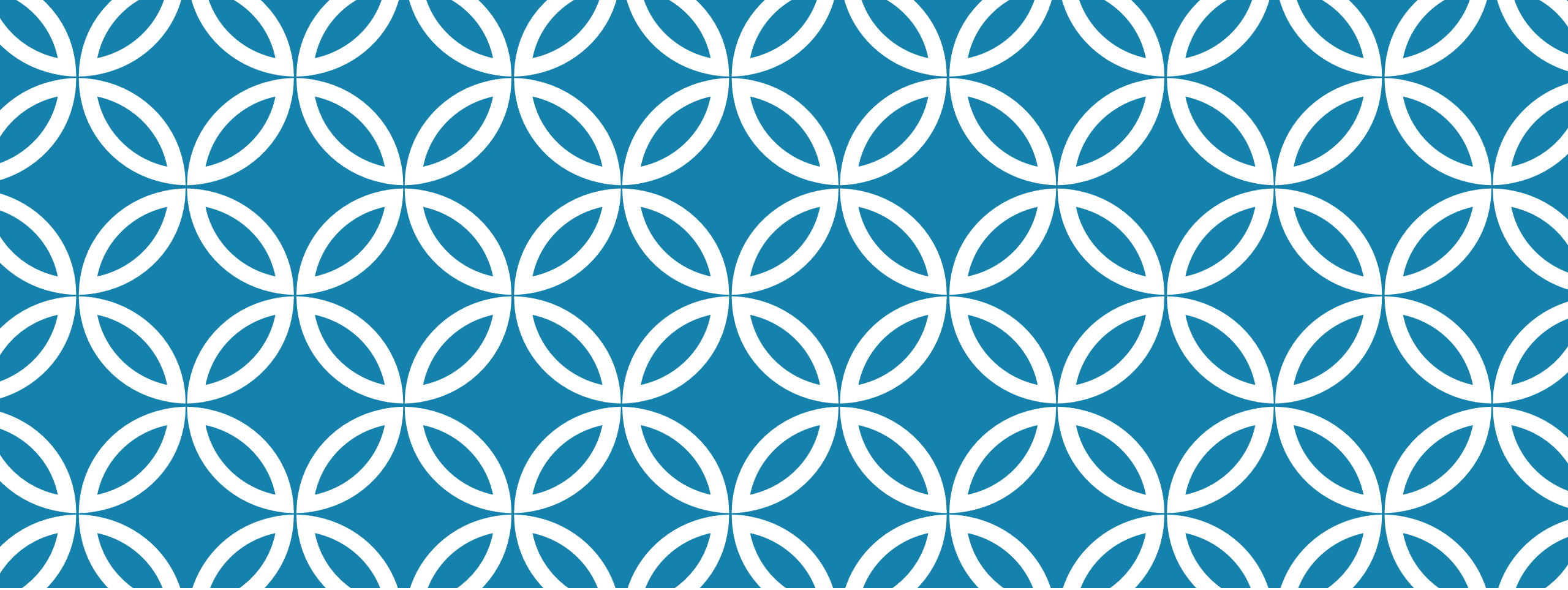


CREDIT EDA ASSIGNMENT REPORT & ANALYSIS

kumar saransh

PROJECT DESCRIPTION

The objective of this case study is to illustrate the practical application of EDA in a genuine business context. Throughout this assignment, you will not only employ the EDA techniques you've acquired but also gain a foundational comprehension of risk analytics within the banking and financial services sector. You will discover how data is leveraged to reduce the risk of financial losses when extending loans to customers.

Objective:

I. To comprehend the factors contributing to challenges in making loan payments for certain individuals.

II. To uncover any additional trends within the data that could aid in identifying elements associated with loan defaults.

BUSINESS UNDERSTANDING

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 specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

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:

I. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

II. 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.

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.

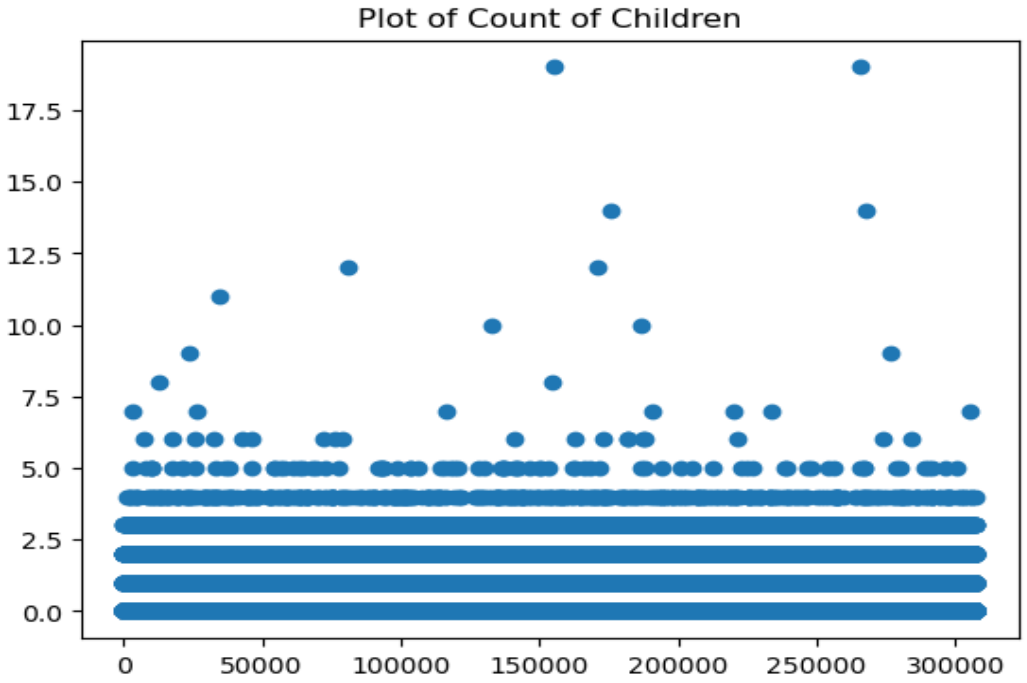
In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

EDA METHODOLOGIES

I. Read and load the csv files (data).
II. Perform Data Quality Checks (treatment of missing values, outliers, binning and type correction)
III. Check data imbalance & split the data
IV. Perform univariate analysis to compare trends between defaulters and non defaulters
V. Perform bivariate analysis, including correlation analysis to find some pattern.
VI. Load the previous application data and merge with current application data (inner join).
VII. Perform the same steps as before to find new trends between previous application and current application, in relation with defaulters and non defaulter.

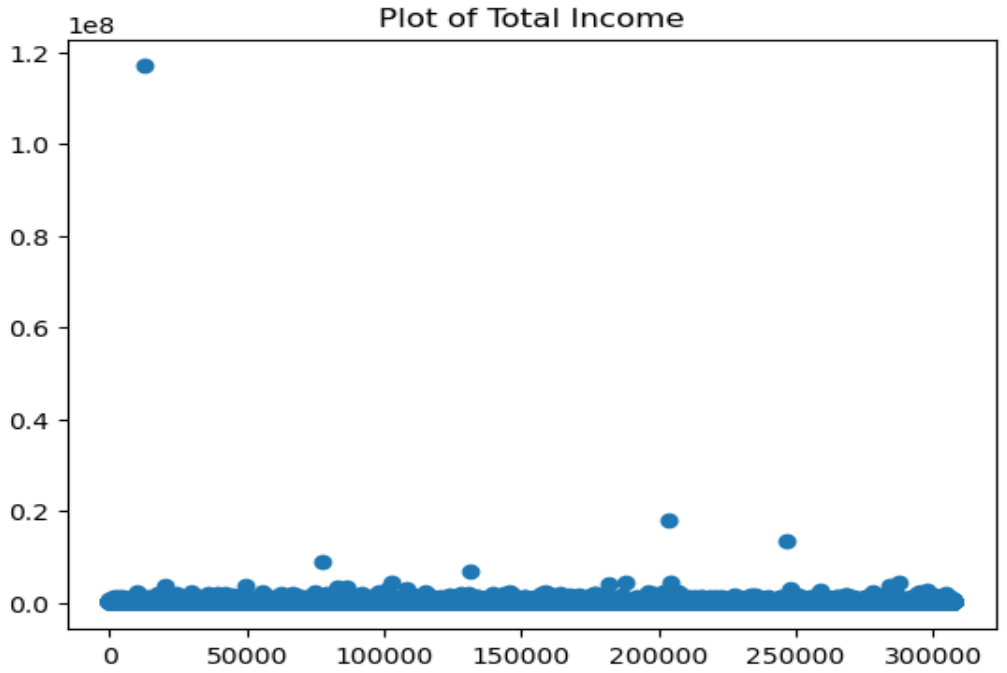
ANALYSIS ON ‘APPLICATION_DATA.CSV’

OUTLIER ANALYSIS



Inferences:

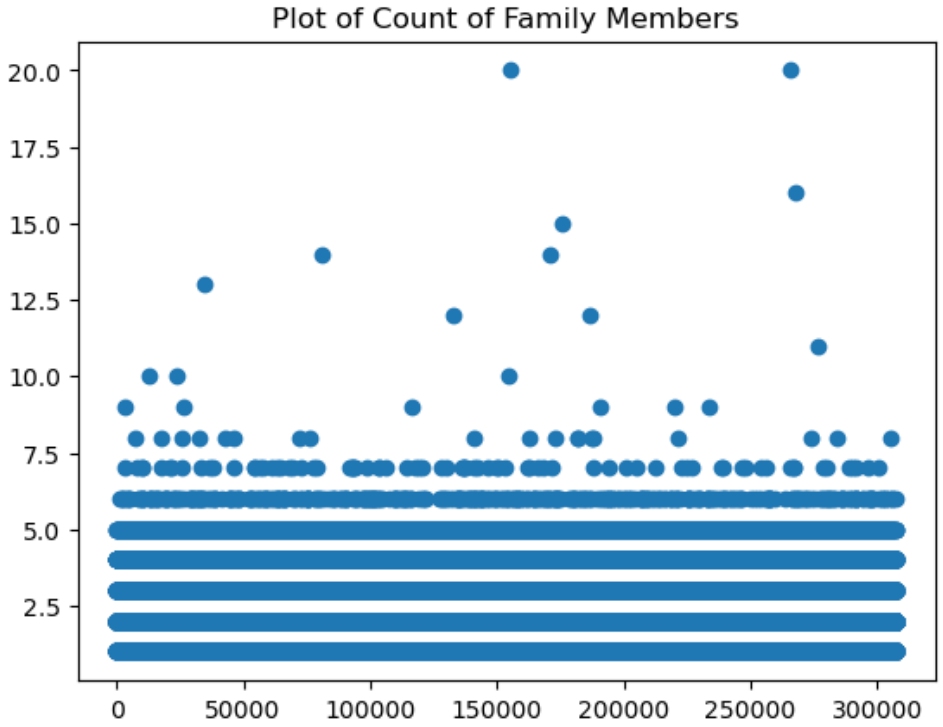
- Majority of the families has children below 5.
- There are 19 children in a family, These outliers may reflect unique family structures or cultural dynamics



Inferences:

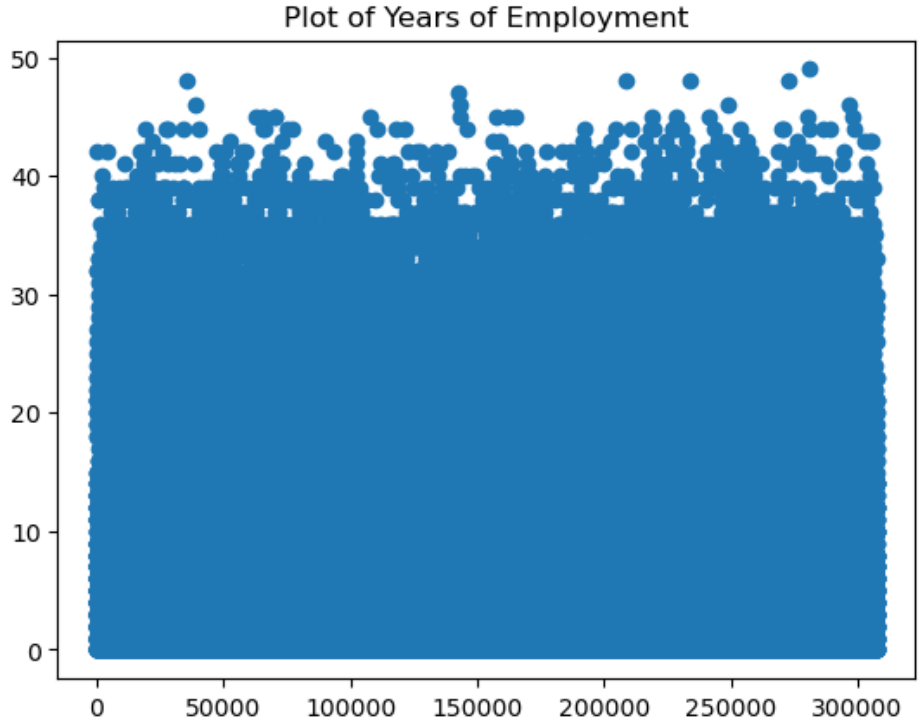
- Total income have outliers, indicating a presence of individuals with significantly higher incomes compared to the majority.
- These outliers may represent the high earning professionals, Founder of some organization.

OUTLIER ANALYSIS



Inferences:

- Majority of the families has below 7 family members.
- There are some families who has more than 20 family members, These outliers may reflect joint family or multi-generational family.



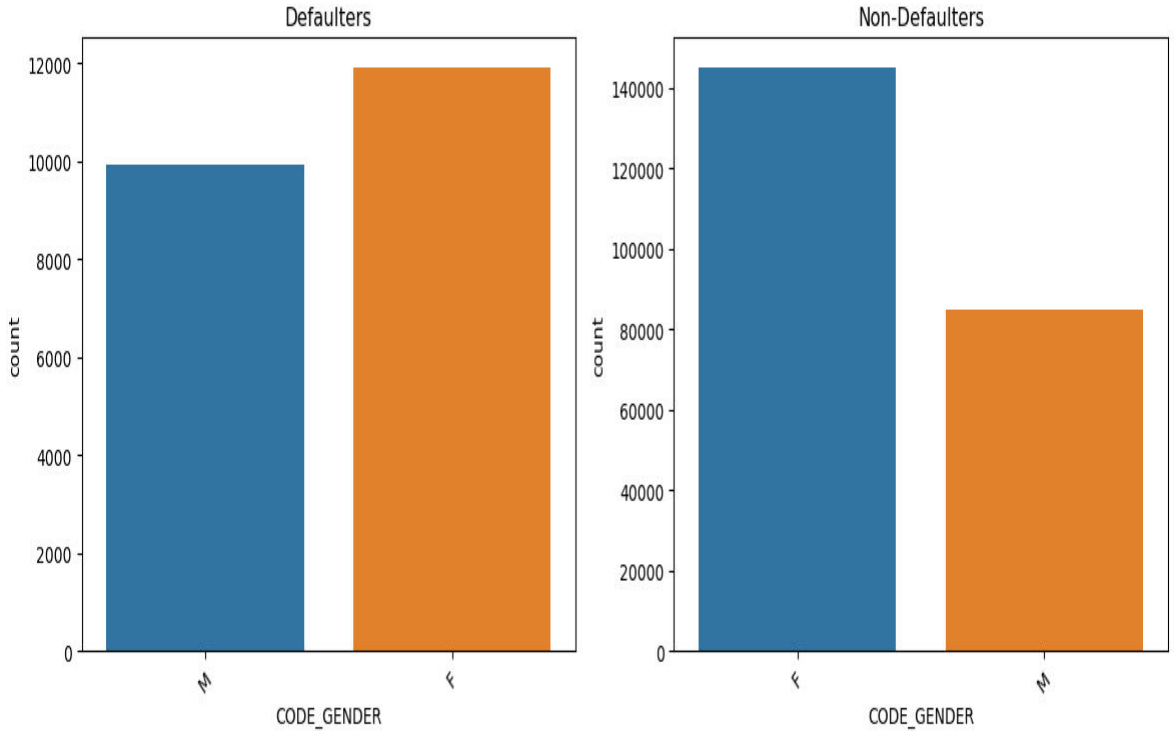
Inferences:

- These outliers indicates that there are certain employees who has been employed from a long time in the same organization.
- This also indicates that there is a lower risk of job loss.

UNIVARIATE ANALYSIS ON CODE_GENDER

Inferences:

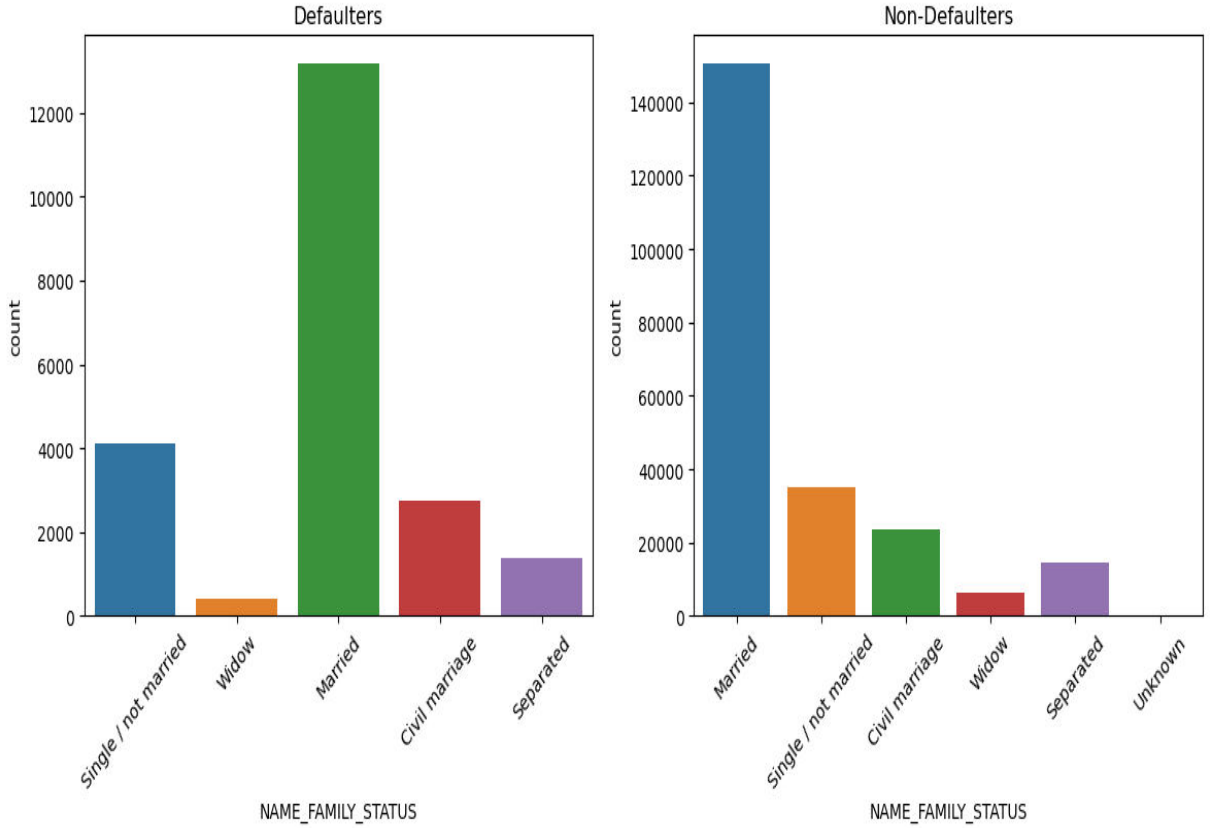
- It is observed that the count of females are slightly more than males as Defaulters.
- It is observed that the count of females are slightly more than males as Non-Defaulters.



UNIVARIATE ANALYSIS ON NAME_FAMILY_STATUS

Inferences:

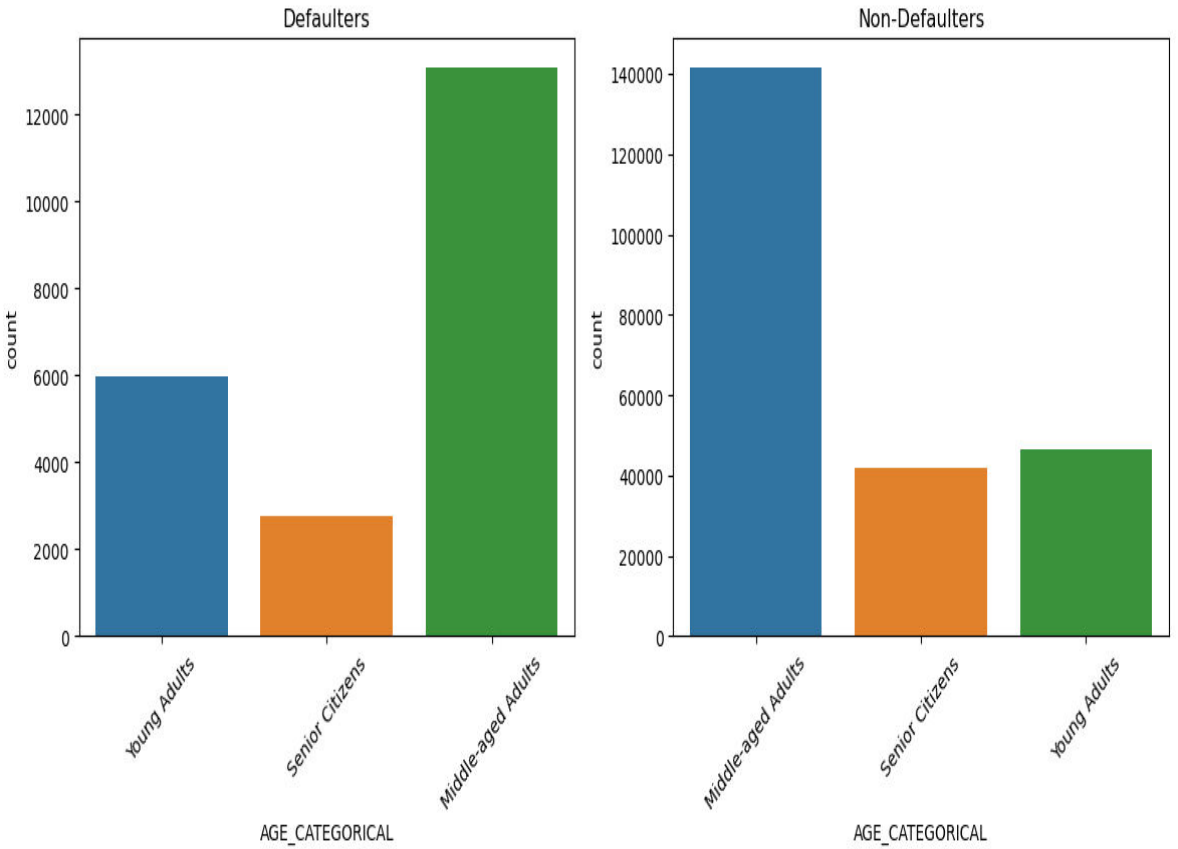
- It is observed that the count of married are significantly higher as Defaulters & Widow are the least defaulters than any other family status.
- It is observed the similar pattern for non-defaulters, married peoples are high as non-defaulters and widow as the least.



UNIVARIATE ANALYSIS ON AGE_CATEGORICAL

Inferences:

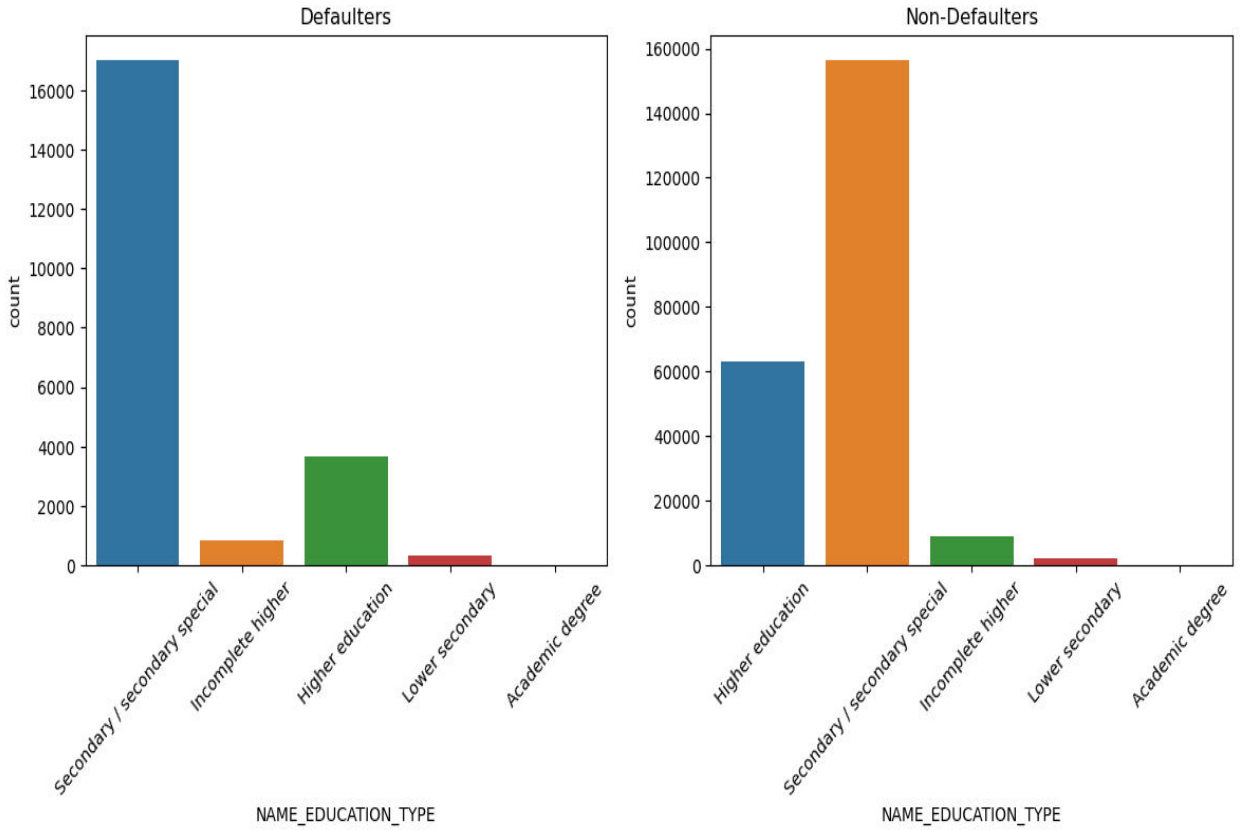
- It is observed that Senior Citizens are compared to be less defaulters than young and middle aged youths.
- Middle Aged youth are the least defaulters.
- Young adults are seems to be more defaulter than the non-defaulter.



UNIVARIATE ANALYSIS ON 'NAME_EDUCATION_TYPE'

Inferences:

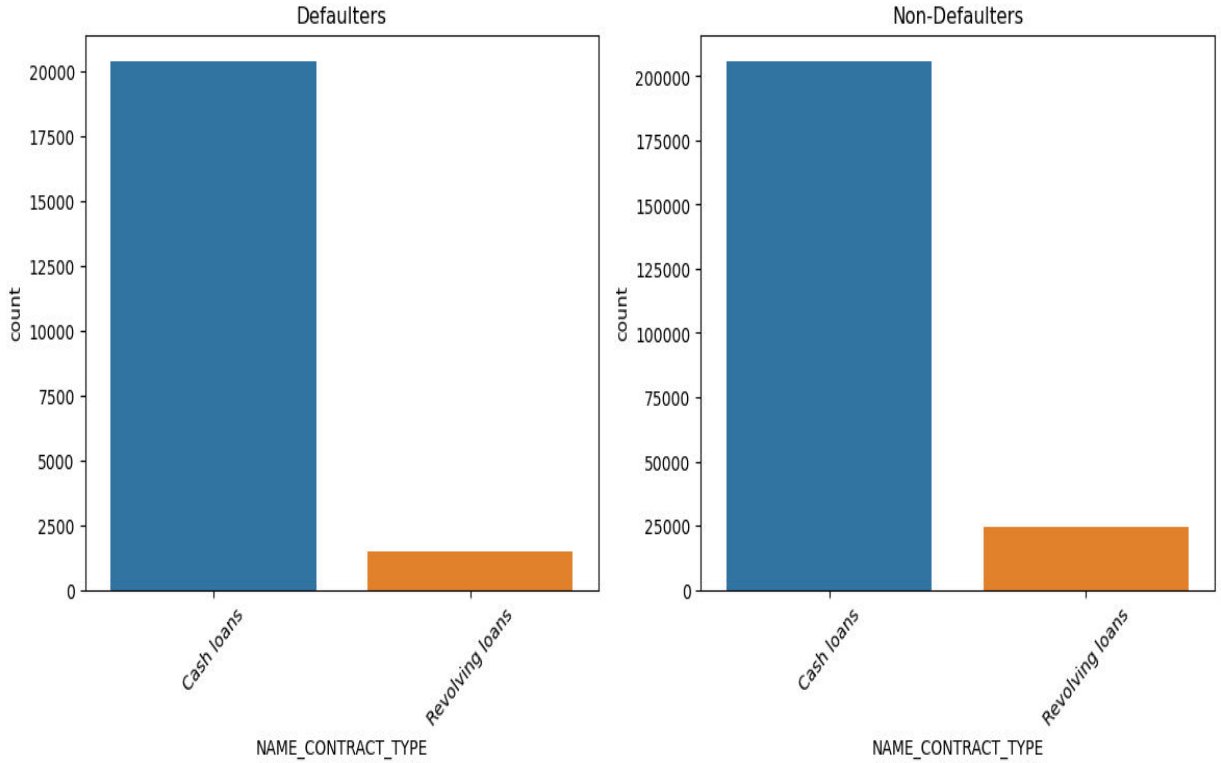
- It is observed that applicants having education type ‘secondary/secondary special’ are more likely to be defaulters than the non-defaulters.
- Applicants having education type ‘Higher Education’ are more likely to be Non-defaulters than the defaulters.
- Applicants with lower education level tends to take less loan.



UNIVARIATE ANALYSIS ON NAME_CONTRACT_TYPE

Inferences:

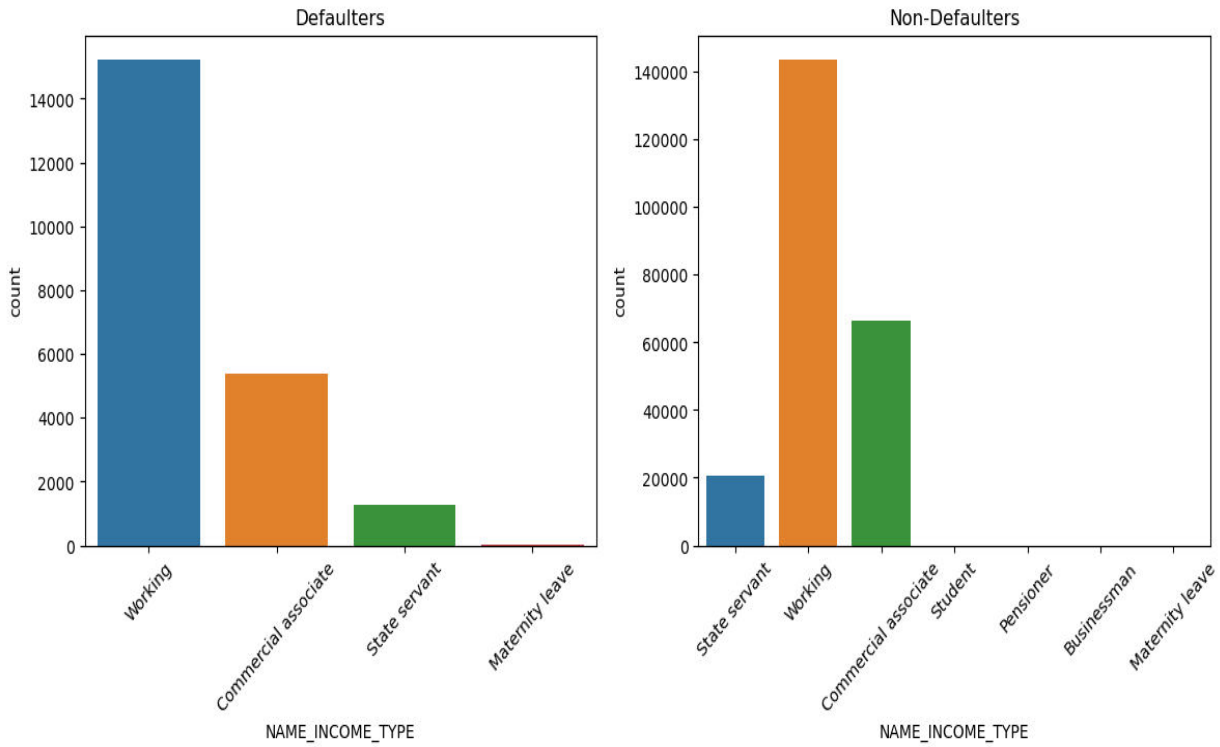
We see, in both the cases Revolving loans are very less in number compared to Cash loans and tends to have less defaulters.



UNIVARIATE ANALYSIS ON NAME_INCOME_TYPE

Inferences:

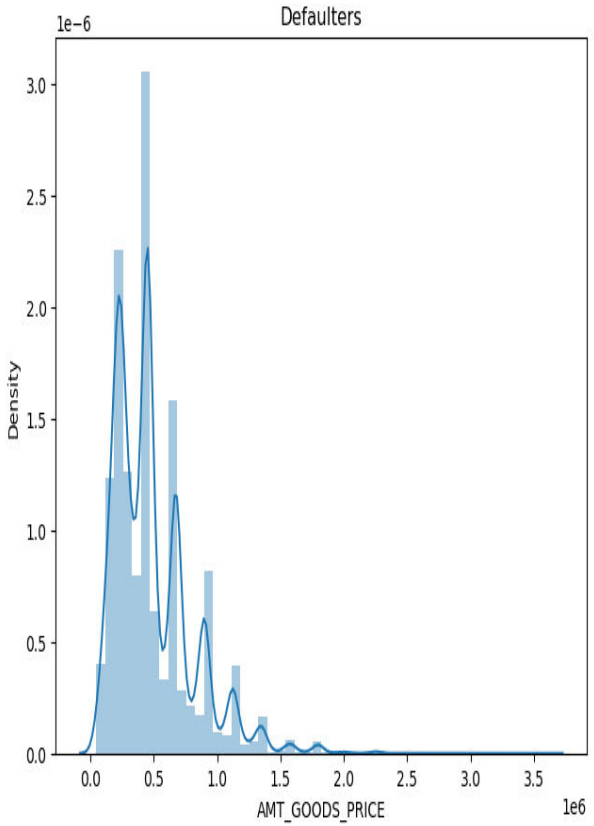
- Defaulters - Working people are mostly defaulted as their numbers are high with compare to other pfrofessions.
- Non-defaulters - Similarly here also working people are more in number who are not defaulted.

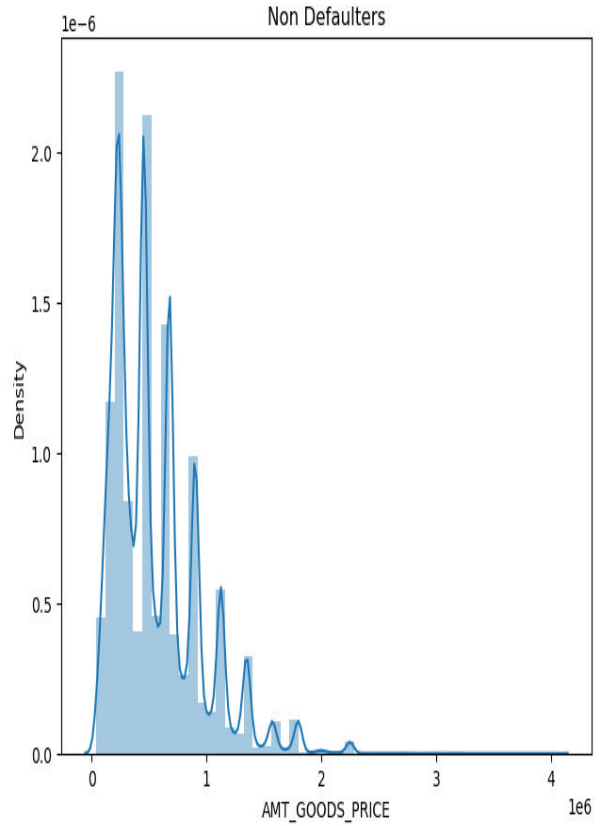


UNIVARIATE ANALYSIS ON AMT_GOODS_PRICE

Inferences:

- It is observed that applicants having higher goods price tends to be more defaulter than non-defaulters.
- When goods price are between 0-5,00,000. Applicants seems to be more non-defaulter than defaulters.

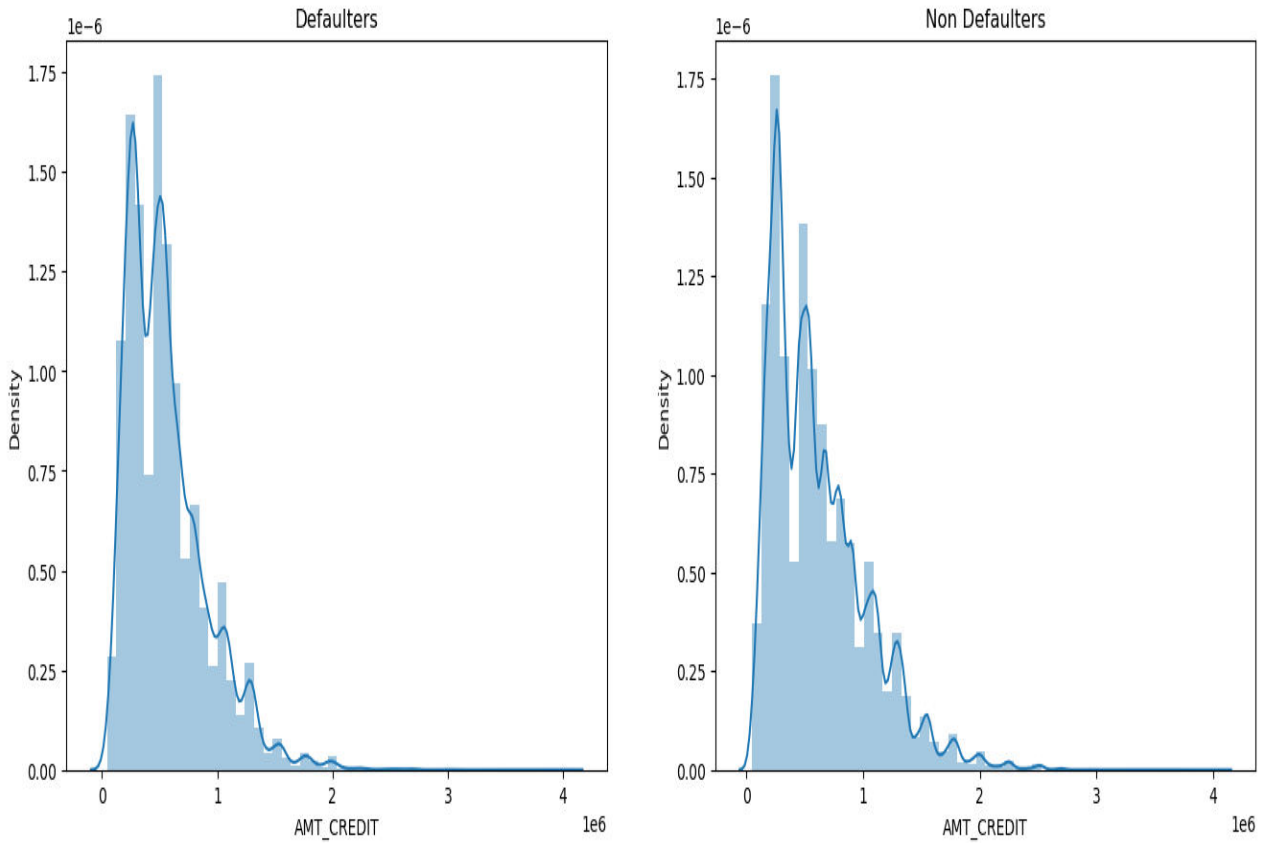




UNIVARIATE ANALYSIS ON AMT_CREDIT

Inferences:

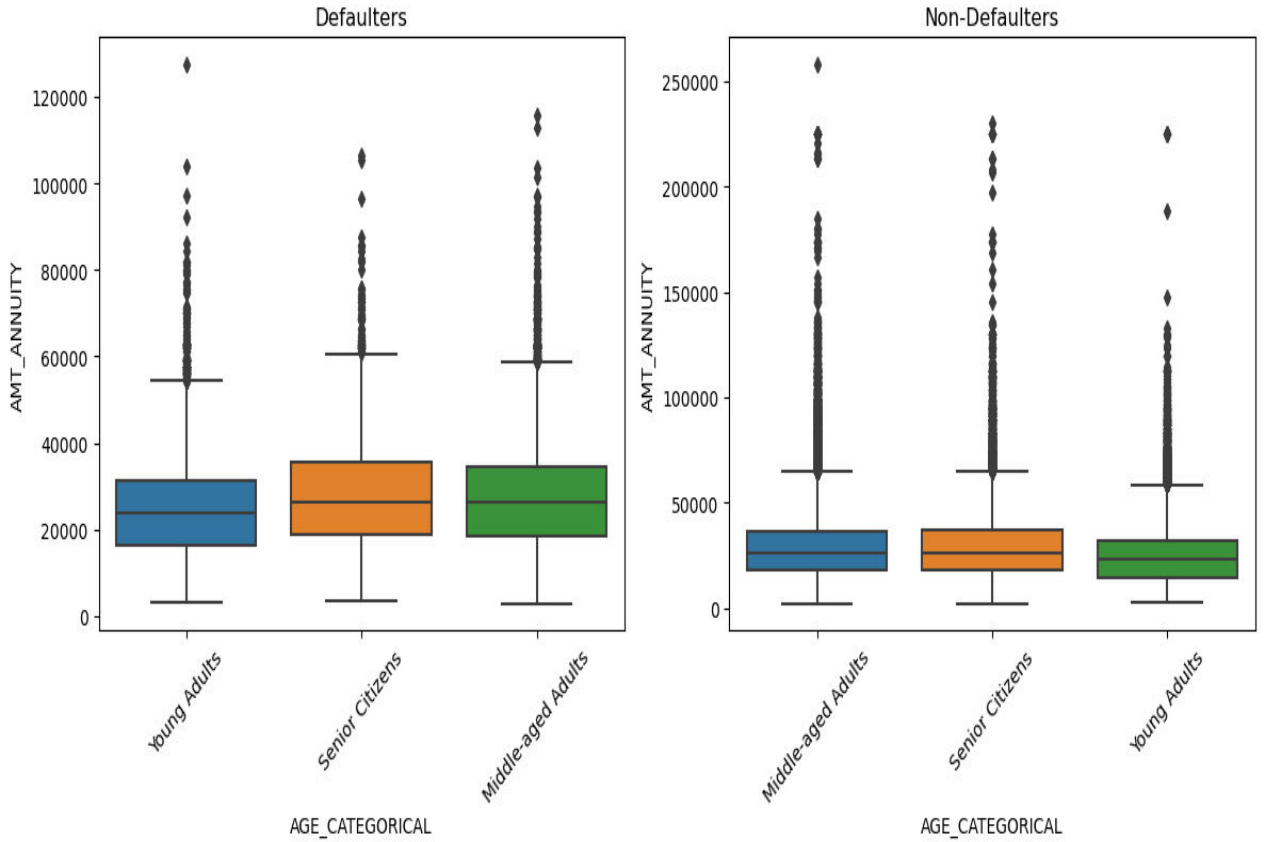
- It is observed that, credit amount is slightly higher for non defaulters.
- applicants having higher credited amount tends to be more defaulter than non-defaulters.



BIVARIATE ANALYSIS ON 'AGE_CATEGORICAL','AMT_ANNUITY'

Inferences:

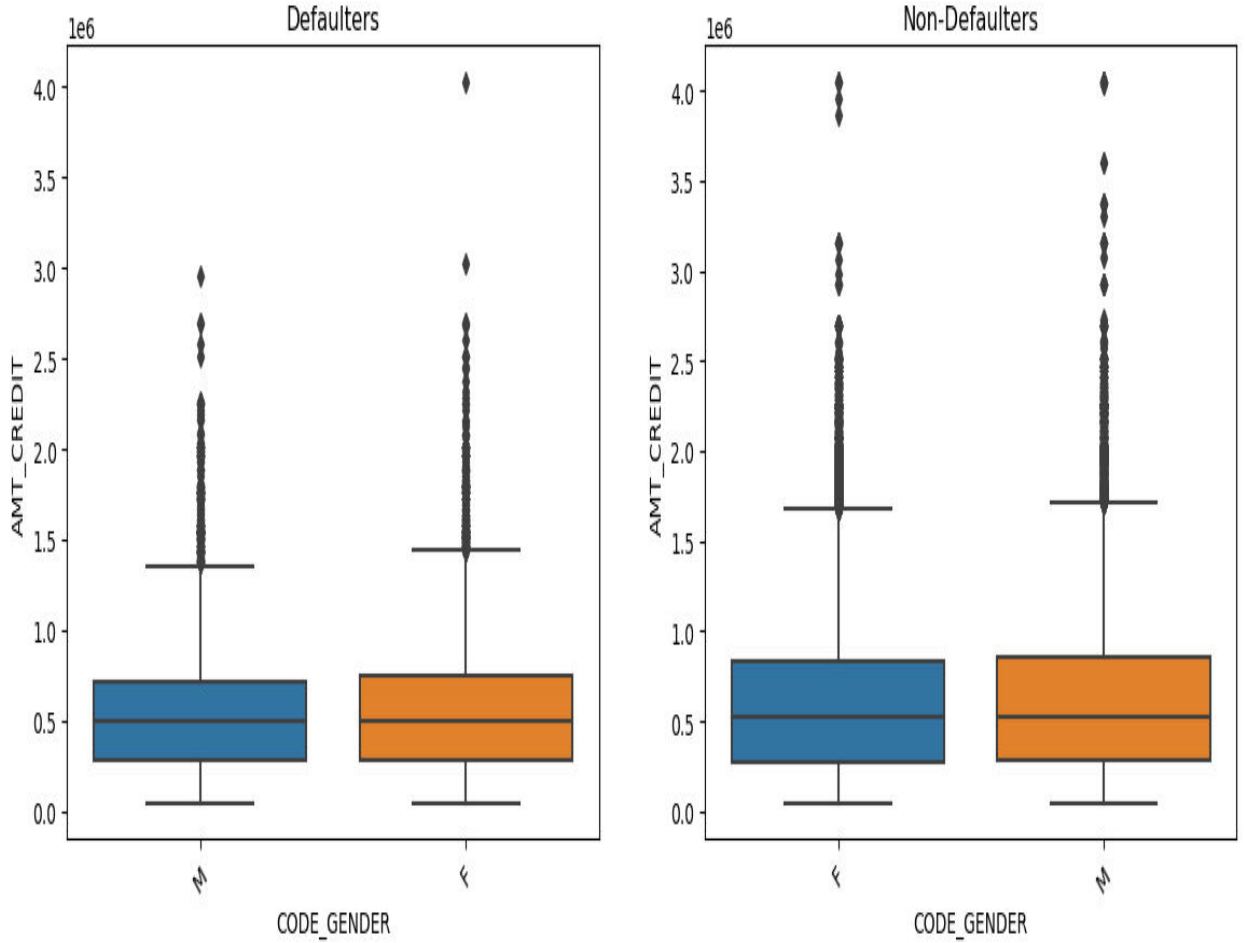
- Here we observe that overall, middle aged adults have the highest annuity loan and tends to be more non-defaulter.
- Whereas, young adults has moderate annuity and tends to be more defaulter.



BIVARIATE ANALYSIS ON 'CODE_GENDER','AMT_CREDIT'

Inferences:

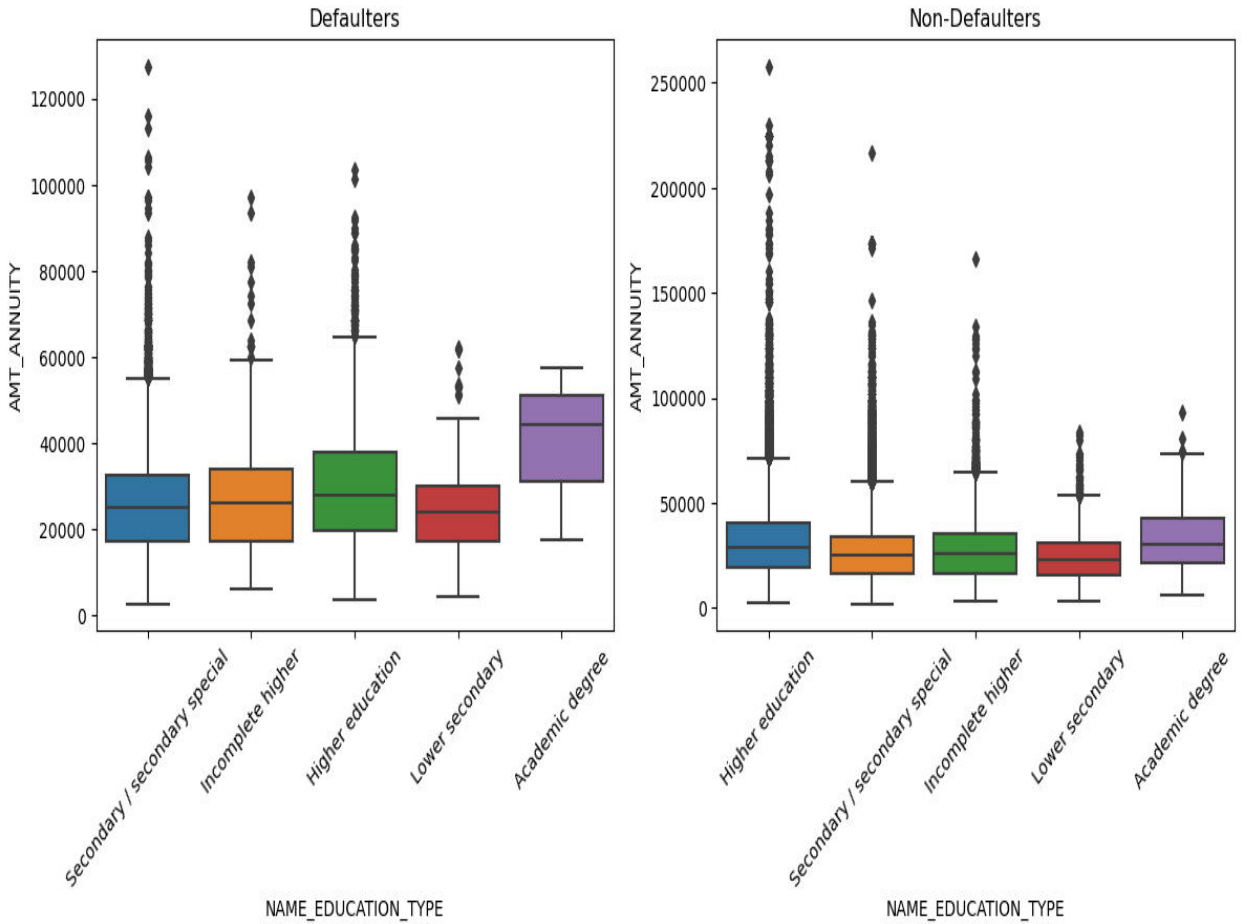
- It is observed females when credited with more amount tends to be a defaulter than males.
- The most amount is credited to males and are least defaulters.



BIVARIATE ANALYSIS ON 'NAME_EDUCATION_TYPE','AMT_ANNUITY'

Inferences:

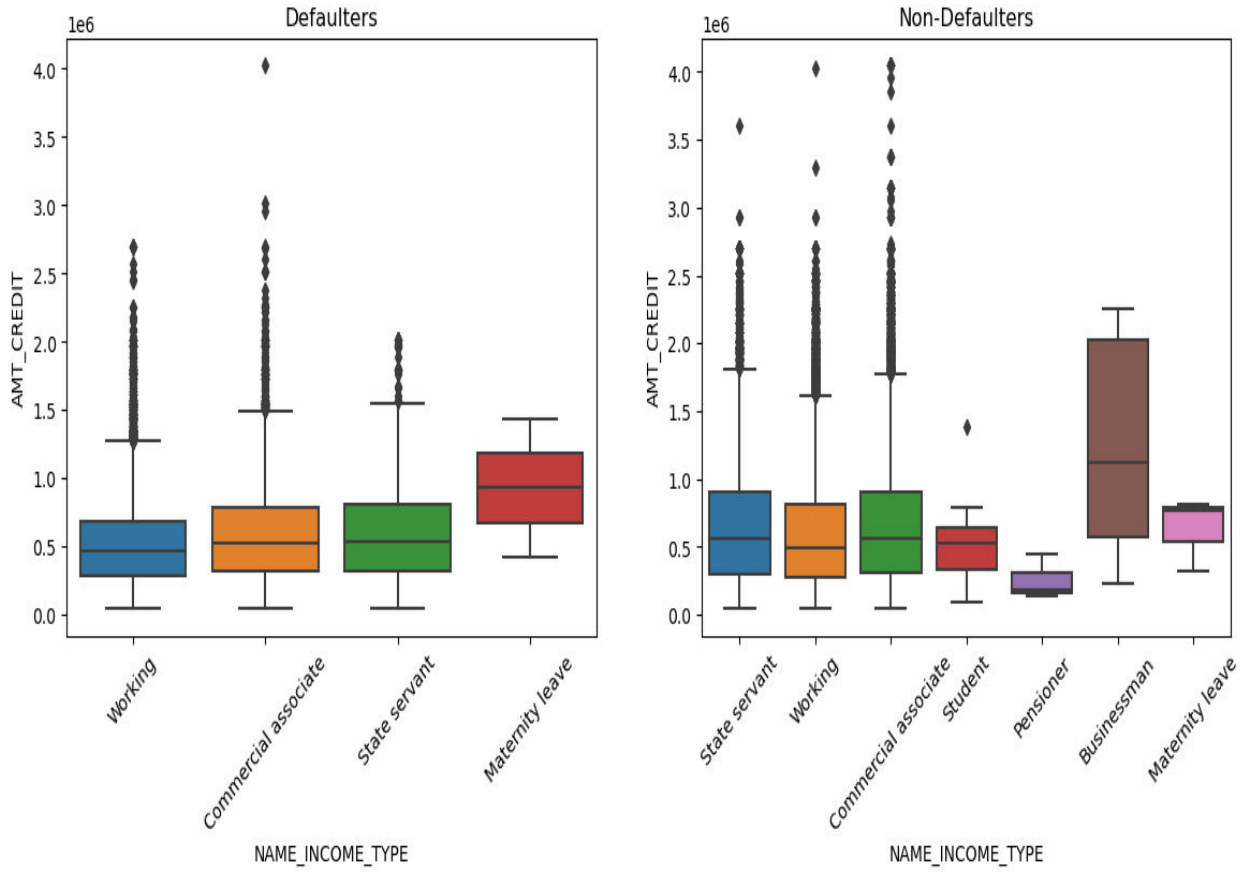
- It is observed that applicants with lower-secondary education has the least annuity and tends to be more defaulter.
- Applicants with Higher education has more annuity and tends to be defaulter.
- Applicants with secondary/secondary special education type has the maximum annuity and tends to be least defaulters



BIVARIATE ANALYSIS ON NAME_INCOME_TYPE','AMT_CREDIT'

Inferences:

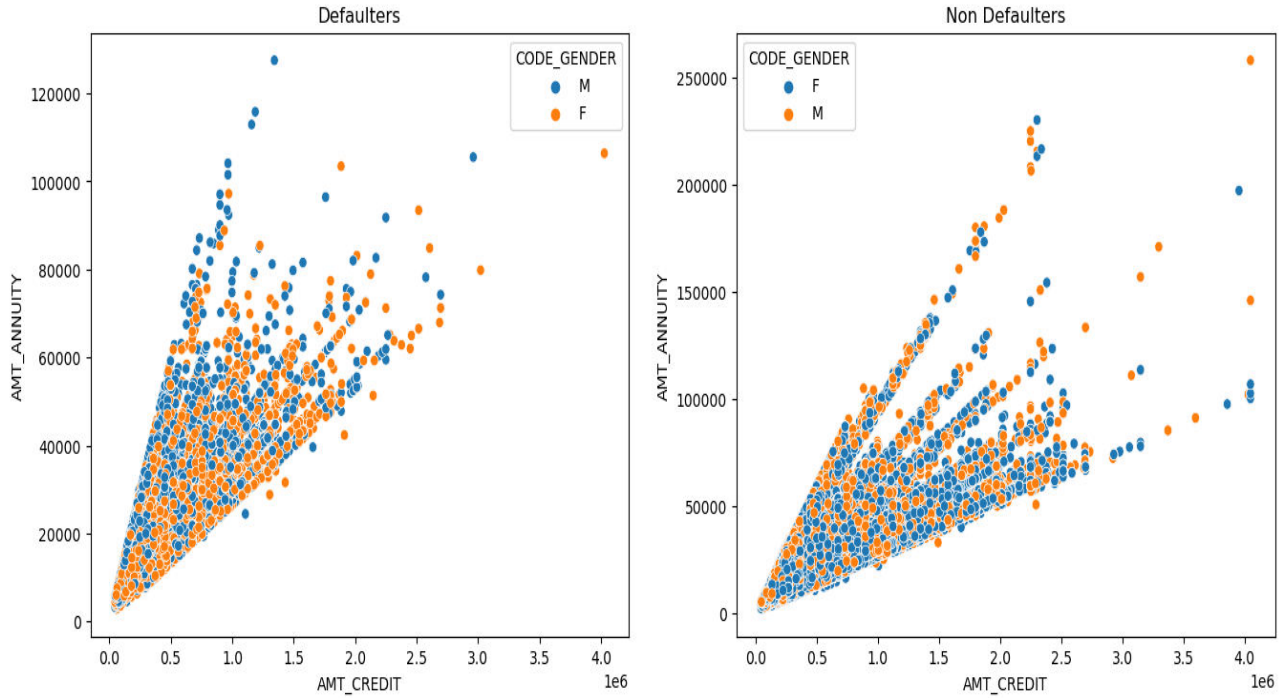
- Here we observe that businessman gets the most amount of loan and also very unlikely to be defaulters



BIVARIATE ANALYSIS ON 'AMT_CREDIT','AMT_ANNUITY','CODE_GENDER'

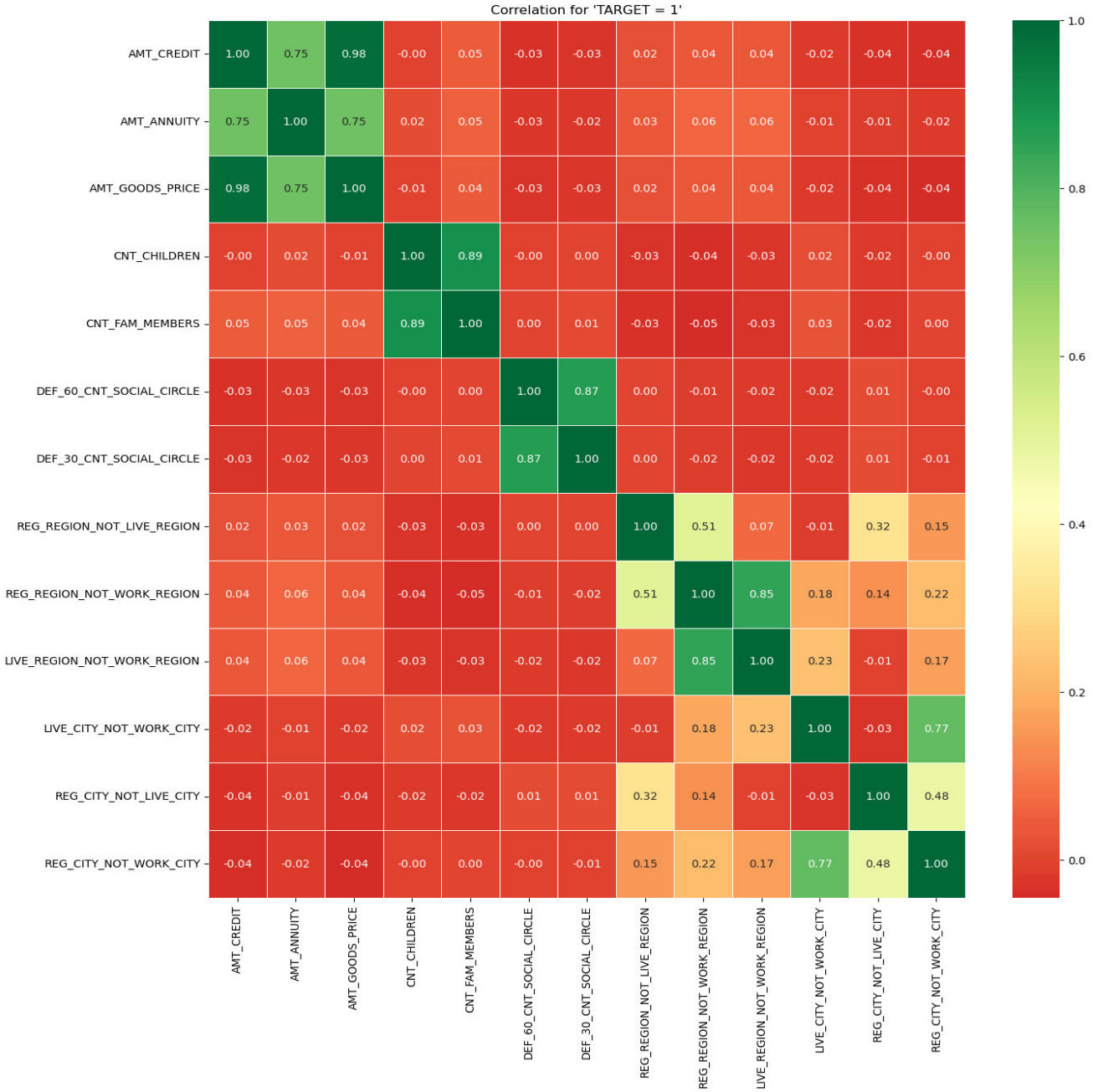
Inferences:

- It is observed that, when credited amount increases annuity also increases.
- when credited amount lies between 200000-400000 and annuity lies between 60000-120000, applicants tends to be more defaulter.
- In comparison to males, females are more likely to be defaulter.



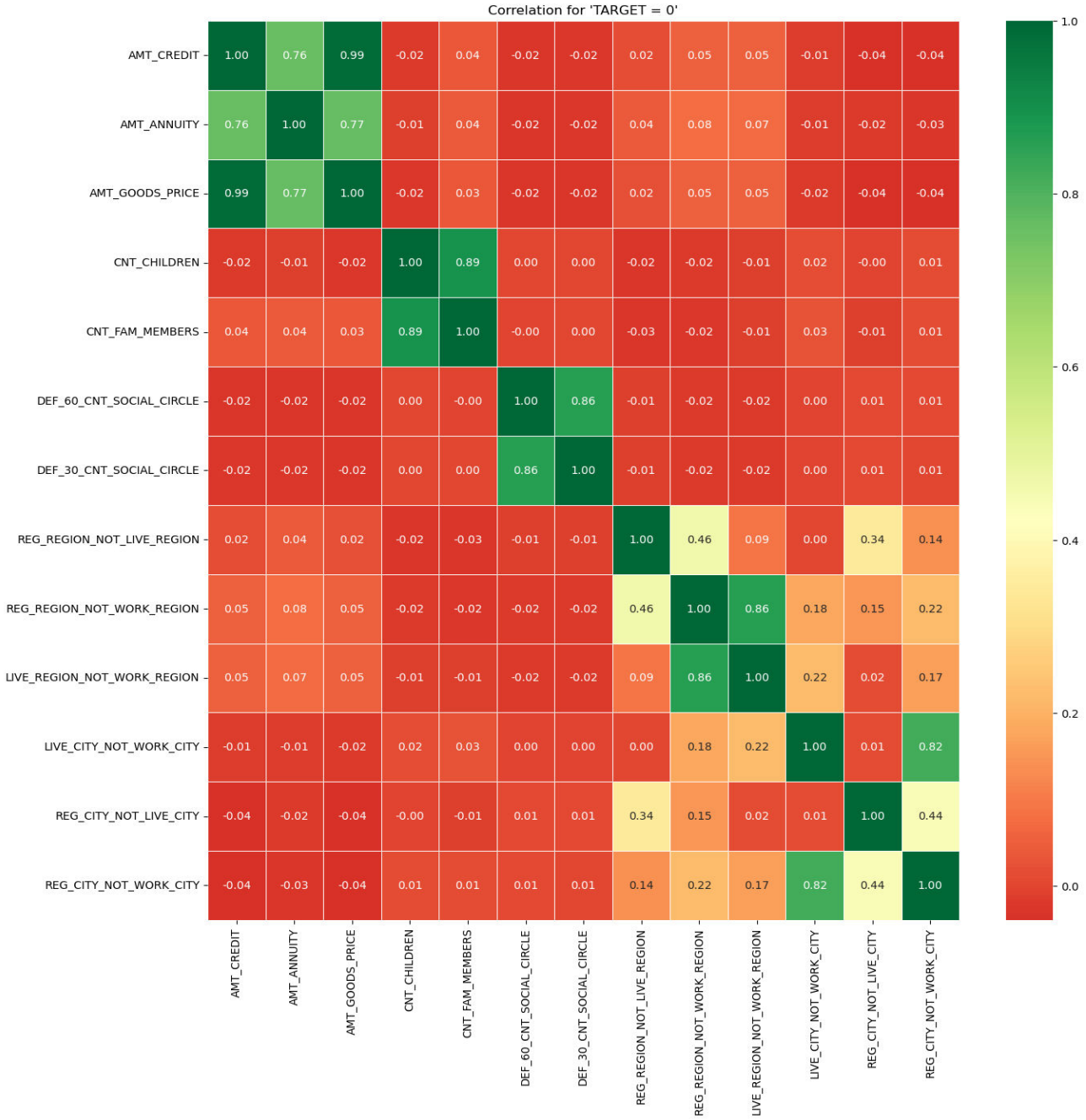
TOP 10 CORRELATIONS ON DEFAULTER (TARGET=1)

1. AMT_ CREDIT and AMT_GOODS_PRICE has a positive correlation of 0.982464
2. CNT_CHILDREN and CNT_FAM_MEMBERS has a positive correlation of 0.893829
3. DEF_60_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE has a positive correlation of 0.867983
4. LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION has a positive correlation of 0.846872
5. REG_CITY_NOT_WORK_CITY and LIVE_CITY_NOT_WORK_CITY has a positive correlation of 0.768247
6. AMT_GOODS_PRICE and AMT_ ANNUITY has a positive correlation of 0.748940
7. AMT_CREDIT and AMT_ANNUITY has a positive correlation of 0.748708
8. REG_REGION_NOT_LIVE_REGION and REG_REGION_NOT_WORK_REGION has a positive correlation of has a positive correlation of 0.506747
9. REG_CITY_NOT_LIVE_CITY and REG_CITY_NOT_WORK_CITY has a positive correlation of 0.478266
10. REG_REGION_NOT_LIVE_REGION and REG_CITY_NOT_LIVE_CITY has a positive correlation of 0.322030



TOP 10 CORRELATIONS ON NON-DEFAULTER (TARGET=0)

1. AMT_ CREDIT and AMT_GOODS_PRICE has a positive correlation of 0.986471
2. CNT_CHILDREN and CNT_FAM_MEMBERS has a positive correlation of 0.893275
3. DEF_60_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE has a positive correlation of 0.861492
4. LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION has a positive correlation of 0.860421
5. REG_CITY_NOT_WORK_CITY and LIVE_CITY_NOT_WORK_CITY has a positive correlation of 0.820828
6. AMT_GOODS_PRICE and AMT_ ANNUITY has a positive correlation of 0.766669
7. AMT_CREDIT and AMT_ANNUITY has a positive correlation of 0.762117
8. REG_REGION_NOT_LIVE_REGION and REG_REGION_NOT_WORK_REGION has a positive correlation of has a positive correlation of 0.461596
9. REG_CITY_NOT_LIVE_CITY and REG_CITY_NOT_WORK_CITY has a positive correlation of 0.442640
10. REG_REGION_NOT_LIVE_REGION and REG_CITY_NOT_LIVE_CITY has a positive correlation of 0.342321

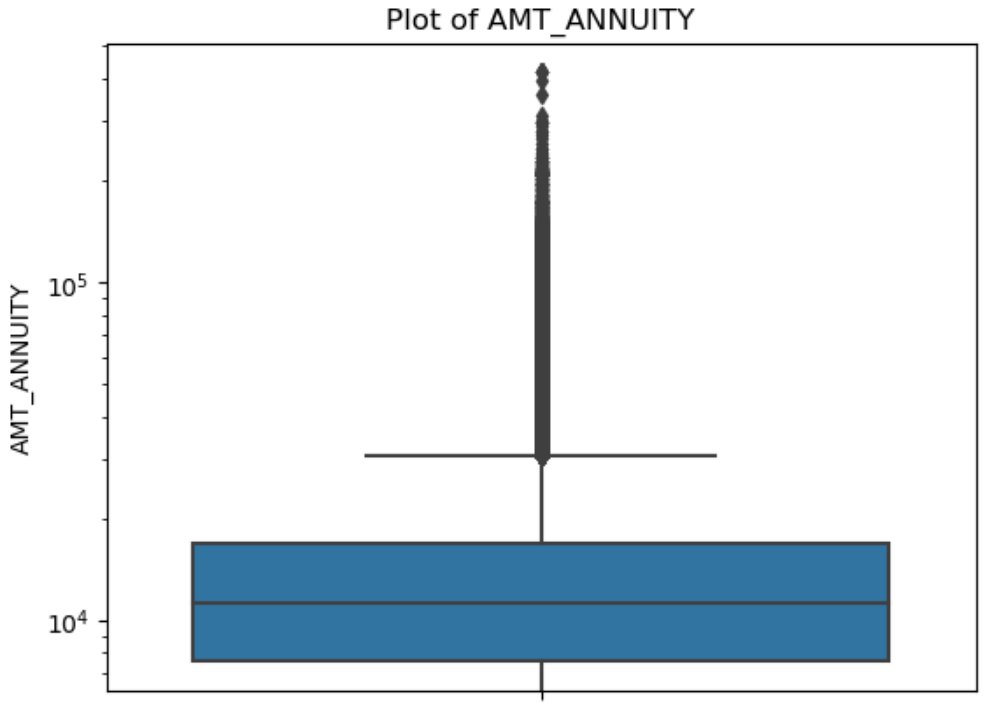


Analysis on ‘previous_application.csv’

OUTLIER ANALYSIS ON AMT_ANNUITY

Insights:

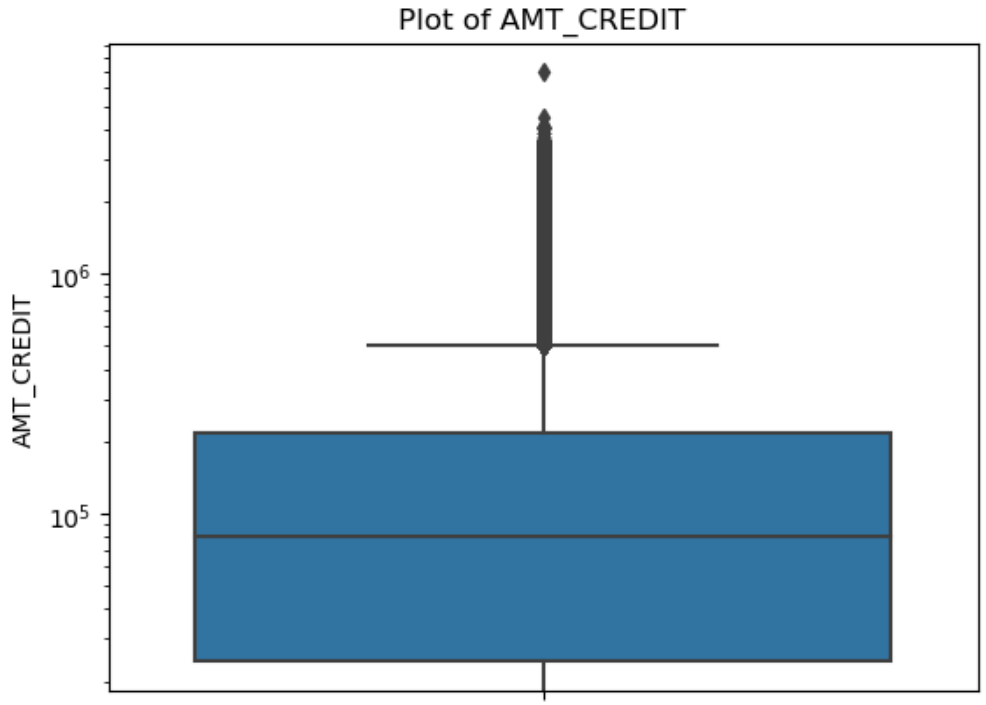
- Some of outliers are noticed in AMT_ANNUITY column.
- This indicates that annuity client are borrowing large amount



OUTLIER ANALYSIS ON AMT_CREDIT

Insights:

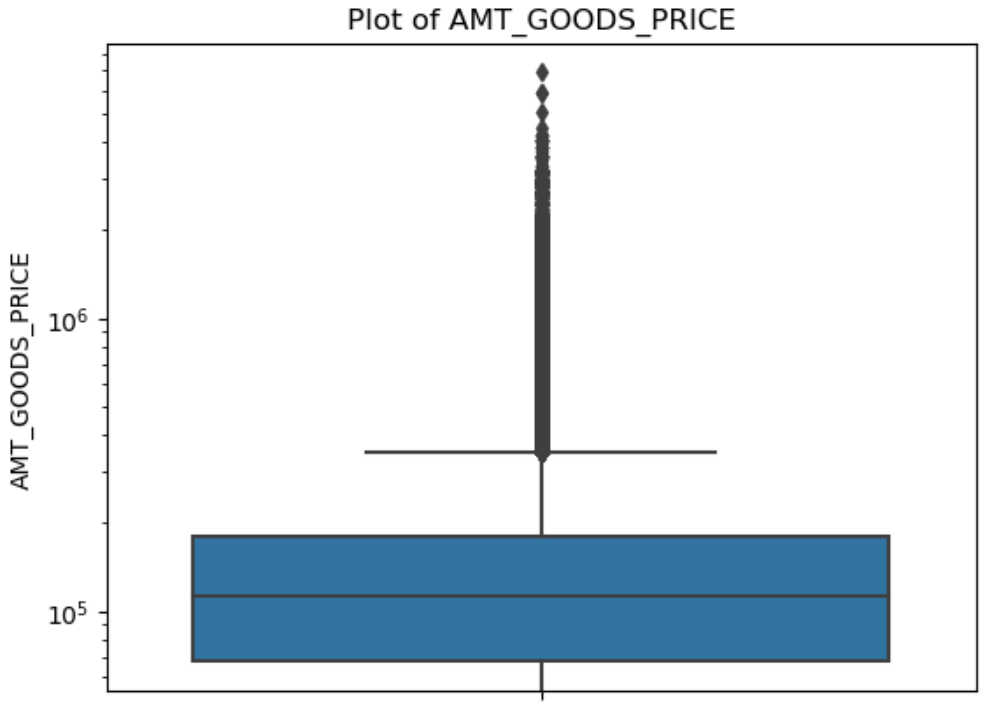
- Some of outliers are noticed in AMT_CREDIT column.
- This indicates that some clients received more credit amount than the amount they applier for.
- The first quartile is bigger than third quartile for credit amount which means most of the clients are from first quartile



OUTLIER ANALYSIS ON AMT_GOODS_PRICE

Insights:

- Some of outliers are noticed in AMT_CREDIT column.
- This indicates that some clients has demanded for higher goods price.

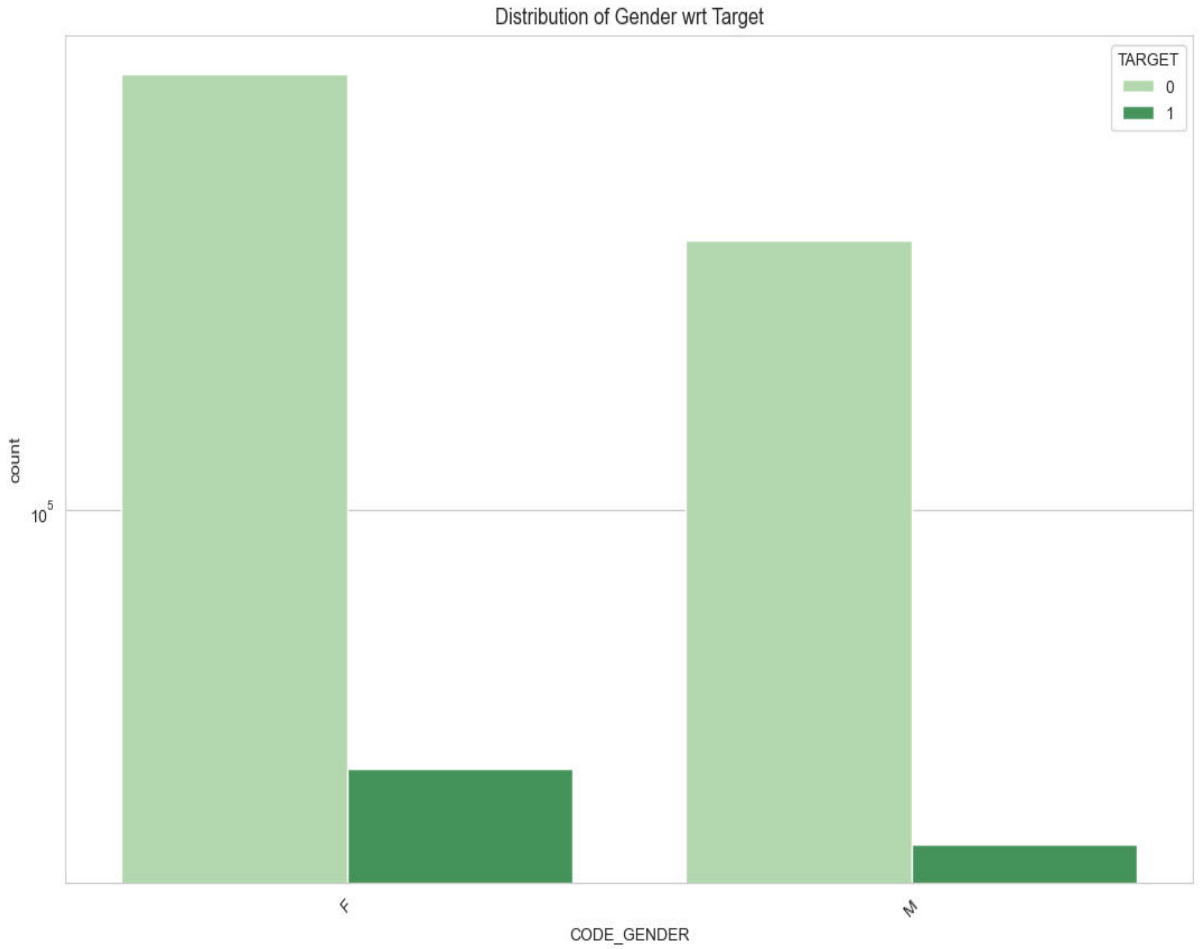


Analysis After Merging Both csv files ‘application_data.csv’ & ‘previous_application.csv’

ANALYSIS ON CODE_GENDER W.R.T TARGET

Inferences:

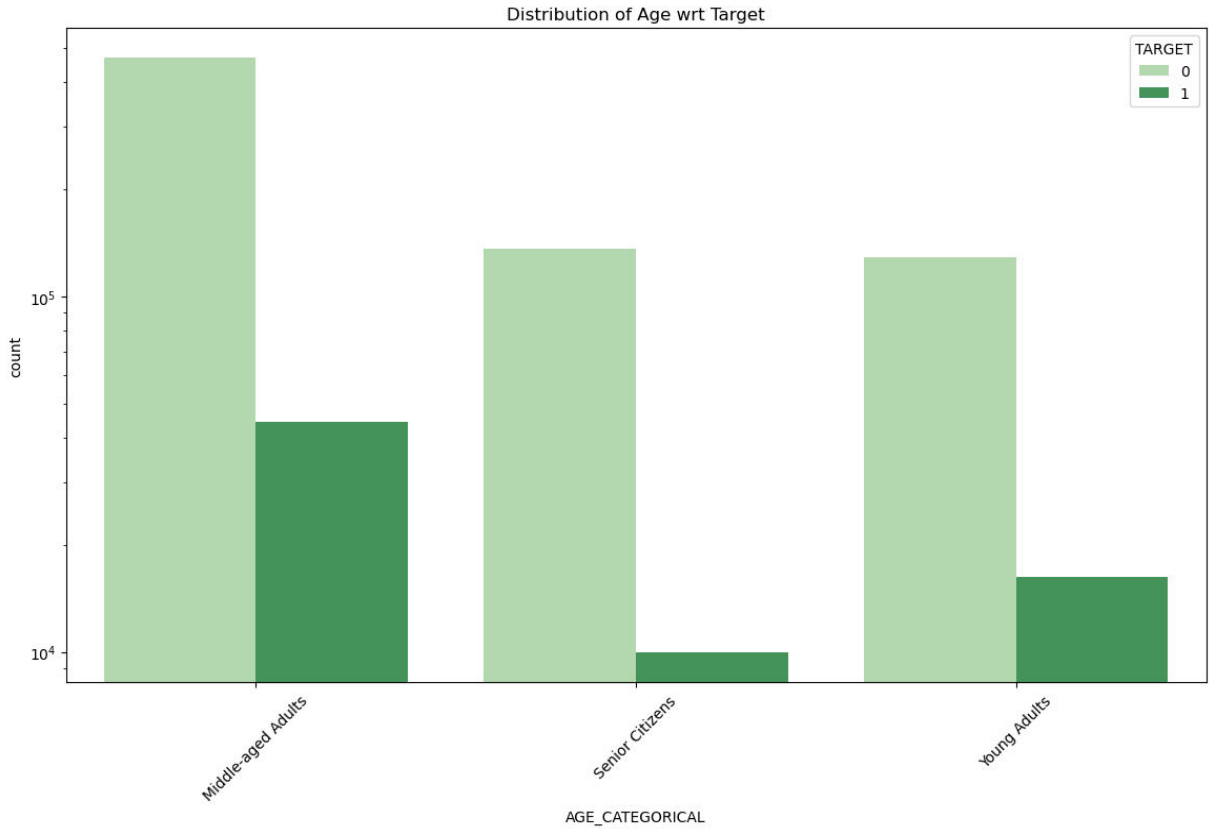
- It is observed that, males are less defaulter than females.
- Count of females are more than males.



ANALYSIS ON AGE_CATEGORICAL W.R.T TARGET

Inferences:

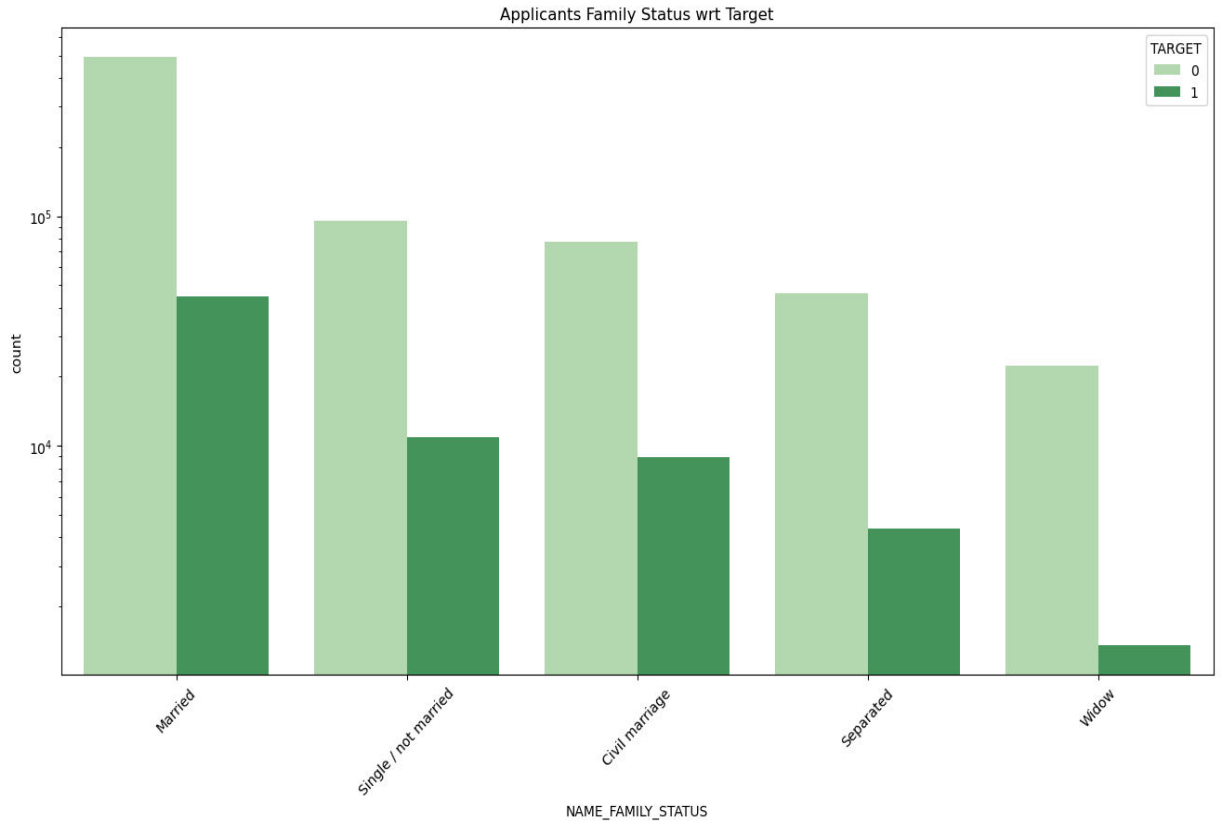
- Senior Citizens are the least defaulters
- Middle aged adults are the most defaulters.



ANALYSIS ON NAME_FAMILY_STATUS W.R.T TARGET

Inferences:

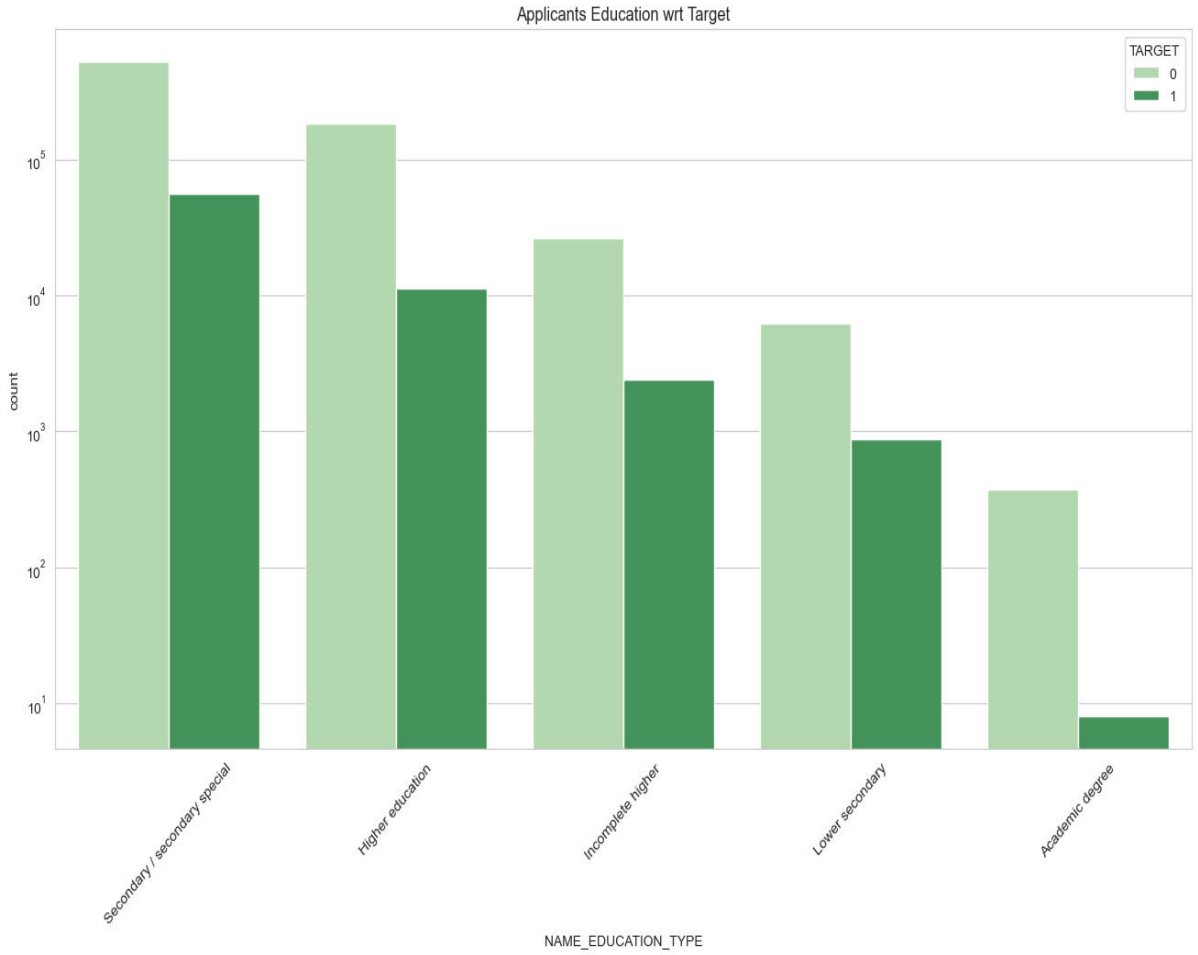
- widows are the least defaulter.
- married are the most defaulter.



ANALYSIS ON NAME_EDUCATION_TYPE W.R.T TARGET

Inferences:

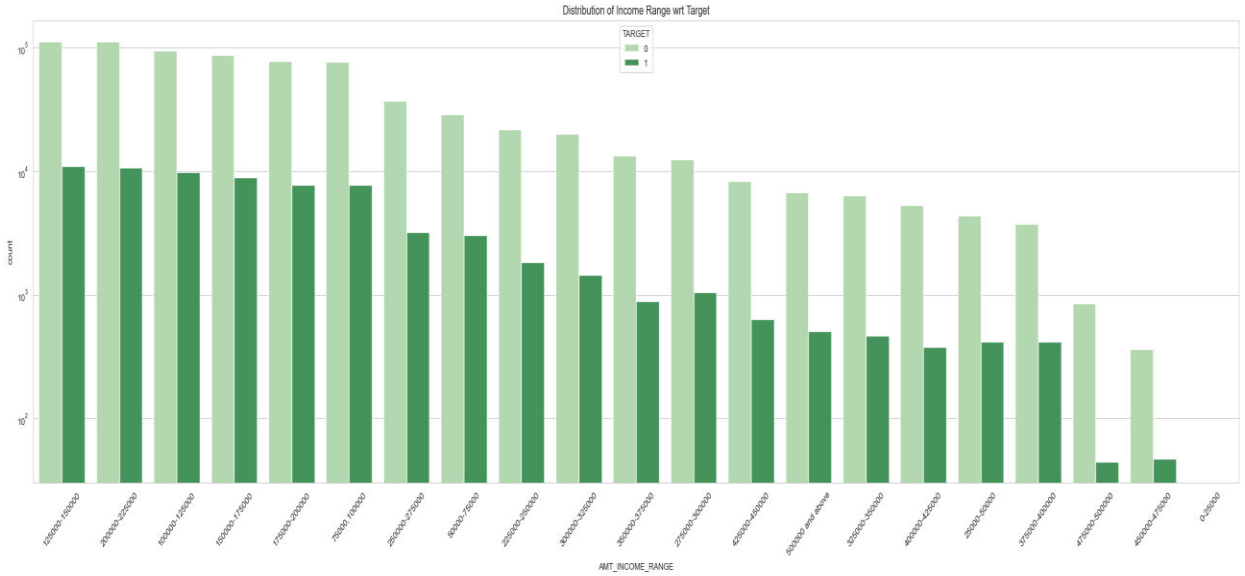
- It is observed that applicants having education type ‘secondary/secondary special’ are more likely to be defaulters than the non-defaulters.
- Applicants having education type ‘Higher Education’ are more likely to be Non-defaulters than the defaulters.
- Applicants with lower education level tends to take less loan.



ANALYSIS ON AMT_INCOME_RANGE W.R.T TARGET

Inferences:

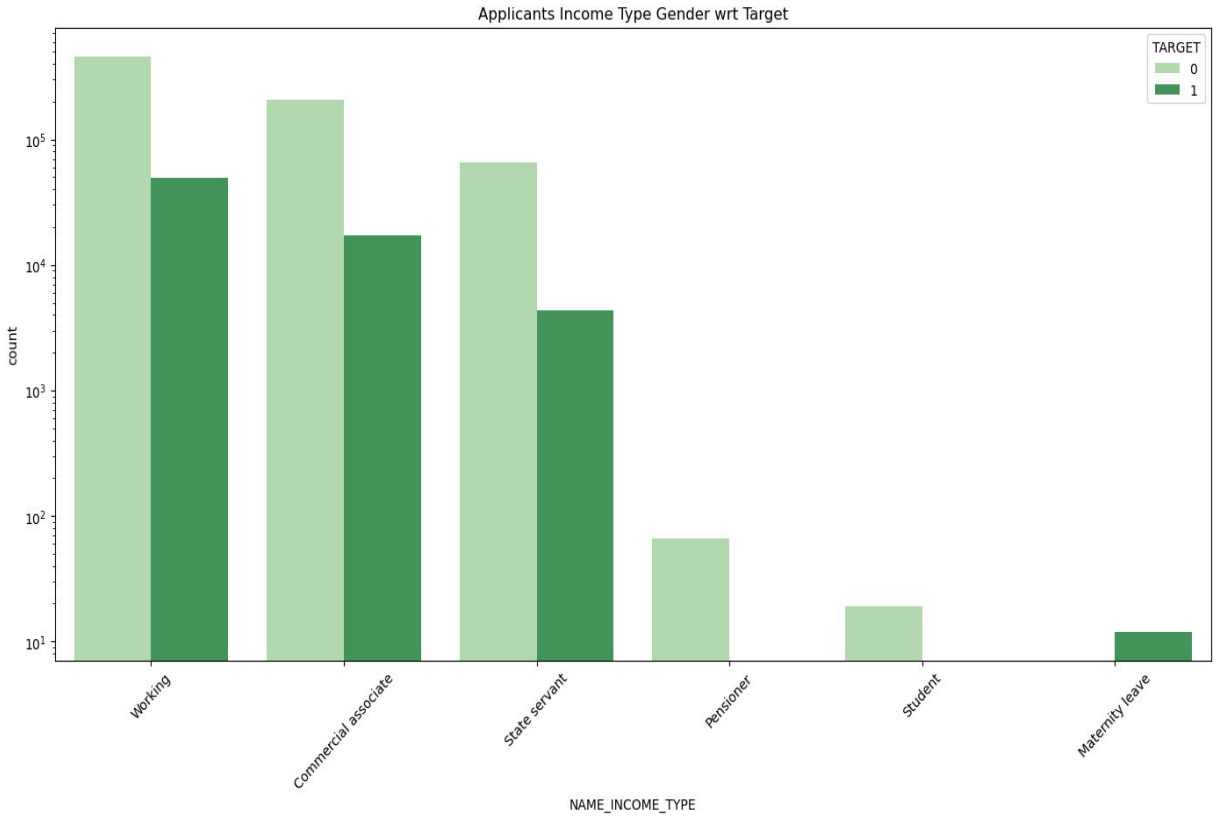
- Applicants with higher income range tends to be less defaulter than lower income range.



ANALYSIS ON NAME_INCOME_TYPE W.R.T TARGET

Inferences:

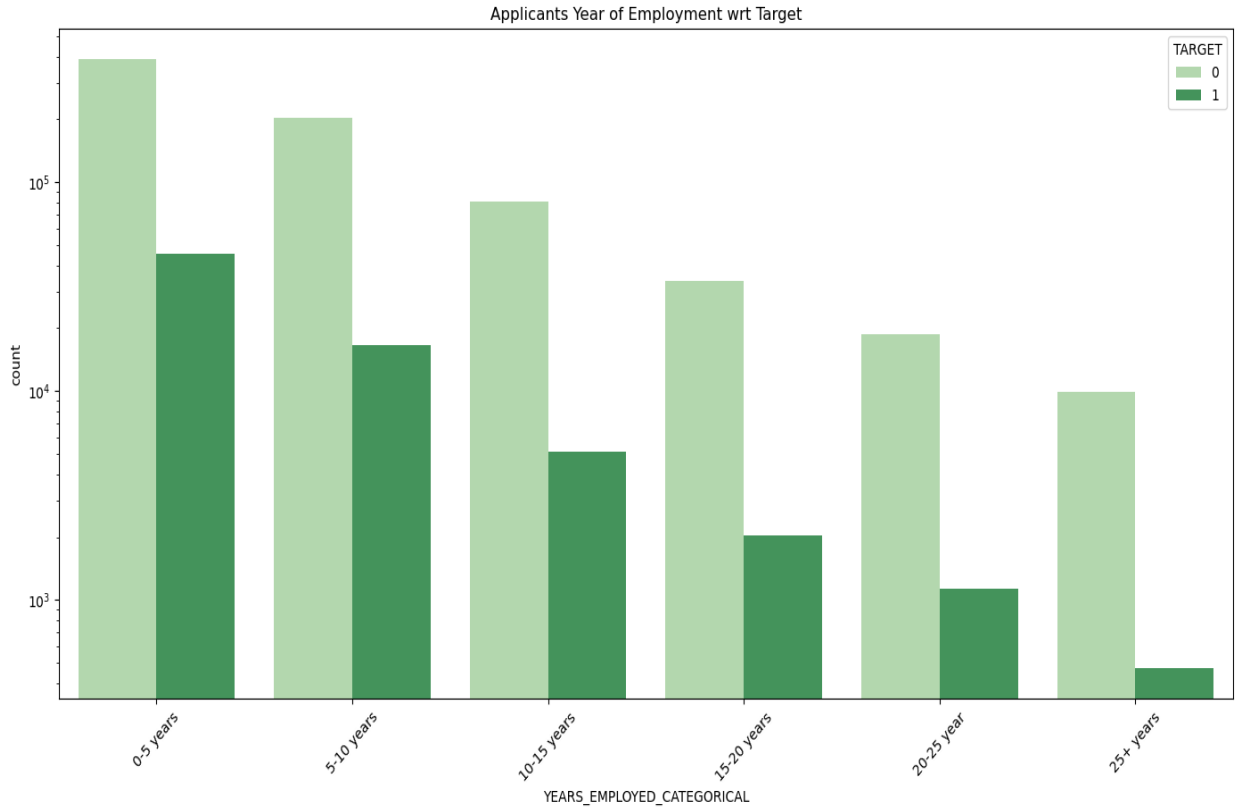
- Working professional are the most defaulters followed by Commercial associate and state servants.
- Students, pensioner are the least defaulters followed by Maternity leave.



ANALYSIS ON YEARS_EMPLOYED_CATEGORICAL W.R.T TARGET

Inferences:

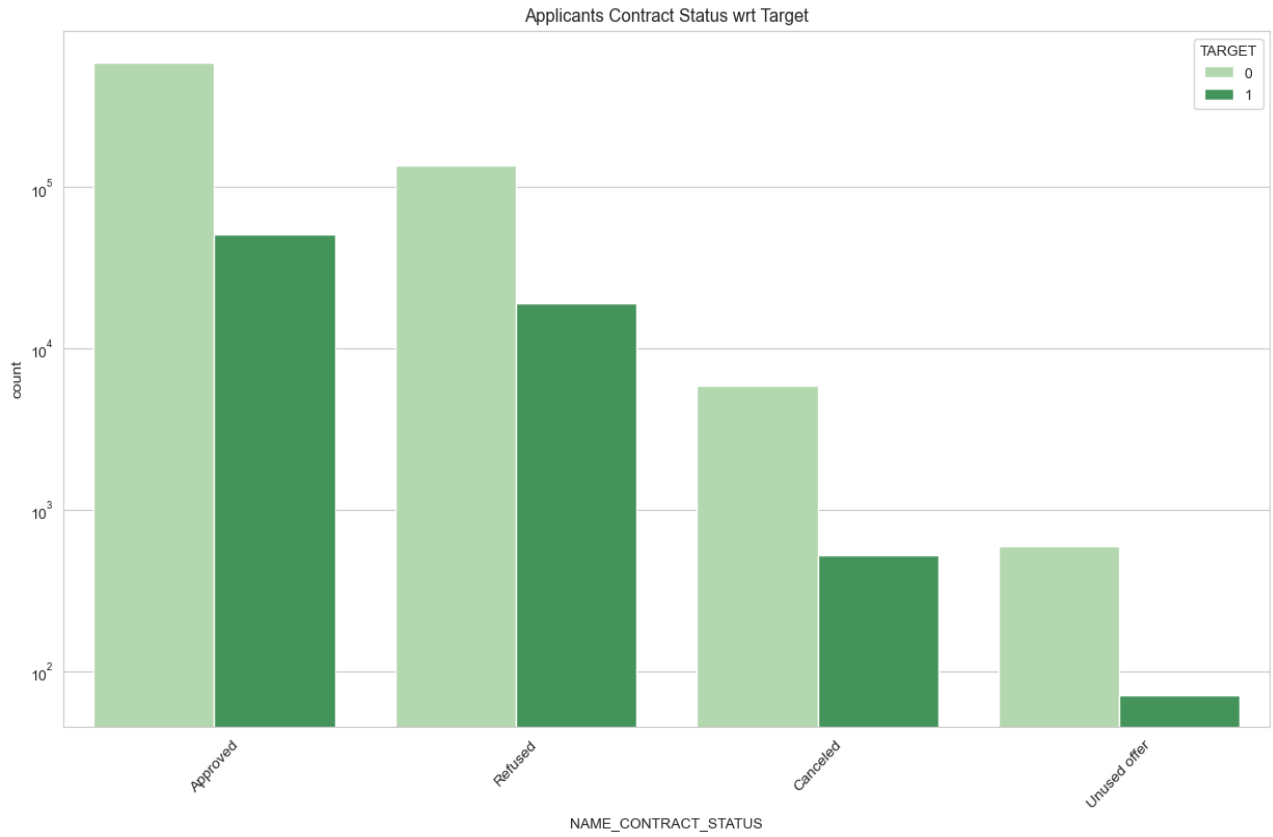
Applicants with more years of employment tends to be less defaulter than with less years of employment



ANALYSIS ON NAME_CONTRACT_STATUS W.R.T TARGET

Inferences:

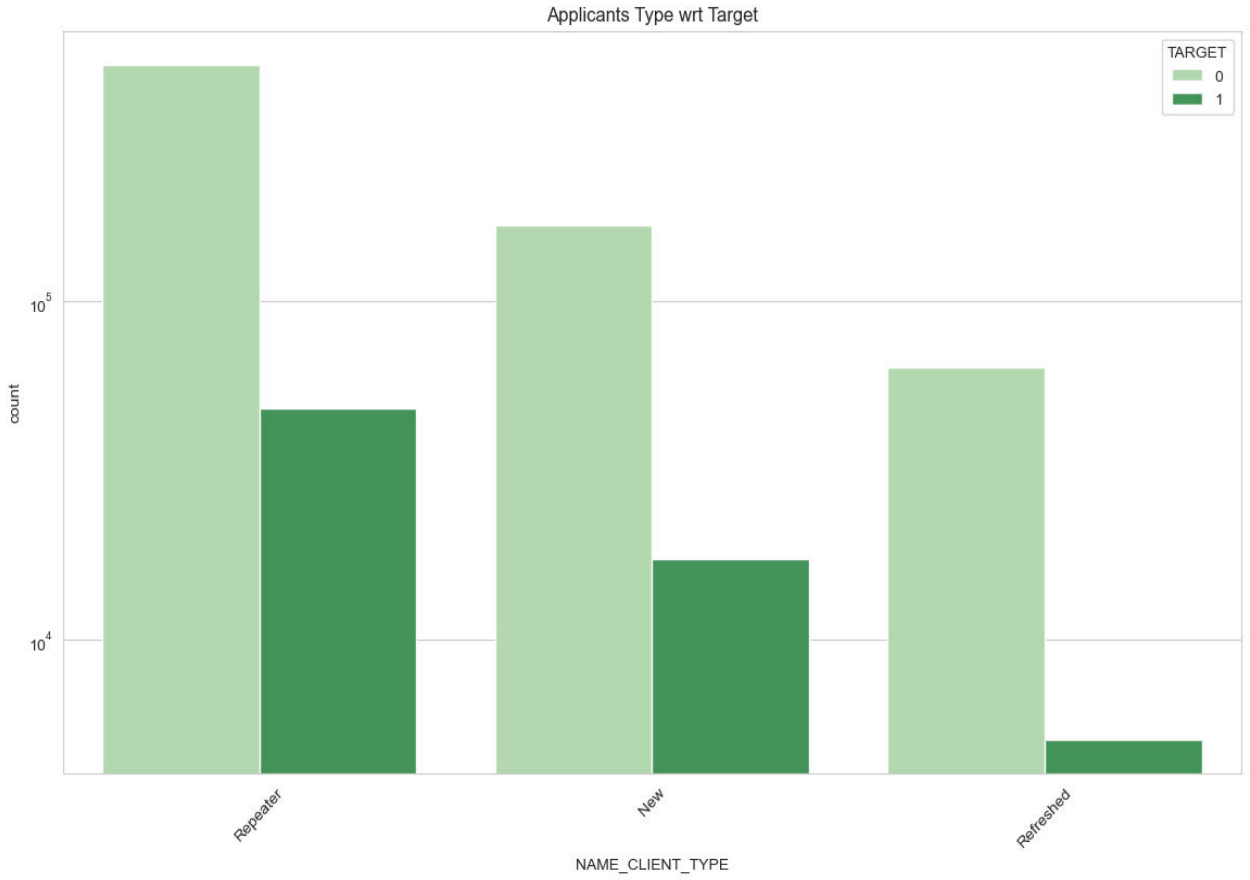
- Applicants when contract status is approved tends to be more defaulter than others.



ANALYSIS ON NAME_CLIENT_TYPE W.R.T TARGET

Inferences:

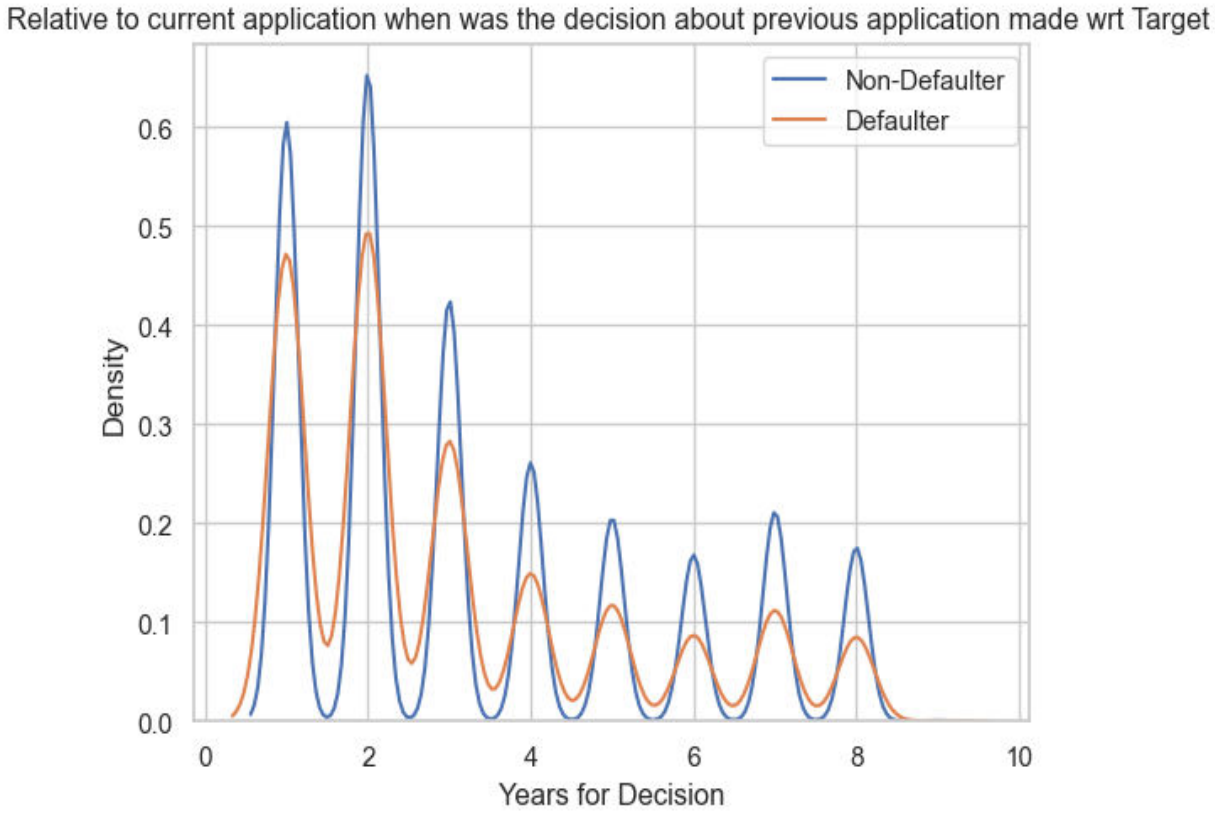
- Applicants of type ‘Refreshed’ are tends to be the least defaulter.
- Applicants of type ‘Repeater’ are tends to be the most defaulter.



ANALYSIS ON YEARS_DECISION W.R.T TARGET

Inferences:

- Relative to current application when the decision about previous application made early are tends to be more defaulter.
- Relative to current application when the decision about previous application made late are tends to be least defaulter.



ANALYSIS ON AMT_ANNUITY, AMT_CREDIT W.R.T TARGET

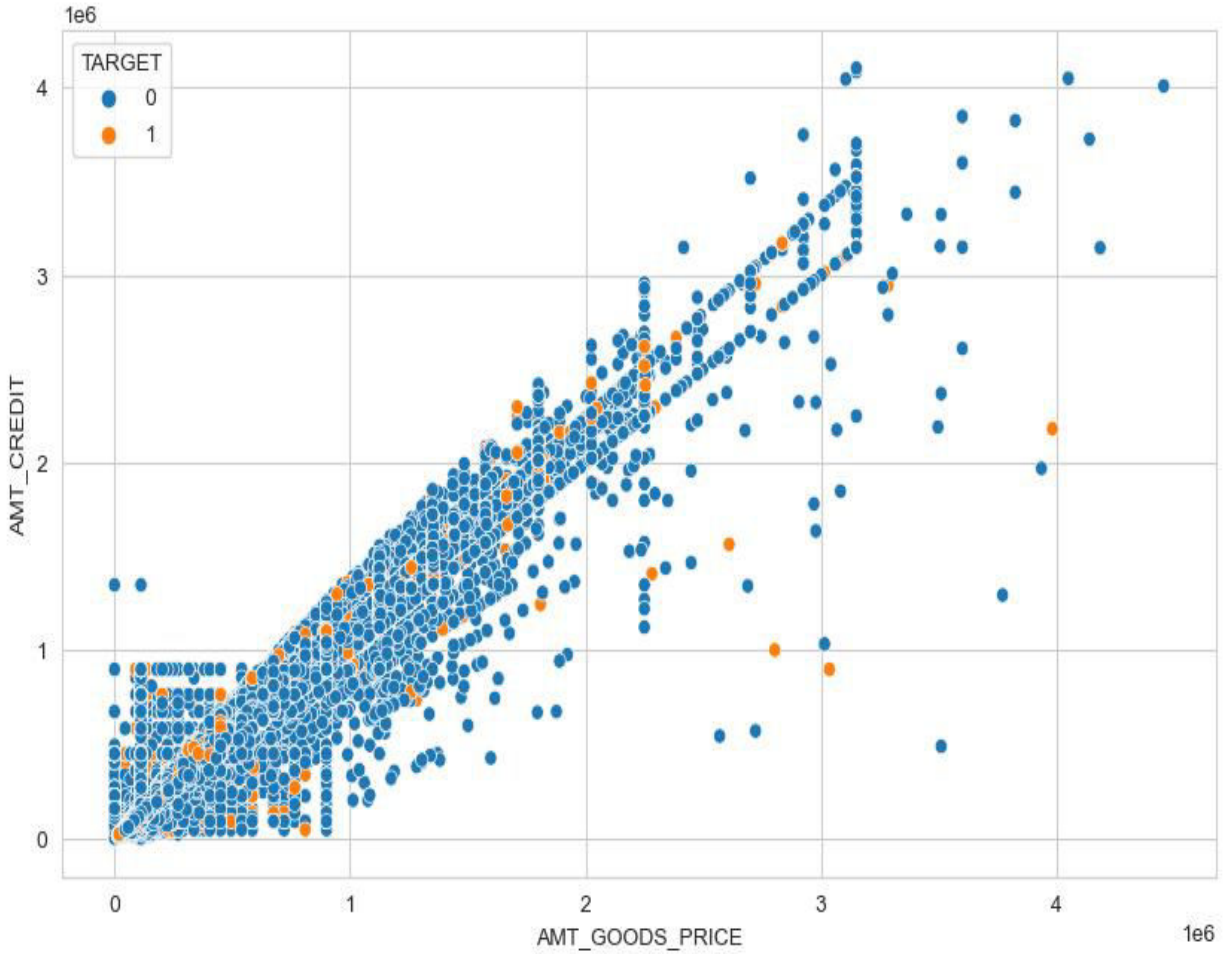
Inferences:

- When credit amount is increase, annuity amount also increases.
- when the amount credited lies between 2000000-3500000 and annuity amount lies between 100000-300000, applicants tends to be more defaulter.

ANALYSIS ON AMT_GOODS_PRICE, AMT_CREDIT W.R.T TARGET

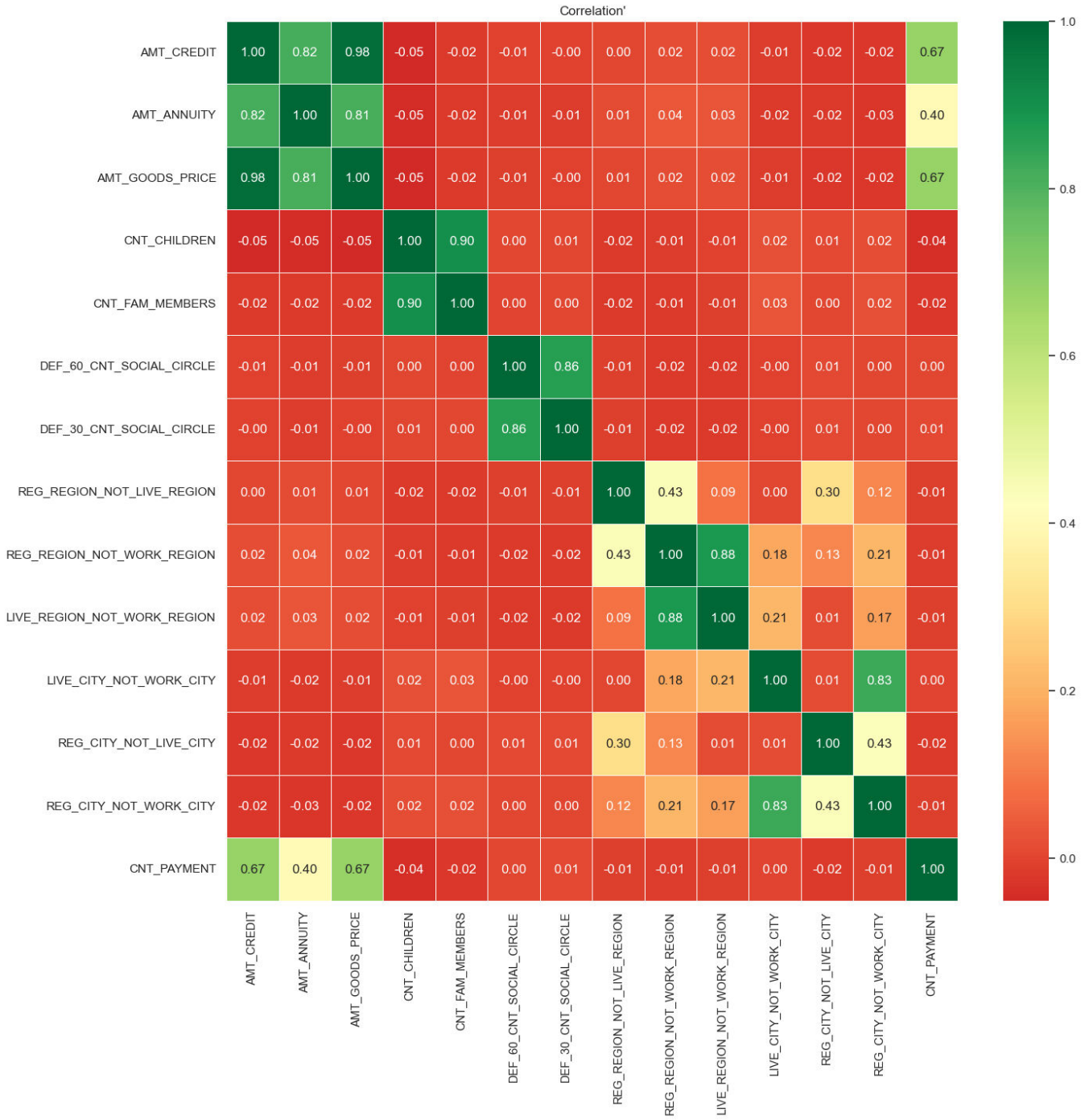
Inferences:

- It is observed that when amount of goods price is increased, amount credit is less.
- Applicants with higher goods price and higher credited amount are tends to be more defaulter.



Top 10 Correlations

1. AMT_ CREDIT and AMT_GOODS_PRICE has a positive correlation of 0.980139
2. CNT_CHILDREN and CNT_FAM_MEMBERS has a positive correlation of 0.898487
3. LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION has a positive correlation of 0.878678
4. DEF_60_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE has a positive correlation of 0.861250
5. REG_CITY_NOT_WORK_CITY and LIVE_CITY_NOT_WORK_CITY has a positive correlation of 0.828603
6. AMT_CREDIT and AMT_ANNUITY has a positive correlation of 0.817685
7. AMT_GOODS_PRICE and AMT_ ANNUITY has a positive correlation of 0.811931
8. AMT_CREDIT and CNT_PAYMENT has a positive correlation of 0.673977
9. AMT_GOODS_PRICE and CNT_PAYMENT has a positive correlation of 0.673027
10. REG_REGION_NOT_LIVE_REGION and REG_REGION_NOT_WORK_REGION has a positive correlation of has a positive correlation of 0.433681



CONCLUSION

In conclusion, the insights gathered from this project shed light on various aspects of loan applicant's characteristics and behaviours. The analysis of income totals revealed the presence of outliers, indicating individuals with significantly higher incomes compared to the majority. This highlights the importance of considering income disparity when assessing borrower risk and tailoring financial products to cater to different income levels.

Overall, these insights contribute to a more nuanced understanding of loan applicants' characteristics, allowing financial institutions to develop tailored loan products, assess borrower risk more effectively, and provide inclusive financial solutions that cater to the diverse needs of their clients. By leveraging these insights, lenders can enhance their decision-making processes, promote responsible lending practices, and support the financial well-being of their customers.