Credit EDA ASSIGNMENT

By Rishabh Gupta & Mandar darekar
C22 Upgrad DATA Science COHORT

Problem Statement

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.
When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

Approach Taken

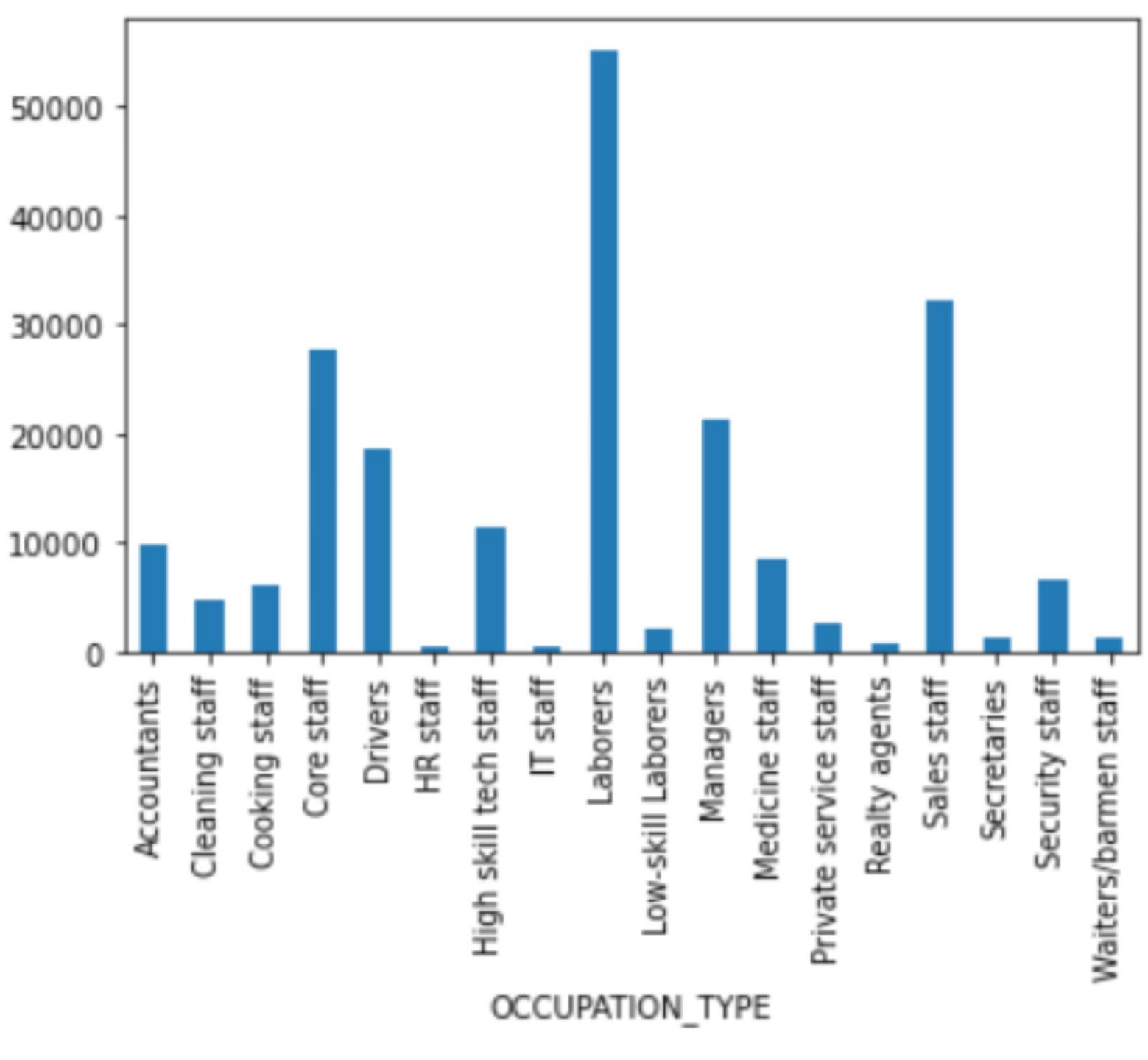
- Cleanse the data to identify nulls and outliers and use proper steps to handle them/ impute the missing values with suitable values based on the assumptions
- Data Preparation: Remove unwanted columns based on valid assumptions and then proceed forward with clean data to univariate analysis, divide the dataset based on target variables to identify differences based on the analysis done
- Univariate analysis: Plot various variables with their count/ mean/ highest values to draw insights
- Bivariate Analysis: Use correlations, trend lines etc to draw insights on dependancy and causation of events based on multiple variables

Data Preparation

- Data Cleaning
- Devising relevant assumptions
- Choosing relevant columns
- Dividing the data set based on target variables
- Creating bins and derived columns

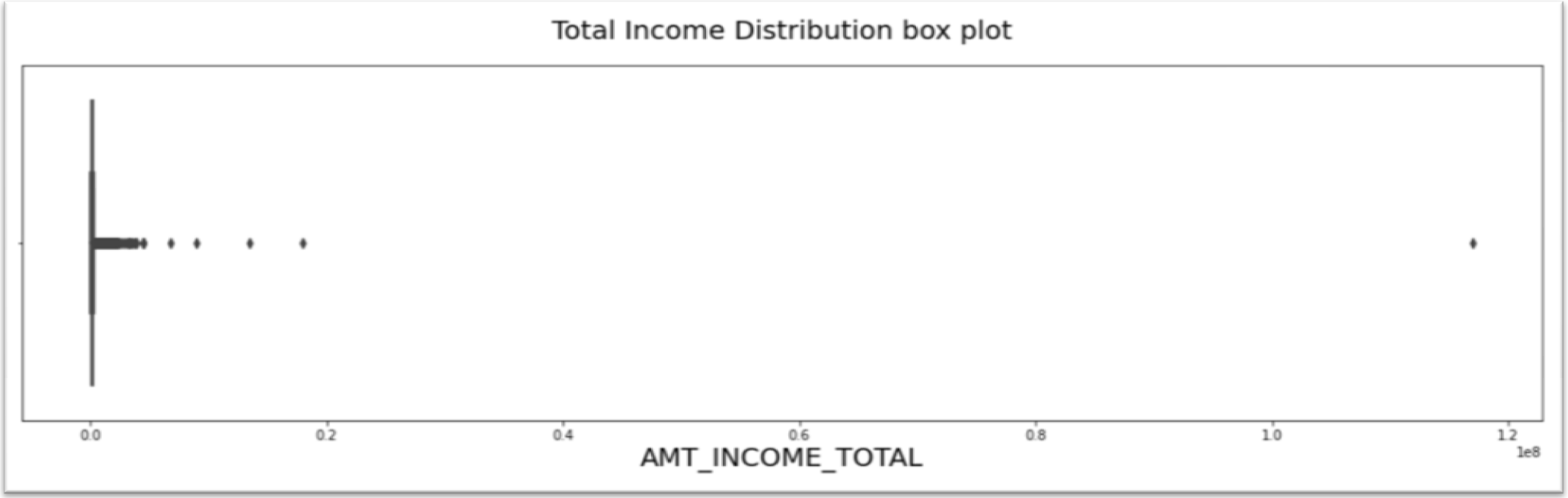
OCCUPATION TYPE

OCCUPATION_TYPE could have been a useful metric to use for our analysis but >30% of the values are null and there is not 1 extremely common occupation applying for loans, Labourers is highest of course but we can not fill everything which is null by labourers since it is not that high also. Hence, for now we would remove this column as well



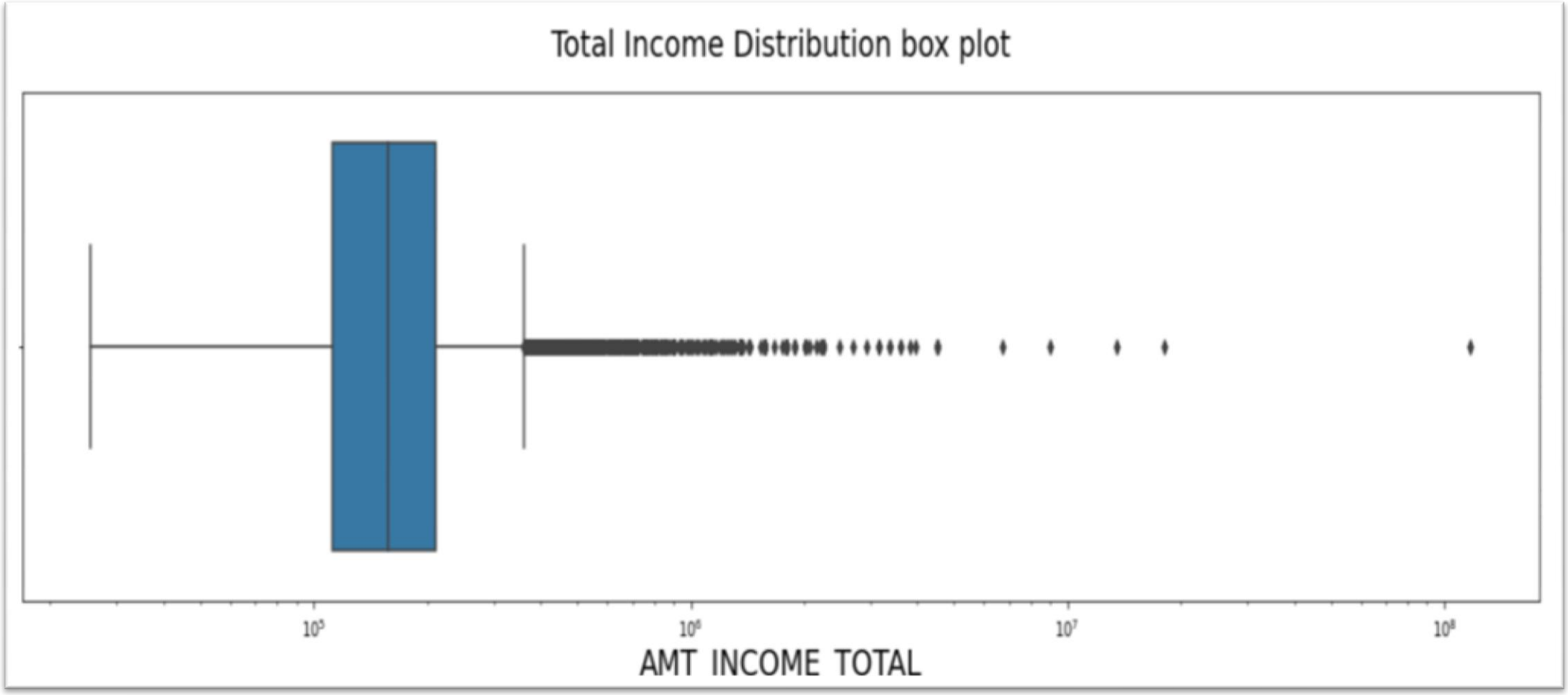
ANALYSIS THE OUTLIERS IN TOTAL INCOME DISTRIBUTION AND REMOVING THEM

We have a couple of outliers here which we can remove, but I would prefer to do analysis by creating buckets over this, so I am leaving the step of removing outliers. Also Lets plot this graph on a log axis to check the outlier details further



TOTAL INCOME DISTRIBUTION

This is now a better representation to see the outliers and devise inferences from.



IDENTIFYING TRASH VALUES AND CORRETING THEM

We see there are some XNA values which are not expected in a couple of fileds. let go forward and remove those before proceeding to the analysis state

In [43]:
```
for x in categoric_col:
    r,c = df1[df1[x] =='XNA'].shape
    if r>0:
        print(x)
```

```
CODE_GENDER
ORGANIZATION_TYPE
```

We have 2 columns with XNA values. We have to go ahead and remove them. We would replace them with either mode or end up removing the complete rows if the percentage is high

In [44]:
```
# Changing the XNA for CODE_GENDER
df1[df1['CODE_GENDER']=='XNA'].shape
```

Out[44]:
```
(4, 32)
```

In [45]:
```
df1['CODE_GENDER'].value_counts()
```

Out[45]:
```
F      202448
M      105059
XNA         4
Name: CODE_GENDER, dtype: int64
```

we would replace those 4 values with F as F is higher substantially

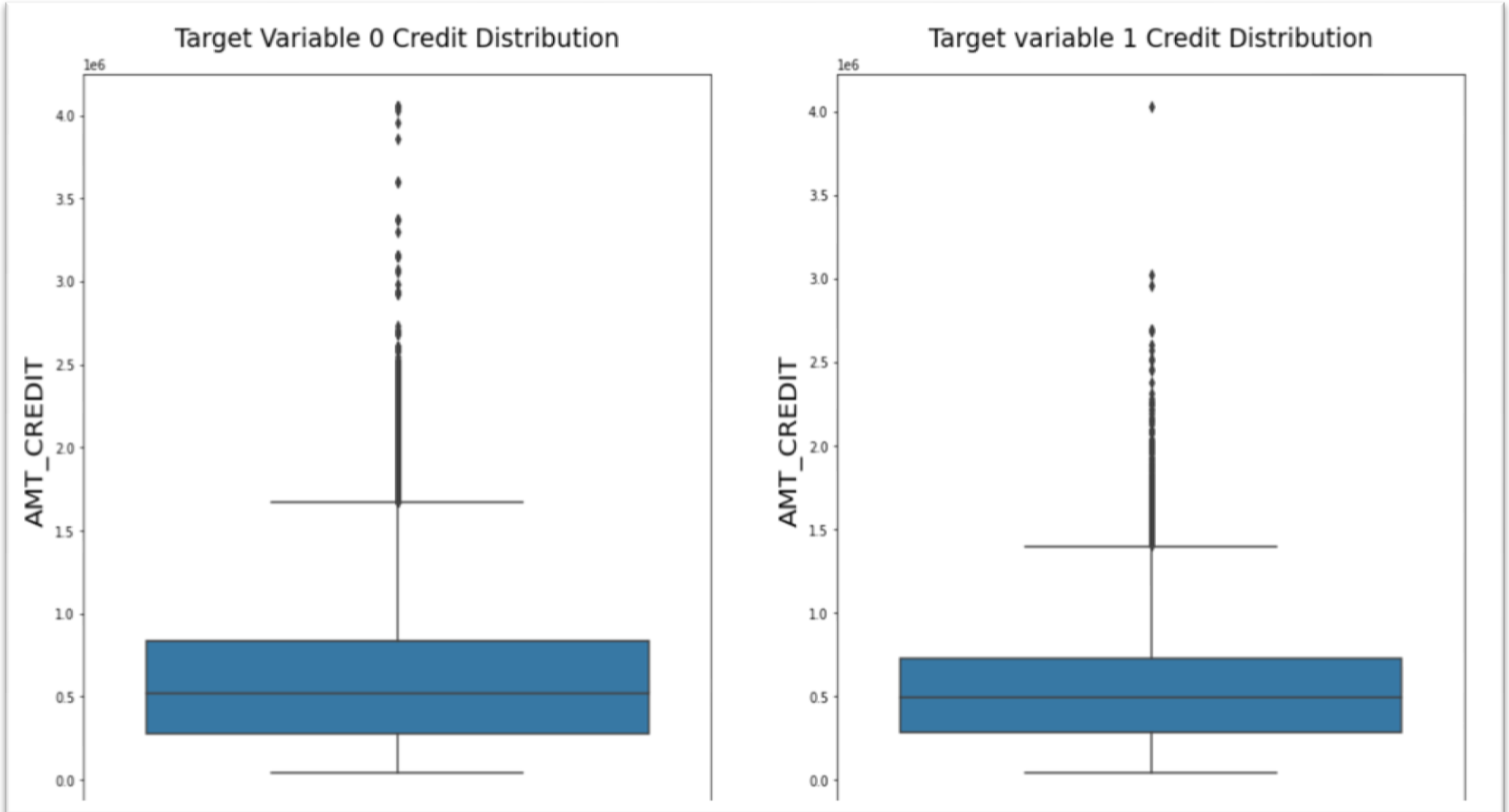
In [46]:
```
# Updating the column 'CODE_GENDER' with "F" for the dataset

df1.loc[df1['CODE_GENDER']=='XNA','CODE_GENDER']='F'
df1['CODE_GENDER'].value_counts()
```

Univariate Analysis

- Below are some sample plots we created as a part of univariate analysis. To look at the complete analysis please refer to the notebook.

CREDIT AMOUNT DISTRIBUTION FOR TARGET

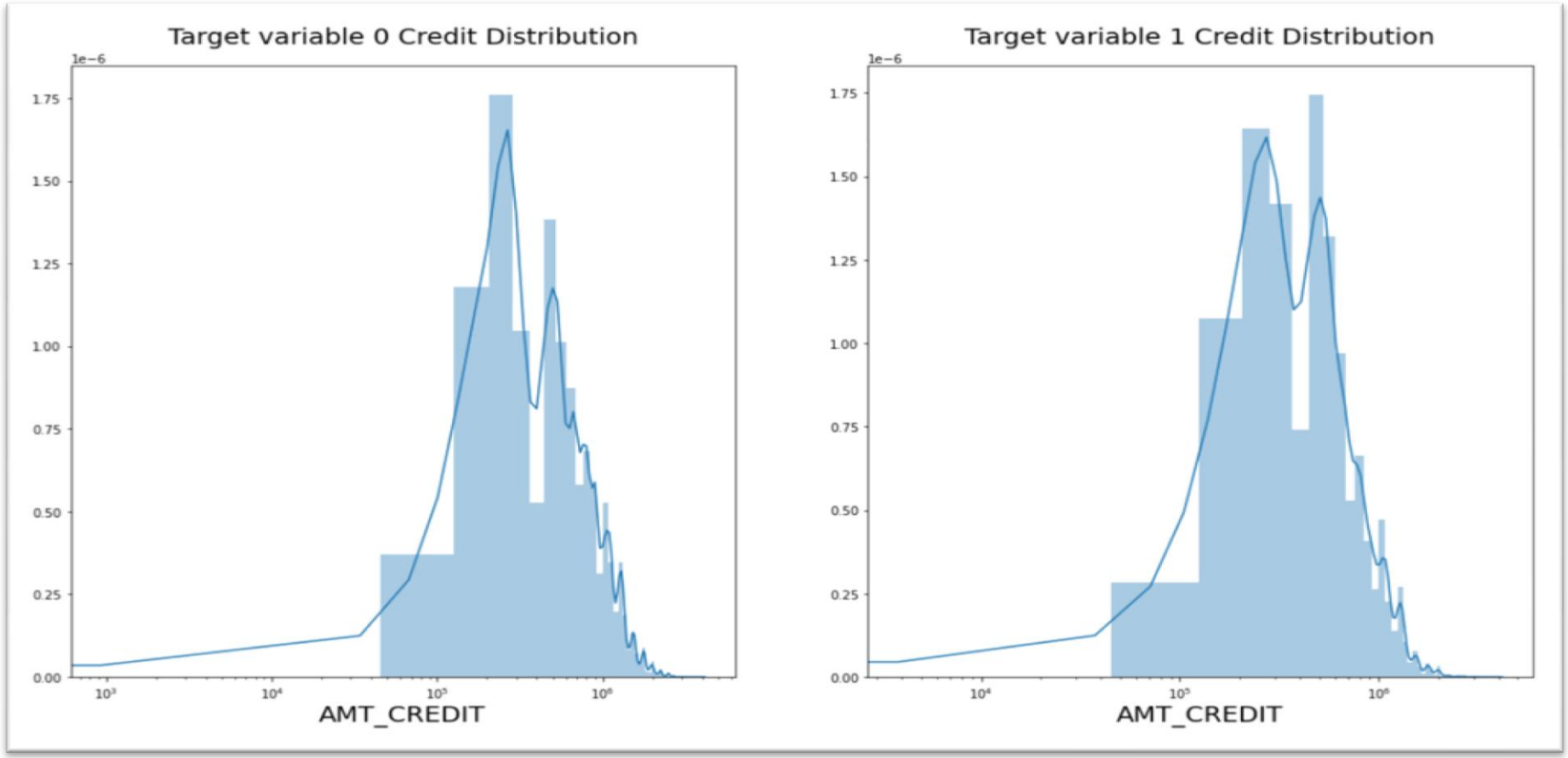
- Median for both target variables seem to be same
- The percentage of people in IQR- 25 to 75 percentile is more for the people who have never defaulted a loan payment



Lets do some analysis on the major peak values for all numeric variables to check what is the most common income people who take loans, what's the most common credit amount and annuity amount for both target variables

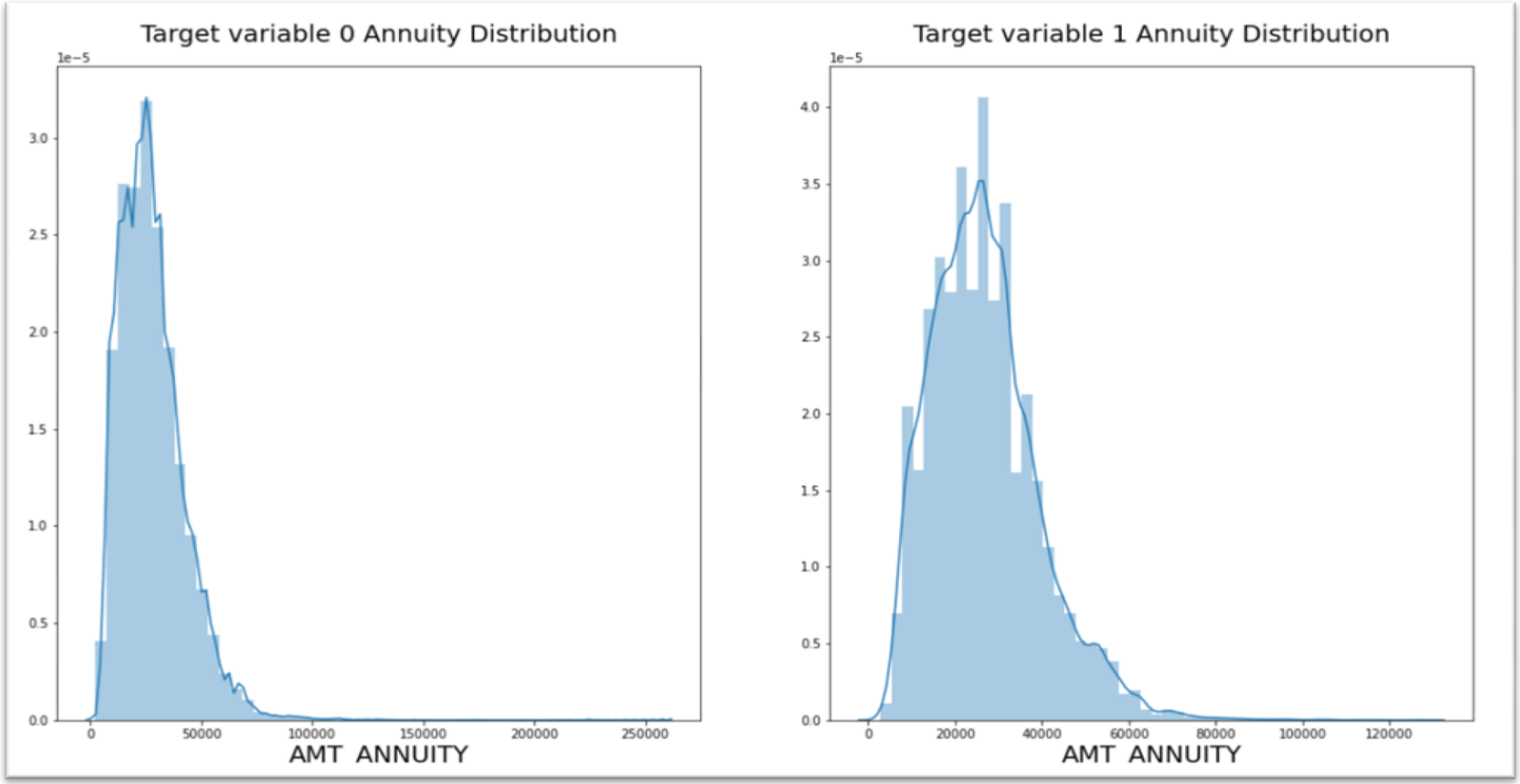
CREDIT AMOUNT DISTRIBUTION

- The peak for credit amount of target variable 1 is higher than that of target variable 0.
- But using the peak we can not create a general scenario statement, because the distribution seem to be similar for both of them



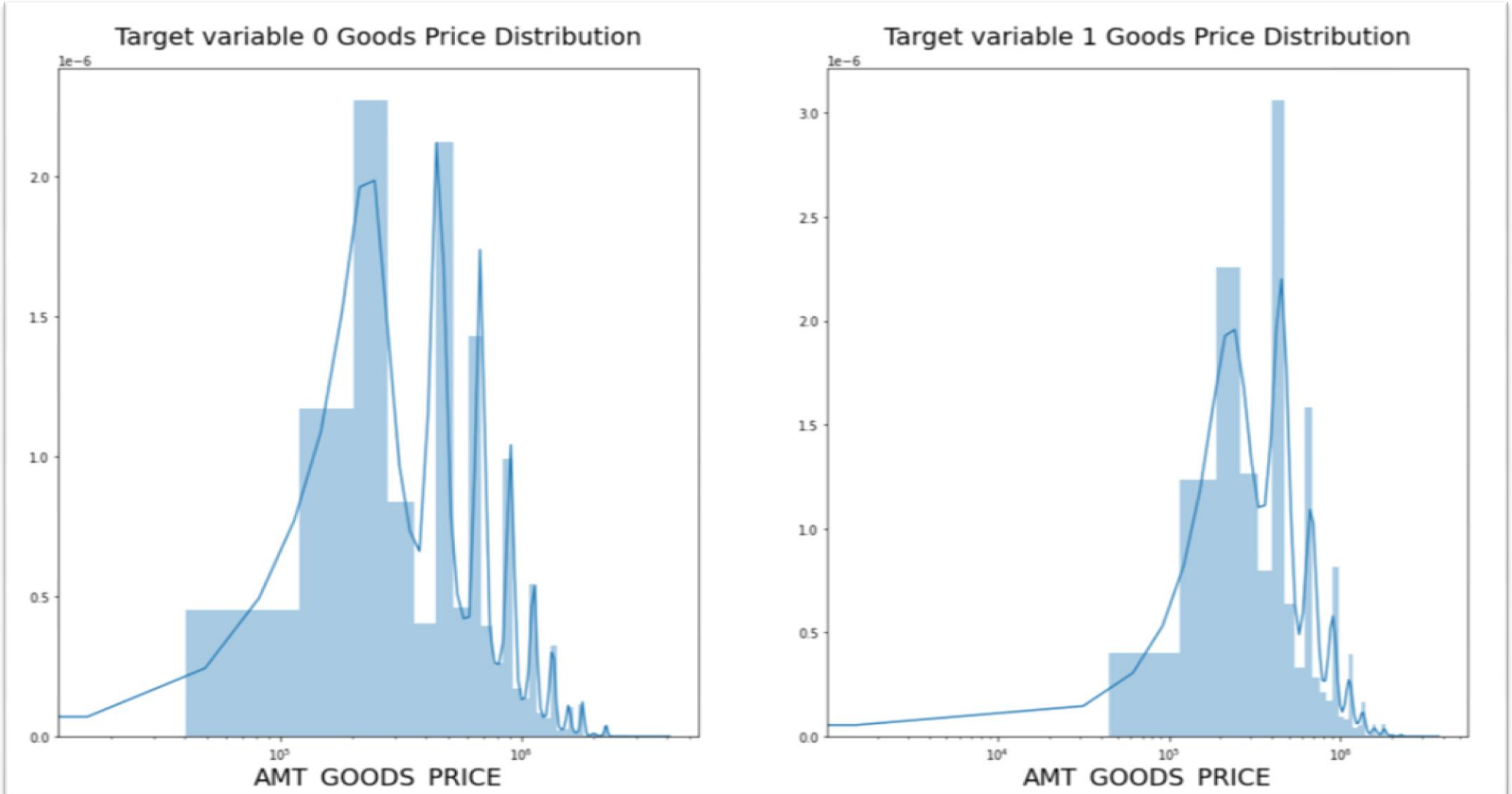
ANNUITY DISTRIBUTION

Again not a lot of difference that can be drawn from both the graphs as the distribution is pretty much the same



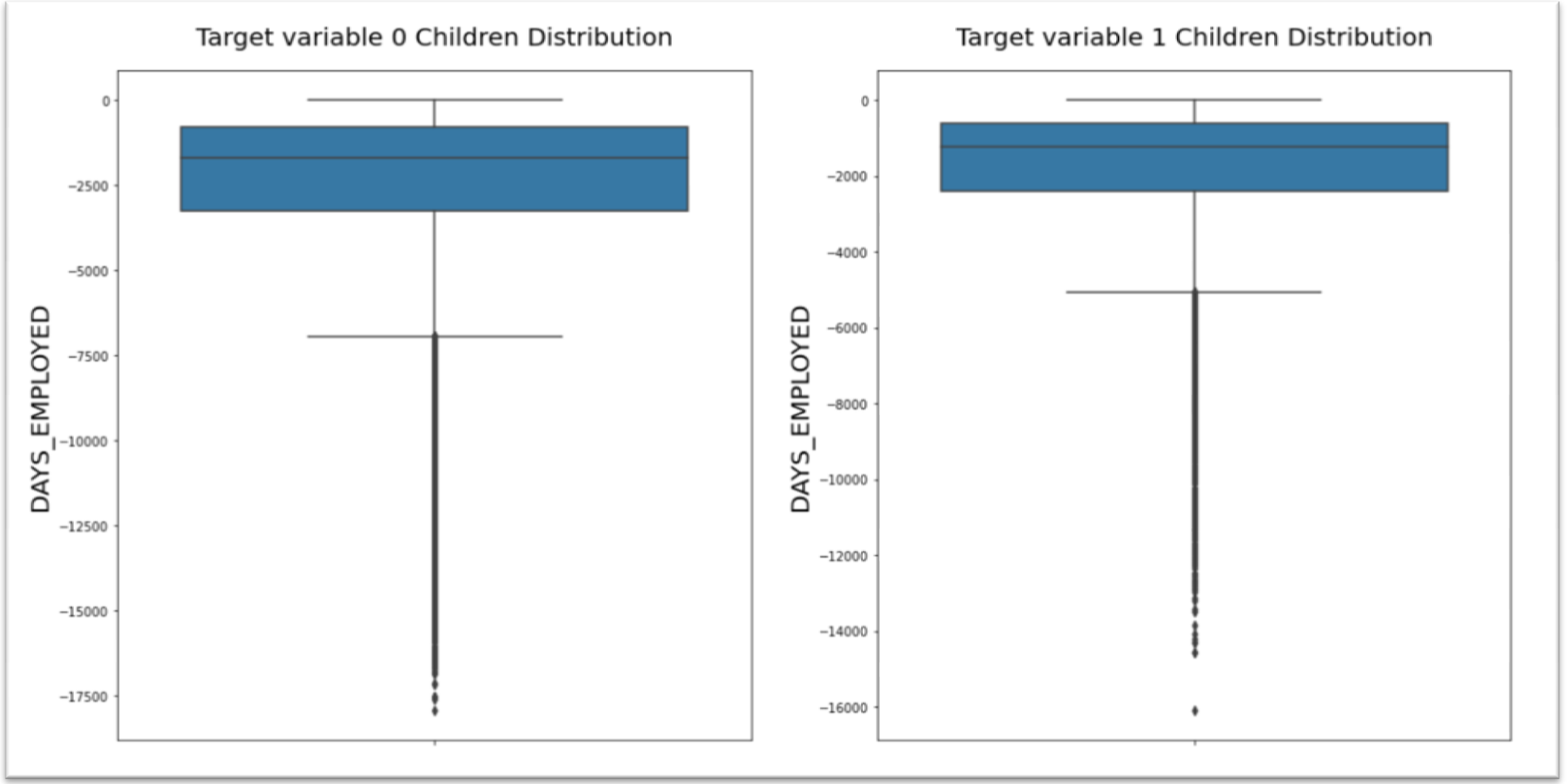
GOODS PRICE DISTRIBUTION

Even though the peak for the goods price for target variable 1 is higher the distribution is more or less the same



GOODS DISTRIBUTION FOR CHILDREN

we see there is a slight factor of people defaulting more if their days of employment is < 2500 days. This is just a very slight trend and not a lot of decisions should be taken based on this



Imbalance ratio

- Imbalance ratio comes out to be ~10.55

Imbalance Percentage

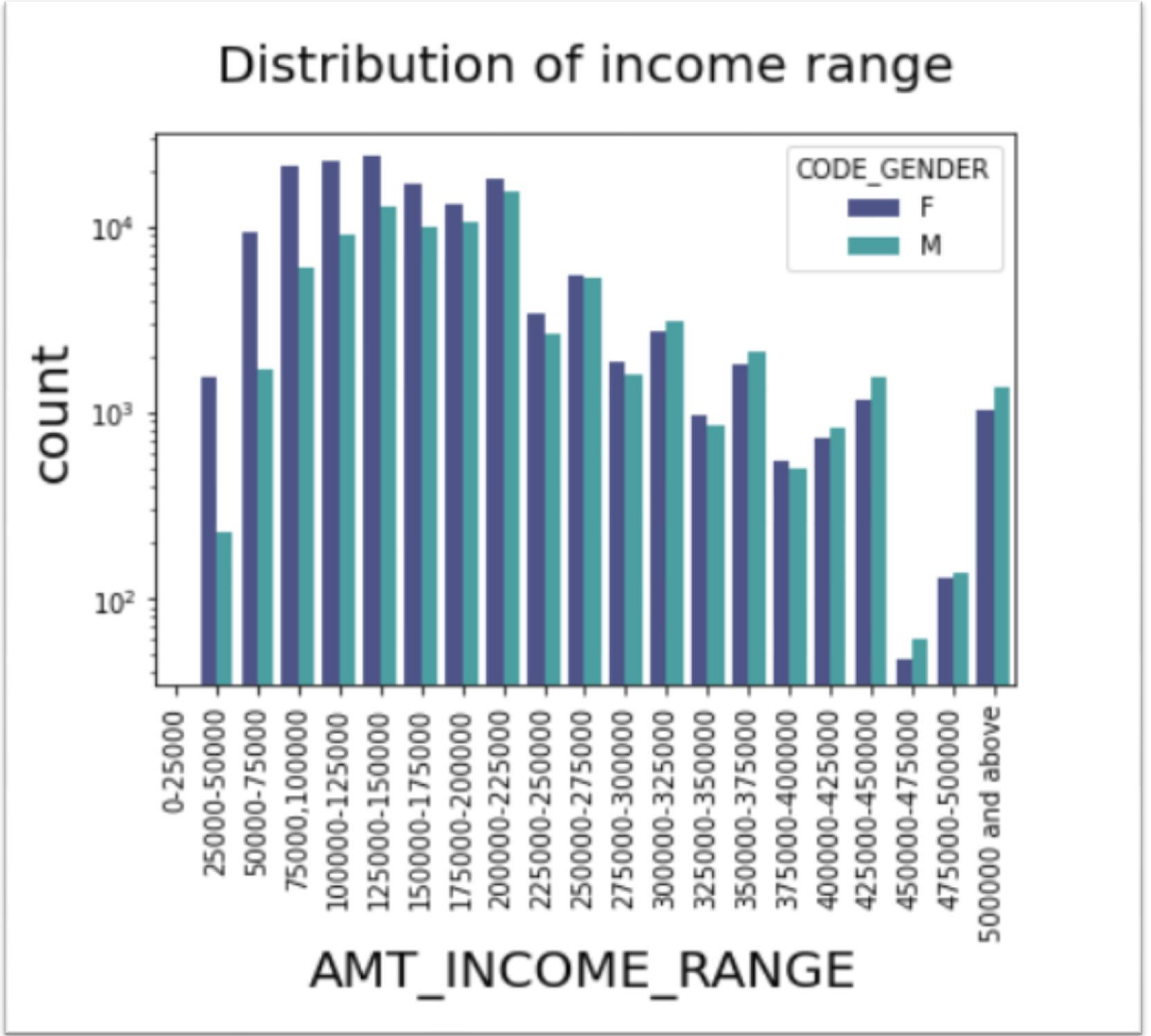
```
In [67]: # Lets go forward and calculate the imbalance percentage of teh target variable


         len(d0)/len(d1)
```

```
Out[67]: 10.547378062743302
```

INCOME RANGE

Inferences from the graph are as follows:

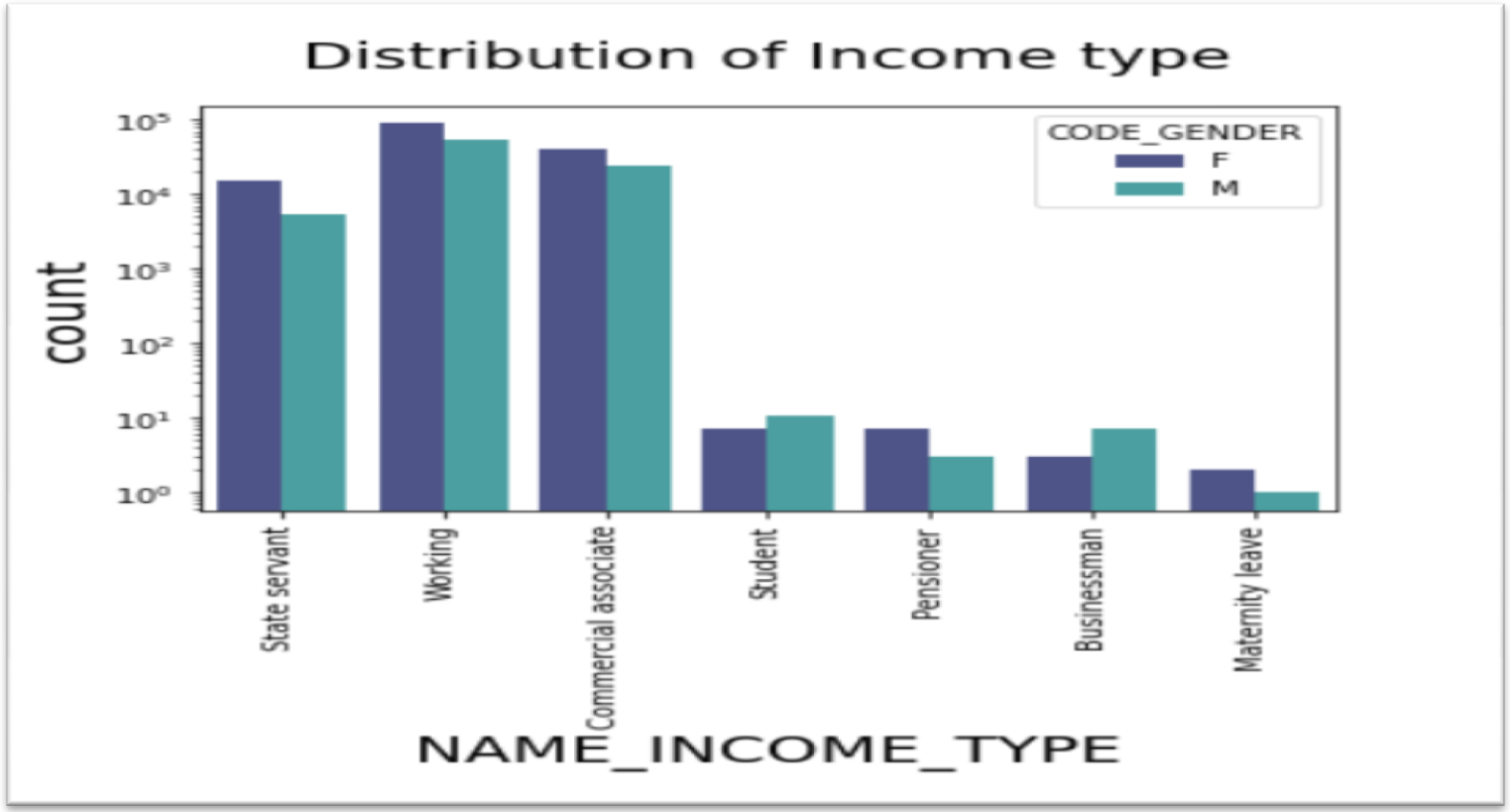
- Count of females are higher than males in most of the buckets of Income Range where they don't have any problems in payment of loan
- Income range from 125000 to 150000 is having most number of loan applications
- On a general trend the count of both males and females decreases as their income increases



INCOME TYPE : WORKING

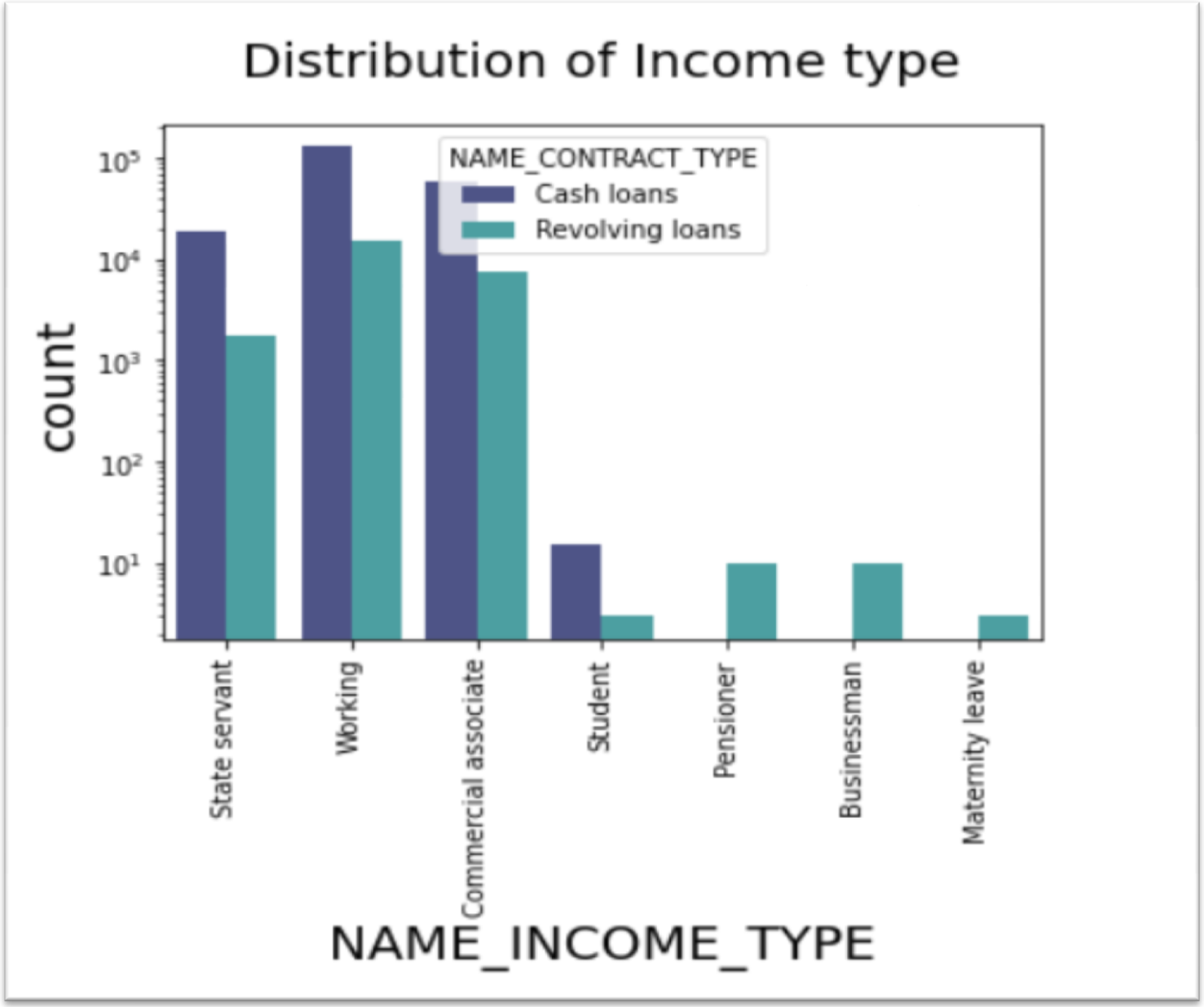
Inference:

- For income type 'Working' the number of loan applications are highest
- For the Females belonging to this category are having more number of loan applications than males do



CONTRACT TYPE

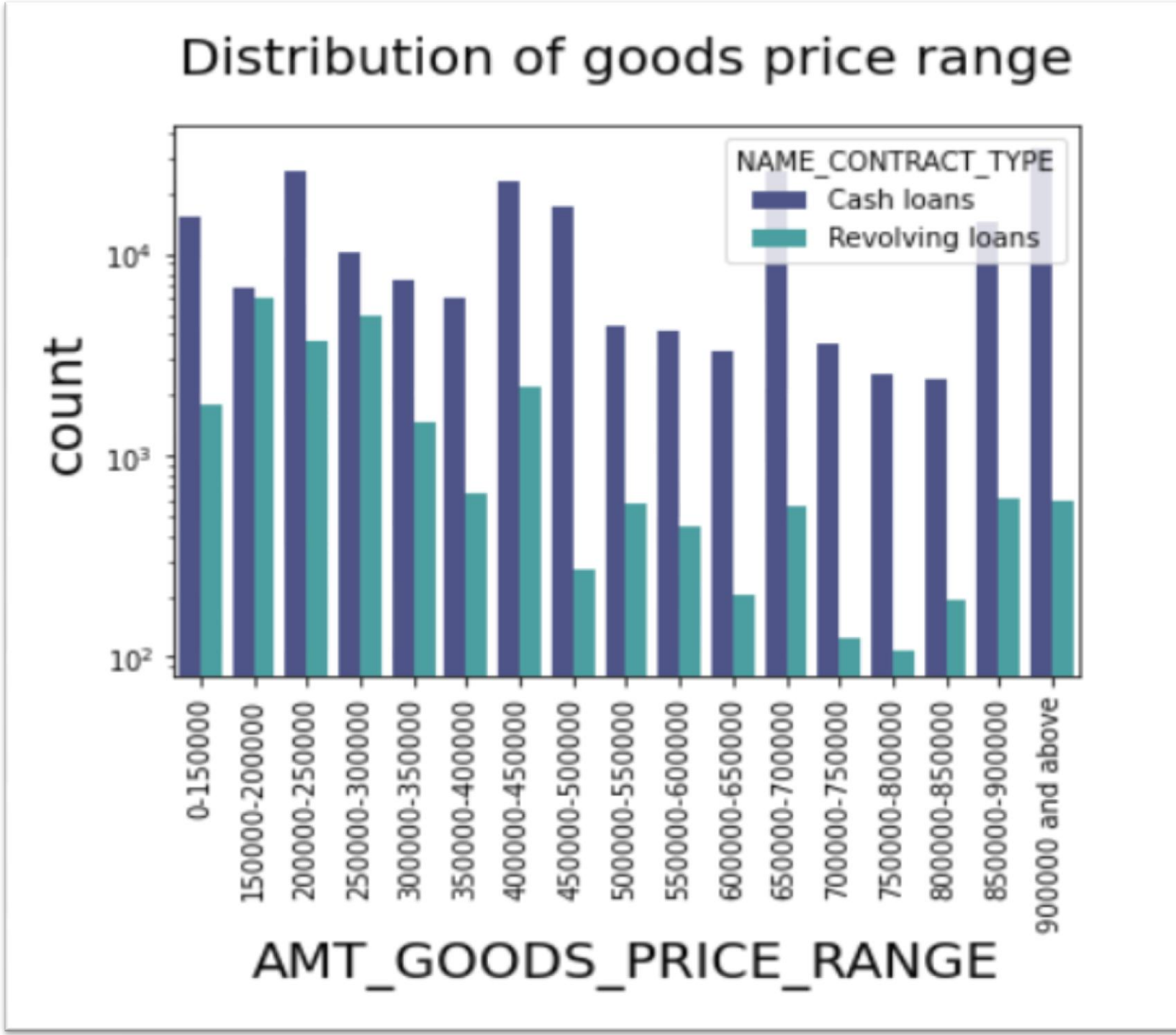
Inferences: Cash loans are more preferred by State Servants, working, Commercial Associates and Students whereas Pensioners, Businessmen, Maternity Leave people do not prefer cash loans at all



PRICE OF GOODS & CATEGORIZING BY CONTRACT TYPE

Inferences:

- Cash loans are preferred regardless of price of goods
- We see a general trend of revolving loans decreasing as the value of goods increases
- There is no trend as such in cash loans with changes in price of the goods

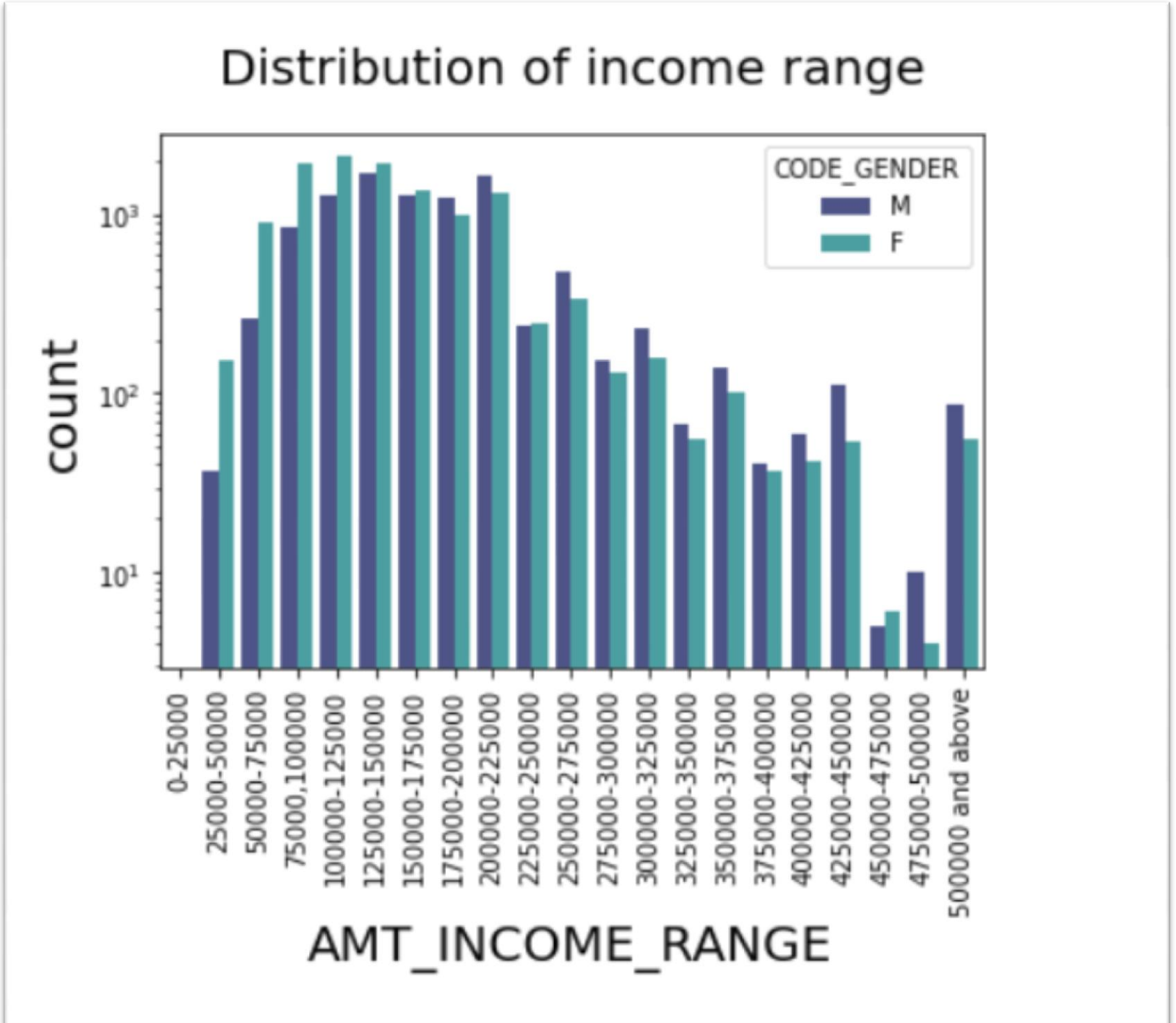


Now that we have done this analysis for target 0 variable, we will do the same analysis for target 1 to see how the trends and inferences vary. This will help us see the difference and also to identify the population we should aim for giving loans

INCOME RANGE

Inferences:

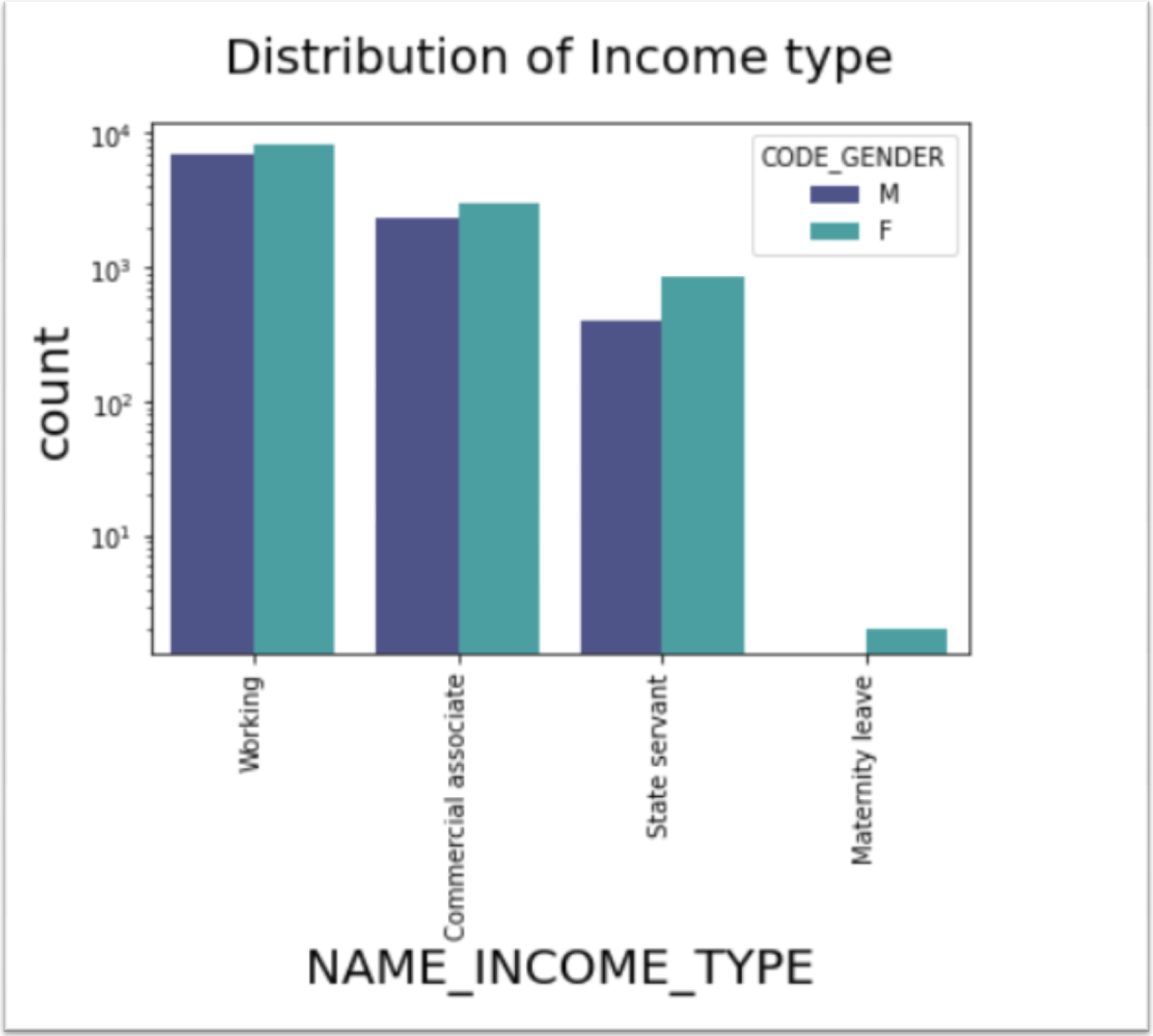
- The general trend remains the same as that of target variable = 0
- Male count applications increase as the income range increases
- At the income ranges where there is a peak of applications, both males and females count are same
- Not a lot of difference as compared to the one where target variable =0, so no major difference pattern observed



GENDER CLASSIFICATION REPRESENTATION

Inferences:

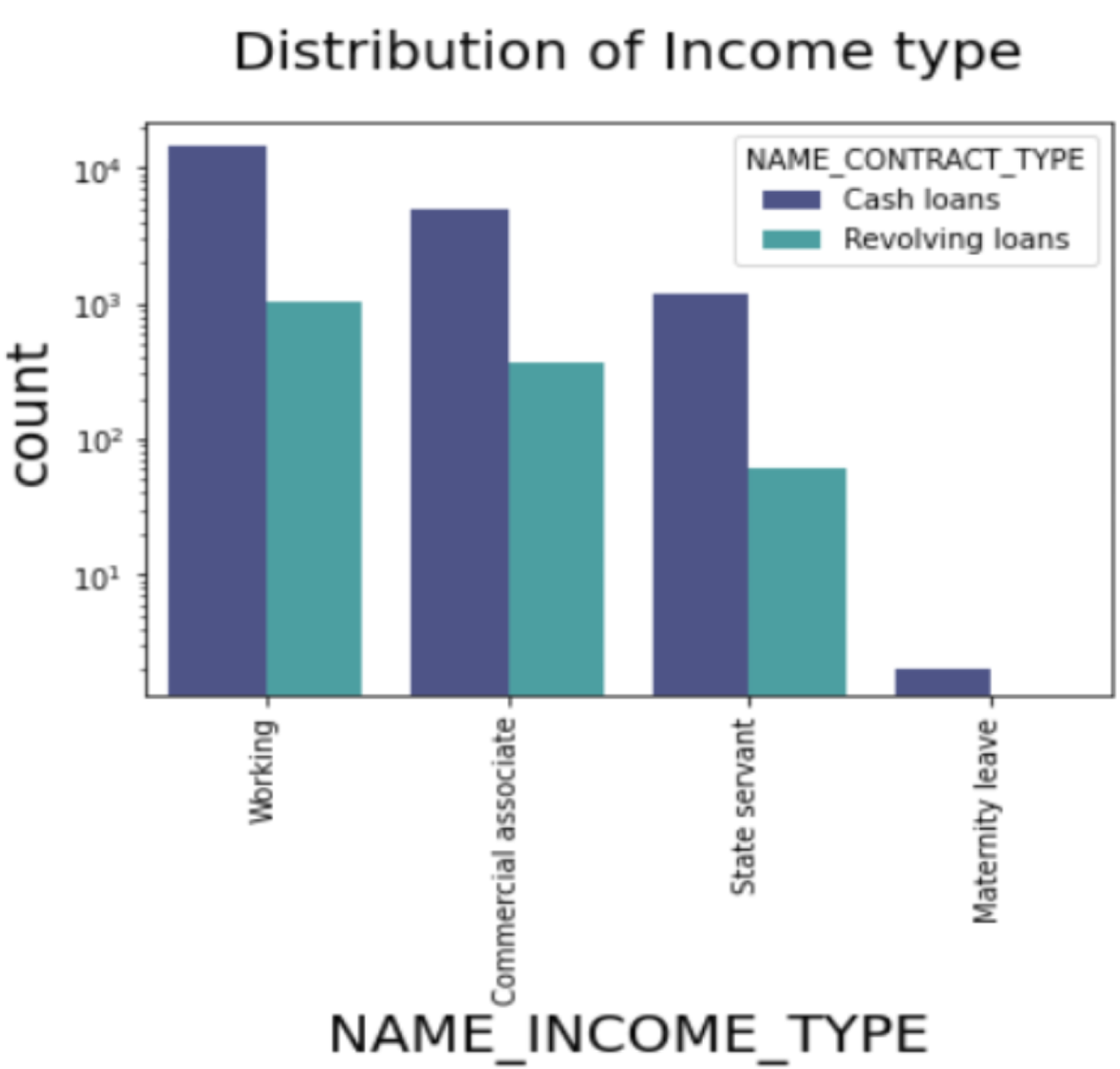
- We see quite some interesting inferences here as there are no Students, Pensioners, Businessmen and we find they never default their instalments, which also means they should be given loan a bit more easily than the other categories
- Even Maternity Leave females also default less, so they could also be considered with the above population
- At last, again we see their are higher female applications with a history of default



Contract type classification

Inferences:

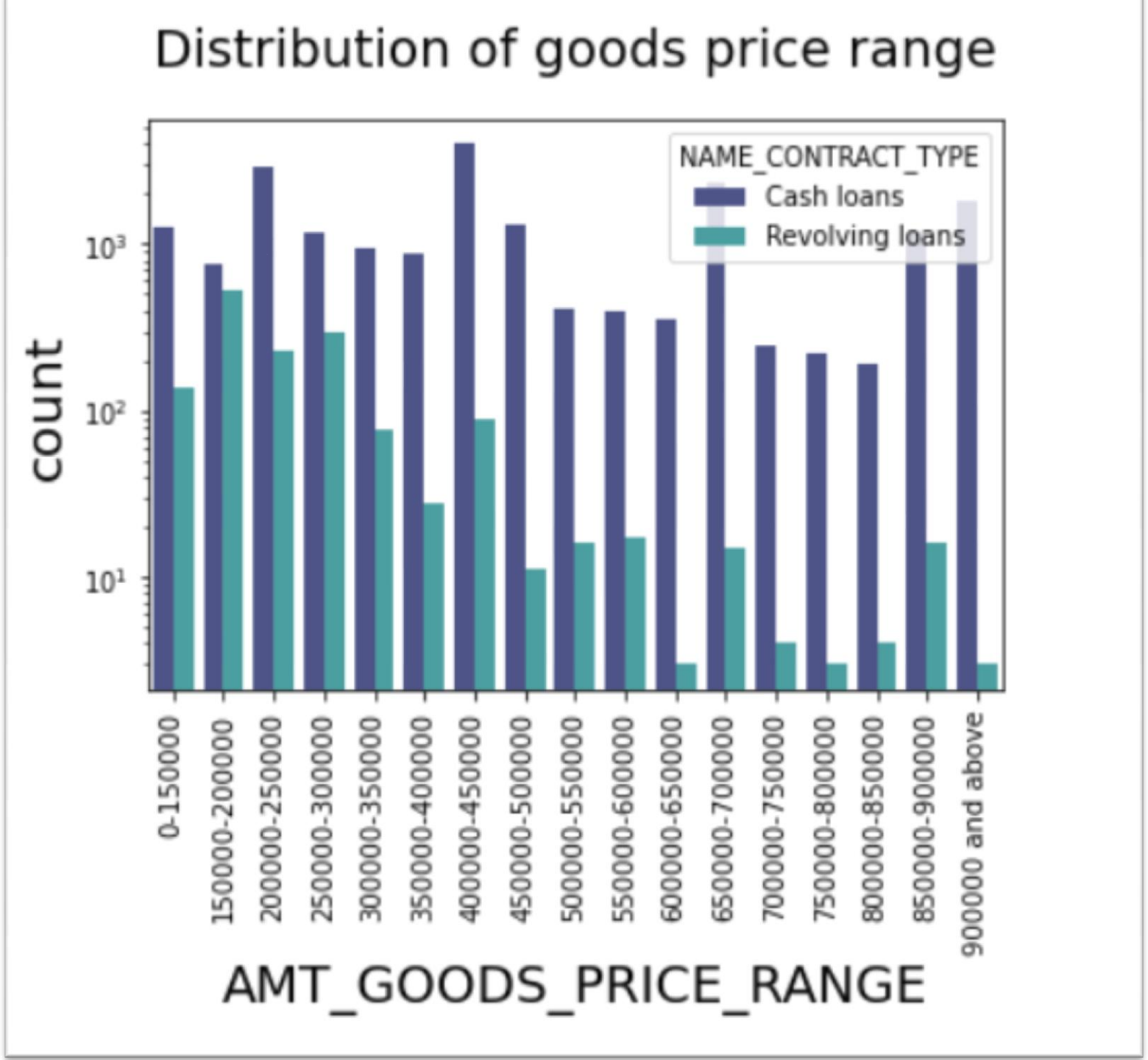
For the categories who default there is more chances of default on the cash loans then on revolving loans



Price of goods and categorizing by contract type

Inferences:

Again no solid trend on whether the price of goods impact the count of people defaulting at least one instalment, but a general trend is seen in decrease of revolving loans as the price of goods increase



Recommendations based on Univariate Analysis

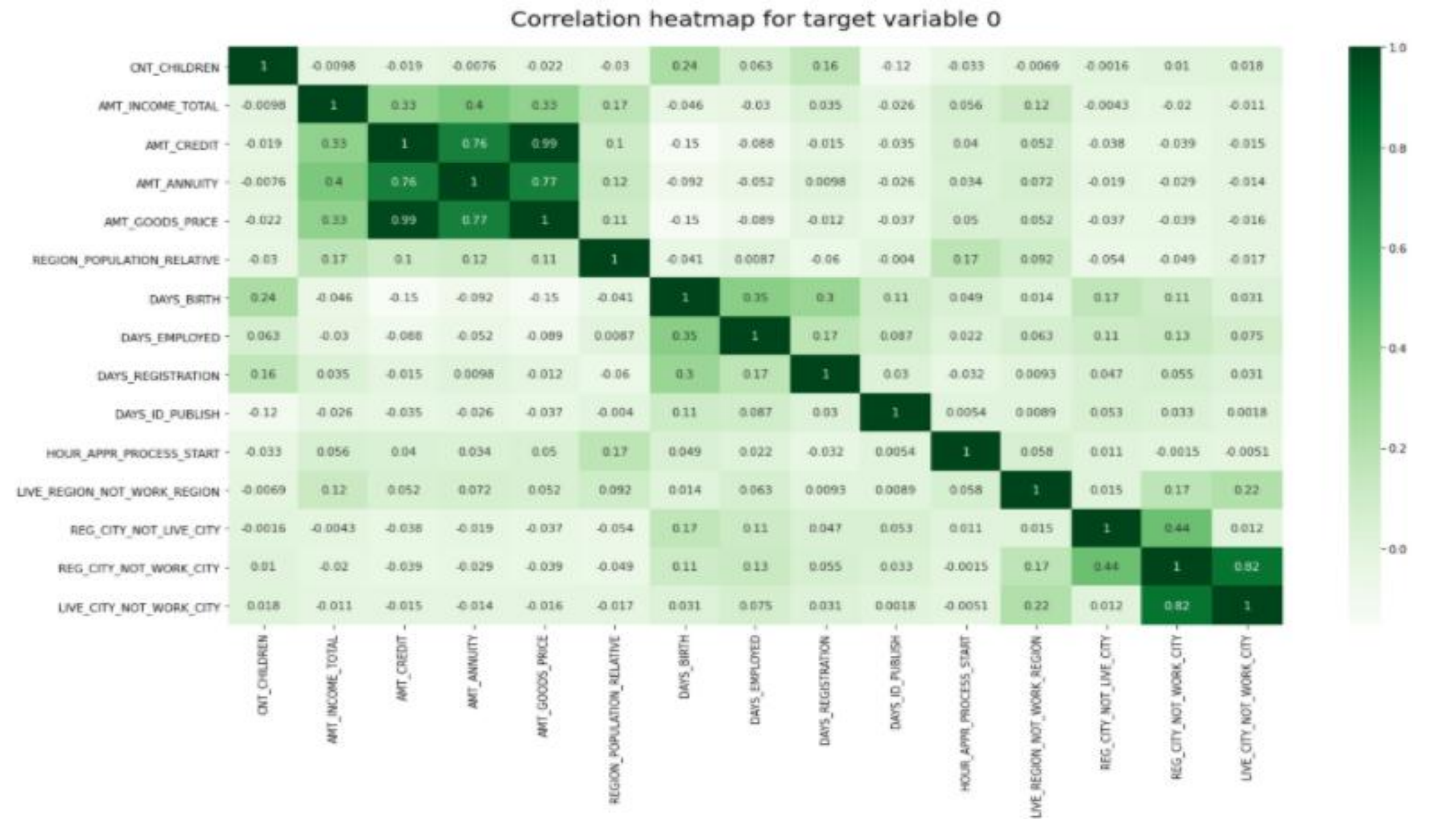
So major differences in target variable = 0 and those with target variable = 1 are that Students, Pensioners, Businessmen have never defaulted the instalment, also, female members on maternity leave have rarely defaulted on paying. So we can draw our decisions based on this analysis for future

Bivariate Analysis

Lets now go ahead and plot the correlations and scatter plots and heat maps and see which variables are more closely related. This would also tell us the dependency and causation of variables here and there

Correlation matrix

- Correlation matrix are created for both target 1 and target 0
- Heatmap are created on these correlation matrix to identify top 10 correlated variables

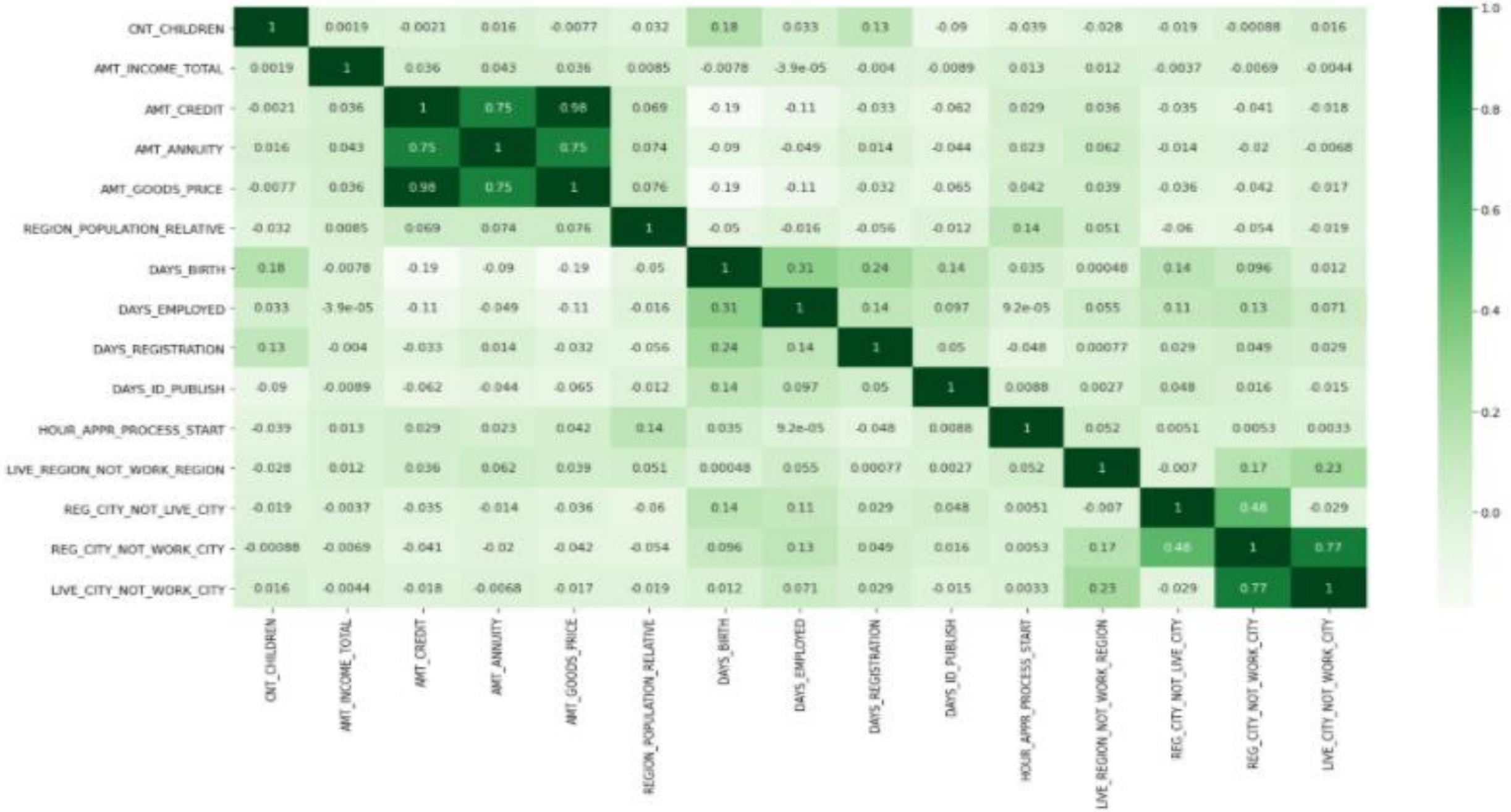

Correlation heatmap for target variable 0
CNT_CHILDREN 1 -0.0098 -0.019 -0.0076 -0.022 -0.03 0.24 0.063 0.16 -0.12 -0.033 -0.0069 -0.0016 0.01 0.018
AMT_INCOME_TOTAL -0.0098 1 0.33 0.4 0.33 0.17 -0.046 -0.03 0.035 -0.026 0.056 0.12 -0.0043 -0.02 -0.011
AMT_CREDIT -0.019 0.33 1 0.76 0.99 0.1 -0.15 -0.068 -0.015 -0.035 0.04 0.052 -0.038 -0.039 -0.015
AMT_ANNUITY -0.0076 0.4 0.76 1 0.77 0.12 -0.092 -0.052 0.0098 -0.026 0.034 0.072 -0.019 -0.029 -0.014
AMT_GOODS_PRICE -0.022 0.33 0.99 0.77 1 0.11 -0.15 -0.089 -0.012 -0.037 0.05 0.052 -0.037 -0.039 -0.016
REGION_POPULATION_RELATIVE -0.03 0.17 0.1 0.12 0.11 1 -0.041 0.0087 -0.06 -0.004 0.17 0.092 -0.054 -0.049 -0.017
DAYS_BIRTH 0.24 -0.046 -0.15 -0.092 -0.15 -0.041 1 0.35 0.3 0.11 0.049 0.014 0.17 0.11 0.031
DAYS_EMPLOYED 0.063 -0.03 -0.068 -0.052 -0.089 0.0087 0.35 1 0.17 0.087 0.022 0.063 0.11 0.13 0.075
DAYS_REGISTRATION 0.16 0.035 -0.015 0.0098 -0.012 -0.06 0.3 0.17 1 0.03 -0.032 0.0093 0.047 0.055 0.031
DAYS_ID_PUBLISH -0.12 -0.026 -0.035 -0.026 -0.037 -0.004 0.11 0.087 0.03 1 0.0054 0.0089 0.053 0.033 0.0018
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 0.058 0.011 -0.0015 -0.0051
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 1 0.015 0.17 0.22
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 0.015 1 0.44 0.012
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 0.17 0.44 1 0.82
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 0.22 0.012 0.82 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
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
1.0
0.8
0.6
0.4
0.2
0.0

From the heatmap above, we observe the following conclusions:

```
Top 10 correlated variables:
1. AMT_GOODS_PRICE and AMT_CREDIT
2. LIVE_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY
3. REG_CITY_NOT_WORk_CITY and REG_CITY_NOT_LIVE_CITY
4. AMT_CREDIT and AMT_ANNUITY
5. AMT_GOODS_PRICE and AMT_ANNUITY
6. AMT_GOODS_PRICE and AMT_INCOME_TOTAL
7. AMT_CREDIT and AMT_INCOME_TOTAL
8. AMT_GOODS_PRICE and AMT_INCOME_TOTAl
9. DAYS_EMPLOYED and DAYS_BIRTH
10. DAYS_REGISTRATION and DAYS_BIRTH
```

All the above variables are positively correlated and hence, as one increase other increases.

Credit Amount/ Annuity Amount is negetively correlated to the count of children, days of birth, days of employment and days of registration. This makes sense as the credit amount will be more for people who are younger and have lesser kids and more income.


Correlation heatmap for target variable 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
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
1.0
0.8
0.6
0.4
0.2
0.0

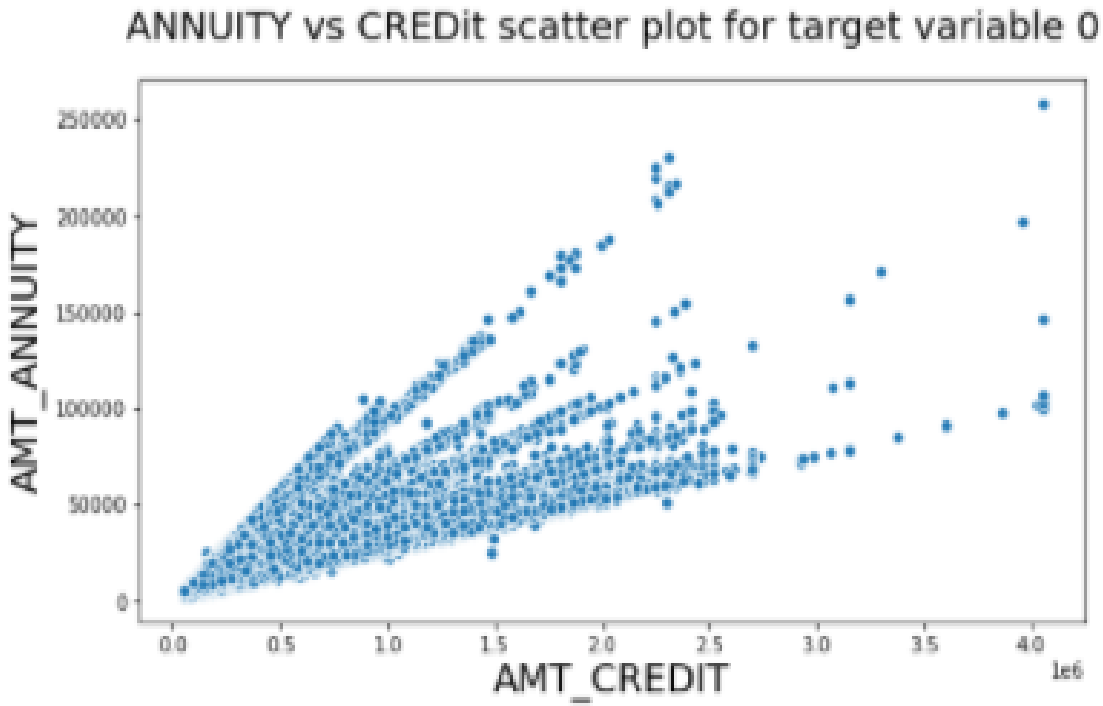
From the heatmap above, we observe the following conclusions:

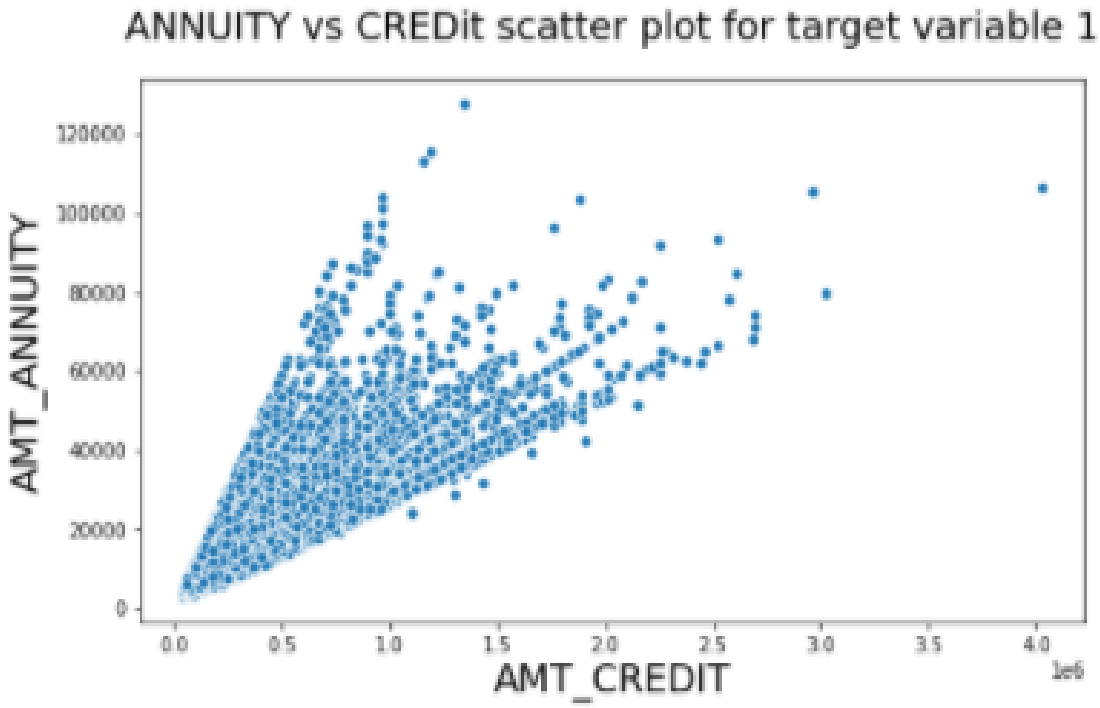
Top 10 correlated variables:

1. AMT_GOODS_PRICE and AMT_CREDIT
2. LIVE_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY
3. REG_CITY_NOT_WORk_CITY and REG_CITY_NOT_LIVE_CITY
4. AMT_CREDIT and AMT_ANNUITY
5. AMT_GOODS_PRICE and AMT_ANNUITY
6. AMT_GOODS_PRICE and AMT_INCOME_TOTAL
7. AMT_CREDIT and AMT_INCOME_TOTAL
8. AMT_GOODS_PRICE and AMT_INCOME_TOTAI
9. DAYS_EMPLOYED and DAYS_BIRTH
10. DAYS_REGISTRATION and DAYS_BIRTH All the above variables are positively correlated and hence, as one increase other increases.

Credit Amount/ Annuity Amount is negetively correlated to the count of children, days of birth, days of employment and days of registration. This makes sense as the credit amount will be more for people who are younger and have lesser kids and more income. The trend for those who default are also similar to the ones who do not default. So major difference here to distinguish a pattern.

Scatter plot between variables to get correlation trends

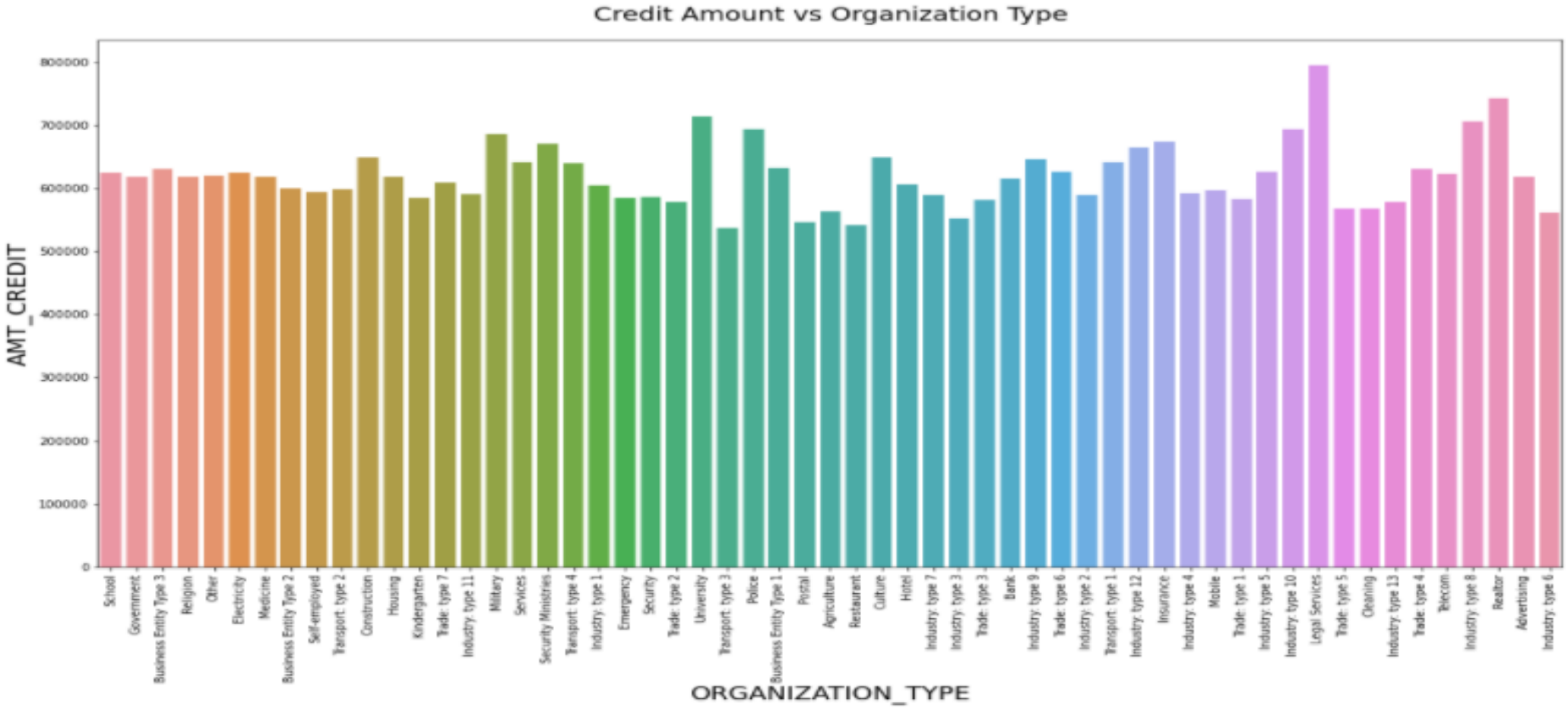




Inferences:

1. Based on these graphs we see a positive correlation between Annuity and Credit Amount for both the target variables
2. We do see a steeper line for target variable 1, which means as a general trend Credit and Annuity correlation is slightly higher for the users who have defaulted

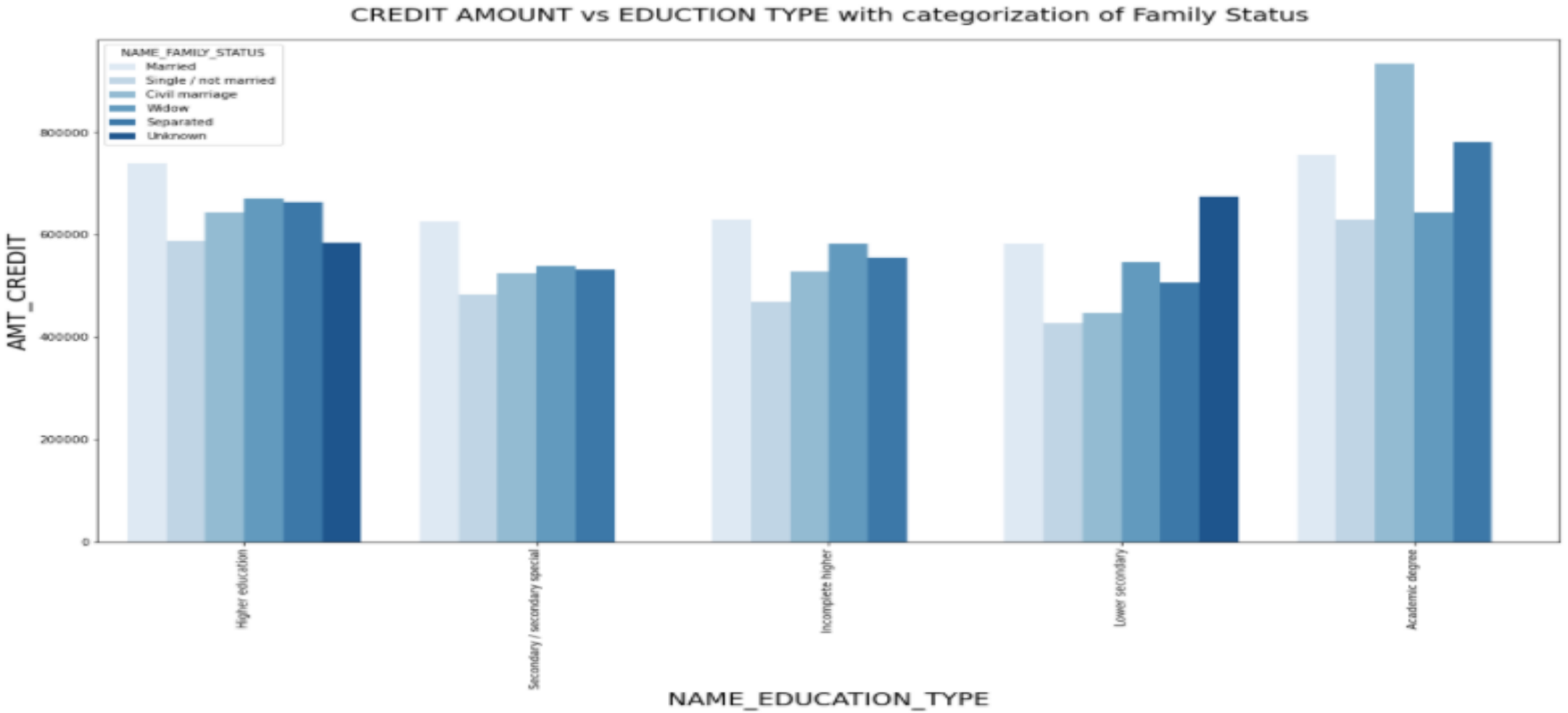
Bar Plot to show which Organization Type has highest Credit amount



Inferences:

Legal Services Organization Type people are taking the highest value loans and Police are taking loans with least Amount

Bar plot to Analyse Family Status with Education Type



Inferences:

Civil married people with academic degree take the highest value loeans and single people with lower secondary education take the lowest value loans. This could also be due to the fact that lower secondary people might not have a great credit history.

Joining Previous Application dataset

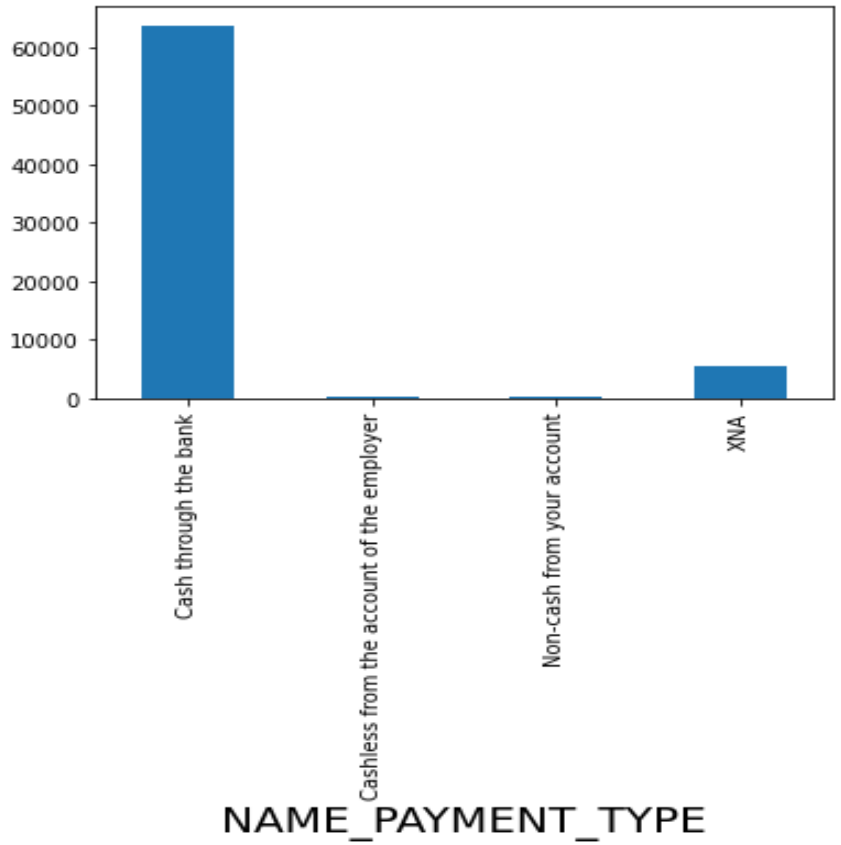
- Same steps of data preparation are involved in this dataset as well before joining and moving ahead

As we go through the data we see there are multiple trash values like XNA and XAP which would effect the overall analysis. Its best we remove them based on the number of rows effected

```
#Checking any fake value or trash values im df2

cat_col = ['NAME_CONTRACT_TYPE','WEEKDAY_APPR_PROCESS_START','FLAG_LAST_APPL_PER_CONTRACT','NFLAG_LAST_
          ,'NAME_CASH_LOAN_PURPOSE','NAME_CONTRACT_STATUS', 'NAME_PAYMENT_TYPE','CODE_REJECT_REASON','N
          'NAME_GOODS_CATEGORY','NAME_PORTFOLIO','NAME_PRODUCT_TYPE','CHANNEL_TYPE',
          'NAME_SELLER_INDUSTRY','PRODUCT_COMBINATION']

for x in cat_col:
    r,c = df2[df2[x] =='XNA'].shape
    if r>0:
        print(x)
```

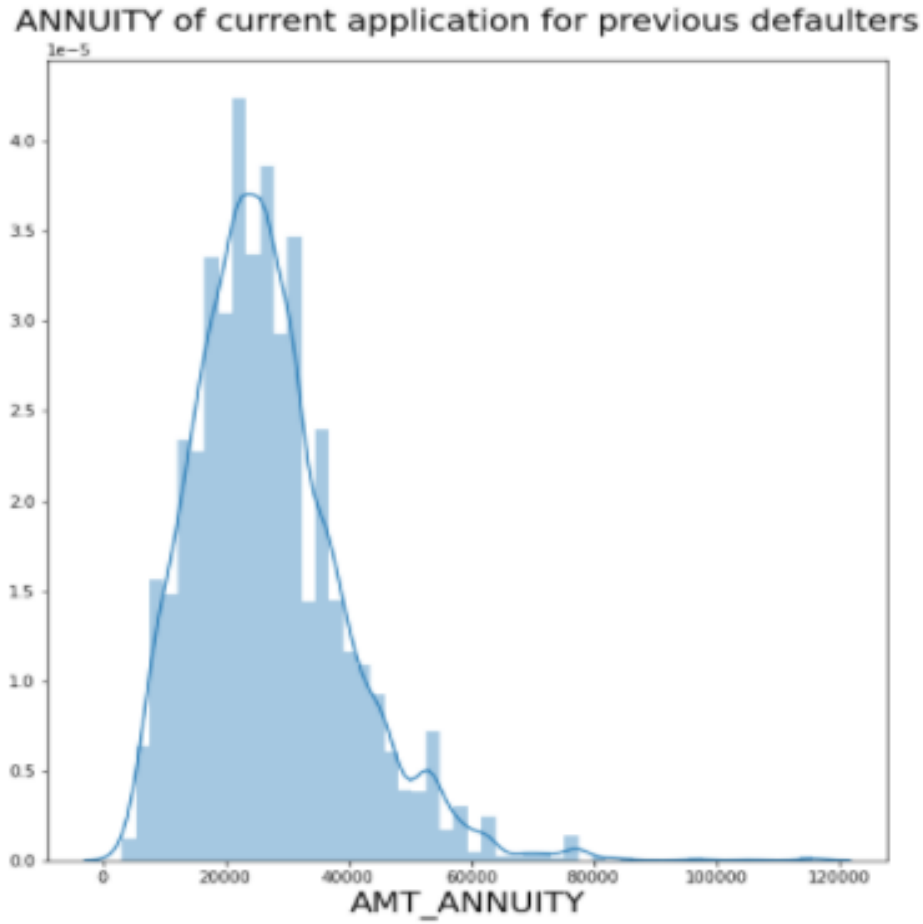
```
NAME_CONTRACT_TYPE
NAME_CASH_LOAN_PURPOSE
NAME_PAYMENT_TYPE
CODE_REJECT_REASON
NAME_CLIENT_TYPE
NAME_GOODS_CATEGORY
NAME_PORTFOLIO
NAME_PRODUCT_TYPE
NAME_SELLER_INDUSTRY
```

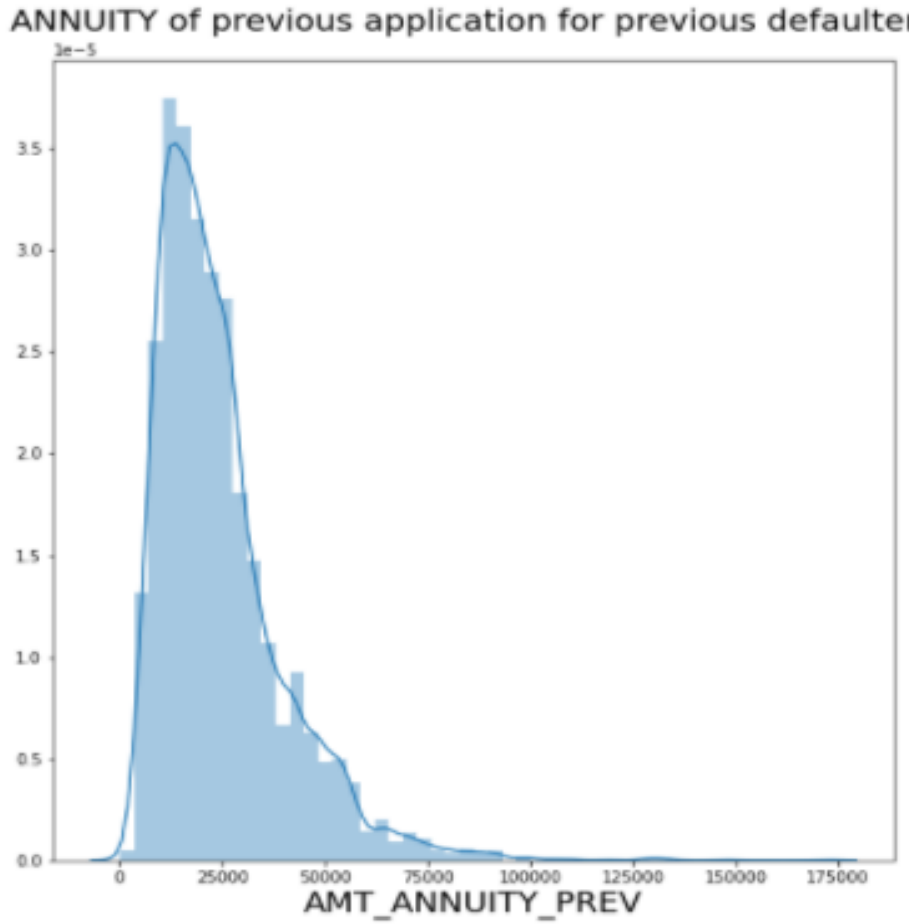
Data Preparation



Since the value Cash through the bank has highest we can replace the XNA by this value. Also for column 'CODE_REJECT_REASON' we do not have proper business definiton for what all these codes mean so we will do the analysis just based on the info provided. 'NAME_CLIENT_TYPE' is an important field so we will try to impute correct values to this as well. But for other columns like 'NAME_GOODS_CATEGORY', 'NAME_PORTFOLIO','NAME_PRODUCT_TYPE','NAME_SELLER_INDUSTRY', we would drop these as we do not necessarily need them for our analysis

Univariate Analysis post merging and cleaning of data sets



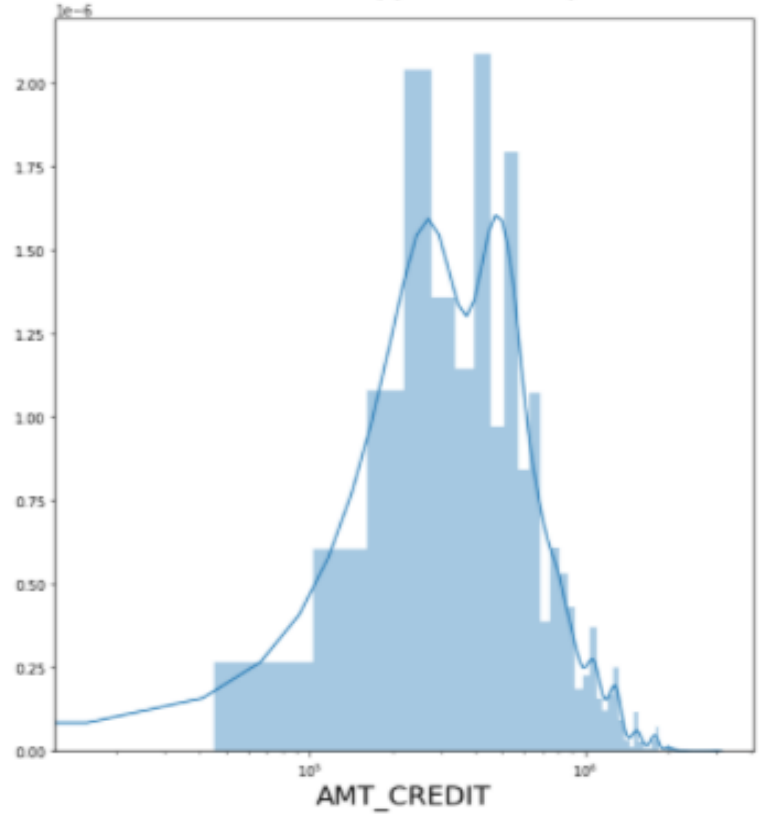


Inferences:

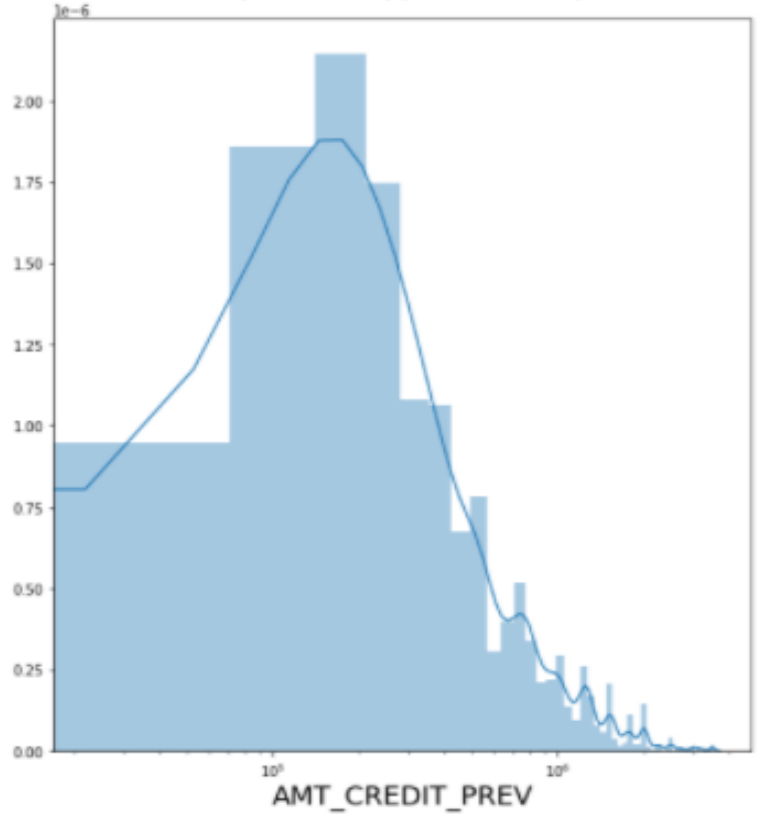
1. General peak is around 25000 for both current and previous application

Univariate Analysis post merging and cleaning of data sets

CREDIT AMOUNT of current application for previous defaulter



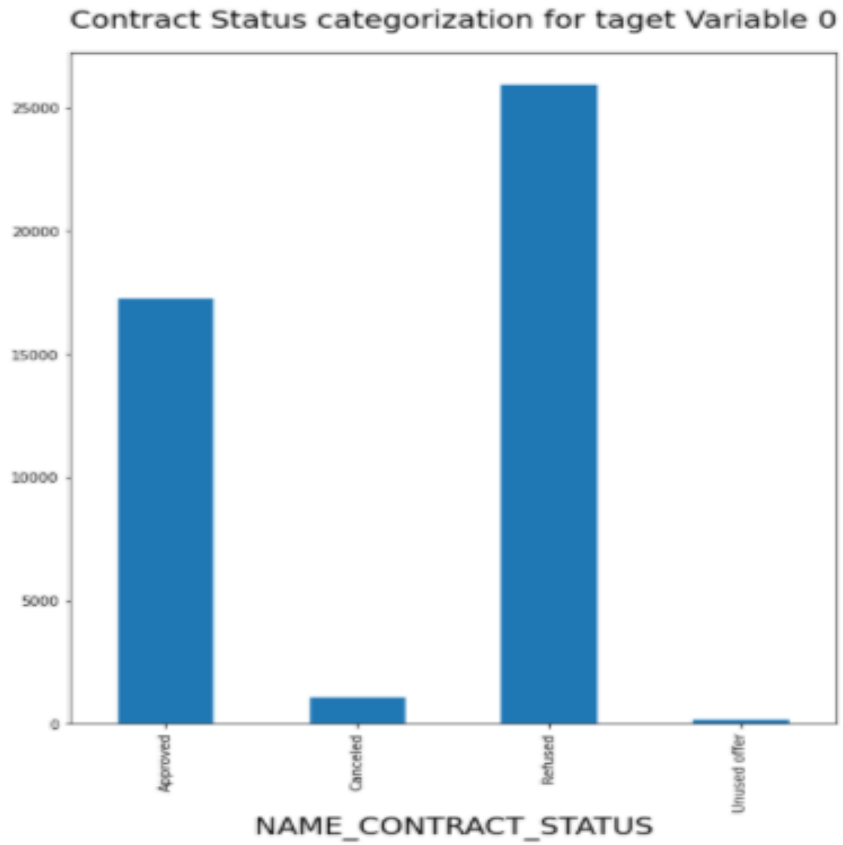
CREDIT AMOUNT of previous application for previous defaulter

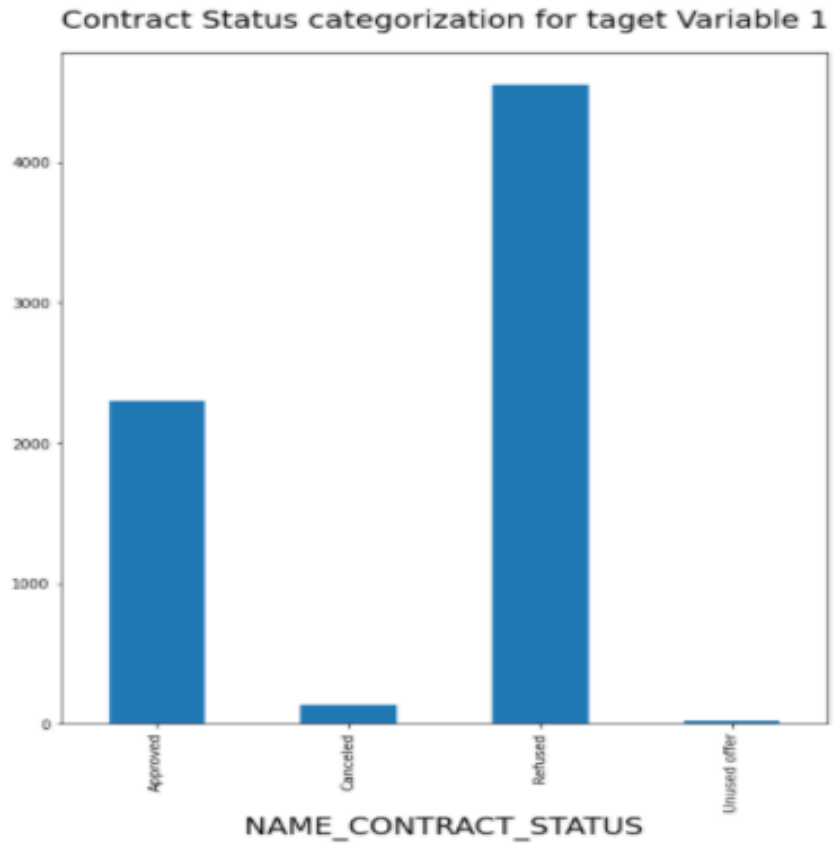


Inferences:

Again the trend remains similar even though their are 2 peaks in current application whereas one is the previous application

Univariate Analysis post merging and cleaning of data sets

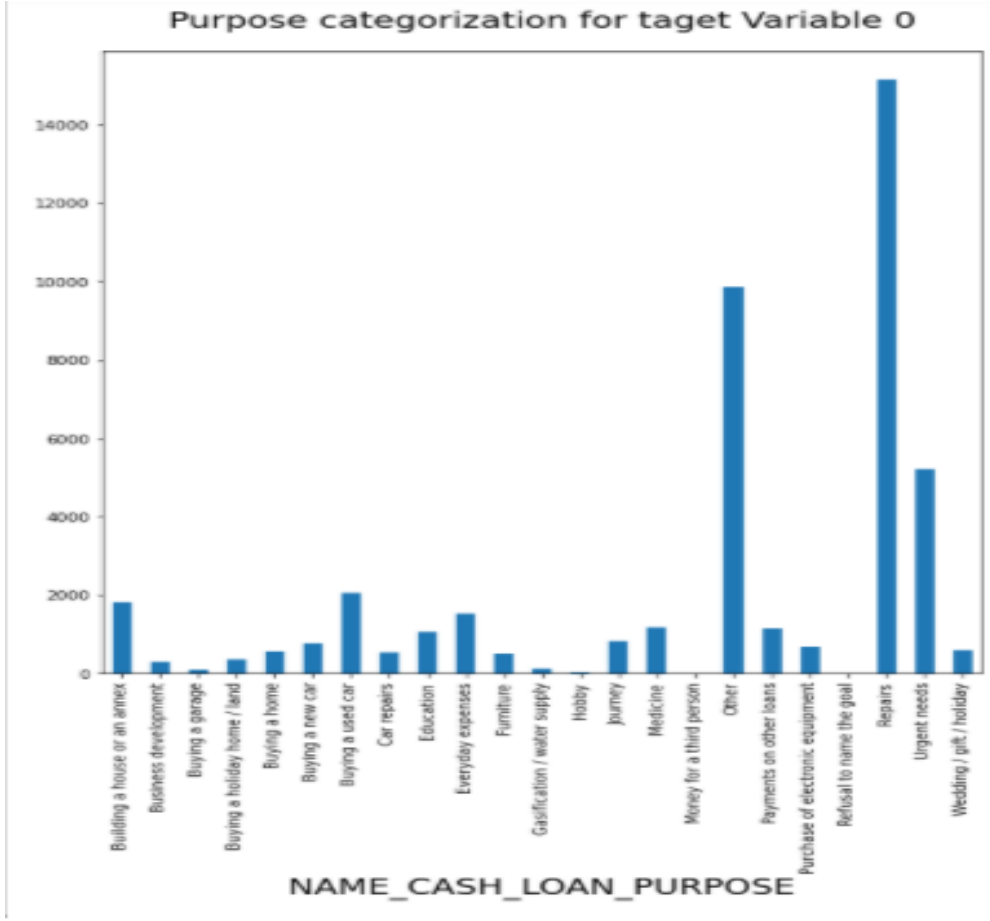


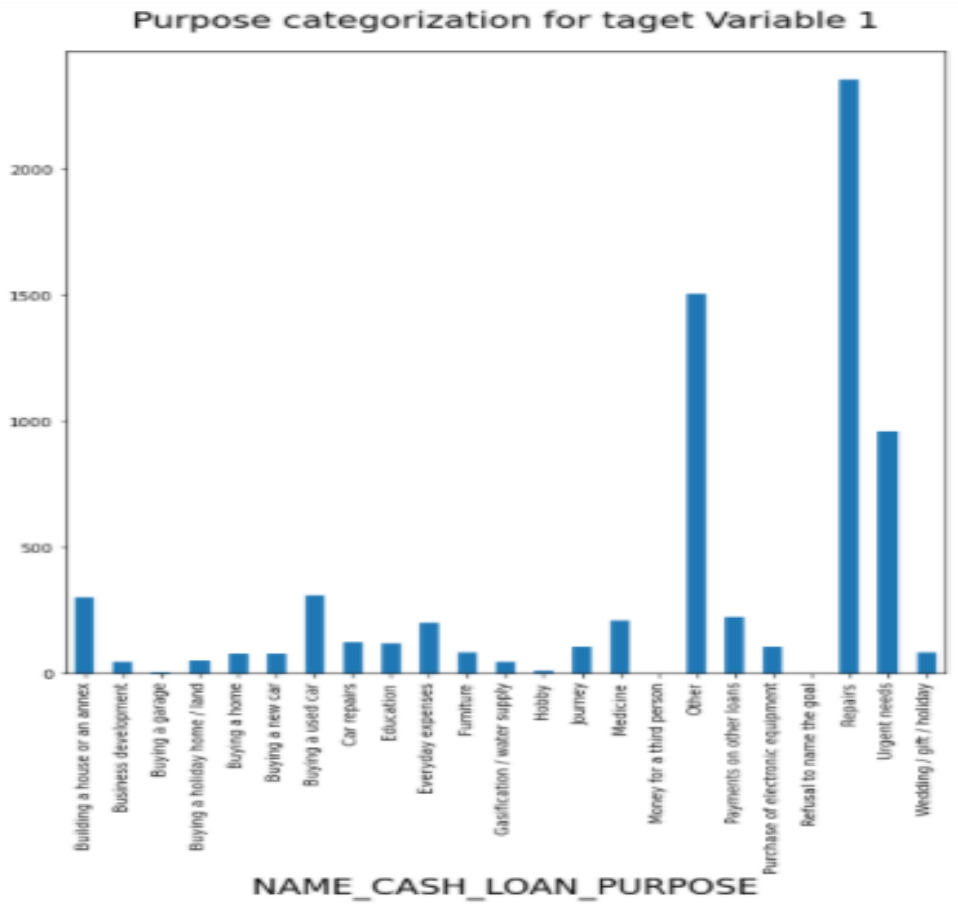


Inferences:

Even though the trend remains the same between defaulters and non defaulters, The refused people reapply for loans again the most and Unused Offers people reapply the least. This makes sense as an applicant might want to improve his credit score and reapply after getting declined against those who have already been offered but they did not use the loan.

Univariate Analysis post merging and cleaning of data sets



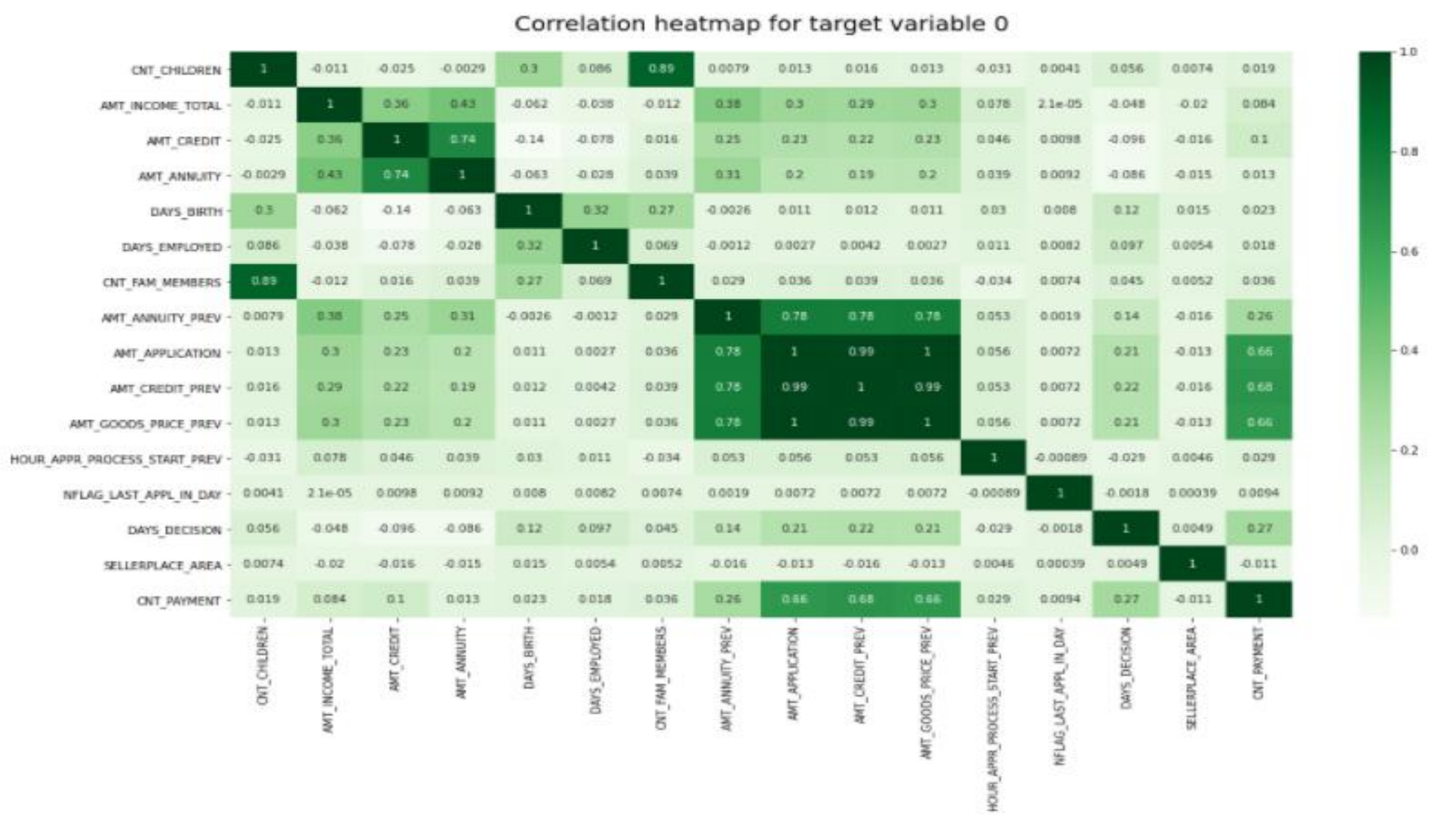


Inferences:

The trend again remains the same for both defaulters and non defaulters but interstingly Purpose for which most loans are taken are Repairs. Ideally we would have expected the highest amount of loans would have been taken for buying homes/cars etc

Bivariate Analysis

- Following the same procedures as we did for the previous data set
- Developing correlation matrix and then plotting heat map on that correlation matrix


Correlation heatmap for target variable 0
CNT_CHILDREN 1 -0.011 -0.025 -0.0029 0.3 0.086 0.89 0.0079 0.013 0.016 0.013 -0.031 0.0041 0.056 0.0074 0.019
AMT_INCOME_TOTAL -0.011 1 0.36 0.43 -0.062 -0.038 -0.012 0.38 0.3 0.29 0.3 0.078 2.1e-05 -0.048 -0.02 0.084
AMT_CREDIT -0.025 0.36 1 0.74 -0.14 -0.078 0.016 0.25 0.23 0.22 0.23 0.046 0.0098 -0.096 -0.016 0.1
AMT_ANNUITY -0.0029 0.43 0.74 1 -0.063 -0.028 0.039 0.31 0.2 0.19 0.2 0.039 0.0092 -0.086 -0.015 0.013
DAYS_BIRTH 0.3 -0.062 -0.14 -0.063 1 0.32 0.27 -0.0026 0.011 0.012 0.011 0.03 0.008 0.12 0.015 0.023
DAYS_EMPLOYED 0.086 -0.038 -0.078 -0.028 0.32 1 0.069 -0.0012 0.0027 0.0042 0.0027 0.011 0.0082 0.097 0.0054 0.018
CNT_FAM_MEMBERS 0.89 -0.012 0.016 0.039 0.27 0.069 1 0.029 0.036 0.039 0.036 -0.034 0.0074 0.045 0.0052 0.036
AMT_AANNUITY_PREV 0.0079 0.38 0.25 0.31 -0.0026 -0.0012 0.029 1 0.78 0.78 0.78 0.053 0.0019 0.14 -0.016 0.26
AMT_APPLICATION 0.013 0.3 0.23 0.2 0.011 0.0027 0.036 0.78 1 0.99 1 0.056 0.0072 0.21 -0.013 0.66
AMT_CREDIT_PREV 0.016 0.29 0.22 0.19 0.012 0.0042 0.039 0.78 0.99 1 0.99 0.053 0.0072 0.22 -0.016 0.68
AMT_GOODS_PRICE_PREV 0.013 0.3 0.23 0.2 0.011 0.0027 0.036 0.78 1 0.99 1 0.056 0.0072 0.21 -0.013 0.66
HOUR_APPR_PROCESS_START_PREV -0.031 0.078 0.046 0.039 0.03 0.011 -0.034 0.053 0.056 0.053 0.056 1 -0.00089 -0.029 0.0046 0.029
NFLAG_LAST_APPL_IN_DAY 0.0041 2.1e-05 0.0098 0.0092 0.008 0.0082 0.0074 0.0019 0.0072 0.0072 0.0072 -0.00089 1 -0.0018 0.00039 0.0094
DAYS_DECISION 0.056 -0.048 -0.096 -0.086 0.12 0.097 0.045 0.14 0.21 0.22 0.21 -0.029 -0.0018 1 0.0049 0.27
SELLERPLACE_AREA 0.0074 -0.02 -0.016 -0.015 0.015 0.0054 0.0052 -0.016 -0.013 -0.016 -0.013 0.0046 0.00039 0.0049 1 -0.011
CNT_PAYMENT 0.019 0.084 0.1 0.013 0.023 0.018 0.036 0.26 0.66 0.68 0.66 0.029 0.0094 0.27 -0.011 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
DAYS_BIRTH
DAYS_EMPLOYED
CNT_FAM_MEMBERS
AMT_ANNUITY_PREV
AMT_APPLICATION
AMT_CREDIT_PREV
AMT_GOODS_PRICE_PREV
HOUR_APPR_PROCESS_START_PREV
NFLAG_LAST_APPL_IN_DAY
DAYS_DECISION
SELLERPLACE_AREA
CNT_PAYMENT
1.0
0.8
0.6
0.4
0.2
0.0

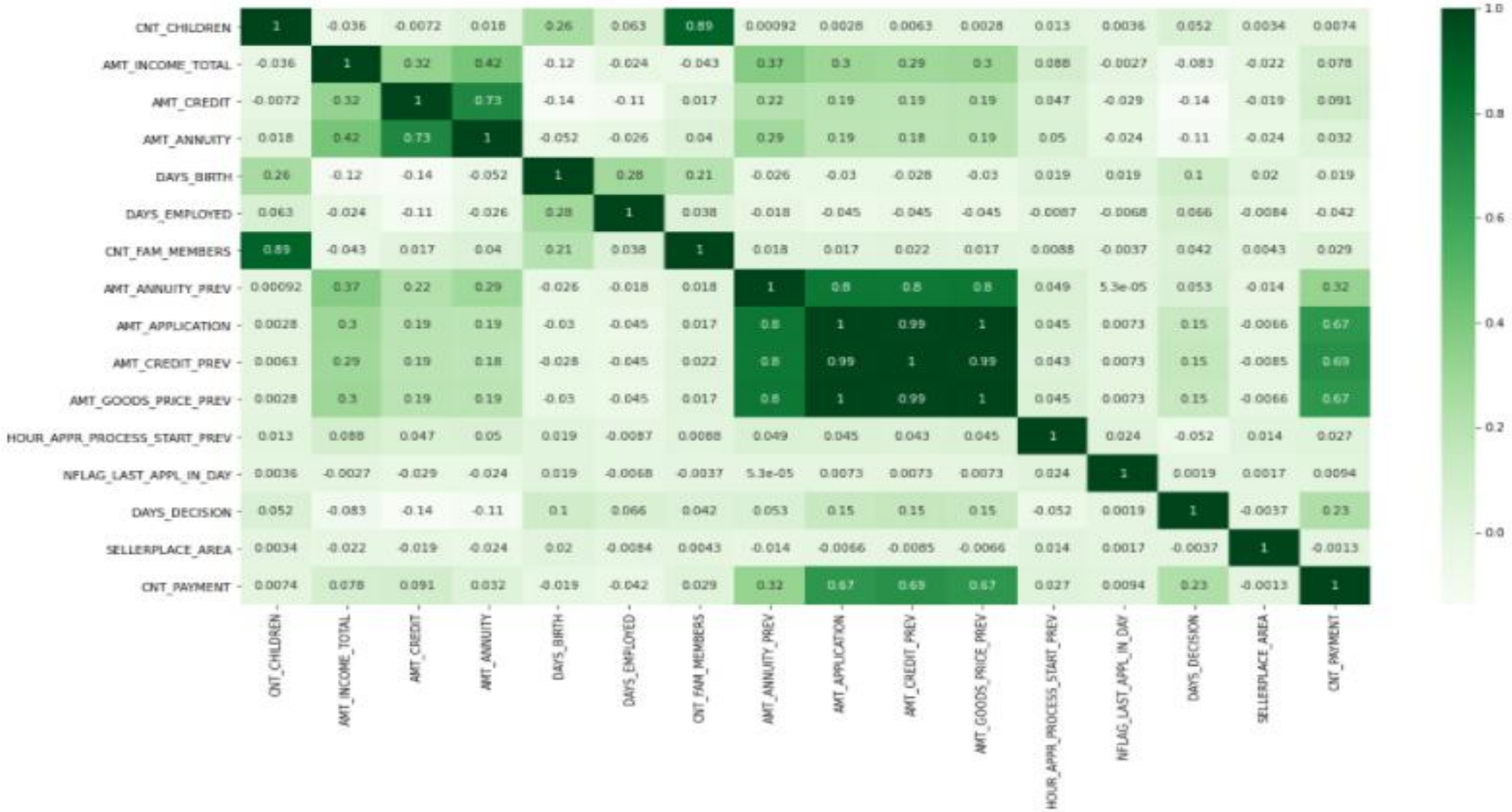
Inferences:

Highest correlated variables include:

1. CNT_CHILDREN AND CNT_FAM_MEMBERS
2. AMT_GOODS_PRICE_PREV and AMT_CREDIT_PREV
3. AMT_CREDIT_PREV and AMT_APPLICATION
4. AMT_GOODS_PRICE_PREV and AMT_CREDIT_PREV
5. AMT_GOODS_PRICE_PREV and AMT_ANNUITY_PREV
6. AMT_ANNUITY_PREV and AMT_APPLICATION
7. AMT_GOODS_PRICE_PREV and AMT_ANNUITY_PREV
8. AMT_GOODS_PRICE_PREV and CNT_PAYMENT
9. CNT_PAYMENT and AMT_APPLICATION
10. AMT_GOODS_PRICE_PREV and CNT_PAYMENT

Similar to what is seen in the current application, income, credit, goods price are all highly positively correlated

Correlation heatmap for target variable 1



Inferences:

Highest correlated variables include:

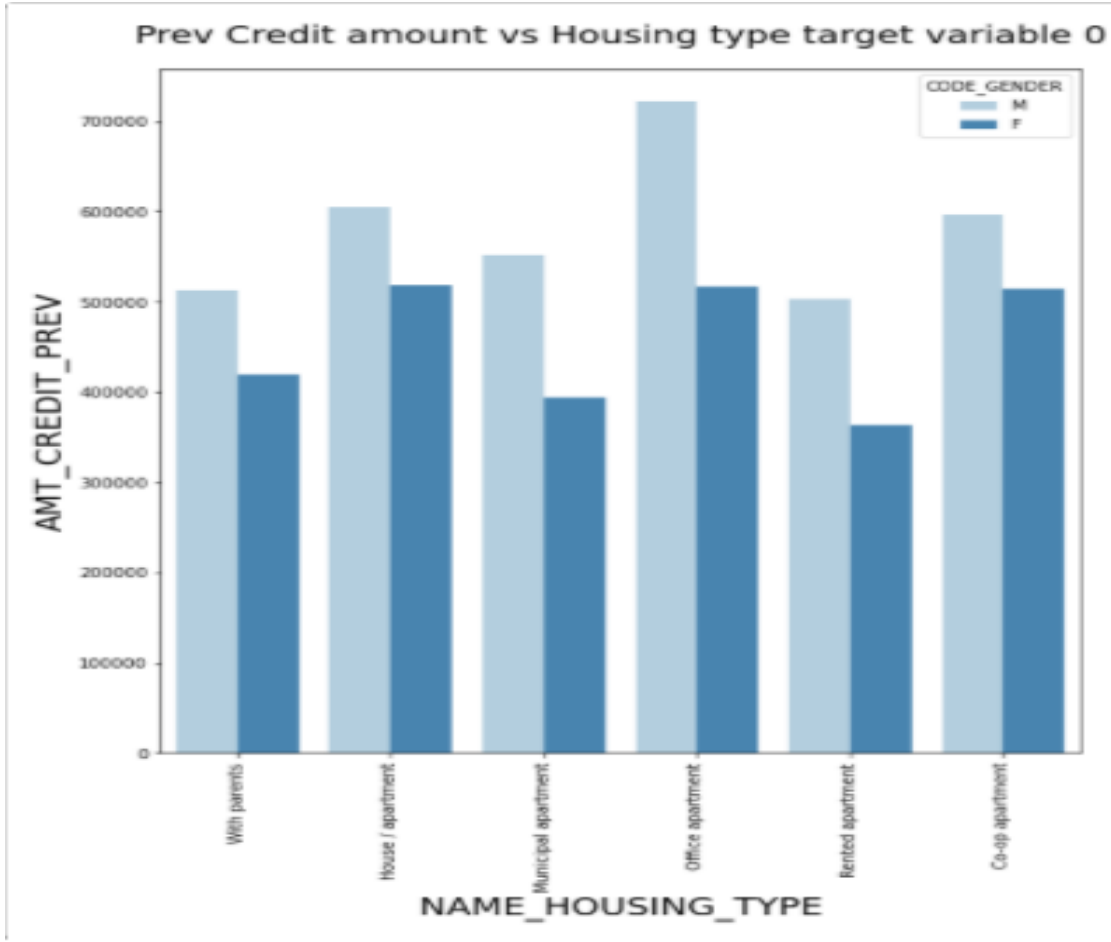
1. CNT_CHILDREN AND CNT_FAM_MEMBERS
2. AMT_GOODS_PRICE_PREV and AMT_CREDIT_PREV
3. AMT_CREDIT_PREV and AMT_APPLICATION
4. AMT_GOODS_PRICE_PREV and AMT_CREDIT_PREV
5. AMT_GOODS_PRICE_PREV and AMT_ANNUITY_PREV
6. AMT_ANNUITY_PREV and AMT_APPLICATION
7. AMT_GOODS_PRICE_PREV and AMT_ANNUITY_PREV
8. AMT_GOODS_PRICE_PREV and CNT_PAYMENT
9. CNT_PAYMENT and AMT_APPLICATION
10. AMT_GOODS_PRICE_PREV and CNT_PAYMENT

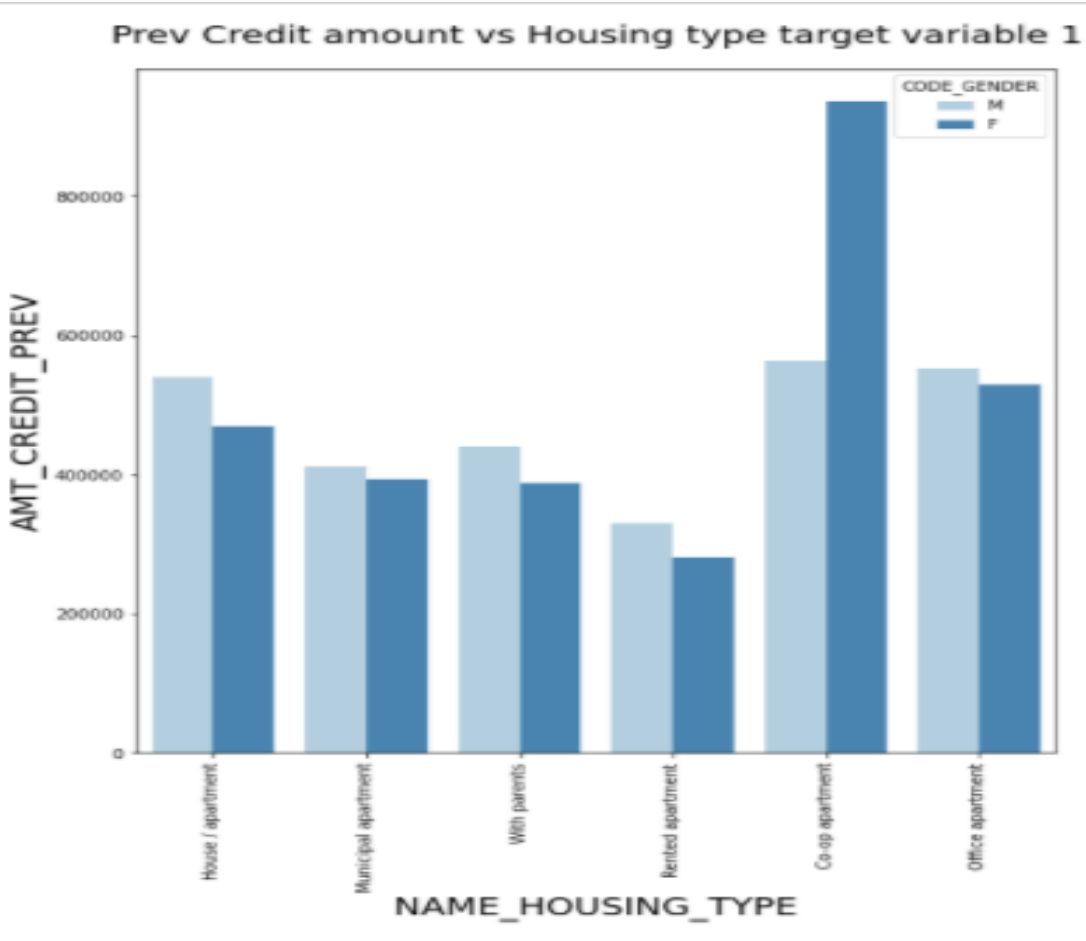
The trend remains same as targetr variable 0

Inferences:

1. The credit amount of Loan purposes like 'Buying a home','Buying a land','Buying a new car' 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.

Housing Type Analysis based on Gender





Inferences:

1. Males living in office apartment have high credit amounts
2. Males overall regardless of the housing type have higher credit amounts than females
3. Females with rented apartment have the lowest credit amounts

Recommendations after joining dataset

- Here for Housing type, office apartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1. So, 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 Recommendations

- So after all these analysis and completing of EDA there a few points which can be kept in mind for approving/disapproving the loan by the bank.
- People with housing type as With parents should be given loans easily as they are having lowest chance of defaulting. Females with co op housing should be refrained from giving loans as they have highest default rate.
- Banks should focus more on contract type Student, Businessmen and Pentioners with housing type other than 'Co-op apartment' for successful repayment of loans
- Loan purpose Repair is having higher number of unsuccessful payments on time so that reason should be approved less
- Also people who have been refused before are more prone to be defaulters