Credit EDA Case Study

Submitted by: Abhinav Sharma, Cassia Rodrigues

Business Understanding

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.

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:

- 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 data provided for the study contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1. **Approved:** The Company has approved loan Application
2. **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

Business Objectives

This case study aims 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 using EDA is the aim of this case study.

Data Understanding

Two data sets were provided as part of this case study

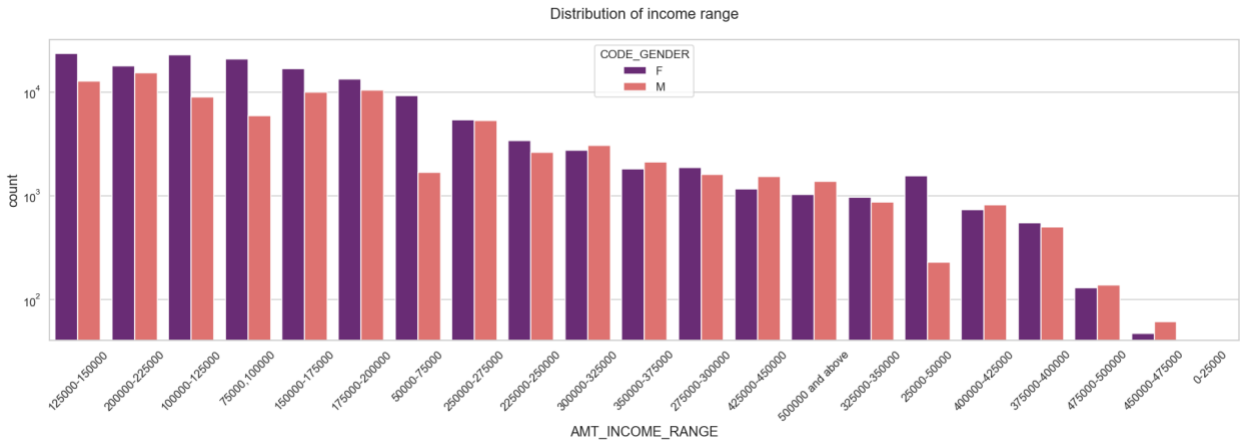
- Application Data
- Previous Application Data

Approach

1. Reading the Application Data Set
2. Identifying the missing data
3. Cleaning the data by deleting certain values and imputing some based on studying the fields
4. Understanding the Categorical Columns
5. Performing the Univariate Analysis for the categories
6. Plotting graphs to understand the correlation between the target variables
7. Performing the Univariate Analysis for the variables
8. Plotting distributions to understand the outliers
9. Performing the Bivariate Analysis for the numerical variables
10. Plotting graphs to understand the correlation between the variables
11. Reading the Previous Application Data Set
12. Identifying & Cleaning the missing data
13. Merging the two data sets for further analysis
14. Performing Univariate & Bivariate Analysis on the combined larger data set
15. Plotting graphs to understand the correlation between categories and variable on the larger data set

Inference - Categorical Univariate analysis for Target 0

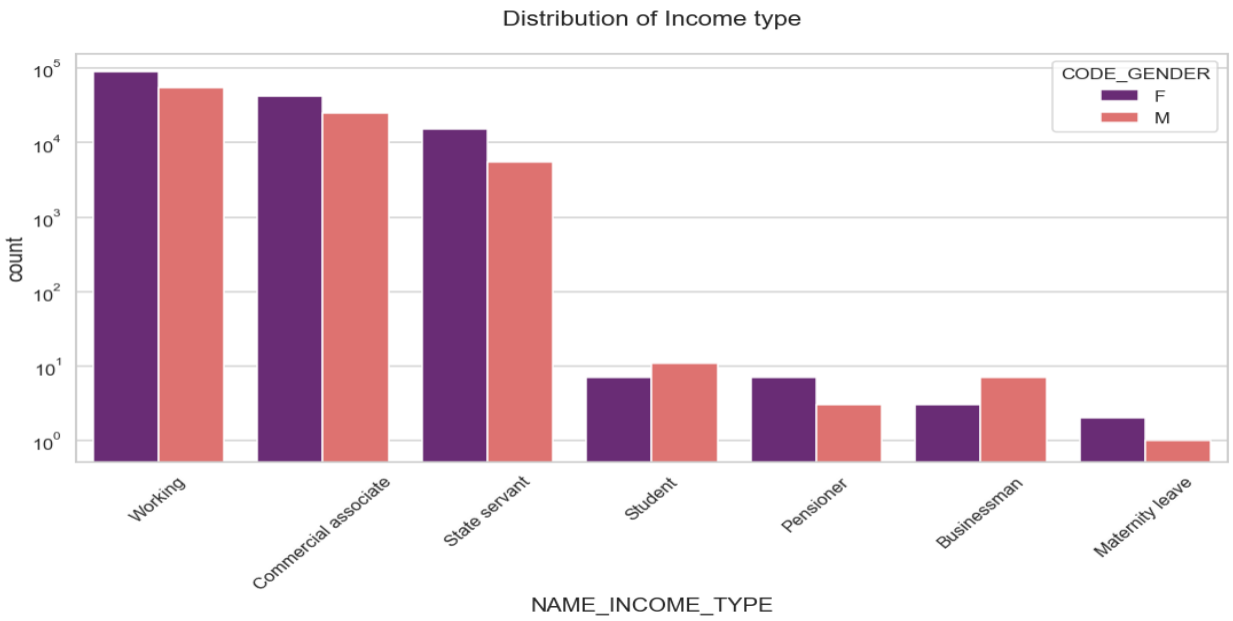
Application Data Set - Distribution of Income Range



Points to be concluded from the graph on the right side.

- Female counts are higher than male.
- Income range from 100000 to 200000 is having more number of credits.
- This graph show that females are more than male in having credits for that range.
- Very less count for income range 400000 and above.

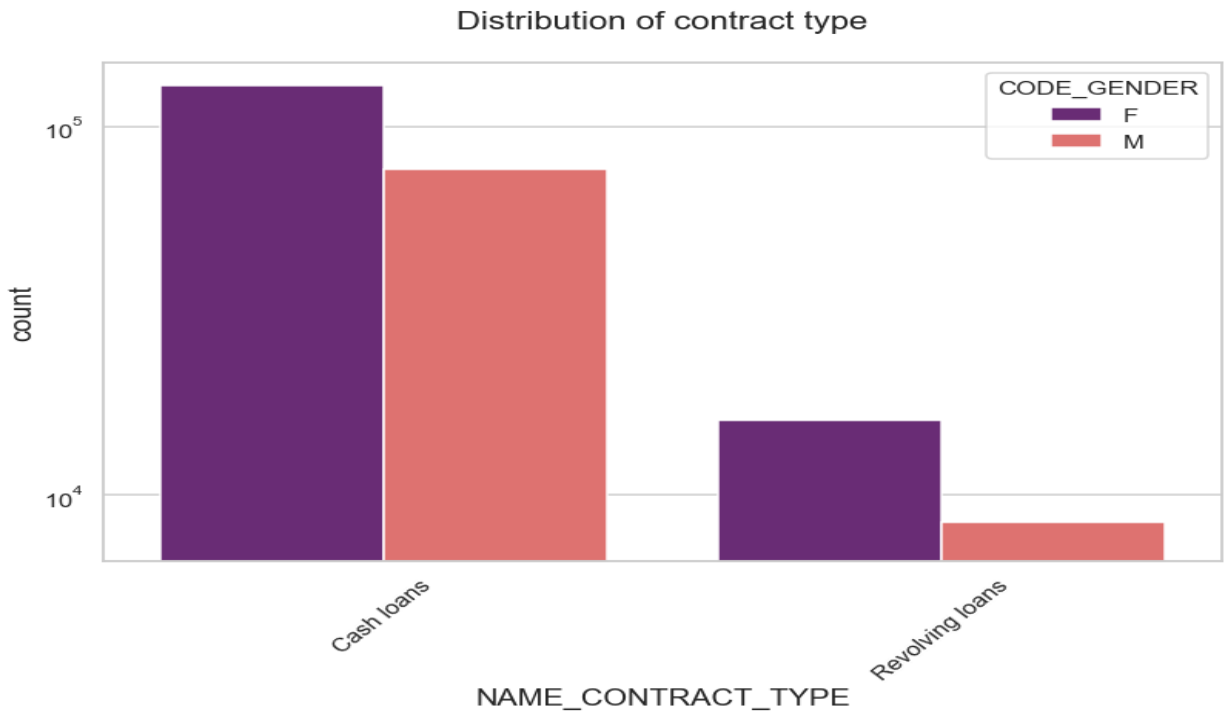
Application Data Set - Distribution of Income Type



Points to be concluded from the graph on the right.

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

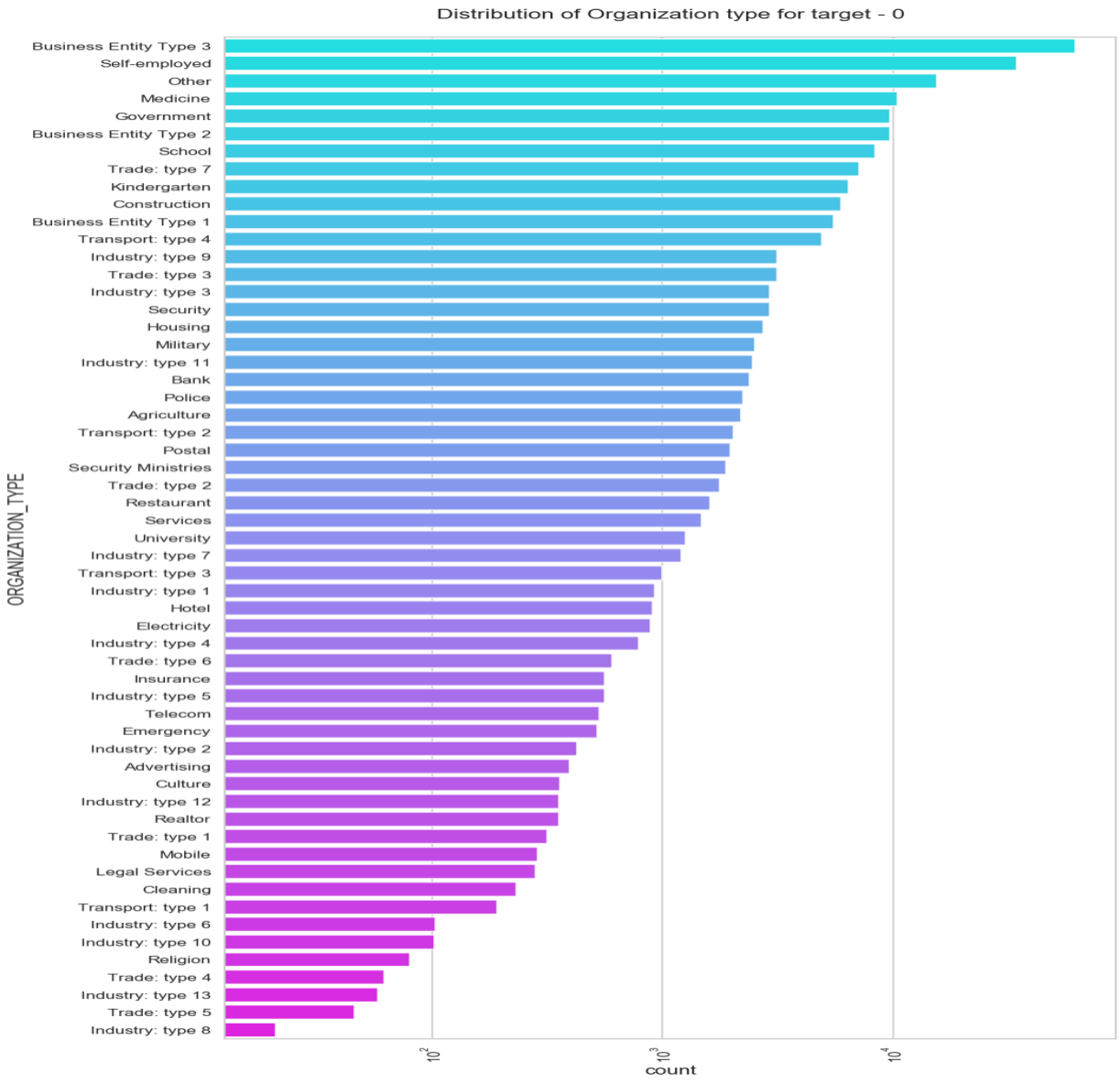
Application Data Set - Distribution of Contract Type



Points to be concluded from the graph on the right.

- For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.
- For this also Female is leading for applying credits.

Application Data Set - Distribution of Organization Type

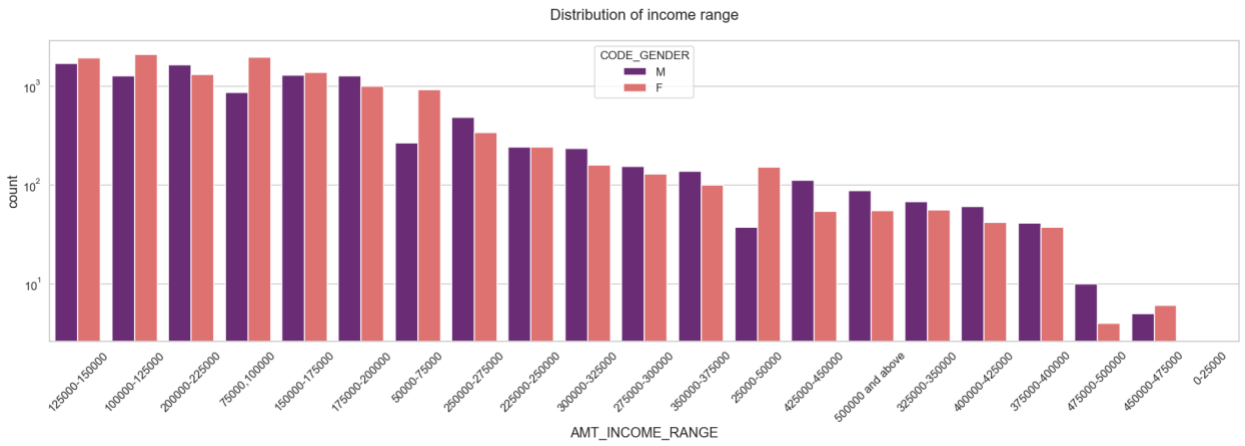


Points to be concluded from the graph on the right.

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

Inference - Categorical Univariate analysis for target 1

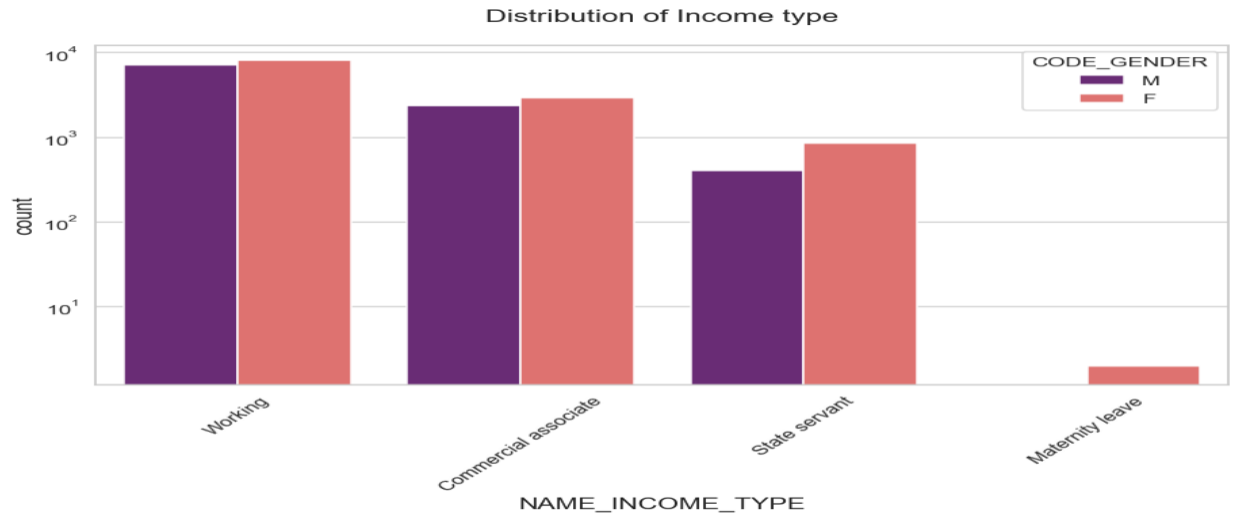
Application Data Set - Distribution of Income Range



Points to be concluded from the graph on the right side.

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

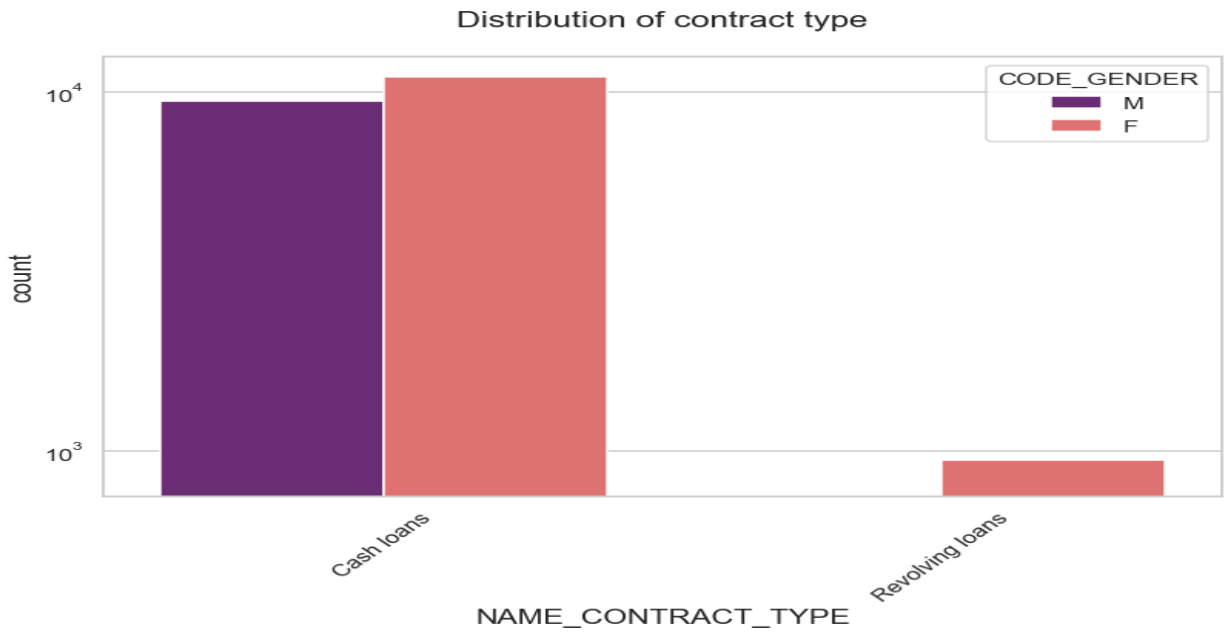
Application Data Set - Distribution of Income Type



Points to be concluded from the graph on the right side.

- 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 number of credits than male.
- Less number of credits for income type ‘Maternity leave’.
- For type 1: There is no income type for ‘student’ , ’pensioner’ and ‘Businessman’ which means they don’t do any late payments.

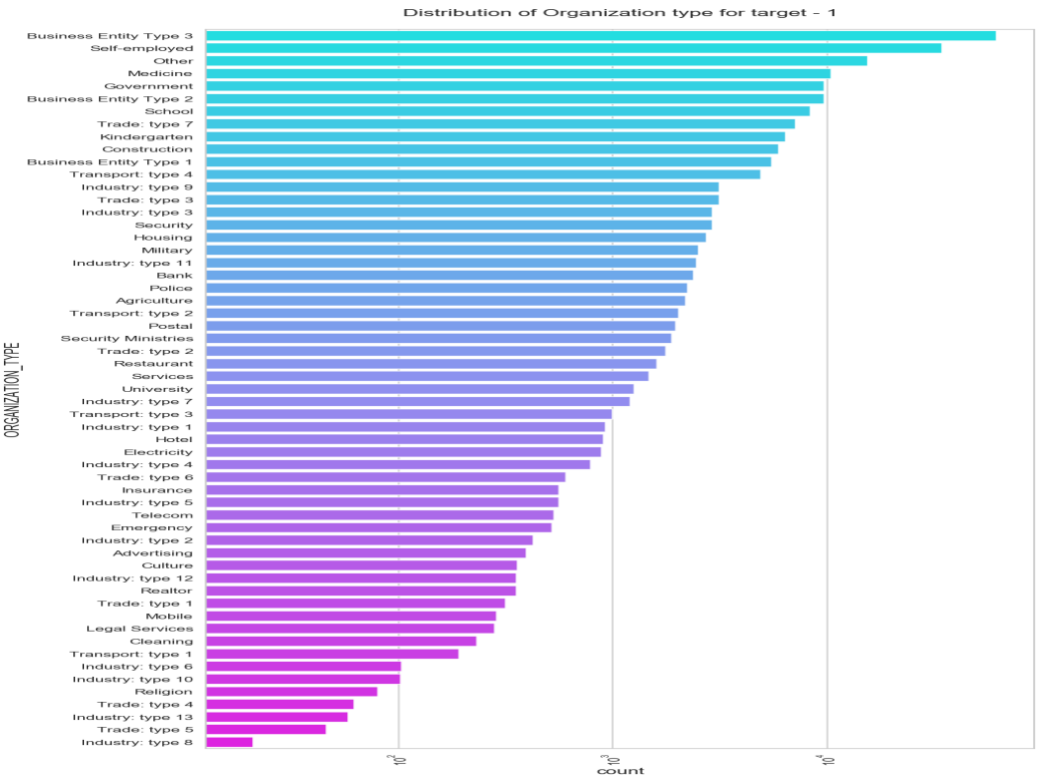
Application Data Set - Distribution of Contract Type



Points to be concluded from the graph on the right.

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

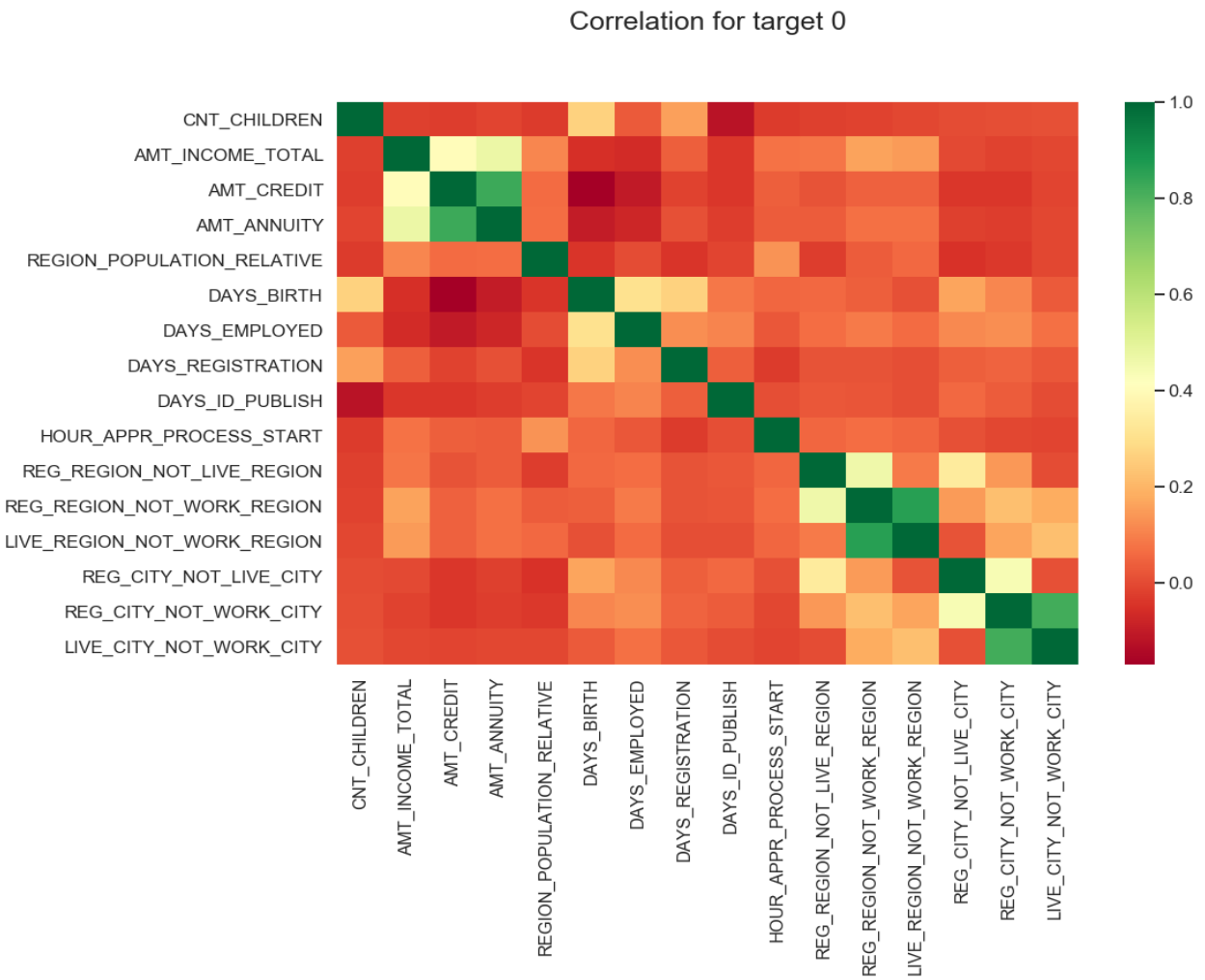
Application Data Set - Distribution of Organization Type



Points to be concluded from the graph on the right.

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

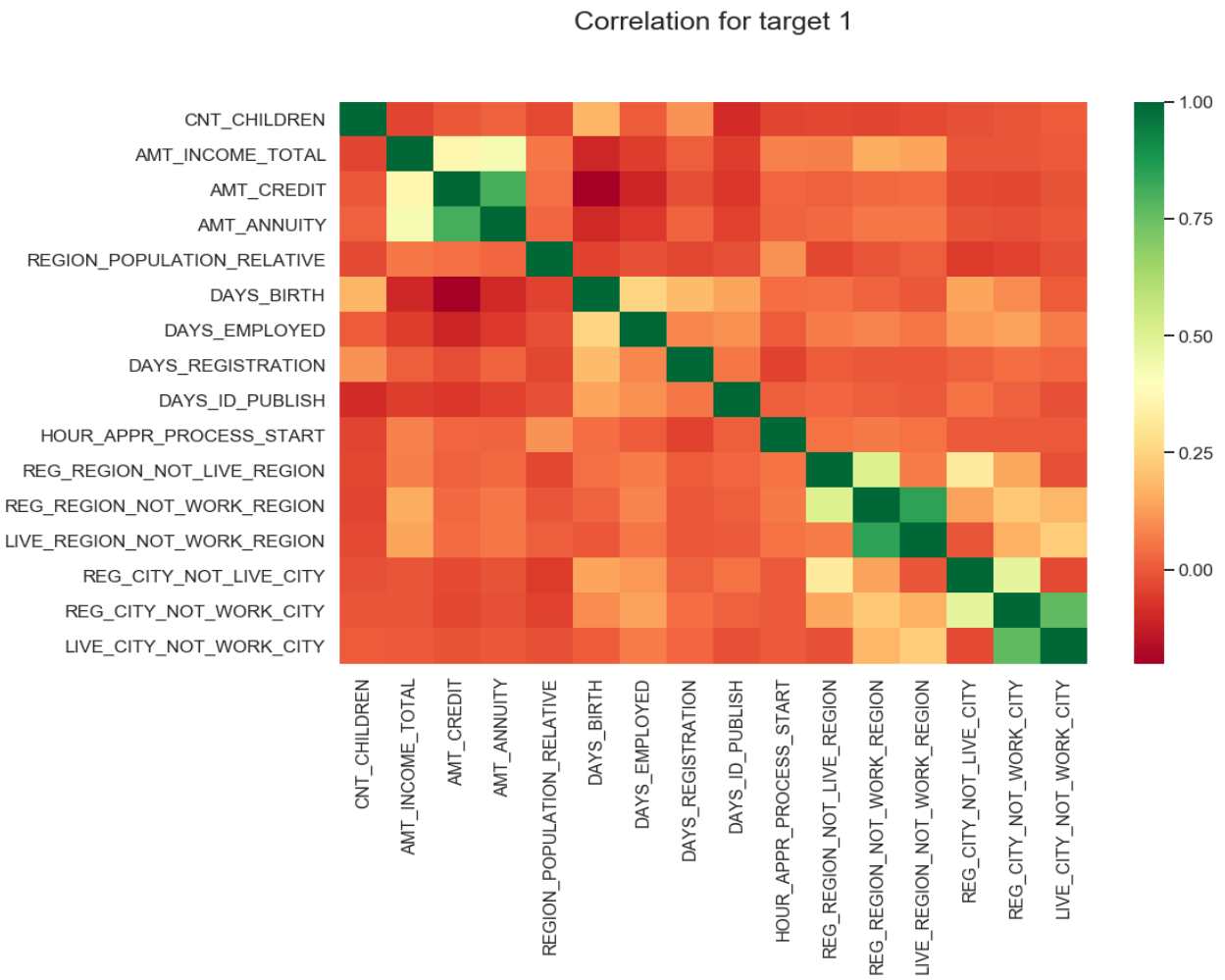
Inference - Correlation of Target 0



Points to be concluded from the graph presented before.

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

Inference - Correlation of Target 1

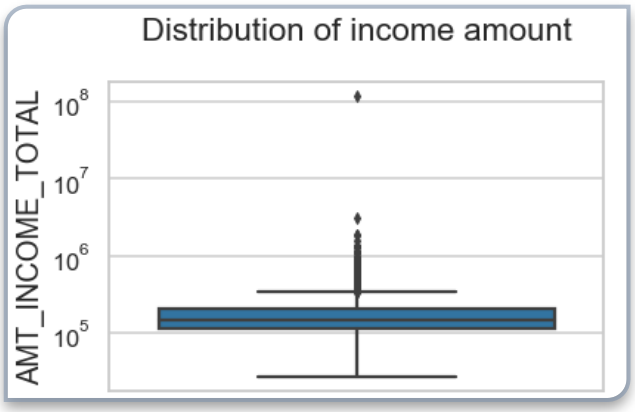


This heat map for Target 1 is also having quite a same observation just like Target 0. But for few points are different. They are listed below.

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

Inference - Categorical Univariate Analysis for Variable Target 0

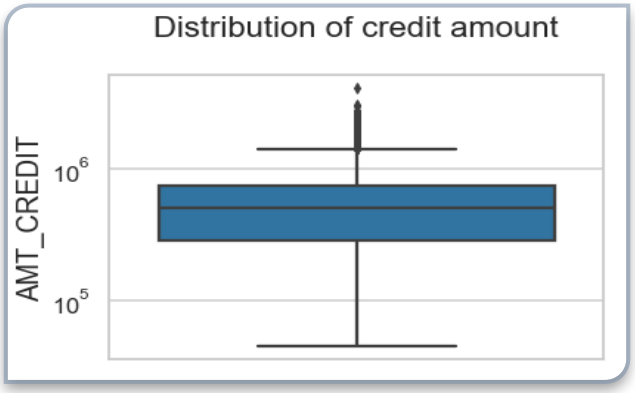
Application Data Set - Boxplot for Income Amount



Few points can be concluded from the graph

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

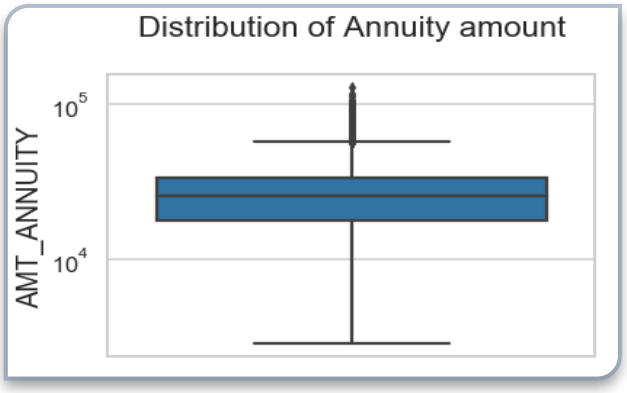
Application Data Set - Boxplot for Credit Amount



Few points can be concluded from the graph

- Some outliers are noticed in credit amount.
- 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.

Application Data Set - Boxplot for Annuity Amount

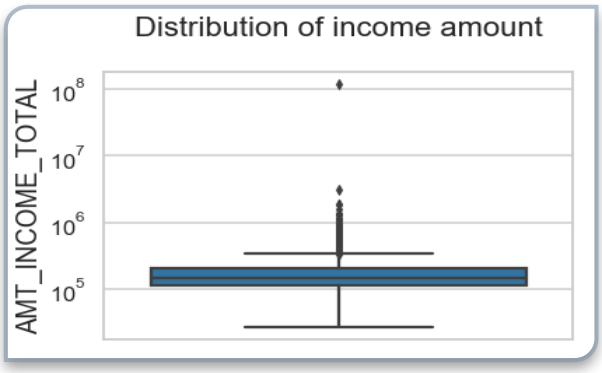


Few points can be concluded from the graph

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

Inference - Categorical Univariate Analysis for Variable Target 1

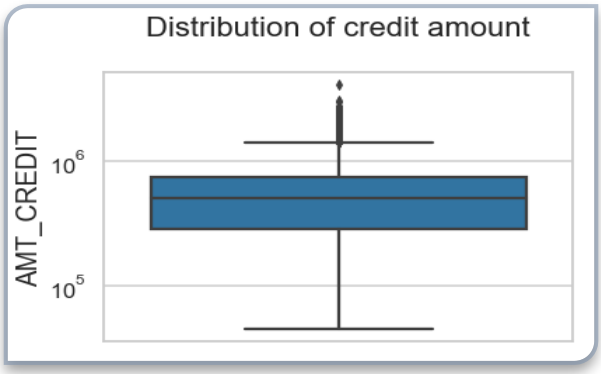
Application Data Set - Boxplot for Income Amount



Few points can be concluded from the graph

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

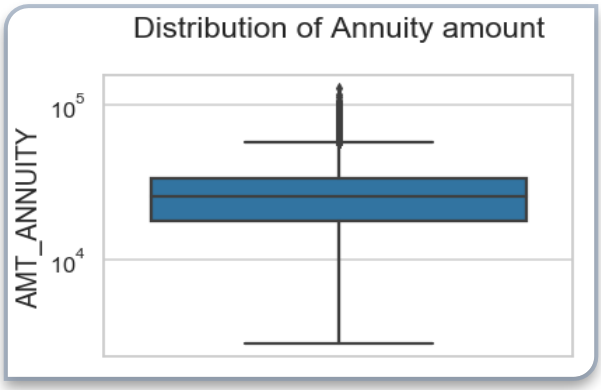
Application Data Set - Boxplot for Credit Amount



Few points can be concluded from the graph

- Some outliers are noticed in credit amount.
- 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.

Application Data Set - Boxplot for Annuity Amount

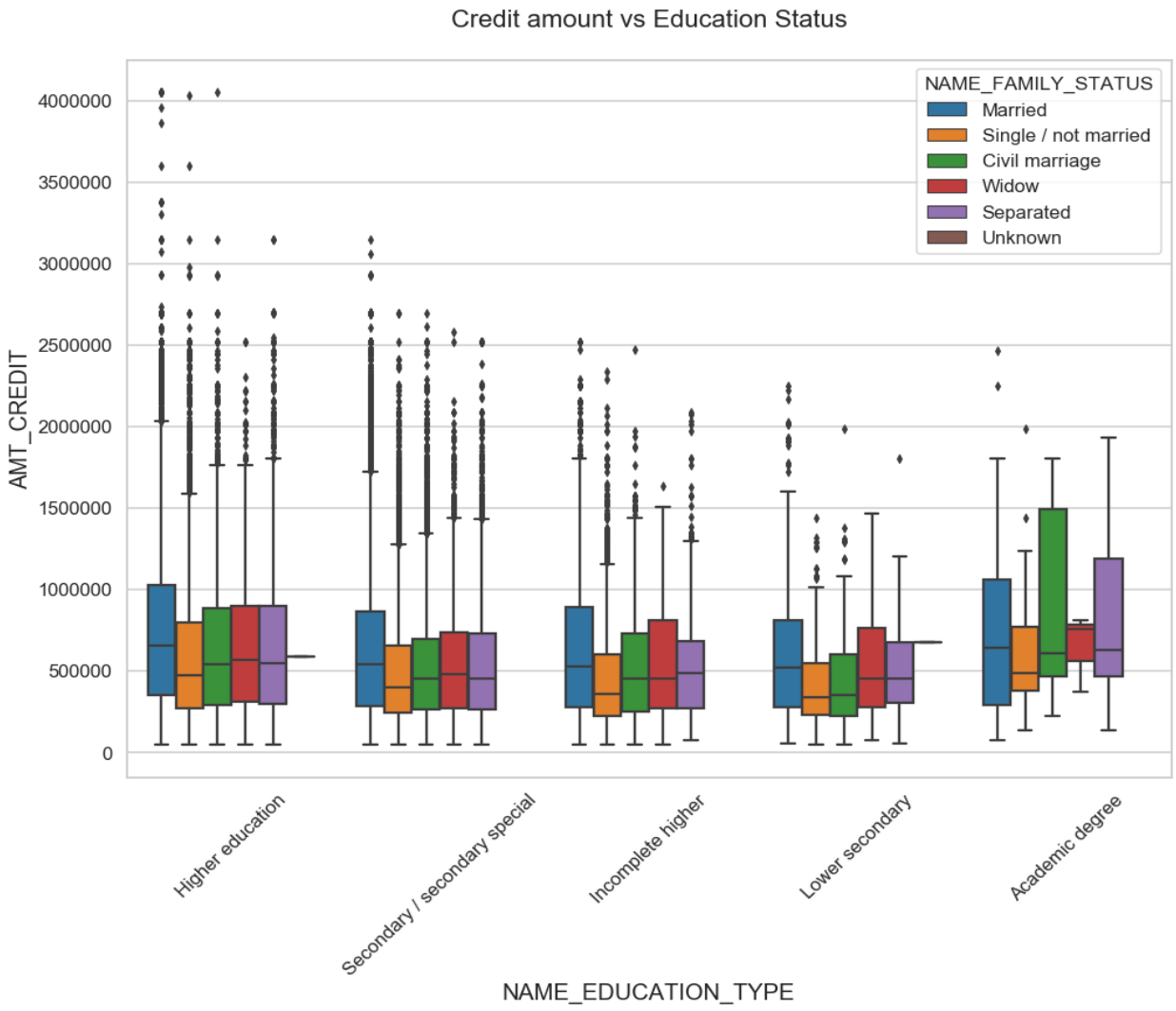


Few points can be concluded from the graph

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

Inference - Bivariate Analysis for Target 0

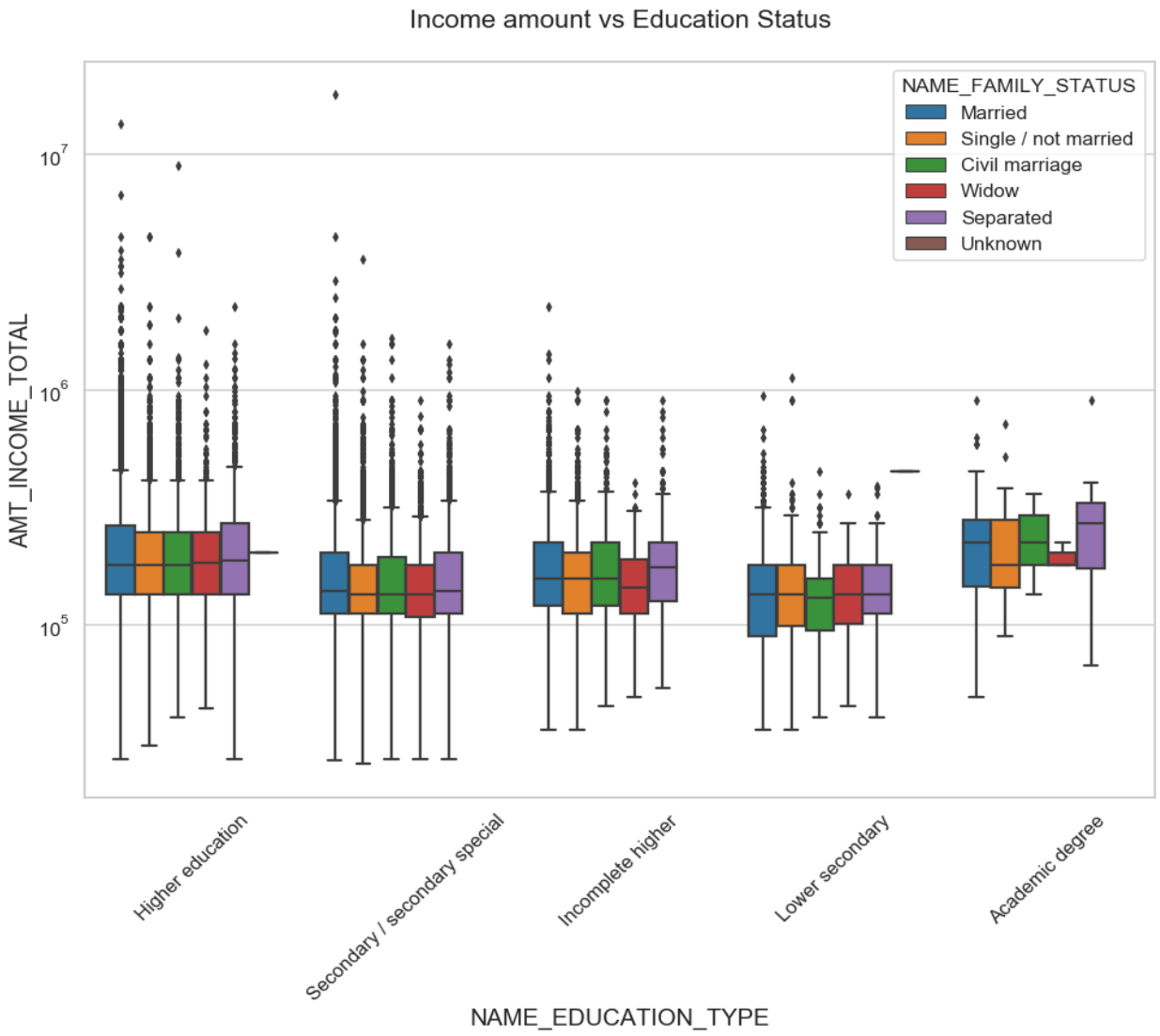
Application Data Set - Credit Amount vs Education Status



Few points can be concluded from the graph.

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

Application Data Set - Income Amount vs Education Status

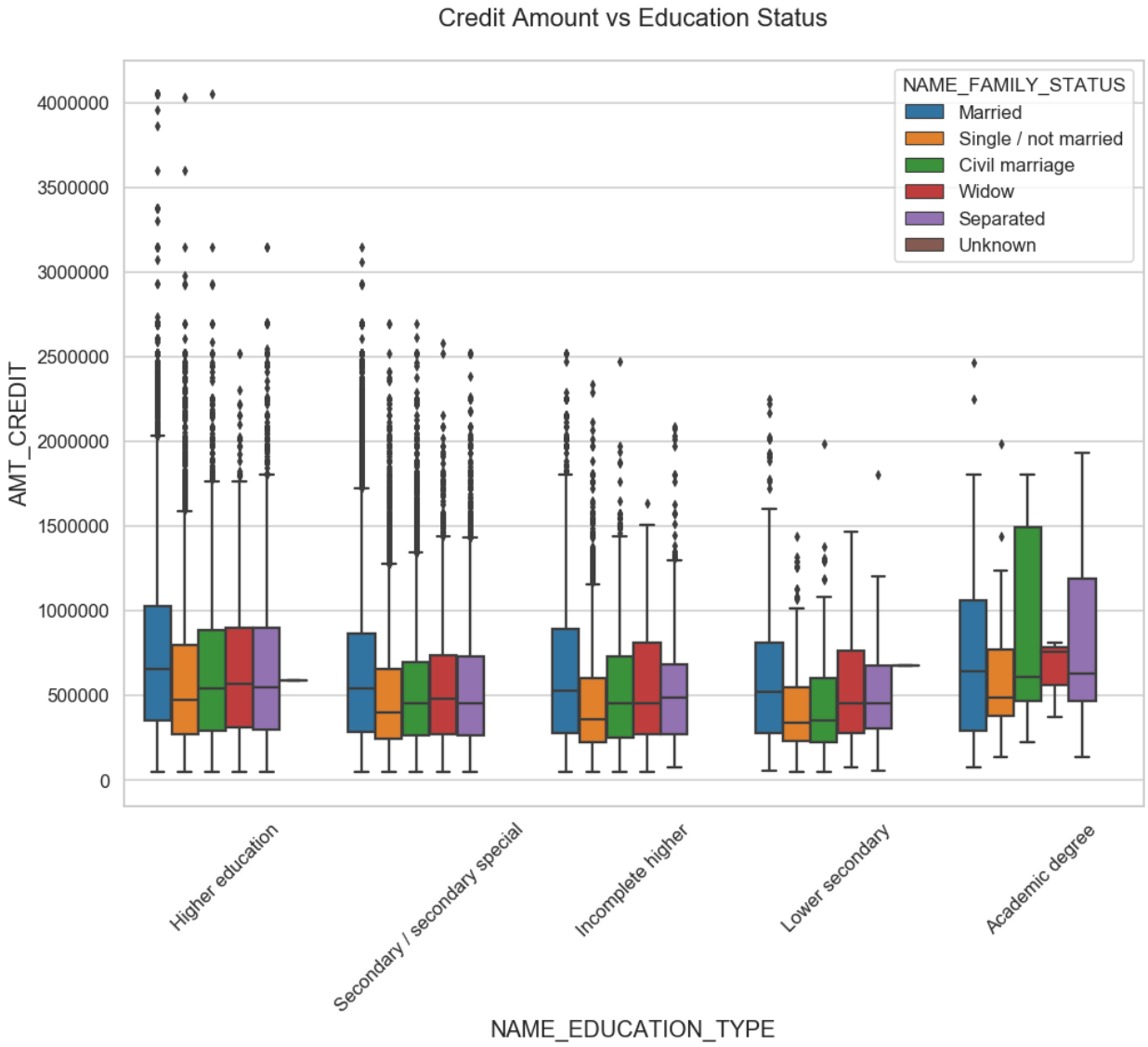


Few points can be concluded from the graph.

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

Inference - Bivariate Analysis for Target 1

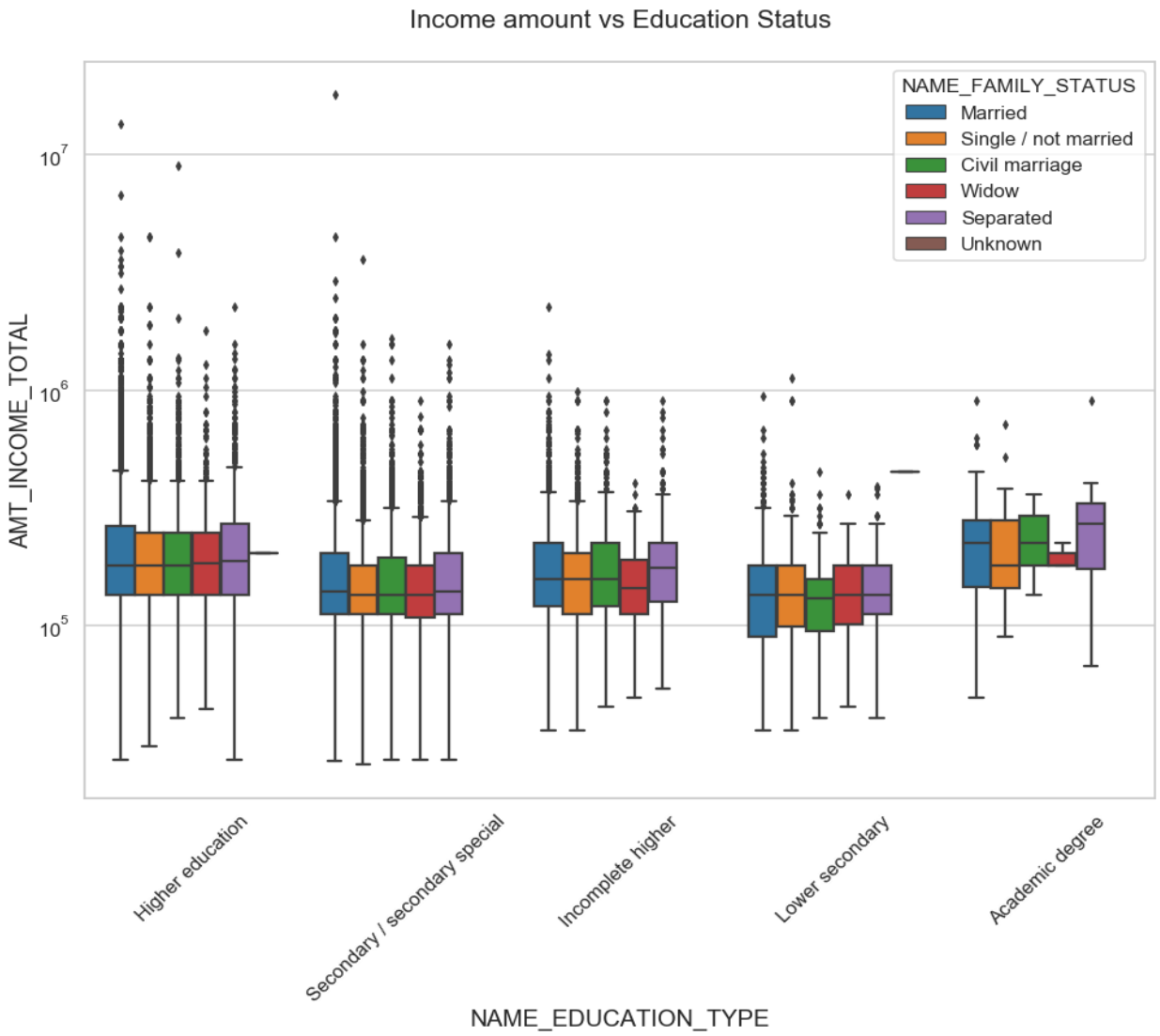
Application Data Set - Credit Amount vs Education Status



Few points can be concluded from the graph.

- Quite similar from Target 0, we can say that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree 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.

Application Data Set - Income Amount vs Education Status

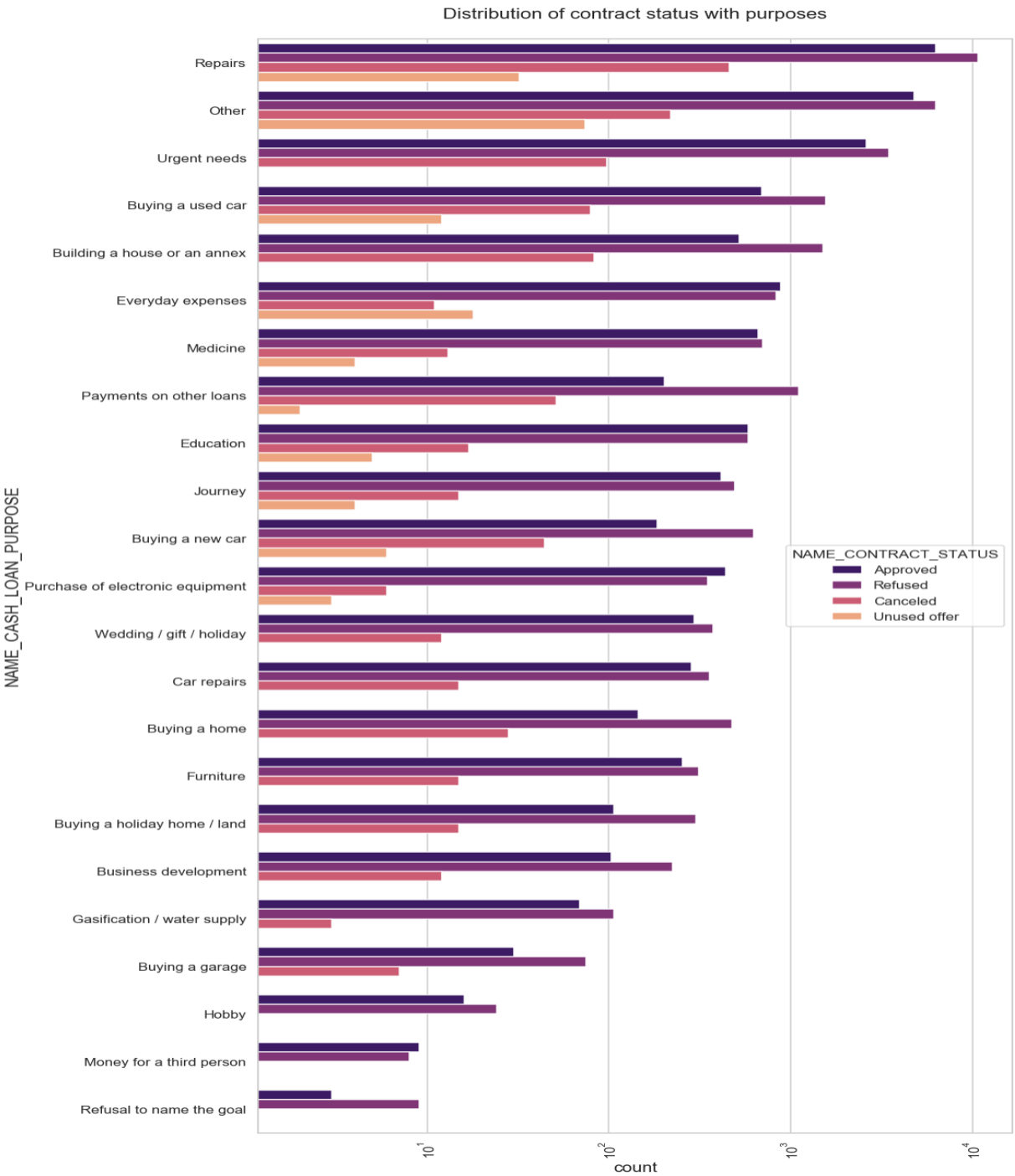


Few points can be concluded from the graph

- Have some similarity with Target0, From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status.
- Less outlier are having for Academic degree but there income amount is little higher that Higher education.
- Lower secondary are have less income amount than others.

Inference - Univariate Analysis after merging previous data

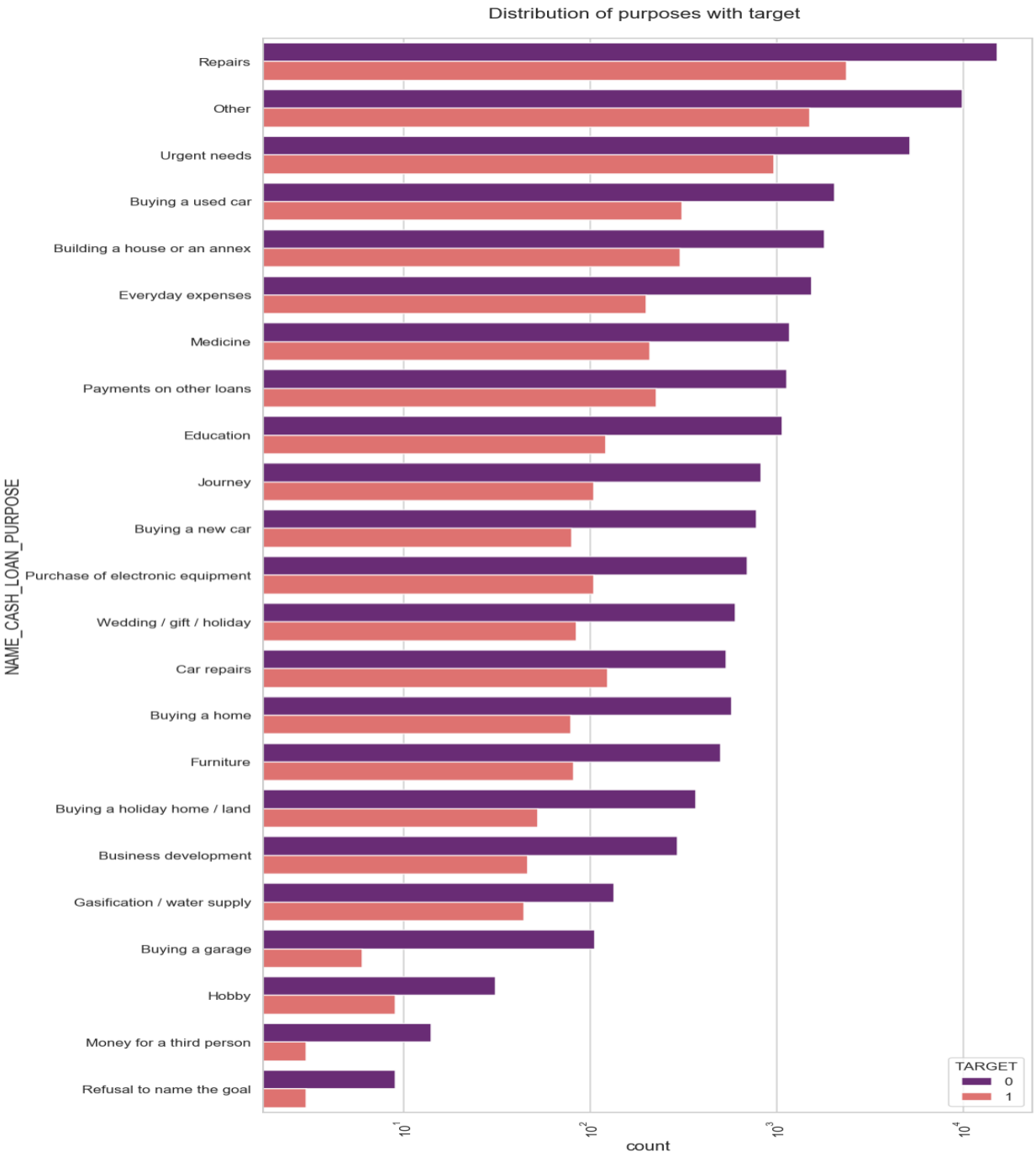
Merged Data Set - Distribution of Contract Status with purposes



Few points can be concluded from the graph

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

Merged Data Set - Distribution of purposes with target

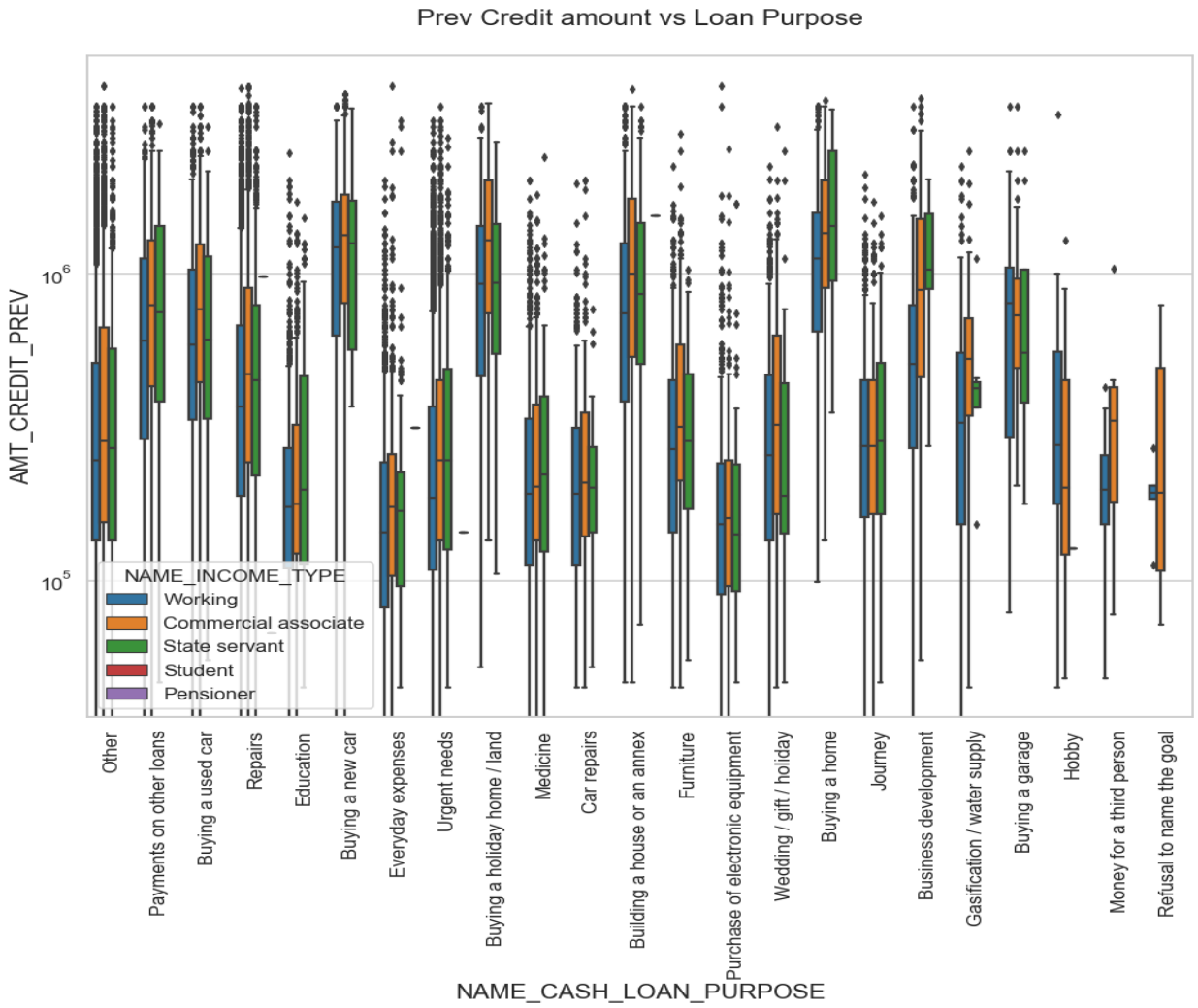


Few points can be concluded from the graph

- 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.

Inference - Performing Bivariate Analysis

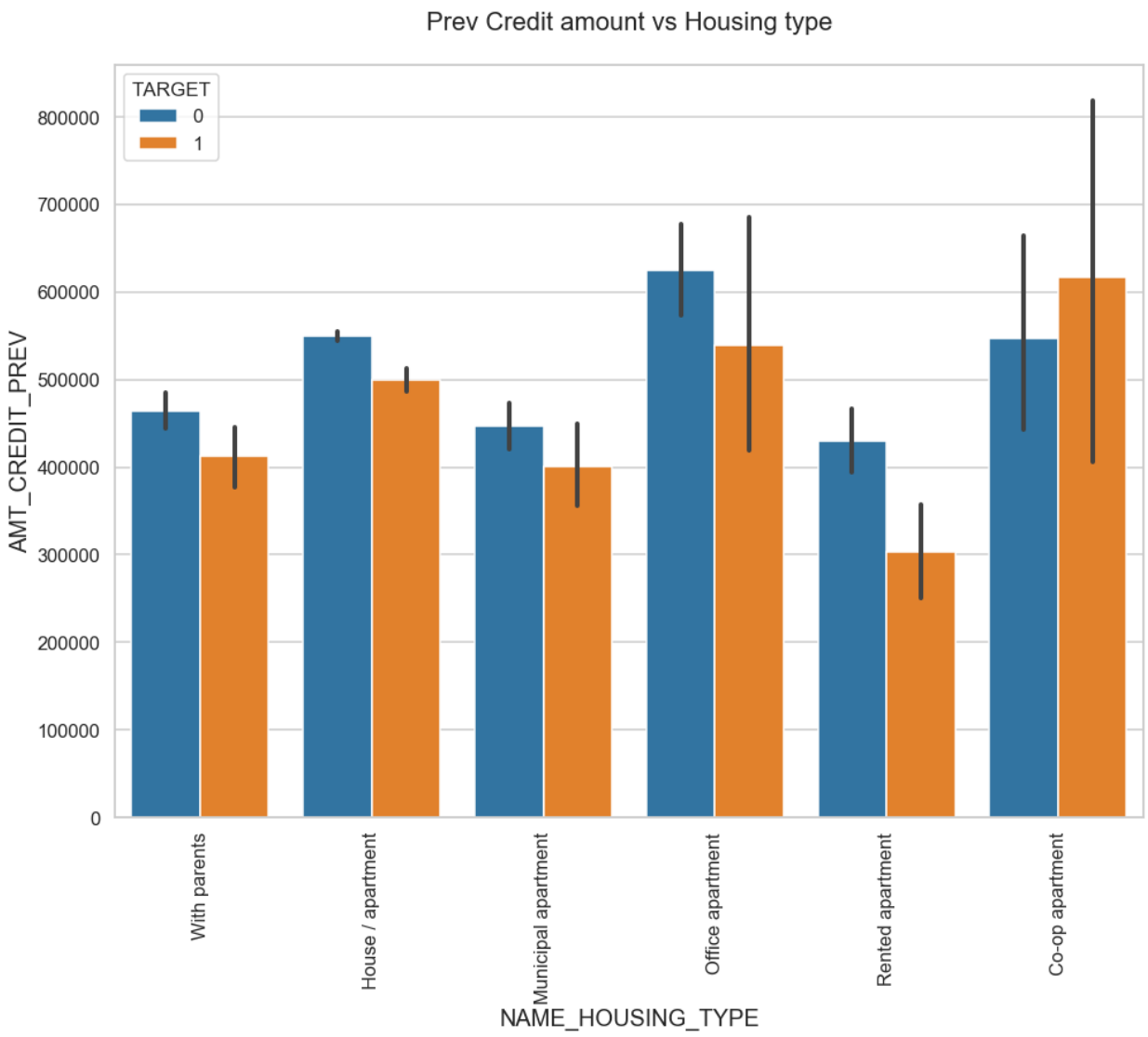
Merged Data Set - Previous Credit Amount vs Loan Purpose



From the previous graph we can conclude the below points:

- 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.

Merged Data Set - Previous Credit Amount vs Housing Type



Few points can be concluded from the graph.

- 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

- Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.
- Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.
- Also with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.
- Get as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.