CREDIT EDA Assignment

- Poojitha Sigiri

First Dataset Explanation

- The first data set contains information of the client at the time of application of current application
- The data tells whether the client had payment difficulties

Data Cleaning

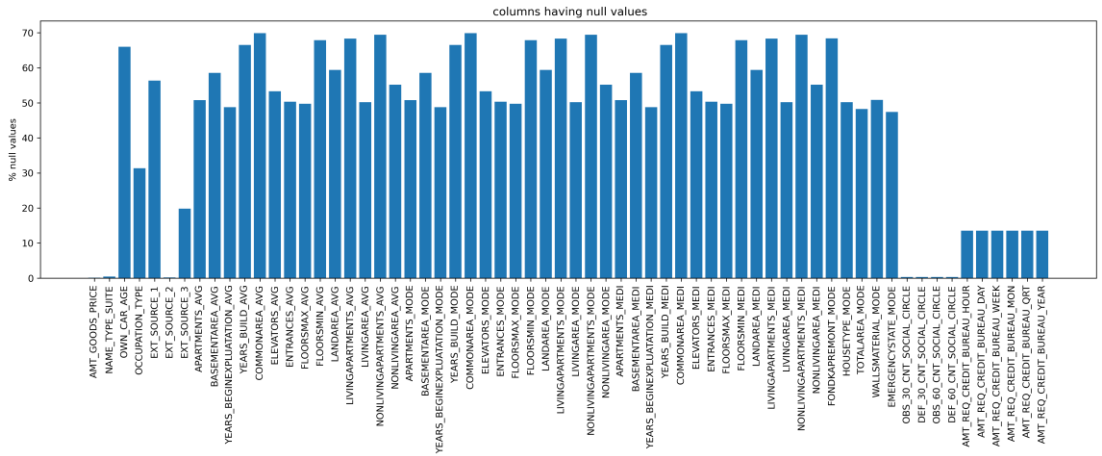
- Data cleaning is the major step in analyzing any data , as it plays the vital role in selecting the correct data in the right format.
- Data cleaning includes majorly few steps below:
  - Handling Null values or missing values .This can be either dropping the columns or imputing them with central tendency values(mean, median, mode)
  - Next step includes data correction , that is to make sure that the variables are in correct format.
  - Checking for the outliers and capping them to maximum range using the quartiles.

Data cleaning in the application data set

- **Checking for Null or missing values:**
  - To check if there are any duplicate data by using "df.duplicated().sum().
  - This will return of there are any duplicate rows
  - Next Is to find percentage of Null or NaN values in each column
  - (df.isna().mean()*100) this gives the percentage of null values in each column

Results of finding Null values

- The Null values over 50% are need to be dropped
- Here the Column with null values above 50% are around :41
- Here we can drop them as df.drop().
- After that columns with null values less than 13% are identified and imputed with relevant values



Columns with null value % less than 13.

- The columns that are imputed with values are :
- They are seven in number.

AMT_GOODS_PRICE , NAME_TYPE_SUITE , EXT_SOURCE_2 ,

OBS_30_CNT_SOCIAL_CIRCLE ,DEF_30_CNT_SOCIAL_CIRCLE ,

OBS_60_CNT_SOCIAL_CIRCLE ,DEF_60_CNT_SOCIAL_CIRCLE,

Handling imputations

- Out of them the categorical columns are handling them by imputing them with higher frequency values (mode)
- The numerical columns are analyzed and imputed with relevant values .(mean(),median(),mode())

Removing unwanted Columns.

- The columns which doesn't add any value or removing them wont affect anything in analysis.
- Had removed almost number of columns to reduce Dataframe size and make it easy for analysis
- We can list all those columns in a list and do below
- Have dropped about 37 columns , which doesn't add that much of impact to analysis.

List of columns that are dropped as they are unwanted

['FLAG_DOCUMENT_2',
'FLAG_DOCUMENT_3', 'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5',
'FLAG_DOCUMENT_6', 'FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8',
'FLAG_DOCUMENT_9', 'FLAG_DOCUMENT_10', 'FLAG_DOCUMENT_11',
'FLAG_DOCUMENT_12', 'FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14',
'FLAG_DOCUMENT_15', 'FLAG_DOCUMENT_16', 'FLAG_DOCUMENT_17',
'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20',
'FLAG_DOCUMENT_21','FLAG_MOBIL','FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE',
'FLAG_EMAIL','CNT_FAM_MEMBERS','REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY','FLOORSMAX_AVG',
'YEARS_BEGINEXPLUATATION_MODE', 'FLOORSMAX_MODE',
'YEARS_BEGINEXPLUATATION_MEDI', 'FLOORSMAX_MEDI', 'TOTALAREA_MODE',
'EMERGENCYSTATE_MODE','YEARS_BEGINEXPLUATATION_AVG',]

Data Correction

- Data correction includes setting the datatypes in correct format like numerical values to be in int or floats, dates in datetime and others in object
- The numeric values which are to be positive they should be in positive values
- The below columns are in negative , they should be change to positive using .abs()
- There are few columns with XNA values .
- Considering CODE_GENDER , OCCUPATION_TYPE

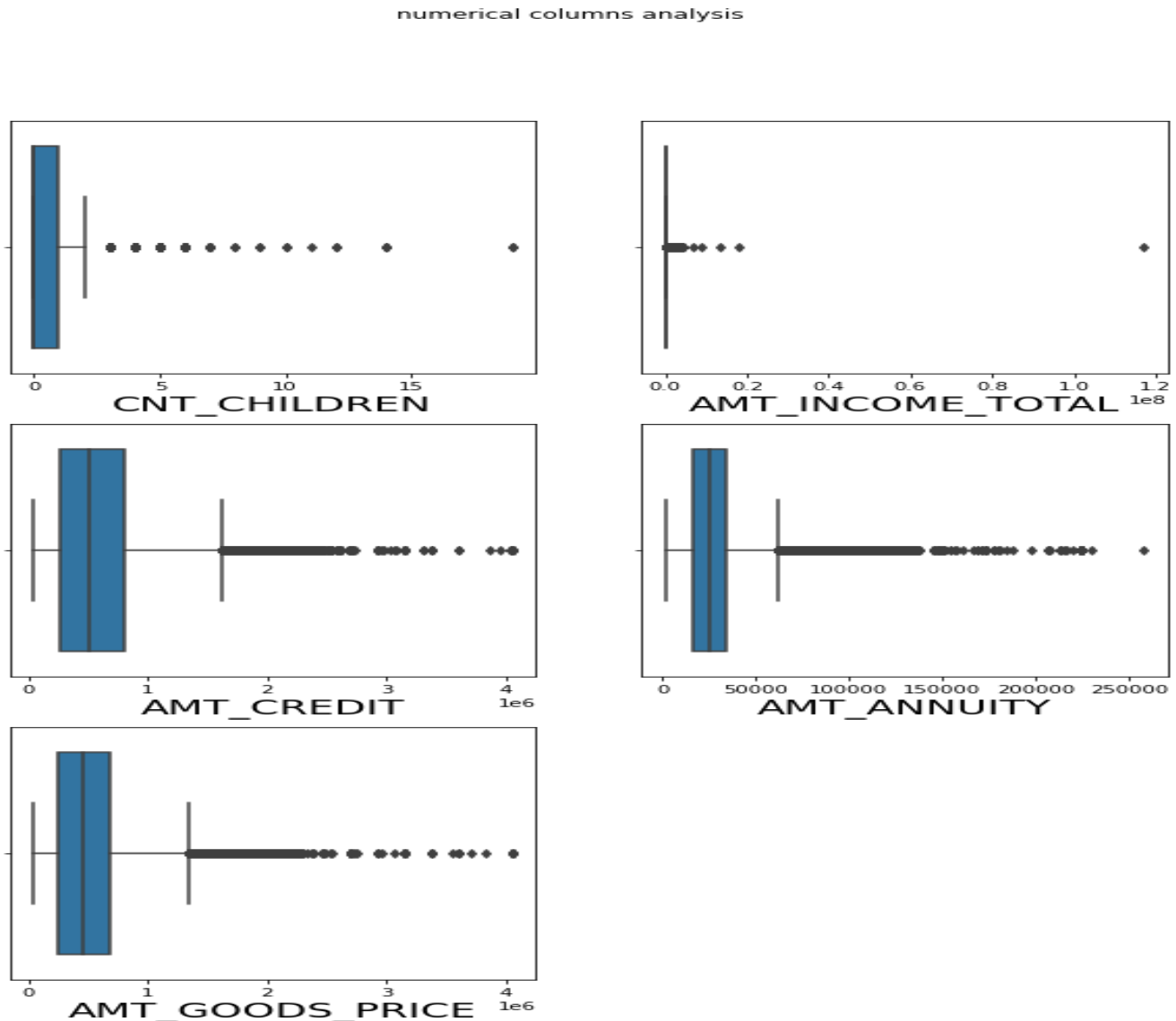
Handling 'XNA' values.

- For code gender the 'XNA' percentage is about very less only few values .So it can be predicted that they could be of the higher frequency values as there are only two values 'M' and 'F'
- For the other column , the XNA percentage is 18% and there are many values .
- So , it is better to replace them with NaN values

Outliers for numerical variables

- The numerical variables we got are:
  - AMT_INCOME_TOTAL , CNT_CHILDREN
  - AMT_CREDIT,AMT_ANNUITY,
  - AMT_GOODS_PRICE
- These are plotted with boxplots and outliers are analysed

Subplots and boxplots for numerical variables outliers.



Handling Outliers.

- The outliers are handles by capping the to max range that is the upper whisker
- The upper whisker is found by below process

```
q1 = app_df2.AMT_INCOME_TOTAL.quantile(0.25)
q3 = app_df2.AMT_INCOME_TOTAL.quantile(0.75)
iqr = q3-q1
upper_whisker = q3 + 1.5*iqr
lower_whisker = q3-1.5*iqr
```

Binning

- The process of converting a continuous column to categorical by diving the column range values into categories
- Here two values are chosen which can be done:
- AMT_CREDIT,AMT_INCOME_TOTAL
- And they are divided based on the quantile range using pd.qcut()
- These values are put into new columns with out altering original ones

Dividing the dataset on particular value

- Here if we observe the data set has a column TARGET with values 0 and 1
- 0- represents the clients who are Non defaulters
- 1- clients who are defaulters facing difficulties in loan repayment
- So here clients are divided based on this
- So , we can create two new datasets based on target values
- Data sets created are: app_Target0, app_Target1

Univariate Analysis

- Univariate analysis done for single columns
- This can be for categorical or numerical
- Analysis done for the columns below:

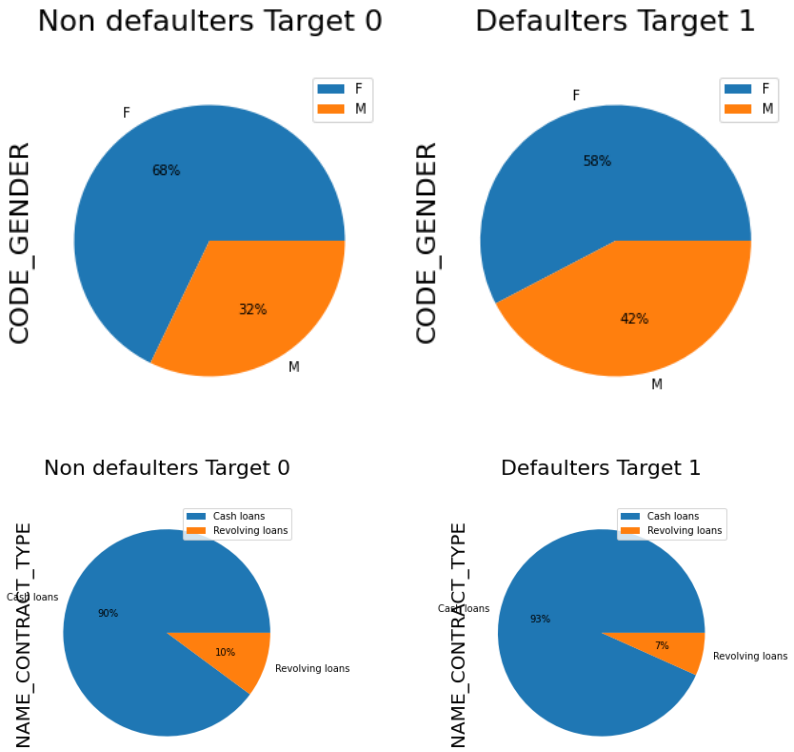
NAME_CONTRACT_TYPE,CODE_GENDER,NAME_TYPE_SUITE,NAME_INCOME_TYPE,NAME_EDUCATION_TYPE,

NAME_FAMILY_STATUS,NAME_HOUSING_TYPE,ORGANIZATION_TYPE

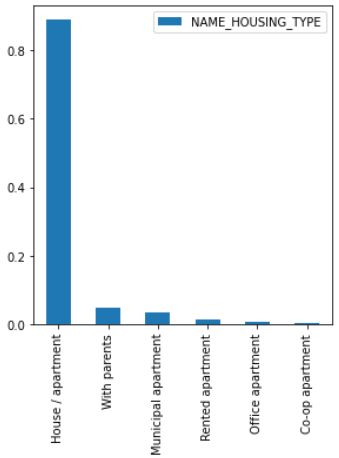
Univariate Analysis

- This shows that in both the cases , For defaulters and Non defaulters, Female percentage is more.

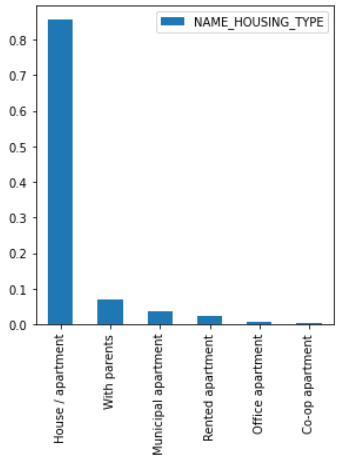
- For the CONTRACT_TYPE , The Cash loans percentage increases or is more in defaulter's category and revolving loans is less, So it can be concluded that in case of defaulters most of them are in contract type - 'Cash loans'



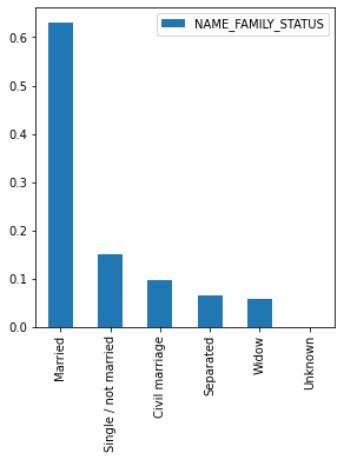
Non defaulters Target 0



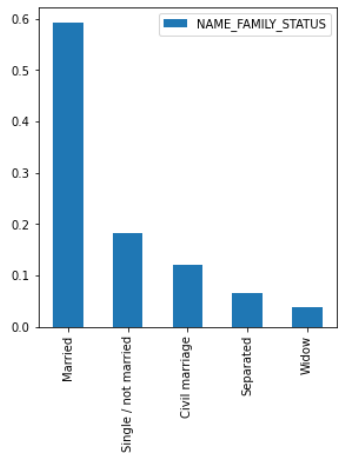
Defaulters Target 1



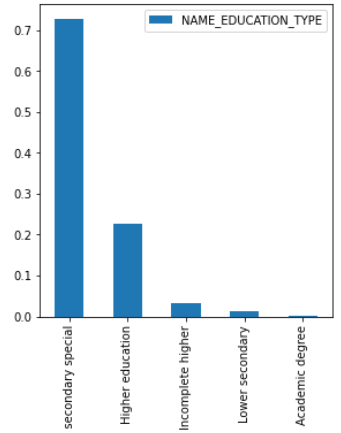
Non defaulters Target 0



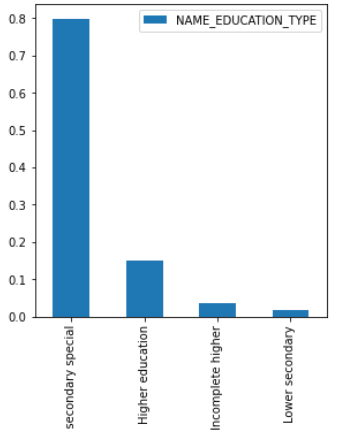
Defaulters Target 1



Non defaulters Target 0



Defaulters Target 1



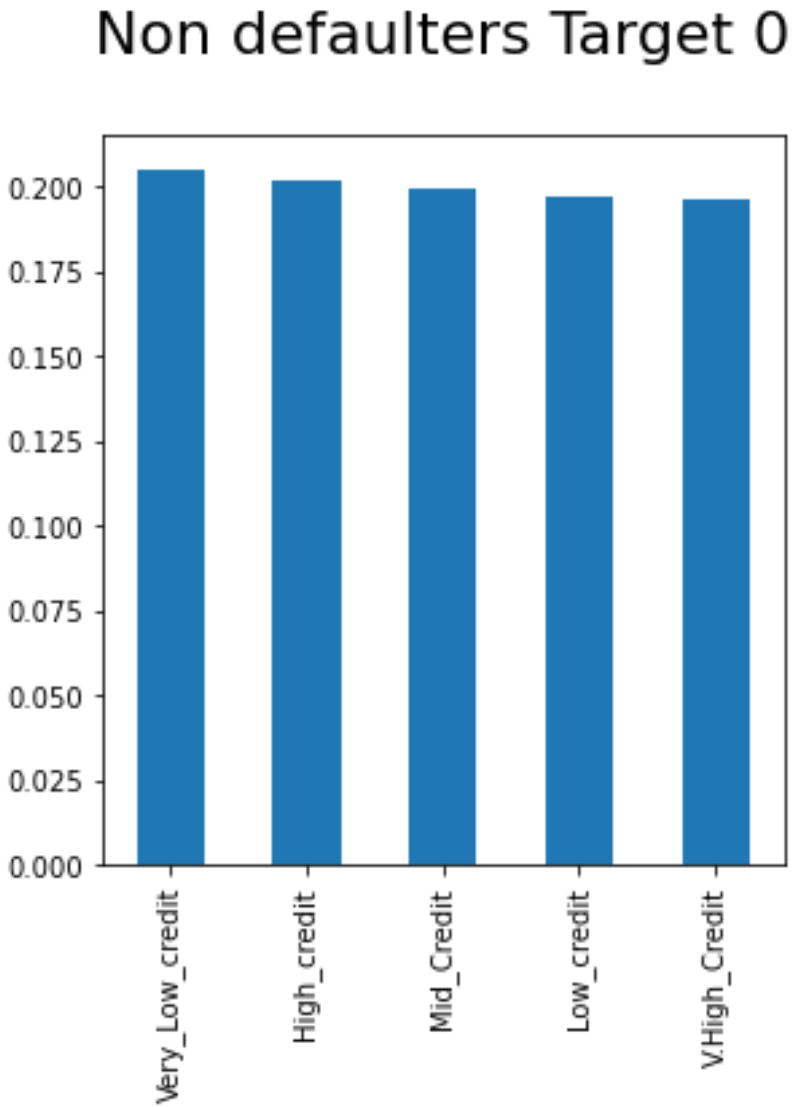
Univariate Numerical Analysis

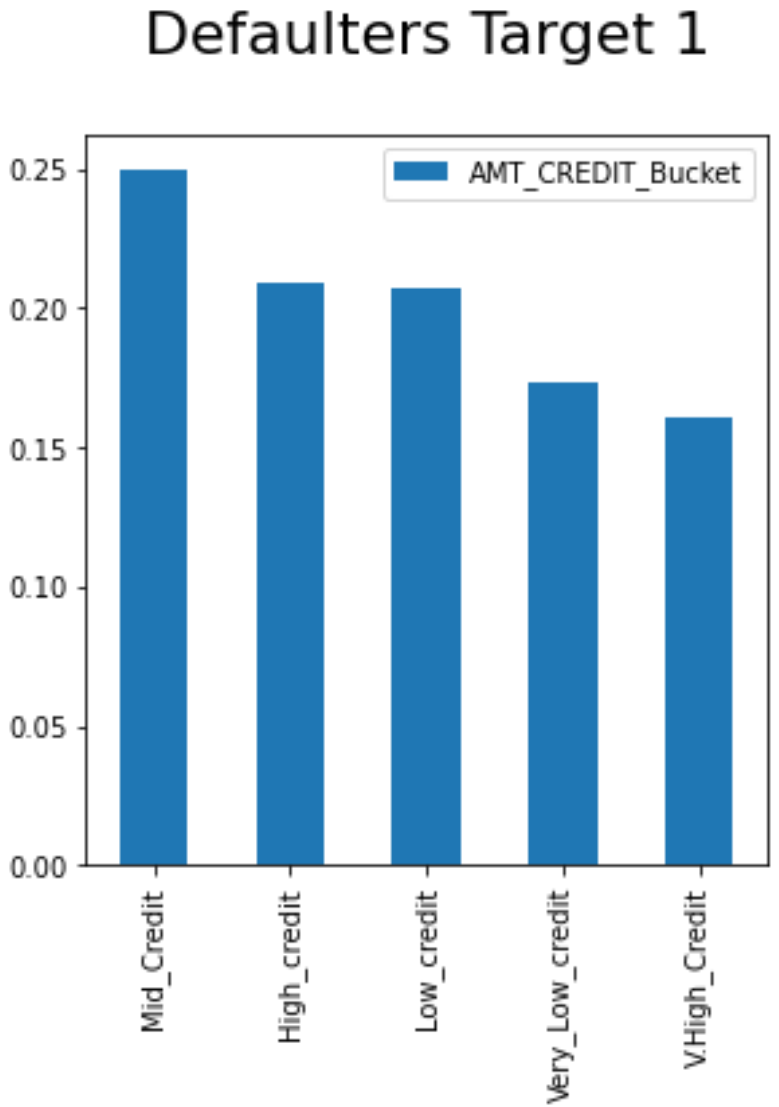
- # From the above observation , there is an increase in working category and decrease in others in Defaulter's case.
- # So we can here deduct there is a possibility that people who are working income type can be defaulters.

Univariate-Numerical

from the above observations we can see that there is high credit value for mid_credit in defaulters set, also

And the other ranges seems to be decreased.





Correlation Between Numerical Columns

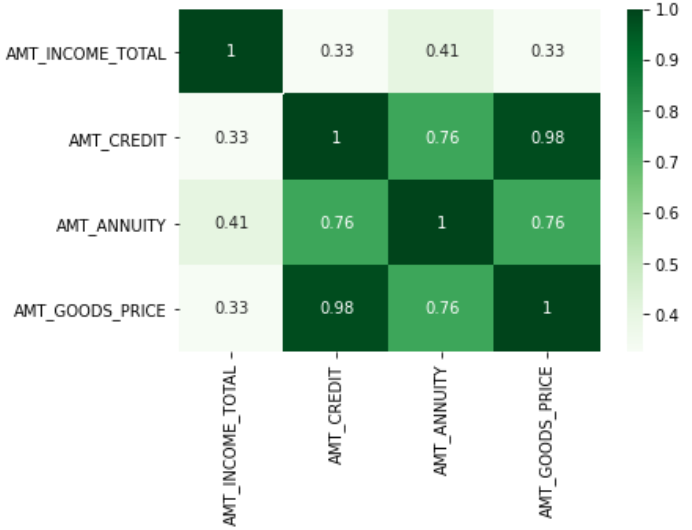
- correlation for Targeto set

- first the highest correlation is between themselves but apart from that

- we can see there is high correlation between AMT_CREDIT and AMT_INCOME_TOTAL - 0.98

- Next comes AMT_ANNUITY vs AMT_GOODS_PRICE , AMT_ANNUITY vs AMY_CREDIT -0.76

- and the least seems to be between AMT_INCOME_TOTAL vs AMT_INCOME_TOTAL and AMT_INCOME_TOTAL vs AMY_CREDIT - 0.33



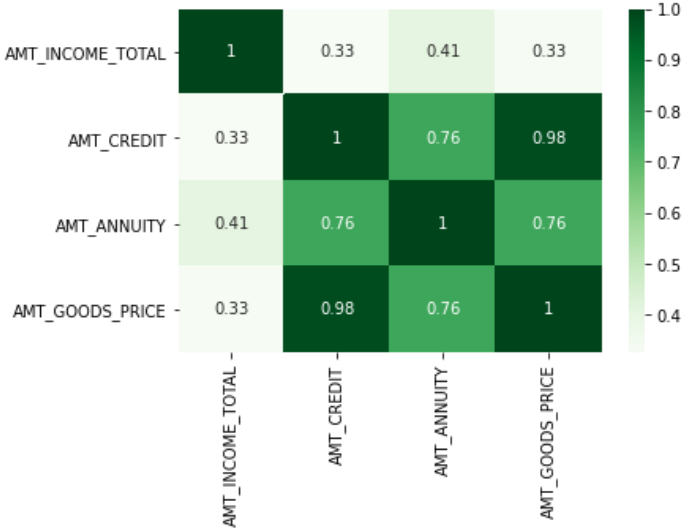
- correlation for Target1 set - correlation follows same trend but slightest difference in values

- first the highest correlation is between themselves but apart from that

- - we can see there is high correlation between AMT_CREDIT and AMT_INCOME_TOTAL - 0.98

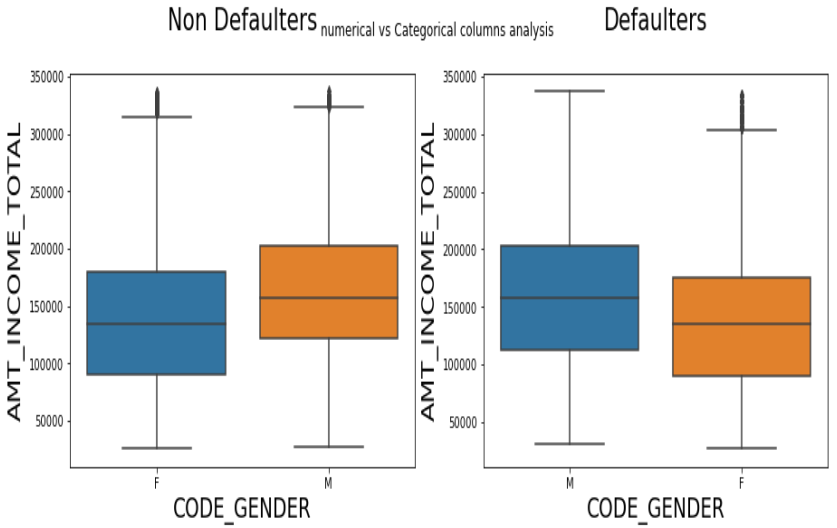
- - Next comes AMT_ANNUITY vs AMT_GOODS_PRICE , AMT_ANNUITY vs AMY_CREDIT -0.74

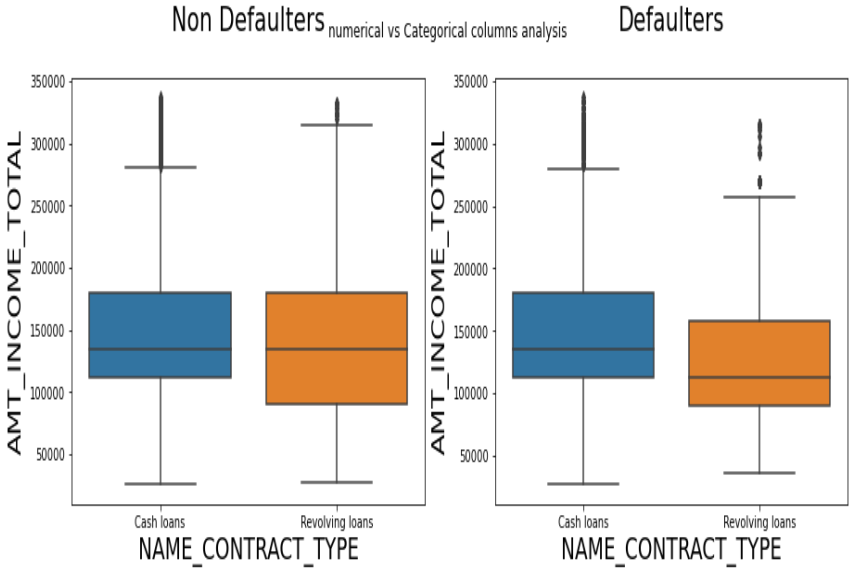
- - and the least seems to be between AMT_INCOME_TOTAL vs AMT_INCOME_TOTAL and AMT_INCOME_TOTAL vs AMY_CREDIT - 0.3



Bivariate Analysis

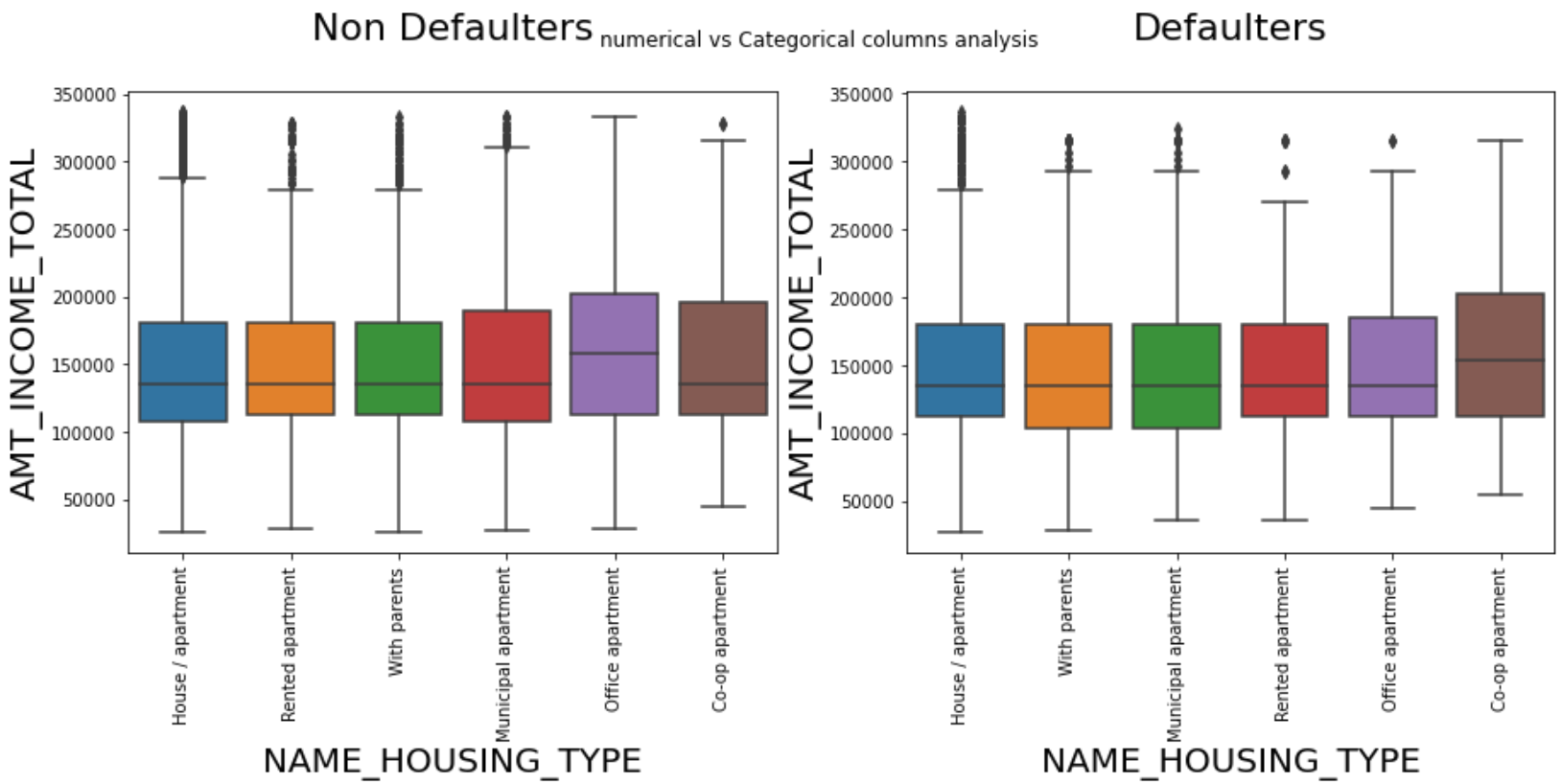
- female lower income are most part of defaulter
- for defaulters are mostly on lower income range - revolving loans and cash loand in higher quartile





Bi variate Analysis

people who live in co - apartment tends to increase with higher quartile range in income for defaulters



Previous_application Dataset

- This is the dataset with details of client with previous application , their status , reasons and other details.
- Here also we do the same data cleaning ,checking for null values and dropping above 50%
- The number of columns dropped are below

['AMT_DOWN_PAYMENT', 'RATE_DOWN_PAYMENT',
'RATE_INTEREST_PRIMARY',

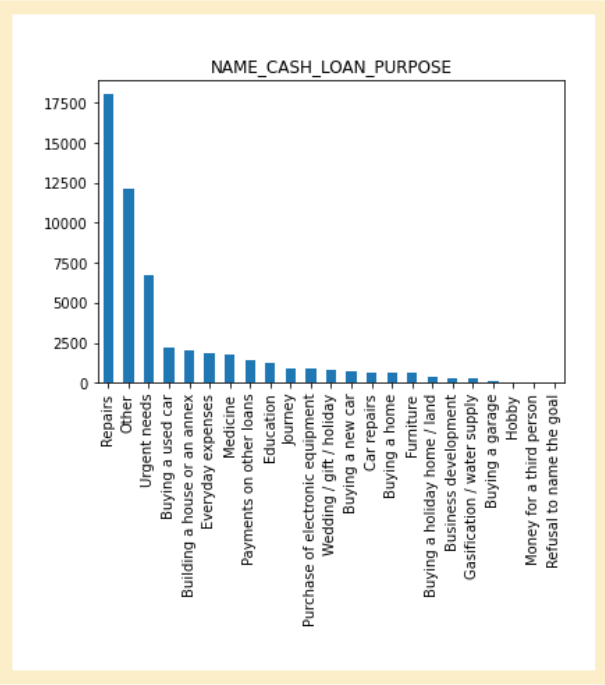
'RATE_INTEREST_PRIVILEGED']

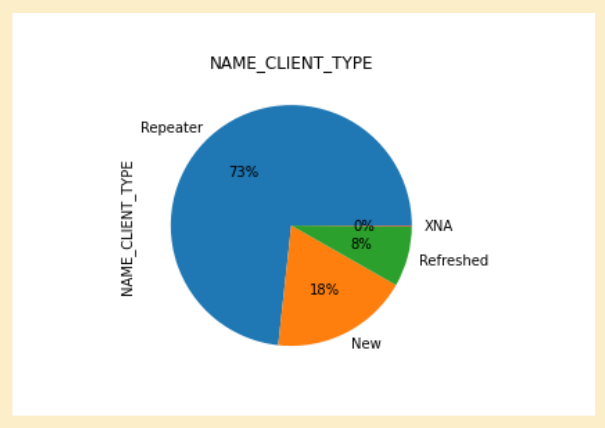
Merging two data frames

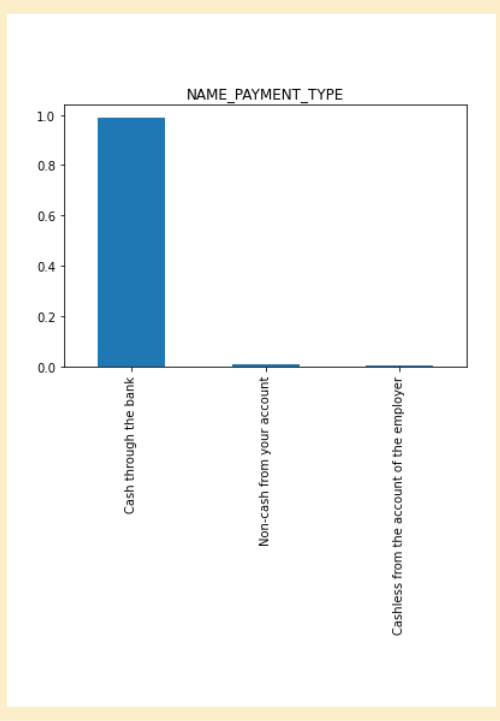
- We have to merge the two data frames application and previous application
- Using pd.merge and inner join on ‘SK_ID_CURR’ which is common in both
- After merging shape should be checked
- Here also we can remove unwanted columns

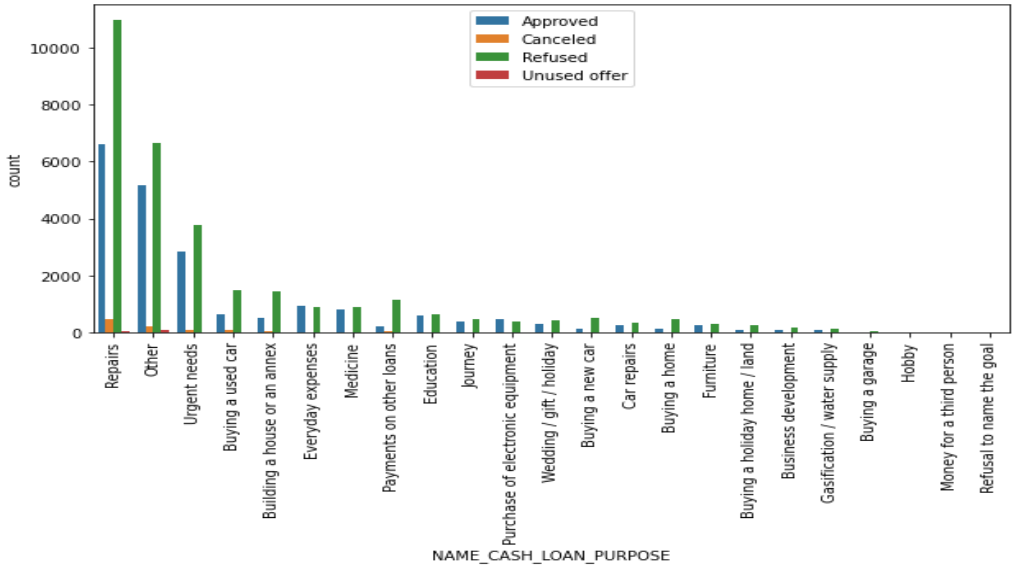
Univariate analysis for combined dataset.

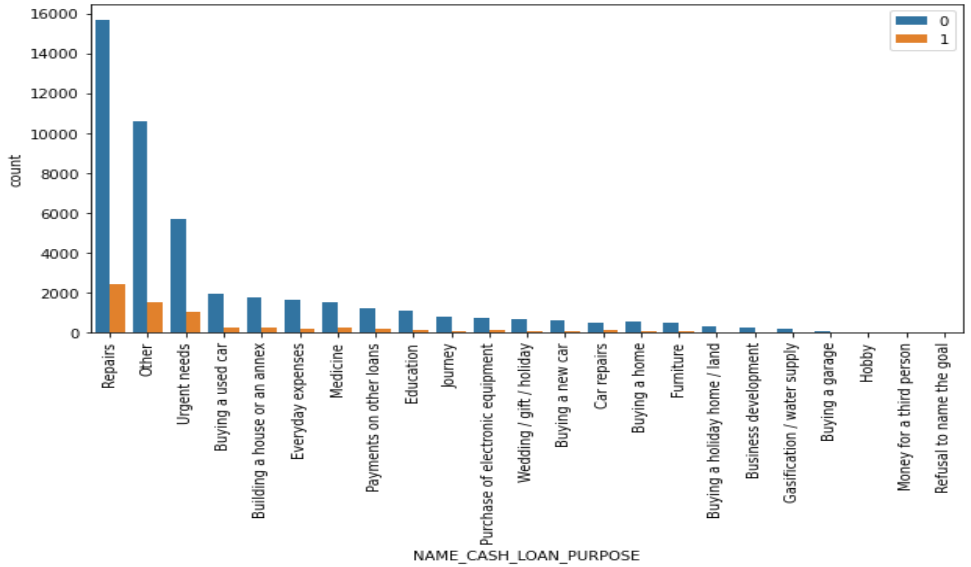
- If we observe here:
- In 'NAME_CASH_LOAN_PURPOSE , the frequency of repairs is more.
- For NAME_CLIENT_TYPE , the Repeaters are more in number
- For NAME_PAYMENT_TYPE , the Repeaters are more in number









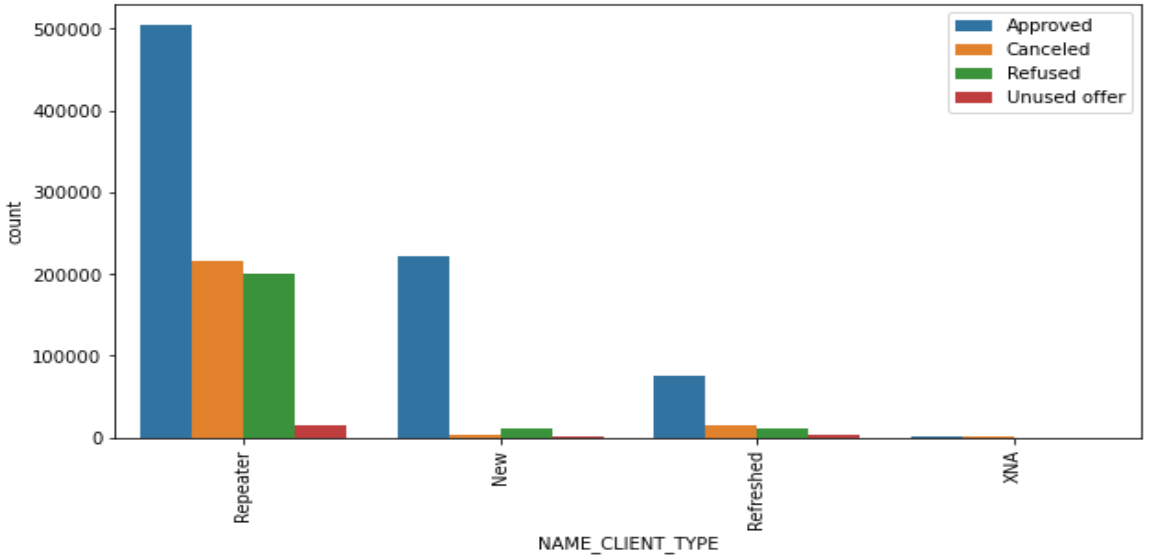


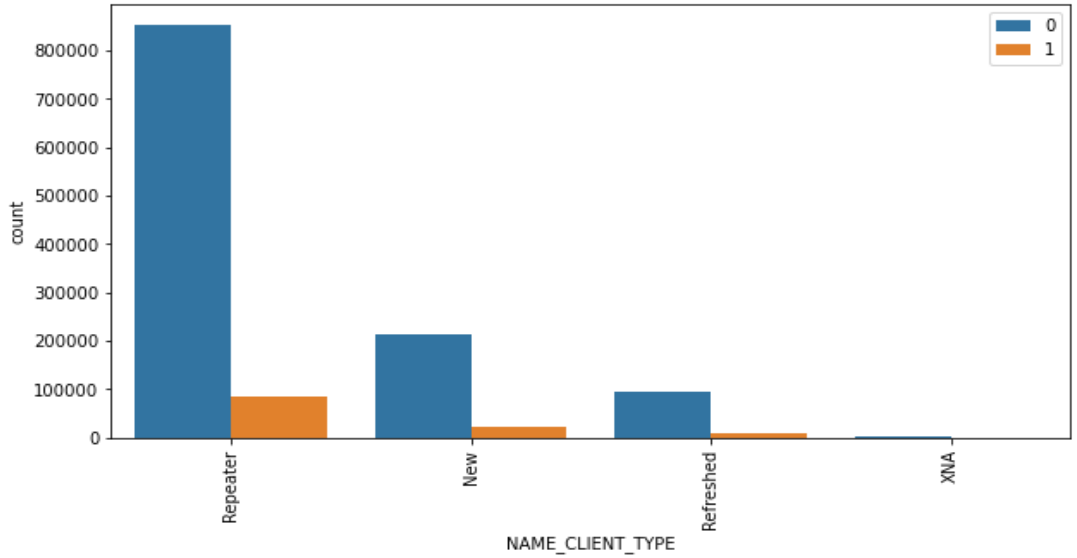
Bivariate Analysis

If we see here when we divided the cash loan purpose with target and status , in both cases

"Repairs" seems to be a possible defaulter or got refused application.

- And Buying Garage,Hobby seems to be opposite.

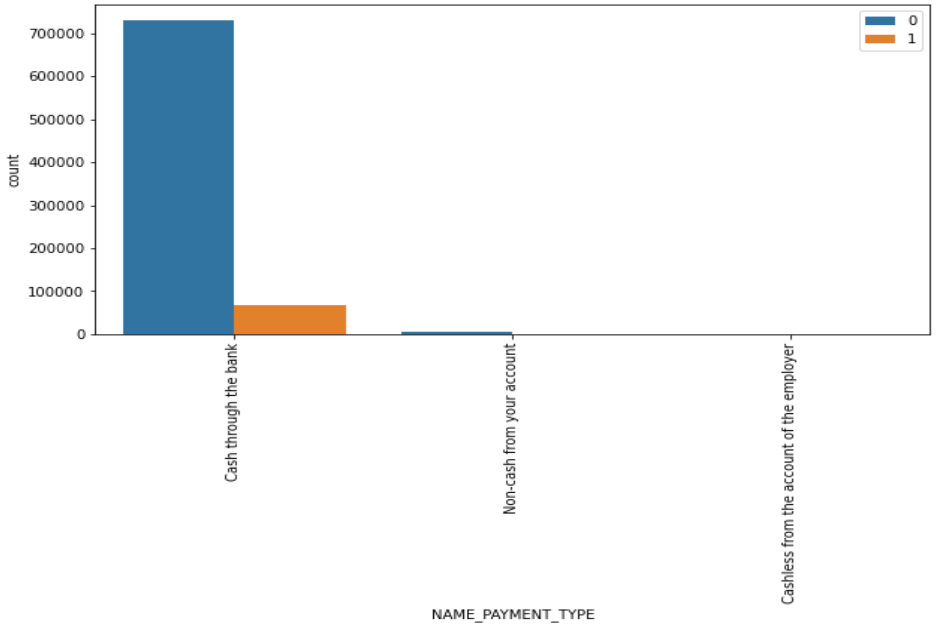


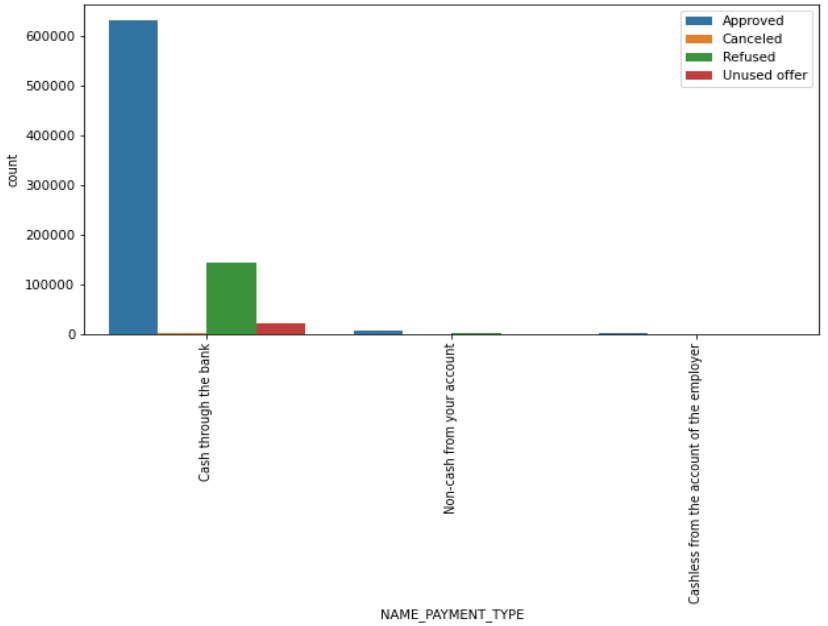


Bivariate Analysis

From the above observations , the client type ,

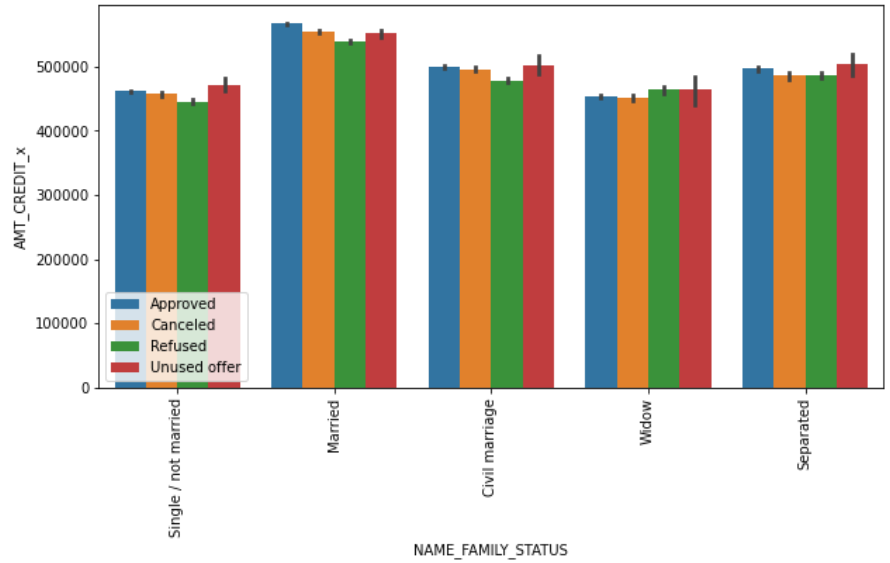
Repeater is comparatively , more in refusal and and more chance of defaulter

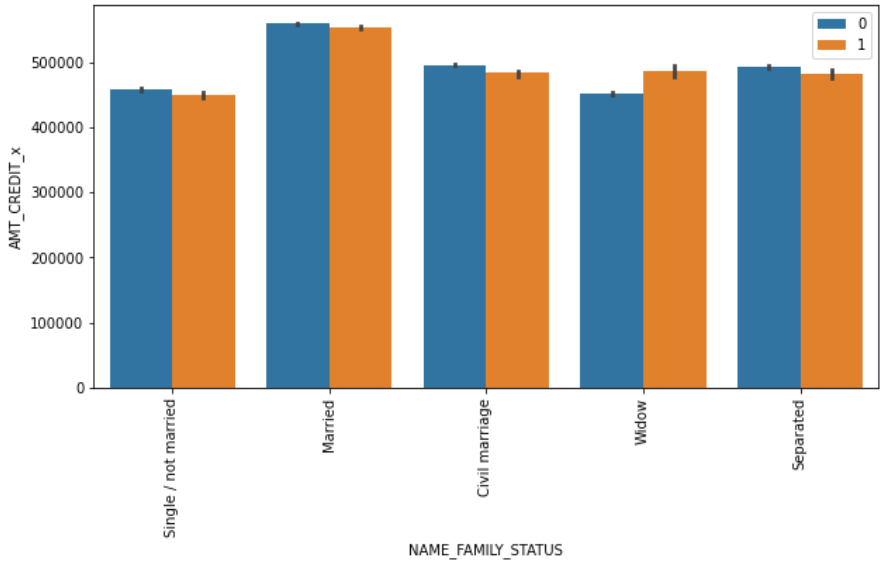




Bivariate Analysis

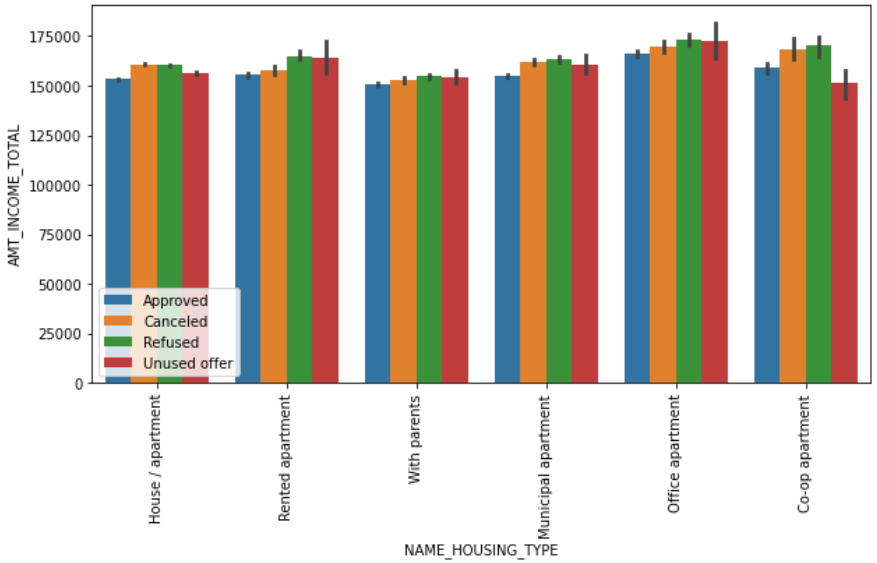
- Here for the 'NAME_PAYMENT _TYPE' , the clients with payment type as "cash through banks" , is more prone to be a defaulter or is a applicant with Refusal status.

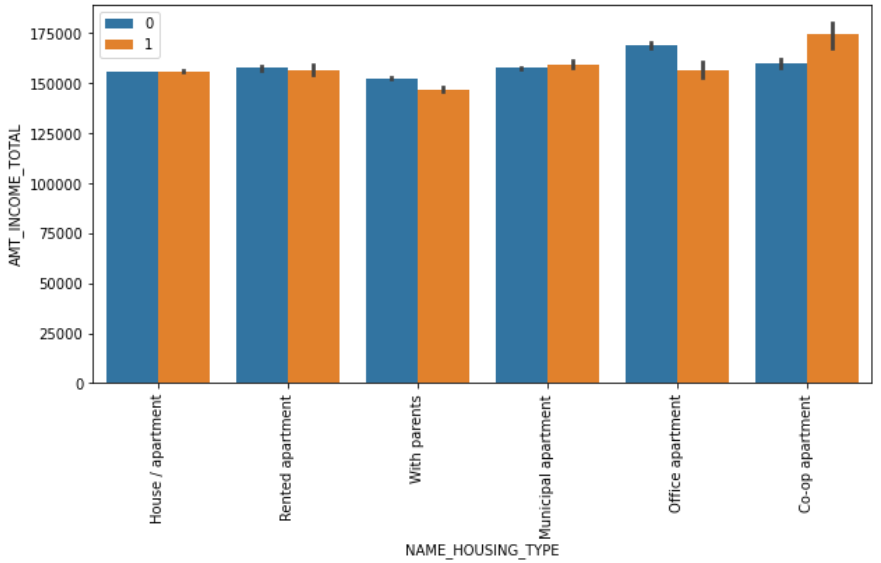




Bivariate Analysis

- Here for the 'NAME_FAMILY _STATUS' , the clients with family status as "Married" , is more prone to be a defaulter or is an applicant with Refusal status.
- Where as Single/not married seems to be successful player





Bivariate Analysis

- From the above observations , the clients with "co-apartment " housing type is more prone to defaulters.
- Where as House/apartment and with parents is less chance to fail to repay loans.

Final Conclusions and observations

- Primary observations
- - In gender we see that female are more in both cases , in defaulters female percentage is more may be with a little lower in come range comparitevely
- - And clients with contract type 'cash loans' also seems tend to be a defaulter, here also income ranges is a littel lower.
- - and considering the occupation - working categeory and educatinal qualification of seconday/secondary special seems to be defaulters

After combining two data sets

- Considering the loan purposes:
- - reason with repairs seems to fail to repay loan and be a defaulter and the loan application gets rejected
- - So banks/credit companies should avoid them
- - Purposes with buying Garage , Hobby , money for 3rd person seem to be less of a defaulter
- - If we see the Housing type , House/apartments or with parents seems to find no issue in paying loans and less refusal rates

After combining two data sets

- - at the same time co-apartment seems to be defaulter so Banks/credit company can avoid them.
- - Seeing the Marital status marries people are more towards defaulters than single/not marries .Banks can also look into this.
- Here for the 'NAME_PAYMENT _TYPE' , the clients with payment type as "cash through banks" , is more prone to be a defaulter or is a applicant with Refusal status.
- From the above observations , the client type ,
- Repeater is comparatively , more in refusal and more chance of defaulter.

After combining two data sets

- From the above observations , the client type ,
- Repeater is comparatively , more in refusal and and more chance of defaulter
- So Banks can consider the above scenarios to validate if they can give loans to the particular client and not miss the ones who can pay successfully.