CREDIT- EDA CASE STUDY

NAME: *Sumit Kumar*

Problem Statement

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.

Business Understanding

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.

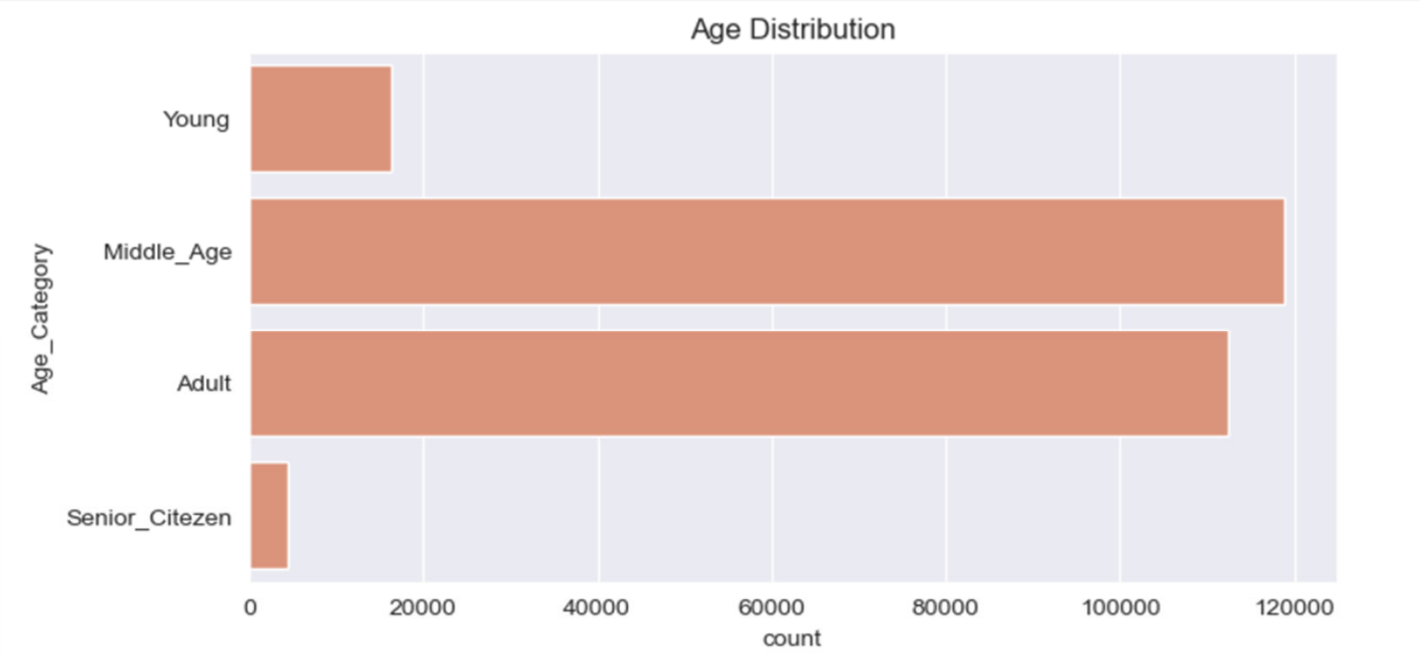
To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

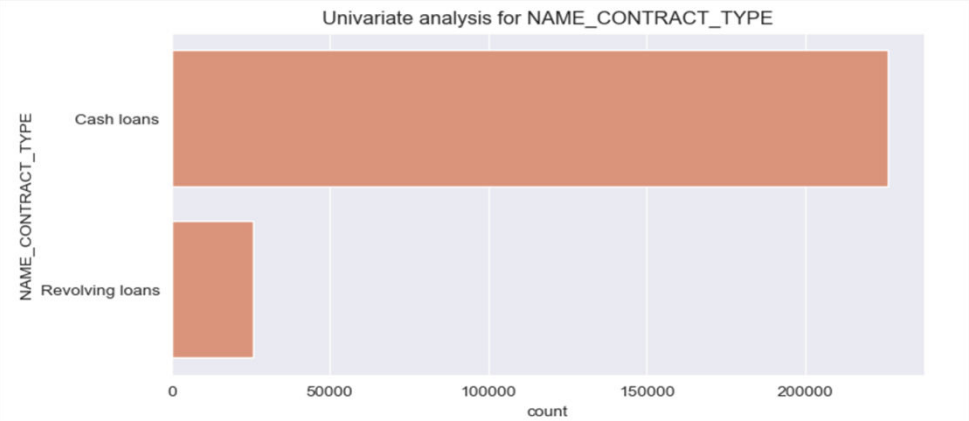
Imputing Missing Values

- I have dropped the columns which are having null values more then 20% in it because imputing those values will decrease efficiency for us to find better trends or correlations .
- I have imputed values as medians in loan annuity columns because it has outliers in it.
- In categorical columns I have imputed missing values with mode as they are maximum. For ex XNA Values replaced with “F” and dropped rows in occupation type column(because of huge number).
- Some columns are not in use for this case-study so I simply dropped those columns. For ex: Flag columns.
- Some columns have very few missing values so Imputing those with Mean, Median .
- AMT_REQ_CREDIT_BUREAU columns I took the median because number of enquires to credit bureau are 0 in most cases. So that we can impute those with 0.

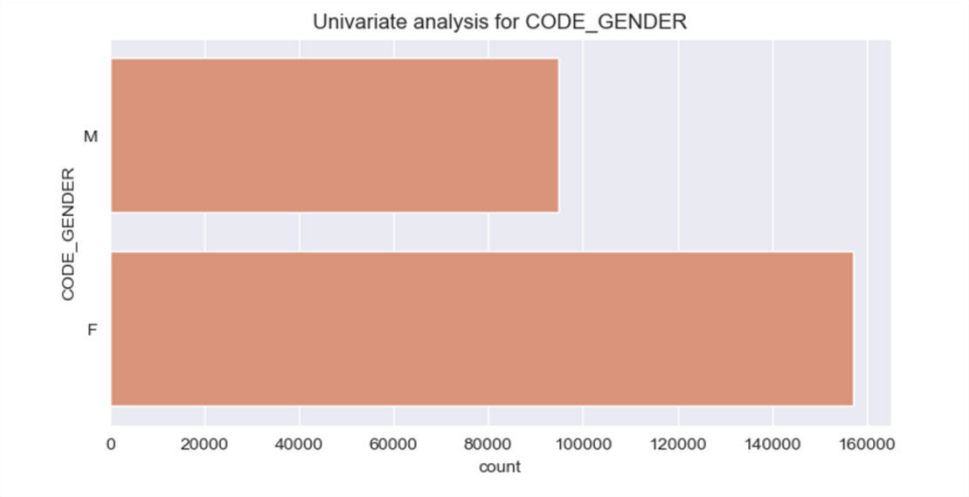
Univariate Analysis for Categorical variables

1. There are more number of middle aged and adult applicants from Age_Category column.

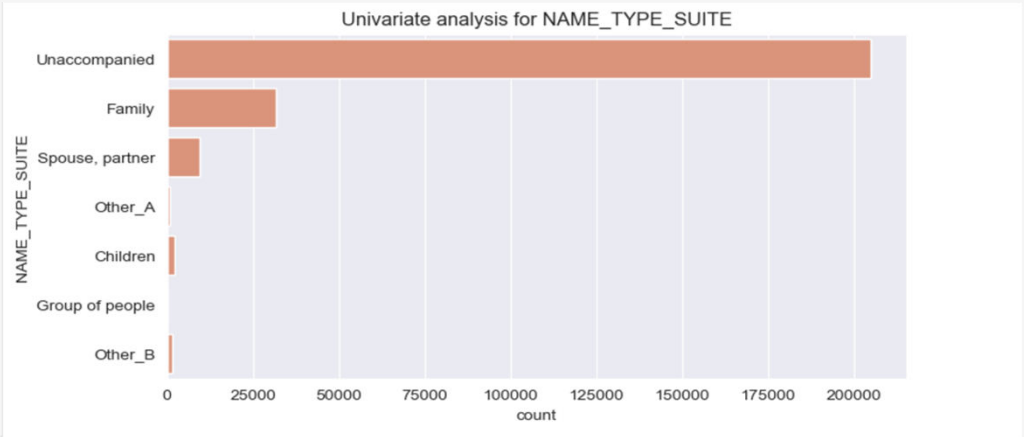


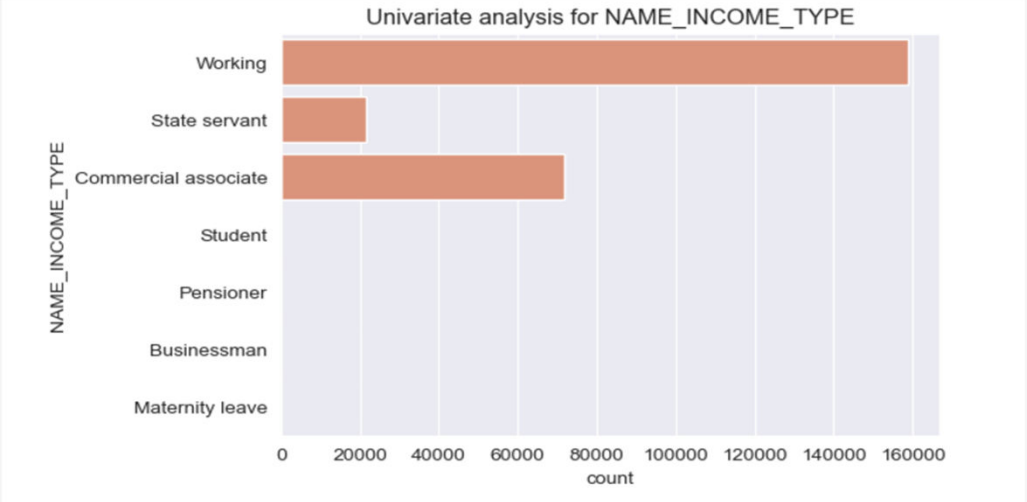


1. There are maximum number of aplicants who opted cash loans.



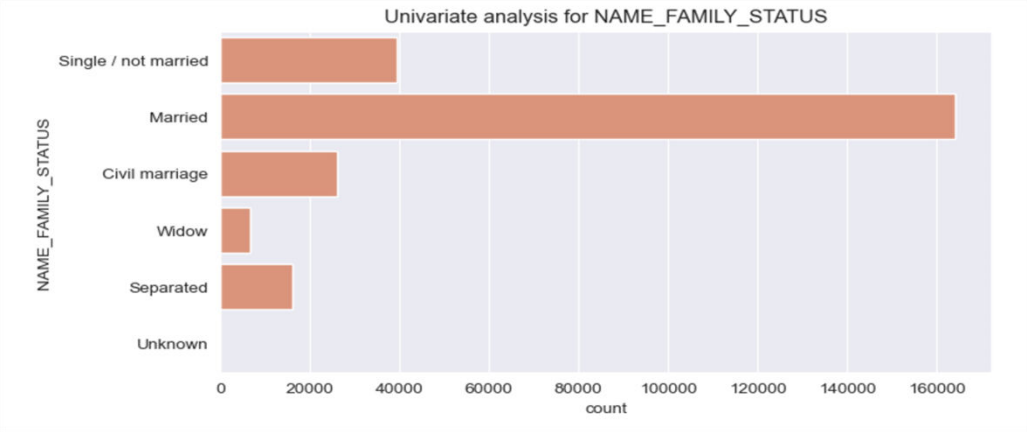
2. There are more number of female applicants then male applicants



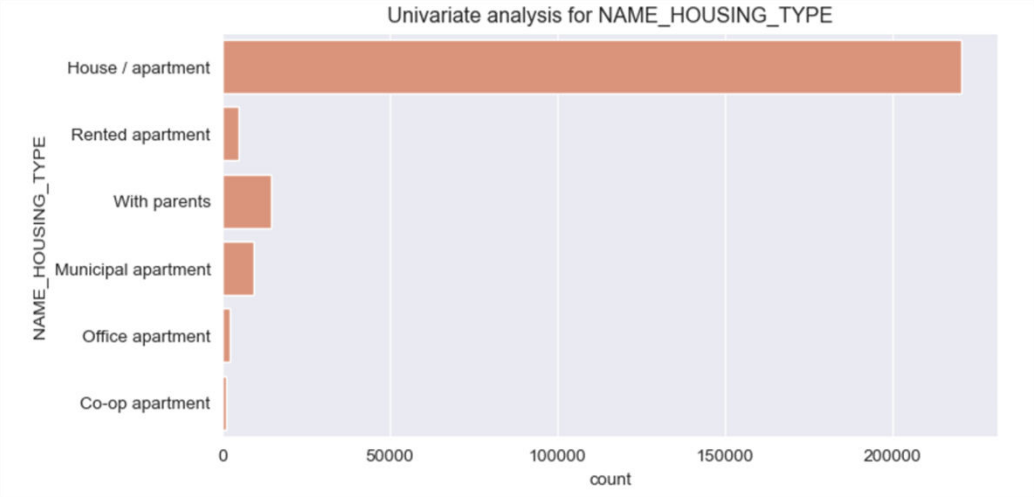


1. Most of the applicants who applied loan were accompanied at that time of applying.

2. There are more number of working applicants who applied for loan.



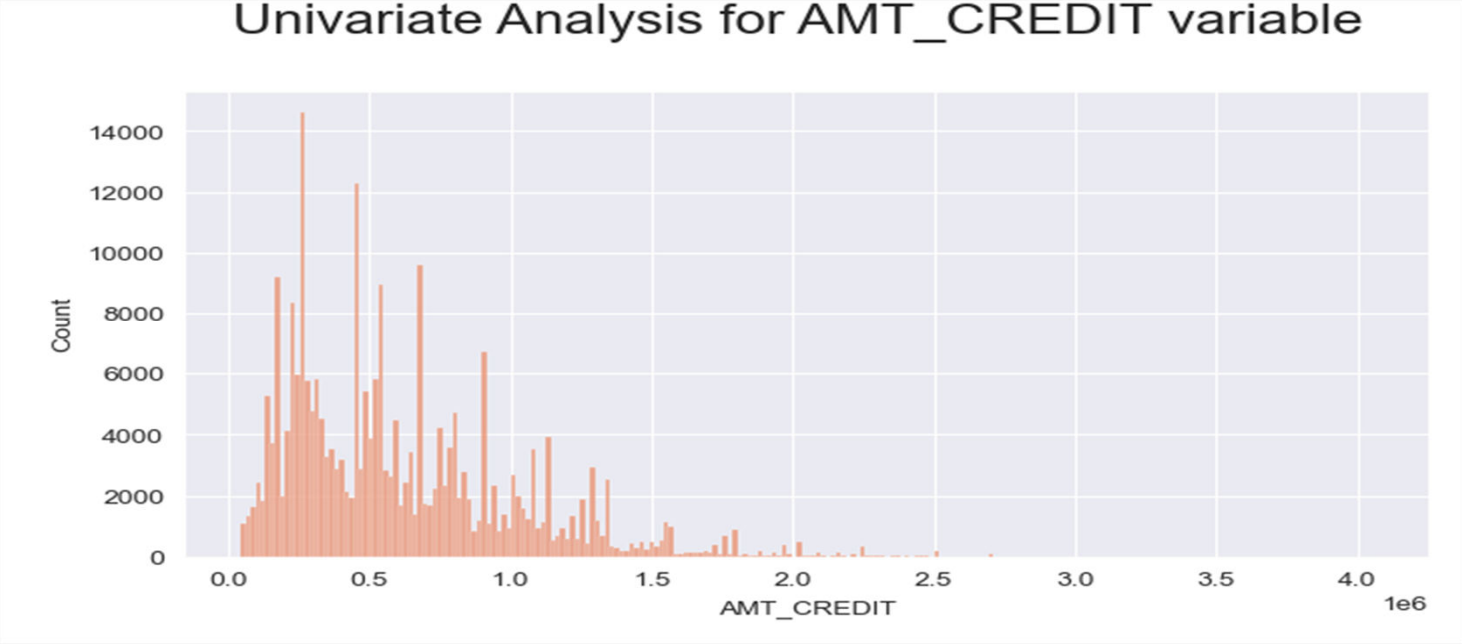
1. Married applicants are more in number who applied for loan.

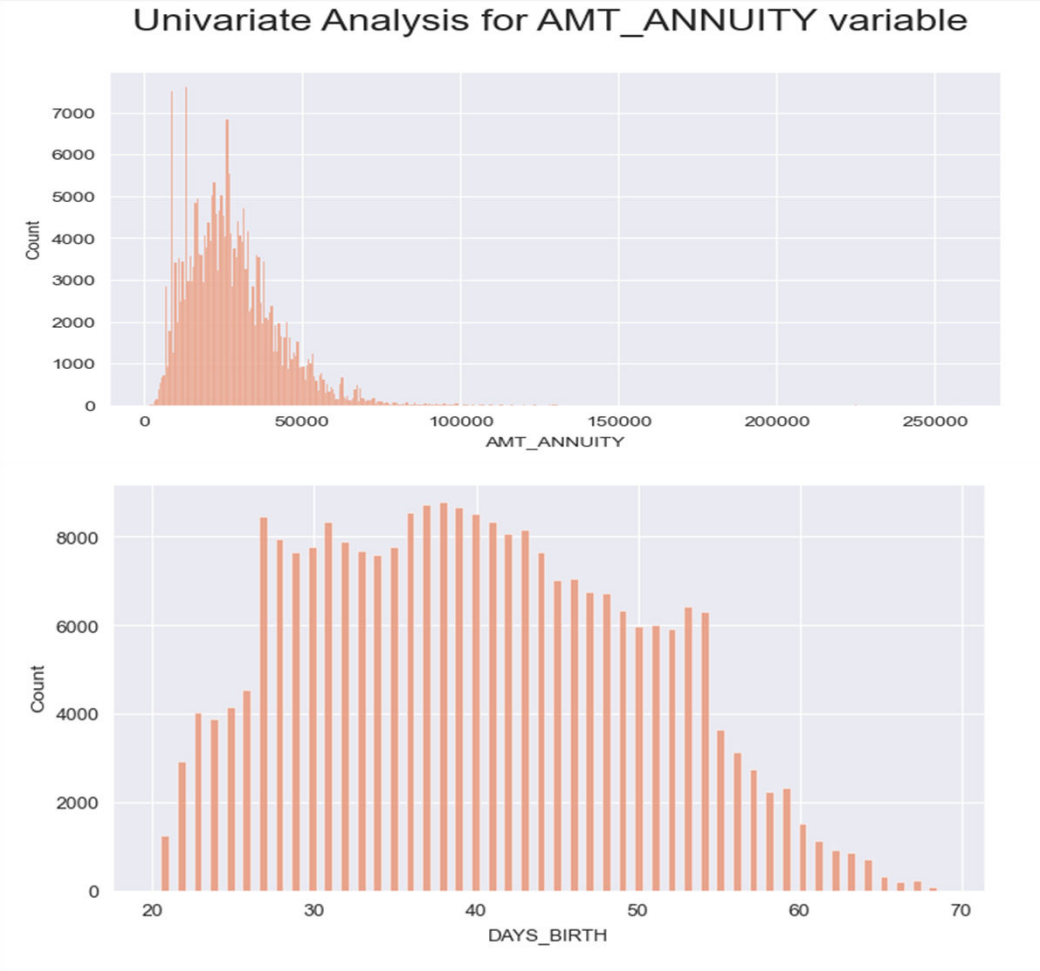


2. Applicants who has house/apartment are maximum in number

Univariate Analysis for Numerical variables

1. Most of the credit amount of the loan of applicants ranges between 0.2 to 0.9.

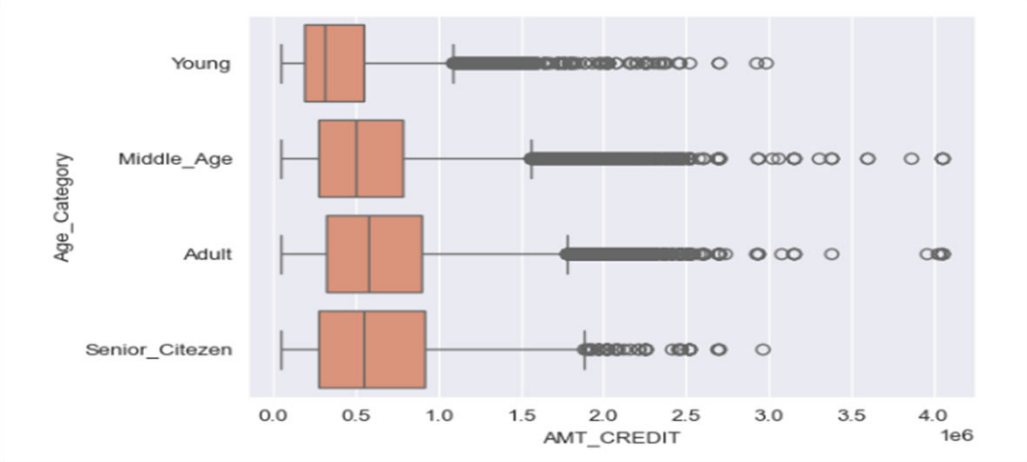




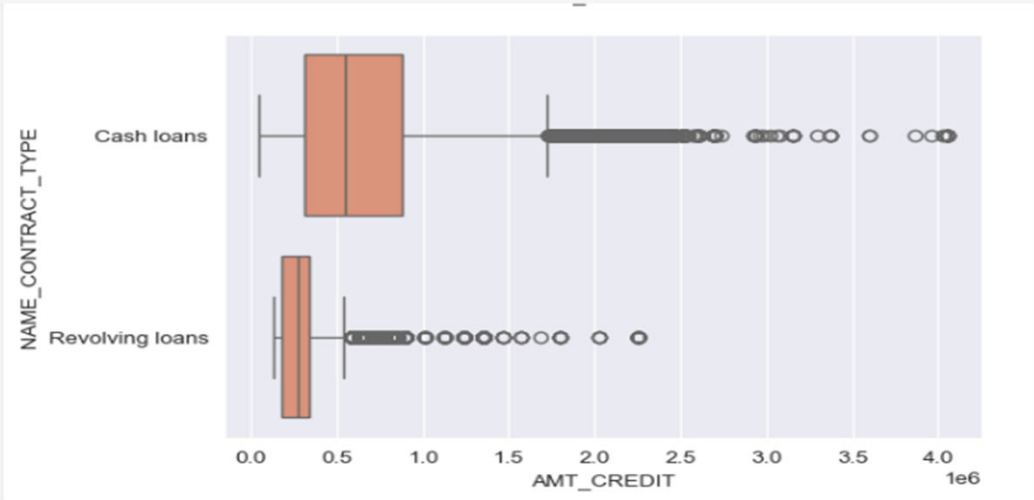
1. Loan annuity ranges 15k to 45k who applied for loan.

2. Maximum no of Applicants who applied for loan are in the age between 25 to 55.

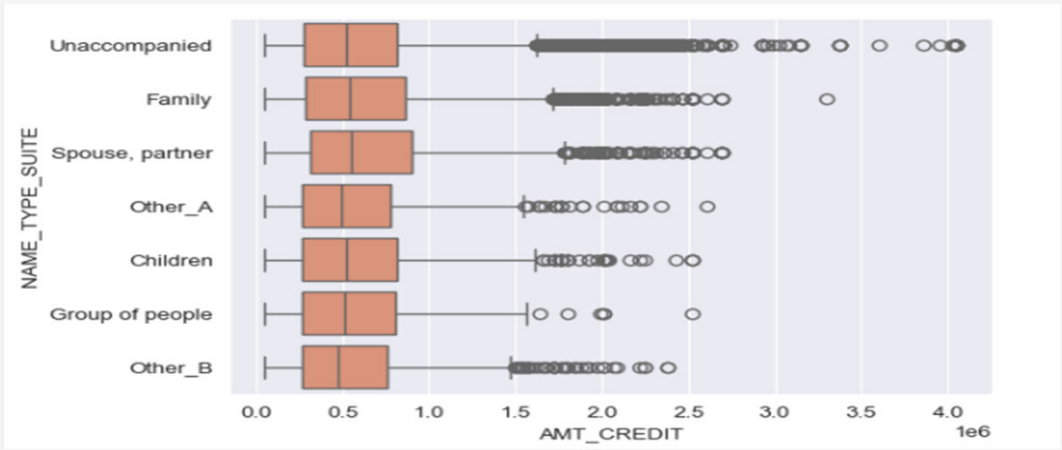
Bivariate Analysis



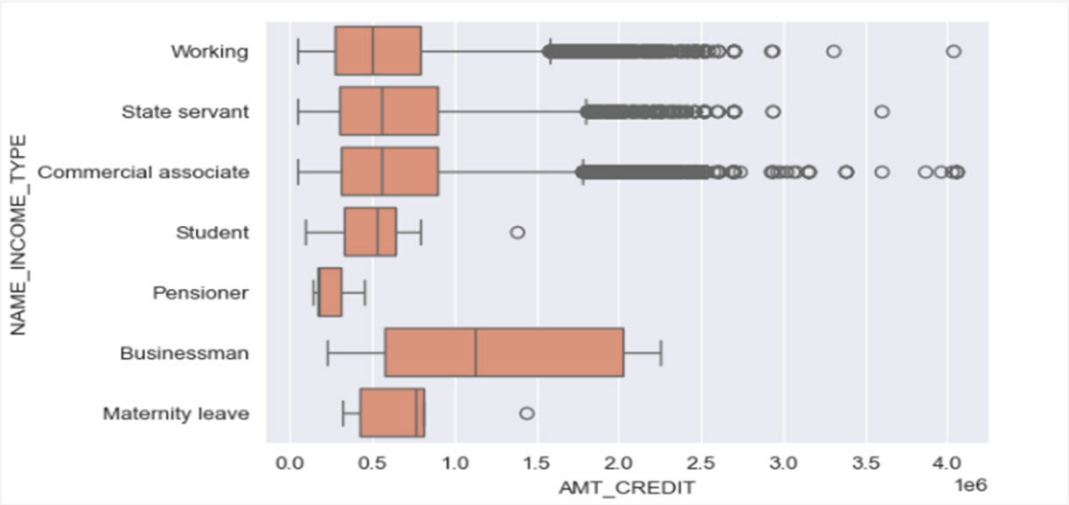
1. Adult and senior citezens have more credit amount of the loan who applied.



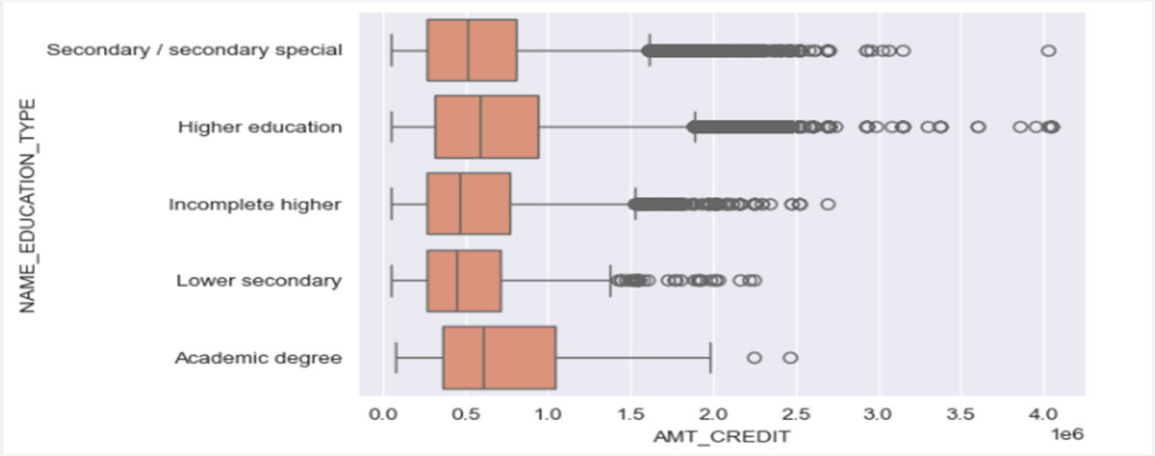
2.Cash loans have more credit amount of the loan for applied applicants



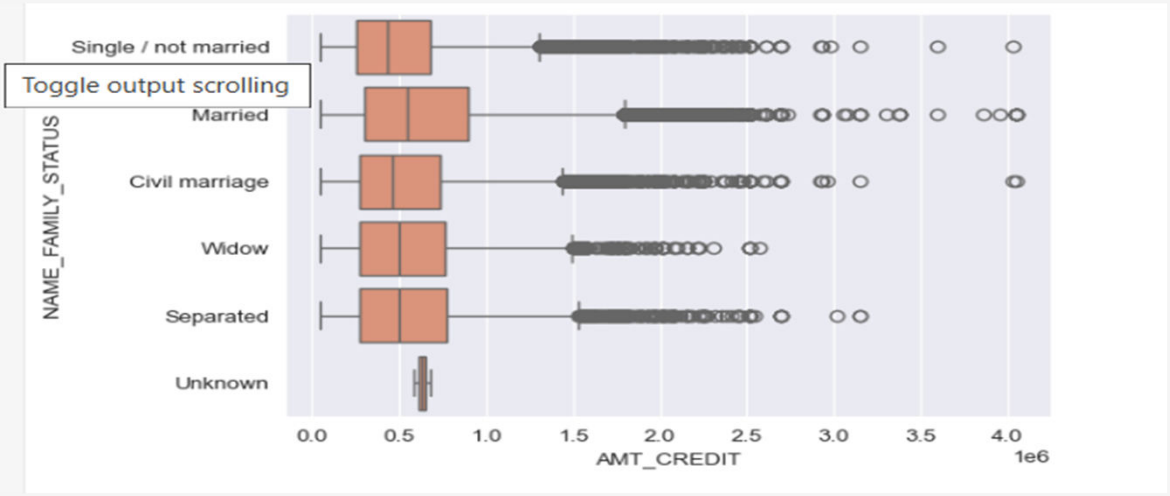
1. Spouse partner and family member who accompanied have more credit amount of the loan.



2. Businessman who applied have more credit amount of the loan.

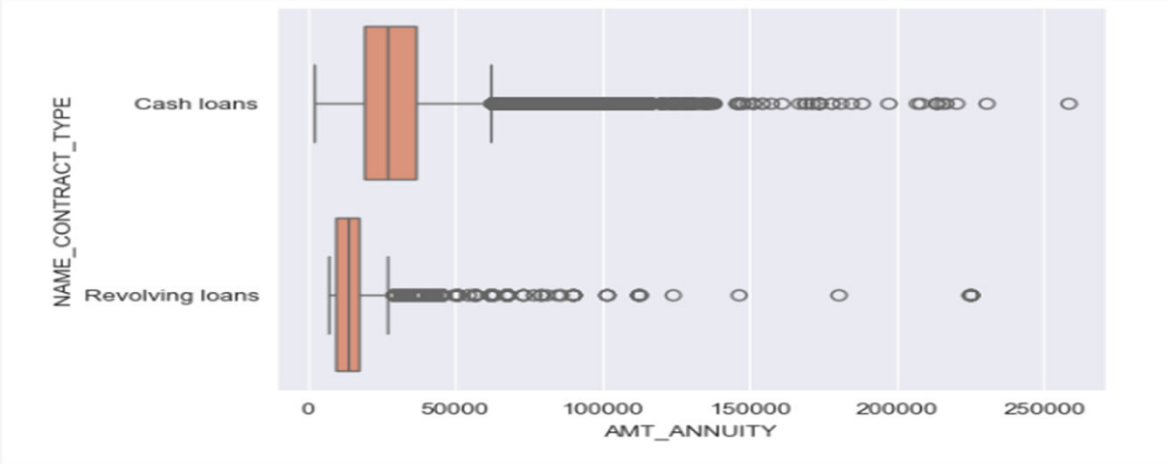


1. Applicants with academic degree and in higher education have more credit amount of the loan

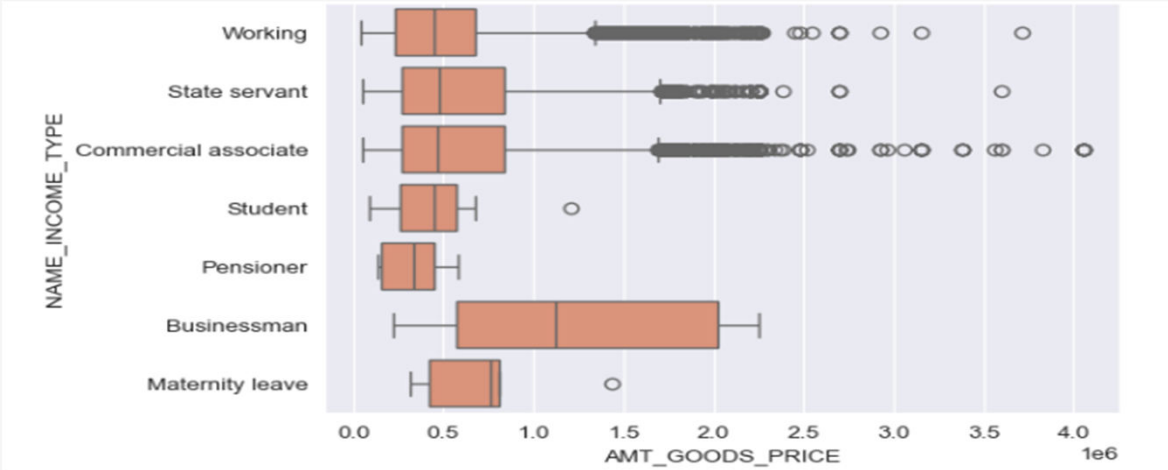


Toggle output scrolling

2. Married Applicants have more credit amount of the loan then others



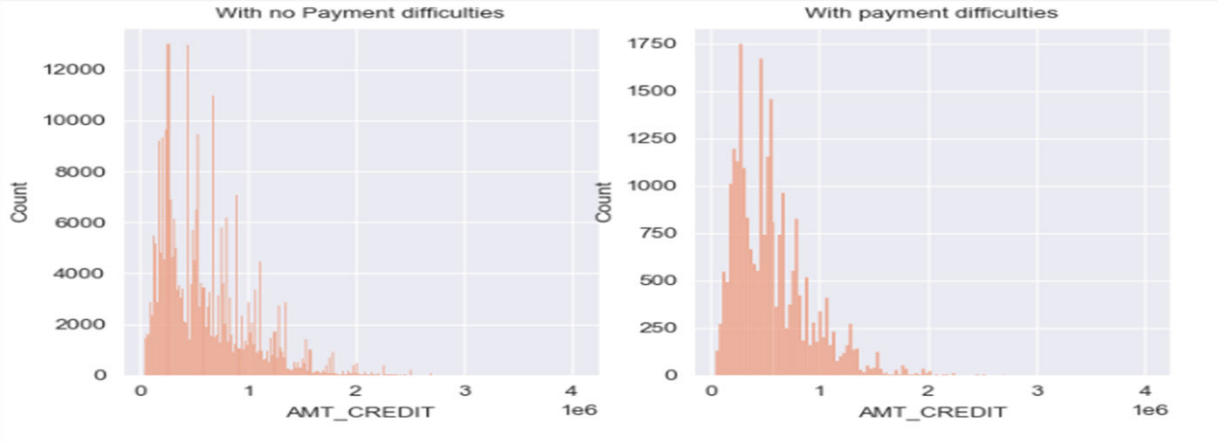
1. Cash loans have high loan annuity.

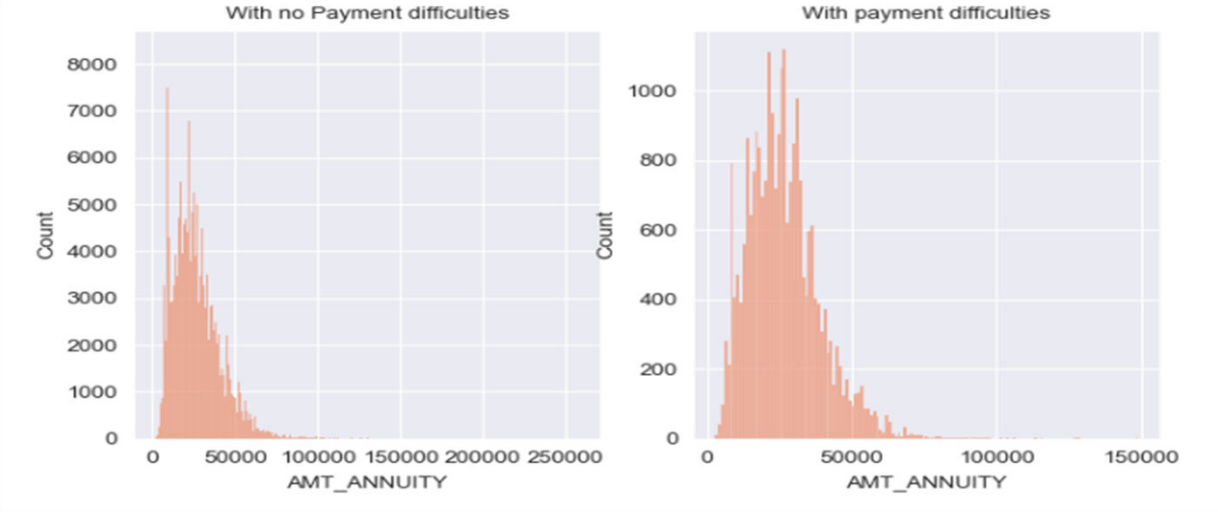


2. Businessman Applicants have higher consumer loans.

1. There is a strong correlation between AMT_GOODS_PRICE and AMT_CREDIT

Numerical Univariate Analysis for target variable 0(with no payment difficulties),1(with payment difficulties)

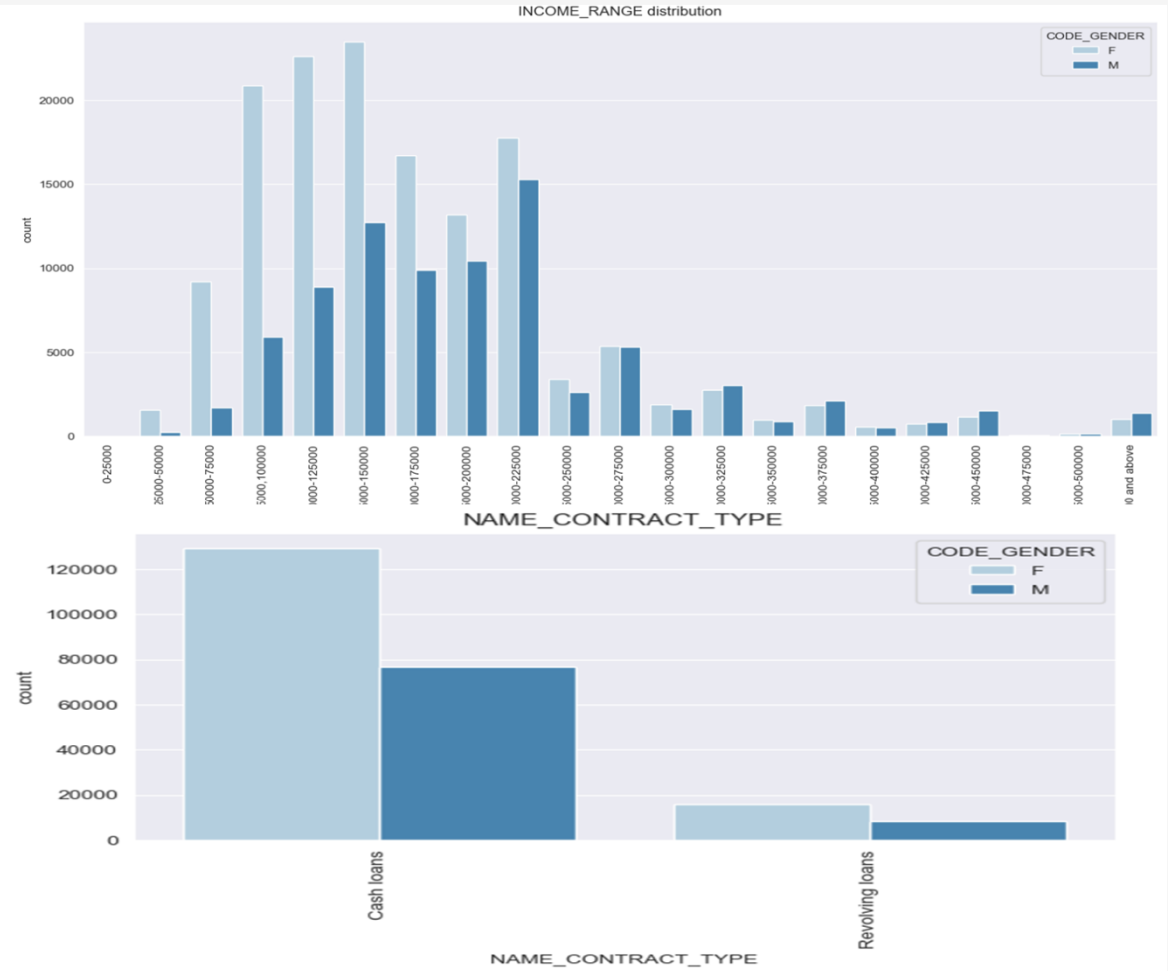




1. Applicants with payment difficulties exist between the 0.2 to 0.9 range of AMT_CREDIT.

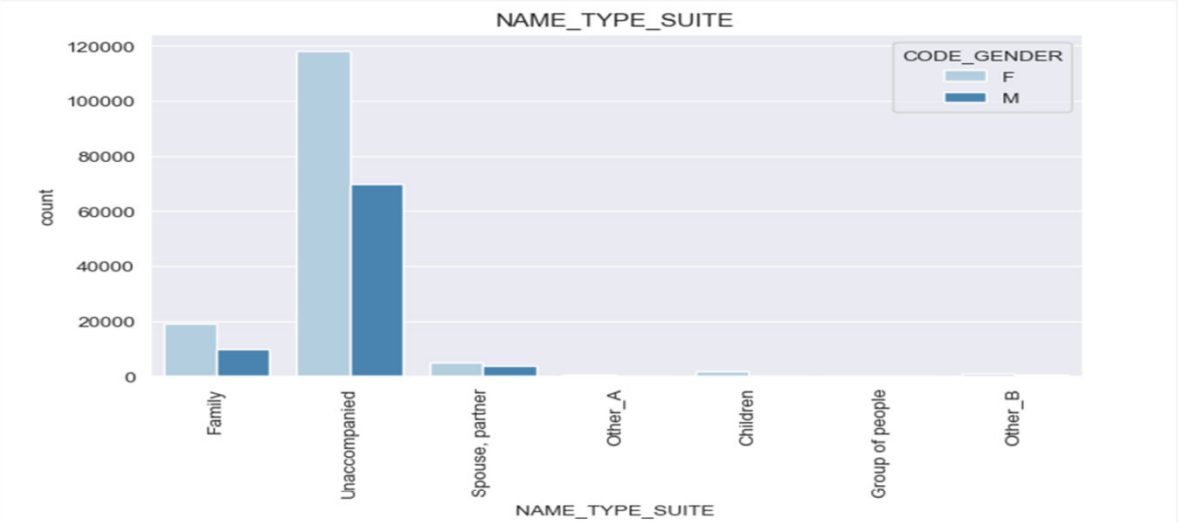
2. Applicants with Loan annuity in range 15k to 45k usually defaults more as compare to other applicants

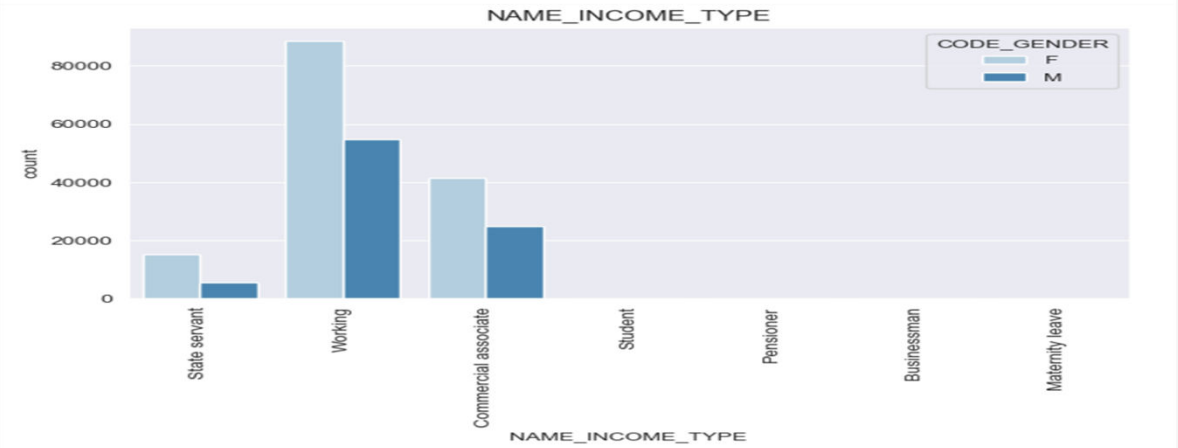
Categorical Univariate Analysis for target variable 0(with no payment difficulties)



1. Very less applicants above 400000 and above
2. Most of the Applicants are female.
3. Most of the applicants have income in range 1 lakh to 2 lakh.

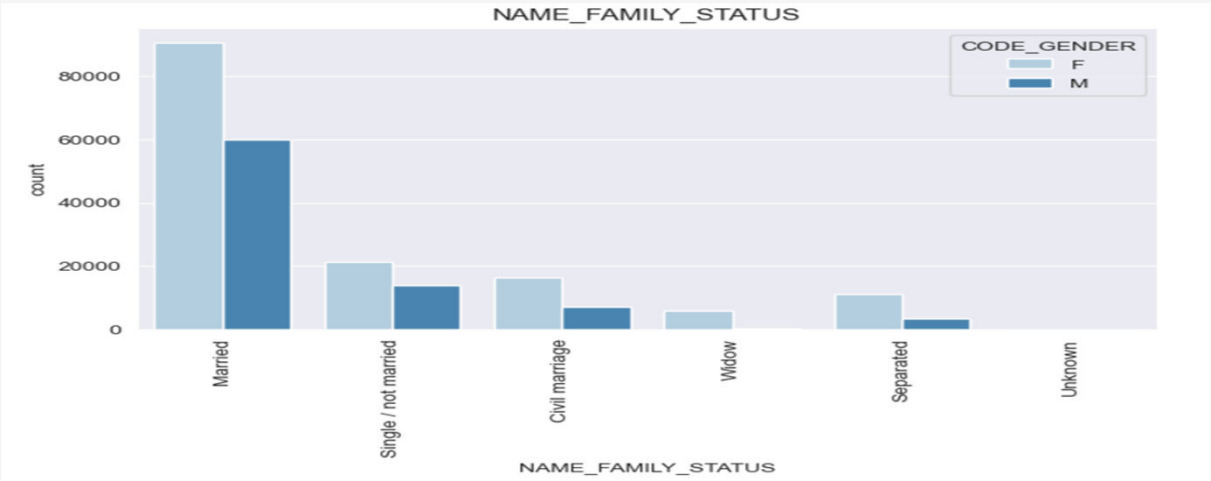
1.Most of the Applicants are having cash loans and females are more in number.

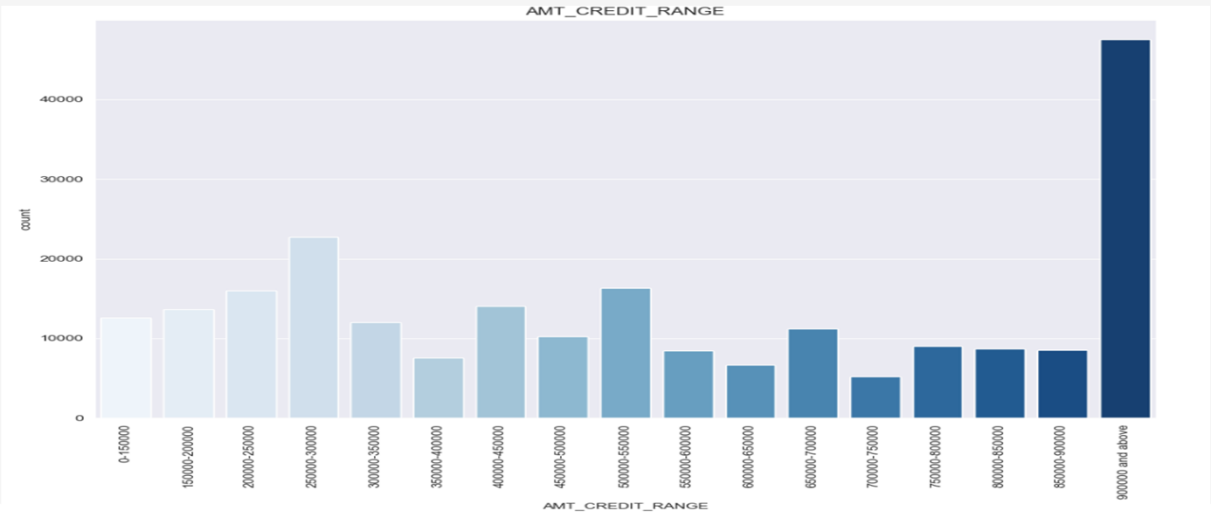




1. Most of the Applicants are unaccompanied at the time of applying for loan and are Females in all categories.

2. Most of the Apllicants are Working and are Females in all categories

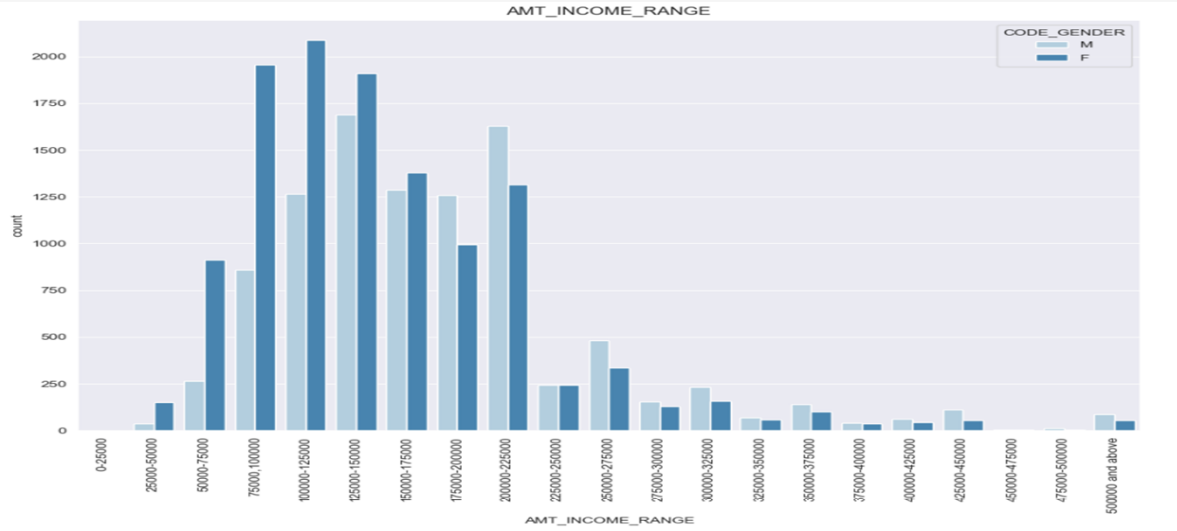


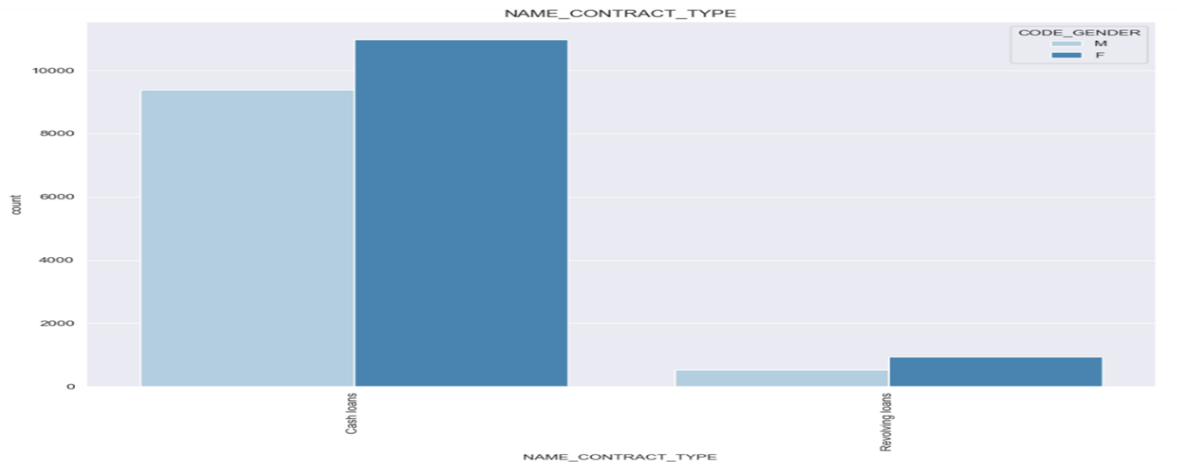


1. Most of the Apllicants are Married and are Females in all categories

2. 2. We can clearly see more number of credits amount of loan are above range 900000 and above

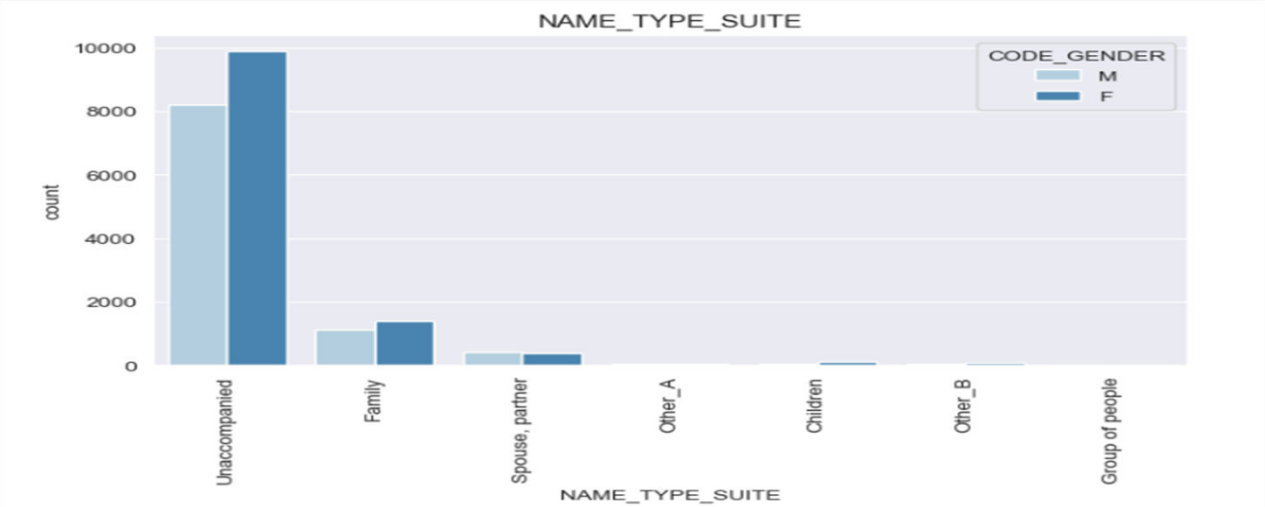
Categorical Univariate Analysis for target variable 1(With payment difficulties)





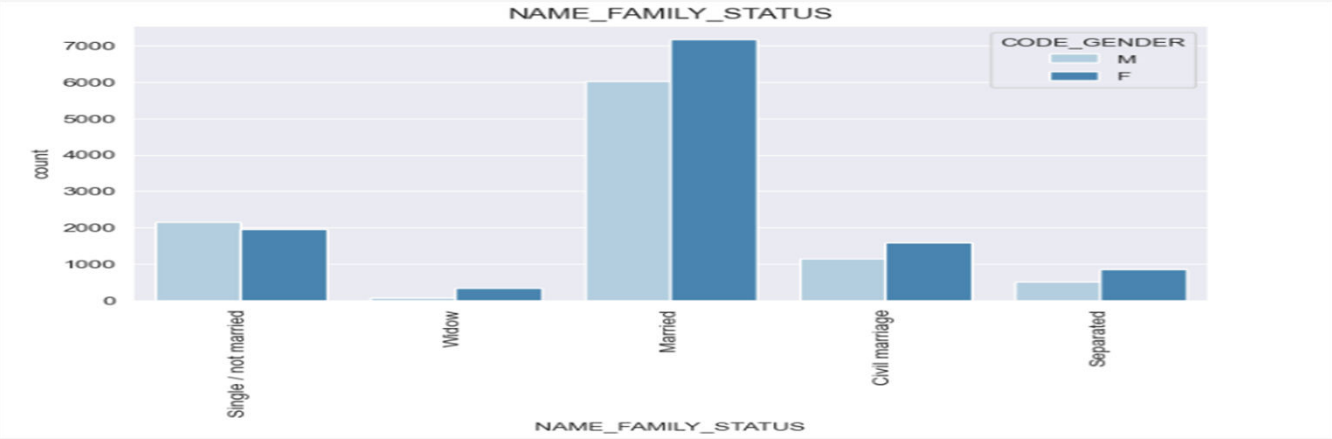
1. Most of the Apllicants who is defaulting has a income range between 50000 to 225000 thousand.
2. Male % increased in all categories who is defaulting.

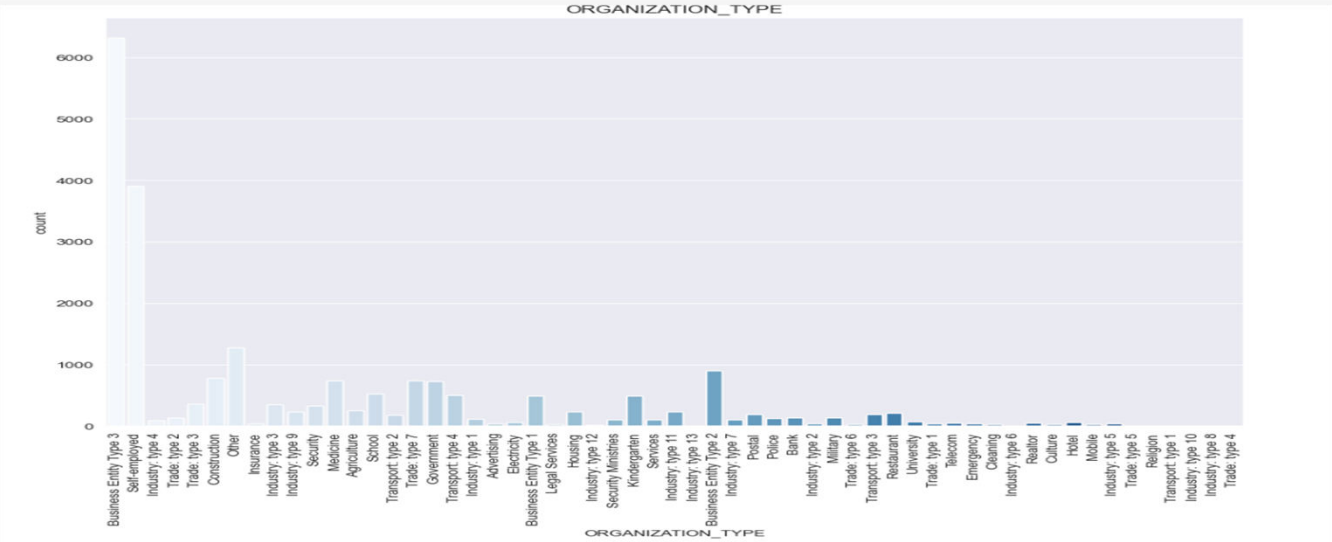
3. Most of the Apllicants are having cash loans Who tends to default.
4. Females are more in number.



1. Most of the Apllicants are Unaccompanied at the time of applying loan Who tends to default.
2. Male % increased in all categories who is defaulting.

3. Most of the Apllicants are Working and Commercial associate Who tends to default.
4. Male % increased in all categories who is defaulting.

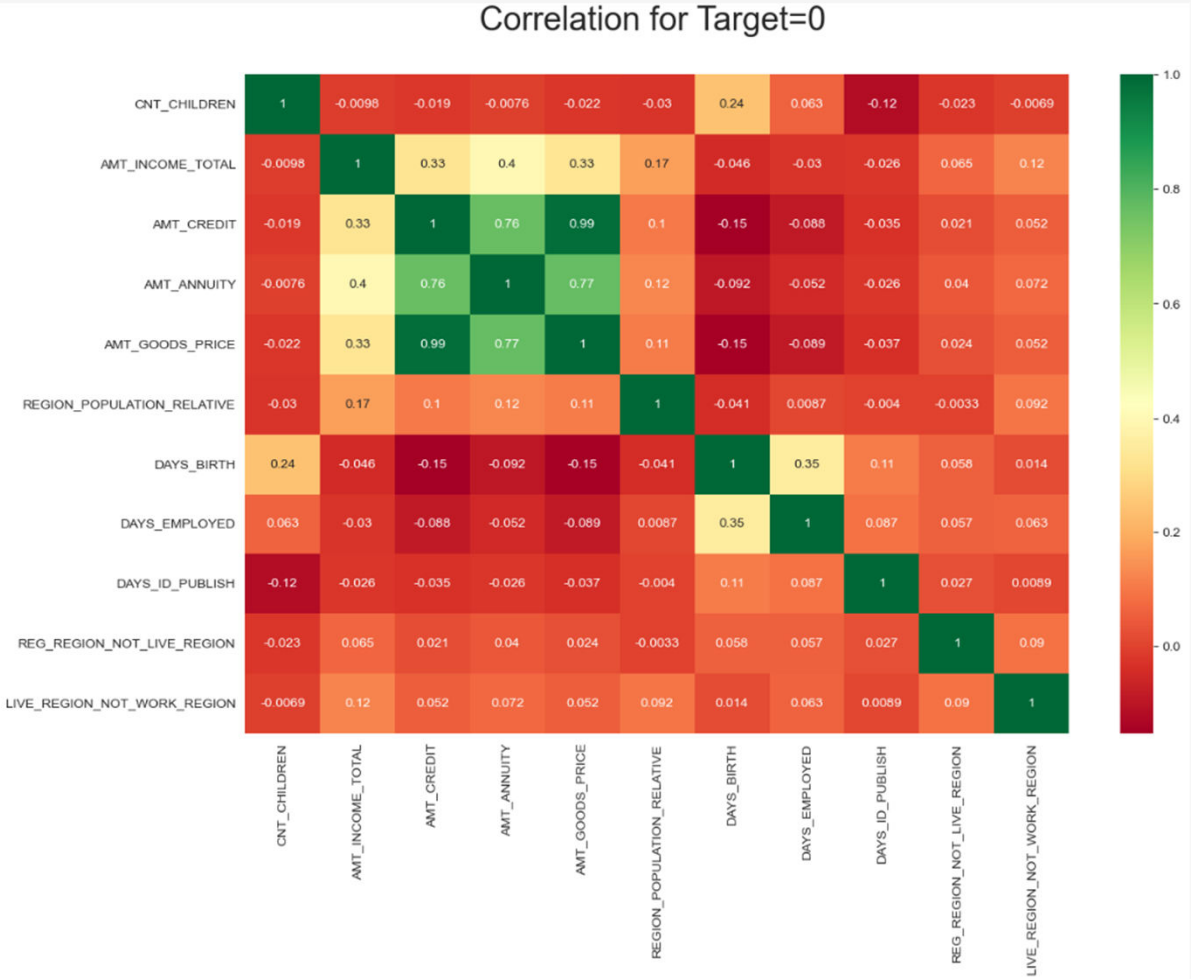




1. Most of the Apllicants are Married Who tends to default.
2. Male % increased in all categories who is defaulting.

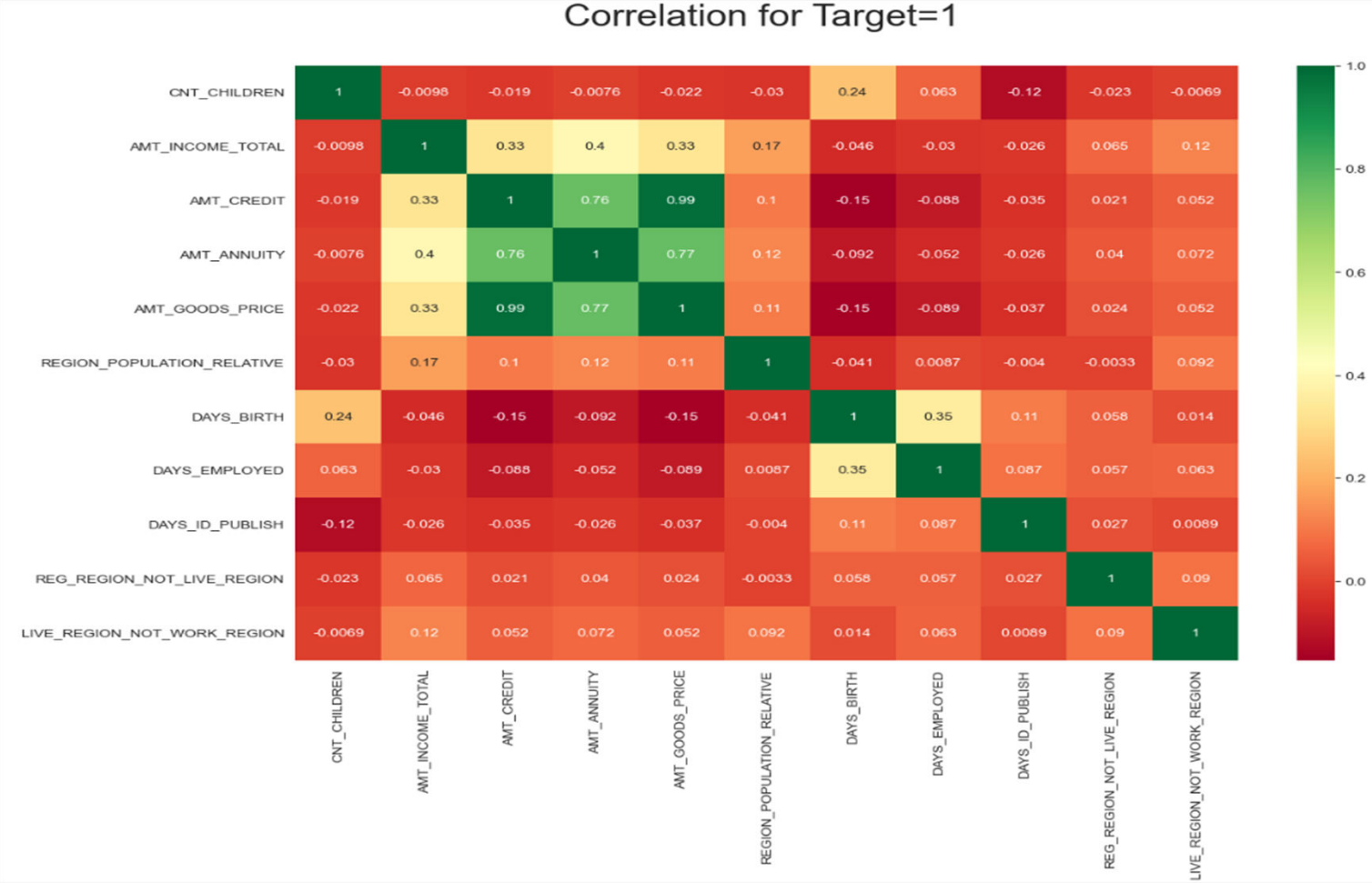
3. Business Entity Type 3, Business Entity Type 2,Self-Employed, other and Medicine are defaulting more.

Correlation For Target Variable 0



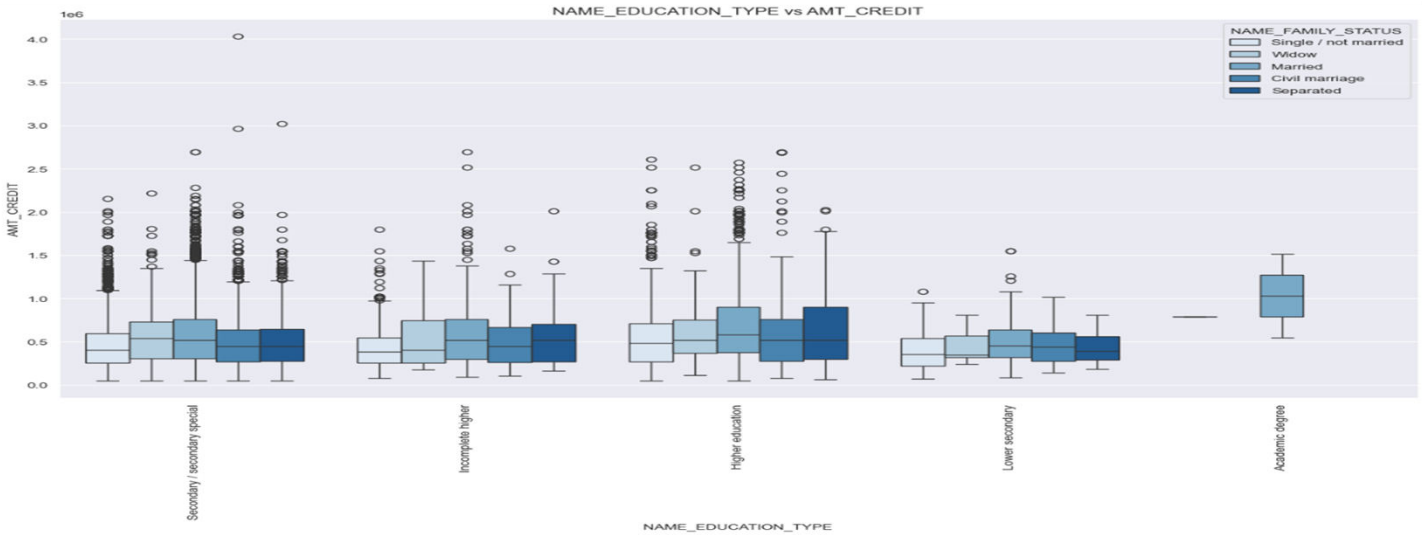
1. There is a high correlation(+VE) between Credit_amount & Goods_Price.
2. As Credit amount increases Days_Birth decreases i.e credit amount of the loan is higher for low age applicants and vice versa
3. As Credit amount increases DAYS_EMPLOYEMENT decreases i.e credit amount of the loan is higher and DAYS_EMPLOYEMENT is less ,vice versa.
4. As Credit amount increases count of children is less and vice versa.
5. More income applicants have less children they will have.
6. More income less age and vice versa.

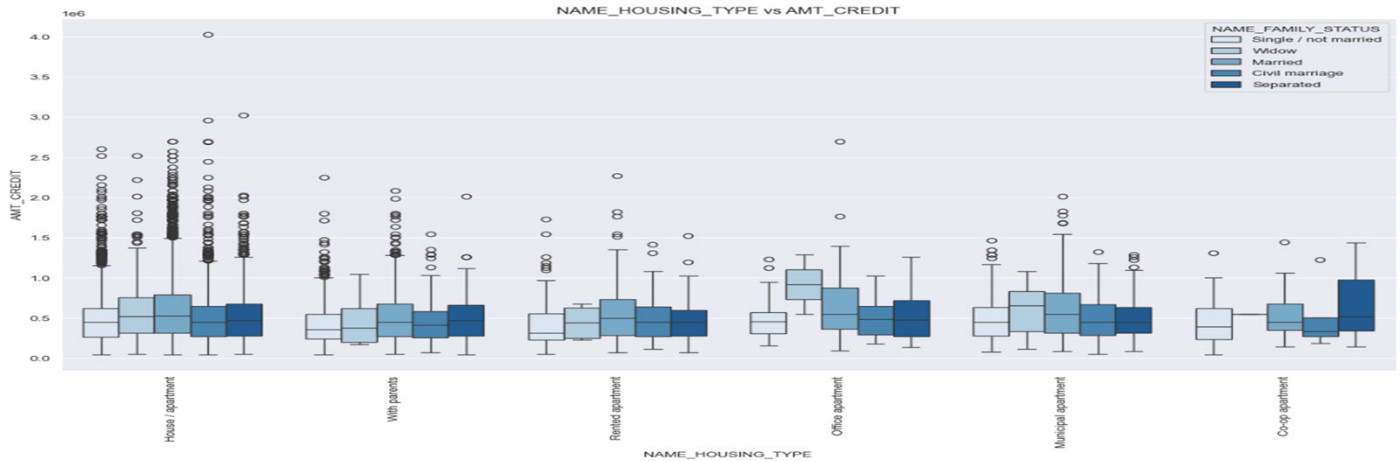
Correlation For Target Variable 1



similar as target_0 variable

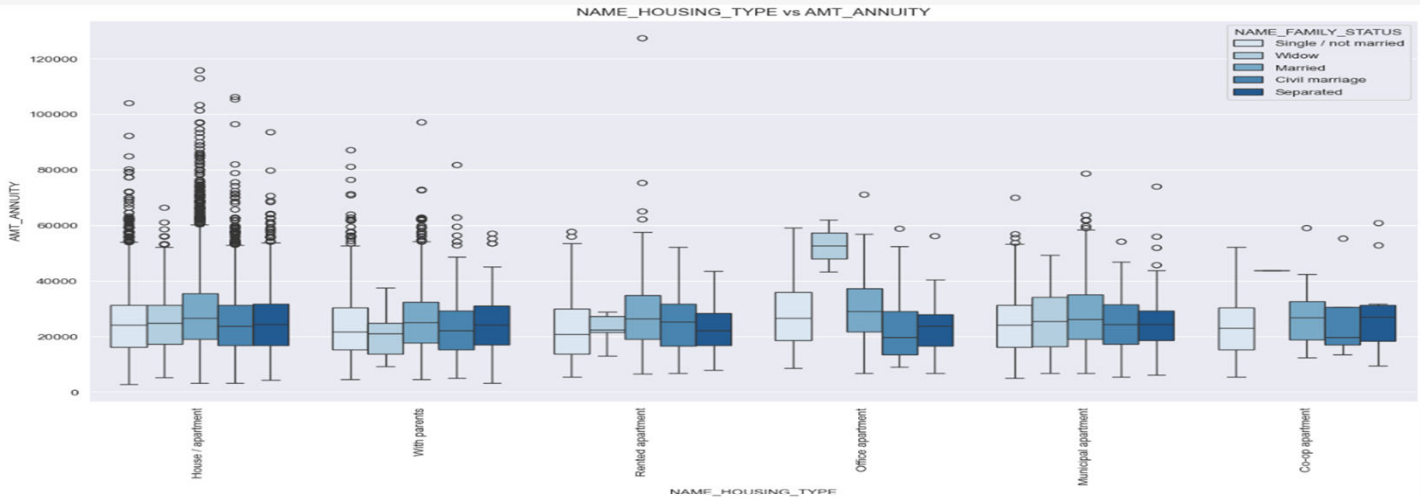
Bivariate Analysis For Categorical and Numerical variables(For Target variable 0)

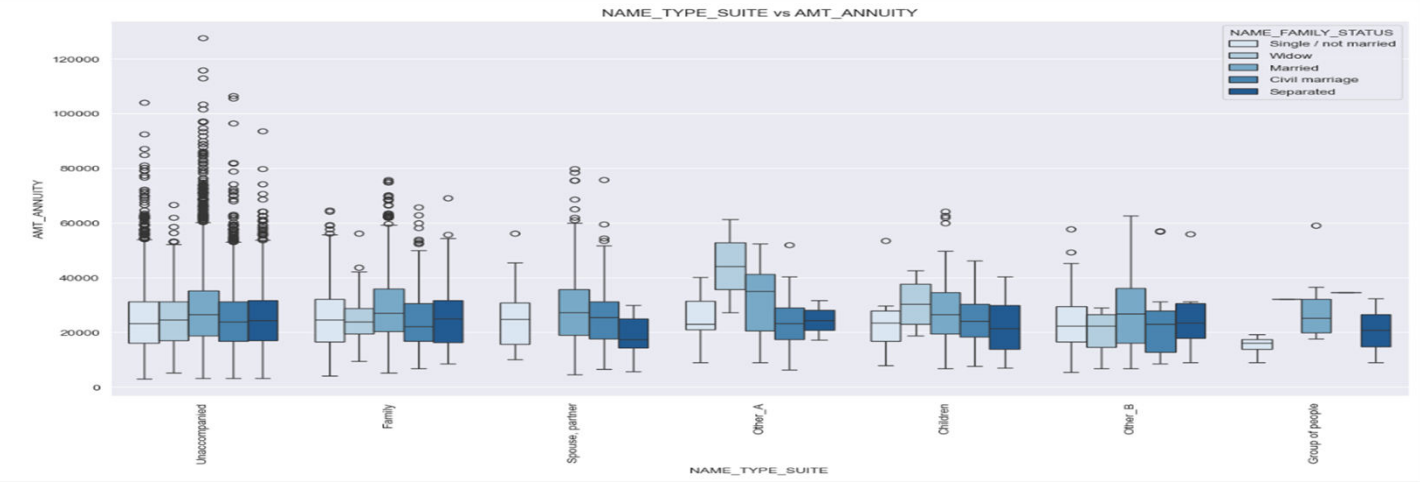




1. In Academic Degree Married has most number of credit amount of the loan(Defaulters)
2. Married,seperated and civil married applicants have more number of credits.
3. secondary special have more outliers in it

4. In Office appartment Widow has most number of credit amount of the loan
5. Married,seperated and civil married applicants have more number of credits.
6. House apartment have more outliers in it.

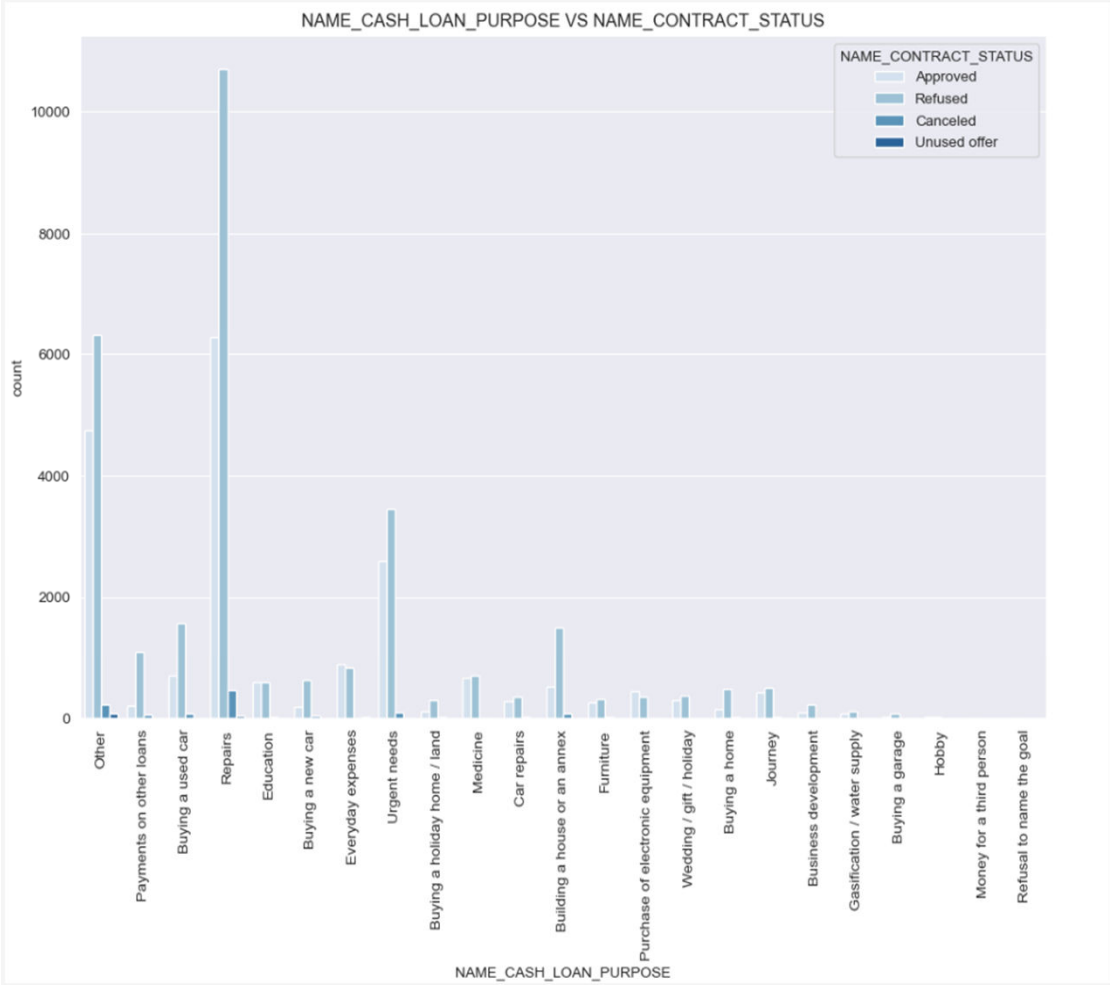


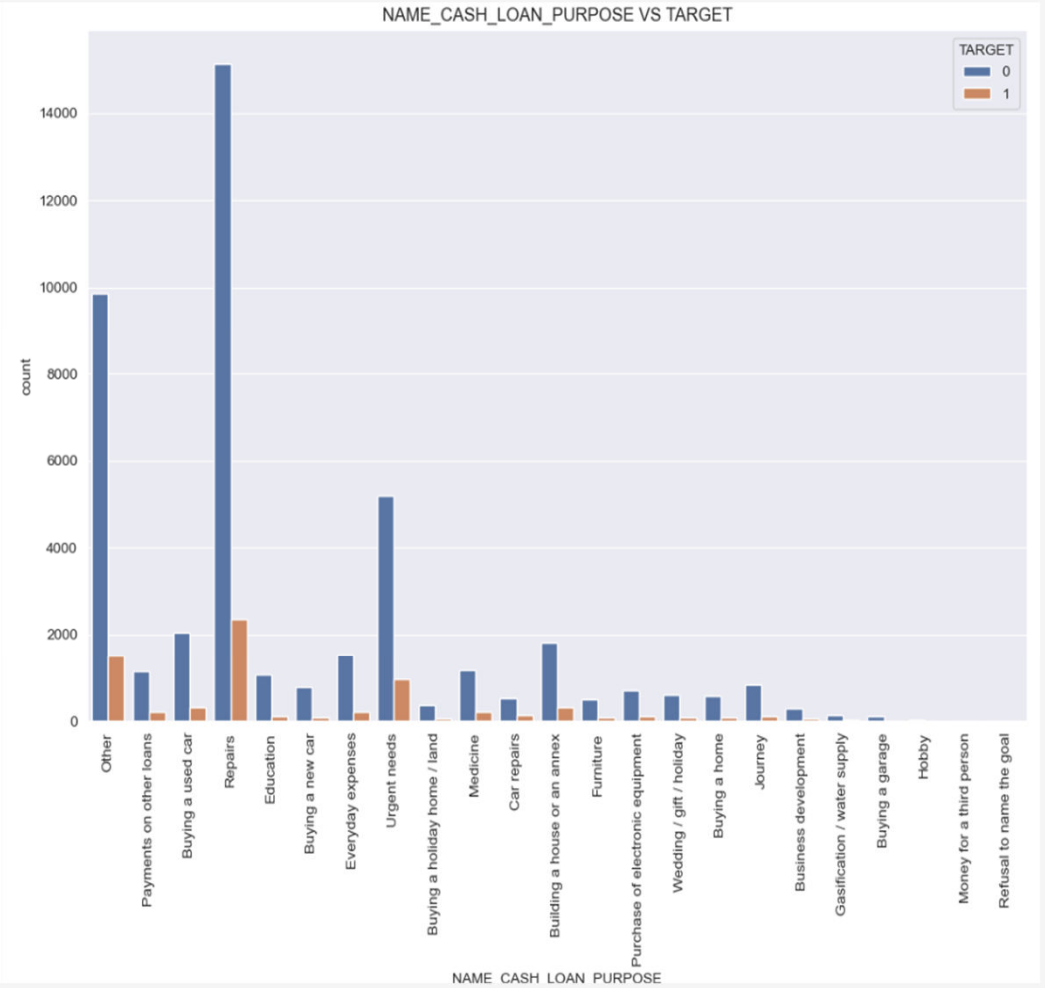


1. In office appartment Widow has highest Loan annuity as compare to others
2. Loan annuity for Married, seperated and civil married applicants are higher as compare to others.
3. House Apartment has more outliers in it.

4. In other_A Widow has highest Loan annuity as compare to others
5. Loan annuity for Married, seperated and civil married applicants are higher as compare to others.
6. Unaccompanied has more outliers in it.

Univariate Analysis after combining of both DataFrames





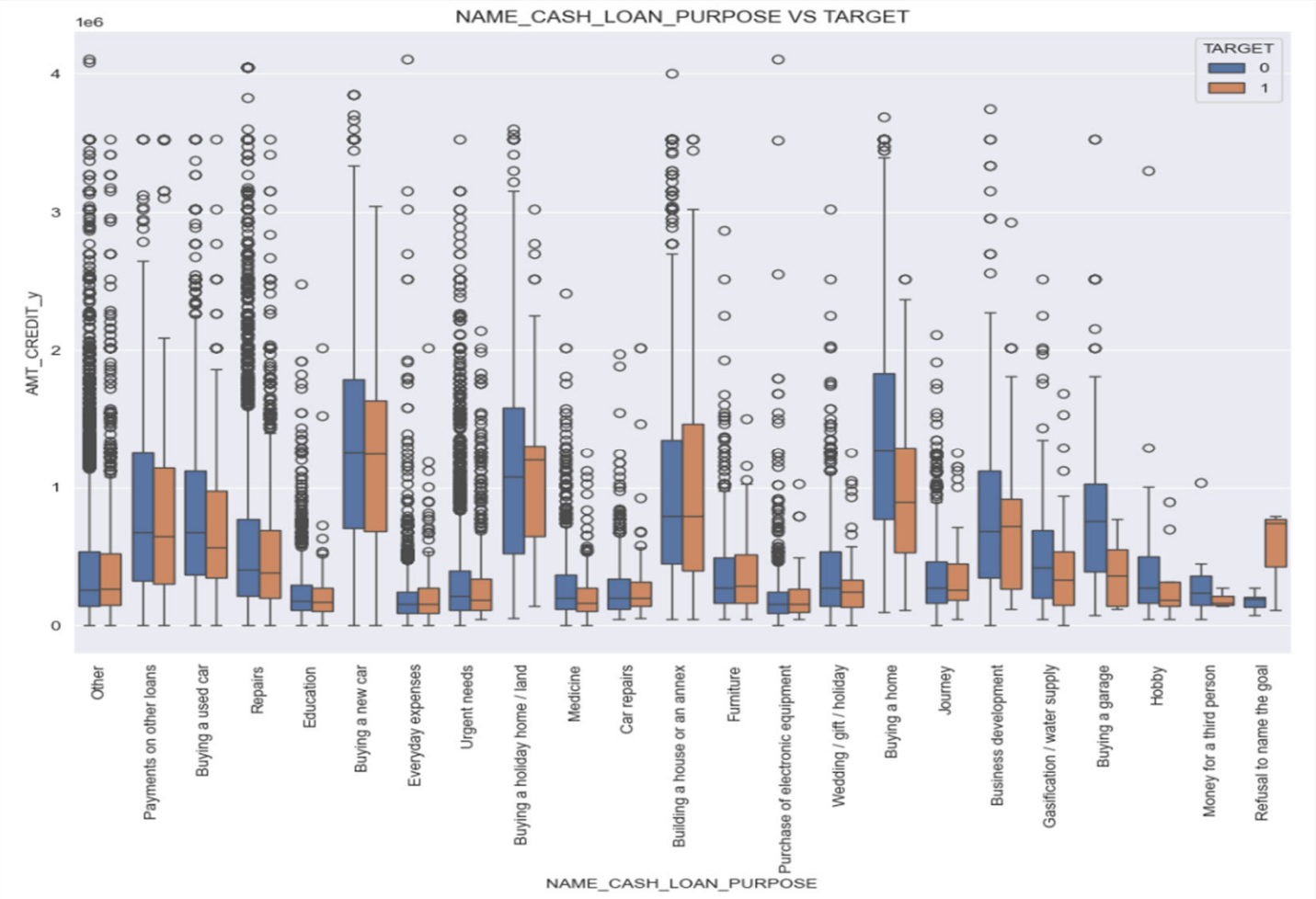
NAME_CASH_LOAN_PURPOSE VS NAME_CONTRACT_STATUS

1. Maximum number of approvels and disapprovels came from Repair category
2. Buying a used or new car has disapprovels more then approvels
3. Urgent needs has almost equal number of approvels and disapprovels.

NAME_CASH_LOAN_PURPOSE VS TARGET

There is difficulty in repaying the loan on time for Repairs, other and Urgent needs category

Bivariate Analysis Of Combined data



1. Applicants Buying a new car, Buying a holiday home, Building a house, Buying a home, Business development are not paying loans on time(Defaulters).