EXPLORATORY DATA ANALYSIS CREDIT ASSIGNMENT

UPGRAD & IIITB, DATA SCIENCE DSC65 – FEB 2024

BY G B SHRUTHI

PROBLEM STATEMENT

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- The client with payment difficulties: He/She had late payment more than X days on at least one of the first Y instalments of the loan in our sample.(Target = 1)
- All other cases: All other cases when the payment is paid on time. (Target = 0)

STEPS OF EDA:

- Application Data Analysis
- Previous Application Data Analysis
- Conclusion

APPLICATION DATA ANALYSIS

- Import necessary modules
- Reading and Analyzing Data
- Checking the Null Percentage of values
- Imputing the values
- Standardization of Data
- Identifying the Outliers
- Categorical and Numerical Analysis
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

APPLICATION DATA ANALYSIS STEPS

- At first, we need to analyze the data of the given excel like checking column headers etc.
- Check the shape of the data frame. (307511, 122)
- Check the no of null value rows present in each and every column and also its data types using info() function.
- Check the statistical information of data frame using describe() function.
- First we need to check the percentage of null values of columns and drop the columns with 50% null values as their presence does impact the statistics.
- Remaining columns who have less percentage of null values , we will impute the columns with Mean()/Median() – Numerical columns and Mode() – Categorical columns.

APPLICATION DATA ANALYSIS STEPS

- Standardize the values of columns like for Time columns if all the values should be converted to either Seconds or Minutes or Hours and if any columns have negative values convert it into positive values.
- Irrelevant columns are dropped because they are not meant for analysis if required.
- Identify the Outliers using boxplot, if any outliers calculate the IQR for them, calculate the upper bounds and lower bounds.

IQR=Q3-Q1 ;

lower_bound=Q1-1.5*IQR ;

upper_bound=Q3+1.5*IQR

app[col]=np.where(app[col]>upper_bound,upper_bound,app[col])

app[col]=np.where(app[col]<lower_bound,lower_bound,app[col])

APPLICATION DATA ANALYSIS STEPS

- Segregate all the columns of data frame based on this data type into Categorical and Numerical Variables for Data Visualization using MatplotLib and seaborn libraries.
- Now Univariate Analysis is done on the Categorical variables using BAR Plot and it is also done on the Numerical variables using HIST Plot and their insights are taken accordingly.
- Bivariate Analysis is done using SCATTER Plot and BAR plots and their insights are taken.
- Multivariate Analysis is done between Continuous Numerical Variables using Heat Maps.

UNIVARIATE ANALYSIS

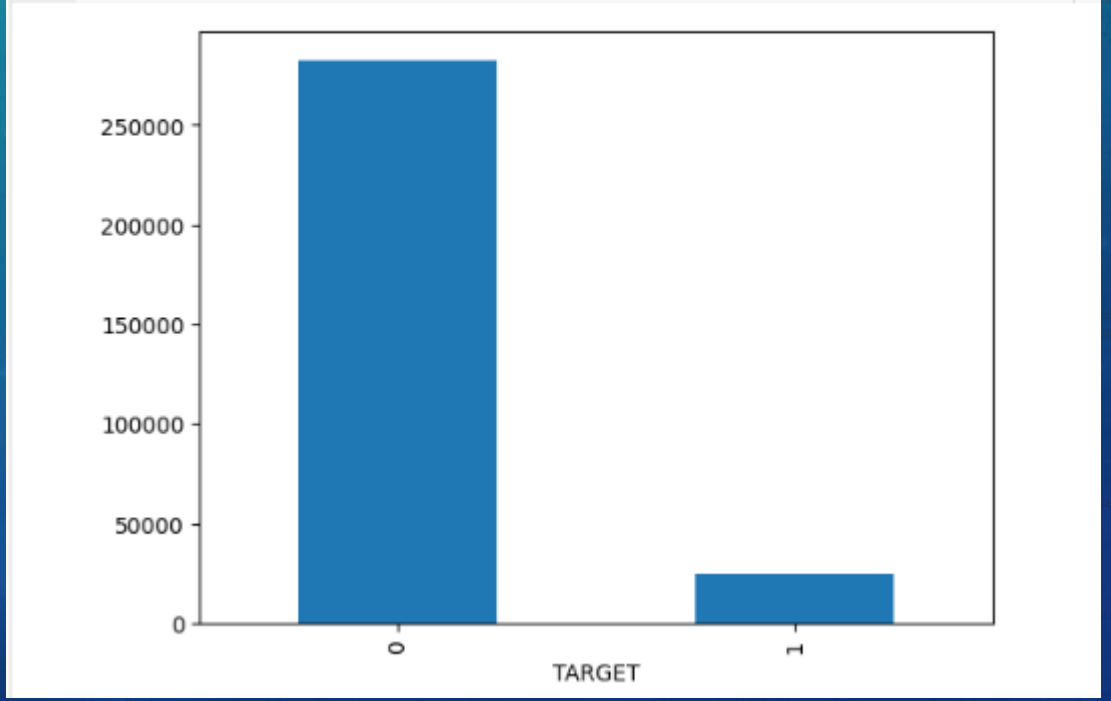
Check the TARGET Variable of Data Frame which the depicts the percentage of Loan Defaulters in Data.

```
#Lets check the percentage of Loan Defaulters from the Data
app['TARGET'].value_counts(normalize=True)*100
```

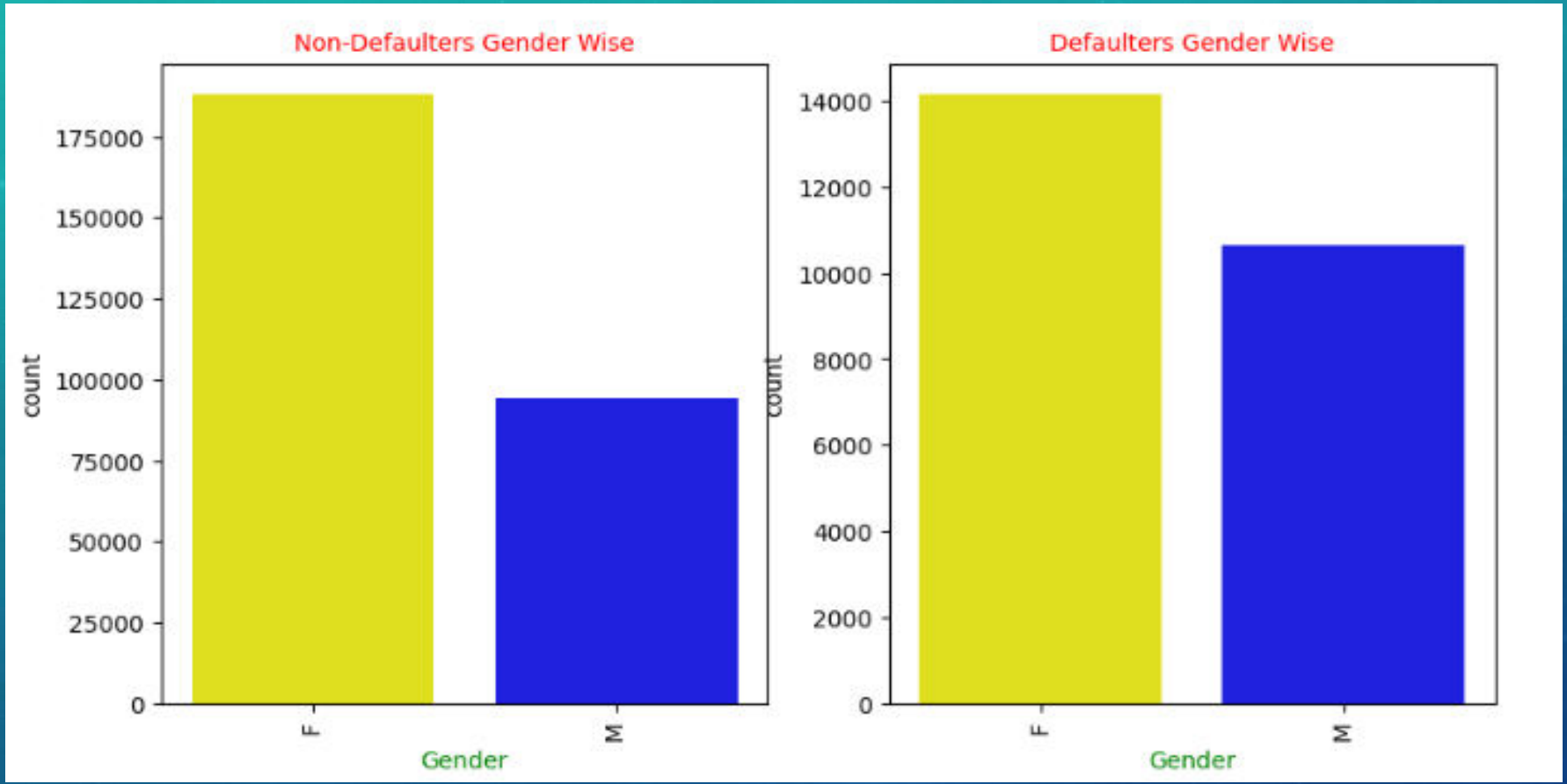
```
TARGET
0    91.926649
1     8.073351
Name: proportion, dtype: float64
```

We can see that out of 100 only almost 8% of the people are Loan Defaulters from our given Data in Application File.

```
#We can see that out of 100 only almost 8% of the people are Loan Defaulters
app['TARGET'].value_counts().plot.bar()
plt.show()
```



CODE_GENDER

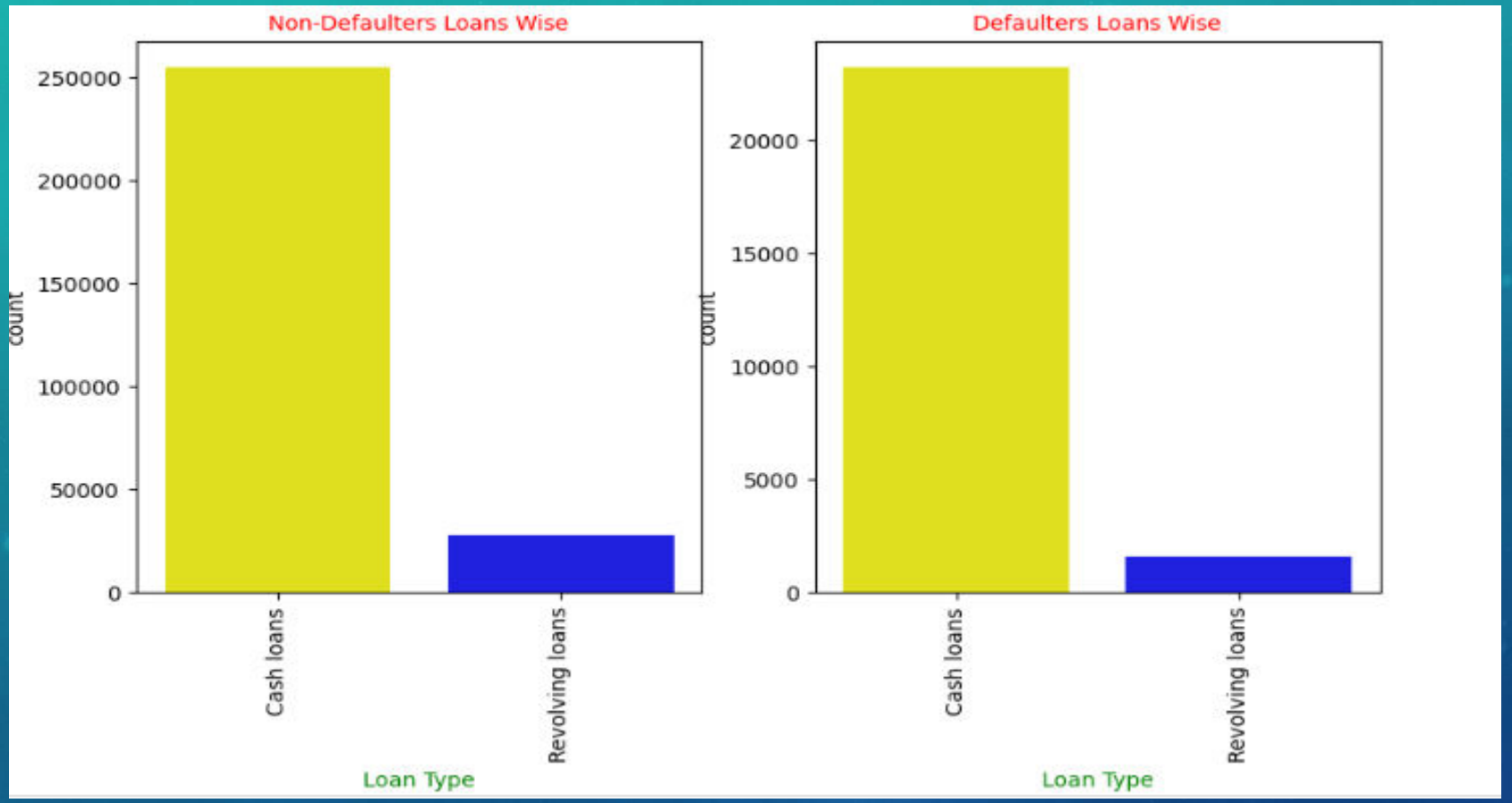


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Females are in higher number than males.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also the Females are in higher number than males.

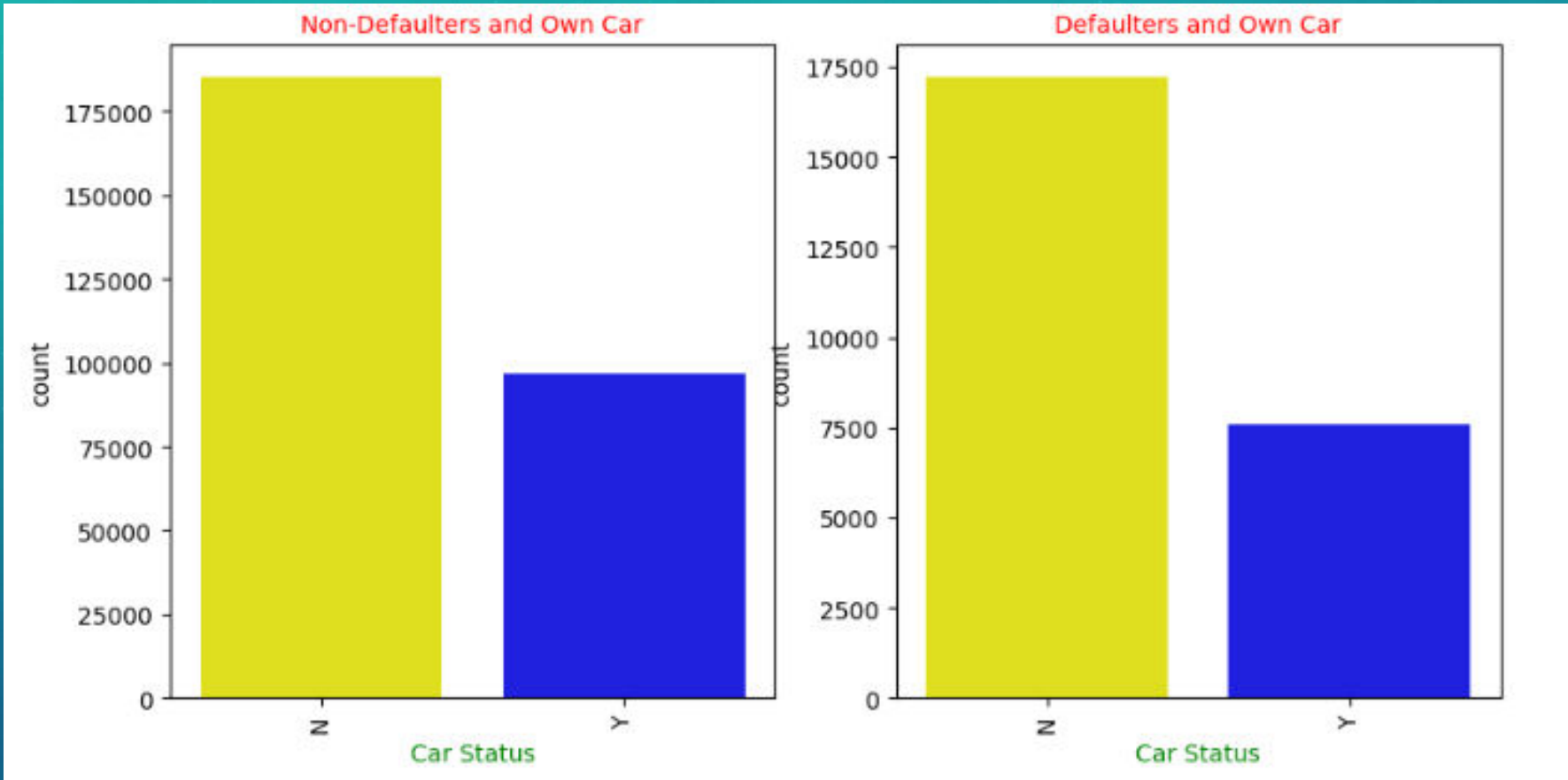
NAME_CONTRACT_TYPE



INSIGHTS:

1. Firstly, Among Non-Defaulters, i.e., TARGET == 0, Cash loans are in higher number than Revolving loans.
2. Next, Among Defaulters, i.e., TARGET == 1, Here also the Cash loans are in higher number than Revolving loans.

FLAG_OWN_CAR

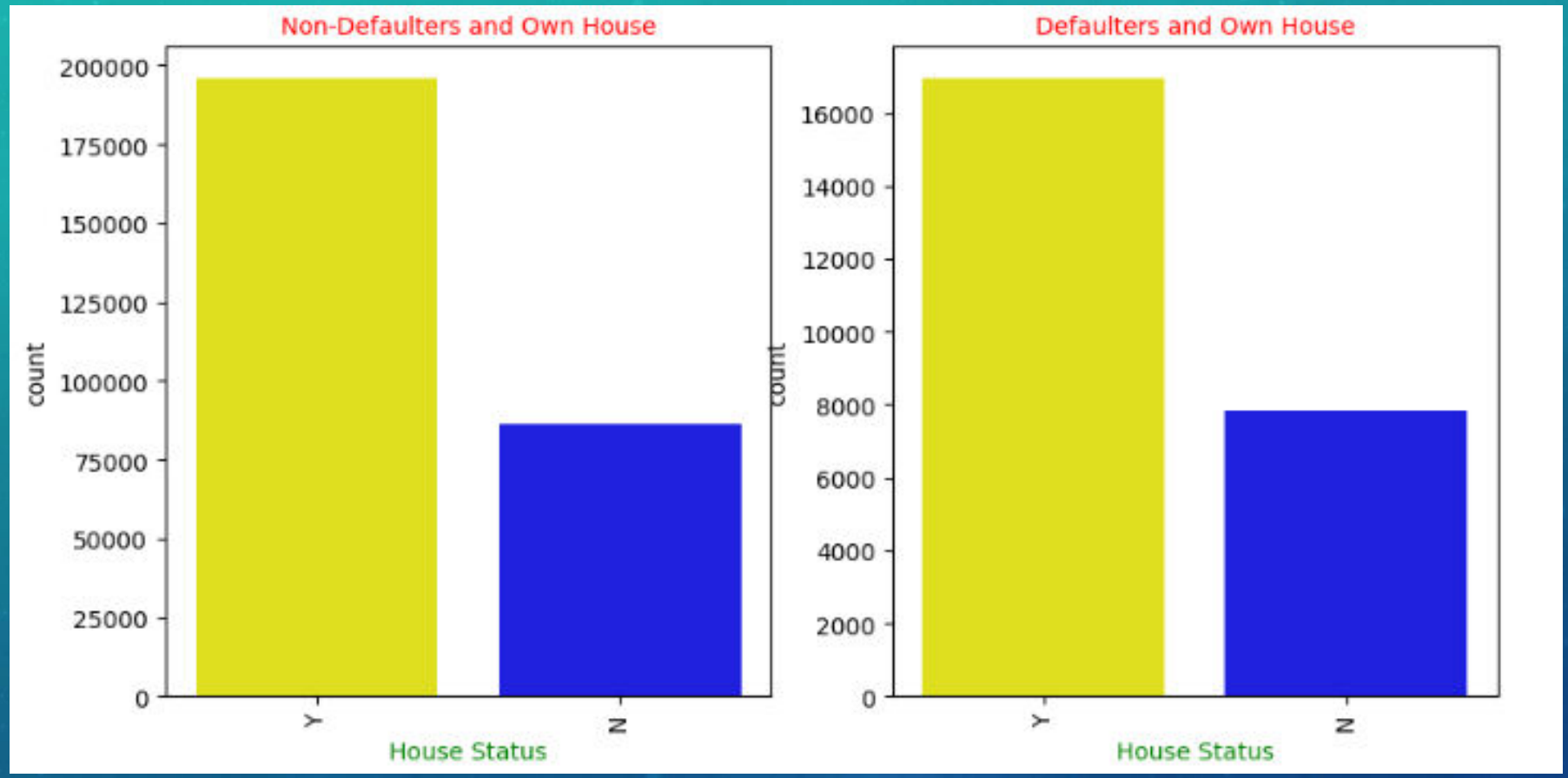


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Maximum of them are not having a own car with them.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also Maximum of them are not having a own car with them.

FLAG_OWN_REALTY

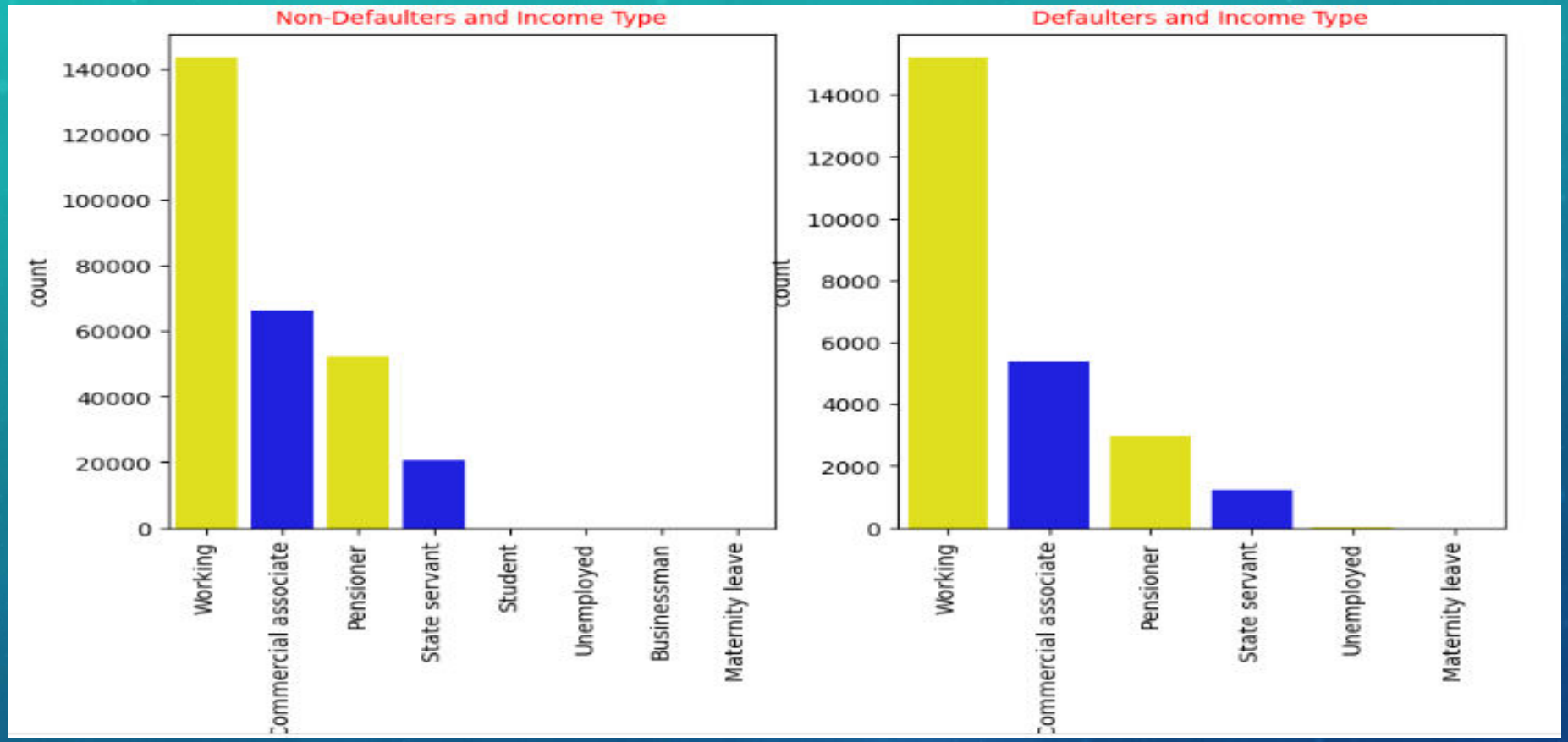


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Maximum of them are having a own house with them.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also Maximum of them are having a own house with them.

NAME_INCOME_TYPE

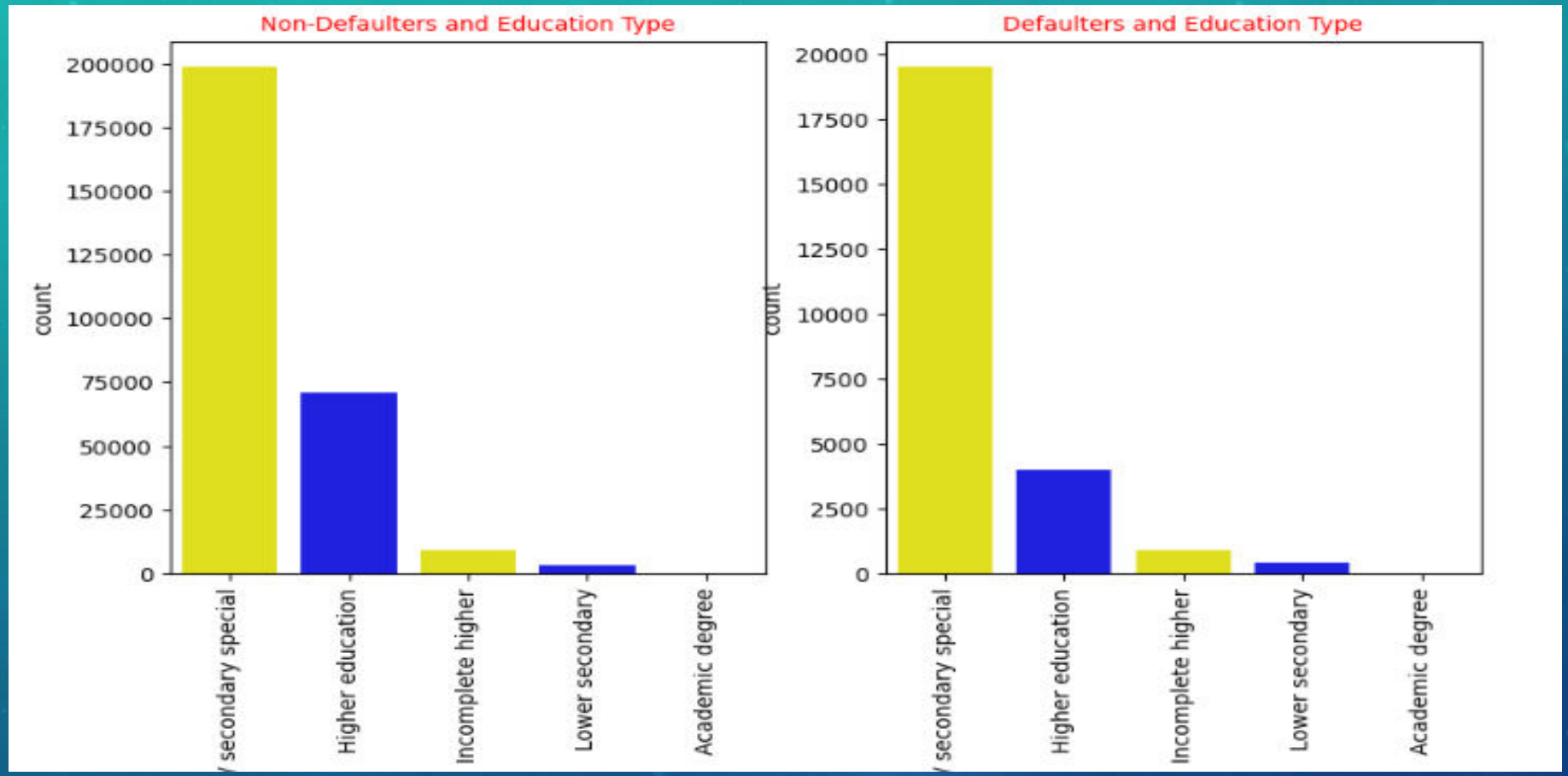


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Working professionals are maximum followed by commercial associates.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also Working professionals are maximum followed by commercial associates.

NAME_EDUCATION_TYPE

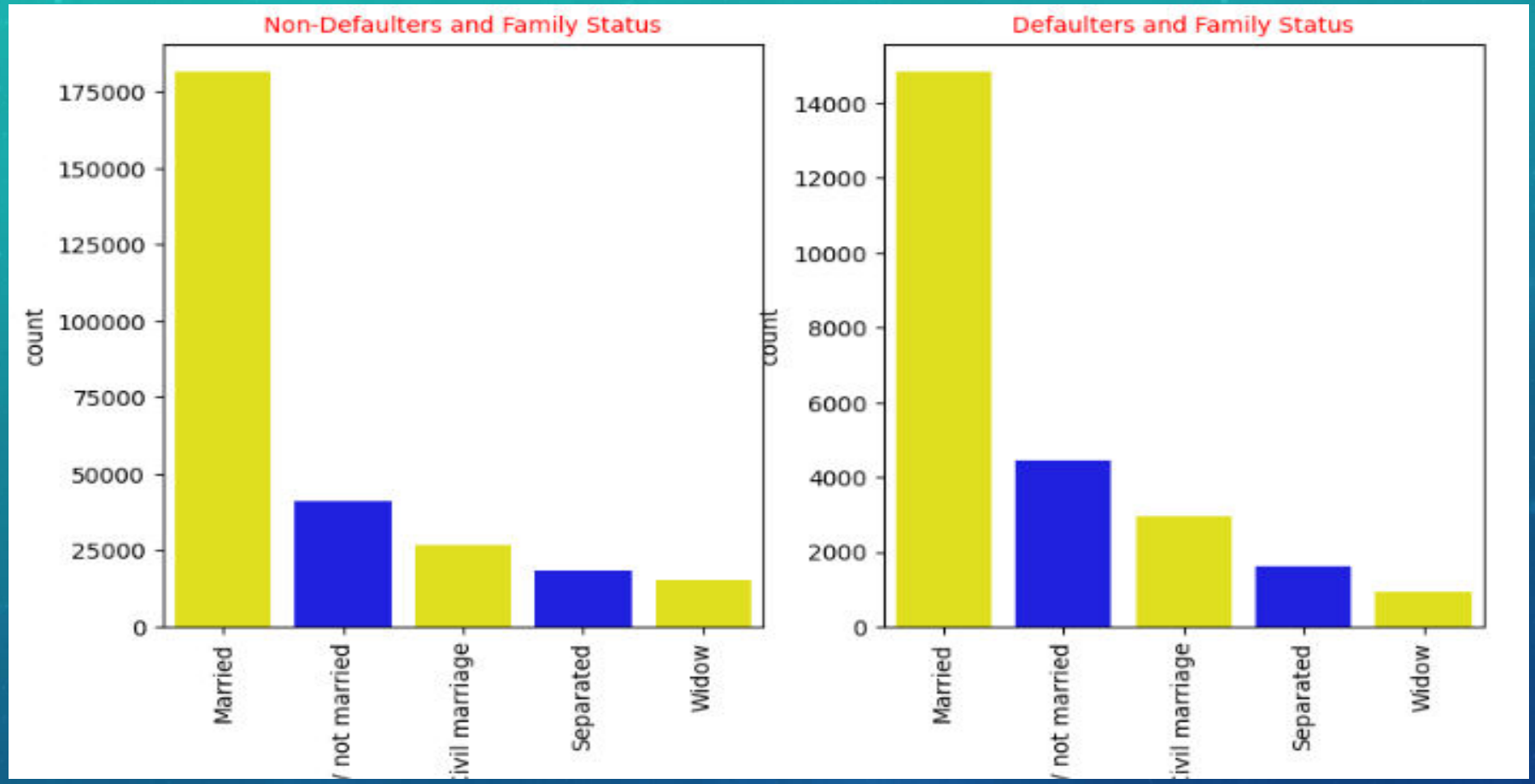


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Most of them are secondary education followed by higher education.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also Most of them are secondary education followed by higher education.

NAME_FAMILY_STATUS

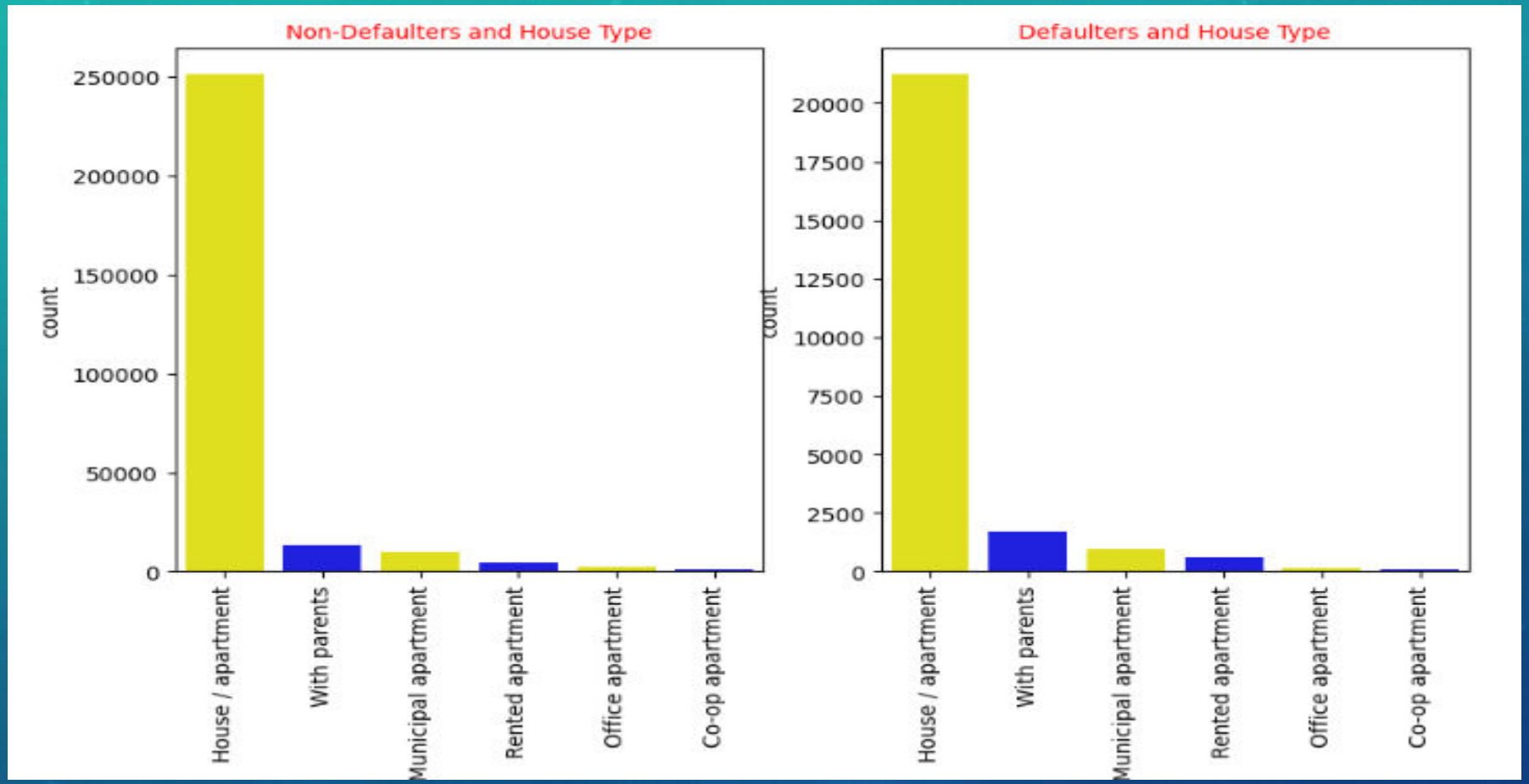


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Most of them are Married.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also Most of them are Married.

NAME_HOUSING_TYPE

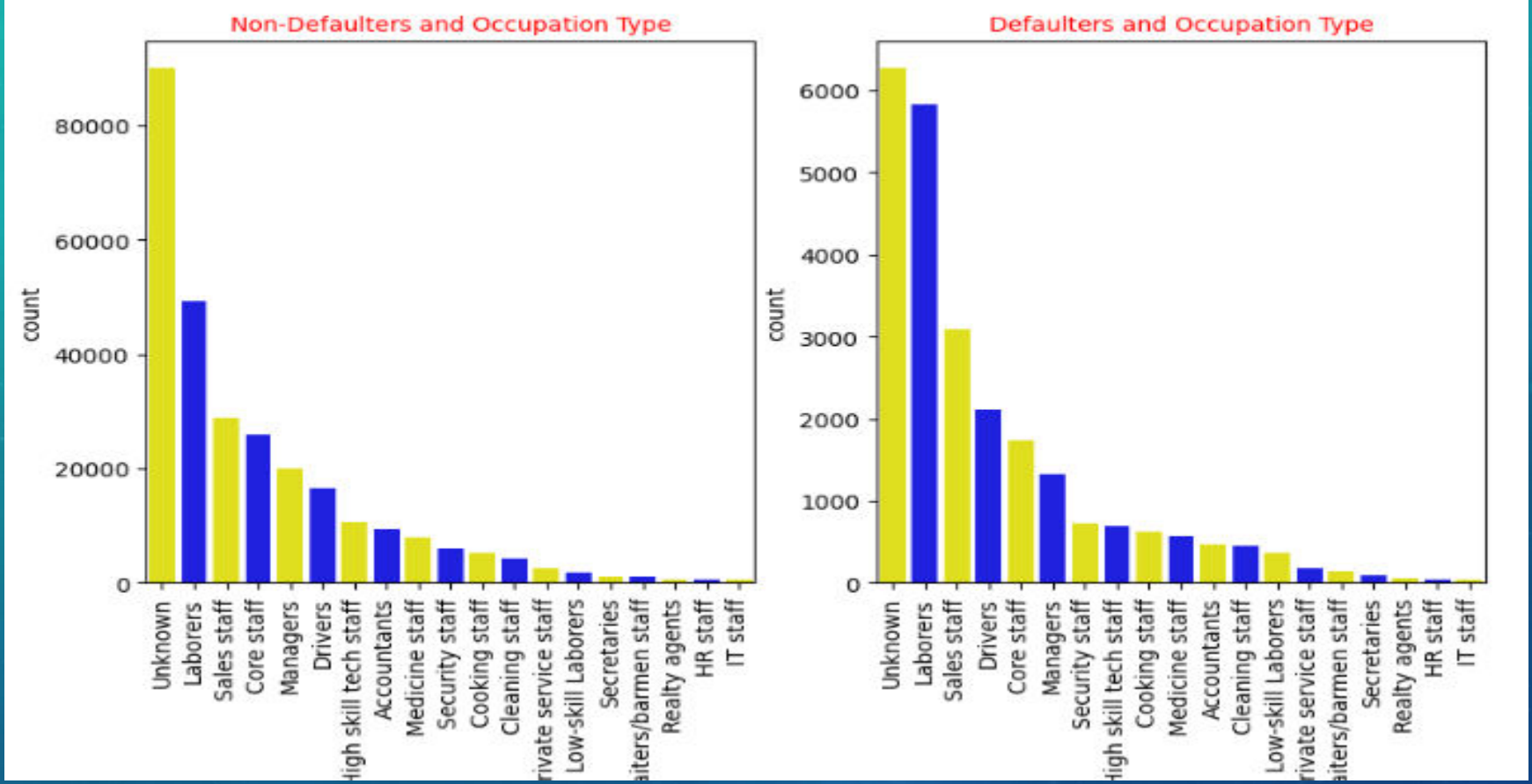


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Most of them are living in House/Apartment.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also Most of them are living in House/Apartment.

OCCUPATION_TYPE

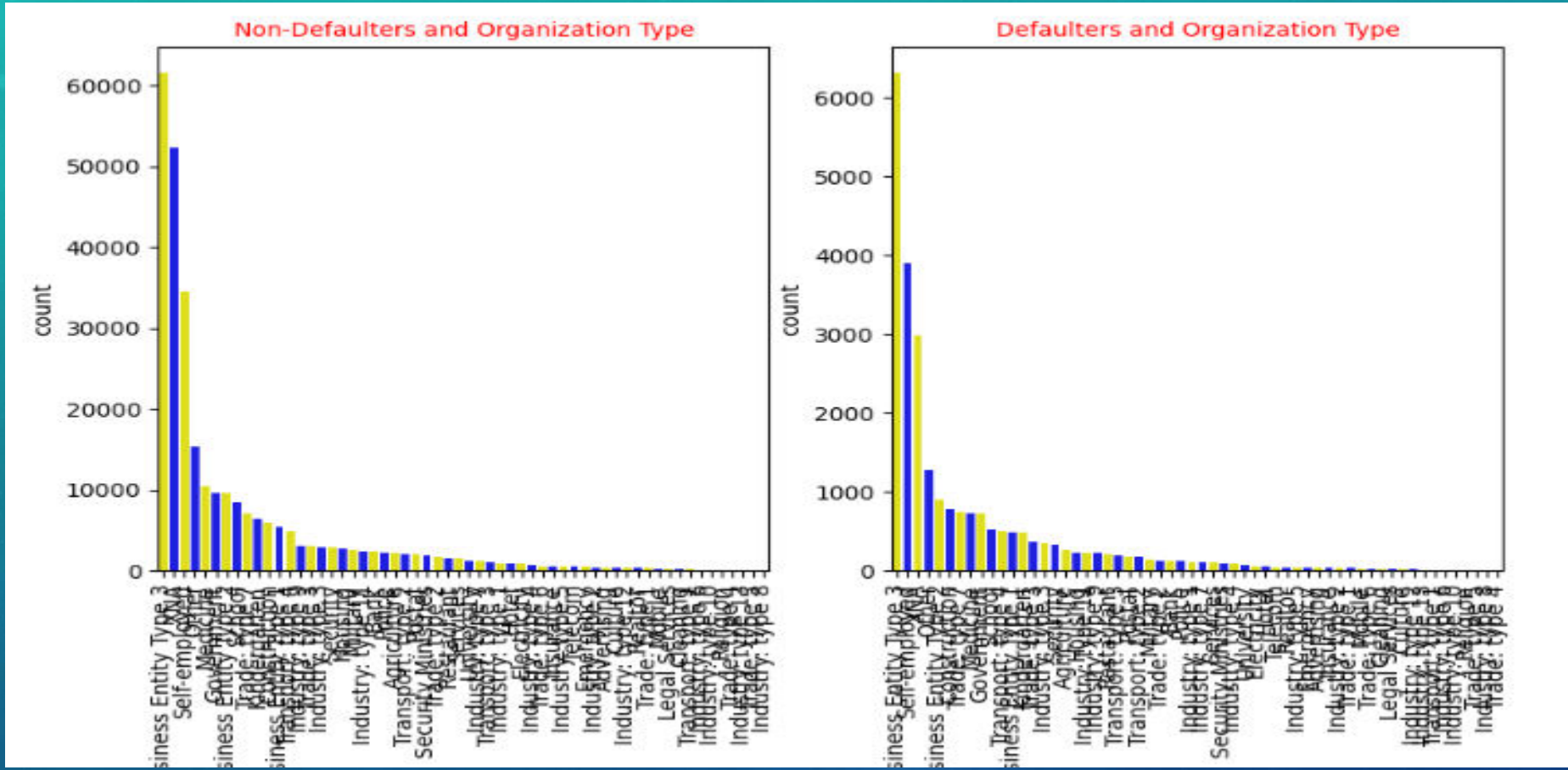


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Most of them are Labourers apart from Unknown occupation type.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also Most of them are Labourers apart from Unknown occupation type.

ORGANIZATION_TYPE

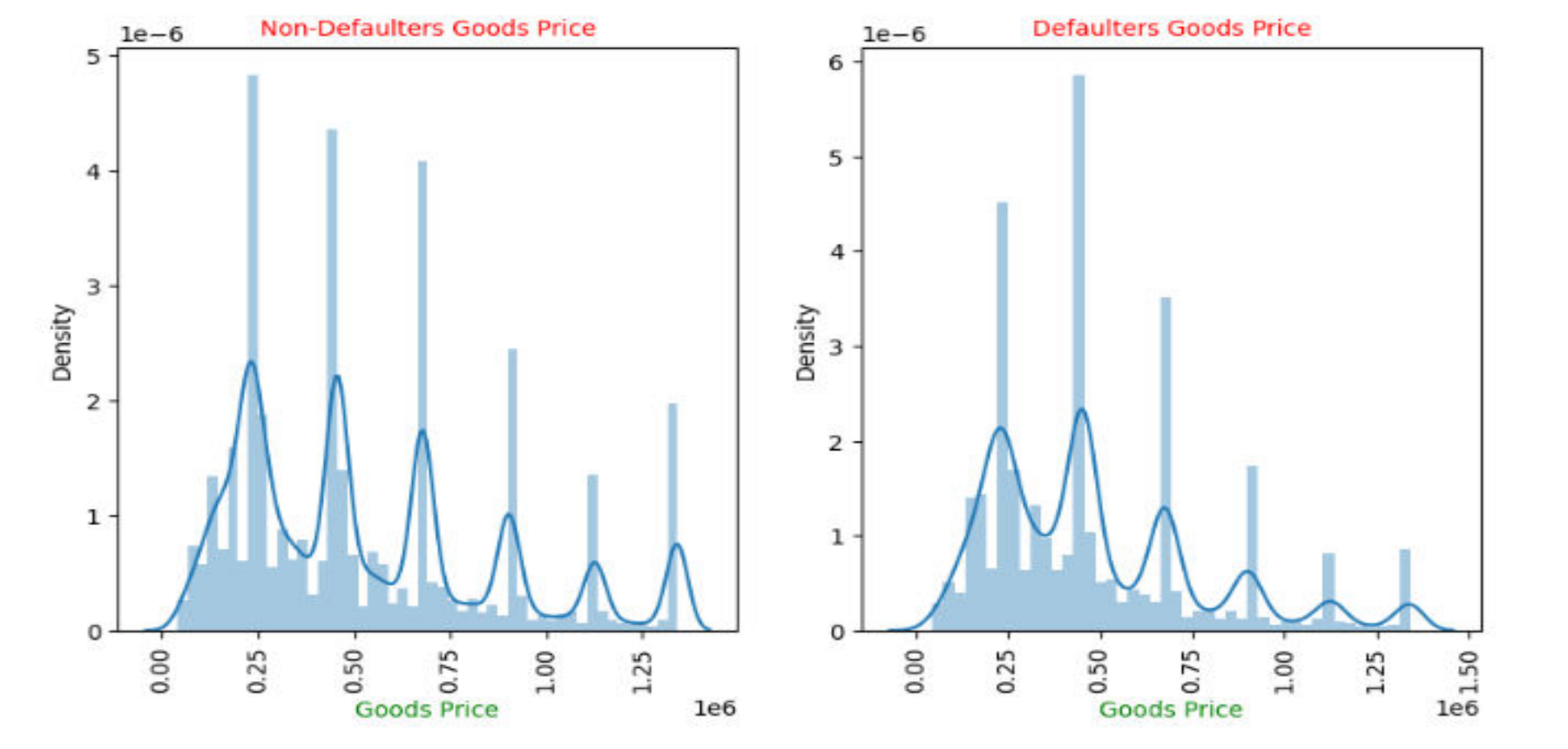


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Most of them belongs to Business Entity Type3.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also belongs to Business Entity Type3.

AMT_GOODS_PRICE

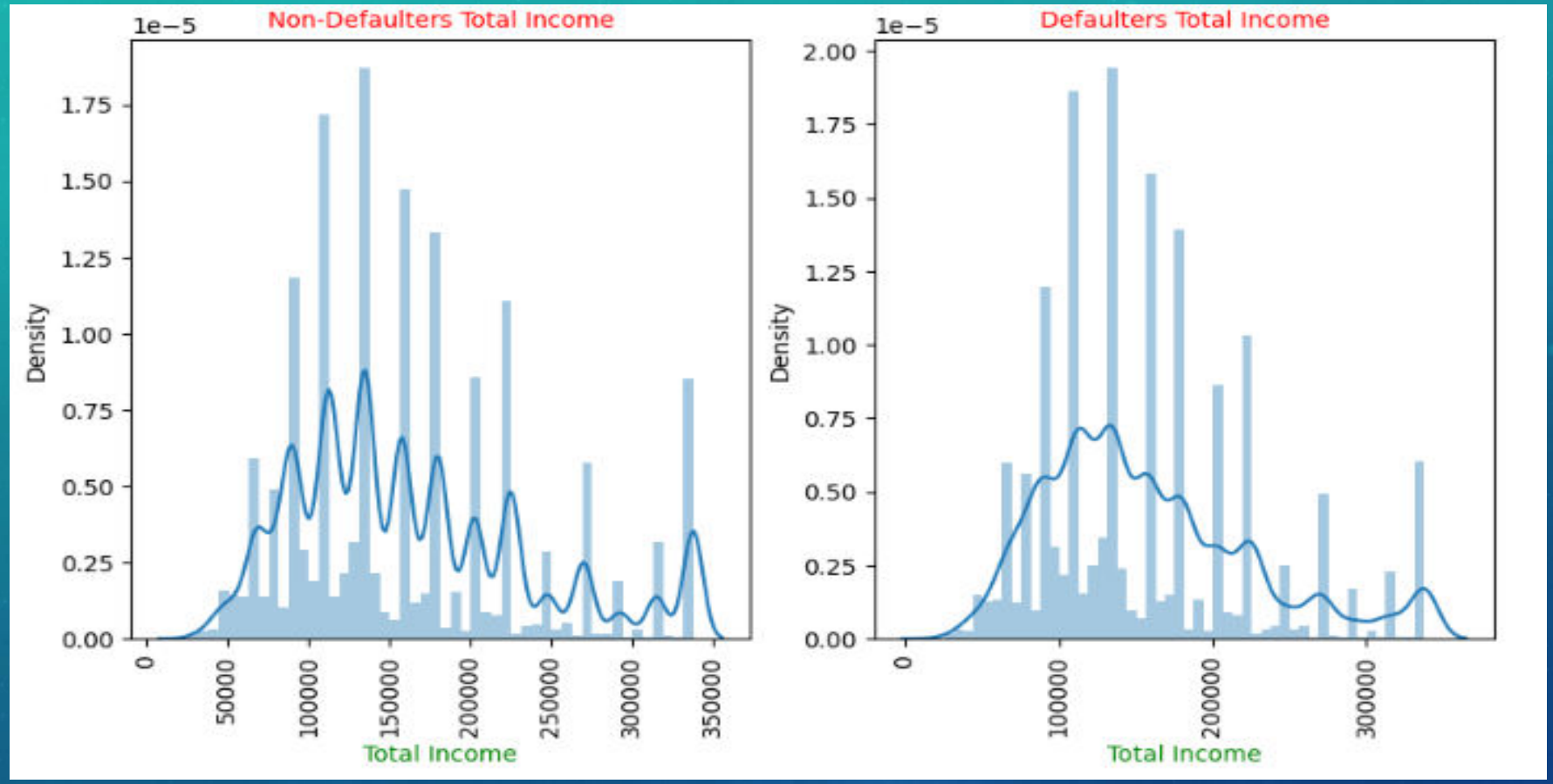


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Most of the goods price falls under 0.2 to 0.5.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also Most of the goods price falls under 0.2 to 0.5.

AMT_INCOME_TOTAL

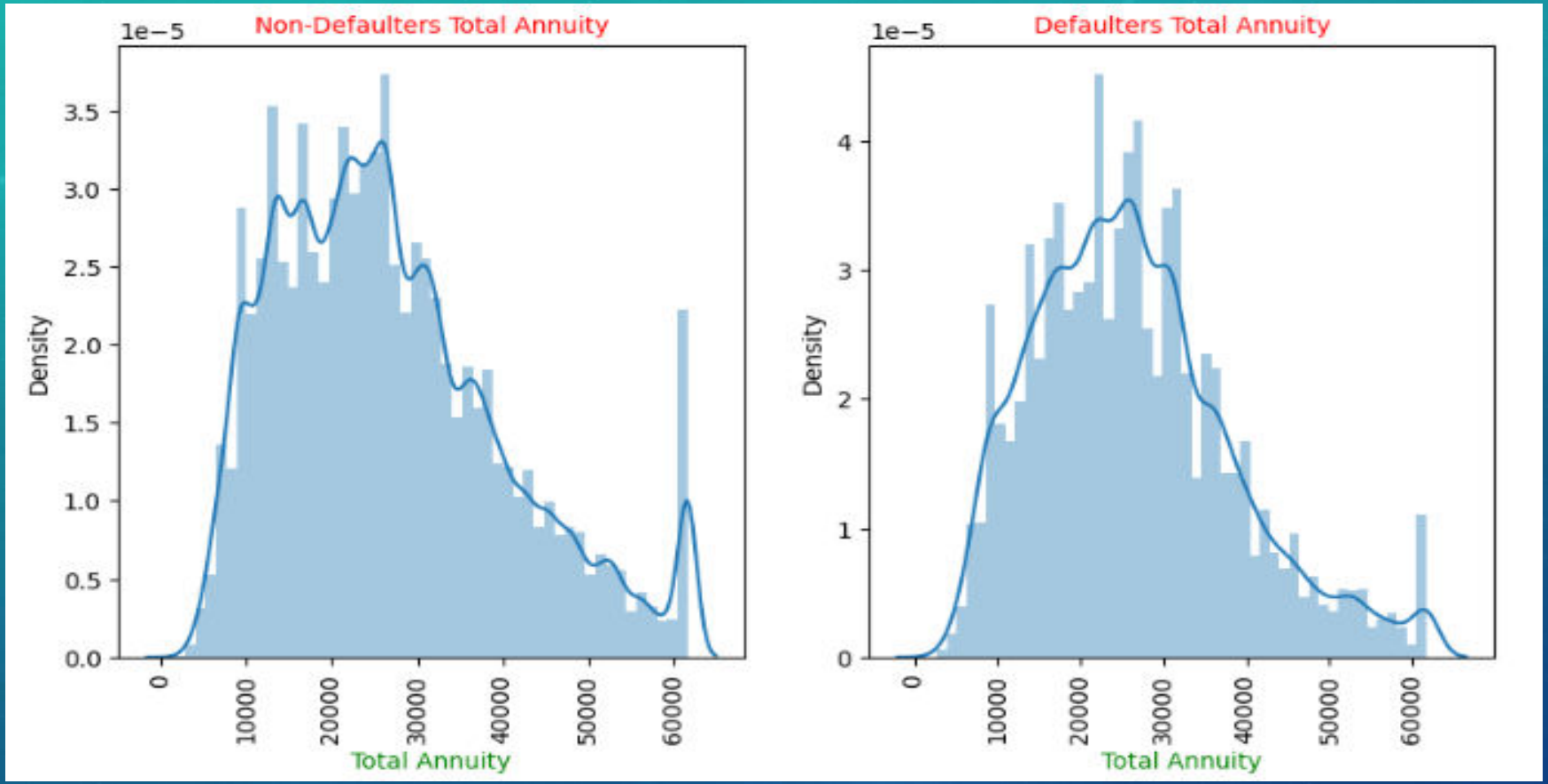


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Most of them earn from 1Lakh to 1.5 Lakh

2.Next, Among Defaulters, i.e., TARGET == 1, Here also Most of them earn from 1Lakh to 1.5 Lakh

AMT_ANNUITY

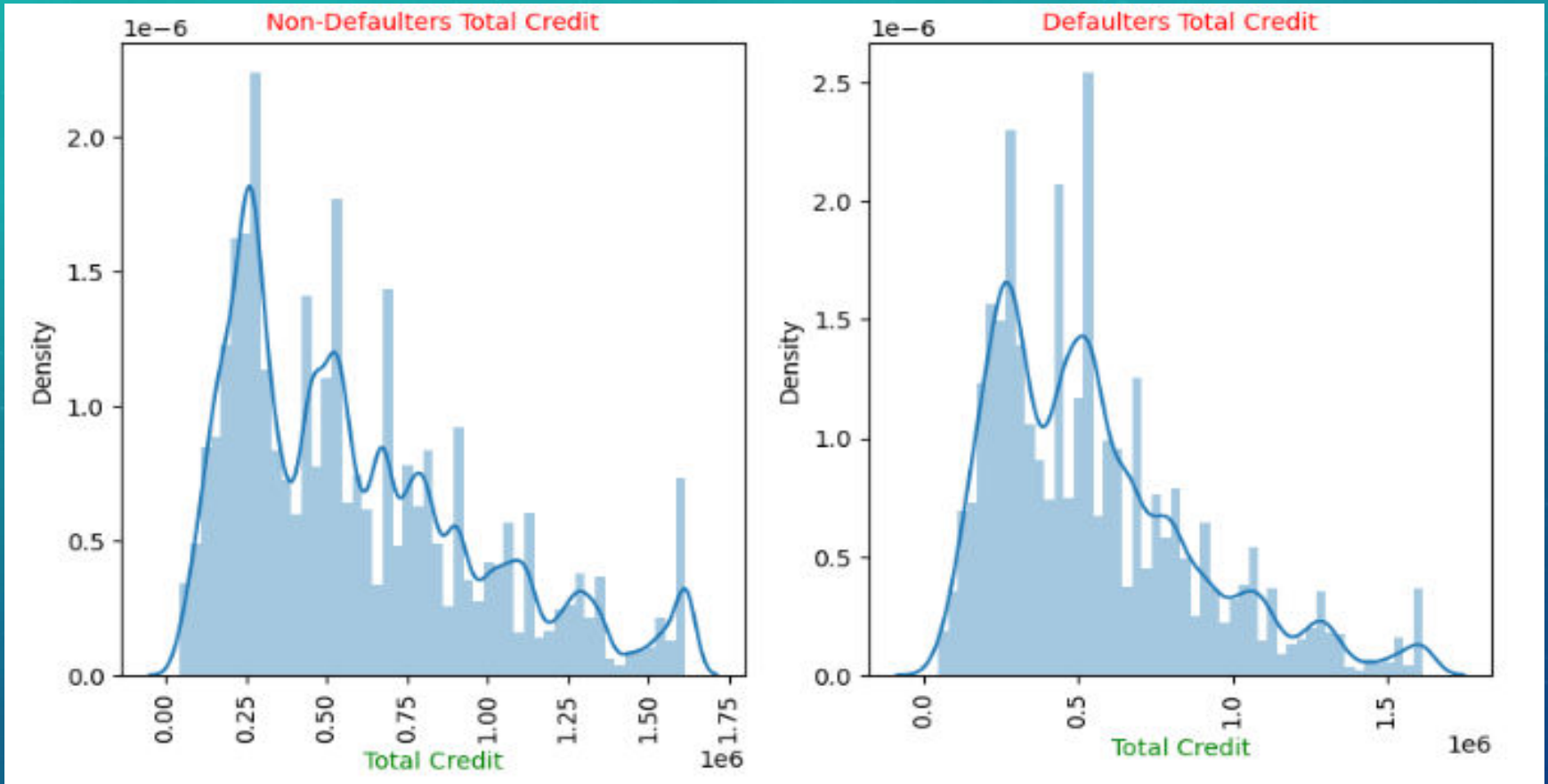


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, total annuity lies from 20k to 30k.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also total annuity lies from 20k to 30k.

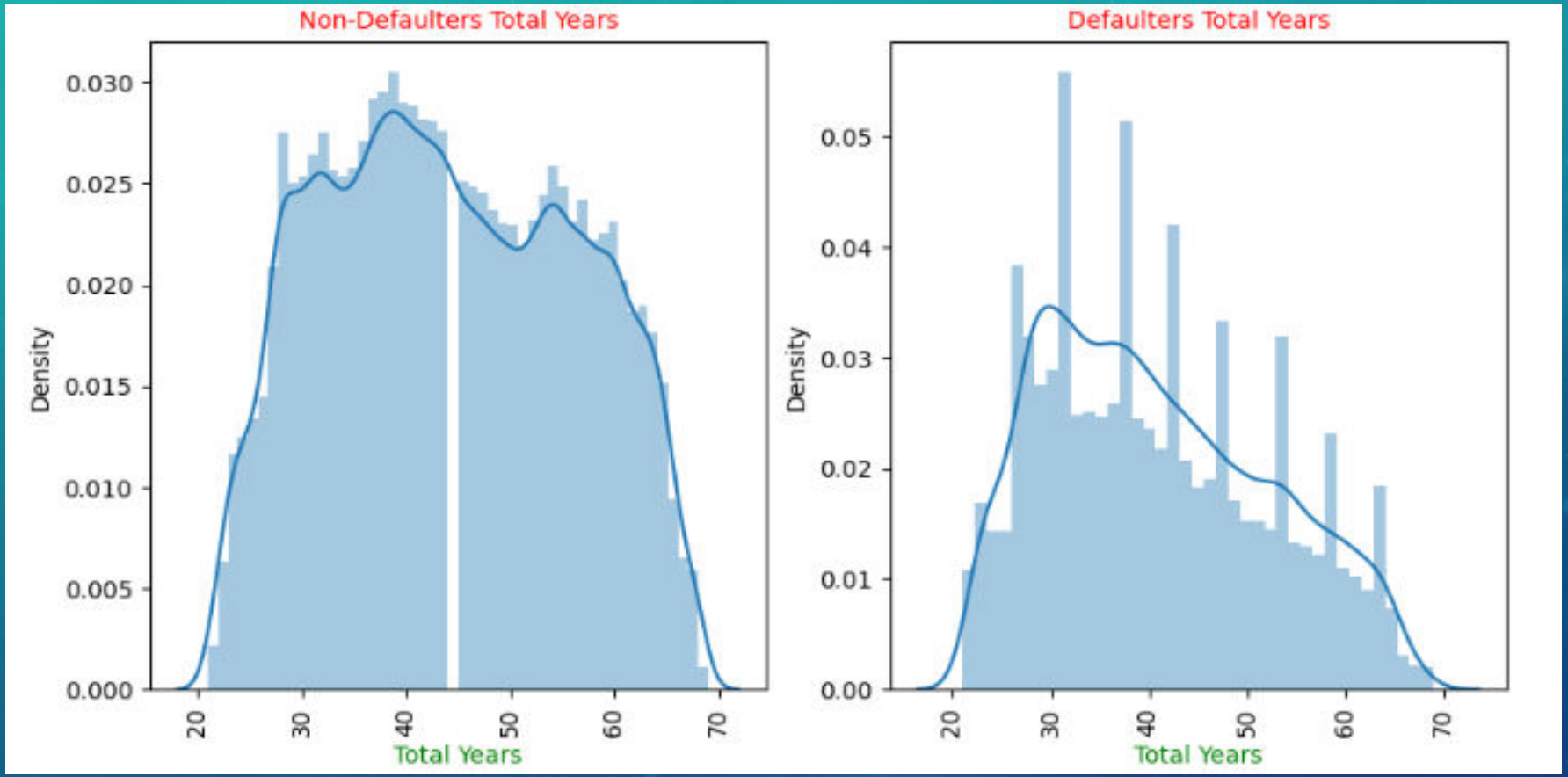
AMT_CREDIT



INSIGHTS:

We can see that the lesser the credit amount of the loan, the more chances of being a defaulter.

YEARS



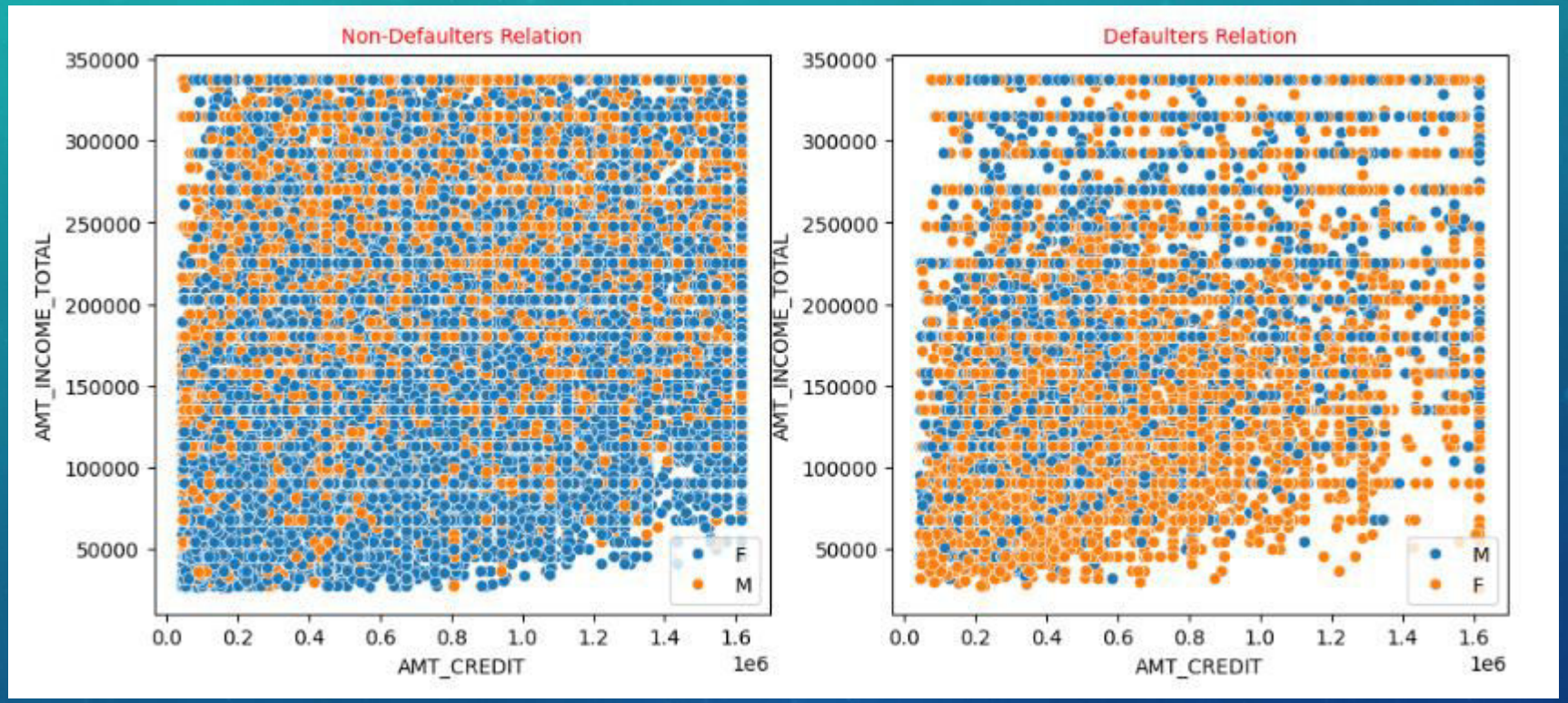
INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, total age lies from 30-40.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also total age lies from 30-40.

BIVARIATE ANALYSIS

AMT_CREDIT and AMT_INCOME_TOTAL Grouped by CODE_GENDER

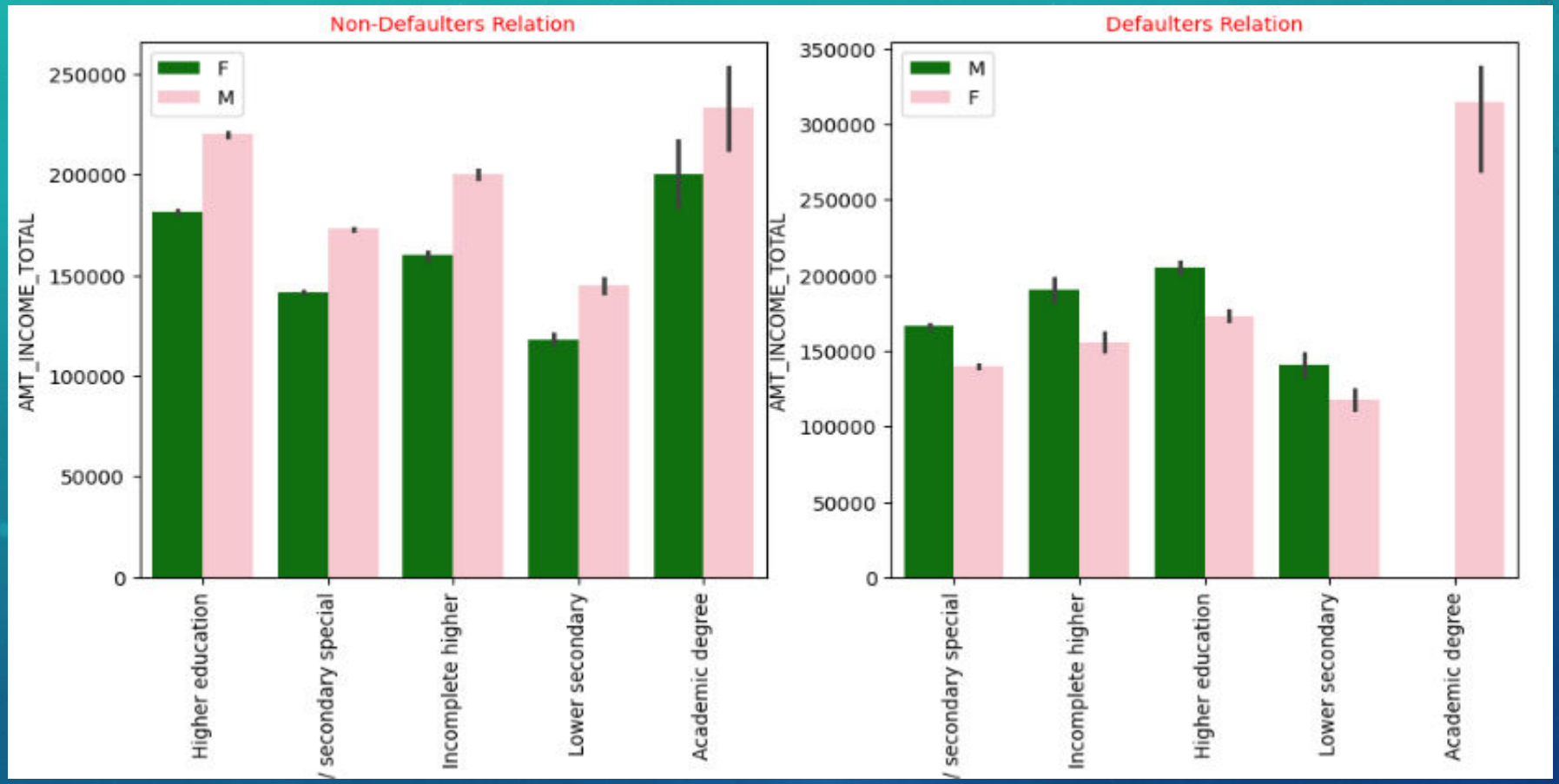


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Female and males are in same ratio.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also Female and males are in same ratio in the initial stage of complete graph.

EDUCATION_TYPE and AMT_INCOME_TOTAL Grouped by CODE_GENDER

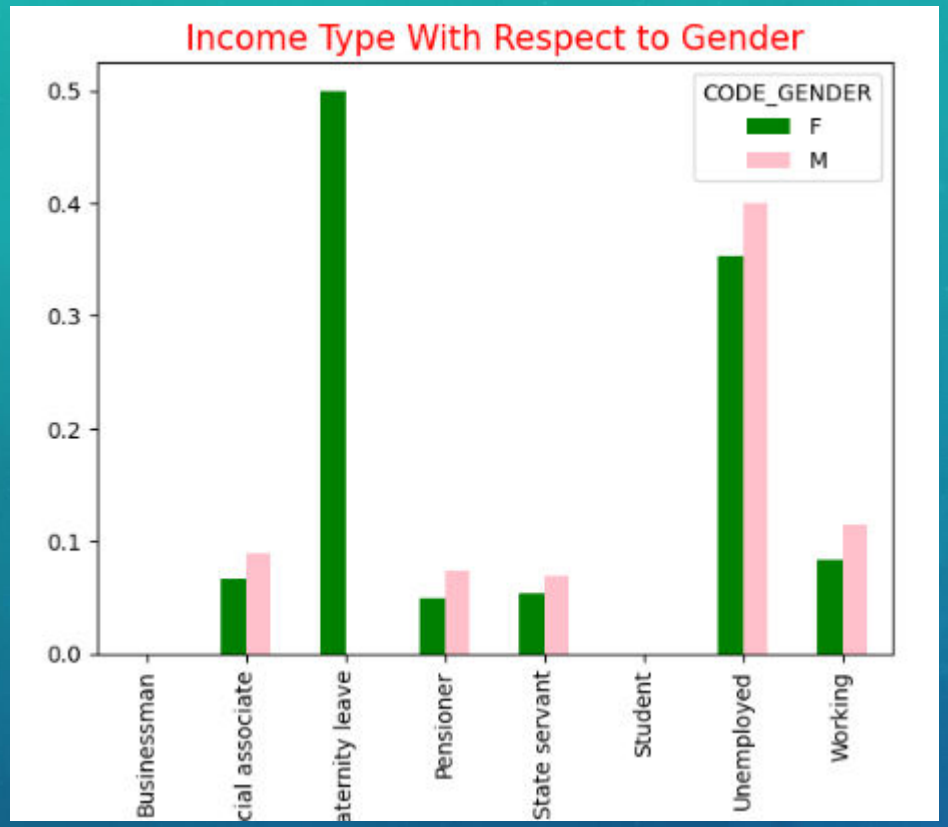


INSIGHTS:

1.Firstly, Among Non-Defaulters, i.e., TARGET == 0, Female in all educational aspects are more non-defaulters than male.

2.Next, Among Defaulters, i.e., TARGET == 1, Here also we can see that males are earning more as well being the more defaulters.

AMT_INCOME_TYPE and CODE_GENDER



INSIGHTS:

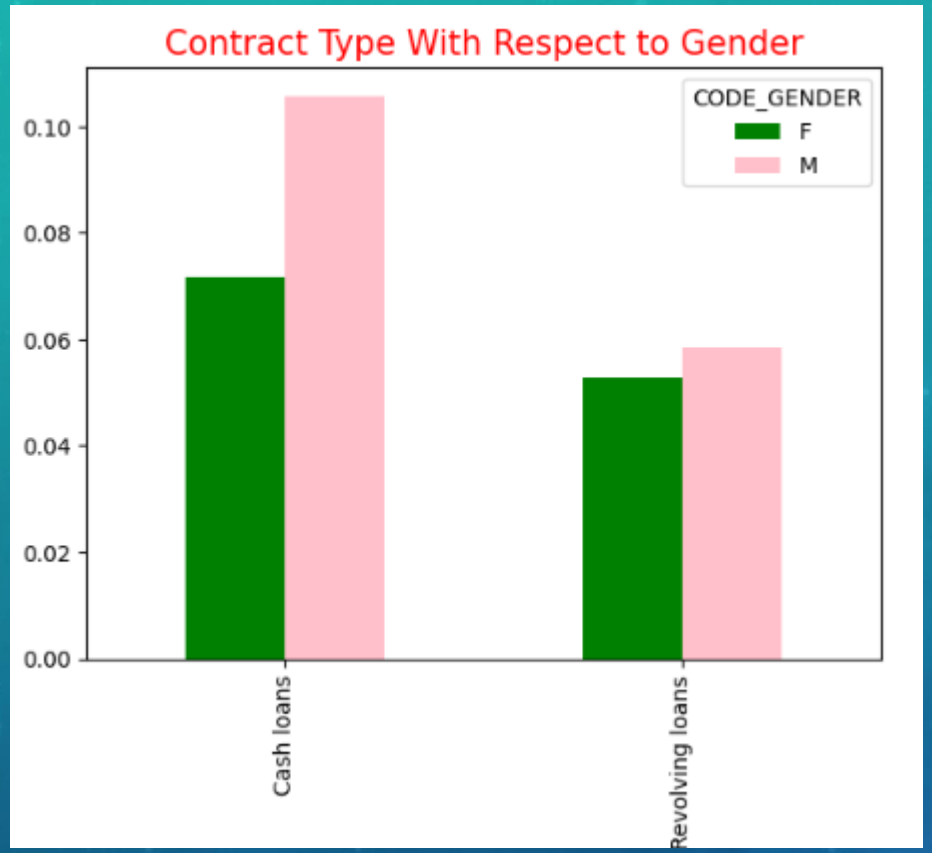
1.We can see that Unemployed people are more defaulters in both male and female case.

2.Males are more unemployed than female.

3.Maternity leave females are also in higher no in defaulters list.

4.Male number’s are more compare to female in defaulters list.

NAME_CONTRACT_TYPE and CODE_GENDER



INSIGHTS:

1. For Cash loans, males are more in number than females and similarly with Revolving loans.

MULTI-VARIATE ANALYSIS

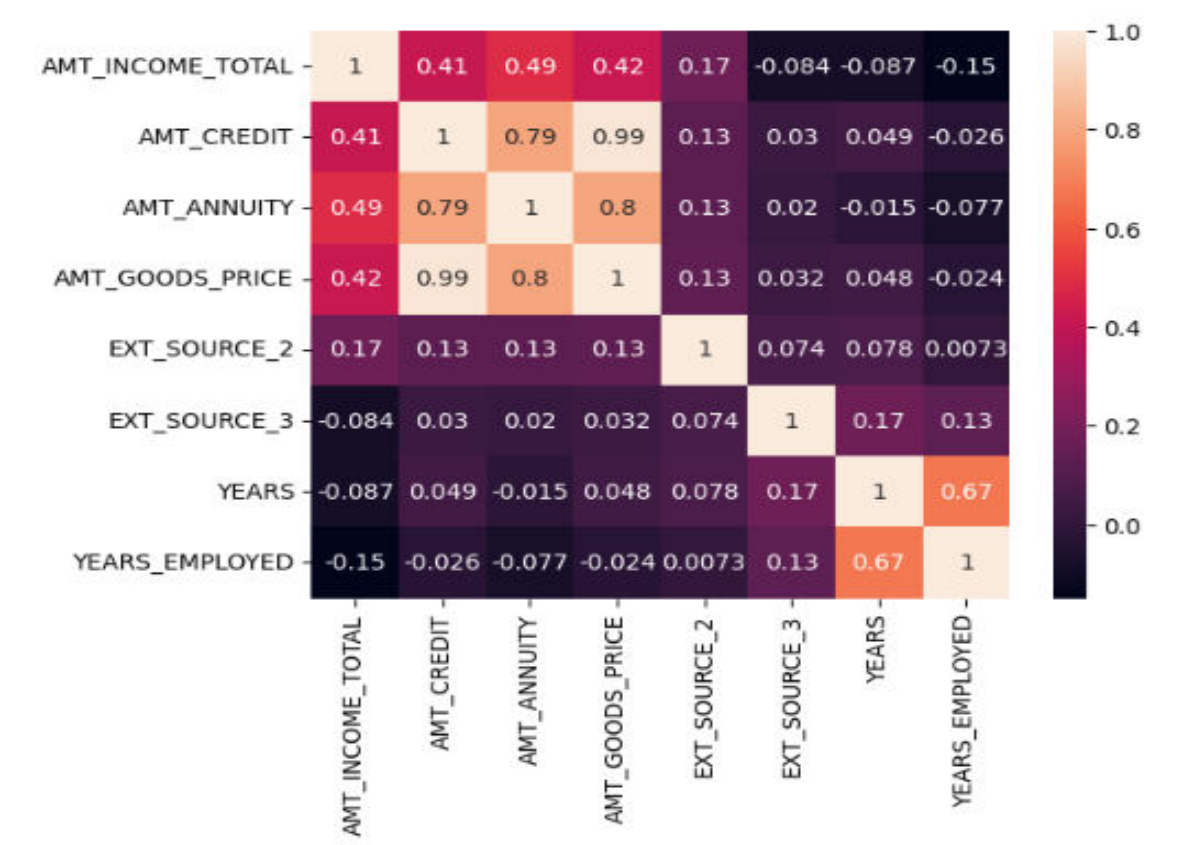
Continuous columns = ['AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY', 'AMT_GOODS_PRICE','EXT_SOURCE_2','EXT_SOURCE_3','YEARS','YEARS_EMPLOYED']

Heat Maps are plotted against the continuous numerical variables for Non-Defaulters

Heatmap for Non-Defaulters (Target=0)

```
T0 = app[(app.TARGET == 0)][cont_col].corr()
T0
```

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	EXT_SOURCE_2	EXT_SOURCE_3	YEARS	YEARS_EMPLOYED
AMT_INCOME_TOTAL	1.000000	0.414053	0.492669	0.417680	0.172204	-0.083589	-0.086546	-0.149292
AMT_CREDIT	0.414053	1.000000	0.794646	0.985841	0.128217	0.029890	0.049049	-0.026400
AMT_ANNUITY	0.492669	0.794646	1.000000	0.797660	0.127629	0.020186	-0.015176	-0.077156
AMT_GOODS_PRICE	0.417680	0.985841	0.797660	1.000000	0.134877	0.031937	0.047632	-0.024422
EXT_SOURCE_2	0.172204	0.128217	0.127629	0.134877	1.000000	0.073858	0.078081	0.007296
EXT_SOURCE_3	-0.083589	0.029890	0.020186	0.031937	0.073858	1.000000	0.173505	0.133088
YEARS	-0.086546	0.049049	-0.015176	0.047632	0.078081	0.173505	1.000000	0.674715
YEARS_EMPLOYED	-0.149292	-0.026400	-0.077156	-0.024422	0.007296	0.133088	0.674715	1.000000



INSIGHTS:

The more the correlated percentage means the high value for being a Non-defaulter.

* 0.99% = AMT_CREDIT & AMT_GOODS_PRICE.
* 0.79% = AMT_CREDIT & AMT_ANNUITY.
* 0.80% = AMT_GOODS_PRICE & AMT_ANNUITY.
* 0.67% = Age & DAYS_EMPLOYED

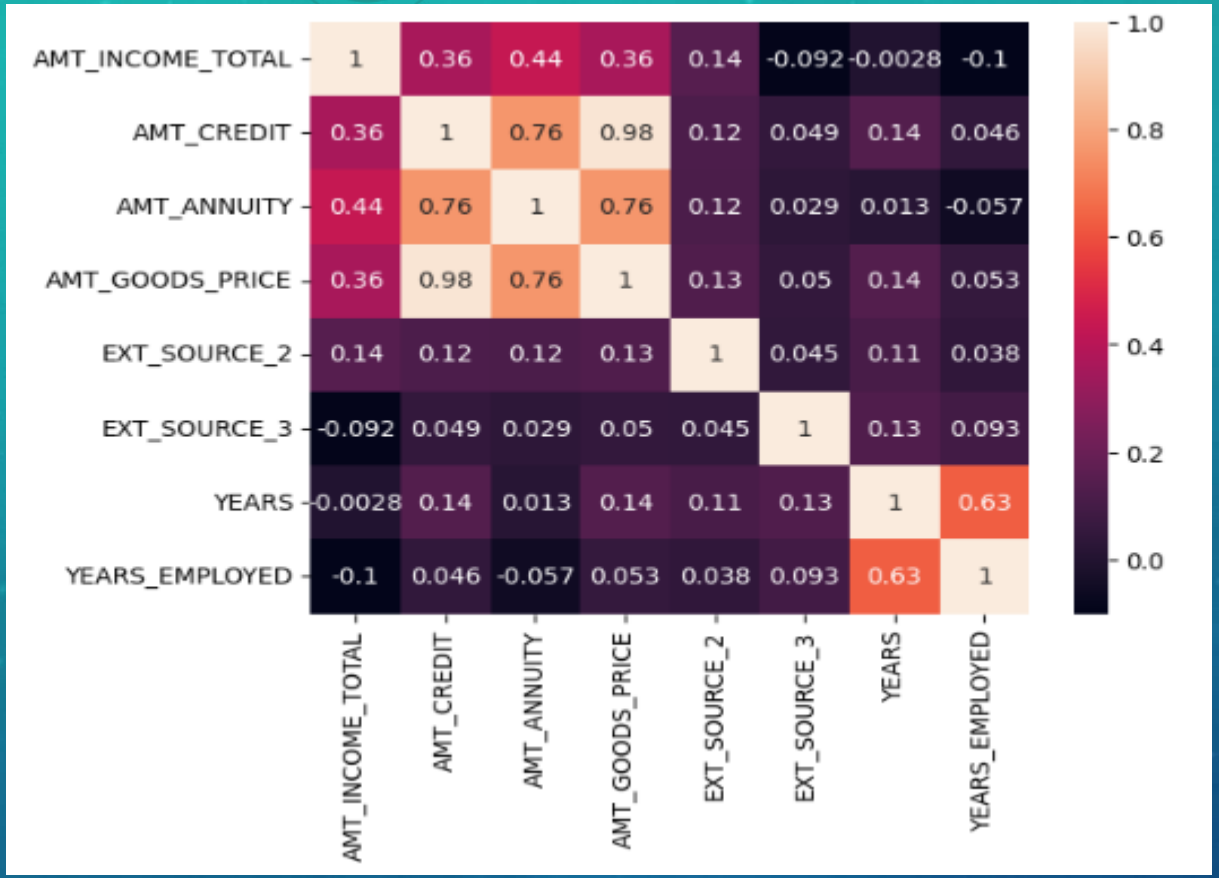
Continuous columns = ['AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY', 'AMT_GOODS_PRICE','EXT_SOURCE_2','EXT_SOURCE_3','YEARS','YEARS_EMPLOYED']

Heat Maps are plotted against the continuous numerical variables for Defaulters

Heatmap for Defaulters (Target=1)

```
[ ]  T1 = app[(app.TARGET == 1)][cont_col].corr()
     T1
```

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	EXT_SOURCE_2	EXT_SOURCE_3	YEARS	YEARS_EMPLOYED
AMT_INCOME_TOTAL	1.000000	0.355803	0.436508	0.357775	0.144791	-0.092406	-0.002804	-0.100297
AMT_CREDIT	0.355803	1.000000	0.759897	0.982183	0.120322	0.048835	0.137802	0.045815
AMT_ANNUITY	0.436508	0.759897	1.000000	0.760749	0.115430	0.028884	0.013191	-0.056697
AMT_GOODS_PRICE	0.357775	0.982183	0.760749	1.000000	0.130279	0.049924	0.139150	0.053134
EXT_SOURCE_2	0.144791	0.120322	0.115430	0.130279	1.000000	0.044654	0.108342	0.037757
EXT_SOURCE_3	-0.092406	0.048835	0.028884	0.049924	0.044654	1.000000	0.130197	0.093096
YEARS	-0.002804	0.137802	0.013191	0.139150	0.108342	0.130197	1.000000	0.626325
YEARS_EMPLOYED	-0.100297	0.045815	-0.056697	0.053134	0.037757	0.093096	0.626325	1.000000



INSIGHTS:

The more the correlated percentage means the high value for being a defaulter.

* 0.98% = AMT_CREDIT & AMT_GOODS_PRICE
* 0.76% = AMT_CREDIT & AMT_ANNUITY
* 0.76% = AMT_GOODS_PRICE & AMT_ANNUITY
* 0.58% = YEARS & DAYS_EMPLOYED

PREVIOUS APPLICATION DATA ANALYSIS

- Import necessary Modules
- Reading and Analyzing Data
- Checking the Null Percentage of values
- Imputing the values
- Standardization of Data
- Identifying the Outliers
- Merging the 2 Data Frames
- Univariate Analysis
- Bivariate and Multivariate Analysis

PREVIOUS APPLICATION DATA ANALYSIS STEPS

- Same like Application Data Analysis, analyze the data of the given excel like checking column headers.
- Check the shape of the data frame. (1670214, 37)
- Check the no of null value rows present in each and every column and also its data types using info() function and for statistical information of data frame using describe() function.
- First we need to check the percentage of null values of columns and drop the columns with 50% null values as their presence does impact the statistics.
- Remaining columns who have less percentage of null values , we will impute the columns with Mean()/Median() – Numerical columns and Mode() – Categorical columns.
- Standardize the values of columns like for Time columns if all the values should be converted to either Seconds or Minutes or Hours and if any columns have negative values convert it into positive values.

PREVIOUS APPLICATION DATA ANALYSIS STEPS

- Identify the Outliers using boxplot, if any outliers calculate the IQR for them, calculate the upper bounds and lower bounds.
- Merge the Application data Frame and Previous Application data Frame together on SK_ID_CURR and draw the insights from the combined data frame.
- Segregate all the columns of combined data frame based on this data type into Categorical and Numerical Variables for Data Visualization using MatplotLib and seaborn libraries.
- Now Univariate Analysis is done on the Categorical variables using BAR Plot and it is also done on the Numerical variables using DIST Plot and their insights are taken accordingly.
- Bivariate Analysis is done using Count Plot and Dist plots and their insights are taken.
- Multivariate Analysis is done between Continuous Numerical Variables using Heat Maps.

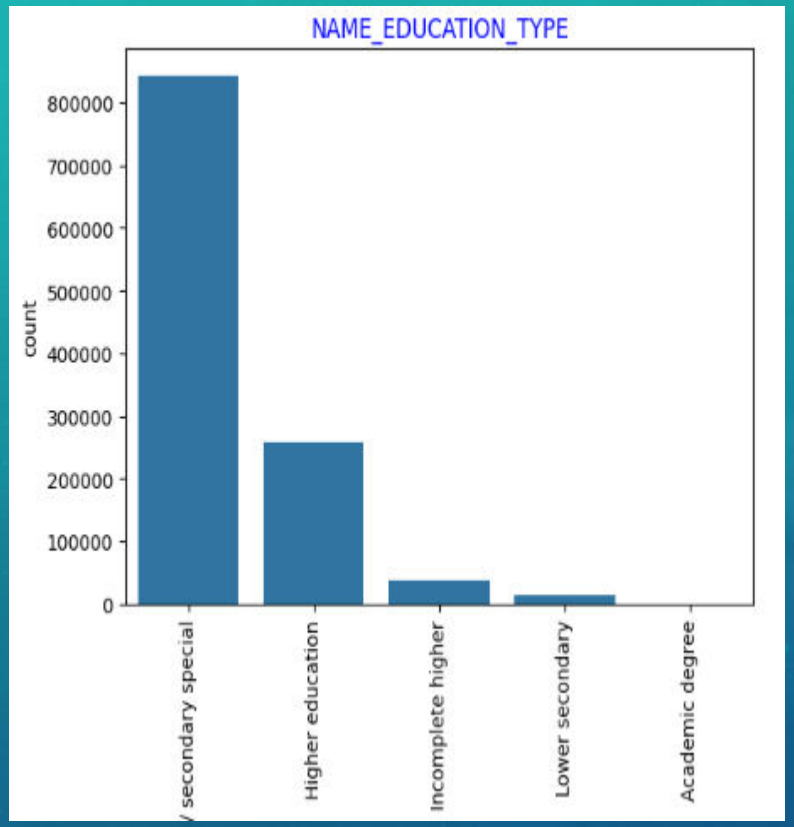
UNIVARIATE ANALYSIS

```
[178] #Univariate Analysis for Categorical Columns
      for col in cat_merged:
        sns.countplot(x=merged[col])
        plt.title(col,color='blue')
        plt.xticks(rotation=90)
        plt.show()
```

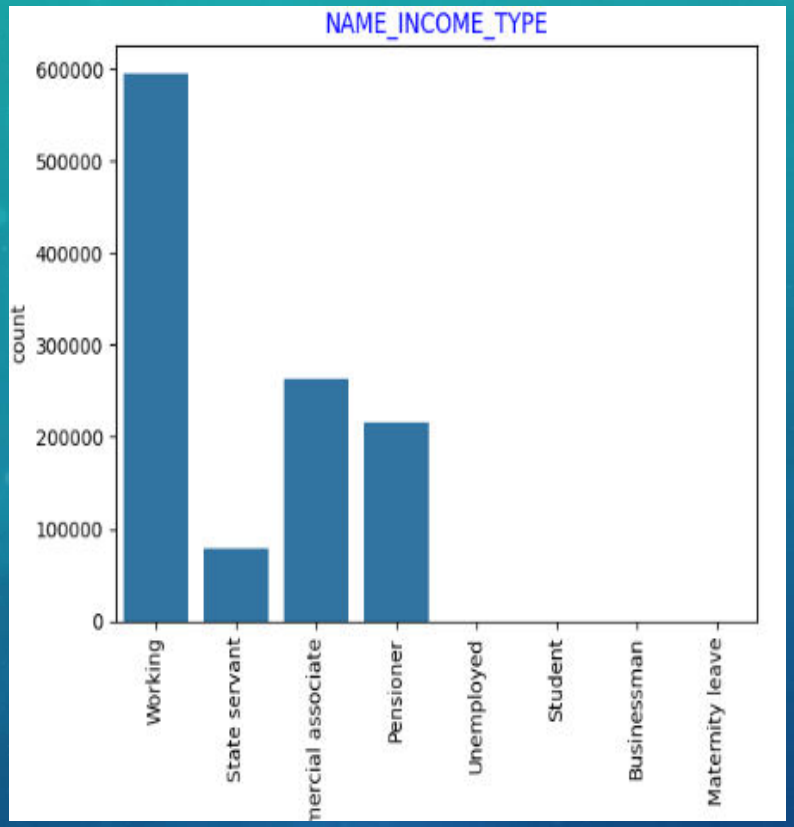
Univariate Analysis is done for Categorical columns using Count Plot.

```
[179] #Univariate Analysis for Numerical Columns
      for col in num_merged:
        sns.distplot(x=merged[col])
        plt.title(col,color='blue')
        plt.xticks(rotation=90)
        plt.show()
```

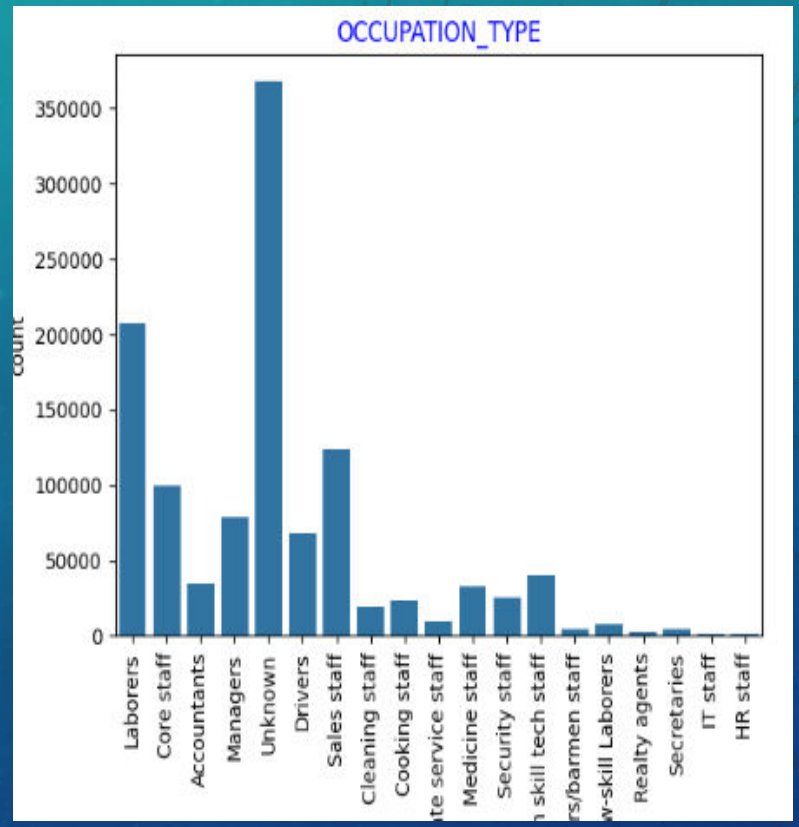
Univariate Analysis is done for Numerical columns using Dist Plot.



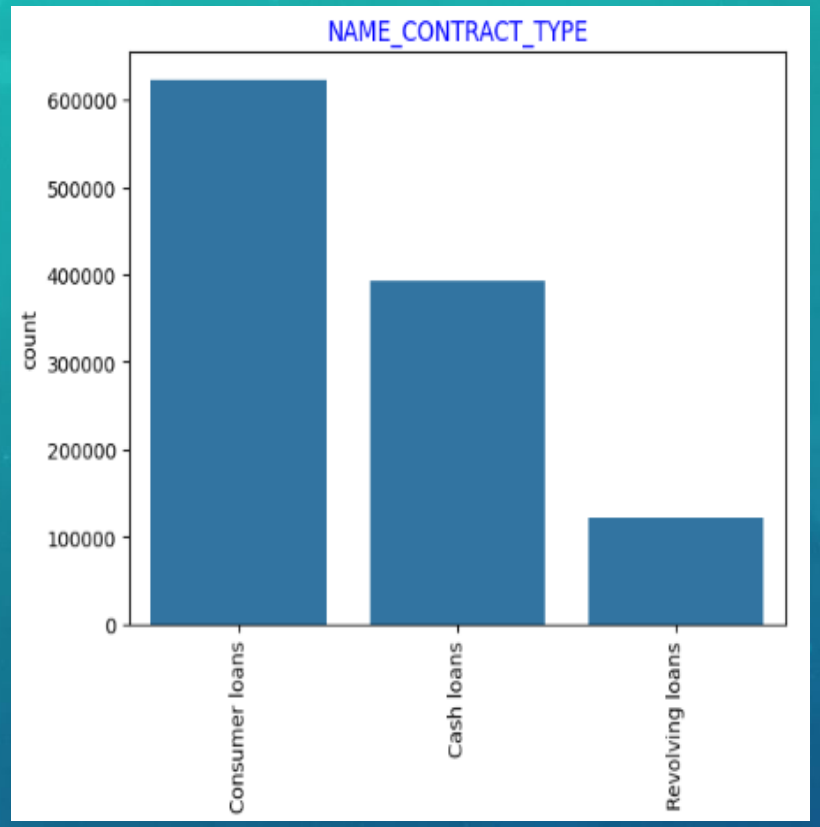
NAME_EDUCATION_TYPE

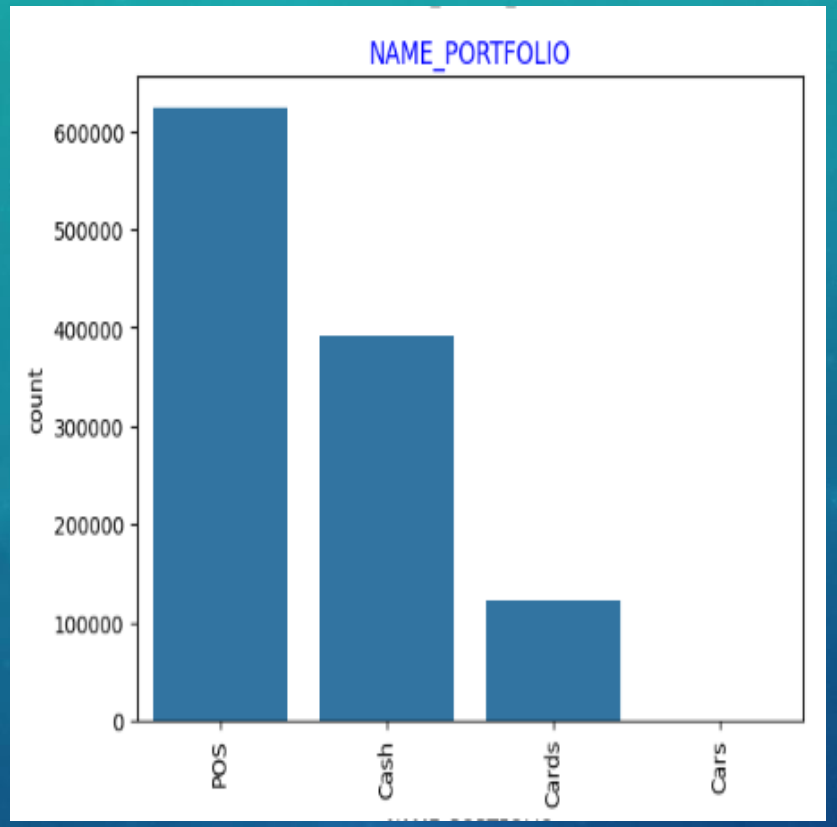


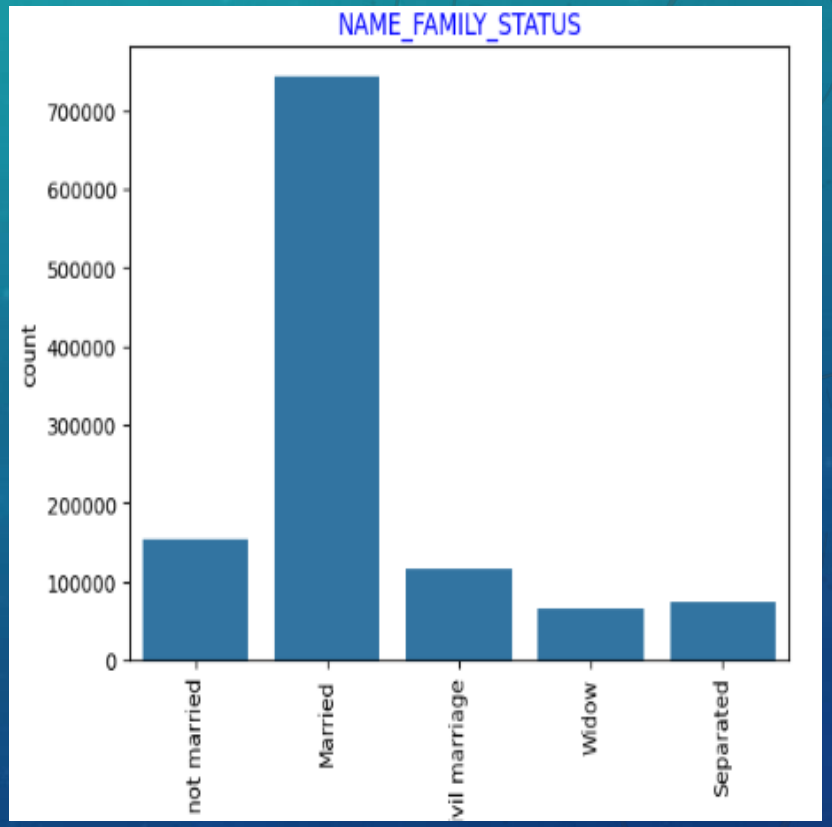
NAME_INCOME_TYPE



OCCUPATION_TYPE



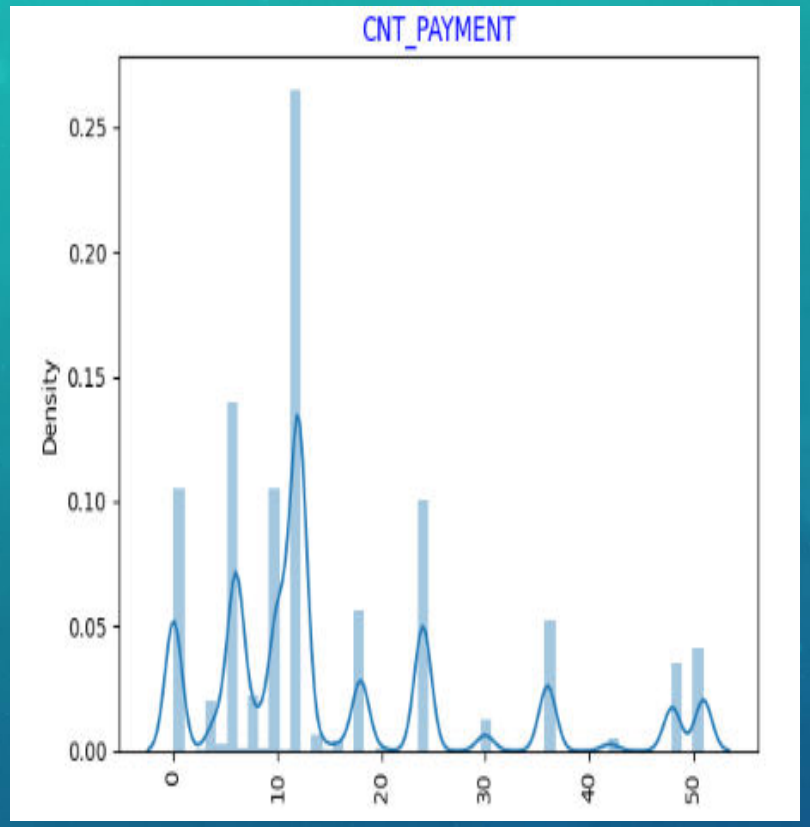


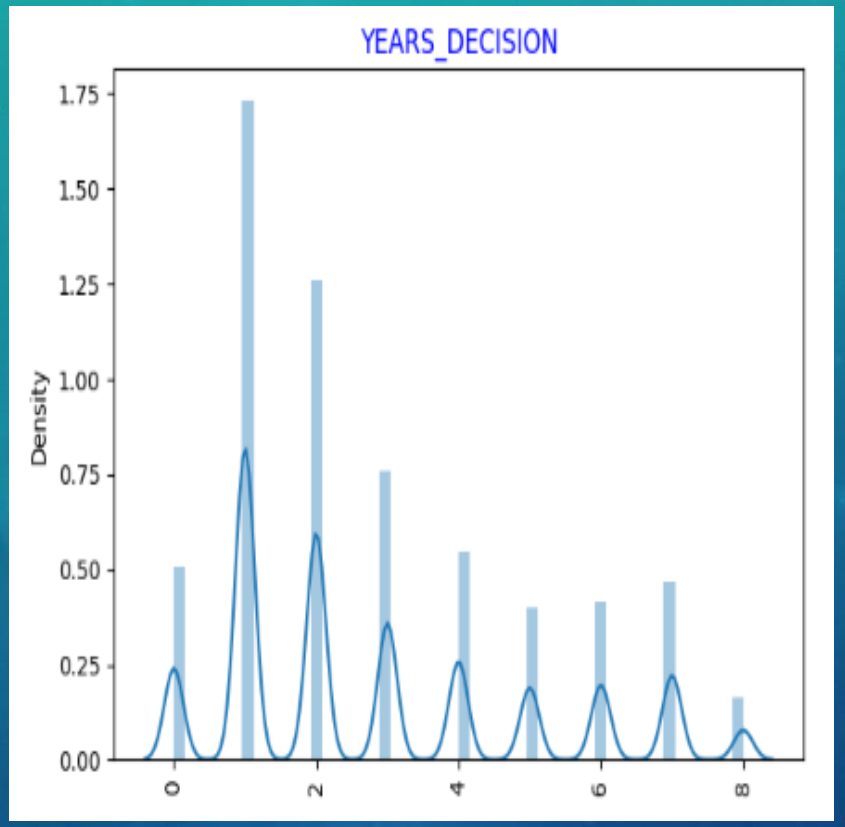


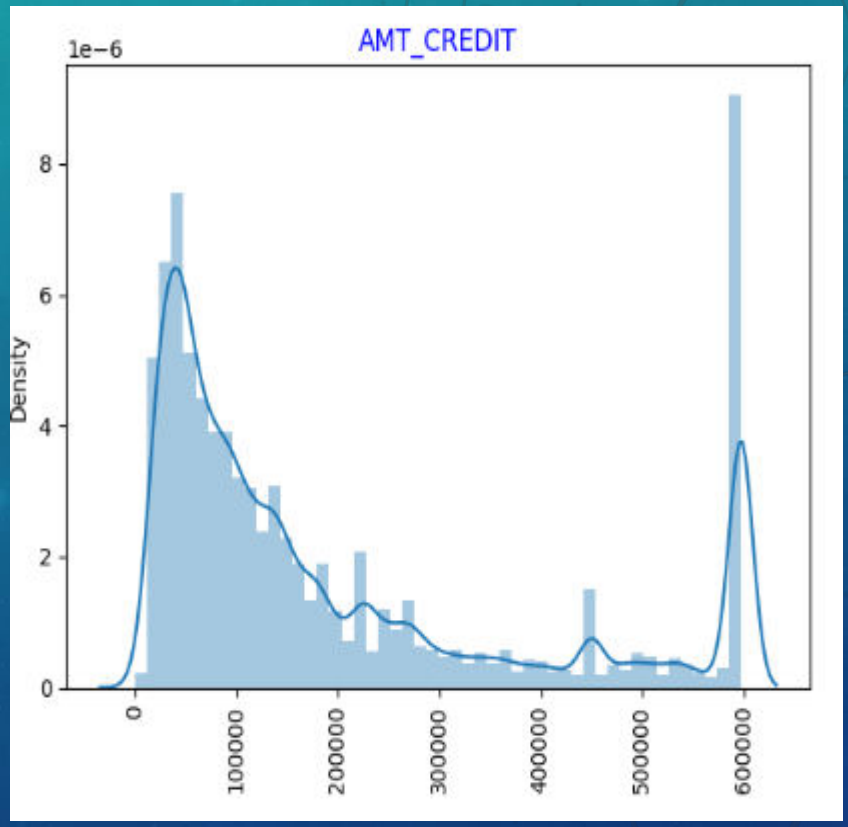
NAME_CONTRACT_TYPE

NAME_PORTFOLIO

NAME_FAMILY_STATUS



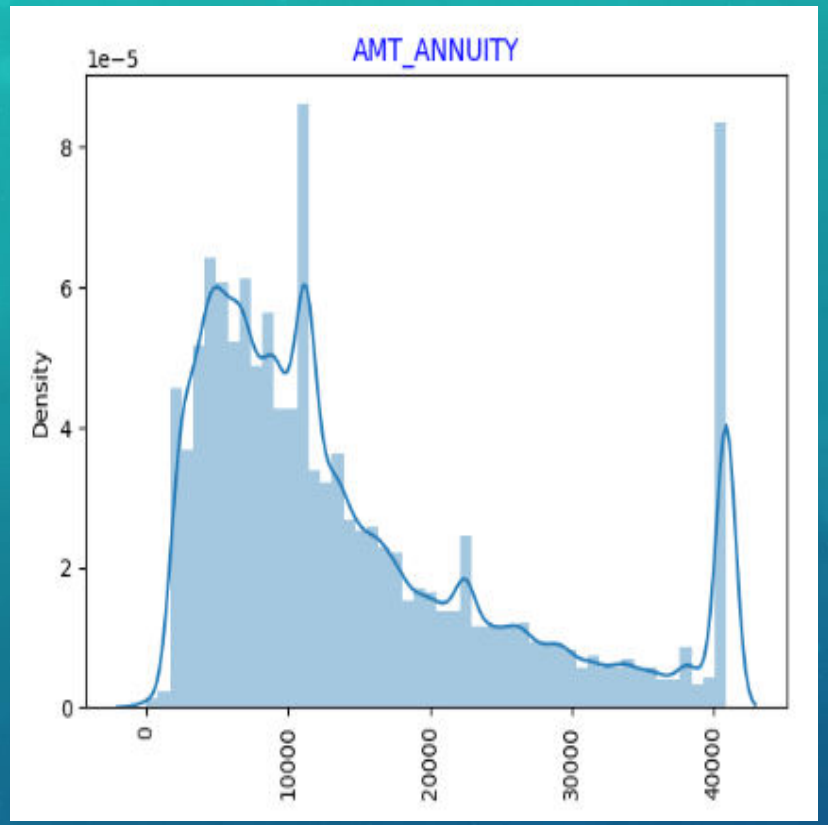


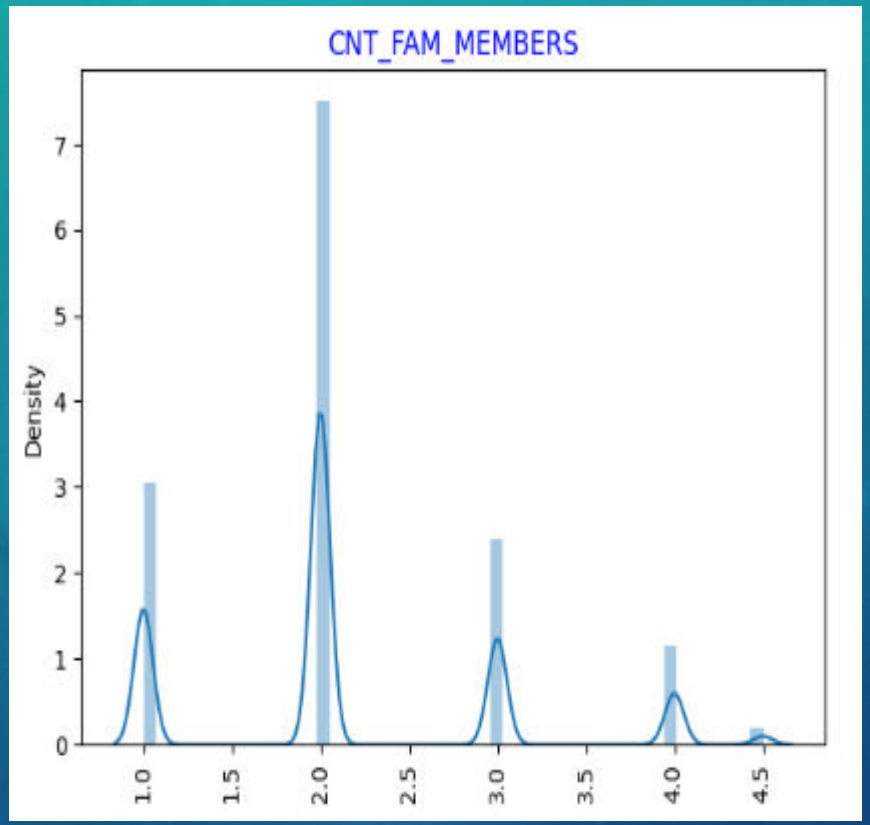


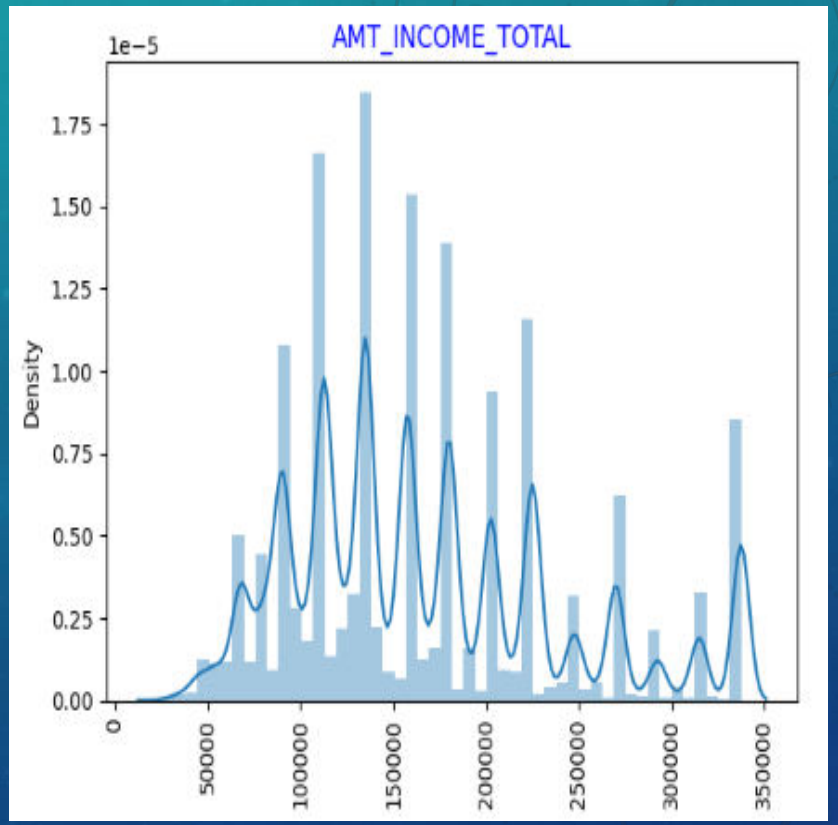
CNT_PAYMENT

YEARS_DECISION

AMT_CREDIT





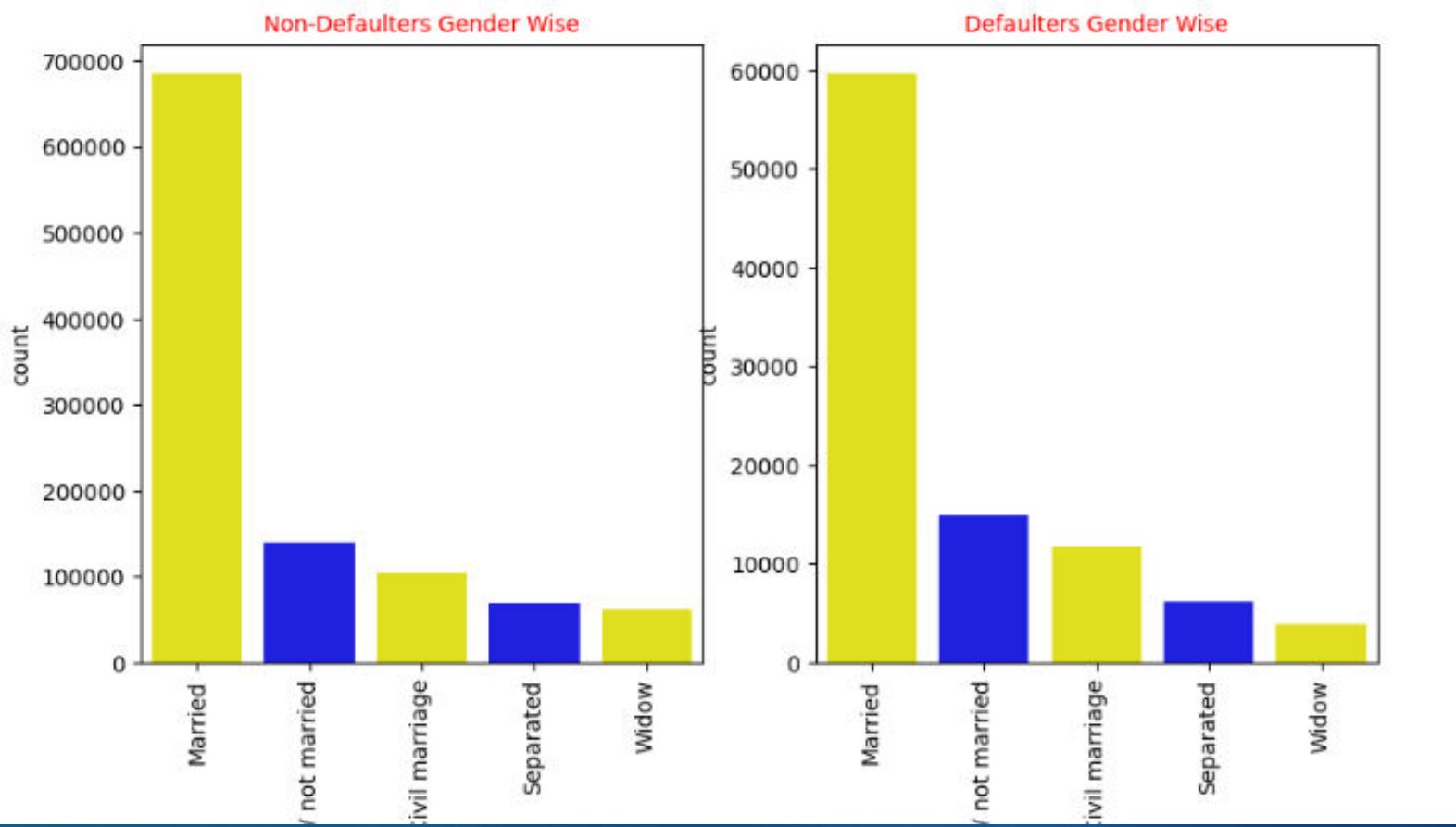


AMT_ANNUITY

CNT_FAM_MEMBERS

AMT_INCOME_TOTAL

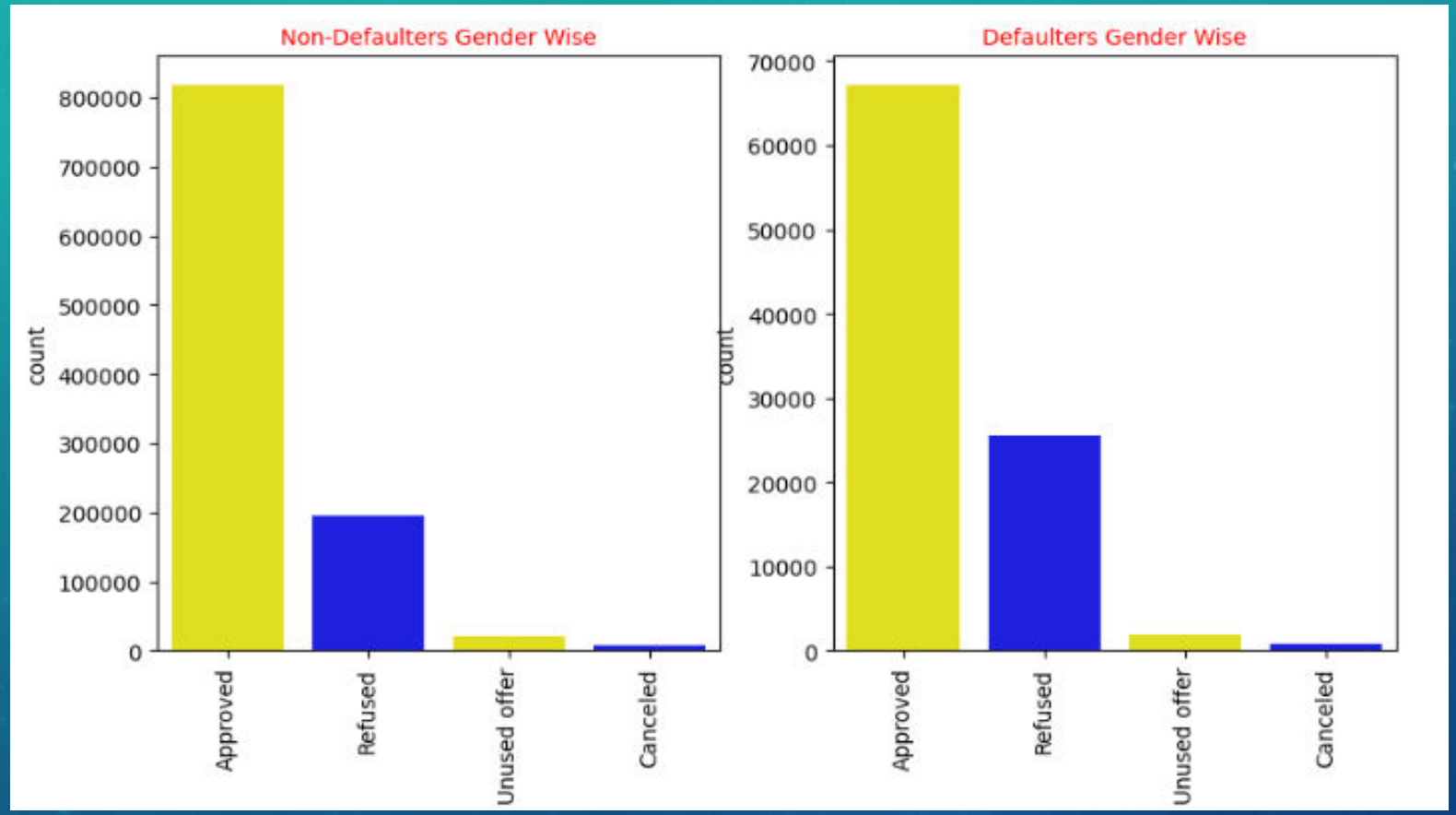
TARGET and CODE_GENDER



INSIGHTS:

1. For Cash loans, males are more in number than females and similarly with Revolving loans.

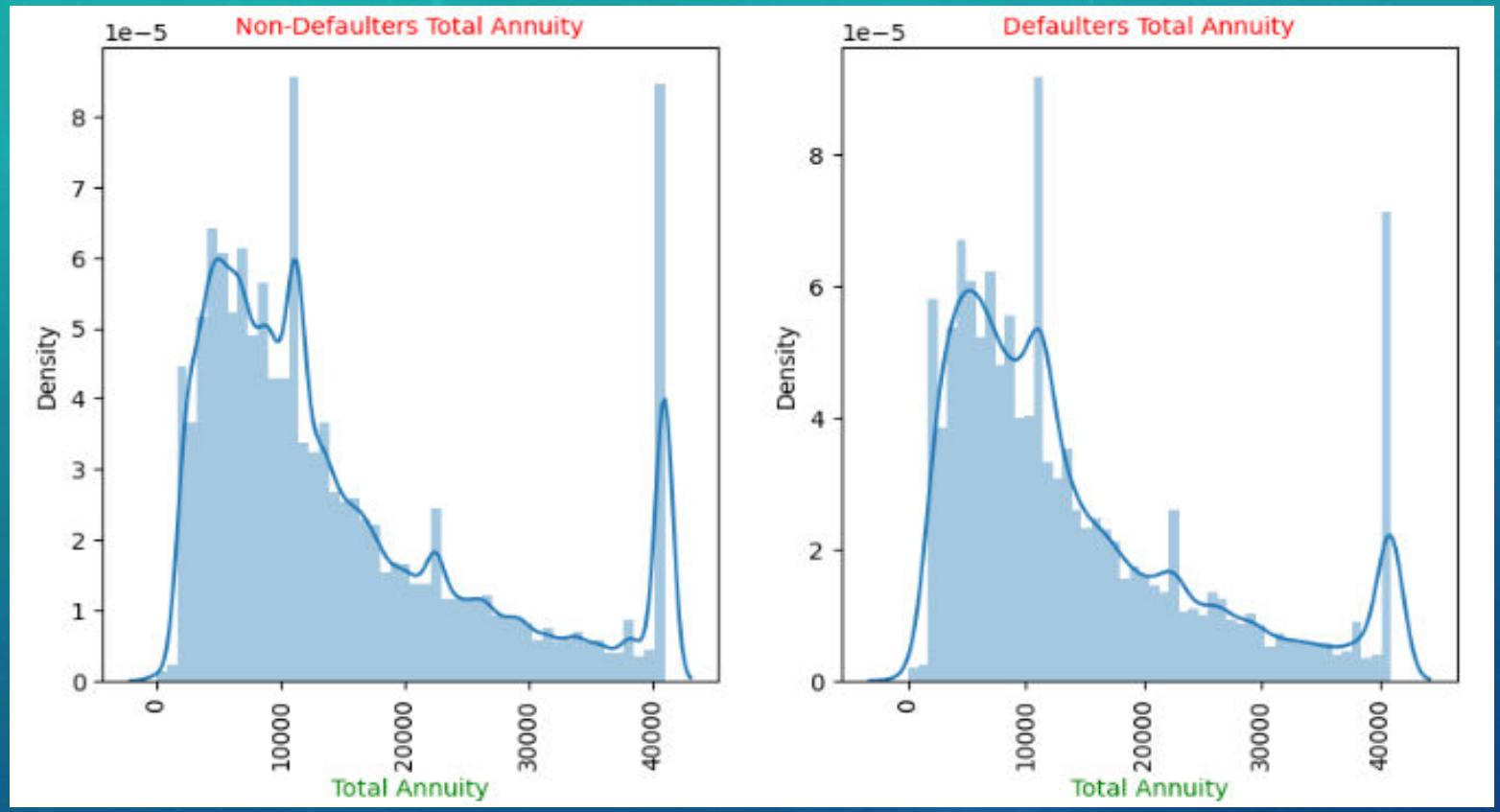
TARGET and CODE_GENDER



INSIGHTS:

1. Among Defaulters and Non-Defaulters, Approved loans are of high percentage.

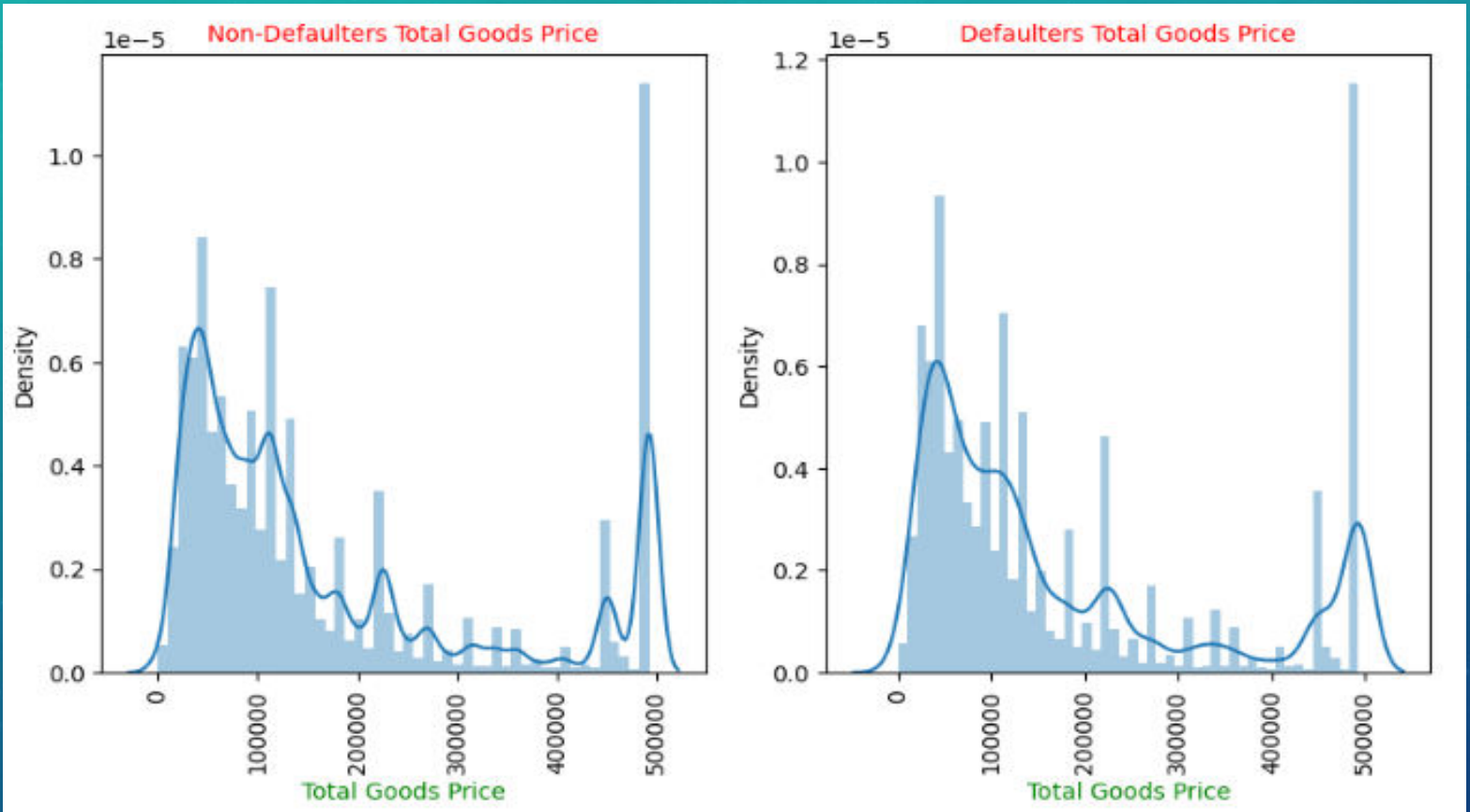
TARGET and AMT_ANNUITY



INSIGHTS:

1. Among Defaulters and Non-Defaulters, Total Annuity is more in between 5000 to 15000.

TARGET and AMT_TOTAL_CREDIT

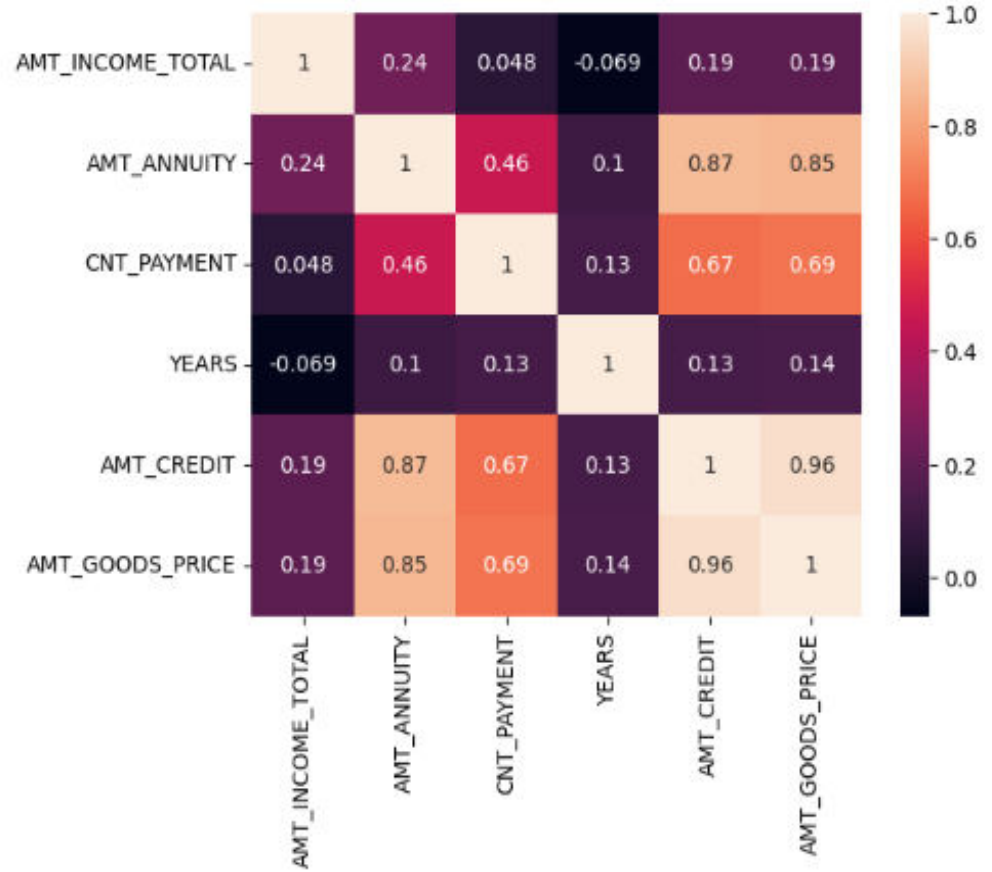


INSIGHTS:

1. Among Defaulters and Non-Defaulters, Total Goods Price is between 5000 to 10000.

BIVARIATE ANALYSIS

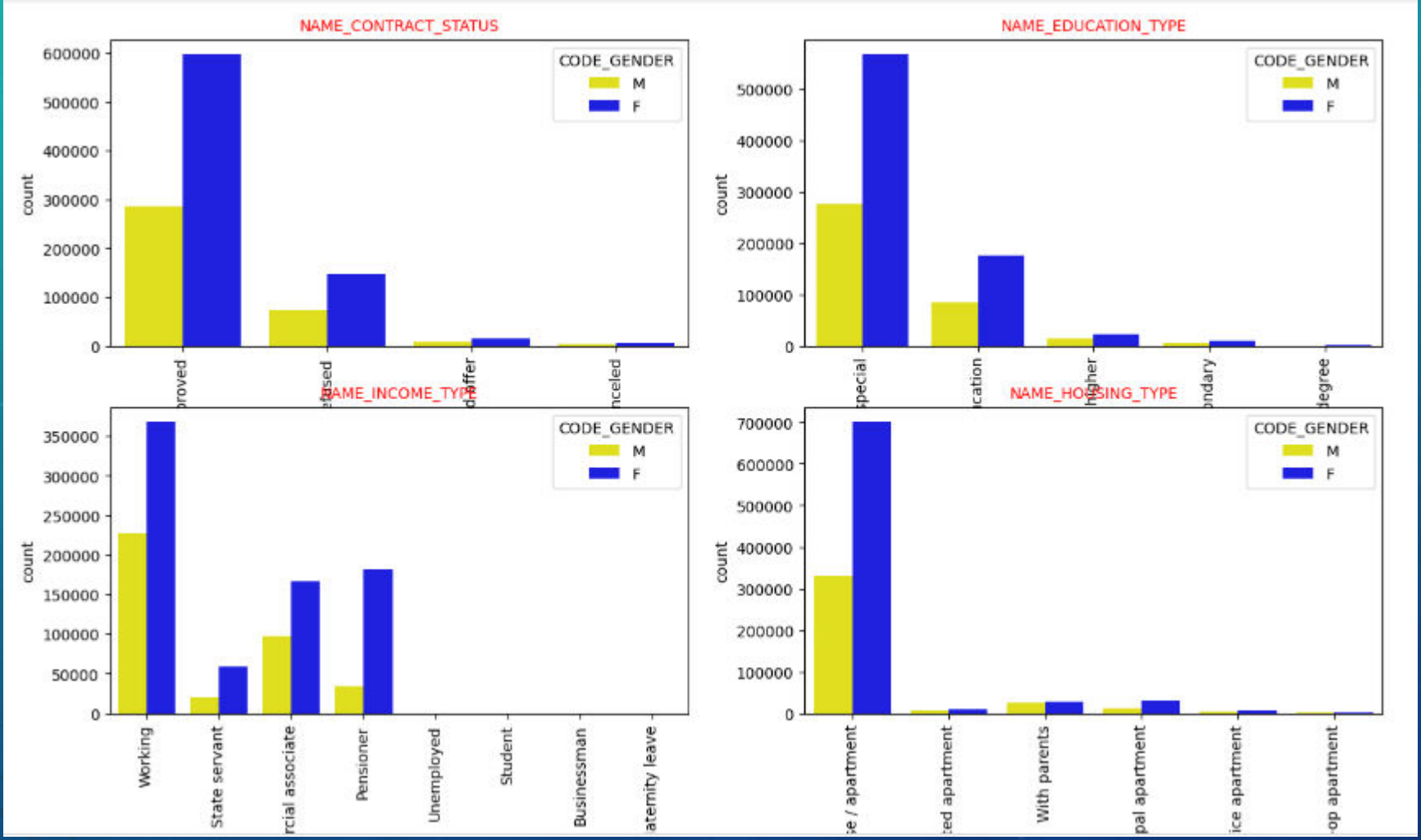
```
merged1 = merged[['AMT_INCOME_TOTAL','AMT_ANNUITY','CNT_PAYMENT','YEARS','AMT_CREDIT','AMT_GOODS_PRICE']]
sns.heatmap(merged1.corr(),annot=True)
plt.show()
```



INSIGHTS:

* 0.96% = AMT_CREDIT & AMT_GOODS_PRICE.
* 0.87% = AMT_CREDIT & AMT_ANNUITY.
* 0.85% = AMT_GOODS_PRICE & AMT_ANNUITY.

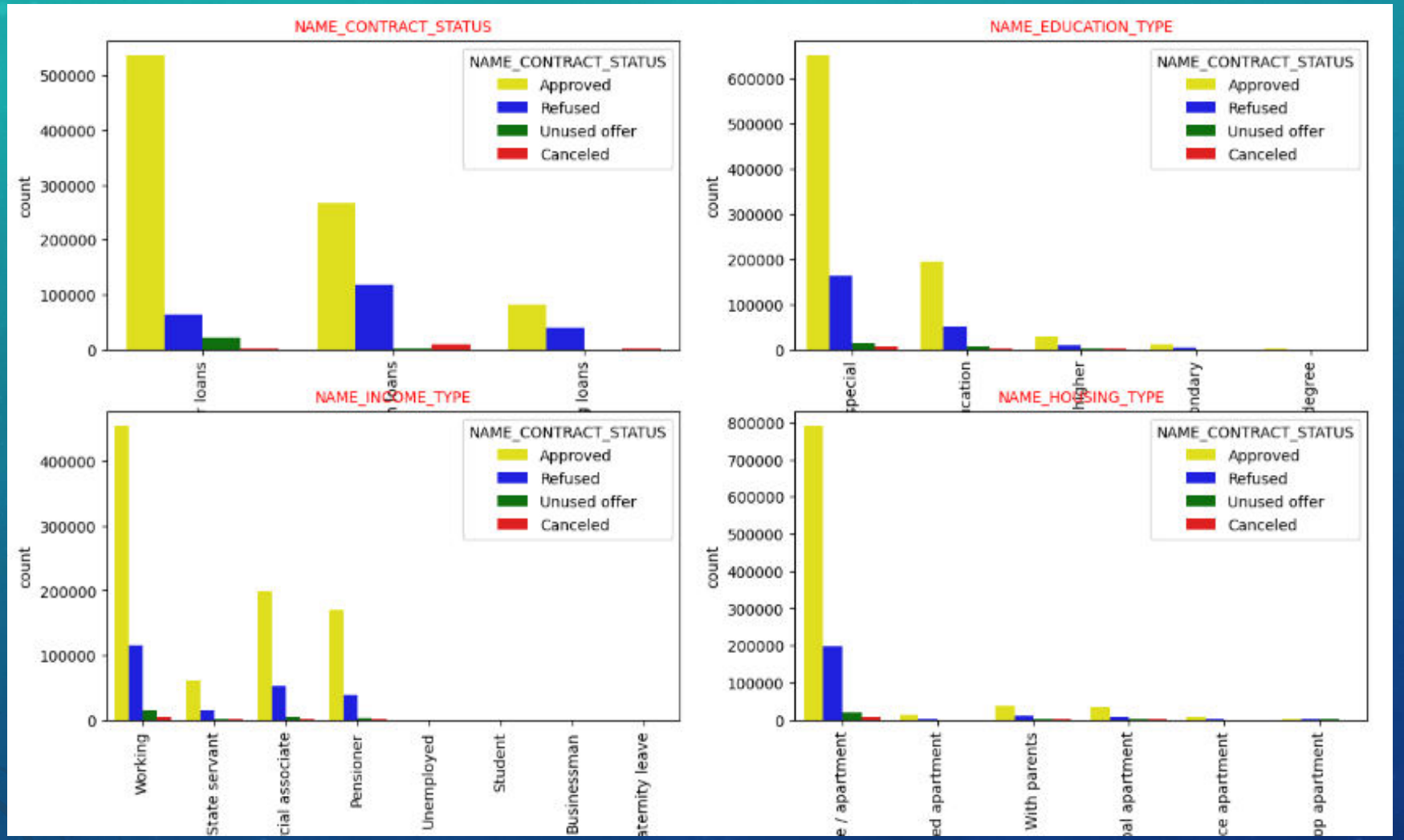
CODE_GENDER is Combined with NAME_CONTRACT_STATUS, NAME_EDUCATION_TYPE, NAME_INCOME_TYPE and NAME_HOUSING_TYPE



INSIGHTS:

Every where the females are high in number than males.

NAME_CONTRACT_STATUS is Combined with NAME_CONTRACT_STATUS, NAME_EDUCATION_TYPE, NAME_INCOME_TYPE and NAME_HOUSING_TYPE



INSIGHTS:

Every where the no of loans approved is higher.

CONCLUSION FOR DEFAULTERS

- People with Lower Secondary / Secondary education are likely to be Defaulters followed by higher education.
- The percentage of Cash Loans are higher than Revolving Loans that too males are greater in number.
- Maximum Defaulters don't have an own car.
- Maximum Defaulters of them do have an own house and it is either a house or apartment.
- Most of the Defaulters are working professionals followed by Commercials.
- Most of the Defaulters are Married.
- Defaulters comes under the age of 30-40 years.
- We can see that the lesser the credit amount of the loan, the more chances of being a defaulter.
- Total annuity lies from 20k to 30k.

CONCLUSION FOR DEFAULTERS

- Business Entity type 3 from Organization Type comes under Defaulters at a maximum rate.
- Males ratio is at relatively higher default rate than Females.
- We can see that males are earning more as well being the more defaulters.
- We can see that Unemployed people are more defaulters in both male and female case.
- Males are more unemployed than females.
- Maternity leave females are also in higher number in defaulters list.
- Male nos are more when compared with female in defaulters list.

CONCLUSION FOR NON-DEFAULTERS

- Rating with 1 in REGION_RATING_CLIENT column comes under Non-Defaulters.
- Student and Businessmen have no defaults.
- People above age of 50 have high probability of Non-Defaulters.
- Females are more in number for Non-Defaulters.
- The percentage of Cash Loans are higher than Revolving Loans that too males are greater in number.
- Maximum Non-Defaulters don't have an own car.
- Maximum Non-Defaulters of them do have an own house and it is either a house or apartment.

THANK YOU