CREDIT EDA CASE STUDY

PRAVEER TIWARI

PRERNA RAVIRAJ

Problem statement- To identify consumer attributes and loan attributes that indicate the tendency to default and identify patterns that lead to this default. This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Analysis approach- Exploratory Data Analysis is used to analyze various factors that lead to loan default and identify the biggest contributors.

DATA CONDITIONING:

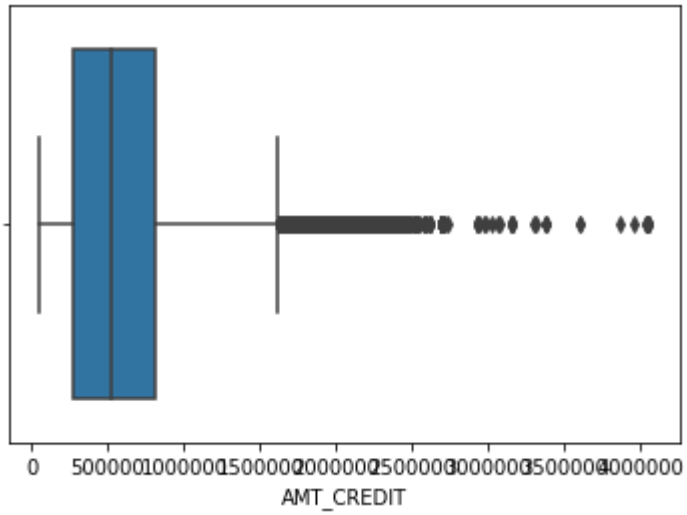
- Uploading the necessary files on jupyter notebook.
- Search for any attributes which have very less data and removing it.
- Finding and removing outliers by using various methods such as filtering data by adding a threshold, substituting by mean if the data field is numerical.
- For empty cells substituting by mean if the data field is numerical.

Types of analysis used:

Univariate, segmented univariate and bivariate analysis along with correlation are the methods adopted to do the analysis. This is done so that banks can make informed choices while accepting and rejecting applicants in order to identify risks and minimize losses. Plots like histograms scatter plots , heat plots are being used for comparision

UNIVARIATE ANALYSIS

1) Amount credited



Box plot analysis:

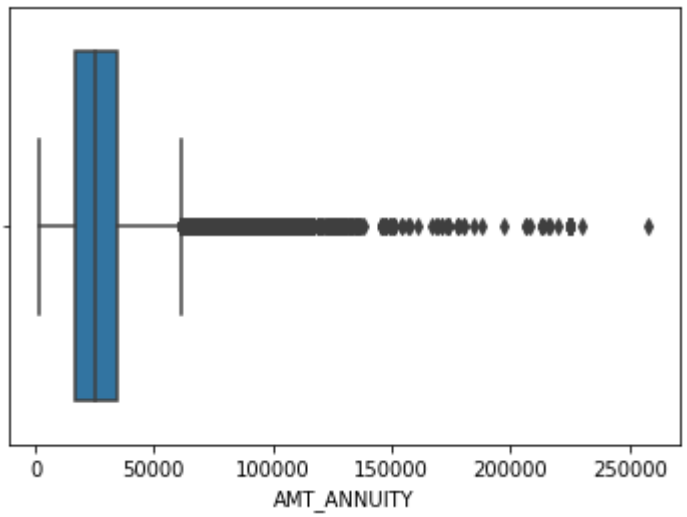
a) Minimum value=0
b) 25 percentile=250000
c) Median=500000
d) 75 percentile=750000

Presence of many outliers also detected

we can observe in the above plot that the most common credit amount falls in the range 250000-750000.

This gives us the major part of applicants.

2) Annuity

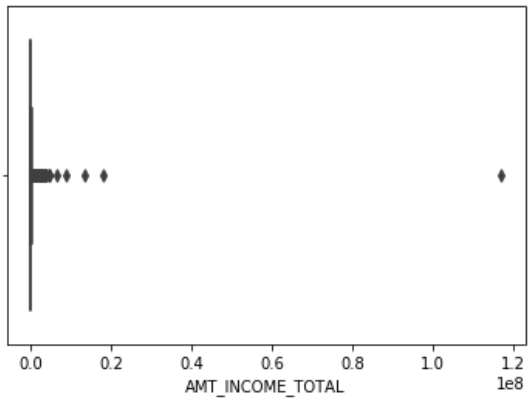


Box plot analysis:

a) Minimum value=0
b) Median=25000
c) 75 percentile=750000

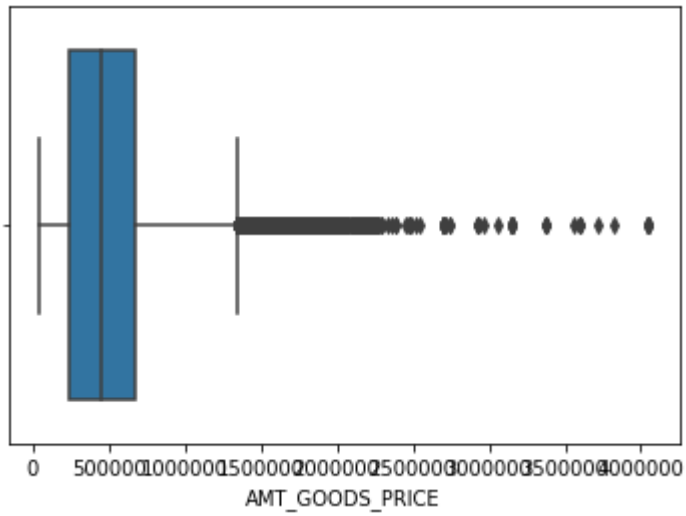
Presence of many outliers detected. Outliers are very spread out.

3) Income total



In this box plot we see that there is an extreme outlier at 1.2 and the rest of the values are very concentrated near 0.

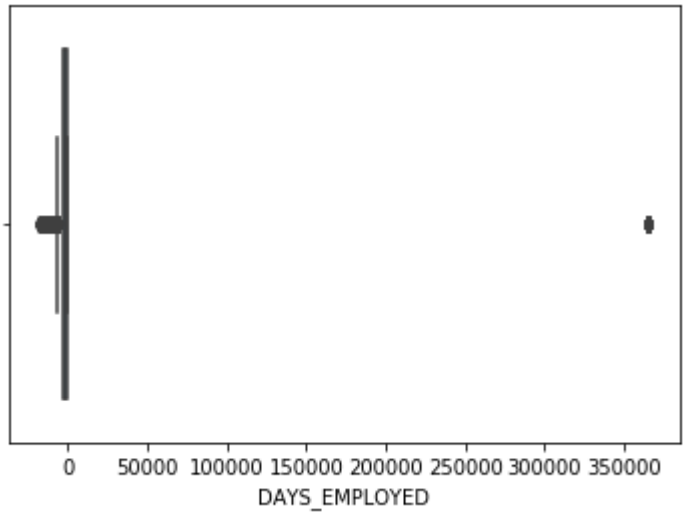
4) Goods price



e) Minimum value=0
f) 25 percentile=250000
g) Median=500000
h) 75 percentile=750000

Presence of many outliers also detected

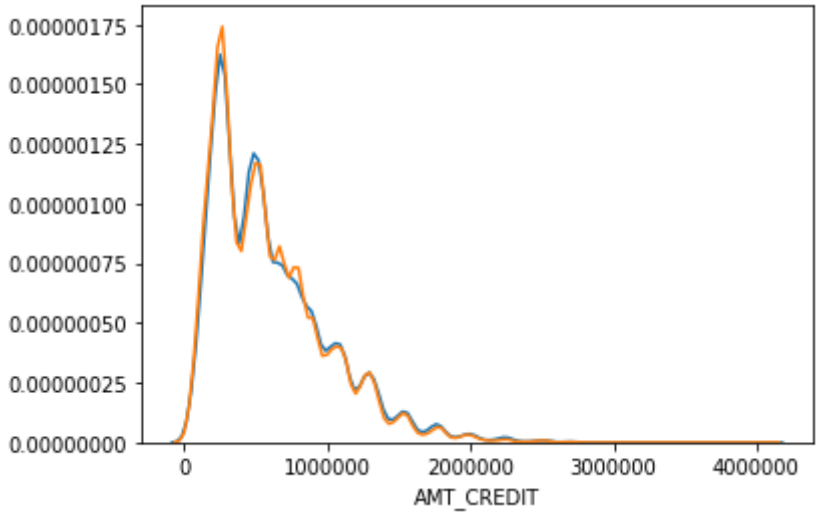
5) Days employed



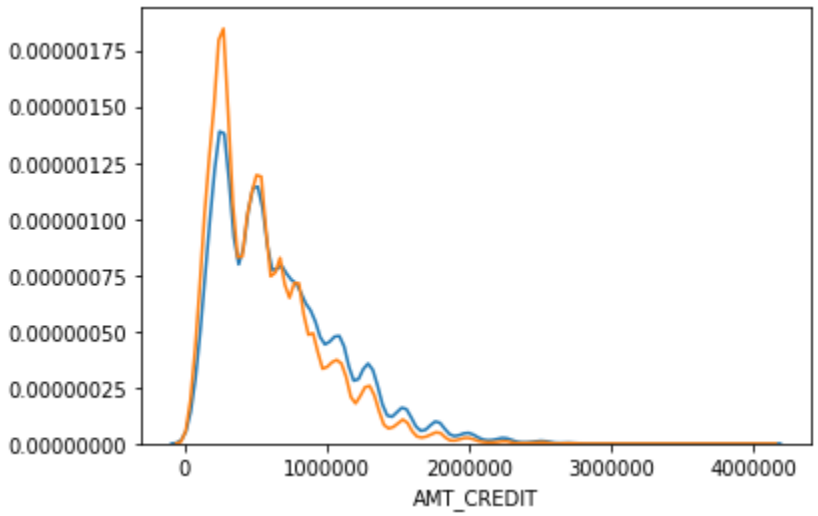
This box plot shows us that almost all values lie in a small concentrated area. There is only one outlier which should be dropped while doing the analysis.

SEGMENTED UNIVARIATE ANALYSIS

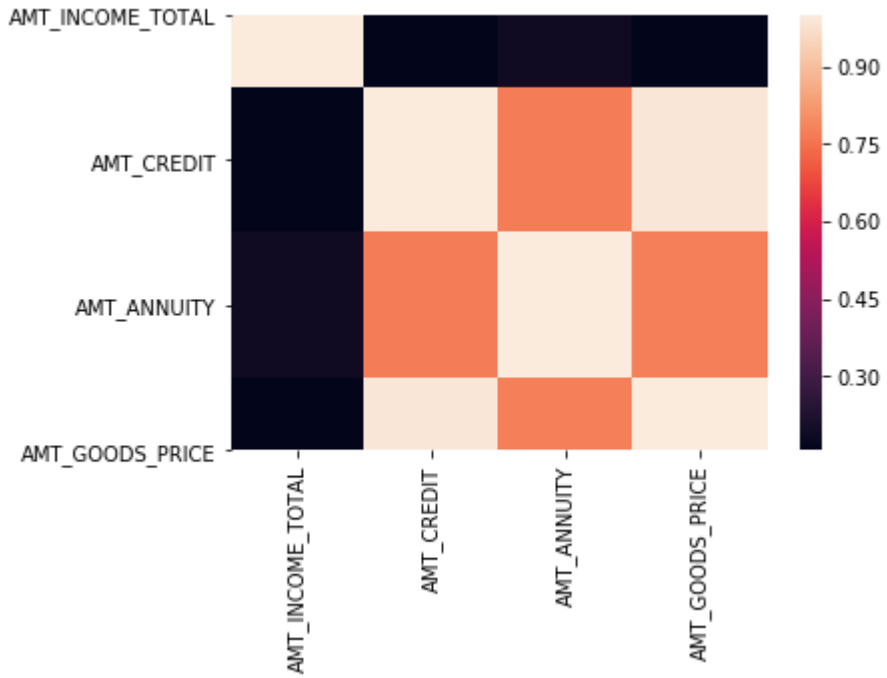
1) Relationship between amount credited and gender



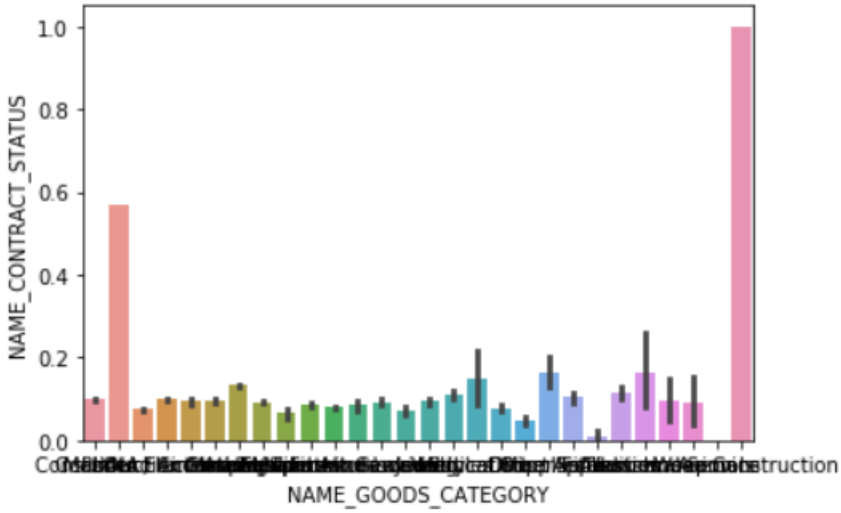
2) Relationship between car ownership and amount credited

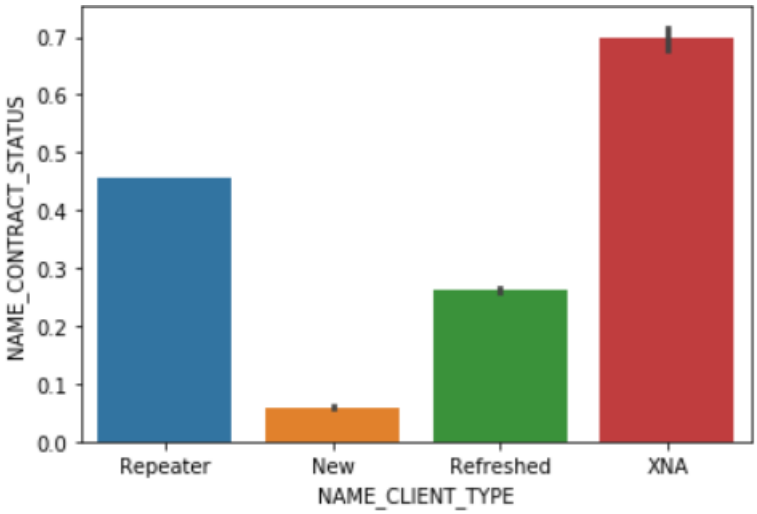


BIVARIATE ANALYSIS



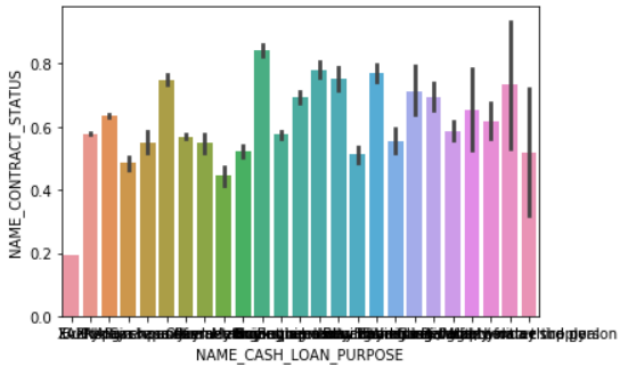
We can see that the variables goods price and amount credit have high values of correlation. Highest among all the four variables. The value of correlation between the other variables are almost equal and similar to each other

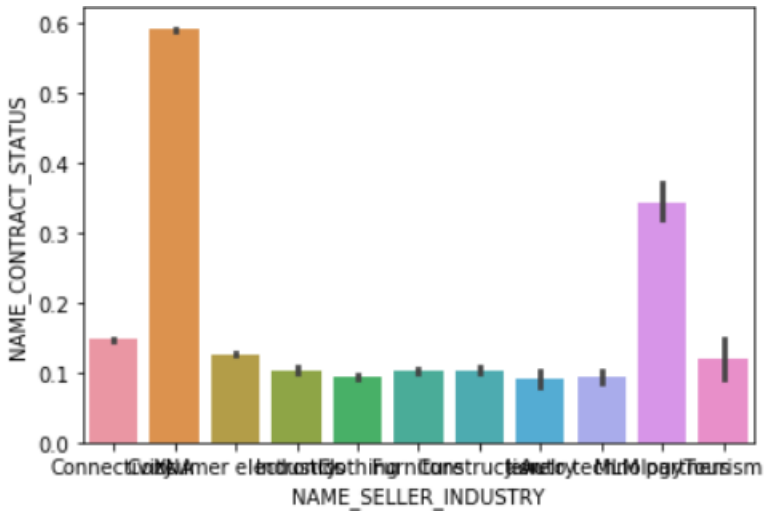




From the above plot it is observed that highest approved loans are not from a

Particular category xna, Repeaters are quite trusted by the bank and hence high approvals are there to them

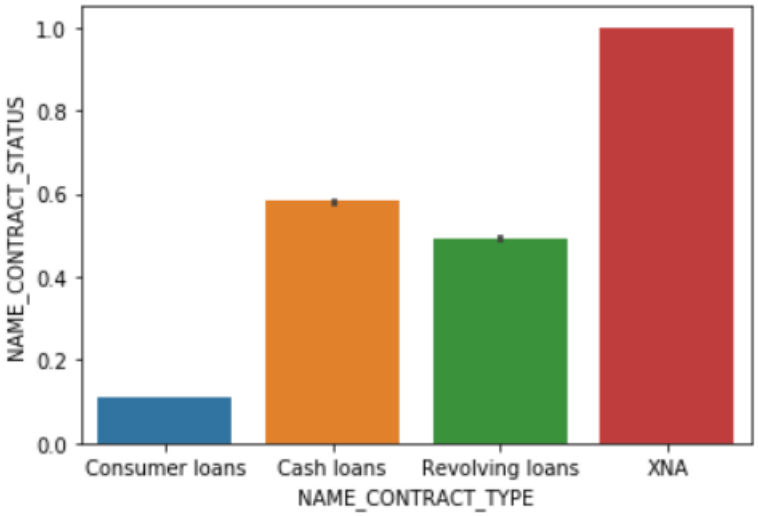




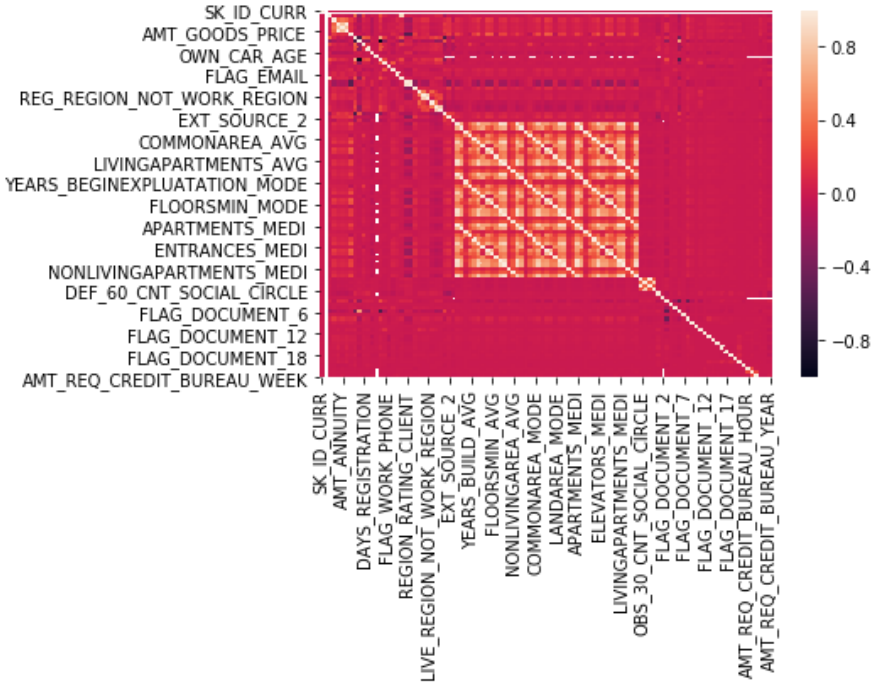
Electronics industries gets highest loans approval because of the current tech boom going on.

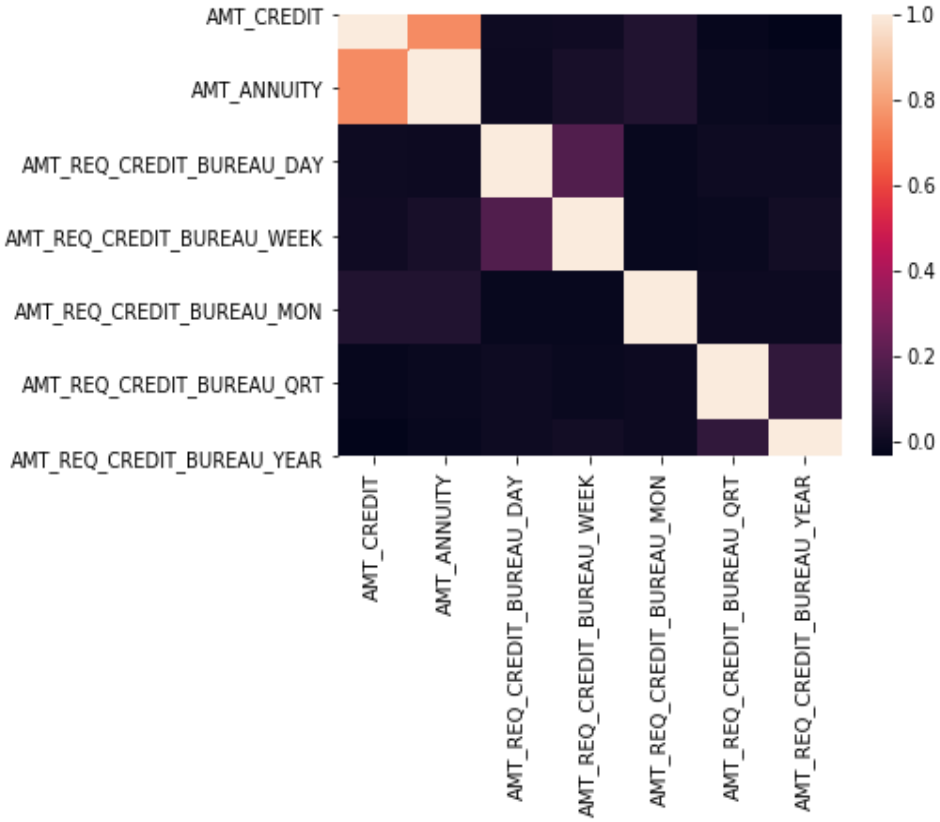
	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT
AMT_ANNUITY	1.000000	0.808856	0.816414	0.267749
AMT_APPLICATION	0.808856	1.000000	0.975765	0.482427
AMT_CREDIT	0.816414	0.975765	1.000000	0.299967
AMT_DOWN_PAYMENT	0.267749	0.482427	0.299967	1.000000

Credit and annuity are high in correlation . Also amount for application and credit amount are high in correlation.

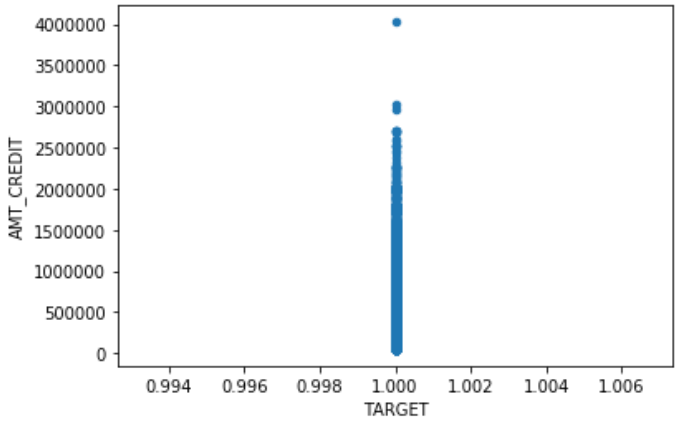


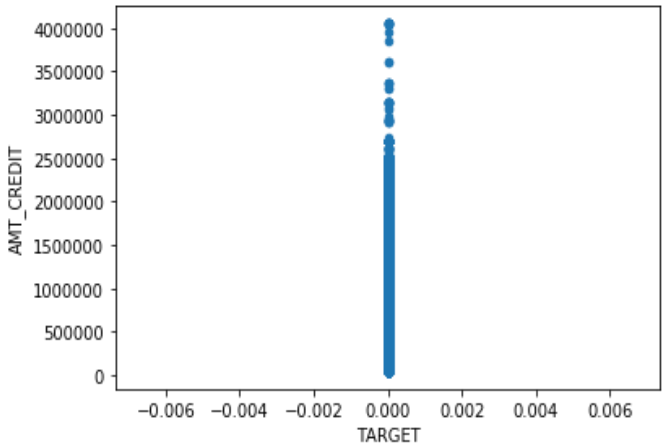
Cash loans are highest because of liquidity It is easy to give loans and get it back in cash.Consumer loans are least.

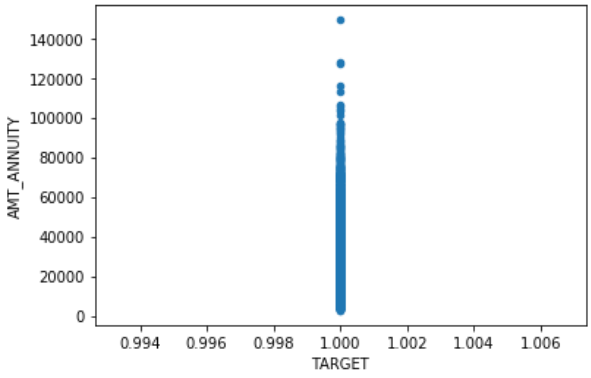

SK_ID_CURR
AMT_GOODS_PRICE
OWN_CAR_AGE
FLAG_EMAIL
REG_REGION_NOT_WORK_REGION
EXT_SOURCE_2
COMMONAREA_AVG
LIVINGAPARTMENTS_AVG
YEARS_BEGINEXPLUATATION_MODE
FLOORSMIN_MODE
APARTMENTS_MEDI
ENTRANCES_MEDI
NONLIVINGAPARTMENTS_MEDI
DEF_60_CNT_SOCIAL_CIRCLE
FLAG_DOCUMENT_6
FLAG_DOCUMENT_12
FLAG_DOCUMENT_18
AMT_REQ_CREDIT_BUREAU_WEEK
SK_ID_CURR
AMT_ANNUITY
DAYS_REGISTRATION
FLAG_WORK_PHONE
REGION_RATING_CLIENT
LIVE_REGION_NOT_WORK_REGION
EXT_SOURCE_2
YEARS_BUILD_AVG
FLOORSMIN_AVG
NONLIVINGAREA_AVG
COMMONAREA_MODE
LANDAREA_MODE
APARTMENTS_MEDI
ELEVATORS_MEDI
LIVINGAPARTMENTS_MEDI
OBS_30_CNT_SOCIAL_CIRCLE
FLAG_DOCUMENT_2
FLAG_DOCUMENT_7
FLAG_DOCUMENT_12
FLAG_DOCUMENT_17
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_YEAR
0.8
0.4
0.0
−0.4
−0.8

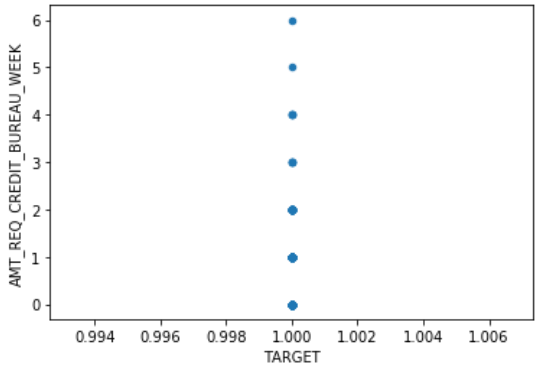

AMT_CREDIT
AMT_ANNUITY
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
AMT_CREDIT
AMT_ANNUITY
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
1.0
0.8
0.6
0.4
0.2
0.0

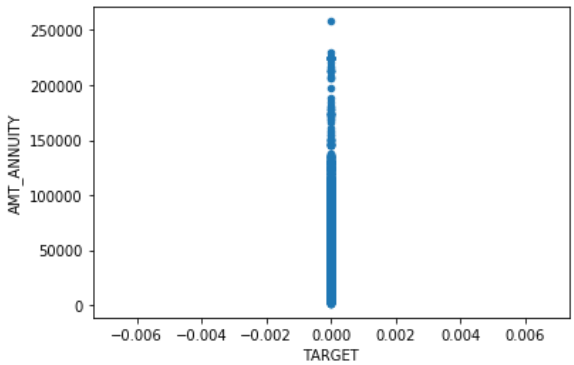
Target 0 and Target 1 comaprisions

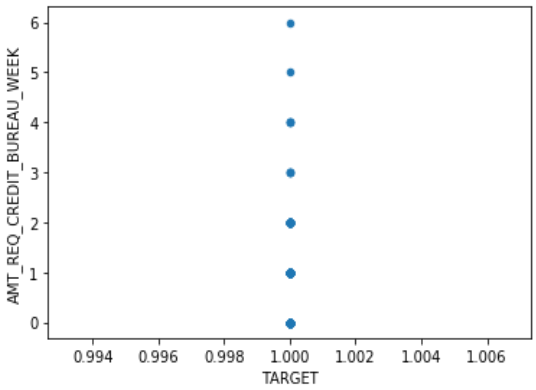


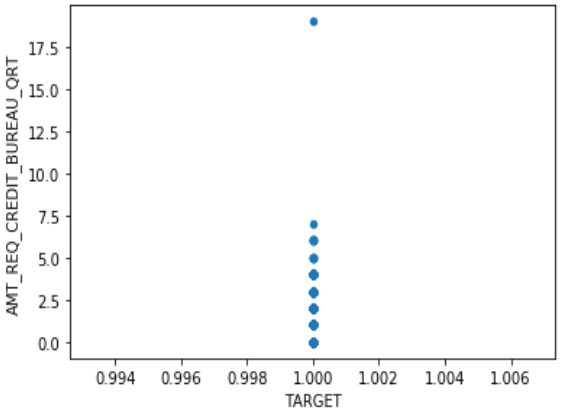


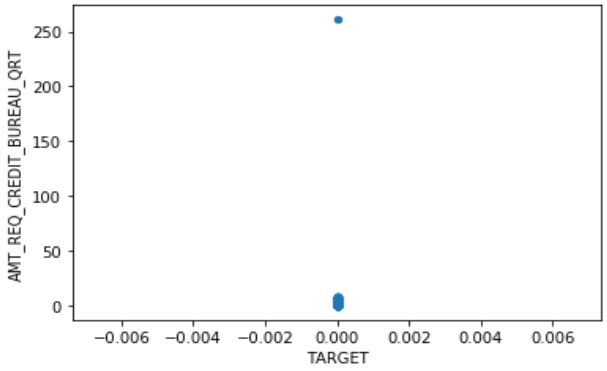


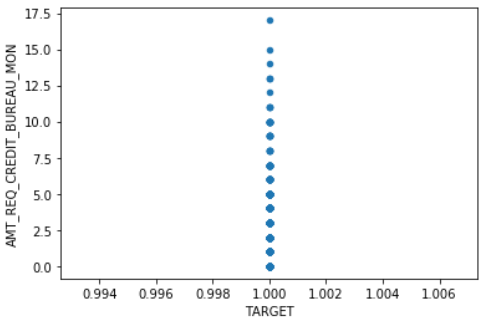


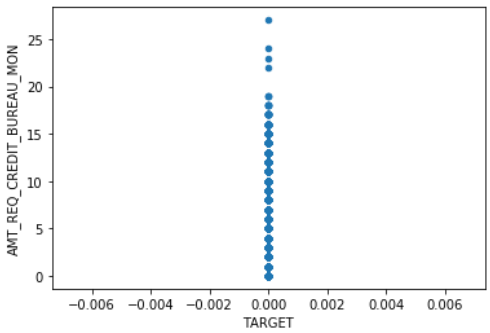



AMT_REQ_CREDIT_BUREAU_WEEK
TARGET


AMT_REQ_CREDIT_BUREAU_QRT
TARGET


AMT_REQ_CREDIT_BUREAU_QRT
TARGET


AMT_REQ_CREDIT_BUREAU_MON
TARGET


AMT_REQ_CREDIT_BUREAU_MON
TARGET

RESULT OF THE ANALYSIS AND CONCLUSION:

The loan defaulters are mostly from self-employed followed by industries .The male and female have same pattern of taking amount of loans.The electronics industries and type 3 businesses have highest loan approvals.

Even though the self employed population has bad approval history they have third highest applicants.

Once the loans are fully paid it is easier to get another one given in the analysis.

Loans credit amount lies in a majority of amount 250000-750000.The transaction preferred is mostly cash because of its ease availability.

Goods price have high correlation with annuity and credit amount because if the person owns a lot goods then it is very likely that he can return the amount taken.