CREDIT EXPLORATORY DATA ANALYSIS CASE STUDY

Executive Post Graduate Programme in Data Science
By Dharshak Chandra P
DS C52 Batch

Business Objective

- 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. Identification of such applicants using EDA is the aim of this case study.

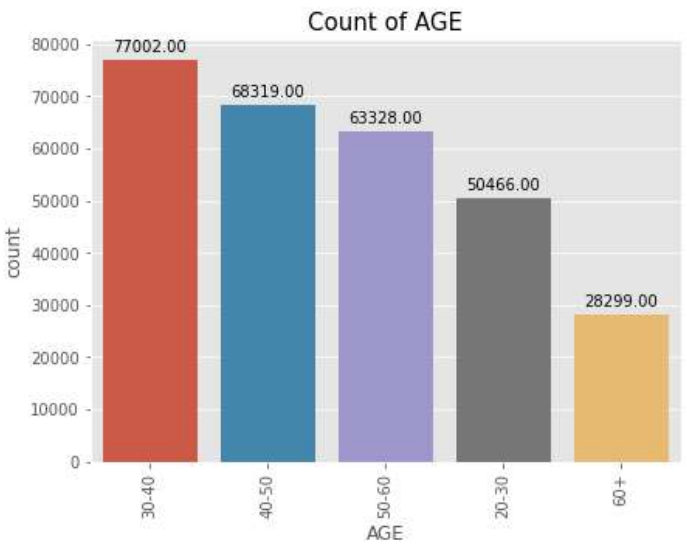
Function created to do Univariate Analysis on Categorical columns

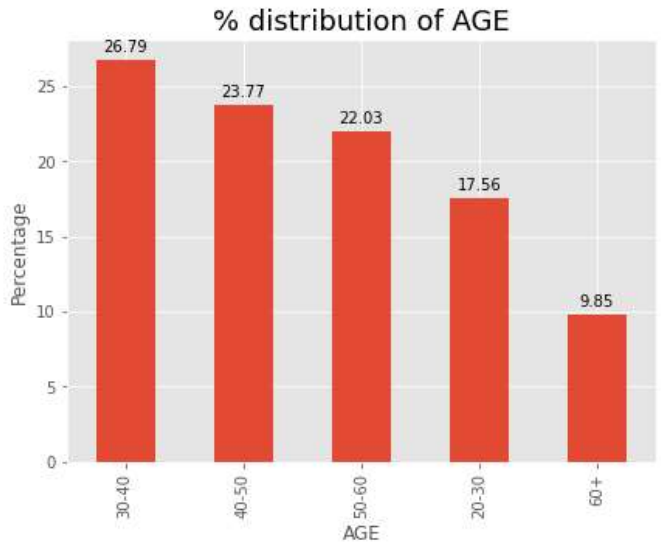
```
def cat_uni_analysis(x):
    print('Categorical Univariate Analysis of {}'.format(x))

    plt.figure(figsize=[15,5])
    plt.subplot(1,2,1)
    plots = sns.countplot(app0[x],order = app0[x].value_counts().index)
    plt.xticks(rotation = 90)
    plt.title('Count of {}'.format(x), fontdict={'fontsize':15})
    for bar in plots.patches:
        plots.annotate(format(bar.get_height(), '.2f'),
                   (bar.get_x() + bar.get_width() / 2,
                    bar.get_height()), ha='center', va='center',
                   size=10, xytext=(0, 8),
                   textcoords='offset points')

    plt.subplot(1,2,2)
    plt.title('% distribution of {}'.format(x), fontdict={'fontsize':18})
    plt.xlabel(x)
    plt.ylabel('Percentage')
    plots = (app0[x].value_counts()*100/len(app0[x])).plot.bar()
    for bar in plots.patches:
        plots.annotate(format(bar.get_height(), '.2f'),
                   (bar.get_x() + bar.get_width() / 2,
                    bar.get_height()), ha='center', va='center',
                   size=10, xytext=(0, 8),
                   textcoords='offset points')

    fig.tight_layout()
    plt.show()
```

Function created to do Univariate Analysis on Numerical columns

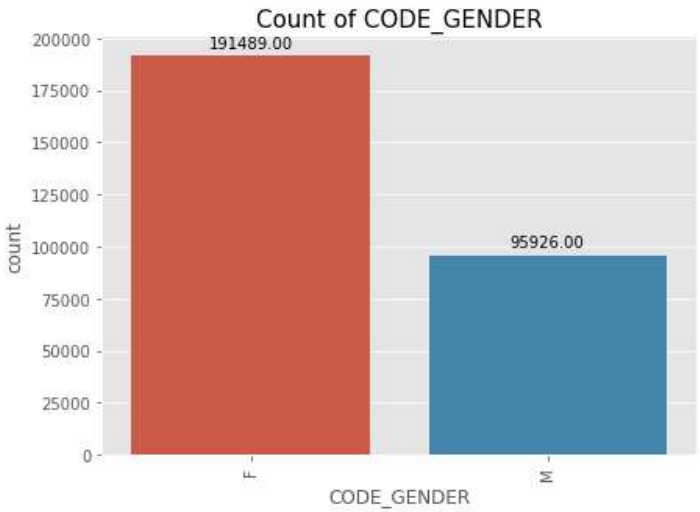
```
def num_uni_analysis(x):
    print('Numerical Univariate Analysis of {}'.format(x))
    print(app0[x].describe())
    plt.figure(figsize=[15,7])
    plt.subplot(1,2,1)
    sns.boxplot(app0[x])
    plt.title('Box Plot of {}:'.format(x),fontdict={'fontsize':18})

    plt.subplot(1,2,2)
    sns.distplot(app0[x],hist=False)
    plt.title('Distplot of {}:'.format(x),fontdict={'fontsize':18})

    fig.tight_layout()
    plt.show()
```

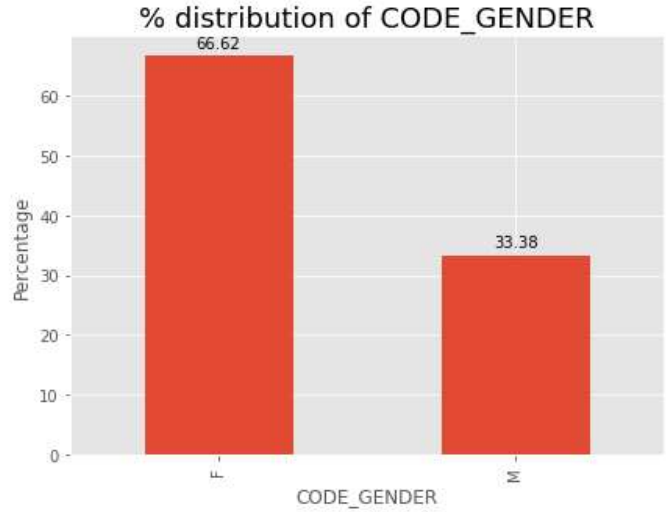
Categorical Univariate Analysis of AGE





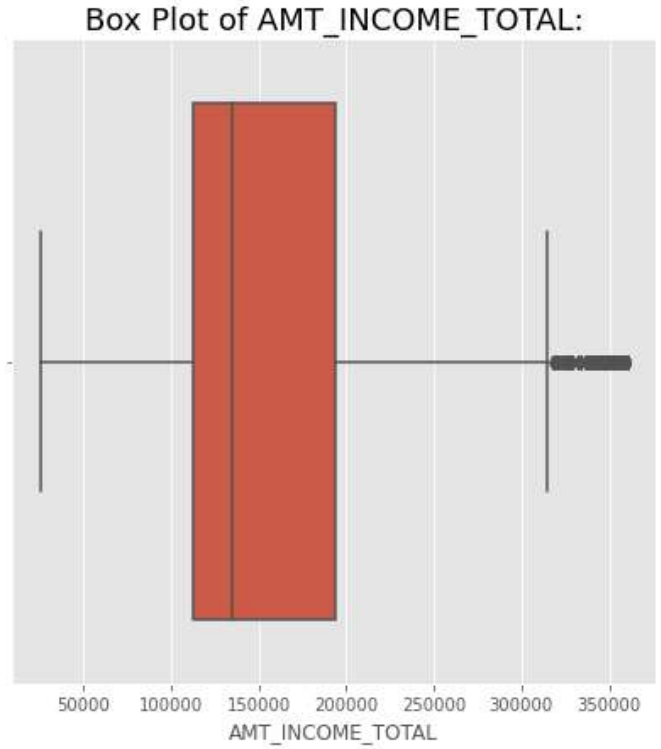
Categorical Univariate Analysis of CODE_GENDER

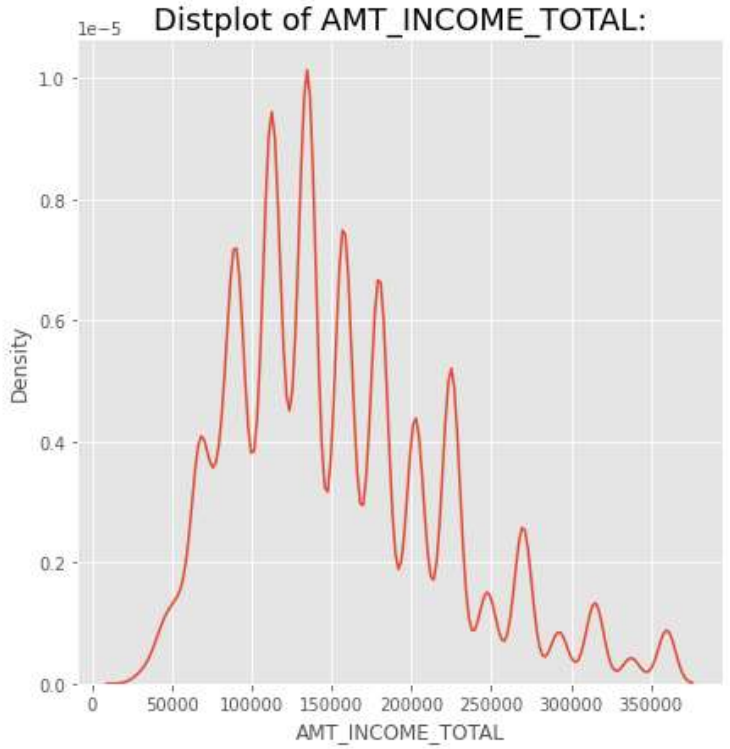




Examples of Visualization after applying the Function to Categorical columns

```
Numerical Univariate Analysis of AMT_INCOME_TOTAL
count    287415.000000
mean     154470.901144
std       66518.029770
min       25650.000000
25%      112500.000000
50%      135000.000000
75%      193500.000000
max      360000.000000
Name: AMT_INCOME_TOTAL, dtype: float64
```

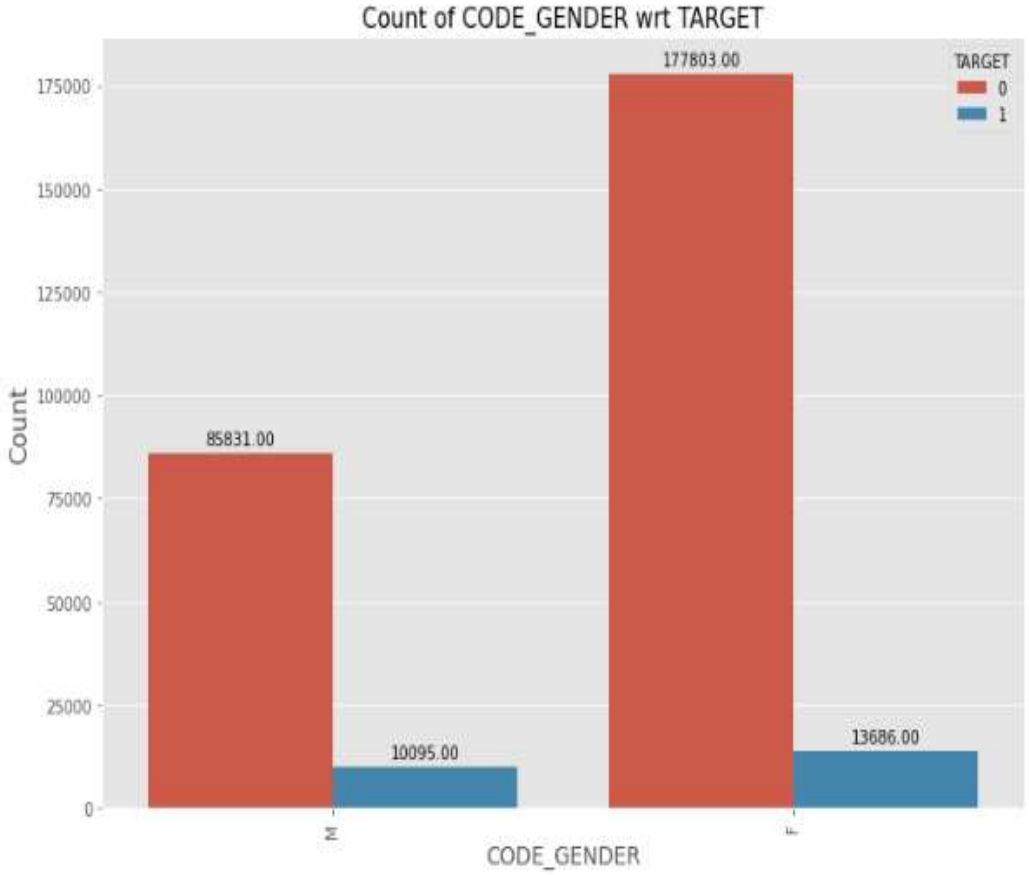


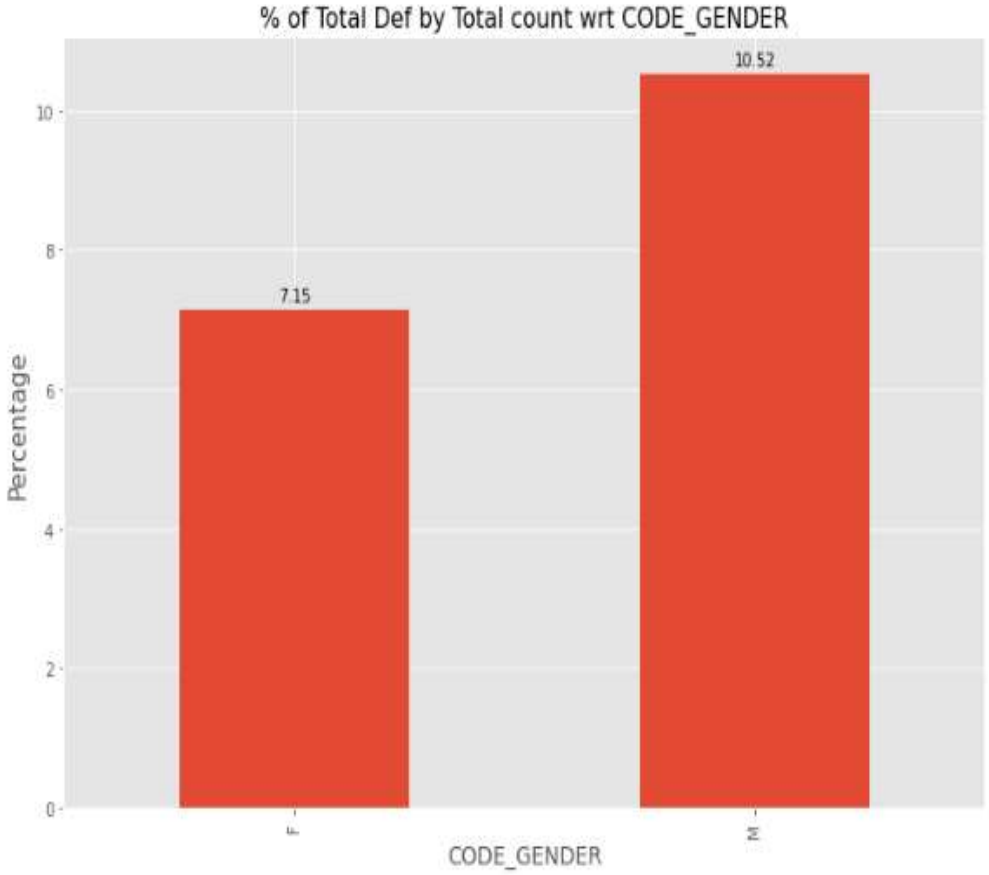


Examples of Visualization after applying the Function to Numerical columns

Segmented Univariate Analysis and Bivariate Analysis

Segmented Categorical Univariate Analysis of CODE_GENDER

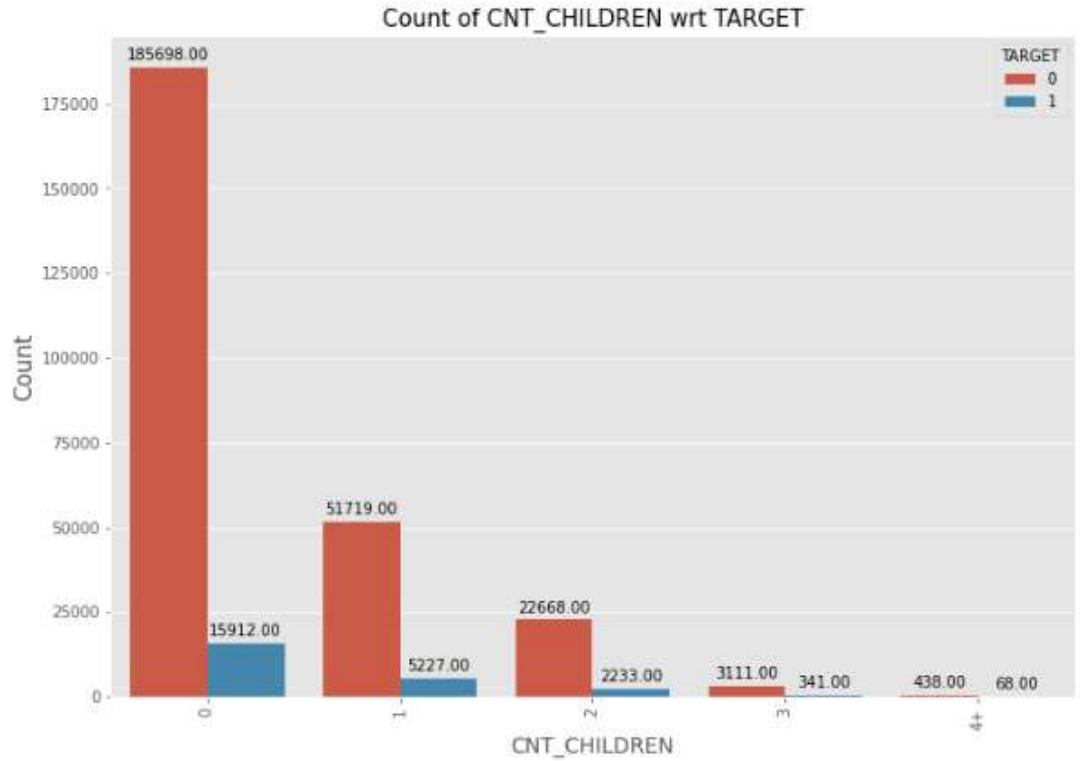


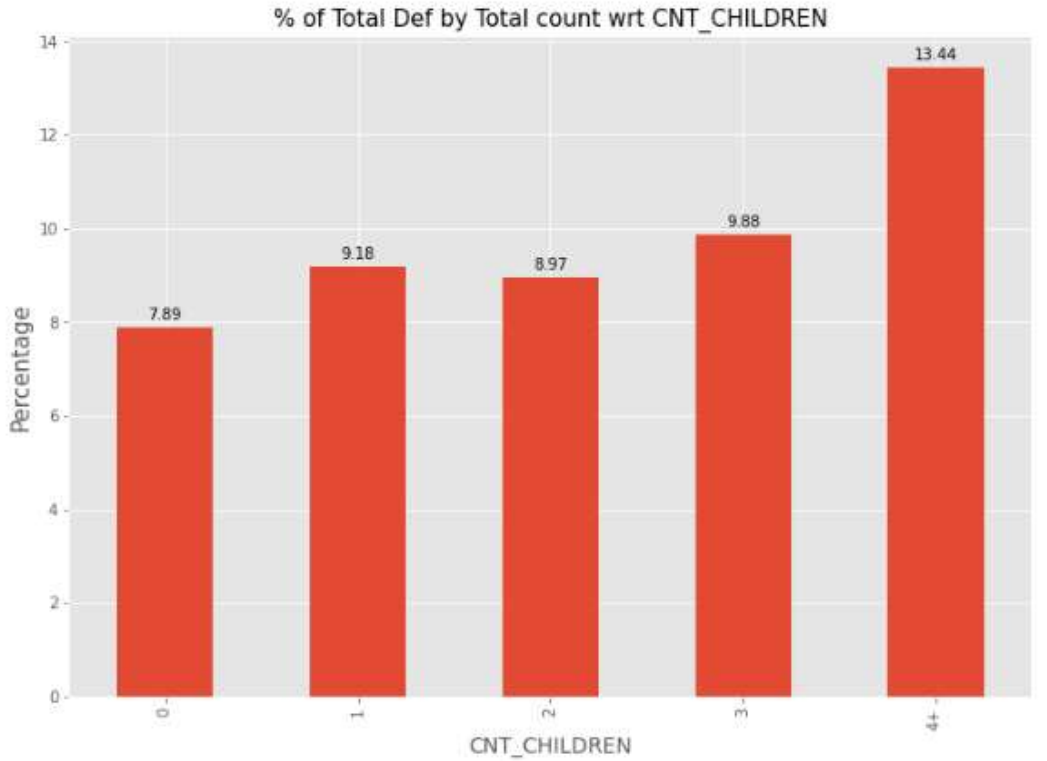


Inference:

- Male are high Defaulters compared to Female

Segmented Categorical Univariate Analysis of CNT_CHILDREN

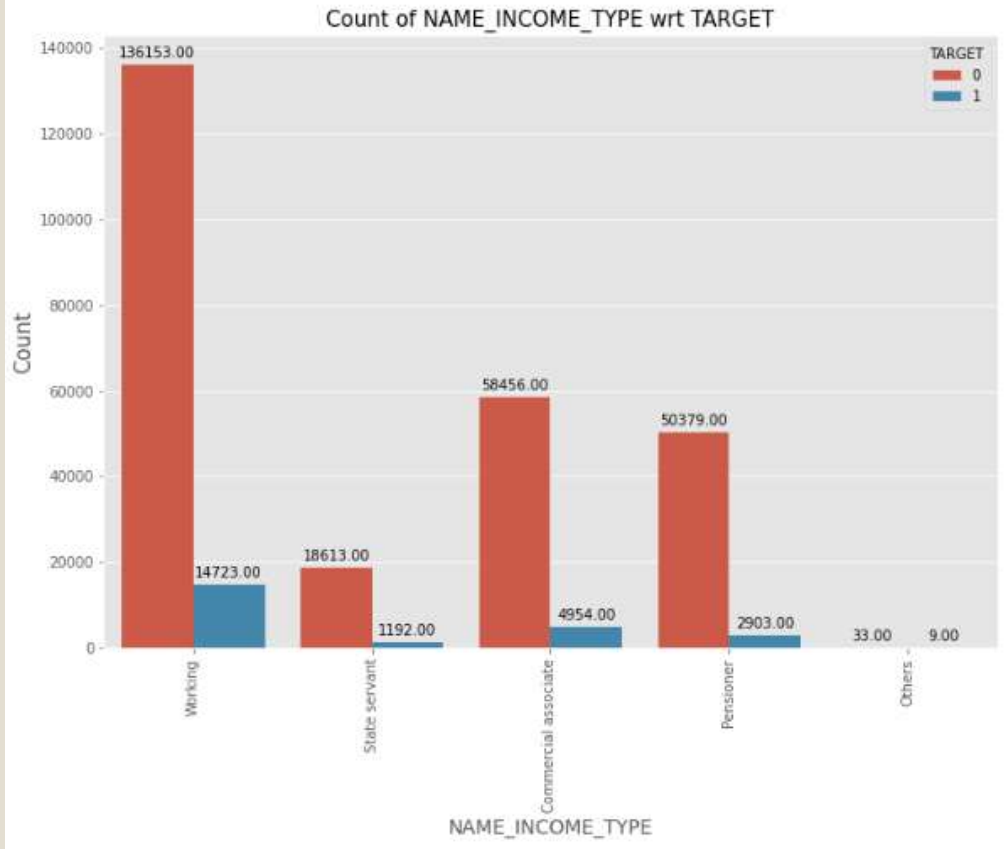


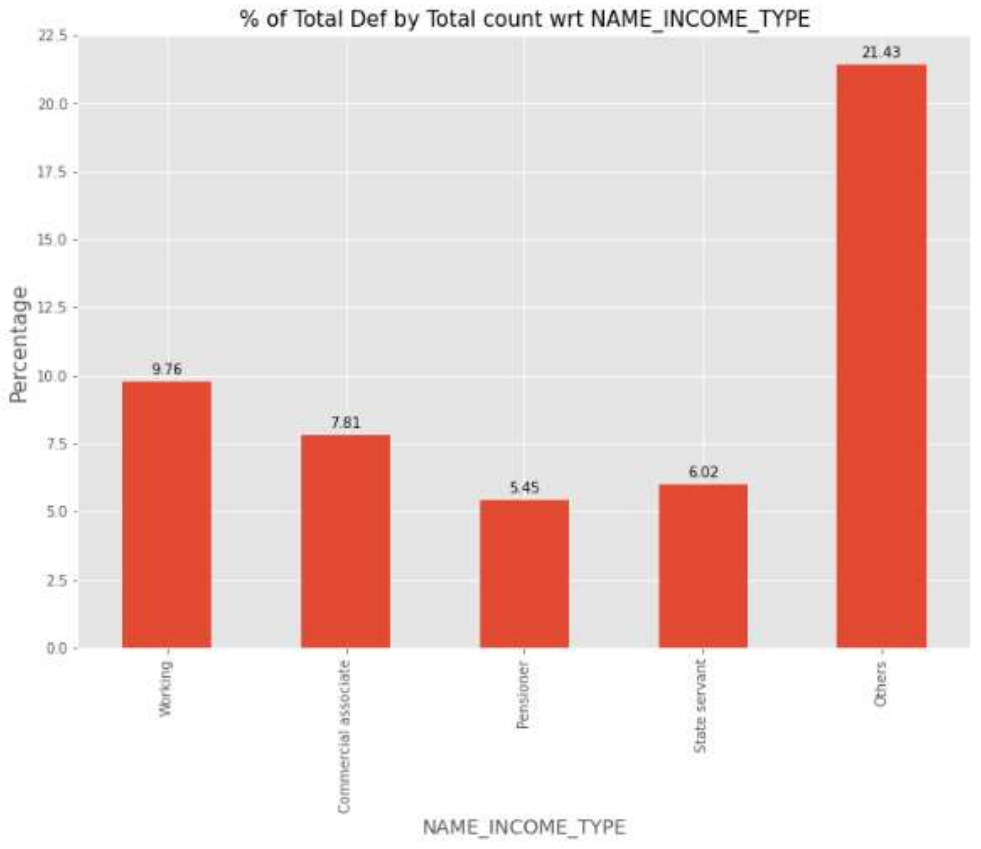


Inference:

- As the number of children increase, the rate of defaulters also increases.

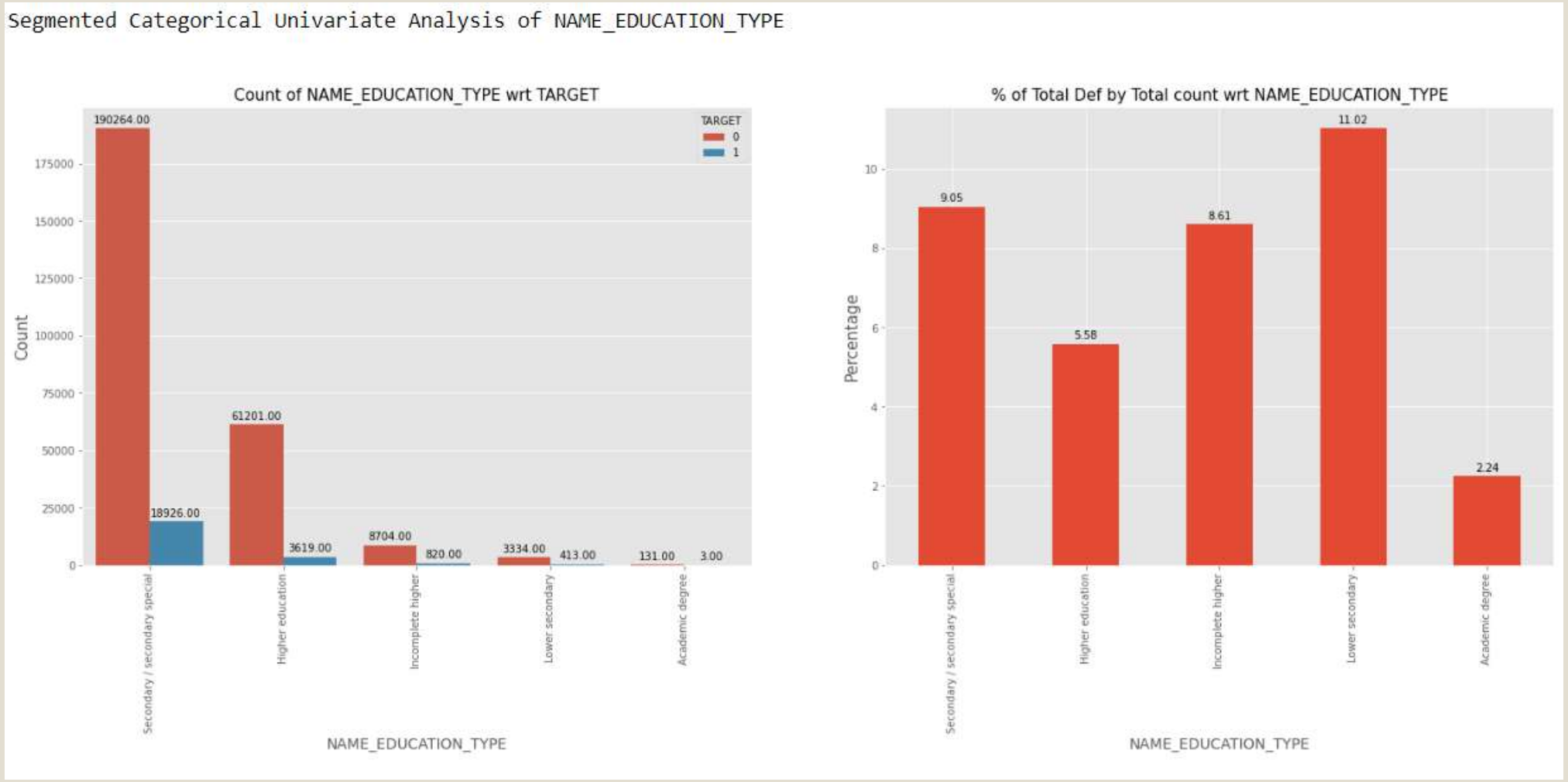
Segmented Categorical Univariate Analysis of NAME_INCOME_TYPE





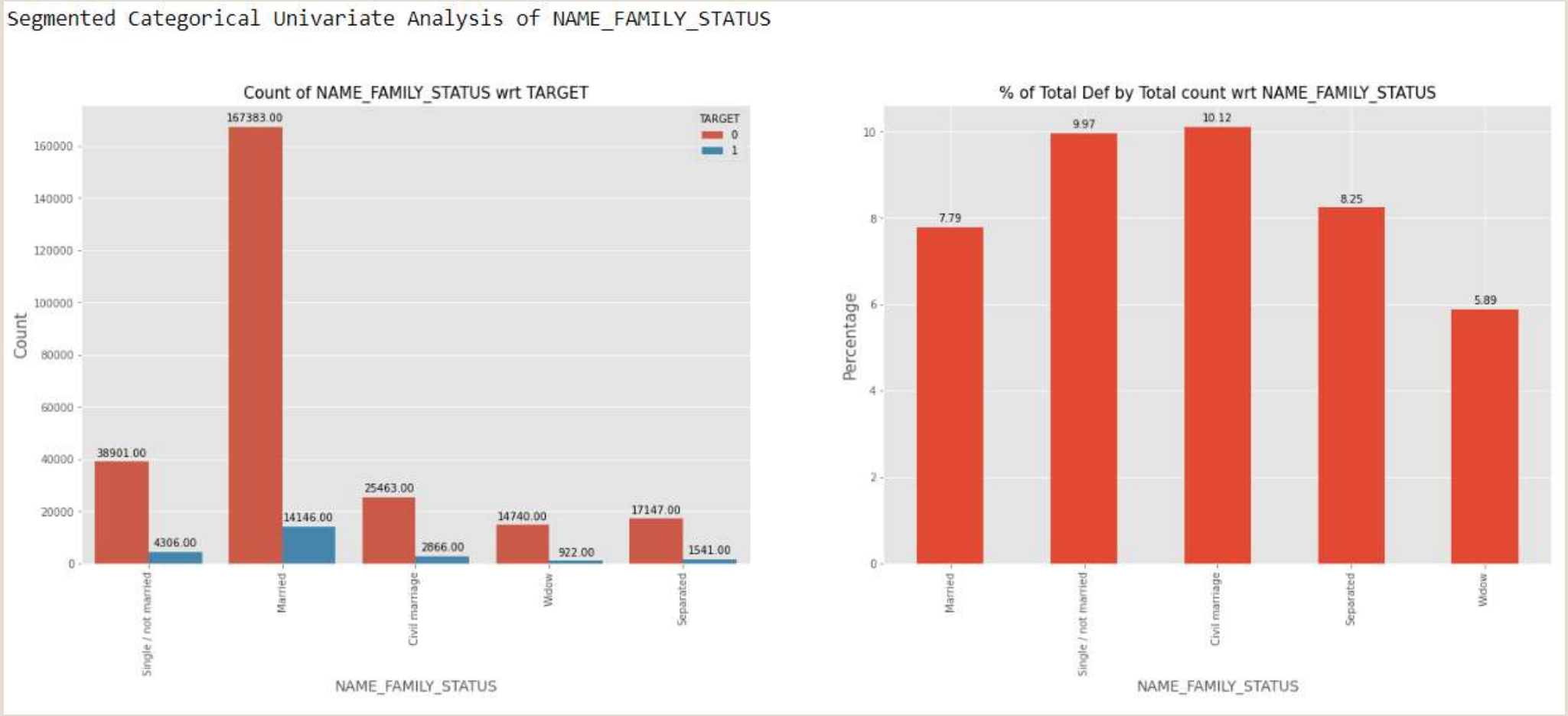
Inference:

- In the NAME_INCOME_TYPE, Others are the high defaulter. This others contain ['Unemployed', 'Student', 'Businessman', 'Maternity leave']. This is followed by Working and Commercial Associate as high defaulters. Pensioners and State servant are the least defaulters.



Inference:

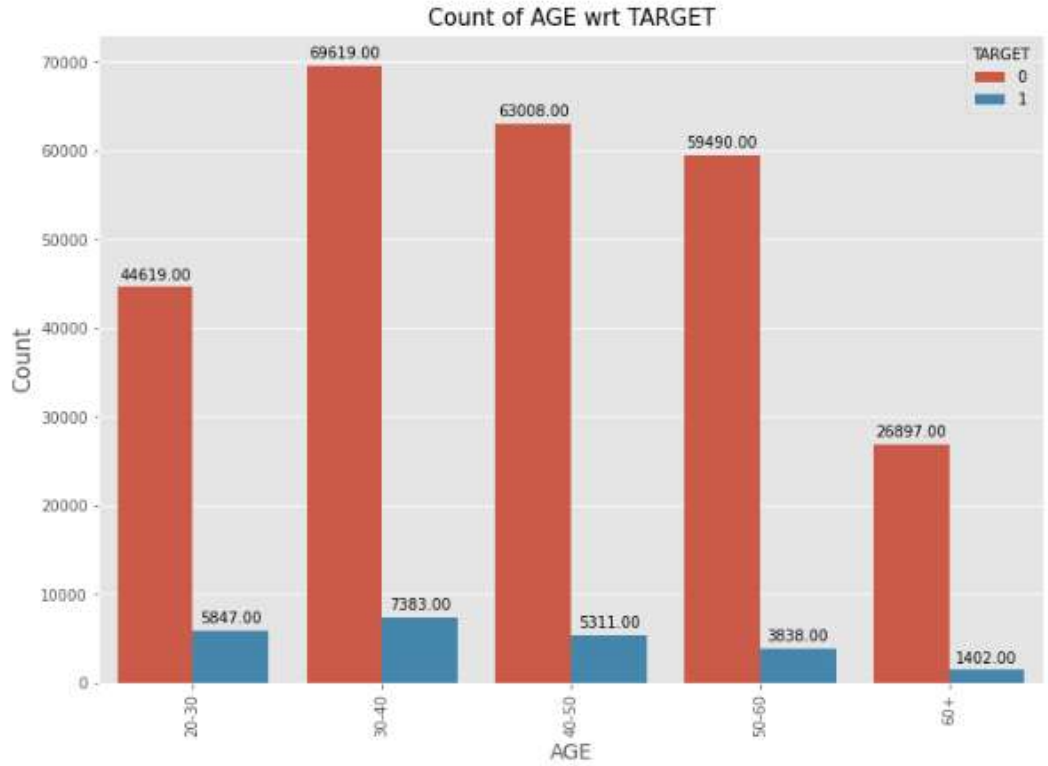
- Under NAME_EDUCATION_TYPE we see that Lower Secondary and Seconadary special are the group of people who are high defaulters. People with Academic degree amount to the least amount of defaulters.

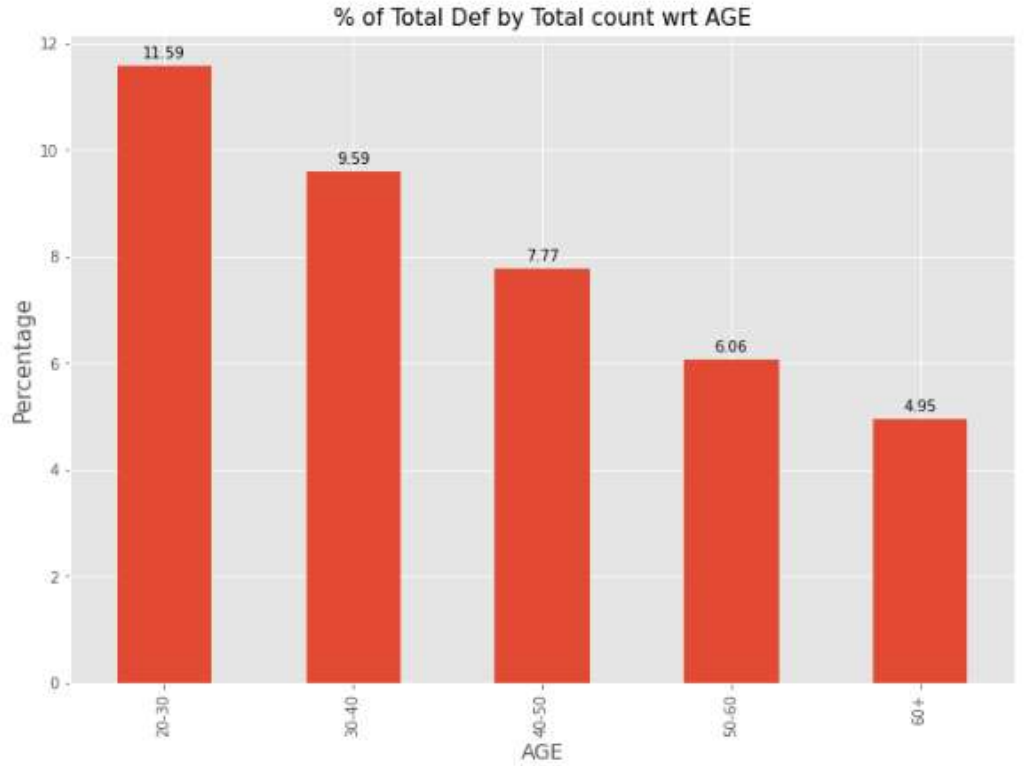


Inference:

- For NAME_FAMILY_STATUS we see that Civil Marriage and Single/ not married have higher defaulting rates. Widows are the least amount of defaulters.

Segmented Categorical Univariate Analysis of AGE

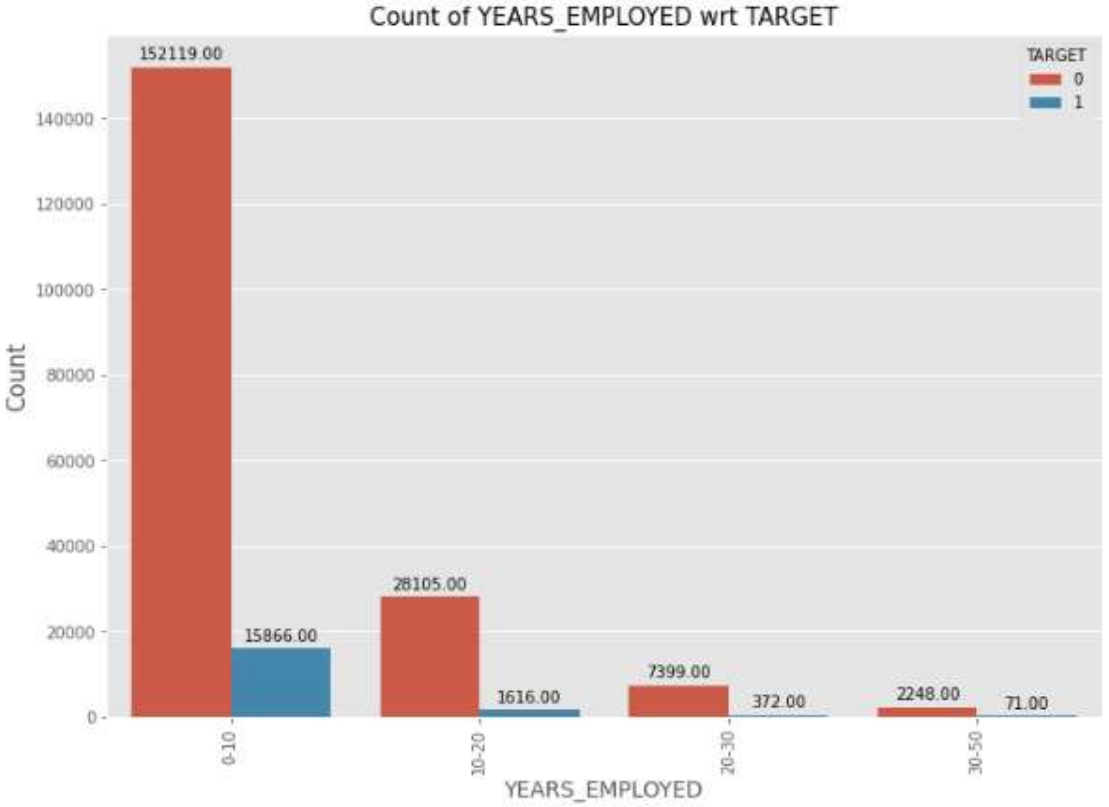


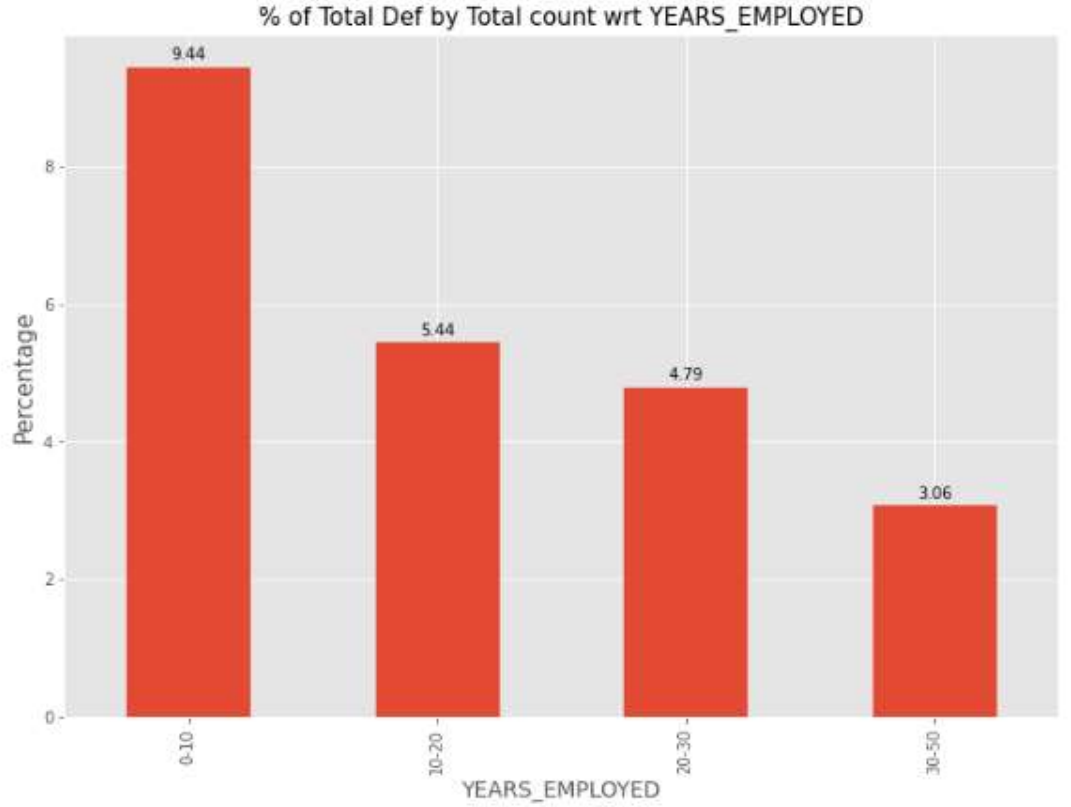


Inference:

- For AGE column there is a clear trend which shows that young people have difficulties to pay back the loans compared to the old people. Therefore the rate of defaulting decreases as the age increases.

Segmented Categorical Univariate Analysis of YEARS_EMPLOYED





Inference:

- For YEARS_EMPLOYED column there is a clear trend which shows that people with less experience have difficulties to pay back the loans compared to the highly experienced people. Therefore the rate of defaulting decreases as the amount of experience increases. This could be possible as people who are more experienced could be paid higher compared to fresher.

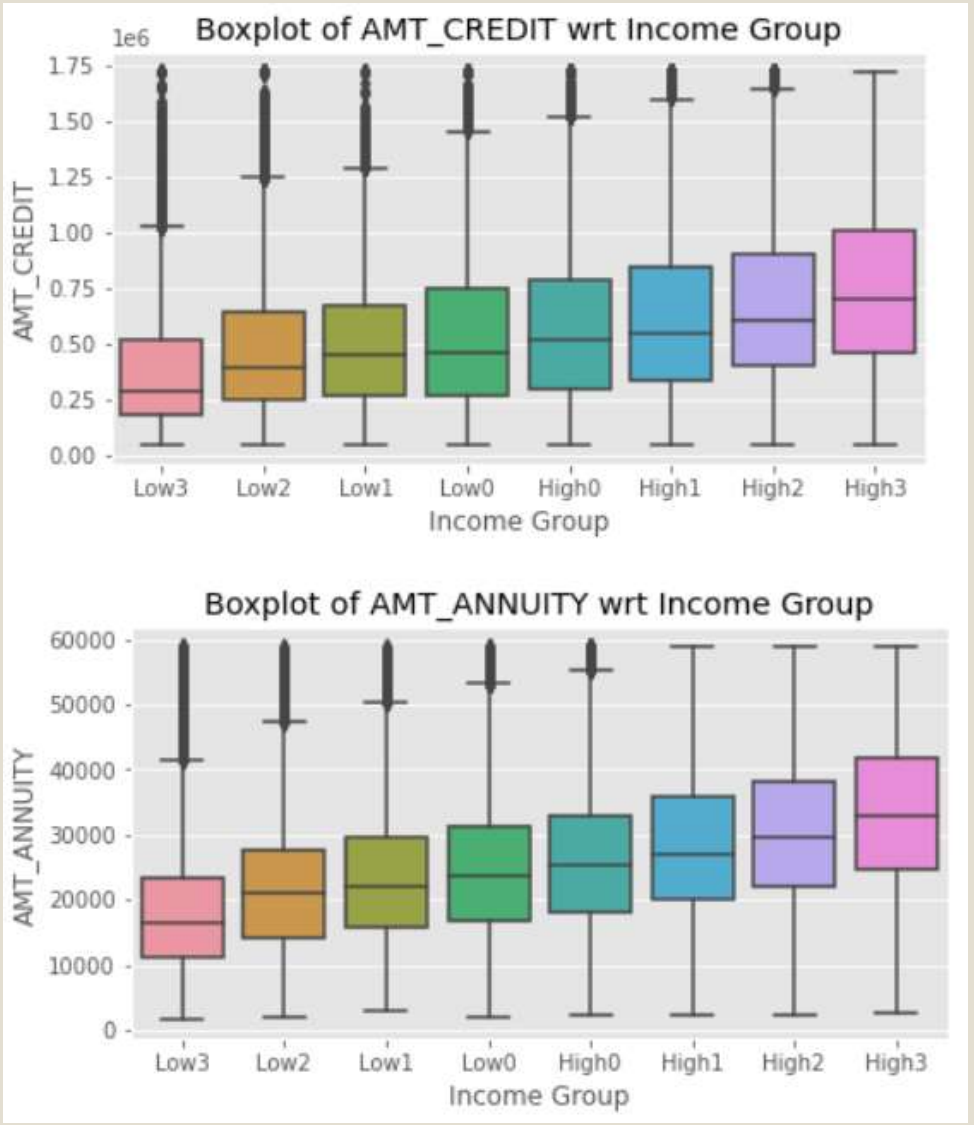
List of all the Inference made on these Categorical columns

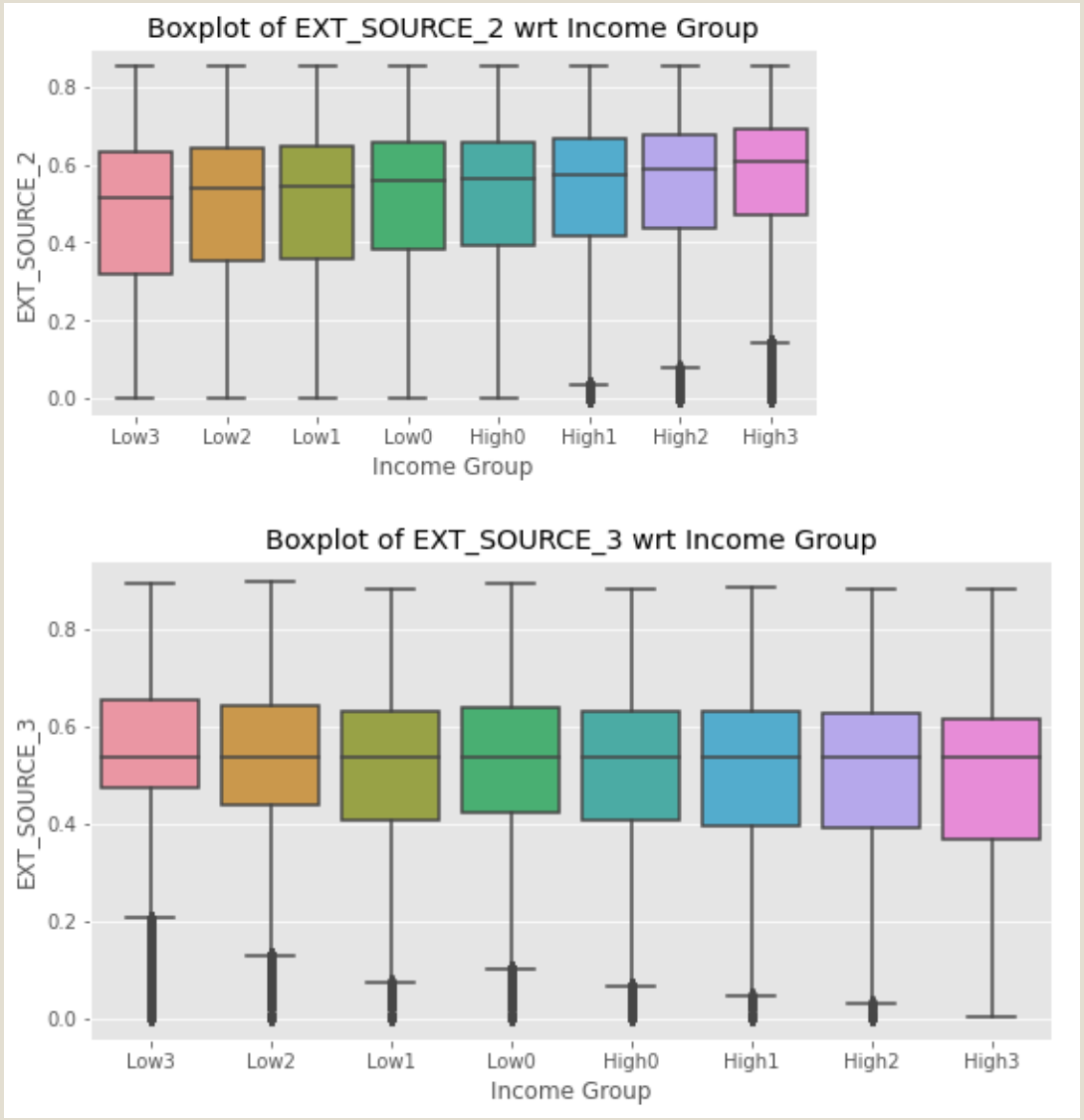
- Male are high Defaulters compared to Female
- As the number of children increase, the percentage of defaulters also increases.
- In the NAME_INCOME_TYPE, Others are the high defaulter. This others contain ['Unemployed', 'Student', 'Businessman', 'Maternity leave']. This is followed by Working and Commercial Associate as high defaulters. Pensioners and State servant are the least defaulters.
- Under NAME_EDUCATION_TYPE we see that Lower Secondary and Secondary special are the group of people who are high defaulters. People with Academic degree amount to the least amount of defaulters.
- For NAME_FAMILY_STATUS we see that Civil Marriage and Single/ not married have higher defaulting rates. Widows are the least amount of defaulters.
- For NAME_HOUSING_TYPE we see that people living in rented apartments or people living with their parents have higher defaulting rates. Whereas people living in Office Apartments and House have lesser defaulting rates.
- For REGION_POPULATION_RELATIVE there is a clear trend which shows that as the population of the place increase, the defaulting percentage decrease. This could be because since there are more people there is a possibility to have more jobs available and more opportunity to earn money.
- For AGE column there is a clear trend which shows that young people have difficulties to pay back the loans compared to the old people. Therefore the rate of defaulting decreases as the age increases.

List of all the Inference made on these Categorical columns

- For YEARS_EMPLOYED column there is a clear trend which shows that people with less experience have difficulties to pay back the loans compared to the highly experienced people. Therefore the rate of defaulting decreases as the amount of experience increases. This could be possible as people who are more experienced could be paid higher compared to fresher.
- YEARS_REGISTRATION also follows the trend of AGE and YEARS_EMPLOYED.
- For OWN_CAR_AGE we see that people who have very old cars, that is over 12 years have higher defaulting rates compared to people who have cars of age 0 to 9. We also see that people who don't own cars also have higher defaulting rates.
- For OCCUPATION_TYPE we see that people who are low skilled like- Low skilled labours, Drivers, Waitress have higher defaulting percentage. Whereas people who are in jobs that require high educations like- Accountants, HR staff, IT staff, Managers have the least defaulting percentage.
- From Total_documents we see that people who have submitted 3 documents to have high defaulting percentage.(Can't really tell the reason for it).
- From Income Group we see that people having the least amount of income are finding it hard to pay back compared to people who have highest income. It is expected, so need to be careful while giving loans to the people who have low income.
- AMT_CREDIT_RANGE follows trend of a normal distribution curve where the extremes that is 0-200000 and 800000 and above have the least defaulting percentage and the middle 400000-600000 has highest defaulting rates.

Plotting Boxplot for Income Groups against other Numerical columns





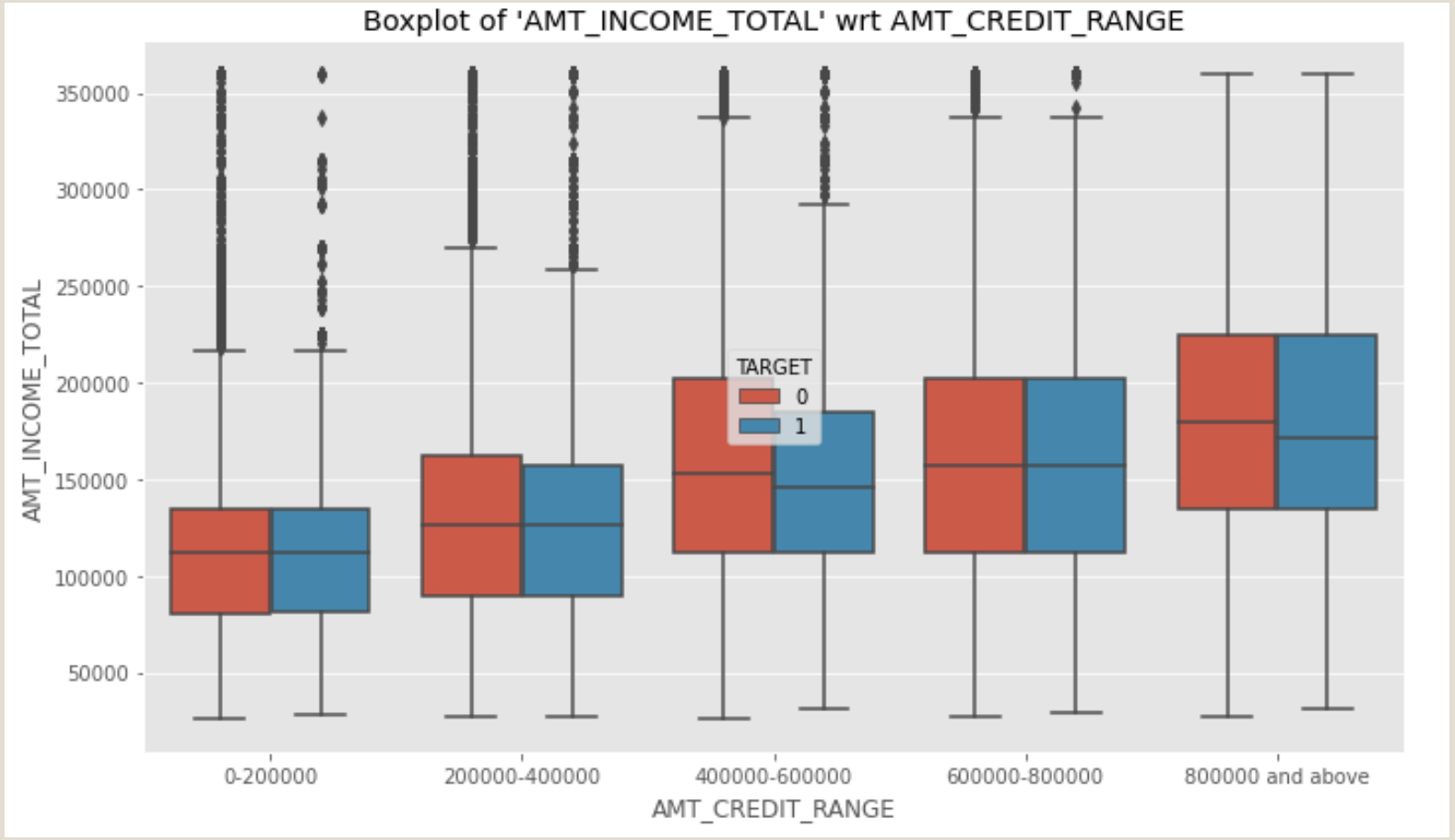
Inference

- As people start earning more, people take higher amount of credit.
- As people start earning more, people pay higher amount of annuity. This is because Credit and Annuity have high corelation.
- As people start earning more, people take credit for higher amount of goods price. This is because Credit and goods price have very high corelation as expected.
- As Income increses, EXT_SOURCE_2 also increases slightly.
- As Income increses, EXT_SOURCE_3 decreases slightly. That is, they are inversely proportional.

Plotting a box plot for 'AMT_INCOME_TOTAL' wrt 'AMT_CREDIT_RANGE' to understand why defaulters are struggling to pay back the loan.

Inference:

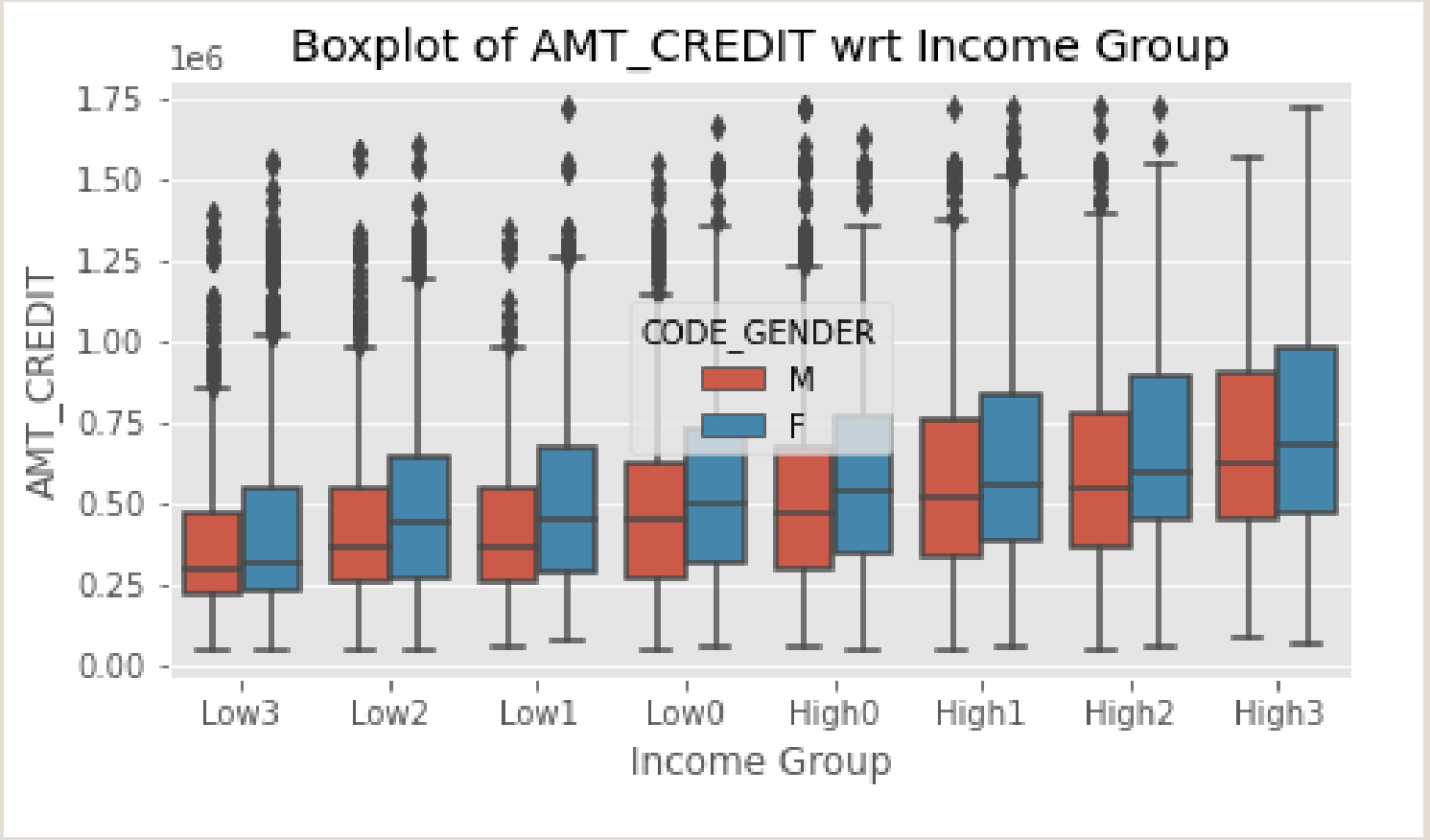
- From the given plot we can clearly understand that from the high defaulting range that is 200000-400000 and 400000-600000, the defaulters have less amount of income compared to non-defaulters and they have taken loans which are slightly above their range to pay it back on time. Therefore they are defaulting. To stop this we can increase the interest rates for people having lesser income but wanting higher credit. This way we can make sure that people who really believe they can pay on time will take the loan at those rates and also bank can minimize the loss in case of defaulting.



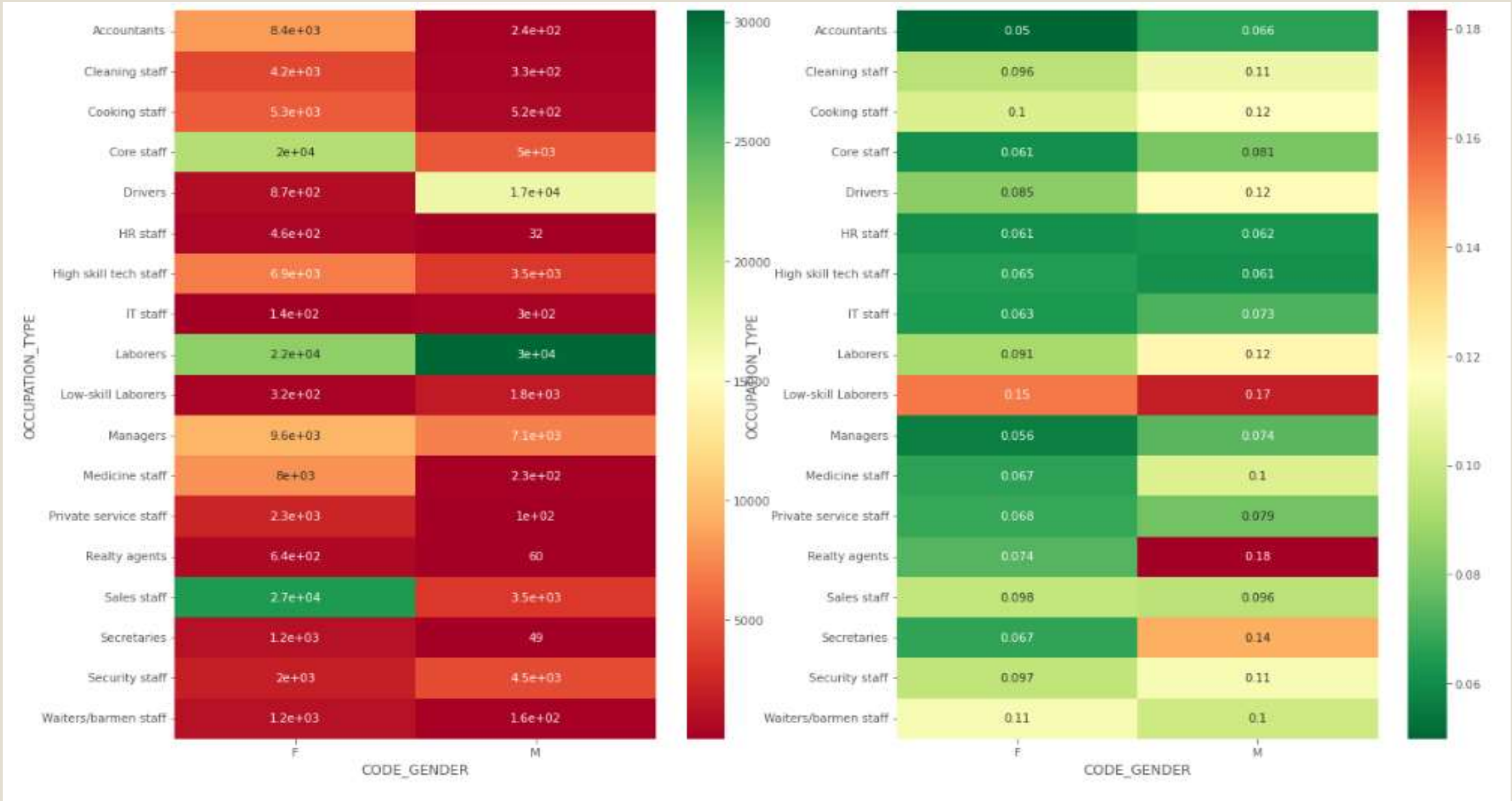
Plotting a box plot for 'AMT_CREDIT' wrt 'Income Group' wrt 'CODE_GENDER'

Inference:

- From this plots we see that Men in spite of taking lower credit compared to women for same income range are struggling to pay back the loan on time.

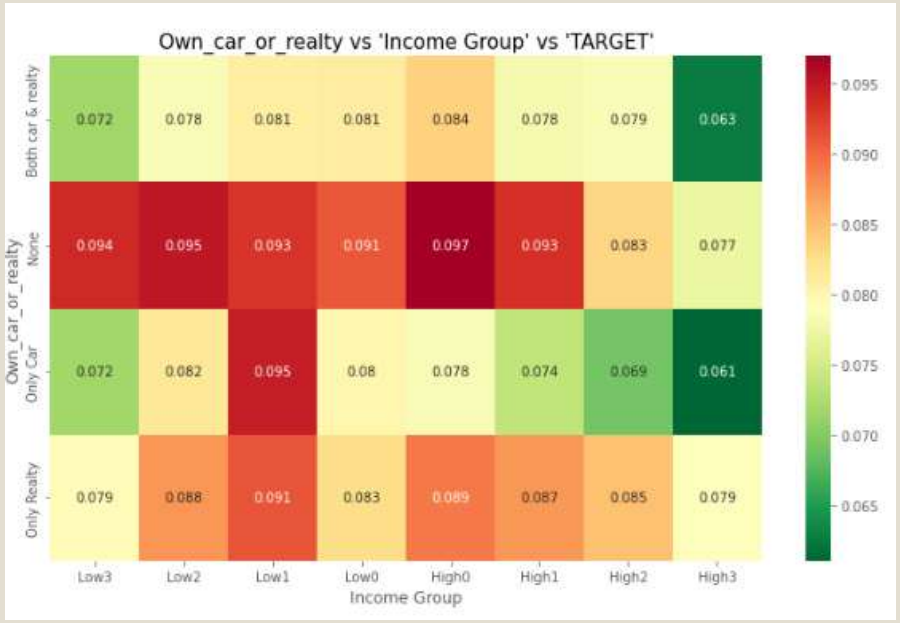


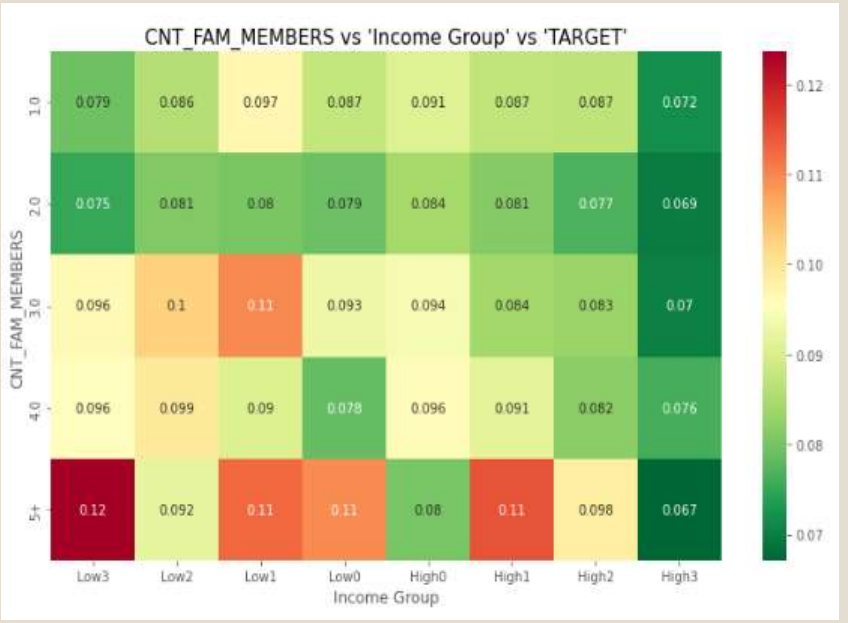
Plotting Heatmap to understand why Men are higher Defaulters than Women



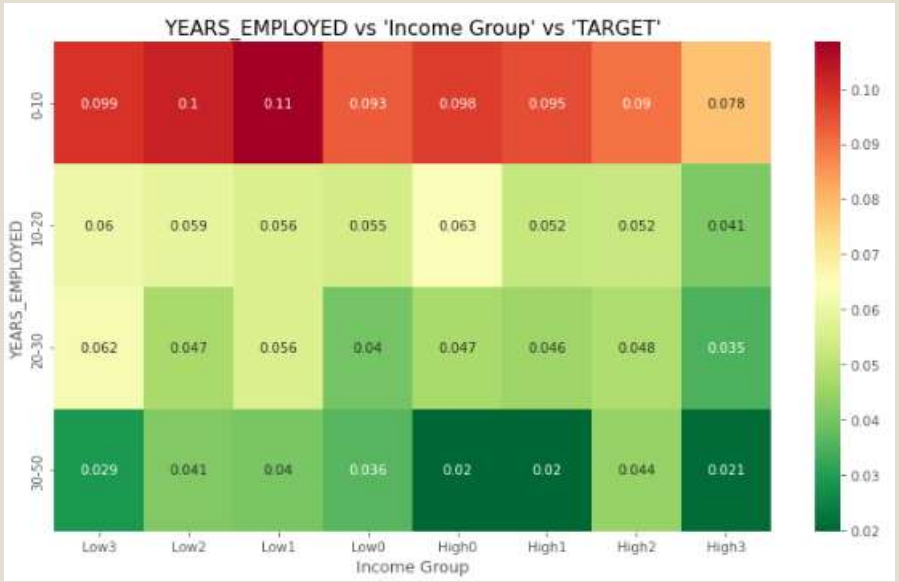
Inference:

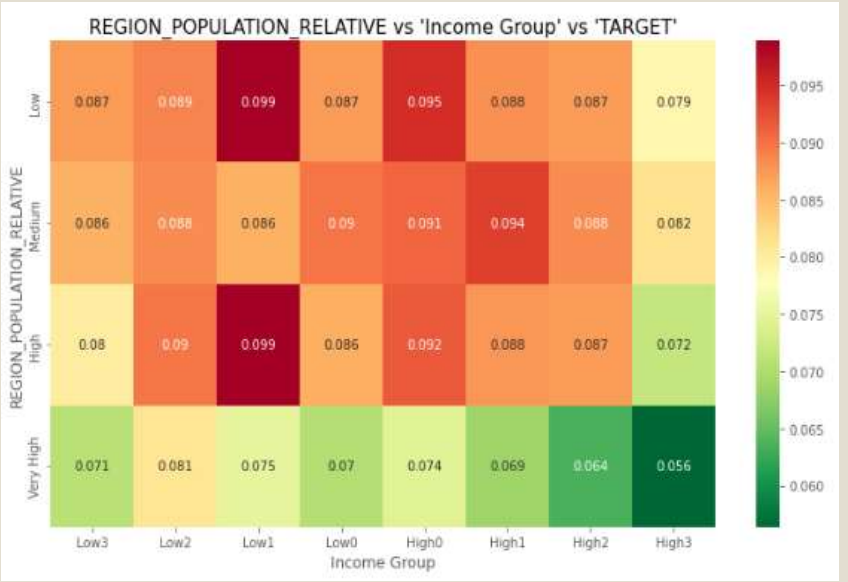
- As we can see here, there are more men in the job roles like "Labourers, Low-skill Labourers and Driver" categories compared to women. These categories make up the highest default making categories. Therefore, men have more default rates than women.
- We also see that the count of women in high paying jobs like "Accountants, HR Staff, Managers, Medicine Staff" are higher than Men in these categories.
- Most loans were provided to Men Laborers who are high defaulters. Thus, increasing the default rate for Men.
- Banks need to be careful while providing loans to these high defaulting categories of people.





Few Examples of Heatmap plotted to find the relation between Categorical columns vs 'Income Group' vs 'TARGET'



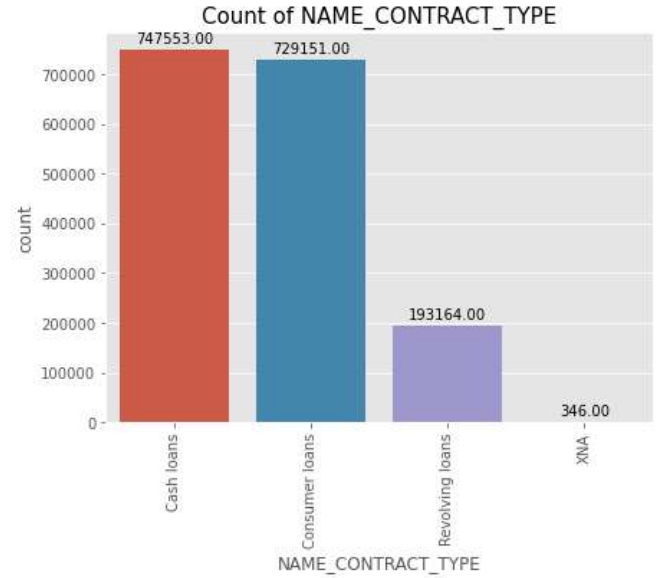


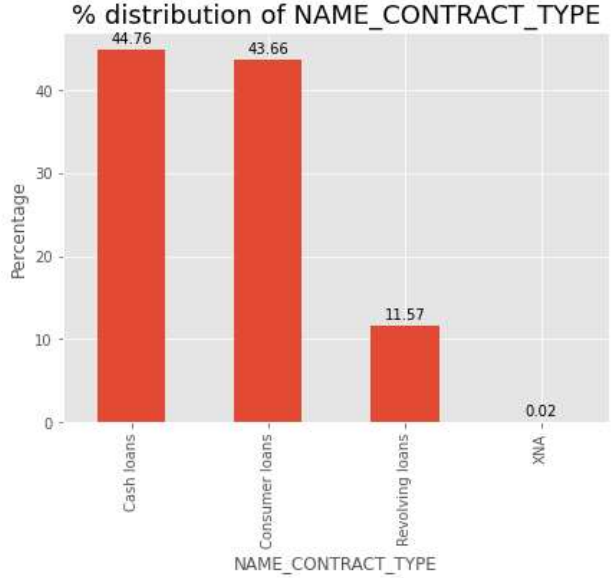
Inference made after Plotting Heatmap to find the relation between Categorical columns vs 'Income Group' vs 'TARGET'

- From heatmap of CODE_GENDER, we observe Men are high defaulters in spite of having income up to the range High1
- From heatmap of CNT_CHILDREN, we observe that as the number of children increase, the rate of defaulting also increases.
- From heatmap of NAME_INCOME_TYPE, we learn that category of Others have less income and are high defaulters.
- From heatmap of NAME_EDUCATION_TYPE, we observe that People with lower education in spite of earning well tend to default.
- From heatmap of NAME_FAMILY_STATUS, we can see that Widow are the safest category to provide loans as the rate of default is very less even for people having low income. Also people belonging to the category of Civil marriage and Single are high defaulters.
- From heatmap of NAME_HOUSING_TYPE, we see that people who live with their parents and living in Rented apartment have high default rates.
- From heatmap of REGION_POPULATION_RELATIVE, we observe that default rate decreases with increase in population of city. Therefore they are inversely proportional.
- From heatmap of AGE, we can see that people ranging from 20-40 have higher rate of default. Whereas people who are above 50 are the best range of people to provide loans as their rate of default is very low.
- From heatmap of YEARS_EMPLOYED, similar to AGE people with less experience tend to default more than people with higher experience.
- From heatmap of OWN_CAR_AGE, we observe that people who have very old cars, that is above 12 years of age and have income in the range of low2 to high2 default a lot compared to others.
- From heatmap of OCCUPATION_TYPE, we see that Low-Skill labourers are high defaulters. This is true to people who are earning very well as well. People with high paying jobs that have high education requirements tend to default very less.
- From heatmap of CNT_FAM_MEMBERS, we see that people with family of more than or equal to 3 and salary less that Low0 tend to default a lot. So, it is better to provide loans to people less than 3 members in their family and for people who have 3 or more members but with Income greater than High1.
- From heatmap of Own_car_or_realty, we see that people with No car or Realty and Only realty tend to have higher defaulting rates.

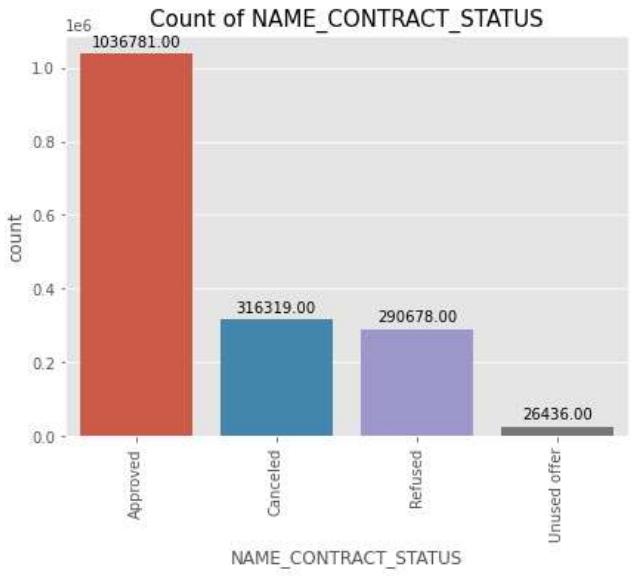
Univariate Analysis for Previous_Application Dataset

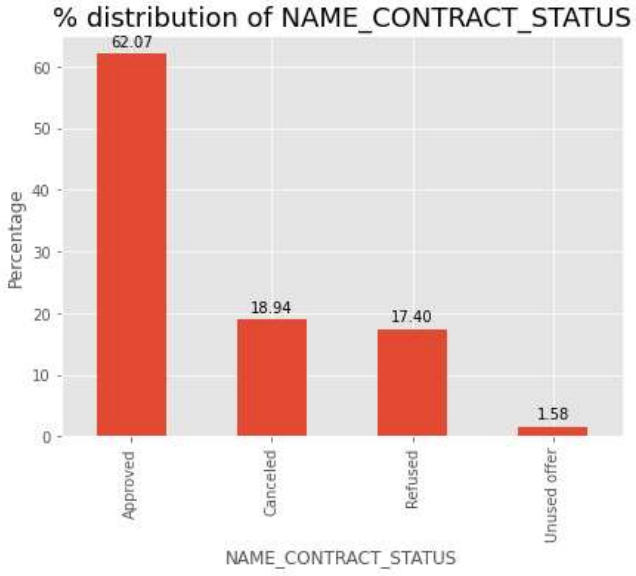
Categorical Univariate Analysis of NAME_CONTRACT_TYPE



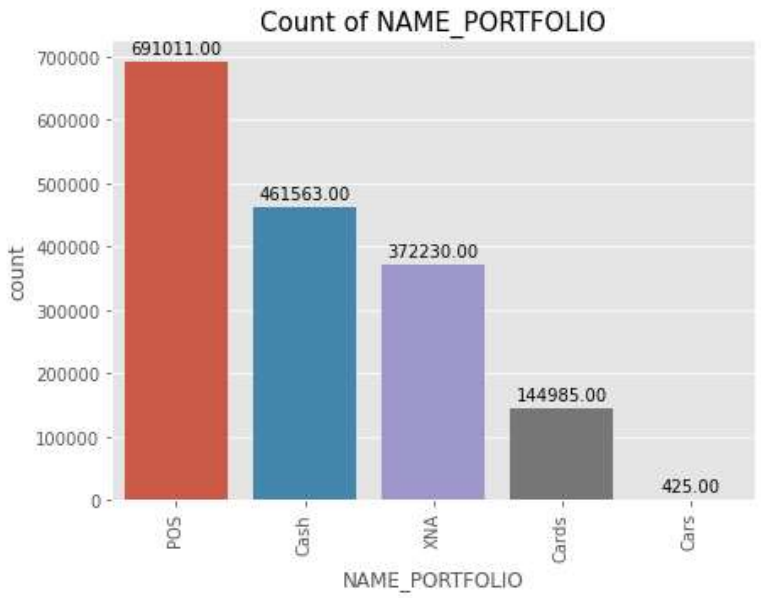


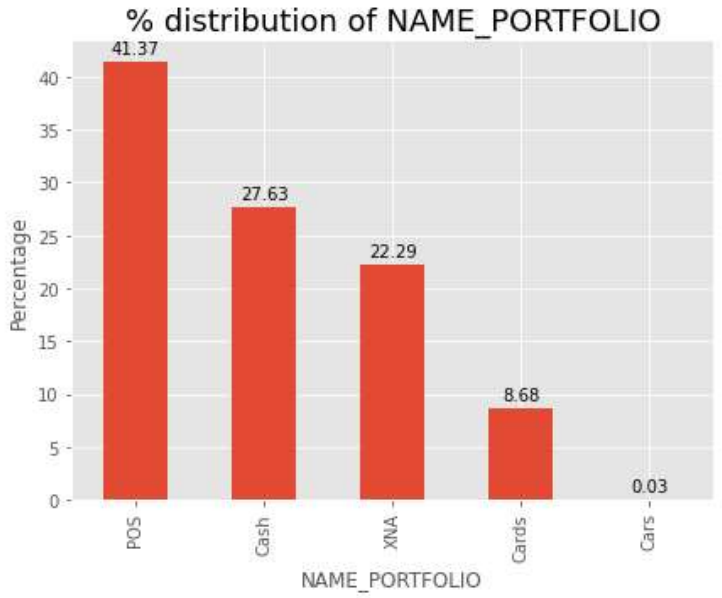
Categorical Univariate Analysis of NAME_CONTRACT_STATUS



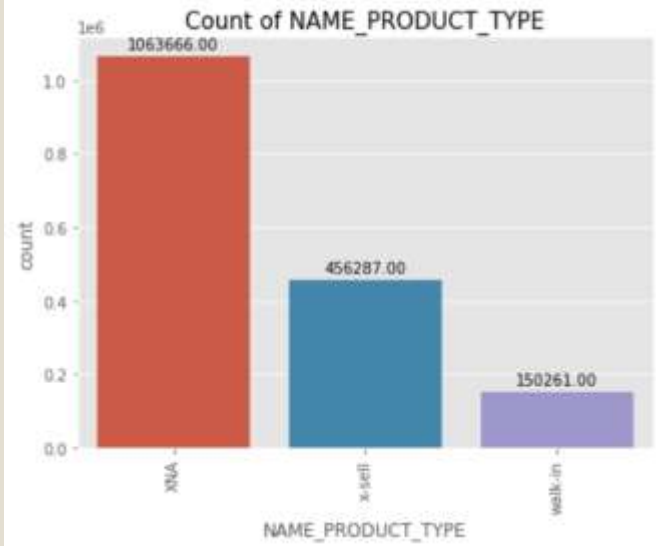


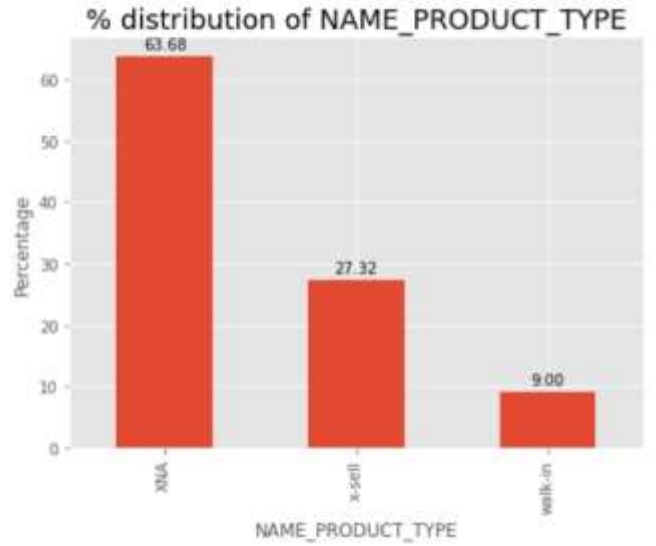
Categorical Univariate Analysis of NAME_PORTFOLIO





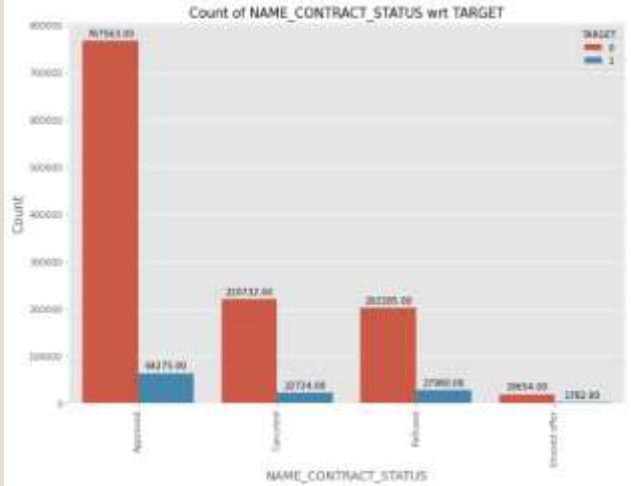
Categorical Univariate Analysis of NAME_PRODUCT_TYPE

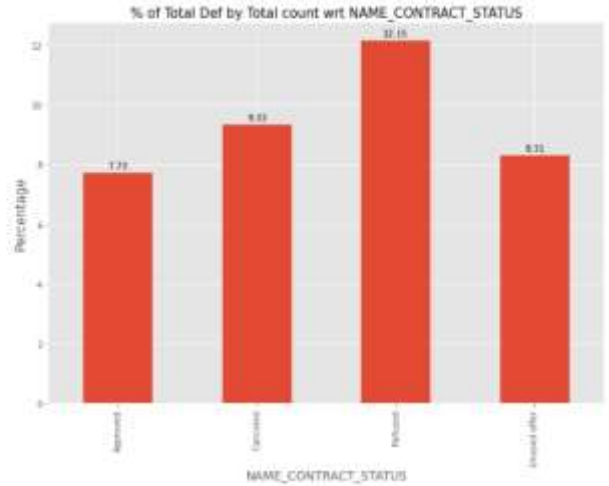




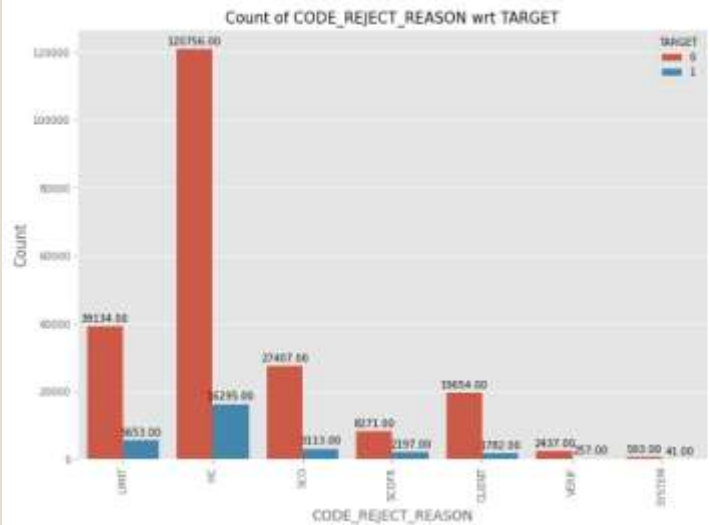
Bivariate Analysis for Merged Dataset

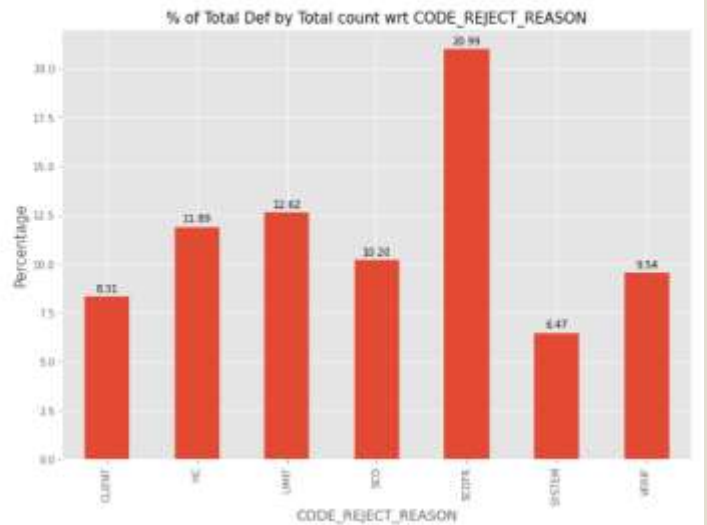
Segmented Categorical Univariate Analysis of NAME_CONTRACT_STATUS



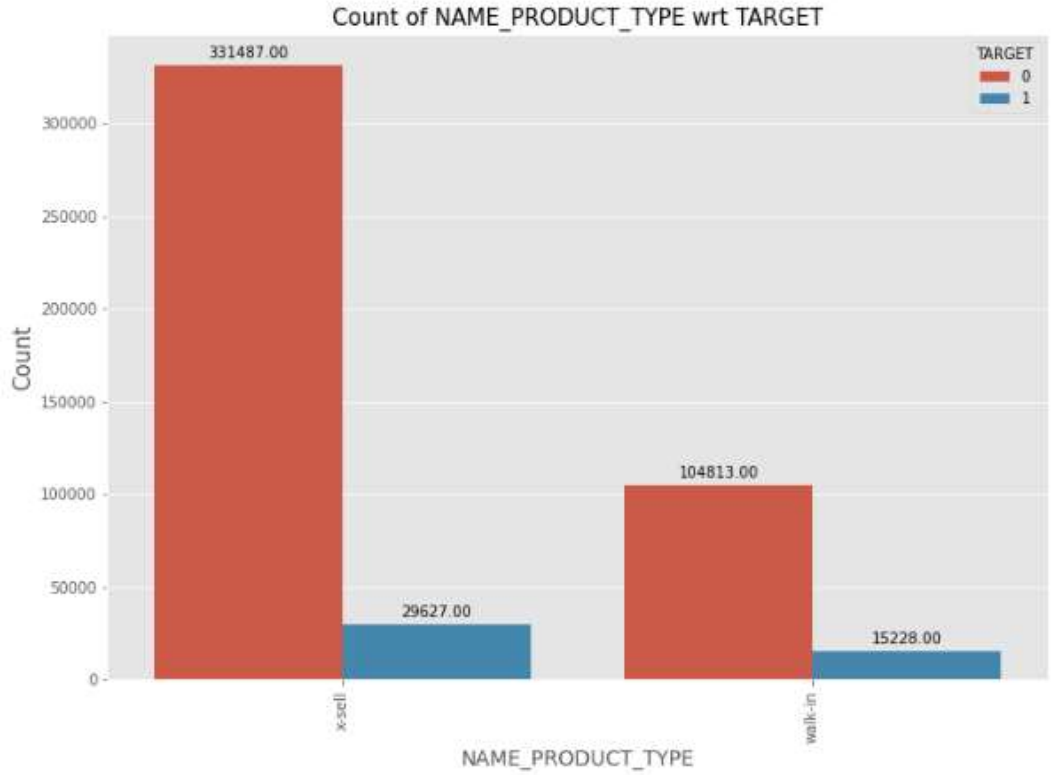


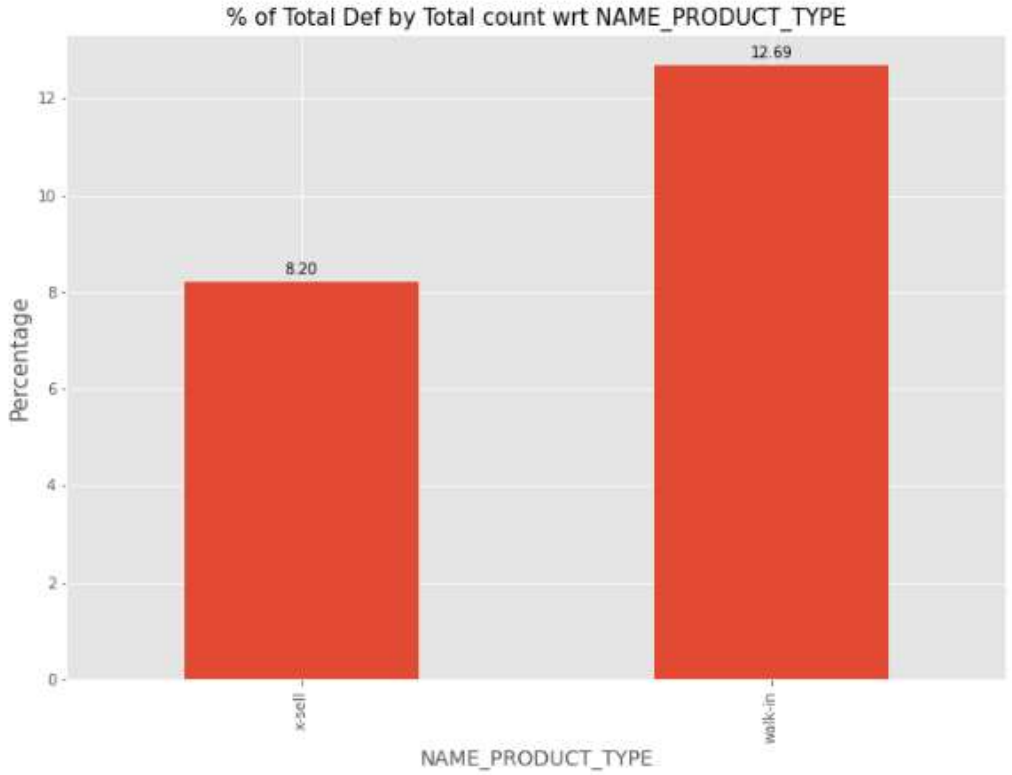
Segmented Categorical Univariate Analysis of CODE_REJECT_REASON



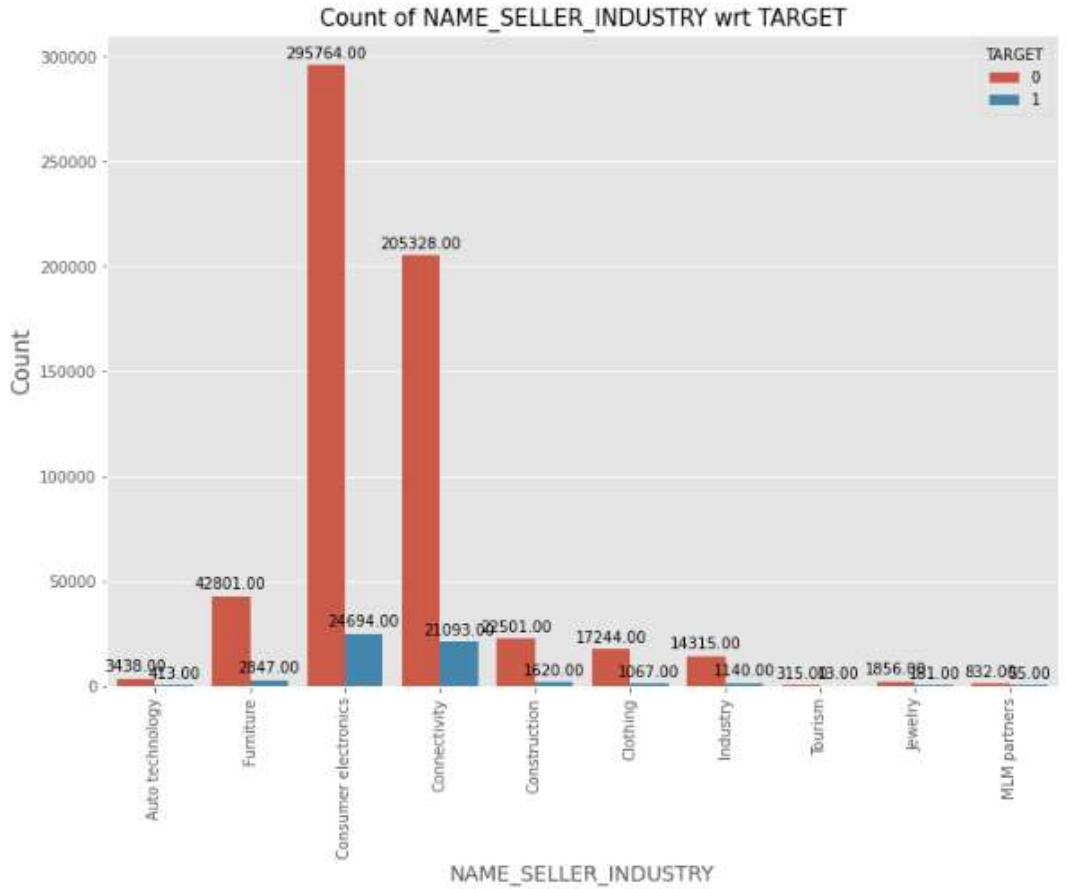


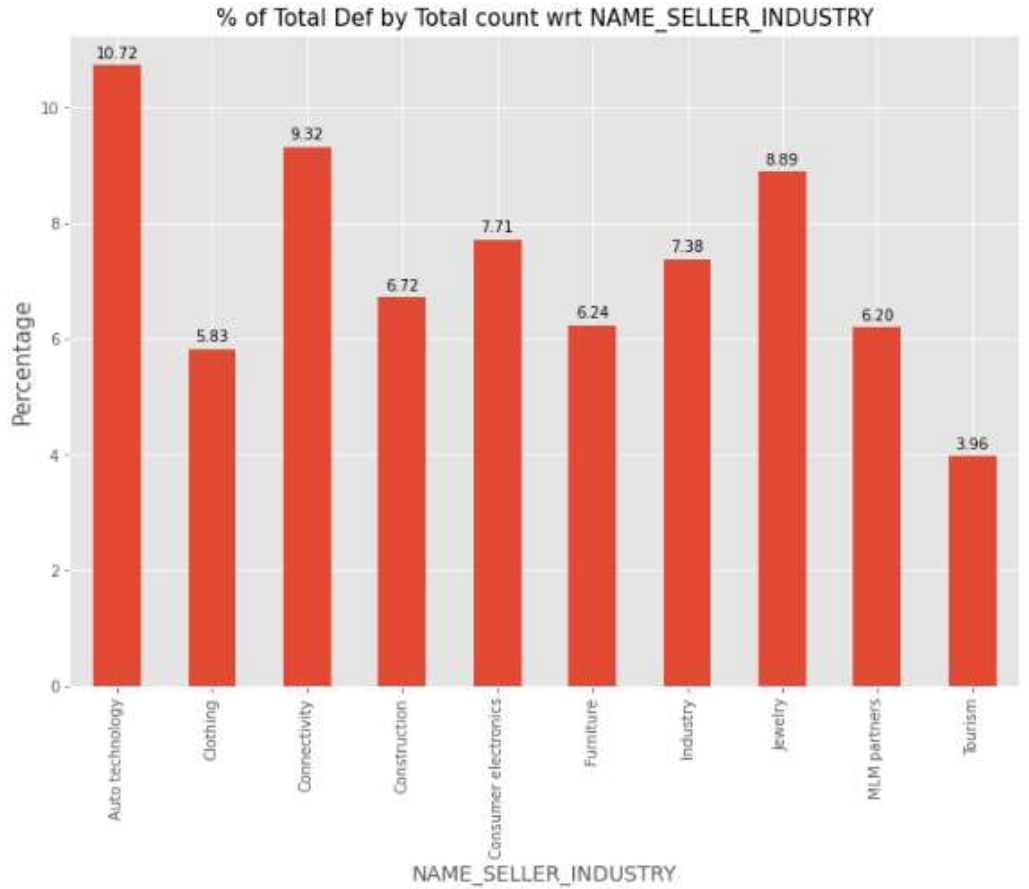
Segmented Categorical Univariate Analysis of NAME_PRODUCT_TYPE



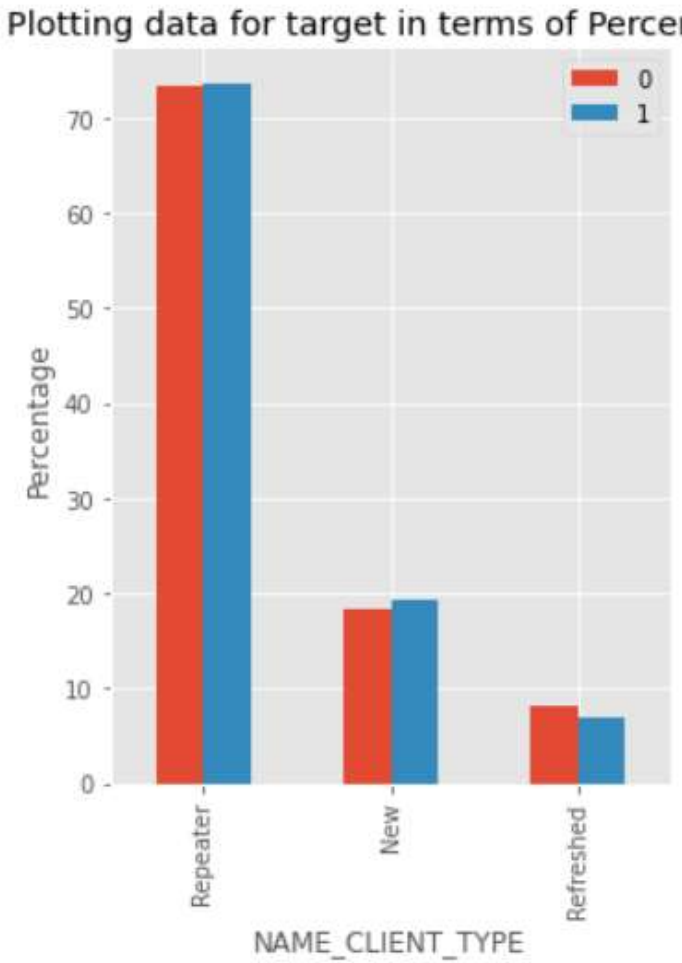


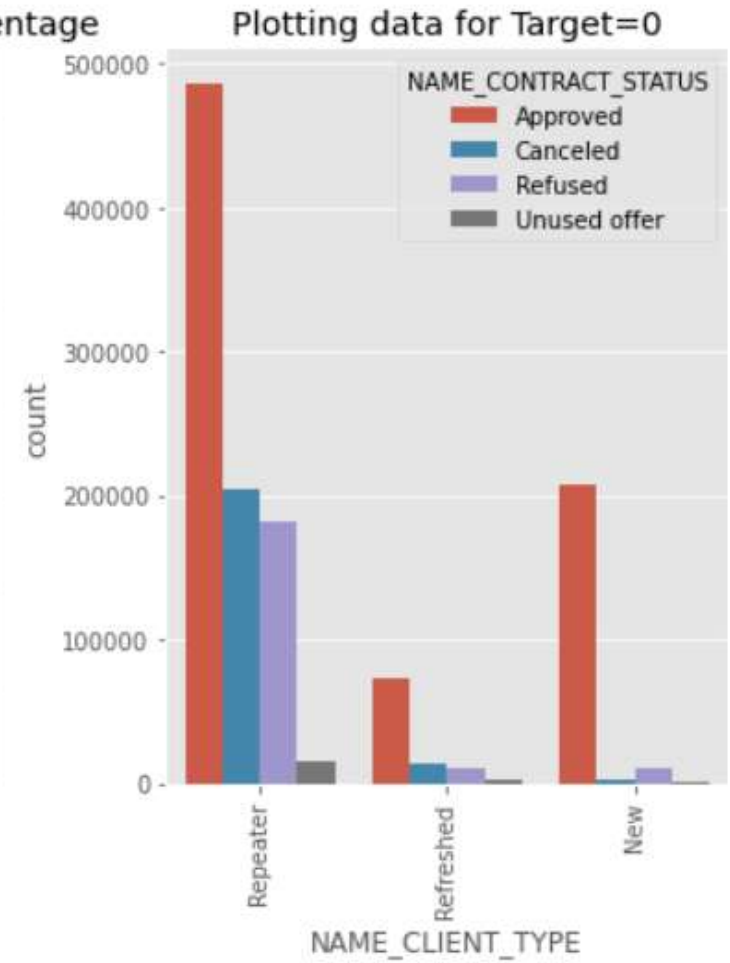
Segmented Categorical Univariate Analysis of NAME_SELLER_INDUSTRY

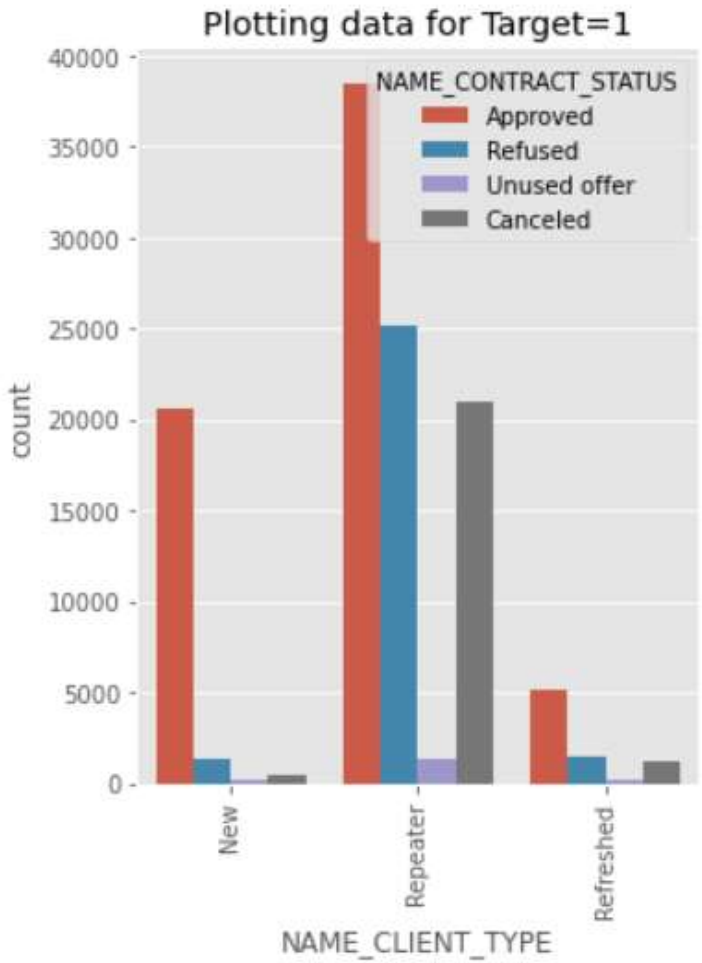




Plotting for NAME_CLIENT_TYPE having hue = NAME_CONTRACT_STATUS

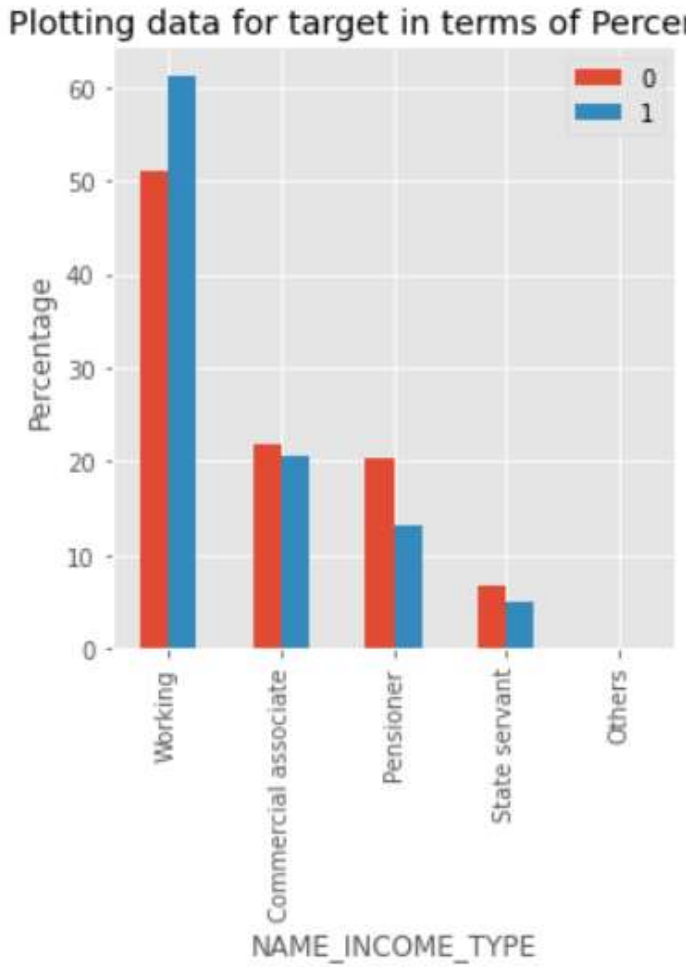


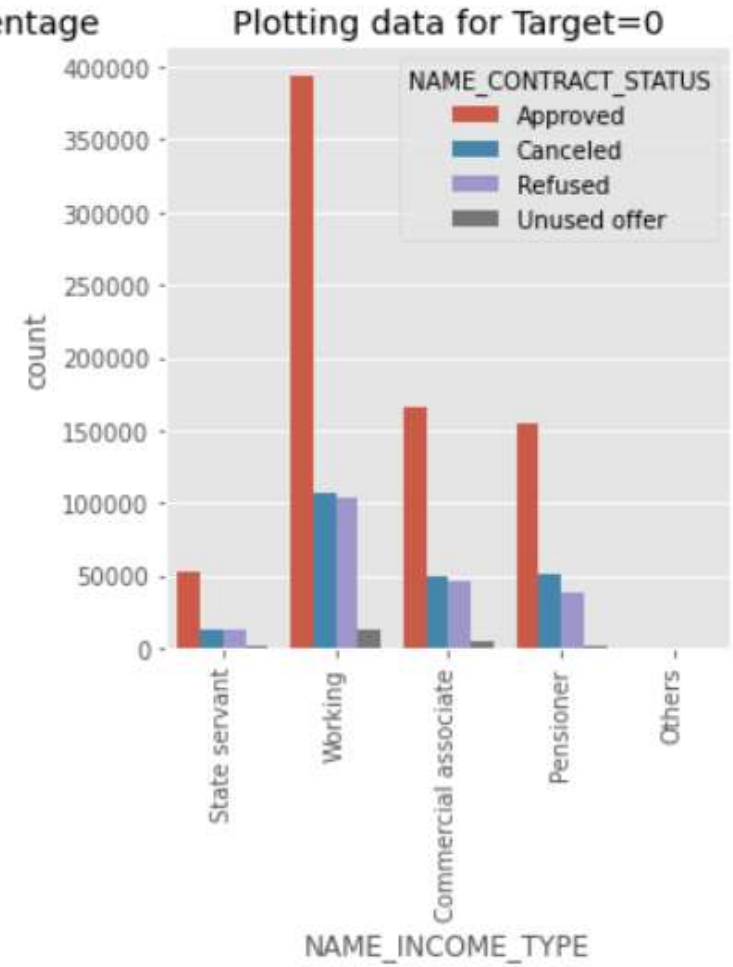


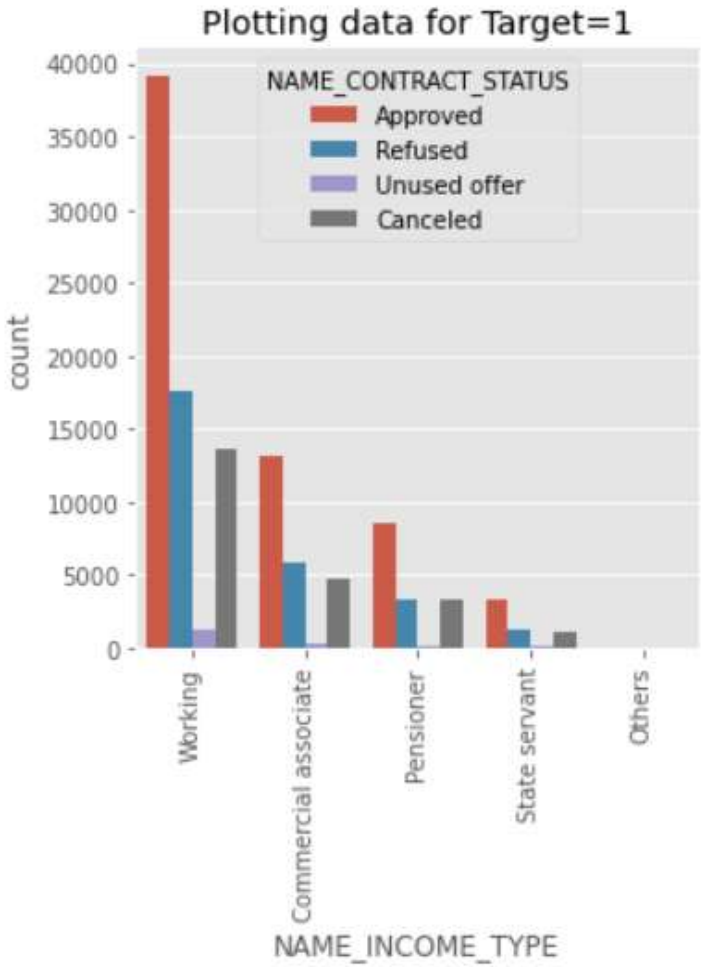


Plotting for CODE_GENDER having hue = NAME_CONTRACT_STATUS

Plotting for NAME_INCOME_TYPE having hue = NAME_CONTRACT_STATUS



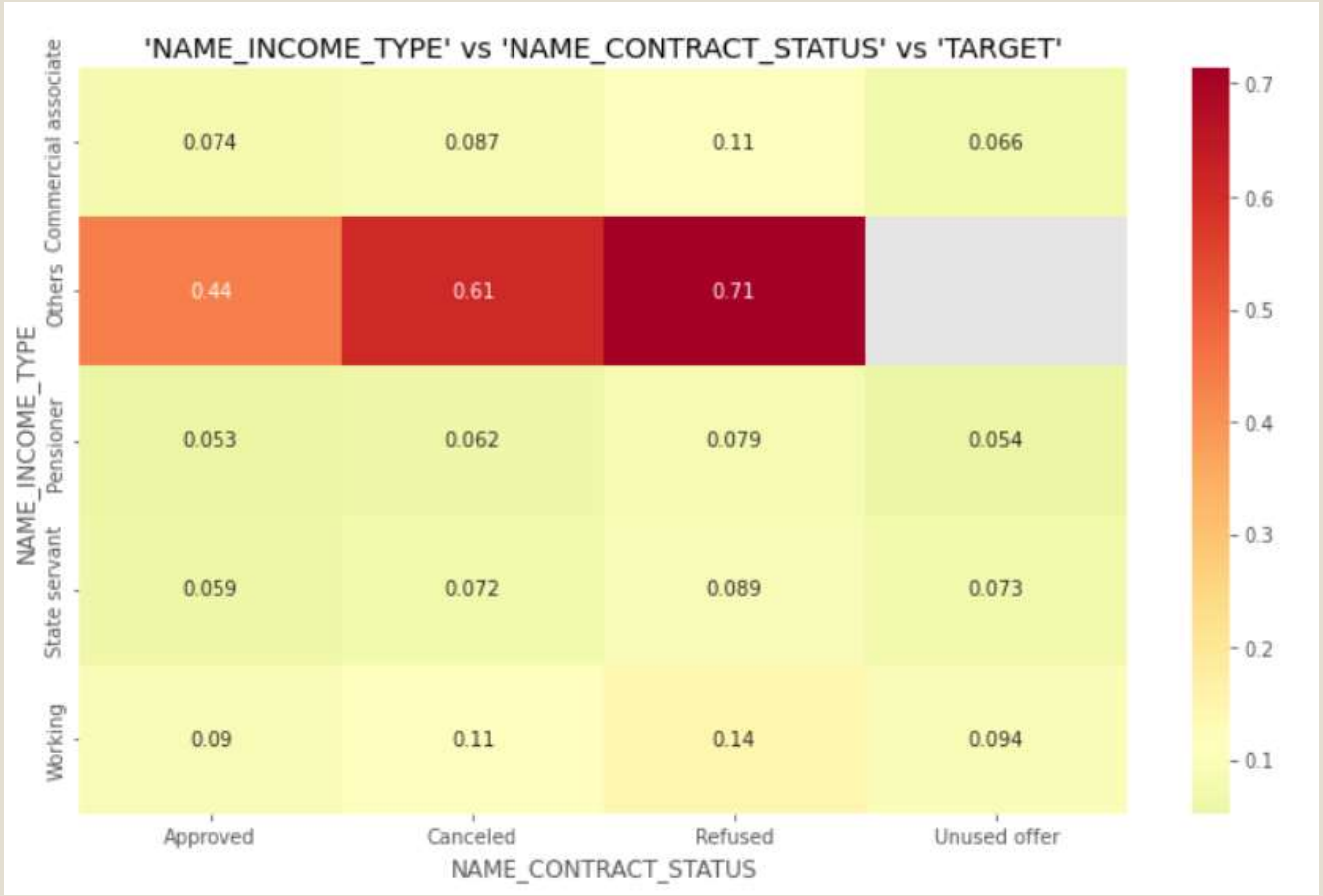


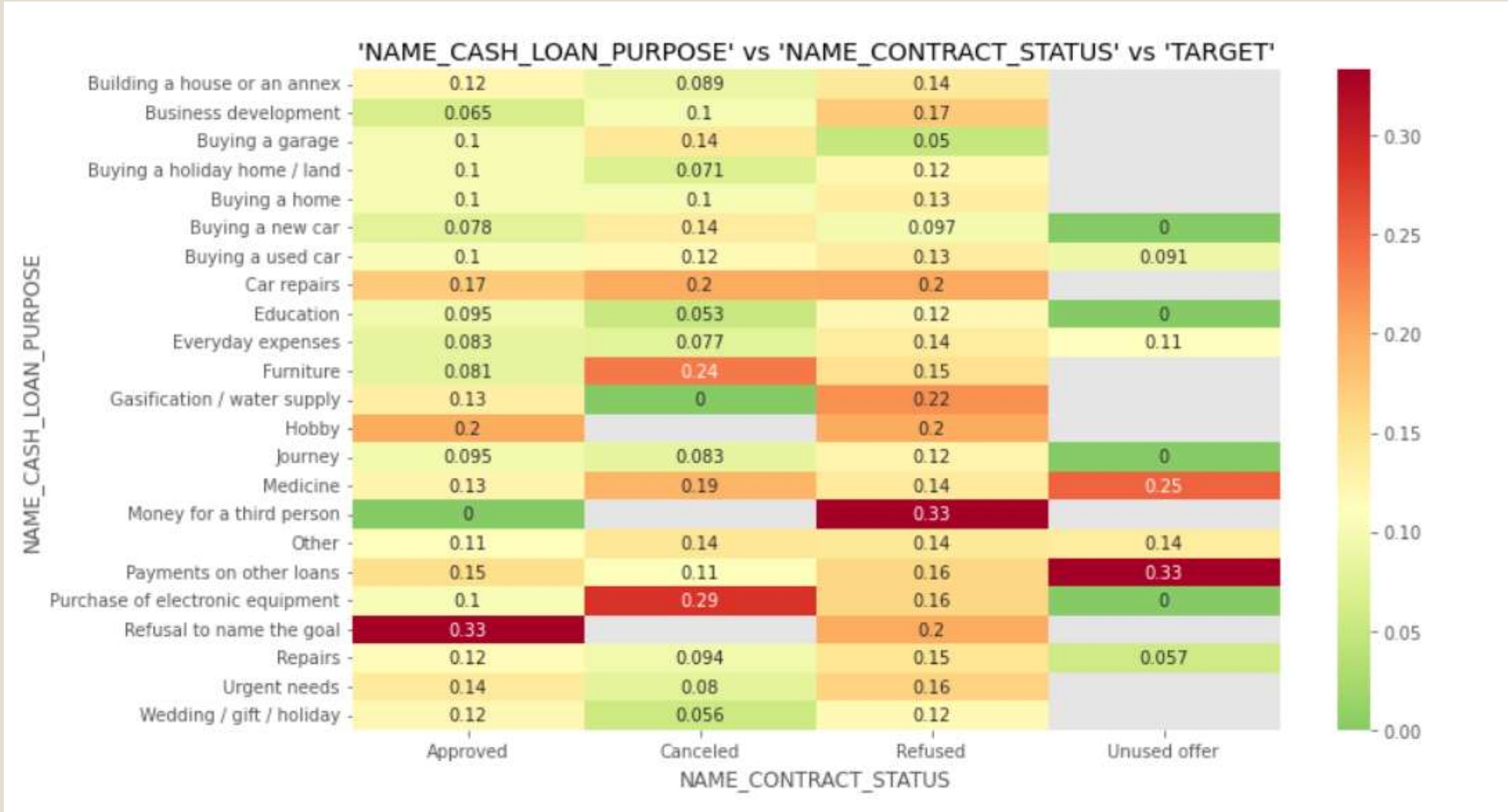


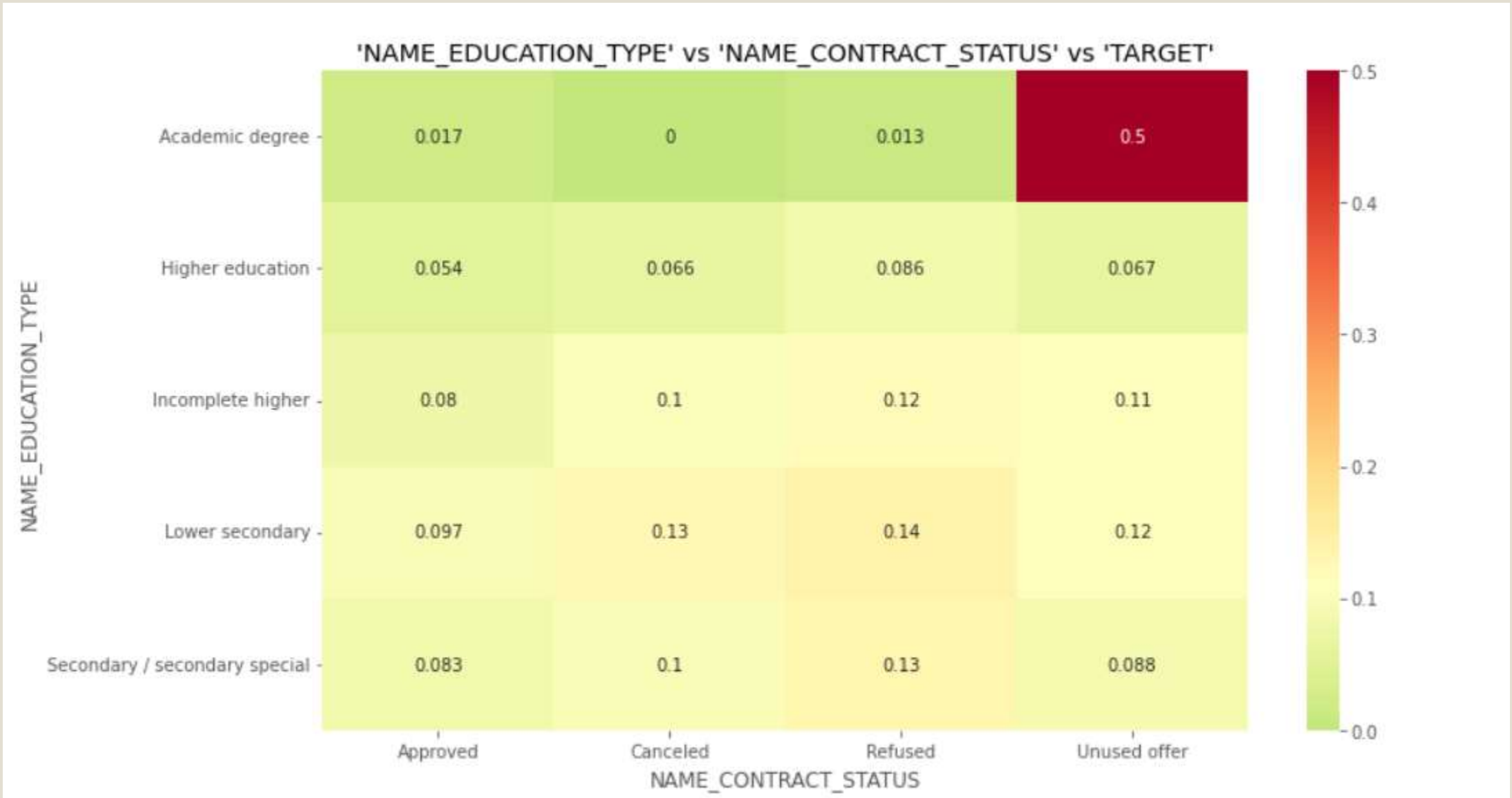
Multivariate Analysis

Plotting heatmap to see the relationship between few key categorical columns vs 'NAME_CONTRACT_STATUS' vs 'TARGET'

Desired result is to have Green boxes over Approved and Red boxes over Refused. This indicates that people who are known high defaulters are being refused and low defaulters are being approved for loans




'NAME_CASH_LOAN_PURPOSE' vs 'NAME_CONTRACT_STATUS' vs 'TARGET'
NAME_CASH_LOAN_PURPOSE
Building a house or an annex 0.12 0.089 0.14
Business development 0.065 0.1 0.17
Buying a garage 0.1 0.14 0.05
Buying a holiday home / land 0.1 0.071 0.12
Buying a home 0.1 0.1 0.13
Buying a new car 0.078 0.14 0.097 0
Buying a used car 0.1 0.12 0.13 0.091
Car repairs 0.17 0.2 0.2
Education 0.095 0.053 0.12 0
Everyday expenses 0.083 0.077 0.14 0.11
Furniture 0.081 0.24 0.15
Gasification / water supply 0.13 0 0.22
Hobby 0.2 0.2
Journey 0.095 0.083 0.12 0
Medicine 0.13 0.19 0.14 0.25
Money for a third person 0 0.33
Other 0.11 0.14 0.14 0.14
Payments on other loans 0.15 0.11 0.16 0.33
Purchase of electronic equipment 0.1 0.29 0.16 0
Refusal to name the goal 0.33 0.2
Repairs 0.12 0.094 0.15 0.057
Urgent needs 0.14 0.08 0.16
Wedding / gift / holiday 0.12 0.056 0.12
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS
0.30
0.25
0.20
0.15
0.10
0.05
0.00


'NAME_EDUCATION_TYPE' vs 'NAME_CONTRACT_STATUS' vs 'TARGET'
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
NAME_EDUCATION_TYPE
0.017
0
0.013
0.5
0.054
0.066
0.086
0.067
0.08
0.1
0.12
0.11
0.097
0.13
0.14
0.12
0.083
0.1
0.13
0.088
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS
0.5
0.4
0.3
0.2
0.1
0.0

Inference:

- **From the heatmaps, we can identify the category of people in spite of having very less defaulting rates being rejected. These are few of missed business opportunity to the bank and we can also see in some cases people having higher default rates being approved. Therefore, to improve the business/profits banks should try to provide loans to people having lesser defaulting rates.**

Thank You.