DATA ANALYSIS (EDA CASE) PROJECT

BY,

KUNAL MORE

AIM

- To understand the driving factors behind load default.
- 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 (too risky applicants) at a higher interest rate, etc.
- Identification of such applications using EDA is the basic aim of this case study.

Analysis & Approach:

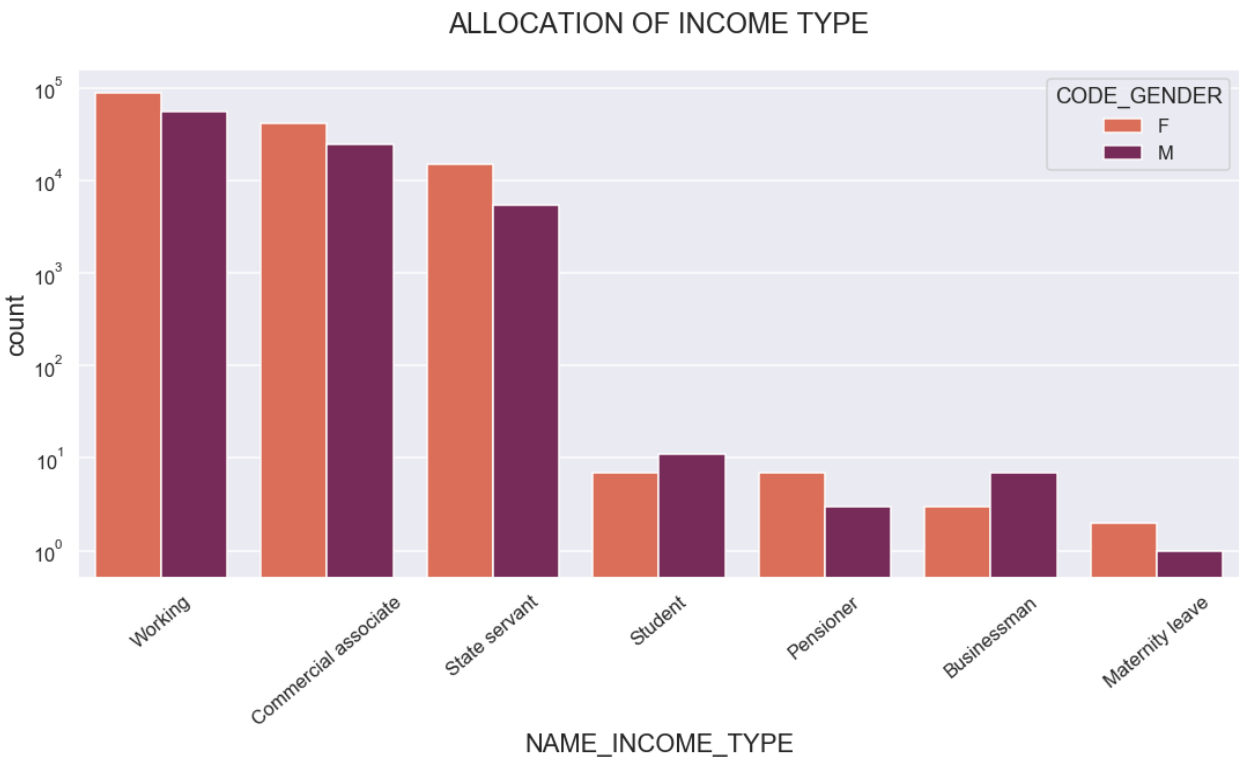
- Importing all the necessary library.
- Reading the dataset from local.
- Finding the shape of dataset.
- So, the concerned column is having very few null values rows. Hence let's try to impute the missing values
- Since this column is having an outlier which is very large it will be inappropriate to fill those missing values with mean, Hence Median comes to rescue for this and we will fill those missing banks with median value.
- Searching the column for null values.
- Removing rows having null values greater than or equal to 30%.
- Removing unwanted columns from this dataset.
- 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 and implement suitable techniques on them to fill those missing values or to delete them.
- Finding categorical columns having these 'XNA' values.
- Since, Female is having the majority and only 4 rows are having NA values, we can update those columns with Gender 'F' as there will be no impact on the dataset.
- For column 'ORGANIZATION_TYPE', we have total count of 307511 rows of which 55374 rows are having 'XNA' values. Which means 18% of the column is having this values. Hence, if we drop the rows of total 55374, will not have any major impact on our dataset.
- Hence, dropping the rows of total 55374 have 'XNA' values in the organization type column.

-:Categorical Univariate Analysis in logarithmic scale for target = 0:-

- **Points to be concluded from the above graph.**

Female counts are higher than male. This graph show that females are more than male in having credits for that range. Income ranges from 100000 to 200000 is having higher number of credits. Very less count for income ranges from 400000 and above. Also, in some ranges Males counts is higher (400000 - 425000). There's a huge disparity in between females and males in the range (25000-50000).

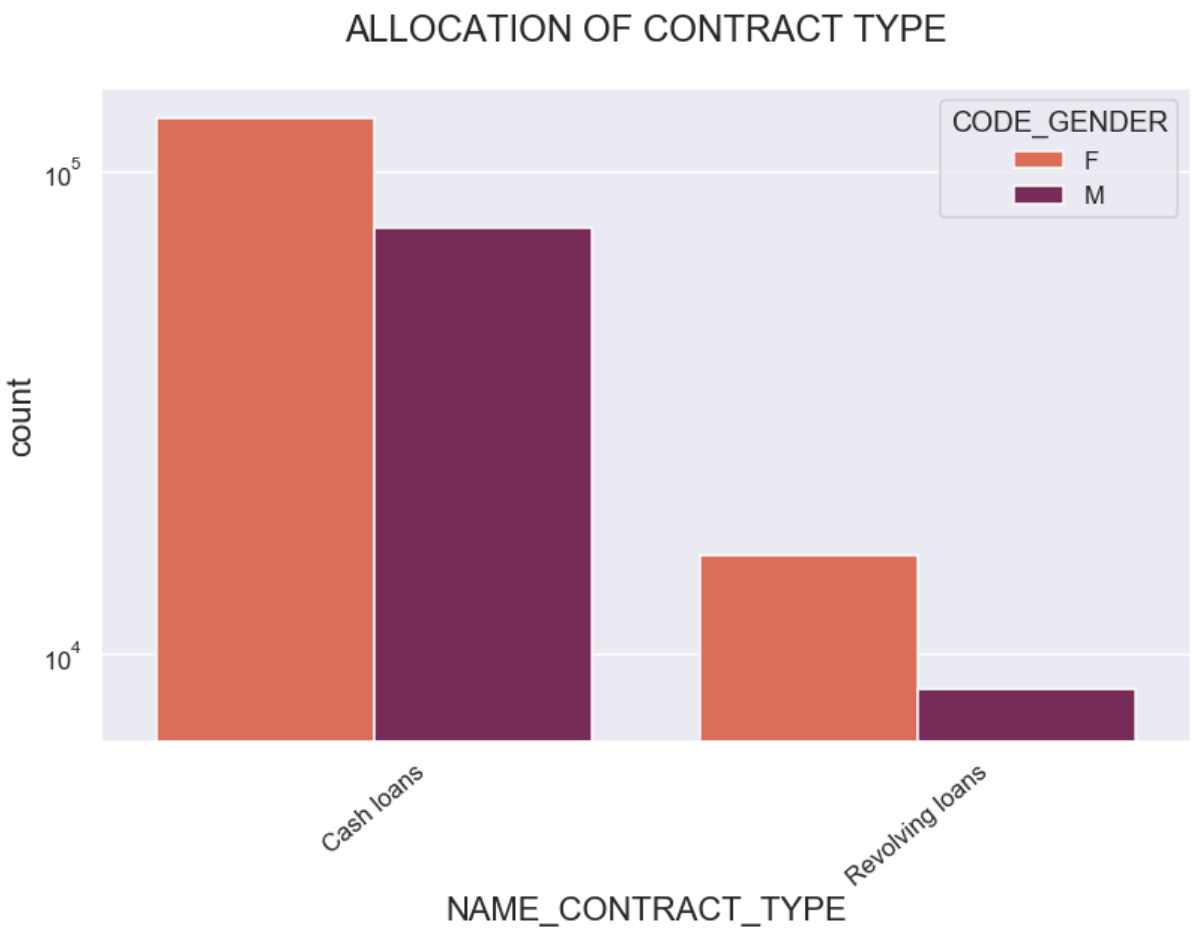
-:ALLOCATION OF INCOME TYPE:-



- **Points to be concluded from the above graph.**

For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits were higher than others. Also, we can observe that the Females are having more number of credits than male for this three income types. Maternity leave having the least incomes among others. Lesser number of credits were present for income type ‘student’, ’pensioner’, ‘Businessman’ and ‘Maternity leave’. Only in income types ‘student’ and ‘Businessman’ the male has more no. of credits than female.

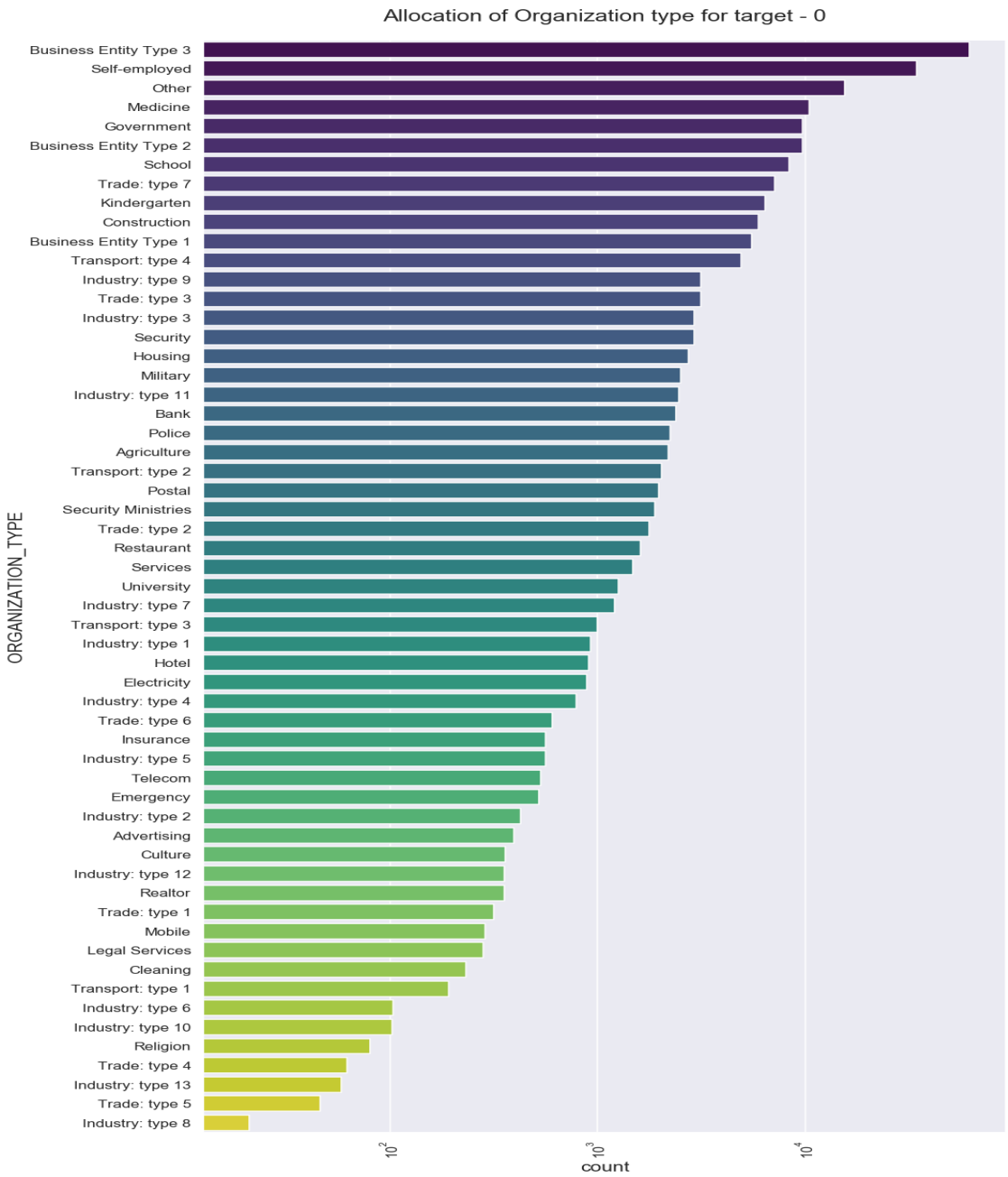
-:ALLOCATION CONTRACT TYPE GRAPH:-



- **Points to be concluded from the above graph**.

  As we can observe from this graph that the contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type. Females were having higher count in cash loans as well as in revolving loans.

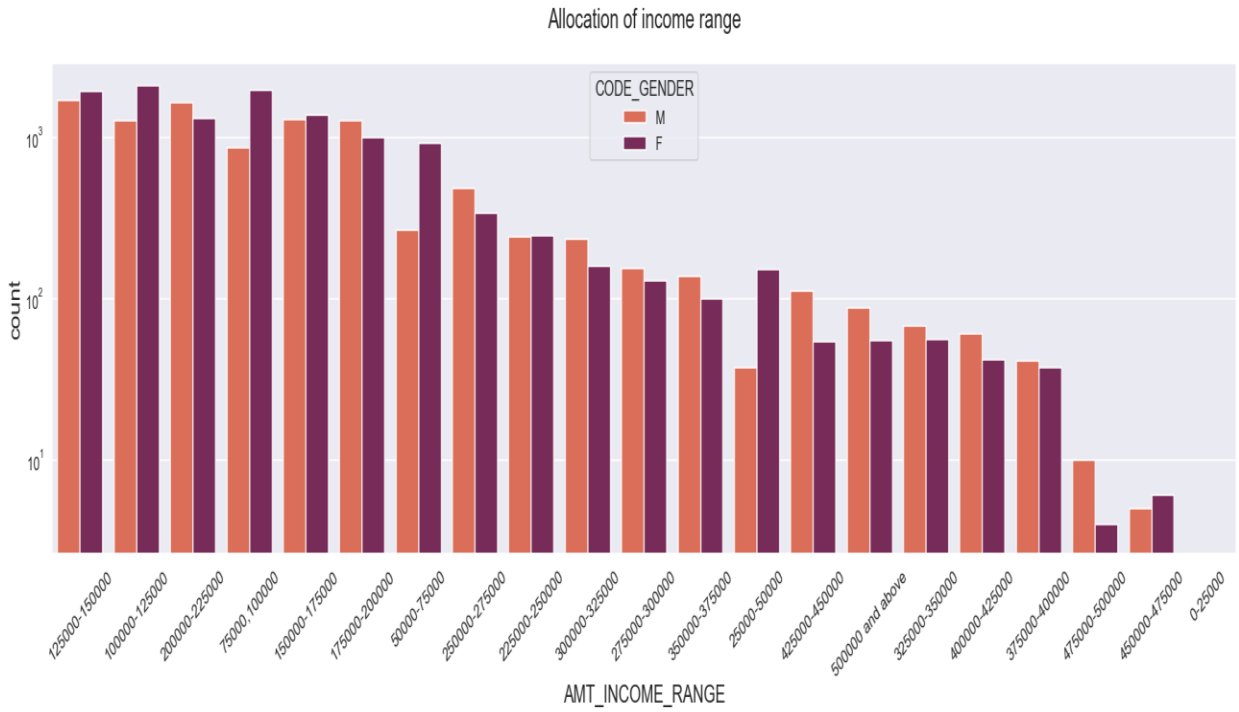
-:Distribution of Organization type Target - 0:-



- **Points to be concluded from the above graph.**

From this graph we can see that the Clients which have applied for credits are from the organization type ‘Business entity Type 3’, ‘Self-employed’, ‘Other’, ‘Medicine’ and ‘Government’. Also, Industry: type 6 & 10 have the same no of Clients. Lesser clients are from Industry type 8, type 6, type 10, religion and trade type 5, type 4.

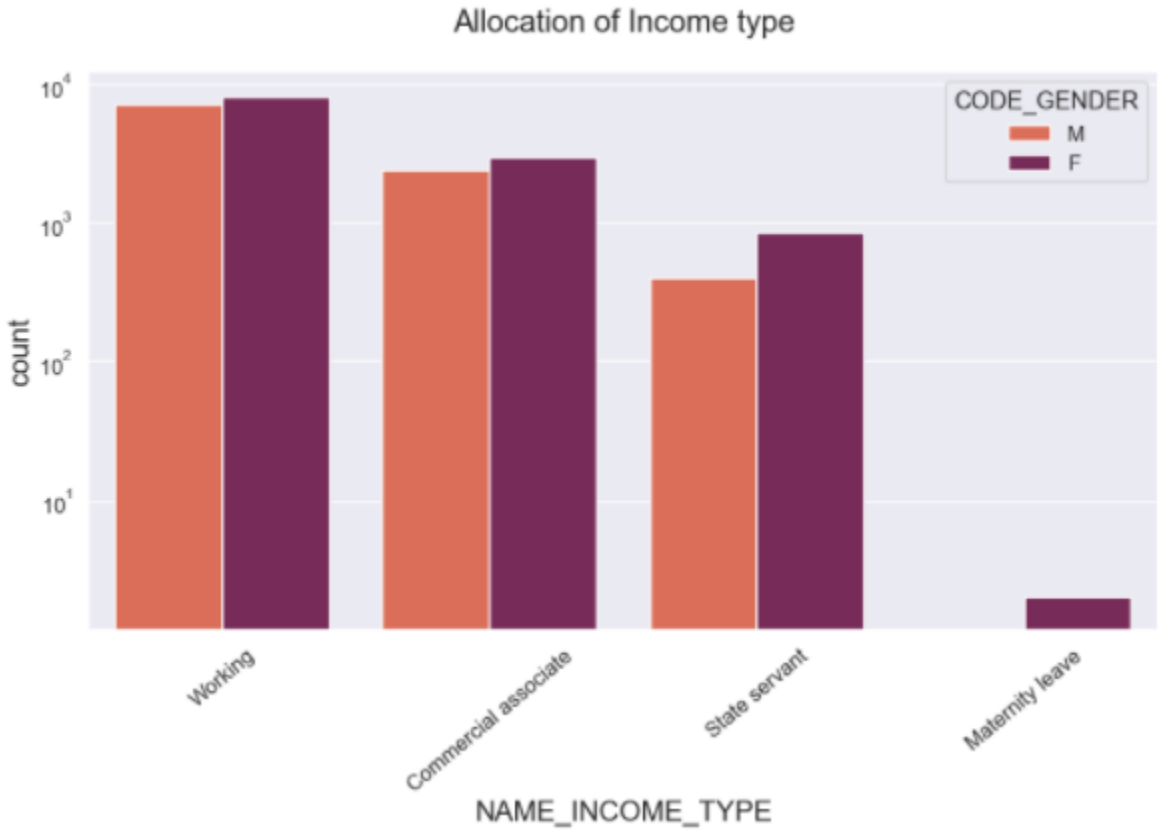
-:Categoroical Univariate Analysis in logarithmic scale for target = 1:-



- **Points to be concluded from the above graph.**

  Male counts are higher than female. This graph show that males are more than female in having credits for that range. Income ranges from 100000 to 200000 is having higher number of credits. Lesser counts for income ranges 400000 and above. Unlike for target 0 this one is completely male dominated. Male counts have been higher and we seen earlier as the income range increases male counts comes into place.

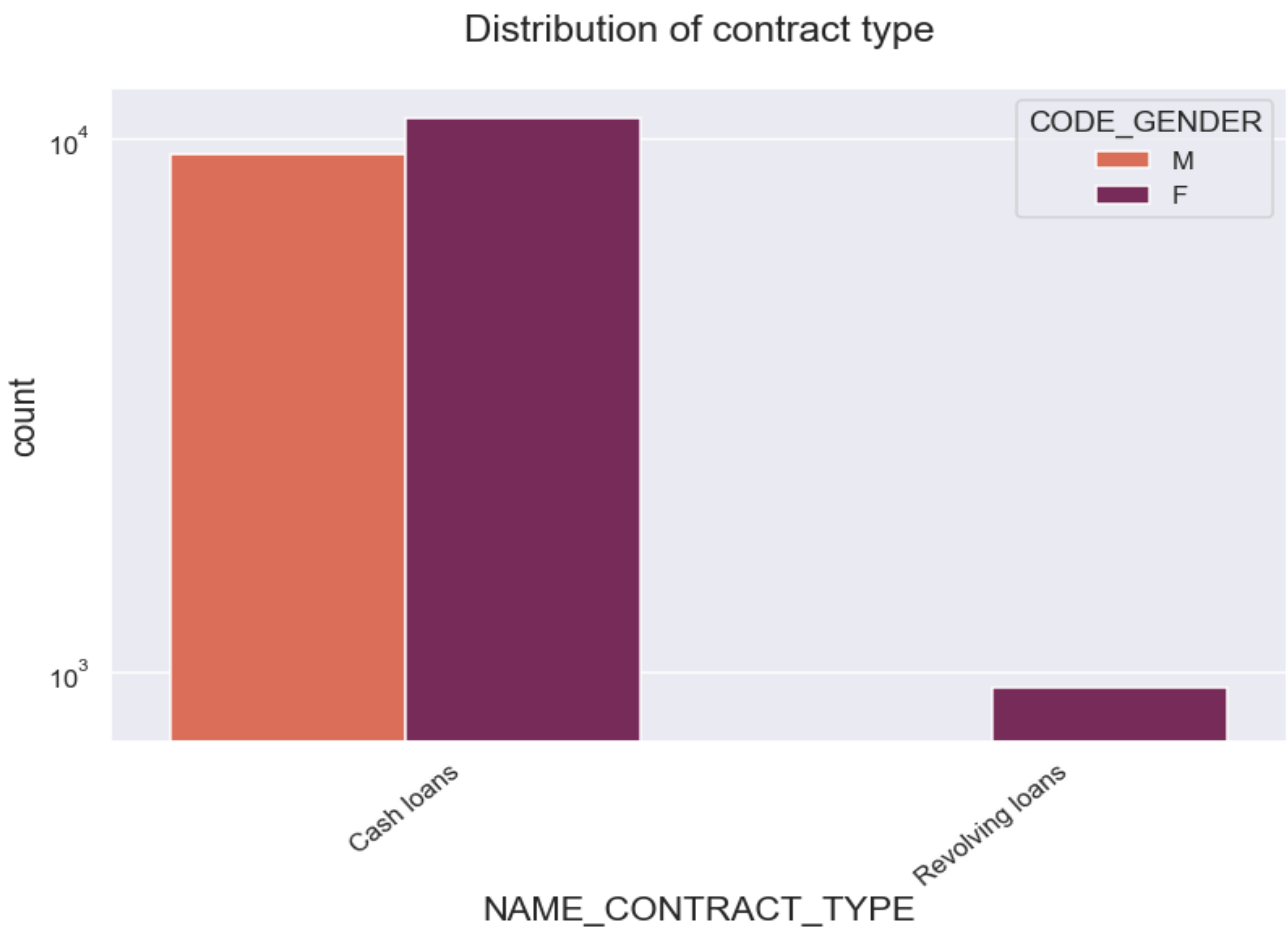
-:ALLOCATION OF INCOME TYPE:-



- **Points to be concluded from the above graph.**

  In the above graph we can clearly observe that for income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than other i.e. ‘Maternity leave. For this Females are having more numbers of credits than male. Also, Maternity leave having lesser number of credits for income types and there were no male counts present here. For type 1: There is no income type for ‘student’, ’pensioner’ and ‘Businessman’ which means they don’t do any late payments.

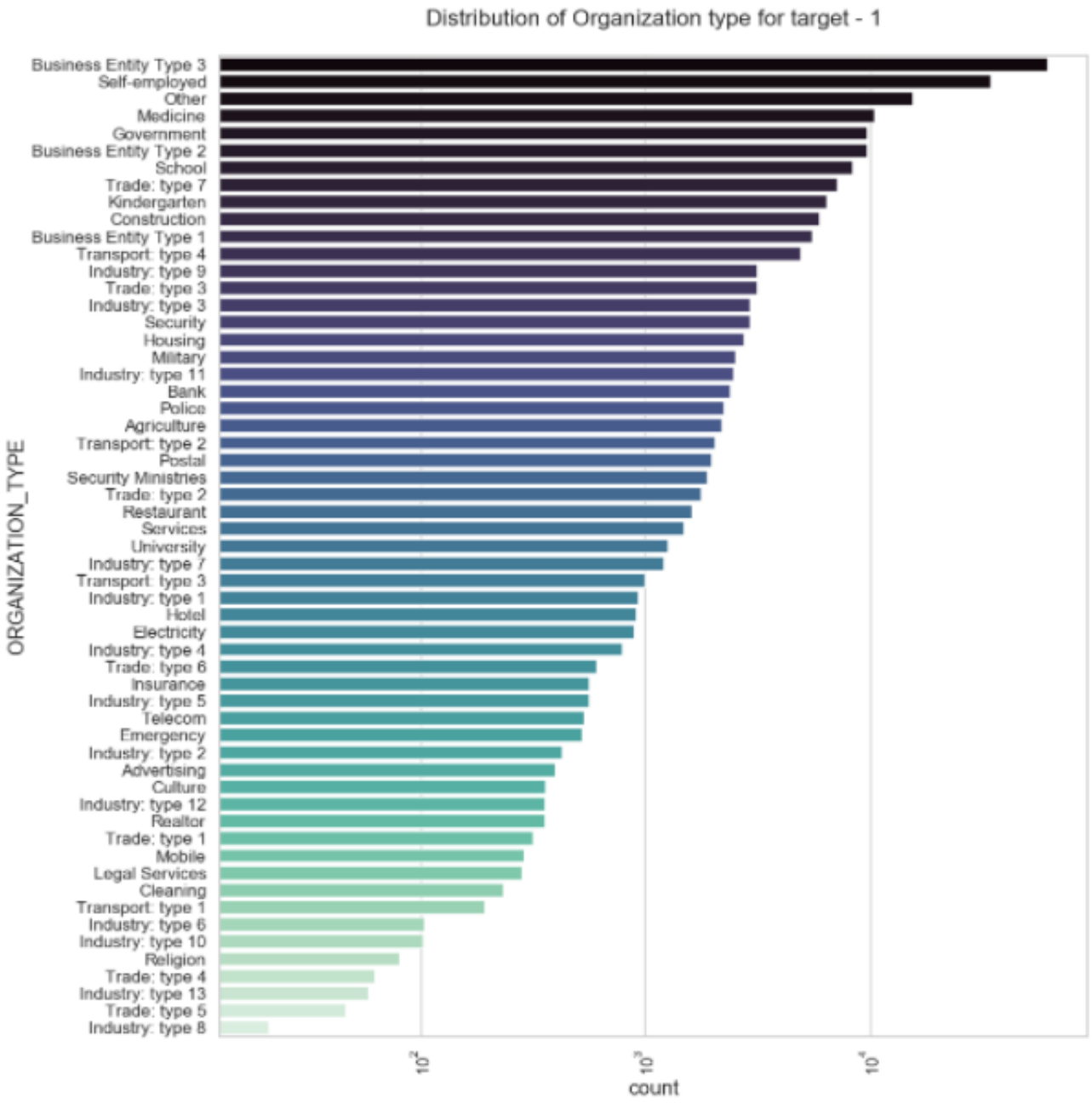
-:Distribution of Contract Type:-



- **Points to be concluded from the above graph.**

  In this graph the contract type 'Cash loans' is having higher number of credits than 'Revolving loans' contract type. Again, the Female is leading for applying more number of credits. Also, there is no Males counts present in Revolving loans. For type 1 have only Female Revolving loans.

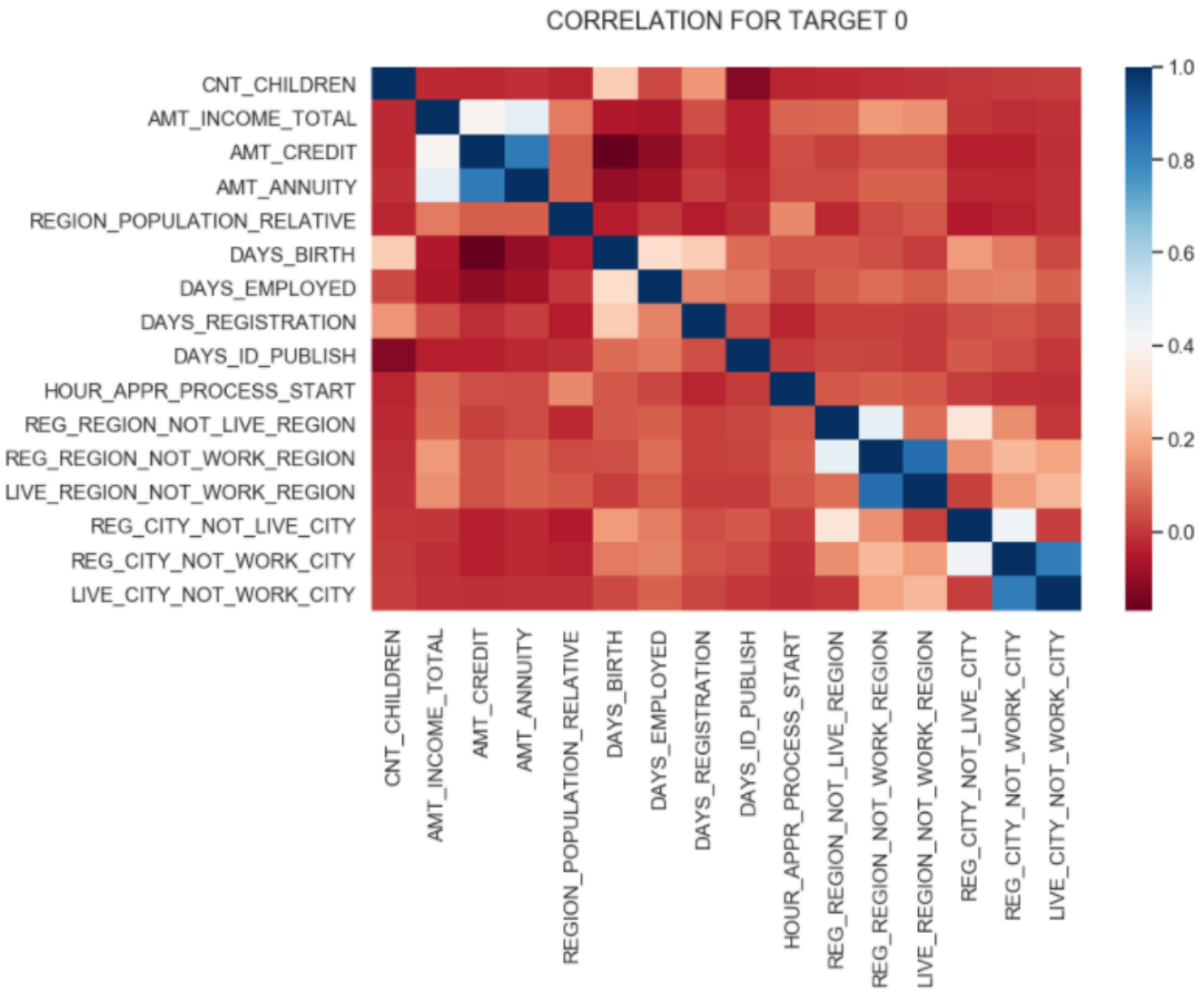
-:Distribution of Organization type for target – 1:-



- **Points to be concluded from the above graph.**

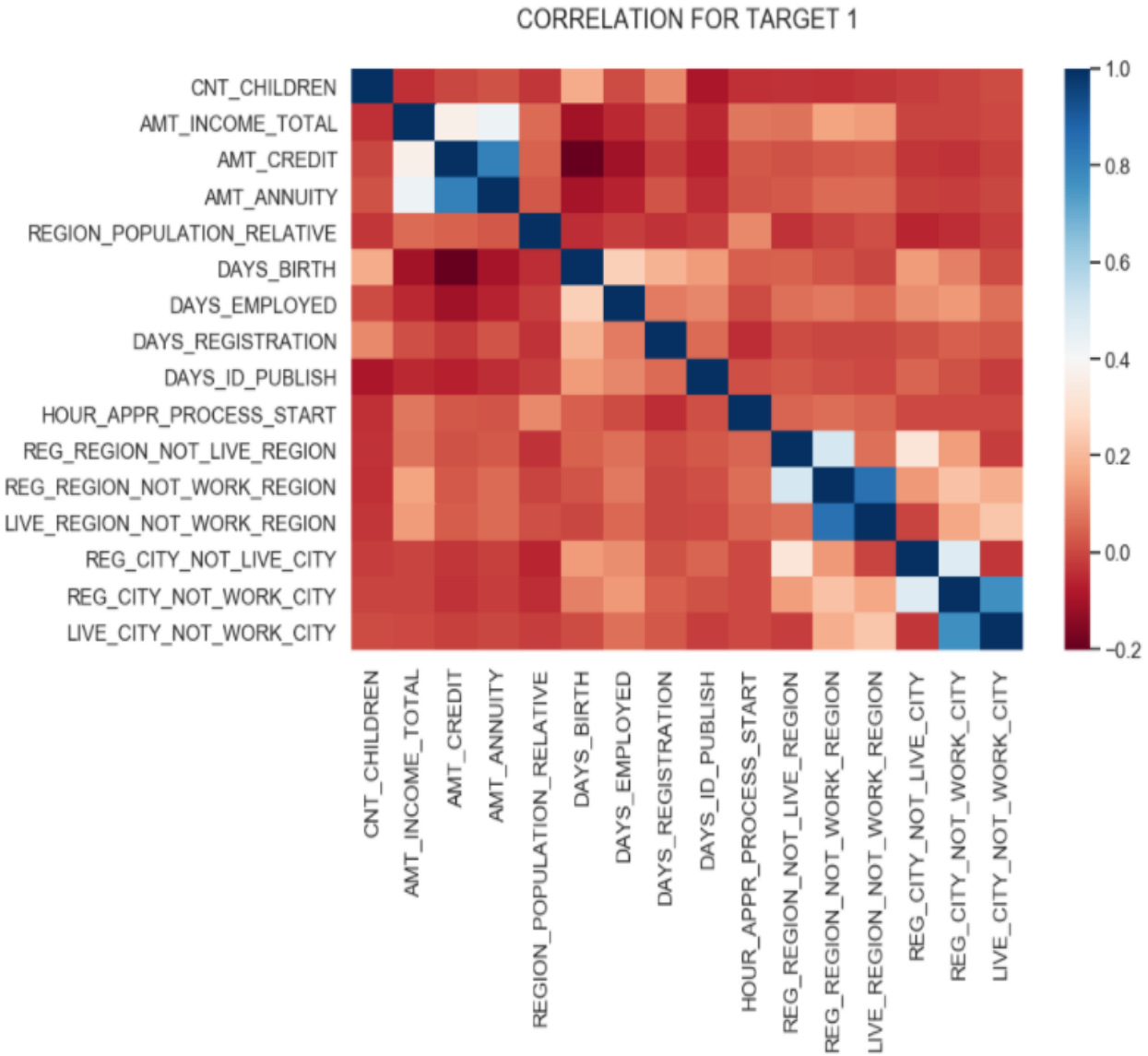
From the above observation we can see that the clients which have applied for credits are from most of the organization type ‘Business entity Type 3’, ‘Self-employed’, ‘Other’, ‘Medicine’ and ‘Government’. Lesser clients are from Industry type 8, type 6, type 10, religion and trade type 5, type 4. Second least is Trade: type 5 Same as type 0 in distribution of organization type. Also, this graph is mostly similar from Target 0.

-:CORRELATION FOR TARGET – 0:-



HEATMAP FOR TARGET- 0

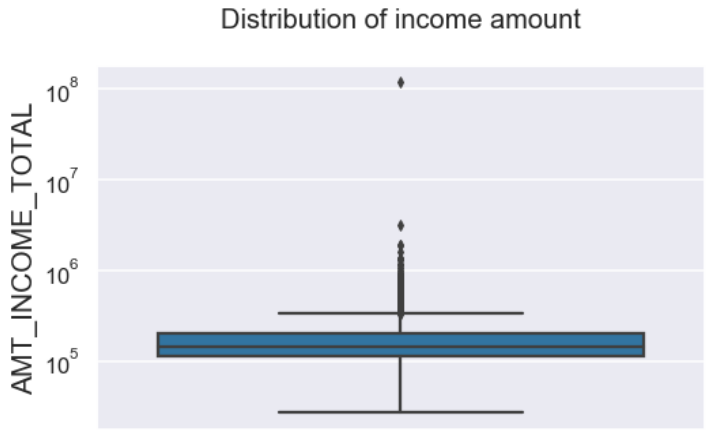
-:CORRELATION FOR TARGET – 1:-

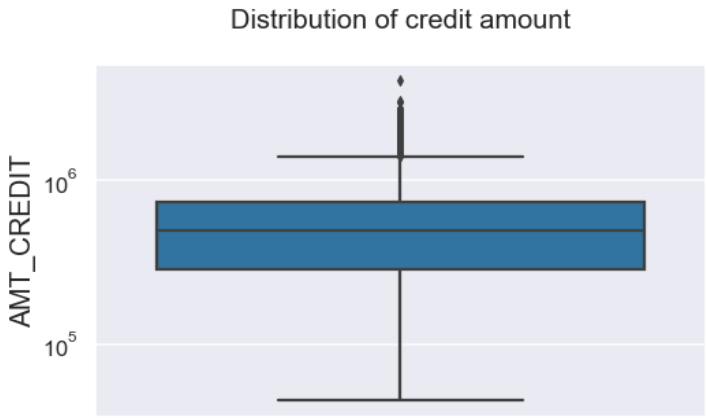


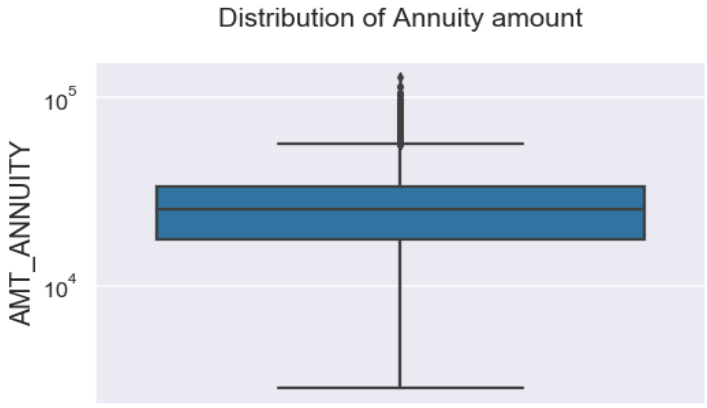
HEATMAP FOR TARGET- 1

-:UNIVARIATE ANALYSIS:-

TARGET – 0

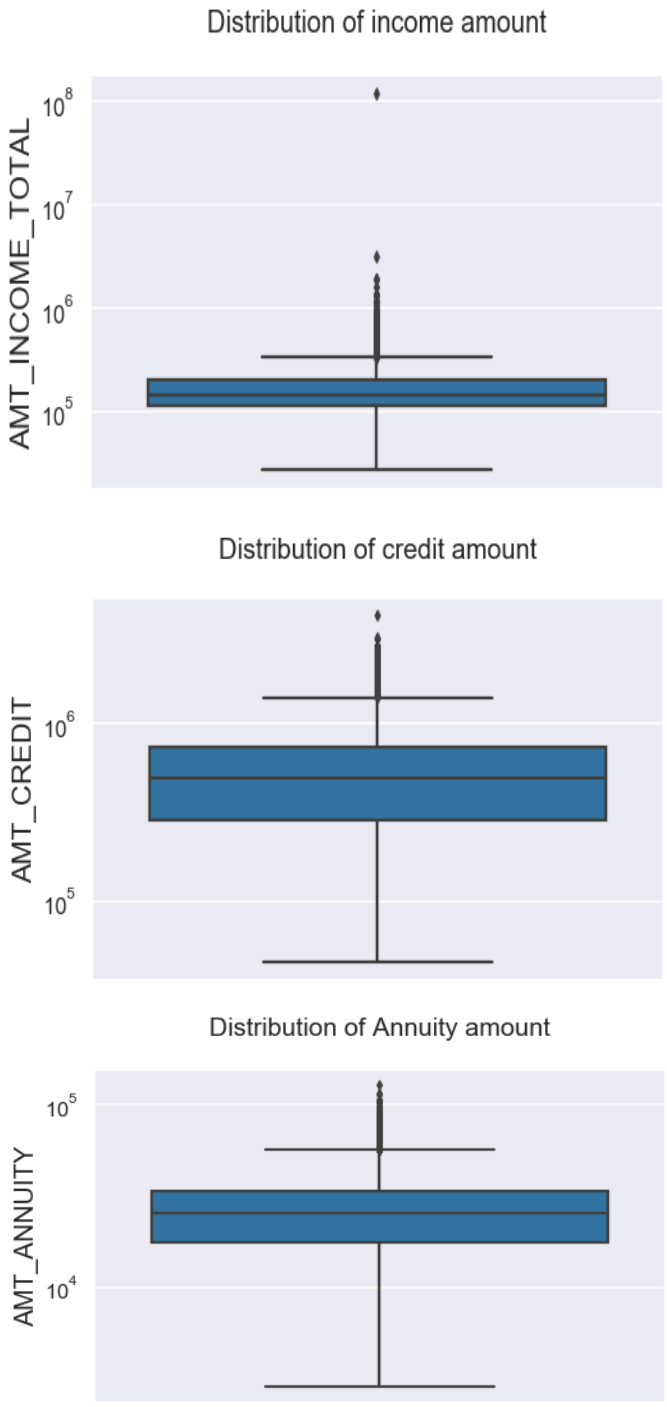






In all of the above boxplots there were some outliers presents, which can be seen as it is outside of both the IQR(Inter Quartile Range).

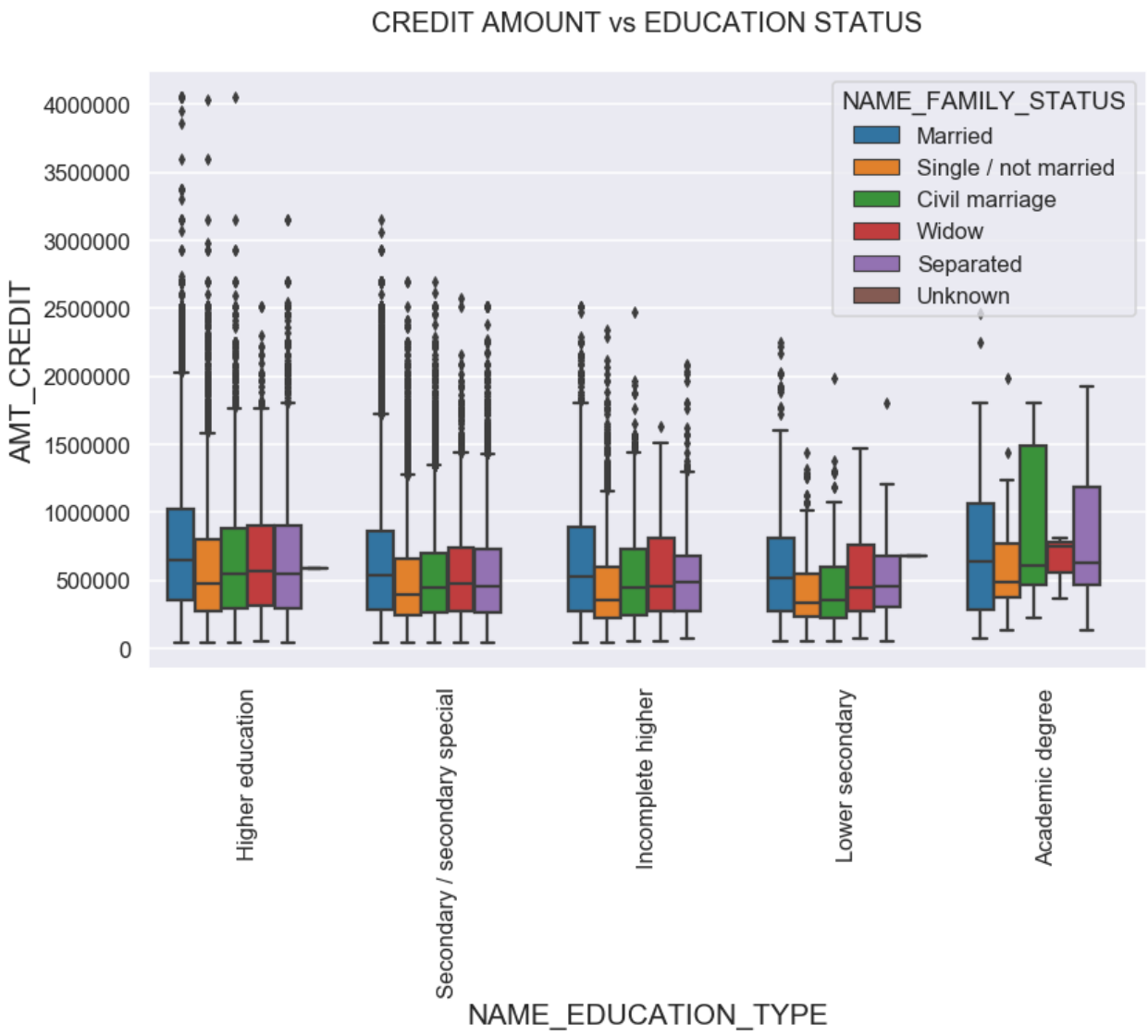
TARGET – 1



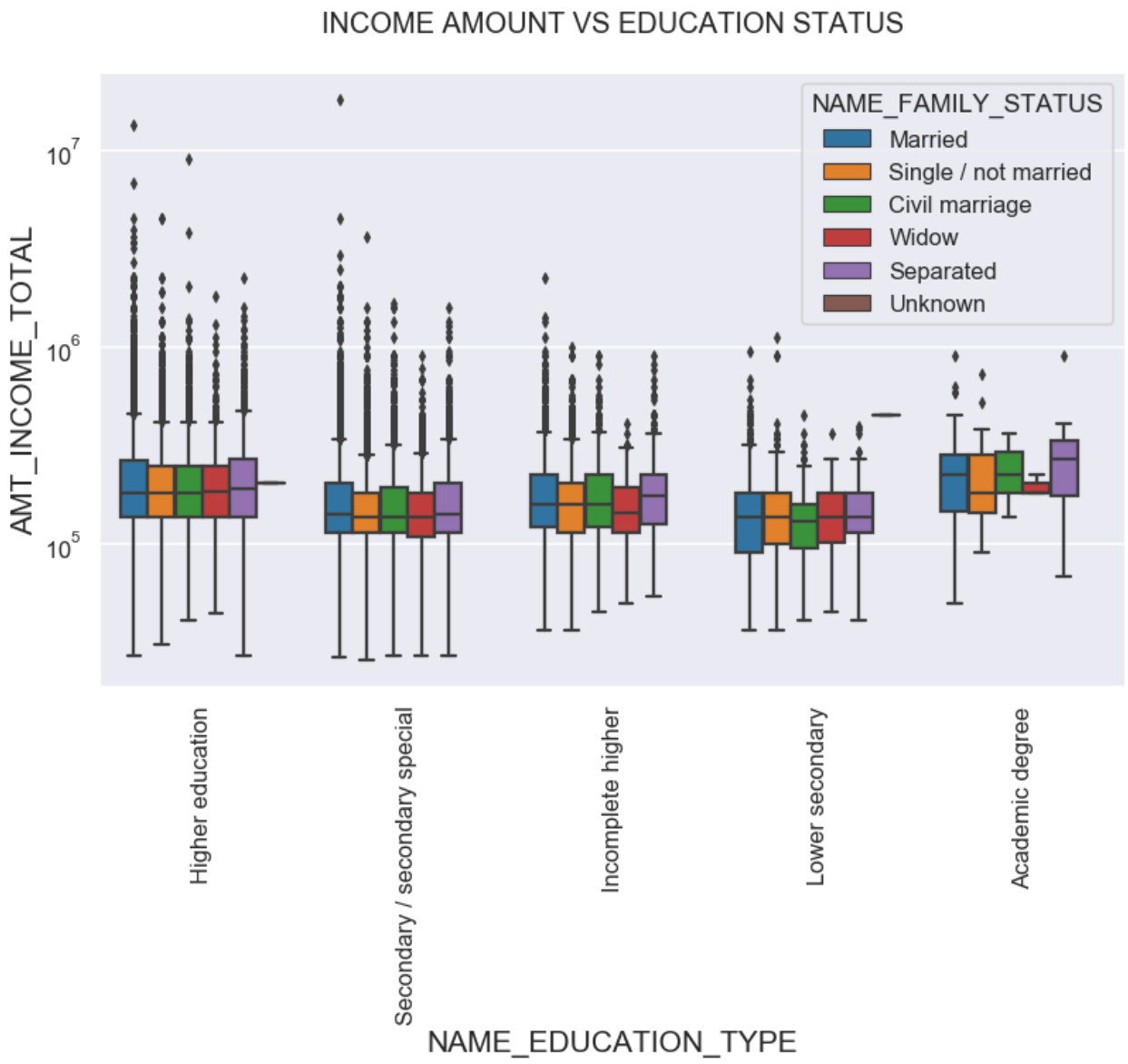
In all of the above boxplots there were some outliers which can be seen as it is outside of both the IQR(Inter Quartile Range). The boxplots are for distribution of income, credit and annuity amounts for target-1.

BIVARIATE ANALYSIS

TARGET – 0

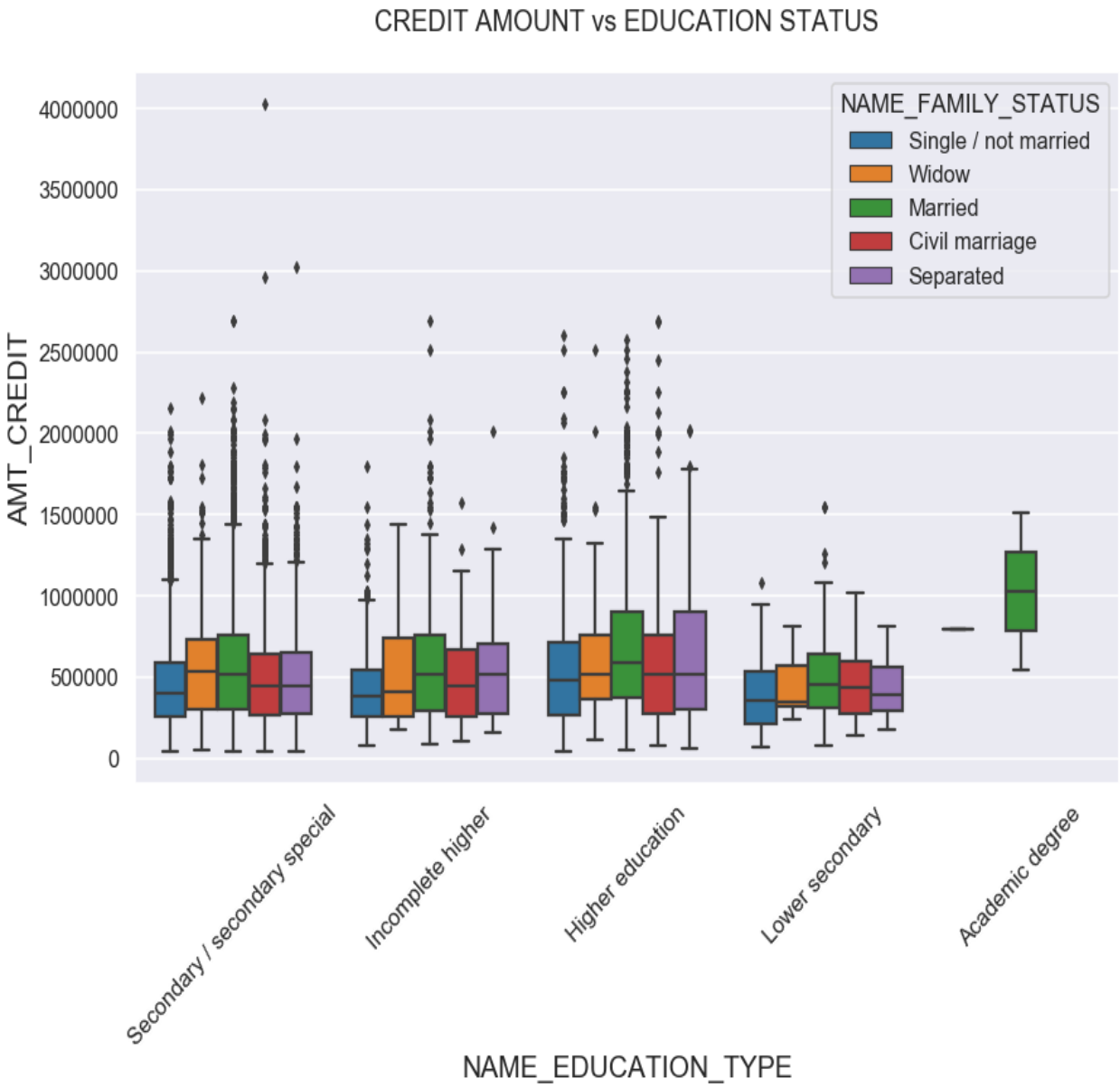


The above boxplot represents the Credit amt vs Edu Stats. From the above box plot we can conclude that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others. Civil marriage for Academic degree is having most of the credits in the third quartile. Also, higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers.

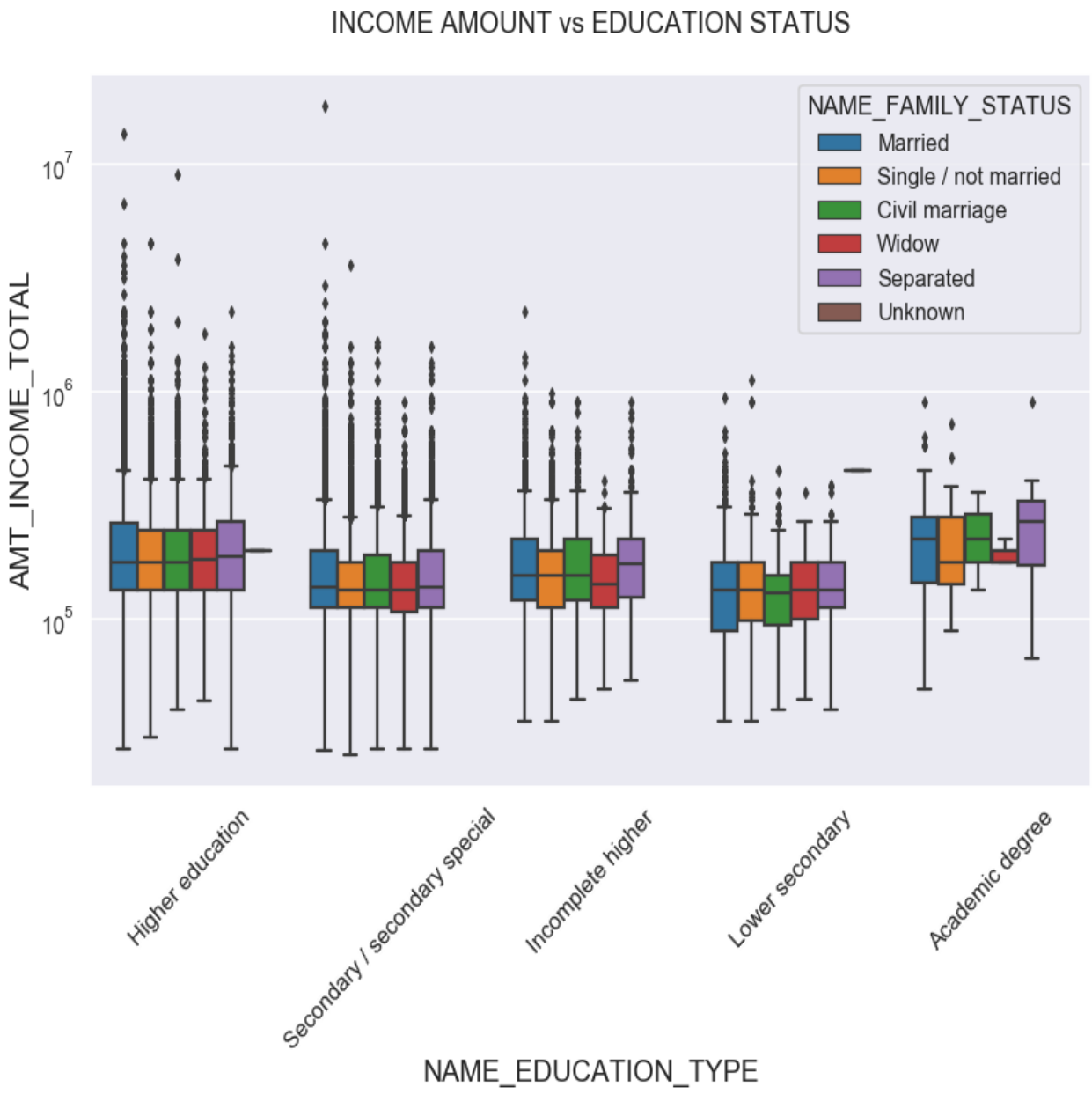


The above boxplot represents the Income amt vs Edu Stats. From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status. It does contain many outliers. Less outliers are having Academic degree but there income amount is little higher that Higher education. Lower secondary of civil marriage family status are having less income amount than others. From the above graph we can conclude that the more educated people are there the more they tend to earn a good income which results the reason why lower secondary has not been faring well in this area.

TARGET – 1

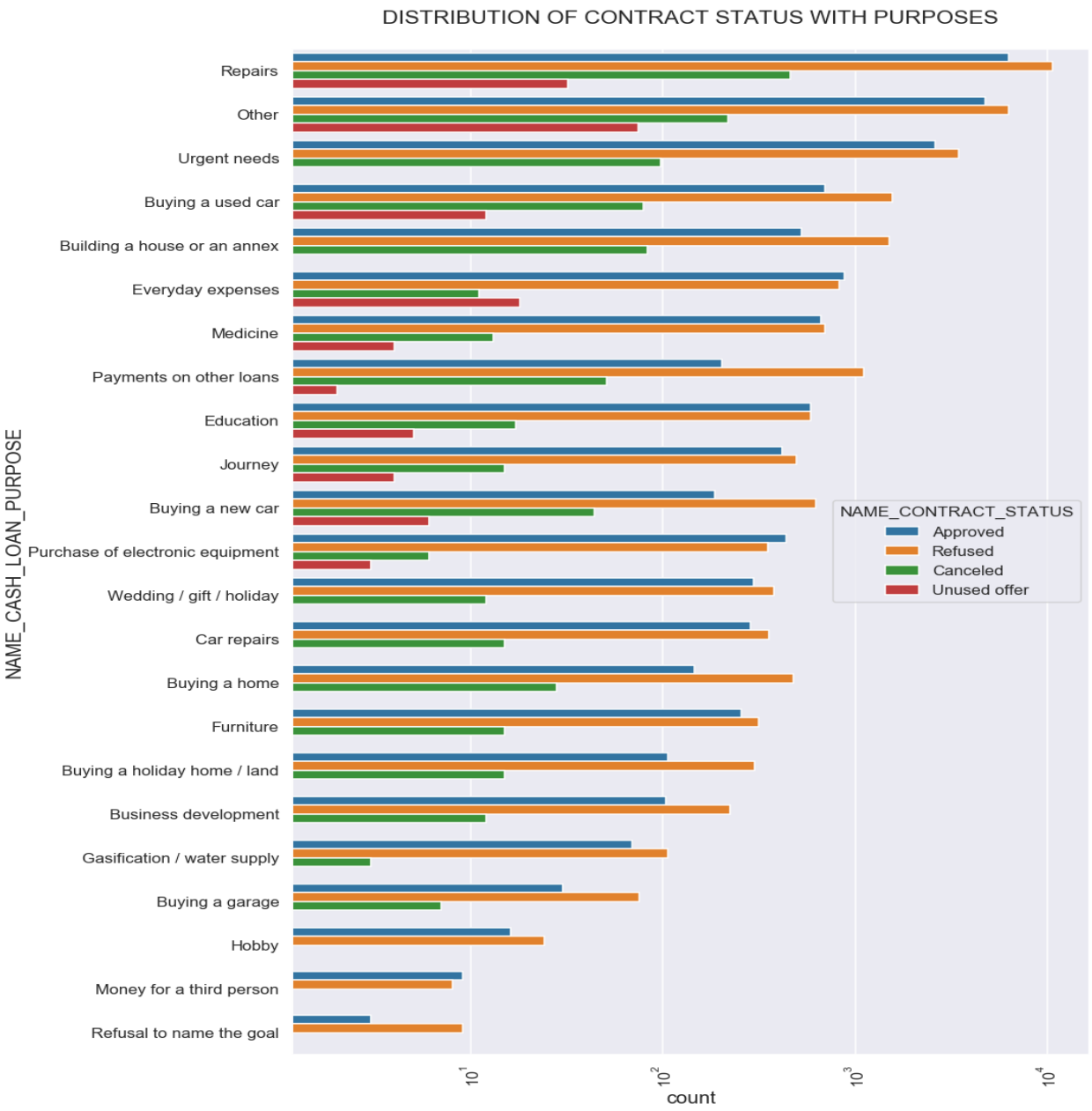


The above boxplot is for credit amount and education status for target 1. It can be seen that in the academic degree field only married people have done fairly well which was not the case with target 0. The common thing although between both the target variables will be that most of the outliers are from the higher education and secondary/secondary special segment had most of the outliers mainly in the status of married and single. It would also be fair to say that the married family status segment has been the most consistent in each of the five segments.



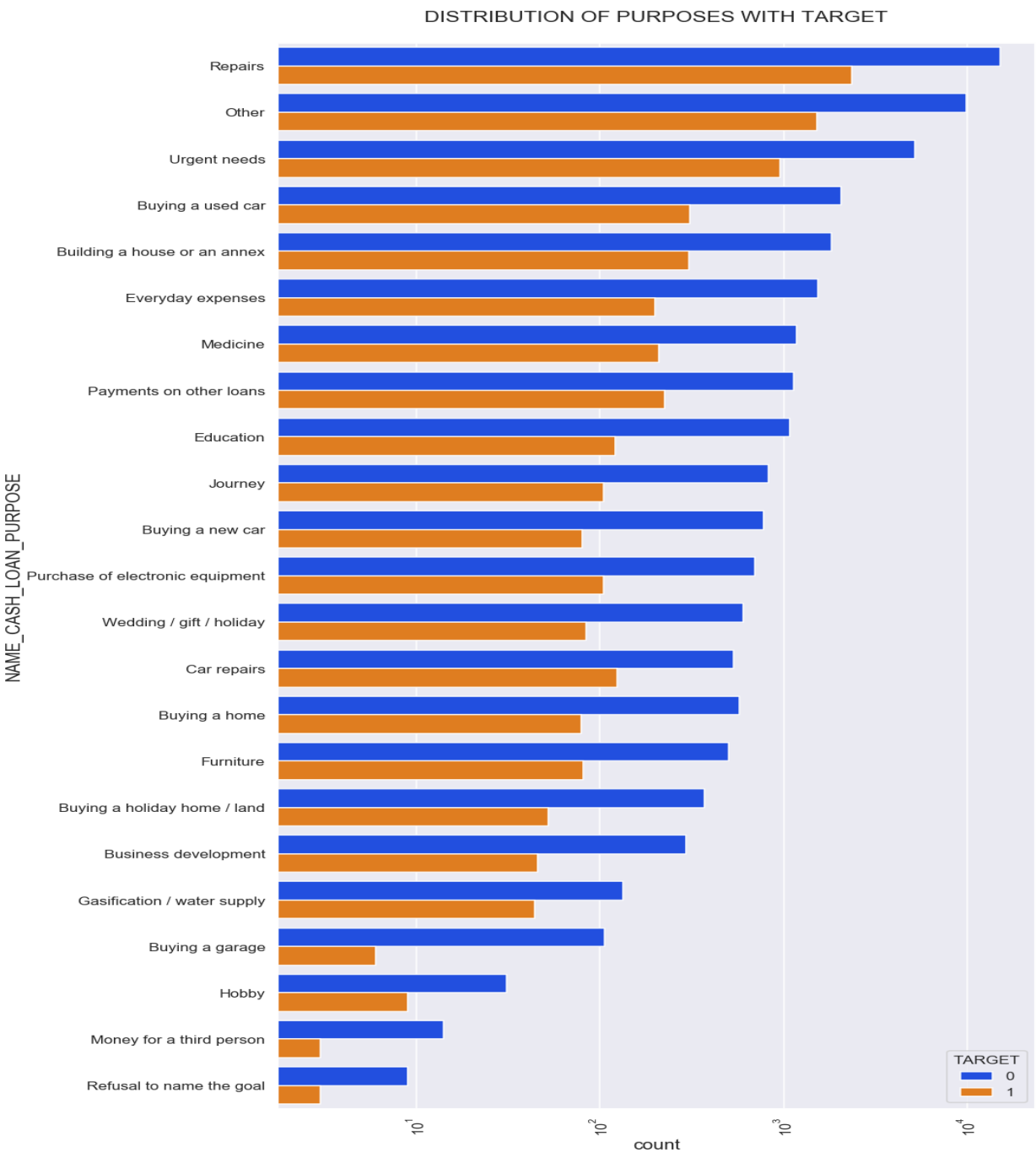
This is the boxplot for income amount and education status for target-1. There are many similarities between the target variables in this case from above boxplot for Education type 'Higher education' the income amount is mostly equal with all family status. Academic degree has the lowest number of outliers but the income is also on par with the other segments except for Lower secondary which has less income amount than others. This basically results that as the more educated people are there the more they tend to earn a good income which is the reason why lower secondary has not been faring well in this area.

-:UNIVARIATE ANALYSIS (After Merging the Datasets):-



- **Points to be concluded from above plot:**

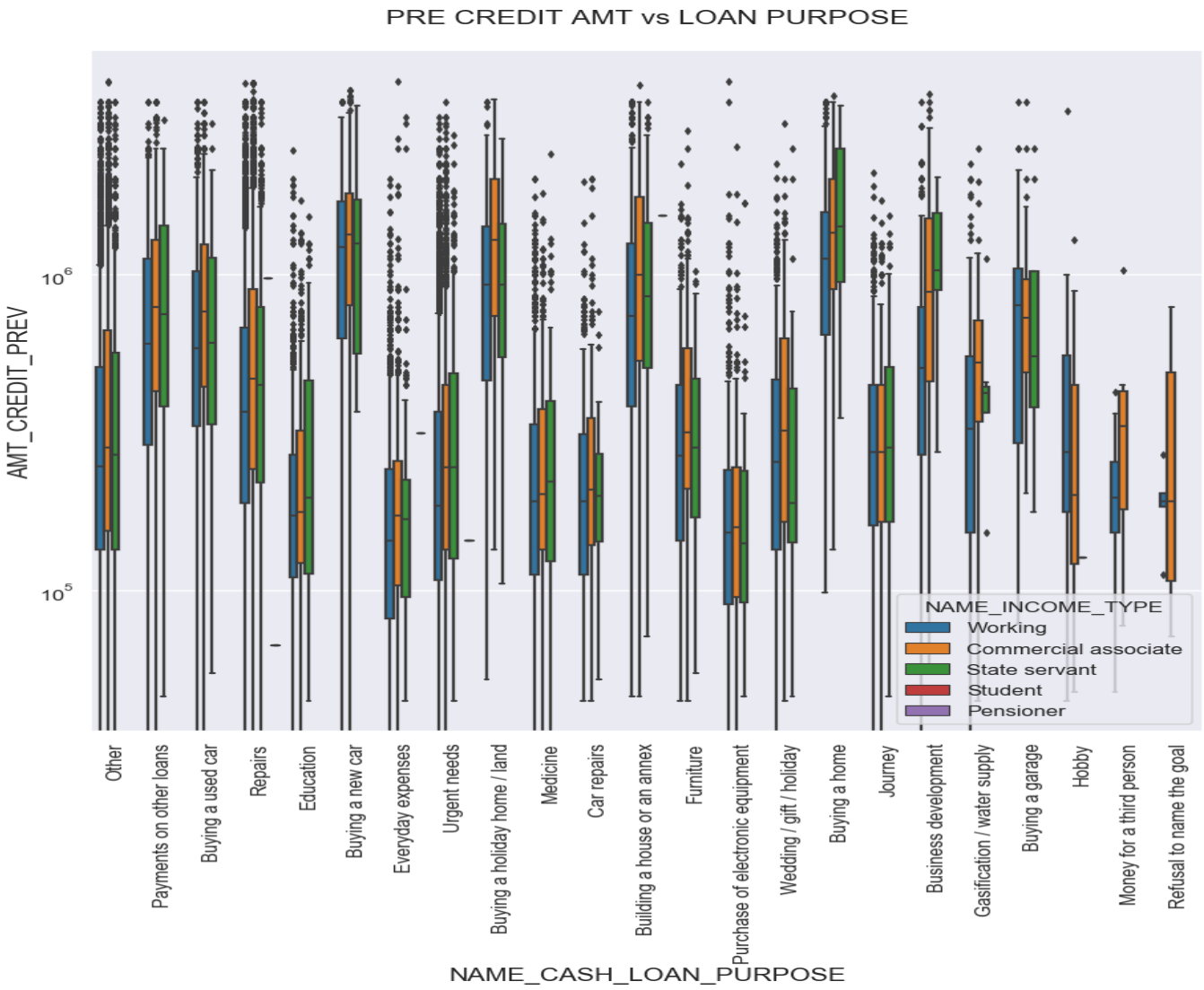
This is the plot for Distribution of contract status with purposes. Most refusal of loans have come from the segment 'repairs'. For education purposes we have equal number of approval and refused. Paying other loans, buying a home and buying a new car has significantly higher rejection compared to approvals. Segment 'Business Development' have more rejections compared to approvals. 'Everyday Expenses' and 'Purchase of Electronic Equipment' are the only two segments which have higher number of approvals than rejections, hence these segments need to be given special concern.



➢ Few points we can conclude from above plot:

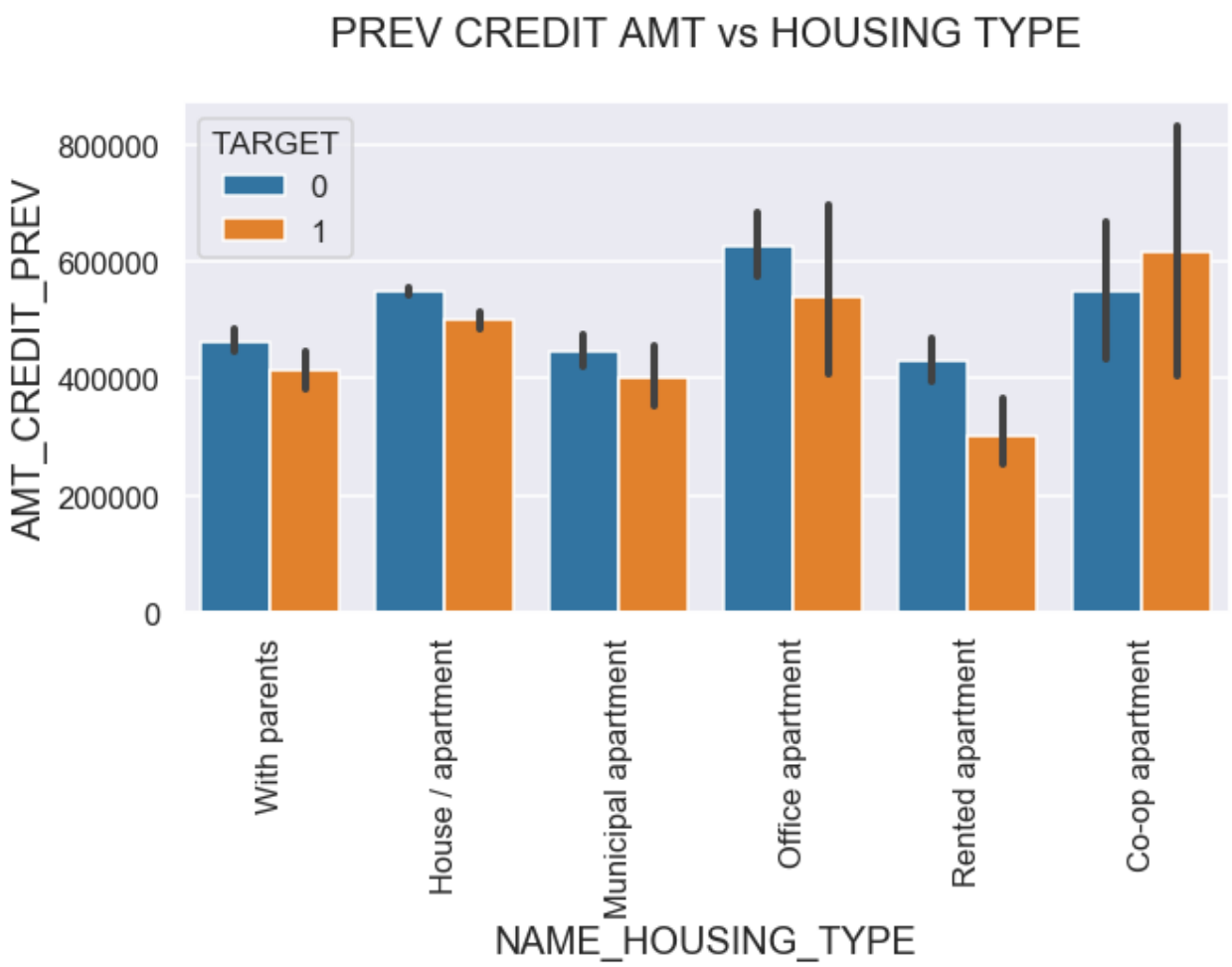
From the above graph we can clearly observe that Loan purposes with 'Repairs' are facing more difficulties in payment on time. There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car'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

-:BIVARIATE ANALYSIS (After merging the Datasets):-



- **From the above plot we can conclude some points:**

The above boxplot represents for Previous 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 high. The State Servant of the segment 'Buying a home' is the highest among all the segments. Income type of state servants have a significant amount of credit applied. The Commercial Associate of the segment 'Purchasing of electronic equipment' is the least compared to the other segments. Among all the segments the least would be the working income type of ‘everyday expenses’ loan purpose.



The above plot represents the Previous 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 making payment. Bank can focus mostly on housing type with parents or House/apartment or municipal apartment for making successful payments. As we can conclude from the above plot, 'Office Apartment' is higher compared to 'House/Apartment'. So, the focus of the bank should be preferring them over House/Apartment for providing loans as they will have minimal difficulties.

-:***CONCLUSION:-***

- From the case study we can conclude that the Banks should focus less on income type ‘Working’ as they are having most numbers of unsuccessful payments.
- Banks supposed to given more focus on contract type ‘Student’, ’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.
- Banks also have to given more focus on the people who are more educated as we have seen that people with lower secondary education have not been doing well in both income amount and credit amount.
- Everyday Expenses' and 'Purchase of Electronic Equipment' are the only 2 segments which have more numbers of approvals than rejections. Hence, these segments need to be given special consideration.
- Getting as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.
- Also, with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.