Bank Loan Case Study

Project Description –

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

This case study focuses on the same by analysing the consumer behaviours both past and present to find out whether that consumer is able to repay the loan or not.

This project identifies 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, 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.

Approach –

As an employee in a consumer finance company, which specialises in lending various types of loans to urban customers, it is my duty to ensure that the applicants capable of repaying the loan are not rejected.
And this analysis is done using Exploratory Data Analysis (EDA).

Before providing loan to a customer, the analyst team should keep in mind the two risks associated with the bank:

- 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.

Tech-Stack used –

I have used **Google Spreadsheet** and **Microsoft Excel 2019** for this project.
It makes it easy to visualize the results of the queries.

Insights –

I have derived the following information/insights using Excel:

Data cleaning is one of the most important practices before analysing a given data.

Let's start with the data set application_data.

The columns containing more than 50 % of NULL values are deleted.
The deleted columns are given below:

'OWN_CAR_AGE', 'EXT_SOURCE_1', 'APARTMENTS_AVG', 'BASEMENTAREA_AVG', 'YEARS_BUILD_AVG', 'COMMONAREA_AVG', 'ELEVATORS_AVG', 'ENTRANCES_AVG', 'FLOORSMIN_AVG', 'LANDAREA_AVG', 'LIVINGAPARTMENTS_AVG', 'LIVINGAREA_AVG', 'NONLIVINGAPARTMENTS_AVG', 'NONLIVINGAREA_AVG', 'APARTMENTS_MODE', 'BASEMENTAREA_MODE', 'YEARS_BUILD_MODE', 'COMMONAREA_MODE', 'ELEVATORS_MODE', 'ENTRANCES_MODE', 'FLOORSMIN_MODE', 'LANDAREA_MODE', 'LIVINGAPARTMENTS_MODE', 'LIVINGAREA_MODE', 'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAREA_MODE', 'APARTMENTS_MEDI', 'BASEMENTAREA_MEDI', 'YEARS_BUILD_MEDI', 'COMMONAREA_MEDI', 'ELEVATORS_MEDI', 'ENTRANCES_MEDI', 'FLOORSMIN_MEDI', 'LANDAREA_MEDI', 'LIVINGAPARTMENTS_MEDI', 'LIVINGAREA_MEDI', 'NONLIVINGAPARTMENTS_MEDI', 'NONLIVINGAREA_MEDI', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE', 'WALLSMATERIAL_MODE'

The columns containing less than 15 % NULL values are listed below:

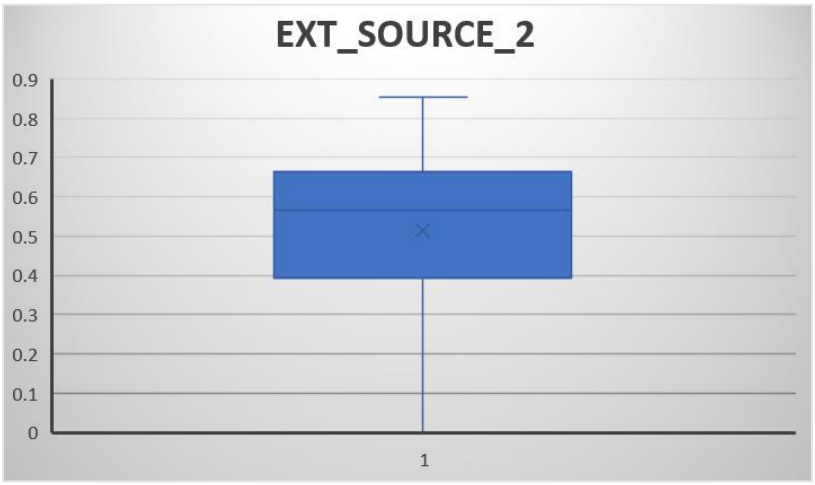
'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE','EXT_SOURCE_2','OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE','AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON','AMT_REQ_CREDIT_BUREAU_QRT', 'AMT_REQ_CREDIT_BUREAU_YEAR'

The cells containing NULL values in each column should be replaced with some suitable values.
The procedure for the same is explained below:

The seven columns whose NULL values are to be replaced are:

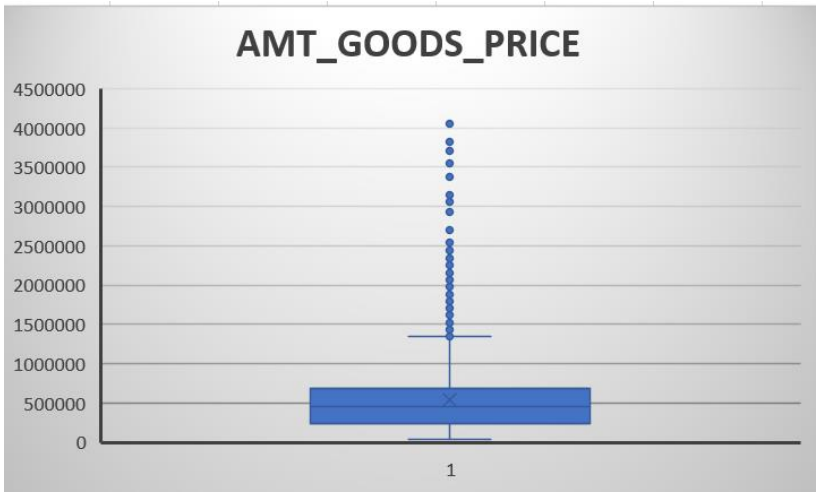
'EXT_SOURCE_2','AMT_GOODS_PRICE', 'OBS_30_CNT_SOCIAL_CIRCLE','OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE','DEF_30_CNT_SOCIAL_CIRCLE','NAME_TYPE_SUITE'

The columns 'EXT_SOURCE_2' and 'AMT_GOODS_PRICE' have continuous data. So, box plot is used to identify the median value for replacement in the columns
The rest of the columns have categorical data. So, mode of the data is calculated and is used to replace the NULL values.



For 'EXT_SOURCE_2' there is no outliers present. And there is no significant difference observed between mean and median. However, data look to be right skewed.

So missing values can be imputed with median value: 0.565



For 'AMT_GOODS_PRICE' there is significant number of outlier present in the data.

So data should be imputed with median value: 450000

For categorical variables / columns the value which should be imputed with should be maximum in frequency (mode).

So, the value to be imputed are:

NAME_TYPE_SUITE: Unaccompanied
OBS_30_CNT_SOCIAL_CIRCLE: 0.0
DEF_30_CNT_SOCIAL_CIRCLE: 0.0
OBS_60_CNT_SOCIAL_CIRCLE: 0.0
DEF_60_CNT_SOCIAL_CIRCLE: 0.0

I have removed some unwanted columns from the application_data dataset which is not needed in this analysis.

The dropped columns are:

'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL', 'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY', 'FLAG_EMAIL', 'CNT_FAM_MEMBERS', 'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY', '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', 'EXT_SOURCE_2', 'EXT_SOURCE_3', 'YEARS_BEGINEXPLUATATION_AVG', 'FLOORSMAX_AVG', 'YEARS_BEGINEXPLUATATION_MODE', 'FLOORSMAX_MODE', 'YEARS_BEGINEXPLUATATION_MEDI', 'FLOORSMAX_MEDI', 'TOTALAREA_MODE', 'EMERGENCYSTATE_MODE'

There are some columns where the value is mentioned as 'XNA' which means 'Not Available'. So, we have to find the number of rows and columns.

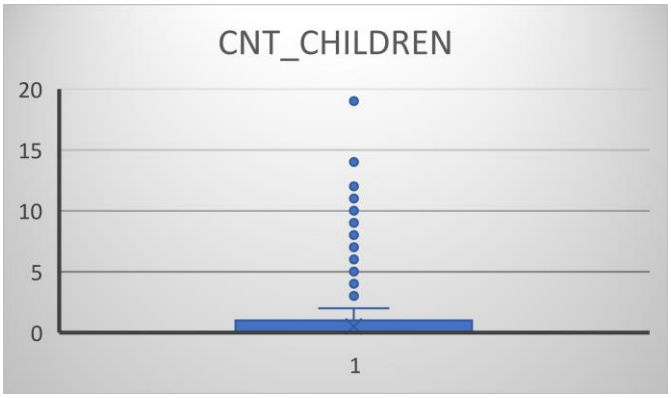
Since, Female is having the majority and only 4 rows are having XNA values, we can impute those with Gender 'F' as there will be no impact on the dataset. Also, there will no impact if we drop those rows.

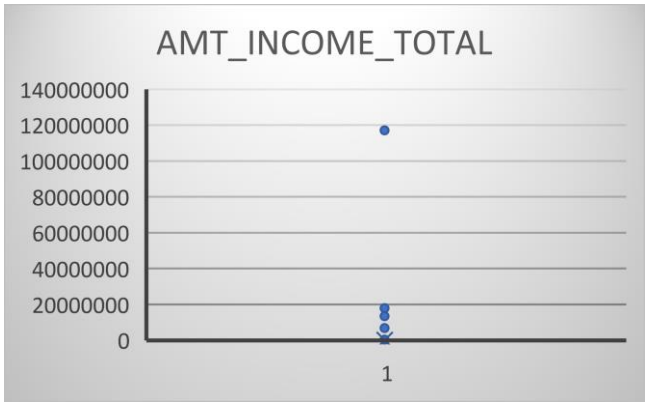
So, for column 'ORGANIZATION_TYPE', we have total count of 307511 rows of which 55374 rows are having 'XNA' values.
Dropped the rows having XNA values.

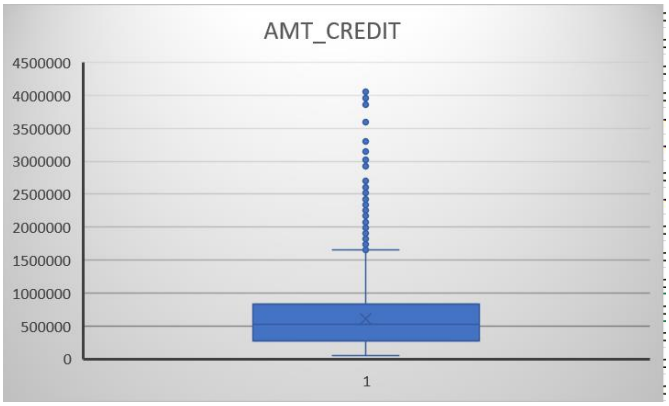
Following age/days columns are having -ve value, which needs to converted to +ve value.

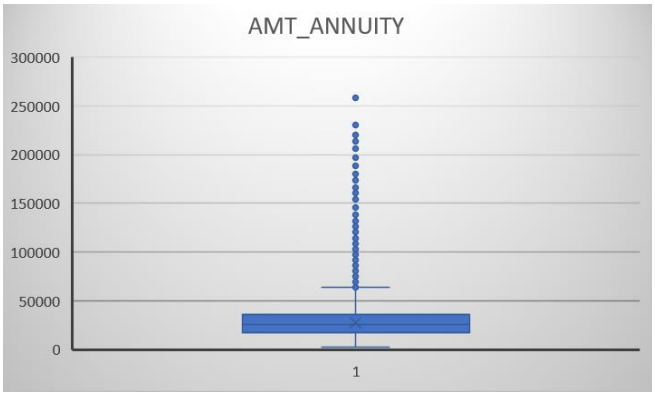
'DAYS_BIRTH'
'DAYS_EMPLOYED'
'DAYS_REGISTRATION'
'DAYS_ID_PUBLISH',
'DAYS_LAST_PHONE_CHANGE'

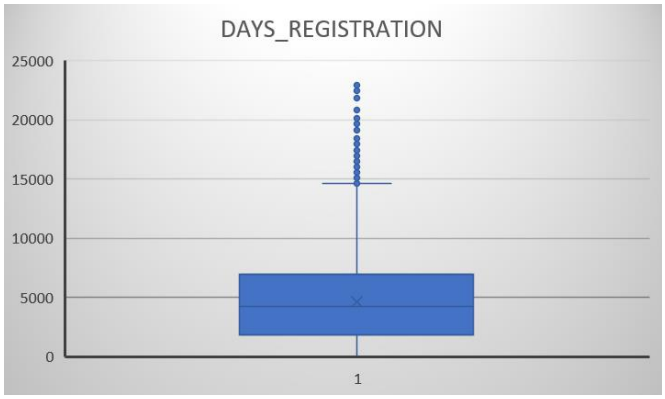
Box Plot for the following columns:
CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, DAYS_EMPLOYED, DAYS_REGISTRATION







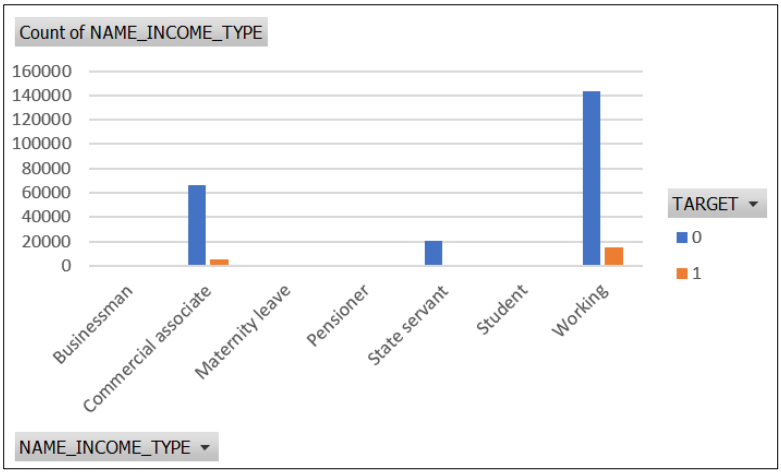




- The first quartile almost missing for CNT_CHILDREN that means most of the data are present in the first quartile.
- There is single high value data point as outlier present in AMT_INCOME_TOTAL and Removal this point will drastically impact the box plot for further analysis.
- The first quartiles are slim compare to third quartile for AMT_CREDIT, AMT_ANNUITY, DAYS_EMPLOYED, DAYS_REGISTRATION. This mean data is skewed towards first quartile.

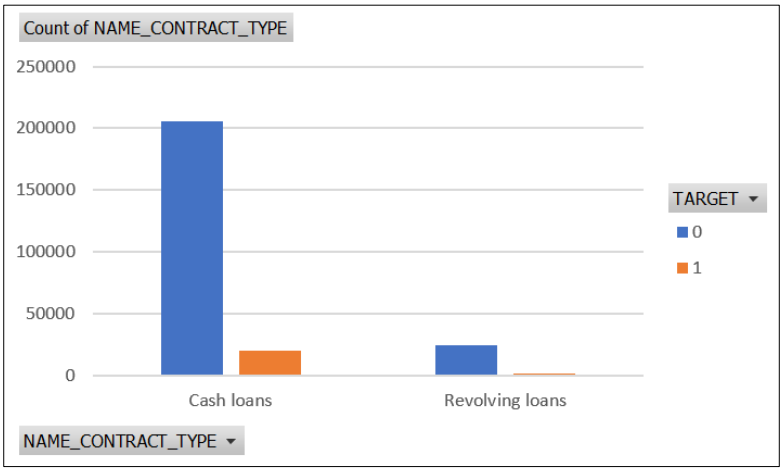
UNIVARIATE ANALYSIS

Categorical Univariate Analysis



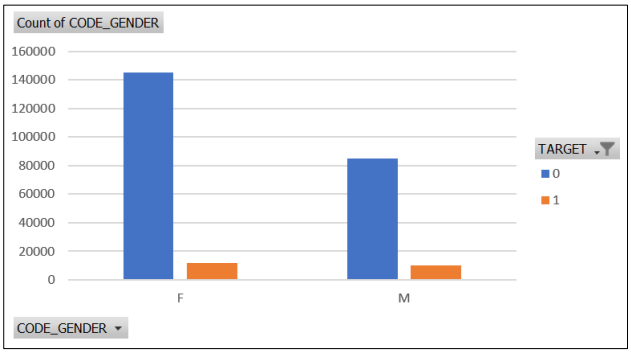
NAME_INCOME_TYPE:

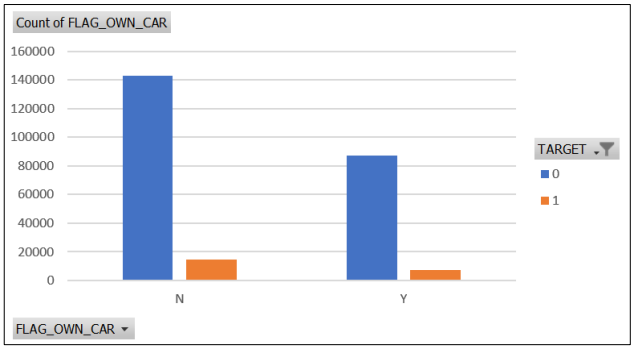
- Student pensioner and business have higher percentage of loan repayment.
- Working, State servant and Commercial associates have higher default percentage.
- Maternity category is significantly higher problem in repayments.



NAME_CONTRACT_TYPE

- For contract type 'cash loans' is having higher number of credits than 'Revolving loans' contract type.
- From the above graphs we can see that the Revolving loans are small amount compared to Cash loans but the % of non-payment for the revolving loans are comparatively high.

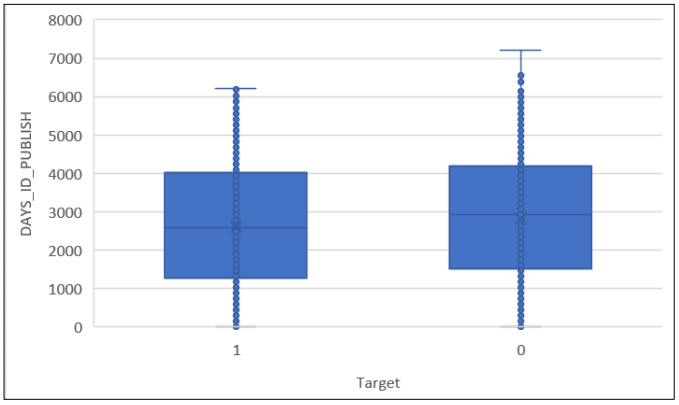


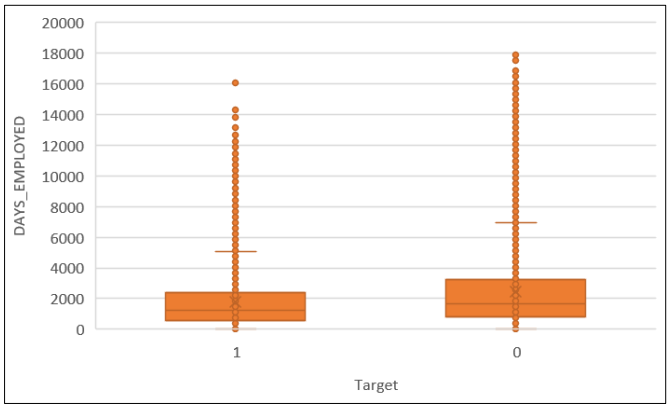


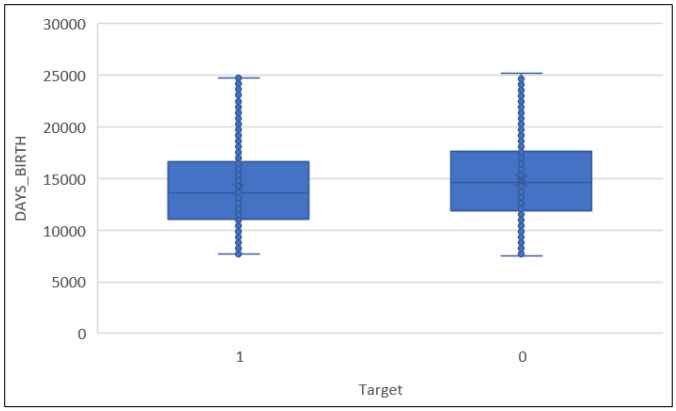
CODE_GENDER:
The % of defaulters are more in Male than Female

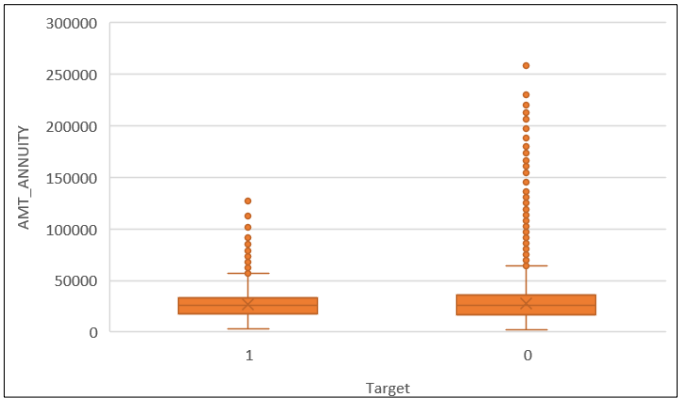
FLAG_OWN_CAR:
The person owning car is having higher percentage of defaulter.

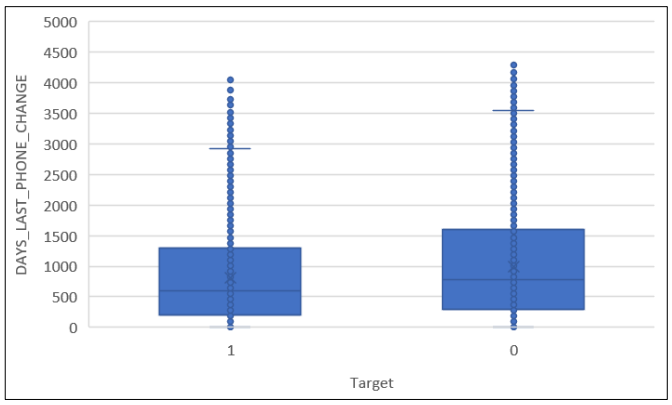
Continuous Univariate Analysis

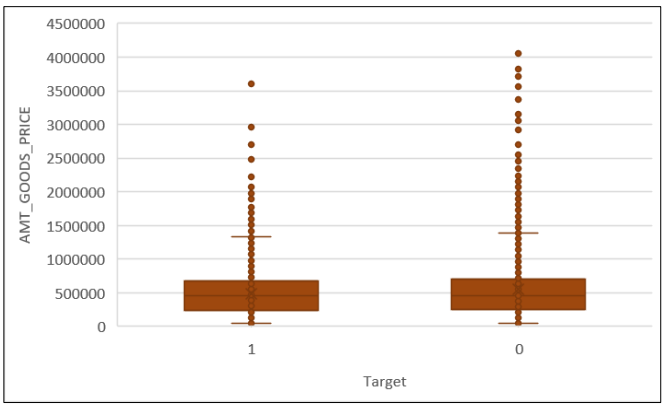












- In DAYS_BIRTH, the people having higher age are having higher probability of repayment.
- Some outliers are observed in 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'DAYS_EMPLOYED', ‘DAYS_LAST_PHONE_CHANGE’ in the dataset.
- Less outlier observed in DAYS_BIRTH and DAYS_ID_PUBLISH.
- 1st quartile is smaller than third quartile in In 'AMT_ANNUITY','AMT_GOODS_PRICE', DAYS_LAST_PHONE_CHANGE.
- In DAYS_ID_PUBLISH, people changing ID in recent days are relatively prone to be default.
- There is single high value data point as outlier present in DAYS_EMPLOYED. Removal this point will drastically impact the box plot for further analysis.

TOP 10 CORRELATIONS BETWEEN NUMERICAL VARIABLES

Target 0

Target 0	var1	var2	correlation
1	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1
2	AMT_GOODS_PRICE	AMT_CREDIT	0.99
3	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
4	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86
5	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.83
6	AMT_GOODS_PRICE	AMT_ANNUITY	0.78
7	AMT_ANNUITY	AMT_CREDIT	0.77
8	DAYS_EMPLOYED	DAYS_BIRTH	0.63
9	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.45
10	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.44

Target 1

Target 1	var1	var2	correlation
1	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1
2	AMT_GOODS_PRICE	AMT_CREDIT	0.98
3	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
4	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.85
5	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.78
6	AMT_GOODS_PRICE	AMT_ANNUITY	0.75
7	AMT_ANNUITY	AMT_CREDIT	0.75
8	DAYS_EMPLOYED	DAYS_BIRTH	0.58
9	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.5
10	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.47

From the above correlation analysis it is inferred that the highest corelation (1.0) is between (OBS_60_CNT_SOCIAL_CIRCLE with OBS_30_CNT_SOCIAL_CIRCLE) and (FLOORSMAX_MEDI with FLOORSMAX_AVG) which is same for both the data set.

Result –

1. Banks should focus more on contract type ‘Student’, ’pensioner’ and ‘Businessman’ with housing type other than ‘Co-op apartment’ for successful payments.

2. Banks should focus less on income type ‘Working’ as they are having the greatest number of unsuccessful payments.

3. In loan purpose ‘Repairs’:

a. Although having higher number of rejections in loan purposes with 'Repairs' there are observed difficulties in payment on time.
b. There are few places where loan payment is delay is significantly high.
c. Bank should keep continue to caution while giving loan for this purpose.

4. Bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment.

5. Bank can focus mostly on housing type ‘with parents’, ‘House\apartment’ and ‘municipal apartment’ for successful payments.