Loan Default Risk Analysis using Exploratory Data Analysis (EDA)

By Jaymeen Jethva

INDEX

- Problem Understanding
- Analysis Approach
- Steps followed for EDA
- Target Variable Analysis
- Insights of Analysis
- Correlation Matrix
- Conclusion

Problem Understanding

- 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 (too 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.
- Our main objective is to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Analysis Approach

Data Cleaning	Handling missing values , Binning Outliers
Univariate Analysis	Categorical Univariate, Numerical Univariate, Segmented Univariate
Bivariate Analysis	Numerical-Numerical , Numerical – Categorical, Categorical - Categorical
Multivariate Analysis	Correlation Matrix

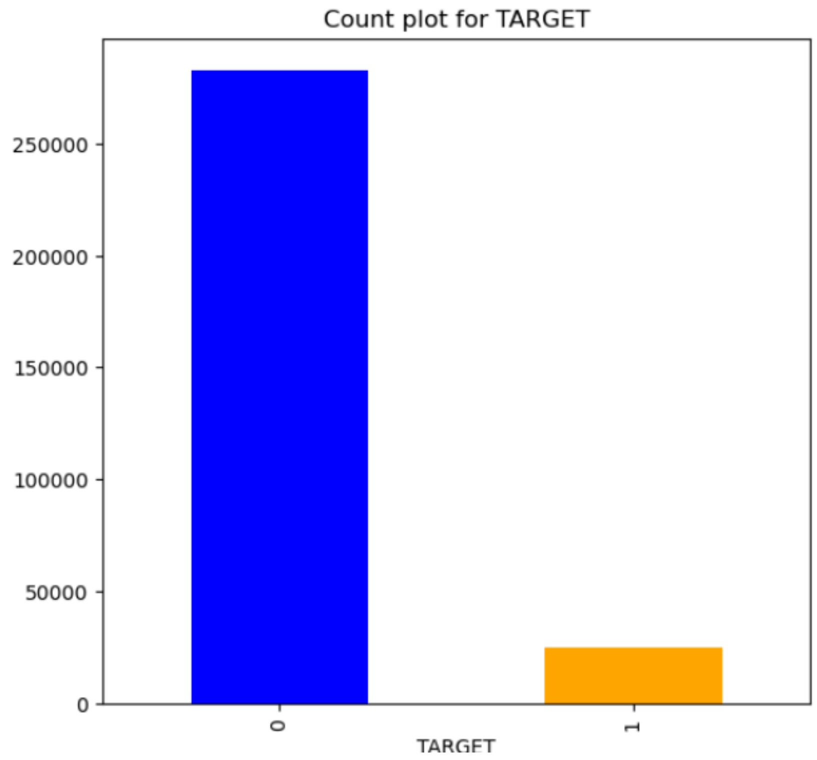
Steps followed for EDA

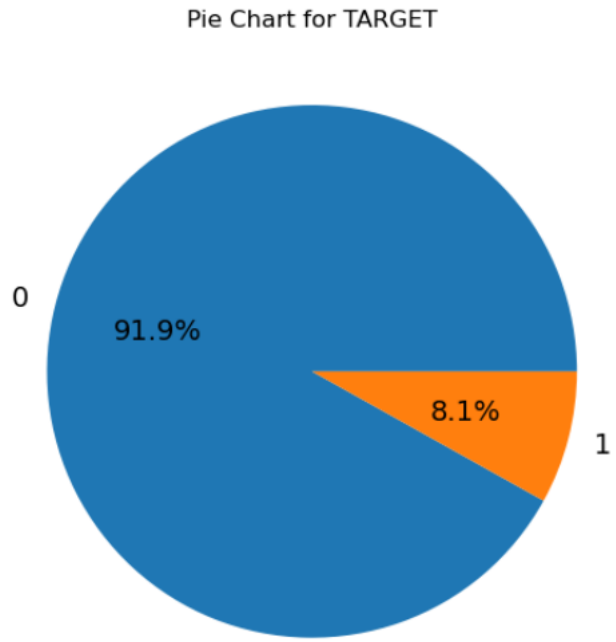
- Start with importing libraries , loading and understanding the dataset
- Check for the quality of the data and missing values
- Drop the columns having significant missing values(>45%) also check that data being dropped is not important for analysis
- Impute the missing values with Mean, Median, Mode or with "Missing" or "Others"
- Identify the Outliers in numerical variables and bin the necessary variable
- Check the imbalanced data

Steps followed for EDA

- Perform univariate analysis for Categorical and Numerical variables
- Perform segmented Univariate analysis by segmenting data into two parts
    1. Defaulter (Target 1)
    2. Non-Defaulter (Target 2)
- Perform Bivariate analysis with
    1. Numerical-Numerical Variables
    2. Numerical-Categorical Variables
    3. Categorical-Categorical Variables
- Find a correlation between variables and identify variables with higher correlation

Target Variable (Data Imbalance)

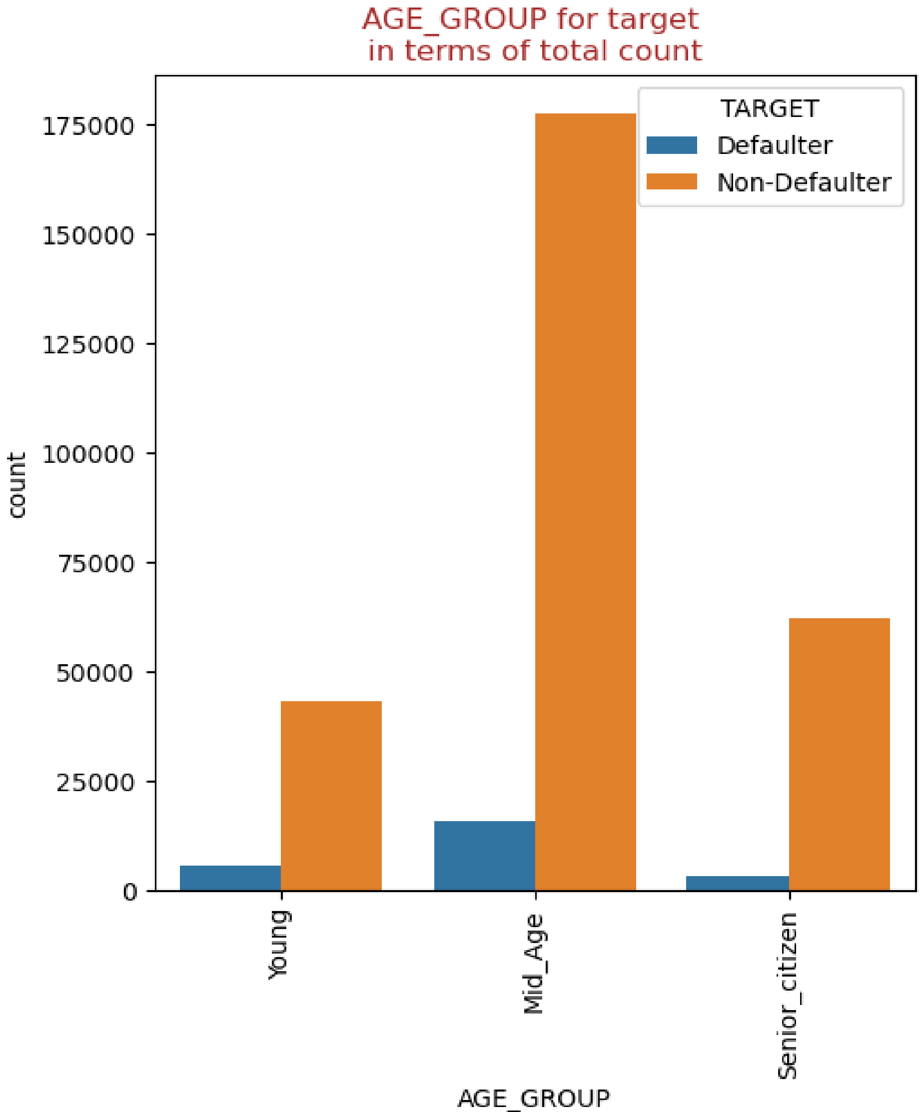


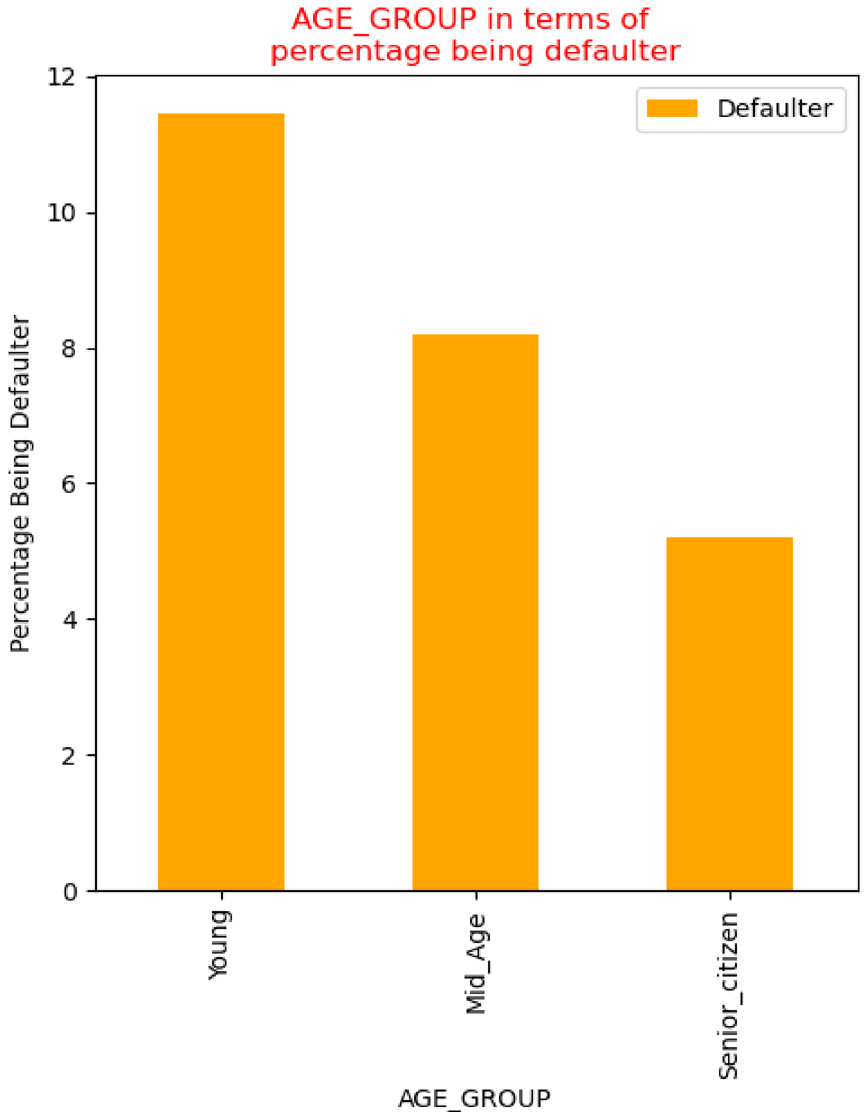


Insights :

- The Target variable is highly imbalanced Where 8.1% of clients are Defaulters and 91.9% of clients are Non-Defaulters
- Imbalanced ratio is (91.9 / 8.1) = 11.34

Insights of Segmented Univariate Analysis

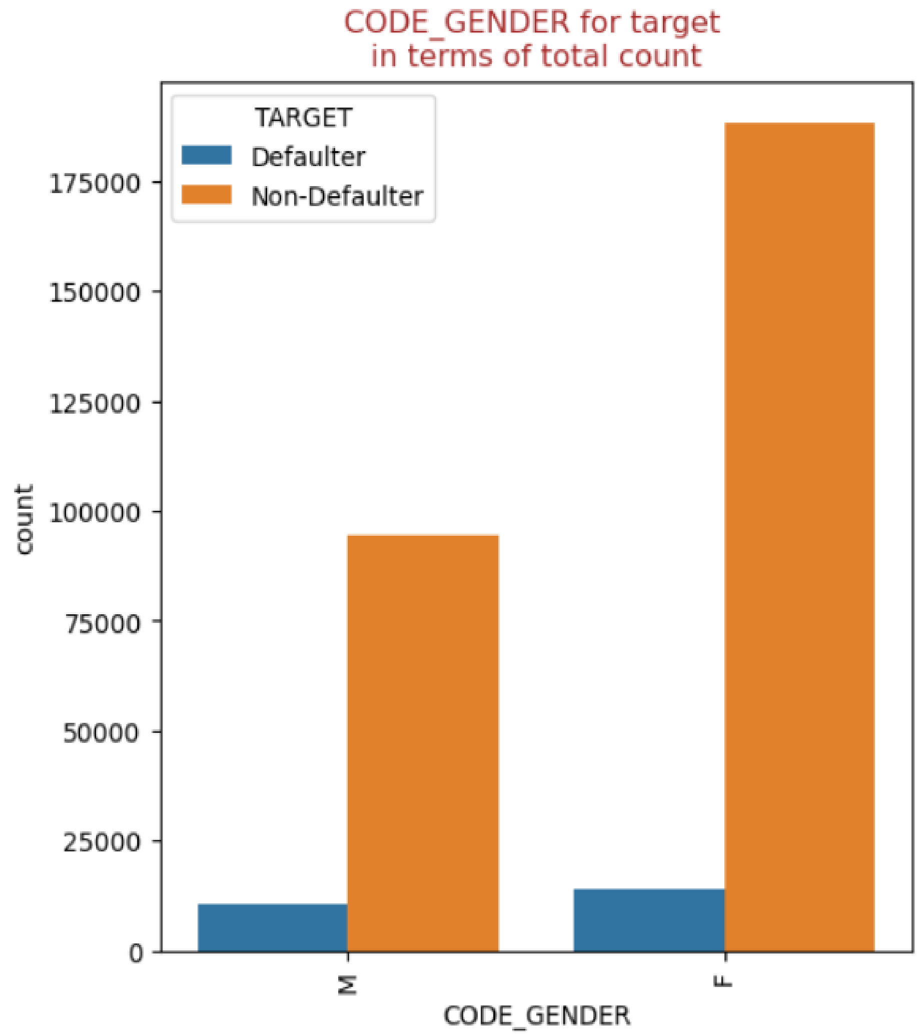


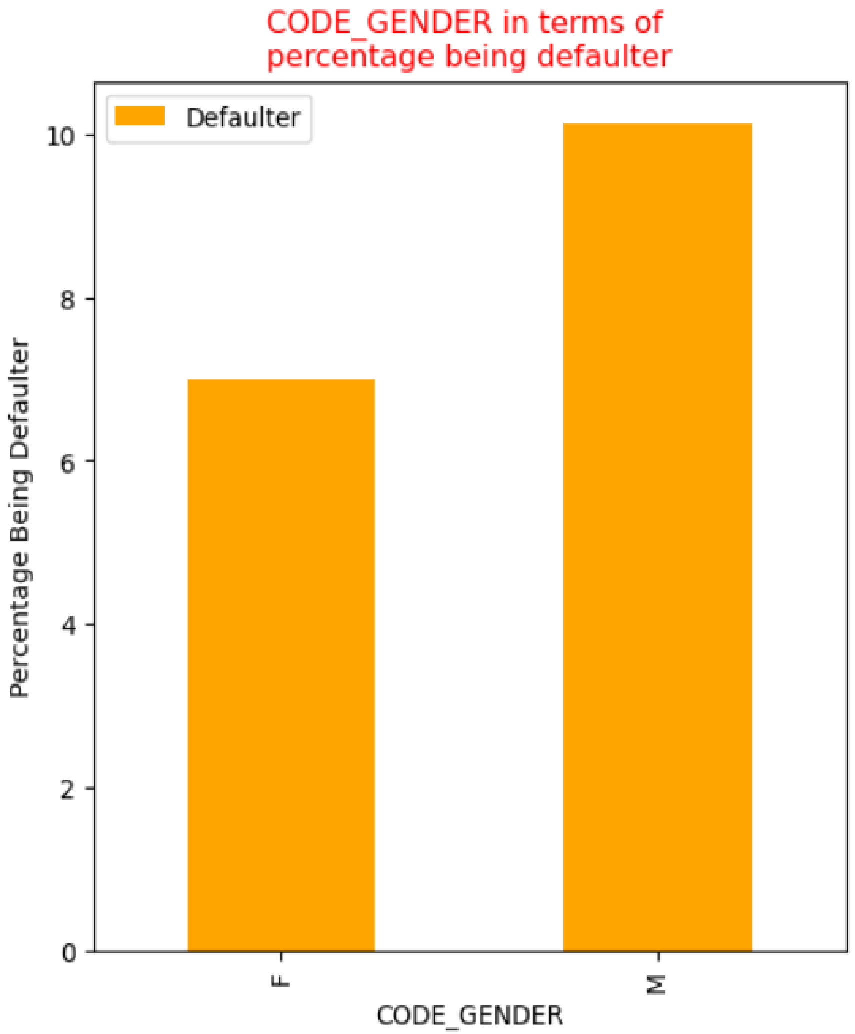


Insights :

- Clients who falls into Young Age Group (i.e. 19-30 yrs) are having highest percentage of being a defaulter among all three Age groups

Insights of Segmented Univariate Analysis

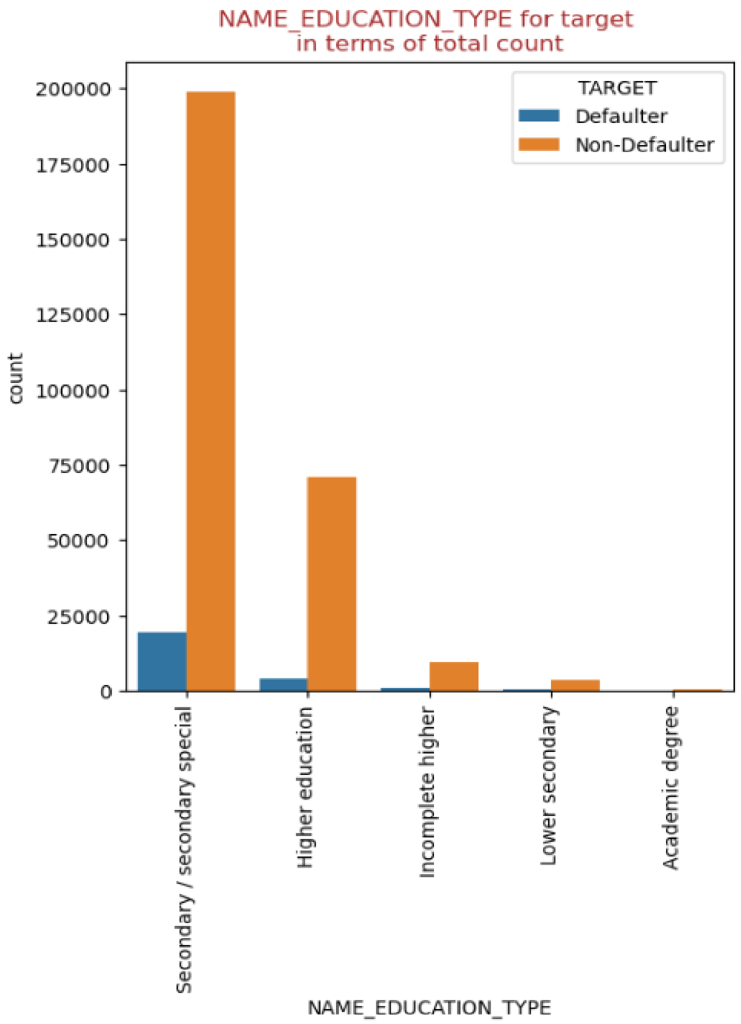


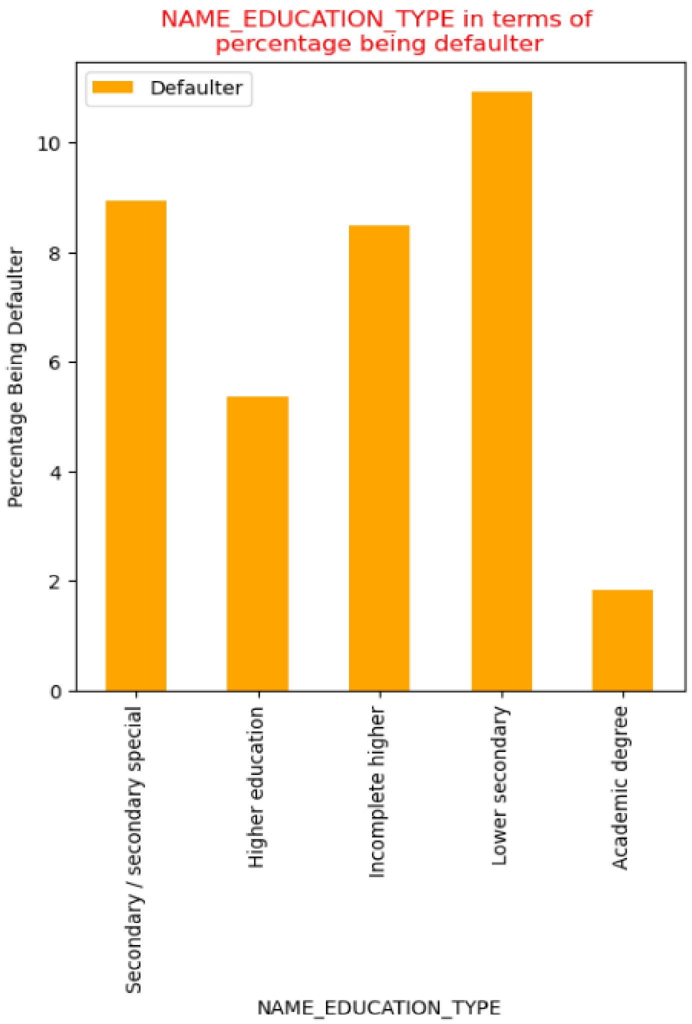


Insights :

- Female Applicants are higher than the males but among all the males and female, Percentage of male being defaulter is higher than female

Insights of Segmented Univariate Analysis

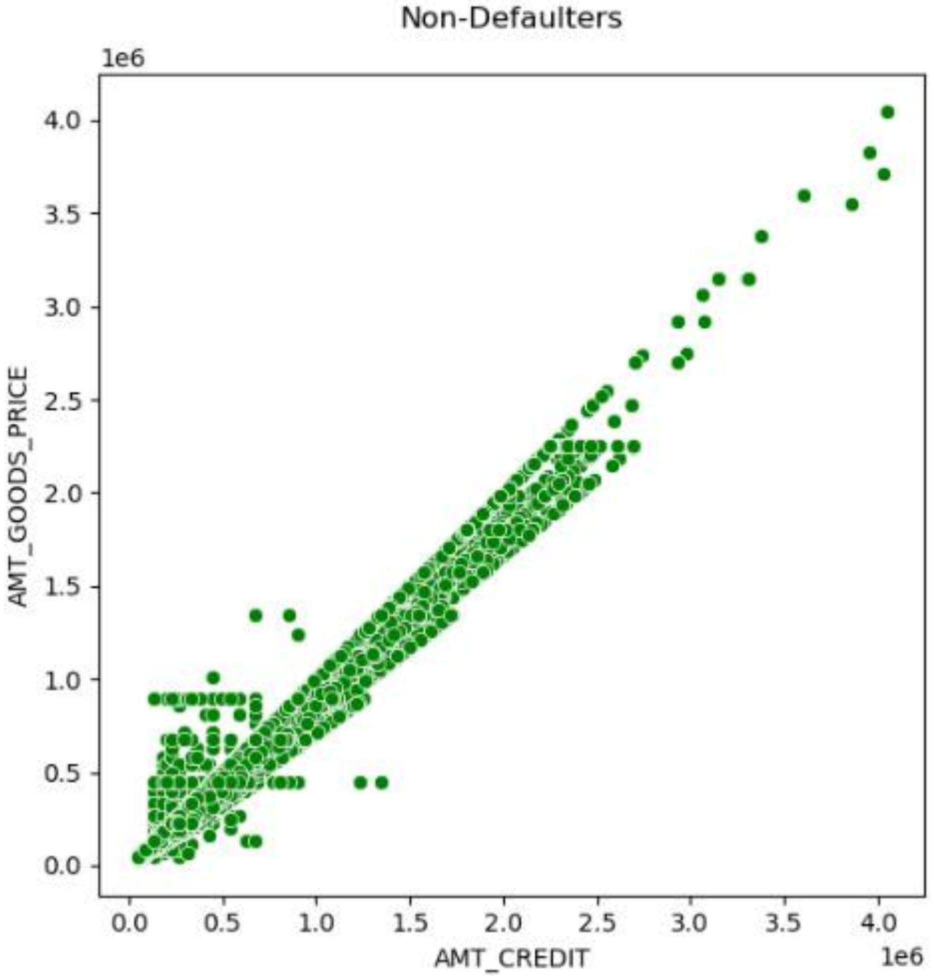


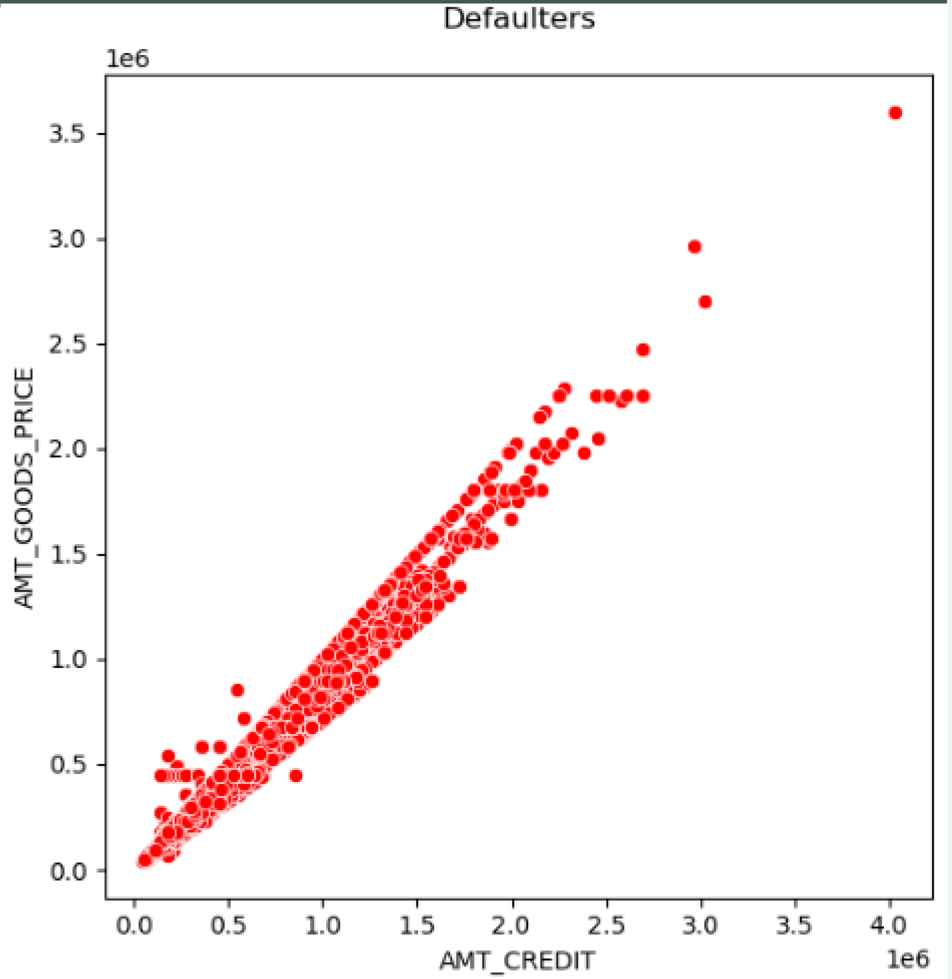


Insights :

- Clients having Higher Education are having least percentage of being defaulter

Insights of Bivariate Analysis

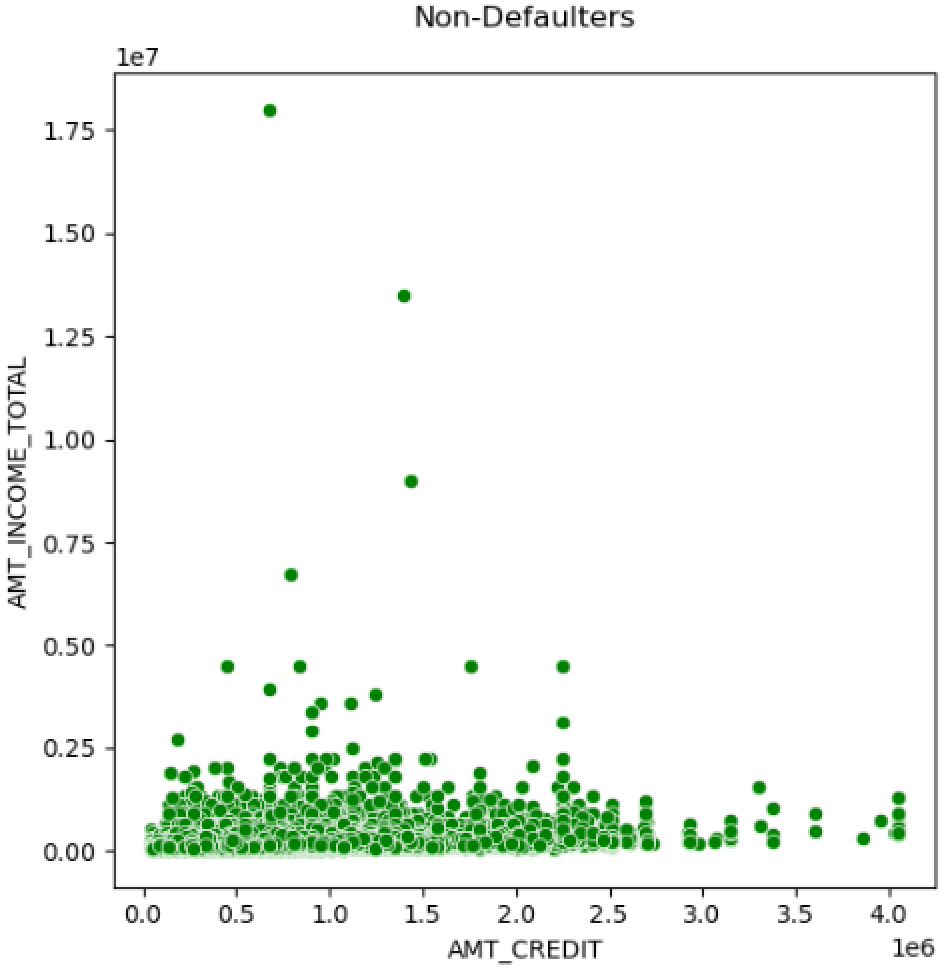


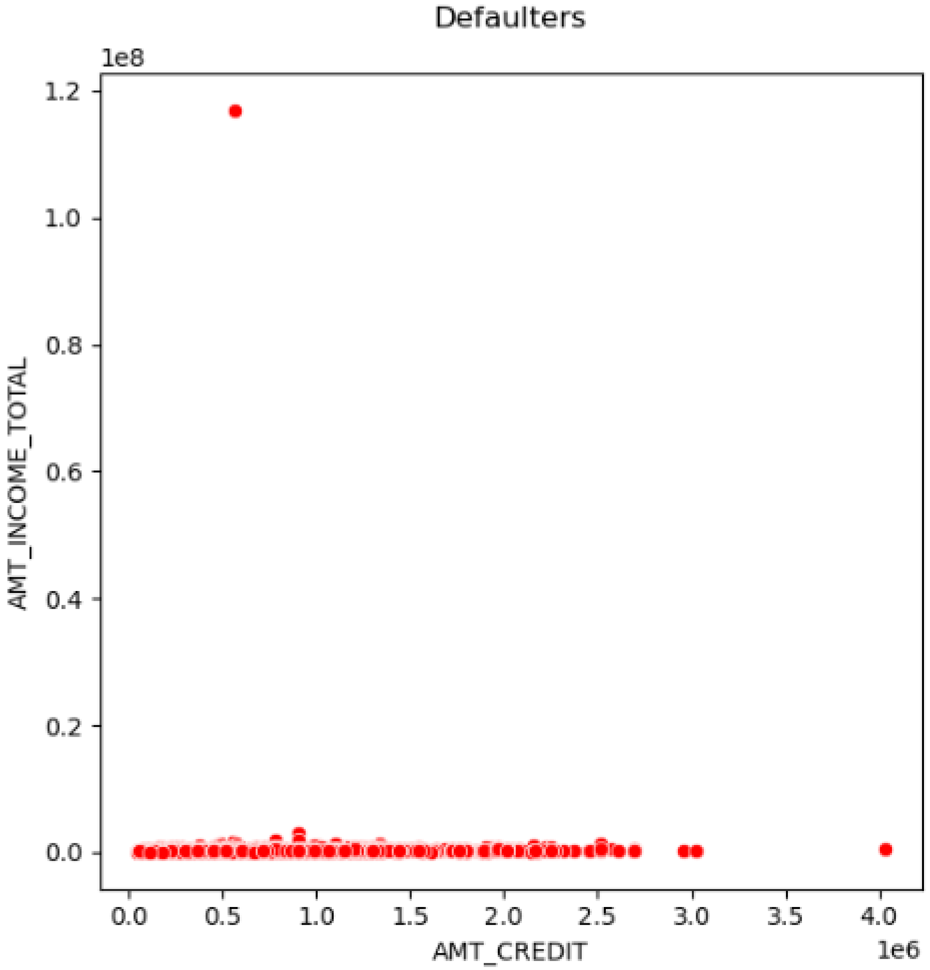


Insights :

- Amount credited and Amount of price of goods are showing same trend for both the cases Non-defaulter & defaulter
- AMT_CREDIT and AMT_GOODS_PRICE are having high correlation

Insights of Bivariate Analysis

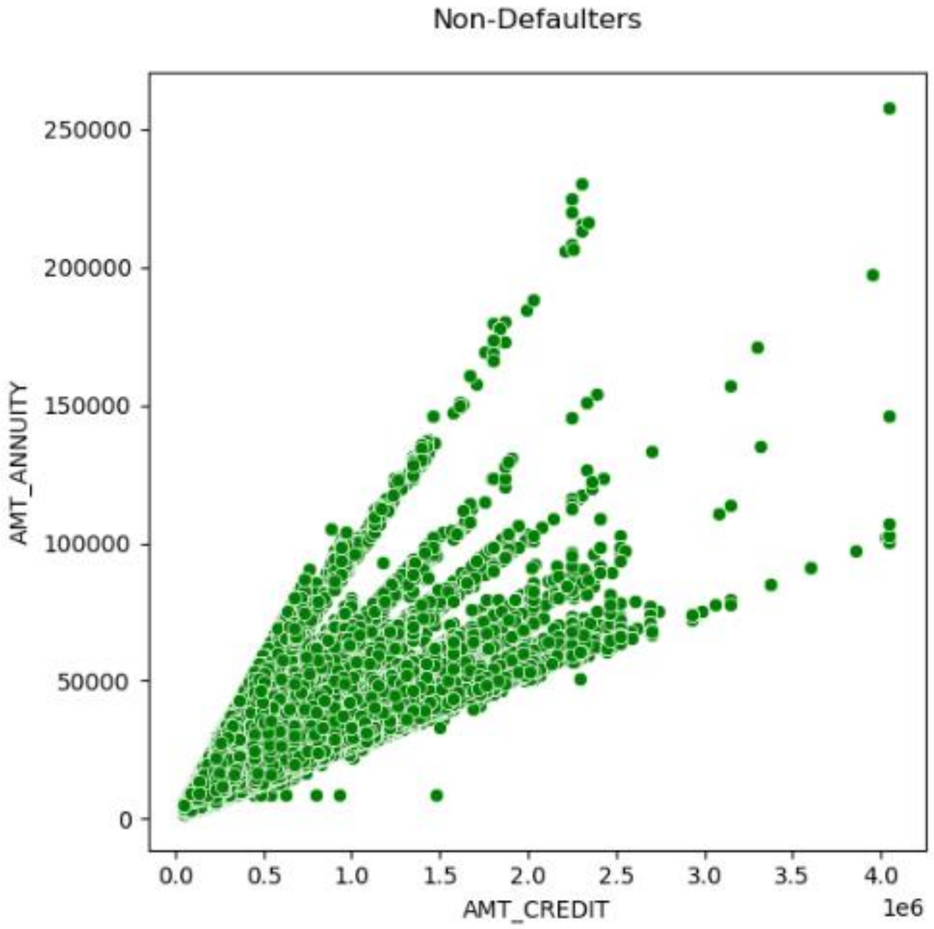


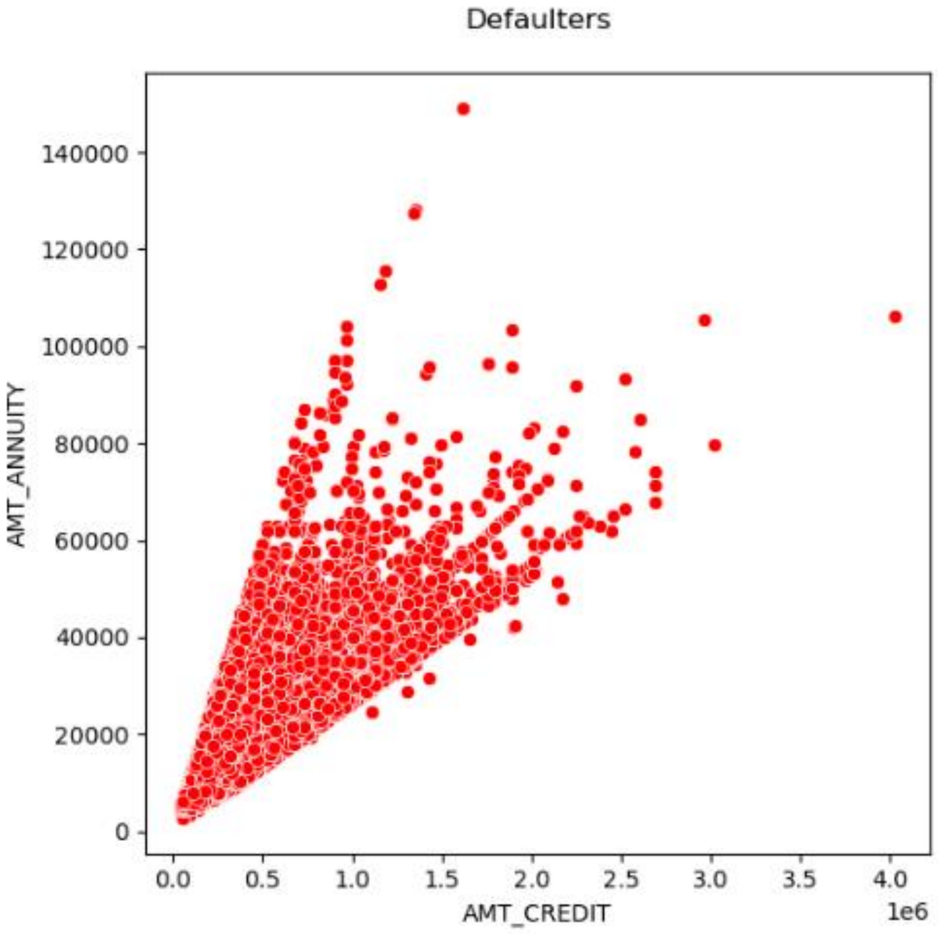


Insights :

- Clients with low income being more likely to default, regardless of the credit amount.

Insights of NUmerical –Numerical Bivariate Analysis

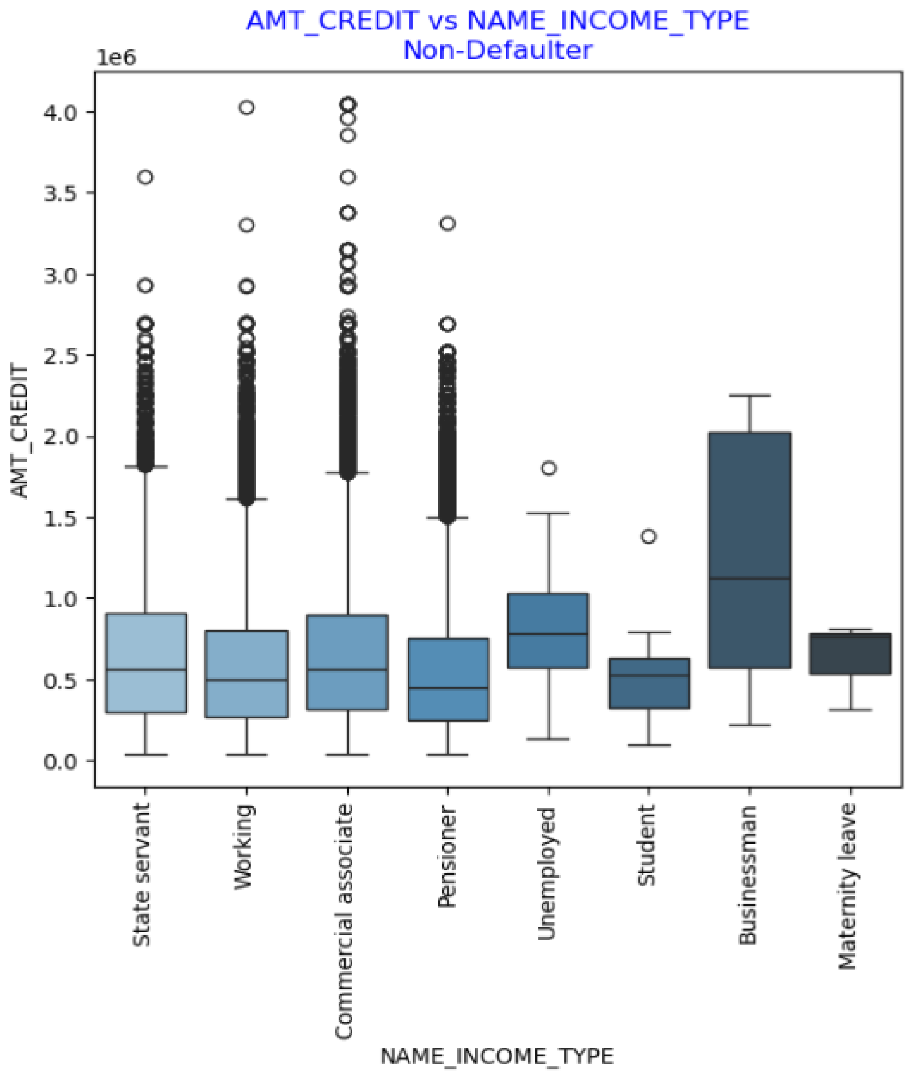


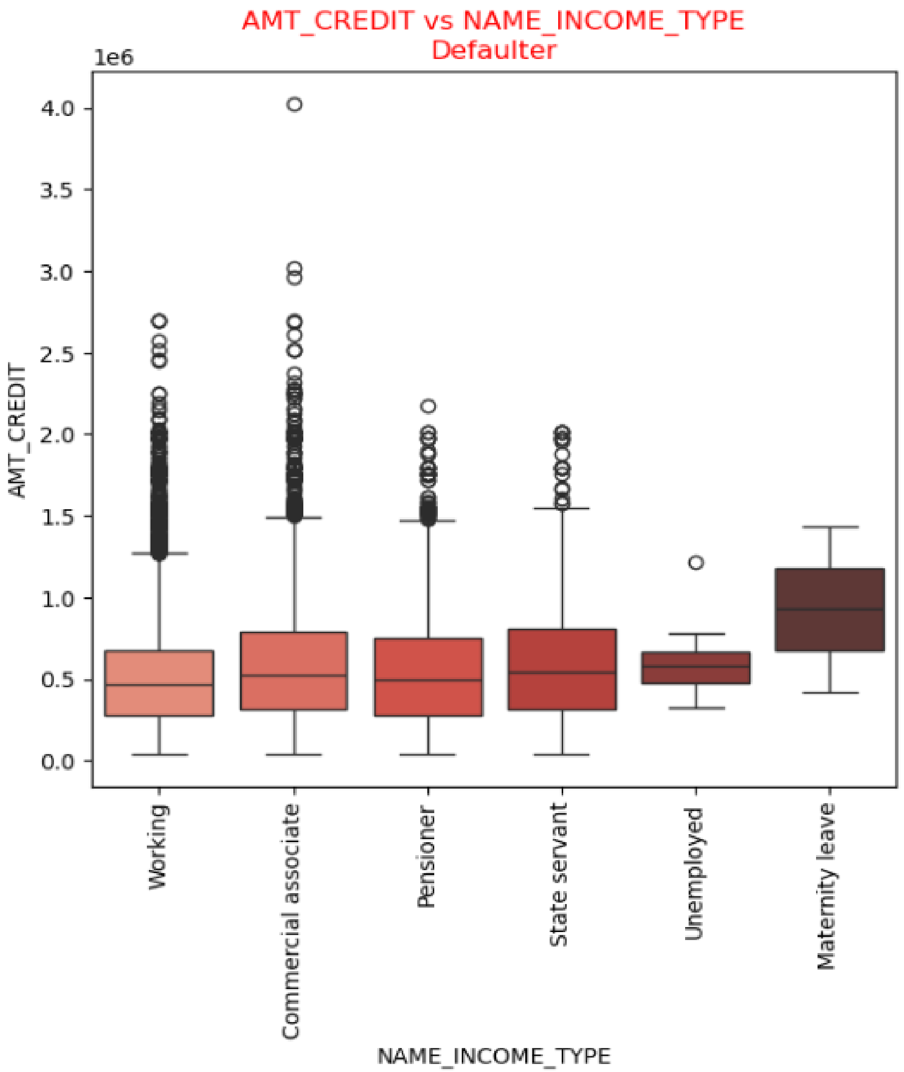


Insights :

- There is positive relationship between AMT_ANNUITY and AMT_CREDIT for both the cases but we can see for Defaulters the slope is slightly more than the Non-defaulter so we can say that clients having more Annuity amount for low credit are more likely to be defaulters

Insights of Numerical – Categorical Bivariate Analysis

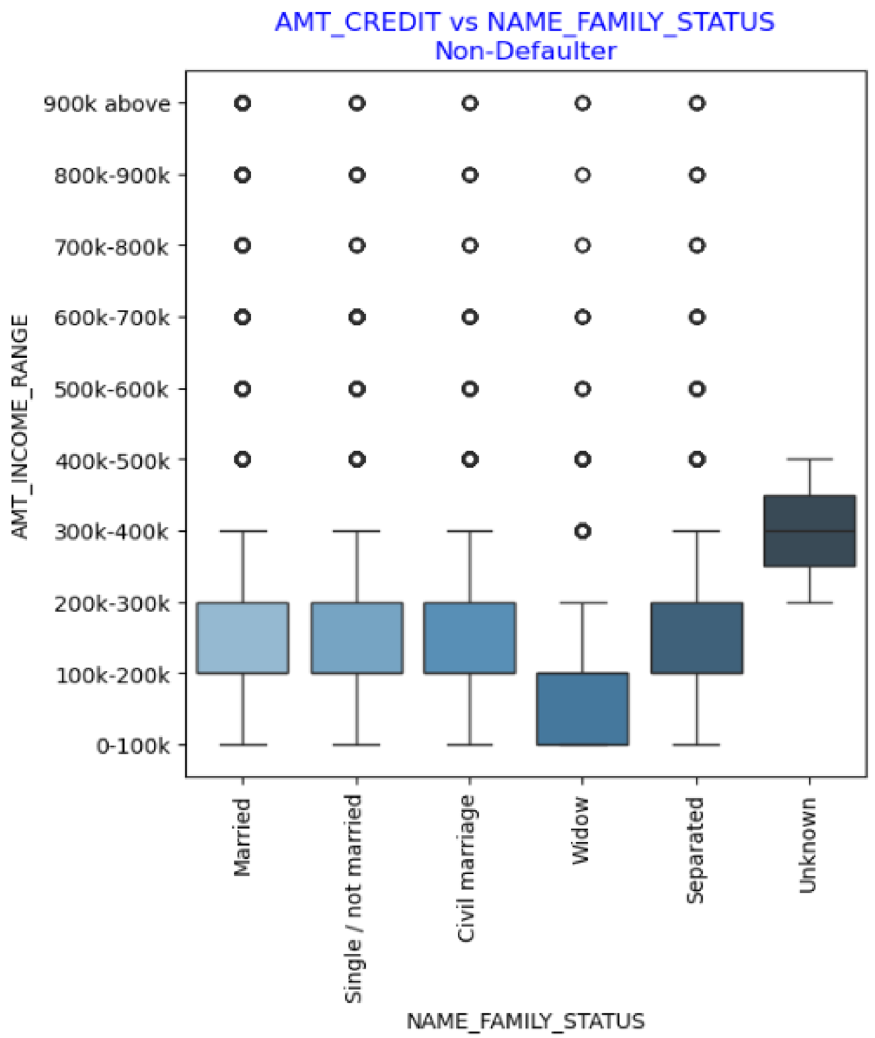


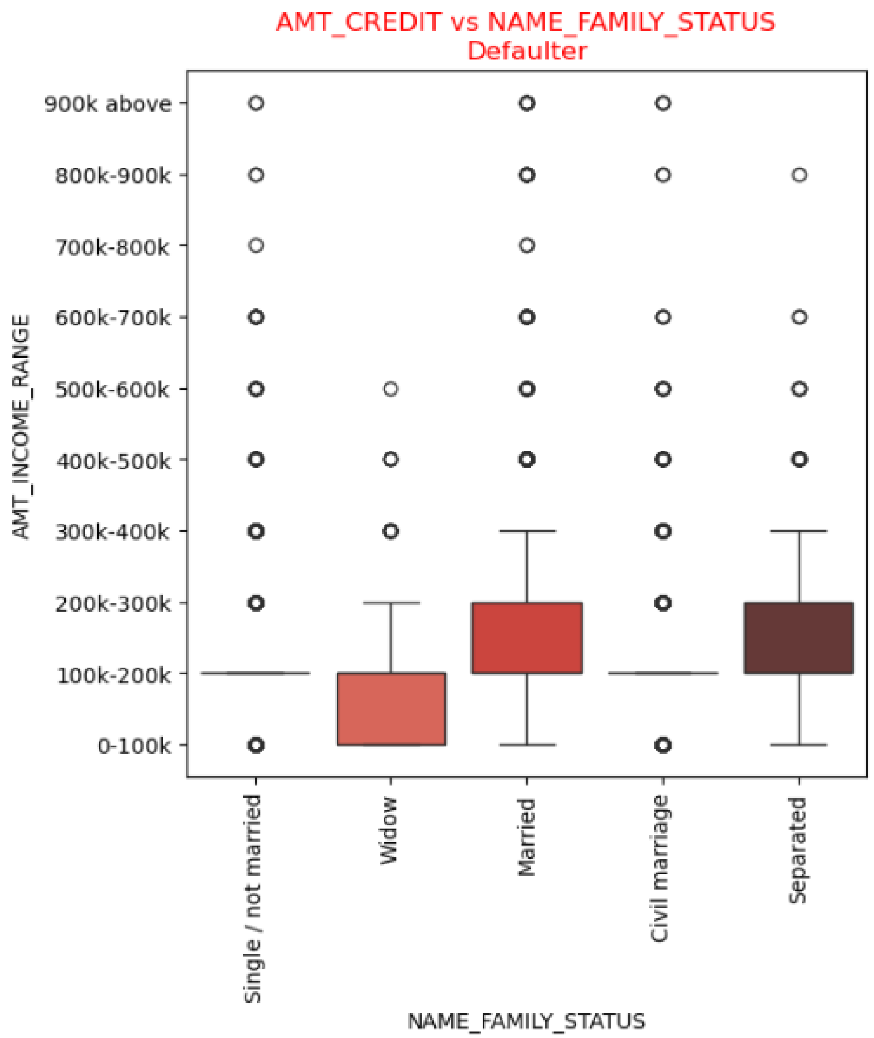


Insights :

- The Amount Credited for businessman is mostly high
- None of the Businessman and Student are defaulter

Insights of Numerical – Categorical Bivariate Analysis





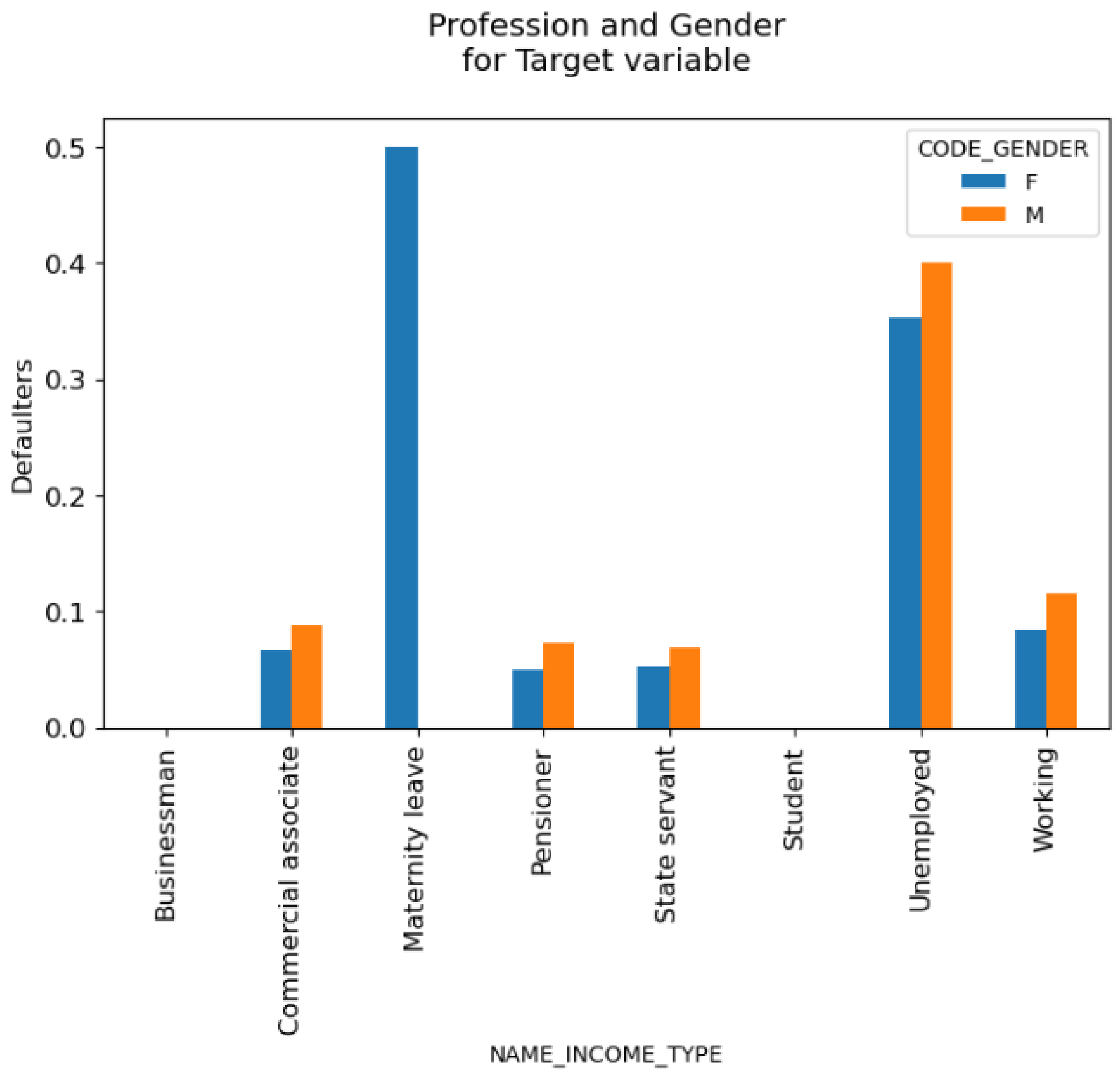
Insights :

- Clients who are Single/Unmarried or having civil marriage with lower income are more likely to be a defaulter

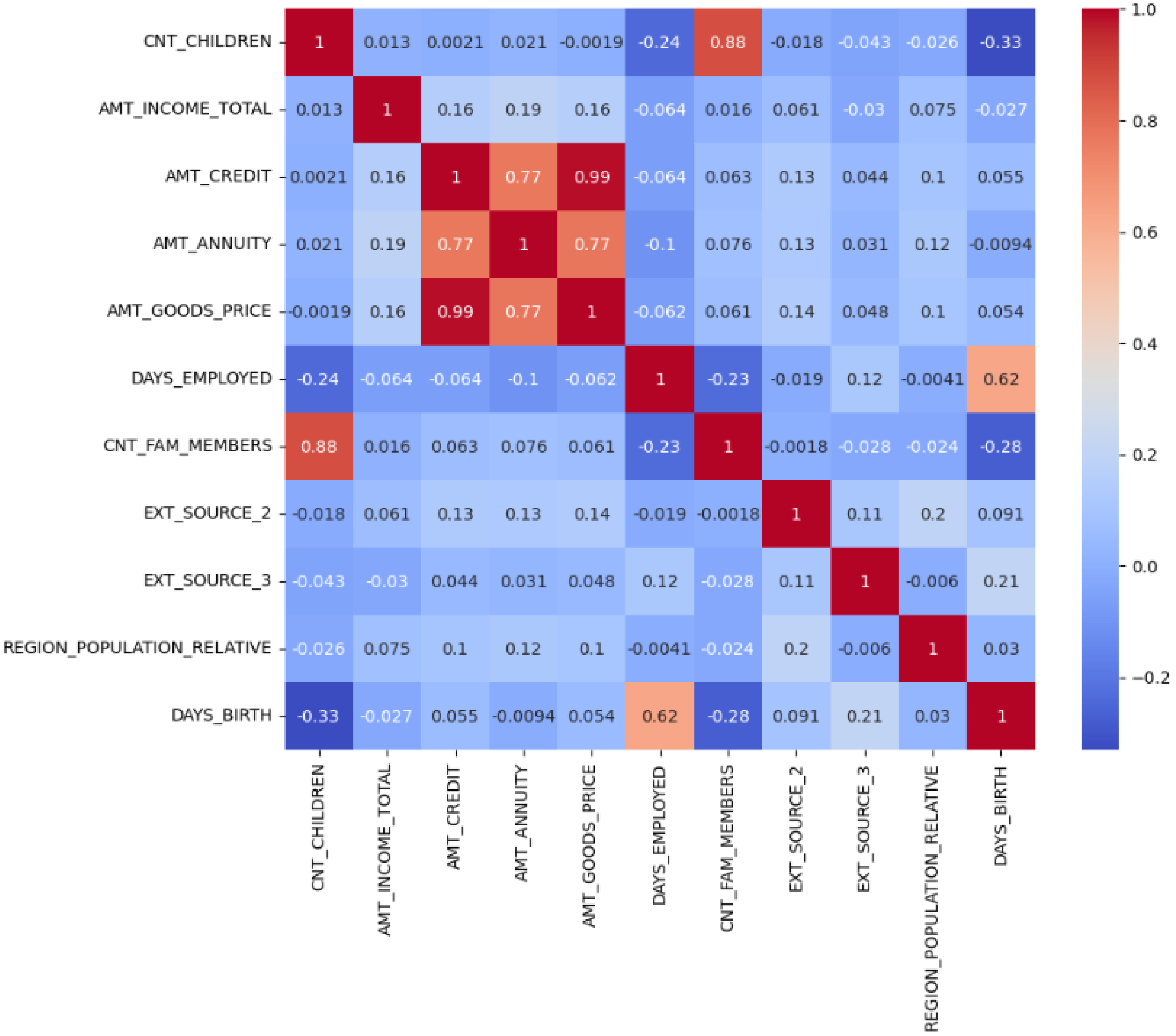
Insights of Categorical- Categorical Bivariate Analysis

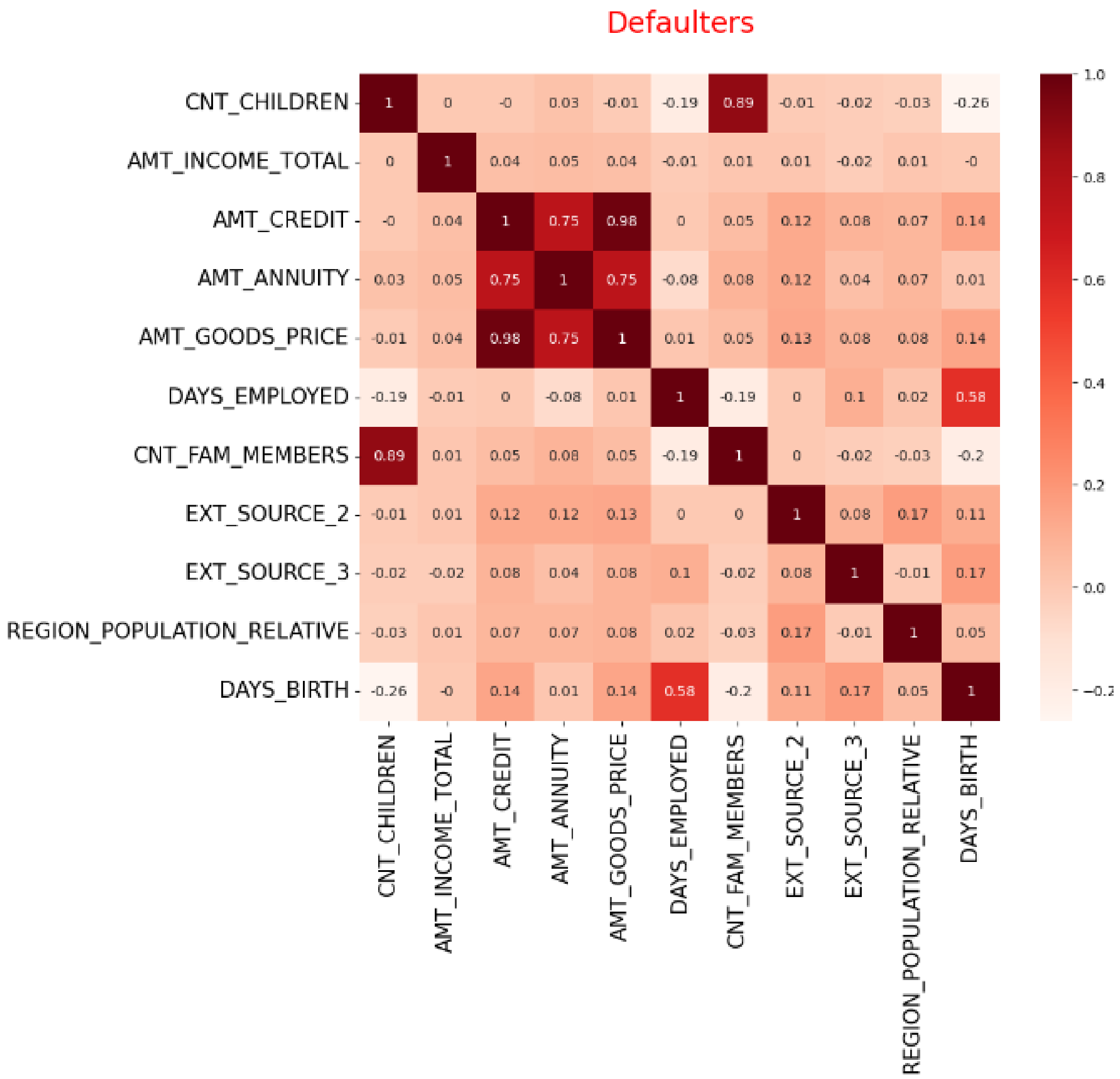
Insights :

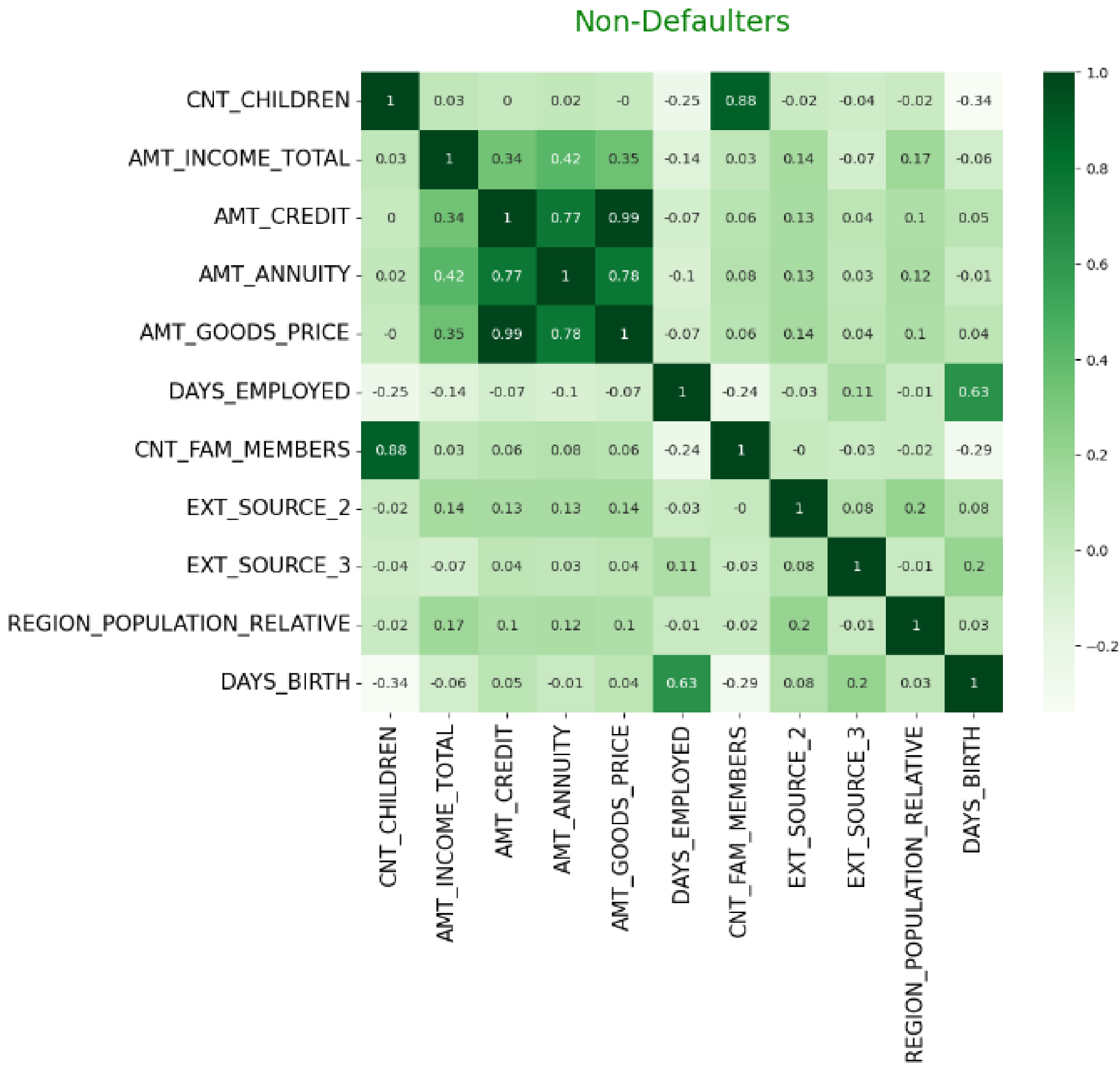
- Clients either unemployed or on maternity leave are more likely to be a defaulter
- Males are more defaulted with their respective professions compared to females except from maternity leave



Correlation Matrix




Defaulters
CNT_CHILDREN 1 0 -0 0.03 -0.01 -0.19 0.89 -0.01 -0.02 -0.03 -0.26
AMT_INCOME_TOTAL 0 1 0.04 0.05 0.04 -0.01 0.01 0.01 -0.02 0.01 -0
AMT_CREDIT -0 0.04 1 0.75 0.98 0 0.05 0.12 0.08 0.07 0.14
AMT_ANNUITY 0.03 0.05 0.75 1 0.75 -0.08 0.08 0.12 0.04 0.07 0.01
AMT_GOODS_PRICE -0.01 0.04 0.98 0.75 1 0.01 0.05 0.13 0.08 0.08 0.14
DAYS_EMPLOYED -0.19 -0.01 0 -0.08 0.01 1 -0.19 0 0.1 0.02 0.58
CNT_FAM_MEMBERS 0.89 0.01 0.05 0.08 0.05 -0.19 1 0 -0.02 -0.03 -0.2
EXT_SOURCE_2 -0.01 0.01 0.12 0.12 0.13 0 0 1 0.08 0.17 0.11
EXT_SOURCE_3 -0.02 -0.02 0.08 0.04 0.08 0.1 -0.02 0.08 1 -0.01 0.17
REGION_POPULATION_RELATIVE -0.03 0.01 0.07 0.07 0.08 0.02 -0.03 0.17 -0.01 1 0.05
DAYS_BIRTH -0.26 -0 0.14 0.01 0.14 0.58 -0.2 0.11 0.17 0.05 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
DAYS_EMPLOYED
CNT_FAM_MEMBERS
EXT_SOURCE_2
EXT_SOURCE_3
REGION_POPULATION_RELATIVE
DAYS_BIRTH
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Non-Defaulters
CNT_CHILDREN 1 0.03 0 0.02 -0 -0.25 0.88 -0.02 -0.04 -0.02 -0.34
AMT_INCOME_TOTAL 0.03 1 0.34 0.42 0.35 -0.14 0.03 0.14 -0.07 0.17 -0.06
AMT_CREDIT 0 0.34 1 0.77 0.99 -0.07 0.06 0.13 0.04 0.1 0.05
AMT_ANNUITY 0.02 0.42 0.77 1 0.78 -0.1 0.08 0.13 0.03 0.12 -0.01
AMT_GOODS_PRICE -0 0.35 0.99 0.78 1 -0.07 0.06 0.14 0.04 0.1 0.04
DAYS_EMPLOYED -0.25 -0.14 -0.07 -0.1 -0.07 1 -0.24 -0.03 0.11 -0.01 0.63
CNT_FAM_MEMBERS 0.88 0.03 0.06 0.08 0.06 -0.24 1 -0 -0.03 -0.02 -0.29
EXT_SOURCE_2 -0.02 0.14 0.13 0.13 0.14 -0.03 -0 1 0.08 0.2 0.08
EXT_SOURCE_3 -0.04 -0.07 0.04 0.03 0.04 0.11 -0.03 0.08 1 -0.01 0.2
REGION_POPULATION_RELATIVE -0.02 0.17 0.1 0.12 0.1 -0.01 -0.02 0.2 -0.01 1 0.03
DAYS_BIRTH -0.34 -0.06 0.05 -0.01 0.04 0.63 -0.29 0.08 0.2 0.03 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
DAYS_EMPLOYED
CNT_FAM_MEMBERS
EXT_SOURCE_2
EXT_SOURCE_3
REGION_POPULATION_RELATIVE
DAYS_BIRTH
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Variables with High Correlation

- **Variables with high correlation for Non-Defaulter:**
  - CNT_FAM_MEMBERS & CNT_CHILDREN -> <u>0.88</u>
  - AMT_GOODS_PRICE & AMT_CREDIT -> <u>0.99</u>
  - AMT_CREDIT & AMT_ANNUITY -> <u>0.77</u>
  - AMT_GOODS_PRICE & AMT_ANNUITY -> <u>0.78</u>
  - DAYS_BIRTH & DAYS_EMPLOYED -> <u>0.63</u>

- **Variables with high correlation for Defaulter:**
  - CNT_FAM_MEMBERS & CNT_CHILDREN -> <u>0.89</u>
  - AMT_GOODS_PRICE & AMT_CREDIT -> <u>0.98</u>
  - AMT_CREDIT & AMT_ANNUITY -> <u>0.75</u>
  - AMT_GOODS_PRICE & AMT_ANNUITY -> <u>0.75</u>
  - DAYS_BIRTH & DAYS_EMPLOYED -> <u>0.58</u>

Conclusion:

- As the data was highly imbalanced so the insights that we get from it aren't much conclusive but still there are few key factor that we can consider :

1. **Education** – It is observed that people with higher education level are least likely to be a defaulter so Bank should be more focusing on clients having Higher education
2. **Age** – Young people(19-30 years old) are observed more likely of being a defaulter so Bank should be less focusing on young clients
3. **Income** – People having low amount of income are found to be a defaulter. It is expected as people with lower income can't repay their loan
4. **Occupation** – It is observed that none of the business man is defaulter and they takes a loan of higher credit amount so Bank should be more focusing on clients having business
5. **Gender** – Among male and female it is observed that males are more likely to be a defaulters

THANK YOU