upGrad

CREDIT EDA ASSIGNMENT

upGrad & IIITB | Data Science Program - October 2023
Batch ID: 5673

Name: G. Arun Kumar

arunreturns04@gmail.com

Introduction

We have gained foundational insights into risk analytics within the banking and financial services sector, comprehending the utilization of data to mitigate the potential loss of funds when extending loans to customers.

Business Understanding

- Loan-providing companies encounter difficulty when granting loans to individuals with limited or non-existent credit histories.
- This challenge creates an opportunity for certain consumers to strategically default on loans for their advantage.
- The main task involves utilizing Exploratory Data Analysis (EDA) to examine patterns within the data.
- The ultimate goal is to employ EDA as a means of ensuring that loan applicants with the capability to repay are not unfairly rejected.
- Upon receiving a loan application, the company is tasked with making a decision on loan approval based on the applicant's profile.

Two distinct risks are associated with the bank's decision-making process:

i. If the applicant is deemed likely to repay the loan, rejecting the loan application results in a business loss for the company.

ii. Conversely, if the applicant is assessed as not likely to repay the loan, indicating a potential default, approving the loan could lead to a financial loss for the company.

Business Objectives

The provided dataset encompasses information regarding loan applications at the time of applying for the loan, consisting of two distinct scenarios.

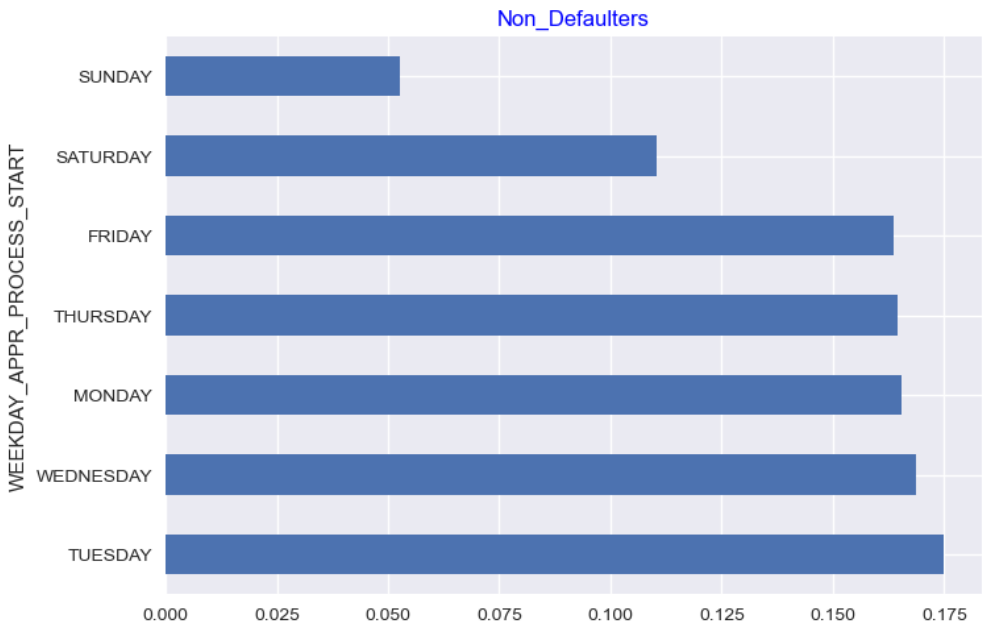
1) Scenario A involves clients facing payment difficulties, specifically those who experienced late payments exceeding X days on at least one of the initial Y instalments in our sample.

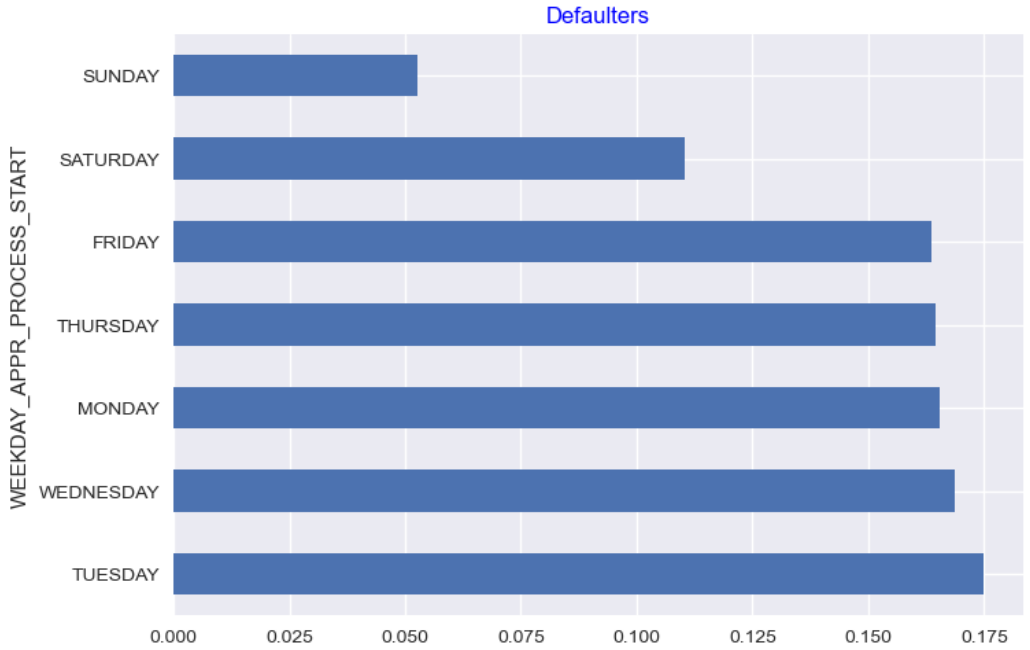
2) Scenario B encompasses all other cases where payments are made punctually, excluding instances falling under the criteria specified in Scenario A.

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, 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 will ensure that the consumers capable of repaying the loan are not rejected.
- Identification of such applicants using EDA is the aim of this case study.

Univariate Analysis on Application data

WEEKDAY_APPR_PROCESS_START

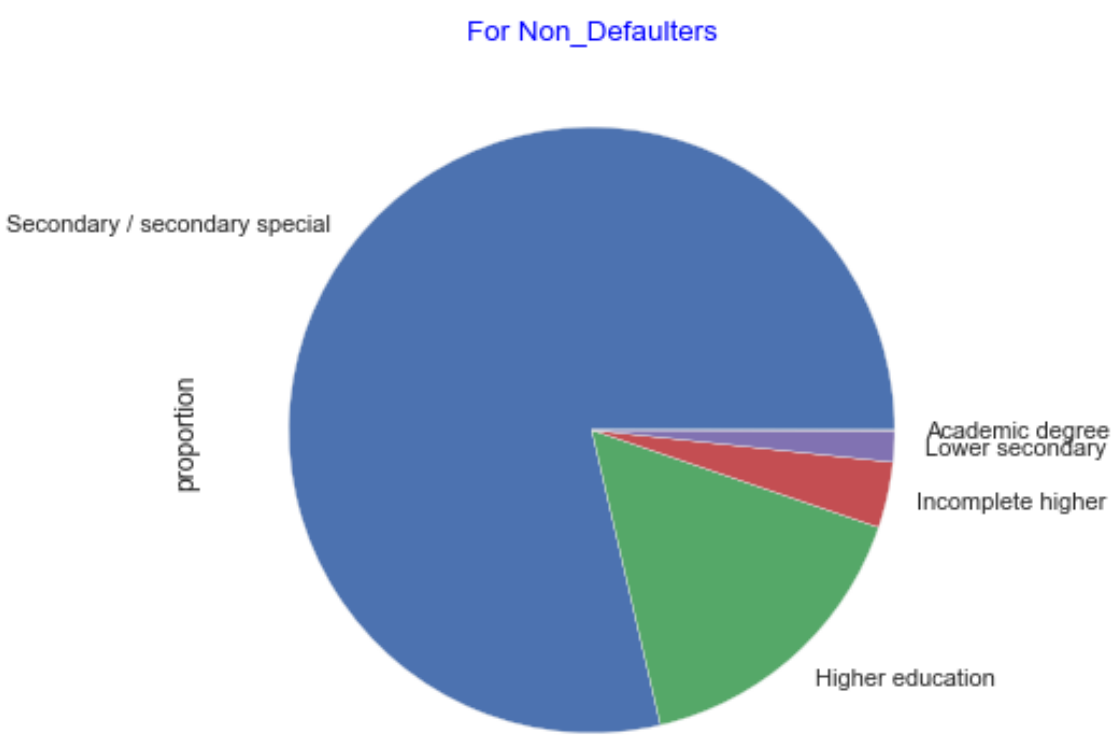


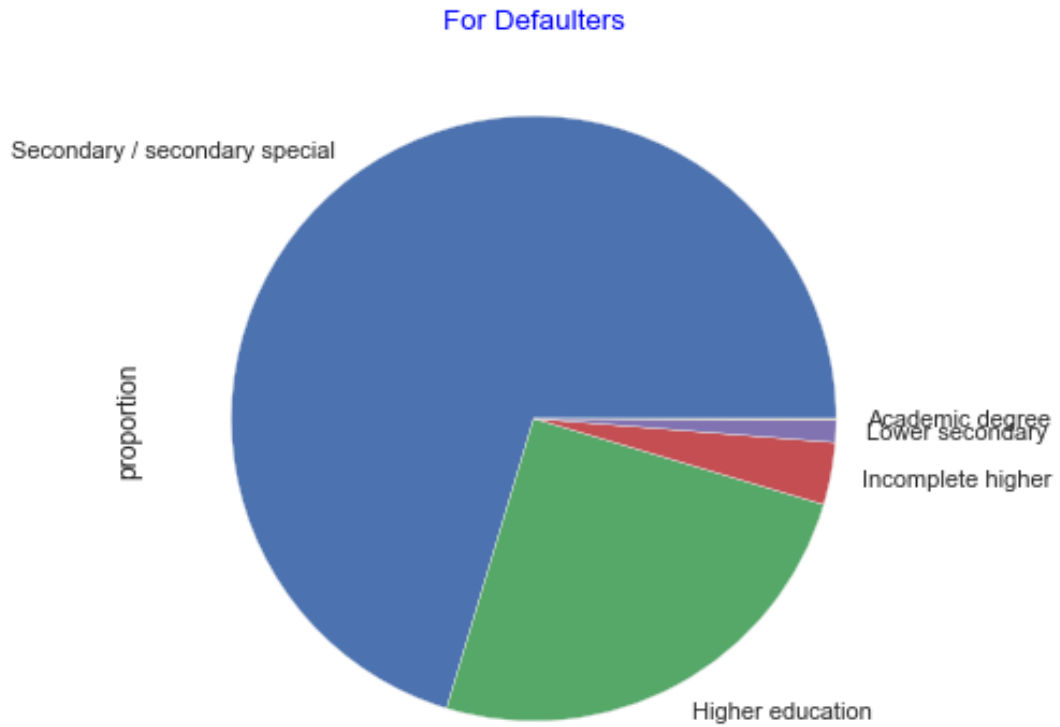


Inferences

- From the graph we can conclude that application starting processes are less on Saturday and Sunday for both Defaulters and Non-Defaulters

NAME_EDUCATION_TYPE

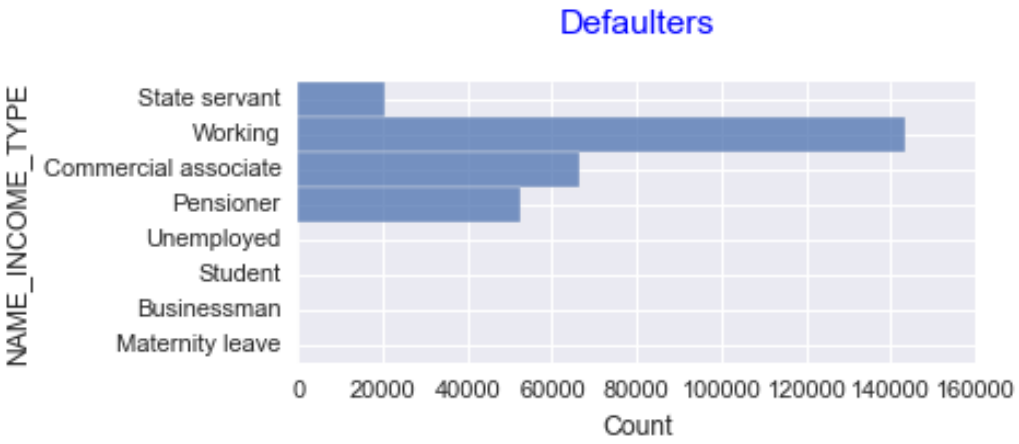


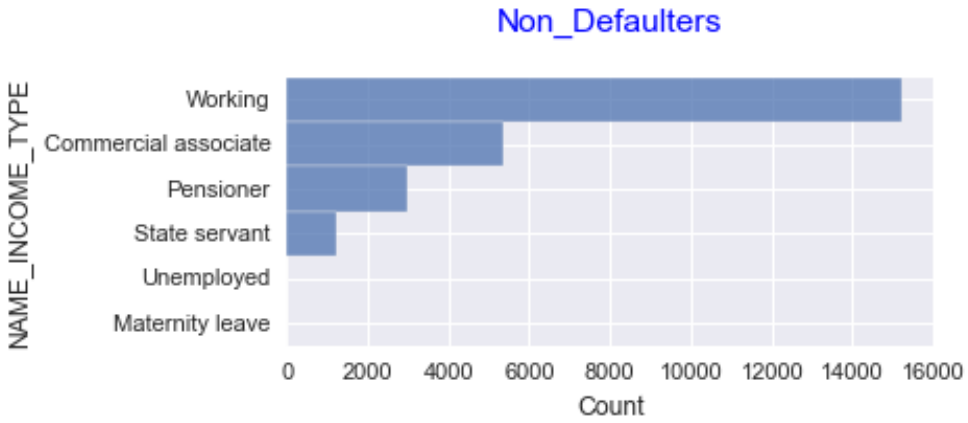


Inferences

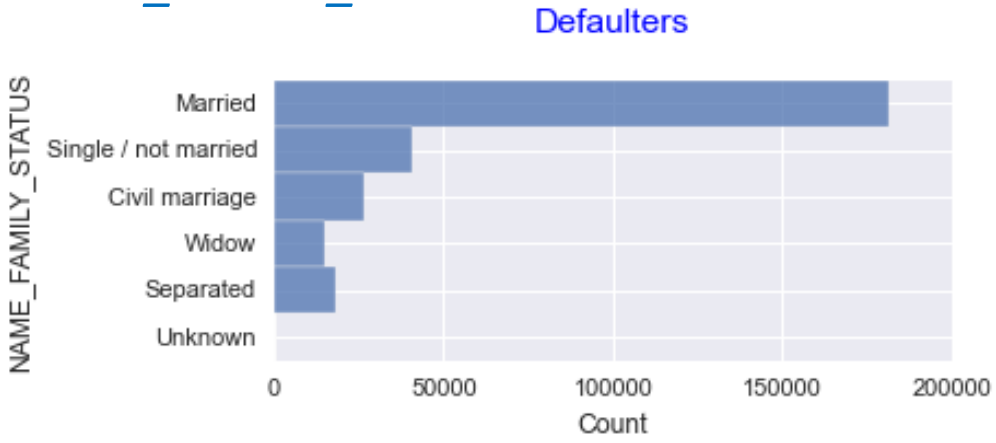
- Secondary/special educated people are applying loans in high number and people with Academic degree education are the least in both Defaulters and Non-Defaulters.

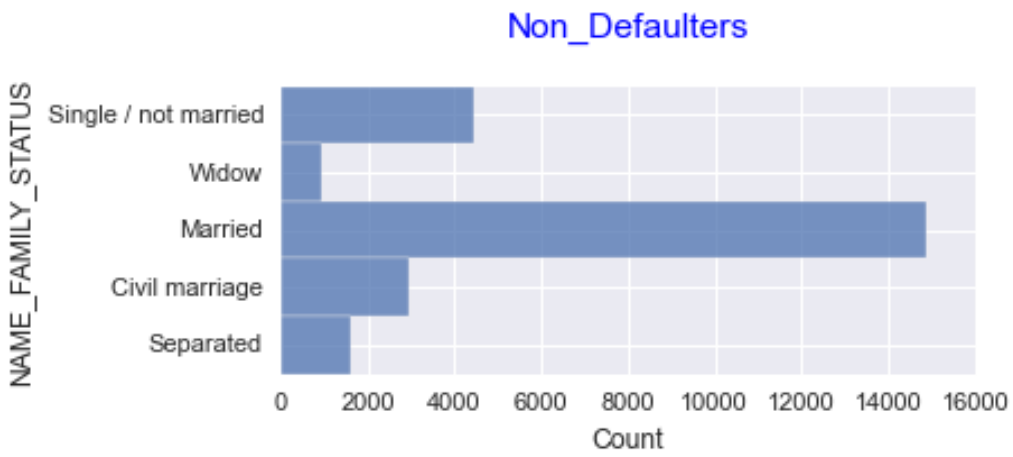
NAME_INCOME_TYPE





NAME_FAMILY_STATUS

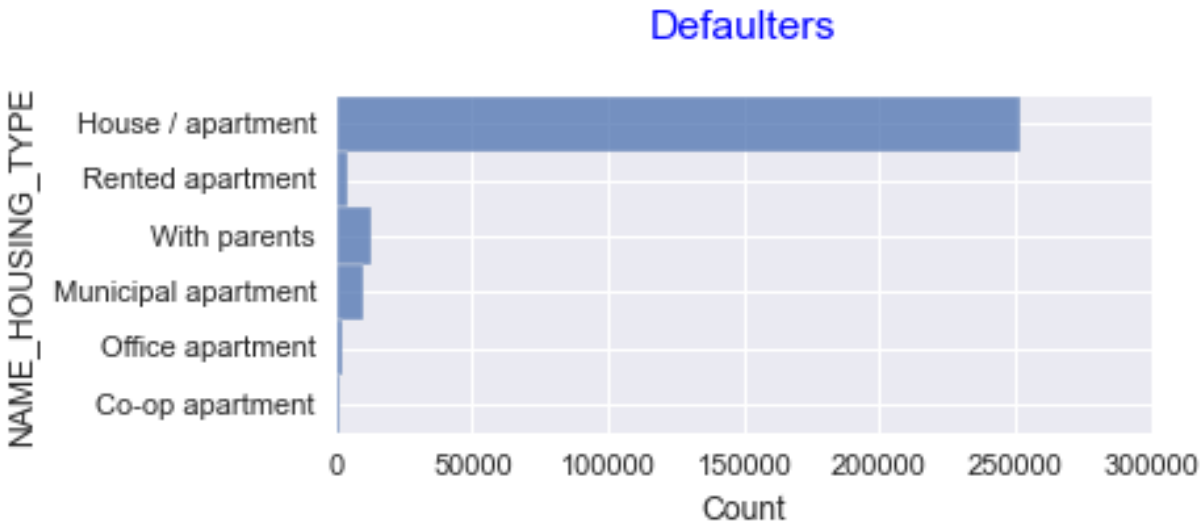


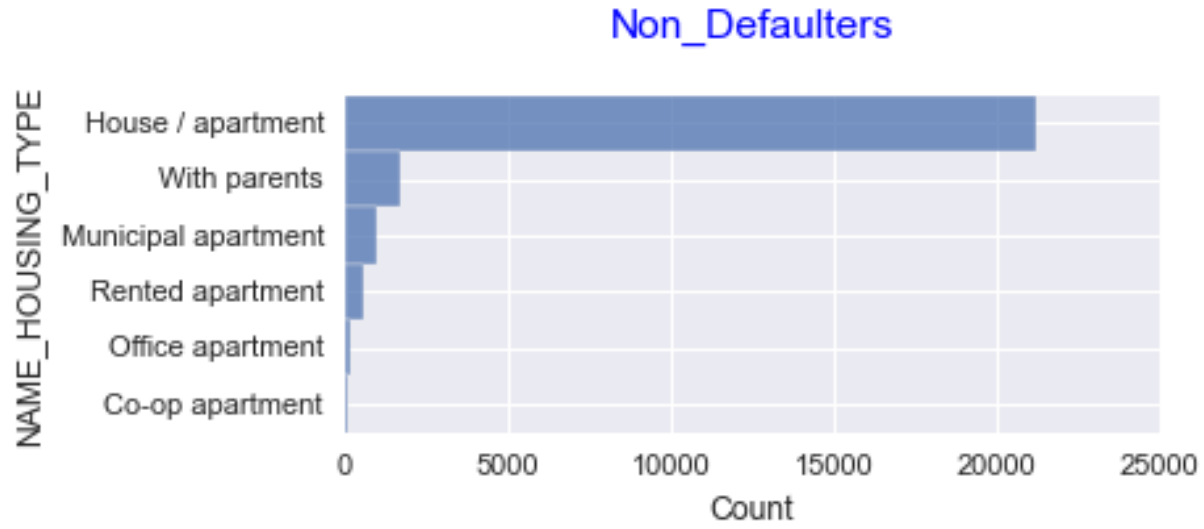


Inferences

- Majority of the Defaulters and Non-Defaulters, both come from working class
- Married people tend to take more loans as compared to other categories but it doesn't impact on whether they are Defaulters or Non-Defaulters

NAME_HOUSING_TYPE



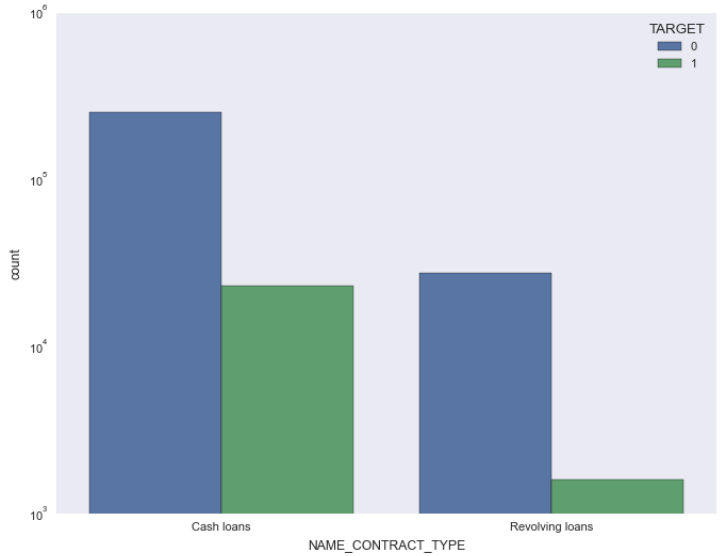


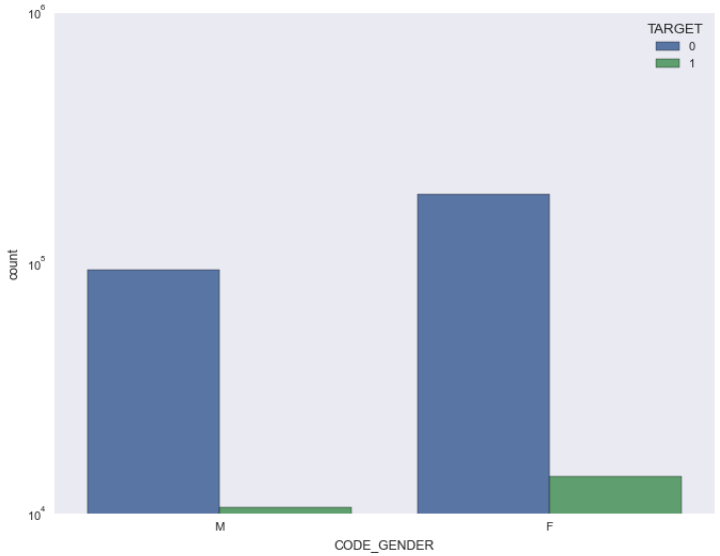
Inferences

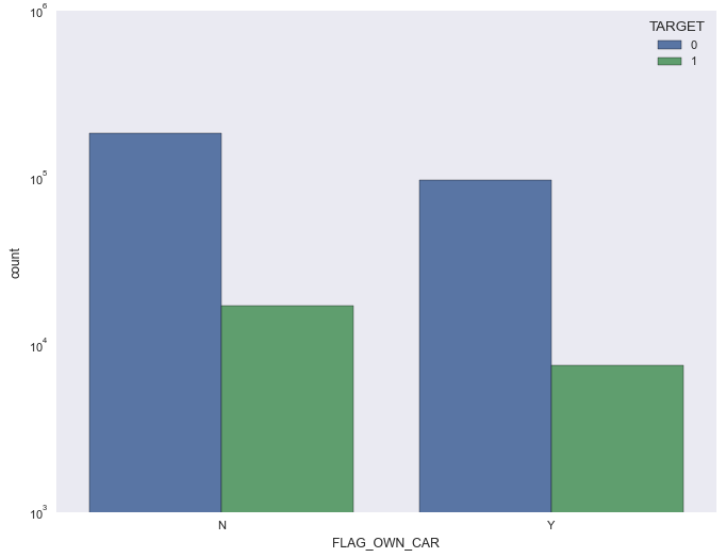
- People having their own house/apartment tend to take more loans and fall in both Defaulters and Non-Defaulters category

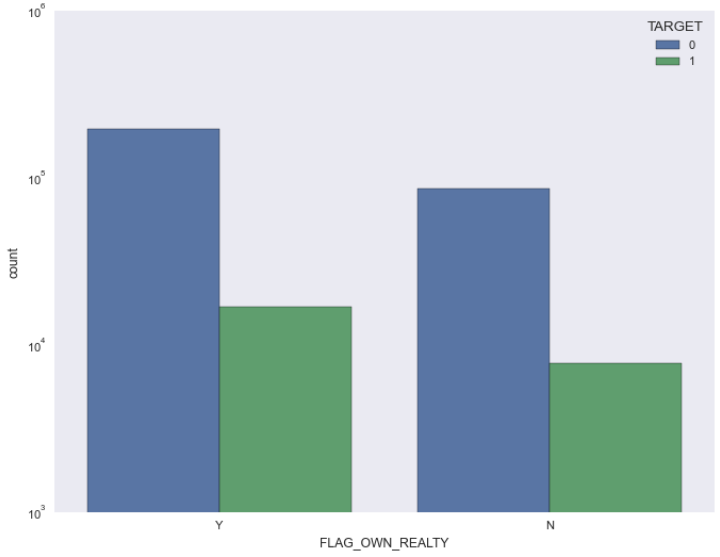
Categorical Columns

['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'AMT_CREDIT_RANGE', 'AMT_INCOME_RANGE']



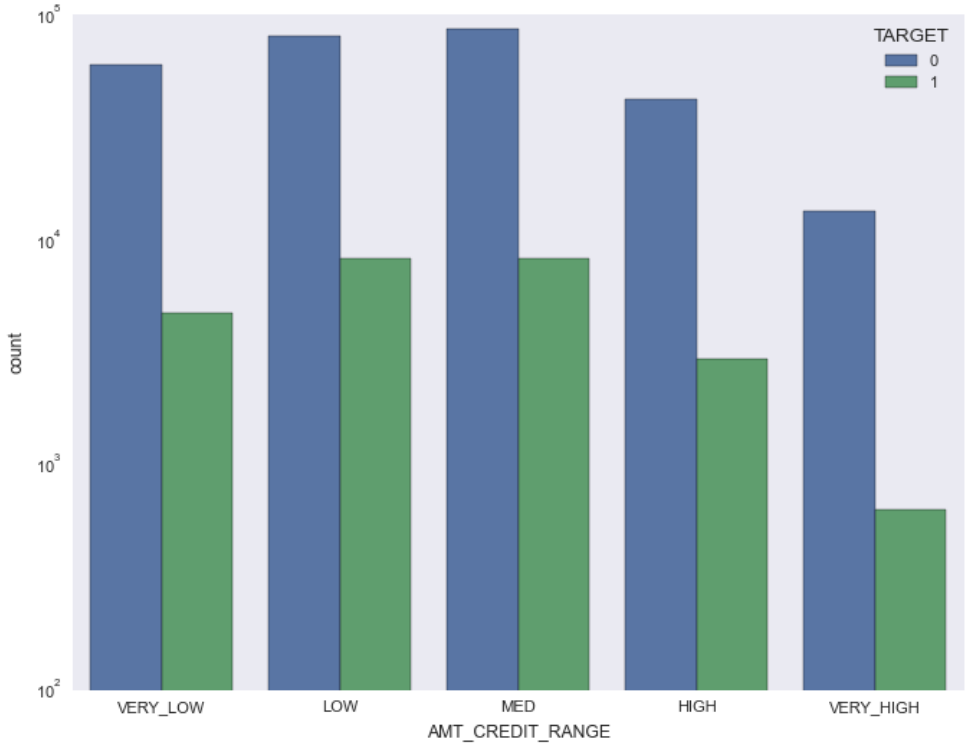


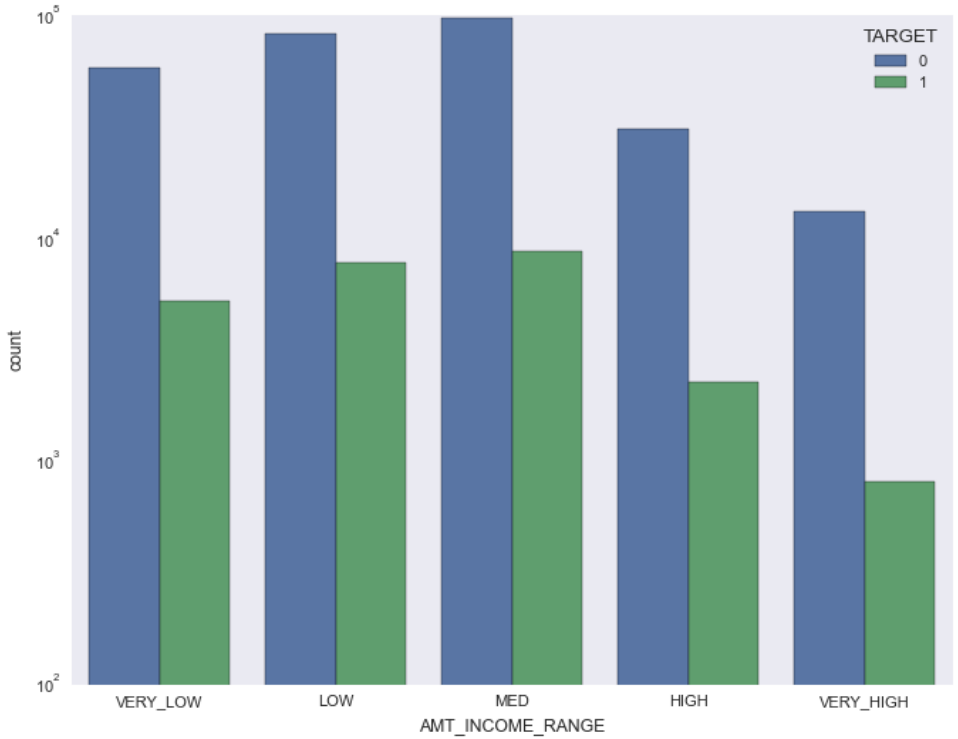




Inferences

- People tend to take more cash loans than revolving, and Defaulters' percentage of revolving loans are less
- Females opt for loans more than Males
- Real estate people tend to take more loans
- People who don't own a car, tend to take more loans



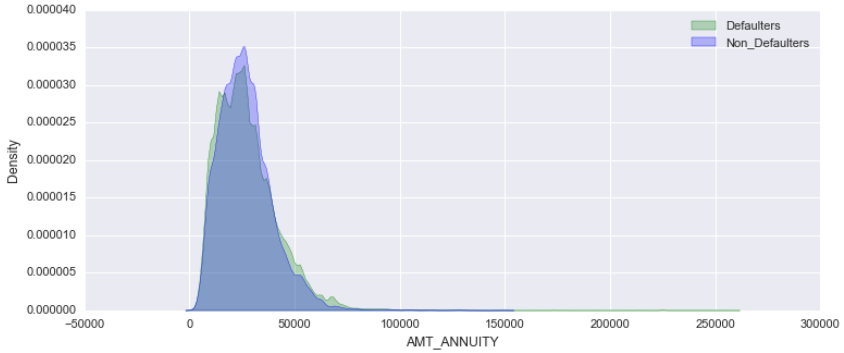


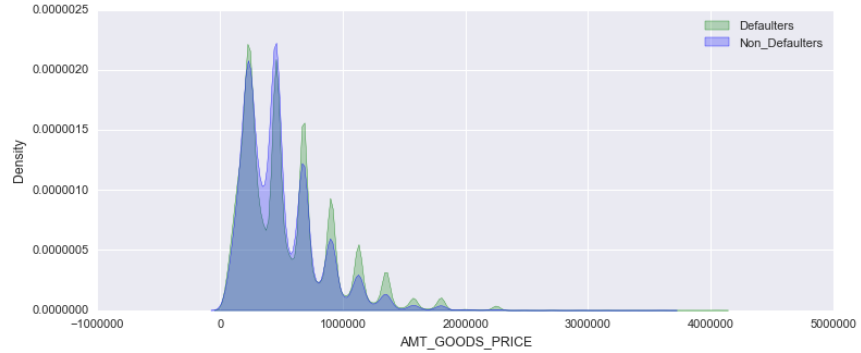
Inferences

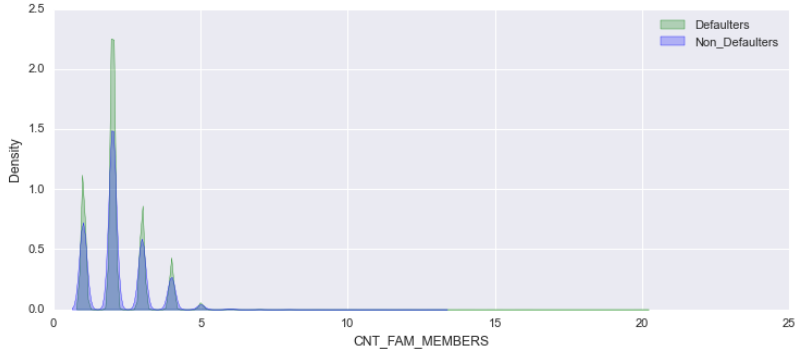
- People with Medium total income are more likely to default
- People with high Credit amount are less likely to default

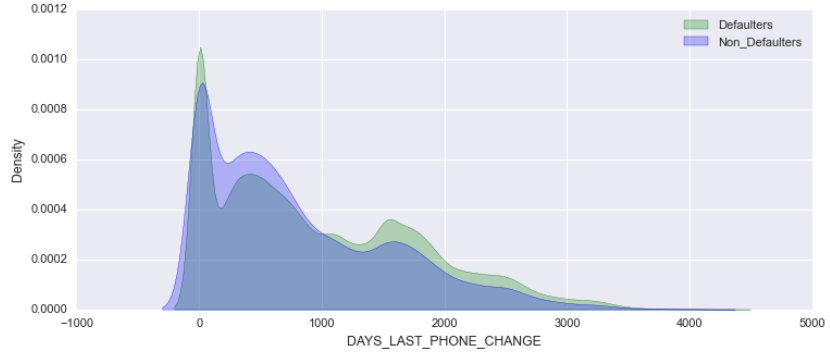
Continuous columns

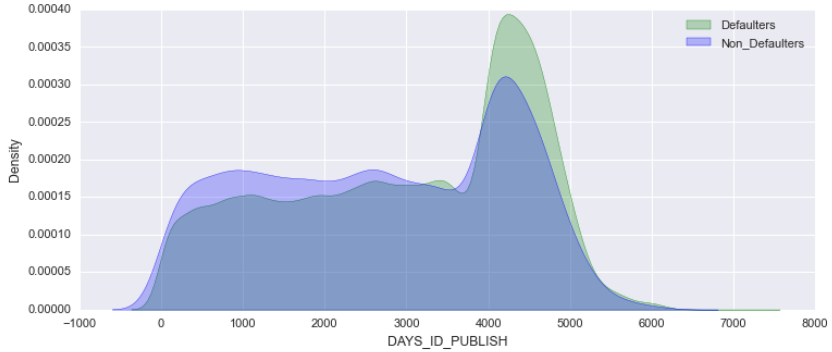
['AMT_ANNUITY','AMT_GOODS_PRICE','CNT_FAM_MEMBERS','DAYS_LAST_PHONE_CHANGE','DAYS_ID_PUBLISH','DAYS_BIRTH','HOUR_APPR_PROCESS_START','DAYS_EMPLOYED','AMT_CREDIT','AMT_INCOME_TOTAL']

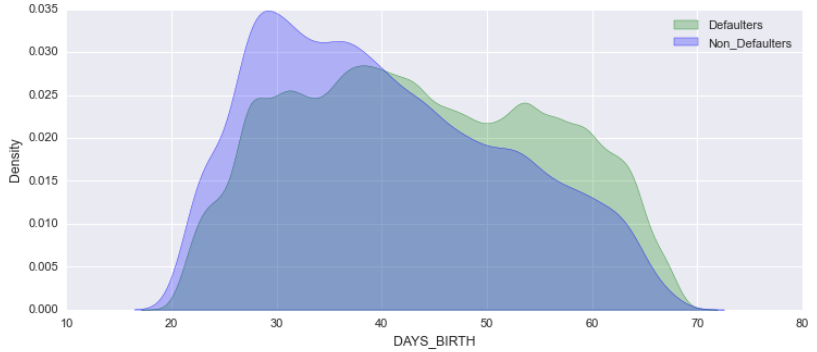






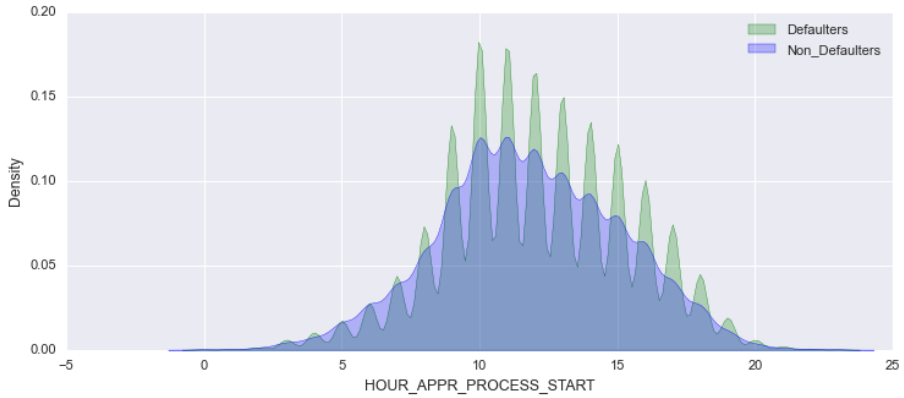


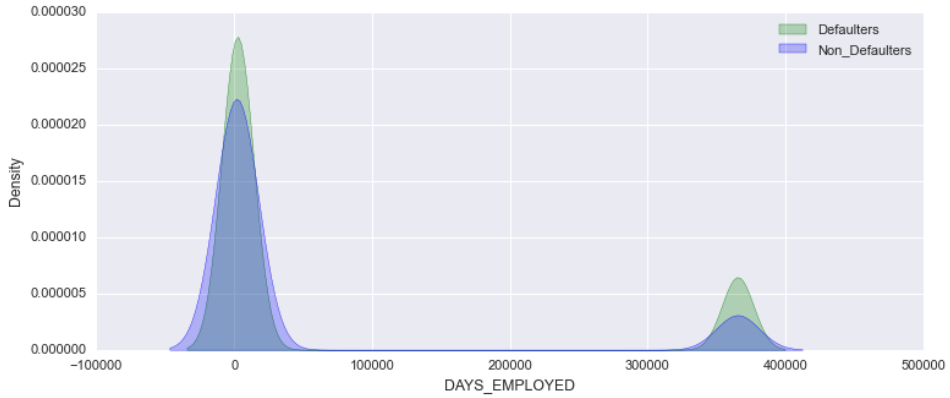


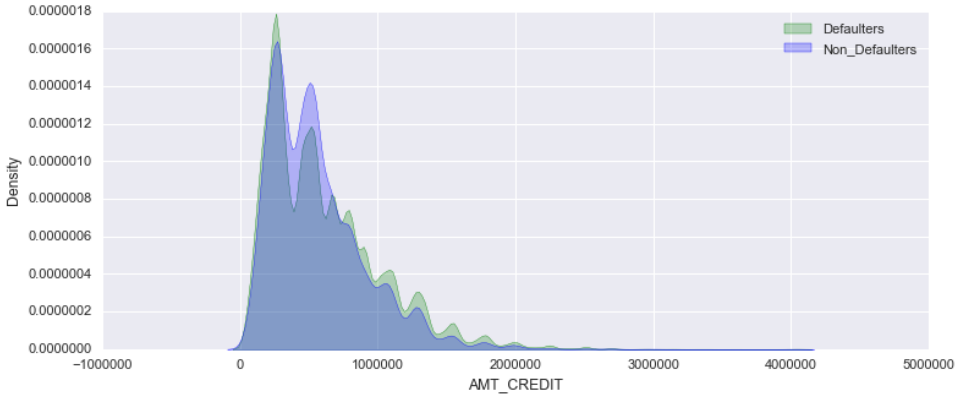


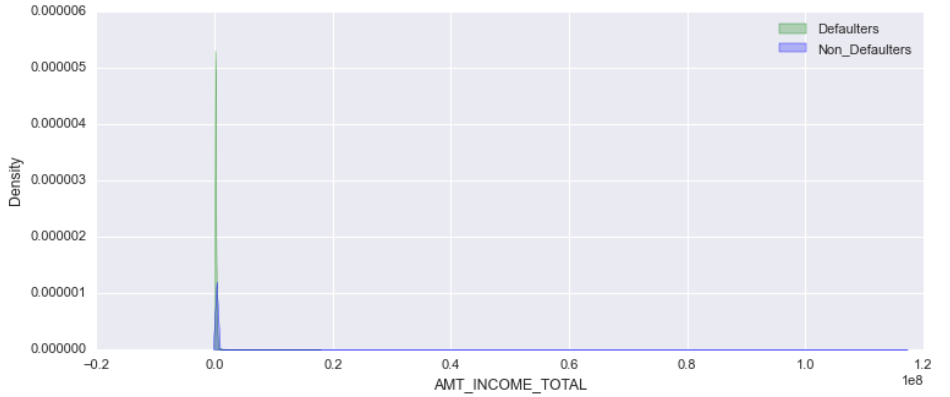
Inferences

- Individuals with lower annuity amounts tend to have a higher incidence of loans
- Loans are more commonly sought for smaller goods amounts
- Nuclear families tend to take more loans
- People whose IDs were published between 4,000 and 5,000 days ago tend to take more loans
- Individuals aged between 27 years (10,000 days) and 41 years (15,000 days) exhibit a higher likelihood of seeking loans
- Retired individuals(penisoners) show a tendency to take out loans







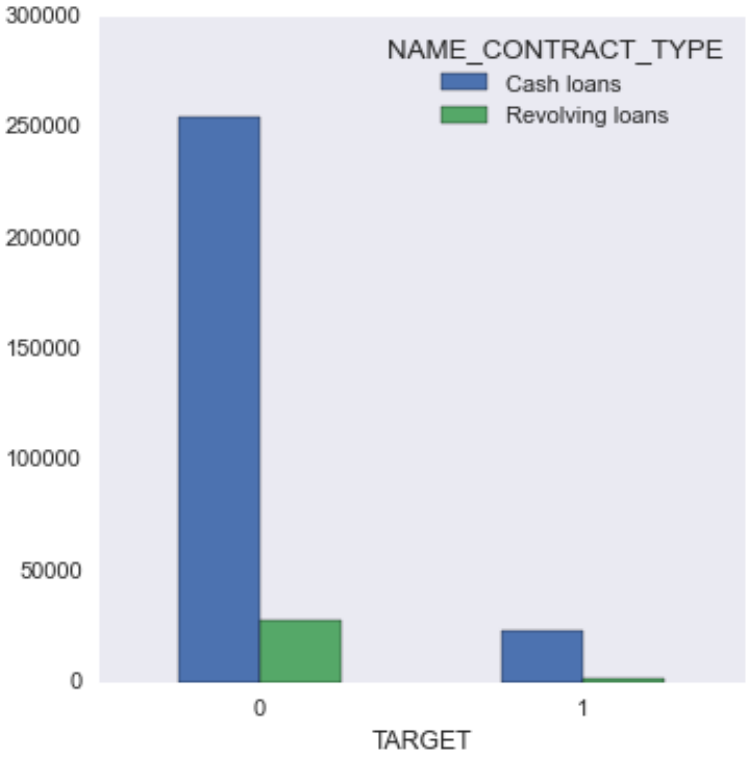


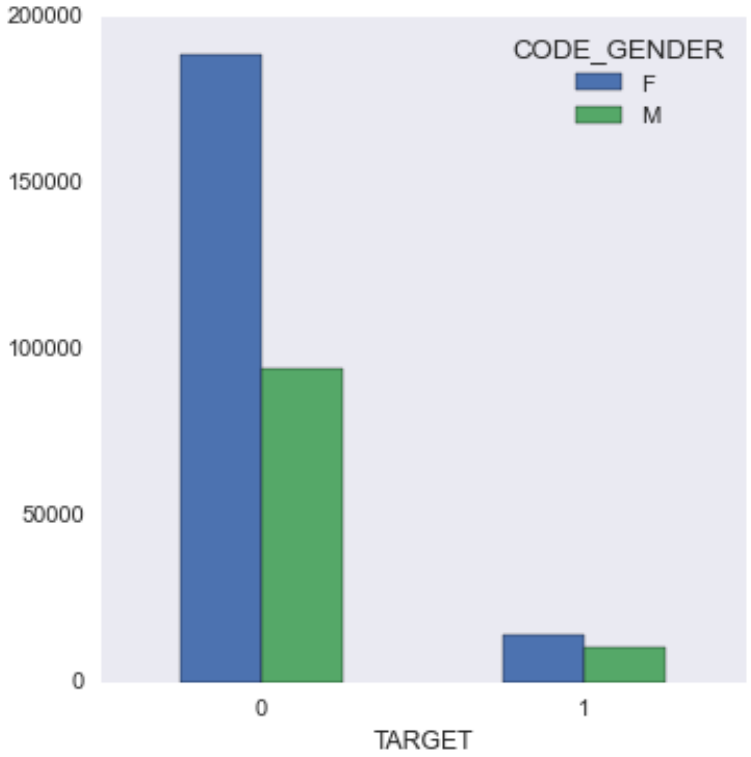
Inferences

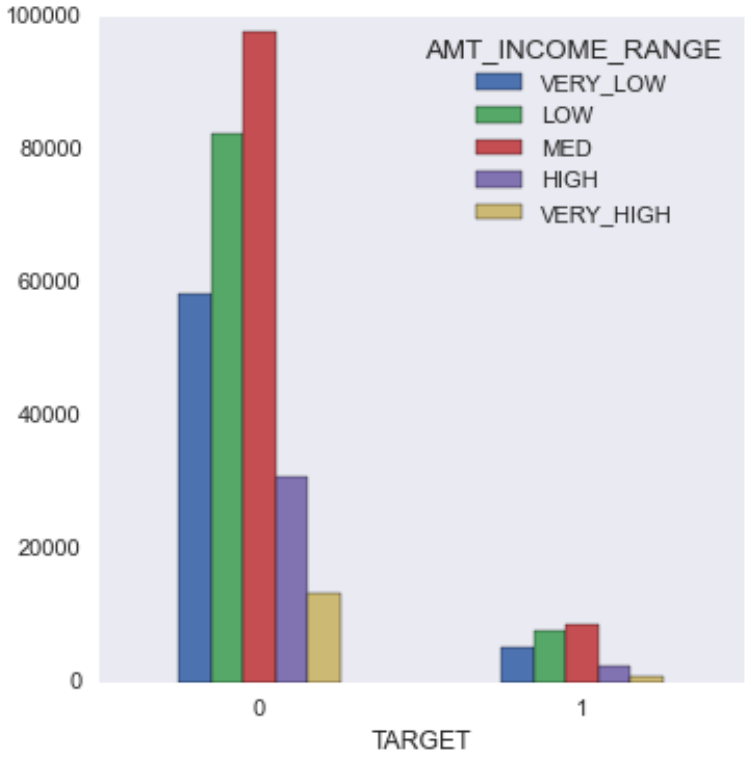
- A significant volume of loan applications is submitted during the time frame of 10 AM to 2 PM.
- Recently employed individuals are inclined to borrow more, possibly to address initial financial needs associated with starting a new job
- Loans are more commonly sought for smaller goods amounts
- Individuals with lower total income are more prone to default on loans.

Bivariate Analysis on Application data

Categorical Columns



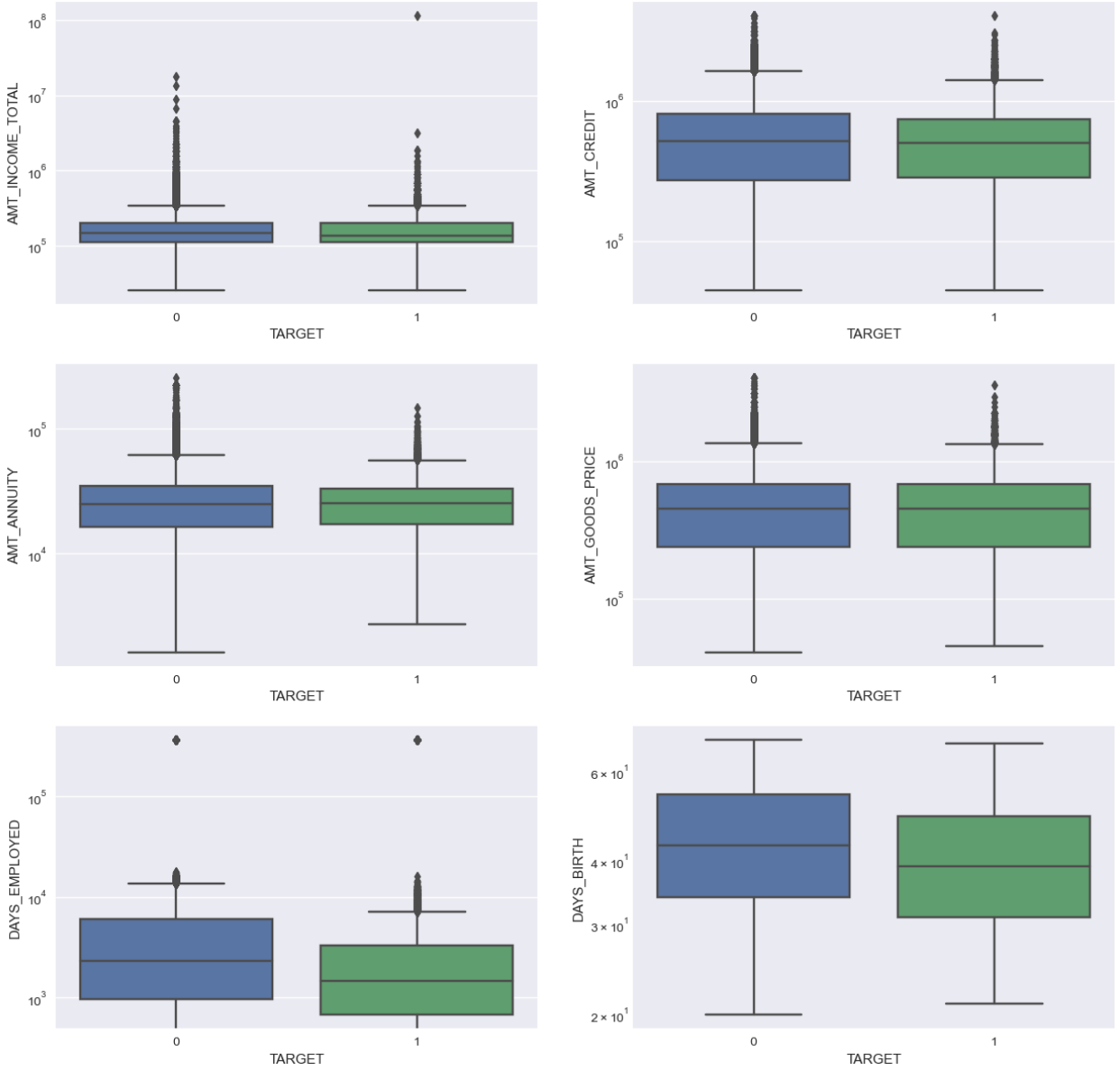




Inferences

- Higher credit amounts are generally cash loans
- Females generally tend to take more loans and men are more likely to default.
- Income segment >500000 (very-high) have less defaulters.

Continuous Columns



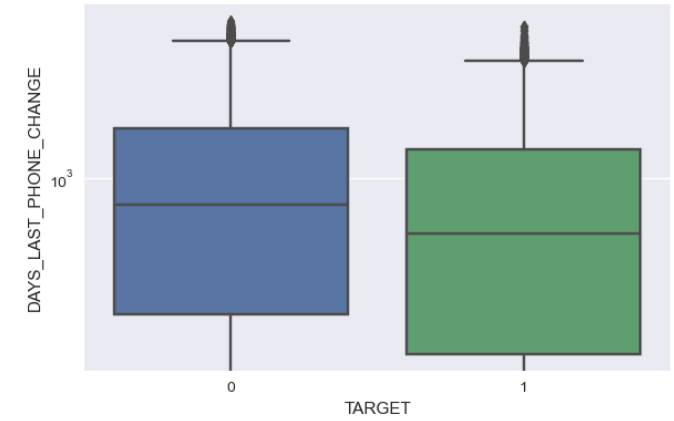
Insights

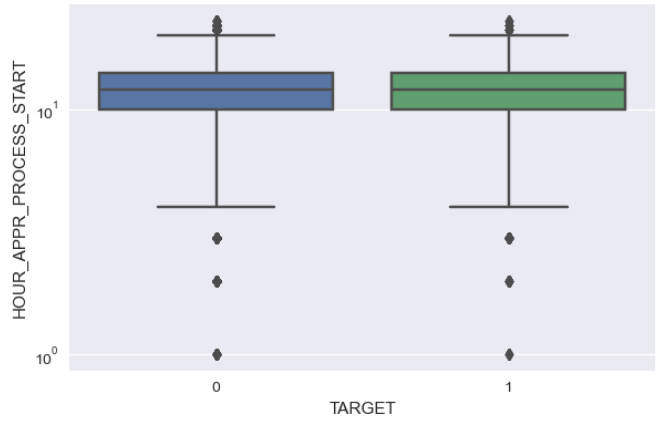
(Target-1 : Defaulters and Target-0: Non-Defaulters)

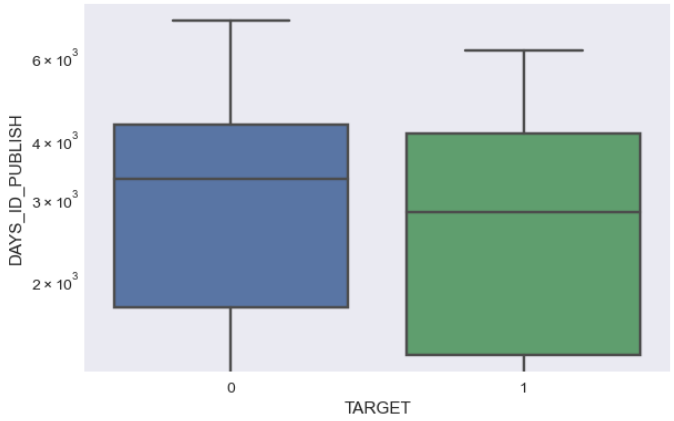
- The majority of individuals experiencing default situations have a lower total income.
- Clients with credit amounts exceeding 50000 are less likely to default compared to those with lower amounts, and vice versa.
- Individuals with a higher number of employment days are less likely to default
- In default cases, a majority of clients have an annuity amount greater than 25000 (median value).
- More aged Clients are less prone to default

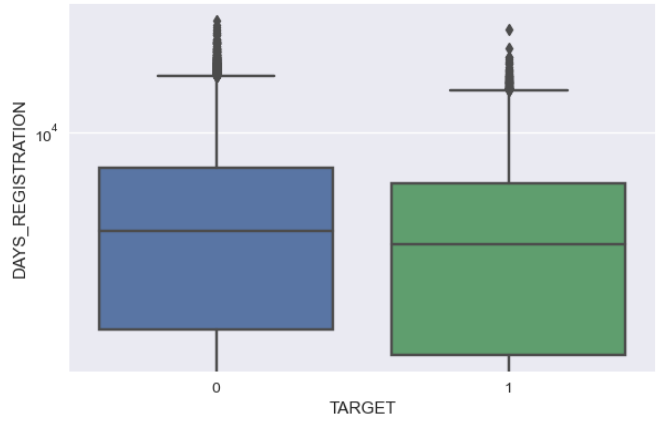
Inferences

- There is a higher occurrence of clients who updated their registration details more than 4000 days after loan approval
- The application process initiation hours are similar for both default and non-default cases.

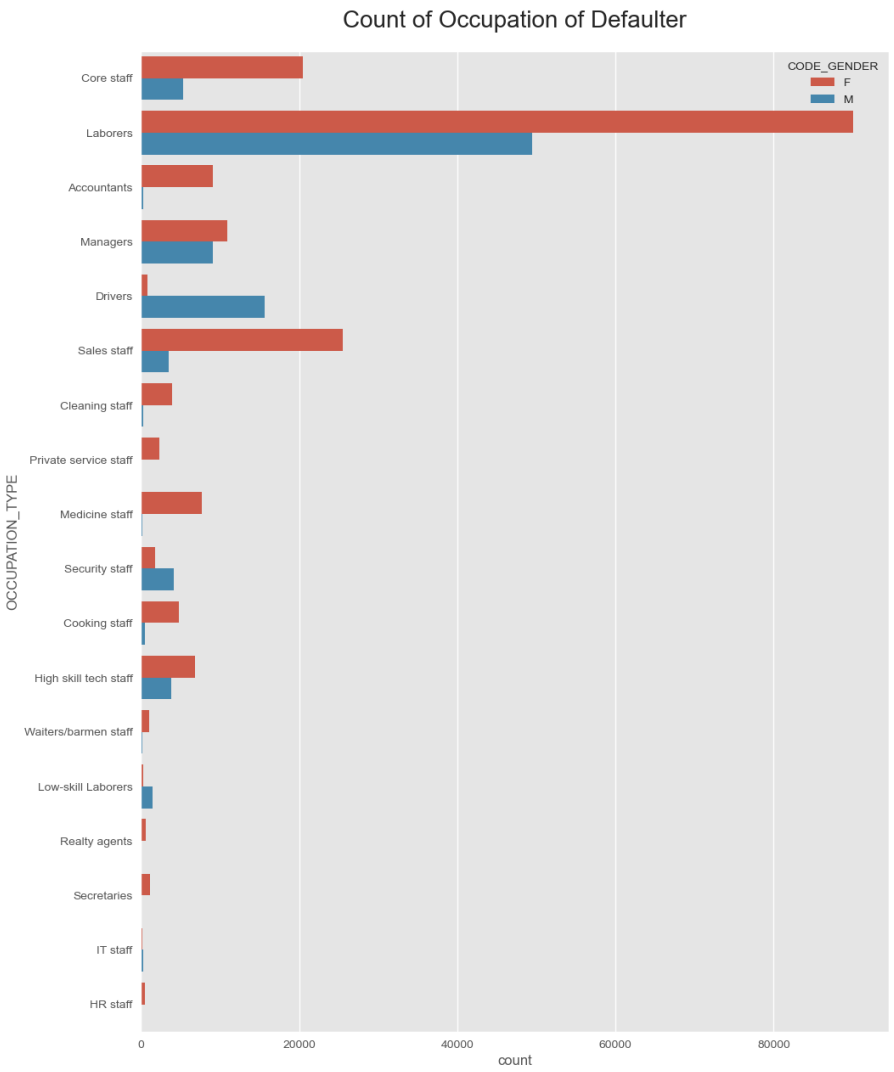


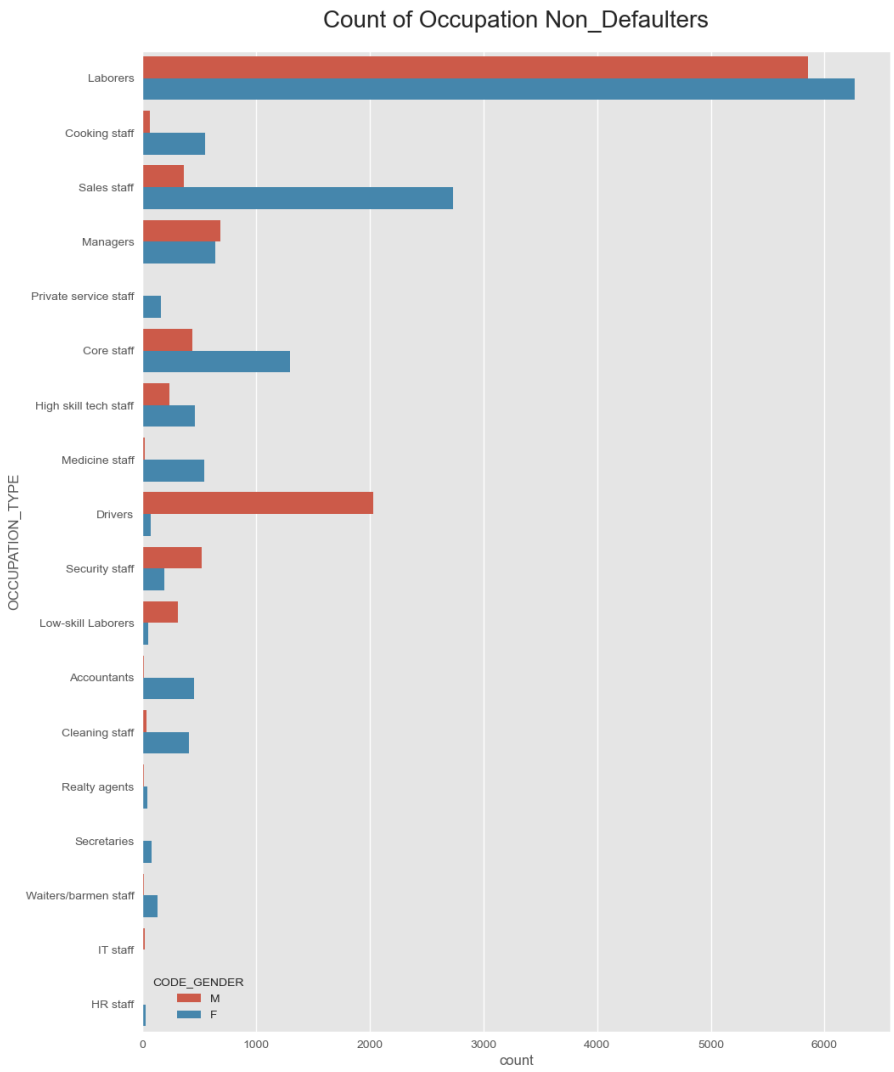






Type of Occupation with CODE_GENDER





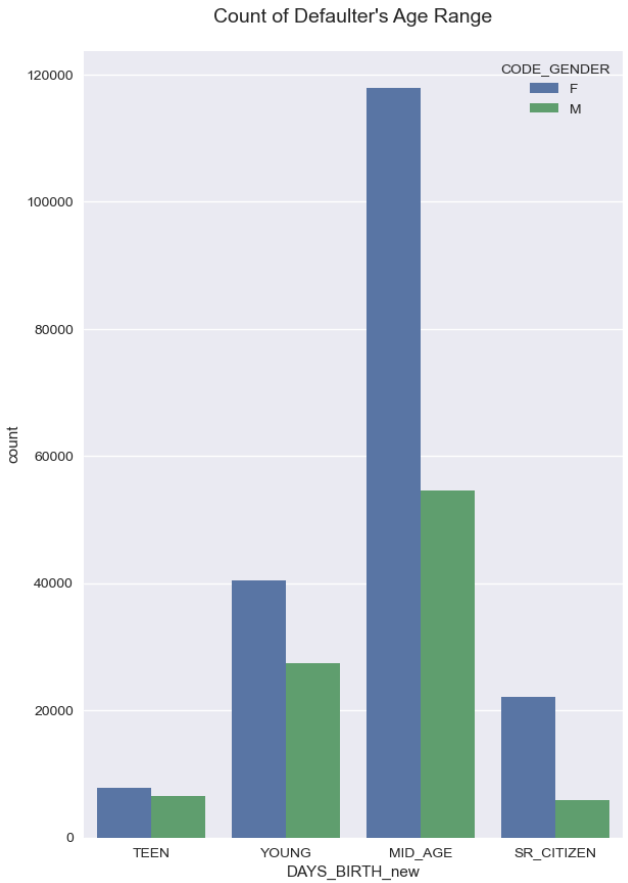
Inferences

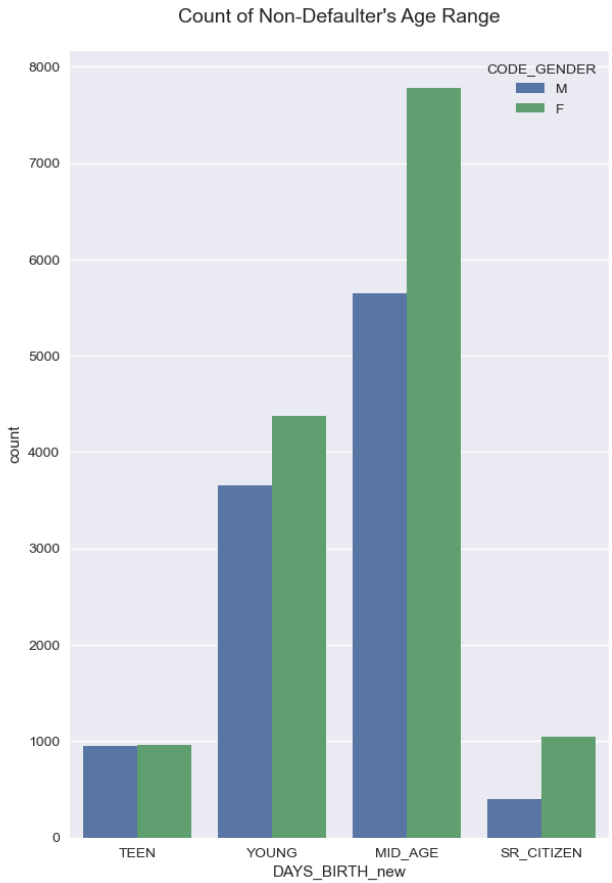
- Laborers are the highest number of population in both defaulter and non-defaulters.
- Male Laborers are more in count of defaulter than non-defaulters.
- Female Laborers are more in count of non-defaulters than defaulters.
- Female category better than Male for repaying loans on time.

Age of Clients with CODE_GENDER

Inferences

- Middle Aged Males are highest in Defaulter plot.
- Middle Aged Females are highest in Non-Defaulters plot.
- Teens and Senior Citizens are lower than Young and Middle aged population in both Defaulter and Non-Defaulters plots.

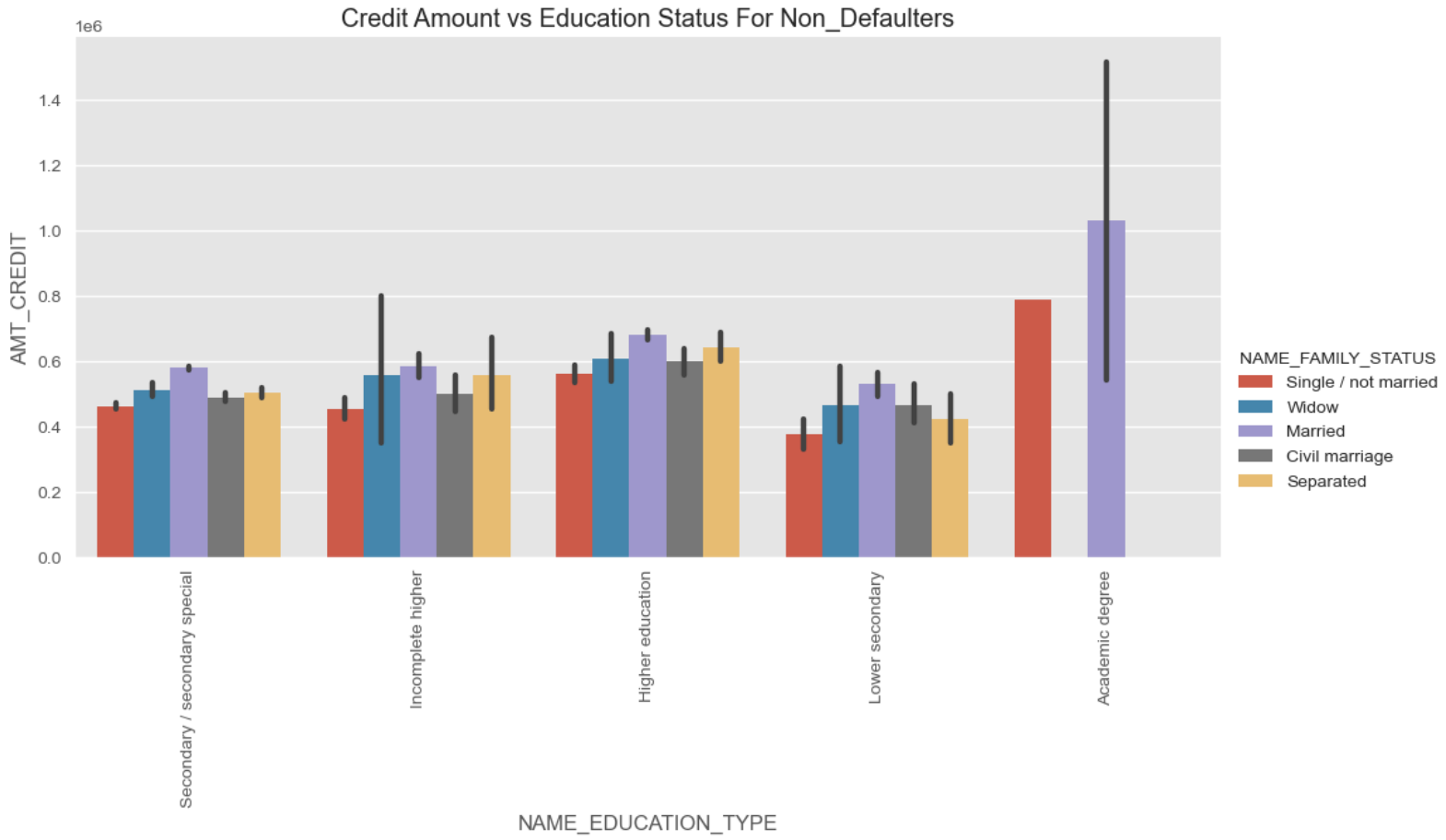


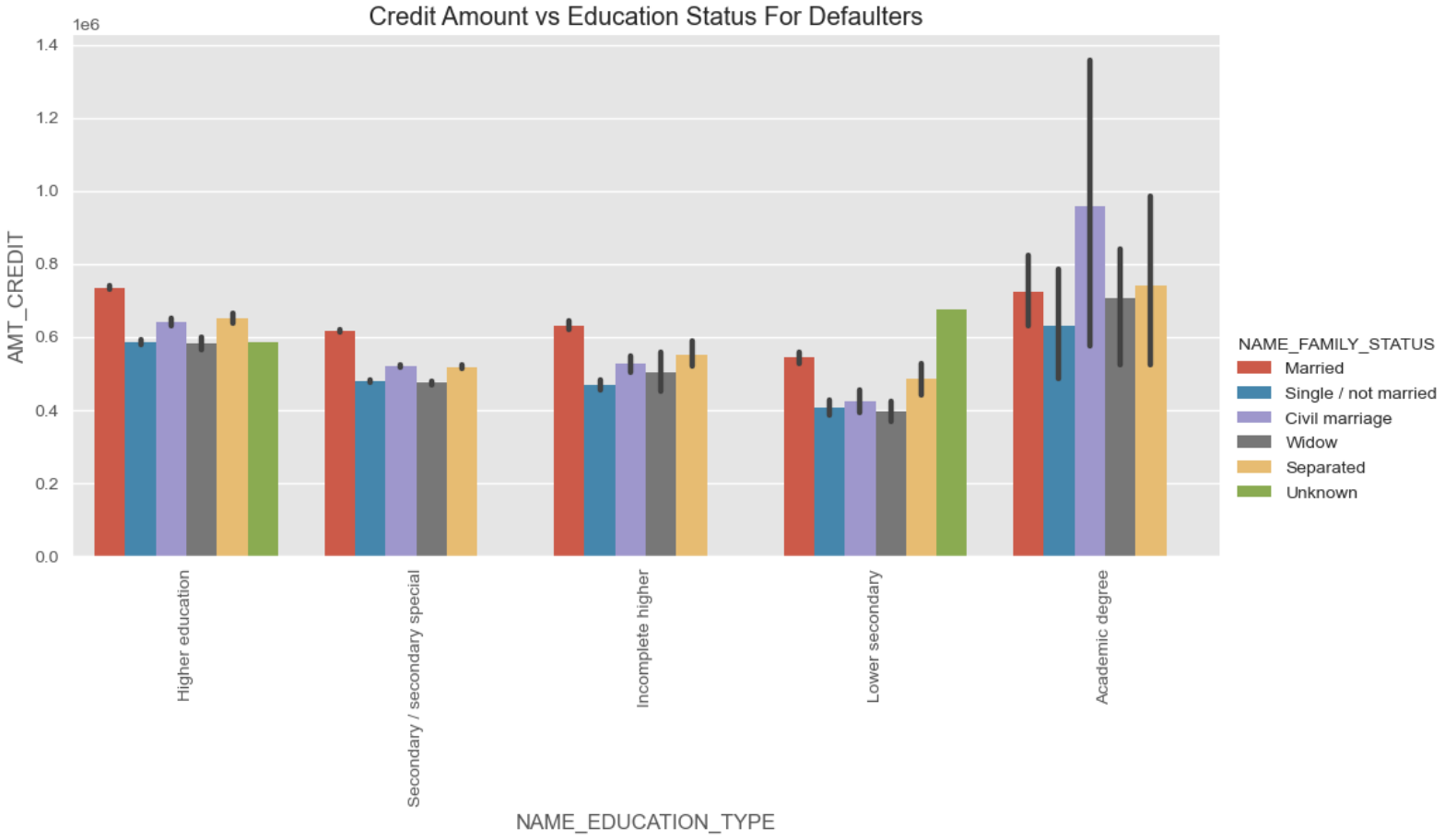


Analysis of Credit amount with respect to Education status

Inferences

- Academic degree holders have only two family types: Single and Married.
- Married customers with academic degrees typically exhibit higher credit amounts
- Across all education segments, married customers tend to have higher credit amounts.
- Customers with lower education levels generally have lower credit amounts.

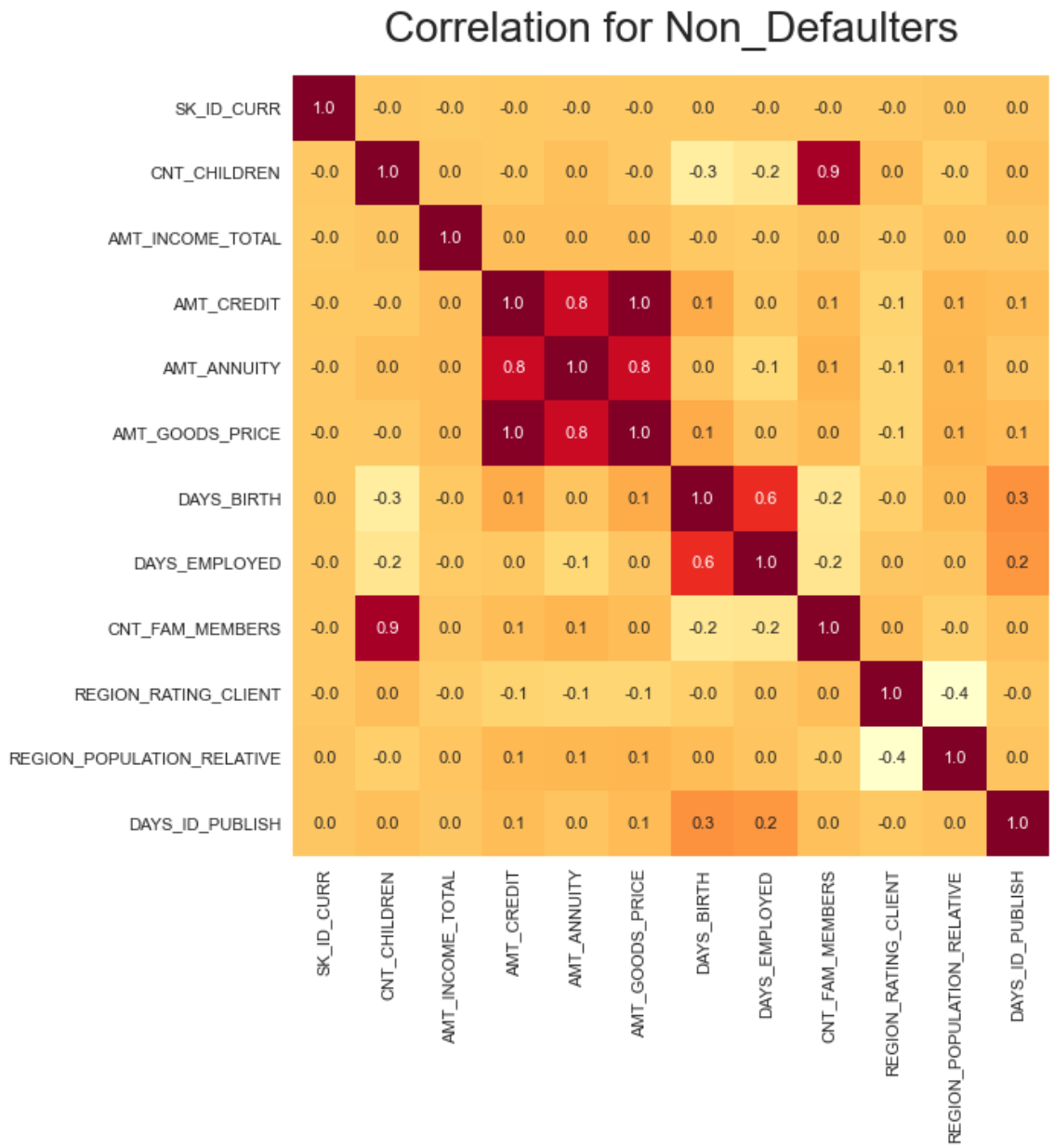


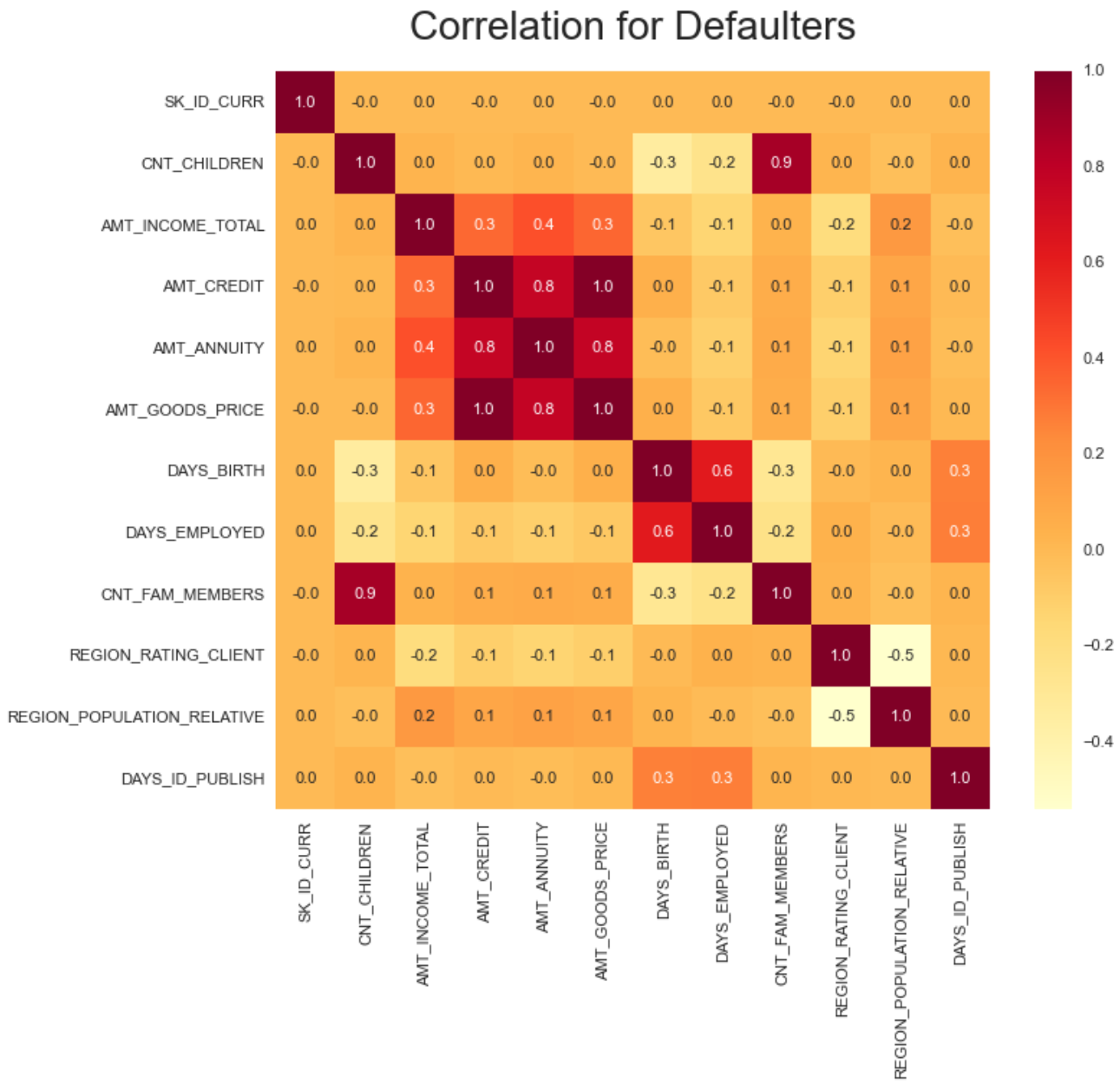


Inferences

- Individuals with academic degrees generally have higher credit amounts, with the civil marriage segment exhibiting the highest credit amounts among them.
- Customers with lower education levels tend to have lower credit amounts, with widows having the lowest credit amounts among this group.
- Married customers, across nearly all education segments except lower secondary and academic degrees, tend to have higher credit amounts.

Correlation





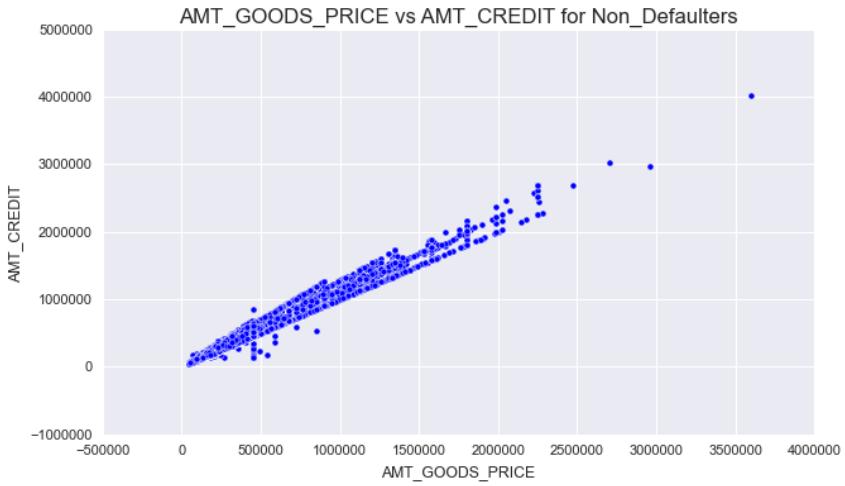
Inferences from the correlation matrix

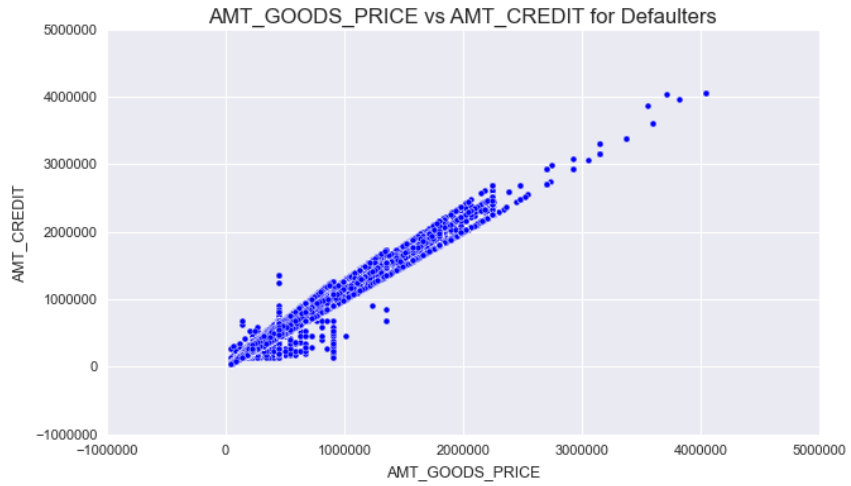
- Correlation represents strength of relationship between variable.
- Correlation lies in the range of -1 to 1

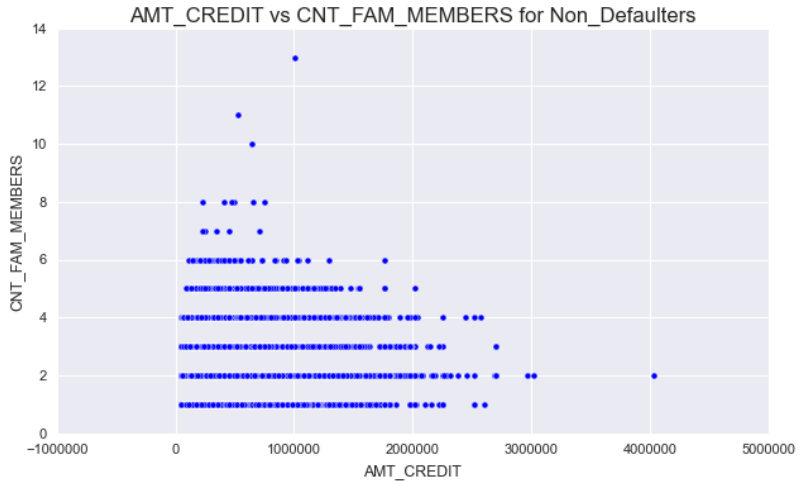
Both for Non-Defaulters(Target 0) and Defaulters(Target 1)the following columns have high correlation values.

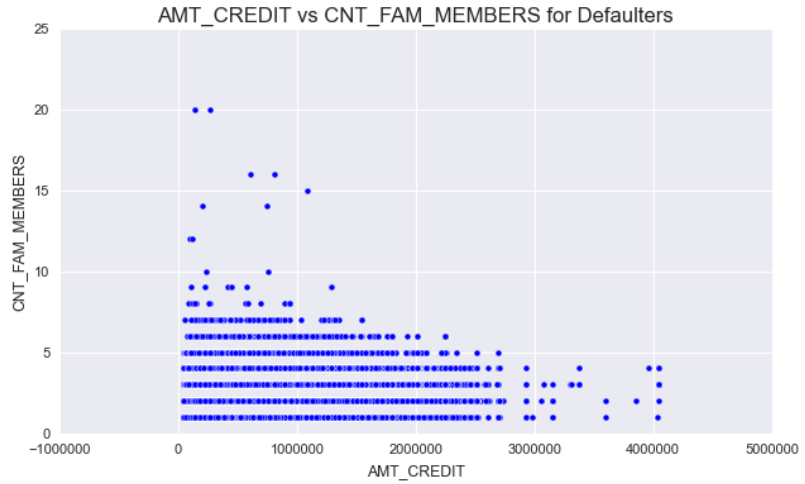
- AMT_GOODS_PRICE and AMT_CREDIT
- AMT_ANNUITY and AMT_INCOME_TOTAL
- AMT_ANNUITY and AMT_GOODS_PRICE
- AMT_ANNUITY and AMT_CREDIT
- AMT_INCOME_TOTAL and AMT_GOODS_PRICE
- CNT_FAM_MEMBER and CNT_CHILDREN

Bivariate analysis on the correlated columns





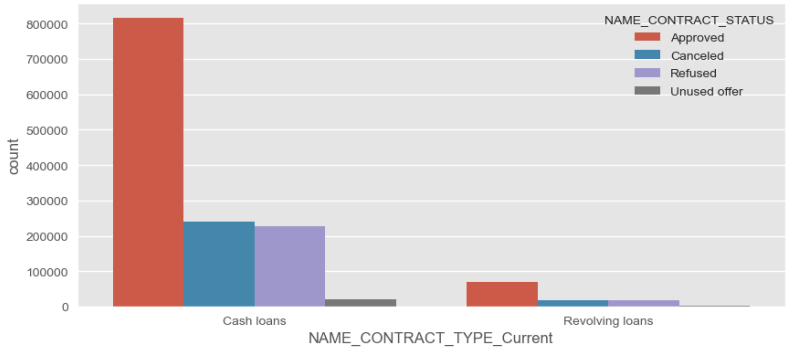


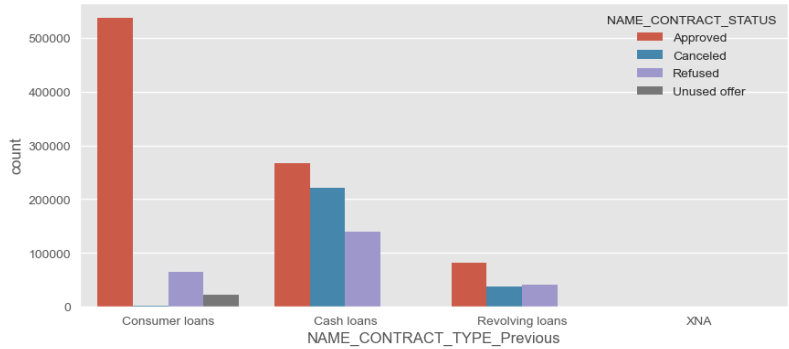


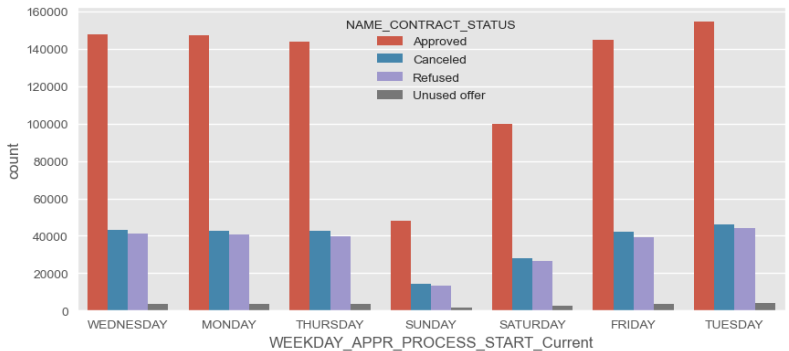
Inferences

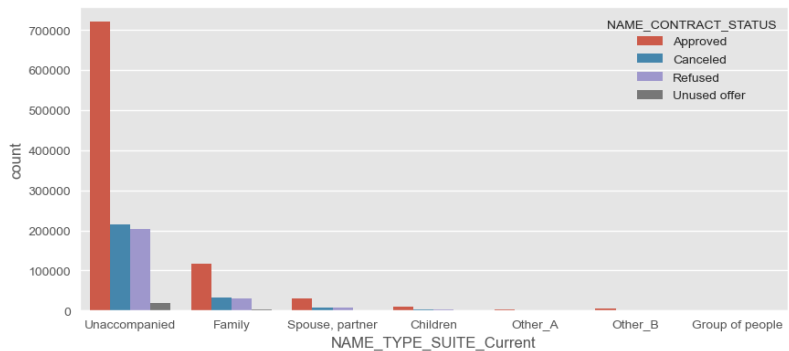
- We can see that the density in the lower left corner is similar in both the case, so the people are equally likely to default if the family is small and the AMT_CREDIT is low. We can observe that larger families and people with larger AMT_CREDIT default less often.

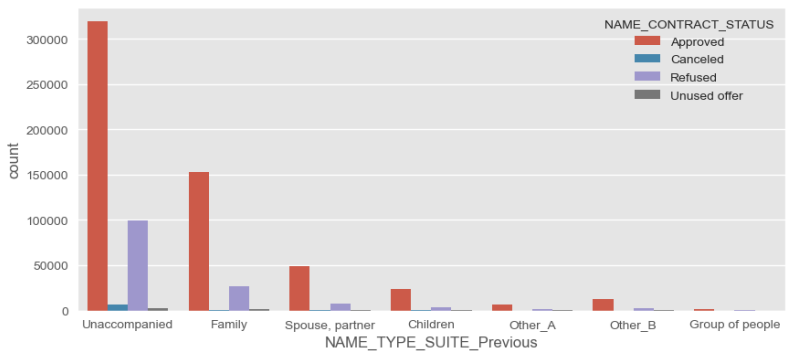
Univariate Analysis after merging ‘application_data’ and ‘previous_application’ datasets

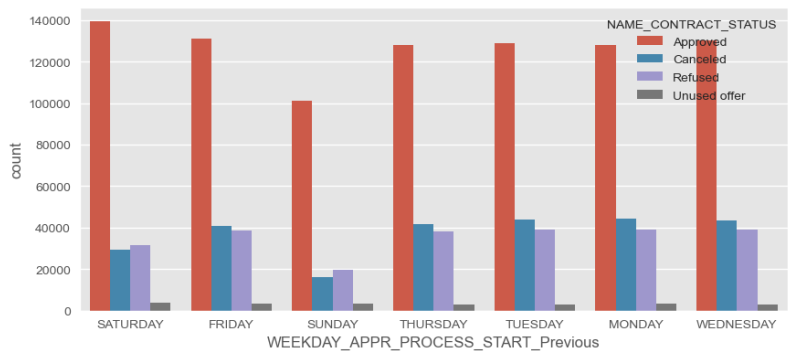






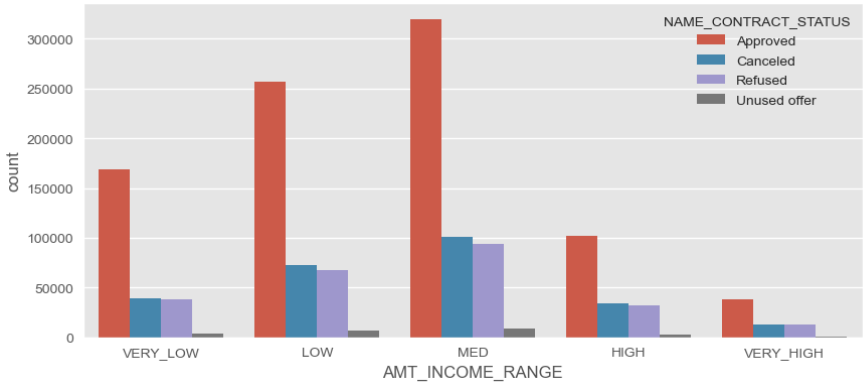


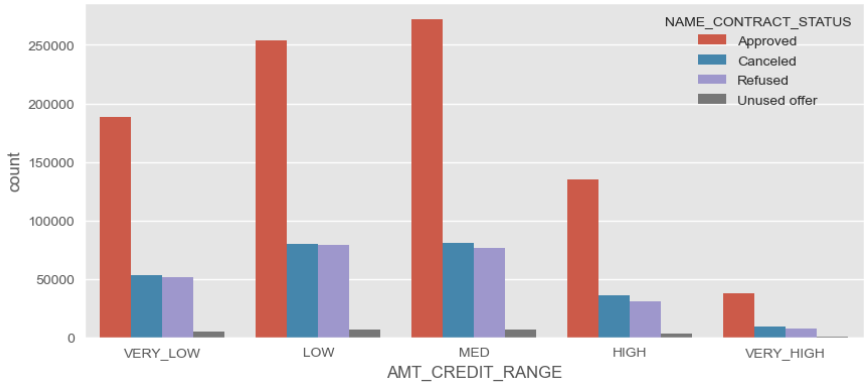


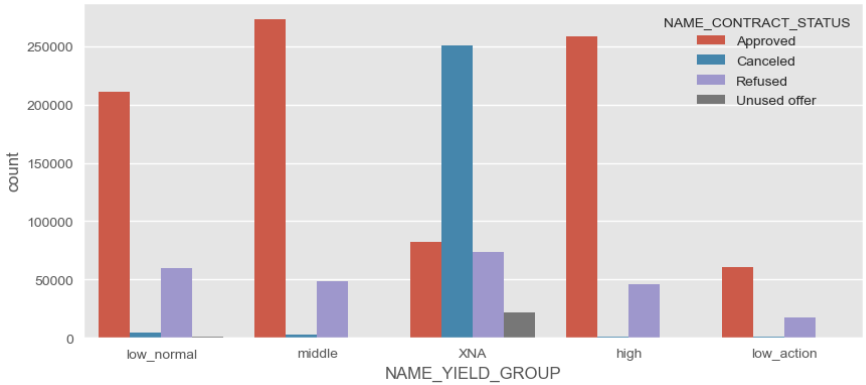


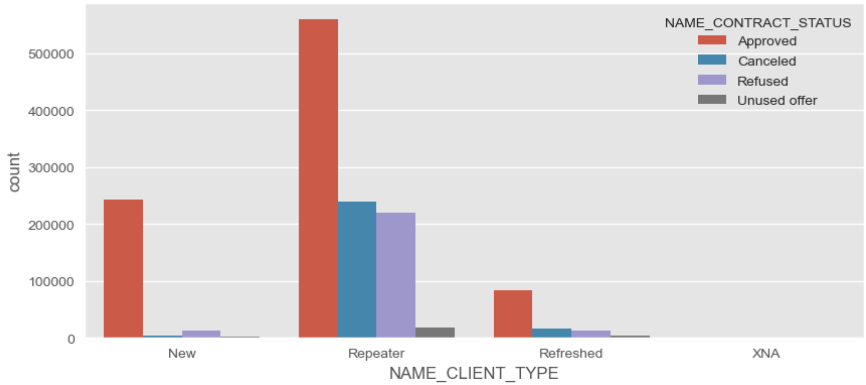
Inferences

- Currently, the bank offers only two types of loans: Cash and Revolving Loans
- In the past, the bank provided Cash, Revolving, and Consumer loans, with Consumer loans having the highest count. Now, Cash loans have the highest count.
- In previous applications, Saturday boasted the highest approval rate, while in the current application, it is Tuesday.



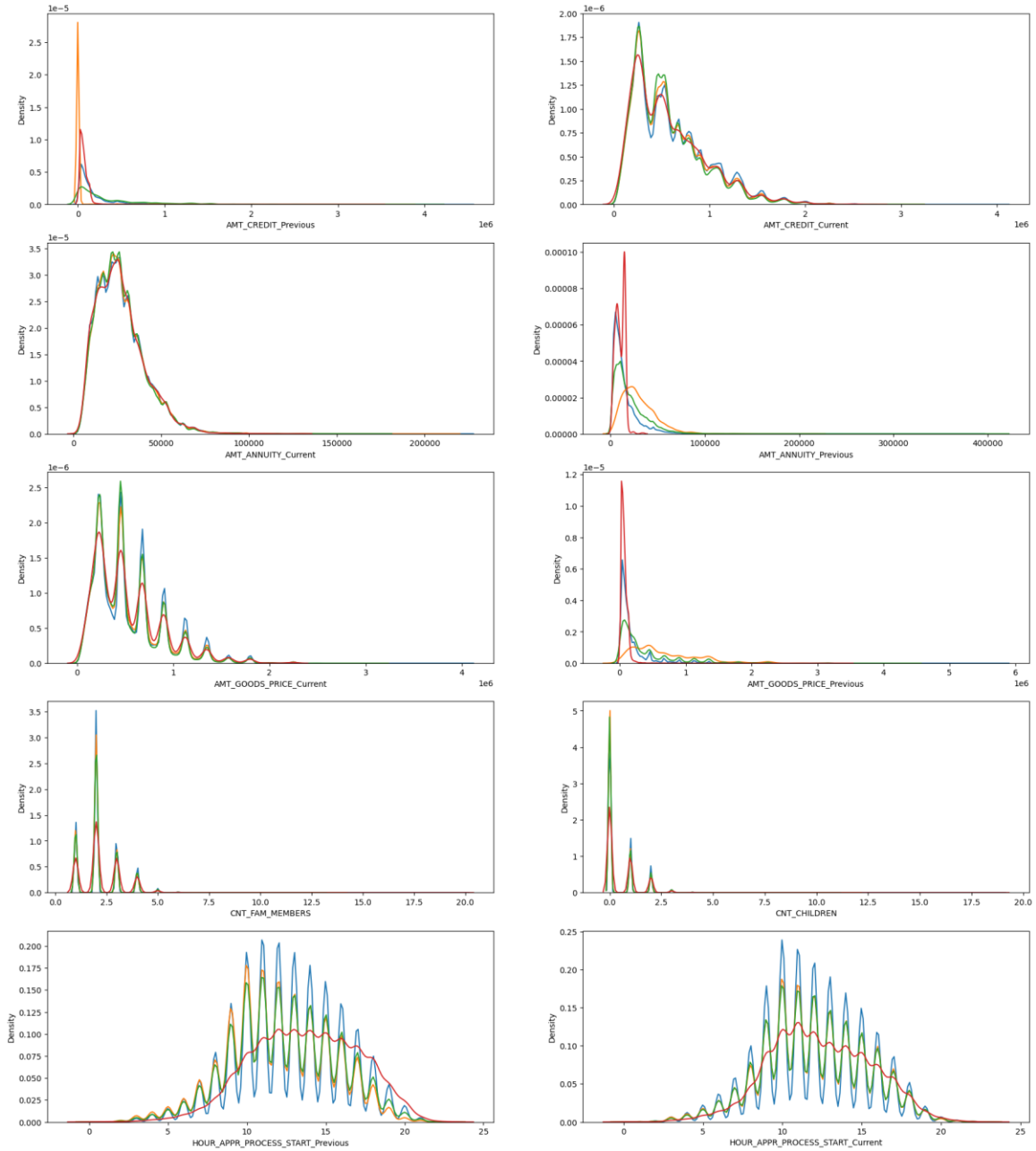






Insights

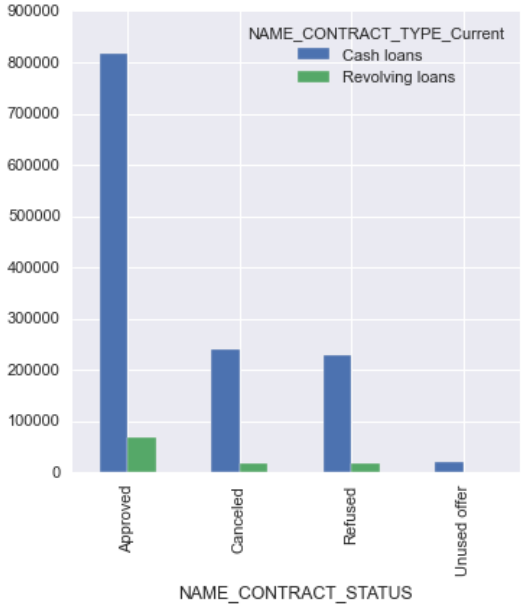
- The value of AMT_CREDIT_RANGE doesn't impact the approval of loans
- Unaccompanied individuals hold the highest number in both NAME_CONTRACT_TYPE_Previous and NAME_CONTRACT_TYPE_Current.
- The NAME_YIELD_GROUP in the middle category records the highest approval rates.
- Repeater applicants secure the highest number of approved loans
- Low AMT_INCOME_RANGE exhibits the highest approval rates.

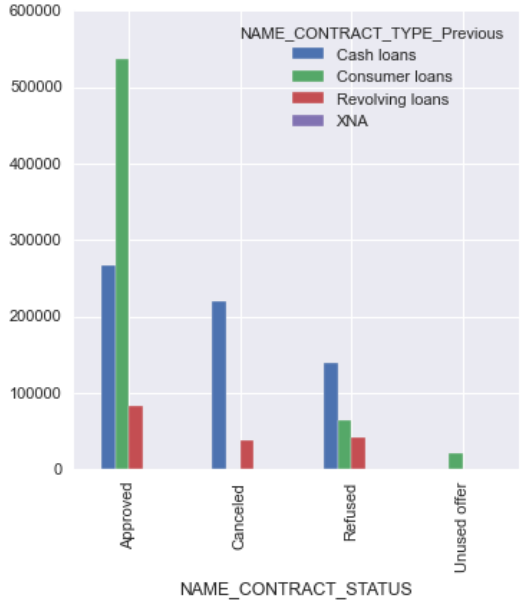


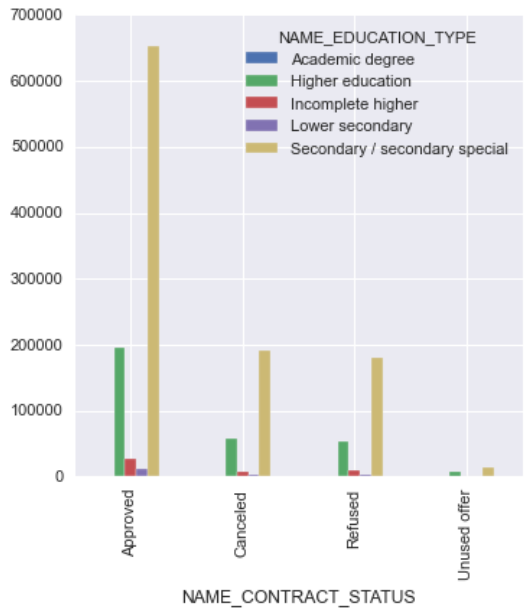
Inferences

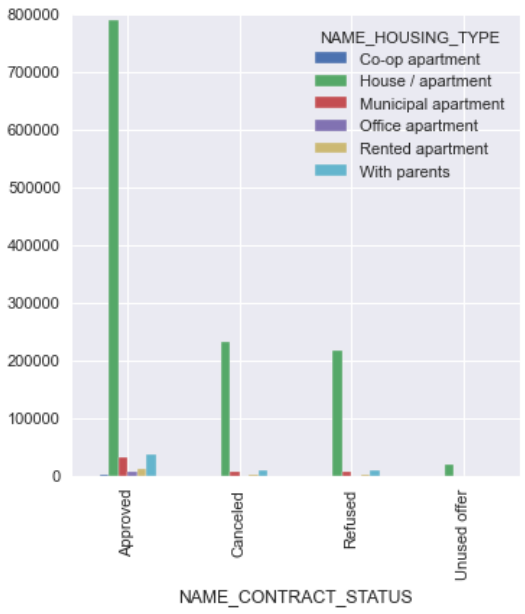
- Nuclear families demonstrate a tendency to apply for more loans.
- Previously, the bank experienced a high number of unused offers, whereas currently, refusals are predominant, especially concerning AMT_GOODS_PRICE.
- Previously, the bank had a substantial number of unused offers; presently, cancellations/refusals are comparable, specifically regarding AMT_ANNUITY.
- It is observed that, a significant volume of applications is submitted between 9 AM to 2 PM in both Current and Previous datasets.
- Consequently, the busiest hours for the bank are from 9 AM to 2 PM.

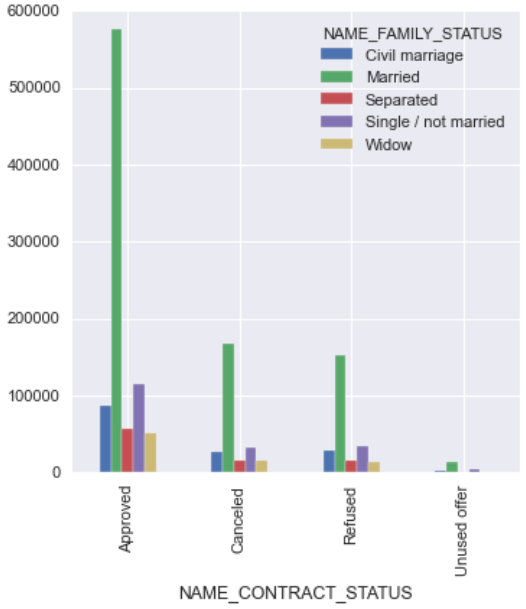
Bivariate Analysis after merging ‘application_data’ and ‘previous_application’ datasets

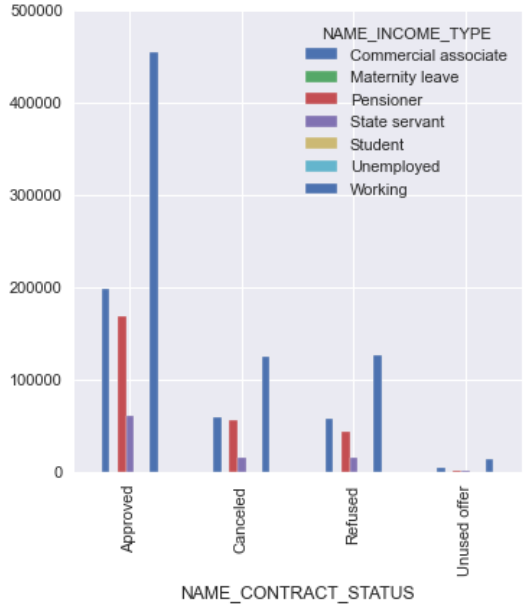


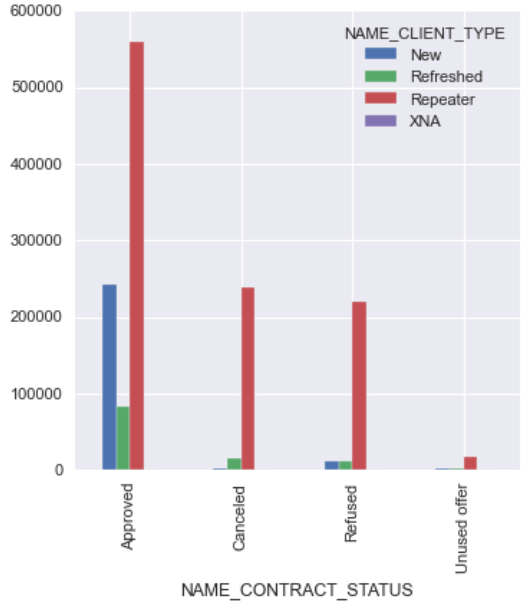






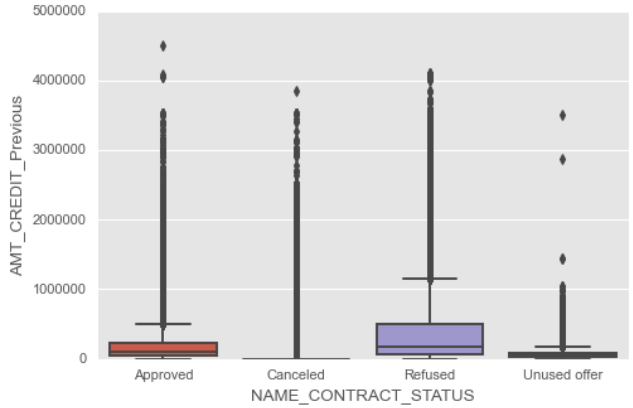


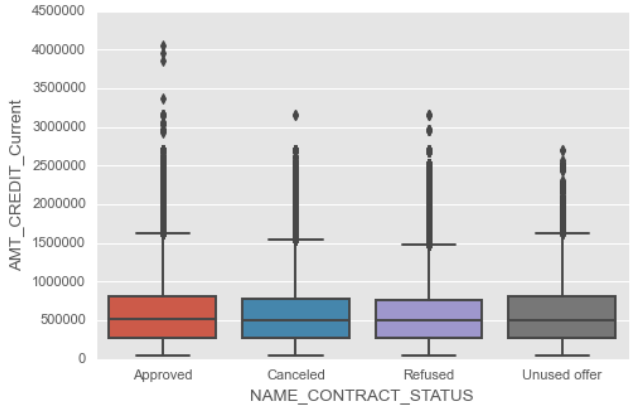


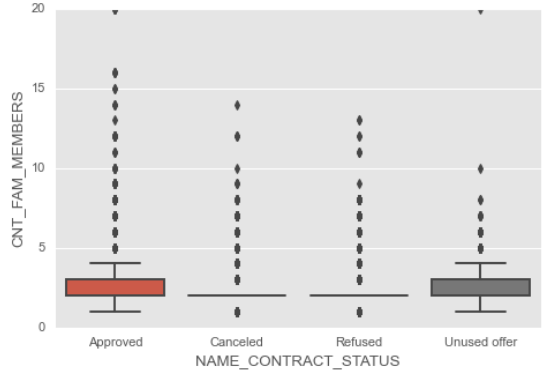


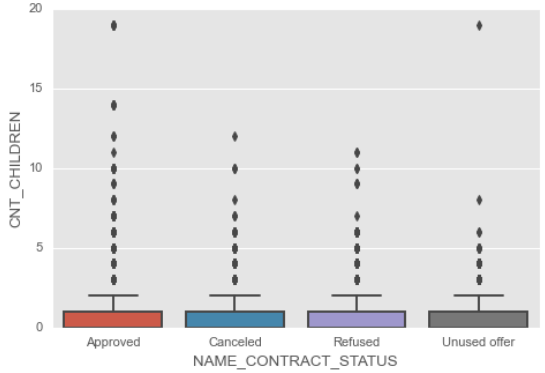
Inferences

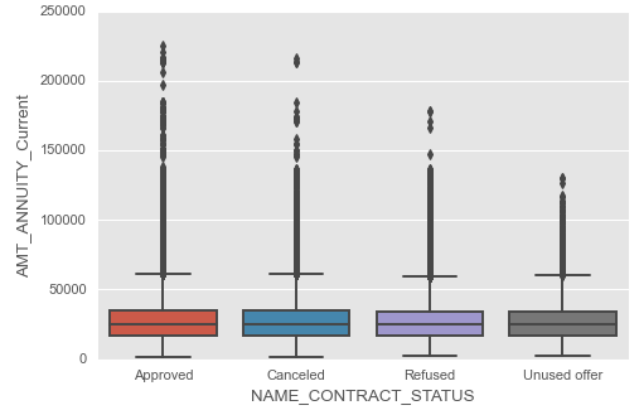
- Generally cash loans were more approved
- Commercial associates, married people, people living in own house/apartment, consumers loans, and Repeaters are majorly approved for loans

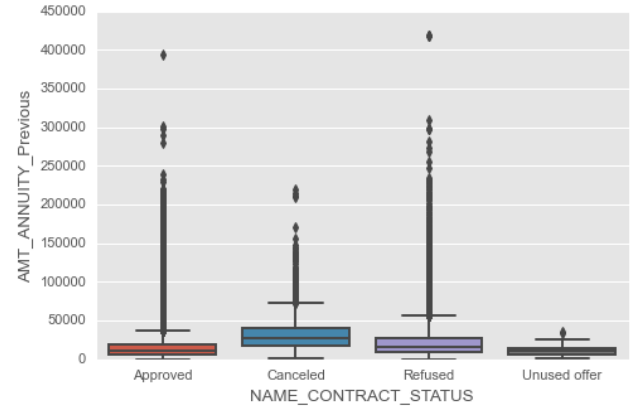


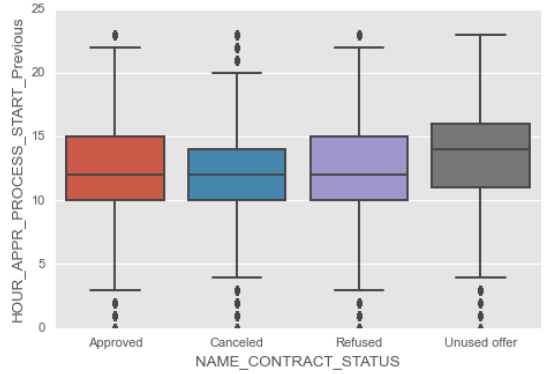


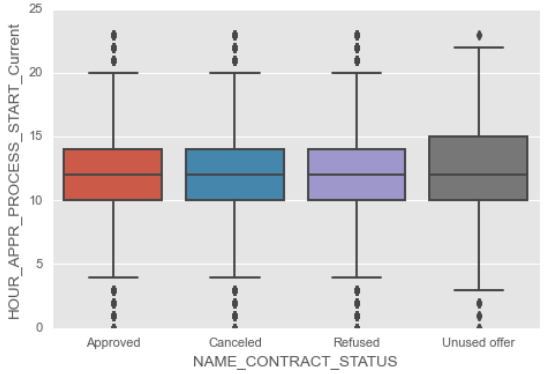


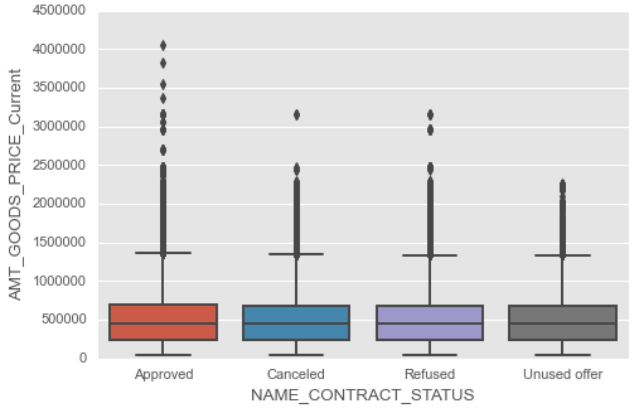


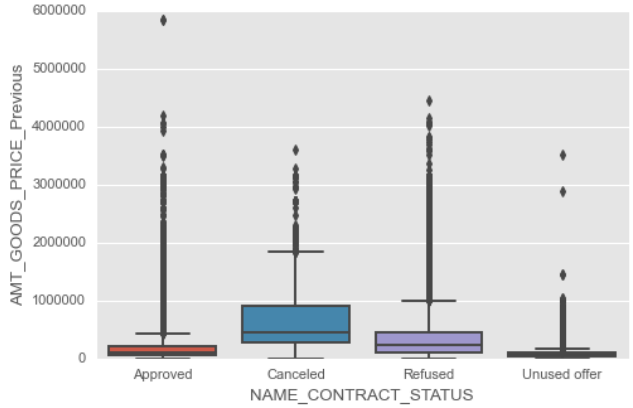






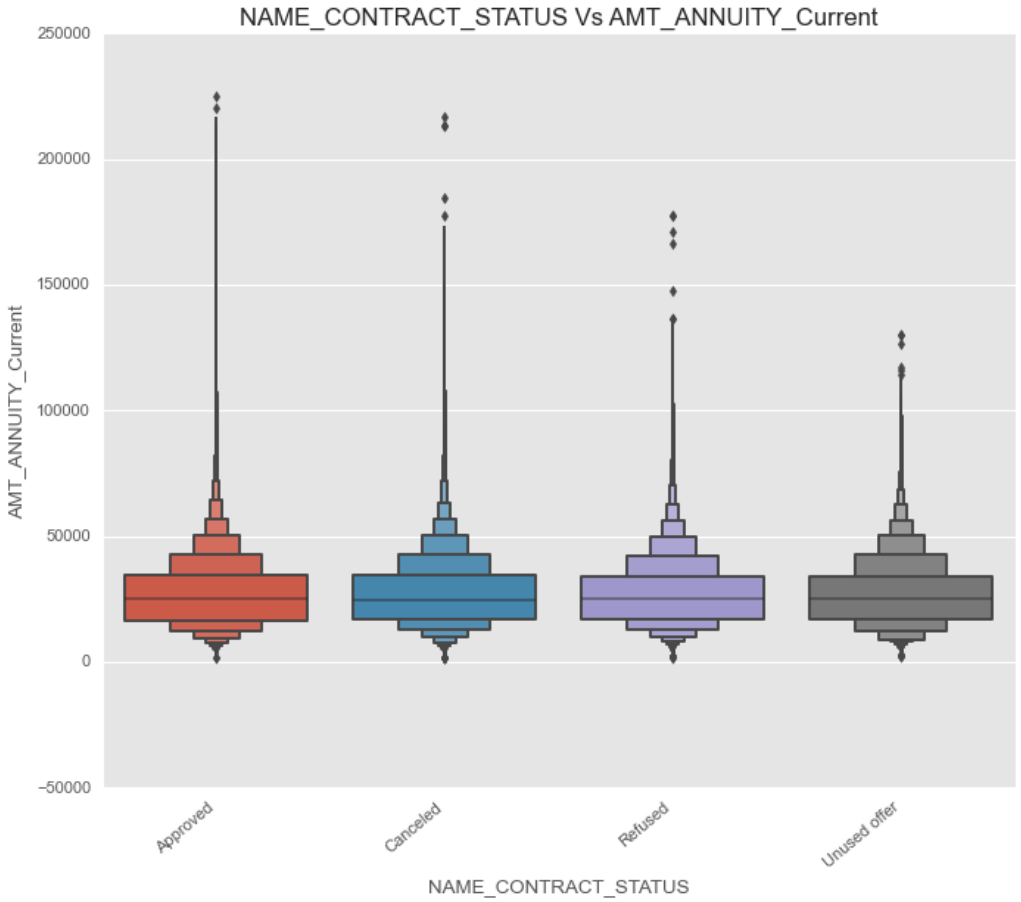






Inferences

- The highest number of refused cases is associated with AMT_CREDIT_Previous, while AMT_CREDIT_Current shows similarity across all four cases.
- The duration spent on unused offers surpasses that in other categories.
- Nuclear families (composed of 2-3 people) attain the highest approval rates.



Inferences

- Previously, a majority of applications were either cancelled or refused.
- Currently, Refused/Cancelled/Approved/Unused all four have similar situation for AMT_ANNUITY.
- Previously, a majority of applications were either cancelled or refused,but in the current scenario for AMT_GOODS_PRICE indicates a similar distribution among Refused, Cancelled, Approved, and Unused categories.
- Yet, the present situation for AMT_ANNUITY reflects a comparable distribution among Refused, Cancelled, Approved, and Unused categories.

CONCLUSION

Main variable for Application dataset - **TARGET**
Main variable for Previous dataset - **NAME_CONTRACT_STATUS**

Key factors to prioritize in predicting loans:

- NAME_EDUCATION_TYPE
- AMT_INCOME_TOTAL
- DAYS_BIRTH
- AMT_CREDIT
- DAYS_EMPLOYED
- AMT_ANNUITY
- NAME_INCOME_TYPE
- CODE_GENDER
- NAME_HOUSING_TYPE

Key Inferences from the analysis

- Banks should target more on women as they are less defaulter than men and make payments on time.
- Banks should avoid giving loans to middle-aged men as there are riskier in becoming defaulter and can focus on middle-aged women.
- They could target the population with academician degree as they have highest income level than other education level population.