Credit EDA Assignment Case Study

Sai Vaibhav Naidu
Email: svnhpt.vaibhav@gmail.com
Phone: +917829999128

Problem Statement

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 to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

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:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. 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.

Business Objective

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.

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.

Steps for Analysis

1. Data Sourcing
    a. We need to read the 2 csv files given to us.
    b. application_data.csv will be stored in dataset inp0
    c. previous_application.csv will be stored in dataset inp1
    d. Joining the 2 datasets into a single dataset after the analysis to compare the data
2. Data Cleaning
    a. Drop the columns or rows with more than 30% NULL values present it.
    b. Impute or replace the missing data in the dataset based on the amount of missing values.
    c. Removing unwanted columns which is unnecessary for the analysis
3. Performing Univariate analysis
    a. Dividing the dataset into 2. Target_0 and Target_1 which represent no payment issues and payment issues respectively
    b. Performing the analysis on AMT_ANNUITY, AMT_CREDIT, AMT_INCOME_TOTAL by box plot, bar graphs, and creating the buckets
    c. Calculating the correlation between the two sets to plotting the heatmaps
    d. Checking for the outliners in the data and making the final judgement

Notes

1. I did the coding in Google collabs and have commented on the lines as and when it is required throughout the code cells
2. The coding file is attached along with this presentation in .ipynb file
3. I have presented the important graphs here in he slides and will be taking through the final conclusion based on it.
4. To get the proper understanding, please go through the coding file as the majority of the steps will not be included here in the slides

Data Cleaning in brief

1. We can observe that there are 307511 rows and 122 columns
2. We found that there are 64 columns with more than 30% NULL values
3. Those columns are dropped

```
[4] #Checking the number of rows and columns
    inp0.shape

    (307511, 122)
```

```
[5] #Creating a variable to store the columns names which has more than 30% NULL values
    emptyCol = inp0.isnull().sum()
    emptyCol=emptyCol[emptyCol.values>(0.3*len(emptyCol))]
    len(emptyCol)

    64
```

```
[6] #Dropping those columns whose data has more than 30% NULL values
    emptyCol = list(emptyCol[emptyCol.values>=0.3].index)
    inp0.drop(labels=emptyCol,axis=1,inplace=True)
    print(len(emptyCol))

    64
```

Similarly columns and rows with null values or XNA values are found and either removed or replaced with the most common value or the median value(in case of numeric column)

Categorical Univariate Analysis

Analysis for Target - 0: Clients with no payment difficulty

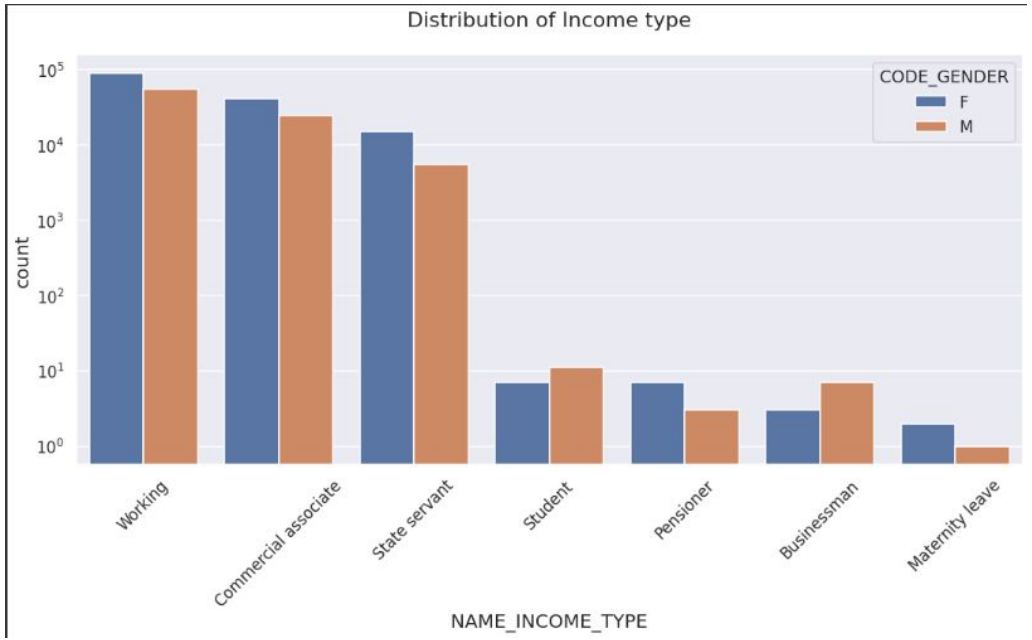
Bar plot for different income ranges

1. Female counts are higher than male.
2. Income range from 100000 to 200000 is having more number of credits.
3. Very less count for income range 400000 and above

Analysis for Target - 0: Clients with no payment difficulty

Bar plot for different income types

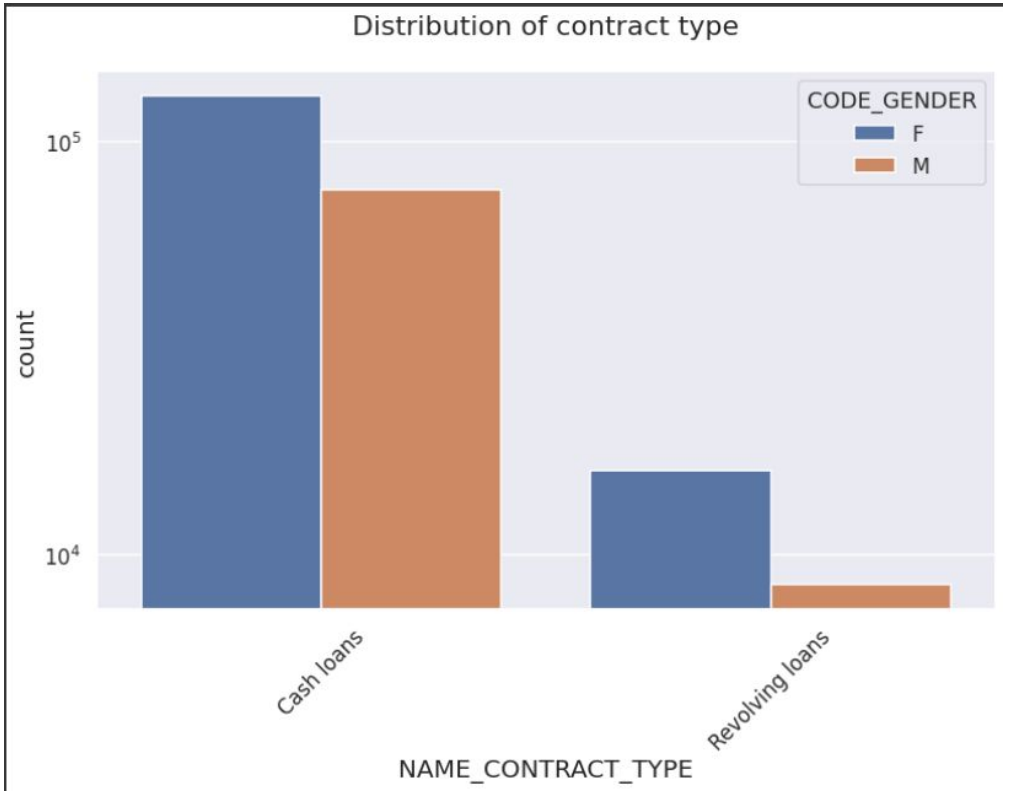
1. For income type ‘working’, 'commercial associate’, and ‘State Servant’ the number of credits are higher than others.
2. For this Females are having more number of credits than male.
3. Less number of credits for income type ‘student’ ,’pensioner’, ‘Businessman’ and ‘Maternity leave’.



Analysis for Target - 0: Clients with no payment difficulty

Bar plot for different contract types

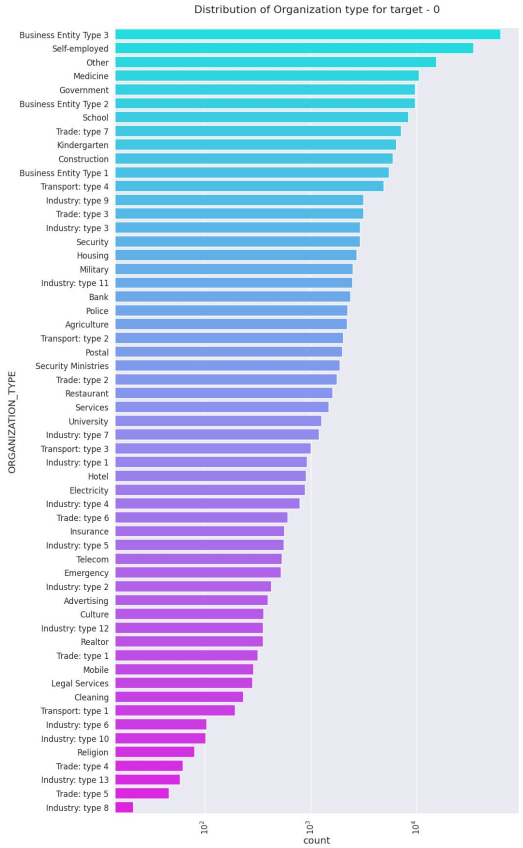
1. For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.
2. In this category too, Females are leading for applying credits.



Analysis for Target - 0: Clients with no payment difficulty

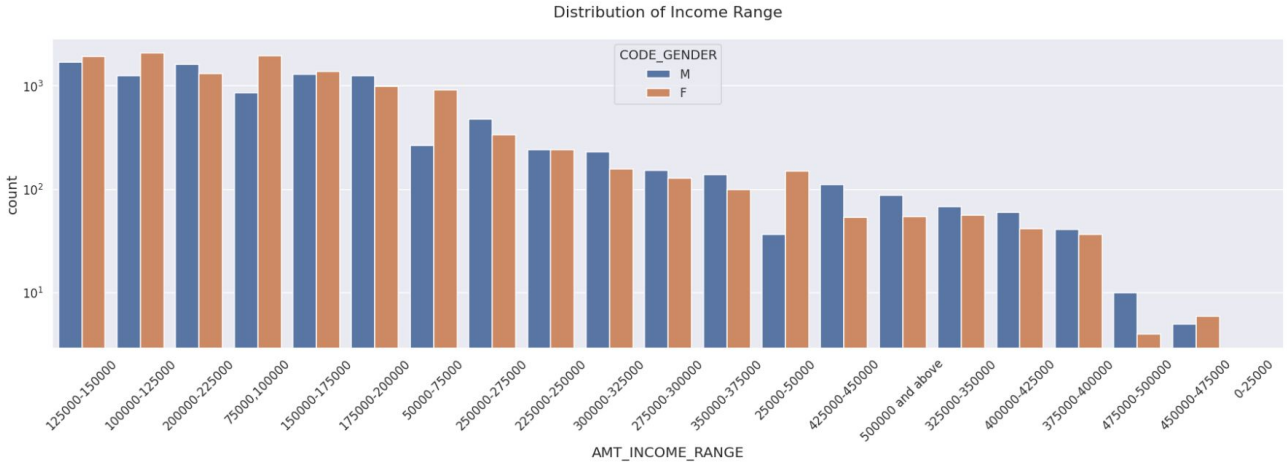
Horizontal Bar plot for different Organization types

1. Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’, ‘Other’ , ‘Medicine’ and ‘Government’.
2. Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4



Analysis for Target - 1: Clients with payment difficulty

Bar plot for different income ranges

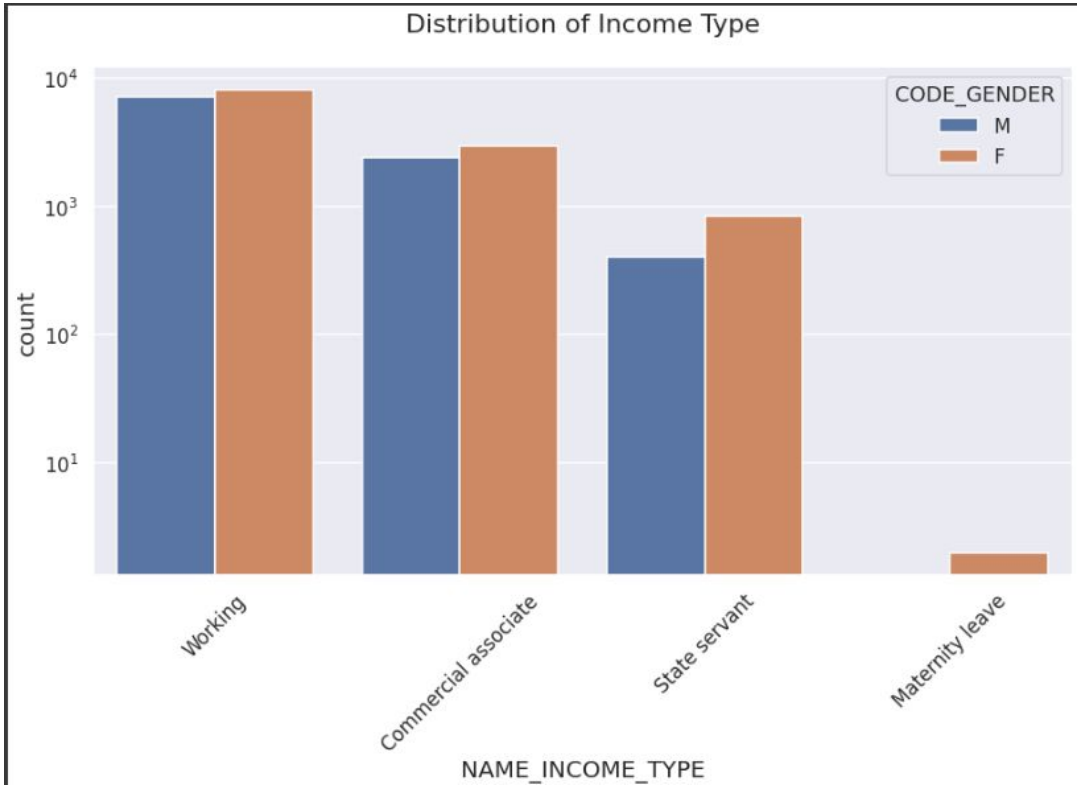


1. Male counts are higher than female.
2. Income range from 100000 to 200000 is having more number of credits.
3. This graph show that males are more than female in having credits for that range.
4. Very less count for income range 400000 and above.

Analysis for Target - 1: Clients with payment difficulty

Bar plot for different income types

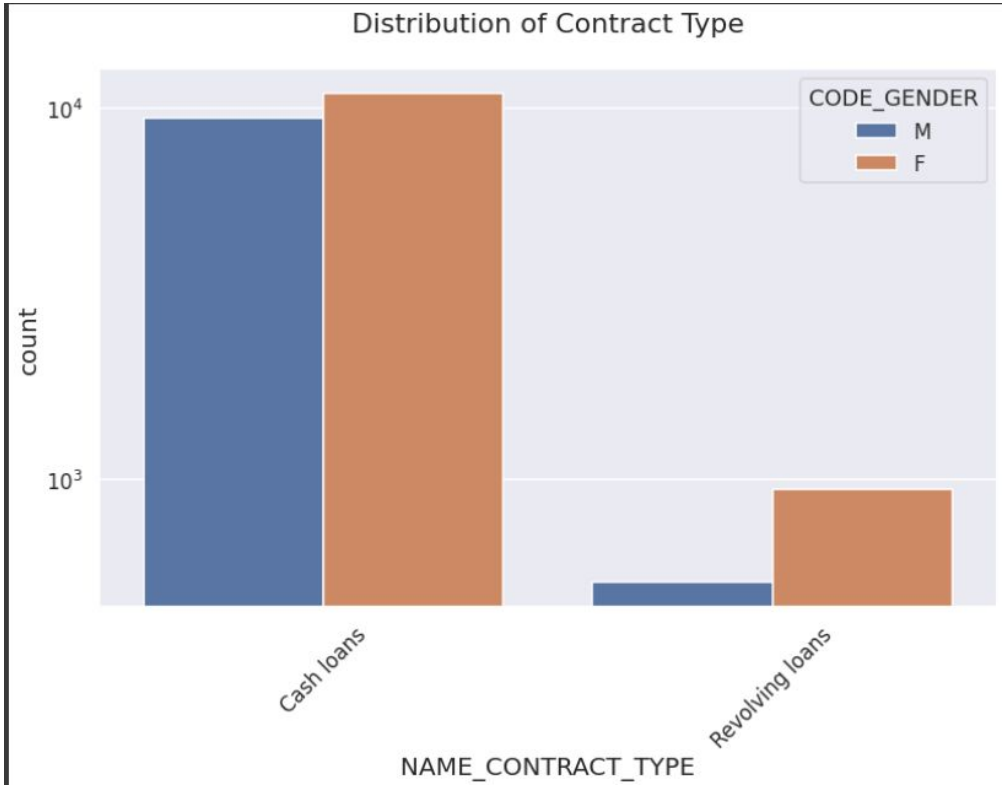
1. For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than other i.e. ‘Maternity leave.
2. For this Females are having more number of credits than male.
3. Less number of credits for income type ‘Maternity leave’.
4. For type 1: There is no income type for ‘student’ , ’pensioner’ and ‘Businessman’ which means they don’t do any late payments.



Analysis for Target - 1: Clients with payment difficulty

Bar plot for different contract types

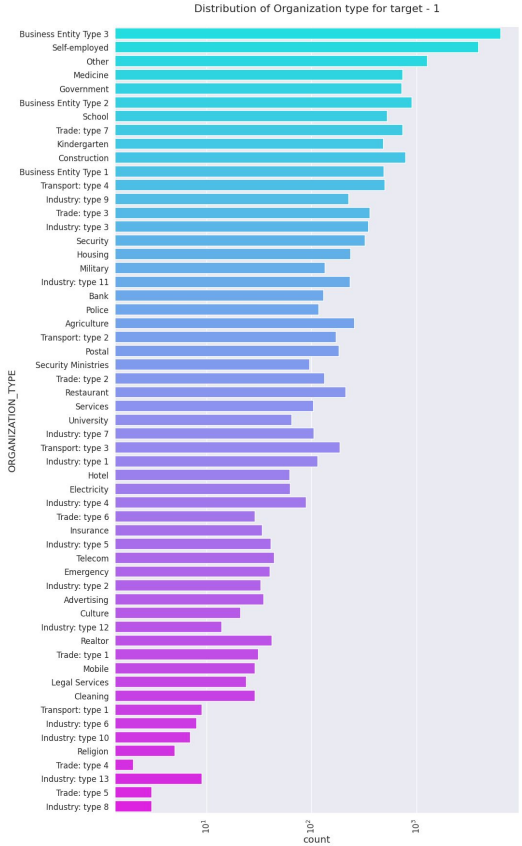
1. For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.
2. For this also Female is leading for applying credits.
3. For type 1 : there is only Female Revolving loans.



Analysis for Target - 1: Clients with payment difficulty

Horizontal Bar plot for different Organization types

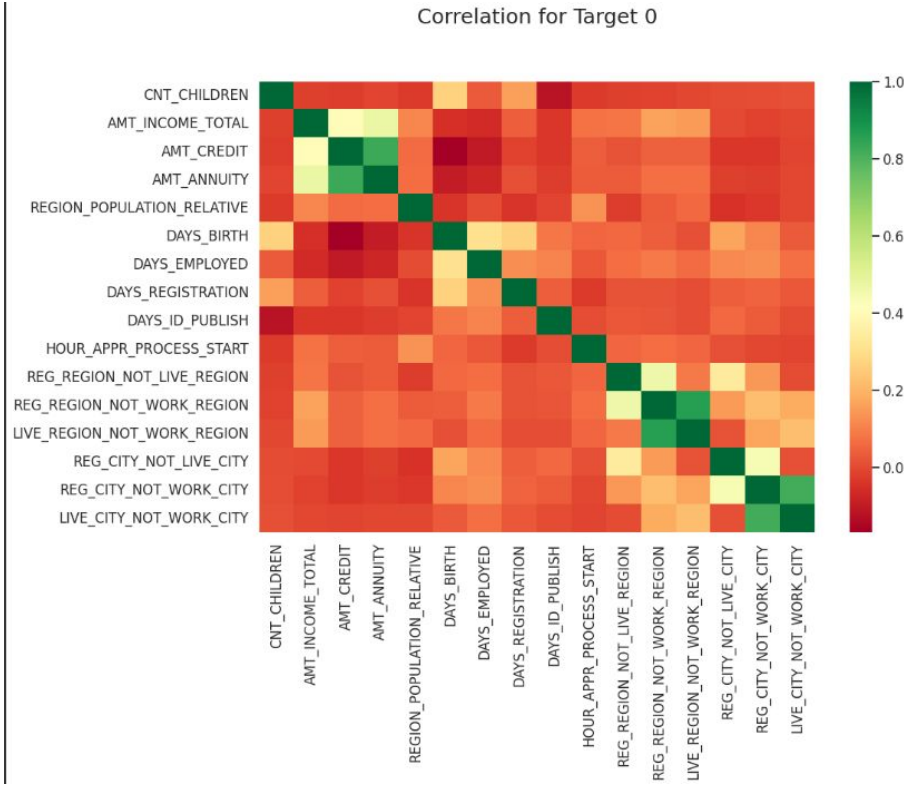
1. Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’ , ‘Other’ , ‘Medicine’ and ‘Government’.
2. Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.
3. Same as type 0 in distribution of organization type.



Correlation between Target 0 and 1

Heat maps plotting correlation for target 0

1. Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa.
2. Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.
3. Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.
4. less children client have in densely populated area.
5. Credit amount is higher to densely populated area



Correlation between Target 0 and 1

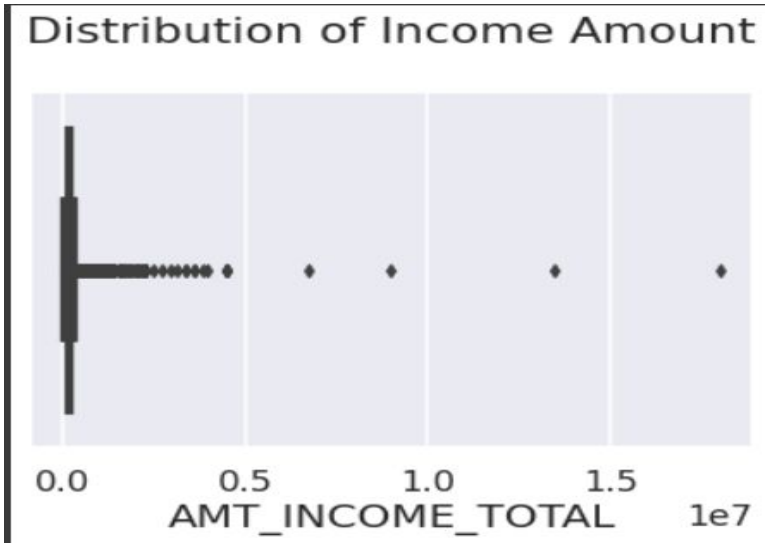
Heat maps plotting correlation for target 1

Here the heat map readings is almost same as observed for Target 0

1. The client's permanent address does not match contact address are having less children and vice-versa
2. The client's permanent address does not match work address are having less children and vice-versa

Checking for Outliners - Target 0

For Income Amount




```
target0_df.AMT_INCOME_TOTAL.describe()

count    2.303020e+05
mean     1.764984e+05
std      1.154998e+05
min      2.565000e+04
25%      1.125000e+05
50%      1.575000e+05
75%      2.160000e+05
max      1.800009e+07
Name: AMT_INCOME_TOTAL, dtype: float64
```

1. Some outliers are noticed in income amount.
2. The third quartiles is very slim for income amount.

Checking for Outliners - Target 0

For Credit Amount

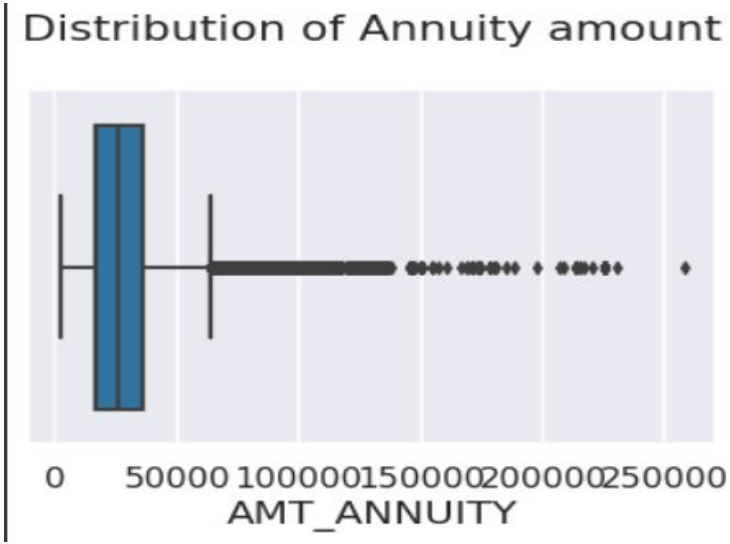

```
target0_df.AMT_CREDIT.describe()

count    2.303020e+05
mean     6.164879e+05
std      4.114378e+05
min      4.500000e+04
25%      2.762775e+05
50%      5.212800e+05
75%      8.353800e+05
max      4.050000e+06
Name: AMT_CREDIT, dtype: float64
```

1. Some outliners are noticed in credit amount
2. The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.

Checking for Outliners - Target 0

For Annuity Amount



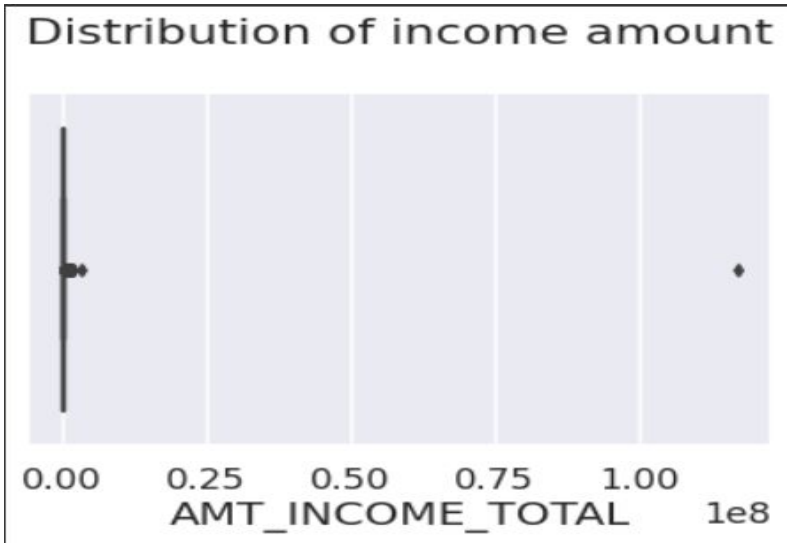

```
target0_df.AMT_ANNUITY.describe()
```

```
count    230302.000000
mean      27902.554759
std       14833.644504
min        1980.000000
25%       16969.500000
50%       25843.500000
75%       35743.500000
max      258025.500000
Name: AMT_ANNUITY, dtype: float64
```

1. Some outliers are noticed in annuity amount.
2. The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.

Checking for Outliners - Target 1

For Income Amount



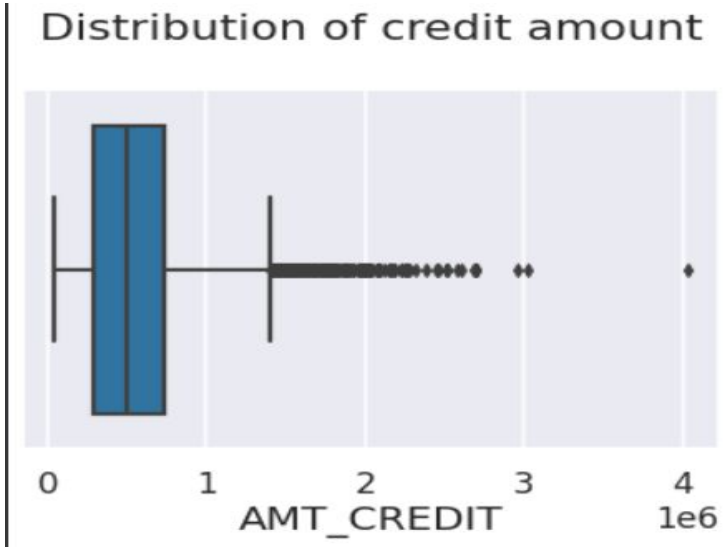

```
target1_df.AMT_INCOME_TOTAL.describe()
```

```
count    2.183500e+04
mean     1.697506e+05
std      7.956149e+05
min      2.700000e+04
25%      1.125000e+05
50%      1.440000e+05
75%      2.025000e+05
max      1.170000e+08
Name: AMT_INCOME_TOTAL, dtype: float64
```

1. Some outliers are noticed in income amount.
2. The third quartiles is very slim for income amount.
3. Most of the clients of income are present in first quartile.

Checking for Outliners - Target 1

For Credit Amount



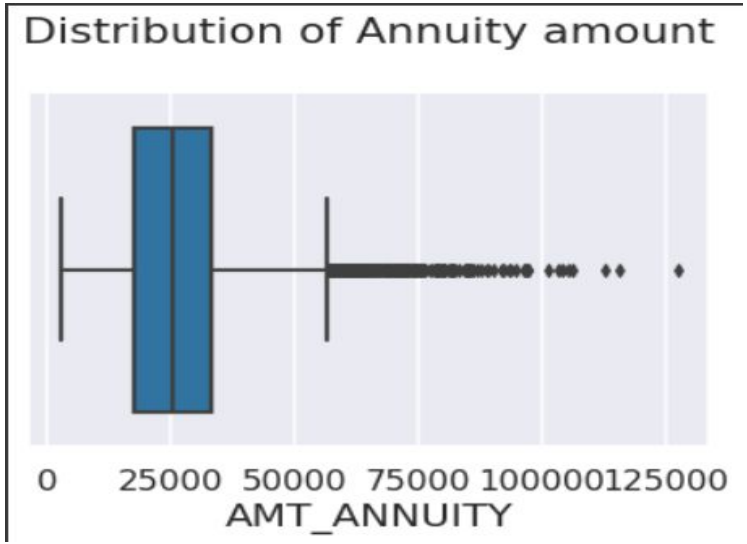

```
target1_df.AMT_CREDIT.describe()

count    2.183500e+04
mean     5.577178e+05
std      3.460483e+05
min      4.500000e+04
25%      2.844000e+05
50%      4.959855e+05
75%      7.290000e+05
max      4.027680e+06
Name: AMT_CREDIT, dtype: float64
```

1. Some outliers are noticed in credit amount.
2. The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.

Checking for Outliners - Target 1

For Annuity Amount




```
target1_df.AMT_ANNUITY.describe()

count     21835.000000
mean      26859.040669
std       12476.177108
min        2844.000000
25%       17732.250000
50%       25578.000000
75%       33394.500000
max      127507.500000
Name: AMT_ANNUITY, dtype: float64
```

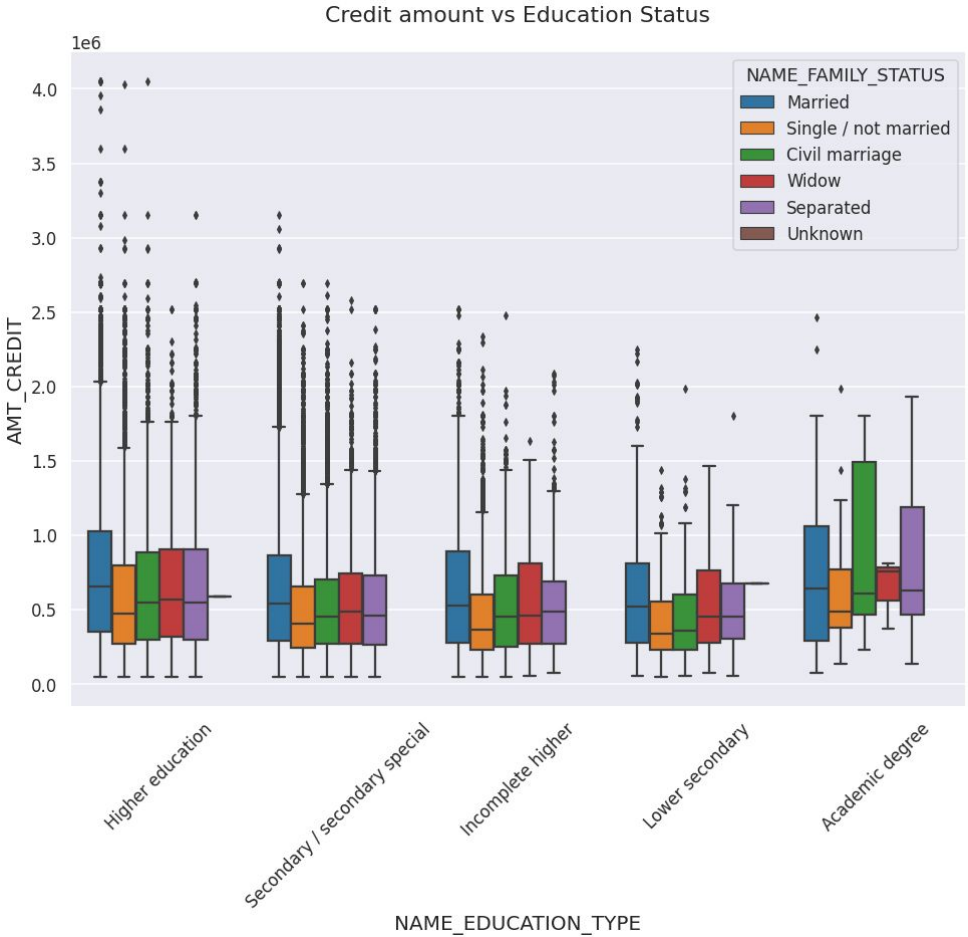
1. Some outliers are noticed in annuity amount.
2. The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.

Bivariate Analysis

For Target 0

Credit Amount vs Education Status

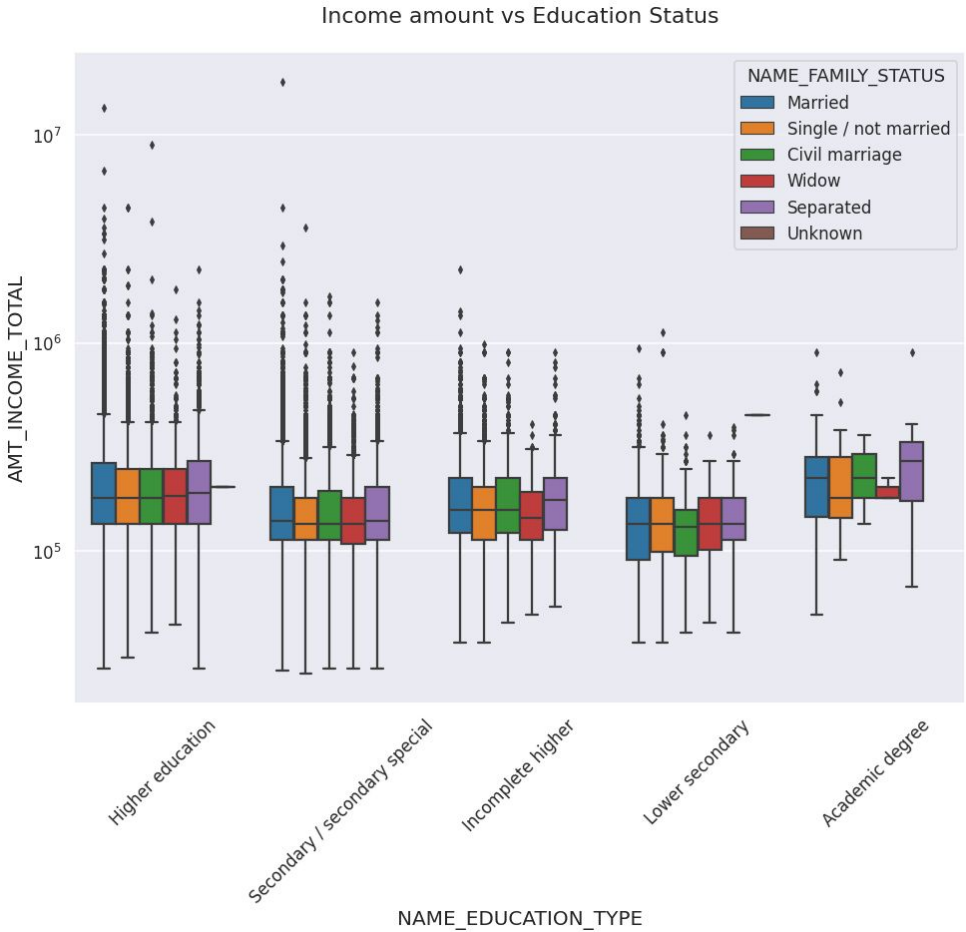
1. Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
2. Also, higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers.
3. Civil marriage for Academic degree is having most of the credits in the third quartile.



For Target 0

Income Amount vs Education Status

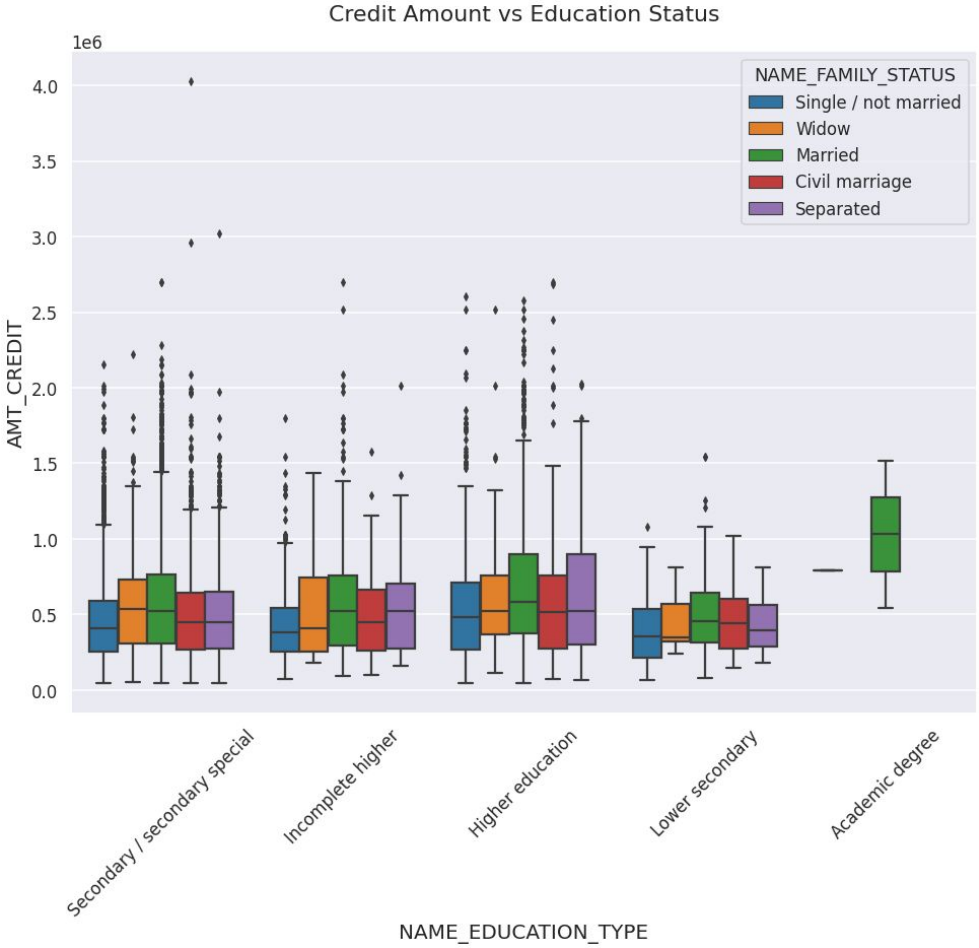
1. For Education type 'Higher education' the income amount is mostly equal with family status.
2. It does contain many outliers. Less outlier are having for Academic degree but there income amount is little higher that Higher education.
3. Lower secondary of civil marriage family status are have less income amount than others.



For Target 1

Credit Amount vs Education Status

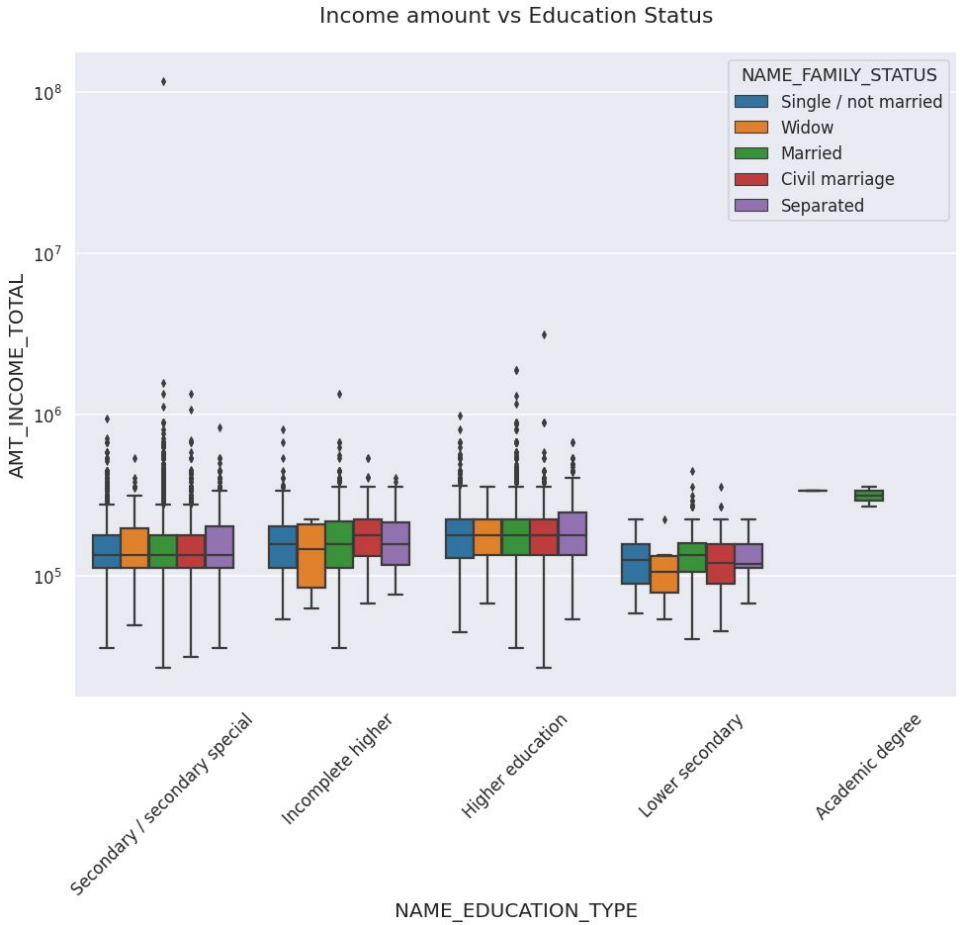
1. This is seen quite similar with Target 0
2. Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
3. Most of the outliers are from Education type 'Higher education' and 'Secondary'.
4. Civil marriage for Academic degree is having most of the credits in the third quartile.



For Target 1

Income Amount vs Education Status

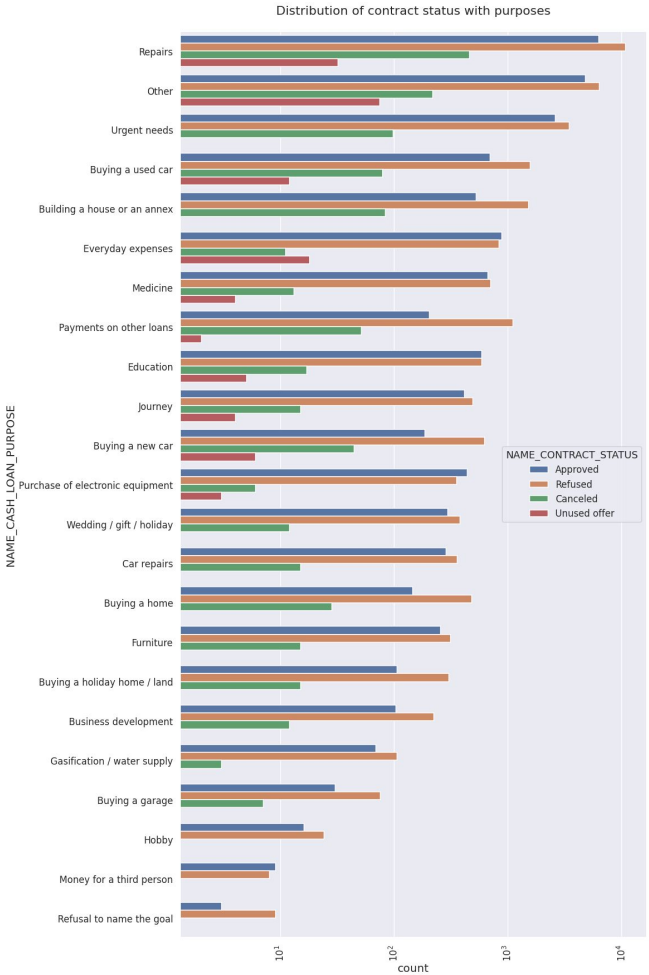
1. Have some similarity with Target0,
2. For education type 'Higher education', the income amount is mostly equal with family status.
3. Less outlier are having for Academic degree but there income amount is little higher that Higher education.
4. Lower secondary are have less income amount than others.



Univariate Analysis after merging the Previous Application Data

Contract status with Purpose Plot

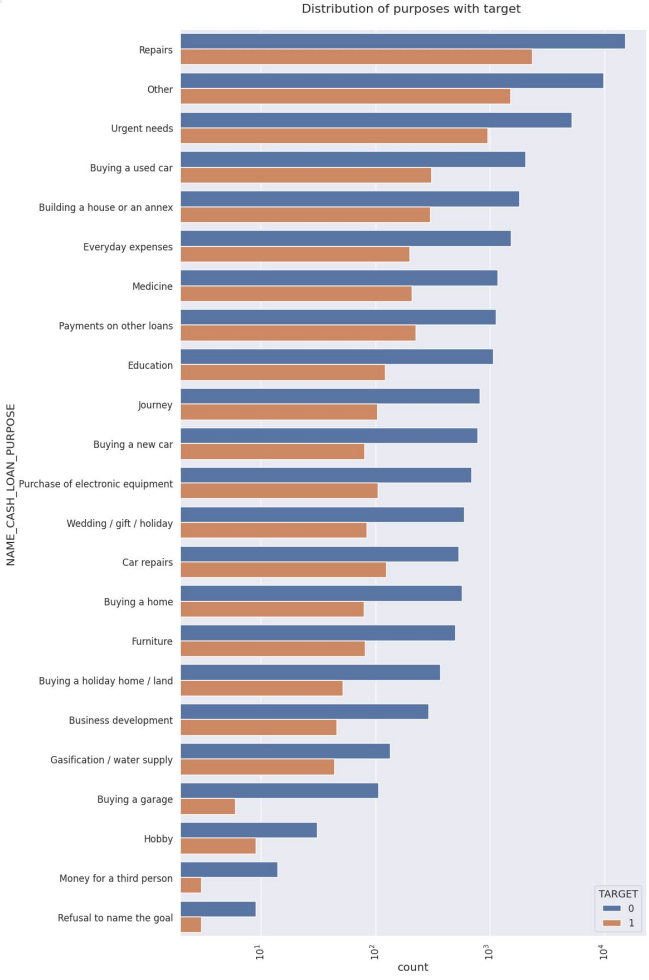
1. Most rejection of loans came from purpose 'repairs'.
2. For education purposes we have equal number of approves and rejection
3. Paying other loans and buying a new car is having significant higher rejection than approves.



Target With Purpose Plot

1. Loan purposes with 'Repairs' are facing more difficulties in payment on time.
2. There are few places where loan payment is significant 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 for minimal payment difficulties.



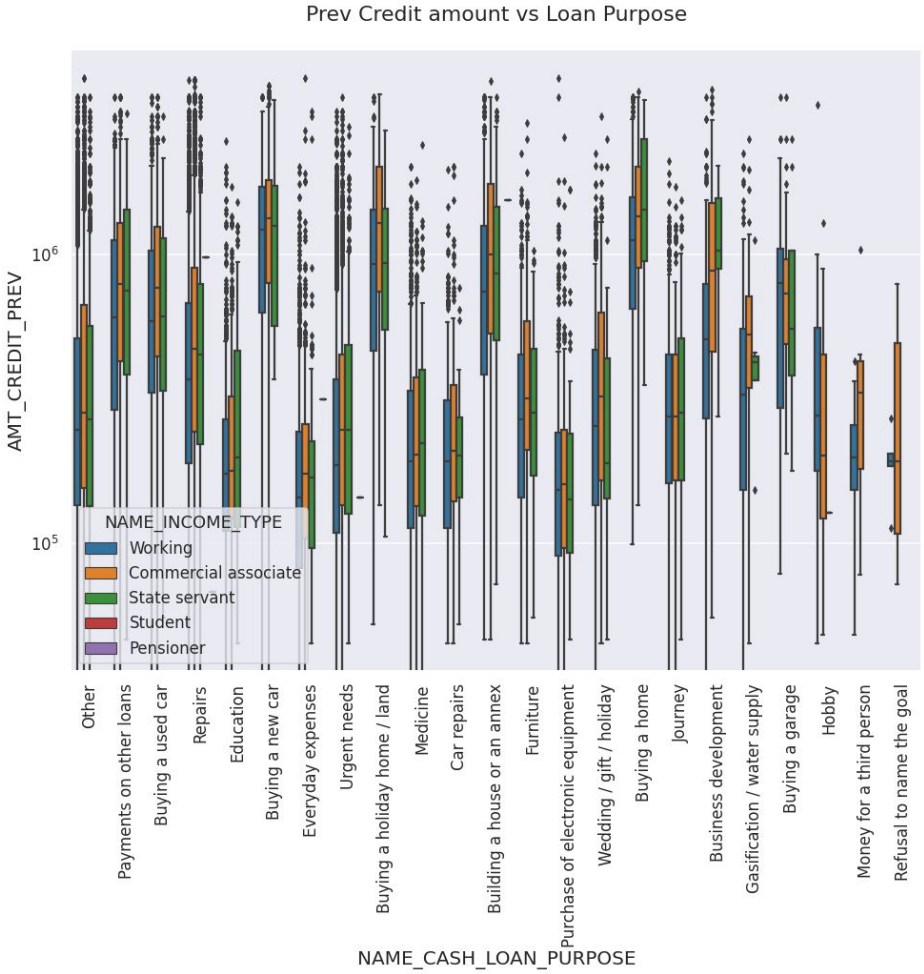
Bivariate Analysis after merging the Previous Application Data

Prev Credit amount vs Loan Purpose

The credit amount of Loan purposes like 'Buying a home','Buying a land','Buying a new car' and 'Building a house' is higher.

Income type of state servants have a significant amount of credit applied

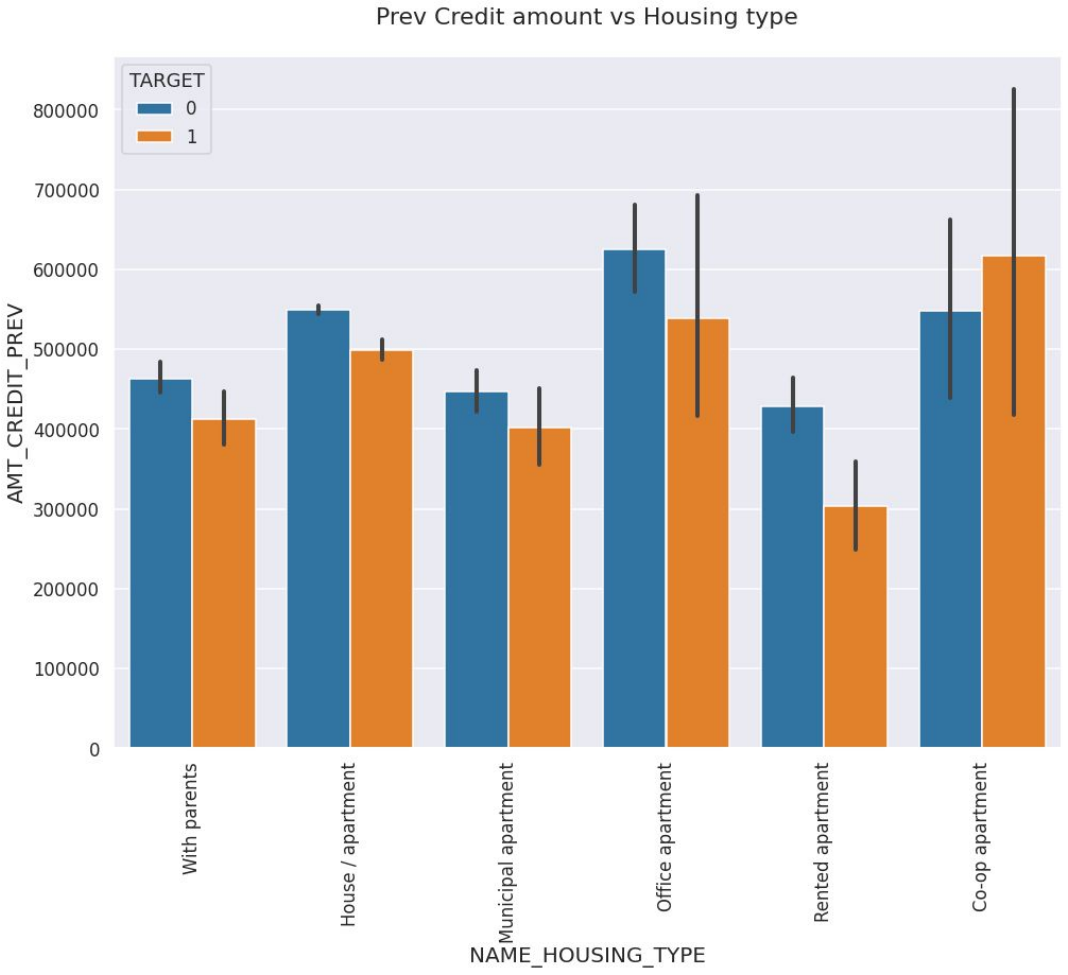
Money for third person or a Hobby is having less credits applied for.



Prev Credit amount vs Housing type

Here for Housing type, office apartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1.

So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank 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 ‘Co-op apartment’ for successful payments (loan clears).
2. Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.
3. Also with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time. Have to try and avoid them as much as possible
4. Get as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.
5. Females have majority of credits from the bank and higher rates of successful payments. Although this can’t be a clear indication, this proves that they repay in time.

THANK YOU