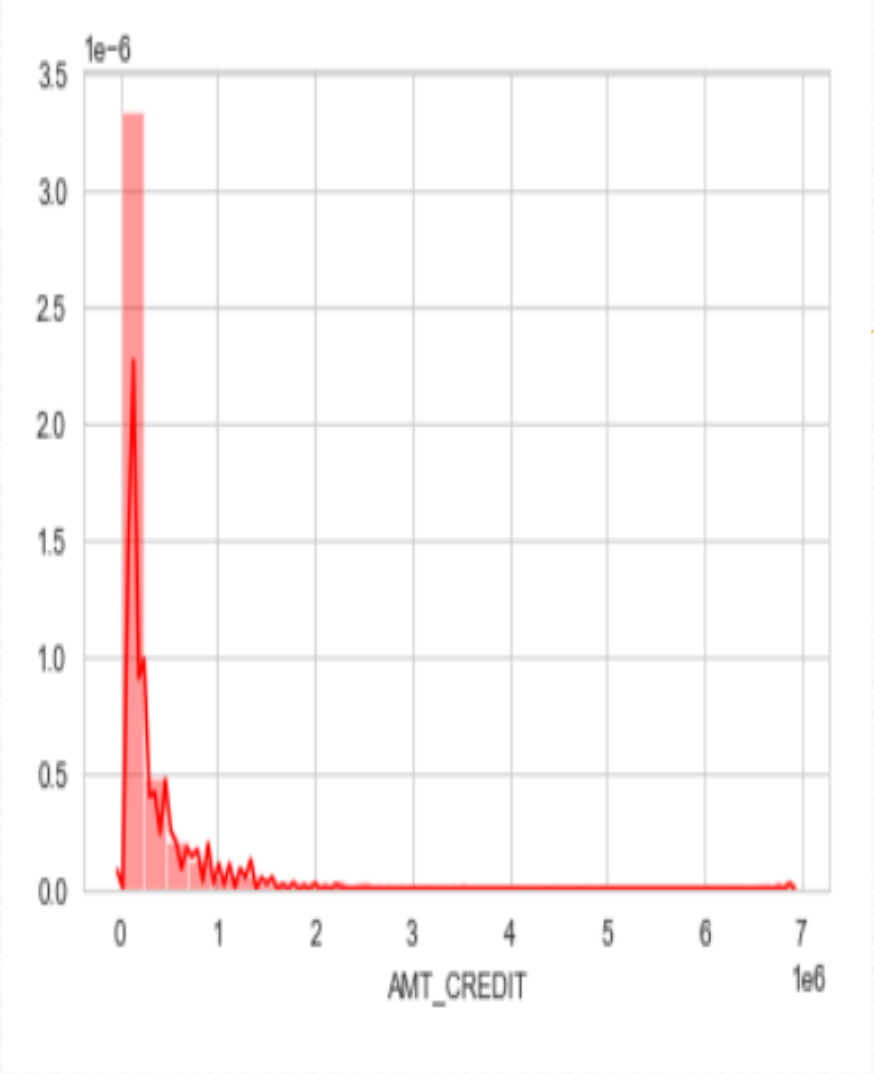
Data Cleaning and Manipulation

- Dropping columns with more than 90% of values as null values.
- Dropping unwanted columns such as WEEKDAY_APPR_PROCESS_START, HOUR_APPR_PROCESS_START
- We can observe by plotting boxplots that AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, DAYS_DECISION, CNT_PAYMENT all have outliers. But these values can be present in real life scenario so we can keep them for now. For example, AMT_APPLICATION have outlier which is obvious because some people can apply for higher loans and AMT_CREDIT also show outliers which verifies that some people who applied for higher loans has received the loans as per their request.

- Analyzing AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT columns

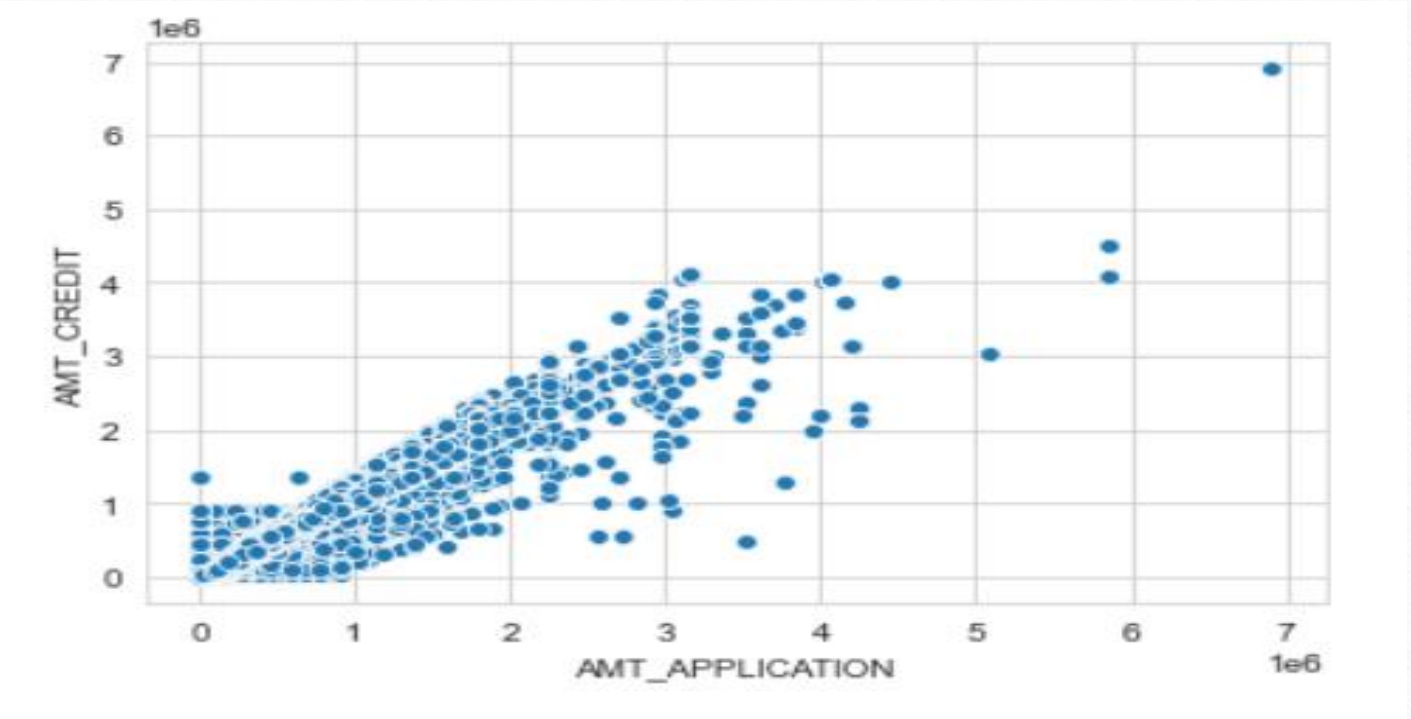






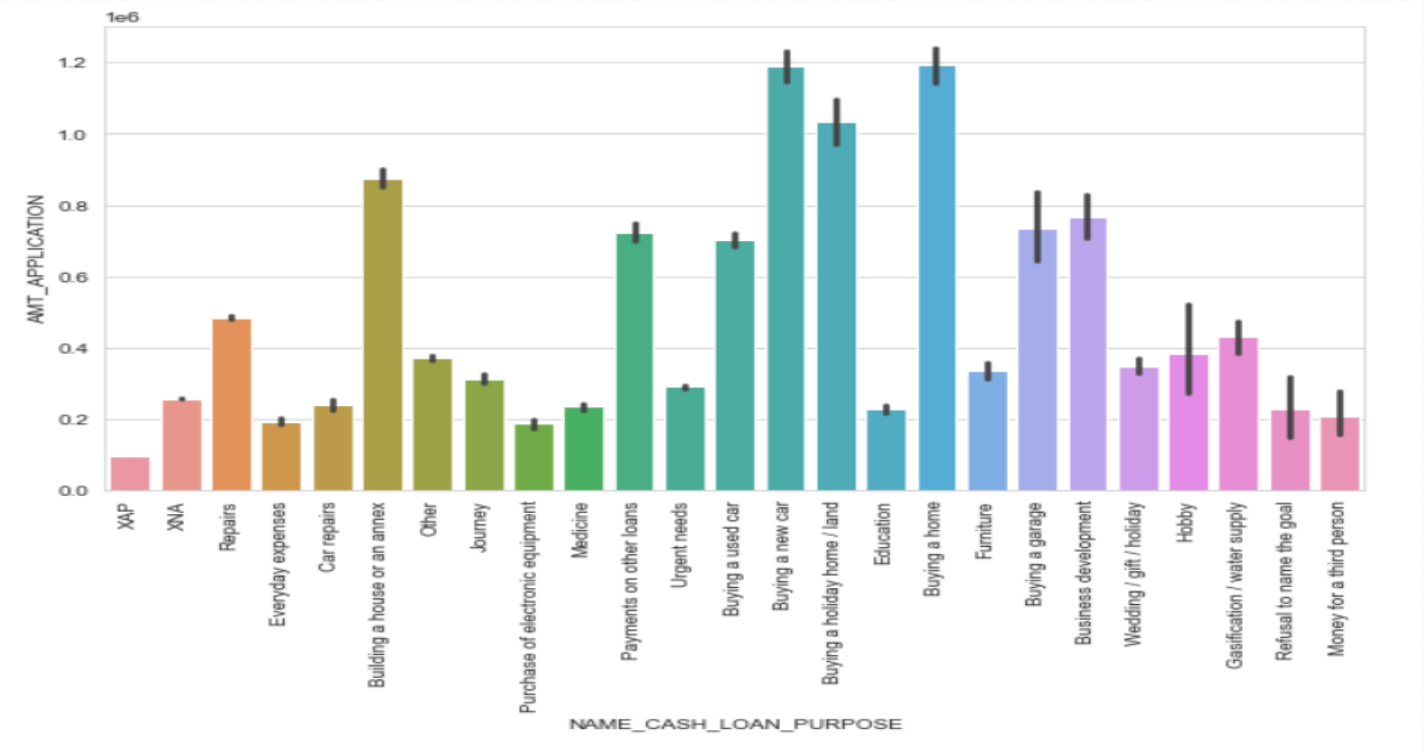
It can be observed that AMT_APPLICATION and AMT_CREDIT both have more distribution in lower area which means majority of people have applied for lower loans amounts and these have been granted. There are few who have applied for higher loans and have been granted requested amounts.

- Relationship between AMT_CREDIT and AMT_APPLICATION



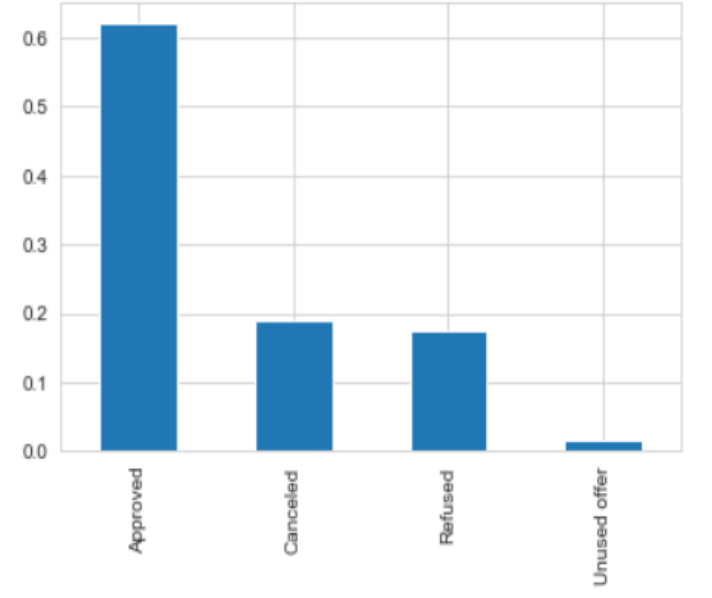
Here we can confirm that AMT_CREDIT and AMT_APPLICATION have kind of linear relationship. Most of the people have received the loan amounts which they have requested for. Where as there are few people who requested for more loan amounts but received lower than requested

➢ Analyzing the purpose of loan



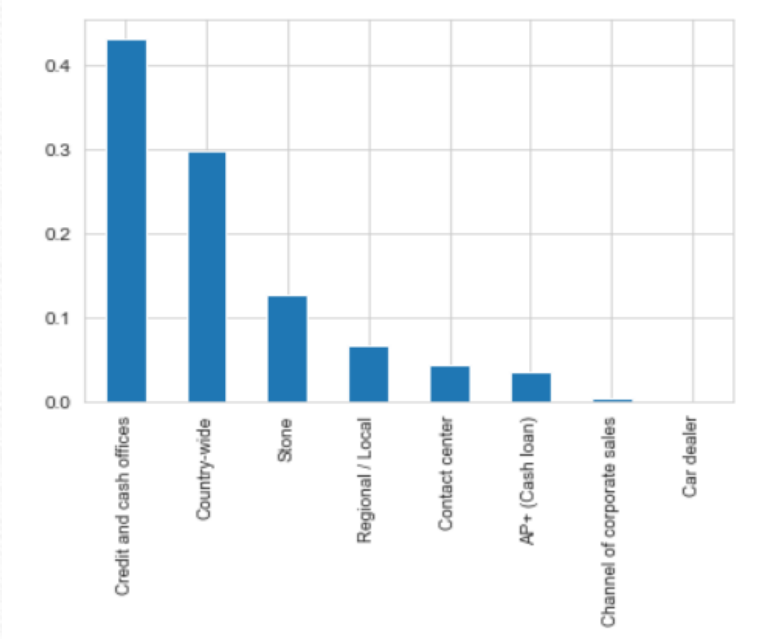
We can observe that majority of applicants are seeking for loan for the "Buying a home", "Buying a new car" and "Buying a holiday home/land" purposes.

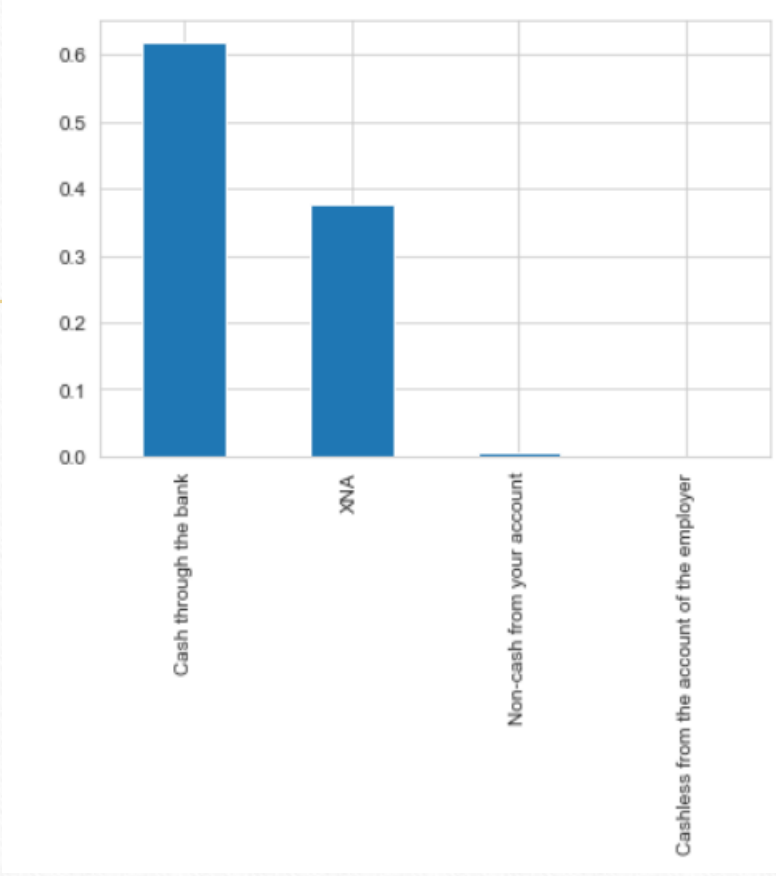
Further analyzing certain columns



- Most of the application received are approved by bank

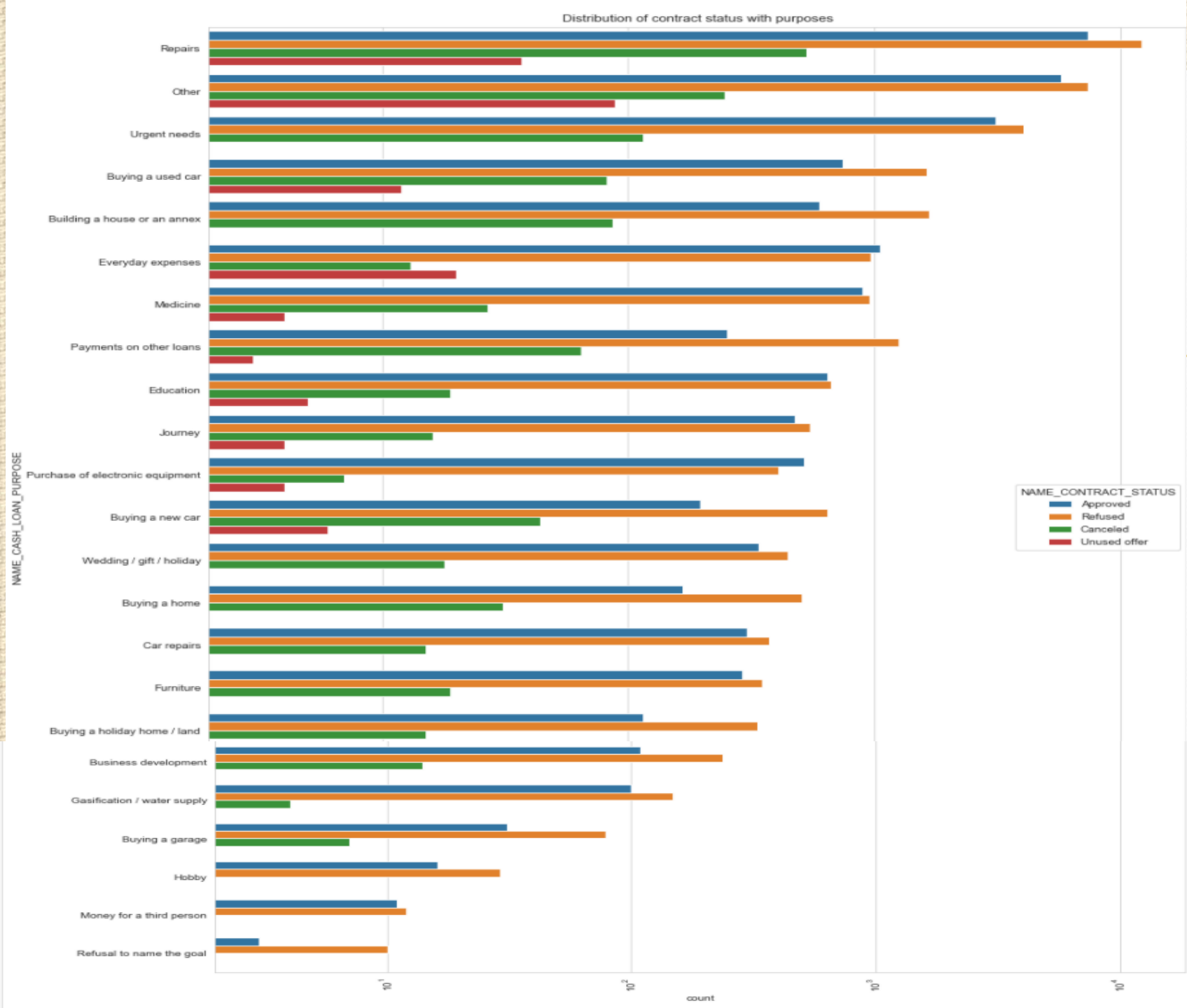
- Most of the applications are received through Credit and cash offices





- Client chooses mostly Cash through the bank method to pay.

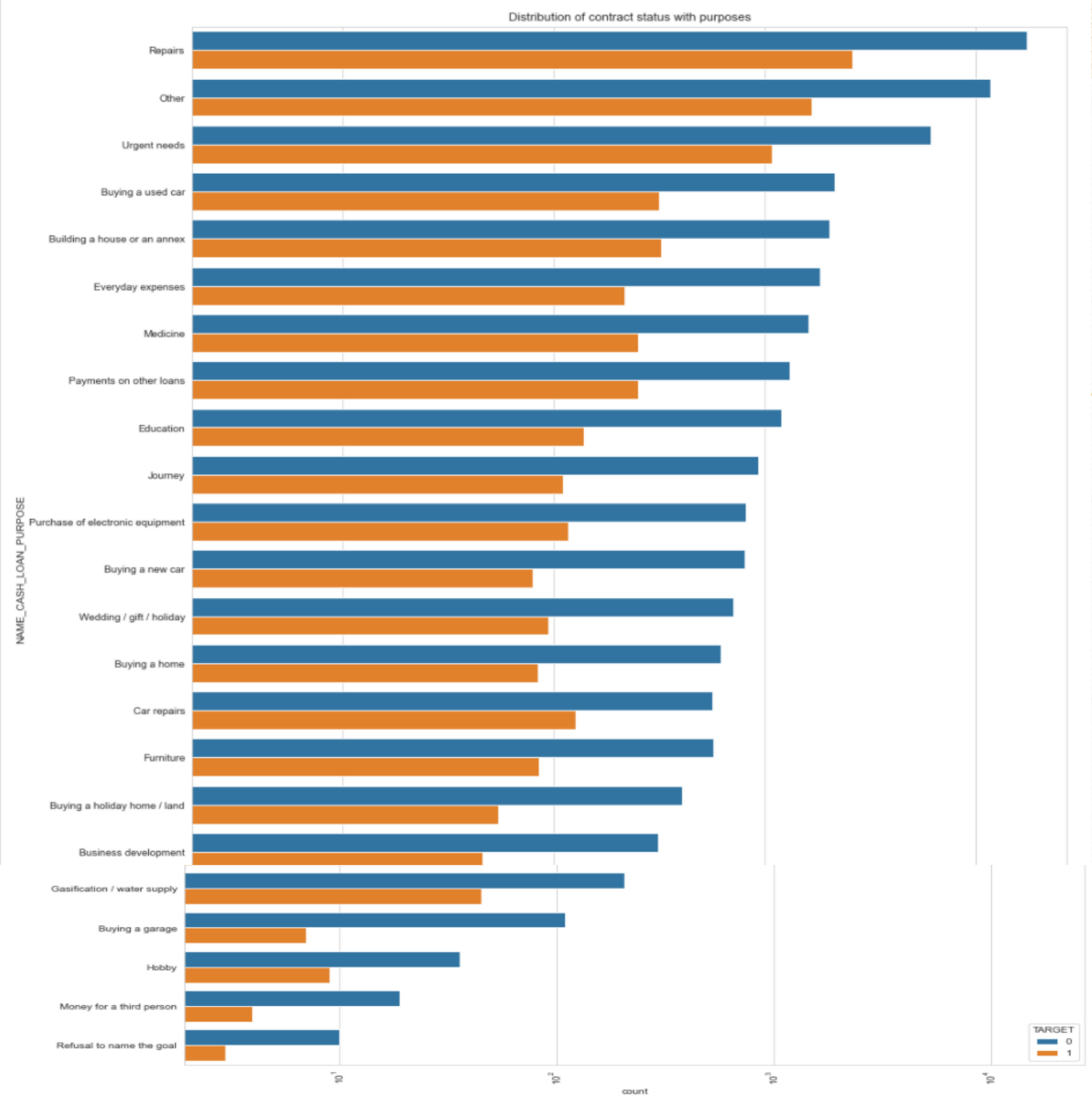
Merging Applications dataset with Previous Application dataset and finding interesting patterns.



Distribution of contract status with purposes.

Observation :

- There is significant rejection in loan of the people who have refused to name the goals
- People applying for loan for the repair purposes are facing highest rejection
- There are more approval than rejections on Everyday expenses and purchase of electronic equipment's loan
- Education and medicine purpose are having almost equal approvals and rejections

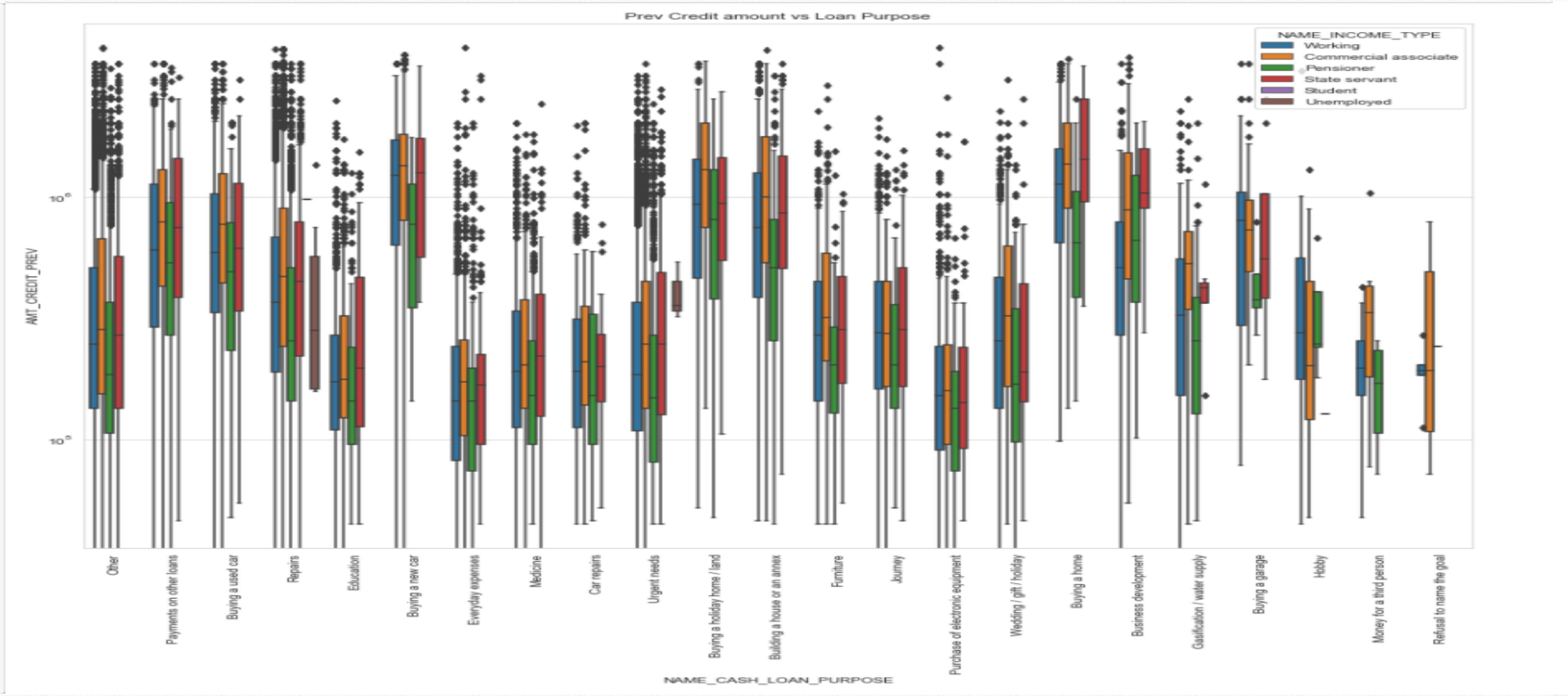


1.People seeking loan for "buying a garage" are less likely to default.

2."Repairs" purposes are more likely to default and hence their loan rejection ratio was also higher

3.Banks can focus on lending loans to people with "business development" "education" "buying a home/land" "buying a new car" as they are less likely to default

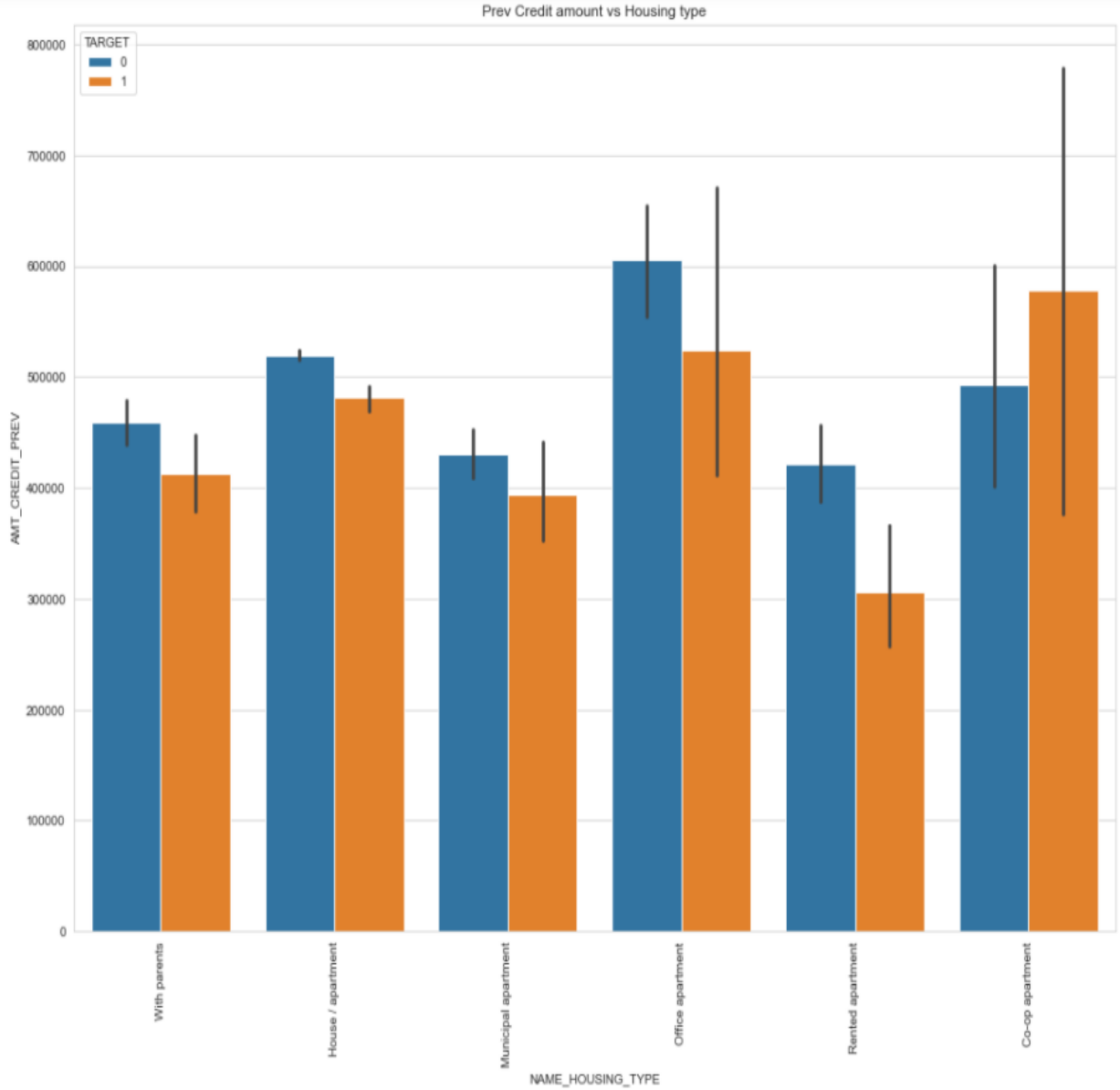
Analyzing loan purpose of previous loan amounts



Drawing observations from the last slide :

From the previous figure it can be observed that :

- Credits are higher for people with "buying a new car" "buying a holiday home/land" "buying a home" "business development" purposes
- Students and Unemployed have overall low credit amounts as compared to other income type people
- State servants gets higher credits for education, journey and buying a home as compared to others
- Working class people have more credits for buying a new car
- Pensioners have more credits for buying a holiday home as compared to other purposes



Analyzing Previous Credit amount vs Housing type

OBSERVATIONS :

- Bank should focus on lending loans to people with housing type "office apartments“
- Banks should analyze risk more closely before lending to people with housing type "Co-op apartment“.

Conclusion

- Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.
- Banks should focus less on income type ‘Working’ as they have most number of unsuccessful payments.
- In loan purpose ‘Repair’ is having highest number of unsuccessful payments on time.
- Target as much clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.

Thank you