# CREDIT EDA CASE STUDY

**Presenters:**

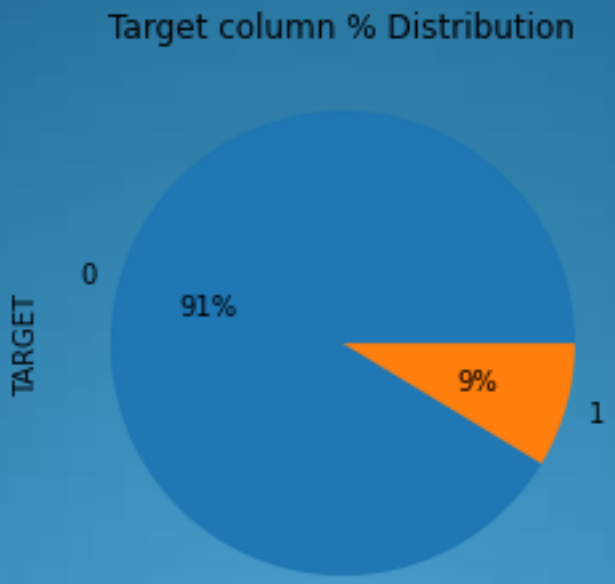
- Pradhan Nayak
- Vijay Kumar Singh

# Categorical Univariate analysis for overall population vs. Target 0 vs. Target 1 (Application Dataset)

# Target column % Distribution

Points to be concluded from the graph.

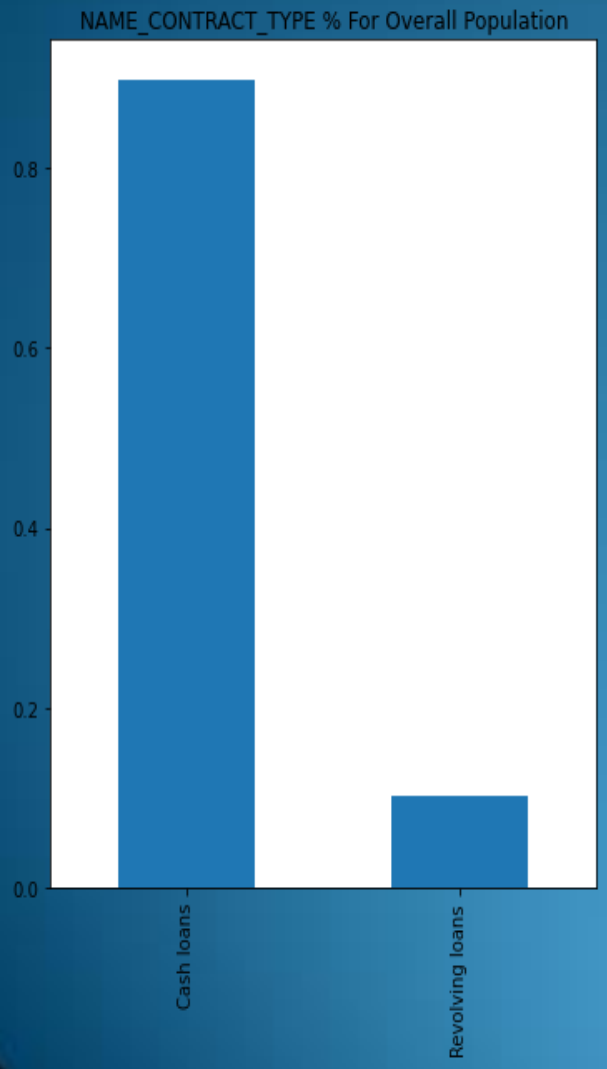
1. Of the overall population in application dataset, close to 91% applicants had no payment difficulties and remaining others had.

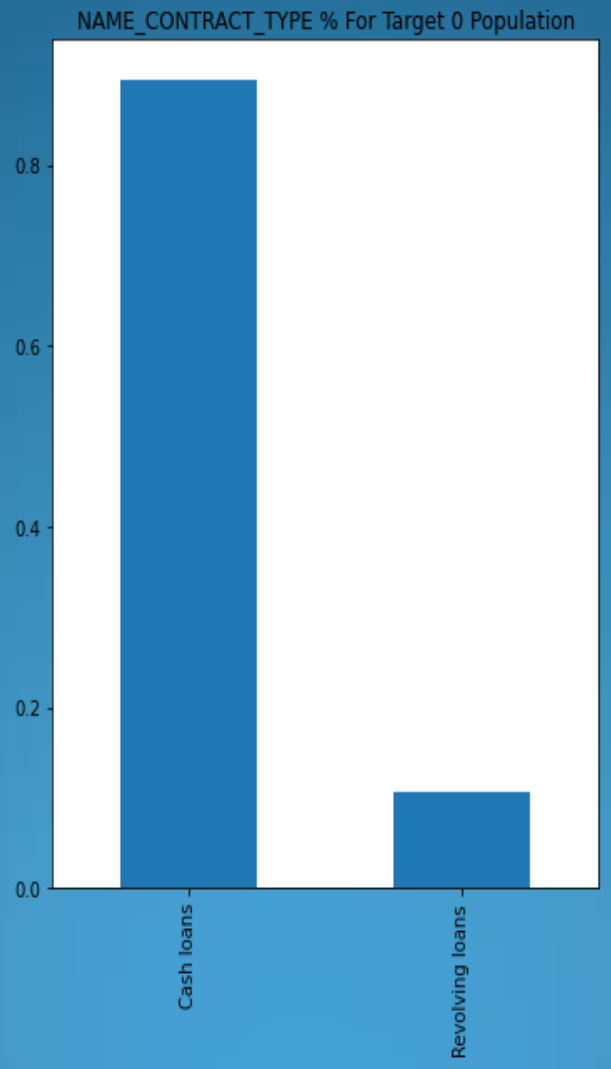


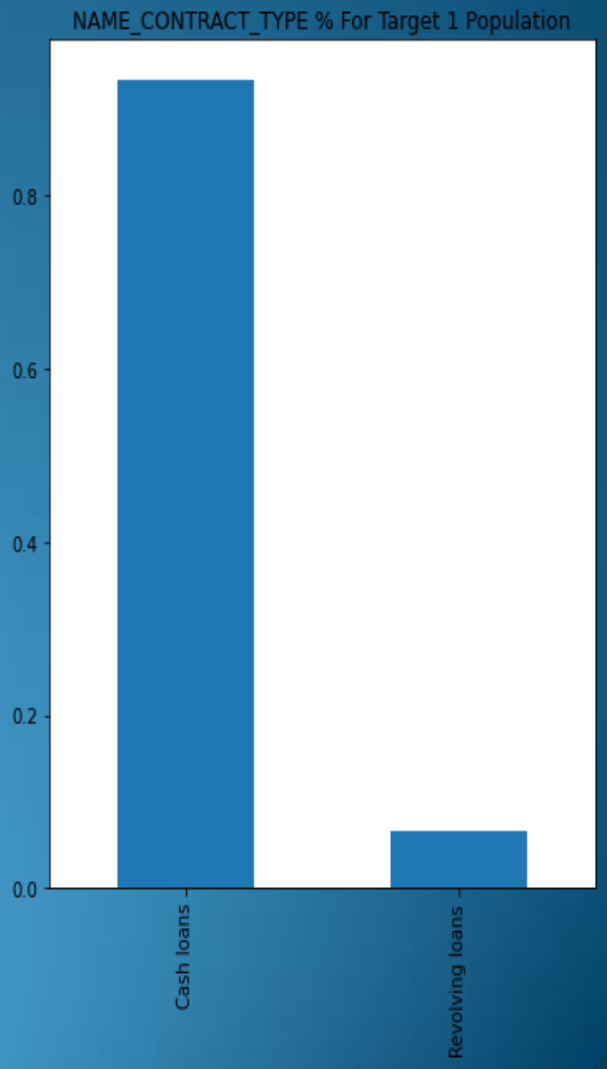
# Contract Type % Distribution

Points to be concluded from the graph.

1. Credit was given mainly for cash loan



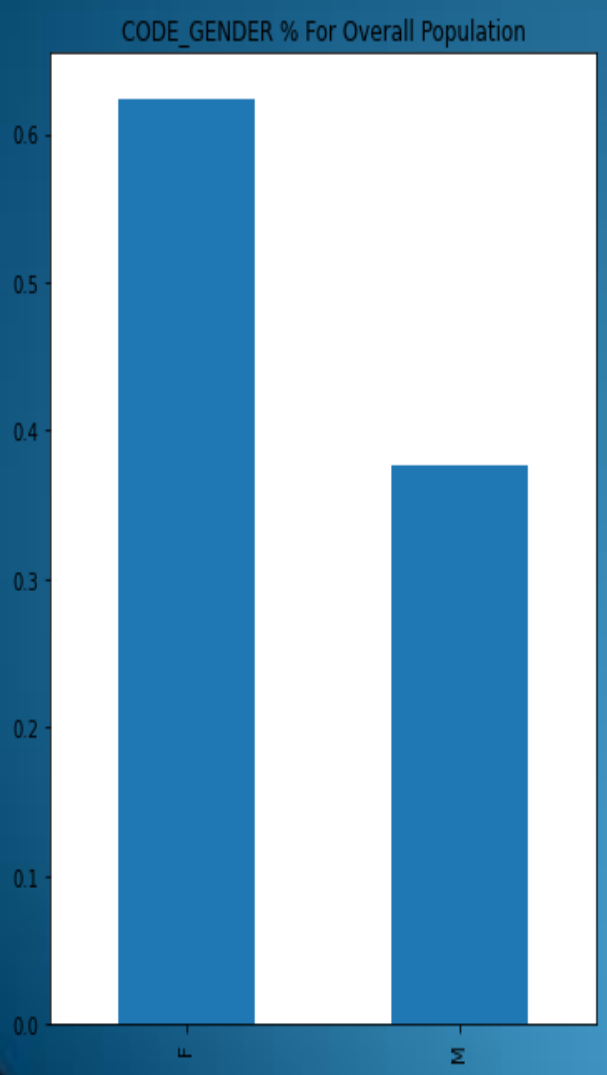



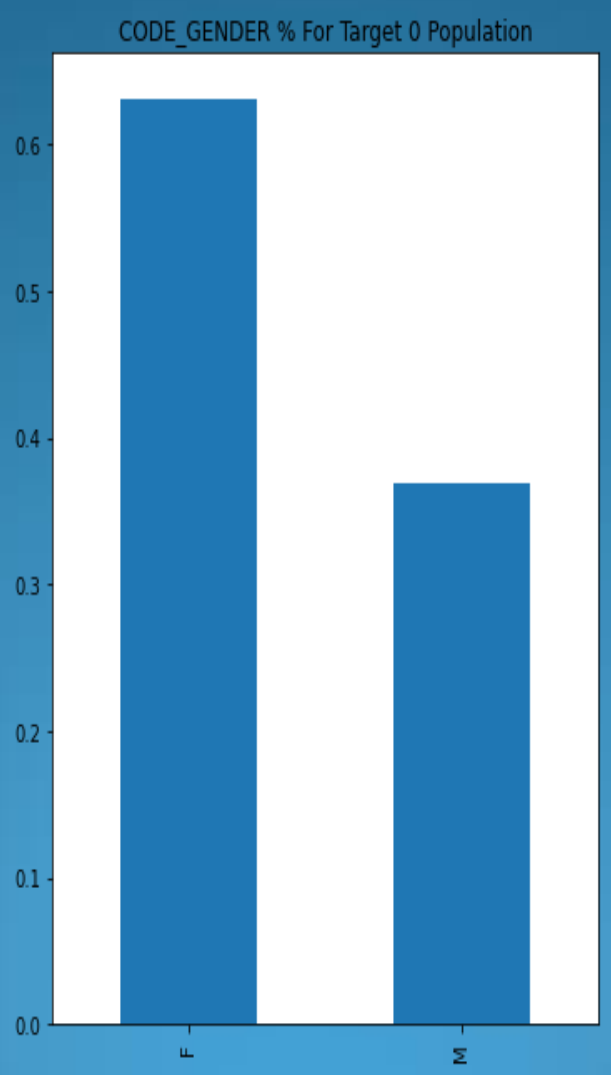



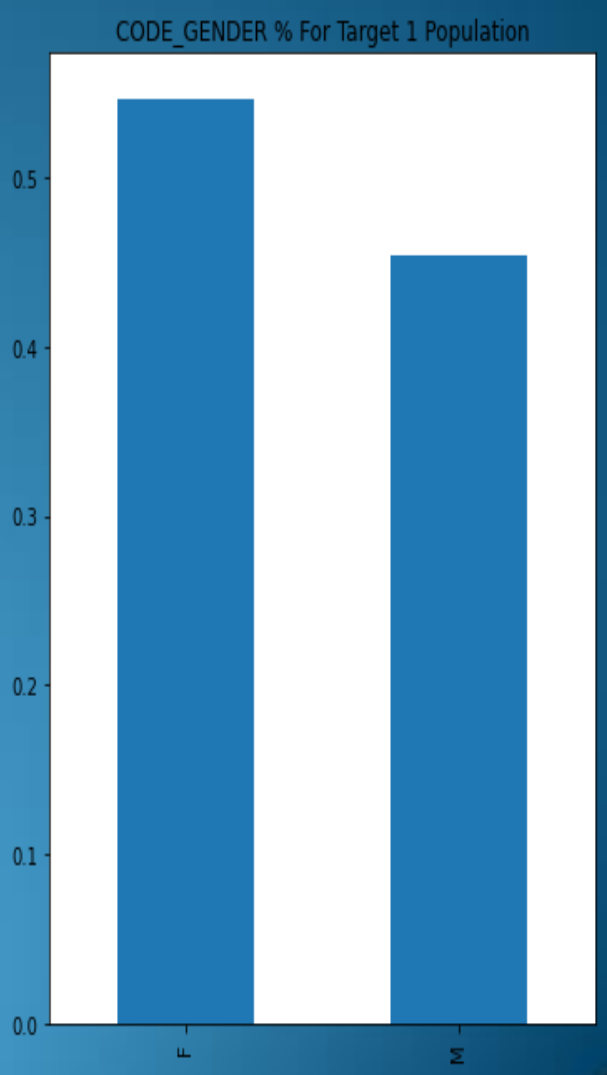
# Gender % Distribution

Points to be concluded from the graph.

1. Credit was given mainly to females
2. Male population have higher chance of defaulted than female



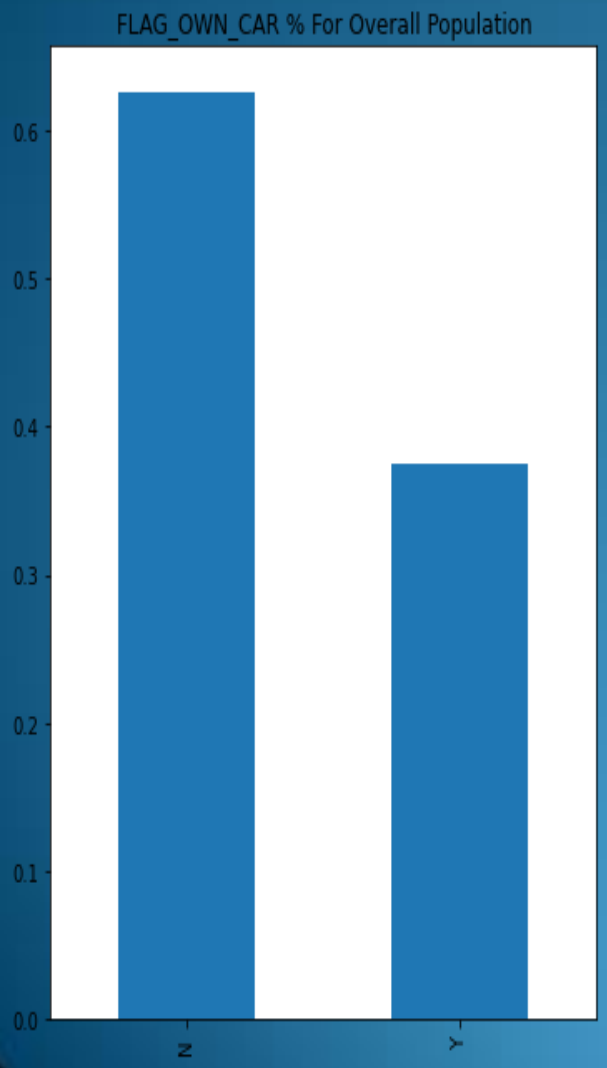



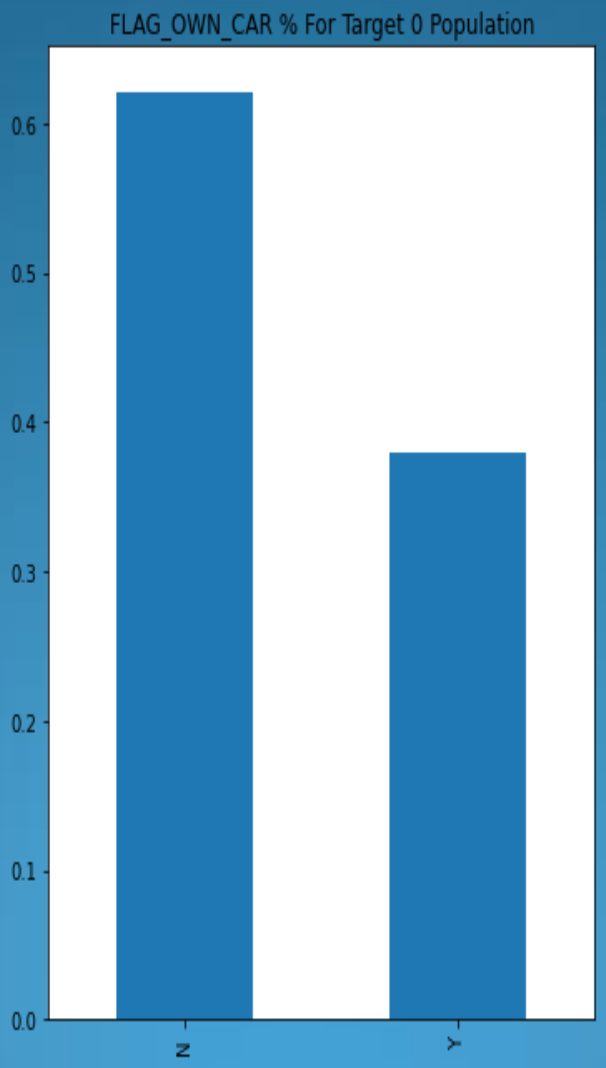



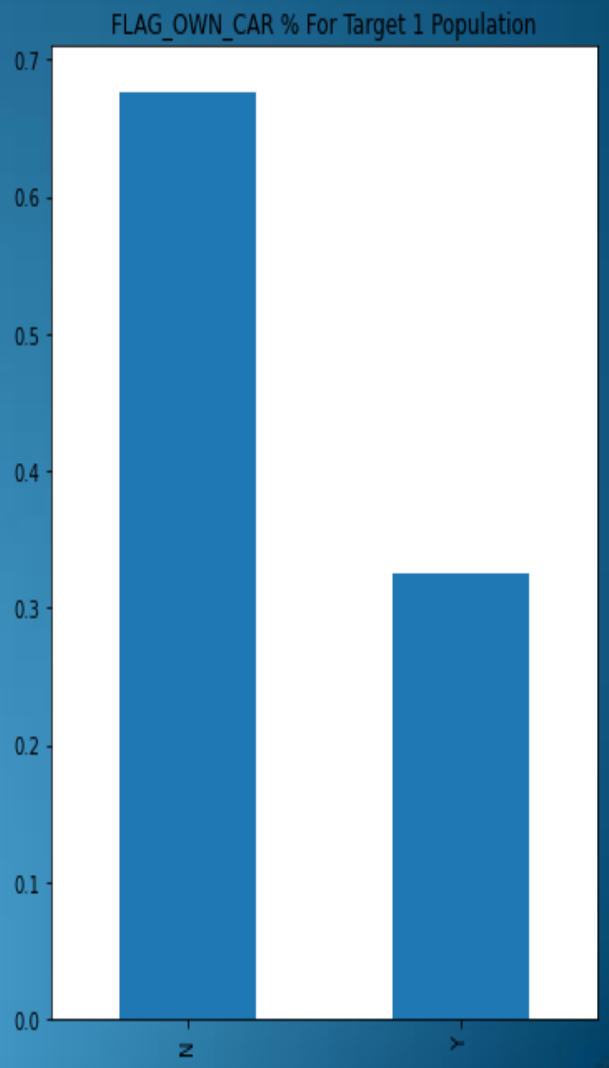
# Car Owner % Distribution

Points to be concluded from the graph.

1. Applicants without own car represents higher % of overall credit then with own car
2. Applicants without own car have higher chance of defaulted than with own car



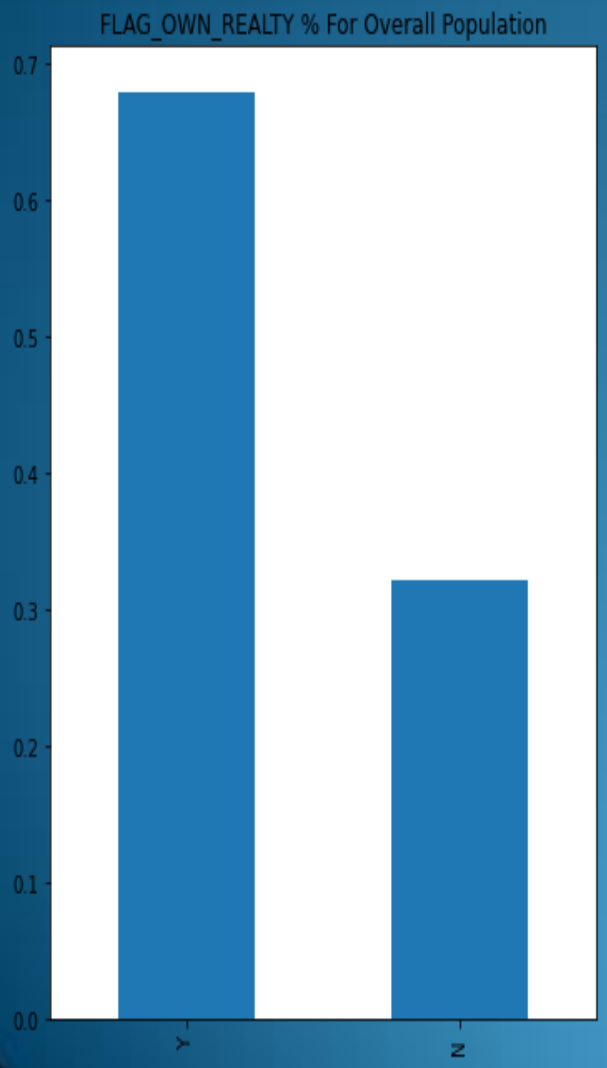



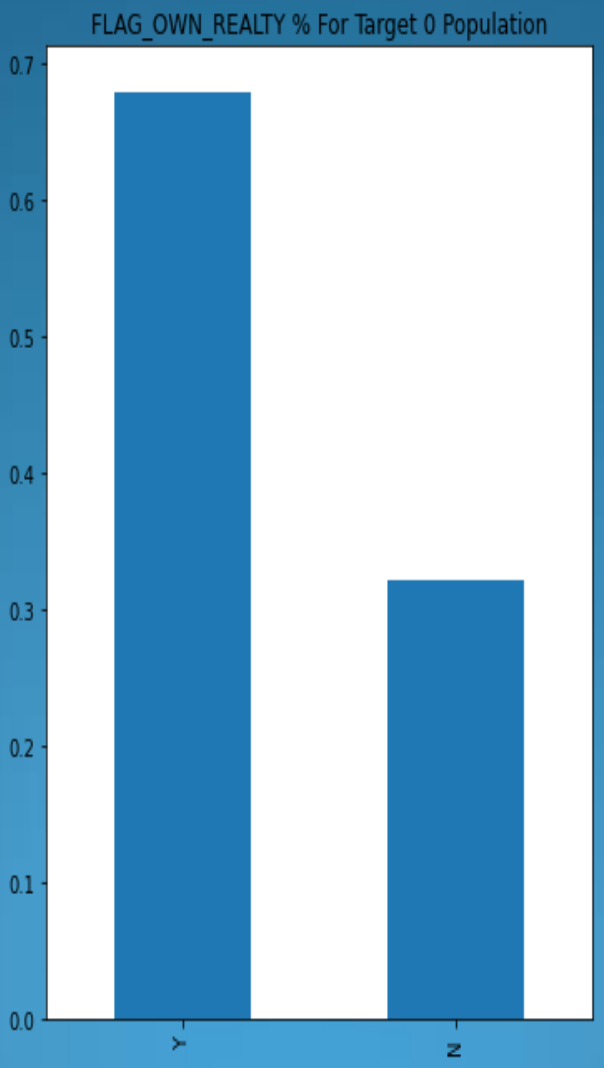



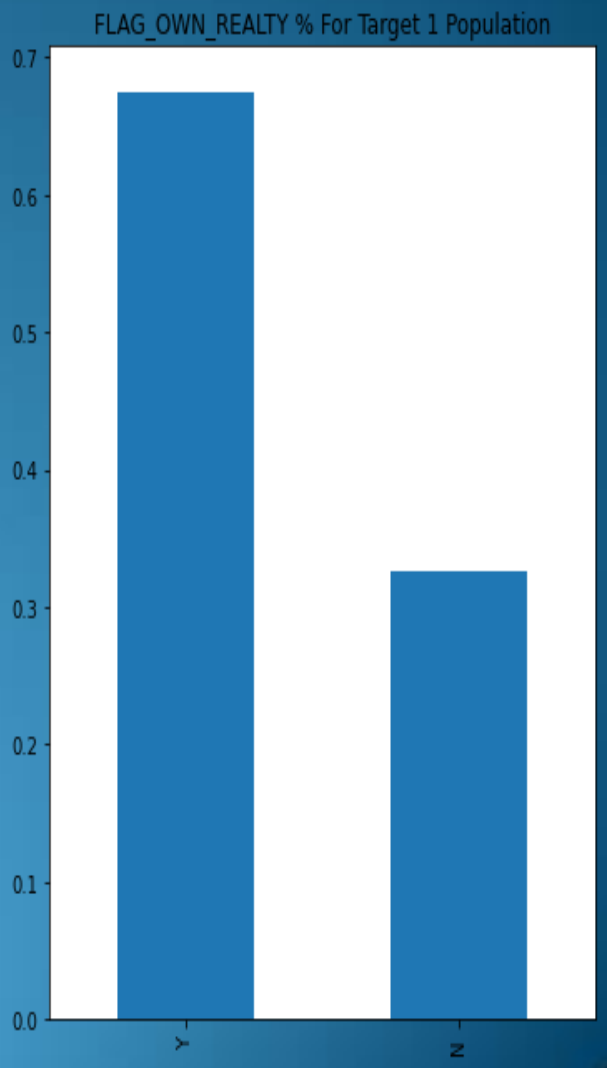
# Realty Owner % Distribution

Points to be concluded from the graph.

1. Applicants with realty represents higher % of overall credit then without realty



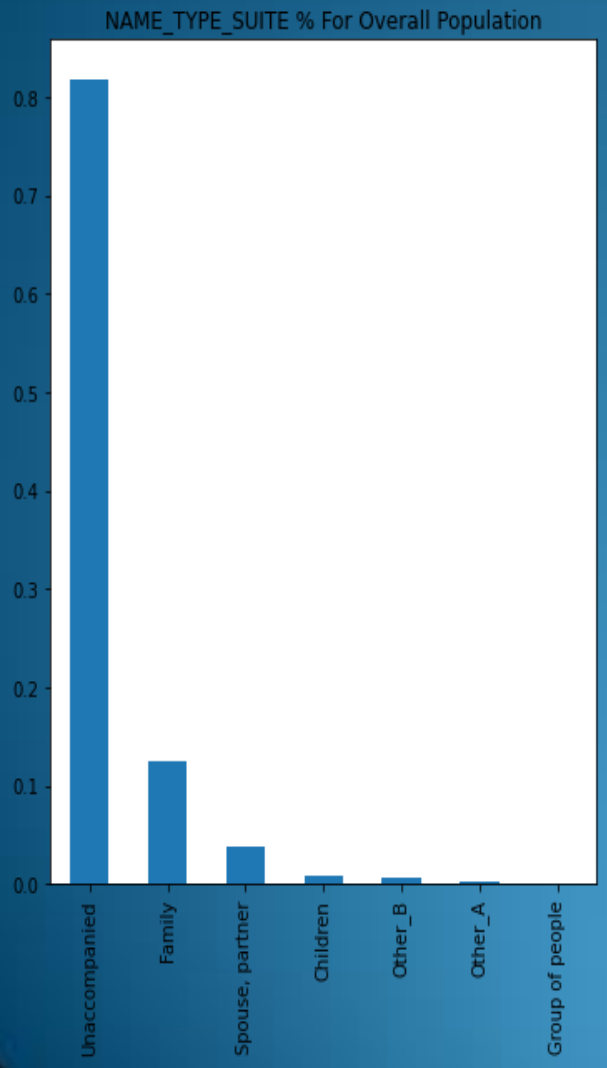



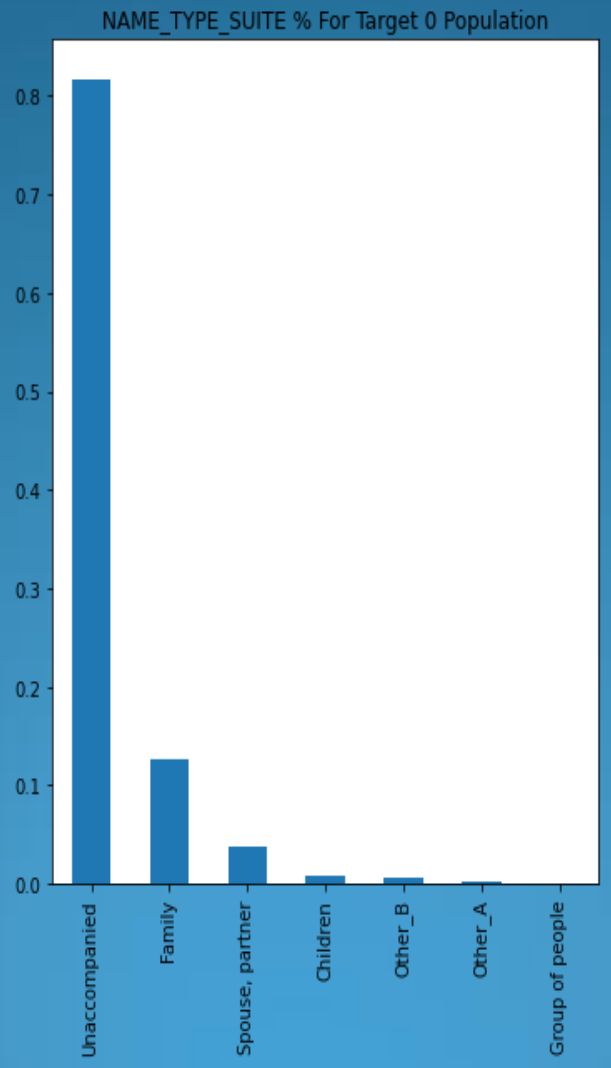



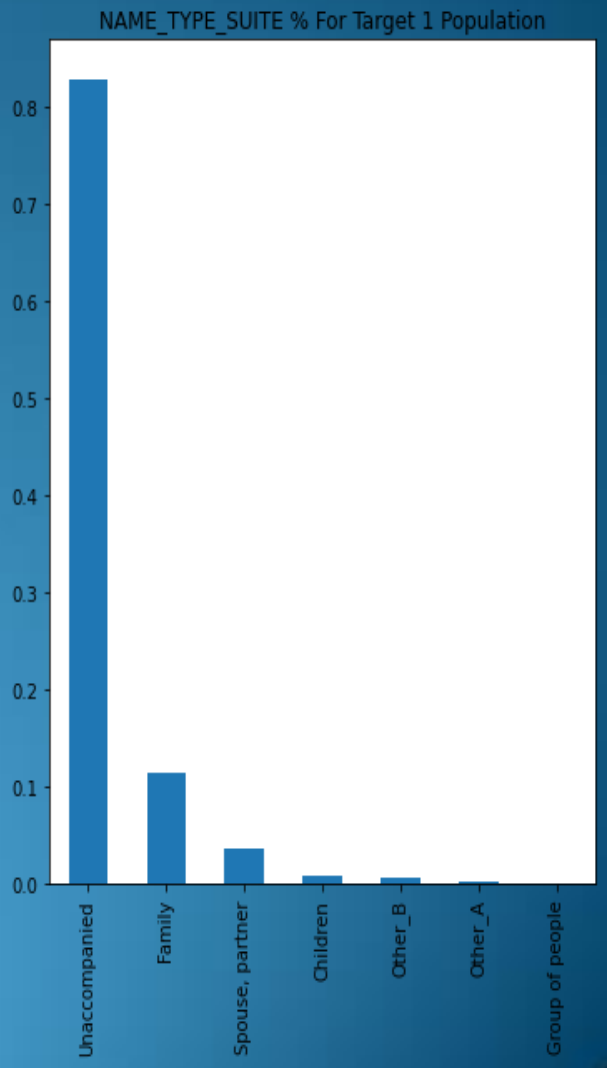
# Suite Type % Distribution

Points to be concluded from the graph.

1. "Unaccompanied" represents higher % credit



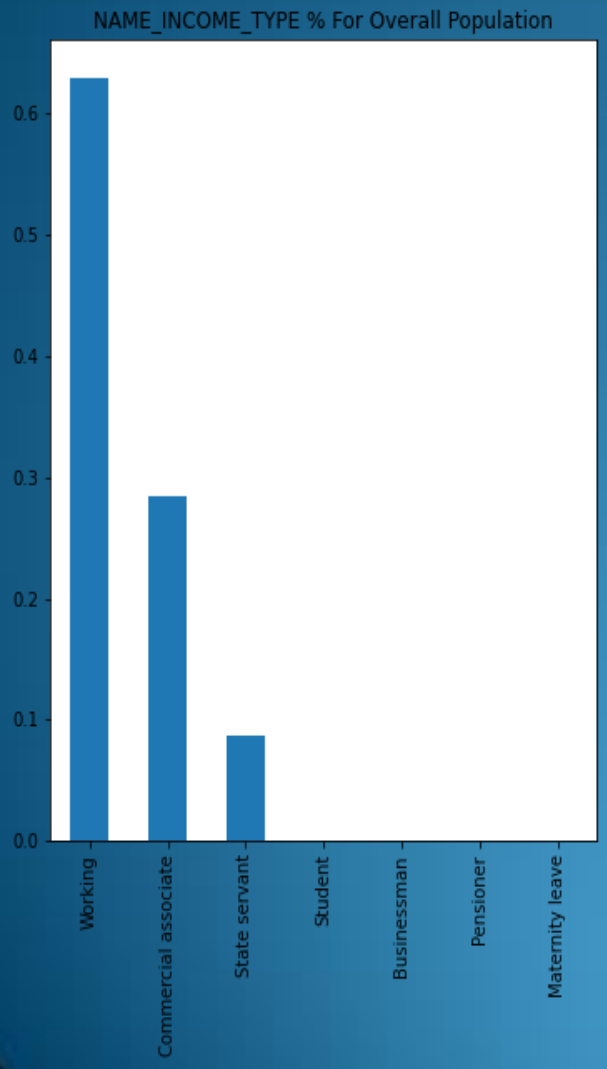



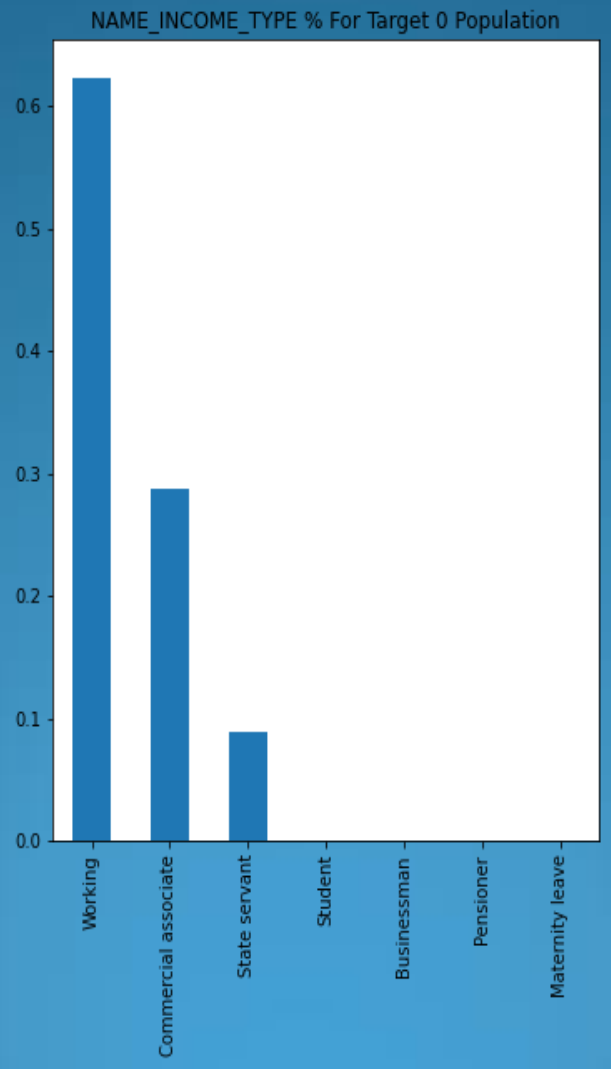



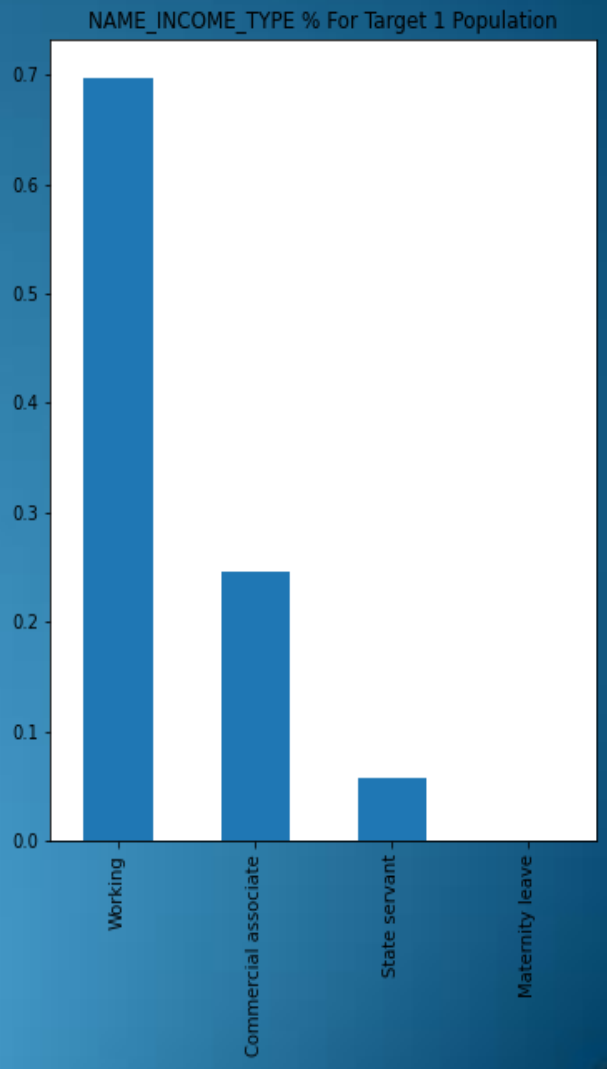
# Income Type % Distribution

Points to be concluded from the graph.

1. Applicants with "Working", "Commercial Associate“ & State Servant" status represents higher % of overall credit then others
2. Applicants with "Student", "Businessman" & "Pensioner" status have less chance to being defaulted



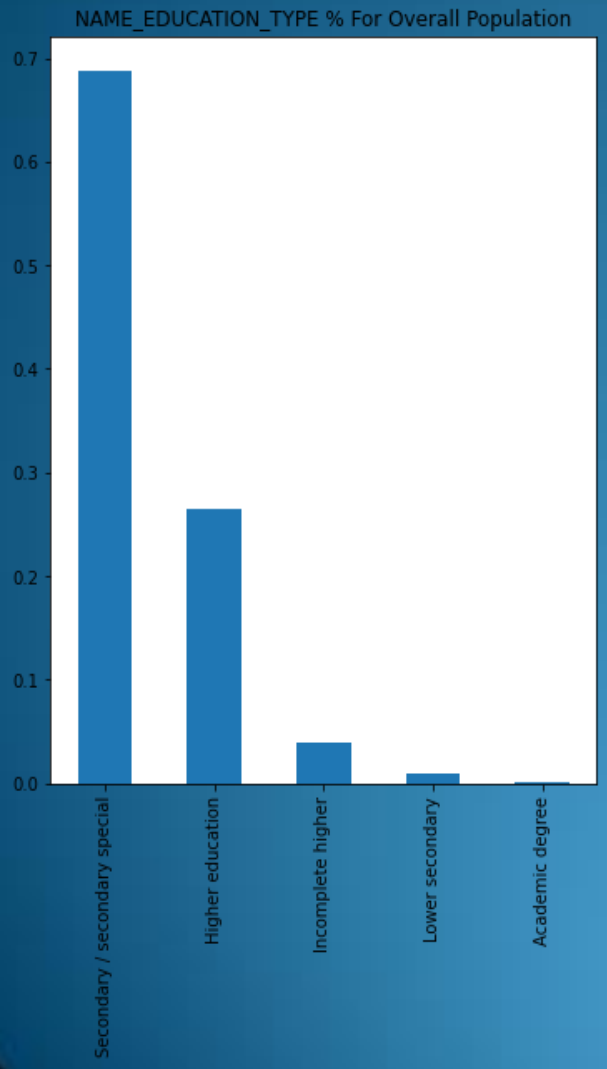



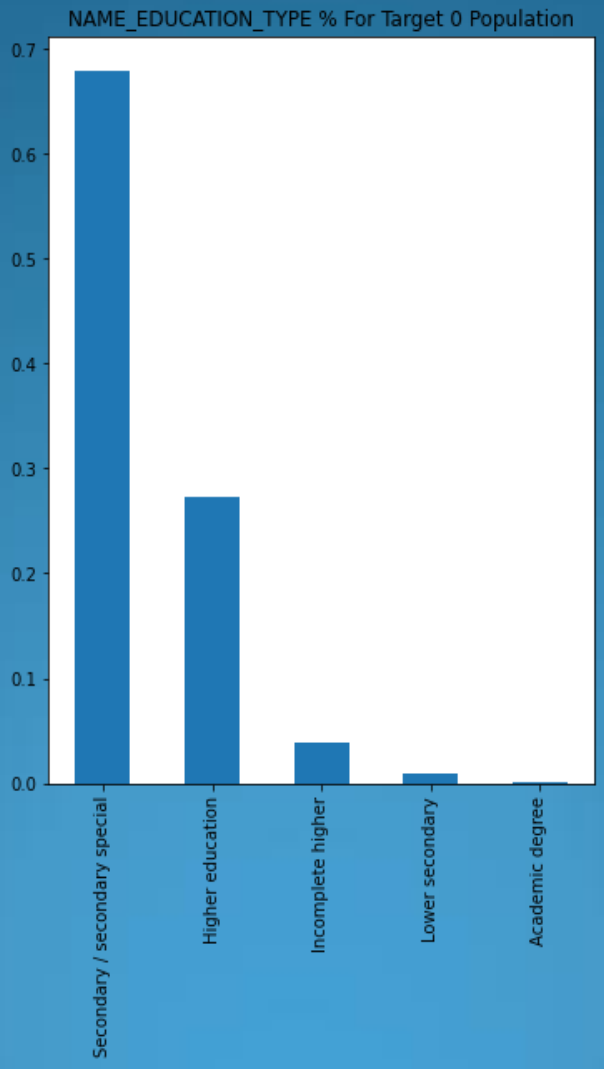



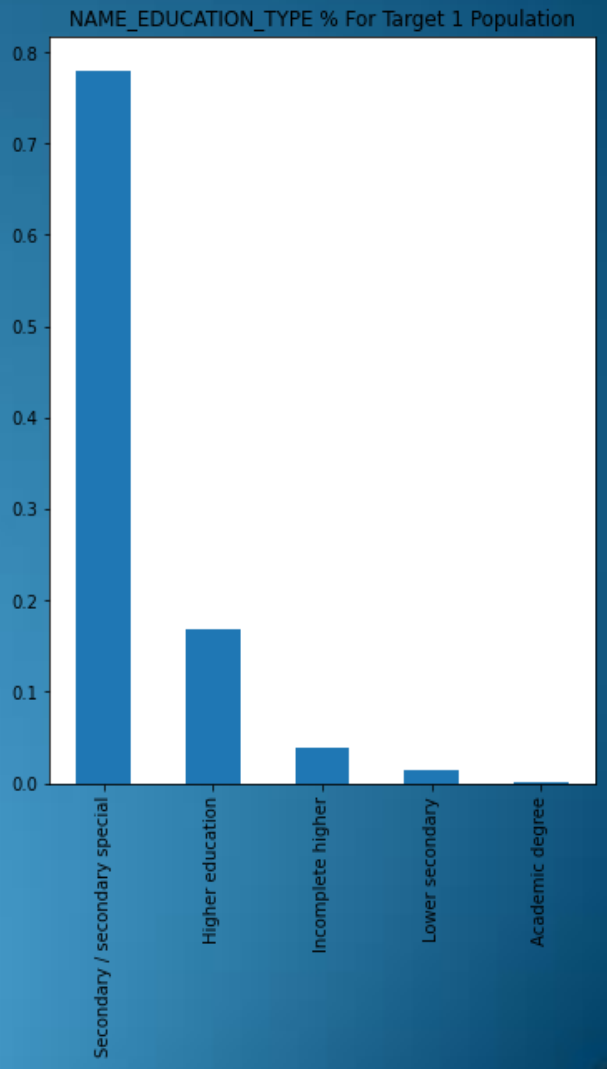
# Education Type % Distribution

Points to be concluded from the graph.

1. Applicants with "Secondary/Secondary Special" & "Higher Education" status represents higher % of overall credit then others
2. Applicants with "Secondary/Secondary Special" have higher chance to being defaulted compare to other



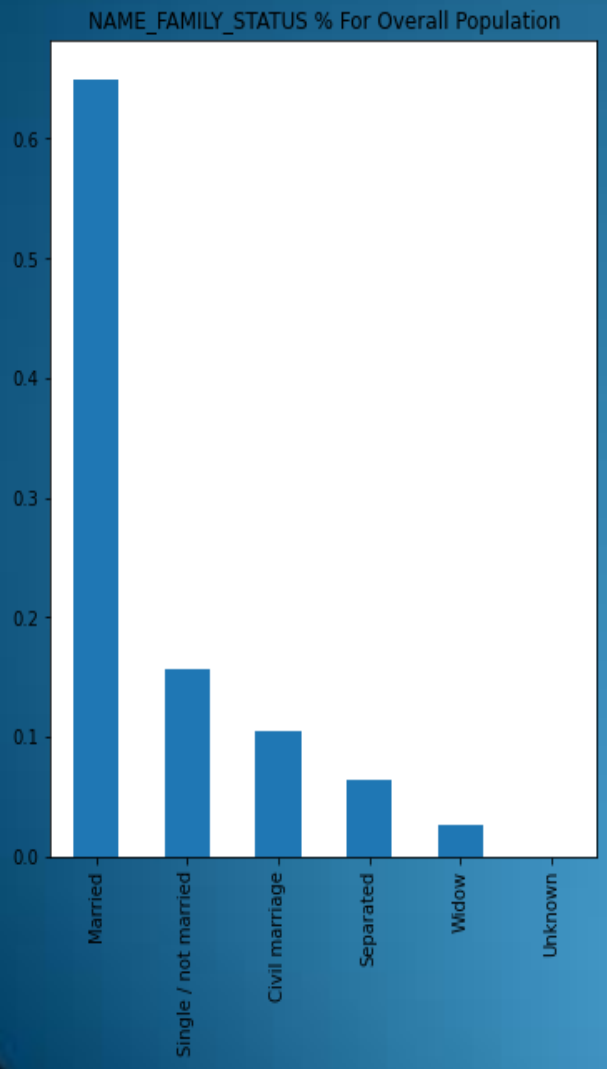



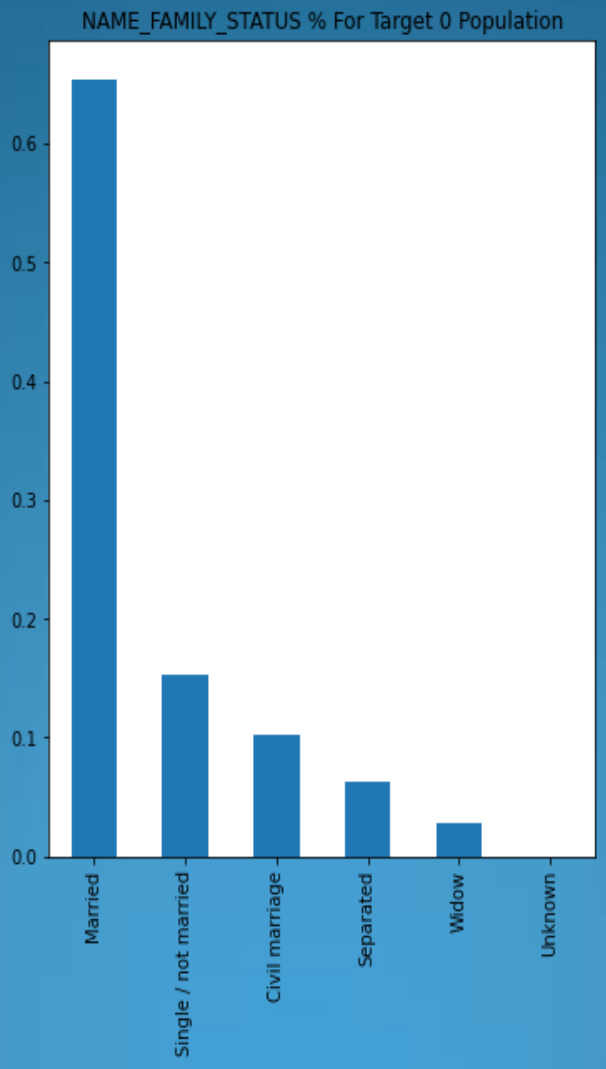



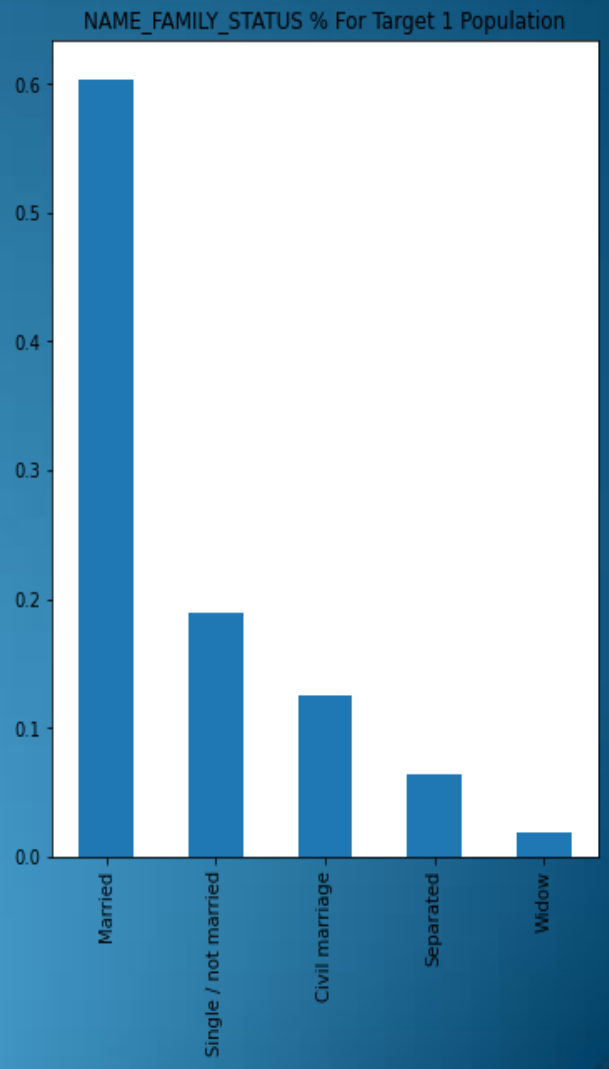
# Family Status % Distribution

Points to be concluded from the graph.

1. Applicants with “Married" & “Single/Not Married" status represents higher % of overall credit then others
2. Applicants with "Married" status have lesser chance to being defaulted compare to other



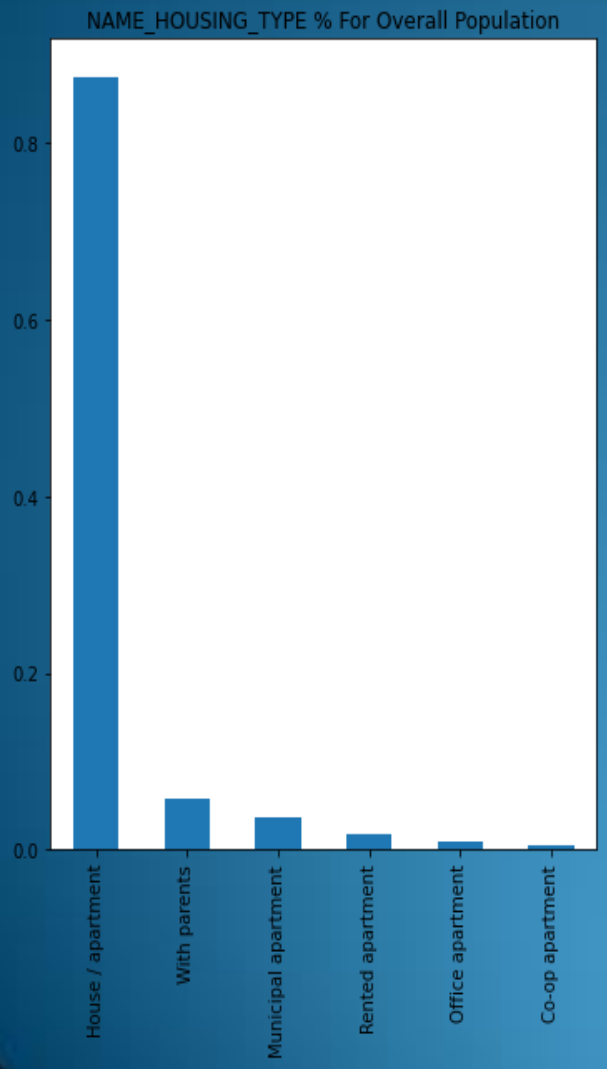



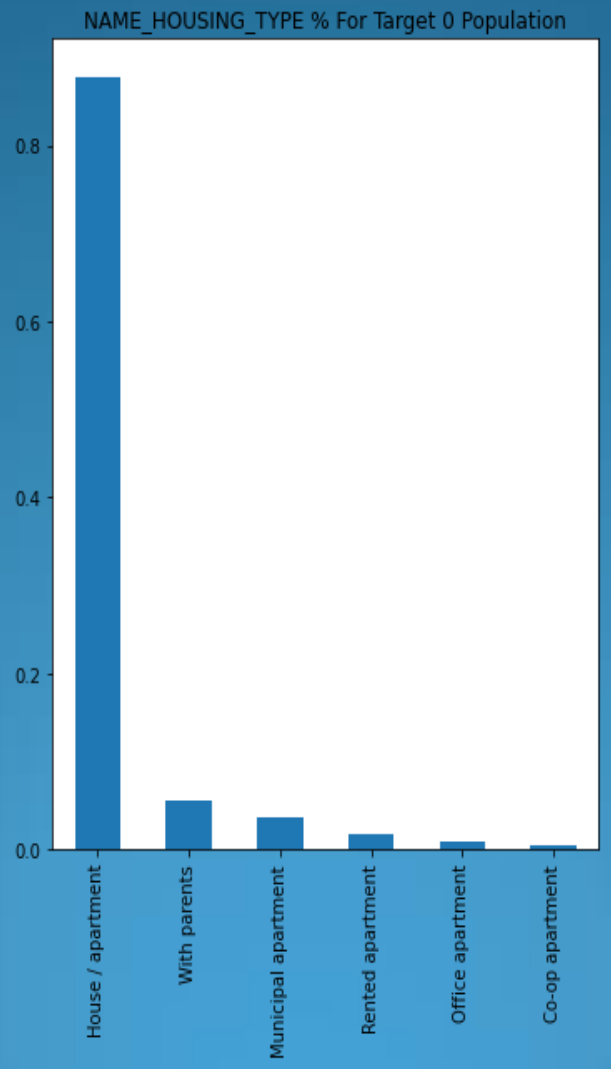



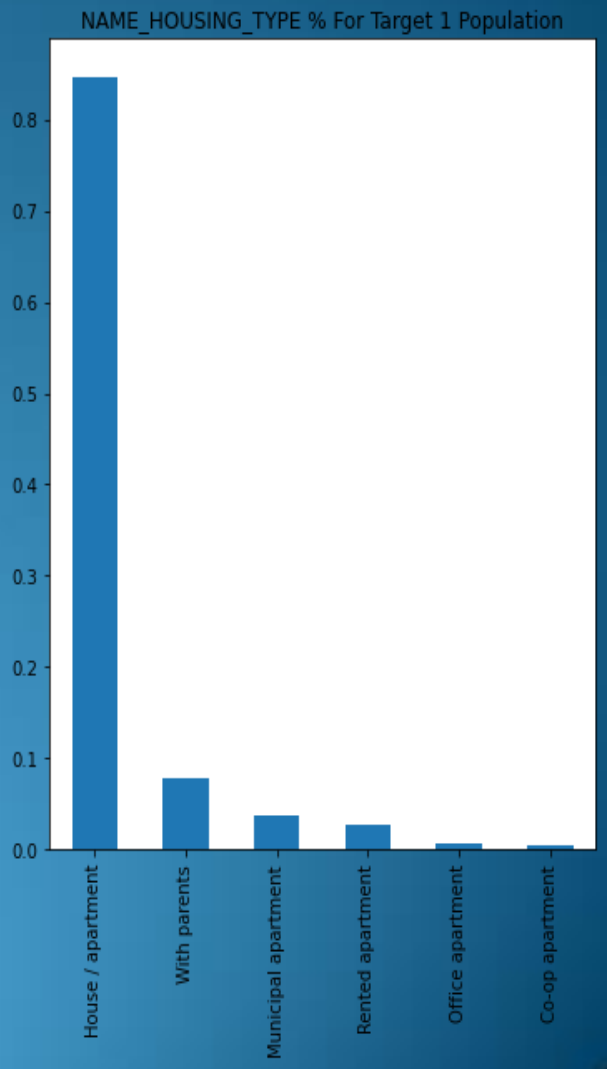
# Housing Type % Distribution

Points to be concluded from the graph.

1. Applicants with "House/Apartment" status represents higher % of overall credit then others
2. Applicants "With Parent" status have higher chance to being defaulted compare to other



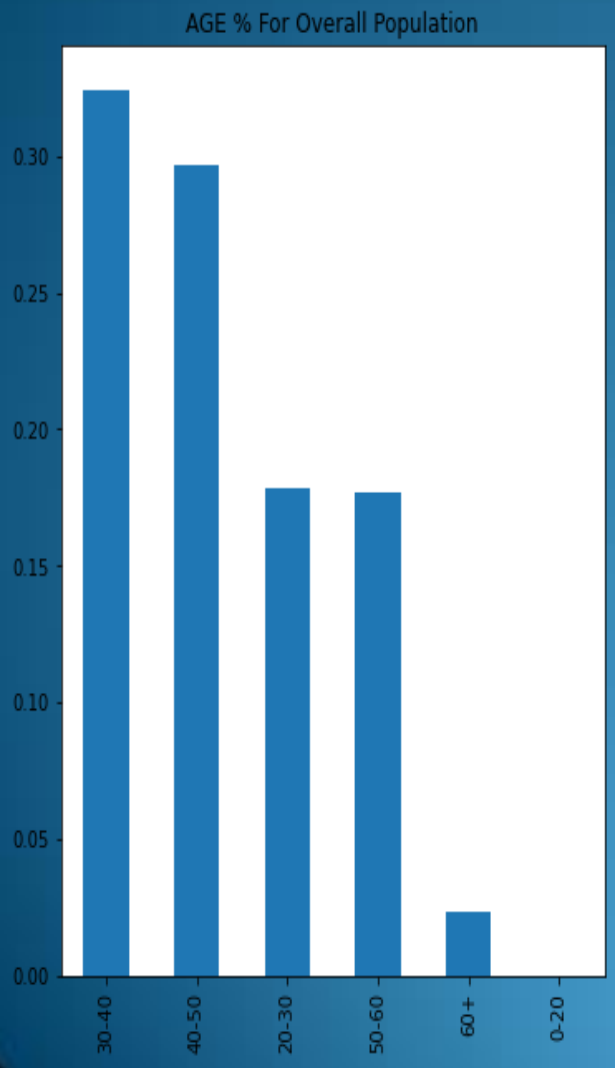



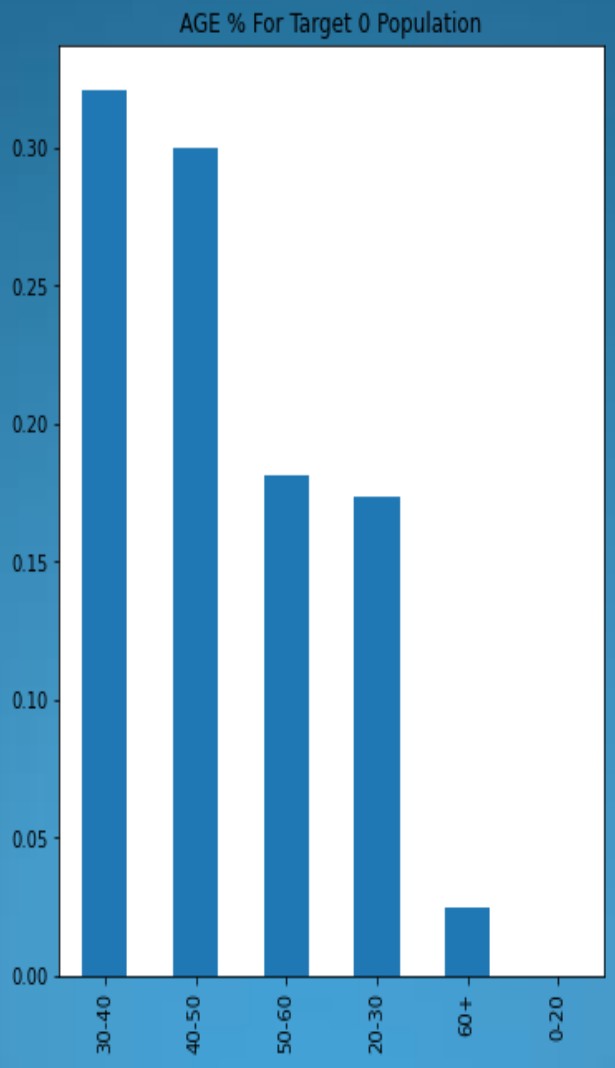



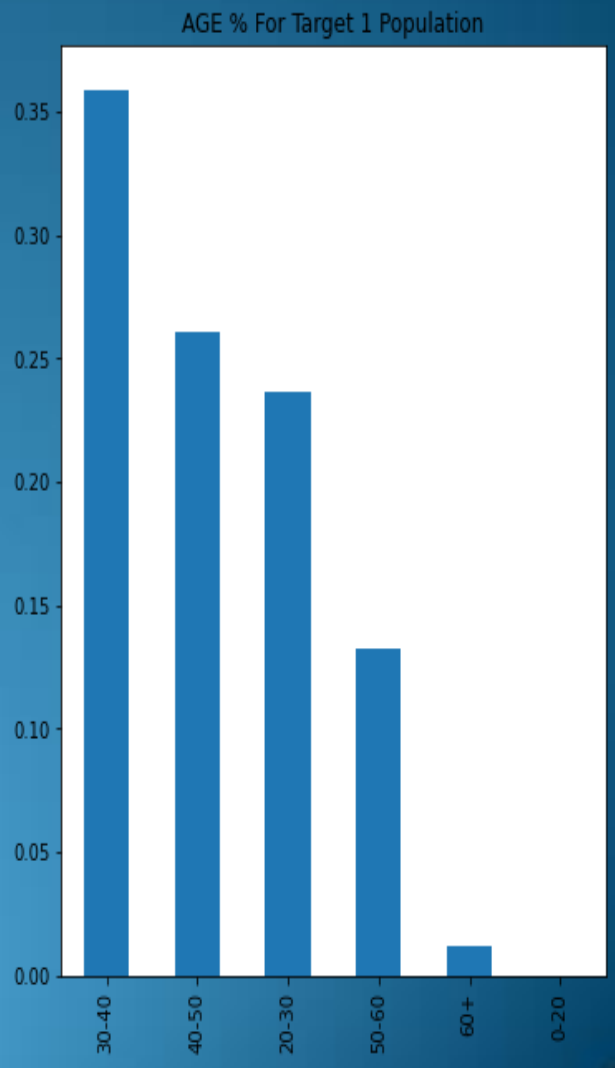
# Age % Distribution

Points to be concluded from the graph.

1. The max credit is given to applicants above age 30
2. The age group 20-30 have higher chance of default than others



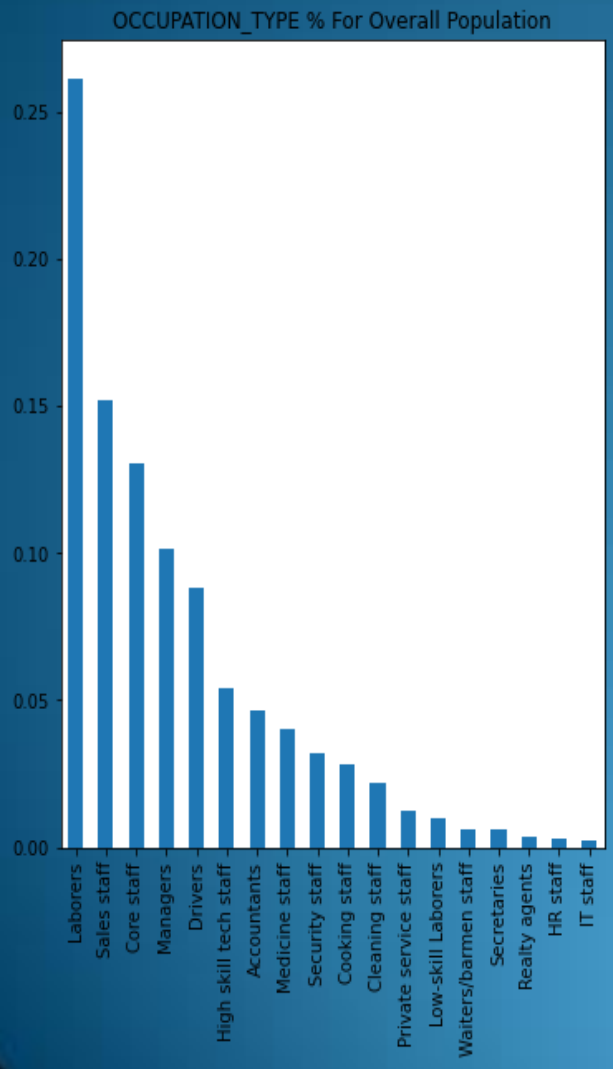



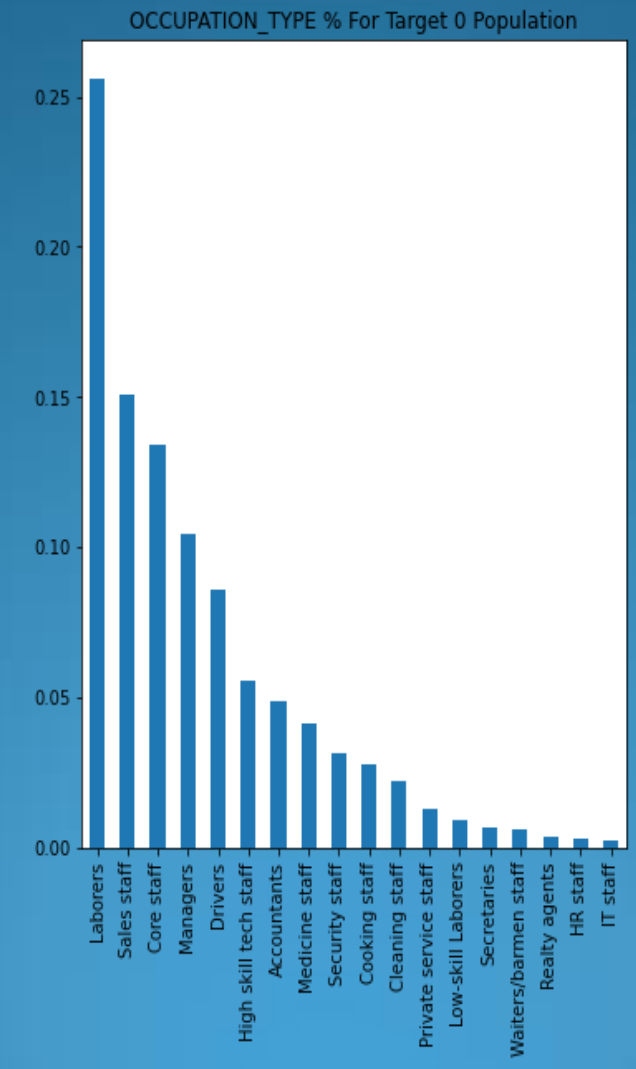



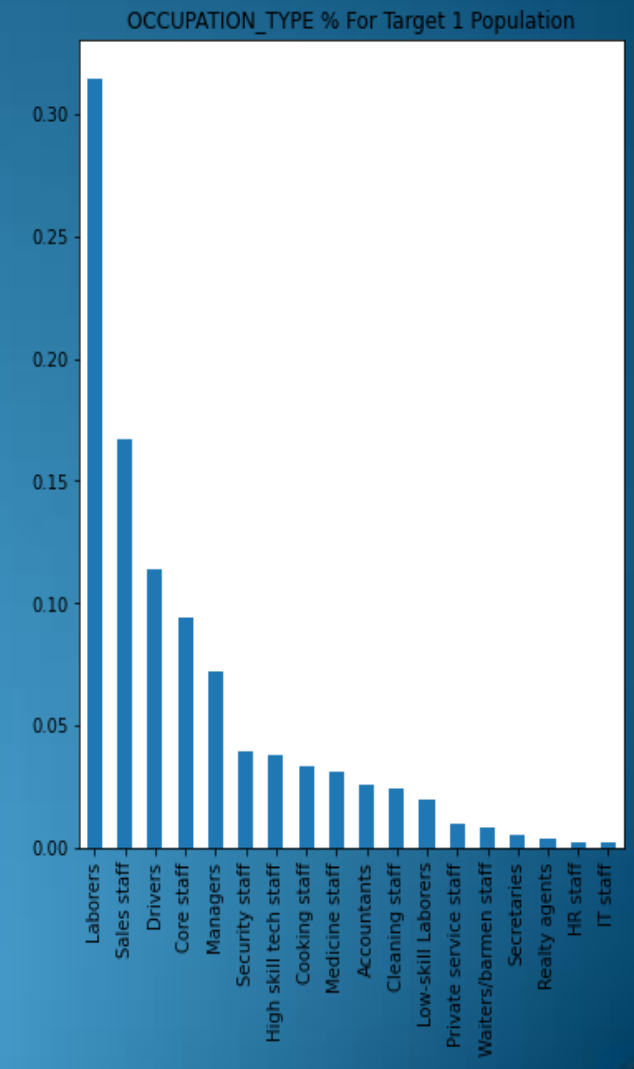
# Occupation Type % Distribution

Points to be concluded from the graph.

1. "Laborers", "Sales Staff", "Core Staff", "Managers" & "Drivers" represents higher % of overall credit then others
2. “Laborers" & "Sales Staff" have higher chance to being defaulted compare to other



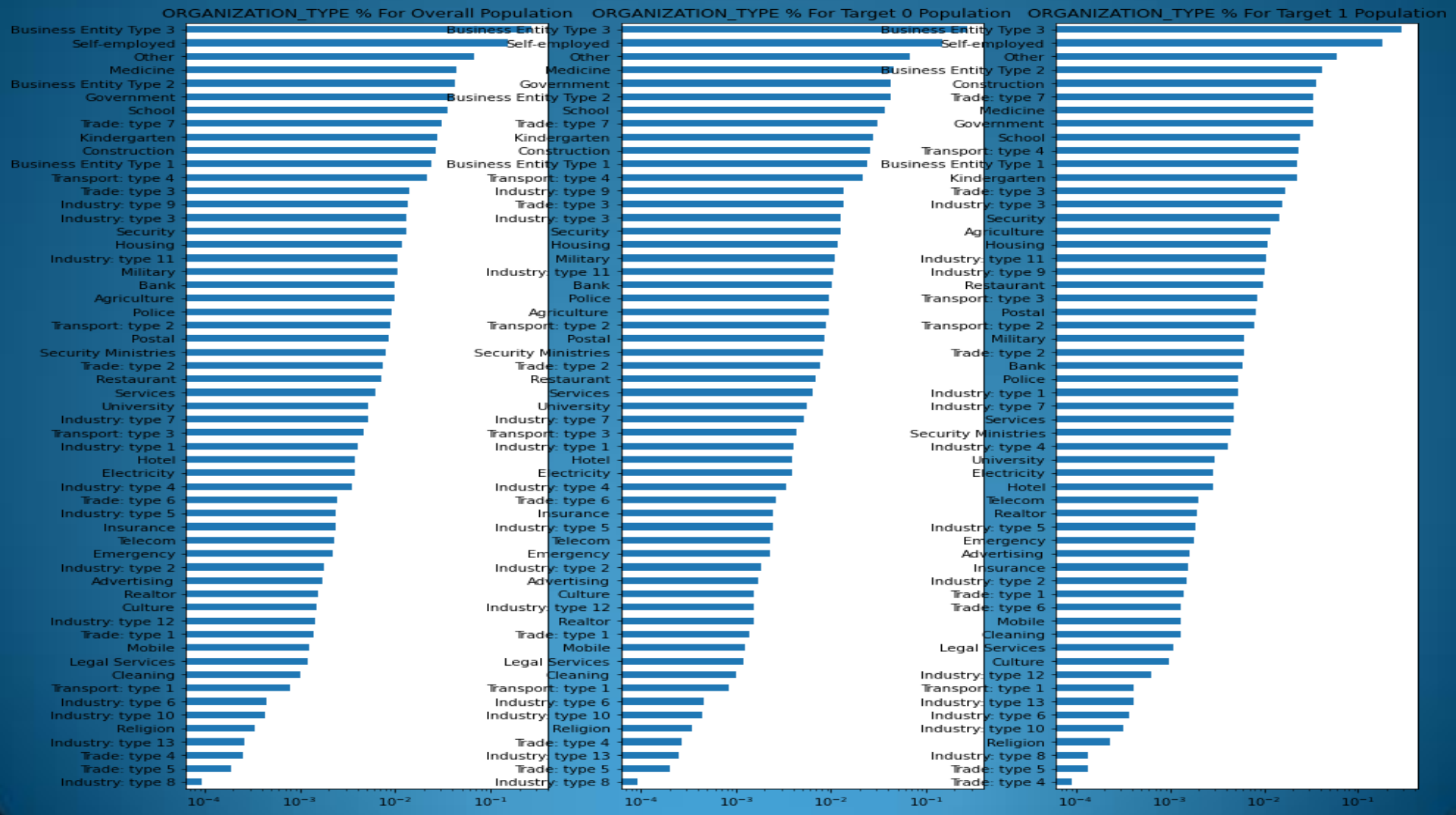



# Organization Type % Distribution

Points to be concluded from the graph.

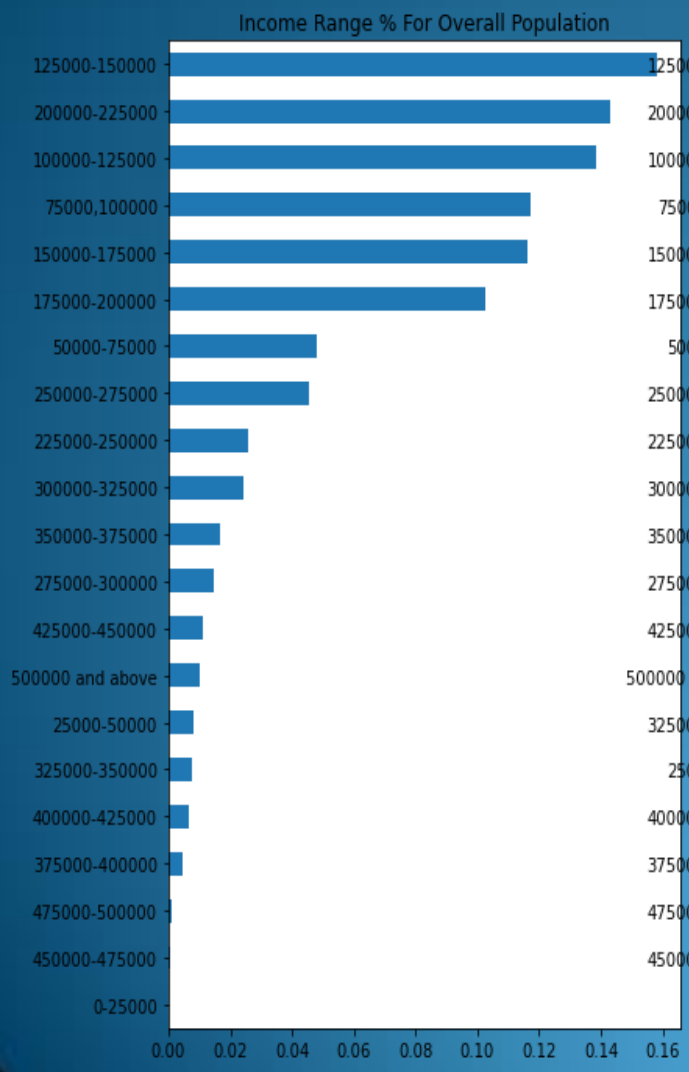
1. Applicants with "Business Entity Type 3" status represents higher % of overall credit then others
2. Applicants "Business Entity Type 3" & "Self-employed" status have higher chance of being defaulted compare to other

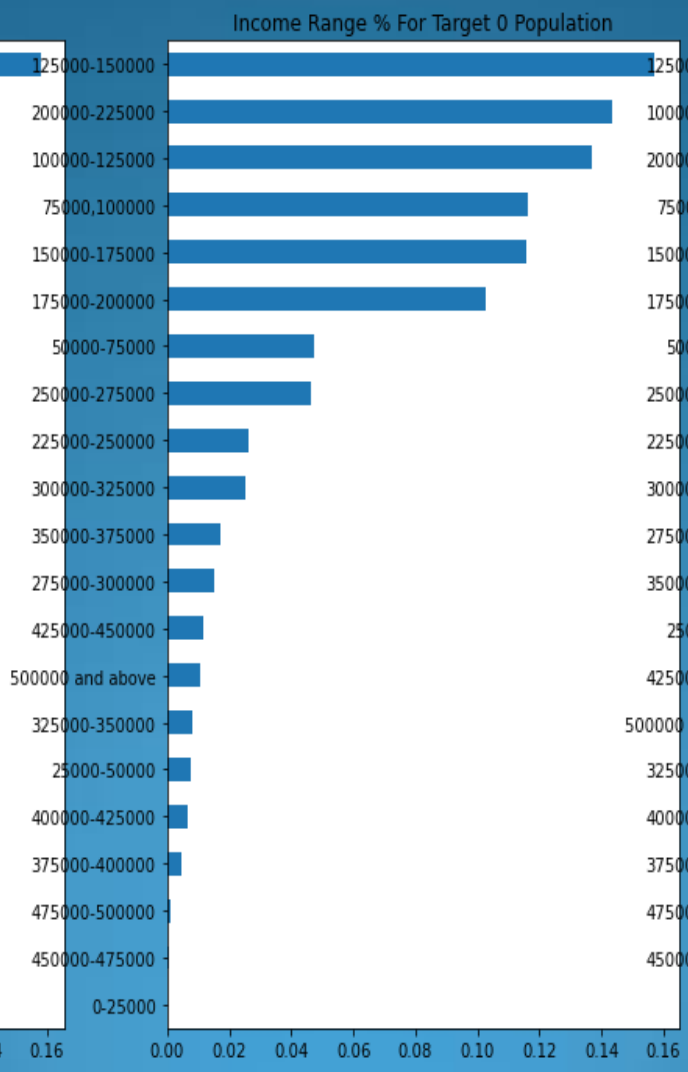


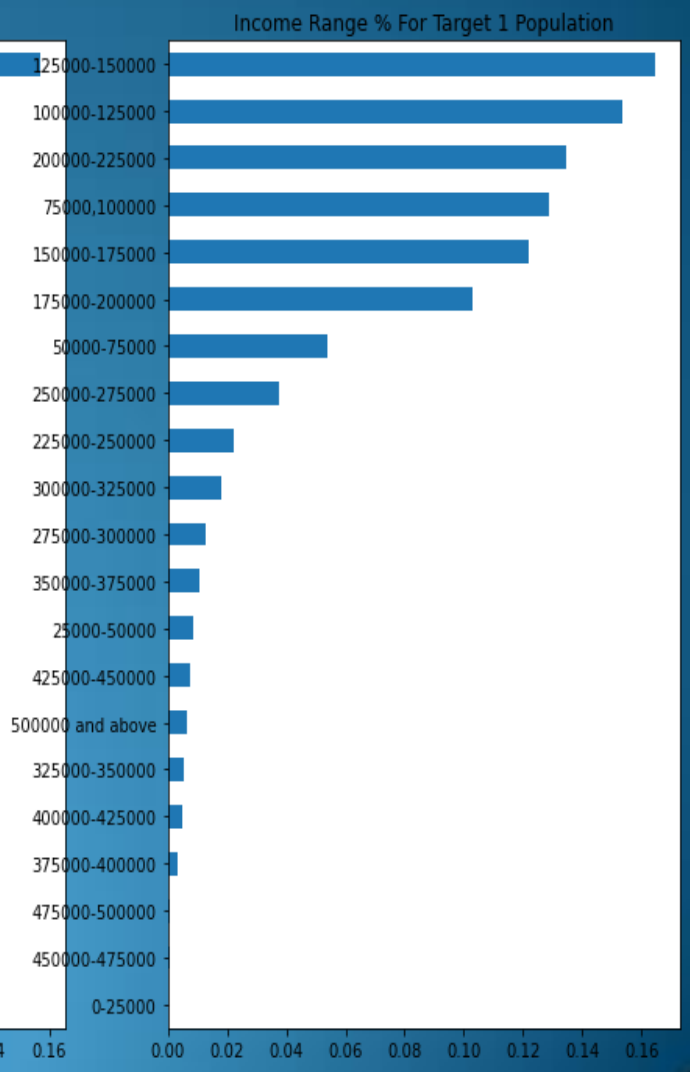
# Income Range % Distribution

Points to be concluded from the graph.

1. Applicants with income range "100000-200000" has maximum credit
2. Applicants with income range 400000 and above has less credit



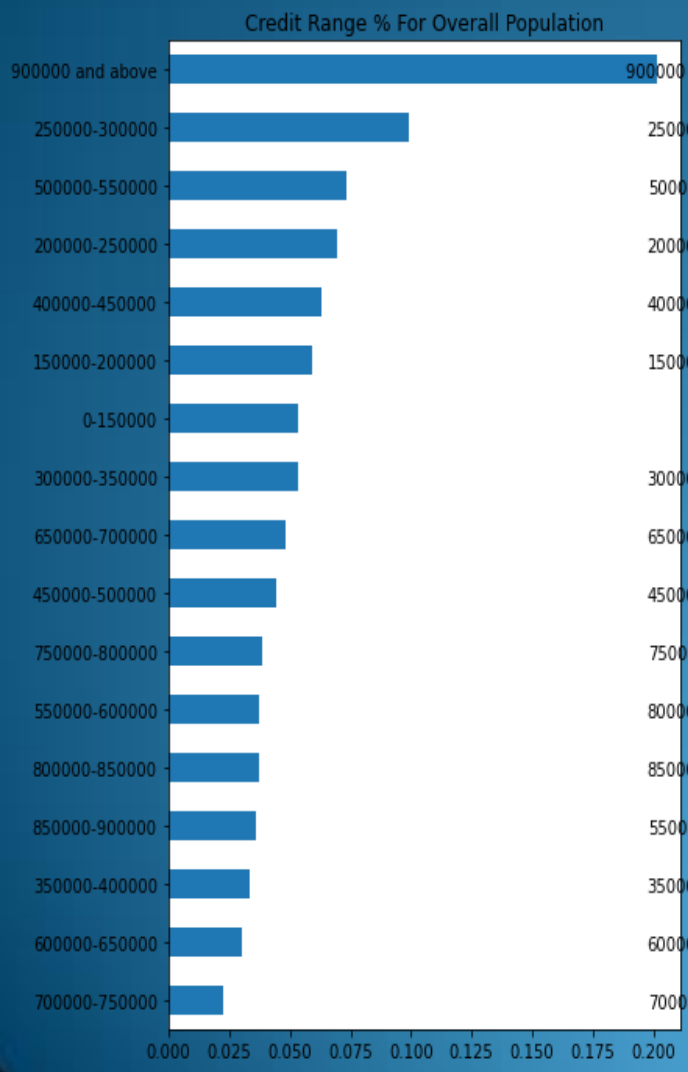



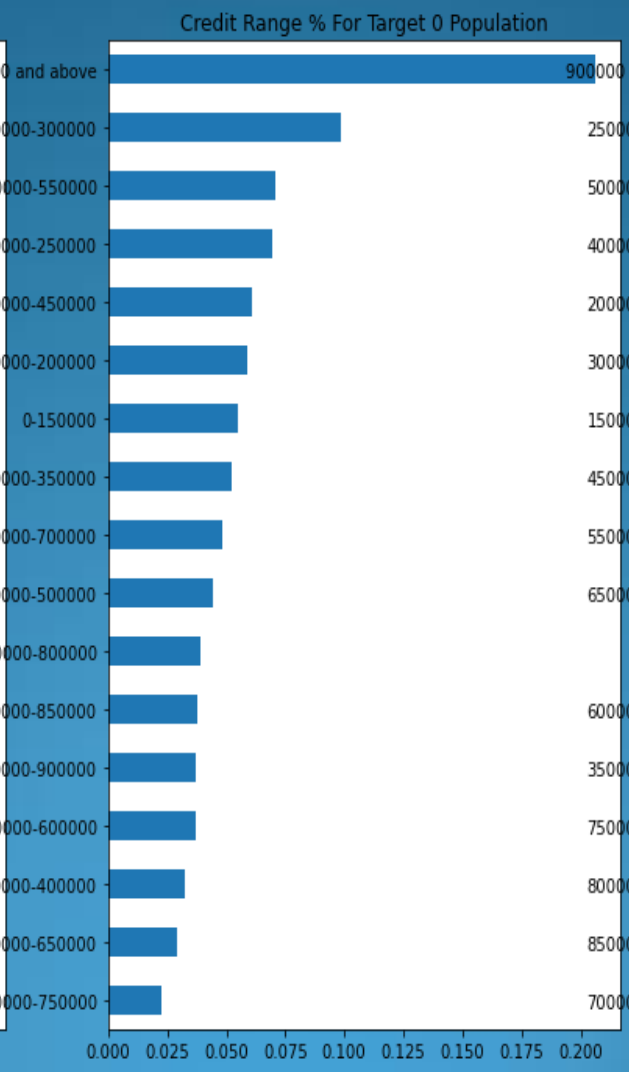



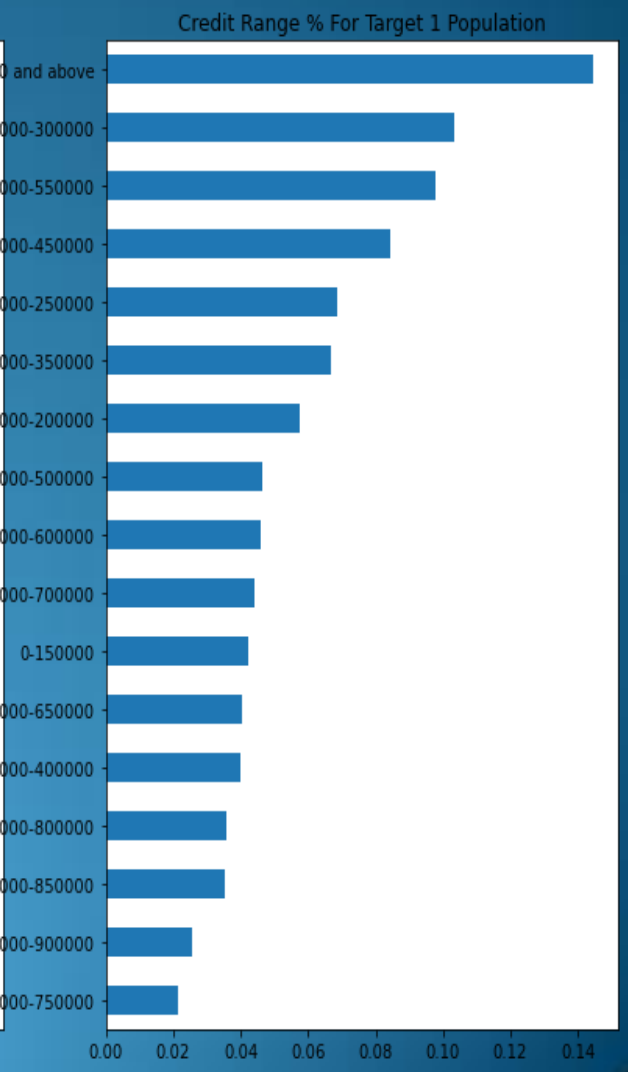
# Credit Range % Distribution

Points to be concluded from the graph.

1. The max credit is of the range of 900000 and above
2. The credit range "500000-550000" have higher chance of default than others



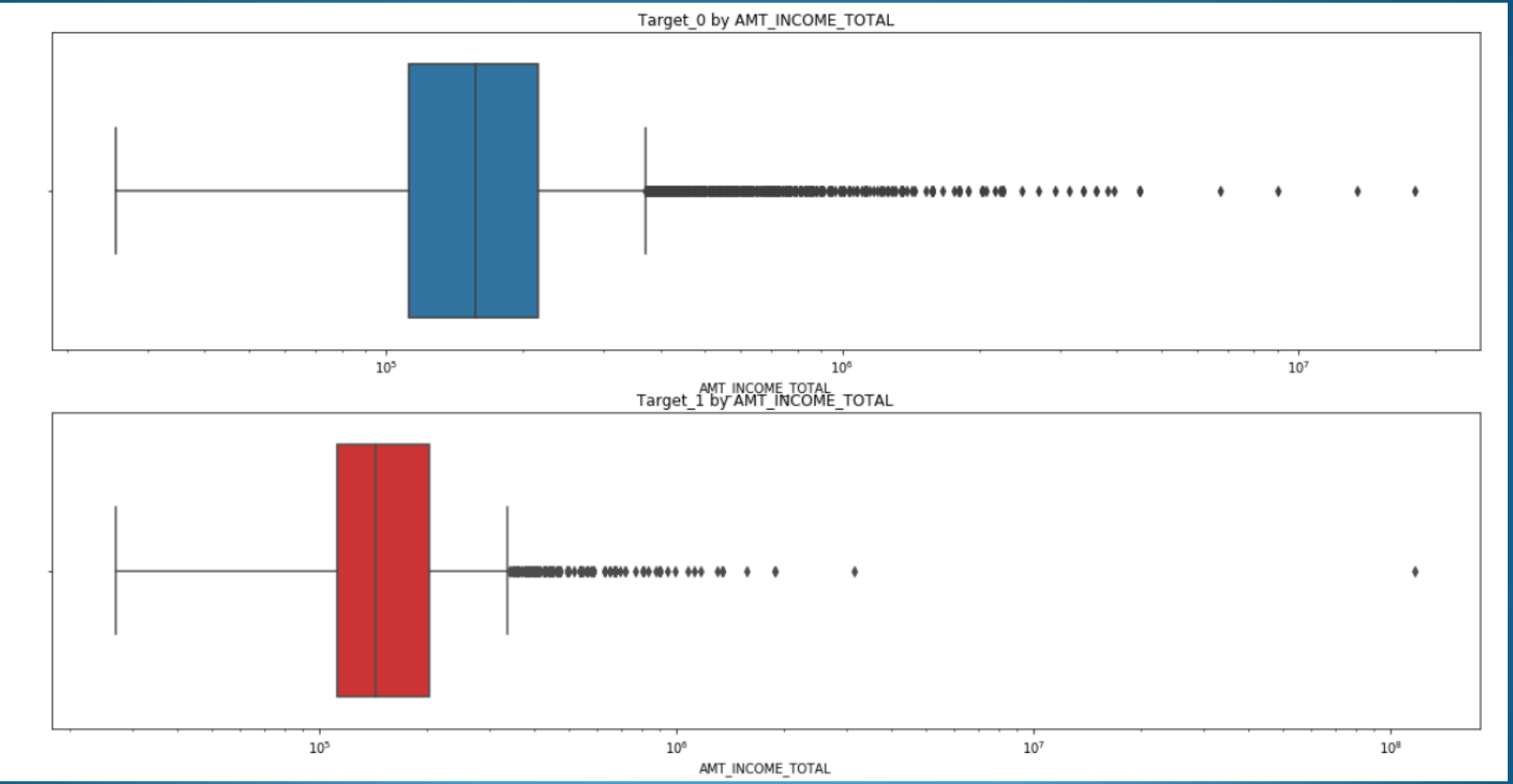



# Checking Outliers For Target0 vs Target1

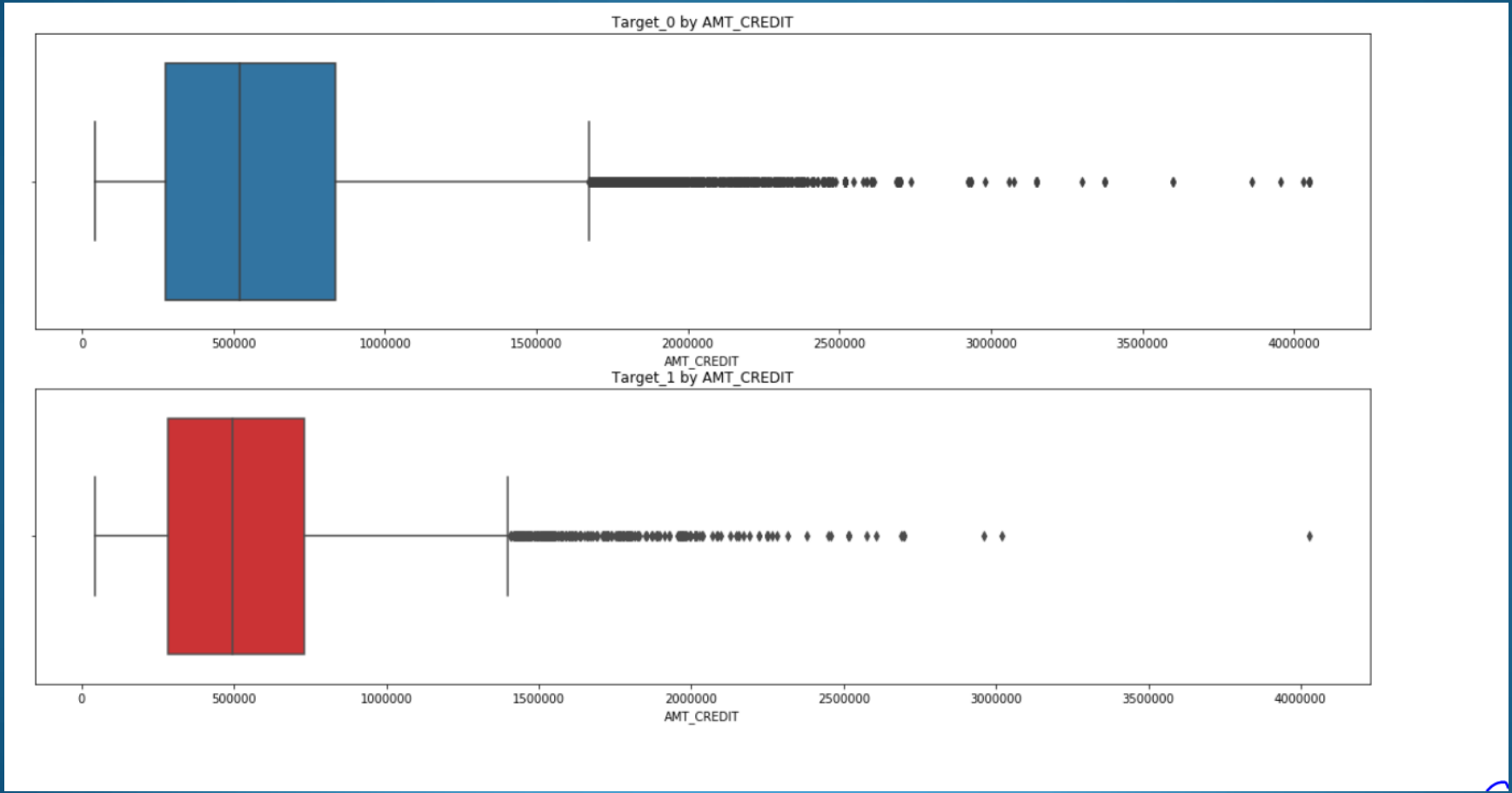
# Target 0 vs Target 1 : Amount Income Total

1.Some outliers are noticed for Target 1
2.Target 1 has the maximum income amount value compared to Target 0.
3.For Target 1: most of the clients income are in between median and third quartile region.



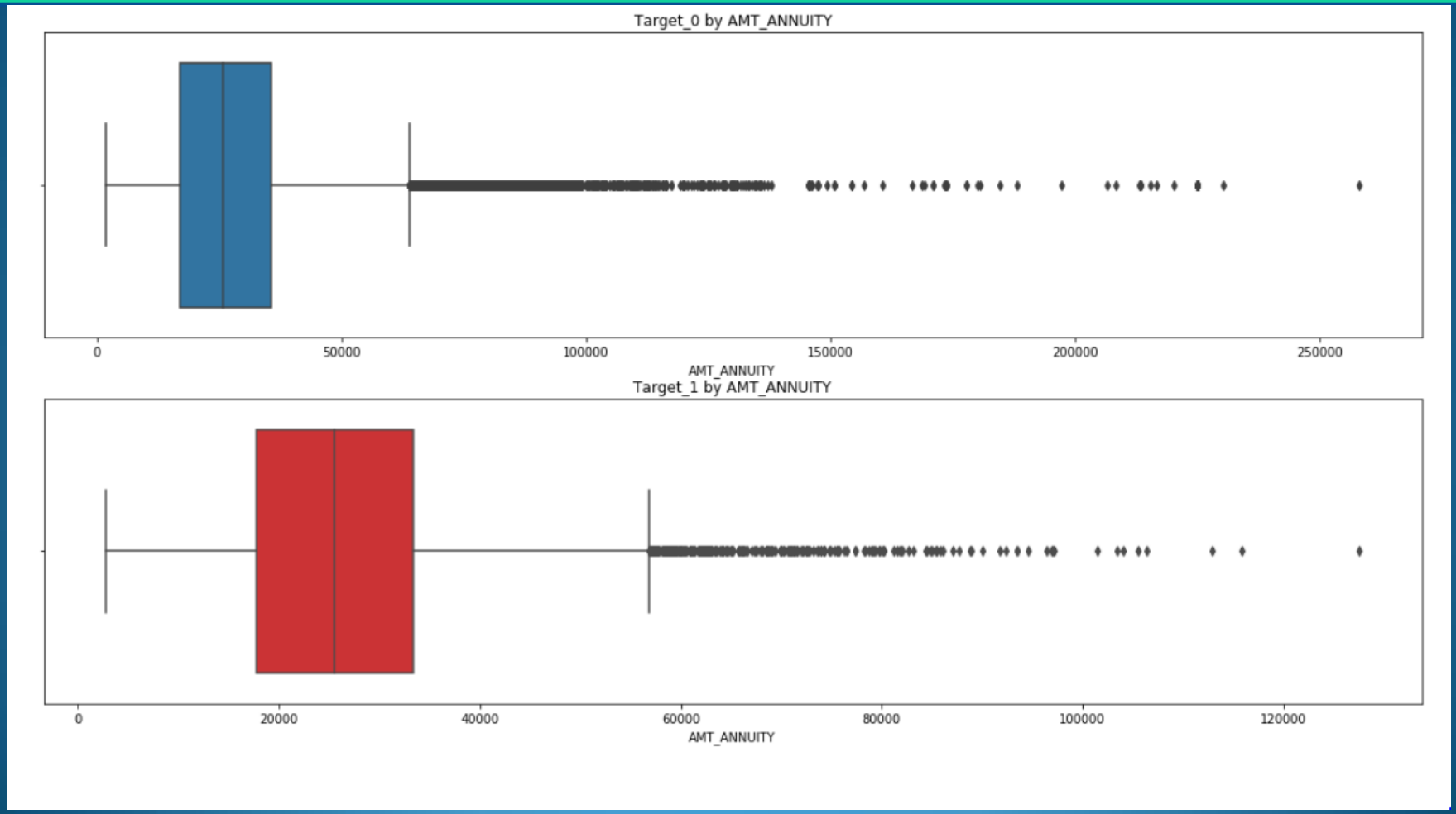
# Target 0 vs Target 1 : Amount Credit

1.Some outliers are noticed for Target 1.
2.For Target 0 and Target 1: most of the credit amounts of clients are present in between maximum to median compared to values between minimum and median.
3.For Target 0 and Target 1: most of the clients have been given higher credit amount



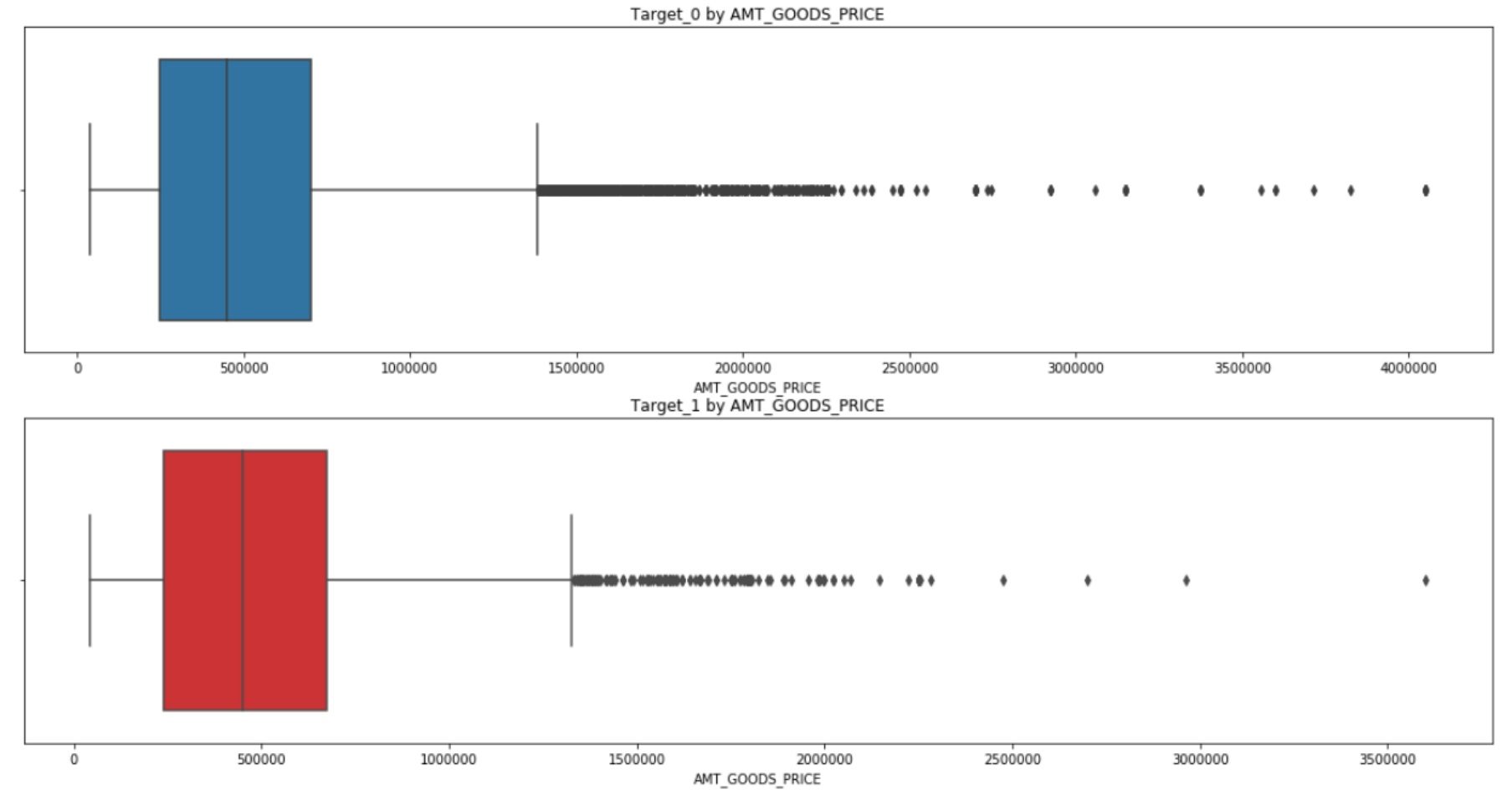
# Target 0 vs Target 1 : Amount  Annuity

1. Some outliers are noticed for Target 1 and Target 0
2. Target 0 has a higher AMT_ANNUITY' value compared to Target 1.
3. For Target 1 and Target 0: most of the clients amount annuity lies between maximum to median compared to values between minimum to median.



# Target 0 vs Target 1 : Amount  Goods Price

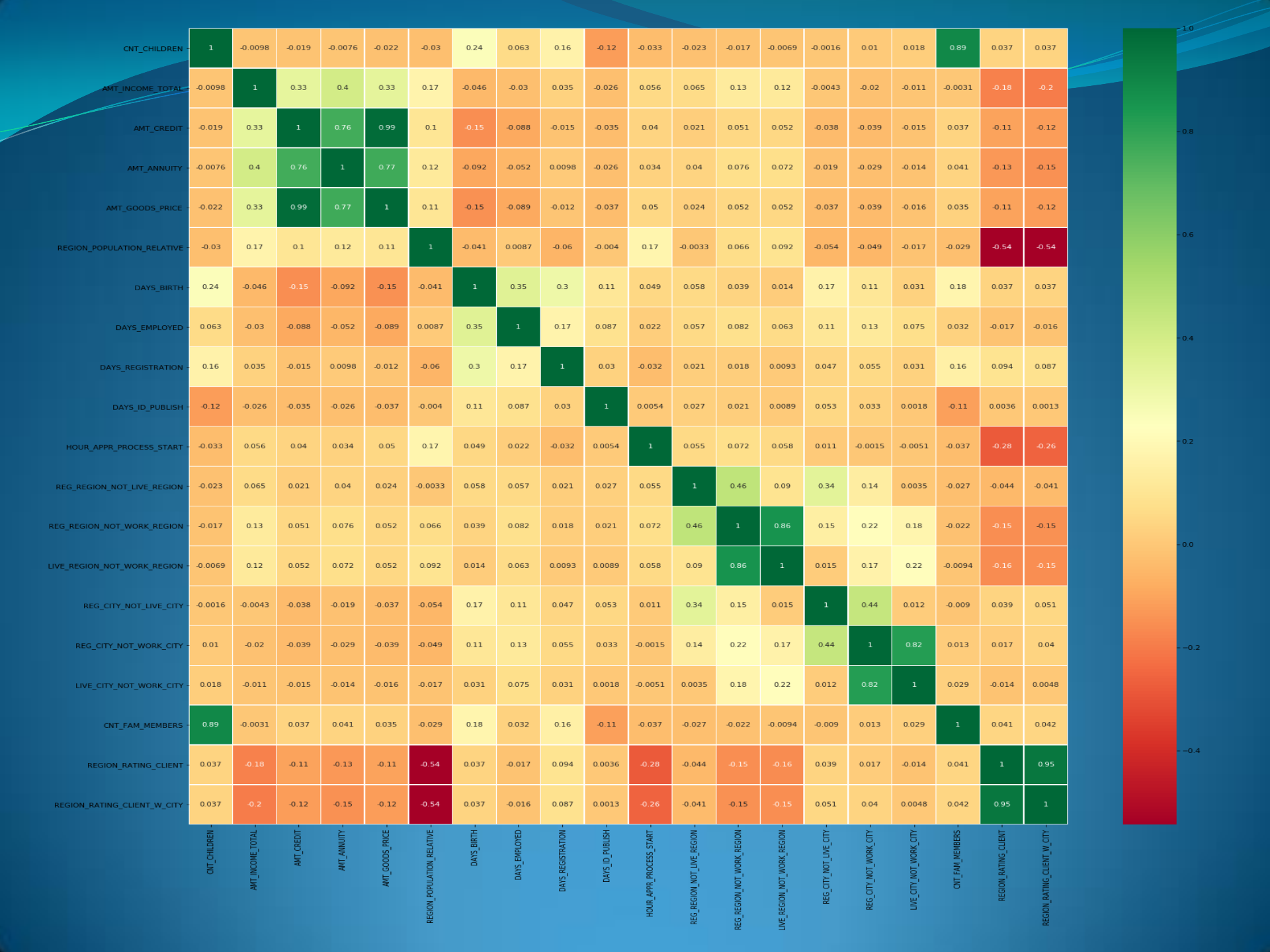
1.Some outliers are noticed for Target 1 and Target 0.
2.For Target 1 and Target 0: most of the clients amount good's price lies between maximum to median compared to values between minimum to median.
3.Target 0 has received the maximum amount value for the goods compared to Target 1.



## Correlation for numerical columns for Target 0

## Correlation for numerical columns for Target 0.

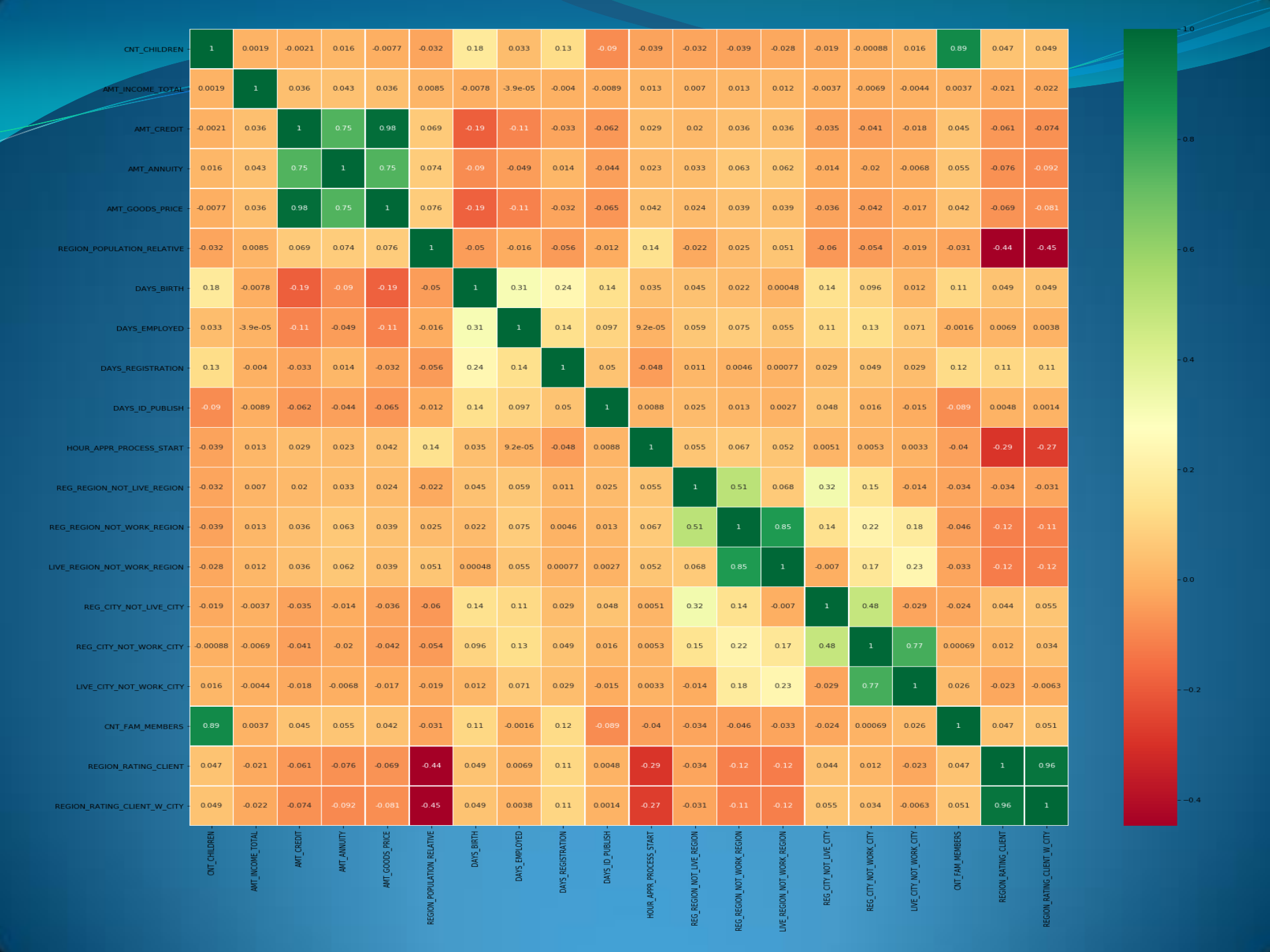
1.Amount annuity has a higher correlation with respect to Amount credit meaning clients higher credit amount tend to take higher EMI.
2.client's permanent address does not match contact address(REG_REGION_NOT_WORK_REGION) has a higher correlation with client's contact address does not match work address(LIVE_REGION_NOT_WORK_REGION).
3.client's permanent address does not match work address(REG_CITY_NOT_WORK_CITY) has a higher correlation with client's contact address does not match work address(LIVE_CITY_NOT_WORK_CITY).
4.Amount annuity has a higher correlation with respect to Amount Goods price meaning clients with higher good's amount tend to take higher EMI.
5.Amount credit has a higher correlation with respect to Amount Goods price meaning clients with higher credit amount tend to take higher consumer good's amount.
6.Region rating client is inversely proportional to the client lives in more populated region.
7.Region rating client has a higher correlation to region where client lives with taking city into account.
8.family members clients have a higher correlation to Number of children the client has.


CNT_CHILDREN 1 -0.0098 -0.019 -0.0076 -0.022 -0.03 0.24 0.063 0.16 -0.12 -0.033 -0.023 -0.017 -0.0069 -0.0016 0.01 0.018 0.89 0.037 0.037
AMT_INCOME_TOTAL -0.0098 1 0.33 0.4 0.33 0.17 -0.046 -0.03 0.035 -0.026 0.056 0.065 0.13 0.12 -0.0043 -0.02 -0.011 -0.0031 -0.18 -0.2
AMT_CREDIT -0.019 0.33 1 0.76 0.99 0.1 -0.15 -0.088 -0.015 -0.035 0.04 0.021 0.051 0.052 -0.038 -0.039 -0.015 0.037 -0.11 -0.12
AMT_ANNUITY -0.0076 0.4 0.76 1 0.77 0.12 -0.092 -0.052 0.0098 -0.026 0.034 0.04 0.076 0.072 -0.019 -0.029 -0.014 0.041 -0.13 -0.15
AMT_GOODS_PRICE -0.022 0.33 0.99 0.77 1 0.11 -0.15 -0.089 -0.012 -0.037 0.05 0.024 0.052 0.052 -0.037 -0.039 -0.016 0.035 -0.11 -0.12
REGION_POPULATION_RELATIVE -0.03 0.17 0.1 0.12 0.11 1 -0.041 0.0087 -0.06 -0.004 0.17 -0.0033 0.066 0.092 -0.054 -0.049 -0.017 -0.029 -0.54 -0.54
DAYS_BIRTH 0.24 -0.046 -0.15 -0.092 -0.15 -0.041 1 0.35 0.3 0.11 0.049 0.058 0.039 0.014 0.17 0.11 0.031 0.18 0.037 0.037
DAYS_EMPLOYED 0.063 -0.03 -0.088 -0.052 -0.089 0.0087 0.35 1 0.17 0.087 0.022 0.057 0.082 0.063 0.11 0.13 0.075 0.032 -0.017 -0.016
DAYS_REGISTRATION 0.16 0.035 -0.015 0.0098 -0.012 -0.06 0.3 0.17 1 0.03 -0.032 0.021 0.018 0.0093 0.047 0.055 0.031 0.16 0.094 0.087
DAYS_ID_PUBLISH -0.12 -0.026 -0.035 -0.026 -0.037 -0.004 0.11 0.087 0.03 1 0.0054 0.027 0.021 0.0089 0.053 0.033 0.0018 -0.11 0.0036 0.0013
HOUR_APPR_PROCESS_START -0.033 0.056 0.04 0.034 0.05 0.17 0.049 0.022 -0.032 0.0054 1 0.055 0.072 0.058 0.011 -0.0015 -0.0051 -0.037 -0.28 -0.26
REG_REGION_NOT_LIVE_REGION -0.023 0.065 0.021 0.04 0.024 -0.0033 0.058 0.057 0.021 0.027 0.055 1 0.46 0.09 0.34 0.14 0.0035 -0.027 -0.044 -0.041
REG_REGION_NOT_WORK_REGION -0.017 0.13 0.051 0.076 0.052 0.066 0.039 0.082 0.018 0.021 0.072 0.46 1 0.86 0.15 0.22 0.18 -0.022 -0.15 -0.15
LIVE_REGION_NOT_WORK_REGION -0.0069 0.12 0.052 0.072 0.052 0.092 0.014 0.063 0.0093 0.0089 0.058 0.09 0.86 1 0.015 0.17 0.22 -0.0094 -0.16 -0.15
REG_CITY_NOT_LIVE_CITY -0.0016 -0.0043 -0.038 -0.019 -0.037 -0.054 0.17 0.11 0.047 0.053 0.011 0.34 0.15 0.015 1 0.44 0.012 -0.009 0.039 0.051
REG_CITY_NOT_WORK_CITY 0.01 -0.02 -0.039 -0.029 -0.039 -0.049 0.11 0.13 0.055 0.033 -0.0015 0.14 0.22 0.17 0.44 1 0.82 0.013 0.017 0.04
LIVE_CITY_NOT_WORK_CITY 0.018 -0.011 -0.015 -0.014 -0.016 -0.017 0.031 0.075 0.031 0.0018 -0.0051 0.0035 0.18 0.22 0.012 0.82 1 0.029 -0.014 0.0048
CNT_FAM_MEMBERS 0.89 -0.0031 0.037 0.041 0.035 -0.029 0.18 0.032 0.16 -0.11 -0.037 -0.027 -0.022 -0.0094 -0.009 0.013 0.029 1 0.041 0.042
REGION_RATING_CLIENT 0.037 -0.18 -0.11 -0.13 -0.11 -0.54 0.037 -0.017 0.094 0.0036 -0.28 -0.044 -0.15 -0.16 0.039 0.017 -0.014 0.041 1 0.95
REGION_RATING_CLIENT_W_CITY 0.037 -0.2 -0.12 -0.15 -0.12 -0.54 0.037 -0.016 0.087 0.0013 -0.26 -0.041 -0.15 -0.15 0.051 0.04 0.0048 0.042 0.95 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

# Correlation for numerical columns for Target 1

## Correlation for numerical columns for Target 1.

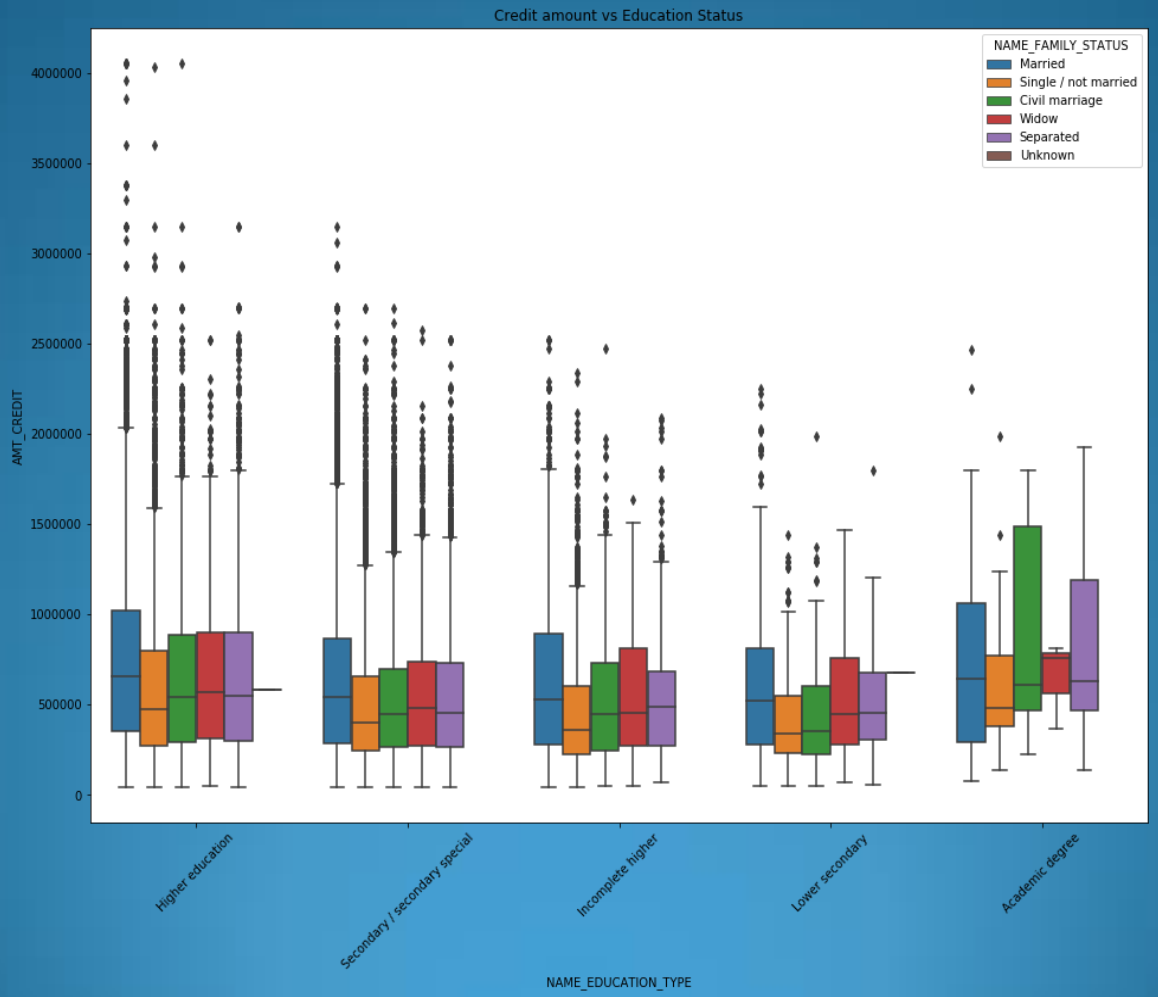
1.Amount annuity has a higher correlation with respect to Amount credit meaning clients higher credit amount tend to take higher EMI.
2.client's permanent address does not match contact address(REG_REGION_NOT_WORK_REGION) has a higher correlation with client's contact address does not match work address(LIVE_REGION_NOT_WORK_REGION).
3.client's permanent address does not match work address(REG_CITY_NOT_WORK_CITY) has a higher correlation with client's contact address does not match work address(LIVE_CITY_NOT_WORK_CITY).
4.Amount annuity has a higher correlation with respect to Amount Goods price meaning clients with higher good's amount tend to take higher EMI.
5.Amount credit has a higher correlation with respect to Amount Goods price meaning clients with higher credit amount tend to take higher consumer good's price.
6.Region rating client is inversely proportional to the client lives in more populated region.
7.Region rating client has a higher correlation to region where client lives with taking city into account.
8.family members clients have a higher correlation to Number of children the client has.
9.region where client lives with taking city into account is inversely proportional to the client lives in more populated region.


CNT_CHILDREN 1 0.0019 -0.0021 0.016 -0.0077 -0.032 0.18 0.033 0.13 -0.09 -0.039 -0.032 -0.039 -0.028 -0.019 -0.00088 0.016 0.89 0.047 0.049
AMT_INCOME_TOTAL 0.0019 1 0.036 0.043 0.036 0.0085 -0.0078 -3.9e-05 -0.004 -0.0089 0.013 0.007 0.013 0.012 -0.0037 -0.0069 -0.0044 0.0037 -0.021 -0.022
AMT_CREDIT -0.0021 0.036 1 0.75 0.98 0.069 -0.19 -0.11 -0.033 -0.062 0.029 0.02 0.036 0.036 -0.035 -0.041 -0.018 0.045 -0.061 -0.074
AMT_ANNUITY 0.016 0.043 0.75 1 0.75 0.074 -0.09 -0.049 0.014 -0.044 0.023 0.033 0.063 0.062 -0.014 -0.02 -0.0068 0.055 -0.076 -0.092
AMT_GOODS_PRICE -0.0077 0.036 0.98 0.75 1 0.076 -0.19 -0.11 -0.032 -0.065 0.042 0.024 0.039 0.039 -0.036 -0.042 -0.017 0.042 -0.069 -0.081
REGION_POPULATION_RELATIVE -0.032 0.0085 0.069 0.074 0.076 1 -0.05 -0.016 -0.056 -0.012 0.14 -0.022 0.025 0.051 -0.06 -0.054 -0.019 -0.031 -0.44 -0.45
DAYS_BIRTH 0.18 -0.0078 -0.19 -0.09 -0.19 -0.05 1 0.31 0.24 0.14 0.035 0.045 0.022 0.00048 0.14 0.096 0.012 0.11 0.049 0.049
DAYS_EMPLOYED 0.033 -3.9e-05 -0.11 -0.049 -0.11 -0.016 0.31 1 0.14 0.097 9.2e-05 0.059 0.075 0.055 0.11 0.13 0.071 -0.0016 0.0069 0.0038
DAYS_REGISTRATION 0.13 -0.004 -0.033 0.014 -0.032 -0.056 0.24 0.14 1 0.05 -0.048 0.011 0.0046 0.00077 0.029 0.049 0.029 0.12 0.11 0.11
DAYS_ID_PUBLISH -0.09 -0.0089 -0.062 -0.044 -0.065 -0.012 0.14 0.097 0.05 1 0.0088 0.025 0.013 0.0027 0.048 0.016 -0.015 -0.089 0.0048 0.0014
HOUR_APPR_PROCESS_START -0.039 0.013 0.029 0.023 0.042 0.14 0.035 9.2e-05 -0.048 0.0088 1 0.055 0.067 0.052 0.0051 0.0053 0.0033 -0.04 -0.29 -0.27
REG_REGION_NOT_LIVE_REGION -0.032 0.007 0.02 0.033 0.024 -0.022 0.045 0.059 0.011 0.025 0.055 1 0.51 0.068 0.32 0.15 -0.014 -0.034 -0.034 -0.031
REG_REGION_NOT_WORK_REGION -0.039 0.013 0.036 0.063 0.039 0.025 0.022 0.075 0.0046 0.013 0.067 0.51 1 0.85 0.14 0.22 0.18 -0.046 -0.12 -0.11
LIVE_REGION_NOT_WORK_REGION -0.028 0.012 0.036 0.062 0.039 0.051 0.00048 0.055 0.00077 0.0027 0.052 0.068 0.85 1 -0.007 0.17 0.23 -0.033 -0.12 -0.12
REG_CITY_NOT_LIVE_CITY -0.019 -0.0037 -0.035 -0.014 -0.036 -0.06 0.14 0.11 0.029 0.048 0.0051 0.32 0.14 -0.007 1 0.48 -0.029 -0.024 0.044 0.055
REG_CITY_NOT_WORK_CITY -0.00088 -0.0069 -0.041 -0.02 -0.042 -0.054 0.096 0.13 0.049 0.016 0.0053 0.15 0.22 0.17 0.48 1 0.77 0.00069 0.012 0.034
LIVE_CITY_NOT_WORK_CITY 0.016 -0.0044 -0.018 -0.0068 -0.017 -0.019 0.012 0.071 0.029 -0.015 0.0033 -0.014 0.18 0.23 -0.029 0.77 1 0.026 -0.023 -0.0063
CNT_FAM_MEMBERS 0.89 0.0037 0.045 0.055 0.042 -0.031 0.11 -0.0016 0.12 -0.089 -0.04 -0.034 -0.046 -0.033 -0.024 0.00069 0.026 1 0.047 0.051
REGION_RATING_CLIENT 0.047 -0.021 -0.061 -0.076 -0.069 -0.44 0.049 0.0069 0.11 0.0048 -0.29 -0.034 -0.12 -0.12 0.044 0.012 -0.023 0.047 1 0.96
REGION_RATING_CLIENT_W_CITY 0.049 -0.022 -0.074 -0.092 -0.081 -0.45 0.049 0.0038 0.11 0.0014 -0.27 -0.031 -0.11 -0.12 0.055 0.034 -0.0063 0.051 0.96 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

# Bivariate Analysis for Target 0

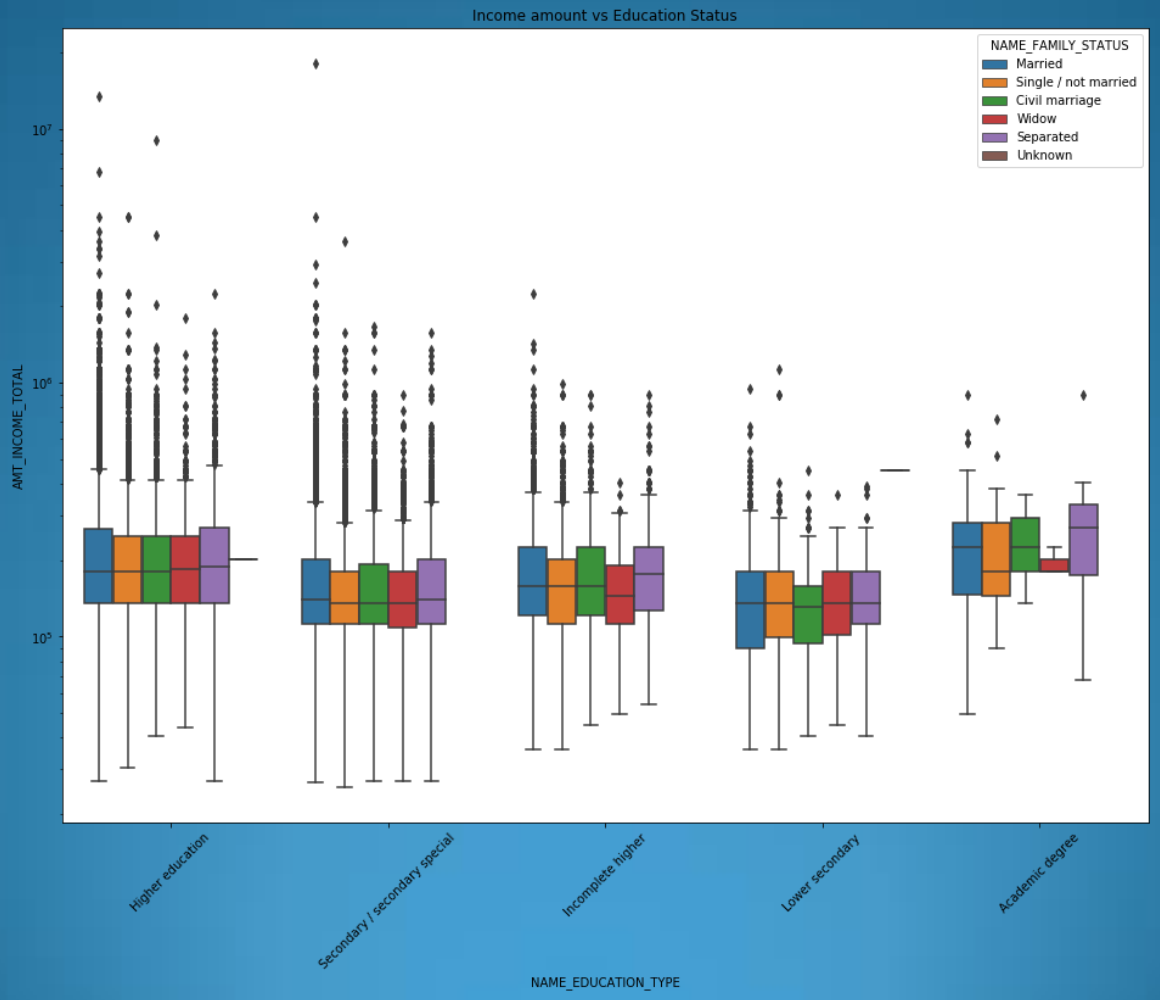
# Target 0 : Credit amount vs Education type

1.Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
2.higher education of family status of 'marriage', 'single','civil marriage', separated are having more outliers.
3.Civil marriage for Academic degree is having most of the credits in the third quartile.



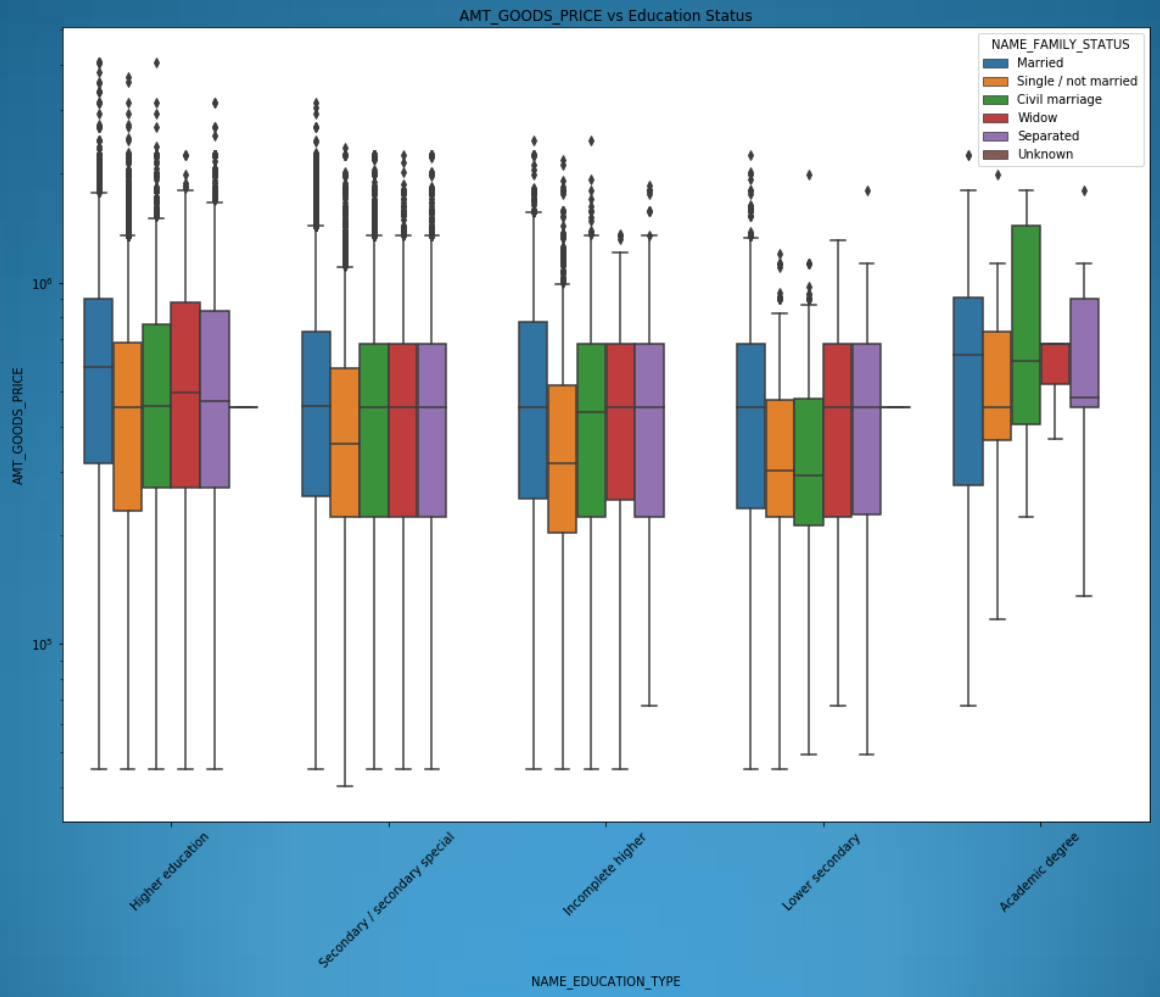
# Target 0 : Income amount vs Education Type

1.for Education type 'Higher education' the income amount is mostly equal with family status. It does contain many outliers.
2.Less outlier are having for Academic degree but there income amount is little higher that Higher education.
3.Lower secondary of civil marriage family status are have less income amount than others.



# Target 0 : Amount Goods Price vs Education Type

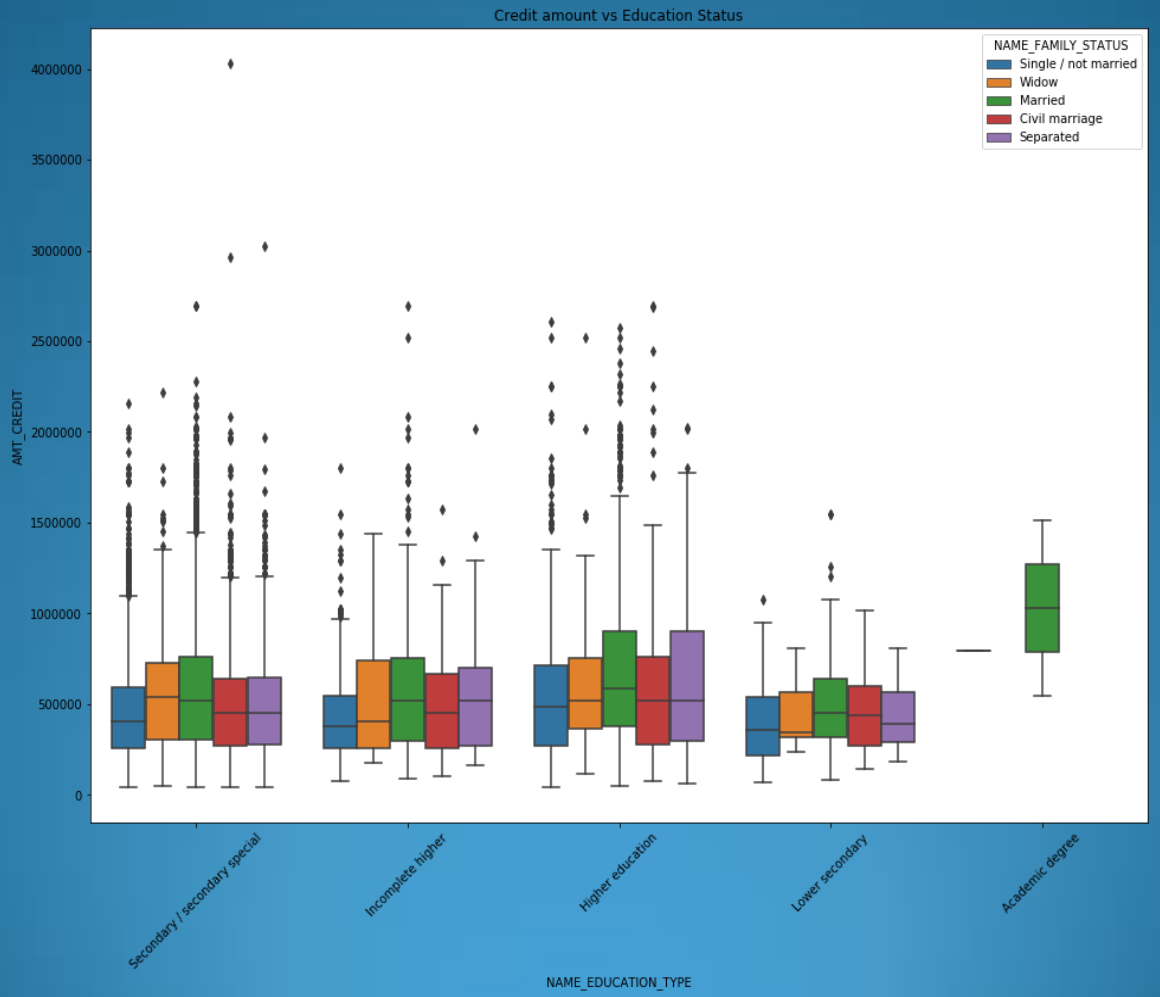
1.For Education type Higher education tend to have more outliers.
2.For education type Academic degree and family status of civil marriage tend to have opted for higher amount Good's price.
3.For education type Academic degree all family status seem to have less number of outliers.



# Bivariate Analysis for Target 1

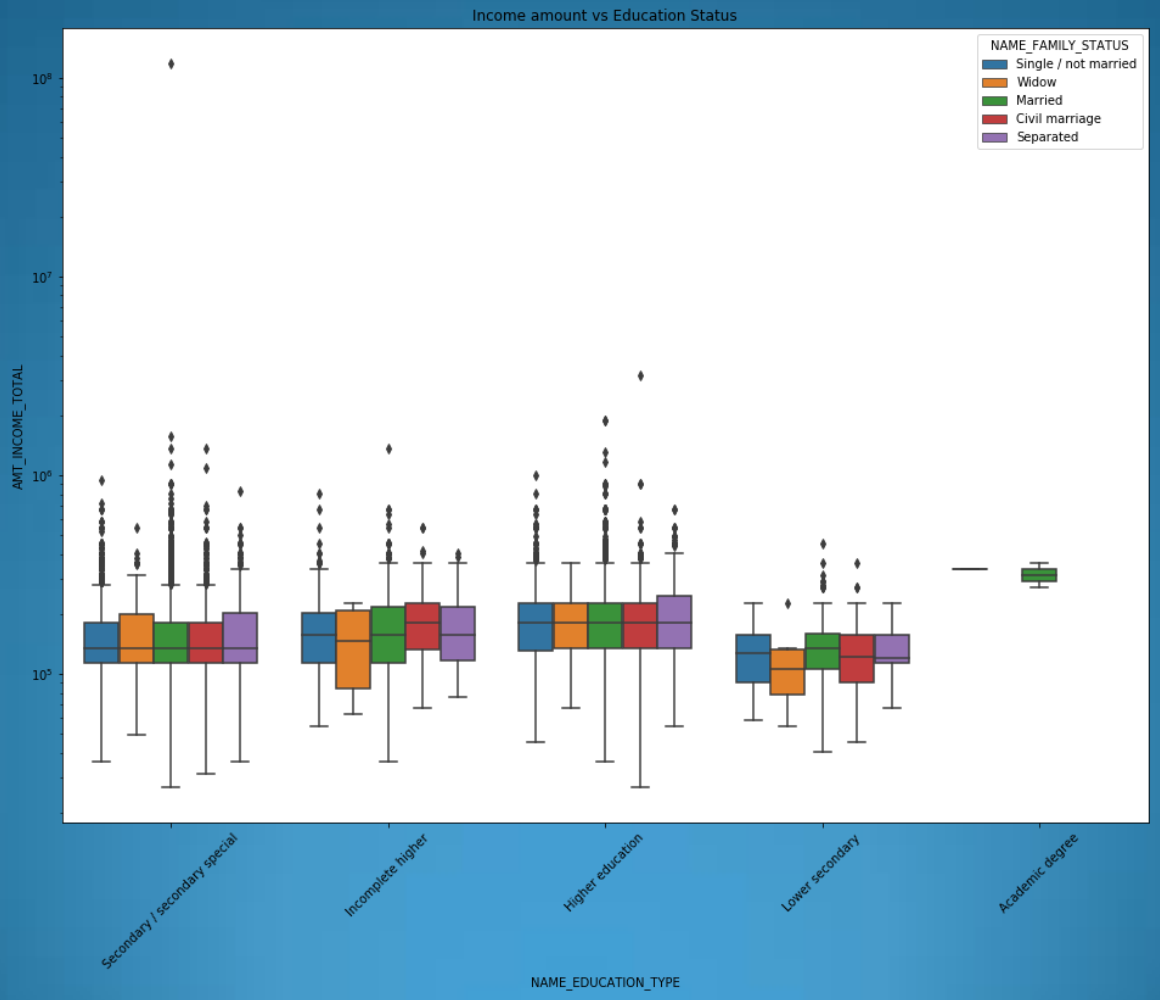
# Target 1 : Credit amount vs Education type

1.From education type Academic degree the family status with single , widow, civil marriage, separated have not opted for Amount credit.
2.Most of the outliers are from Education type 'Higher education' and 'Secondary'.
3.married for Academic degree is having most of the credits compared to other education type and family status.



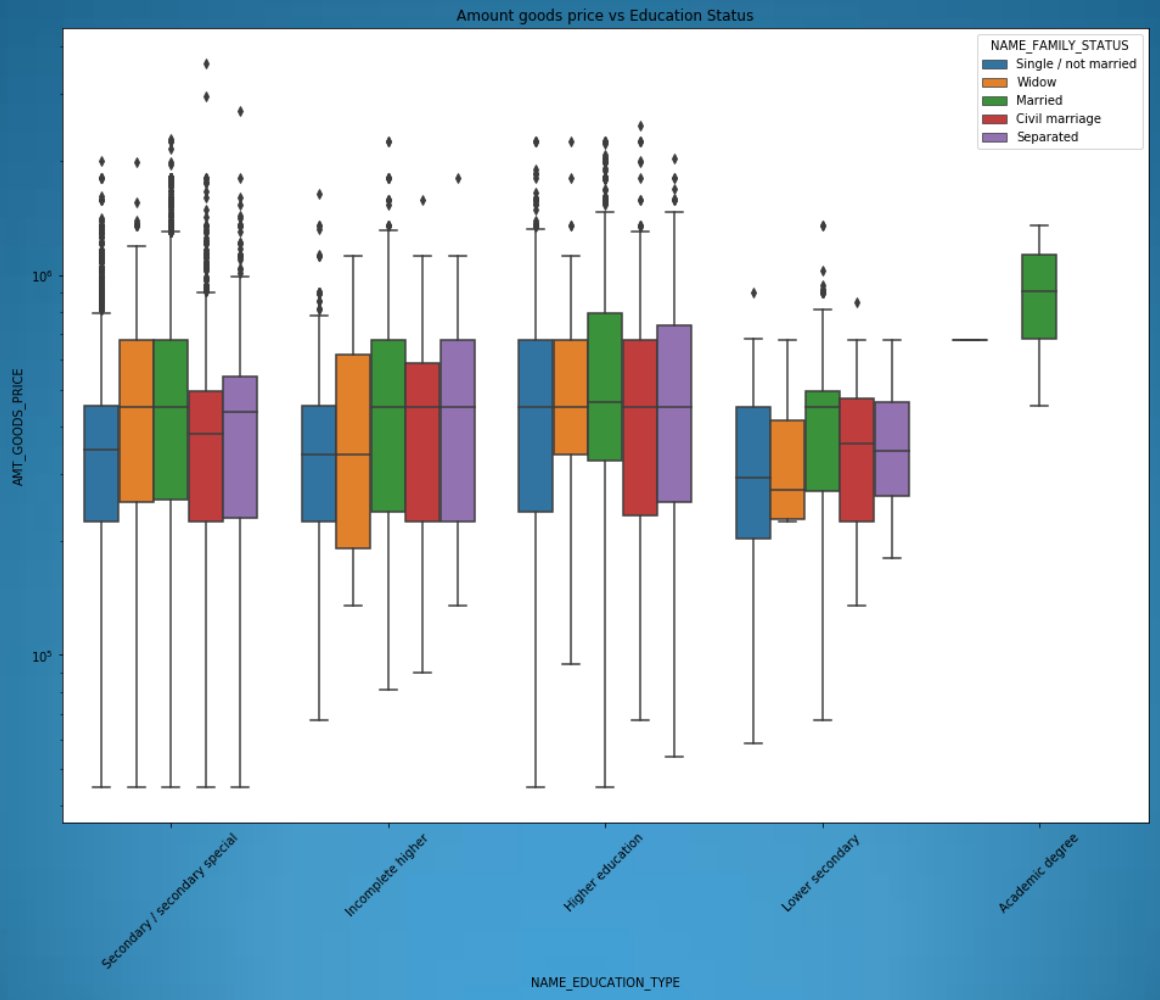
# Target 1 : Income amount vs Education Type

1.For Education type 'Higher education' seem to have the income amount is mostly equal with family status.
2.Less outlier are having for Academic degree but there income amount for family status 'married' is little higher that Higher education.
3.Lower secondary are have less income amount than others.



# Target 1 : Amount Goods Price vs Education Type

1.From above boxplot for Education type 'Higher education' and "secondary /secondary special seem to have more outliers than the rest."
2.Less outlier are having for Academic degree but there income amount for family status 'married' is higher than rest of the education type and rest of the family status.
3.Lower secondary are have less income amount than others.

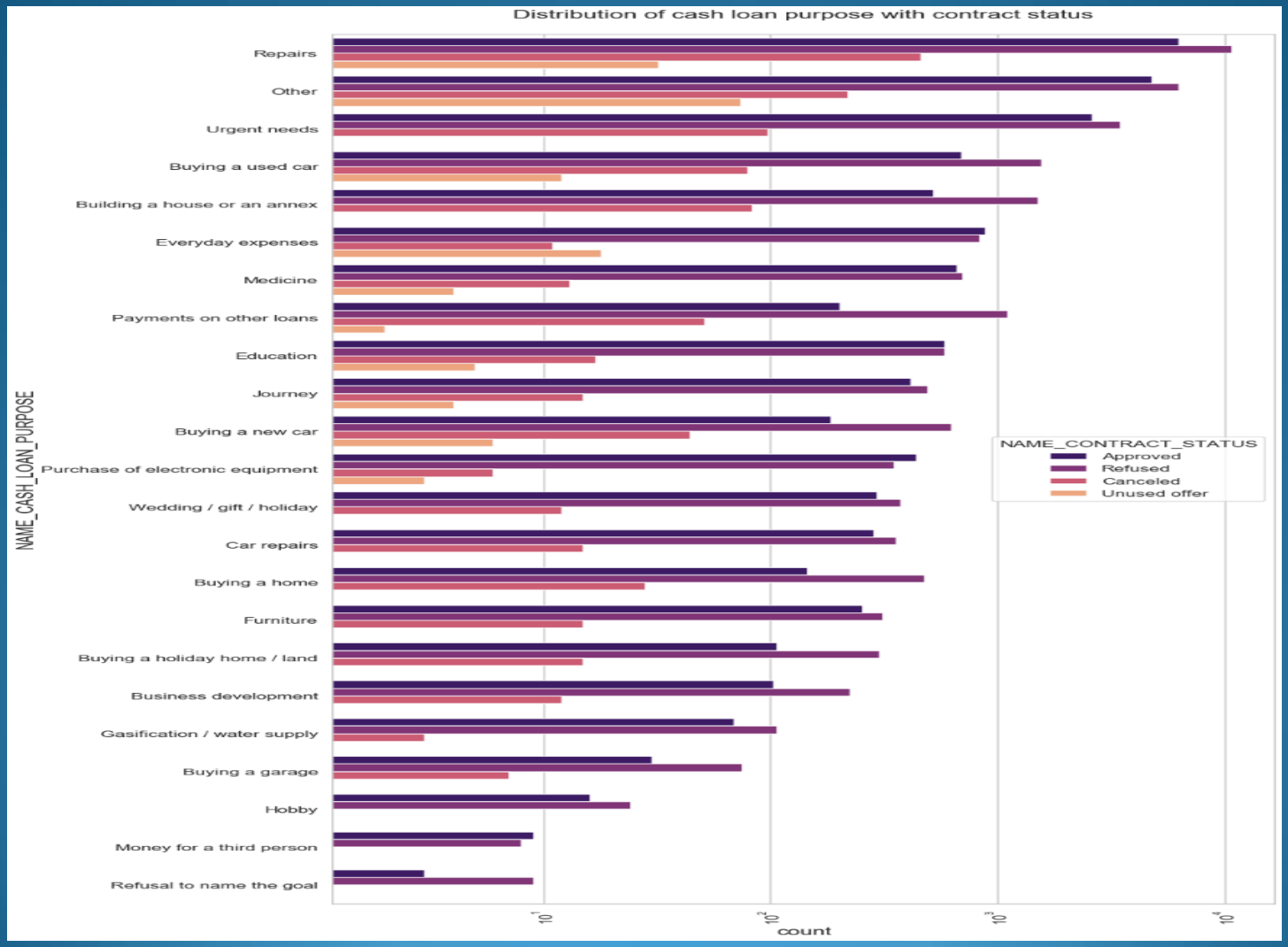


# Previous application data

## Univariate analysis on combined dataset

# Distribution of cash loan purpose with contract status

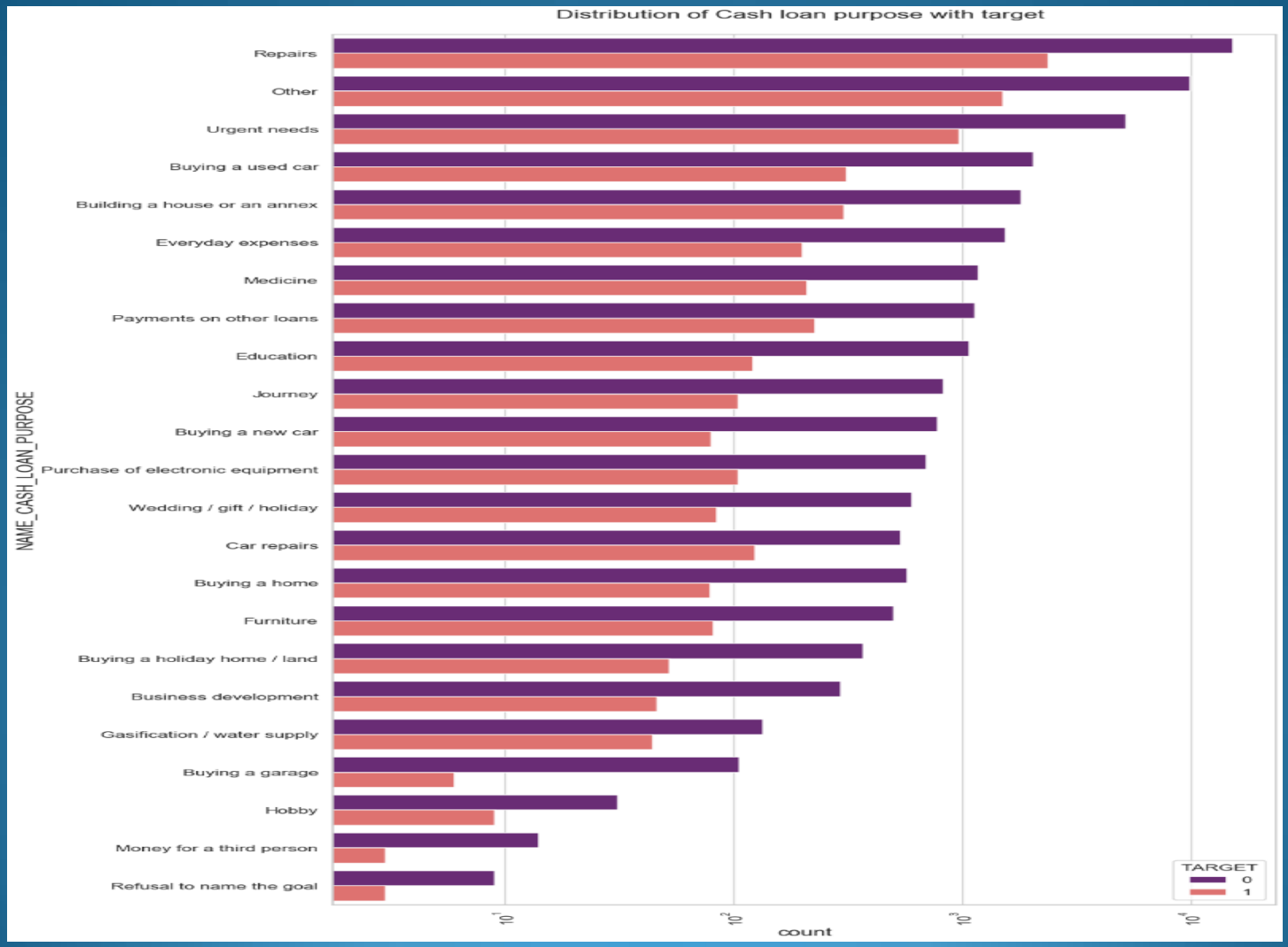
1.Most refusal are for the loans came from purpose 'repairs'.
2.For education purposes we have equal number of approves and refusal.
3.Paying other loans and buying a new car is having significant higher refusals than approvals.
4.Most of the cancellation are for the loan purpose repairs'.



# Distribution of cash loan purpose with Target column

1.Loan purposes with 'Repairs' are facing more difficulties in payment on time.
2.Following Loan purposes that are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education' where loan payment on time is significant higher than facing difficulties.

Hence the Banks can focus on these purposes for which the client is having significant higher chances of repaying the loan amount on time.

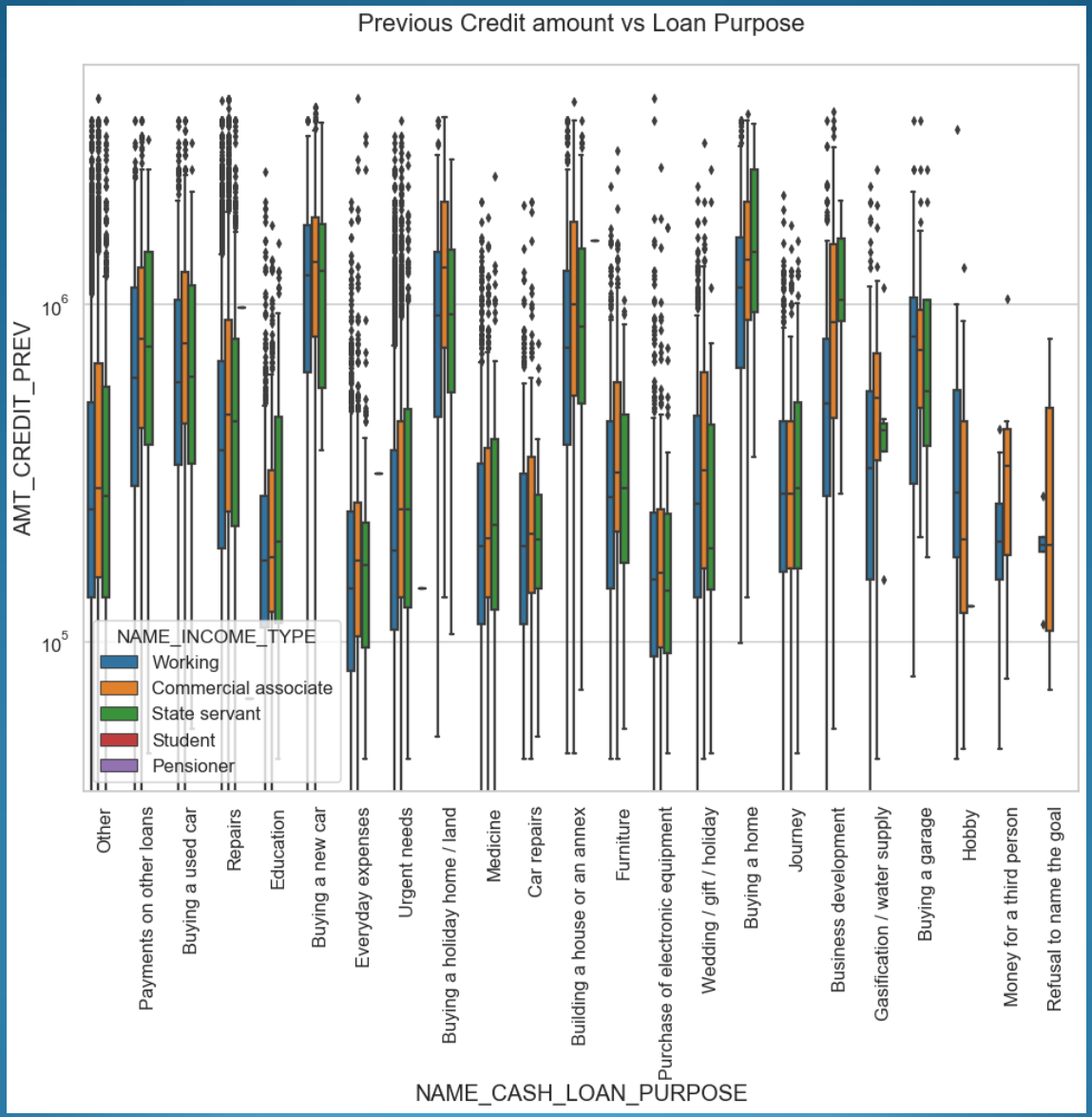


# Previous application data

## Bivariate analysis on combined dataset

# Previous Credit amount vs Loan Purpose

1.The credit amount of Loan purposes like 'Buying a new car', 'Buying a land', 'Buying a home' and 'Building a house' is higher.
2.Income type of state servants have a significant amount of credit applied
3.Money for third person or a Hobby is having less credits applied for.

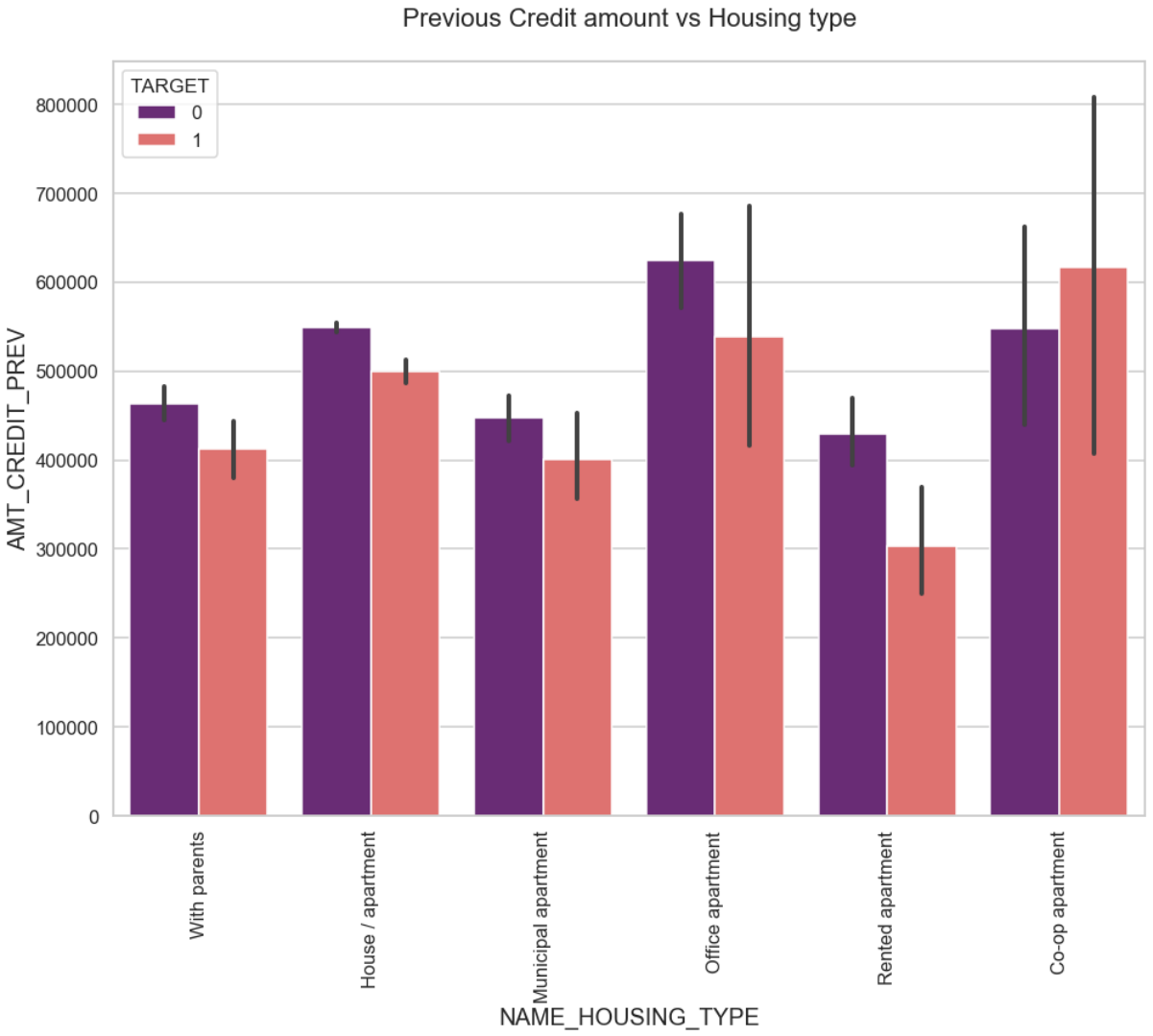


# Previous application data

## Categorical vs Numeric Bivariate Analysis

# Previous Credit amount vs Housing type

1.Office apartment is having higher credit of target 0
2.Co-op apartment is having higher credit of target 1.
From the above inferences we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment.
Bank can focus mostly on housing type with parents or House apartment or municipal apartment for successful payments.



## Final Inferences from the Case study

1. Banks should focus more on contract type ’pensioner’ ,‘Businessman’ and ‘Student’ with housing ‘type other than ‘Co-op apartment’ for successful payments.
2. Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.
3. Also bank should avoid giving loans with purpose ‘Repair’ as it is having a higher number of unsuccessful payments on time.
4. Get as much as clients from housing type ‘With Rented apartment’ as they are having least number of unsuccessful payments.