Credit EDA Case Study

By Aparna Sarkar (DS C70 batch)

Problem Statement

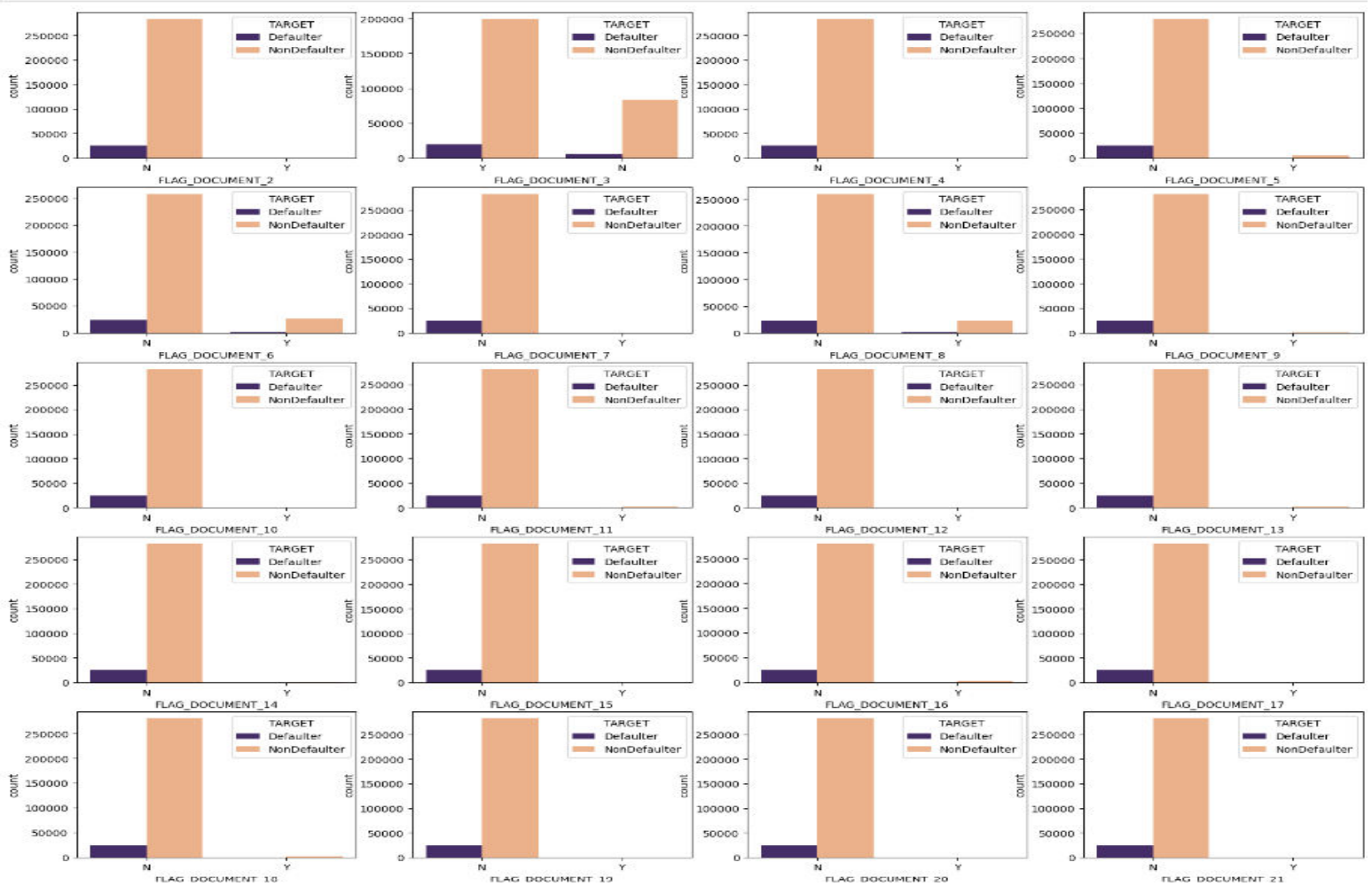
- This case study focuses on application of Exploratory Data Analysis to identify patterns in loan application data for a consumer finance company.
- The main objective is to balance the risk of loan defaults with the potential loss of business from declining deserving applications.
- By analysing the datasets, the goal is to gain insights regarding various factors influencing repayment ability , which would help the company take informed decisions that minimize the financial risks while maximizing the loan approvals for deserving customers.

Approach

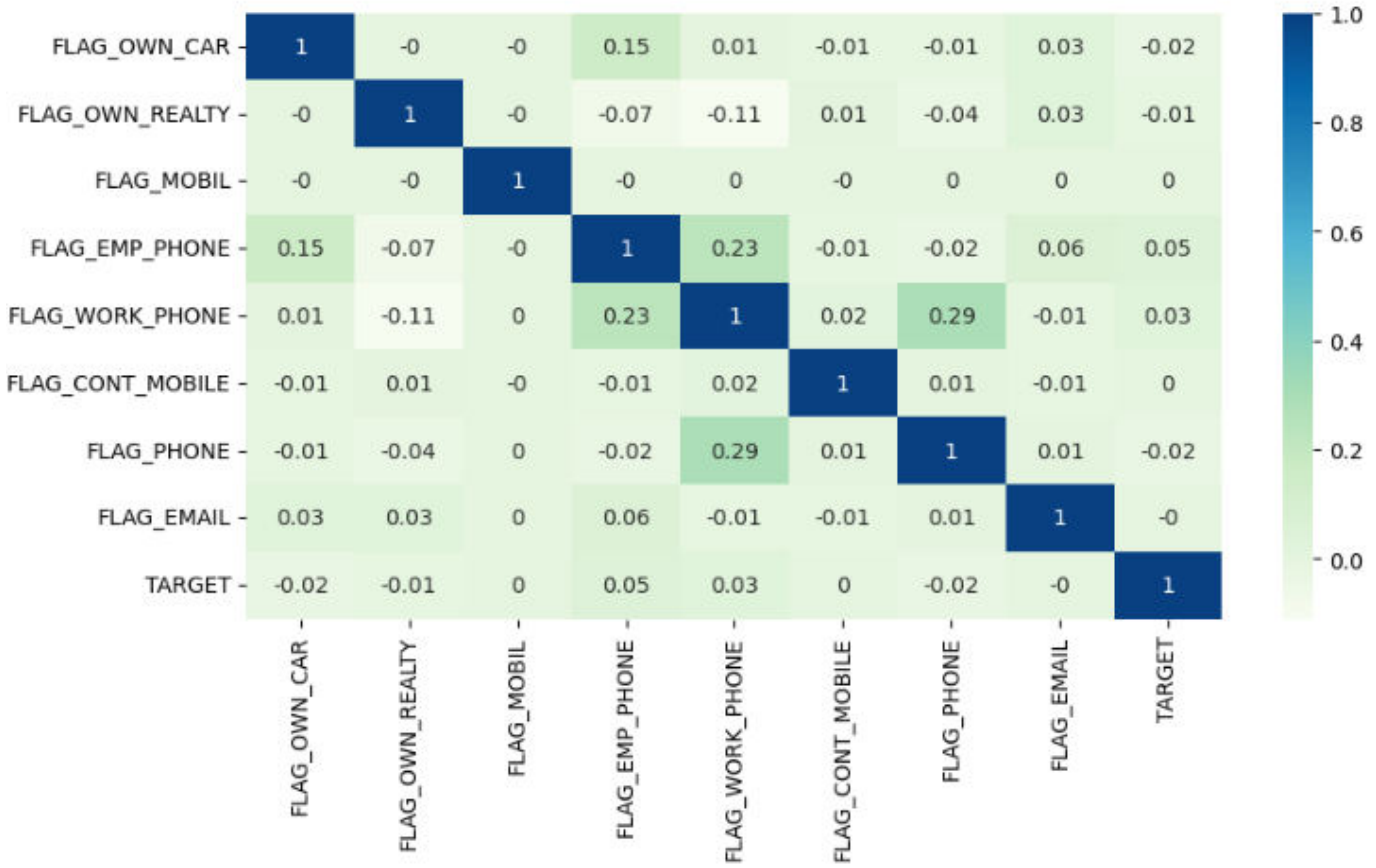
- There are two datasets given ‘application_data’ and ‘previous_application’ regarding loans.
- First, we need to assess the quality of data. We will identify the missing values, drop irrelevant columns, detect outliers and standardize the datasets. The ‘column_description’ file will help us understand the meanings of the columns.
- Once we ensure the datasets are reliable, we will examine data imbalances, perform univariate and bivariate analysis for gaining insights regarding the various factors influencing repayment of loans.
- Next, we will find out top 10 correlation for the Client with payment difficulties and all other cases
- Finally , we will merge both the datasets and use various plots to draw conclusions and make recommendations

Missing value Handling & Standardization

- For handling missing values, we will drop the columns having more than 40% missing values.
- Countplots and heatmaps have been used to understand the relevant flag columns and irrelevant ones have been dropped.
- For columns having missing values, missing value imputations have been performed. For imputation , mean , median or mode have been used wherever applicable.
- ‘AMT_REQ_CREDIT_BUREAU_%’ columns are filled with mode, OCCUPATION_TYPE has been filled with ‘Unknown’,’% CNT_SOCIAL_CIRCLE’ columns are filled with mode etc.
- The days columns were having negative values, they have been corrected as part of data standardization.



Plots to understand flag documents' relation with target variable



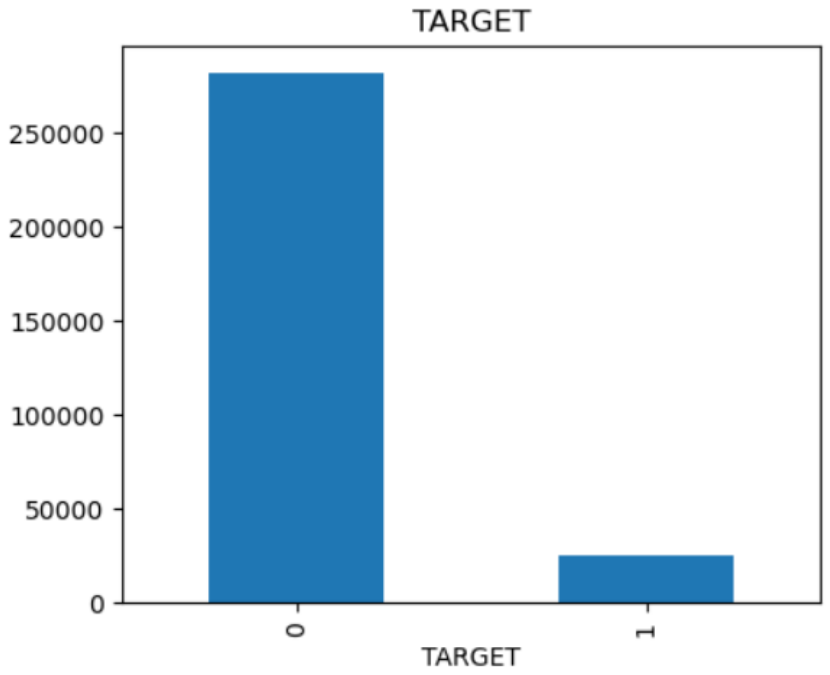
Heatmap to understand flag columns' correlations with target variable

Outliers Detection & Binning

- For understanding which columns might have outliers, unique counts have been checked.
- For each column having a very high count of unique values, percentile distributions, min ,max have been assessed and boxplots are also applied to understand outliers.
- AMT_GOODS_PRICE, AMT_INCOME_TOTAL, AMT_CREDIT has been converted to lakhs values as the values were very large.
- For columns having outliers, binning have been applied. E.g. New binned columns have been derived for AMT_GOODS_PRICE, AMT_INCOME_TOTAL, AMT_CREDIT etc.
- ´New Columns AGE, YEARS_EMPLOYED, YEARS_ID_PUBLISH have been derived like for better readability and binning has also been applied.

Data Imbalance Check

- Almost 92% of applicants are not defaulters whereas only 8% are defaulters
- Data Imbalance ratio is around 11%




```
TARGET
0    91.927118
1     8.072882
Name: proportion, dtype: float64
```

Other insights

- Almost 65% of applicants are Female whereas only 34% are male.
- Almost 90% of the loans are of cash type, whereas only 10% are revolving loans
- 70% of the loan applicants have 'Secondary / secondary special' level of education
- 63% of the loan applicants are married , only 14% are single.
- 88% of the loan applicants have a house or an apartment, and only 1% are living on rent.

```
CODE_GENDER
F     65.835694
M     34.164306
Name: proportion, dtype: float64
```

```
NAME_CONTRACT_TYPE
Cash loans         90.478715
Revolving loans     9.521285
Name: proportion, dtype: float64
```

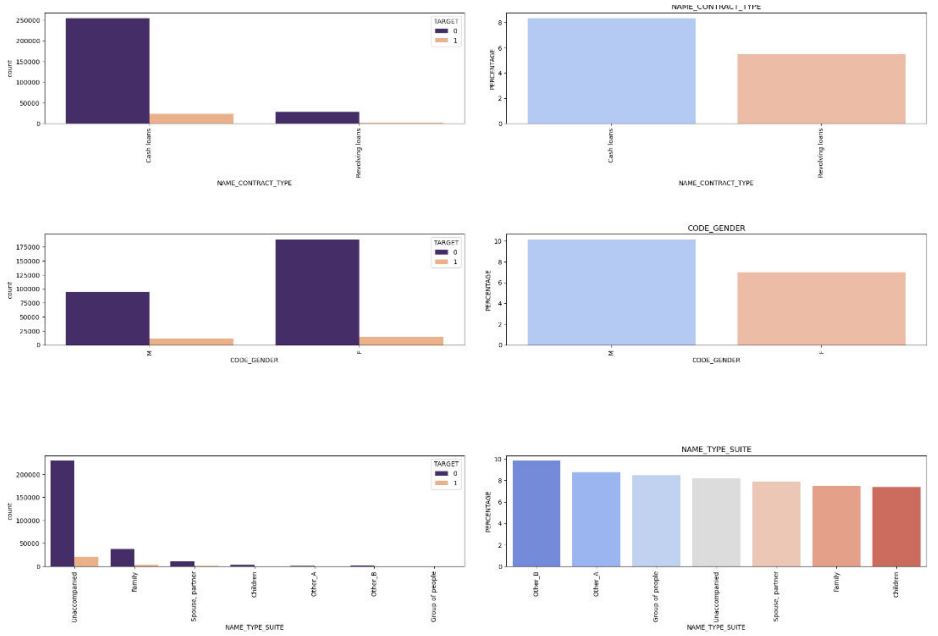
```
NAME_EDUCATION_TYPE
Secondary / secondary special    71.018923
Higher education                 24.344820
Incomplete higher                 3.341994
Lower secondary                   1.240931
Academic degree                   0.053331
Name: proportion, dtype: float64
```

```
NAME_FAMILY_STATUS
Married                 63.878040
Single / not married    14.778008
Civil marriage           9.682580
Separated                6.429038
Widow                    5.231683
Unknown                  0.000650
Name: proportion, dtype: float64
```

```
NAME_HOUSING_TYPE
House / apartment      88.734387
With parents            4.825844
Municipal apartment     3.636618
Rented apartment        1.587260
Office apartment        0.851026
Co-op apartment         0.364865
Name: proportion, dtype: float64
```

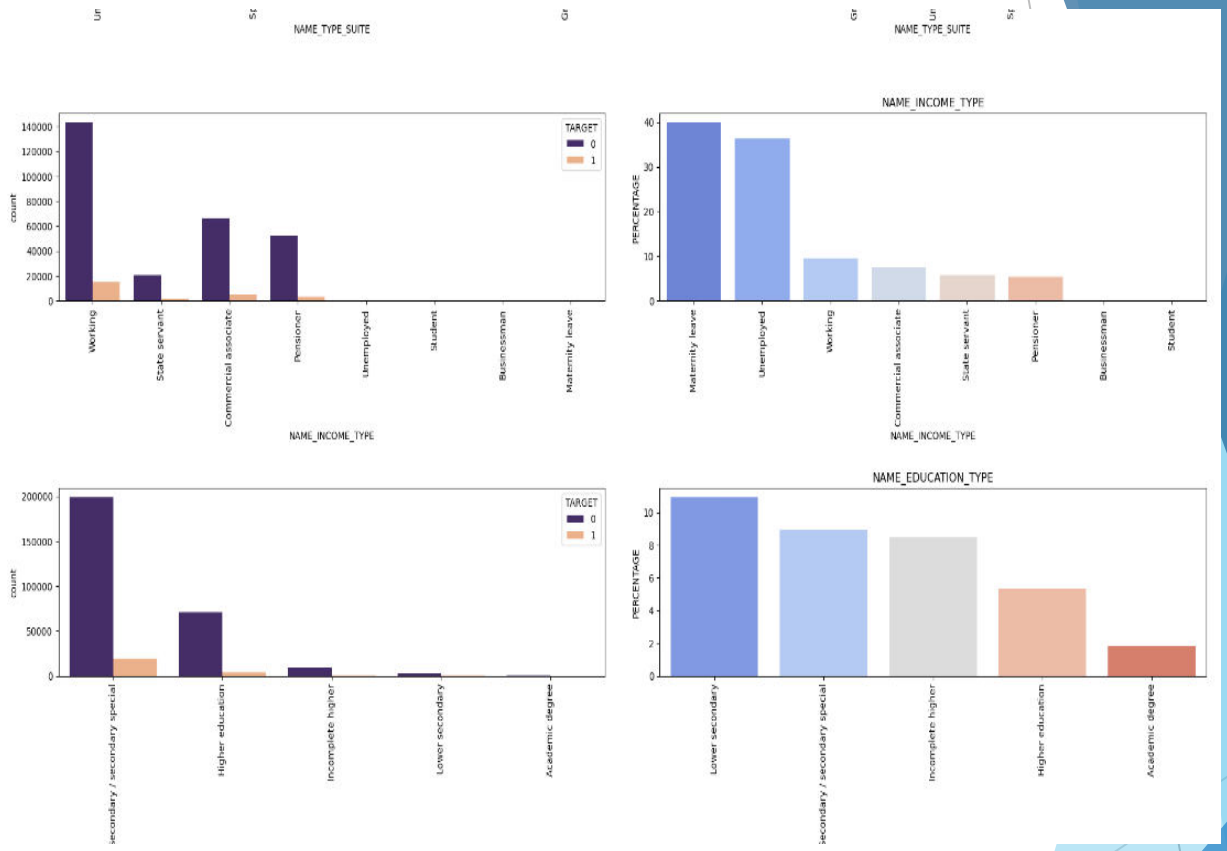
Categorical Univariate Analysis

- NAME_CONTRACT_TYPE : More clients have taken 'cash loan'. People who have taken cash loans are less likely to be defaulter considering the proportion.
- CODE_GENDER :Females have taken more loans. Females are less likely to be defaulters.
- NAME_TYPE_SUITE : Most of the people were unaccompanied while taking a loan, defaulting rate is around 9%.'The other B' category has high percentage of defaulters. The people accompanied by children are less likely to default, but the number of loans taken is very less.



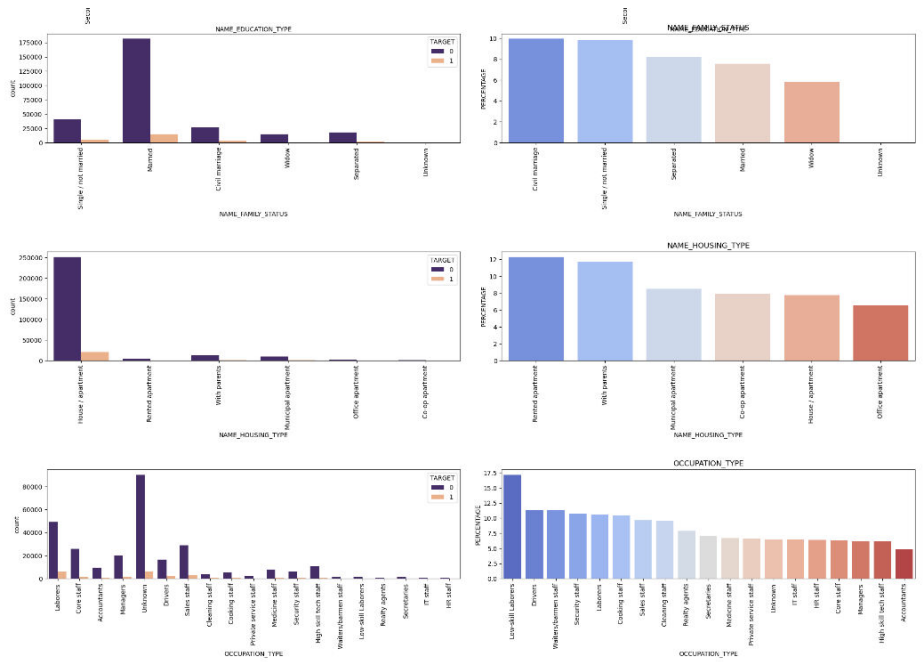
Categorical Univariate Analysis

- NAME_INCOME_TYPE : Most of the loan were given to working professionals, defaulting rate for working professionals is 10%. The people in maternity leave has high defaulting rate . Pensioners, Commercial Professionals have low default rates.
- NAME_EDUCATION_TYPE : Most of the loan were given to Secondary education. Academic degree and Higher education has lowest defaulting rate of less than 5%.



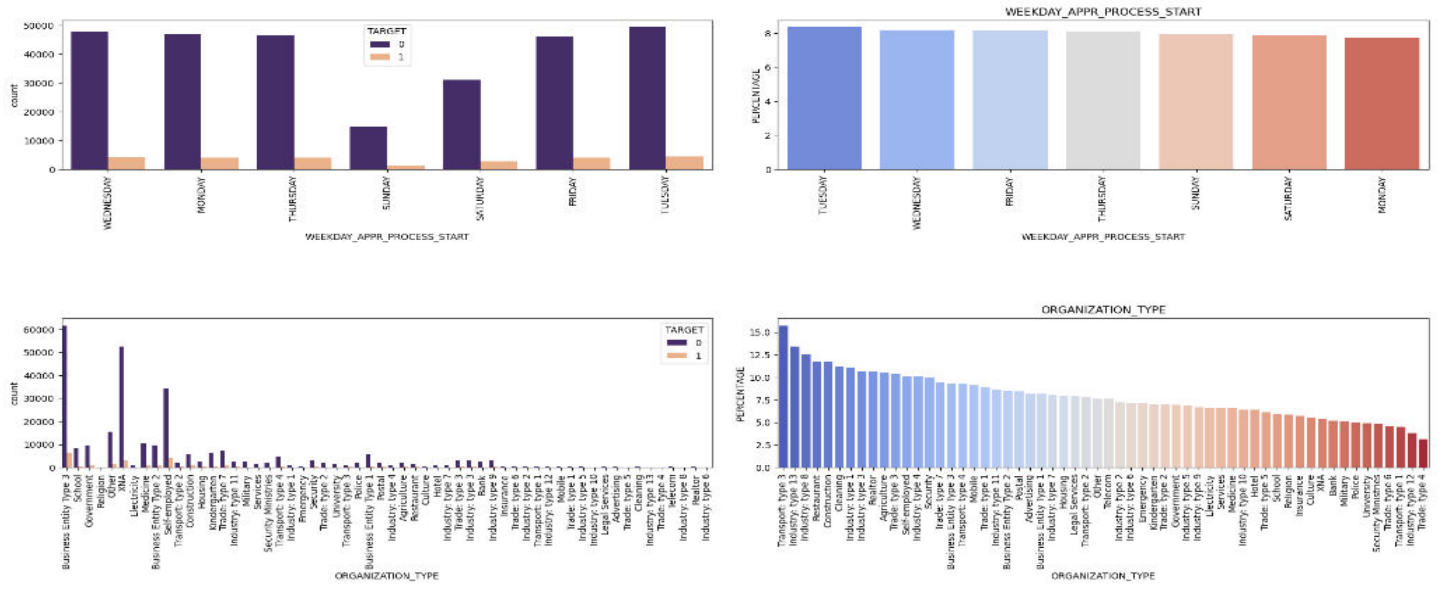
Categorical Univariate Analysis

- NAME_FAMILY_STATUS : Most of the loan has been given to married people . Civil marriage has the highest defaulter rate. Married people have lower defaulting rate (8%)
- NAME_HOUSING_TYPE : Most of the loans are given to house owners . Rented people has highest defaulting rate, whereas house owners , office apartment people have lower defaulting rate.
- OCUPATION_TYPE : Laborers have taken most of the loans. Low skilled laborers and drivers have highest default rate. Accountants , Core staff, Managers have low rate of defaulters.

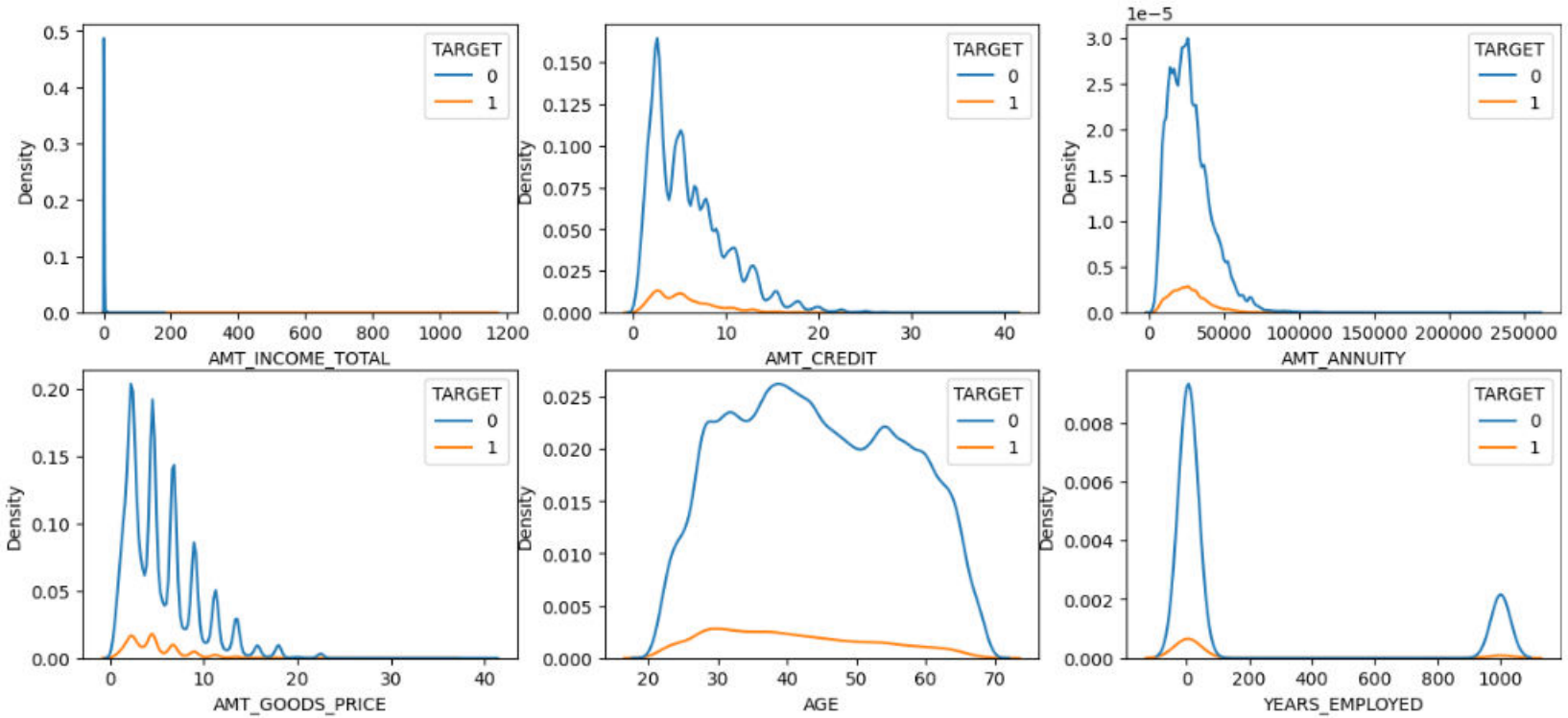


Categorical Univariate Analysis

- WEEKDAY_APPR_PROCESS_START : We can ignore this column as this wont make an impact on our analysis.
- ORGANIZATION_TYPE : Most of the people who have taken loan are working in 'Business Entity Type 3','Self Employed', 'Business Entity Type 2'.Transport Type 3 has highest rate of defaulters but number of loans



Numerical Univariate Analysis

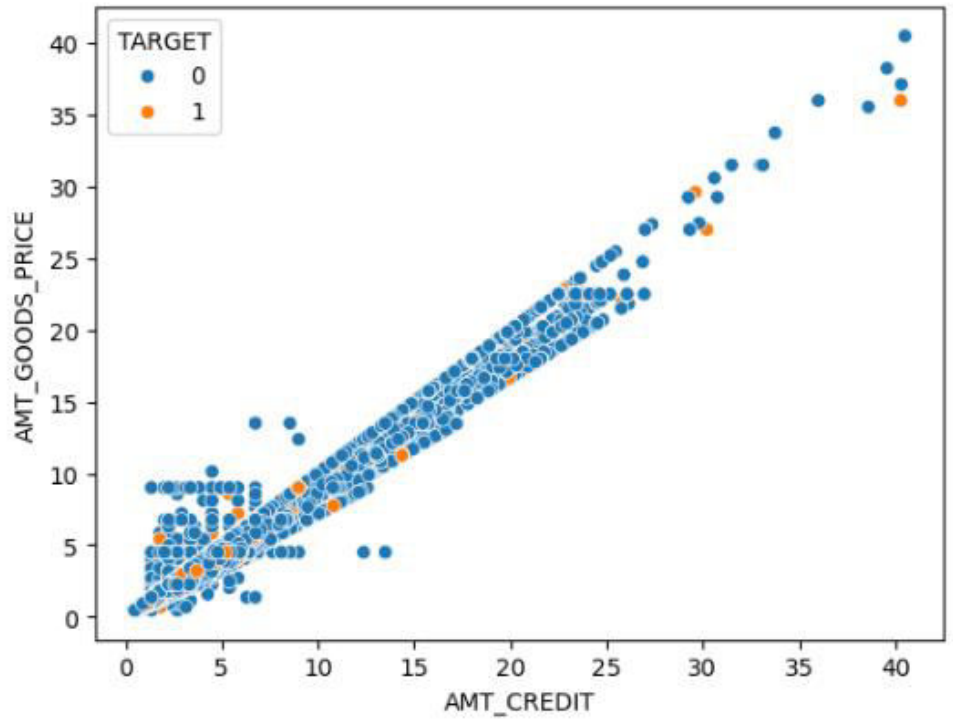


- ✓ Most clients have income between 0-1 million
- ✓ Most clients are paying annuity of 0-50,000.
- ✓ Mostly the loans given had product price ranging between 0-1 million.
- ✓ Amount credit also follows similar curve as expected (0 -1 million)
- ✓ Most clients have age between 20-70 range.
- ✓ There is not much to conclude on the defaulters.

Numerical Bivariate Analysis

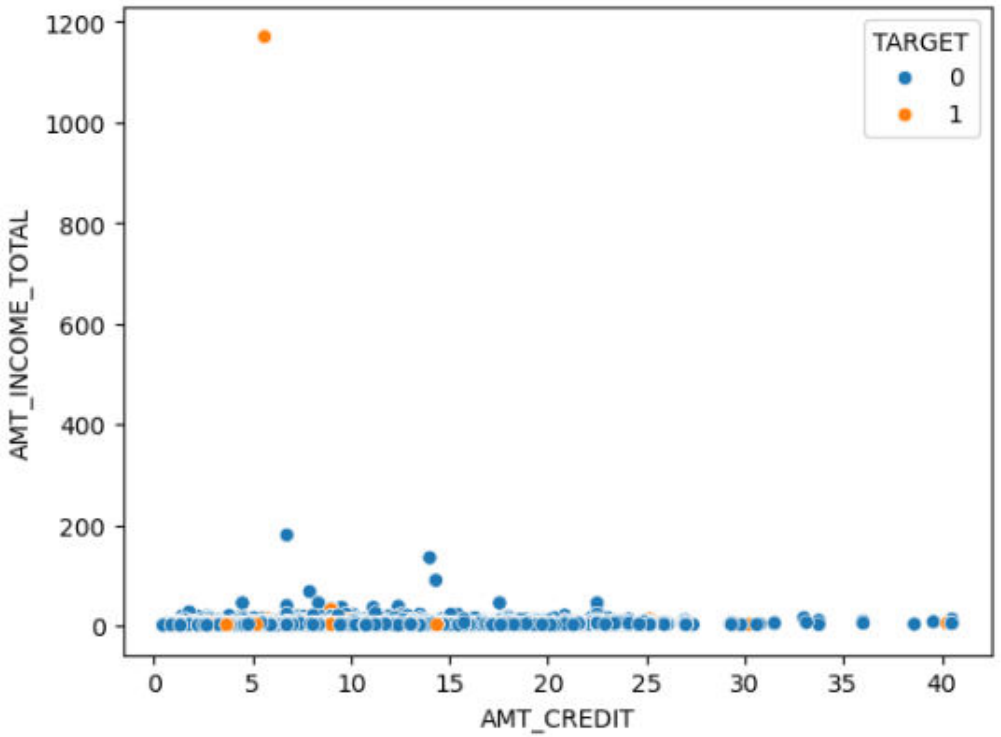
- As we have seen in univariate also, amount credit follows similar trend as amount goods price . We see here there is a perfect linear correlation .
- Most of the defaulters belong where amount credit is less than 1.5 million

Scatter plot of Credit Amount vs. Goods Price



Numerical Bivariate Analysis

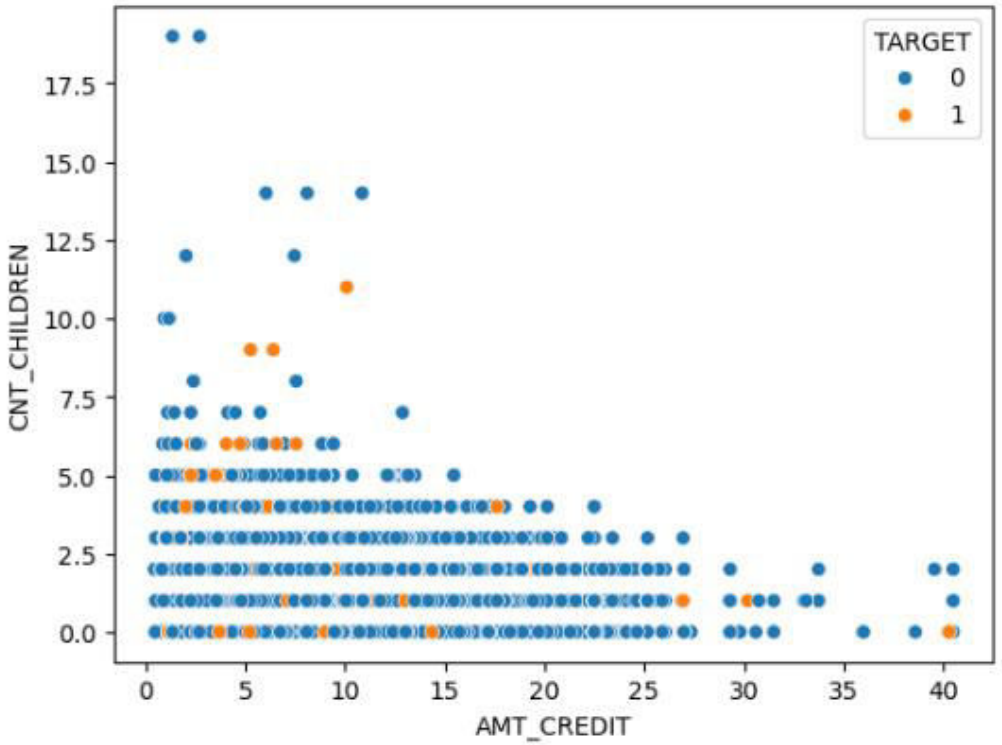
- People who are earning less are inclined to take loans
- Most of the defaulters belong where amount credit is less than 1.5 million



Scatter plot of Credit Amount vs. Income

Numerical Bivariate Analysis

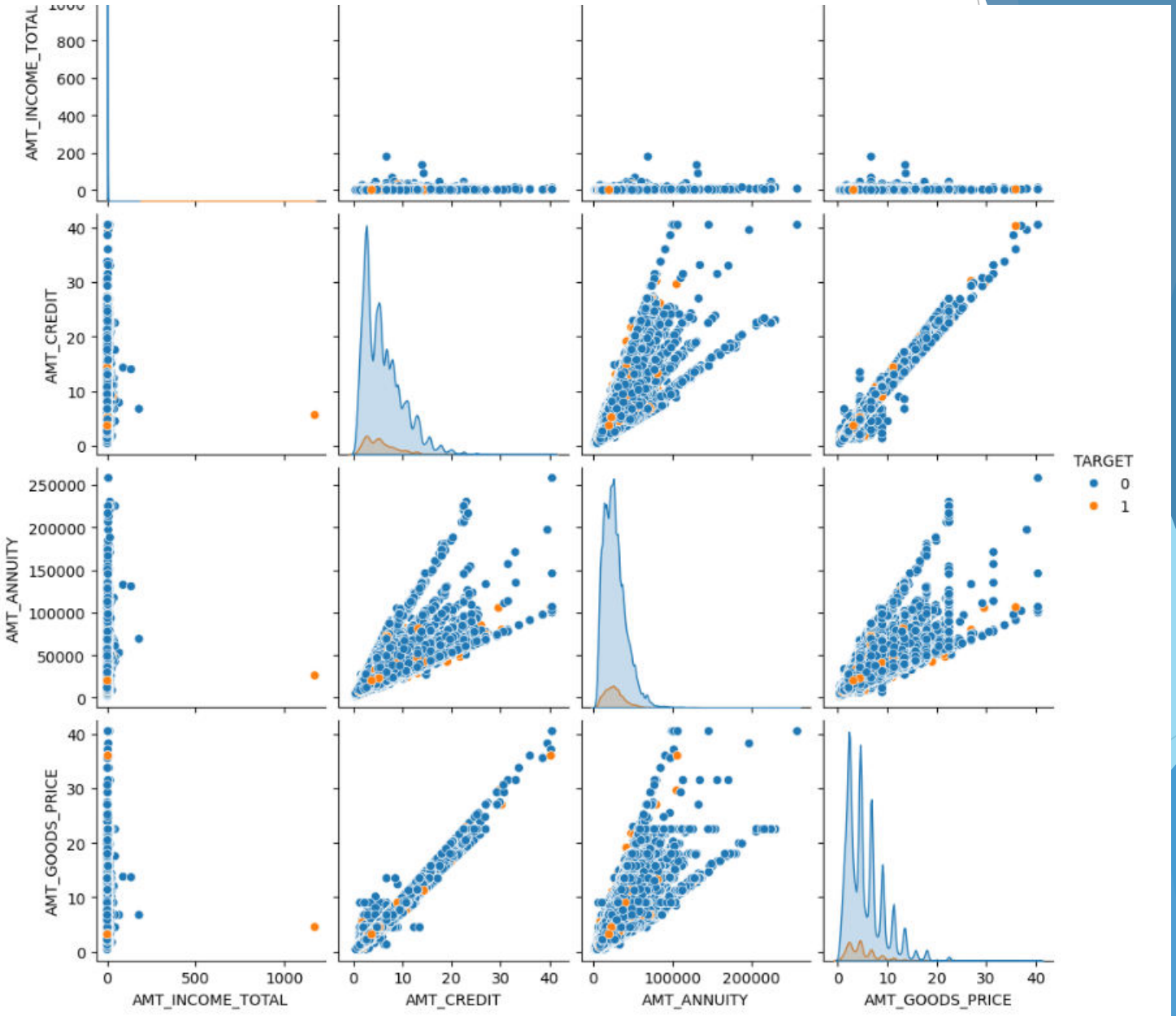
- People having children 1-3 have low number of defaulters.



Scatter plot of Credit Amount vs. Children Count

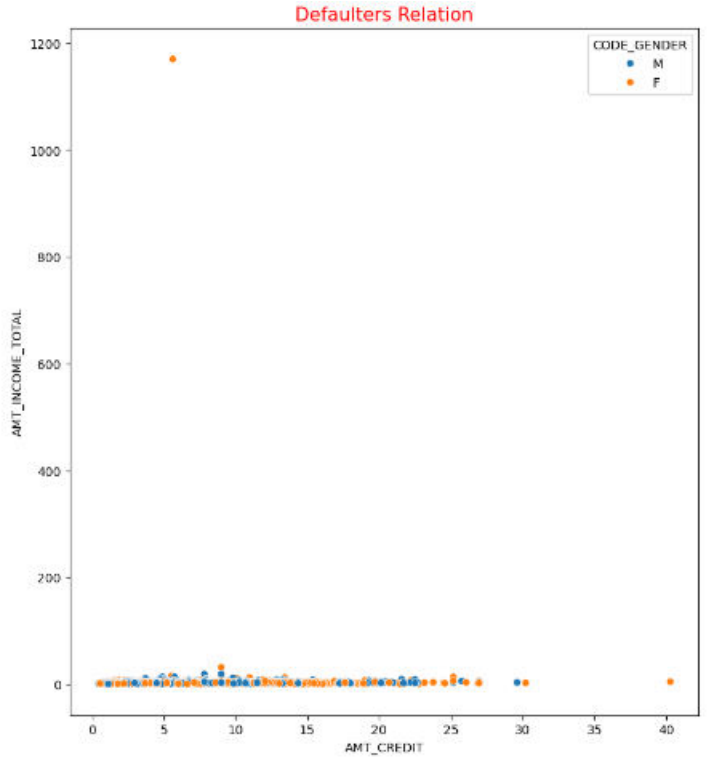
Numerical Bivariate Analysis

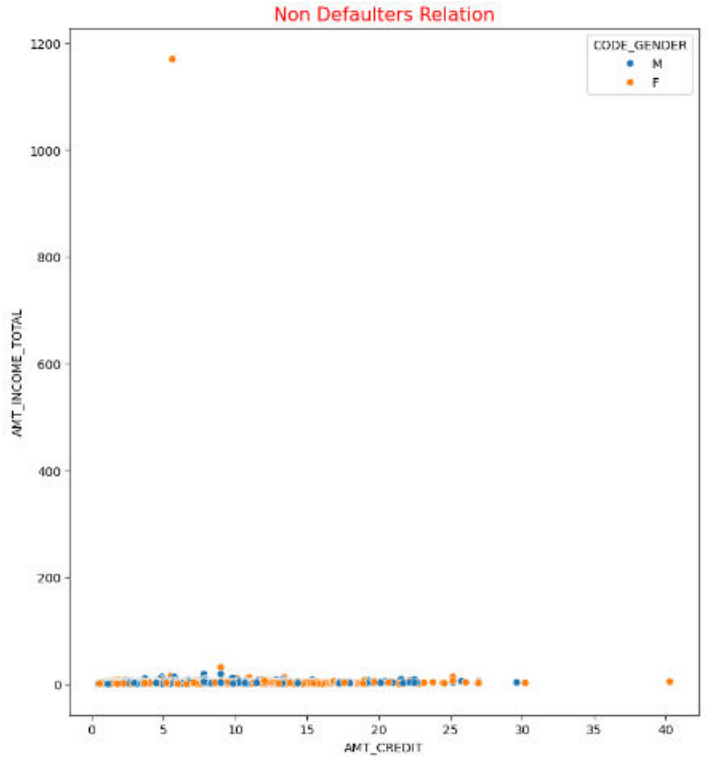
- Amount goods price and Amount credit has liner relation as we have previously seen.
- People who are paying annuity of 100000 are getting credit of around 2 million . This also looks linearly related.



Plots comparing defaulters and non-defaulters

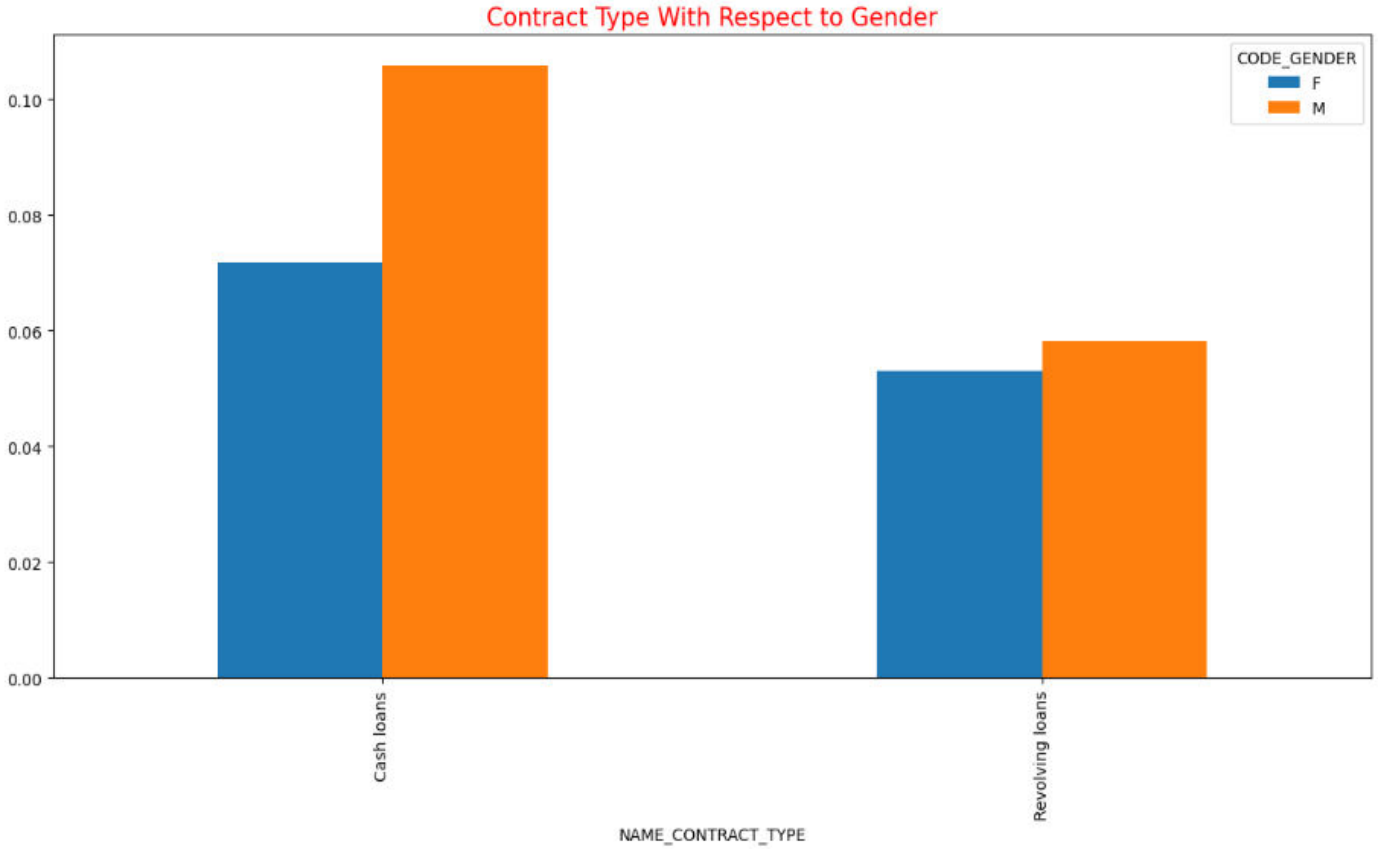
- For both defaulter and non defaulters values are most congested in the lower ranges.





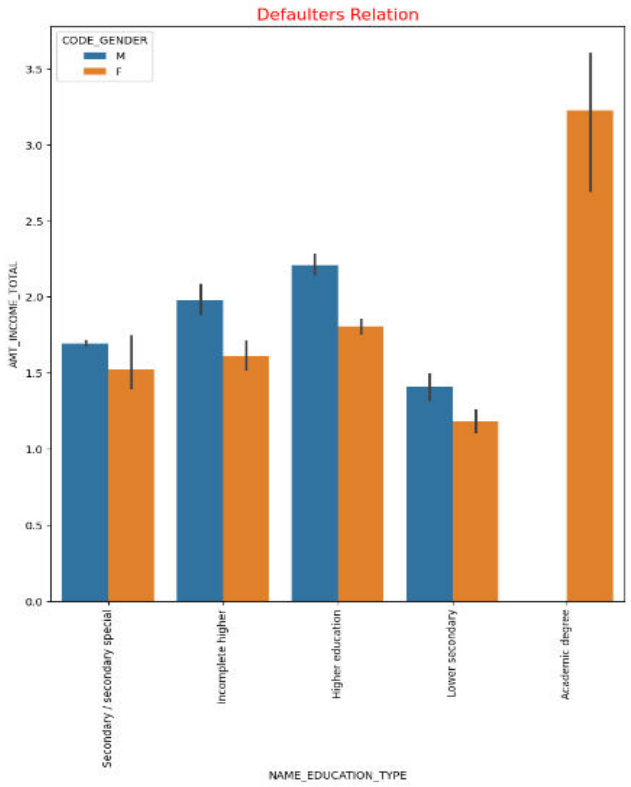
Plot showing Contract Type w.r.t Gender

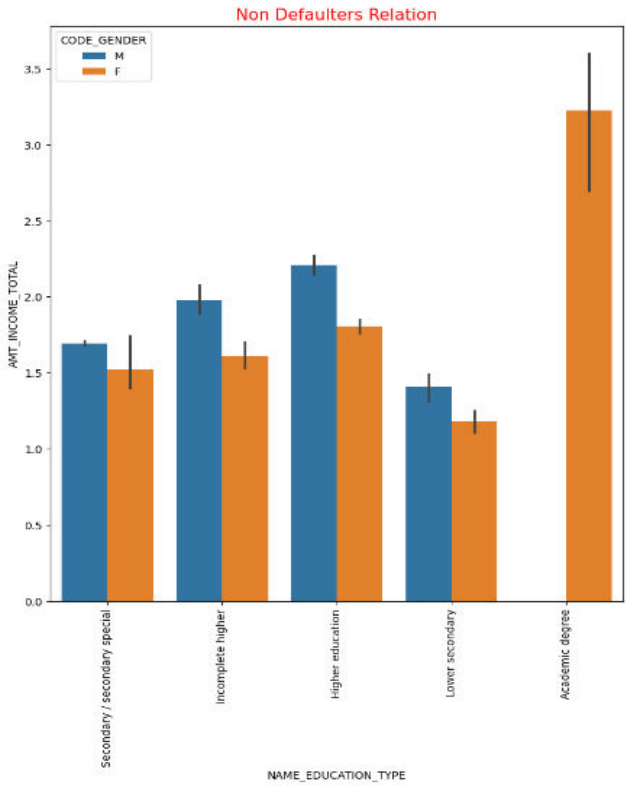
- From the plot we can see for both 'Cash loan' and 'Revolving loan' Males have taken more loans than Females.



Plots comparing defaulters and non-defaulters

- From the plots we can see that female academic degree holders are paid highest for both defaulter and non-defaulter category



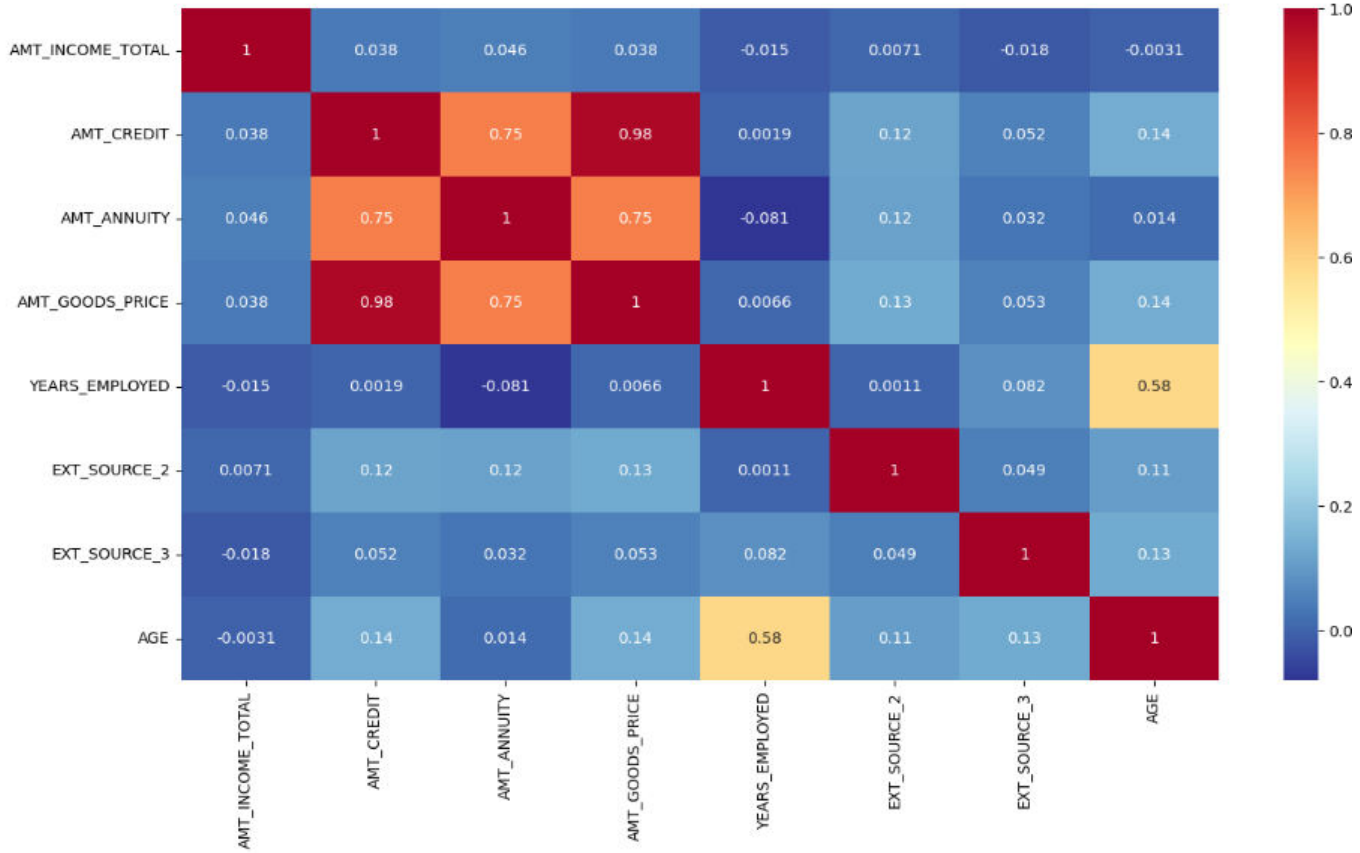


Finding top 10 correlations with target variable

- We can see the top 10 correlations with the target variables listed
- Next, we will see top 10 correlations with defaulters and non-defaulters

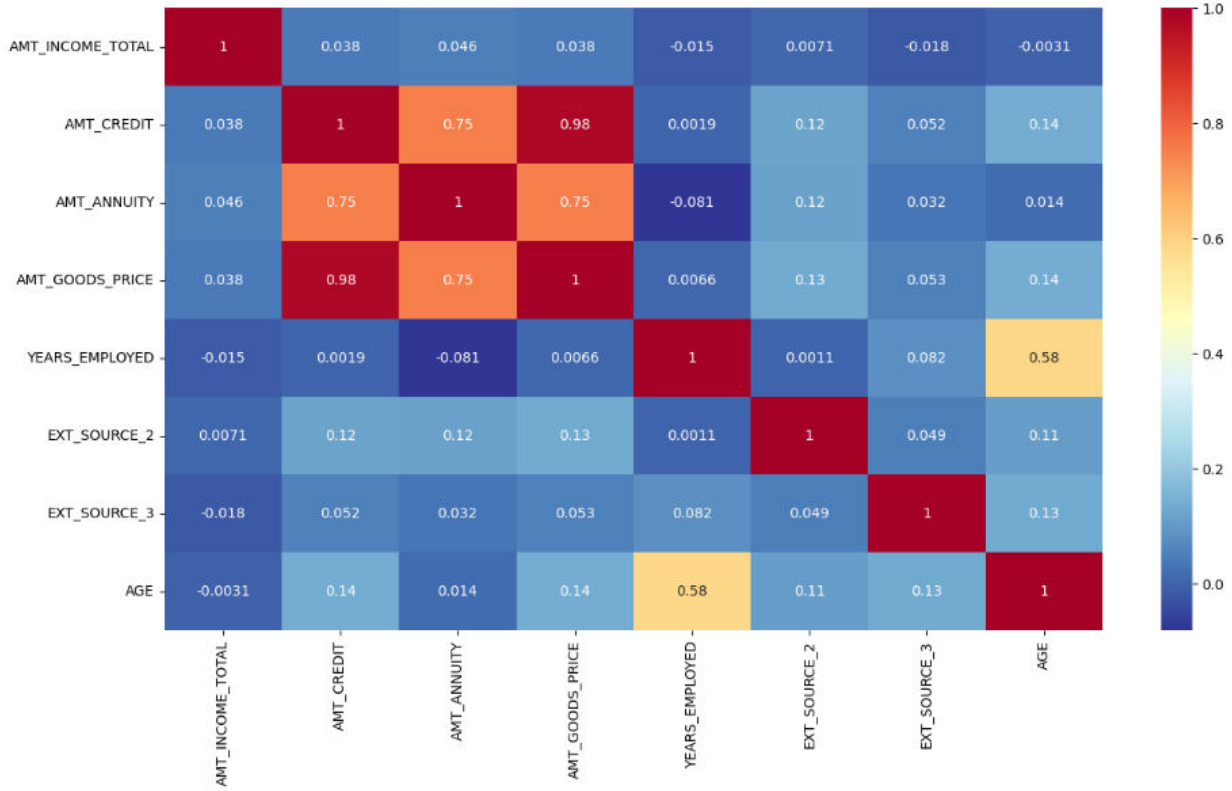
```
EXT_SOURCE_2                  0.160303
EXT_SOURCE_3                  0.157397
AGE                           0.078239
DAYS_BIRTH                    0.078239
REGION_RATING_CLIENT_W_CITY   0.060893
REGION_RATING_CLIENT          0.058899
DAYS_LAST_PHONE_CHANGE        0.055218
YEARS_ID_PUBLISH              0.051457
DAYS_ID_PUBLISH               0.051457
REG_CITY_NOT_WORK_CITY        0.050994
Name: TARGET, dtype: float64
```

Finding top 10 correlations with defaulters



1090	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998270
250	AMT_GOODS_PRICE	AMT_CREDIT	0.982783
587	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
494	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
1132	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869016
755	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
881	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
251	AMT_GOODS_PRICE	AMT_ANNUITY	0.752295
209	AMT_ANNUITY	AMT_CREDIT	0.752195
1629	YEARS_EMPLOYED	FLAG_DOCUMENT_6	0.617646
1239	FLAG_DOCUMENT_6	DAYS_EMPLOYED	0.617646
1567	AGE	DAYS_EMPLOYED	0.582185
377	DAYS_EMPLOYED	DAYS_BIRTH	0.582185
1607	YEARS_EMPLOYED	DAYS_BIRTH	0.582185

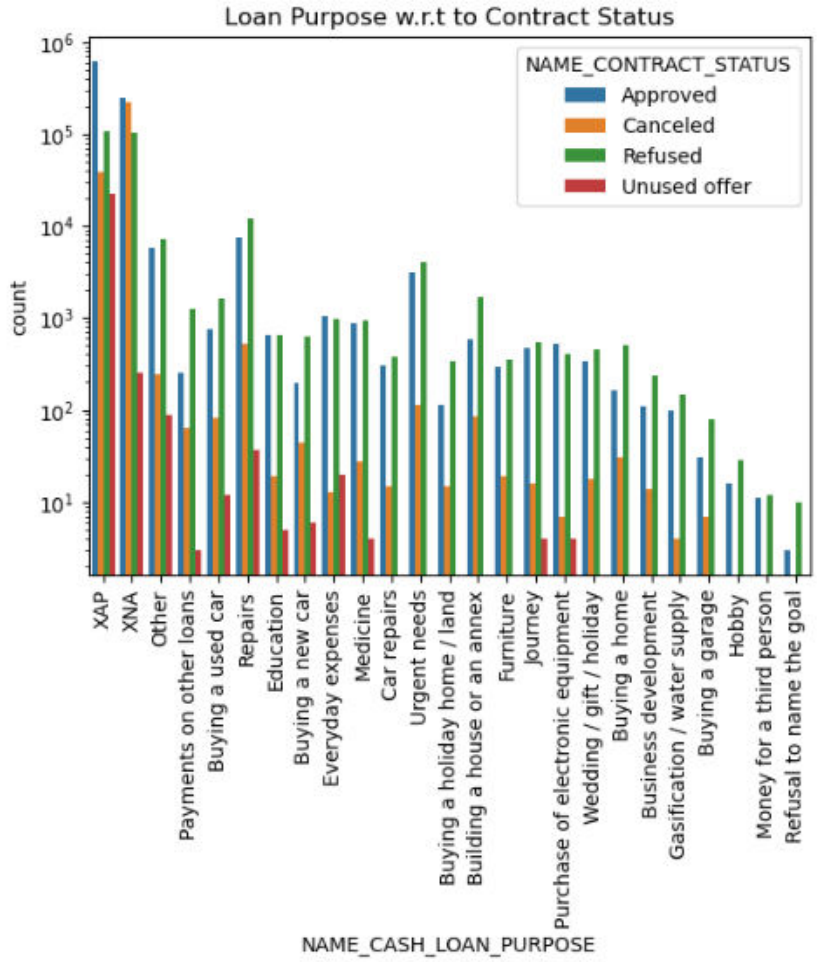
Finding top 10 correlations with non-defaulters



1090	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998510
250	AMT_GOODS_PRICE	AMT_CREDIT	0.987022
587	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
494	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
755	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
1132	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859371
881	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
251	AMT_GOODS_PRICE	AMT_ANNUITY	0.776421
209	AMT_ANNUITY	AMT_CREDIT	0.771297
377	DAYS_EMPLOYED	DAYS_BIRTH	0.626114

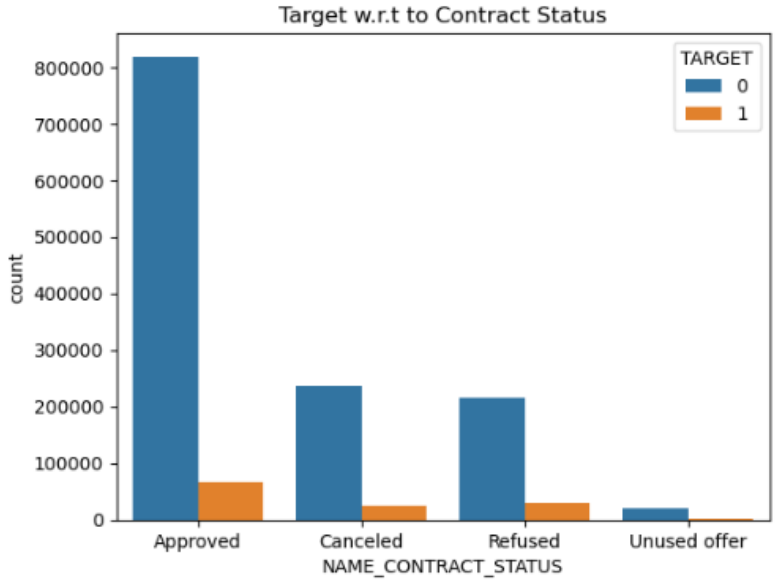
Visualisations

- After cleaning and standardizing the previous_application data and merging with application_data ,let us have a look at some plots which would help us derive some useful insights.
- ‘Repairs' category had the highest number of loan applications previously as well as highest amount of refusals and cancellations
- Most of the unused loan offers belong to 'others' category.



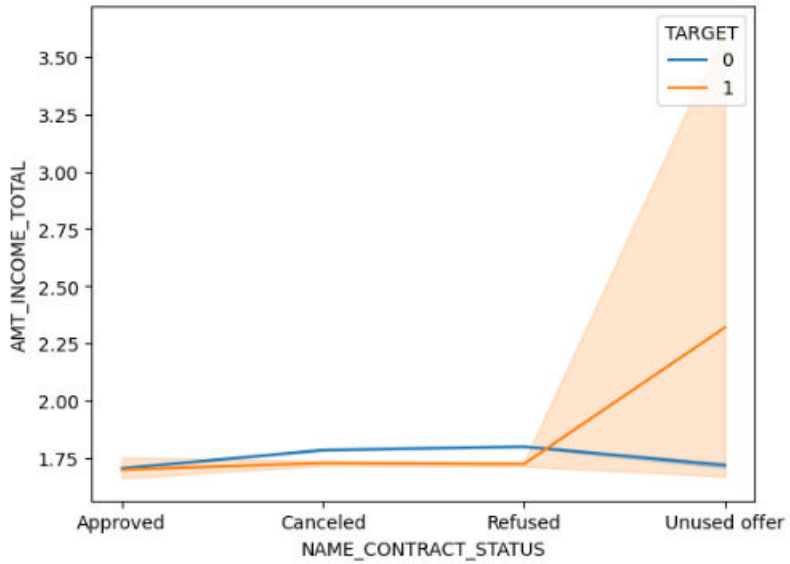
Insights

- Even if previously the loans were cancelled or refused , more than 80-90% of those were non-defaulters in current application So Bank should reanalyze those data to offer a loan .



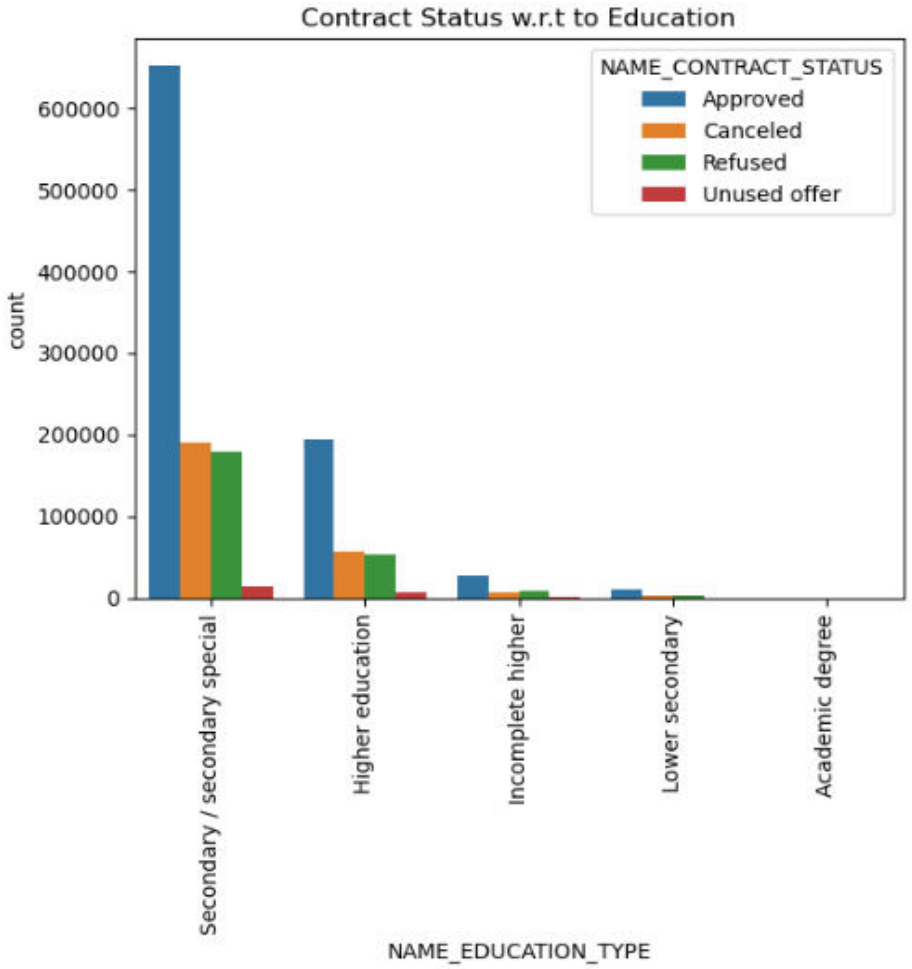
Insights

- The previous offers which were unused in current application have maximum number of defaulters despite having high income



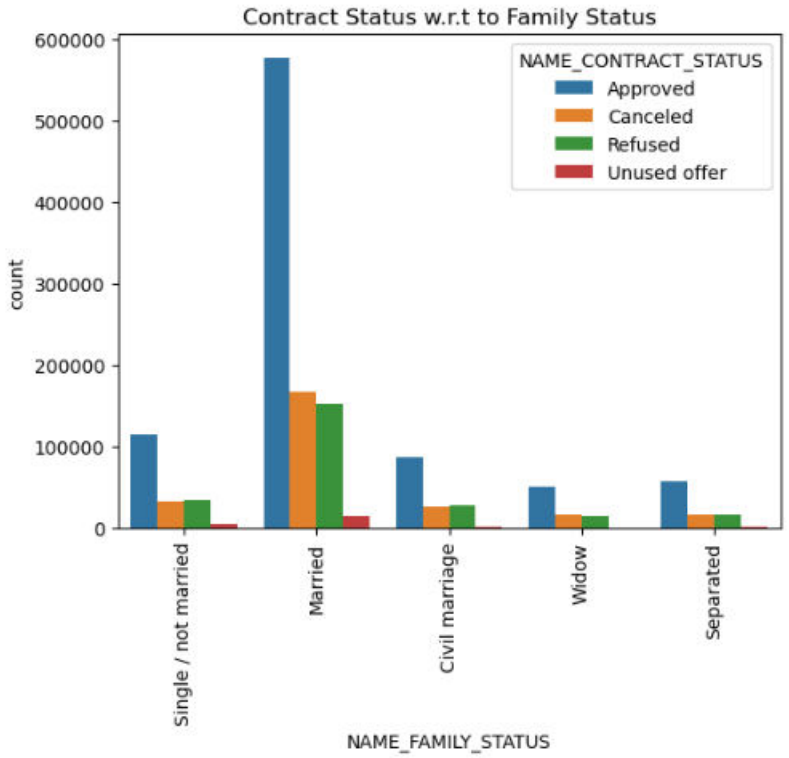
Insights

- Most of the loans were approved for Secondary education previously



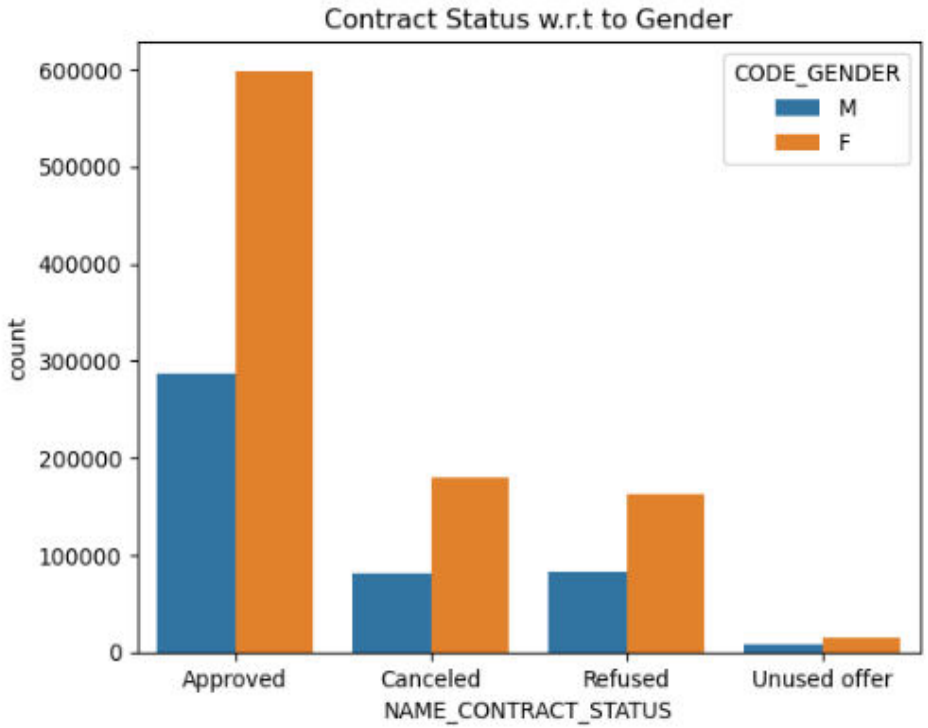
Insights

- Most of the approvals were given to married people previously.



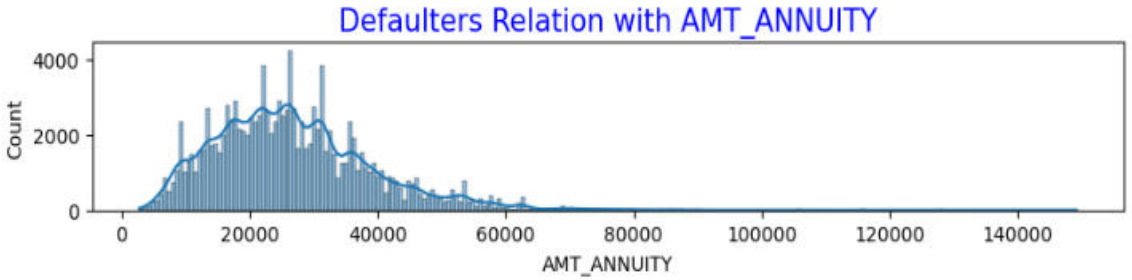
Insights

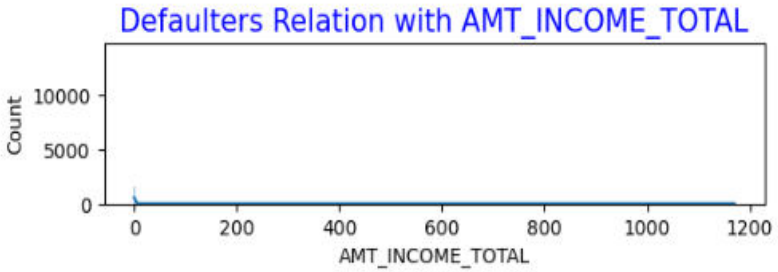
- Female contracts got approved more compared to male.
- Female contracts also got refused more compare to male but that is nearly negligible and Less in numbers compare to approval.

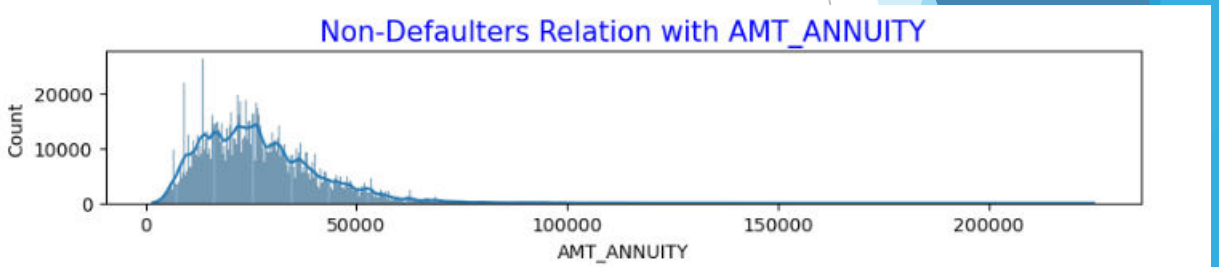


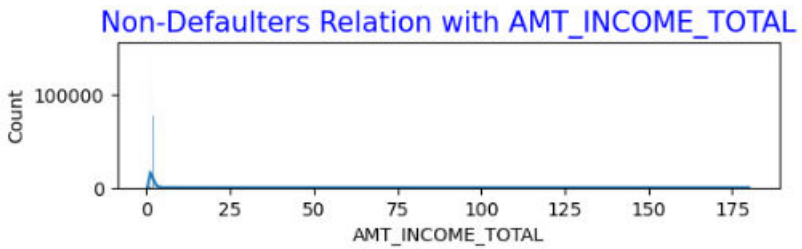
Insights

- For defaulters annuity lies between 0-50,000.
- Most of the non-defaulters have annuity less that 50k and income less that 1 million



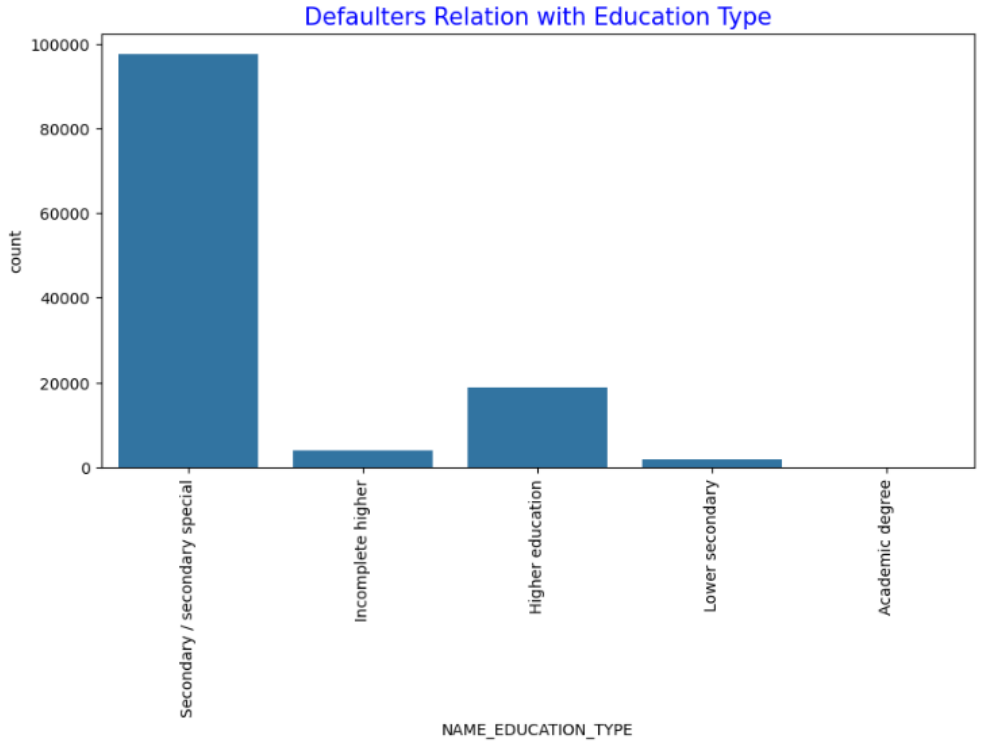






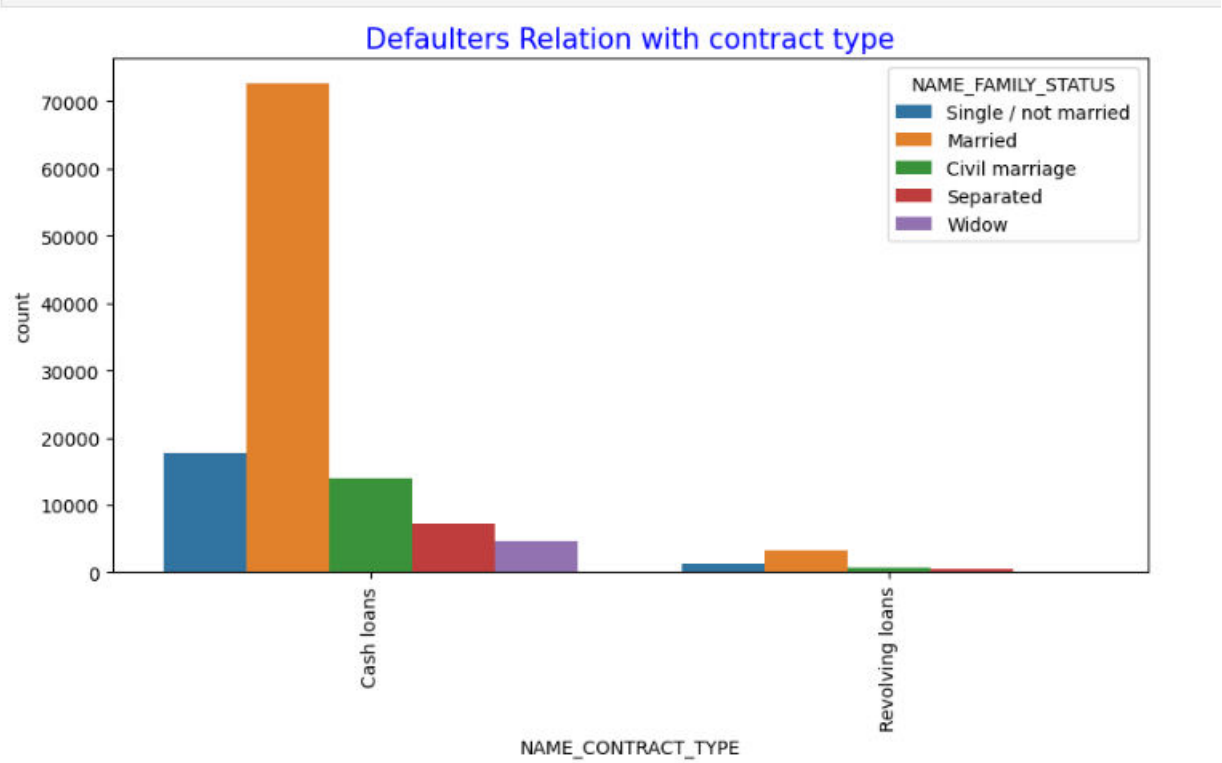
Insights

- Chances of Secondary/ secondary special people education type becoming defaulters is more.



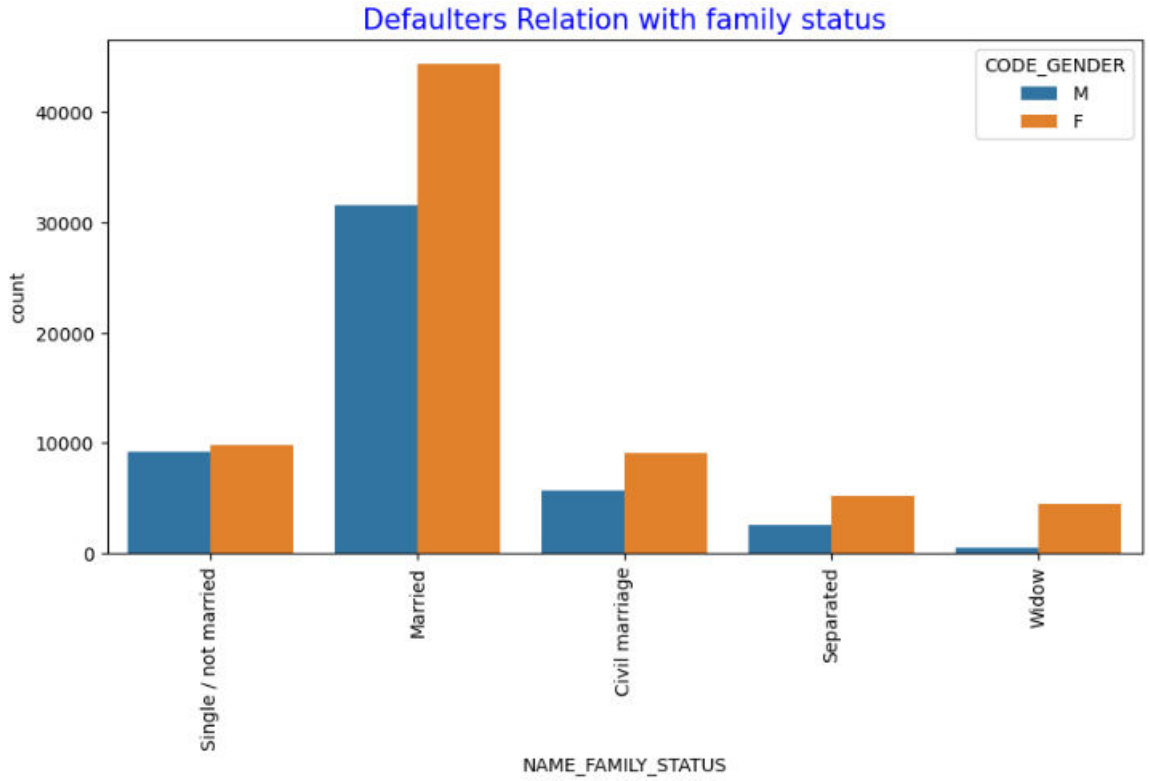
Insights

- Most of the defaulters are married and have taken cash loans



Insights

- Out of defaulters, married people are most in numbers.



Conclusions

- **NAME_CONTRACT_TYPE**: People who have taken cash loans are less likely to be defaulter considering the proportion.
- **CODE_GENDER**: Females have taken more loans . Females are less likely to be defaulters.
- **NAME_TYPE_SUITE**: Most of the people were unaccompanied while taking a loan, defaulting rate is around 9%.'The other B' category has high percentage of defaulters. The people accompanied by children are less likely to default, but the number of loans taken is very less.
- **NAME_INCOME_TYPE**: Most of the loan were given to working professionals, defaulting rate for working professionals is 10%. The people in maternity leave has high defaulting rate . Pensioners, Commercial Professionals have low default rates.
- **NAME_EDUCATION_TYPE**: Academic degree and Higher education has lowest defaulting rate of less than 5%.Female academic degree holders are paid highest for both defaulter and non-defaulter category
- **NAME_FAMILY_STATUS**: Civil marriage has the highest defaulter rate . Married poeple have lower defaulting rate
- **NAME_HOUSING_TYPE**: Rented people has highest defaulting rate, whereas house owners , office apartment people have lower defaulting rate.
- **OCCUPATION_TYPE**: Low skill laborers and drivers have highest default rate. Accountants , Core staff, Managers have low rate of defaulters.
- **ORGANIZATION_TYPE**: Clients with Trade Type 4 and Industry type 12 have less number of defaulters . Transport Type 3 has highest rate of defaulters but number of loans given were low.
- **CNT_CHILDREN**: People having children 1-3 have low number of defaulters.
- **AMT_CREDIT** : Most of the defaulters belong where amount credit is less than 1.5 million
- **AMT_INCOME_TOTAL** : Most of the non-defaulters have annuity less that 50k and income less that 1 million
- **NAME_CONTRACT_STATUS** : More than 80-90% of those were non-defaulters in current application So Bank should reanalyze those data to offer a loan .The previous offers which were unused in current application have maximum number of defaulters despite having high income

Recommendations

- **Recommendations for Applicants who should be targeted/preferred based on the insights :**
  - Applicants having own housing apartment
  - Females should be preferred over males.
  - Applicants having higher education.
  - Applicants who are married and having children 1-3
  - widows
  - Applicants having credit amount less that 1.5 million
  - Applicants having annuity less that 50K
  - Applicants having income less that 1 million
  - Applicants who have been Previously refused, cancelled offers should be reassessed for loan offers.
- **Recommendations for Applicants who should be avoided:**
  - Applicants who are single/civil married should be less targeted
  - Low skill laborers/drivers should be avoided
  - Applicants who have previous offer unused / rejected should be avoided
  - Applicants having low education level should be avoided