Credit EDA Assignment

The procedure followed while carrying out this data analysis are as follows.

1. **Data Cleaning-**

- The columns having more than 40 percent null values were dropped.
- The remaining columns having null values were either replaced with mean/median depending on the skewness or left as it is.
- Then the irrelevant, dependent or variables with a large proportion of missing values are dropped.
- The data was then checked for any outliers.
- The values of certain columns were standardized.
- Some of the columns were binned and converted to categorical data.
- Finally, the data was checked for any null values or wrong data types.

2. Univariate analysis

- The various variables were plotted to get better understanding of the data.

3. Bivariate analysis

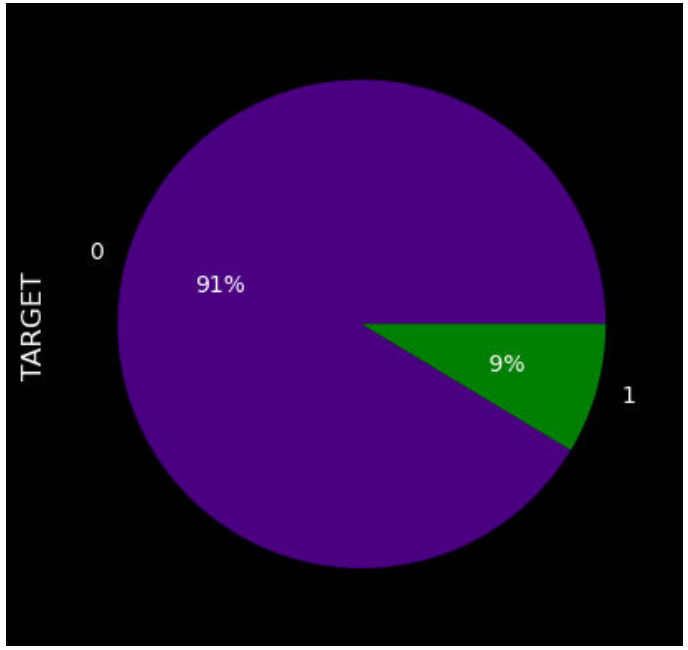
- The variables are then plotted the against the target variable to check for any kind of relation between them.

4. Correlation matrix of target and all other cases were plotted.

5. Variables that drive the target and top 10 correlation for target and all other cases are derived.

DATA IMBALANCE

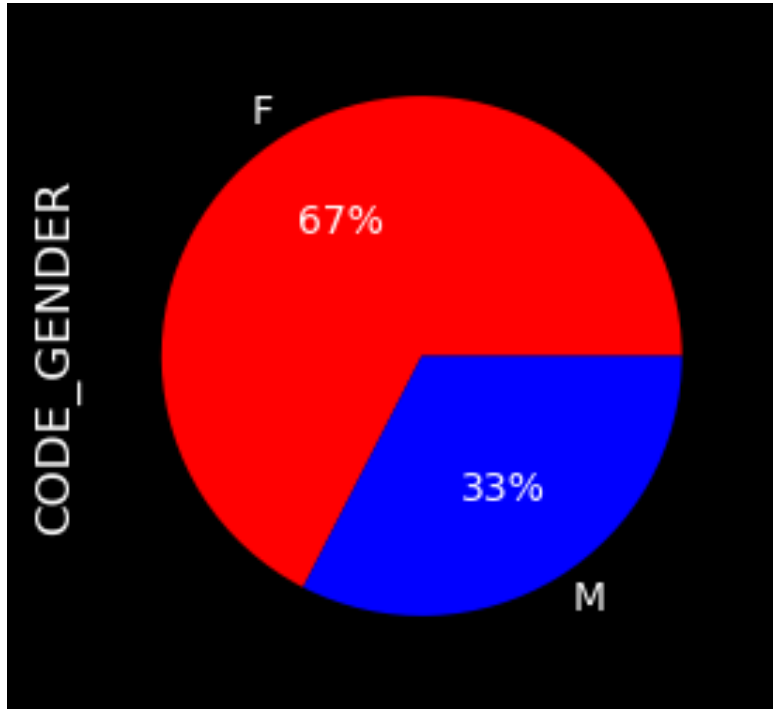
An imbalance ratio of 10.59 was observed in this data



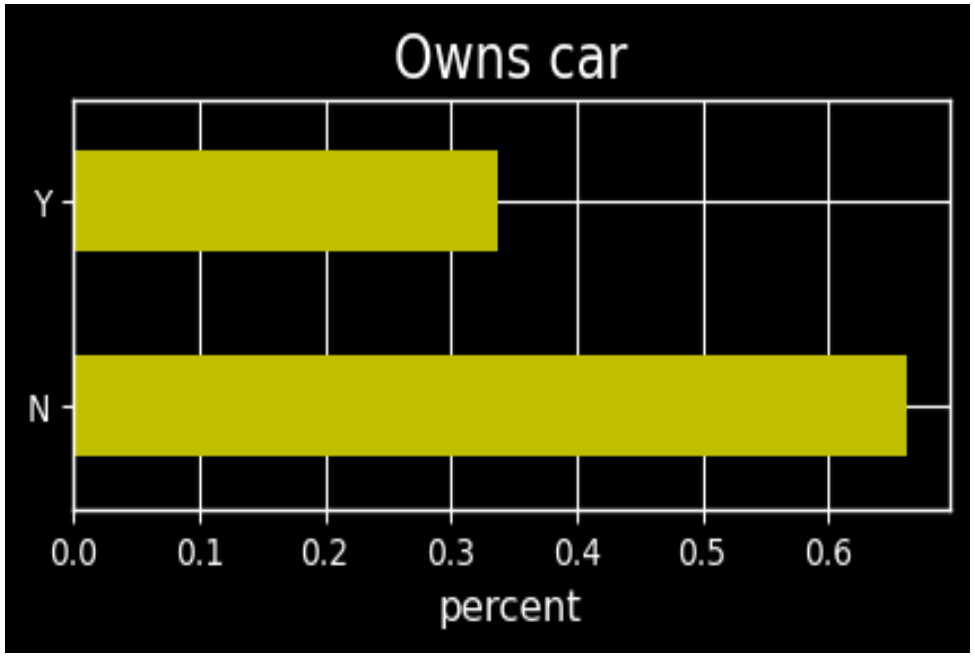
- **UNIVARIATE ANALYSIS**

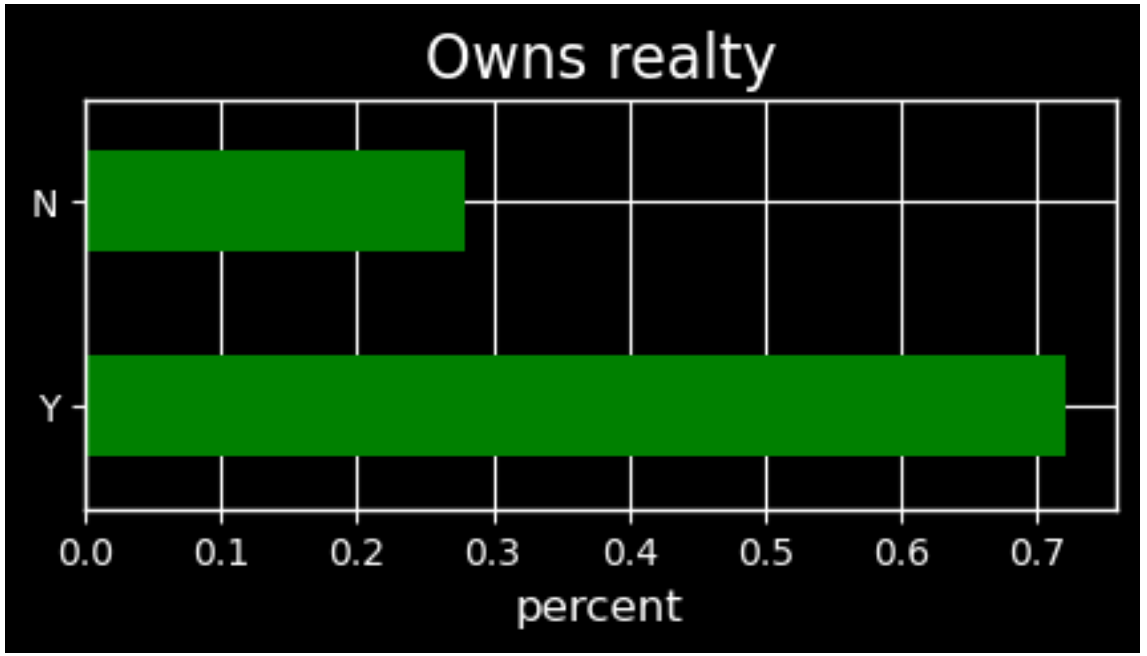
Some notable facts discovered from univariate analysis are presented below.

1. 67% OF THE CLIENTS ARE FEMALE WHILE 33% ARE MALES

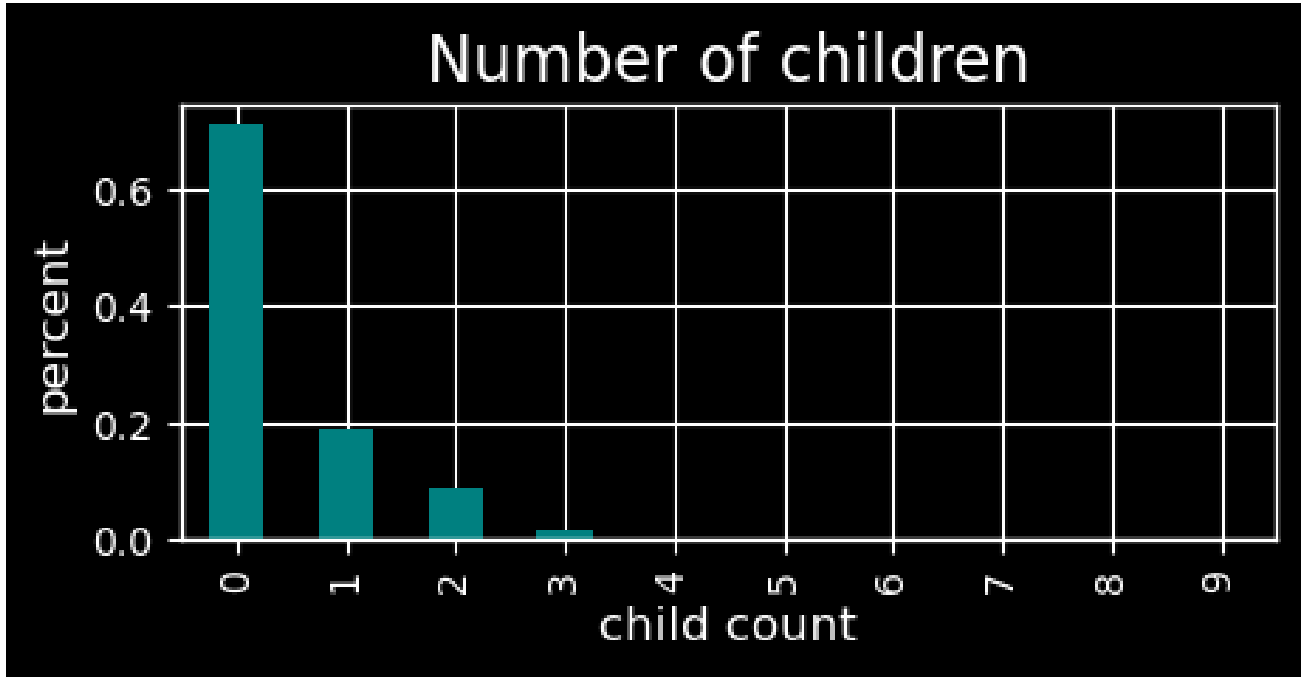


2. MAJORITY OF THE CLIENTS OWN CAR AND REALTY.

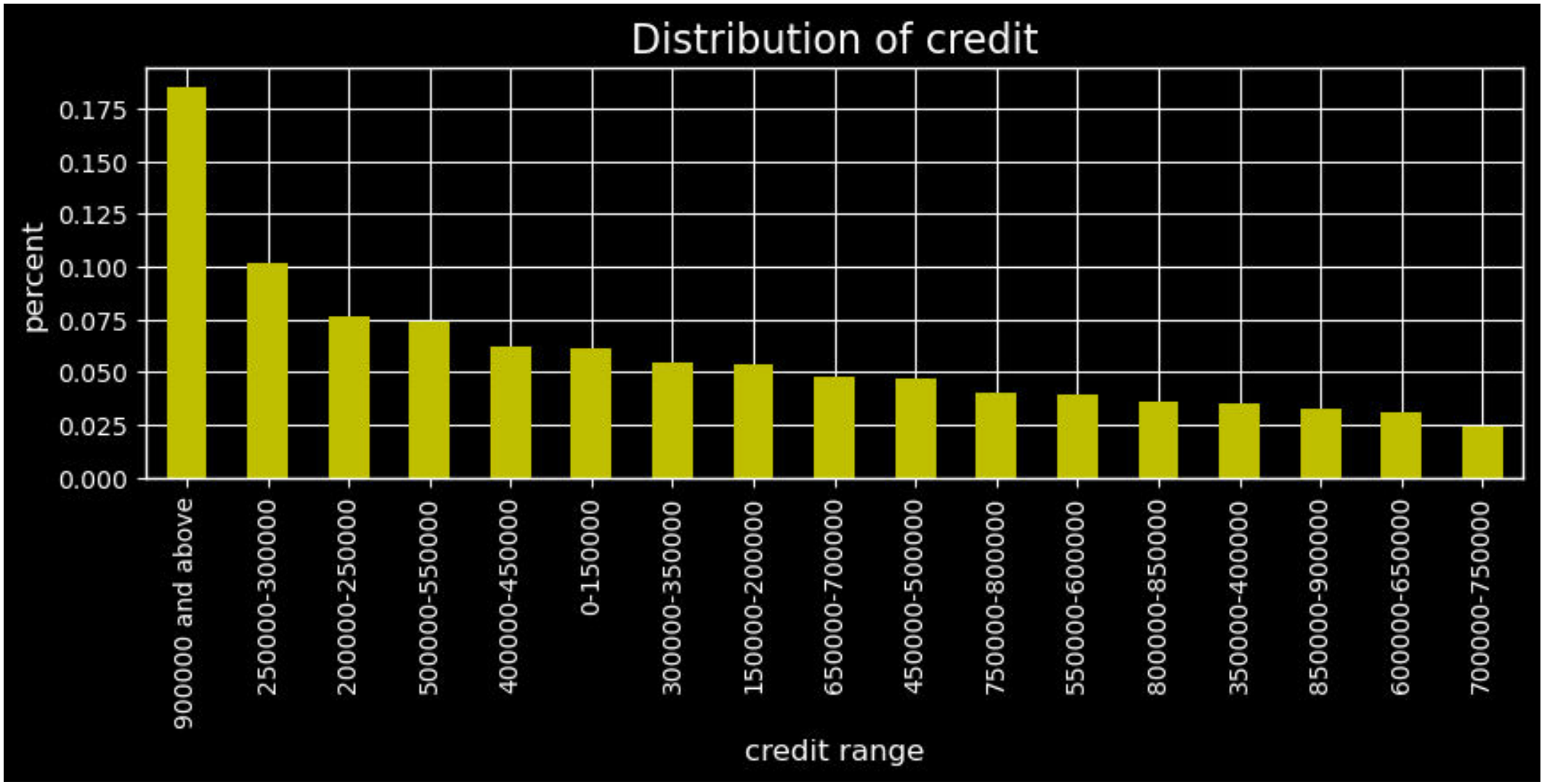




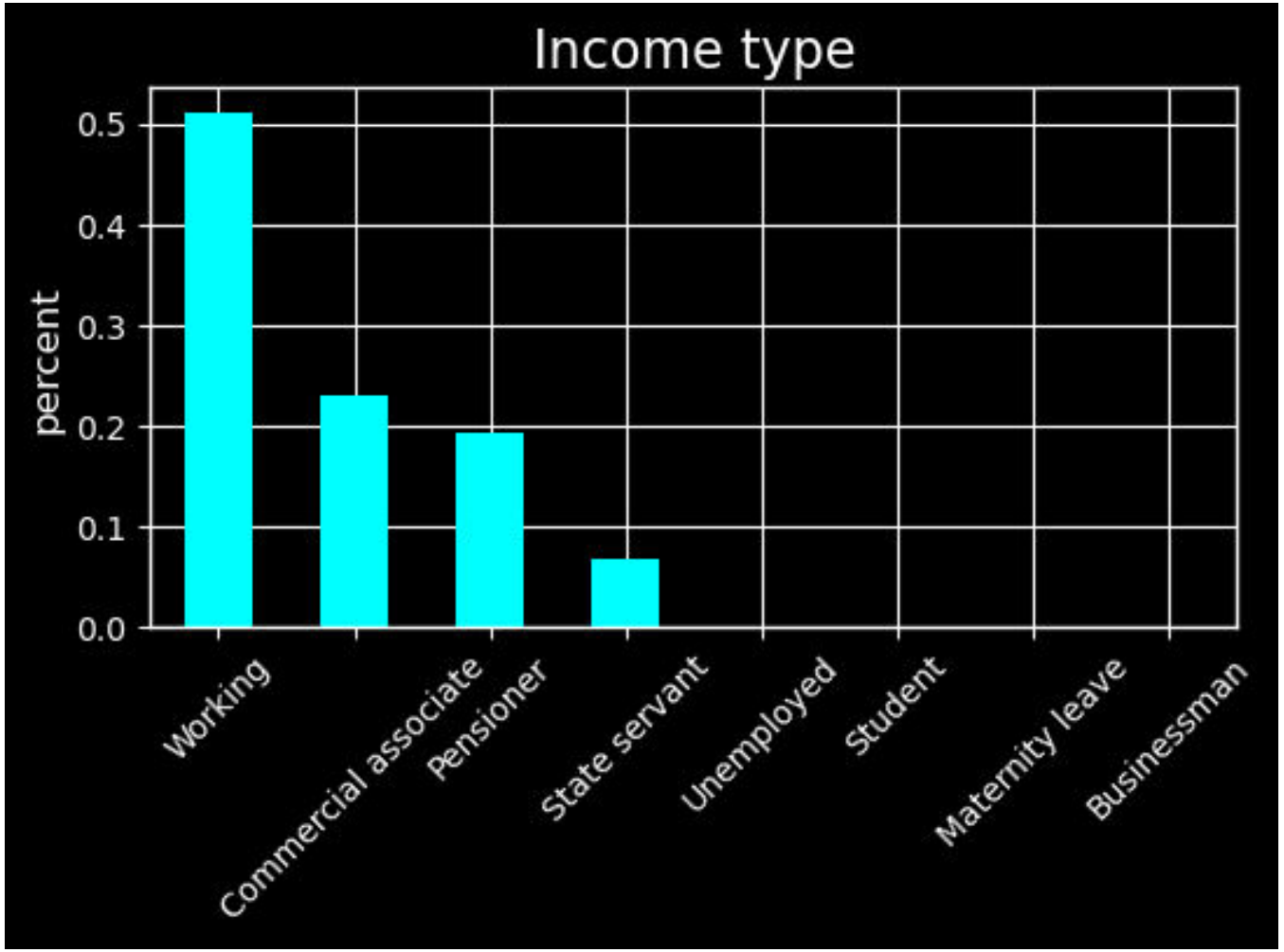
3. ALMOST 70 PERCENT OF THE CLIENTS HAS NO CHILD.



4. AMONG THE CREDIT ASKED , 9LAKHS AND ABOVE TOPS THE CHART

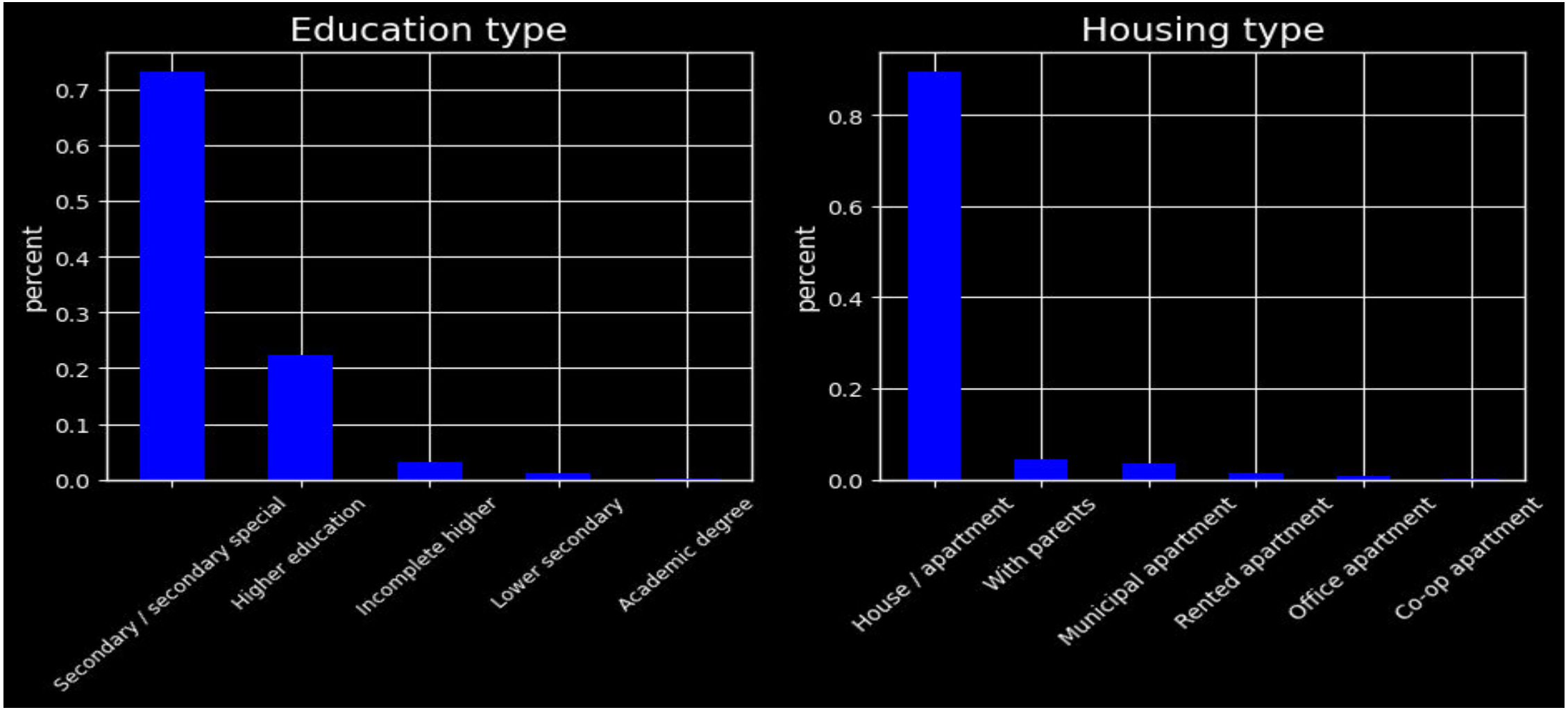


5. THE MAJORITY OF THE CLIENTS ARE WORKING(51%) OR COMMERCIAL ASSOCIATE(22%)



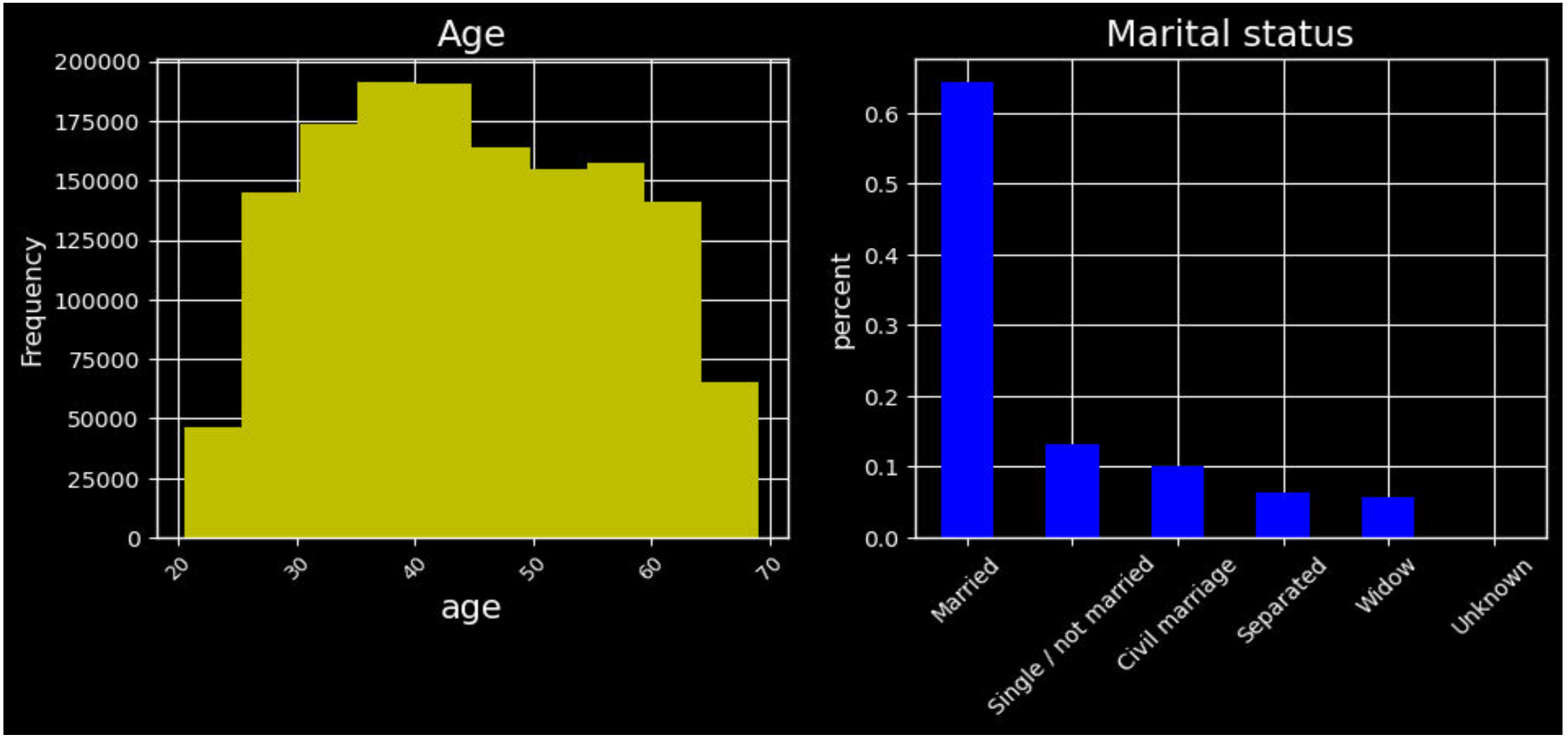
6. MOST OF THE CLIENTS HAVE SECONDARY EDUCATION(75%),FOLLOWED BY HIGHER EDUCATION(22%),AND VERY FEW HAVE ACADEMIC DEGREE.

7. MAJORITY OF THE CLIENTS LIVE IN HOUSE/APARTMENTS(91%)

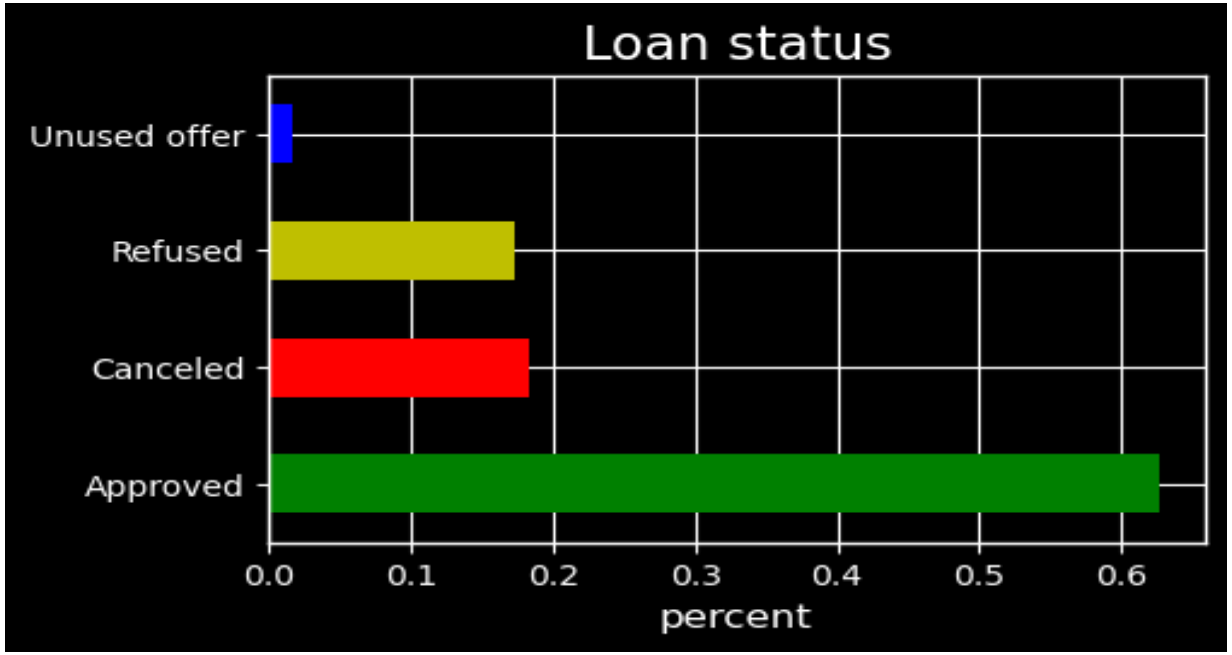


8. AGE IS NORMALLY DISTRIBUTED AND PEAKS AT 35-45 YEARS.

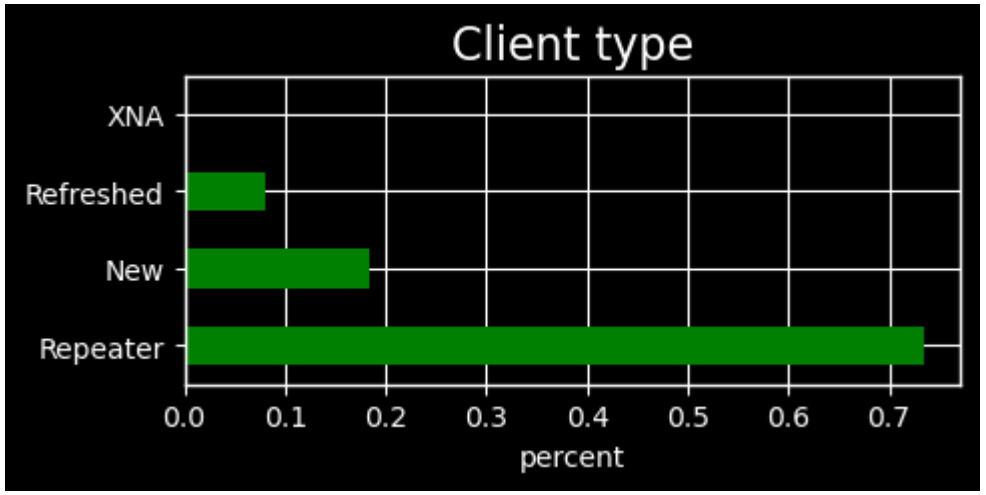
9. MAJORITY OF THE CLIENTS ARE MARRIED (66%)



10. MOST OF THE PREVIOUS LOAN APPLICATION WERE APPROVED



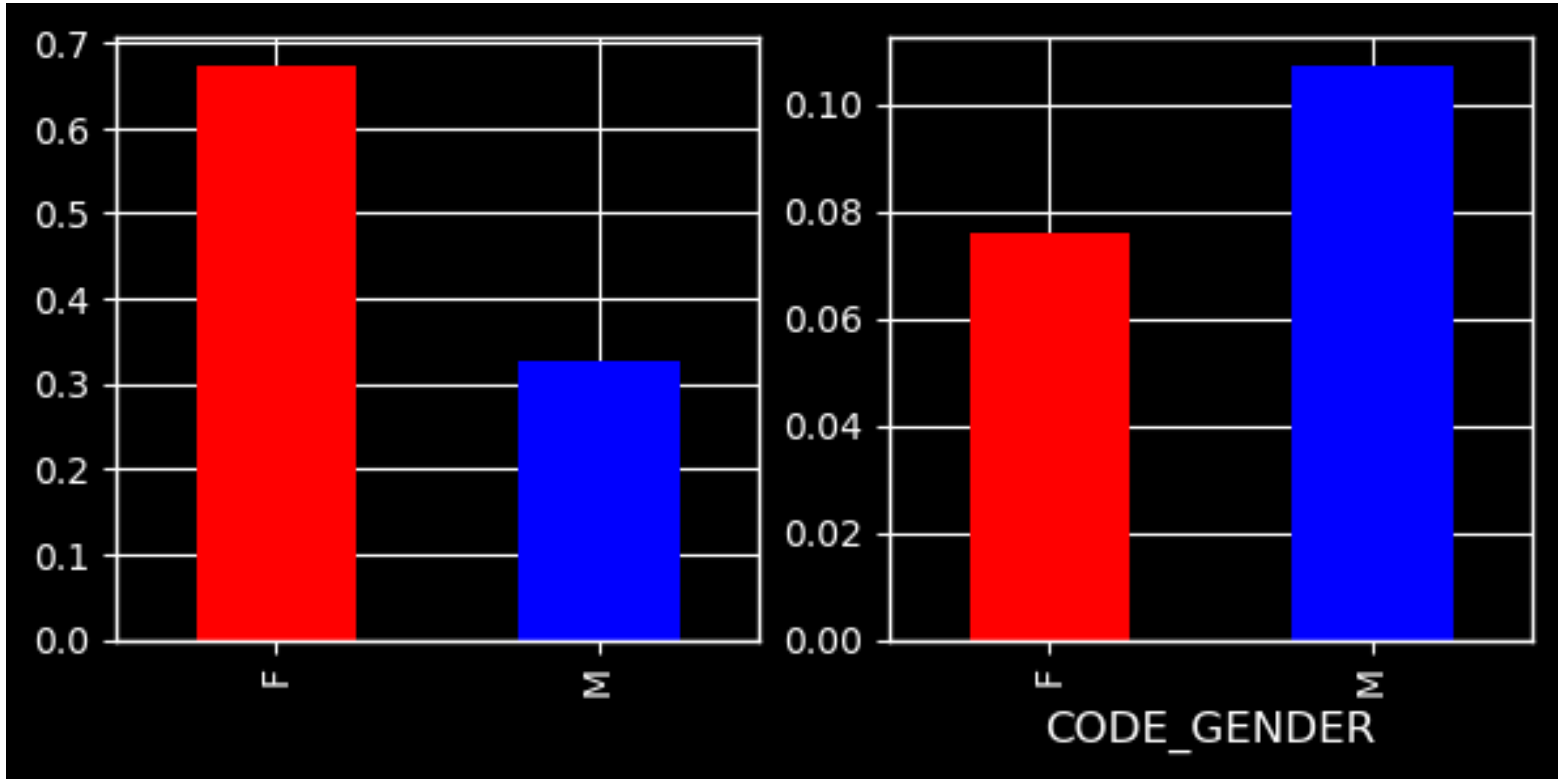
11. MOST OF THE CLIENTS WERE REPEATERS IN CASE OF PREVIOUS LOAN APPLICATION.



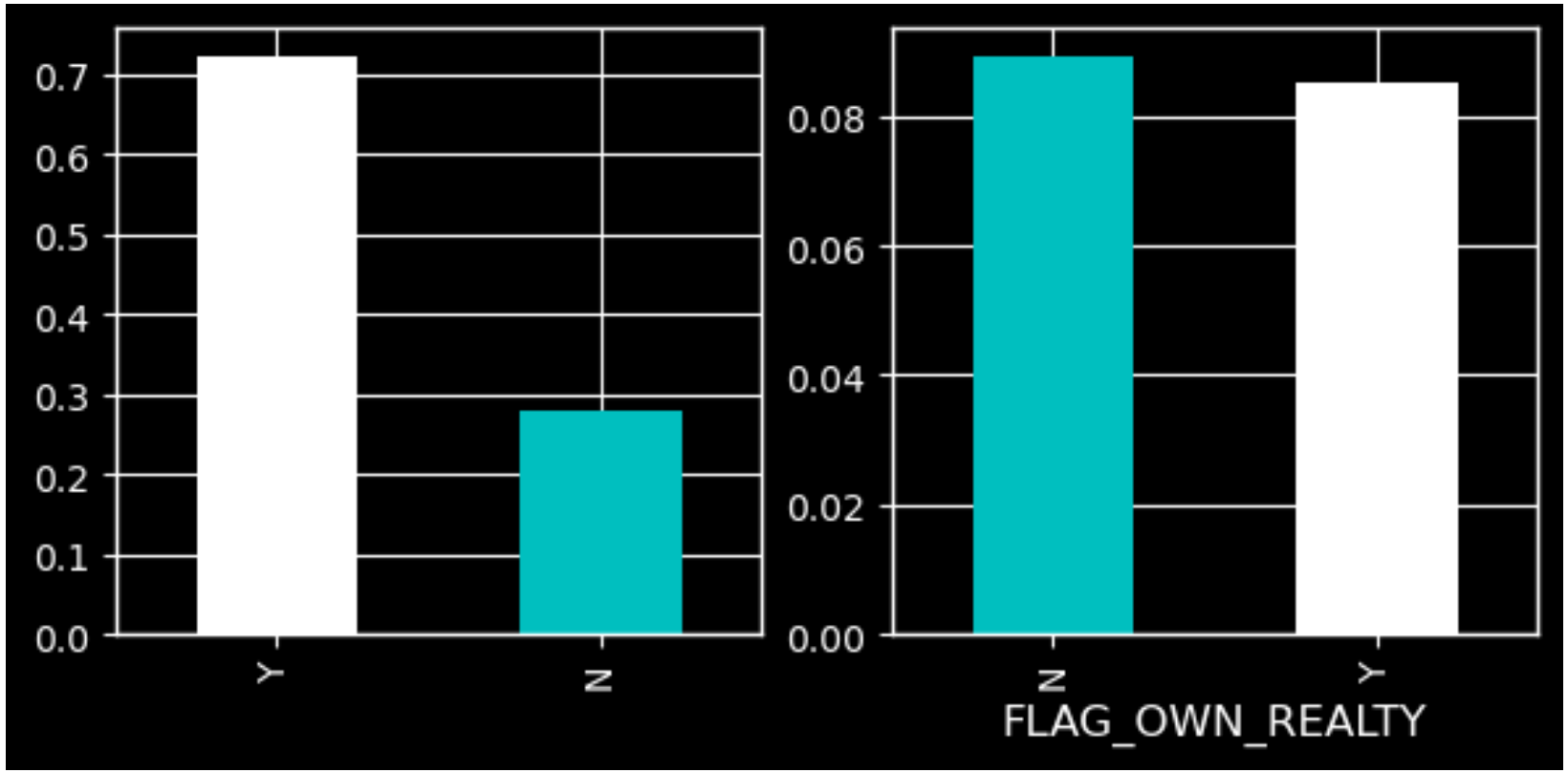
- **BIVARIATE ANALYSIS**

Note: The left graph shows the percentage distribution of the variable & right graph the percentage difficulty in repaying the loan.

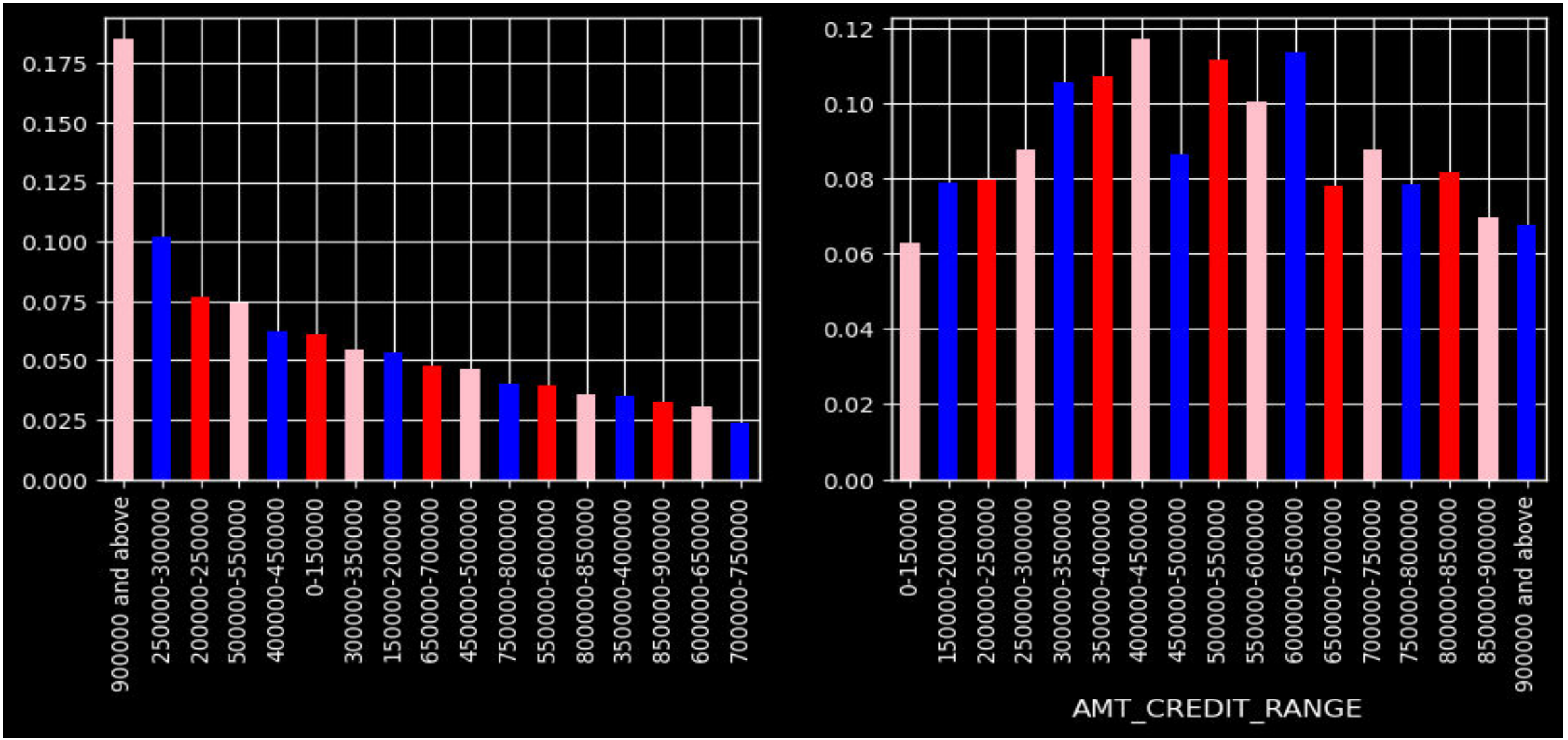
1. ALTHOUGH FEMALE CLIENTS ARE MORE BUT IT THE MALE GENDER WHO IS FACING MORE DIFFICULTY IN REPAYING LOAN COMPARED TO FEMALES.



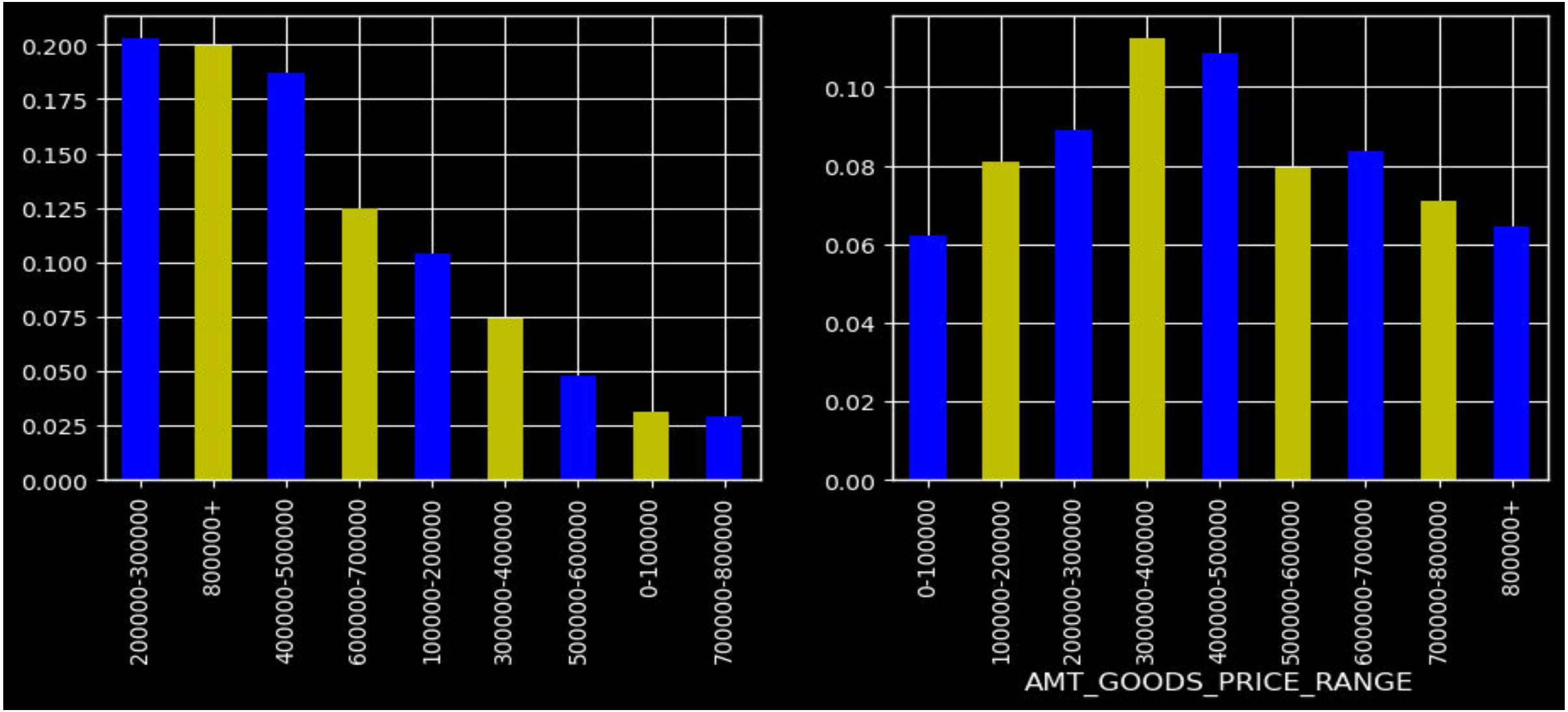
2. MOST OF THE CLIENTS HAVE OWN REALTY (71%),AND THE ONES WHO DON'T ARE FACING MORE DIFFICULTY IN REPAYING THE LOANS.



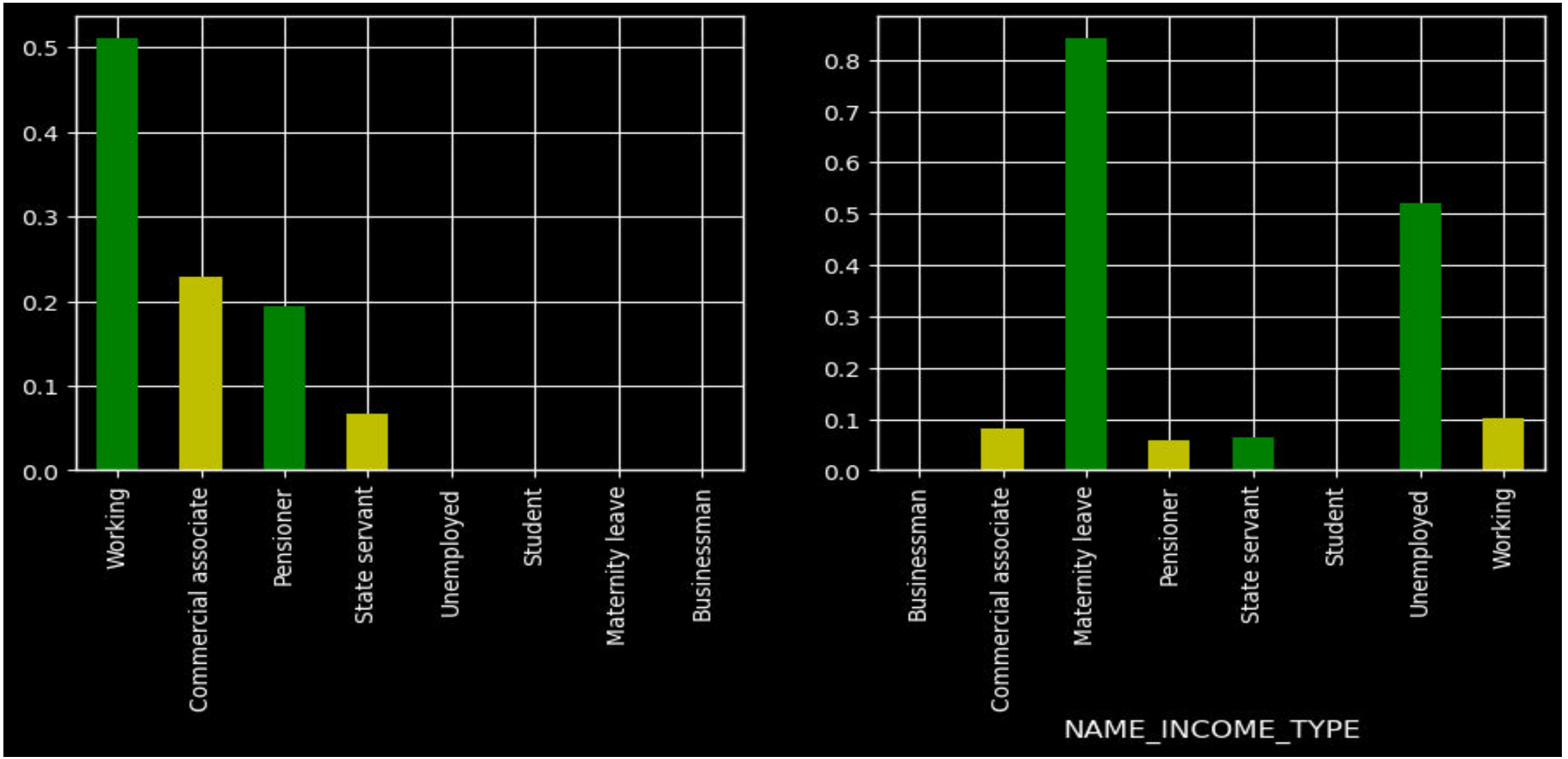
3. AS THE CREDIT AMOUNT INCREASES THE DIFFICULTY INCREASES AND ABOVE 6,50,000 CREDIT THE DIFFICULTY SEEMS TO GET LESS.



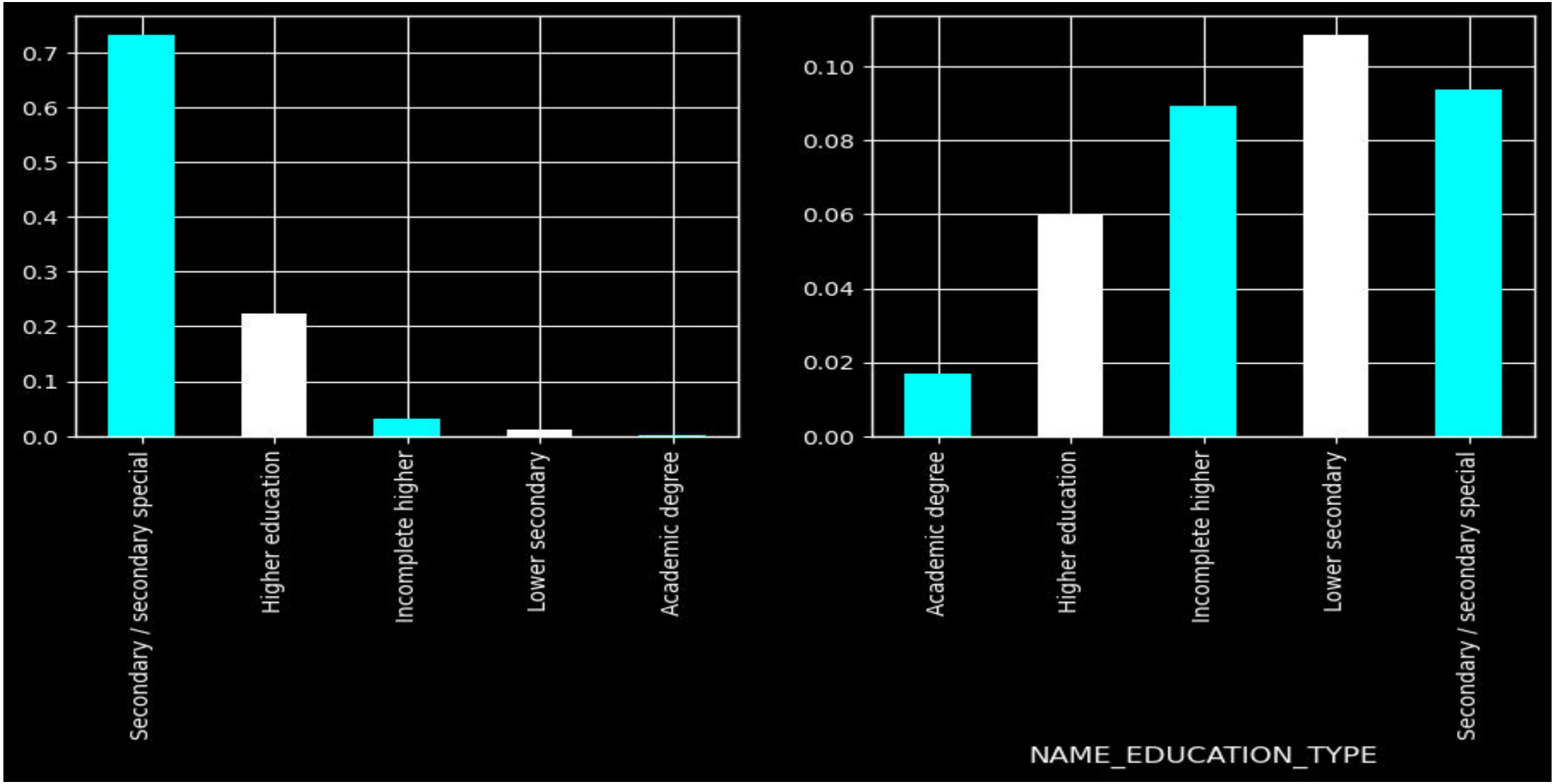
4. ONE INTERESTING THING TO SEE HERE IS THAT THE ONES HAVING LESS DIFFICULTY IN REPAYING LOANS HAVE APPLIED FOR GOODS IN THE LOWER PRICE RANGE AND THE GOODS IN THE HIGHER PRICE RANGE. THE MID PRICED GOODS APPLIERS ARE GETTING MORE DIFFICULTY IN REPAYING



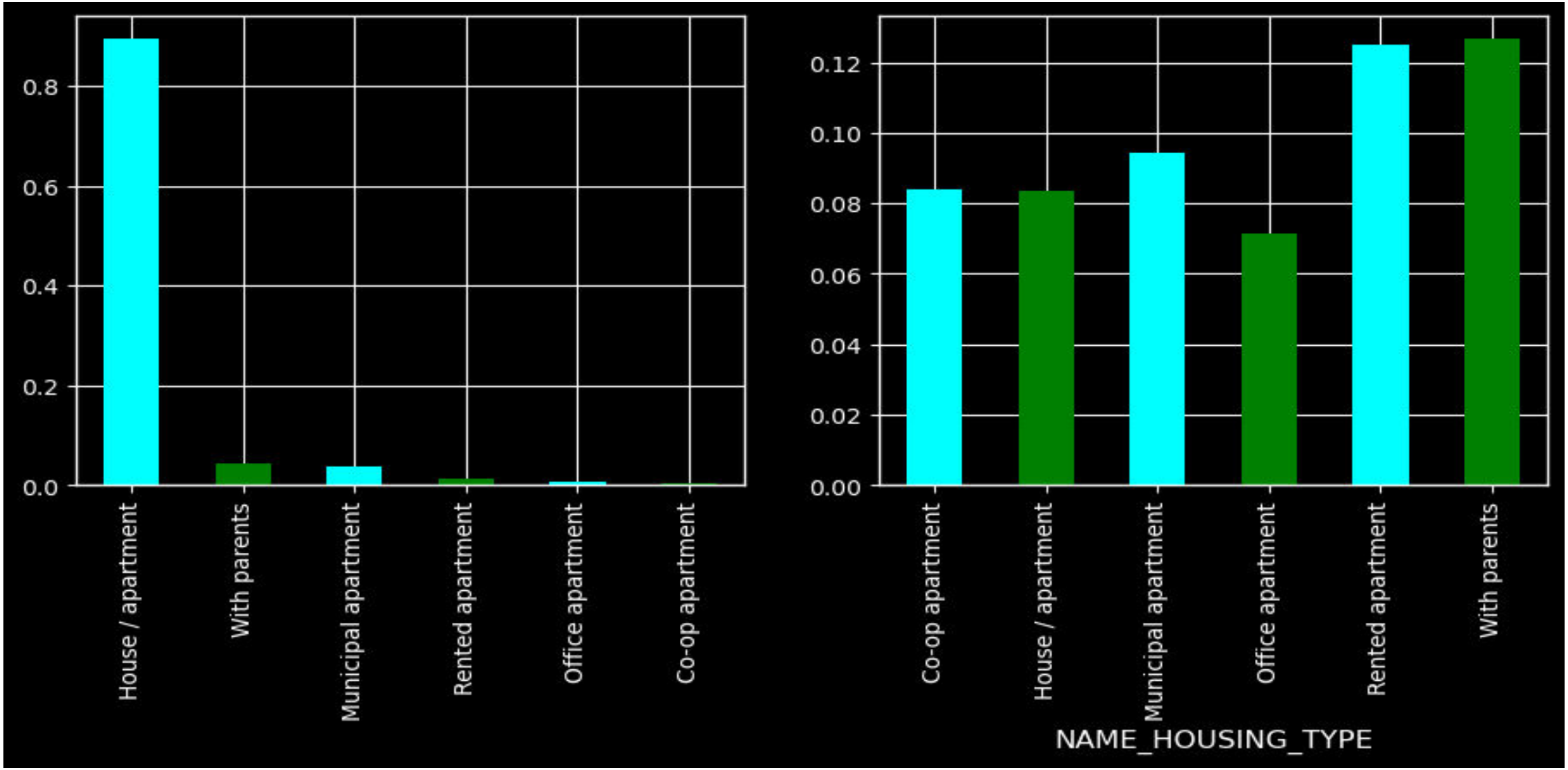
5. UNEMPLOYED AND MATERNITY LEAVE CLIENTS ARE HAVING DIFFICULTY IN PAYING THE LOANS.



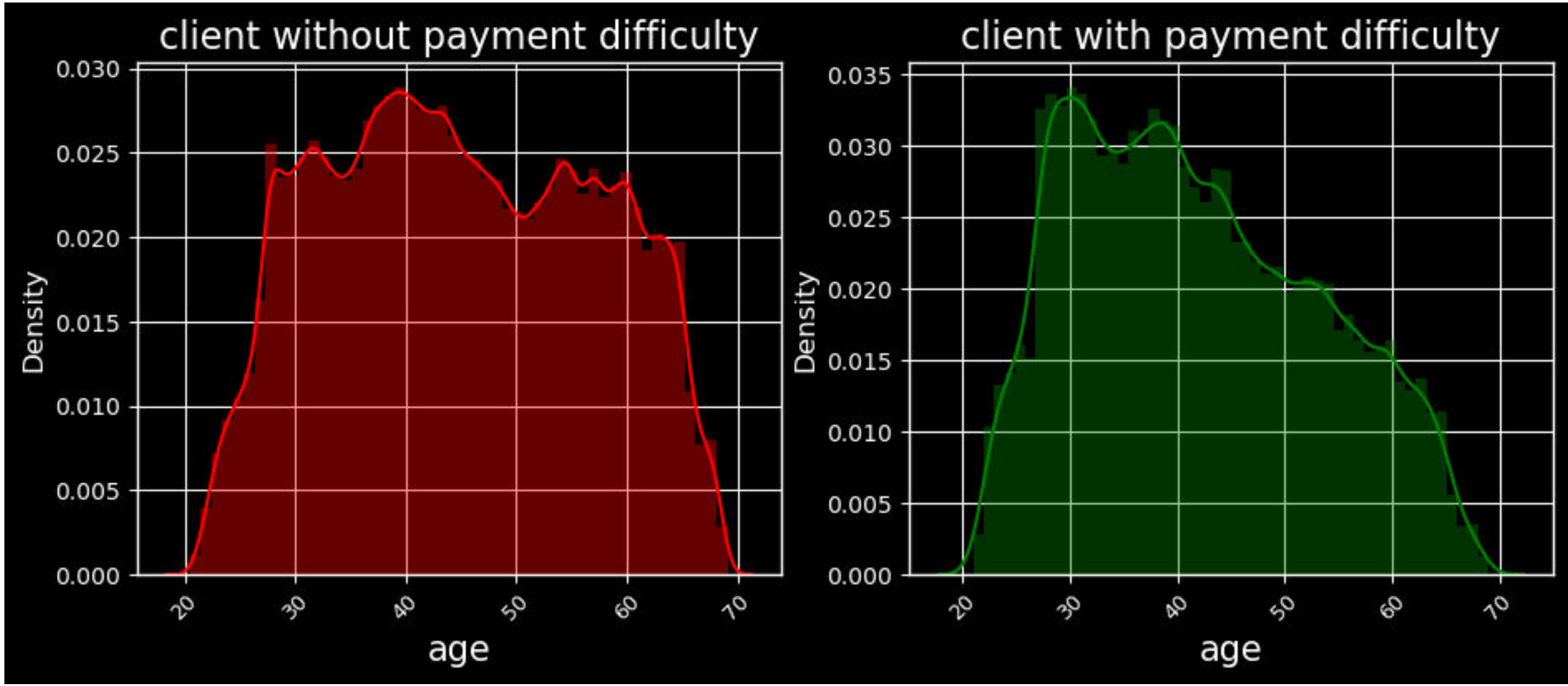
6. HERE WE CAN SEE THE ABILITY TO REPAY LOAN IS DIRECTLY PROPORTIONAL TO THE EDUCATION LEVEL. THE LOAN PAYING DIFFICULTY ORDER IS LOWER SEC> SECONDARY> INCOMPLETE HIGHER> HIGHER> ACADEMIC



7. THE ONES LIVING WITH PARENTS OR RENTED APARTMENTS ARE HAVING THE MAXIMUM PROBLEM OF LOAN REPAYMENT.

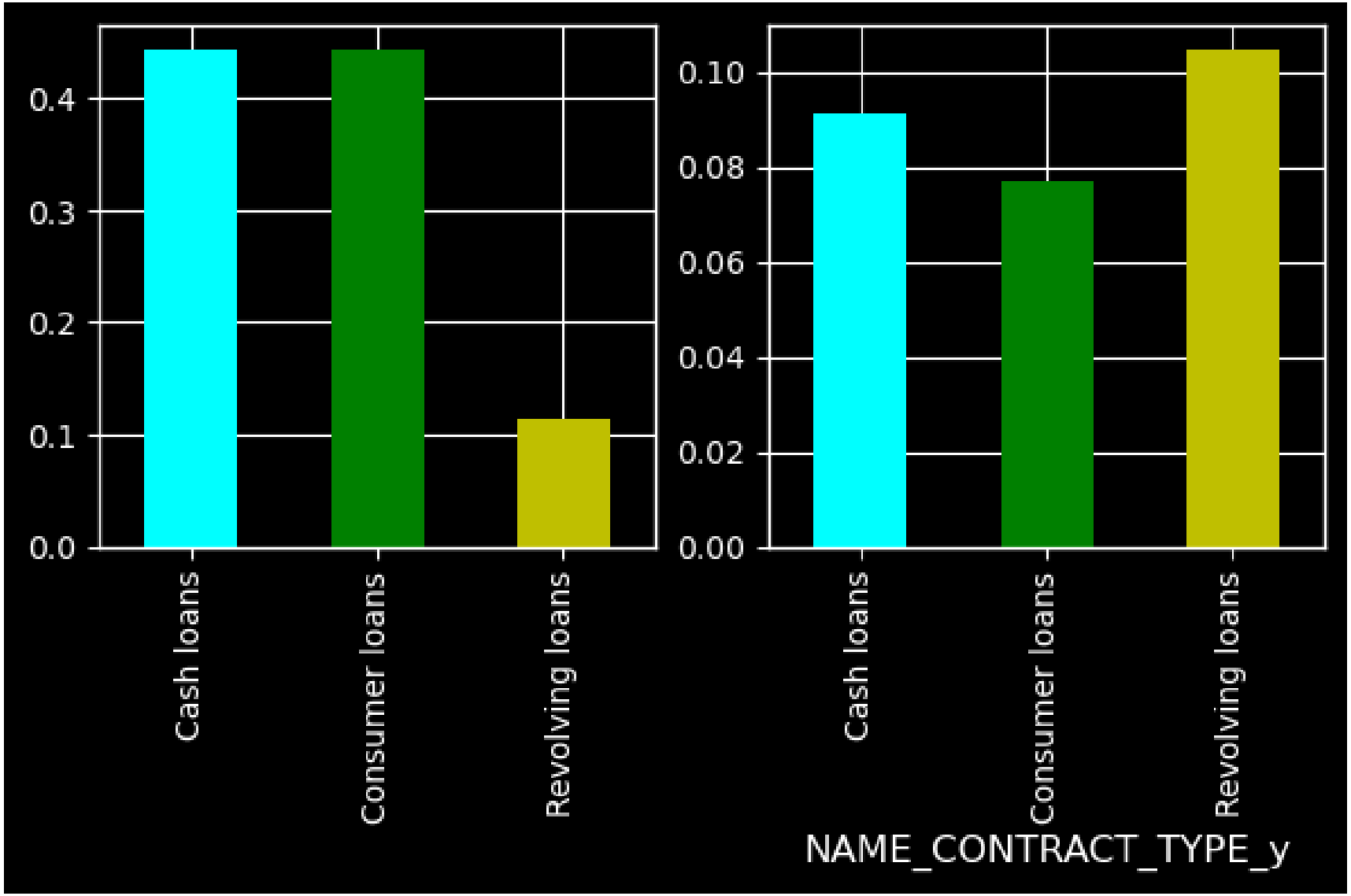


8. AS THE AGE INCREASES BEYOND 40YEARS THE DIFFICULTY ALSO DECREASES.

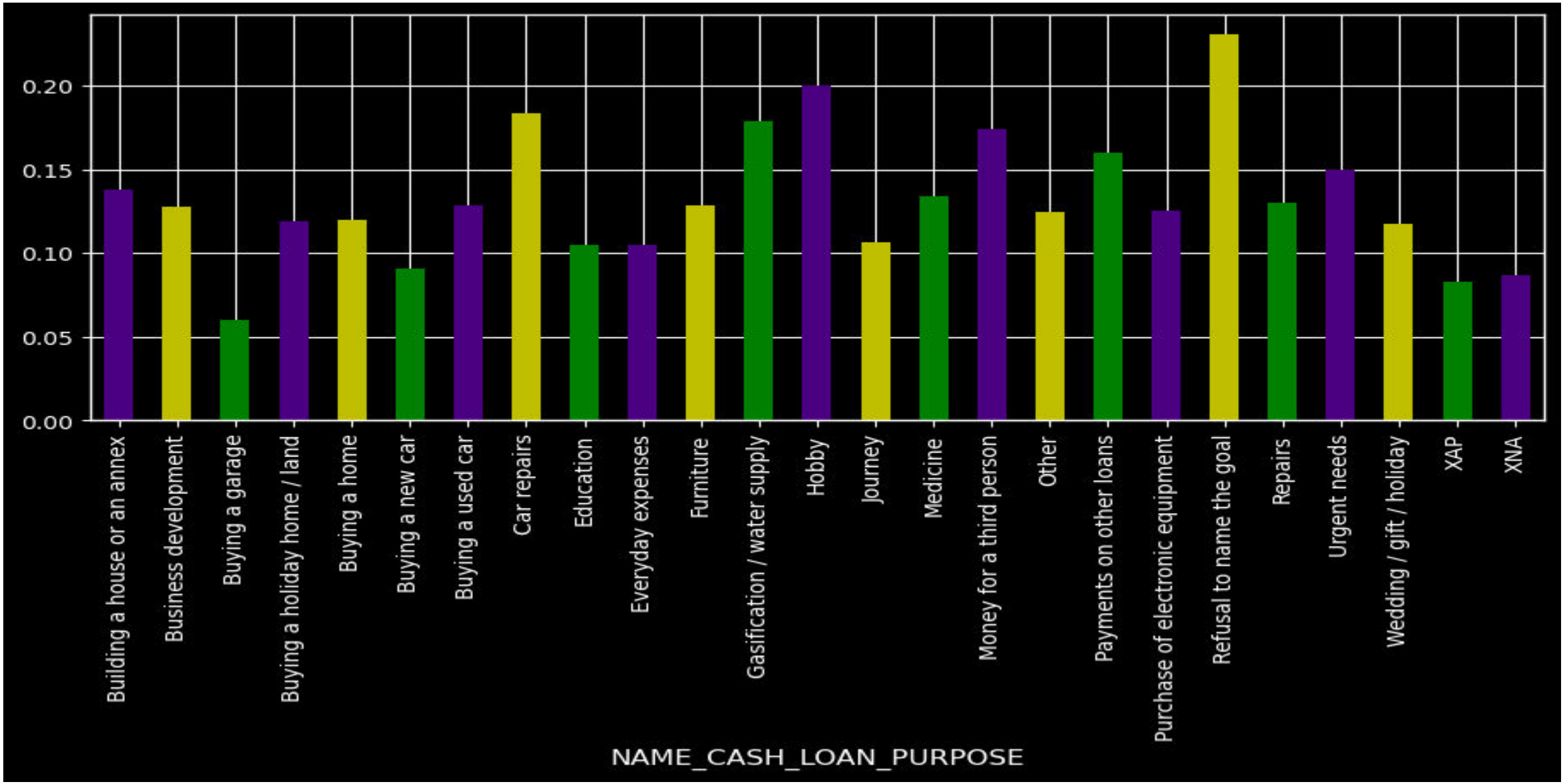


9. ON COMPARING WE CAN SEE (A). CLIENTS LIVING IN RATING 3 ARE HAVING MORE DIFFICULTY. (B). CLIENTS LIVING IN RATING 1 ARE HAVING LESS DIFFICULTY

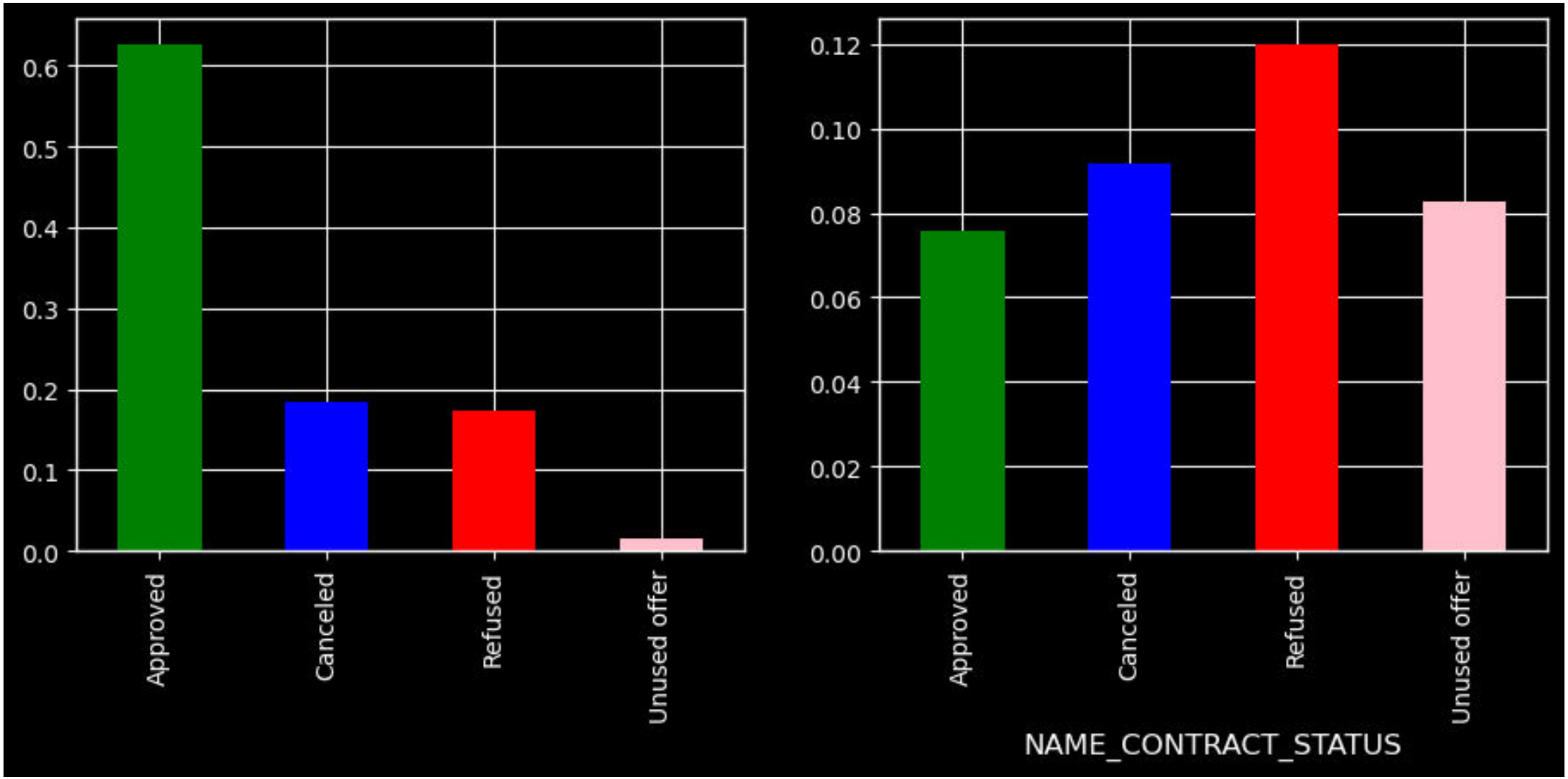
10. REVOLVING LOANS WERE THE LEAST APPLIED FOR , AND THE ONES WHO APPLIED FOR IT ARE HAVING MORE DIFFICULTY TO REPAY THE CURRENT LOAN.



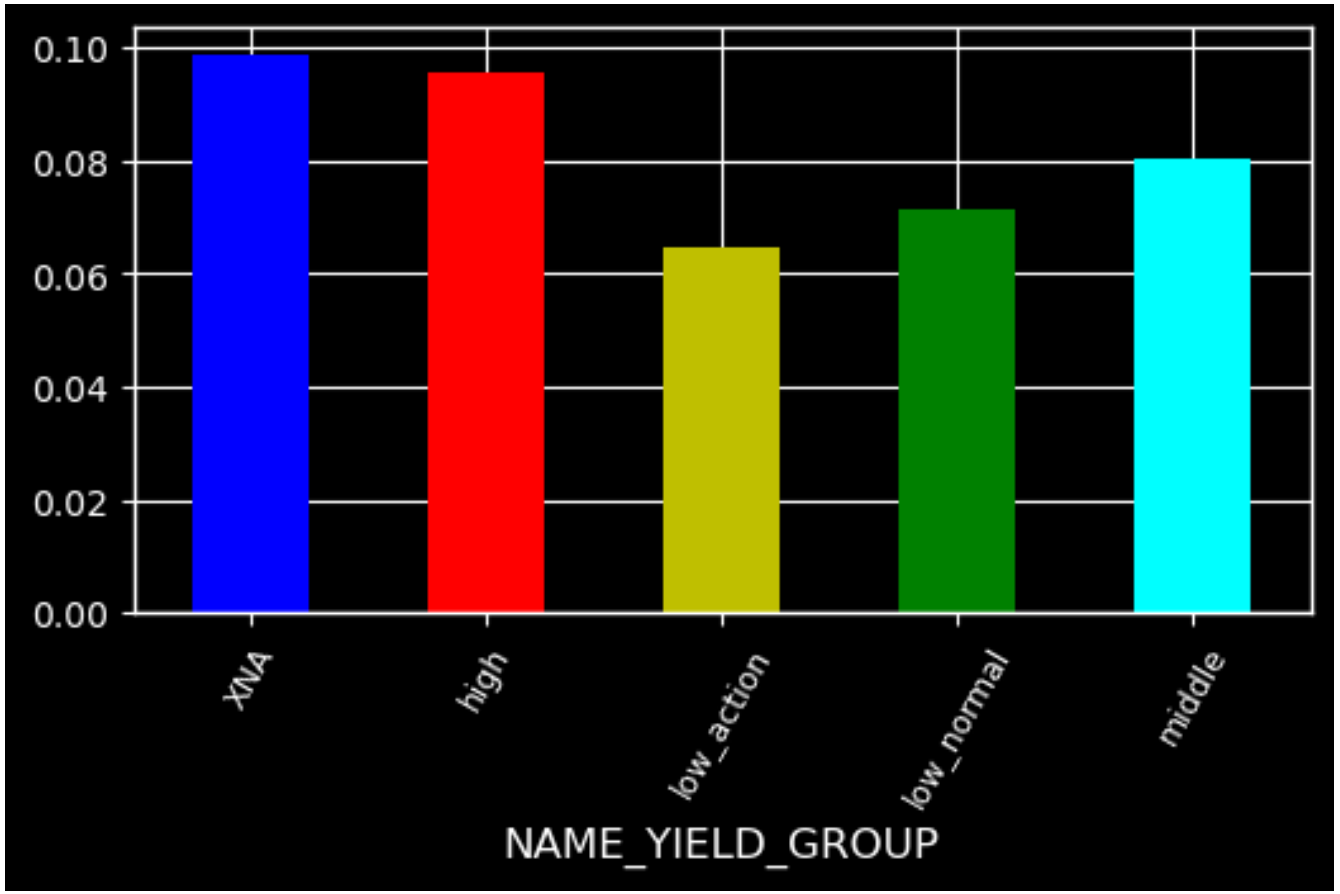
11. THE CLIENTS WHO REFUSED TO NAME THE PURPOSE FOR THE PREV LOAN ARE THE ONES HAVING MOST DIFFICULTY IN PAYING THE CURRENT ONE, FOLLOWED BY HOBBY AND CAR REPAIRS.



12. ONE IMP THING TO SEE HERE IS THAT CLIENTS WHOSE PREV APPN WAS REJECTED ARE HAVING DIFFICULTY IN PAYING FOR THE CURRENT LOAN, FOLLOWED BY CANCELED AND UNUSED OFFER.

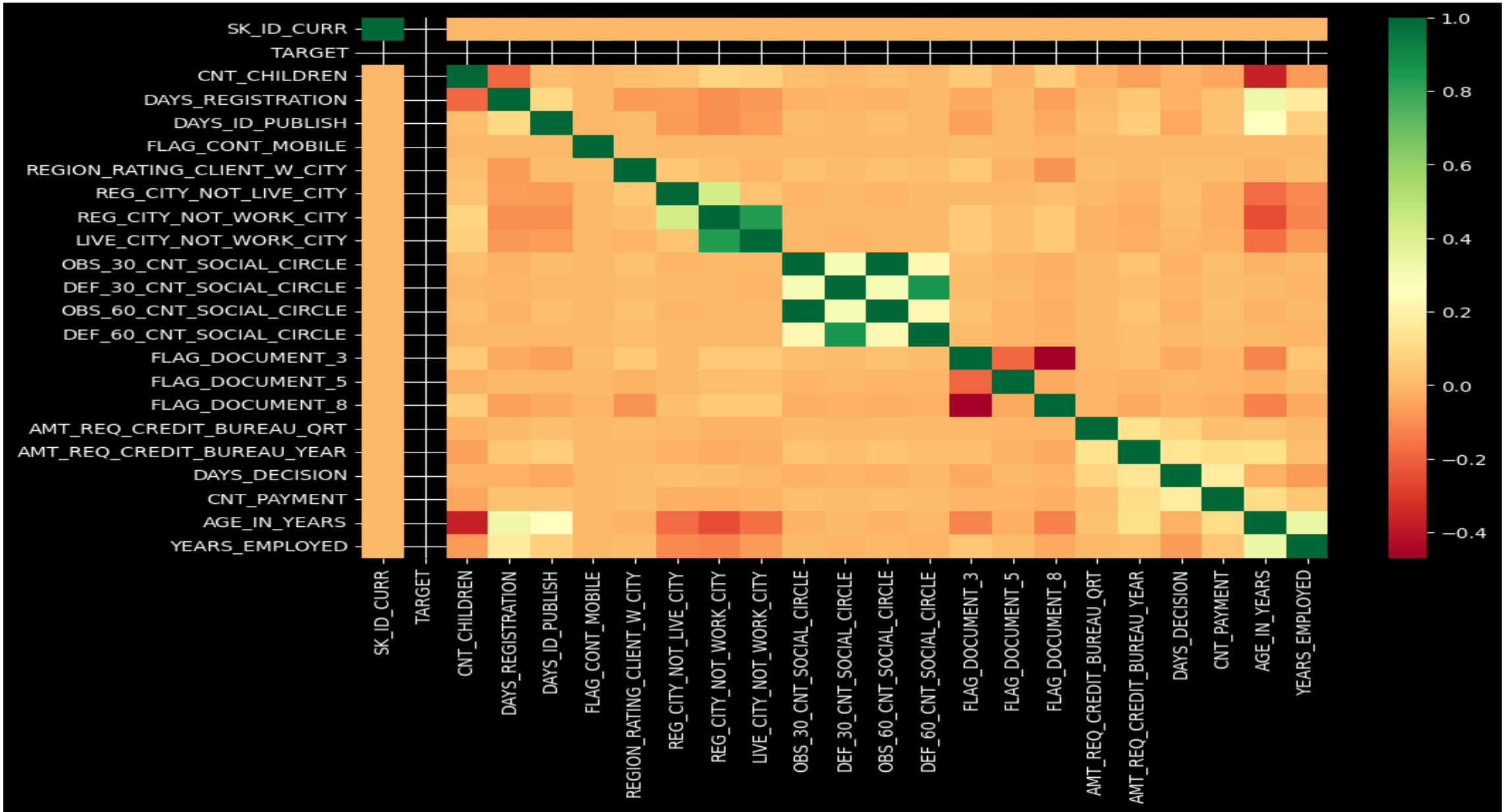


13. AS PER THIS PLOT THE LOAN REPAYMENT DIFFICULTY IS DEPENDENT ON THE INTEREST RATE OF THE PREV LOAN. THE HIGH INTEREST RATE PAYING CLIENTS ARE FACING MOST DIFFICULTY AND THE LOW ACTION INTEREST RATE THE LEAST.

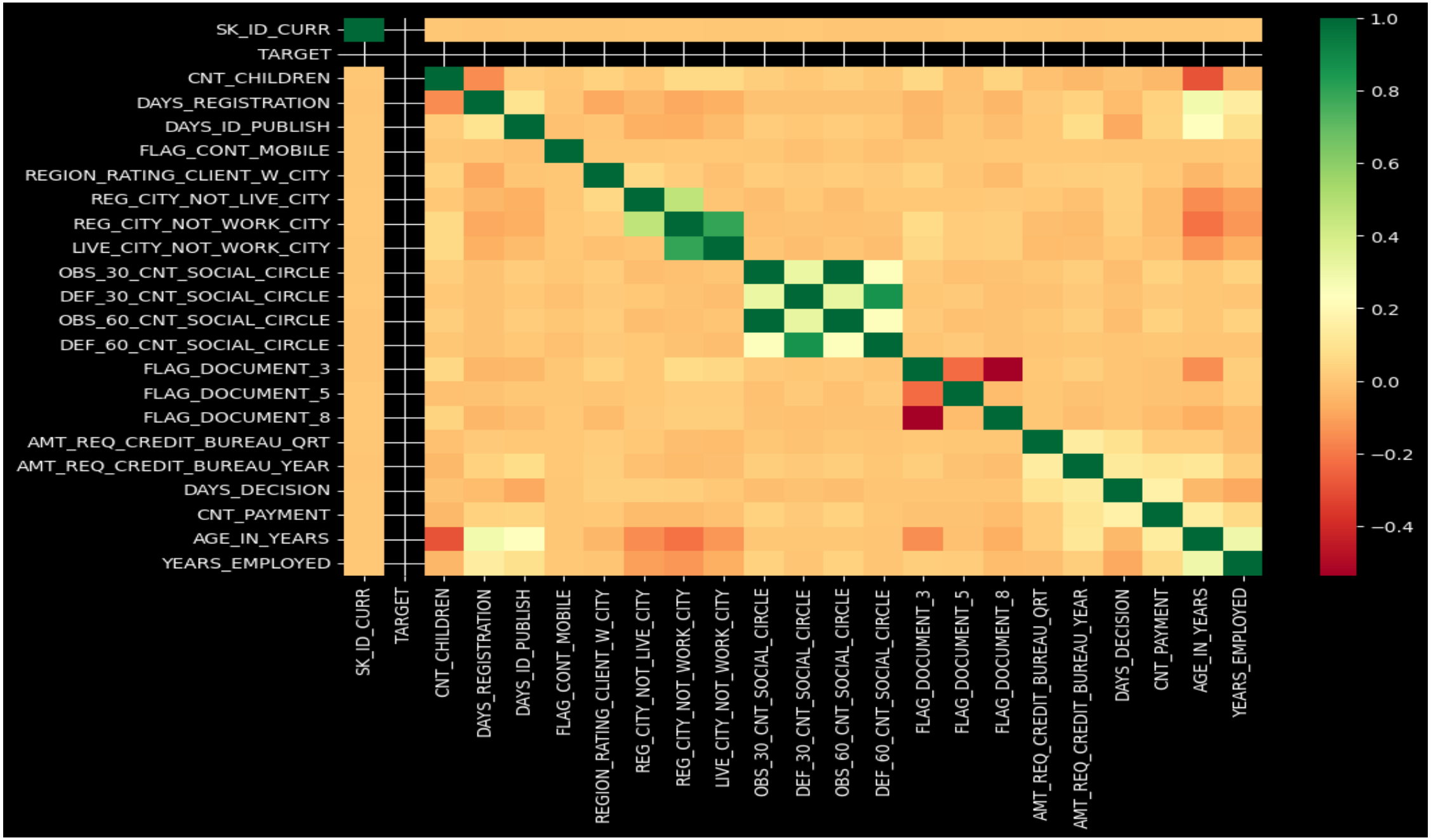


CORRELATION MATRIX

- For all other cases.



- For clients with payment difficulties.



- TOP 10 CORRELATION FOR ALL OTHER CASES

1. DEF_30_CNT_SOCIAL_CIRCLE - DEF_60_CNT_SOCIAL_CIRCLE
2. DEF_30_CNT_SOCIAL_CIRCLE - OBS_60_CNT_SOCIAL_CIRCLE
3. DEF_60_CNT_SOCIAL_CIRCLE - OBS_60_CNT_SOCIAL_CIRCLE
4. DEF_30_CNT_SOCIAL_CIRCLE - OBS_30_CNT_SOCIAL_CIRCLE
5. OBS_30_CNT_SOCIAL_CIRCLE - OBS_60_CNT_SOCIAL_CIRCLE
6. DEF_60_CNT_SOCIAL_CIRCLE - OBS_30_CNT_SOCIAL_CIRCLE
7. REG_CITY_NOT_WORK_CITY - REG_CITY_NOT_LIVE_CITY
8. LIVE_CITY_NOT_WORK_CITY - REG_CITY_NOT_WORK_CITY
9. FLAG_DOCUMENT_3 - FLAG_DOCUMENT_5
10. FLAG_DOCUMENT_3 - FLAG_DOCUMENT_8

- TOP 10 CORRELATION FOR CLIENTS WITH PAYMENT DIFFICULTIES

1. OBS_30_CNT_SOCIAL_CIRCLE - OBS_60_CNT_SOCIAL_CIRCLE
2. LIVE_CITY_NOT_WORK_CITY - REG_CITY_NOT_WORK_CITY
3. REG_CITY_NOT_WORK_CITY - REG_CITY_NOT_LIVE_CITY
4. DEF_30_CNT_SOCIAL_CIRCLE - DEF_60_CNT_SOCIAL_CIRCLE
5. DEF_30_CNT_SOCIAL_CIRCLE - OBS_30_CNT_SOCIAL_CIRCLE
6. DEF_30_CNT_SOCIAL_CIRCLE - OBS_60_CNT_SOCIAL_CIRCLE
7. AGE_IN_YEARS - YEARS_EMPLOYED
8. FLAG_DOCUMENT_3 - FLAG_DOCUMENT_5
9. FLAG_DOCUMENT_3 - FLAG_DOCUMENT_8
10. DEF_60_CNT_SOCIAL_CIRCLE - OBS_30_CNT_SOCIAL_CIRCLE

- DRIVER VARIABLES FOR LOAN DEFAULT

1. CODE_GENDER
2. NAME_INCOME_TYPE
3. NAME_EDUCATION_TYPE
4. AGE_IN_YEARS
5. REGION_RATING_CLIENT_W_CITY
6. NAME_CONTRACT_STATUS
7. NAME_YIELD_GROUP
8. NAME_HOUSING_TYPE
9. FLAG_OWN_REALTY
10. NAME_CONTRACT_TYPE_y

CONCLUSION

1. The company should look out for the clients who falls in the category of most of the driver variables for loan defaults.
2. The categories are **male**, income type is **unemployed** or **maternity leave**, **lower education**, **young in age**, belongs to **region rating 3**, **doesn't own realty** and **live with parents** or **rented apartment**, **paying higher interest rate for previous revolving loan**, and whose previous loan application was either **refused** or **cancelled**.
3. For example, a young male client who is unemployed and lives in region rating 3 with his parents has higher chances of default.