CREDIT EDA CASE STUDY

Kushagra Vikram Jaiswal & Ishar Mohapatra

From the application data csv we have cleaned columns with more than 40% records empty and on the basic of few irrelevant columns based on description in the data dictionary.

We have done our analysis majorly on the following columns.

Ordered Categorical Fields

1. NAME_EDUCATION_TYPE
2. REGION_RATING_CLIENT
3. OCCUPATION_TYPE
4. REGION_RATING_CLIENT_W_CITY

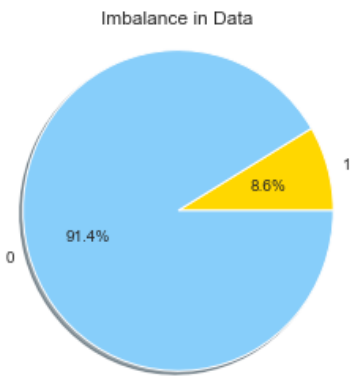
Unordered Categorical Fields

1. ORGANIZATION_TYPE
2. NAME_INCOME_TYPE
3. NAME_TYPE_SUITE
4. REGION_RATING_CLIENT_W_CITY
5. NAME_FAMILY_STATUS
6. CODE_GENDER

Quantitative Columns

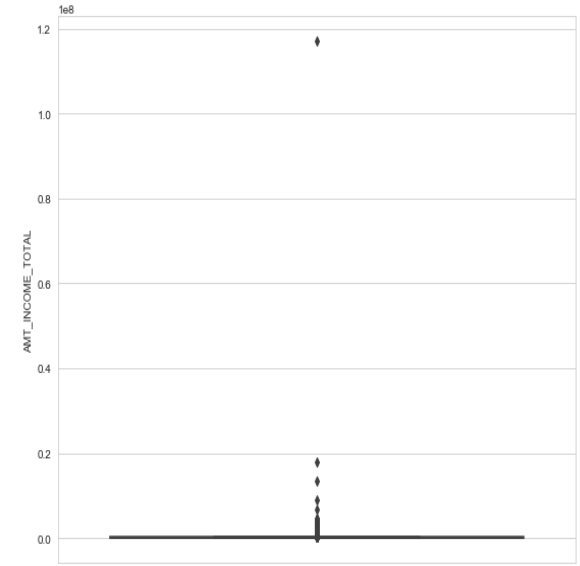
1. EXT_SOURCE_2
2. DAYS_BIRTH
3. DAYS_REGISTRATION
4. AMT_ANNUITY
5. DAYS_EMPLOYED
6. DAYS_ID_PUBLISH
7. AMT_CREDIT
8. AMT_INCOME_TOTAL
9. AMT_GOODS_PRICE
10. EXT_SOURCE_3
11. REGION_POPULATION_RELATIVE

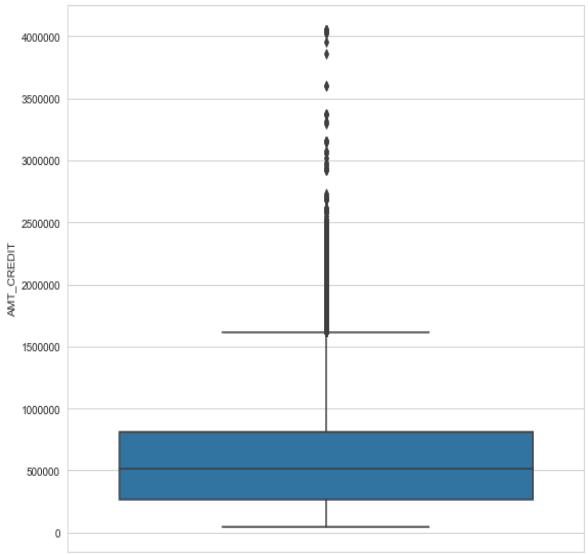
And the TARGET Column for finding Defaulters if any.

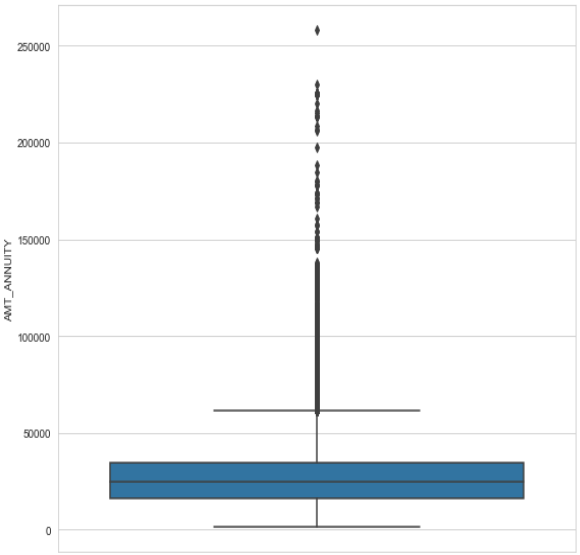


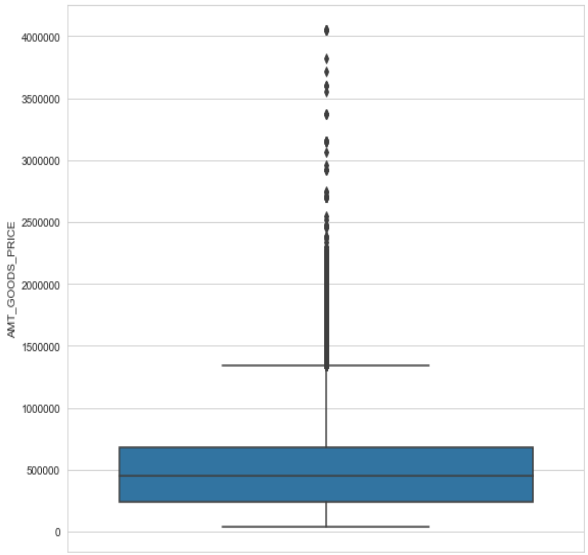
We can see that there are 91.4% data where there is no default in loans, whereas only 8.6% data where we have defaulters, which leads to imbalance in data during analysis.

We have also noticed that there are a lot of outliers for the quantitative analsyis of columns like 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE'





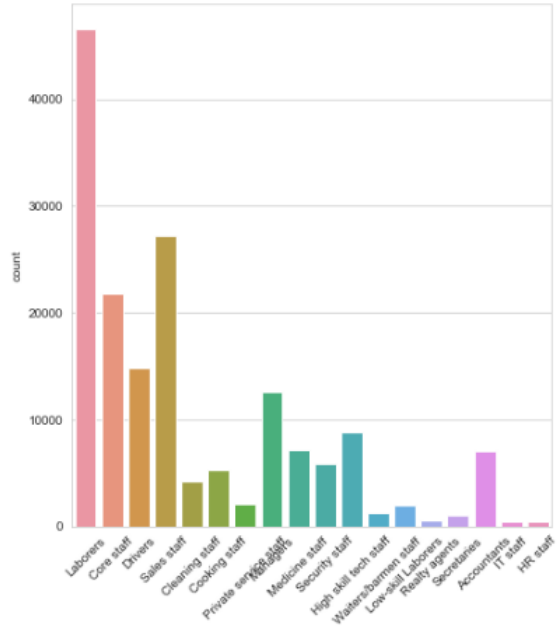


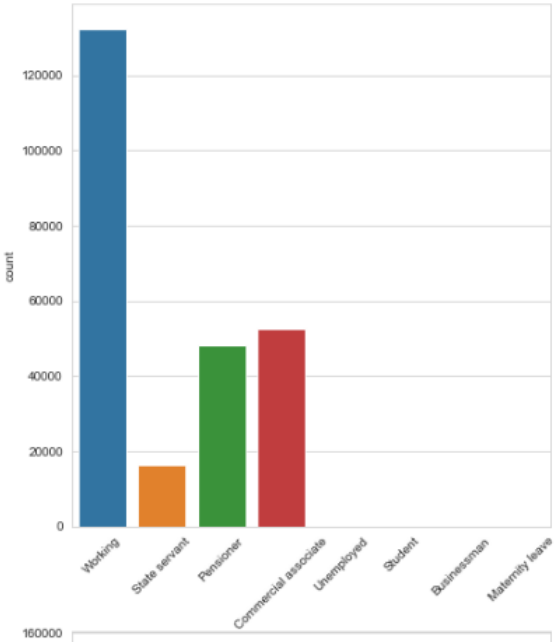


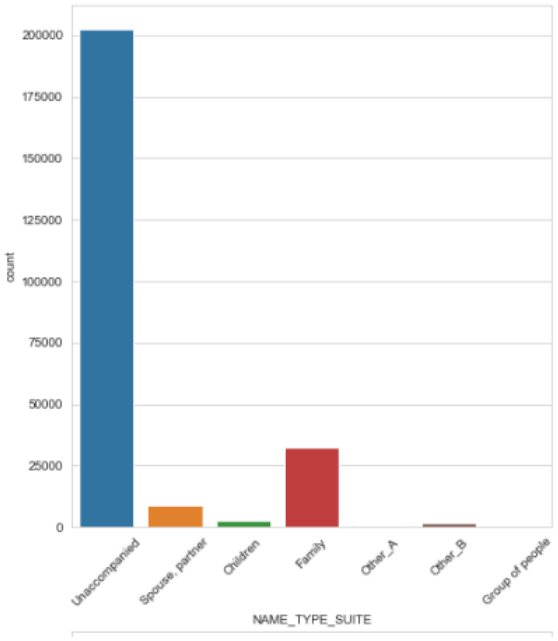
So as part of the cleaning process we have done outliers treatment while removing top 5 % data for further analysis.

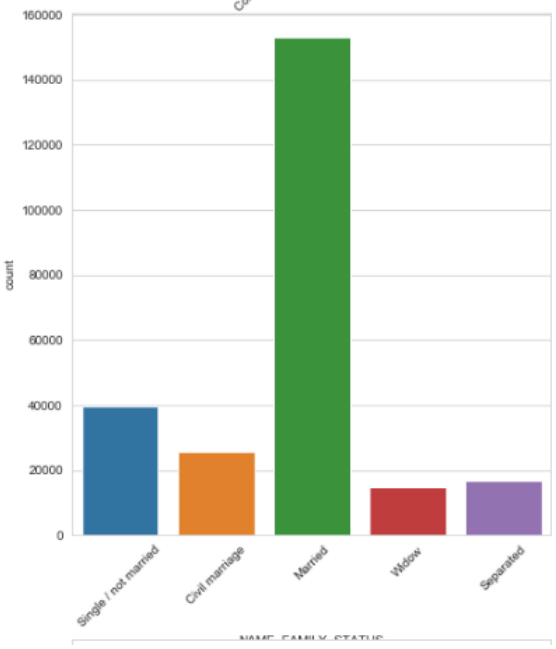
Univariate Analysis

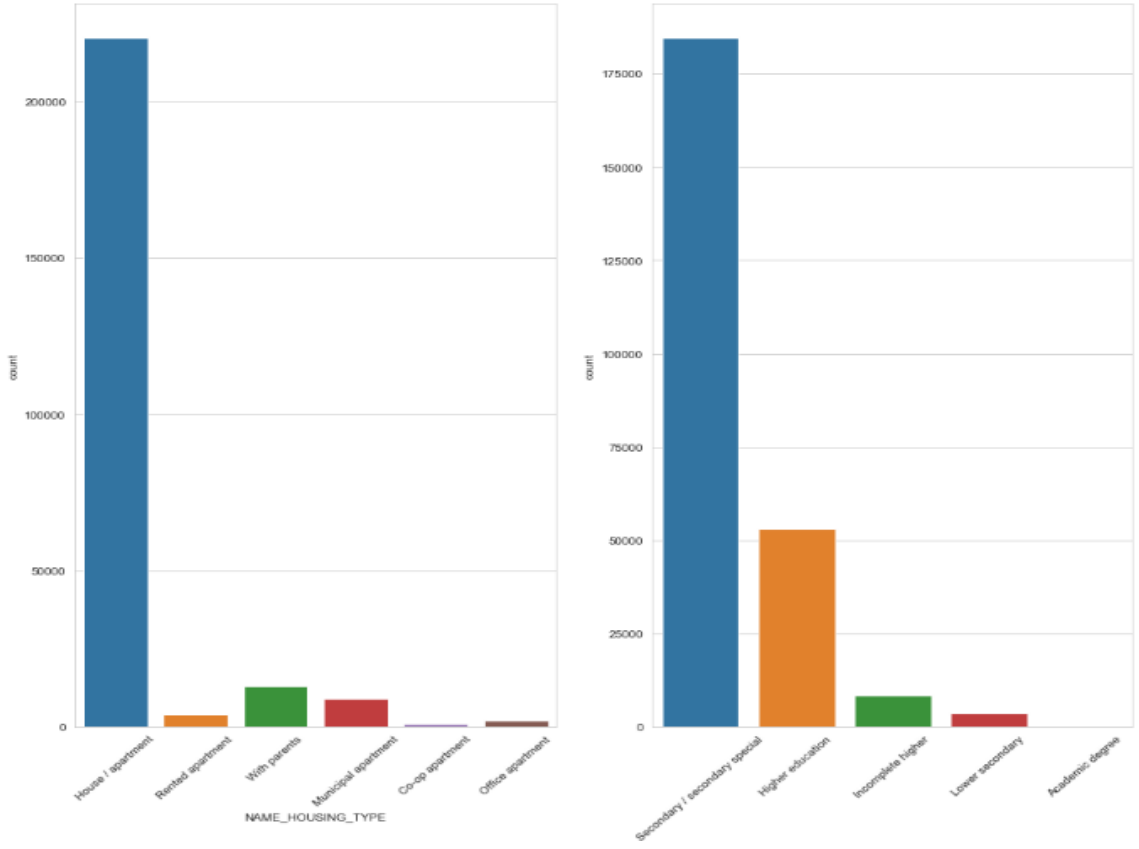
For categorical columns







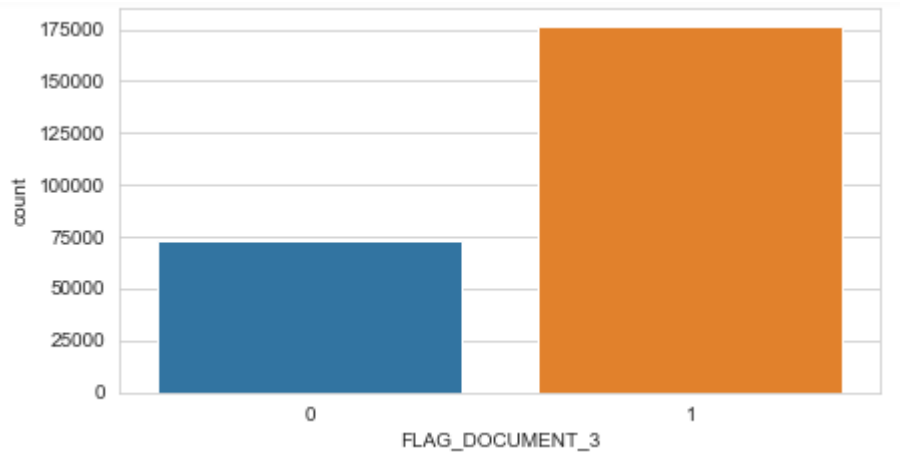


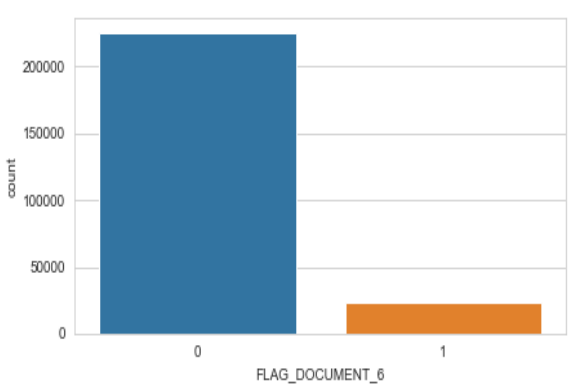


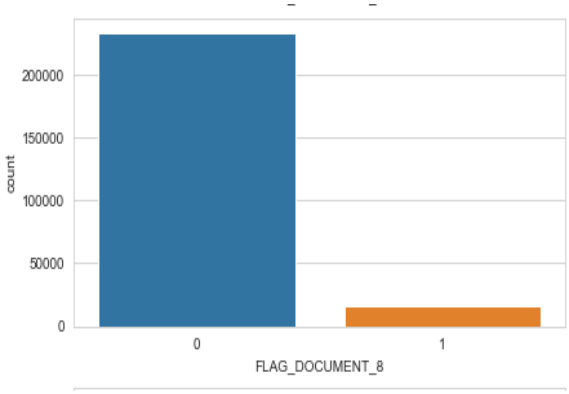
Above plots shows how many customers lie in particular category and here are insights

1. Clearly Laborers apply more for loans, followed by Sales Staff and Core Staff
2. Working Clients apply more loans followed by Commercial associates then by Pensioners
3. Married people apply for more loans followed by Single/ Not Married
4. Most people who apply for loan have a House /apartment
5. Females apply more loans than Males
6. Cash Loans applications are more than Revolving Loans
7. Most people applying for loan don't own a Car, but most own house/flat, phone and email.
8. Most of the people are not defaulting on their instalments.
9. Secondary/ Secondary Special educated people apply for more loan followed by Higher education and Incomplete Higher.

On plotting FLAG_DOCUMENTS columns we could see most people provided FLAG_DOCUMENT_3 and very minimal Document_6 and 8 while rest documents were not submitted.

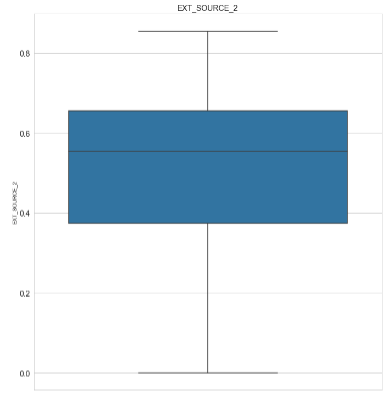


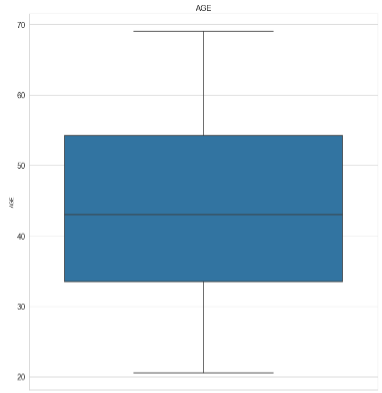


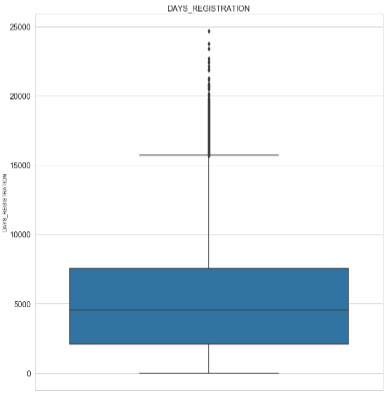


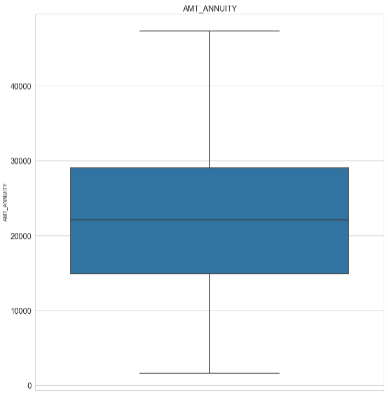
As we can see that except FLAG_DOCUMENT_3 all the columns have negligible count of 1s. So we are removing all the FLAG_DOCUMENT columns after making the observation

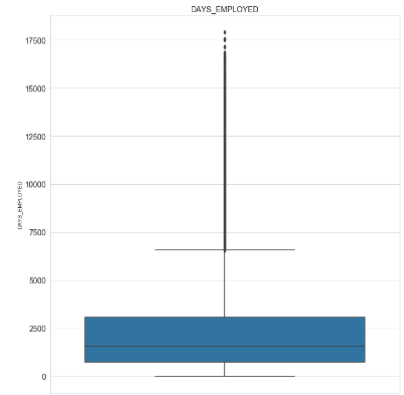
For Quantitative Columns

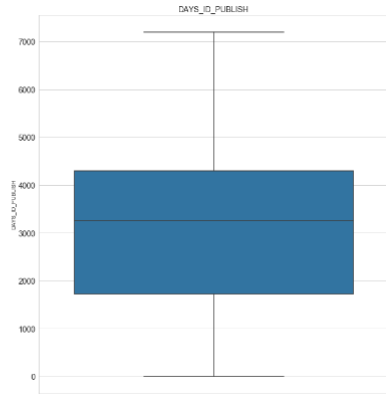

EXT_SOURCE_2
0.8
0.6
0.4
0.2
0.0

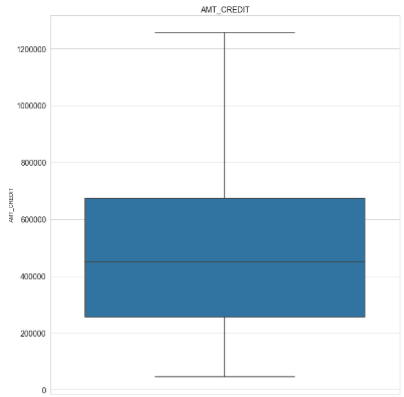

AGE
70
60
50
40
30
20

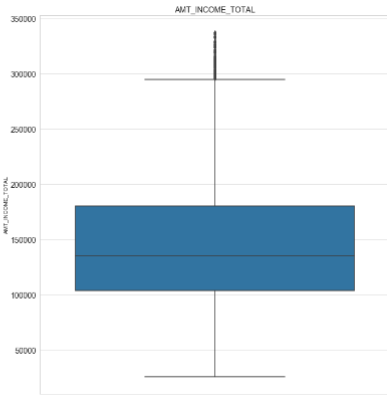

DAYS_REGISTRATION
25000
20000
15000
10000
5000
0


AMT_ANNUITY
40000
30000
20000
10000
0


DAYS_EMPLOYED
17500
15000
12500
10000
7500
5000
2500
0


DAYS_ID_PUBLISH
7000
6000
5000
4000
3000
2000
1000
0


AMT_CREDIT
1200000
1000000
800000
600000
400000
200000
0


AMT_INCOME_TOTAL
350000
300000
250000
200000
150000
100000
50000

Clearly, we can make out few observations from the above plots

1. Majority of Normalized score of the population from external data source 2 lies around in between 0.4 - 0.65, clearly External source 3 is also having near about similar values.
2. Majority of Age of Loan applicants lies around in between 35 - 55 years.
3. Majority of DAYS_REGISTRATION column values lies around in between 2500 - 5500 days, clearly there are lot of outliers as well who haven't changed their registration since long.
4. DAYS_ID_PUBLISH columns we can notice majority of the client change their identity document nearly 1800 - 4200 days before applying loan.
5. REGION_POPULATION_RELATIVE column spread is nearly between 0.01 - 0.029.
6. We notice that AMT columns and DAYS_EMPLOYED are having a lot of outliers we need to analyse further.

Now we created 2 new categorical columns based on Age (AGE_CATEGORY) and Income Group (INCOME_GROUP).

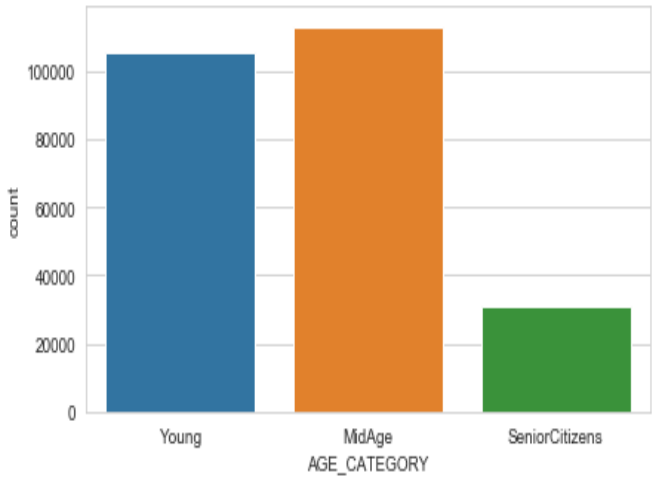
Based on AMT_INCOME_TOTAL column we categorized as

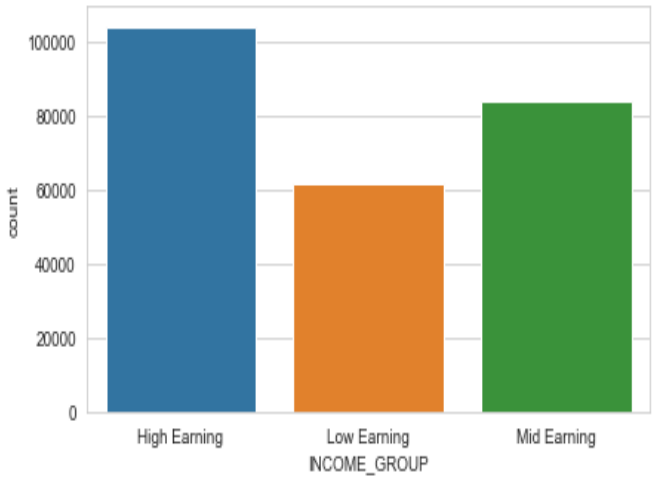
- Low Earning (0 - 100000)
- Mid Earning (100000 - 150000)
- High Earning (more than 150000)

Now, for Age we converted DAYS_BIRTH column to AGE column and on checking the age column we noticed the client age spread from 20 - 69

- Young (20-39 yrs.)
- Mid Age (40-59 yrs.)
- Senior Citizens (60 - 69 yrs.)

Below is the plot from the derived categorical columns.

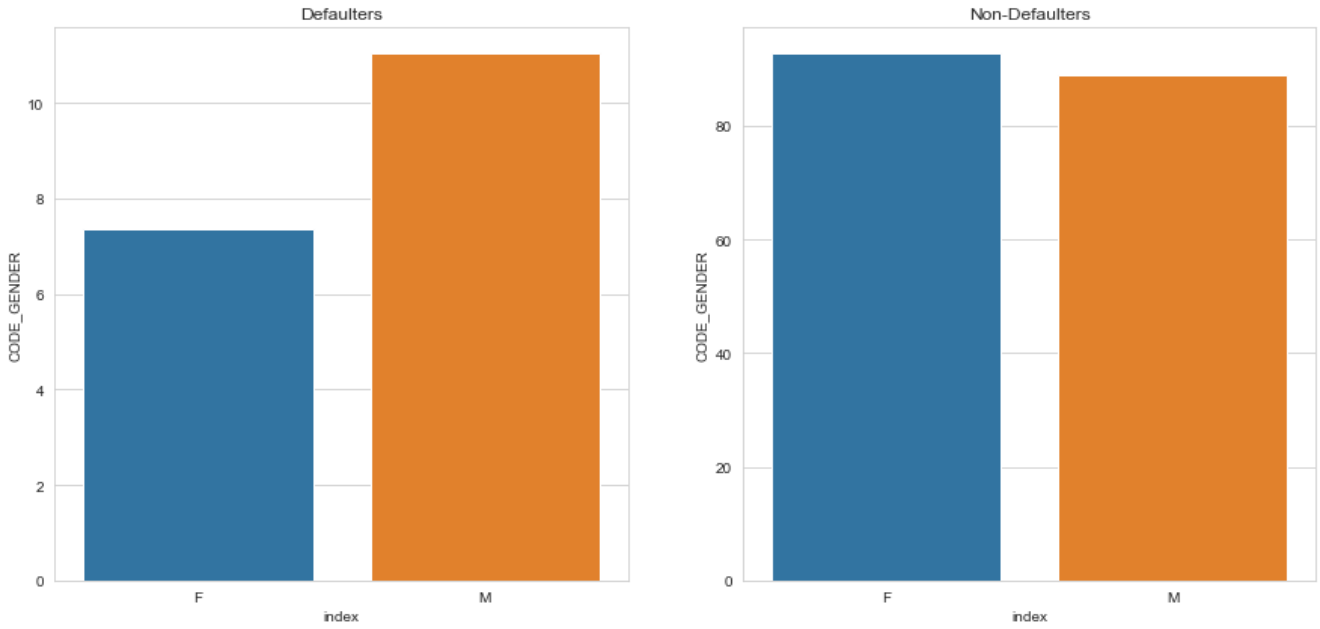




Insights

1. Mid Age Group apply most for the loan followed by Young people.
2. High Earning followed by Mid Earning apply for loan most

Univariate Analysis on two sets of data (Defaulter and Non-Defaulters)



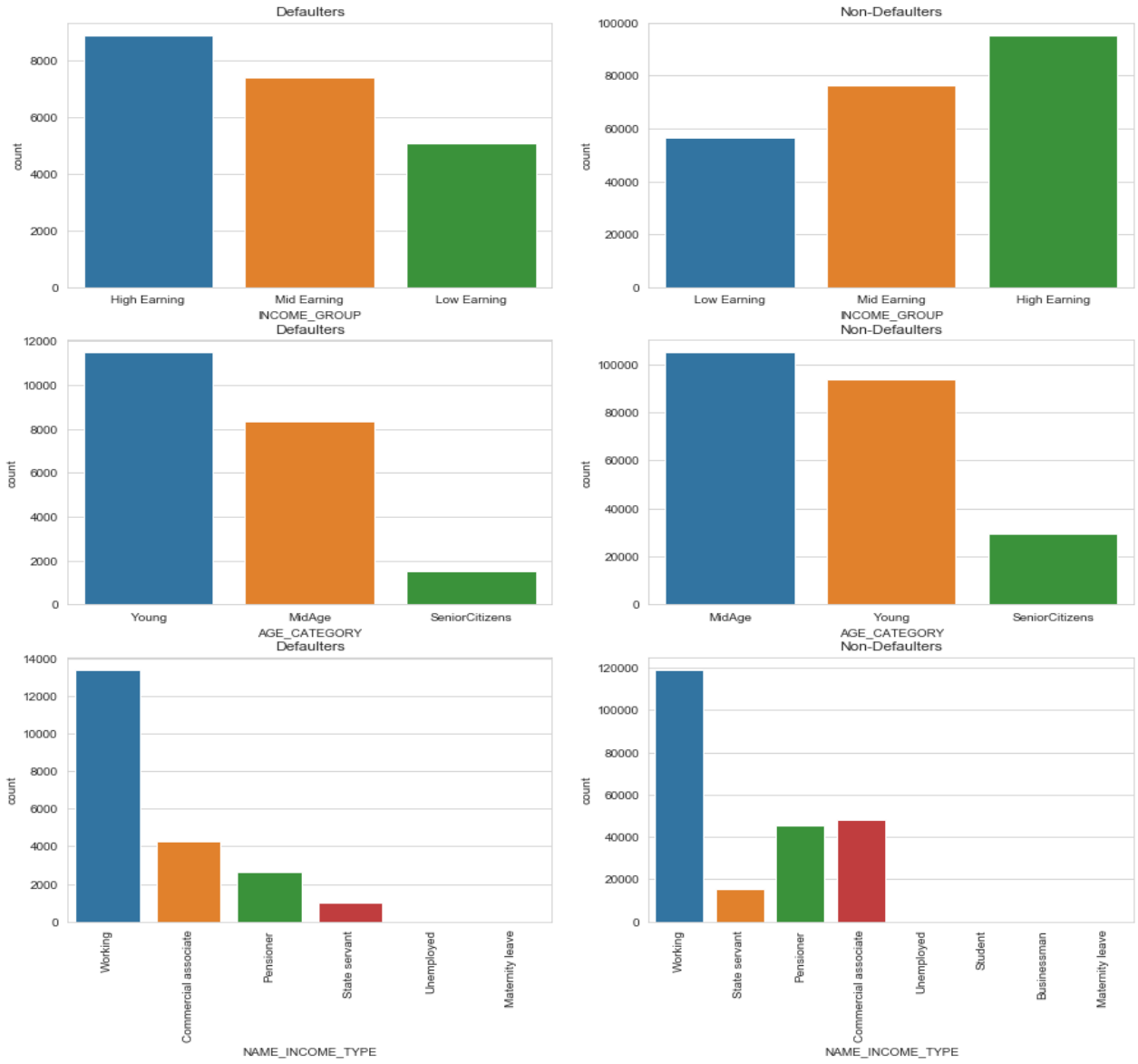
Insights:

1. Since female clients are more in number than male clients, so the defaulters count of female is more than males.
2. So we have taken the percentage of Female who defaults and percentage of Male who defaults.

Clearly we can see Male Clients default more than the Female Clients.

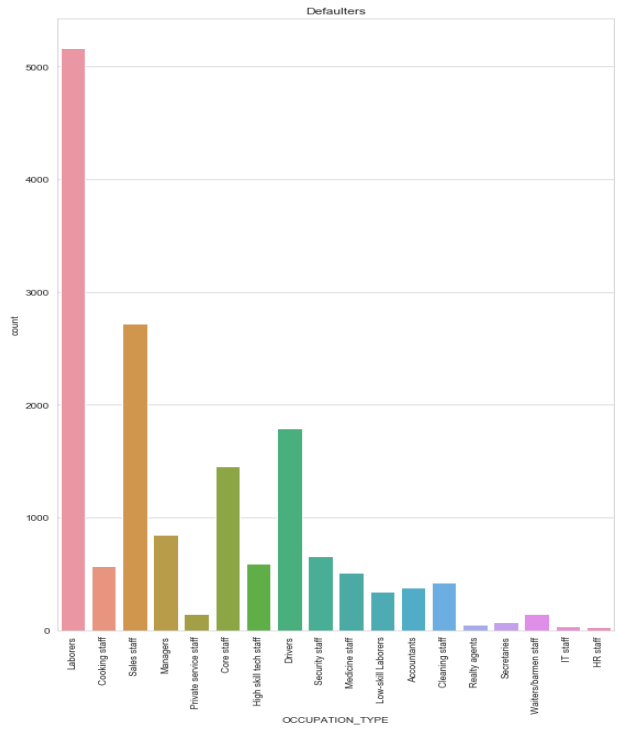
Male Defaulters (11.03%)
Female Defaulters (7.36%)

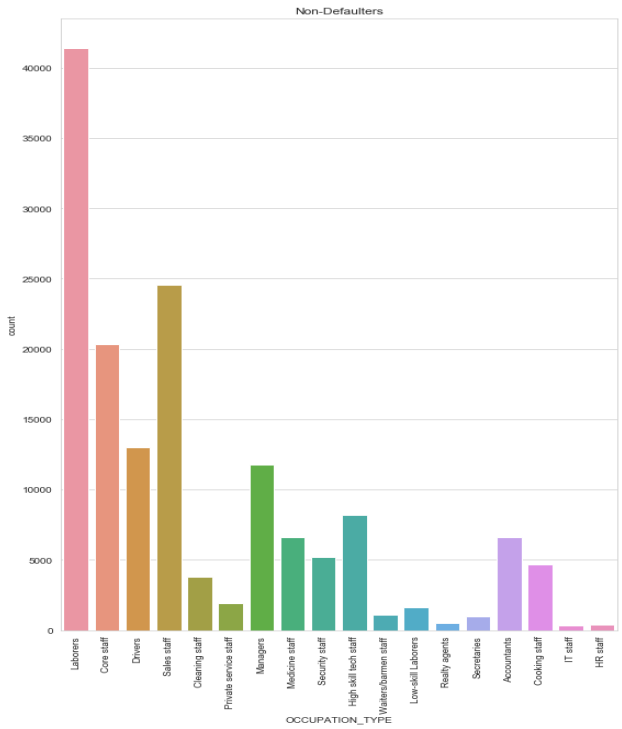
So more loans should be provided to Females.

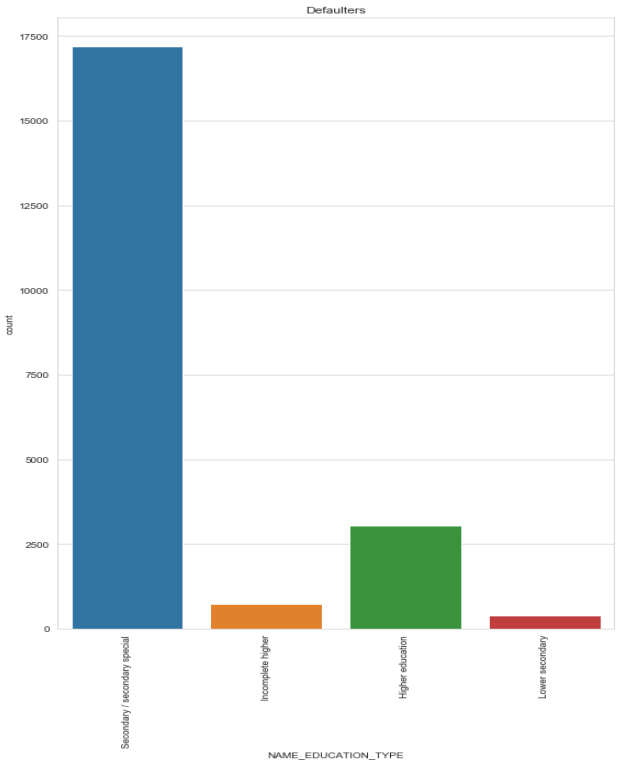


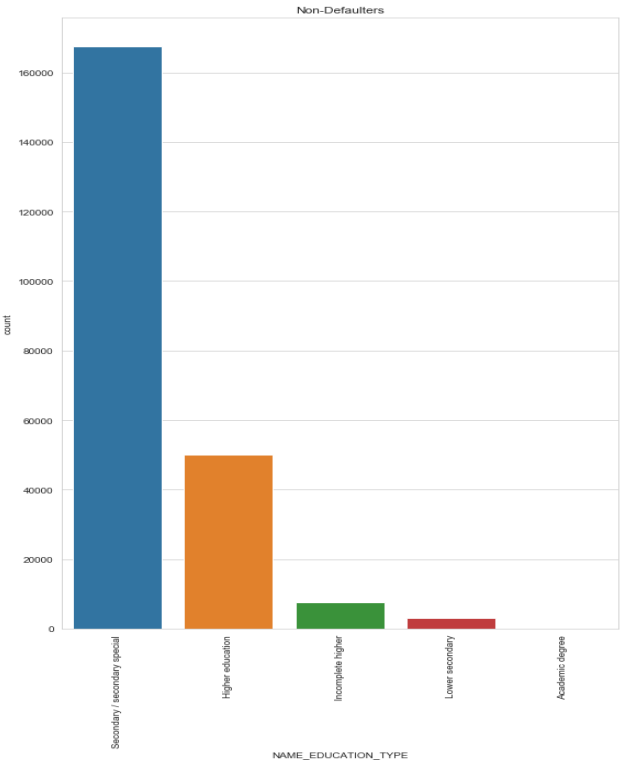
Insights:

1. Previously we noticed that Mid Age Group apply most for the loan followed by Young people but in defaulters list we notice young people default more than that of Mid Age Group.
2. High Earning followed by Mid Earning apply for loan most, so defaulters in high earning people are more than mid earning people.
3. Working Clients apply more loans followed by Commercial associates then by Pensioners and same trends follows for defaulters and non-defaulters also





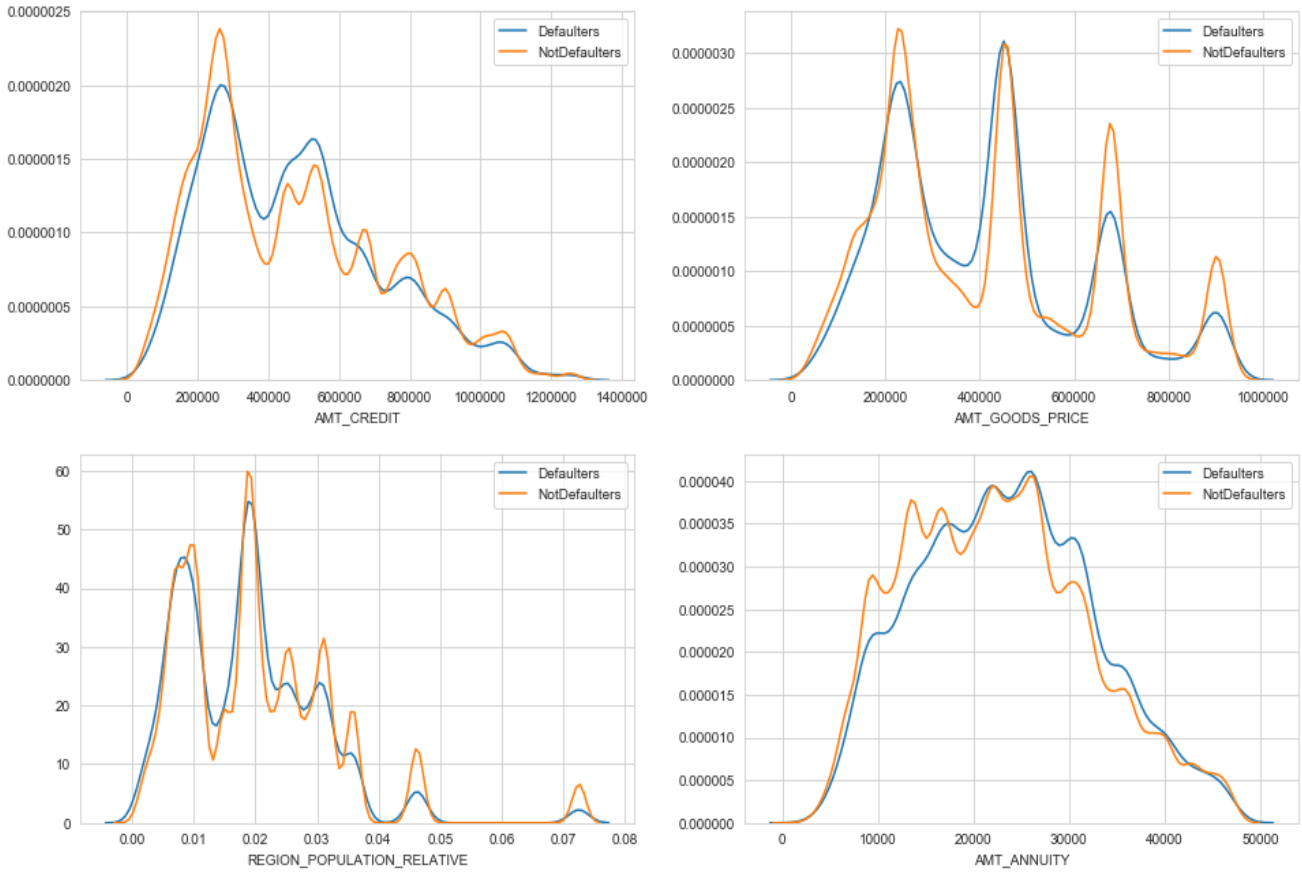




Insights:

1. Previously we noticed Laborers apply more loan and hence they top the list in both defaulters as well as non-defaulters followed by Sales Staff.
2. Previously we noticed Secondary/Secondary Special apply more loan and hence they top the list in both defaulters as well as non-defaulters followed by Higher Secondary.

Univariate for Continuous variables

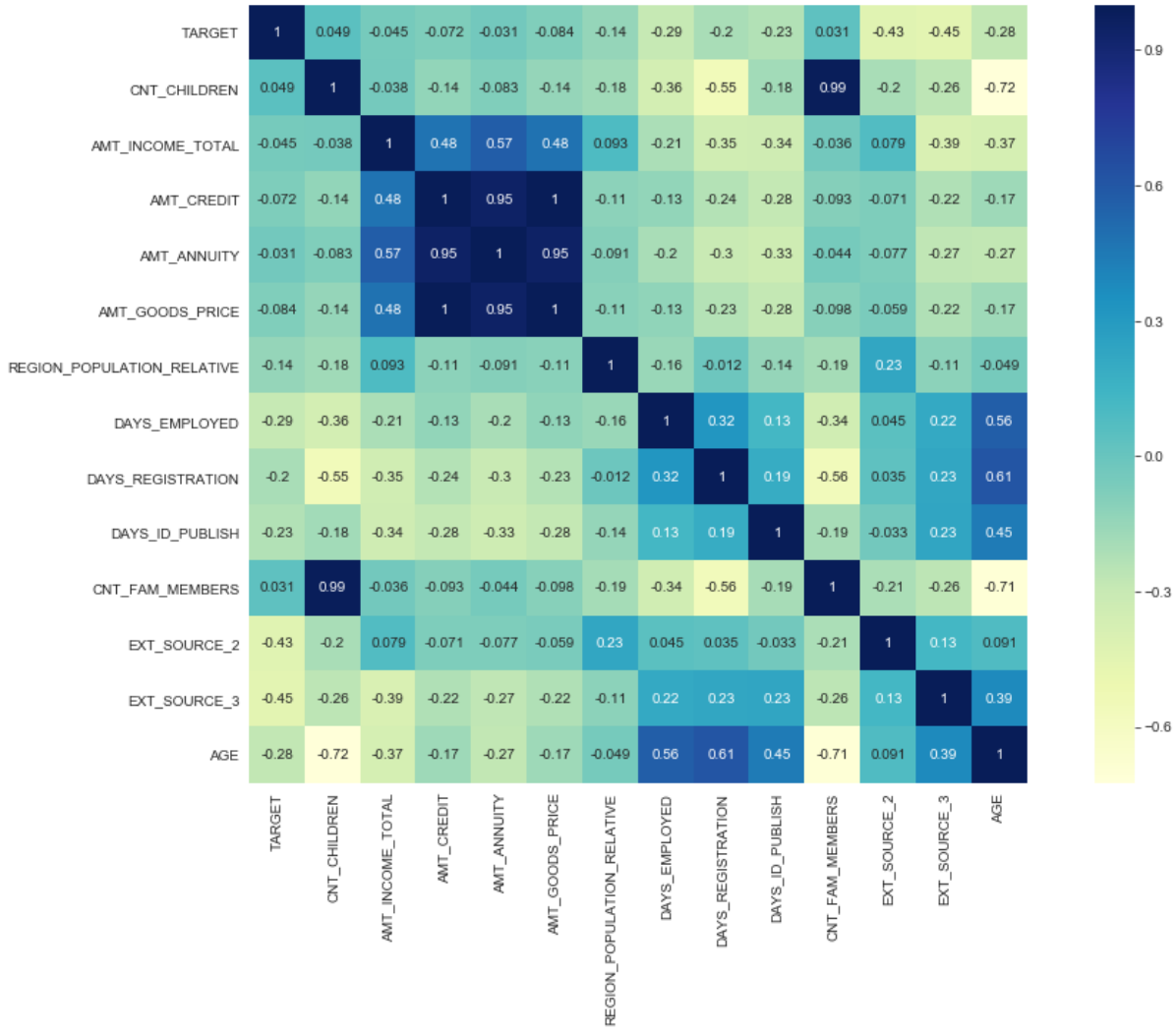


Insights:

1. When comparing Defaulters to Non Defaulters for AMT_CREDIT both go hand in hand except in the middle we see a sudden spike of more Defaulters than Non-Defaulters when the amount of Credit is around 500000
2. Due to outliers we are not able to make out any insights in INCOME plot.
3. When comparing Defaulters to Non Defaulters for AMT_GOODS_PRICE similar trend can be seen we see a sudden spike of more Defaulters than Non-Defaulters when the amount of Credit is around 500000
4. For REGION_POPULATION_RELATIVE we can notice there is good amount of gap and profit area for banks where non defaulters are more than defaulters where population relative is between 0.04 - 0.05 and 0.07 - 0.08. Banks should concentrate more in those areas.
5. For AMT_ANNUITY, we can see loss area where defaulters are more when annuity is around 25000.

Bivariate analysis

For continuous variables

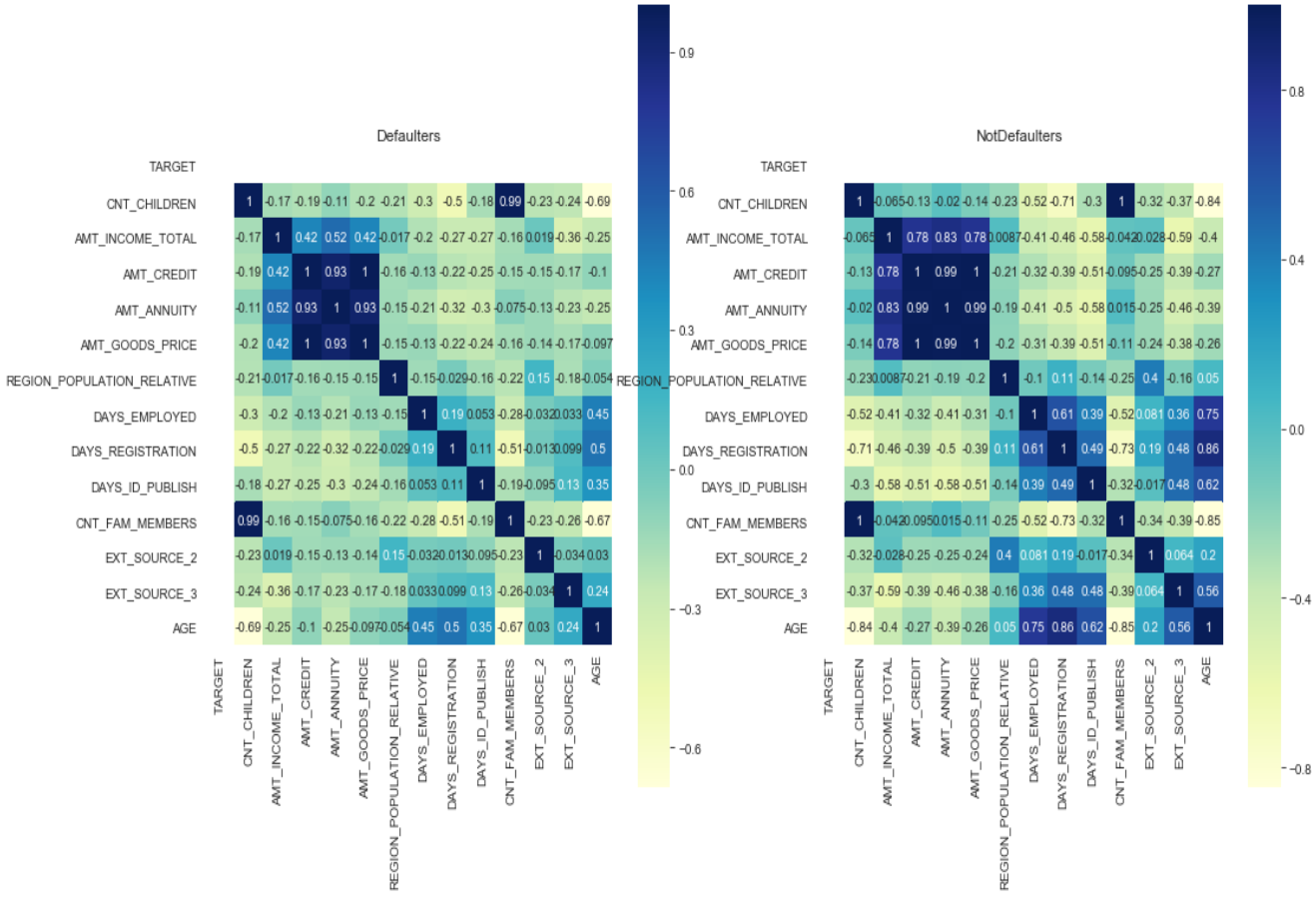


Insights:

Important observation for correlation to make are

1. AMT_GOODS_PRICE and AMT_CREDIT have a good positive correlation of 0.976080
2. AMT_GOODS_PRICE and AMT_ANNUITY have a good positive correlation of 0.748504
3. AMT_ANNUITY and AMT_CREDIT have a good positive correlation of 0.748642
4. CNT_FAM_MEMBERS and CNT_CHILDREN have a good positive correlation of 0.875547 which is quite obvious

For continuous variables after splitting dataframe based on TARGET column.



Insights:

Important observation for correlation to make are which holds true for both for the non-defaulter as well as defaulters

1. AMT_GOODS_PRICE and AMT_CREDIT have a good positive correlation of 0.976553 which was 0.971472 for defaulters
2. AMT_GOODS_PRICE and AMT_ANNUITY have a good positive correlation of 0.750547 which was 0.733866 for defaulters
3. AMT_ANNUITY and AMT_CREDIT have a good positive correlation of 0.749907 which was 0.737421 for defaulters
4. CNT_FAM_MEMBERS and CNT_CHILDREN have a good positive correlation of 0.874701 which was 0.883714 for defaulters

So Overall Correlations remain the same even after data is split

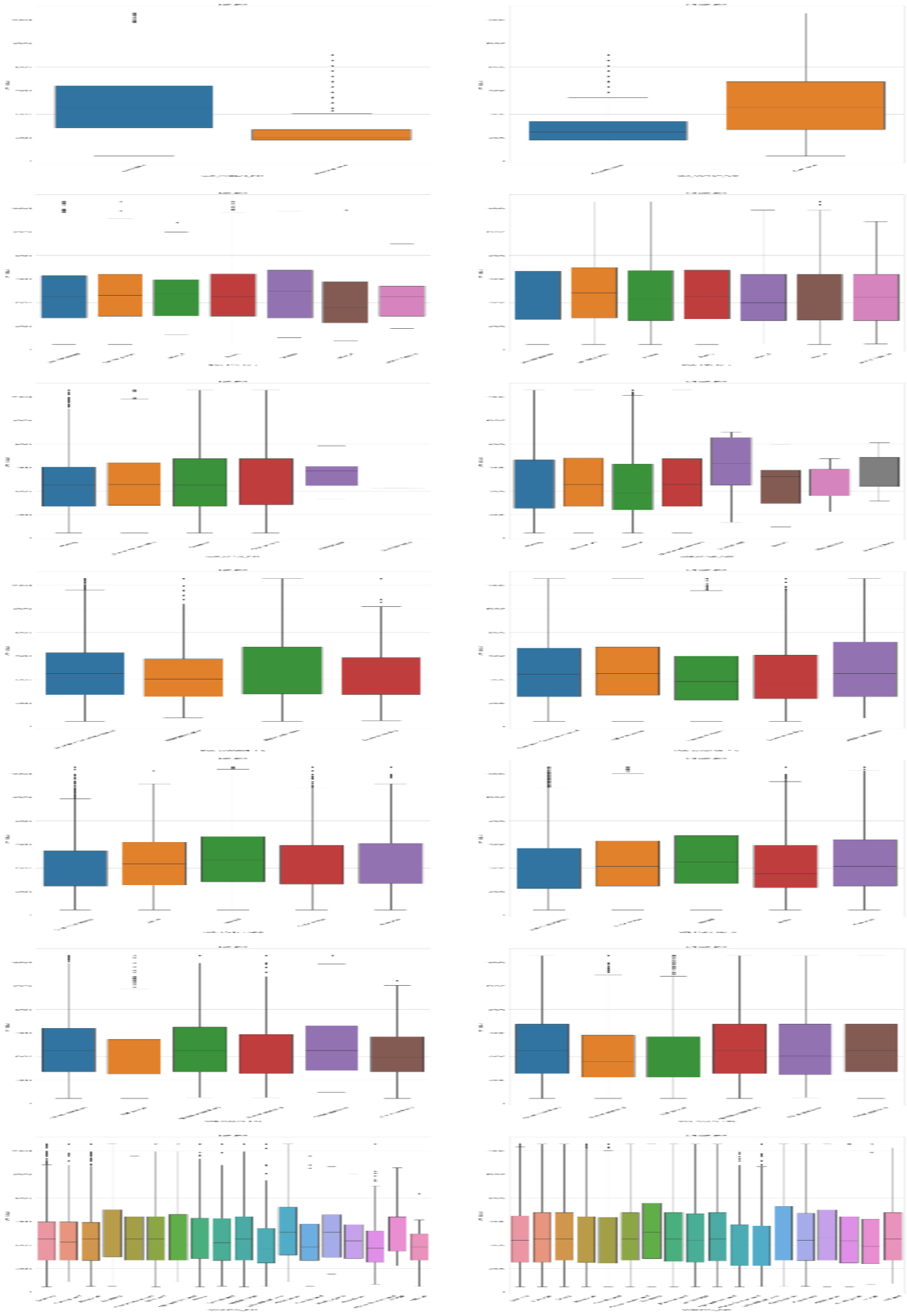
We can clearly see both Defaulter's heat map and Non Defaulter's heat map is almost similar.

Since we noticed AMT_CREDIT is most correlated so let's plot AMT_INCOME_TOTAL and AMT_CREDIT with Gender for Both Defaulters and Not Defaulters

Insights:

1. On the first scatter plot we can see on the lower income side we a lot of female client across AMT_CREDIT whereas as we go on to the higher side of AMT_INCOME we see male clients default more than females.
2. We can see the same after the spikes where orange line of Non-Defaulters is crossing way over blue line that is for defaulters.

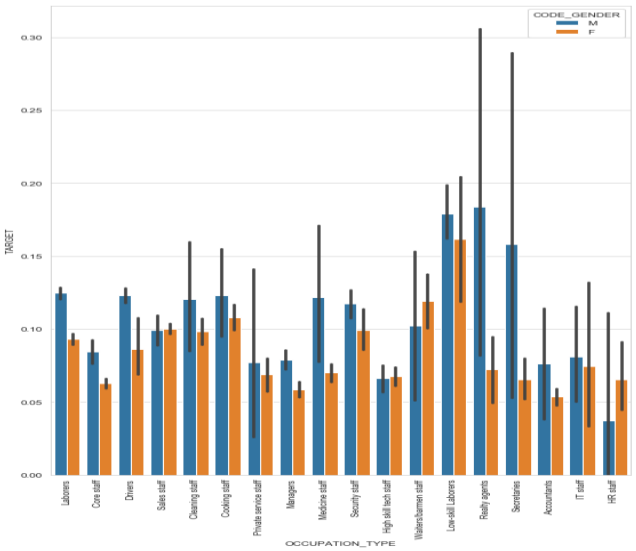
For Categorical variables

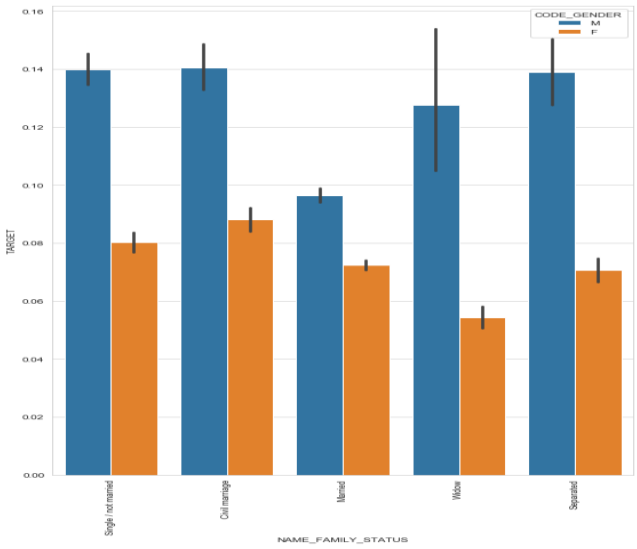


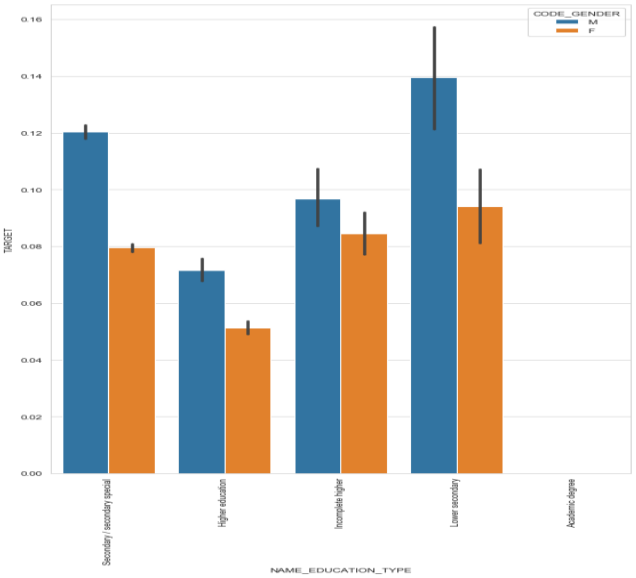
Insights:

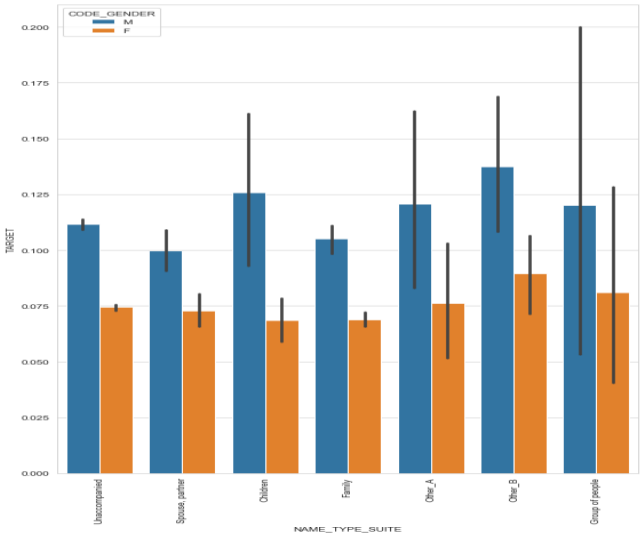
1. Revolving loans are much lower than Cash loans for both Defaulters and Not Defaulters.
2. Column NAME_TYPE_SUIT has no difference for both Defaulters and Not Defaulters.
3. State Servants and Pensioners have a high amount credit in case of defaulters. In comparison, in case of Not Defaulters, Unemployed has the highest amount credit. This could be due to fact that unemployed people might be needing more funds to cover expenses.
4. In Education, people in Higher Education have the most amount credited and are Defaulters. For Not Defaulter, Academic degree has the highest amount credit.
5. Married people take the highest amount credit in case of both Defaulters and Not Defaulters.
6. Almost all NAME_TYPE_SUITES apply for similar AMT_CREDIT with similar medians for both Defaulters and Not Defaulters.
7. Accountants and Managers take the highest amount credit for Defaulters. Same is true for Not Defaulters as well.

Analysis of two segmented variables





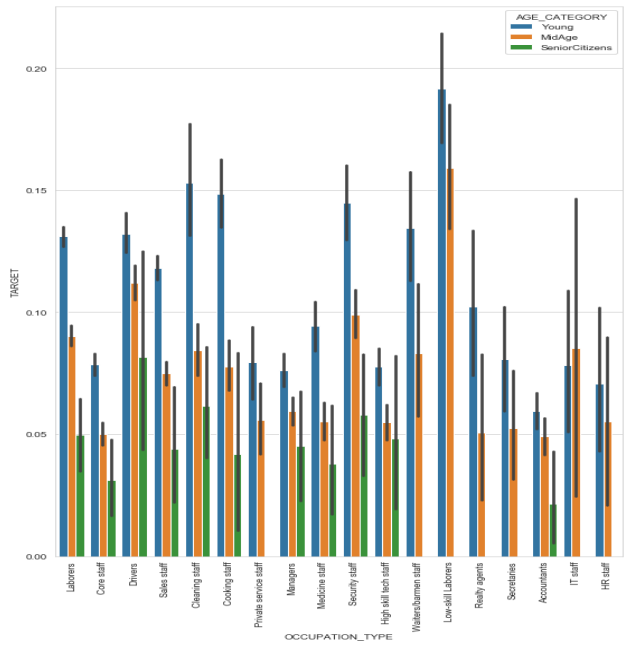


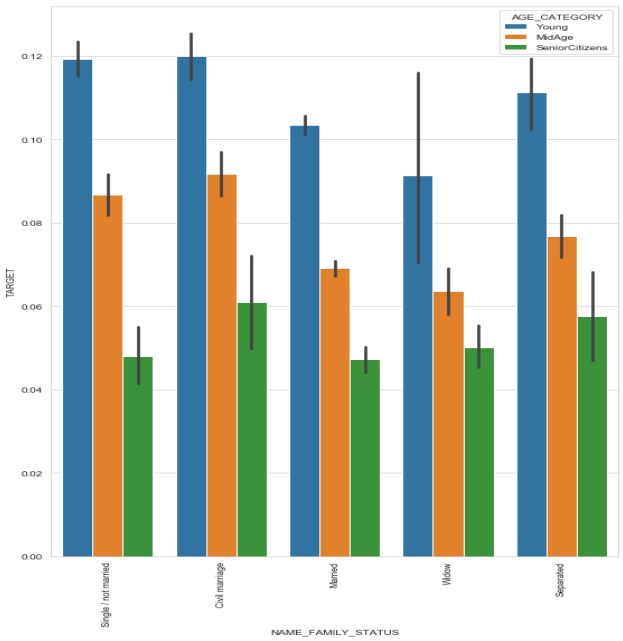


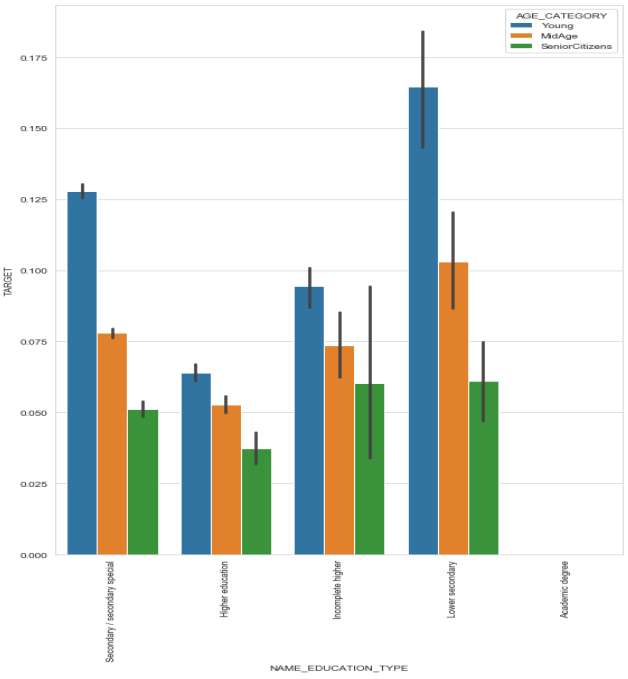
Insights:

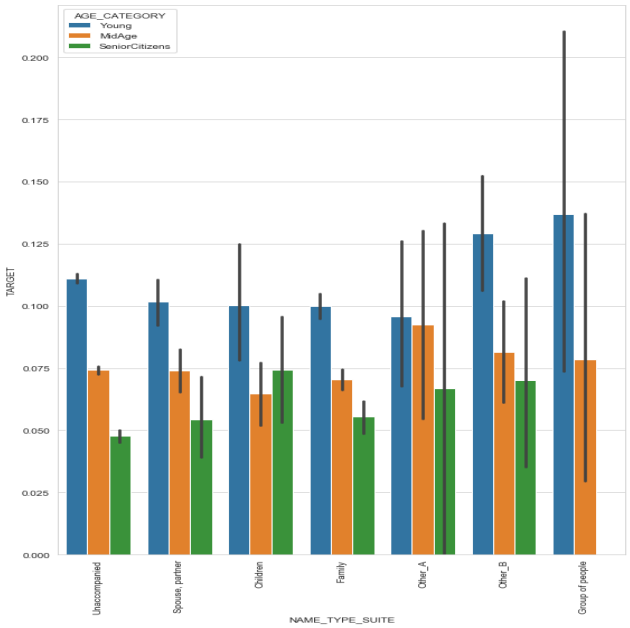
These above plots shows us defaulters mean in each category

1. Male working as Realty Agent, Low skilled laborers and Secretaries are the most among defaulters.
2. Female working as Low skilled laborers and Waitress /Barmen Staff are the most among defaulters
3. Single/not married and Civil Married male and separated are most to default on their payment whereas Females who are Civil Married and Single/not married default more.
4. Clients with Lower Secondary education are most to default among men and women then by Secondary / Secondary special.
5. Clients both male and female who are accompanied by Other_B default the most.









Insights:

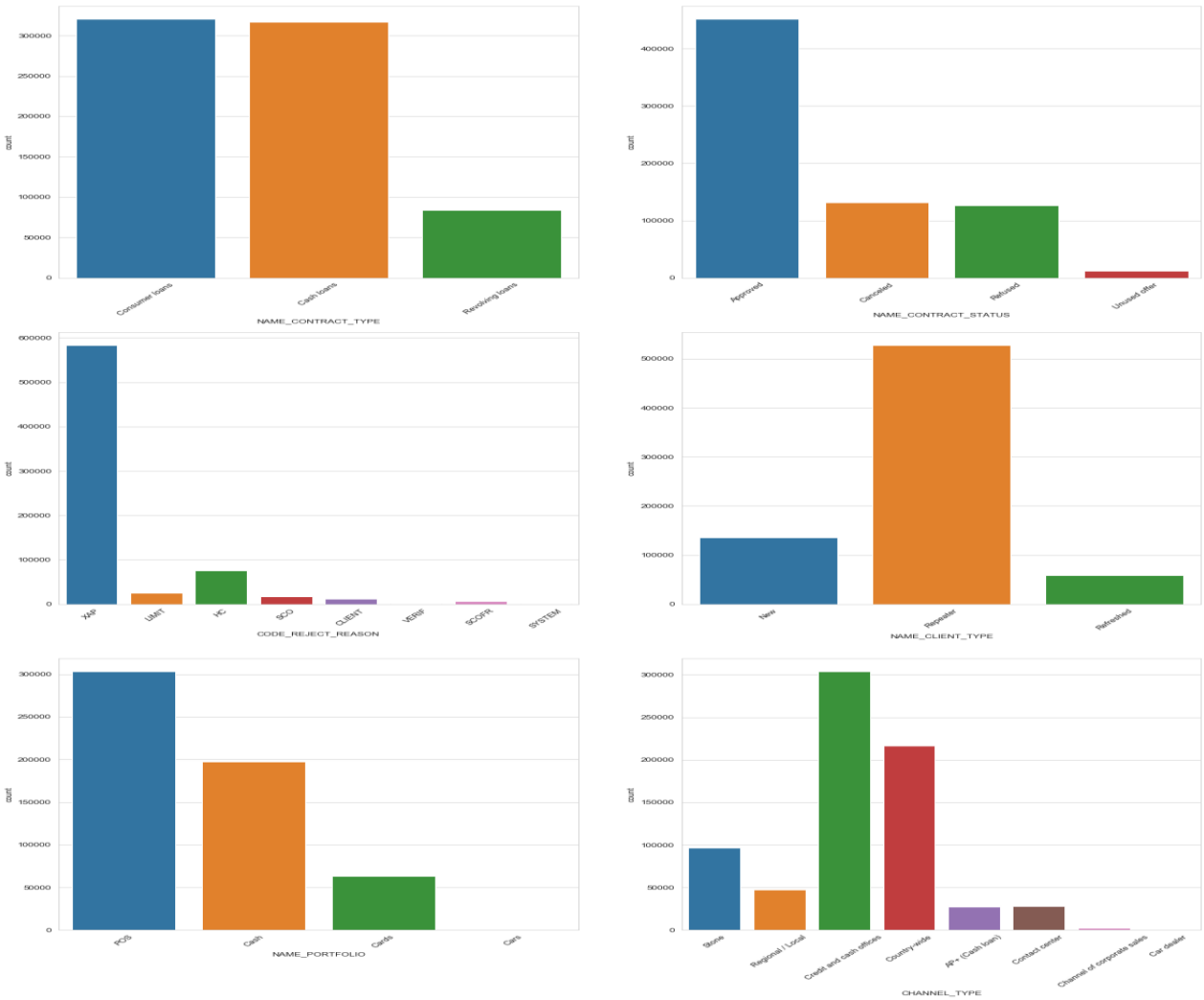
These above plots shows us defaulters mean in each category

1. People who work as Low Skilled Laborers tends to default more among all age groups, probably Seniors don't work in that role to
2. Single and Young people default the most followed by Mid-Age and Single.
3. Young and Mid-Age people with Lower Secondary tends to default more.
4. Young people accompanied by Group of people defaults more.

Merge application_data with prev_application

Univariate Analysis on merge data

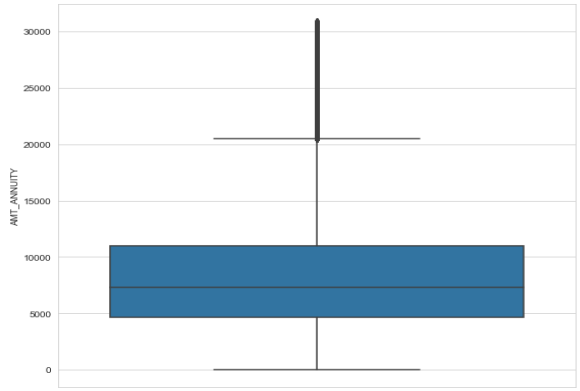
For Categorical columns

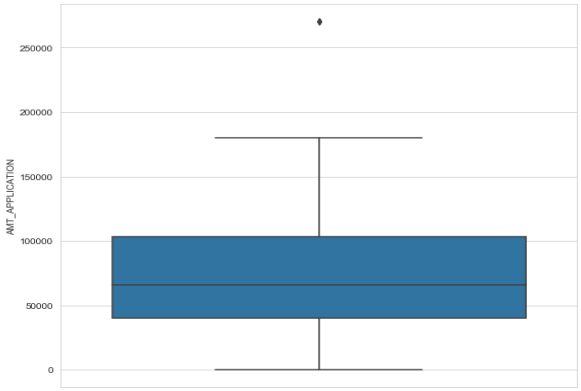


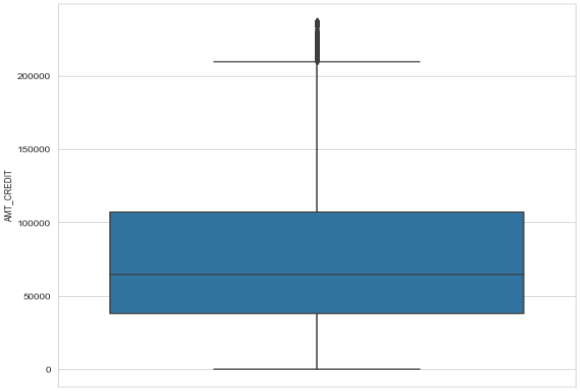
Insights:

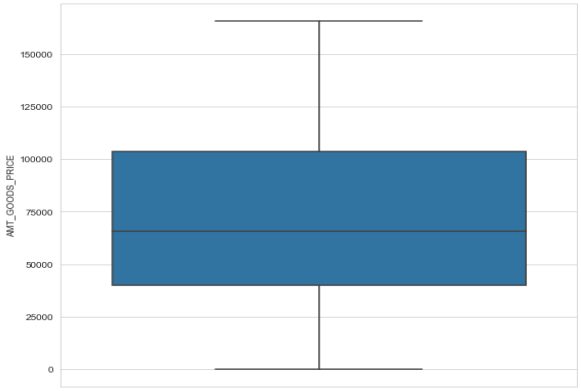
1. NAME_CONTRACT_TYPE: Most of the type of loans applied for are Consumer loans, followed by Cash loans.
2. NAME_CONTRACT_STATUS: Approx. 63% of the loans are "approved”, whereas 1.63% loans are in "unused offer"
3. CODE_REJECT_REASON: Majority(81%) of the times, the loan is rejected because of the reason code 'XAP'
4. NAME_CLIENT_TYPE: Most of the applicants are Repeaters (73%), whereas only 18% of the applicants are New.
5. NAME_PORTFOLIO: Most of the portfolios applied for is "POS" ,followed by "CASH"
6. CHANNEL_TYPE: Credit and cash offices(42%) bring in most of the clients then followed by Country-wide and Stone

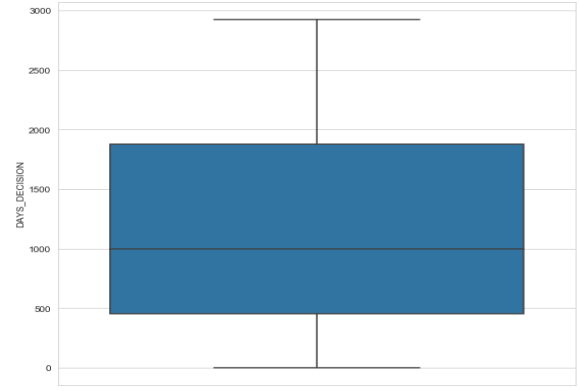
For continuous variables





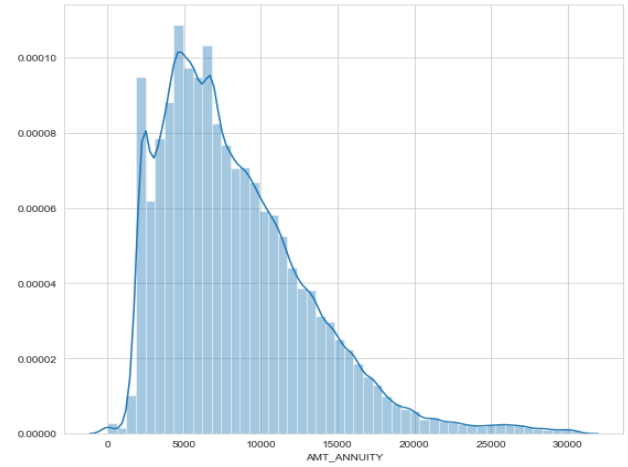


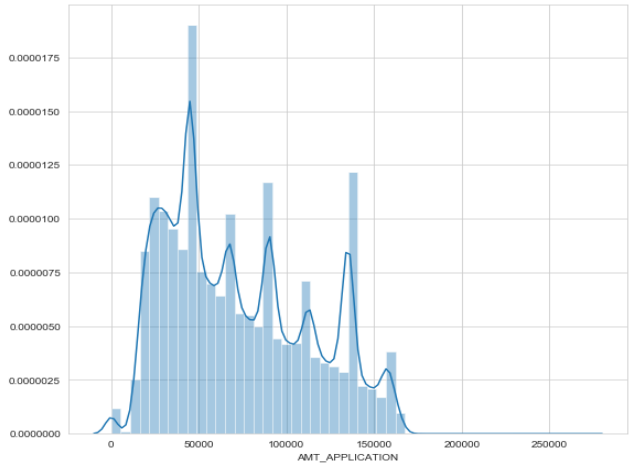


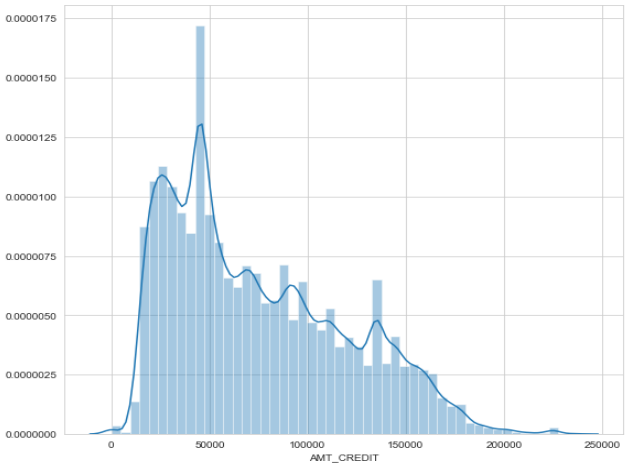


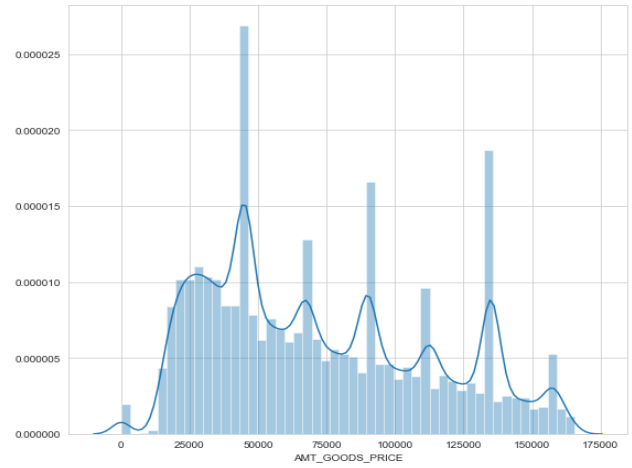
Insights:

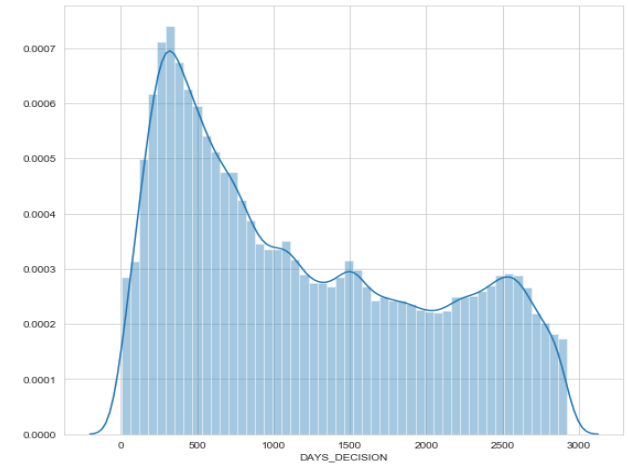
1. After cleaning of Outliers top 10 percent
2. AMT_ANNUITY: We observe clearly majority lies between 5000 and 12000 with a lot of outliers.
3. AMT_APPLICATION: Majority lies around 40000 - 100000 with a small amount of outliers.
4. AMT_CREDIT: Is almost similar to AMT_APPLICATION with slightly higher outliers.
5. AMT_GOODS_PRICE: Majority lies around 40000 - 115000
6. DAYS_DECISION: Majorly around 500 -1900 days spent on Decision









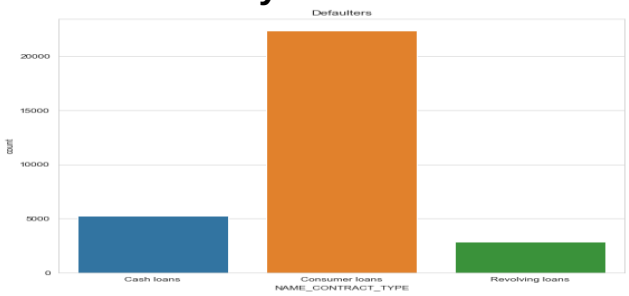


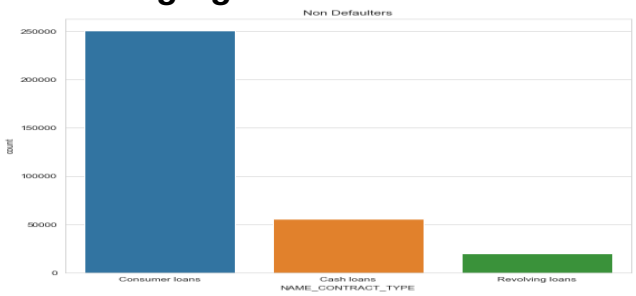
Insights

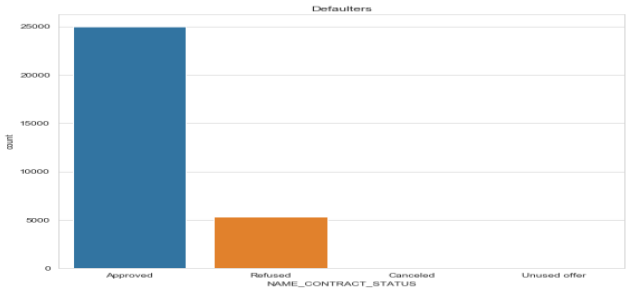
After cleaning of Outliers top 10 percent

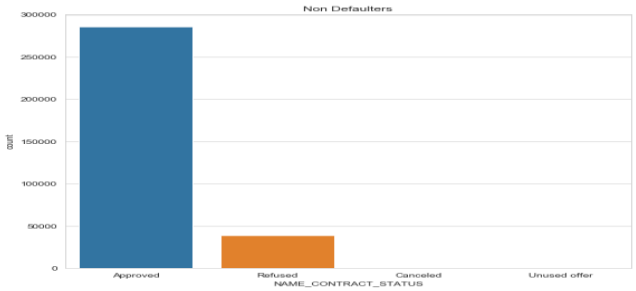
1. AMT_ANNUITY: We observe similar trends in distribution plot as well majority of annuity is towards the lower from around 4000 - 12000.
2. AMT_APPLICATION: We observe similar trends in distribution plot as well majority distribution around 20000 - 200000.
3. AMT_CREDIT: We can see a clear spike around 50000
4. AMT_GOODS_PRICE: we can see a spike after regular intervals.
5. DAYS_DECISION: Majorly distributed from 0-2800 days, maximum around 400 days
6. Let's see if similar trends can be seen after splitting of data.

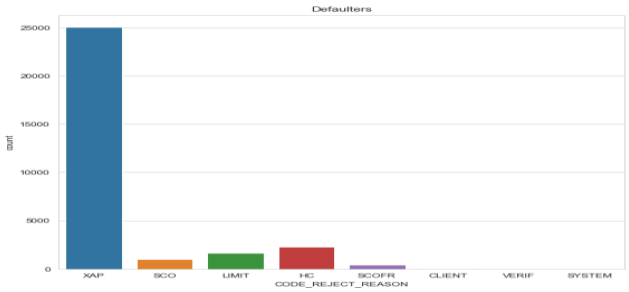
Univariate Analysis on two sets of data after merging

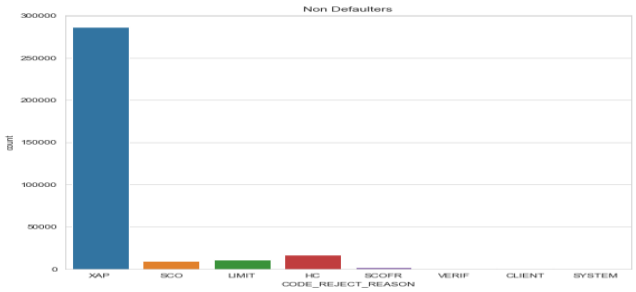


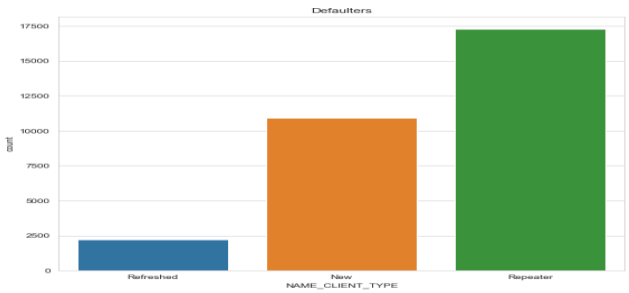


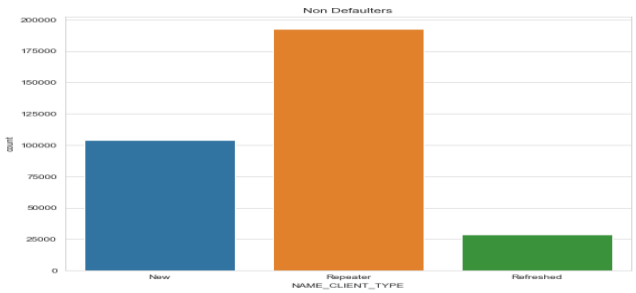


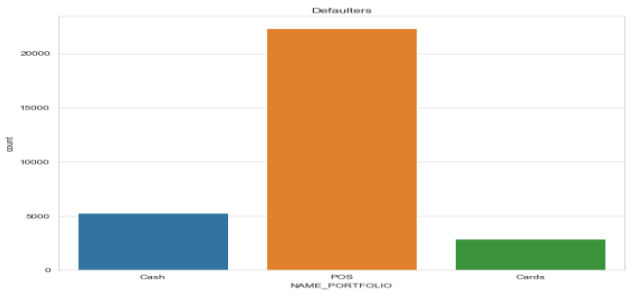


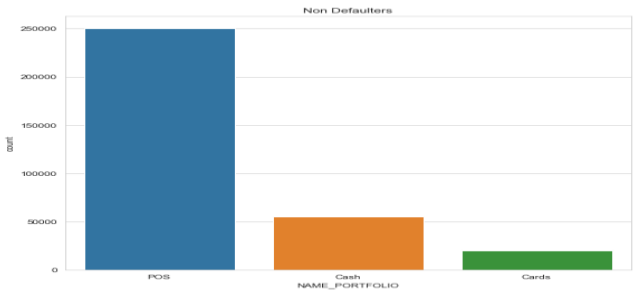


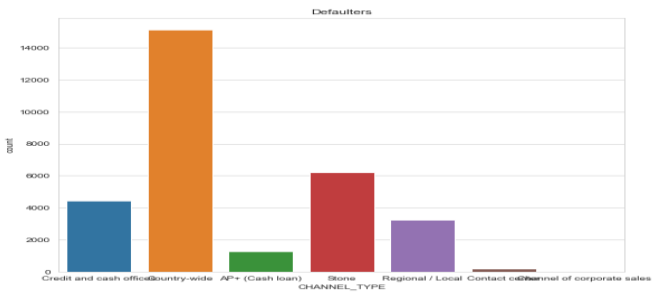


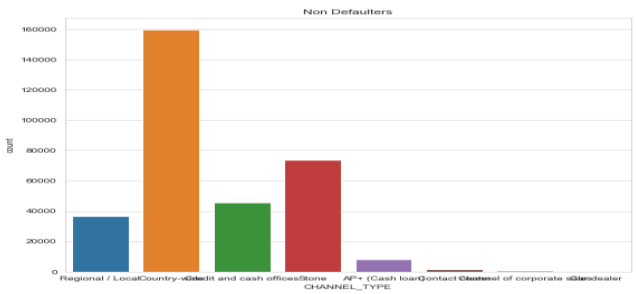








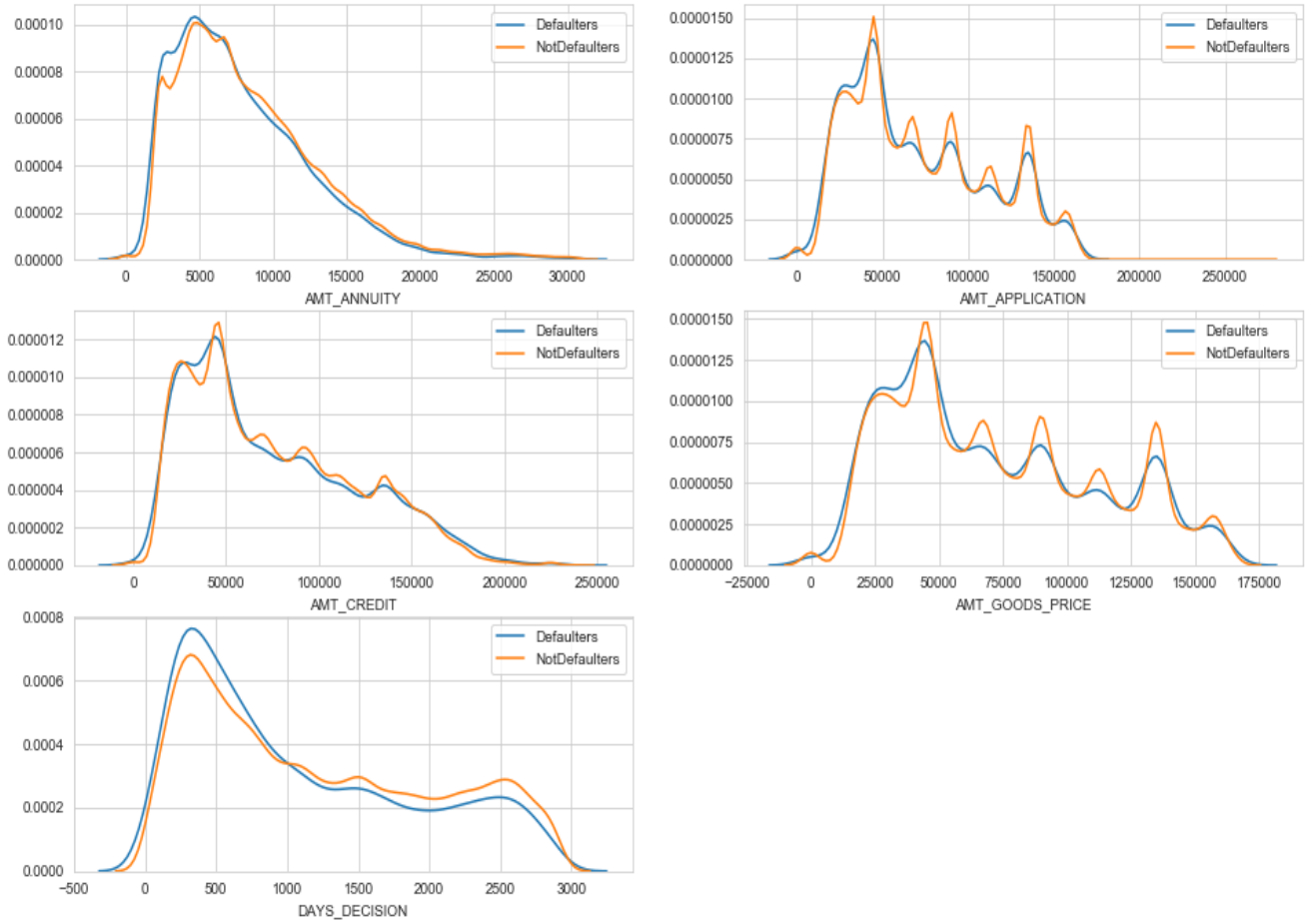




Insights from comparing the graphs:

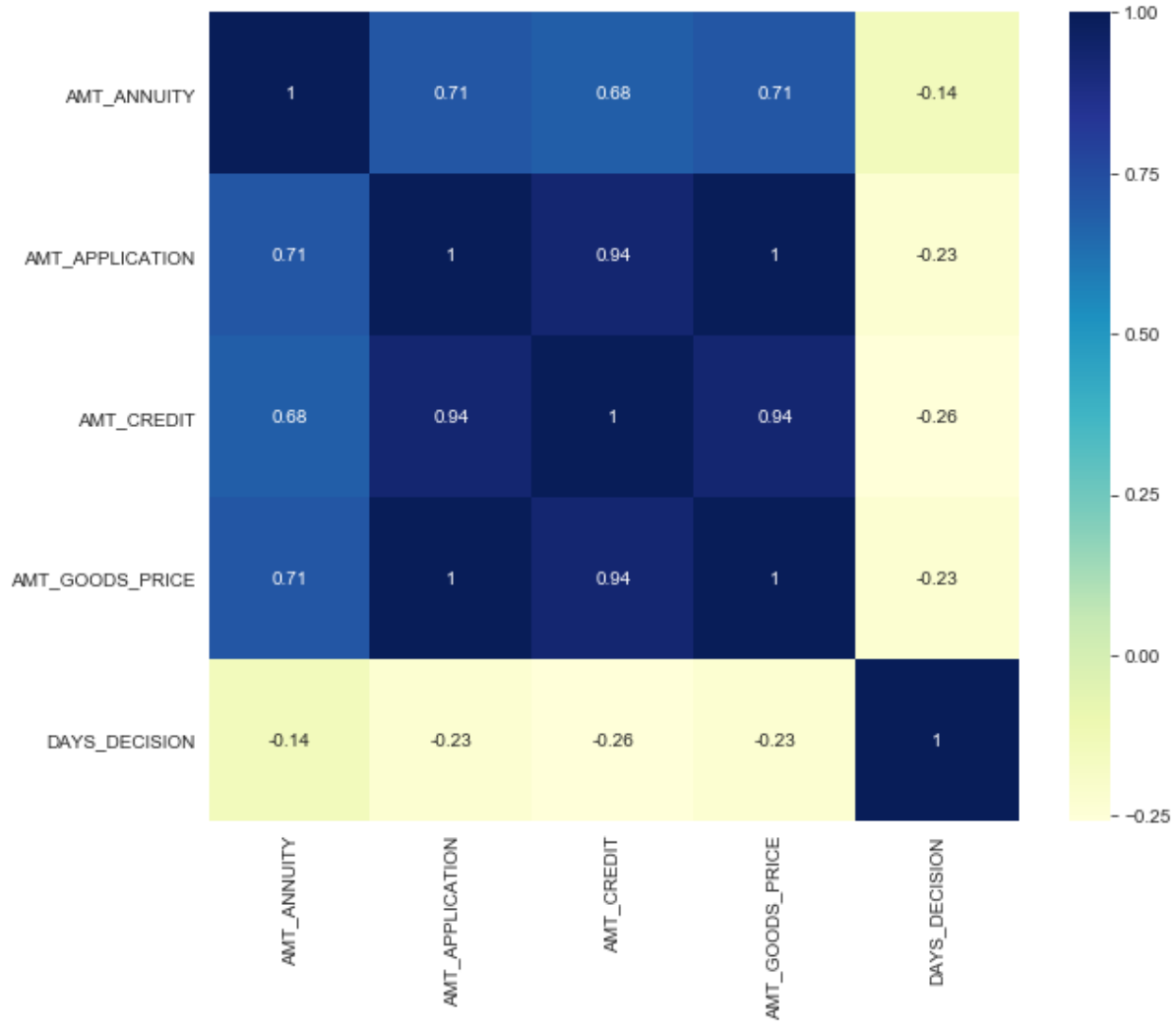
1. NAME_CONTRACT_TYPE: Most of the type of loans applied for are Consumer loans, followed by Cash loans, hence consumer loans tops the list in defaulters and non-defaulters.
2. NAME_CONTRACT_STATUS: Approx. 63% of the loans are "approved", hence comes out top on both the plots
3. CODE_REJECT_REASON: Majority(81%) of the times, the loan is rejected because of the reason code 'XAP', hence tops the list in both categories
4. NAME_CLIENT_TYPE: Clearly new clients are more Non Defaulters than defaulters, whereas repeaters tops both the list
5. NAME_PORTFOLIO: Most of the portfolios applied for is "POS", hence tops the list in both categories.
6. CHANNEL_TYPE: Credit and cash offices (42%) bring in most of the clients then followed by Country-wide and Stone, whereas most number of defaulters are from Countrywide then by Stone.

Univariate analysis for continuous variables after splitting dataframe



Insights

1. AMT_CREDIT, AMT_GOODS_PRICE and AMT_APPLICATION works very well unless it less than 50000
2. Interestingly we see a pattern here there are more defaulters when DAYS_DECISION is less, min should be around 1000 days
3. AMT_ANNUITY should be greater than 5500 for Bank's profit

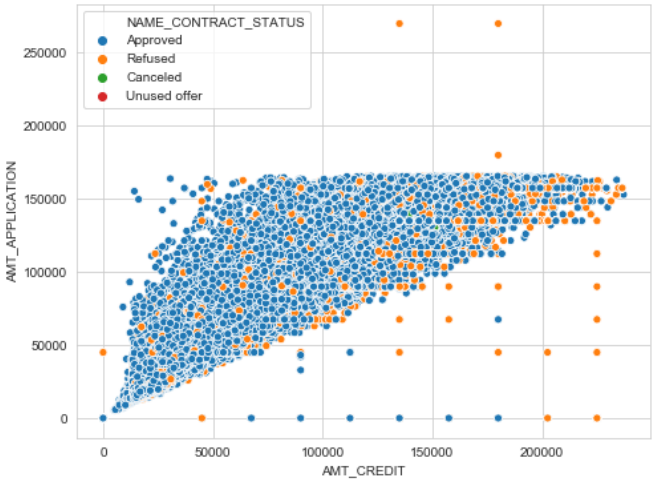


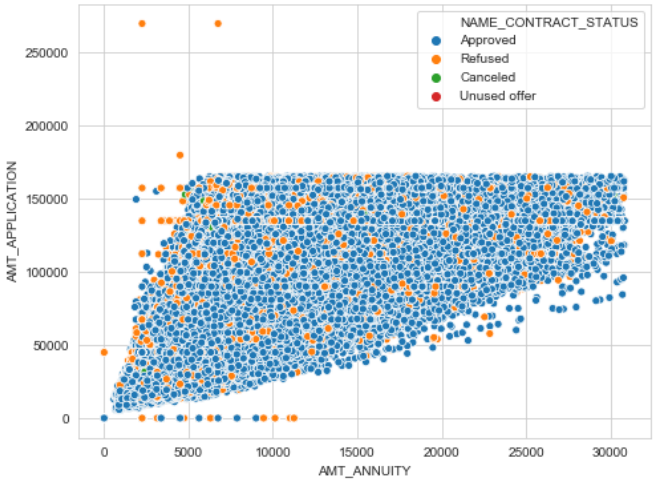
Insights

Highly correlated columns

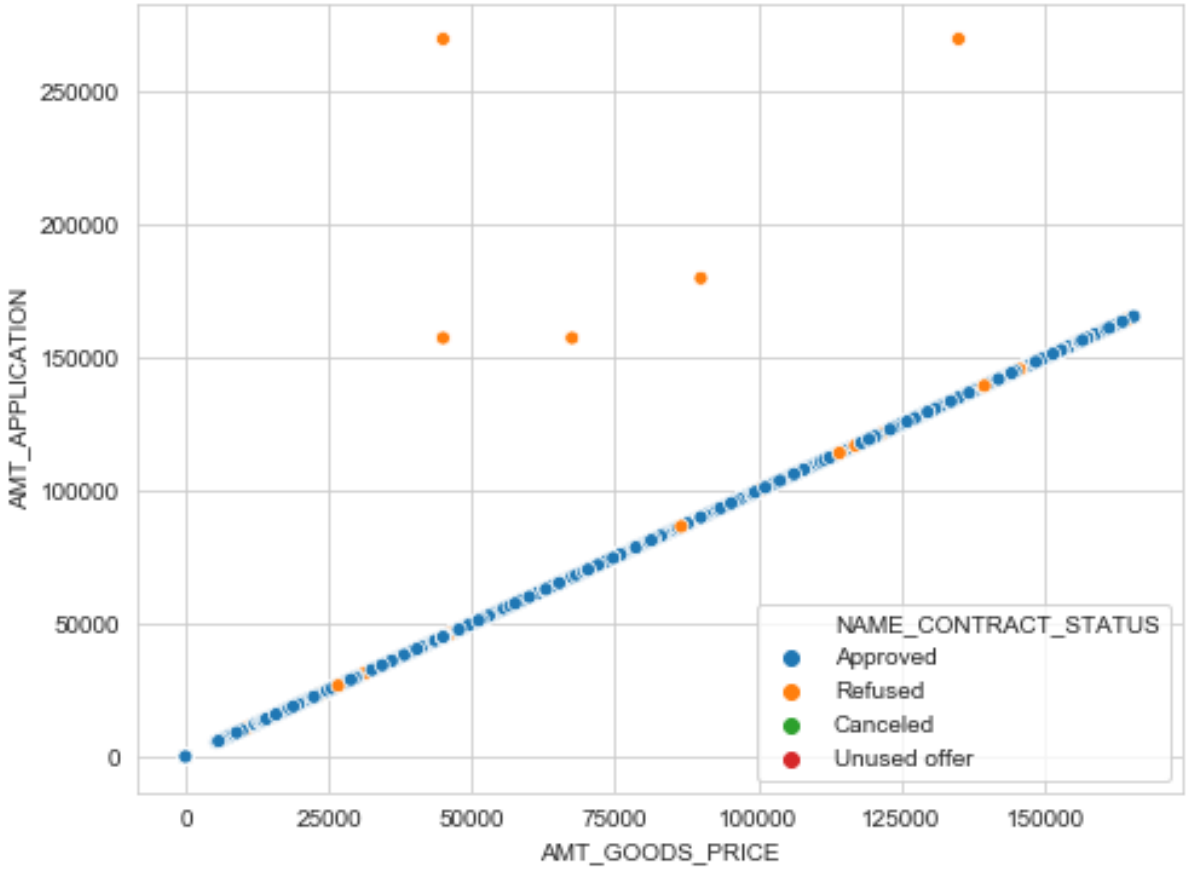
1. AMT_APPLICATION and AMT_CREDIT
2. AMT_APPLICATION and AMT_ANNUITY
3. AMT_CREDIT and AMT_ANNUITY

Bivariate analysis on continuous variable, since most correlated are AMT_APPLICATION and AMT_CREDIT then by AMT_ANNUITY





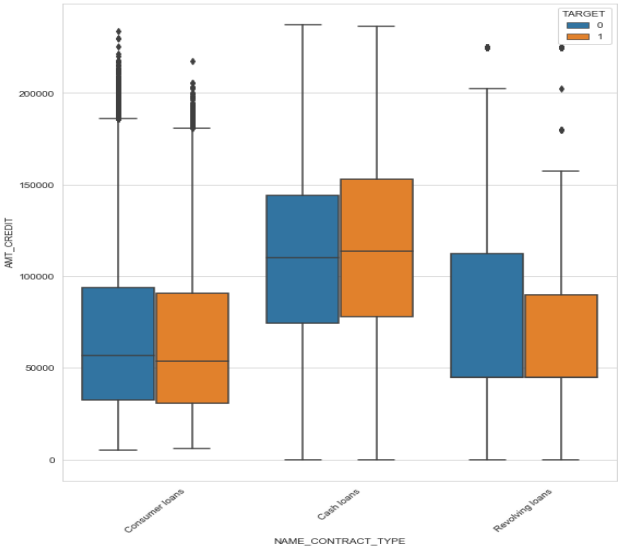
Insights:
Maximum AMT_APPLICATION most certainly that was approved was below 200000, more than that it is always rejected and also Approval chances increases when Annuity is increased.

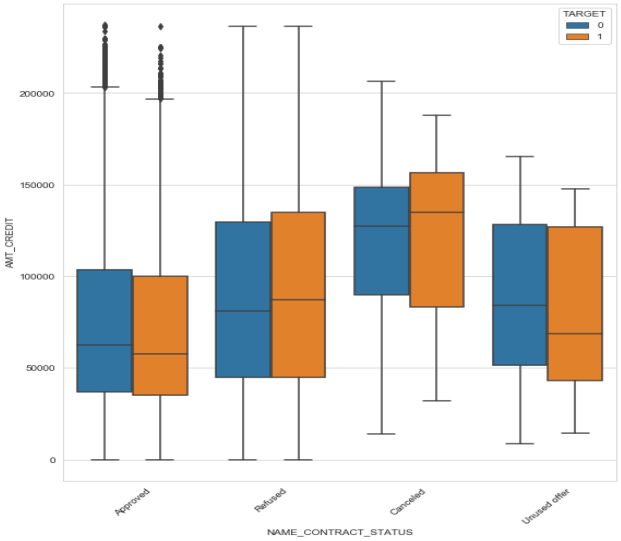


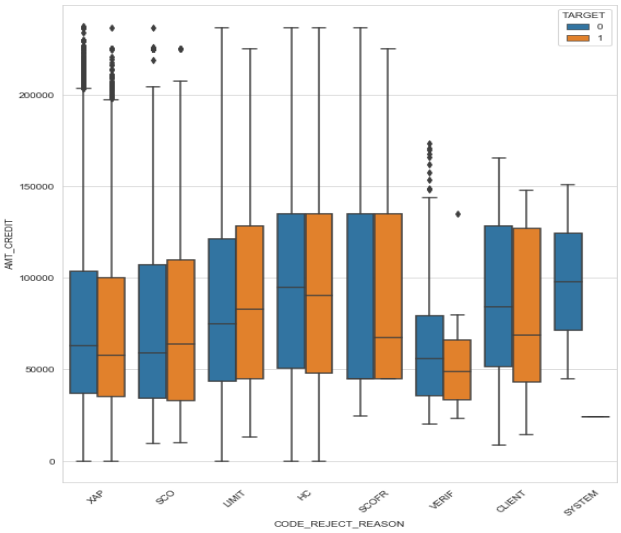
Insights:
Interestingly AMT_GOODSPRICE always increased when AMT_APPLICATION is increased.

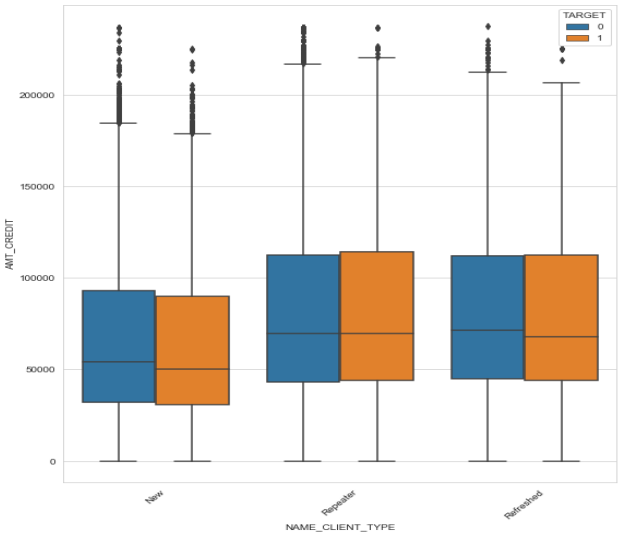
Bivariate analysis on categorical variable

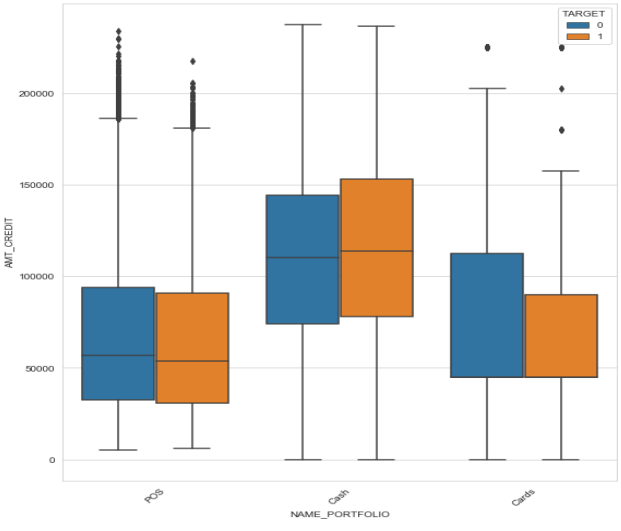
Categorical columns vs AMT_CREDIT

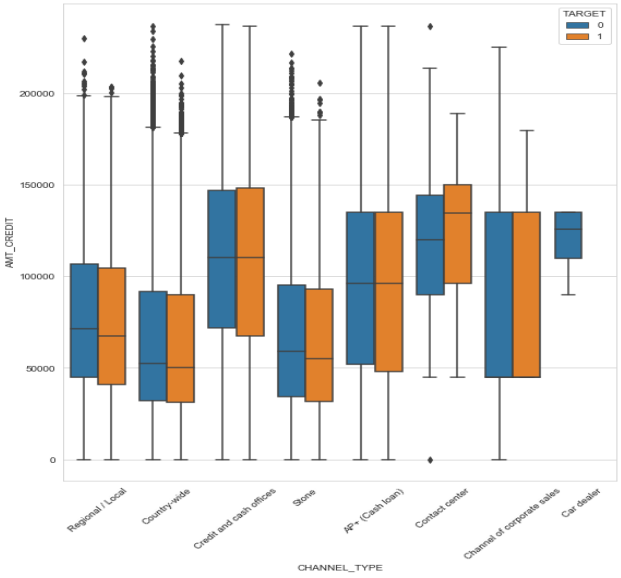












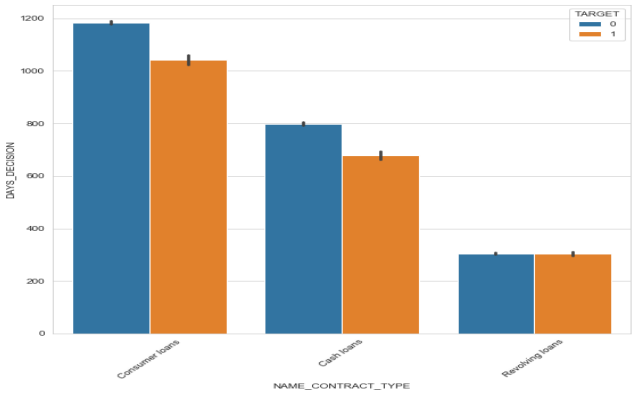
Insights:

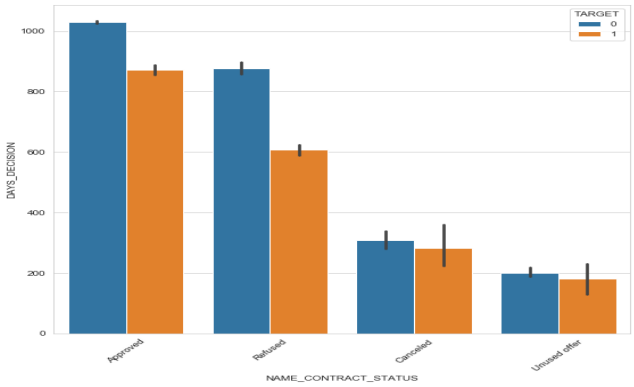
1. AMT_CREDIT for Cash loans are more than Revolving and Consumer loans for both Defaulters and Not Defaulters
2. We can see a lot of loans was cancelled in between process than refused and approved for higher credit. The cancelled loans have higher defaults in the current loan
3. For Higher credits Rejection reason were mostly SCOFR & HC for both Defaulters and Not Defaulters

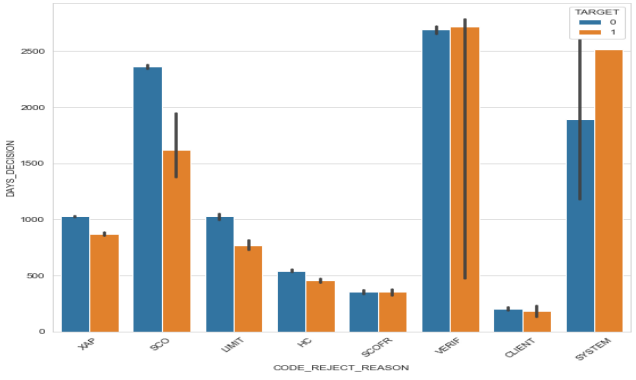
Interestingly, rejection reason, 'System' is not present for Defaulters which means none of them rejected by System got Current loan

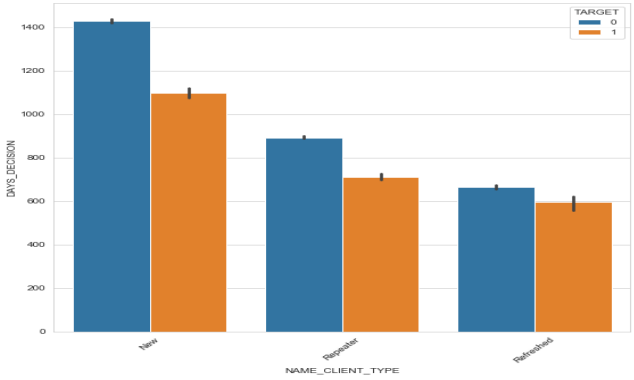
4. AMT_CREDIT for Repeater is almost same for refreshed clients and both are more than new clients for higher credits both Defaulters and Not Defaulters
5. The loan with portfolio Cash have more amount credited followed by Cards. Cash amount credit is higher for Defaulters than Not Defaulters.
6. The credit amount of the loan is more from Credit and Cash Offices and Contact Centers for both Defaulters and Not Defaulters

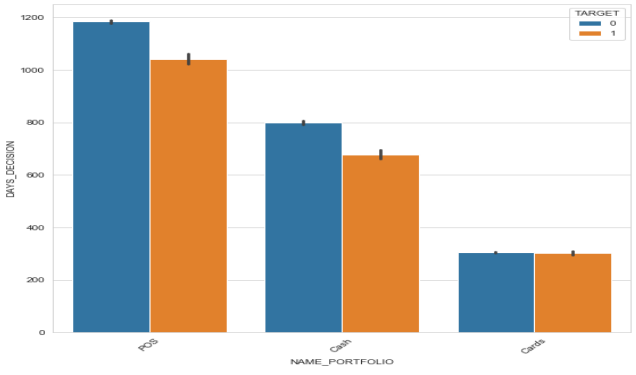
Let's analyze the categorical columns vs DAYS_DECISION

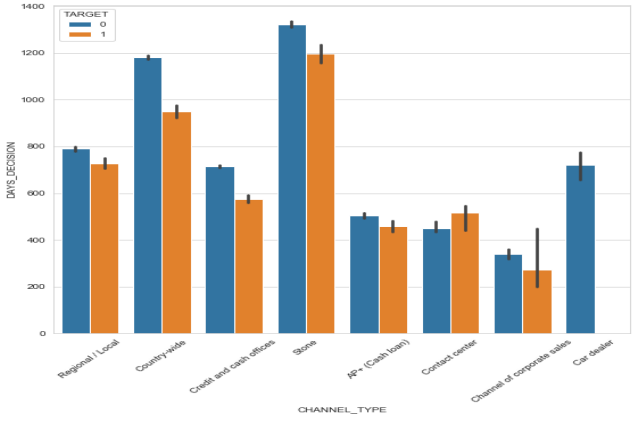








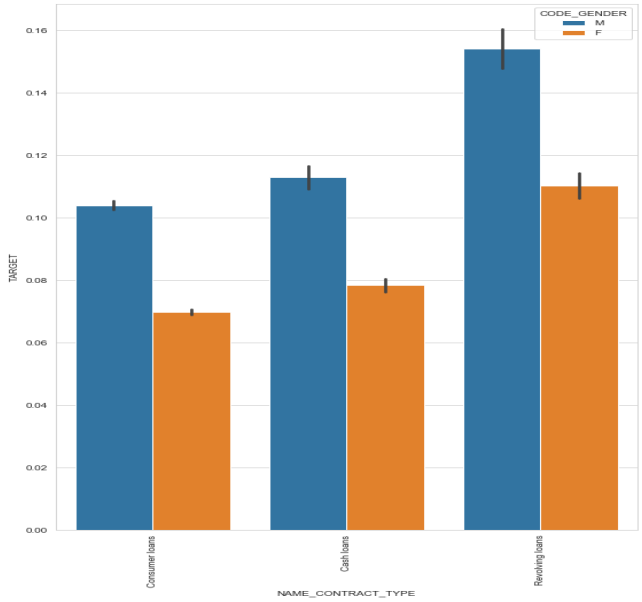


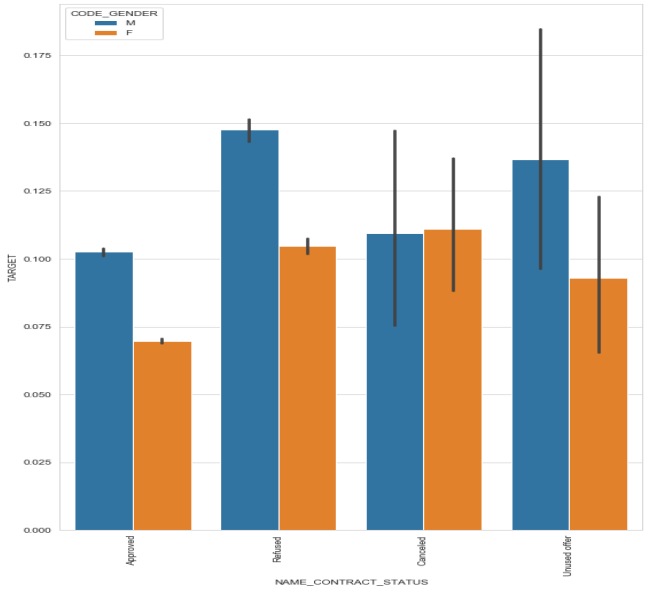


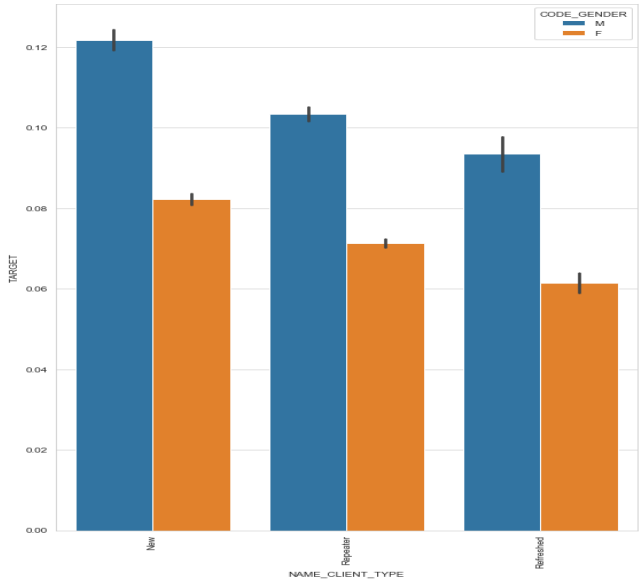
Insights:

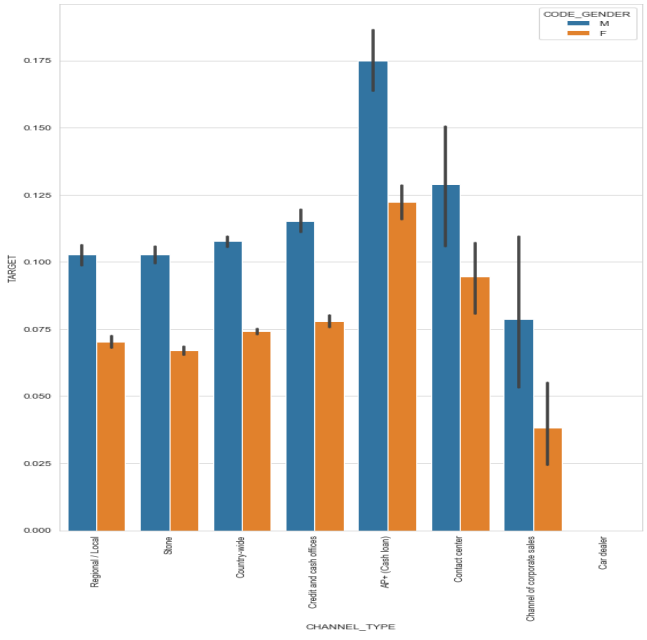
1. Consumer loans take the most number of days for decision of loans for both Defaulters and Not Defaulters.
2. Interestingly, Unused offer (NAME_CONTRACT_STATU) take the least number of days for decision for both Defaulters and Not Defaulters.
3. Code Reject reason 'Verif' takes the most number of days, followed by 'System' for both Defaulters and Not Defaulters
4. Client Type, New takes the most number of days for loan processing decision (approx. double than Refreshed) which is understandable as the credit company will be more cautions and do its KYC properly for new customers.
5. Portfolio type 'POS' has the highest number of days for decision, while 'cards ' have the lowest for both Defaulters and Not Defaulters.
6. Channel Type 'Stone' takes the most number of days for decision for both Defaulters and Not Defaulters.

Analysis of two segmented variables after merge





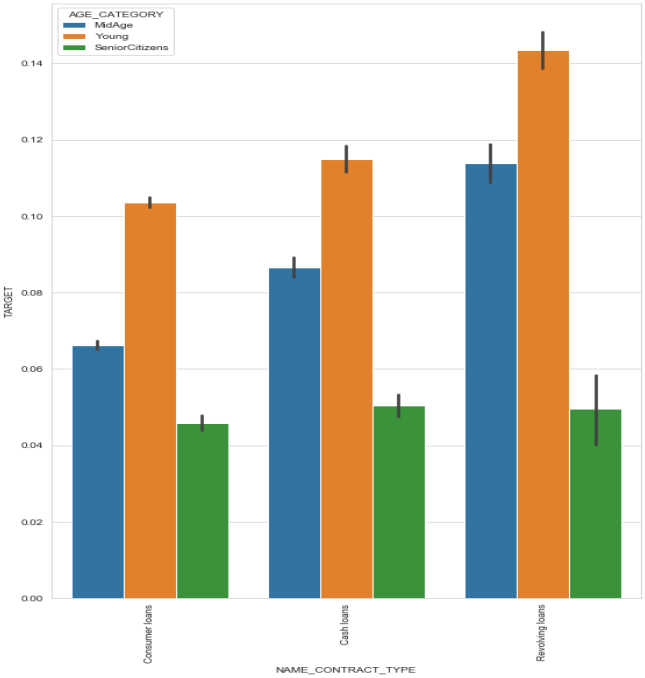


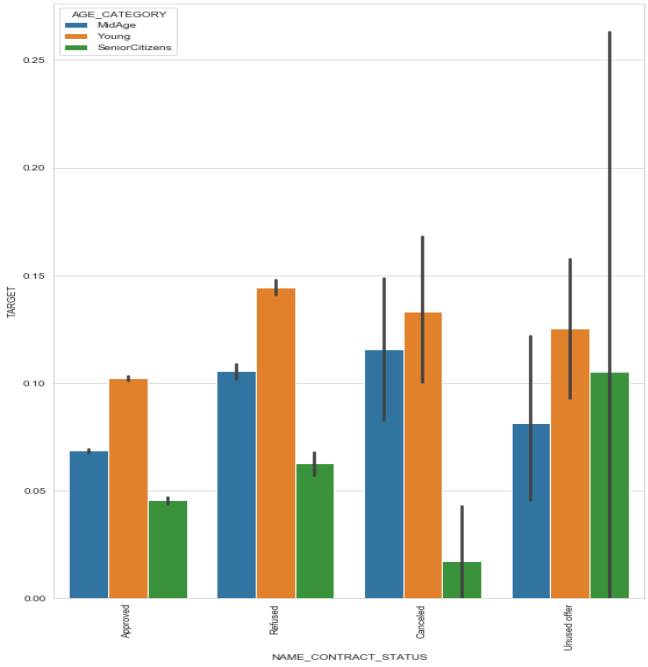


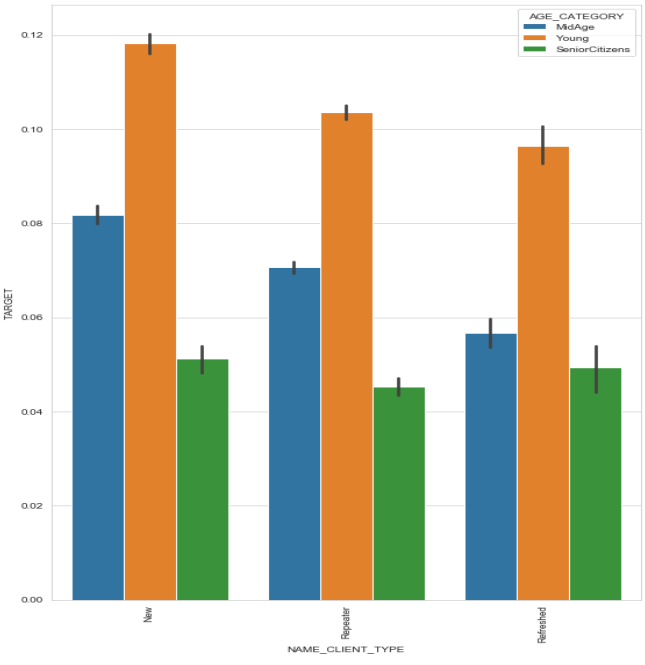
Insights:

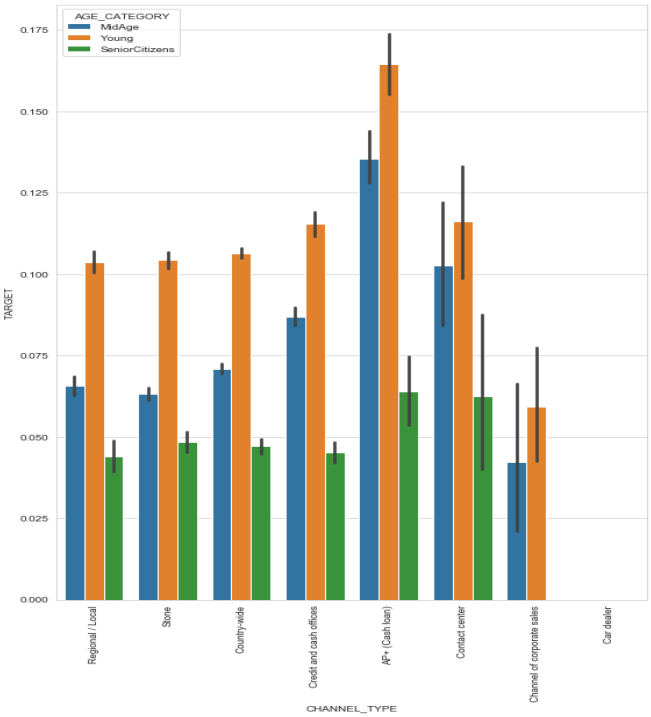
These above plots shows us defaulters mean in each category

1. Male and Female clients both with Revolving Loans tends to default more than other loans.
2. Male clients for whom the loan was previously refused will default more followed by Unused Offer whereas for females for whom application was cancelled tend to default more followed by Refused.
3. New clients default more than repeater and refreshed
4. Both Male and Female clients with AP+(Cash Loans) default more than other channel types.





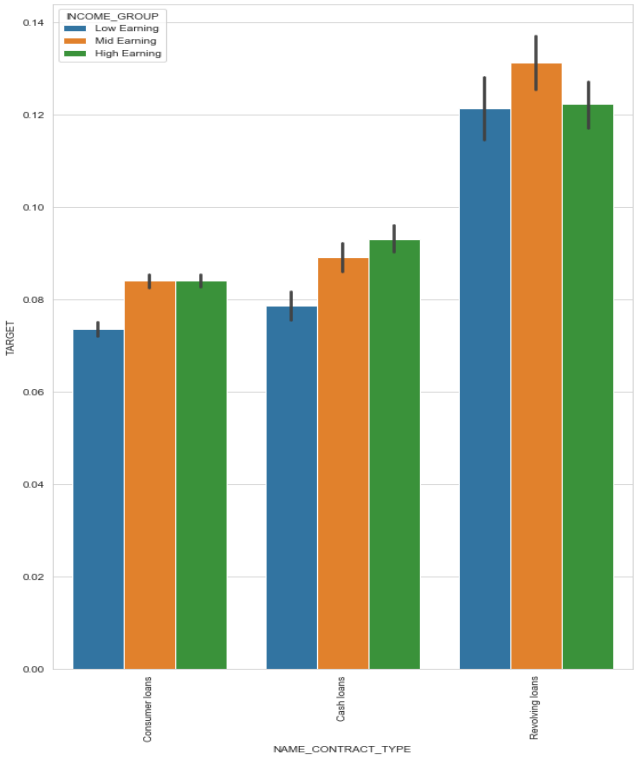


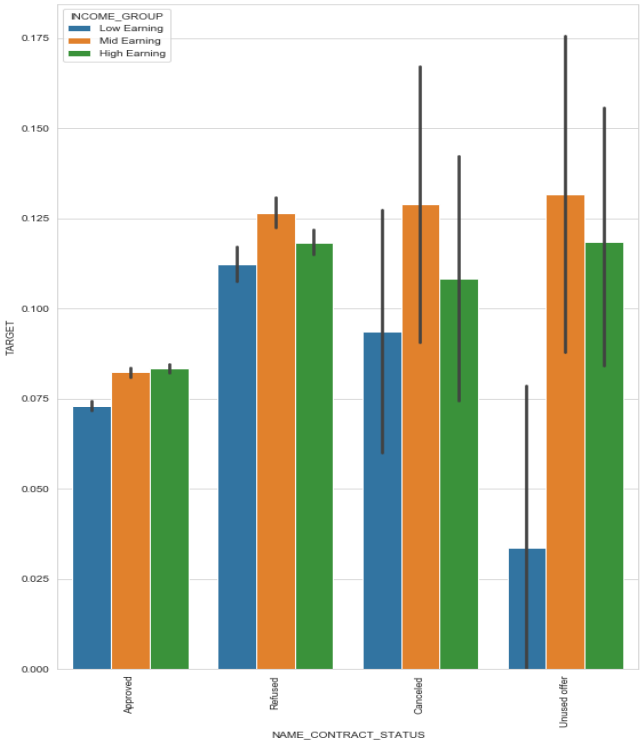


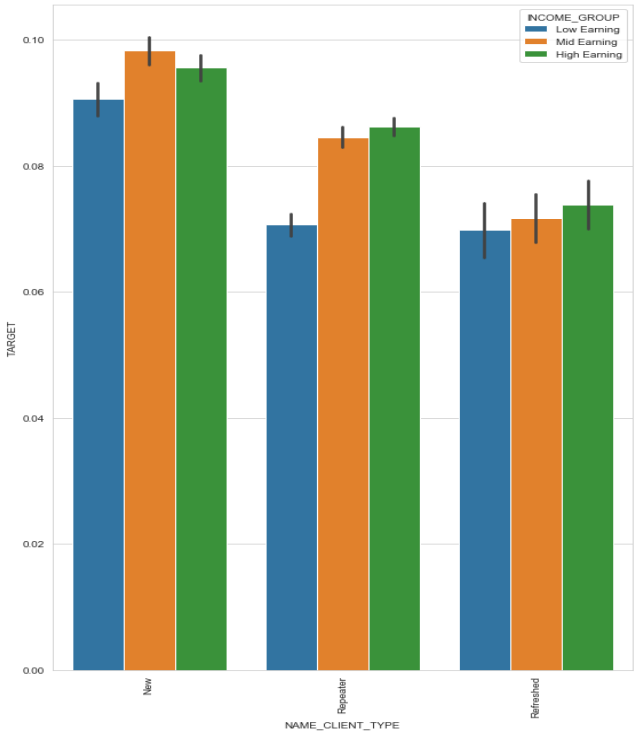
Insights:

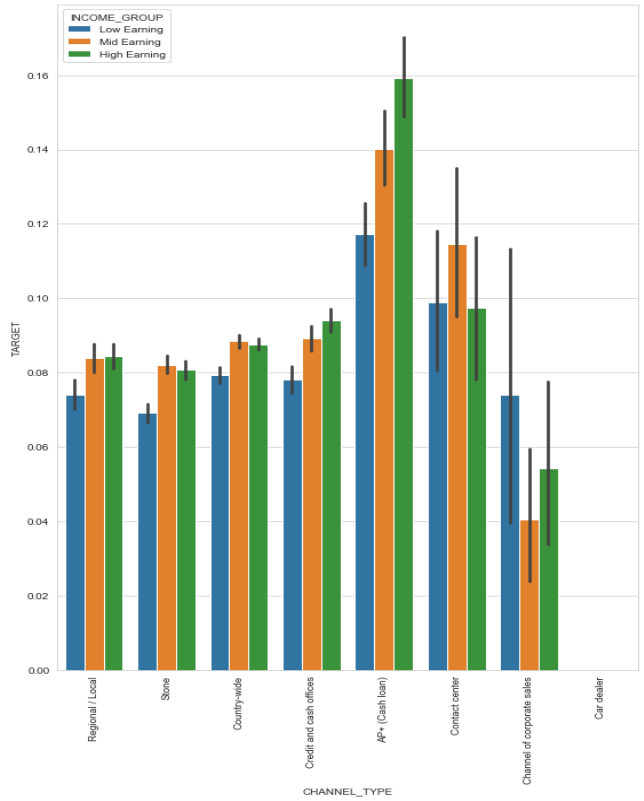
These above plots shows us defaulters mean in each category

It's clear that young clients default more than mid-Age and seniors in all types of loans irrespective of Client Types, Channel Types and Contract status.









Insights

1. Clients with any Income Groups tend to default more in Revolving Loans.
2. Clients with Mid Income Group default more if their was application was cancelled previously then by Refused where as for other income groups if application was refused previously chances of defaulters are more.
3. New Clients default more than repeater and refreshed for all income Groups
4. Clients with AP+(Cash Loans) default more than other channel types for all income groups.

SUMMARY:

1. Male Clients default more than female clients.
2. Male Clients who are young, Single/Unmarried and especially New have a high rate of defaults.
3. Female clients default for lesser amount of credit as compared to male clients.
4. Female working as Low skilled laborers and Waitress /Barmen Staff are the most among defaulters
5. Young clients have higher risk associated as compared to middle age or older clients.
6. State Servants and Pensioners having high amount credit in loans default more.
7. Clients with Lower Secondary education are most to default among both men and women
8. Clients for whom the loan was previously refused or cancelled default more.
9. Clients requiring AP+(Cash Loans) default more.

THANK YOU!!!