CREDIT RISK ANALYSIS (EDA)

By-Sanskriti Kandpal

BUSINESS OBJECTIVE

This case study aims to identify patterns that indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the loan amount, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. This case study aims to identify such applicants using EDA.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables that are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

PROBLEM STATEMENT

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

METHODOLOGY

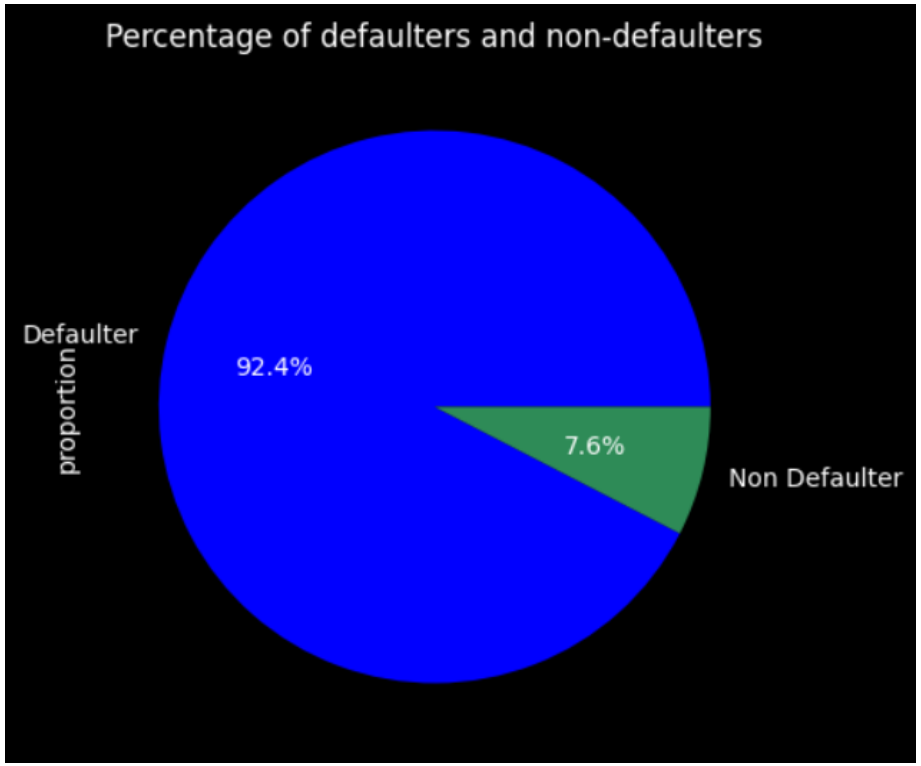
- First we uploaded the required libraries for the analysis ie pandas, numpy, matplotlib, and seaborn in ipynb file.
- Then we upload both the required datasets in the ipynb file.
- First we dropped any column in the dataset that was of no use to the analysis.
- After this we looked for null values in the dataset, we replaced the null values with the closest value possible, or the null values were dropped.
- Then we looked for outliers in the dataset, we kept most of the outliers as it is for the analysis.
- We then proceeded with the analysis part, by showing visualization for univariate, bivariate, and multivariate analysis and getting the required insights from the datasets.
- Finally we concluded, on the factors that may affect the applicants to turn into defaulters.

STEPS USED FOR CLEANING:

- The dataset was checked for the null values, the columns with a high number of null values were dropped if it does not align with the analysis.
- Then for the rest of the columns, first the meaning of the column was understood, then the null values were filled using the help of the values present in the column and the values corresponding to the other columns for the same row.
- Once all the null values were filled, we moved to the next part, identifying the outliers.
- The outliers were identified using boxplot, and were capped if necessary, otherwise they were kept as it is.
- Once all this was done, it was time to move to the analysis part.

UNIVARIATE ANALYSIS:

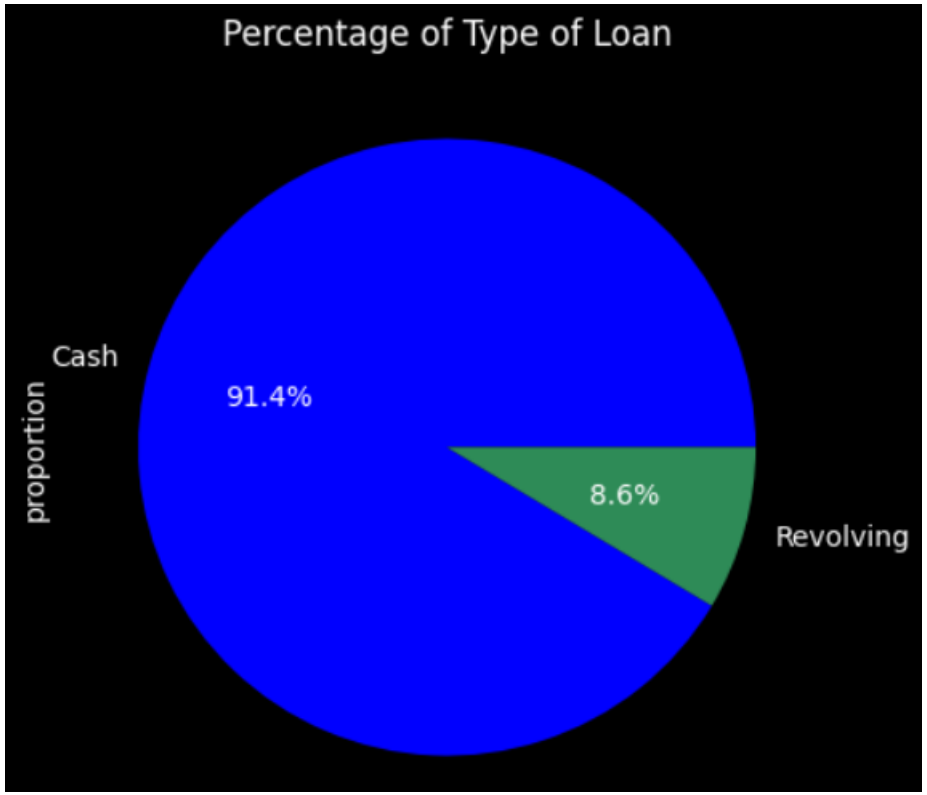
Percentage of Defaulters and Non-Defaulters



Findings:

- Majority of the applicants are non-defaulter (92.4%).

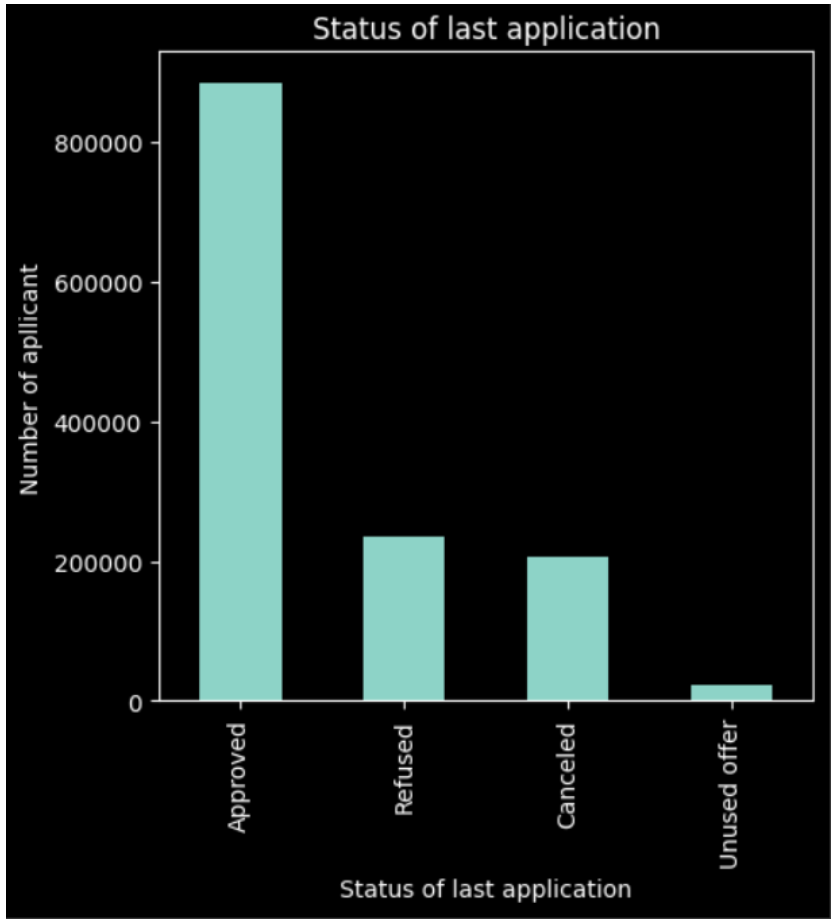
Percentage of Type of Loan



Findings:

- Majority of loans in current applications are Cash Loans.
- Very few applicants applied for revolving loans.

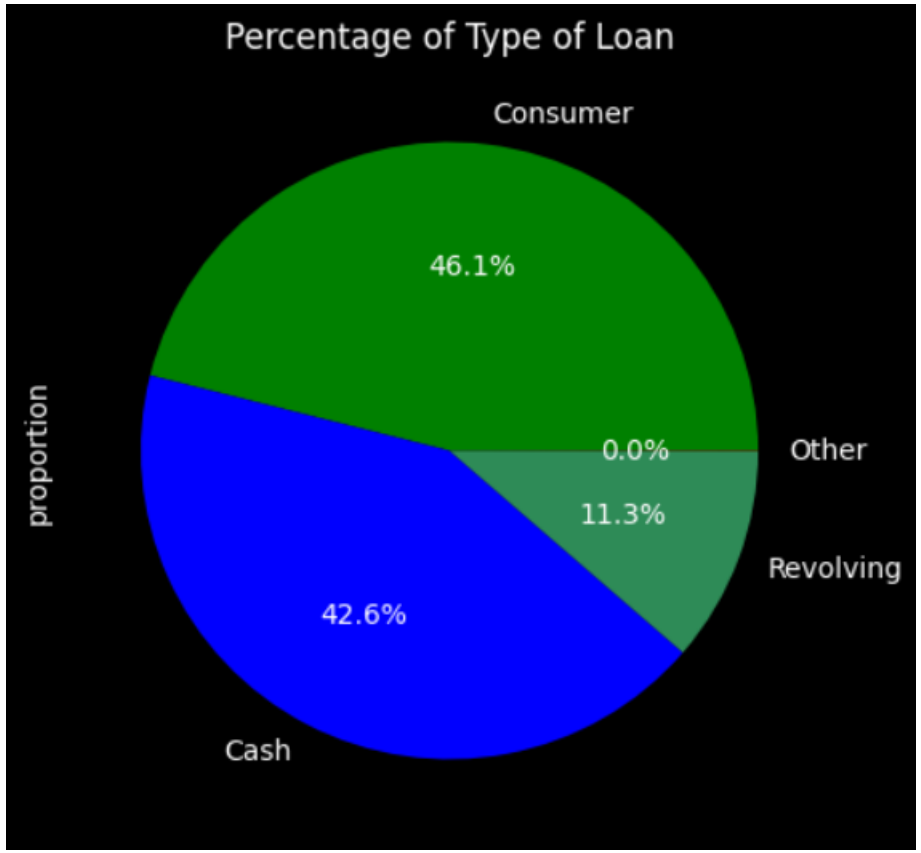
Status of last application



Findings:

- The majority of loans in previous applications were approved.
- The least were the unused offers.

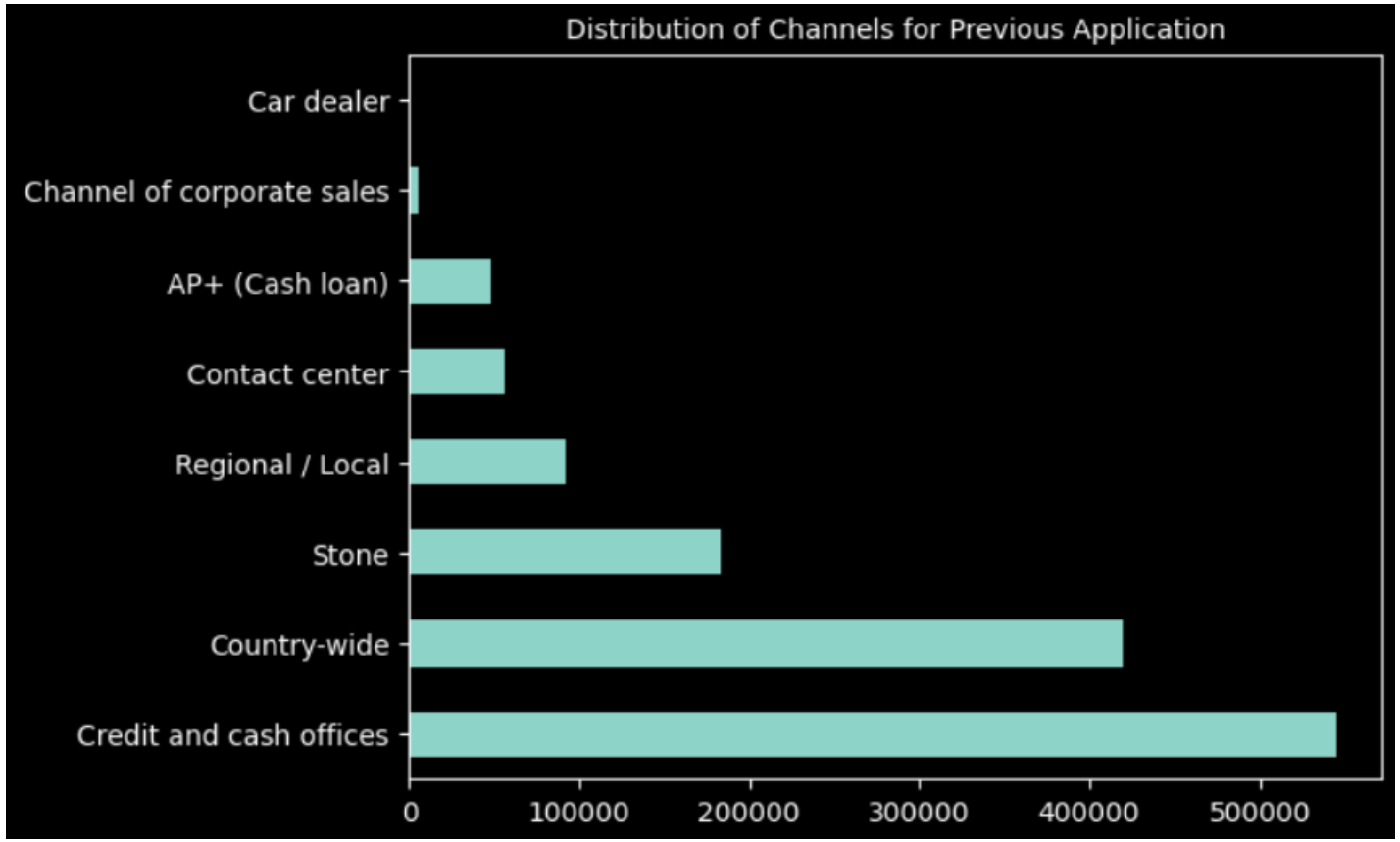
Percentage of Type of Loan for Previous Application



Findings:

- There were 4 types of loans in previous applications, namely cash loans, consumer loans revolving loans, and others.
- In previous applications the highest number of applications was for consumer loans and the least were for others.

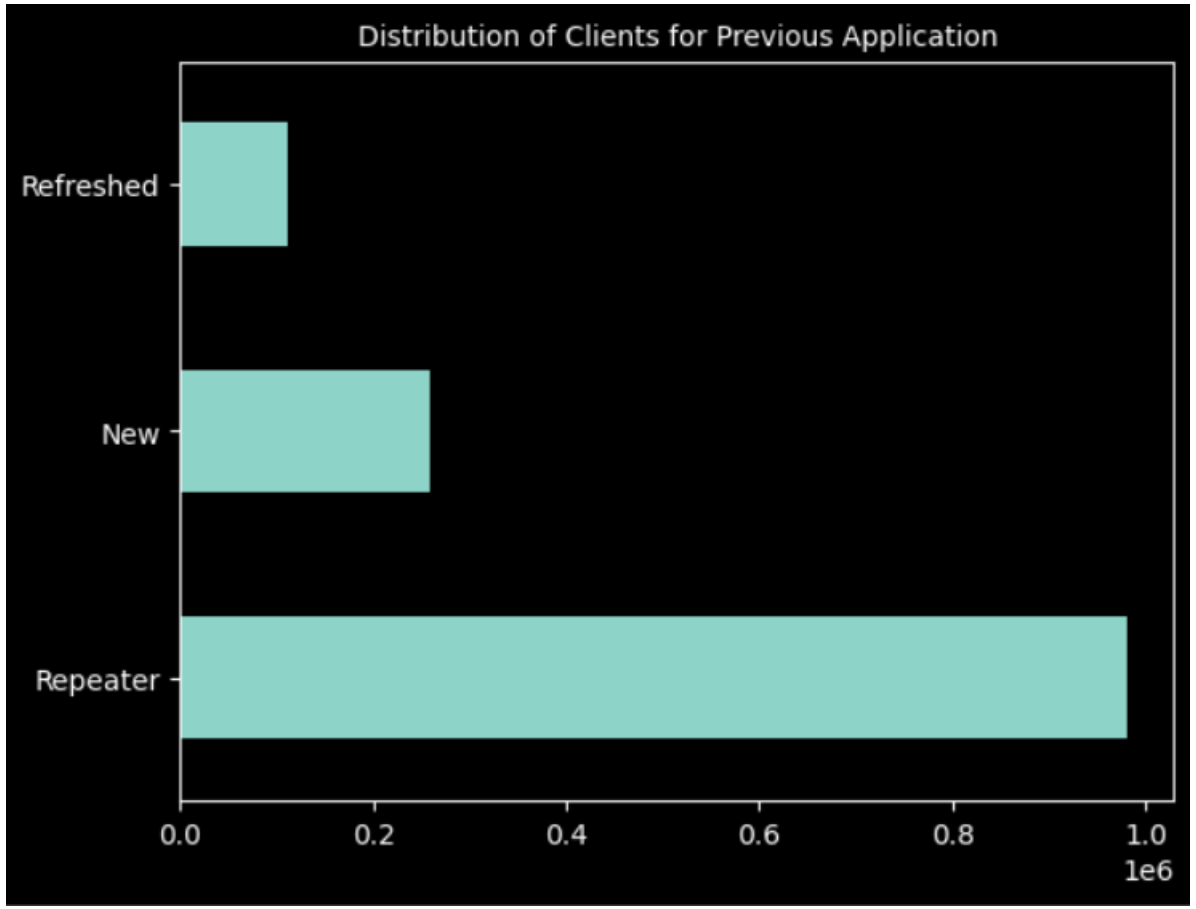
Distribution of Channels for Previous Application



Findings:

- The highest number of clients were acquired from credit and cash offices.
- The least number of clients were acquired from car dealers.

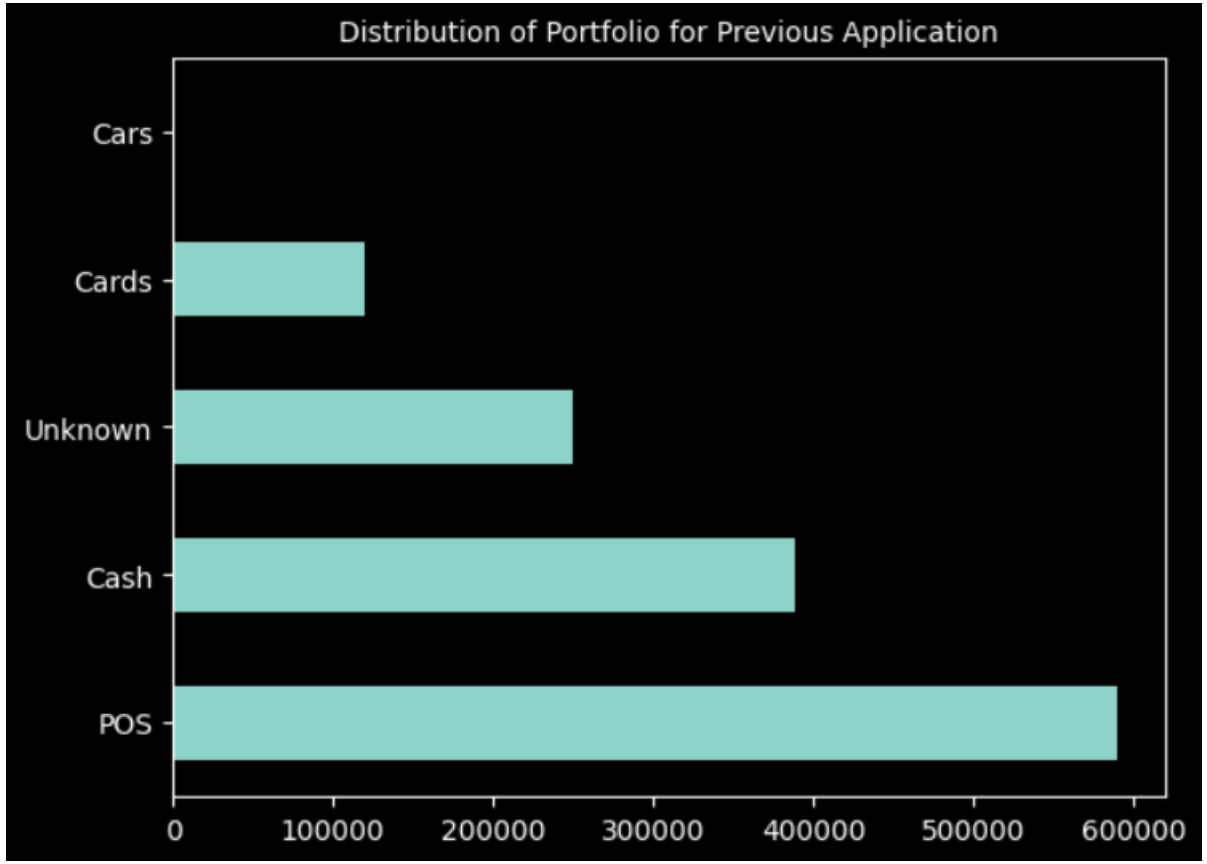
Distribution of Clients for Previous Applications



Findings:

The majority of the clients were repeaters and the least were refreshed clients.

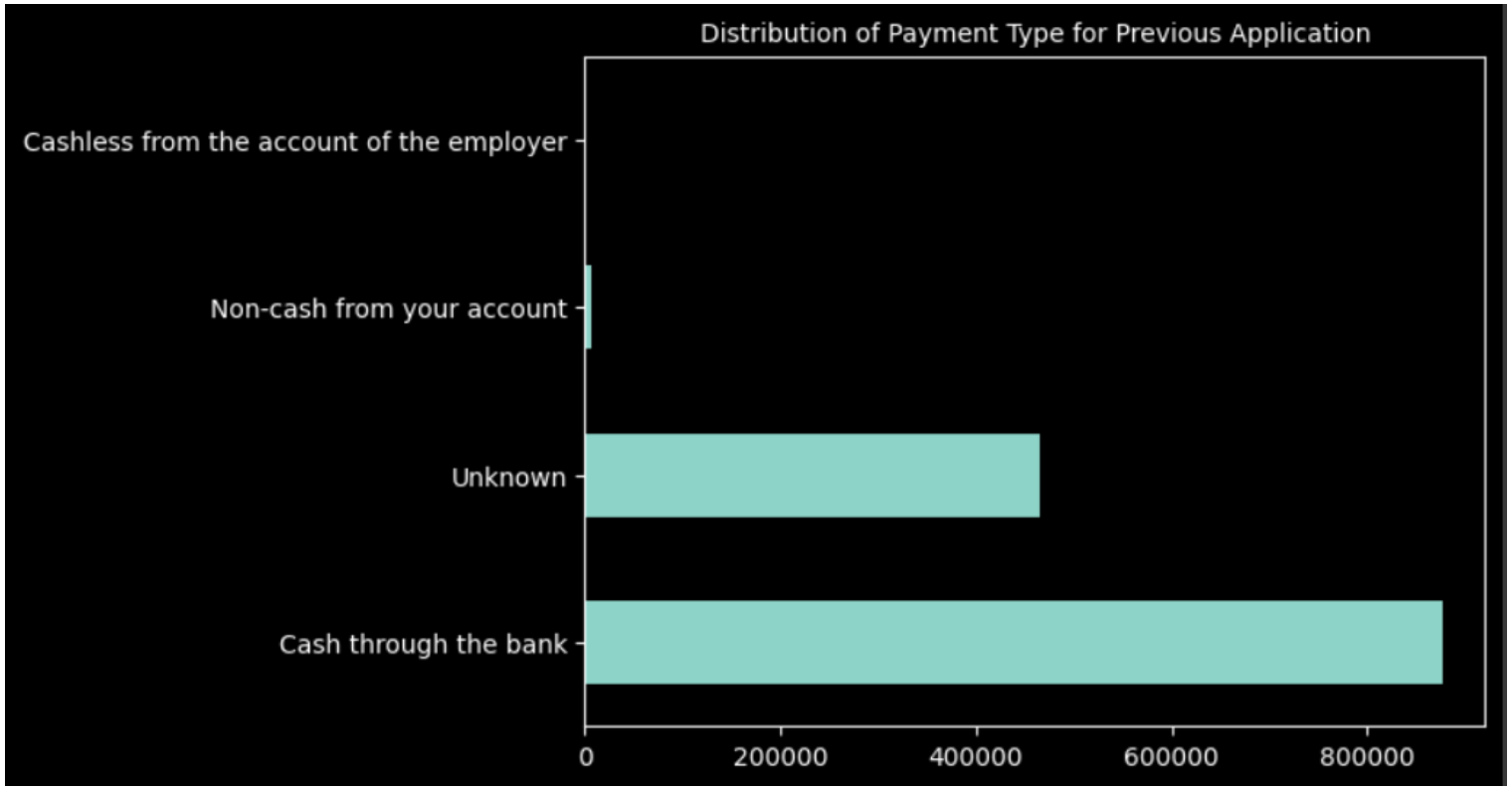
Distribution of Portfolio for Previous Application



Findings:

- The majority of the Portfolios were from POS for Previous applications.
- The least portfolios were from car for previous applications.

Distribution of Payment Type for Previous Application

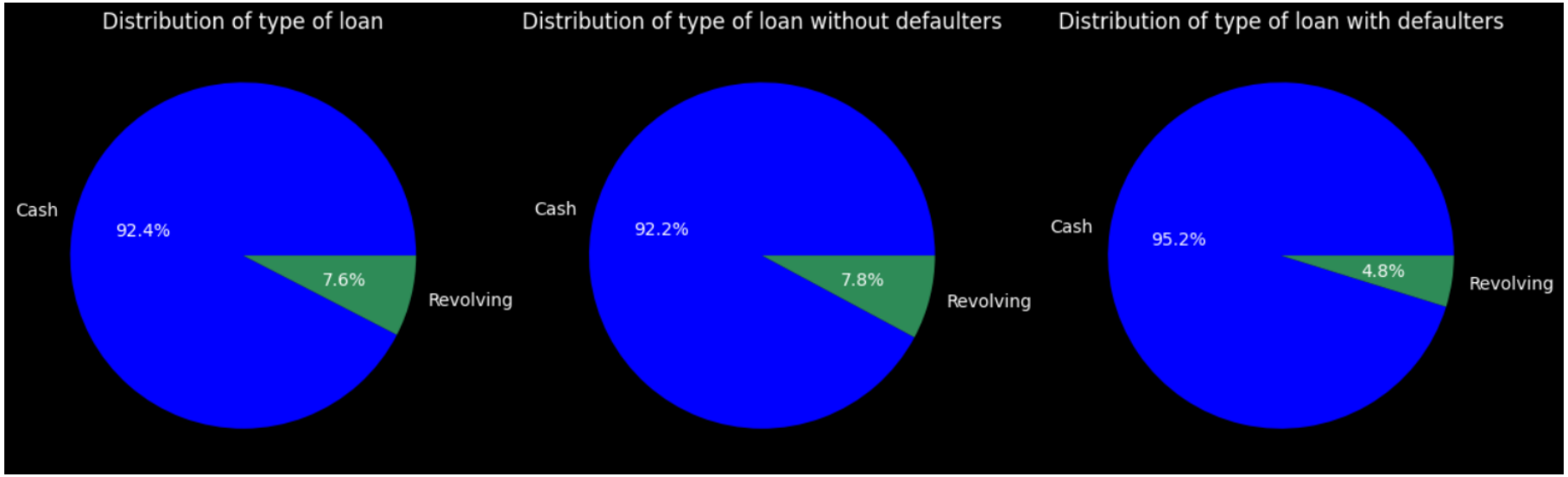


Findings:

- Most of the payments were done by Cash through the bank for previous applications.
- Least payments were made by Cashless from the account of the employer for previous applications.

BIVARIATE ANALYSIS:

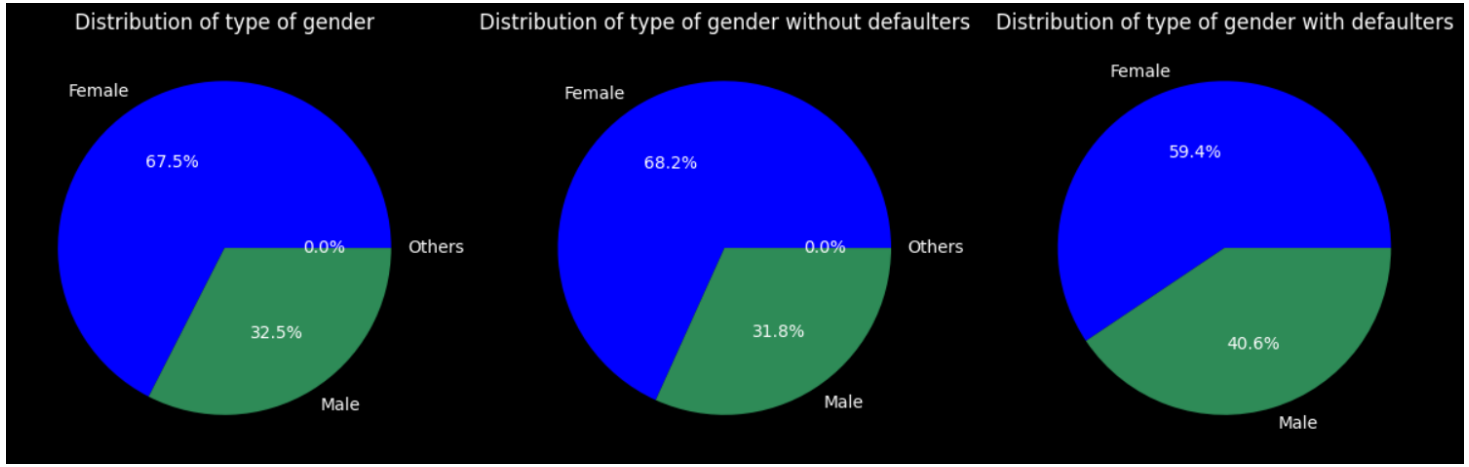
Distribution of Type of Loan for Defaulter and Non Defaulters

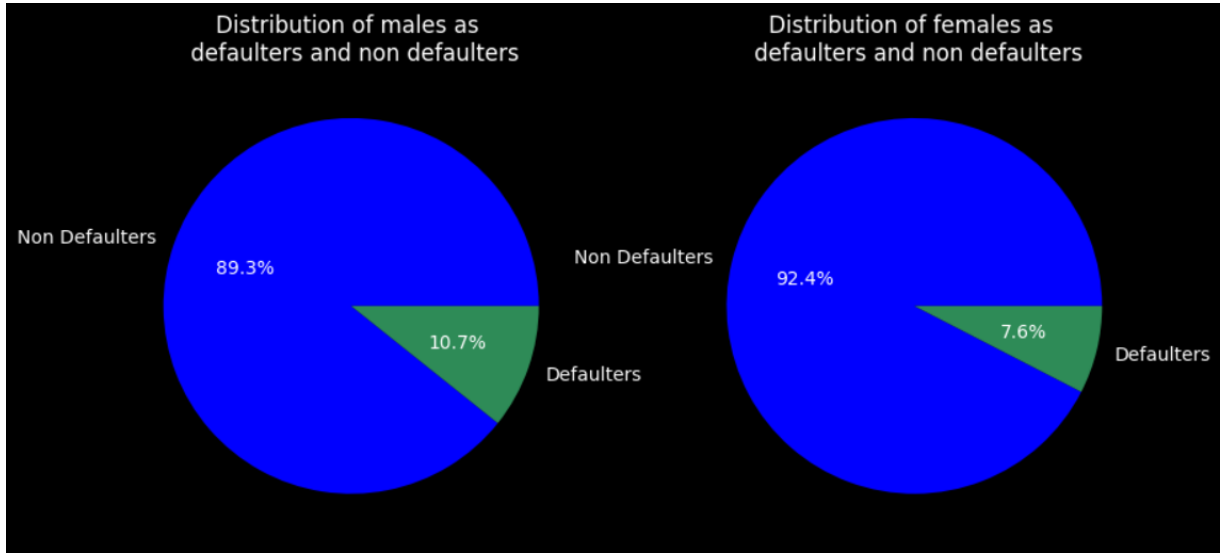


Findings:

- The majority of the loans were cash loans, even if the applicant was defaulter or not.
- For defaulters the proportion of revolving loans is little less than non-defaulter, this means defaults are occuring more in cash loans than revolving loans.

Gender for Defaulter and Non-Defaulters

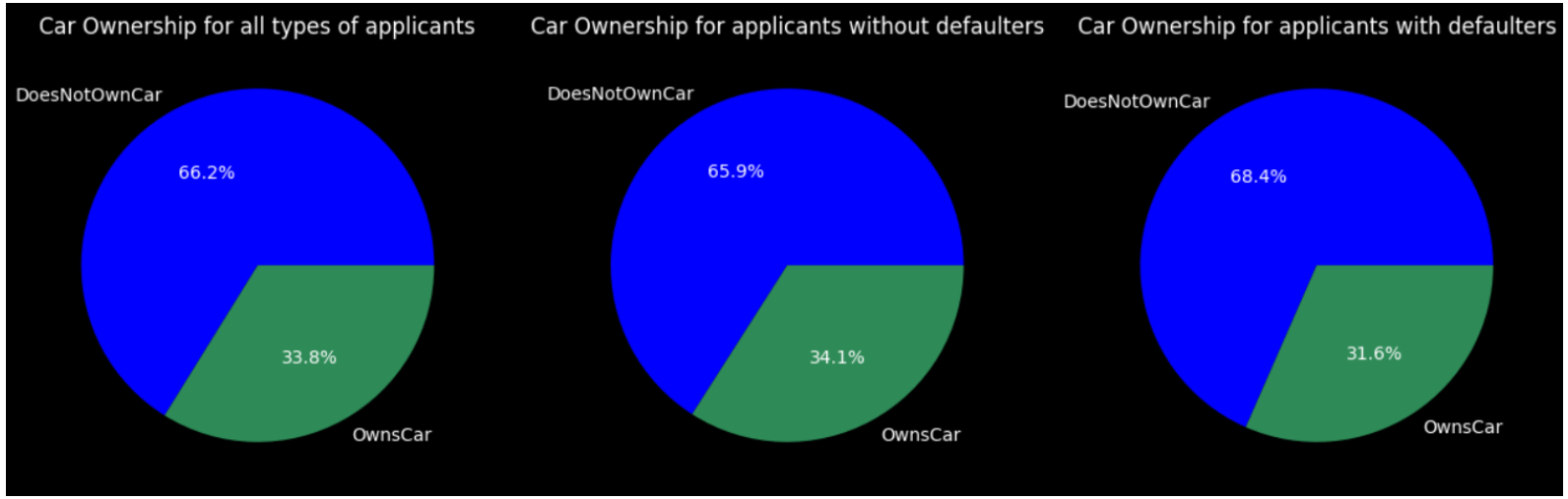


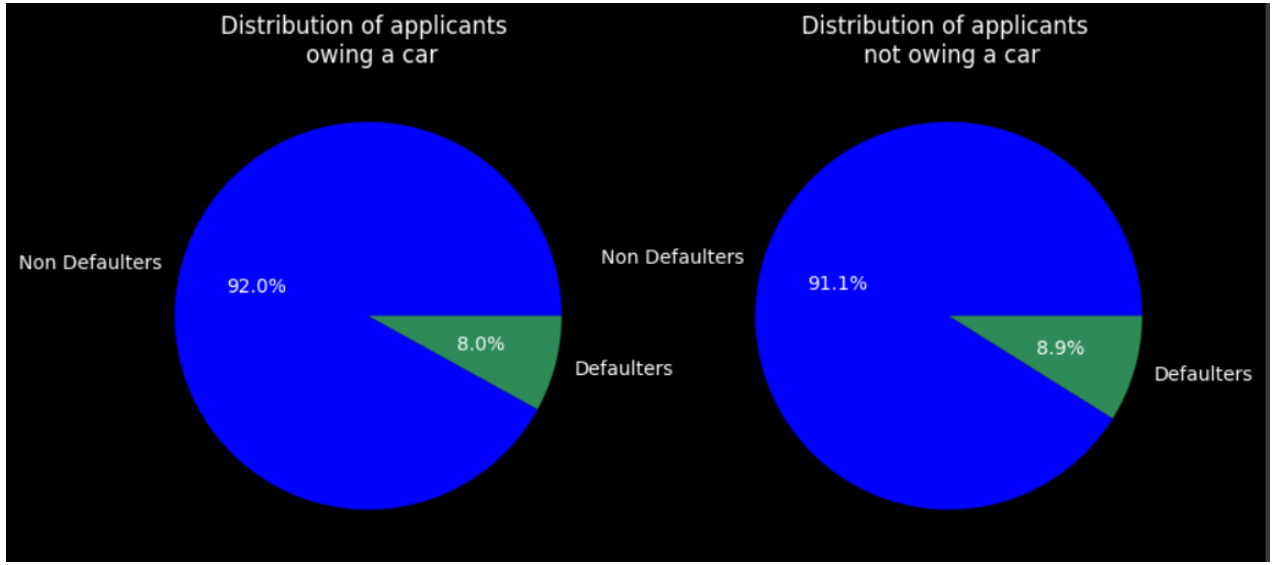


Findings:

- The majority of the applicants are female be they combined, for non defaulters or defaulters.
- Even though the proportion of male is less for defaulters and non defaulters, but the proportion of male defaulters are more than female defaulter at the individual level.

Car Ownership

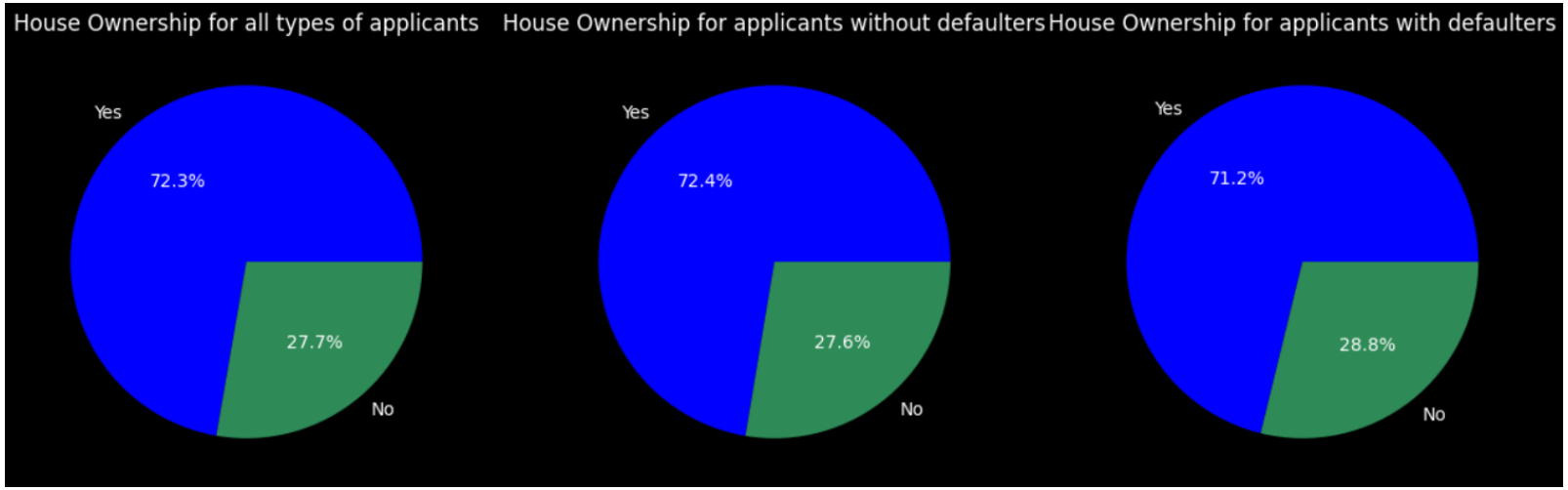


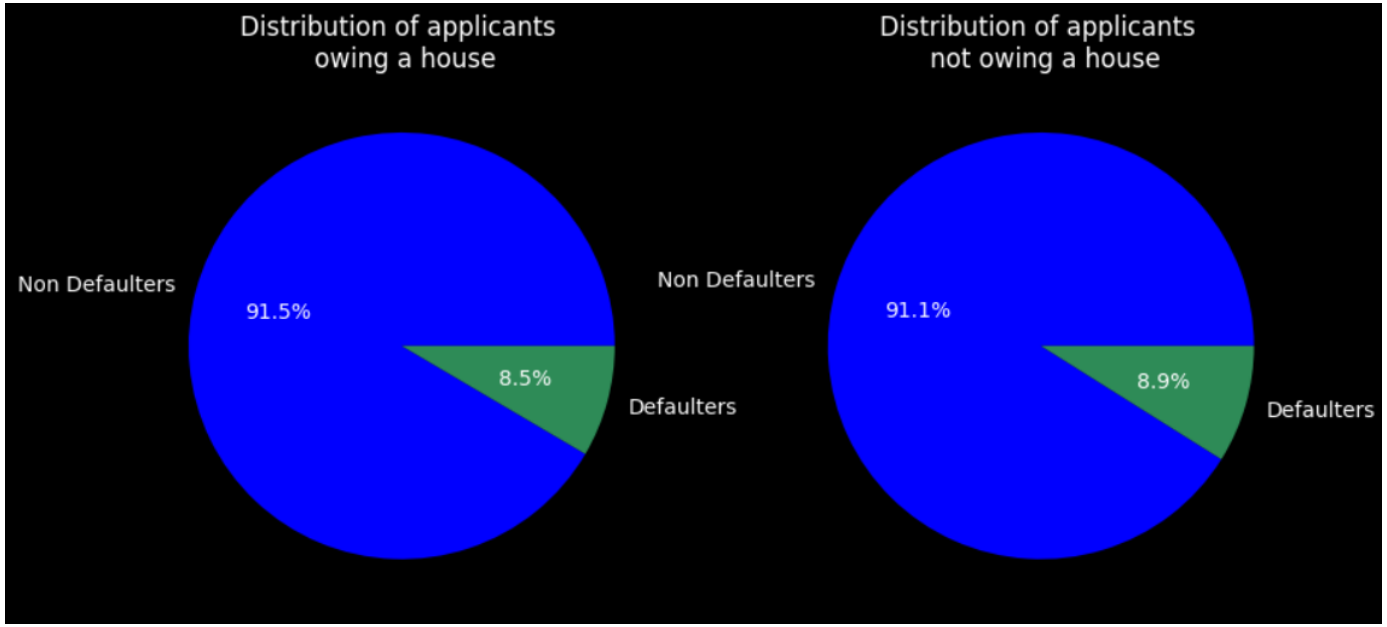


Findings:

The proportion of applicant owing a car is greater for defaulters and as well as non defaulters.

House Ownership

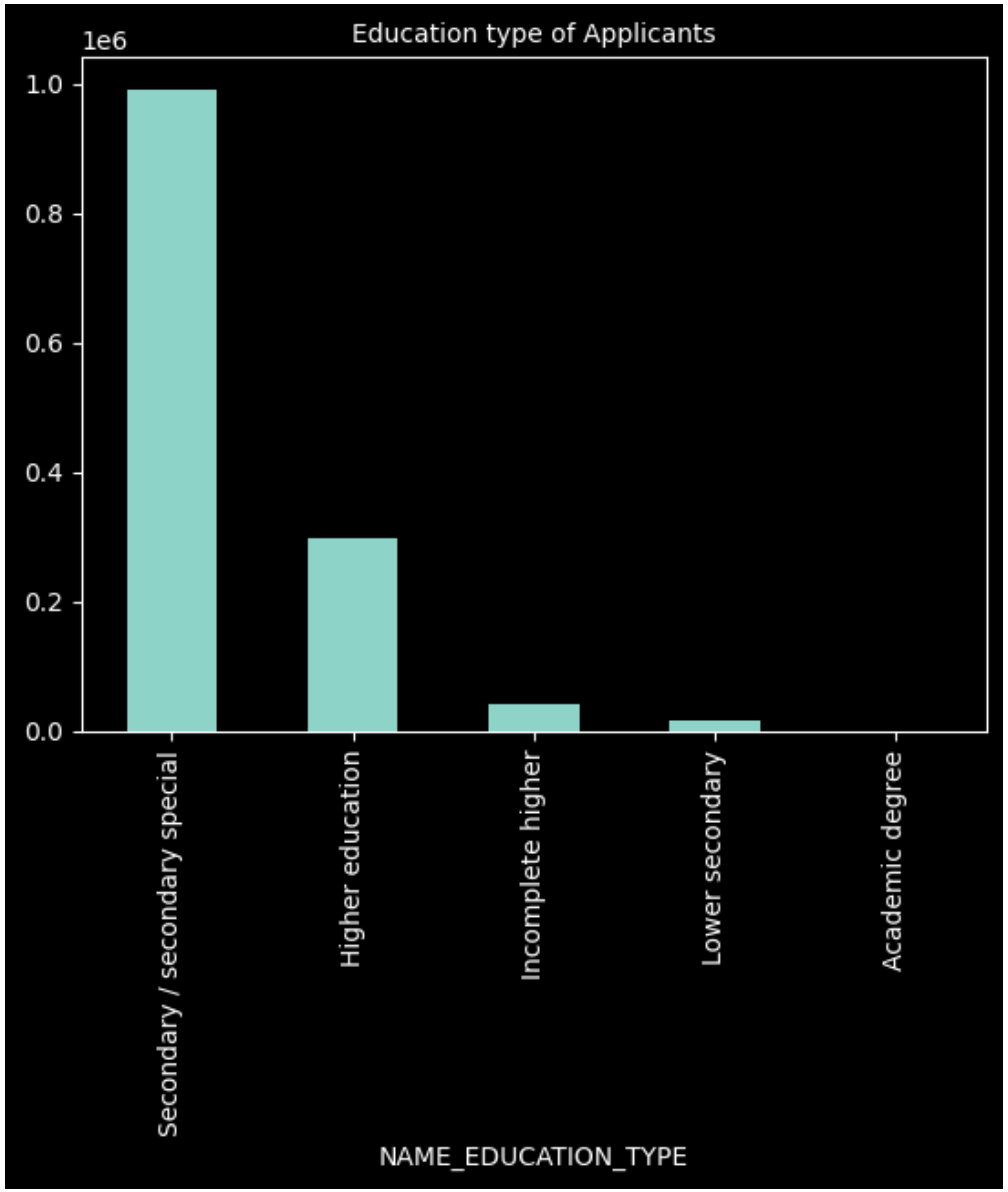




Findings:

The proportion of applicants owing a house is greater for defaulters and as well as non-defaulters.

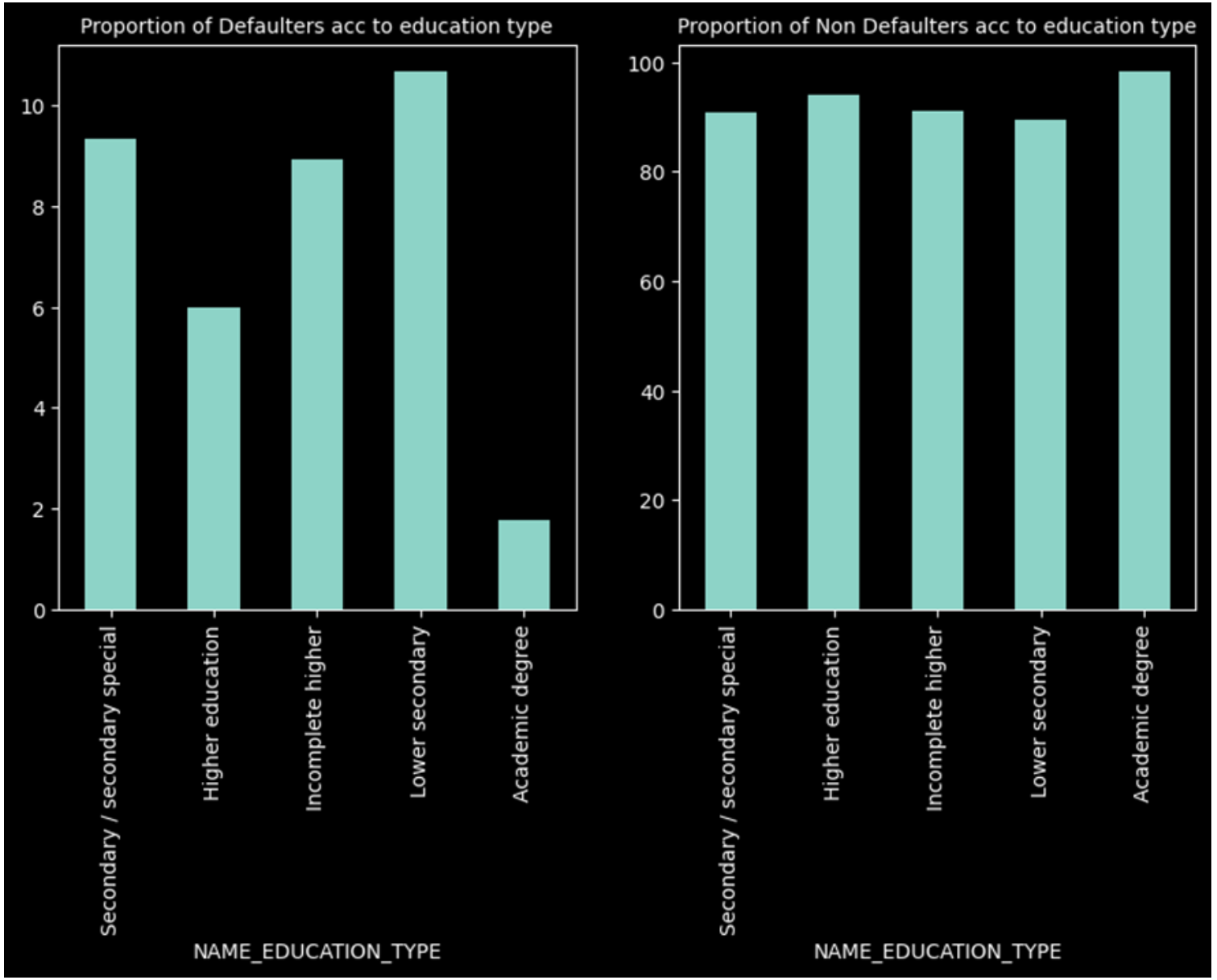
Education Type



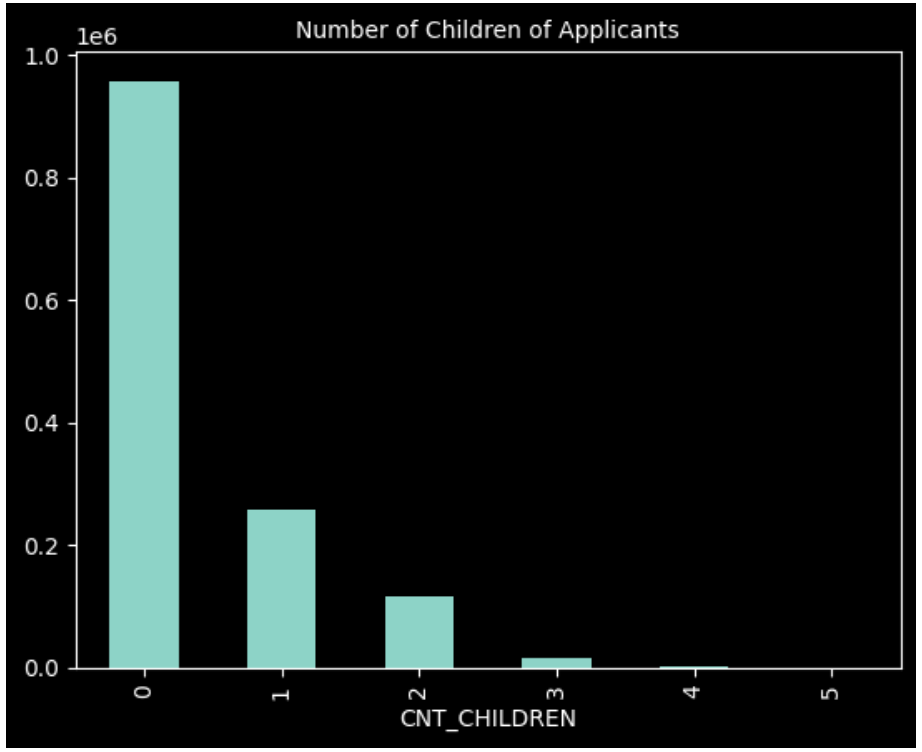
Findings:

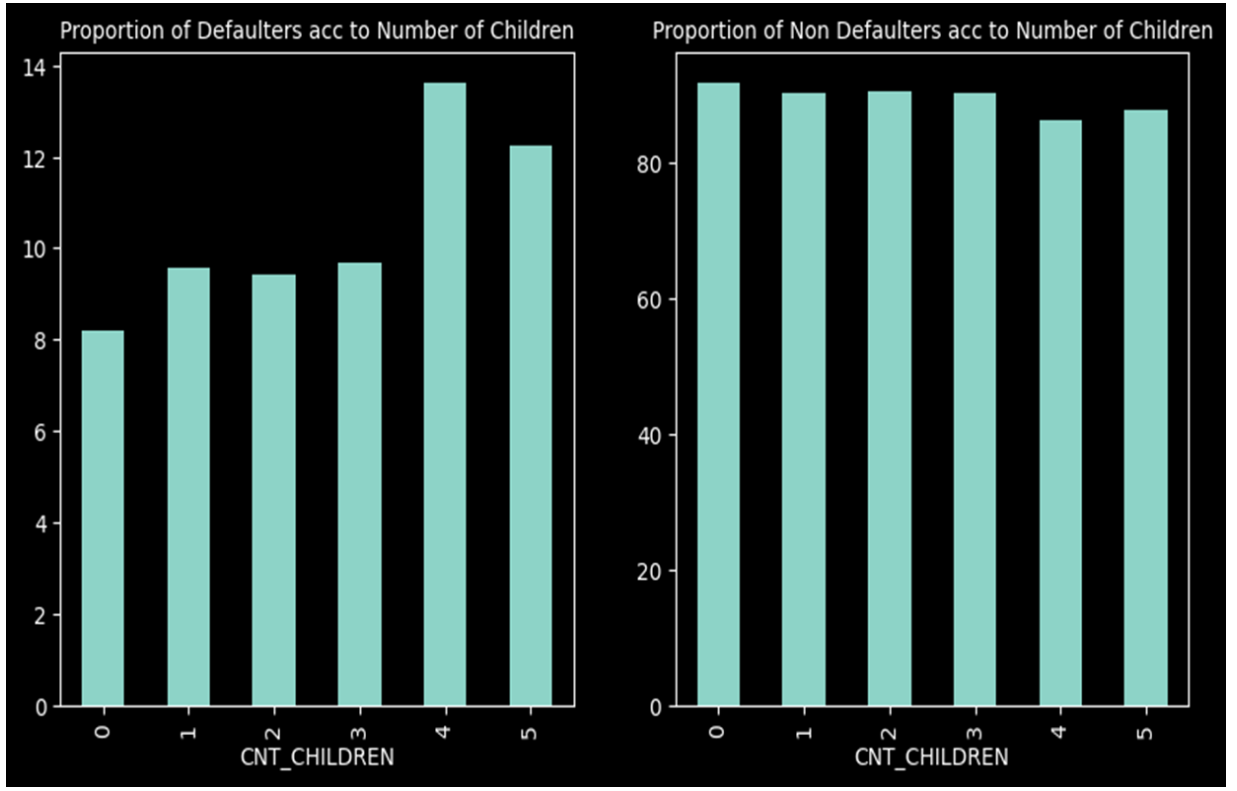
- The majority of applicants have a Secondary/ Secondary special education level (73.40% approx.) and the least number of applicants are from an Academic Degree education level (0.04% approx).
- The majority of defaulters are from the secondary/ Secondary special education level (9.33% approx) and the least number of applicants are from the Academic degree education level (1.78% approx).

- The majority of non-defaulters are from academic degree education level (98.22% approx.) and the least number of applicants are from Secondary/ Secondary special education level (90.67% approx.).
- So in conclusion we can say that the higher the education level, the lower the chance of an applicant defaulting.



Number of Children

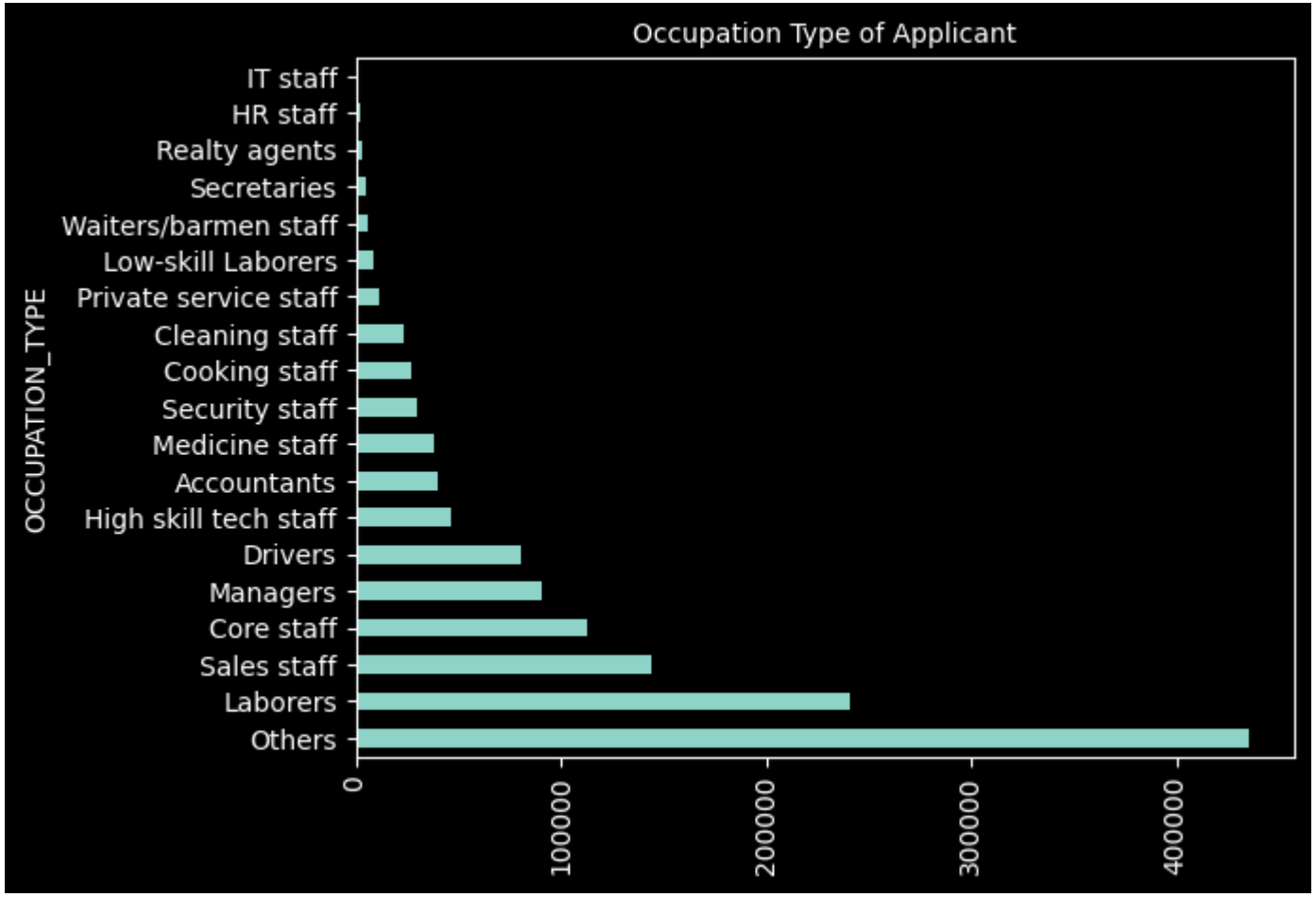




Findings:

- The majority of the applicants have no child (70.93% approx) and the least number of applicants have 5 or more kids (0.04% approx).
- The majority of the defaulters have 4 children (13.61% approx) or 5 or more children (12.23% approx) while the least number of defaulters have zero children (8.12% approx).
- . The majority of non-defaulters have zero children (91.8% approx.) and the least number of non-defaulters have 4 children (86.39% approx) or 5 or more children (87.77% approx).
- Therefore we can conclude that the applicants having 3 or fewer children have a lower chance of defaulting

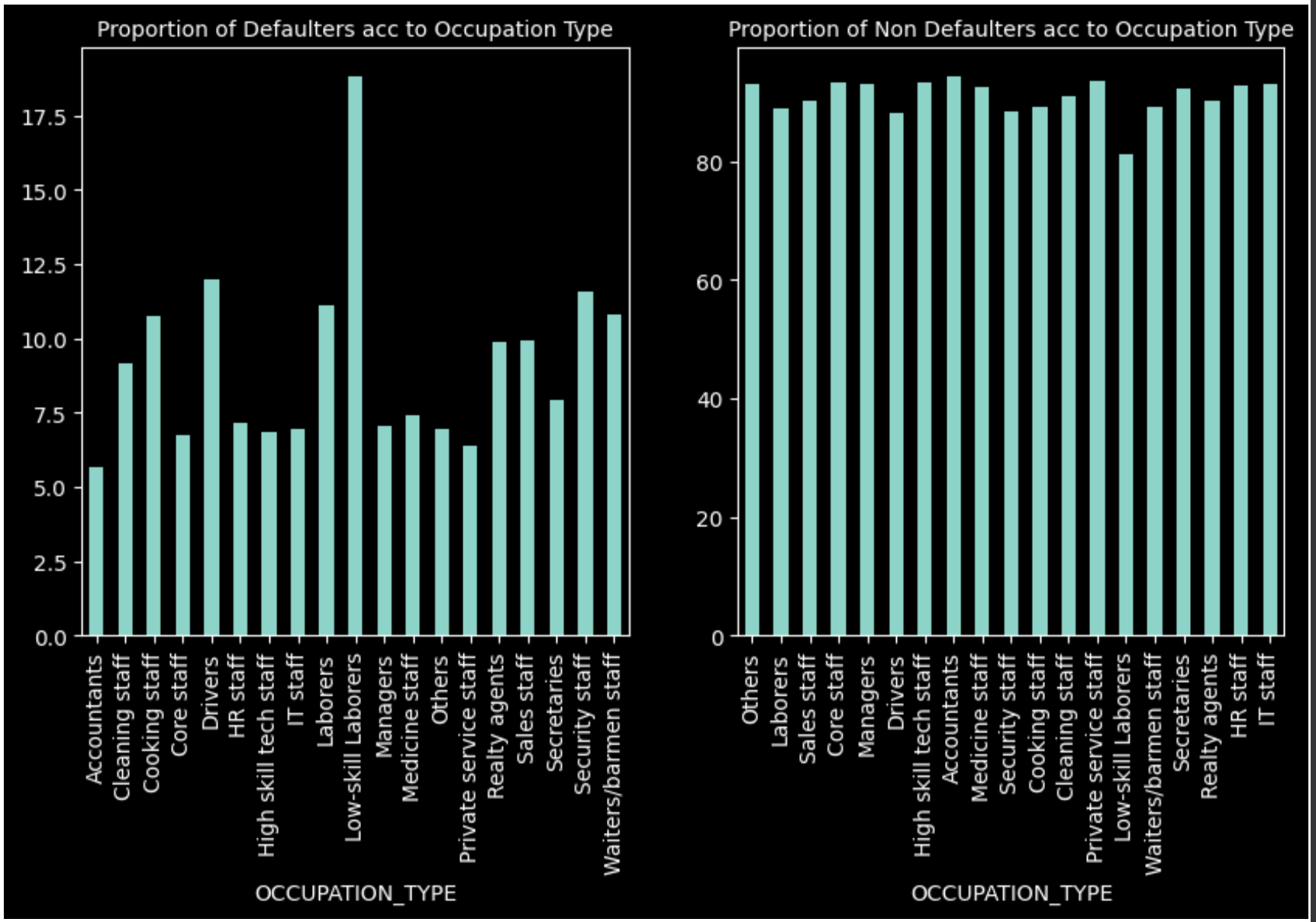
Occupation Type



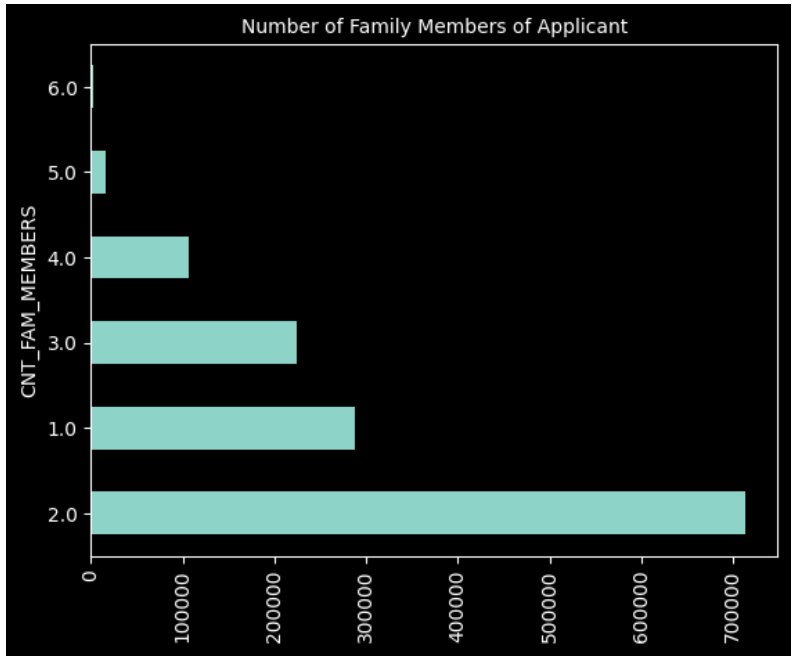
Findings:

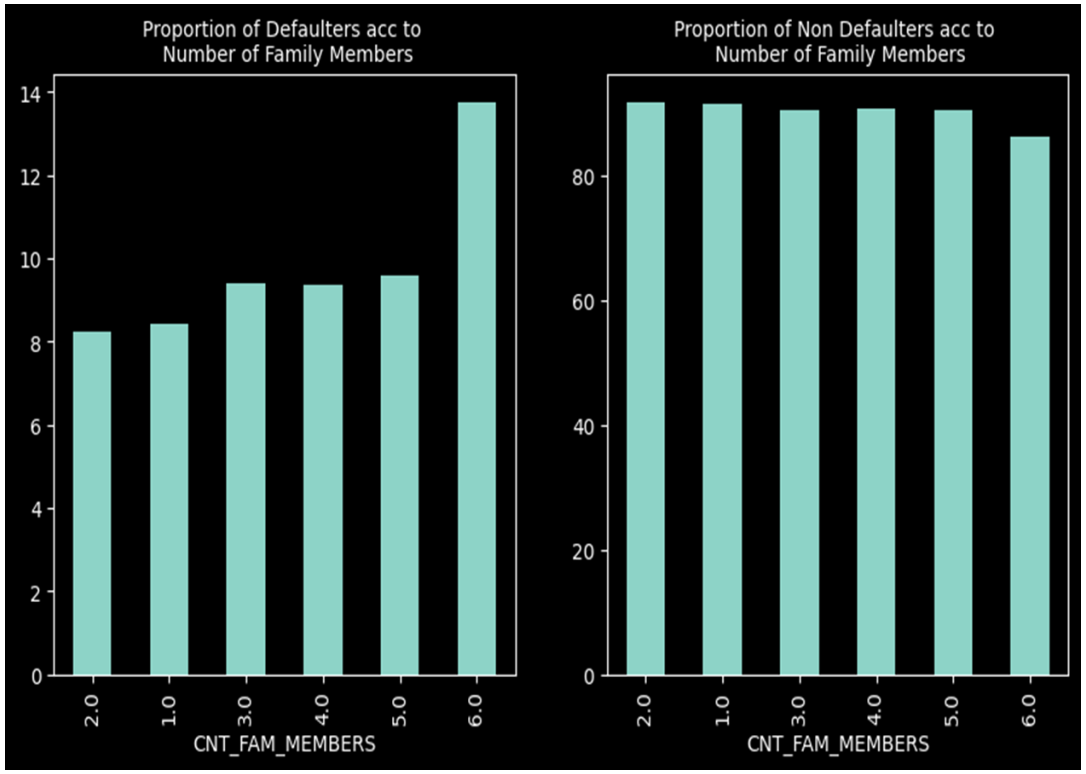
- The majority of the applicants are from other occupation types (32.23% approx) and the least number of applicants are IT staff (0.12% approx).
- The majority of the defaulters are Low-skill Laborers (18.83% approx) and the least number of defaulters are accountants (5.67% approx).

- The majority of the non-defaulters are from Accountants (94.33% approx) and the least number of non-defaulters are from LOW-Skill Laborere (81.18% approx).
- By looking at the graphs we can conclude that people with high levels of jobs are less likely to default.



Number of Family Members

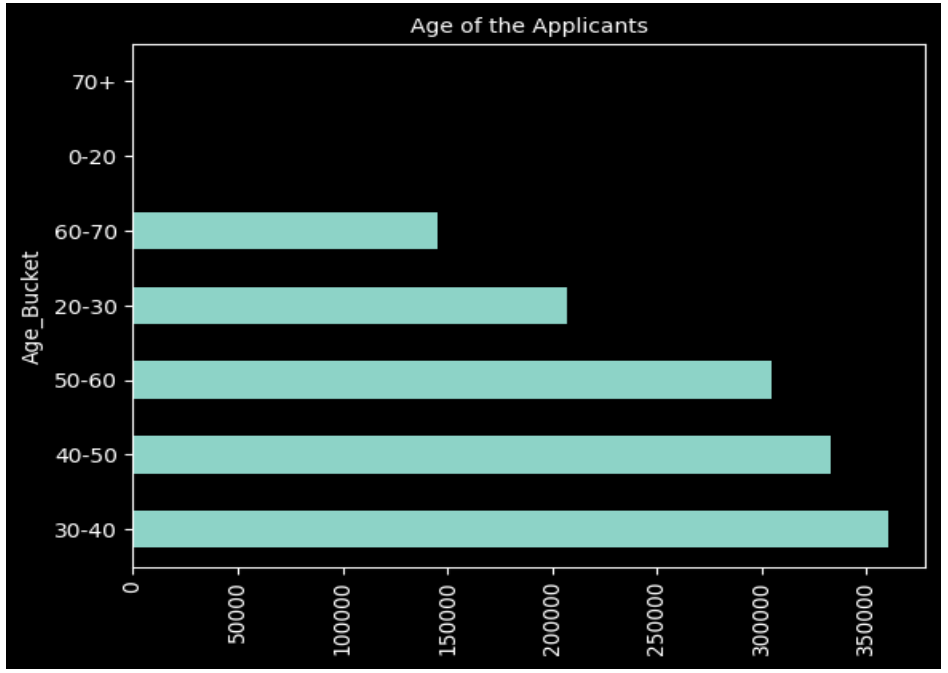


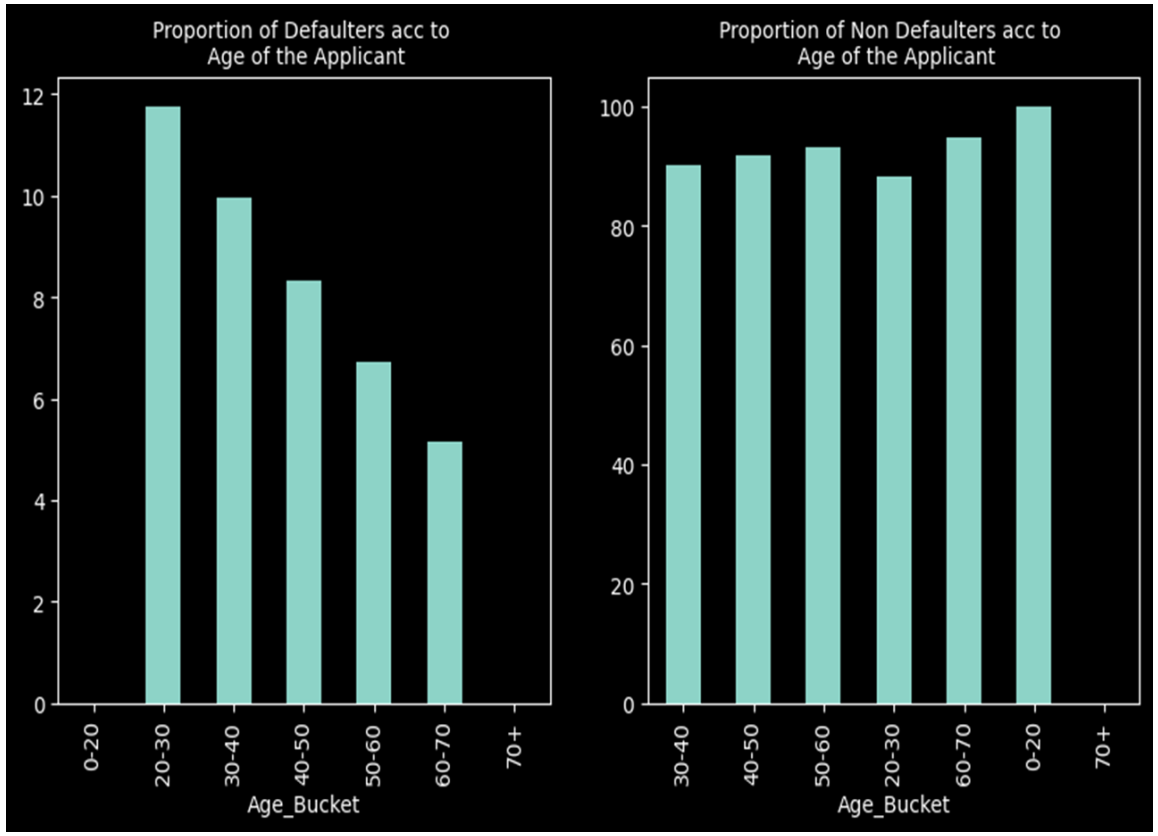


Findings:

- The most number of applicants have 2 family members (52.8% approx) and the least number of applicants have 6 or more family members (0.17% approx).
- The largest number of applicants that defaulted have 6 or more family members (13.75% approx) and the least that defaulted have 2 family members (8.25% approx).
- The most number of non-defaulters have 2 family members (91.75% approx) and the least number of non-defaulters have 6 or more family members (86.25% approx).
- Therefore we can conclude that the fewer the applicant has family members, the lesser there is a chance of default. This may be due to less financial dependents on the applicant.

Age





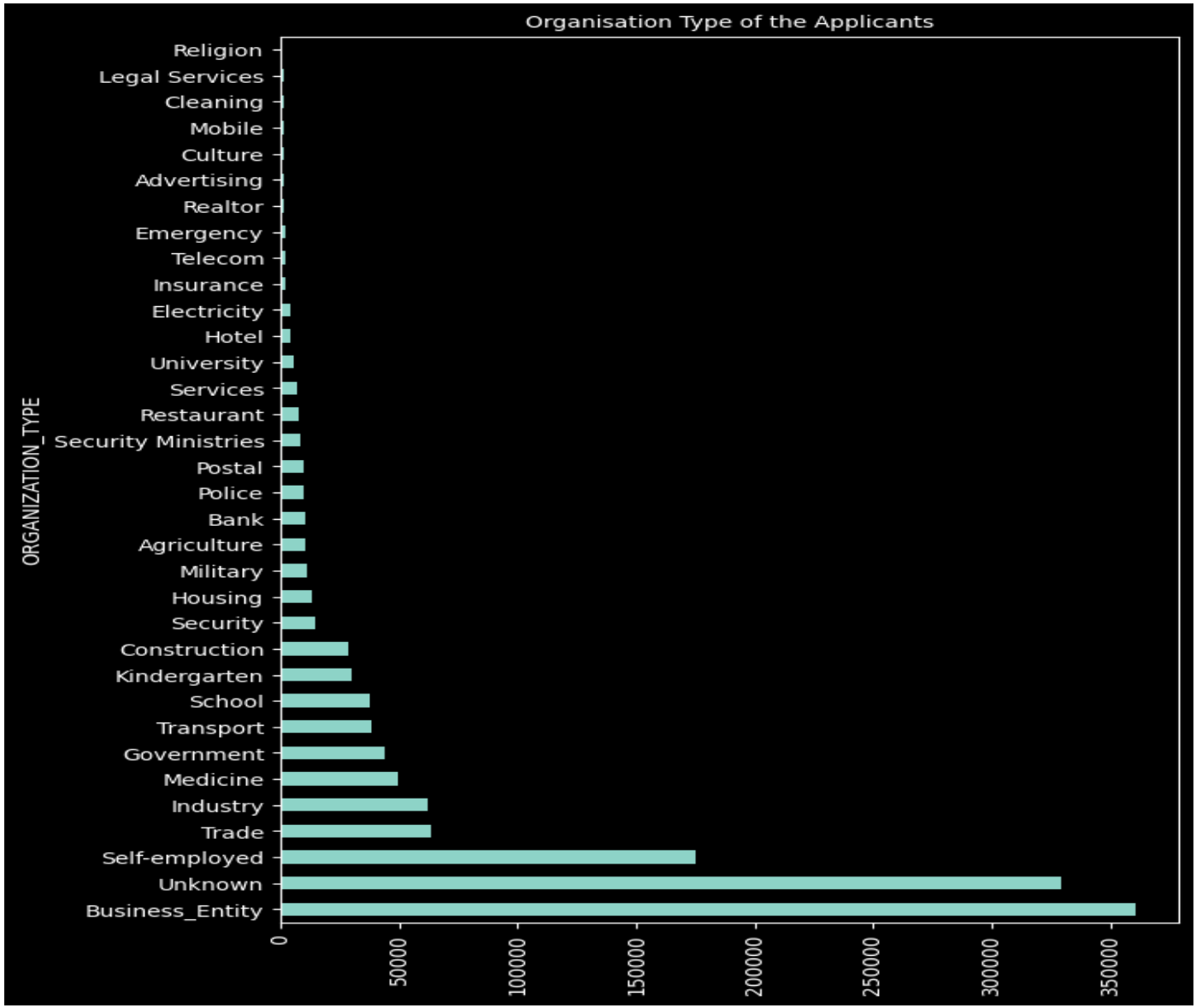
Findings:

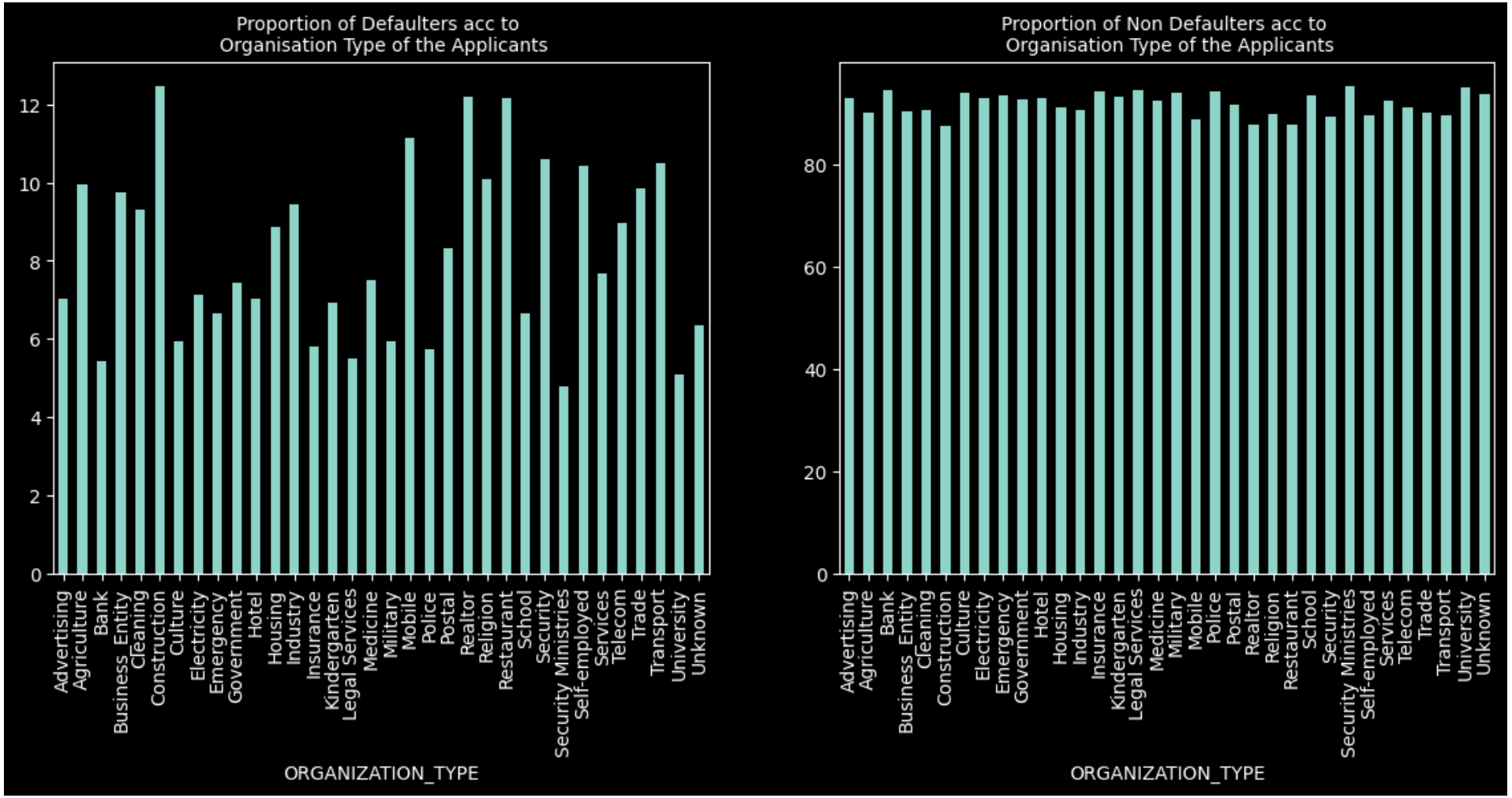
- The majority of the applicants are from age 30-40 (26.7% approx) and the least number of applicants are from 0-20 (0.0003% approx).
- The majority of defaulters are from 20-30 age (11.76% approx) and the least number of applicants are from 0-20 age (0%) which may be due to very less applicants and none of them defaulting, so we take 60-70 age (5.14% approx) as lowest.
- There is a general trend that with age increasing the number of defaulters is decreasing.

Organization Type

Findings:

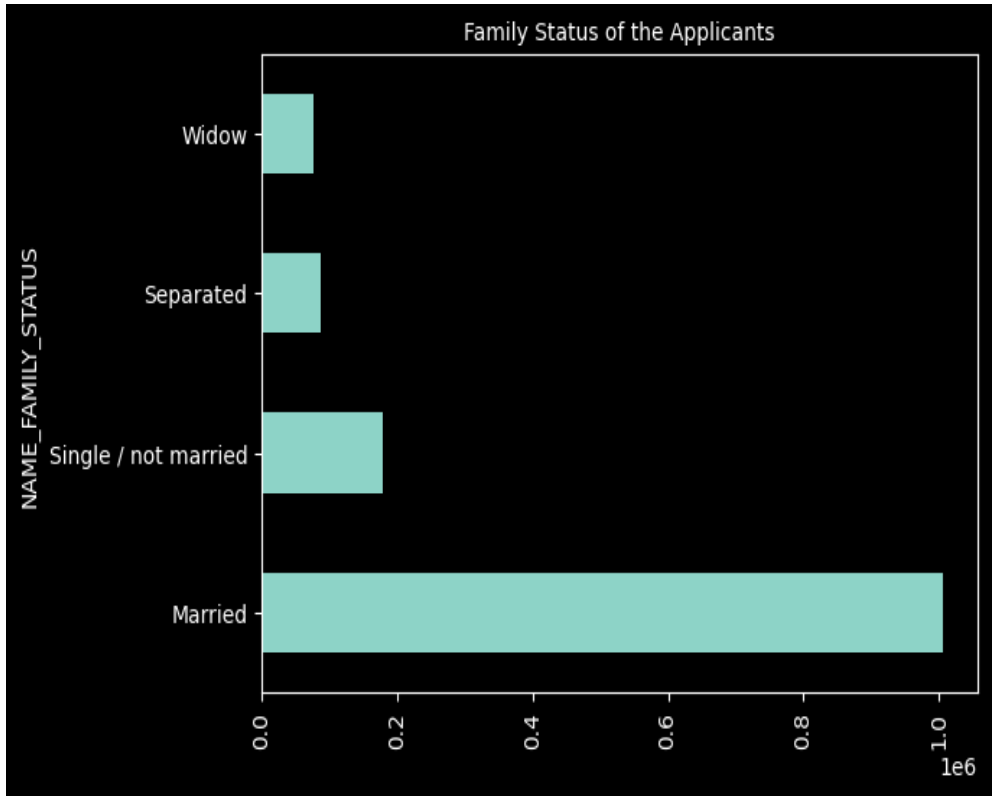
- The majority of the applicants are from Business Entity (26.71% approx) and the least number of applicants are from Religion (0.031% approx).
- The majority of the defaulters are from Construction (12.46% approx), Realtor (12.19% approx), and Restaurant (12.17% approx), and the least number of defaulters are from Security Ministries (4.8% approx).

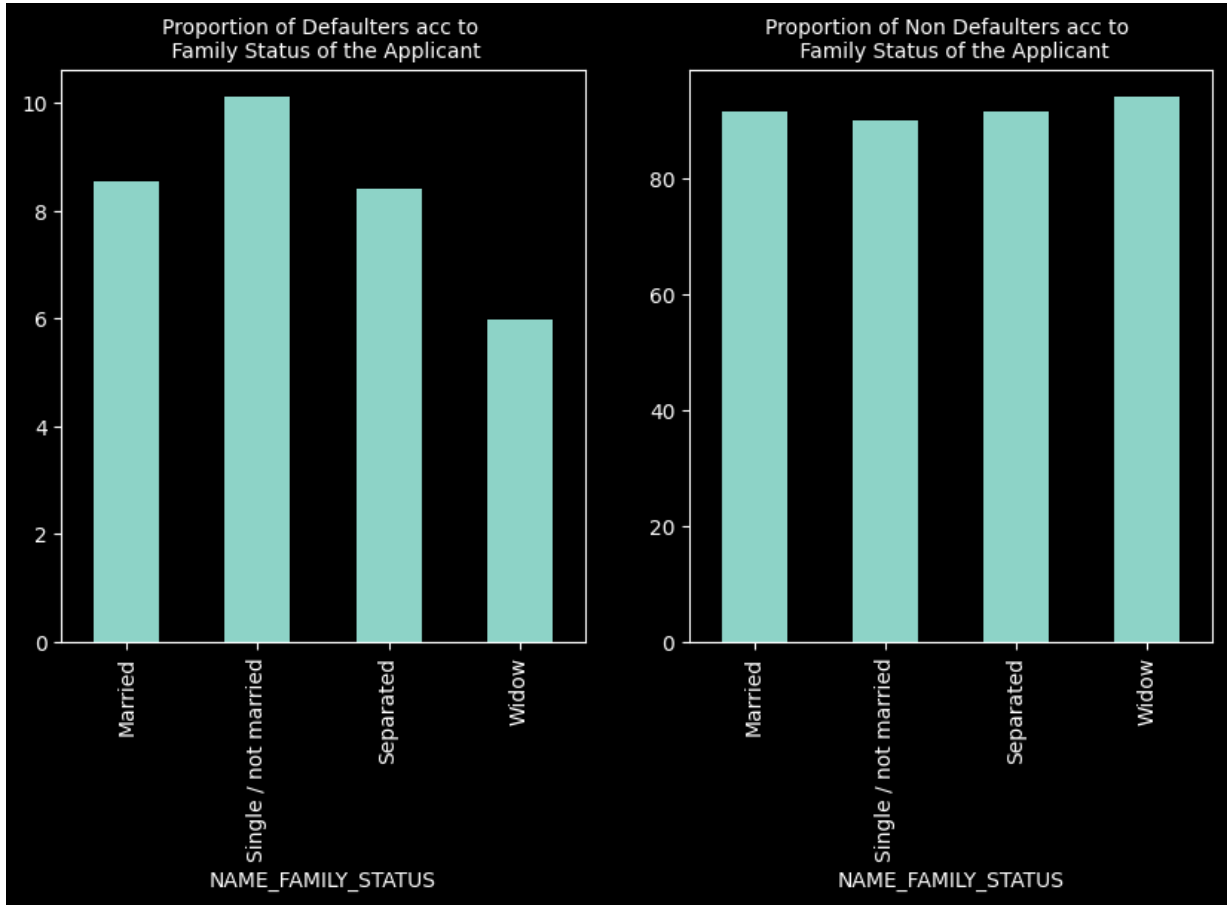




- We can see that those jobs that generally come under tier 1 or tier 2 jobs have less number of defaulters as compared to tier 3 jobs.

Family Status

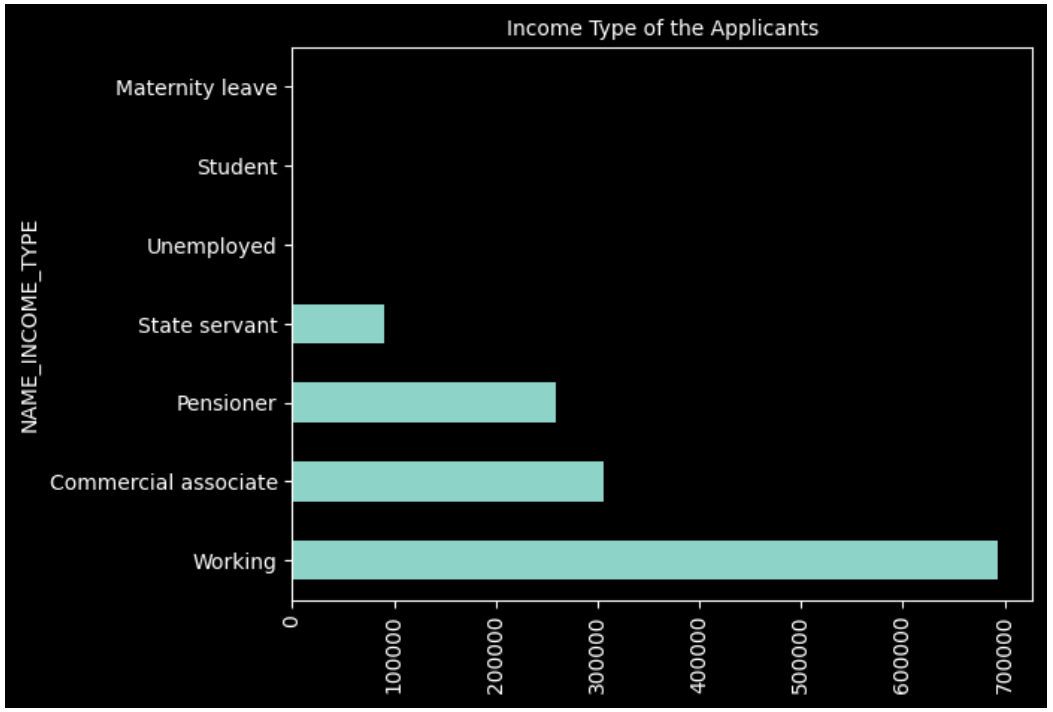


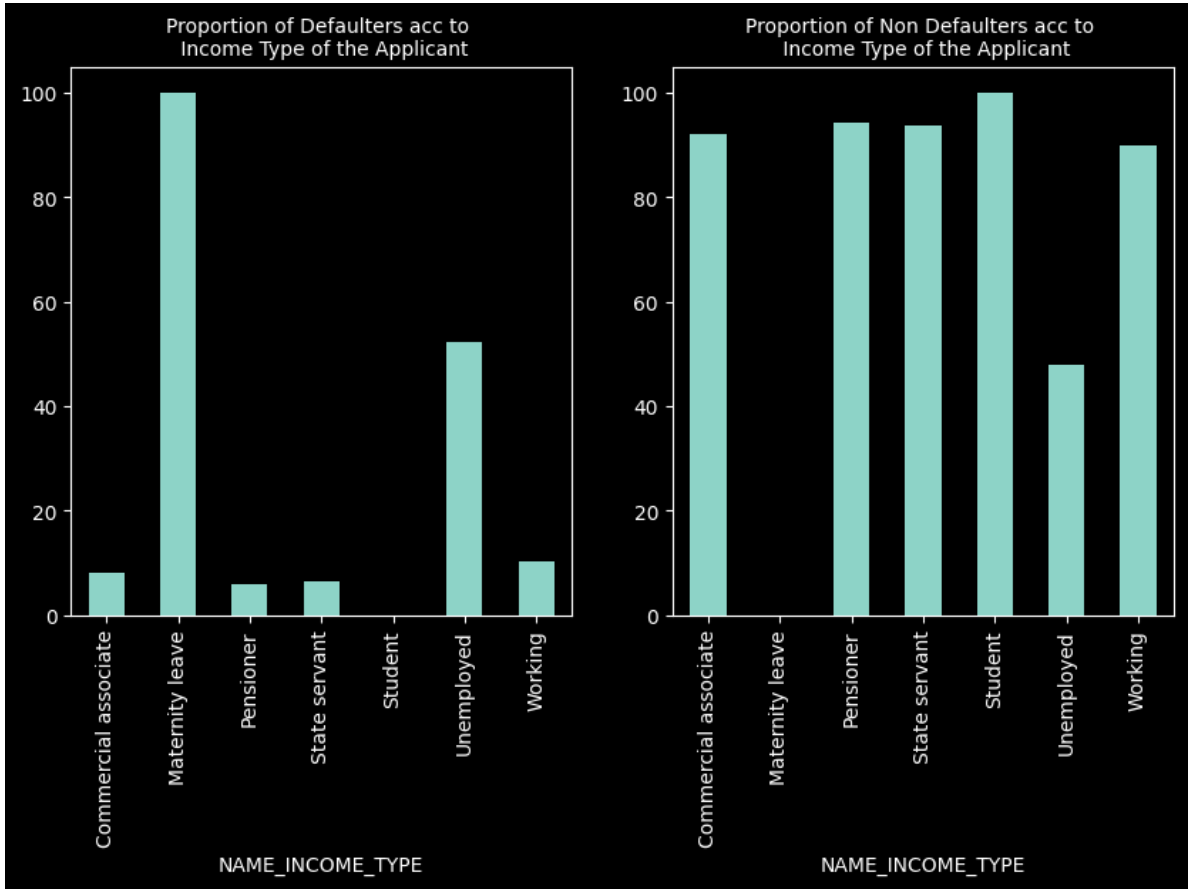


Findings:

- The majority of the applicant are married (74.58% approx) and the least number of applicants are widows (5.74% approx).
- The majority of defaulters are Single / not married (10.1% approx) and the least number of defaulters are widows (5.99% approx).

Income Type

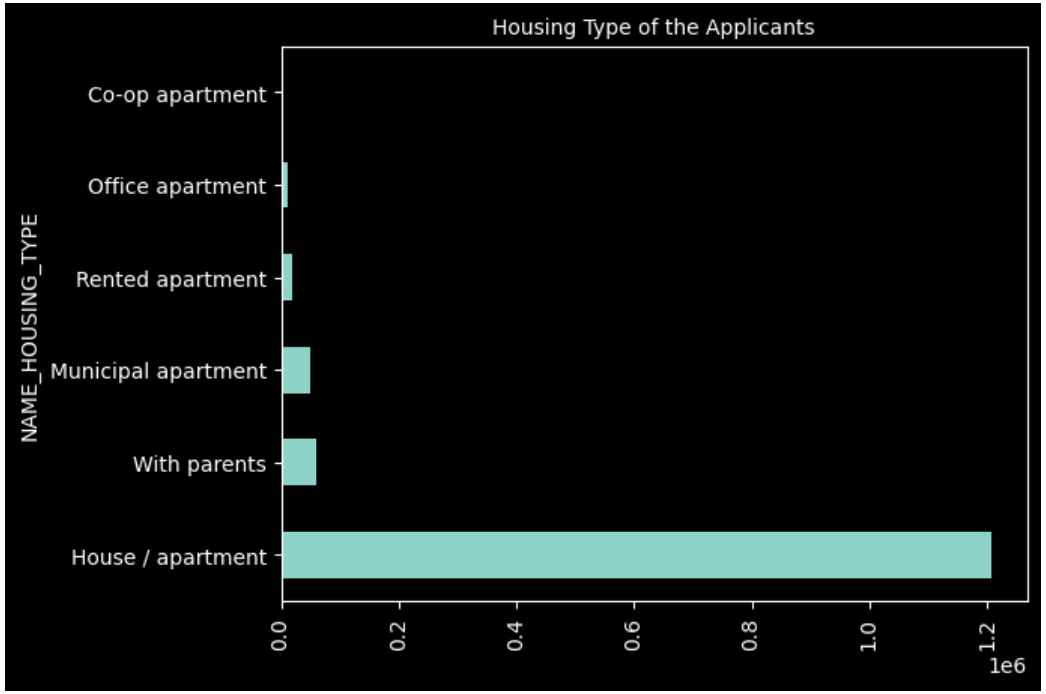


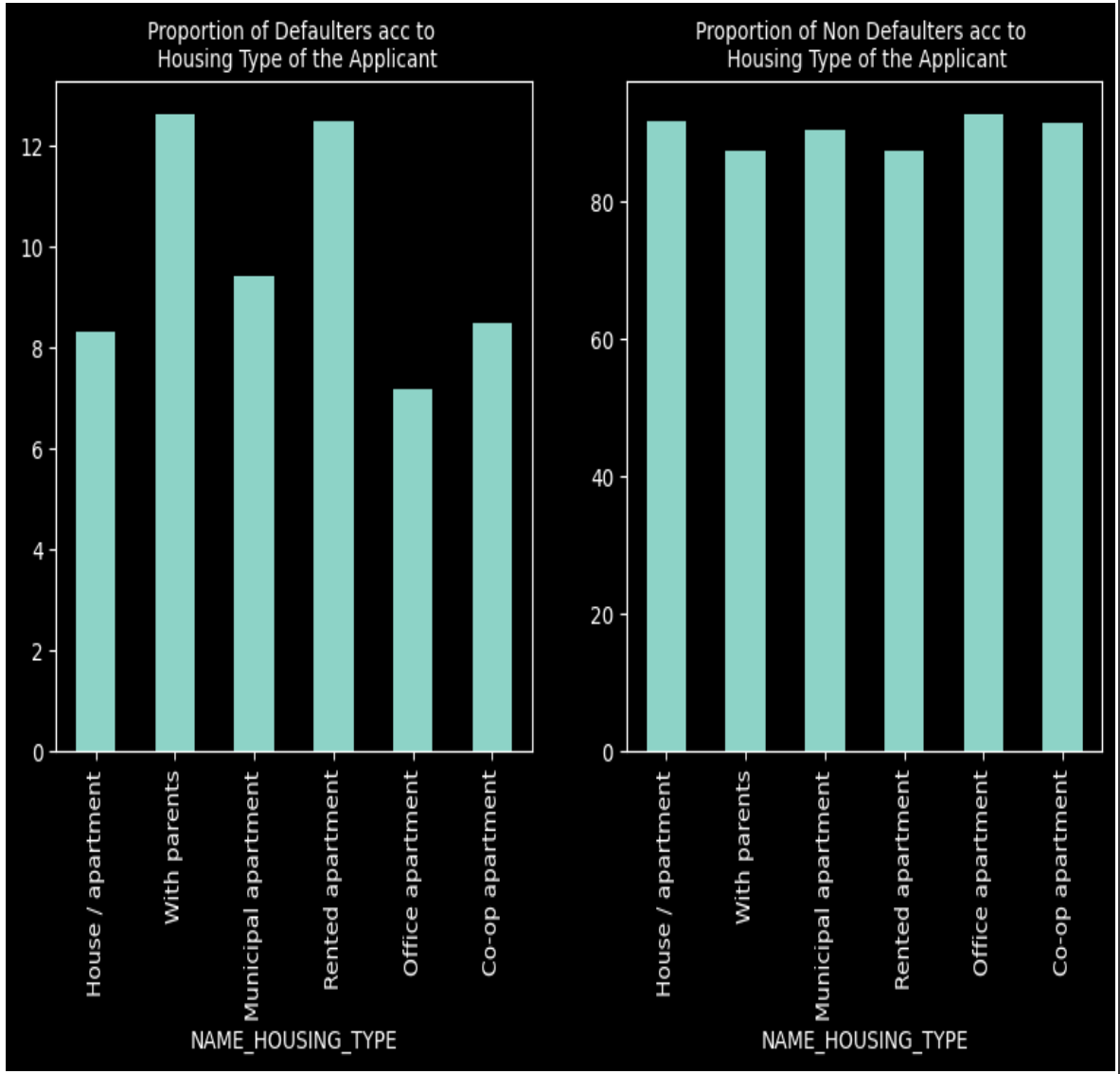


Findings:

- The majority of the applicants are working (51.34% approx) and the least number of applicants are on Maternity leave (0.001%).
- The majority of the defaulters are on maternity leave (100%), second highest are unemployed (52.14% approx). The lowest number of defaulters are students(0%) and the second lowest are pensioners (5.80% approx).
- We can see that those applicants who do not have a regular source of income except students tend to default.

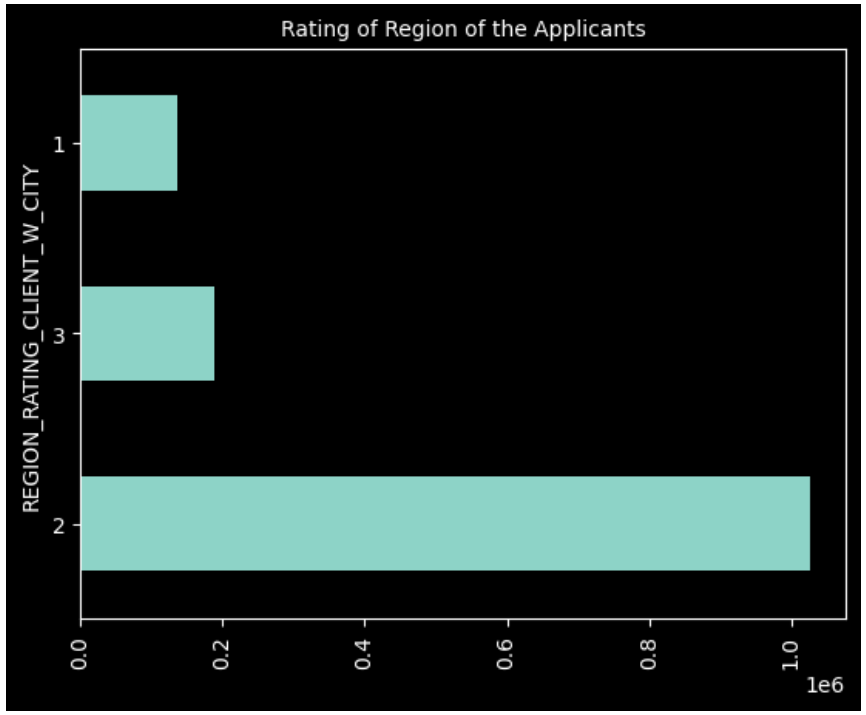
Housing Type

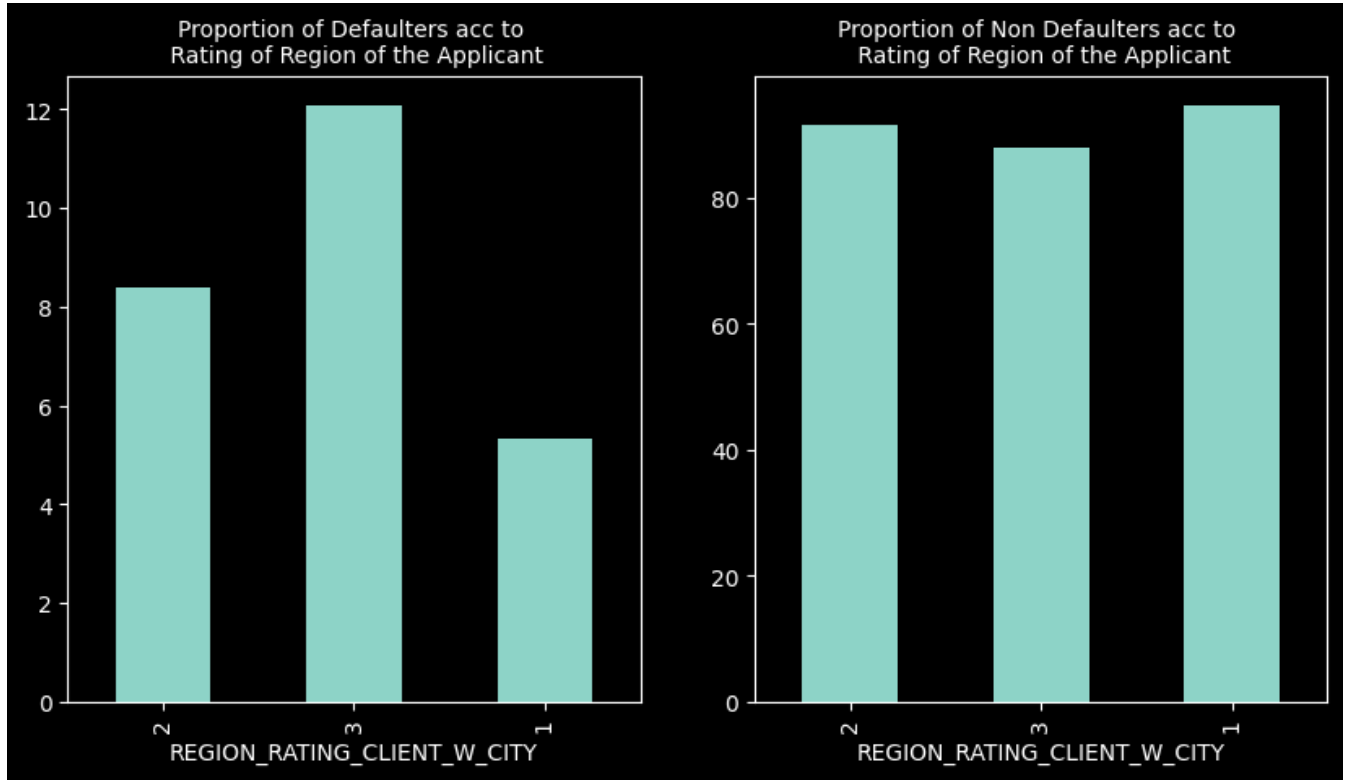




Findings:
The majority of the applicants live in houses/ apartment (89.43% approx) and the least number of applicants live in Co-op/apartments (0.31% approx).
The majority of Defaulters are living with parents (12.63% approx) or in a rented apartment (12.48% approx), while the least number of defaulters live in House/apartment (8.31% approx) or in a co-op apartment (8.48% approx).

Rating of Region

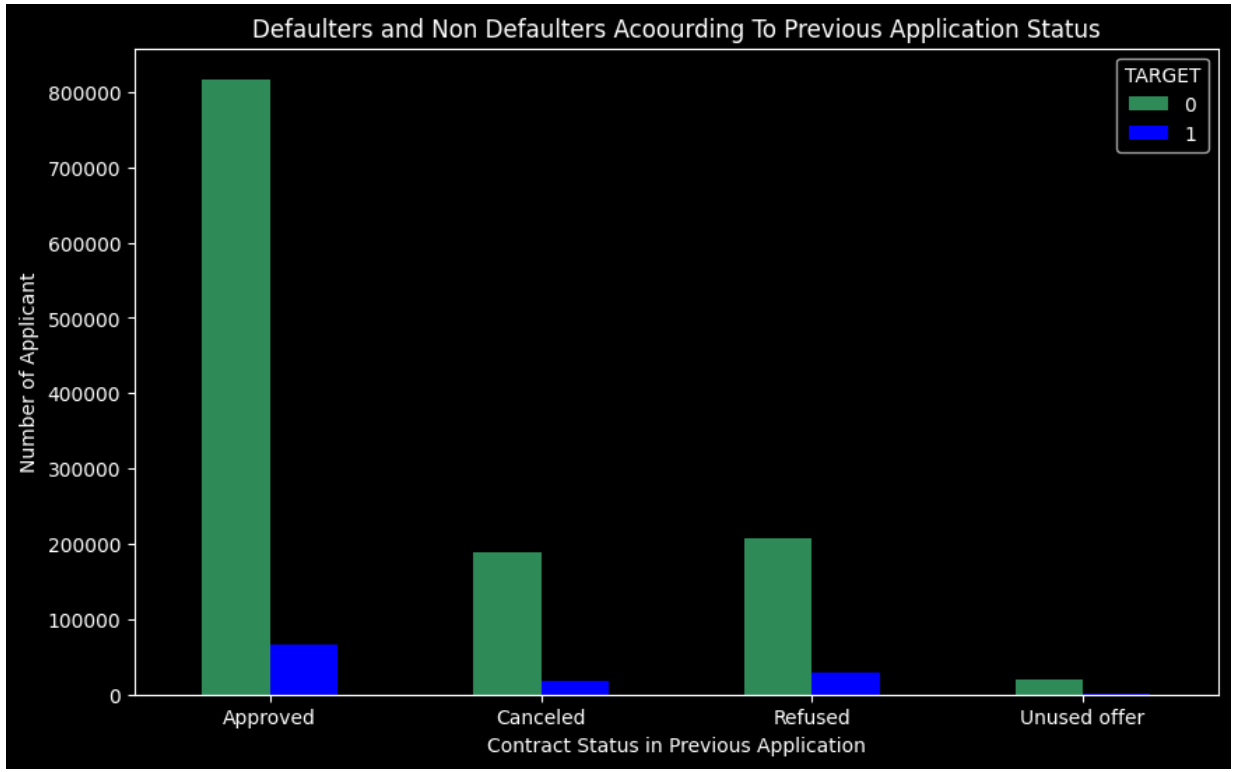


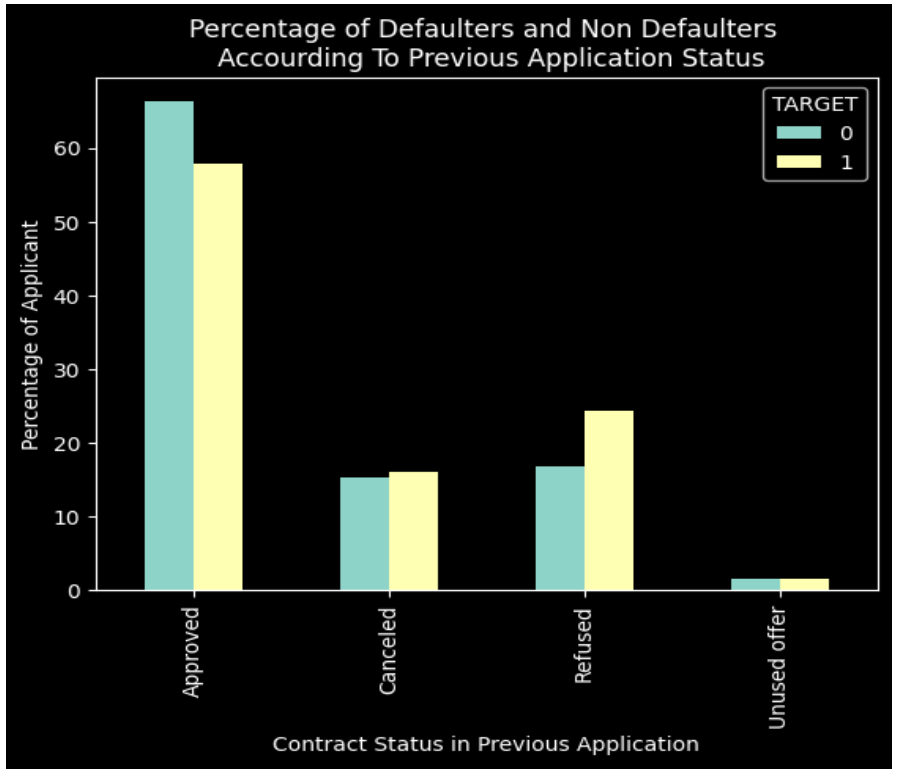


Insights:

- The majority of applicants are from tier 2 cities (75.97% approx), and the least are from tier 1 cities (10.08% approx).
- The majority of applicants that default are from tier 3 cities (12.07% approx) and the least that default are from tier 1 cities (5.33% approx.).
- There is a general trend that with an increase in the tier of the city the number of defaulters is increasing.

Contract Status in Previous Application

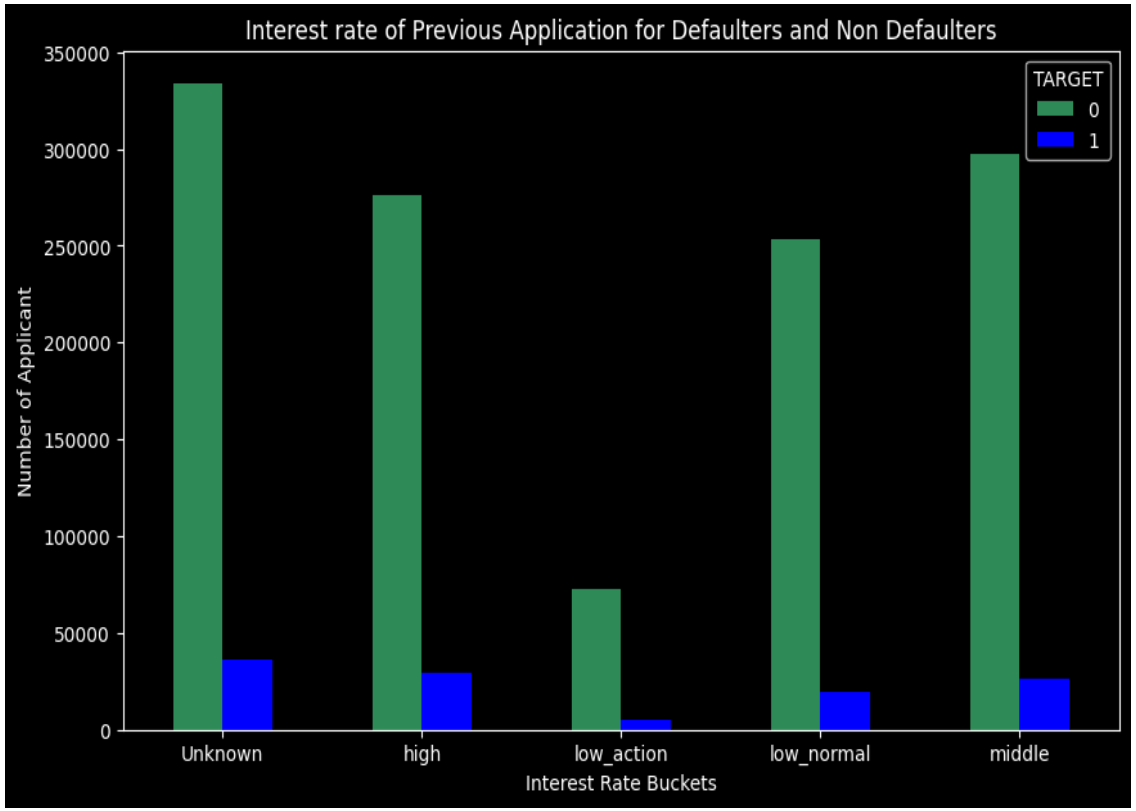


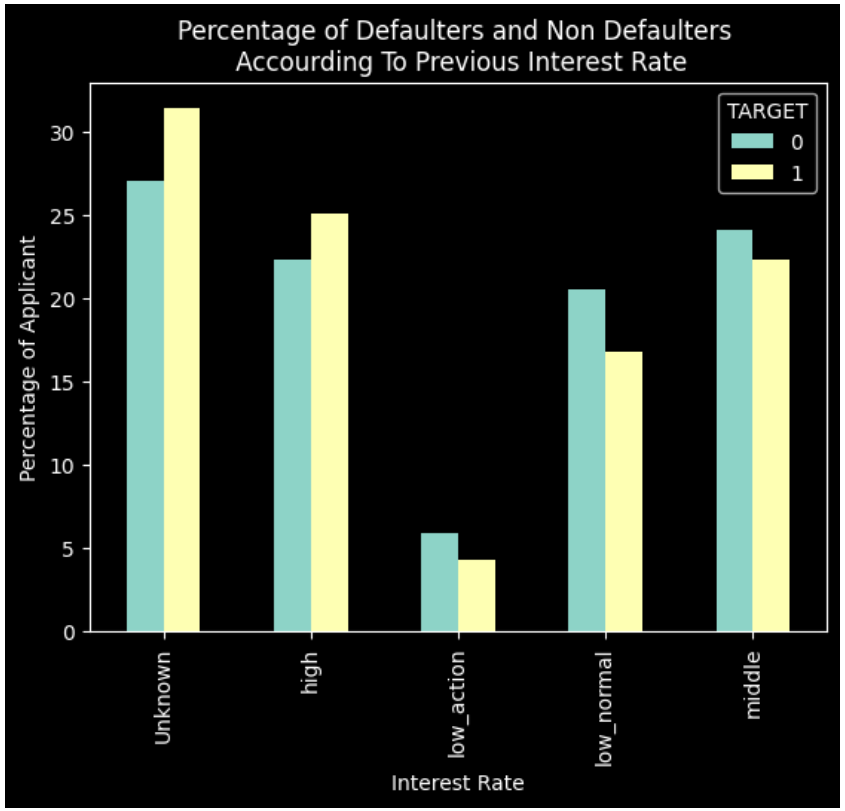


Insights:

- The majority of previous applications of every status type were non-defaulters in numbers.
- Even though the number of refused applicants who turned out to be defaulters is less than the number who turned out to be non-defaulters, the proportion of defaulters is greater than the proportion of non-defaulters. This means that in the total number of defaulters, the applicants who were previously rejected held a significant proportion. The same goes for previously Cancelled applications.

Interest Rate of Previous Application

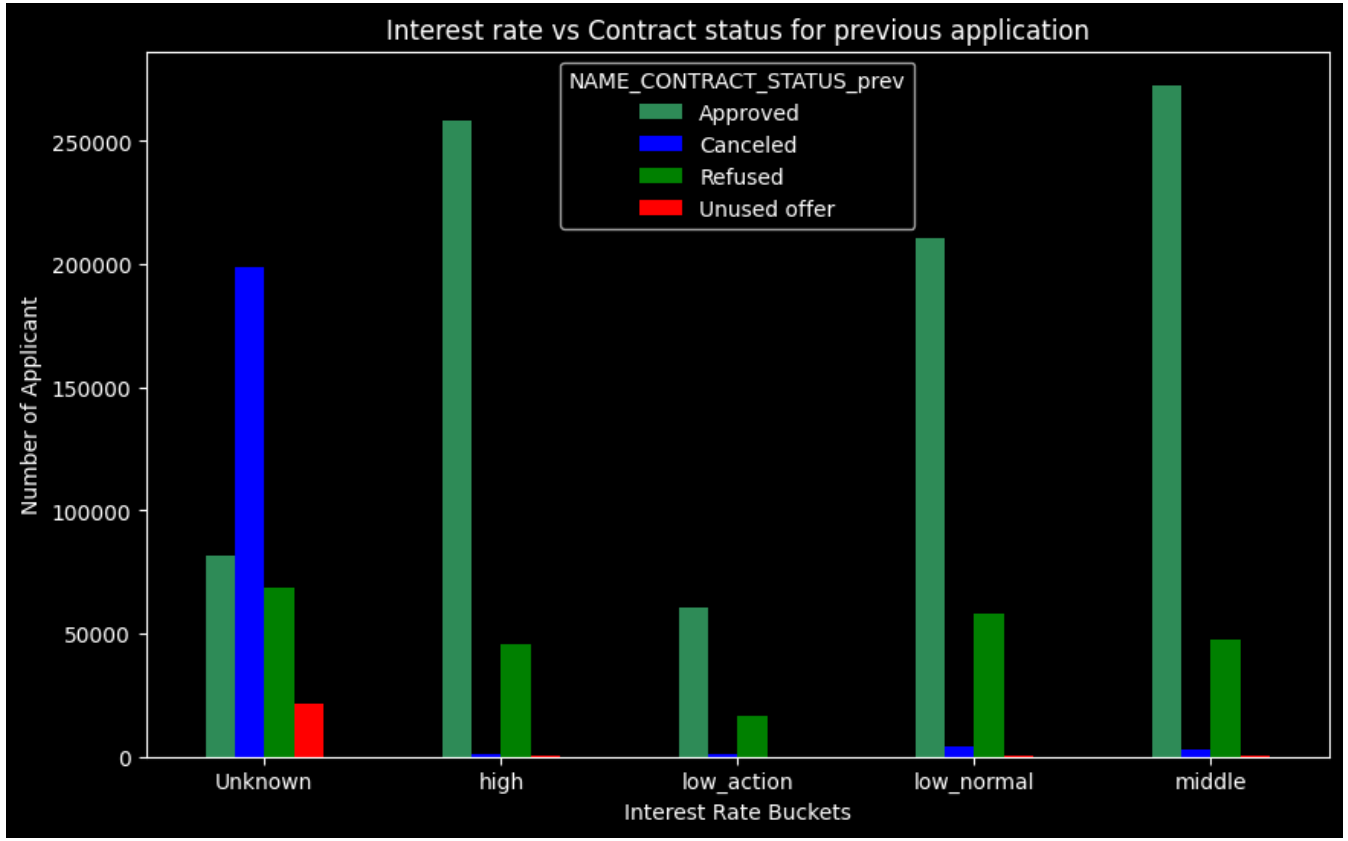




Insights:

- The number of non-defaulters is greater than defaulters for every interest rate in numbers.
- For unknown interest rate the proportion in defaulter is greater than the proportion in non-defaulter. The same goes for high interest rates. This means the chances of an applicant turning default is higher for unknown and high interest rates.

Interest Rate vs Contract Status for the Previous Application

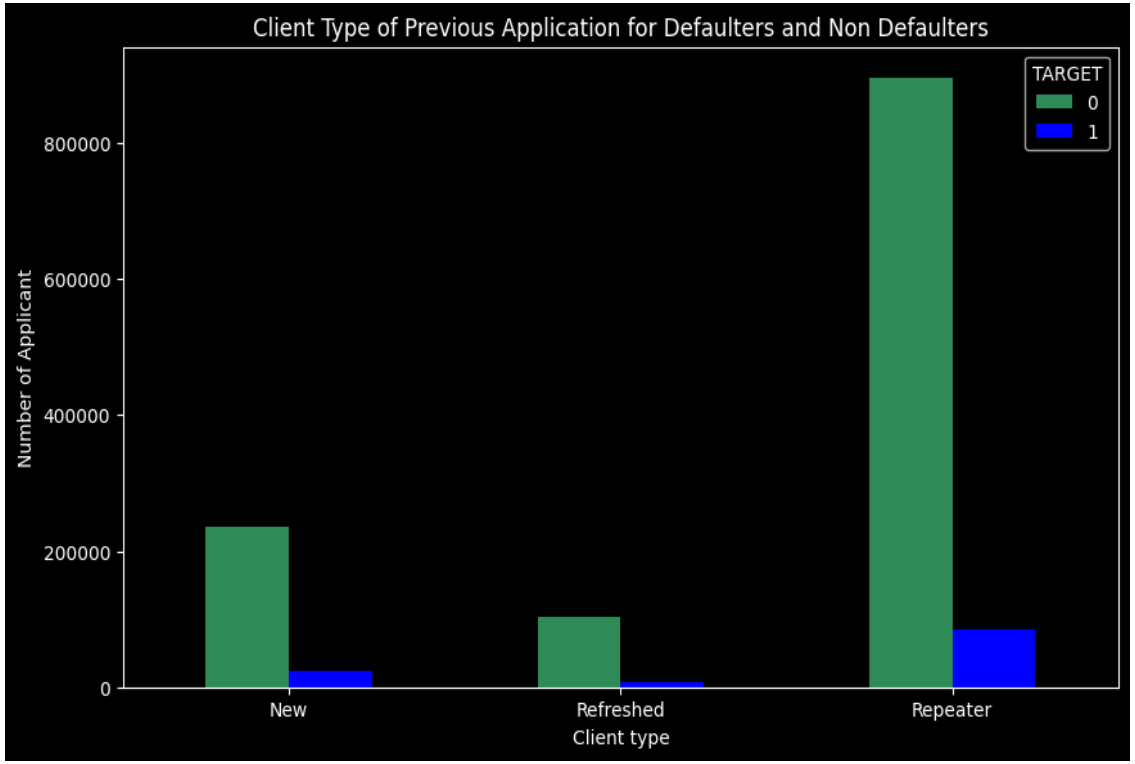


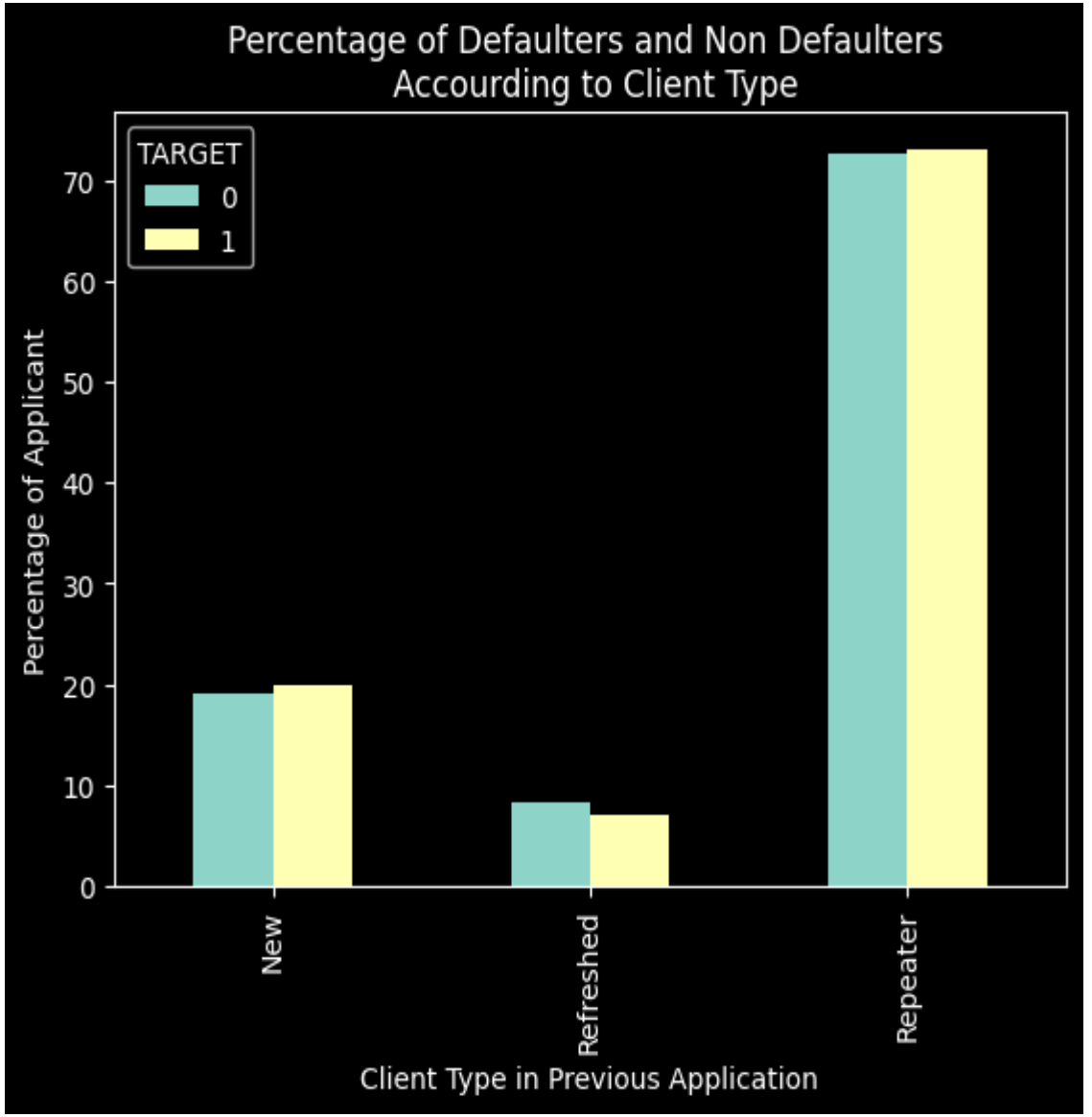
Insights:

- For Unknown Interest Rate: The majority of the previous applications were canceled during the process. Significantly lower amounts of applications were approved. A significant amount of previous applications that were unused offers belonged to unknown interest rates.
- For High Interest Rate: The majority of applications from high interest rates were approved. A very low amount of applications were canceled or refused offers.
- For Low_action Interest Rate: Even though the majority of applications were approved, the number relative to other interest rates is low. There were negligent amounts of canceled or refused orders. A few of the applications were refused, which relative to refused offers of other interest rates, is of significant amount.

- For Low_normal Interest Rate: Most of the applicants with low_normal interest get approved, but a significant amount out of refused applications also belongs to low_normal interest. Very few applicants are canceled or unused.
- For Middle Interest Rate: Most of the applicants with low_normal interest get approved, but a significant amount out of refused applications also belong to low_normal interest. Very few applicants are canceled or unused.

Client Type of Previous Application

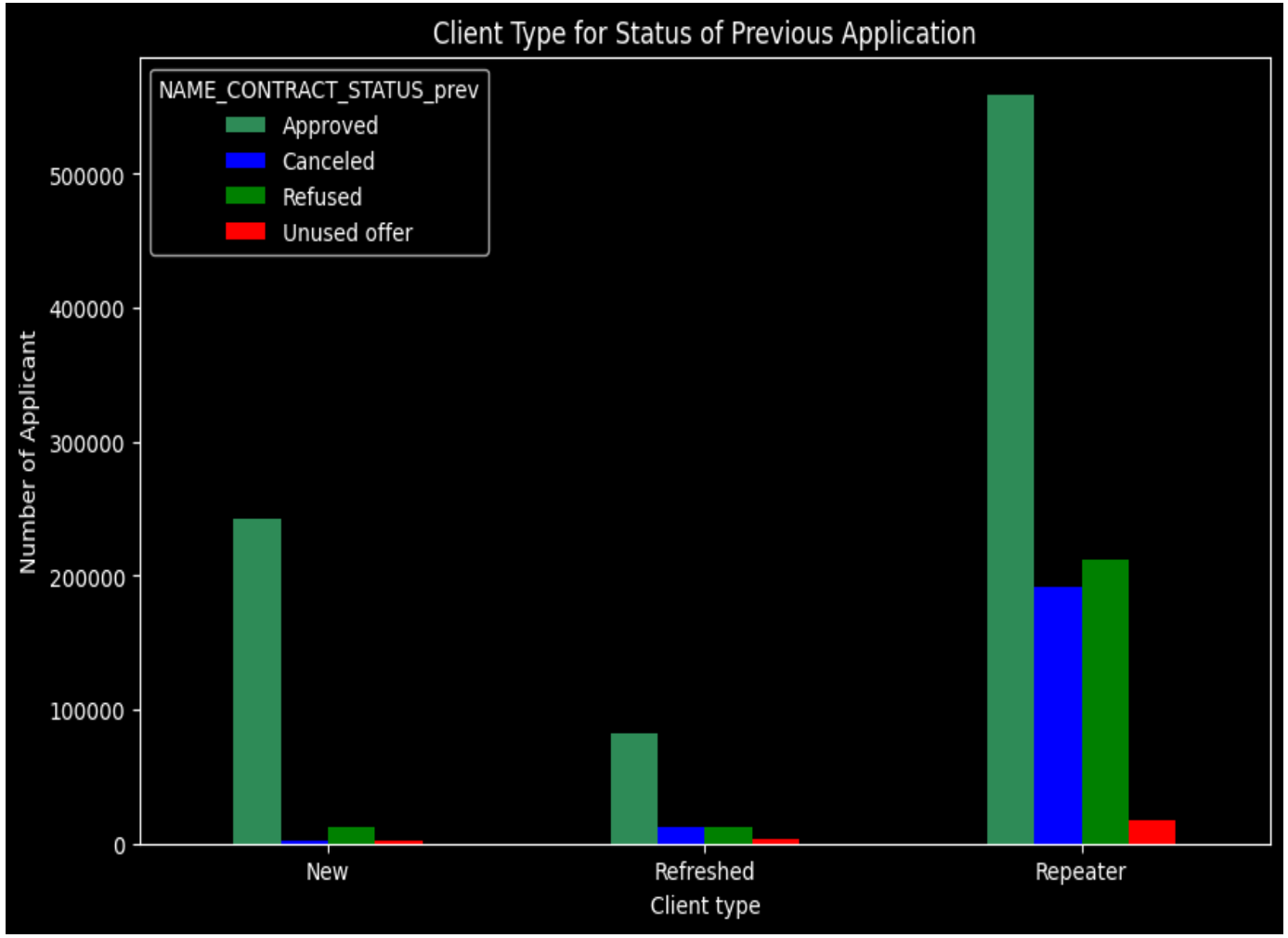




Insights:

- The majority of applicants are repeaters, and the least number of applicants are refreshed.
- For both defaulters and non-defaulters, the composition of client type remains the same.

Client Type for Status of Previous Application

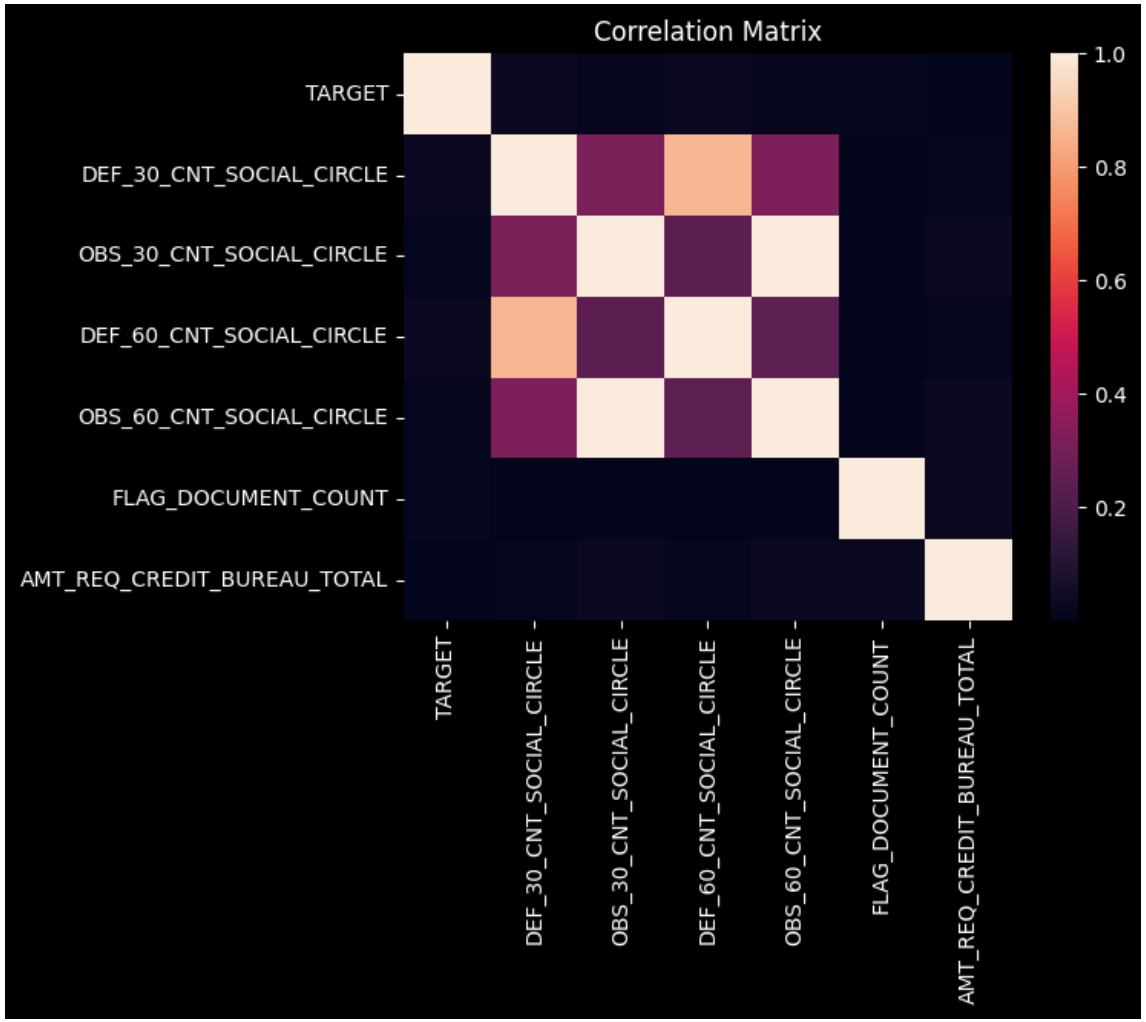


Insights:

- New Applicants: The majority of the new applicants get approval on their applications. With very few getting refused and almost negligible getting canceled or unused.
- Refreshed Applicant: Majority of the refreshed applicants get approval with very few getting refused, canceled, or unused.
- Repeater: Majority of the applicants are repeaters therefore is has a major portion in every status be it approval, refused, canceled, or unused, with a major portion in approved followed by refused, then canceled, and then unused.

MULTIVARIATE ANALYSIS:

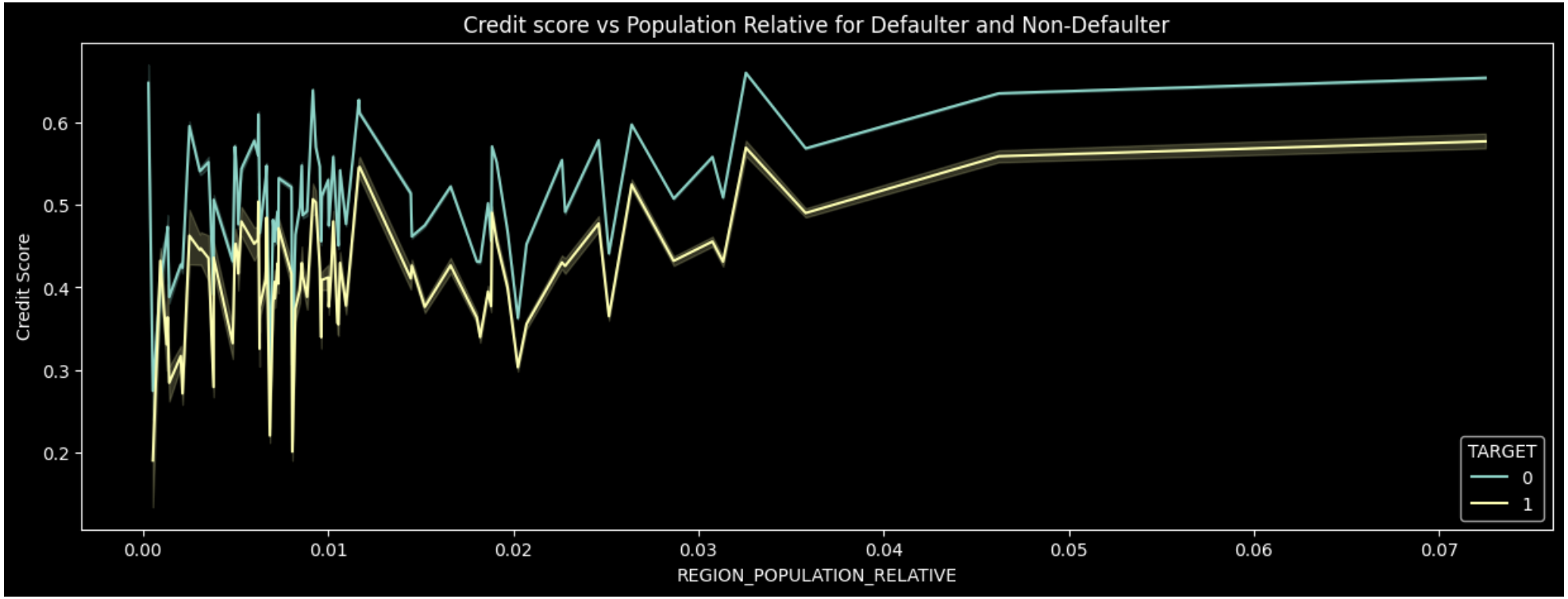
Correlation Matrix



Insights:

Target variable is not correlated to any of these variables namely DEF_30_CNT_SOCIAL_CIRCLE, OBS_30_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, FLAG_DOCUMENT_COUNT, AMT_REQ_CREDIT_BUREAU_TOTAL.

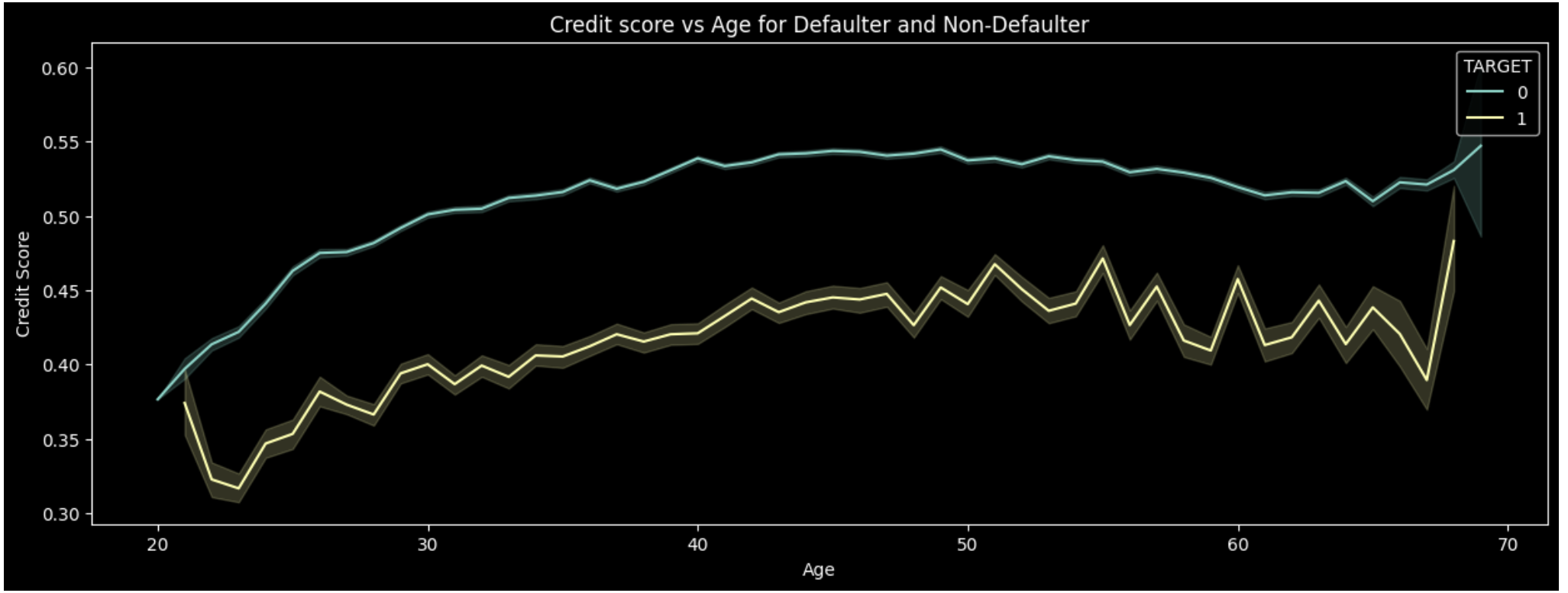
Credit Score vs Population Relative for Defaulter and Non-Defaulter



Insights:

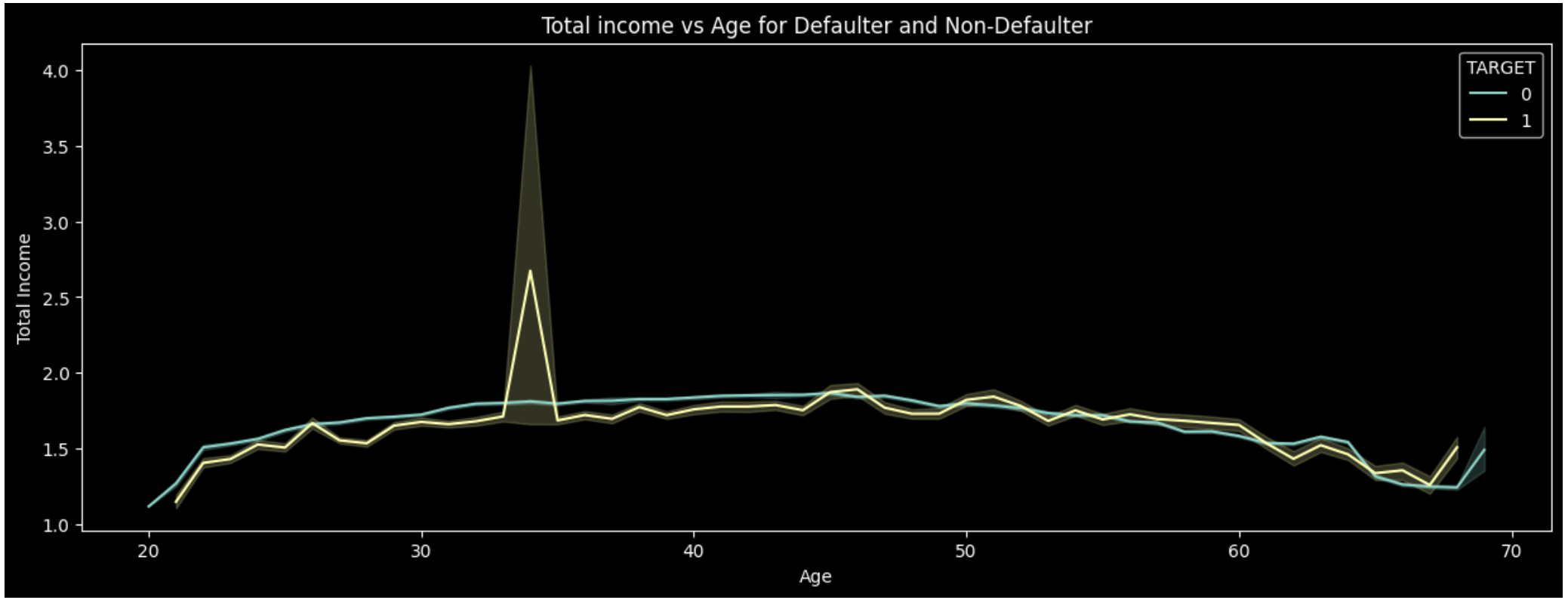
- The higher the credit score and more region population relative, the lower the chances for default.
- Higher Credit score in general means fewer chances of getting a default.

Credit Score vs Age for Defaulter and Non-Defaulter



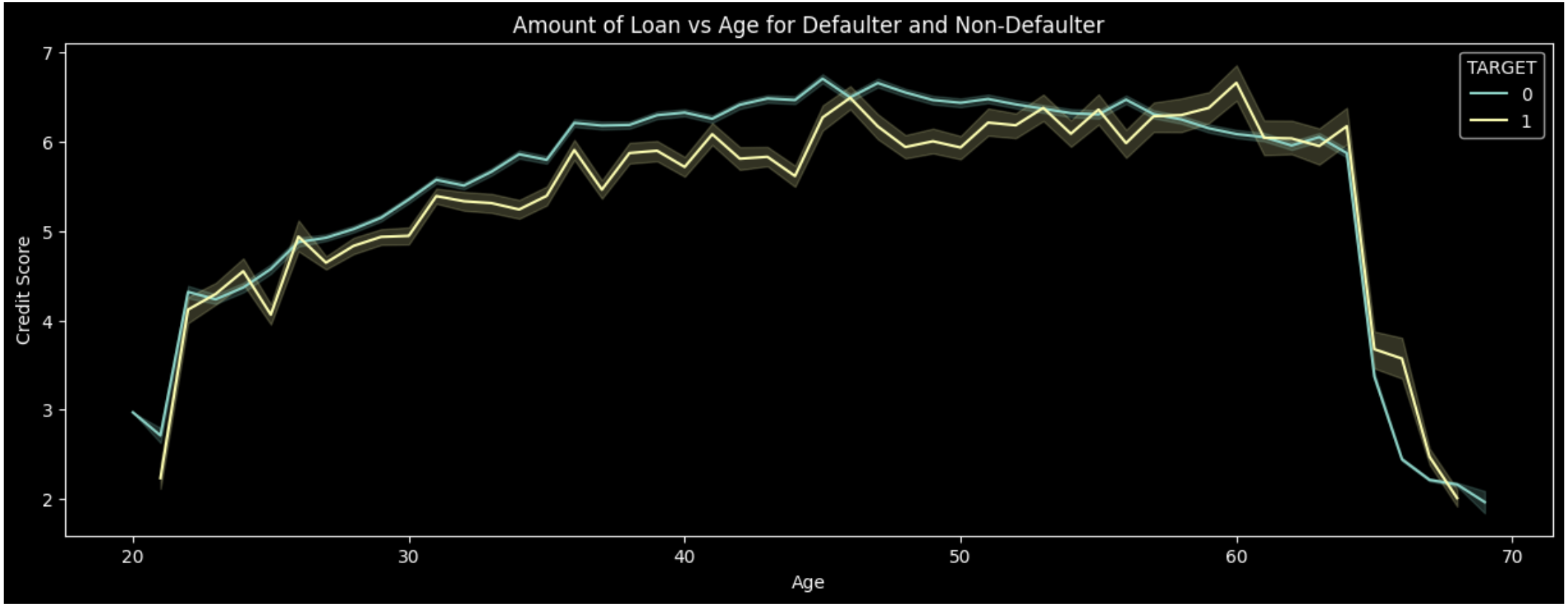
Insight:
With age, the credit score increases. Applicants with default generally have a lower credit score than people of the same age group.

Total Income vs Age for Defaulter and Non-Defaulter



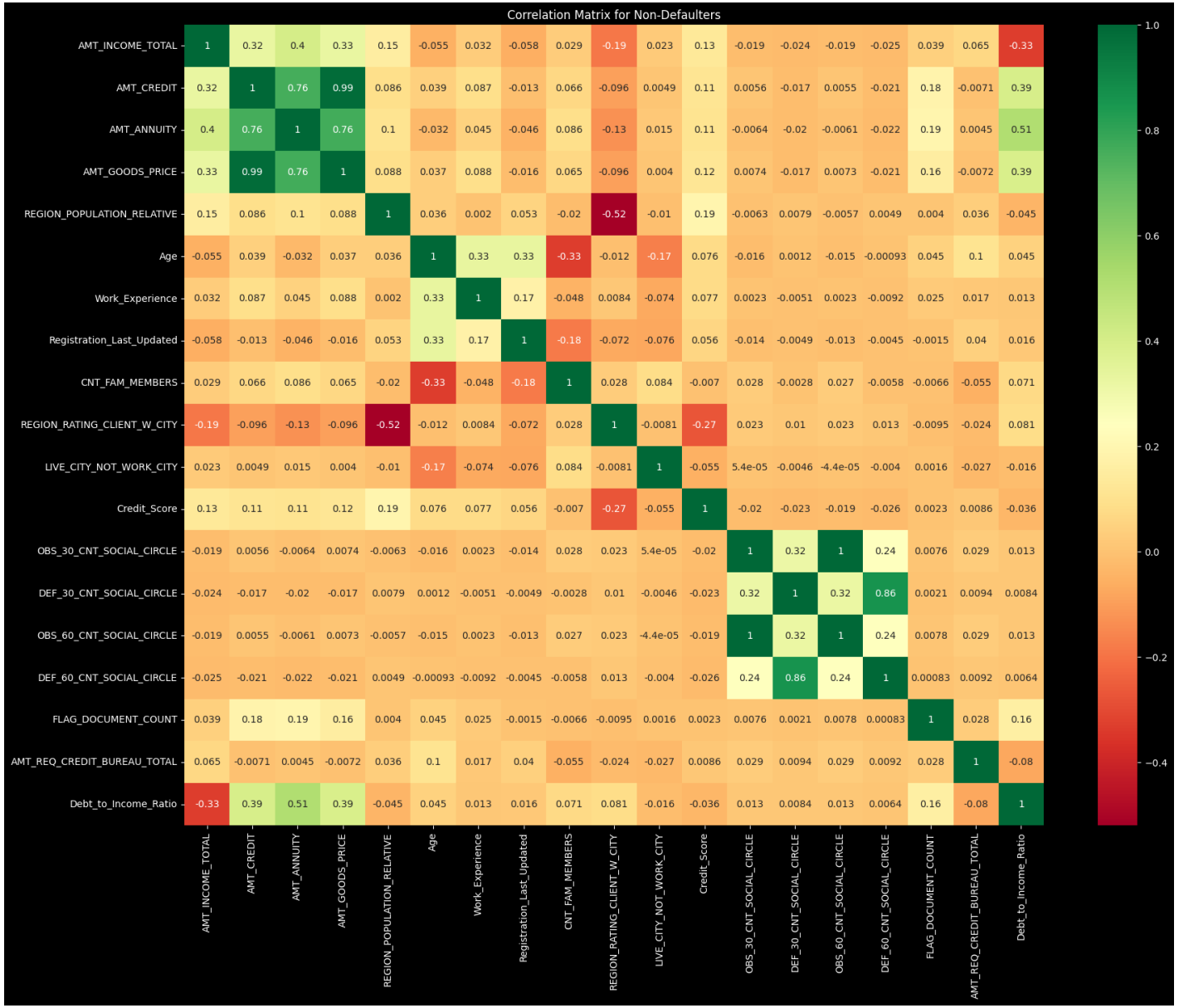
Insight:
There is not much of a trend between age and total income on an applicant getting default.

Amount of Loan vs Age for Defaulter and Non-Defaulter

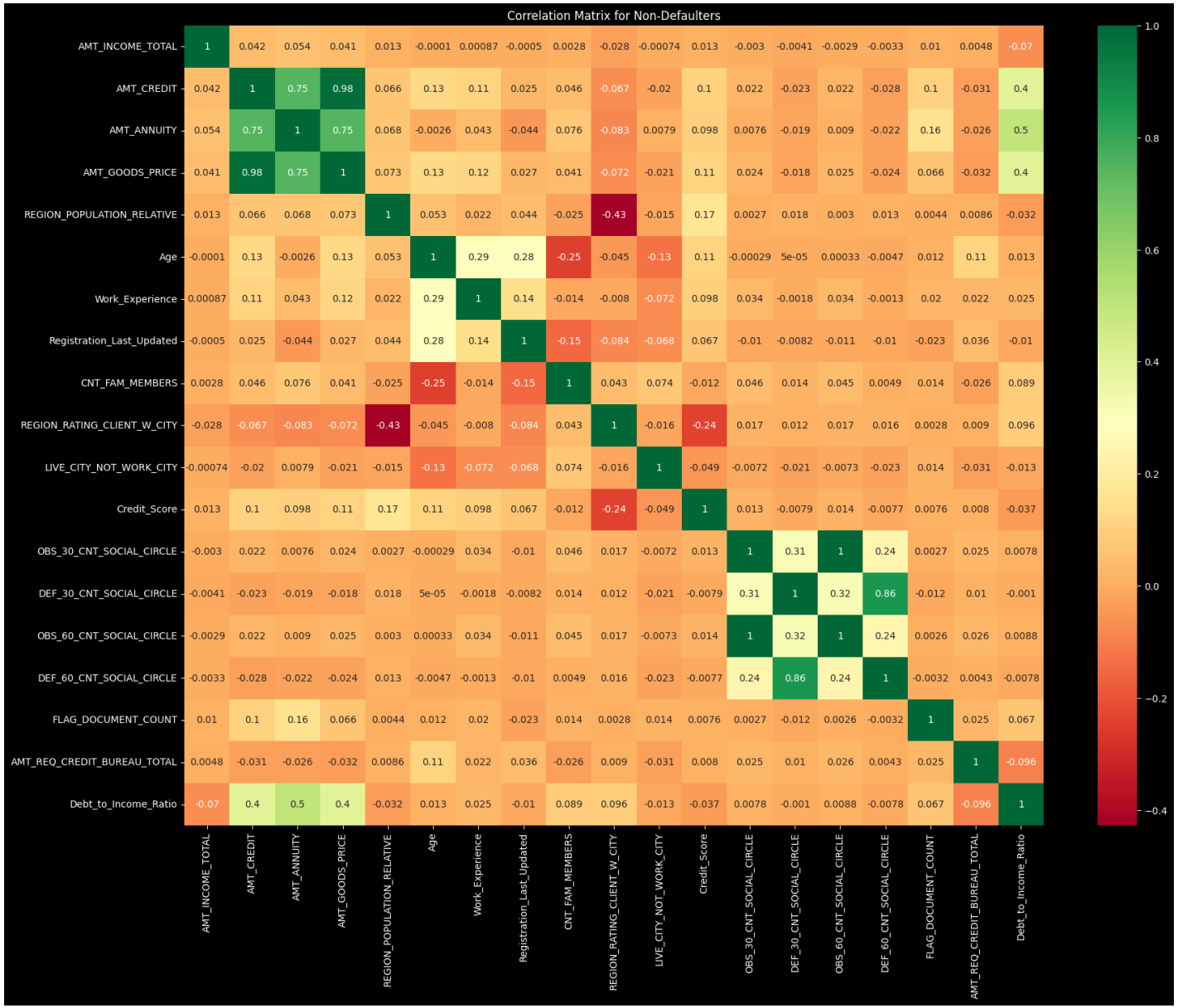


Insight:
The trend for loan amount first increases with an increase in age, maybe due to increasing salary, and then suddenly decreases, maybe due to retirement.

Correlation Matrix for Non-Defaulters



Correlation Matrix for Non-Defaulters

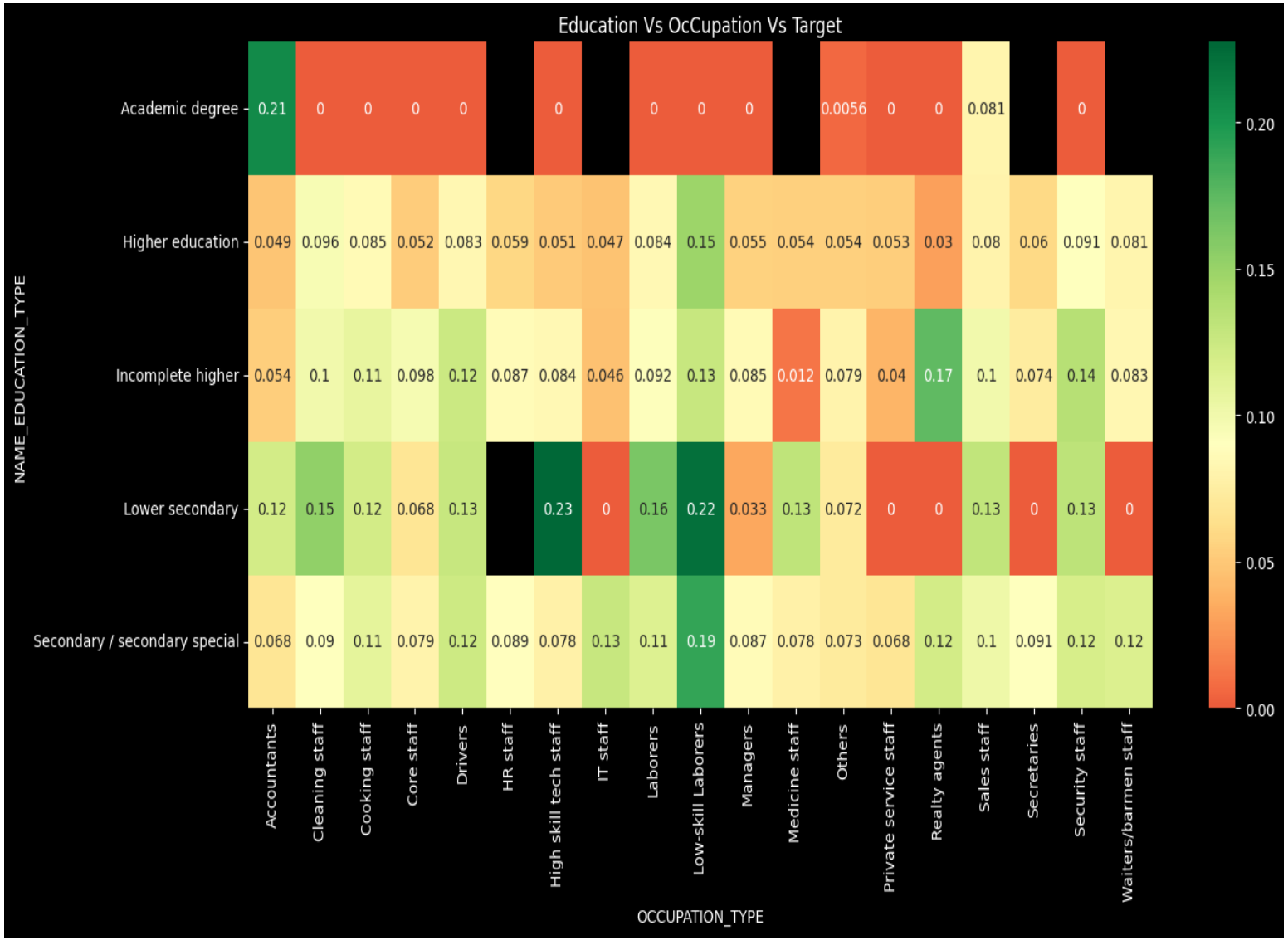


Insights:

- The debt-to-income ratio is more negatively correlated for defaulters than non-defaulters.
- The debt-to-income ratio is more positively correlated with the amount of credit and annuity for non-defaulters than for defaulters.

CORRELATION MATRIX FOR TARGET VARIABLE:

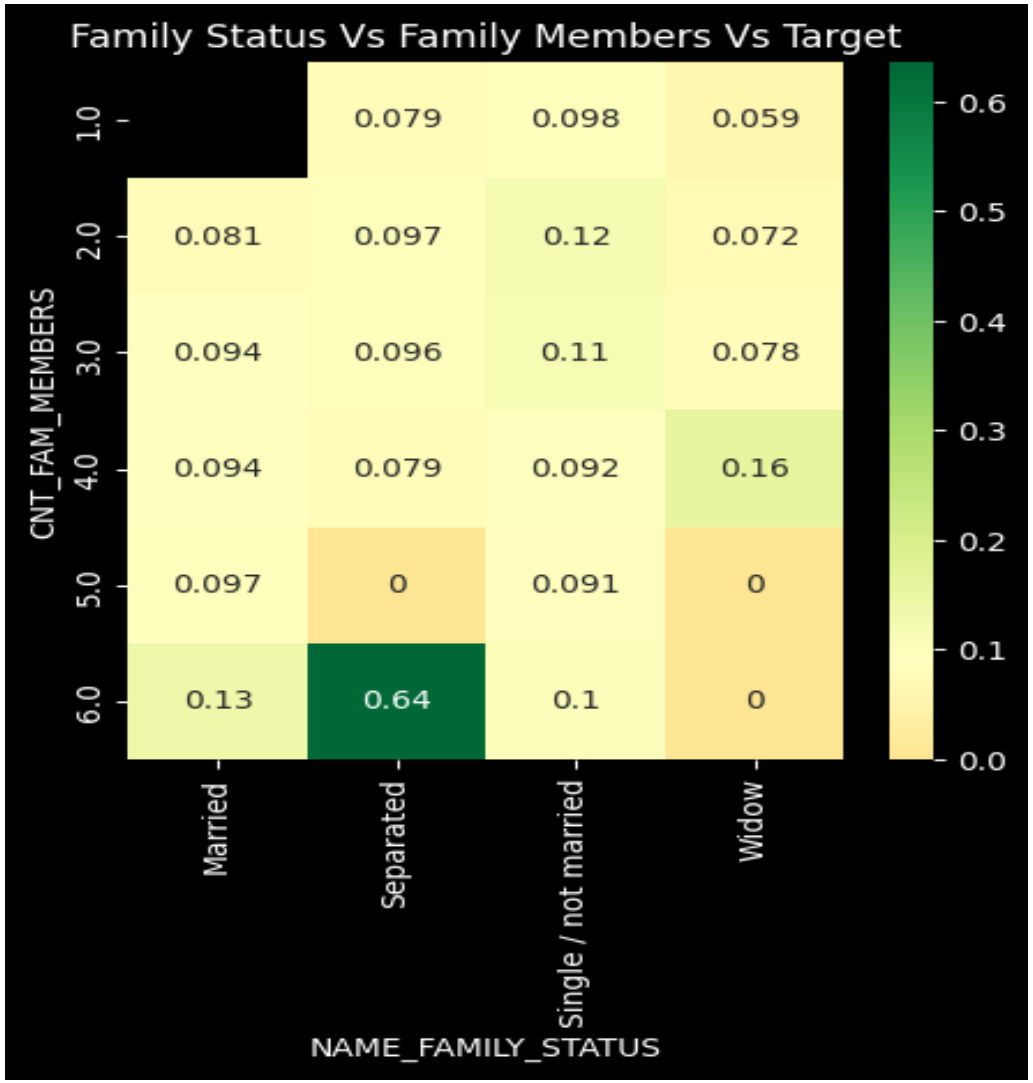
Education Vs Occupation Vs Target



Insights:

- In Academic degrees the majority of defaults are from Accountants, with a few from sales staff, this may be due to less number of applicants with academic degrees and the majority belonging to accountants.
- Occupations like drivers, cleaning staff, cooking staff, low-skill laborers, Realty agents, security staff or waiters/barmen staff are likely to default when they have an education level equal to or less than incomplete higher.
- This can show educational background and the type of job they are doing, have a direct impact on whether the applicant will default.

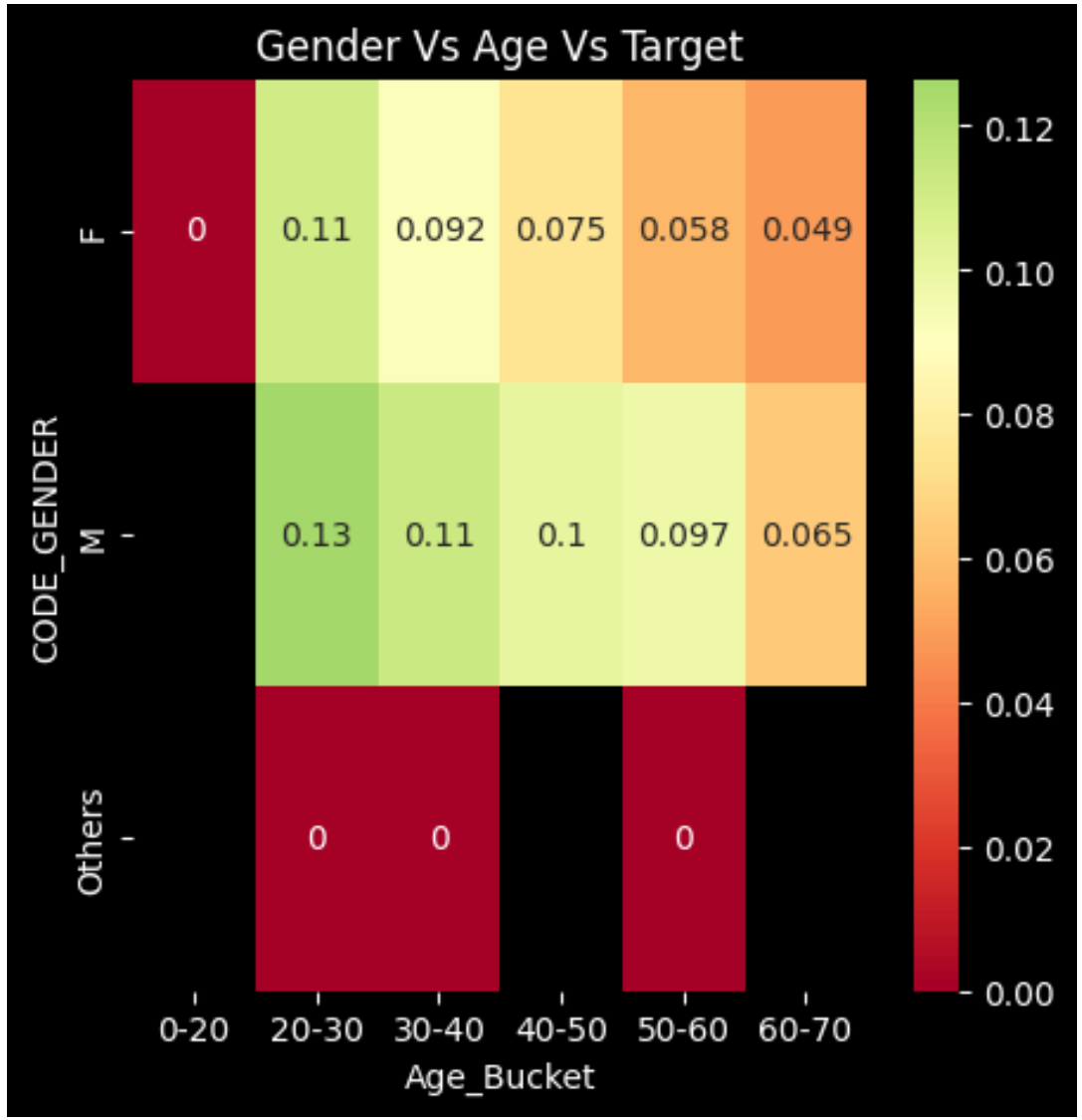
Family Status Vs Family Members Vs Target



Insights:

- There were a very small number of families with 6 or more members, out of which the majority defaults were from separated couples.
- Even for widows, the maximum defaults were with 4 family members.
- This might show that the bigger the size of the family and the family status of the applicants might indicate whether it will default.
- Specially, separated couples with big families are likely to default.

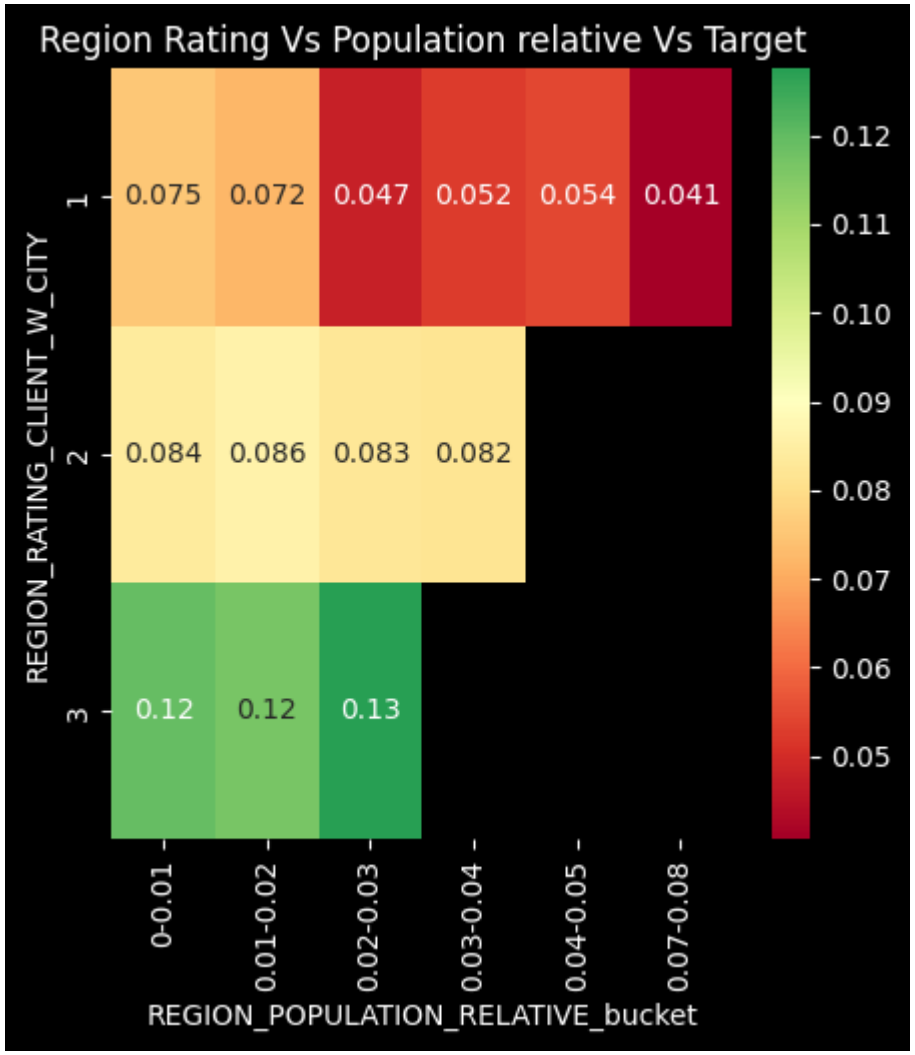
Gender Vs Age Vs Target



Insights:

- The majority of defaults are from younger age groups ie 20-30 and 30-40, which keeps on decreasing for older applicants.
- As we discussed earlier the majority of the defaults are from females, but we can see that the chances of a male applicant turning out a defaulter is more than female, even though it decreases with age, but the chances are still high.

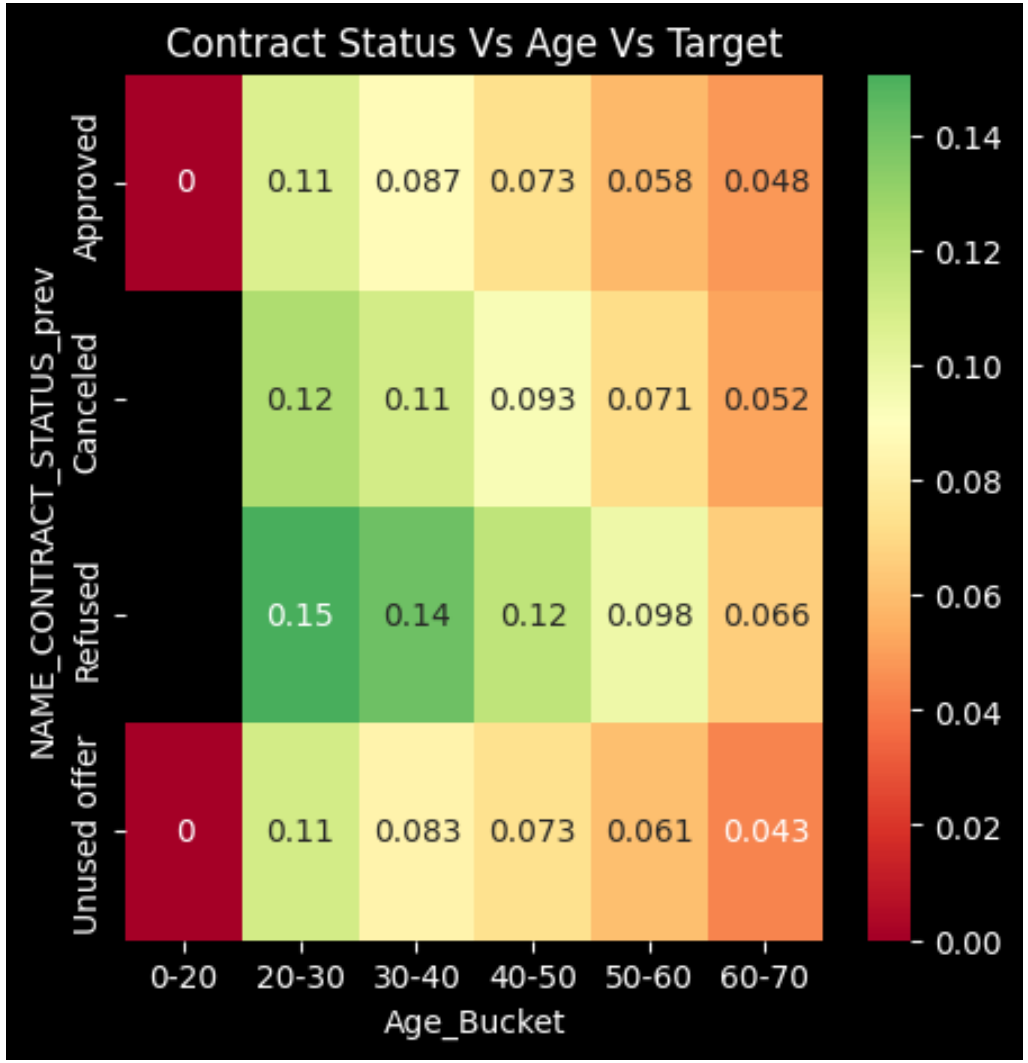
Region Rating Vs Population Relative Vs Target



Insights:

- We can see that the applicants coming from tier 3 cities and especially from regions with low populations are most likely to default.
- On the other hand the applicants coming from tier 1 cities are very less likely to default, which decreases more with the population increases.

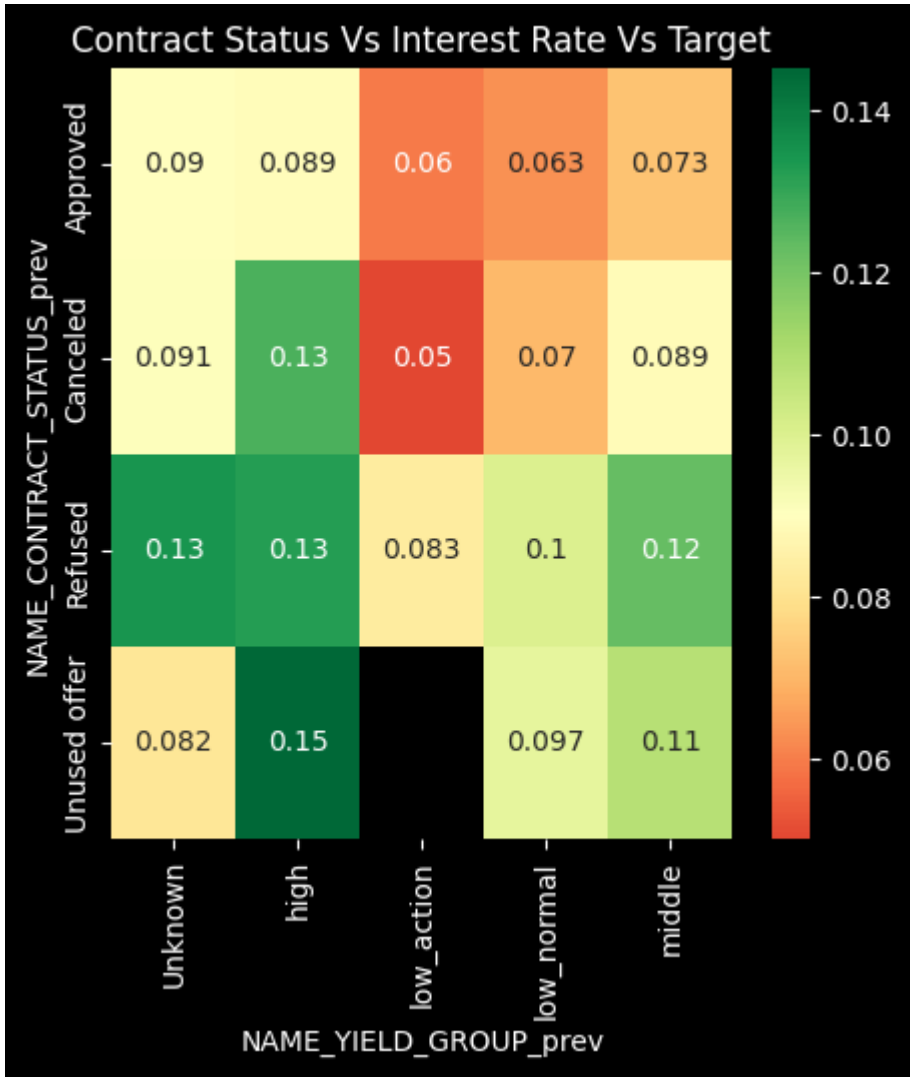
Contract Status Vs Age Vs Target



Insights:

- We can see that the majority of defaults belong to the 20-30 age group, which decreases as the age increases.
- Also the majority of defaults occur from applicants who were previously refused the loan, the defaults in the refused category only significantly decrease after the age of 50.
- Therefore the bank should carefully look at the previous application status of the applicants, with special importance to applicants coming with a previously refused applicant.

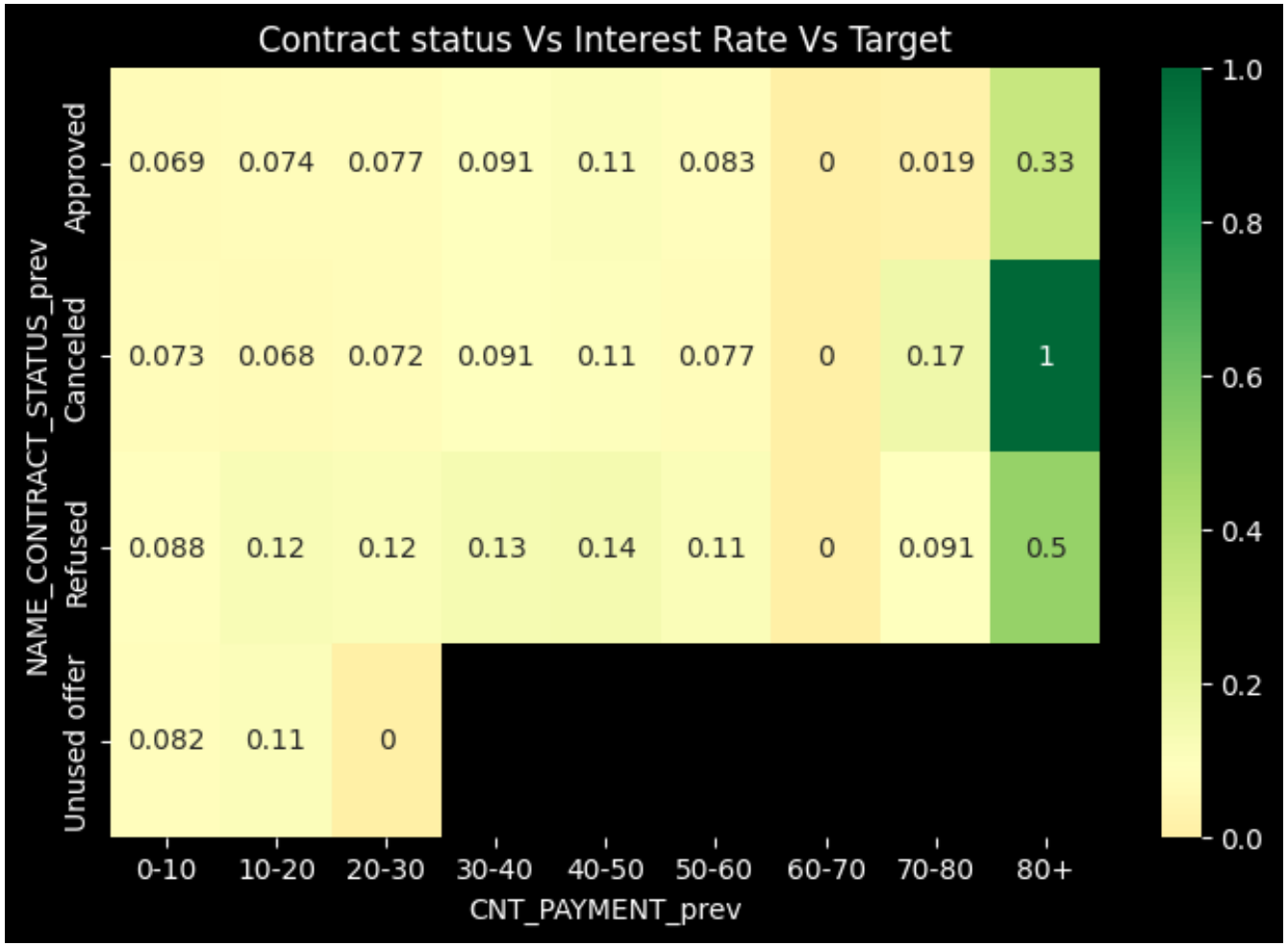
Contract Status Vs Interest Rate Vs Target



Insights:

- The majority of the defaults are from high interest rate followed by middle interest rate on previous applications, and the least defaults are from low_action interest rates.
- The people with previously refused offers default a lot along with people with unused offers.
- Special attention should be given to applicants coming from refused offers, especially from high or middle interest rates, and applicants with refused offers with unknown interest rates in previous applications.

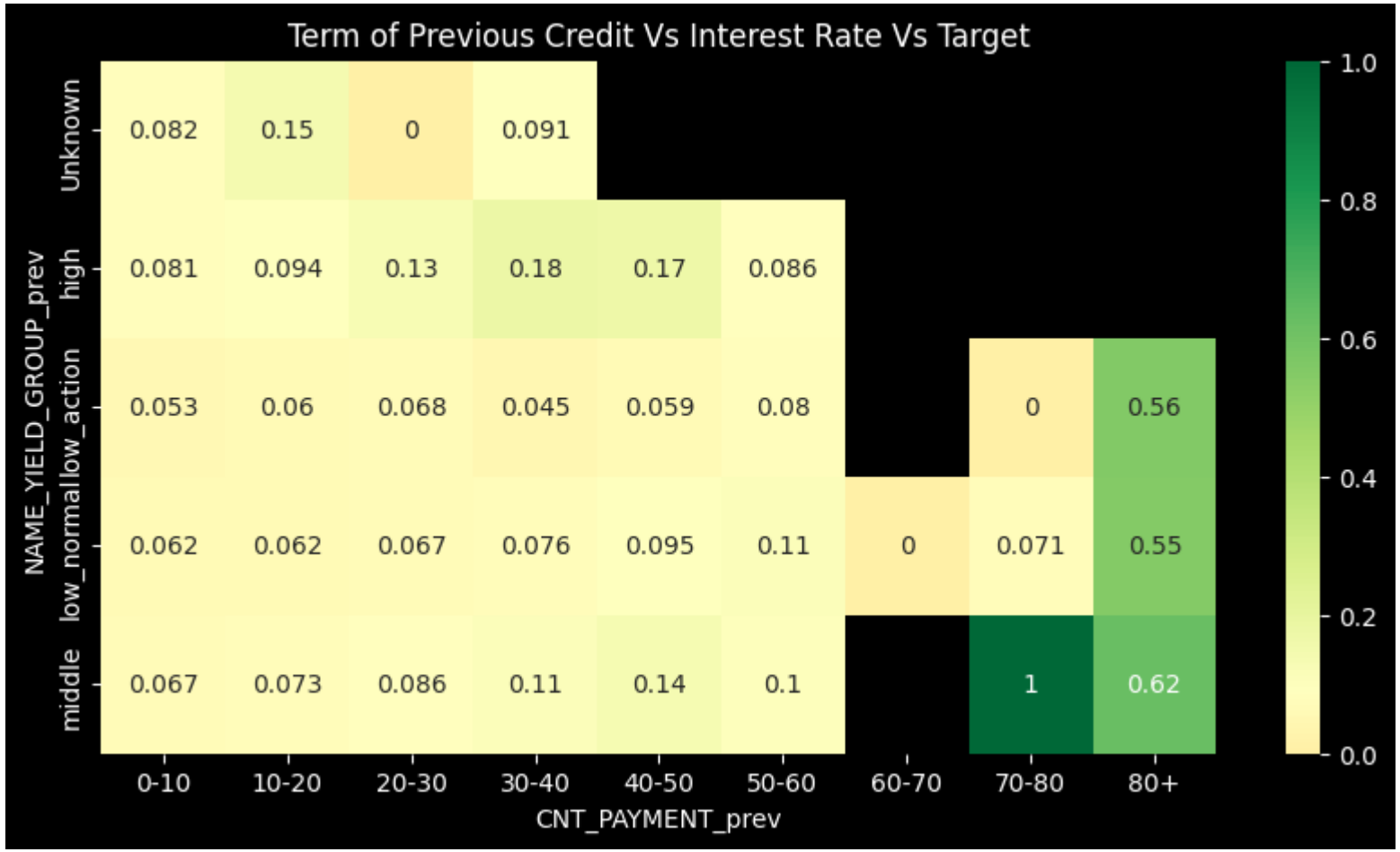
Contract status Vs Interest Rate Vs Target



Insights:

- The applicants with more terms of credit in previous applications are more likely to default.

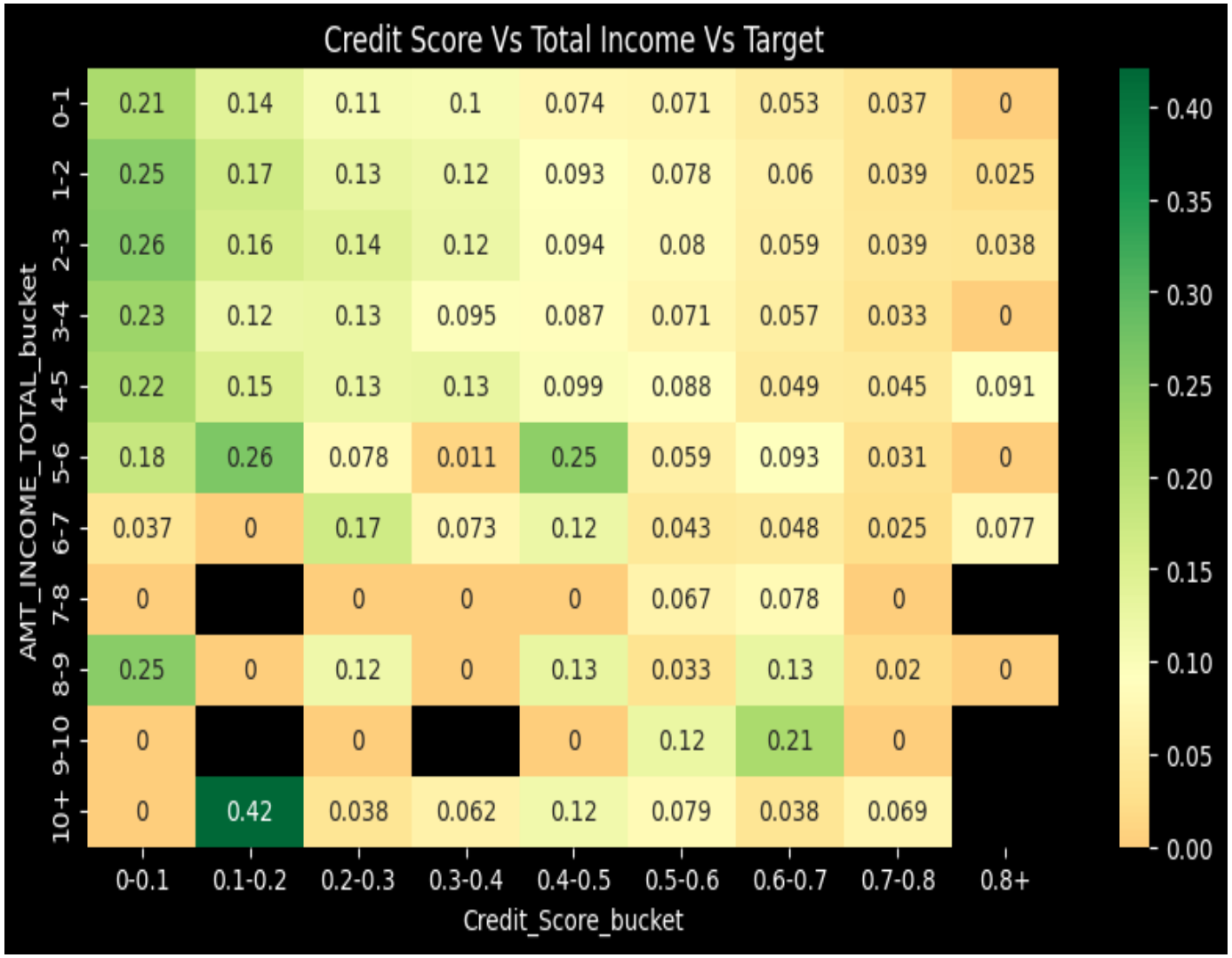
Term of Previous Credit Vs Interest Rate Vs Target



Insights:

This correlation graph also shows that the applicants with higher terms of previous loans are more likely to default.

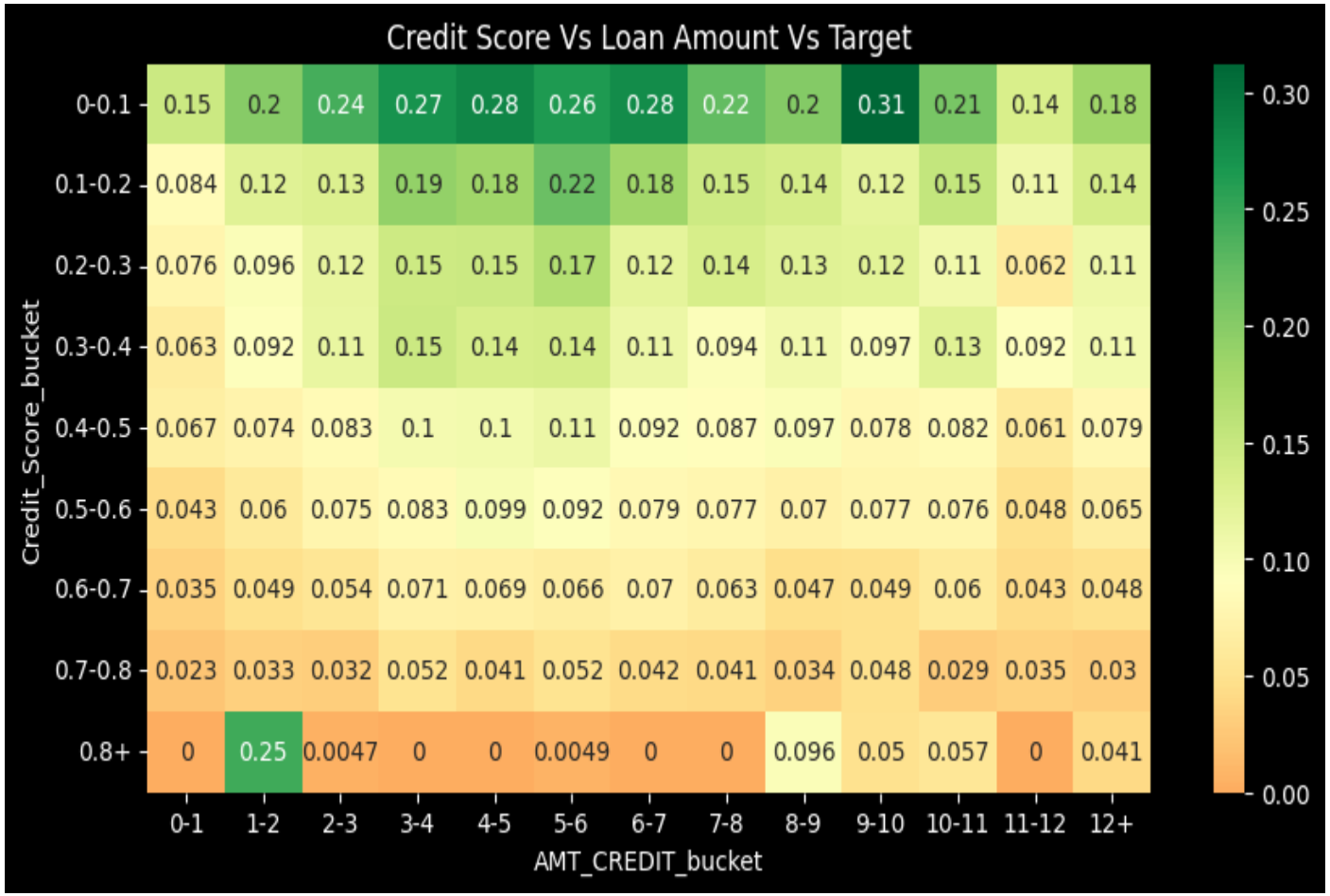
Credit Score Vs Total Income Vs Target



Insights:

- The majority of the defaults are made by people with low credit scores and low income. This means special attention should be given to people with low credit scores and people with low incomes.
- Credit score is a good indicator of whether an applicant with default or not because the applicant with a high salary is also likely to default.

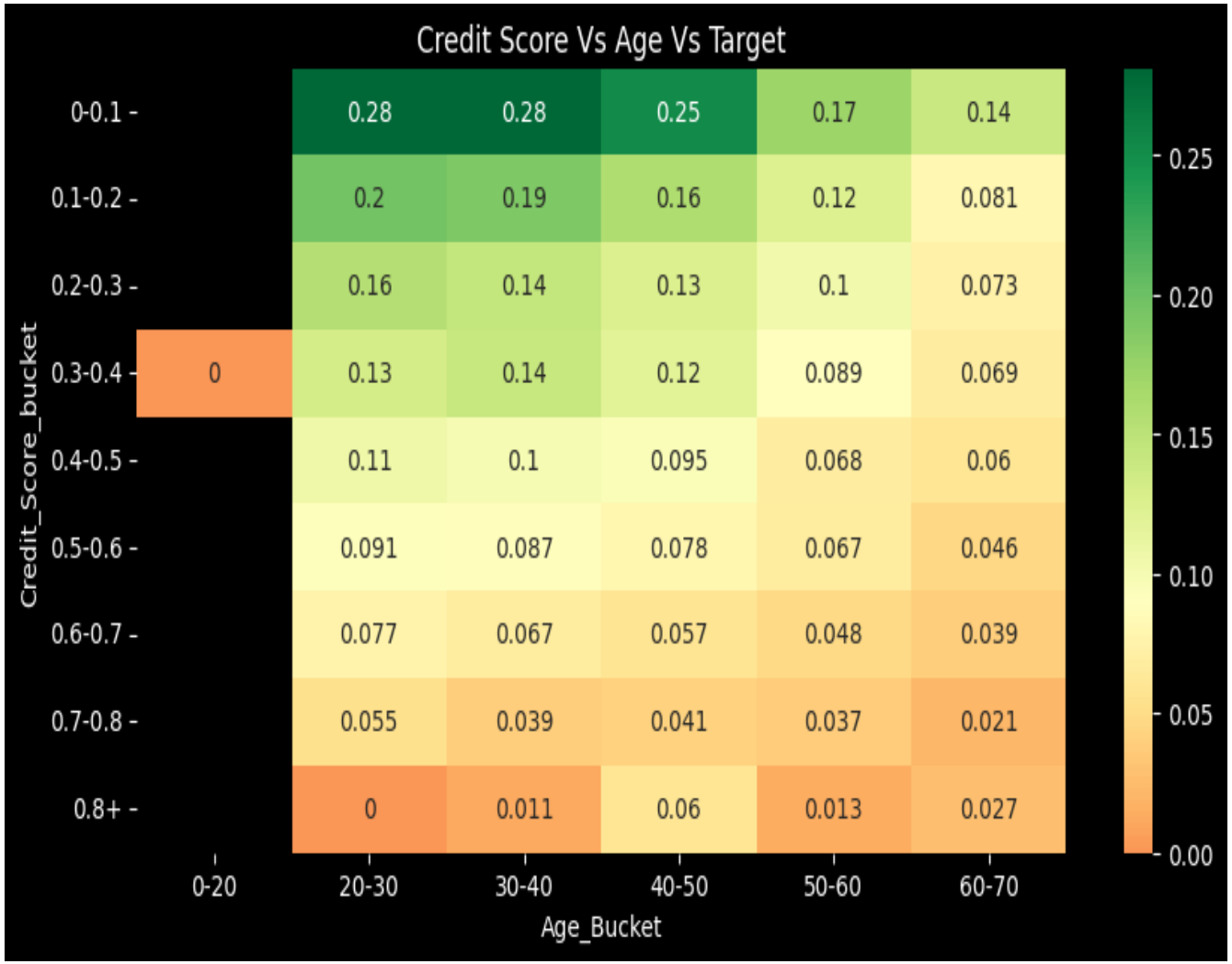
Credit Score Vs Loan Amount Vs Target



Insights:

- The applicants with low credit scores are most likely to default even with low loan amounts. The chances of an increase with an increase in loan amount and a decrease in credit score.
- Majority of the defaults are done for loan amounts of 3-6 lakhs. Also, people with high credit scores and loan amounts of 1-2 lakhs are more likely to default.
- There the bank gives attention to applicants with low credit scores and loan amounts of 2-6 lakhs as they are more likely to default.

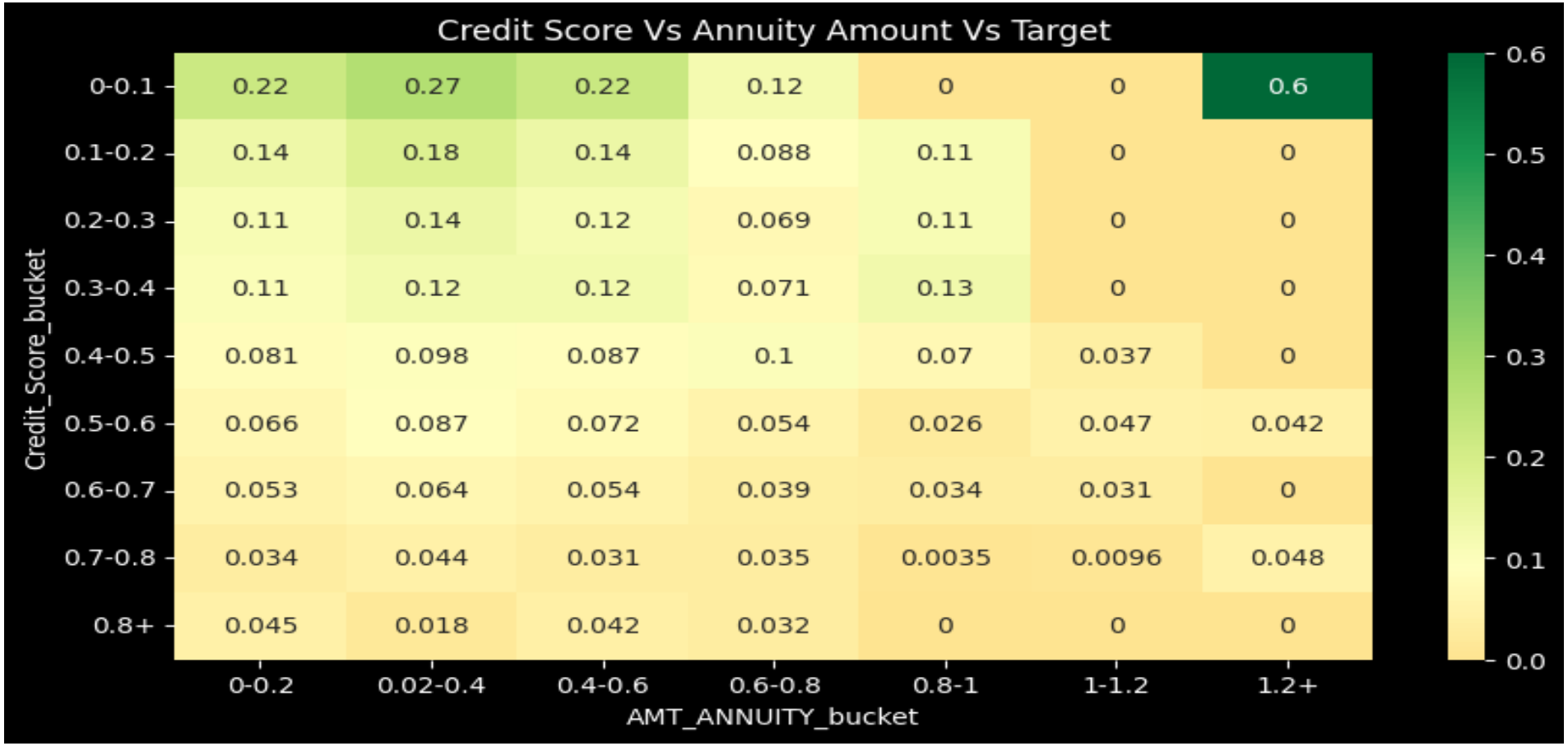
Credit Score Vs Age Vs Target



Insights:

- People with low credit scores are likely to default for any age group.
- People aged between 20-50 are likely to default when their credit score is low.

Credit Score Vs Annuity Amount Vs Target



Insights:
People with low credit income and high annuity amounts are to default the most.

FINAL ANALYSIS:

Final Results

1. Age: Age of the applicant defines a lot whether the applicant will default or not, applicant with lower age specially 20-40 years are to default a lot as compared to older applicant.

2. Occupation: The applicants who are at high skill jobs are less likely to default than applicants coming from low skill jobs.

3. Education: The applicants who are highly educated are less likely to default that applicants from low education level.

4. Gender: Even if the number of defaulter are more for female, the chances of a male applicant turn out to be defaulter is more.

5. Credit Score: People with low credit scores are more likely to default and the people with high credit score are very less likely to default.

6. Previous Application Status: Those applicants who had been refused the previous loan are more likely to default in the next.

7. Interest Rate of Previous Applications: The applicants who had high, unknown, or middle interest rates are more likely to default than lower interest rates.

8. Region Rating and Population Relative: The applicants living in tier 3 cities are more likely to default than applicants living in tier 1 cities. Also, the applicants living in areas with lower-population relatives are more likely to default.

9. Count Of family members and Children: The applicants with a high number of family members and children are likely to default, this may be due to the number of people dependent on the applicant.

10. Family Status: Applicants who are single or separated are more likely to default than married and single.

11. Type of Income: Applicants who are unemployed or on maternity leave are more likely to default.

12. Living Status: People living with parents or in rented apartments have a high chance of getting defaulted.