



CREDIT EDA CASE STUDY

- VEDANT PATIL

Problem Statement

To identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate

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

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

Company wants to identify the driving factors behind loan default.

Company can utilize this knowledge for further risk assessment.

Analysis Approach

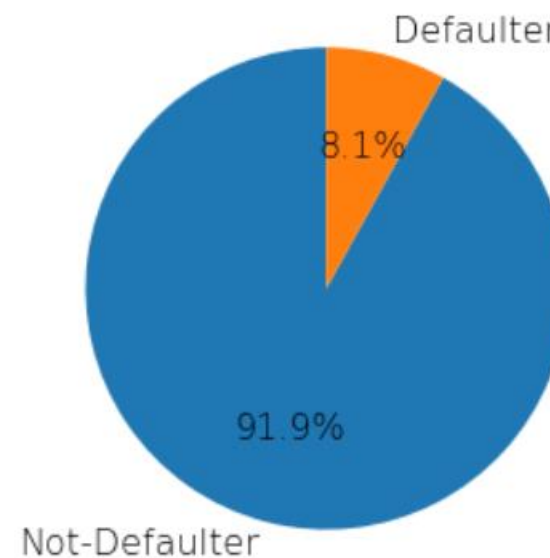
- ▶ Understanding the variables
- ▶ Importing/loading the data
- ▶ Checking the structure of the data
- ▶ Data cleaning and Imputation
- ▶ Outliers Analysis
- ▶ Data Analysis
- ▶ Conclusion

Data Imbalance Analysis

Given dataset has 91.9% Non defaulter and 8.1% of defaulter

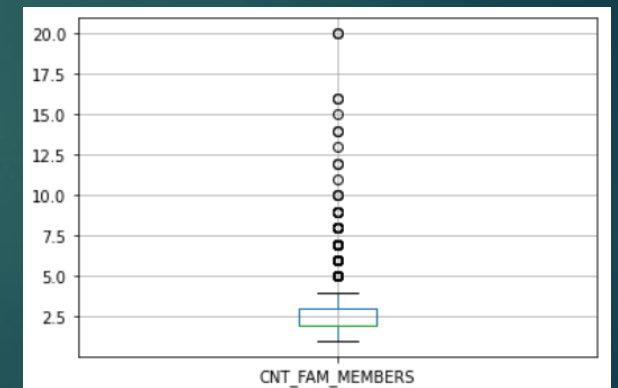
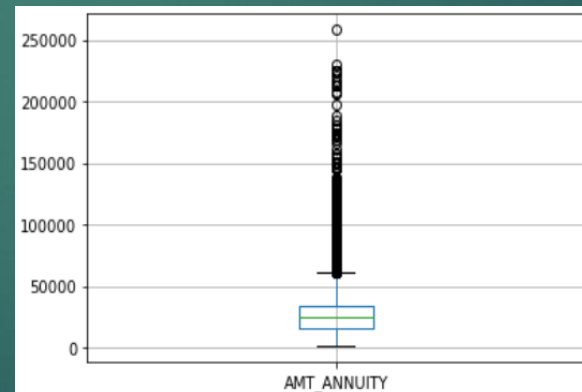
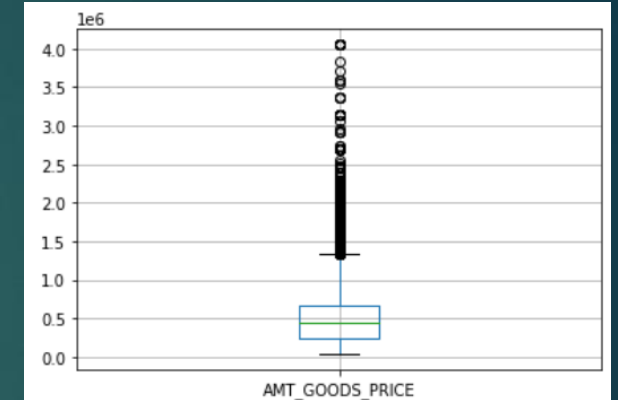
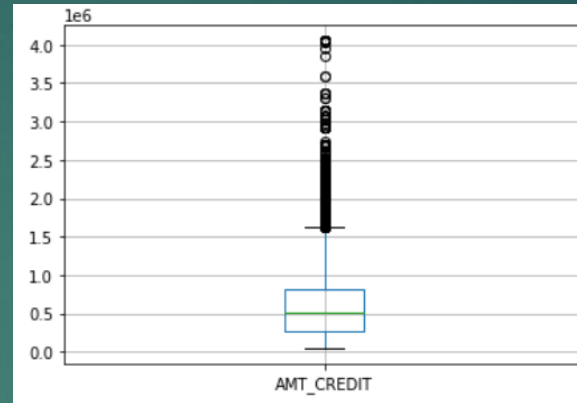
Hence the Given dataset is Imbalanced

Data Imbalance on Target Variable

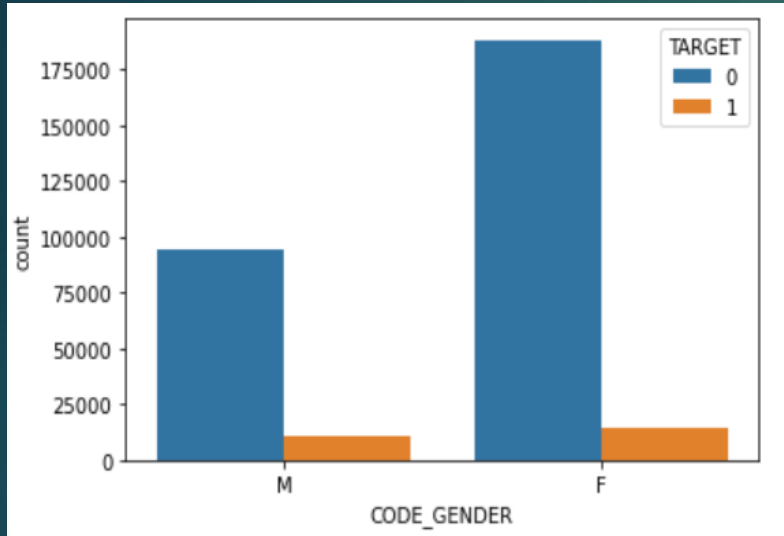


Outliers Analysis

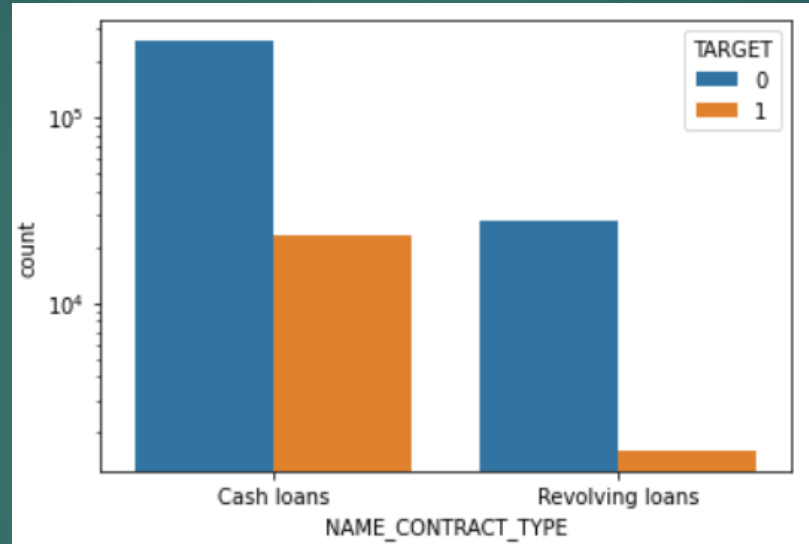
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, CNT_CHILDREN these columns have some outliers.
- AMT_INCOME_TOTAL has large number of outliers that shows few of the loan applicants have high income as compared to the others.
- EXT_SOURCE_2 has no outliers that means the data available is reliable.
- DAYS_EMPLOYED has outlier values around 3,50,000 (days) that is around 958 years which is not possible and hence this has to be incorrect entry.



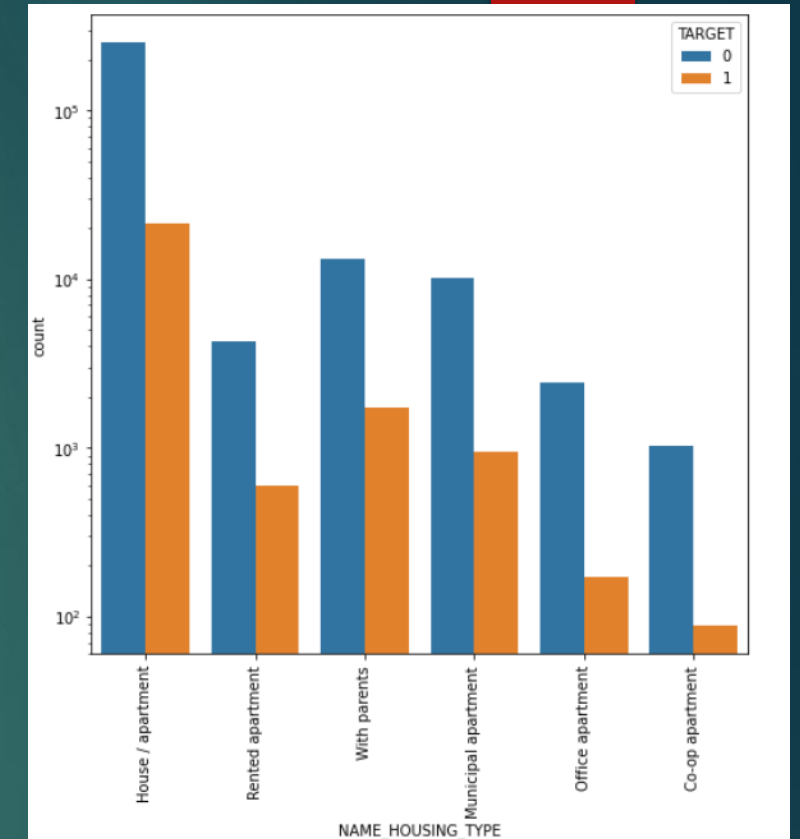
Univariate Analysis



Observation: The number of female clients is more than the number of male clients. Also males have a higher chance of not returning their loans when compared with females

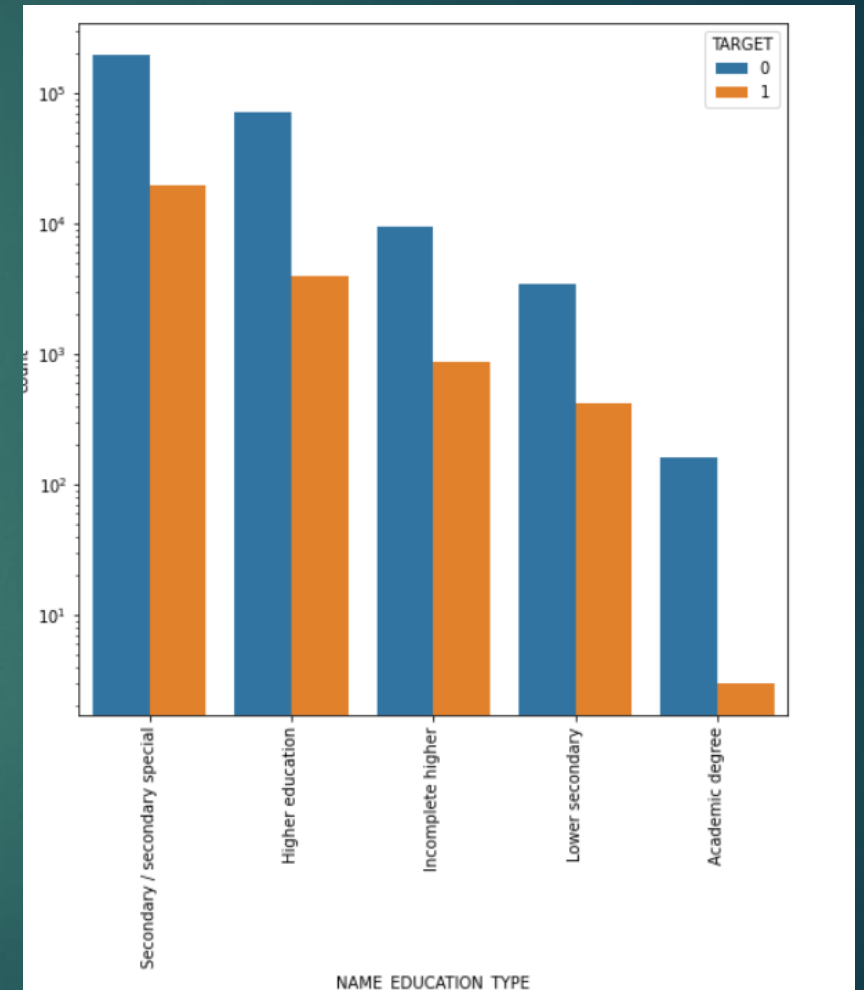


Observation for Name_Contract_type: Revolving loans are just a small fraction from the total number of loans



Observations: Most of the people live in House/apartment People living in Co-op apartments have lowest default rate And People living with parents and living in rented apartments have chances of defaulting

Observations: Most of the clients have Secondary / secondary special education which are followed by clients with Higher education. Clients having academic degree are less in numbers. Though the Lower secondary category are lower in numbers but may have the chances of defaulters.



Observations from Univariate Categorical Analysis:

Female tends to take more loans than Males.

People with Medium total income may default more.

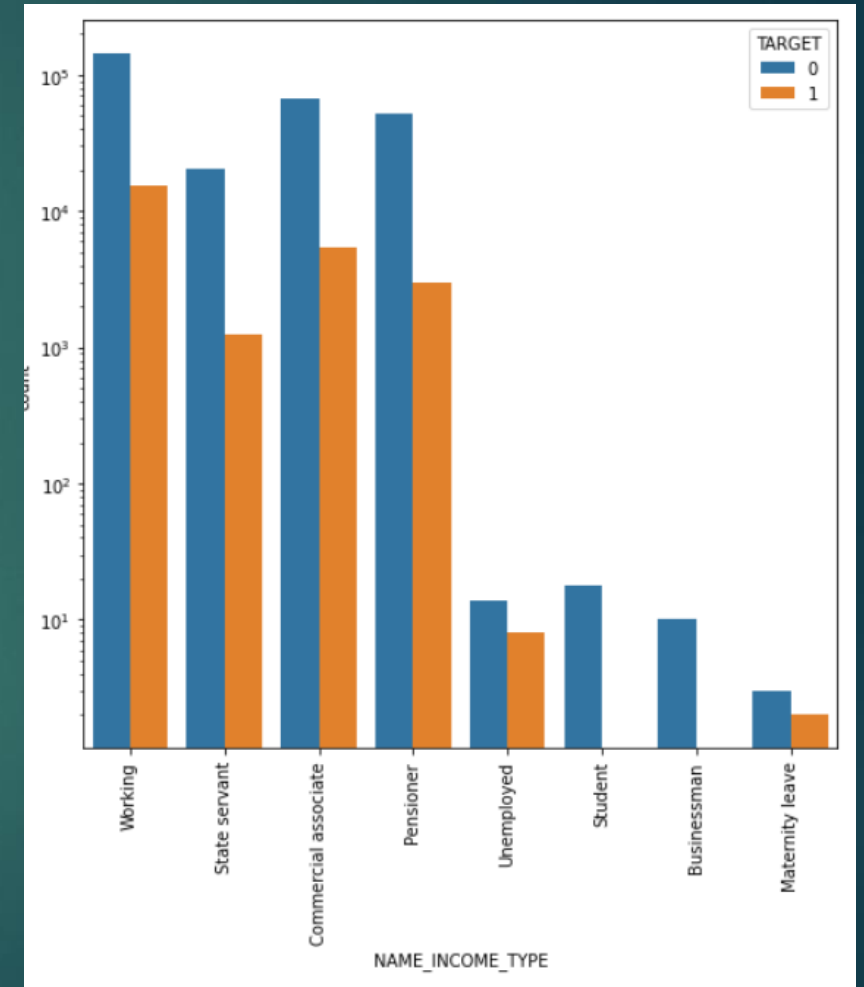
Clients having high Credit amount are less likely to default.

People who lives in house / apartment are taking more loans.

Married people are taking more Loan as compared to other categories.

we can conclude that secondary/special educated loan applyer are in high in number

People who have started application process on sunday are less chances to default.

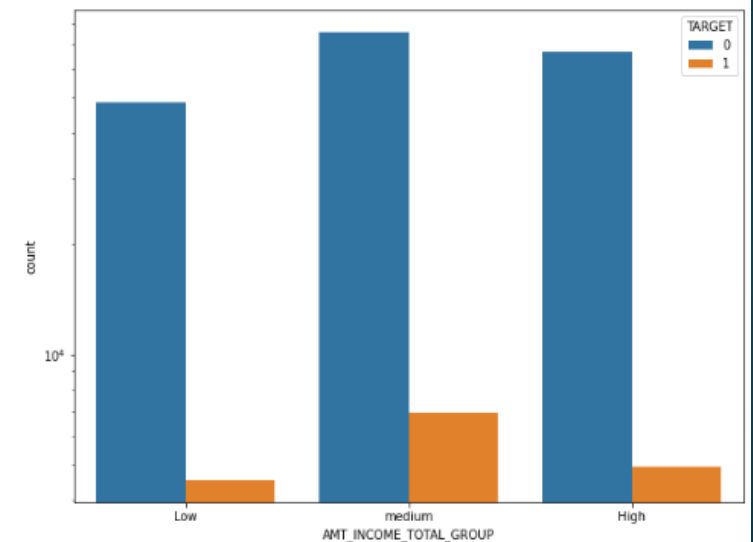
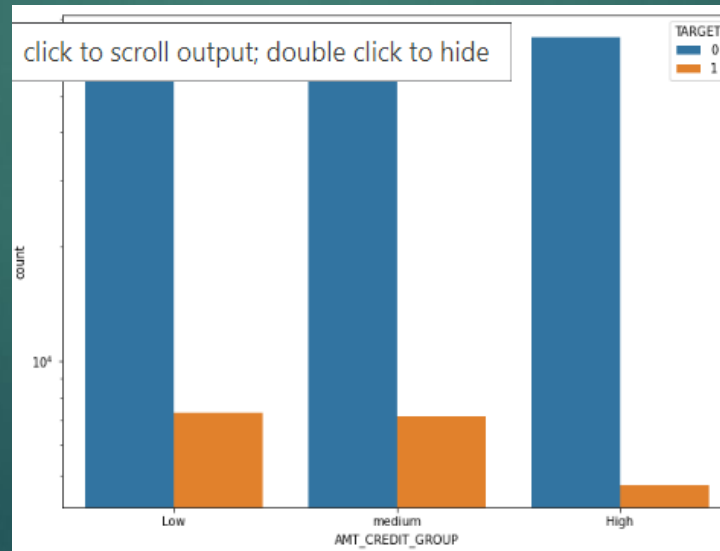
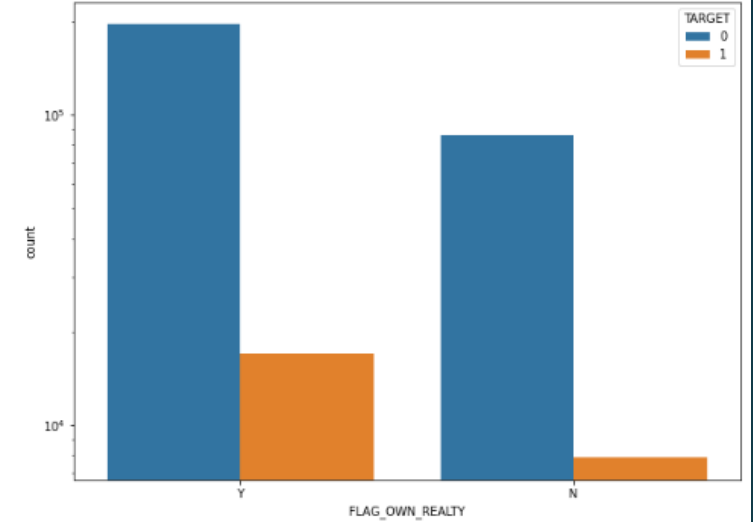
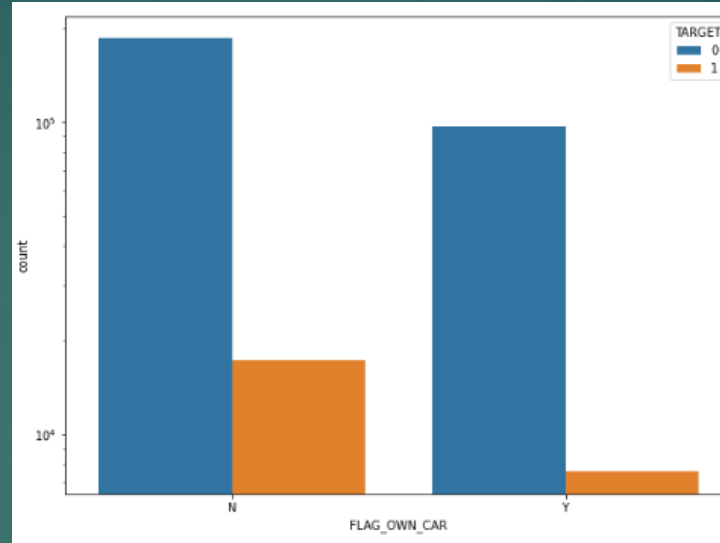


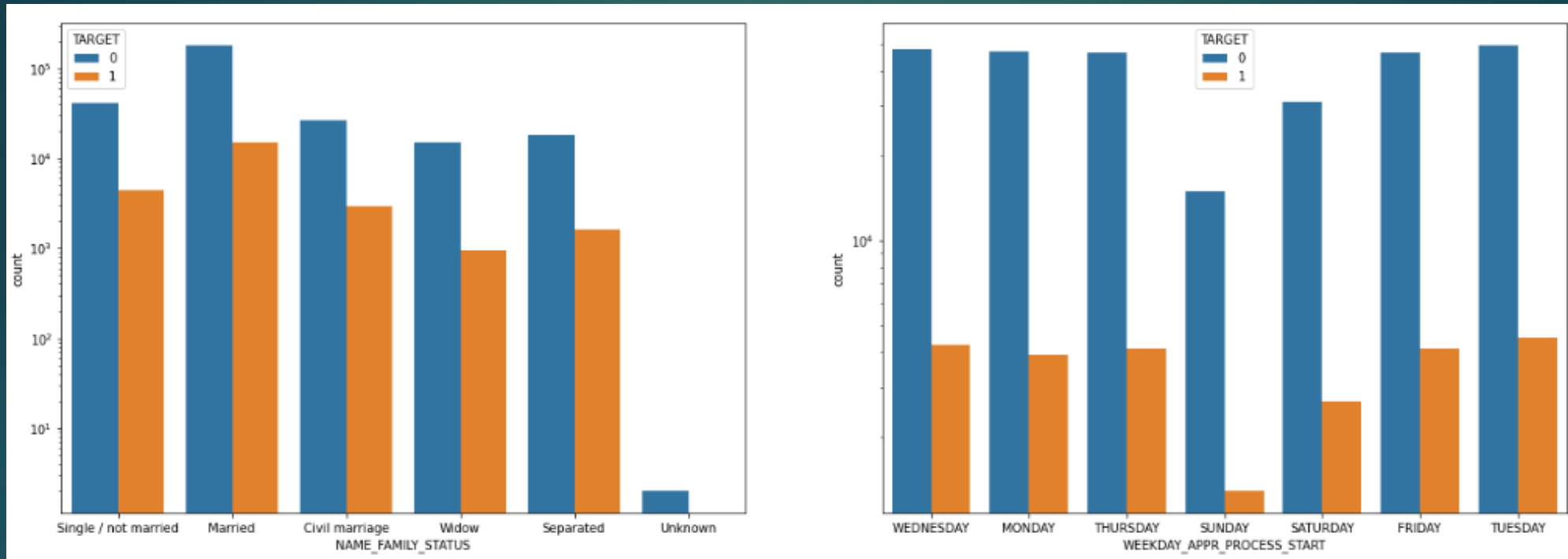
People with real estate are likely to take more loans.

People who don't have a car tends to take more loans.

People tend to take more cash loans, and default chances of revolving loans is less.

Saturday and sunday are not too busy for bank in terms of loan applications.



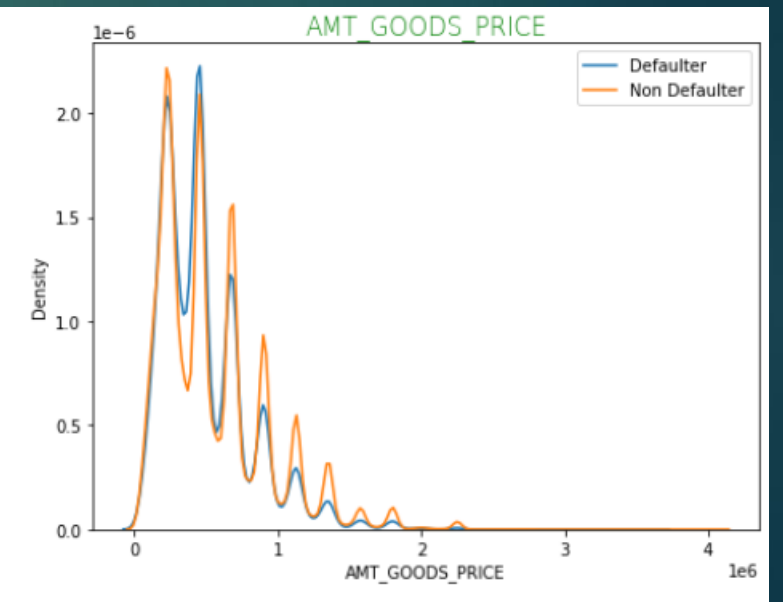
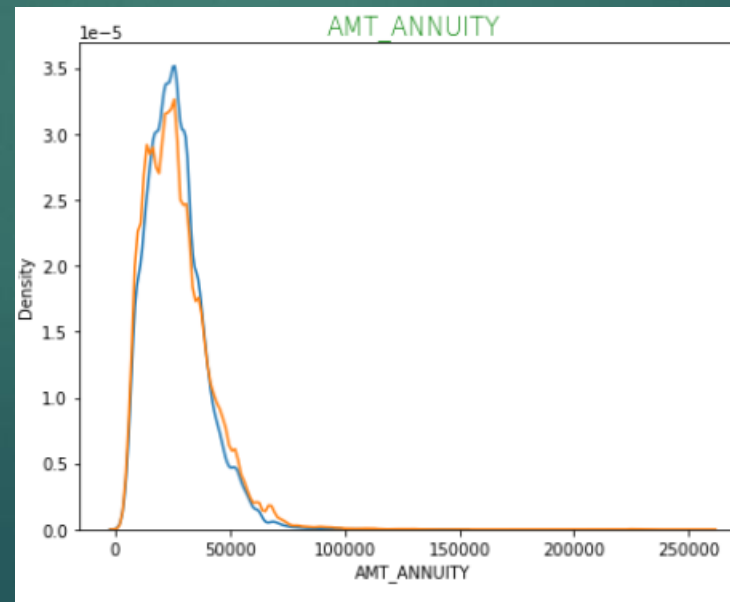
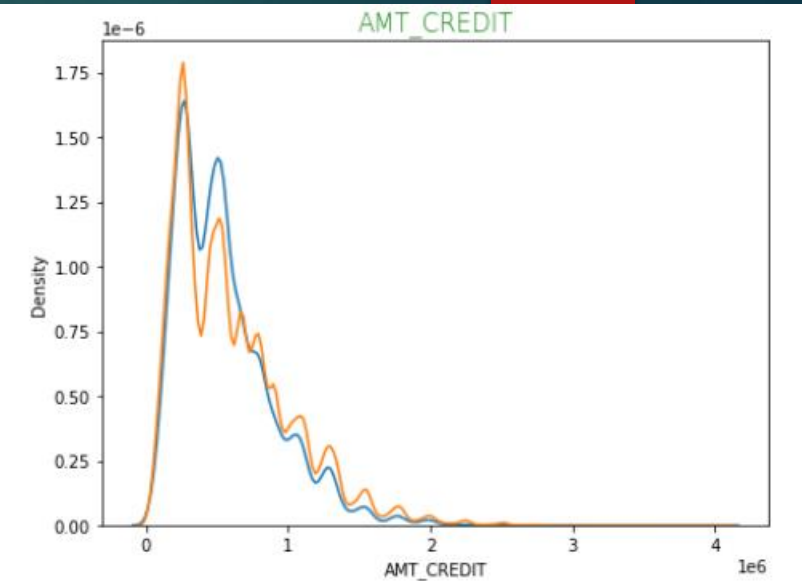
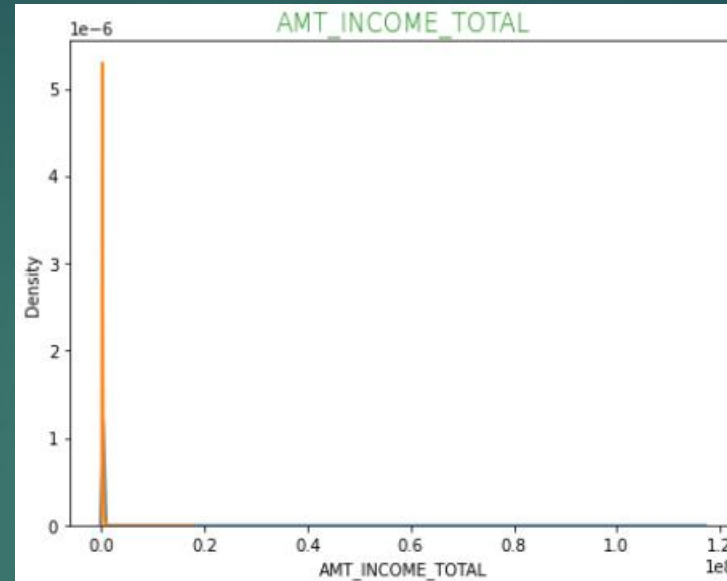


People who have started application process on sunday are less chances to default.

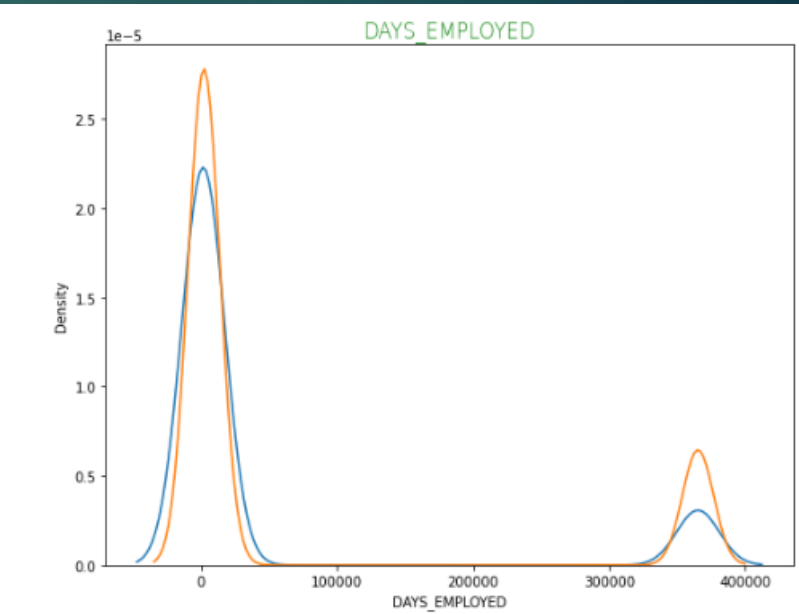
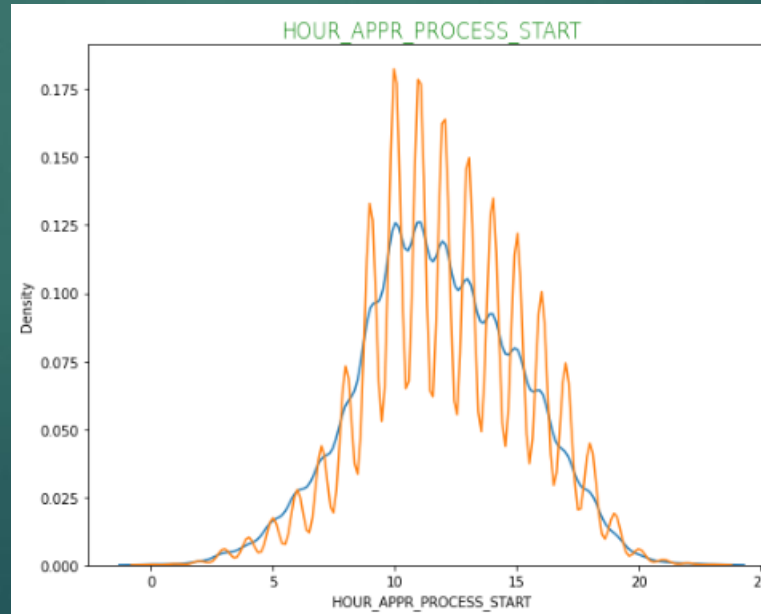
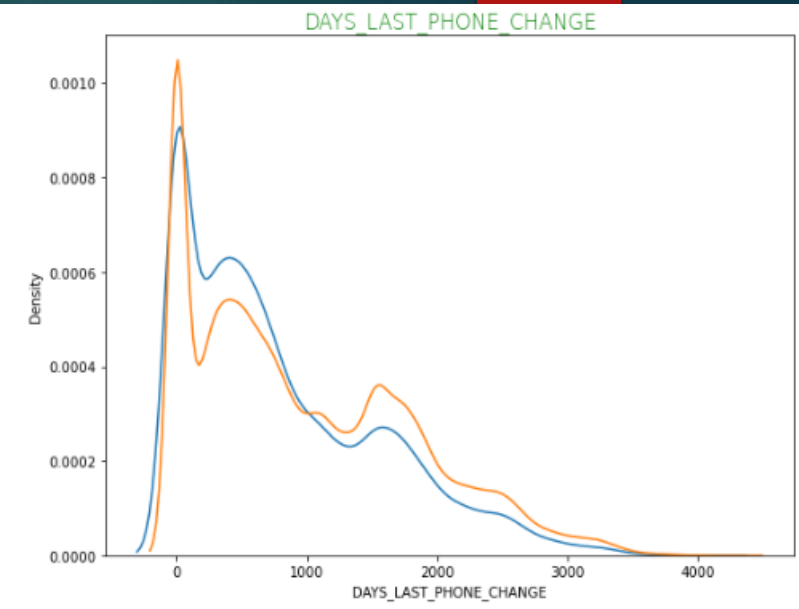
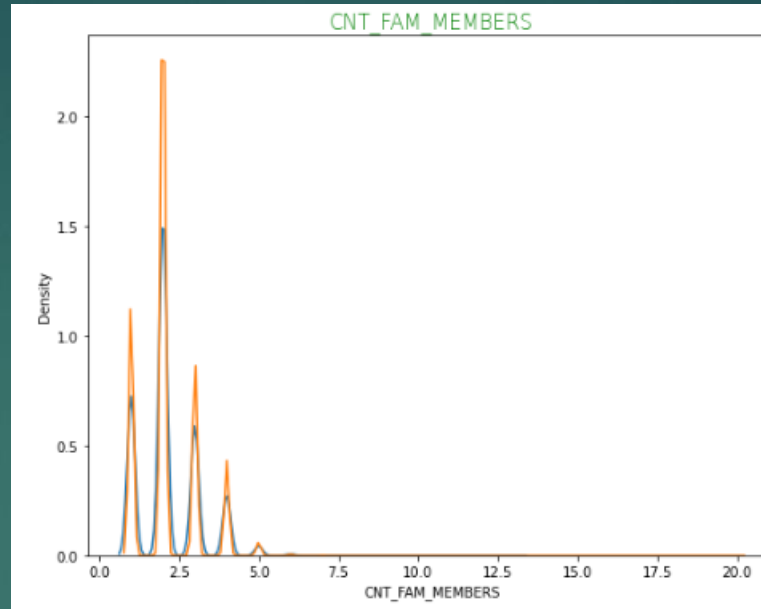
Saturday and sunday are not too busy for bank in terms of loan applications.

Numerical Analysis

- **Observations:**
- Most number of loans are given for goods price below 10 lac mount.
- Most number of people are paying annuity less than 50000 for the credit loan.
- The Non defaulters and defaulters distribution overlap in all the plots and hence we cannot use any of these variables to make a decision.
- Credit amount of the loan is less than 10 lacs amount.



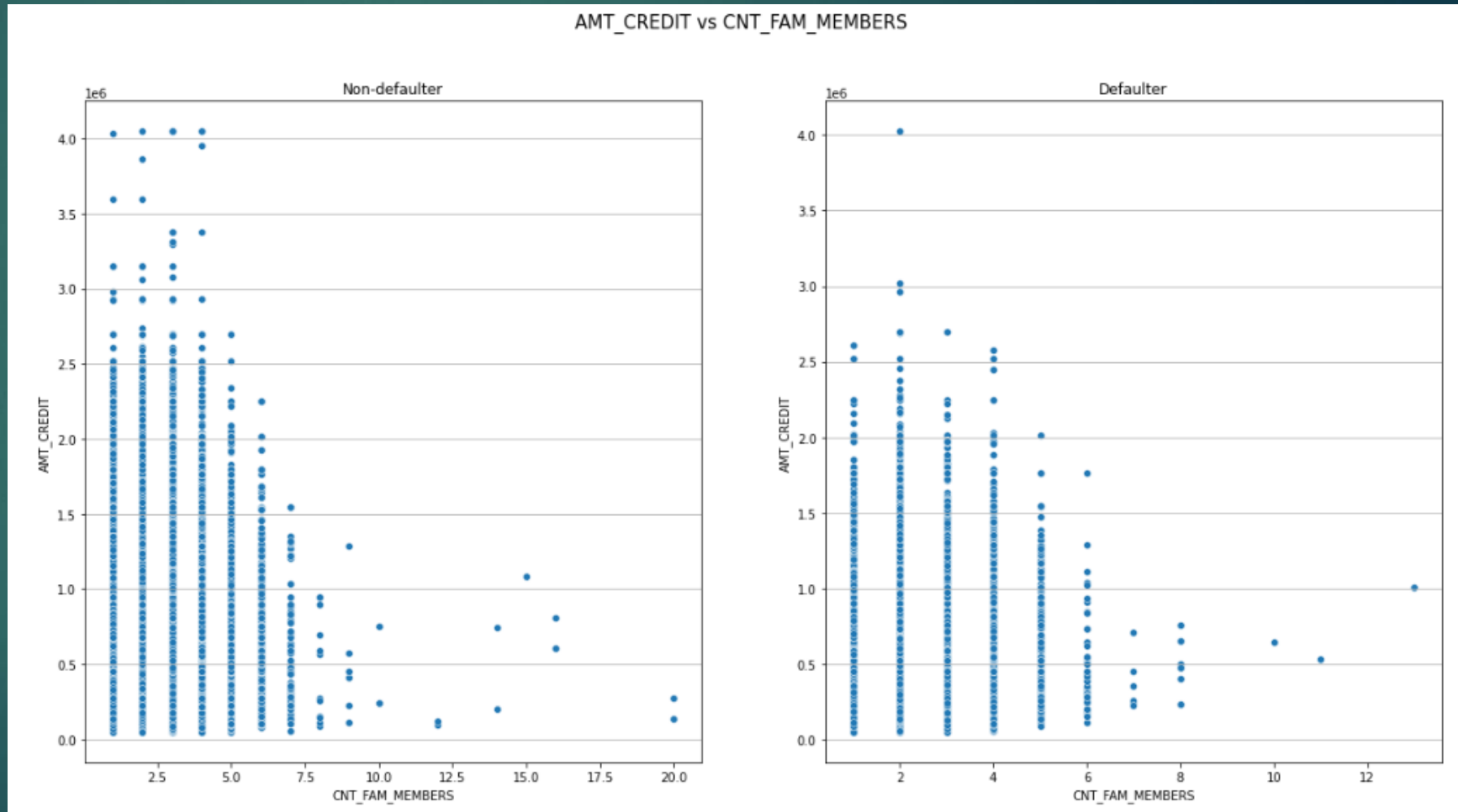
- ▶ Observations: We can see from above graphs
- ▶ Nuclear family are taking more loans.
- ▶ People who just got employed likely to take more loans.
- ▶ People with age between 10000-days and 15000-days yrs likely to take more loans.
- ▶ People who retired likely to take more loans.
- ▶ People whose ids got published between 4000 days and 5000 days ago are likely to take more loans.
- ▶ High number of applications are filed in 10 AM to 2 PM



Observations:

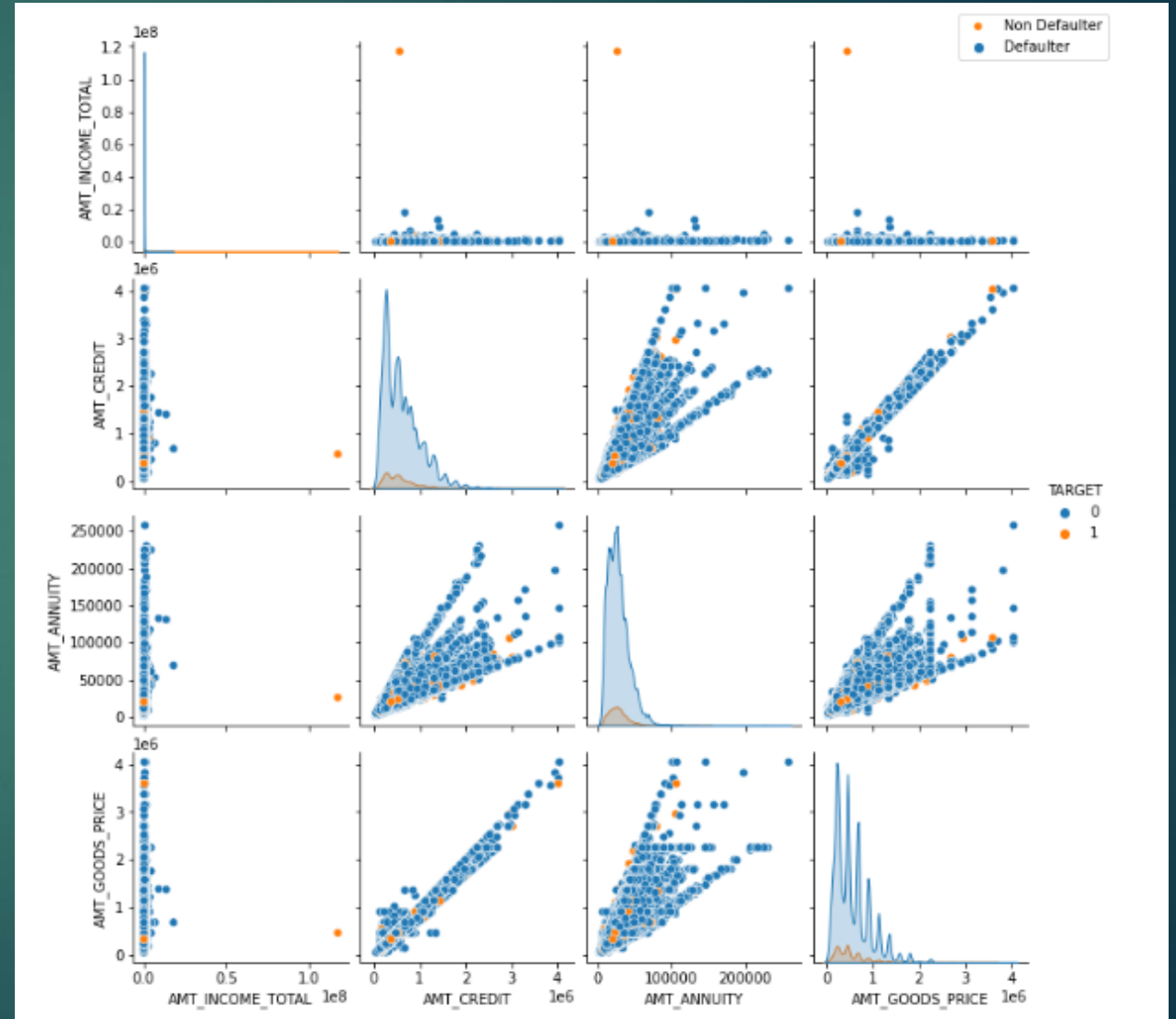
Clients who has Big family size and low credit amount may default less.

Clients who has small family size and higher Amount credit may default less.

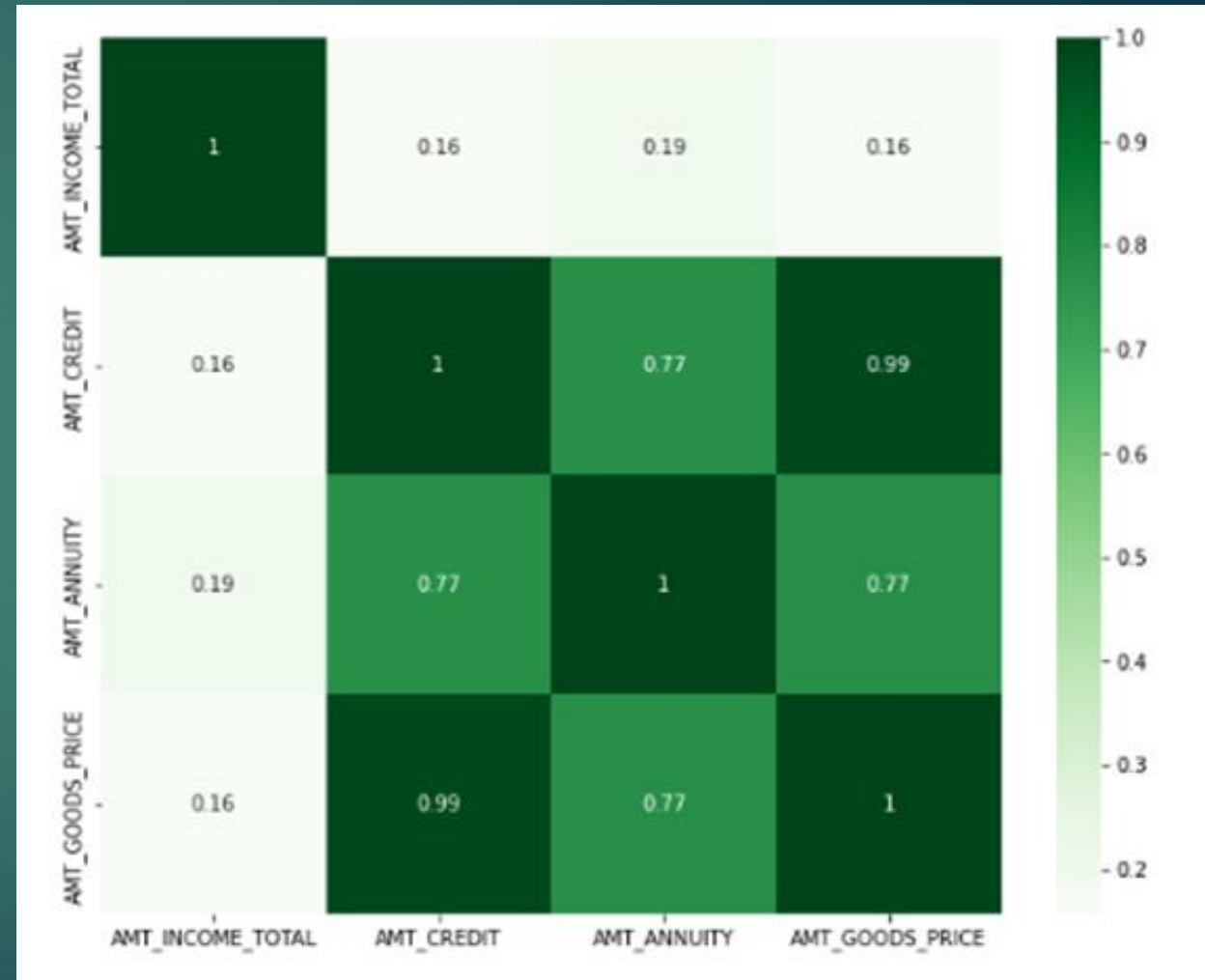


Multivariate Analysis

- ▶ **Observations:**
- ▶ AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as based on the scatterplot.
- ▶ There are very less defaulters where AMT_CREDIT > 3M.
- ▶ When AMT_ANNUITY > 15000 and AMT_GOODS_PRICE > 3M then there is low chance of defaulter.




- ▶ Observations:
- ▶ Applicants owning goods of higher value may go for loans of higher amounts.
- ▶ High correlation between AMT_CREDIT and AMT_GOODS_PRICE.



Conclusion

- ▶ Banks should more focus on contract type 'Student' , 'pensioner' and 'Businessman' and having housing type other than 'Co-op apartment' for successful payments.
- ▶ Go for clients from housing type 'With parents' as thees people are having least number of unsuccessful payments.
- ▶ Banks should focus less on income type 'Working' as they are having most number of unsuccessful payments
- ▶ Target variable for Application dataset should be - TARGET
- ▶ Target variable for Previous dataset should be - NAME_CONTRACT_STATUS



▶ Following are the Major variables which are to be considered before approving application to minimize risk of loss:

- ▶ AMT_ANNUIITY
- ▶ CODE_GENDER
- ▶ AMT_CREDIT
- ▶ AMT_INCOME_TOTAL
- ▶ NAME_INCOME_TYPE
- ▶ NAME_EDUCATION_TYPE
- ▶ NAME_HOUSING_TYPE
- ▶ DAYS_BIRTH
- ▶ DAYS_EMPLOYED
- ▶ NAME_CONTRACT TYPE