Credit EDA Case Study

By - PRIYANKA SHAH

Problem Statement - I

- 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 specializes in lending various types of loans to urban customers. You have to use EDA to analyze 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.

Problem Statement Understanding - I

- When people ask for a loan, the company has to decide whether to give them the money or not. There are two big risks for the company:
- Risk of Not Approving a Good Applicant: If they say "no" to someone who would have paid back the loan on time, they lose out on making money because they didn't lend to a good customer.
- Risk of Approving a Bad Applicant: If they say "yes" to someone who won't be able to pay the money back, the company loses money because they won't get their money back.
- Now, some people try to trick the company. They might not have a history of borrowing money (like a credit card or previous loans), so the company doesn't know if they're good at repaying loans or not. These people might take advantage of this and not pay the money back.

Problem Statement Understanding - II

- Aim is to look at a bunch of data (information) about these loan applications and to find patterns. These patterns can help the company make better decisions. For example, if people who have been late with payments before are more likely to not pay back their loans, the company might be more cautious about lending to those people.
- The main goal is to figure out what things about the people applying for loans (like their history and characteristics) and the loans themselves (like how big the loan is) make it more likely that they won't pay back. This way, the company can make smarter choices about who to give loans to and how to protect themselves from losing money.
- So, in simple terms, you're trying to use data to find out what makes someone a good or bad borrower, so the company can make better loan decisions and not reject people who can actually repay their loans.

Steps

- Data Understanding
- Data Sourcing
- Data Cleaning
- Univariate Analysis
- Bivariate Analysis
- Correlation
- Merging of the datasets
- Conclusion

Step 1: Data Understanding

Datasets

Three datasets are used for this case-study:

- *'application_data.csv'* contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties.**
- *'previous_application.csv'* contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
- *'columns_description.csv'* is data dictionary which describes the meaning of the variables.

Step 2: Data Sourcing

Steps for Data Sourcing

1. Reading the csv files:

- Reading the csv files for application data and previous application data.

2. Information about the datasets:

- Checking no. of rows and columns of given datasets -df.shape
- Checking the column-wise info of the datasets - df.info(verbose=True)
- Generating statistical information about the numerical columns (or series) of the given datasets, which helps understanding the distribution and characteristics of the data. - df.describe()

Step 3: Data Cleaning

Steps in Data Cleaning

1. Checking the missing values:

- Analysing the missing values to be handled and how to handle those values

2. Checking the outliers:

- Checking whether to remove the outliers, to transform them, or keep them based on goal of our analysis

3. Checking the data imbalance:

- Checking the ratio of data imbalance

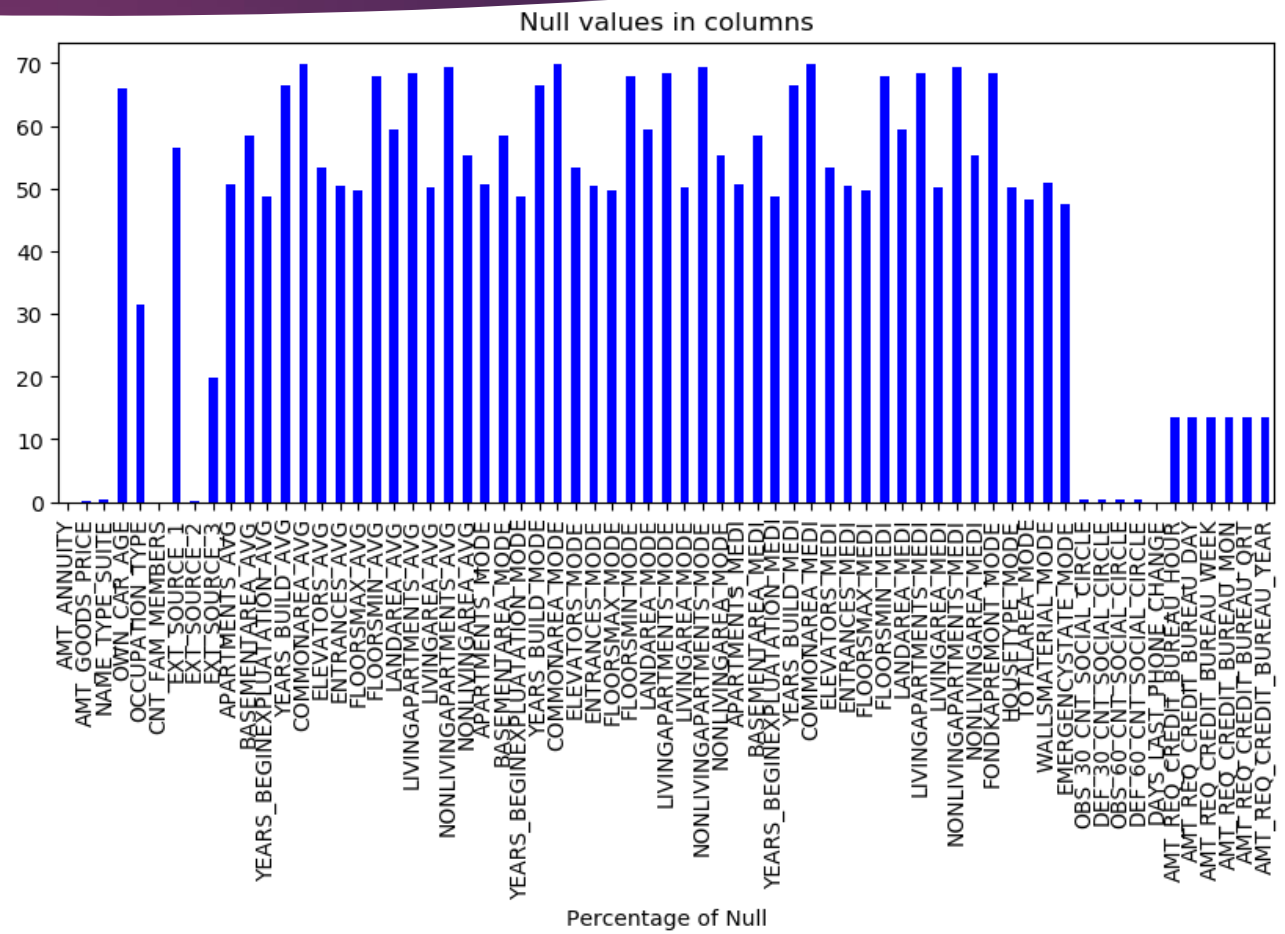
Checking the Missing Values

1. Analyzing the missing data (null values) for both the datasets:

- The figure shows the percentage of null values for application data.

2. Handling the missing values for the datasets:

- Imputing the missing value
- Standardizing the column values (Converting the negative value columns to positive, creating bins)



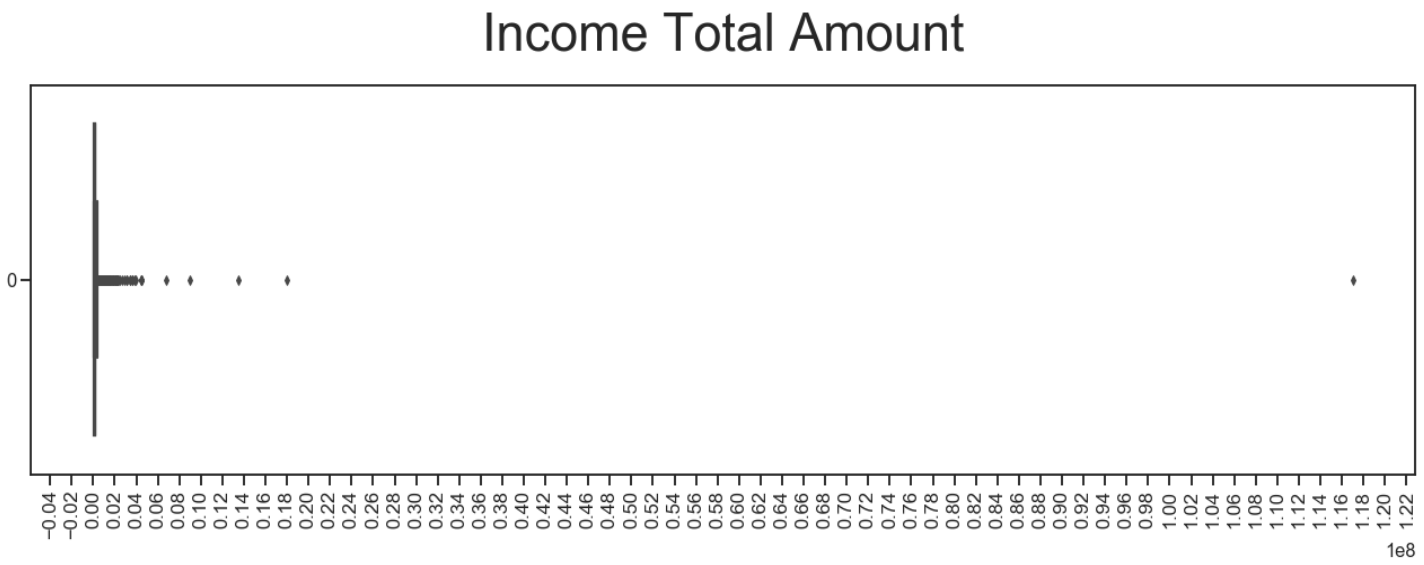
Checking the Outliers

Using Box Plot for Outliers Analysis

Box Plot for AMT_INCOME_TOTAL Variable

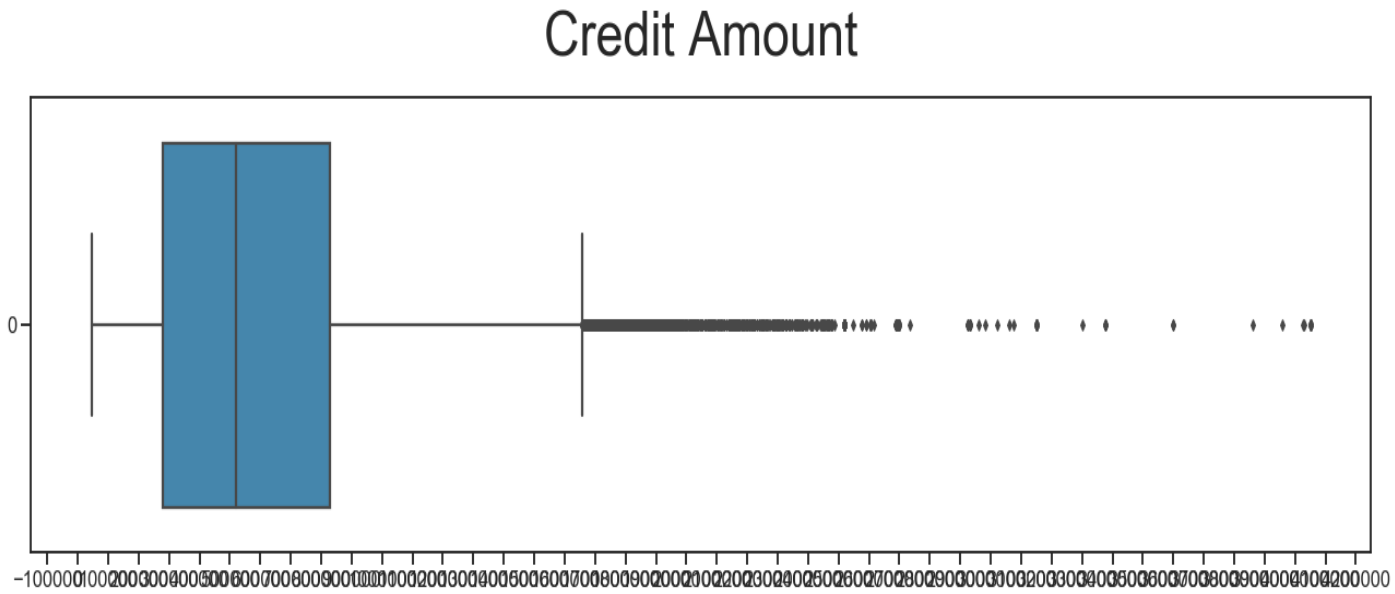
Few points can be concluded from the graph.

- Some outliers are noticed in income total amount.
- Income can vary from people to people. Therefore, no operation is required in this



Box Plot for AMT_CREDIT Variable

Few points can be concluded from the graph.

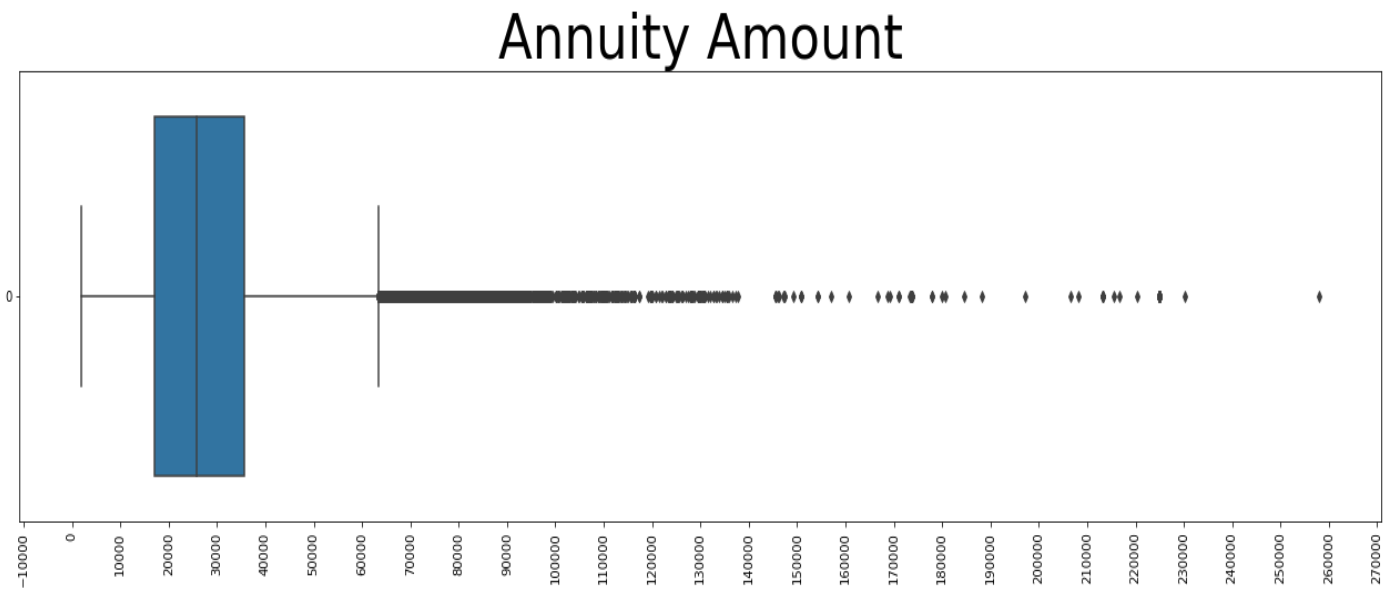


- Some outliers are noticed in credit amount.
- The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.
- The outliers value can handled by converting into bins .

Box Plot for AMT_ANNUITY Variable

Few points can be concluded from the graph.

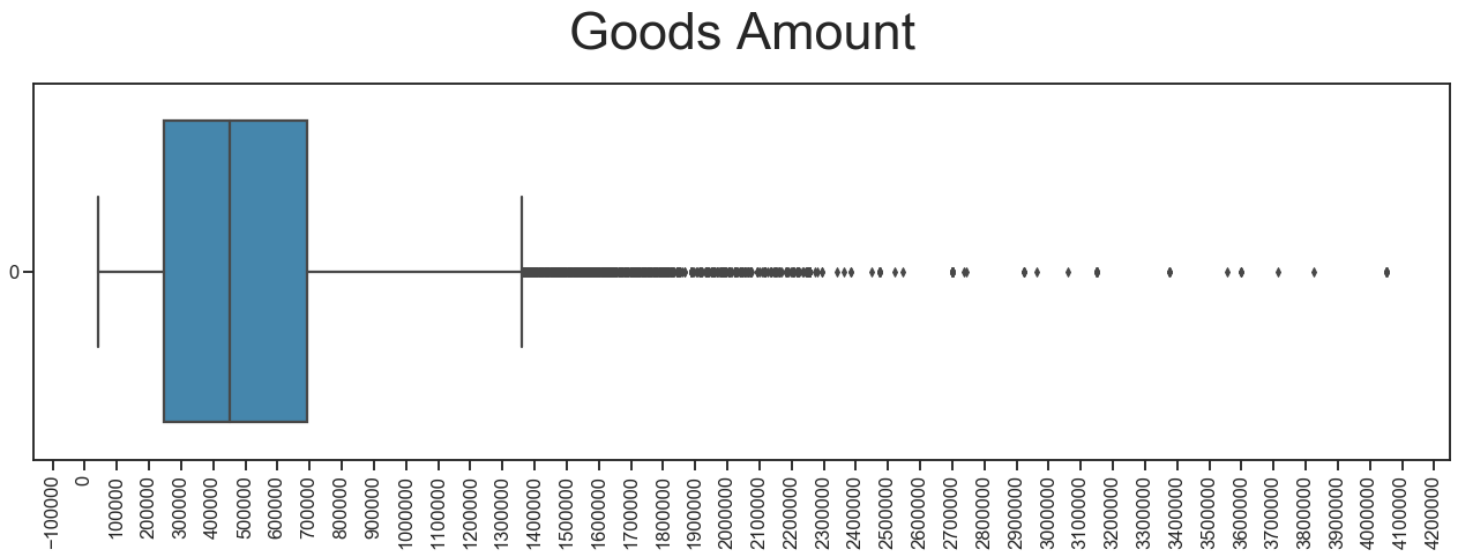
- Some outliers are noticed in annuity amount.
- The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are present in the first quartile.
- The outliers value can handled by converting into bins .



Box Plot for AMT_GOODS_PRICE Variable

Few points can be concluded from the graph.

- Some outliers are noticed in annuity amount.
- Prices can vary from products to products. Therefore, no operation is required in this

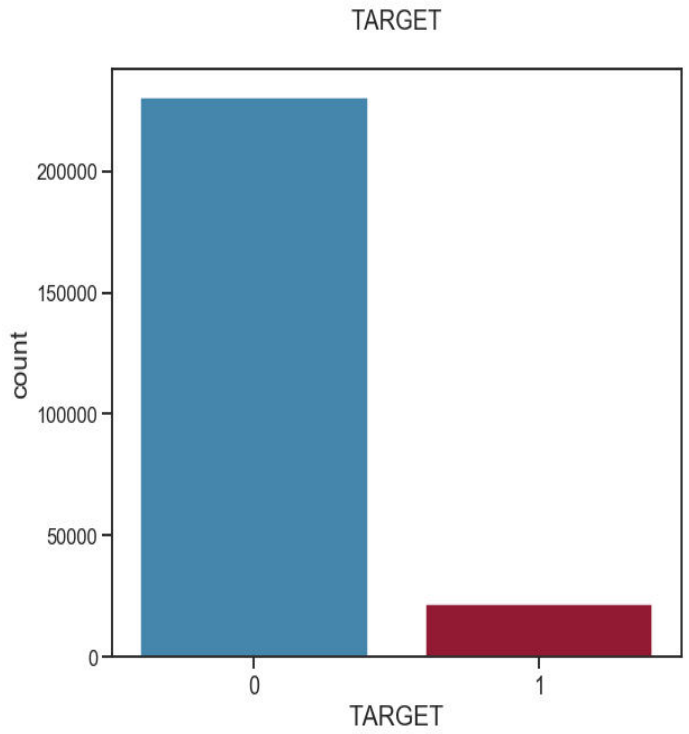


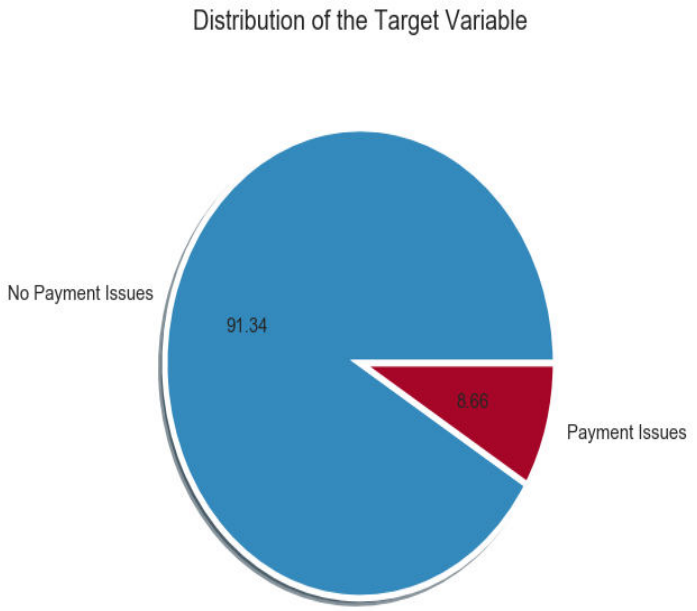
Checking the Data Imbalance

1. We can divide the dataset to two different dataframes:
   - Target=0 (client with no payment difficulties)
   - Target=1 (client with payment difficulties)

2. Defaulters are **8.66**% of total

3. Imbalance Ratio is **10.55**





Step 4: Univariate Analysis

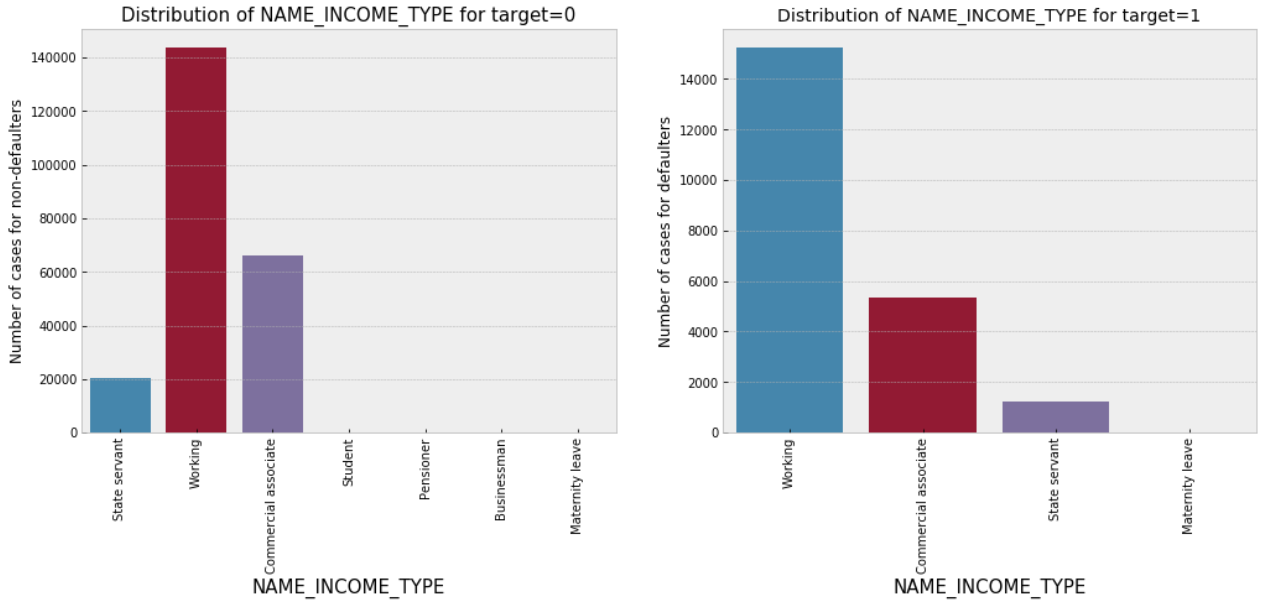
Univariate Analysis

Unordered Categorial Univariate Analysis

Unordered Categorical Data

Graph for NAME_INCOME_TYPE column

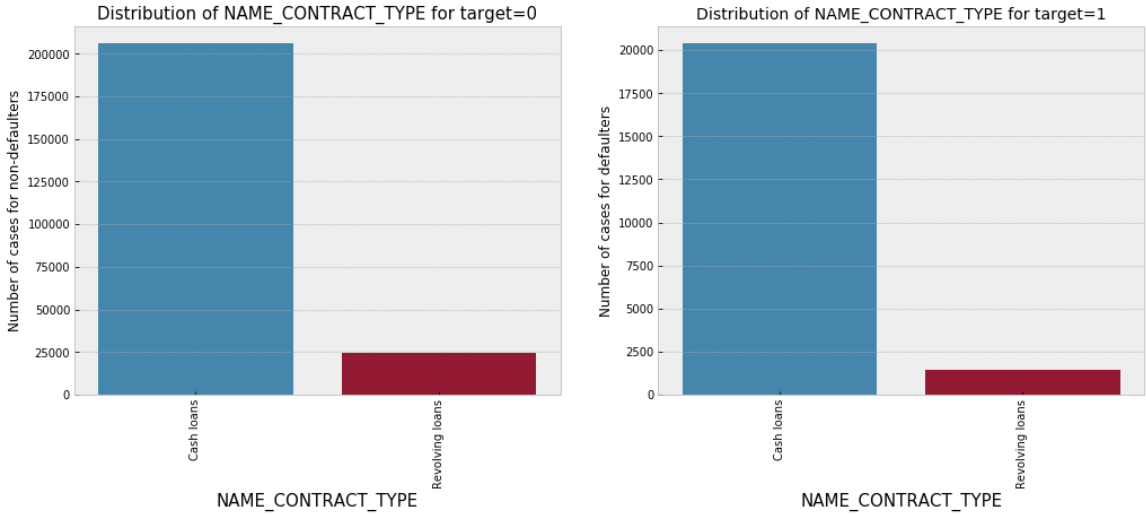
- In the given plot NAME_INCOME_TYPE we can see that most of the defaulters are from the working class



Unordered Categorical Data

Graph for NAME_CONTRACT_TYPE column

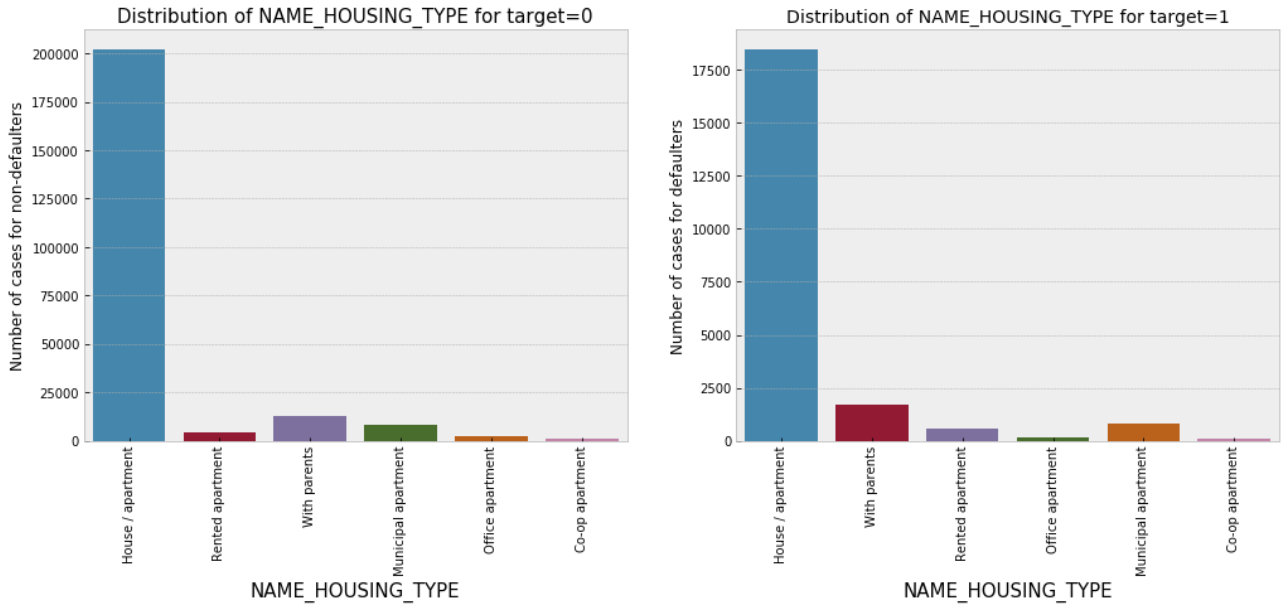
- In the given plot NAME_CONTRACT_TYPE we can see that resolving loans has less distribution in number of cases for defaulters



Unordered Categorical Data

Graph for NAME_HOUSING_TYPE column

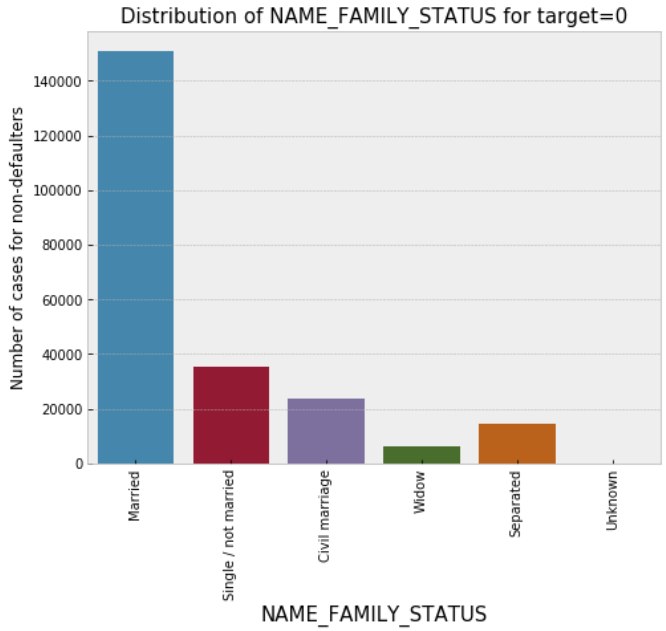
- In the given plot NAME_HOUSING_TYPE people living in the rented apartments and those living with parents have higher default rate

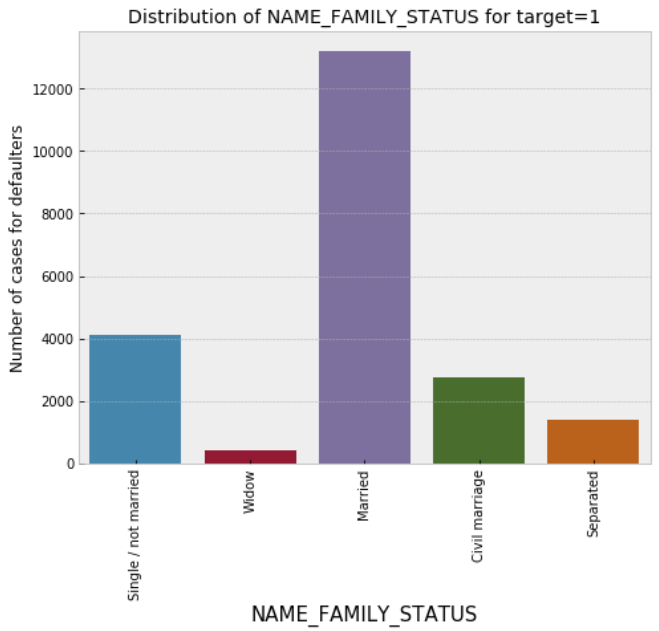


Unordered Categorical Data

Graph for NAME_FAMILY_STATUS column

- In the given plot NAME_FAMILY_STATUS married followed by single/not married has high rate in defaulters and non-defaulters.

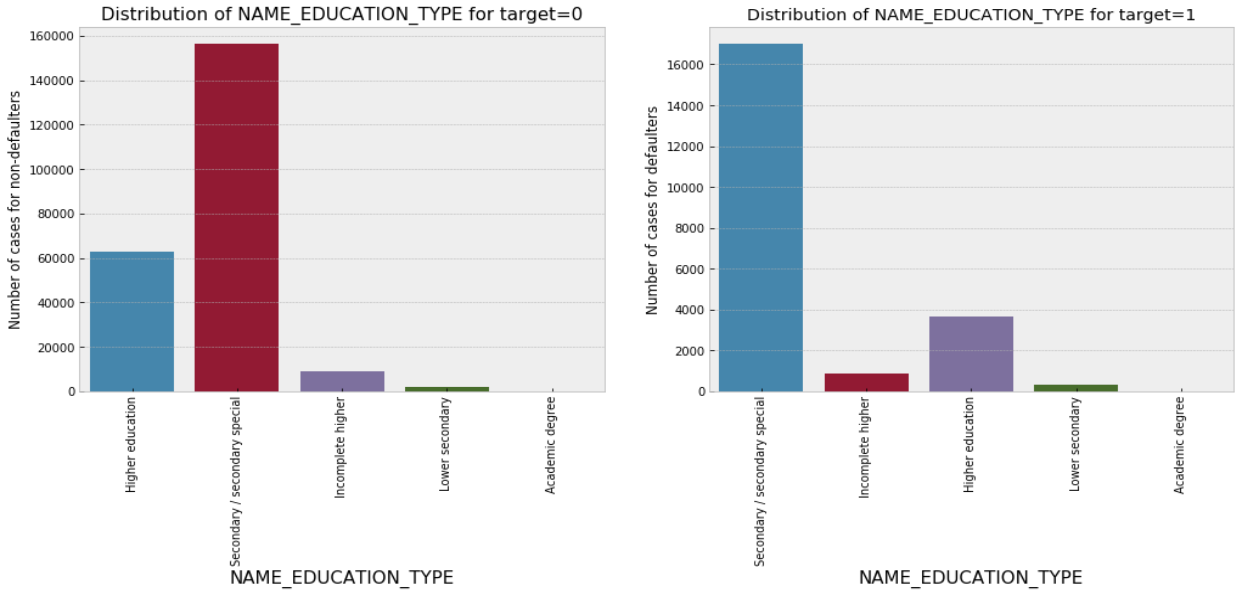




Unordered Categorical Data

Graph for NAME_EDUCATION_TYPE column

- In the given plot NAME_EDUCATION_TYPE Higher education is more in non defaulters as compared to in defaulters category.
- This comes to a conclusion that higher the education, more the individual is capable of paying the loan because of good salary package



- In [80]:

Univariate Analysis

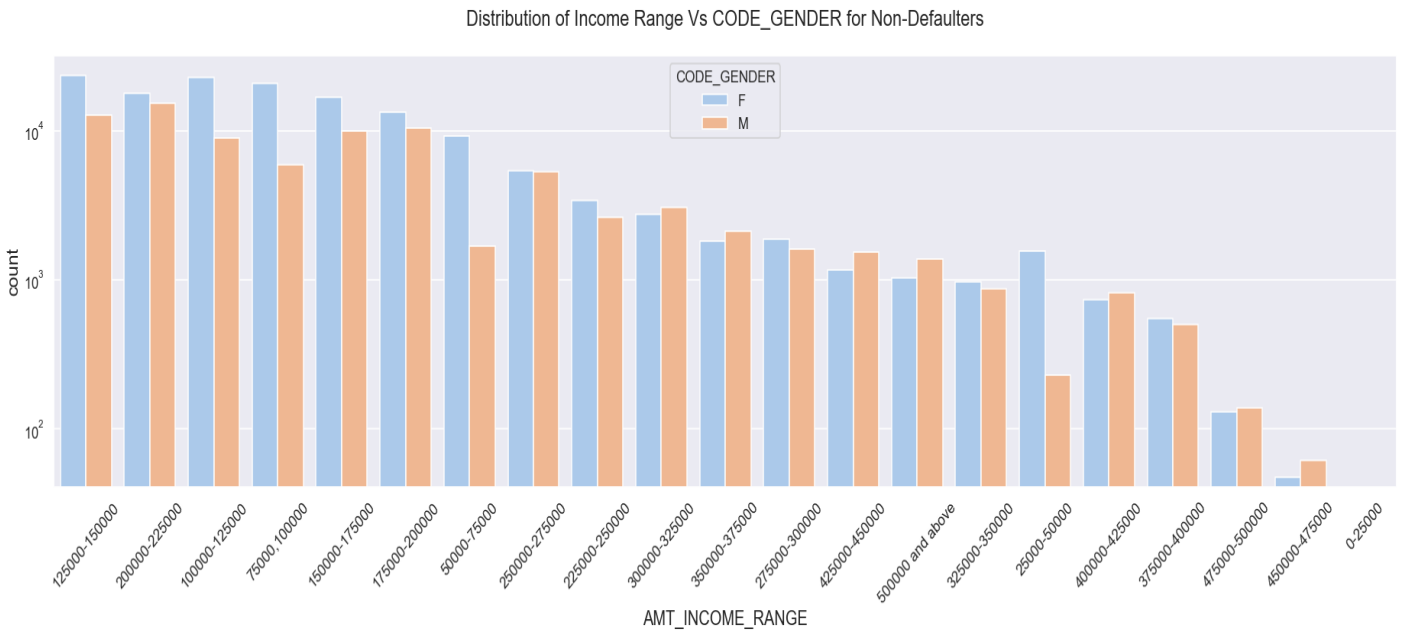
Numerical Categorial Univariate Analysis

Numerical Categorical Data

Distribution of Income Range – CODE GENDER (Target=0)

Points to be concluded from the graph

- 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.

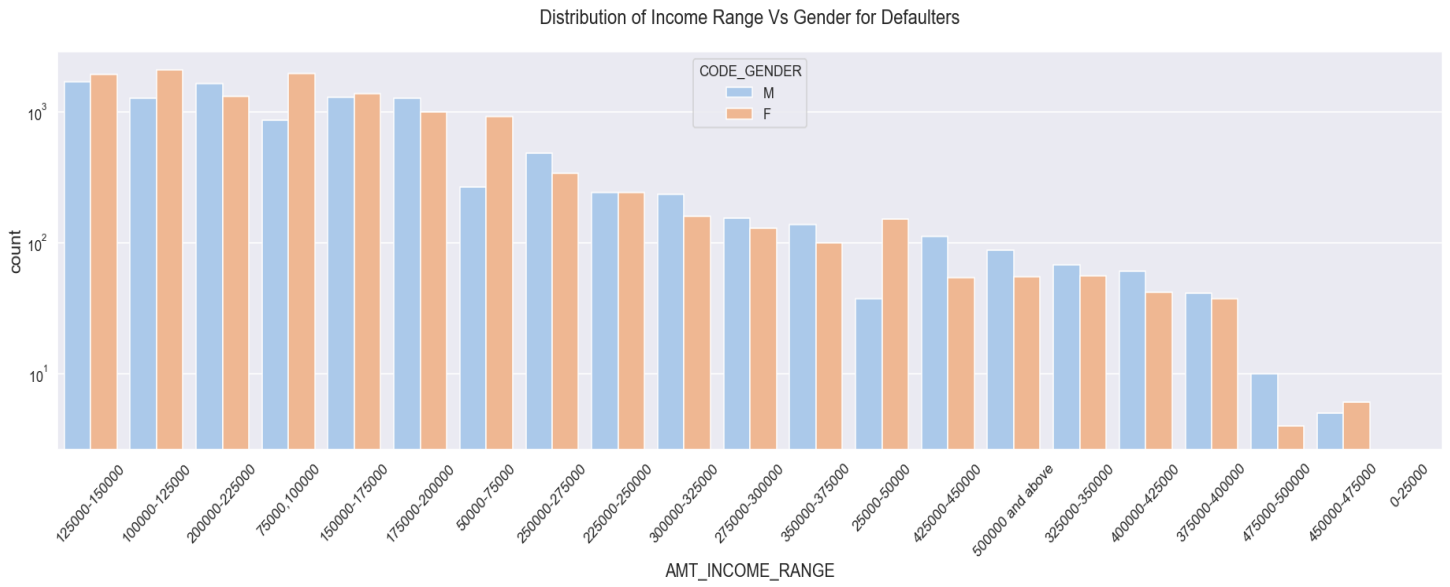


Numerical Categorical Data

Distribution of Income Range – CODE GENDER (Target=1)

Points to be concluded from the graph

- Male counts are higher than female.
- This graph show that males are more than female in having credits for that range.
- Very less count for income range 400000 and above.

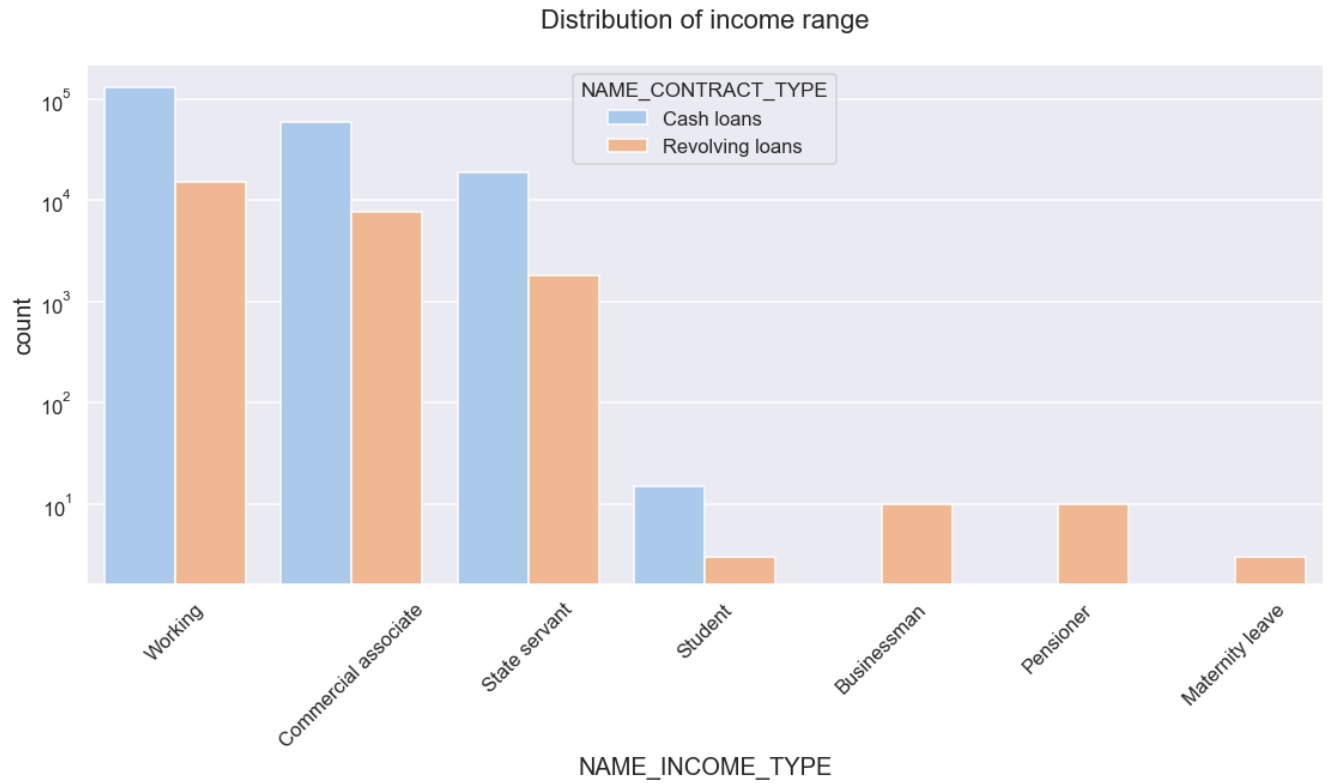


Numerical Categorical Data

Distribution of Income Range – NAME_CONTRACT_TYPE (Target=0)

Points to be concluded from the graph

- Working class has more number of Cash Loans

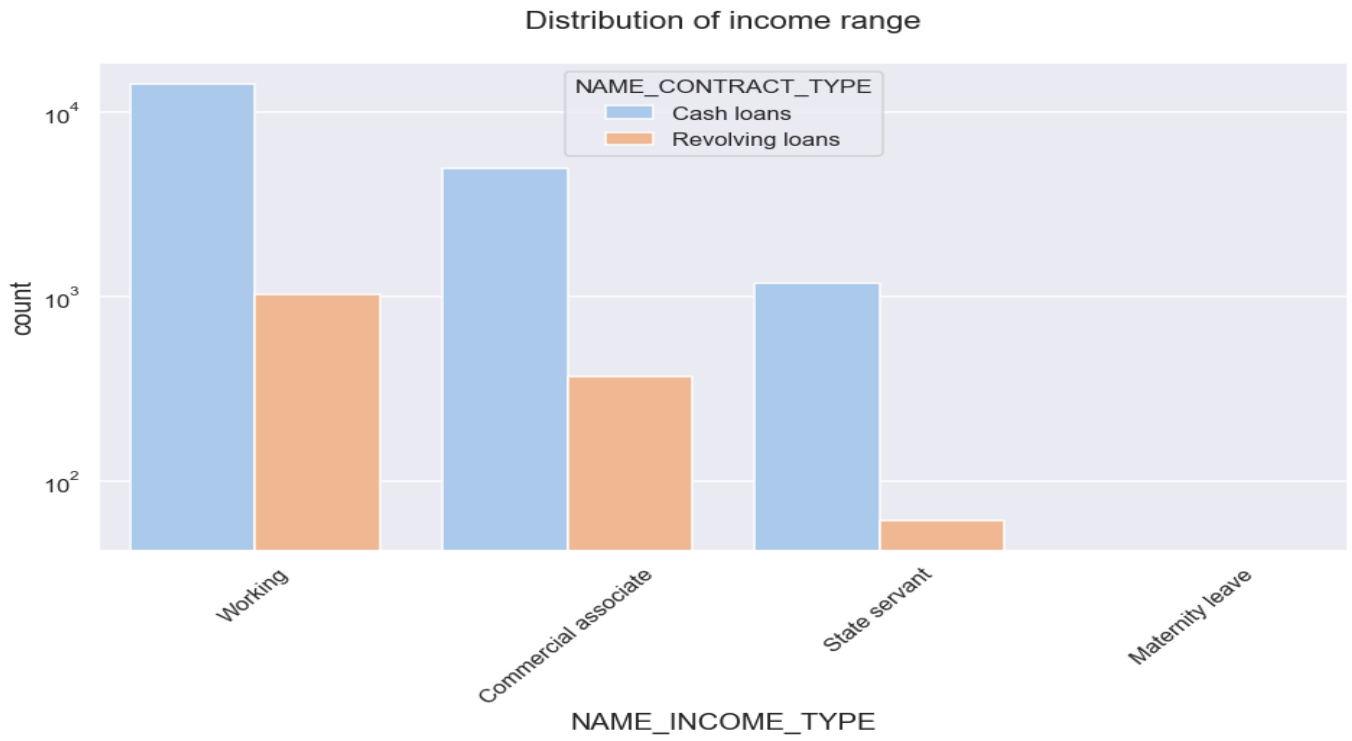


Numerical Categorical Data

Distribution of Income Range – NAME_CONTRACT_TYPE (Target=1)

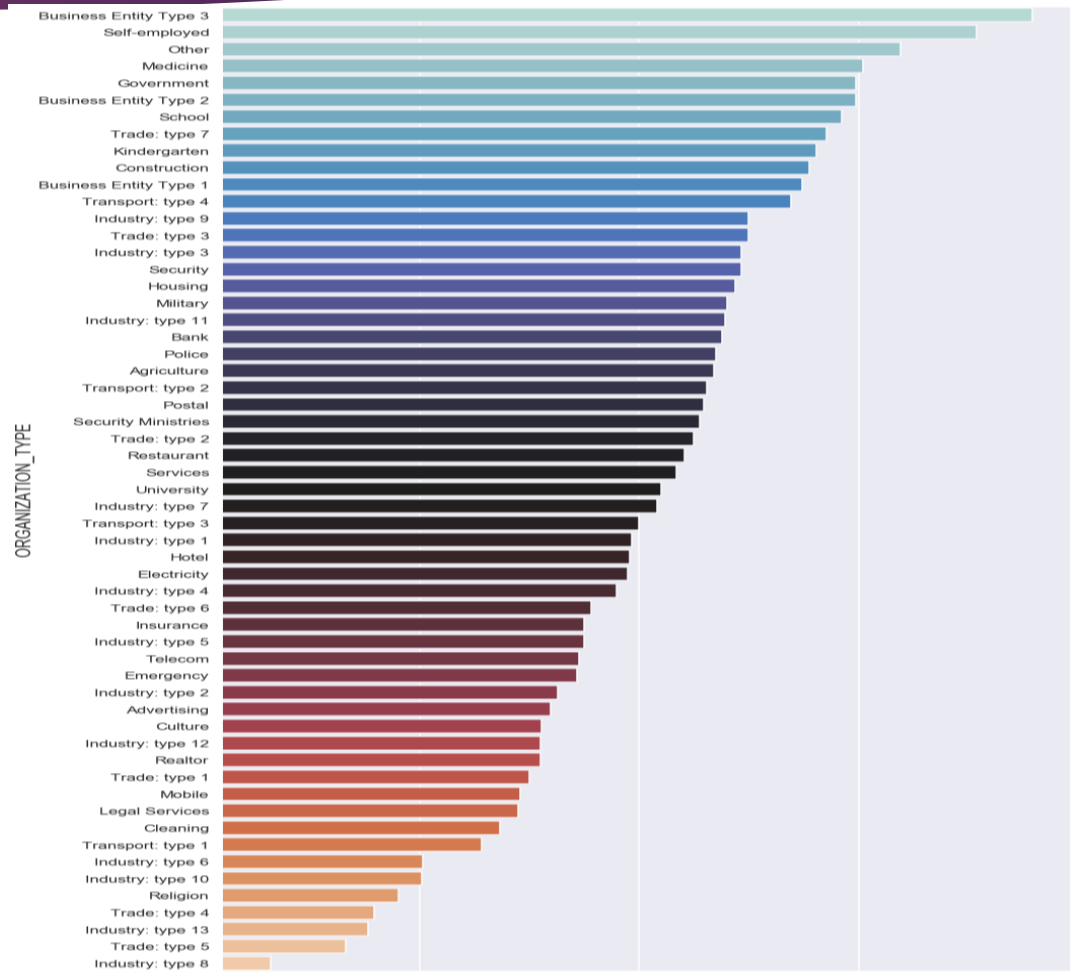
Points to be concluded from the graph

- Working class has more number of Cash Loans



Numerical Categorical Data

Distribution of Organization Type (Target=0)

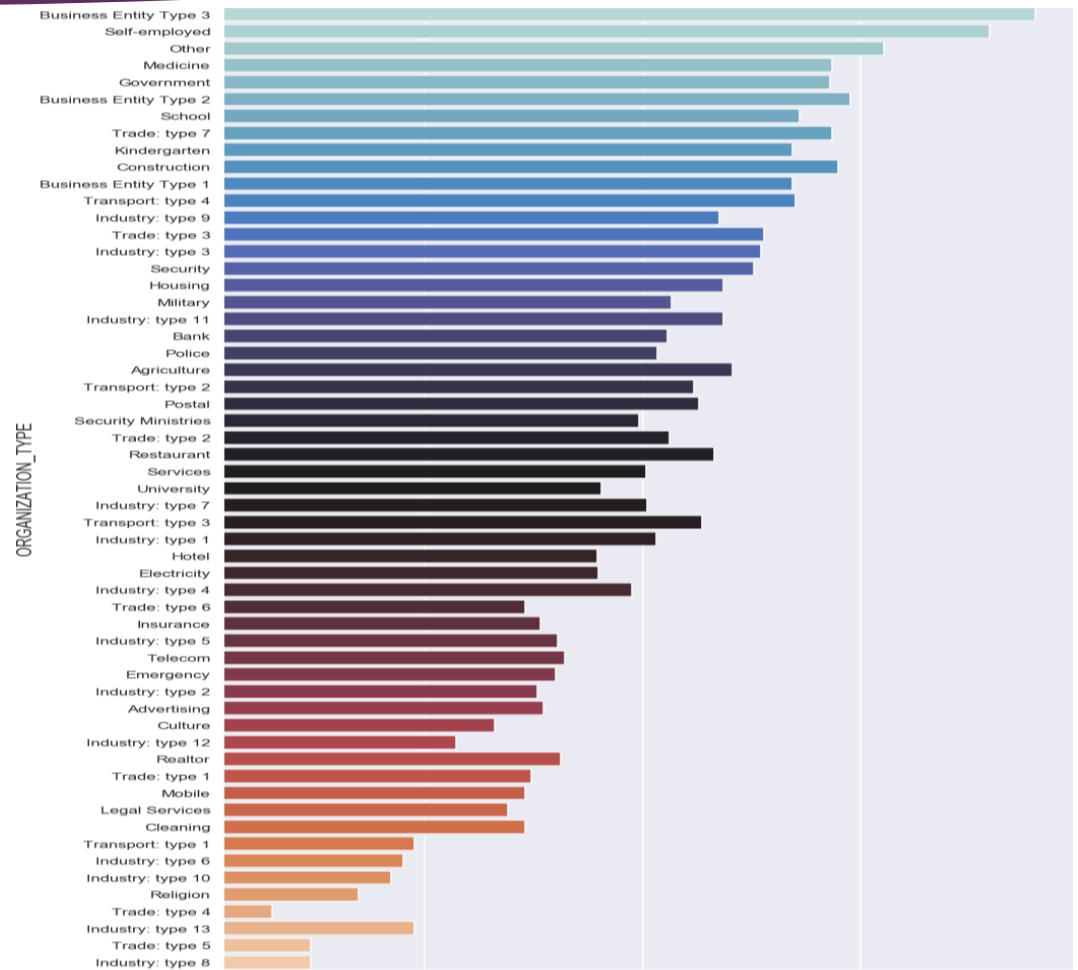


Points to be concluded from the graph

- 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.

Numerical Categorical Data

Distribution of Organization Type (Target=1)



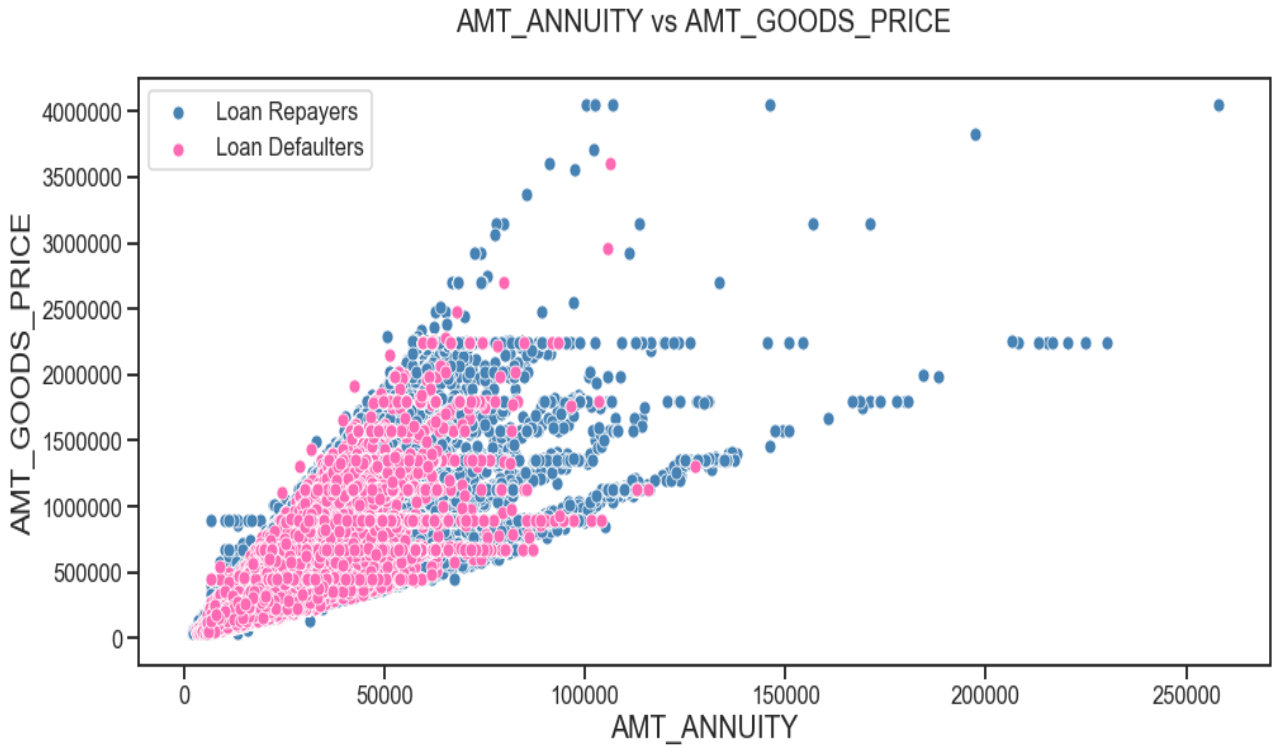
Points to be concluded from the graph

- 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.

Step 5: Bivariate Analysis

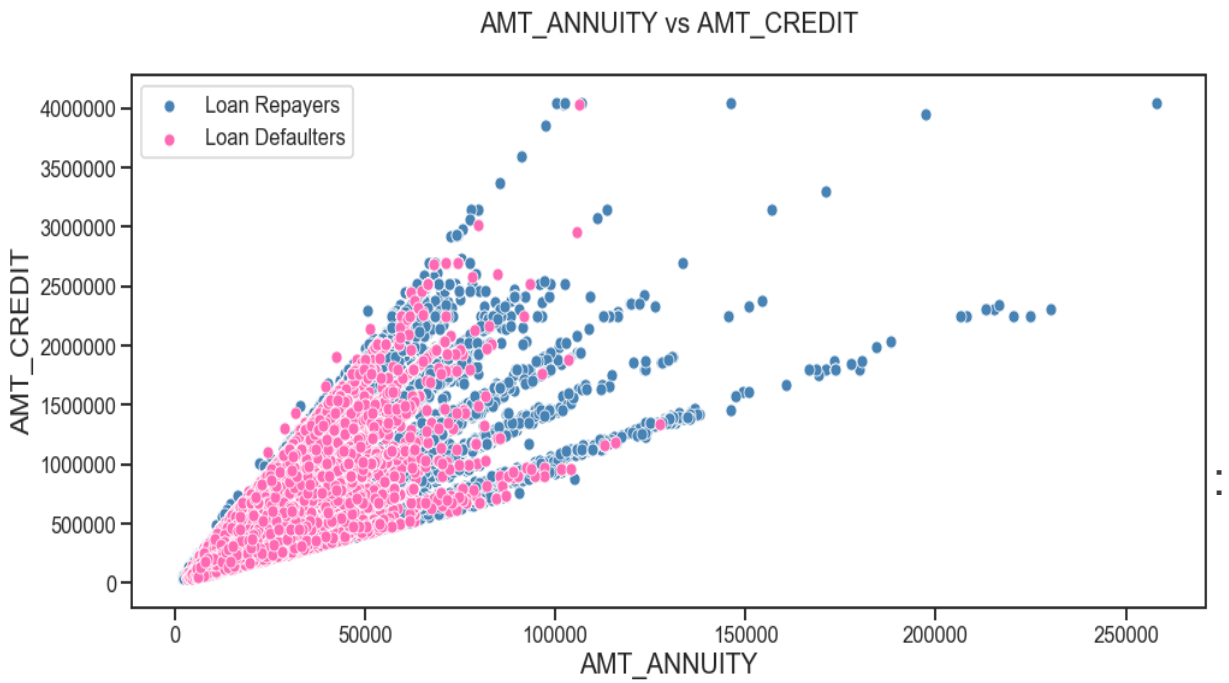
AMT_ANNUITY Vs AMT_GOOD_PRICE

- Correlation between Amount Annuity & Amount Goods Price is pretty moderates but they are not thoroughly correlated because there are above par values for both the columns.



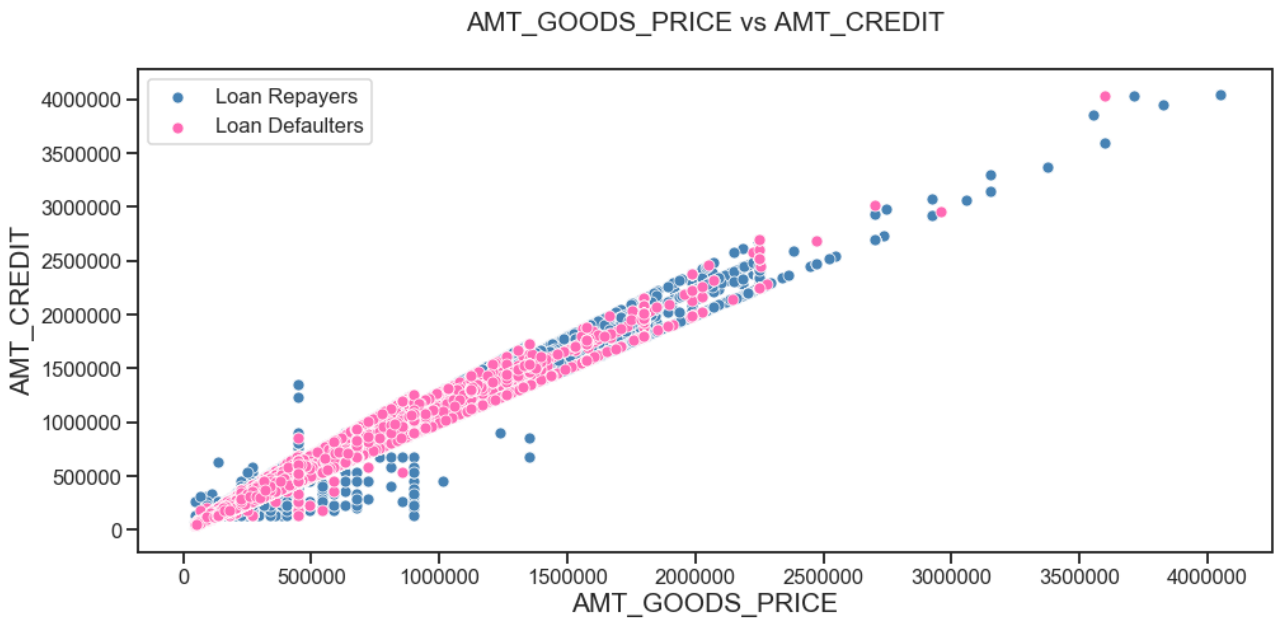
AMT_ANNUITY Vs AMT_CREDIT

- There's a decrease in defaulters when Amount Annuity increases.
- Most of the defaulters are having AMT_ANNUITY values less than 60000.



AMT_CREDIT Vs AMT_GOOD_PRICE

- Clients having good prices will repay their loans.

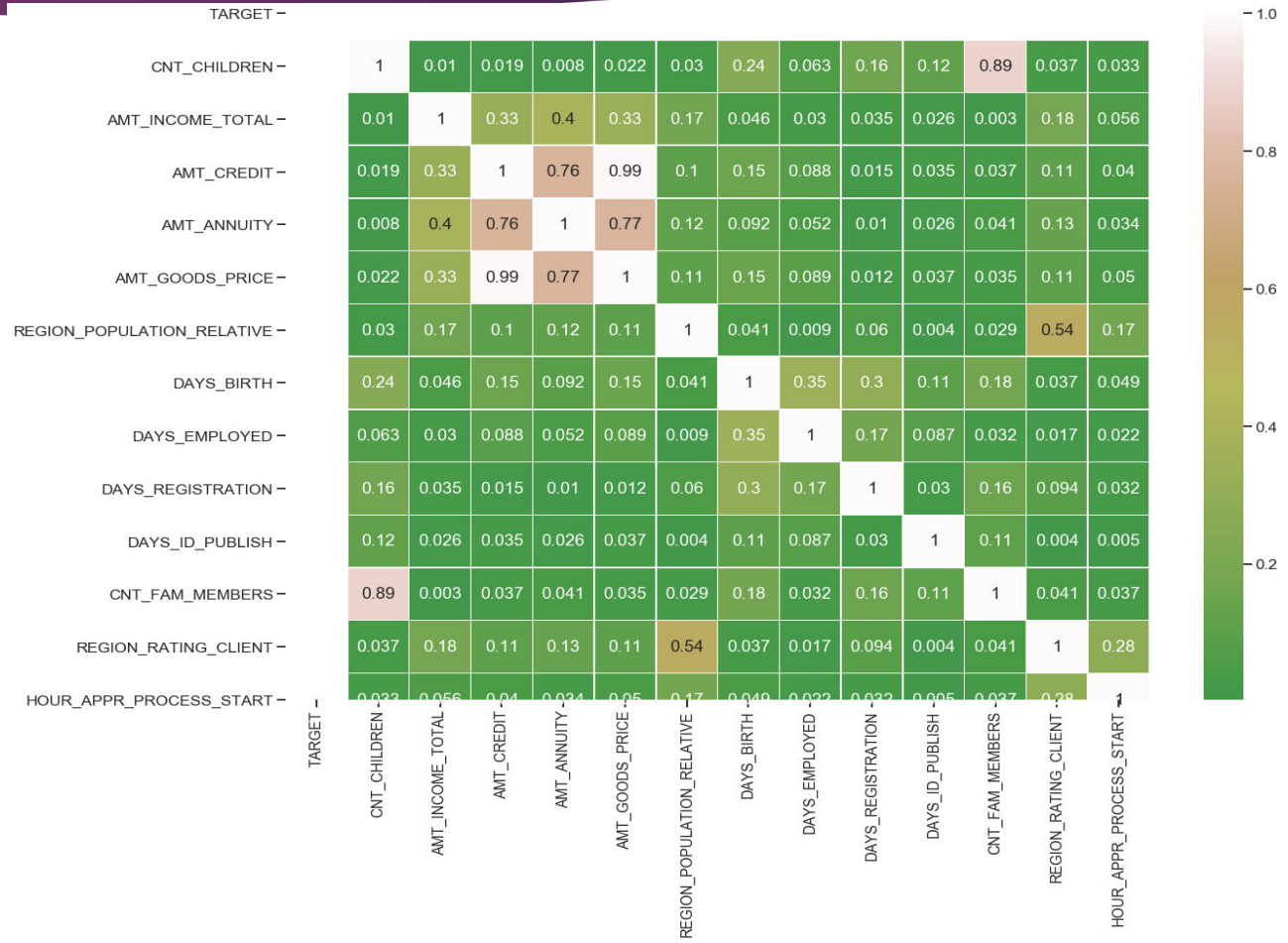


Step 6: Correlation

Correlation for Target=0

Points to be concluded from the graph presented before.

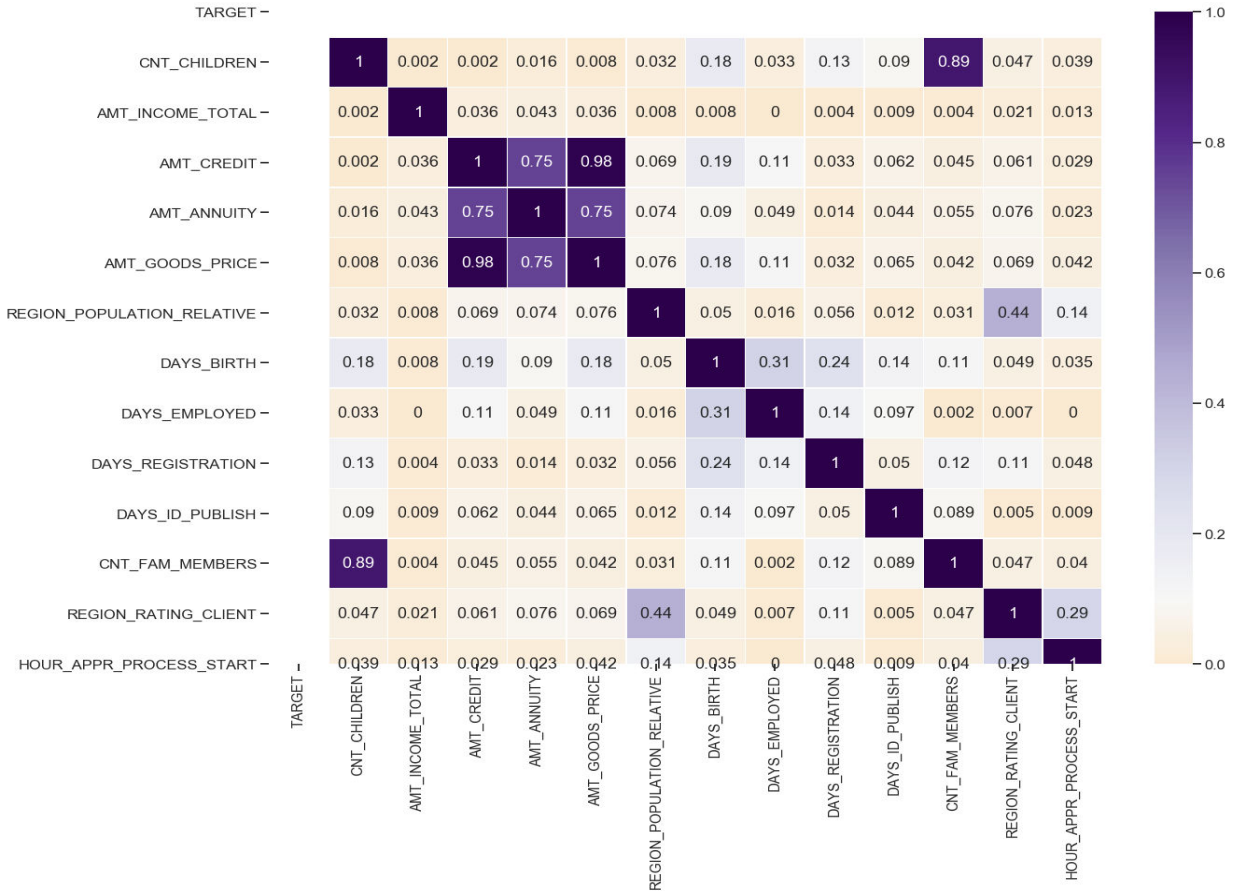
- Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa.
- Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.
- Income amount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



Correlation for Target=1

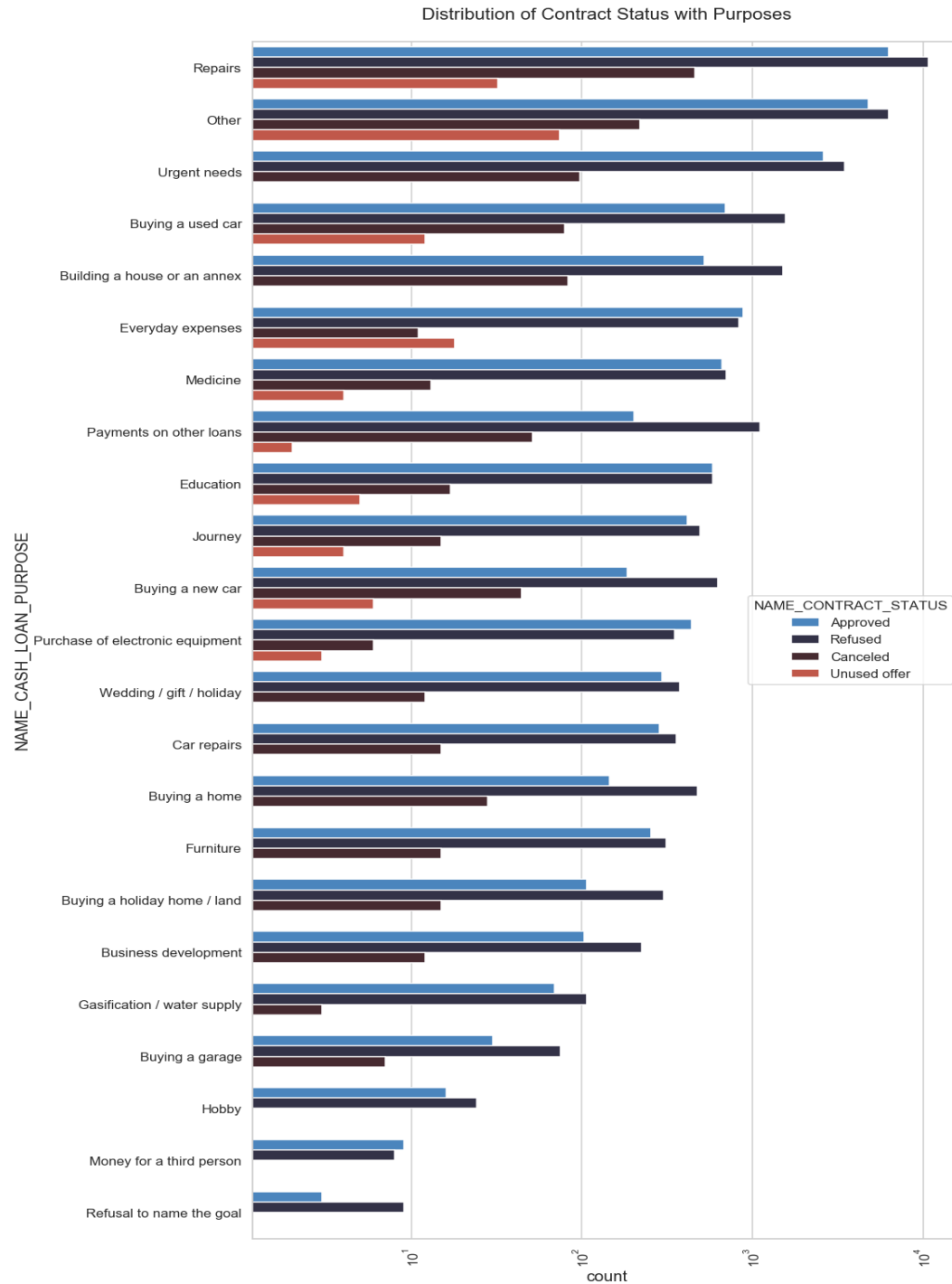
Points to be concluded from the graph presented before.

- The loan annuity correlation with credit amount and also with goods price has slightly reduced in defaulters when compared to the one's who pays on time.
- We can also see that one who pay's on time have high correlation in number of days employed when compared to defaulters



Step 7: Merging the dataset

Distribution of Contract Status with Purposes

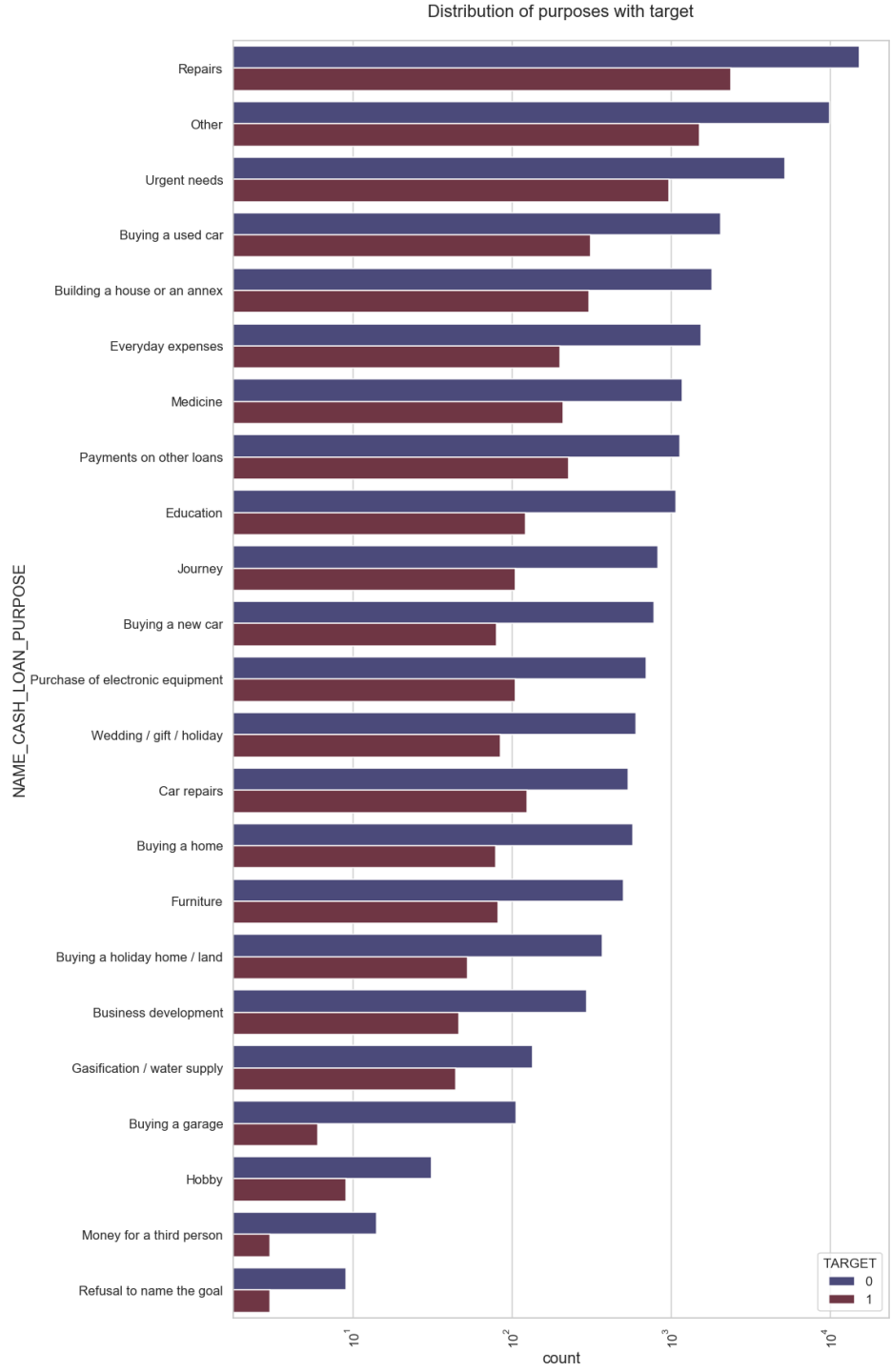


Distribution of Contract Status with Purposes

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.

Distribution of Purpose with Targets



Distribution of Purpose with Targets

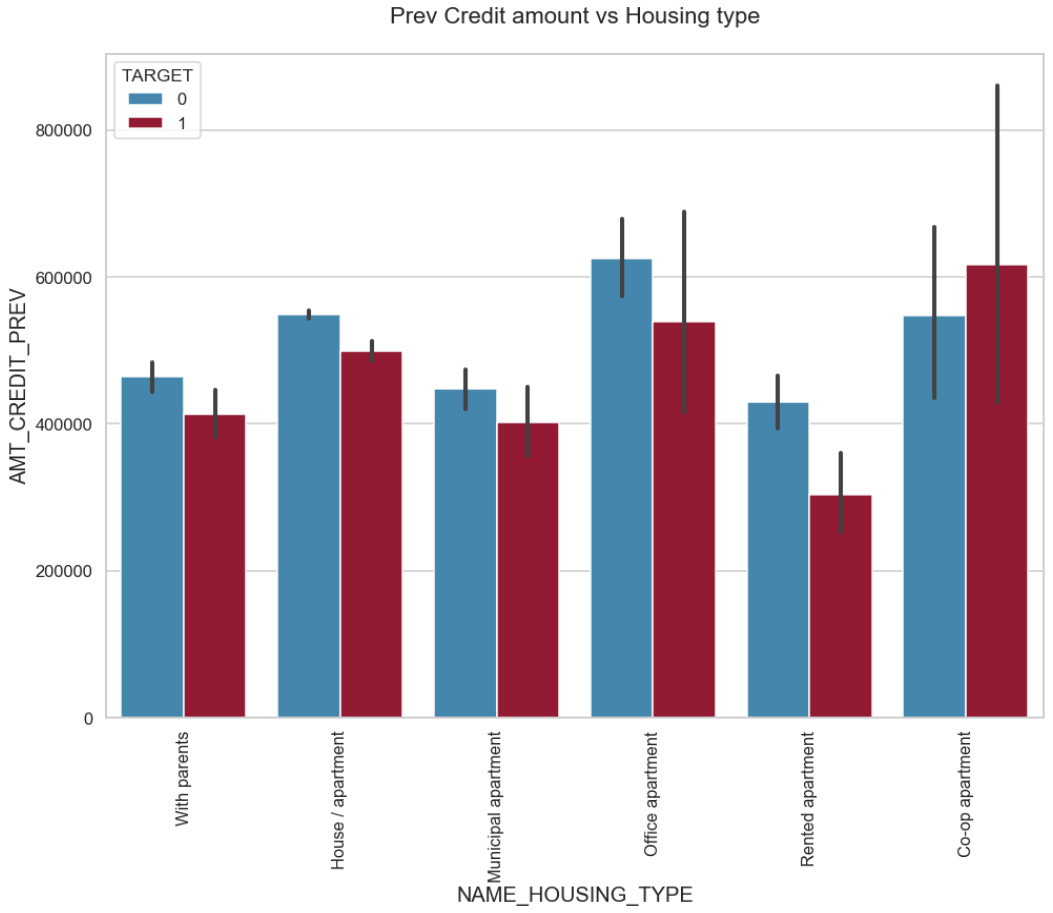
Few points can be concluded from the graph

- 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.

Prev 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, bank should avoid giving loans to the 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.



Step 8: Conclusion

Conclusion

1. Final Conclusion for groups where loan can be provided:

- Old people of any income group
- Female clients with higher education
- Any client who's previous loan was approved
- Clients with high income category
- Giving loans to married is recommended
- Business Entity Type 3 and self employed people
- Clients from housing type 'With parents'.

2. Banks should give more revolving loans

Conclusion

1. Final Conclusion for groups where loan should not be provided:

- Loan purpose “Repair”
- Income type “Working”
- Previously refused loan status group
- Unemployed Clients
- Lower secondary and secondary educated clients