**CREDIT EDA CASE STUDY**

**Submitted by Arun Kumar and Anirudh Mneon**

**Introduction to the problem**

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. 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:

* 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 given below 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.

**The Study and Observations of the Data**

The two data sets received as part of the case study

* Application Data
* Previous Application Data

We have performed Data quality checks on these two data, utilizing the Application data first

**Missing values:**

We have observed that there are a couple of missing values in the Application dataset.

We have dropped columns where the missing values are greater than 19%

We have observed that there are a couple of missing values in the Previous Application Dataset

We have dropped columns where the missing values are greater than 20%

**We have imputed the below columns with mean and mode:**

NAME\_TYPE\_SUITE(Object)

AMT\_GOODS\_PRICE(Integer)

**We have imputed 0 for these columns where more than 90% values are missing:**

AMT\_REQ\_CREDIT\_BUREAU\_HOUR

AMT\_REQ\_CREDIT\_BUREAU\_DAY

AMT\_REQ\_CREDIT\_BUREAU\_WEEK

AMT\_REQ\_CREDIT\_BUREAU\_MON

**Checking Data types of columns**

Filtering between Float and Object columns

We have converted few variables that were not Float variables to Integer

A lot of the int columns are named flag. We can check their unique values

*Flag variables with 0 and 1 values should be converted to Categorical variables*

**Checking for Outliers in the Data using 3 variales**

We can see from the results below that some columns have a significant min or max value. We will probe these

1. AMT\_INCOME\_TOTAL- shows 5 outliers with 4 outliers tween 0-0.2 and one eing at 1.2
2. AMT\_ANNUITY-
3. DAYS\_EMPLOYED-

After the treatment of Outliers:

*Excluding values outside 99%ile in each of the 3 variables*

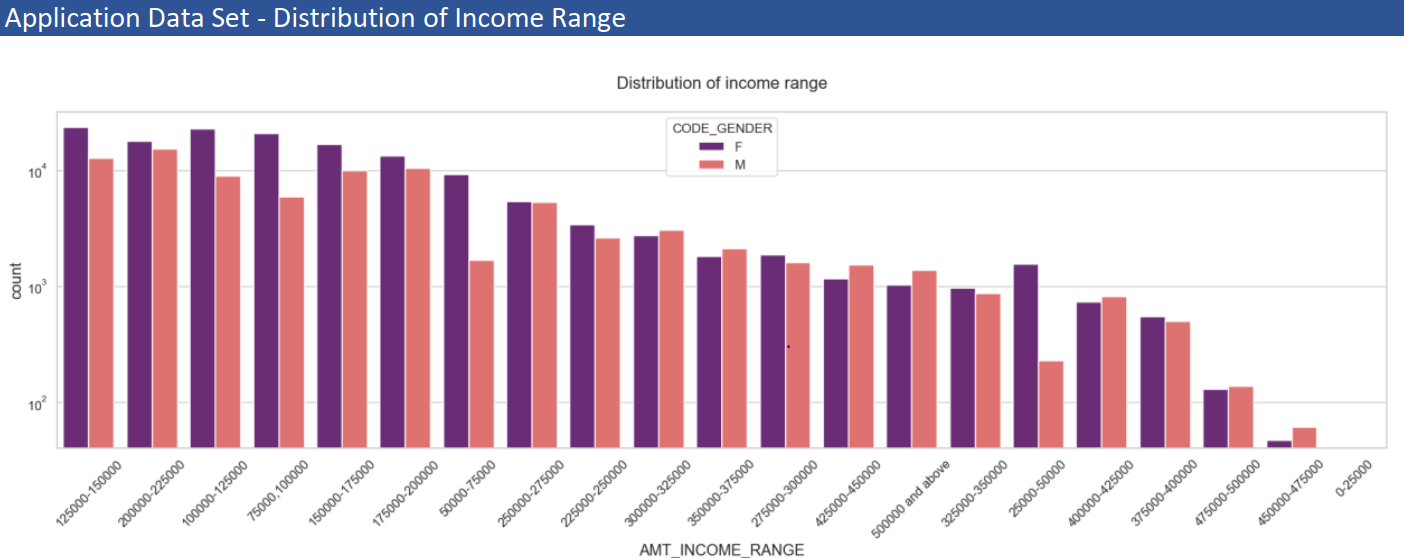
1. AMT\_ANNUITY
2. DAYS\_EMPLOYED
3. AMT\_INCOME\_TOTAL

**Univariate and Bivariate Analysis**

Performing the Univariate Analysis for the variables

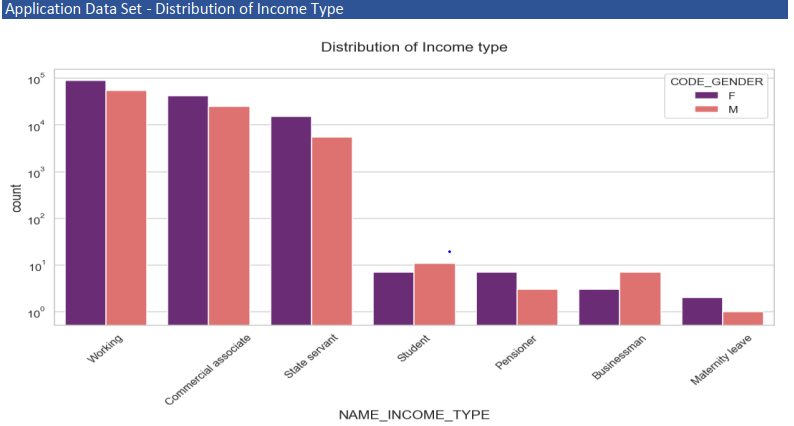
Plotting distributions to understand the outliers  
Performing the Bivariate Analysis for the numerical variables  
Plotting graphs to understand the correlation between the variables  
 Reading the Previous Application Data Set  
 Identifying & Cleaning the missing data  
 Merging the two data sets for further analysis  
 Performing Univariate & Bivariate Analysis on the combined larger data set  
 Plotting graphs to understand the correlation between categories and variable on the larger data set

Inference - Categorical Univariate analysis for Target 0

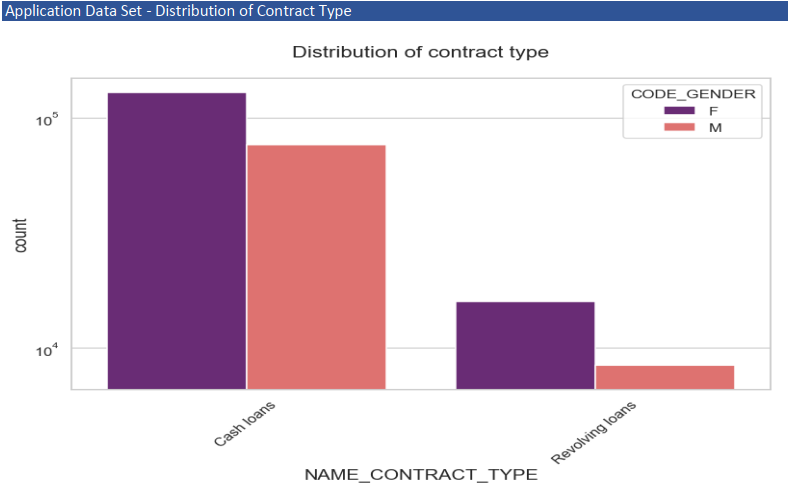


Important points to note include:

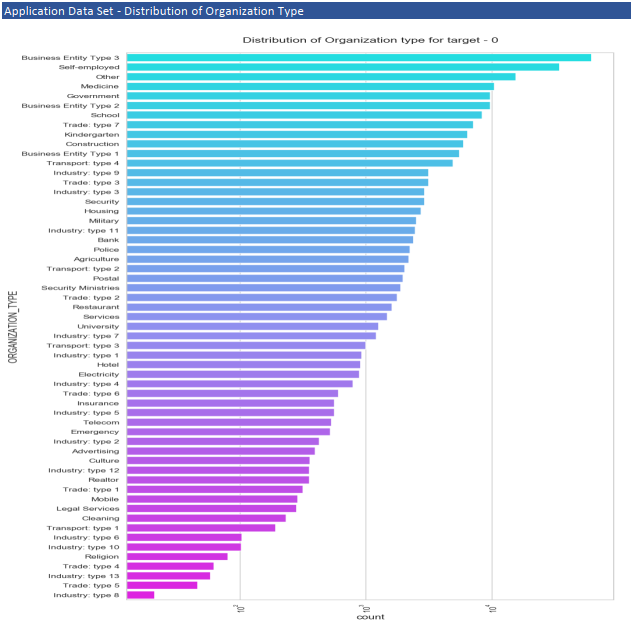
* Females seem to have a consistent higher count than the males in the graph, and are shown to have higher credits than males overall
* Income range from 100000 to 200000 seems to have a higher number of credits



Important points to conclude are:  
• For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher  
than others.  
• For this Females seem to have more number of credits than male.  
• Less number of credits for income type ‘student’ ,’pensioner’, ‘Businessman’ and ‘Maternity leave’.

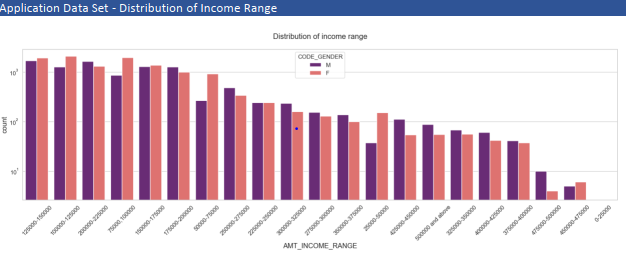


Important point to be concluded from the graph on the right.  
• For contract type ‘cash loans’ has a higher number of credits than ‘Revolving loans’ contract type.  
• Females are faring better for applying credits



Important pointers to be concluded from the graph on the right.  
• Clients who have applied for credits are mostly from the organization type ‘Business entity Type 3’ , ‘Self  
employed’ , ‘Other’ , ‘Medicine’ and ‘Government’.  
• Least of the clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4

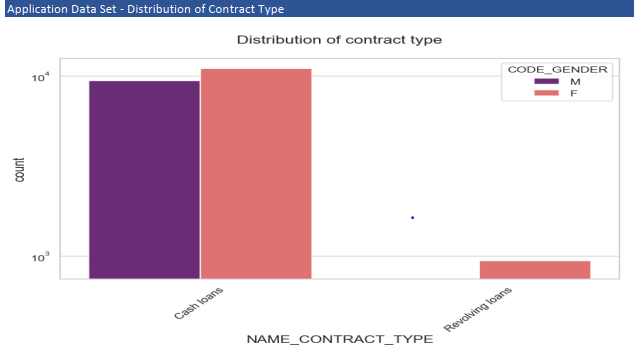
Inference - Categorical Univariate analysis for target 1



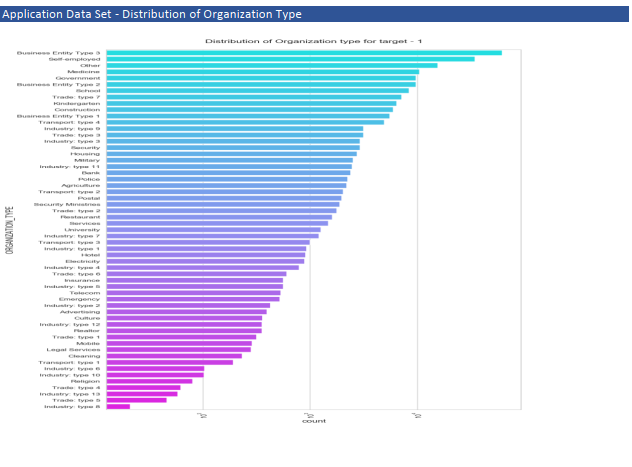
Important points to be concluded from the graph on the right side.  
• Male counts are higher than female, males are leading in the credit realm unlike the previous set.  
• Income range from 100000 to 200000 is has more number of credits.  
• Very less count for income range 400000 and above



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

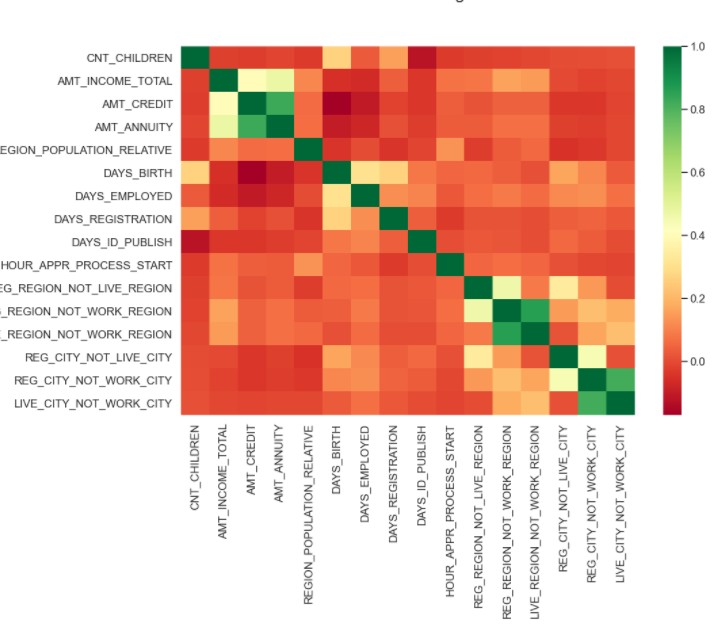


Points to be concluded from the graph on the right.  
• For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.  
• For this also Female is leading for applying credits.  
• For type 1 : there is only Female Revolving loans



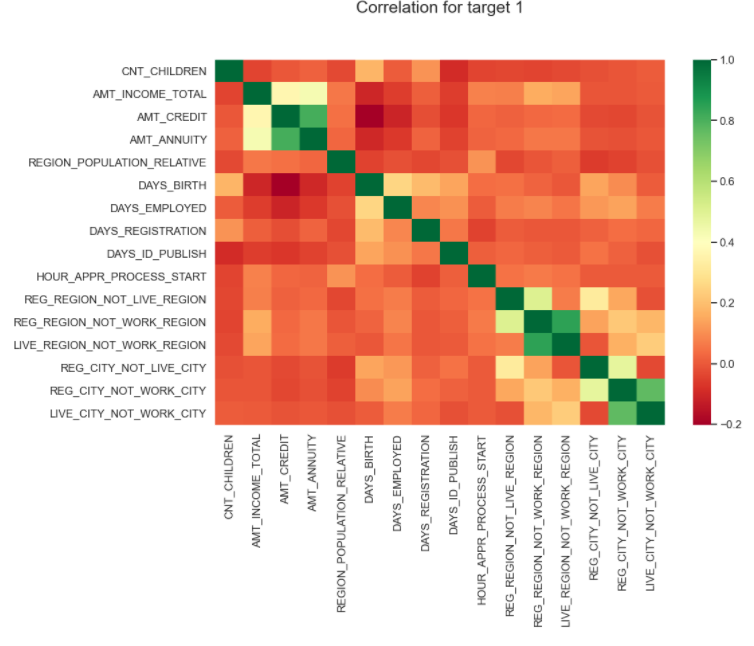
Points concluded:  
• Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self  
employed’ , ‘Other’ , ‘Medicine’ and ‘Government’.  
• Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.  
• Same as type 0 in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

Inference - Correlation of Target 0

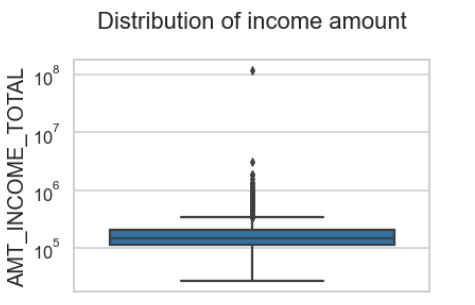


Points to be concluded from the graph presented before.  
• Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low  
age and vice-versa.  
• Credit amount is inversely proportional to the number of children client have, means Credit amount is  
higher 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

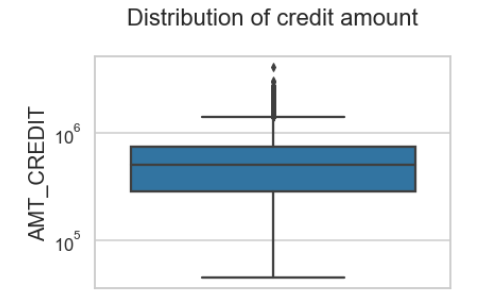
Inference - Correlation of Target 1



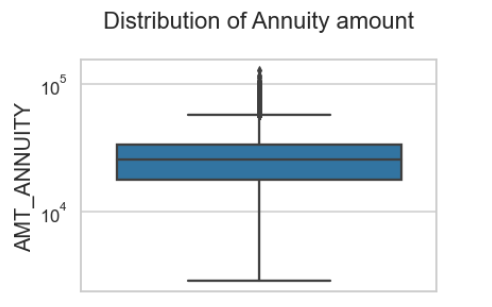
This heat map for Target 1 is also having quite a same observation just like Target 0. But for few points are different.  
They are listed below.  
• The client's permanent address does not match contact address are having less children and vice-versa  
• The client's permanent address does not match work address are having less children and vice-versa

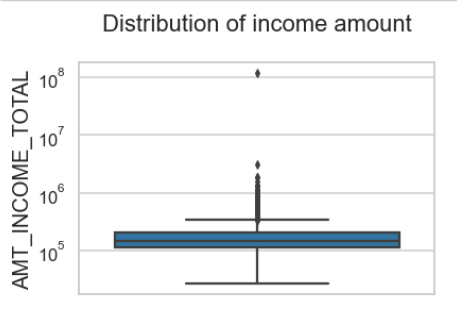
Inference - Categorical Univariate Analysis for Variable Target 0

Few points can be concluded from the graph  
• Some outliers are noticed in income  
amount.  
• The third quartiles is very slim for  
income amount

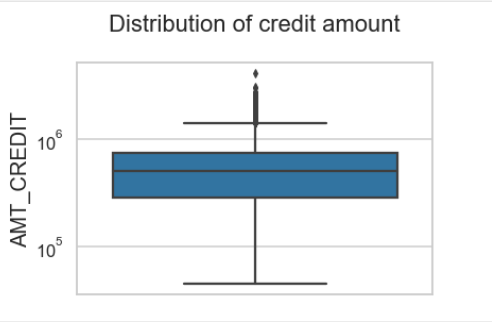


Few points can be concluded from the graph  
• Some outliers are noticed in credit  
amount.  
• The first quartile is bigger than third  
quartile for credit amount which means  
most of the credits of clients are present  
in the first quartile

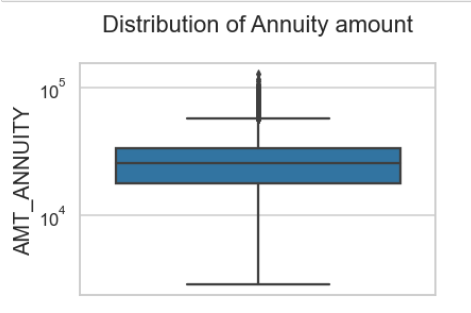
Few points can be concluded from the graph  
• Some outliers are noticed in annuity  
amount.  
• The first quartile is bigger than third  
quartile for annuity amount which  
means most of the annuity clients are  
from first quartile.

Inference - Categorical Univariate Analysis for Variable Target 1 

Few points can be concluded from the graph  
• Some outliers are noticed in income  
amount.  
• The third quartiles is slim for income  
amount.  
• Most of the clients of income are present  
in first quartile

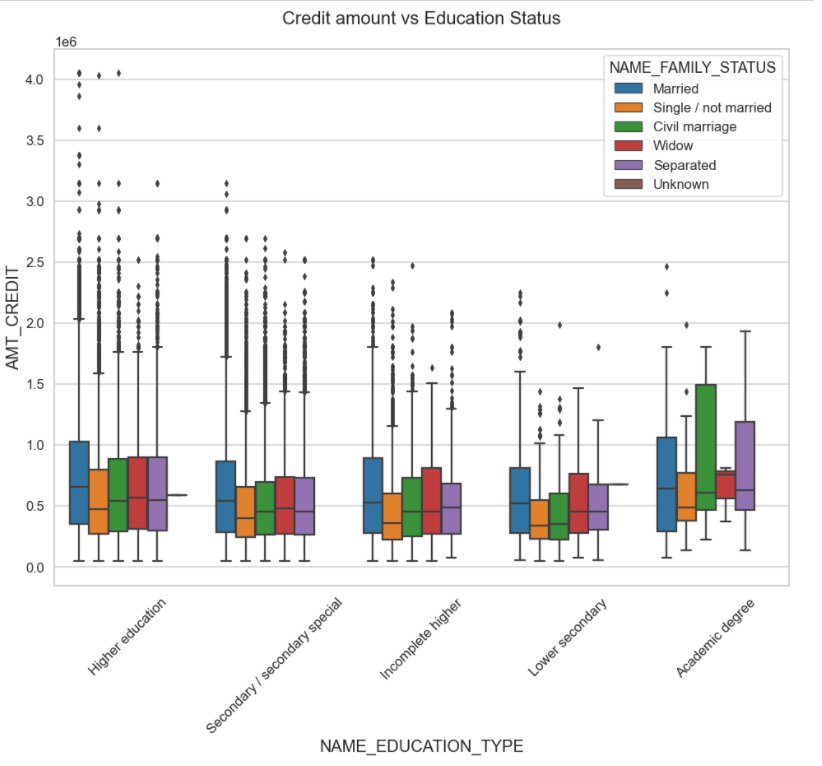


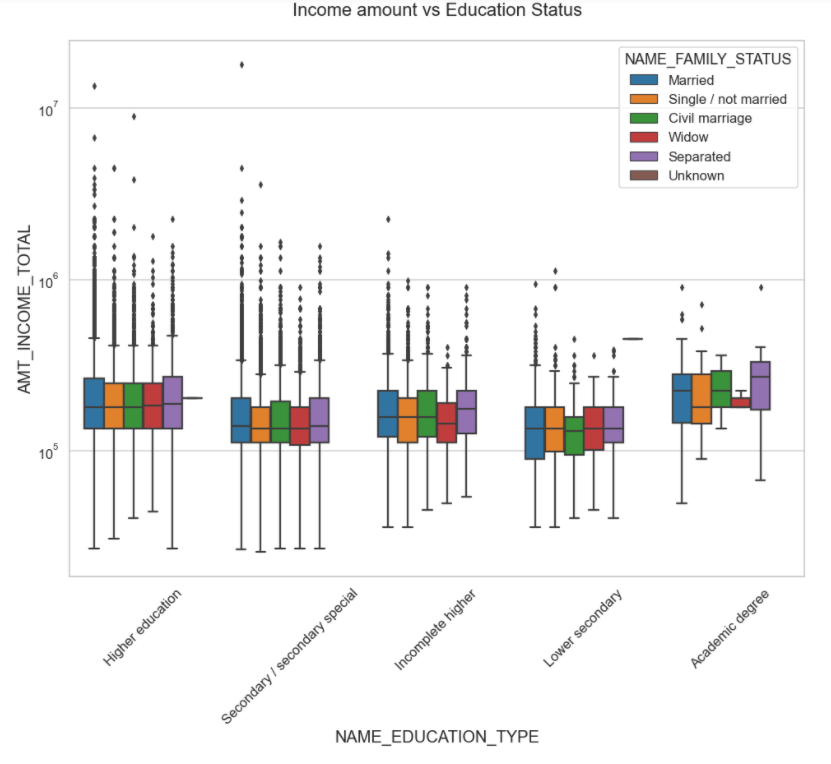
Few points can be concluded from the graph  
• Some outliers are noticed in credit amount.  
• The first quartile is bigger than third  
quartile for credit amount which means  
most of the credits of clients are present in  
the first quartile.

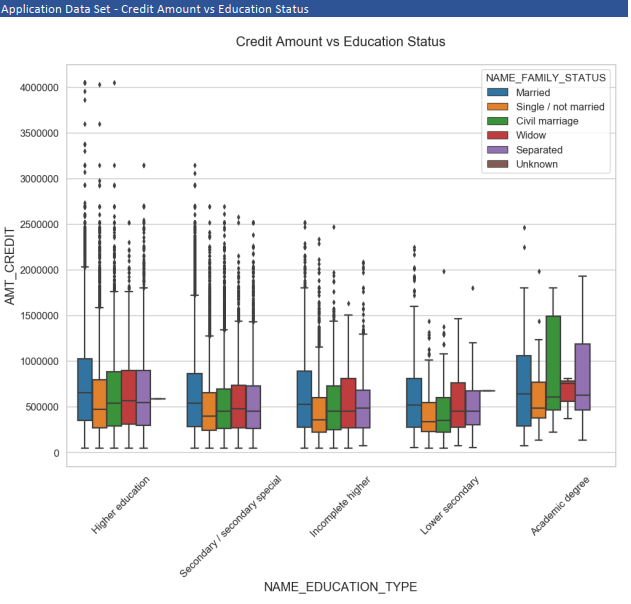


Few points can be concluded from the graph  
• Some outliers are noticed in annuity  
amount.  
• The first quartile is bigger than third  
quartile for annuity amount which means  
most of the annuity clients are from first  
quartile

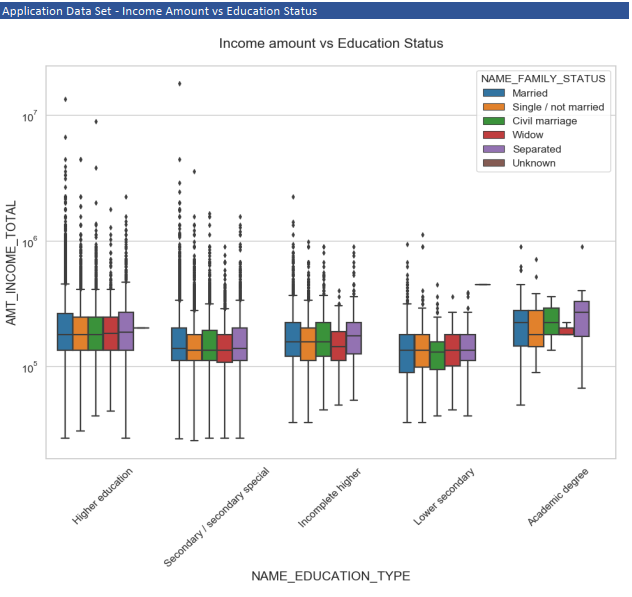
Inference - Bivariate Analysis for Target 0

points can be concluded from the graph.  
• Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education have higher  
number of credits than others.  
• Higher education of family status of 'marriage', 'single' and 'civil marriage' have more outliers.  
• Civil marriage for Academic degree has most of the credits in the third quartile 

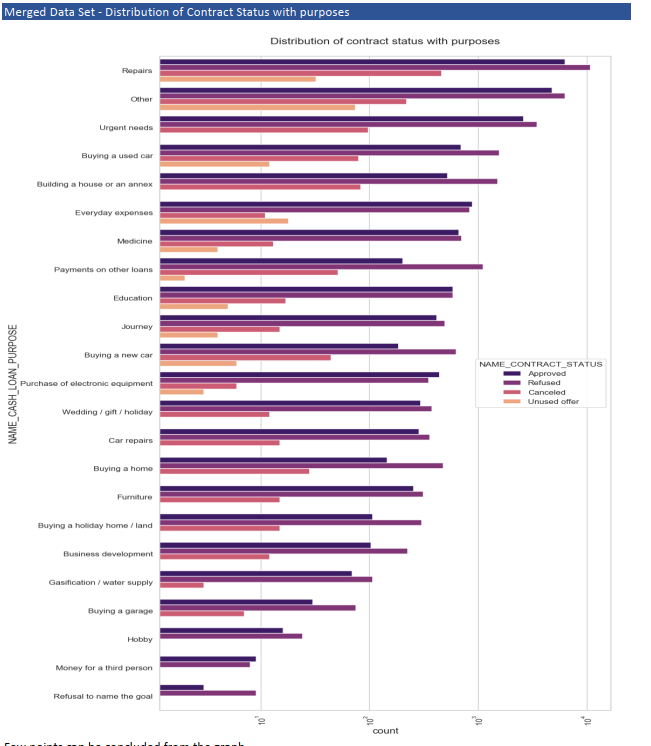
points can be concluded from the graph.  
• For Education type 'Higher education' the income amount mean is mostly equal with family status. It does  
contain many outliers.  
• Less outlier are having for Academic degree but they are having the income amount is little higher that  
Higher education.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

Inference - Bivariate Analysis for Target 1 

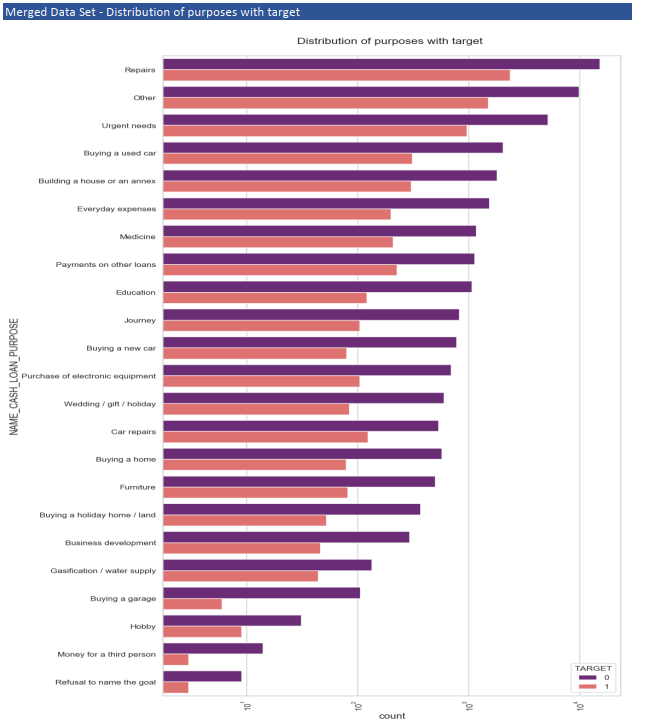
Few points can be concluded from the graph.  
• Quite similar from Target 0, we can say that Family status of 'civil marriage', 'marriage' and 'separated' of  
Academic degree education are having higher number of credits than others.  
• Most of the outliers are from Education type 'Higher education' and 'Secondary’.  
• Civil marriage for Academic degree is having most of the credits in the third quartile



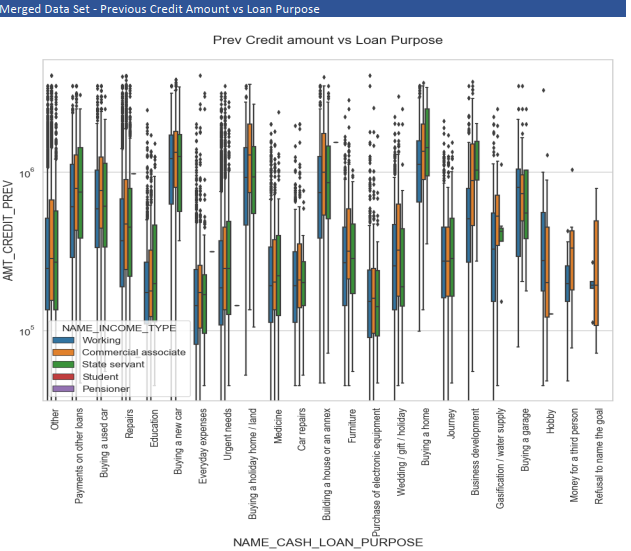
Few points can be concluded from the graph  
• Have some similarity with Target0, From above boxplot for Education type 'Higher education' the income  
amount is mostly equal with family status.  
• Less outlier are for Academic degree but there income amount is little higher that Higher education. 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

Inference - Univariate Analysis after merging previous data 

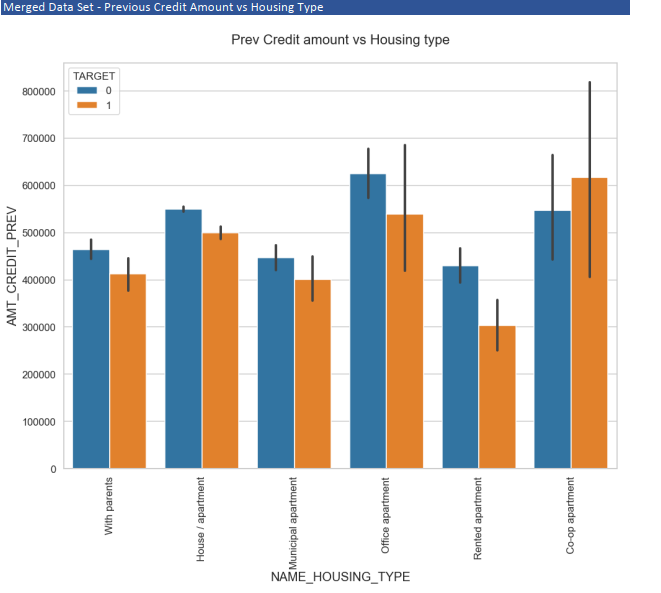
Few points can be concluded from the graph  
• Most rejection of loans came from purpose 'repairs'.  
• For education purposes we have equal number of approves and rejection  
• Paying other loans and buying a new car is having significant higher rejection than approves



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 

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.



Points to be concluded   
• Here for Housing type, office apartment has 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 shouldn’t focus on income type ‘Working’ as they have the most number of unsuccessful  
payments.  
2. Also with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.  
3. More clients from housing type ‘With parents’ should be approached as they have the least number of unsuccessful payments.

4.Banks should focus more on contract type ‘Student’ ,’Pensioner’ and ‘Businessman’ with housing ‘type  
other than ‘Co-op apartment’ for successful payments.