CREDIT EDA ASSIGNMENT

SUBMITTED BY ADITYA VIKRAM

CONTENT

- Problem Statement
- Dataset 1 – Basic Approach
- Dataset 2 – Basic Approach
- Detailed Approach
- Merged Dataset
- Conclusion

PROBLEM STATEMENT

The purpose of the assignment is to understand risk analytics in banking and financial services and thus to identify the individuals who should be allotted loans and who should be denied to avoid any loss in company.

BASIC APPROACH – DATASET 1(APPLICATION DATA)

- Reading Data
- Handling Null Values- Replace/Delete
- Deleting Unimportant Columns
- Standardization
- Handling Outliers
- Binning
- Data Imbalance
- Univariate Analysis
- Bivariate Analysis
- Correlation

BASIC APPROACH – DATASET 2(PREVIOUS APPLICATION)

- Reading Data
- Handling Null Values- Replace/Delete
- Deleting Unimportant Columns
- Standardization
- Handling Outliers
- Bivariate Analysis
- Correlation

DETAILED APPROACH

IMPORTING LIBRARIES

- Imported libraries like NumPy , Pandas , Matplotlib , Seaborn.
- Imported Warnings.
- Set column and rows display so that all columns can be displayed easily.

READING DATASET

- 2 dataset were provide so read both dataset namely application data, previous application.
- Referred column description file to understand the value of column of banking loan dataset.
- Read shape and datatypes information to understand the data which we will be dealing with further.

HANDLING NULL VALUES

- Checked null values for dataset 1 and dropped the columns having null values greater than 30 % as those does not look relevant based on column description and as the column "OCCUPATION_TYPE" looks relevant so kept it and omitted others as replacing the values or keeping the numerical values may skew the result. Hence, dropped the columns for best result.
- Now 18 columns are left for dataset1 and thus dealing with other 6 columns and replacing it with median as it is non continuous numerical value column and the columns are AMT_REQ_CREDIT_BUREAU_YEAR ,AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_MON ,AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_DAY ,AMT_REQ_CREDIT_BUREAU_HOUR .
- Now remains 12 columns, so dealing with next columns; OBS_30_CNT_SOCIAL_CIRCLE ,DEF_30_CNT_SOCIAL_CIRCLE ,OBS_60_CNT_SOCIAL_CIRCLE ,DEF_60_CNT_SOCIAL_CIRCLE , AMT_GOODS_PRICE, AMT_ANNUITY, CNT_FAM_MEMBERS, DAYS_LAST_PHONE_CHANGE , EXT_SOURCE_2 , are numerical value column and have null values less than 1% so imputing it with median value as it will not have any impact on overall analysis and for categorical column replaced it with mode .

HANDLING NULL VALUES

- EXT_SOURCE_3 – dropping it as it does not look important based on column description.
- Only columns remain is OCCUPATION_TYPE and based on description it looks important as it has Job of the targeted sector which will have a big role in determining if they are capable of repaying of loan or not as person with high paying job have a greater chance of repaying the loan and can be granted large amount of loan. Therefore, I am neither deleting or replacing the null value with mode as it will not give clear picture and thus imputing it with 'Unknown". Also, this will give idea who is in requirement of loans.
- For dataset 2 - 11 columns have null value more than 40% so dropped those columns and those column does not look relevant based on description.
- For dataset 2 – 5 columns are left with null values and out of which 4 columns are numerical and based on describe checked and found that there are non continuous values o replaced with median and for remaining 1 column replaced the null values with mode.

DELETING UNIMPORTANT COLUMNS

- For dataset 1 and 2 - Deleted columns of lesser importance based on column description as it will not add any insights for decision making.
- Deleted columns from dataset 1 – “FLAG_MOBIL','FLAG_EMP_PHONE','FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE','FLAG_PHONE','FLAG_DOCUMENT_2','FLAG_DOCUMENT_3','FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5', 'FLAG_DOCUMENT_6', 'FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8', 'FLAG_DOCUMENT_9','FLAG_DOCUMENT_10','FLAG_DOCUMENT_11','FLAG_DOCUMENT_12','FLAG_DOCUMENT_13','FLAG_DOCUMENT_14','FLAG_DOCUMENT_15','FLAG_DOCUMENT_16','FLAG_DOCUMENT_17','FLAG_DOCUMENT_18','FLAG_DOCUMENT_19','FLAG_DOCUMENT_20','FLAG_DOCUMENT_21’,"REGION_POPULATION_RELATIVE","FLAG_EMAIL","REGION_RATING_CLIENT","REGION_RATING_CLIENT_W_CITY","REG_REGION_NOT_LIVE_REGION","REG_REGION_NOT_WORK_REGION","LIVE_REGION_NOT_WORK_REGION","REG_CITY_NOT_LIVE_CITY","REG_CITY_NOT_WORK_CITY","LIVE_CITY_NOT_WORK_CITY","DAYS_LAST_PHONE_CHANGE“.
- Deleted columns from dataset 2- 'NAME_PORTFOLIO','SELLERPLACE_AREA','WEEKDAY_APPR_PROCESS_START','HOUR_APPR_PROCESS_START','FLAG_LAST_APPL_PER_CONTRACT','NFLAG_LAST_APPL_IN_DAY’.

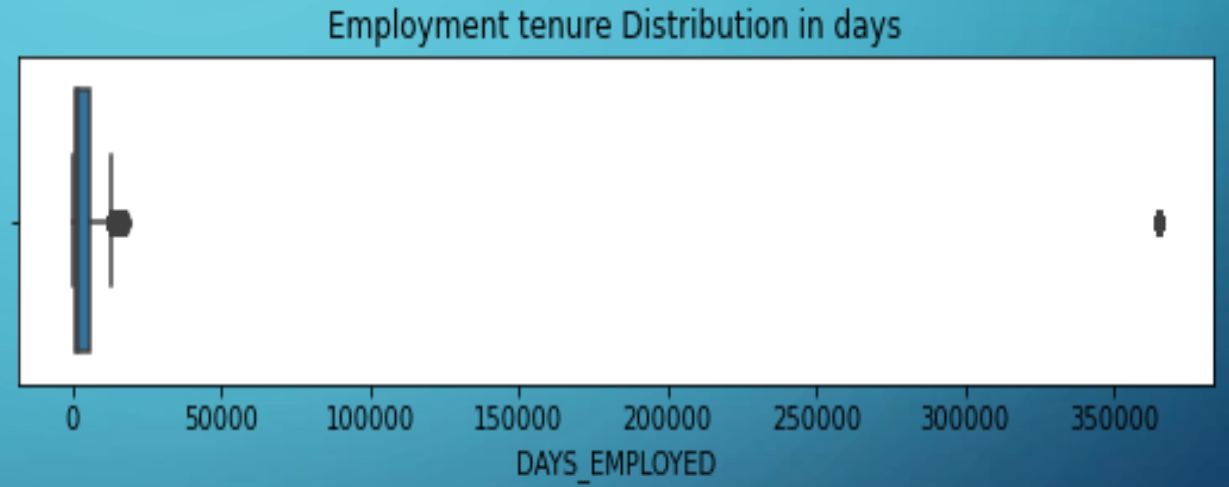
STANDARDIZATION

- For dataset 2 - Dealing with negative values in days column - Columns("DAYS_BIRTH","DAYS_EMPLOYED","DAYS_REGISTRATION","DAYS_ID_PUBLISH","DAYS_LAST_PHONE_CHANGE") are all day's column with negative value which is not possible hence converted all columns to positive using abs.
- For dataset 1- High Value numerical column- Converting High Value numerical column into categorical via binning. Converting days columns to years for standardization and then binning.
- For dataset 1 - Categorization- Converting useful column to categorical for insights.
- For dataset 2 – Column DAYS_DECISION has negative value so converting it to positive.

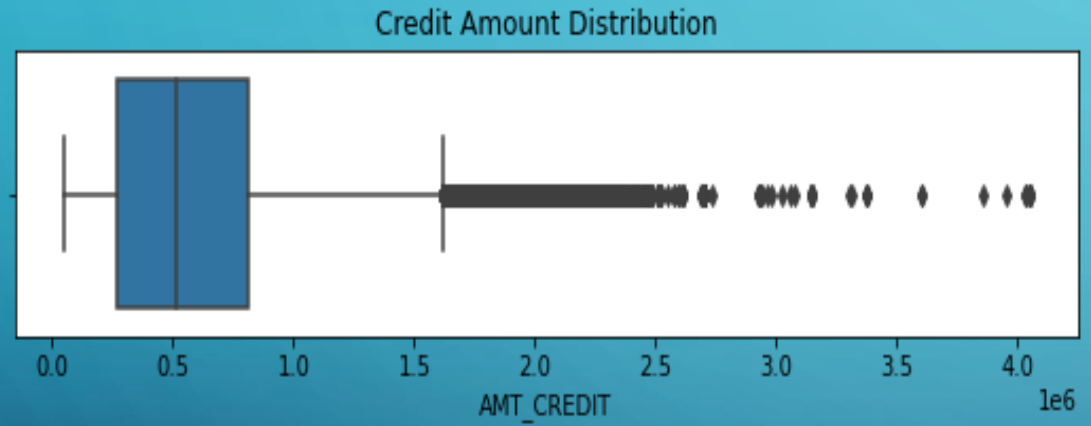
OUTLIERS-DATASET 1

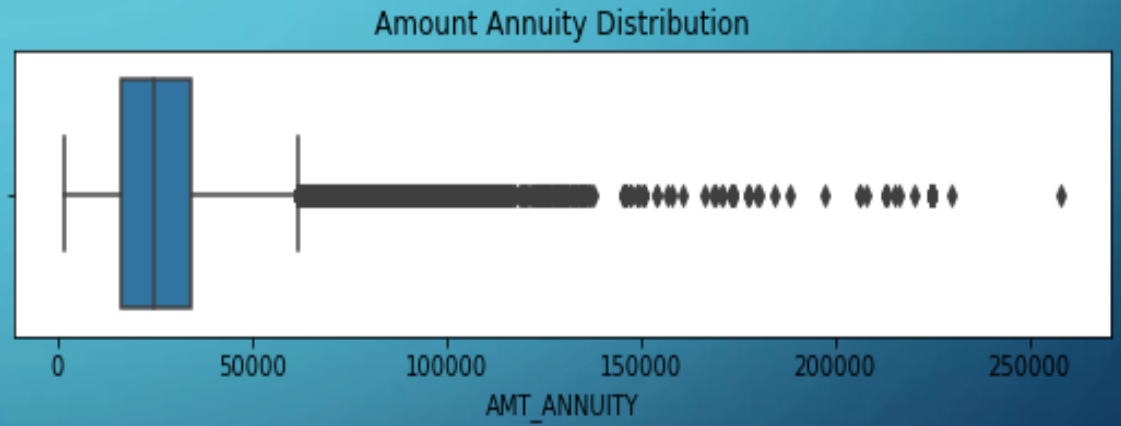
For column - DAYS_EMPLOYED we can see that column has huge outlier.

Similarly , for few other numerical columns and found that there were few outlier but not much and in some columns there were no outliers.



OUTLIERS-DATASET 1

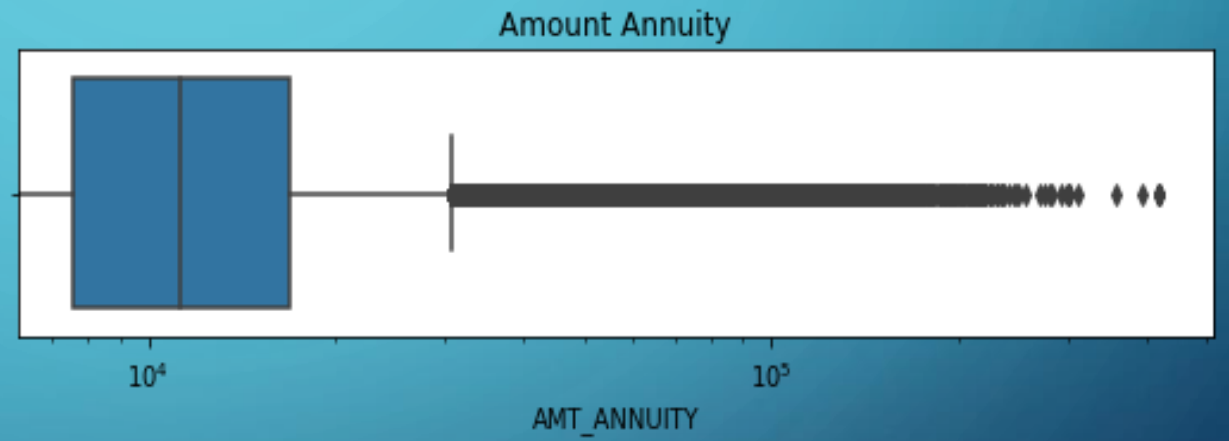




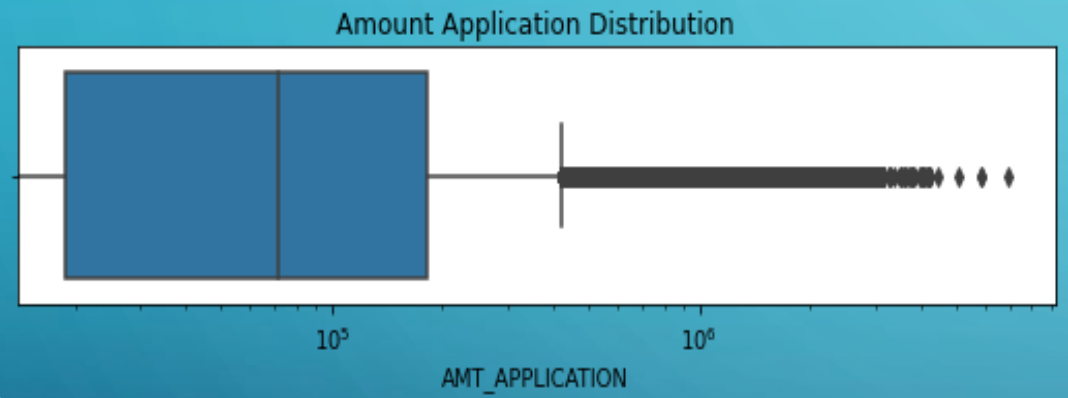
OUTLIERS-DATASET 2

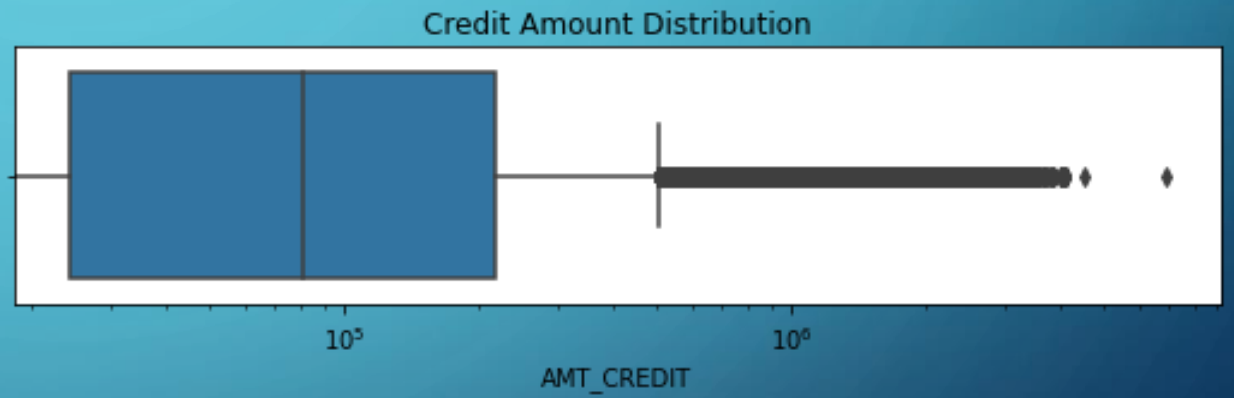
For column - AMT_ANNUITY we can see that column has few outliers.

Similarly , for few other numerical columns and found that there were few outlier but not much and in some columns there were no outliers.



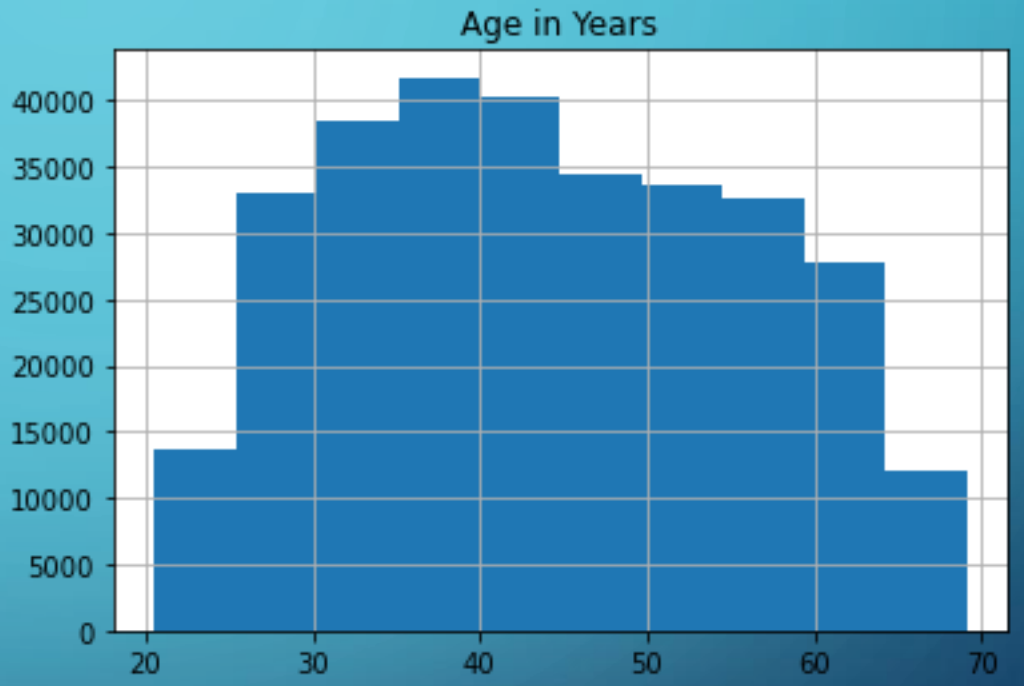
OUTLIERS DATASET-2





BINNING

For few columns as the values were high so binning them would make more sense such as age group, employee experience, income – all these can be grouped.



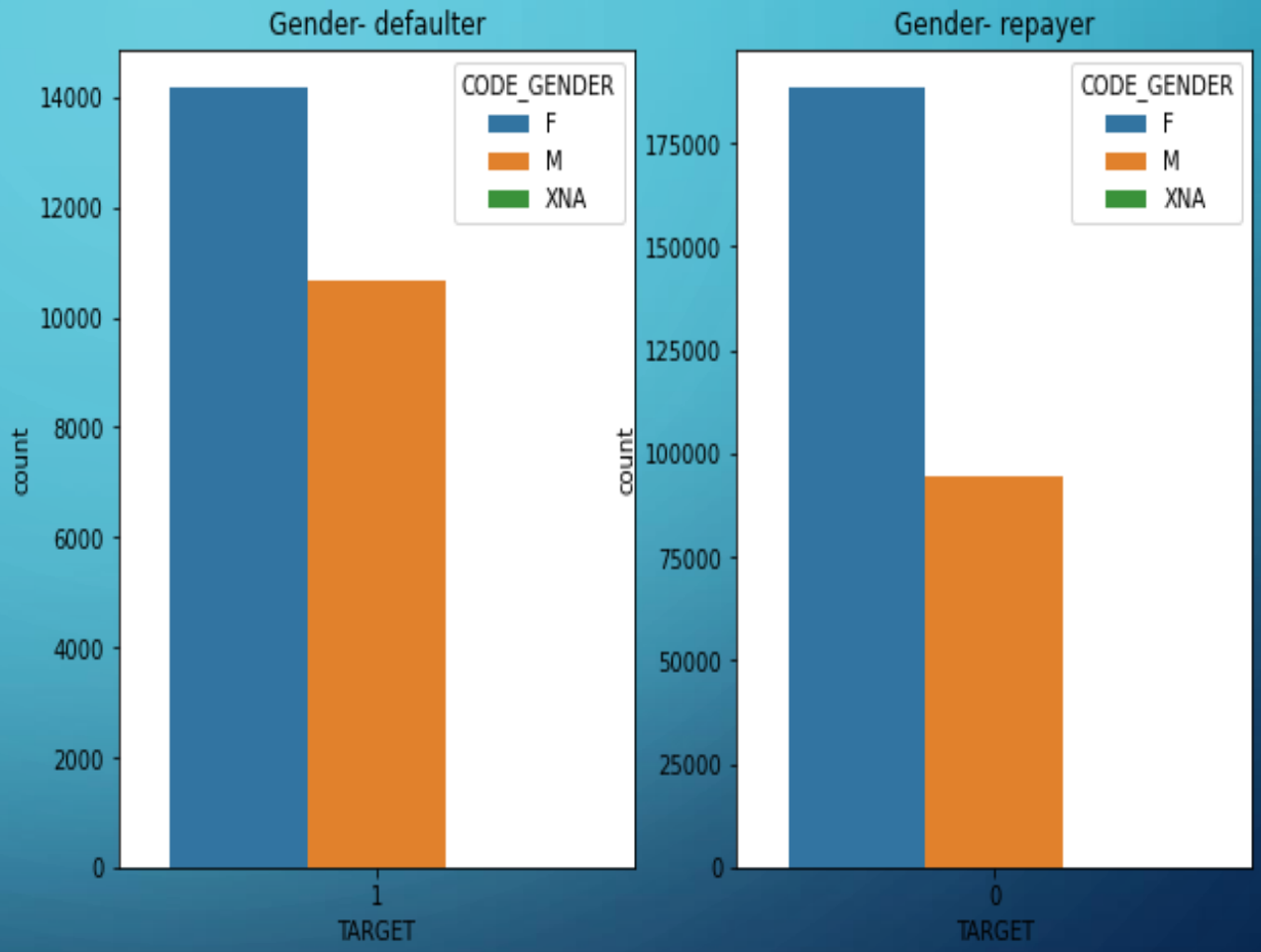
DATA IMBALANCE

The ratio with respect to Repayer and defaulter is 11.39/1 which signifies that 91.93 % people repays the loan on time and only 8.07% people faces difficulty in repaying the loan.

UNIVARIATE ANALYSIS- DATASET 1

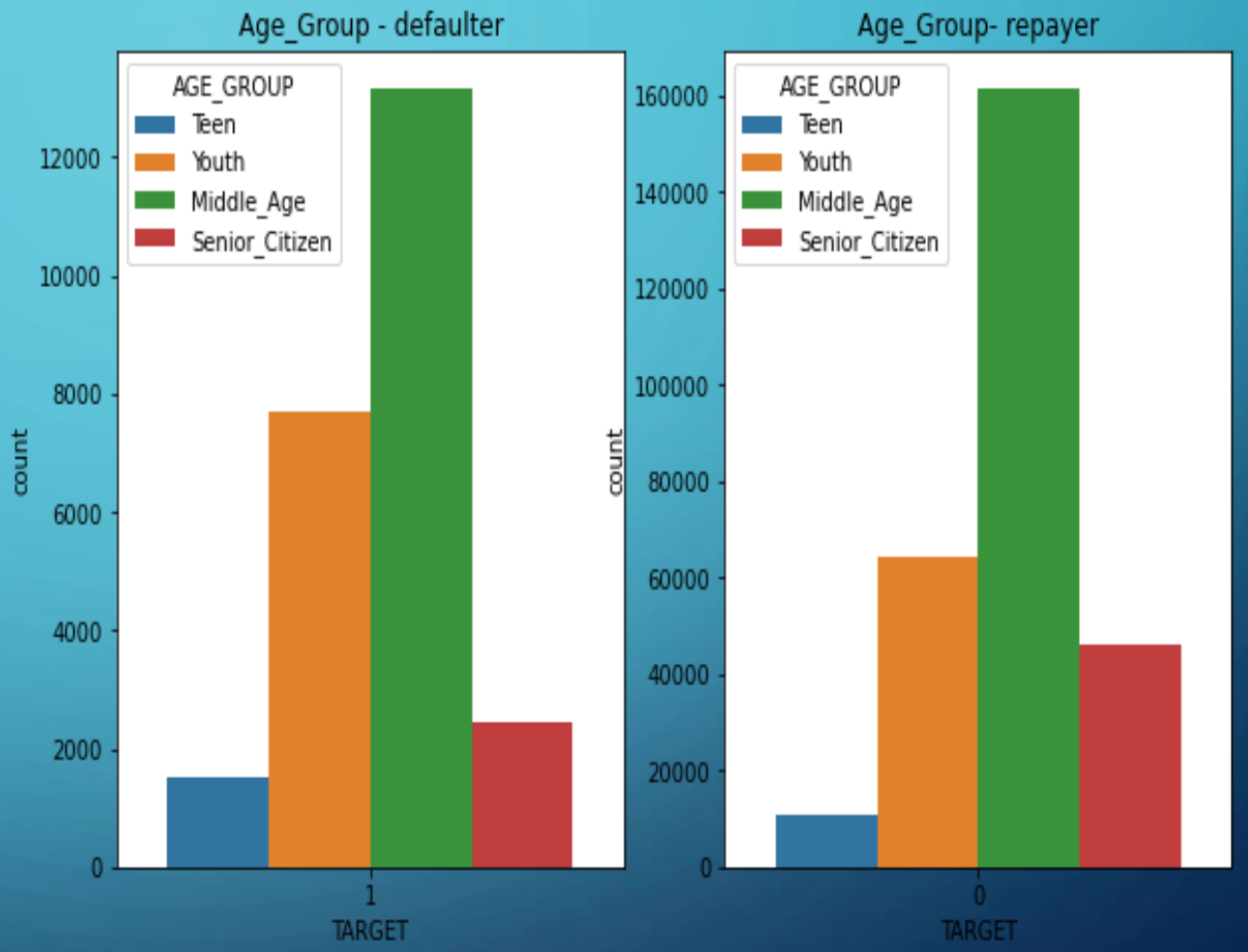
Univariate Analysis to understand different aspects such as who is more likely to pay the loans and which will default . Also , which section of customer applies for loan mostly .

Here, we can see that Female applies for loan most and they are likely to repay the loan amount.



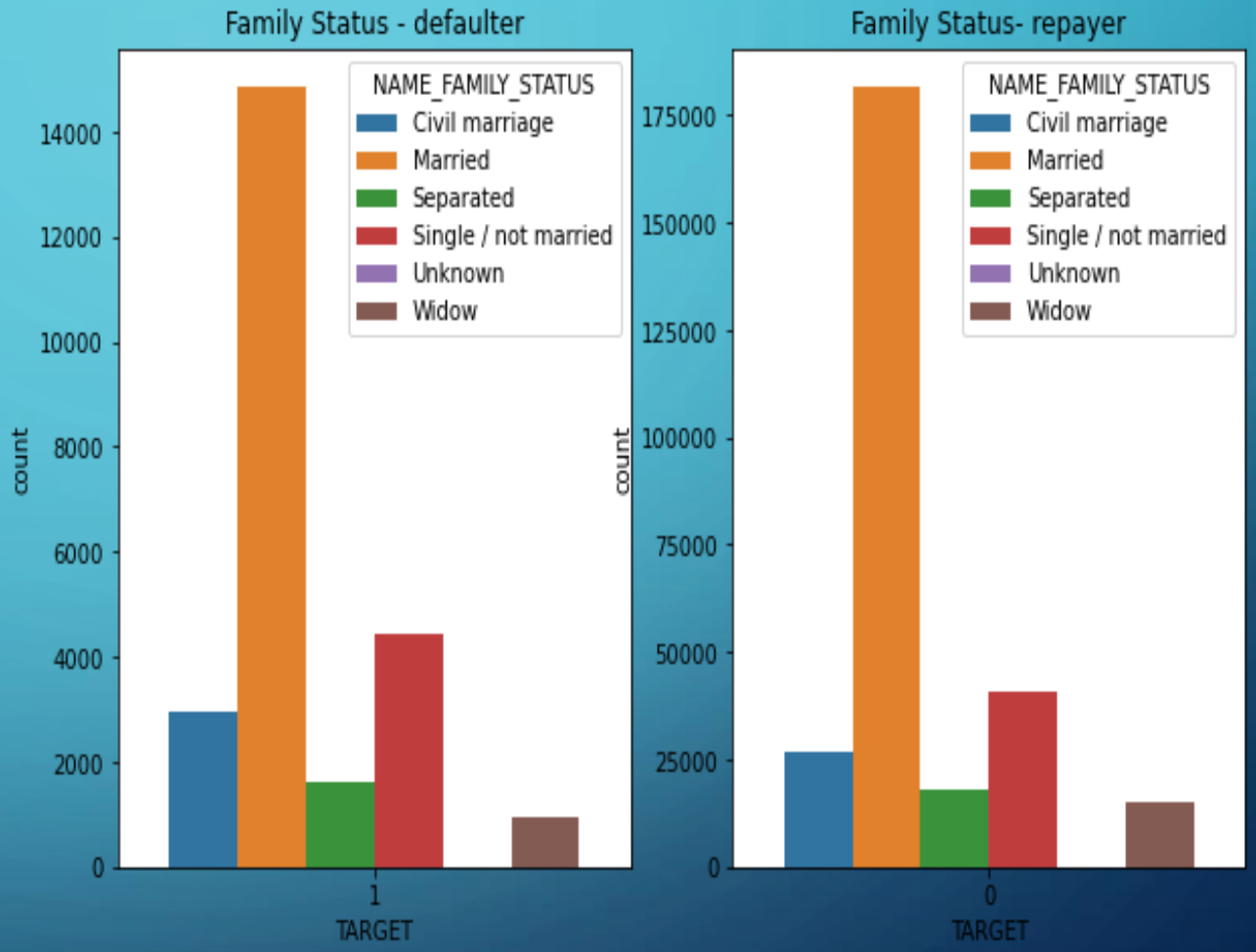
UNIVARIATE ANALYSIS-DATASET 1

Here, we can see that Middle age group people applies for loan most and they are likely to repay the loan amount.



UNIVARIATE ANALYSIS-DATASET 1

Here, we can see that Married people applies for loan most and they are likely to repay the loan amount.



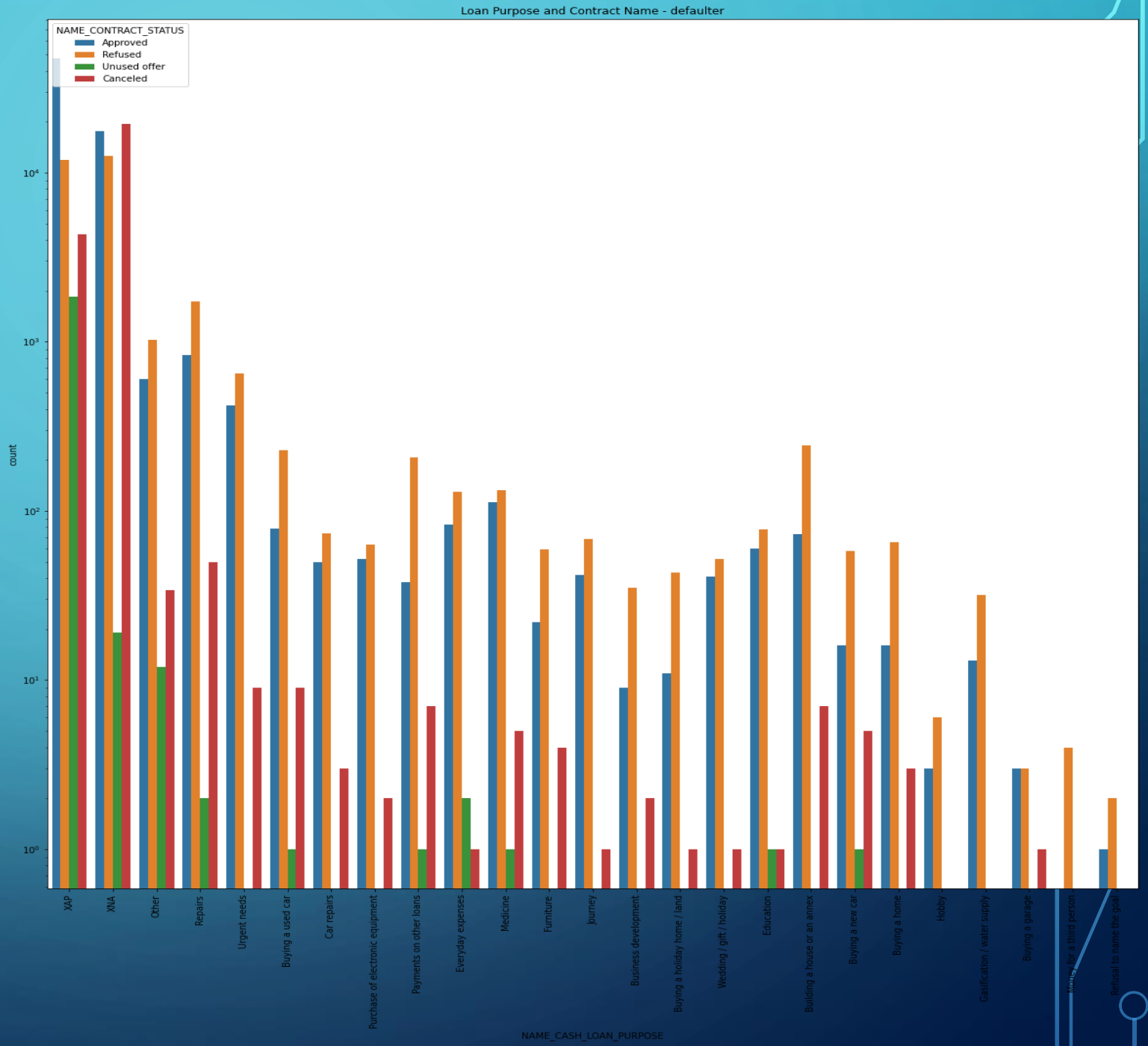
UNIVARIATE ANALYSIS-DATASET 2

1)Unknown reason (XNA ,XNP) are higher for both defaulter and repayer.

2)Repairing , Buying used car and Urgent needs are the top reason for applying loans.

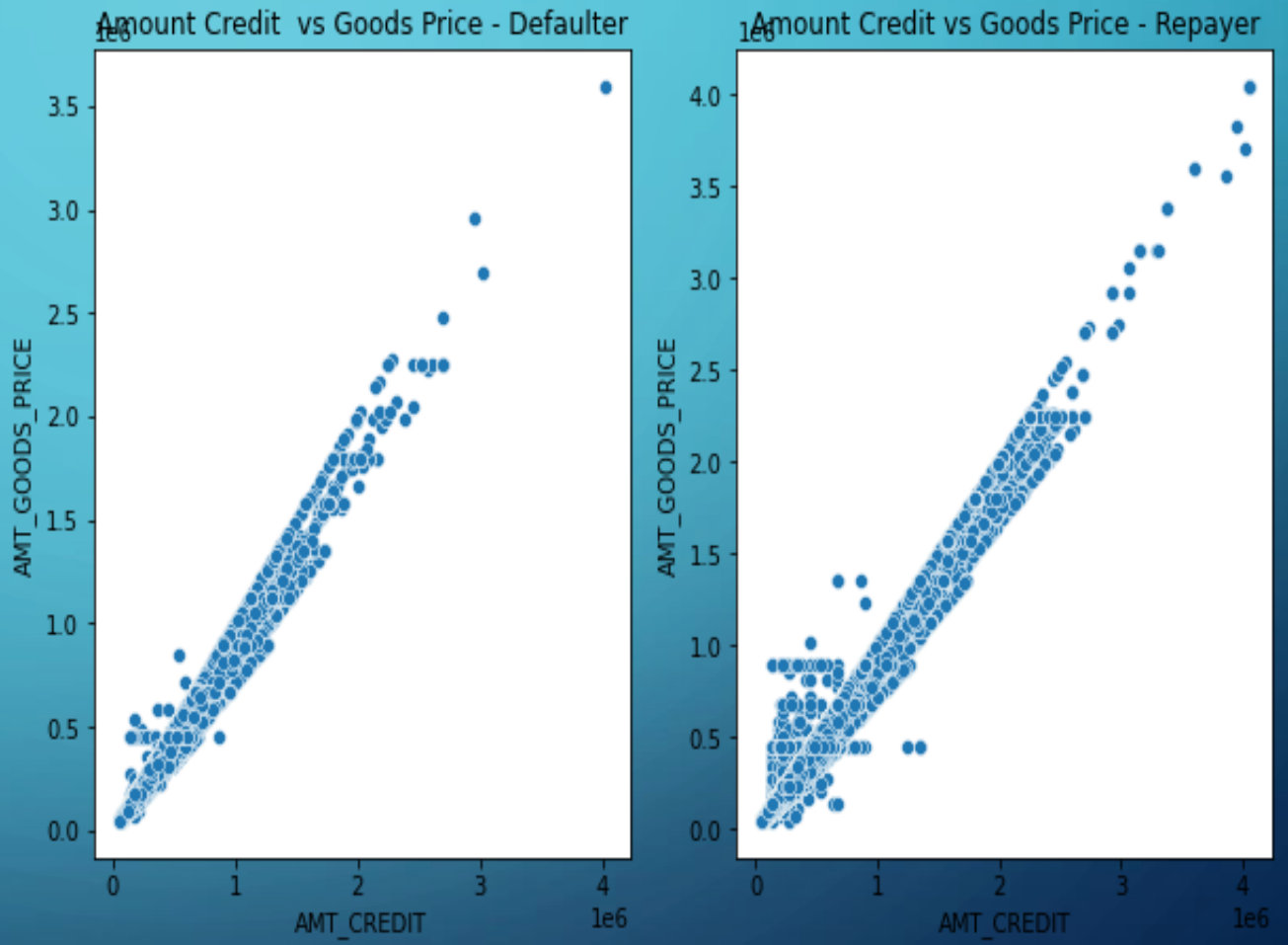
3)Loan applied with "repairing" reason are mostly rejected followed by loan applied with reason "other" and "building house". Hence, these reasons are considered risky by banks.

3)Buying a garage, Hobby, Money for third purpose are least reason for applying loans.



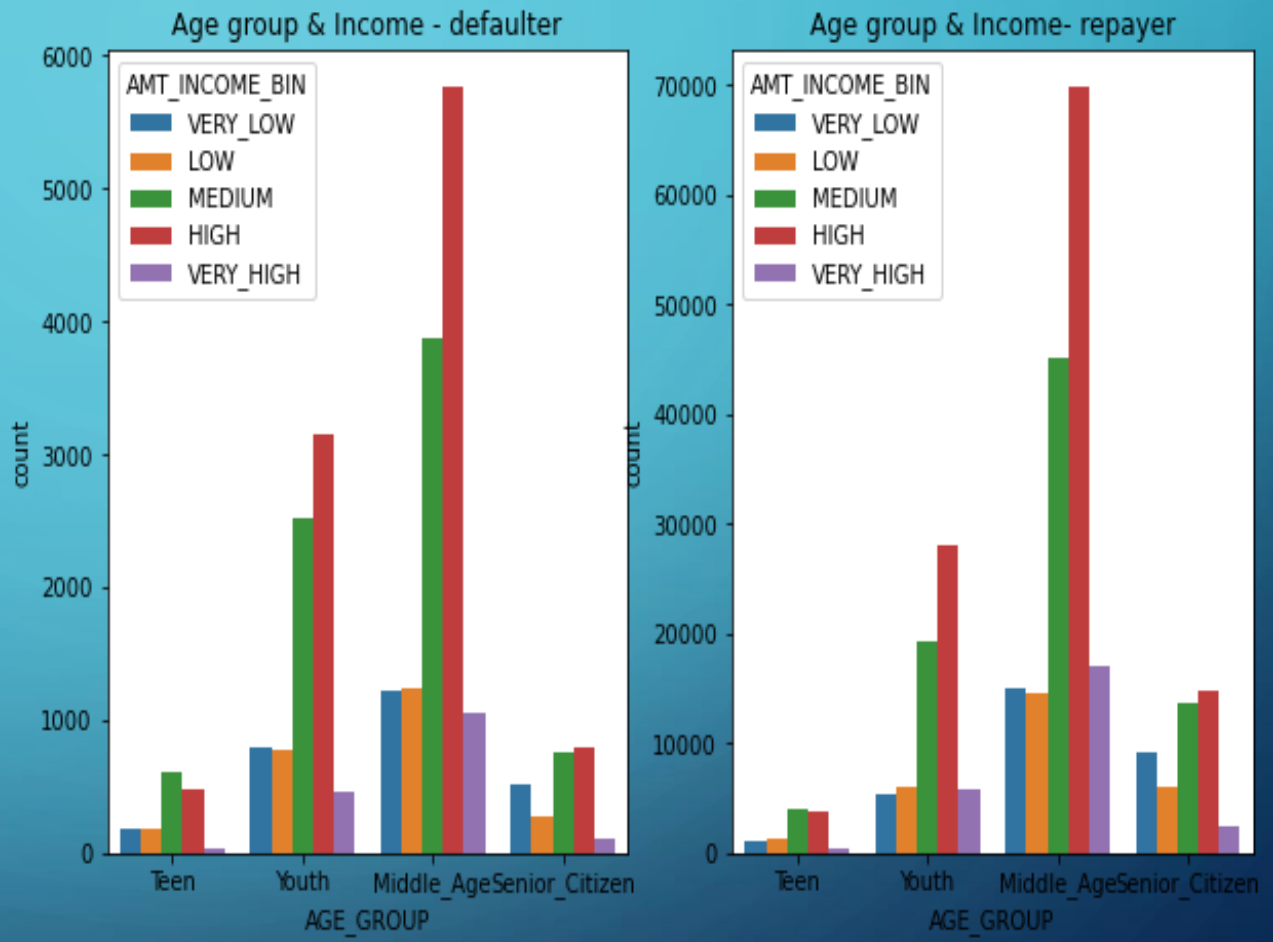
BIVARIATE ANALYSIS- DATASET 1

Here, we can see that people who repay loan on time have high chance of getting credit for goods.



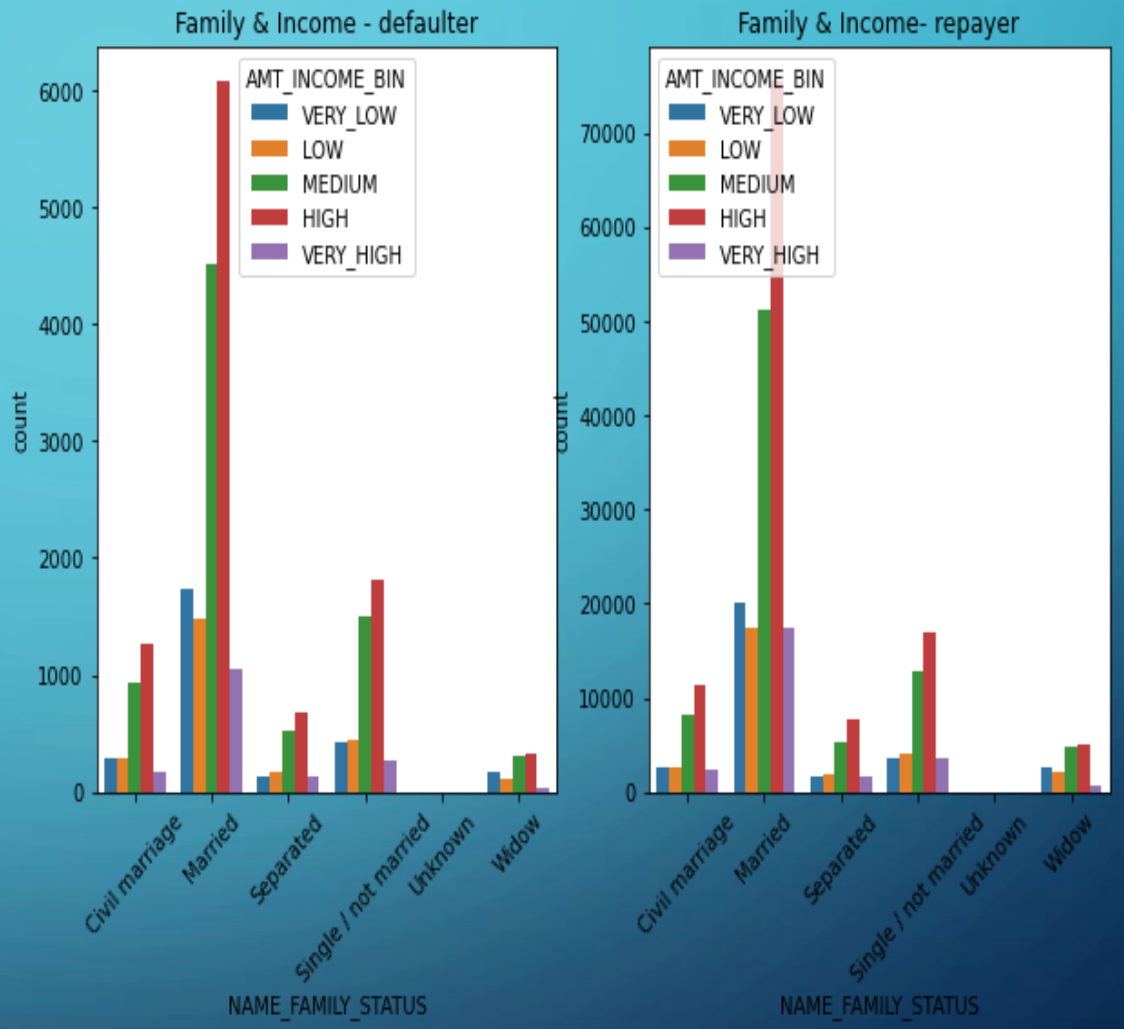
BIVARIATE ANALYSIS- DATASET 1

Here we can see that the best target people will be senior citizen with high income range.



BIVARIATE ANALYSIS- DATASET 1

Here we can see that married people with high income range is best for loan sanction.



CORRELATION- DATASET 1

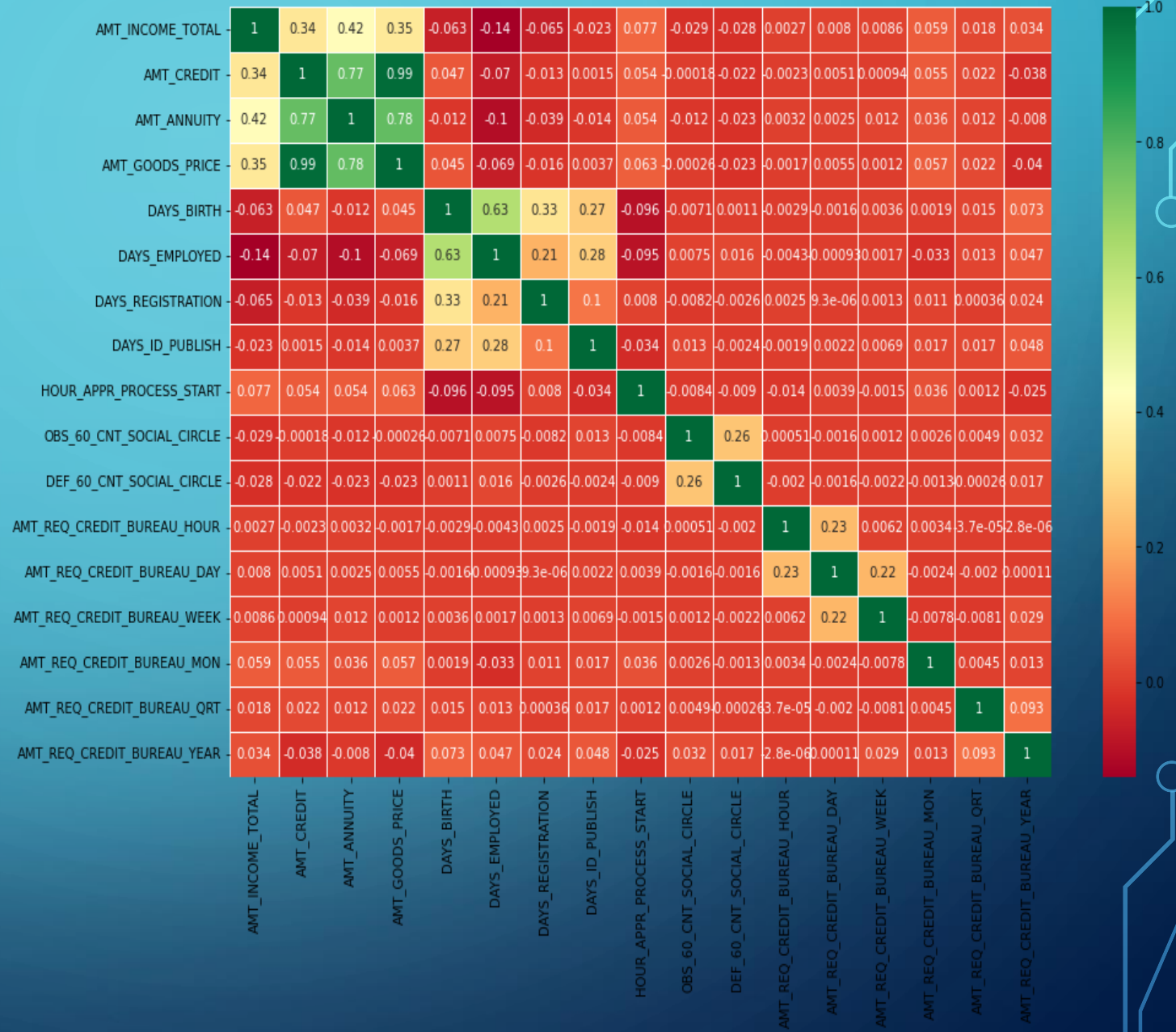
1)Here Credit amount ,Goods Price Amount, Loan Annuity are highly correlated.

2)Credit amount correlation with days employed has drastically reduced in defaulters(0.019) when compared to repayers(0.47).

3)There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among repayers.

4)Credit amount and Good price amount are highest corelated category.

5)We do correlation to find how different variable are corelated and then we can find causation.



CONCLUSION

- Revolving loan is the bank should offer more as people choose it over cash loans.
- People with academic degree has less defaults so bank should target this section.
- People with higher education repays so bank should target this section.
- Youth are most likely to default to bank can avoid them.
- Single person are also most likely to default so bank can avoid them.
- Senior citizen apply for loan most and they are more likely to repay if they have high income range.
- Person who has house/apartments are also more likely to apply for loan so bank should target them.
- Person with experience in job are less likely to default.

THANK YOU !!