CREDIT EDA ASSIGNMENT

By KHUSHI SOMAIYA

PROBLEM STATEMENT

Data analysis for bank to identify variables which can act as risk to bank business and avoid business loss based client application profile and previous application profile.

Here, The data given contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

1. The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,

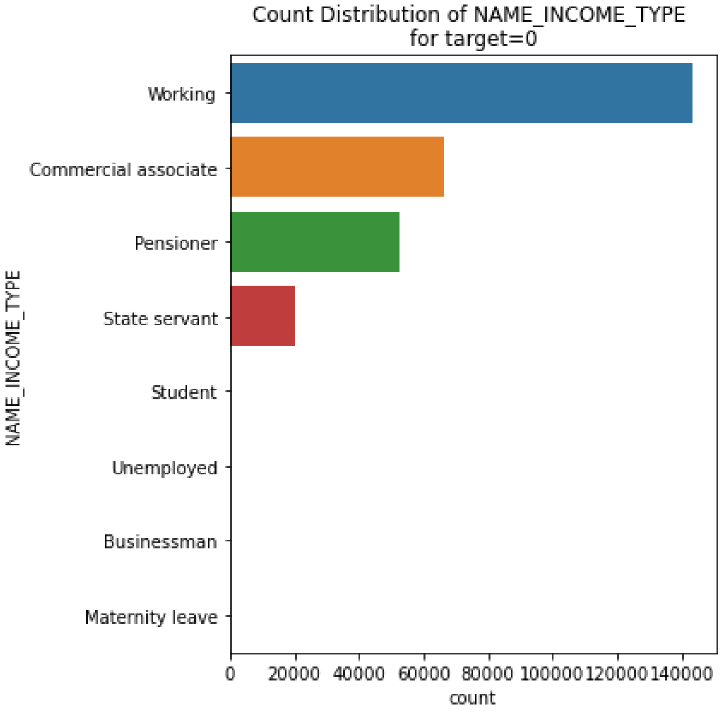
2. All other cases: All other cases when the payment is paid on time.

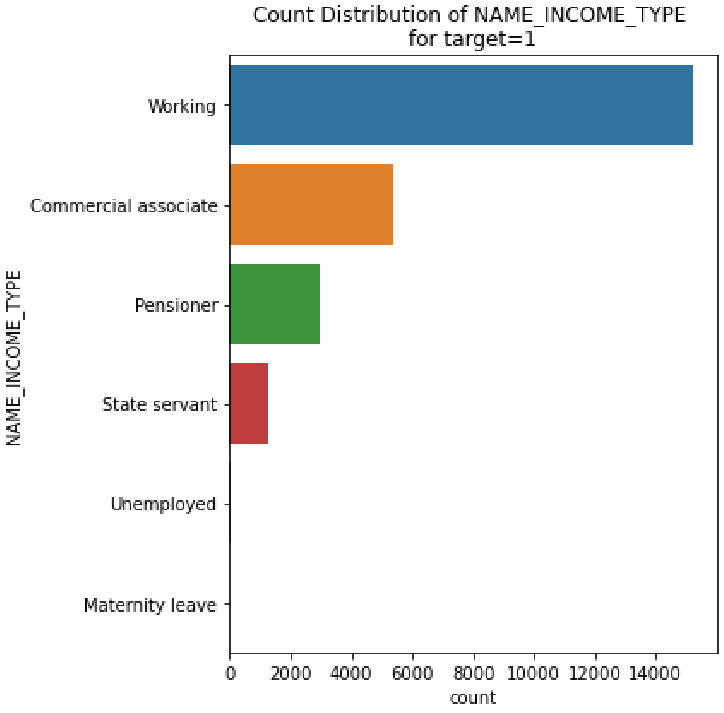
STEPS TAKEN:

1. Importing necessary libraries
2. Reading application dataset and data understanding
3. Data cleaning of application dataset (finding missing values , disguised missing values and outliers and performing necessary imputation)
4. Finding data imbalance and creating 2 new dataframes based on target values which is equal to **11.38**
5. Performing univariate,bivariate and multivariate analysis on both dataframes
6. Reading previous application dataset and data understanding
7. Data cleaning of previous application dataset (finding missing values , disguised missing values and performing necessary imputation)
8. Merging application and previous application dataset and removing unwanted columns
9. Performing bivariate analysis on merged dataset

UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES FOR EACH TARGET

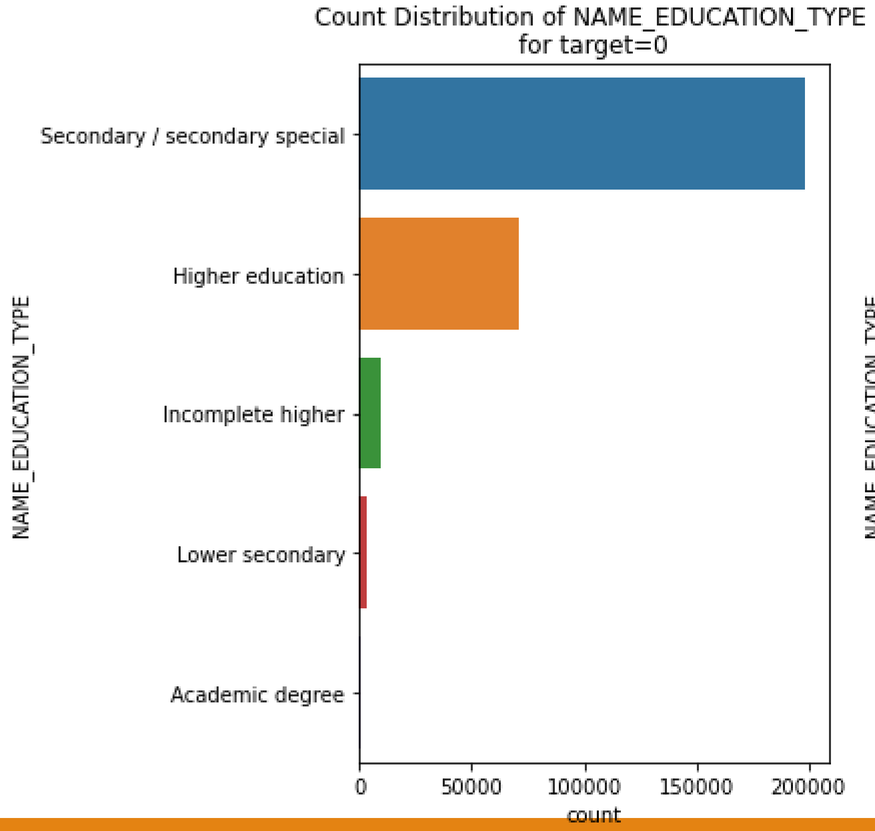
From graph we can observe that most of the people who applied for loans are working.

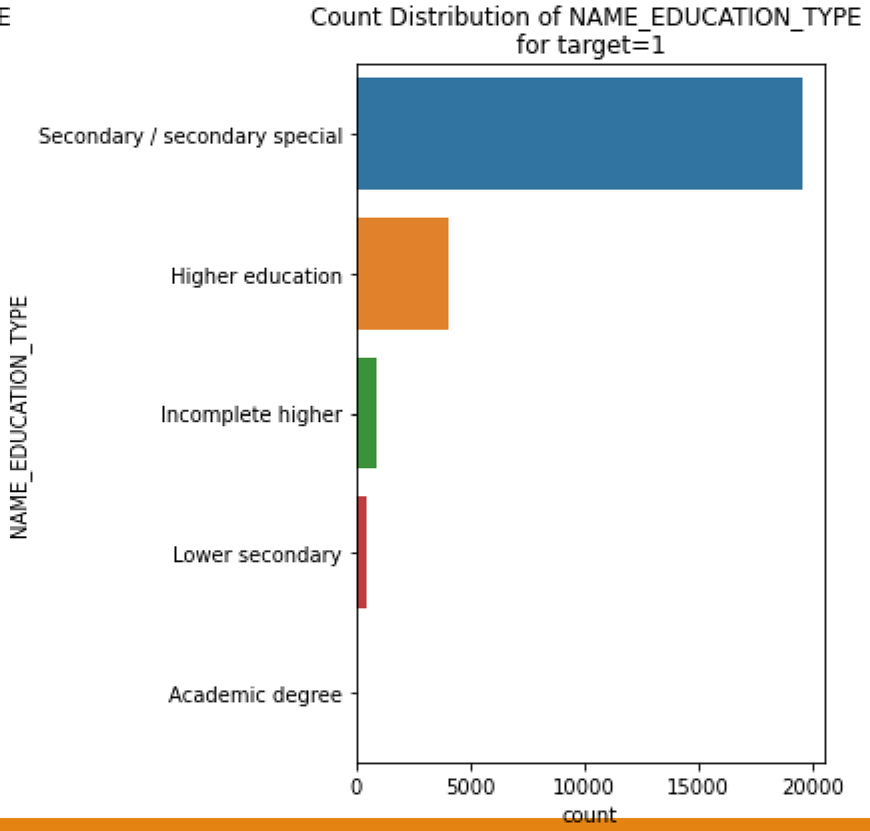

Count Distribution of NAME_INCOME_TYPE for target=0


Count Distribution of NAME_INCOME_TYPE for target=1

UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES FOR EACH TARGET

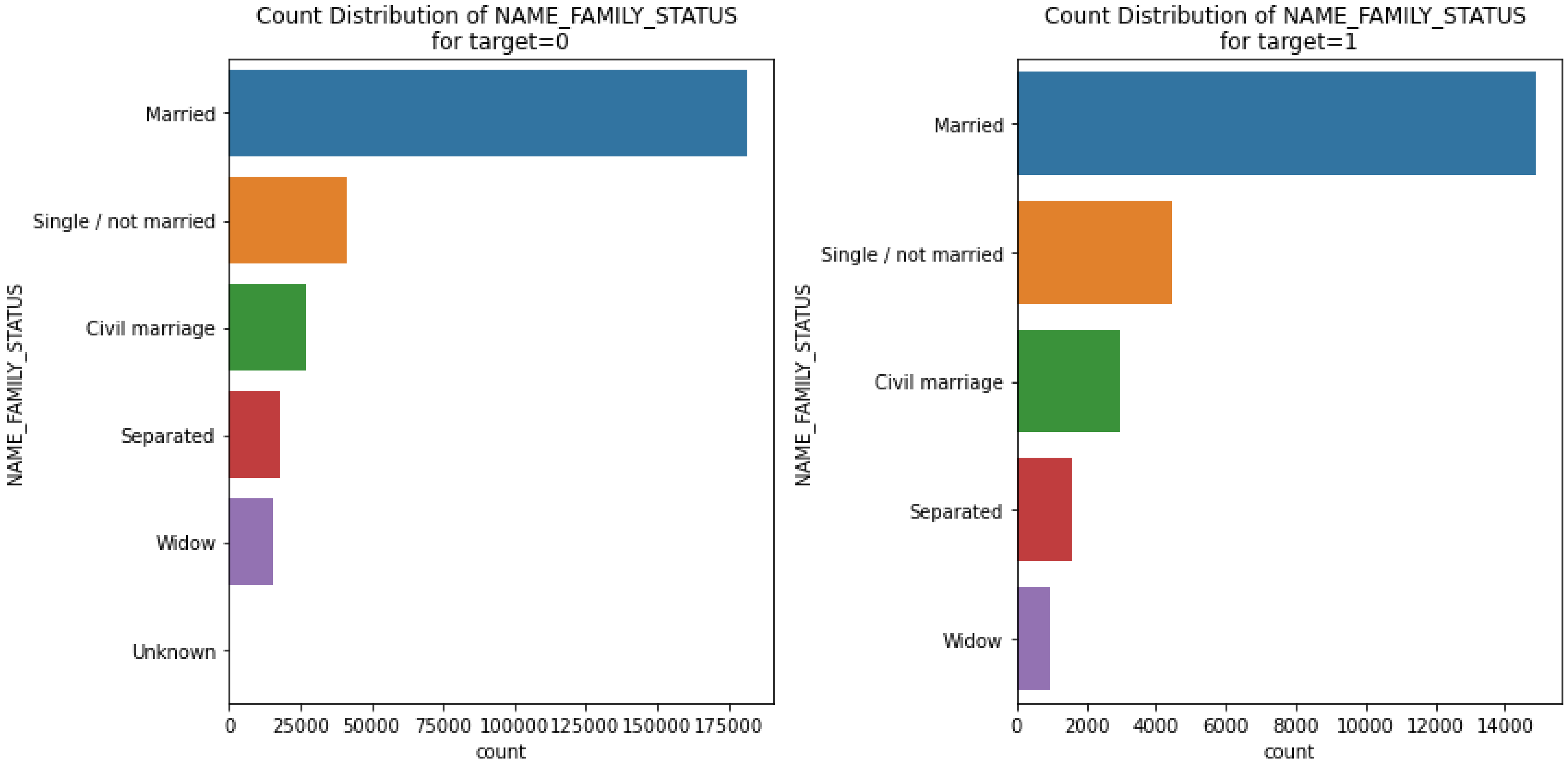
From below graph we can observe that most of the people who applied for loans have secondary/secondary special education.





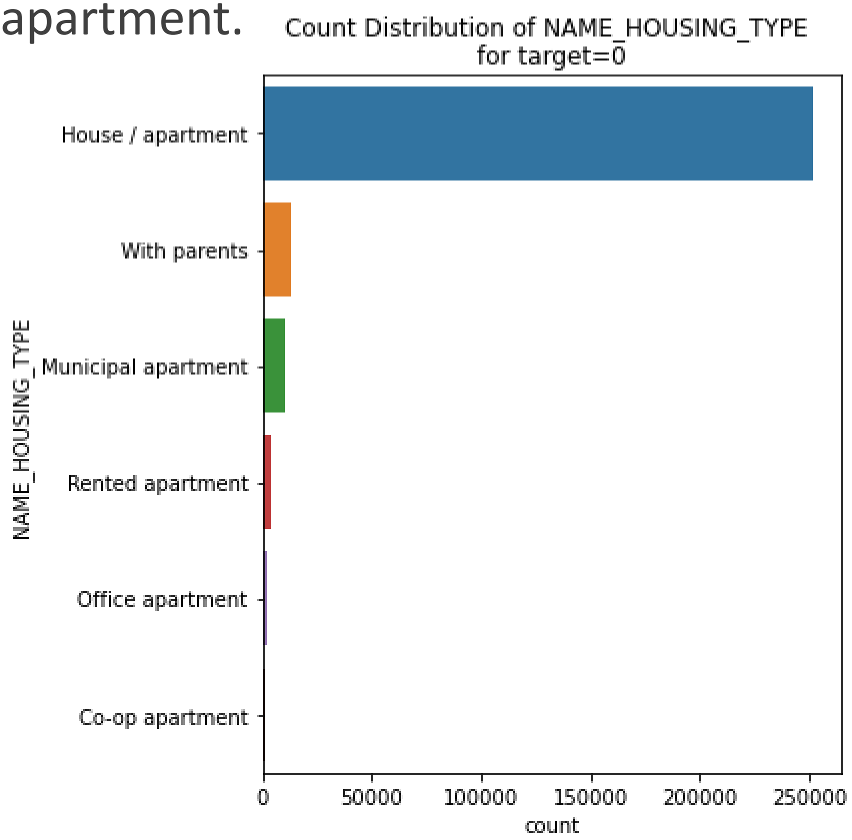
UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES FOR EACH TARGET

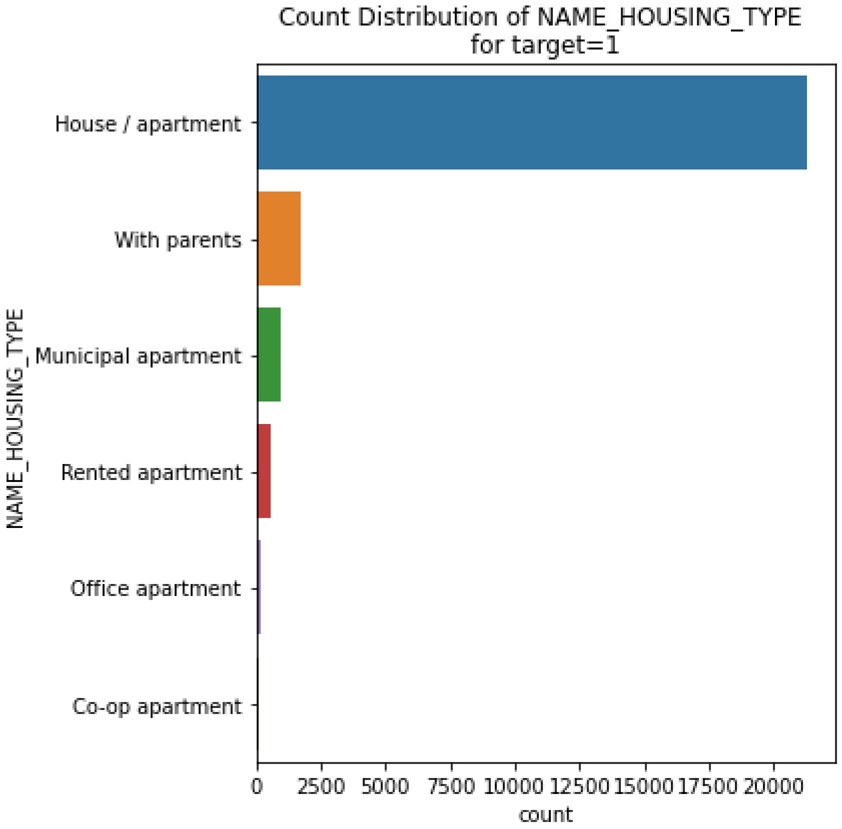
From below graph we can observe that most of the people who applied for loans are married.



UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES FOR EACH TARGET

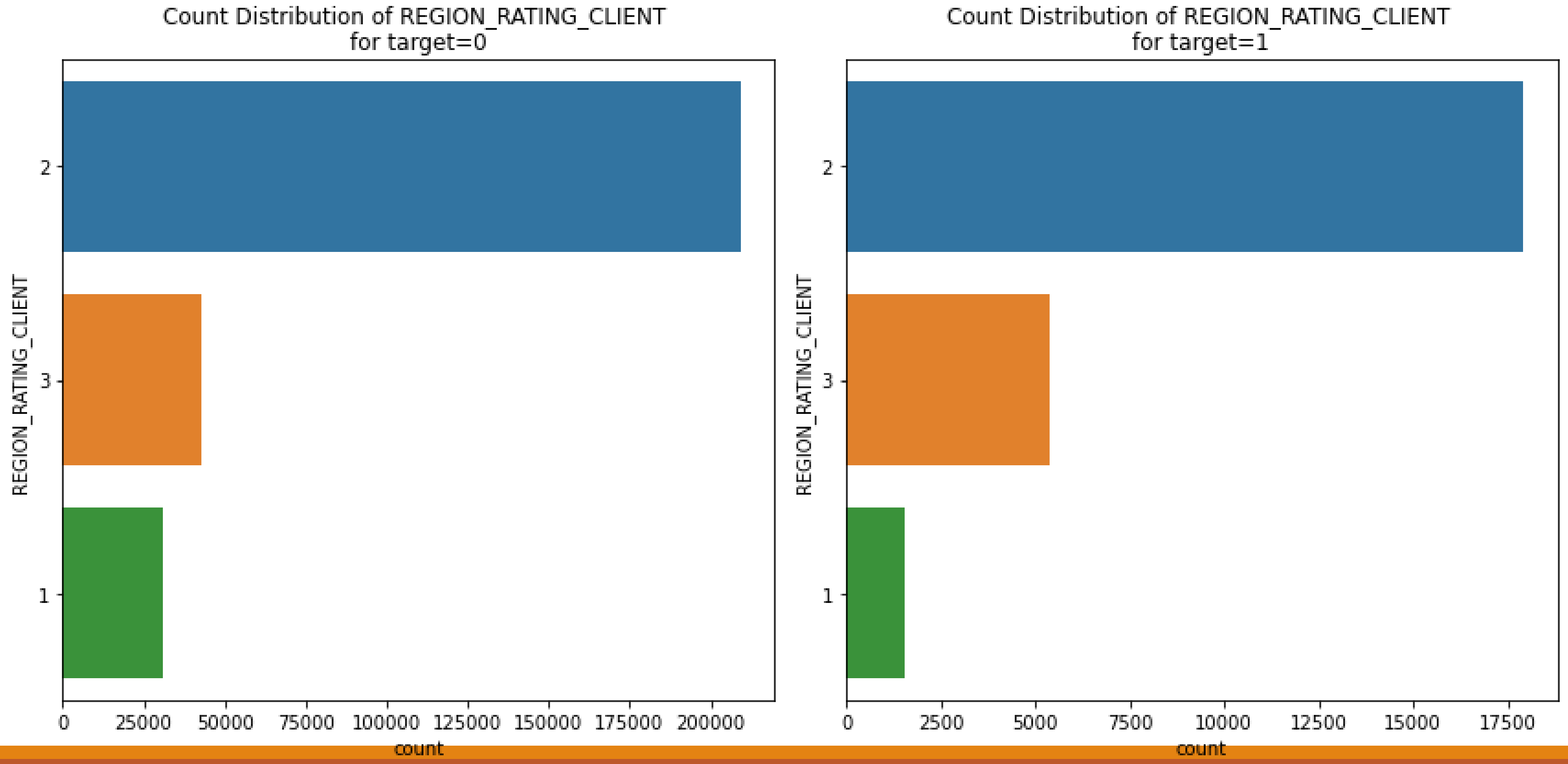
From below graph we can observe that most of the people who applied for loans live in house/apartment.





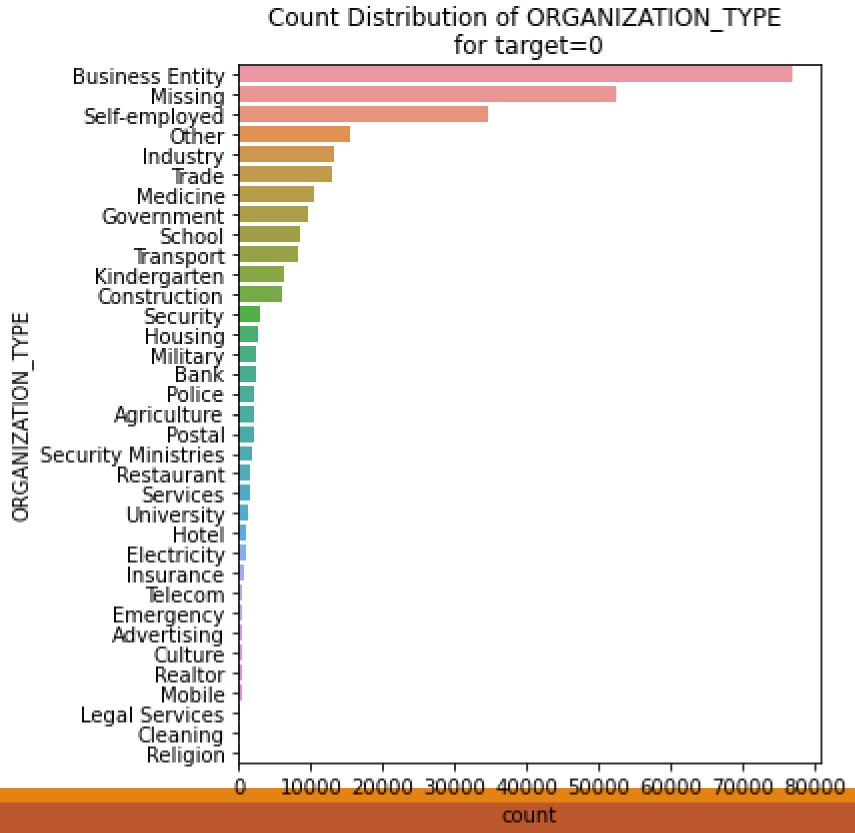
UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES FOR EACH TARGET

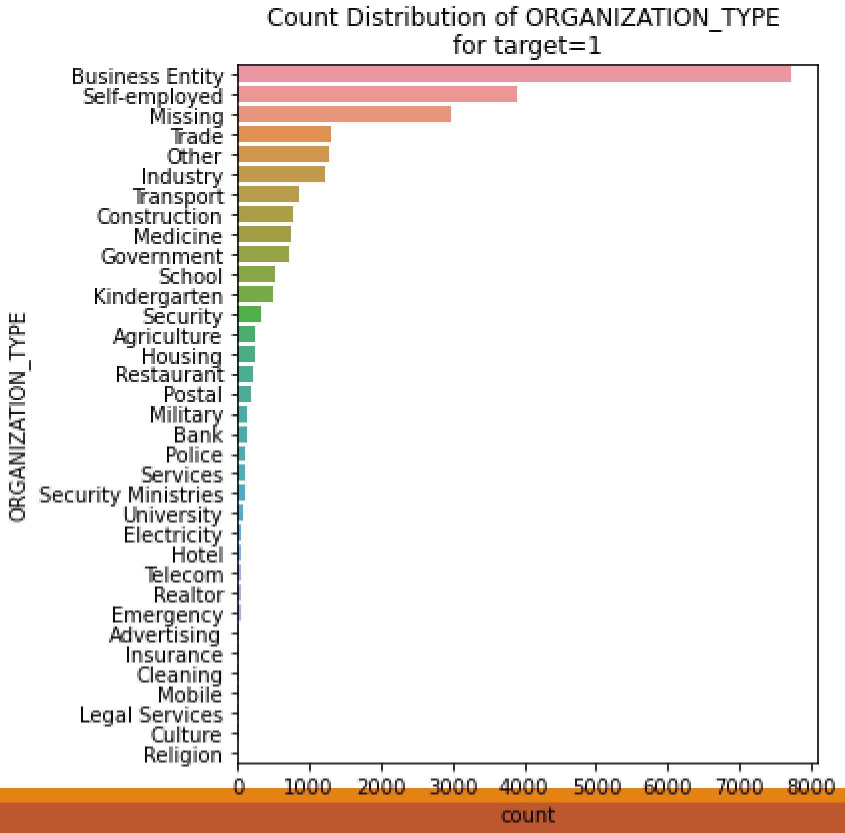
#From below graph we can observe that most of the people who applied for loans are from region 2 whereas least loans are applied from region 1



UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES FOR EACH TARGET

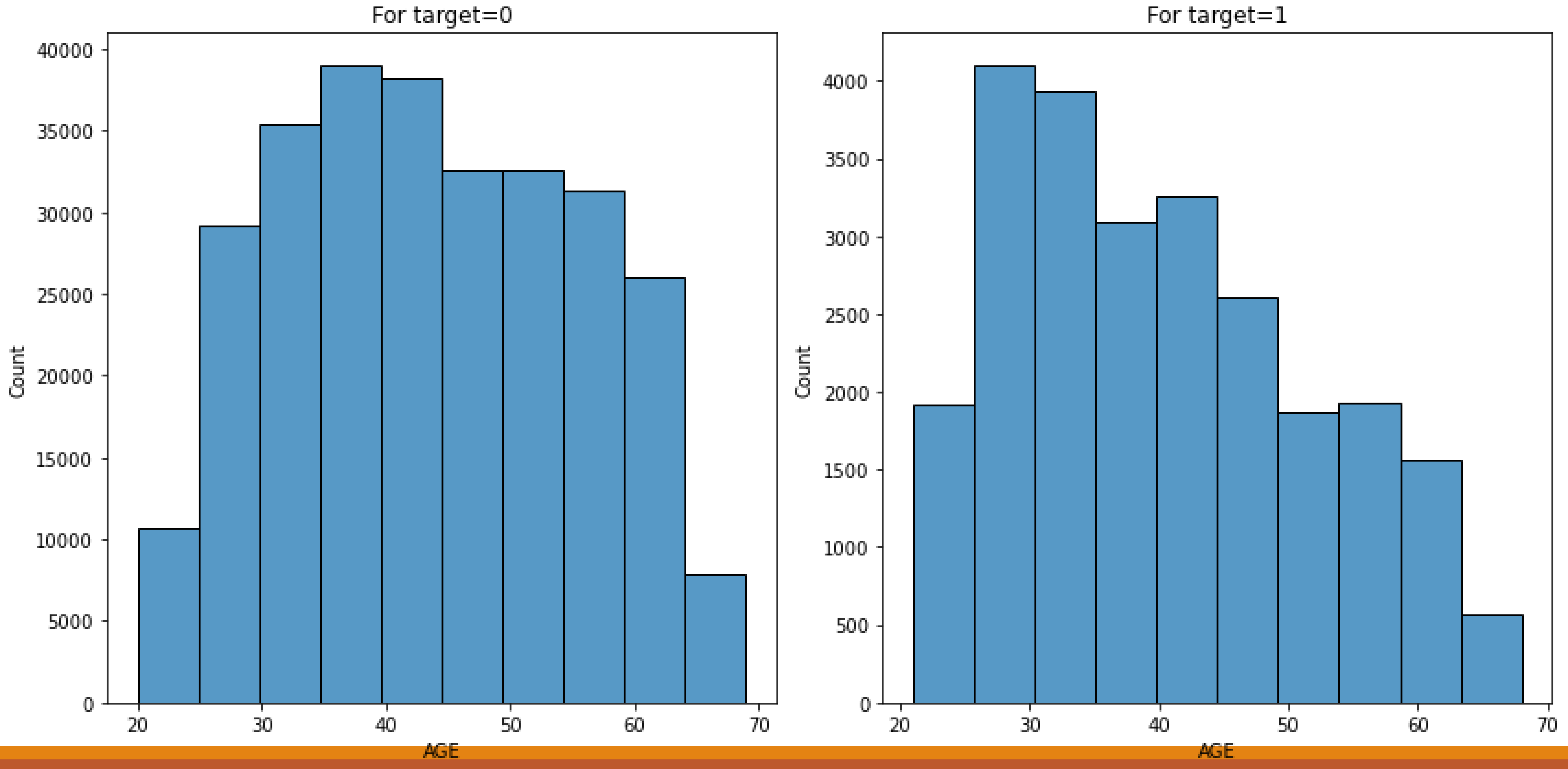
#From below graph we can observe that most of the people who applied for loans are from Business Entity whereas least amount of loans are applied from organization of religion





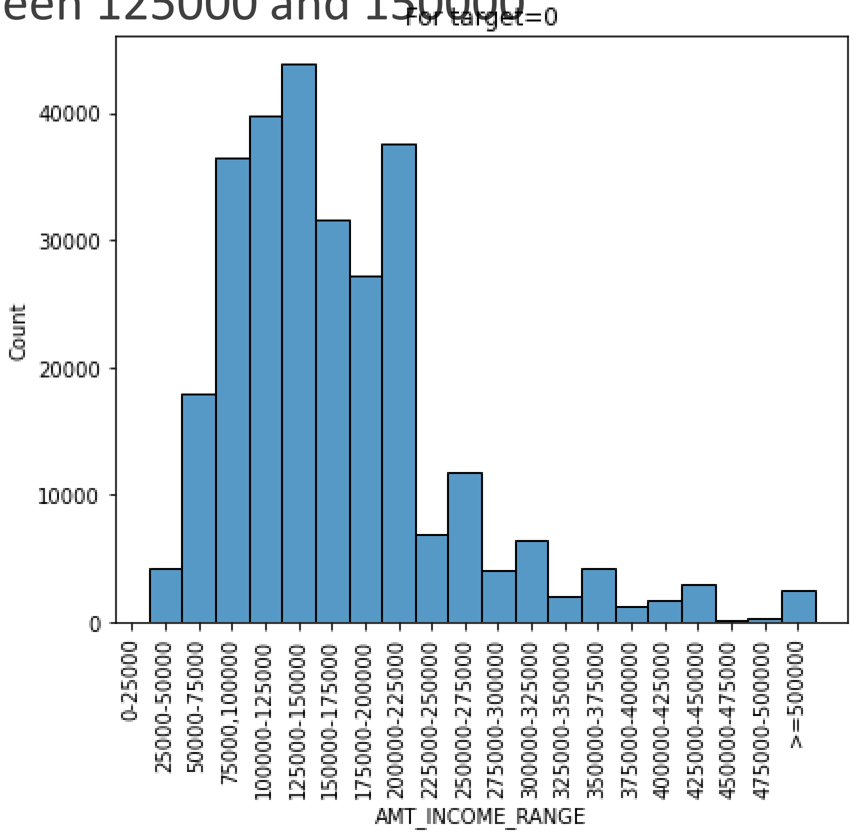
UNIVARIATE ANALYSIS FOR ORDERED CATEGORICAL VARIABLES

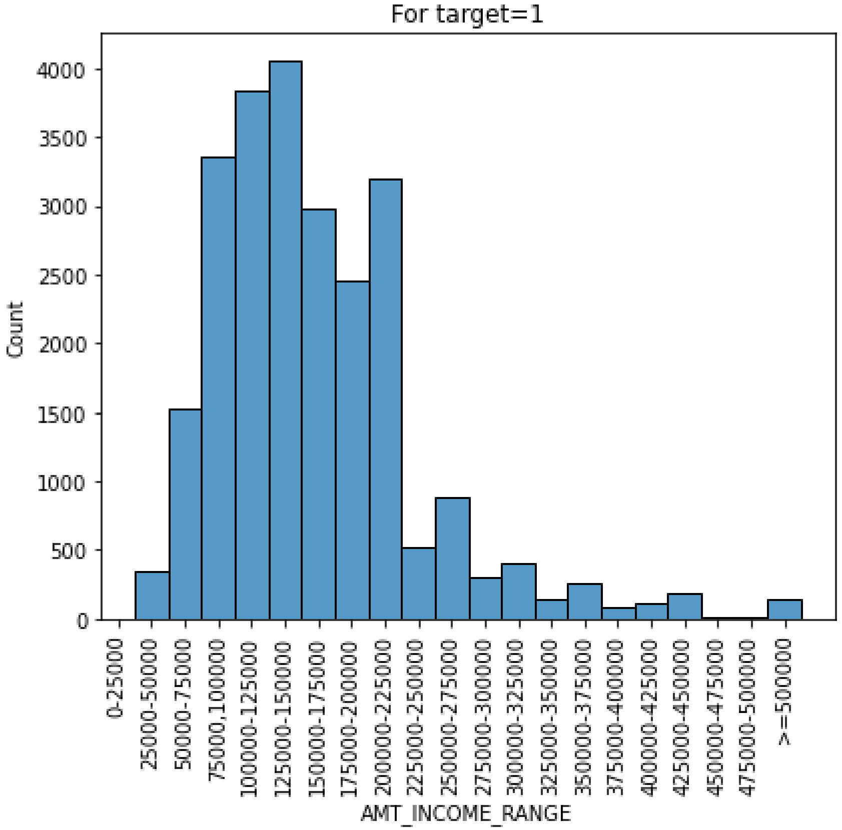
#From above graph we can observe that most of the loans are defaulted by people aged between 25 and 40



UNIVARIATE ANALYSIS FOR ORDERED CATEGORICAL VARIABLES

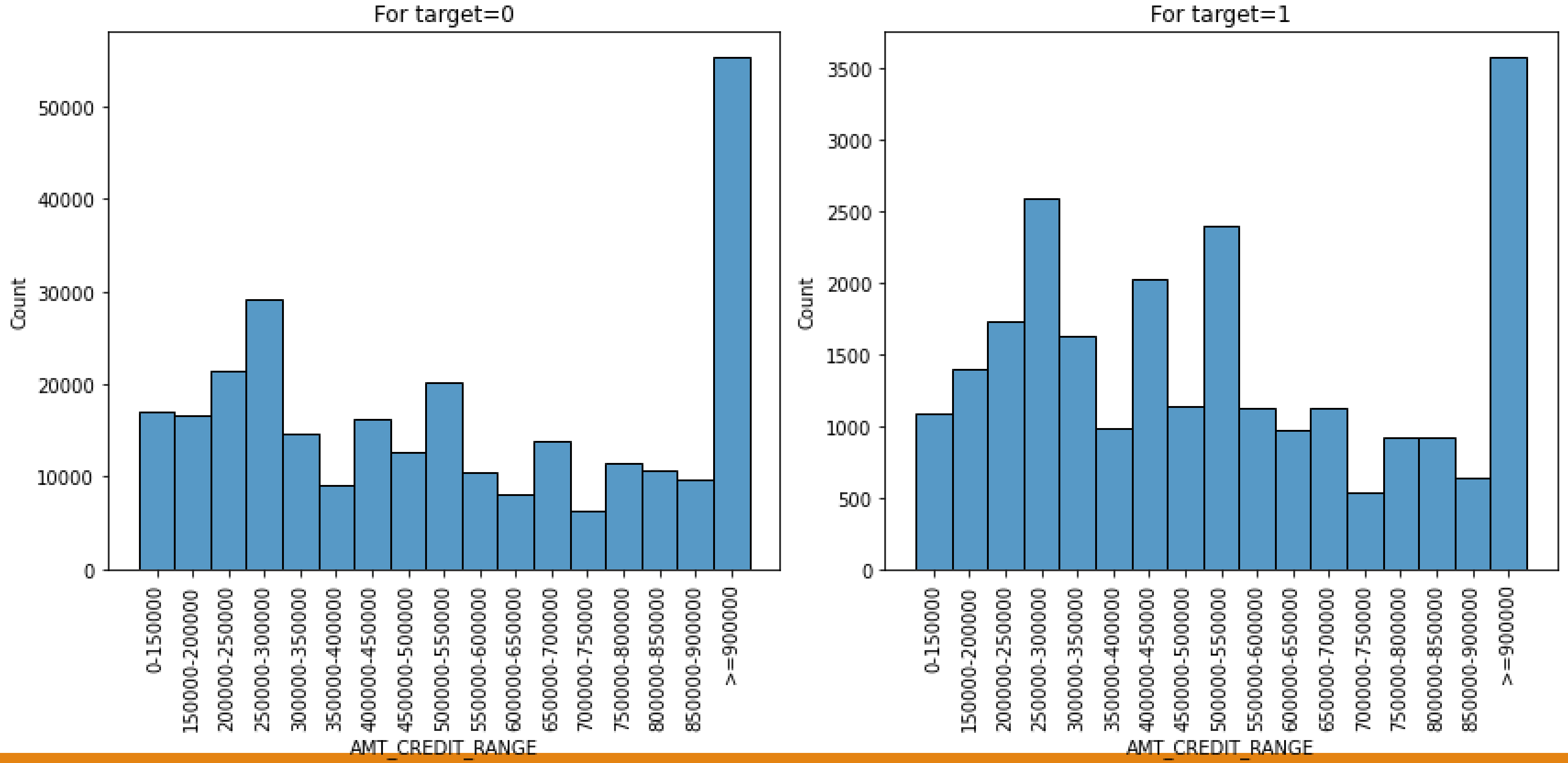
#From above graph we can observe that most of the loans are applied by people whose income is between 125000 and 150000


For target=0


For target=1

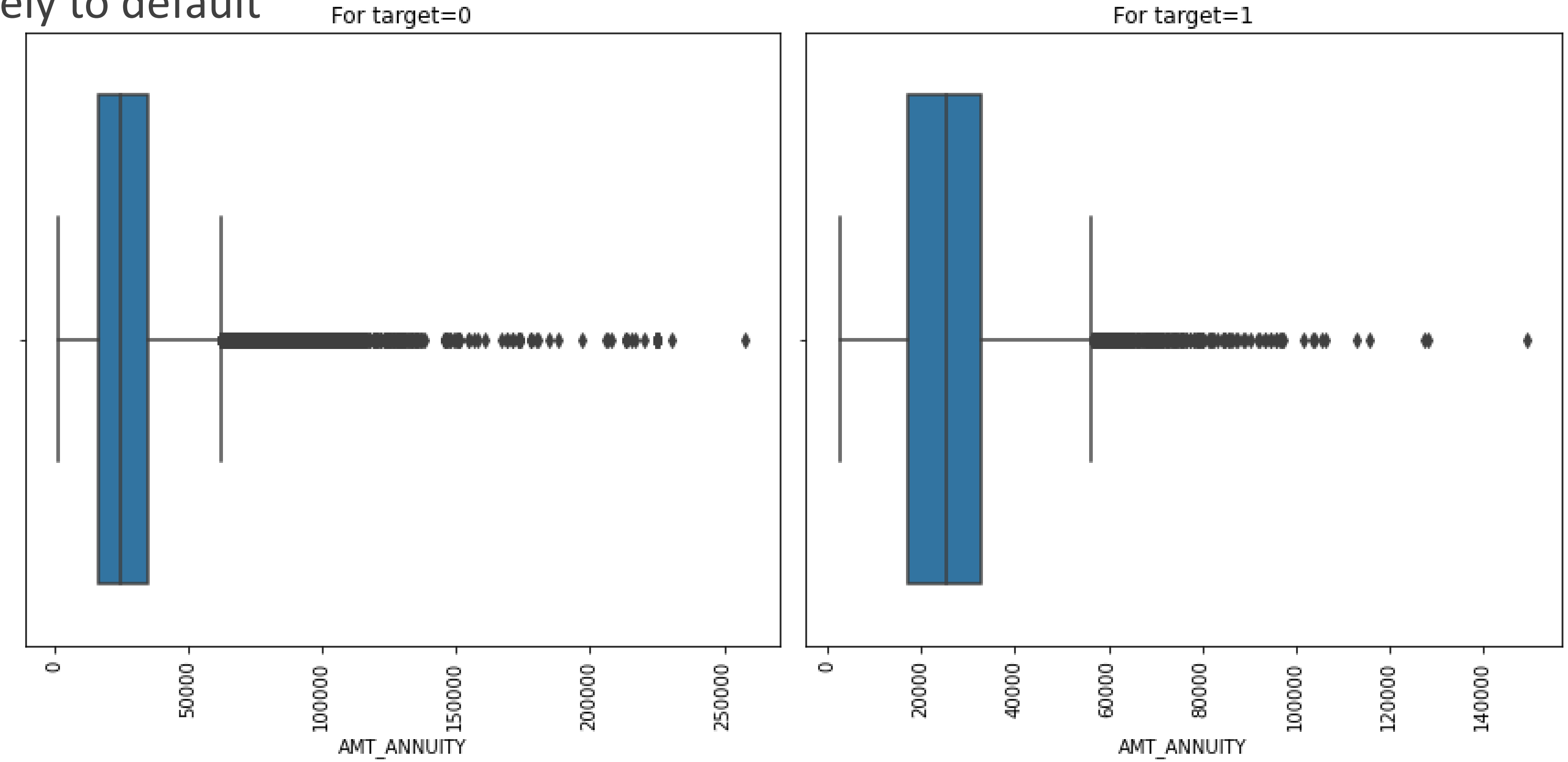
UNIVARIATE ANALYSIS FOR ORDERED CATEGORICAL VARIABLES

#From above graph we can observe that most of the loans are applied by people has credit amount greater than or equal to 900000



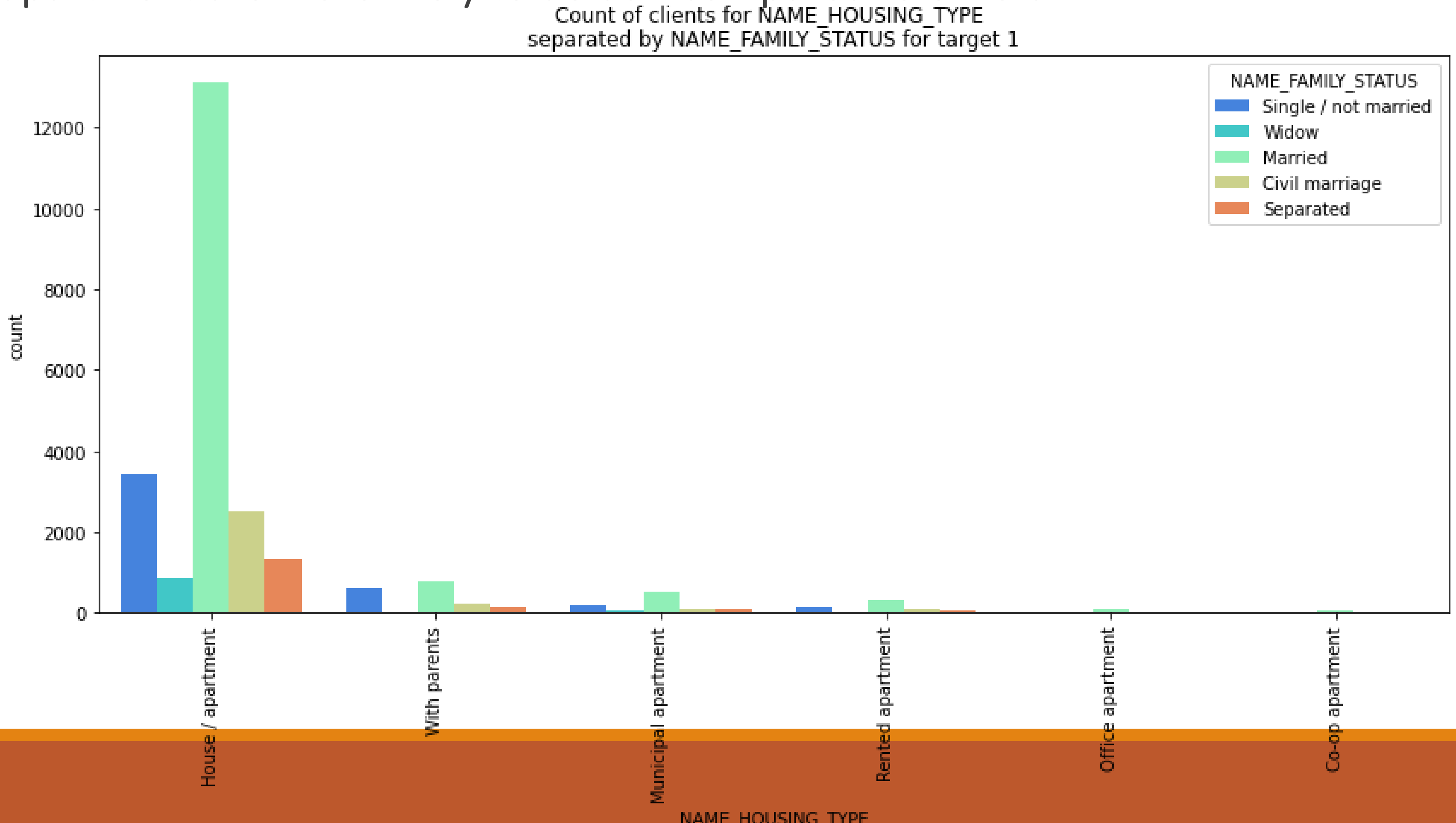
UNIVARIATE ANALYSIS FOR NUMERICAL VARIABLES

#From above graph, we can say that clients who have AMT_ANNUITY between 20000-40000 are most likely to default



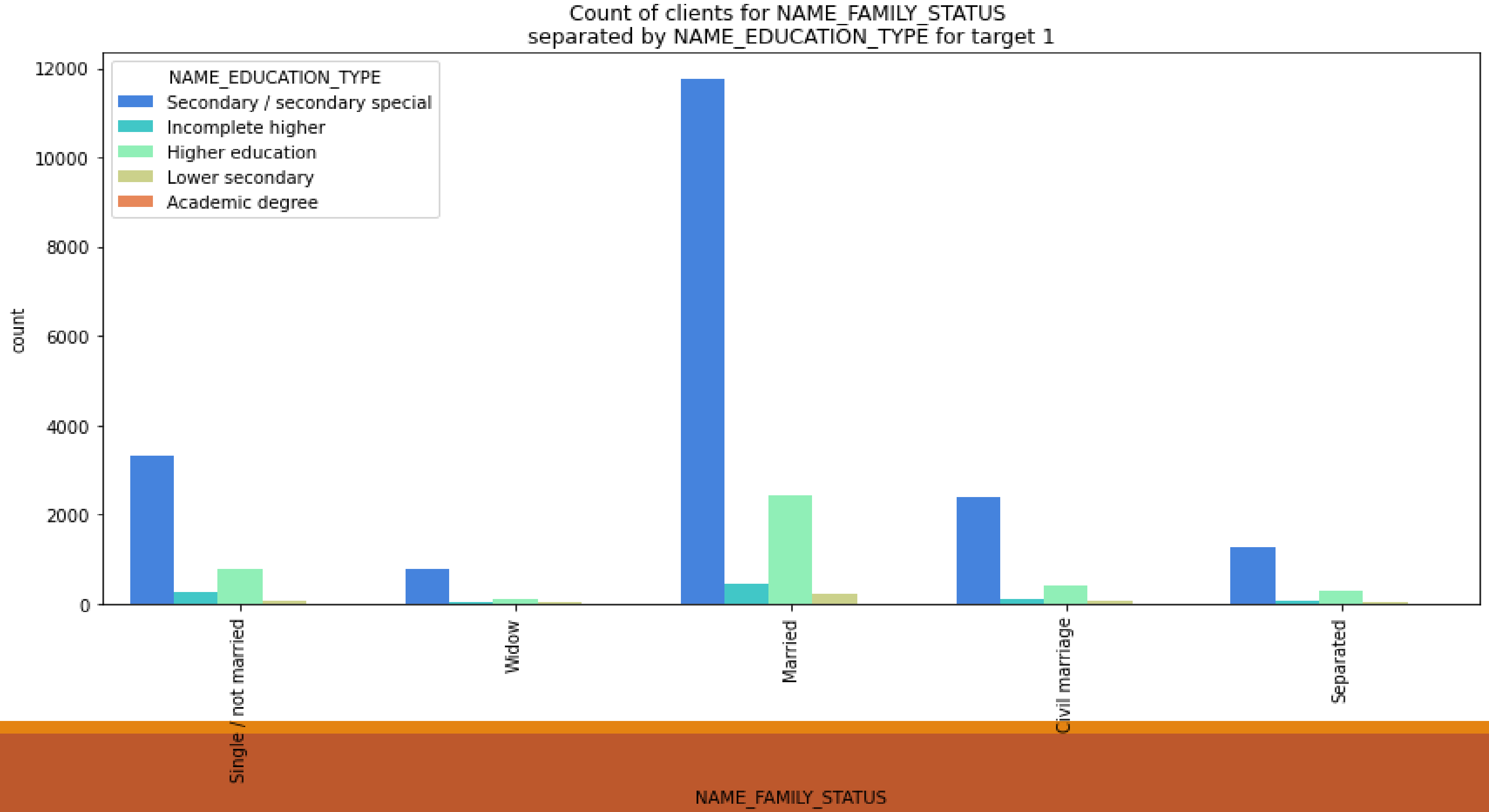
BIVARIATE ANALYSIS FOR TARGET VALUE =1

#We can observe from the graph that people who are married and have their own house/apartment are more likely to default compared to others



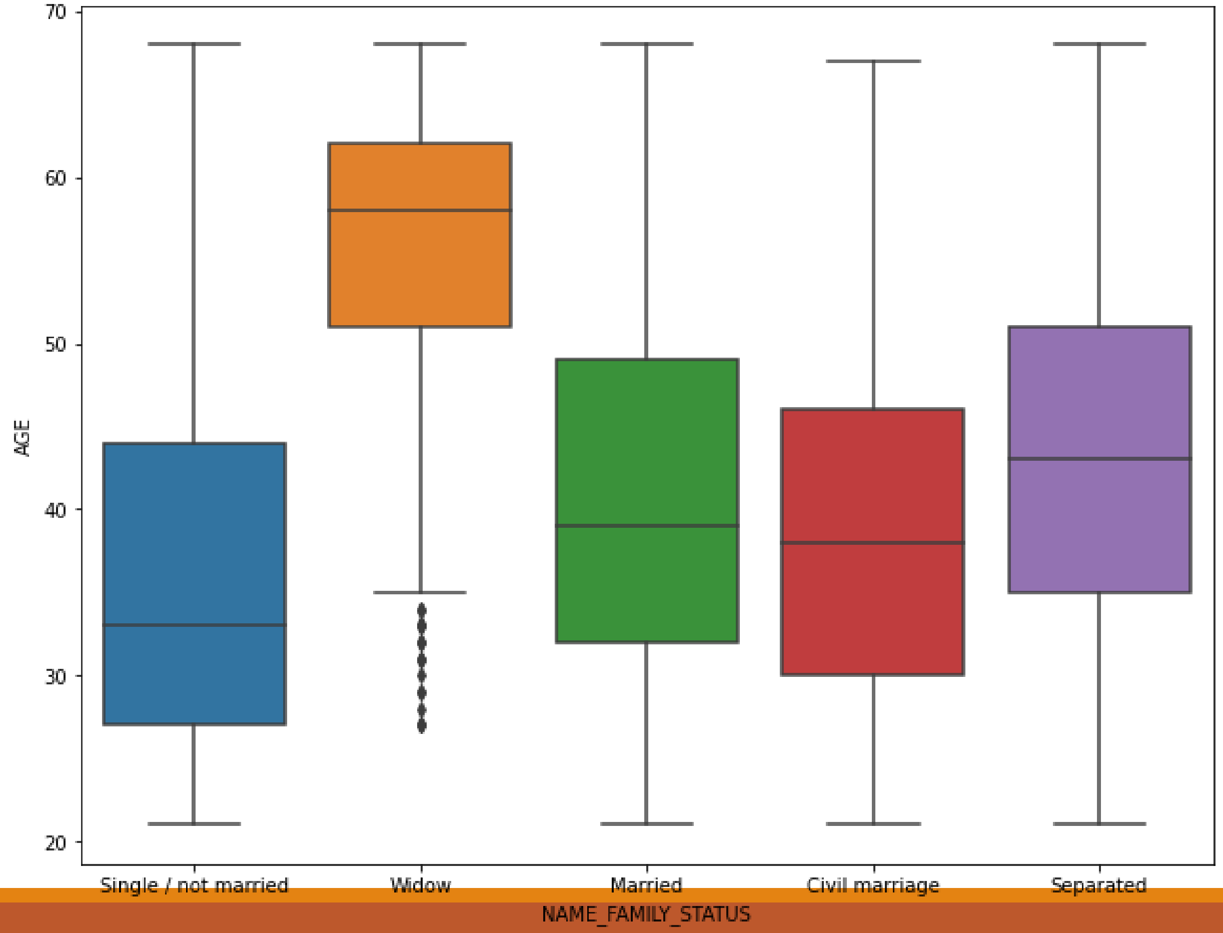
BIVARIATE ANALYSIS FOR TARGET VALUE =1

#We can observe from the graph that people who have secondary/secondary level of education and are married are more likely to default compared to others



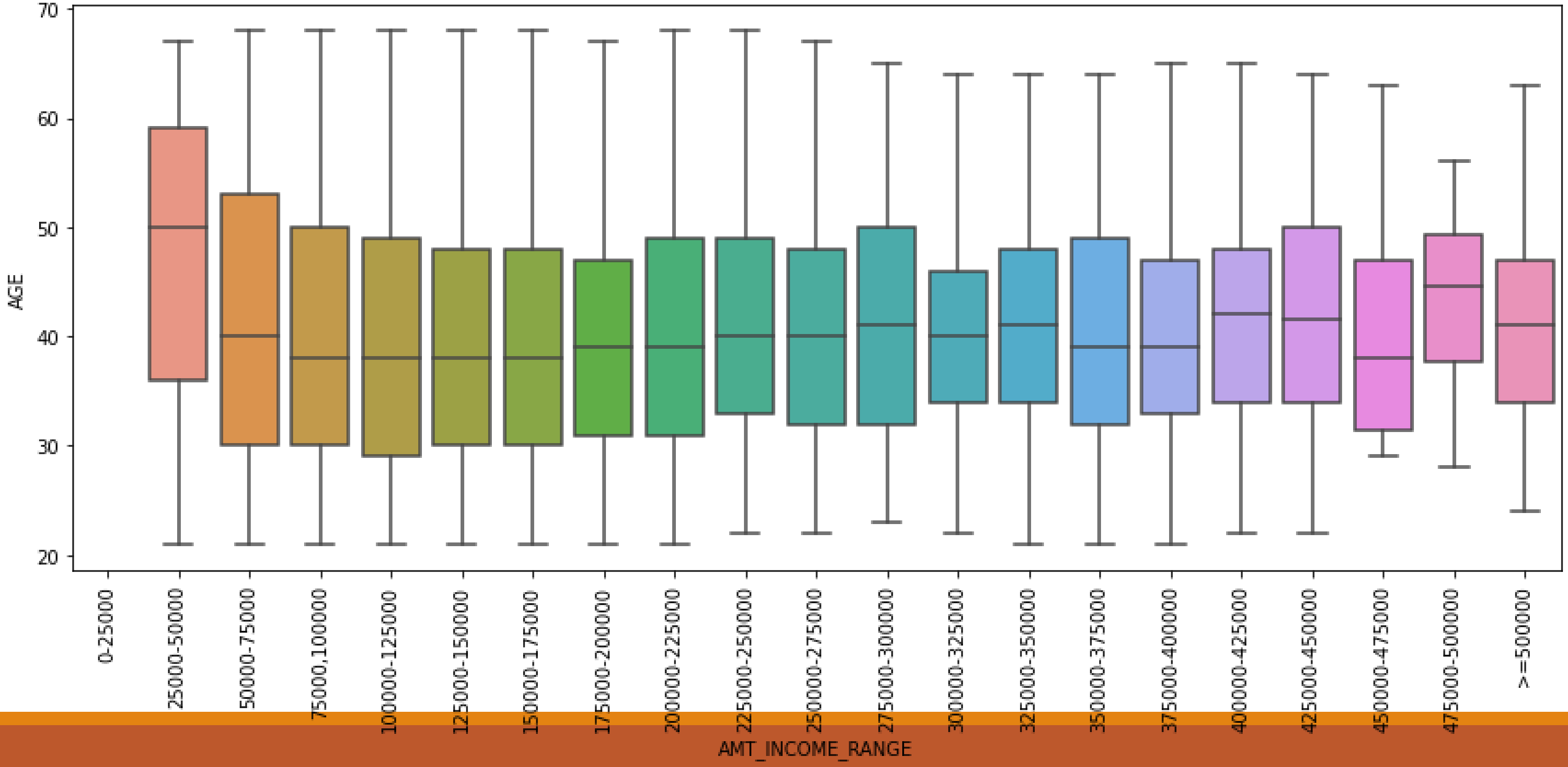
BIVARIATE ANALYSIS FOR TARGET VALUE =1

#We can observe that people who are widow and have age>50 are more likely to default



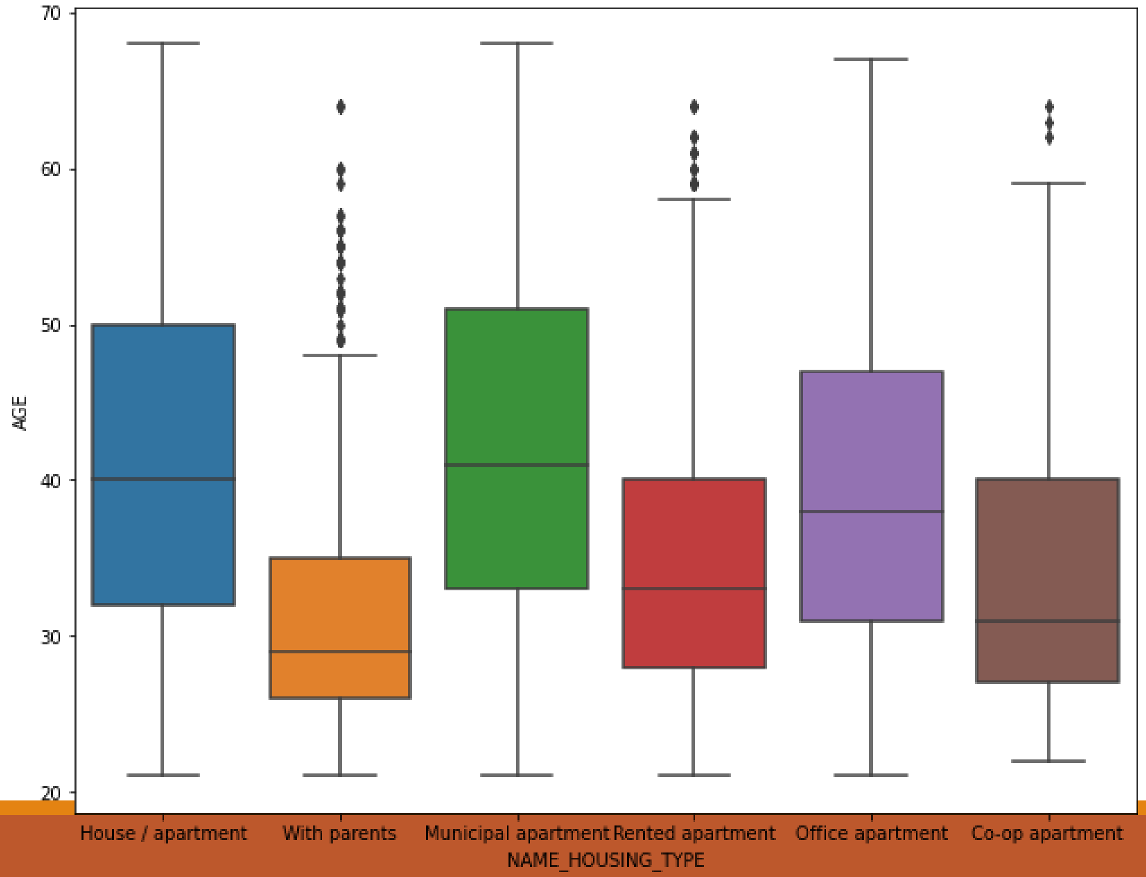
BIVARIATE ANALYSIS FOR TARGET VALUE =1

#We can observe that people who have income between 25000-50000 and have age>50 are more likely to default

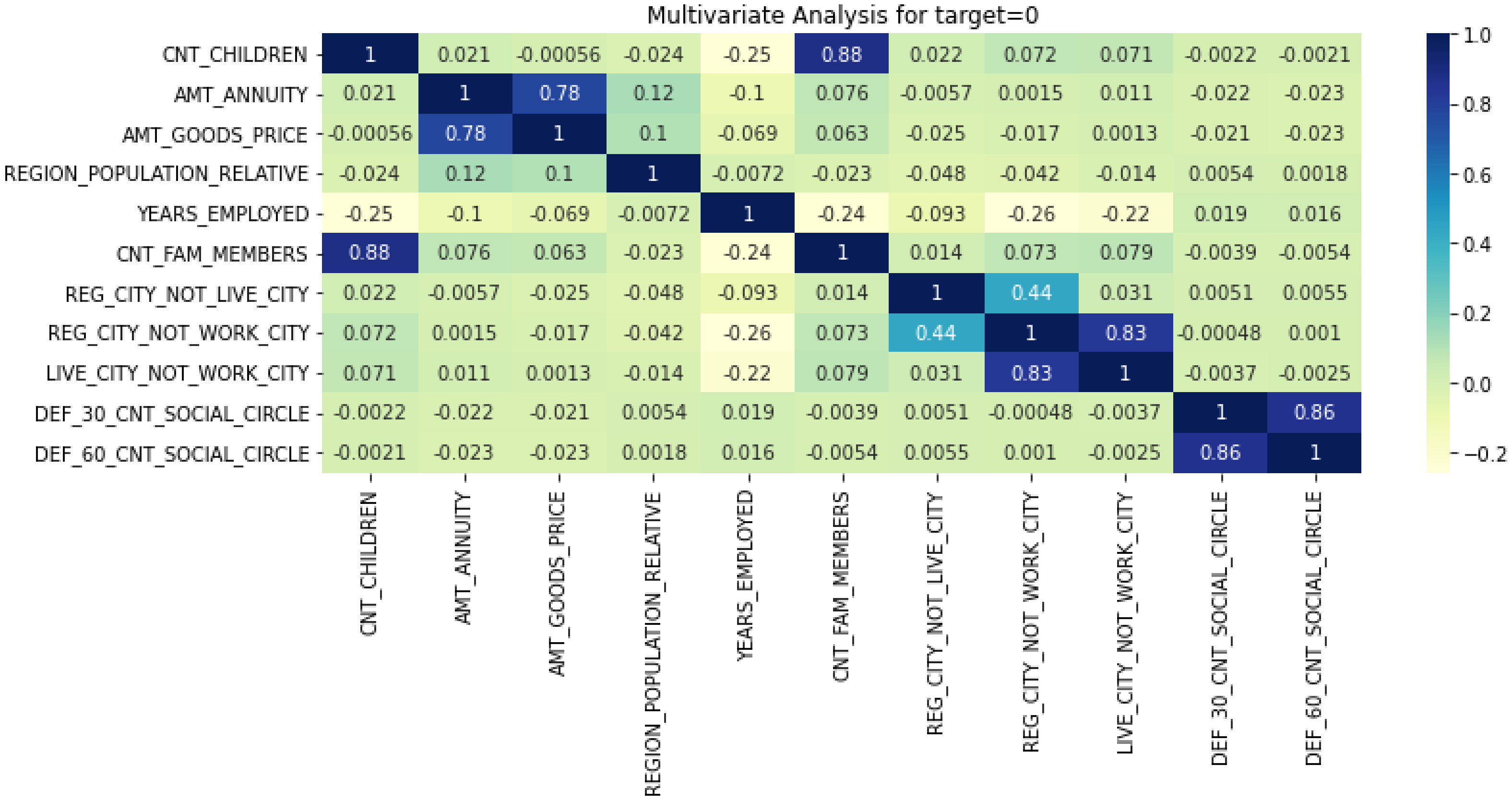


BIVARIATE ANALYSIS FOR TARGET VALUE =1

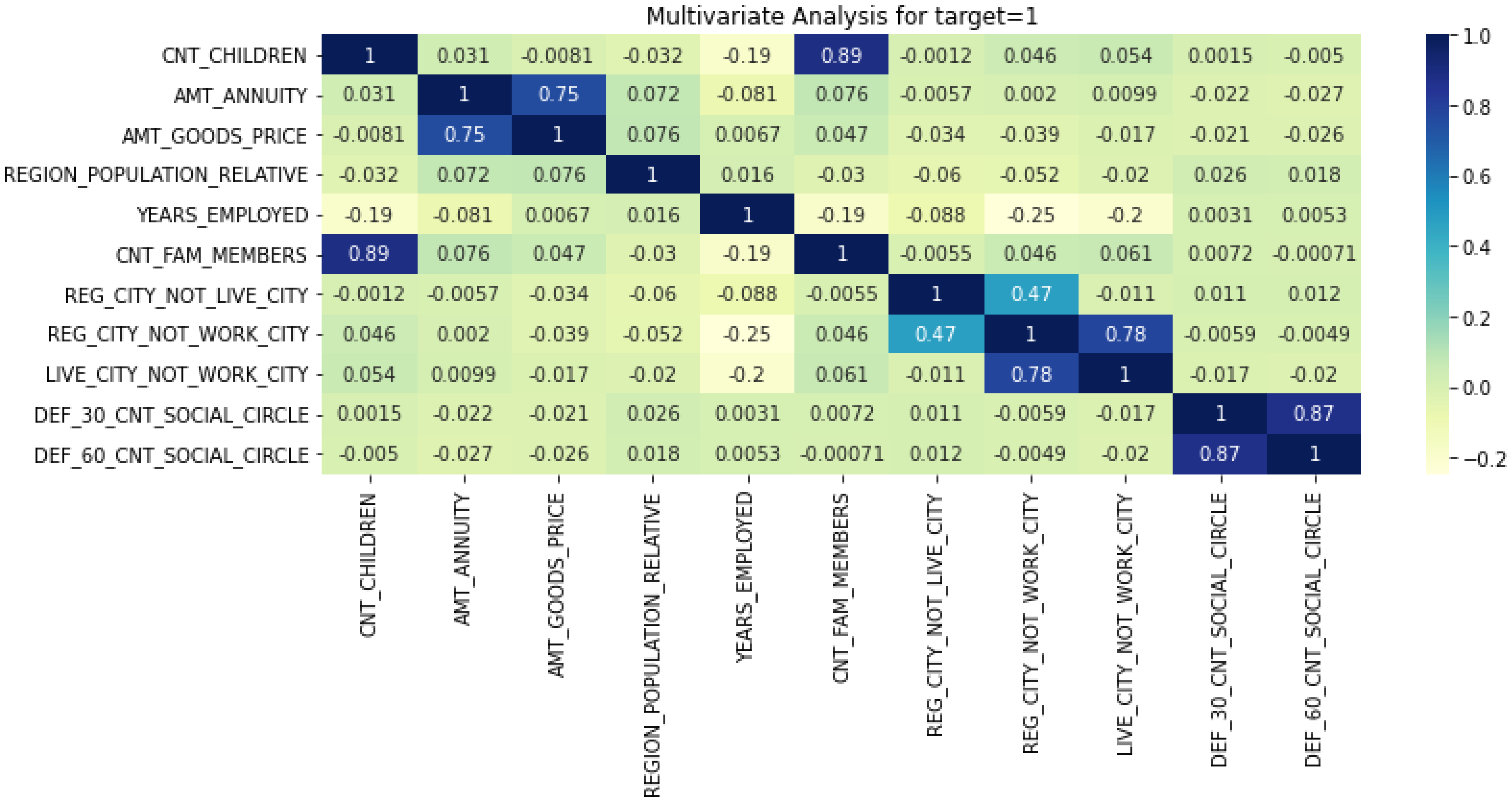
#We can observe that people who live in municipal apartment or own house/apartment and have age>40 are more likely to default



MUTIVARIATE ANALYSIS FOR EACH TARGET VALUE =0



MUTIVARIATE ANALYSIS FOR EACH TARGET VALUE =1


Multivariate Analysis for target=1

	CNT_CHILDREN	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_EMPLOYED	CNT_FAM_MEMBERS	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
CNT_CHILDREN	1	0.031	-0.0081	-0.032	-0.19	0.89	-0.0012	0.046	0.054	0.0015	-0.005
AMT_ANNUITY	0.031	1	0.75	0.072	-0.081	0.076	-0.0057	0.002	0.0099	-0.022	-0.027
AMT_GOODS_PRICE	-0.0081	0.75	1	0.076	0.0067	0.047	-0.034	-0.039	-0.017	-0.021	-0.026
REGION_POPULATION_RELATIVE	-0.032	0.072	0.076	1	0.016	-0.03	-0.06	-0.052	-0.02	0.026	0.018
YEARS_EMPLOYED	-0.19	-0.081	0.0067	0.016	1	-0.19	-0.088	-0.25	-0.2	0.0031	0.0053
CNT_FAM_MEMBERS	0.89	0.076	0.047	-0.03	-0.19	1	-0.0055	0.046	0.061	0.0072	-0.00071
REG_CITY_NOT_LIVE_CITY	-0.0012	-0.0057	-0.034	-0.06	-0.088	-0.0055	1	0.47	-0.011	0.011	0.012
REG_CITY_NOT_WORK_CITY	0.046	0.002	-0.039	-0.052	-0.25	0.046	0.47	1	0.78	-0.0059	-0.0049
LIVE_CITY_NOT_WORK_CITY	0.054	0.0099	-0.017	-0.02	-0.2	0.061	-0.011	0.78	1	-0.017	-0.02
DEF_30_CNT_SOCIAL_CIRCLE	0.0015	-0.022	-0.021	0.026	0.0031	0.0072	0.011	-0.0059	-0.017	1	0.87
DEF_60_CNT_SOCIAL_CIRCLE	-0.005	-0.027	-0.026	0.018	0.0053	-0.00071	0.012	-0.0049	-0.02	0.87	1

CONCLUSION DRAWN FROM MUTIVARIATE ANALYSIS

#1.CNT_CHILDREN has most positive correlation with CNT_FAM_MEMBERS, this maybe because count of family members includes number of children

#2.YEARS_EMPLOYED has slight negative correlation with AMT_ANNUITY this shows that clients who worked for less years have greater value of AMT_ANNUITY

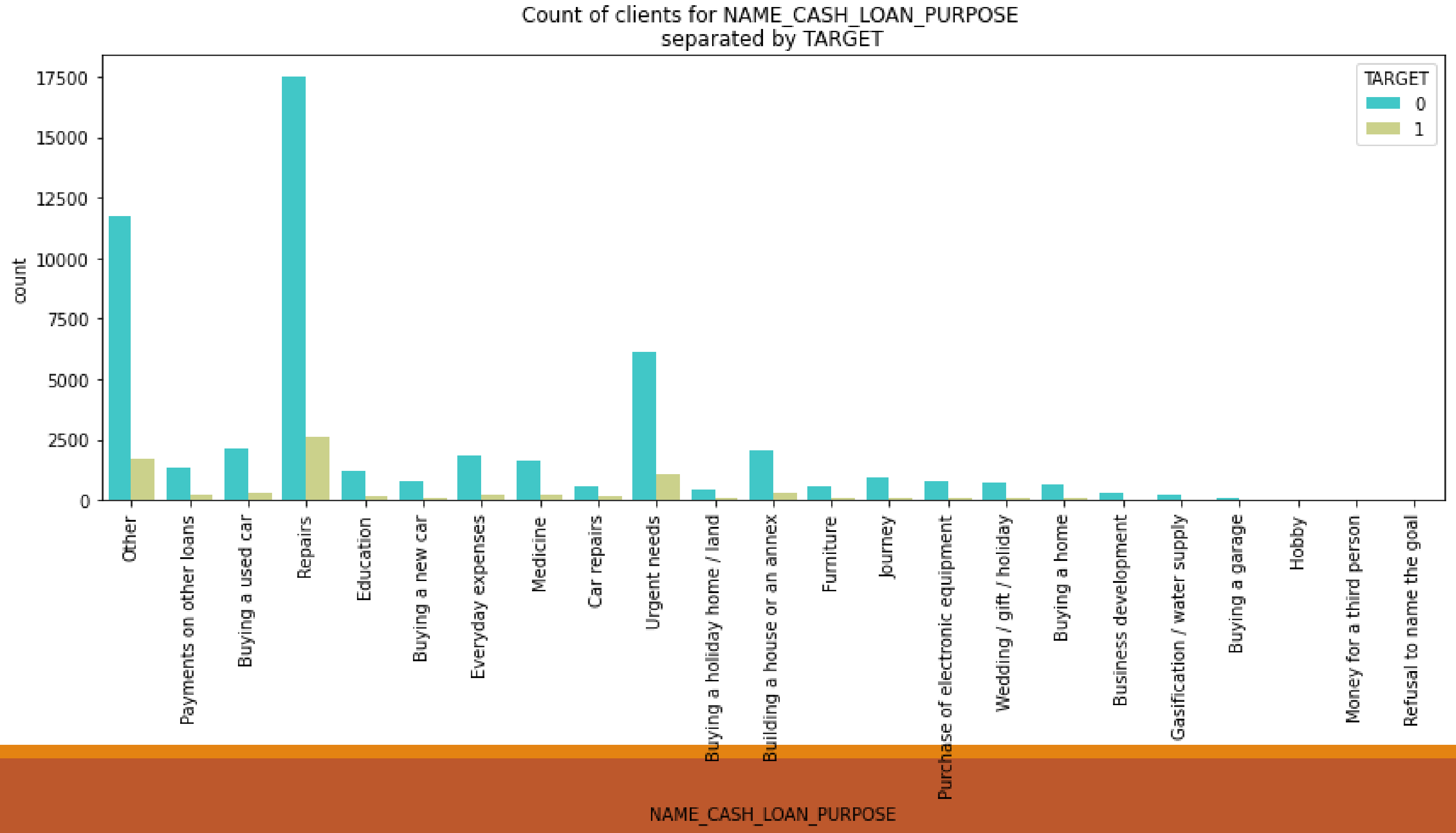
#3. CNT_FAMILY has most positive correlation with AMT_ANNUITY other than CNT_CHILDREN this maybe because clients with big family may require more loan amount

#4. REGION_POPULATION_RELATIVE has positive correlation=0.12 with AMT_ANNUITY in target=0 and positive correlation=0.072 with AMT_ANNUITY in target=1 this shows that more populated areas are likely to payback loan compared to others

#5. REG_CITY_NOT_WORK_CITY has negative correlation with YEARS_EMPLOYED this shows that client who lives in city other than their hometown are most likely to have started working recently.

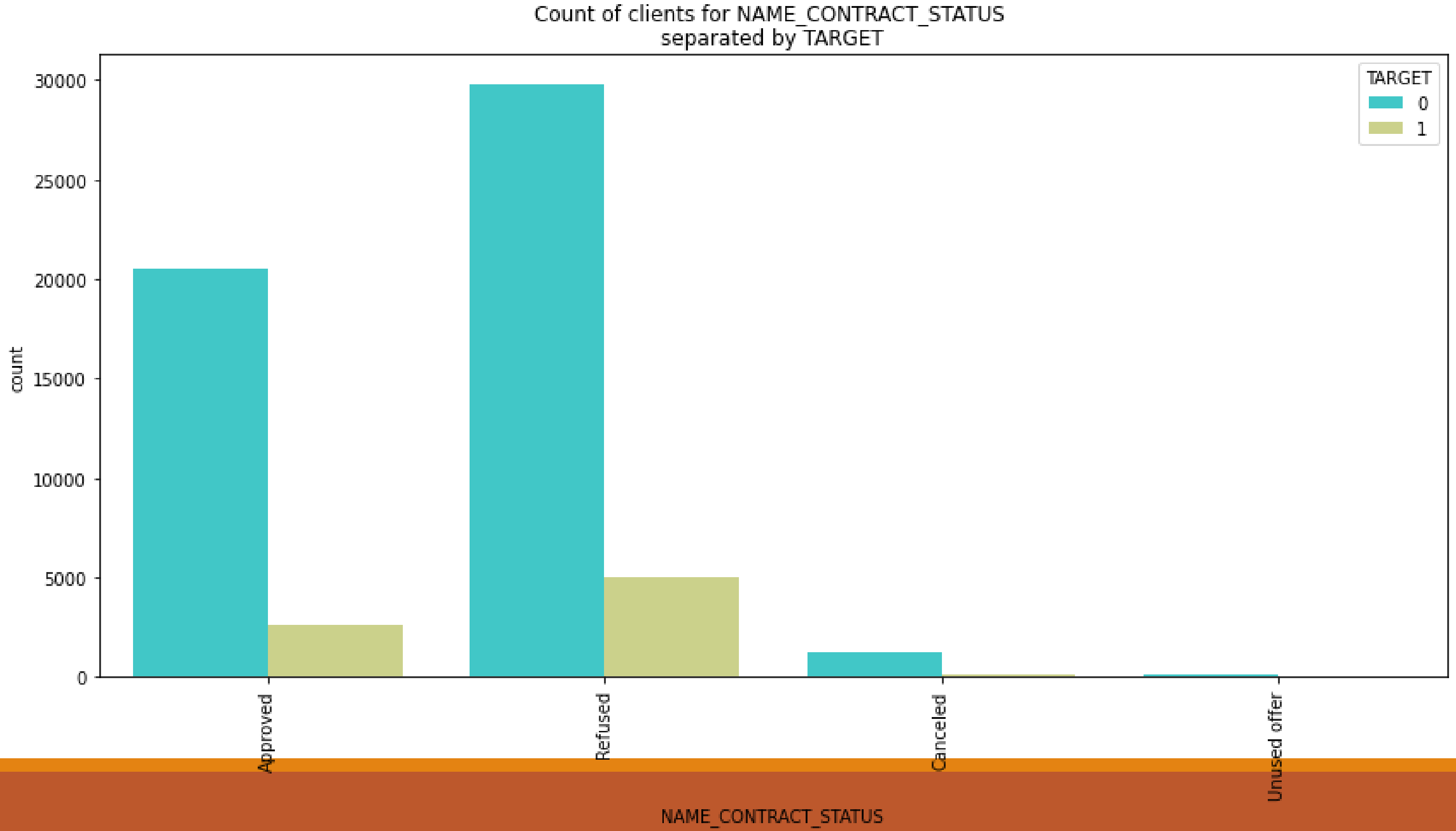
BIVARIATE ANALYSIS AFTER MERGING DATASETS

#From below graph we can observe that most loans are applied for purpose of repairs and therefore the chances of default also increases



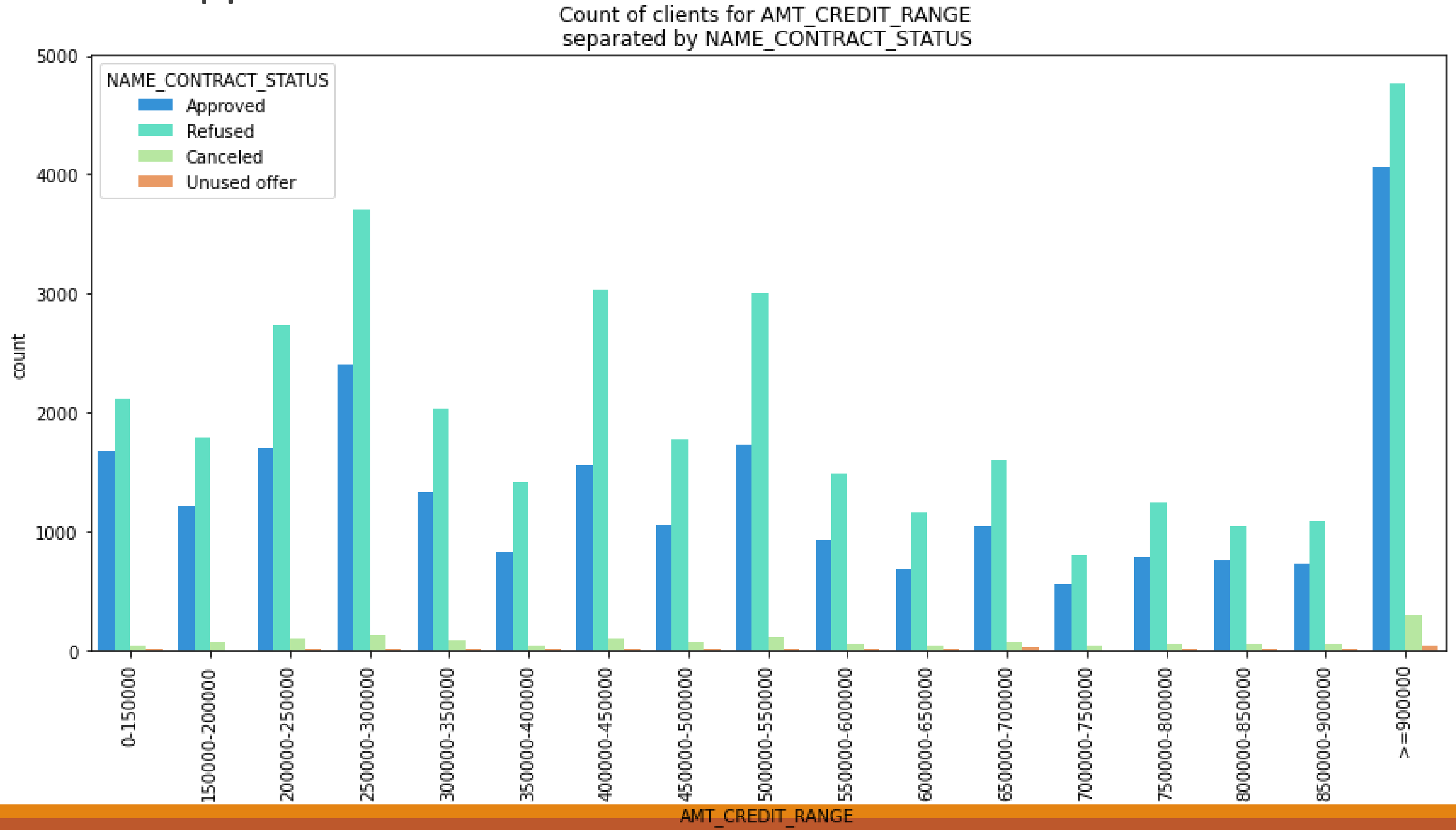
BIVARIATE ANALYSIS AFTER MERGING DATASETS

From below graph we can observe that loans which were previously refused have higher chances of default compared to those whose previous application was approved



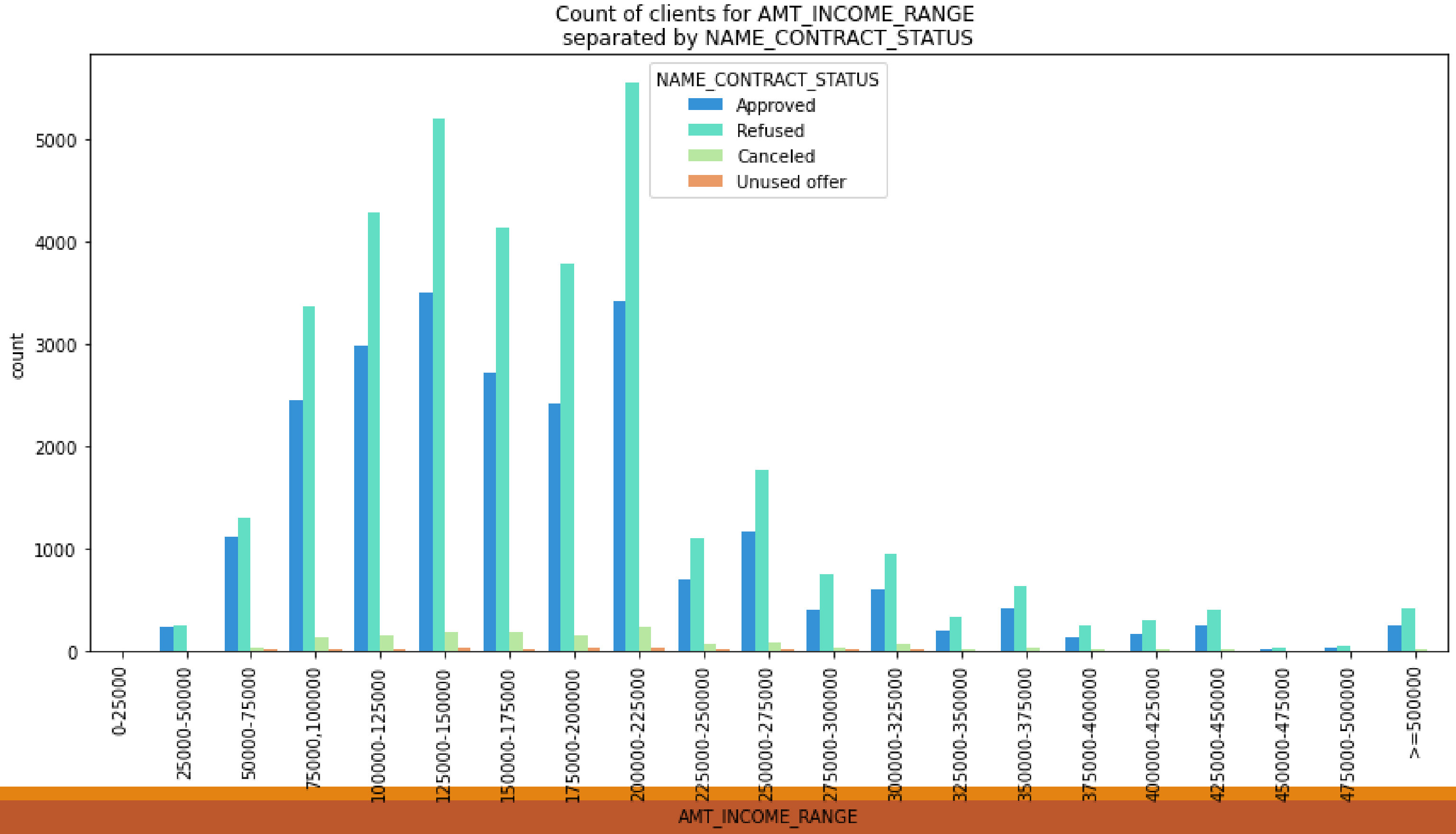
BIVARIATE ANALYSIS AFTER MERGING DATASETS

#From below graph we can observe that clients who have credit >=900000 are also the ones whose previous application was refused



BIVARIATE ANALYSIS AFTER MERGING DATASETS

#From below graph we can observe that clients who have income between range 200000-225000 are also the ones whose previous application was refused


Count of clients for AMT_INCOME_RANGE
separated by NAME_CONTRACT_STATUS
NAME_CONTRACT_STATUS
Approved
Refused
Canceled
Unused offer
count
>=500000
AMT_INCOME_RANGE