Bank Loan Case Study

BY

Rahul Patil

Table of Contents

Project Description:

1. 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.

2. When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile.

   Two types of risks are associated with the bank's decision:

   a. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
   b. 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.

3. Loan case study help to understand risk analytics in banking and financial services & help to understand how data is used to minimize the risk of losing money while lending to customers.

Approach:

- To successfully carry out this project we are going to use **SIX STEP** of Data Analysis Process i.e (Ask, Prepare, Process, Analyze, Share, Act)
- Ask step include asking right set of question which justify goal and give motivation to carry out analysis

- We have following Objective (reasons) to justify goal of this project.

    - **To identify 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** is the aim of this case study.

- Prepare: We have data in excel format which need to first clean, transform and load into correct format to make it suitable for analysis purpose.
- This step includes selecting right data, tools, data source to make project successful
- Process: Data we have in excel format we need to clean data such as removing null values, identifying data type, removing outliers which affect the analysis.
- Analyze: We are using excel itself which come with inbuilt statistical formulae and visualization tool to analyze data to draw insight.
- Share: we are showing data obtain from analysis in the form of row and column as well as chart wherever required for better and easy understanding.
- Act: Step include taking decision based on insight opt from this project.

Tech-Stack Used:

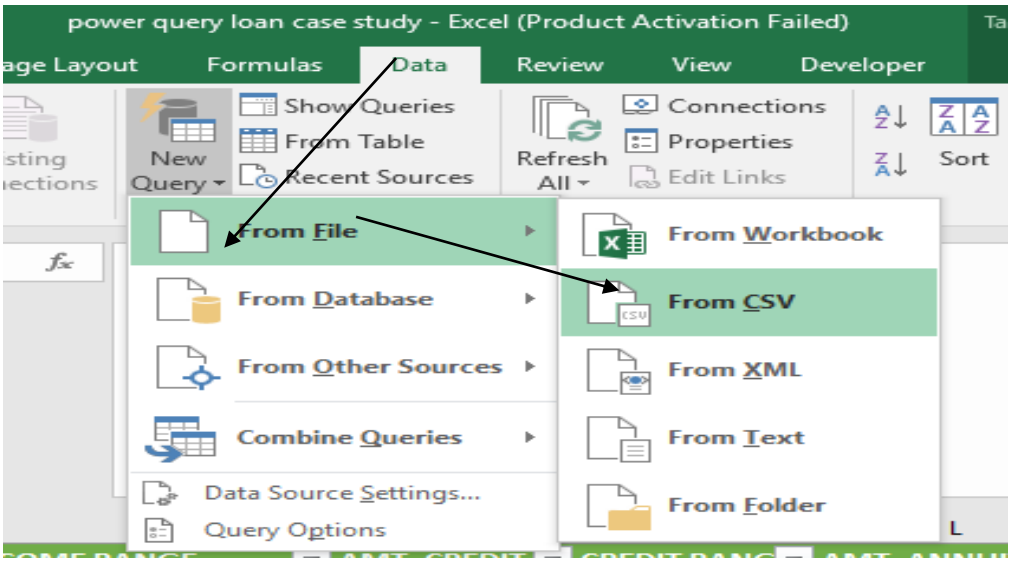
- Initially We are going to use excel which come with inbuilt statistical formulae and visualization tool to analyze data to draw insight.
- As this is bulky dataset above 3 lacs of record, with addition of formulae's, charts it become bulkier and start crashing frequently so as soon as it reached to this step further analysis we will move to Jupyter notebook.
- In this project we are going to perform medium level complex analysis using excel power query, pivot, formulae, chart are suitable tool to carry out analysis without investing in high technology and free open source tool Jupyter notebook.
- For using Jupyter notebook one must know the python programming language.
- Excel is not suitable for large dataset so it is advice to use good BI tool for large dataset for better analysis.

Insights:

Brief about data
Total number of columns=122
Total number of original record=307511

Data Cleaning:

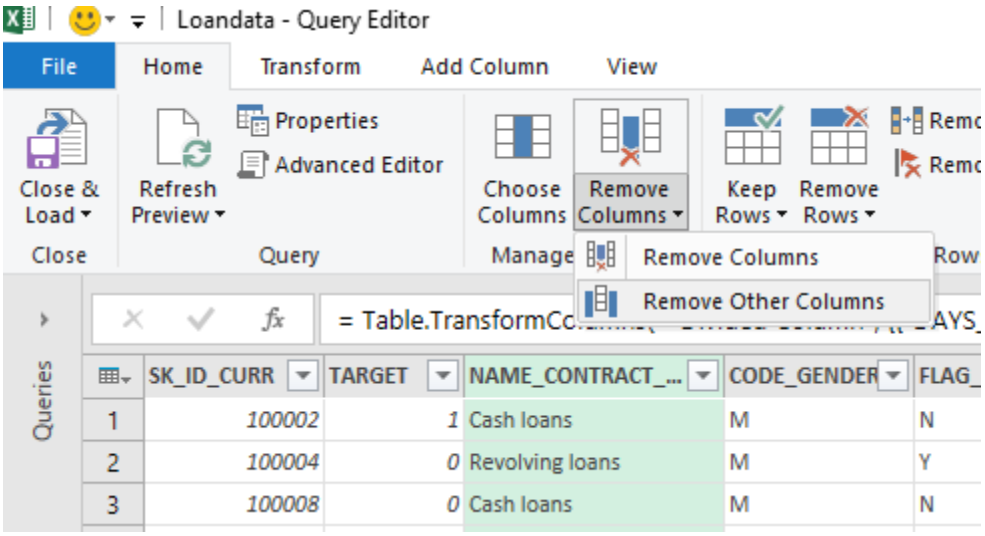
1. Open blank excel workbook refer below Image to load data using power query.



2. Select file format accordingly from the folder where data is stored.

3. Drop the unwanted column to reduce complexity of data.

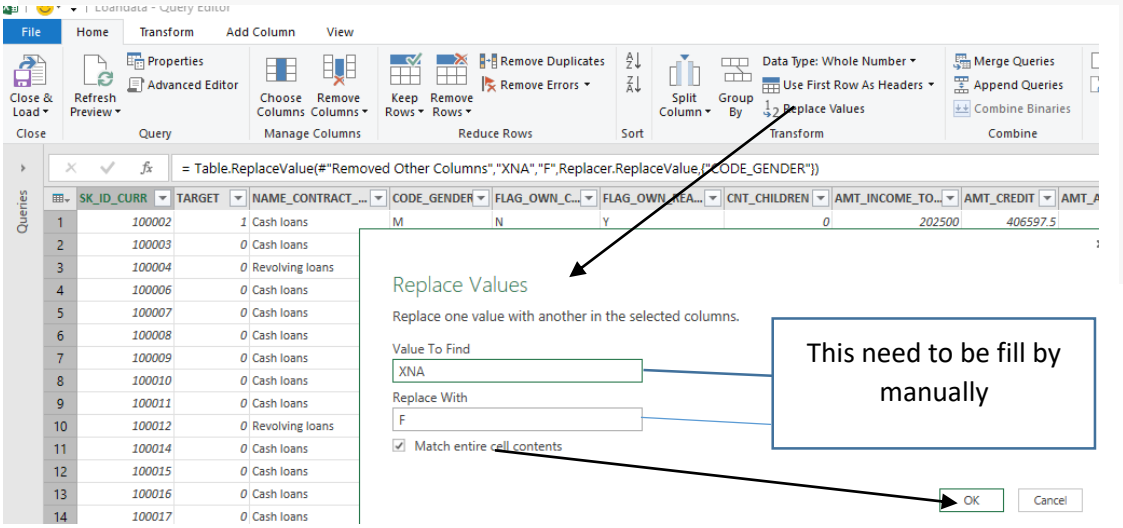
4. If number of column you want to remove is less then select column and click on remove column, if number of column you want to remove is more than column you want to keep then select column you want to keep and click on remove other columns option.



5. There are few Undefined values in column CODE_GENDER
6. Fill that values with mode of columns

   With F value

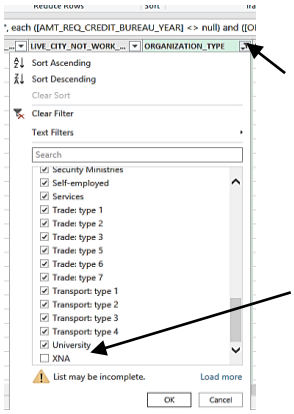
   Select column & Refer image below to perform operation



7. There are Undefined values in column ORGANIZATION_TYPE

8. We can't make conclusion on this values so we will select only those record h aving define ORGANIZATION_TYPE we will drop the record containing value XNA

9. Click on column arrow to expand option uncheck value want to remove.



10. To remove null value, perform same operation **explain in point 9.**

Brief about data after cleaning

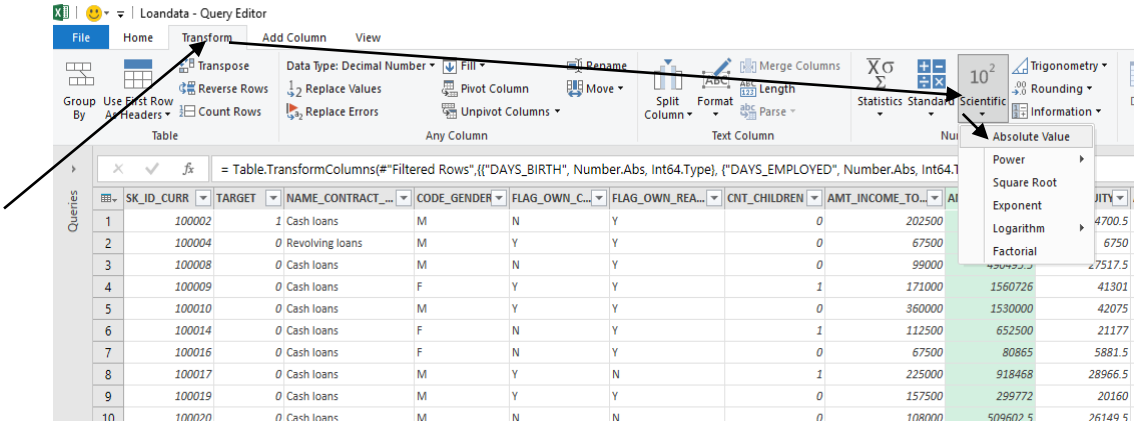
Total number of columns=42
Total number of record after cleaning data=202241
We have 66% data left after cleaning data

EDA:

1. 1.Following column data have negative values days cannot be negative s o convert this values to absolute values

    a. DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS _ID_PUBLISH

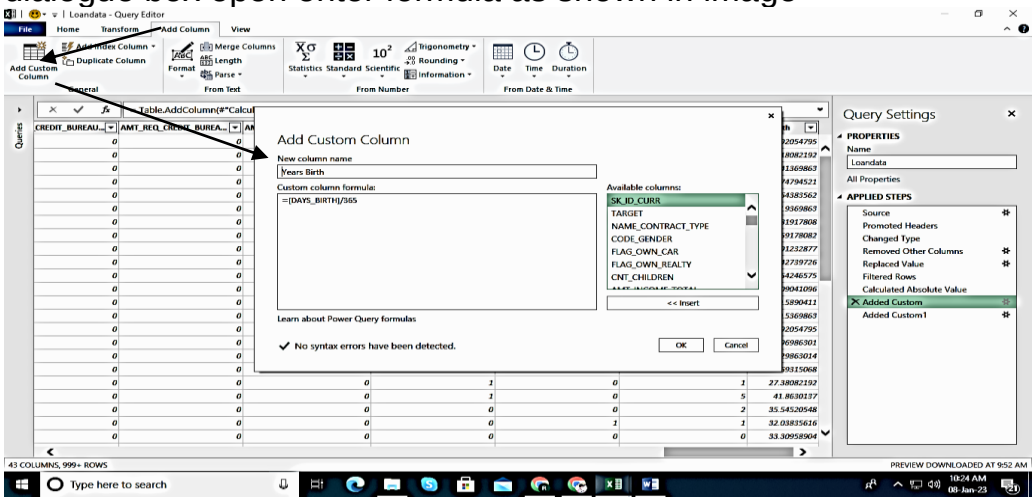
        i. Select columns go to Transform Tab click on Scientific select absolute value



2. Convert number of days of birth and employed into years by considering average days of a year as 365

DAYS_BIRTH / 365
DAYS_EMPLOYED / 365

- Click on Add Column tab then click on Add custom column new dialogue box open enter formula as shown in image



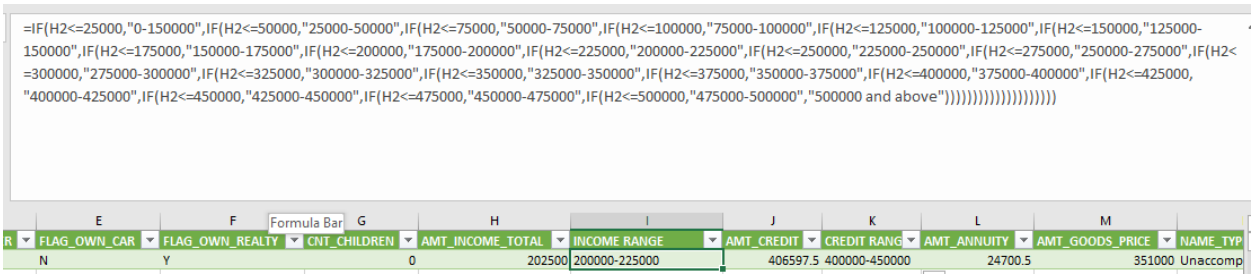
3. Create new column called **INCOME RANGE** and **CREDIT RANGE**

Minimum income =26550
Maximum income = 117000000

To Divide data in following range

0-25000,25000-50000,50000-75000,75000,100000,100000-125000, 125000-150000, 150000-175000,175000-200000,200000-225000,225000-250000,250000-275000,275000-300000,300000-325000,325000-350000,350000-375000,375000-400000,400000-425000,425000-450000,450000-475000,475000-500000,500000 and above

Use following formula to create new column INCOME RANGE

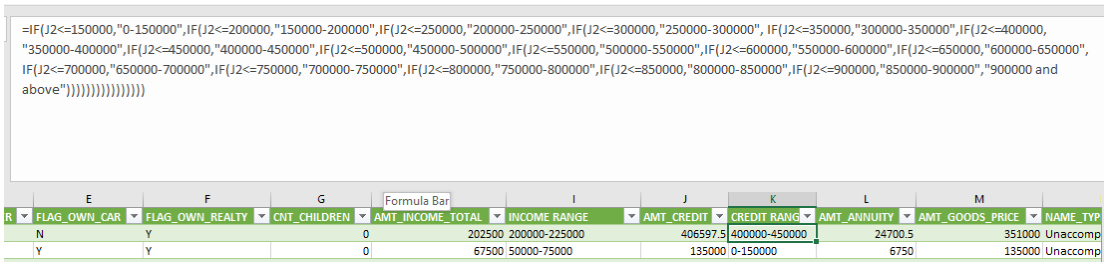


Minimum Credit amount = 45000
Maximum Credit amount = 4050000

To Divide data in following range

0-150000, 150000-200000,200000-250000, 250000-300000, 300000-350000, 350000-400000,400000-450000,450000-500000, 500000-550000,550000-600000,600000-650000,650000-700000, 700000-750000,750000-800000,800000-850000,850000-900000,900000 and above

Use following formula to create new column CREDIT RANGE

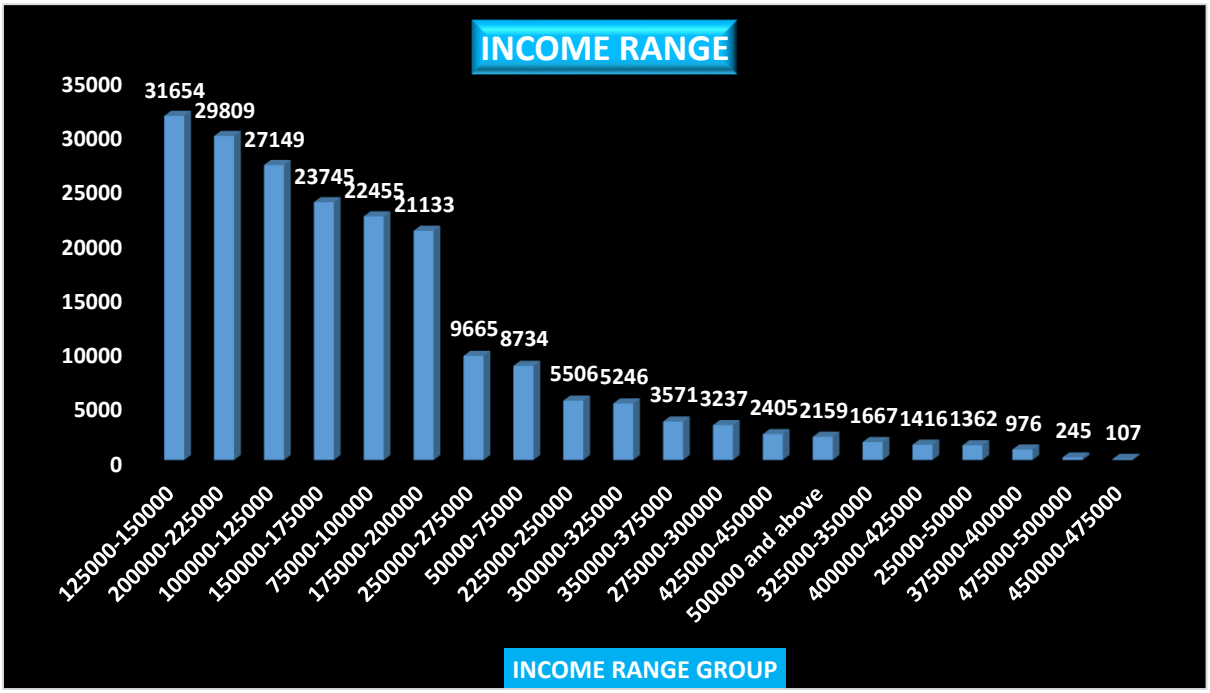


4. Find the ratio of data imbalance?

0	1	Grand Total	Defaulter per
185379	16862	202241	8.34%

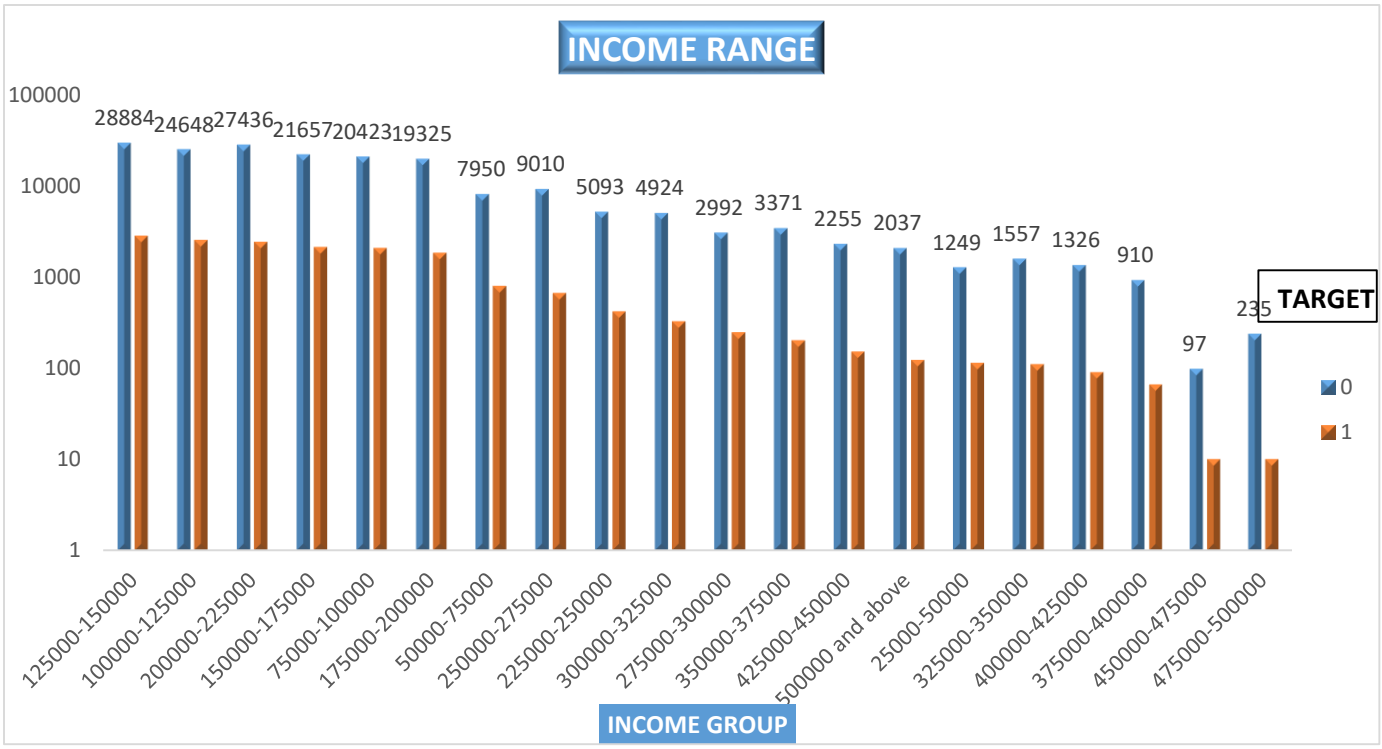
a. Almost 92 % people paying their EMI on time there are very less percentage of Defaulter.
b. Let's find out how we can minimize the defaulter percentage.

5. which income range of people are mostly applying for loans?



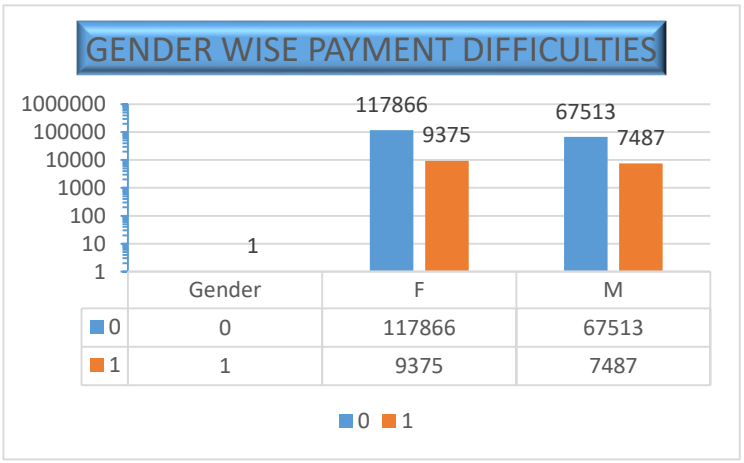
Most of people applying for loan are lies in following income group

INCOME GROUP	NUMBER OF APPLICATION
125000-150000	31654
200000-225000	29809
100000-125000	27149
150000-175000	23745
75000-100000	22455
175000-200000	21133



- From the above plot we can infer that the maximum number of clients without payment difficulties lie in the income range 1.2lac-1.5lac and immediate next income range is 2lac-2.25 lac
- Most clients with payment difficulties lie in the income range is 4.5lac-4.75lac

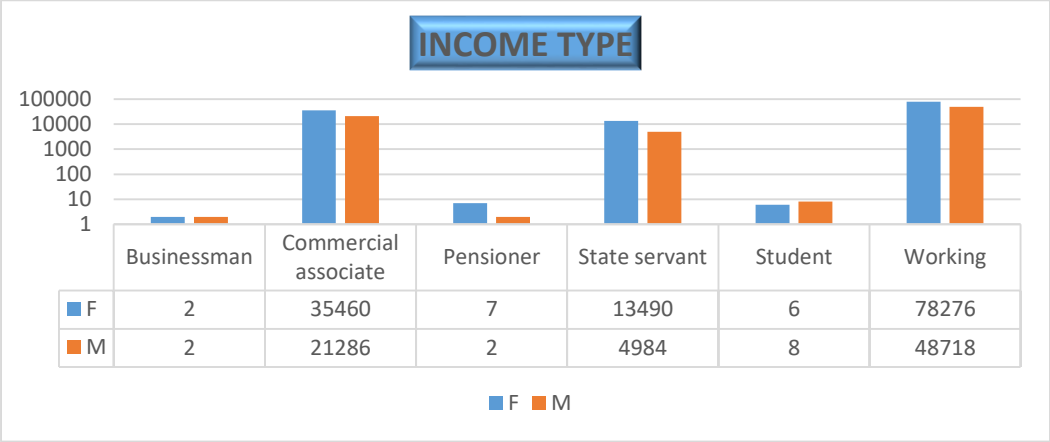
6.IF specific gender having payment difficulties?



- It can be seen that both number of males and females is almost same for having payment difficulties.

- It can also be seen that number of females is more than males for not having payment difficulties.

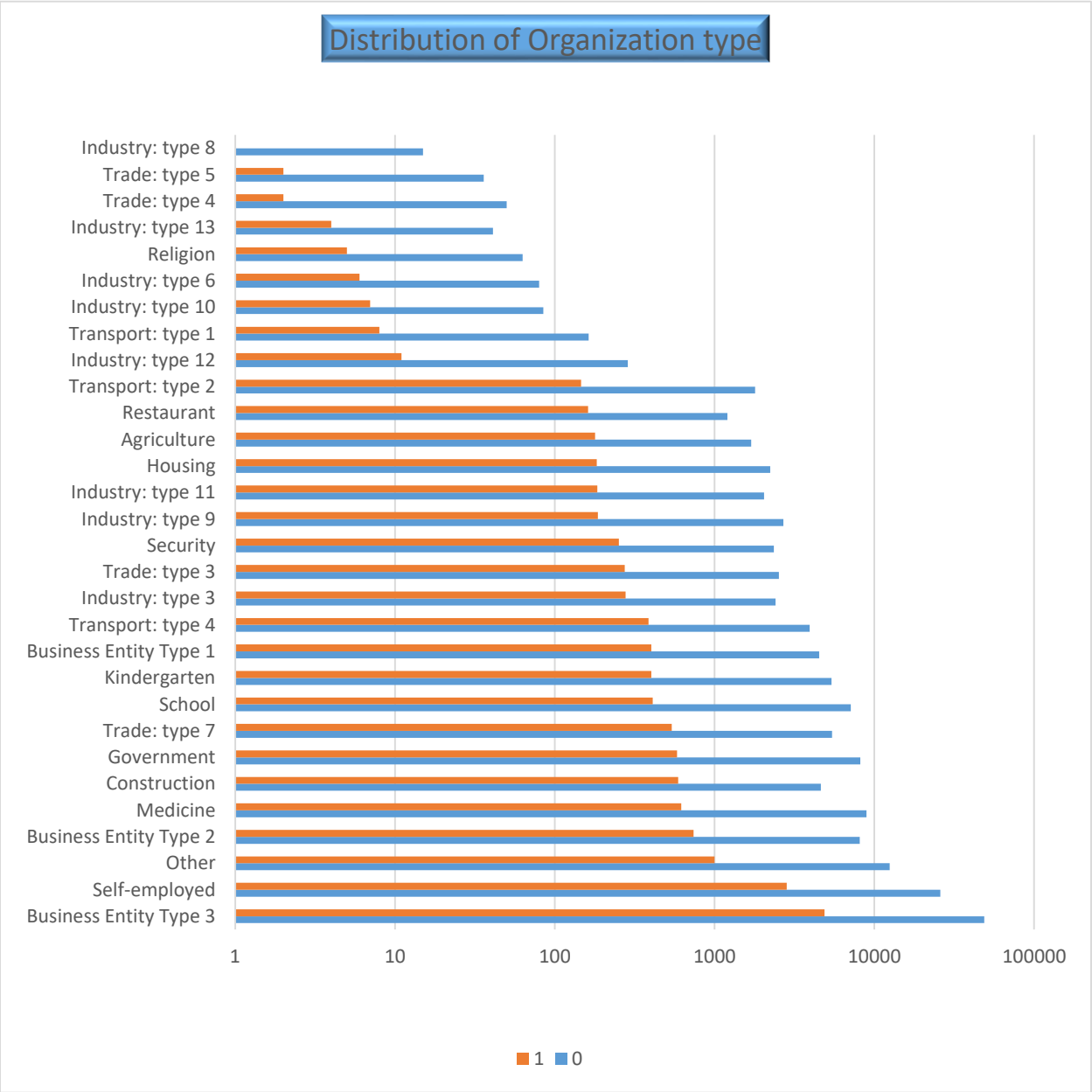
7.Which income type mostly applying for loan



	Businessman	Commercial associate	Pensioner	State servant	Student	Working
F	2	35460	7	13490	6	78276
M	2	21286	2	4984	8	48718

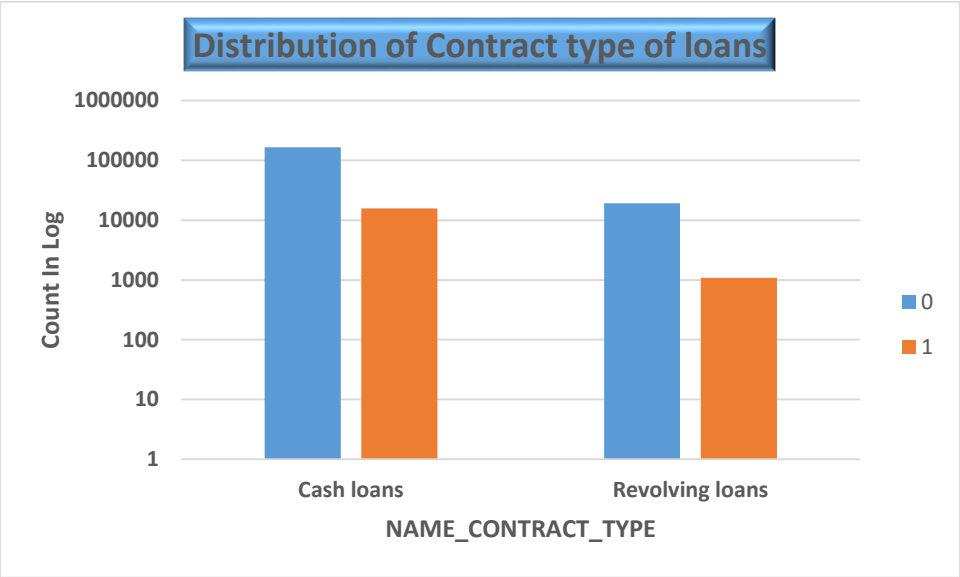
- We can conclude that most clients who fall in working type income category are applying for loans

8.what organization type people are applying for loans and are getting the loans actually.



- It can be seen that Trade: type 4, organization type has least count of payment difficulty clients
- Most clients with no payment difficulties lie in organization type named Business Entity Type 3

9.which contract type of loans more applications are pouring in for?

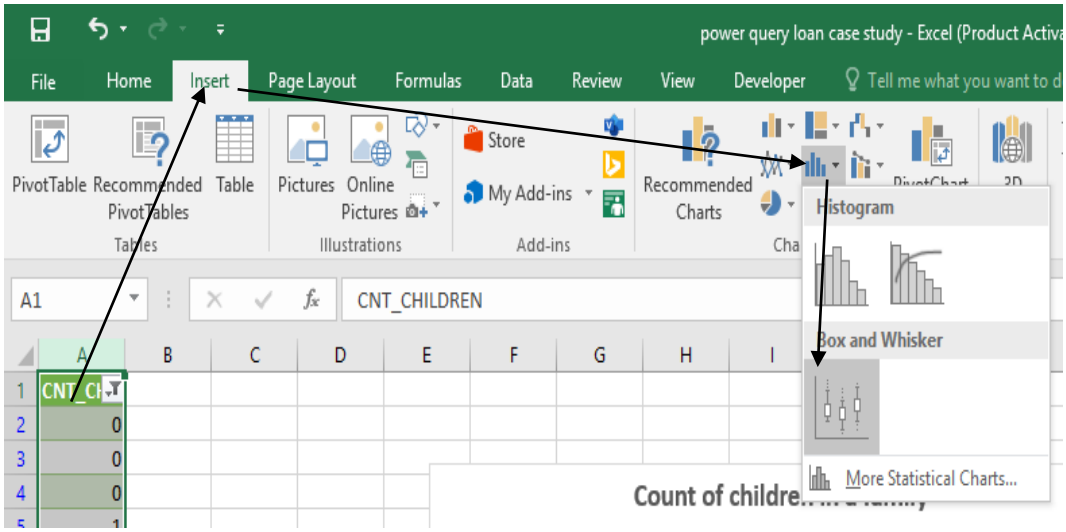


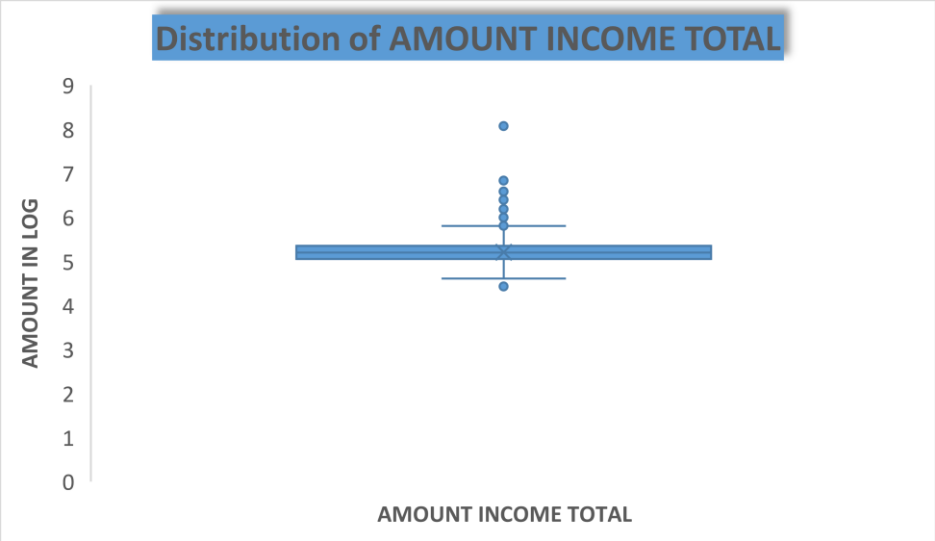
- Most cash loans applicants don't have payment difficulties
- The same type of loans also have the most applicants with payment difficulties

Identify if there are outliers in the data:

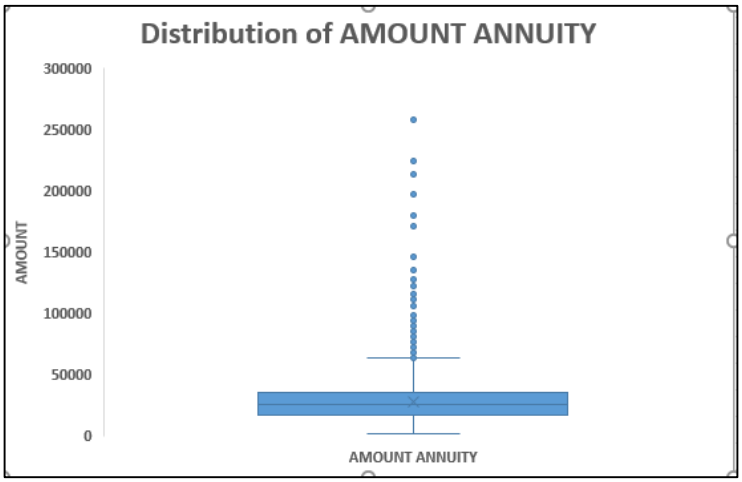
- Checking for outliers in AMT_INCOME_TOTAL

  Select range & follow step in following image to create box-plot

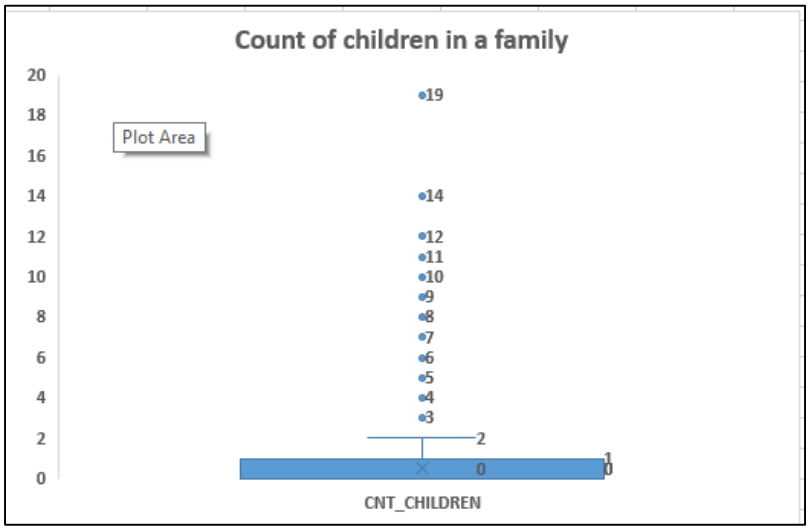




- Checking outliers for Credit Amount

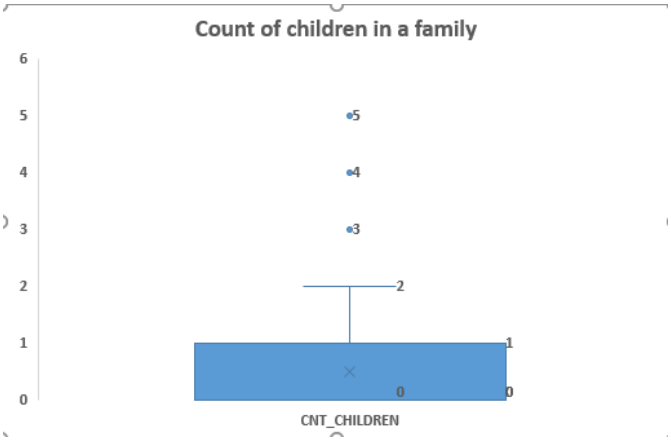


- checking the count of children in each family

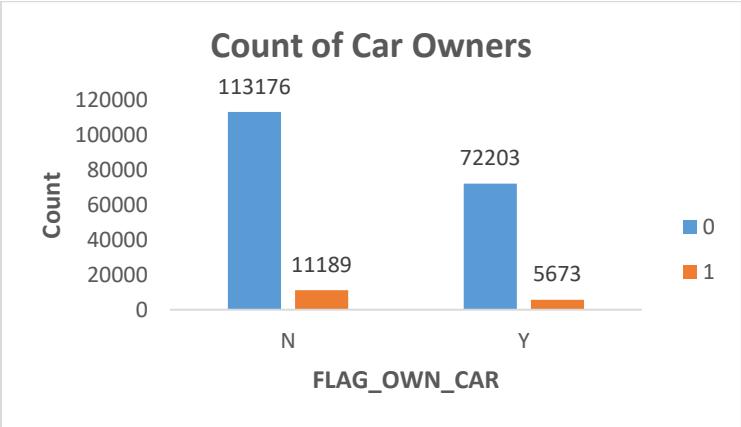


- It can be seen in the above boxplot that the count of children column clearly has outliers that are to be dealt with

- `Count of children in a family after dropping some rows`

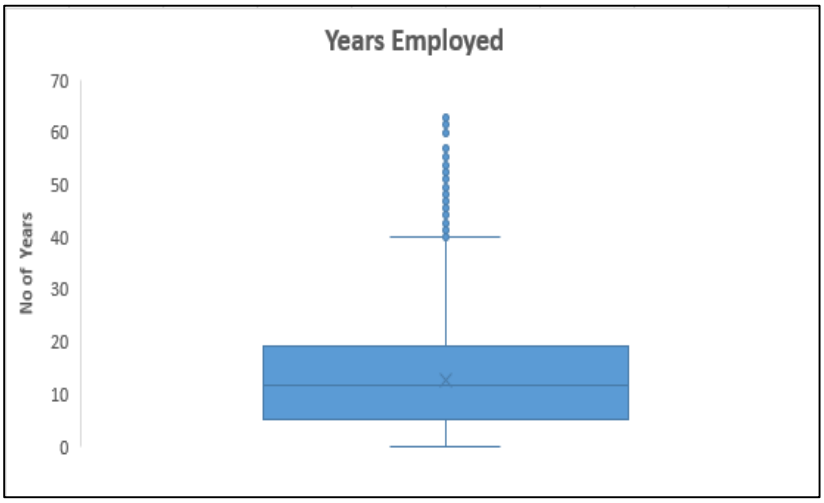


- Checking count of car owners with their capabilities to make a payment

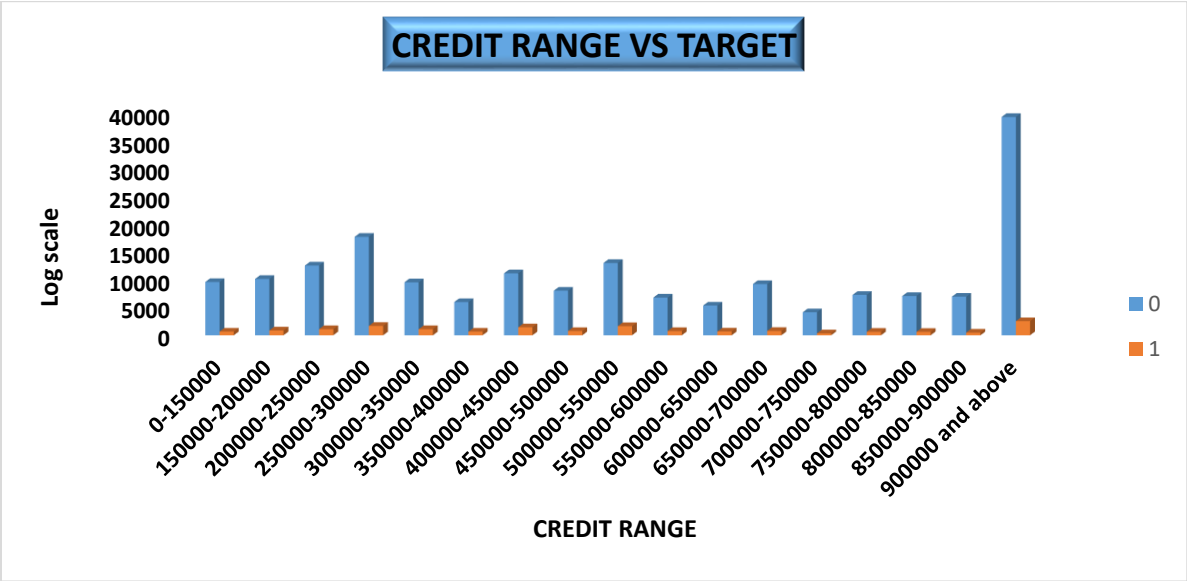


- Checking for outliers for the column DAYS_REGISTRATION

- Checking for outliers in column named Years_EMPLOYED



- Checking for Credit range that clients are getting that are likely to pay and not pay?

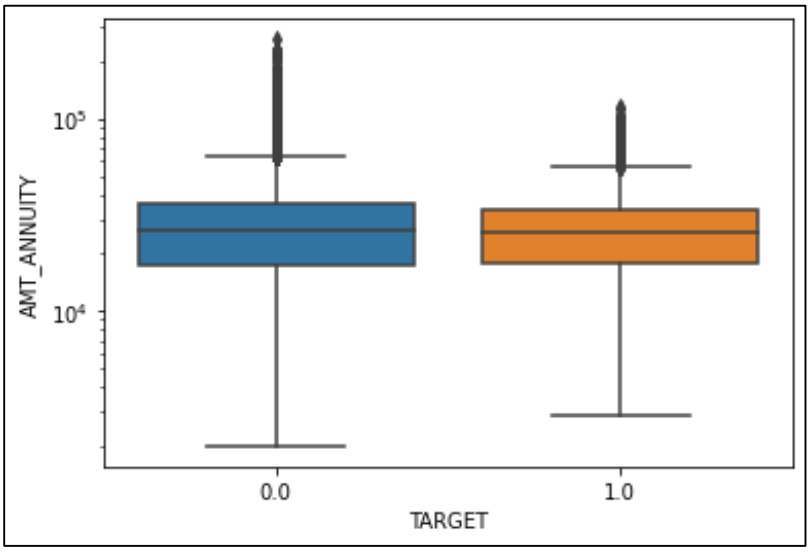


  - Clients with credit range lying in 900000 and above are the ones who are capable of paying the loans back
  - Least number of clients lying in income range 7lac- 7.5 lac are not capable of paying

Excel started crashing from this stage so we will move our rest of analysis on jupyter notebook.

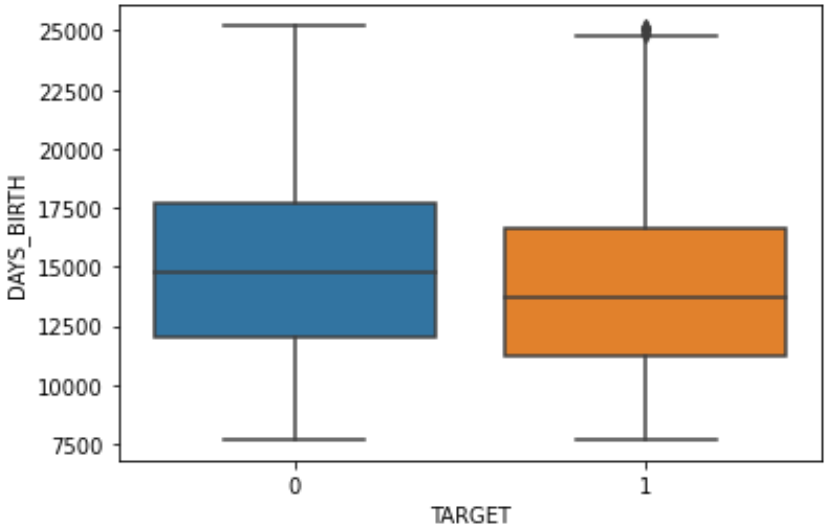
- Checking for annuity amount for Target variable

```
sns.boxplot(data=bank_loan_df, x='TARGET',y="AMT_ANNUITY")
plt.yscale('log')
plt.show()
```



- Checking for DAYS_BIRTH for Target variable

```
sns.boxplot(data=bank_loan_df, x='TARGET', y='DAYS_BIRTH')
plt.show()
```



- Dividing the dataset into two dataset of target=1(client with payment difficulties) and target=0(all other)

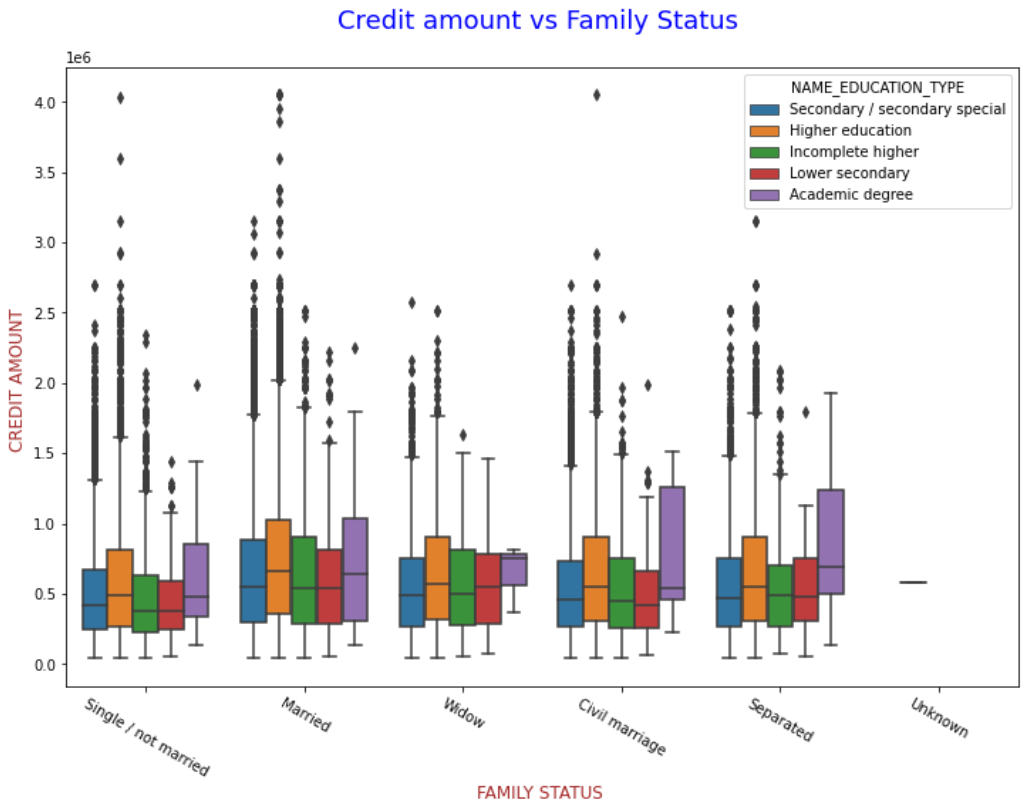
```
target0=bank_loan_df.loc[loanapp["TARGET"]==0]
target1=bank_loan_df.loc[loanapp["TARGET"]==1]
```

BIVARIATE ANALYSIS:

For Dataframe named target0:

- **Checking for Credit amount provided to the customers based on their Family type and Eduaction status**

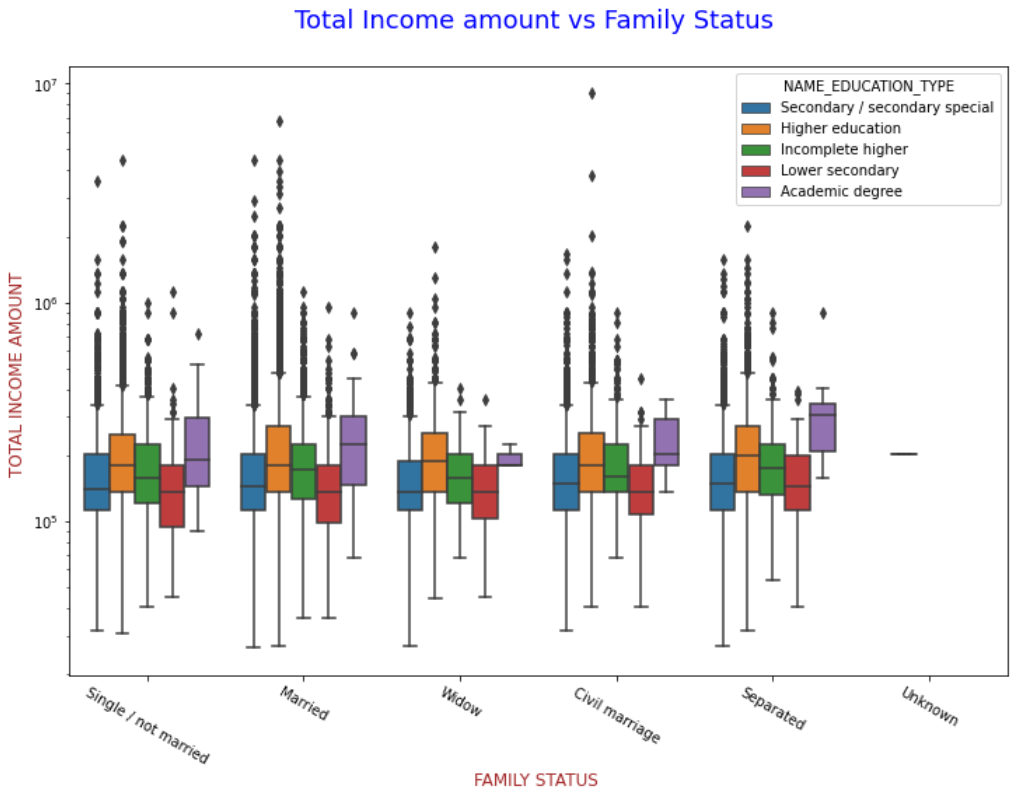
```
plt.figure(figsize=(12,8))
scale_factor=5
sns.boxplot(data=target0,x=target0.NAME_FAMILY_STATUS,y=target0.AMT_CREDIT,hue=target0.NAME_EDUCATION_TYPE)
plt.xticks(rotation=-30)
plt.xlabel("FAMILY STATUS ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.ylabel("CREDIT AMOUNT ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.title('Credit amount vs Family Status \n',fontdict={'fontsize': 18, 'fontweight' : 10, 'color' : 'Blue'})
plt.show()
```



- 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 or not' and 'civil marriage' are having more outliers. Civil marriage for Academic degree is having most of the credits in the third quartile.

- **Checking for Income amount of the customers based on their Family type and Eduaction status**

```
plt.figure(figsize=(12,8))
scale_factor=5
sns.boxplot(data=target0,x=target0.NAME_FAMILY_STATUS,y=target0.AMT_INCOME_TOTAL,hue=target0.N
AME_EDUCATION_TYPE)
plt.xticks(rotation=-30)
plt.xlabel("FAMILY STATUS ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.ylabel("TOTAL INCOME AMOUNT ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brow
n'})
plt.yscale('log')
plt.title('Total Income amount vs Family Status \n',fontdict={'fontsize': 18, 'fontweight' : 1
0, 'color' : 'Blue'})
plt.show()
```

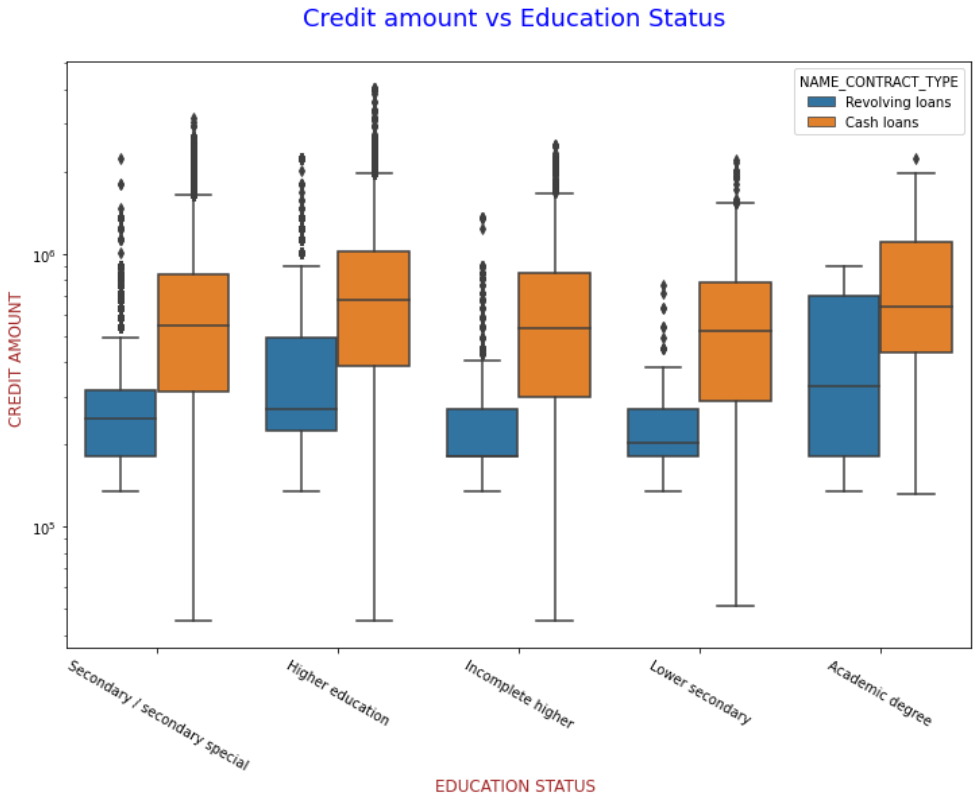


- Family status of 'civil marriage', 'marriage' and 'separated' of Higher education are having higher number of income than others.

- Also, higher education and secondary/second special education statuses with family status of 'marriage', 'single or not' and 'civil marriage' are having more outliers. Married for Higher education is having most of the incomes in the lower bound.

- **Checking for Credit amount provided to the customers based on their education and Contract types of loans they are applying for**

```
plt.figure(figsize=(12,8))
scale_factor=5
sns.boxplot(data=target0,x=target0.NAME_EDUCATION_TYPE,y=target0.AMT_CREDIT,hue=target0.NAME_C
ONTRACT_TYPE)
plt.xticks(rotation=-30)
plt.xlabel("EDUCATION STATUS ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brow
n'})
plt.ylabel("CREDIT AMOUNT ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.yscale('log')
plt.title('Credit amount vs Education Status \n',fontdict={'fontsize': 18, 'fontweight' : 10,
'color' : 'Blue'})
plt.show()
```

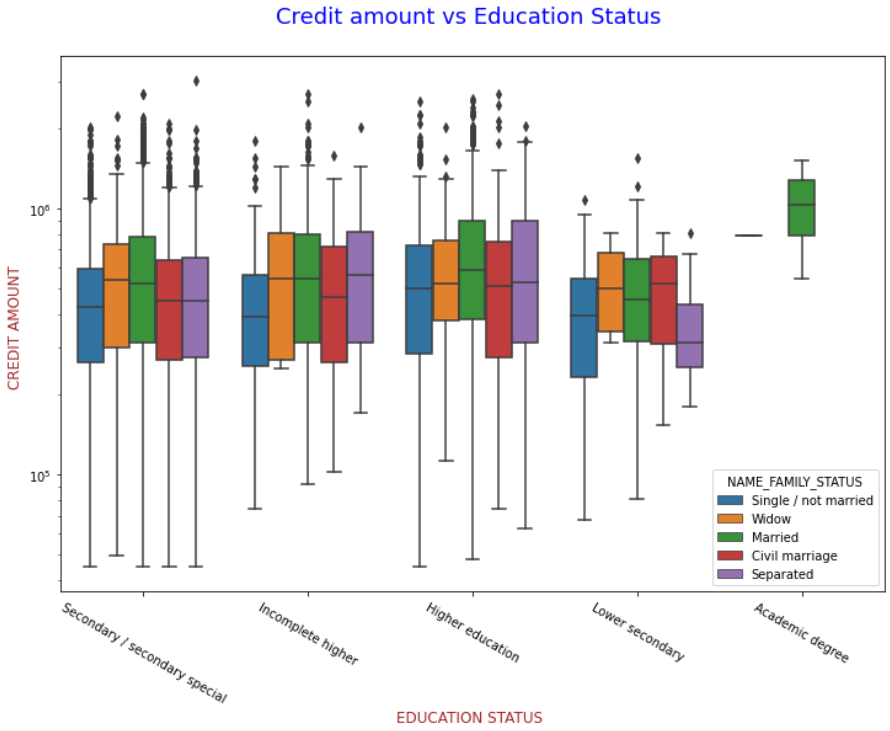


- It can be seen that the clients with education status of Higher education are the maximum credit seekers with highest in terms of contract type of cash loans of contract type.
- It can also be observed that the contract type revolving loans issued maximum credit amount holders education status Higher education.
- The highest credit amount in cash loans is given to a client with education level secondary/secondary special.
- basically education level or type is not playing much role as of who gets what amount of credit.

For Dataframe named target1:

- **Checking for Credit amount provided to the customers based on their education and according their family status**

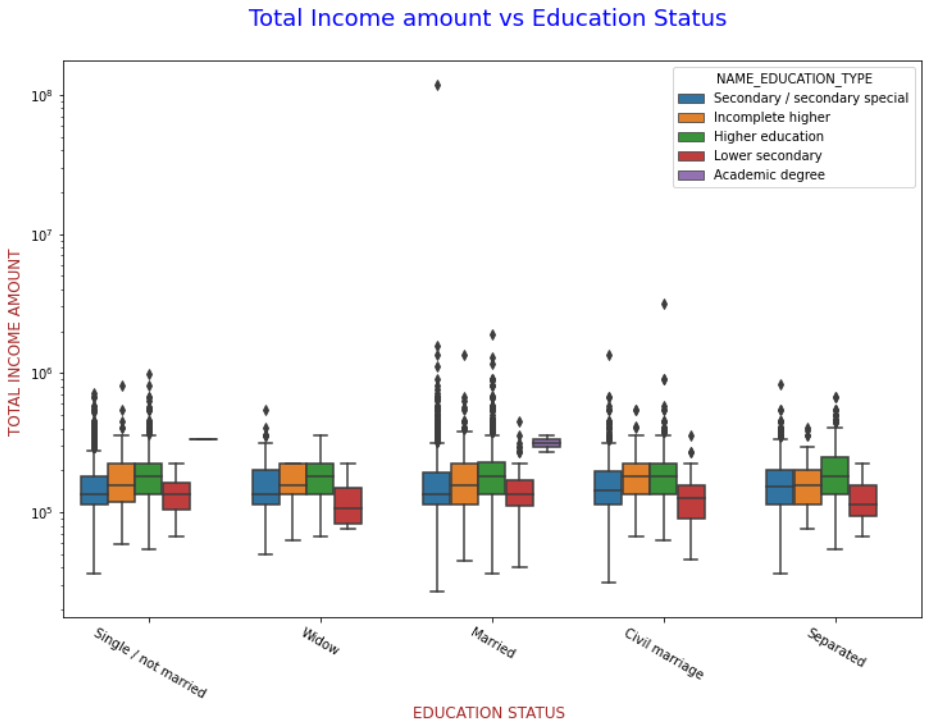
```
plt.figure(figsize=(12,8))
scale_factor=5
sns.boxplot(data=target1,x=target1.NAME_EDUCATION_TYPE,y=target1.AMT_CREDIT,hue=target1.NAME_FAMILY_STATUS)
plt.xticks(rotation=-30)
plt.xlabel("EDUCATION STATUS ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.ylabel("CREDIT AMOUNT ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.yscale('log')
plt.title('Credit amount vs Education Status \n',fontdict={'fontsize': 18, 'fontweight' : 10, 'color' : 'Blue'})
plt.show()
```



1. Observations are Quite similar with Target 0
2. Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are --having less number of credits than others.
3. Most of the outliers are from Education type 'Higher education' and 'Secondary'.
4. Most number of all types of education as well as family lie in lower bound

- **Checking for Income amount of the customers based on their education and according to their family status**

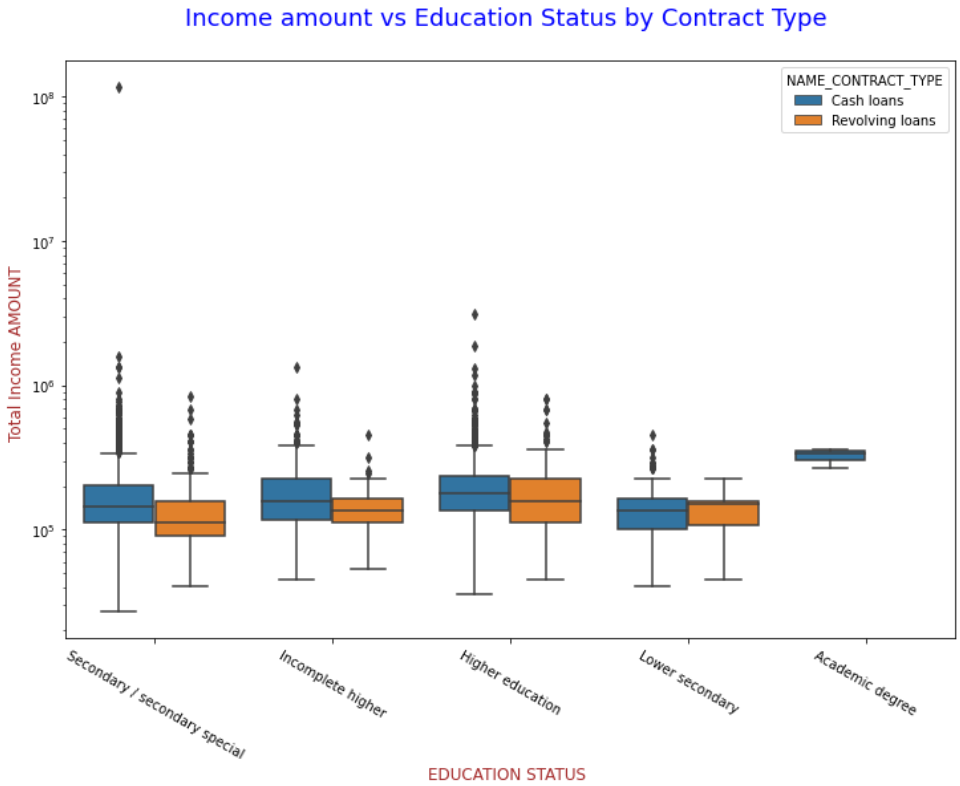
```
plt.figure(figsize=(12,8))
scale_factor=5
sns.boxplot(data=target1,x=target1.NAME_FAMILY_STATUS,y=target1.AMT_INCOME_TOTAL,hue=target1.NAME_EDUCATION_TYPE)
plt.xticks(rotation=-30)
plt.xlabel("EDUCATION STATUS ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.ylabel("TOTAL INCOME AMOUNT ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.yscale('log')
plt.title('Total Income amount vs Education Status \n',fontdict={'fontsize': 18, 'fontweight' : 10, 'color' : 'Blue'})
plt.show()
```



1. This is also having same similarity with target0,
2. Almost all Family types and education types have the income amount mostly equal.
3. Least outlies are for Lower secondary and their income amount is also little lesser than that of all other education types.
4. In Academic degree very less number of people are in payments with difficulty from dataframe named target1

- **Checking for Credit amount provided to the customers based on their education and according the Contract types of loans they are applying for.**

```
plt.figure(figsize=(12,8))
scale_factor=5
sns.boxplot(data=target1,x=target1.NAME_EDUCATION_TYPE,y=target1.AMT_INCOME_TOTAL,hue=target1.NAME_CONTRACT_TYPE)
plt.xticks(rotation=-30)
plt.xlabel("EDUCATION STATUS ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.ylabel("Total Income AMOUNT ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.yscale('log')
plt.title('Income amount vs Education Status by Contract Type \n',fontdict={'fontsize': 18, 'fontweight' : 10, 'color' : 'Blue'})
plt.show()
```



- It can be seen that the maximum income amount holders are with education levels secondary and higher education levels
- It's strange that, there are no revolving amount loans are demanded by clients with education level academic degree

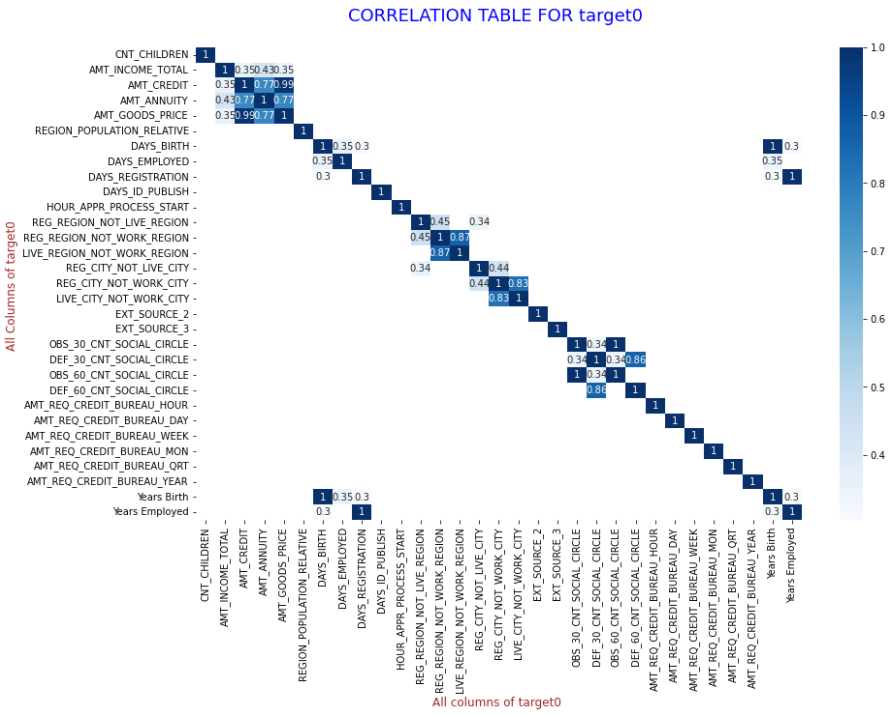
Multivariate Analysis:

Creating a correlation table between all the variables of target0 dataframe

```
target0_correlation=target0.iloc[:,2:].corr()
target0_correlation
```

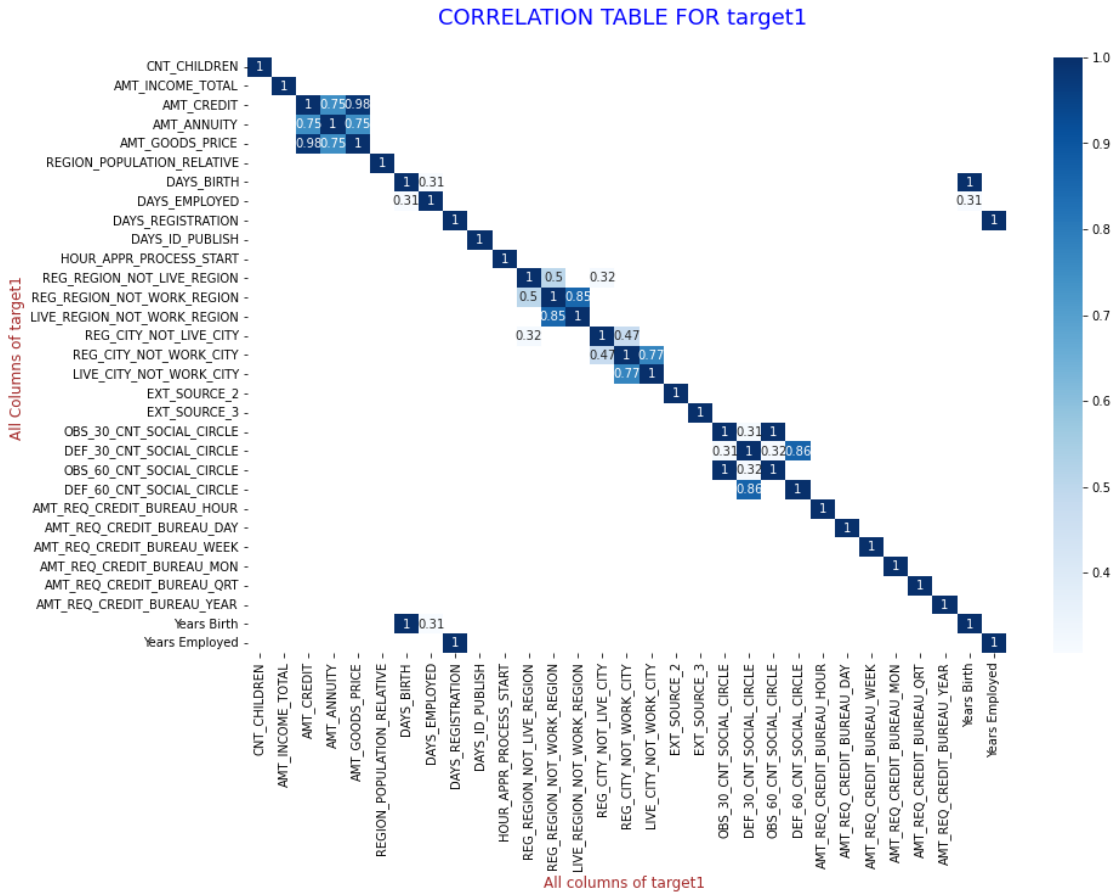
Table is large to display we are going to use heatmap to visualize it

```
#plotting the correlation table of all variables of target0 dataframe
plt.figure(figsize=(14,9))
sns.heatmap(target0_correlation[target0_correlation>0.3],cmap='Blues',annot=True)
plt.title('CORRELATION TABLE FOR target0 \n',fontdict={'fontsize':18, 'fontweight' : 10, 'color' : 'Blue'})
plt.xlabel("All columns of target0 ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.ylabel("All Columns of target0 ", fontdict={'fontsize': 12, 'fontweight' : 5, 'color' : 'Brown'})
plt.show()
```



MOST RELEVANT CORRELATIONAL PAIRS with no difficulties of payment

- It is observed that the maximum correlation is between the variables listed below in pairs
    1. AMT_CREDIT to AMT_ANNUITY
    2. AMT_CREDIT to AMT_TOTAL_INCOME
    3. AMT_ANNUITY to AMT_INCOME_TOTAL
    4. REG_CITY_NOT_WORK_CITY to REG_CITY_NOT_LIVE_CITY
    5. DAYS_EMPLOYED to DAYS_BIRTH
    6. DAYS_BIRTH to DAYS_REGISTRATION



Almost all of the correlational pairs of variables are as same as target0 dataframe

Summary:

1. Banks should focus more on contract type ‘Student’, ‘pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ and 'office apartment' for successful payment
2. Banks should focus less on income types maternity leave and working as they have most number of unsuccessful payments
3. Bank can focus mostly on housing type with parents, House or apartment and municipal apartment with purpose of education, buying land, buying a garage, purchase of electronic equipment and some other purposes with target0 significantly more than target1 for successful payments.
4. People living with their parent or having own house don’t need to pay rent its less likely to default.
5. Banks can offer more offers to clients who are students and pensioners as they take all offers and are more likely to pay back.