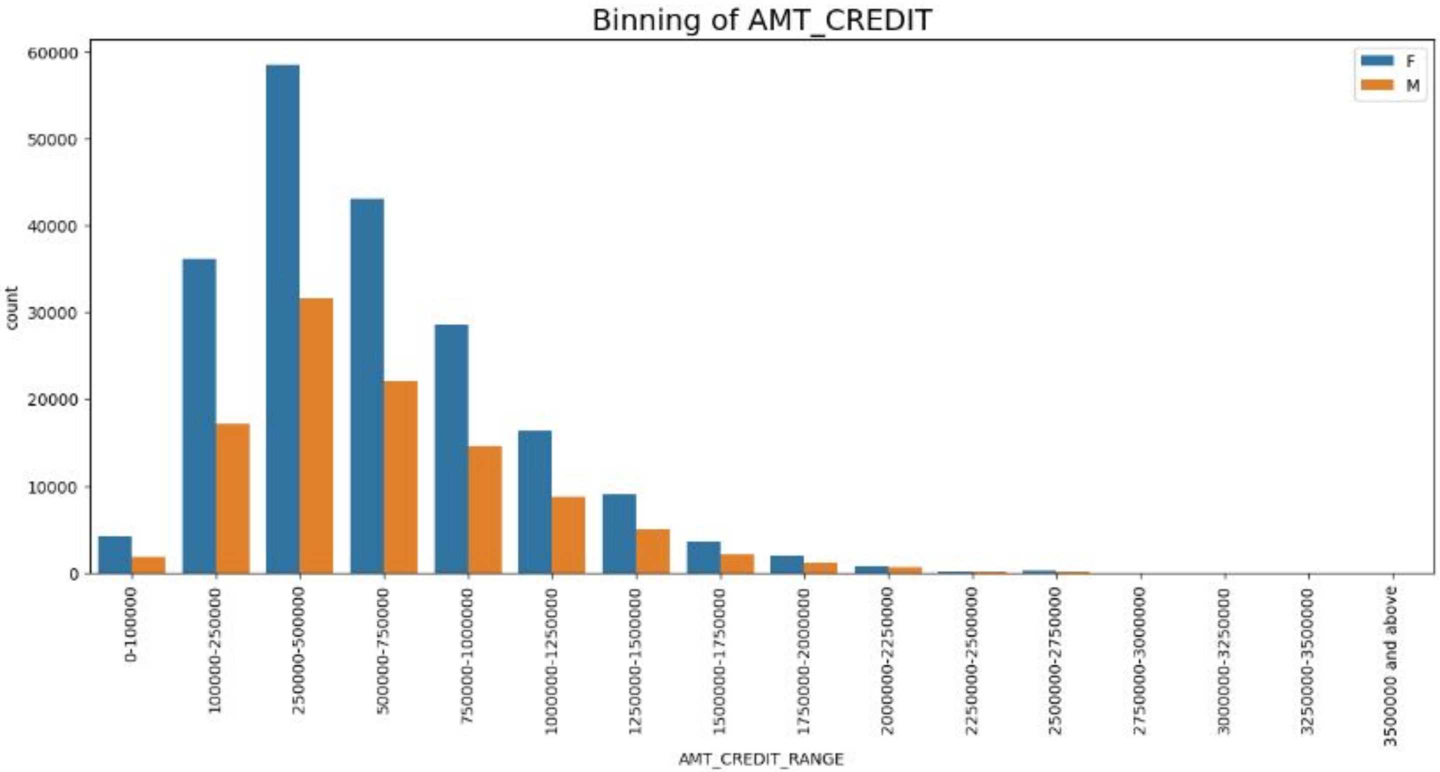
CREDIT EDA CASE STUDY ASSIGNMENT

- Shweta Alukuru Trikutam

AIM:

1. 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.
2. Identification of such applicants using EDA is the aim of this case study.
3. The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.



CURRENT APPLICATION -

Insights - The above plot shows the Binning of AMT_CREDIT in ranges.

1. Female counts are higher than Male counts.
2. The credit range from 250000 - 500000 have more number of counts.
3. Very less count for credit range from 1250000 and above.

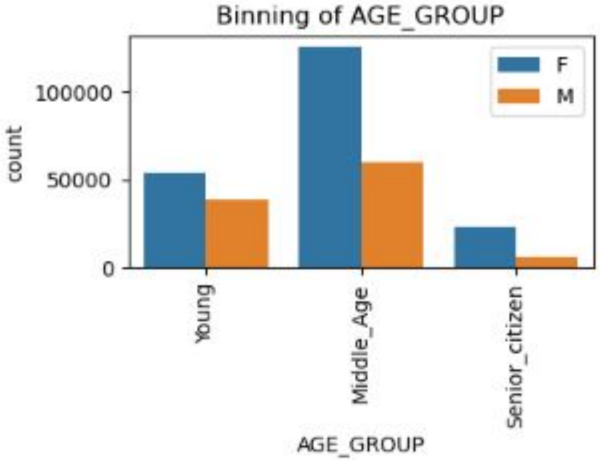
Insights-

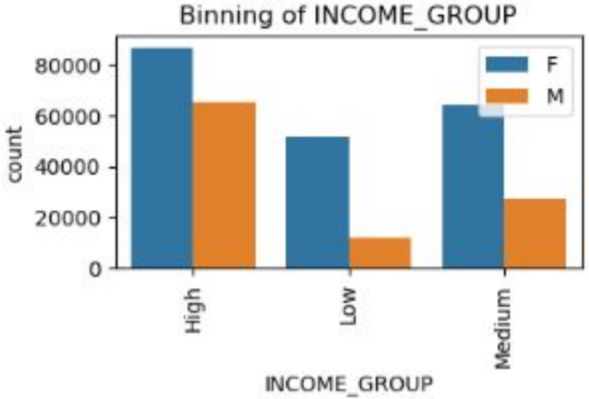
For Plot 1 - AGE_GROUP

1. There are more number of 'Females' than 'Males' in all age group.
2. Clients from the 'Middle Age' group are more in number and Clients who belong to 'Senior Citizen' group are less.

For Plot 2 - INCOME_GROUP

1. Clients who are 'Females' are more in number than 'Males' in all income group.
2. Clients with 'High Income' are more in number and Clients with 'Low Income' are less in number.



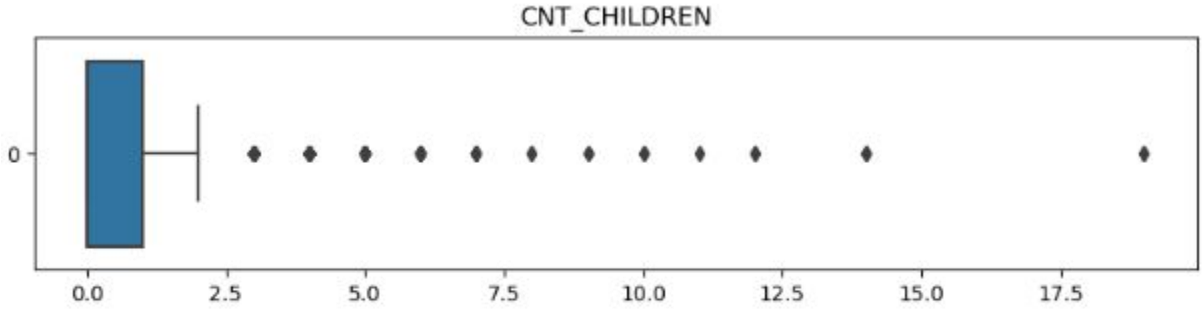


OUTLIERS

Insights -

- From the plot, we can see that count of number of children goes more than 17.5, which is not possible in general case scenario. Hence, this variable has an outlier.
- We can also notice that the 1st Quartile is missing for 'CNT_CHILDREN'. Hence, we can say that most of the data is present in the 1st quartile.

(Note - Only one of the variable having Outliers is shown here. There are more outliers identified in notebook.)



Dealing with Outliers -

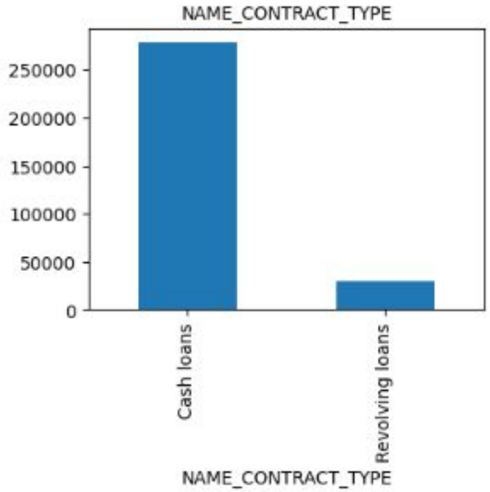
Finding outliers in all the numerical columns with 1.5 IQR rule and removing the outlier records

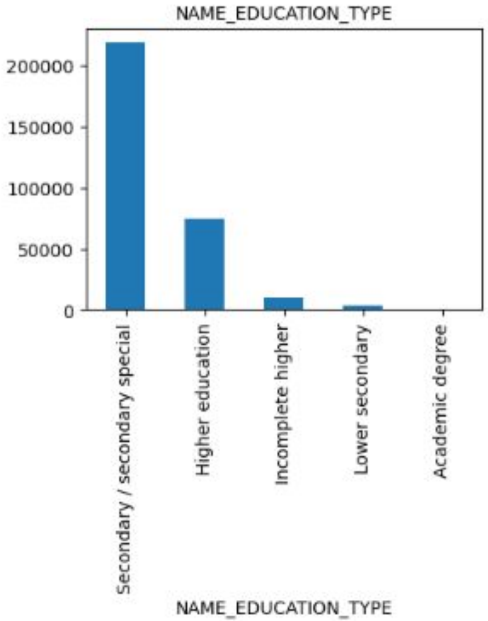
```
outlier_col = ['AMT_INCOME_TOTAL' , ‘AMT_CREDIT’]

for col in outlier_col:

    q1 = df_app[col].quantile(0.25)

    q3 = df_app[col].quantile(0.75)

    iqr = q3-q1

    range_low  = q1-1.5*iqr

    range_high = q3+1.5*iqr

    df_app = df_app.loc[(df_app[col] > range_low) & (df_app[col] < range_high)]
```

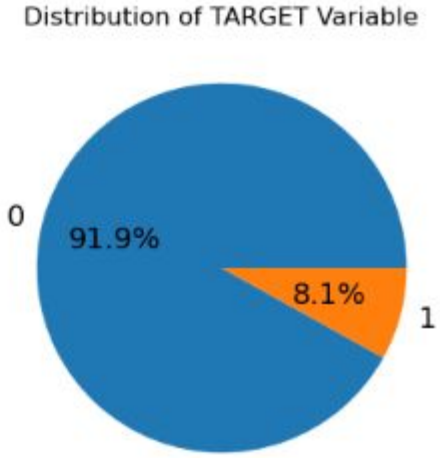
IMBALANCE DATA

Insights -

1. **NAME_CONTRACT_TYPE** - There are very few 'Revolving loans' than 'Cash loans'.
2. **NAME_EDUCATION_TYPE** - Most of the loans are applied by 'Secondary/Secondary special' educated people.
3. The 3rd plot, shows the distribution of the TARGET variable.



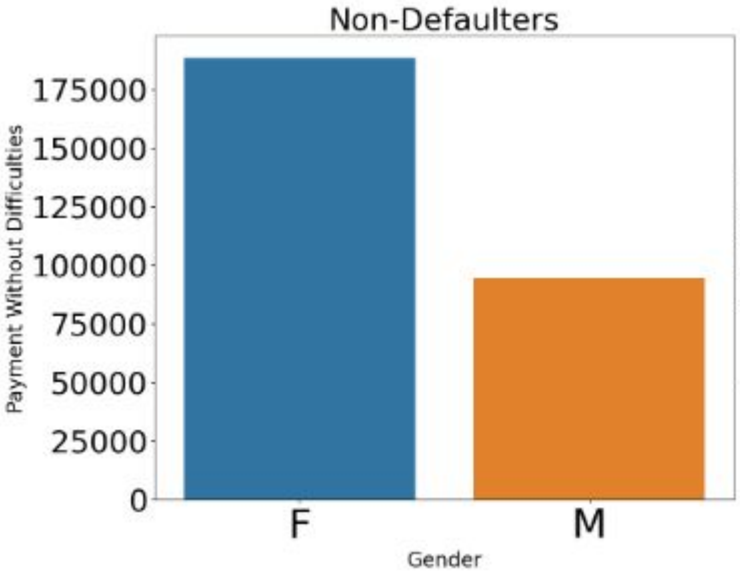


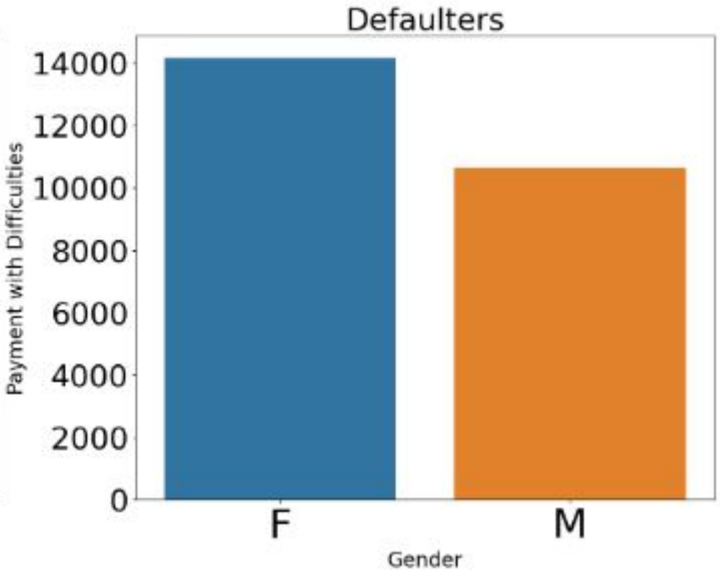


Univariate Analysis

Insights-

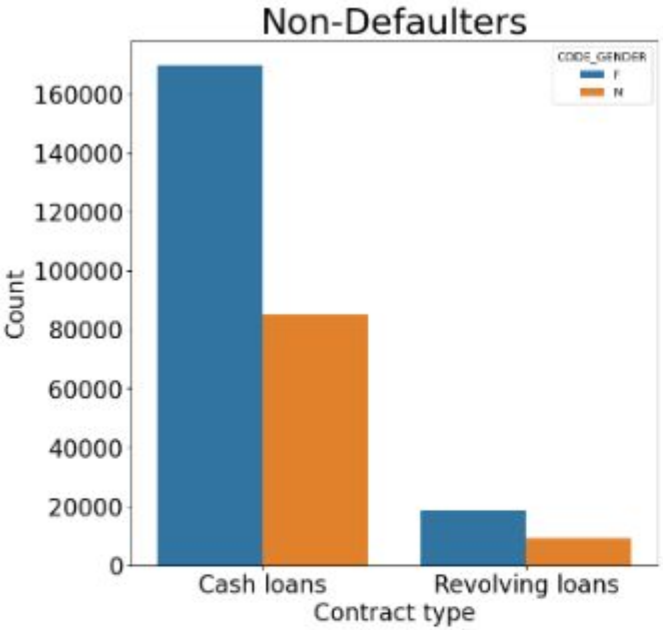
1. Female counts are higher than Male counts in both the targets.
2. There are higher number of Female Clients who have no payment difficulties.

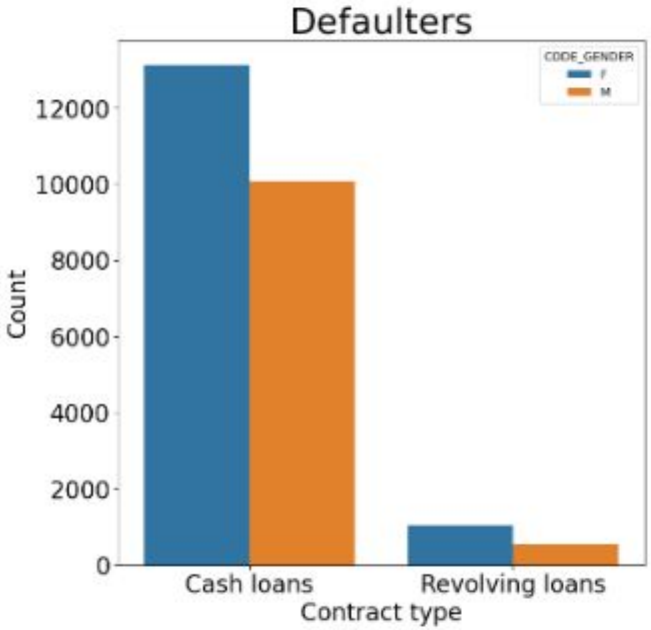




Insights -

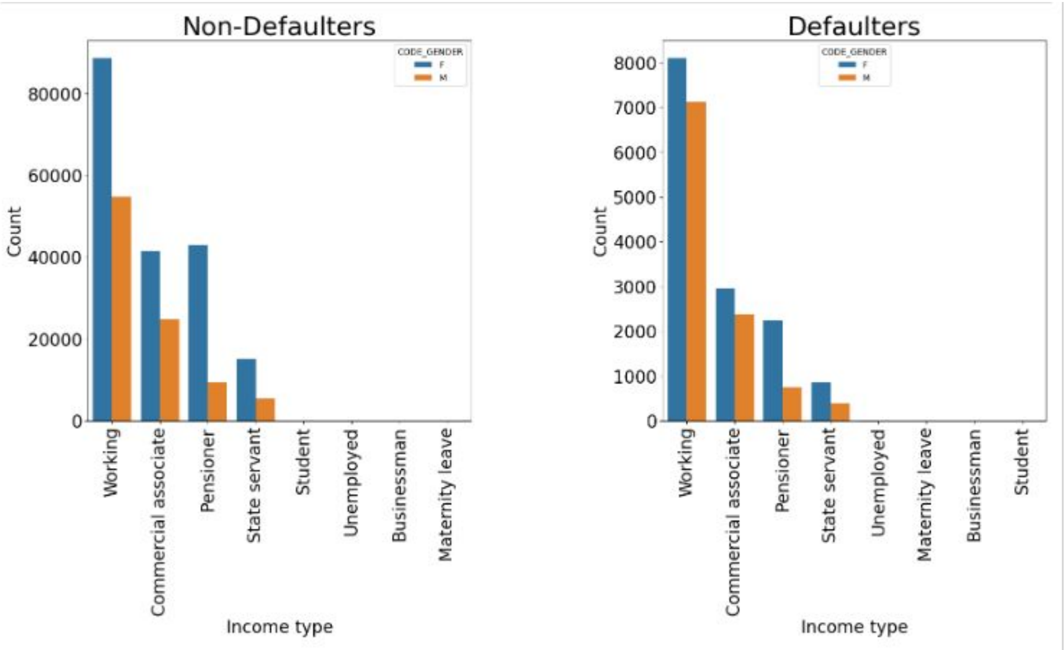
1. The Contract type 'Cash loans' have higher number of credits than 'Revolving loans' contract type.
2. The Female have more number of counts.





Insights -

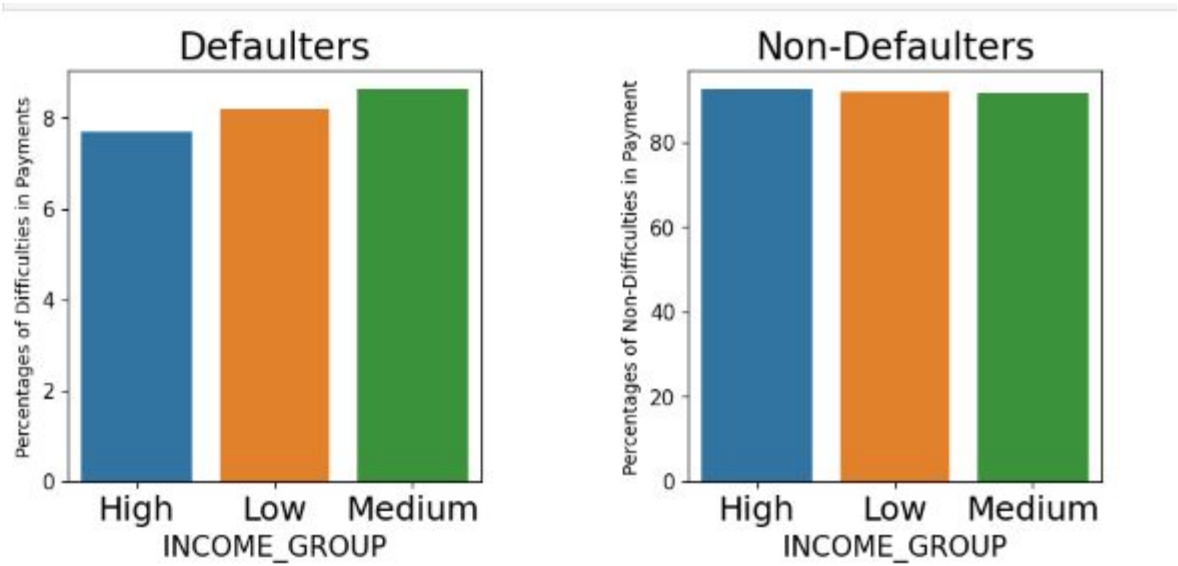
1. For Income Type 'Working' , 'Commercial Associate' and 'Pensioner' has higher number of credit counts compared to others.
2. Also, Females have more in number than Males.



Segmented Univariate Analysis

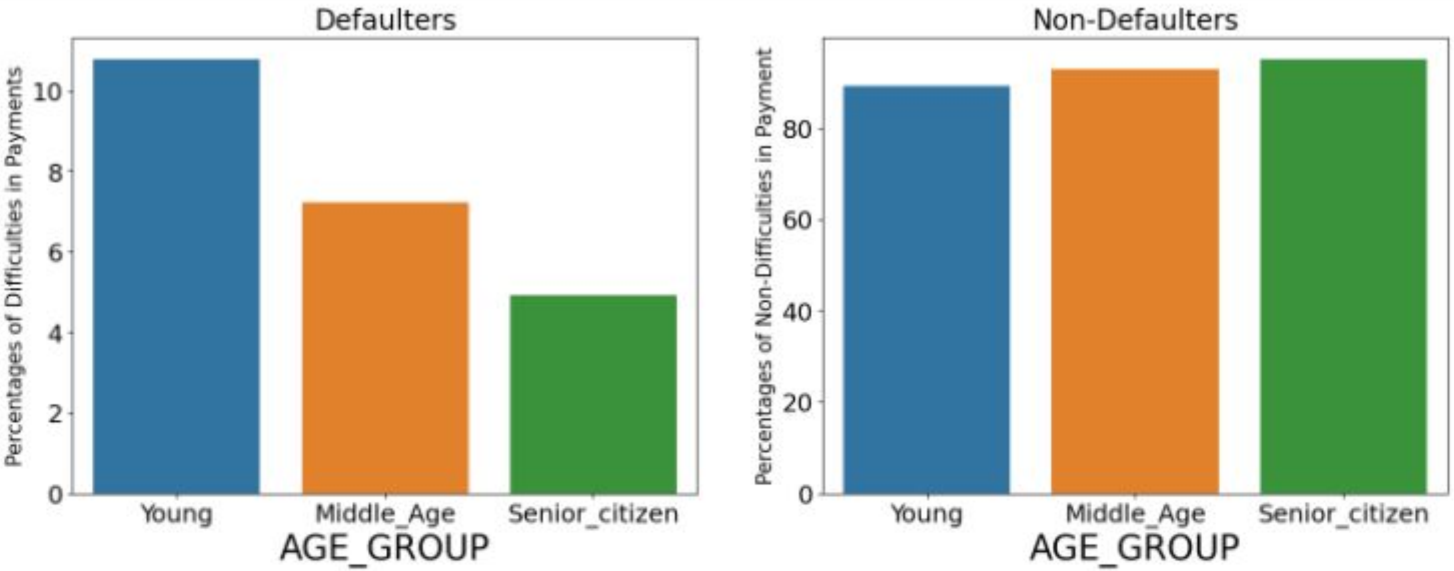
Insights

1. All income groups have almost same number of counts in no payments difficulties.
2. Clients with Medium Income have higher payment difficulties, followed by Low Income clients.



Insights

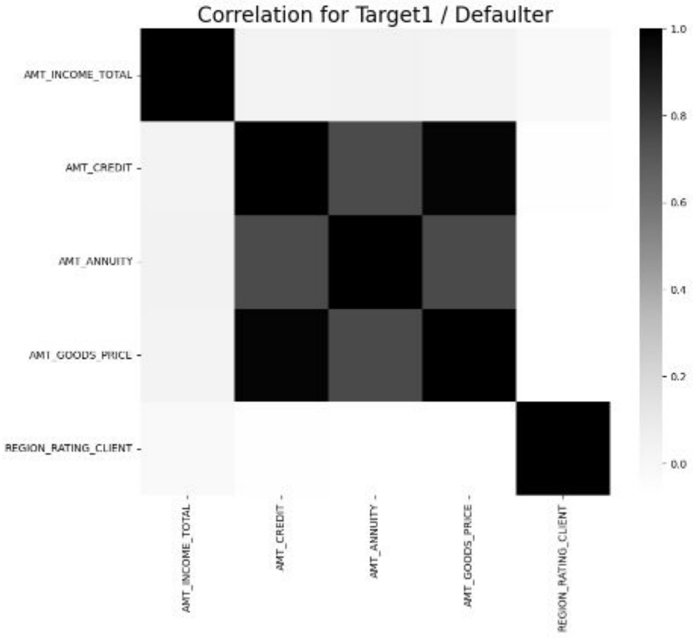
1. ‘Young’ age group clients have higher payment difficulties compared to other age groups.
2. There is no much difference in clients who are able to pay without difficulties, but carefully observing, we can see that ‘Senior Citizens’ age group have higher counts.

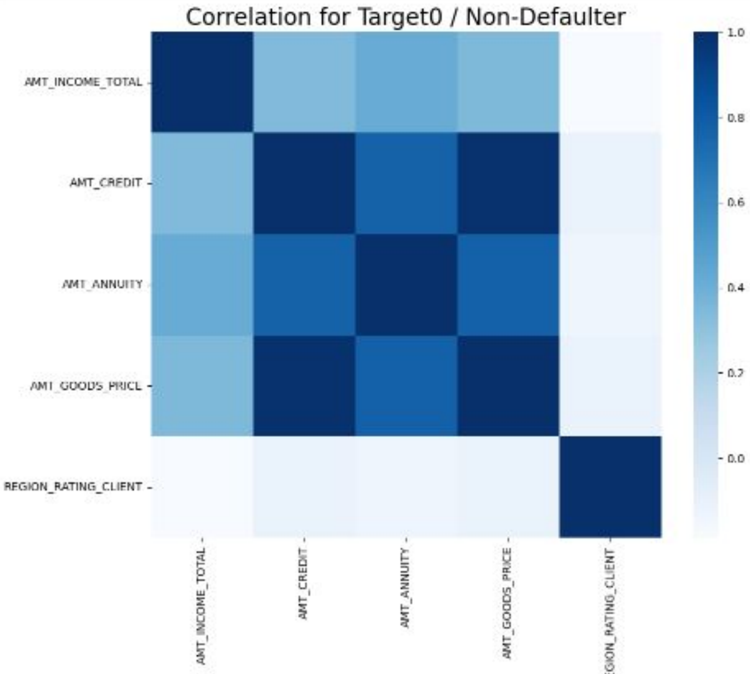


Correlation Matrix

Insights - High Corelate Columns for Target0 and Target1 ->

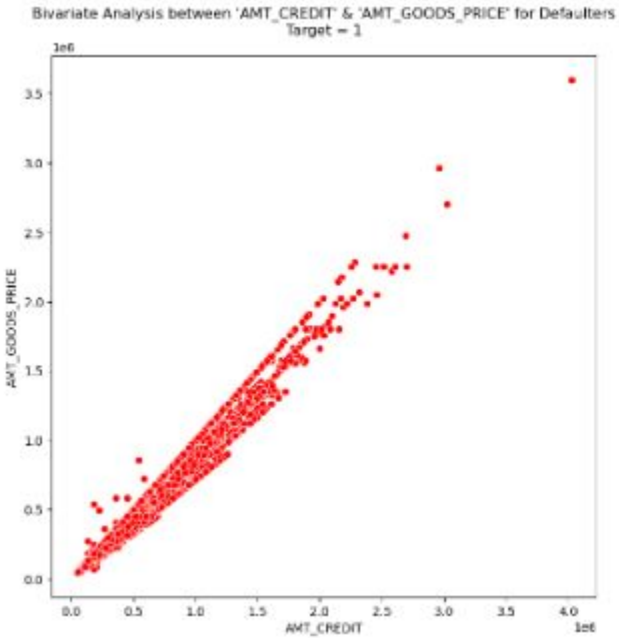
1. AMT_CREDIT & AMT_ANNUITY
2. AMT_CREDIT & AMT_GOODS_PRICE
3. AMT_ANNUITY & AMT_GOODS_PRICE

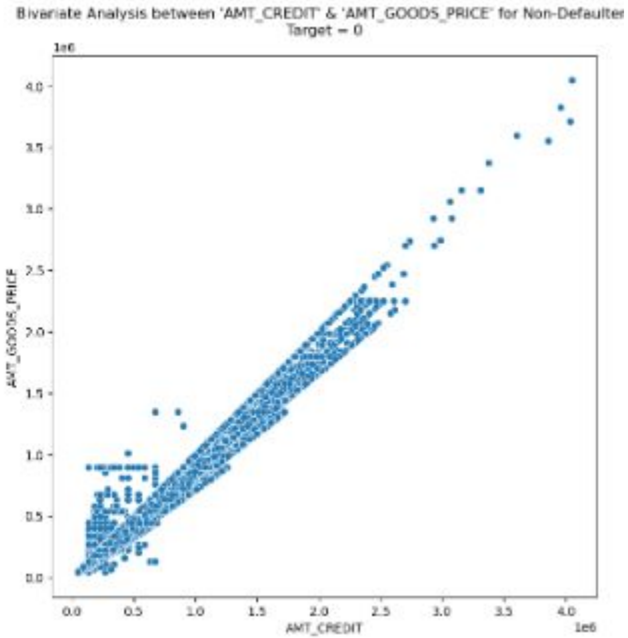




Bivariate Analysis

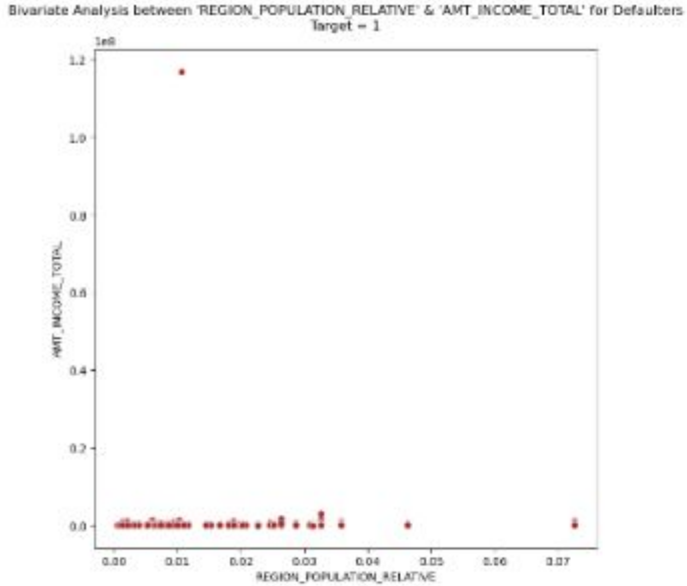
Insights - 'AMT_CREDIT' and 'AMT_GOODS_PRICE' are showing the same kind of trend as the credit amount maybe same or less than goods price.

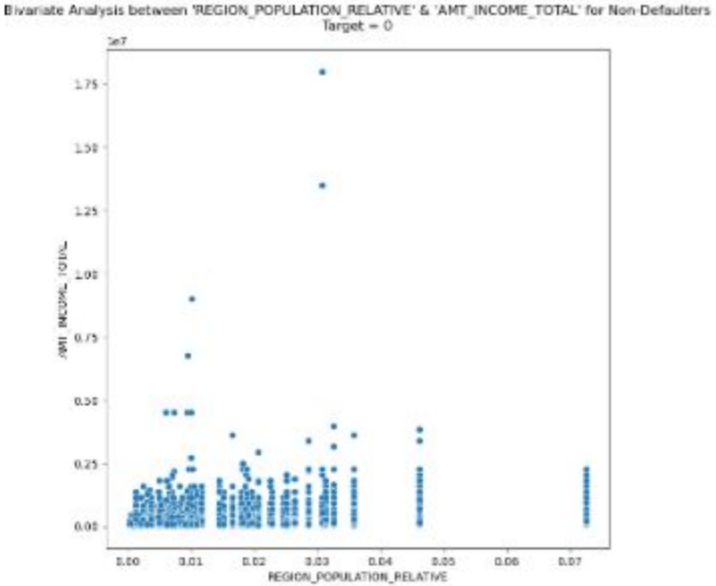




Insights -

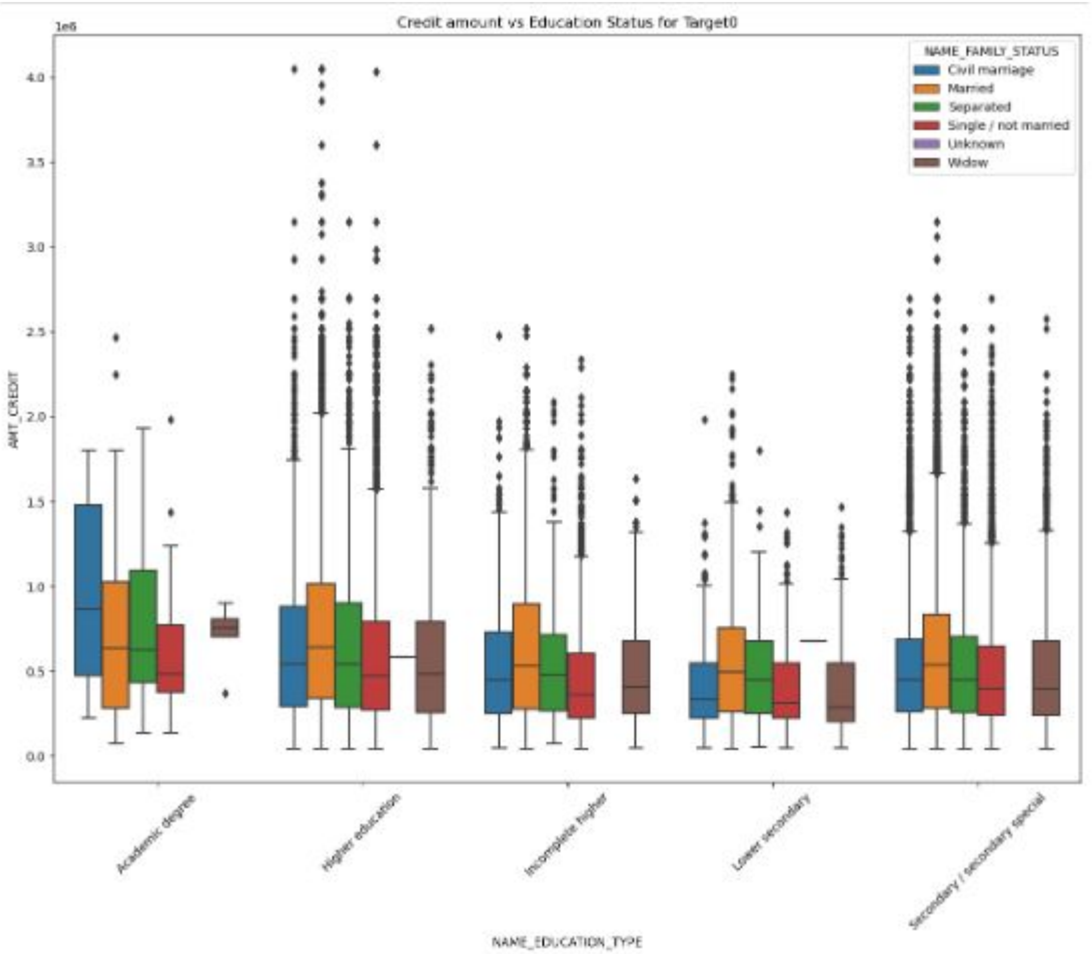
1. Defaulters have very low income where region population is less dense.
2. Non-defaulters have higher income than defaulters in the same population regions.





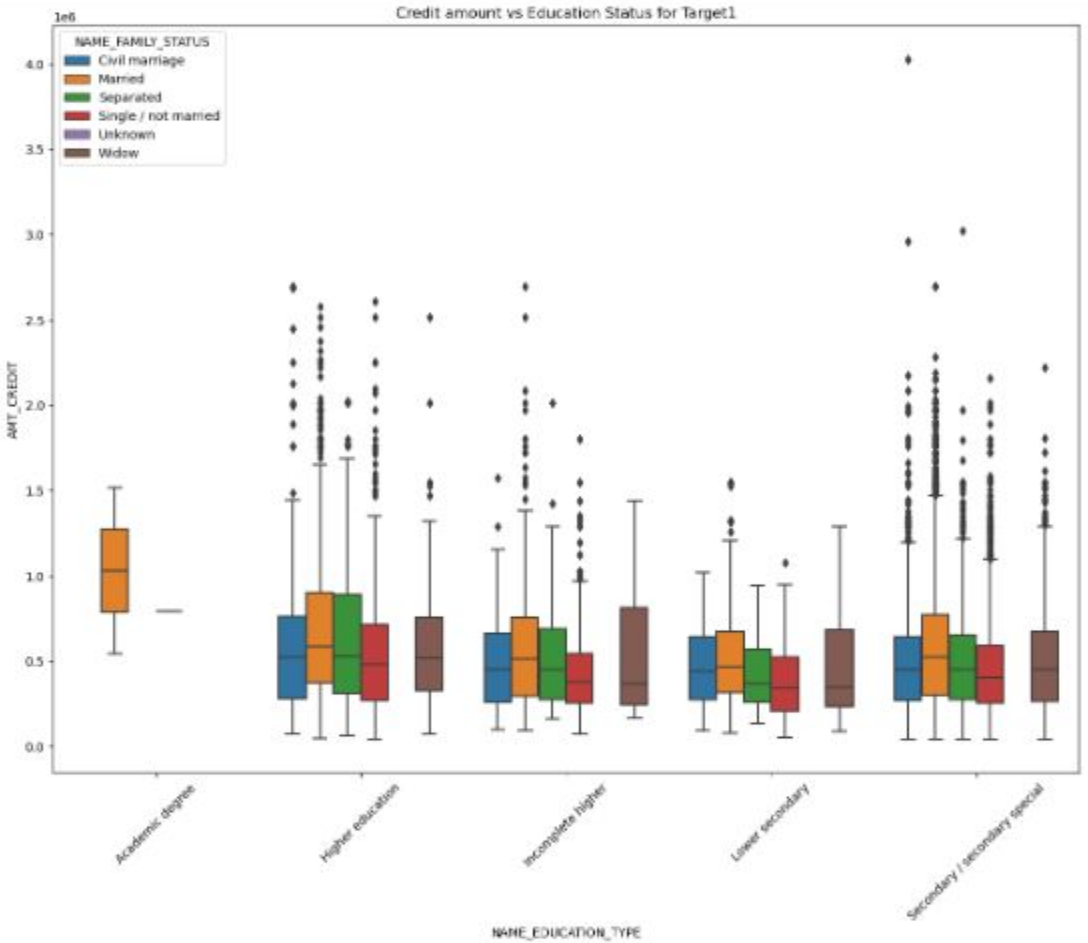
Insights - Credit amount vs Education Status for Target0

1. Family status with 'Civil Marriage','Separated' and 'Married' from the 'Academic Degree' have higher number of credits than others.
2. Also, 'Higher Education' has more number of outliers.
3. The 'Civil Marriage' family status in the 'Academic Degree' have most of the credits in the third Quartile (Q3).



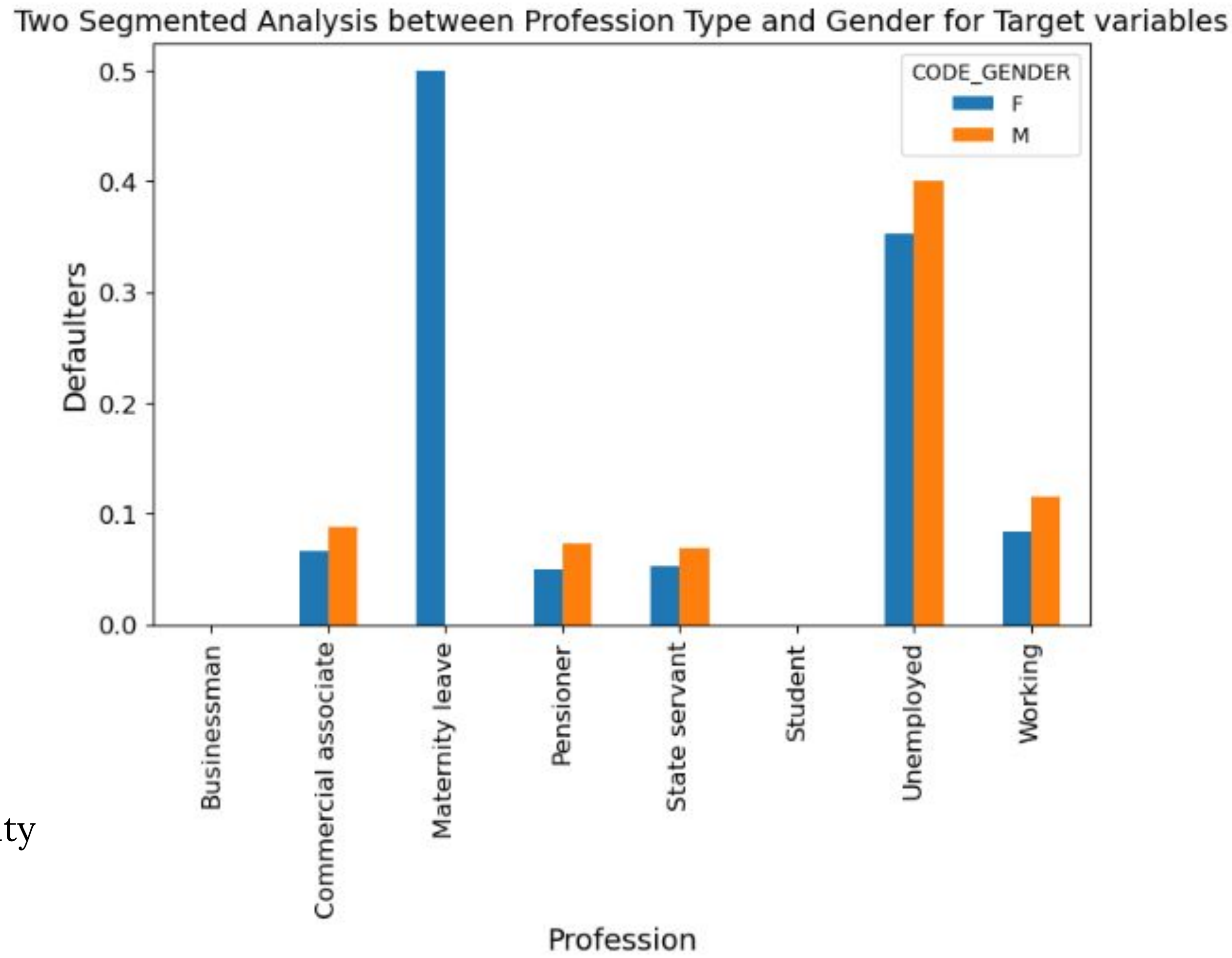
Insights - Credit amount vs Education Status for Target1

1. 'Married' family status of the 'Academic degree' education type has higher number of credits than others.
2. 'Secondary/Secondary Special' education type has higher number of outliers.
3. 'Academic degree' education type has no outliers and has no family status except for 'Married'.



Insights - Two Segmented Analysis

1. The 'Maternity leave' and 'Unemployed' clients are more defaulted.
2. The default rate is lesser in all other professions.
3. Males are more defaulted with their respective professions compared to females other than Maternity leave.

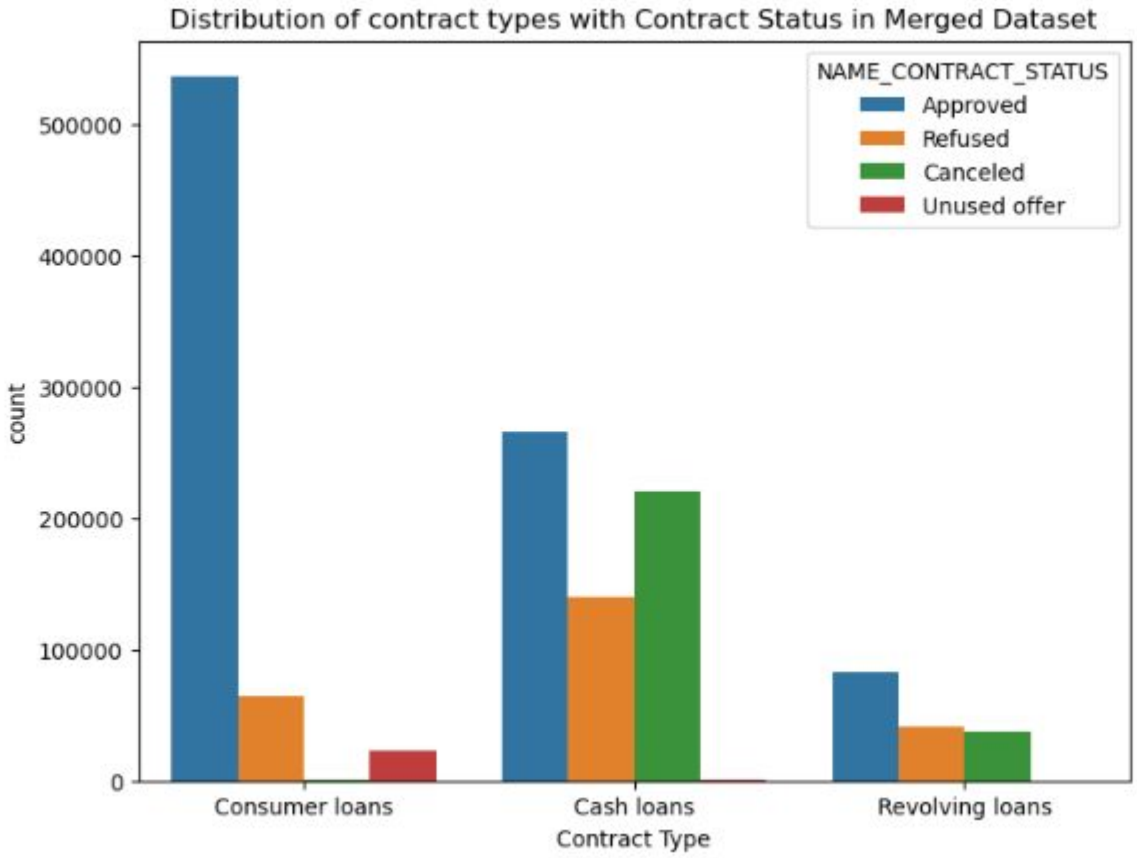


MERGED DATASET -

Univariate Analysis -

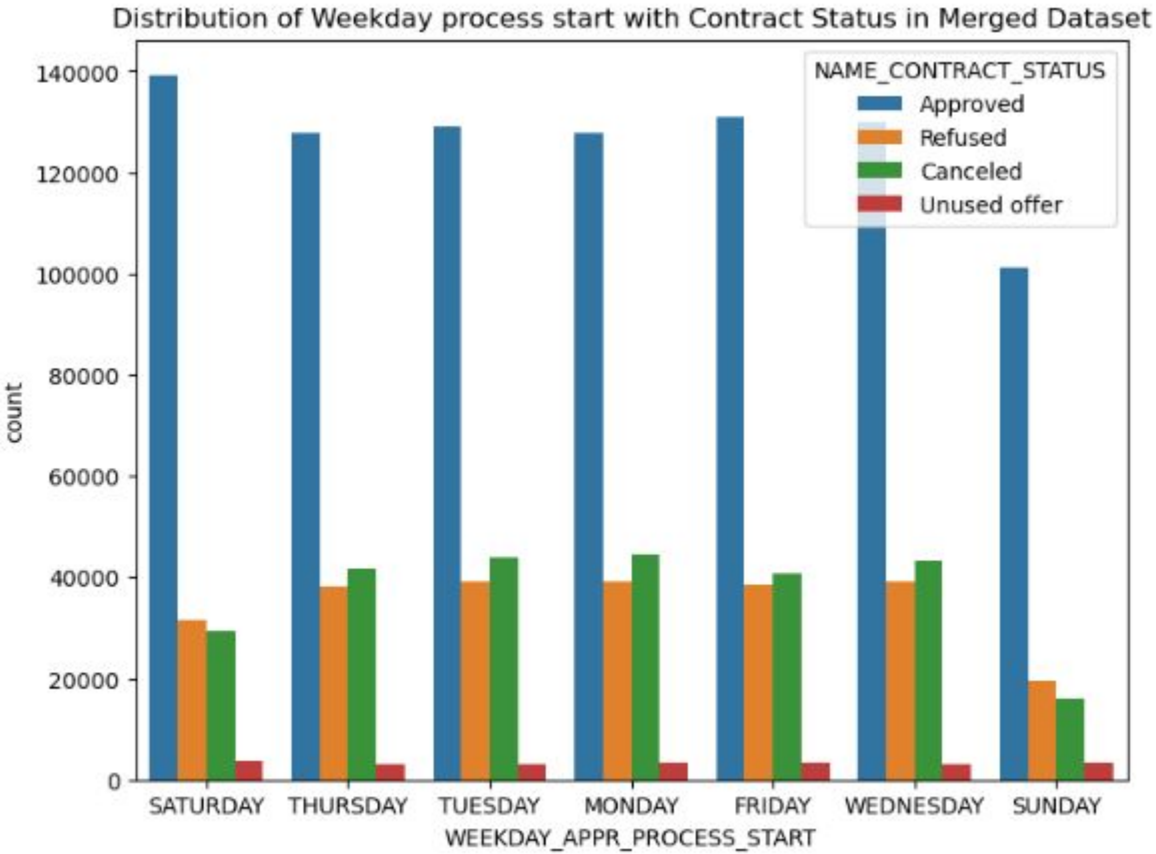
Insights -

1. 'Consumer' type of loan has the highest number of 'Approved' counts.
2. 'Consumer' type of loan has the least or almost none number of 'Cancelled' loan status.



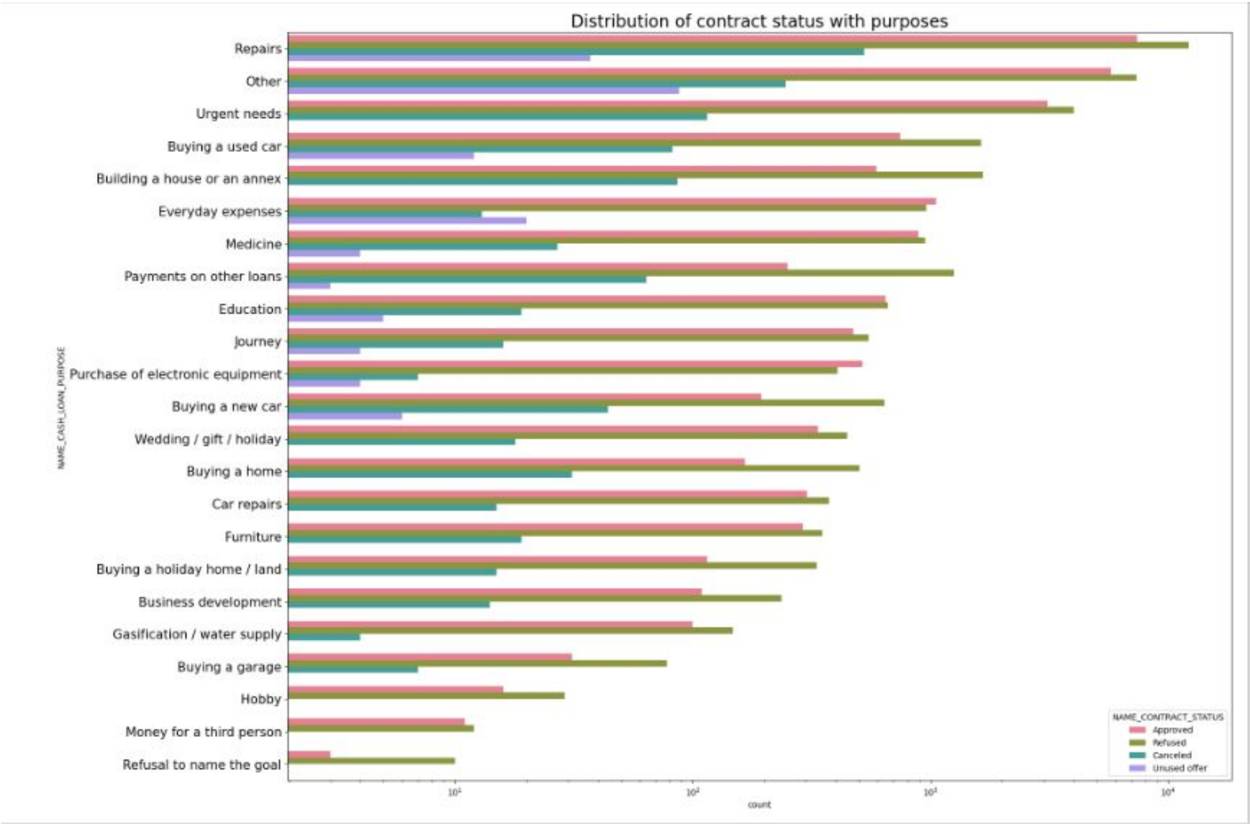
Insights -

Saturday has the highest Approval rate.



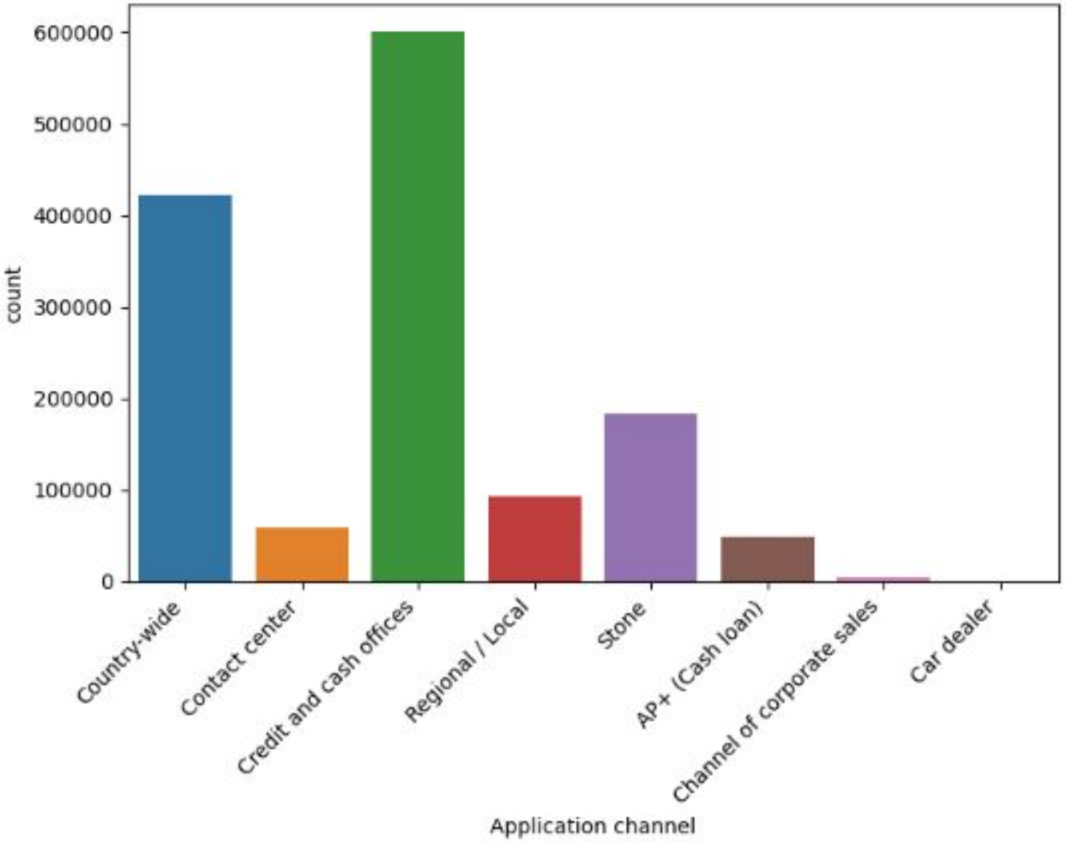
Insights -

1. There are higher number of 'Refused' counts for loans with 'Repair' purpose.
2. For 'Education' purpose, there are almost equal number of counts of 'Approved' and 'Refused'.
3. 'Payments on other loans' and 'Buying a new car' have higher Rejections than Approvals.



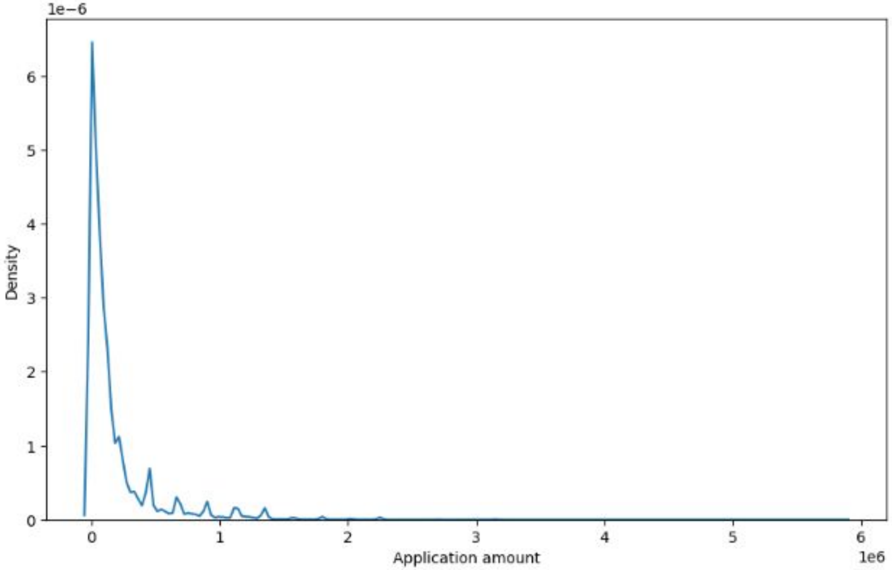
Insights -

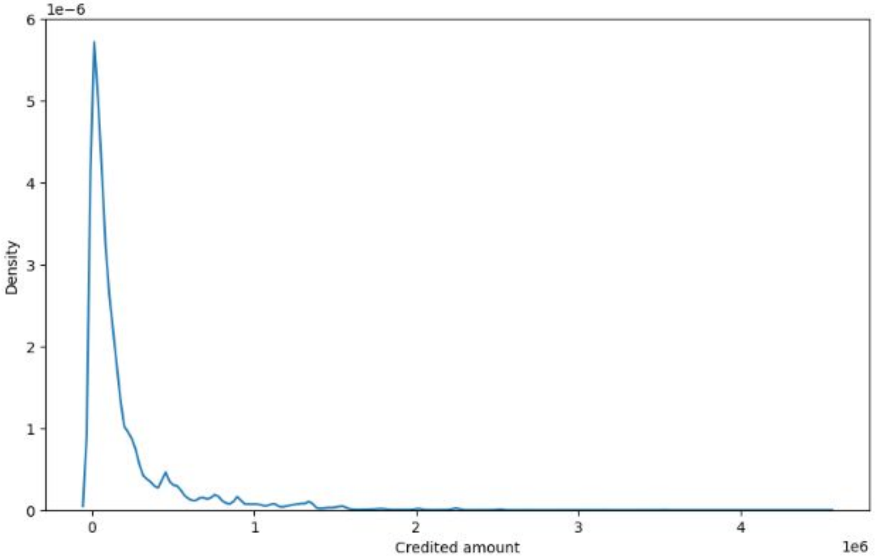
'Credit and Cash Offices' have higher number of counts.



Insights -

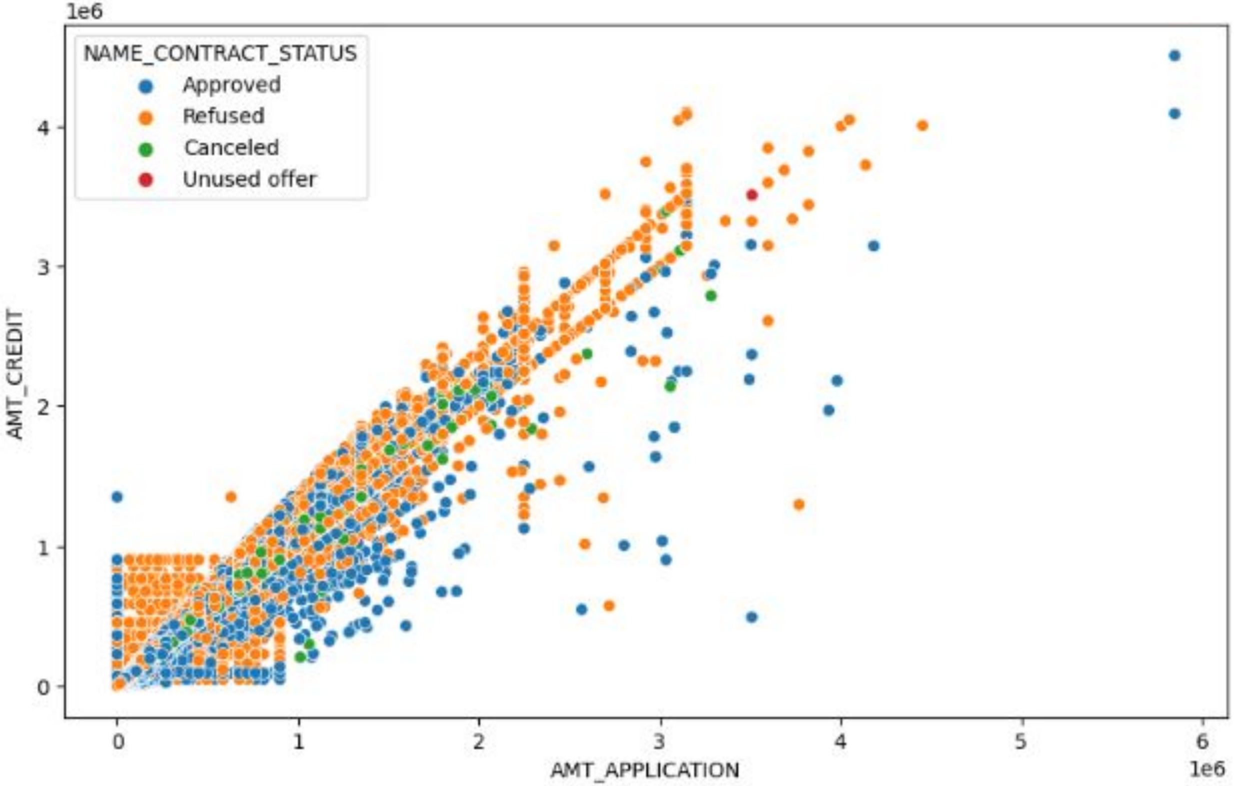
Plot 1 - Most of the applications were the amount of below 2000000 as we see from the above distribution plot.
Plot 2 - The distribution of the credit amount of loan was mostly in 2000000 range.





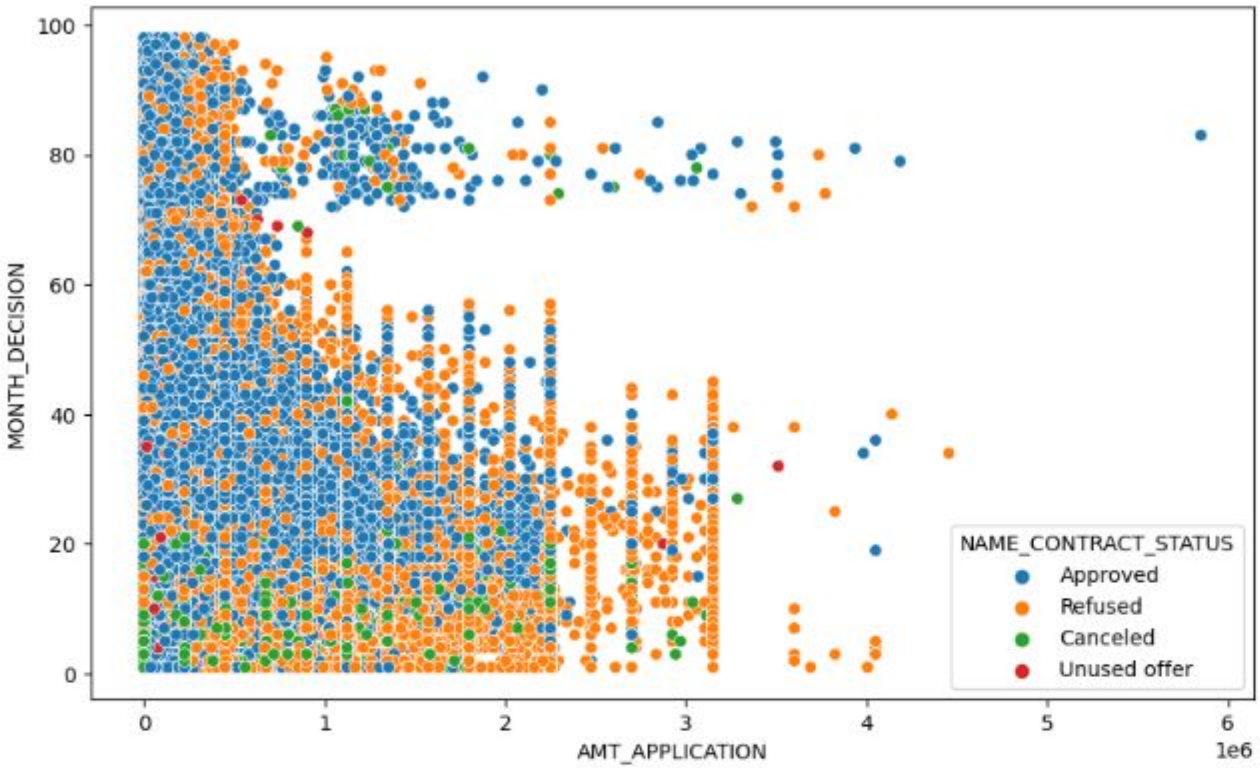
Insights -

1. The Amount Credited is increased with respect to the Application Amount.
2. The 'Refused' and 'Approved' status of the Applications for the Credit amounts are more in number.
3. There is a higher concentration in the lesser amount of applications and lesser amount of Credit.



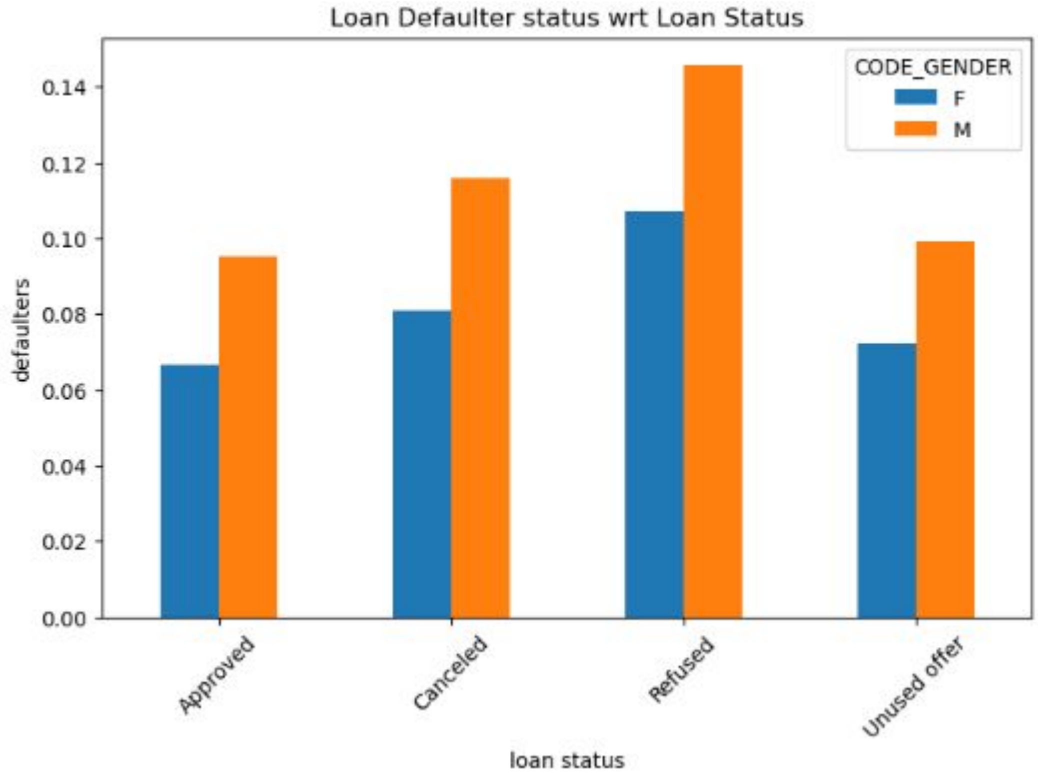
Insights -

1. There is a higher concentration for lower Application amounts.
2. More the amount of Application loans, lesser the Months Decision.
3. Most of the higher amount of loan Application Decision is made in the recent time.



Insights -

1. 'Refused' client is more defaulted than 'Approved' client.
2. 'Males' are more than 'Females'.



CONCLUSION

Banks should focus less on the following groups as they have higher unsuccessful payments rate -

1. Clients opting for cash Loans.
2. Clients with 'Secondary/Secondary Special' education qualification.
3. Clients who are 'Working' and 'UnEmployed' Professionals.
4. Clients who are married.
5. Clients who are 'Young', especially with 'Low' Income.
6. Loan purposes with 'Repairs' have more rejections.
7. Clients whose loans have been 'Refused', 'Cancelled'.

The following have successful payment rates and highly recommended groups -

1. Clients with 'Approved Consumer' loans
2. The Approval rates are higher on Saturdays
3. Clients with children
4. 'Tourism' goods category have highest success payment rate.
5. 'Non-cash' payment type has higher payment success.
6. 'Senior Citizen' age group have the most success payments.

Top Major variables to consider for loan prediction before approving application to minimize risk of loss:

1. NAME_EDUCATION_TYPE
2. AMT_INCOME_TOTAL
3. DAYS_BIRTH
4. AMT_CREDIT
5. DAYS_EMPLOYED
6. AMT_ANNUITY
7. NAME_INCOME_TYPE
8. CODE_GENDER

END