DCS-52 : Data Science (January-2023)

Credit EDA Case Study

By : Digmakumari Tarunkumar Patel

Introduction

The assignment study is designed in a way to applied the knowledge of EDA into the real scenario. This study not only aims at applying learned methods and tricks into the analysis but also be helpful for understanding the banking sector for its risk analysis, finance to get a broad idea of risk associated to bank while lending money to its clients.

Business Statement

With inadequate or nonexistent credit histories, loan providers find it challenging to grant loans to individuals. Because of this, some customers take advantage of it by default. Imagine you work for a consumer finance firm that specializes in providing urban customers with several kinds of loans. To analyze the patterns found in the data, you must employ EDA. By doing this, it will be ensured that only those applicants who can repay the loan would be accepted.

Aim for EDA case study

1. 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.
2. In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

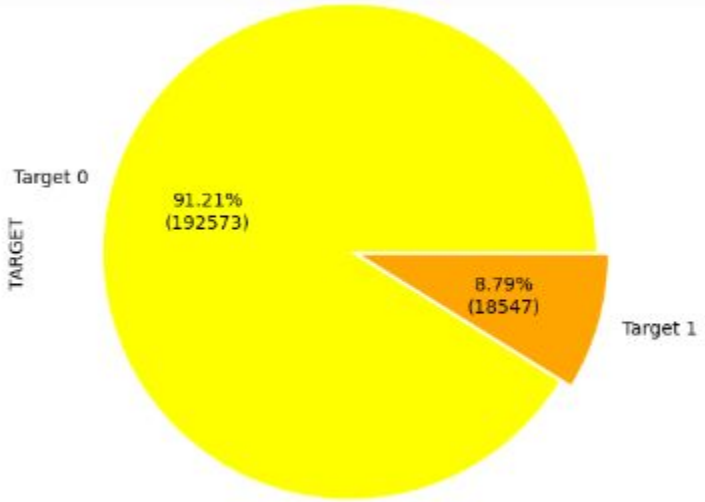
Data files

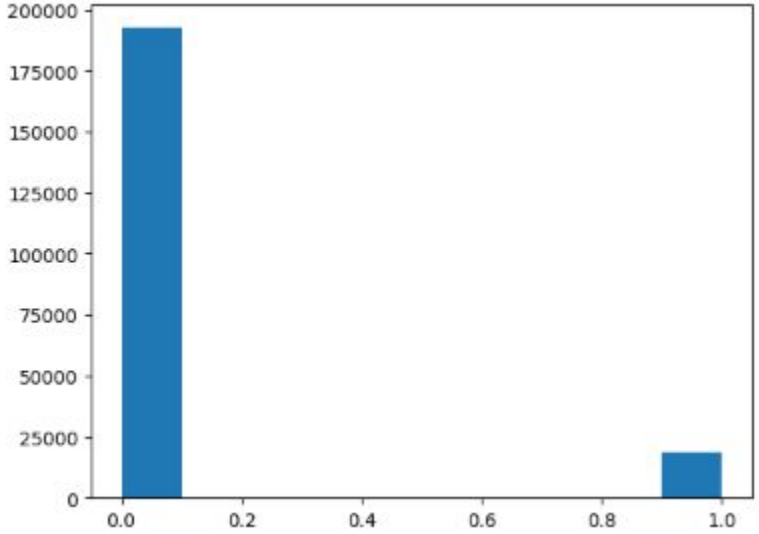
1. Application_Data.csv
2. Previous_data.csv

Procedure

1. Data cleaning : Removing unwanted columns or data, replacing values to appropriate substitute as mean , median or mode , locating outlier and filling the missing values
2. Analysing the cleaned data : Following analysis process of univariate or bivariate analysis , checking imbalance among the highlighted columns
3. Making conclusion and drawing insights

Analysing Imbalance





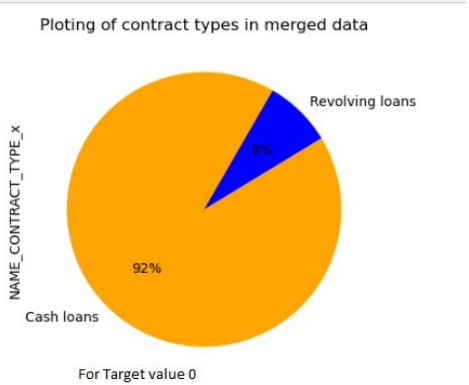
- Their is highly huge difference between targets 1 and 0
- Target 0 reflects the one with low repayment difficulties
- Target 1 reflects the one with high payment difficulties

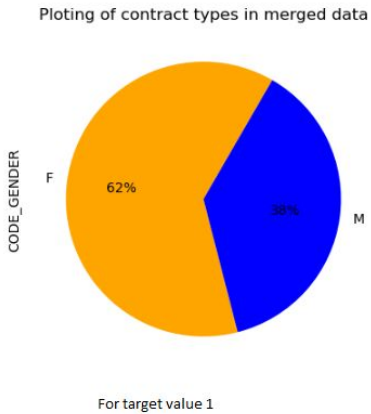
Univariate Analysis for category variable

Analysis included data related to :

1. NAME_CONTRACT_TYPE
2. CODE_GENDER
3. NAME_EDUCATION_TYPE
4. NAME_FAMILY_STATUS
5. CNT_CHILDREN

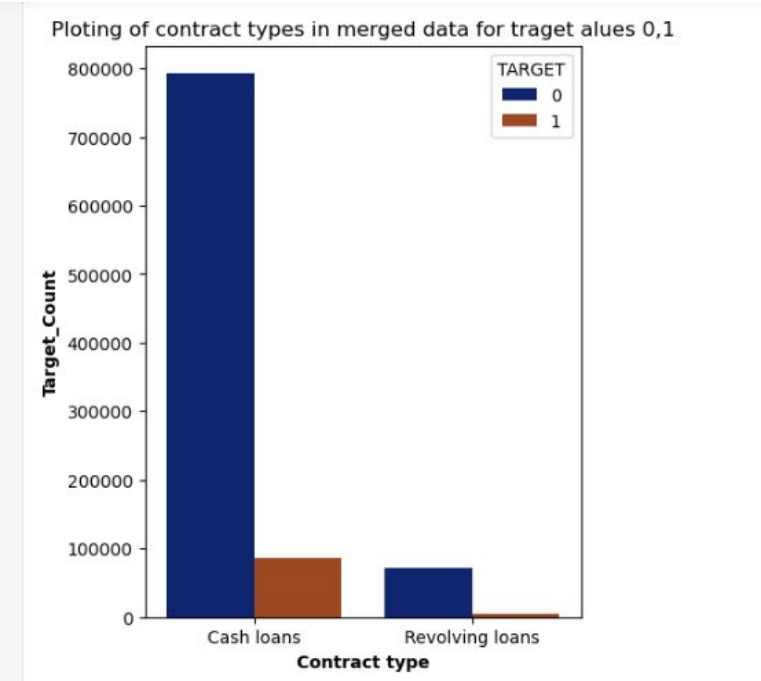
1. NAME_CONTRACT_TYPE





Both set of client consider cash loans over revolving loans ; Moreover by target 0 high preference is cash loans.

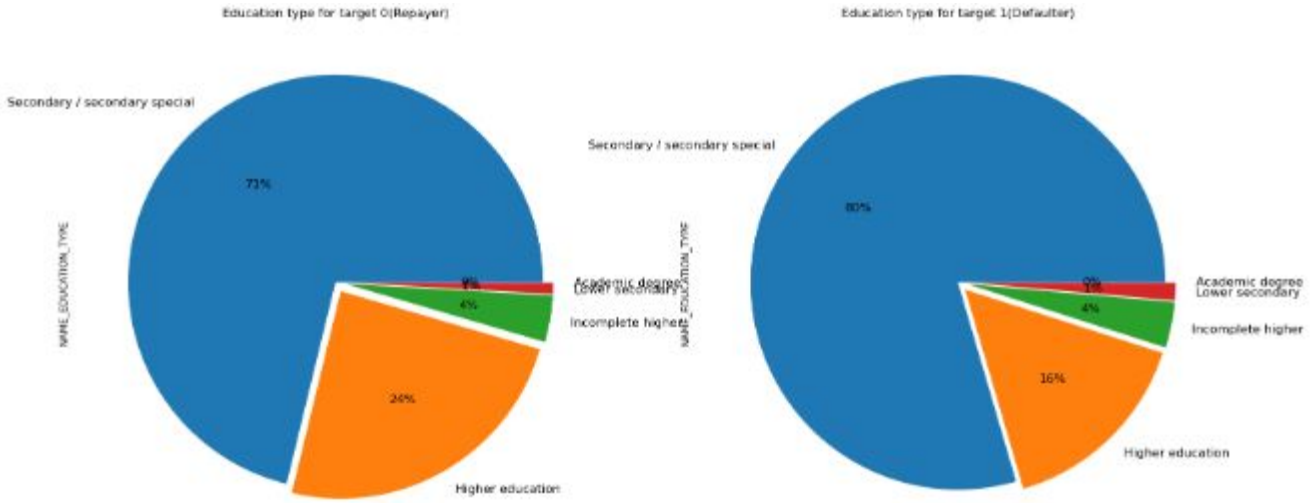
2. CODE_GENDER



The female client seems more responsible repayers among male among both the set

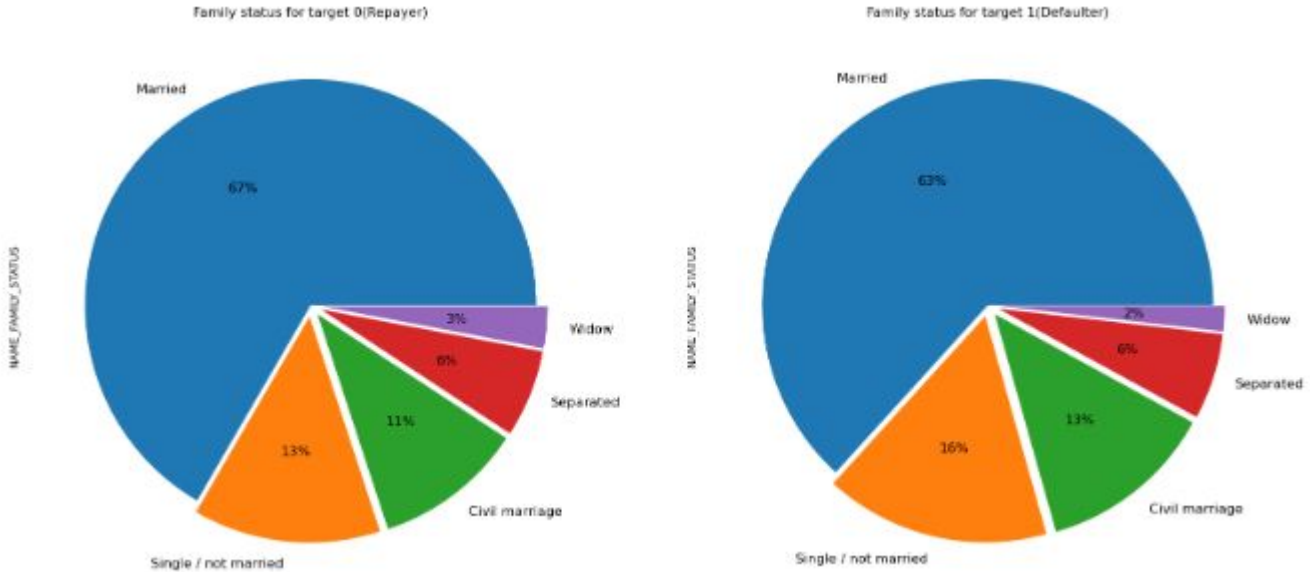
Also for contract type as cash loan and revolving loans maximum client prefer cash loans.

3. NAME_EDUCATION_TYPE



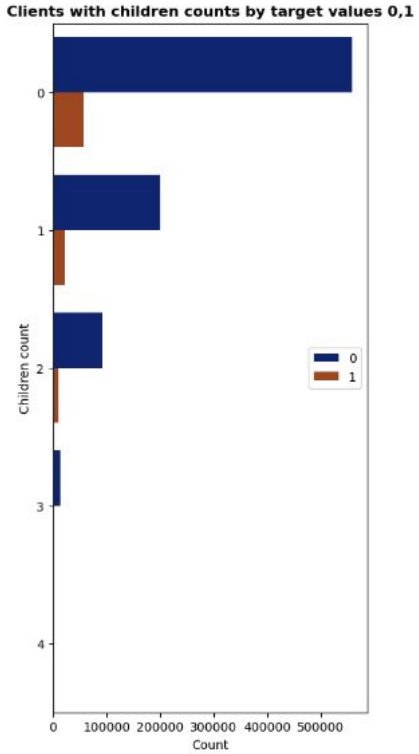
- The defaulters are 9 times higher than those with repayer having secondary education
- On other side the count for defaulter is reduced by 8 % in higher education , whereas other category has minority level for repayment status

4. NAME_FAMILY_STATUS



- Negligible different among both the client for facing difficulty in repayment for married one
- Where as an even distribution for family status in both cases for repayment

5. CNT_CHILDREN



- The client with no children are likely to repay than that of having more than 1 child.

Correlated variable for numeric analysis

Top 10 correlated variables

The clients with no payment difficulties

AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.985660
AMT_CREDIT_y	AMT_APPLICATION	0.973159
DAYS_TERMINATION	DAYS_LAST_DUE	0.930055
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.876816
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.865745
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.825342
AMT_ANNUITY_x	AMT_GOODS_PRICE_x	0.750534
AMT_CREDIT_x	AMT_ANNUITY_x	0.747118
AMT_APPLICATION	CNT_PAYMENT	0.648360

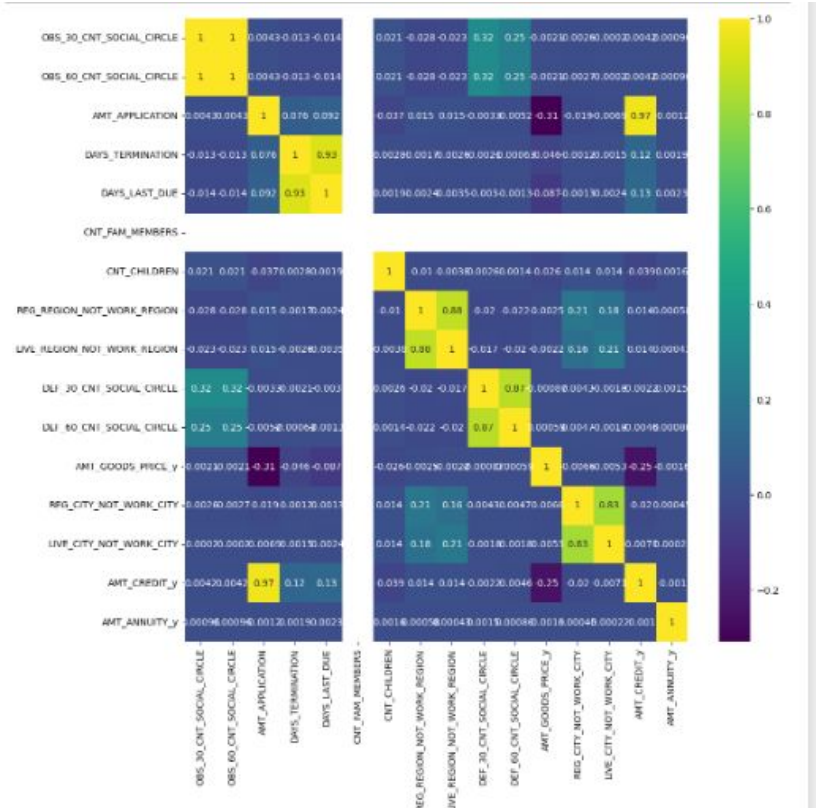
The clients who are defaulters

AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.981656
AMT_CREDIT_y	AMT_APPLICATION	0.973343
DAYS_LAST_DUE	DAYS_TERMINATION	0.945315
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.866704
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.856626
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.776536
AMT_GOODS_PRICE_x	AMT_ANNUITY_x	0.737564
AMT_CREDIT_x	AMT_ANNUITY_x	0.737129
CNT_PAYMENT	AMT_APPLICATION	0.664802

Heatmap analysis for Correlative values

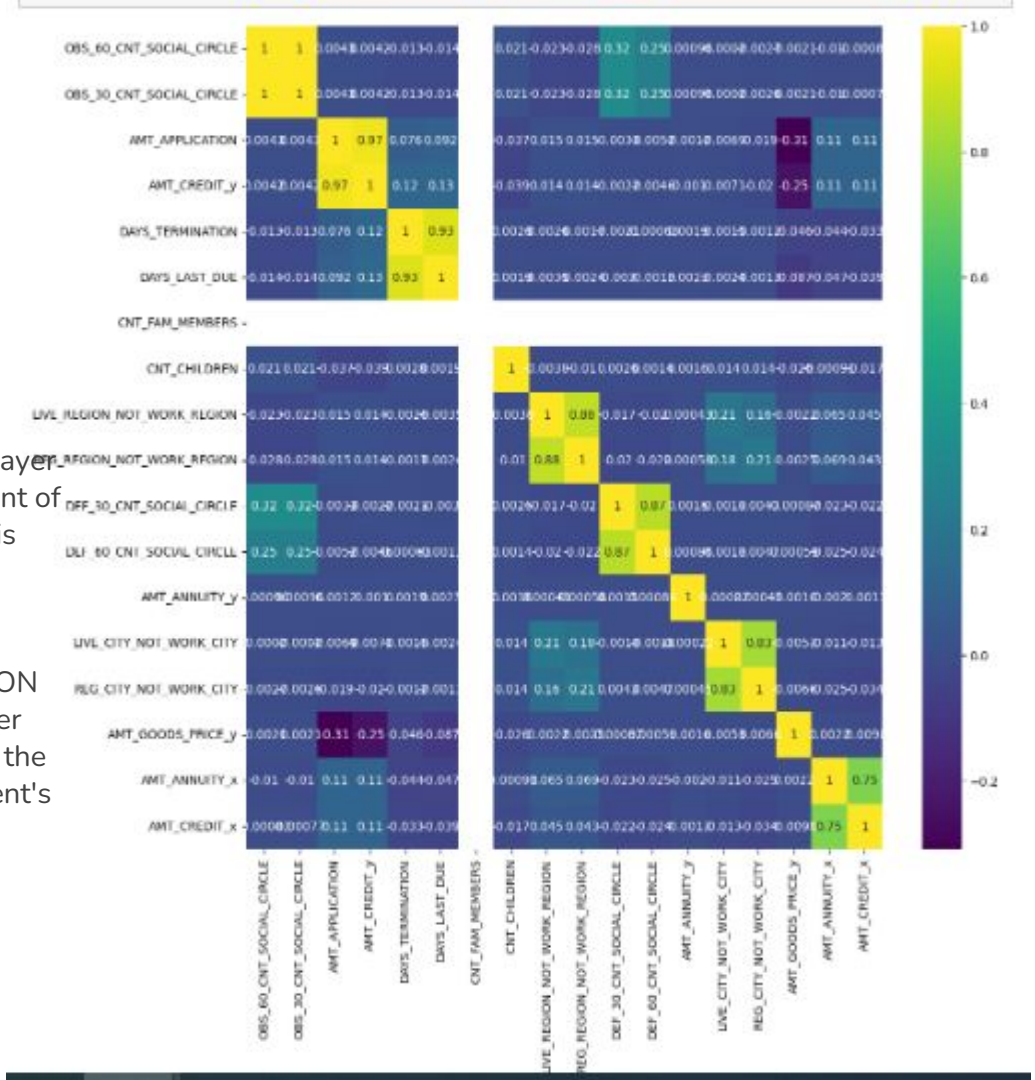
Heatmap for repayers

- The fact that AMT_APPLICATION and AMT_GOODS_PRICE have a high correlation indicates that the amount of credit a client previously requested is proportional to the goods price.
- CNT_CHILDREN and CNT_FAM_MEMBERS have profoundly corresponded which implies a client with kids is highly liable to have other family members too.
- AMT_ANNUITY and AMT_APPLICATION also have a high correlation, which means loan annuity is directly proportional to the goods price that the client asked for previously which is higher as shaded.



Heatmap for defaulter

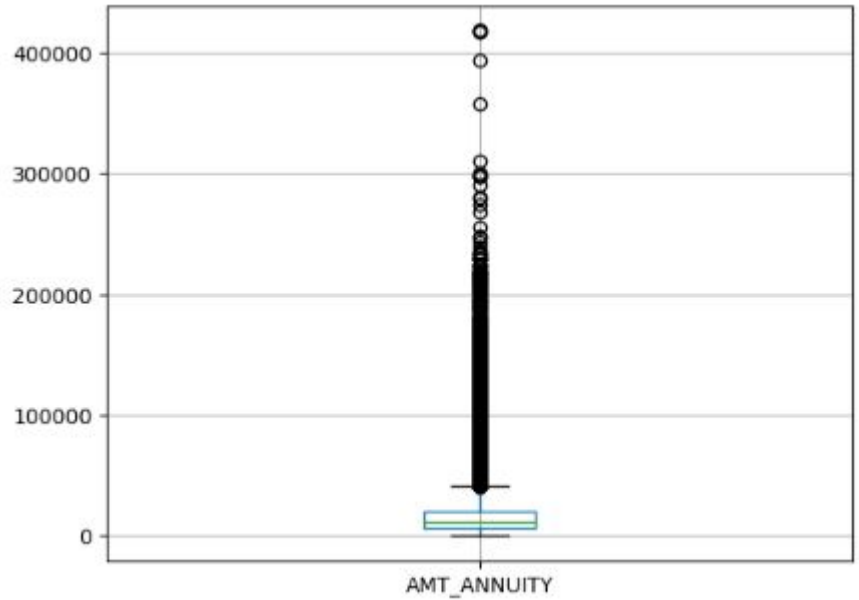
- AMT_GOODS_PRICE and AMT_APPLICATION also have a high correlation here also as that of the repayer heatmap. This indicates that the amount of credit the client requested in the past is proportional to the goods price.
- As goods price is higher so the client request for higher credit
- AMT_ANNUITY and AMT_APPLICATION have a high correlation with the repayer heatmap, which states that the higher the loan annuity issued, the higher the client's previously requested goods price.



Outlier

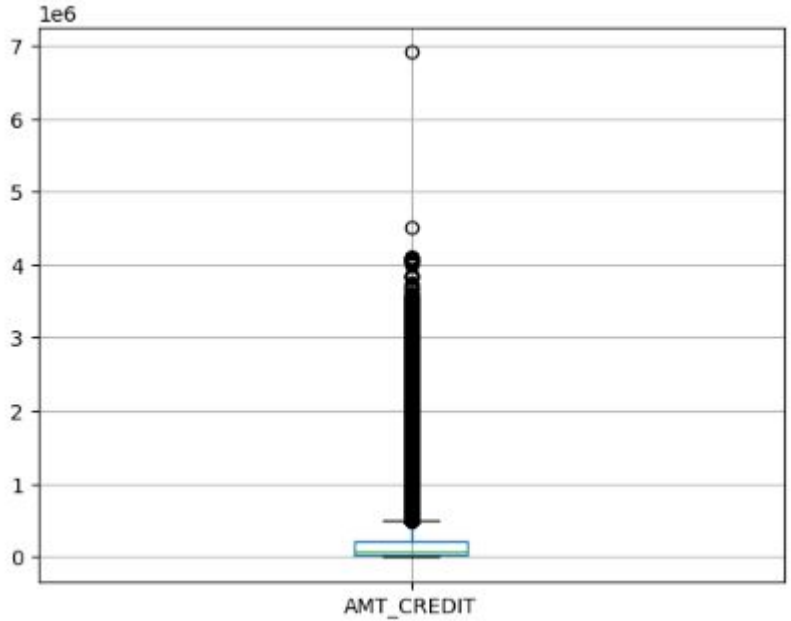
- AMT_ANNUITY
- CNT_FAM_MEMBERS
- AMT_CREDIT
- CNT_CHILDREN

AMT_ANNUITY



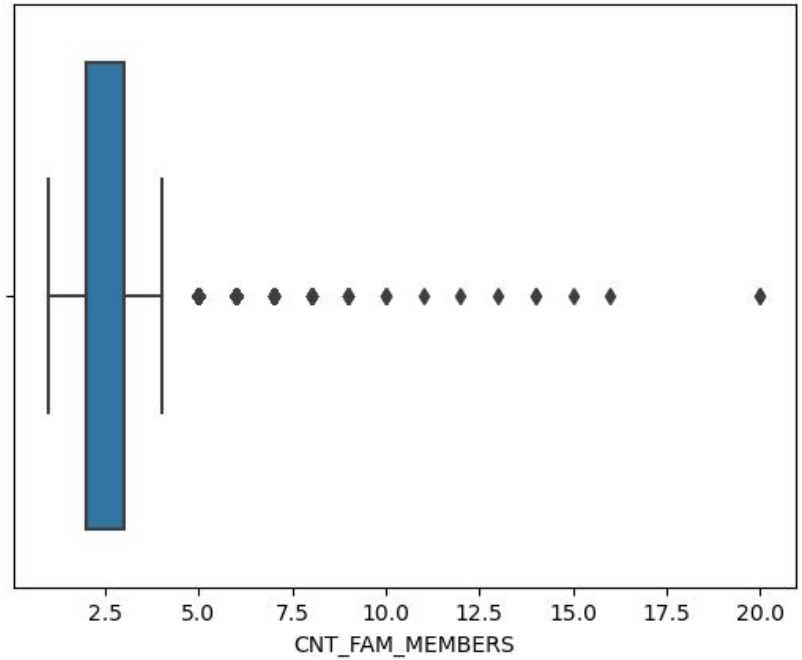
Values after 400000 shows outliers

AMT_CREDIT



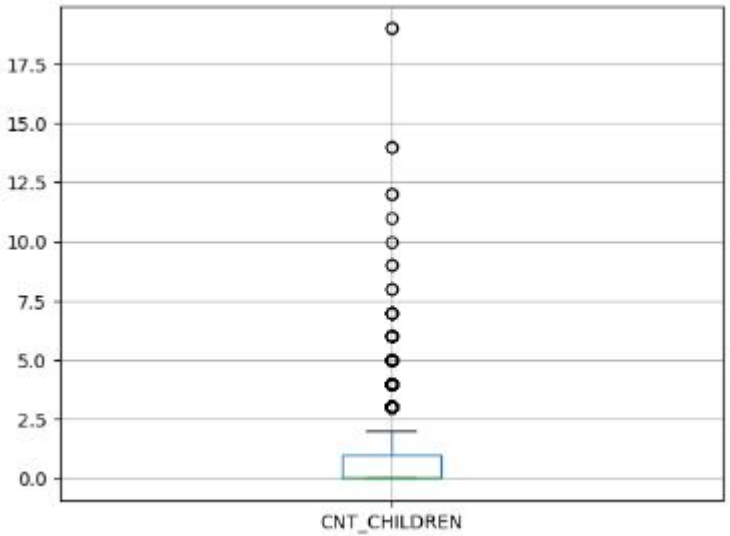
- Above 4 and more are outlier

CNT_FAM_MEMBERS



Members more than 5 in family according to box plot indicates outliers

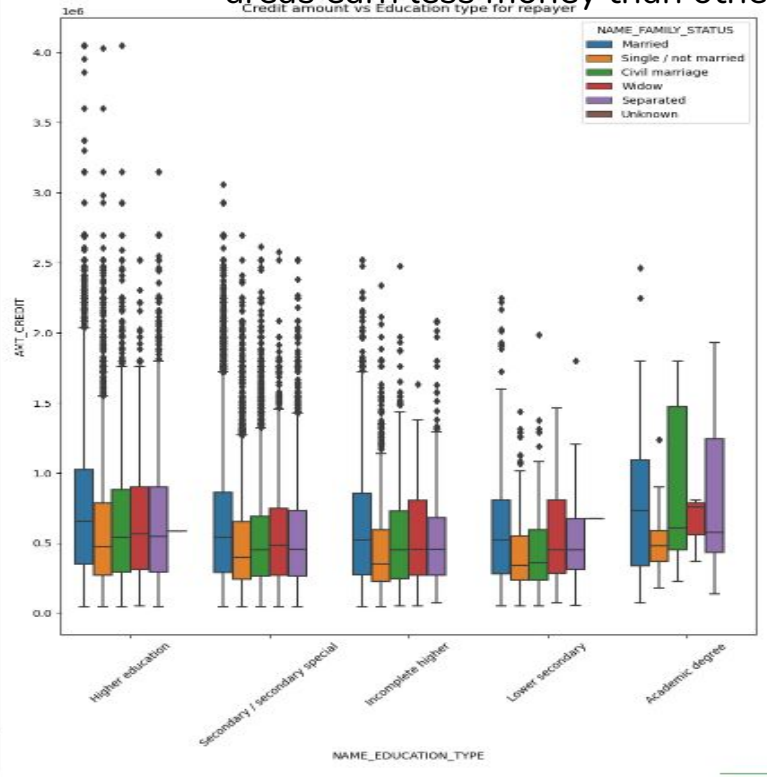
CNT_CHILDREN

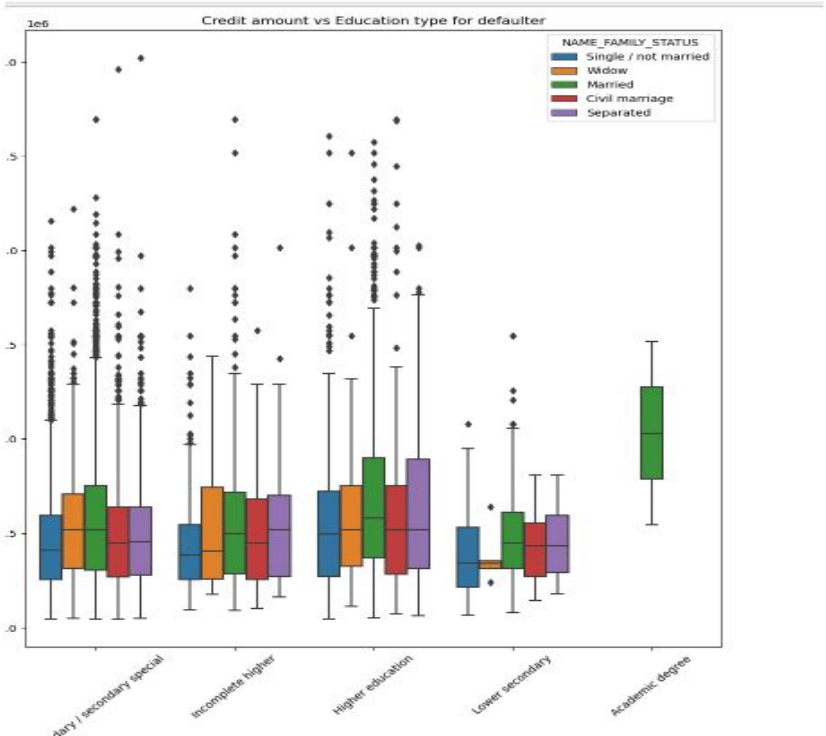


- Client whose values is above 2 or more are into outlier category
- Above 7 the application counts also decreases

Bivariate analysis

Likewise T0 i.e repayer plot. According to the boxplot above, for the education type "higher education," the income is roughly equal to family status. Academic degrees have fewer outliers, but their income is slightly higher than that of higher education. Lower secondary areas earn less money than other areas.





Conclusion

- Revolving loans depicts a not so opted loan in terms of loan preference by clients
- Females bets the repayment race by male in all area like business types, contract types , income .
- The clients with no children are likely to face less difficulty in repayment compared to singles.
- The bank needs to focus less on working customers because they have the most defaulters.