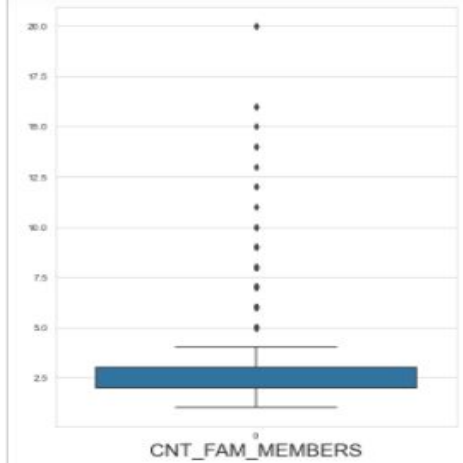
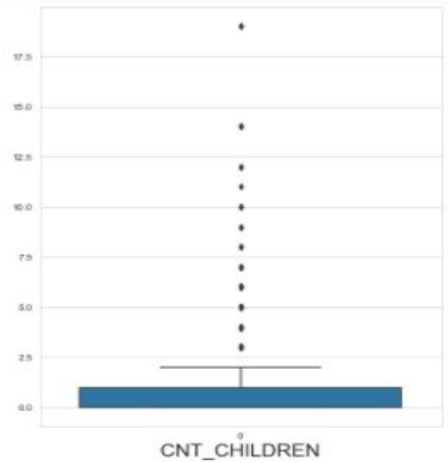
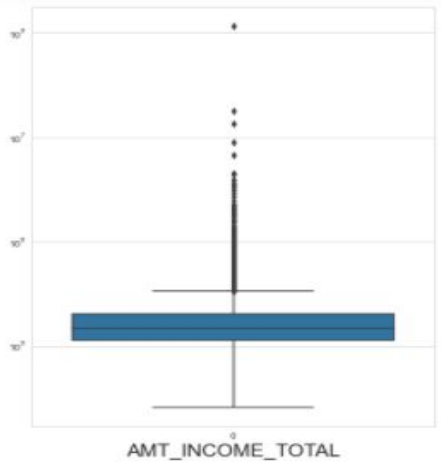
# Credit EDA Case Study

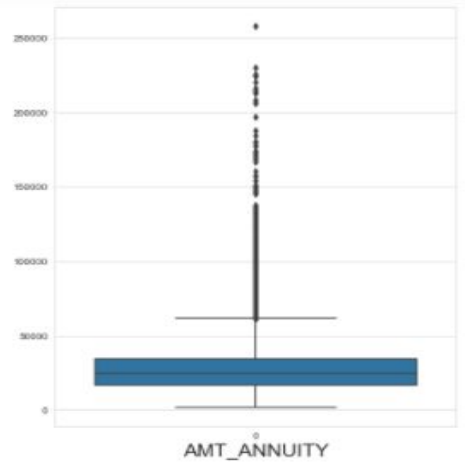
By Dhruv Sharma and Prajwal Gunjikar

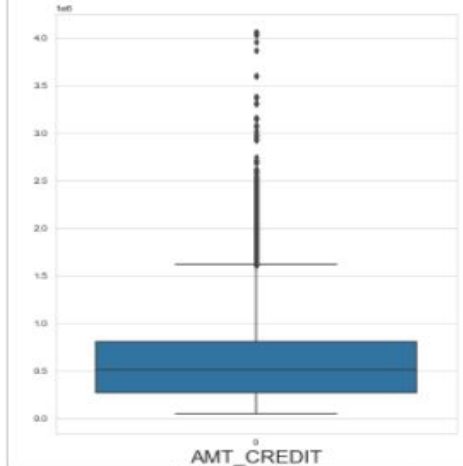
# 01| Outlier Detection

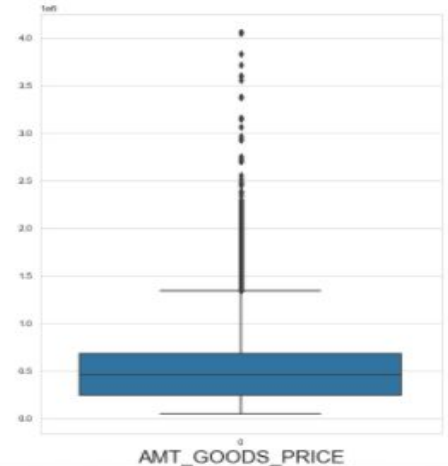


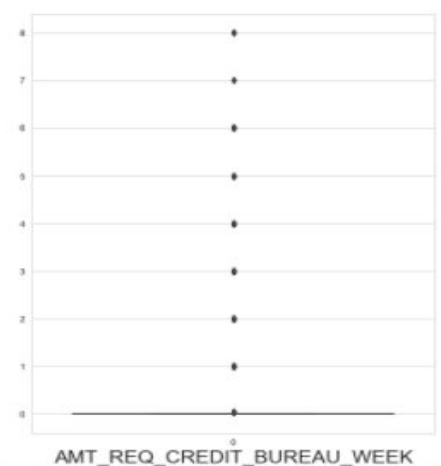



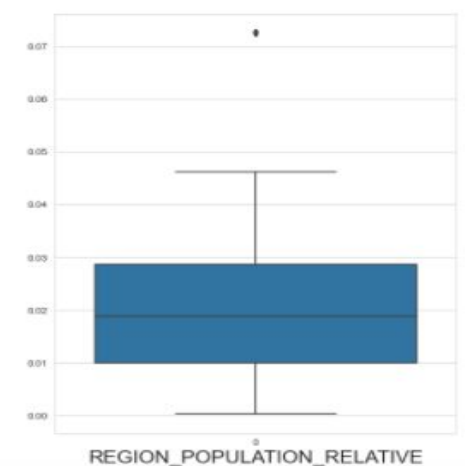



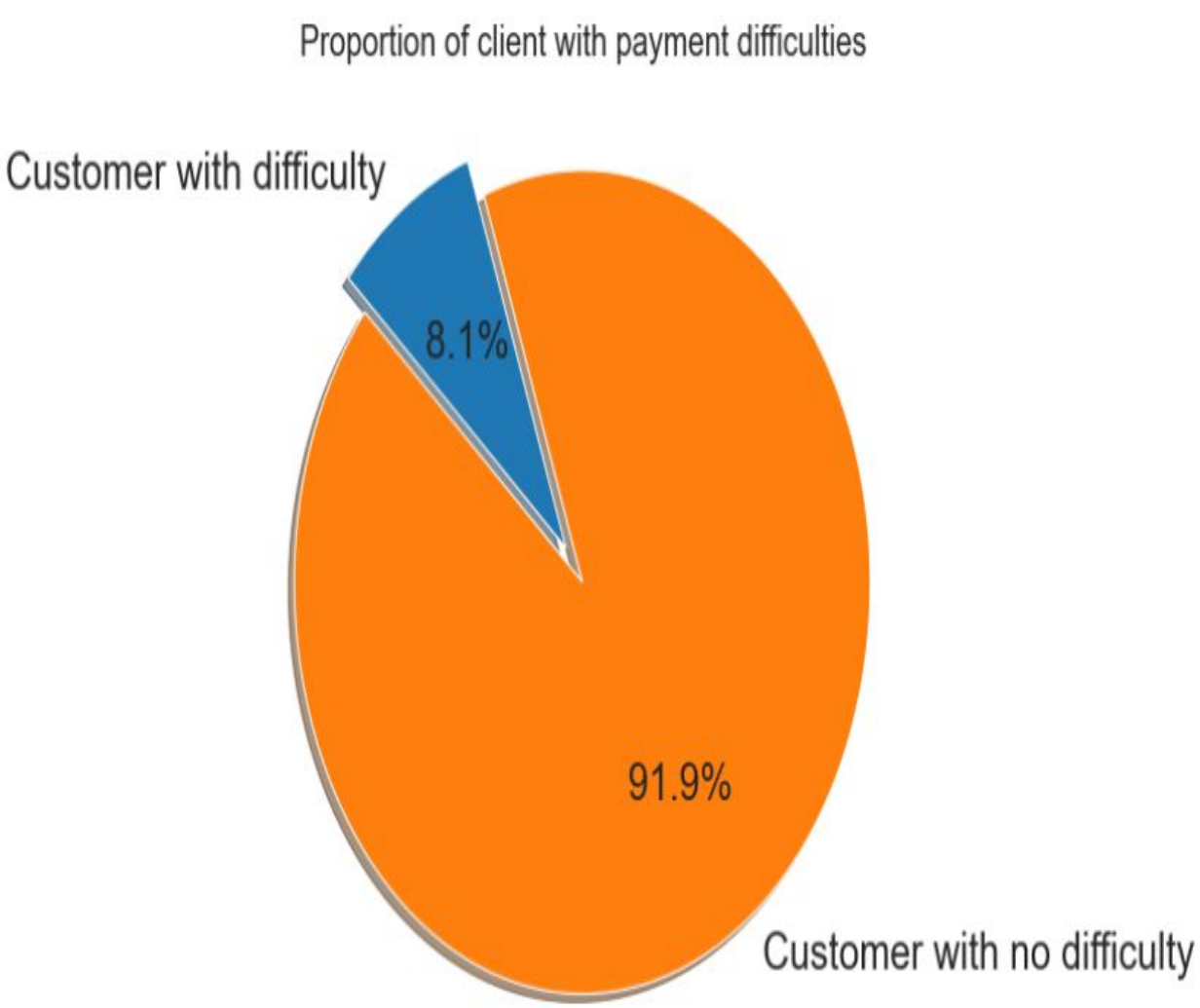
## Inferences :

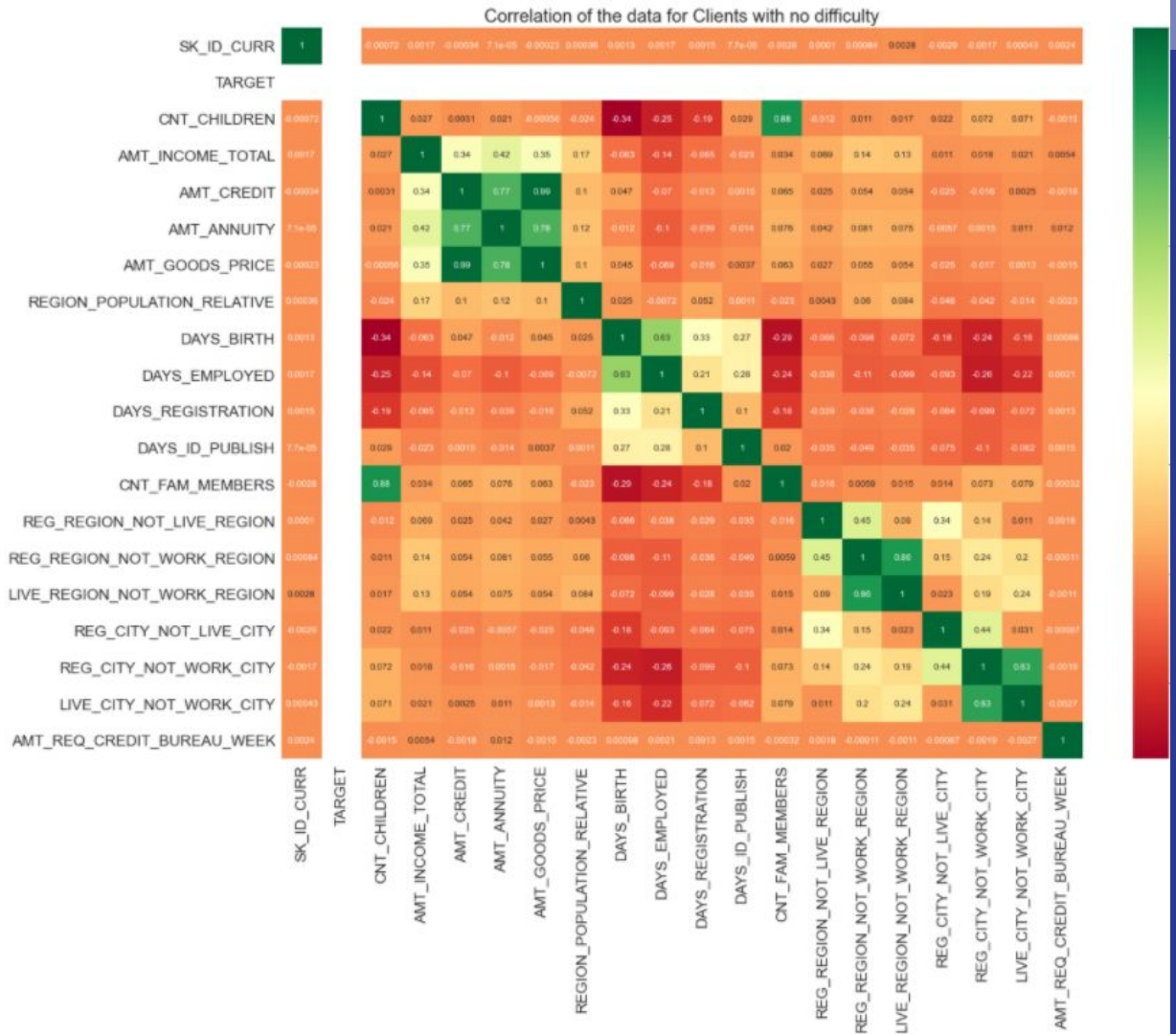
- **From the diagram above, we can see that the column CNT_FAM_MEMBERS does contain outliers as there are some people with large number of family members i.e. (members > 20)**
- **The column CNT_CHILDREN does contain outliers as there are some people with large number of children (i.e. children > 18)**
- **The column AMT_INCOME_TOTAL does contain outliers. But this is an advantage for the bank because more the income of the applicant, more likely is he/she going to repay the loan without any difficulties**
- **The column AMT_ANNUITY does contain outliers. This is an area of concern for the bank**
- **The column AMT_CREDIT has outliers present and this means that some people have asked for a huge credit amount while applying for the loan**
- **The column AMT_GOODS_PRICE has outliers present. This means that for some cases the amount of goods is higher than the credit of loan given.**
- **The column AMT_REQ_CREDIT_BUREAU_WEEK has few outliers present meaning that the number of enquiries to credit bureau for some applicants is higher than others**
- **The column REGION_POPULATION_RELATIVE has no major outliers present.**

## 02| Data Visualization Of Clients With payment Difficulties

**Inferences :**

- From the pie chart , we can see that around 8.1 % of the client are having payment difficulties whereas 91.9% clients are facing no issues. This is an imbalanced condition and we need to deal with the data imbalance. First we calculate the imbalance ratio in the TARGET column






## 03 | 10 Highly Correlated Variable From Target 0

1. AMT_GOODS_PRICE - AMT_CREDIT
2. CNT_FAM_MEMBERS - CNT_CHILDREN
3. LIVE_REGION_NOT_WORK_REGION - REG_REGION_NOT_WORK_REGION
4. LIVE_CITY_NOT_WORK_CITY - REG_CITY_NOT_WORK_CITY
5. AMT_GOODS_PRICE - AMT_ANNUITY
6. AMT_ANNUITY - AMT_CREDIT
7. DAYS_EMPLOYED - DAYS_BIRTH
8. REG_REGION_NOT_WORK_REGION - REG_REGION_NOT_LIVE_REGION
9. REG_CITY_NOT_WORK_CITY - REG_CITY_NOT_LIVE_CITY
10. AMT_ANNUITY - AMT_INCOME_TOTAL

## Inferences :

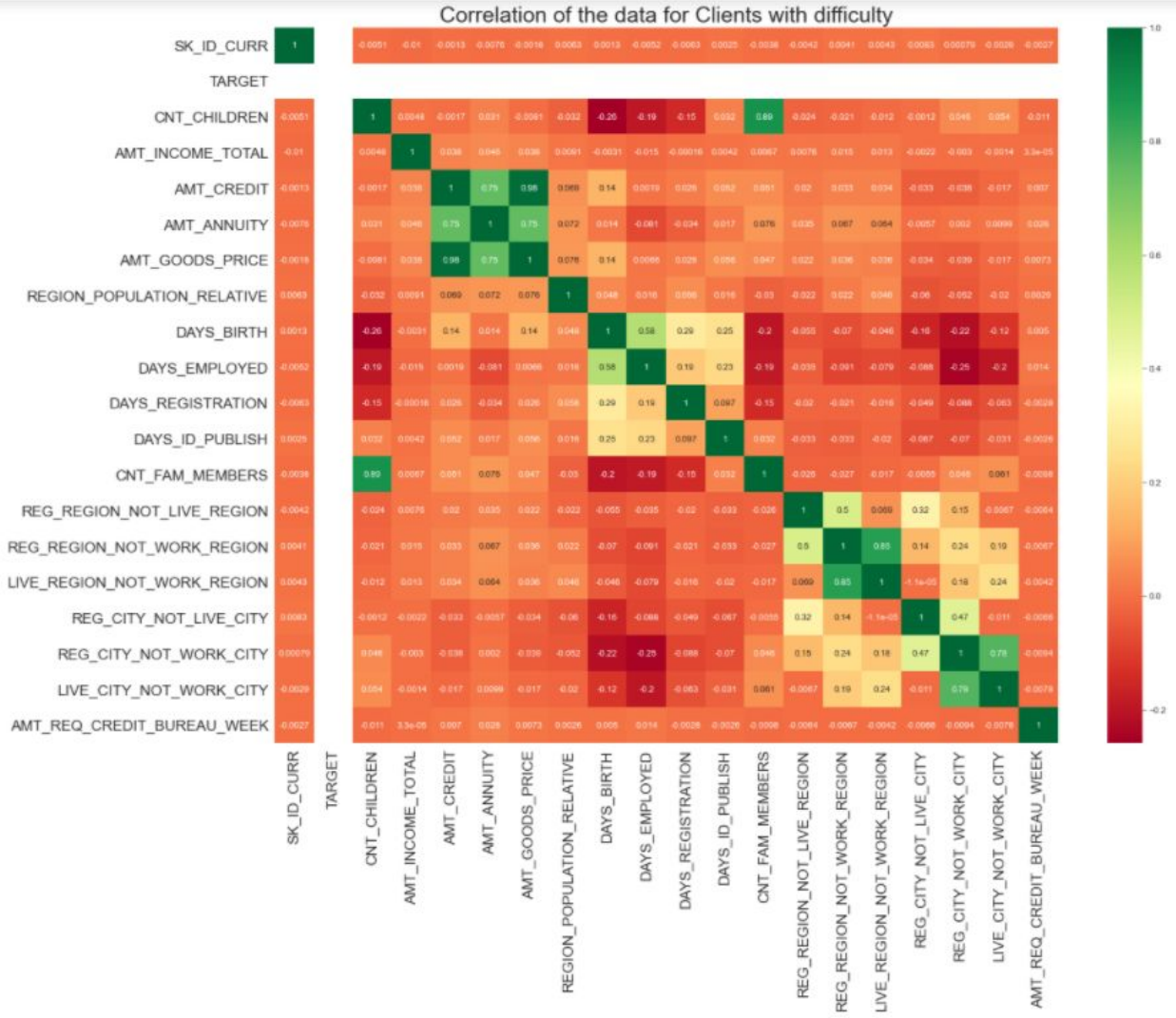
- **From the correlation heatmap, it can be seen that as the price of goods increases, loan credit amount increases and vice versa**
- **Count of family members is proportional to the count of children in the family**
- **The annuity amount is closely correlated to the goods price and the credit amount**

# 04 | 10 Highly Correlated Variable From Target 1

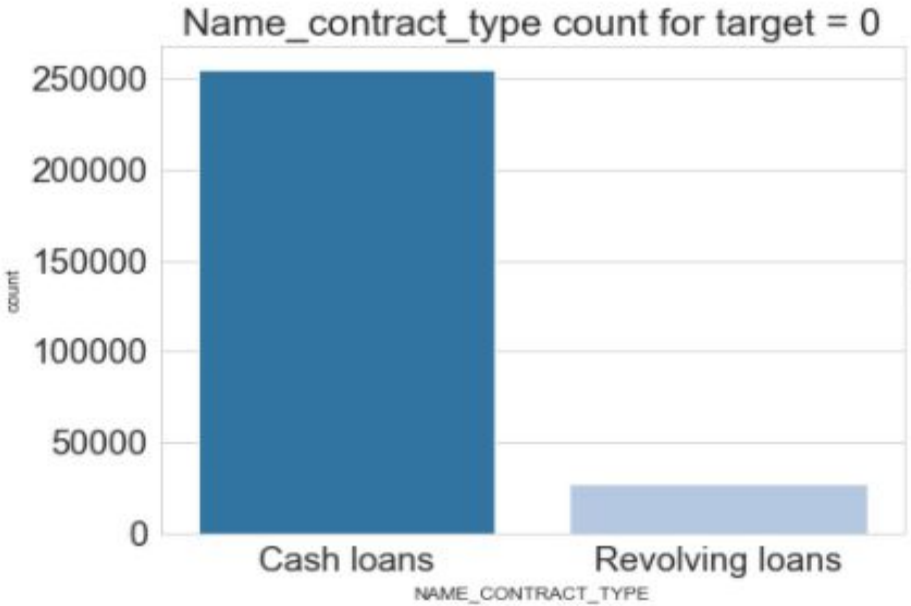
1. AMT_GOODS_PRICE - AMT_CREDIT
2. CNT_FAM_MEMBERS - CNT_CHILDREN
3. LIVE_REGION_NOT_WORK_REGION - REG_REGION_NOT_WORK_REGION
4. LIVE_CITY_NOT_WORK_CITY - REG_CITY_NOT_WORK_CITY
5. AMT_GOODS_PRICE - AMT_ANNUITY
6. AMT_ANNUITY - AMT_CREDIT
7. DAYS_EMPLOYED - DAYS_BIRTH
8. REG_REGION_NOT_WORK_REGION - REG_REGION_NOT_LIVE_REGION
9. REG_CITY_NOT_WORK_CITY - REG_CITY_NOT_LIVE_CITY
10. AMT_ANNUITY - AMT_INCOME_TOTAL

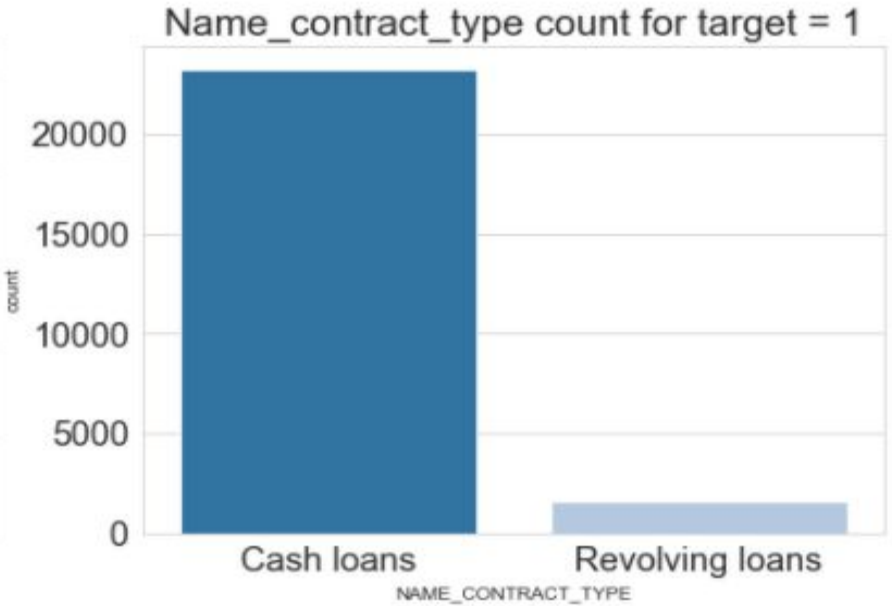
Inferences :

- From the correlation heatmap, it can be seen that as the price of goods increases, loan credit amount increases and vice versa
- Count of family members is proportional to the count of children in the family
- The annuity amount is closely correlated to the goods price and the credit amount



## 05 | Visualizing the contract type for both type of applicant (i.e. applicant with and without payment difficulties) | Univariate Analysis

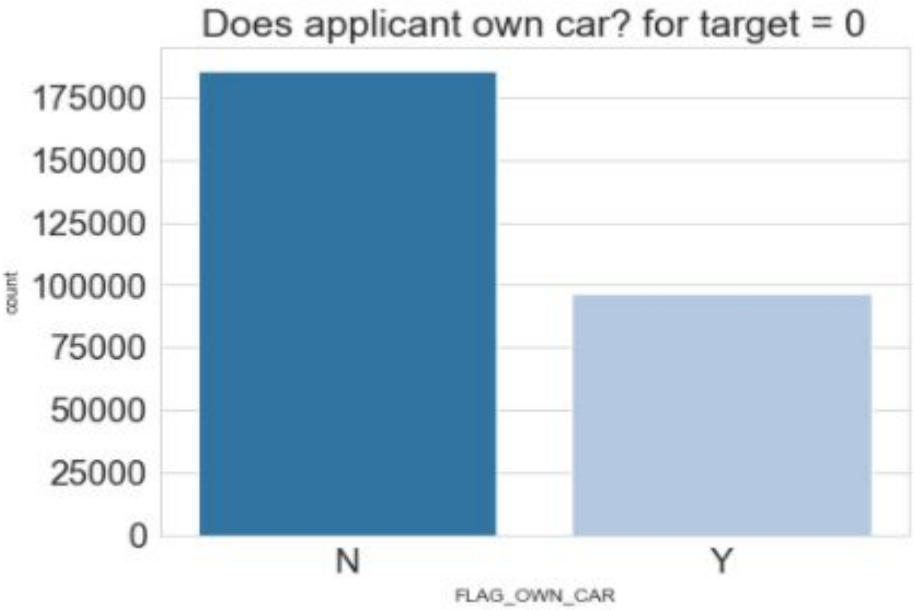


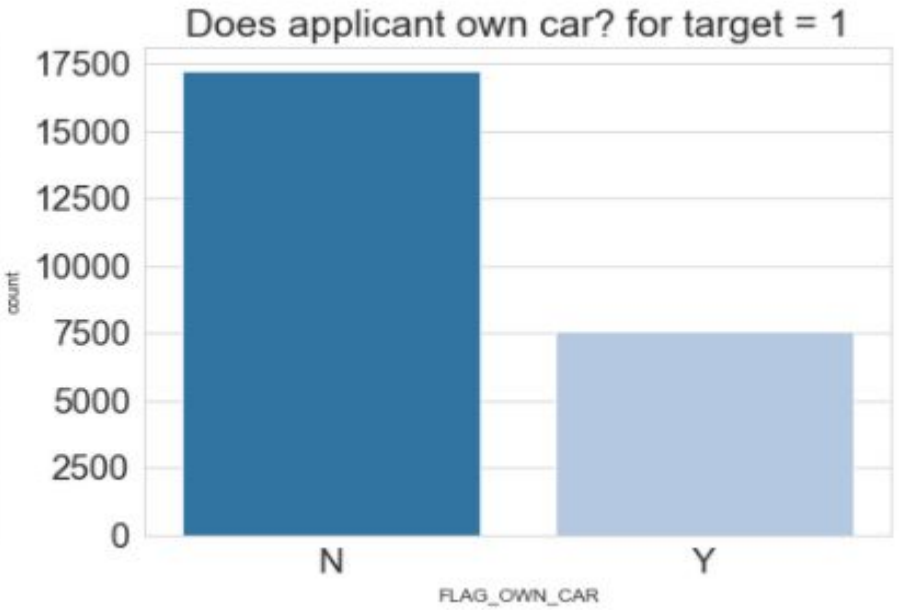




**Inferences :**

- **Cash loans are significantly more than revolving loans for both the applicants with and without payment difficulties.**

# 06 | Displaying count of applicant who own a car for both type of applicant (i.e. applicant with and without payment difficulties) | Univariate Analysis

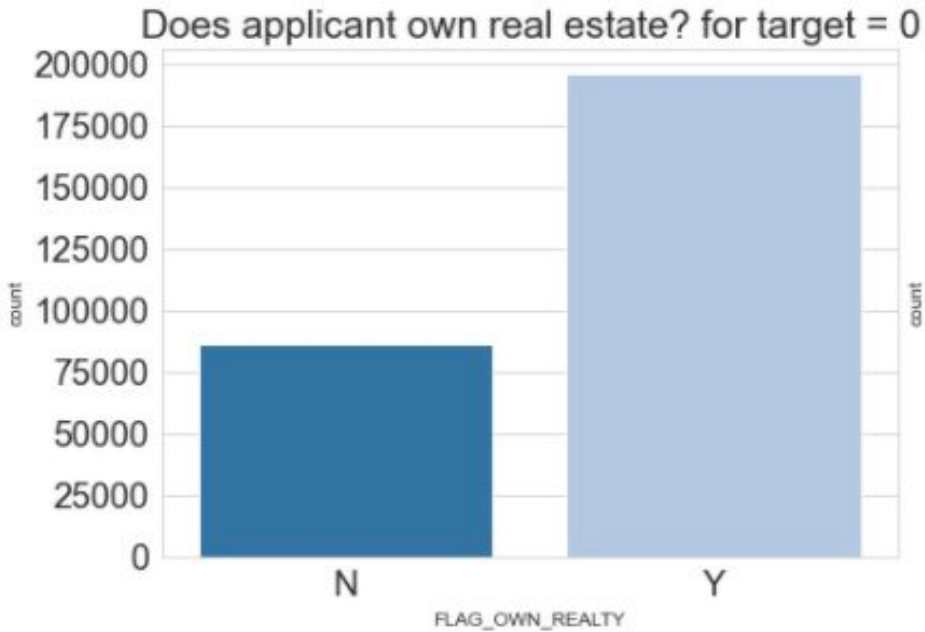


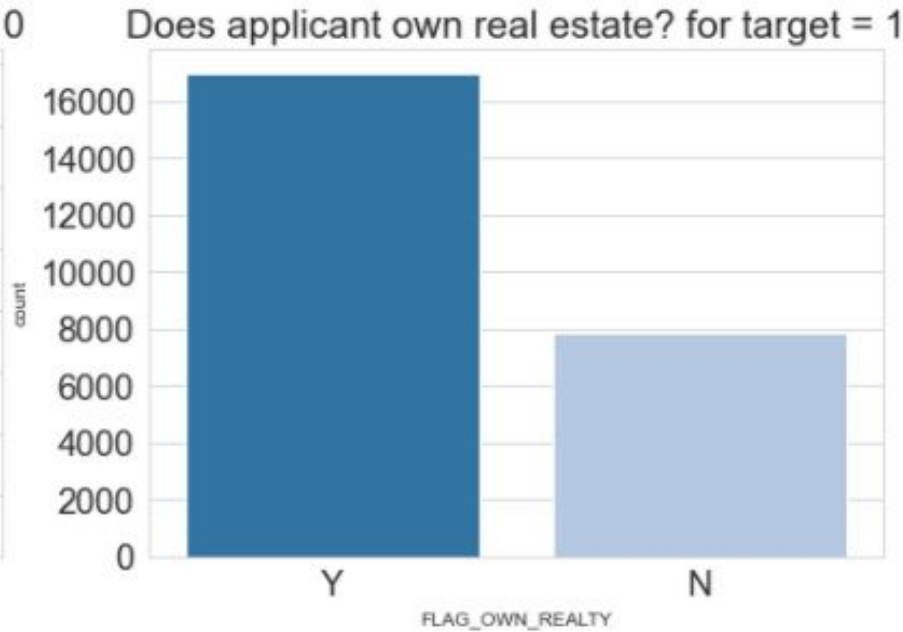




**Inferences :**

- **Most applicants do not own a personal car for both applicants with and without payment difficulties as we can see that the distribution is similar for both cases.**

# 07 | Displaying count of applicant who own a real estate for both type of applicant (i.e. applicant with and without payment difficulties) | Univariate Analysis

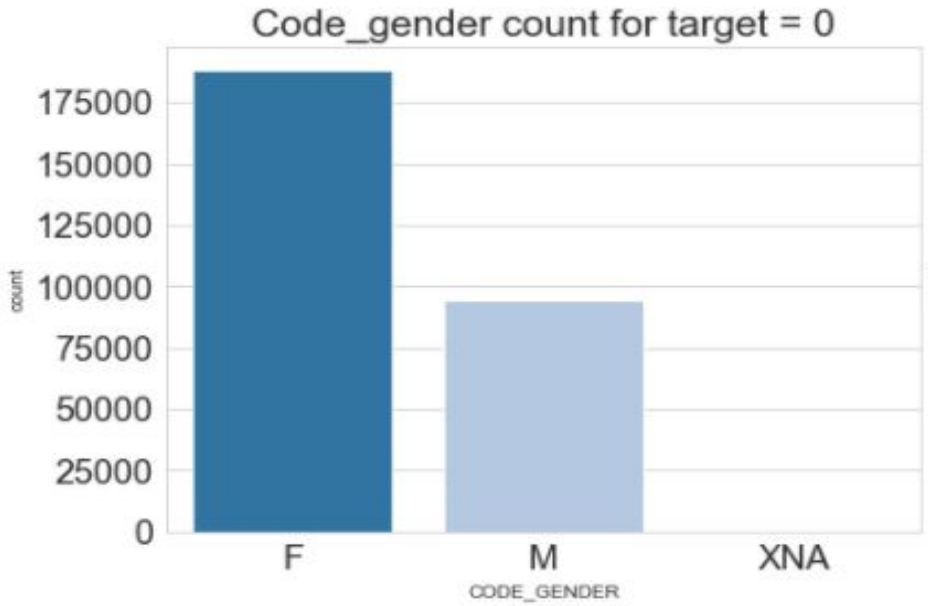



**Inferences :**

- **Most applicants own real estate for both applicants with and without payment difficulties as we can see that the distribution is similar for both cases.**

# 08 | Displaying  gender count of applicant for both type of applicant (i.e. applicant with and without payment difficulties) | Univariate Analysis

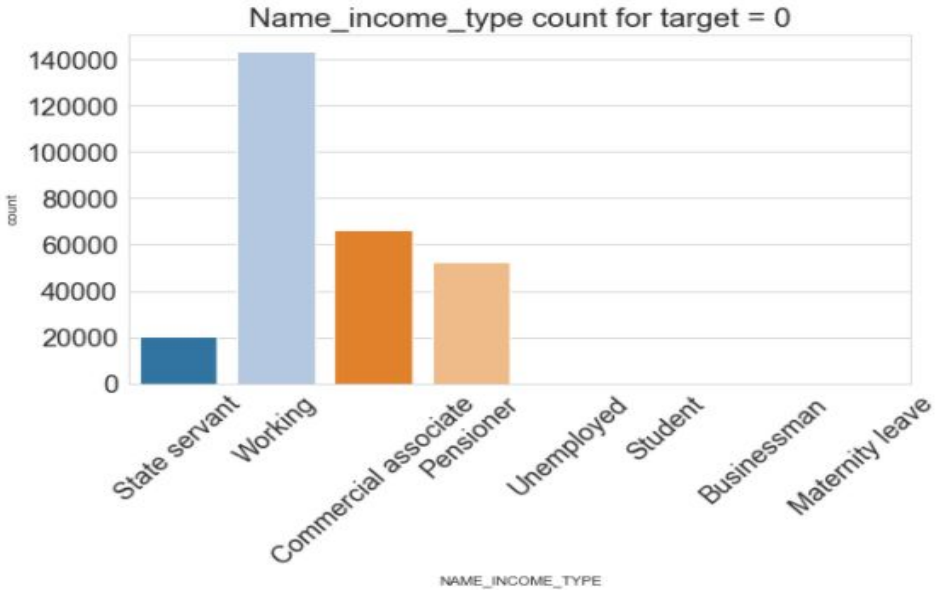


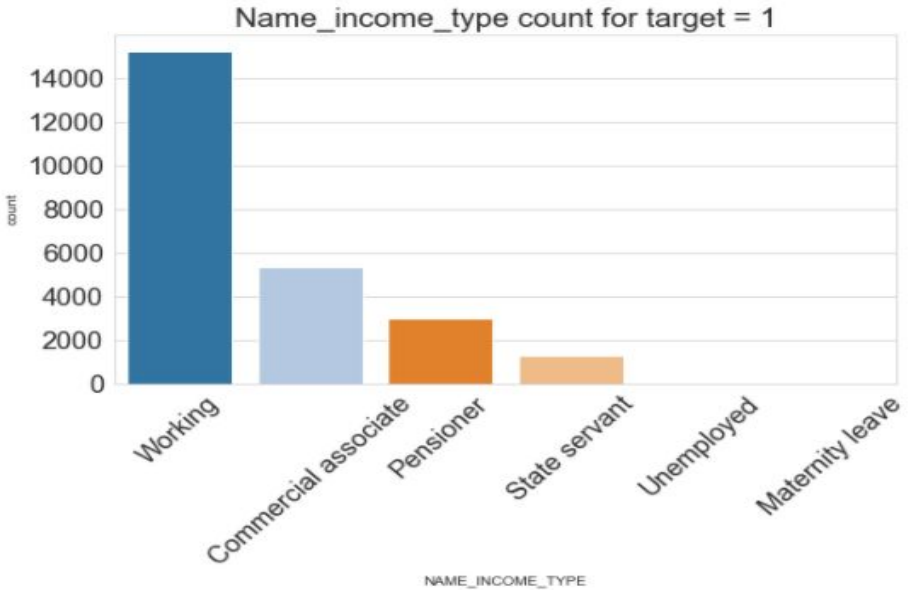


**Inferences :**

- **Most applicants are Female for both with and without payment difficulties**
- **The proportion of males having payment difficulties is greater than males having no payment difficulties.**
- **Also there are no applicants with gender not mentioned (in this case XNA) who are defaulting.**

# 09 | Displaying the income type of applicant for both type of applicants(i.e. applicant with and without payment difficulties) | Univariate Analysis

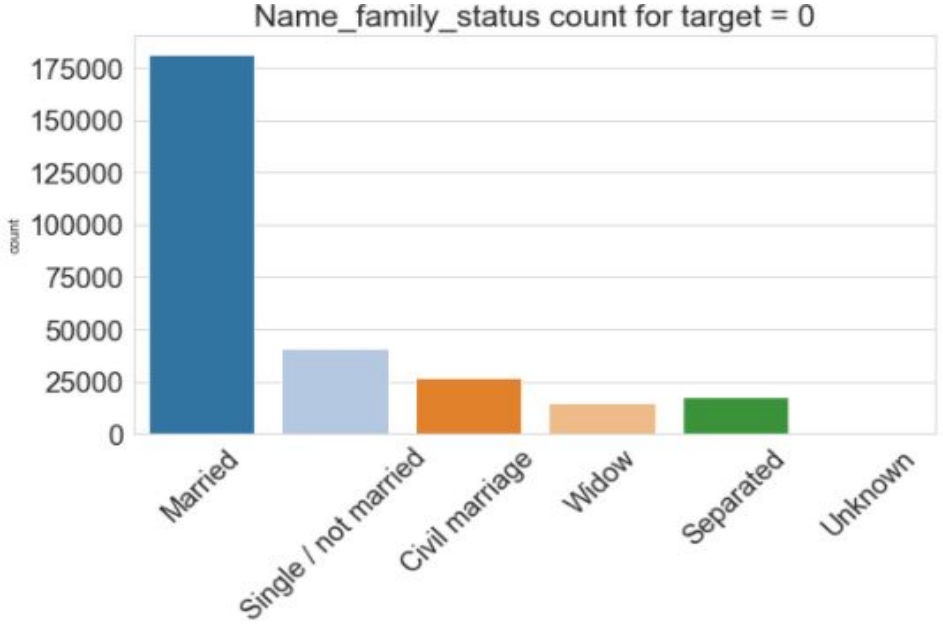


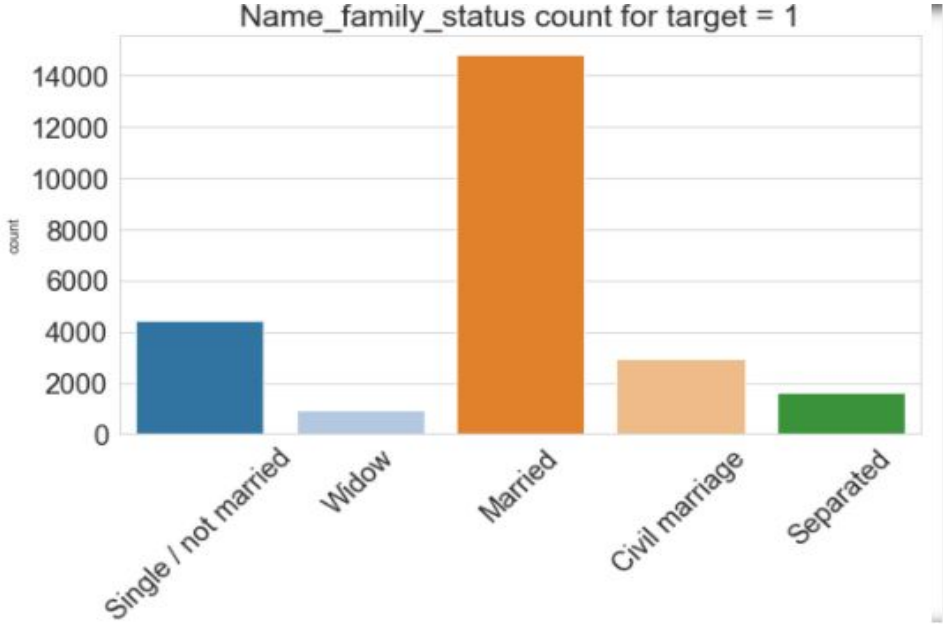




**Inferences :**

- **Most applicants have income type as Working for both cases followed by the Commercial Associate.**
- **The distribution appears similar for both the applicants with and without payment difficulties**
- **There are no applicants with income type as Businessmen and Students who have payment difficulties**

# 10 | Displaying the family status of applicant for both type of applicants(i.e. applicant with and without payment difficulties) | Univariate Analysis

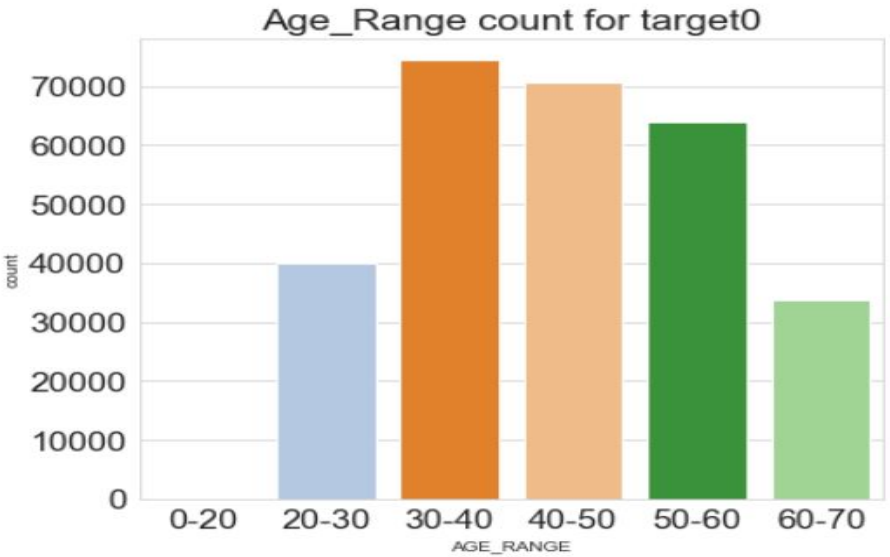


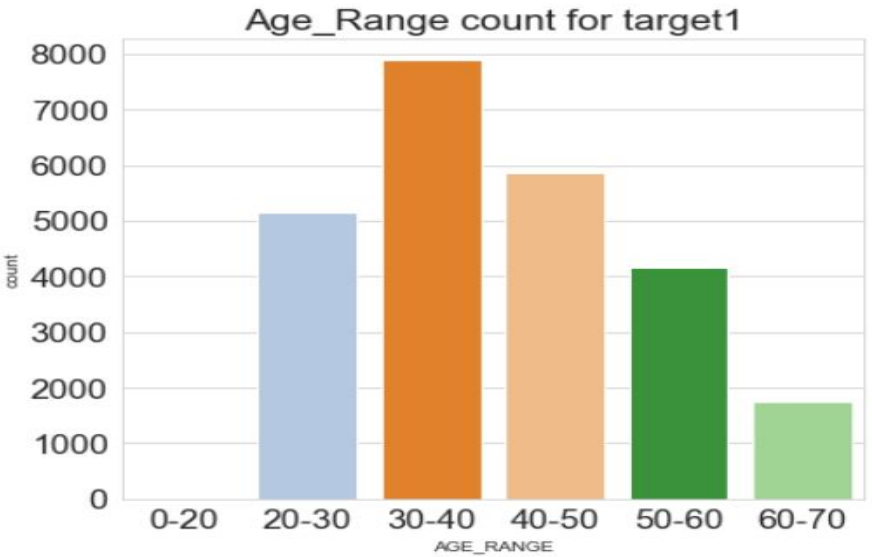




**Inferences :**

- **Most applicants applying for loan are married followed by Single/not married people.**
- **The applicants having family status as widow are the least**
- **The distribution appears to be similar among both the applicants with and without payment difficulties.**

# 11 | Displaying the age range of applicant for both type of applicants(i.e. applicant with and without payment difficulties) | Univariate Analysis

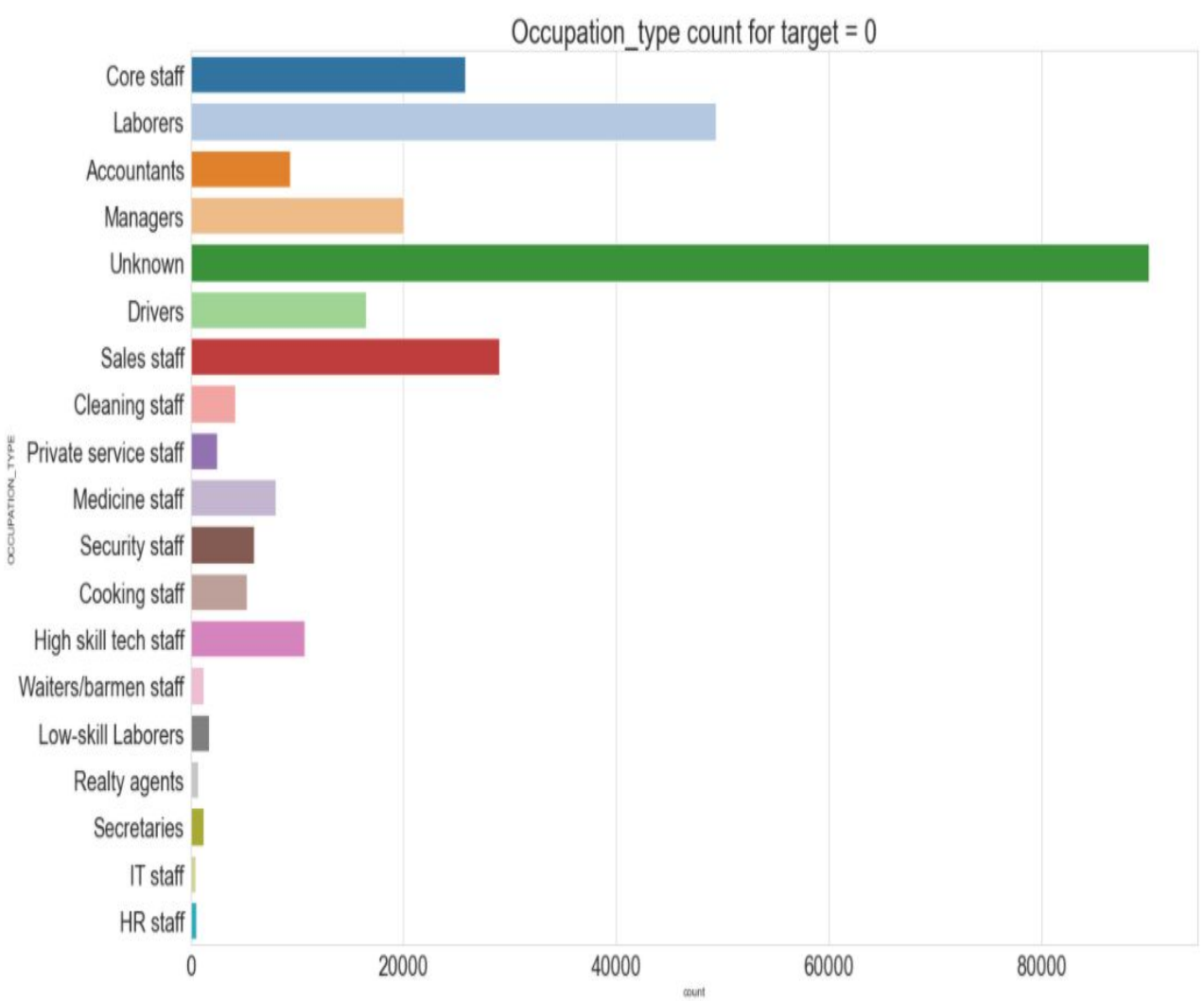



**Inferences :**

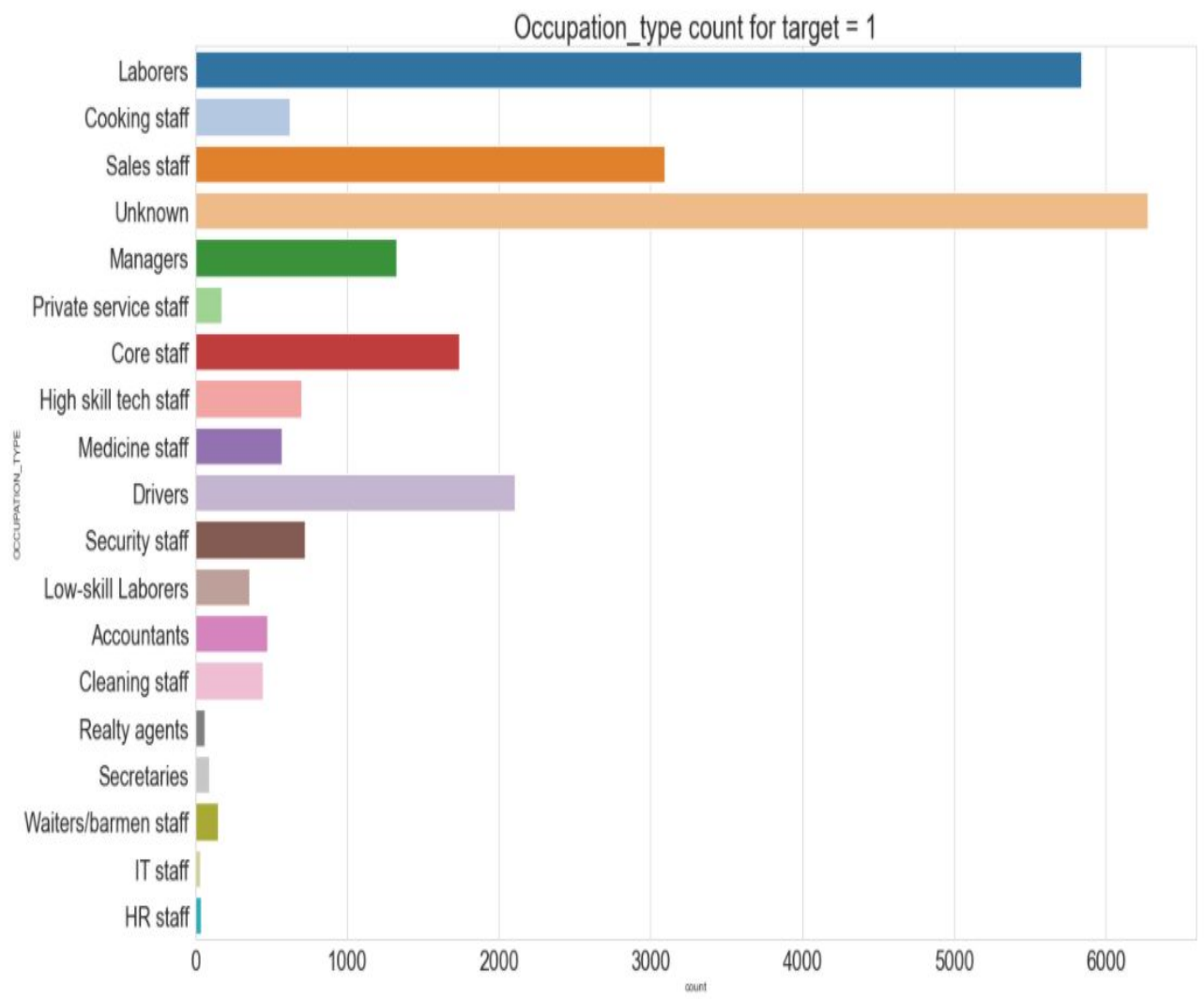
- **There are no applicants aged 0 to 20 for both cases**
- **The age group 30-40 have the highest number of applicants for both cases**
- **It can be seen that the age group 40-50 has almost similar count to age group 30-40 for applicants with no difficulties. Whereas for clients with difficulty, the count seems to have gone down. Similar is the case for age group 50-60.**
- **Hence bank should focus on the age group 30-40 as they are most likely to have issues regarding payment.**

# 12 | Displaying the occupation type of applicant for applicant without payment difficulties | Univariate Analysis




## Inferences :

- Since there are 31.3% missing values in the column OCCUPATION_TYPE, we can impute them with Unknown and work with the rest of the data
- The majority applicants having no payment difficulties are Laborers followed by Sales staff and Core staff.

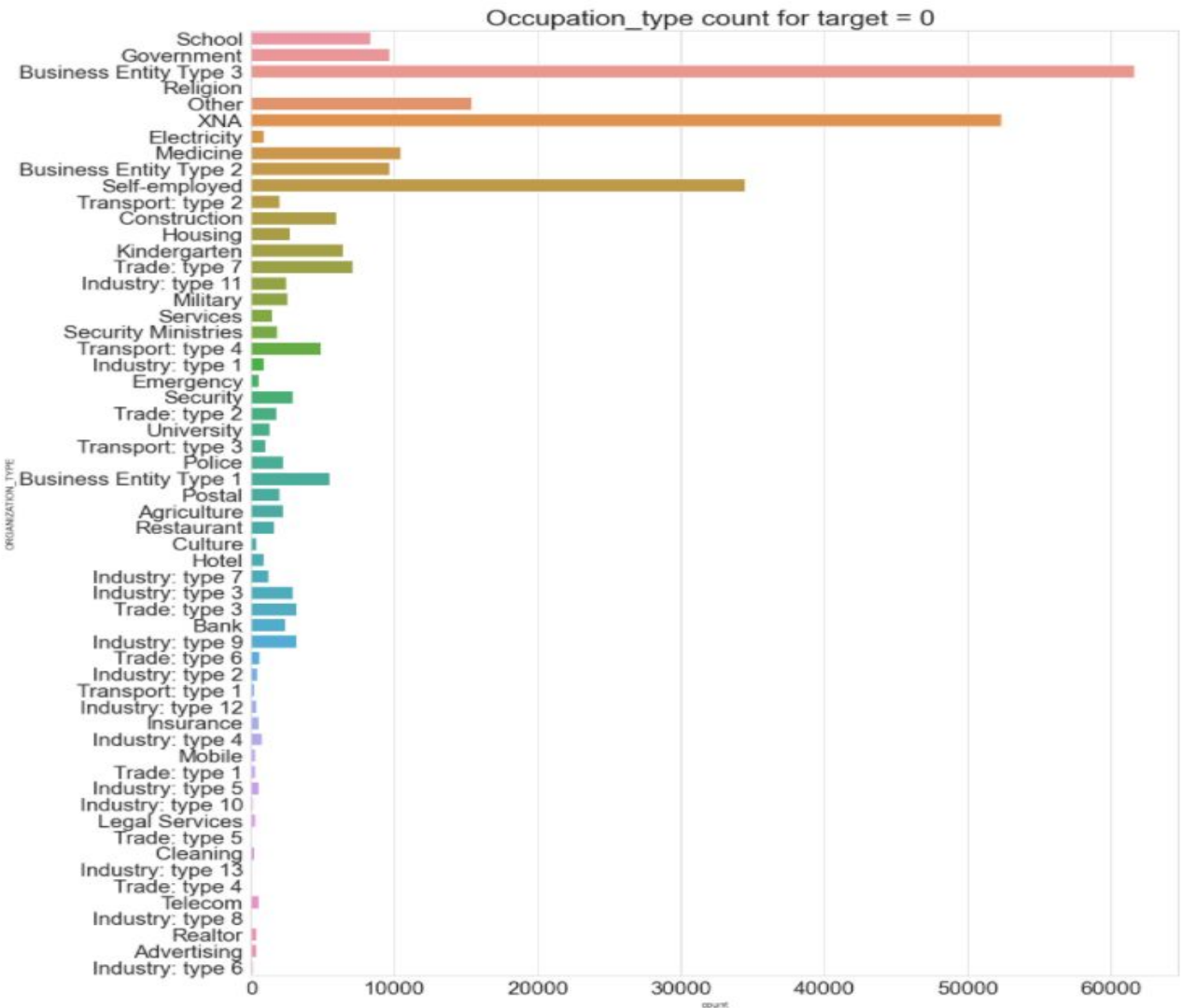


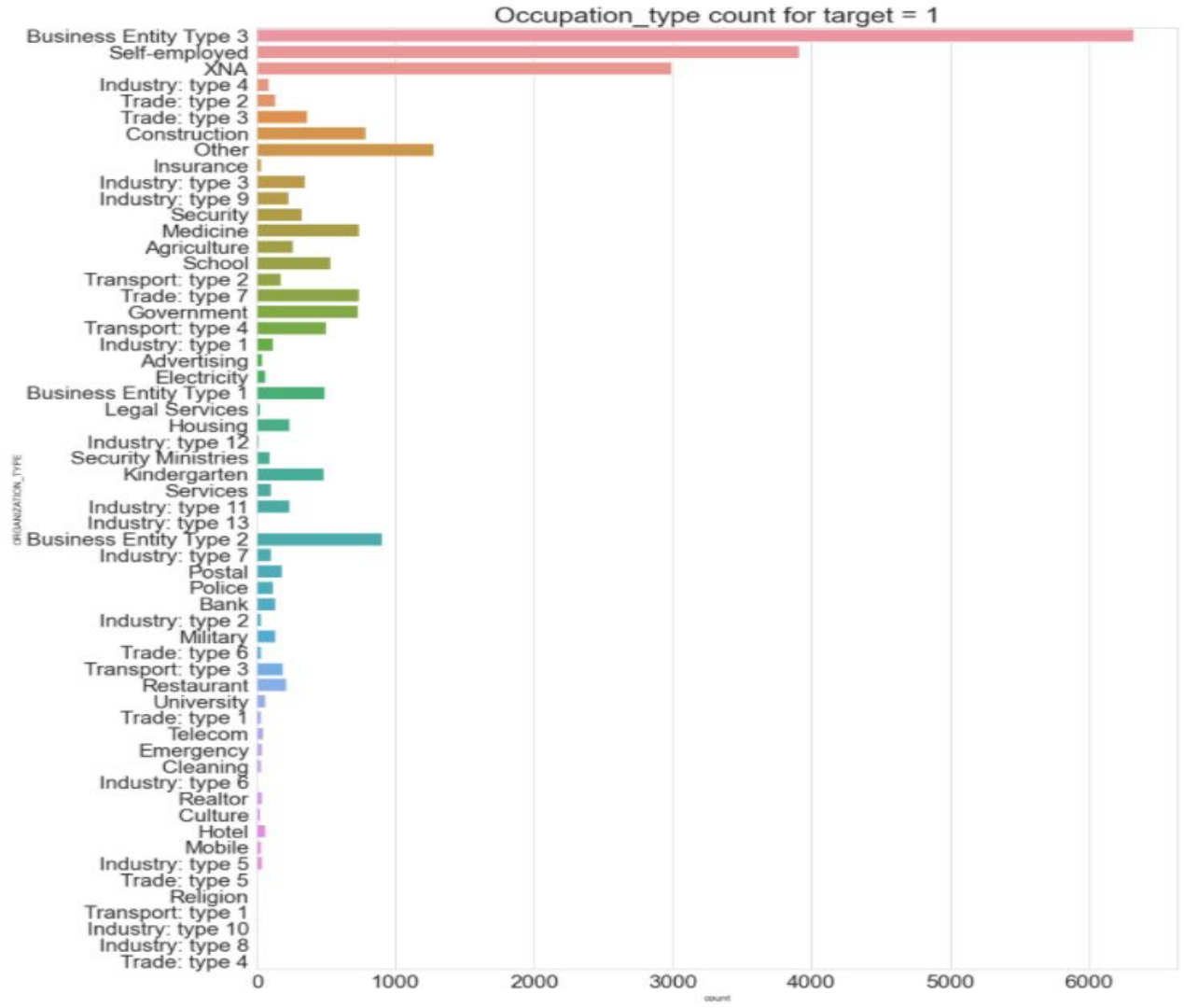

## 13 | Displaying the occupation type of applicant for applicant without payment difficulties | Univariate Analysis

**Inferences :**

- **Laborers have the most difficulty while repaying loan and hence bank should be aware of the laborer occupation type in particular.**
- **Sales staff follows the laborer in the difficulties faced while repaying the loan.**
- **The Drivers seem to have a significant increase in the number of difficulties in repaying loan.**

## 14 | Displaying the organization type of applicant for applicant without payment difficulties | Univariate Analysis



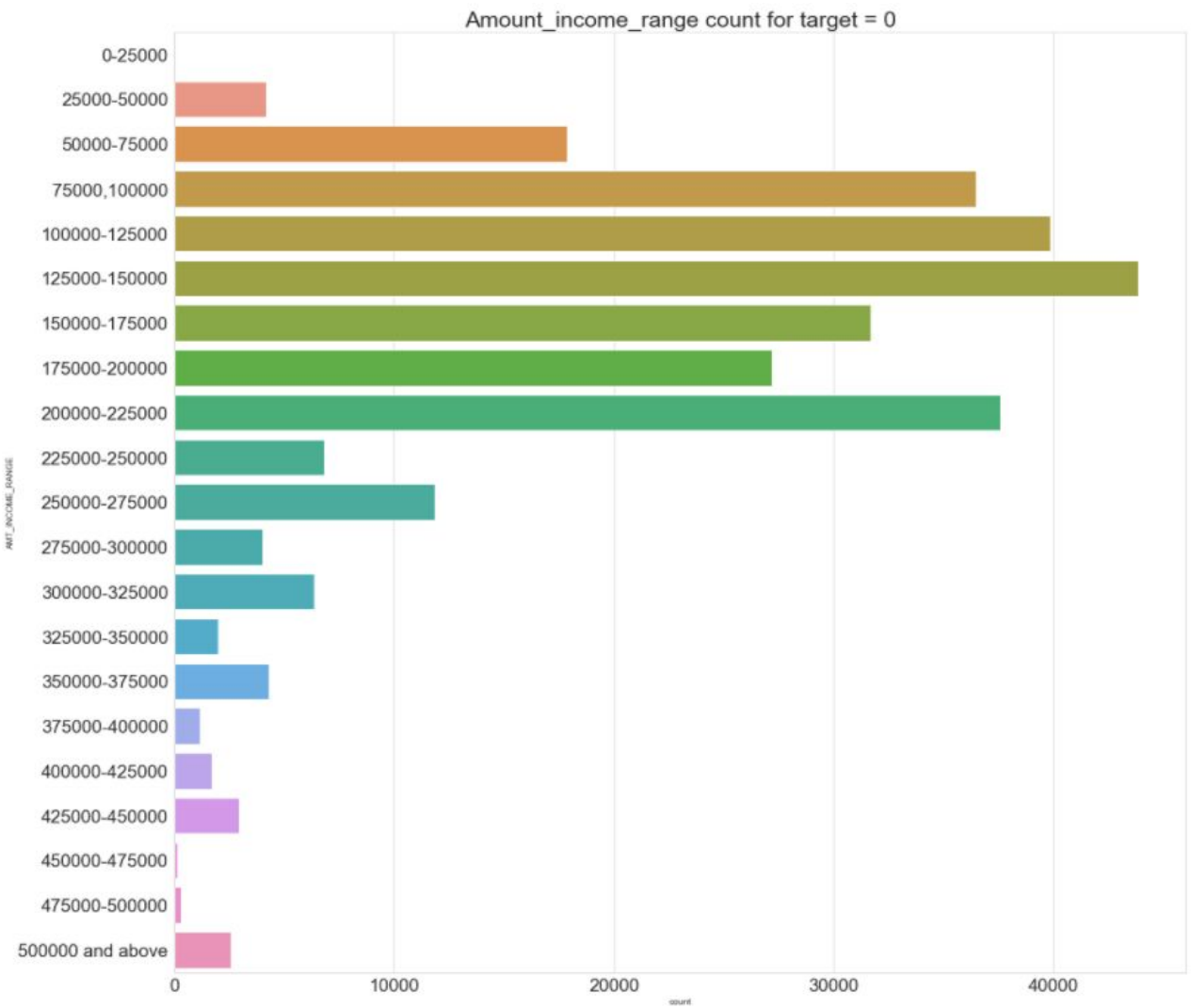


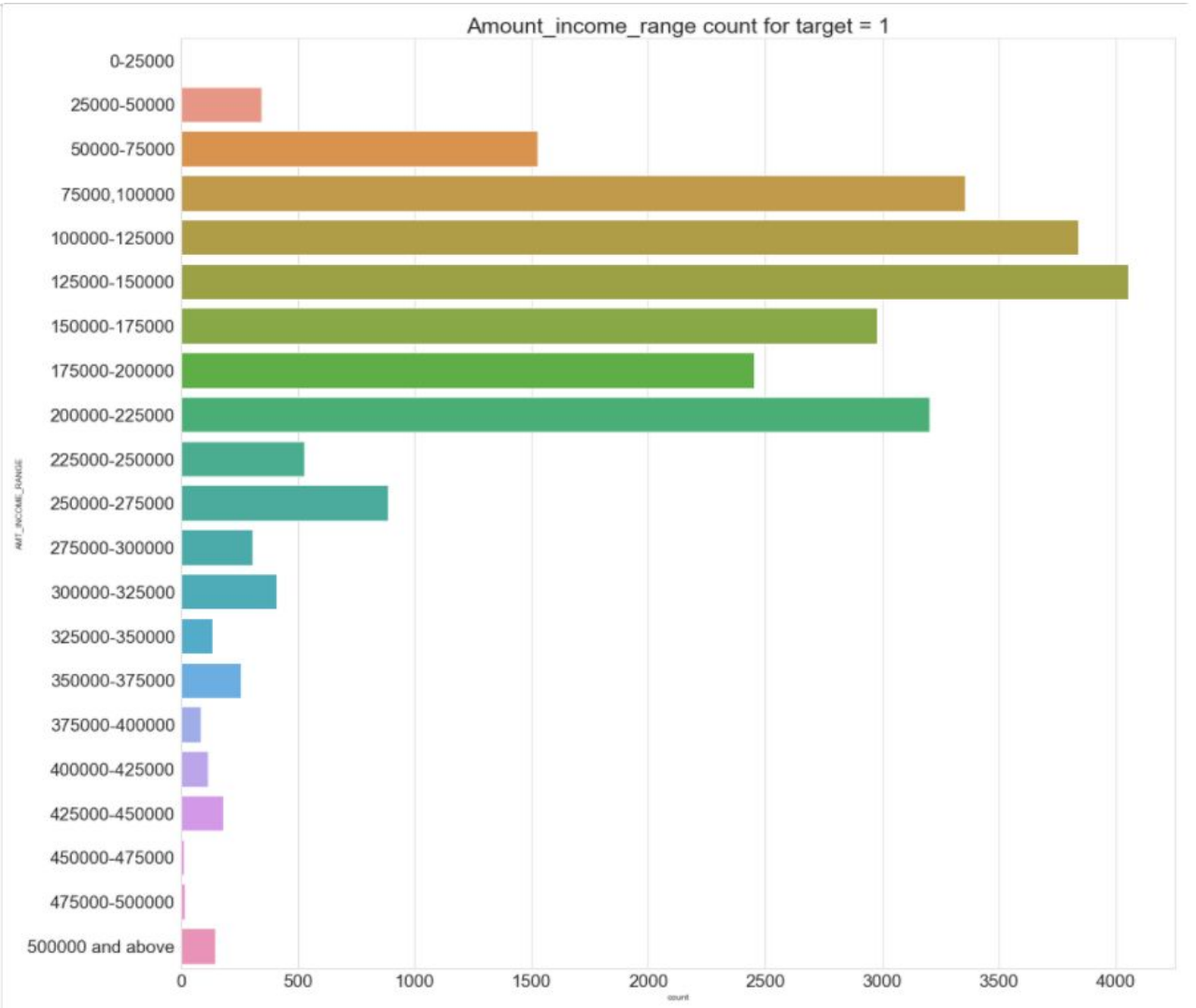

## 15 | Displaying organization type of applicant for applicant without payment difficulties | Univariate Analysis

**Inferences :**

- The applicants having occupation type as Business Entity type 3 are having most difficulty in repaying loan
- Self employed applicants are second on the list of not being able to repay loan. Bank should consider these types of applicants while approving the loan.

## 16 | Displaying the income range of applicant for applicant without payment difficulties | Univariate Analysis






# 17 | Displaying the income range of applicant for applicant without payment difficulties | Univariate Analysis

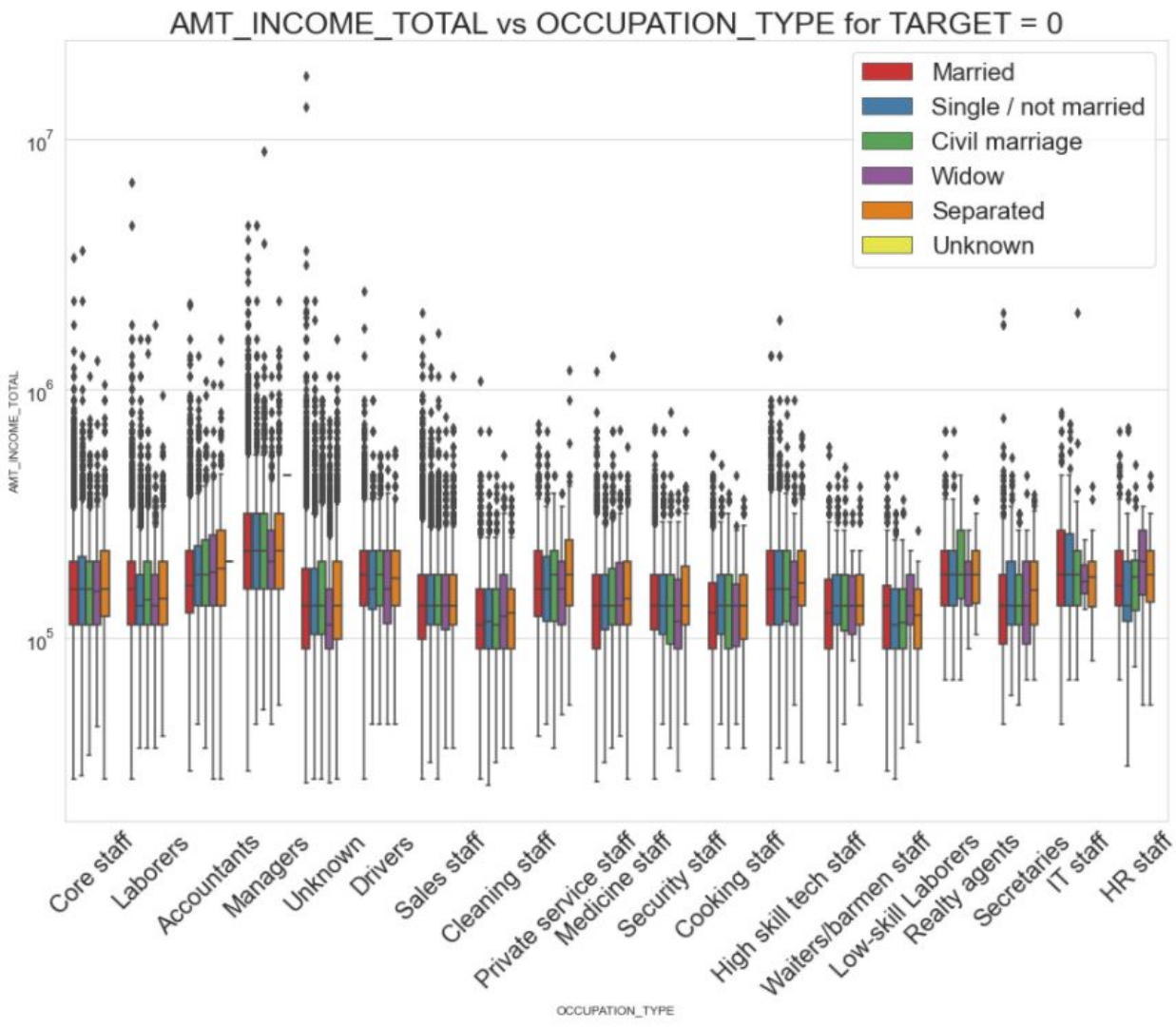
## Inferences :

- Most the people who are not able to repay the loans on time are having income somewhere between 125k to 150k
- Applicants having income between 100k to 125k are also more likely to not repay loans.
- Majority applicants fall between income range of 100k to 150k and 200k to 225k

# 18 | Displaying the total income for the various occupation types sorted by the family status for applicant with no payment difficulties | Bivariate Analysis

## Inferences :

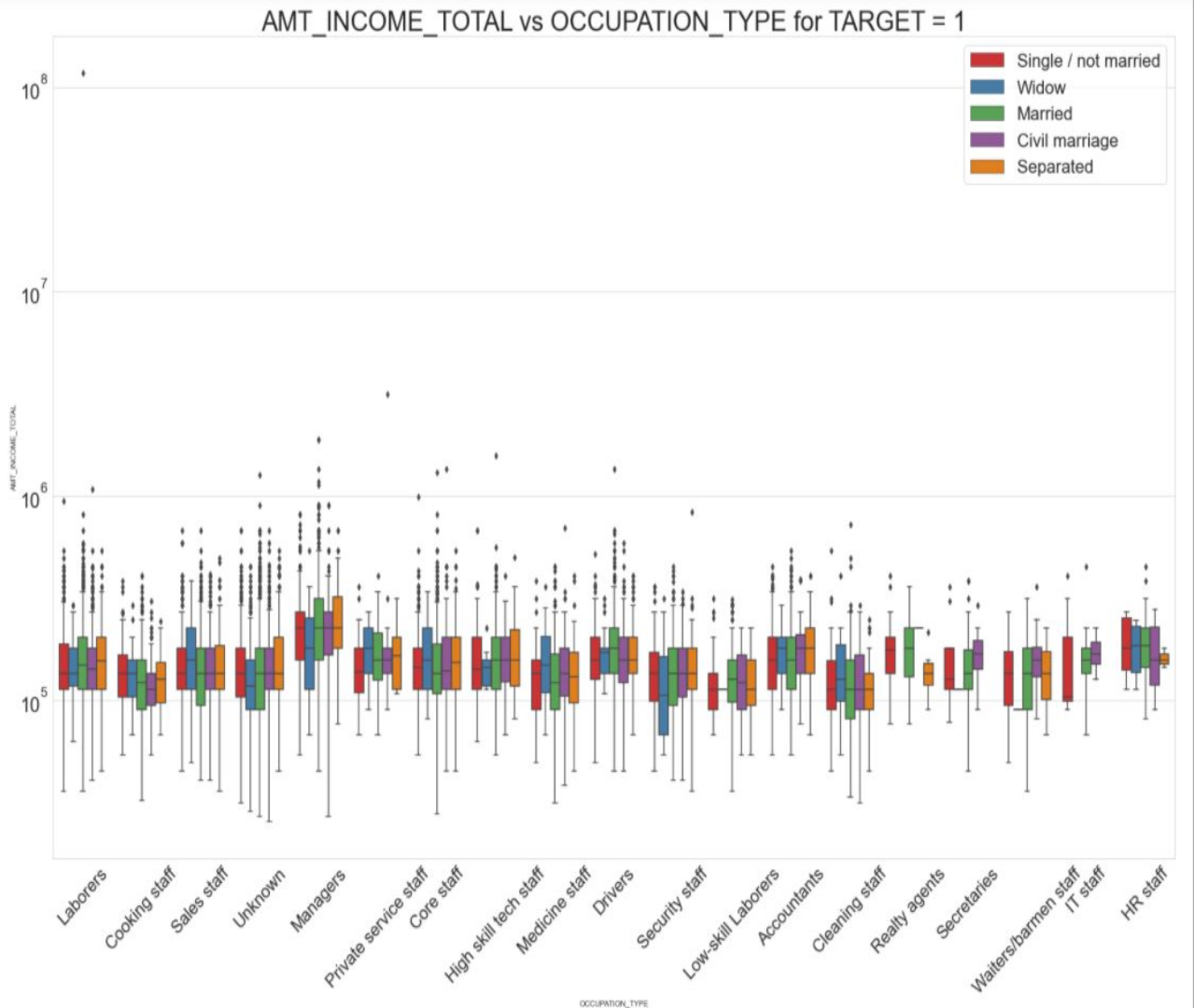
- Managers are having the highest income amount for all family status except widowed for applicants having no payment difficulties
- Accountants,IT staff and HR staff are also having higher income compared to occupation types
- The cleaning staff has the least income amount for all family types(Married,Single,Civil Marriage,Widow,Separated) with all the 25% quartiles falling below 100k income.

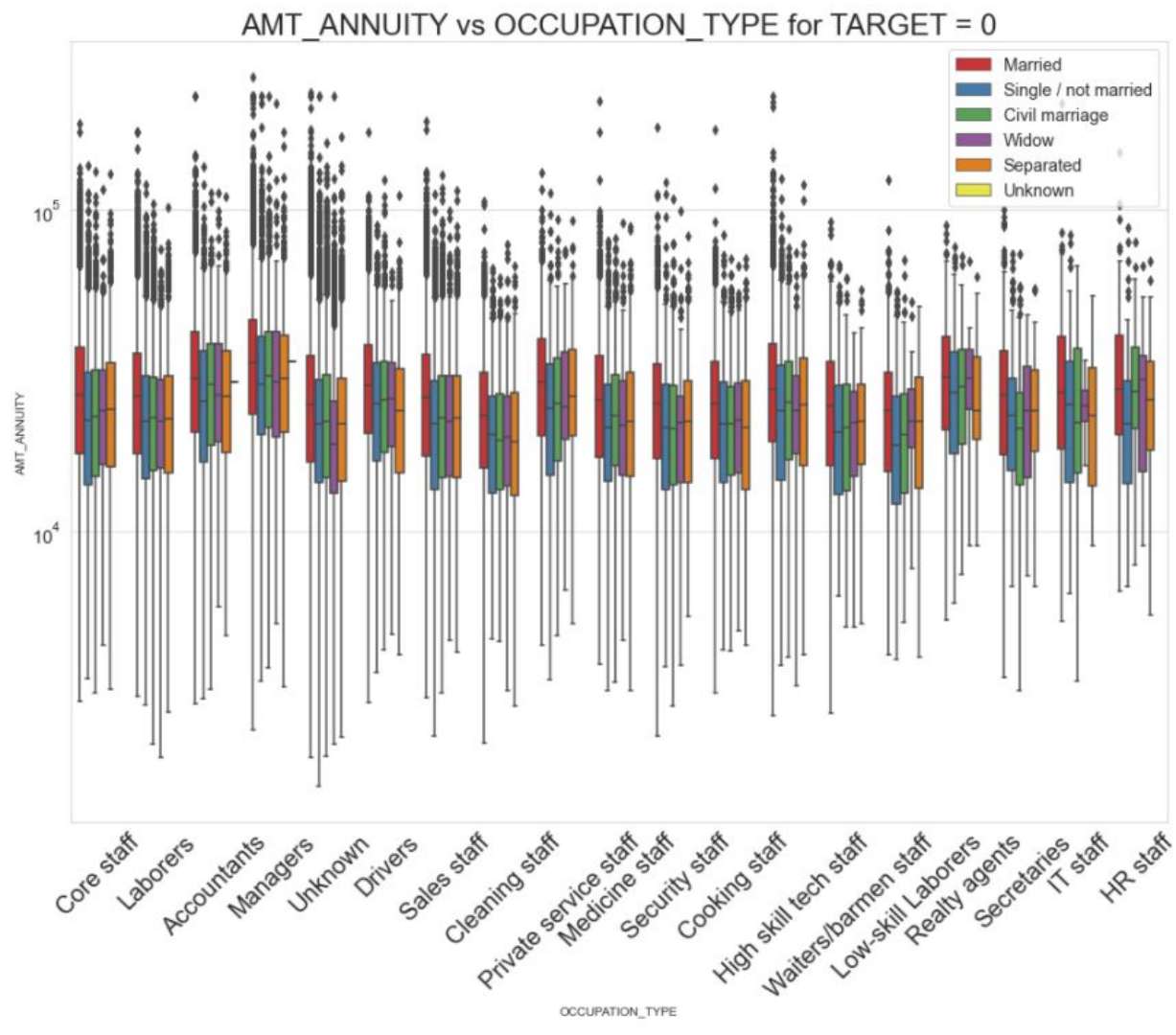


# 19 | Displaying the total income for the various occupation types sorted by the family status for applicant with payment difficulties | Bivariate Analysis

## Inferences :

- Managers are having the highest income amount among rest of the occupation types.
- There are outliers present in fields such as laborers, unknown, managers, core staff and private service staff. Despite their income being very high, these applicants are having difficulty in repaying the loan. This is an area of interest for the bank and further enquiry can be done.
- The income amount for widowed state servants seems to be lowest



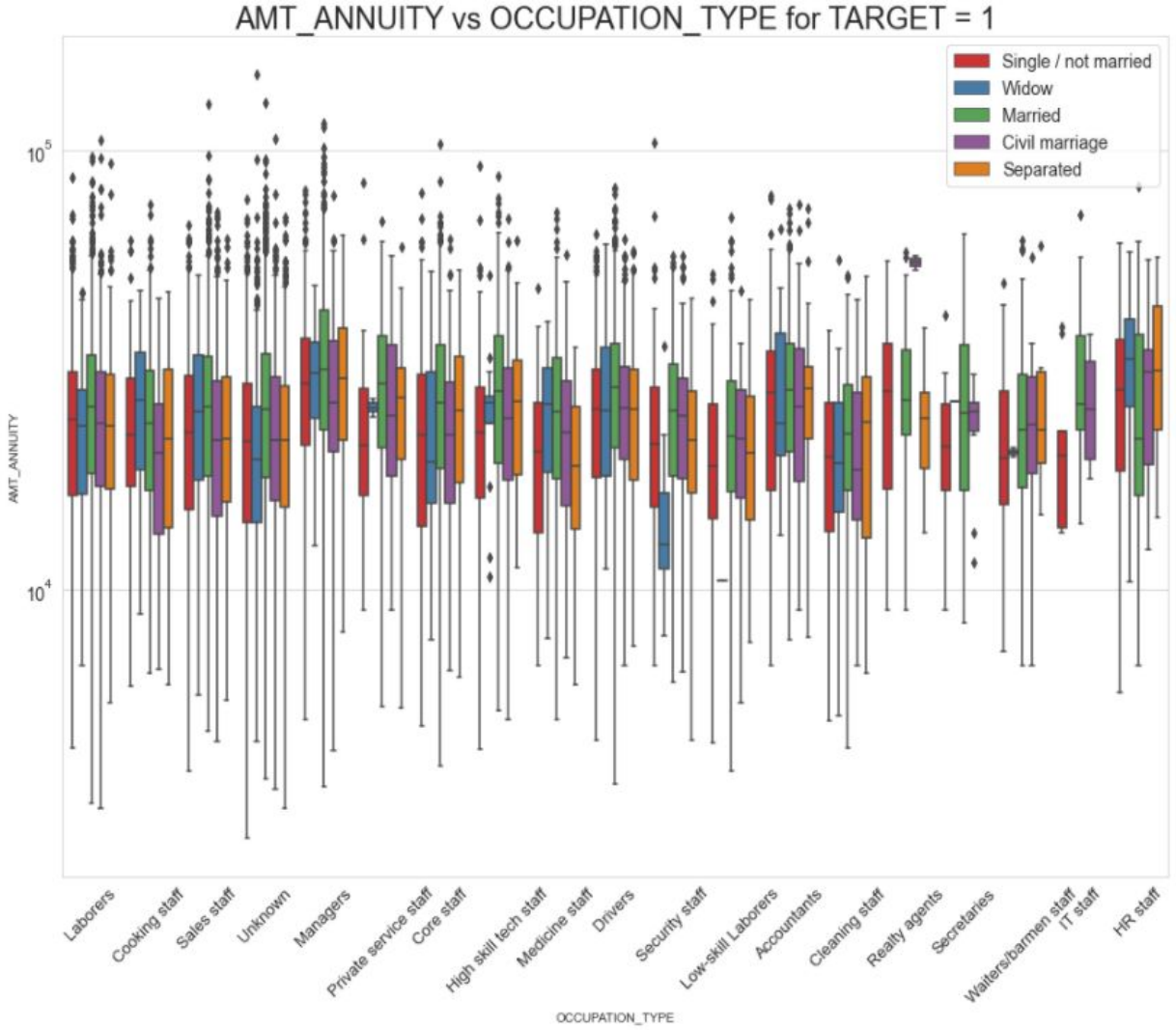


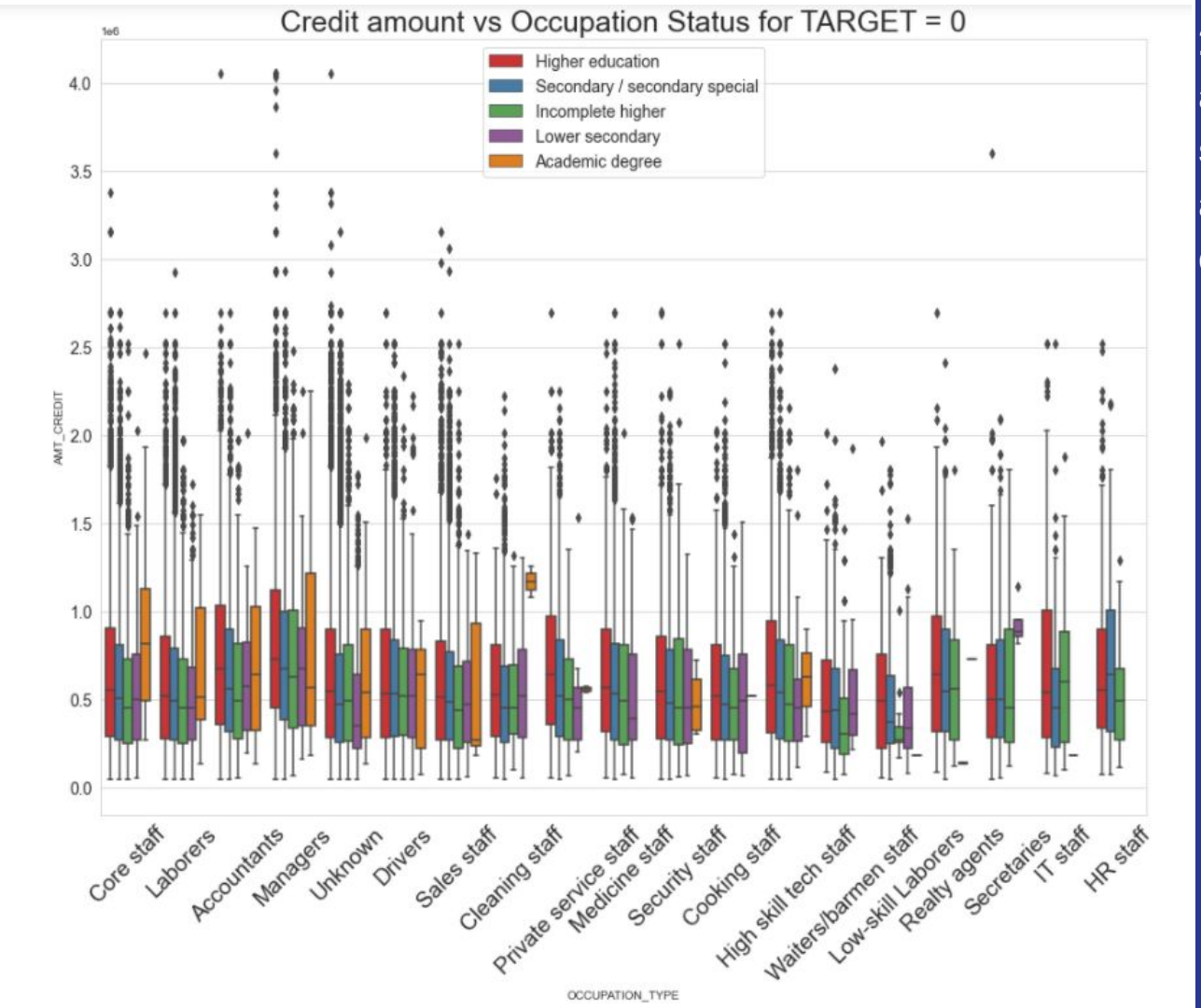

**20 | Displaying the annuity amount for all the occupation types sorted by the family status for all applicants with no payment difficulties | Bivariate Analysis**

# 21 | Displaying the annuity amount for all the occupation types sorted by the family status for all applicants with payment difficulties| Bivariate Analysis

## Inferences :

- The annuity type of all kinds of Managers appears to be the highest among peers.
- HR Staff is at par with Managers having a high annuity amount for all family types.
- Widowed Security staff has the lowest annuity amount as 3 quartiles ( 25,50 and 75) are the lowest.






## 22 | Displaying the credit amount for all occupation types sorted by the education type for all applicants with no payment difficulties| Bivariate Analysis

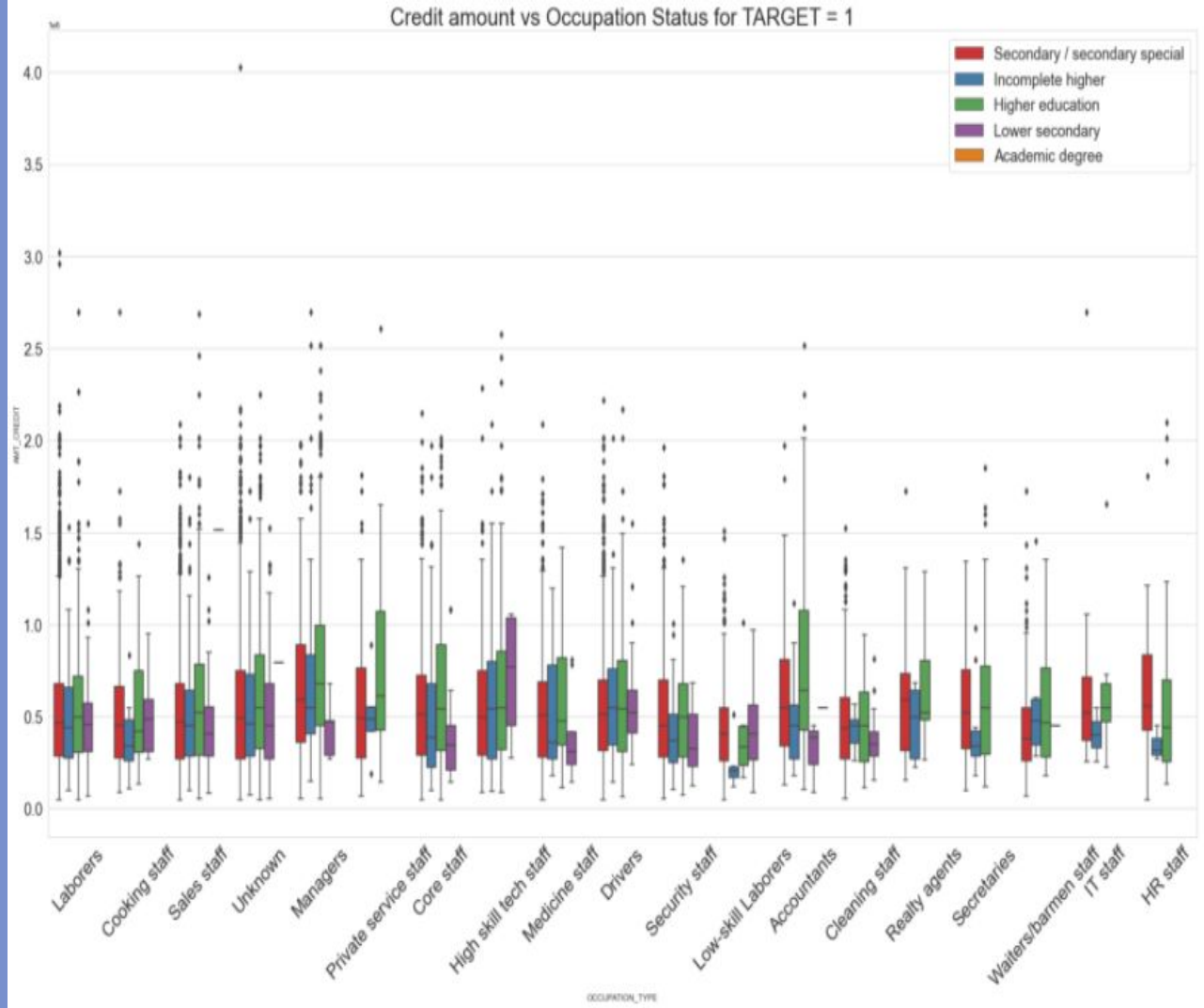
**Inferences :**

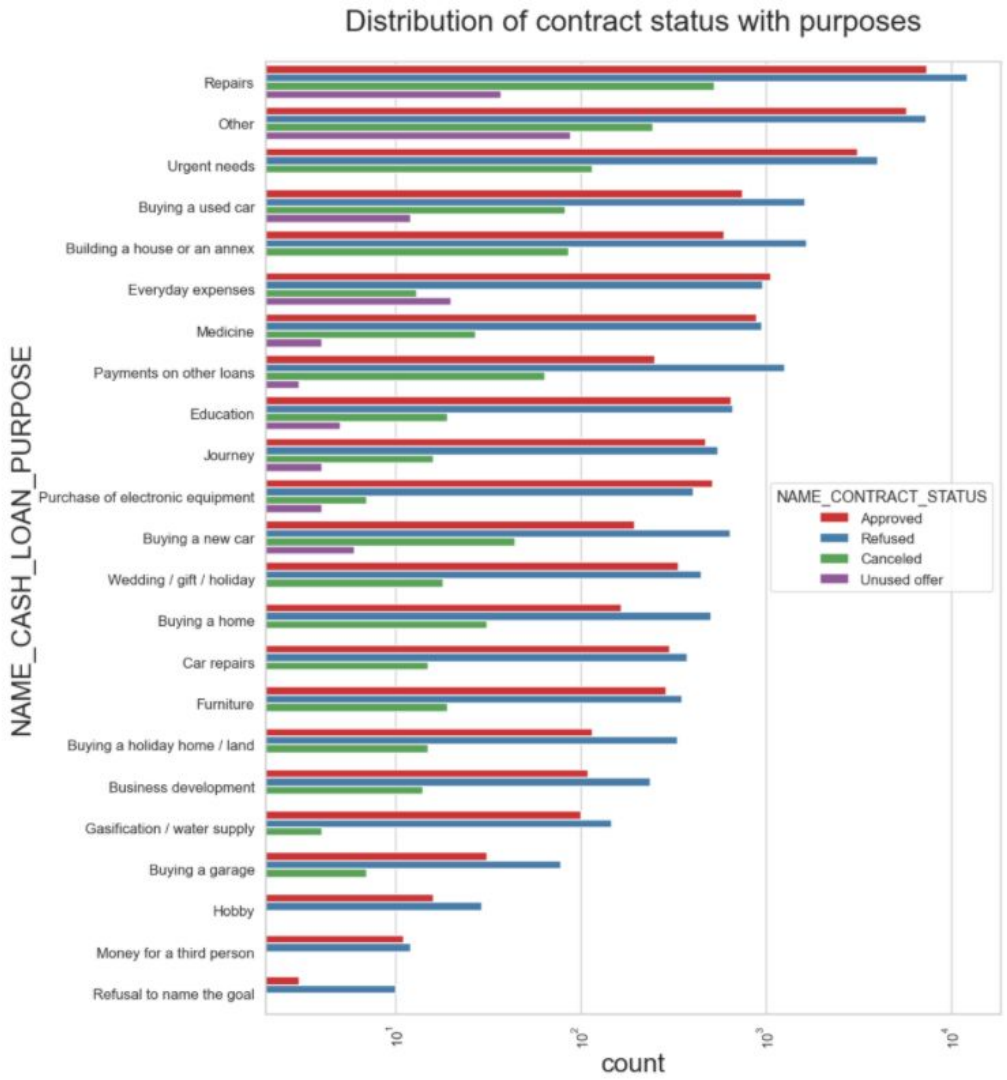
- Managers having higher education have asked for the highest credit amount for loan.
- There are outliers present in managers,unknown,core staff and sales staff.
- The bank should keep in mind the lower income groups and disperse the credit amount accordingly.

# 23 | Displaying the credit amount for all occupation types sorted by the education type for all applicants with payment difficulties| Bivariate Analysis

## Inferences :

- Applicants with an academic degree regardless of their occupation, do not have difficulties in repaying the loan
- There is an unusually high credit amount asked by an applicant with unknown occupation. Bank should be aware of such anomalies.¶
- Accountants having higher education find it difficult to repay loans as their 25th and 75th quartiles are far apart.






## 24 | Displaying the contract status with purposes| Univariate Analysis

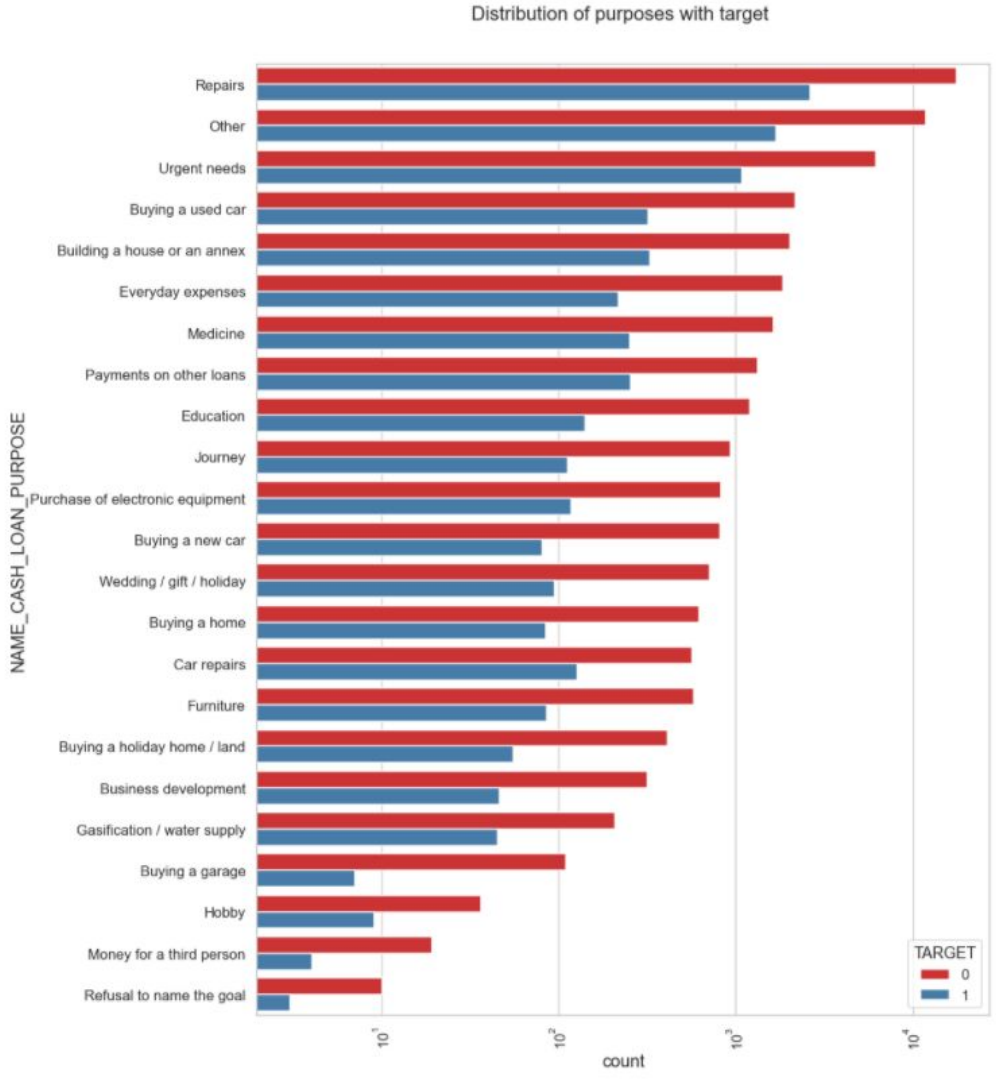
**Inferences :**

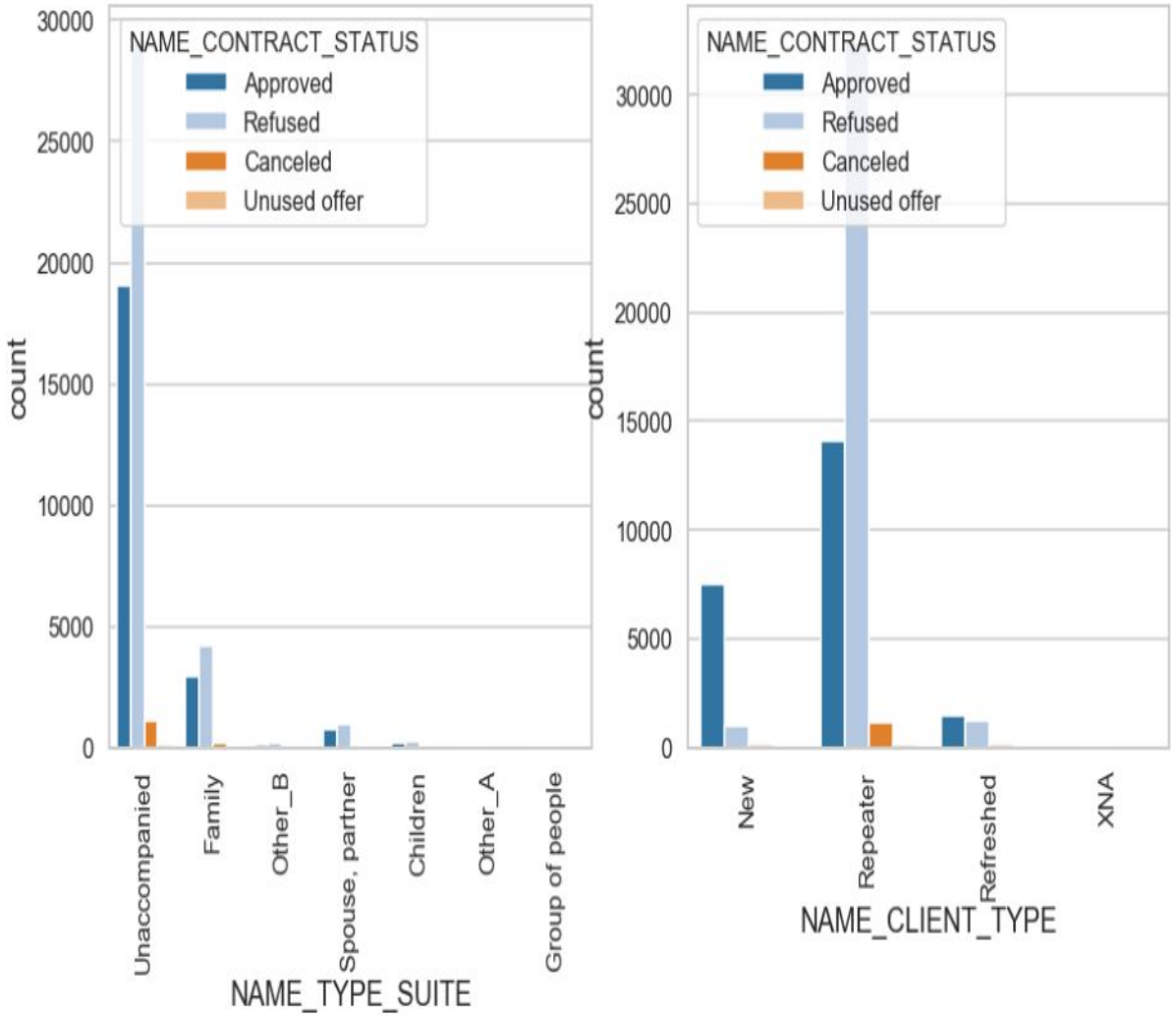
- From the figure , we can see that on an average, the number of loans rejected are more than loans approved.
- Repairs have the most number of loans which are rejected as well as approved
- Everyday expenses and purchase of electronic equipment have their approval rate slightly higher than their rejection rate
- Hobby, Money for a third person and refusal to name goal have no cancelled loans.

# 23 | Displaying the Distribution of purposes with target | Univariate Analysis

## Inferences :

- Applicants who have taken loan for 'Repairs' are facing more difficulties in payment on time.
- Followed by repairs are Buying a used car or buying a house.
- There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land','Buying a new car' and 'Education' Hence bank can focus on these purposes for which the client is having for minimal payment difficulties.









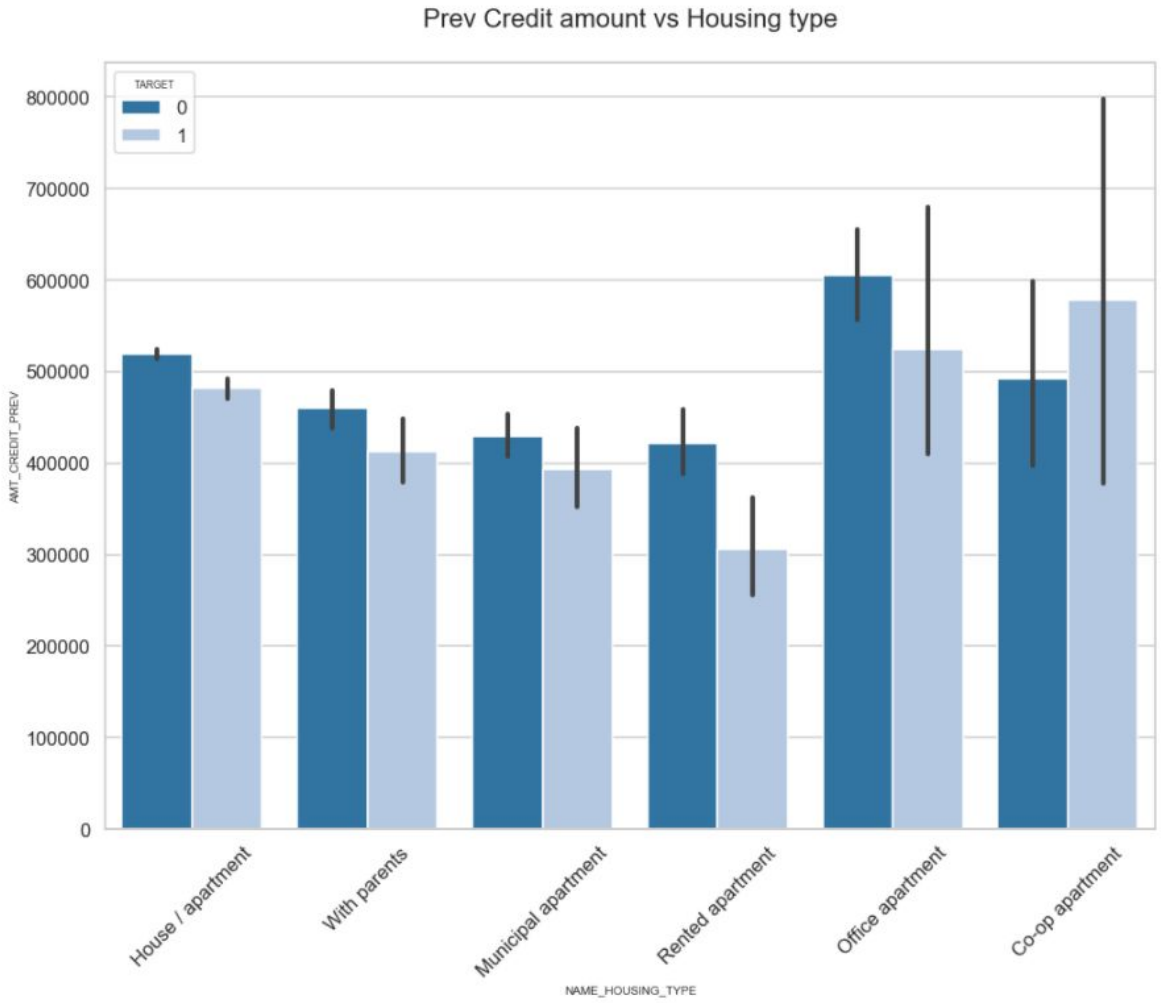
## Inferences :

- Most of the applicants who were applying for loan are unaccompanied.
- Rejection rate is higher for these unaccompanied applicants.
- Next in line are applicants who were accompanied by their family. these applicants also have a rejection rate which is only slightly higher than the approval rate.
- Most of the client types are repeater who also have the highest rejection rate
- New applicants have a significantly higher approval rate than its peers.
- Approval and rejection rate seem to be similar for Refreshed clients

# 25 | Displaying the Prev Credit amount vs Housing type | Bivariate Analysis

## Inferences :

- The applicants having house type as office apartment are less likely to face difficulties while repaying loan. Also the previous credit amount for such applicants is the highest.
- From the figure above we can see that applicants with Co-Op apartment have more difficulties while repaying loan. Hence bank should keep this in mind while lending loan to such applicants.
- Rented apartments have the least difficulty in repaying loan.



# CONCLUSION

1. Bank Should focus on Contract Status "Businessmen", and "Pensioners" for the successful payment.
2. In the housing type other then "Co-op apartments", all other housing types are preferable for the loan as they have least chances of unsuccessful payment.
3. When it comes to income type bank should focus less on "working" type as they have most number of unsuccessful payments.
4. And the loan purpose "Repair" is having higher number of unsuccessful payments
5. Overall on the basis of this dataset analysis, we can say that banks should focus on Contract Status "Businessmen", and "Pensioners" and housing type other then "Co-op apartments" for the successful payment.

# Thank you