CREDIT EDA CASE STUDY

Rama Mishra
Nidhi Sharma

CREDIT EDA CASE STUDY

Introduction:-

In this case study, after applying EDA(Exploratory Data Analysis) :-

- ✓ We need to have basic understanding of risk analytics in banking and financial services.
- ✓ Understand how data is used to minimize the risk of losing money while lending to customers.

CREDIT EDA CASE STUDY

Business Understanding:-

You work for a consumer finance company which specialises in lending various types of loans to urban customers.

- ✓ When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

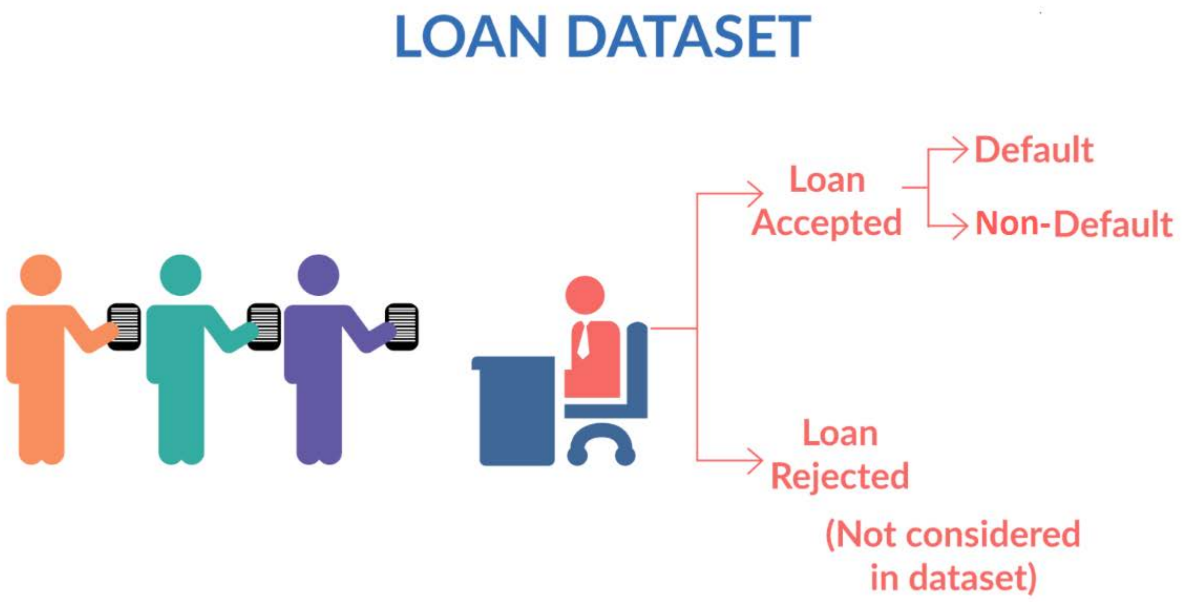
- ✓ Two types of risks are associated with the bank's decision:

  - ➢ **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

CREDIT EDA CASE STUDY

The loan data contains the information about:-

- ✓ Past loan applicants and whether they ‘defaulted’ or not.
- ✓ The aim is to identify patterns which indicate if a person is likely to default, 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.
- ✓ In this case study, Using EDA to understand how consumer attributes and loan attributes influence the tendency of default.



Steps involved are :-

1. After importing and checking the structure of the 'application‘ data.
2. Dealing with the Quality Check and Missing values of the data.
3. 3.1) Finding the percentage of missing values of all the columns.

```
ENTRANCES_MODE                      50.35
FLOORSMAX_MODE                      49.76
FLOORSMIN_MODE                      67.85
LANDAREA_MODE                       59.38
LIVINGAPARTMENTS_MODE               68.35
LIVINGAREA_MODE                     50.19
NONLIVINGAPARTMENTS_MODE            69.43
NONLIVINGAREA_MODE                  55.18
APARTMENTS_MEDI                     50.75
BASEMENTAREA_MEDI                   58.52
YEARS_BEGINEXPLUATATION_MEDI        48.78
YEARS_BUILD_MEDI                    66.50
COMMONAREA_MEDI                     69.87
ELEVATORS_MEDI                      53.30
ENTRANCES_MEDI                      50.35
FLOORSMAX_MEDI                      49.76
FLOORSMIN_MEDI                      67.85
LANDAREA_MEDI                       59.38
LIVINGAPARTMENTS_MEDI               68.35
LIVINGAREA_MEDI                     50.19
NONLIVINGAPARTMENTS_MEDI            69.43
NONLIVINGAREA_MEDI                  55.18
FONDKAPREMONT_MODE                  68.39
HOUSETYPE_MODE                      50.18
TOTALAREA_MODE                      48.27
WALLSMATERIAL_MODE                  50.84
EMERGENCYSTATE_MODE                 47.40
OBS_30_CNT_SOCIAL_CIRCLE             0.33
DEF_30_CNT_SOCIAL_CIRCLE             0.33
OBS_60_CNT_SOCIAL_CIRCLE             0.33
DEF_60_CNT_SOCIAL_CIRCLE             0.33
DAYS_LAST_PHONE_CHANGE               0.00
```

3.2) Removing columns with high missing percentage(i.e more than 60%).

Subtask 3.2: Removing columns with high missing percentage(i.e more than 60%).

```
# Removing columns with high missing percentage(i.e more than 60%).
application_data= application_data.drop('OWN_CAR_AGE', axis=1)
application_data= application_data.drop('YEARS_BUILD_AVG', axis=1)
application_data= application_data.drop('COMMONAREA_AVG', axis=1)
application_data= application_data.drop('FLOORSMIN_AVG', axis=1)
application_data= application_data.drop('LIVINGAPARTMENTS_AVG', axis=1)
application_data= application_data.drop('NONLIVINGAPARTMENTS_AVG', axis=1)
application_data= application_data.drop('YEARS_BUILD_MODE', axis=1)
application_data= application_data.drop('COMMONAREA_MODE', axis=1)
application_data= application_data.drop('FLOORSMIN_MODE', axis=1)
application_data= application_data.drop('LIVINGAPARTMENTS_MODE', axis=1)
application_data= application_data.drop('NONLIVINGAPARTMENTS_MODE', axis=1)
application_data= application_data.drop('YEARS_BUILD_MEDI', axis=1)
application_data= application_data.drop('COMMONAREA_MEDI', axis=1)
application_data= application_data.drop('FLOORSMIN_MEDI', axis=1)
application_data= application_data.drop('LIVINGAPARTMENTS_MEDI', axis=1)
application_data= application_data.drop('NONLIVINGAPARTMENTS_MEDI', axis=1)
application_data= application_data.drop('FONDKAPREMONT_MODE', axis=1)

application_data
```

Subtask 3.3: Checking the best metric to imput the missing values for columns with less missing percentage(i.e less than 13%) (for 5 columns)

We are selecting following 5 columns which has less percentage(around 13%) of missing values to impute them,

1) 'NAME_TYPE_SUITE'
2) 'AMT_GOODS_PRICE'
3) 'EXT_SOURCE_2'
4) 'OBS_60_CNT_SOCIAL_CIRCLE'
5) 'OBS_30_CNT_SOCIAL_CIRCLE'

1) The most common value(**mode**) of **'NAME_TYPE_SUITE'** is 'Unaccompanied' (dtype is int), so we have impute the NaNs by that.

2) We have observed that there is not so much difference between mean and median values of **'EXT_SOURCE_2'** and for **'AMT_GOODS_PRICE'** the difference between mode and median values is not so much. Thus, we will impute the missing values of 'EXT_SOURCE_2' and 'AMT_GOODS_PRICE' by the **median**(the central value).

3) We have observed that there is an outlier '348' and '344' respectively in both the columns **'OBS_30_CNT_SOCIAL_CIRCLE**', **'OBS_60_CNT_SOCIAL_CIRCLE'** respectively. Median and mode values of both columns are same only. Thus, I will impute the missing values of both by the **mode**(the most occurring value).¶

Subtask 3.4: Checking the datatypes of all columns and changing the datatype if required.

The columns 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'DAYS_LAST_PHONE_CHANGE' contains days count, so it cannot be of negative value. Hence, replacing it with its absolute values.

In [43]:
```
# making columns as a string
application_data['DAYS_BIRTH']= abs(application_data['DAYS_BIRTH'])
application_data['DAYS_EMPLOYED']= abs(application_data['DAYS_EMPLOYED'])
application_data['DAYS_REGISTRATION']= abs(application_data['DAYS_REGISTRATION'])
application_data['DAYS_ID_PUBLISH']= abs(application_data['DAYS_ID_PUBLISH'])
application_data['DAYS_LAST_PHONE_CHANGE']= abs(application_data['DAYS_LAST_PHONE_CHANGE'])
```

In [44]:
```
# checking the datatypes
d_type= application_data.dtypes
d_type[17:21]
```

Out[44]:
```
DAYS_BIRTH             int64
DAYS_EMPLOYED          int64
DAYS_REGISTRATION    float64
DAYS_ID_PUBLISH        int64
dtype: object
```

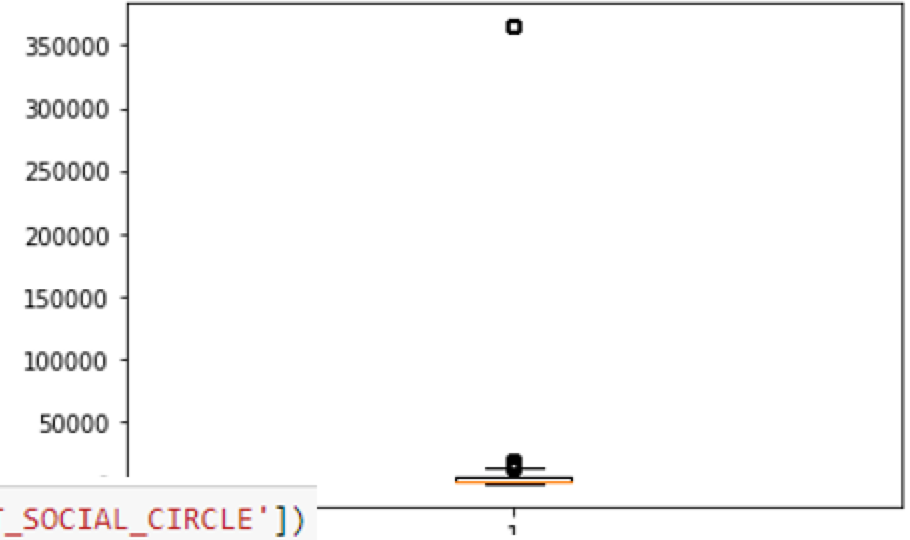
In [45]:
```
# checking whether negative sign is replaced or not.
application_data[['DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'DAYS_LAST_PHONE_CHANGE']]
```

Out[45]:

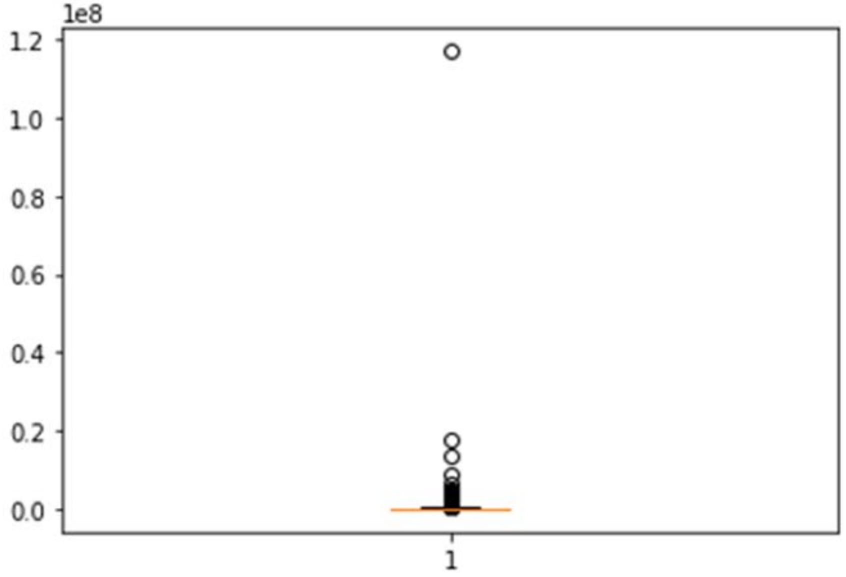
	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
0	9461	637	3648.0	2120	1134.0
1	16765	1188	1186.0	291	828.0
2	19046	225	4260.0	2531	815.0
3	19005	3039	9833.0	2437	617.0
4	19932	3038	4311.0	3458	1106.0

Subtask 3.5: For numerical columns checking for the outliers and reporting them for at least 3 variables.
Treating them and analysing it.
We are choosing the below variables for outlier treatment.
1) 'DAYS_EMPLOYED'
2)'AMT_INCOME_TOTAL'
3) 'OBS_30_CNT_SOCIAL_CIRCLE'¶

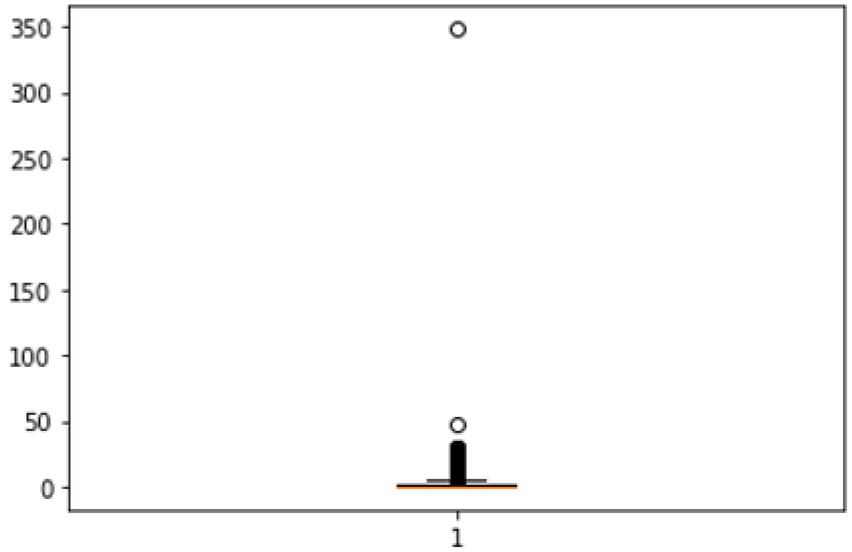
```
plt.boxplot(application_data['DAYS_EMPLOYED'])
plt.show()
```




```
plt.boxplot(application_data['AMT_INCOME_TOTAL'])
#plt.yscale('log')
plt.show()
```




```
plt.boxplot(application_data['OBS_30_CNT_SOCIAL_CIRCLE'])
plt.show()
```



1) **DAYS_EMPLOYED** has a maximum value of 365243 which does not make sense as this corresponds to more than 1000 years (i.e., no one can be employed for a 1000 years) and around 18% (55374) rows are having this values. so we cant drop it. So **iam replacing it with NAN.**

2) **'AMT_INCOME_TOTAL**' i.e income of the client has a maximum value of 117000000 and it is greater than (q3+1.5*iqr) so it act as an outlier which i have confirmed from IQR method. **So iam going to deal with this variable by capping it.**

3) The maximum value in **'OBS_30_CNT_SOCIAL_CIRCLE'** is 348.0 and only one row is having 348.0 value. So iam going to **deal with this variable by dropping it.**

Subtask 3.6: Binning of continuous variables, checking if we need to bin any variable in different categories (doing this for 1 or two columns).

we have done binning for one column 'DAYS_BIRTH'

```
# Since 'DAYS_BIRTH' are in days so iam dividing it by 365 days to convert it into years and storing it in a new column so that r
application_data['YEARS_BIRTH']= application_data['DAYS_BIRTH'].apply(lambda x: "{}".format((x/365)))
```

```
# checking the column
application_data['YEARS_BIRTH']
```

```
# Changing the datatype of the column
application_data['YEARS_BIRTH']= application_data['YEARS_BIRTH'].astype('float')
```

```
# checking the datatype of column
application_data['YEARS_BIRTH']
```

```
# doing the round off the column
application_data.YEARS_BIRTH= round(application_data.YEARS_BIRTH)
```

```
#Checking the column
application_data['YEARS_BIRTH']
```

```
# Binning the 'DAYS_BIRTH' and forming a new column 'age_group'
application_data['YEARS_BIRTH']= pd.cut(application_data.YEARS_BIRTH, [0, 30, 40, 50, 60, 9999], labels=['<30', '30-40', '40-50',
```

```
application_data['YEARS_BIRTH']
```

Task 4: Analysis

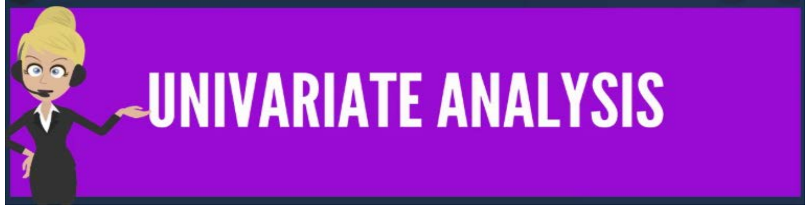
Subtask 4.1: Checking the imbalance percentage.

```
In [ ]: #Imbalance percentage
        application_data['TARGET'].value_counts(normalize=True)
```

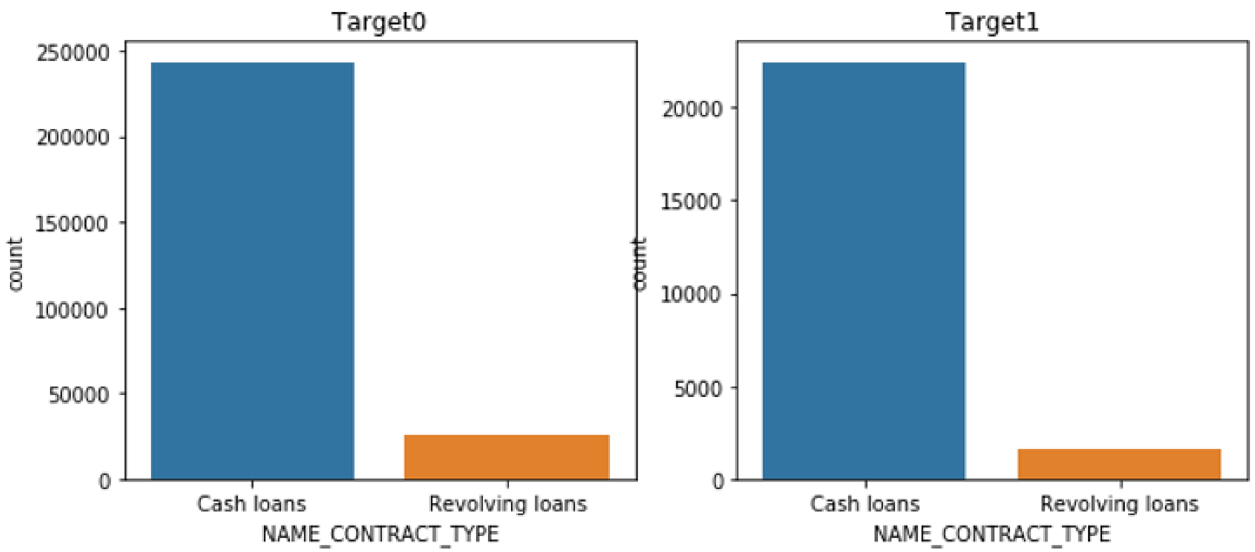
Subtask 4.2: Dividing the data into two sets, i.e Target=1 and Target=0.

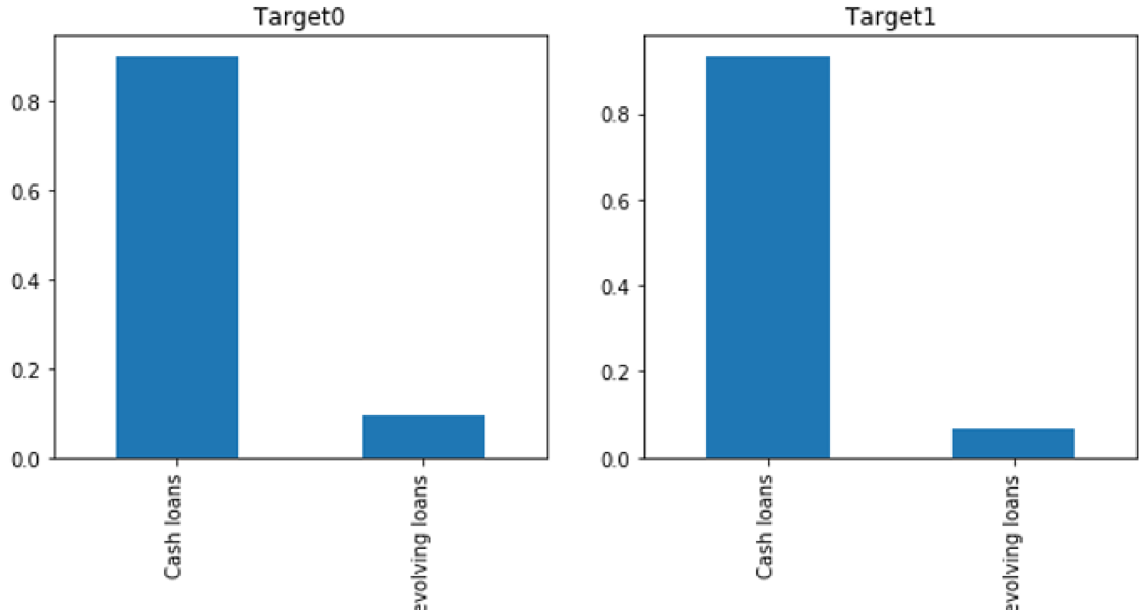
```
In [ ]: Target1= application_data[application_data['TARGET']== 1]
        Target1
```

```
In [ ]: Target0= application_data[application_data['TARGET']== 0]
        Target0
```



Comparing the target variable across categories of categorical variables.





As per the graphical representation, within Target1(client with payment difficulties) and Target0 number of people preferred cash loans is higher than revolving loans.

There is no great difference between the pattern of target0 and target1

As per the graphical representation,

1) Target0(client with payment difficulties): The percentage of Cash loans is 90.33% and Revolving loans is 9.67%

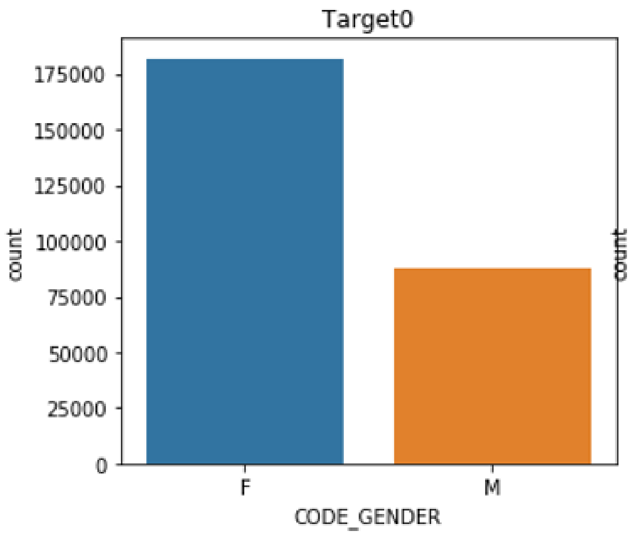
2) Target1 : The percentage of Cash loans is 93.48% and Revolving loans is 6.52%

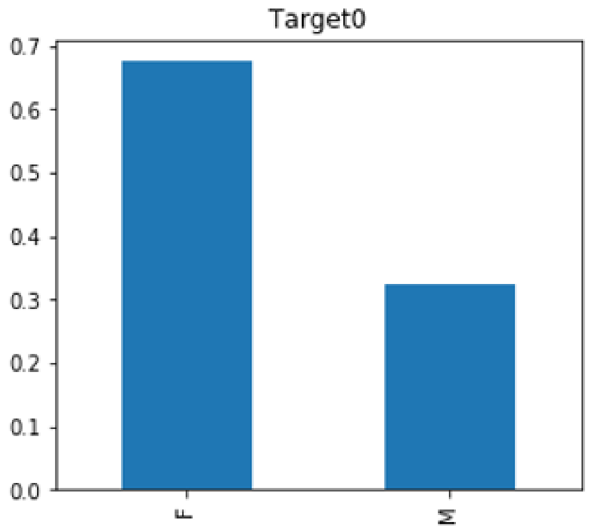
Hence number of people preferred cash loans is higher than revolving loans.

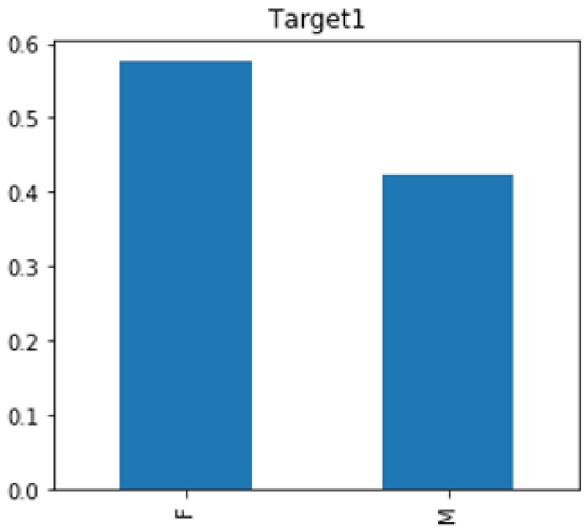
CODE_GENDER

As per below representation, within Target1(client with payment difficulties) and Target0 count of females are more as compare to males.

As per the graphical representation,
1) Target0(client with payment difficulties): The percentage of Females is 67.53% and Males is 32.47%
2) Target1 : The percentage of Females is 57.62% and Males is 42.38%
Hence count of females is higher than Males.

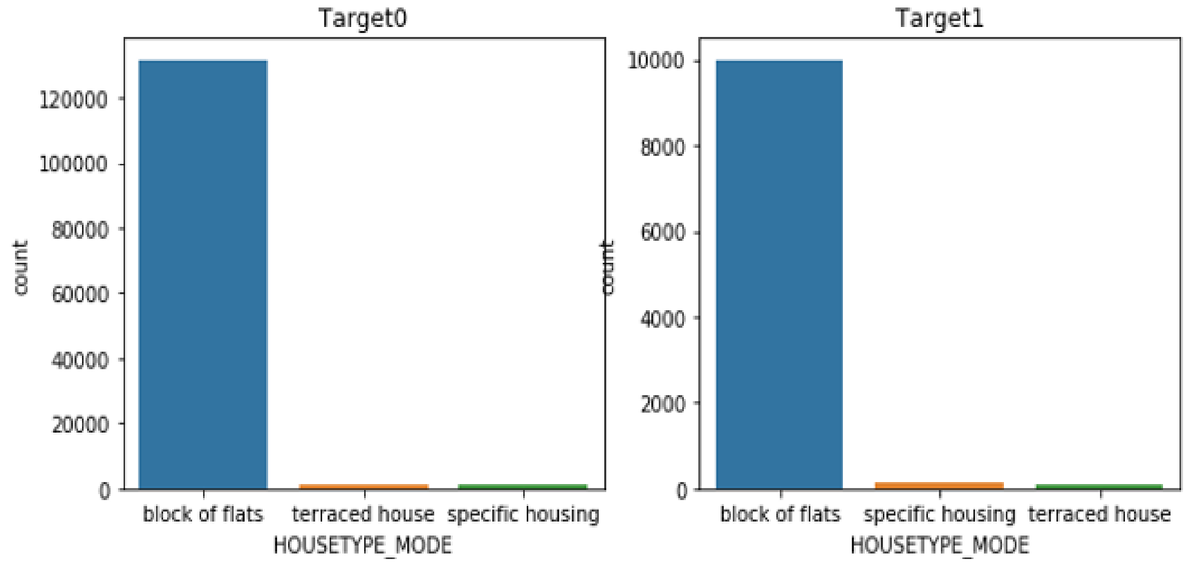


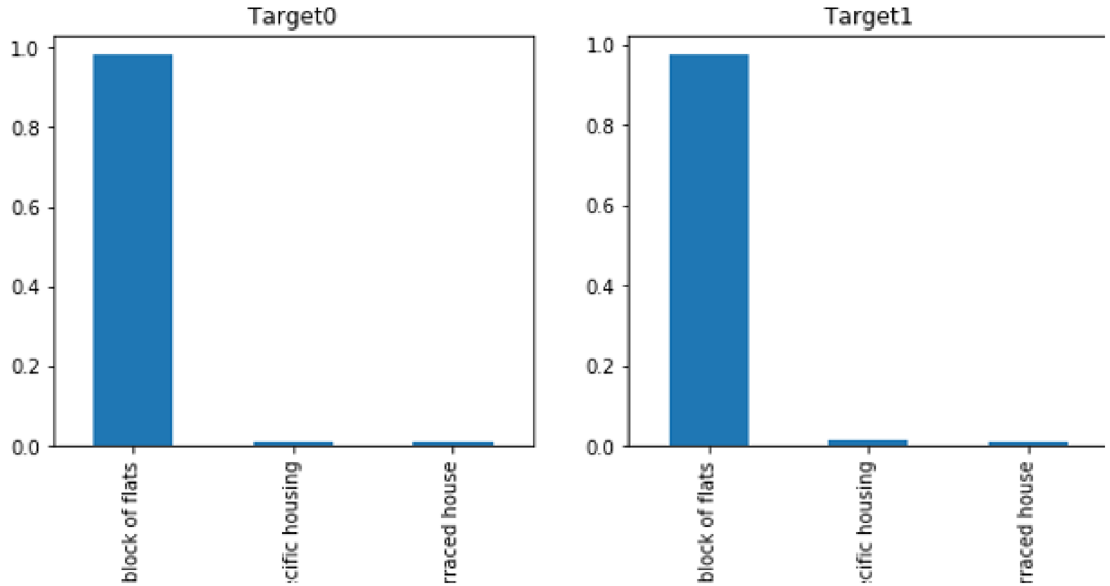




In target0 the females rate are double of males but in target1 there is only 15.24% difference between rates of females and males

HOUSE TYPE_MODE





We found that the number of Clients staying:-

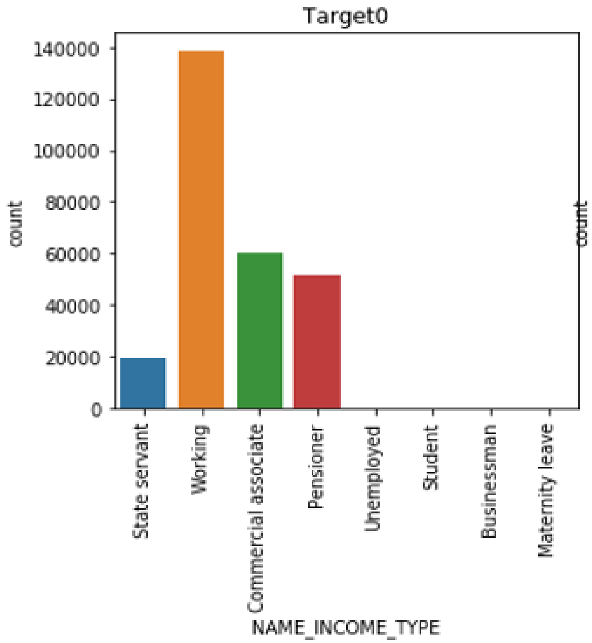
- In block of flats are higher as in both cases of Target 0 & Target1.
- Block of Flats % is higher in both the cases.

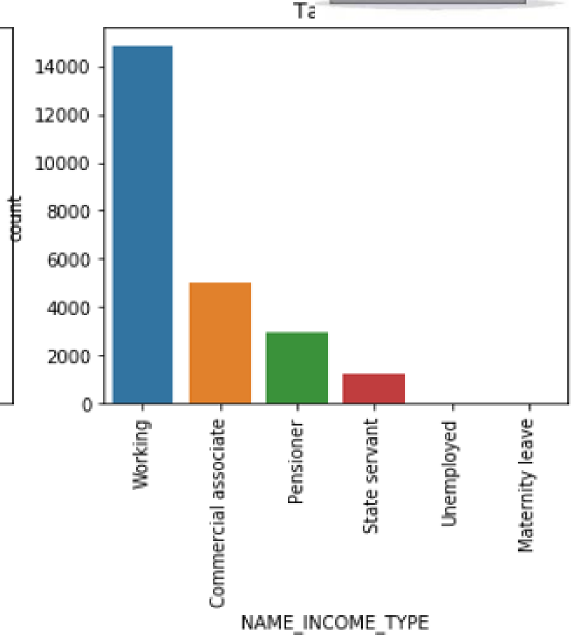
There is no great difference between the pattern of target0 and target1

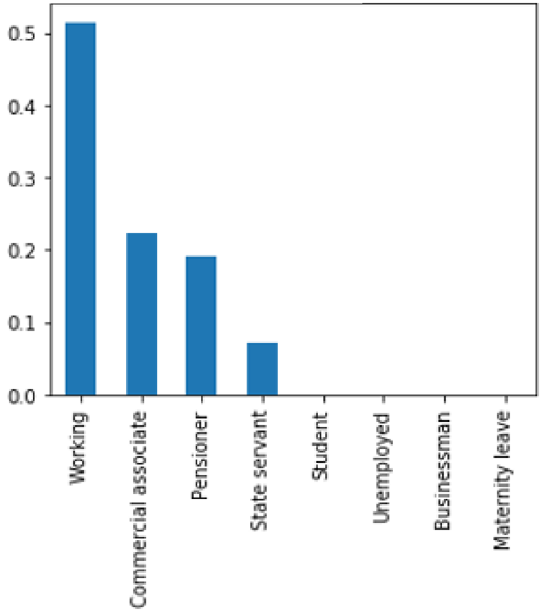
As per above representation derivation mention below:-

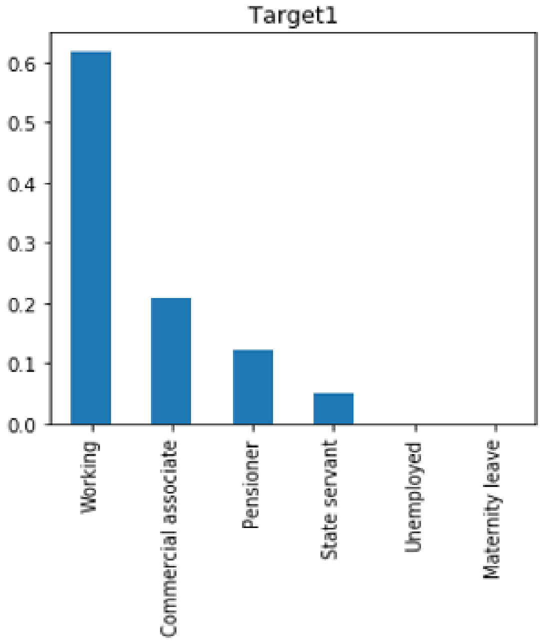
- Target 0(Client with payment difficulties) - % against block of flats is 98.20%, specific housing and terraced house is 1.0% & 0.9% respectively.
- Target1 :-- % against block of flats is 97.60%, specific housing and terraced house is 1.5% & 0.9% respectively.

INCOME_TYPE





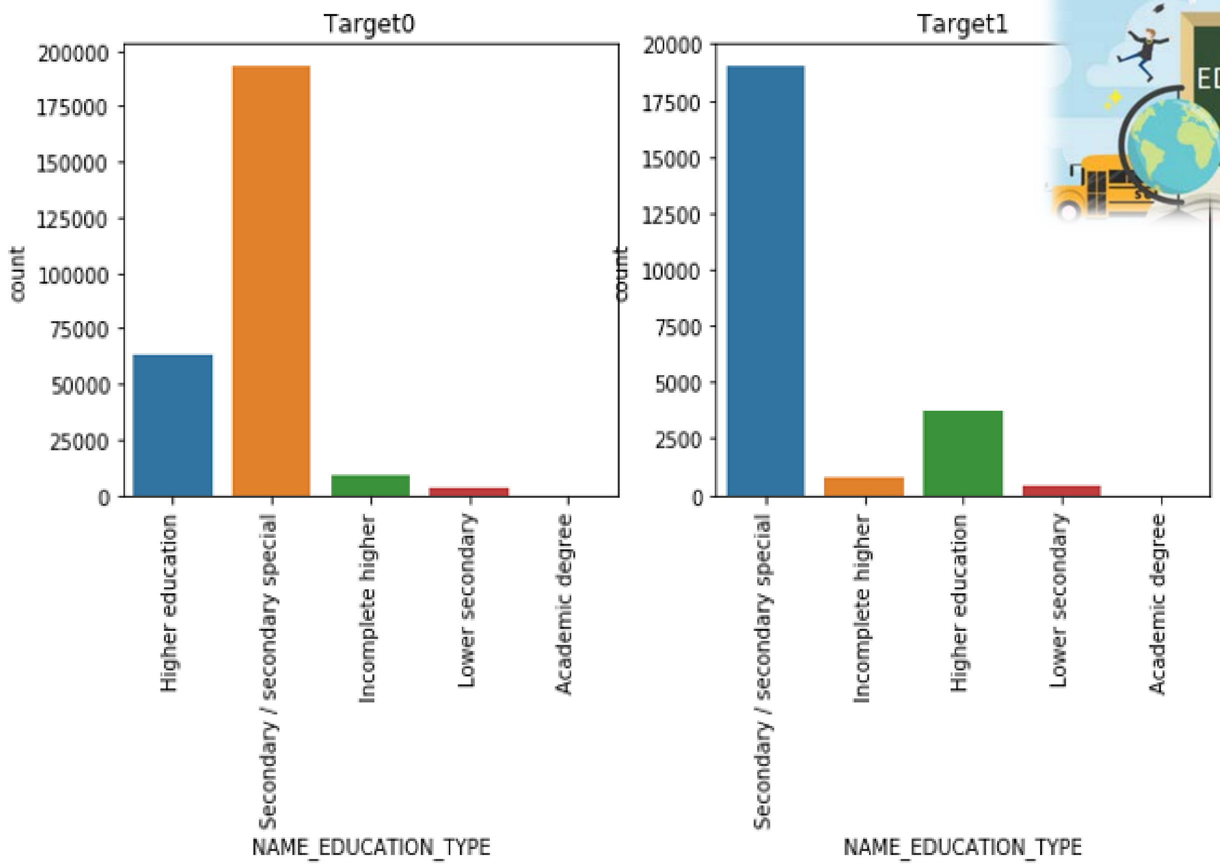


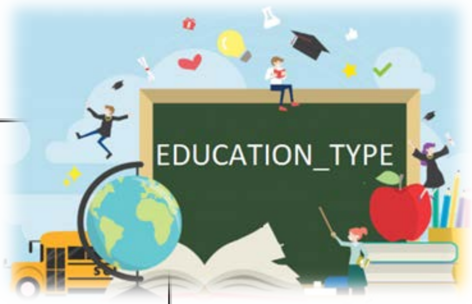


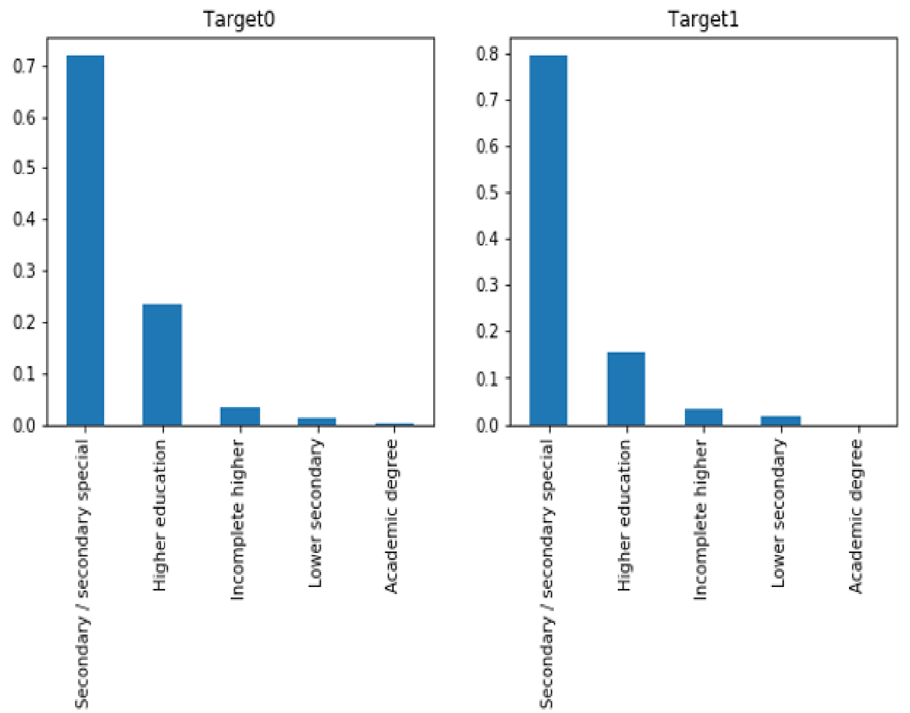
As per the graphical representation, within Target1(client with payment difficulties) and Target0 shows more number of working income type.

In Target0 the percentage of unemployed and maternity leave is less as compare to unemployed and maternity leave percentage in Target1. In target1 there are no student and businessman variables¶and hence there is less risk for them.

As per the graphical representation,
1) Target0(client with payment difficulties): The percentage of Working is 51.47%, commercial associate is 22.31% and pensioner is 19.05%
2) Target1 : The percentage of Working is 61.88%, commercial associate is 20.86% and pensioner is 12.18%
Hence number of people preferred cash loans is higher than revolving loans.







As per the graphical representation, within Target1(client with payment difficulties) and Target0 shows people's highest education level is from secondary / secondary special.

In Target0 the percentage of Higher education and Academic degree is quite more than in target1

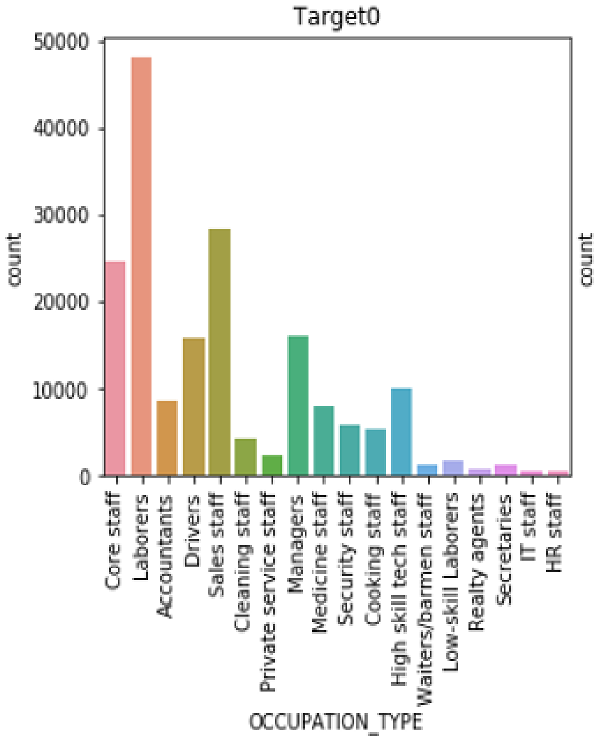
As per the graphical representation,

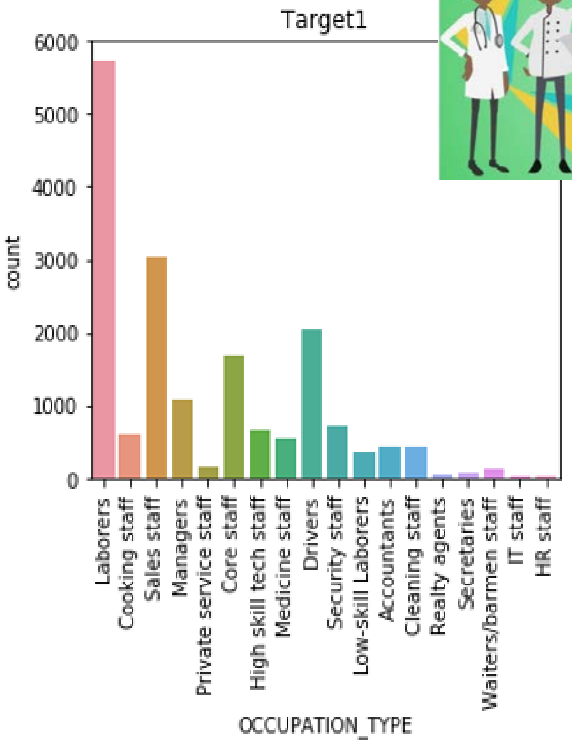
1) Target0(client with payment difficulties): The percentage of Secondaryeducation is 71.81%, higher education is 23.60% and incomplete higher education is 3.3%

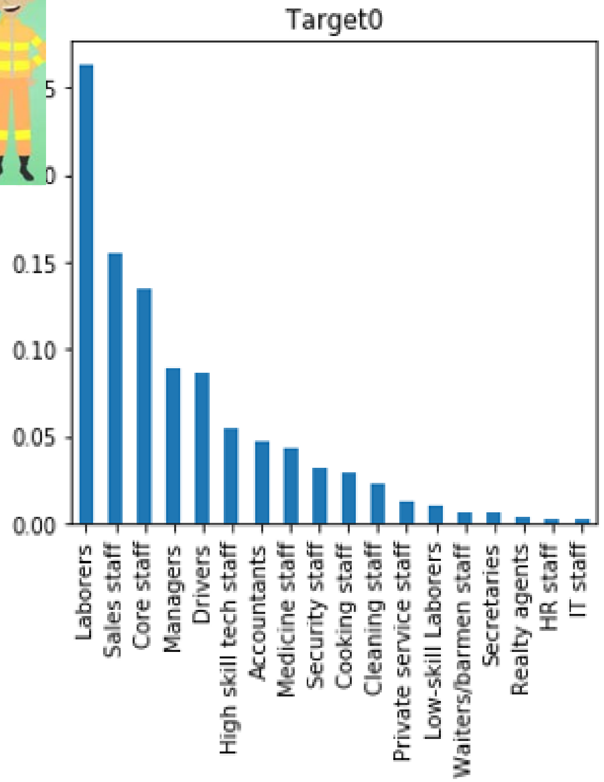
2) Target1 : The percentage of Secondaryeducation is 79.42%, higher education is 15.41% and incomplete higher education is3.4 %

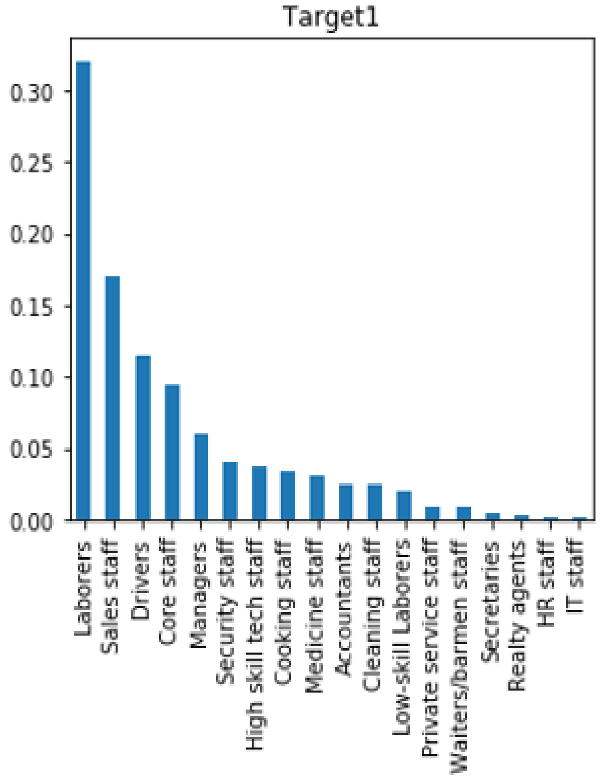
Hence people's highest education level is from secondary / secondary special.

OCCUPATION_TYPE









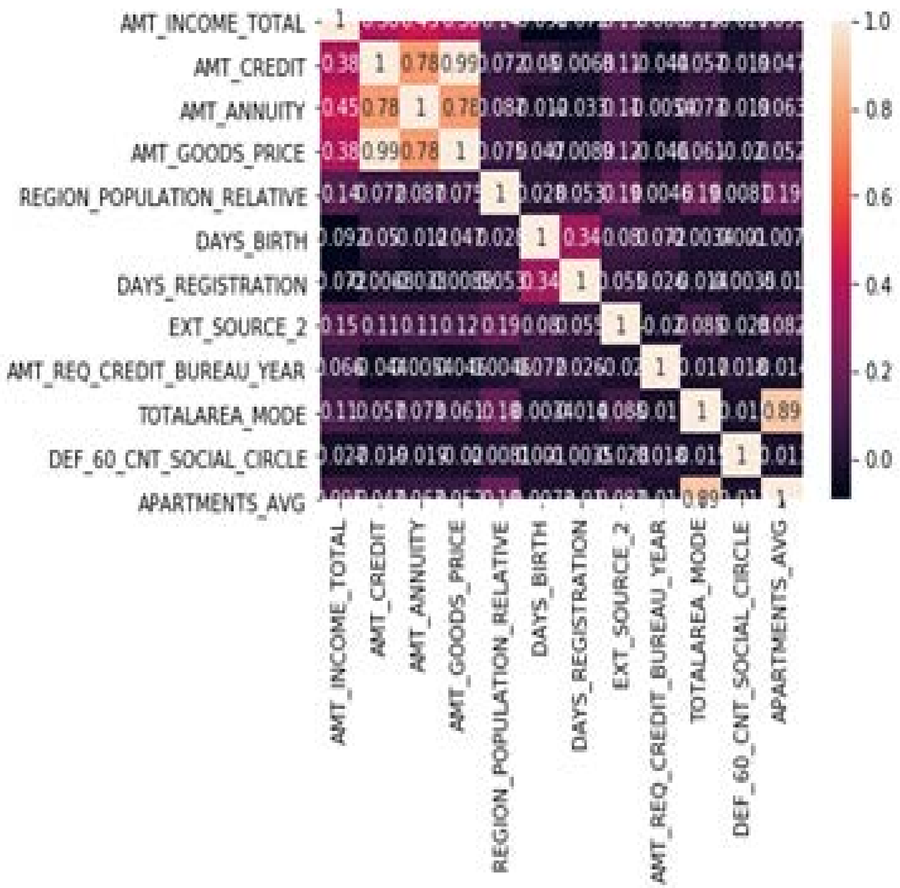
As per the graphical representation, within Target1 (client with payment difficulties) and Target0 shows that type of occupation is highest i.e laborers.

In target0 Accountants percentage is double than the Percentage of Accountants in Tareget1. In Target1 the percentage of drivers are more than Core staff, Managers

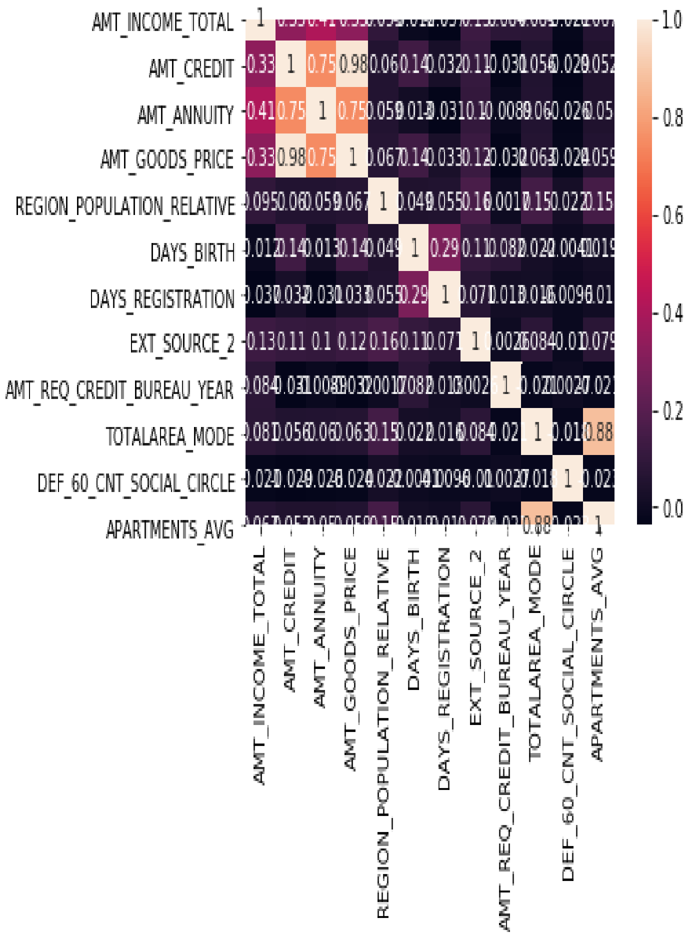
As per the graphical representation,
1) Target0(client with payment difficulties): The percentage of laborers is 26.34%, sales staff is 15.49% and drivers is 8.62%
2) Target1 : The percentage of laborers is 32.0%, sales staff is 16.93% and drivers is 11.37%
Hence laborers is the highest type of occupation

Subtask 4.4: Finding correlation for numerical columns for both the cases, i.e. 0 and 1.

```
In [111]: sns.heatmap(corr, annot=True)
Out[111]: <matplotlib.axes._subplots.AxesSubplot at 0xcaa79c8>
```




```
In [120]: sns.heatmap(corr1, annot=True)
Out[120]: <matplotlib.axes._subplots.AxesSubplot at 0x21026308>
```



Subtask 4.5: Checking the variables with highest correlation are the same in both the files or not?

```
Top10_Target0.head(10)
```

	VAR1	VAR2	Correlation
37	AMT_GOODS_PRICE	AMT_CREDIT	0.986338
141	APARTMENTS_AVG	TOTALAREA_MODE	0.892115
38	AMT_GOODS_PRICE	AMT_ANNUITY	0.777440
25	AMT_ANNUITY	AMT_CREDIT	0.775076
24	AMT_ANNUITY	AMT_INCOME_TOTAL	0.448436
36	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.380162
12	AMT_CREDIT	AMT_INCOME_TOTAL	0.378048
77	DAYS_REGISTRATION	DAYS_BIRTH	0.335233
136	APARTMENTS_AVG	REGION_POPULATION_RELATIVE	0.190690
112	TOTALAREA_MODE	REGION_POPULATION_RELATIVE	0.187044

```
Top10_Target1.head(10)
```

	VAR1	VAR2	Correlation
37	AMT_GOODS_PRICE	AMT_CREDIT	0.981995
141	APARTMENTS_AVG	TOTALAREA_MODE	0.878912
25	AMT_ANNUITY	AMT_CREDIT	0.750779
38	AMT_GOODS_PRICE	AMT_ANNUITY	0.749910
24	AMT_ANNUITY	AMT_INCOME_TOTAL	0.410953
12	AMT_CREDIT	AMT_INCOME_TOTAL	0.332006
36	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.330708
77	DAYS_REGISTRATION	DAYS_BIRTH	0.289886
88	EXT_SOURCE_2	REGION_POPULATION_RELATIVE	0.163350
136	APARTMENTS_AVG	REGION_POPULATION_RELATIVE	0.152971

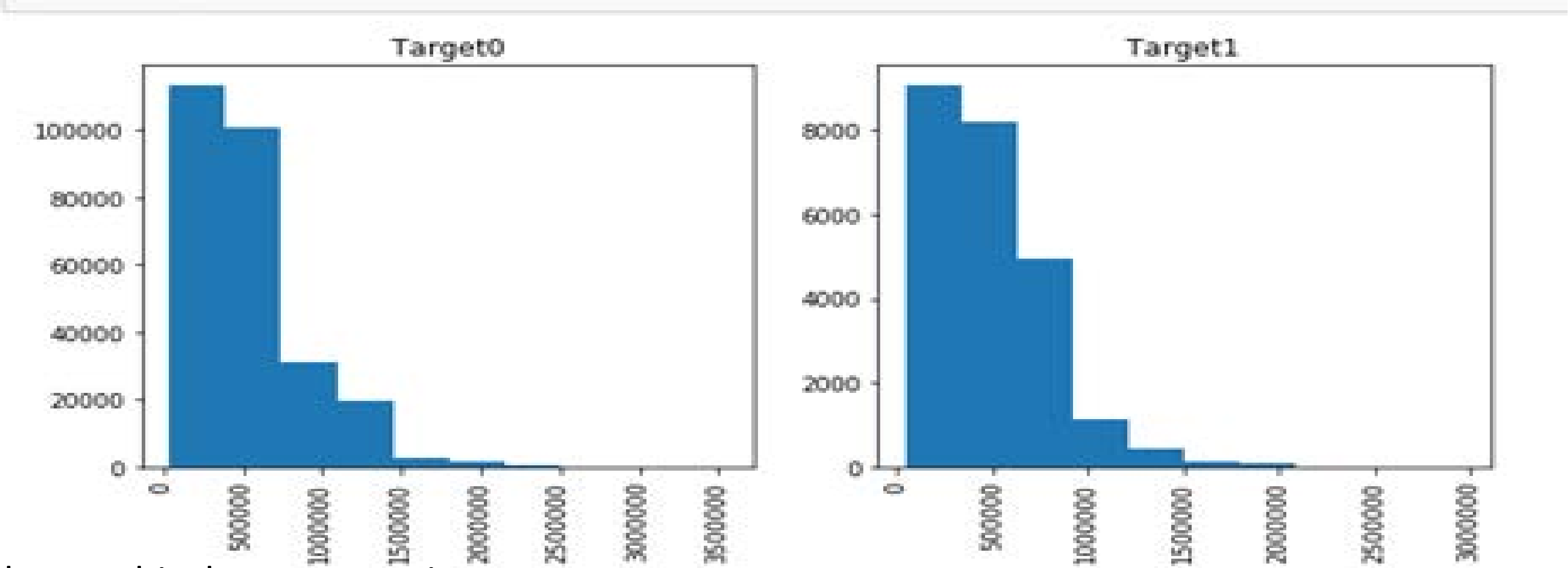
In Both the cases(Target0 and Target1), the top 4 highly correlated variables are same i.e.
1) 'AMT_GOODS_PRICE' 'AMT_CREDIT' 2) 'APARTMENTS_AVG' 'TOTALAREA_MODE' 3) 'AMT_GOODS_PRICE' 'AMT_ANNUITY' 4) 'AMT_ANNUITY' 'AMT_CREDIT

but from 5th highly correlated values are different in both the cases i.e.Target0 :
5) 'AMT_ANNUITY' 'AMT_INCOME_TOTAL'
6) 'AMT_GOODS_PRICE' 'AMT_INCOME_TOTAL'
7) 'AMT_CREDIT' 'AMT_INCOME_TOTAL'
8) 'DAYS_REGISTRATION' 'DAYS_BIRTH'
9) 'APARTMENTS_AVG' 'REGION_POPULATION_RELATIVE'
10) 'TOTALAREA_MODE' 'REGION_POPULATION_RELATIVE'

Target1:
5) 'DAYS_REGISTRATION' 'DAYS_BIRTH' 6) 'EXT_SOURCE_2' 'REGION_POPULATION_RELATIVE' 7) 'APARTMENTS_AVG' 'REGION_POPULATION_RELATIVE' 8) 'TOTALAREA_MODE' 'REGION_POPULATION_RELATIVE' 9) 'DAYS_BIRTH' 'AMT_GOODS_PRICE' 10) 'DAYS_BIRTH' 'AMT_CREDIT

Subtask 4.6: Performing univariate analysis for numerical variables for both 0 and 1. Comparing the target variable across categories of continuous variables.

1) **AMT_GOODS_PRICE**



As per the graphical representation,

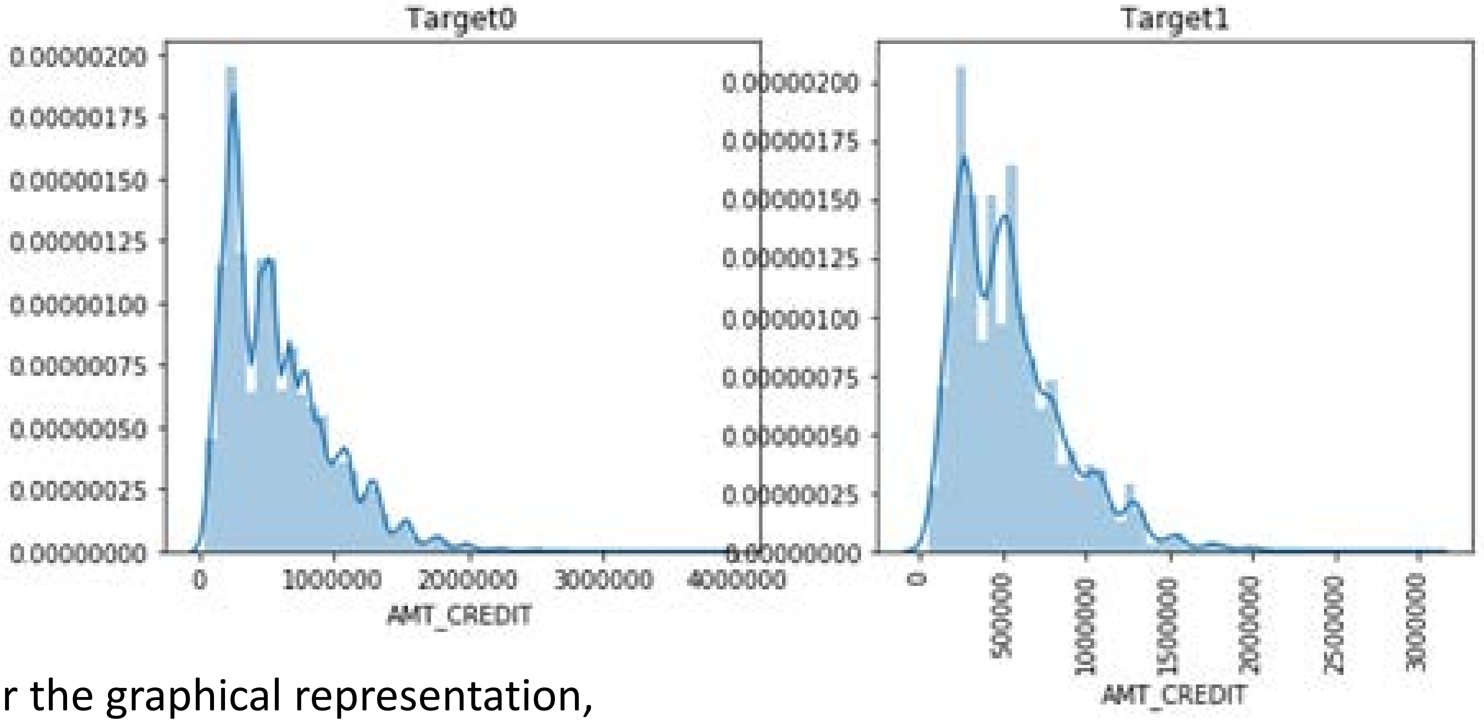
target0 : **AMT_GOODS_PRICE** varies mostly in between 1-4.5lakh

target1 : **AMT_GOODS_PRICE** varies mostly in between 1-4.5lakh. More percentage of people (14%) have got loan for 4.5Lakh as compare to target0

For consumer loans it is the price of the goods for which the loan is given

The spread of Target1 is less as compared to target0.

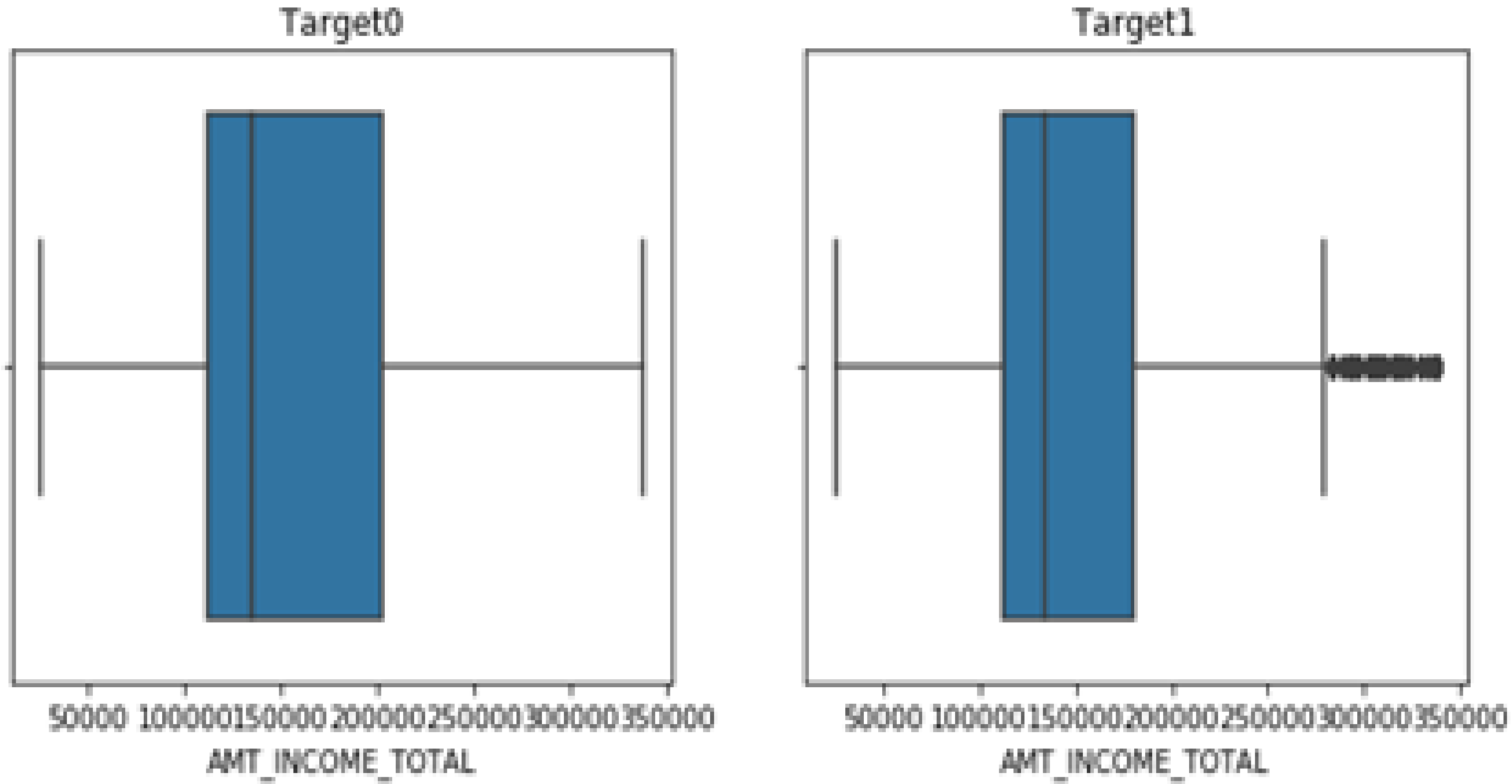
2) 'AMT_CREDIT'



As per the graphical representation,
Final credit amount on the previous application varies between 0-25lakh but in target0: for first 5 lakh it increases and then a down fall and again increases and fall gradually till 25 lakh target1: for first 5 lakh it increases gradually and then fall a little and increase again and then falls till 25 lakh

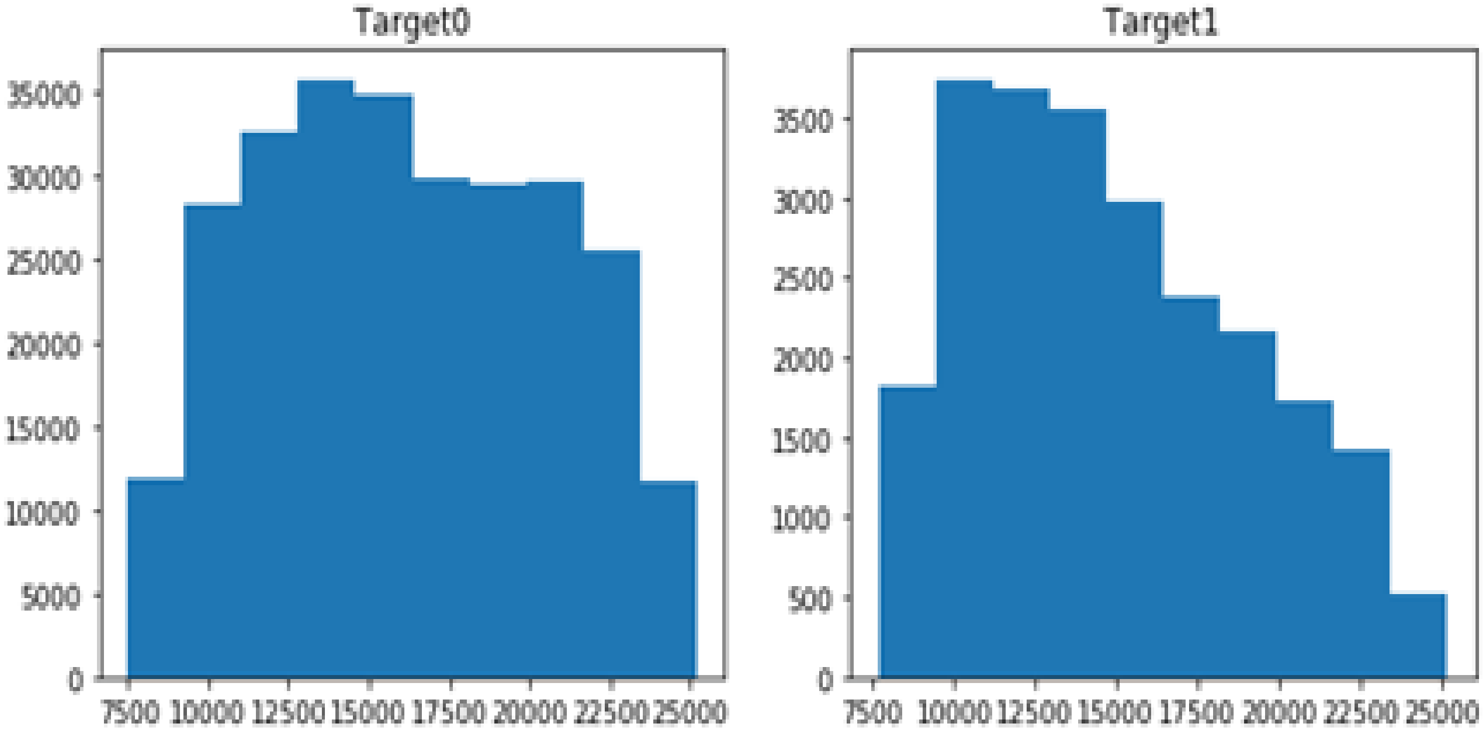
In comparison to both the graphs at around 2 lakh target0 has a deep fall as compared to the target1 graph .

3) 'AMT_INCOME_TOTAL'



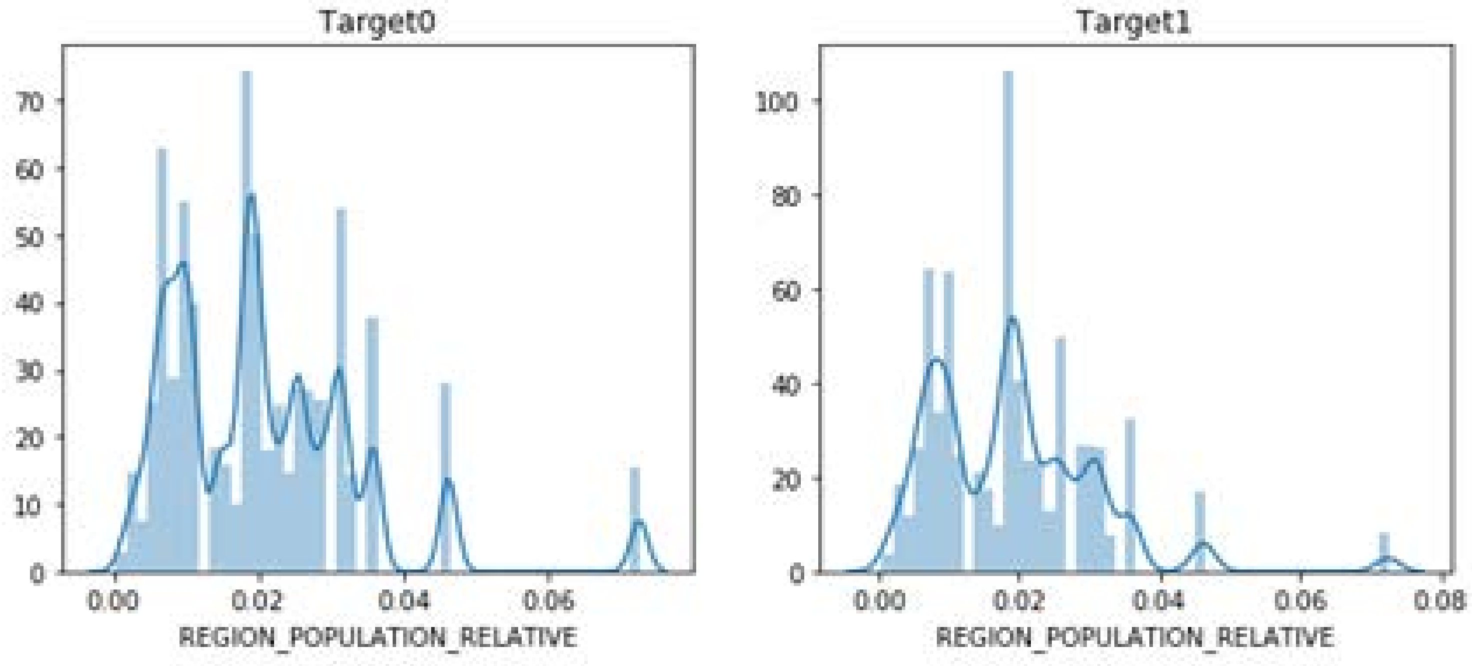
- As per the graphical representation, Target1: has many outliers

4) 'DAYS_BIRTH'



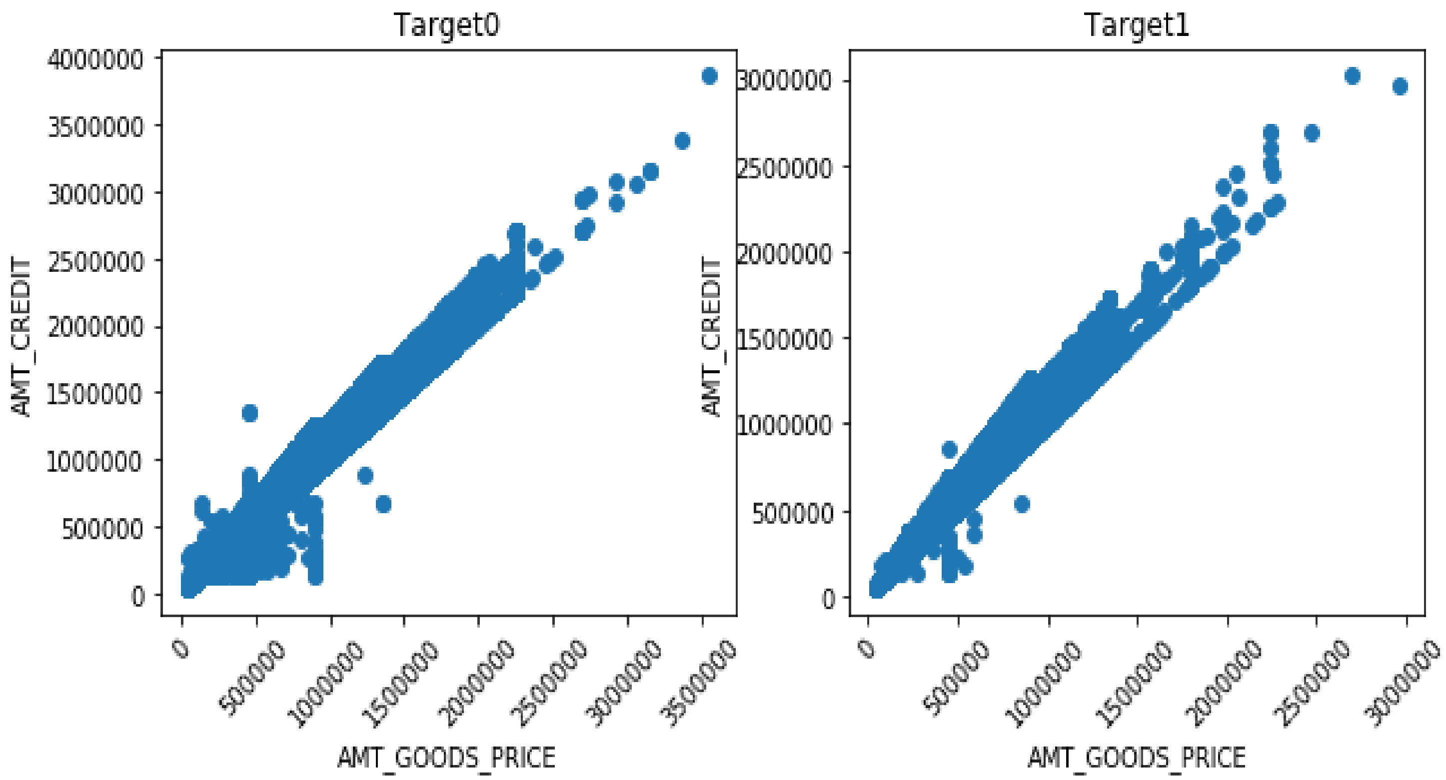
As per the graphical representation, In Target1 more percentage of young peoples are there as compared to Target0

5) 'REGION_POPULATION_RELATIVE'



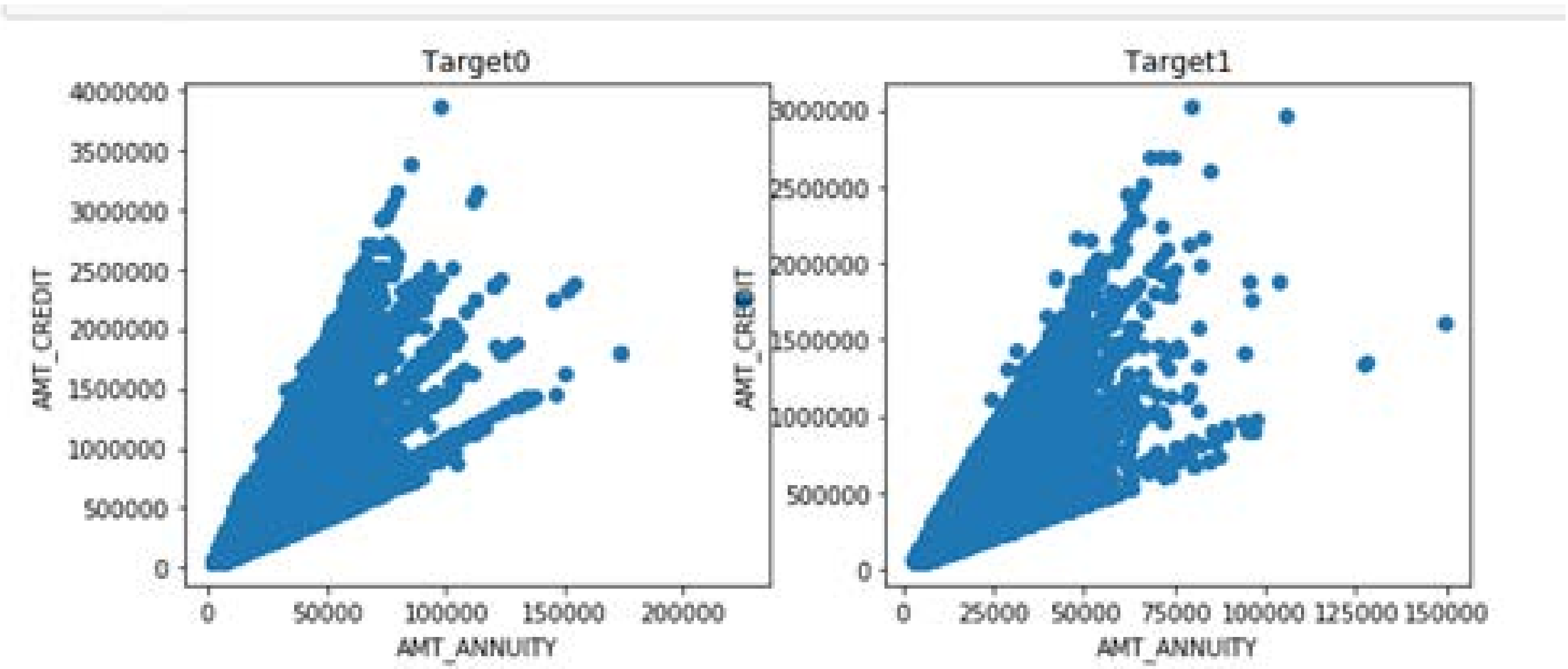
- As per the graphical representation, In both Target0 and Target1 more number of people residing in lesser populated region.

Subtask 4.7: Performing bivariate analysis for numerical variables for both 0 and 1. for continuous-continuous data¶
1) 'AMT_GOODS_PRICE' and 'AMT_CREDIT'



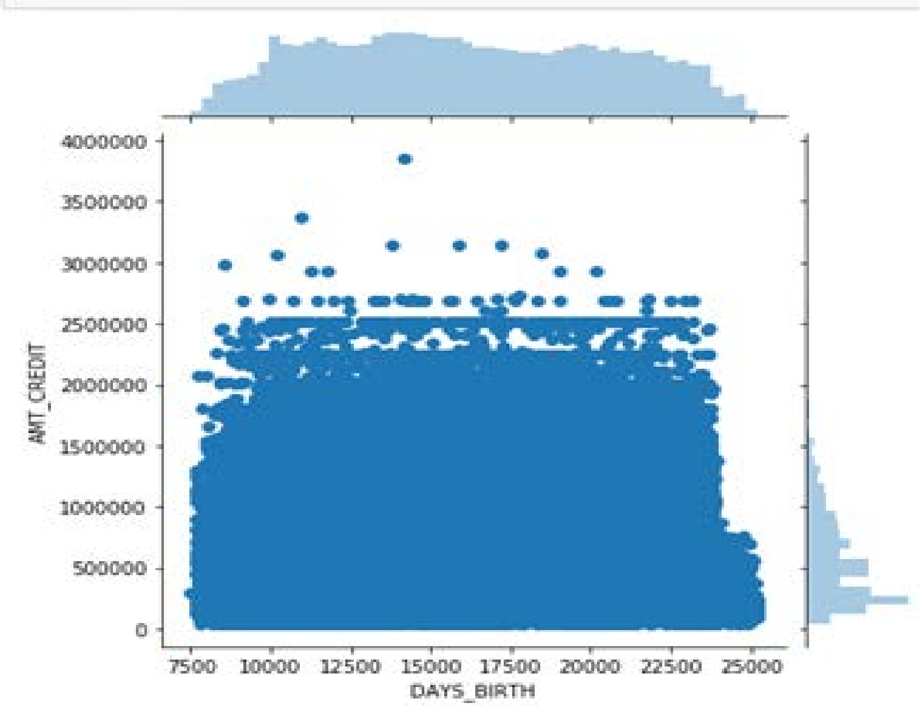
- As per the graphical representation, In Target0 more amount is credited against goods price as compare to Target1

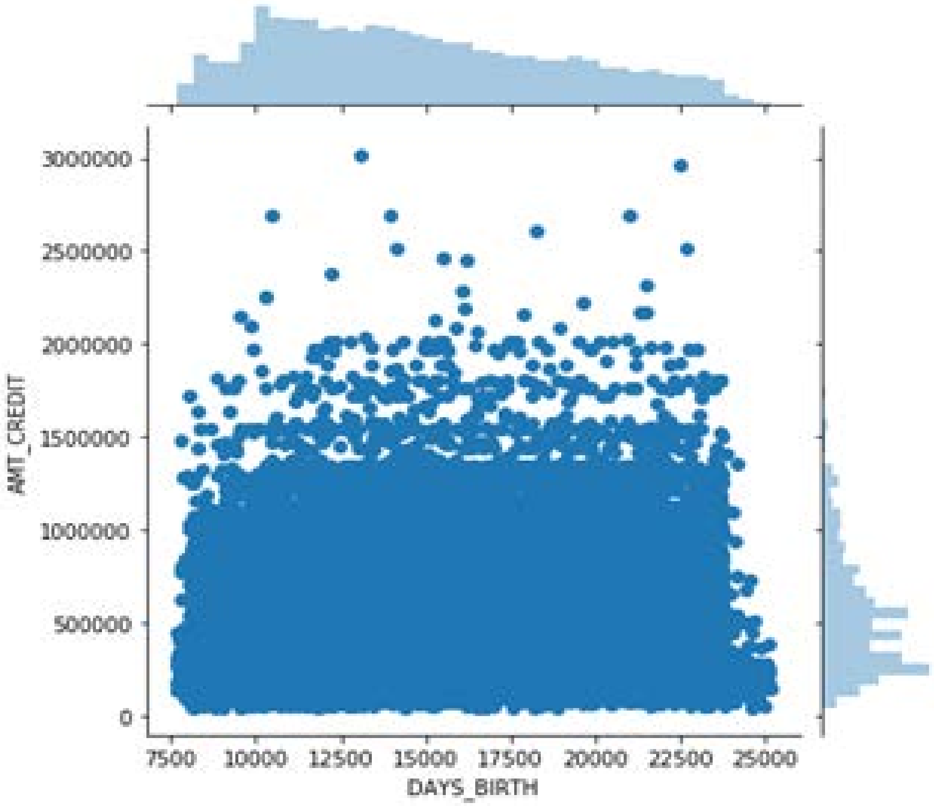
2) 'AMT_ANNUITY' and 'AMT_CREDIT'



As per the graphical representation, In Target0 more amount is credited against annuity amount as compare to Target1

3) 'DAYS_BIRTH' and 'AMT_CREDIT'

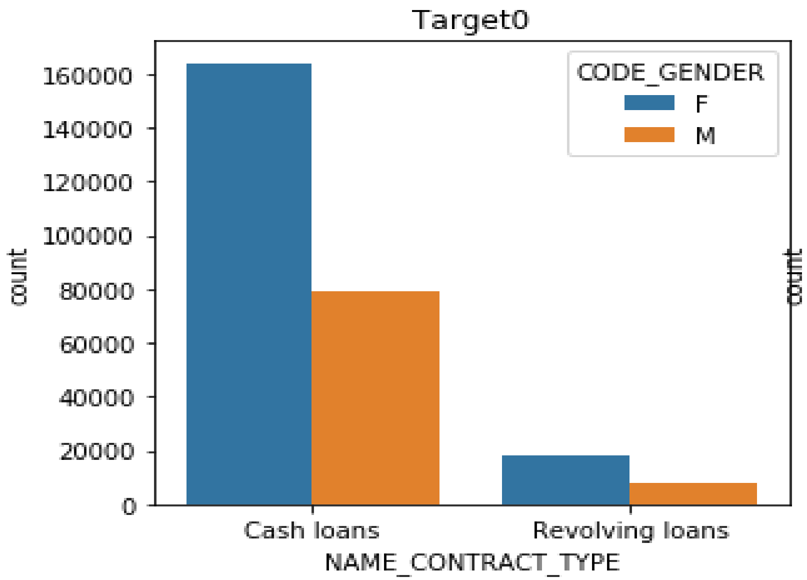


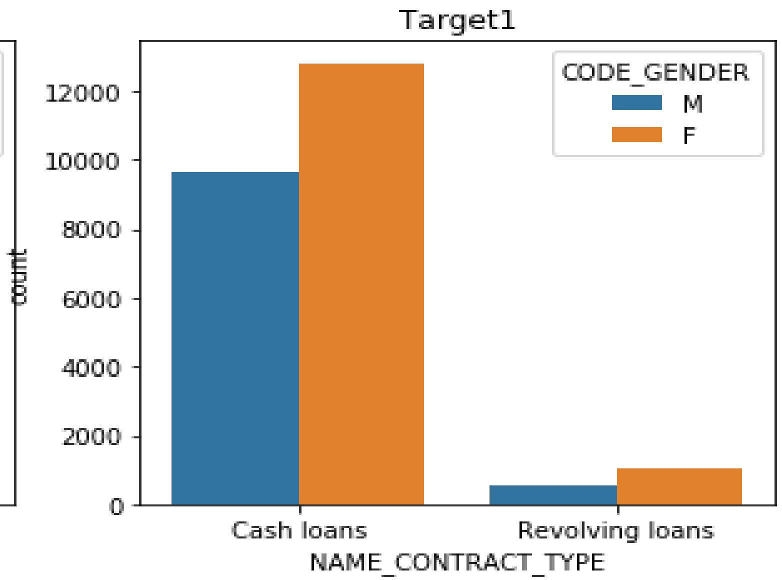


- As per the graphical representation, In Target1 more no of younger people have got amount credited as compare to Target0.

For categorical data - categorical data

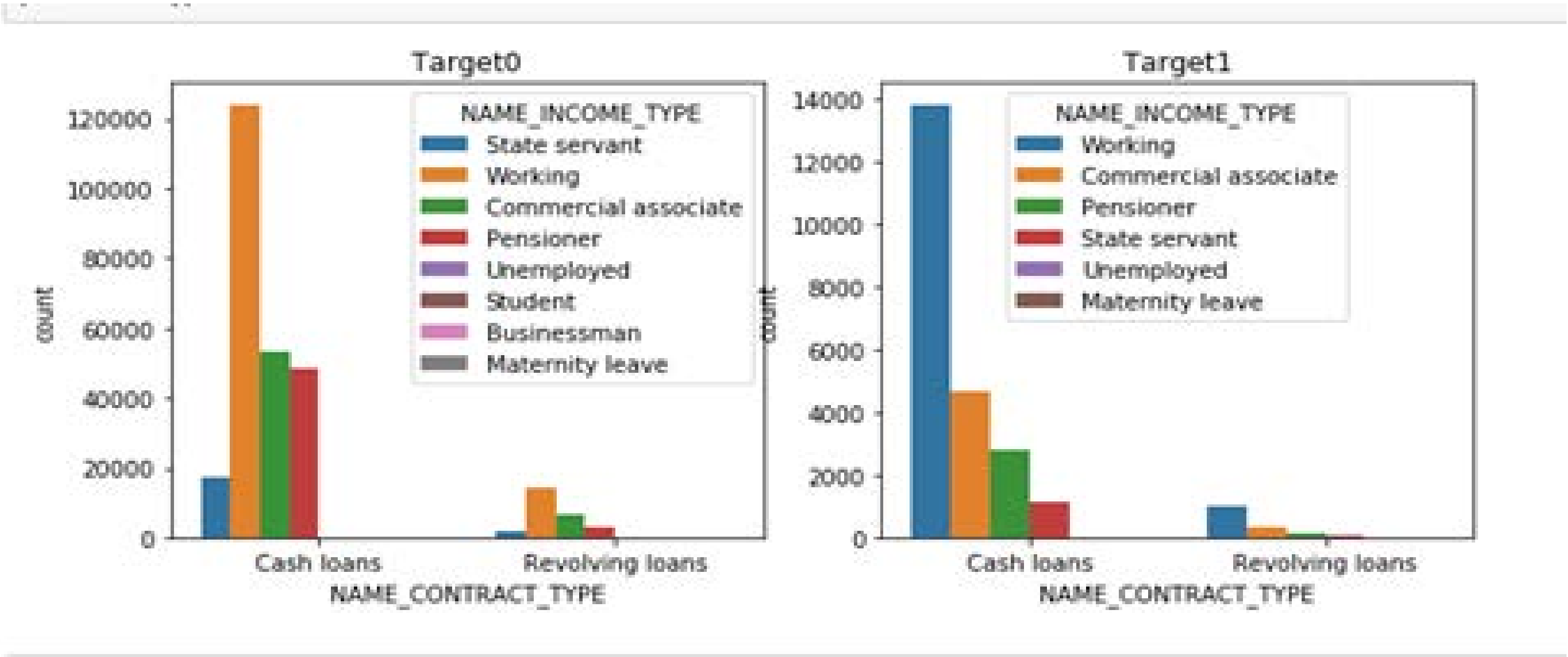
1) 'NAME_CONTRACT_TYPE‘ and 'CODE_GENDER'





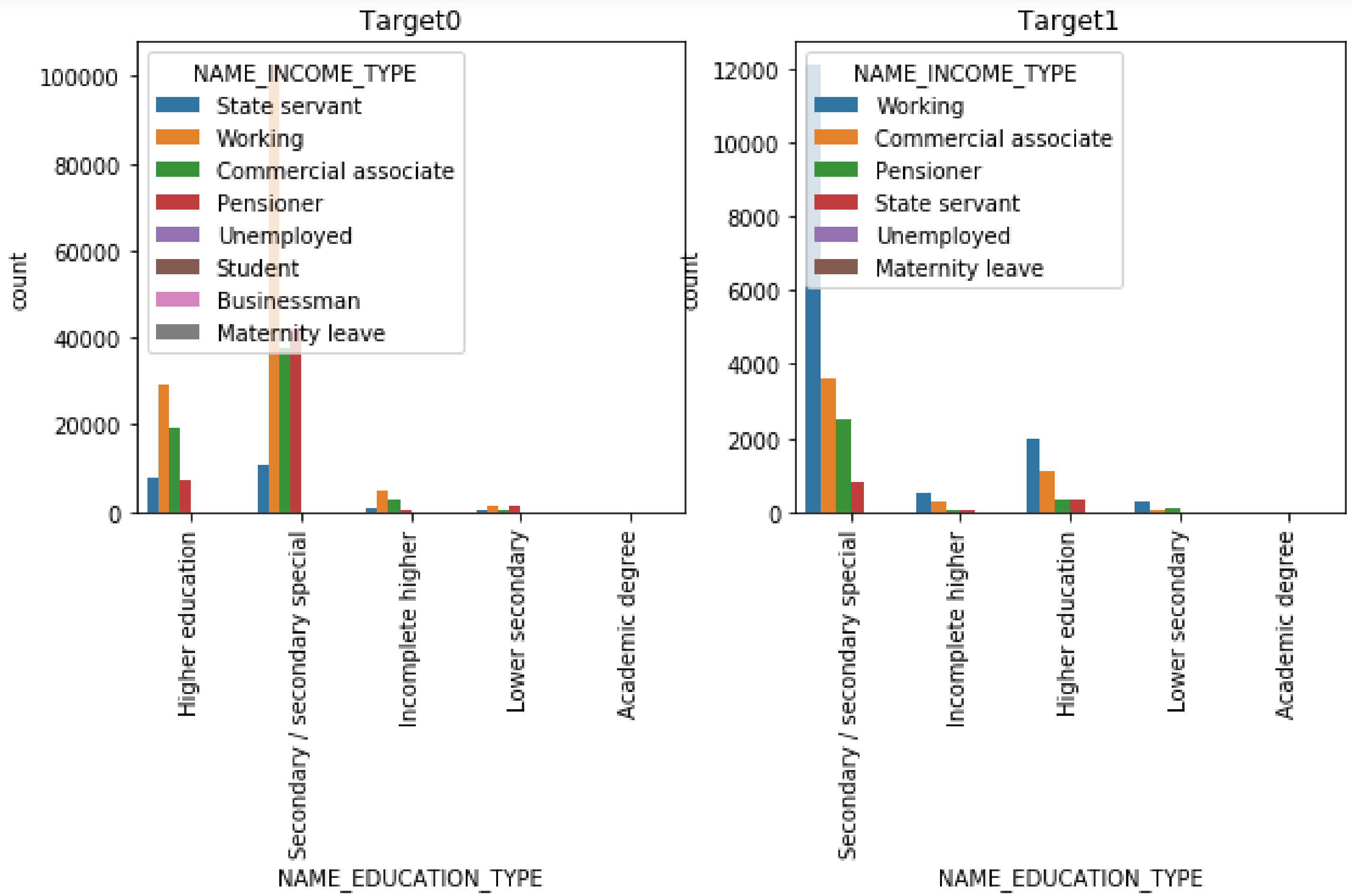
As per the graphical representation, both In Target0 and Target1 more number of females had preferred cash loans as compare to revolving loans.

2) 'NAME_CONTRACT_TYPE' and 'NAME_INCOME_TYPE'



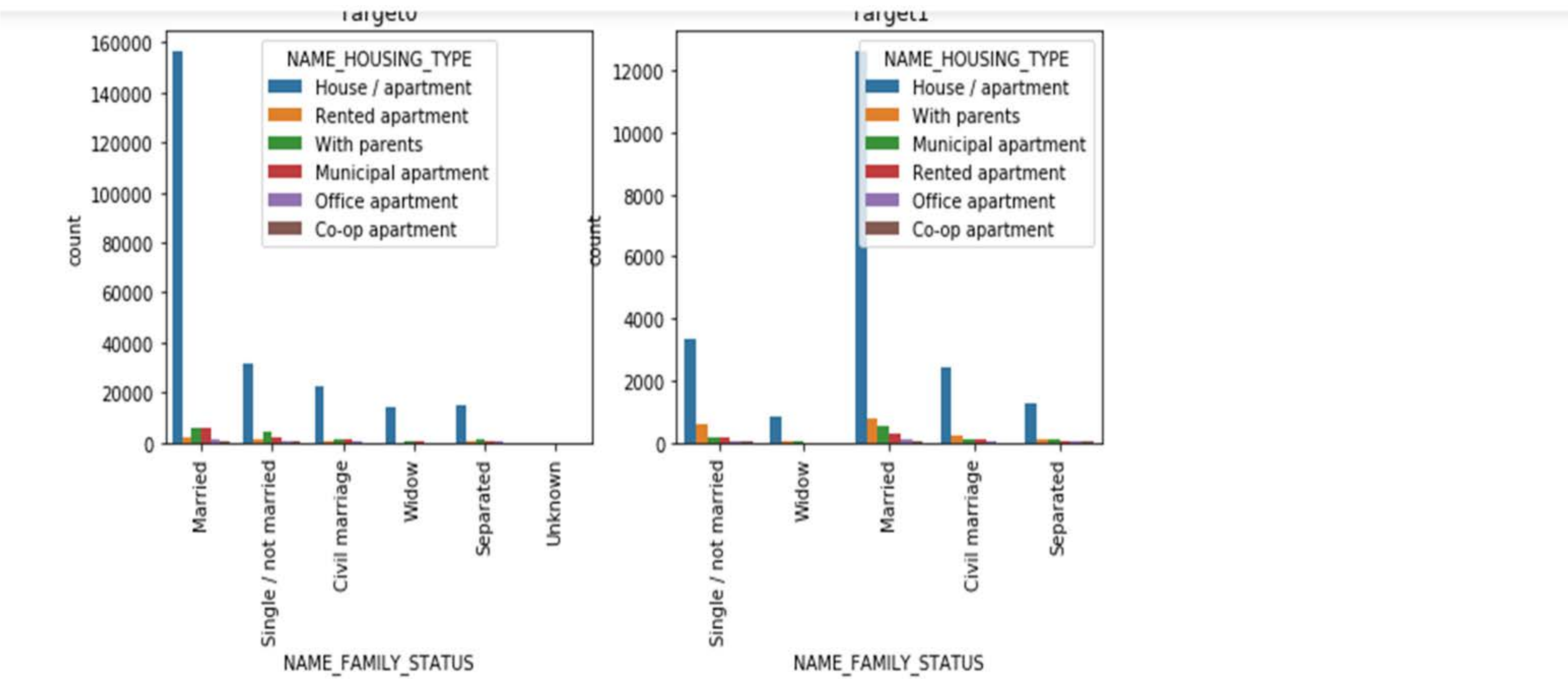
- As per the graphical representation, In Target0 and Target1 more no of working income type had preferred cash loans.

3) 'NAME_EDUCATION_TYPE‘ and 'NAME_INCOME_TYPE'



- As per the graphical representation, Both In Target0 and Target1 more no of people are from working income type and their highest education is from Secondary level. In Target 1 student and businessman variables are not there.

4) 'NAME_FAMILY_STATUS‘ and 'NAME_HOUSING_TYPE'

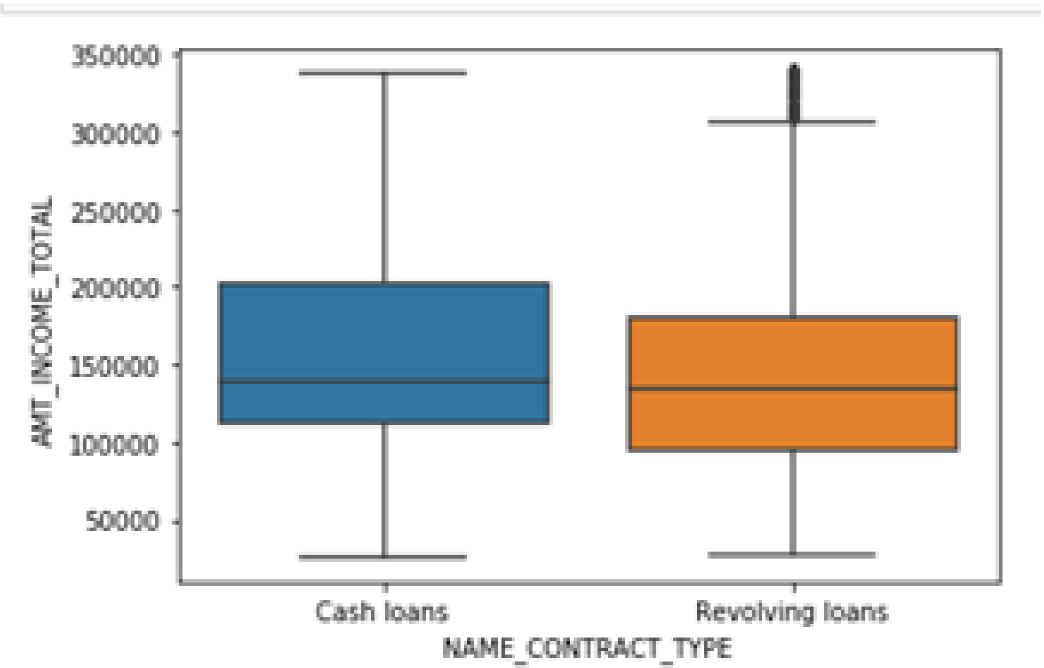


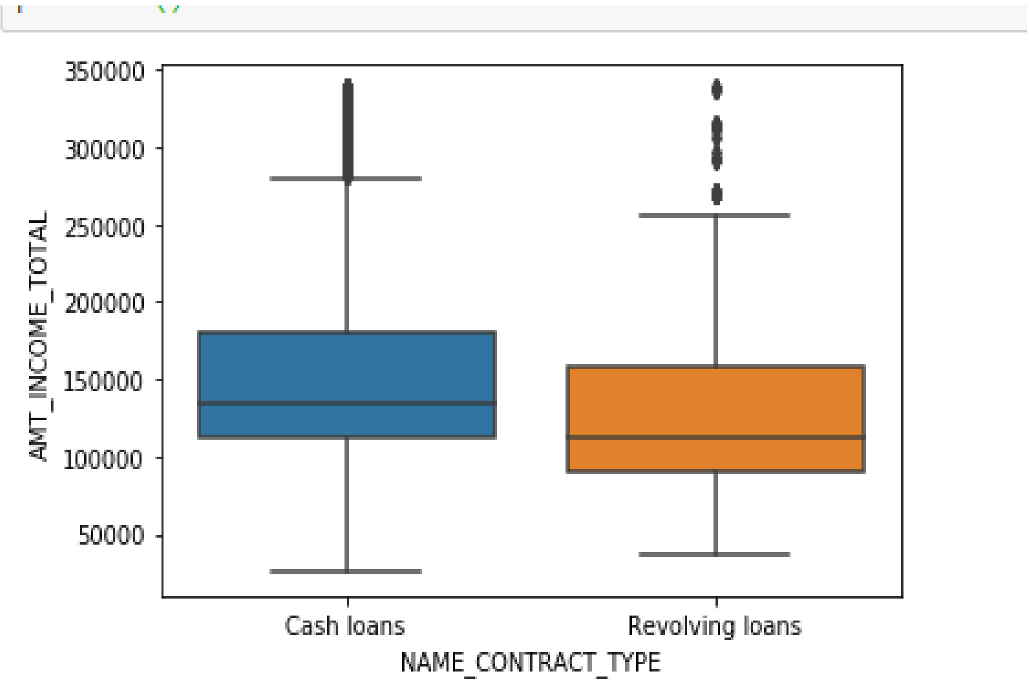
- As per the graphical representation, both In Target0 and Target1 more no of married person are living in apartment type of house

In Target0 number of married person living with parents and living in Municipal apartment is almost same while in Target1 its different.

For continuous - categorical data

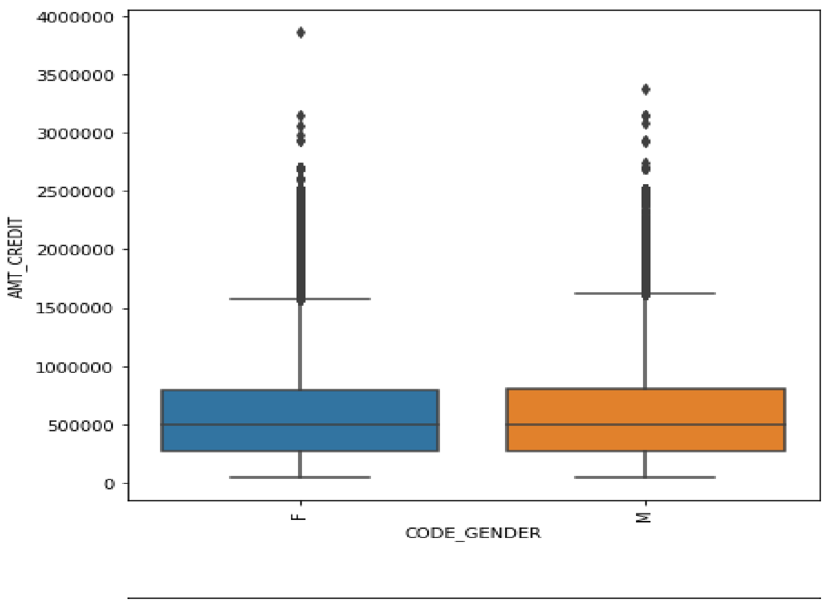
1) 'NAME_CONTRACT_TYPE‘ and 'AMT_INCOME_TOTAL'

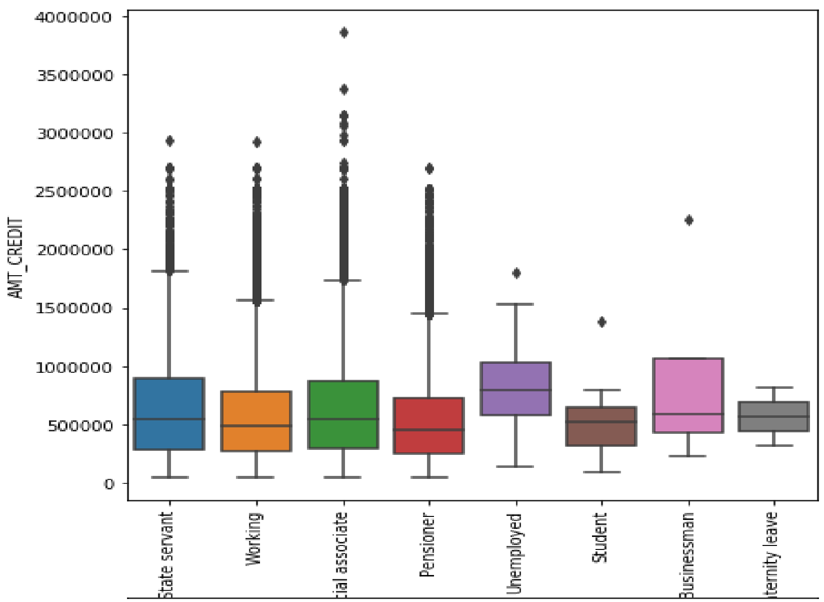


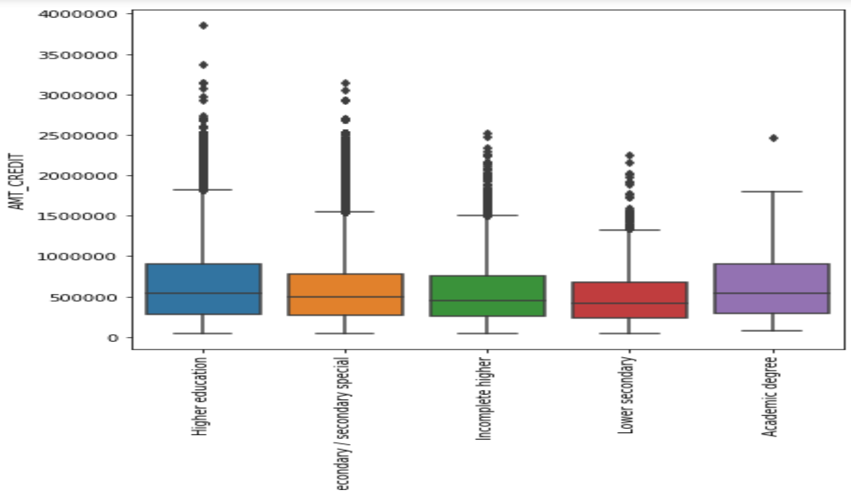


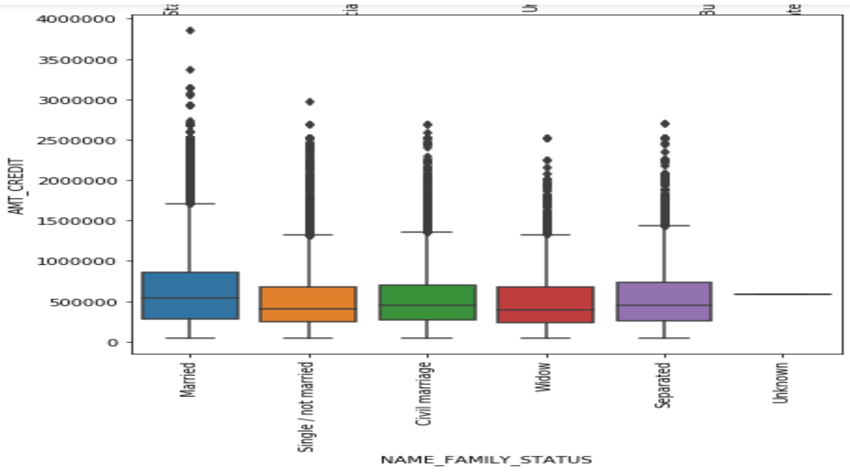
- As per the graphical representation,
In Target0 : Cash loans dont have outliers for total income amount
In Target1 : Cash loans have outliers for total income amount
In Target1 revolving loans have more outliers for total income amount as compared to Target0.

TARGET0 Categorical =['CODE_GENDER','NAME_INCOME_TYPE','NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS']
Continuous = 'AMT_CREDIT'

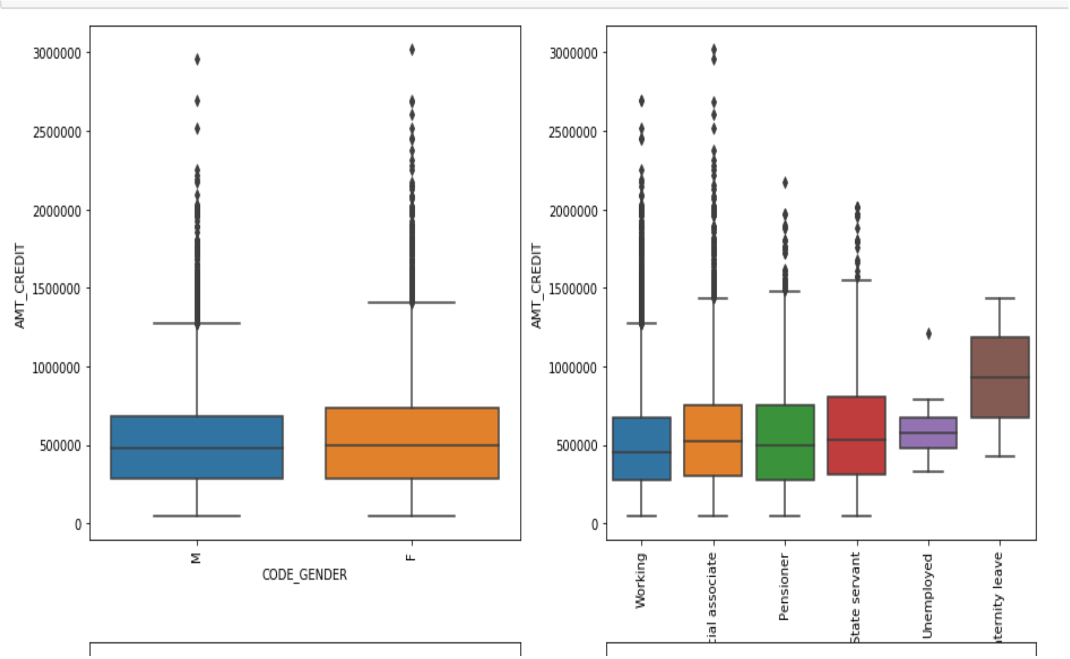


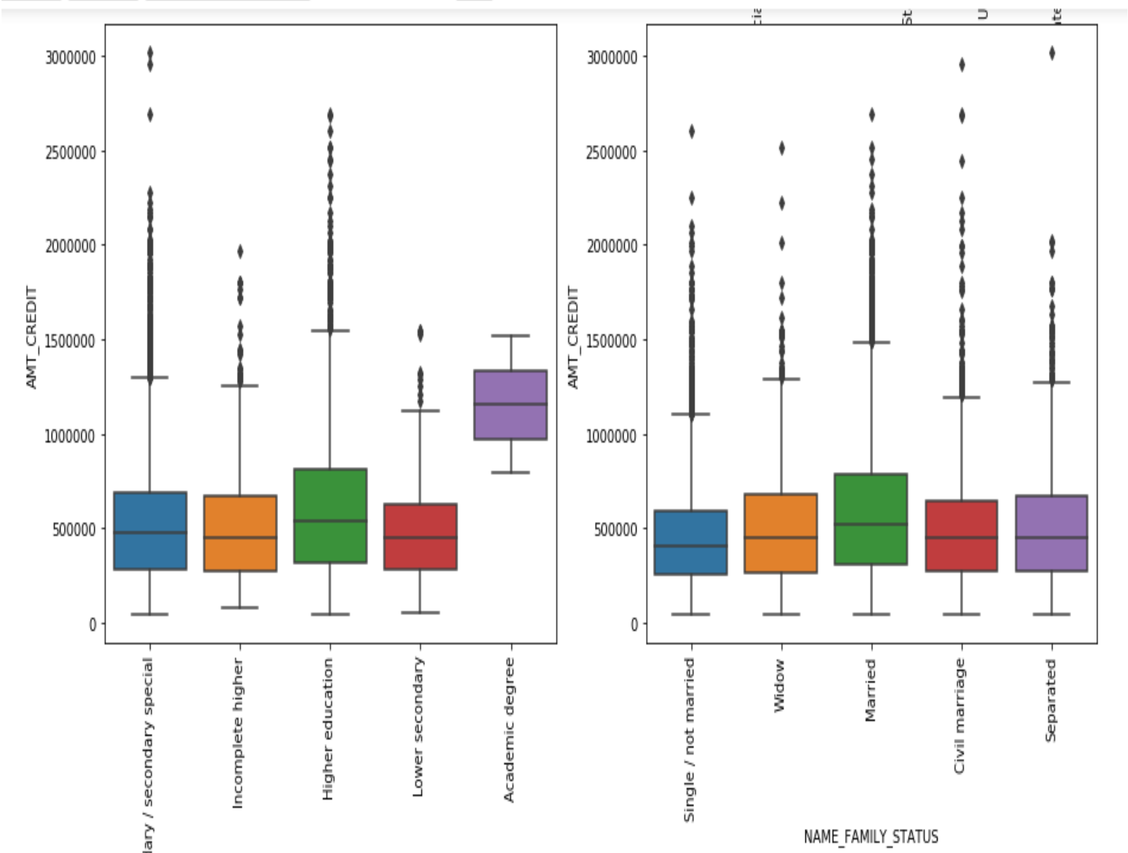






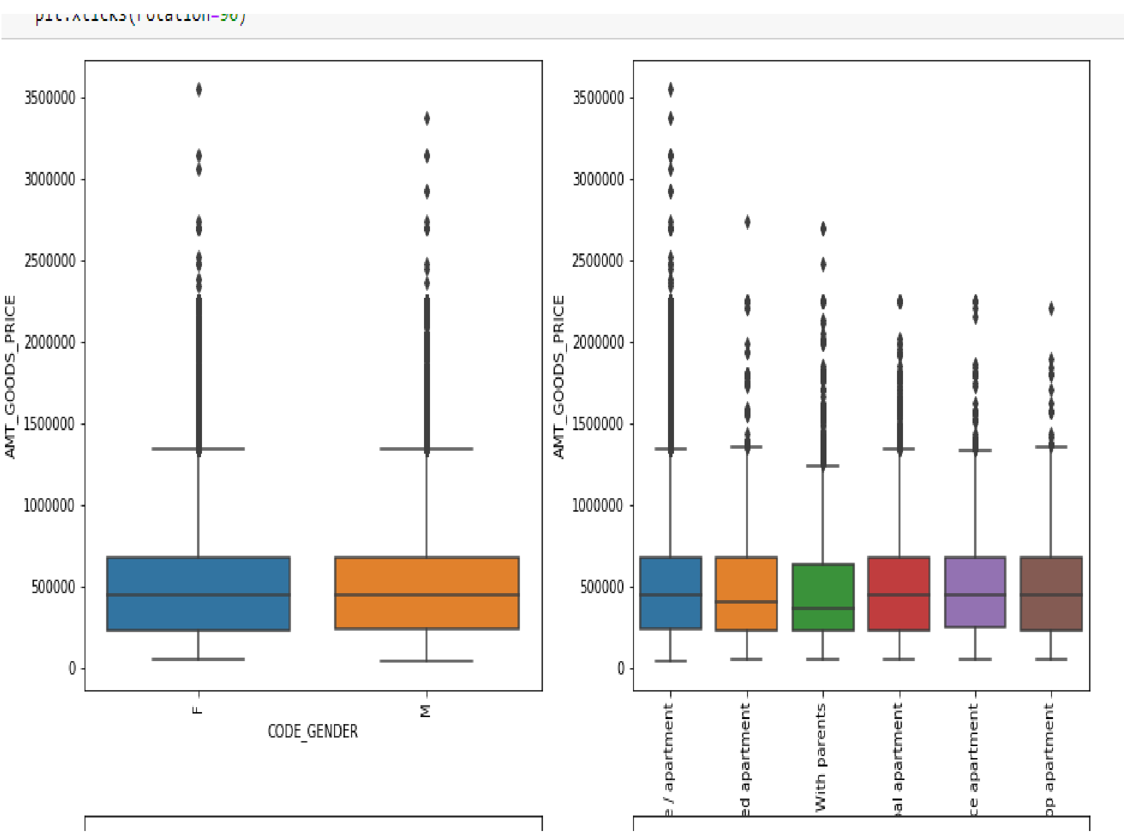
TARGET1 Categorical =['CODE_GENDER','NAME_INCOME_TYPE','NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS']
Continuous = 'AMT_CREDIT

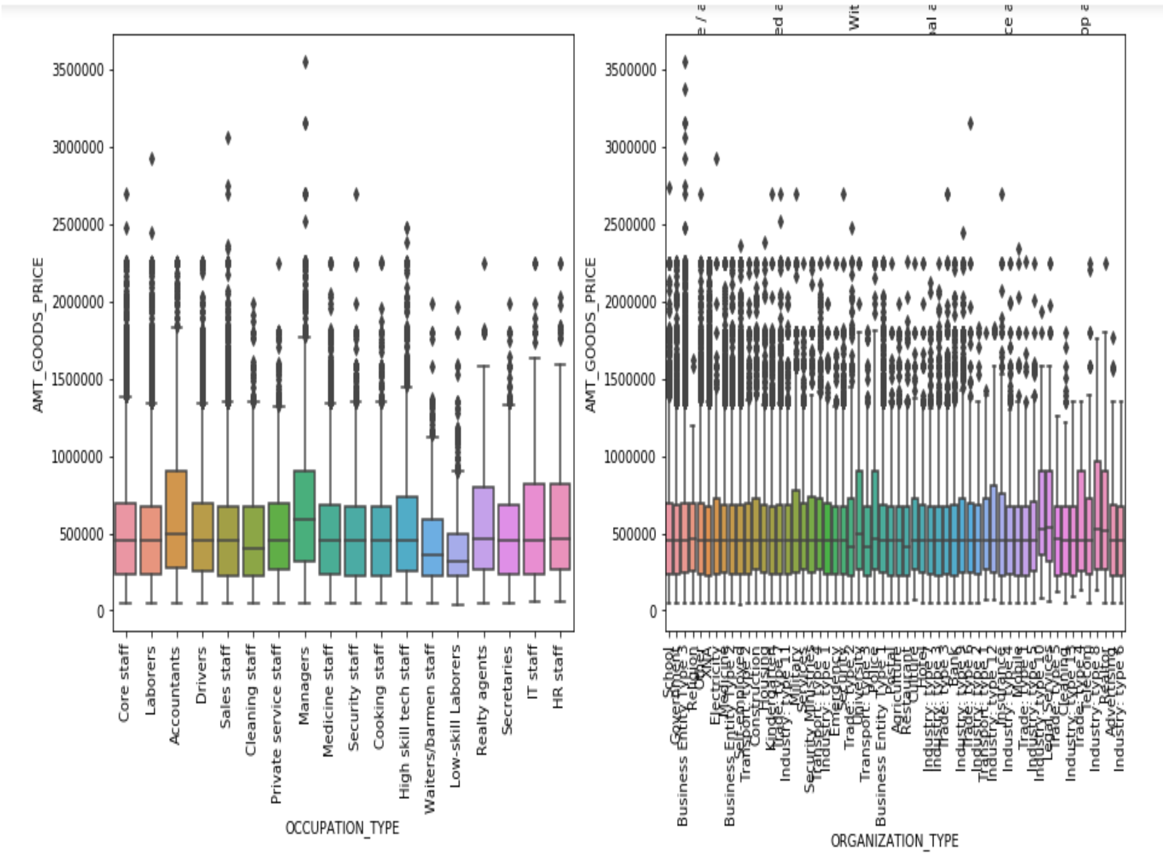




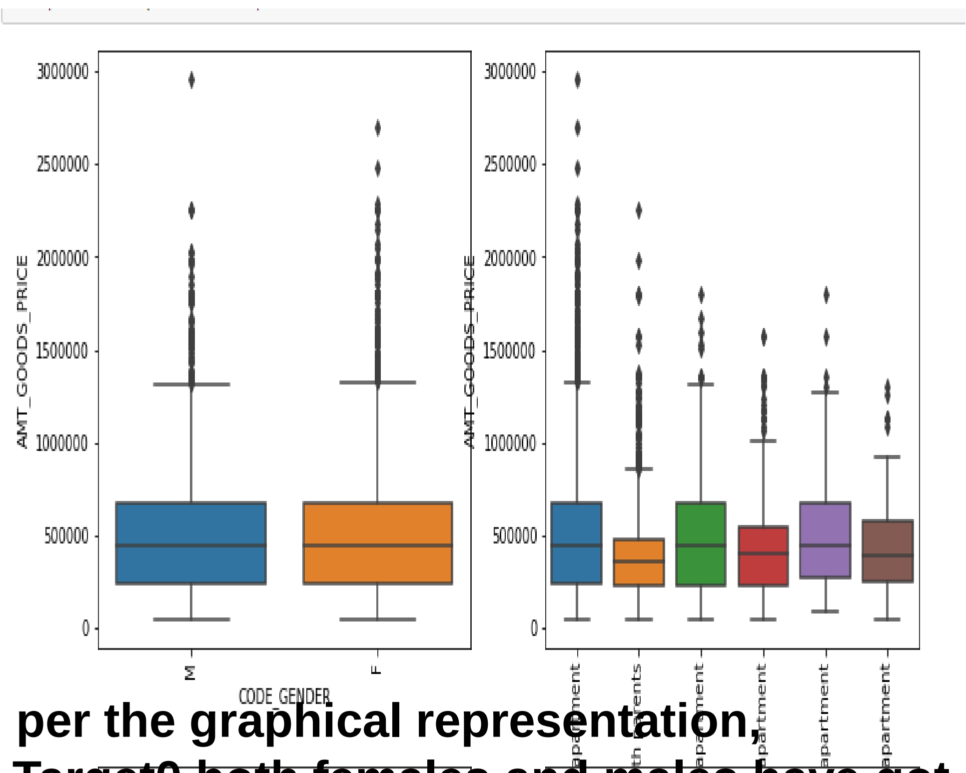
- As per the graphical representation,
1) Both in Target0 and Target1 more number of females have got amount credited
2) Both in Target0 and Target1 commercial associates have more credited amount.
3) In Target0 higher educated people have got more amounts credited while in Target1 secondary educated people have got more amount credited.
4) In Target0 married person have got more amount credited where as in Target1 separated and civil marriage people have got more amount credited.

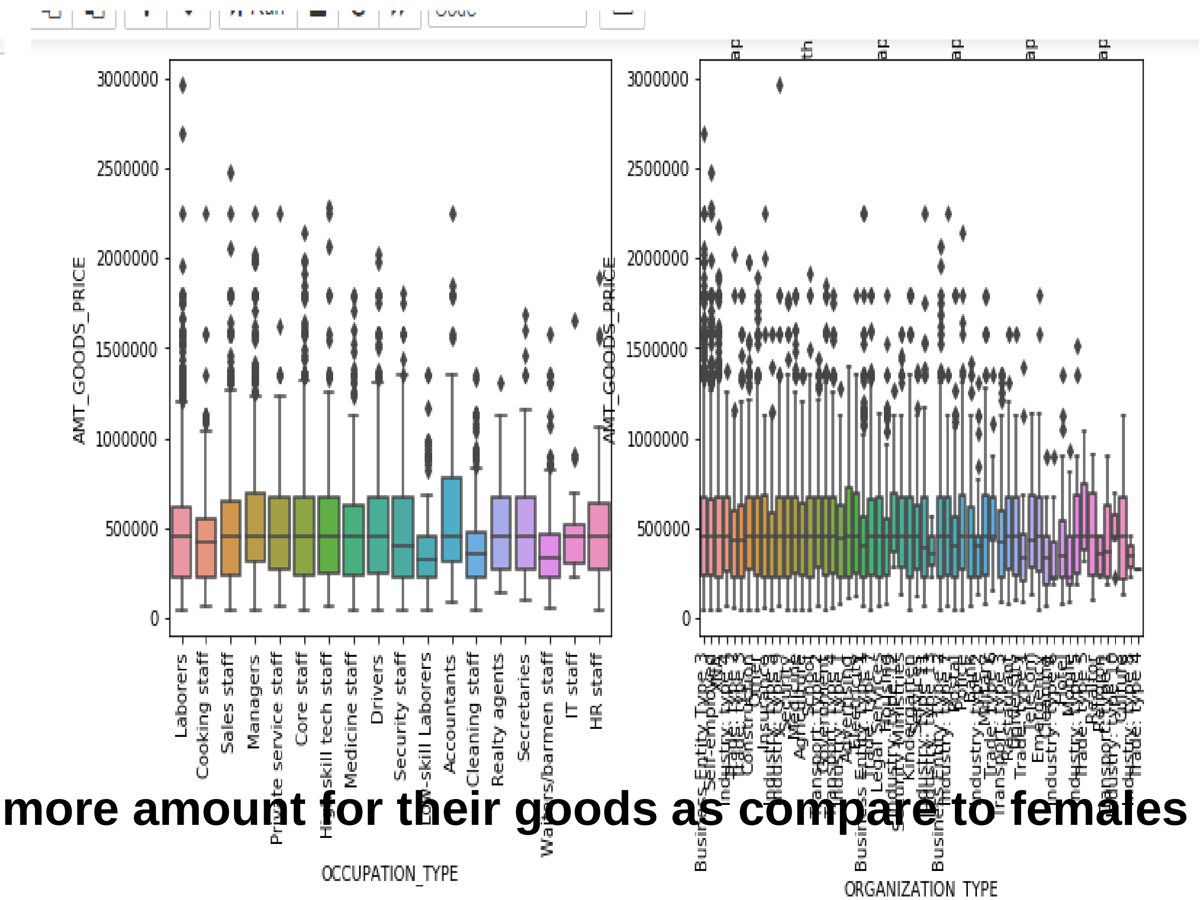
TARGET0 Categorical = 'CODE_GENDER','NAME_HOUSING_TYPE','OCCUPATION_TYPE','ORGANIZATION_TYPE‘
Continuous ='AMT_GOODS_PRICE'





TARGET1 Categorical = 'CODE_GENDER','NAME_HOUSING_TYPE','OCCUPATION_TYPE','ORGANIZATION_TYPE'
Continuous ='AMT_GOODS_PRICE





- As per the graphical representation,
1)In Target0 both females and males have got more amount for their goods as compare to females and males in Target1
2) Both in Target0 and Target1 people are living in apartment type of house have got more amount for their goods.
3) In Target0 managers have got more amount for their goods. In Target1 laborers have got more amount for their goods.
4) In Target0 business Entity Type 3 people have got more amount for their goods where as in Target1 Industry: type and business Entity Type 3 people have got more amount for their goods

TASK 5: After Reading and Analysis of the “Previous Application” data.

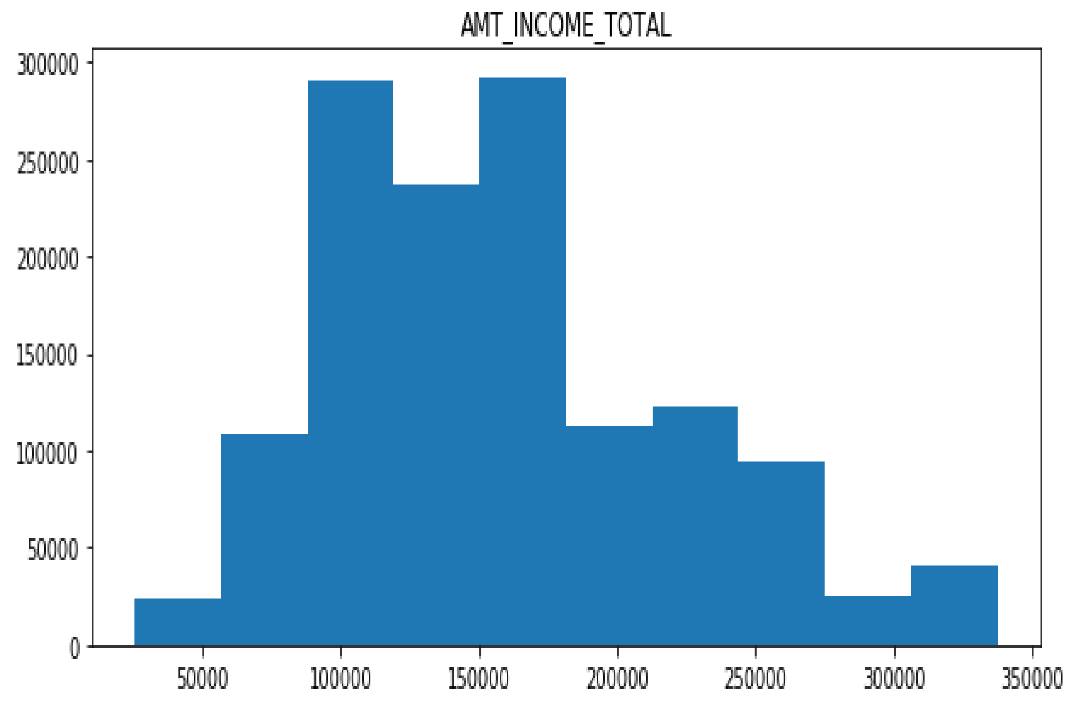
Subtask 5.1: merging the files with application data

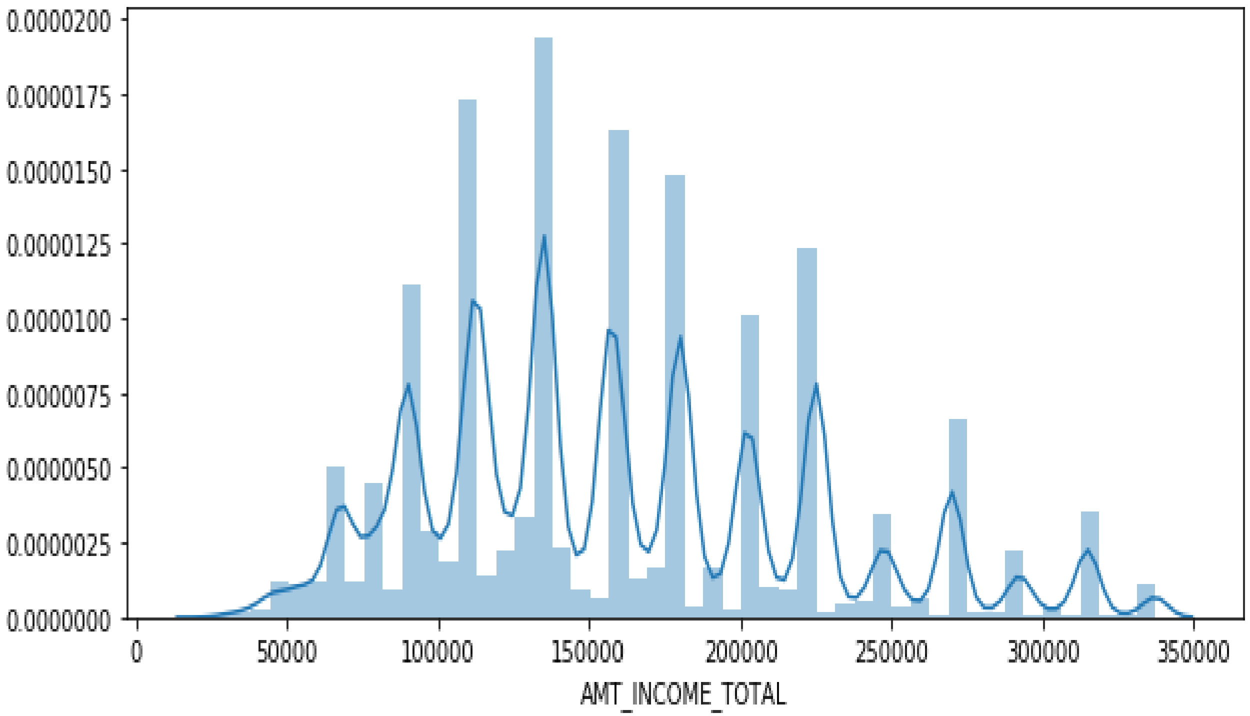
```
application_prev_data = pd.merge(application_data, previous_application, how='inner', on='SK_ID_CURR')
application_prev_data
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL
0	100002	1	Cash loans	M	N	Y	0	202500.0
1	100003	0	Cash loans	F	N	N	0	270000.0
2	100003	0	Cash loans	F	N	N	0	270000.0
3	100003	0	Cash loans	F	N	N	0	270000.0
4	100004	0	Revolving loans	M	Y	Y	0	67500.0
...	...	...	...	...	...	...	...	...
1348412	456255	0	Cash loans	F	N	N	0	157500.0
1348413	456255	0	Cash loans	F	N	N	0	157500.0
1348414	456255	0	Cash loans	F	N	N	0	157500.0
1348415	456255	0	Cash loans	F	N	N	0	157500.0
1348416	456255	0	Cash loans	F	N	N	0	157500.0

1348417 rows × 140 columns

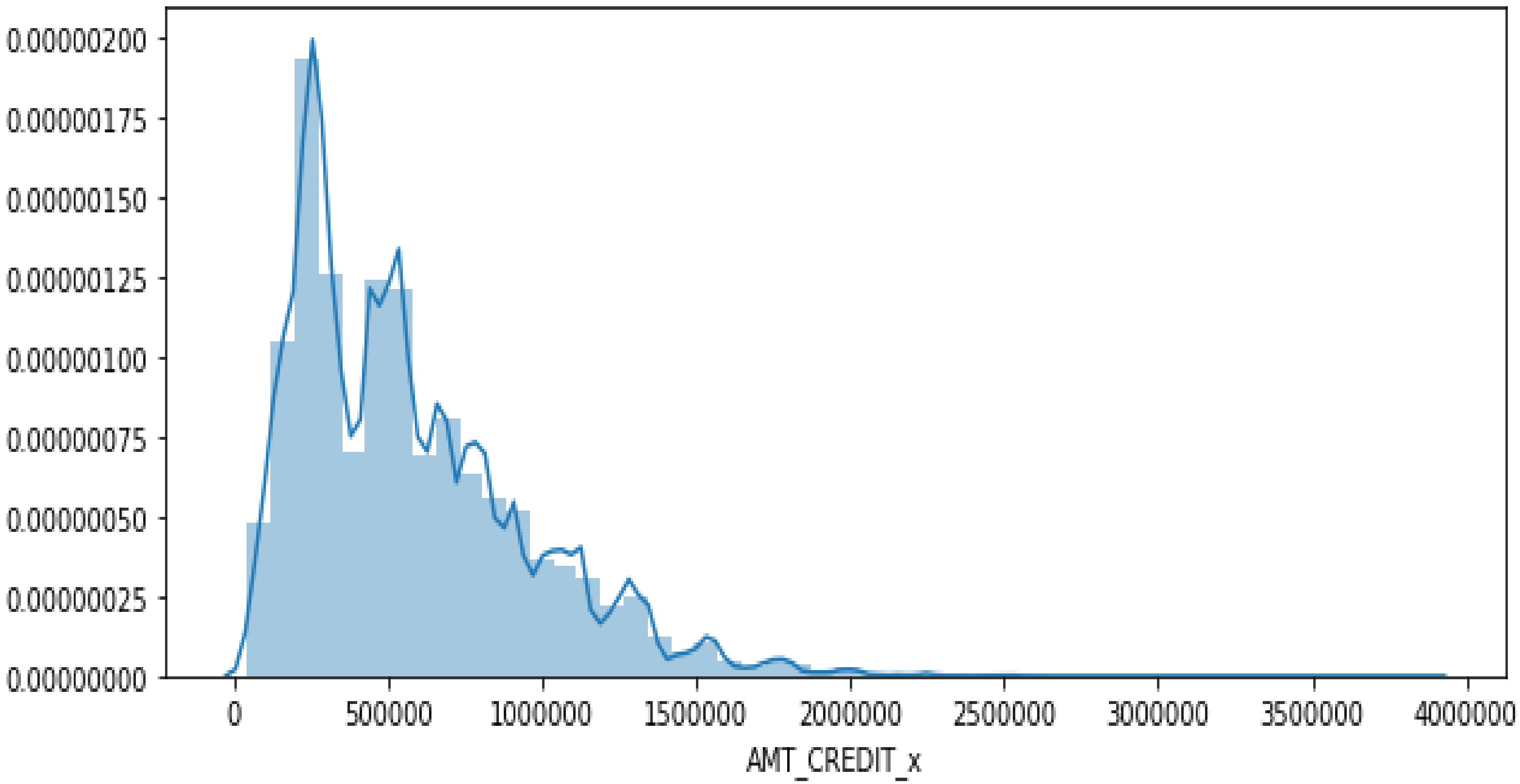
Subtask 5.2: Performing univariate and bivariate analysis to find some pattern.
univariate analysis for numerical(continuous) variables
1) 'AMT_INCOME_TOTAL'





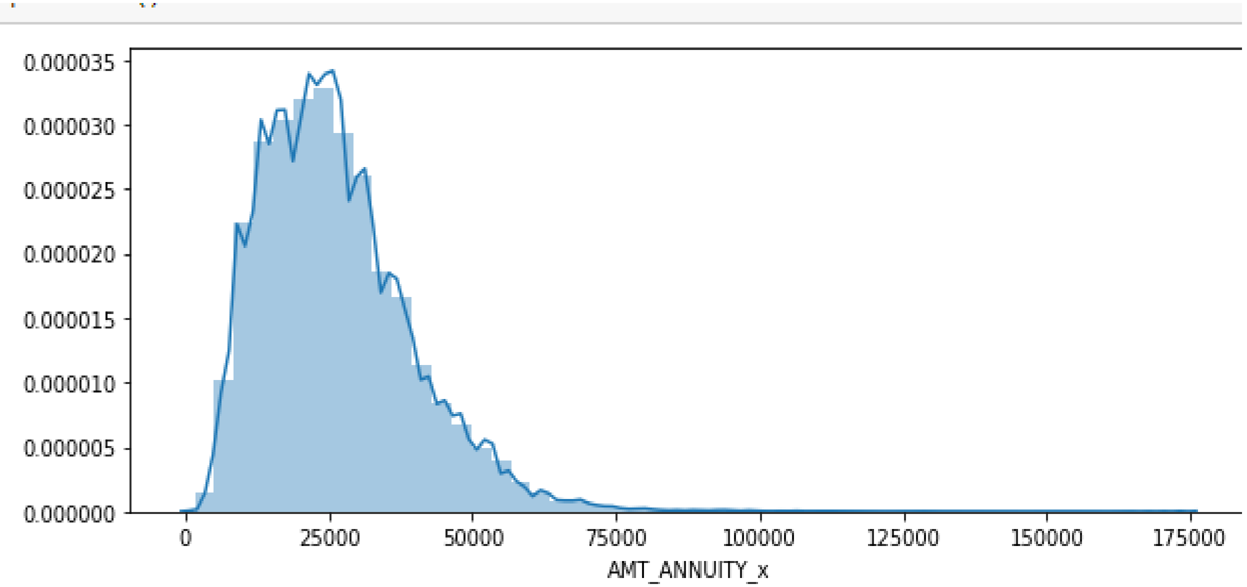
Income ranges from 25k to 350k.Tere are few spikes in between. Income of the client majorly lies between 90 thousand to 18lakhs(middle class) and then another group of clients income are in the range of 18 to 28 lakhs(upper middle class) and then few clients are in range of 28 and above.

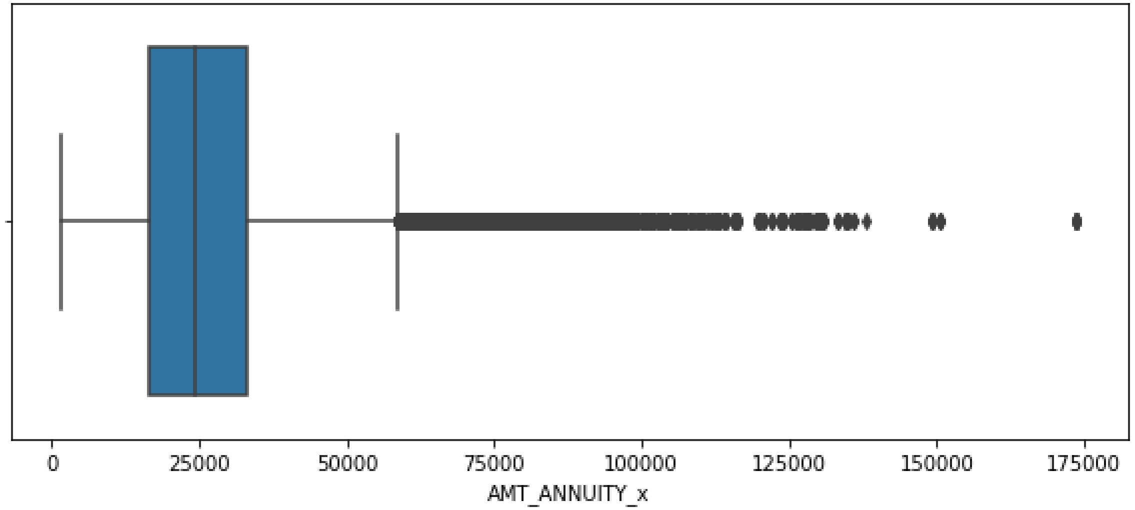
2) 'AMT_CREDIT_x'



Majority loan amount credited to people is in range of 3-4 lakh, another range of credits is around for 5 lakhs and then higher credit amount variations decreases

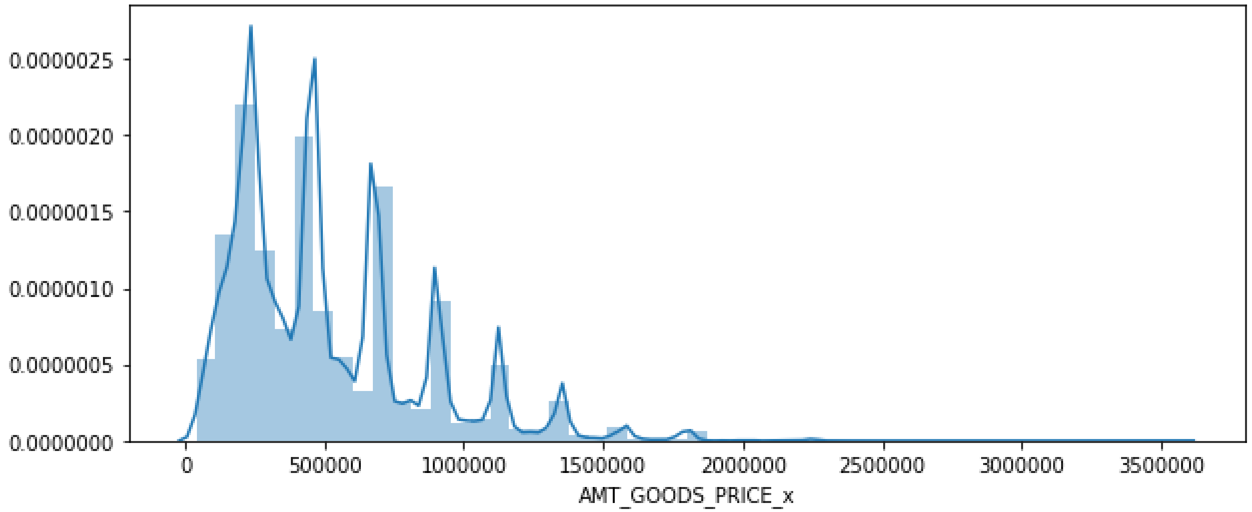
3) 'AMT_ANNUITY_x'

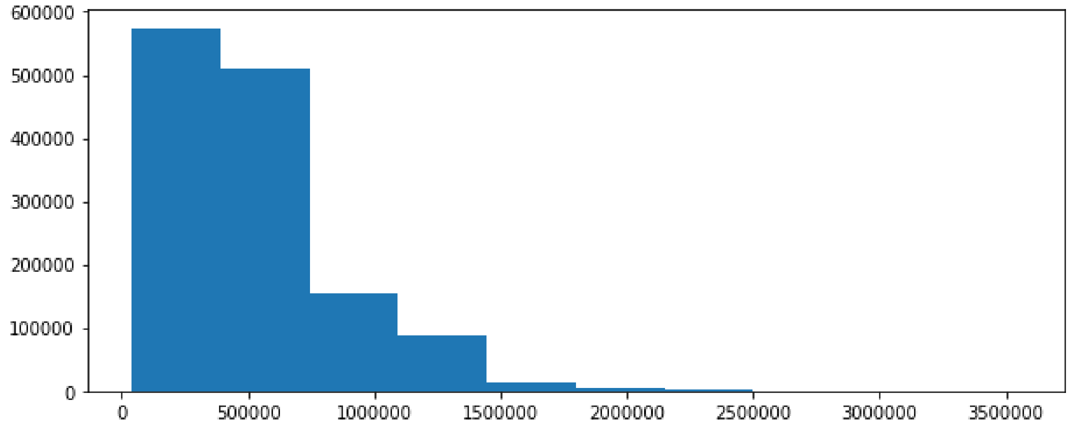




Loan annuity amount mostly is in range of 0 to 26 thousand and then it keeps on decreasing.

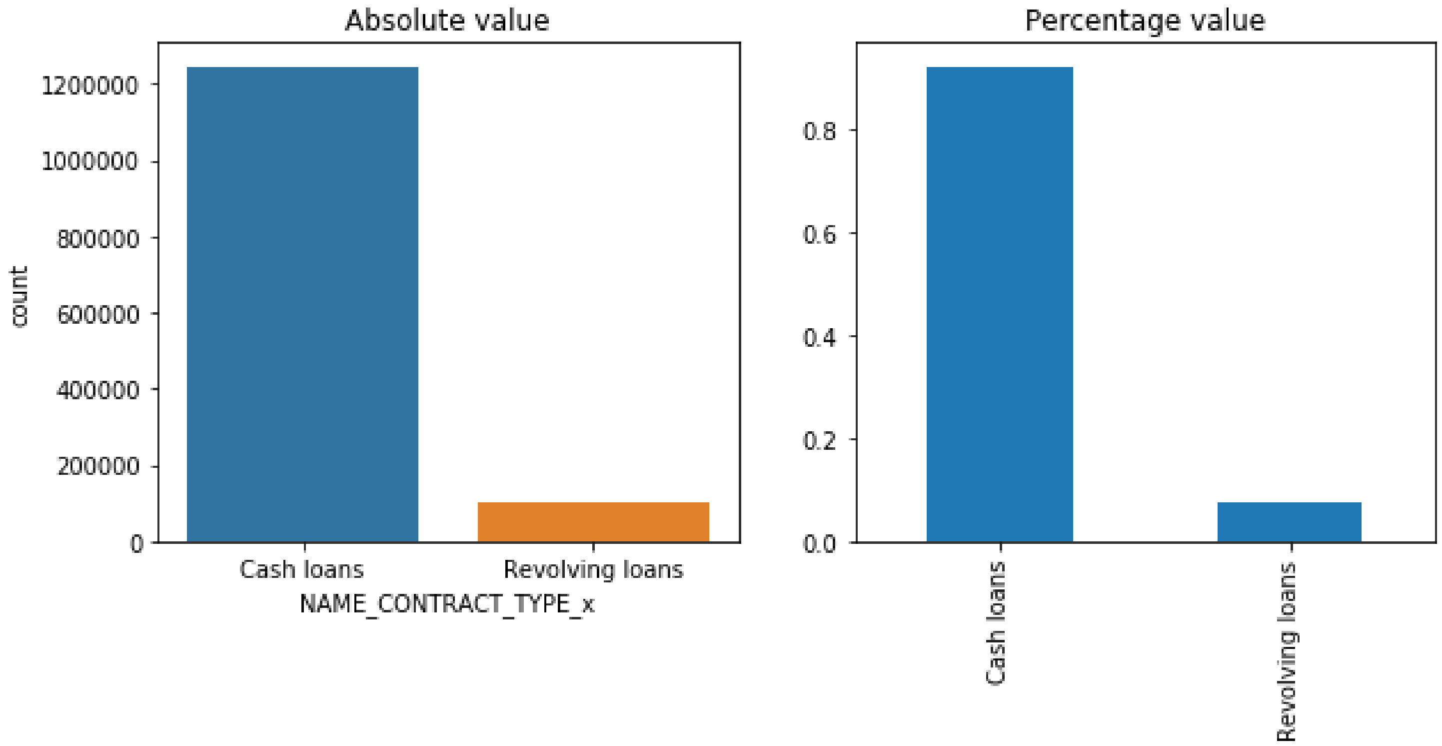
4) 'AMT_GOODS_PRICE_x'





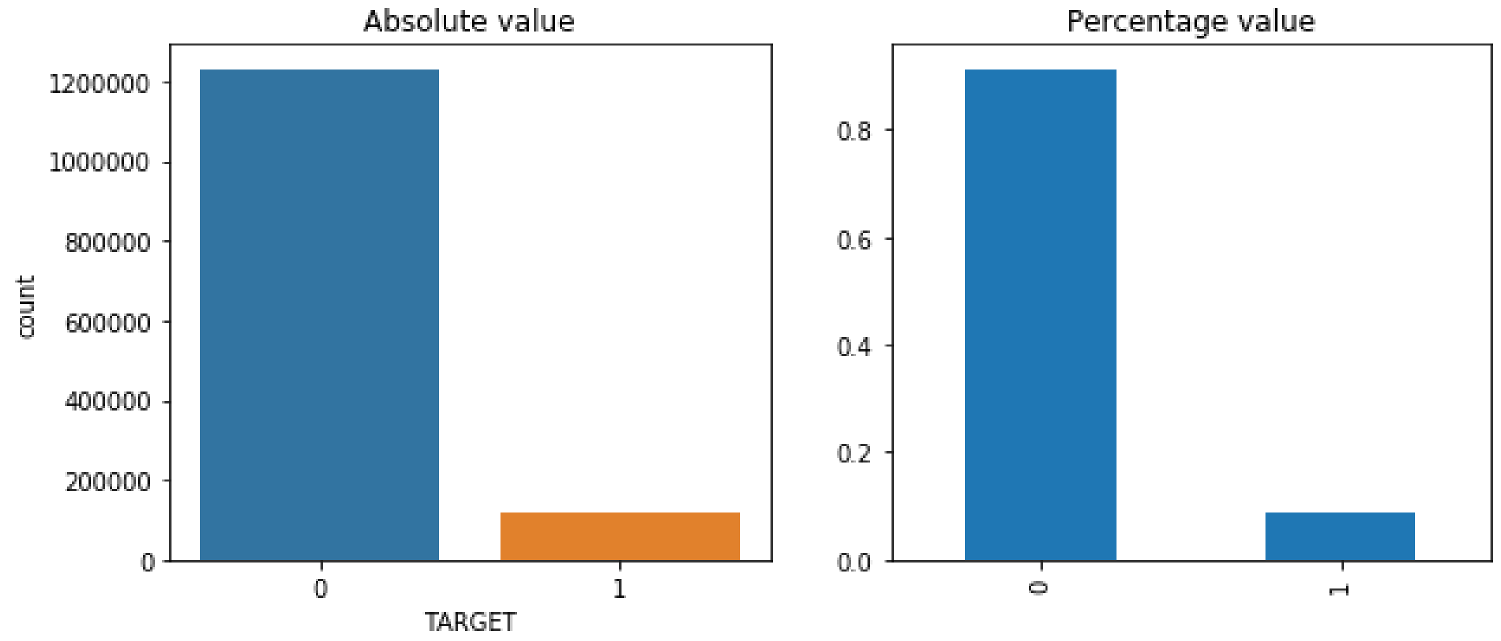
For consumer loans, the price of the goods for which the loan is given, majority people got 50 thousand to 3.5 lakhs and then for higher amount the number of people is gradually decreasing.

univariate analysis for categorical variables
1) NAME_CONTRACT_TYPE_x.



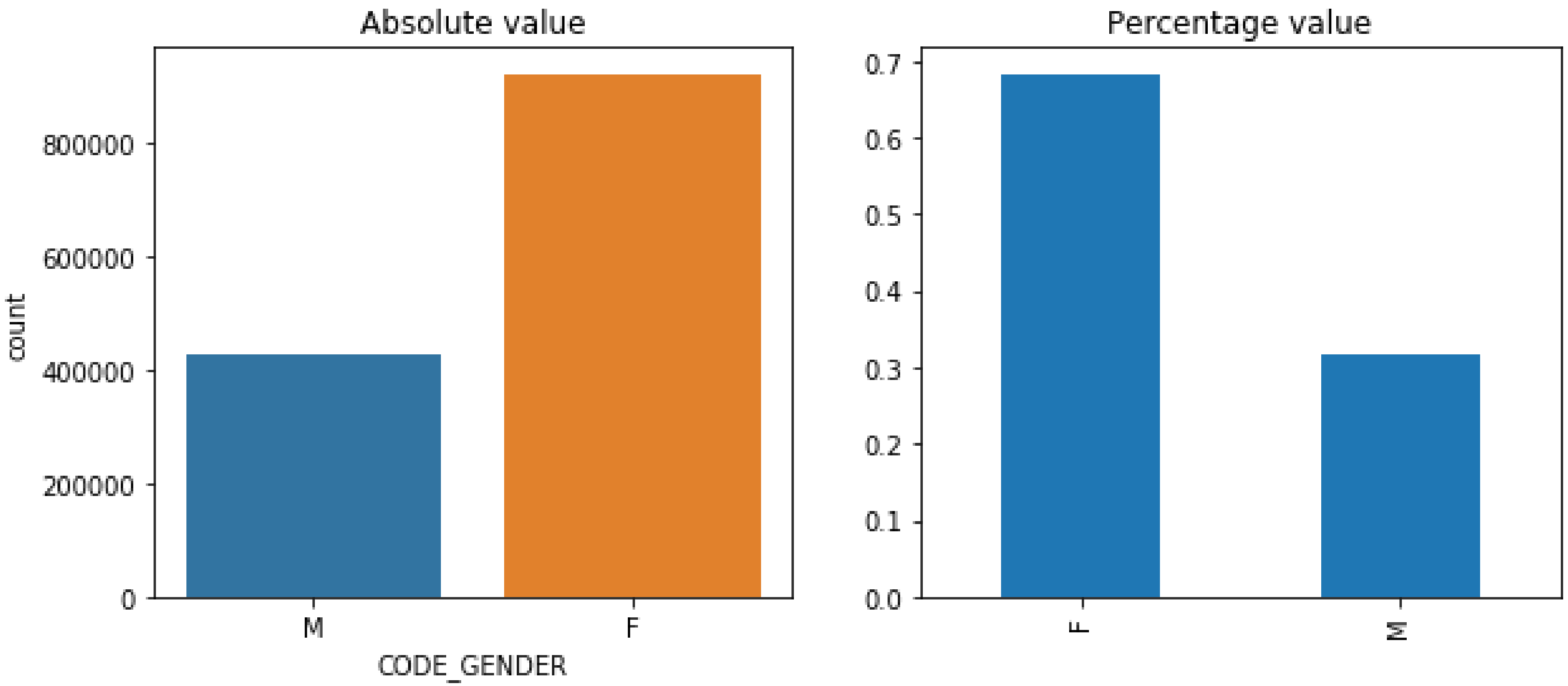
- As per the graphical representation, more number of people have preferred cash loans(92.38%) as compare to revolving loans(7.62%)

2) TARGET.



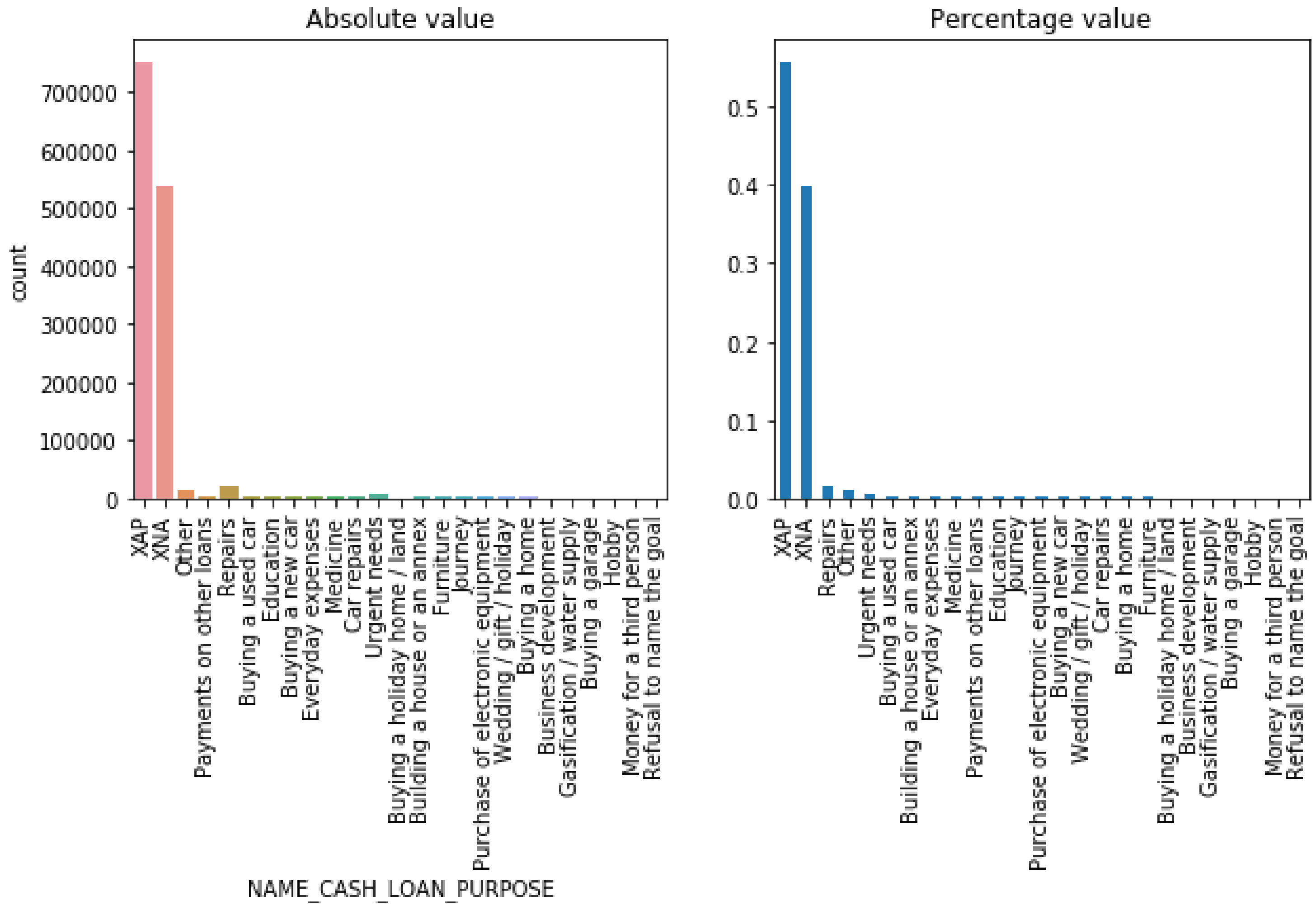
- As per the graphical representation, the clients with payment difficulties (Target value =1) is less (8.72%) than rest (Target value =0) having no problem with payment issues (91.28%).

3) CODE_GENDER



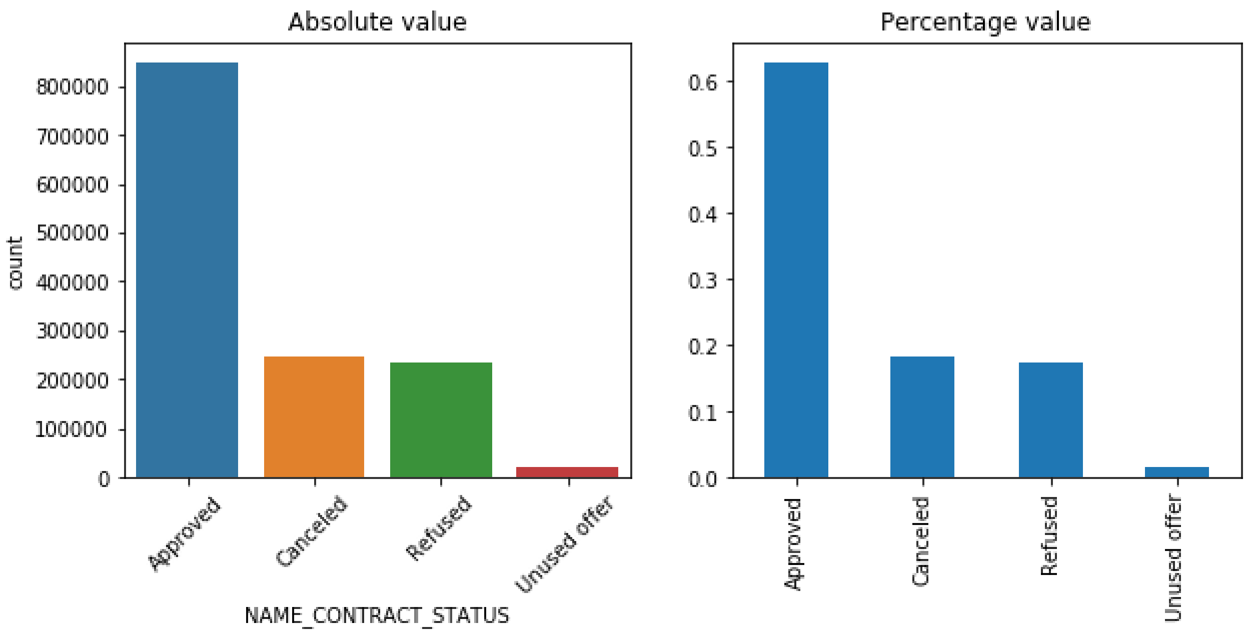
- As per the graphical representation, the rate of female clients(68.37) is double of male clients(31.63%).

4) NAME_CASH_LOAN_PURPOSE



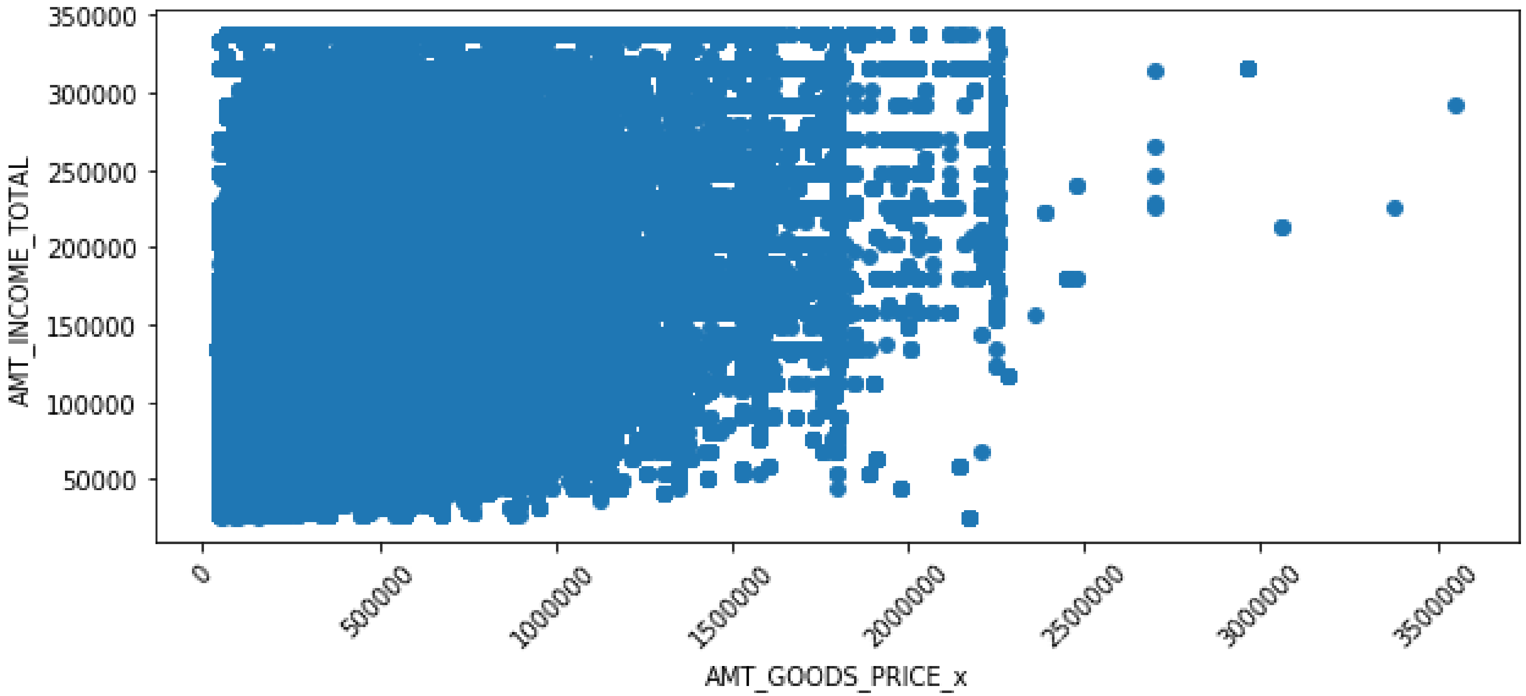
- As per the graphical representation, the purpose of taking loan for XAP amongs the client is higher(56%) than is XNA (40%) and so on

5) NAME_CONTRACT_STATUS



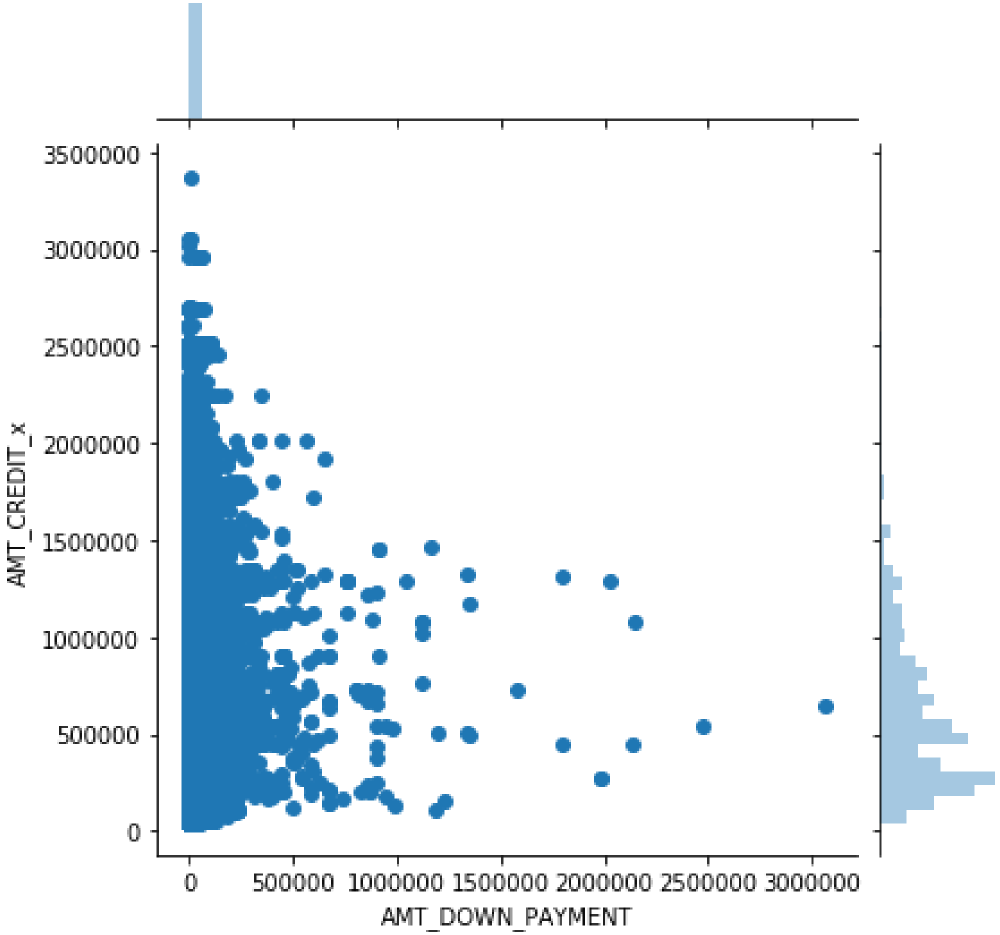
- As per the graphical representation, the contract status of clients are mostly approved(62.84%) and then canceled (18.30%) and refused(17.24%) and then unused (1.62%)

Performing bivariate analysis for continuous - continuous variables
1) 'AMT_GOODS_PRICE_x' & AMT_INCOME_TOTAL'



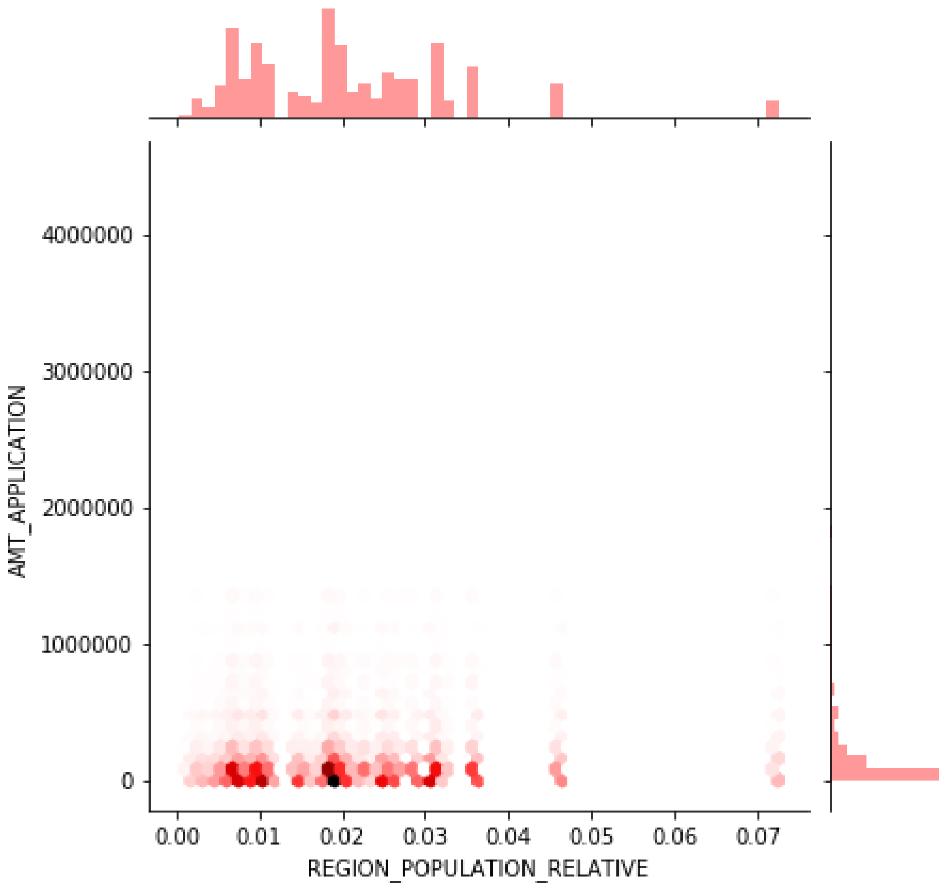
- As per the graphical representation, For consumer loans, more is the income of client more amount of price they have got for their goods

2) 'AMT_DOWN_PAYMENT', 'AMT_CREDIT_x



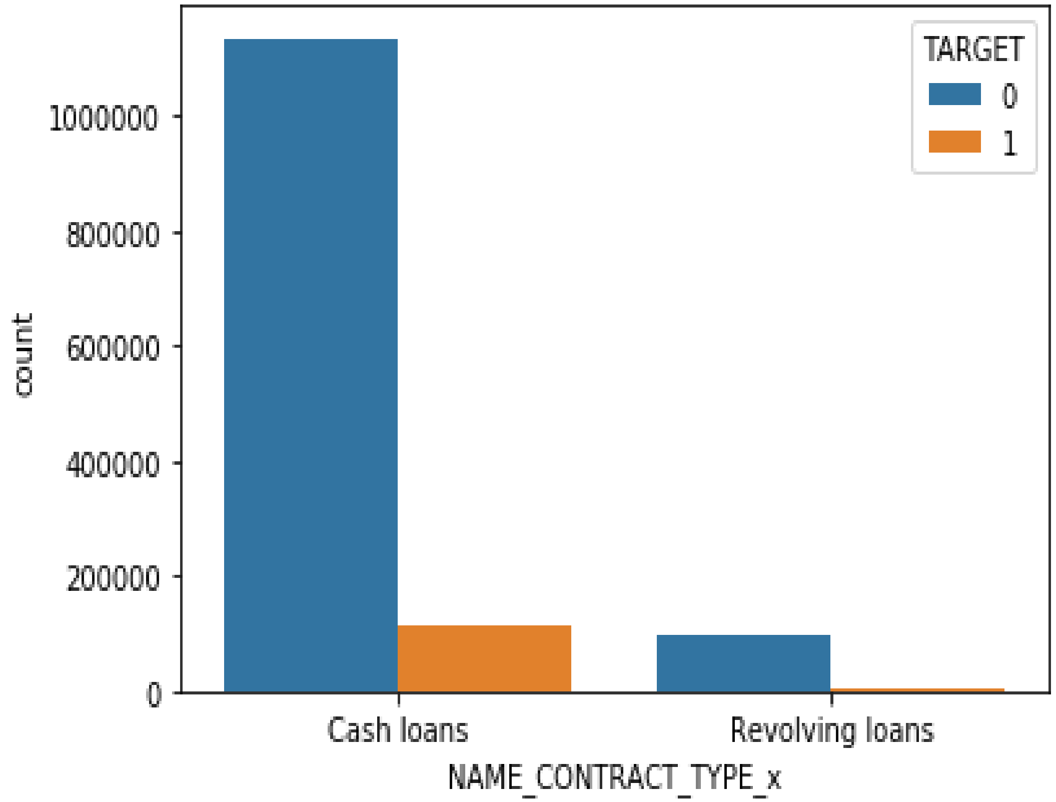
- As per the graphical representation, the amount of down payment done(0 to 4lakh) is more in the range of people who have got amount credited in the range of (0 to 5 lakh)

2) REGION_POPULATION_RELATIVE', 'AMT_APPLICATION'



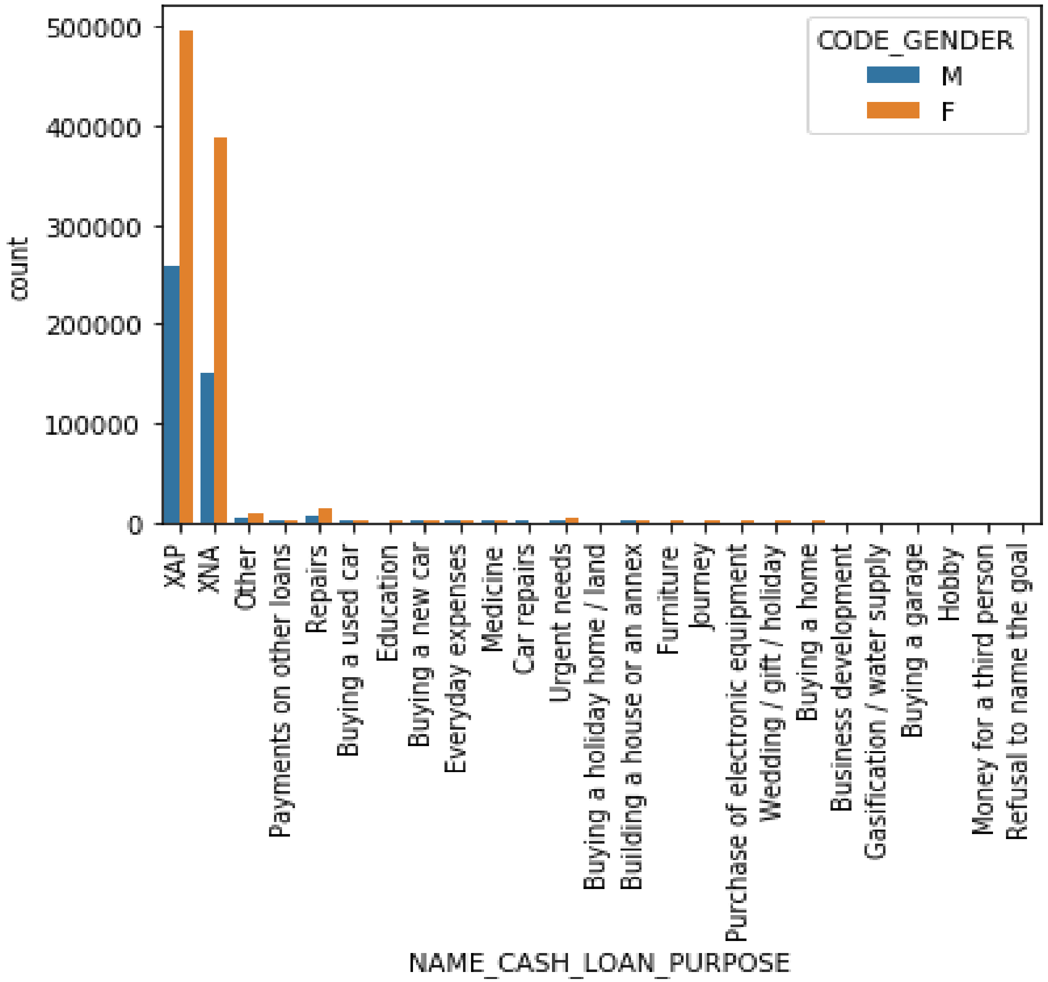
- As per the graphical representation, the amount of credit client ask on the previous application is more in the range of 0 to 4 lakhs and they are mostly from less populated area.

Performing bivariate analysis for categorical - categorical variables
1) 'NAME_CONTRACT_TYPE_x' AND 'TARGET'



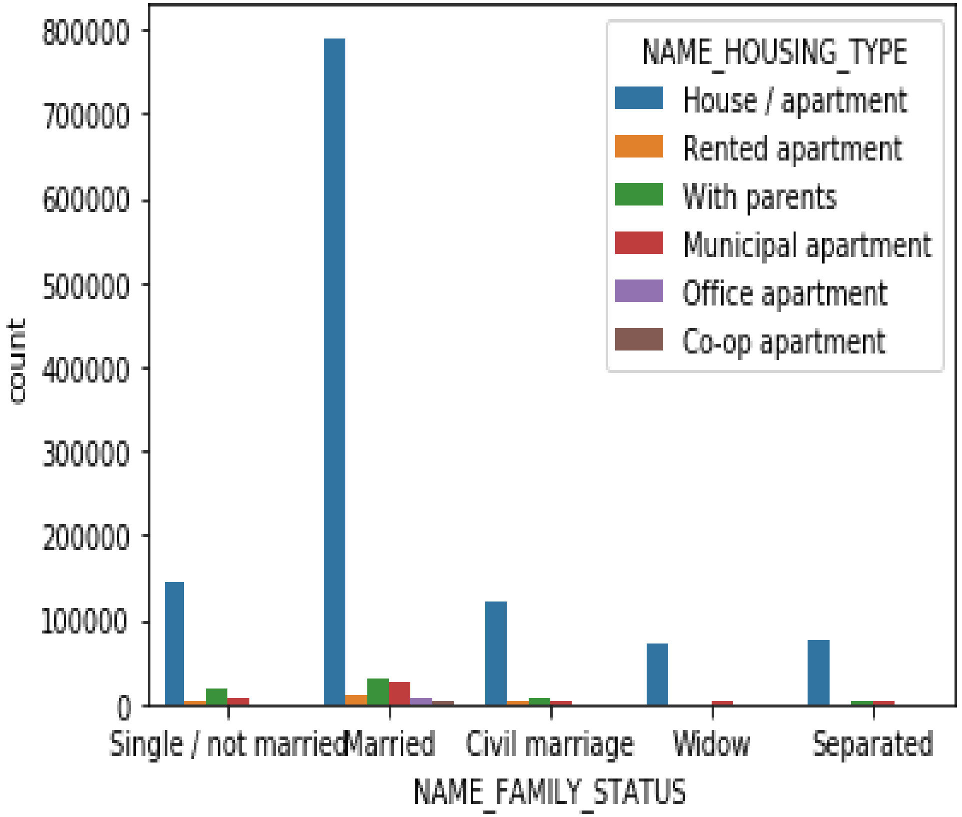
- As per the graphical representation, clients with payment difficulties and rest cases both preferred cash loans as compare to revolving loans.

2) 'NAME_CASH_LOAN_PURPOSE' and 'CODE_GENDER'



- As per the graphical representation, both males and females purpose for taking loan is mostly for XAP and then XNA and so on

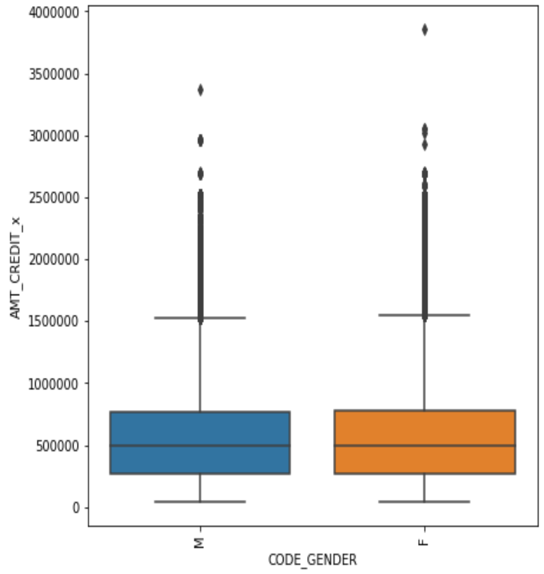
3) 'NAME_FAMILY_STATUS' and 'NAME_HOUSING_TYPE'

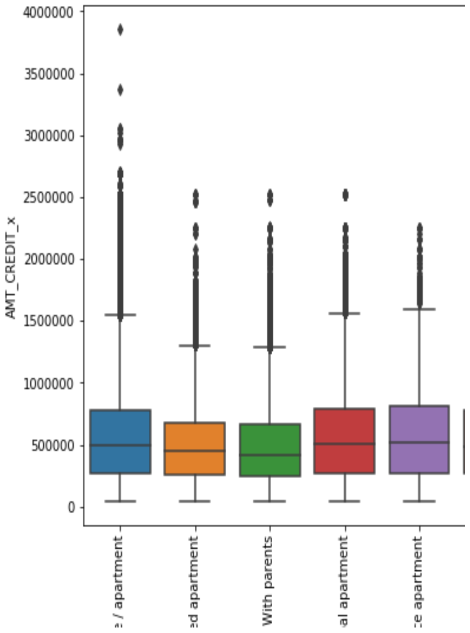


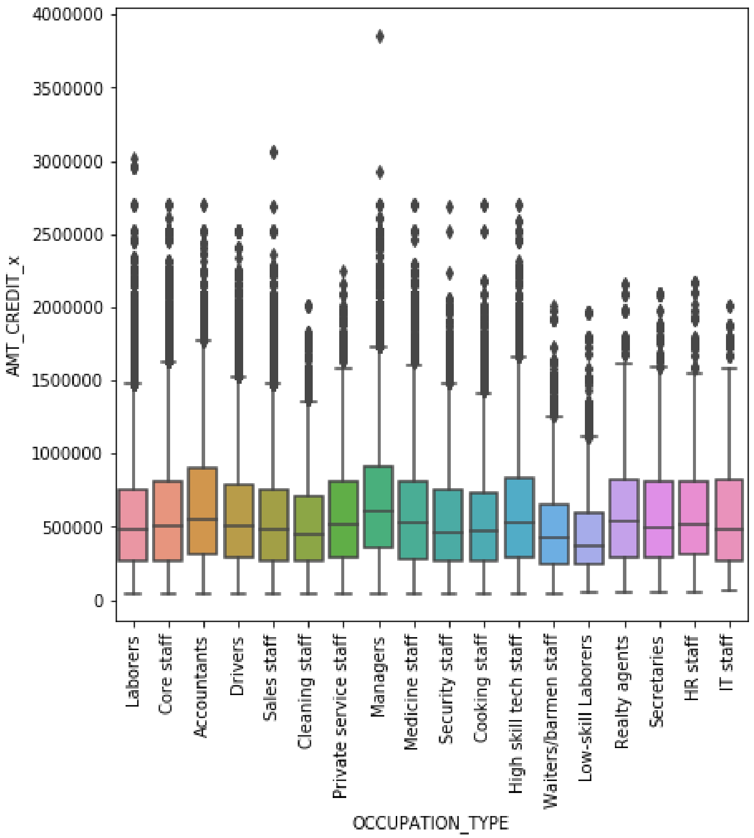
- As per the graphical representation, more number of married person are living in house apartment type then with their parents and in municipal apartments.

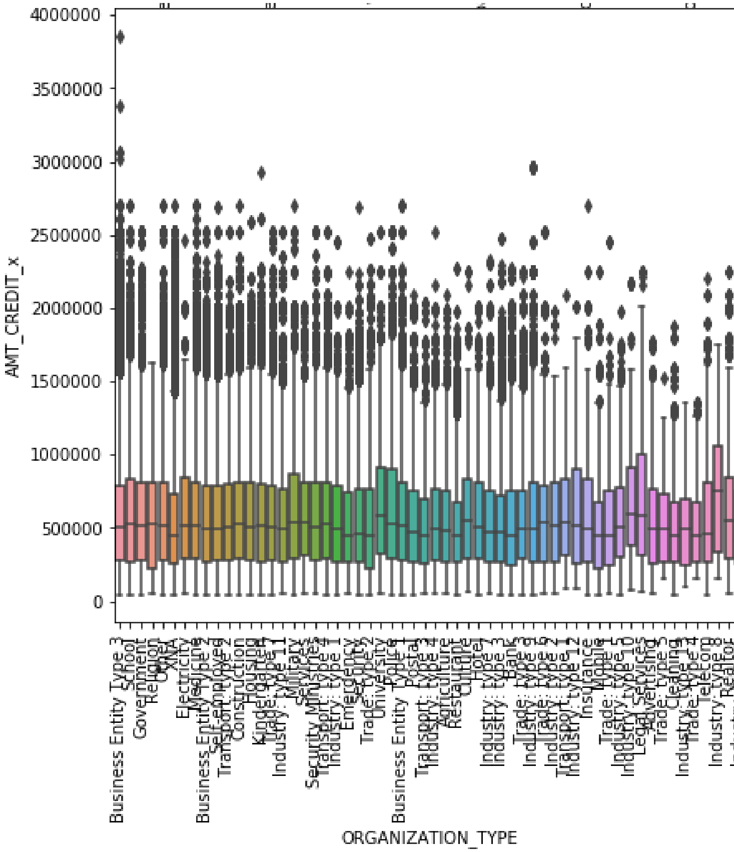
Performing bivariate analysis for categorical - continuous variables¶

Categorical = 'CODE_GENDER','NAME_HOUSING_TYPE','OCCUPATION_TYPE','ORGANIZATION_TYPE'
Continuous = 'AMT_CREDIT_x'



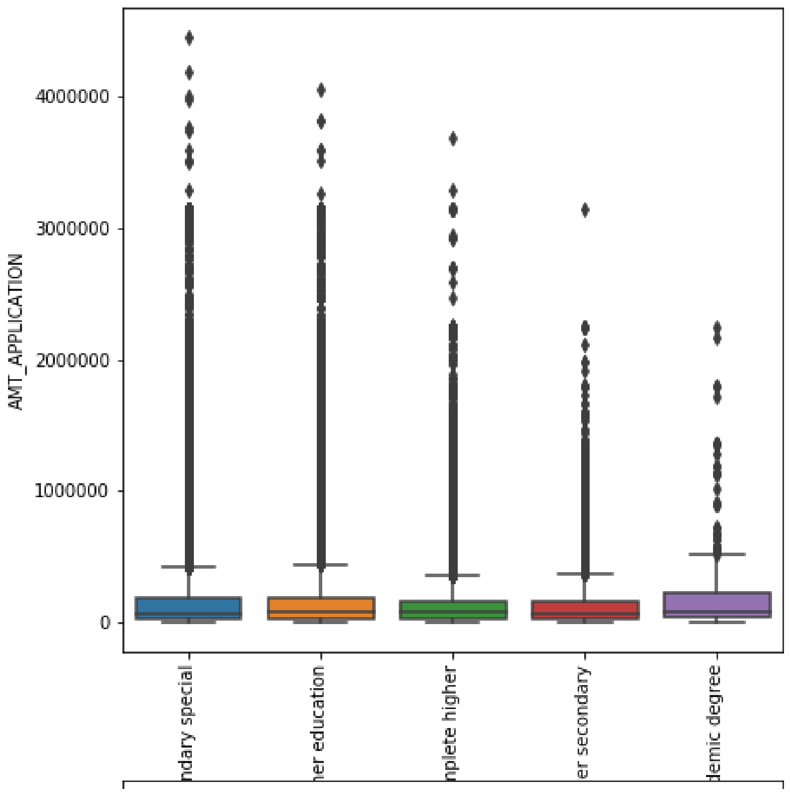


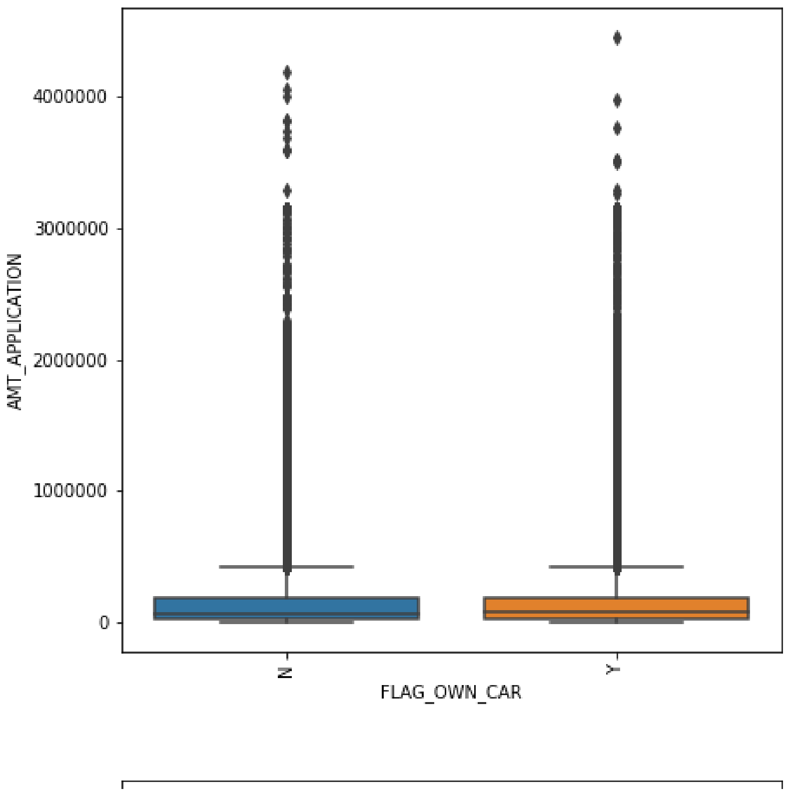


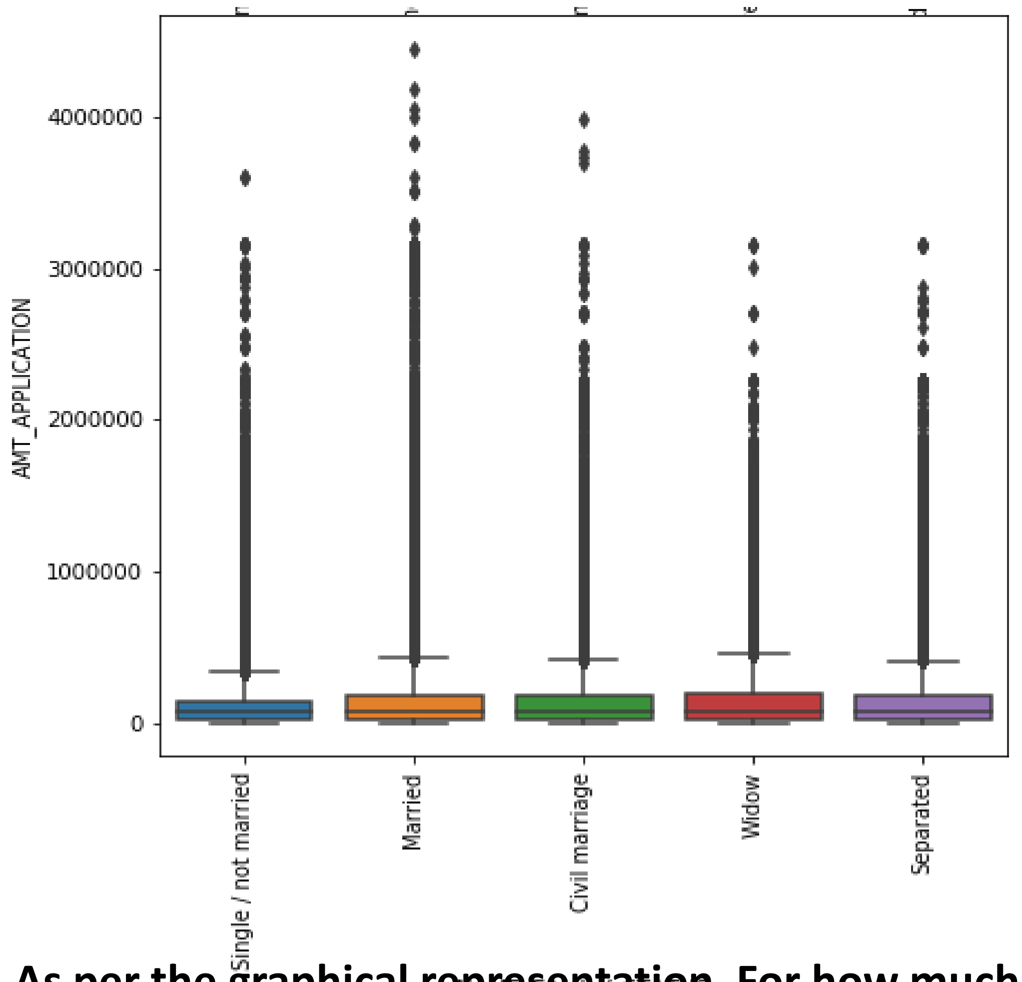


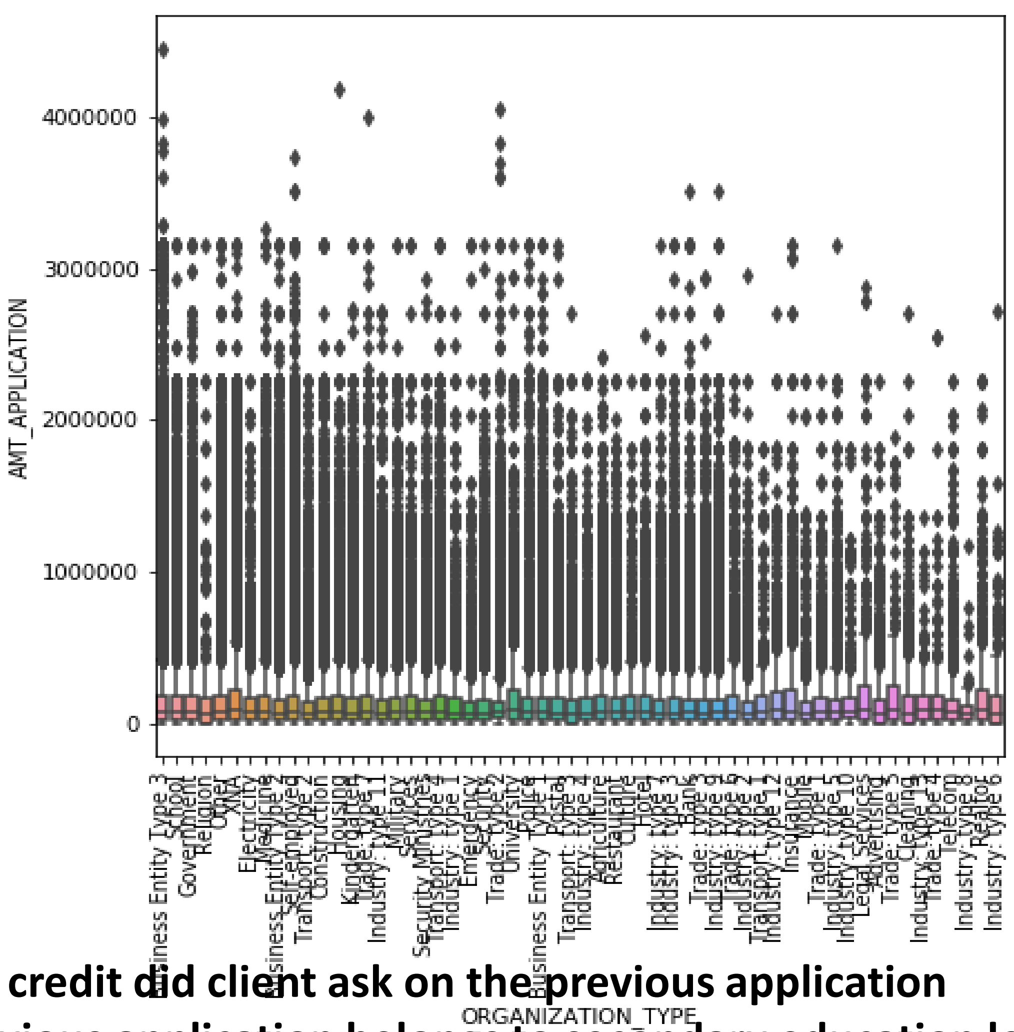
- As per the graphical representation,
1) The loan amount credited amongst Females is higher then males.
2) The loan amount credited amongst the people living in apartment type house is higher than the others
3) The loan amount credited amongst the managers is higher than the others
4) The loan amount credited amongst the business entity type 3 is higher than the others

Categorical = 'NAME_EDUCATION_TYPE','FLAG_OWN_CAR','NAME_FAMILY_STATUS','ORGANIZATION_TYPE'
Continuous = 'AMT_APPLICATION'









As per the graphical representation, For how much credit did client ask on the previous application

1) The larger amount of credit client ask on the previous application belongs to secondary education level as compare to others.

2) The larger amount of credit client ask on the previous application are having thier own cars

3) The larger amount of credit client ask on the previous application are mostly married and then are of civil marriage.

4) The larger amount of credit client ask on the previous application belongs to business entity type 3 than the others.