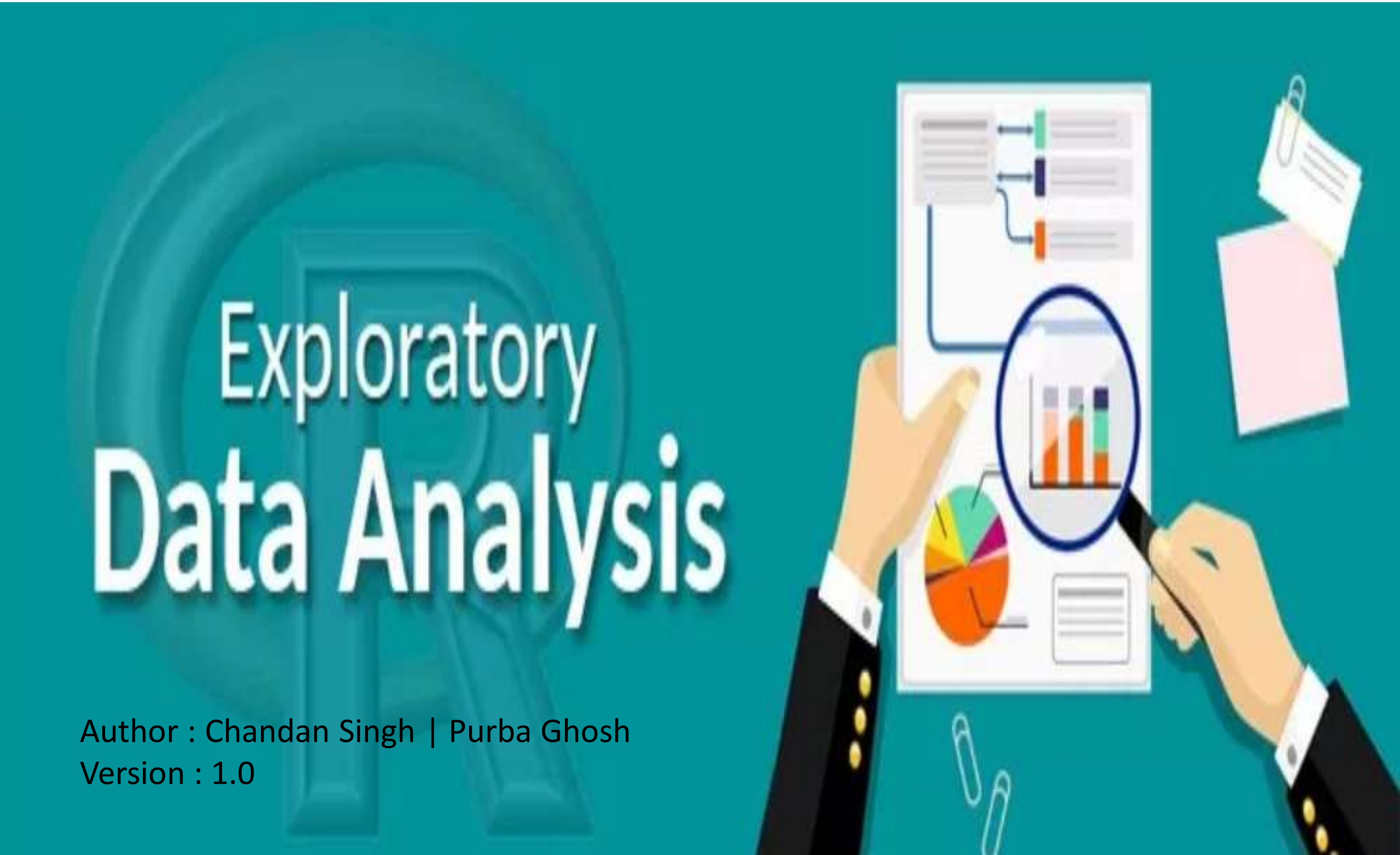
EDA Case Study



Author : Chandan Singh | Purba Ghosh
Version : 1.0

Applying EDA in a real business scenario

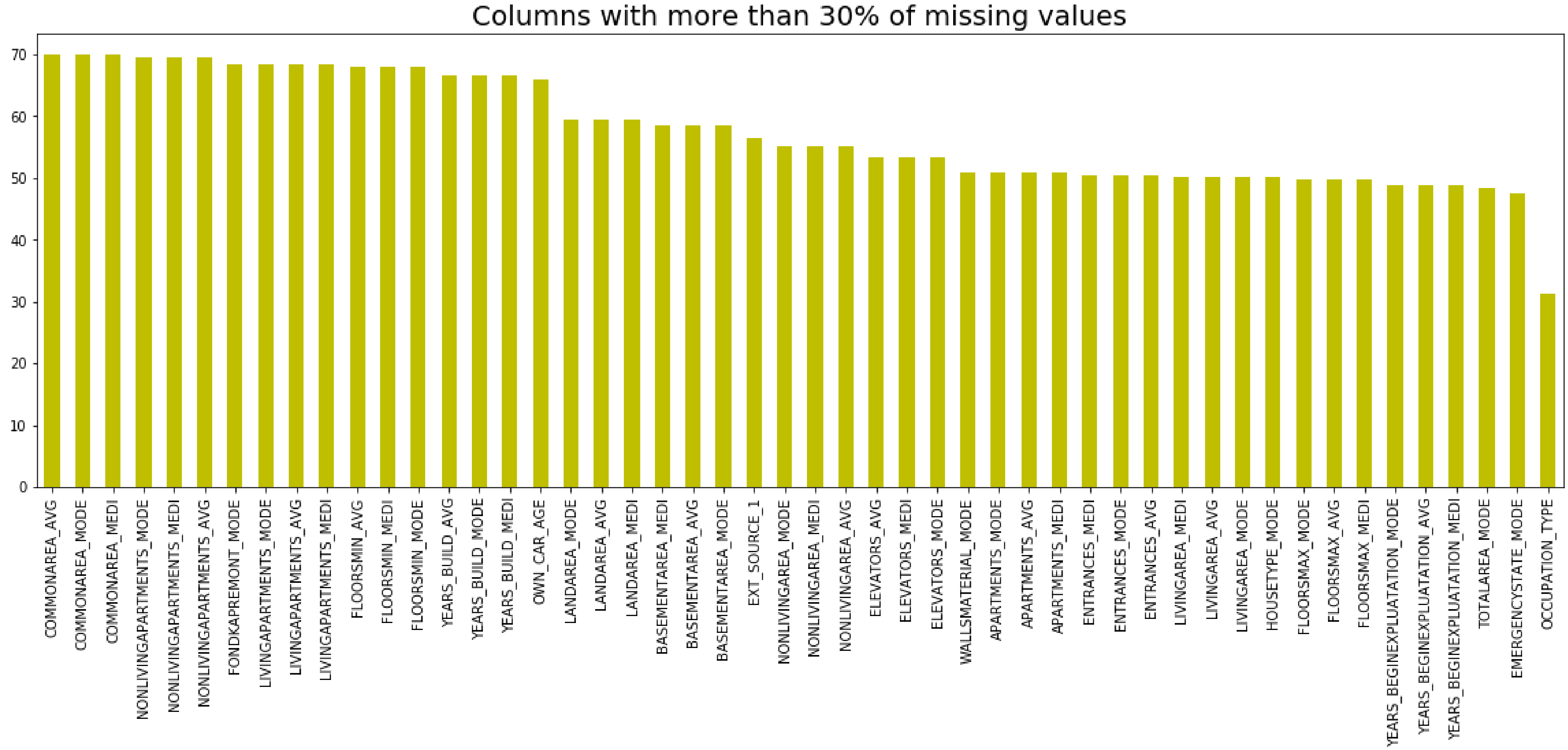
Objective :

1. To use EDA to analyse the patterns present in the data to ensure that the applicants ,capable of repaying the loan are not rejected
2. Understand how consumer attributes and loan attributes influence the tendency of default.
3. Understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default and utilise this knowledge for its portfolio and risk assessment .

Current Application

Identification of columns with more than 30% of missing values

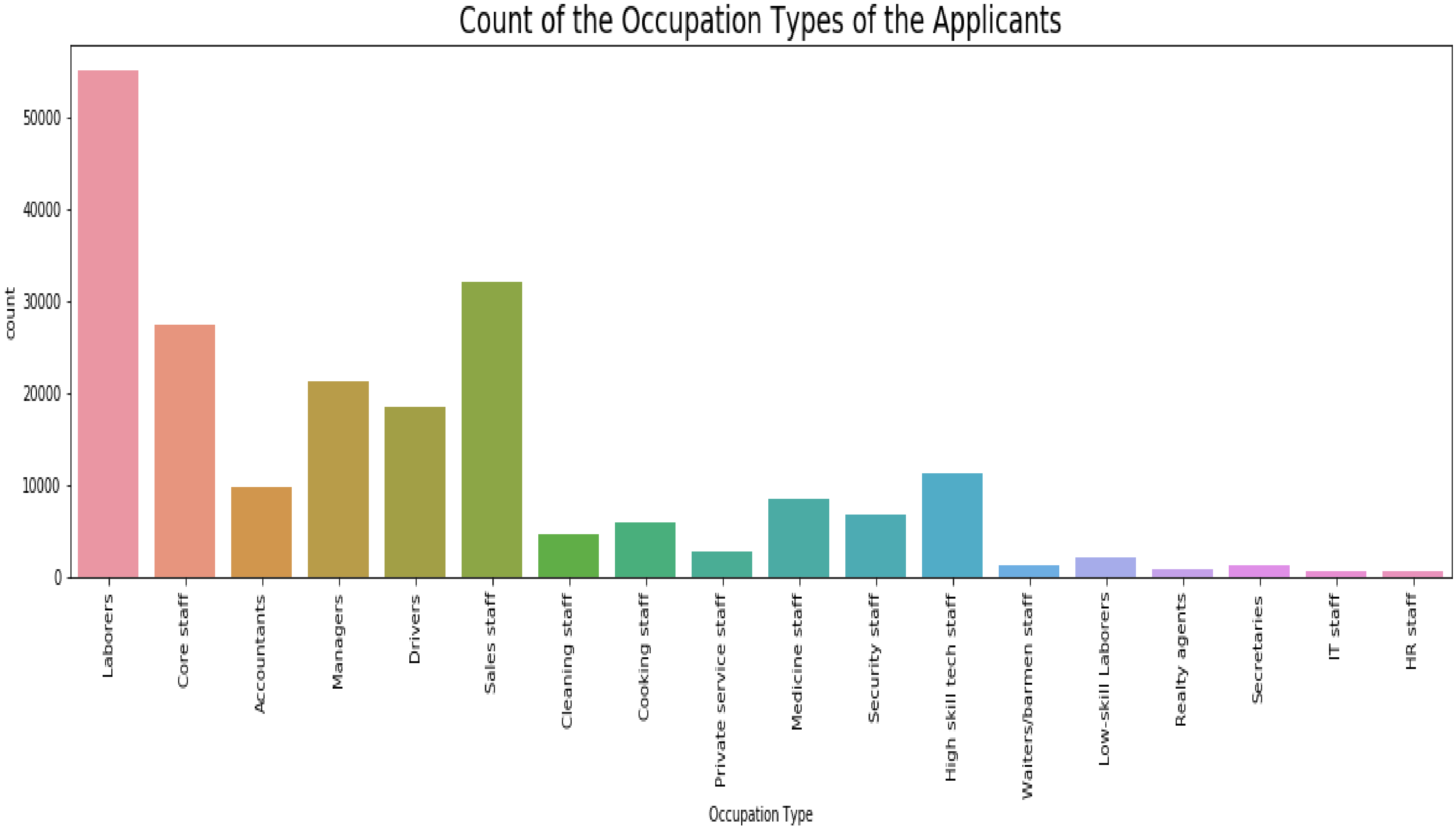
- Application Dataset has 307511 rows and 122 columns. So we identified columns with more than 30% missing values.



Inference from previous graph

- There are 50 columns, that have more than 30% of missing values.
- Some columns have excrement large values such as COMMONAREA_AVG, COMMONAREA_MODE, COMMONAREA_MEDI, FLOORSMIN_AVG, FLOORSMIN_MEDI etc.
- With more than 30% of missing values, these columns are not recommended to use for any analysis, without proper treatment/imputation. Since, the no of columns to be treated are high, hence, imputing these columns with arbitrary values like mean, median, mode etc, would be risky and might produce unexpected results. It is better to drop all these columns, but, for this exercise, we only need to drop columns that are having more than 65% of missing values

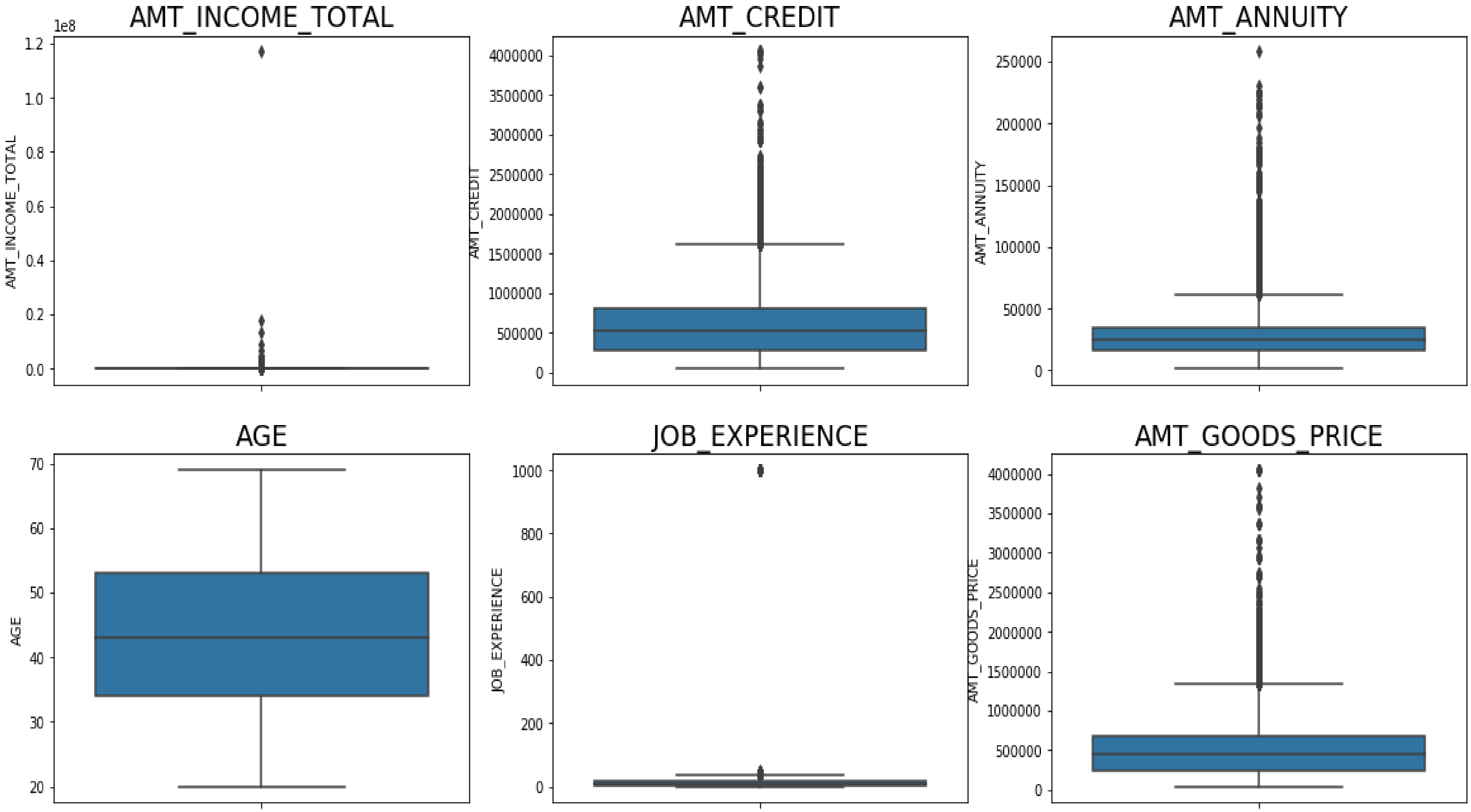
Count of the Occupation Types of the Applicants



Inference from previous graph

- It seems that the "Self Employed" is missing in the Occupation List, which is the major chunk of occupation, hence, we can impute the missing values in the variable with the value "Self Employed"

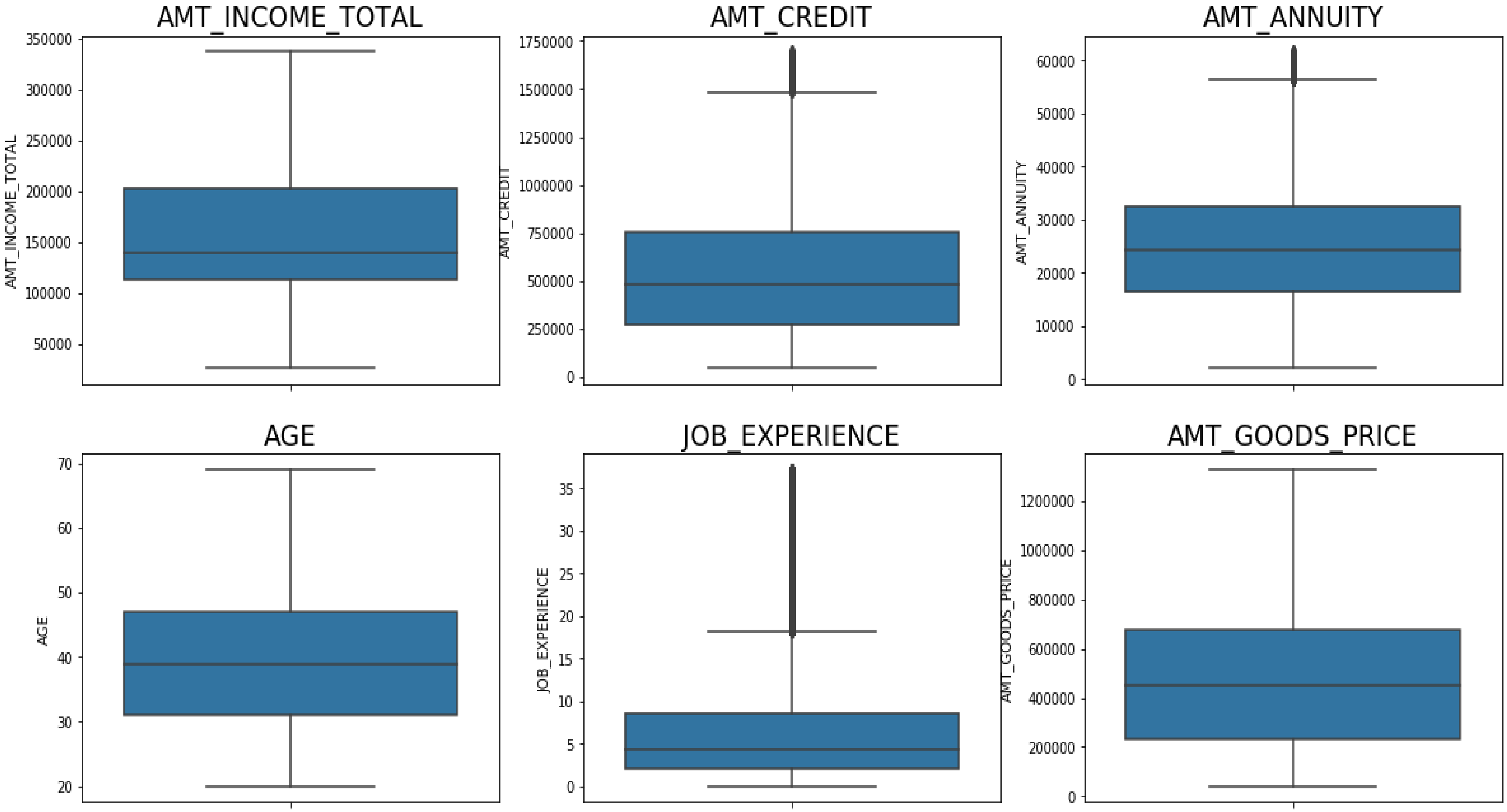
Check for outliers in the interested columns



Inference from previous graph

- There are outliers present in the following columns:
  - AMT_INCOME_TOTAL
  - AMT_ANNUITY
  - JOB_EXERIENCE
  - AMT_GOODS_PRICE
- Due to the presence of the outliers, the boxplot is not properly readable for the above columns
- AGE Column is formly distributed.
- Outliers Can be removed using IQR

Verify the data after removing the outliers, perform analysis and provide insight

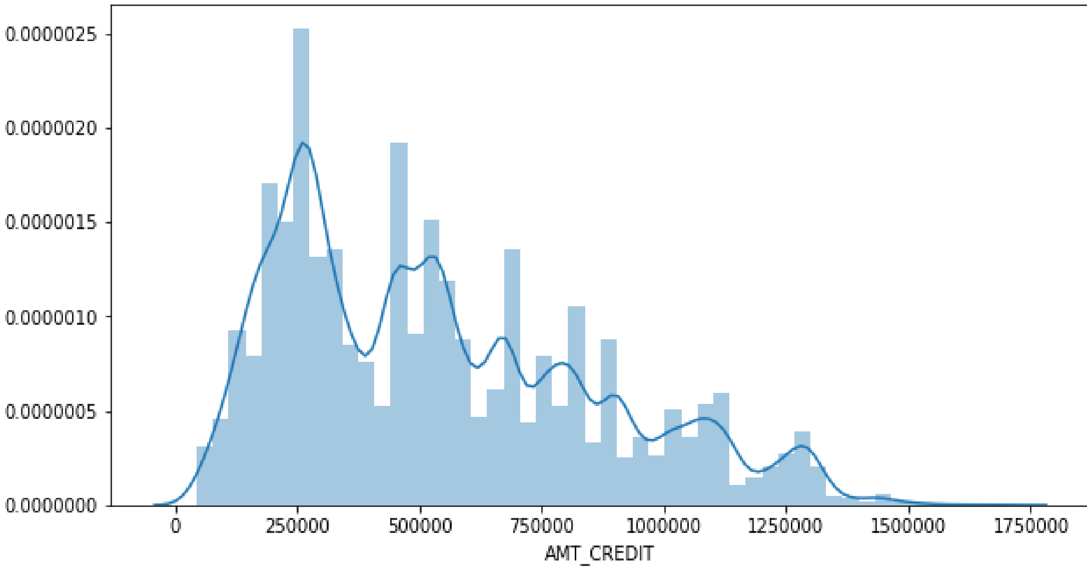


Inference from previous graph

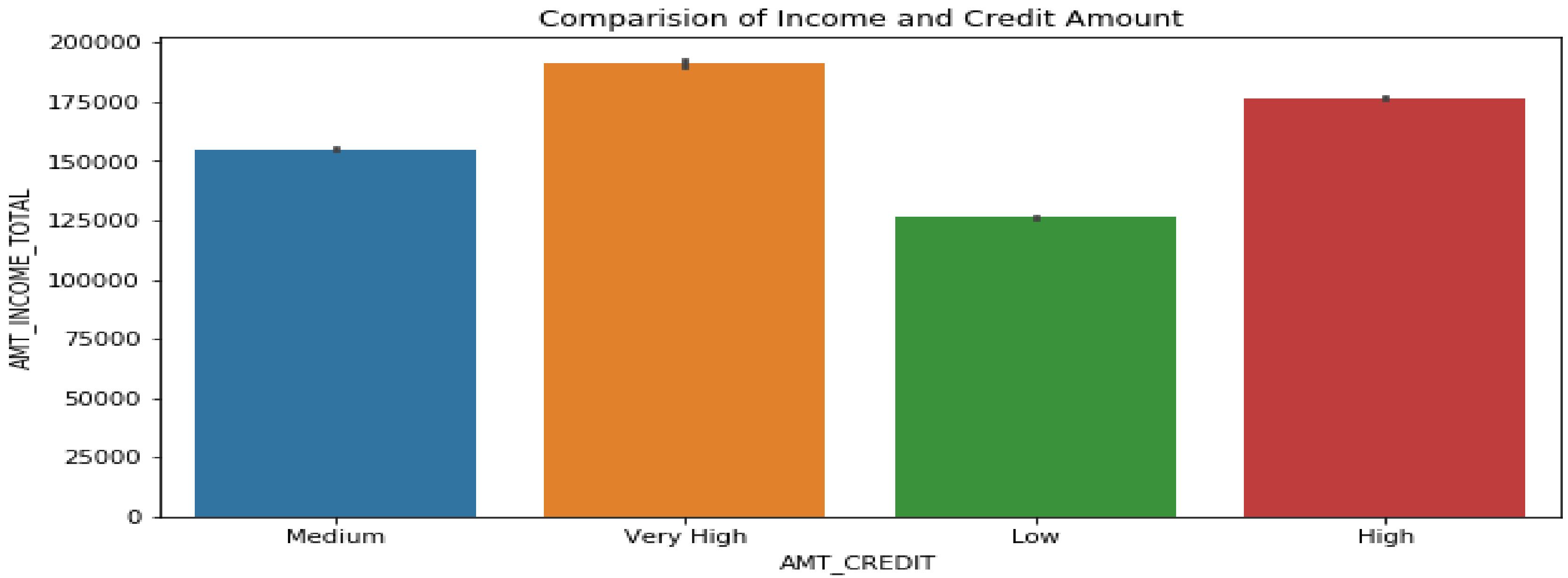
- Outliers(values beyond upper bound i.e. 1.5 * 75 Percentile) have been removed from the dataset for the above mentioned numeric columns
- The no of records have been reduced to 224797
- Most of the people applied for the loans are between 35 and 48 years old. As young as 20 years and as old as 69 years.
- Regarding work experience, majority of the applicants are having approximately 8 years of experience. While the average work experience is 4.4 years.
- The average income of the applicants is approximately 112500 . 75% of the applicants earn 202500.
- Average Loan Amount disbursed is 486000. Minimum and Maximum Loan Amount are 45000 and 1695483 respectively.

Binned continuous variable

- AMT_CREDIT and AMT_INCOME_TOTAL can be binned
- Now, bin the variable into the following 4 discrete categories. This will help us in analyzing - Loan Amount varies across other variables such as Target(Default Rate), Income of the applicant etc
- Low
- Medium
- High
- Very High



Comparison of income and credit amount



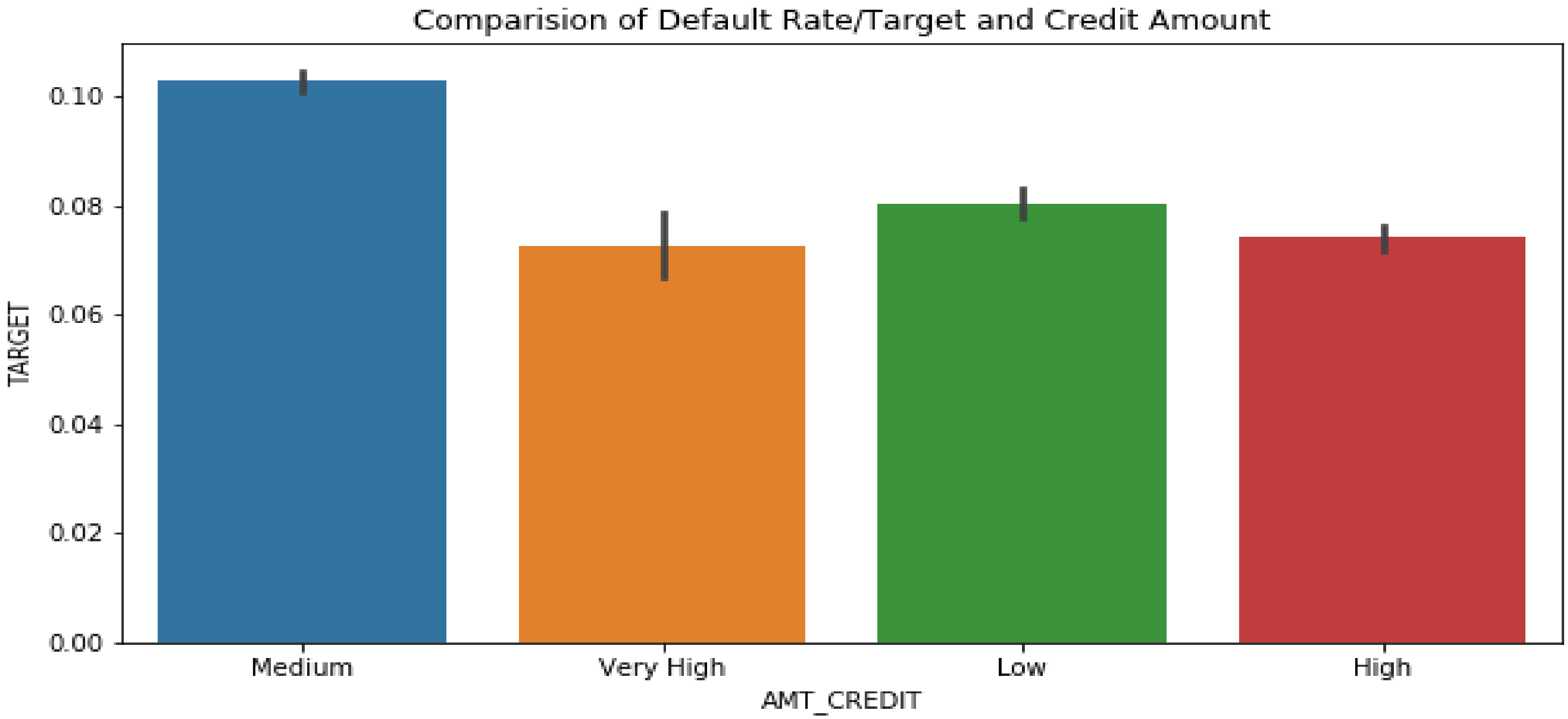
Inference :

1.Higher the income higher the loan amount

2.Lower the income lower the loan amount

3.This clearly suggests that the current Income of the Applicants, plays a vital role in the loan eligibility

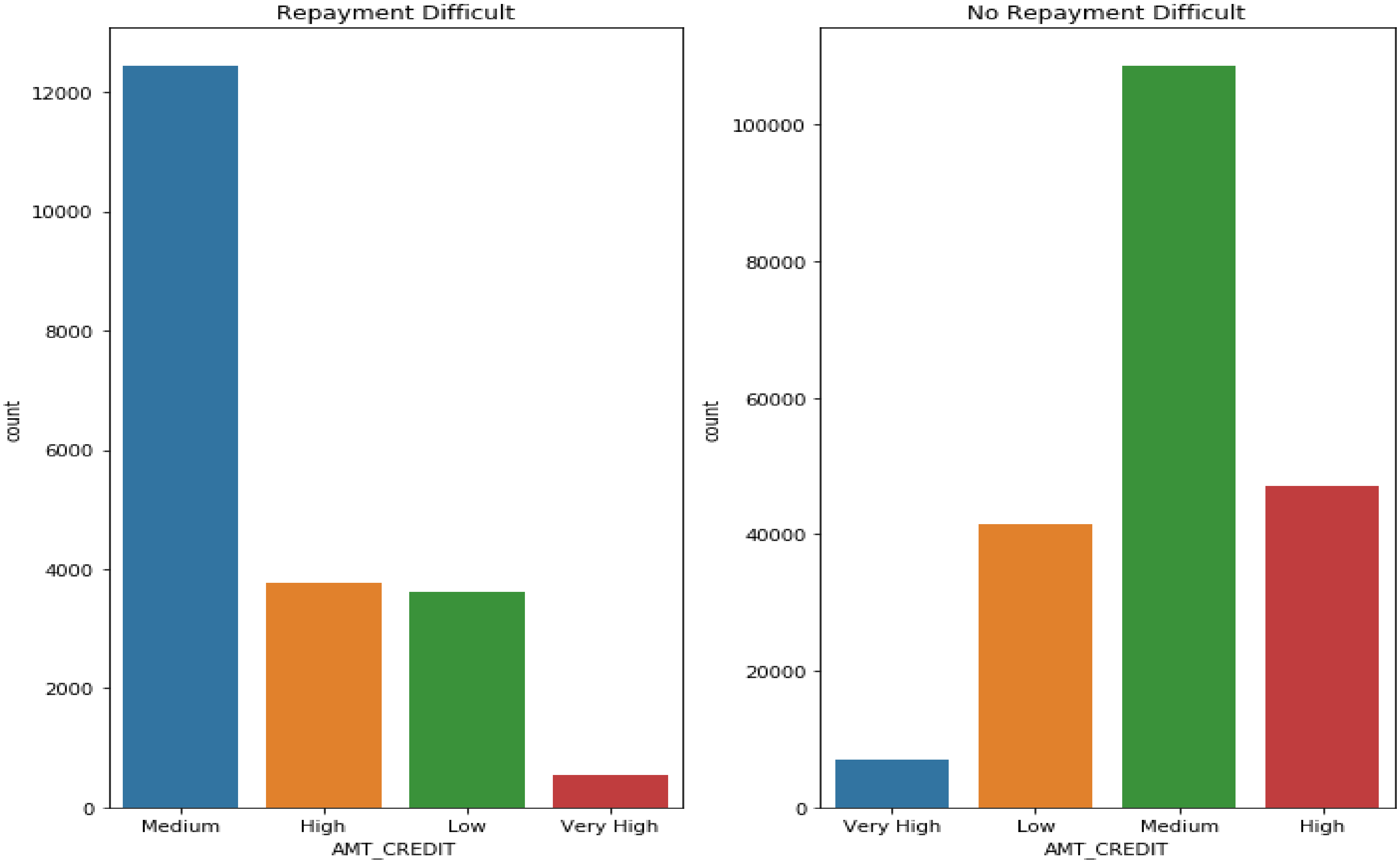
Comparison of default rate and credit amount



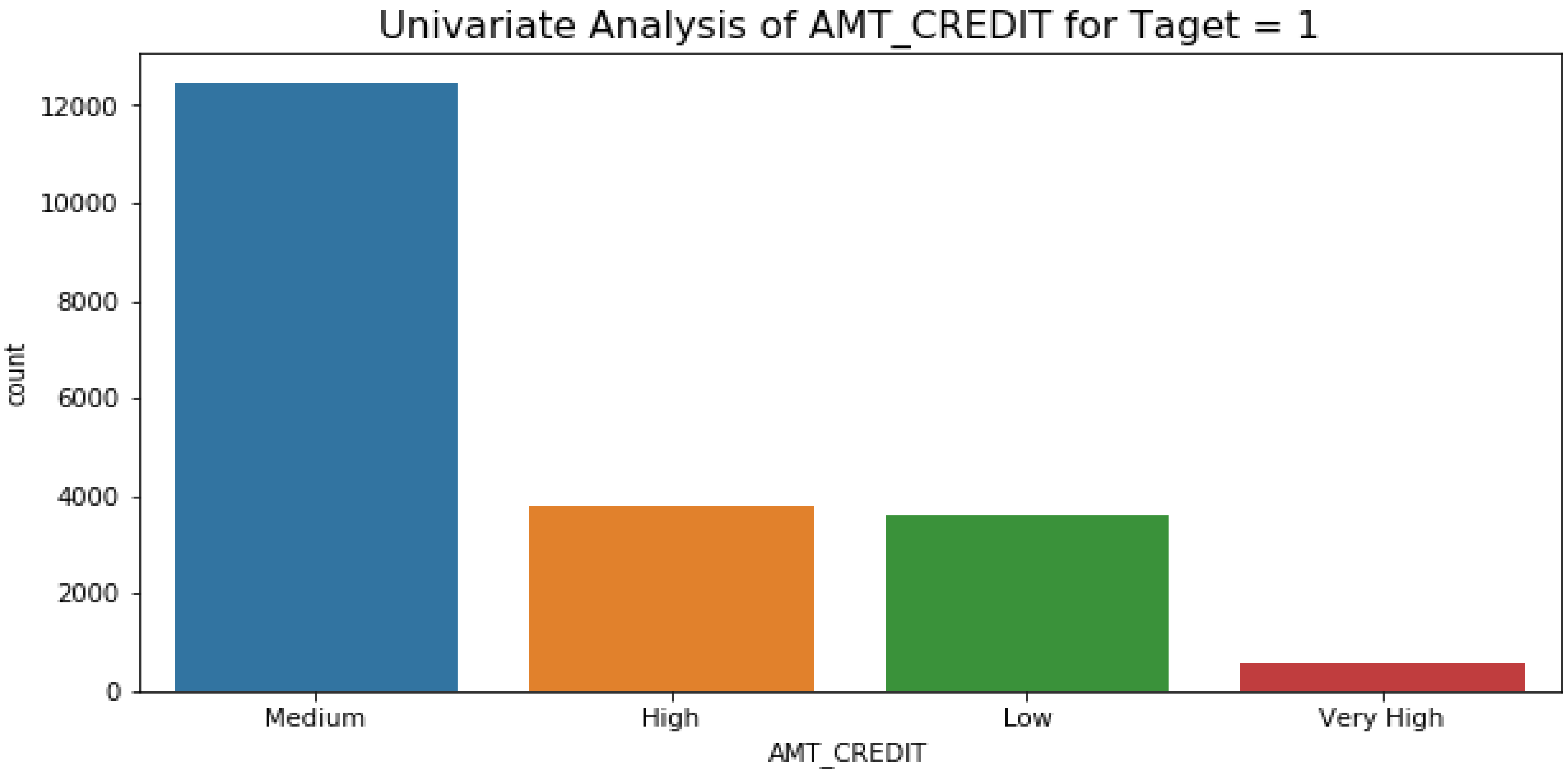
There is not much to choose here, but applicants with Average/Medium Loan Amount are the ones, who have most difficulties in repayment

DATA ANALYSIS
Check the imbalance percentage of target variable

- The imbalance percentage of the TARGET variable is 9.98
- We divided the data into 2 sets , target =0 and target=1
- Next is the univariate analysis for categorical variables for Target = 1 and Target = 0


Repayment Difficult
count
12000
10000
8000
6000
4000
2000
0
Medium
High
Low
Very High
AMT_CREDIT
No Repayment Difficult
count
100000
80000
60000
40000
20000
0
Very High
Low
Medium
High
AMT_CREDIT

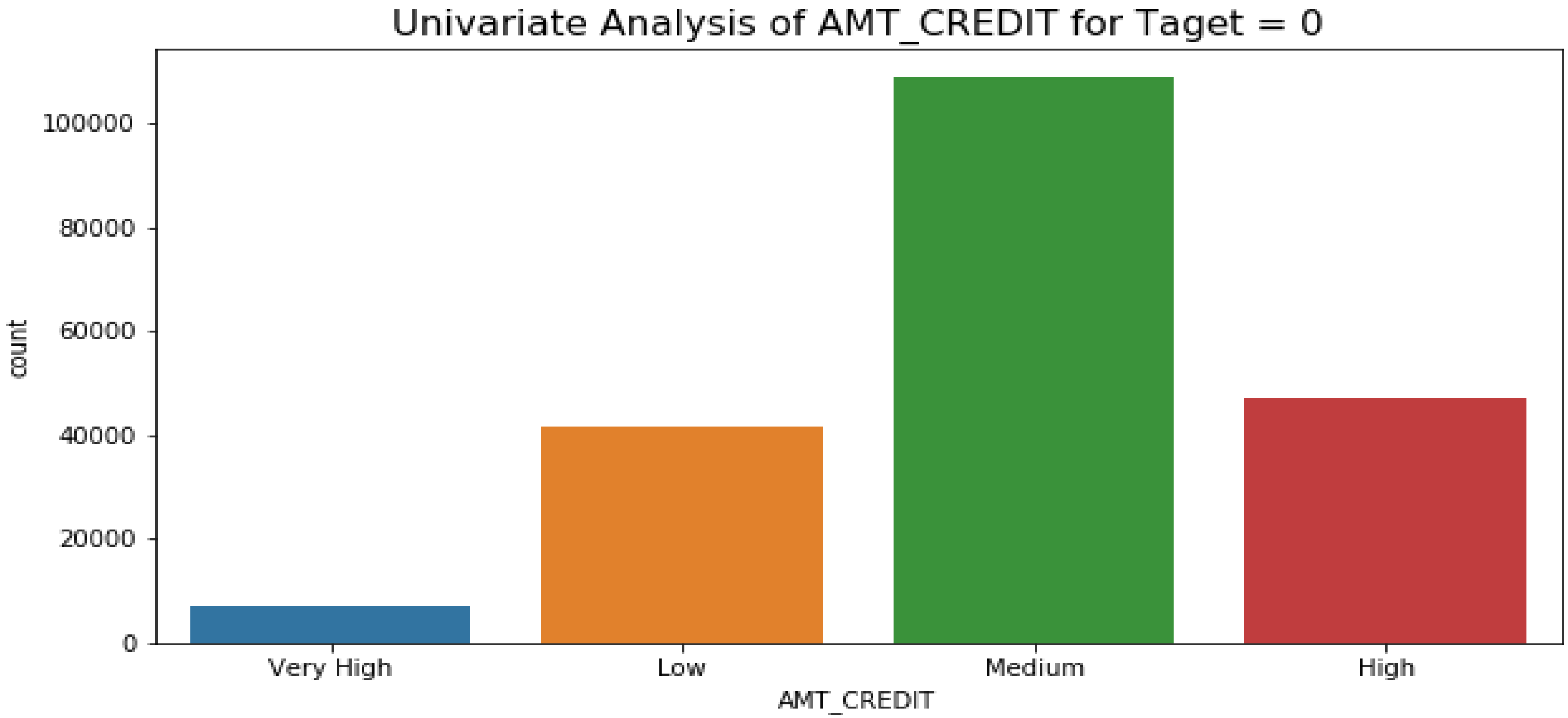
Univariate analysis of amt_credit for target =1



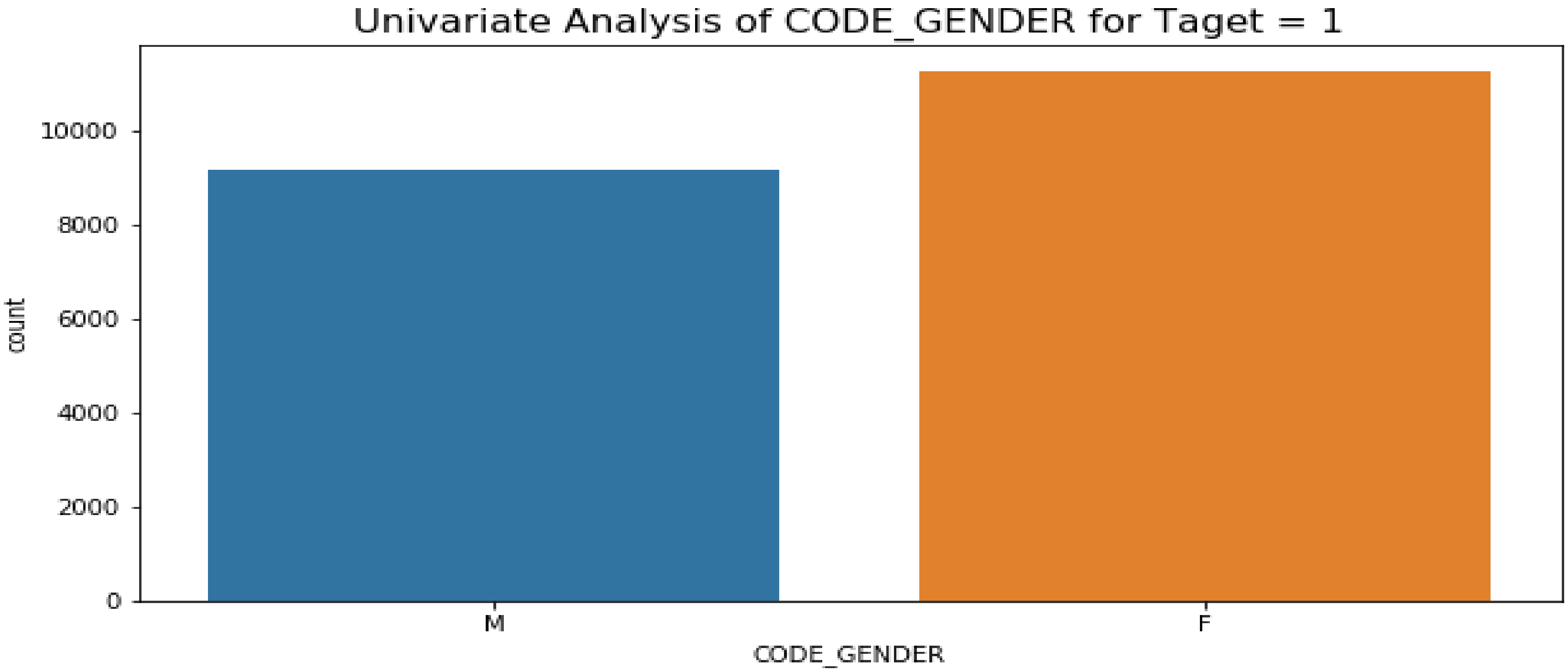
People with very high Credit amount(above 1250000) are having least difficulty in making repayment.

People with medium credit amount(Between 250000 and 750000) are the ones, who are the most defaulters

Univariate analysis of amt_credit for target =0

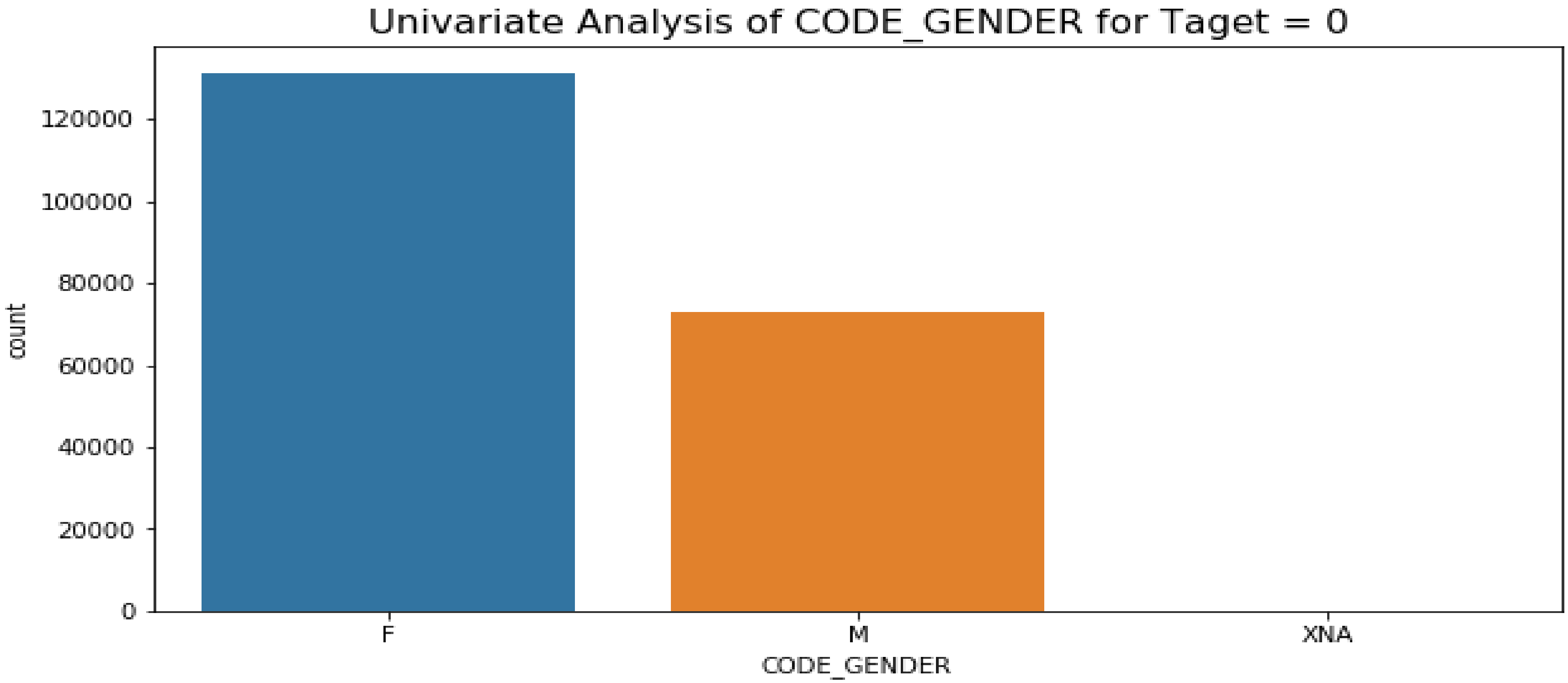


Univariate analysis of code_gender for target =1



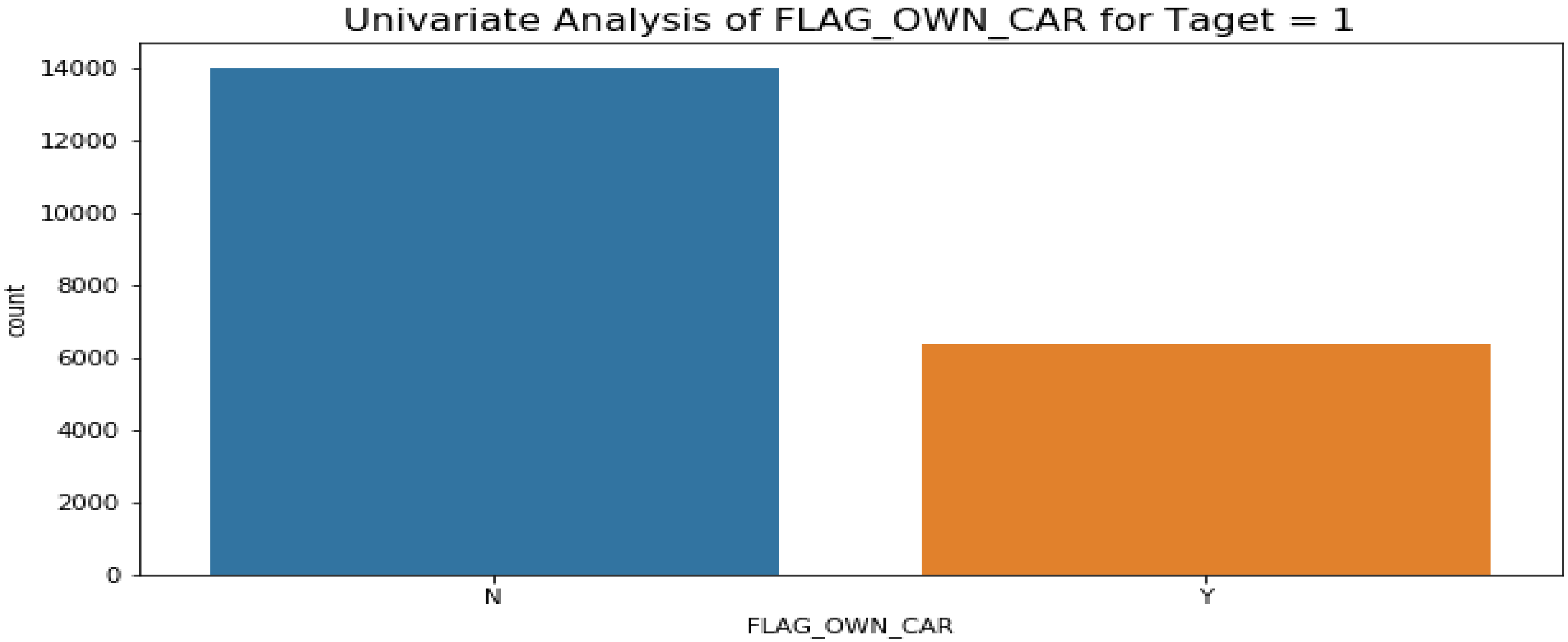
Women have more difficulty in repaying the loan amount as compare to Men

Univariate analysis of code_gender for target =0



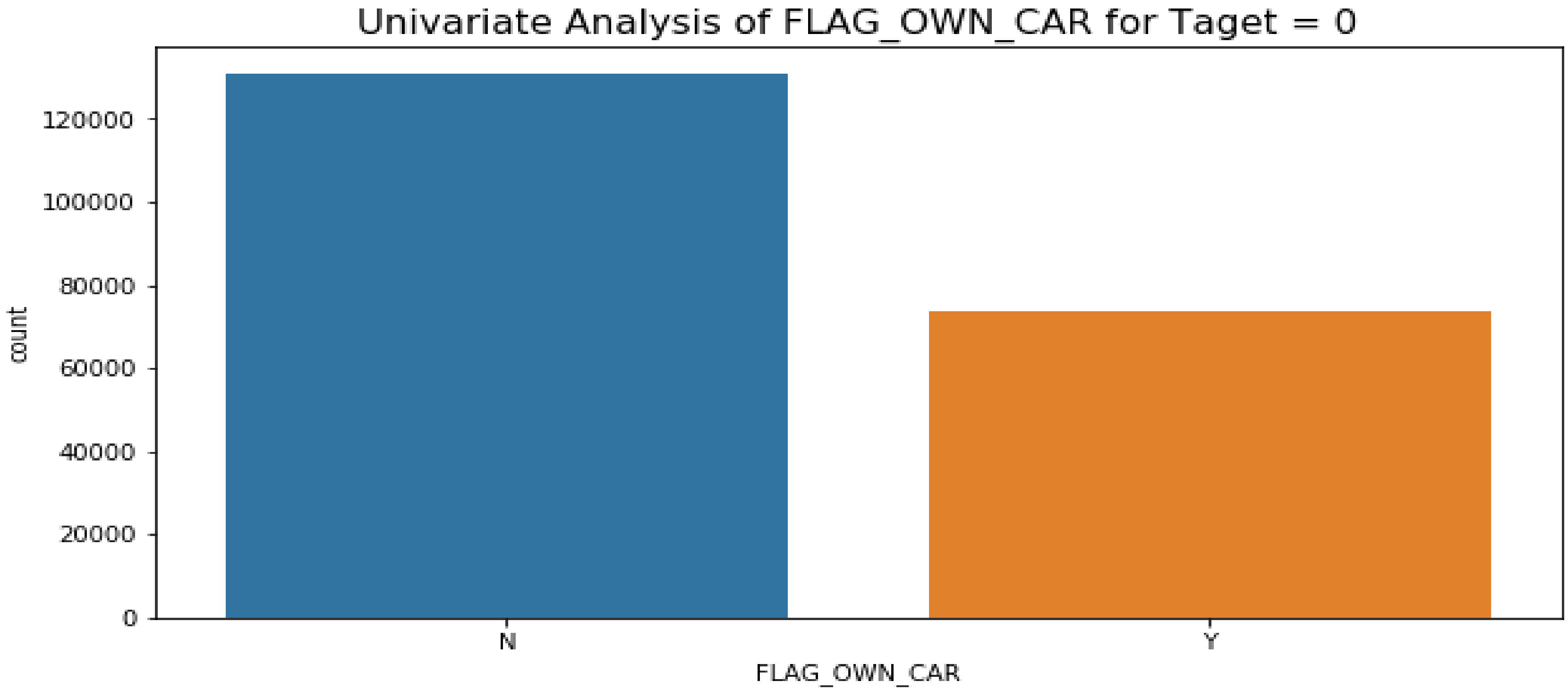
No of women applied for loan are higher than men
This also suggests that females are good in taking care of the loan repayment.

Univariate analysis of flag_own_car for target =1



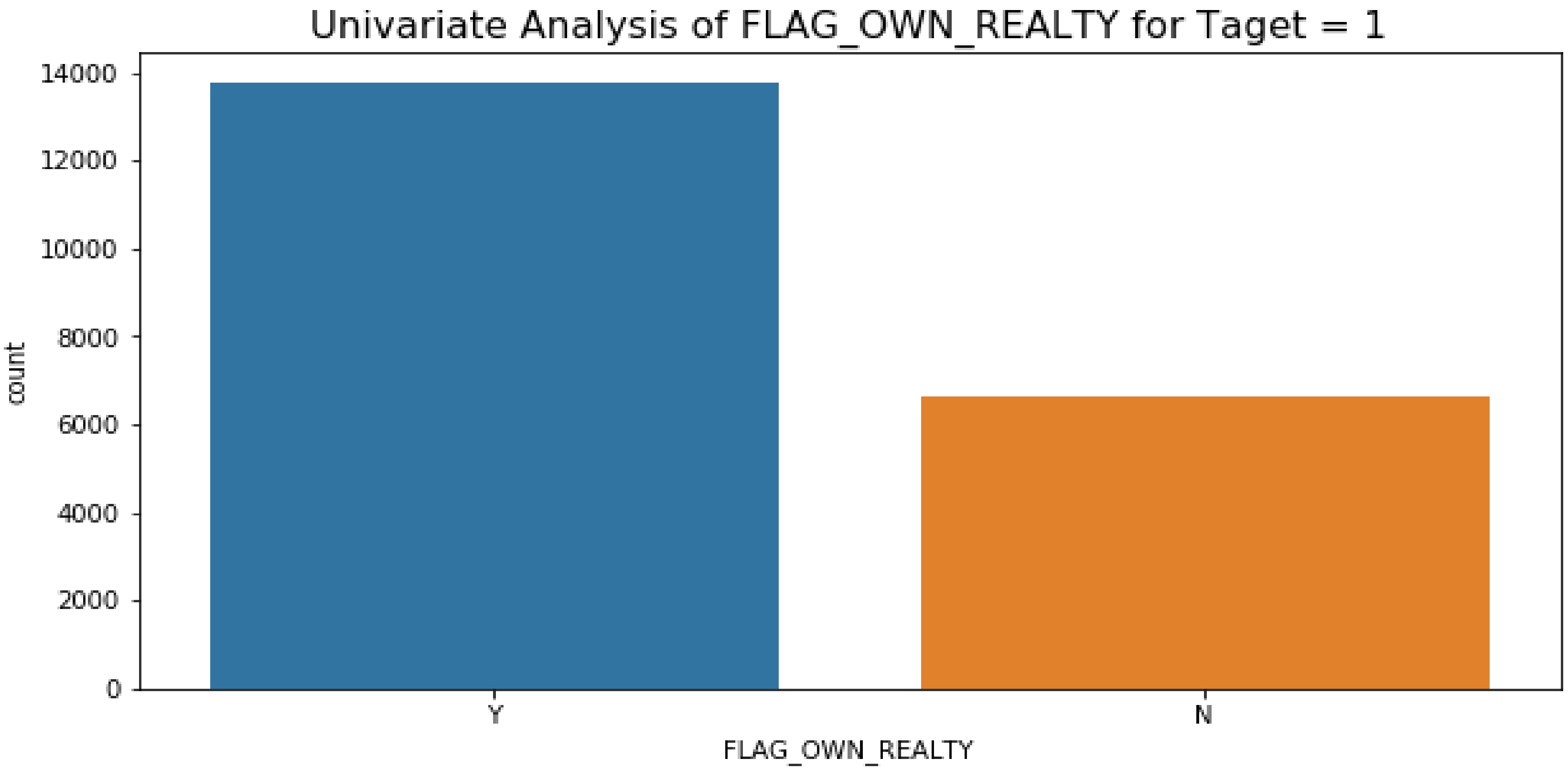
People who do not have own a car have difficulties in making payment, while the ones, who own car/s are having less difficulty in repayment

Univariate analysis of flag_own_car for target =0



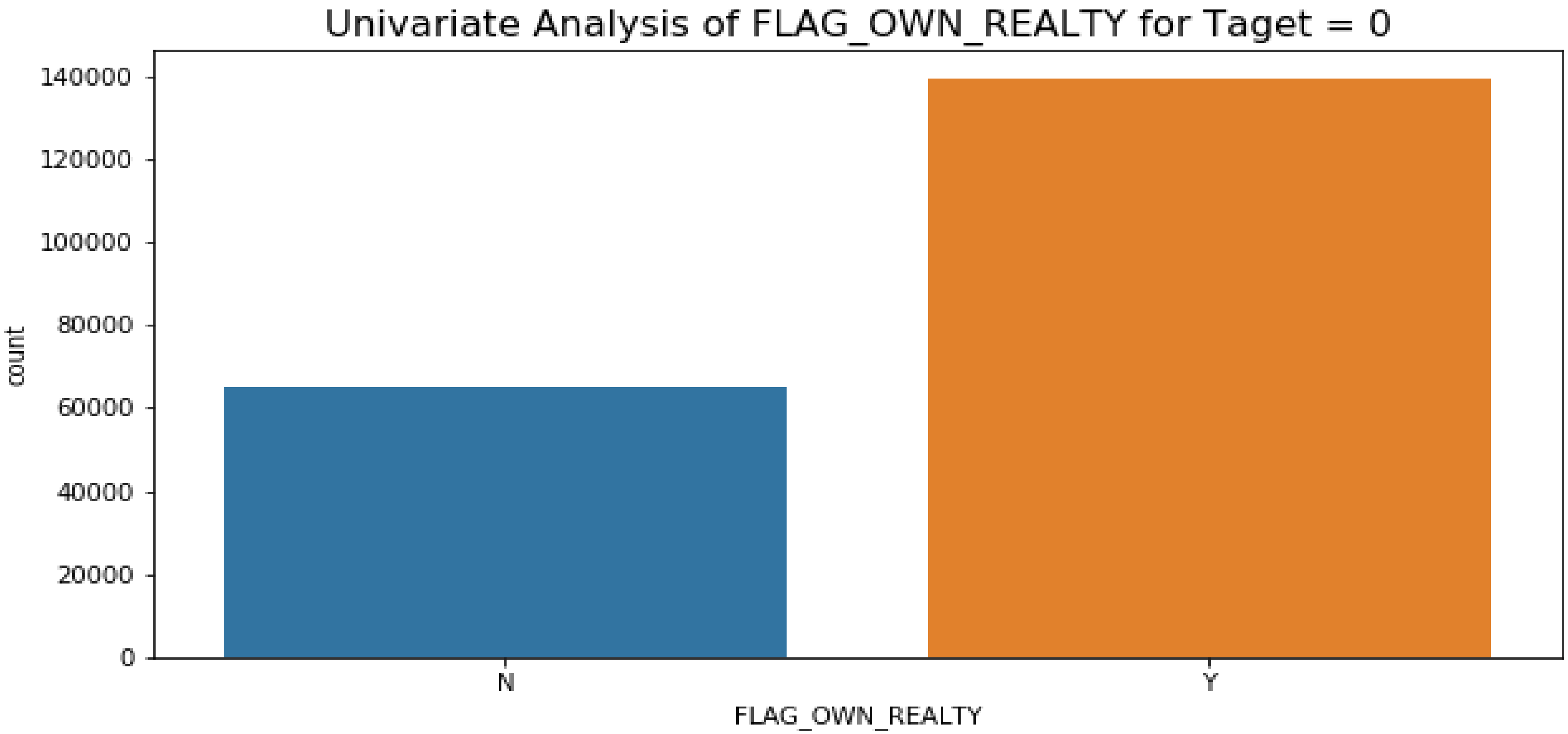
There are ~130000 people who do not have car and applied for a loan. And they do not difficulties in making repayment.
There are ~740000 people who have car and are in need of money

Univariate analysis of flag_own_realty for target =1



People who do not own Realty have less difficulty in making repayment, while the ones who own house or any other realties, have difficulties in repayment.

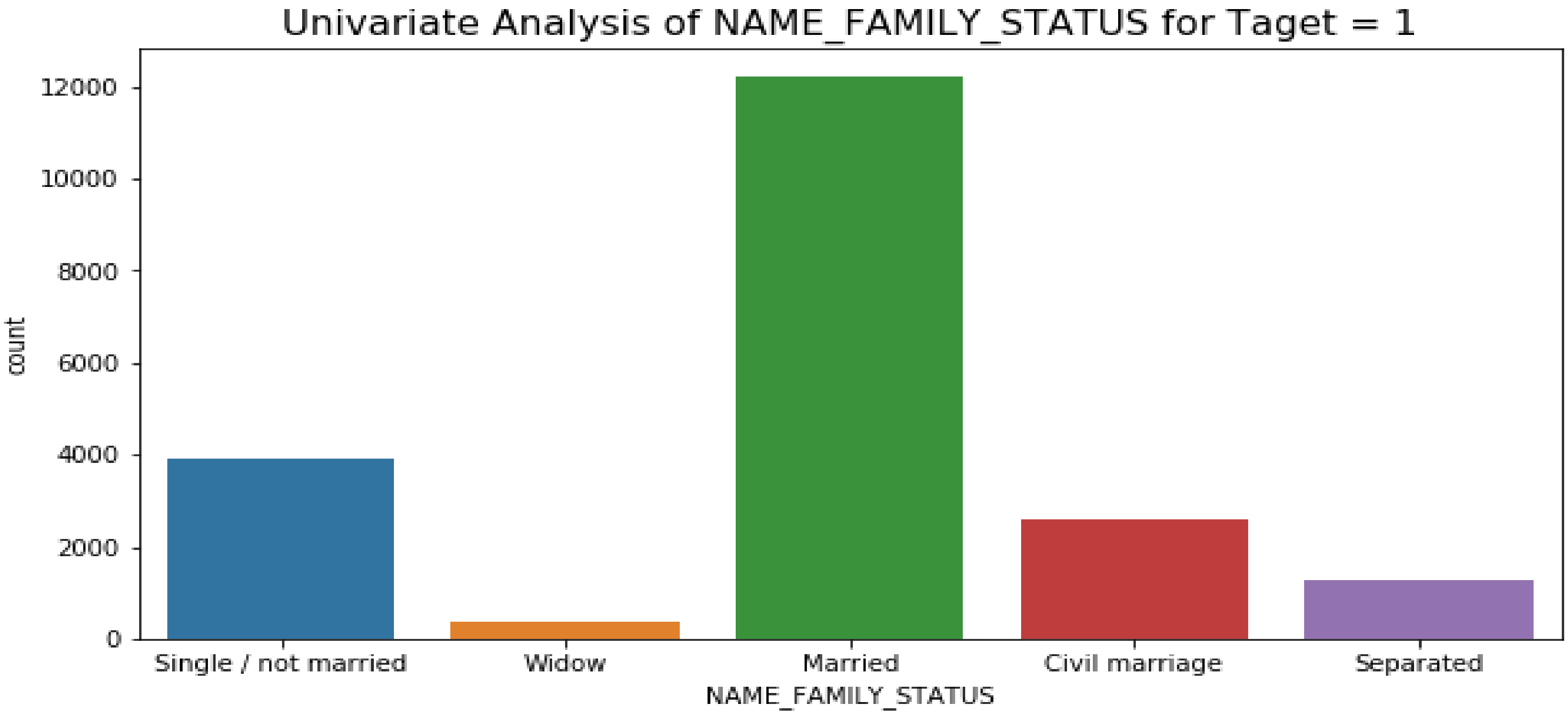
Univariate analysis of flag_own_reality for target =0



There are approximately 140000 applicants, who own a realty and applied for a loan. They are good in making repayment as well, as they do not face difficulty in paying EMIs.

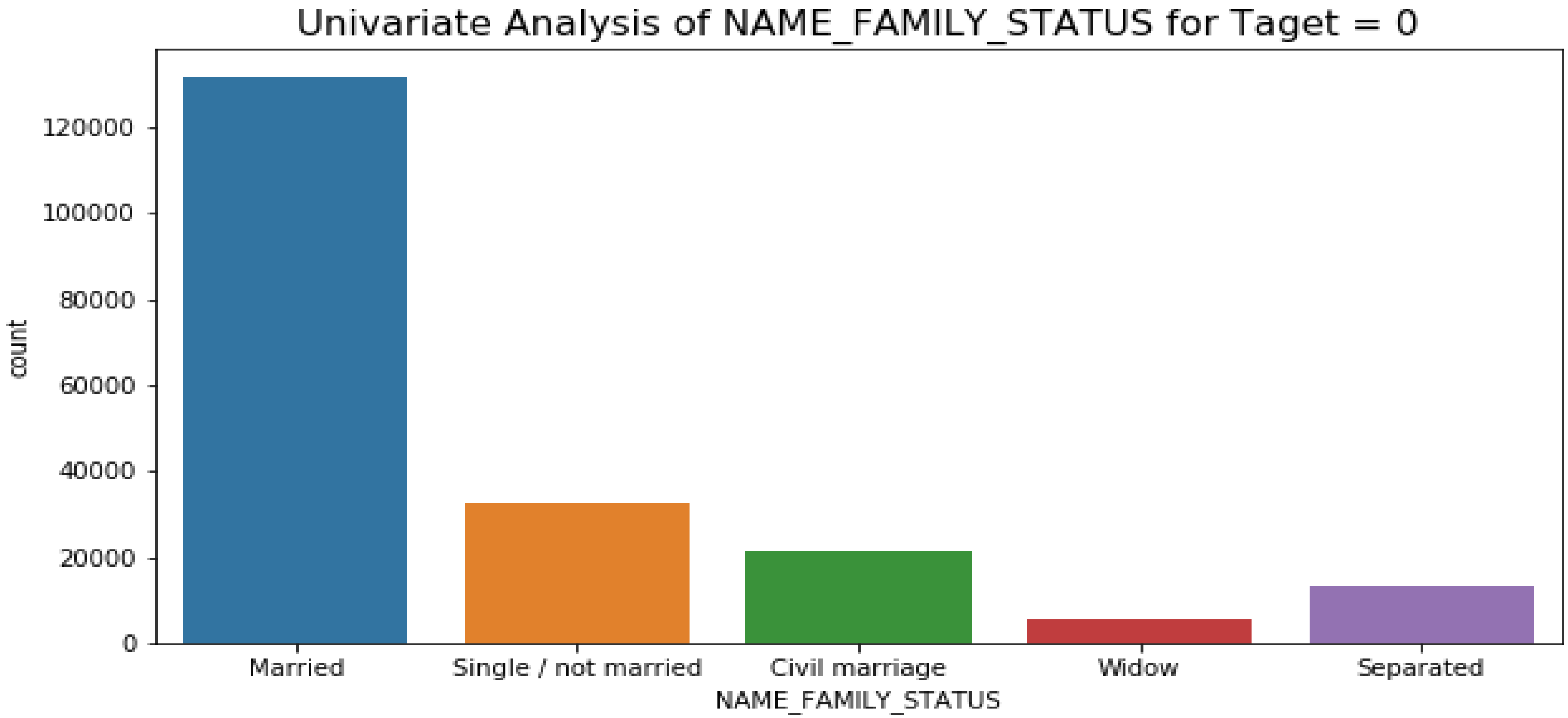
The people who do not own a realty are almost half of the ones, who own.

Univariate analysis of name_family_status for target =1



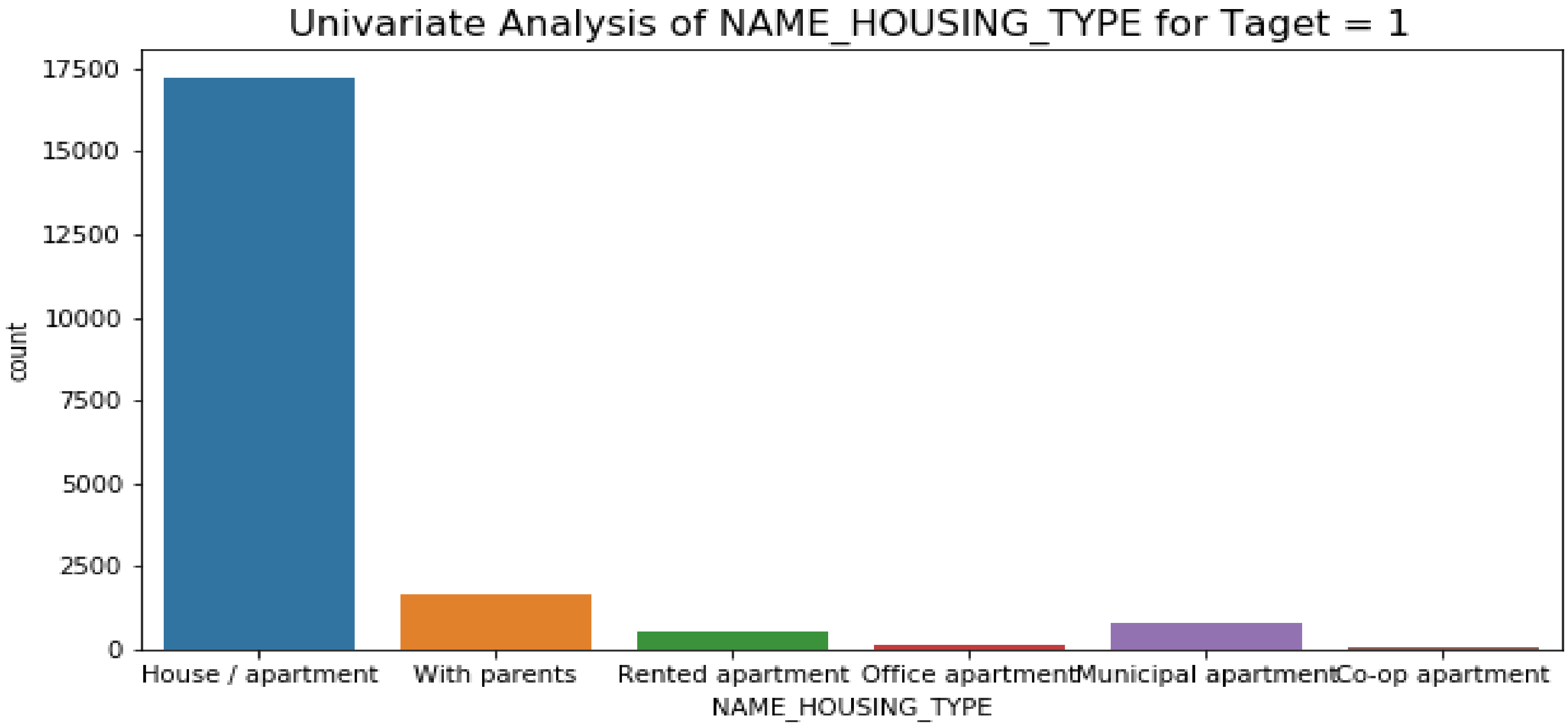
People who are married, have difficulties in making payment, while widows have least difficulties, among the defaulters

Univariate analysis of name_family_status for target =0



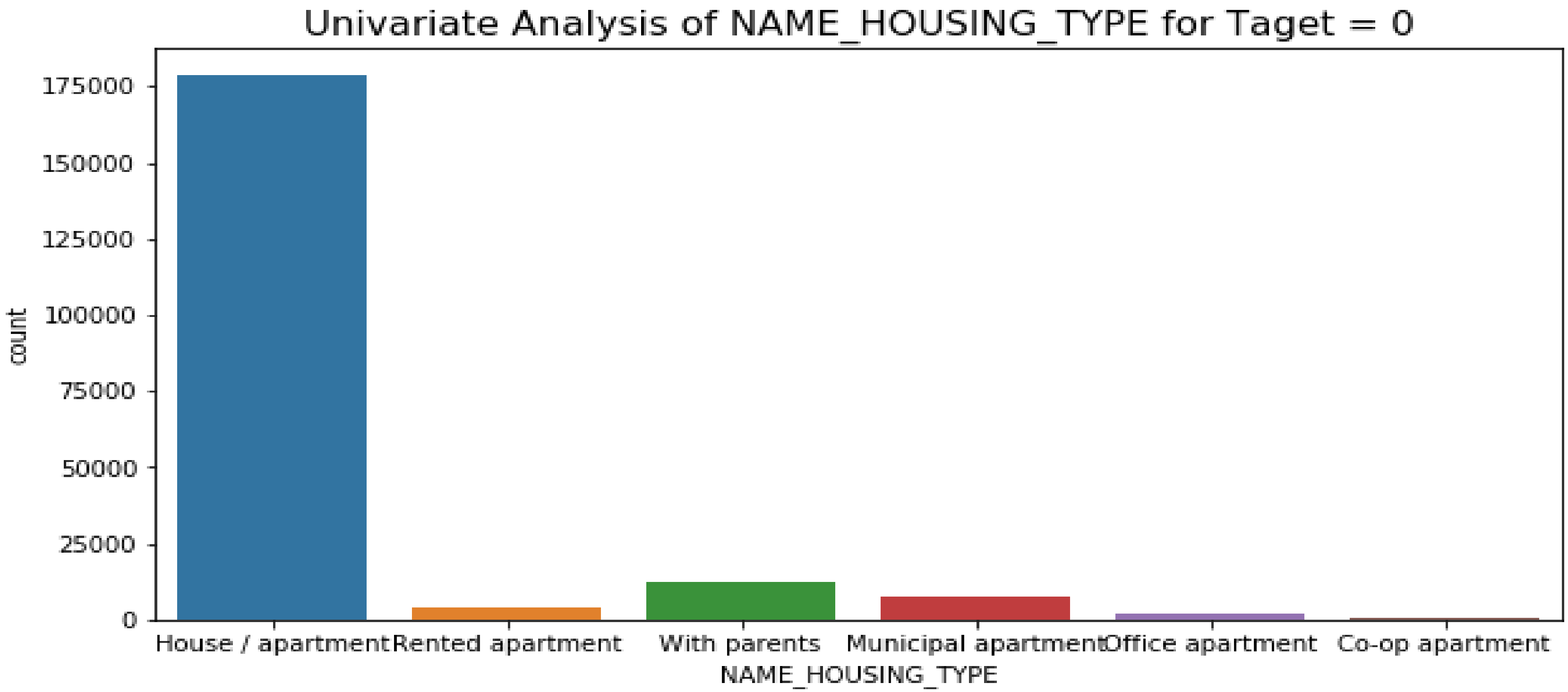
The graph suggests that the people who are married, are the ones who need loans

Univariate analysis of name_housing_type for target =1



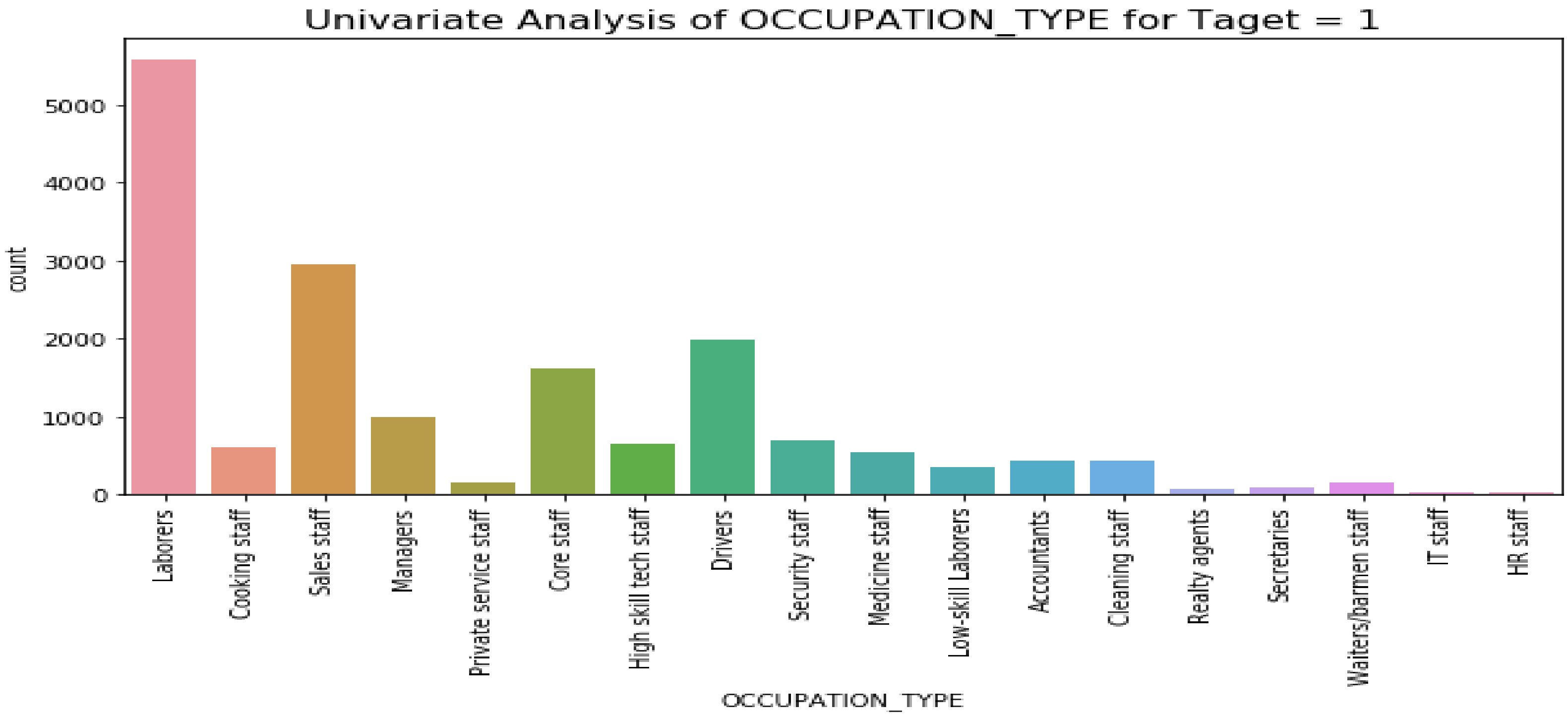
Applicants who live in an apartment or house are having more difficulties than others

Univariate analysis of name_housing_type for target =0



People who live in co-op apartment are the least in numbers, who need a loan

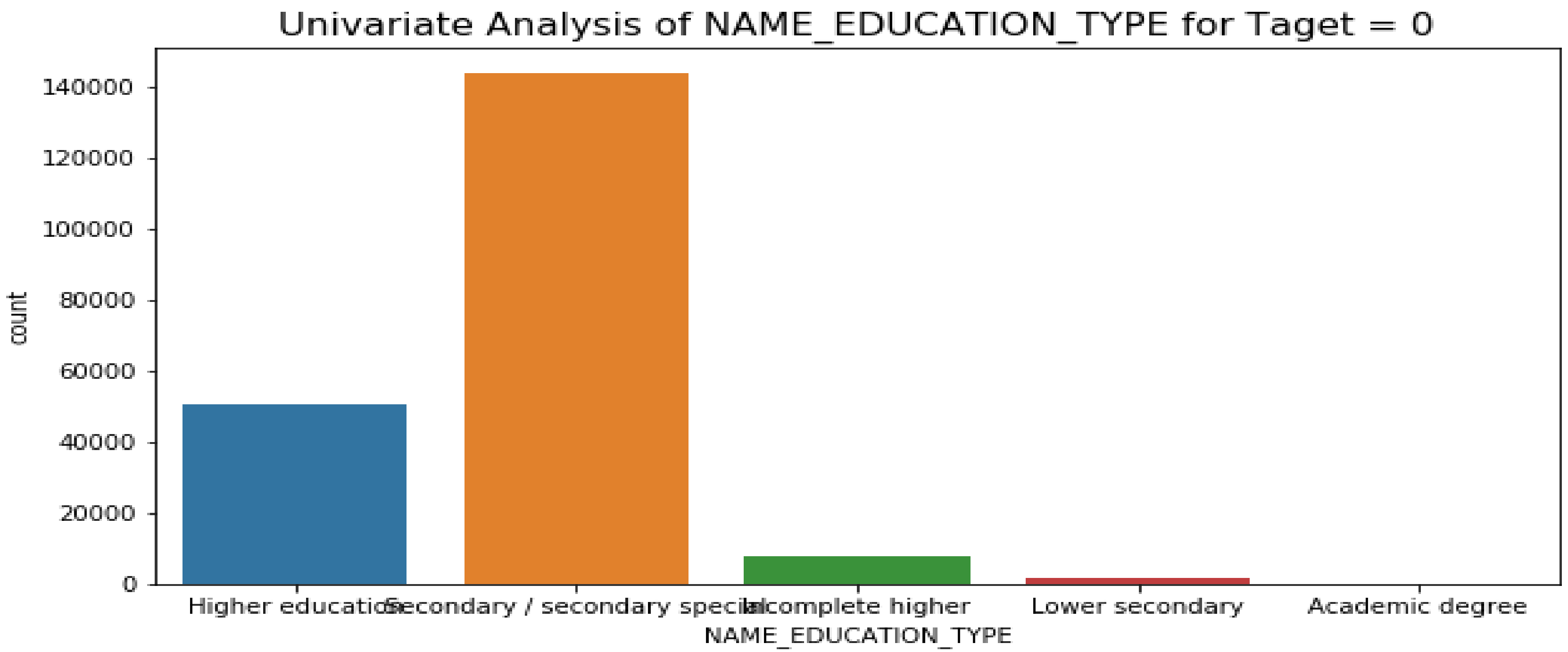
Univariate analysis of name_education_type for target =1



People who are Secondary educated, have more difficulties in making repayments than the rest.

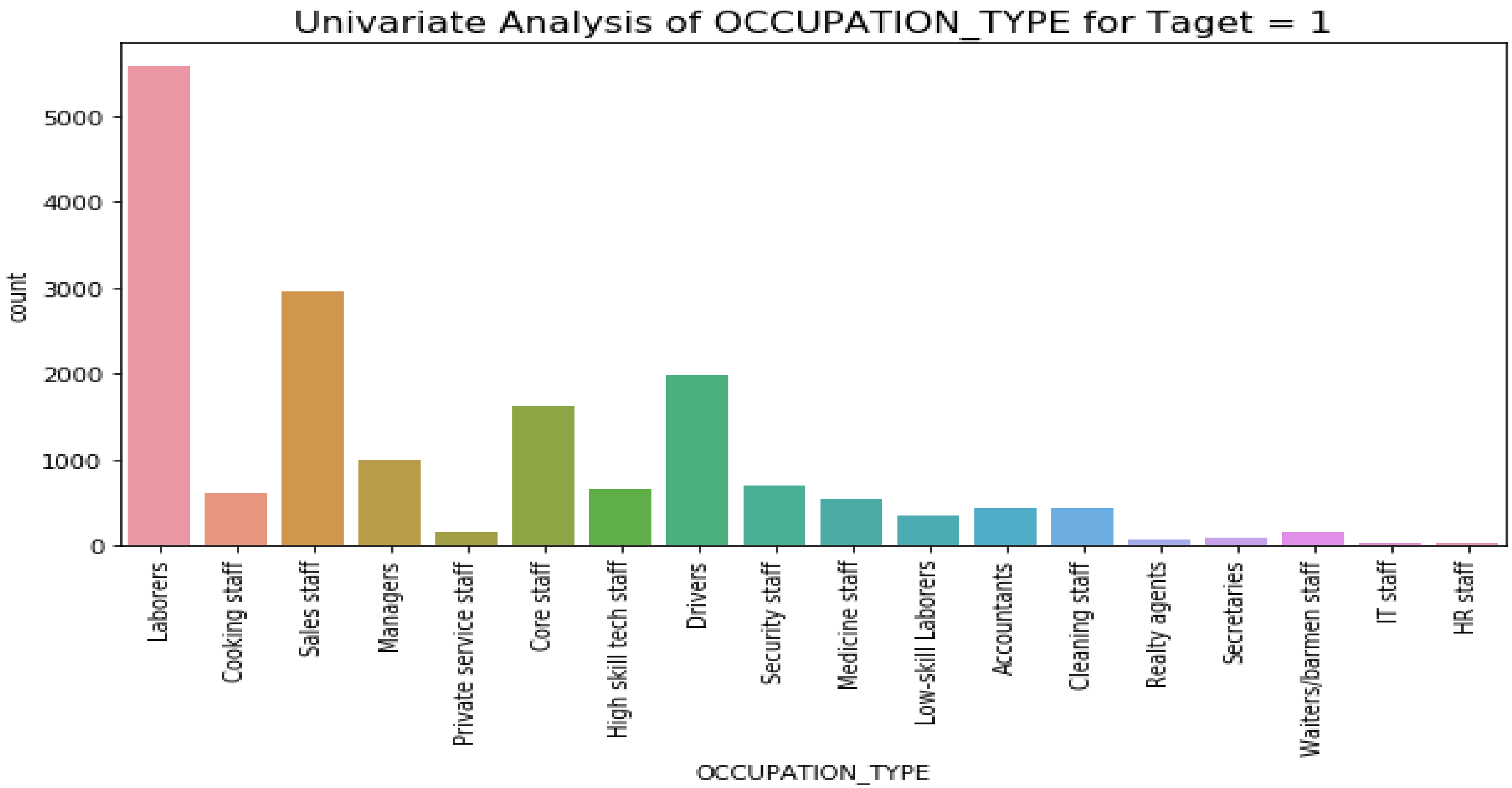
People who are educated till Lower Secondary are the least defaulters

Univariate analysis of name_education_type for target =0



People who have an Academic degree are the least who need a loan.
It seems, people who have higher education or the ones, who have secondary/secondary are higher in numbers for a a loan application

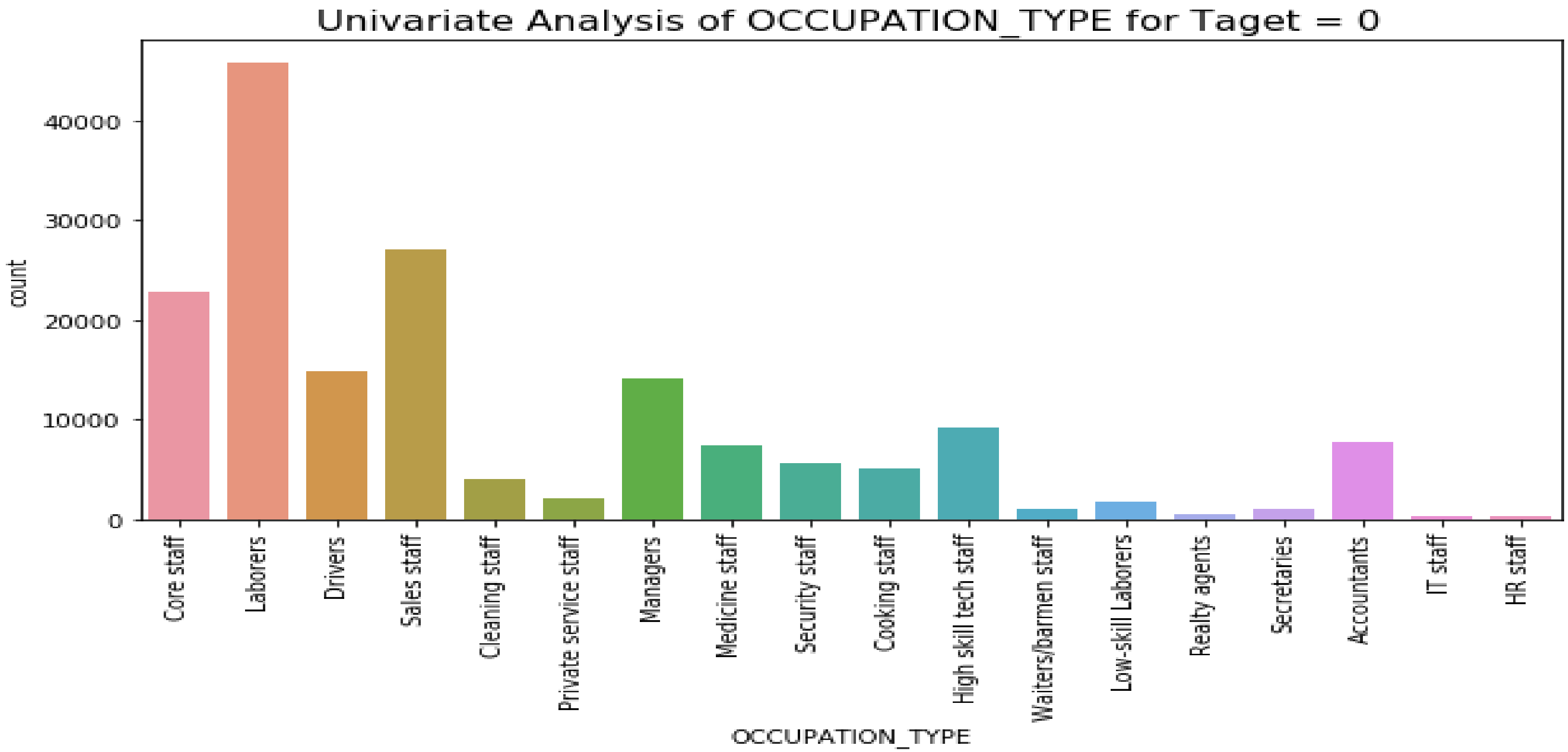
Univariate analysis of occupation_type for target =1



Applicants who are labourers, have more difficulties in making payment, while the IT and HR staffs are the ones, who least difficulties.

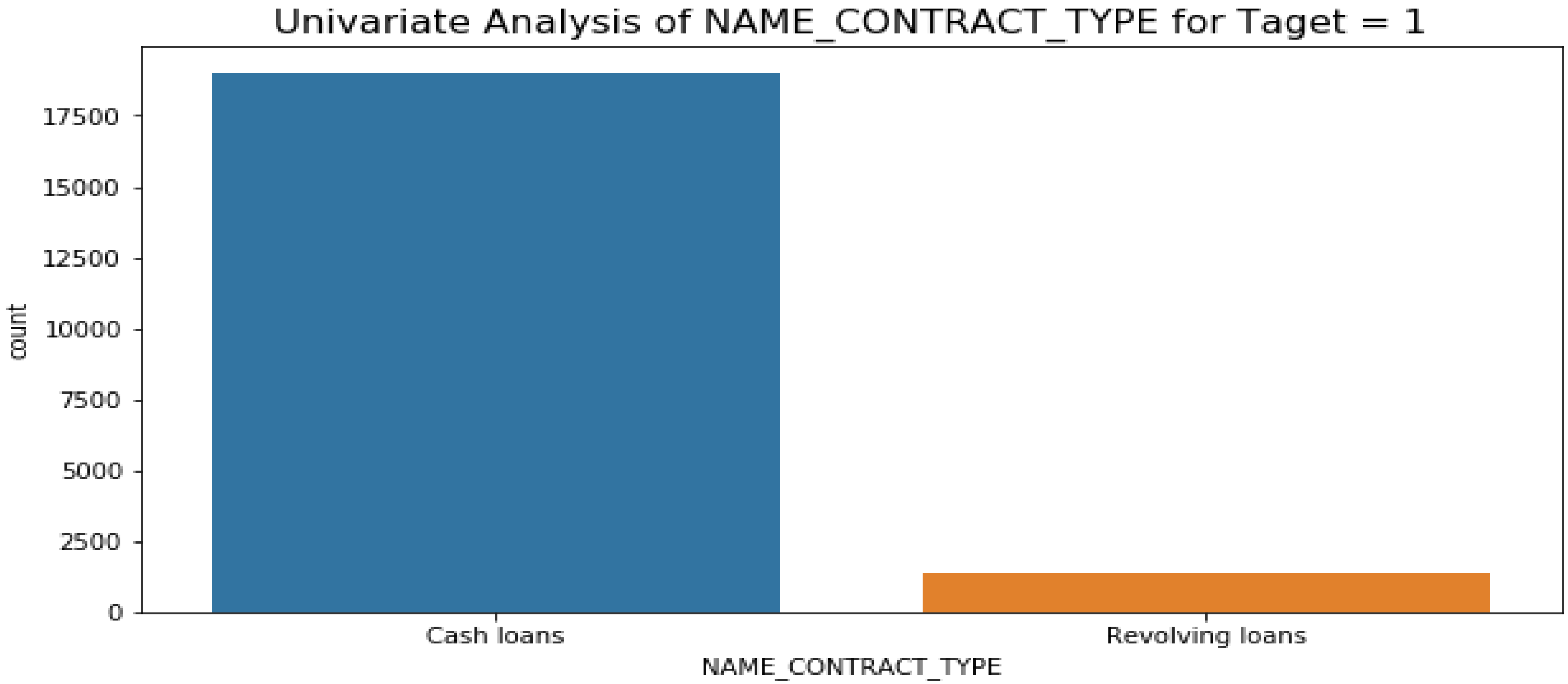
Interestingly, Highly skilled people are also among the defaulters

Univariate analysis of occupation_type for target =0



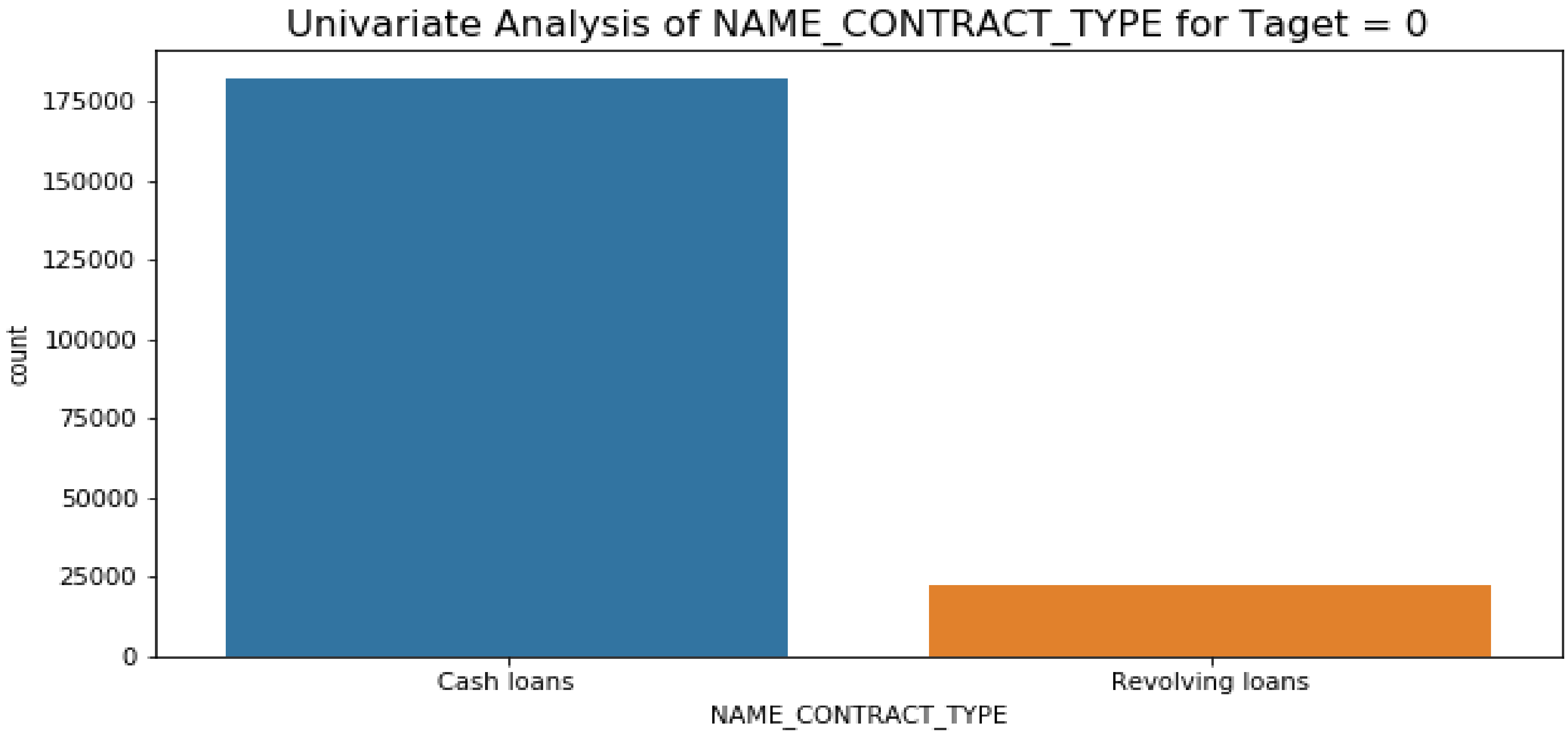
Labourers are very high in numbers, while IT, HR, Realty employee are very less, when it comes to taking a loan

Univariate analysis of name_contract_type for target =1



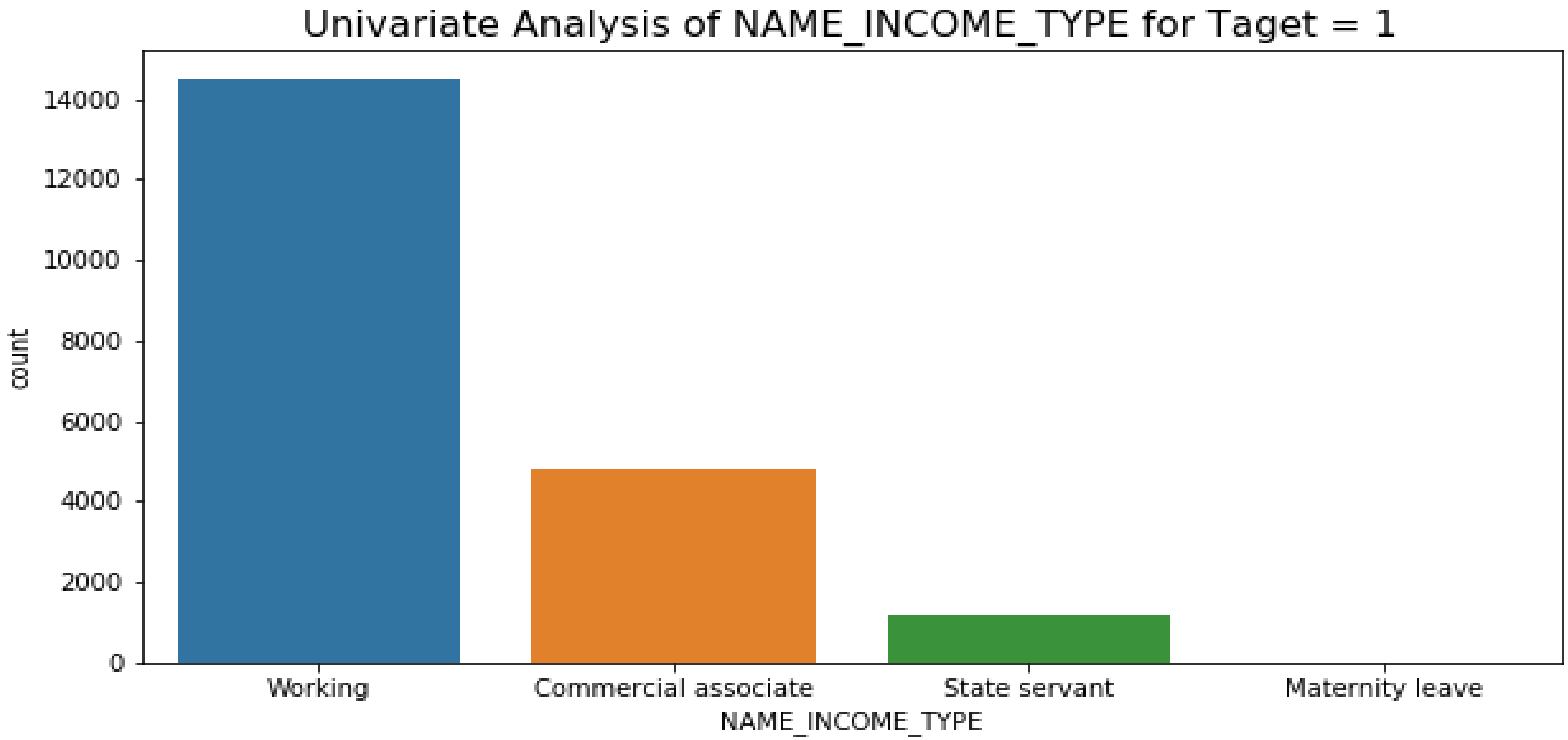
Cash loans have greater payment defaulter rate than any other types of loan

Univariate analysis of name_contract_type for target =0



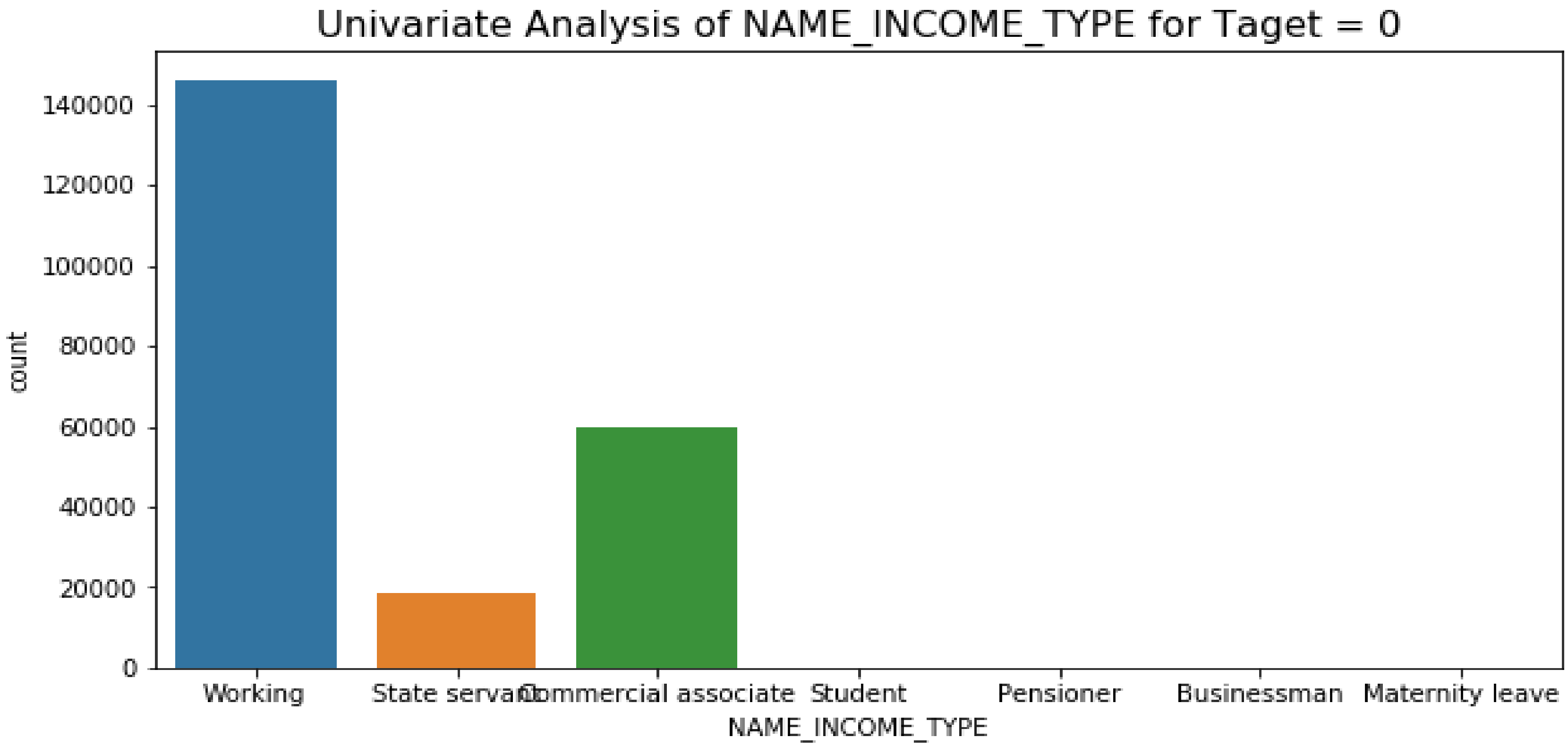
People are more in need of cash loans than any other types of loan.
Cash loan applications are 8 times higher than Revolving loans.

Univariate analysis of name_income_type for target =1



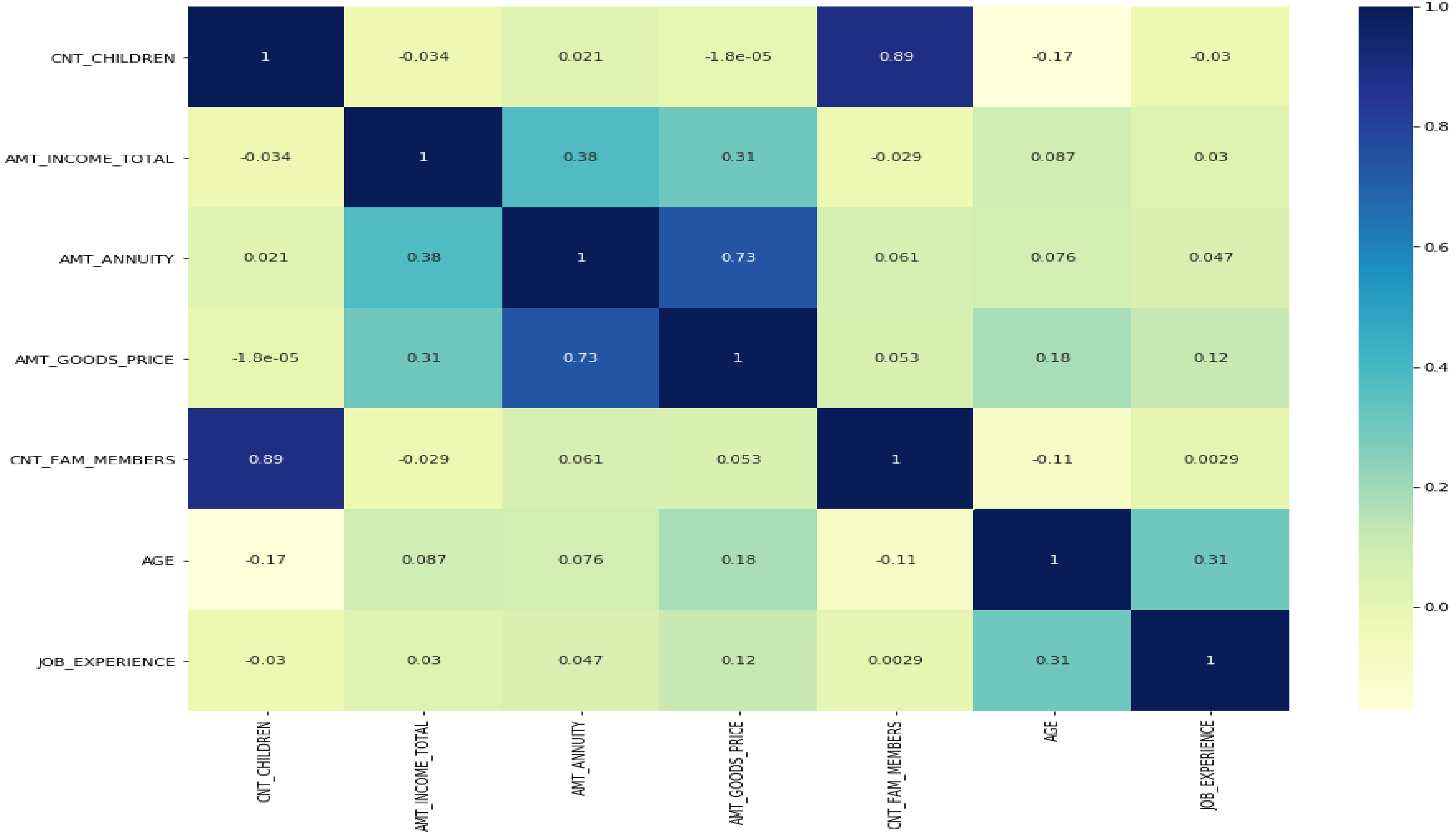
Working clients are ones, who are on the top of the defaulter list

Univariate analysis of name_income_type for target =0



Working People are the ones, who apply most for the loans

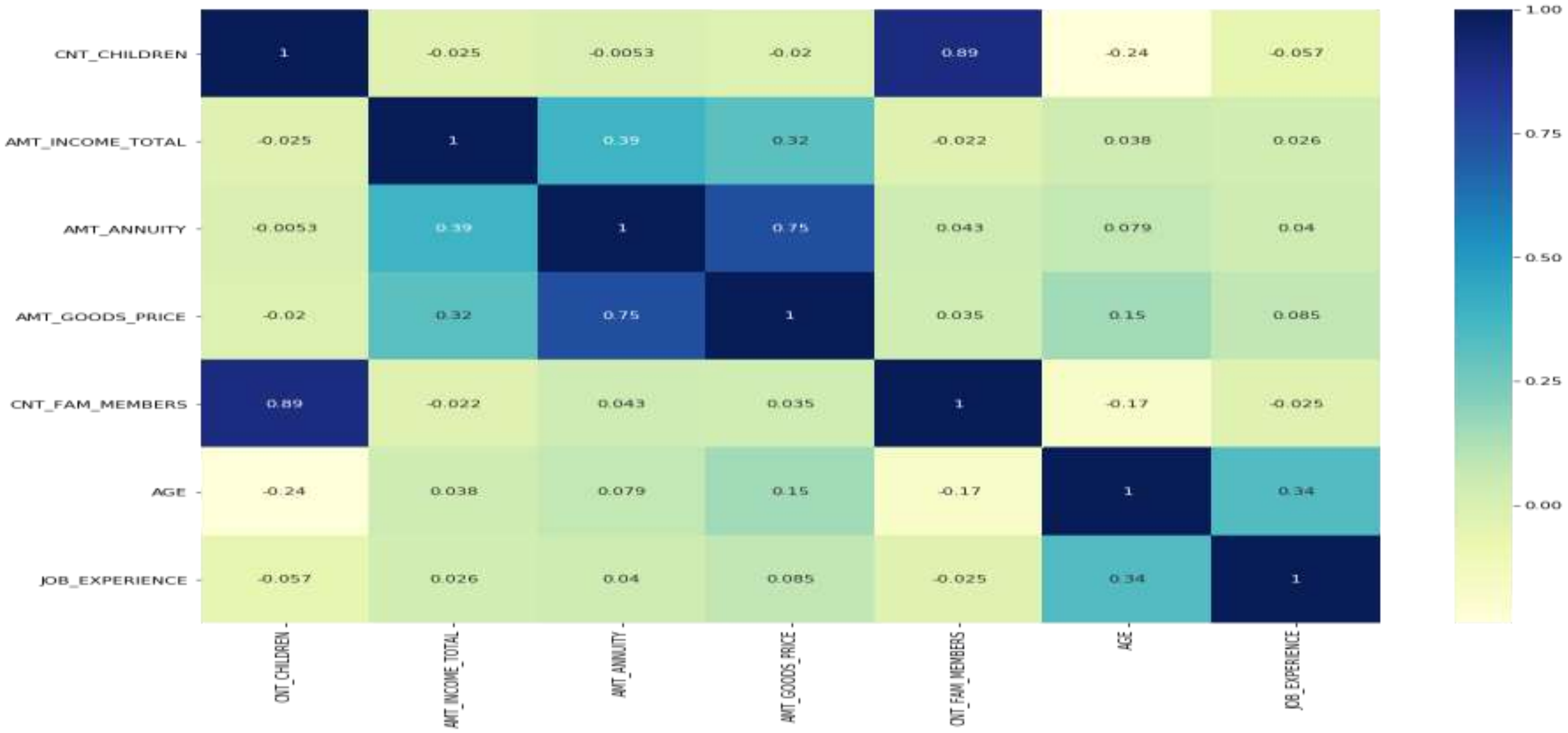
correlations for numeric columns for both the Targets



	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS	AGE	JOB_EXPERIENCE
CNT_CHILDREN	1	-0.034	0.021	-1.8e-05	0.89	-0.17	-0.03
AMT_INCOME_TOTAL	-0.034	1	0.38	0.31	-0.029	0.087	0.03
AMT_ANNUITY	0.021	0.38	1	0.73	0.061	0.076	0.047
AMT_GOODS_PRICE	-1.8e-05	0.31	0.73	1	0.053	0.18	0.12
CNT_FAM_MEMBERS	0.89	-0.029	0.061	0.053	1	-0.11	0.0029
AGE	-0.17	0.087	0.076	0.18	-0.11	1	0.31
JOB_EXPERIENCE	-0.03	0.03	0.047	0.12	0.0029	0.31	1

Inference from the previous graph

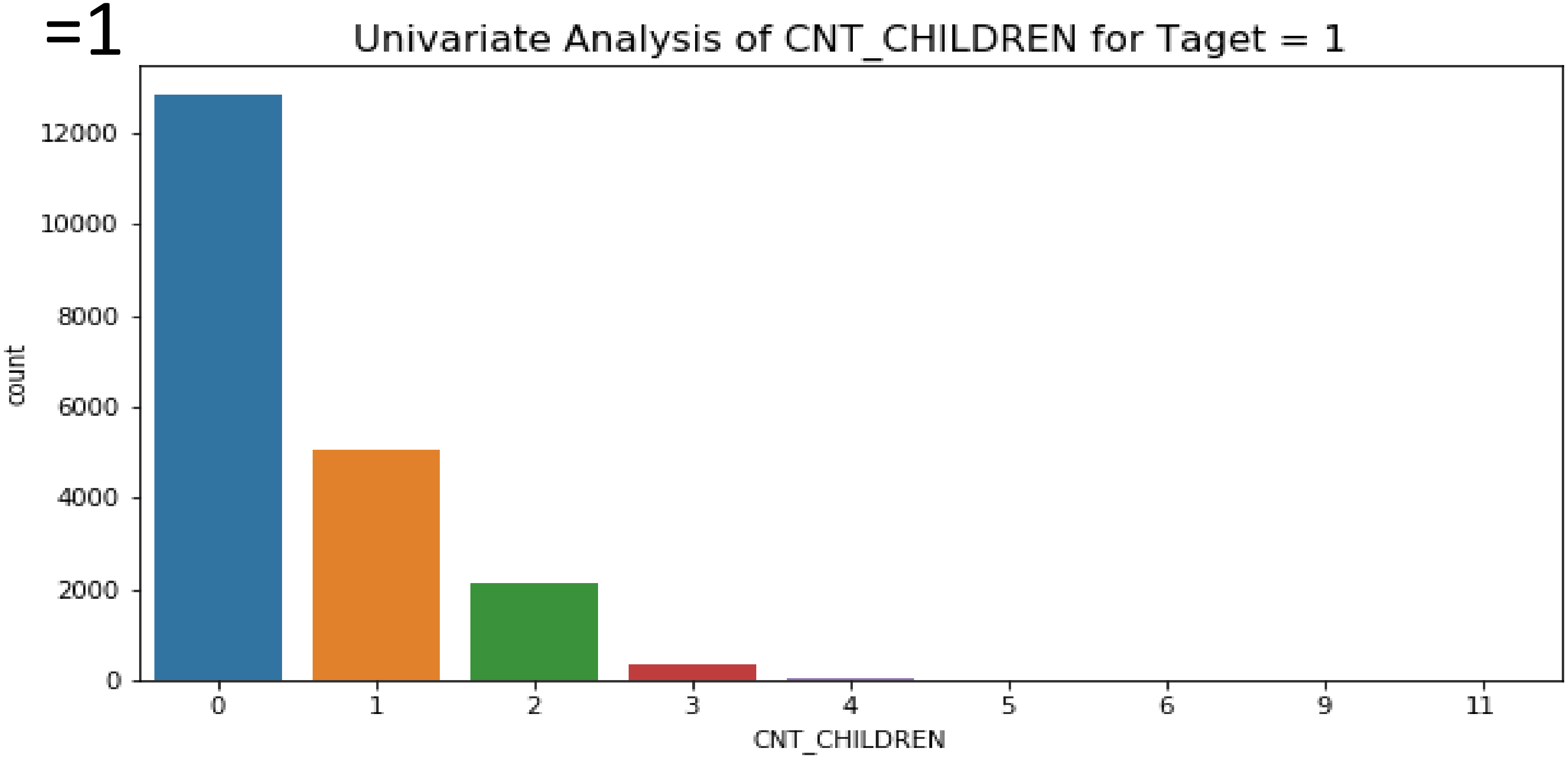
- As per the analysis:
- There is a high correlation between CNT_CHILDREN and CNT_FAM_MEMBERS
- There is good correation between AMT_ANNUITY and AMT_GOODS_PRICE. Higher the price of the goods, higher the loan given to the customers.
- There is negative between AMT_INCOME_TOTAL and CNT_FAM_MEMBERS. As the no of members in the family increases, the total income of the Customer decreases.



There is a high correlation between CNT_CHILDREN and CNT_FAM_MEMBERS
There is good correlation between AMT_ANNUITY and AMT_GOODS_PRICE. Higher the price of the goods, higher the loan given to the customers.
There is negative between AMT_INCOME_TOTAL and CNT_FAM_MEMBERS. As the no of members in the family increases, the total income of the Customer decreases
Check, if the variables with highest correralation are the same
Yes, the variables with highest correlations are the same i.e. CNT_FAM_MEMBERS and CNT_CHILDREN

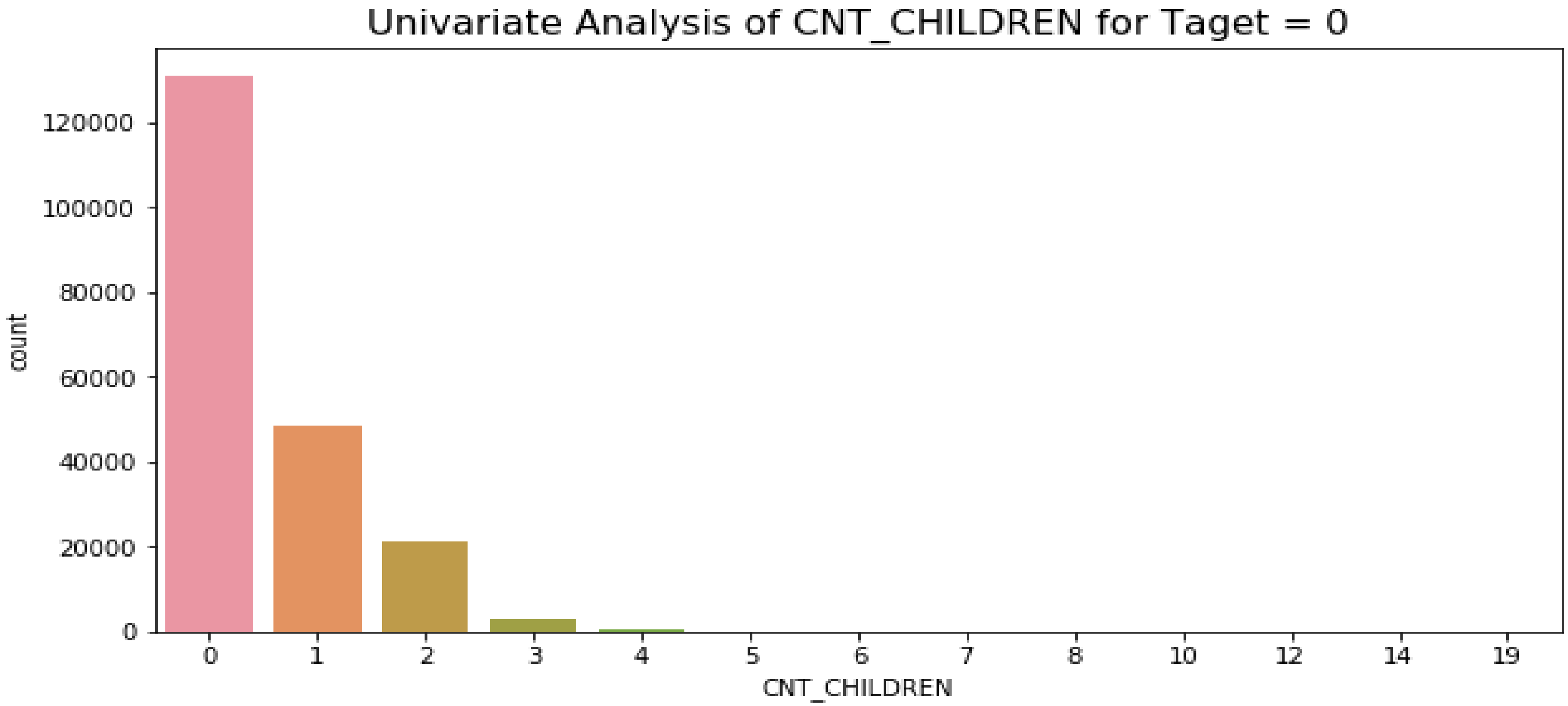
Univariate analysis for categorical variables for Target = 1 and Target = 0

- Univariate analysis of cnt_children for target =1



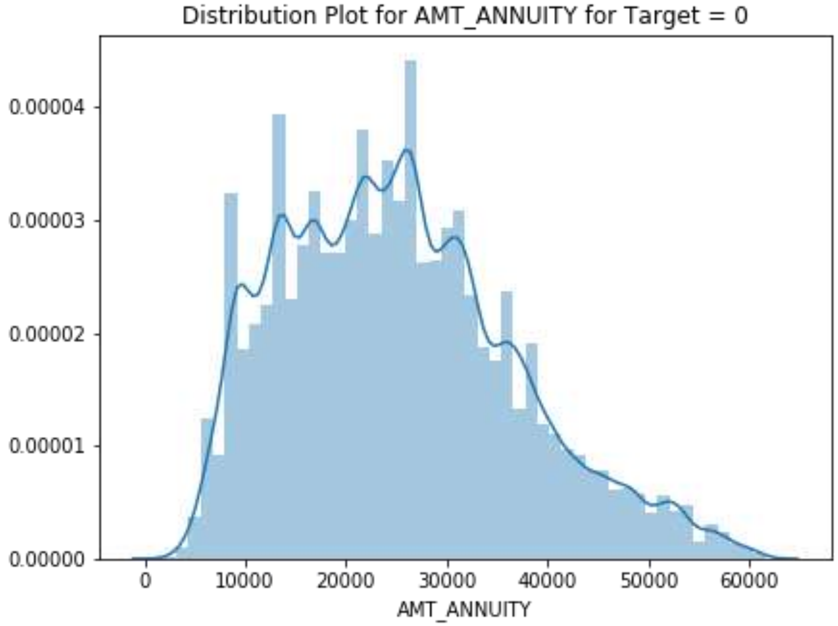
Clients with no children are riskier than clients, having childrens

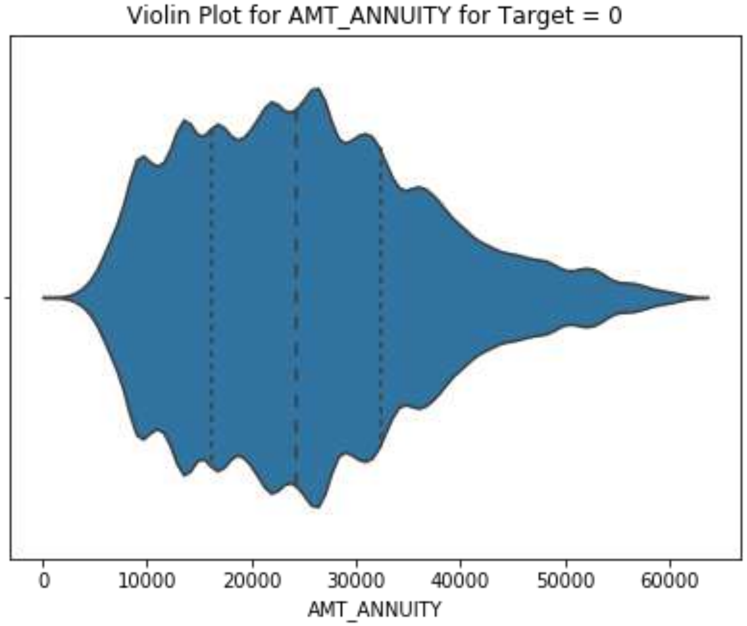
Univariate analysis of cnt_children for target =0



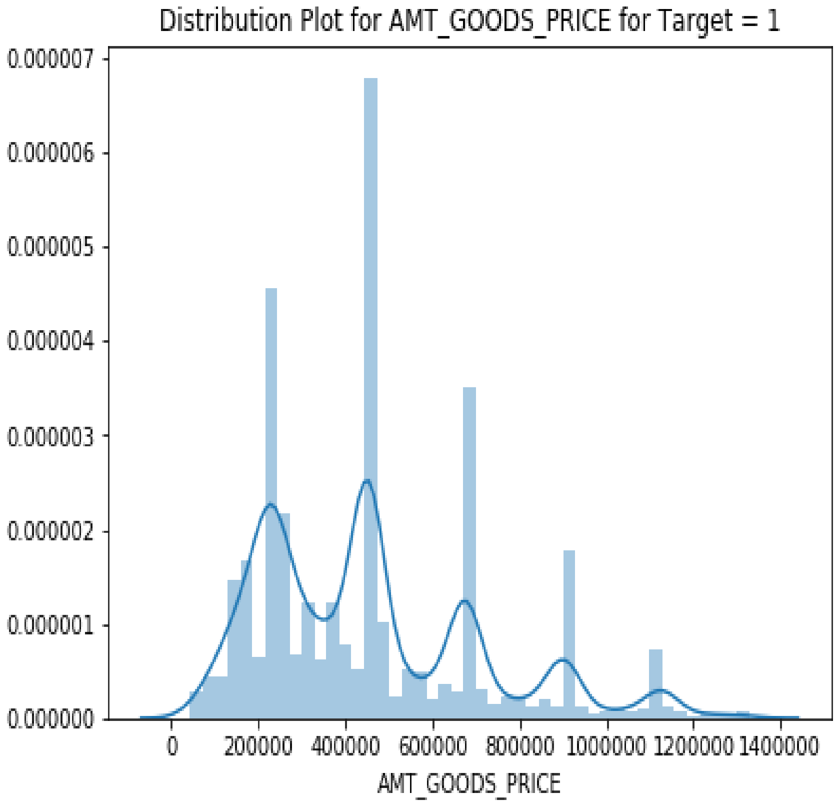
There are approximately ~131000 customers with no children.
There is only one client with the highest no of children(12)
Most of the clients are either having 1 or 2 children

Univariate analysis for categorical variables for Target = 1 and Target = 0

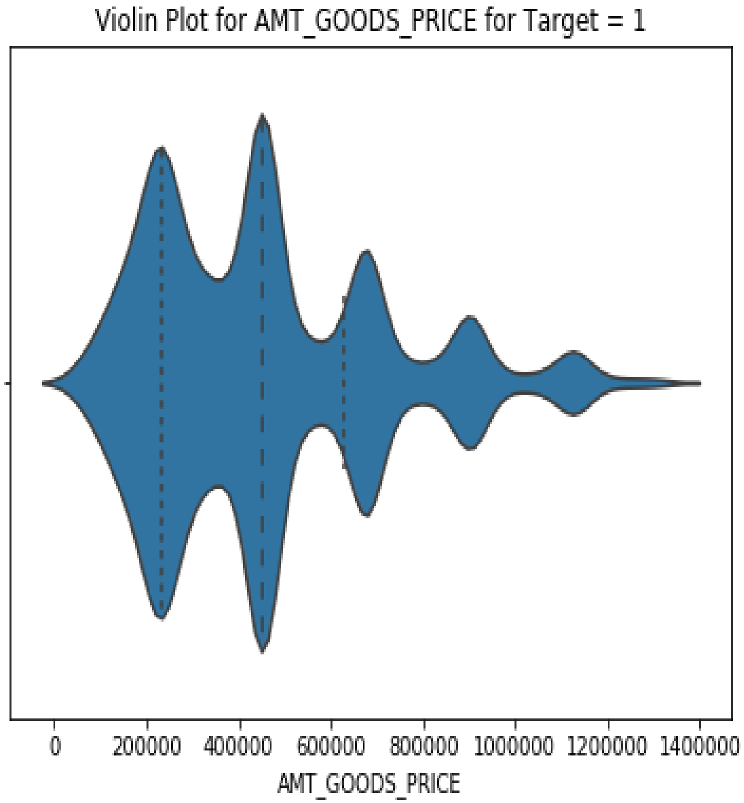




Most of the Amount Annuity is between ~24000 and ~33000.
Minimum annuity is closed to 2000, while the maximum is around 61000



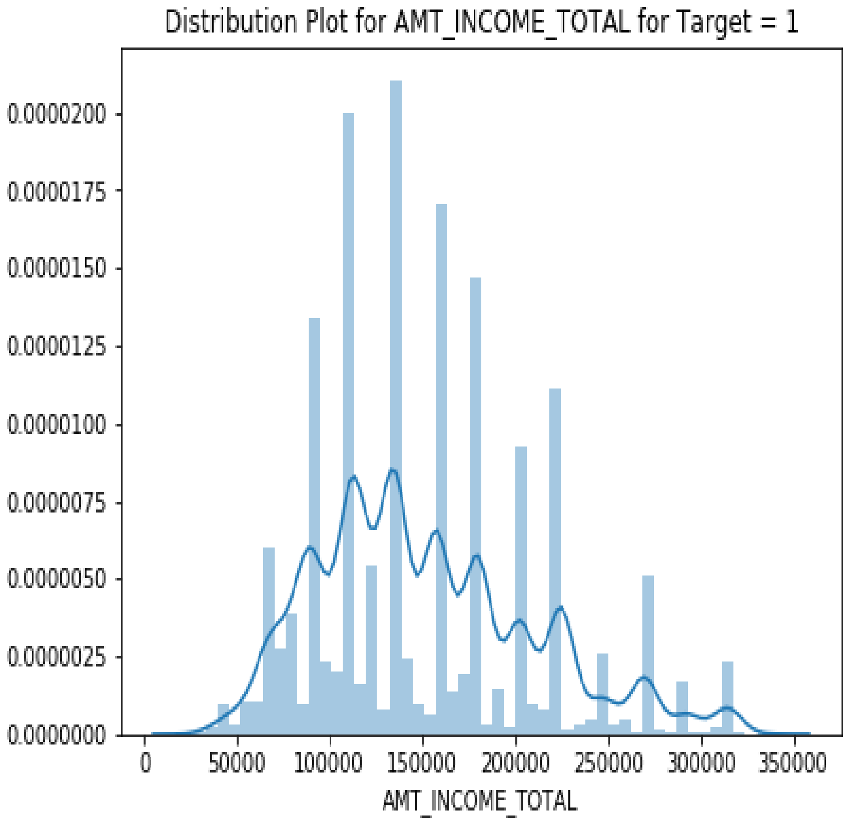
Distribution Plot for AMT_GOODS_PRICE for Target = 1

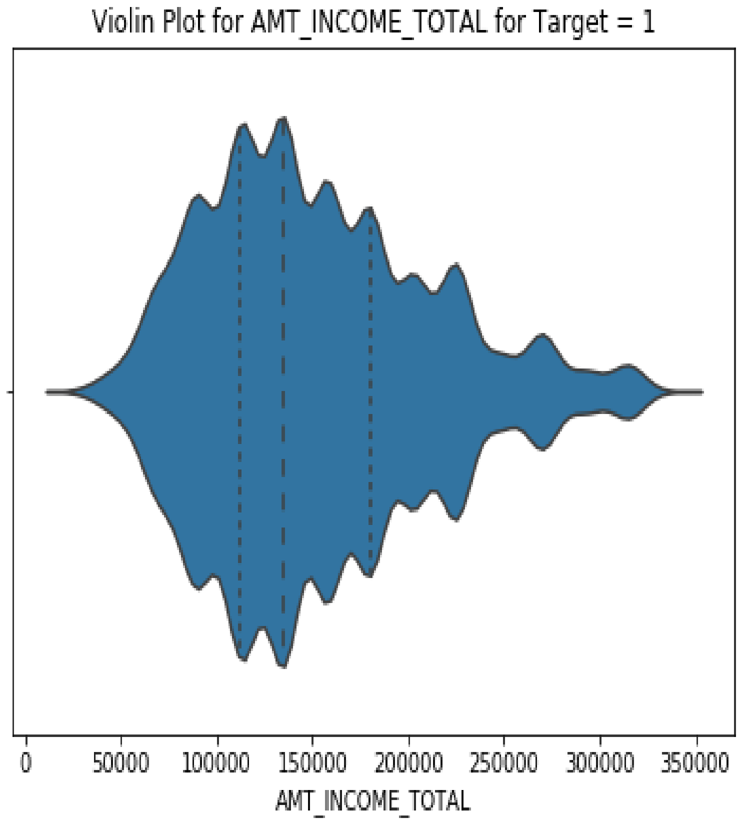


Violin Plot for AMT_GOODS_PRICE for Target = 1

Most of the Goods Prices are between ~450000 and ~630000.

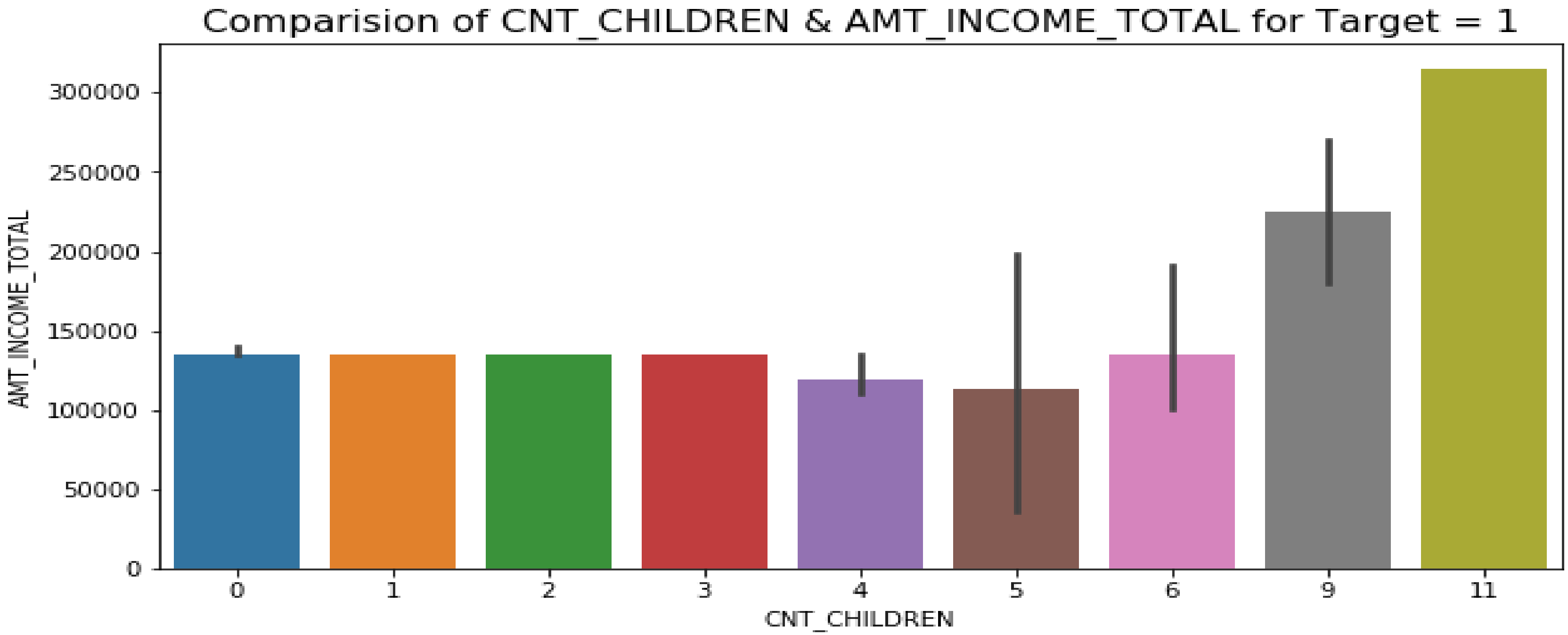
Minimum Goods Price is closed to 45000, while the maximum is around 13200000





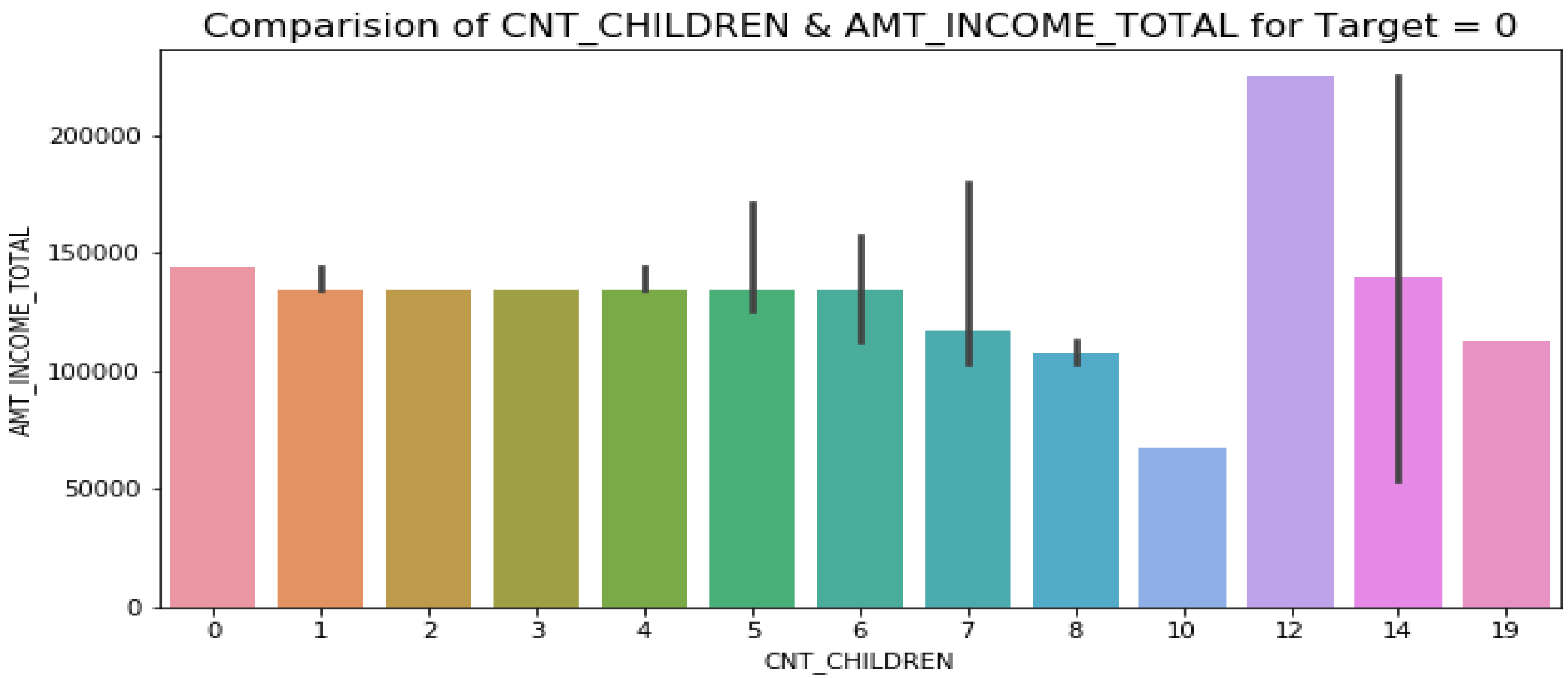
Most of the clients are earning between ~135000 and ~180000
Minimum income of the clients is around 27000, while the maximum is approximately 336000
Average income is approximately 135000

Comparison of cnt_children and income_total for target=1



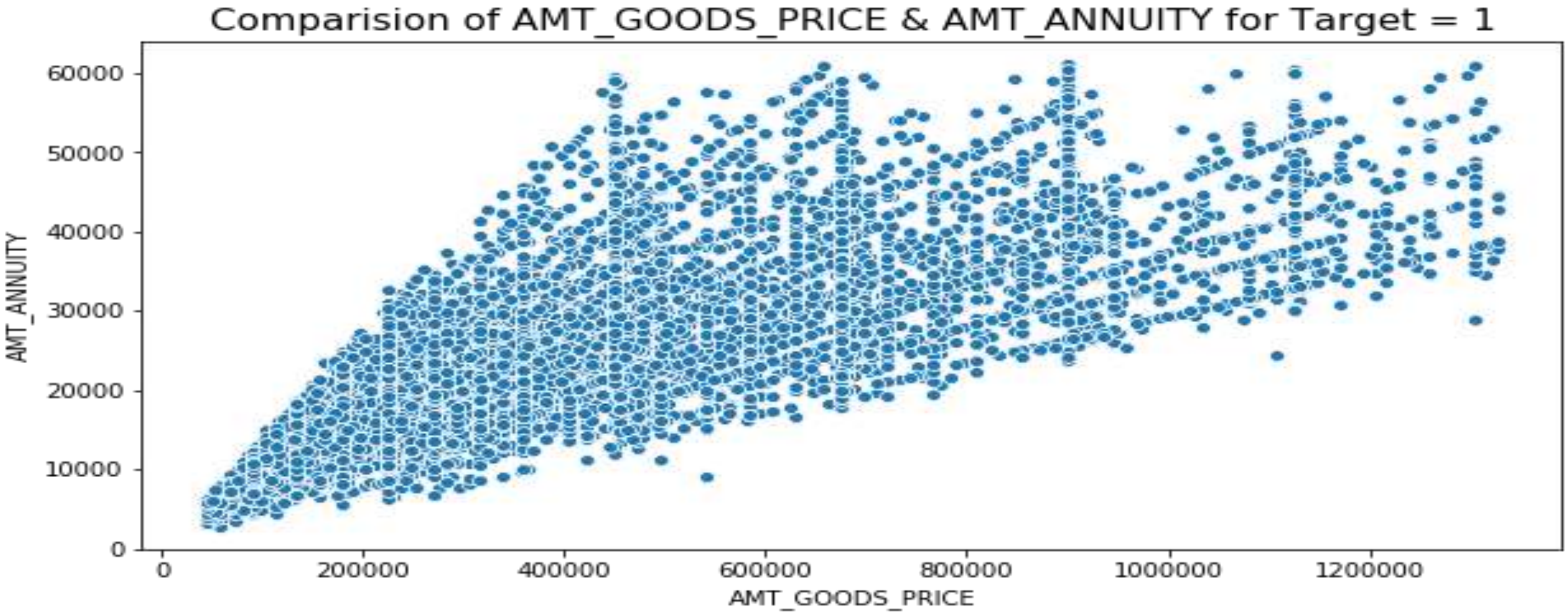
The average Income of the childless applicants is ~135000
The applicants with 5 children, have the lowest average income(112500)
The applicants with 11 children, have the highest average income(315000)

Comparison of cnt_children and income_total for target=0



The average Income of the childless applicants is ~144000
The applicants with 10 children, have the lowest average income(~67500)
The applicants with 12 children, have the highest average income(~225000)
There are applicants with 19 children, this seems unrealistic in today's world. Must be a data entry issue.

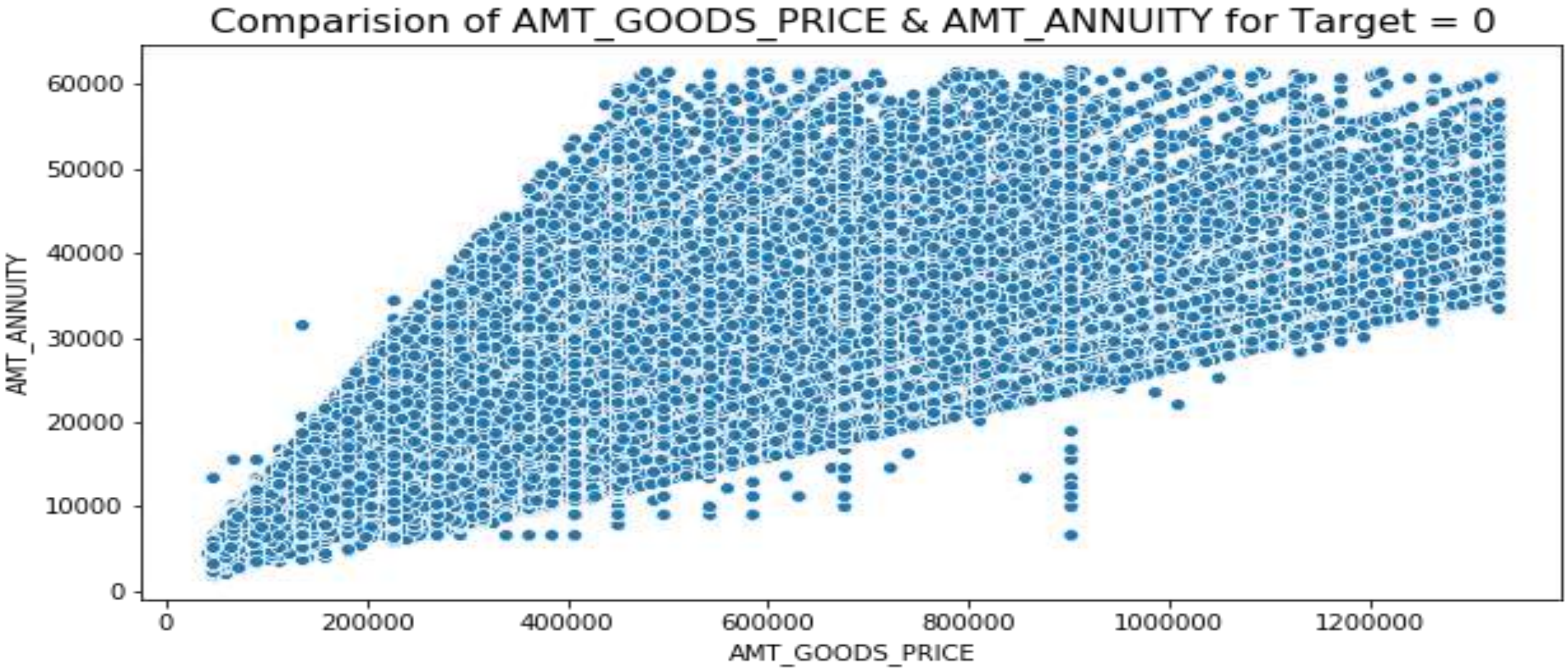
Comparison of amt_goods_price & amt_annuity for target=1



There is a high correlation between Annuity Amount and the price of the goods of the clients.

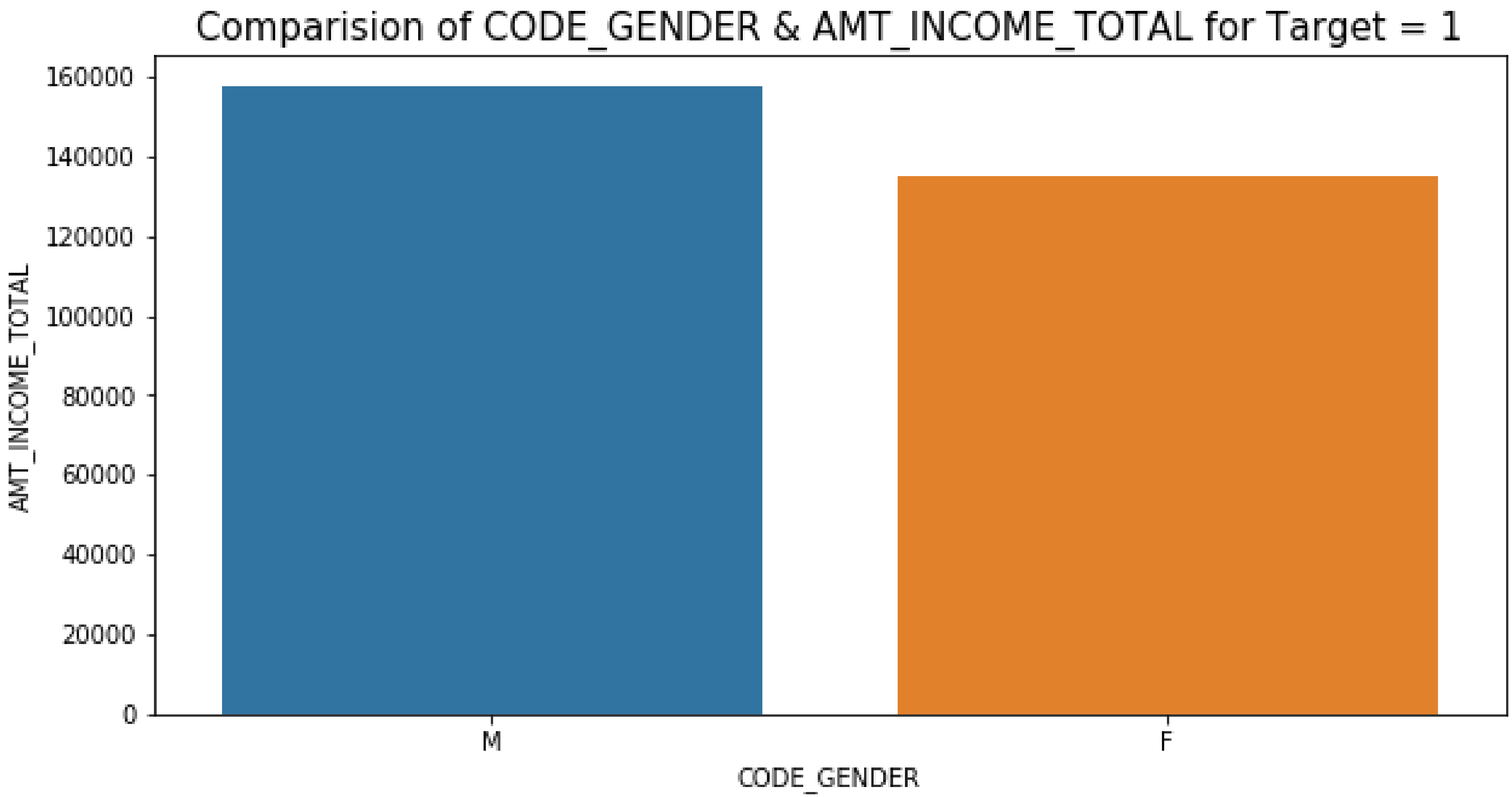
Higher the value of the Goods, higher is the Annuity amount

Comparison of amt_goods_price & amt_annuity for target=0



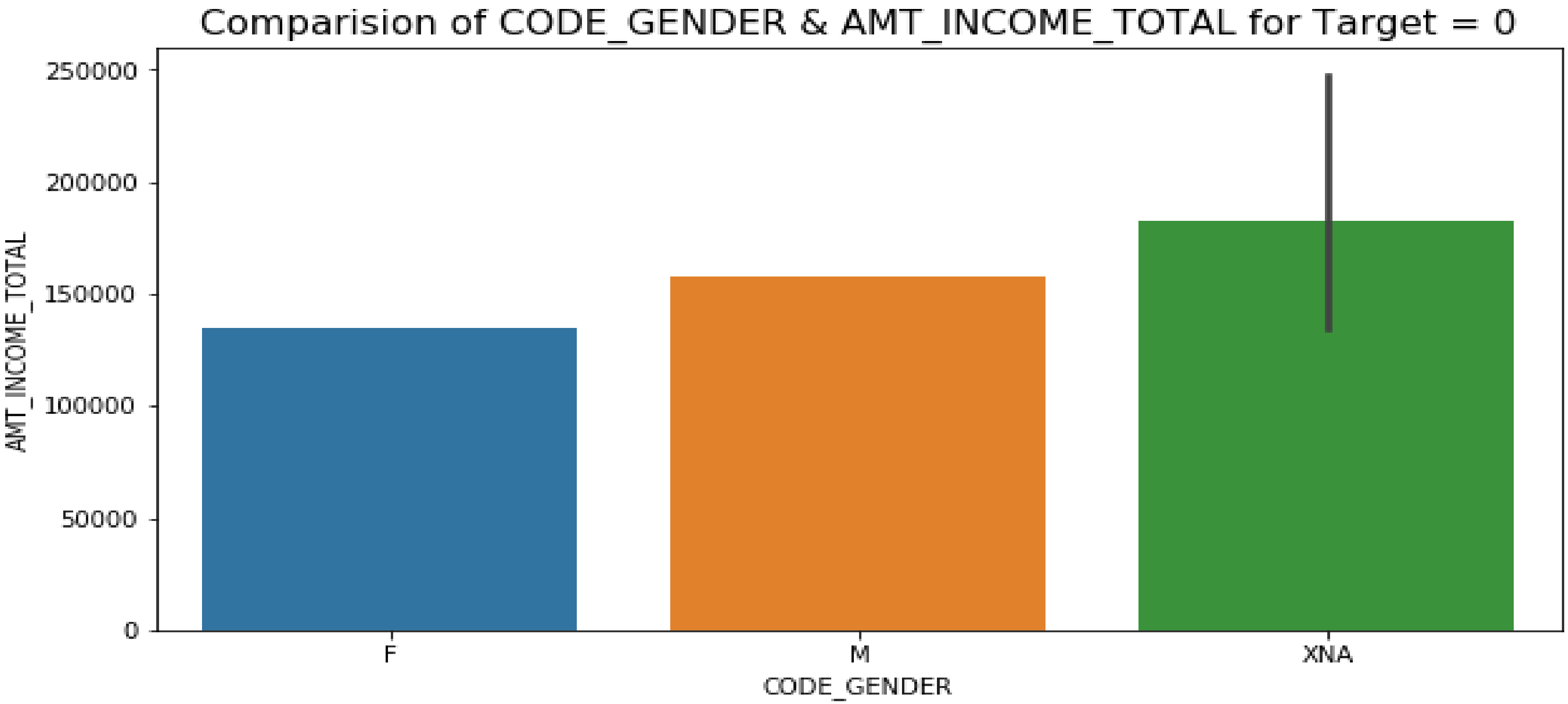
There is a high correlation between Annuity Amount and the price of the goods of the clients.
Higher the value of the Goods, higher is the Annuity amount.

Comparison of code_gender & amt_income_total for target=1



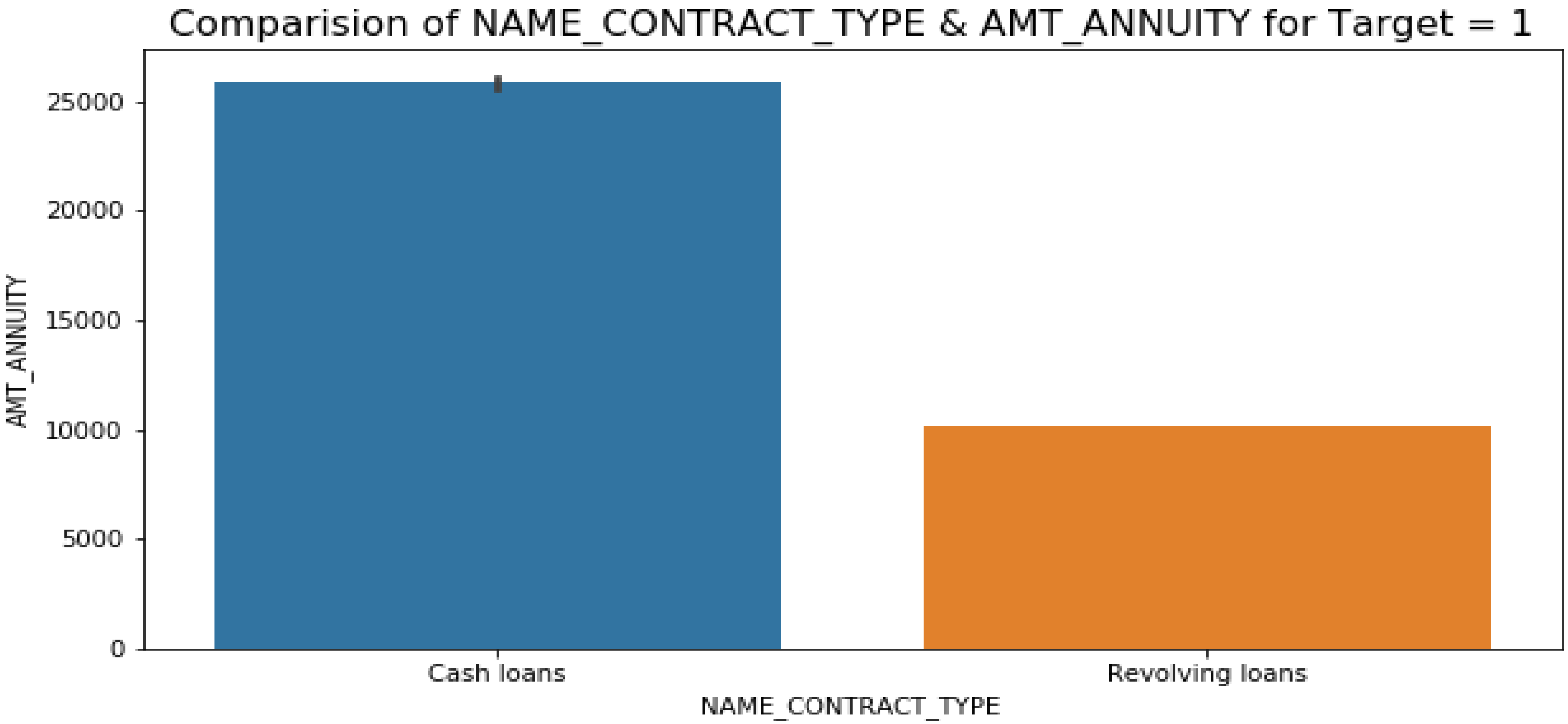
The average income of the Male applicants is ~157000, wheras for Female applicants, it is ~135000

Comparison of code_gender & amt_income_total for target=0



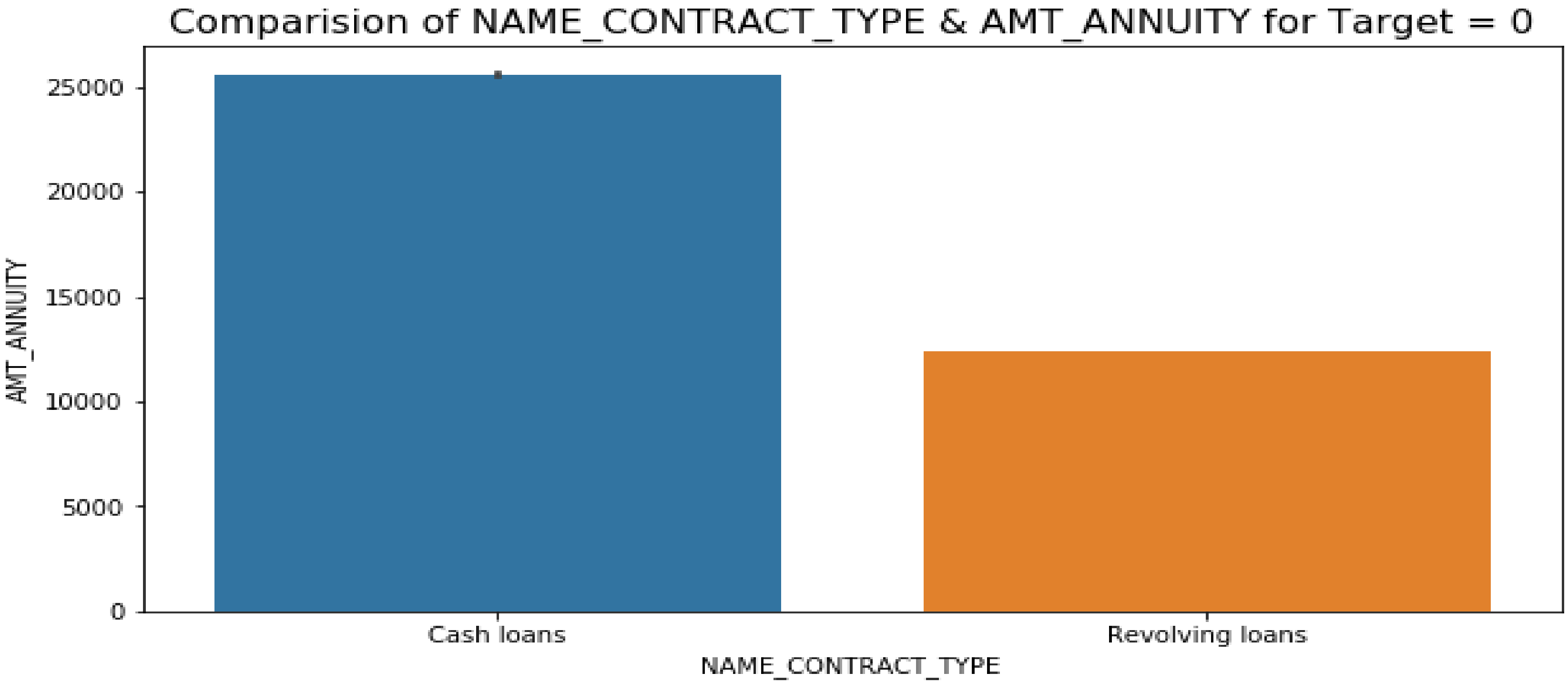
The average income of the Male applicants is ~157000, wheras for Female applicants, it is ~135000
The unknown gender has the highest average income, which is ~182000

Comparison of name_contract_type and amt_annuity for target=1



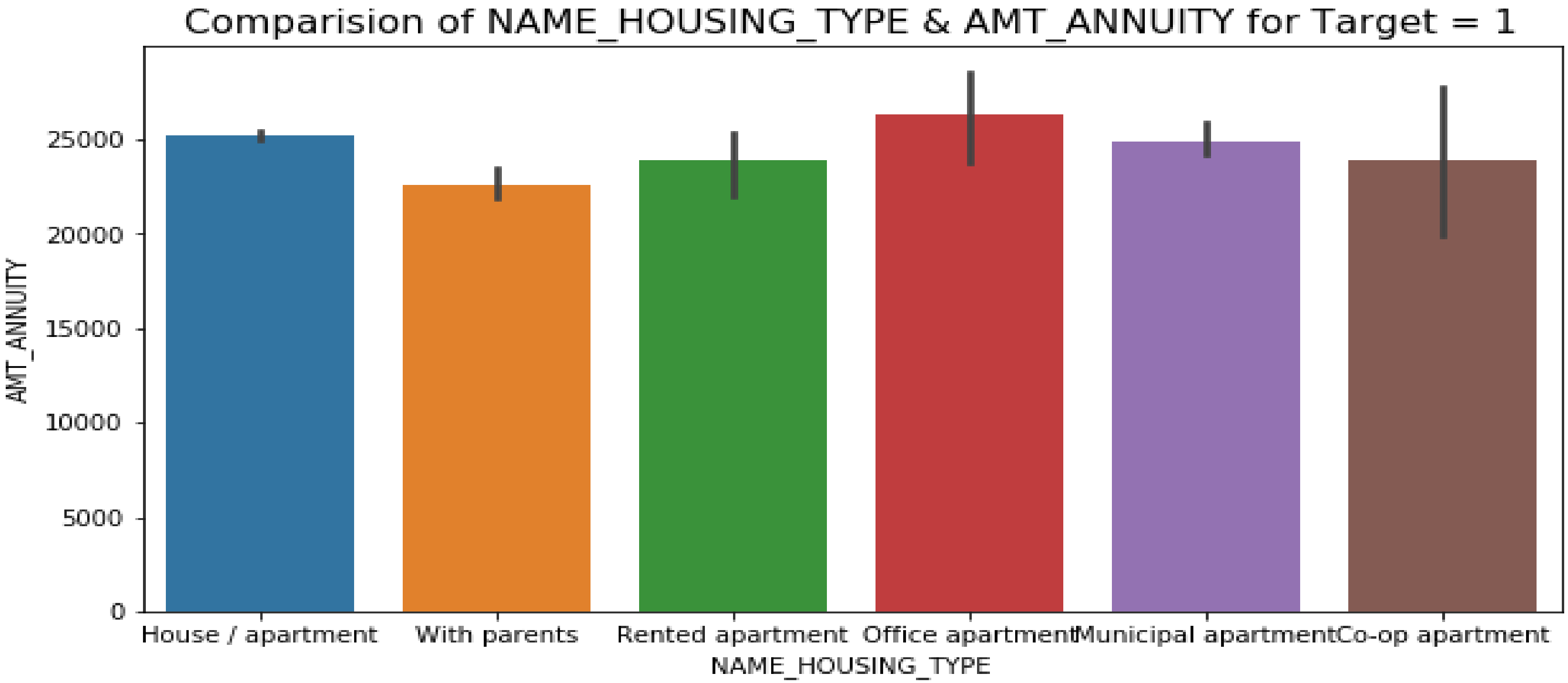
Clients who applied for Cash loans, have average Annuity Amount ~25000
Clients with Resolving loans, have average Annuity amount ~ 10000

Comparison of name_contract_type and amt_annuity for target=0



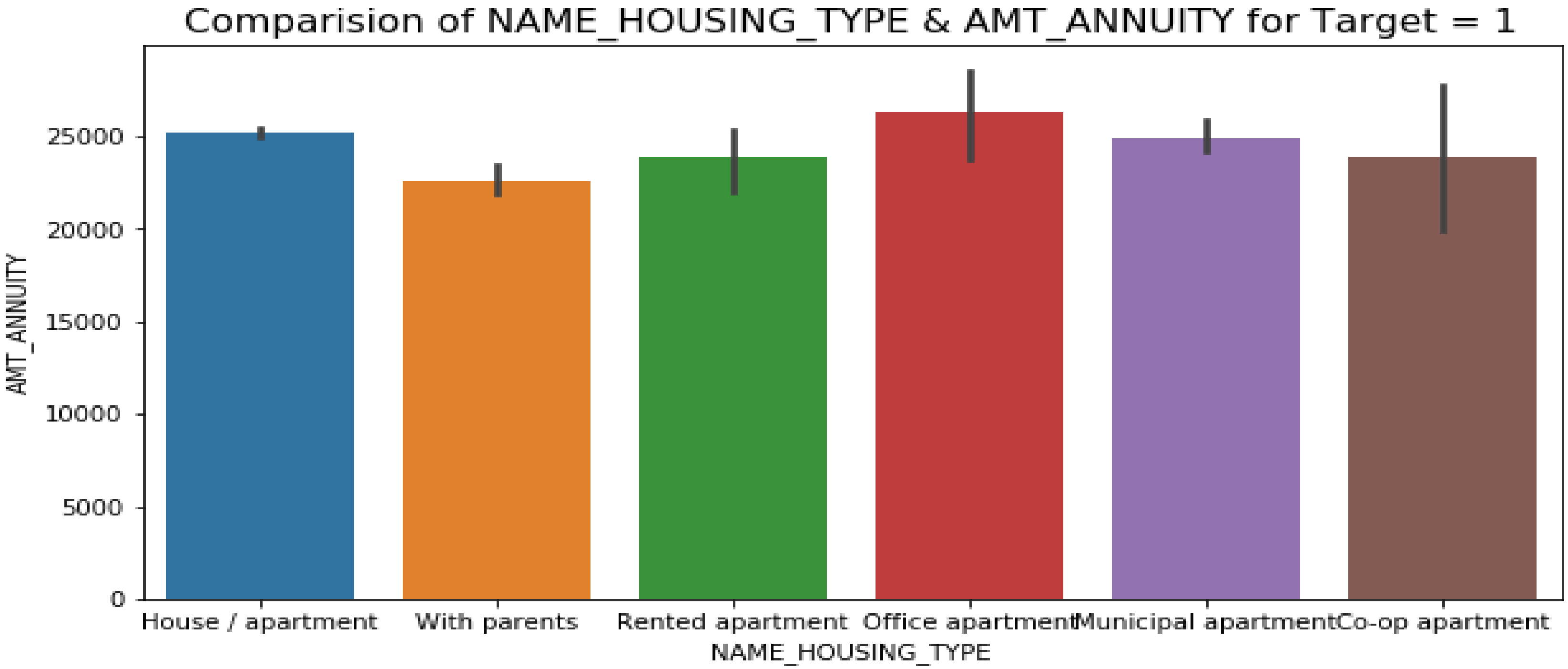
Clients who applied for Cash loans, have average Annuity Amount ~25000
Clients with Resolving loans, have average Annuity amount ~ 12000

Comparison of name_housing_type and amt_annuity for target=1



Clients living with parents, have the lowest average Annuity, and they are having difficulties in repay the loan.
Clients living in office accommodation, have the highest average Annuity, and they also have difficulties in repayment

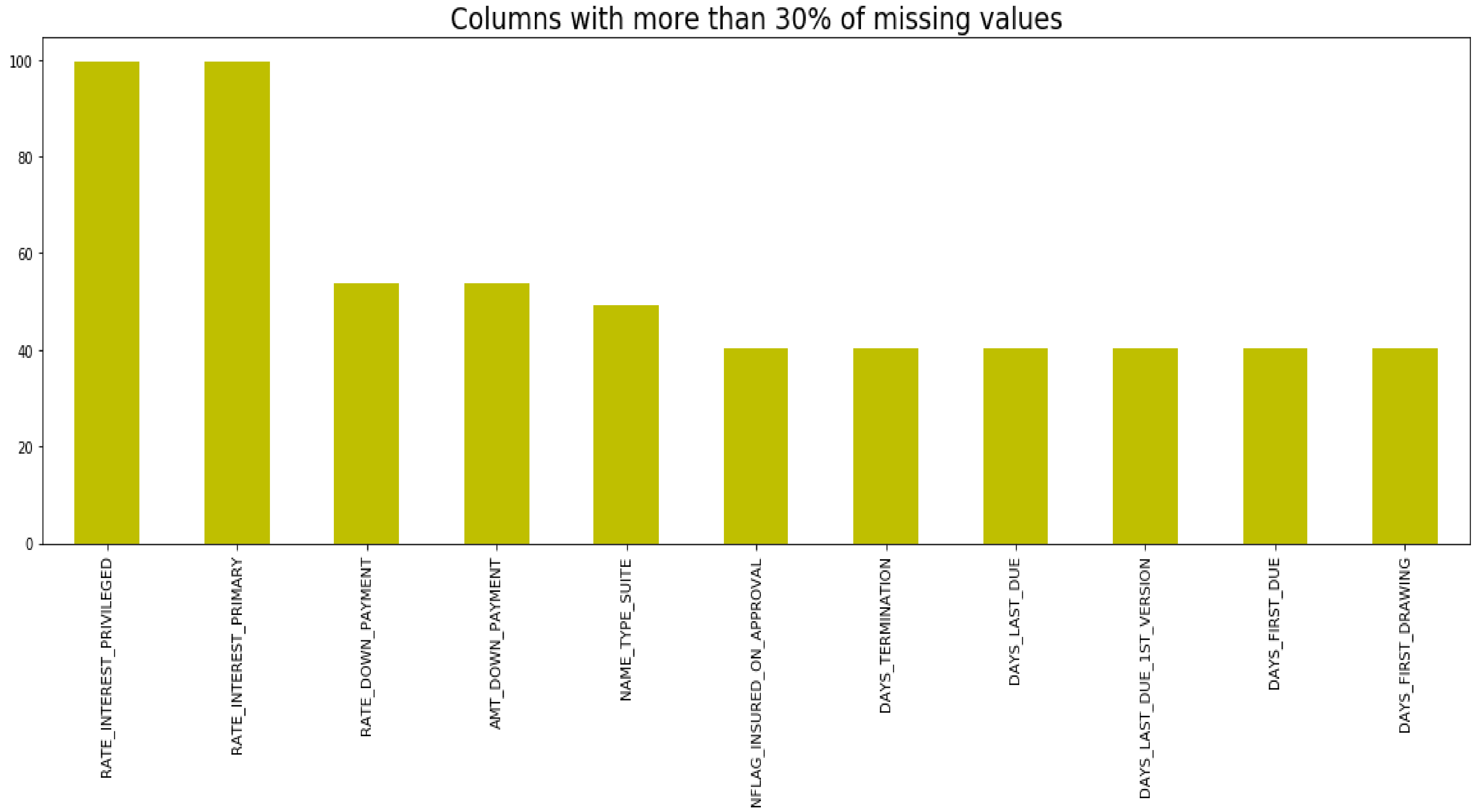
Comparison of name_housing_type and amt_annuity for target=0



Here also, Clients living with parents, have the lowest average Annuity(~22000).
Clients living in office accommodation, have the highest average Annuity(~25000

Previous Application

Identification of columns with more than 30% of missing values



Inference from previous graph

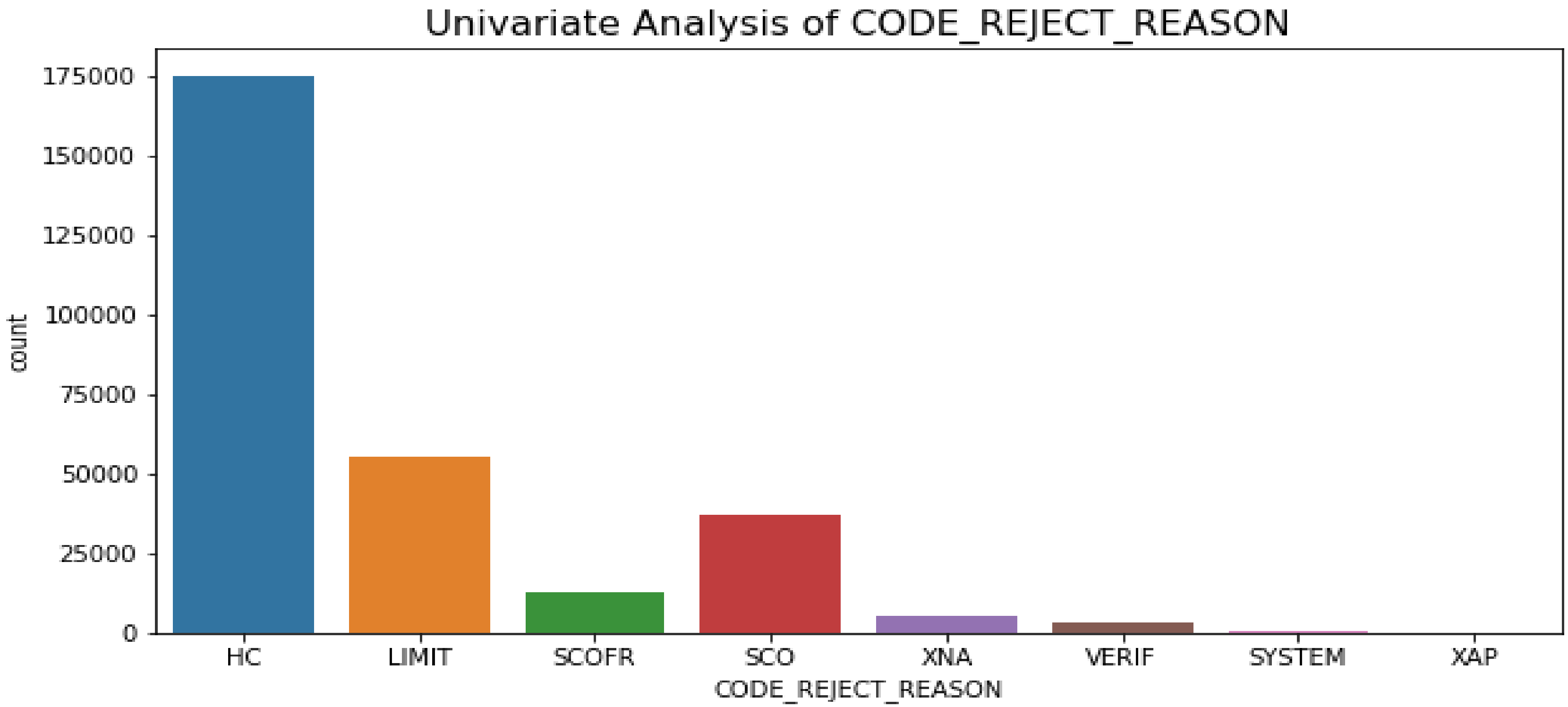
- There are 11 columns, that have more than 30% of missing values.
- Some columns have extrement large values such as RATE_INTEREST_PRIVILEGED, RATE_INTEREST_PRIMARY, RATE_DOWN_PAYMENT etc.
- With more than 30% of missing values, these columns are not recommended to use for any anaylsis, without proper treatment/imputation. Since, the no of columns to be treated are high, hence, imputing these columns with arbitrary values like mean, median, mode etc, would be risky and might produce unexpected results. It is better to drop all these columns, but, for this exercise, we only need to drop columns that are having more than 65% of missing values.

DATA ANALYSIS

- To start with, let's understand the objective of the analysis clearly and identify the variables that we want to consider for analysis.
- The objective is to identify predictors of default so that at the time of loan application, we can use those variables for approval/rejection of the loan.
- Let's have a look at the target variable - NAME_CONTRACT_STATUS. We need to relabel the values to a binary form - 0 or 1, 1 = Late payment/defaulter and 0 otherwise

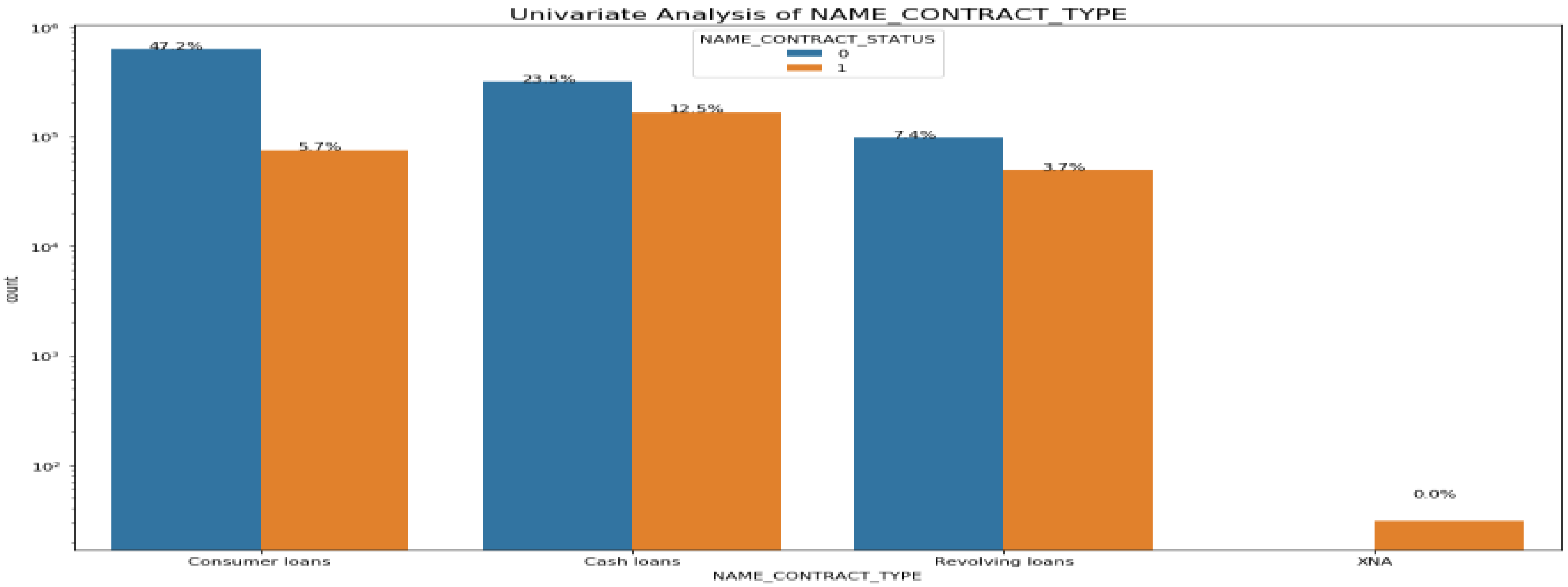
UNIVARIATE ANALYSIS

- Univariate analysis of code_reject_reason.



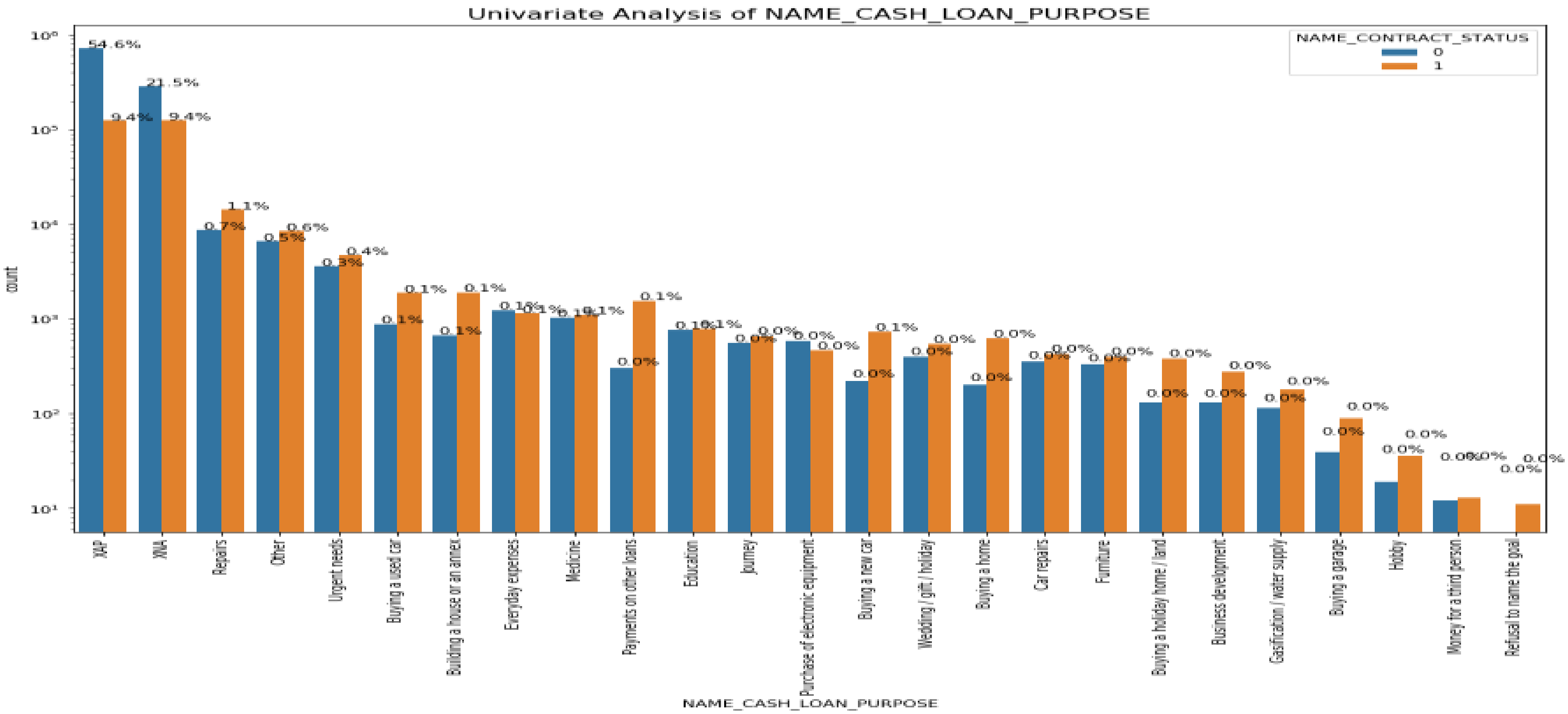
60% applications got rejected due to the Reasone Code "HC"

Univariate analysis of name_contrast_type



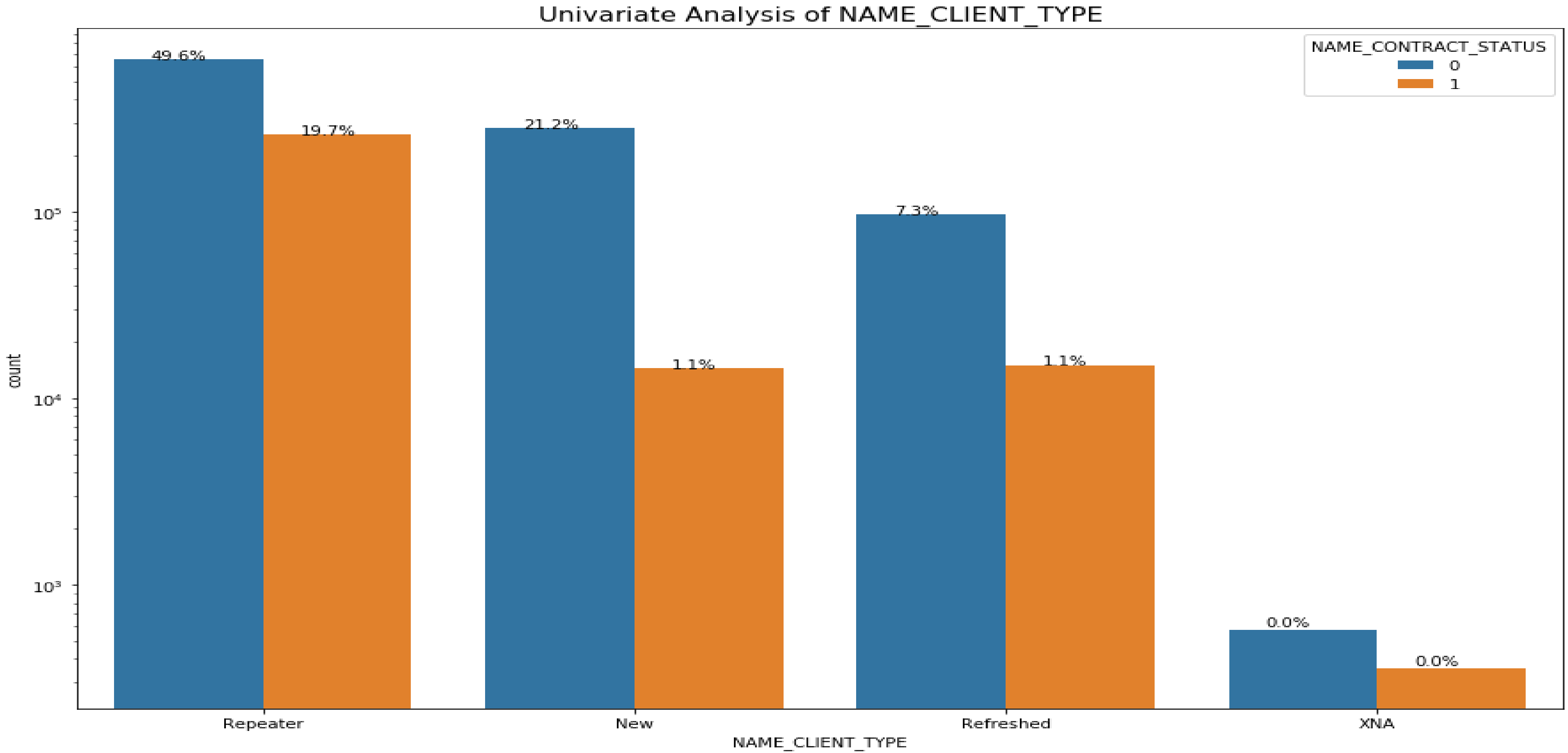
Approximately 600000(47.2%) applications got Approved, while ~80000(5.7%) got rejected for Consumer loans
Approximately 300000(23.5%) applications got Approved, while ~170000(12.5%) got rejected for Cash loans
Approximately 70000(7.4%) applications got Approved, while ~40000(3.7%) got rejected for Revolving loans

Univariate analysis of name_cash_loan_purpose



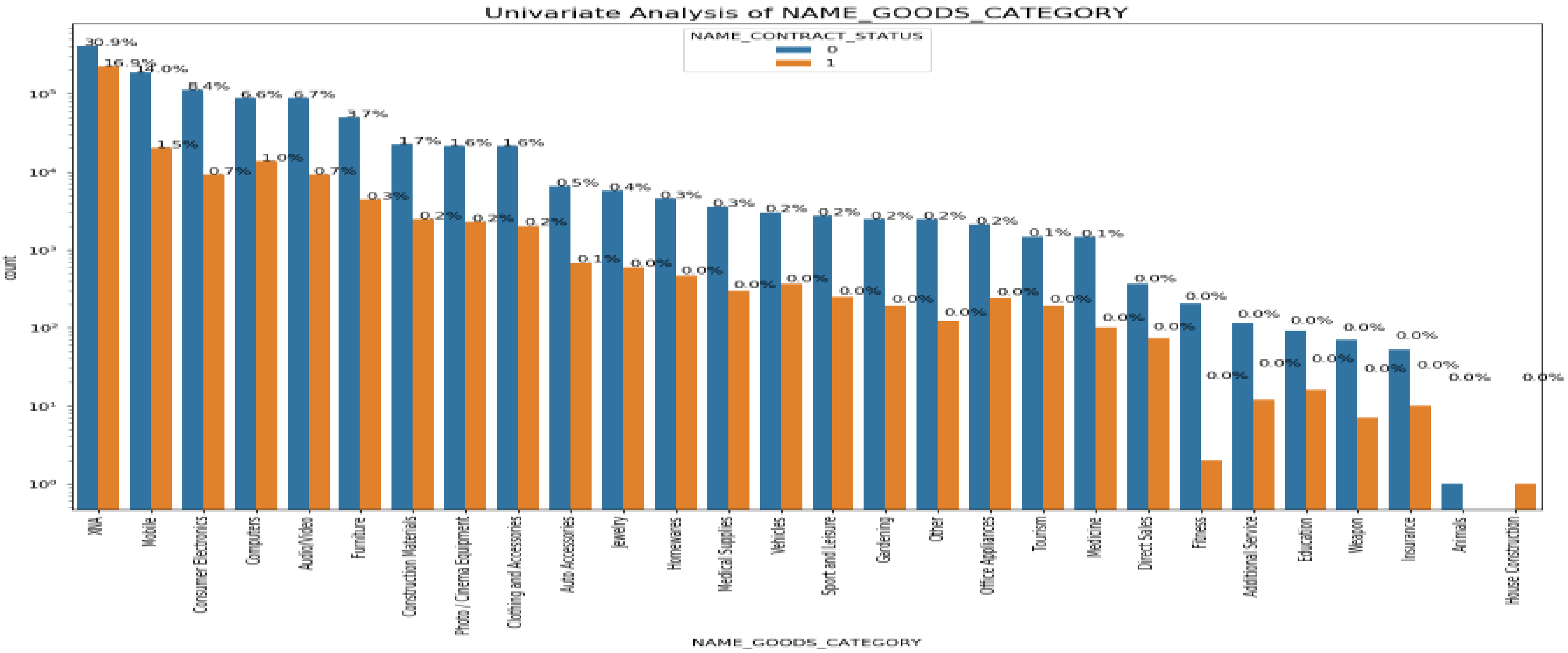
There are loan purposes called XAP, XNA with high default rates or rejected applications. The business meaning of these types are not known at this point in time.
All other loan purposes are neglible.

Univariate analysis of name_client_type



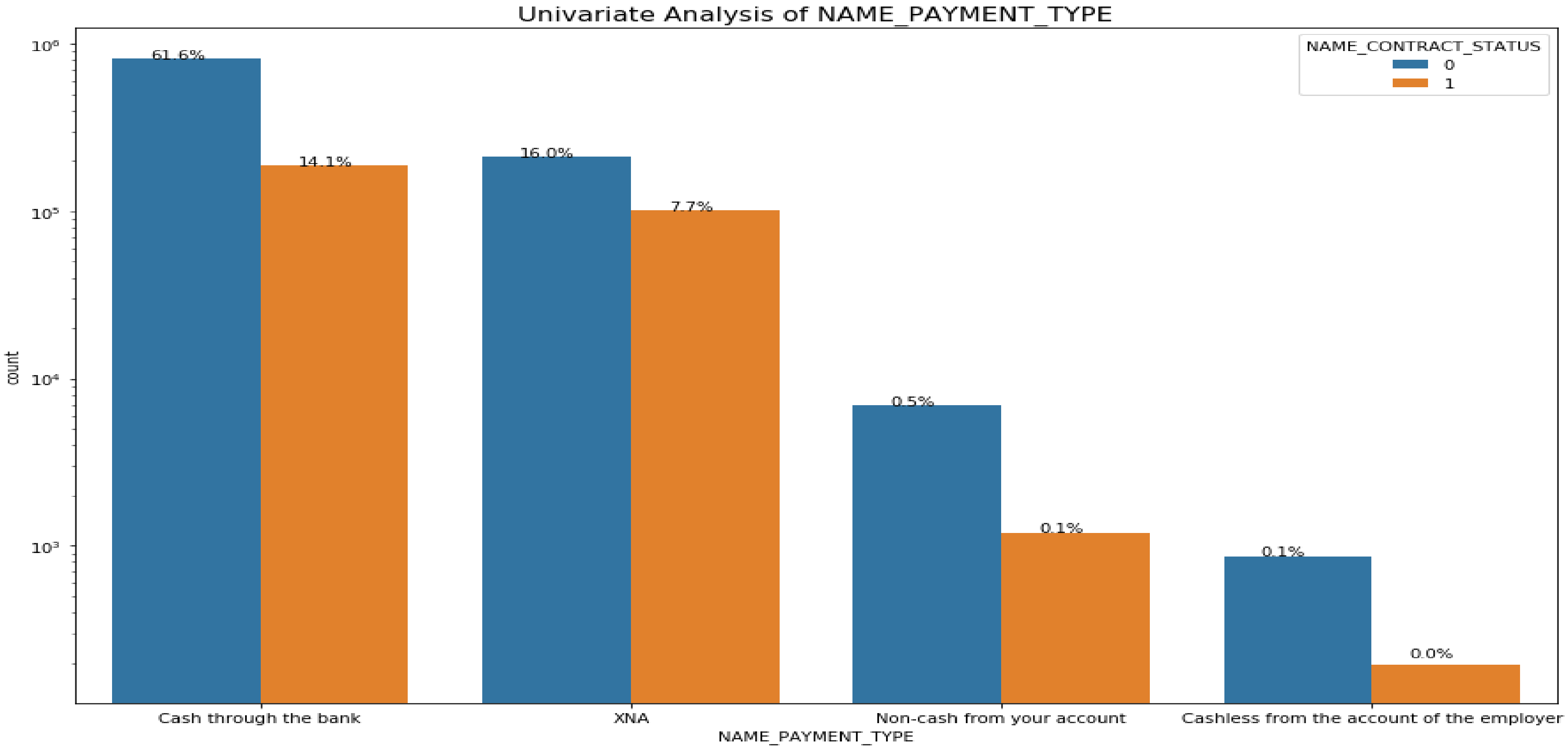
49.6% applications are rejected for the clients, who had previously applied or are repeater.
21.2% new applicants have been rejected.
7.3% of refreshed got rejected.

Univariate analysis of name_goods_category



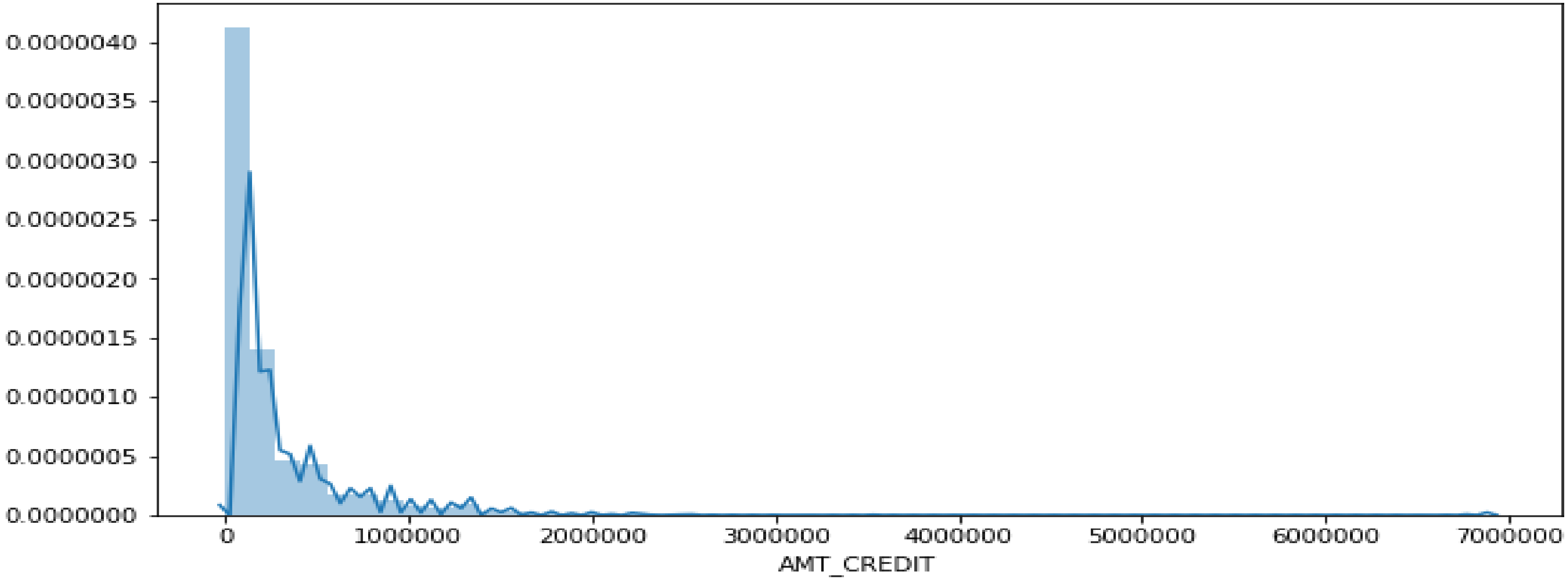
30.9% applications for unknown reasons rejected
14% applicants, who applied for loan to buy a mobile rejected

Univariate analysis of name_payment_type



61.6% applications were rejected, where the applicants chose to repay the loan via "Cash through the bank" payment method.
"Cashless from the employer account" had the lowest rejection rate(0.1%)

Univariate Analysis of Continous/Numeric Variables



Now, bin the variable into the following 4 discrete categories. This will help us in analyzing - Loan Amount varies across other variables such as Target(Default Rate), Income of the applicant etc

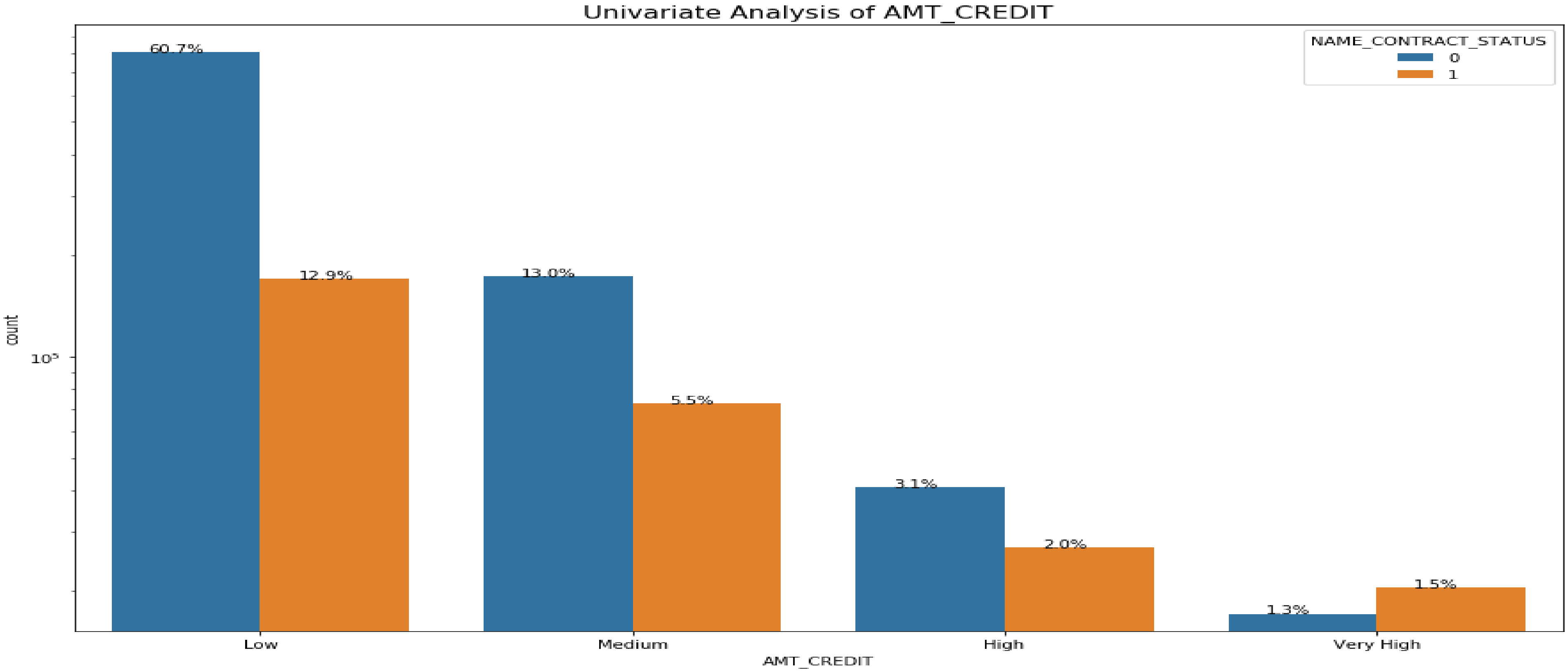
Low

Medium

High

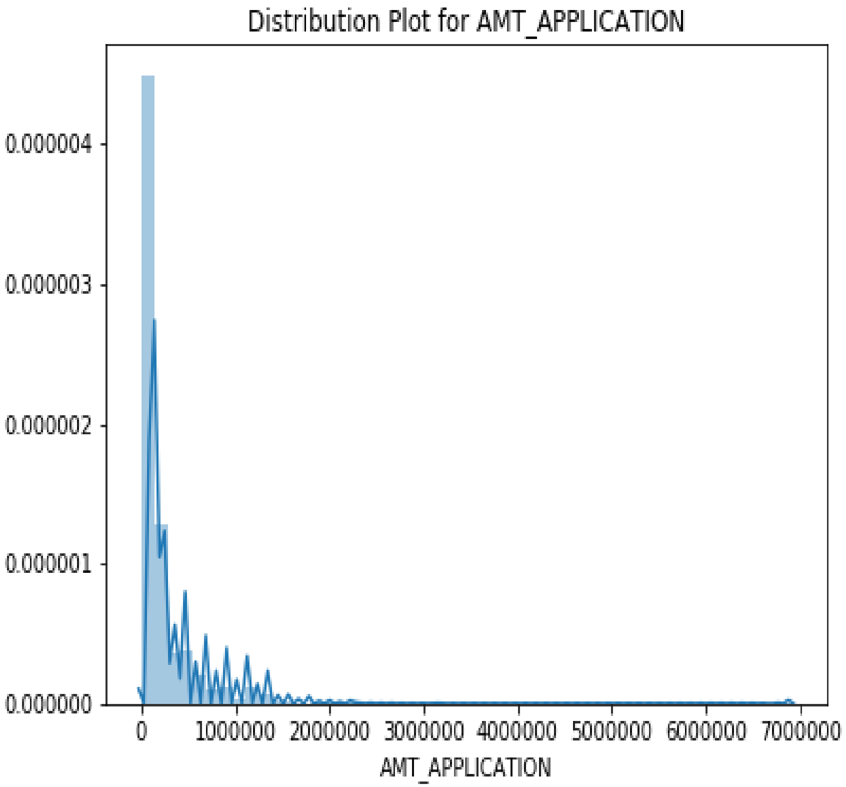
Very High

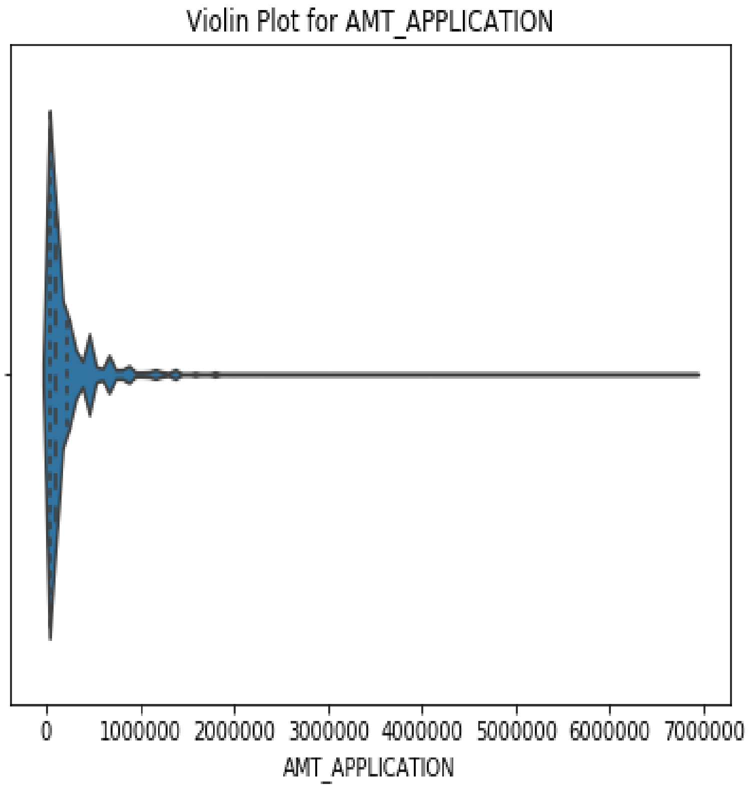
Univariate analysis of amt_credit



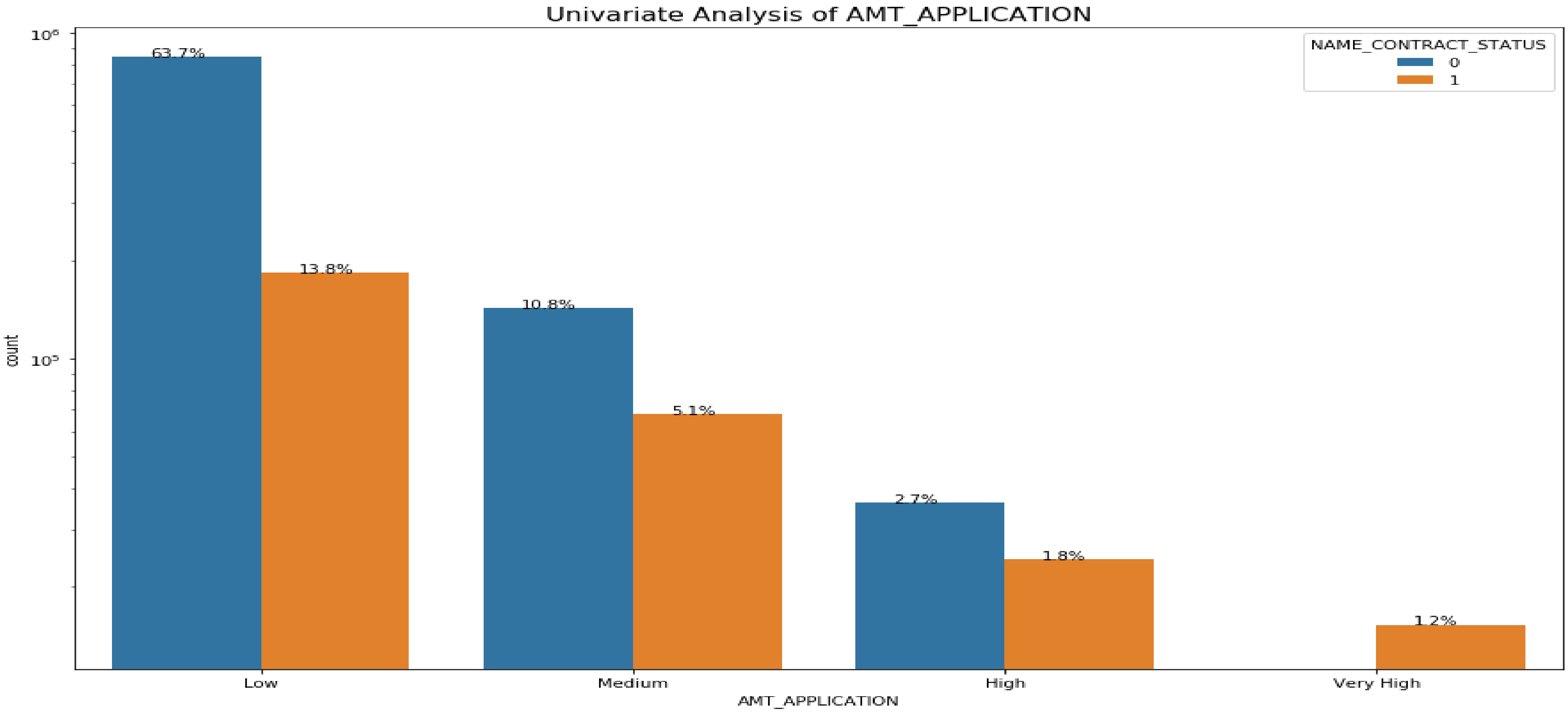
60.7% of applications for low amount loan rejected, while 12.9% were accepted.
1.3% applications for very high low amount rejected

Most of the requested loans are distributed between 5000000 and 8000000


Distribution Plot for AMT_APPLICATION

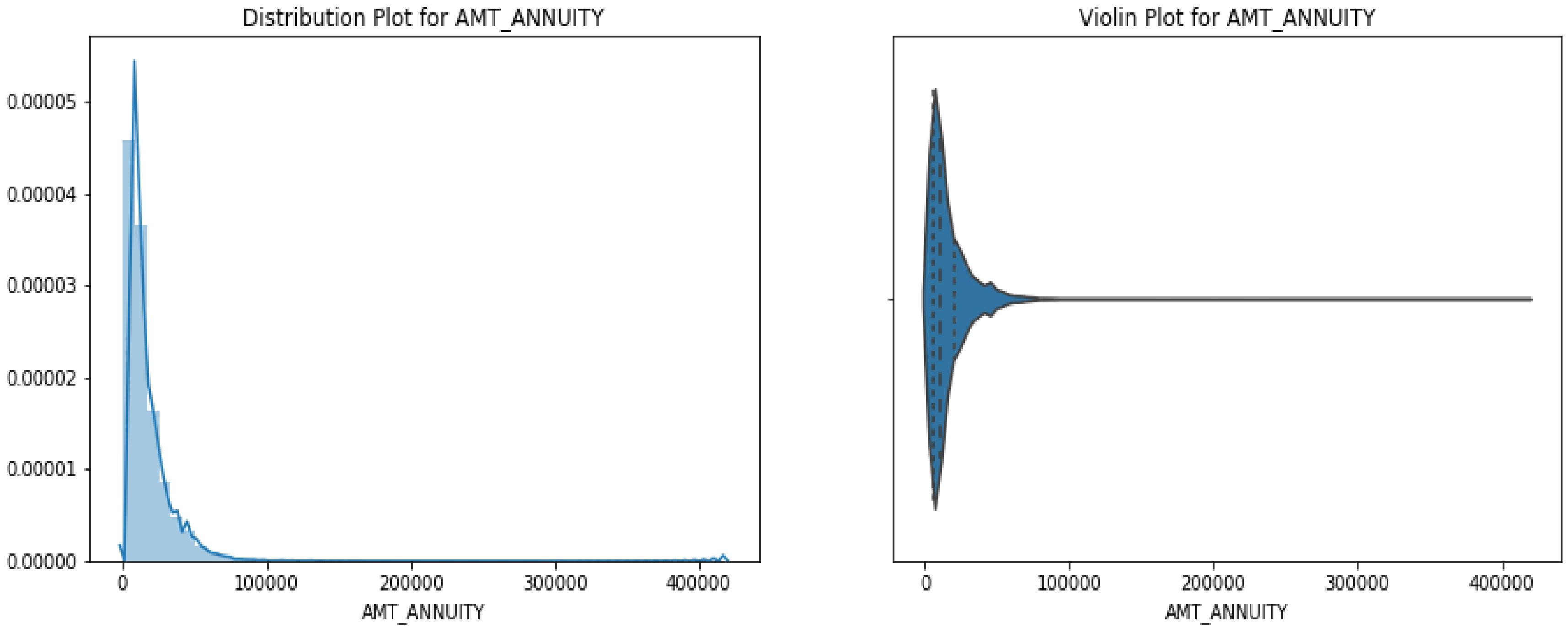

Violin Plot for AMT_APPLICATION

Univariate analysis of Amt_application



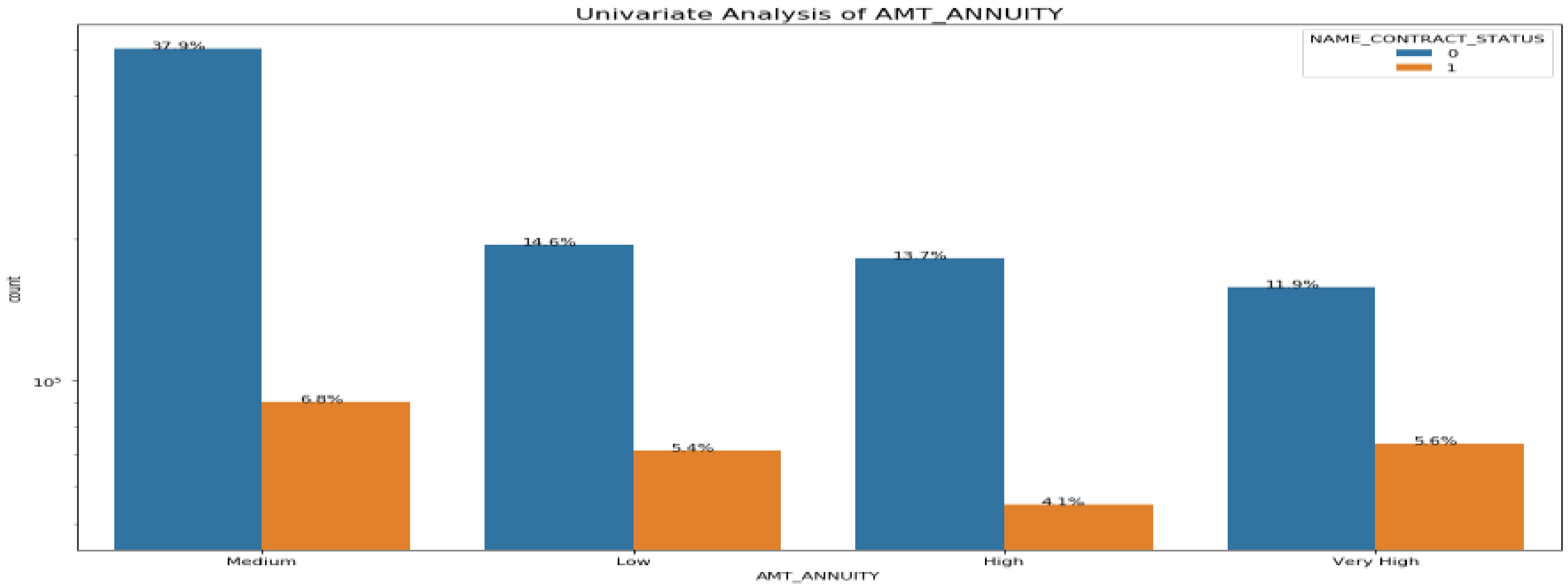
63.7% of the applications for low amount loan rejected.
None of the Very high loan applications was rejected.
All the applications for the requested very high loan were accepted

Distribution plot for amt_annuity



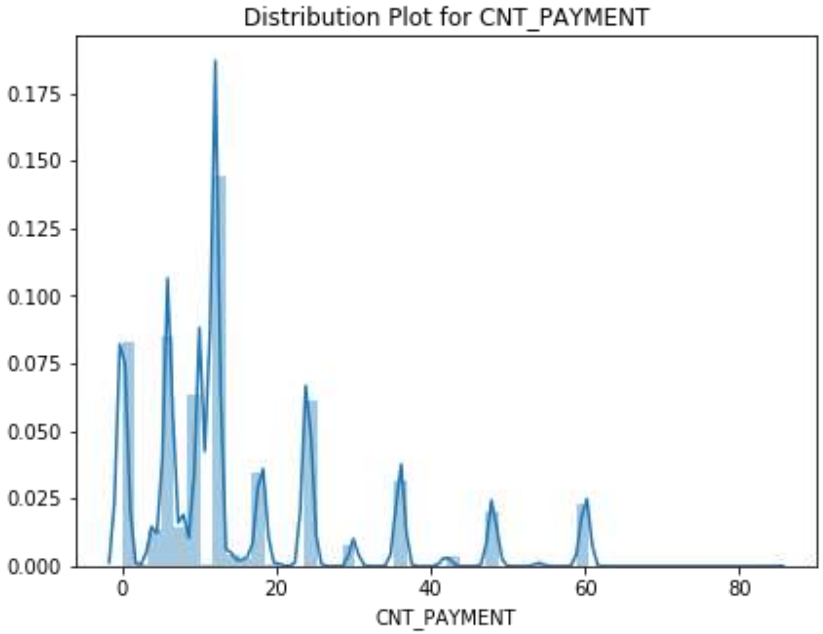
Most of the Annuity Amounts are distributed between 10000 and 20000

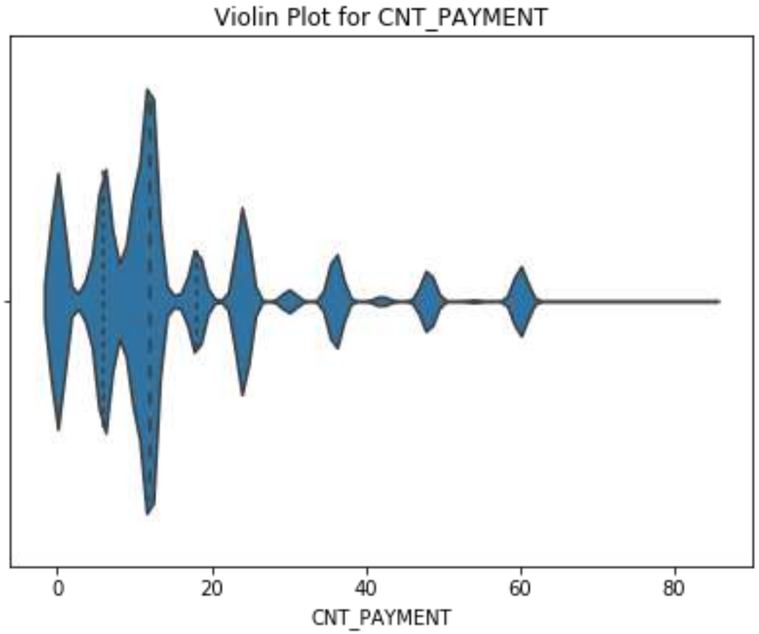
Univariate analysis of amt_annuity



37.9% applications(between 5000 and 15000) with medium Annuity amount were rejected.
There was only 11.9% of refusal in the very high category(>25000)

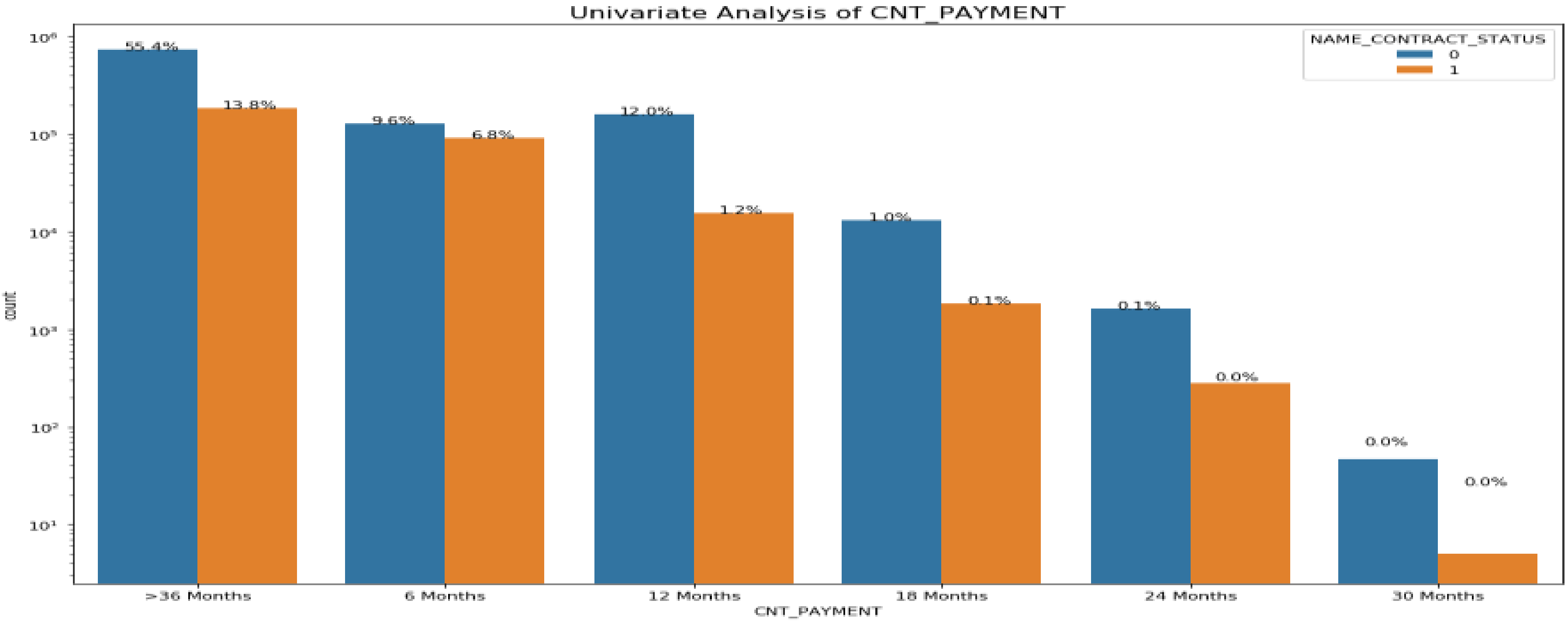
Distribution plot for cnt_payment





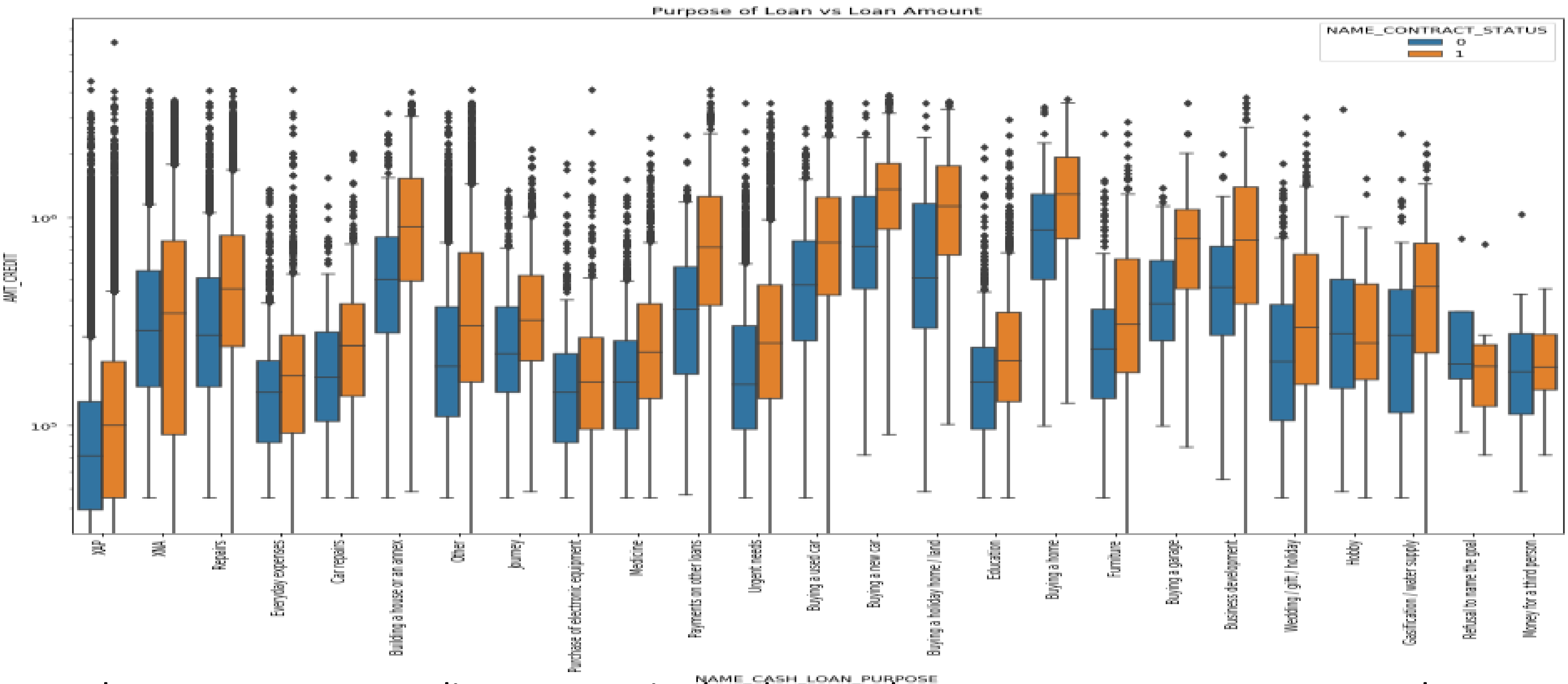
Most of the loan terms are distributed between 12 and 18

Univariate analysis on cnt_payment



Loan term of more than 36 months had the highest rejection rate(55.4%). It also had the highest acceptance rate(13.8%) among other

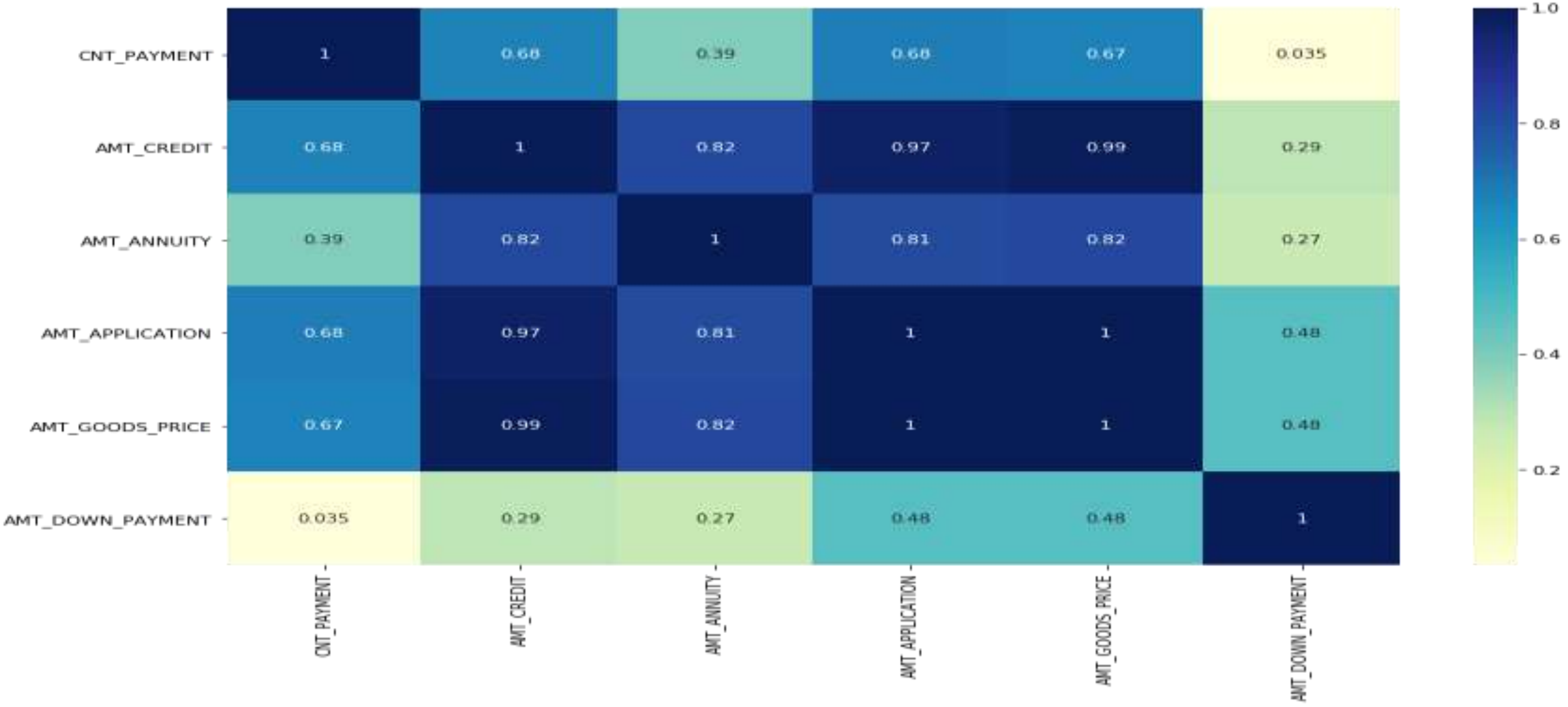
Bivariate/Multivariate Analysis



There are so many outliers present in the dataset, but we are not suppose to treat them as part of this exercise.

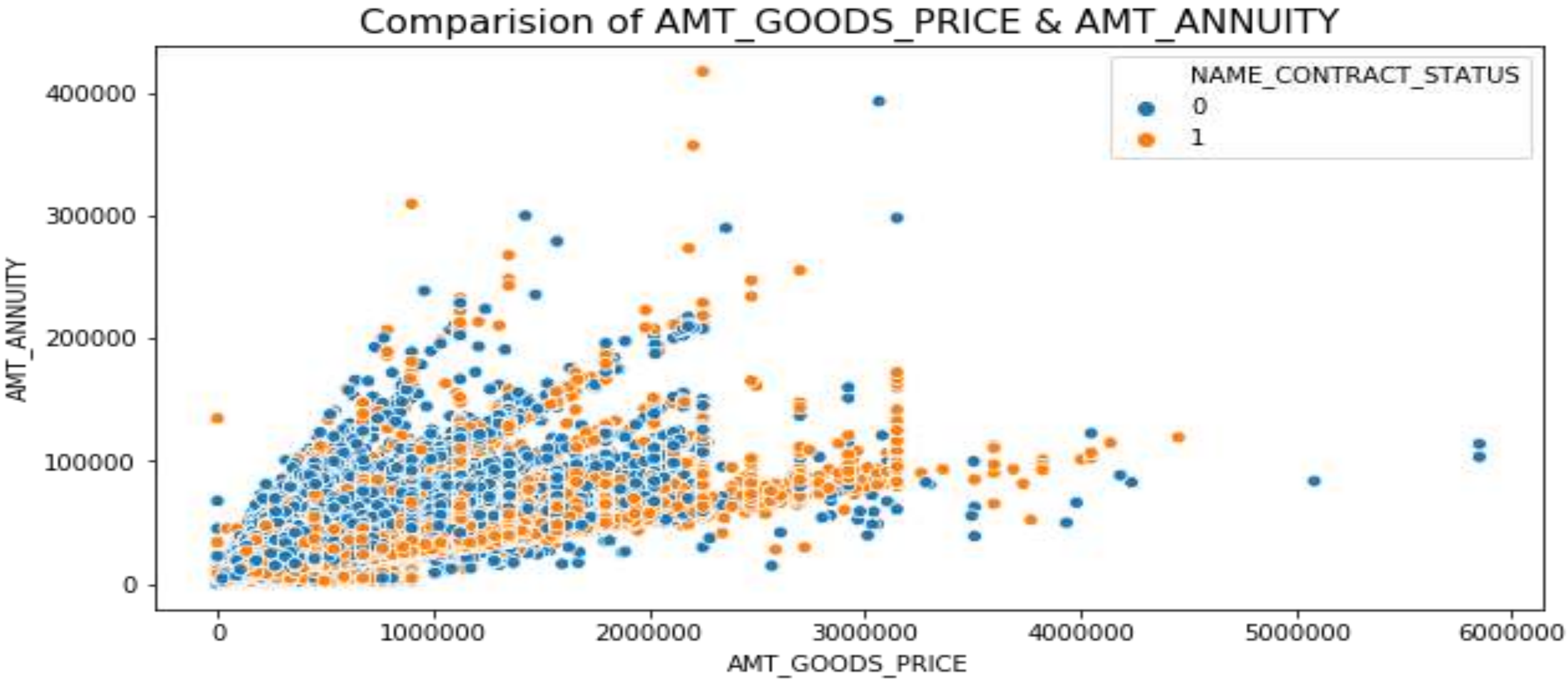
Buying a new car, buying a holiday home/land, Building a new house etc have high loam amounts and refusal rates

Checking co relation between all continuous variable



It is clear from the heatmap that the variables, AMT_GOODS_PRICE, AMT_APPLICATION, AMT_CREDIT, AMT_ANNUITY are very closely interrelated. Therefore, any of these column can be taken for analysis

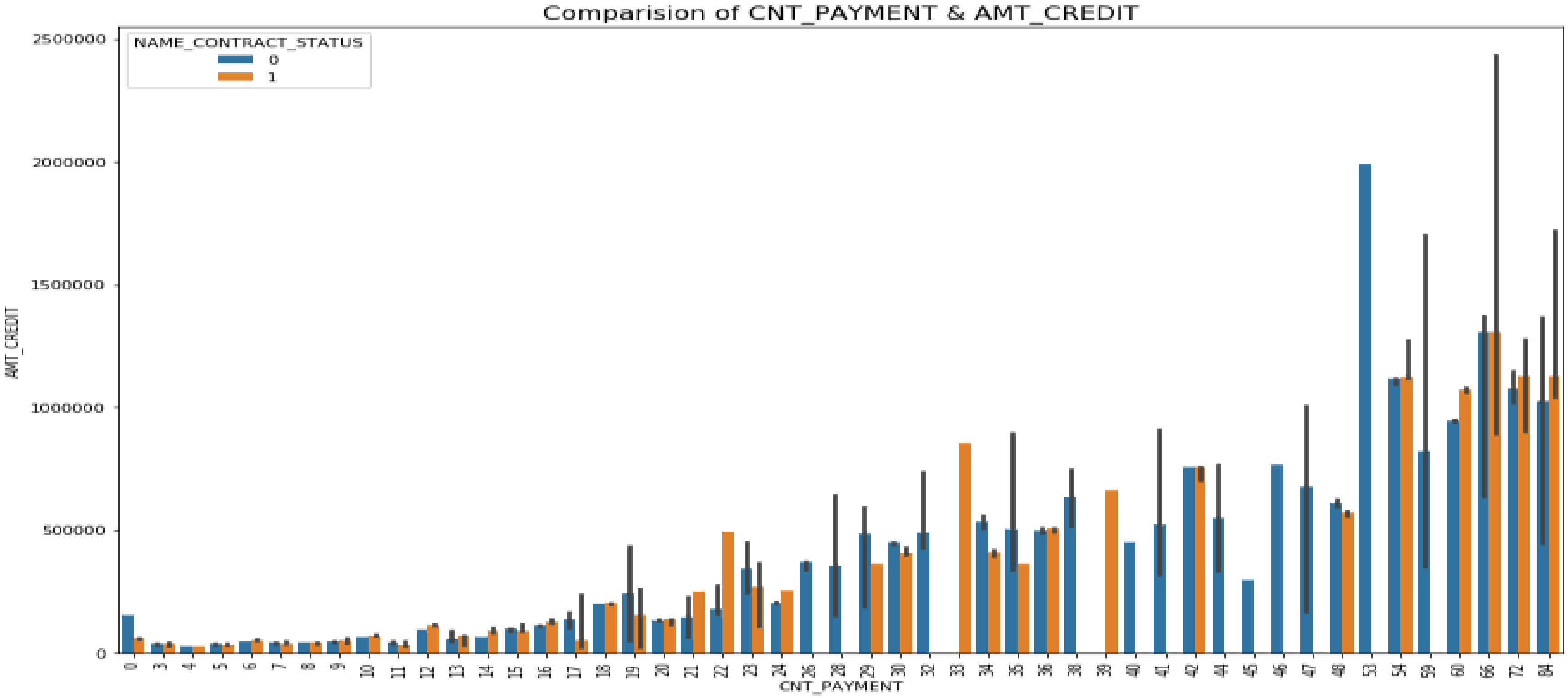
Comparison of amt_goods_price and amt_annuity



There is a high correlation between Annuity Amount and the price of the goods of the clients.

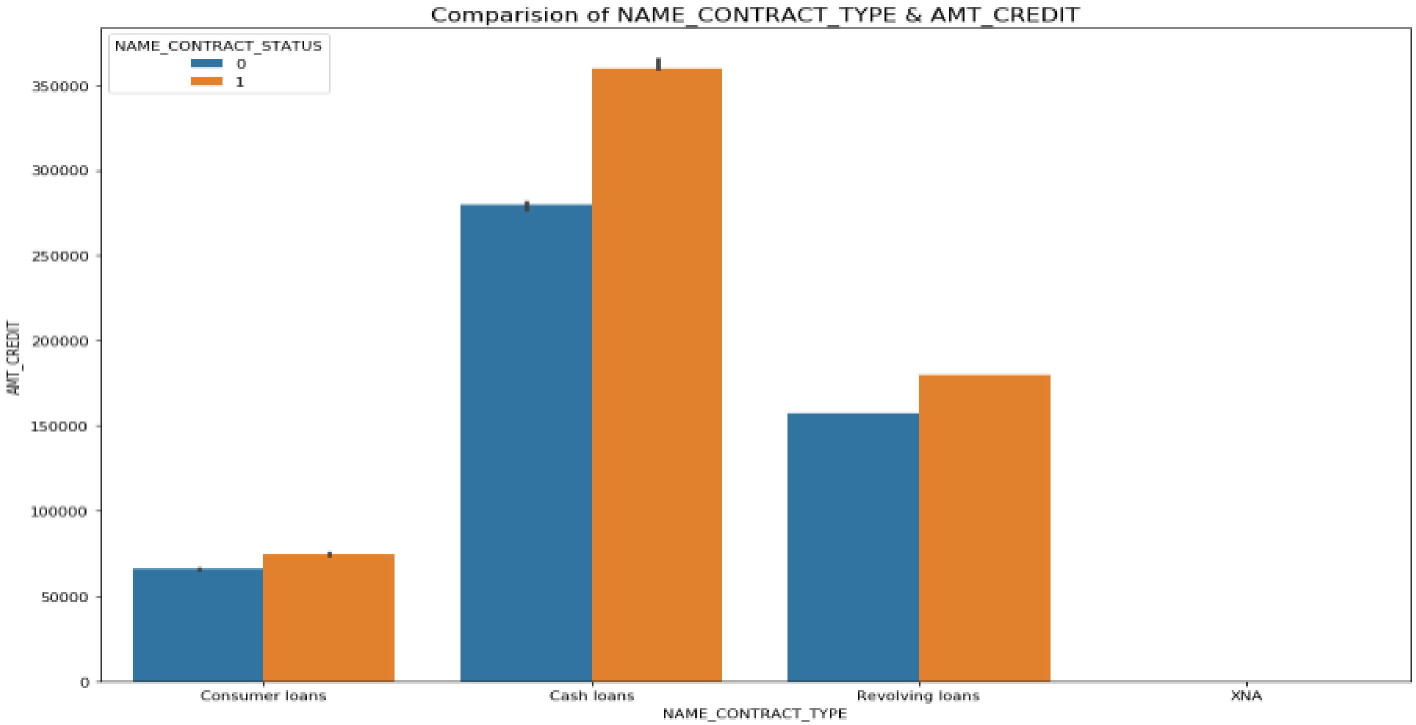
Higher the value of the Goods, higher is the Annuity amount

Comparison of cnt_payment & amt_credit



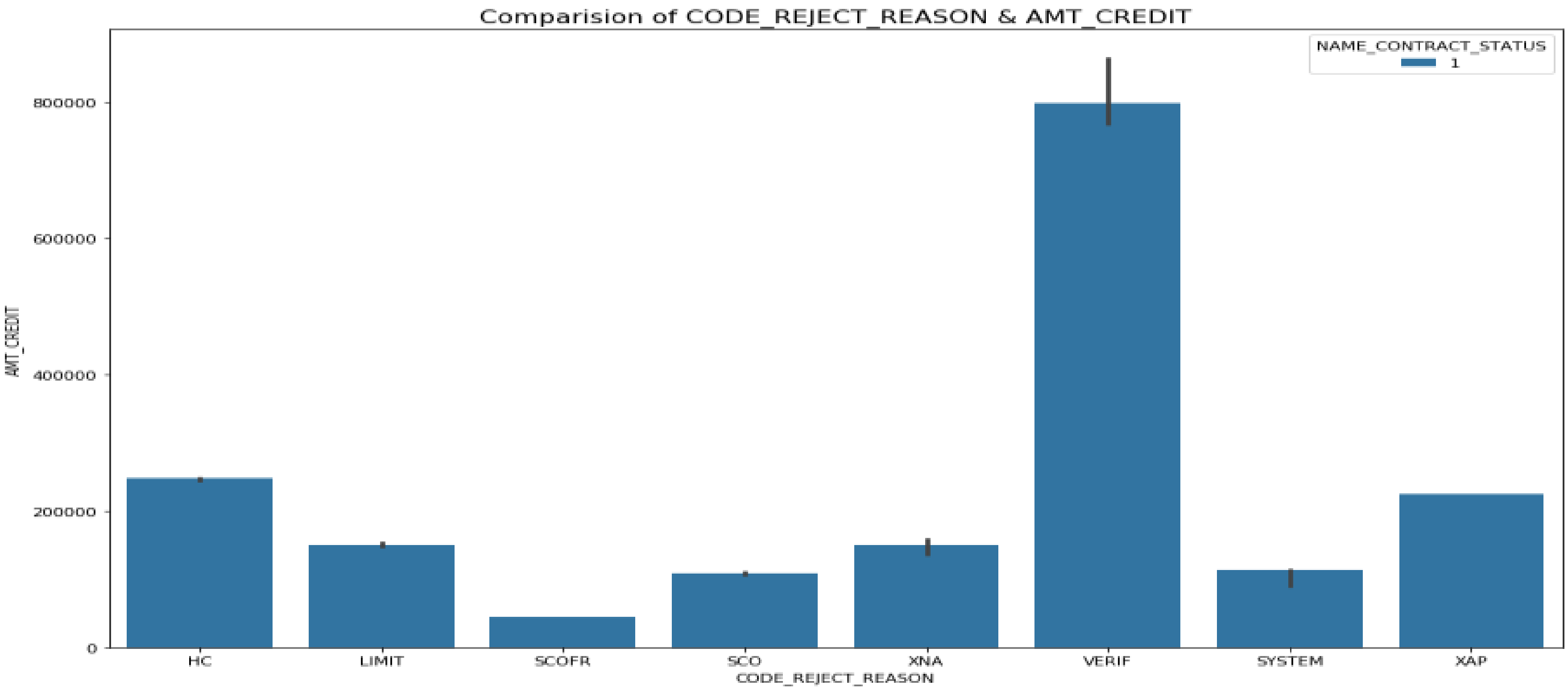
Higher the Loan amount, higher is the loan term or no of payments

Comparison of name_contract_type and amt_credit



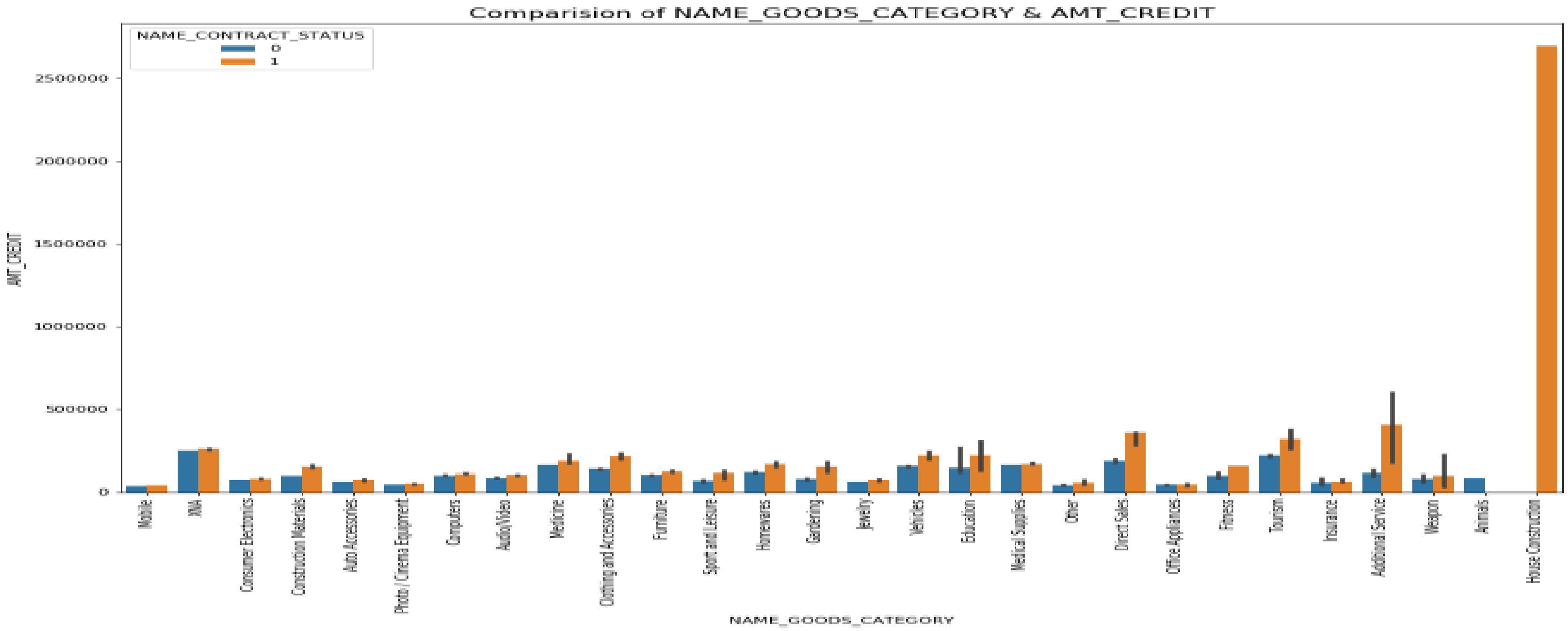
The average cash loan amount, that got rejected is ~340000
The average consumer loan, that got rejects is ~70000
Revolving loans are floating around 170000

Comparison of code_reject_reson & amt_credit



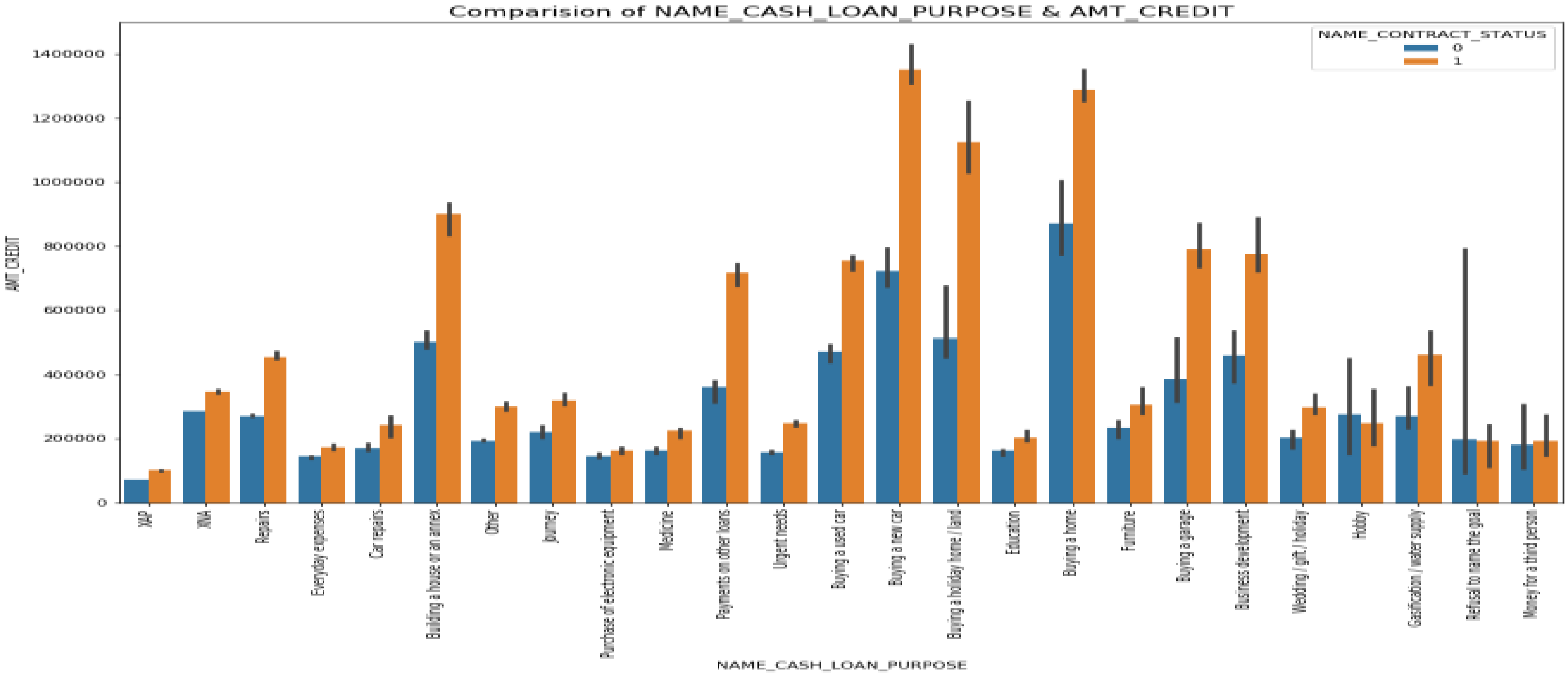
Most of the higher amount loans were rejected with reason code 'VERIF'
Reason code 'SCOFR' had the lowest average loan

Comparison of name_goods_category and amt_credit



House construction has the highest average(~2600000) loan amount and all of them were rejected

Comparison of Name_cash_loan_purpose and amt_credit



Buying a new car, Buying a holiday home, building a house have higher average loan amount varies from ~900000 to ~ 1300000

These type of cash loans have very high refusal rate

Conclusion

- **The top 5 variables to be considered for Loan Prediction are:**

- **Amount of Loan**
  - As seen in the analysis, if the loan amount is high, the application gets rejected.
- **Purpose of Loan**
  - This is very important variable, as observed, if the purpose of cash loan is for buying a car, building a new holiday home, then the application is very likely to get rejected.
- **Term of the loan**
  - More than 55% applicants applied for more than 36 Months of payment and mostly got rejected.
- **Type of employment**
  - As per the analysis - the low skilled labourers have the most difficulty in making repayment
- **Income of the Applicants**
  - If the income of the applicants is low, then he/she is more likely to be defaulter. As seen in the analysis
Working class has the most difficulty in the making repayment