EDA CASE STUDY

Submitted by:
Rajesh Mahendra M
Dada Nanjesha GS

Problem Statement:

- The loan providing companies want to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

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

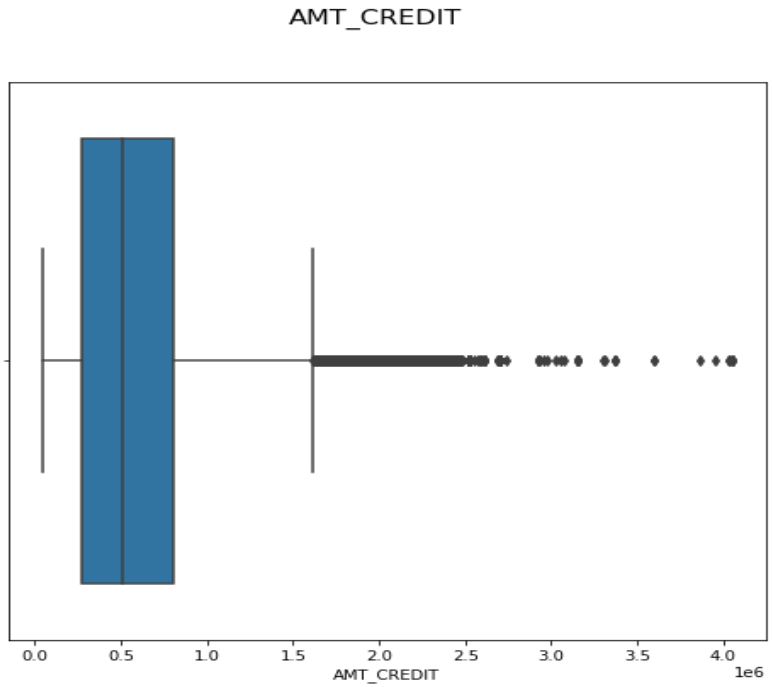
- If the applicant is likely to repay the loan, then not approving the loan results is a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

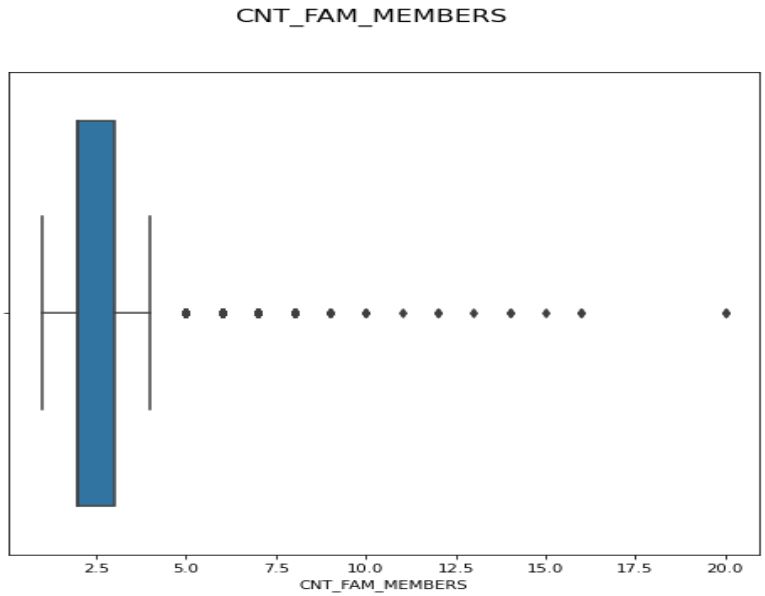
Statement of Purpose:

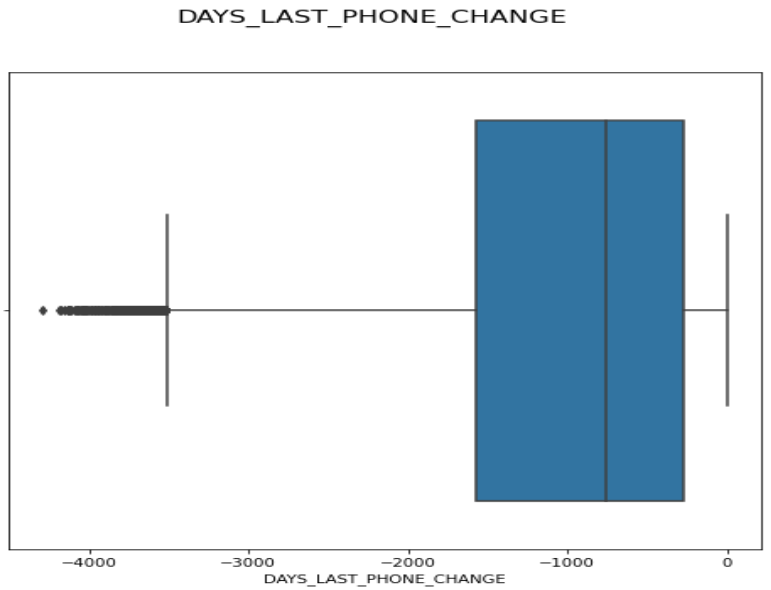
Primary Objective of this analysis is to identify variables which are strong indicators or drivers to find out clients potential for loan credit and customers who aren't and to use these insights in approval / rejection decision making.

Outliers in the given data frame

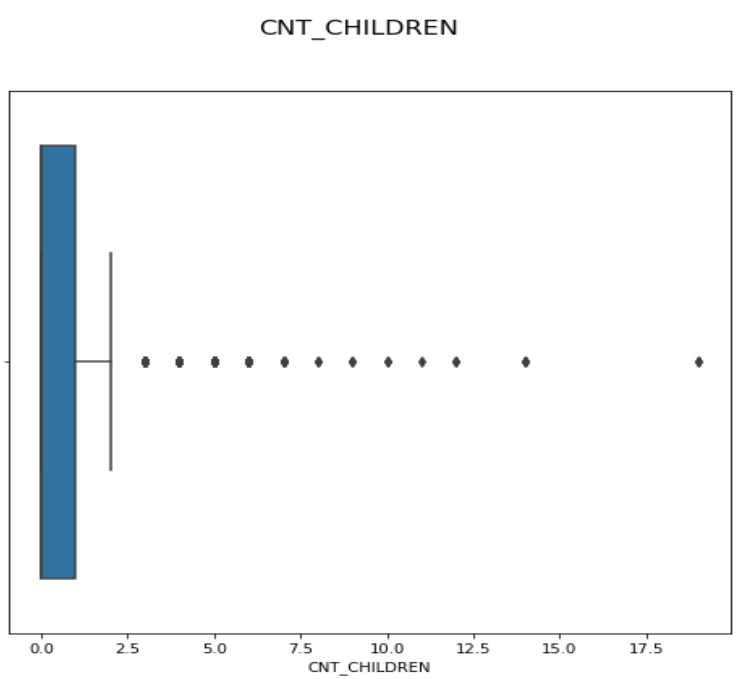
Spot outliers in the columns and find reasons for this outlier value presence. Here are some of the values we could spot as outliers with the help of plots:

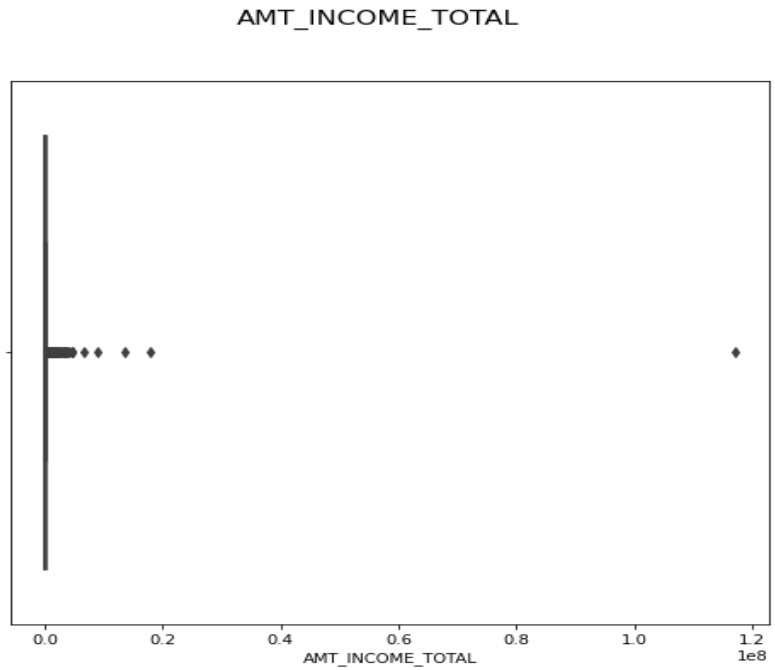




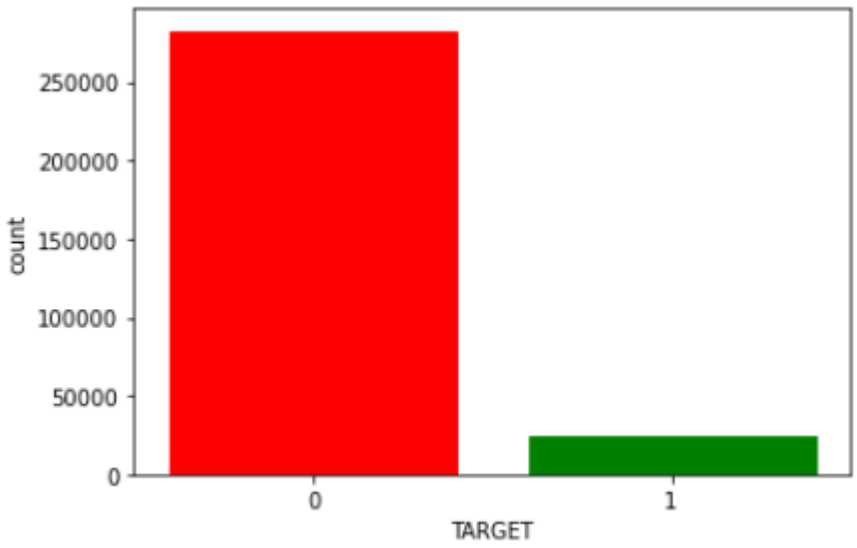


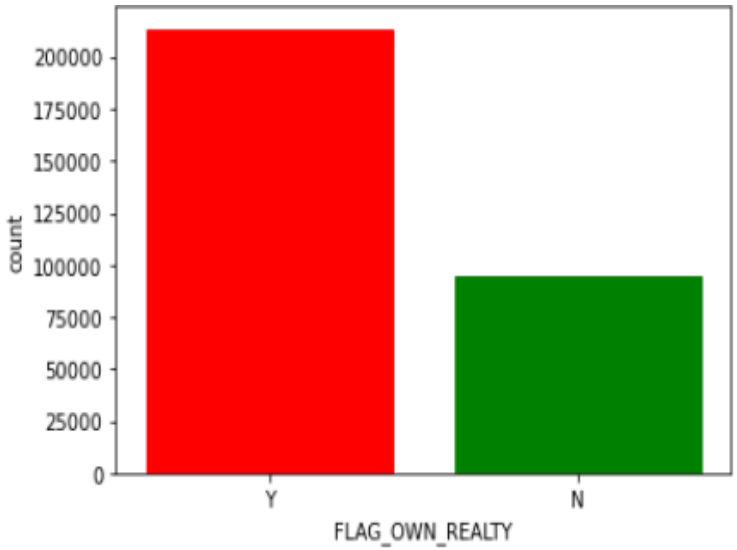
Box plot for CNT_CHILDREN shows a large outlier(19). Since a family cannot or very rare to have 19 children.
In the plot AMT_INCOME_TOTAL, we can visually see that the MAX amount is way largest than the other statistical data

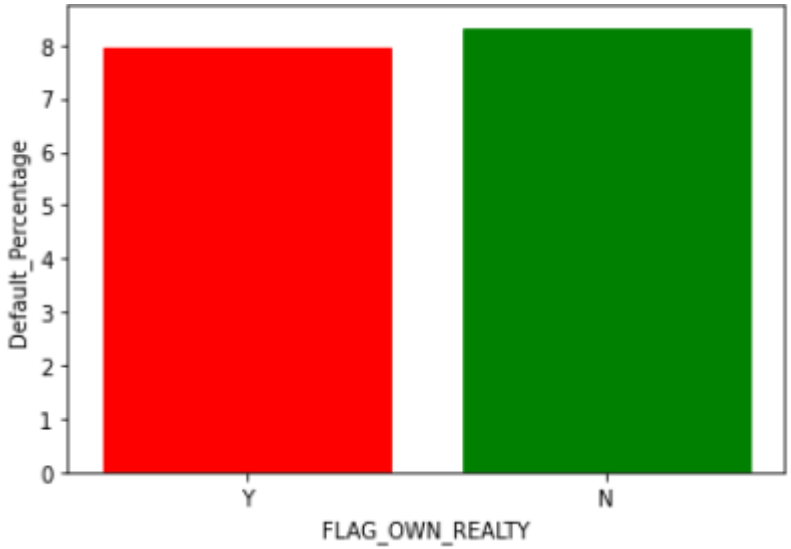




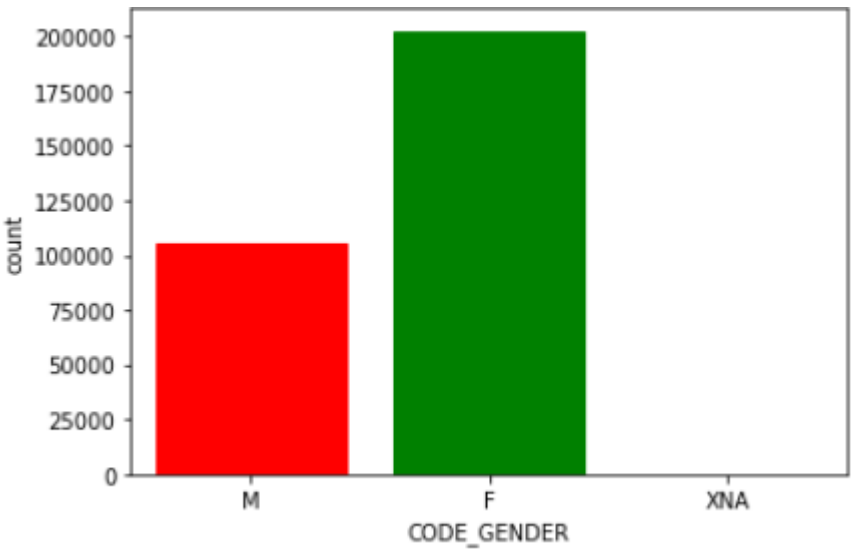
Analyzing the count of Target variables

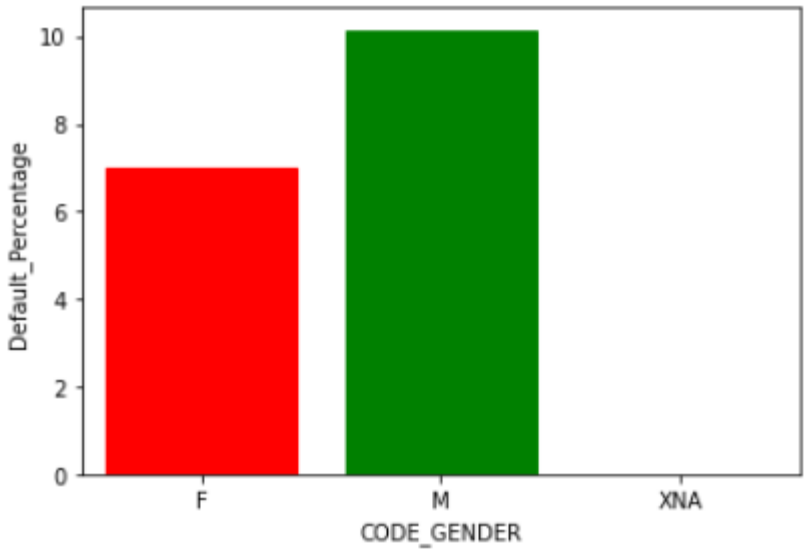




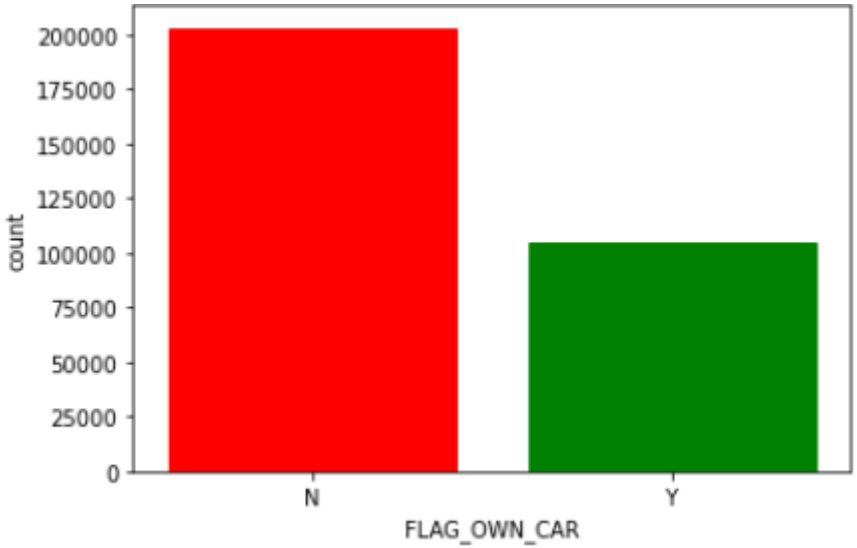


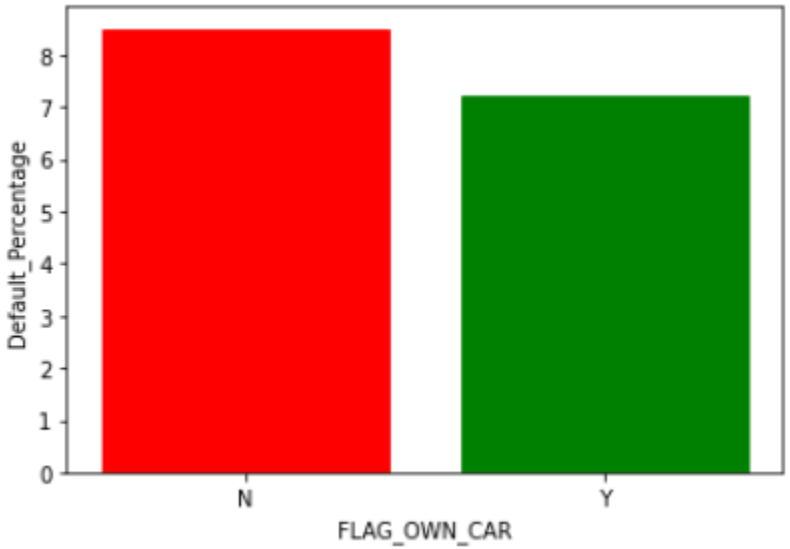
CODE GENDER



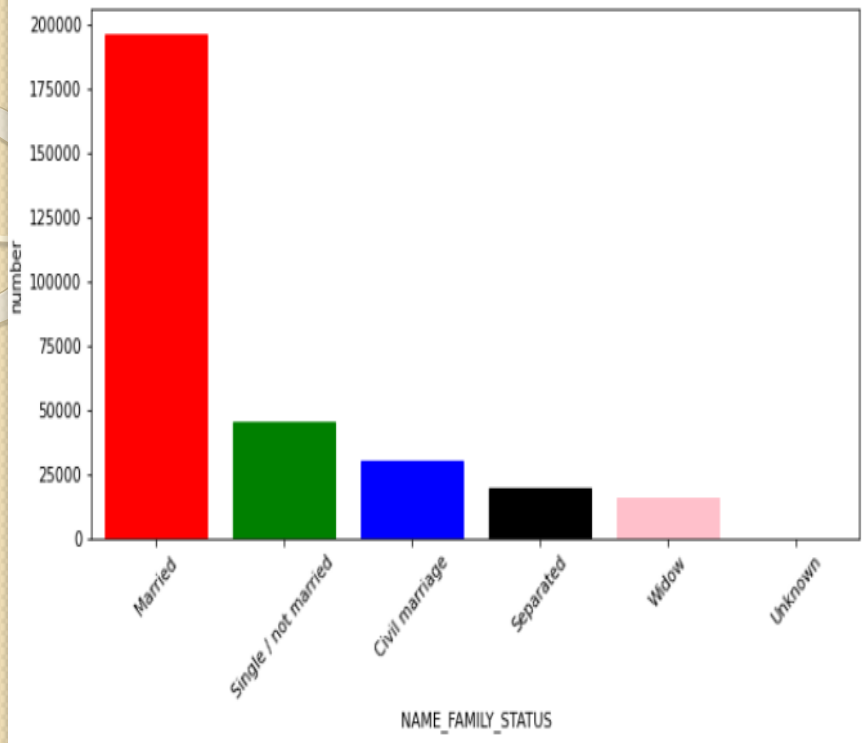


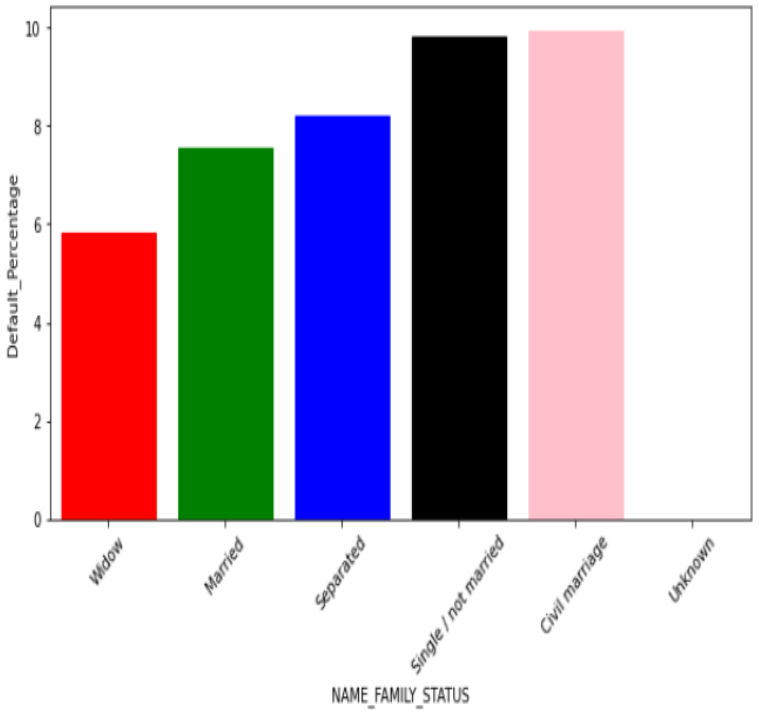
FLAG_OWNCAR



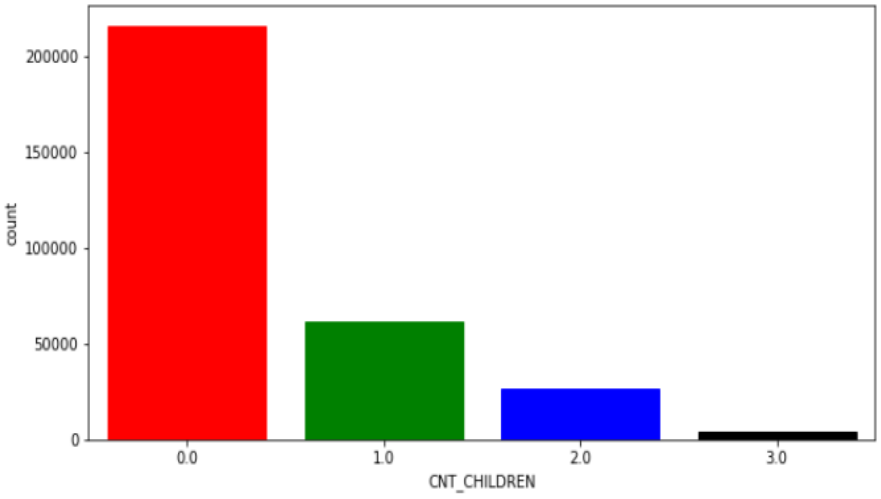


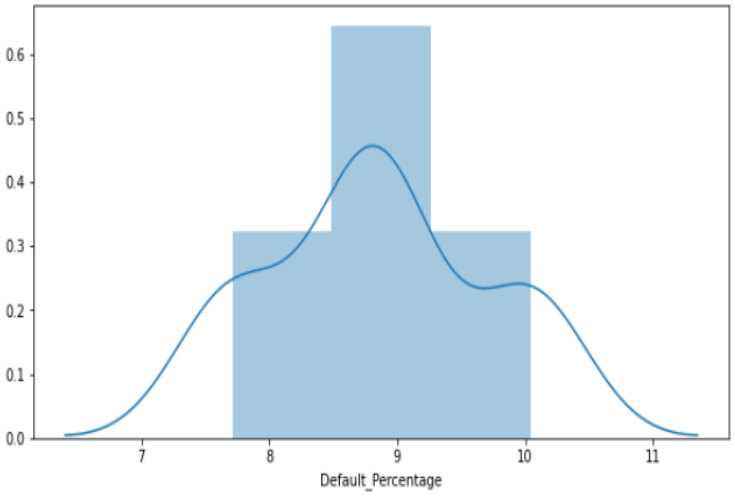
NAME_FAMILY_STATUS



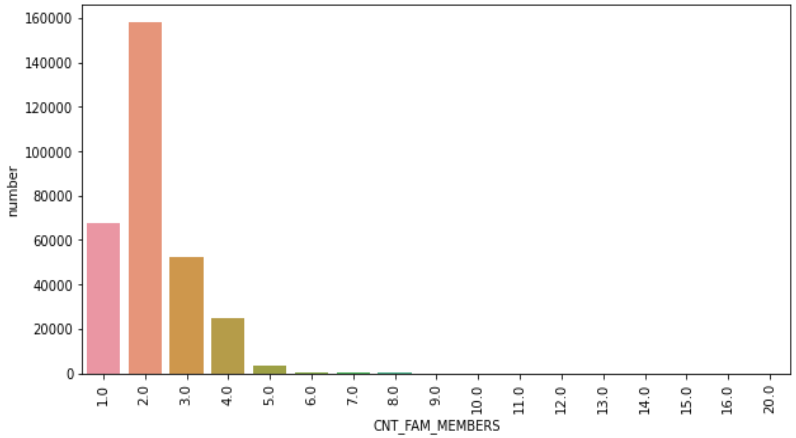


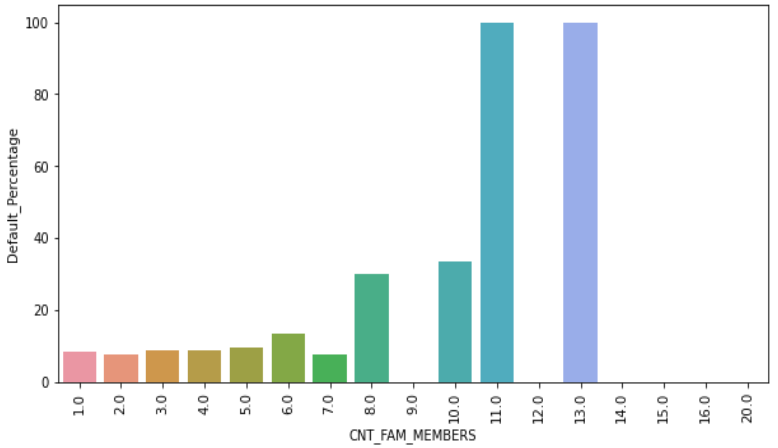
CNT_CHILDREN



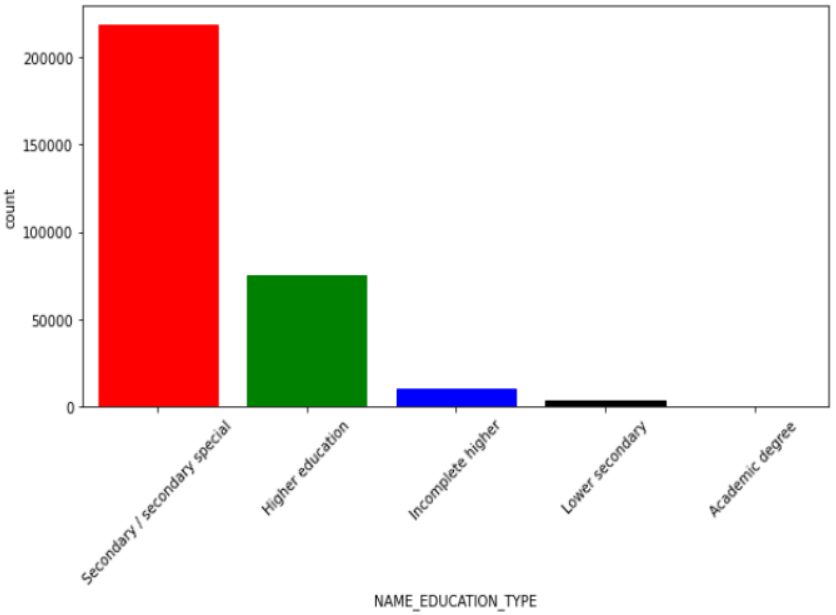


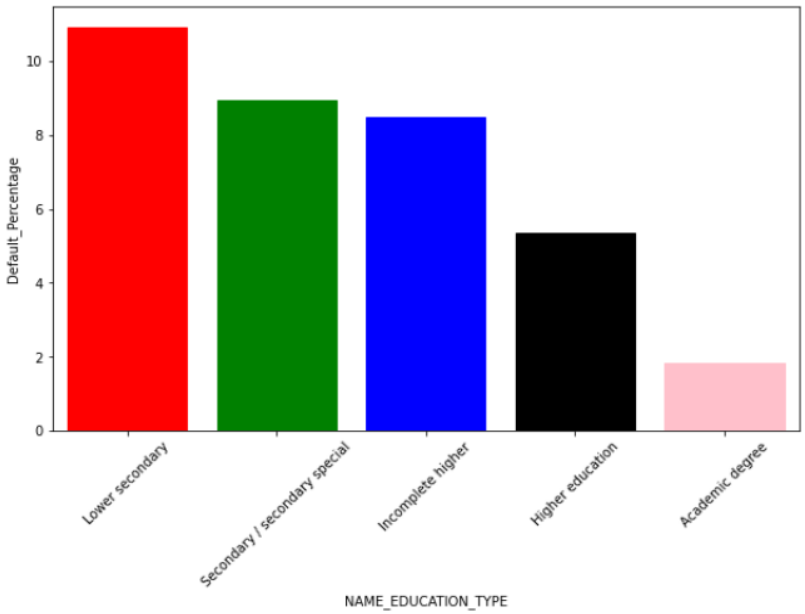
CNT_FAM_MEMBERS



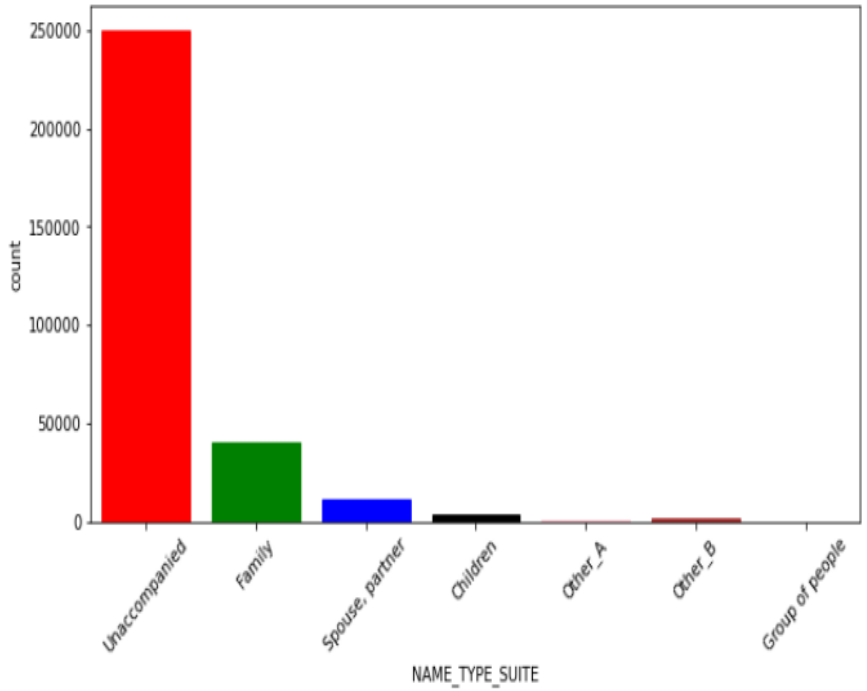


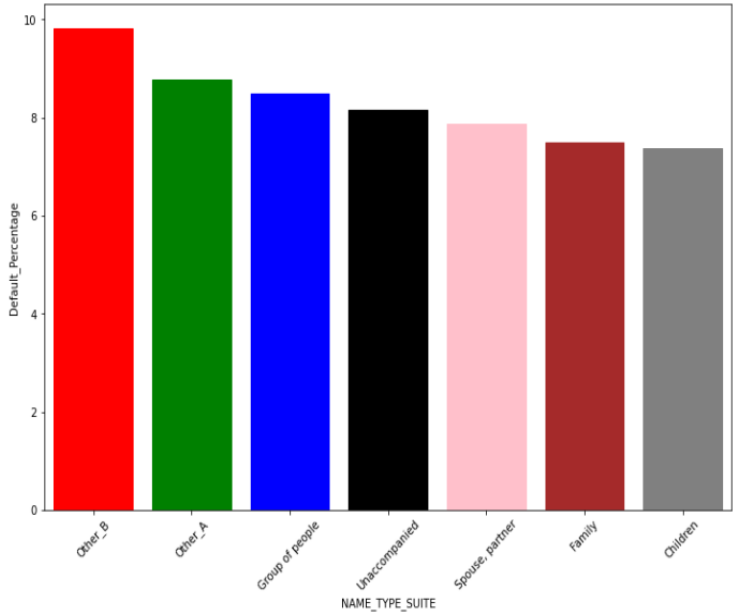
NAME_EDUCATION_TYPE



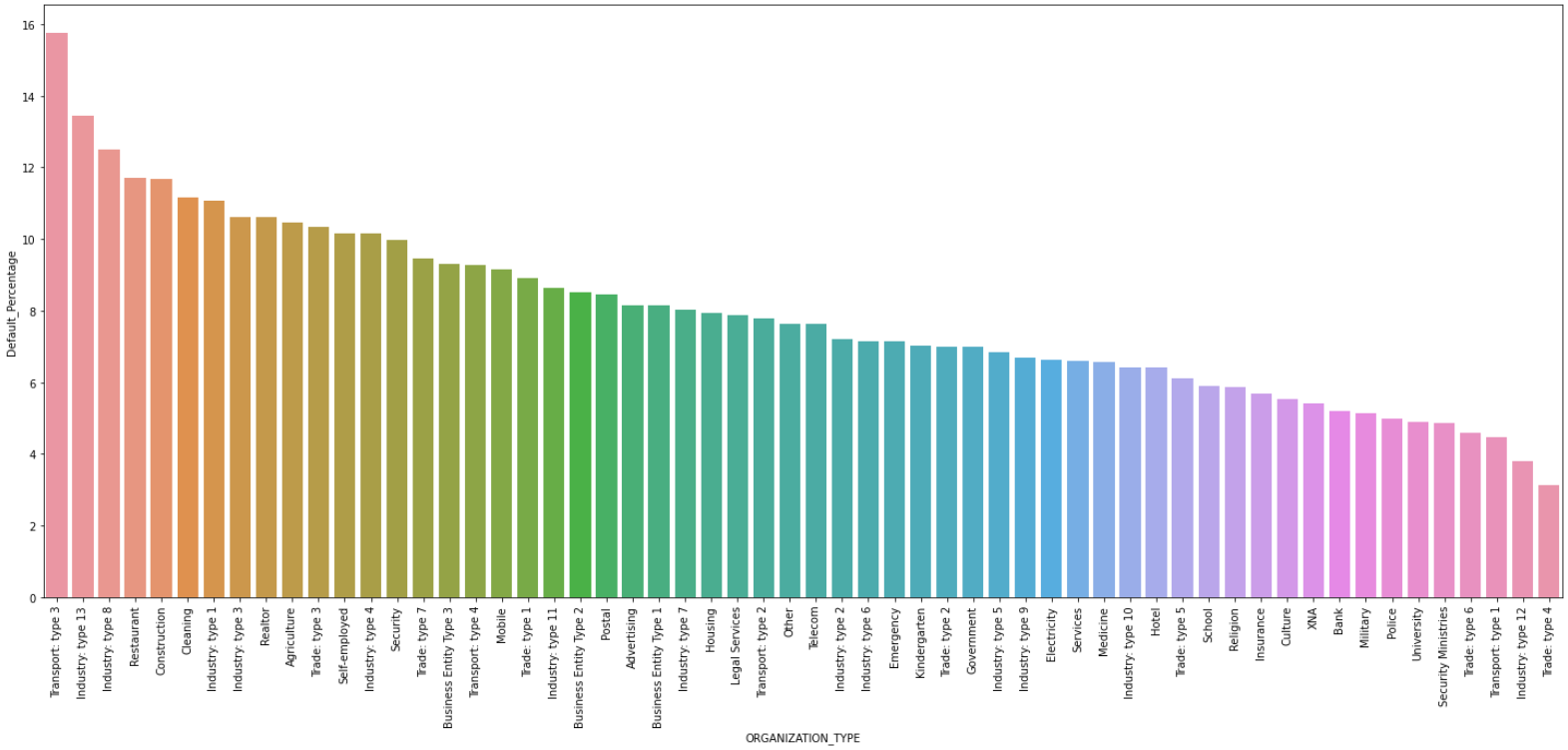


NAME_TYPE_SUITE

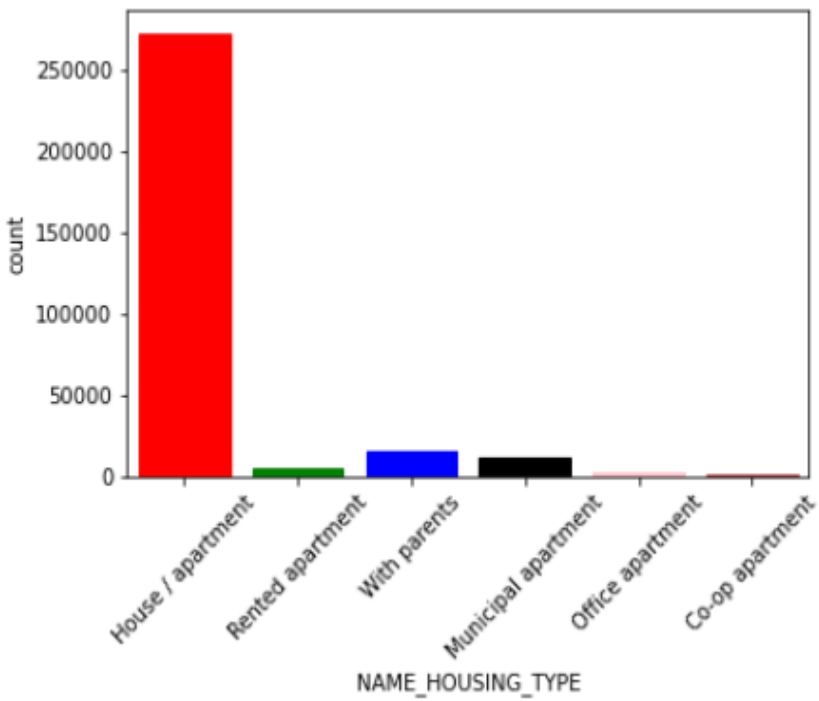


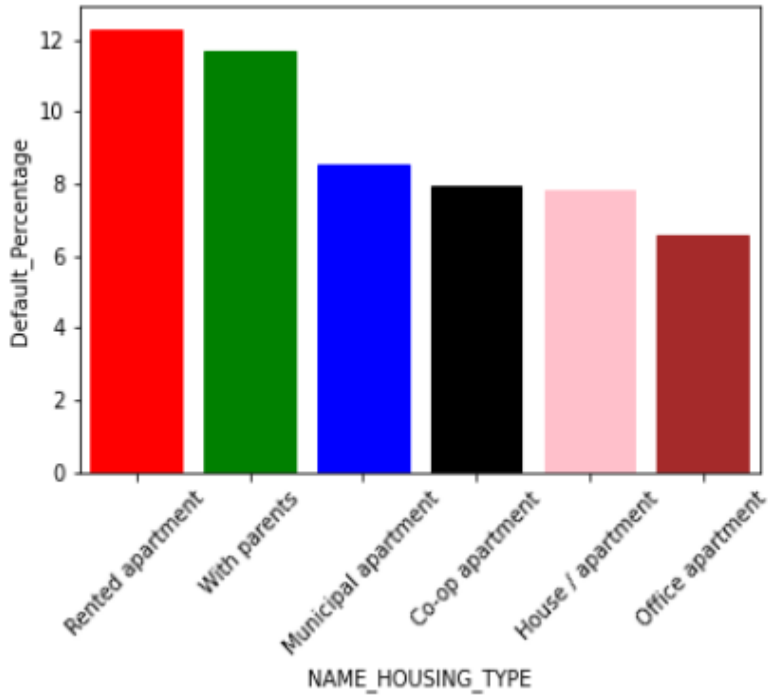


ORGANISATION_TYPE



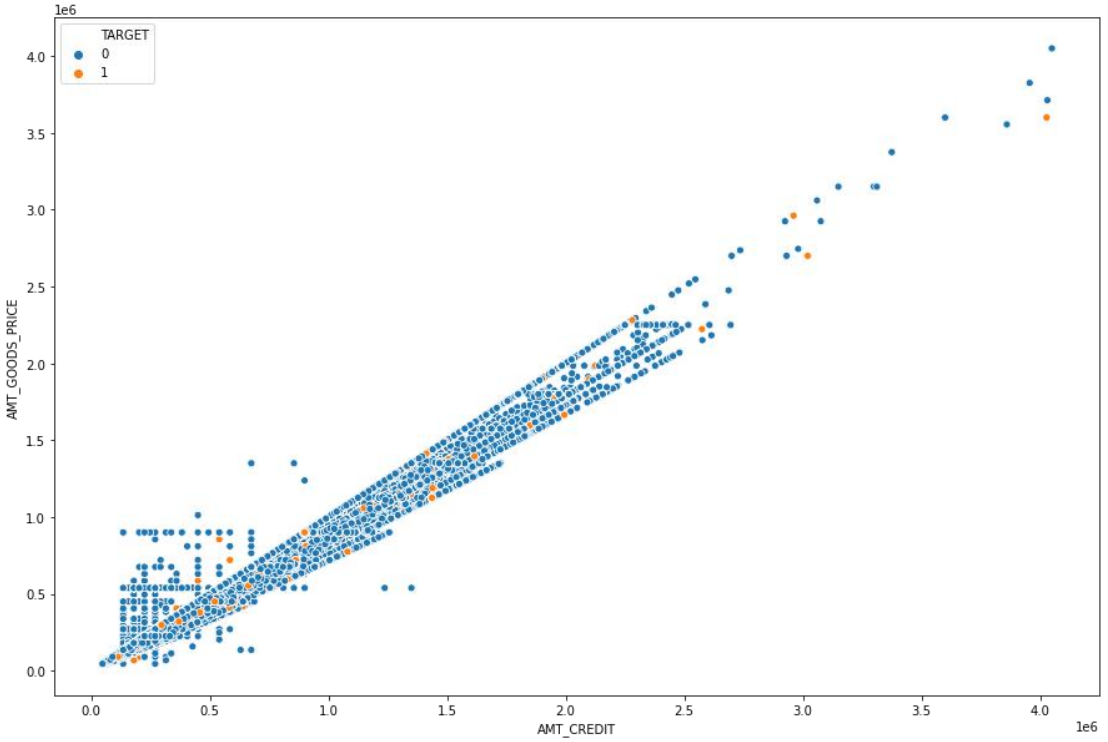
NAME_HOUSING_TYPE





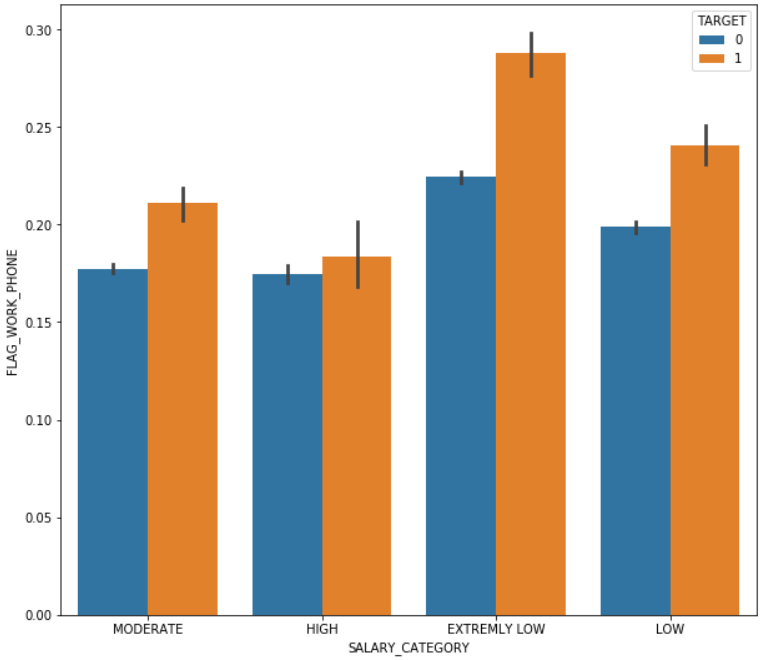
BIVARIATE ANALYSIS

AMT_CREDIT vs AMT_GOODS_PRICE



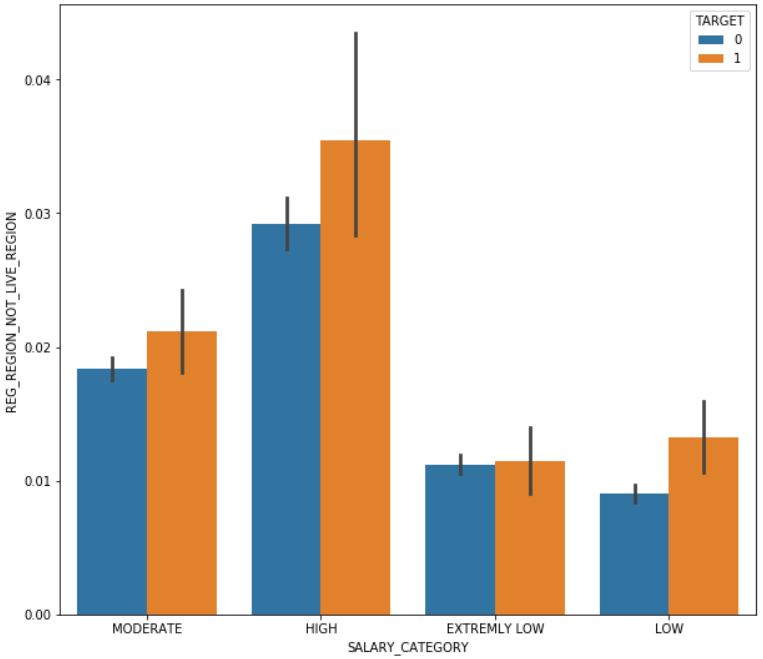
We found that Credit amount and the Amount goods price are more correlated with the Defaulters. The Defaulters are linearly increasing as these both variable increases.

Salary Category vs Client who provided Home Number



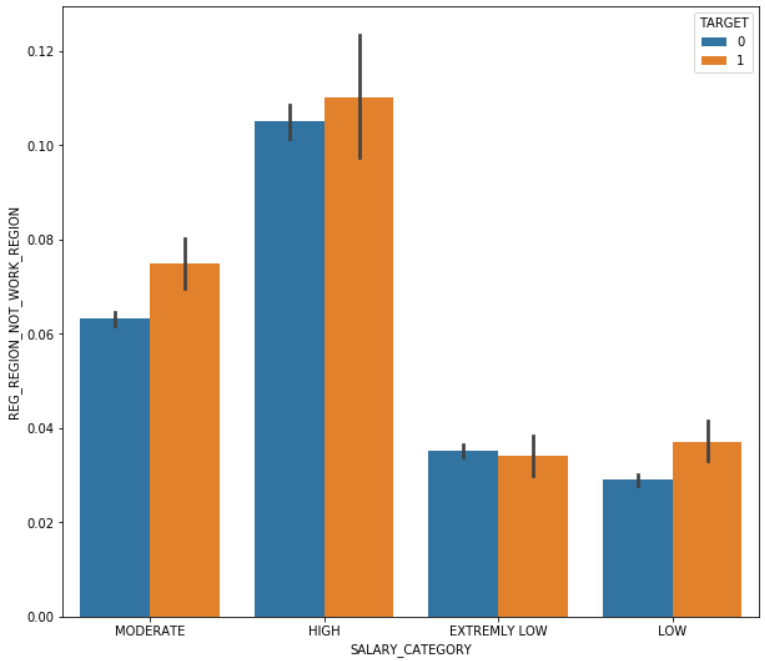
Client with Extremely low salary has more chance to be a Defaulter, when he did not provide the Home phone number. Here approximately 30% people only produced the phone number

SALARY VS CLIENT WHOSE PERMANENT ADDRESS NOT MATCH WITH CONTACT ADDRESS -REGION LEVEL.



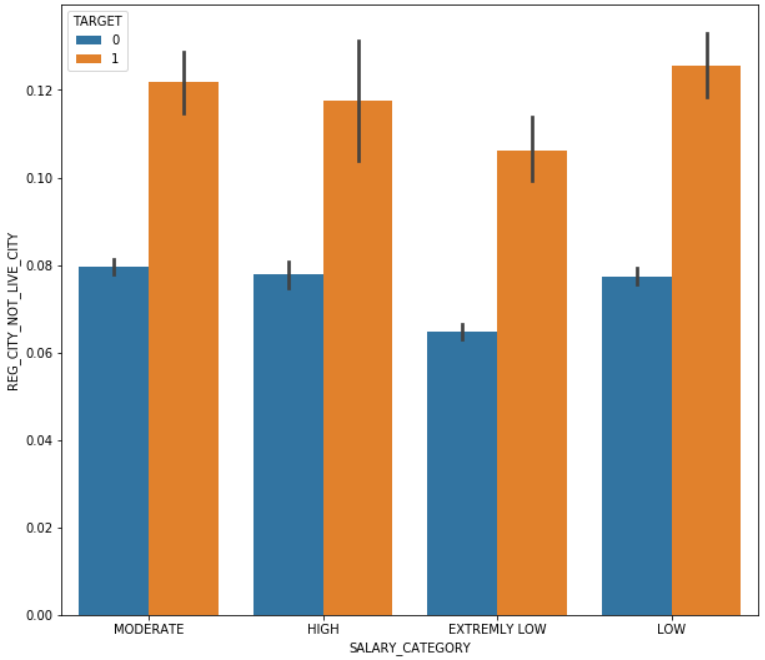
When Client gets Extremely lower salary and if his/her Contact address doest match, then there is a Higher chance for him/her to be defaulter

Salary vs Client whose Permanent Address not match with Work Address - Region Level



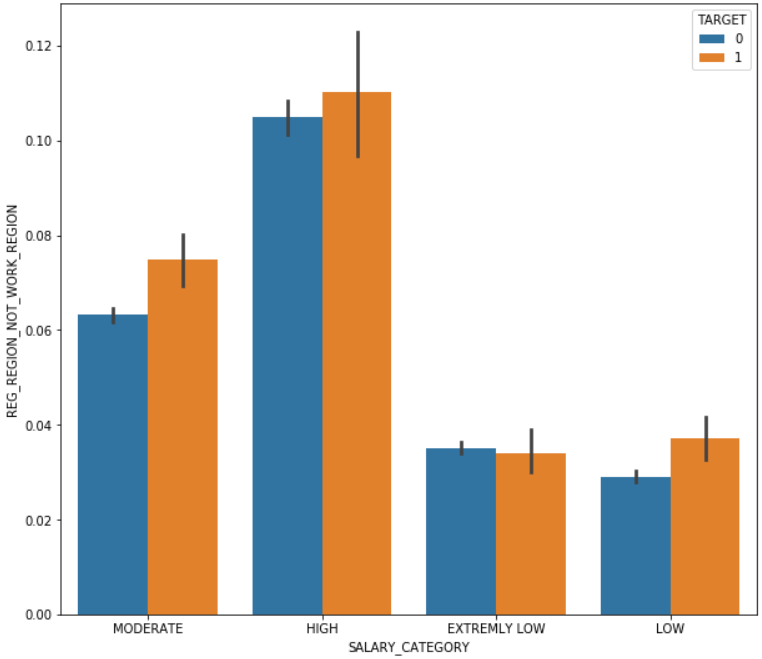
When Client gets Extremely lower salary and if his/her Work address doesn't match, then there is a Higher chance for him/her to be defaulter.

Salary vs Client whose Permanent Address not match with Contact Address -City Level



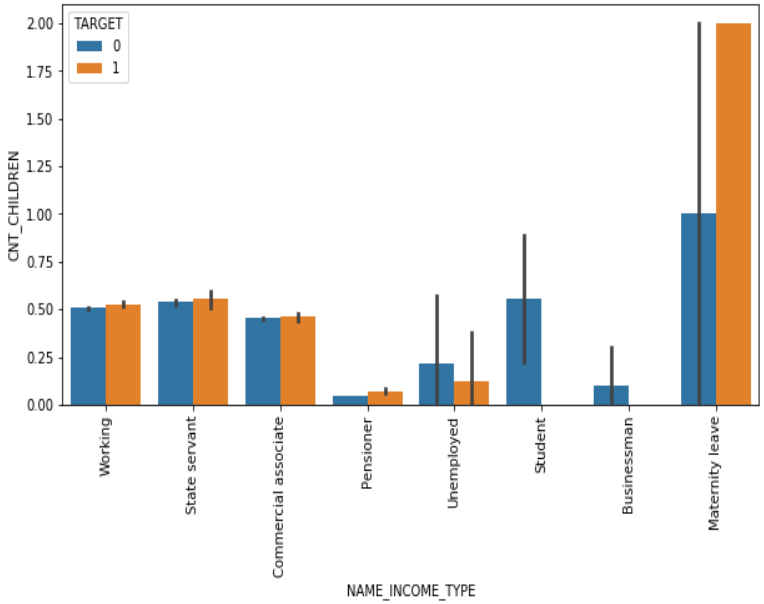
When Client gets LOWER salary and if his/her CONTACT address(CITY-LEVEL) doesn't match, then there is a Higher chance for him/her to be defaulter.

Salary vs Client whose Permanent Address not match with Work Address -City Level



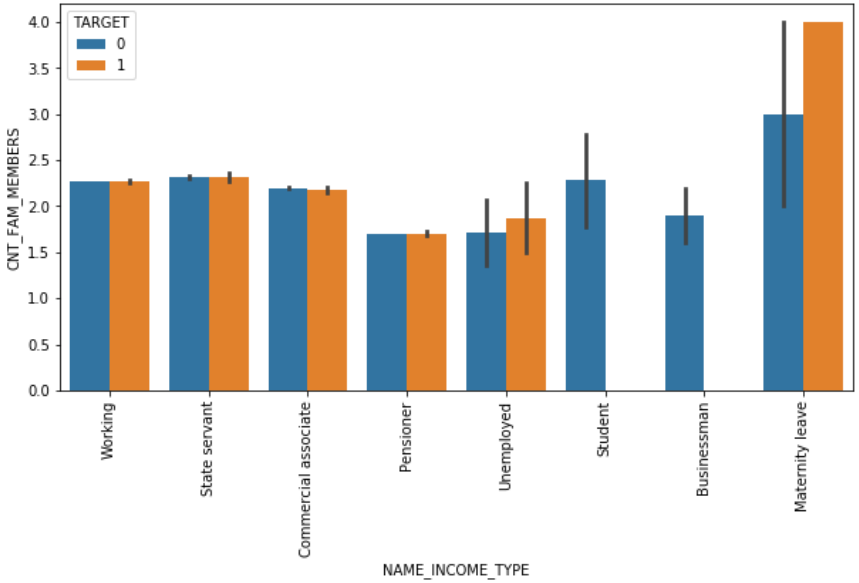
When Client gets HIGH salary and if his/her WORK address(CITY-LEVEL)doesn’t match, then there is a Higher chance for him/her to be defaulter.

INCOME vs CHILDREN Count



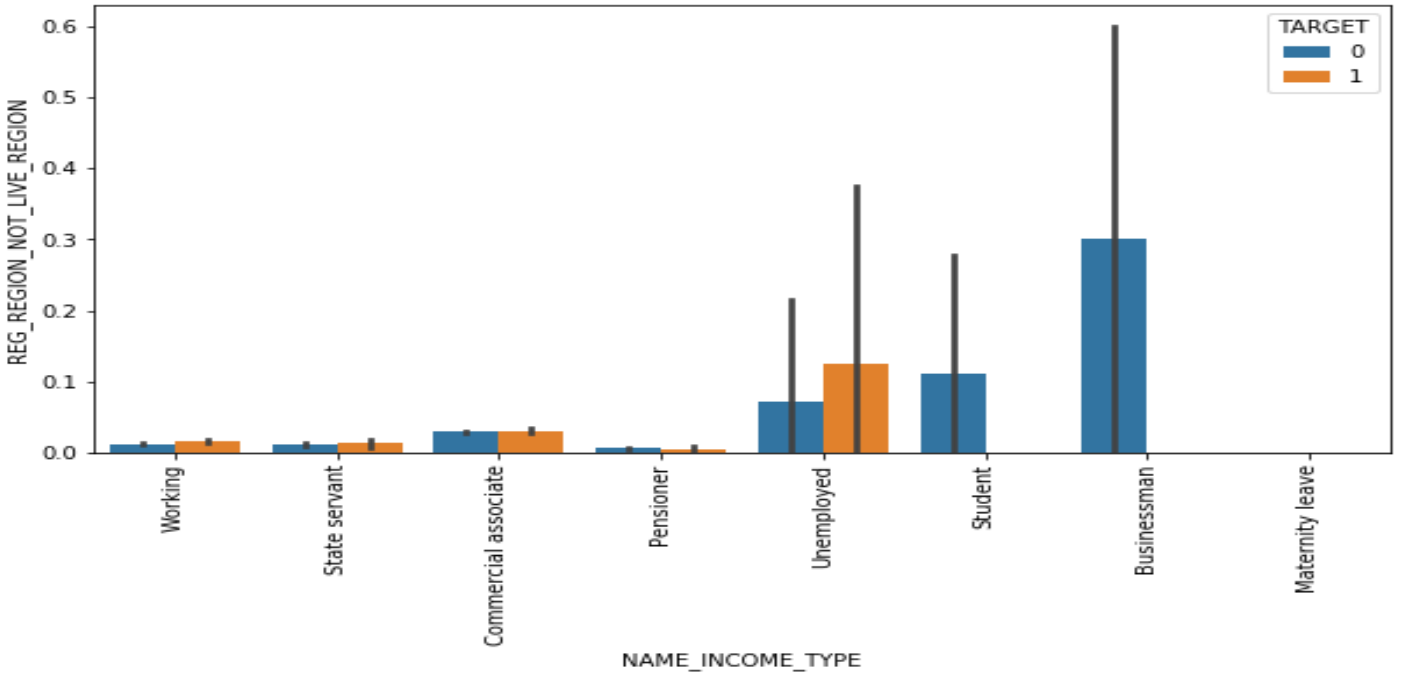
People who getting income via Maternity Leave tends to be more Defaulter when they have more children.

Income vs No.of.FamilyMembers



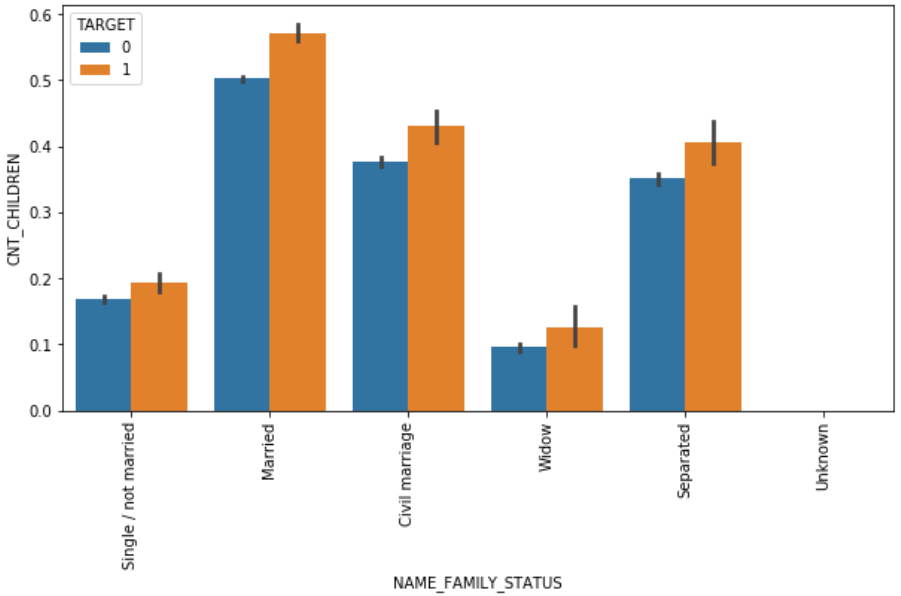
People who getting income via Maternity Leave tends to be more Defaulter when they have more Family Members.

Income Type vs Client whose Permanent Address not match with Contact Address -Region Level



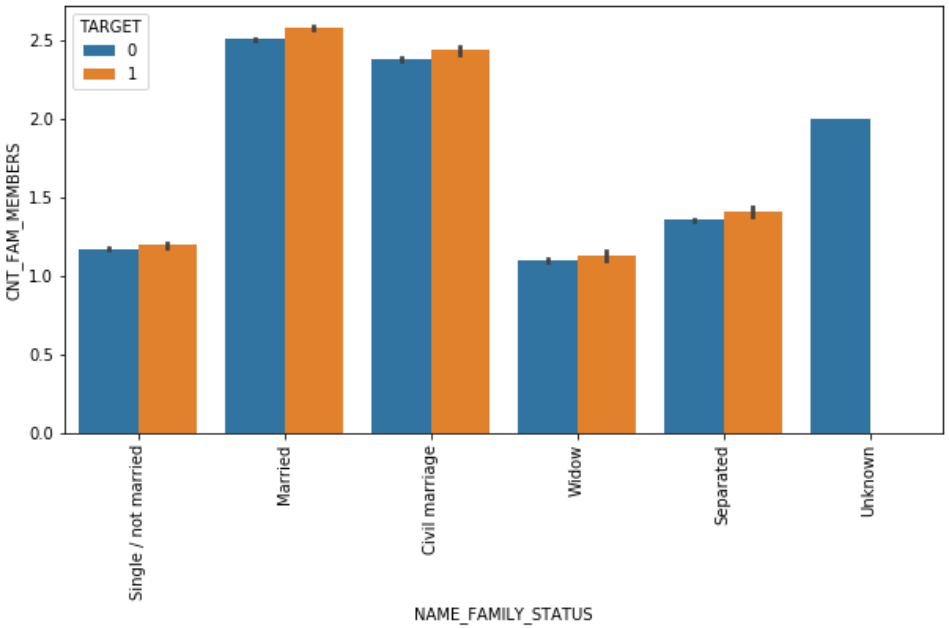
Client who are Unemployed has more chance to be a defaulter , when their Permanent Address does not match with the Contact Address in the Regional Level

Family Status vs Count Of Children



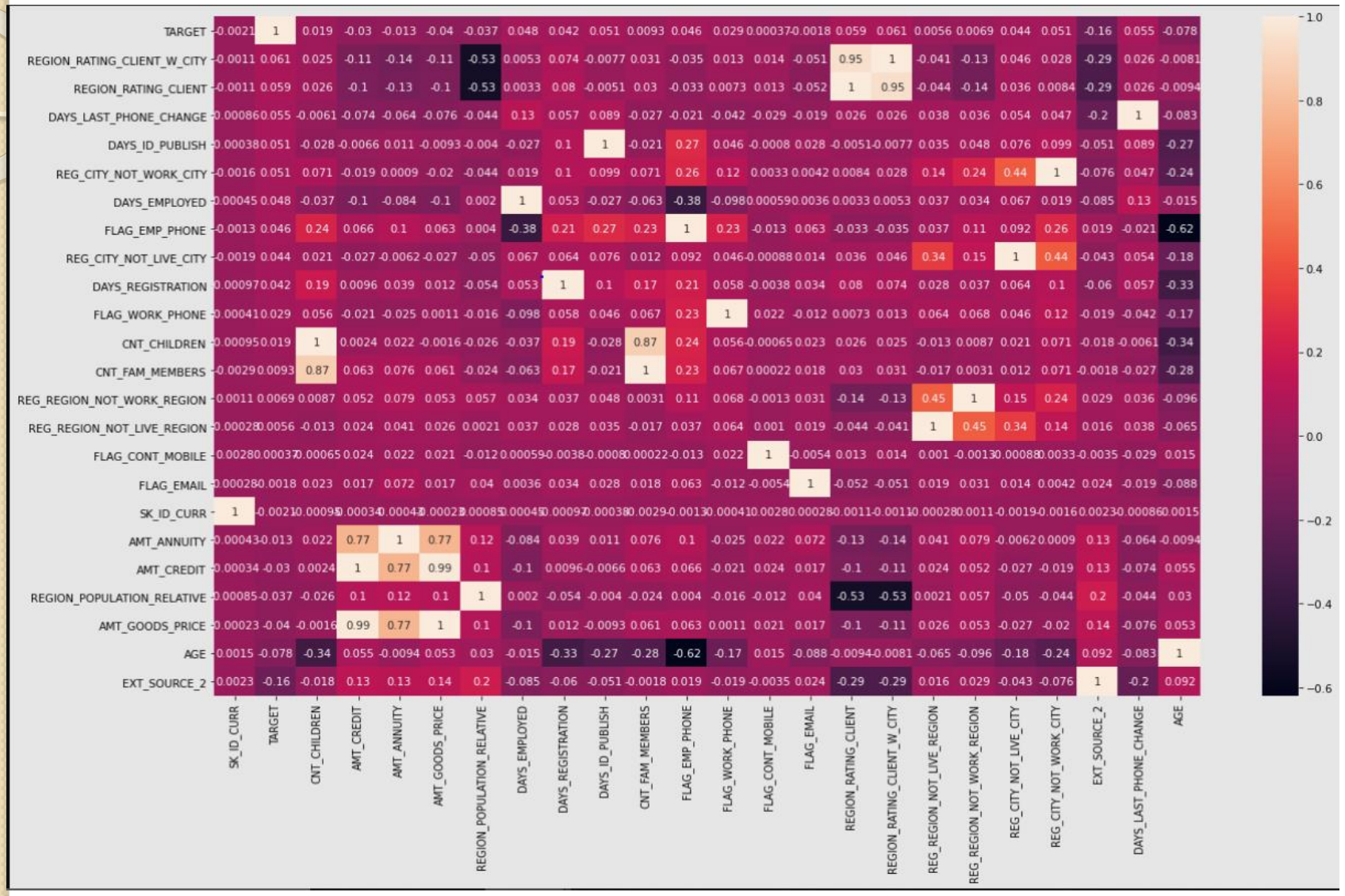
Client who are married and has more children (5+), chances to be a defaulter in High. This may be due to the Economic situation of their family, because of more children.

Family Status vs Count Of Family Members



Client who are married and has more children (5+), chances to be a defaulter in High. This may be due to the Economic situation of their family, because of more children

ANALYSING CORRELATION OF TARGET VARIABLE VS OTHER VARIABLES

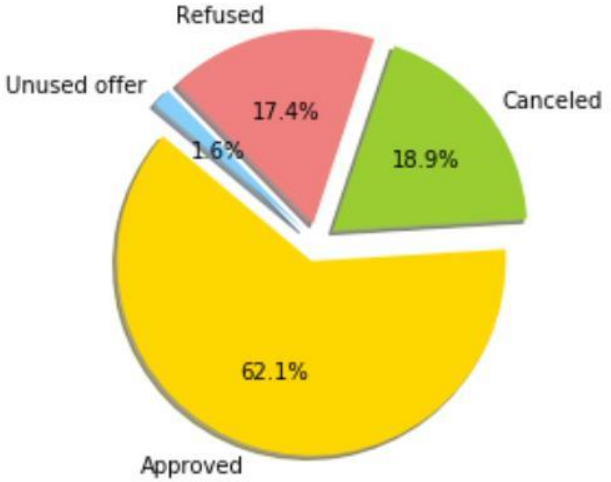


Highly Correlated Variables

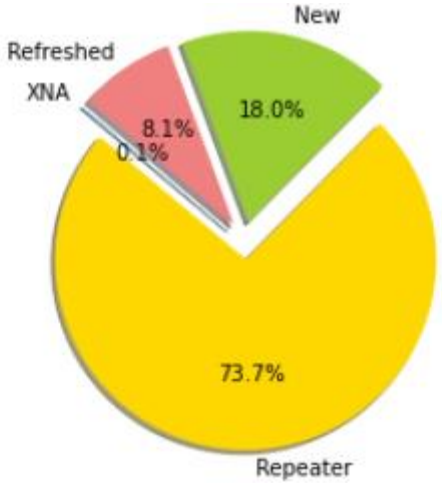
1.AMT_CREDIT and AMT_GOODS_PRICE =0.99
2.REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT = 0.95
3.CNT_FAM_MEMBERS and CNT_CHILDREN = 0.87
4.AMT_ANNUITY and AMT_CREDIT = 0.77

PREVIOUS APPLICATION ANALYSIS

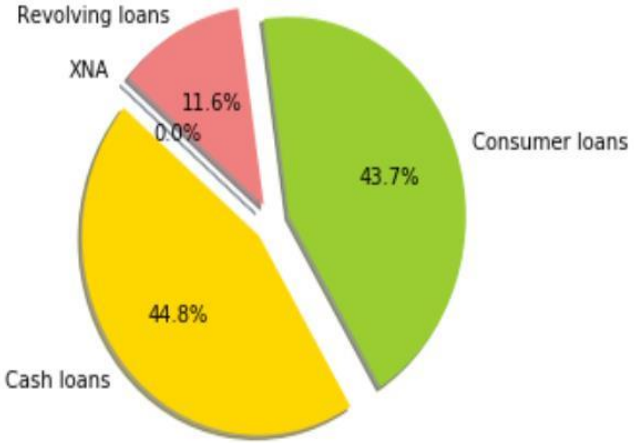
Then we moved on to analysis of the second data set. We performed few data cleaning steps and then moved on to analyzing the data.



Client Type

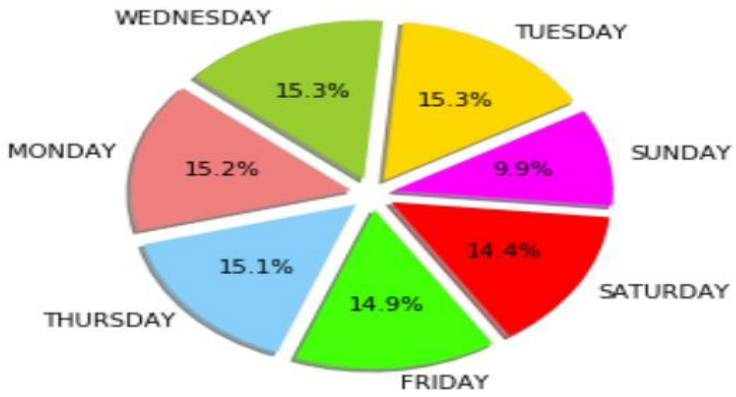


Based on Contract Type



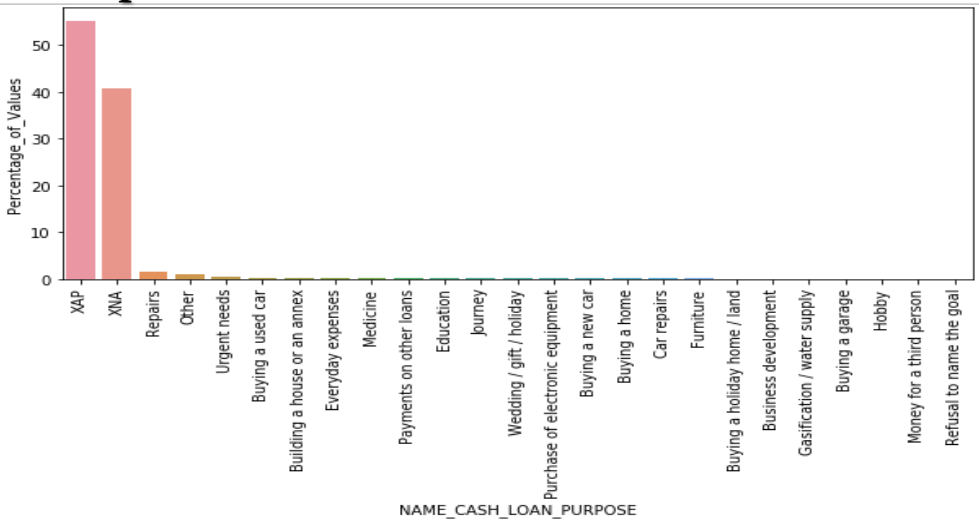
73.4% applicants are repeaters. Only, 18.4% are new clients.

Based on Days of Approval



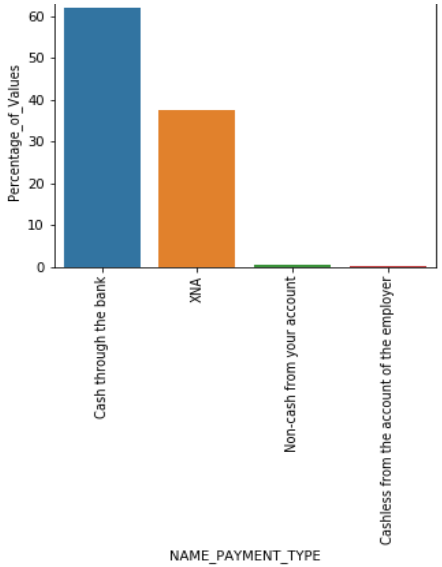
Most of the clients have opted to apply loan on Tuesday. It is very interesting to see that applicants are very low on weekends. We would otherwise assume that the applicants would prefer weekends to apply.

Based on Purpose of Loan



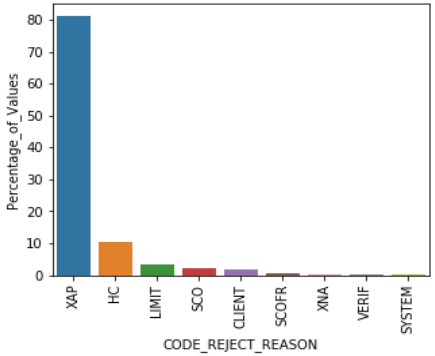
Most Loan purpose was not recorded. **XAP and XNA** values are highest.

Based on Payment Type



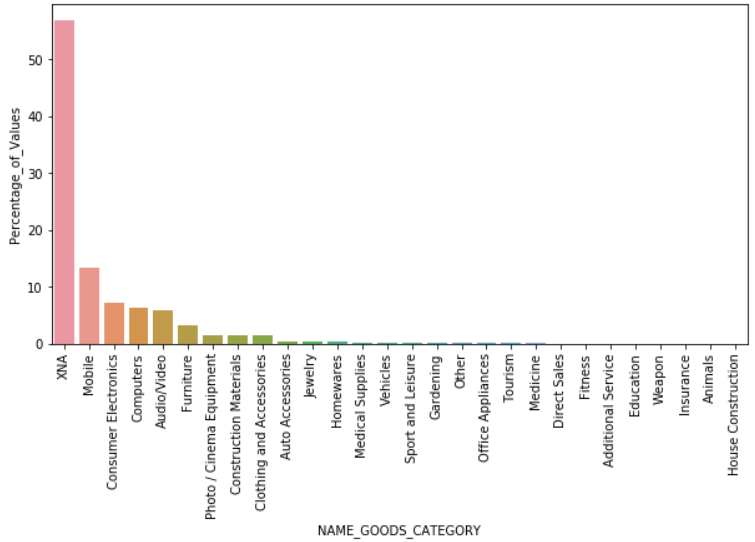
Most people preferred **CASH(62.44%)** as the mode of Payment

Based on Reason of rejection of loan



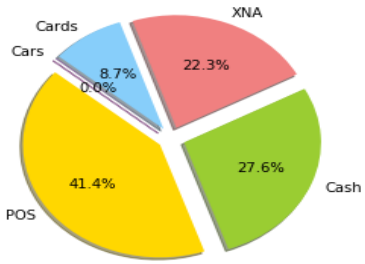
Primary reason for the Loan to get rejected is not recorded(**XAP (81%)**) followed by **HC**.

Based on What kind of goods did the client apply for in the previous application - NAME_GOODS_CATEGORY



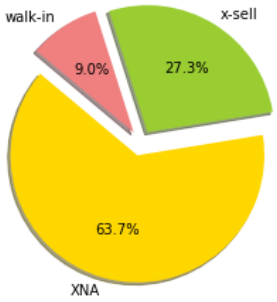
Most clients applied for Mobile and 53.96% of the data is not recorded(XNA).

Based on was the previous application for CASH, POS, CAR, …**NAME_PORTFOLIO**



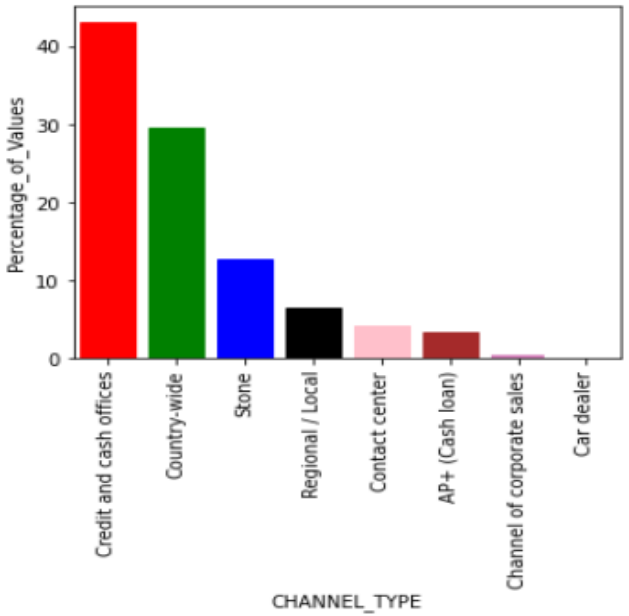
41.4% of the applications were for POS.

Based on Was the previous application x-sell or walk-in - NAME_PRODUCT_TYPE

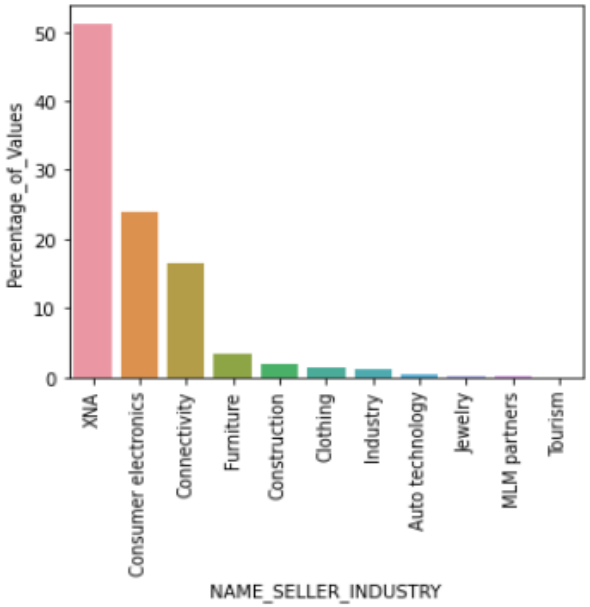


X-sell applications were more than walk-in

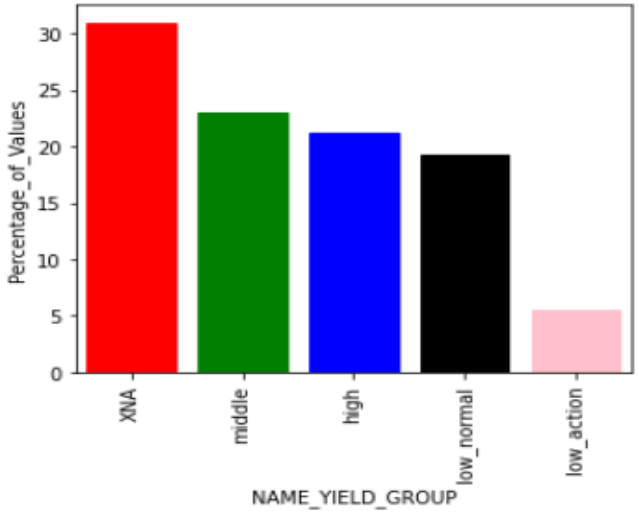
Based on Through which channel we acquired the client on the previous application - CHANNEL_TYPE



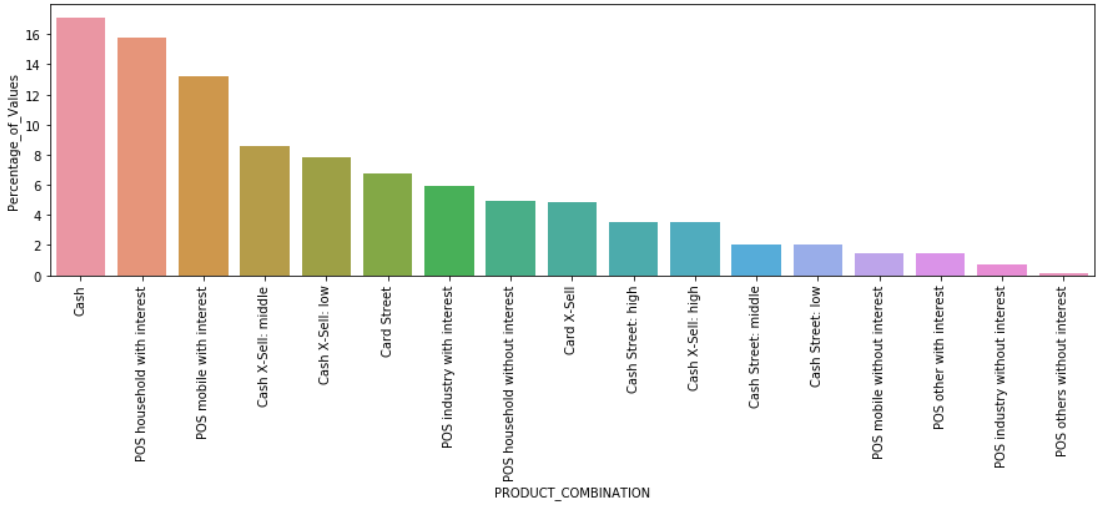
Based on The industry of the seller - NAME_SELLER_INDUSTRY



Based on Grouped interest rate into small medium and high of the previous application - NAME_YIELD_GROUP

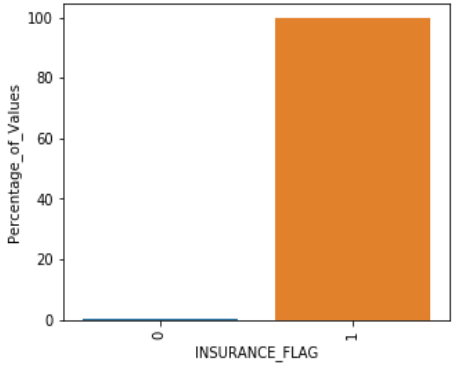


Based on PRODUCT_COMBINATION



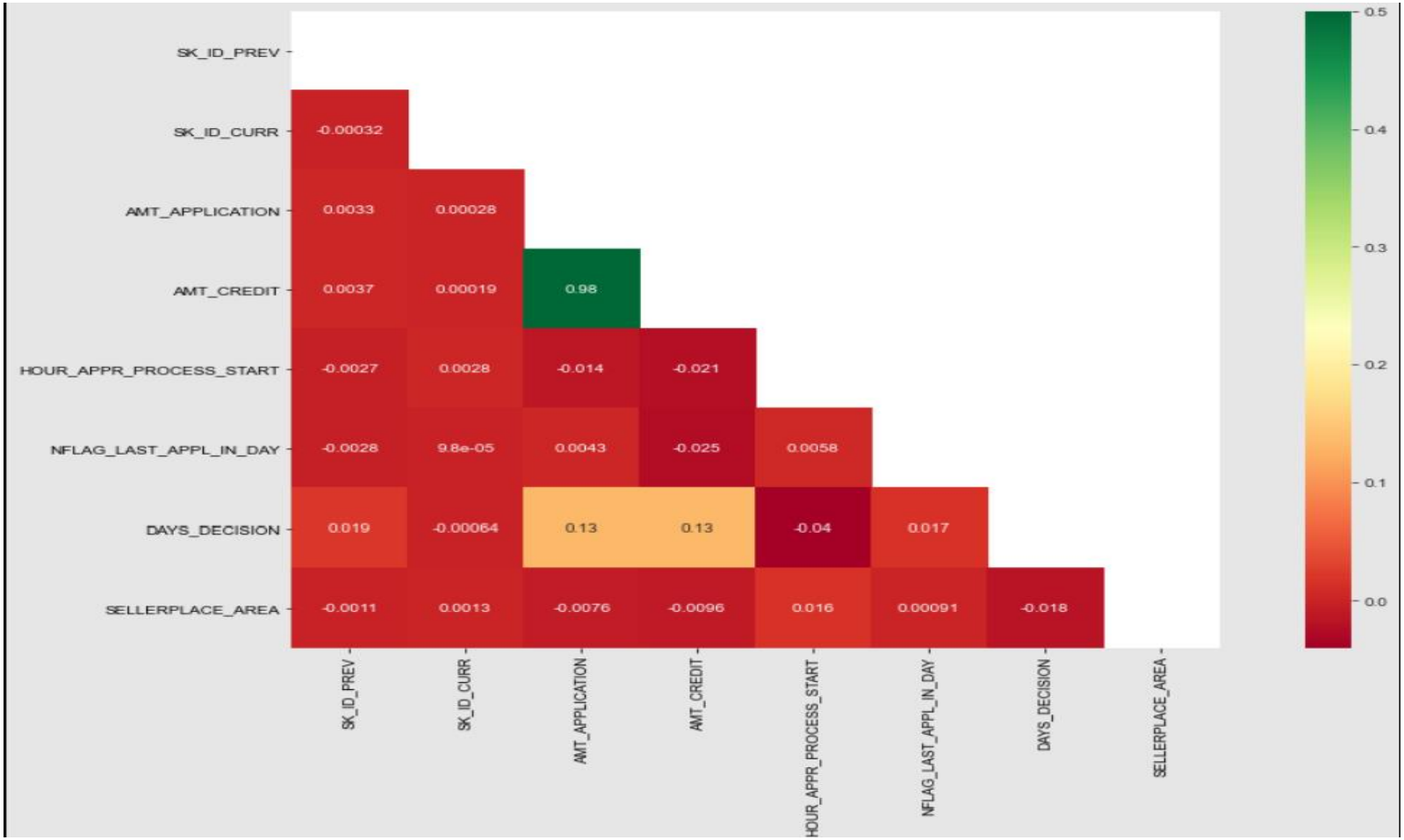
Highest product combination is **Cash** followed by **POS household with interest**

Based on Flag if the application was the last application per day of the client - NFLAG_LAST_APPL_IN_DAY

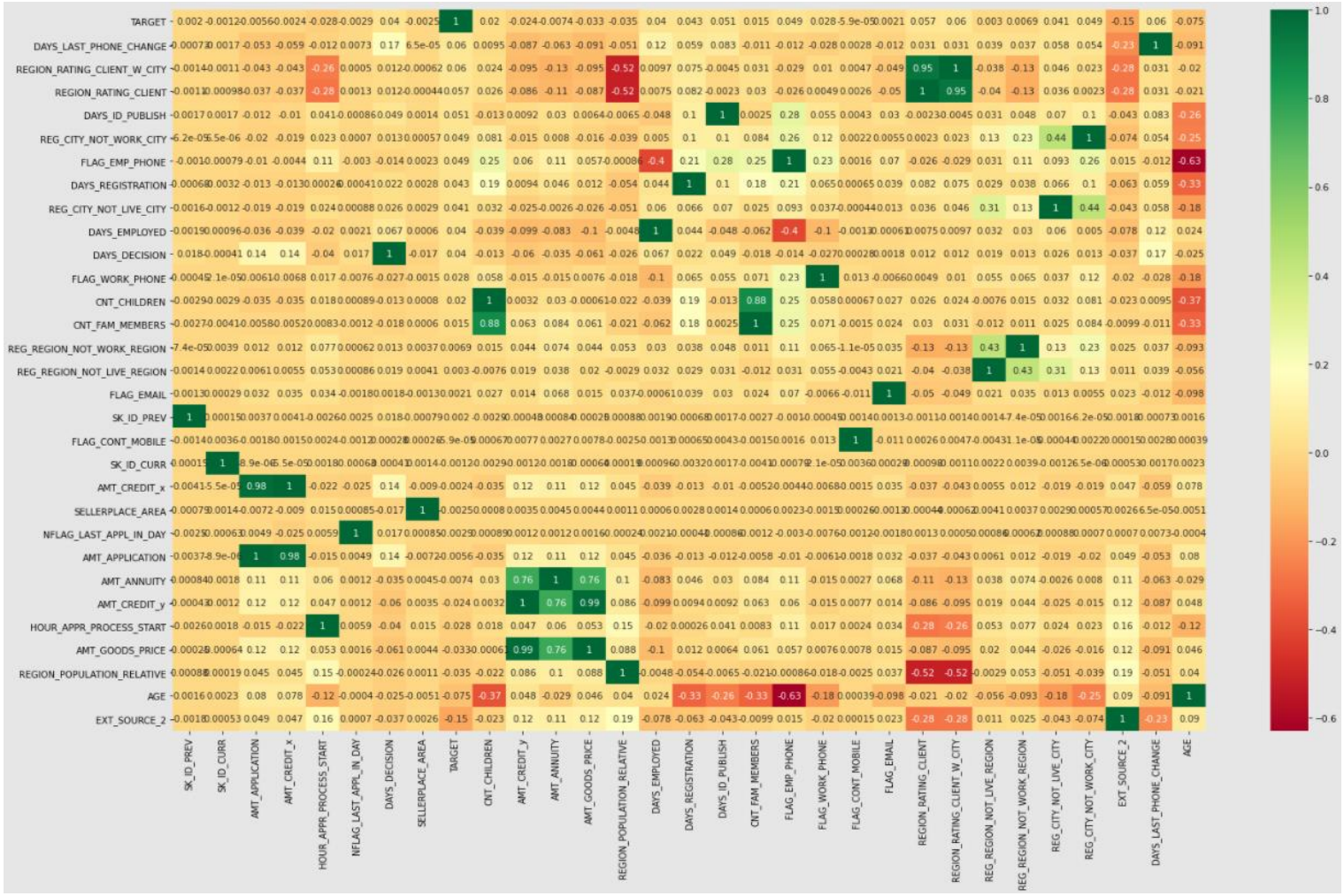


For most clients it was the last application of the day.

Correlation of variables in Previous Application



Correlation between previous_data and application_data dataframes



CONCLUSION

Less chance to be defaulter

- Old females
- Old people of any income group
- Client with high income category
- Widow who has unused previous loan status
- Clients who are working as a state servant
- Client with higher education(Female)
- Any client who's previous loan was approved
- Refreshed client who has unused loan status previously

More chance to be defaulter

- Male clients with civil marriage
- Previously Refused loan
- Secondary education clients are most likely to be defaulted when their previous loans were cancelled or refused

THANK YOU