PRESENTATION
ON
CREDIT EDA CASE STUDY

BY

YASHWANTH S

DS – 42 {MARCH BATCH}

PURPOSE

- credit risk analysis will help the company/bank to make a decision for loan approval based on the applicant’s profile. Which controls loss of business to the company and avoid financial loss for the company/bank.

STEPS

- ➢ Data understanding and sourcing
- ➢ Check for data quality issues and Binning
- ➢ Check for data imbalance and univariate, segmented univariate and bivariate analysis, correlation
- ➢ Merging of application data with previous application data
- ➢ Data analysis by univariate, segmented univariate, bivariate analysis, and correlation
- ➢ Recommendation and risks

DATA QUALITY ISSUES AND BINNING

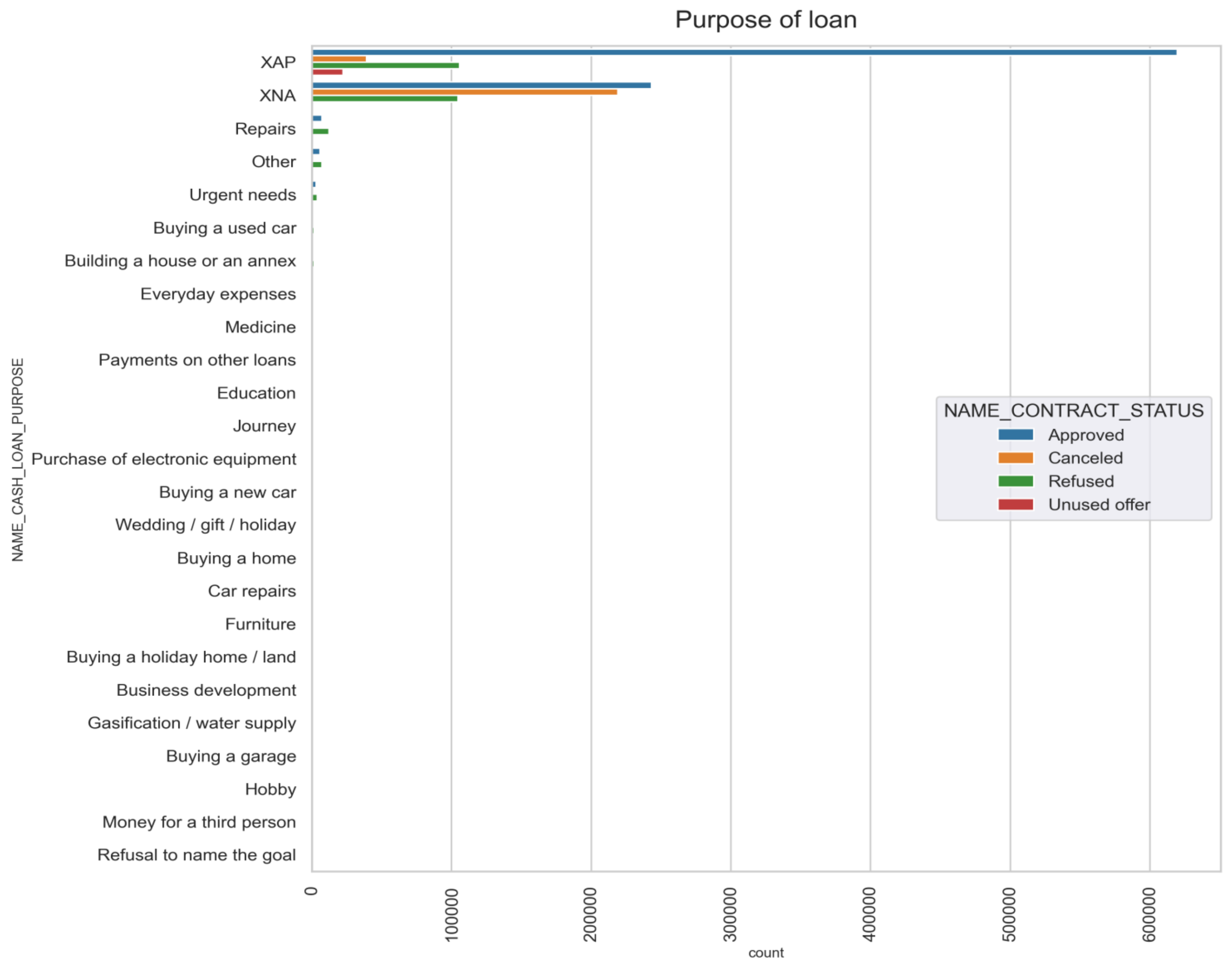
Data quality issues can stem from duplicate data, unstructured data, incomplete data, different data formats, or difficulty accessing the data.

Binning can improve model quality by strengthening the relationship between attributes. Supervised binning is a form of intelligent binning in which an important way to group a number of more or less continuous values into a smaller number of "bins". For example, if you have data about a group of people, you might want to arrange their ages into a smaller number of age intervals.

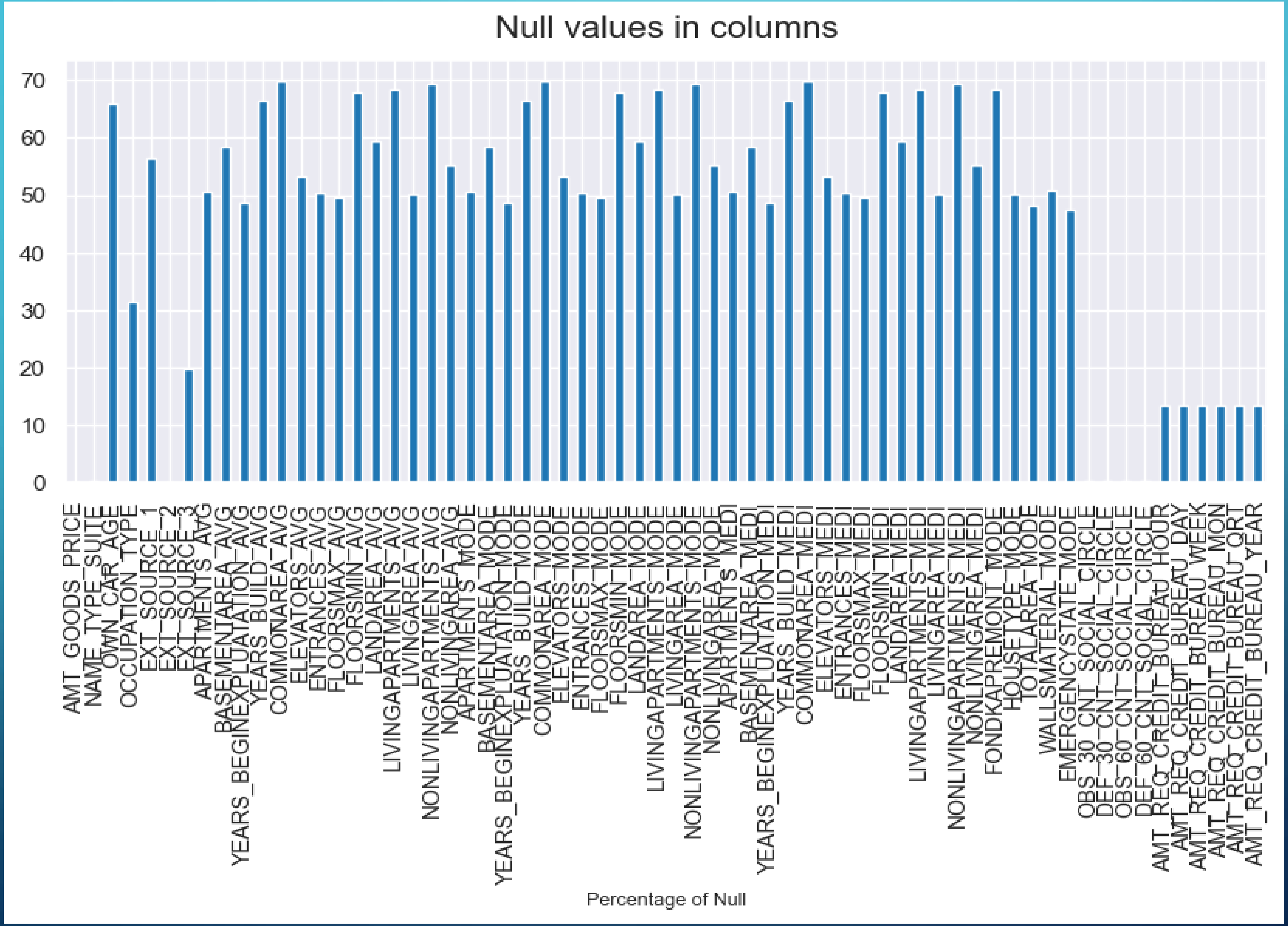
BUSINESS UNDERSTANDING

Business Understanding Performing EDA operation on the given dataset for risk analytics in banking and financial services and understanding how data is used to minimize the risk of losing money while lending to customers.

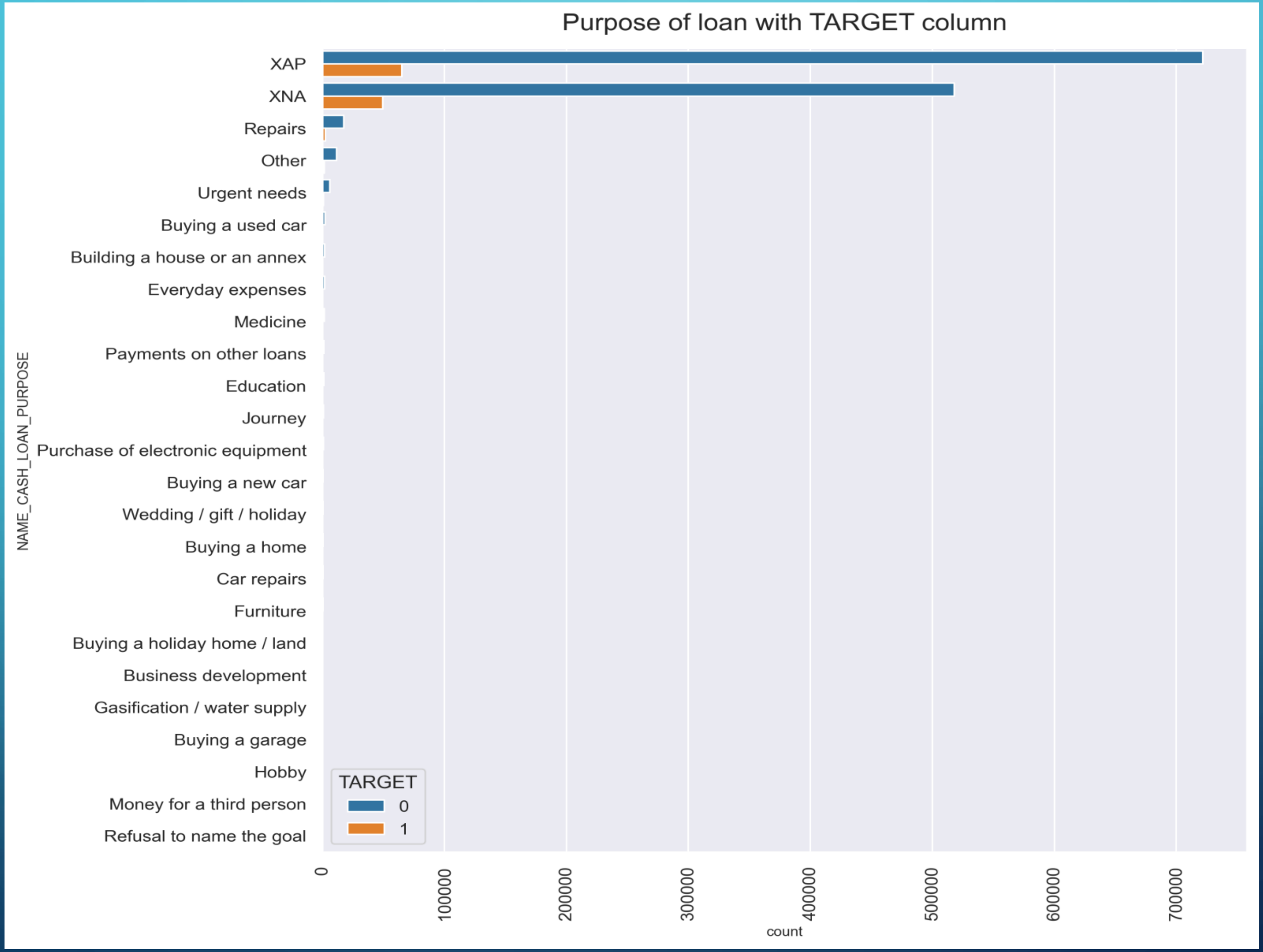
Performing univariate analysis
Purpose of loan



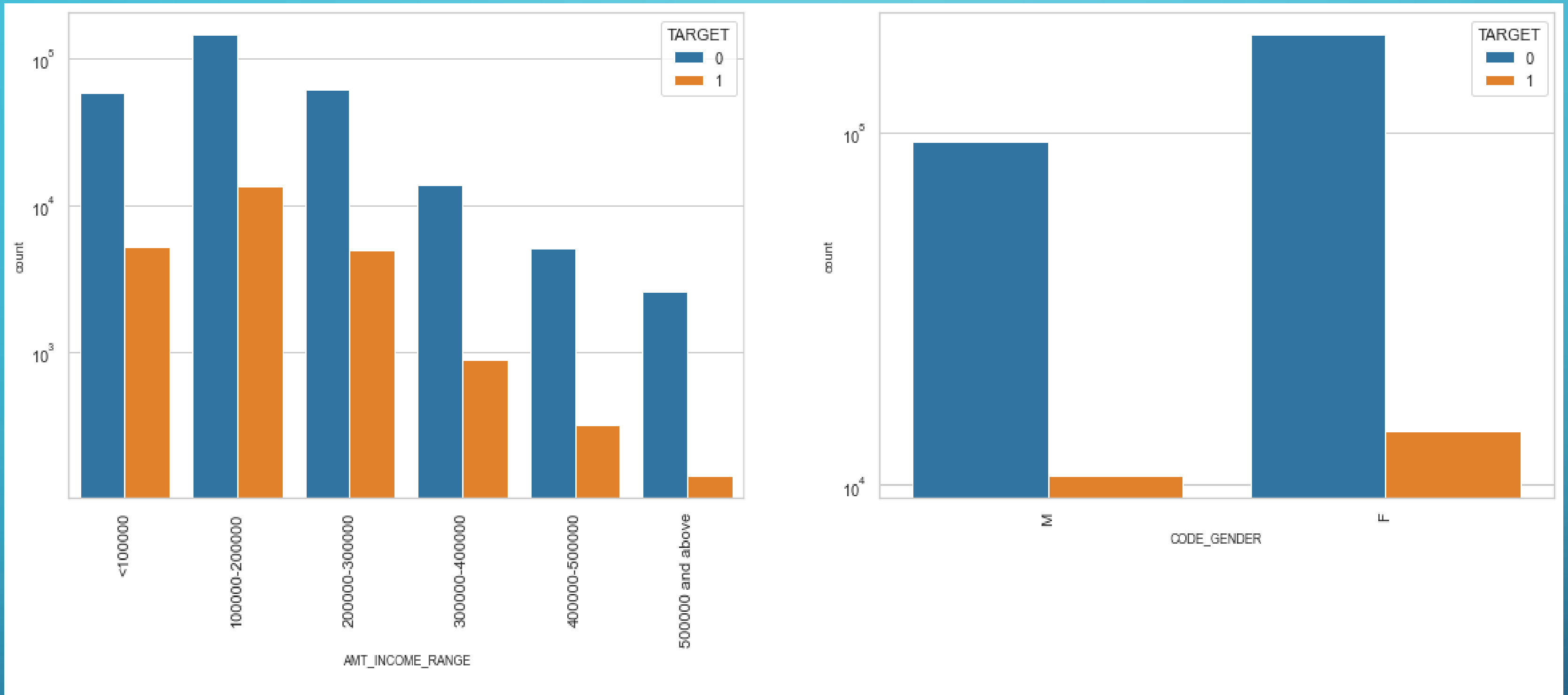
Visualizing Null values of columns in graph



Purpose of the loan with TARGET column



UNIVARIATE ANALYSIS OF CONTINUOUS DATA IN APPLICATION DATA

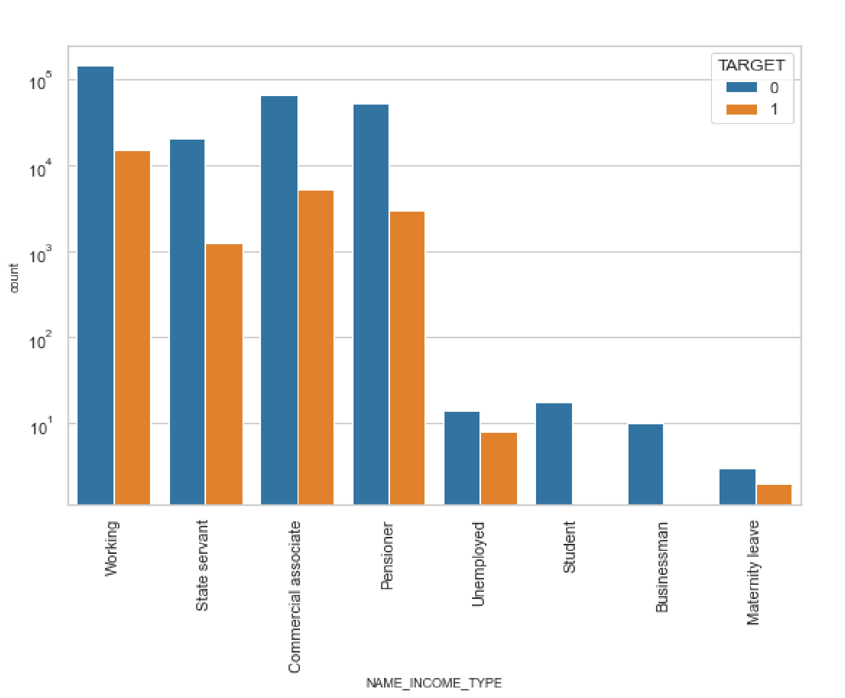


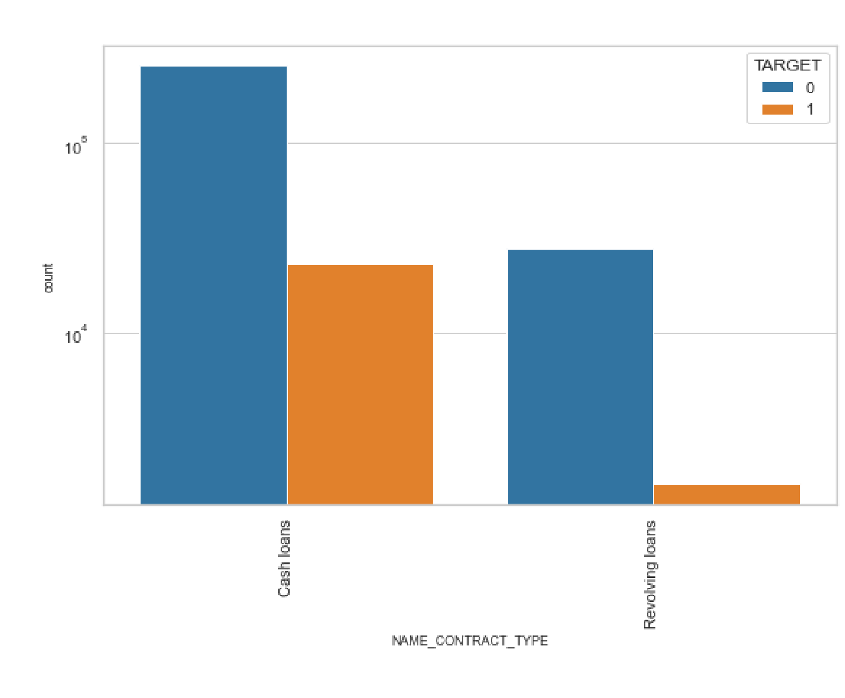
#Observations : AMT_INCOME_RANGE :

- People in the range 100000-200000 have high number of loan and also have a high in defaulter
- Income segment >500000 has less defaulter.

CODE_GENDER:

- The % of defaulters are more in Male than Female





NAME_INCOME_

- Student and business are higher in percentage of loan repayment.
- Working, State servant and Commercial associates are higher in default percentage.
- Maternity category is significantly higher problem in repayment.

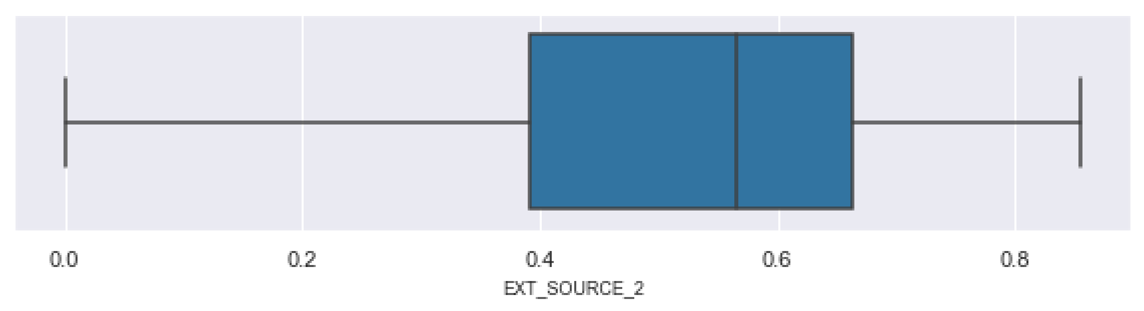
NAME_CONTRACT_TYPE

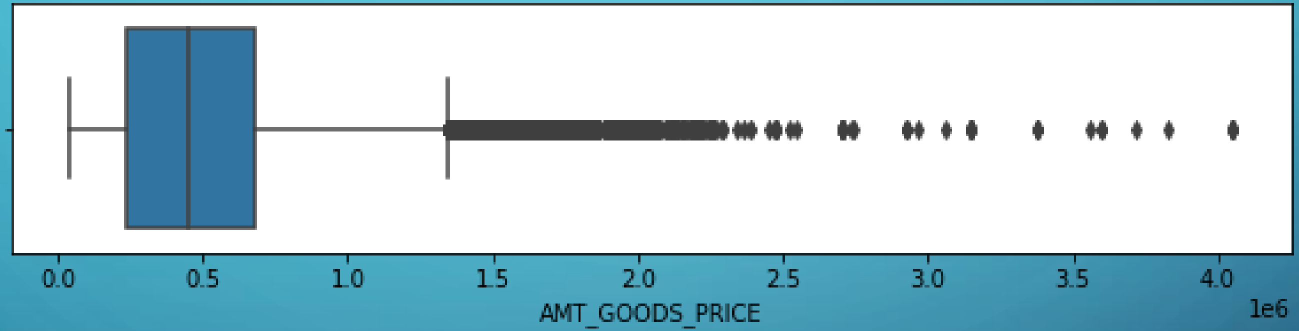
- For contract type ‘Cash loans’ are high in number of credits than ‘Revolving loans’ contract type.
- By above graph 'Revolving loans' is small amount compared to ‘Cash loans’

Correlation among the Continuous variable

The correlation coefficient is a measure of the degree of linear association between two continuous variables, i.e. when plotted together, how close to a straight line is the scatter of points.

Continuous variable





Observation from Boxplots:

For 'EXT_SOURCE_2' no outliers present. So data is rightly present. For the 'AMT_GOODS_PRICE' outlier present in the data. so need to impute with median value: 4

The conclusion of the Analysis

- Banks must target more on contract types ‘Student’, ’Pensioner’ and ‘Businessman’ for profitable business.

- Banks must focus less on income type ‘Working’ as it has the most number of unsuccessful payments in order to get rid of financial loss for the organization.