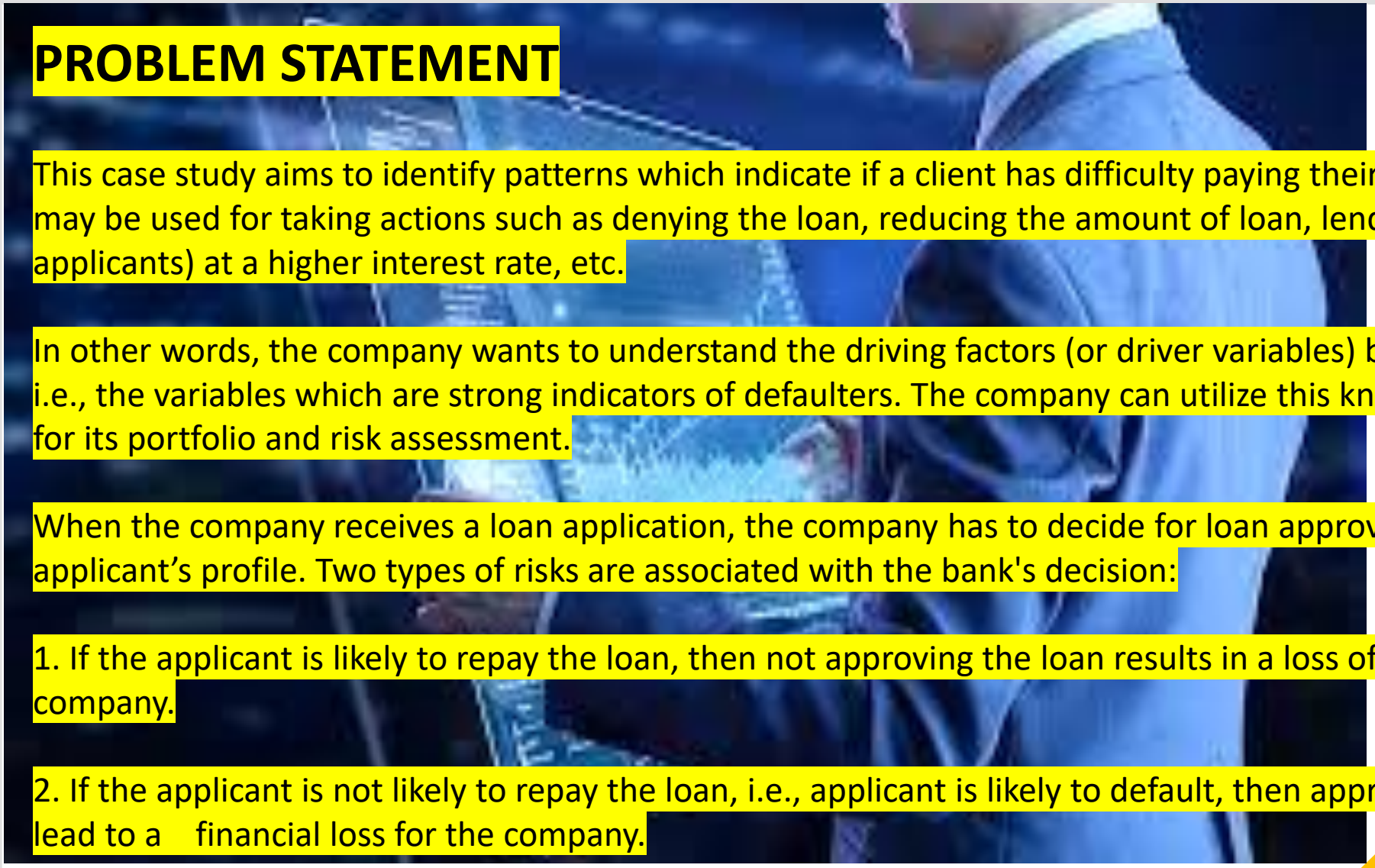
EDA CREDIT ASSIGNMENT

BANK LOAN



PROBLEM STATEMENT

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.

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 defaulters. The company can utilize this knowledge or analysis for its portfolio and risk assessment.

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:

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

2. If the applicant is not likely to repay the loan, i.e., applicant is likely to default, then approving the loan may lead to a financial loss for the company.

STEPS INVOLVED IN THE ANALYSIS

Understanding the data

Data Cleaning and Manipulation

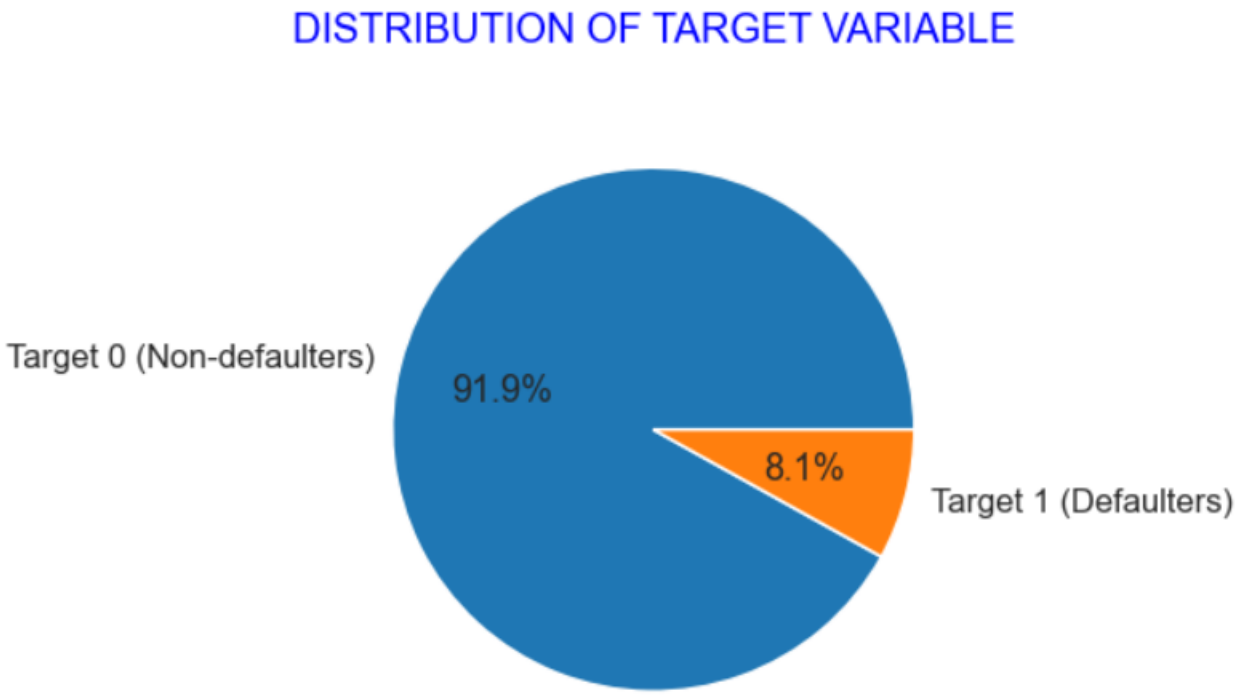
Data Imputation

Finding Outliers

Data Analysis

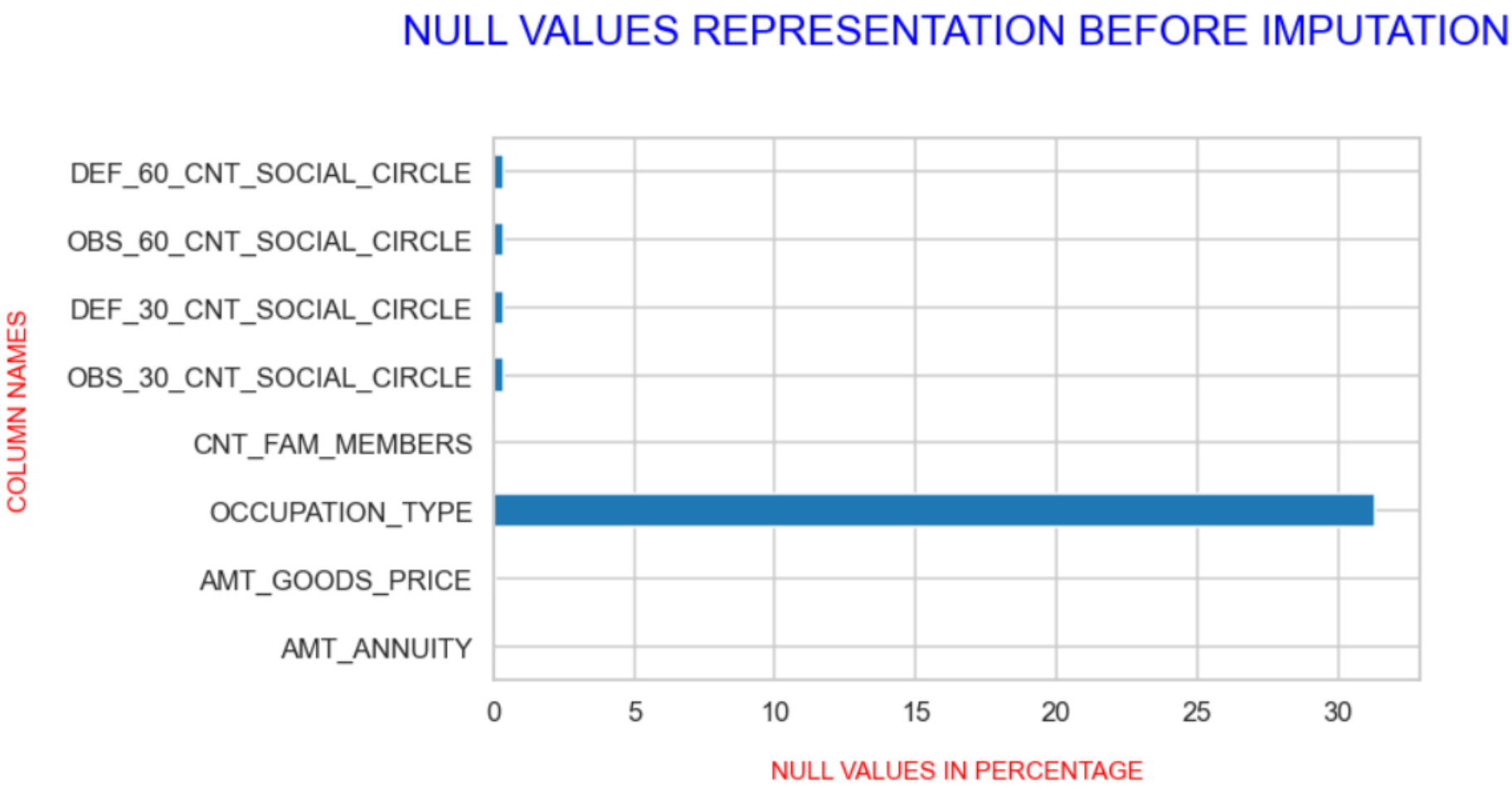
Completion of Analysis!

DATA IMBALANCE



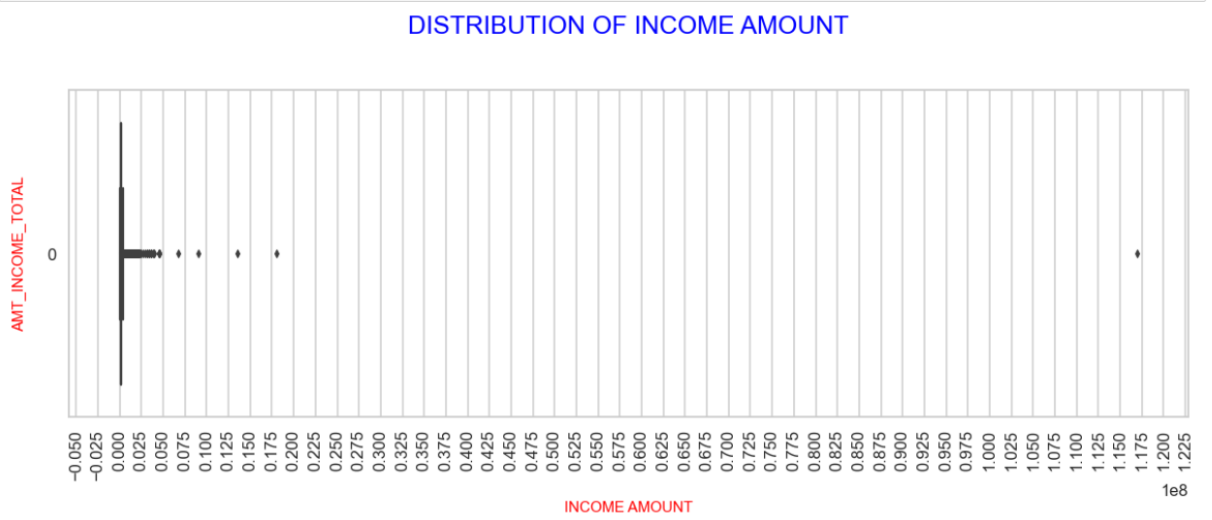
1. The ratio of imbalance between Target 0 (Non-defaulters) and Target 1 (Defaulters)
2. We can clearly see that non-defaulters are dominating the defaulters.
3. The ratio of imbalance = 91.9 / 8.1 = 11.35 : 1

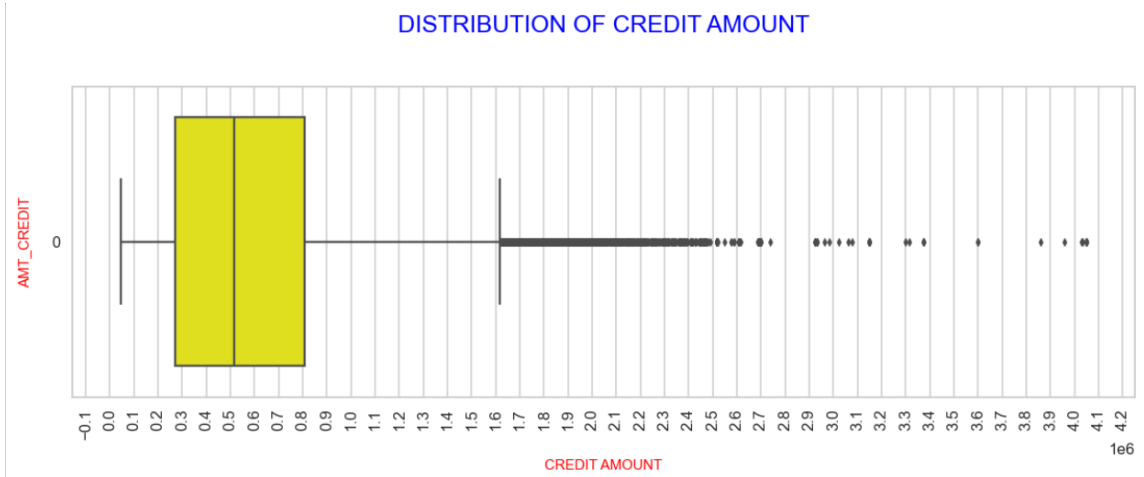
NULL VALUES REPRESENTATION BEFORE IMPUTATION IN APPLICATION DATA SET

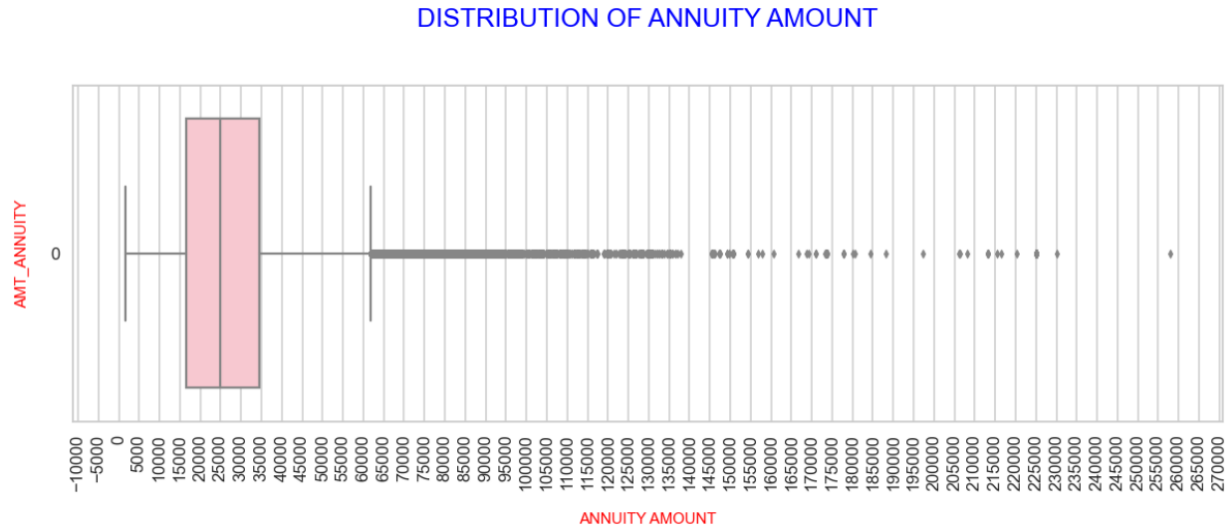


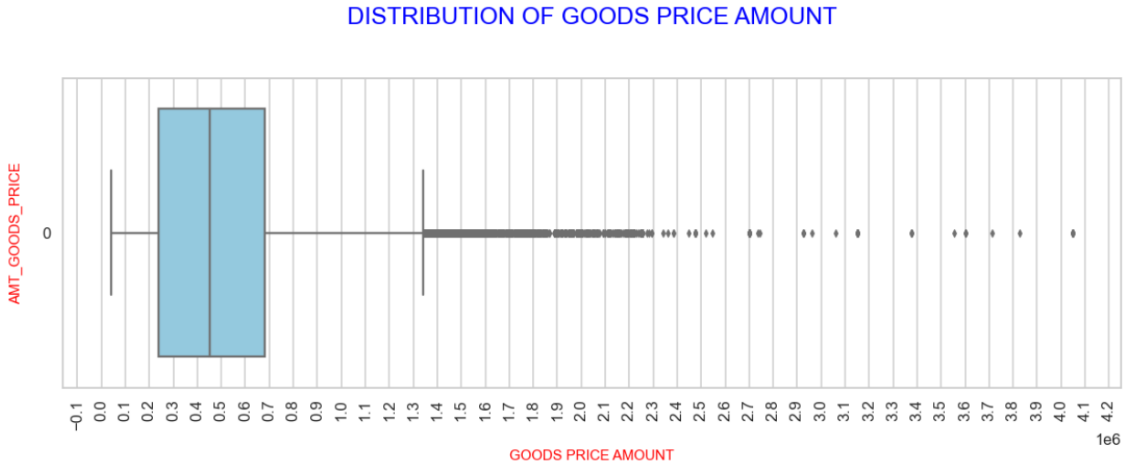
- From the below horizontal bar plot, we can observe that OCCUPATION_TYPE is having the highest percentage of null values

IDENTIFYING OUTLIERS

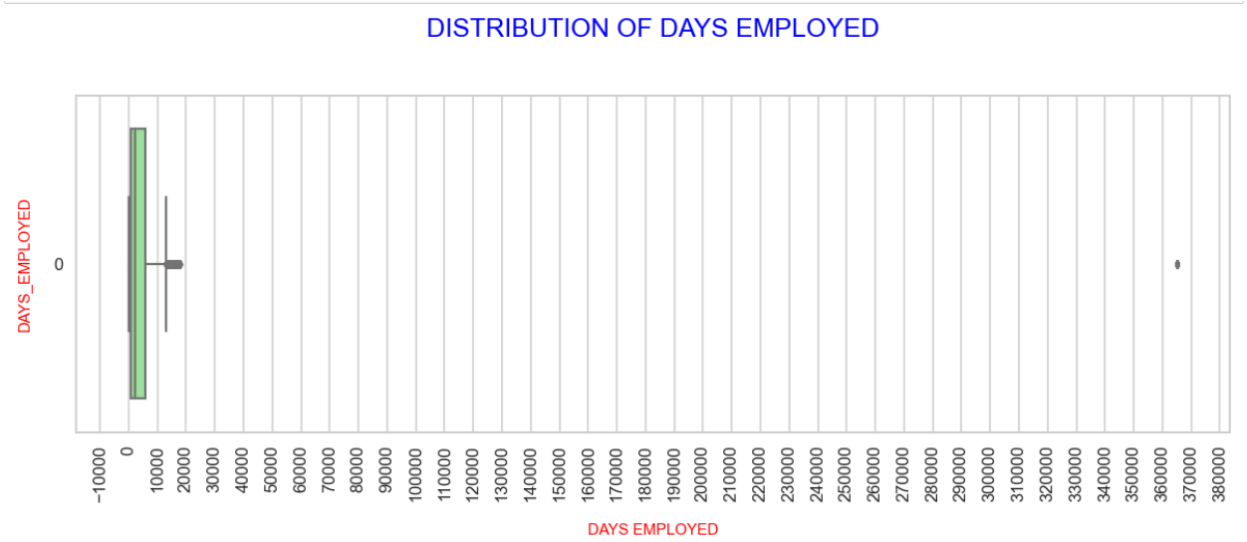


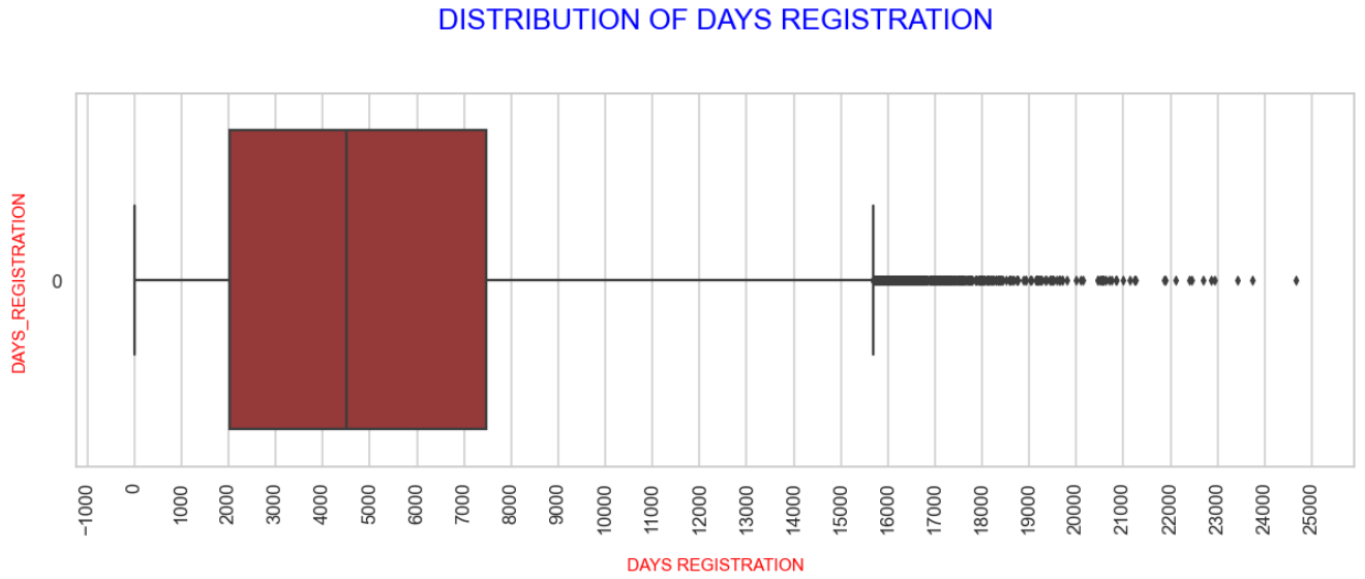






IDENTIFYING OUTLIERS





IDENTIFYING OUTLIERS

1. There are outliers in AMT_INCOME_TOTAL column. The maximum income in the column is 1.17 X (10 ^ 8) which is really a high value, but income differs from person to person.

2. There are outliers in AMT_CREDIT column. The maximum credit amount in the column is 4.05 X (10 ^ 6)

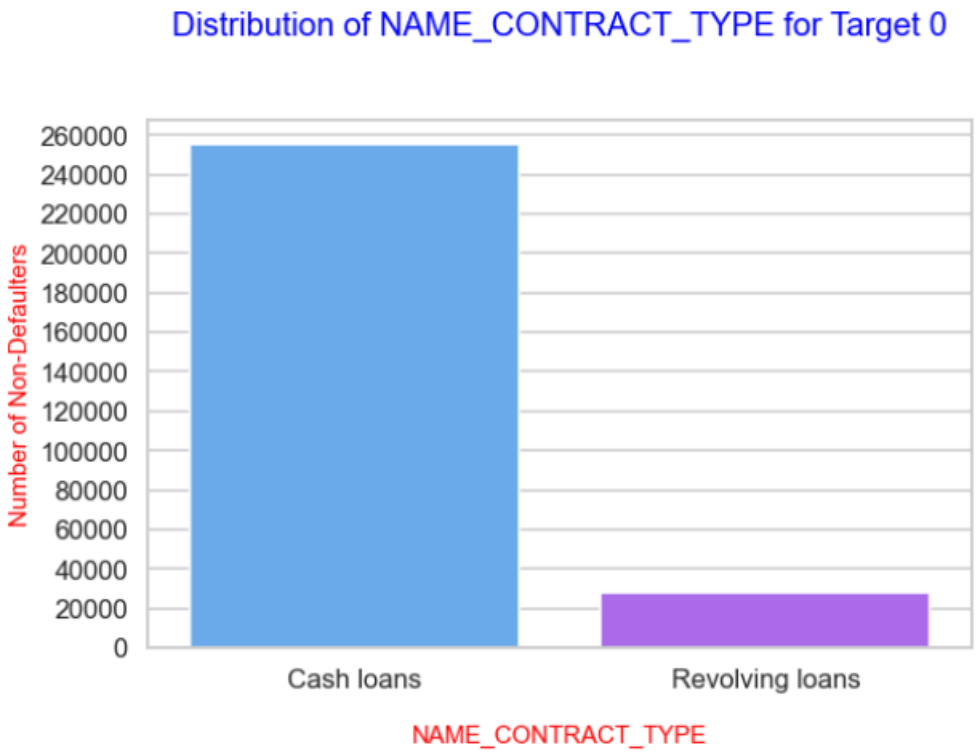
3. There are outliers in AMT_ANNUITY column. The maximum annuity amount in the column is 258025.5

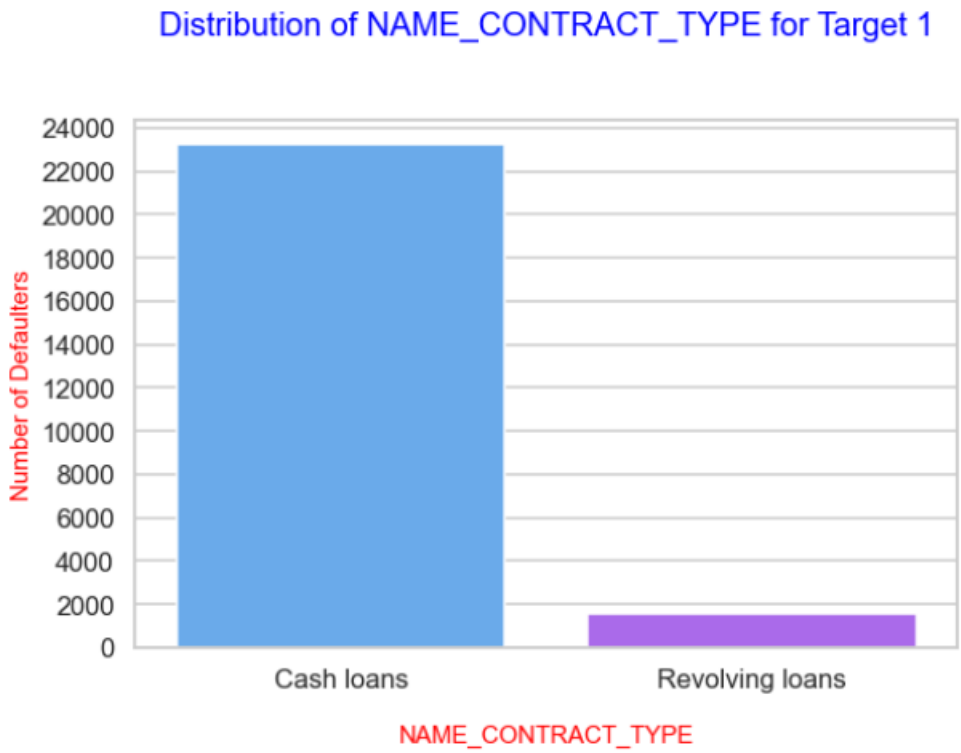
4. There are outliers in AMT_GOODS_PRICE column. The maximum goods price amount in the column is 4.05 X (10 ^ 6) but price varies from product to product

5. There is an outlier in DAYS_EMPLOYED column. The maximum days employed in the column is 365243.0

6. There are outliers in DAYS_REGISTRATION column. The maximum days registration in the column is 24672.0

Univariate analysis for NAME_CONTRACT_TYPE column

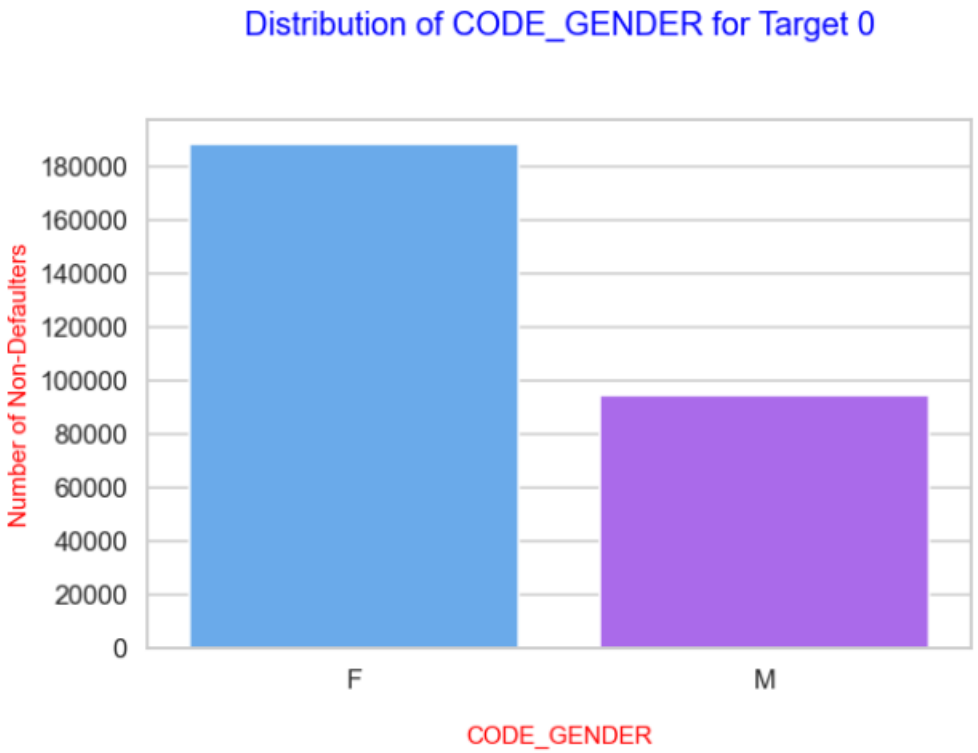


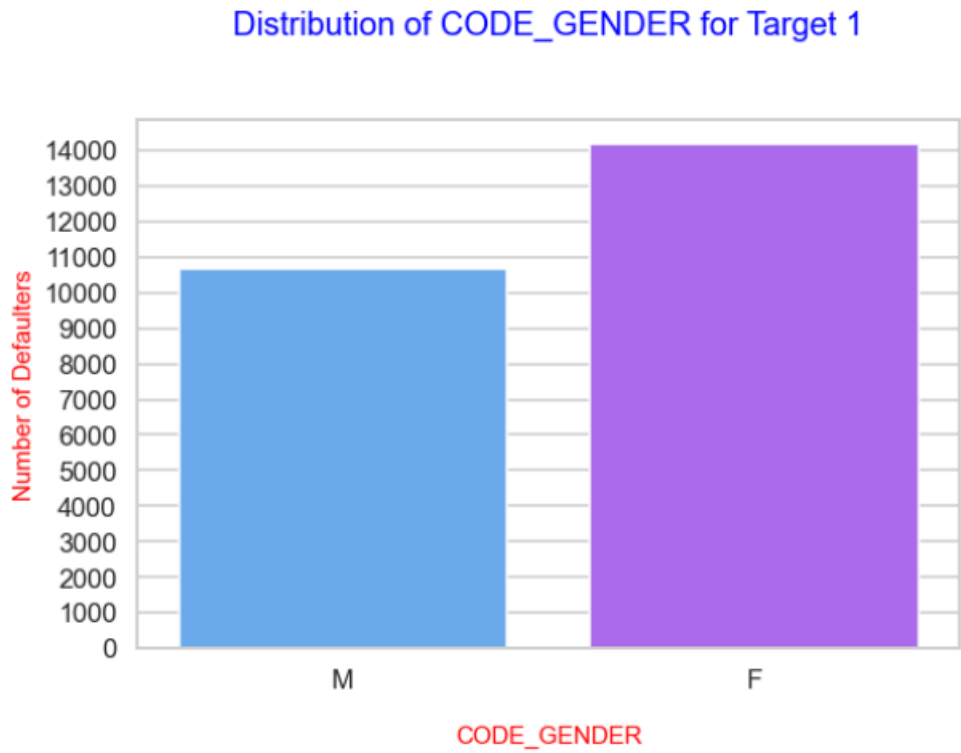


INFERENCES

1.From the above graphs, we can infer that cash loans are dominating.
2.Revolving loans have a smaller number of defaulters.
3.Cash Loans constitute of 90.5% of the loans.
4.Banks should consider giving revolving loans.

Univariate analysis for CODE_GENDER column





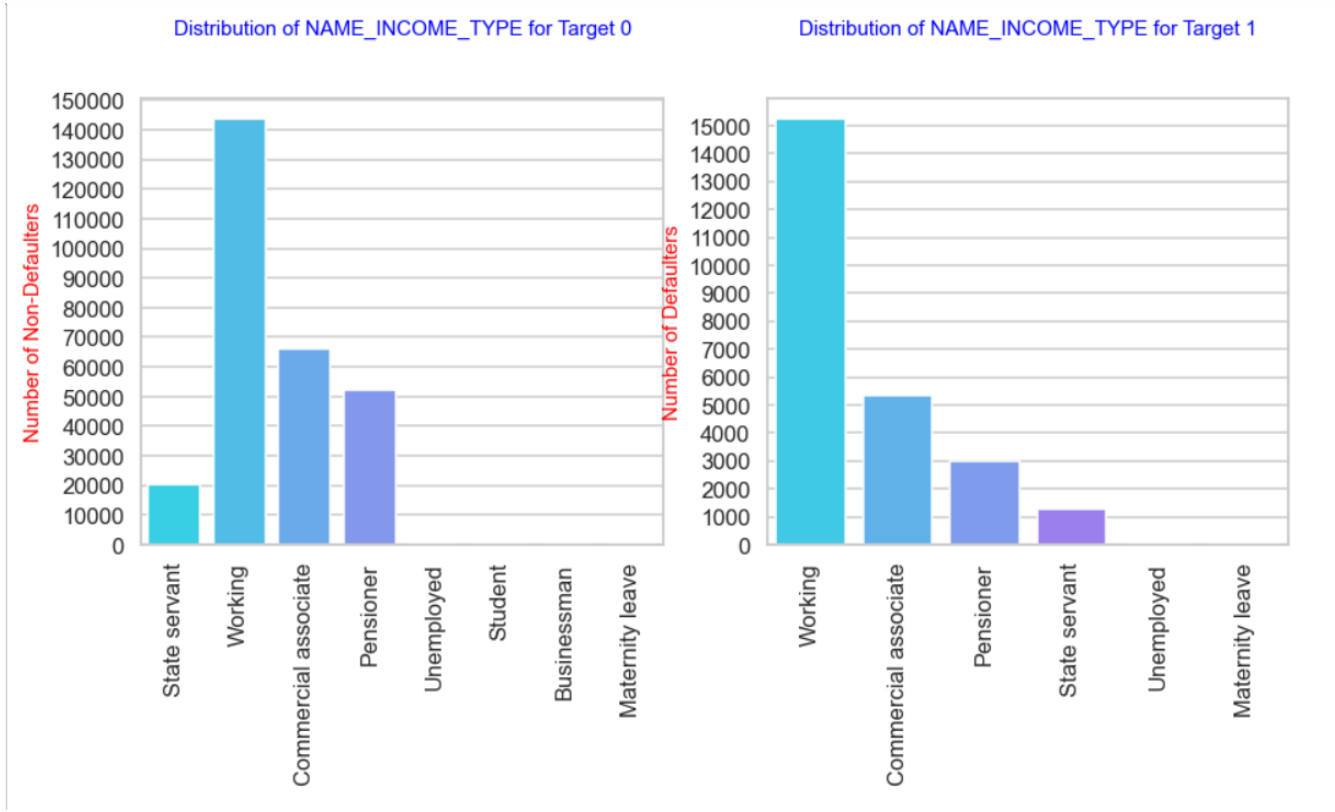
INFERENCES

1.From the above graphs, we can infer that most loan applicants are females.

2.Females mostly tend to pay the loans with respect to the total number of applications in both males and females.

3.Bank should have a higher acceptance rate for loans for females.

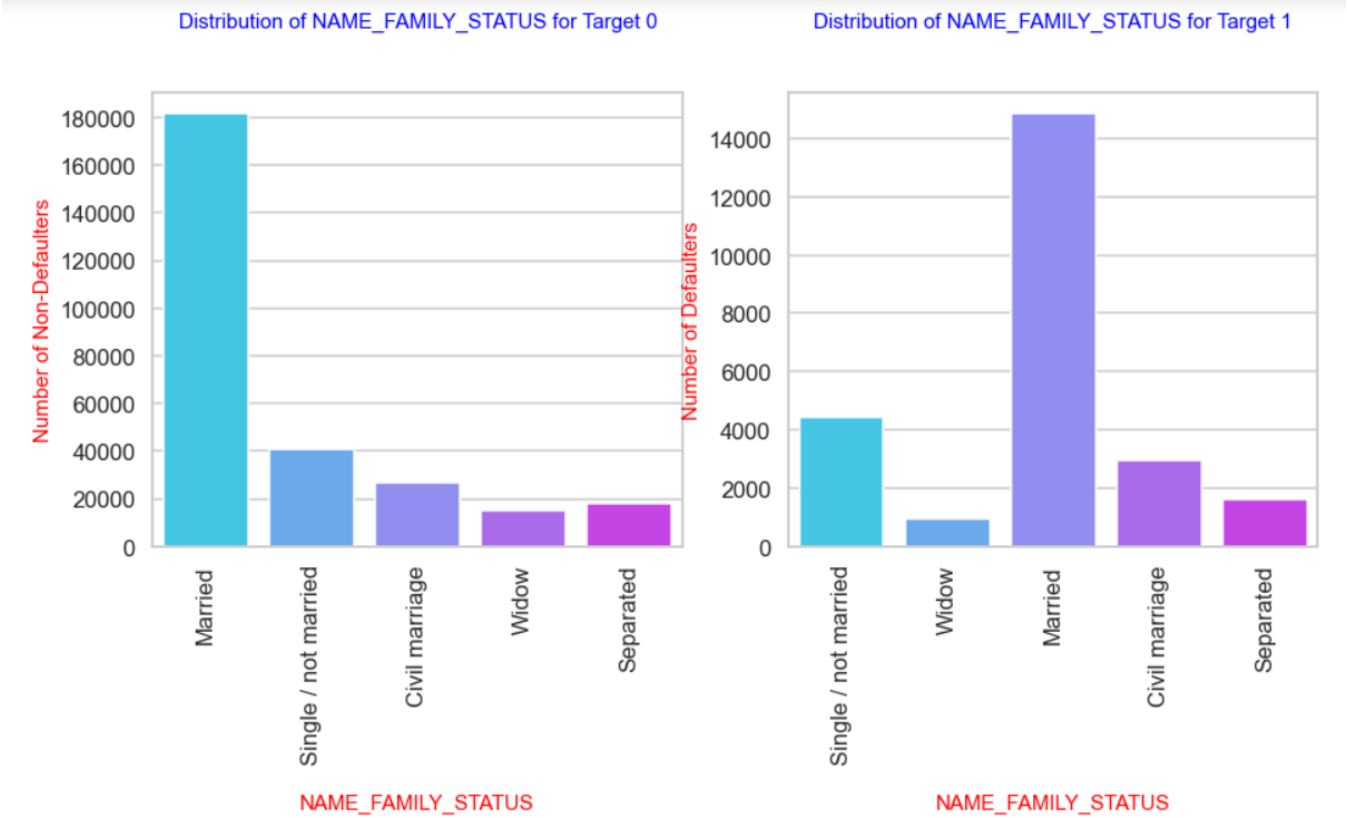
Univariate analysis for NAME_INCOME_TYPE column



INFERENCES

1.From the above graphs, we can infer that working professionals are leading the number of loan applications.
2.Maximum working professionals tend to pay the loans on time as they have regular income.
3.The second highest defaulters are Commercial associates.
4.Banks should consider giving more loans to working professionals.

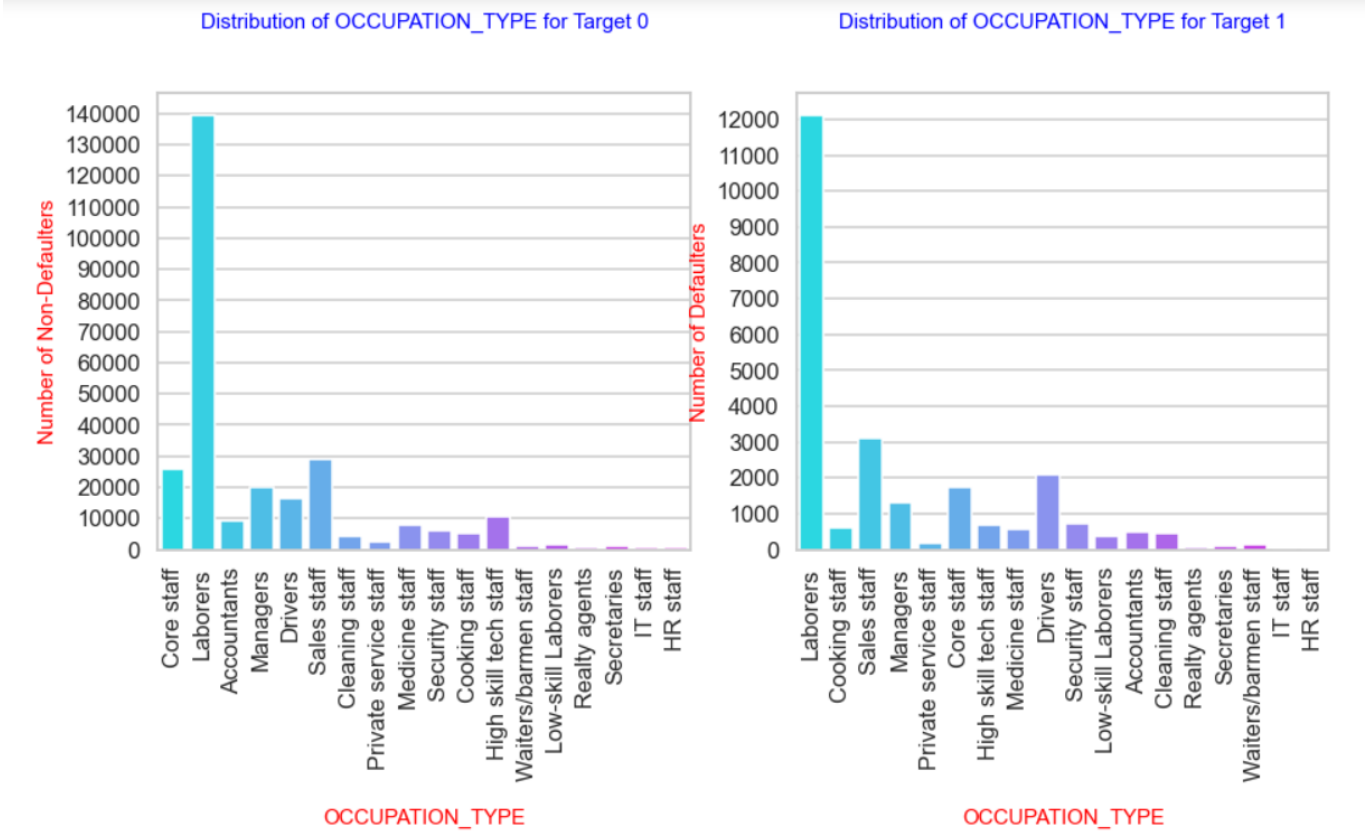
Univariate analysis for NAME_FAMILY_STATUS column



INFERENCES

1.From the above graphs, we can infer that married people are leading the number of loan applications.
2.Married people are the highest in paying the loans on time.
3.Banks should consider giving more loans to married people.

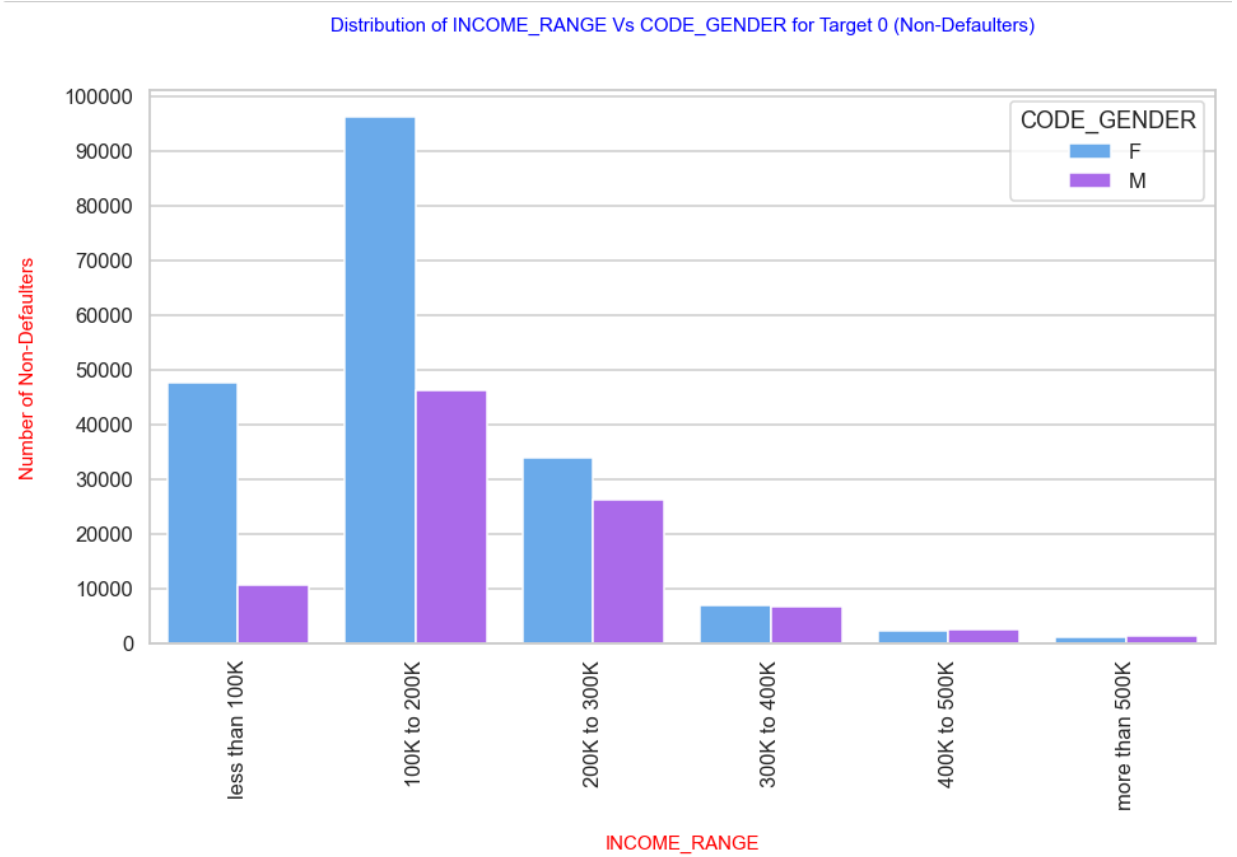
Univariate analysis for OCCUPATION_TYPE column



INFERENCES

1.From the above graphs, we can infer that laborers are leading the number of loan applications.
2.Most of the laborers tend to pay the loans on time.
3.The second highest defaulters are Sales staff.
4.Banks should consider giving more loans to laborers.

Numerical - Categorical Univariate Analysis

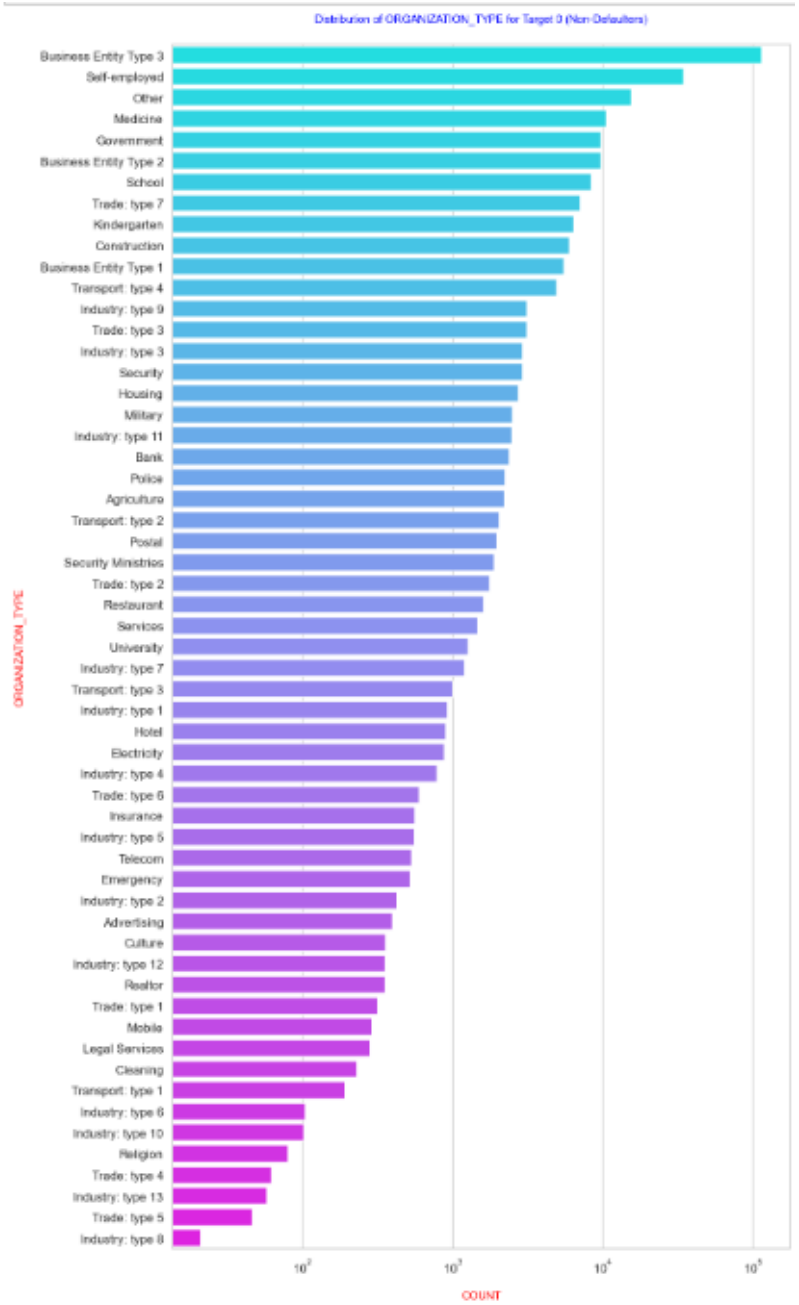


INFERENCES

1.From the above graph, we can infer that income in the range 100K to 200K is the highest in which particularly females tend to pay the loans on time.

2.The proportions of females are bigger than males mostly which means that females are less defaulters than males.

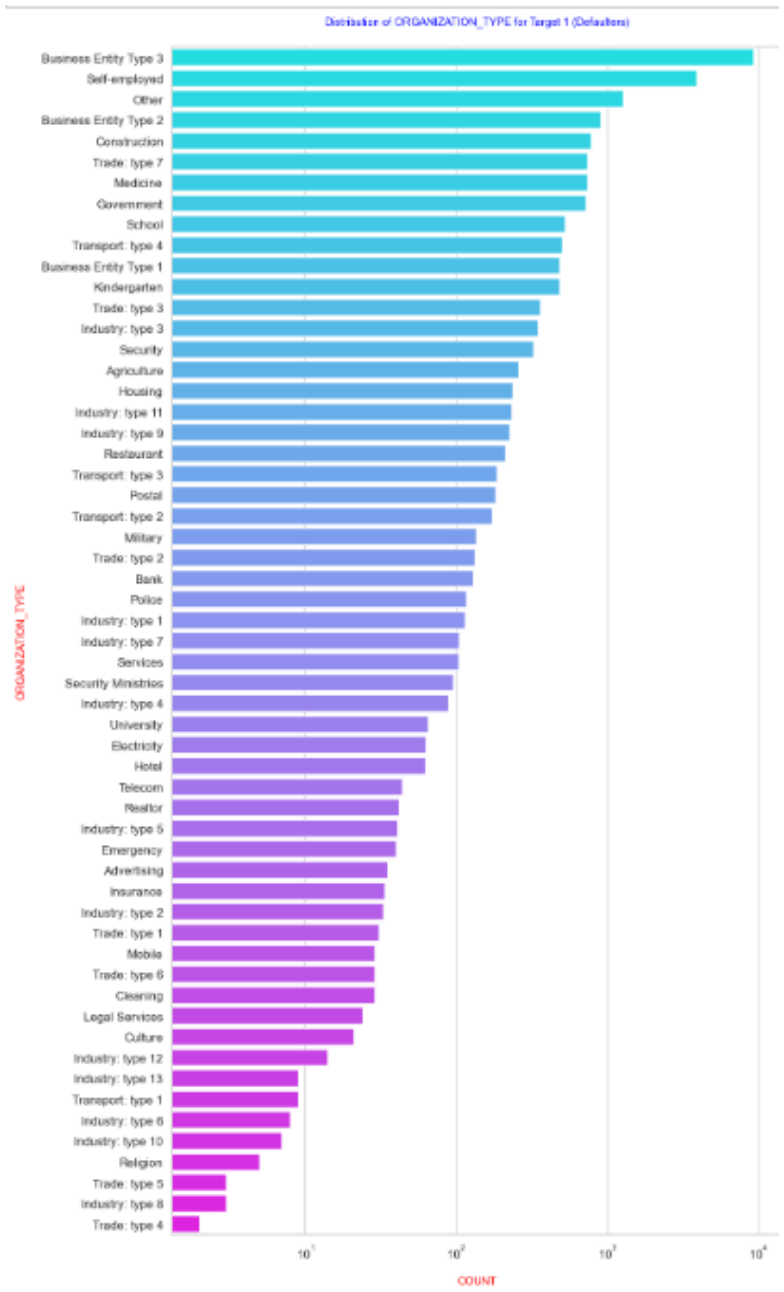
Distribution of ORGANIZATION_TYPE for Target 0 (Non-Defaulters)



INFERENCES

1.Applicants who applied for the loan are mostly from Business Entity Type 3, Self-employed and Other organization types.

2.Applicants are less from Industry: type 8, Trade: type 5, Industry: type 13 organizations types.



Distribution of ORGANIZATION_TYPE for Target 1 (Defaulters)

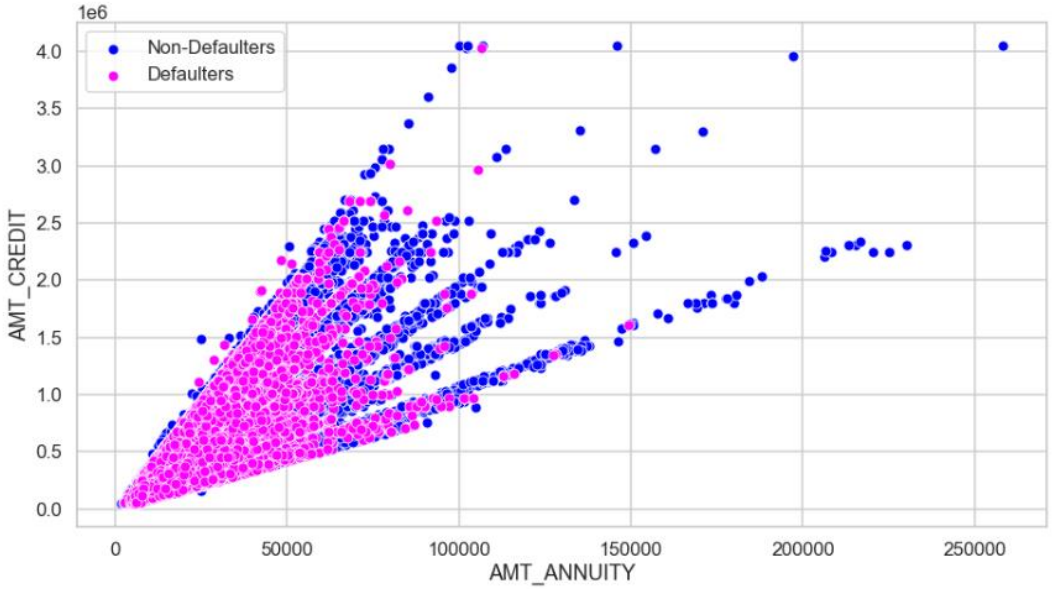
INFERENCES

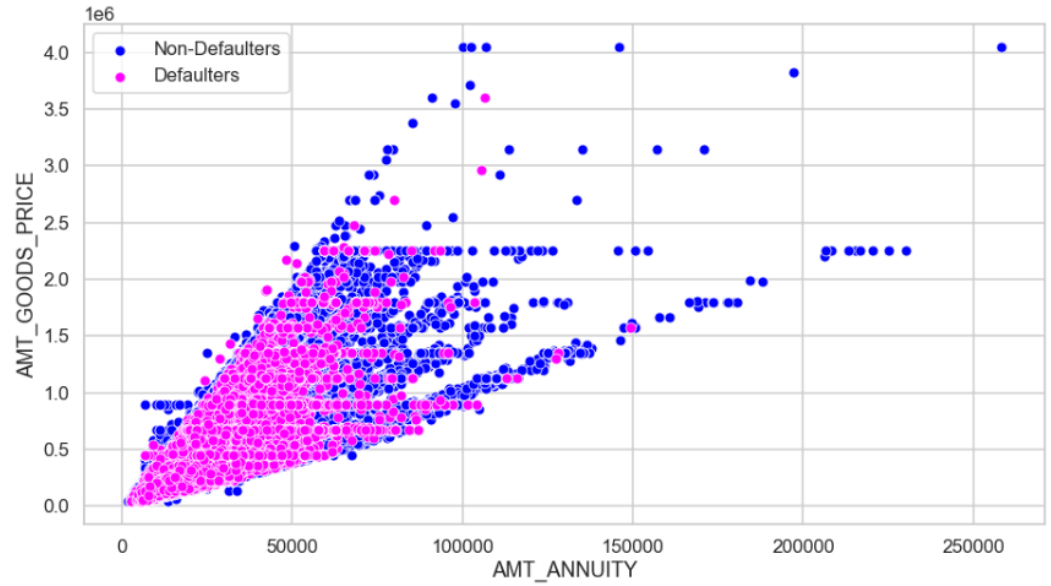
1.Applicants who applied for the loan are mostly from Business Entity Type 3, Self-employed and Other organization types.

2.Applicants are less from Industry: type 8, Trade: type 5, Industry: type 13 organizations types.

BIVARIATE ANALYSIS

FIG 1





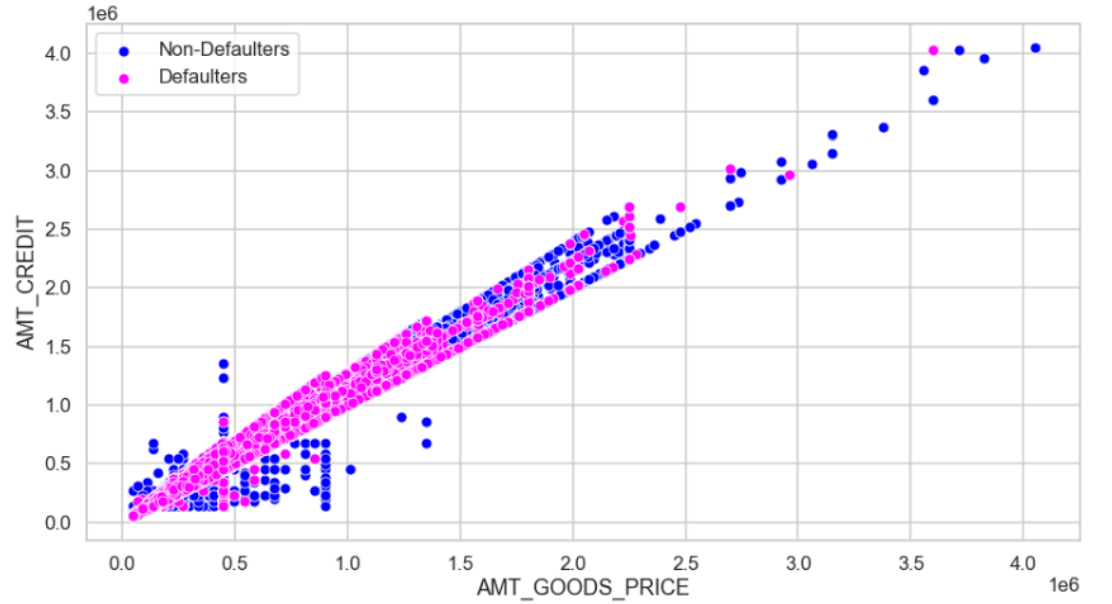


FIG 2

FIG 3

BIVARIATE ANALYSIS INFERENCES

INFERENCES FIG 1

1.From the above scatterplot, we can infer that the two variables (AMT_ANNUITY, AMT_CREDIT) are fairly correlated.
2.Most of the defaulters are having AMT_ANNUITY values less than 80000, but after 80000, there's a decrease in the defaulters.

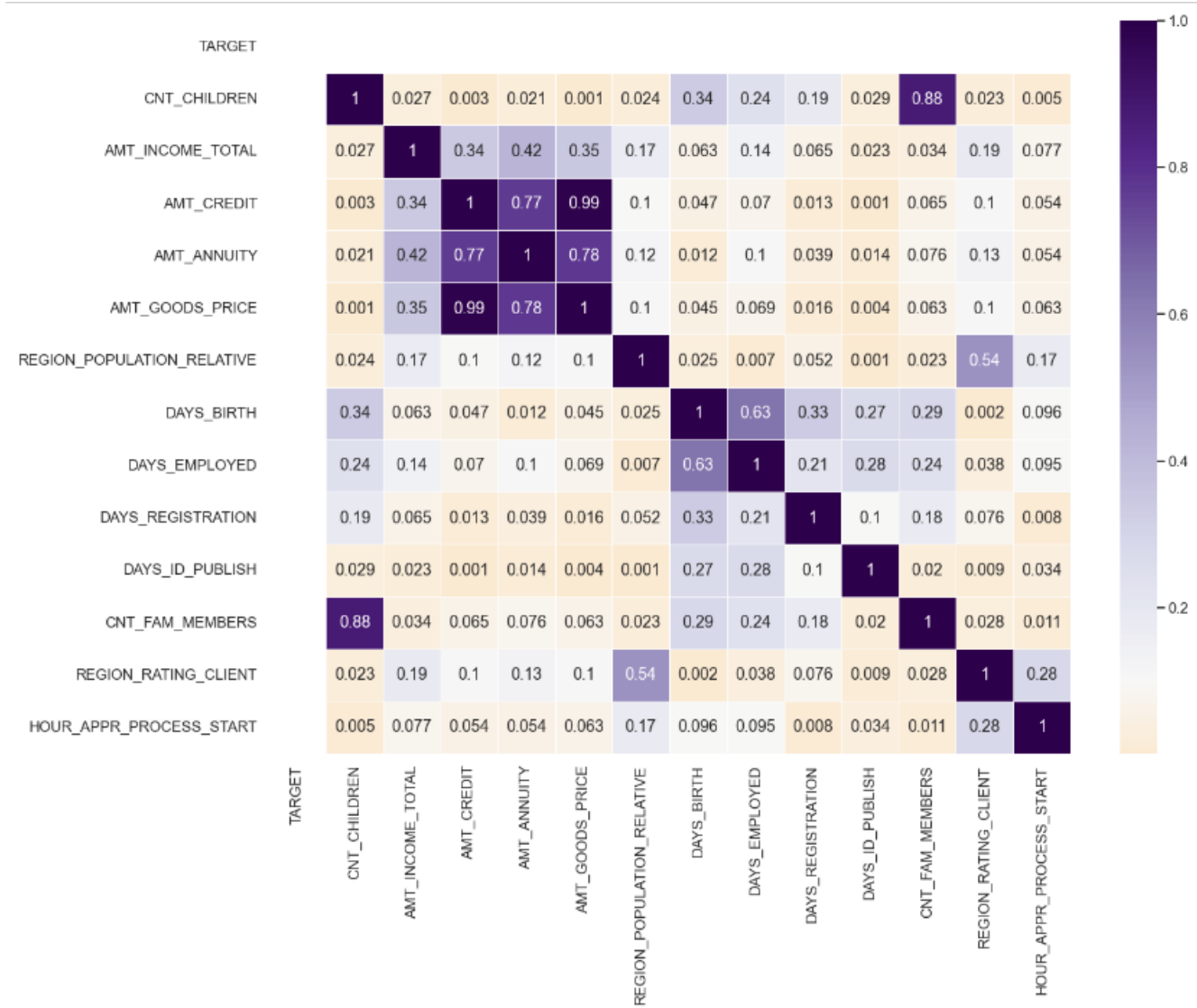
INFERENCES FIG 2

1.From the above scatterplot, we can infer that the two variables (AMT_GOODS_PRICE, AMT_CREDIT) are strongly correlated as there's a liner increase in the values of both the variables.

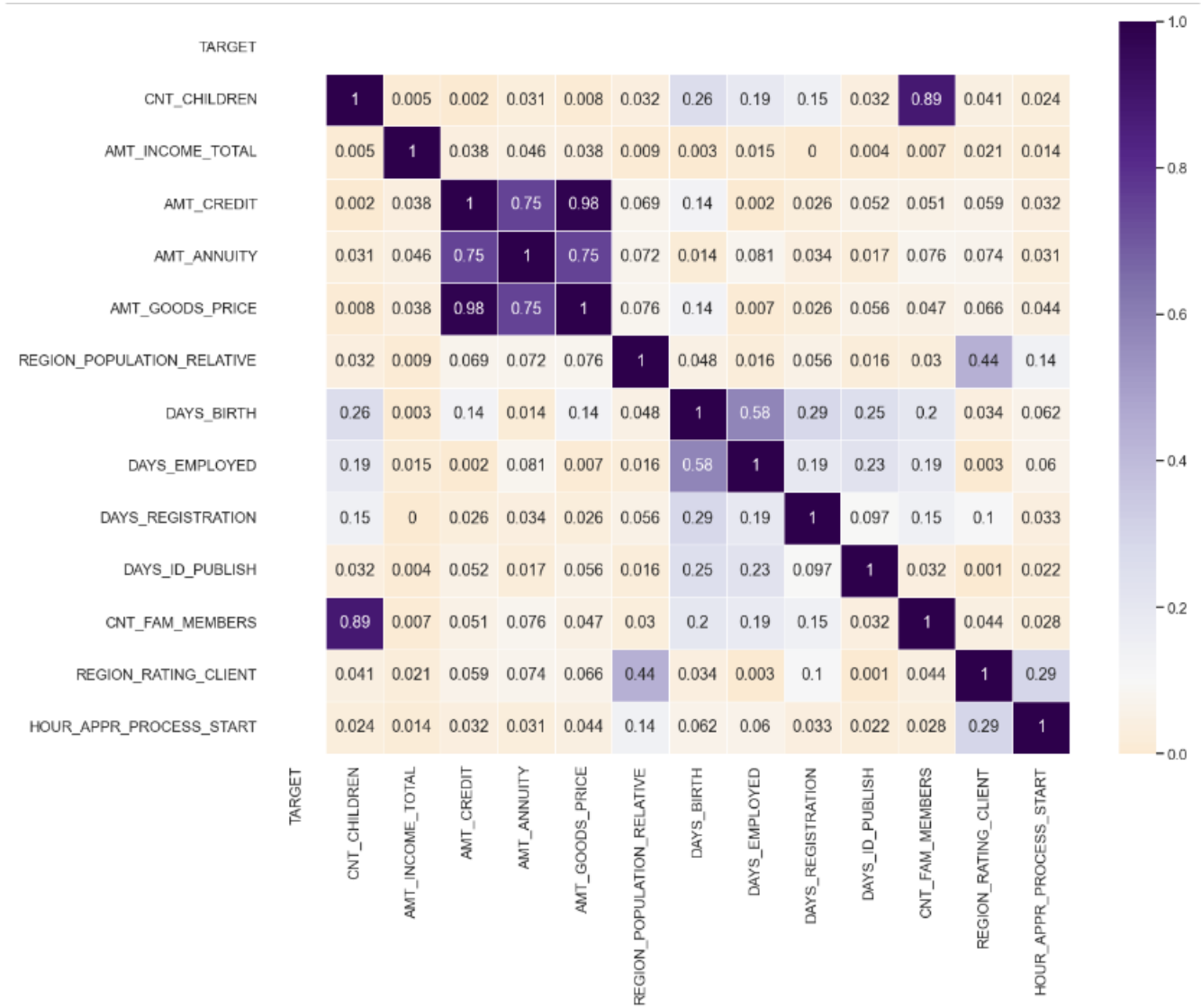
INFERENCES FIG 3

1.From the above scatterplot, we can infer that the two variables (AMT_ANNUITY, AMT_GOODS_PRICE) are moderately correlated.
2.Most defaulters have AMT_ANNUITY values less than 100000. Beyond 100000, there is a noticeable decrease in the defaulters.

CORRELATION TARGET 0



CORRELATION TARGET 1



CORRELATION TARGET 0

	VAR1	VAR2	Correlation
0	AMT_CREDIT	AMT_GOODS_PRICE	0.987
1	CNT_CHILDREN	CNT_FAM_MEMBERS	0.879
2	AMT_ANNUITY	AMT_GOODS_PRICE	0.776
3	AMT_CREDIT	AMT_ANNUITY	0.771
4	DAYS_BIRTH	DAYS_EMPLOYED	0.626
5	REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	0.539
6	AMT_INCOME_TOTAL	AMT_ANNUITY	0.419
7	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.349
8	AMT_CREDIT	AMT_INCOME_TOTAL	0.343
9	CNT_CHILDREN	DAYS_BIRTH	0.337

CORRELATION TARGET 1

	VAR1	VAR2	Correlation
0	AMT_GOODS_PRICE	AMT_CREDIT	0.983
1	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885
2	AMT_CREDIT	AMT_ANNUITY	0.752
3	DAYS_BIRTH	DAYS_EMPLOYED	0.582
4	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.443
5	REGION_RATING_CLIENT	HOUR_APPR_PROCESS_START	0.294
6	DAYS_REGISTRATION	DAYS_BIRTH	0.289
7	CNT_CHILDREN	DAYS_BIRTH	0.259
8	DAYS_BIRTH	DAYS_ID_PUBLISH	0.253
9	DAYS_EMPLOYED	DAYS_ID_PUBLISH	0.229

INFERENCES FROM THE ABOVE CORRELATION

1.Target variable is not present in the correlation as it is a categorical variable and not a continuous variable.

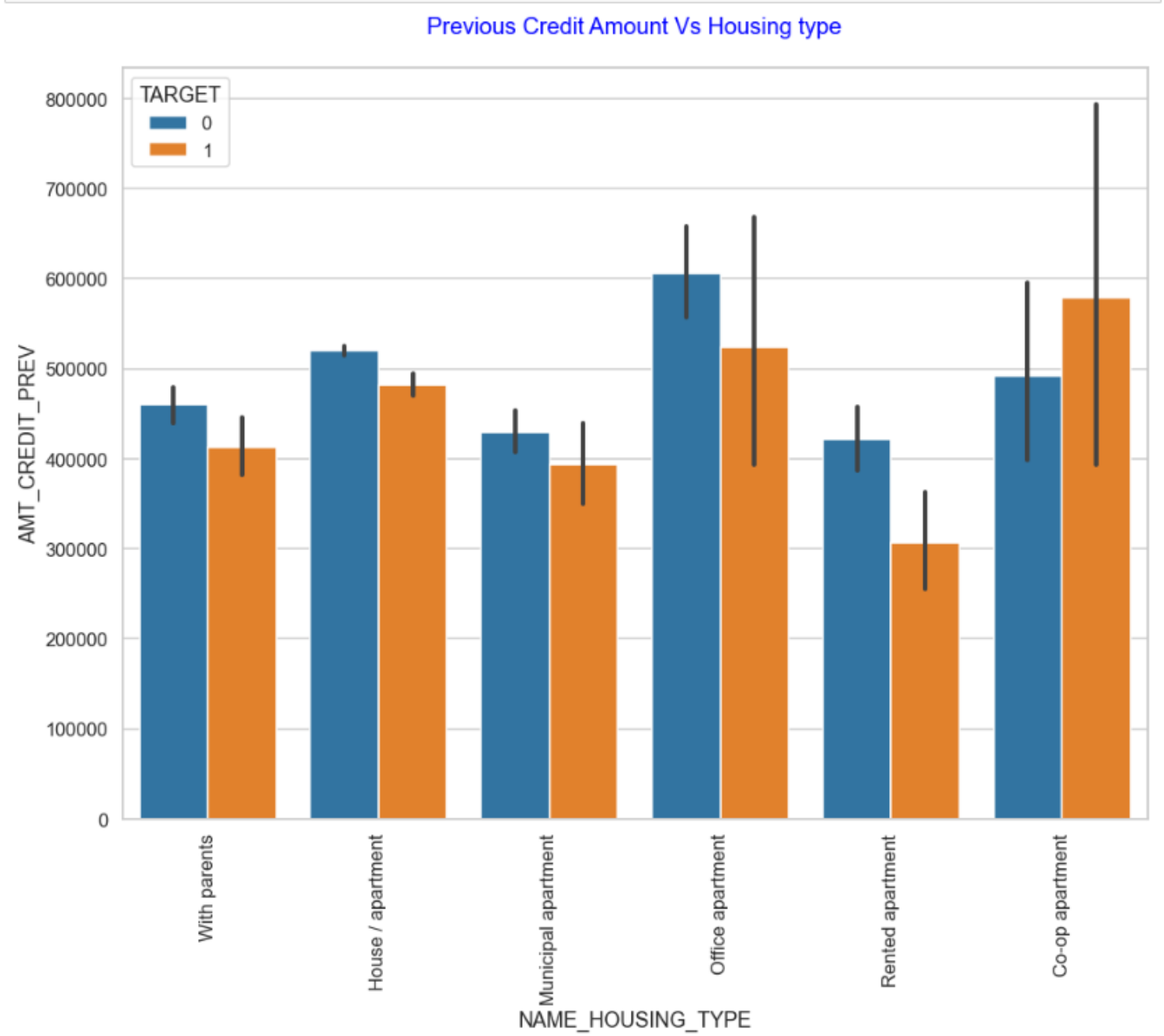
2.AMT_GOODS_PRICE and AMT_CREDIT are highly correlated with a value of 0.98.

3.The correlation between AMT_CREDIT and AMT_ANNUITY is slightly reduced for the defaulters.

4.The correlation is strong between CNT_FAM_MEMBERS and CNT_CHILDREN even though the correlation increases for the defaulters.

5.The correlation of non-defaulters is high for DAYS_EMPLOYED and DAYS_BIRTH (0.626) when compared to the correlation of defaulters (0.582)

MERGED DATA SET



CONCLUSION

1.Banks should focus more on contract type 'Student', 'Pensioner' and 'Businessman' with housing type rather than 'Co-op apartment' for successful payments as 'Co-op apartment' has difficulties in paying the loan.

2.Banks should focus less on income type 'Working' as they are having the most number of unsuccessful payments.

3.Also, loan purpose 'Repair' is having highest number of unsuccessful payments on time.

4.Banks should focus on as much clients from housing type 'With parents' as possible as they are having least number of unsuccessful payments.


Thank you
Ghadiyaram Ramakrishna Vivek