CREDIT EDA ASSIGNMENT

PRESENTATION BY --

M N VAMSI KRISHNA

CONTENTS

- Problem Statement
- Work flow
- Importing libraries and warnings
- Reading datasets
- Outliers handling
- Univariate analysis
- Bivariate analysis
- Conclusion

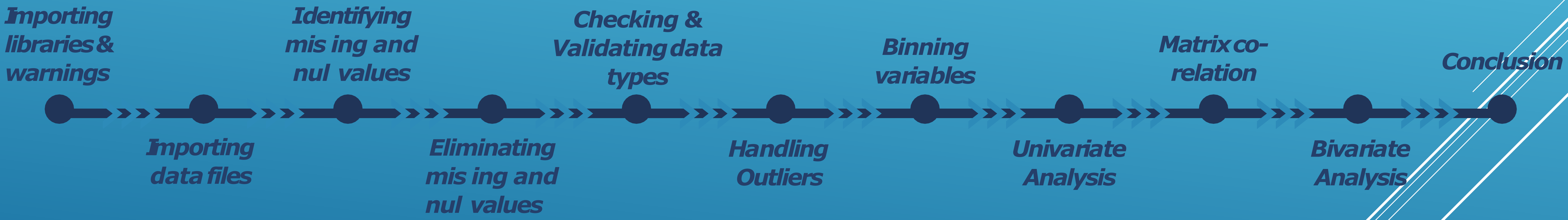
PROBLEM STATEMENT

AIM

Understand how the bank approves and refuses loan. Find out different patterns and represent the outcomes to help the bank reduce the credit risk and interest risk.

The two input files are extracted, cleaned/transformed and few columns are analyzed via different charts generated using different Python libraries. Then some inferences are made based on the outcomes.

WORK FLOW



Importing libraries & warnings

Importing data files

Identifying mis ing and nul values

Eliminating mis ing and nul values

Checking & Validating data types

Handling Outliers

Binning variables

Univariate Analysis

Matrix co-relation

Bivariate Analysis

Conclusion

IMPORTING LIBRARAIES AND WARKINGS

IMPORTING LIBRARIES:

```
IMPORT PANDAS AS PD
IMPORT NUMPY AS NP
IMPORT MATPLOTLIB.PYPLOT AS PLT
%MATPLOTLIB INLINE
IMPORT SEABORN AS SNS
IMPORT PLOTLY.EXPRESS AS PX
IMPORT WARNINGS
```

Imported warnings

Highlights warnings however the program runs.

READING DATASET

- Two data files were extracted from the given dataset. namely - 'application_data.csv' and 'previous_data.csv'
- Highlighted datafile description, shape etc., in the notebook for elaborated experience in reading the data.

HANDLING DATA, NULL & MISSING VALUES

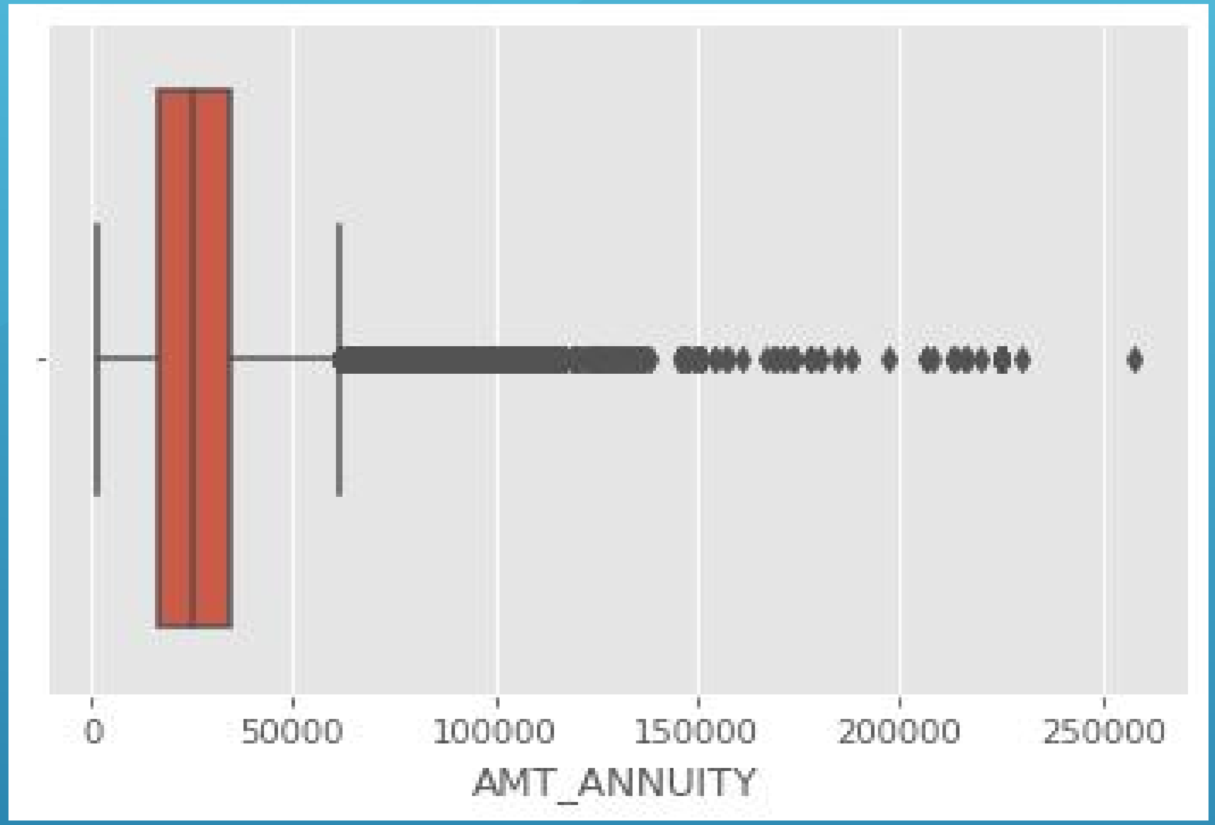
- Checked for null values in aapplication_data.csv and eliminated 49 columns which had null values more than 40%

- Post that, AMT_ANNUITY, AMT_GOODS_PRICE, EXT_SOURCE_2, NAME_TYPE_SUITE,
- had less than 1%of null (& numeric) values. Hence, identified outliers and imputed using the best approach available.

- checked for unique values in columns by the following condition:
  - If the count of unique values <=40, it's a categorical column
  - If the count of unique values >50, it's a continuous column

- AMT_ANNUITY variable

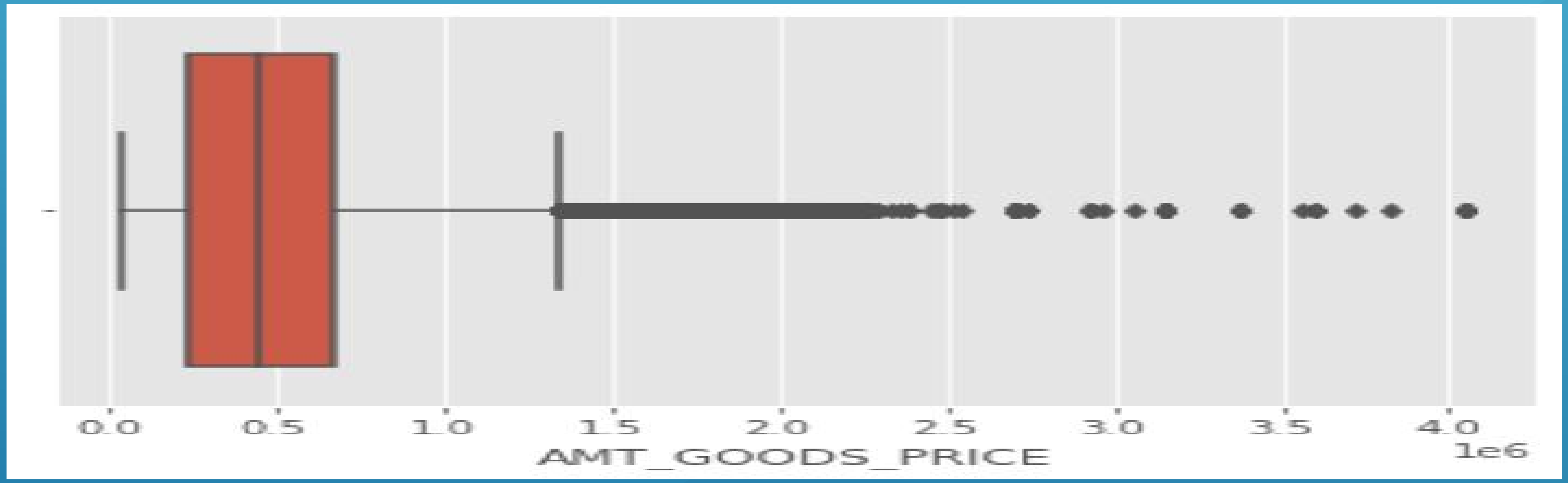
As seen here, outlier is present at 258025. Hence to impute the outlier values, we will use median here.



OUTLIERSHANDLING

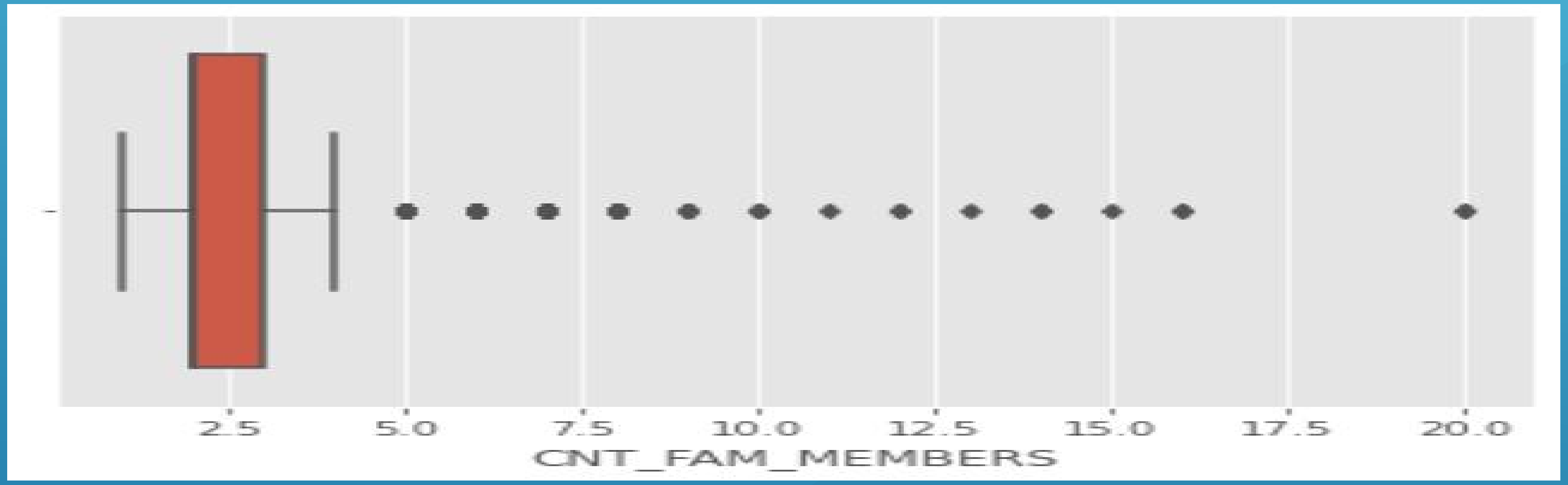
- AMT_GOODS_PRICE variable

The outliers here are present after the 99th quantile.



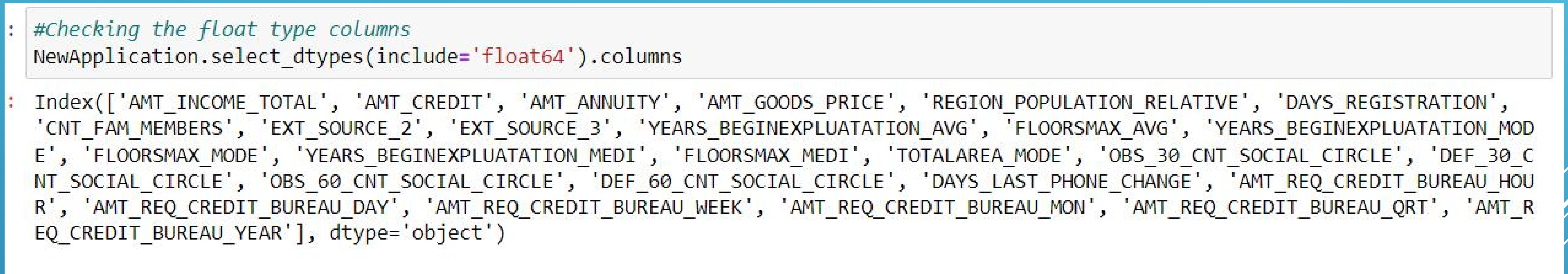
OUTLIERSHANDLING

- **CNT_FAM_MEMBERS**



OUTLIERSHANDLING

check datatypes of columns and modify them appropriately



```
#Checking the float type columns
NewApplication.select_dtypes(include='float64').columns
```

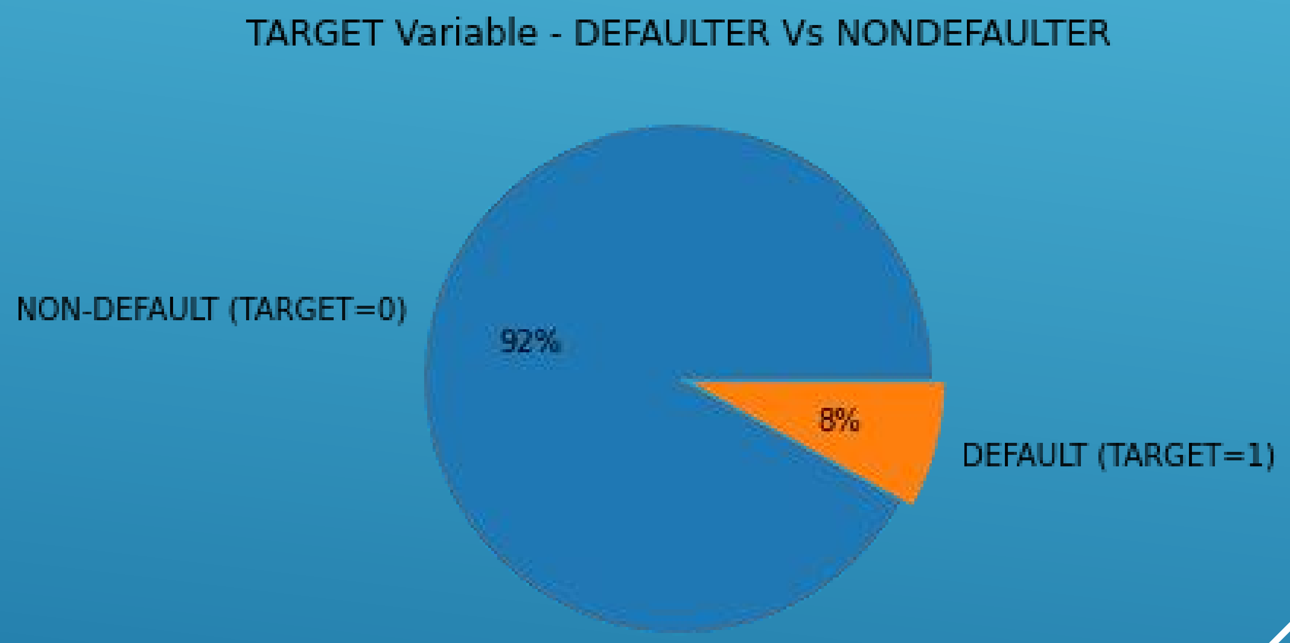
```
Index(['AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'REGION_POPULATION_RELATIVE', 'DAYS_REGISTRATION',
'CNT_FAM_MEMBERS', 'EXT_SOURCE_2', 'EXT_SOURCE_3', 'YEARS_BEGINEXPLUATATION_AVG', 'FLOORSMAX_AVG', 'YEARS_BEGINEXPLUATATION_MOD
E', 'FLOORSMAX_MODE', 'YEARS_BEGINEXPLUATATION_MEDI', 'FLOORSMAX_MEDI', 'TOTALAREA_MODE', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_C
NT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE', 'DAYS_LAST_PHONE_CHANGE', 'AMT_REQ_CREDIT_BUREAU_HOU
R', 'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT', 'AMT_R
EQ_CREDIT_BUREAU_YEAR'], dtype='object')
```

Splitting the dataframe into two separate dfs

Dataframe with all the data related to non-defaulters

Dataframe with all the data related to defaulters

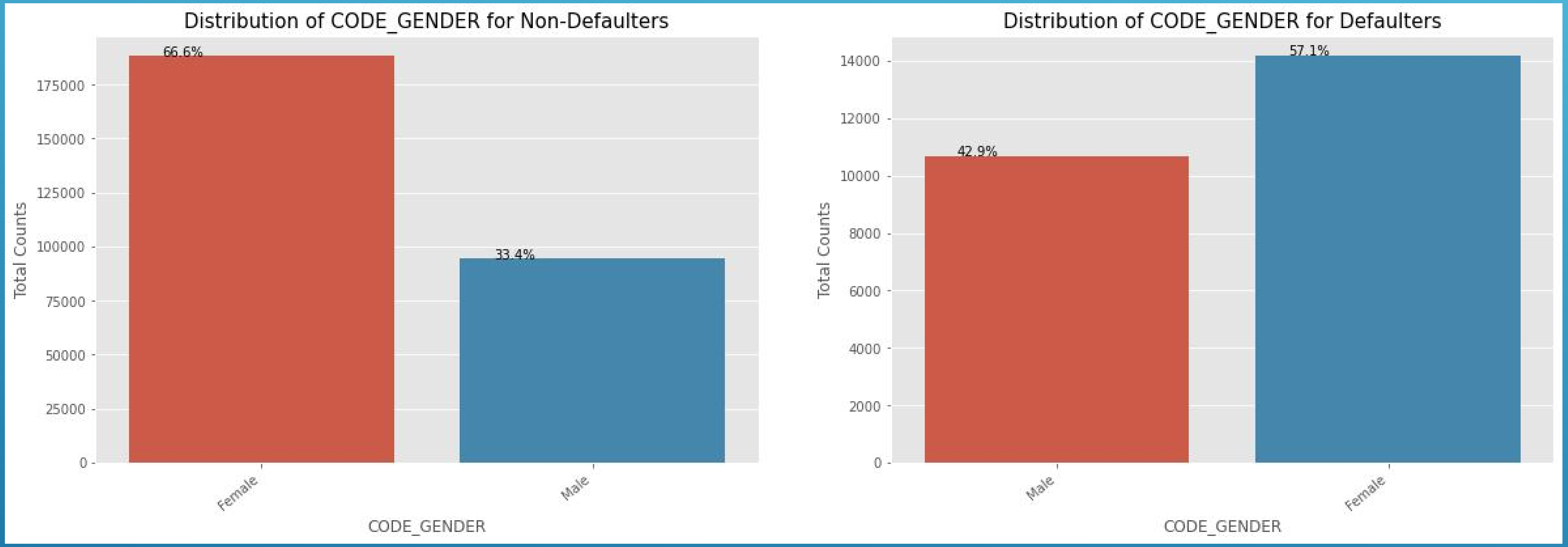
Checking for imbalance in Target



UNIVARIATE CATEGORICAL ORDERED ANALYSIS :

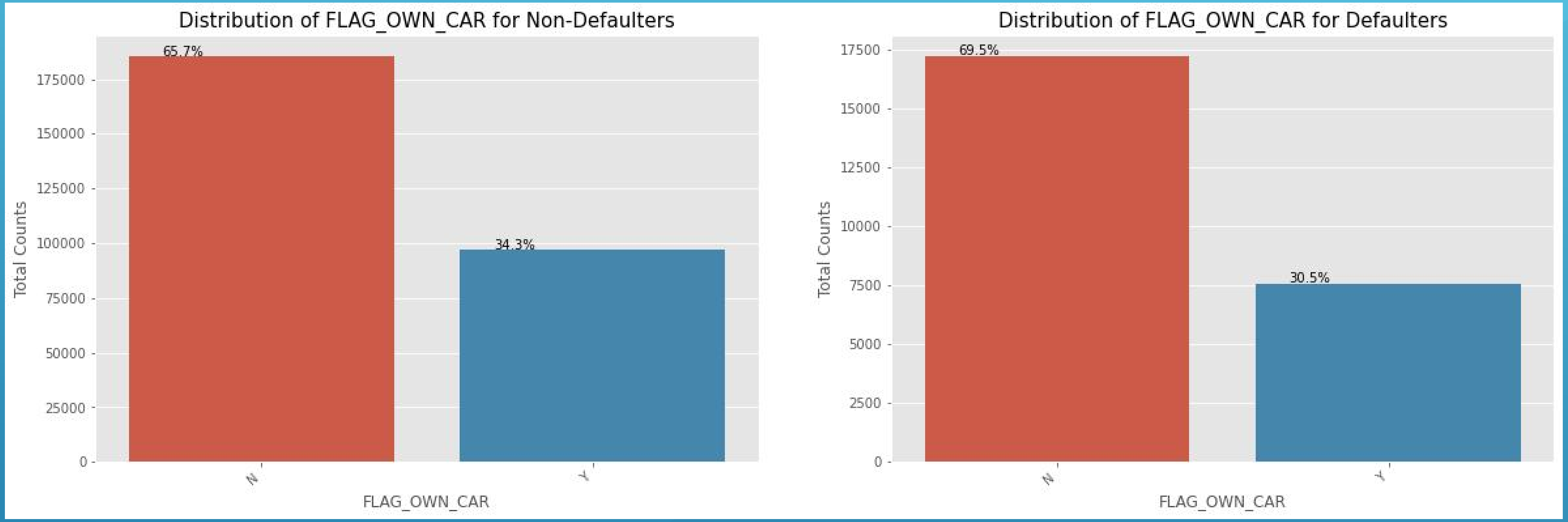
Adding the normalized percentage for easier comparision between defaulter and non-defaulte

Adding the normalized percentage for easier comparision between defaulter and non-defaulterr



We can see that Female contribute 67% to the non-defaulters while 57% to the defaulters. We can conclude that We see more female applying for loans than males and hence the more number of female defaulters as well.
But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts.

UNIVARIATE CATEGORICAL ORDERED ANALYSIS

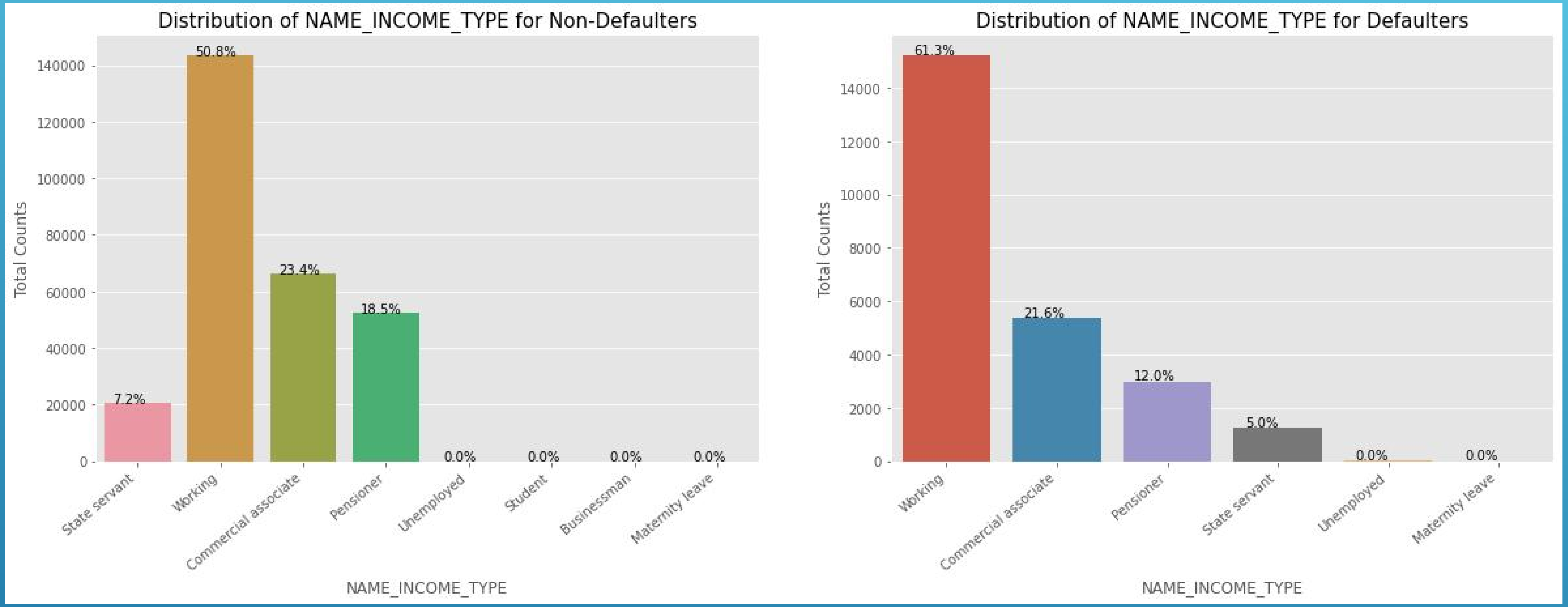


We can see that people with cars contribute 65.7% to the non-defaulters while 69.5% to the defaulters. We can conclude that

While people who have car default more often, the reason could be there are simply more people without cars

Looking at the percentages in both the charts, we can conclude that the rate of default of people having car is low compared to people who don't.

UNIVARIATE CATEGORICAL ORDERED ANALYSIS



Distribution of NAME_INCOME_TYPE for Non-Defaulters

Distribution of NAME_INCOME_TYPE for Defaulters

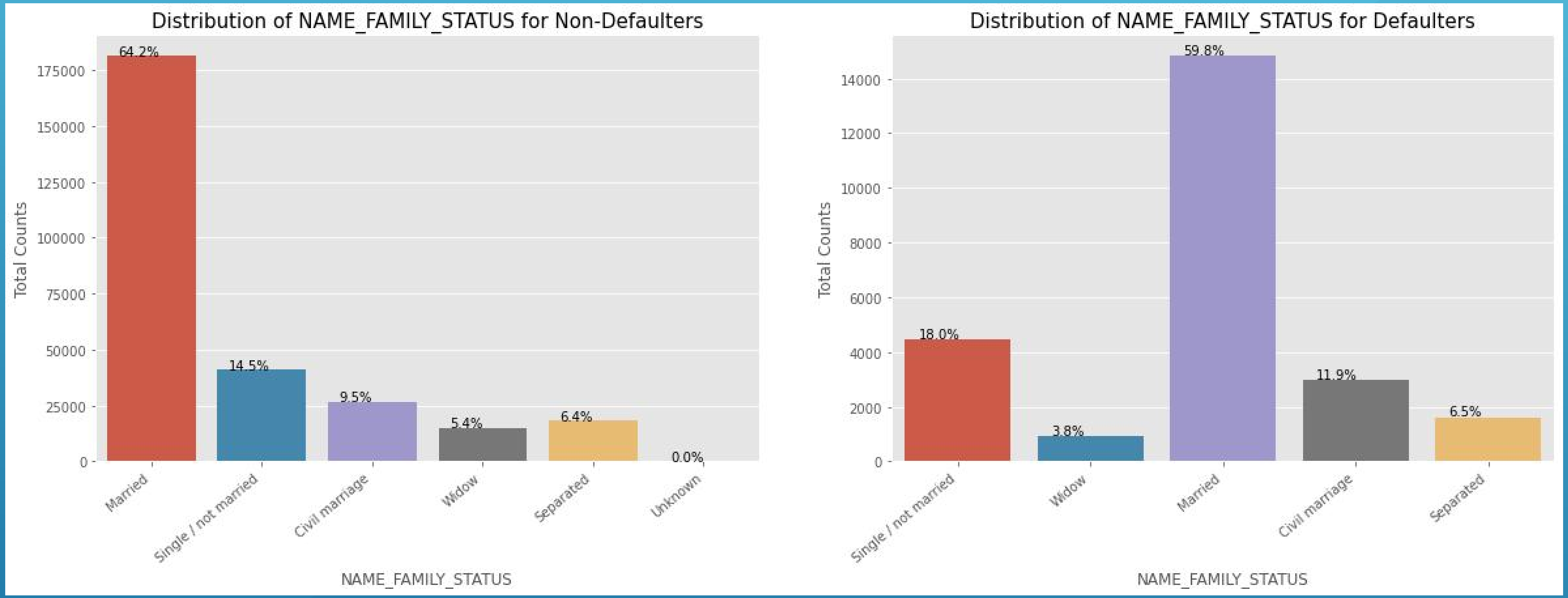
We can notice that the students don't default. The reason could be they are not required to pay during the time they are students.

We can also see that the BusinessMen never default.

Most of the loans are distributed to working class people

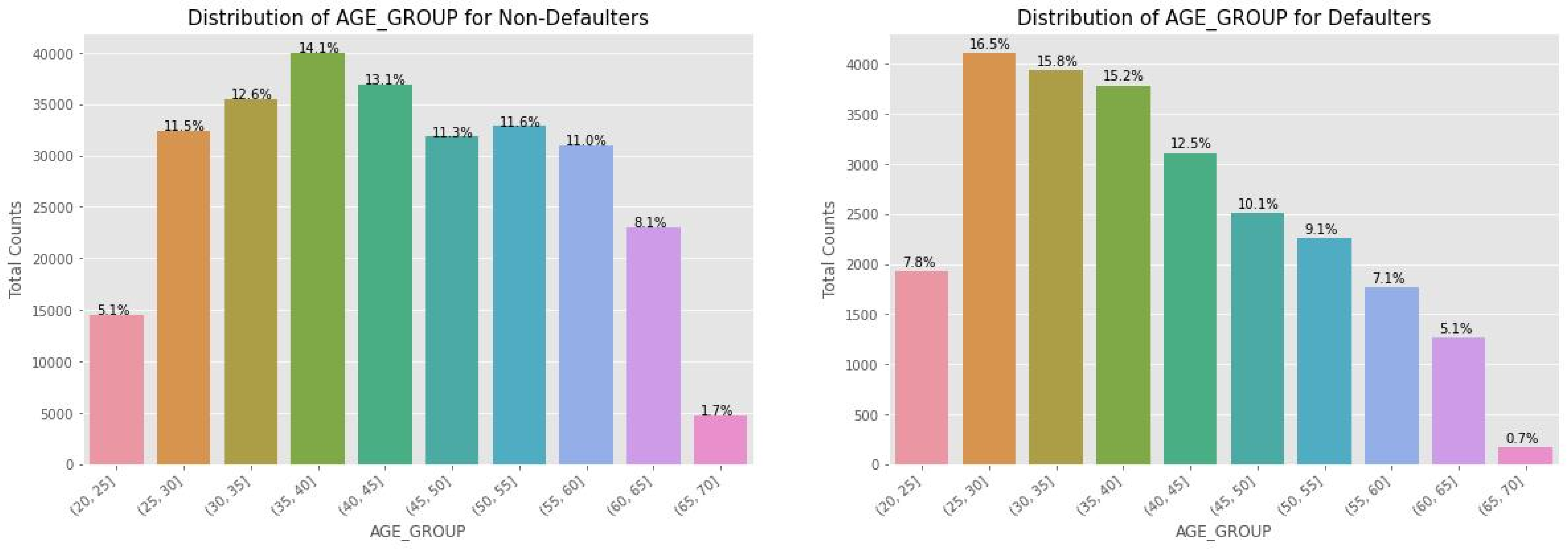
We also see that working class people contribute 51% to non defaulters while they contribute to 61% of the defaulters. Clearly, the chances of defaulting are more in their case.

UNIVARIATE CATEGORICAL ORDERED ANALYSIS



Married people tend to apply for more loans comparatively.
But from the graph we see that Single/non Married people contribute 14.5% to Non Defaulters and 18% to the defaulters. So there is more risk associated with them.

Univariate Categorical Ordered Analysis

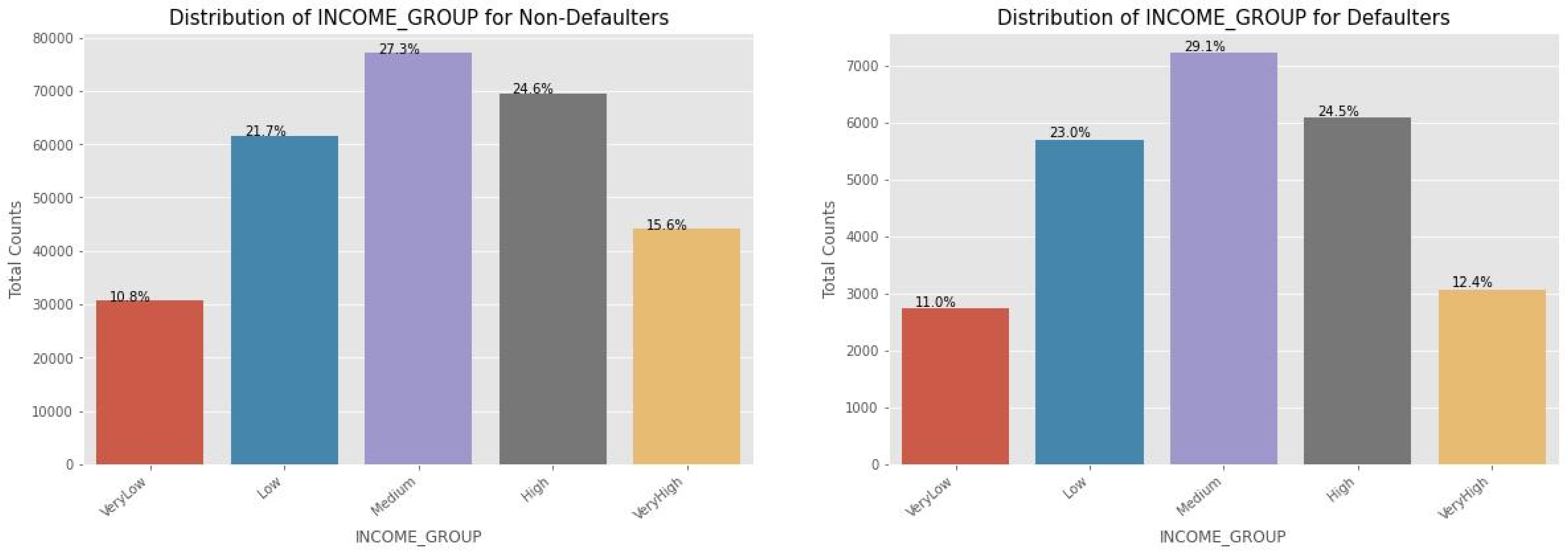


Distribution of AGE_GROUP for Non-Defaulters

Distribution of AGE_GROUP for Defaulters

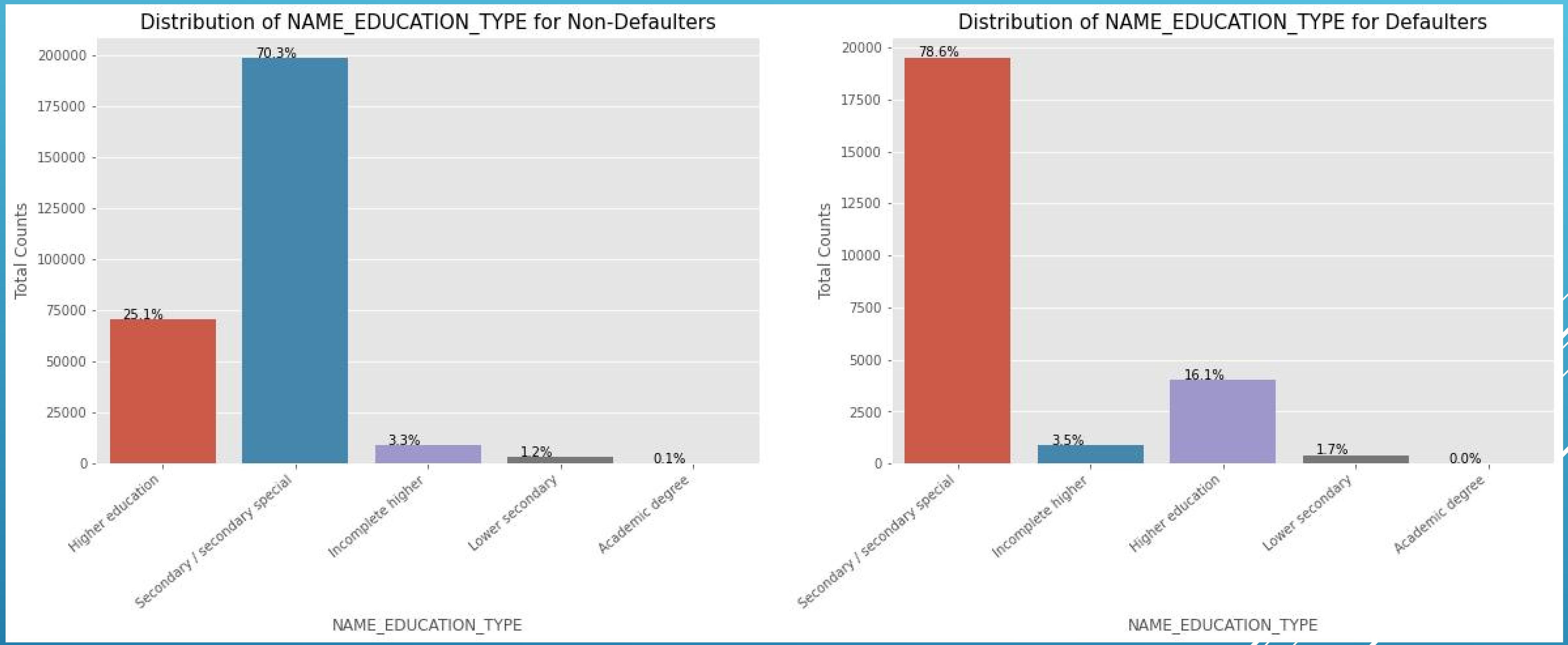
We see that (25,30] age group tend to default more often. So they are the riskiest people to loan to.
With increasing age group, people tend to default less starting from the age 25. One of the reasons could be they get employed around that age and with increasing age, their salary also increases.

Univariate Categorical Ordered Analysis



The Very High income group tend to default less often. They contribute 12.4% to the total number of defaulters, while they contribute 15.6% to the Non-Defaulters.

Univariate Categorical Ordered Analysis

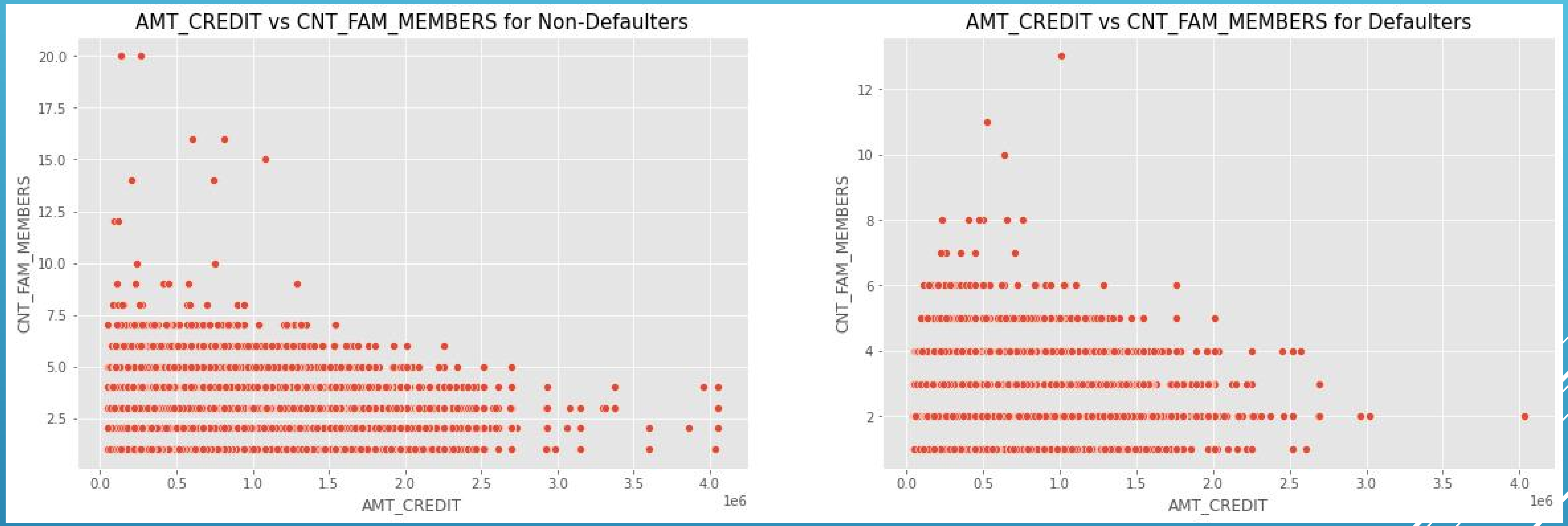


Distribution of NAME_EDUCATION_TYPE for Non-Defaulters

Distribution of NAME_EDUCATION_TYPE for Defaulters

Almost all of the Education categories are equally likely to default except for the higher educated ones who are less likely to default and secondary educated people are more likely to default

BIVARIATE ANALYSIS - NUMERIC - NUMERIC

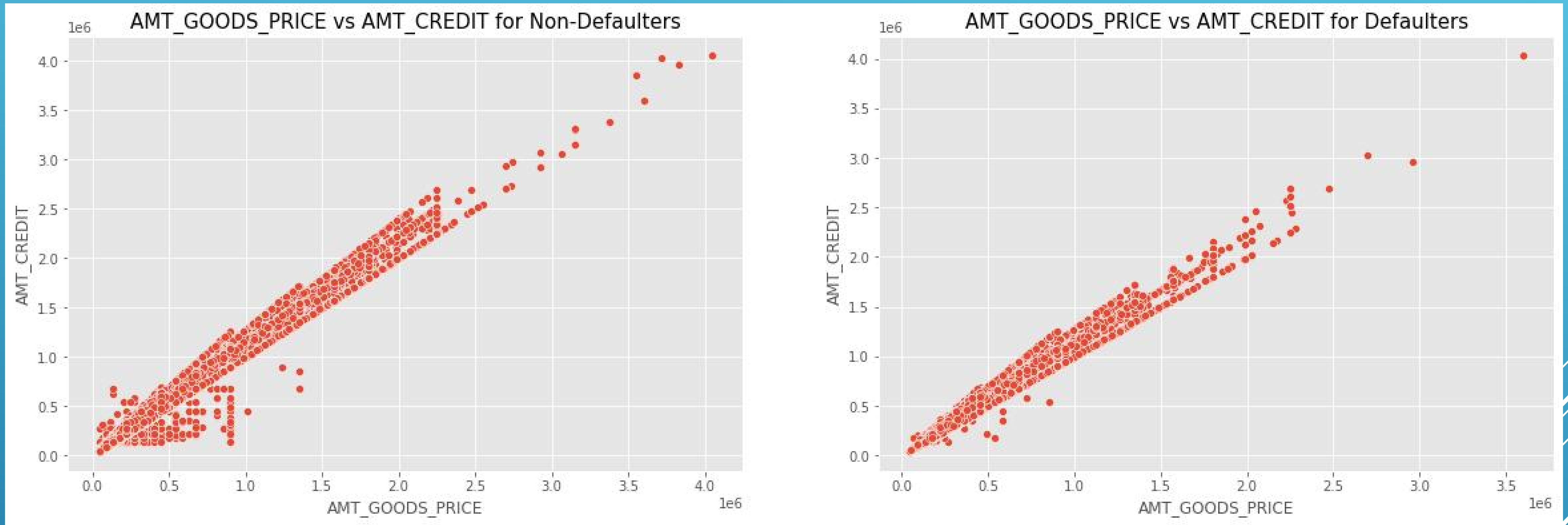


AMT_CREDIT vs CNT_FAM_MEMBERS for Non-Defaulters

AMT_CREDIT vs CNT_FAM_MEMBERS for Defaulters

We can see that the density in the lower left corner is similar in both the case, so the people are equally likely to default if the family is small and the AMT_CREDIT is low. We can observe that larger families and people with larger AMT_CREDIT default less often

BIVARIATE ANALYSIS - NUMERIC - NUMERIC



AMT_GOODS_PRICE vs AMT_CREDIT for Non-Defaulters

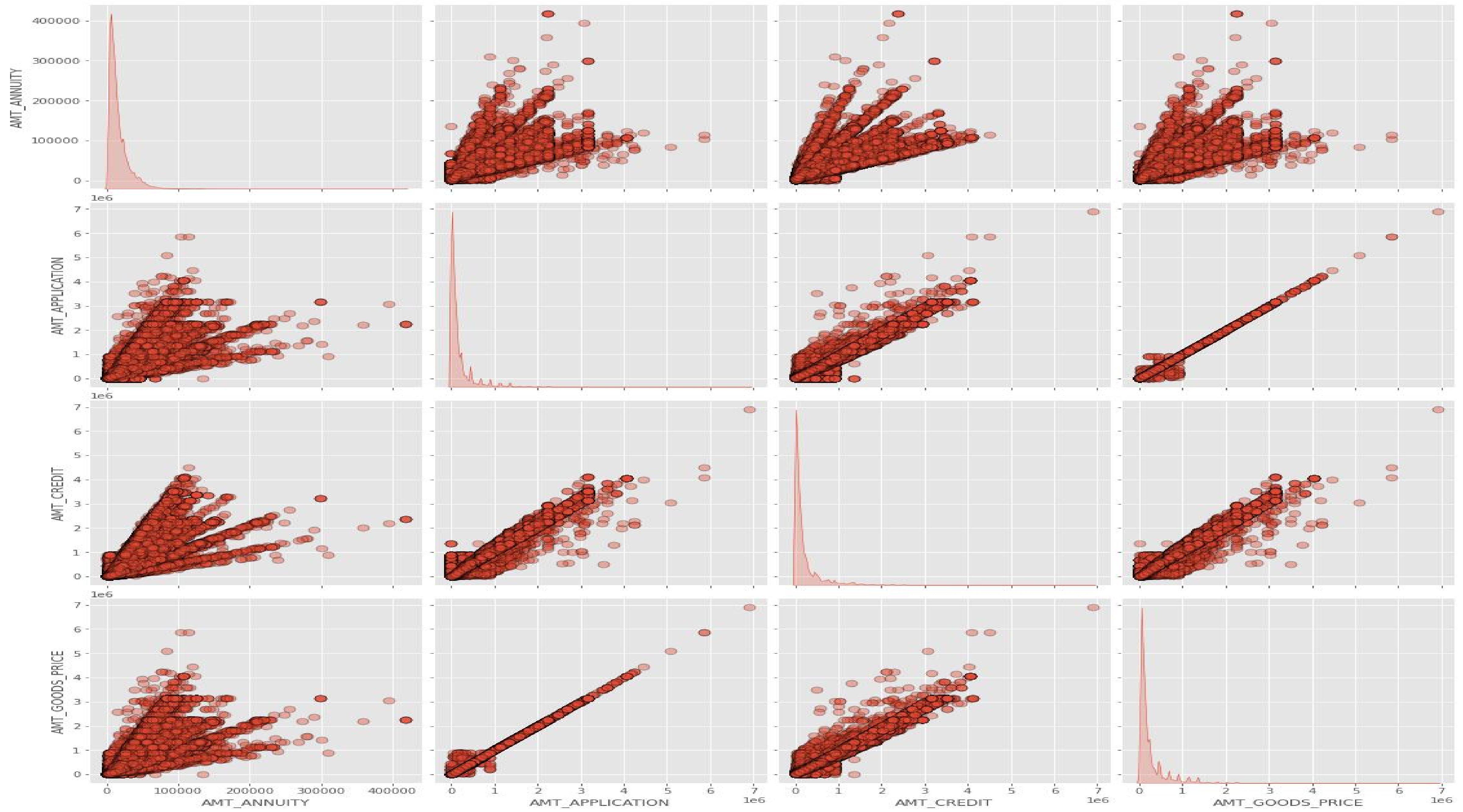
AMT_GOODS_PRICE vs AMT_CREDIT for Defaulters

Using pairplot to perform bivariate analysis on numerical columns

Annuity of previous application has a very high and positive influence over: (Increase of annuity increases below factors)
(1) How much credit did client asked on the previous application
(2)Final credit amount on the previous application that was approved by the bank
(3) Goods price of good that client asked for on the previous application.

For how much credit did client ask on the previous application is highly influenced by the Goods price of good that client has asked for on the previous application

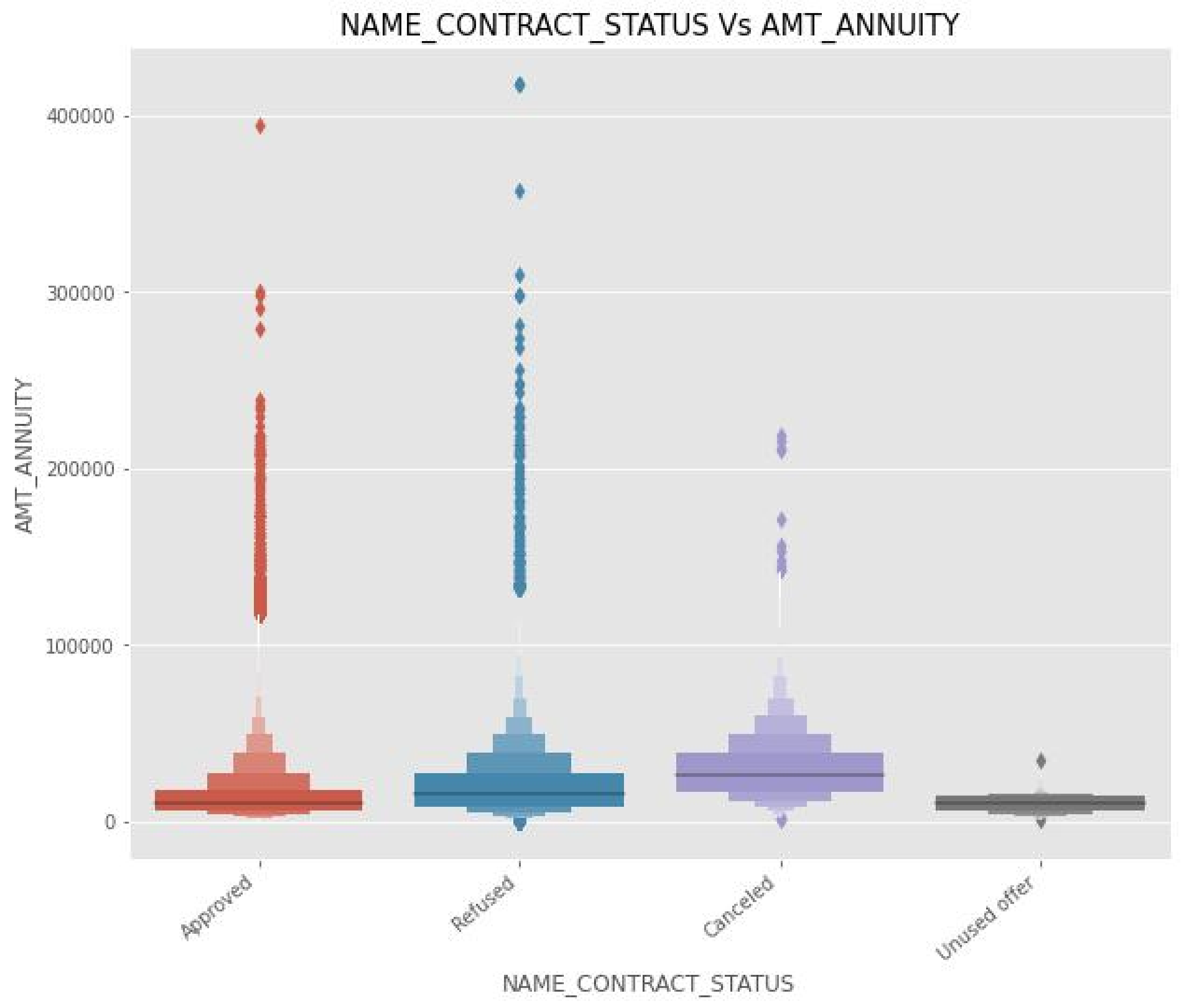
Final credit amount disbursed to the customer previously, after approval is highly influence by the application amount and also the goods price of good that client asked for on the previous application.



Using box plot to do some more bivariate analysis on categorical vs numeric columns

by-varient analysis of Contract status and Annuity of previous application

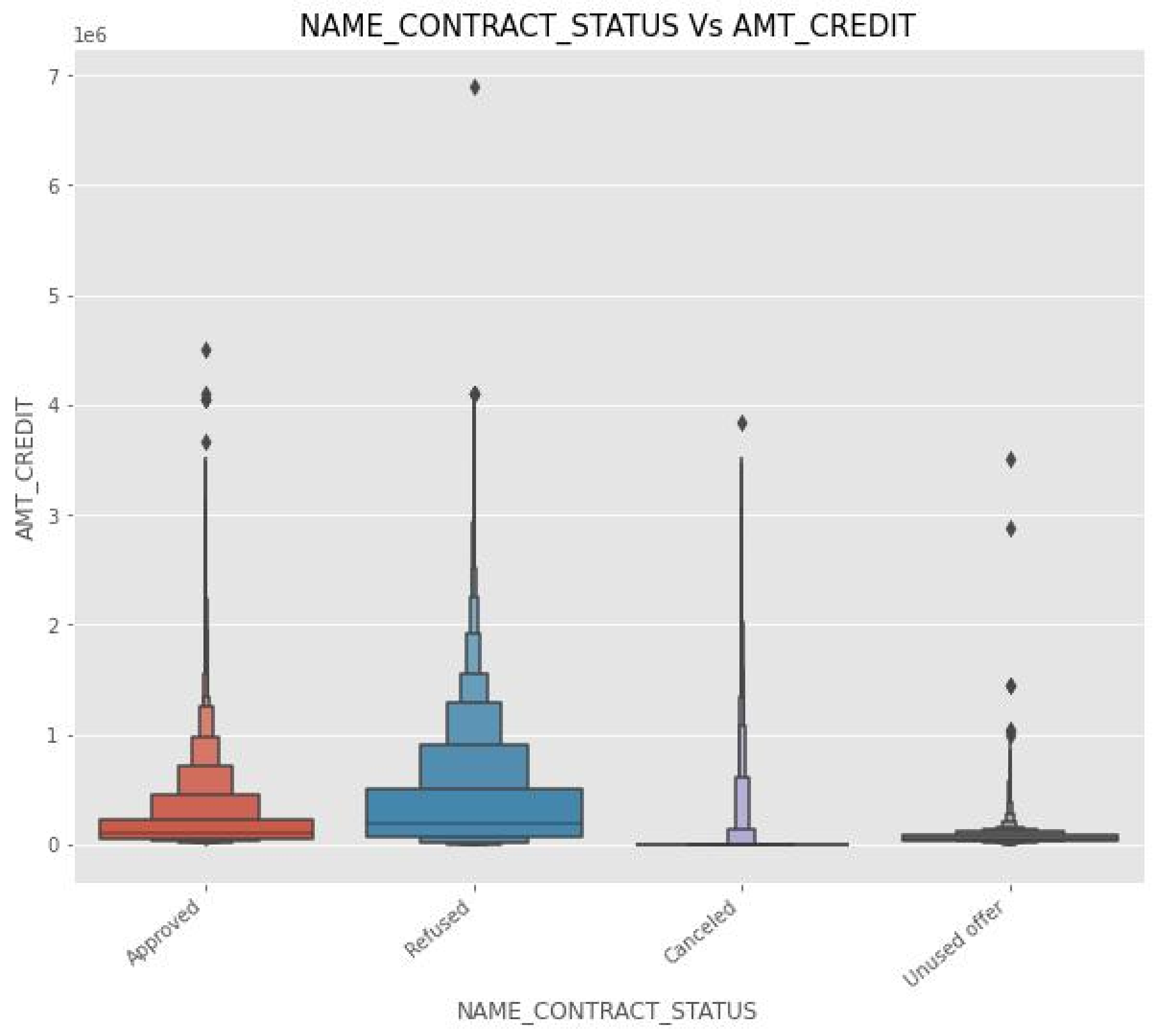
From the above plot we can see that loan application for people with lower AMT_ANNUITY gets canceled or Unused most of the time.
We also see that applications with too high AMT ANNUITY also got refused more often than others.



Using box plot to do some more bivariate analysis on categorical vs numeric columns

by-varient analysis of Contract status and Final credit amount disbursed to the customer previously, after approval

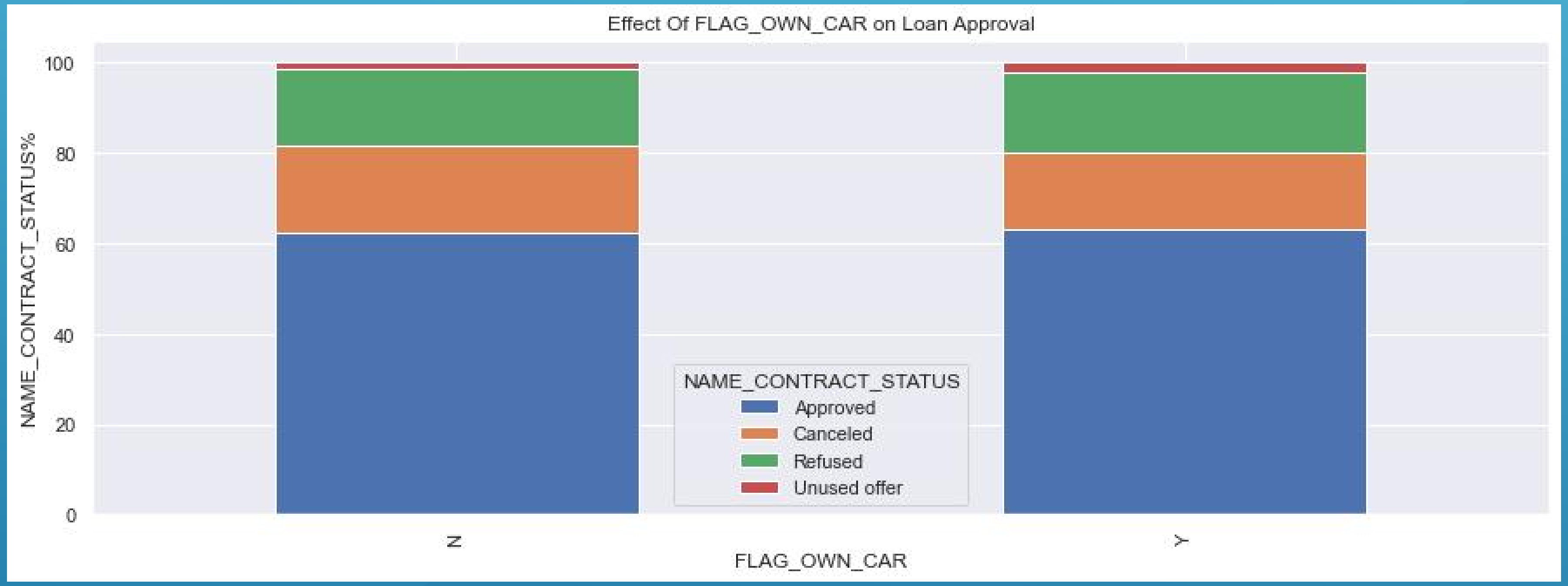
We can infer that when the AMT_CREDIT is too low, it get's cancelled/unused most of the time..



MERGING THE FILES AND ANALYZING THE DATA

Merging the two files to do some analysis

NewLeftPrev = pd.merge(NewApplication_Final, PreviousApplication, how='left', on=['SK_ID_CURR']).

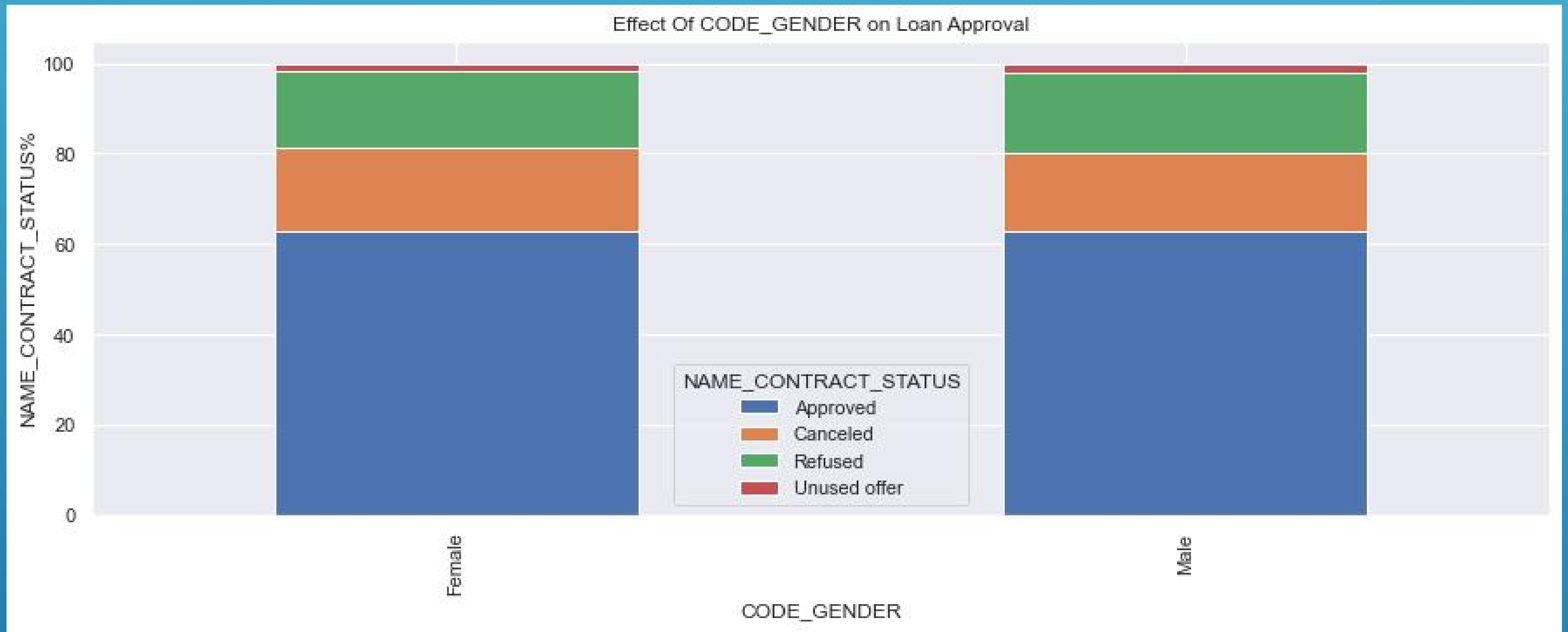


We see that car ownership doesn't have any effect on application approval or rejection. But we saw earlier that the people who has a car has lesser chances of default. The bank can add more weightage to car ownership while approving a loan amount

MERGING THE FILES AND ANALYZING THE DATA

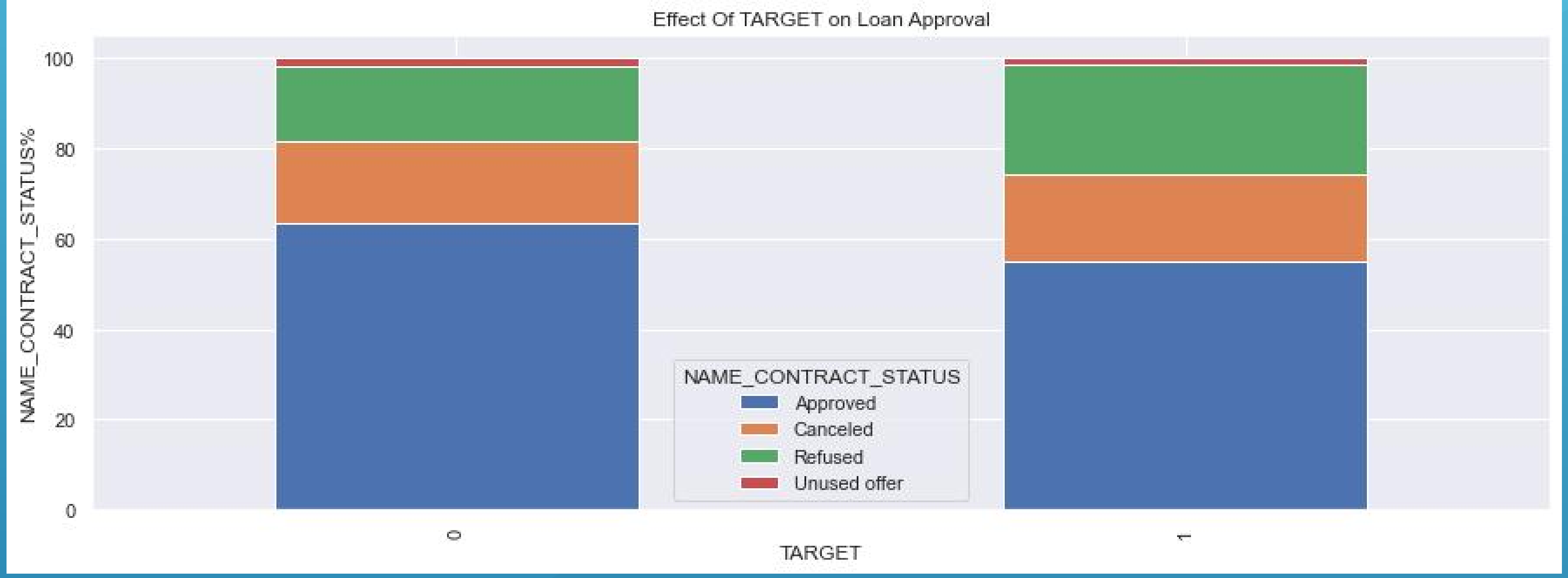
Merging the two files to do some analysis

```
NewLeftPrev = pd.merge(NewApplication_Final, PreviousApplication, how='left', on=['SK_ID_CURR']).
```



We see that code gender doesn't have any effect on application approval or rejection.
But we saw earlier that female have lesser chances of default compared to males. The bank can add more weightage to female while approving a loan amount.

CONCLUSION:



We can see that the people who were approved for a loan earlier, defaulted less often where as people who were refused a loan earlier have higher chances of defaulting.

THANK YOU