AKANKSHA SHINDE

Data Analyst Trainee

Dataset URL: [CLICK HERE](https://drive.google.com/drive/folders/1Mcvc5tJGvgJK8UetCtY1VnMYoVOMzaAe?usp=sharing)

Bank Loan Case Study

# Description

In this bank loan case study project, we aim to explore and analyse a dataset containing information about loan applicants and their loan application outcomes. The dataset includes various variables such as demographic information, financial information, and loan information.

Our objective is to gain insights into the variables that impact the loan application outcome, which is our target variable. We will perform exploratory data analysis (EDA) to understand the distribution and characteristics of each variable. We will then perform univariate, segmented univariate, and bivariate analyses to understand the relationship between each variable and the target variable.

# Approach

1. **Data Collection:** The dataset url is mentioned at the top of this doc. We dropped some columns, performed imputation and replaced negative values with positive ones.
2. **Univariate Analysis:** We performed univariate analysis for both categorical and numerical values.
3. **Segmented Univariate Analysis:** We created datasets out of our given datasets. Some variables were analysed by segments such as age, gender, income, or loan amount. This analysis can help identify patterns or relationships that may not be apparent in the overall dataset.
4. **Bivariate Analysis:** We have box plotted some variables like total income, credits based on educational qualification and family status. Which will help in describing how the loan approval rate varies by loan amount, credit score, or other demographic factors.
5. **Data Visualization:** We have created charts and graphs to visualise data. We also used box plots and whiskers to visualise IQR and outliers.
6. **Query Merging:** We have performed inner merging of two tables on Excel.

# Tech-Stack Used: MS Excel

1. **Data organisation and management:** Excel is an excellent tool for organising and managing large amounts of data. It allows you to create spreadsheets that can contain complex calculations, graphs, and charts to help you make sense of your data.
2. **Time-saving:** Excel can help you save time by automating repetitive tasks, such as calculations, sorting, filtering, and formatting. It also provides built-in functions for common mathematical and statistical calculations.
3. **Flexibility:** Excel is a versatile tool that can be used in a variety of settings and for a range of tasks, from simple household budgeting to complex financial analysis.
4. **Collaboration:** Excel allows for easy collaboration with others by allowing multiple users to access and edit the same spreadsheet simultaneously. It also provides tools for tracking changes and comments.
5. **Accessibility:** Excel is widely used and widely available, making it easy to access and share files with others. It is also relatively affordable compared to other data management tools.

# Insights

## Income Range

The people having 100000-200000 are having higher number of loans and also having higher defaulter. The income segment having >500000 are having less defaulters.

## Credit Range

The people having <100000 loan are less defaulter. Income having more than >100000 are almost equal % of loan defaulter

## Income Type

Student pensioners and businesses have a higher percentage of loan repayment. Working, State servant and Commercial associates have a higher default percentage. Maternity category is significantly higher in repayment.

## Contract Type

The contract type ‘cash loans’ has a higher number of credits than ‘Revolving loans’ contract type. From the above graphs we can see that the Revolving loans are small compared to Cash loans but the % of non payment for the revolving loans are comparatively high.

## Defaulters

The percentage of defaulters is more in Male than females. The person owning a car is having a higher percentage of defaulters.

## Credits and Outliers

Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.

Also, higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers. Civil marriage for Academic degree is having most of the credits in the third quartile.

## Target 0 Analysis

In Education type 'Higher education' the income amount is mostly equal with family status. It does contain many outliers.

Less outliers are for Academic degrees but their income amount is little higher than Higher education.

Lower secondary of civil marriage family status have less income than others.

## Target 1 Analysis

Family status of 'married' and 'separated' of Higher education are having higher number of credits than others.

Most of the outliers are from Education type 'Higher education' and 'Secondary'.

Civil marriage for Academic degree is having most of the credits in the third quartile.

Education type 'Higher education' the income amount is mostly equal with family status.

Less outliers are for Academic degrees but their income amount is little higher than Higher education.

Lower secondary students have less income than others.

## Correlation Analysis

The highest correlation (1.0) is between (OBS\_60\_CNT\_SOCIAL\_CIRCLE with OBS\_30\_CNT\_SOCIAL\_CIRCLE) which is the same for both Target 0 and Target 1.

## Previous Data Analysis

Most rejection of loans came from purpose 'repairs'. For education purposes we have an equal number of approvals and rejection. Paying other loans and buying a new car is having significantly higher rejection than approval.

In loan purpose ‘Repairs’:

a. Although having a higher number of rejection in loan purposes with 'Repairs' there are observed difficulties in payment on time.

b. There are few places where loan payment delay is significantly high.

c. Banks should continue to caution while giving loans for this purpose.

Banks should avoid giving loans to the housing type of office apartment as they are having difficulties in payment.

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

# Results

I have learnt Univariate, Segmented Univariate and Bivariate analysis. I also learnt Data Cleaning and Imputation. Along with that I have learnt Query Merging/Table Merging and this has helped me to upskill myself through this project.

REPORT

# Problem Statement:

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.

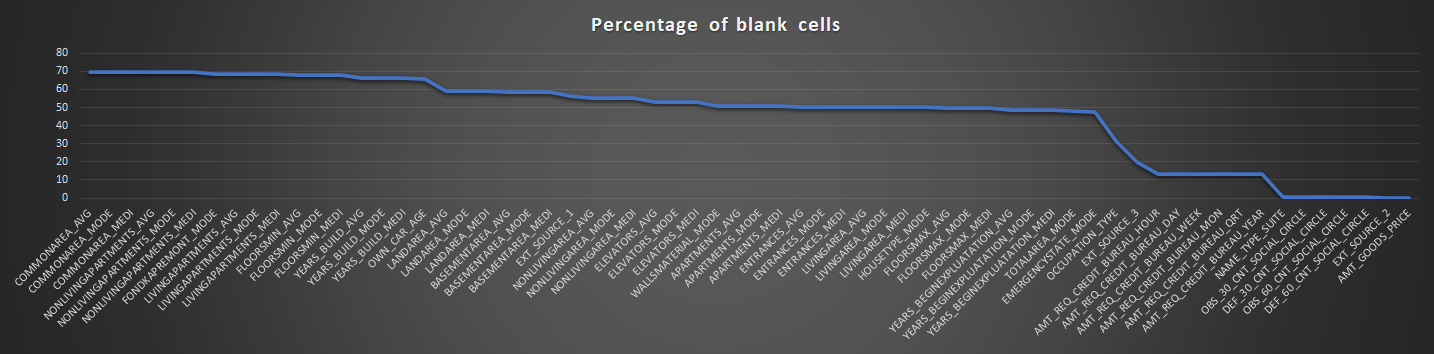
The company wants to understand the driving factors (or driver variables) behind the loan default. i.e the variables which are strong in loan default.

# Columns with missing values:

Below table shows the columns with percentage of missing values-

|  |  |  |
| --- | --- | --- |
| COLUMN NAME | Blank Values | Percentage of blank cells |
| COMMONAREA\_AVG | 214865 | 69.87 |
| COMMONAREA\_MODE | 214865 | 69.87 |
| COMMONAREA\_MEDI | 214865 | 69.87 |
| NONLIVINGAPARTMENTS\_AVG | 213514 | 69.43 |
| NONLIVINGAPARTMENTS\_MODE | 213514 | 69.43 |
| NONLIVINGAPARTMENTS\_MEDI | 213514 | 69.43 |
| FONDKAPREMONT\_MODE | 210295 | 68.39 |
| LIVINGAPARTMENTS\_AVG | 210199 | 68.35 |
| LIVINGAPARTMENTS\_MODE | 210199 | 68.35 |
| LIVINGAPARTMENTS\_MEDI | 210199 | 68.35 |
| FLOORSMIN\_AVG | 208642 | 67.85 |
| FLOORSMIN\_MODE | 208642 | 67.85 |
| FLOORSMIN\_MEDI | 208642 | 67.85 |
| YEARS\_BUILD\_AVG | 204488 | 66.5 |
| YEARS\_BUILD\_MODE | 204488 | 66.5 |
| YEARS\_BUILD\_MEDI | 204488 | 66.5 |
| OWN\_CAR\_AGE | 202929 | 65.99 |
| LANDAREA\_AVG | 182590 | 59.38 |
| LANDAREA\_MODE | 182590 | 59.38 |
| LANDAREA\_MEDI | 182590 | 59.38 |
| BASEMENTAREA\_AVG | 179943 | 58.52 |
| BASEMENTAREA\_MODE | 179943 | 58.52 |
| BASEMENTAREA\_MEDI | 179943 | 58.52 |
| EXT\_SOURCE\_1 | 173378 | 56.38 |
| NONLIVINGAREA\_AVG | 169682 | 55.18 |
| NONLIVINGAREA\_MODE | 169682 | 55.18 |
| NONLIVINGAREA\_MEDI | 169682 | 55.18 |
| ELEVATORS\_AVG | 163891 | 53.3 |
| ELEVATORS\_MODE | 163891 | 53.3 |
| ELEVATORS\_MEDI | 163891 | 53.3 |
| WALLSMATERIAL\_MODE | 156341 | 50.84 |
| APARTMENTS\_AVG | 156061 | 50.75 |
| APARTMENTS\_MODE | 156061 | 50.75 |
| APARTMENTS\_MEDI | 156061 | 50.75 |
| ENTRANCES\_AVG | 154828 | 50.35 |
| ENTRANCES\_MODE | 154828 | 50.35 |
| ENTRANCES\_MEDI | 154828 | 50.35 |
| LIVINGAREA\_AVG | 154350 | 50.19 |
| LIVINGAREA\_MODE | 154350 | 50.19 |
| LIVINGAREA\_MEDI | 154350 | 50.19 |
| HOUSETYPE\_MODE | 154297 | 50.18 |
| FLOORSMAX\_AVG | 153020 | 49.76 |
| FLOORSMAX\_MODE | 153020 | 49.76 |
| FLOORSMAX\_MEDI | 153020 | 49.76 |
| YEARS\_BEGINEXPLUATATION\_AVG | 150007 | 48.78 |
| YEARS\_BEGINEXPLUATATION\_MODE | 150007 | 48.78 |
| YEARS\_BEGINEXPLUATATION\_MEDI | 150007 | 48.78 |
| TOTALAREA\_MODE | 148431 | 48.27 |
| EMERGENCYSTATE\_MODE | 145755 | 47.4 |
| OCCUPATION\_TYPE | 96391 | 31.35 |
| EXT\_SOURCE\_3 | 60965 | 19.83 |
| AMT\_REQ\_CREDIT\_BUREAU\_HOUR | 41519 | 13.5 |
| AMT\_REQ\_CREDIT\_BUREAU\_DAY | 41519 | 13.5 |
| AMT\_REQ\_CREDIT\_BUREAU\_WEEK | 41519 | 13.5 |
| AMT\_REQ\_CREDIT\_BUREAU\_MON | 41519 | 13.5 |
| AMT\_REQ\_CREDIT\_BUREAU\_QRT | 41519 | 13.5 |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | 41519 | 13.5 |
| NAME\_TYPE\_SUITE | 1292 | 0.42 |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.33 |
| DEF\_30\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.33 |
| OBS\_60\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.33 |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.33 |
| EXT\_SOURCE\_2 | 660 | 0.21 |
| AMT\_GOODS\_PRICE | 278 | 0.09 |
| SK\_ID\_CURR | 0 | 0 |
| TARGET | 0 | 0 |
| NAME\_CONTRACT\_TYPE | 0 | 0 |
| CODE\_GENDER | 0 | 0 |
| FLAG\_OWN\_CAR | 0 | 0 |
| FLAG\_OWN\_REALTY | 0 | 0 |
| CNT\_CHILDREN | 0 | 0 |
| AMT\_INCOME\_TOTAL | 0 | 0 |
| AMT\_CREDIT | 0 | 0 |
| AMT\_ANNUITY | 12 | 0 |
| NAME\_INCOME\_TYPE | 0 | 0 |
| NAME\_EDUCATION\_TYPE | 0 | 0 |
| NAME\_FAMILY\_STATUS | 0 | 0 |
| NAME\_HOUSING\_TYPE | 0 | 0 |
| REGION\_POPULATION\_RELATIVE | 0 | 0 |
| DAYS\_BIRTH | 0 | 0 |
| DAYS\_EMPLOYED | 0 | 0 |
| DAYS\_REGISTRATION | 0 | 0 |
| DAYS\_ID\_PUBLISH | 0 | 0 |
| FLAG\_MOBIL | 0 | 0 |
| FLAG\_EMP\_PHONE | 0 | 0 |
| FLAG\_WORK\_PHONE | 0 | 0 |
| FLAG\_CONT\_MOBILE | 0 | 0 |
| FLAG\_PHONE | 0 | 0 |
| FLAG\_EMAIL | 0 | 0 |
| CNT\_FAM\_MEMBERS | 2 | 0 |
| REGION\_RATING\_CLIENT | 0 | 0 |
| REGION\_RATING\_CLIENT\_W\_CITY | 0 | 0 |
| WEEKDAY\_APPR\_PROCESS\_START | 0 | 0 |
| HOUR\_APPR\_PROCESS\_START | 0 | 0 |
| REG\_REGION\_NOT\_LIVE\_REGION | 0 | 0 |
| REG\_REGION\_NOT\_WORK\_REGION | 0 | 0 |
| LIVE\_REGION\_NOT\_WORK\_REGION | 0 | 0 |
| REG\_CITY\_NOT\_LIVE\_CITY | 0 | 0 |
| REG\_CITY\_NOT\_WORK\_CITY | 0 | 0 |
| LIVE\_CITY\_NOT\_WORK\_CITY | 0 | 0 |
| ORGANIZATION\_TYPE | 0 | 0 |
| DAYS\_LAST\_PHONE\_CHANGE | 1 | 0 |
| FLAG\_DOCUMENT\_2 | 0 | 0 |
| FLAG\_DOCUMENT\_3 | 0 | 0 |
| FLAG\_DOCUMENT\_4 | 0 | 0 |
| FLAG\_DOCUMENT\_5 | 0 | 0 |
| FLAG\_DOCUMENT\_6 | 0 | 0 |
| FLAG\_DOCUMENT\_7 | 0 | 0 |
| FLAG\_DOCUMENT\_8 | 0 | 0 |
| FLAG\_DOCUMENT\_9 | 0 | 0 |
| FLAG\_DOCUMENT\_10 | 0 | 0 |
| FLAG\_DOCUMENT\_11 | 0 | 0 |
| FLAG\_DOCUMENT\_12 | 0 | 0 |
| FLAG\_DOCUMENT\_13 | 0 | 0 |
| FLAG\_DOCUMENT\_14 | 0 | 0 |
| FLAG\_DOCUMENT\_15 | 0 | 0 |
| FLAG\_DOCUMENT\_16 | 0 | 0 |
| FLAG\_DOCUMENT\_17 | 0 | 0 |
| FLAG\_DOCUMENT\_18 | 0 | 0 |
| FLAG\_DOCUMENT\_19 | 0 | 0 |
| FLAG\_DOCUMENT\_20 | 0 | 0 |
| FLAG\_DOCUMENT\_21 | 0 | 0 |

Below chart shows columns only with missing values-



Now, we identify columns those have missing values greater than 50%-

|  |  |  |
| --- | --- | --- |
| COLUMN NAME | Blank Values | Percentage of blank cells |
| COMMONAREA\_AVG | 214865 | 69.87 |
| COMMONAREA\_MODE | 214865 | 69.87 |
| COMMONAREA\_MEDI | 214865 | 69.87 |
| NONLIVINGAPARTMENTS\_AVG | 213514 | 69.43 |
| NONLIVINGAPARTMENTS\_MODE | 213514 | 69.43 |
| NONLIVINGAPARTMENTS\_MEDI | 213514 | 69.43 |
| FONDKAPREMONT\_MODE | 210295 | 68.39 |
| LIVINGAPARTMENTS\_AVG | 210199 | 68.35 |
| LIVINGAPARTMENTS\_MODE | 210199 | 68.35 |
| LIVINGAPARTMENTS\_MEDI | 210199 | 68.35 |
| FLOORSMIN\_AVG | 208642 | 67.85 |
| FLOORSMIN\_MODE | 208642 | 67.85 |
| FLOORSMIN\_MEDI | 208642 | 67.85 |
| YEARS\_BUILD\_AVG | 204488 | 66.5 |
| YEARS\_BUILD\_MODE | 204488 | 66.5 |
| YEARS\_BUILD\_MEDI | 204488 | 66.5 |
| OWN\_CAR\_AGE | 202929 | 65.99 |
| LANDAREA\_AVG | 182590 | 59.38 |
| LANDAREA\_MODE | 182590 | 59.38 |
| LANDAREA\_MEDI | 182590 | 59.38 |
| BASEMENTAREA\_AVG | 179943 | 58.52 |
| BASEMENTAREA\_MODE | 179943 | 58.52 |
| BASEMENTAREA\_MEDI | 179943 | 58.52 |
| EXT\_SOURCE\_1 | 173378 | 56.38 |
| NONLIVINGAREA\_AVG | 169682 | 55.18 |
| NONLIVINGAREA\_MODE | 169682 | 55.18 |
| NONLIVINGAREA\_MEDI | 169682 | 55.18 |
| ELEVATORS\_AVG | 163891 | 53.3 |
| ELEVATORS\_MODE | 163891 | 53.3 |
| ELEVATORS\_MEDI | 163891 | 53.3 |
| WALLSMATERIAL\_MODE | 156341 | 50.84 |
| APARTMENTS\_AVG | 156061 | 50.75 |
| APARTMENTS\_MODE | 156061 | 50.75 |
| APARTMENTS\_MEDI | 156061 | 50.75 |
| ENTRANCES\_AVG | 154828 | 50.35 |
| ENTRANCES\_MODE | 154828 | 50.35 |
| ENTRANCES\_MEDI | 154828 | 50.35 |
| LIVINGAREA\_AVG | 154350 | 50.19 |
| LIVINGAREA\_MODE | 154350 | 50.19 |
| LIVINGAREA\_MEDI | 154350 | 50.19 |
| HOUSETYPE\_MODE | 154297 | 50.18 |

Now, we are dropping all the columns having missing values greater than 50%.

Now, we identifying columns with less missing values (<15%)-

|  |  |  |
| --- | --- | --- |
| COLUMN NAME | Blank Values | Percentage of blank cells |
| AMT\_GOODS\_PRICE | 278 | 0.09 |
| EXT\_SOURCE\_2 | 660 | 0.21 |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.33 |
| DEF\_30\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.33 |
| OBS\_60\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.33 |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | 1021 | 0.33 |
| NAME\_TYPE\_SUITE | 1292 | 0.42 |
| AMT\_REQ\_CREDIT\_BUREAU\_HOUR | 41519 | 13.5 |
| AMT\_REQ\_CREDIT\_BUREAU\_DAY | 41519 | 13.5 |
| AMT\_REQ\_CREDIT\_BUREAU\_WEEK | 41519 | 13.5 |
| AMT\_REQ\_CREDIT\_BUREAU\_MON | 41519 | 13.5 |
| AMT\_REQ\_CREDIT\_BUREAU\_QRT | 41519 | 13.5 |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | 41519 | 13.5 |

These columns shall be imputed with suitable values which shall be explained subsequently.

Identifying unique values in the column having <15% null value-

|  |  |  |  |
| --- | --- | --- | --- |
| Columns | Median | Mean | Number of unique values |
| EXT\_SOURCE\_2 | 0.565961 | 0.514392674 | 119831 |
| AMT\_GOODS\_PRICE | 450000 | 538396.2075 | 1002 |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | 0 | 1.422245424 | 33 |
| OBS\_60\_CNT\_SOCIAL\_CIRCLE | 0 | 1.405292179 | 33 |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | 1 | 1.899974435 | 25 |
| AMT\_REQ\_CREDIT\_BUREAU\_MON | 0 | 0.26739526 | 24 |
| AMT\_REQ\_CREDIT\_BUREAU\_QRT | 0 | 0.26547415 | 11 |
| DEF\_30\_CNT\_SOCIAL\_CIRCLE | 0 | 0.143420666 | 10 |
| NAME\_TYPE\_SUITE | #NUM! | #DIV/0! | 10 |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | 0 | 0.100048941 | 9 |
| AMT\_REQ\_CREDIT\_BUREAU\_DAY | 0 | 0.007000211 | 9 |
| AMT\_REQ\_CREDIT\_BUREAU\_WEEK | 0 | 0.034361936 | 9 |
| AMT\_REQ\_CREDIT\_BUREAU\_HOUR | 0 | 0.006402448 | 5 |

For analysis of imputation we have selected 7 columns.

Continuous variables:

'EXT\_SOURCE\_2'

'AMT\_GOODS\_PRICE'

Categorical variables:

'OBS\_30\_CNT\_SOCIAL\_CIRCLE'

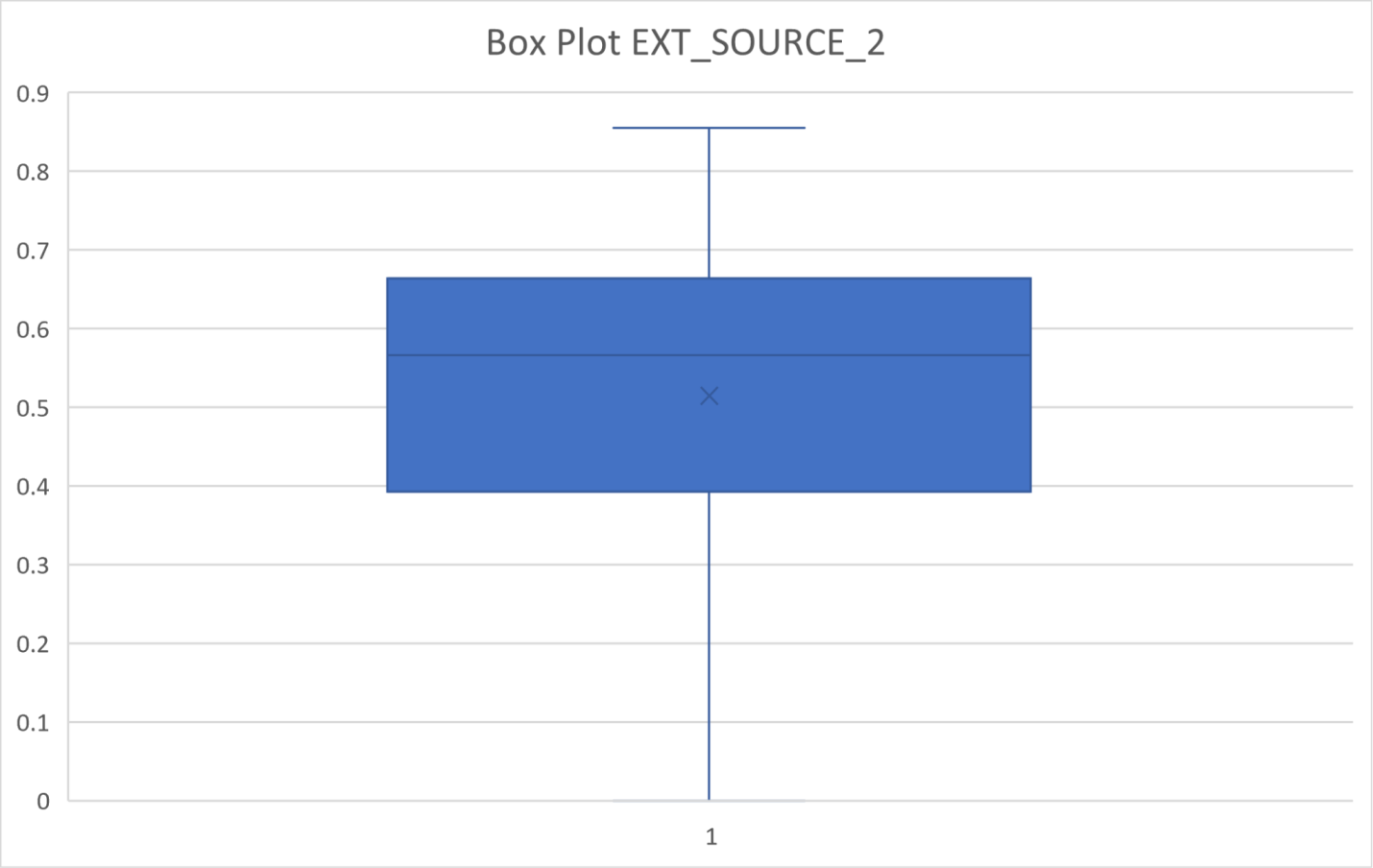
'OBS\_60\_CNT\_SOCIAL\_CIRCLE'

'DEF\_60\_CNT\_SOCIAL\_CIRCLE'

'DEF\_30\_CNT\_SOCIAL\_CIRCLE'

'NAME\_TYPE\_SUITE'

Box plot for continuous variable-





Inference from box plot:

For 'EXT\_SOURCE\_2' there are no outliers present. And there is no significant difference observed between mean and median. However data looks to be right skewed. So missing values can be imputed with median value: 0.565961

For 'AMT\_GOODS\_PRICE' there is a significant number of outliers present in the data. SO data should be imputed with median value: 450000

Maximum Frequency of categorical values are,

NAME\_TYPE\_SUITE: Unaccompanied

OBS\_30\_CNT\_SOCIAL\_CIRCLE: 0

DEF\_30\_CNT\_SOCIAL\_CIRCLE: 0

OBS\_60\_CNT\_SOCIAL\_CIRCLE: 0

DEF\_60\_CNT\_SOCIAL\_CIRCLE: 0

For categorical variables the value which should be imputed with maximum frequency.

So the value to be imputed are:

NAME\_TYPE\_SUITE: Unaccompanied

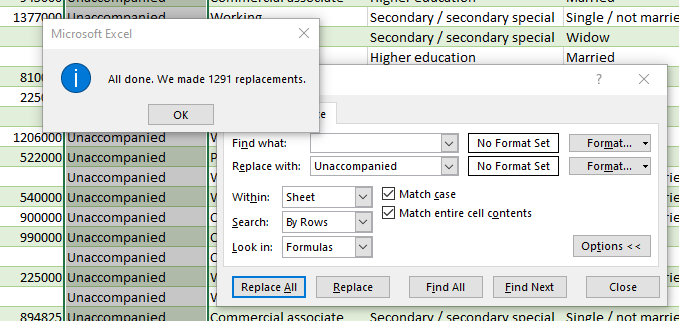
OBS\_30\_CNT\_SOCIAL\_CIRCLE: 0

DEF\_30\_CNT\_SOCIAL\_CIRCLE: 0

OBS\_60\_CNT\_SOCIAL\_CIRCLE: 0

DEF\_60\_CNT\_SOCIAL\_CIRCLE: 0

Here, I have replaced some blank values-



Now, we are dropping unwanted columns below for better analysis-

'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',

'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',

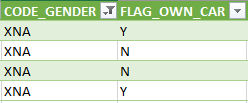
'FLOORSMAX\_MODE','YEARS\_BEGINEXPLUATATION\_MEDI','FLOORSMAX\_MEDI','TOTALAREA\_MODE','EMERGENCYSTATE\_MODE'

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.

|  |  |
| --- | --- |
| CODE\_GENDER | 1 |
| M | 105059 |
| F | 202448 |
| XNA | 4 |
| 0 | 0 |

Maximum Frequency: F

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 be no impact if we drop those rows.



Now, replacing these XNA values to F will have no impact on the dataset.

From Organization\_Type Column-

|  |
| --- |
| Count of XNA, Org\_type |
| 55374 |
| 18.007% |

Highest Frequency of Categorical Value: Business Entity Type 3

Since, the percentage of XNA values is more than 18% we cannot replace it.

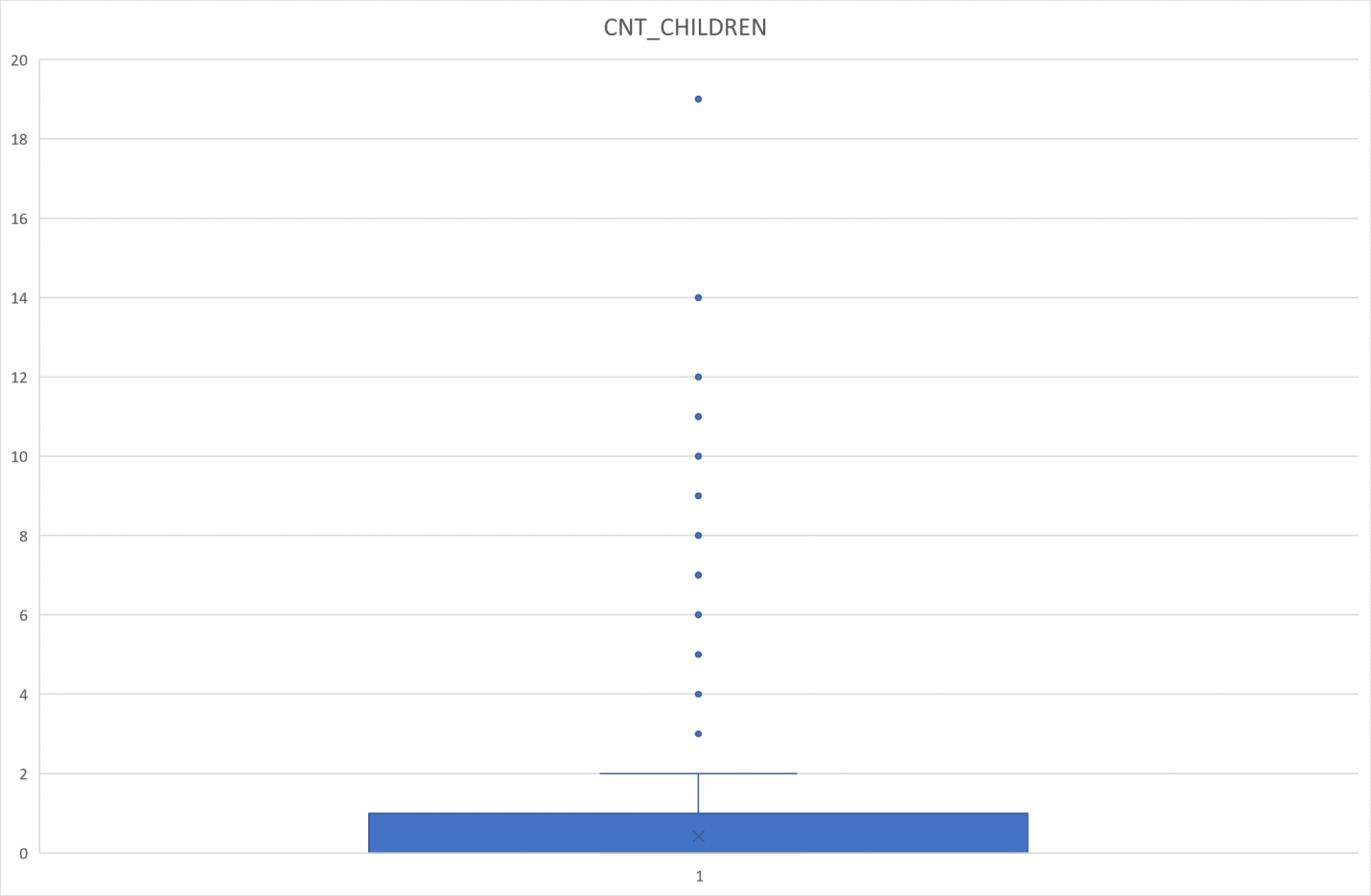
Following age/days columns are having -ve value, which needs to be converted to +ve value.

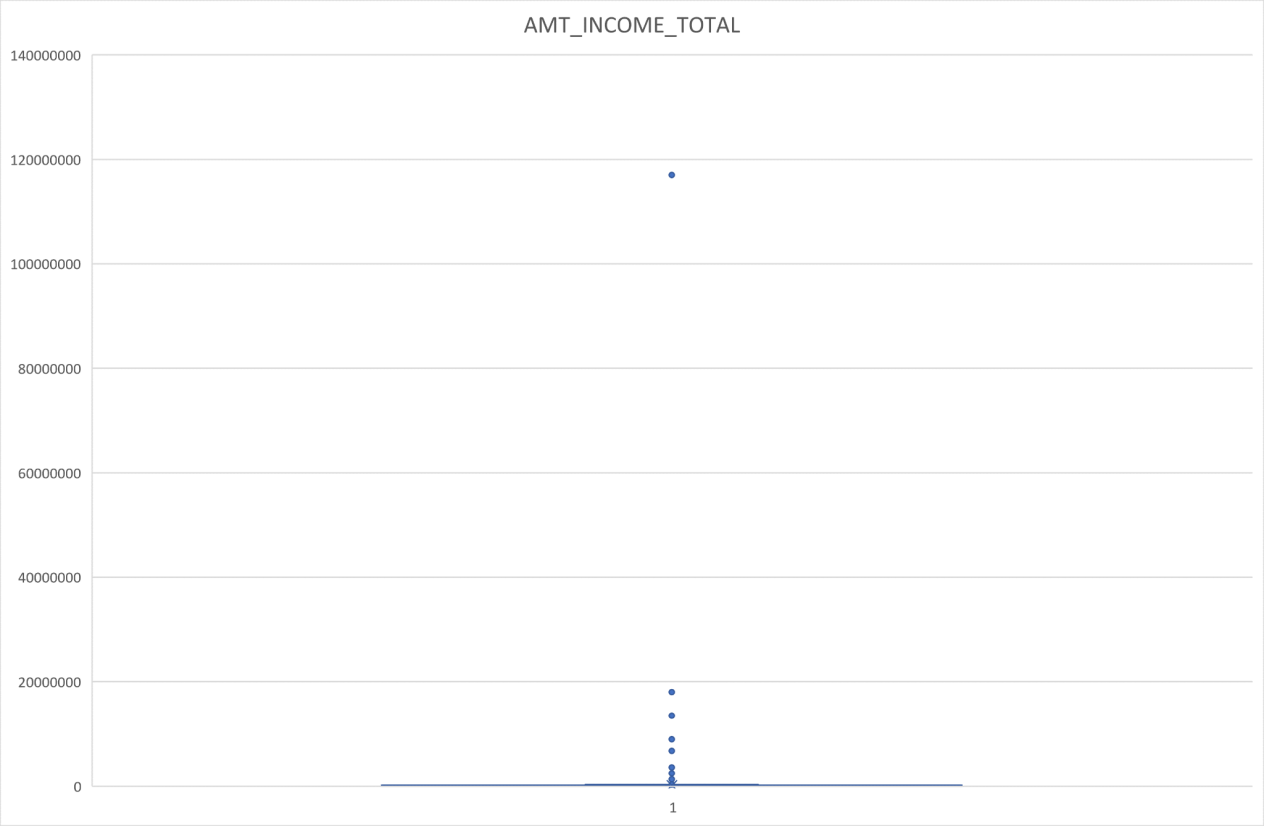
'DAYS\_BIRTH','DAYS\_EMPLOYED','DAYS\_REGISTRATION','DAYS\_ID\_PUBLISH','DAYS\_LAST\_PHONE\_CHANGE'

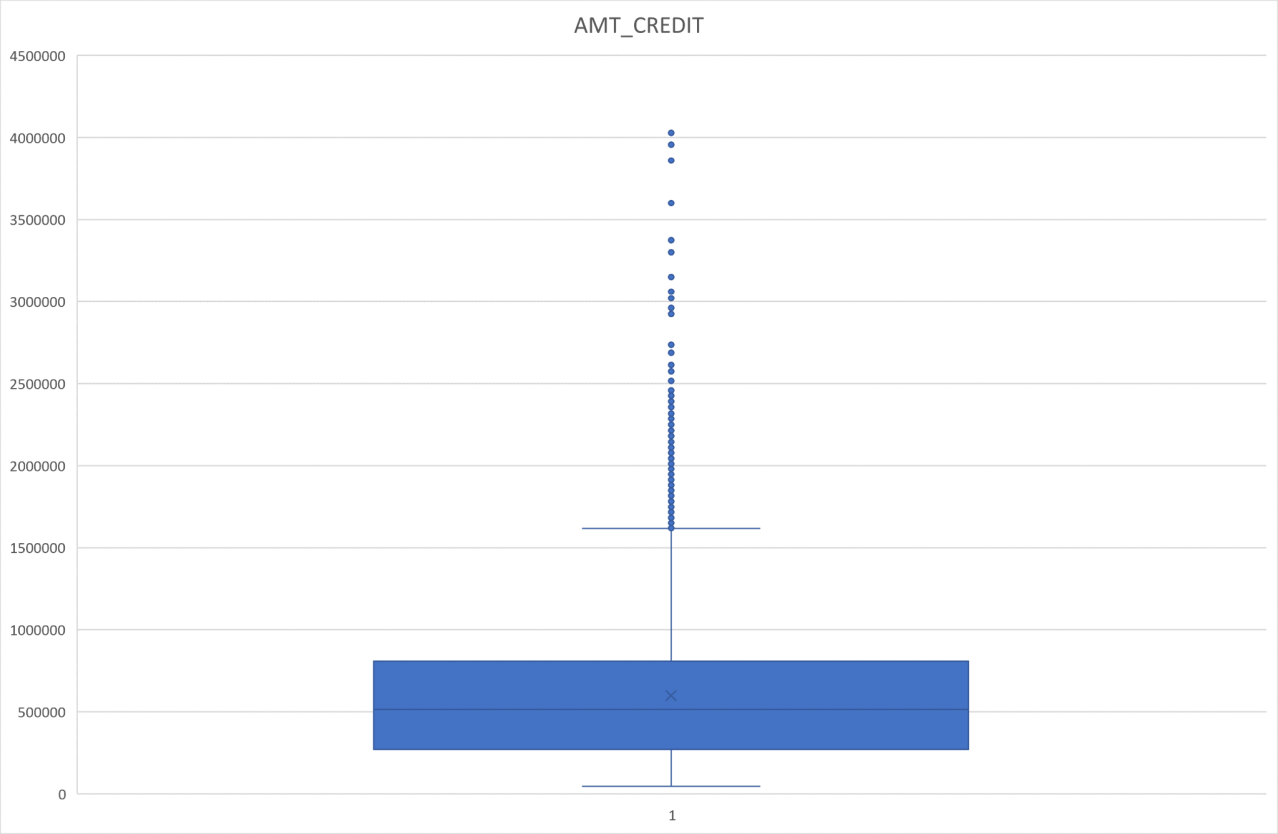
We have created alternate columns next to these columns by prefixing M to all these names, Now, these new columns will contain +ve values by using following formula-

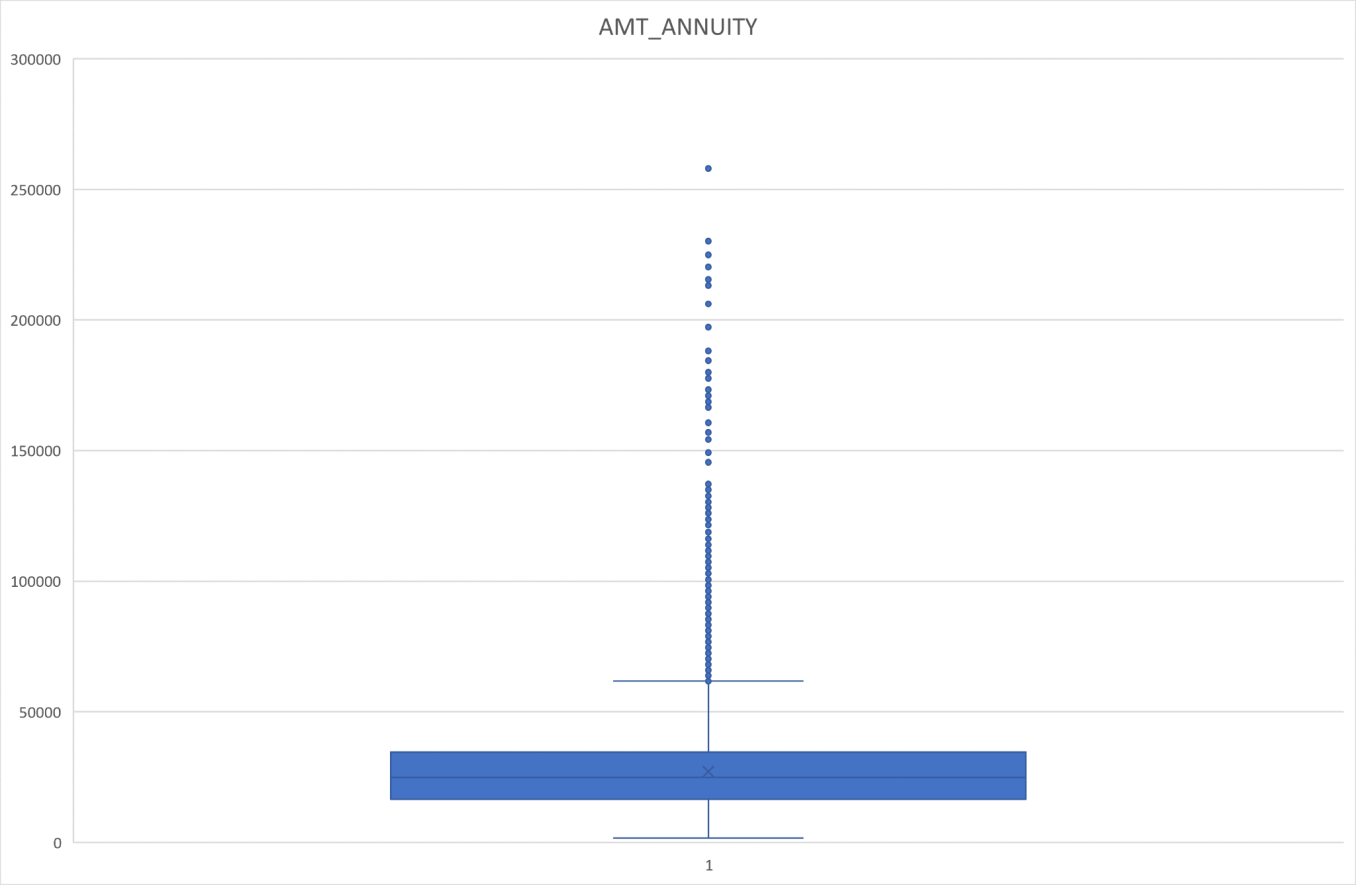
=IF([@[DAYS\_BIRTH]]<0,[@[DAYS\_BIRTH]]\*-1,[@[DAYS\_BIRTH]])

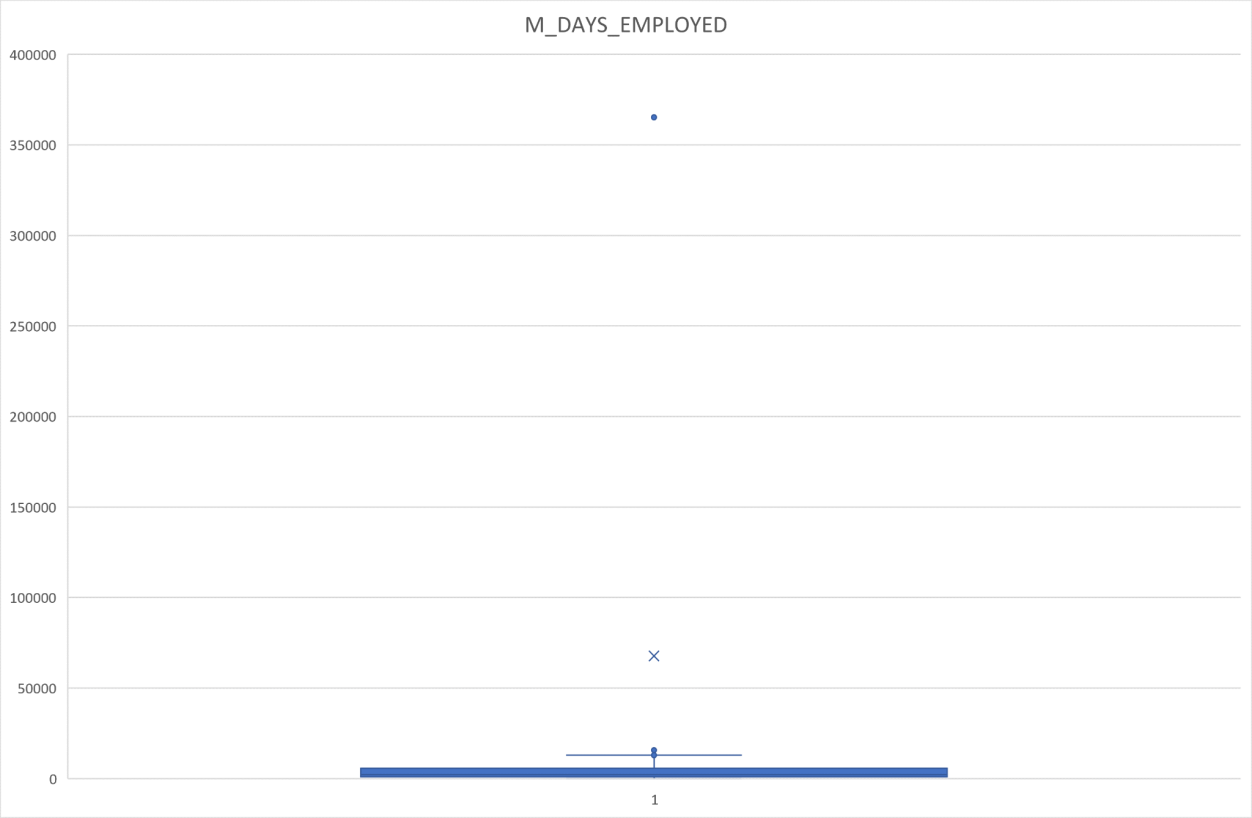
Making Box Plot of our following Numerical column- 'CNT\_CHILDREN', 'AMT\_INCOME\_TOTAL', 'AMT\_CREDIT', 'AMT\_ANNUITY', 'M\_DAYS\_EMPLOYED', 'M\_DAYS\_REGISTRATION'

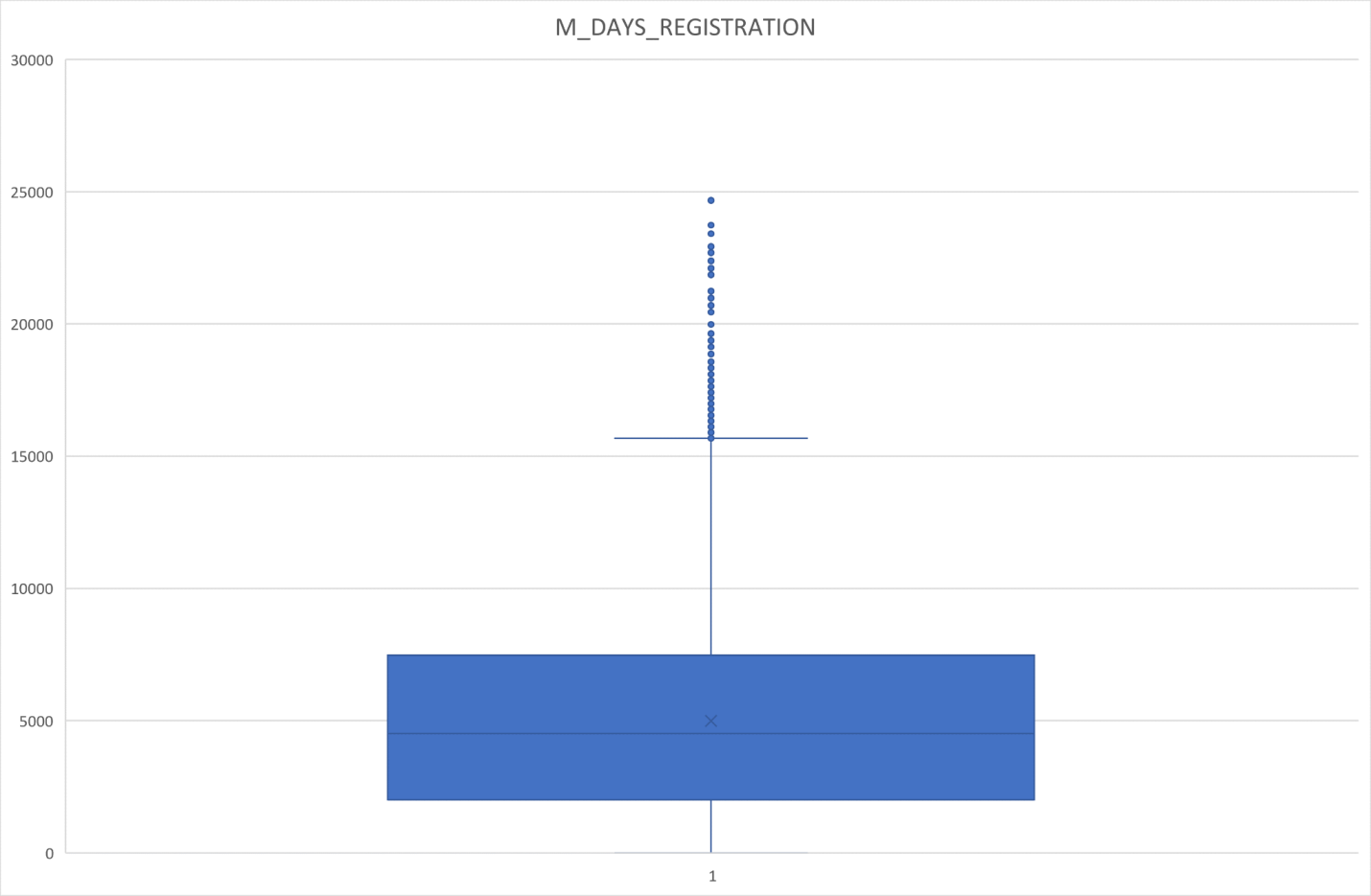












From the above box plot and describe analysis we found that following are the numeric columns are having outliers:

CNT\_CHILDREN, AMT\_INCOME\_TOTAL,AMT\_CREDIT,AMT\_ANNUITY,M\_DAYS\_EMPLOYED, M\_DAYS\_REGISTRATION

1. The first quartile is almost missing for CNT\_CHILDREN that means most of the data are present in the first quartile.
2. There is a single high value data point as an outlier present in AMT\_INCOME\_TOTAL and removing this point will drastically impact the box plot for further analysis.
3. The first quartile is slim compared to the third quartile for AMT\_CREDIT,AMT\_ANNUITY, DAYS\_EMPLOYED, DAYS\_REGISTRATION. This mean data is skewed towards the first quartile.

# 

# Analysis

Insights from number of target values(0 & 1) -

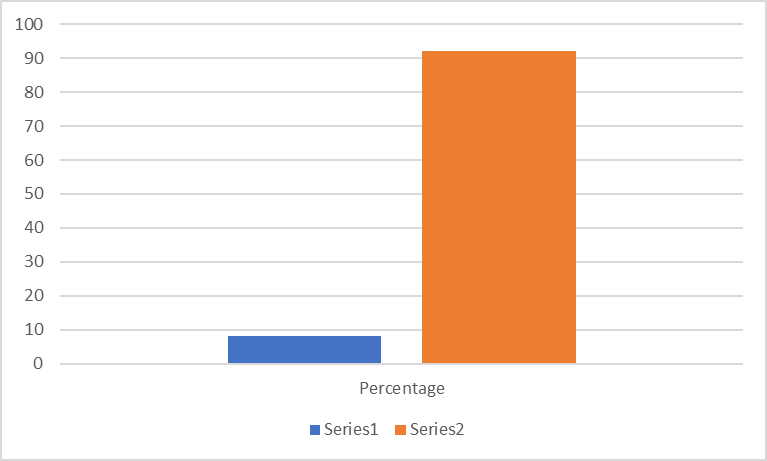
|  |  |  |
| --- | --- | --- |
| Categorical Values | Count | Percentage |
| 1 | 24825 | 8.072881946 |
| 0 | 282686 | 91.92711805 |

Calculating Imbalance percentage

Since the majority is 0 and minority is 1

Imbalance Ratio = 91.92711805/8.072881946= 11.38715

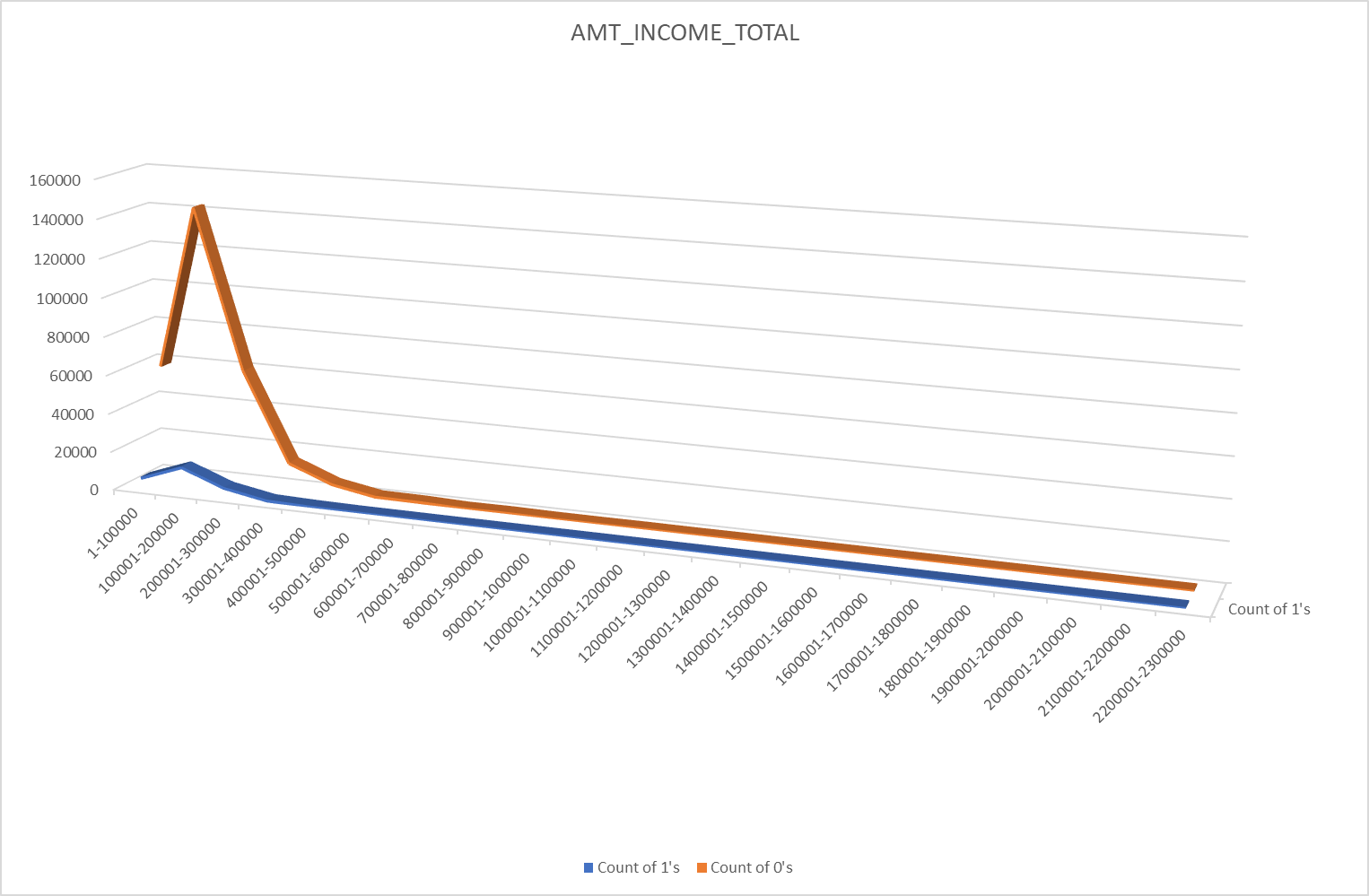
We can graphically represent this-



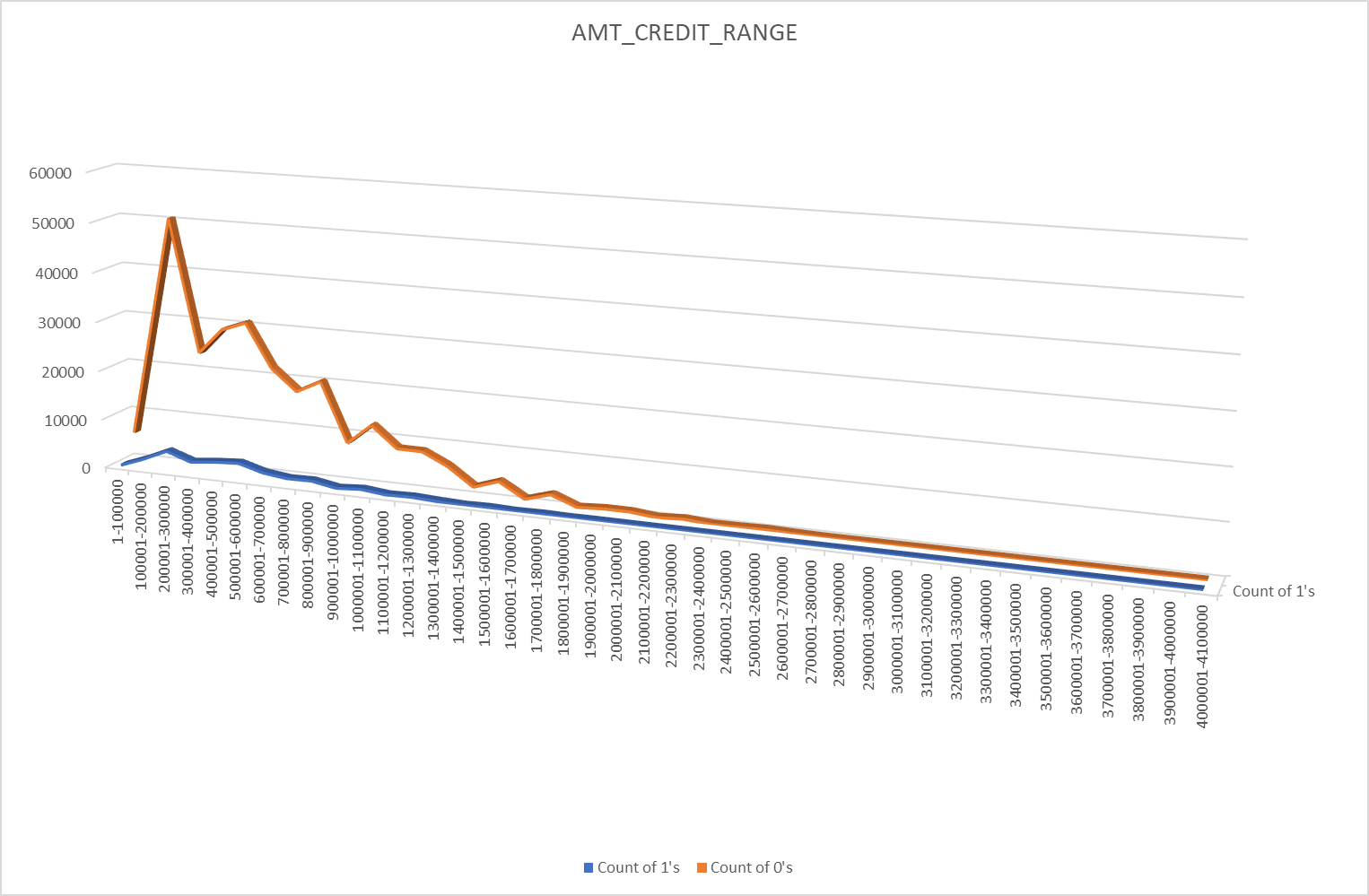
# Univariate analysis

Categorical Univariate Analysis for target= 1 vs 0 (client with payment difficulties vs no payment difficulties)

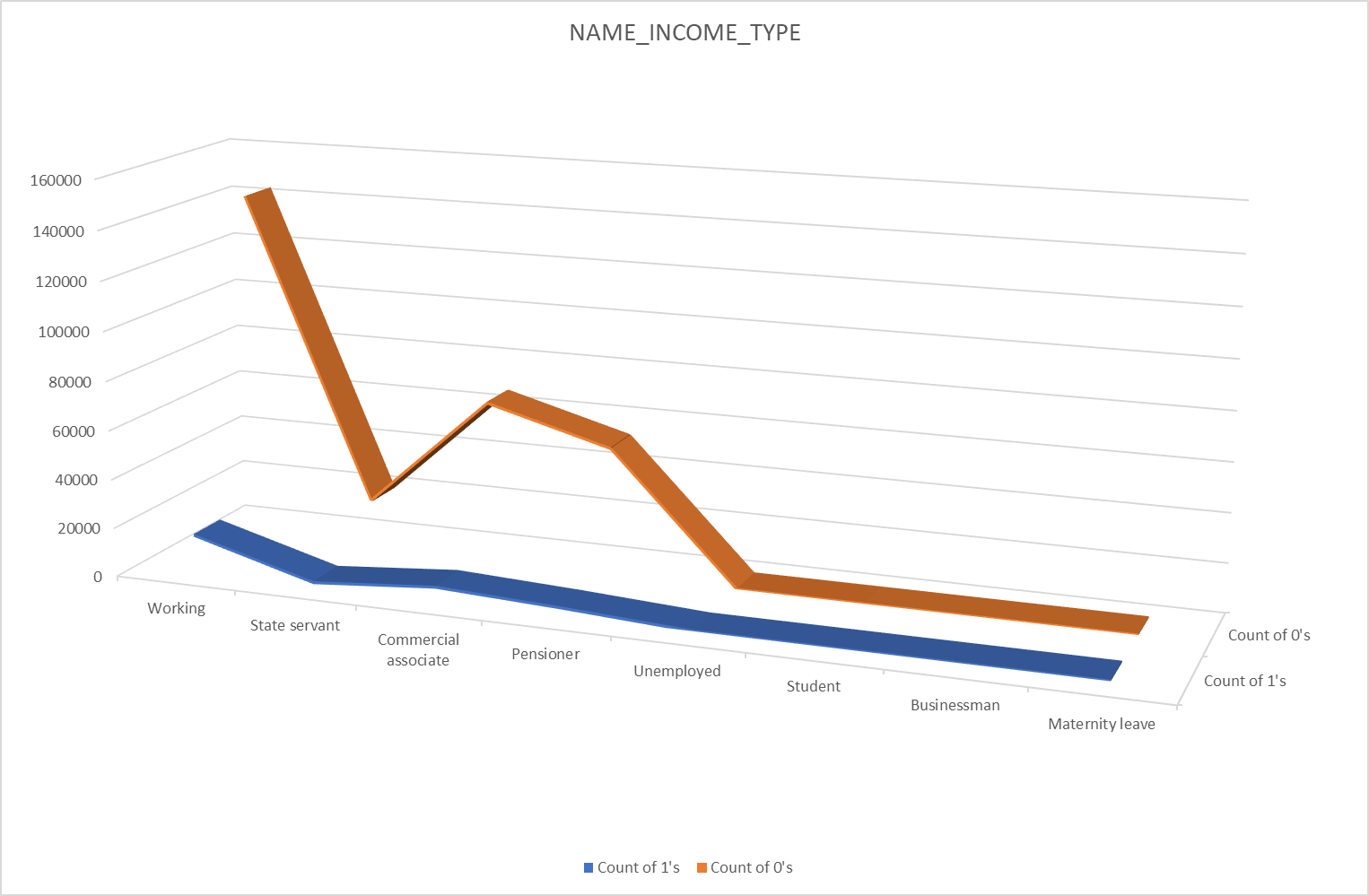
We will group the AMT\_INCOME\_TOTAL by 100000 and show the counts of 0 & 1’s on our chart-



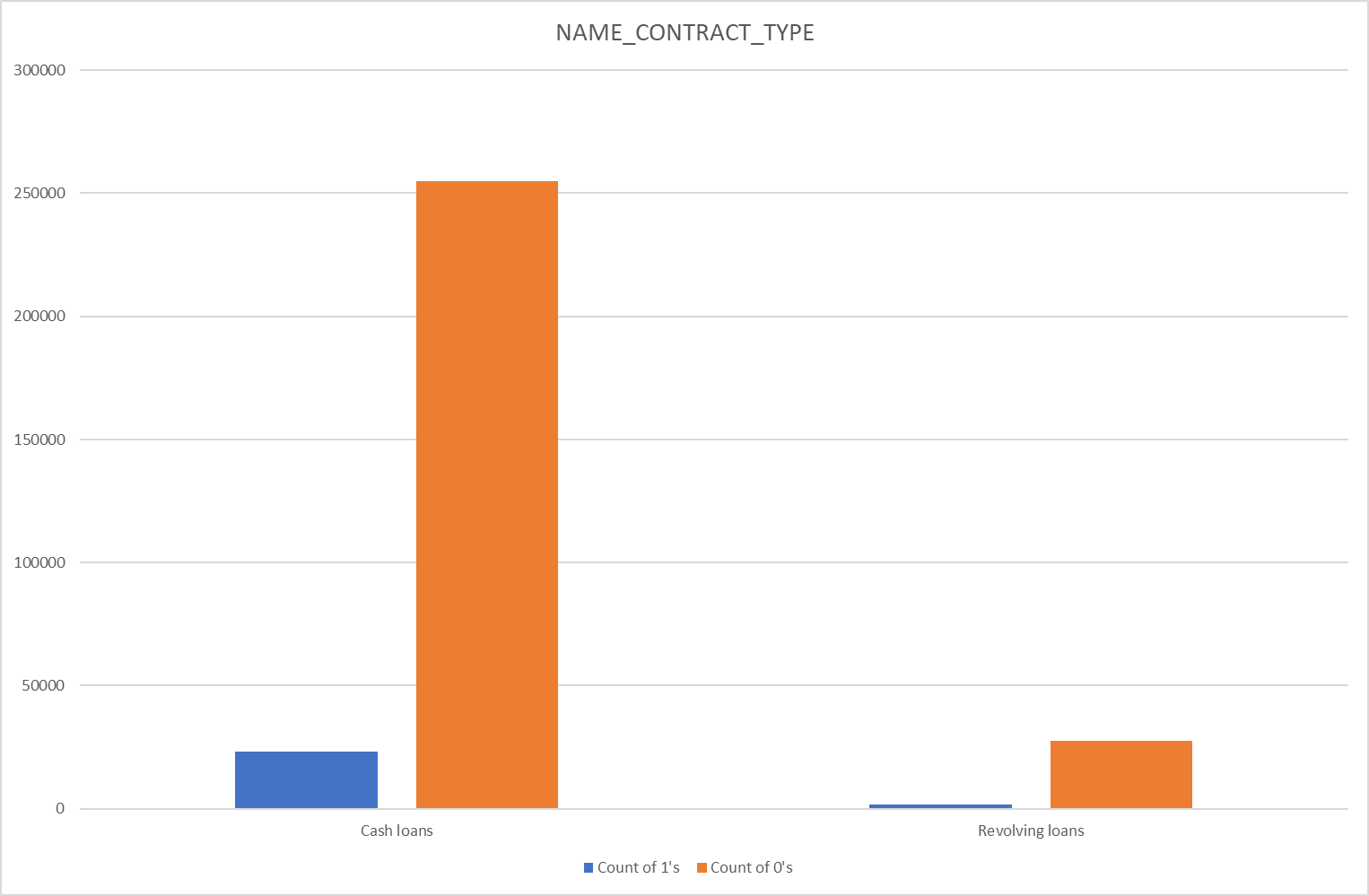
The above graph shows Income of the clients vs payment difficulties



The above graph shows Credit Amount of Loans vs payment difficulties as 0’s and 1’s.



The above graph shows income type of clients vs payment difficulties.



The above graph shows the contract type (revolving or cash) vs client difficulties in payment.

# Summary

AMT\_INCOME\_RANGE :

\* The people having 100000-200000 are having higher number of loan and also having higher in defaulter

\* The income segment having >500000 are having less defaulters.

AMT\_CREDIT\_RANGE:

\* The people having <100000 loan are less defaulter.

\* income having more than >100000 are almost equal % of loan defaulter

NAME\_INCOME\_TYPE:

\* Student pensioners and businesses have a higher percentage of loan repayment.

\* Working, State servant and Commercial associates have a higher default percentage.

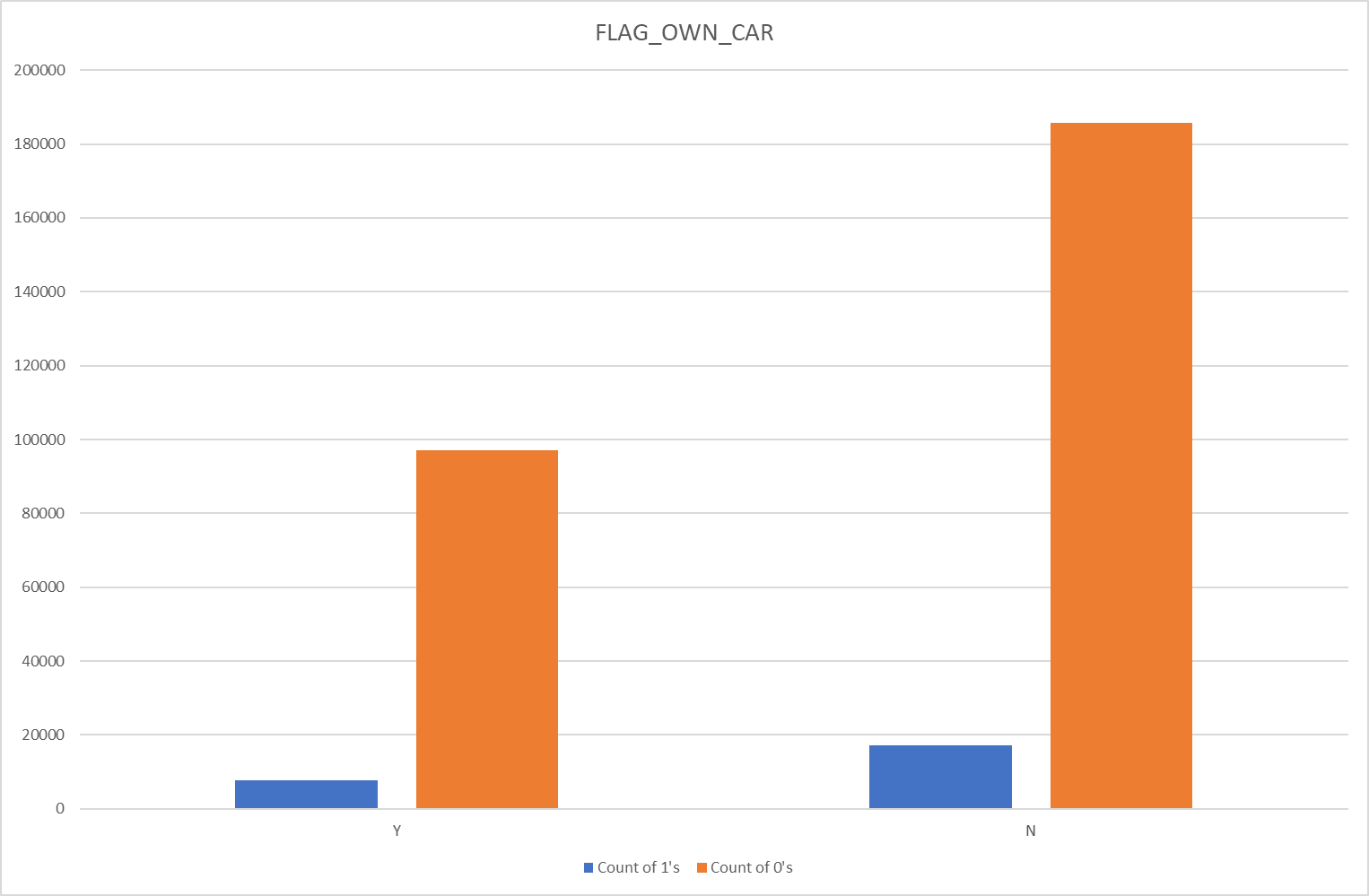
\* Maternity category is significantly higher in repayment.

NAME\_CONTRACT\_TYPE

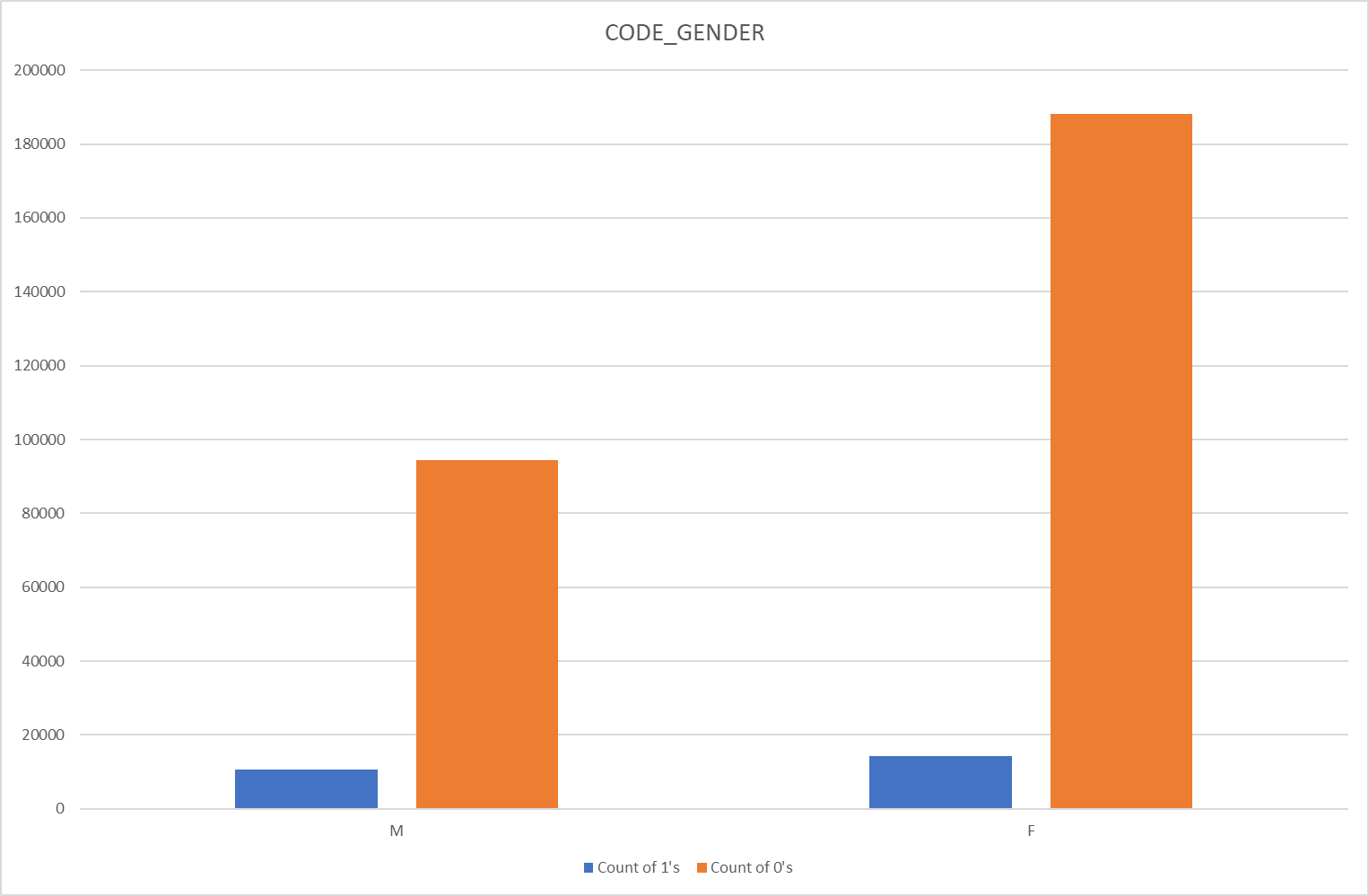
\* The contract type ‘cash loans’ has a higher number of credits than ‘Revolving loans’ contract type.

\* From the above graphs we can see that the Revolving loans are small compared to Cash loans but the % of non payment for the revolving loans are comparatively high.

# Categorical Univariate Analysis in Value scale



The above graph shows the client's difficulty with payment vs whether they have a car or not ( Represented as Y or N).



The above graph shows clients gender vs payment difficulties.

# Summary

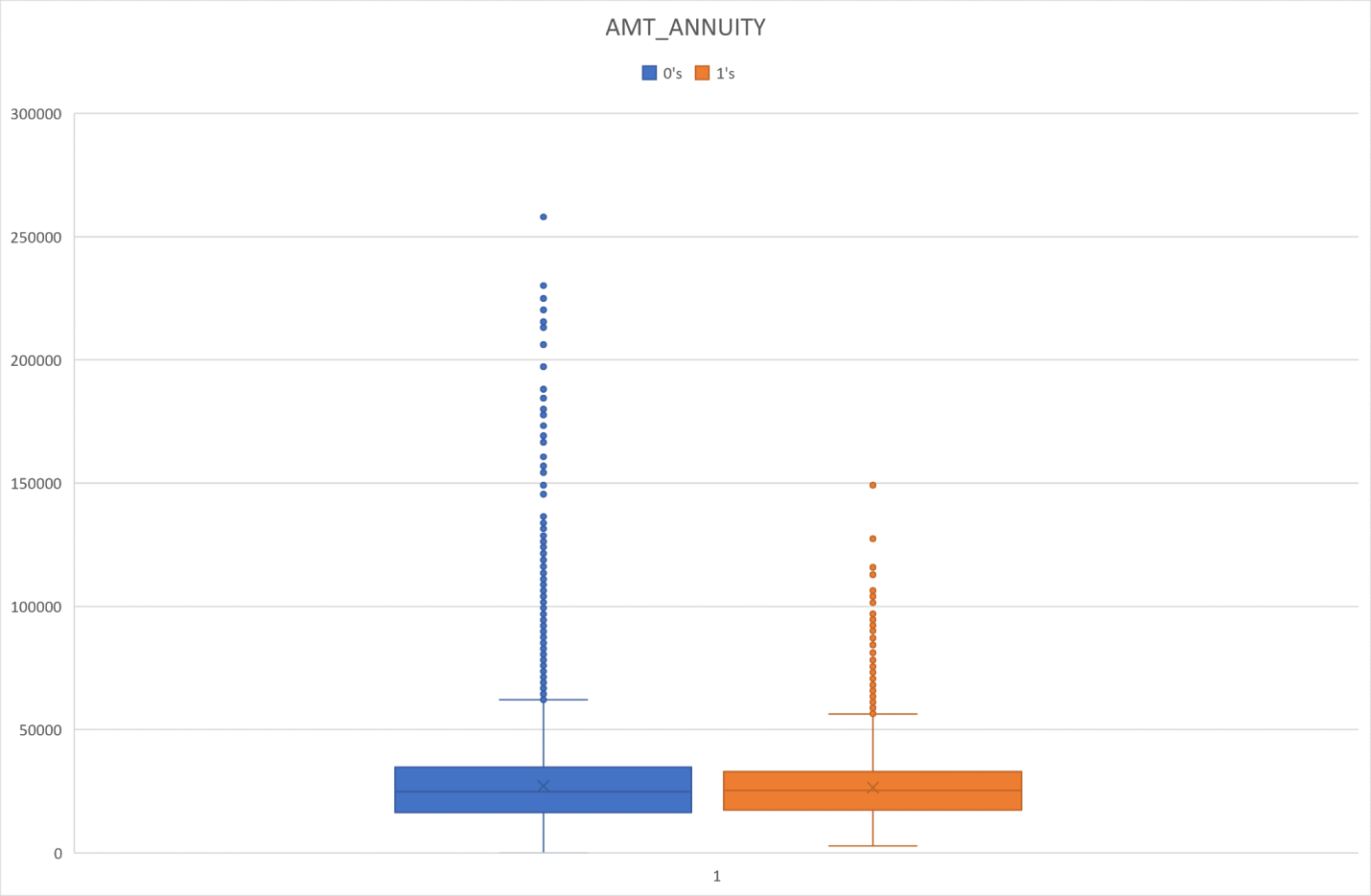
CODE\_GENDER:

The percentage of defaulters are more in Male than Female

FLAG\_OWN\_CAR:

The person owning a car is having a higher percentage of defaulters.

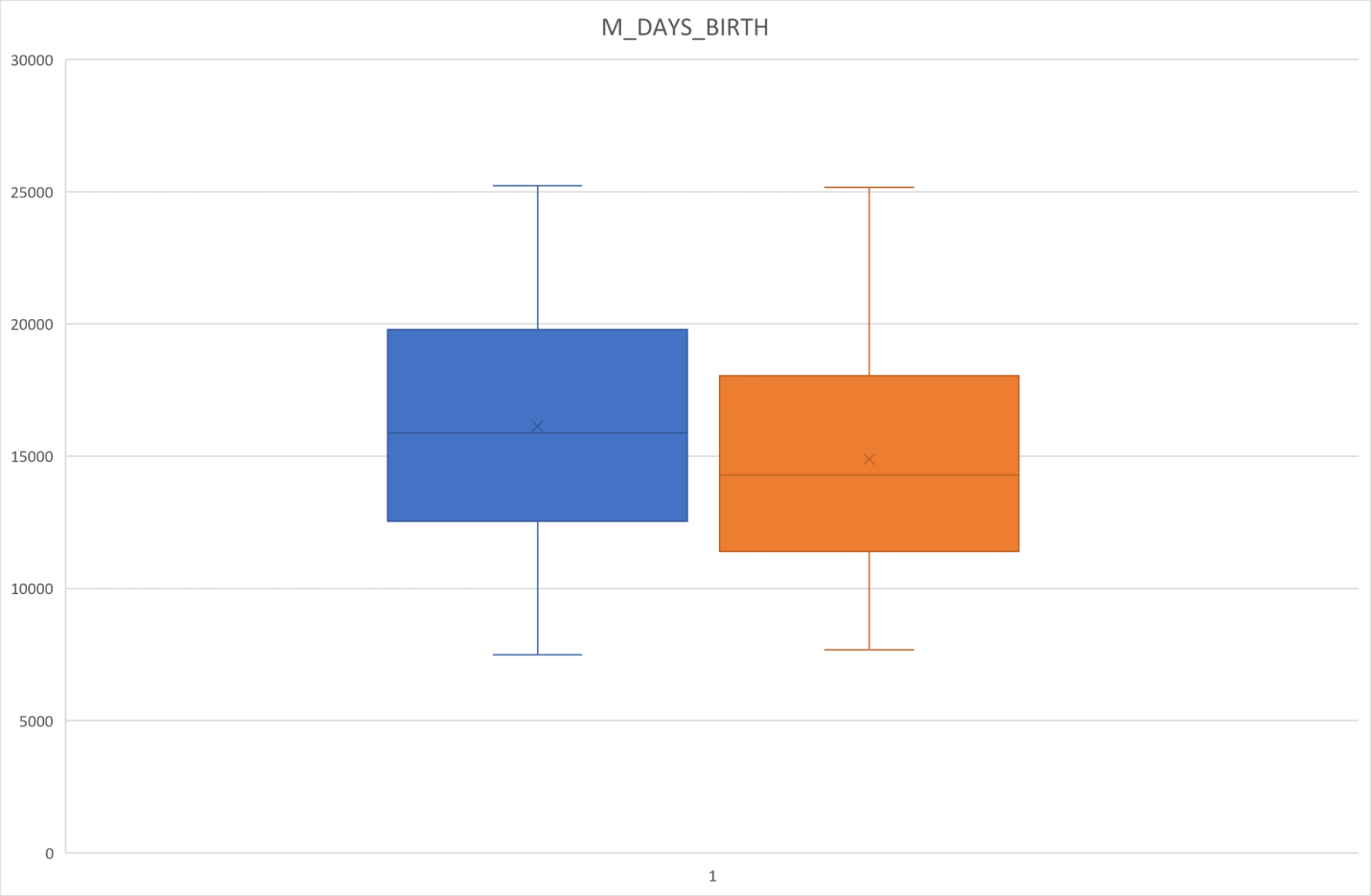
# Univariate Analysis for continuous variable



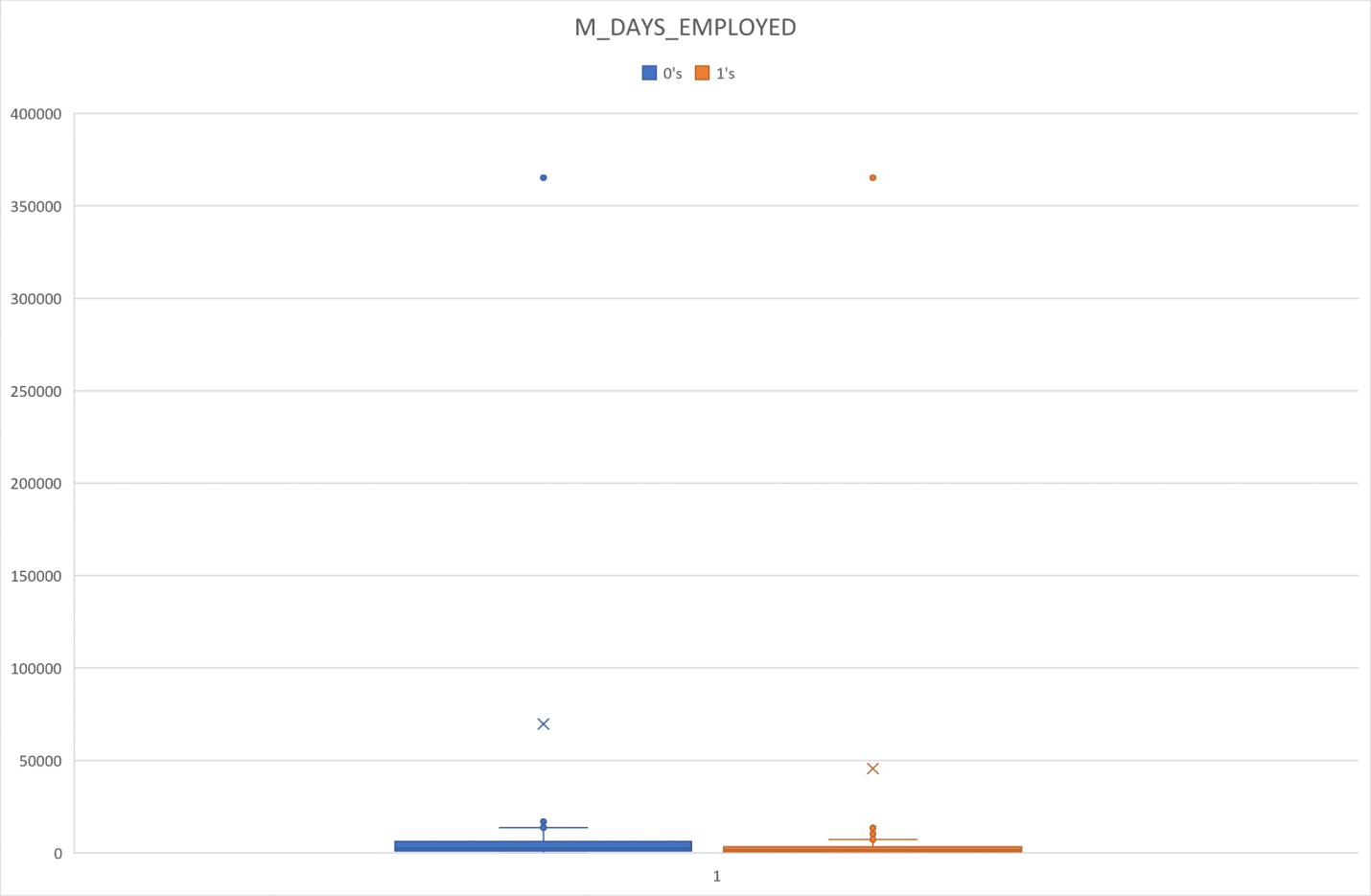
The above box plot shows the client's difficulty vs Loan Annuity.



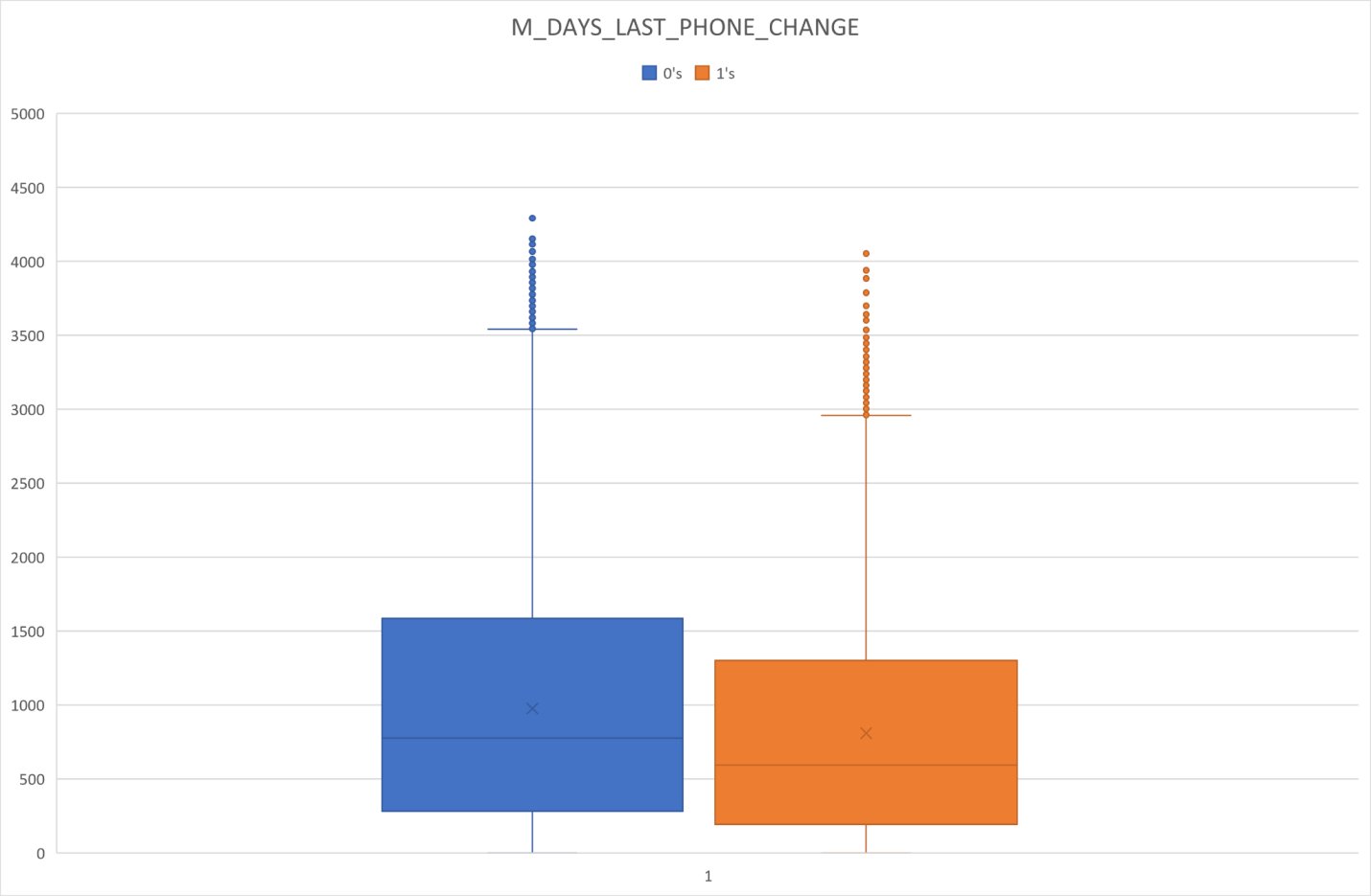
The above graph shows the client's difficulties vs price of goods for which loan is taken.



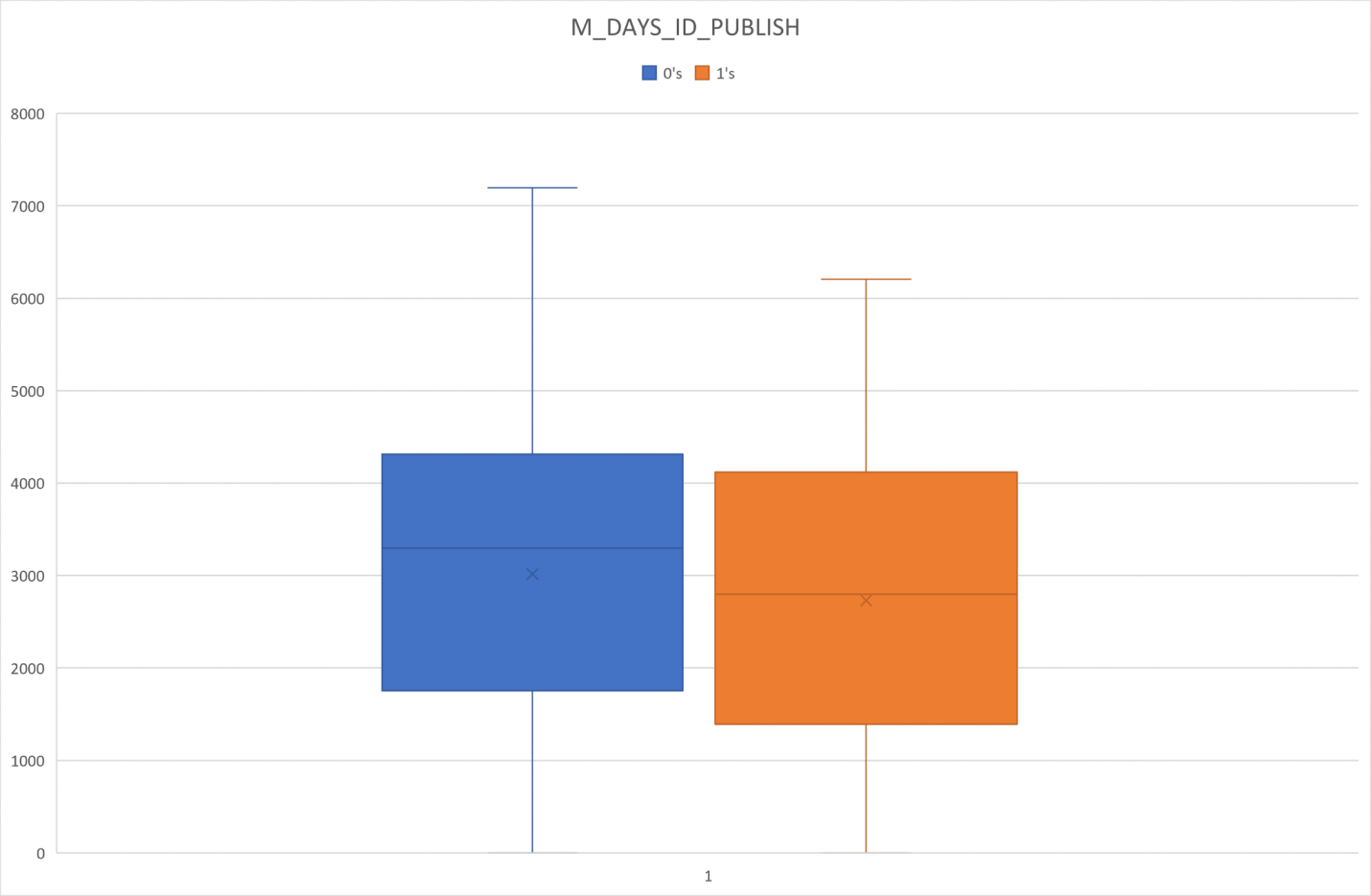
The above graph shows payment difficulty of clients vs age in days.



The above graph shows the client's payment difficulties vs how many days before the application the person started current employment.



The above graph shows the client's difficulties vs how many days before submitting the application did the client change the phone.



The above graph shows the client's difficulties vs how many days before the application the client changed the identity document with which he applied for the loan.

# Summary

M\_Days\_Birth: The people having higher age are having higher probability of repayment.

Some outliers are observed in In 'AMT\_ANNUITY','AMT\_GOODS\_PRICE','DAYS\_EMPLOYED', M\_DAYS\_LAST\_PHONE\_CHANGE in the dataset.

Less outlier observed in Days\_Birth and M\_DAYS\_ID\_PUBLISH

The 1st quartile is smaller than the third quartile in In 'AMT\_ANNUITY','AMT\_GOODS\_PRICE', M\_DAYS\_LAST\_PHONE\_CHANGE.

In DAYS\_ID\_PUBLISH: people changing ID in recent days are relatively prone to be default.

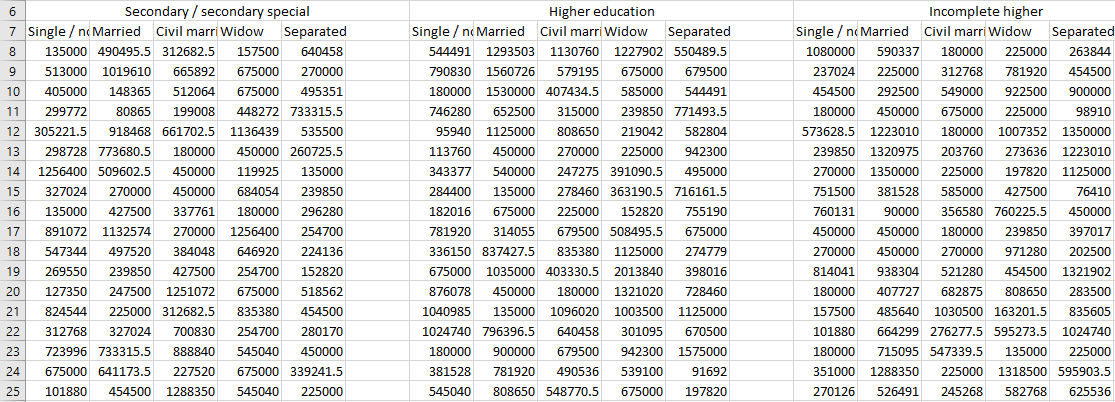
There is a single high value data point as an outlier present in M\_DAYS\_EMPLOYED. Removal this point will drastically impact the box plot for further analysis.

# Bivariate analysis for numerical variables

### For Target 0

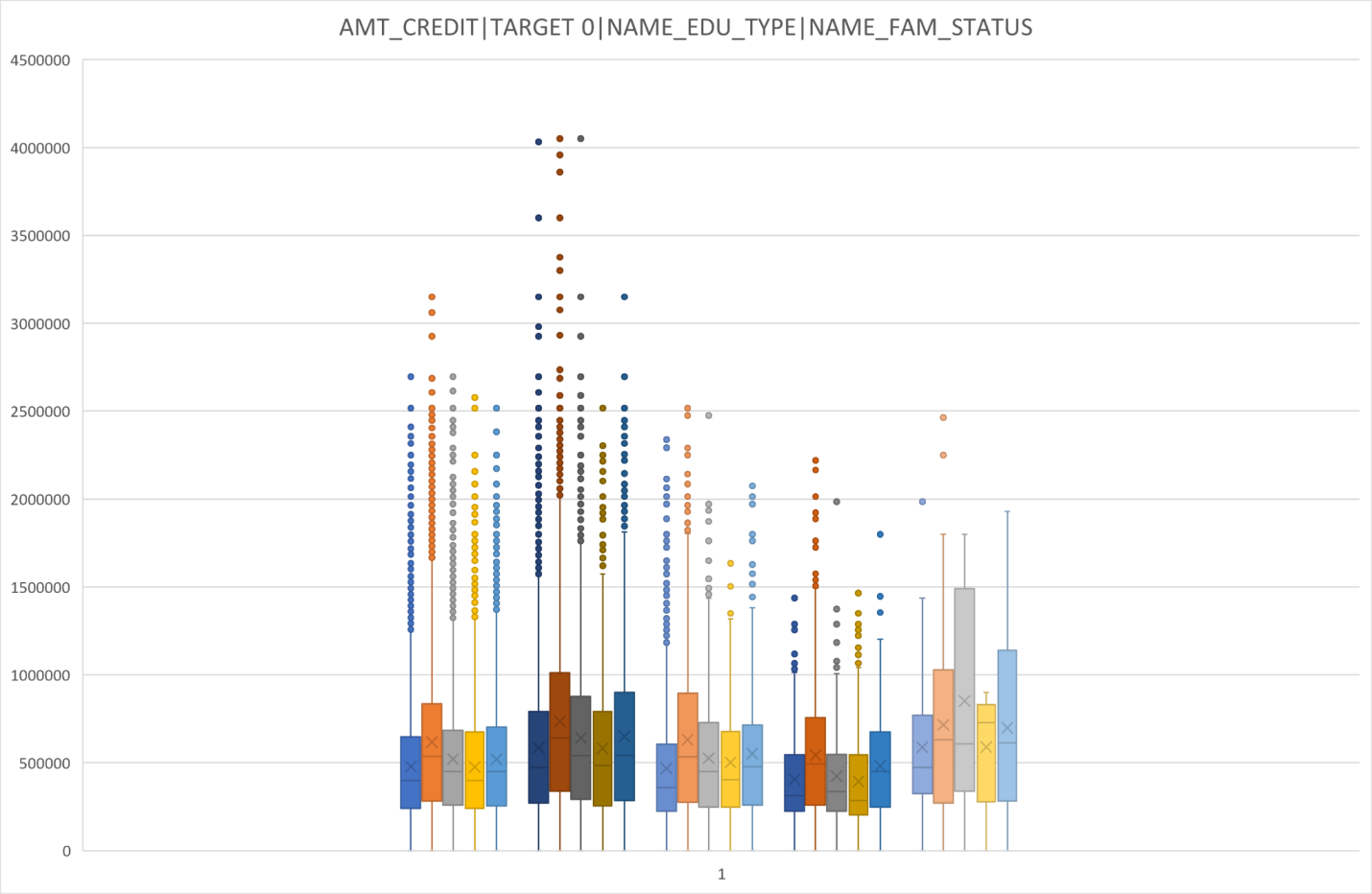
Box plotting for Credit amount-

We have extracted a new group of data using filter function-



=FILTER(application\_data[AMT\_CREDIT],(application\_data[TARGET]=0)\*(application\_data[NAME\_EDUCATION\_TYPE]=Sheet9!C6)\*(application\_data[NAME\_FAMILY\_STATUS]=Sheet9!C7))

The below chart shows box plot of AMT\_CREDIT for Target 0’s based upon Education Type and Family Status-



Legends (From Left) -

Group 1: Secondary/Secondary Special

Group 2: Higher Education

Group 3: Incomplete Higher

Group 4: Lower Secondary

Group 5: Academic Degree

Colours-

Dark Blue: Single/ Unmarried

Orange: Married

Gray: Civil Marriage

Yellow: Widow

Light Blue: Separated

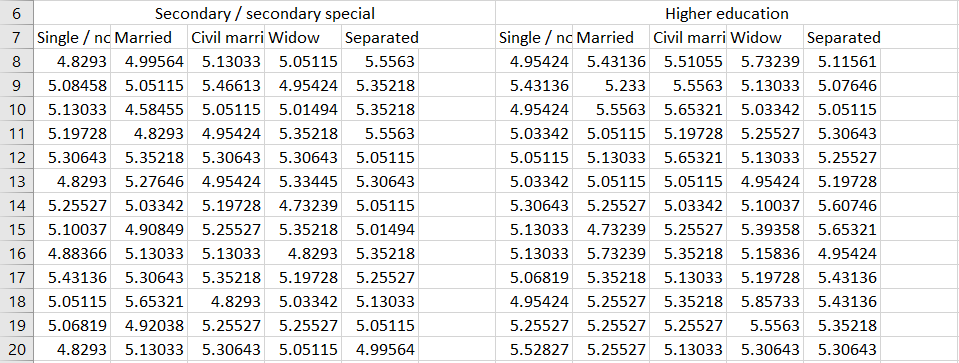
# Summary

Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.

Also, higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers. Civil marriage for Academic degree is having most of the credits in the third quartile.

# Box plotting for Income amount in Logarithmic Scale

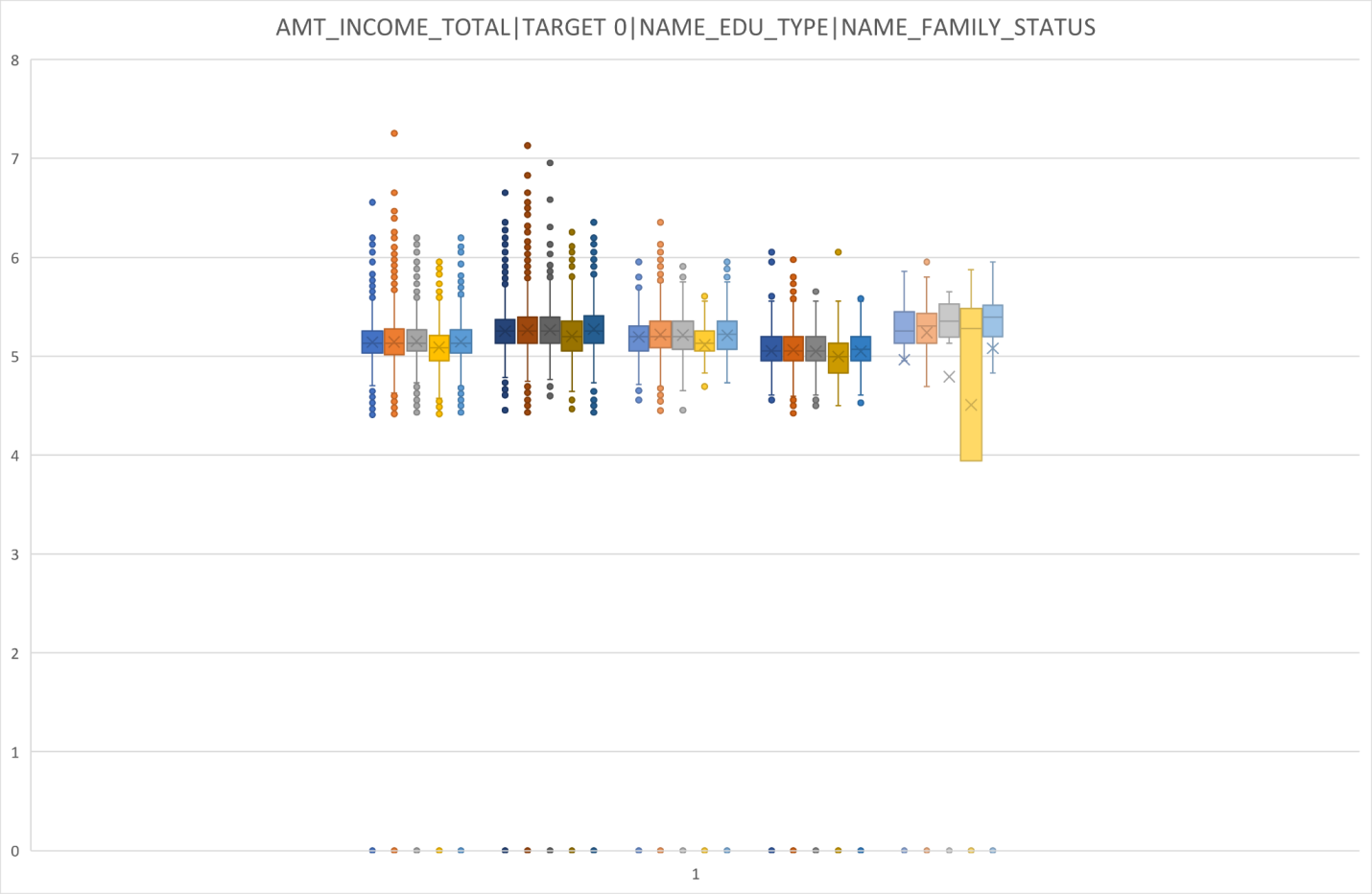
We have extracted logarithmic values of the total income of the clients for those who have no difficulty in payment (Target =0) based on their educational qualification and family status. The dataset we have created is shown below-



The formula we have used-

=LOG10(FILTER(application\_data[AMT\_INCOME\_TOTAL],(application\_data[TARGET]=0)\*(application\_data[NAME\_EDUCATION\_TYPE]=Sheet9!C6)\*(application\_data[NAME\_FAMILY\_STATUS]=Sheet9!C7)))

The box plot we have made based on our dataset-



Legends (From Left) -

Group 1: Secondary/Secondary Special

Group 2: Higher Education

Group 3: Incomplete Higher

Group 4: Lower Secondary

Group 5: Academic Degree

Colours-

Dark Blue: Single/ Unmarried

Orange: Married

Gray: Civil Marriage

Yellow: Widow

Light Blue: Separated

The scaling of Y axis is on power of 10 (5 -> )

# Summary

In Education type 'Higher education' the income amount is mostly equal with family status. It does contain many outliers.

Less outliers are for Academic degrees but their income amount is little higher than Higher education.

Lower secondary of civil marriage family status have less income than others.

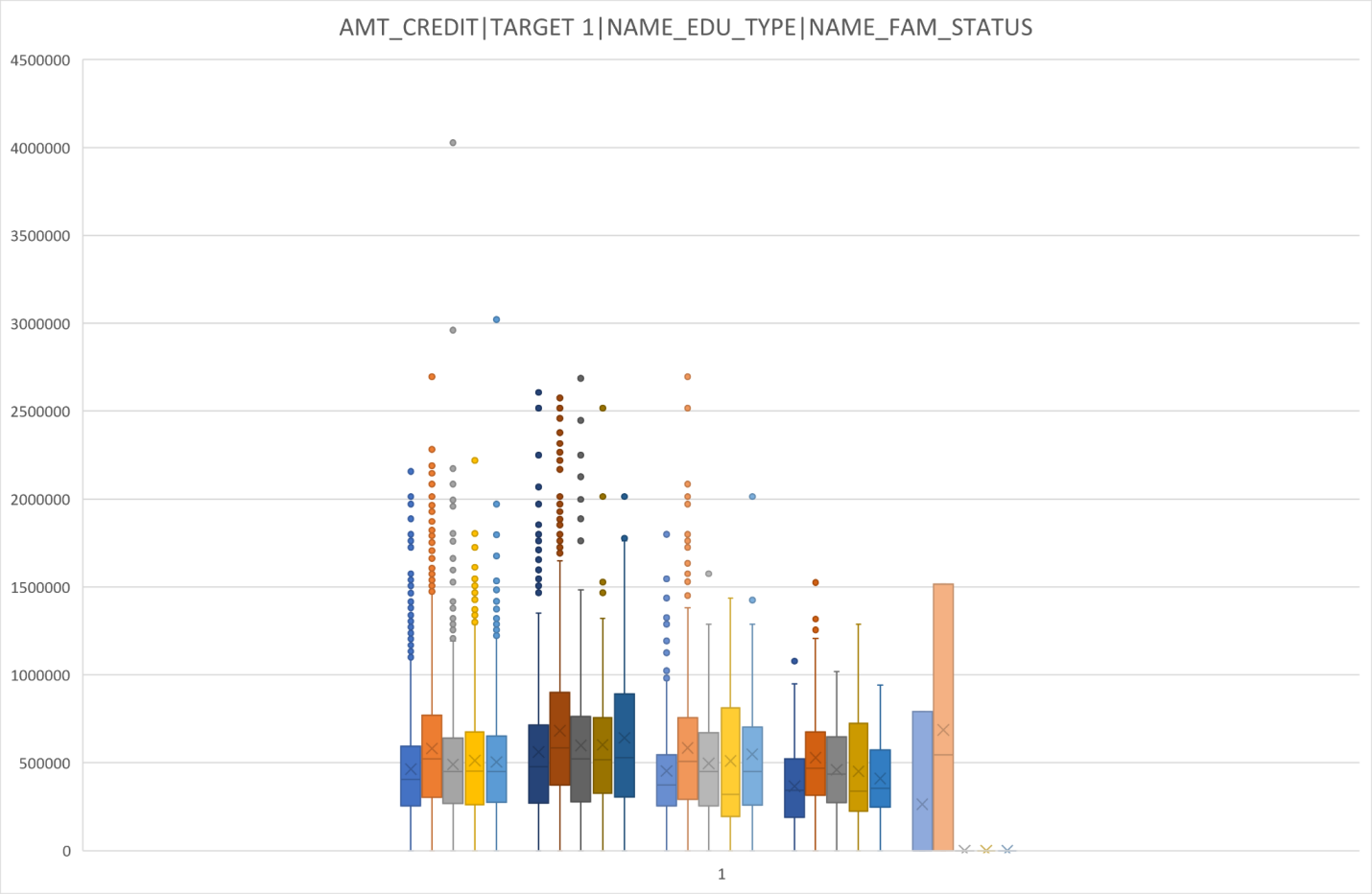
### For Target 1

## Box plotting for credit amount-

The following box plot shows the values of credit amount for those having payment difficulty (Target-1) based on educational qualification and family status.

We have used following formula-

=FILTER(application\_data[AMT\_CREDIT],(application\_data[TARGET]=C3)\*(application\_data[NAME\_EDUCATION\_TYPE]=Sheet9!C6)\*(application\_data[NAME\_FAMILY\_STATUS]=Sheet9!C7))



Legends (From Left) -

Group 1: Secondary/Secondary Special

Group 2: Higher Education

Group 3: Incomplete Higher

Group 4: Lower Secondary

Group 5: Academic Degree

Colours-

Dark Blue: Single/ Unmarried

Orange: Married

Gray: Civil Marriage

Yellow: Widow

Light Blue: Separated

# Summary

Observations are Quite similar with Target 0

Family status of 'married' and 'separated' of Higher education are having higher number of credits than others.

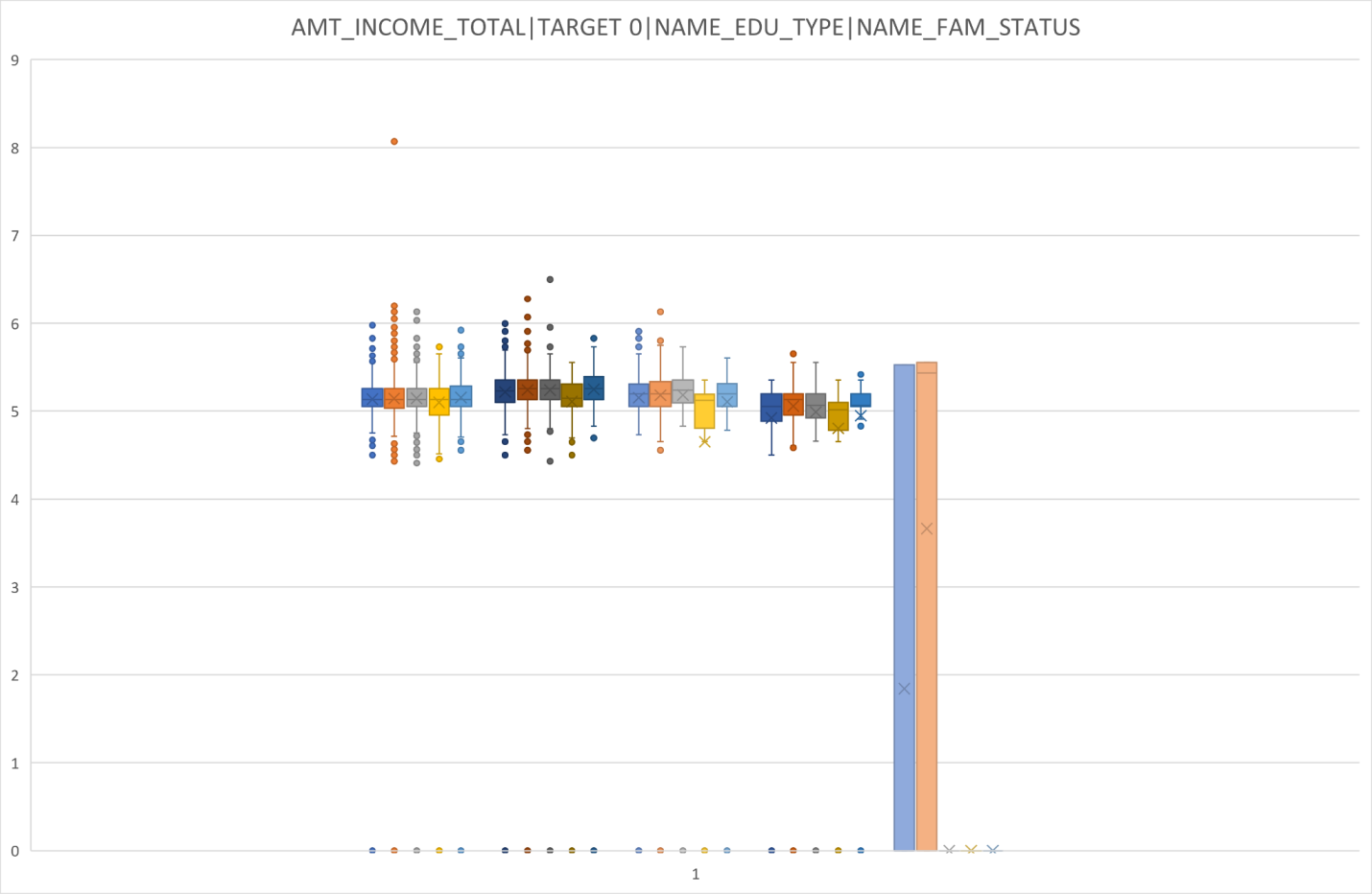
Most of the outliers are from Education type 'Higher education' and 'Secondary'.

## Box plotting for Income amount in logarithmic scale

The following box plot shows the values of income amount in logarithmic scale for those having payment difficulty (Target-1) based on educational qualification and family status.

We have used following formula-

=LOG10(FILTER(application\_data[AMT\_INCOME\_TOTAL],(application\_data[TARGET]=C3)\*(application\_data[NAME\_EDUCATION\_TYPE]=Sheet9!C6)\*(application\_data[NAME\_FAMILY\_STATUS]=Sheet9!C7)))



Legends (From Left) -

Group 1: Secondary/Secondary Special

Group 2: Higher Education

Group 3: Incomplete Higher

Group 4: Lower Secondary

Group 5: Academic Degree

Colours-

Dark Blue: Single/ Unmarried

Orange: Married

Gray: Civil Marriage

Yellow: Widow

Light Blue: Separated

The scaling of Y axis is on power of 10 (5 -> 105)

# Summary

There is also have some similarity with Target0,

Education type 'Higher education' the income amount is mostly equal with family status.

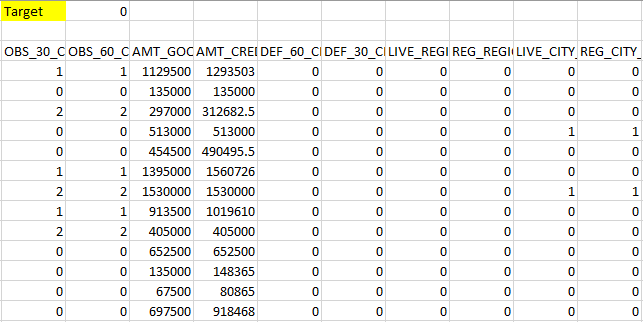
Less outliers are for Academic degrees but their income amount is little higher than Higher education.

Lower secondary students have less income than others.

# Correlation:

We prepared a dataset out of application data table-

=FILTER(application\_data[OBS\_30\_CNT\_SOCIAL\_CIRCLE],application\_data[TARGET]=G1)



Getting top 10 correlation between columns of Target- 0:

We used following formula-

=ROUND(CORREL(F4:F282689,G4:G282689),2)

The correlation table is as follows-

|  |  |  |
| --- | --- | --- |
| Col 1 | Col 2 | Correlation |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | OBS\_60\_CNT\_SOCIAL\_CIRCLE | 1 |
| AMT\_GOODS\_PRICE | AMT\_CREDIT | 0.99 |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | DEF\_30\_CNT\_SOCIAL\_CIRCLE | 0.86 |
| LIVE\_REGION\_NOT\_WORK\_REGION | REG\_REGION\_NOT\_WORK\_REGION | 0.86 |
| LIVE\_CITY\_NOT\_WORK\_CITY | REG\_CITY\_NOT\_WORK\_CITY | 0.83 |
| AMT\_GOODS\_PRICE | AMT\_ANNUITY | 0.78 |
| AMT\_ANNUITY | AMT\_CREDIT | 0.77 |
| M\_Days\_Employed | M\_Days\_Birth | 0.63 |
| REG\_REGION\_NOT\_WORK\_REGION | REG\_REGION\_NOT\_LIVE\_REGION | 0.45 |
| REG\_CITY\_NOT\_WORK\_CITY | REG\_CITY\_NOT\_LIVE\_CITY | 0.44 |

Top 10 correlated columns: target 1

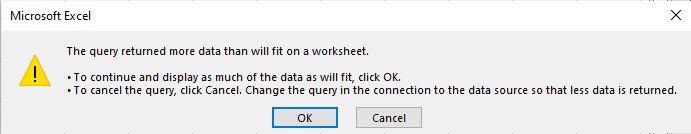
|  |  |  |
| --- | --- | --- |
| Col 1 | Col 2 | Correlation |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | OBS\_60\_CNT\_SOCIAL\_CIRCLE | 1 |
| AMT\_GOODS\_PRICE | AMT\_CREDIT | 0.98 |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | DEF\_30\_CNT\_SOCIAL\_CIRCLE | 0.87 |
| LIVE\_REGION\_NOT\_WORK\_REGION | REG\_REGION\_NOT\_WORK\_REGION | 0.85 |
| LIVE\_CITY\_NOT\_WORK\_CITY | REG\_CITY\_NOT\_WORK\_CITY | 0.78 |
| AMT\_GOODS\_PRICE | AMT\_ANNUITY | 0.75 |
| AMT\_ANNUITY | AMT\_CREDIT | 0.75 |
| M\_Days\_Employed | M\_Days\_Birth | 0.58 |
| REG\_REGION\_NOT\_WORK\_REGION | REG\_REGION\_NOT\_LIVE\_REGION | 0.5 |
| REG\_CITY\_NOT\_WORK\_CITY | REG\_CITY\_NOT\_LIVE\_CITY | 0.47 |

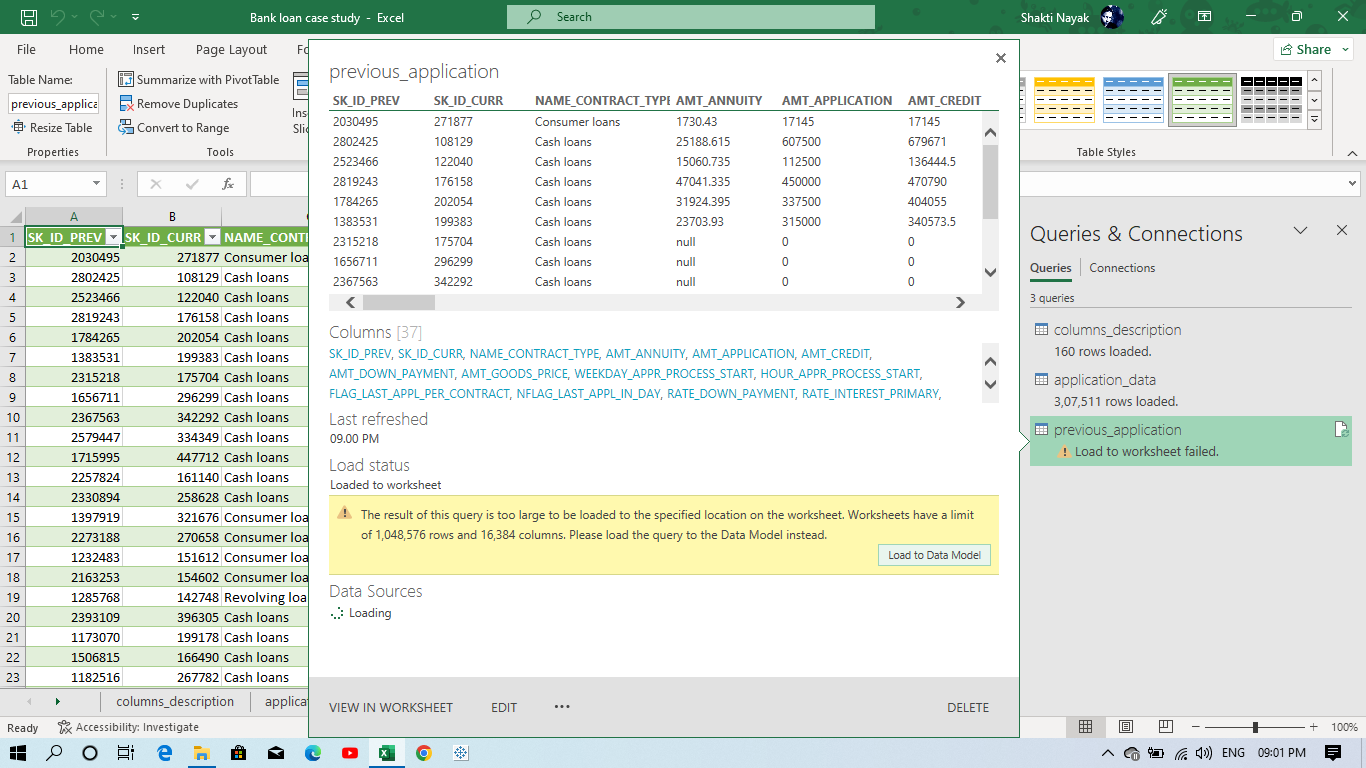
# Summary

From the above correlation analysis it is inferred that the highest correlation (1.0) is between (OBS\_60\_CNT\_SOCIAL\_CIRCLE with OBS\_30\_CNT\_SOCIAL\_CIRCLE) which is the same for both the data set.

# Importing Previous Application Data

We have encountered several issues while importing Previous Application Data because the importing file contains so many rows beyond excel’s limitation.





So, our imported data contains rows within excel’s limitation. That means some of our data is lost and cannot be imported.

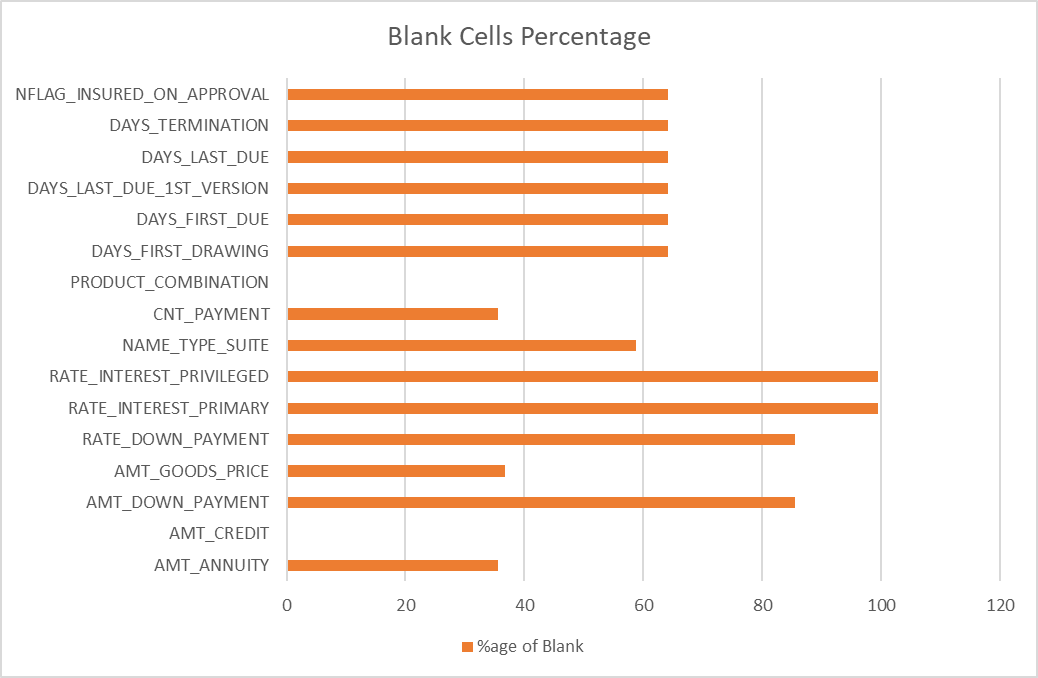
# Read Previous Application data and merging with application data

## Identifying columns only with null values

The below table shows columns with percentage of blank cells on the Previous Application dataset.

|  |  |  |
| --- | --- | --- |
| Col Name | Count Blank | %age of Blank |
| AMT\_ANNUITY | 372232 | 35.49884367 |
| AMT\_CREDIT | 1 | 9.53675E-05 |
| AMT\_DOWN\_PAYMENT | 895844 | 85.43442291 |
| AMT\_GOODS\_PRICE | 385515 | 36.76561047 |
| RATE\_DOWN\_PAYMENT | 895844 | 85.43442291 |
| RATE\_INTEREST\_PRIMARY | 1043561 | 99.52182724 |
| RATE\_INTEREST\_PRIVILEGED | 1043561 | 99.52182724 |
| NAME\_TYPE\_SUITE | 616506 | 58.79464988 |
| CNT\_PAYMENT | 372230 | 35.49865293 |
| PRODUCT\_COMBINATION | 346 | 0.032997163 |
| DAYS\_FIRST\_DRAWING | 673065 | 64.18854159 |
| DAYS\_FIRST\_DUE | 673065 | 64.18854159 |
| DAYS\_LAST\_DUE\_1ST\_VERSION | 673065 | 64.18854159 |
| DAYS\_LAST\_DUE | 673065 | 64.18854159 |
| DAYS\_TERMINATION | 673065 | 64.18854159 |
| NFLAG\_INSURED\_ON\_APPROVAL | 673065 | 64.18854159 |

The below graph shows Columns with percentage of blank cells-



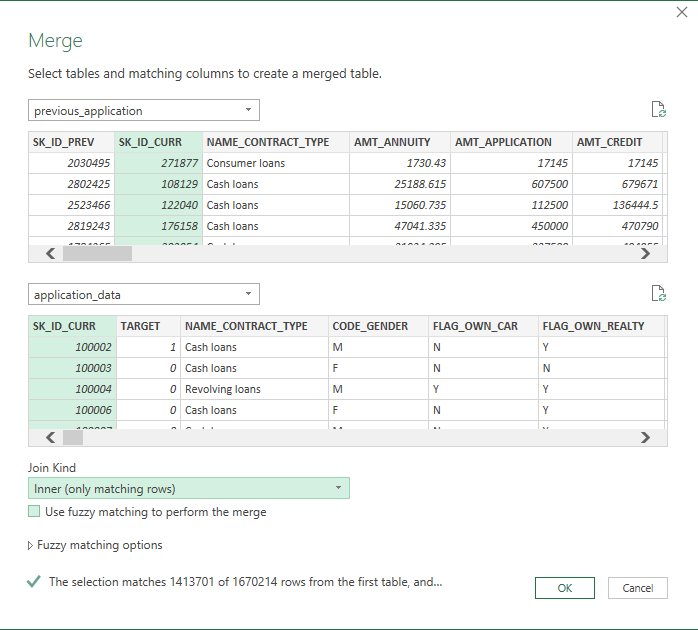
Columns having null value more than 50%-

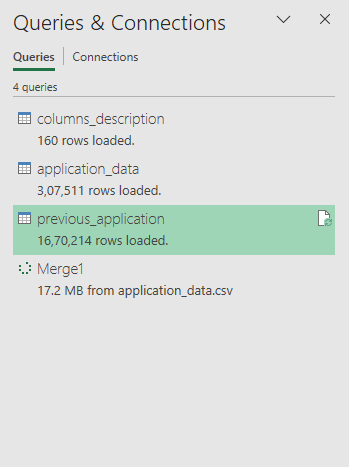
|  |  |  |
| --- | --- | --- |
| Col Name | Count Blank | %age of Blank |
| AMT\_DOWN\_PAYMENT | 895844 | 85.43442291 |
| RATE\_DOWN\_PAYMENT | 895844 | 85.43442291 |
| RATE\_INTEREST\_PRIMARY | 1043561 | 99.52182724 |
| RATE\_INTEREST\_PRIVILEGED | 1043561 | 99.52182724 |
| NAME\_TYPE\_SUITE | 616506 | 58.79464988 |
| DAYS\_FIRST\_DRAWING | 673065 | 64.18854159 |
| DAYS\_FIRST\_DUE | 673065 | 64.18854159 |
| DAYS\_LAST\_DUE\_1ST\_VERSION | 673065 | 64.18854159 |
| DAYS\_LAST\_DUE | 673065 | 64.18854159 |
| DAYS\_TERMINATION | 673065 | 64.18854159 |
| NFLAG\_INSURED\_ON\_APPROVAL | 673065 | 64.18854159 |

Now, we can drop all the columns from the Dataset for which the missing value percentage is more than 50%.

# Merging

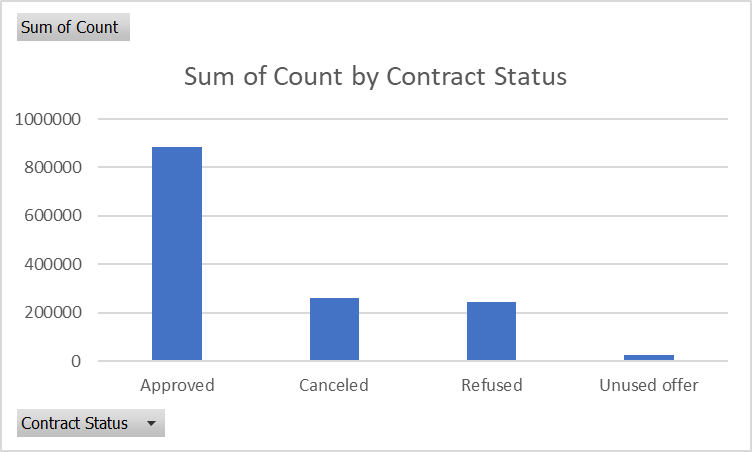
Now, we merge the previous data and application data table into one based on inner merging of SK\_ID\_CURR. Below is the screenshot of merging-



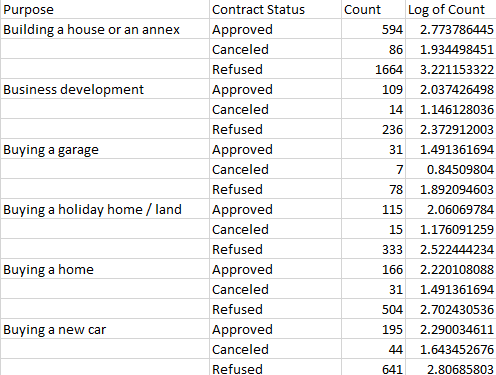


# Distribution of contract status with purposes in logarithmic scale

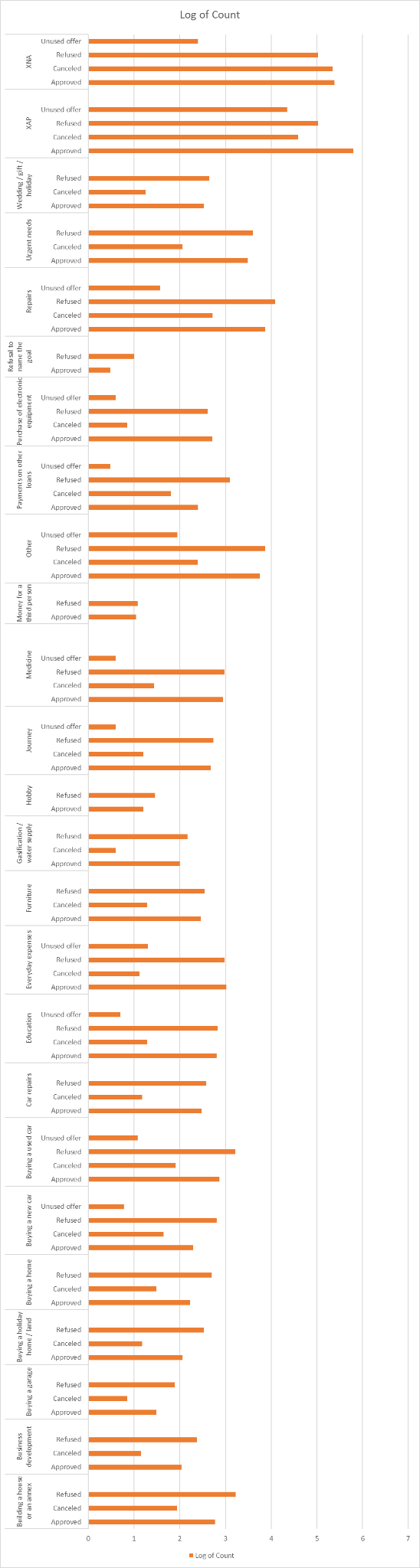
The following graph shows number of applications based on contract status-

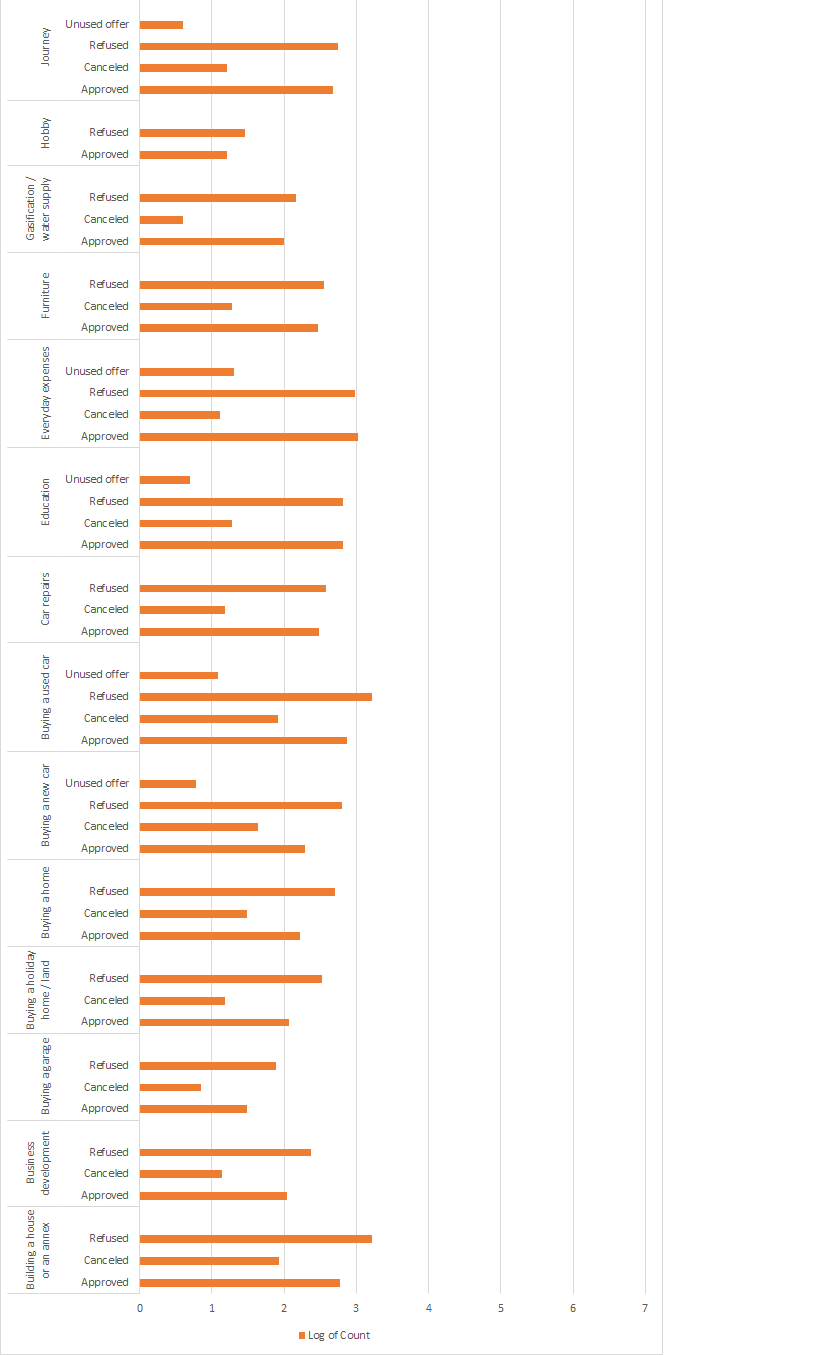


Now we have prepared a dataset of Distribution of contract status with purposes out of the pivot table of our new merged table.



Now out of this dataset plotting a graph of “ Distribution of contract status with purposes in logarithmic scale “





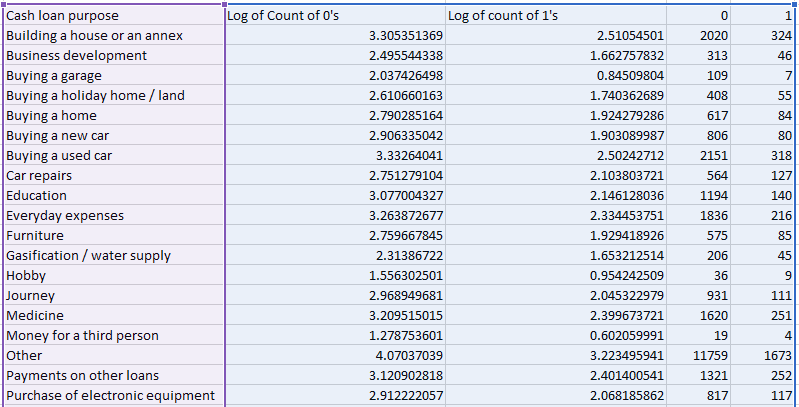
The above graph shows the distribution of contract status with purposes in logarithmic scale.

# Summary:

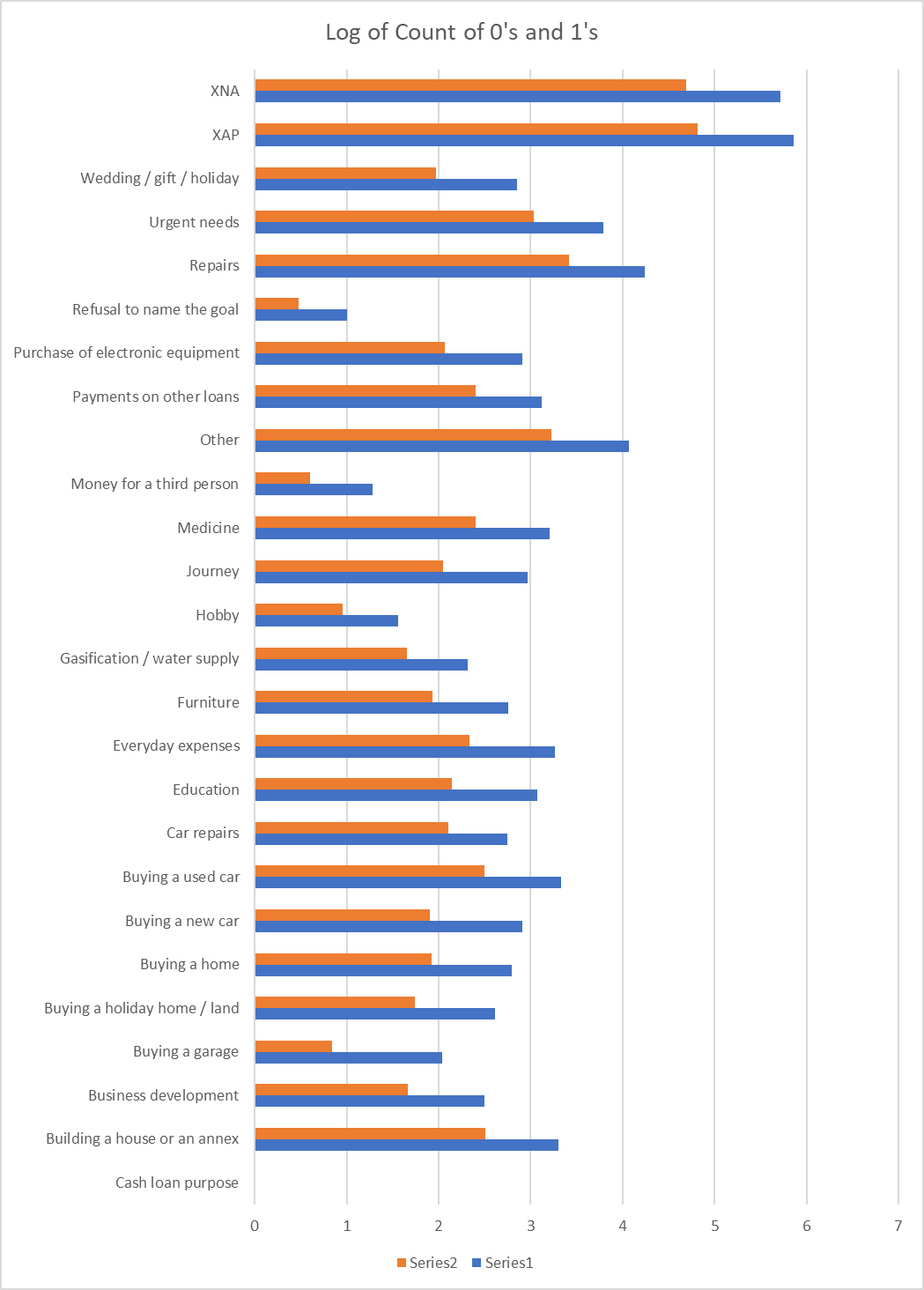
Most rejection of loans came from purpose 'repairs'. For education purposes we have an equal number of approvals and rejection. Paying other loans and buying a new car is having significantly higher rejection than approval.

# Distribution of cash loan purpose in logarithmic scale

We have prepared the dataset below from the pivot table of merged tables, showing logarithms of counts of 0’s and 1’s for different cash loan purposes.



A bar graph plotted out of this table is shown below-



Legend:

Series 2- Orange colour shows log of count of 1’s

Series 1- Blue colour shows log of count of 0’s

# Summary:

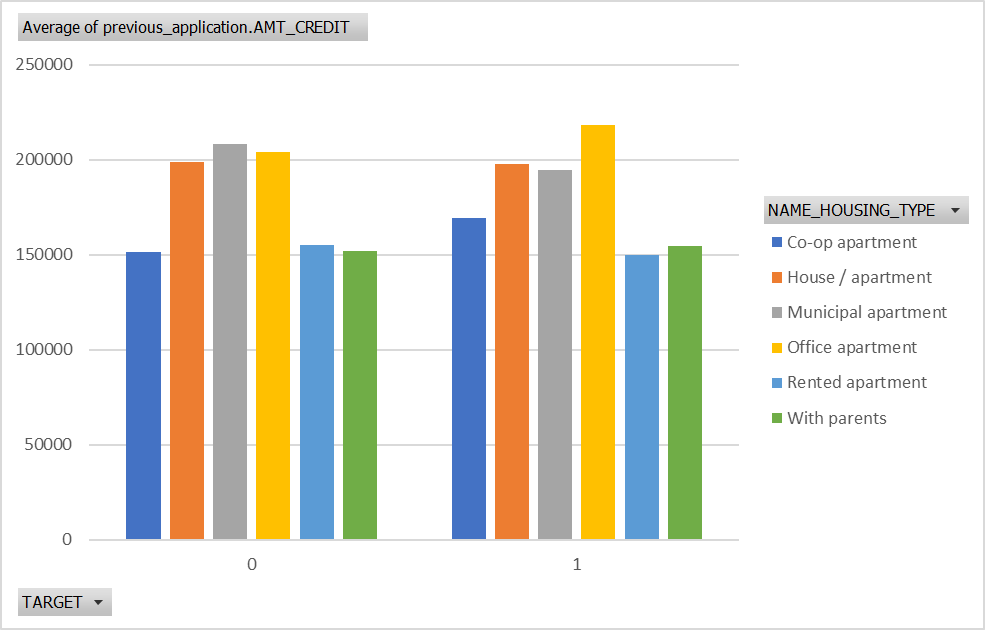
Loan purposes with 'Repairs' are facing more difficulties in payment on time. There are few places where loan payment is significantly higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land','Buying a new car' and 'Education' Hence we can focus on these purposes for which the client is having minimal payment difficulties.

# Bivariate analysis

## Box plotting for Credit amount prev vs Housing type in average

We have used a pivot table to analyse the average of credit amounts from merged previous application columns based on housing types for different targets.

The graph is shown below-



# Summary:

Here for Housing type, municipal apartment is having higher credit of target 0 and office apartment is having higher credit of target 1. So, we can conclude that banks should avoid giving loans to the housing type of office apartment as they are having difficulties in payment. Banks can focus mostly on housing type with parents or House\apartment or municipal apartment for successful payments.

# Conclusion

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

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

3. In loan purpose ‘Repairs’:

a. Although having a higher number of rejection in loan purposes with 'Repairs' there are observed difficulties in payment on time.

b. There are few places where loan payment delay is significantly high.

c. Banks should continue to caution while giving loans for this purpose.

4. Banks should avoid giving loans to the housing type of office apartment as they are having difficulties in payment.

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

END