EDA Case Study

BY: AFZAL AHMAD

Problem Statement:

- **Introduction**

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

- **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 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:

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):
Approved: The Company has approved loan Application
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.
Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
Unused offer: Loan has been cancelled by the client but at different stages of the process.
In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

Business Objectives
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.
To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

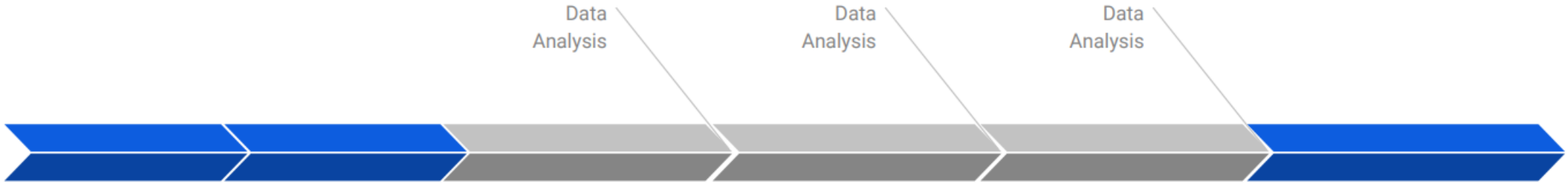
This dataset has 3 files as explained below:

1. '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.

2. '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.

3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Problem solving methodology



Data Cleaning

Removing the null valued columns, unnecessary variables and checking the null value percentage and removing the respective rows.

Data Understanding

Working with the Data Dictionary and getting knowledge of all the columns and their domain specific uses

Univariate Analysis

Analysing each column, plotting the distributions of each column.

Segmented Univariate Analysis

Analysing the continuous data columns with respect to the categorical column

Bivariate Analysis

Analysing the two variable behaviour like term and loan status with respect to loan amount.

Recommendations

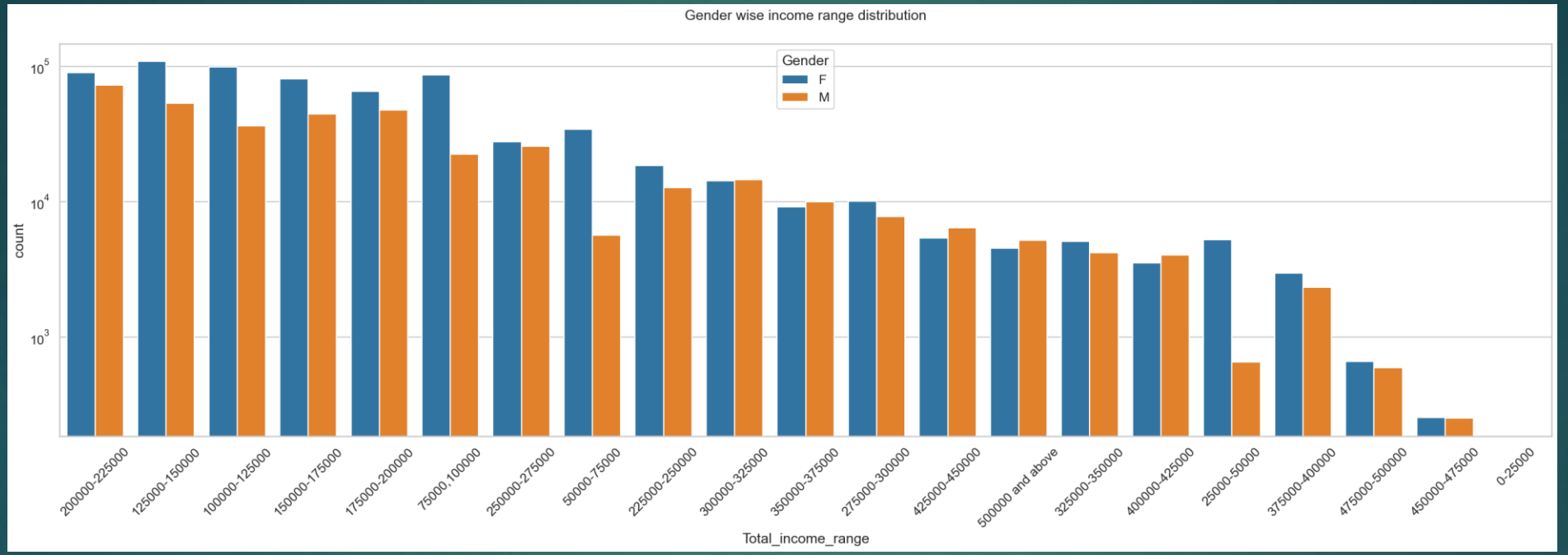
Analysing all plots and recommendations for reducing the loss of business by detecting columns best which contribute to loan defaulters.

Univariate Analysis

- For univariate analysis I merged both the data set *'application_data.csv' and 'previous_application.csv' .*
- *After merging the data I cleaned the data set removed null values and fill the null values with mean, median and mode or appropriate variable.*
- *After that I created two new dataset by name difficulties_df (focusing on Target 1) and ON_time_payment_df (focusing on Target 0).*

Categorical Univariate analysis for target 0

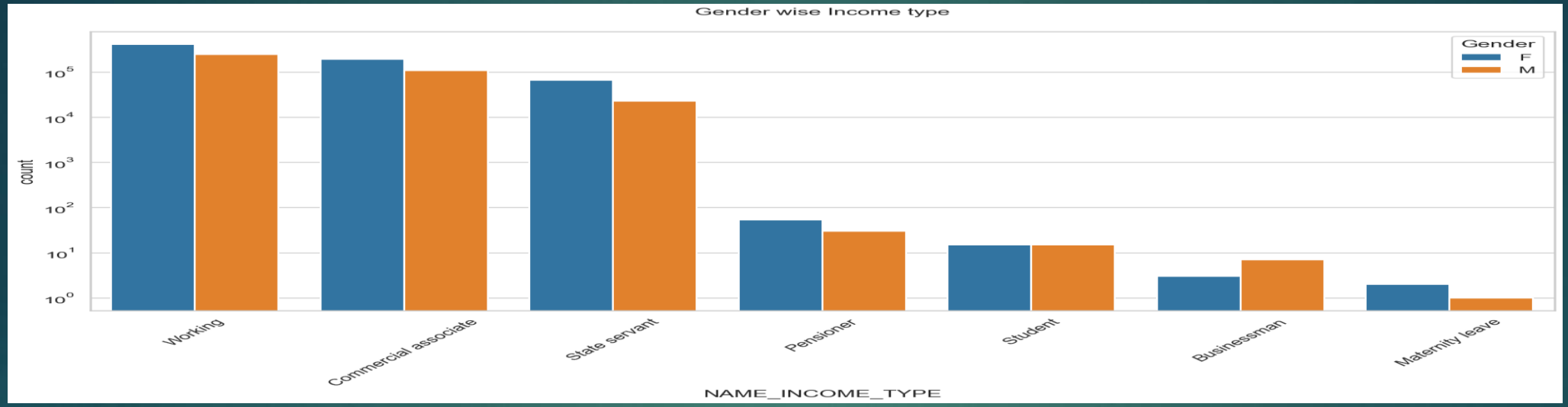
Income Range Distribution:



Points To Be Concluded From Above Graph for total income range:

- Females income are higher than male.
- Income range from 75,000 to 2,25,000 having higher no. of credits.
- 4,00,000 and more are having low no. of credits

Income Type Distribution



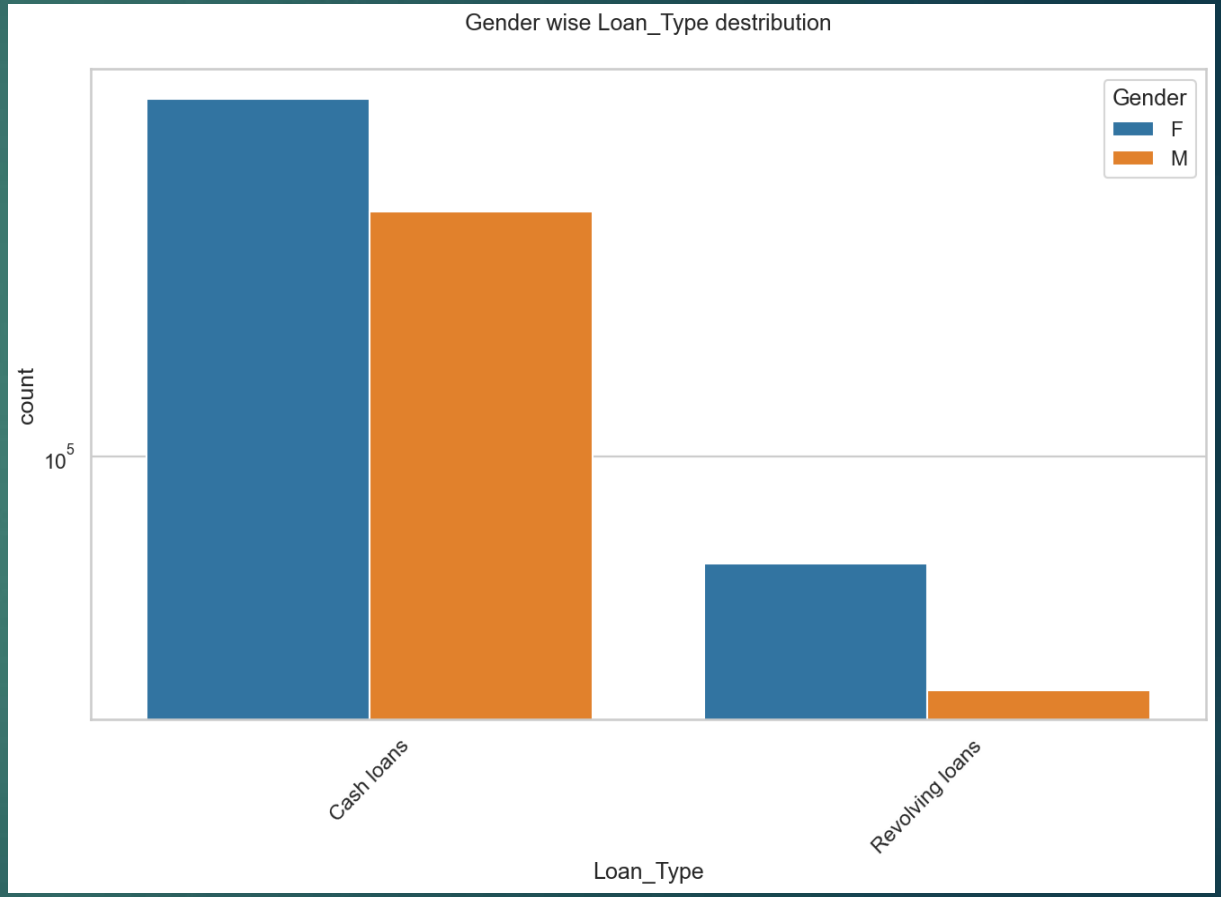
Points To Be Concluded From Above Graph for Income Type Distribution:

- working', 'commercial associate', and 'State Servant' the number of credits are higher than others.
- Low number of credits for income type 'student' ,'pensioner', 'Businessman' and 'Maternity leave'.

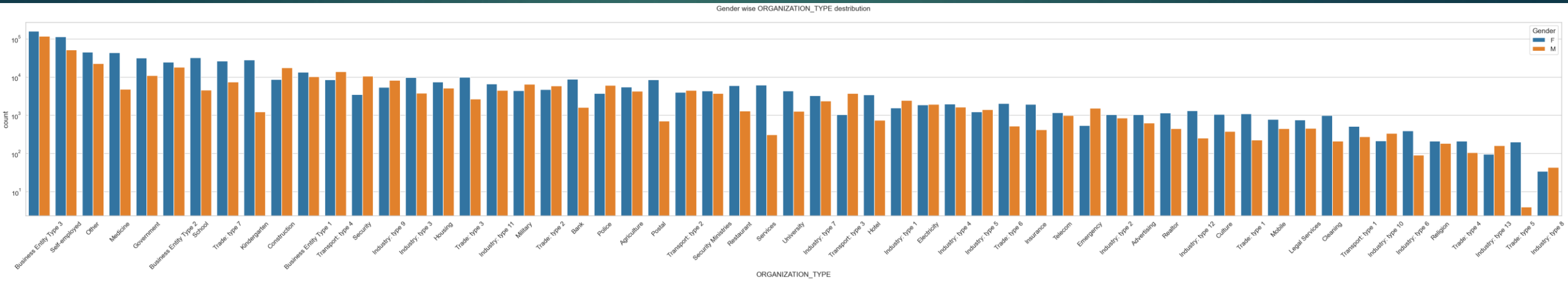
Income Type Distribution:

Points To Be Concluded From Graph for Loan Type Distribution:

- Cash loan is higher than revolving loan.
- Female takes more cash loan than male.



Organization Type Distribution:

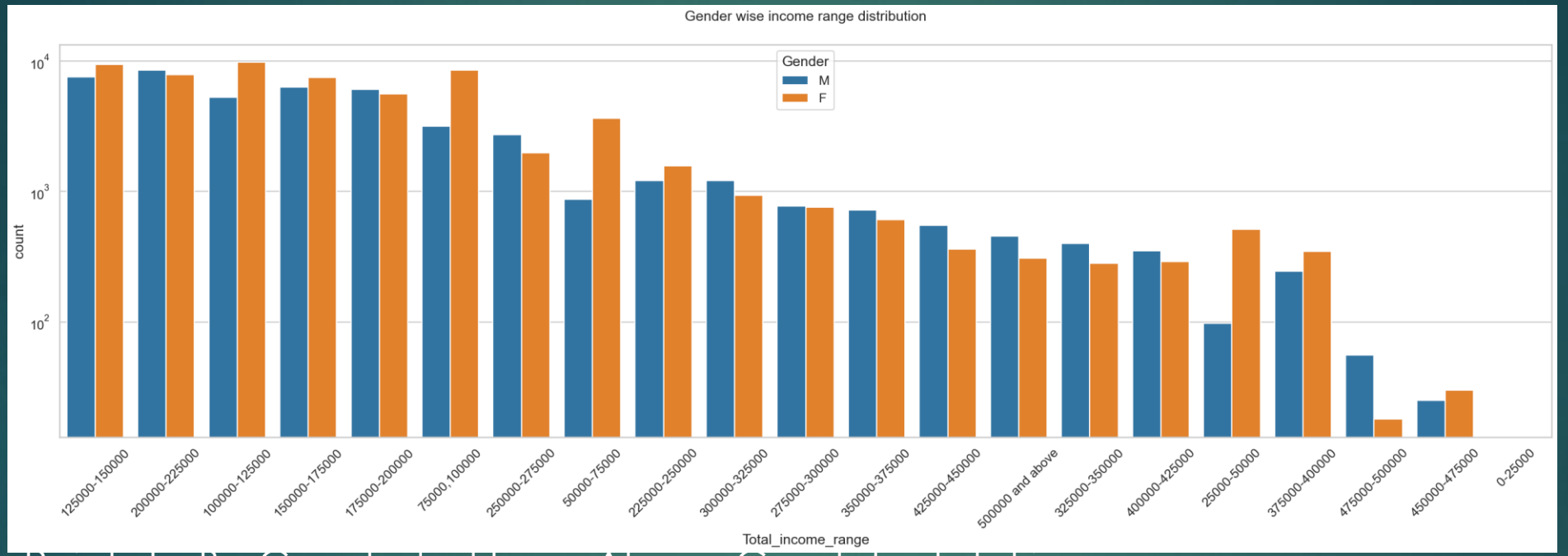


Points To Be Concluded From Graph for Loan Type Distribution:

- 'Business entity Type 3' , 'Self employed', 'Other' , 'Medicine' and 'Government' mostly applied for the credit.
- Industry type 8,type 6, type 10, religion and trade type 5, type 4 lowest applied for the credits.
- Females are applied more for credits.

Categorical Univariate analysis for target 1

Income Range Distribution:



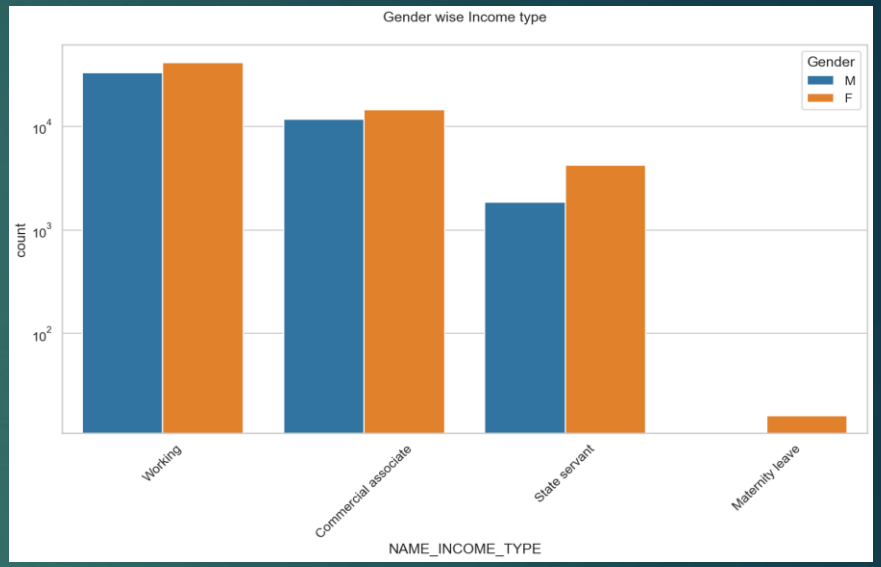
Points To Be Concluded From Above Graph for total income range:

- Females income are higher than male.
- 75,000 to 1,25,000 having more no. of female credit.
- Income range from 75,000 to 2,25,000 having higher no. of credits.
- 4,00,000 and more are having low no. of credits

Income Type Distribution

Points To Be Concluded From Graph for Income Type Distribution:

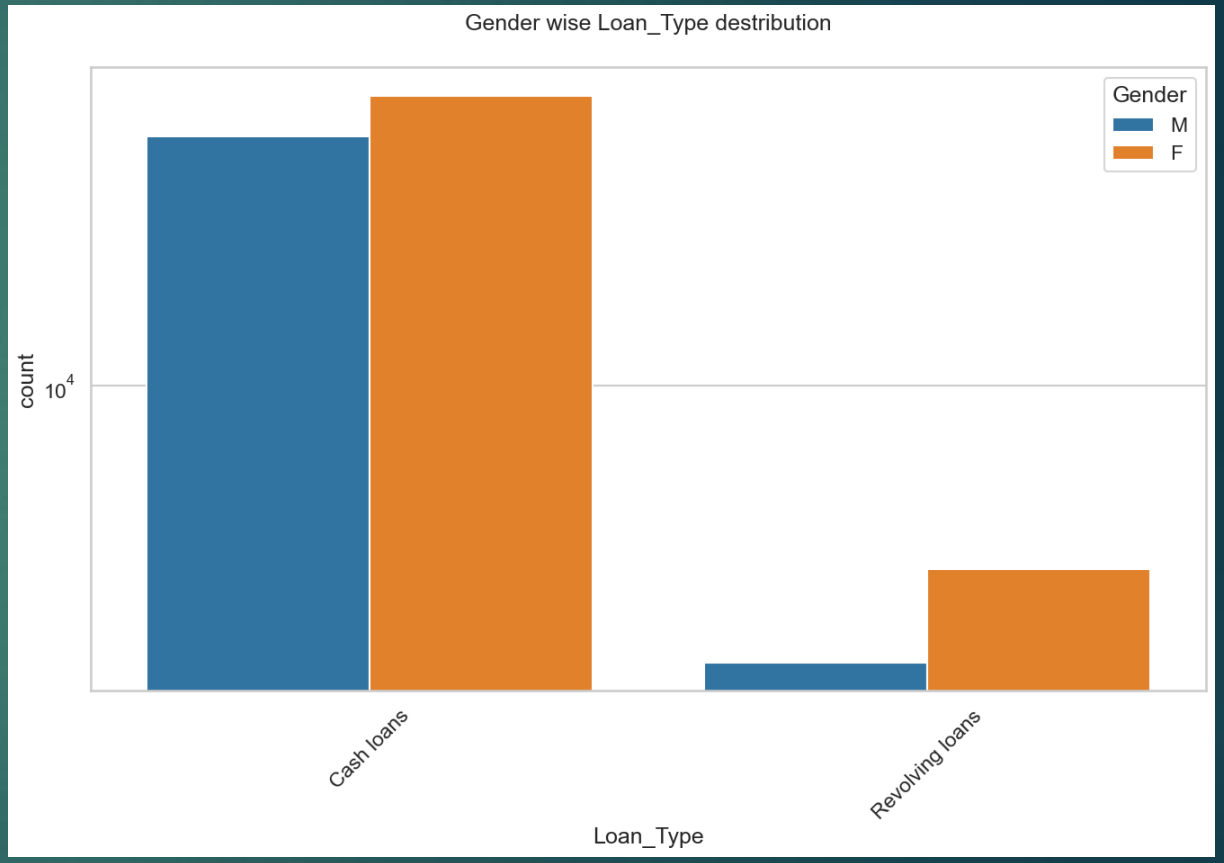
- ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than others.
- “Maternity Leave” has Low number of credit is.
- “Student” “pensioner” and “businessman” having no credits which means they never done the late payments.



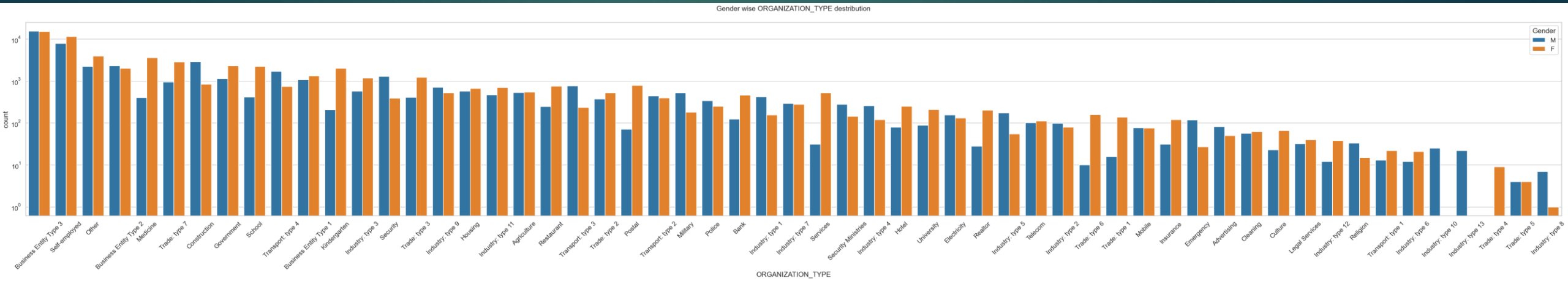
Income Type Distribution:

Points To Be Concluded From Graph for Loan Type Distribution:

- Cash loan is higher than revolving loan.
- Female takes more cash loan than male.



Organization Type Distribution:

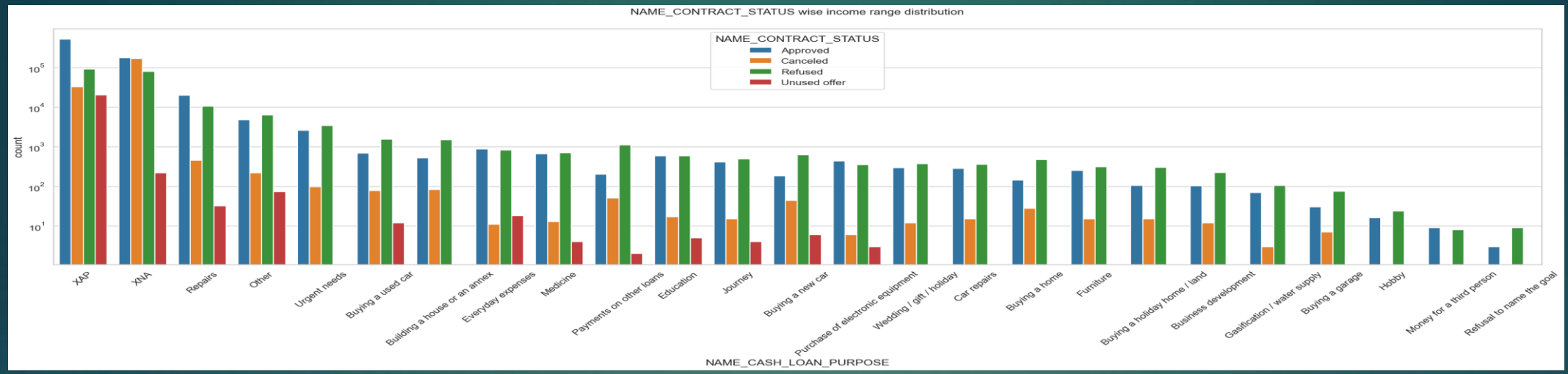


Points To Be Concluded From Graph for Loan Type Distribution:

- ‘Business entity Type 3’ having equal no. of credits in male and female.
- Other aspects are same as on time payment.

Univariate analysis of merged data

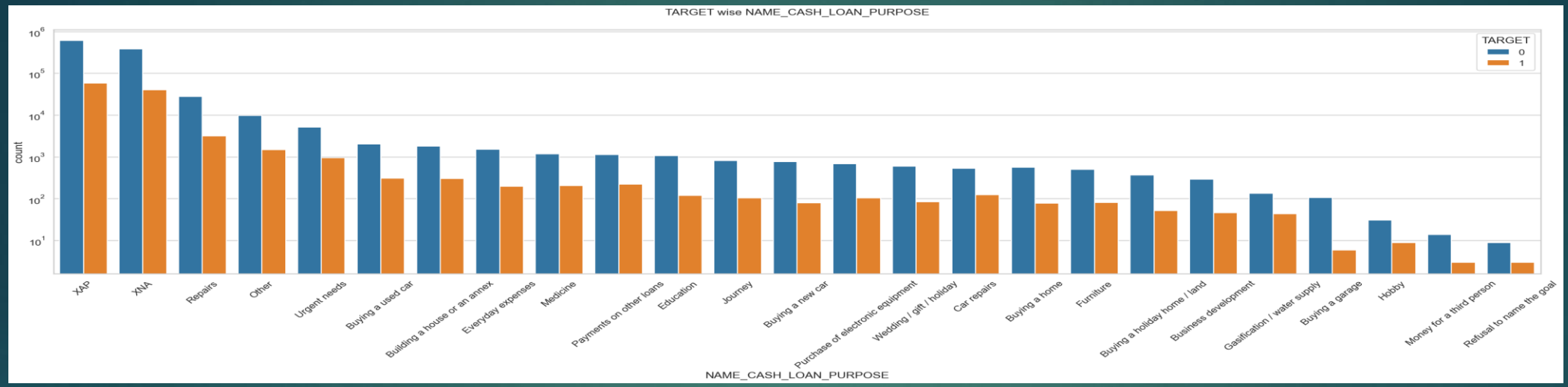
Distribution of contract status with purpose



Points To Be Concluded From Graph for Loan Type Distribution:

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

Distribution of contract status with purpose

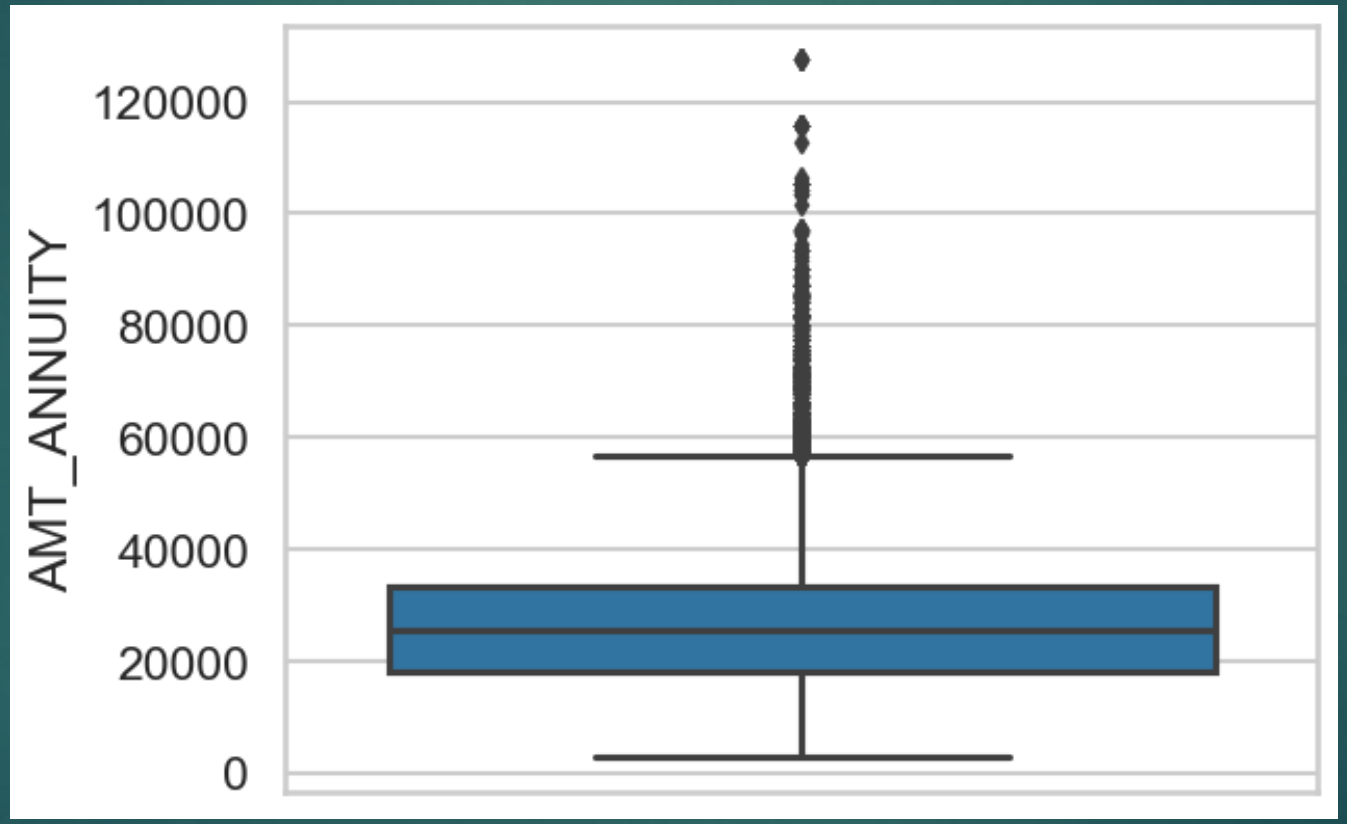


Points To Be Concluded From Graph for Loan Type Distribution:

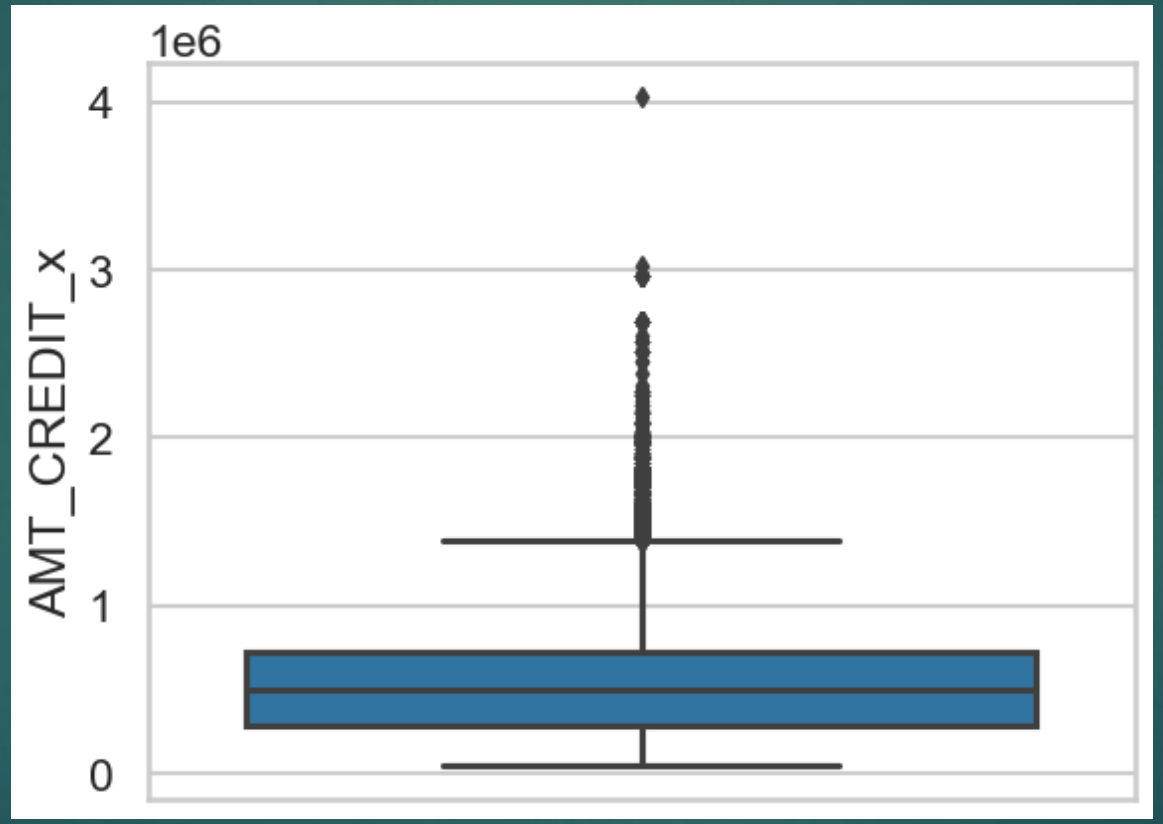
- Xap xna is unavailable category.
- Loan purposes with 'Repairs' are facing more difficulties in payment on time.
- 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education' are few places where loan payment is significant higher than facing difficulties.

Categorical Univariate analysis for variables target 0

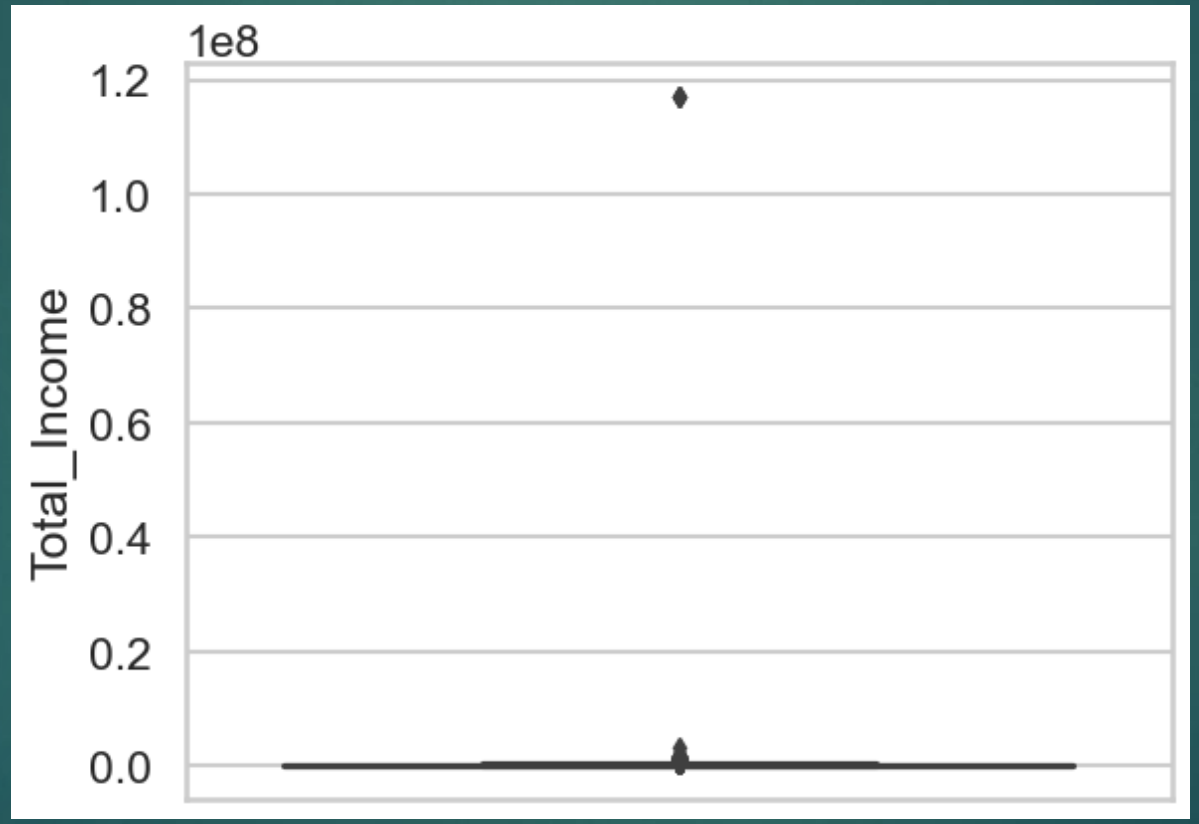
Boxplot for Annuity



Boxplot for Credit

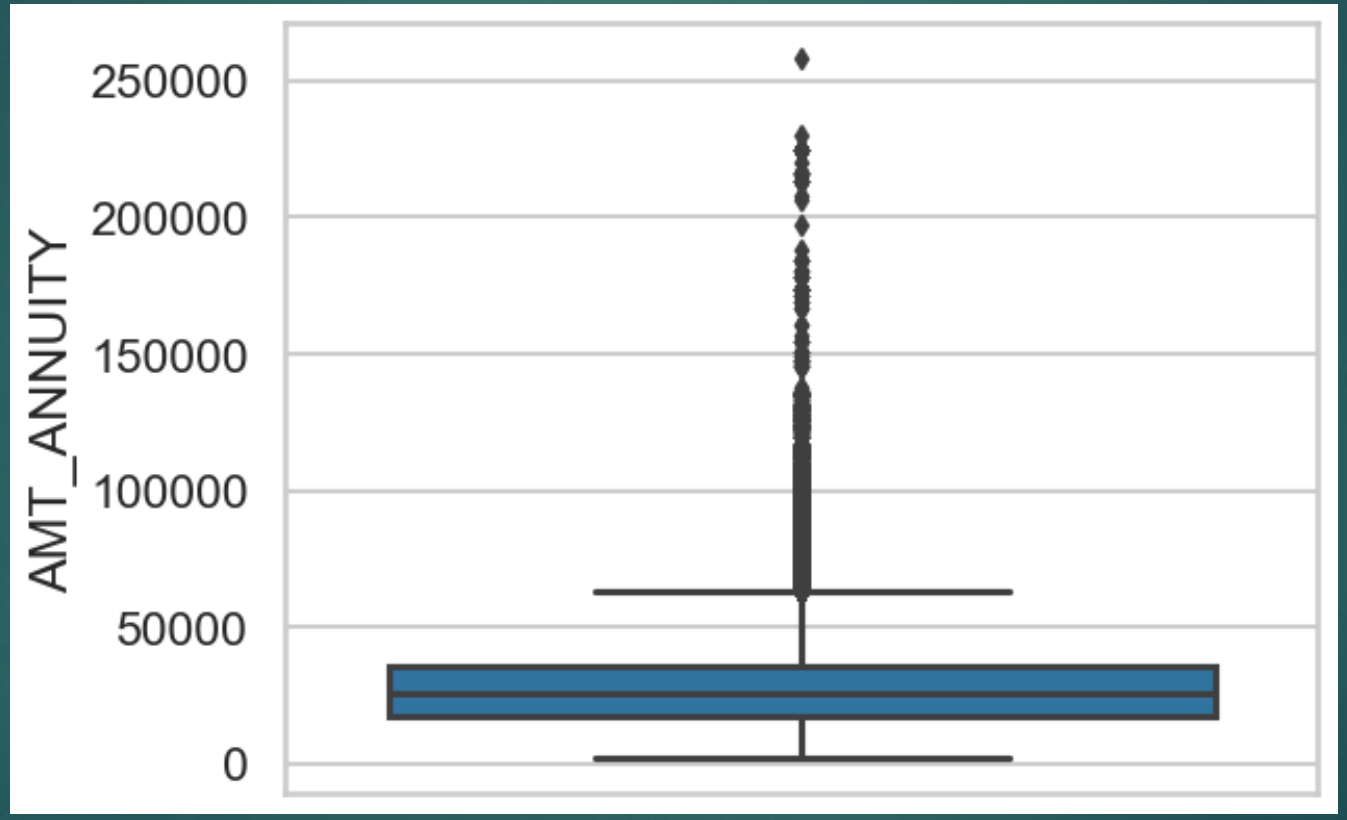


Boxplot for Income

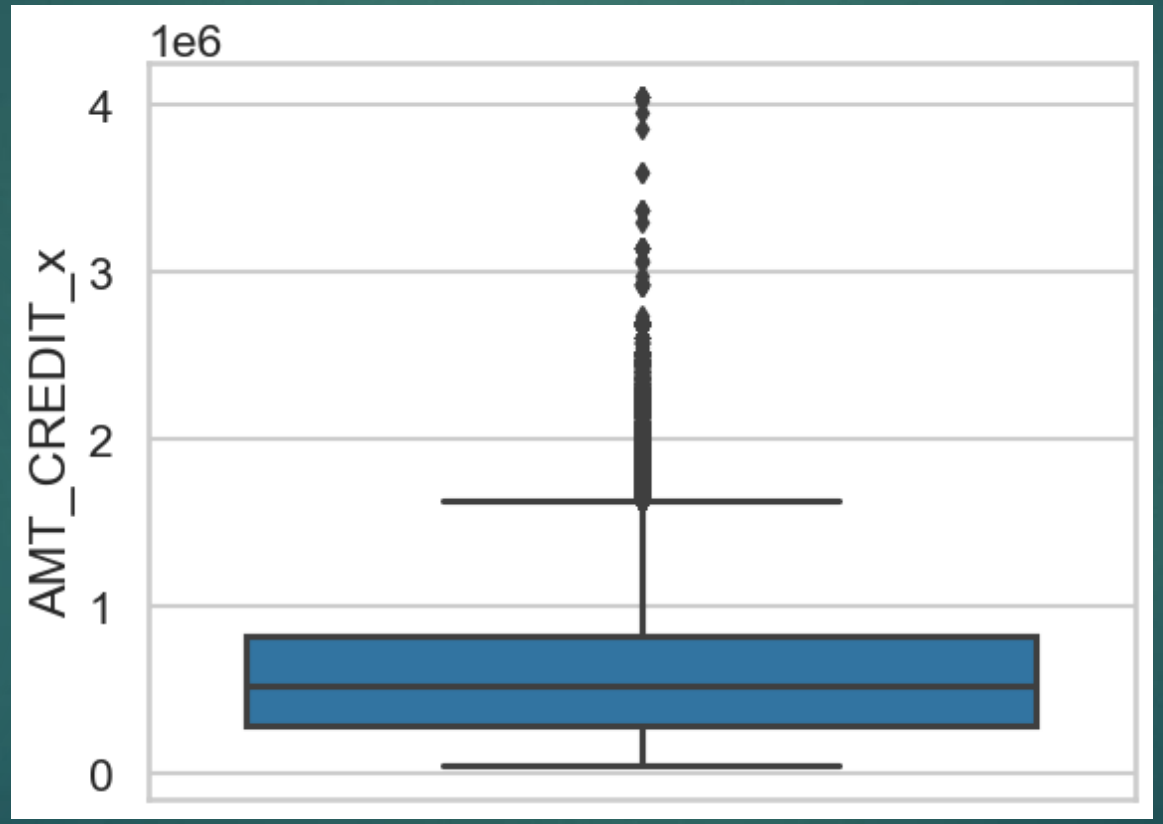


Categorical Univariate analysis for variables target 1

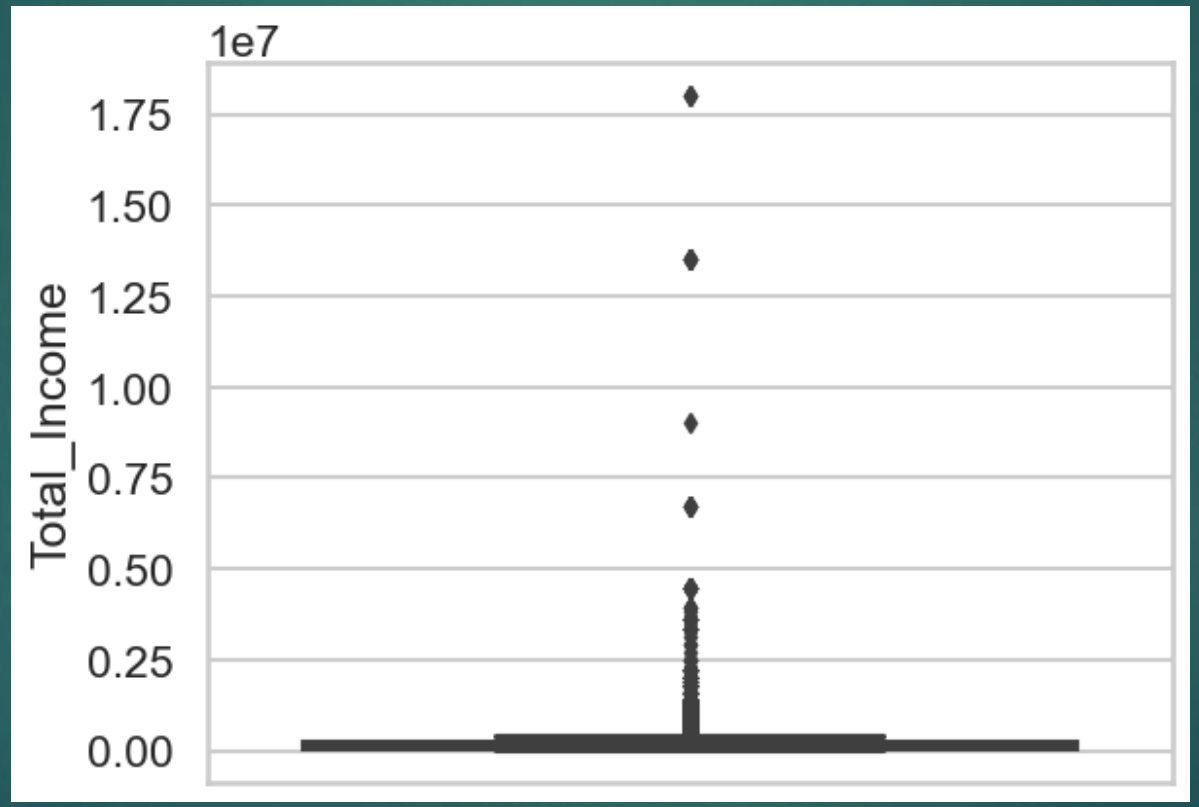
Boxplot for Annuity



Boxplot for Credit



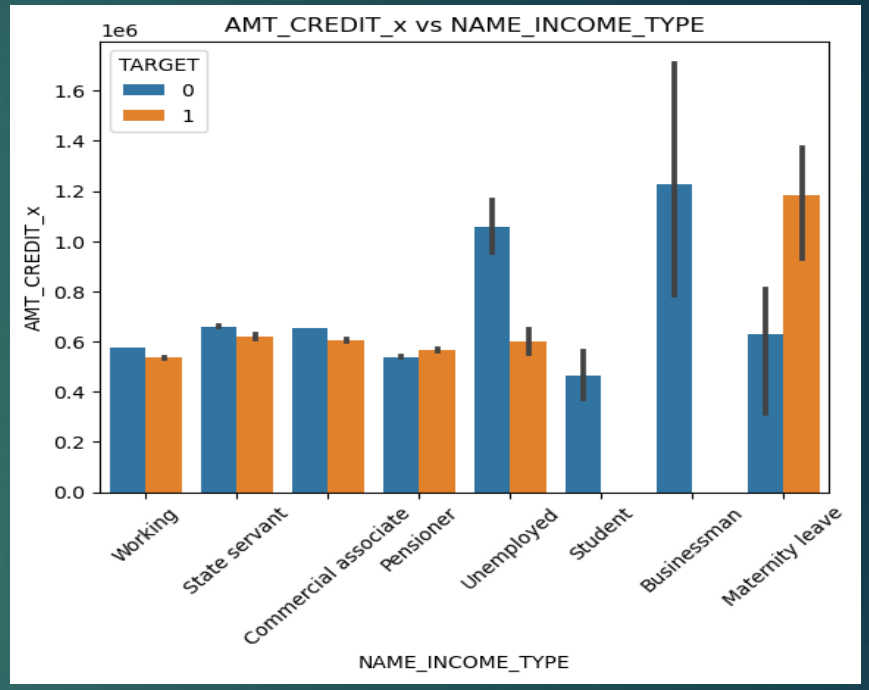
Boxplot for Income



There are outliers in the box plot of target 1 and target 2

Performing bivariate analysis

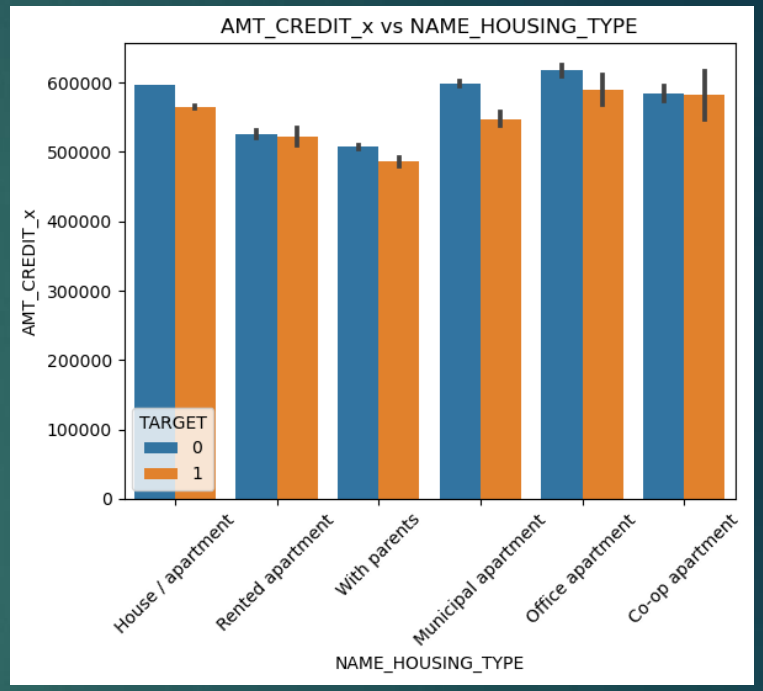
AMT Credit vs Income Type



Points To Be Concluded From Graph:

- in case of income type bank should avoid to give loan to maternity leave as there history is not good and focus on student and businessman for giving loan.

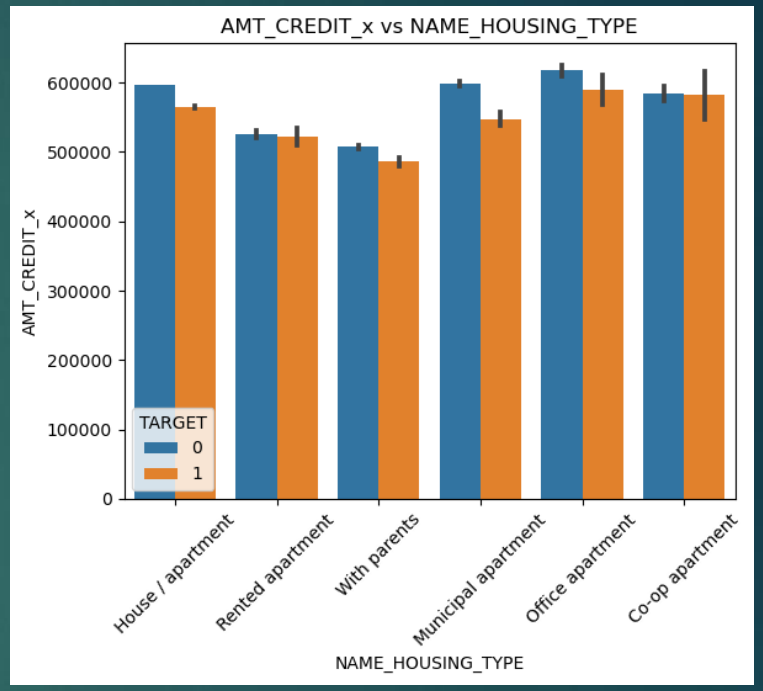
AMT Credit vs Housing Type



Points To Be Concluded From Graph for Loan Type Distribution:

- house/apartment rented apartment muncipal apartment has higher credit of target 0 bank can focus mostly on these categories. in case of co-op apartment there is a high chance of loan defaulting so the bank can ignore them to give loan.

AMT Credit vs Education Type

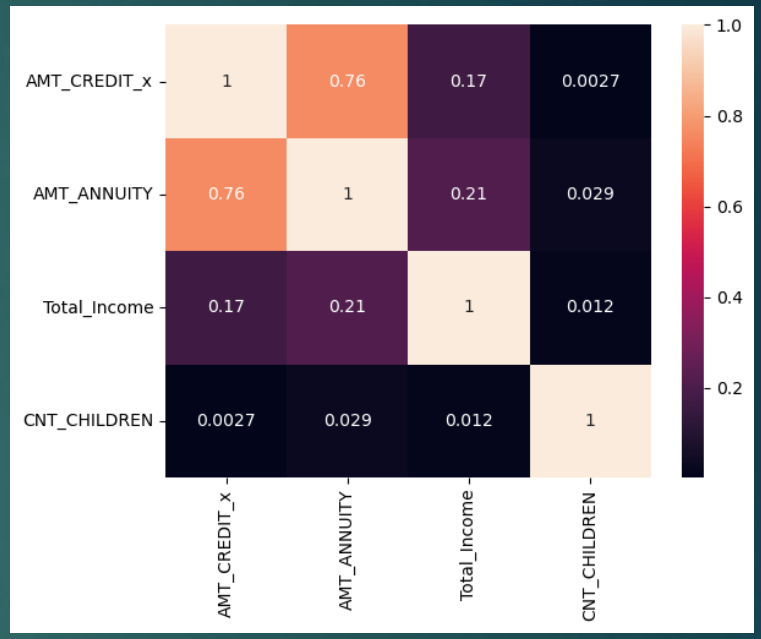


Points To Be Concluded From Graph for Loan Type Distribution:

- in terms of education qualification with Higher education has mostly chance of paying the loan so bank can focus on that and avoid lower secondary education people.

Performing Multivariate analysis

Correlation in variables



Points To Be Concluded From Graph:

- annuity is having correlation with amt credit and vice versa children is having lowest correlation with amt credit and vice versa.

Conclusion:

- So final conclusion is student or businessman with higher education living in house/apartment or with parents are more trustable to give loan and pensioner and maternity leave persons bank should have to avoid for giving loan.
- 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.