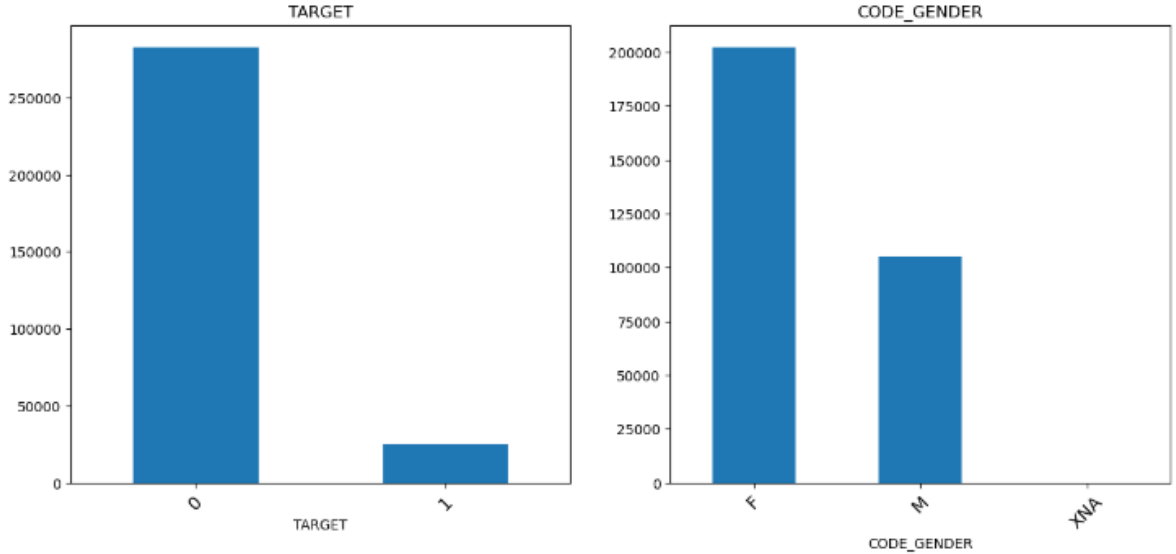
EDA Assignment Submission

By Vyankatesh Kshatriya

Application Dataset

- Importing of libraries and data
- Understanding of data (use of: .head(), .tail(), .shape, .info()
- Data checks:
  - Finding of % null values (Dropping columns having more than 50% null values)
  - 8 column have more than 13% null values
    - Suggesting how data can impute into those columns
  - Understanding of datatypes
    - Changing datatypes for categorical column as ‘category’
  - Checking negative values (converting –ve values to +ve using .abs() method)
  - Finding outliers using boxplot
  - Binning and Bucketing

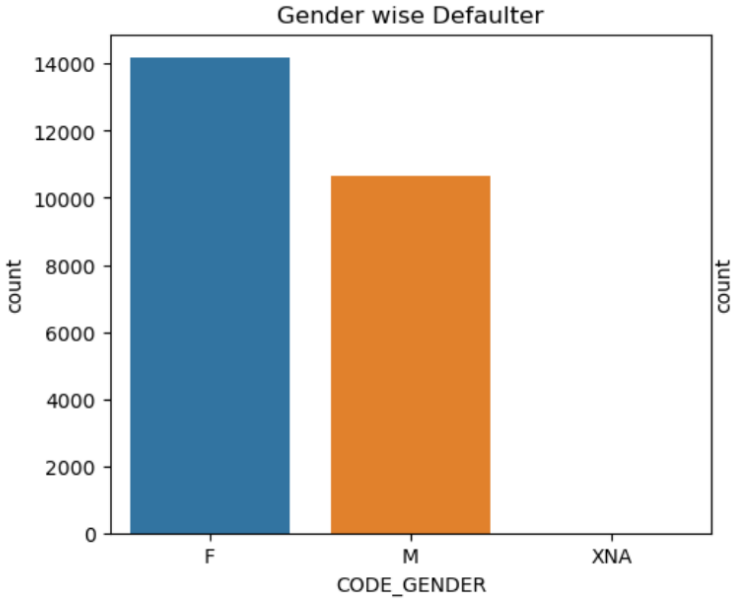
Data Imbalance Study

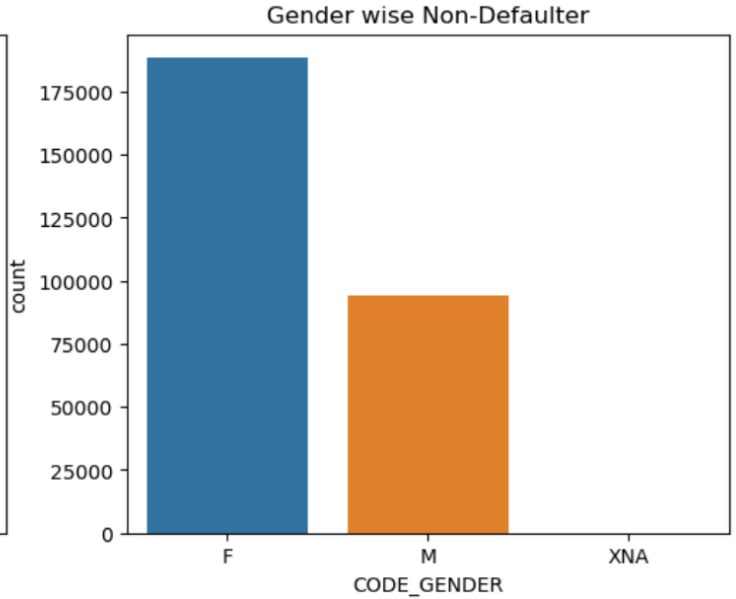


Conclusion-

- The number of defaulters are very less when compared to the repayers
- There are more female clients than male clients.

Univariate Analysis of Target column across categories

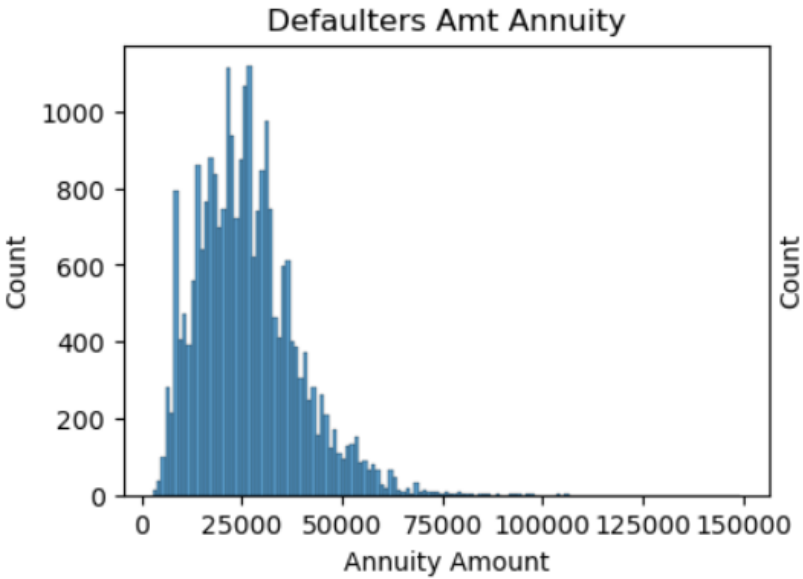


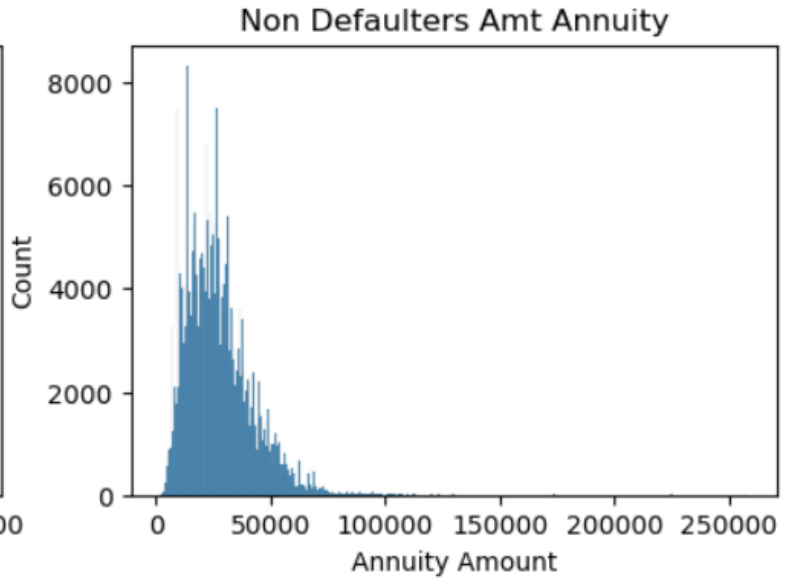


Conclusion-

- Most of the client opting for loans are female whether is it defaulter or non defaulter

Univariate Analysis of Target column and Numeric Columns

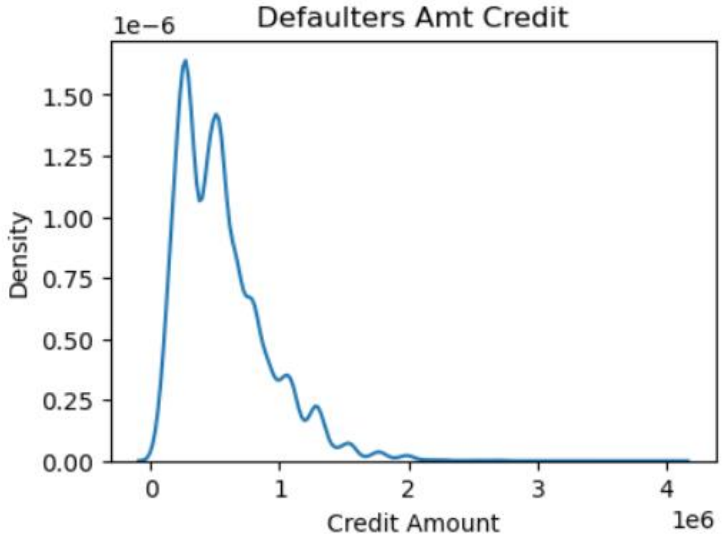


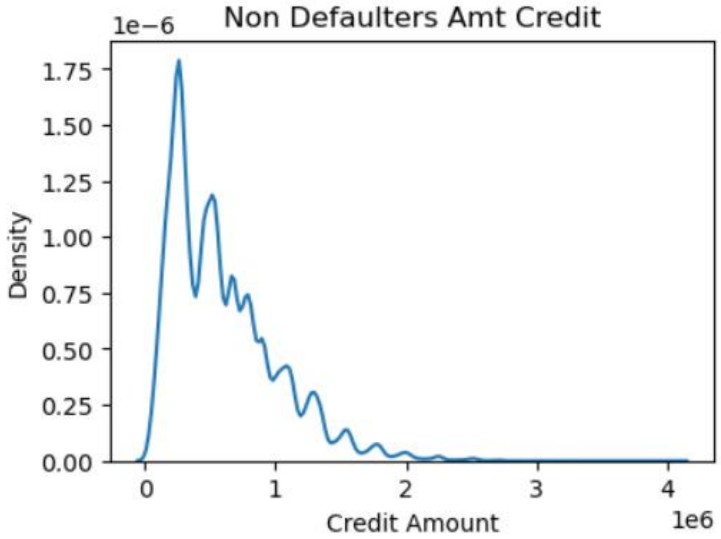


Conclusion-

- As mentioned earlier and after seeing this graph its confirmed that the loan takers have taken secondary education

Univariate Analysis of Target column and Numeric Columns



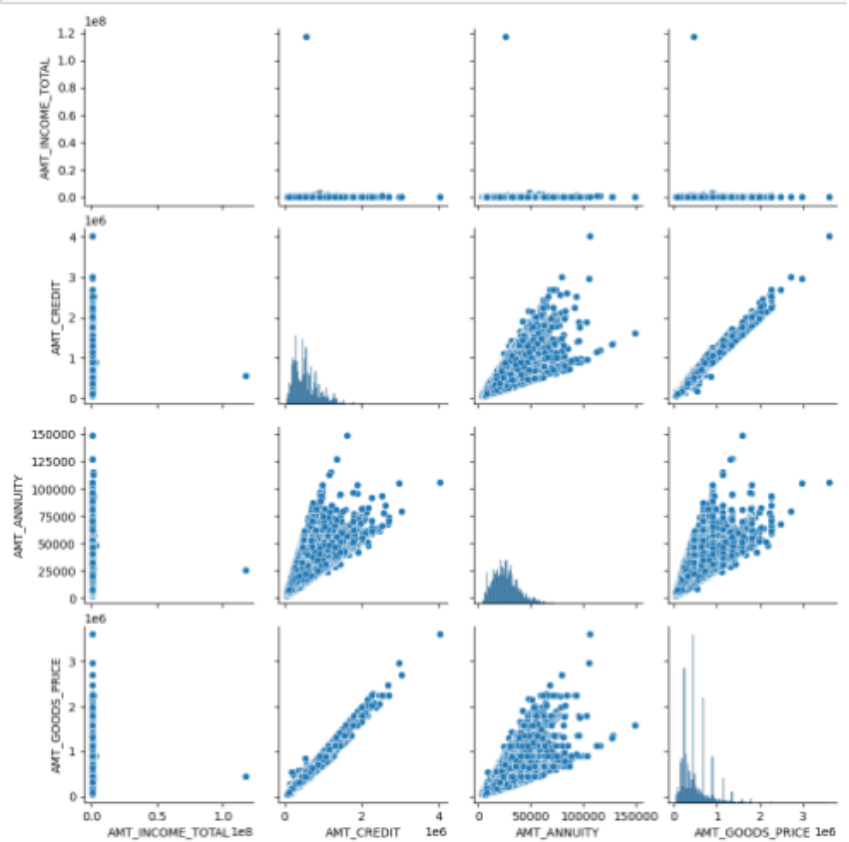


Conclusion-

- Here, as the credit amt decrease the chances of default increase
- In other words, when the credit amt is less the chances of payment gettig default is more

Bivariate Analysis

- Numerical vs Numerical

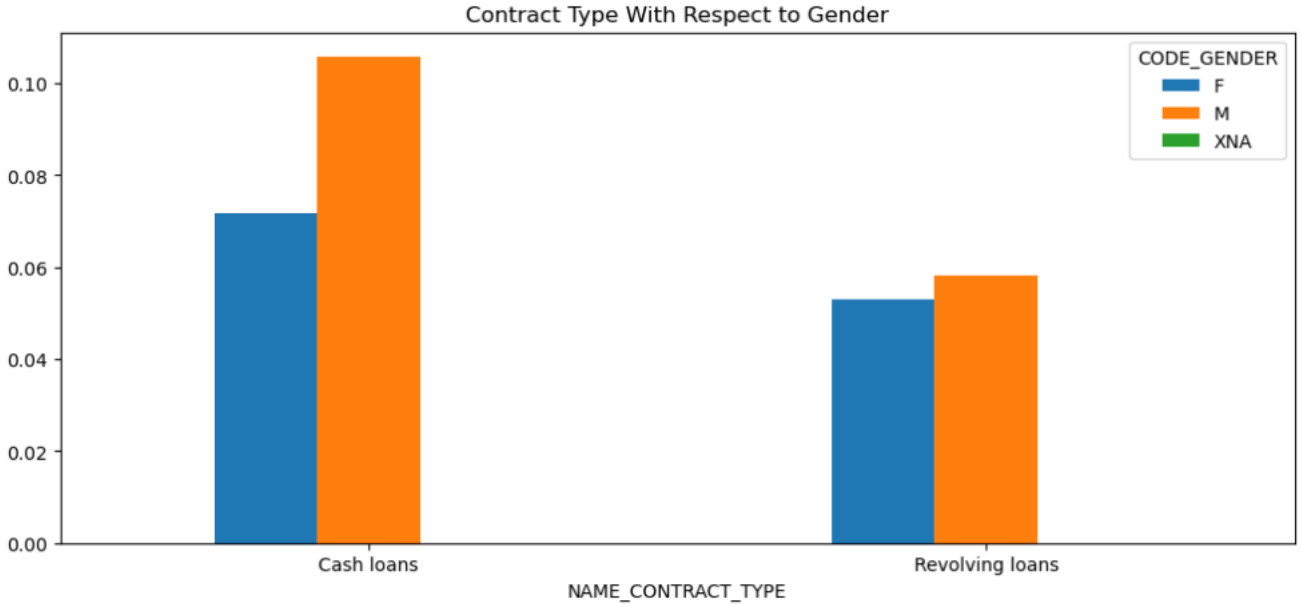


Conclusion-

- This shows strong linear relation of AMT_CREDIT and AMT_GOODS_PRICE
- weak relation of AMT_INCOME_TOTAL and AMT_GOODS_PRICE
- weak relation of AMT_ANNUITY and AMT_GOODS_PRICE
- weak relation of AMT_INCOME_TOTAL with other 3 columns

Bivariate Analysis

- Categorical vs Categorical

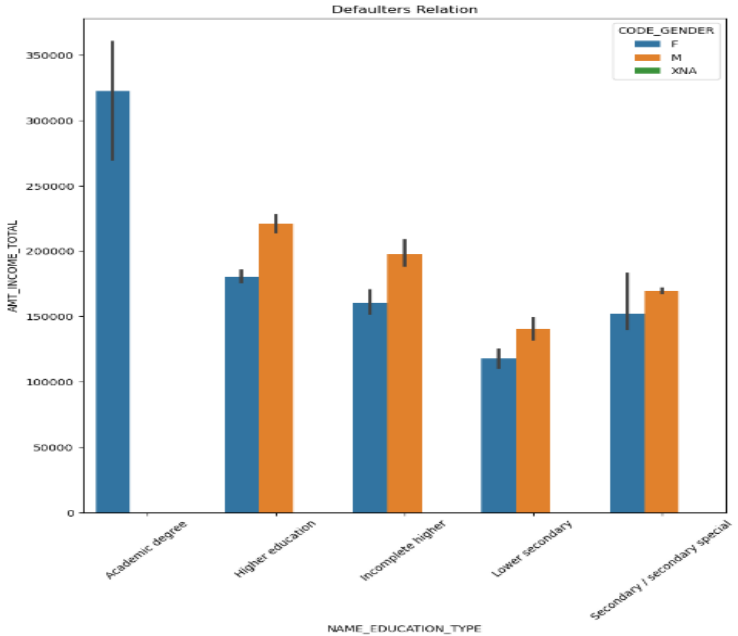


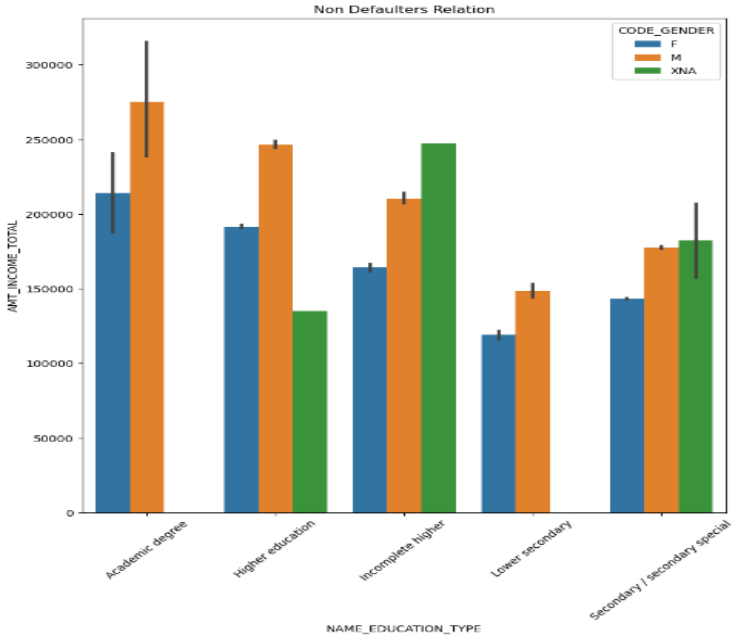
Conclusion-

- Male have higher amount application for both types of loan i.e. cash loans and revolving loans than females

Bivariate Analysis

- Numerical vs Categorical





Conclusion-

Defaulter

- Female academic holders are earning higher amount still there is not a single male defaulter having academic degree
- In all other cases, male are earning more than female in every single education category

Non-Defaulter

- In non-defaulters, in each of the category of education male are earning more than female

Previous Application Dataset

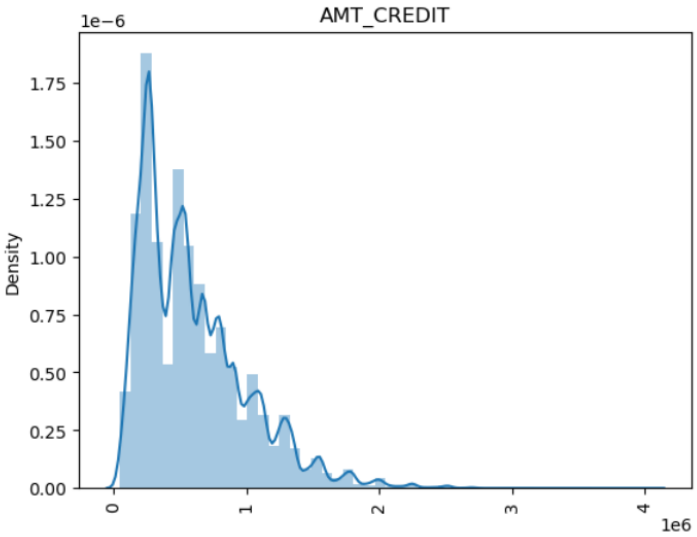
- Importing of libraries and data
- Understanding of data (use of: .head(), .tail(), .shape, .info()
- Data checks:
  - Finding of % null values (Dropping columns having more than 50% null values)
  - 8 column have more than 13% null values
    - Suggesting how data can impute into those columns
  - Understanding of datatypes
    - Changing datatypes for categorical column as ‘category’
  - Checking negative values (converting –ve values to +ve using .abs() method)
  - Finding outliers using boxplot
  - Binning and Bucketing

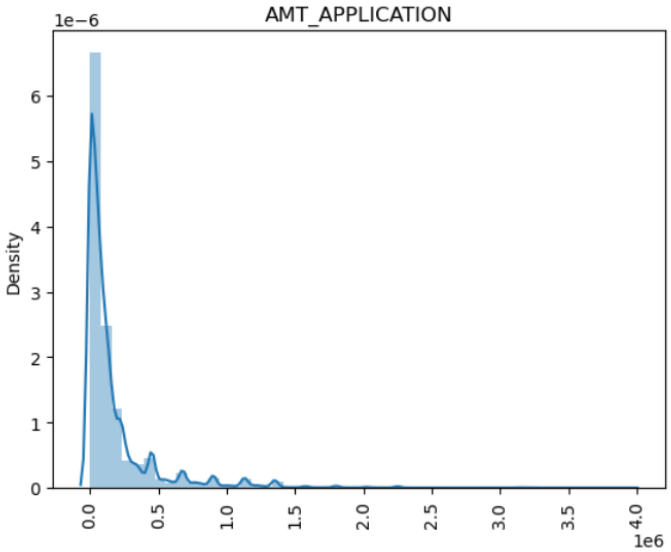
Merge Dataset

- Prevappl and appl, both dataset has one common key SK_ID_CURR
- Hence, both these dataset will be merged into one on this key
- I will choose joining type as left as i want all the columns from appl to be present in the new dataset
- I will choose few column from appl and few from prevappl to be merged
- Merged data into one dataset by using pd.merge()

Merge Data

- **Univariate Analysis**
  - Numerical Column



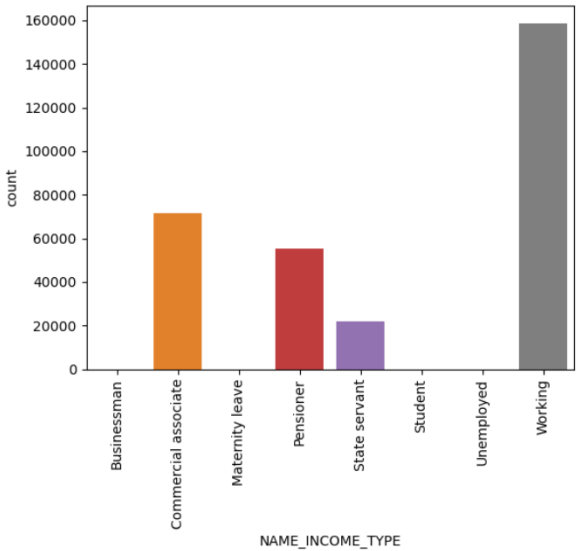


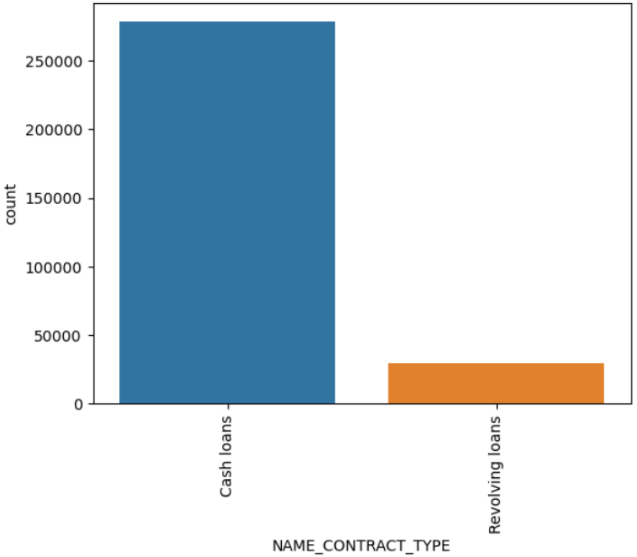
Conclusion-

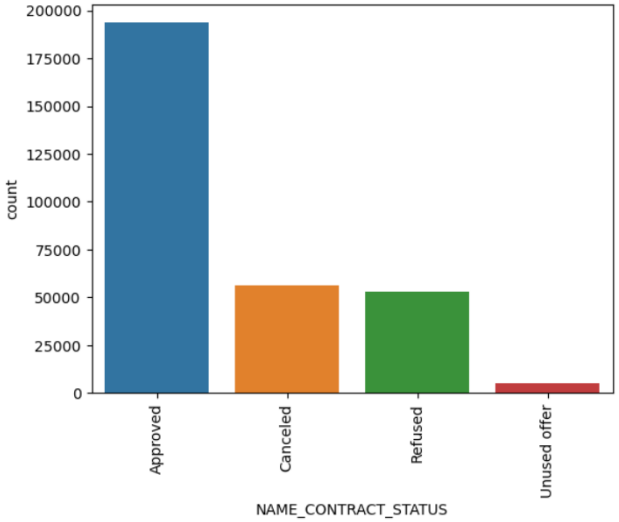
- Most of the credited loan amount lies between 30k-1.25 lakhs
- Mostly clients has asked for the amount in application is for upto 50k

Merge Data

- **Univariate Analysis**
  - Categorical Column



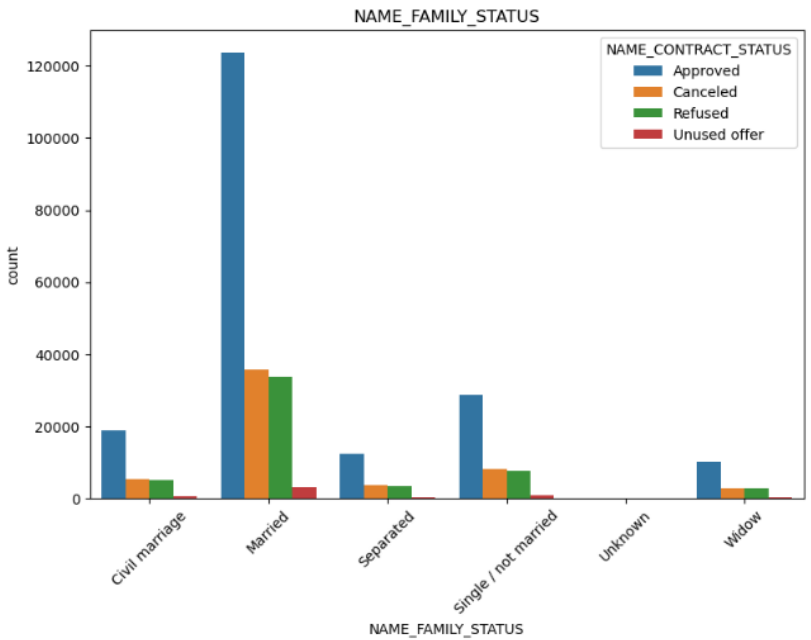


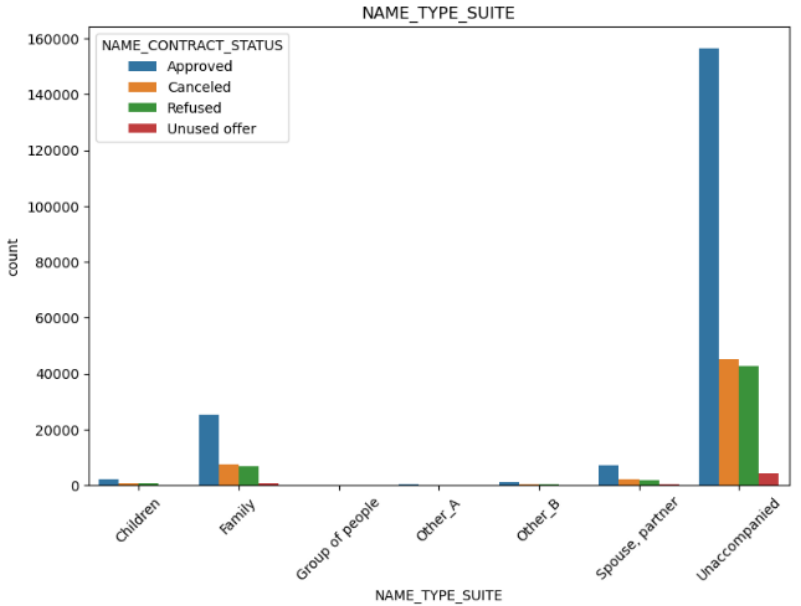


Conclusion-

- The number of loans taken as cashloans is far more than revolving loans
- Most of the customers who has taken loans has income type as worker
- Approved loans has higher number than cancelled or refused loans

Merge Data • Bivariate Analysis



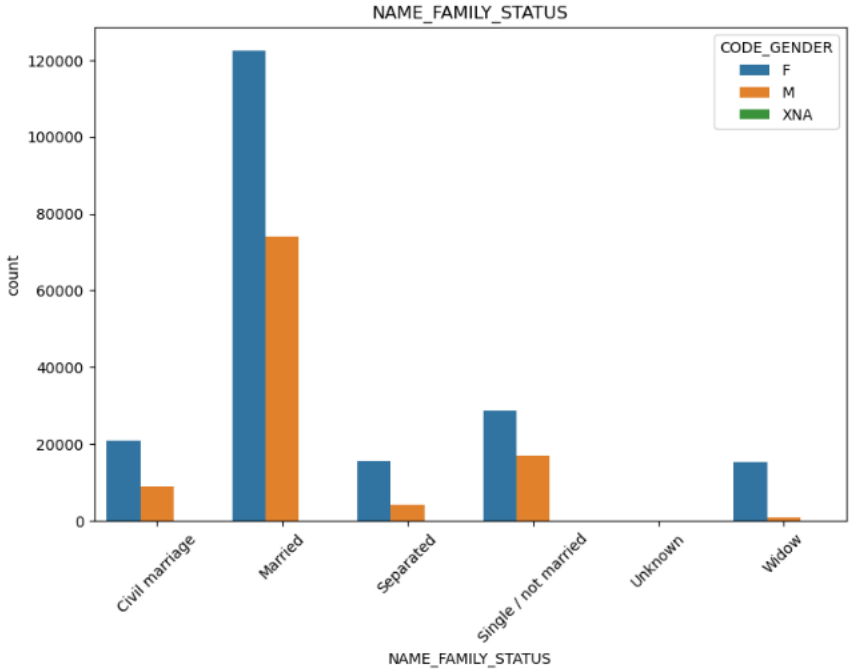


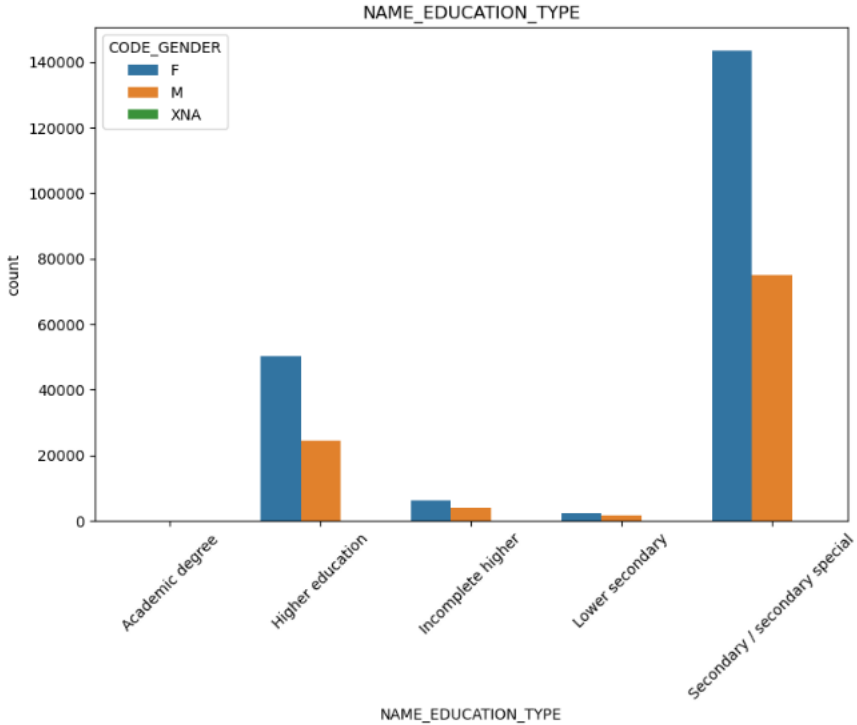
Conclusion-

- Client as unaccompanied type suite has applied loan mostly
- Client as family status married has applied for most of the loans

Merge Data

- **Bivariate Analysis**



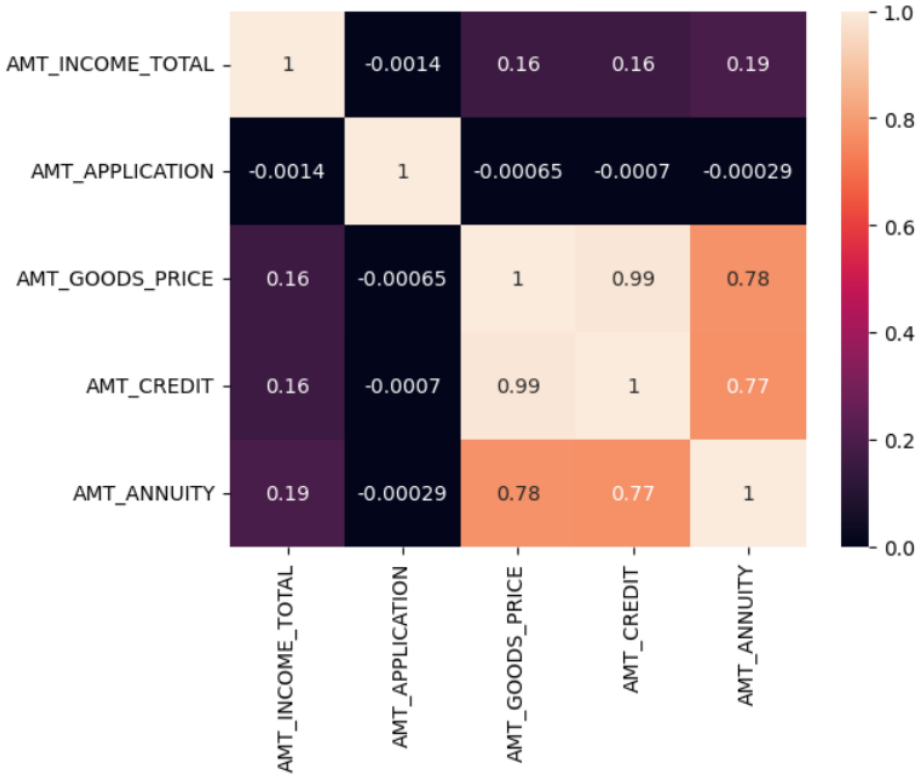


Conclusion-

- In all the above cases/graphs, its clearly visible that females has applied for the more number of loans than male did

Merge Data

- **Multivariate Analysis**



Conclusion-

- AMT_GOODS_PRICE and AMT_CREDIT has shown greater correlation
- AMT_GOODS_PRICE, AMT_CREDIT and AMT_ANNUITY has weak correlation with AMT_INCOME_TOTAL
- AMT_ANNUITY with AMT_GOODS_PRICE and AMT_CREDIT has shown moderate correlation with each other