

BANK DATA ANALYSIS

Swetha sasidharan

Business Understanding and overview

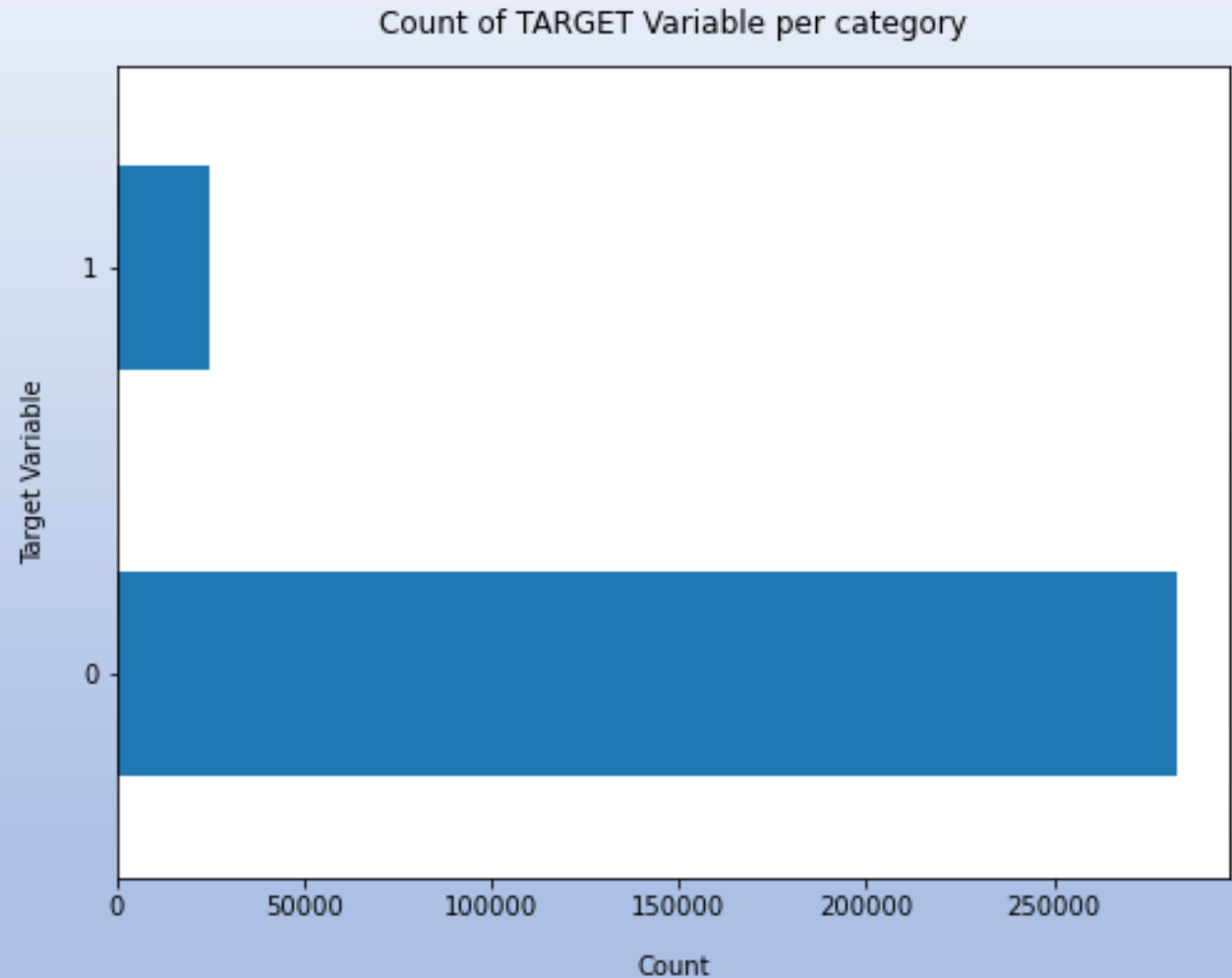
The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

UNDERSTANDING THE DATA

Target variable

Finding

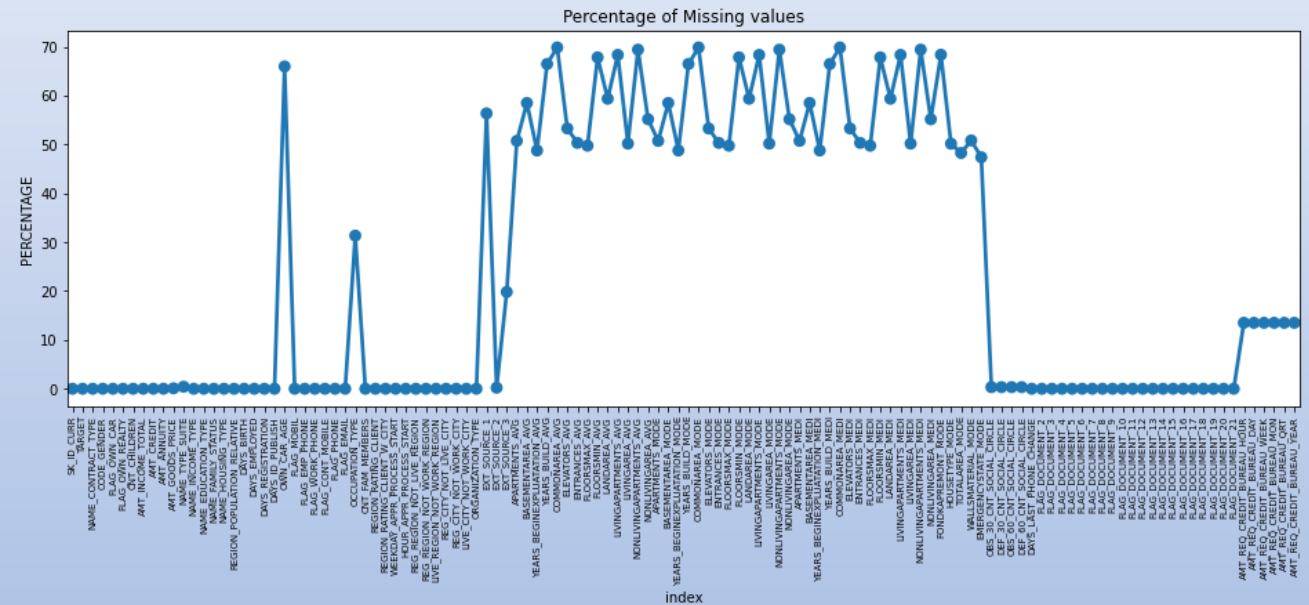
- Data is highly imbalanced, ratio = 92:8
- Most people are repaying the loan.
- 8% people are defaulters.
- So we analyze the data with other features while taking the target values separately to get some insights.



Missing data

Finding

- 64 columns(30%-70%)have null values. Other columns have no null value.
- 49 cloumns have 40% above null value.
- For analysis we drop high number of null value columns.so we dropped the 49 column that have high number of null value.



Missing Value Treatement

- We can see that 'AMT_ANNUITY', 'AMT_GOODS_PRICE' and 'EXT_SOURCE_2' columns having very less percentage of null values and is also a numeric data. We can impute median here.
- 'NAME_TYPE_SUITE' and 'OCCUPATION_TYPE' are categorical data .so we impute mode here.
- Some columns are not carrying any important information for our analysis ,so we deleting that columns(51 columns).
- CODE_GENDER contain 4 'XNA' values. We changed the XNA to the mode.
- ORGANIZATION_TYPE contain 55374 'XNA' value. So dropped the 'XNA' values .

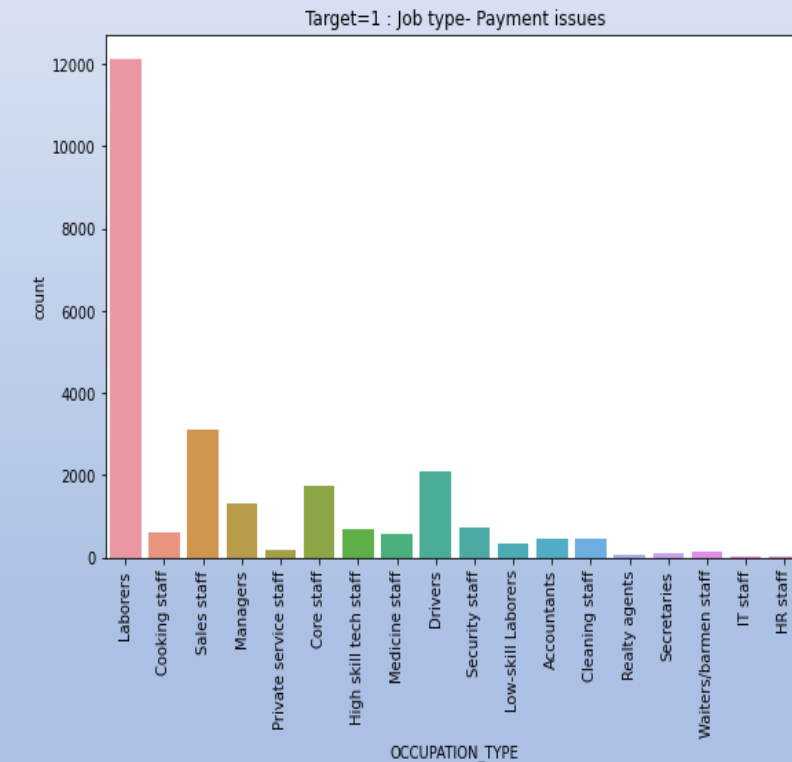
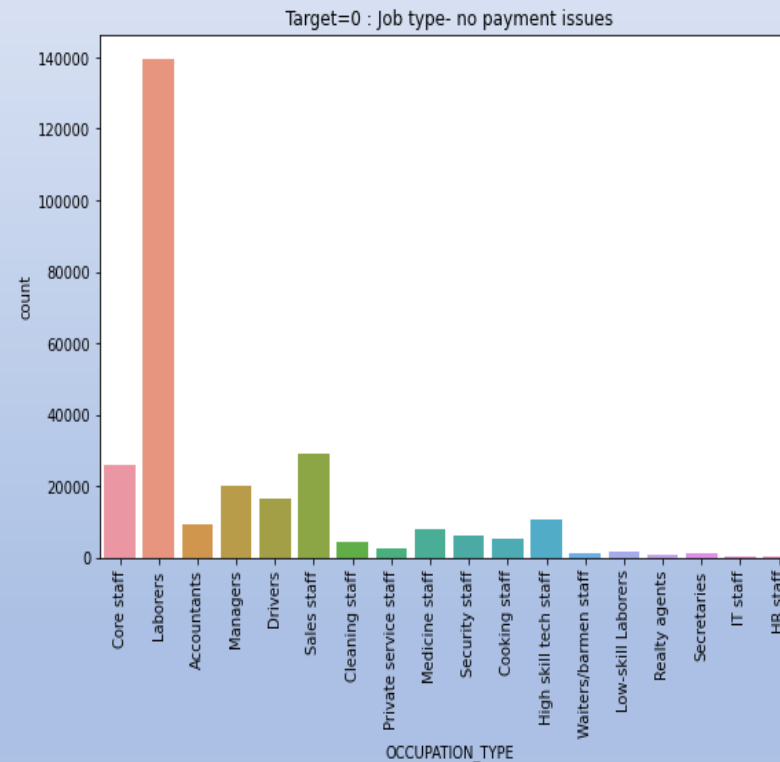
CATEGORICAL ANALYSIS

UNIVARIATE ANALYSIS

OCCUPATION_TYPE vs TARGET VARIABLE

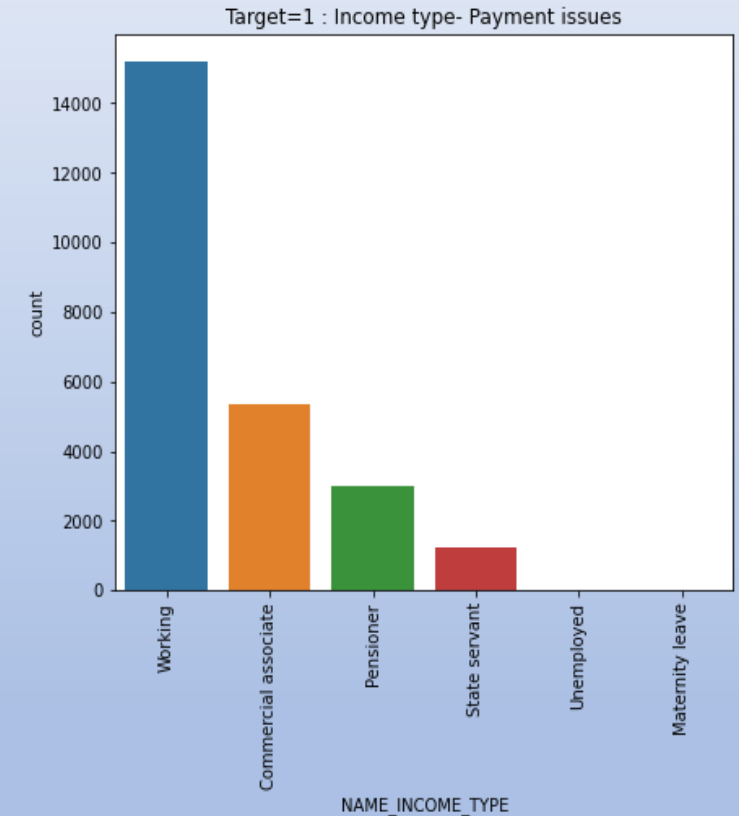
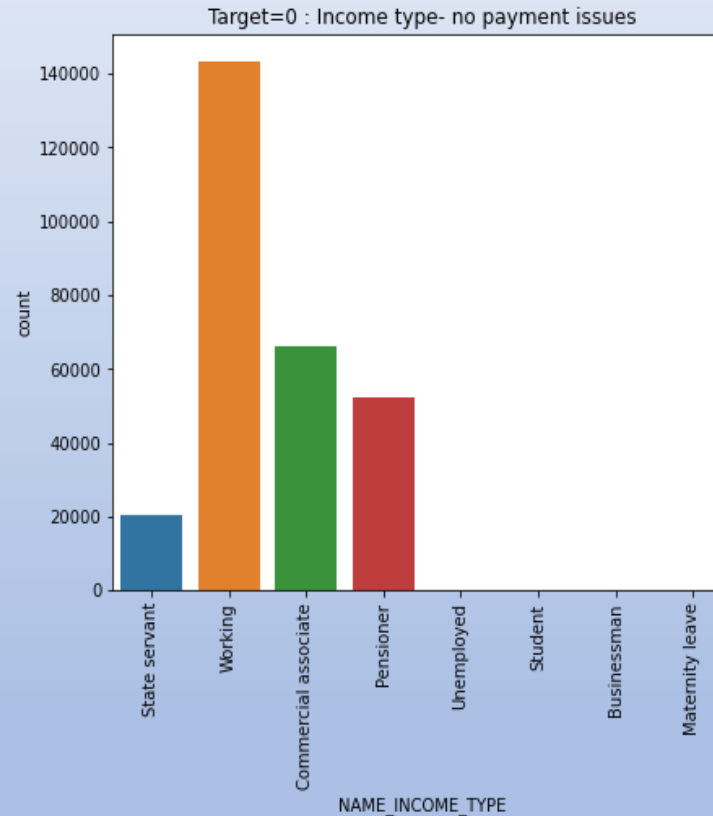
FINDINGS

- Labourers have high repayment percentage.
- Low skill labores ,Sales staff, core staff, drivers, especially people with less income range are the most likely to be loan defaulters than highly paid staffs and accountants.
- Better the occupation, lesser the chance of defaulting.



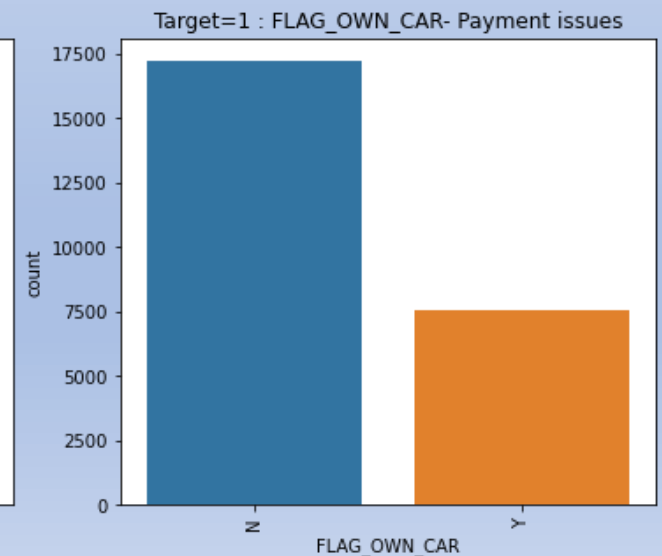
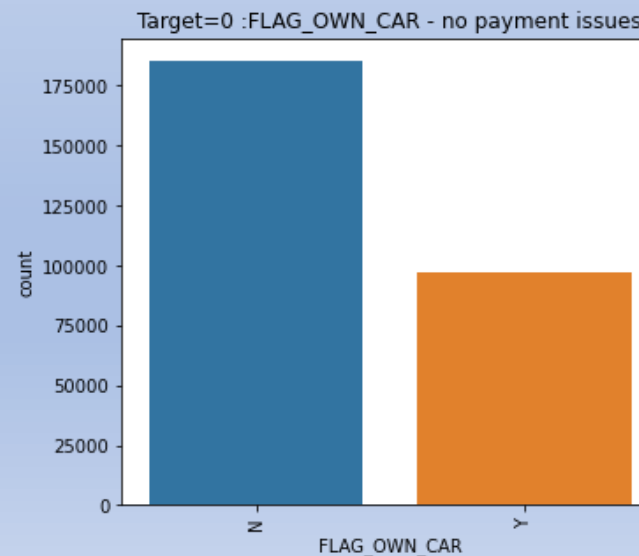
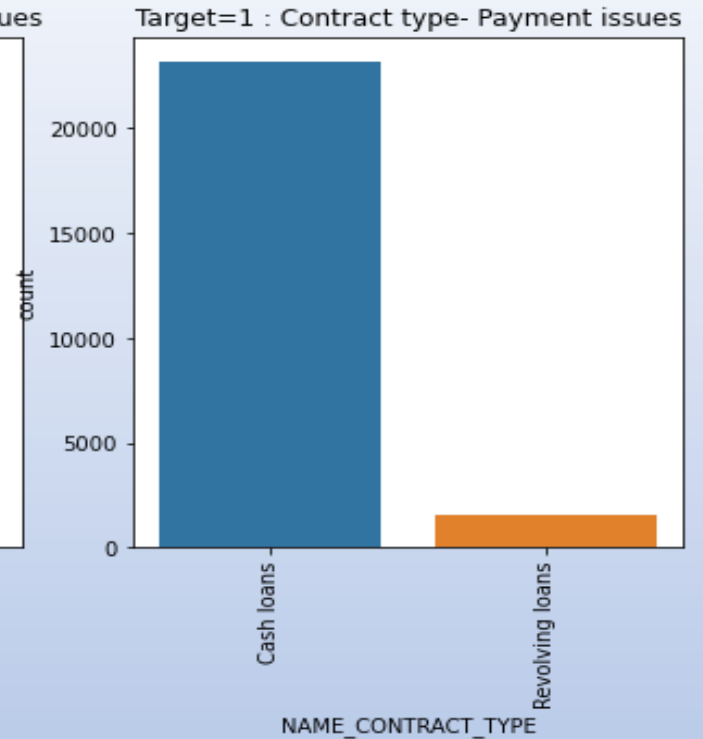
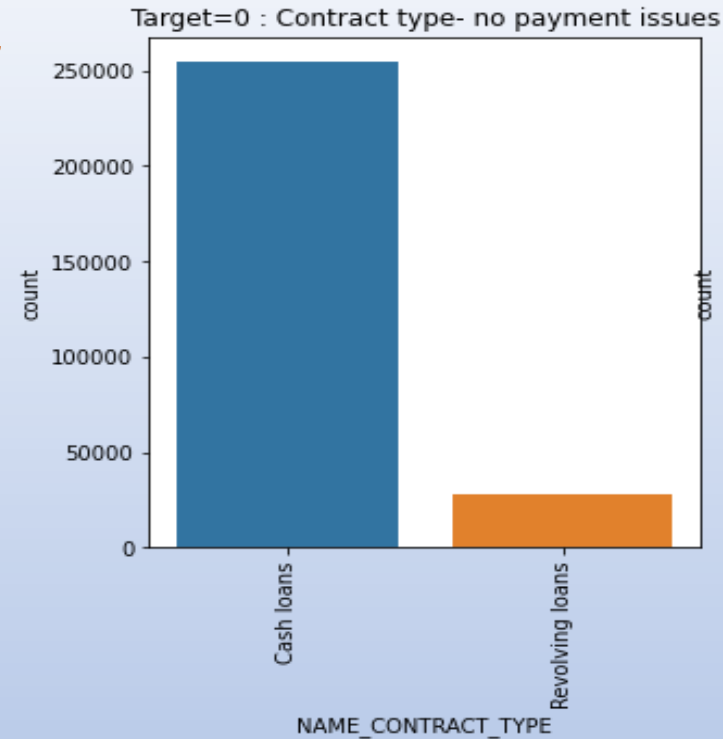
NAME_INCOME_TYPE VS TARGET VARIABLE

- Working customers can be targeted to lend loans as they have higher percentage of making payments on time.
- commercial associates, state servants and pensioners are fairly more reliable.



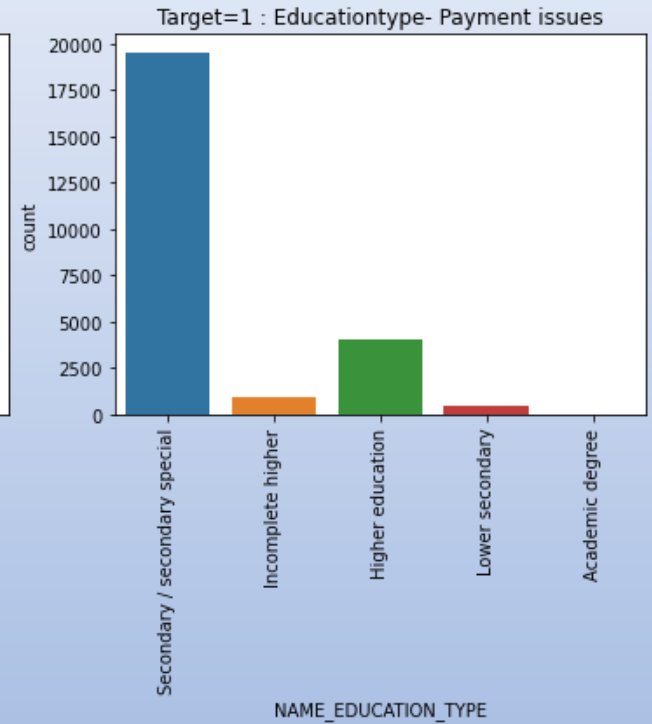
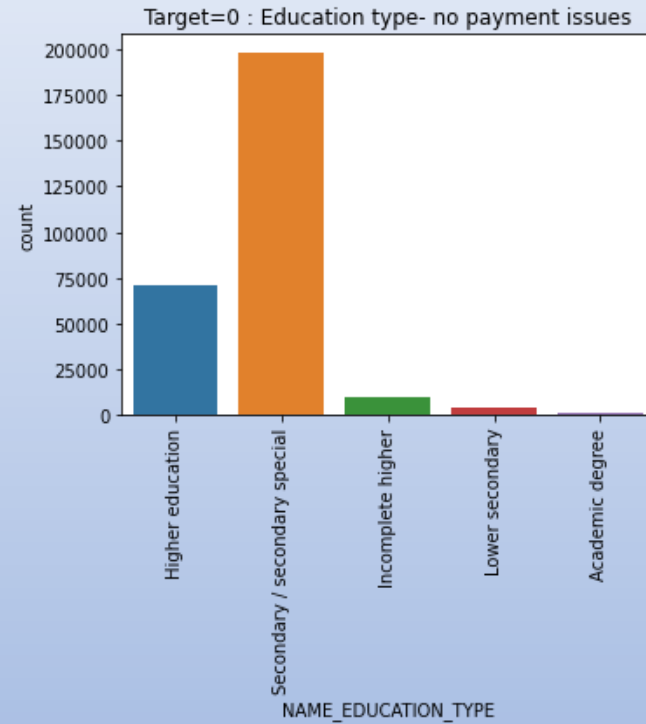
NAME_CONTRACT_TYPE , FLAG_OWN_CAR vs TARGET VARIABLE

- Customers having own cars less likely to not repay the loan.
- The most cash loan are not repayed as compare to revolving loans



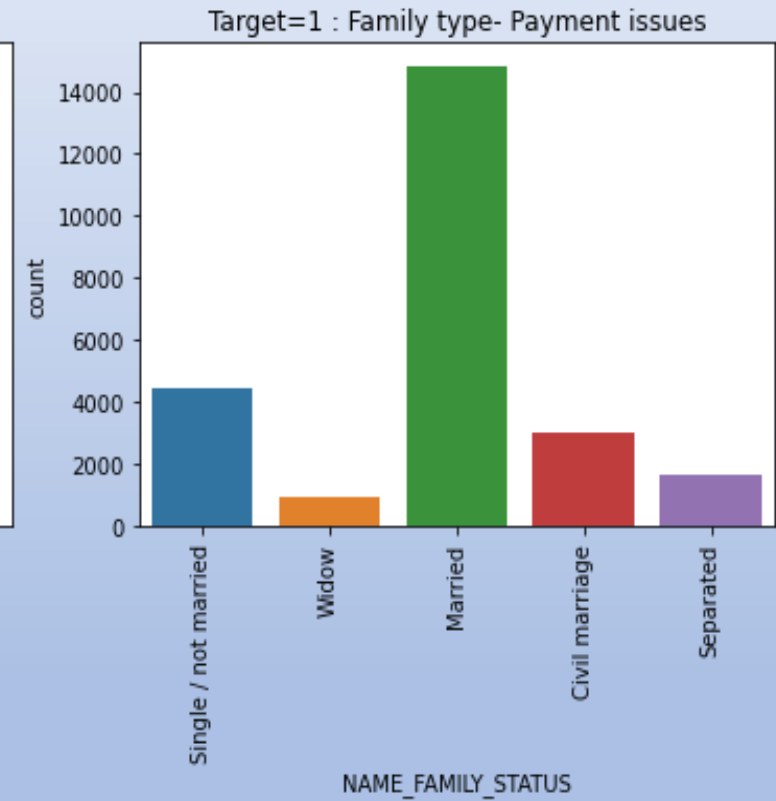
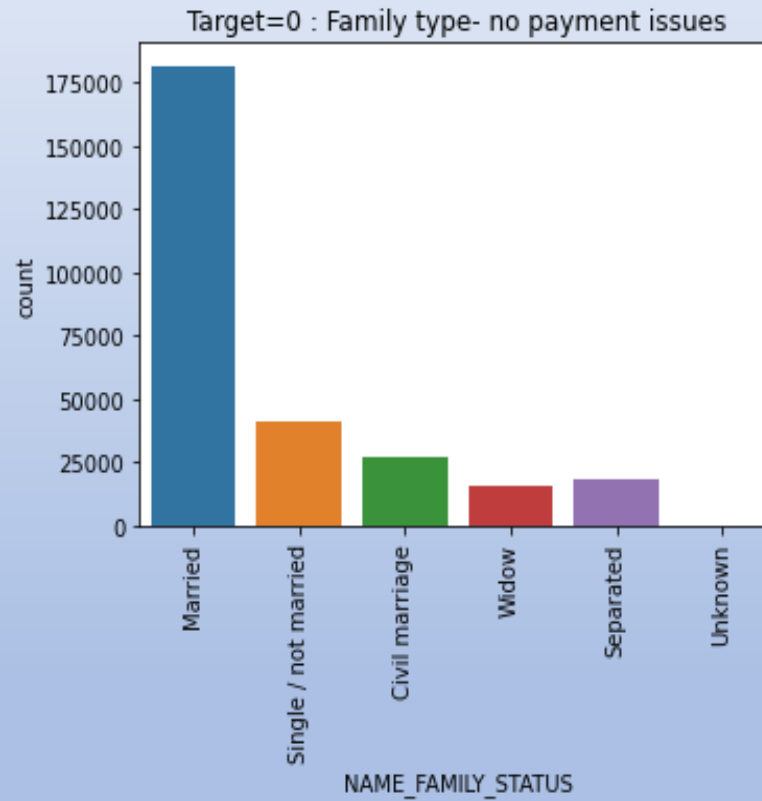
NAME_EDUCATION_TYPE VS TARGET VARIABLES

- Customers with secondary education are most likely to make payments when compared to customers with higher education.
- Lower secondary education loan are more risky compare other education.



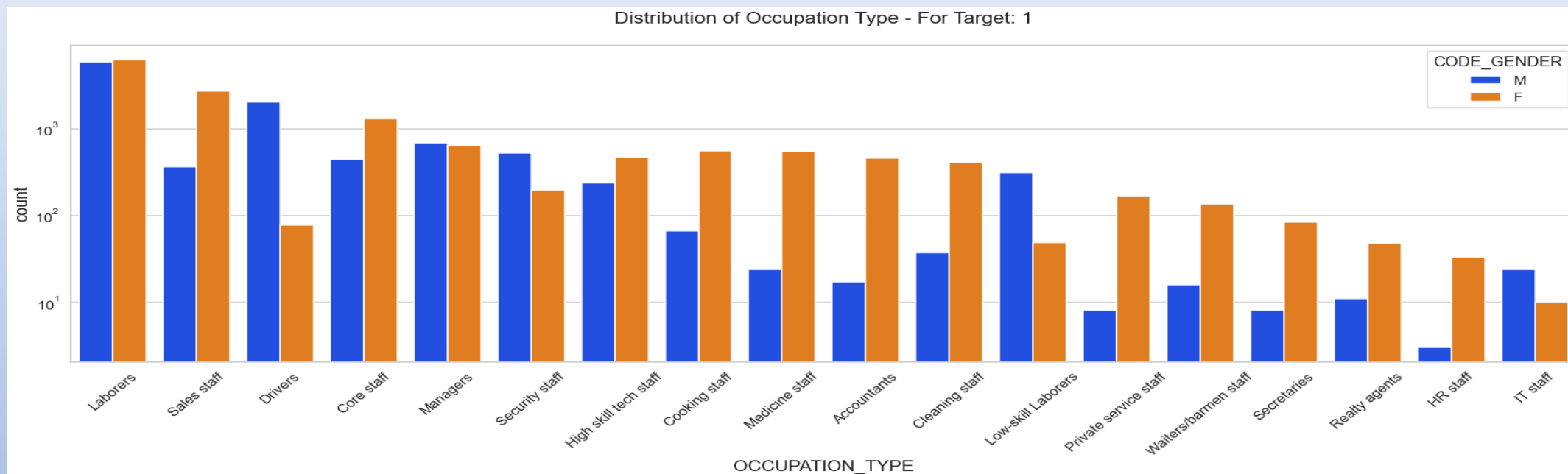
NAME_FAMILY_STATUS VS TARGET VARIABLES

- Most married customers are repaying there loans.
- Civil marraiges and single people have higher chances of defaulting than others.



CATEGORICAL ANALYSIS VS TARGET AND GENDER

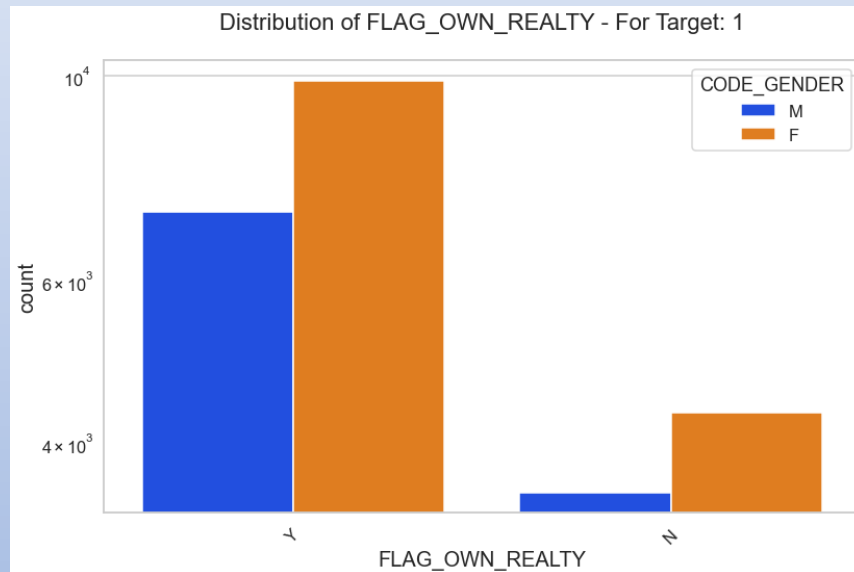
Occupation type vs defaulters(target 1)



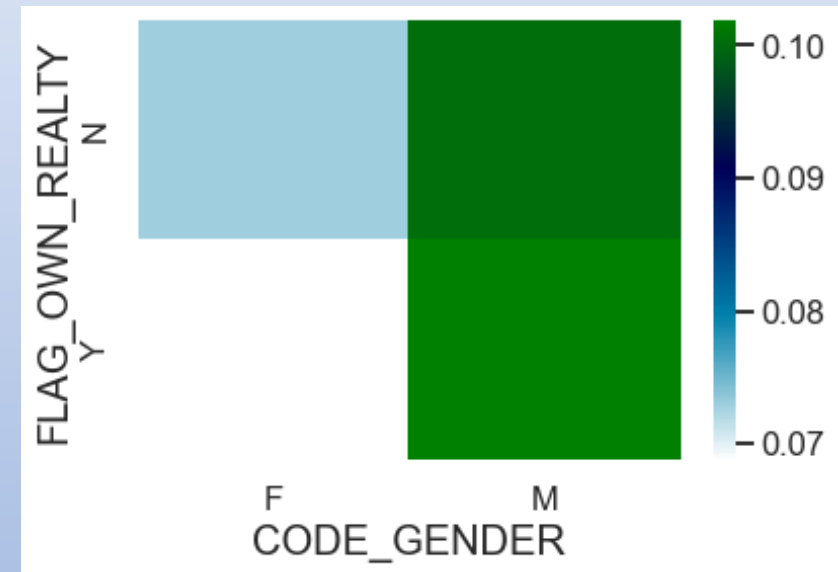
GOOD INSIGHT

- Female under occupation with Accountants, Private Service Staff, Medicine staff ,Secretaries, Realty Agents, HR Staff etc. Are the most defaulted sub categories against their male counter parts.
- Male under occupation with Security staff, Low Skilled Laborer's, Drivers etc. Are the most defaulted sub categories against their female counter parts.

FLAG_OWN_REALTY vs defaulters(target 1)



- Female clients they have own flat they are the most defaulters as compared to male clients

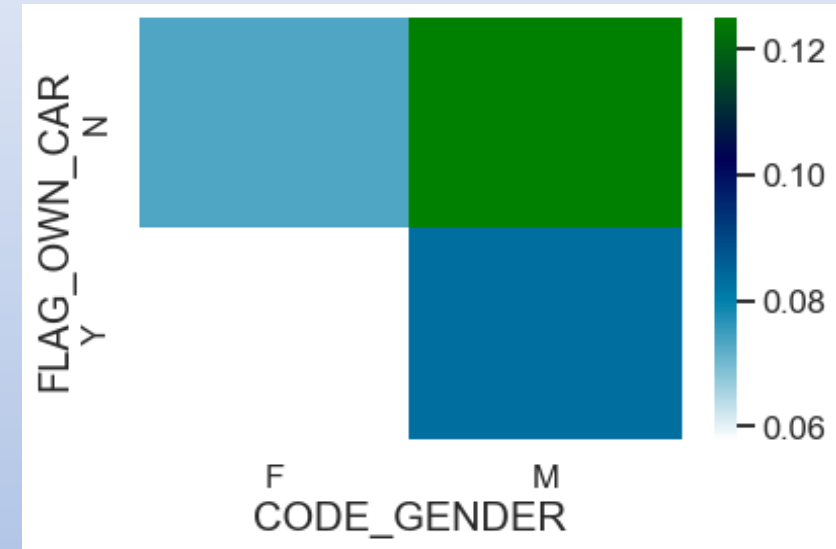


- Male clients they have own flat or no own flat they are highest default rate as compared to other 2 sub categories.

FLAG_OWN_CAR vs defaulters(target 1)

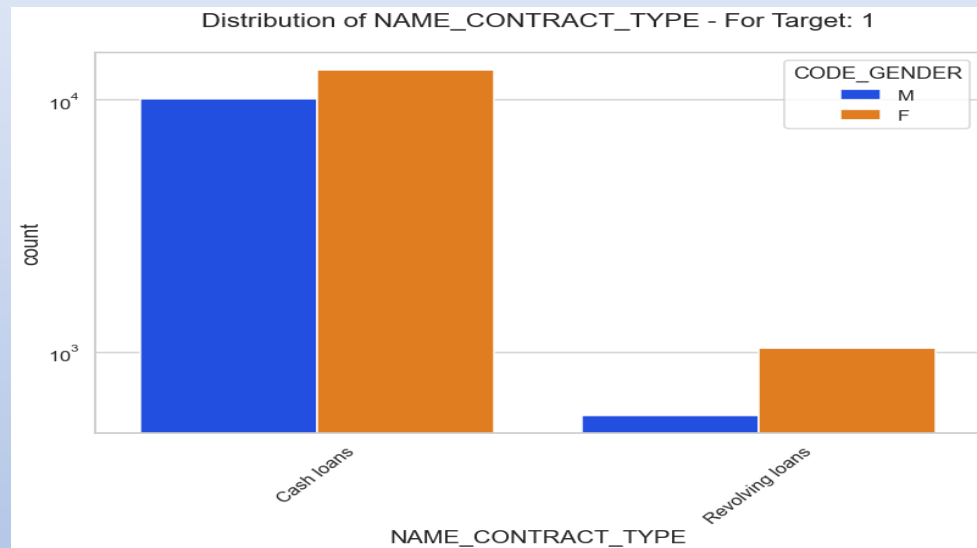


- Overall people without cars are most defaulters but female clients without car are the most defaulters as compared to the male clients without cars.

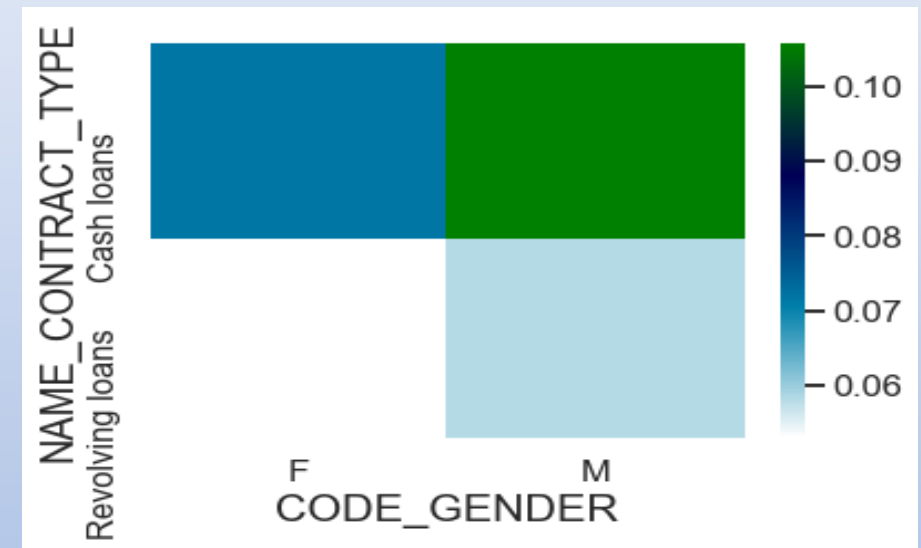


- Male client without have own car are the highest default rate as compared to other 3 sub categories.

NAME_CONTRACT_TYPE vs defaulters(target 1)

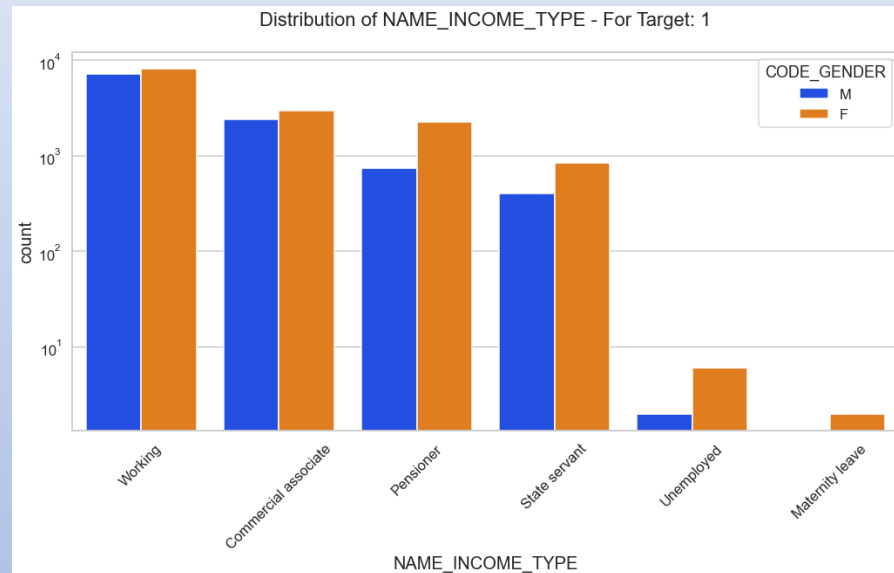


- Overall people took loan is cash they are most defaulters but female clients who took loan is cash are the most defaulters as compared to the male clients

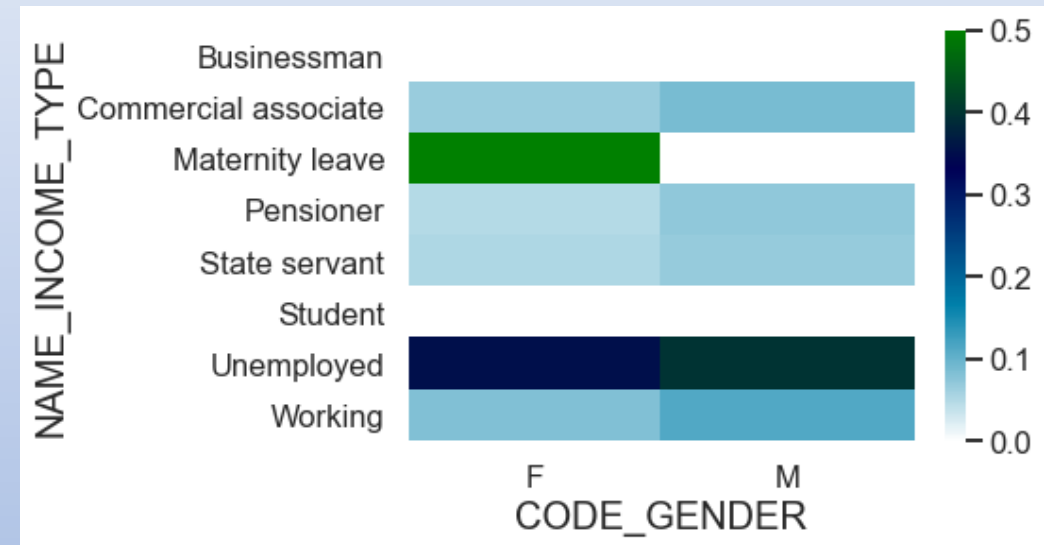


- Male clients took loan is cash they are the highest default rate as compared to other 3 sub categories.

NAME_INCOME_TYPE vs defaulters(target 1)

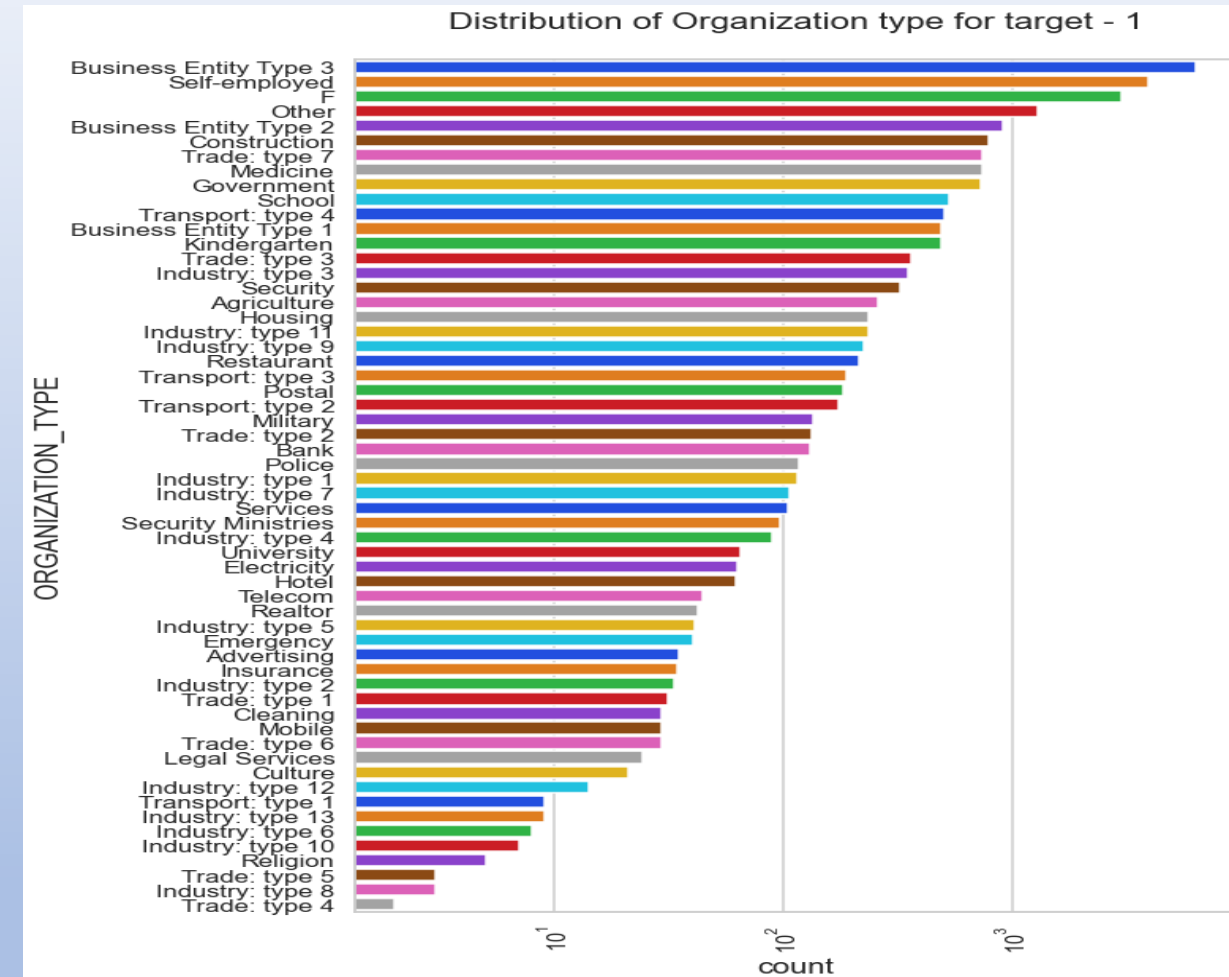
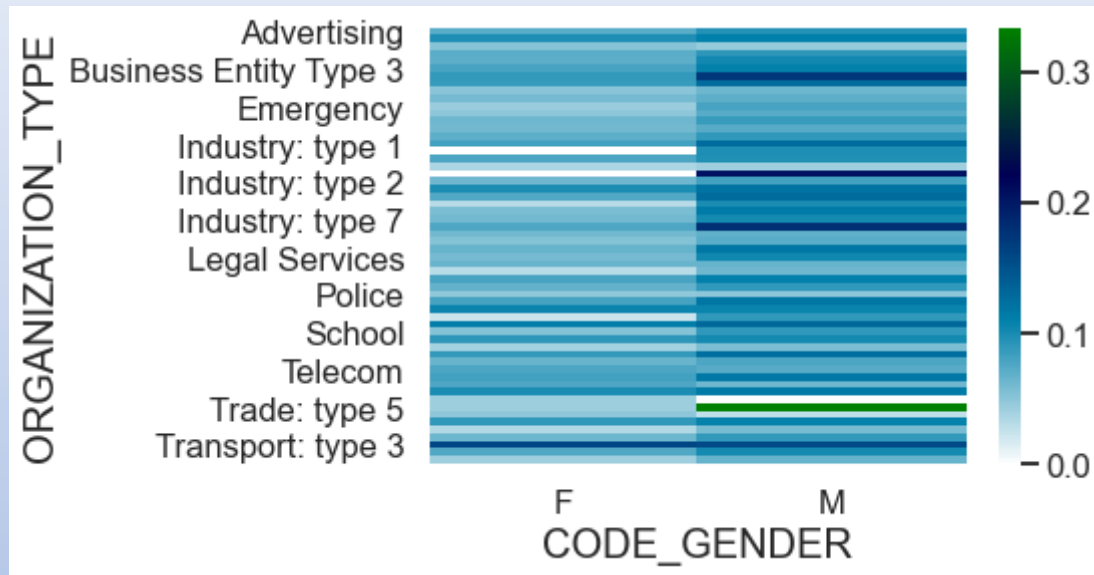


- Females have highest default rate as compared male.



- Female clients have maternity leave they have highest default rate as compared to other sub categories. Both unemployed male and female client's are highest defaulters.

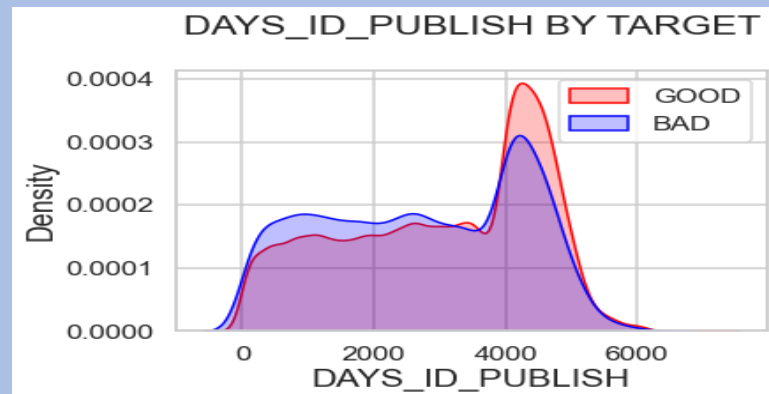
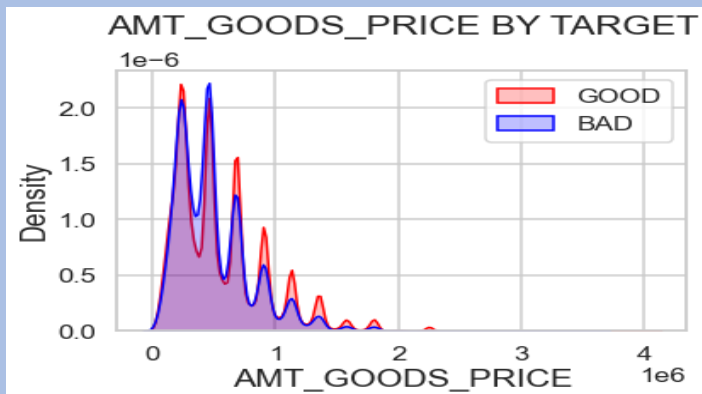
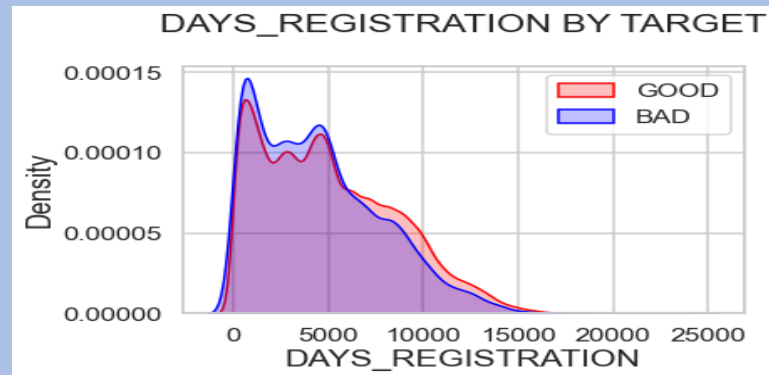
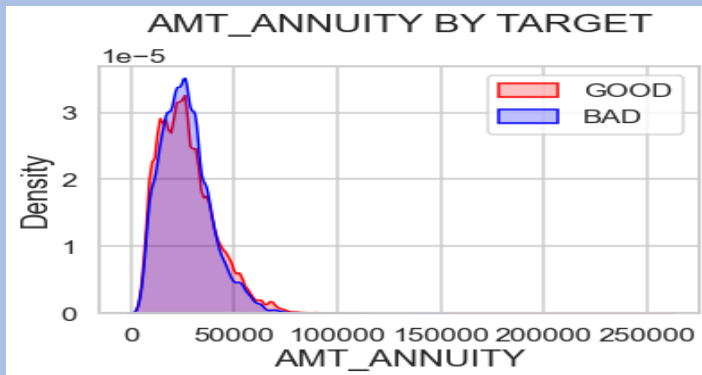
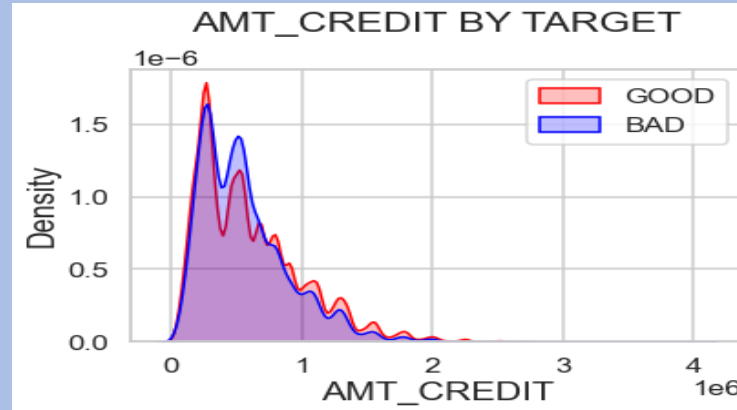
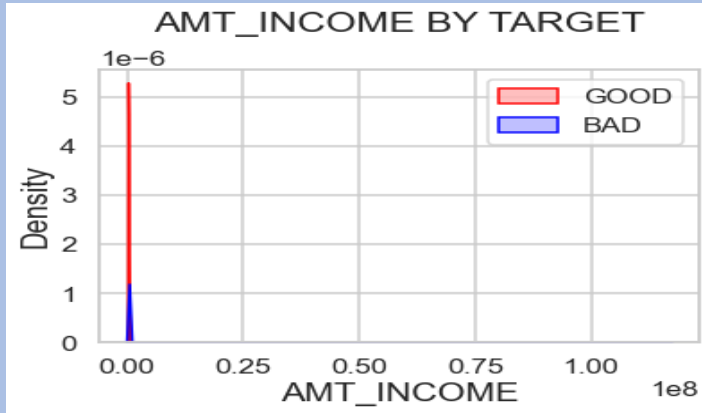
ORGANIZATION_TYPE vs defaulters (target1)



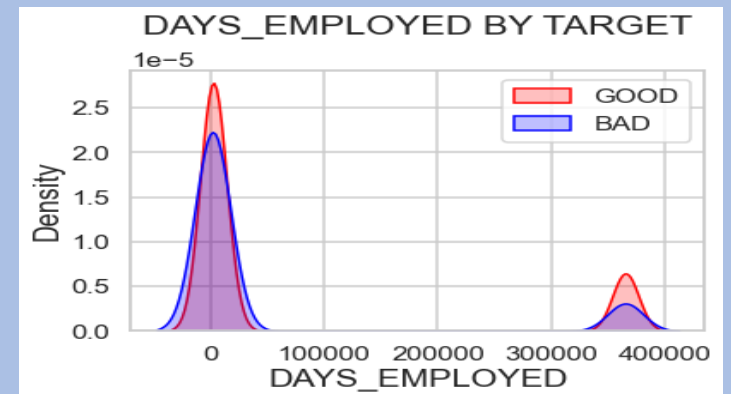
- Male in Trade type 5 are highest default rate as compared to other sub categories. Clients which have applied for credits are from most of the organization type Business entity Type 3 , Self employed , Other , Medicine, Business entity Type 2 and Government. Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.

NUMERICAL ANALYSIS

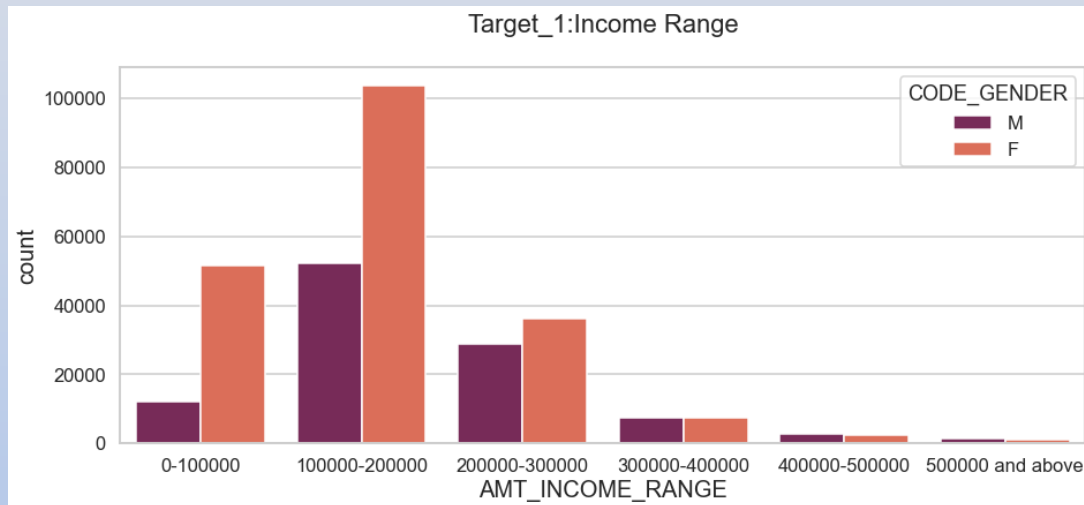
SEGMENTED ANALYSIS



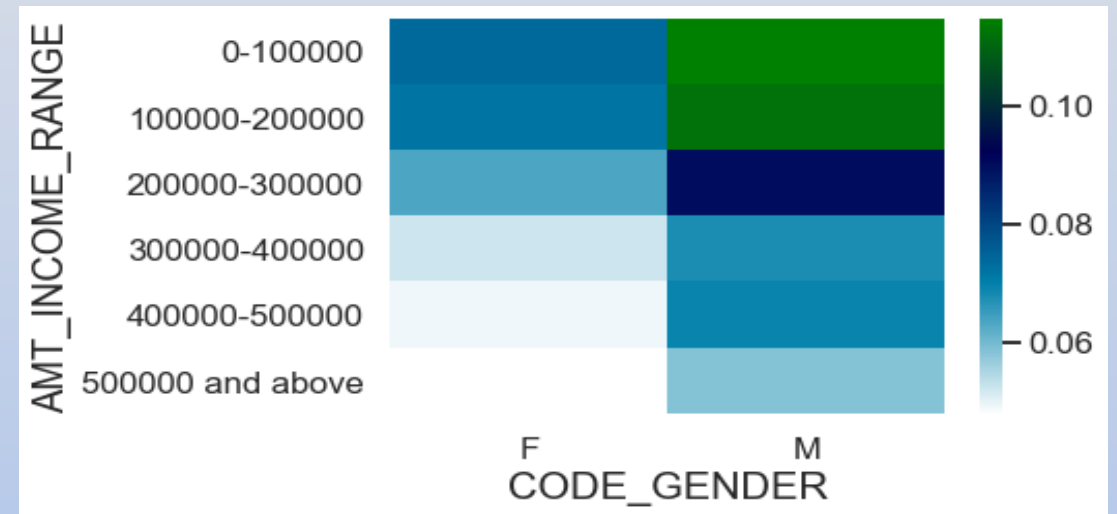
- Clients change their registration closer to application date, they are more likely to default.
- Clients who change their identity documents closer to loan application are less reliable than those who change it in advance. when the loan annuity is low the client are more likely to default.



Income Range vs defaulters(target 1)

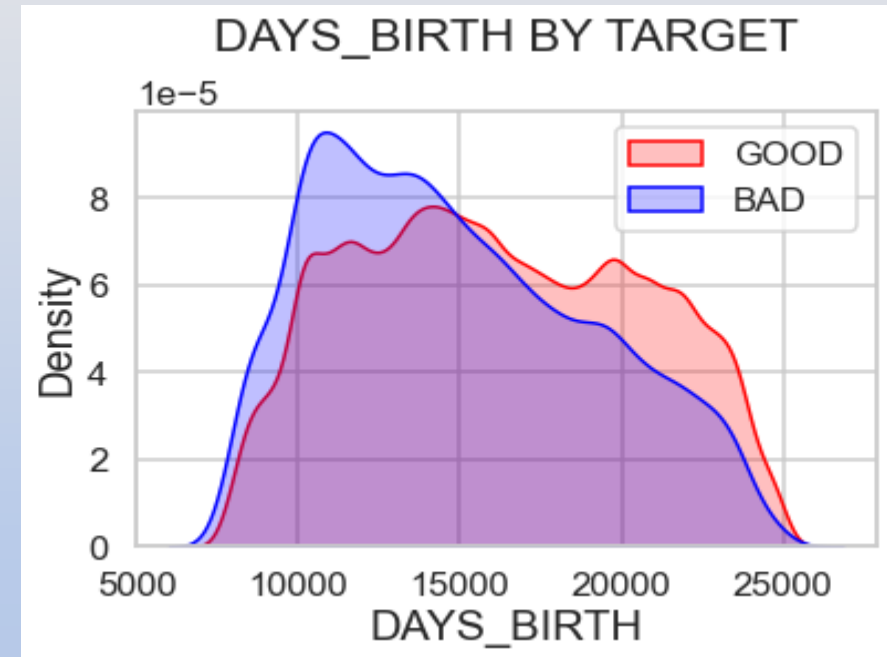
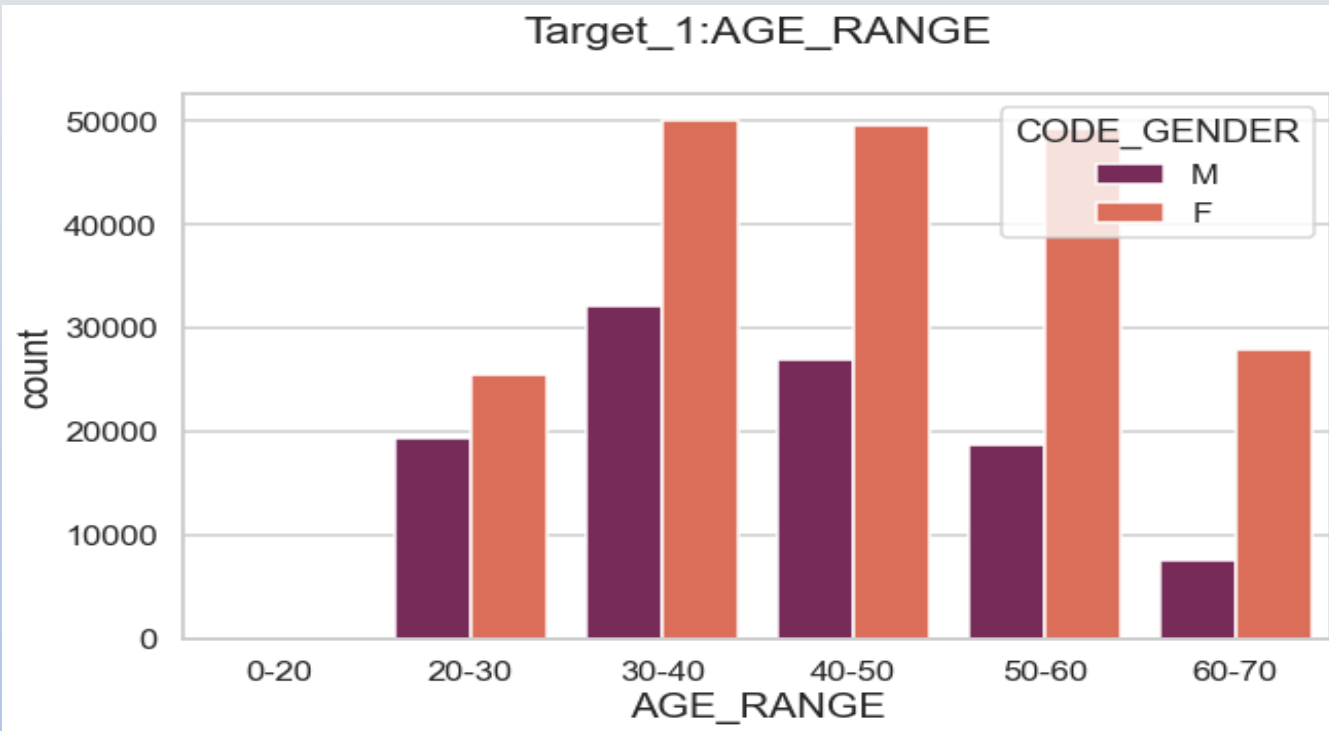


- Female have low income are more likely to default compare to male clients. Below 200000 income is having more credit. Above 400000 income is having less count.



- Male having below 300000 income are more likely to default. Default rate is increasing when income are decrease.

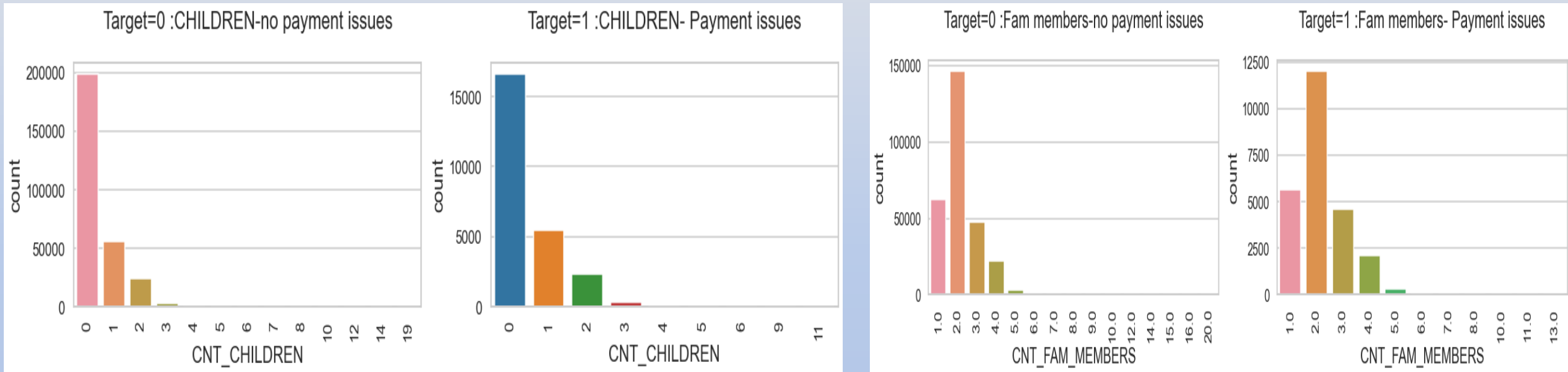
Age vs defaulters(target 1)



FINDINGS

- Here younger clients are less reliable compare to older clients. as client get older more reliable.
- In above bar chart age range from 30 to 60 is having more number of credits.Clients having below 30 age and above 60 year have less credicts
- As per the heat map male having below 40 age and . female having below age 30 are more likely to default. Default rate is increasing when age decrease. Above 55 age female clients are more reliable.

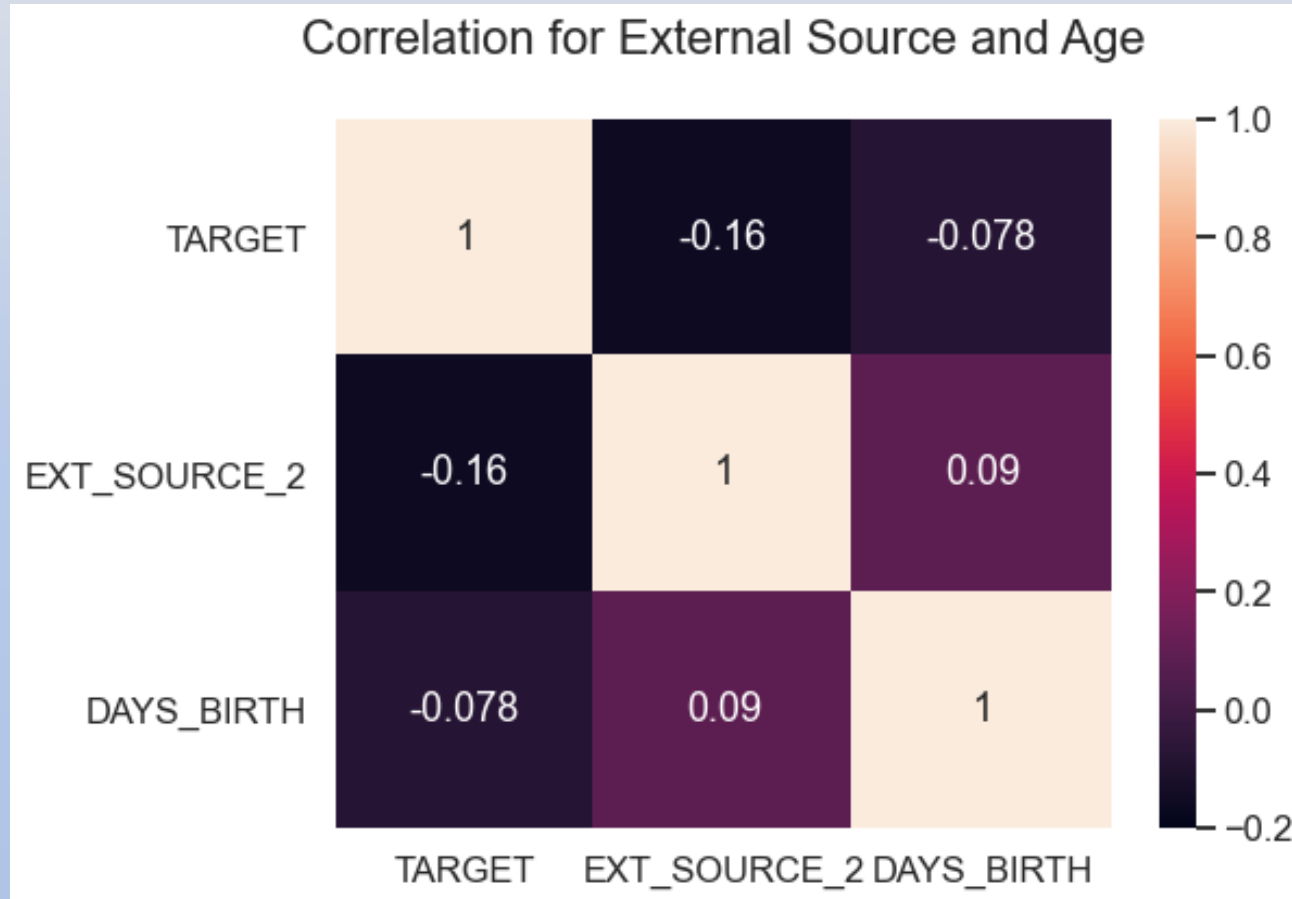
Family features



FINDINGS

- Client have 3 kids they are more likely to default compare to others. Most clients have no or very few children and are likely less member in family are more reliable .
- Increasing members in family increase the deafult. to repay loan on time.

Correlation for External Source and Age

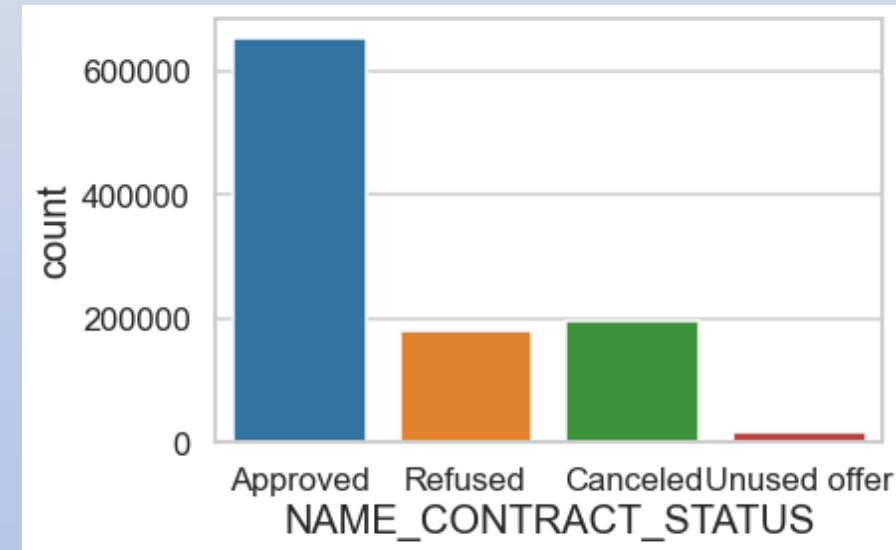
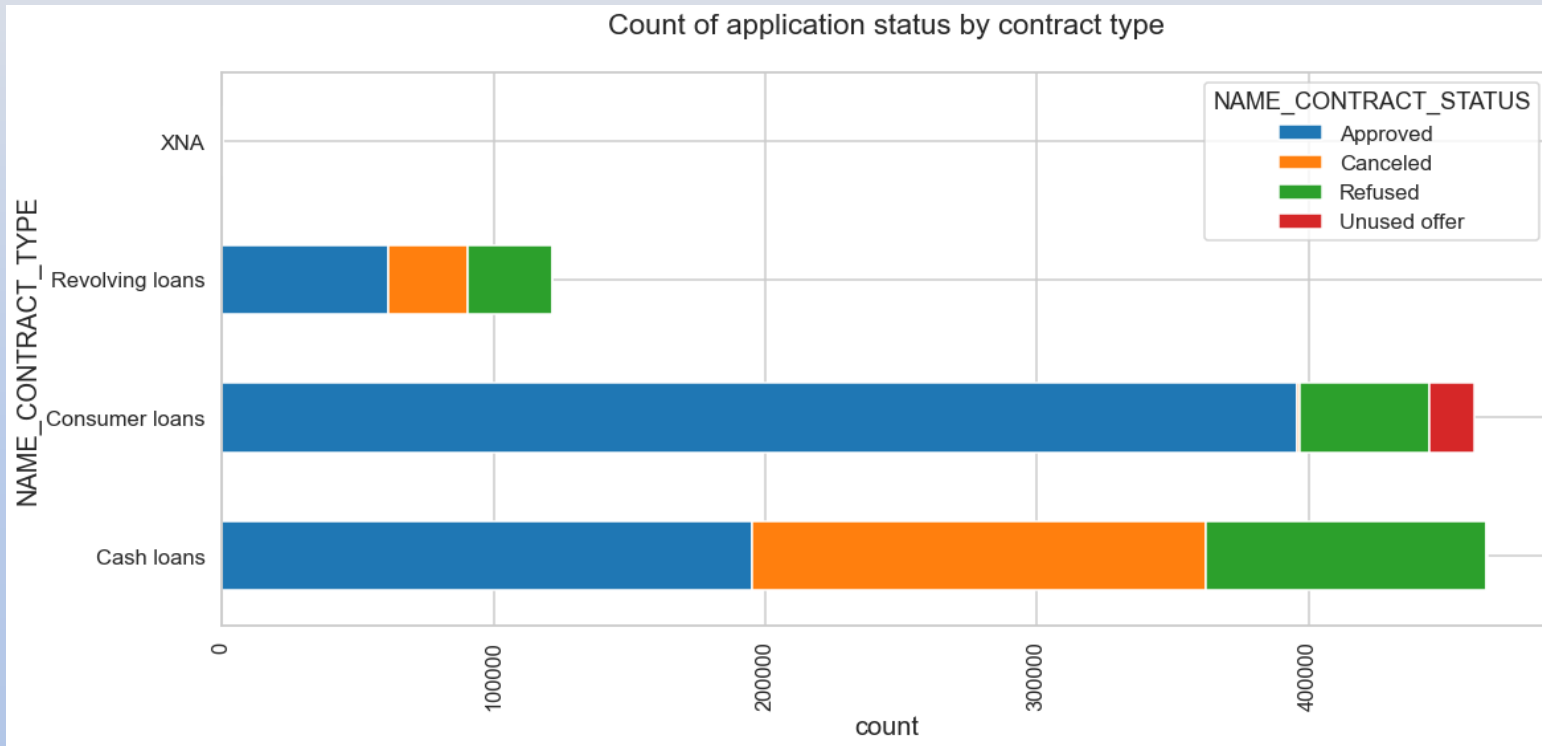


FINDINGS

- External source has negative correlations with the target
- The value of the EXT_SOURCE2 increases, the client is more likely to repay the loan.
- DAYS_BIRTH is positively correlated with EXT_SOURCE_2 ,age of client is increases
Normalized score from external data source.

PREVIOUS DATA VS CURRENT DATA ANALYSIS

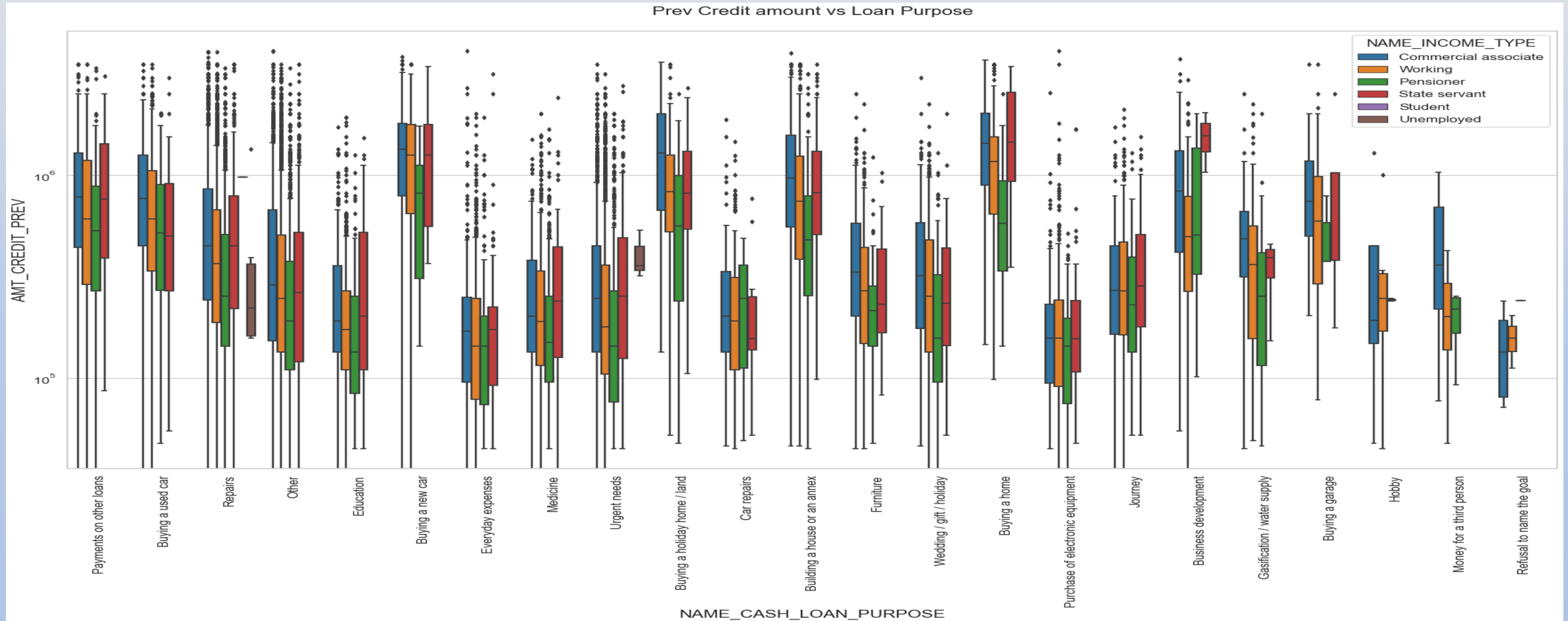
Application status vs contract type



FINDINGS

- Consumer loan applications are most approved loans and cash loans are most cancelled and refused loans.
- Consumer loans also rarely cancel, they are the most reliable type.

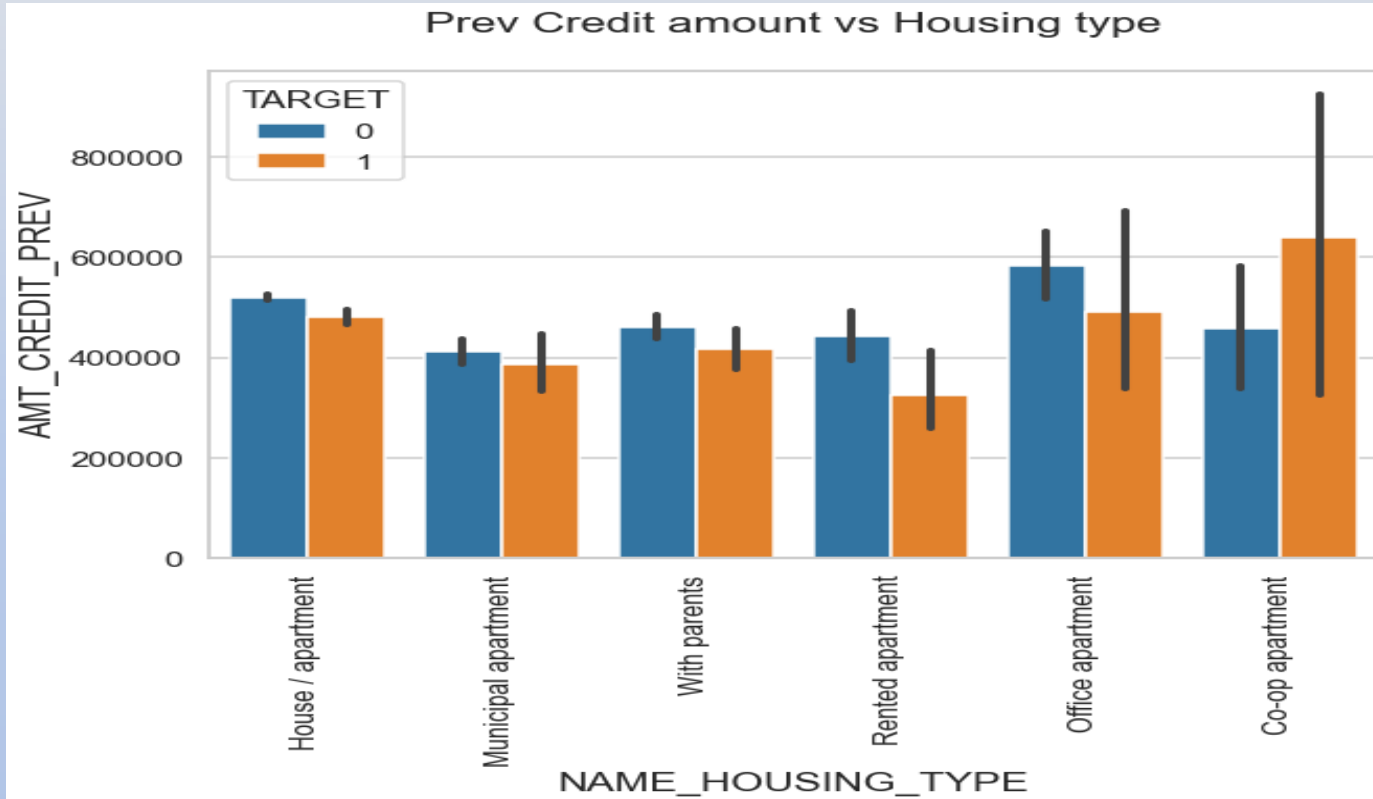
Previous Credit amount vs Loan Purpose



FINDINGS

- Buying a new car, buying a holiday home, buying a home and building a house or an annex are higher credit.
- Money for third person or a Hobby is having less credits applied for.

Previous Credit amount vs Housing type



FINDINGS

- Co-op apartment having higher credit of target 1. House/apartment, office apartment is having higher credit of target 0. That means co-op apartment housing type are more default.
- we can conclude that bank should avoid giving loans to the housing type of co-op apartment.
- Bank can focus mostly on housing type with parents or office apartment or House\apartment or municipal apartment for successful payments.

CONCLUSION

- Banks should focus more on contract type 'Student' , 'Pensioner' and 'Businessman with housing type with parents or office apartment or House\apartment or municipal apartment for successful payments.
- Accountant, high skill tech staff, manager and sales staff are more reliable.
- Clients have more income and better occupation lesser chance to defaulting.
- Male client's are more default than female client's.
- Loan purposes with repairs ,other and urgent needs are facing more difficulties for repay the loan.
- External source has negative correlations with the target, the value of the EXT_SOURCE2 increases, the client is more likely to repay the loan. DAYS_BIRTH is positively correlated with EXT_SOURCE_2 ,age of client is increases Normalized score from external data source.
- client's giving different address are more default.
- client's having own flat and car they are more reliable compare to client's don't have car and falt their own.
- client's are unemployed, Commercial associates, Working & State servants are mostly defaulters.

THANK YOU