CREDIT EDA CASE STUDY

By

Manas Shirsat & Nikhil Jojen

DATA IMBALANCE

Imbalance TARGET: NON-DEFAULTER Vs DEFAULTER
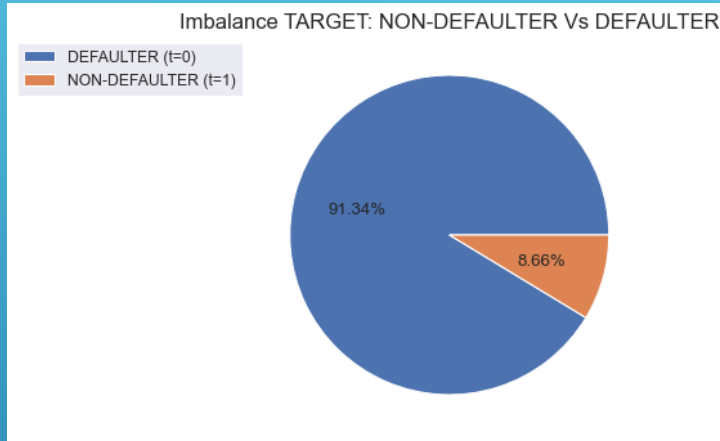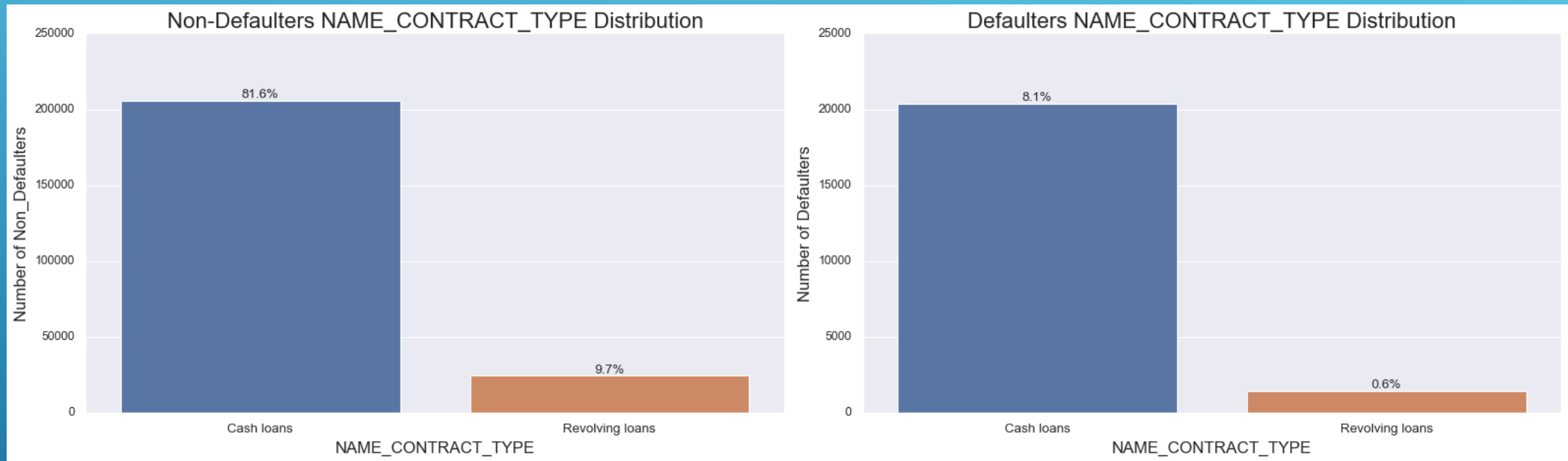
The Data after clean-up is highly imbalanced, around 8.66% of the data for loan 'Defaulters', whereas 91.34% of the data for 'Non-Defaulters'.

Imbalance in CODE_GENDER
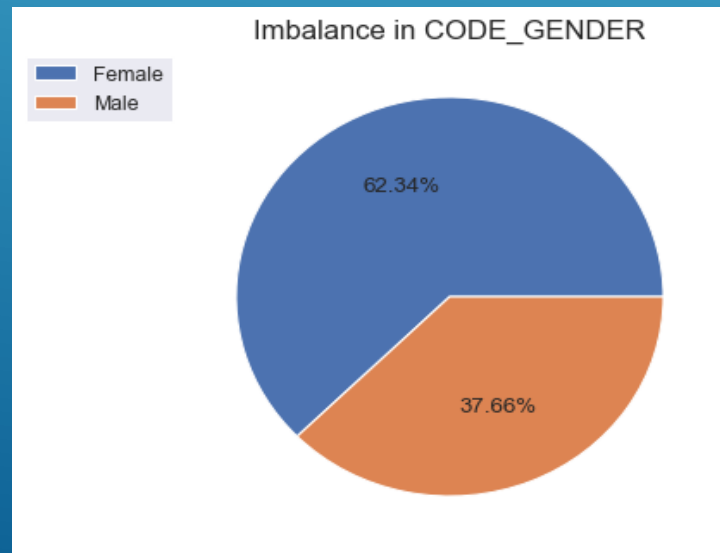
62.34% of the data provided are for Females, and 37.66% of the data are for Males.

Univariate Analysis on Categorical Variables

1. Contract Type

- *Around 89% of the total loans are from 'Cash Loans' and only 10% are 'Revolving Loans'.*
- *8% of the Total Cash Loans were defaulted whereas only 0.6% of the Total Revolving Loans were defaulted.*
- *Default percentage for Revolving loans are very low when compared to that of Cash Loans, however, the frequency of Revolving loans is very less when compared with the Cash Loans.*

2. Gender
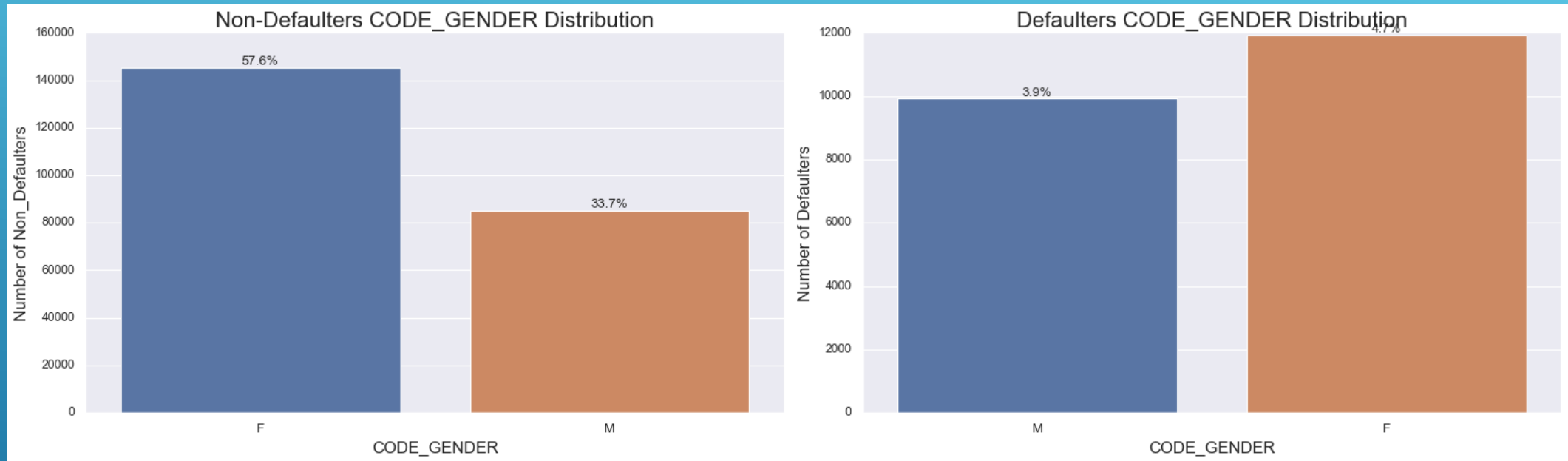
- Around 62% of the total customers are Females, and around 37% are Males.
- 4.7% of the Females and 3.9% of the Males out of all the customers are defaulters.
- Among the Defaulters, the Males are more likely to be a defaulter.
- Hence giving loans to Females is less risky and should be preferred candidates.

3. Family Status
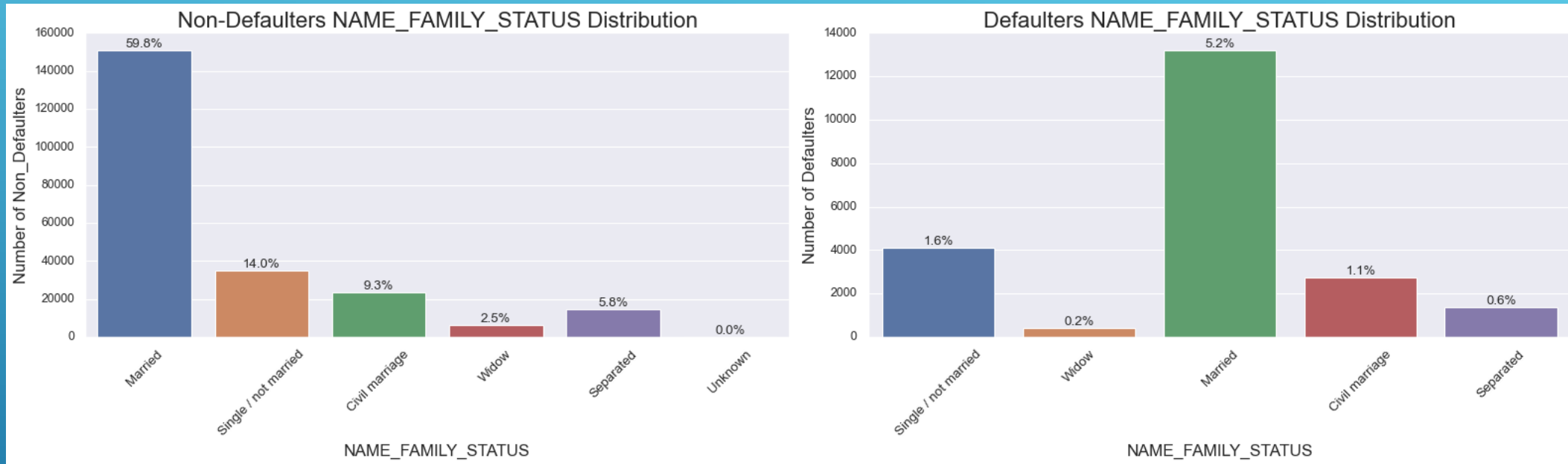
- *Around 65% of the total customers are Married out of which only 5% are defaulters. Hence, they can be preferred for giving loans.*
- *Most defaults are done by Single and Civil Marriage customers.*
- *Widows are less likely to default with only 0.2%, however they constitute only 2.5% of the total population.*

4. Education Type
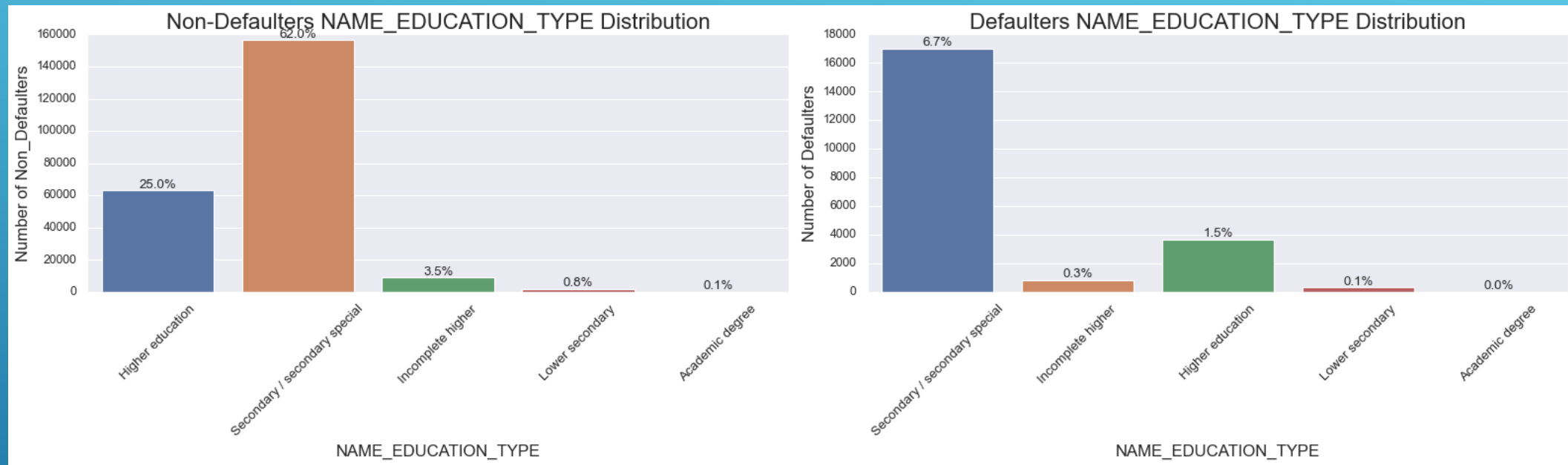
- *Majority of the customers have Secondary Education(which is around 68% of the total population), followed by Higher Education(around 26%).*
- *Lower Secondary, even though having a less percentage of 0.8% of the total population, have a high defaulter percentage.*
- *As we can see, as the education level increases less likely the customer is to default. Hence, when giving loans Banks should give preference to people with Higher education level.*

5. Income Type
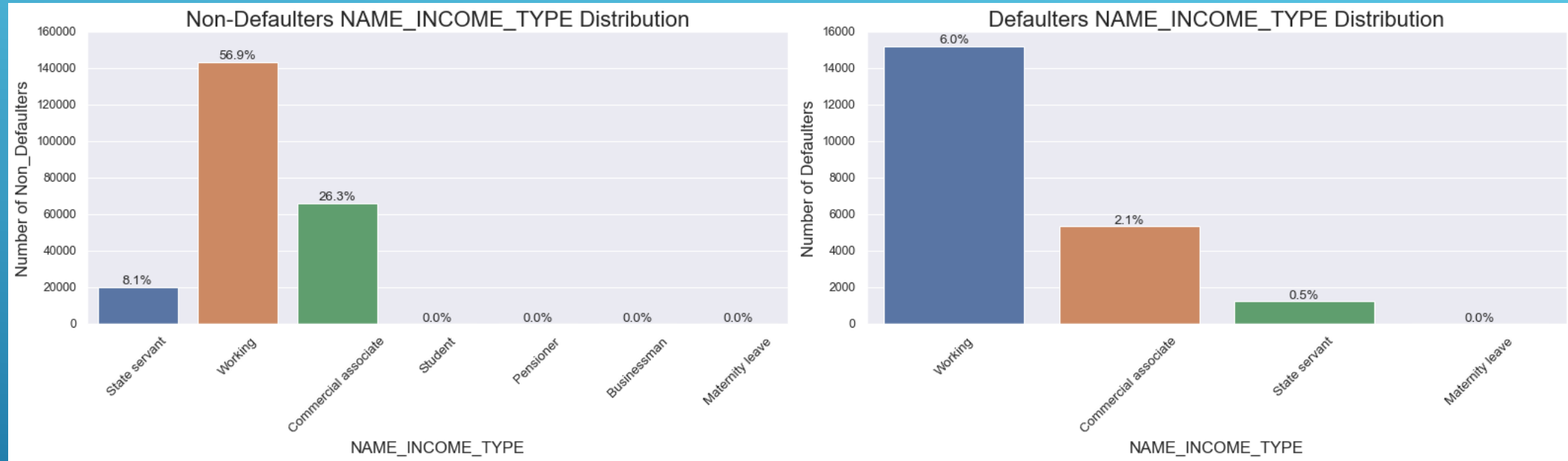
- *Customers who are currently Working have high proportion of defaulters when compared to others.*
- *In comparison, State Servants are less likely to Default.*

6. Housing Type
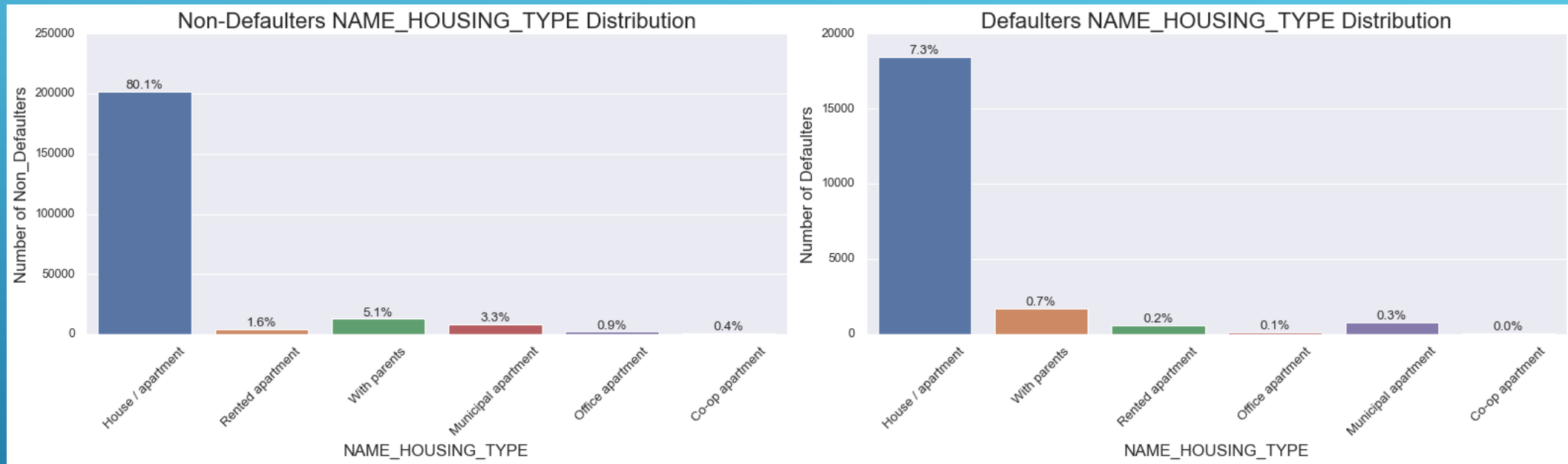
- *Around 87% of the loans were applied by the customers living in House/Apartment.*
- *Customers living With Parents have slightly more proportion of defaulting when compared to non-defaulters.*
- *Furthermore, the percentage of defaults in Municipal and Rented apartment is slightly higher.*

7. Occupation Type

Non-Defaulters OCCUPATION_TYPE Distribution

Defaulters OCCUPATION_TYPE Distribution
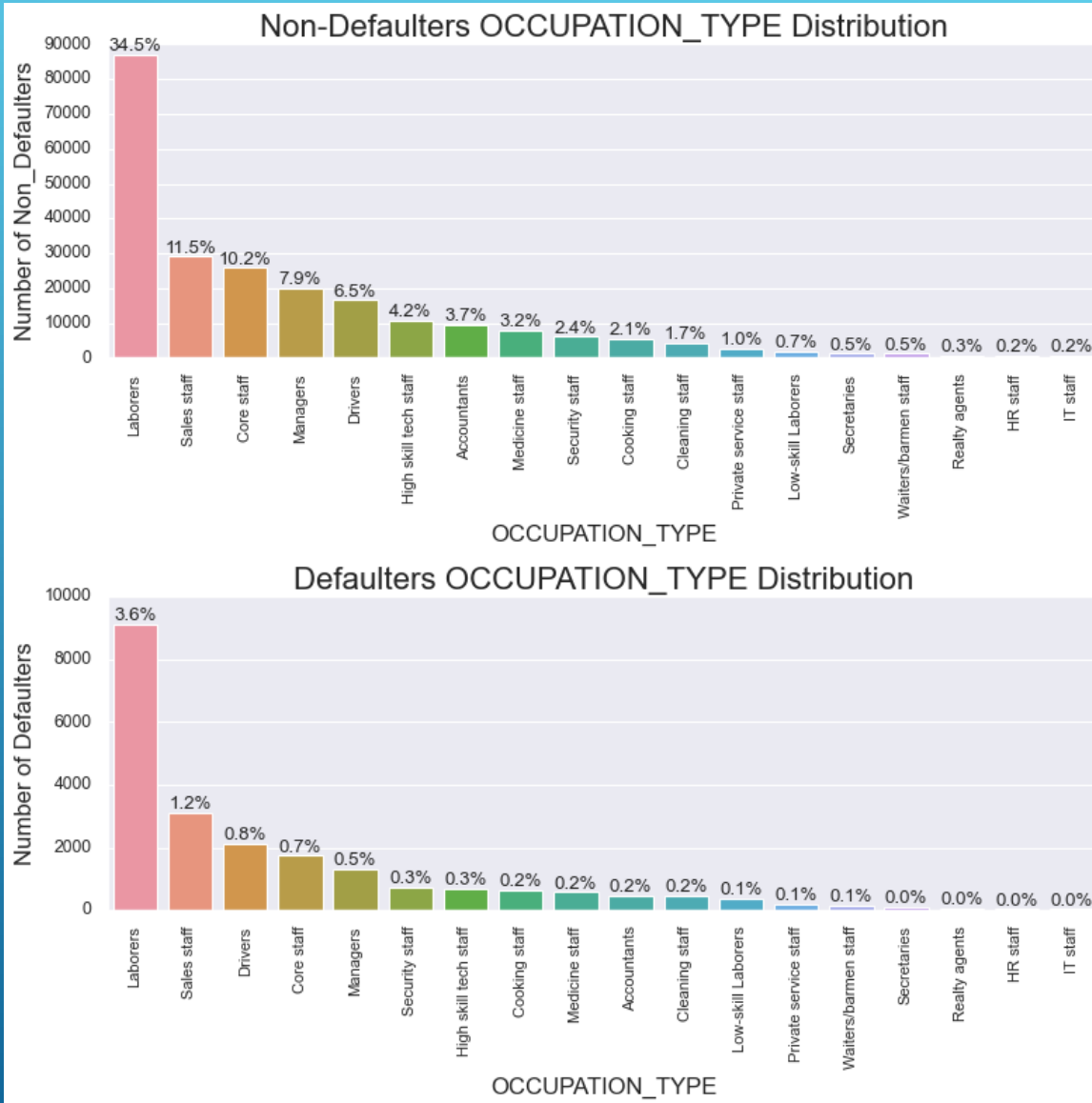
- *Around 38% of the total loans were applied by the customers having profession Labourers. They also have higher proportion of defaulters.*
- *Sales staff also have a slightly higher proportion of defaulters when compared with the others.*
- *We can also observe that IT/HR staff have lower proportion of defaults.*
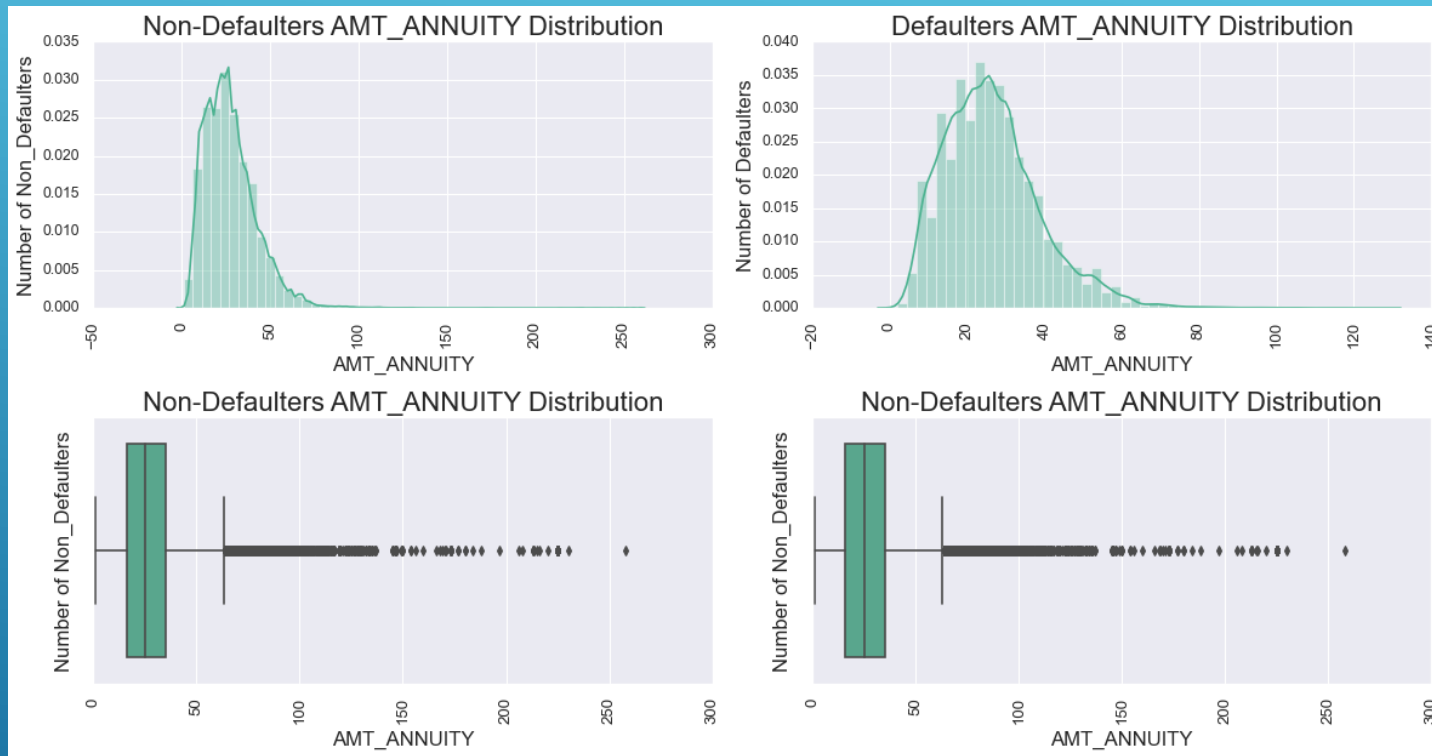
8. Age Group
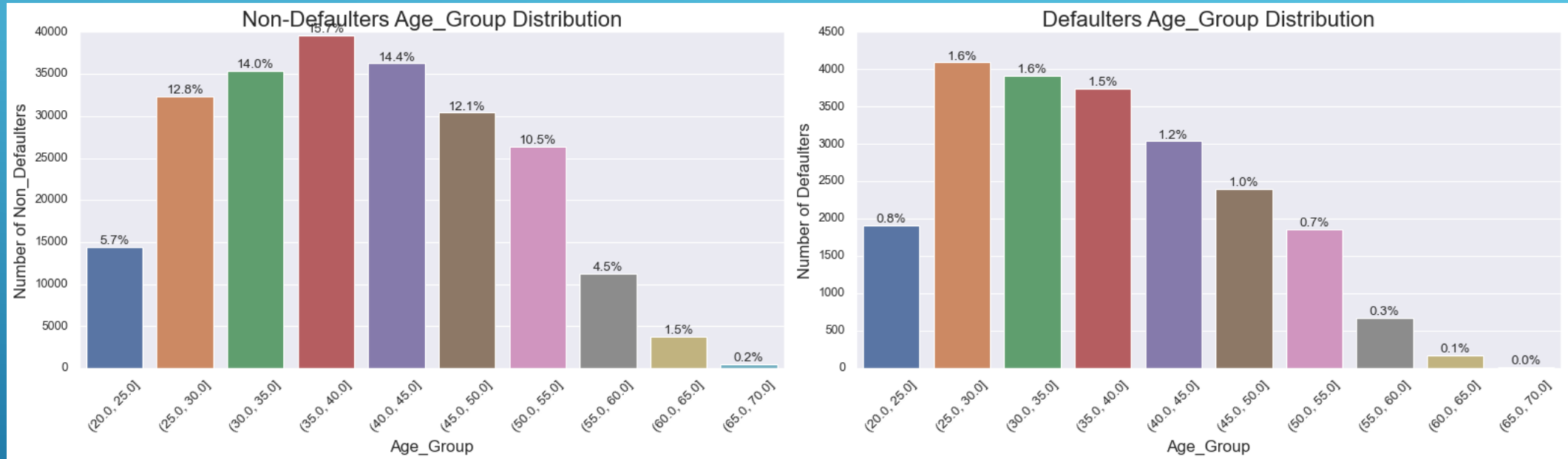
- *Majority of the loans were applied by the customers between the age 25 and 50.*
- *Customers under 30 years are more likely to be defaulters.*
- *Customers above 60 years are less likely to be defaulters.*

Univariate Analysis on Continuous Variables

1. Annuity

- *For Annuity, we observe a slight increase in the proportion of Defaulters when compared with non-defaulters.*
- *The maximum percentage of the defaulters are in the range [0 − 40,000].*
- *The average Annuity for both the Defaulters and Non-Defaulters is similar, which is around 25,000.*
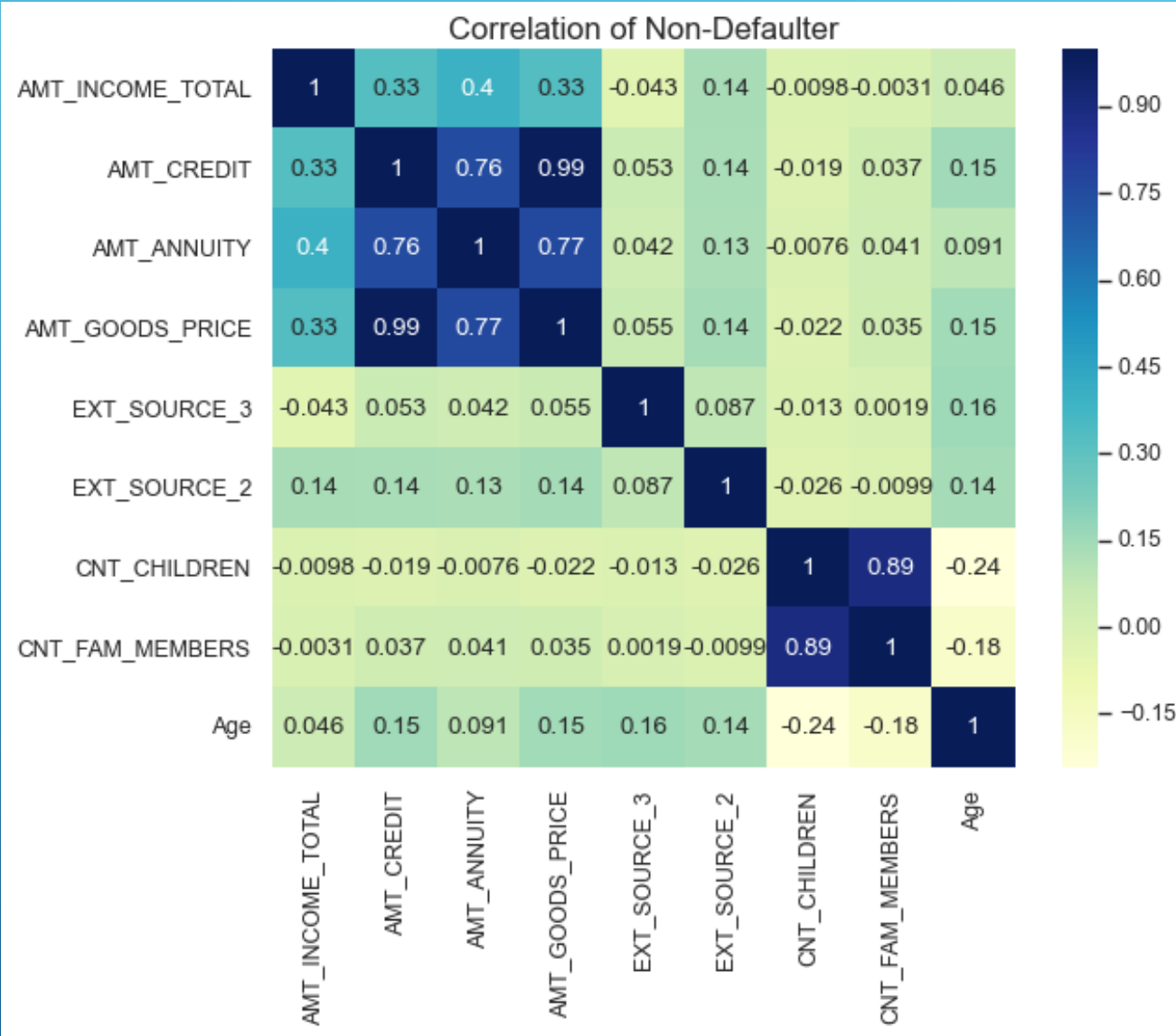
2. Credit
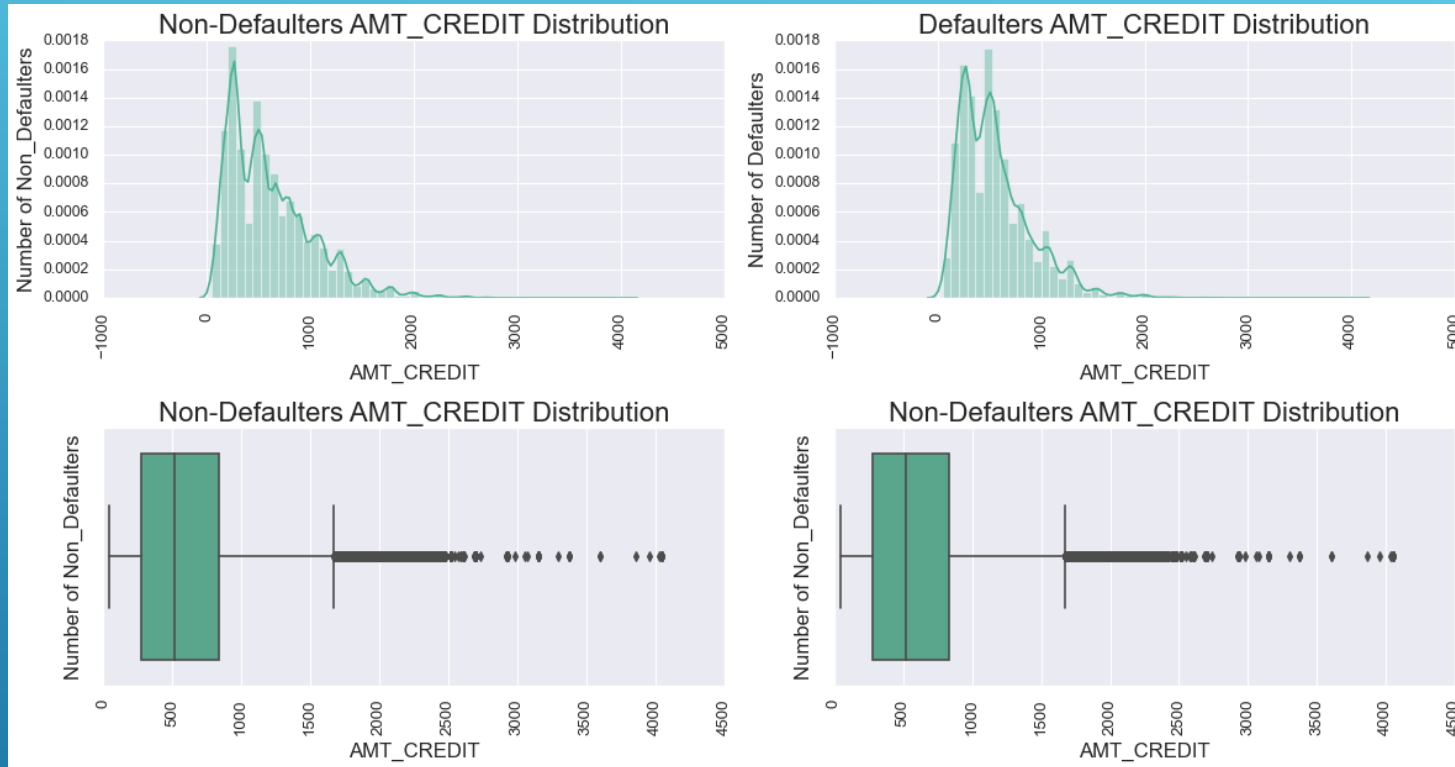
- *For Credit, we observe a slight increase in the proportion of Defaulters when compared to non-defaulters.*
- *The maximum percentage of the defaulters are in the range [0 − 1000K].*

Correlations

1. Target = 0

Correlation of Non-Defaulter

- *Credit amount is inversely proportional to the number of children a customer have.*
- *Income is inversely proportional to the number of children a customer have.*
- *Credit amount is directly proportional to the Goods price, with a correlation 0.99.*
- *Credit amount is directly proportional to the Annuity, with a correlation 0.76.*
- *Income is inversely proportional to the number of family members a customer have.*
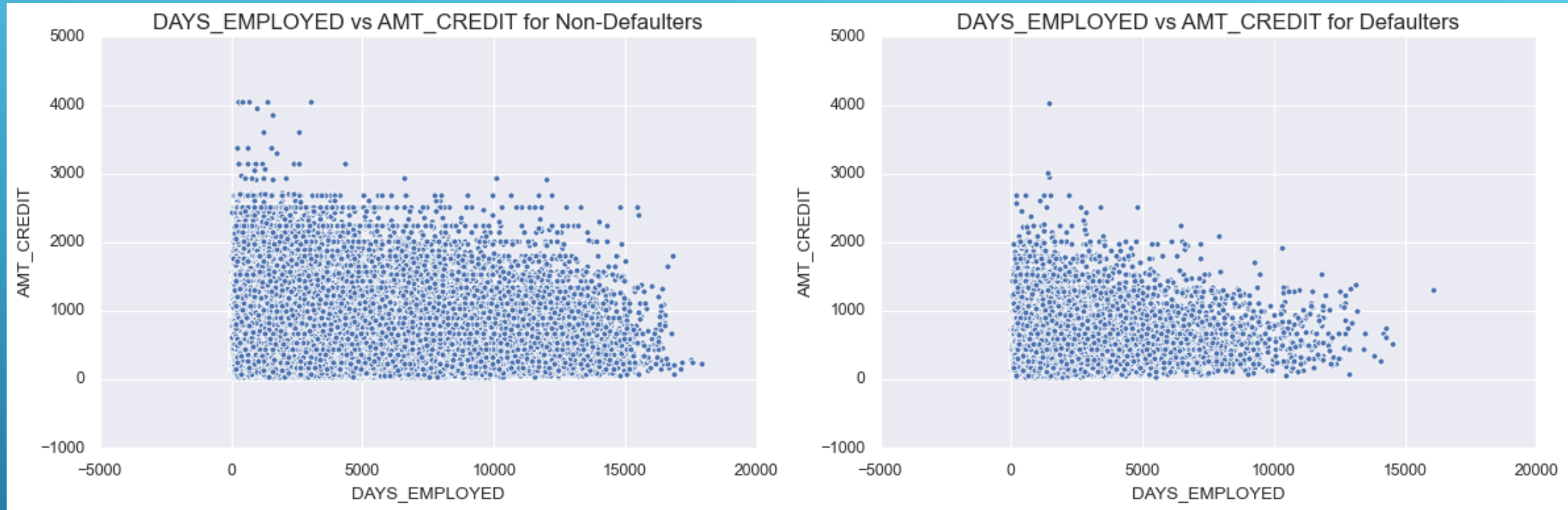
2. Target = 1

Correlation of Defaulter
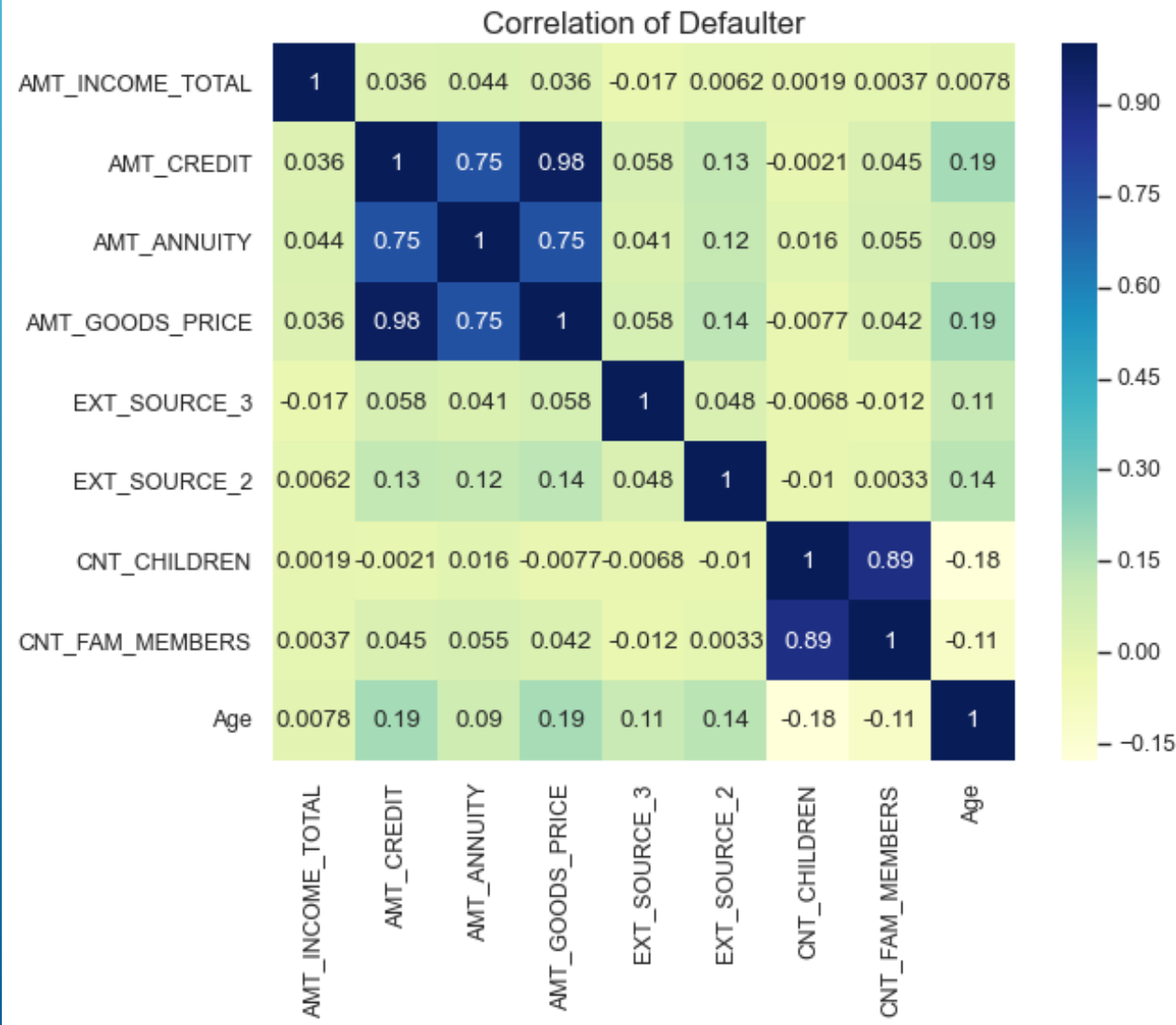
- *The heat-map for Target = 1(Defaulter) is also having the same observations with slight variations in the values.*
- *Correlation between Credit amount and Annuity 0.75.*
- *Correlation between Credit amount and Income amount 0.036.*

Bivariative Analysis

1. Days Employed vs Credit

- *We observe that, as the customer's experience increases the less likely the customers tend to default, customers with less experience tend to be defaulters.*
- *Also the loan default is concentrated below 2000K loan amount and the customers above 25 years of experience.*

2. Age vs Credit
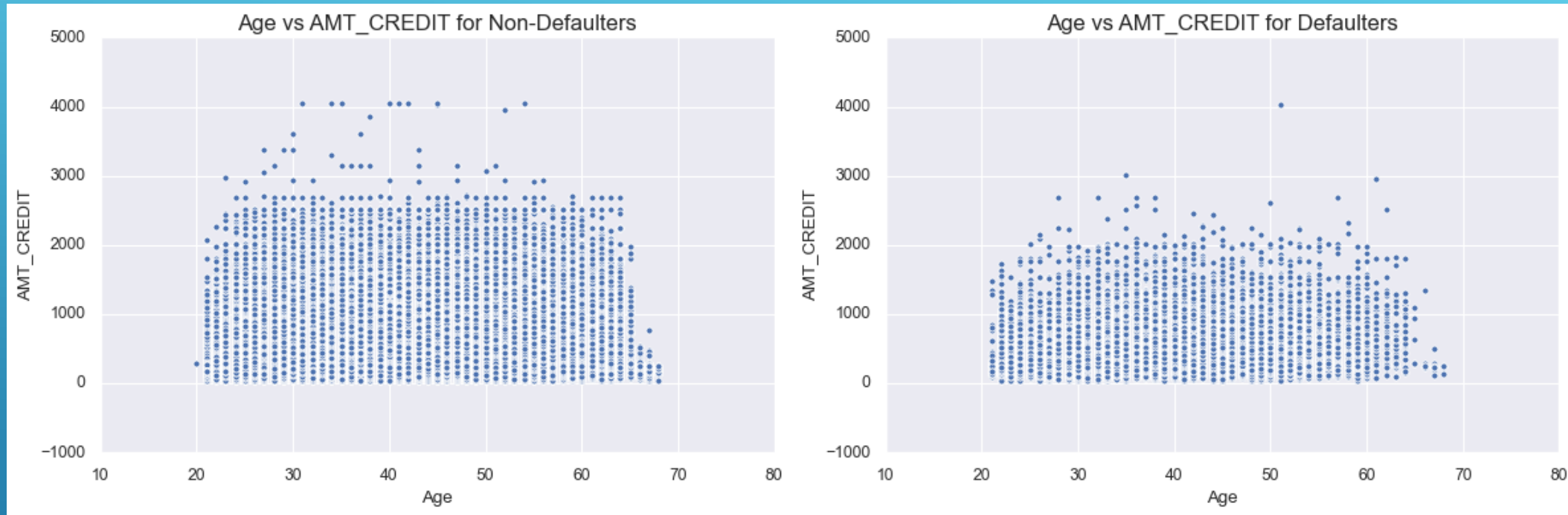
- *We observe that, the customer above 60 years are less likely to be a defaulter.*
- *Also the loan default is concentrated below 2000K loan amount and the customers between 20 and 60 years.*

## 3. Goods Price vs Credit
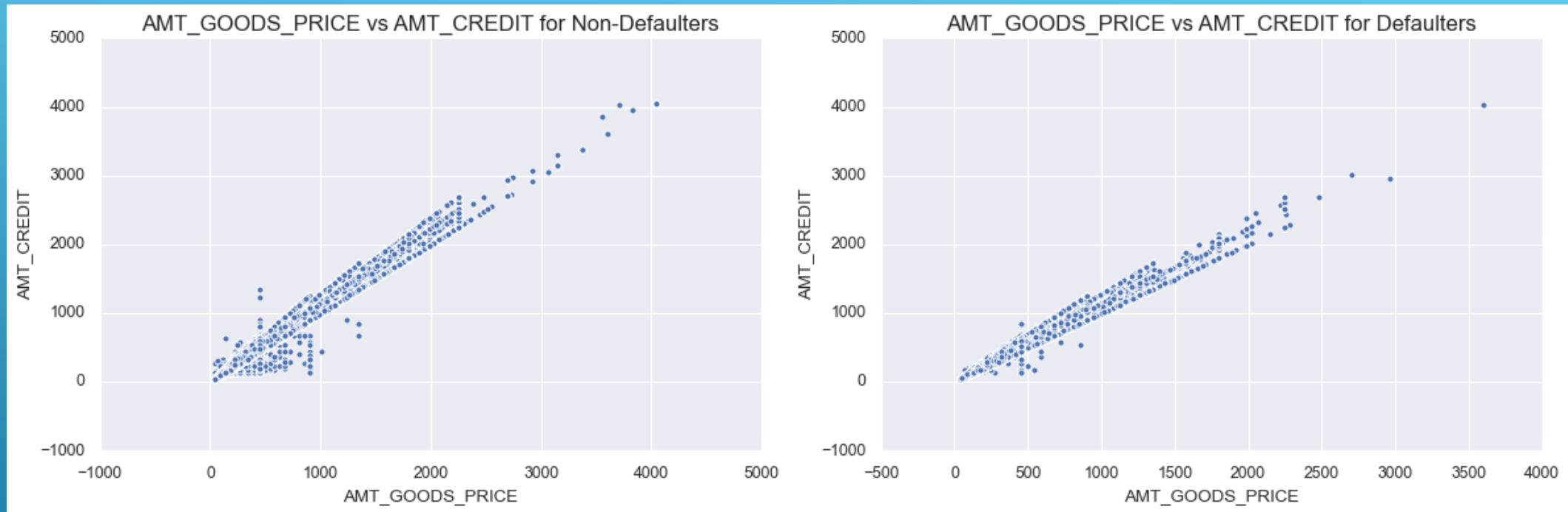
- *As we observed from the heat-map, the correlation between Goods Price and Credit amount for Non-Defaulters is 0.99 and for Defaulters is 0.98.*
- *Hence, we can say that Goods Price is directly proportional to the Credit amount for both Non-Defaulters and Defaulters. So as the Goods Price increases the Credit amount also increases.*

4. Annuity vs Credit
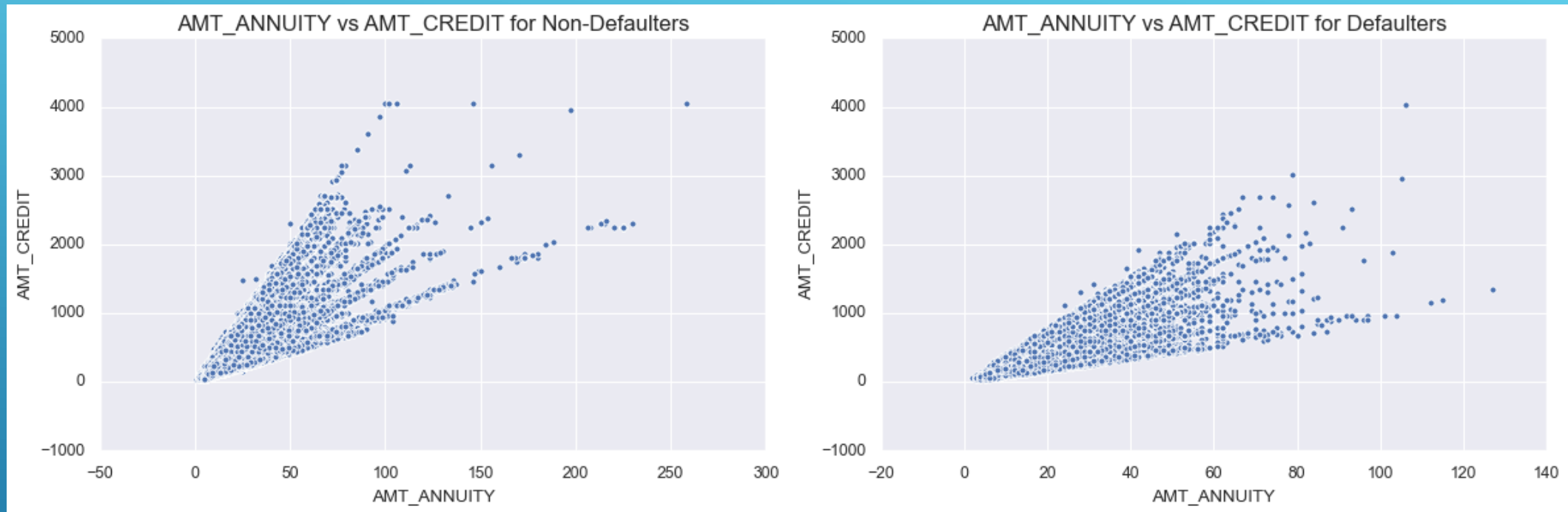
- *As we observed from the heat-map, the correlation between Annuity and Credit amount for Non-Defaulters is 0.76 and for Defaulters is 0.75.*
- *Hence, we can say that Annuity is directly proportional to the Credit amount for both Non-Defaulters and Defaulters. So as the Annuity increases the Credit amount also increases.*

# Analysis on Previous and Merged Application Data
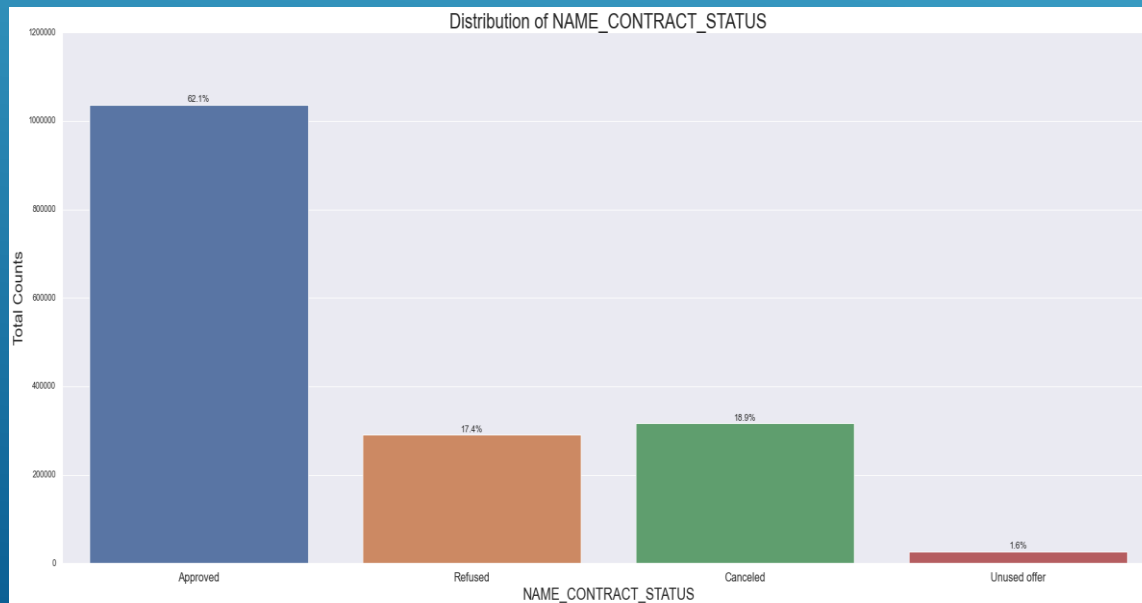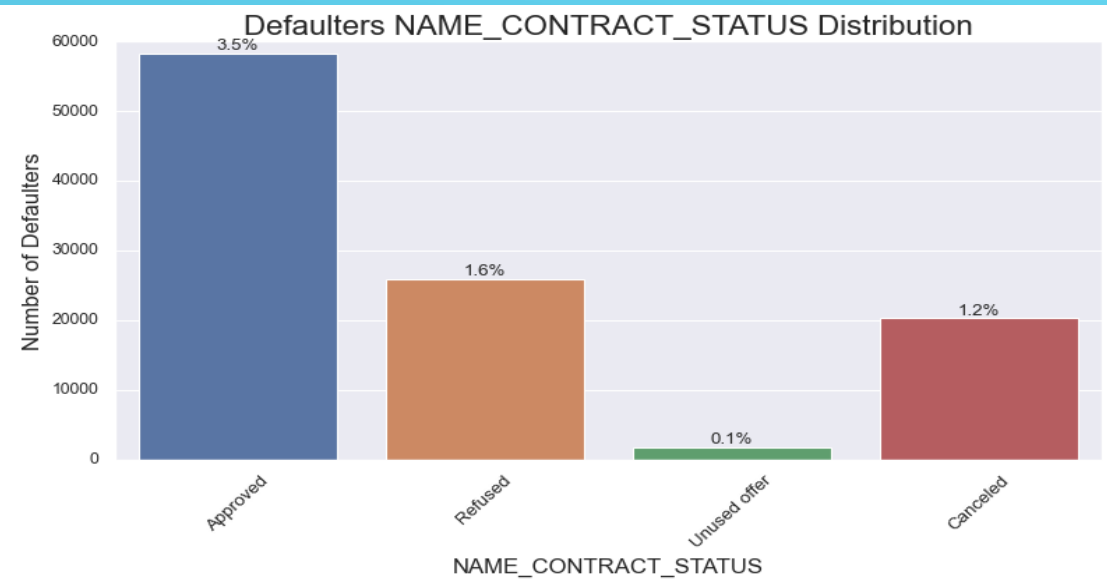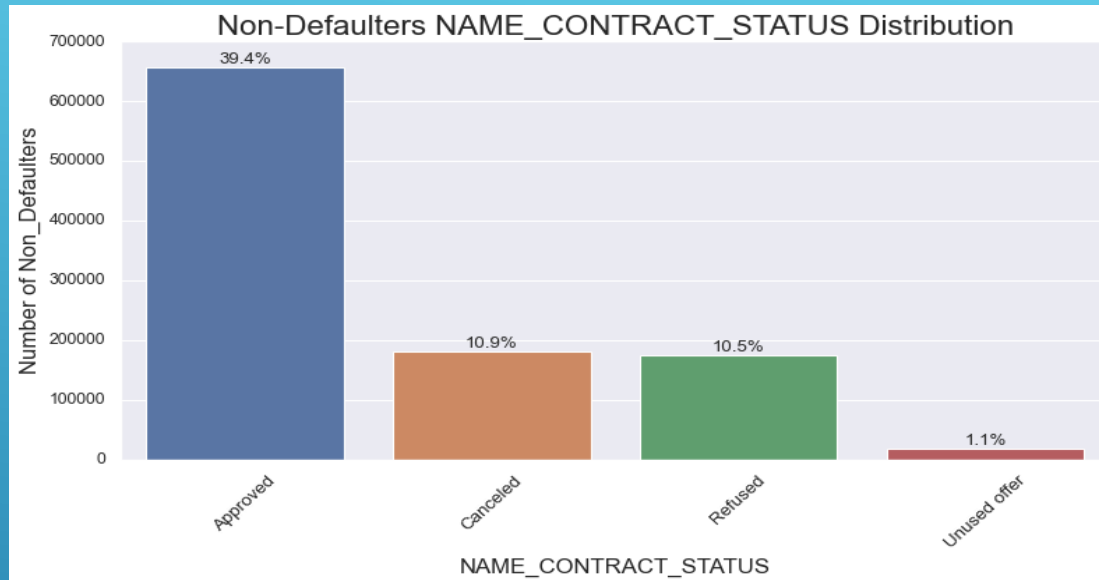
- 3.5% of the total customer's loans were Approved even though they were Defaulters.
- Around 10% of the loans were Refused for Non-Defaulted customers.
- About only 1% of the loans were Refused for the Defaulted customers.
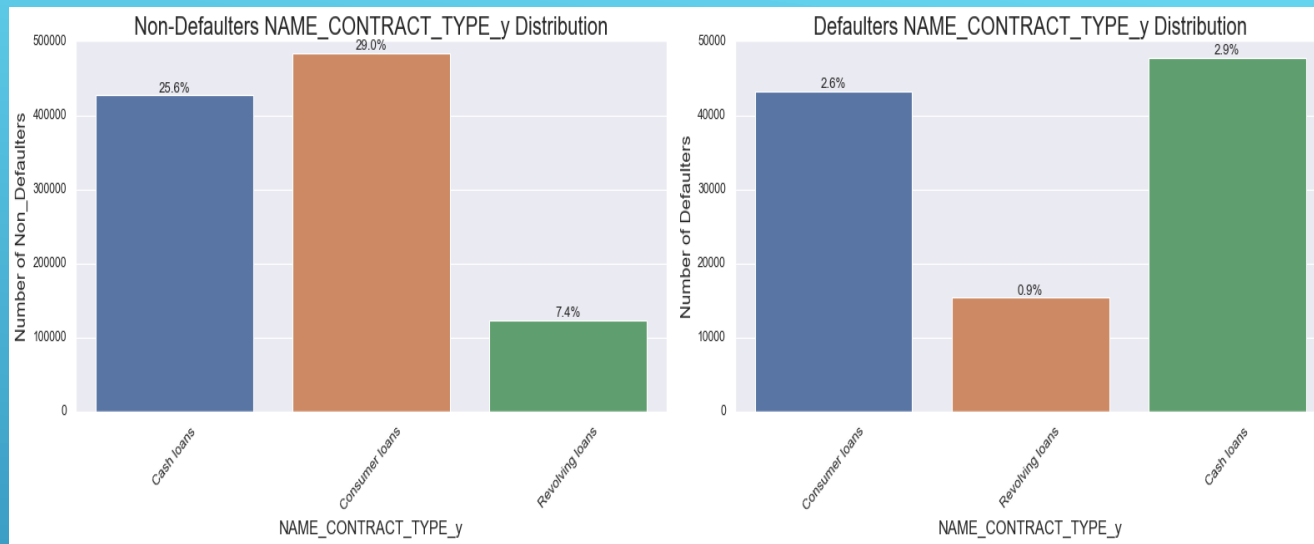
- *It seems that for defaulters more of the previous applications were for Cash Loans.*
- *For non–defaulters the percentage of Cash loans and Consumer loans have a very slight difference, Consumer loans being the max.*
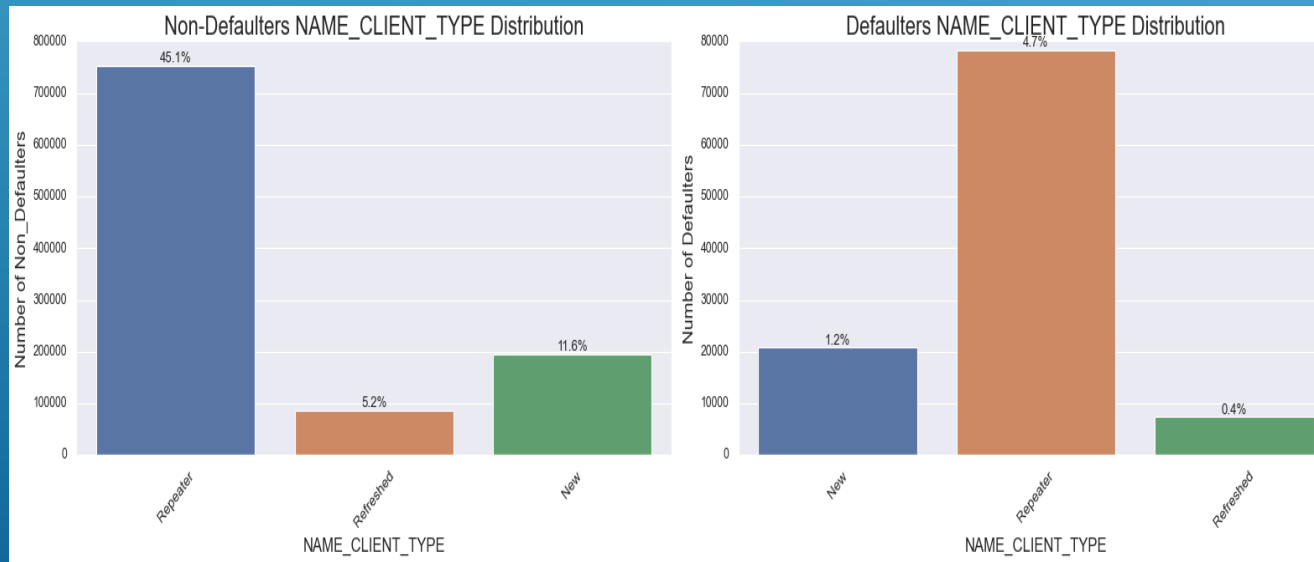
- *As we can observe that, there are around 49% of the Repeater customers.*
- *Only around 4% of the Repeater customers are Defaulters.*
- *And about 12% of the new customers are defaulters.*

Proportion of Defaulters


Defaulers in NAME_CONTRACT_STATUS and NAME_CLIENT_TYPE
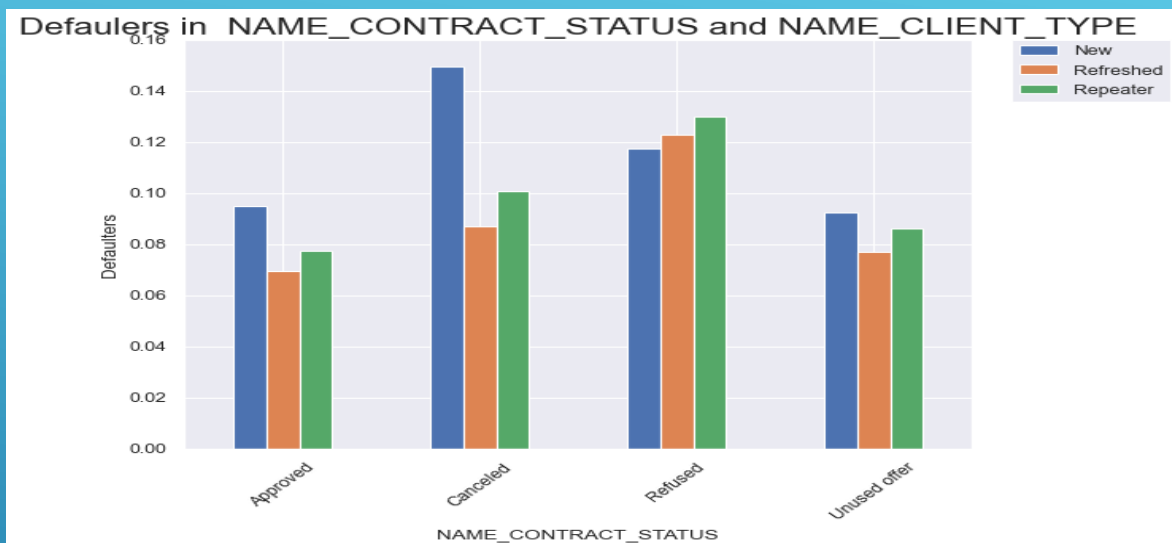
- We can observe that 'New' customers have there previous loan applications cancelled the most, as they tend to be defaulters.
- We observe a very high peak in all the Contract Status in Refused column.
- Repeaters have got there loan application refused the most.
- We also see a high peak in approval for Repeater customers.

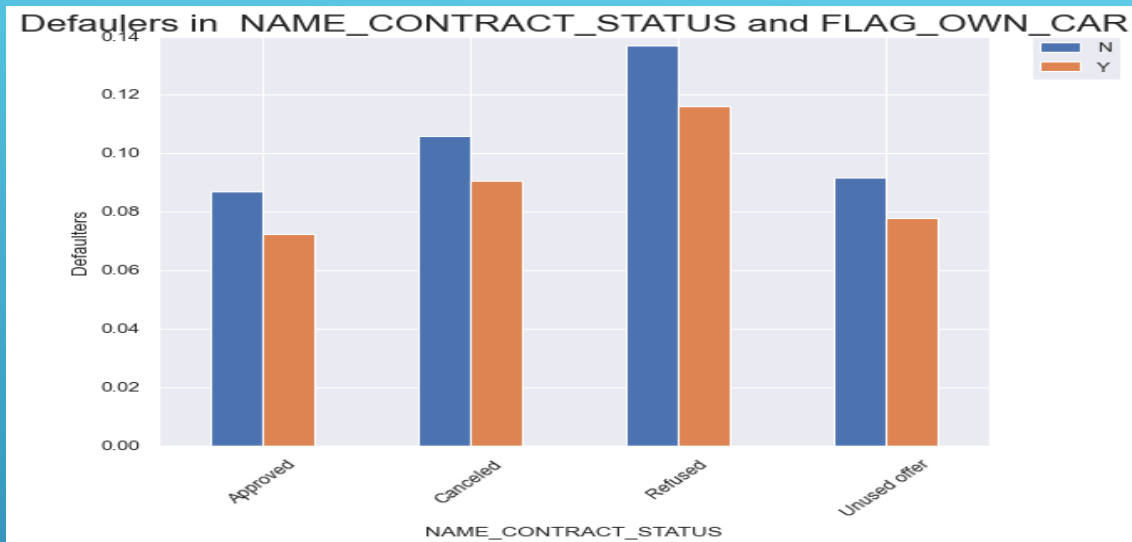

Defaulers in NAME_CONTRACT_STATUS and NAME_CONTRACT_TYPE_y
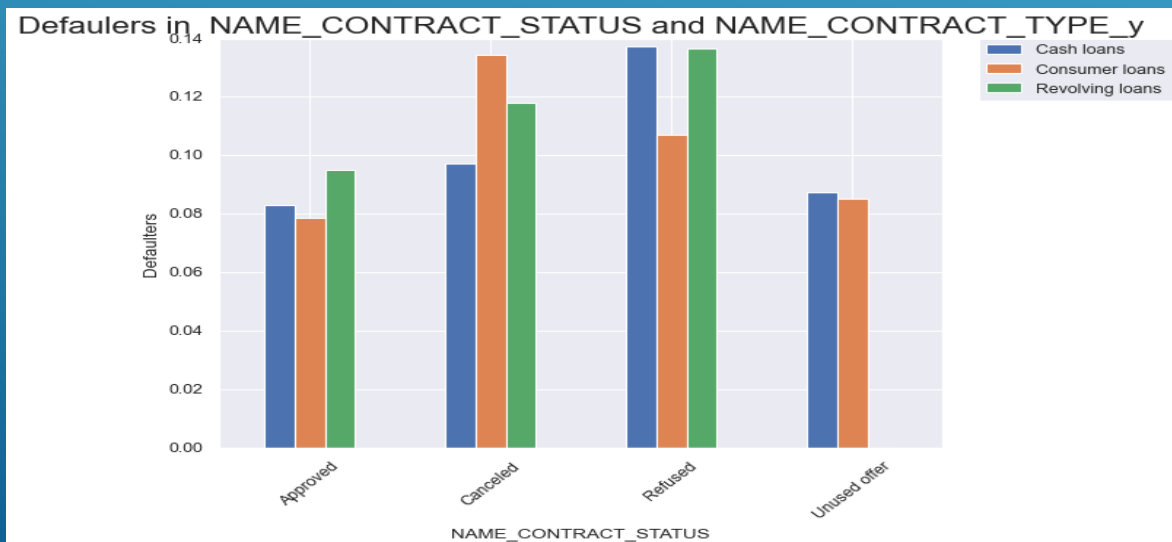
- In the 'Refused' column we can observe a peak on 'Cash loans' and 'Revolving loans'.
- Consumer loans were cancelled the most by the customers in the previous application.
- Most of the defaulters were refused for the loan in the previous application.

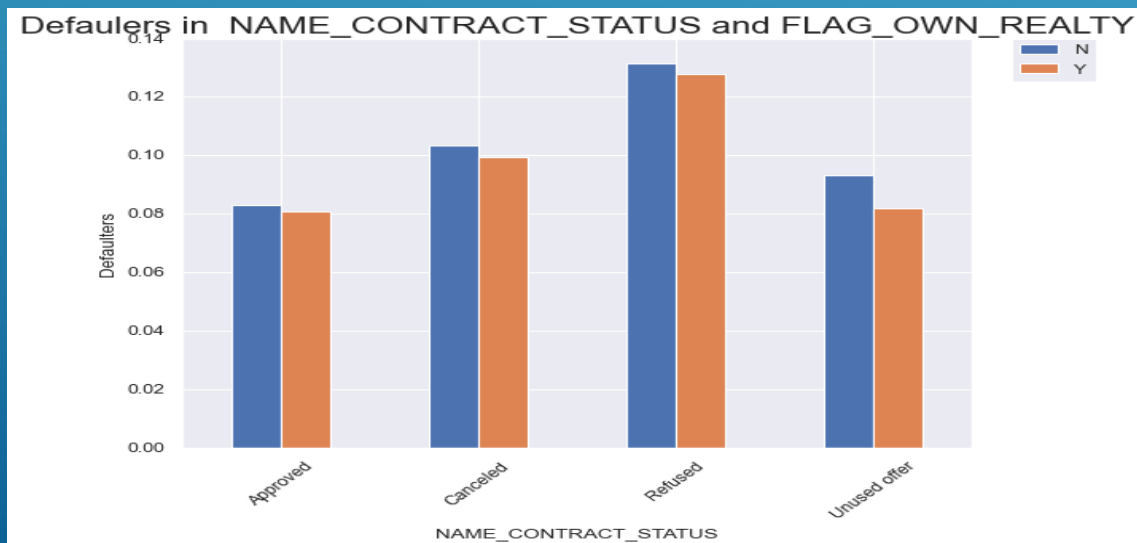Defaulers in NAME_CONTRACT_STATUS and NAME_FAMILY_STATUS
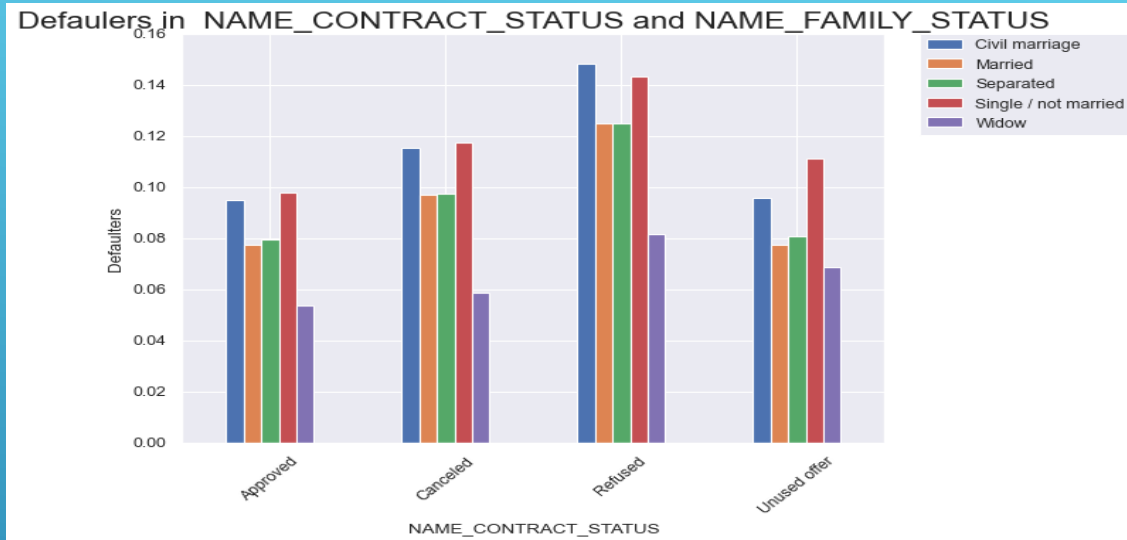
- *We observe that the most % of defaulters are from 'Civil Marriage', followed by 'Single/Not Married'.*
- *Widows are less likely to be defaulters.*
- *We observe that most of the defaulters were refused in the previous application.*

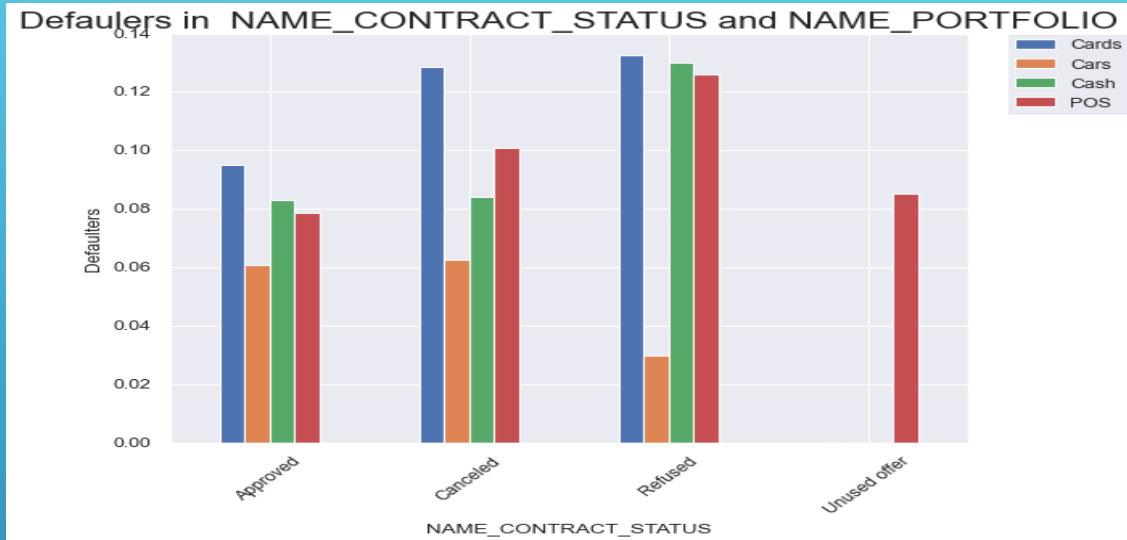Defaulers in NAME_CONTRACT_STATUS and NAME_EDUCATION_TYPE
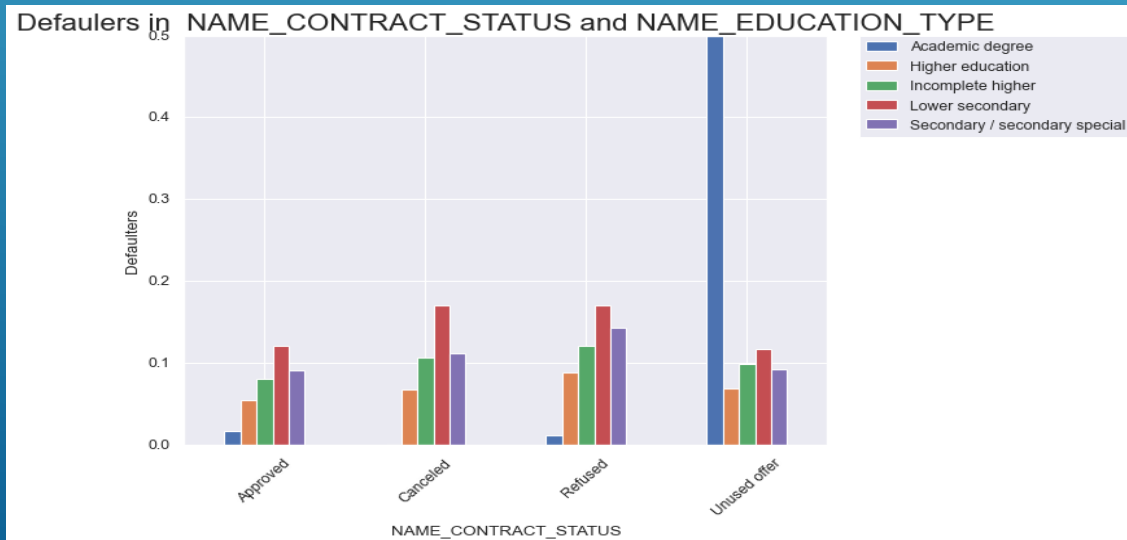
- *We observe that the most percentage of defaulters belong to 'Lower Secondary' and 'Secondary/Secondary Special' education type.*
- *We also observe that 'Academic Degree' have the lowest percentage of defaulters present.*

Defaulers in NAME_CONTRACT_STATUS and FLAG_OWN_CAR

- *We observe that the most percentage of defaulters doesn't own car.*

Defaulers in NAME_CONTRACT_STATUS and FLAG_OWN_REALTY

- *We observe that the most percentage of defaulters doesn't own realty.*

Defaulers in NAME_CONTRACT_STATUS and NAME_PORTFOLIO

- *We see a high Default percentage in the 'Approved' column in all of the Portfolios.*
- *There is a high Refused rate in Cards, Cash and POS, which is a good thing.*

Defaulers in NAME_CONTRACT_STATUS and Income_Range
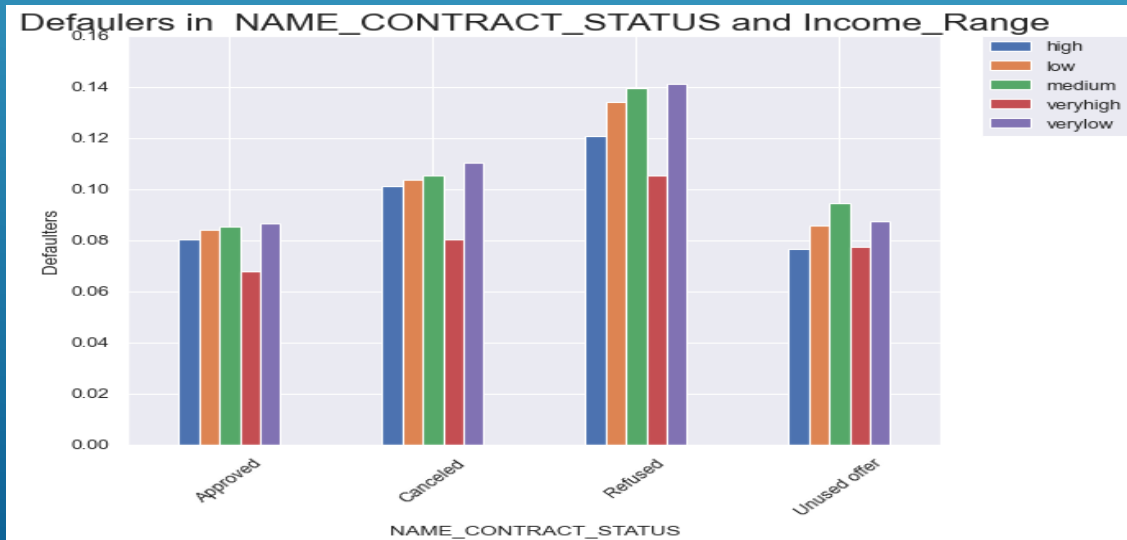
- *We observe that people have Very-low Income, have the maximum percentage of defaulters, followed by Medium Income and Low Income.*
- *We can observe a high refuse rate for all the Income ranges.*

Defaulers in NAME_CONTRACT_STATUS and Credit_Range
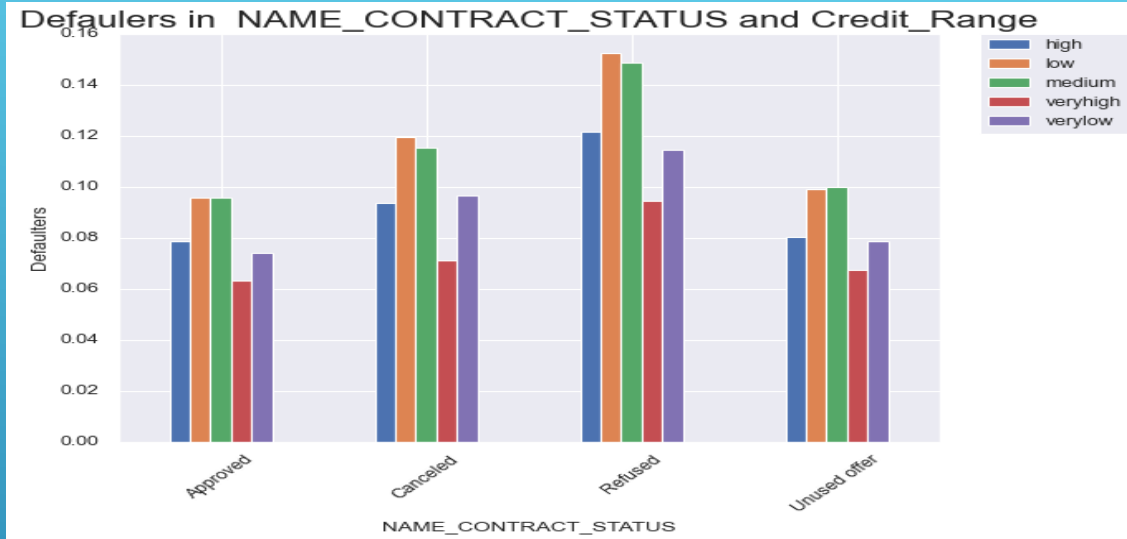
- *We see a high Default percentage for the Low Credit, followed by Medium and High Credit.*
- *We can observe a high refuse rate for all the Credit ranges.*

Defaulers in NAME_CONTRACT_STATUS and Annuity_Range
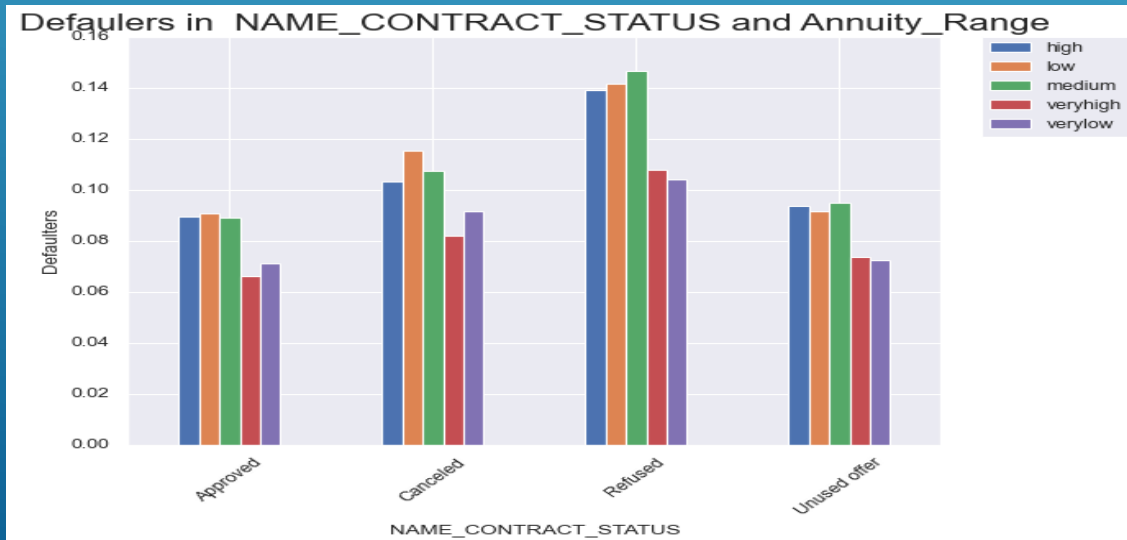
- *We observe that customers having Medium Annuity, have the maximum percentage of defaulters, followed by Low and High.*
- *We can observe a high refuse rate for all the Annuity ranges.*

Defaulers in NAME_CONTRACT_STATUS and CODE_GENDER
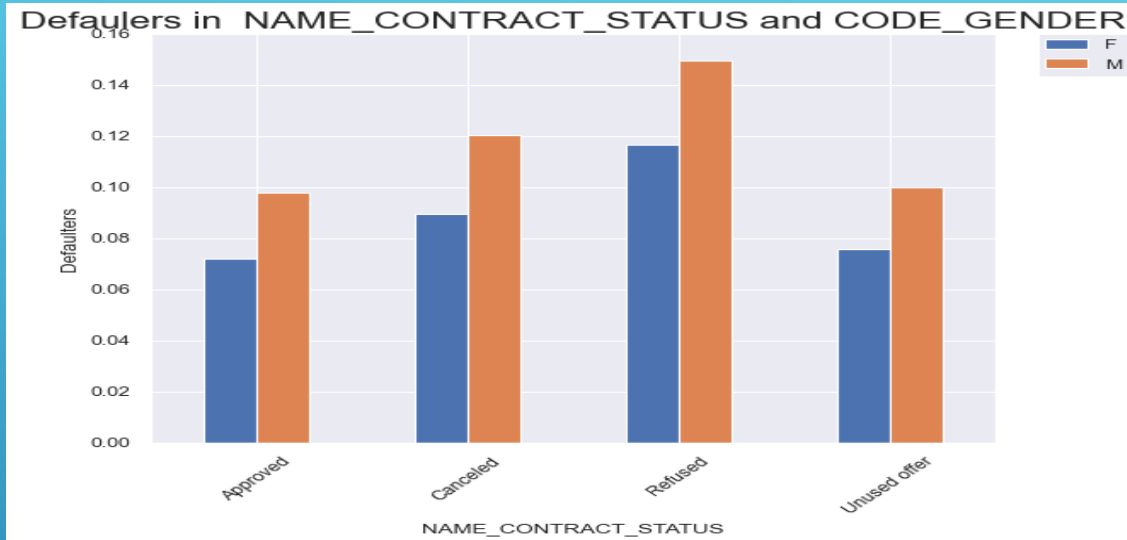
- *We observe that Females have applied more % of loans than that of Males. And have got approved even though they being a defaulter chances are high.*
- *We can observe that most of the defaulters were refused in the previous application.*


Defaulers in NAME_FAMILY_STATUS and CODE_GENDER
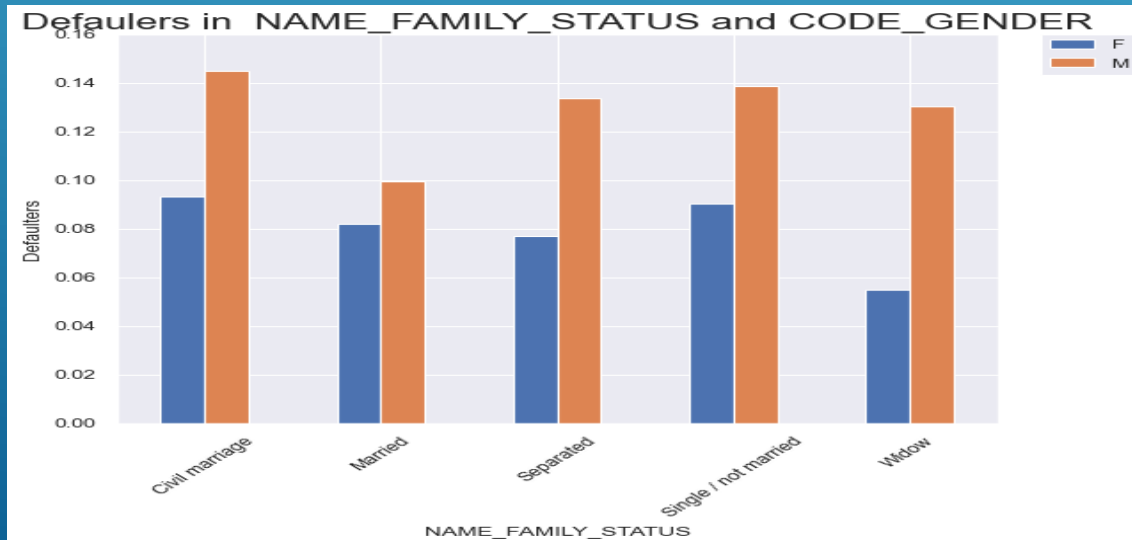
- *We observe that Males with Civil Marriage have the highest peak being a defaulter, followed by Single/Not-Married and Separated.*
- *We also observe a high peak in Civil Marriage and Single/Not-Married for Female customers.*

Defaulers in NAME_EDUCATION_TYPE and CODE_GENDER
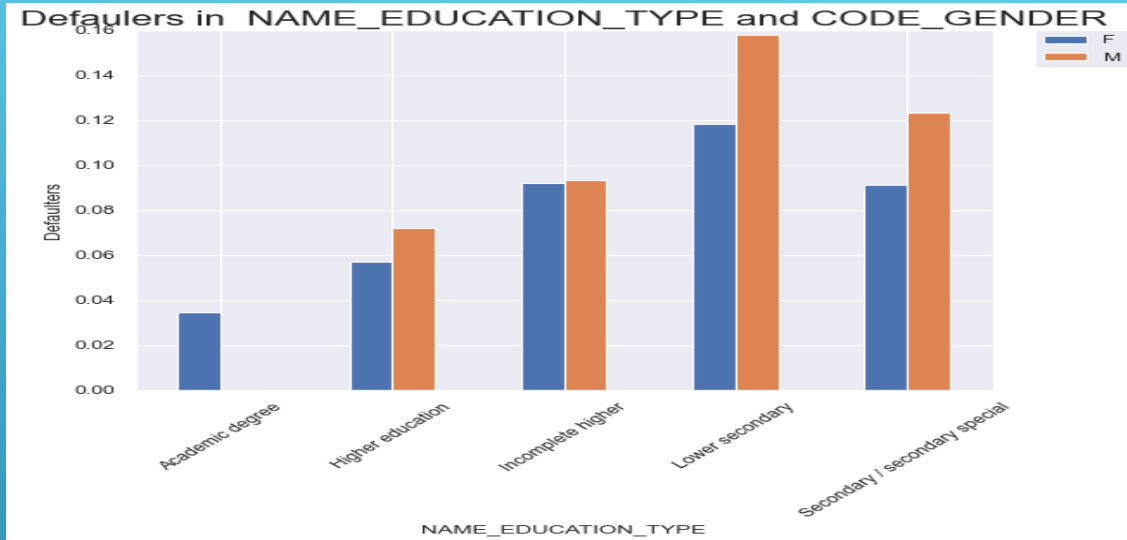
- *Male customers are more educated than that of the Female customers, also male customers tend to be defaulters.*
- *Most of the customers have Lower Secondary education level and also the Default percentage for Lower secondary is the highest.*
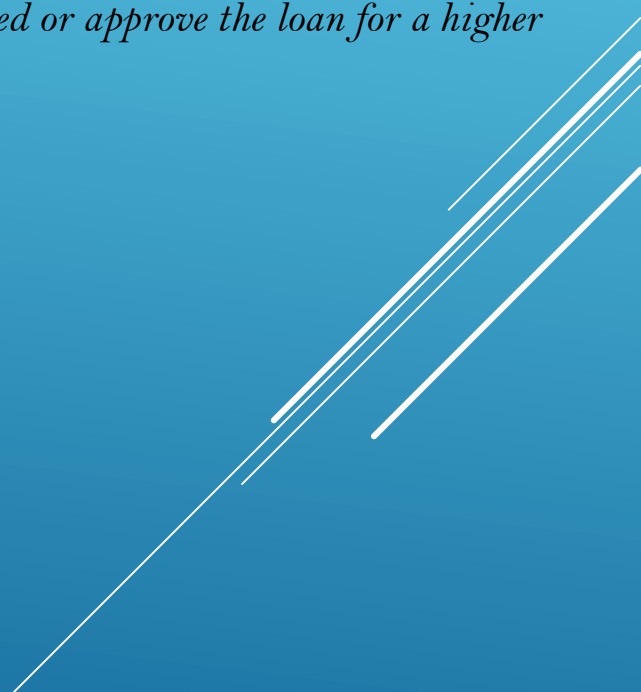

Defaulers in Age_Group and CODE_GENDER
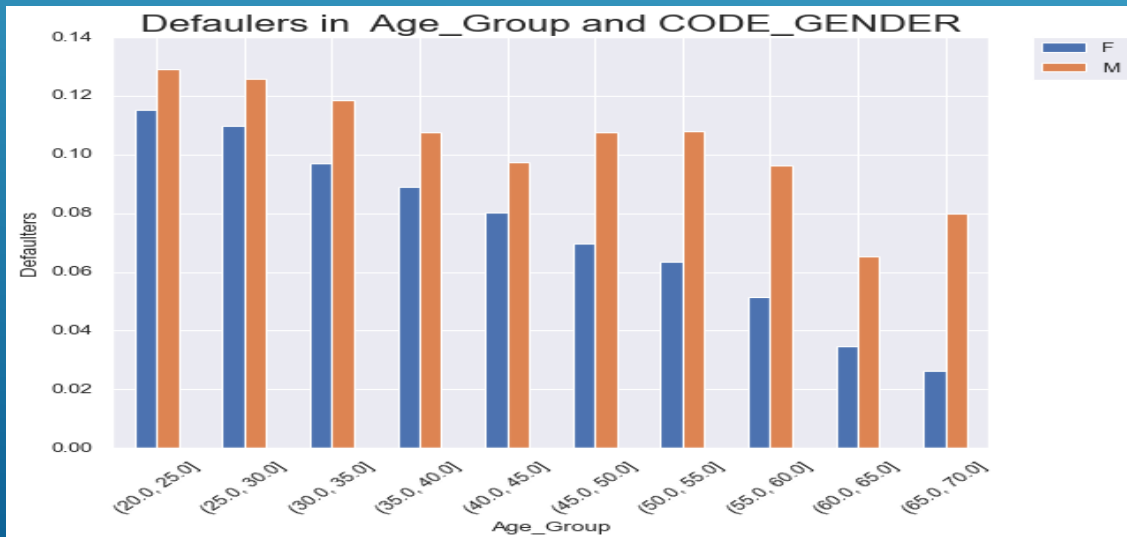
- *Male customers throughout the Age-group have consistently tend to be Defaulters.*
- *We can observe that as the age of the Female customer increases the less likely they tend to be Defaulters.*
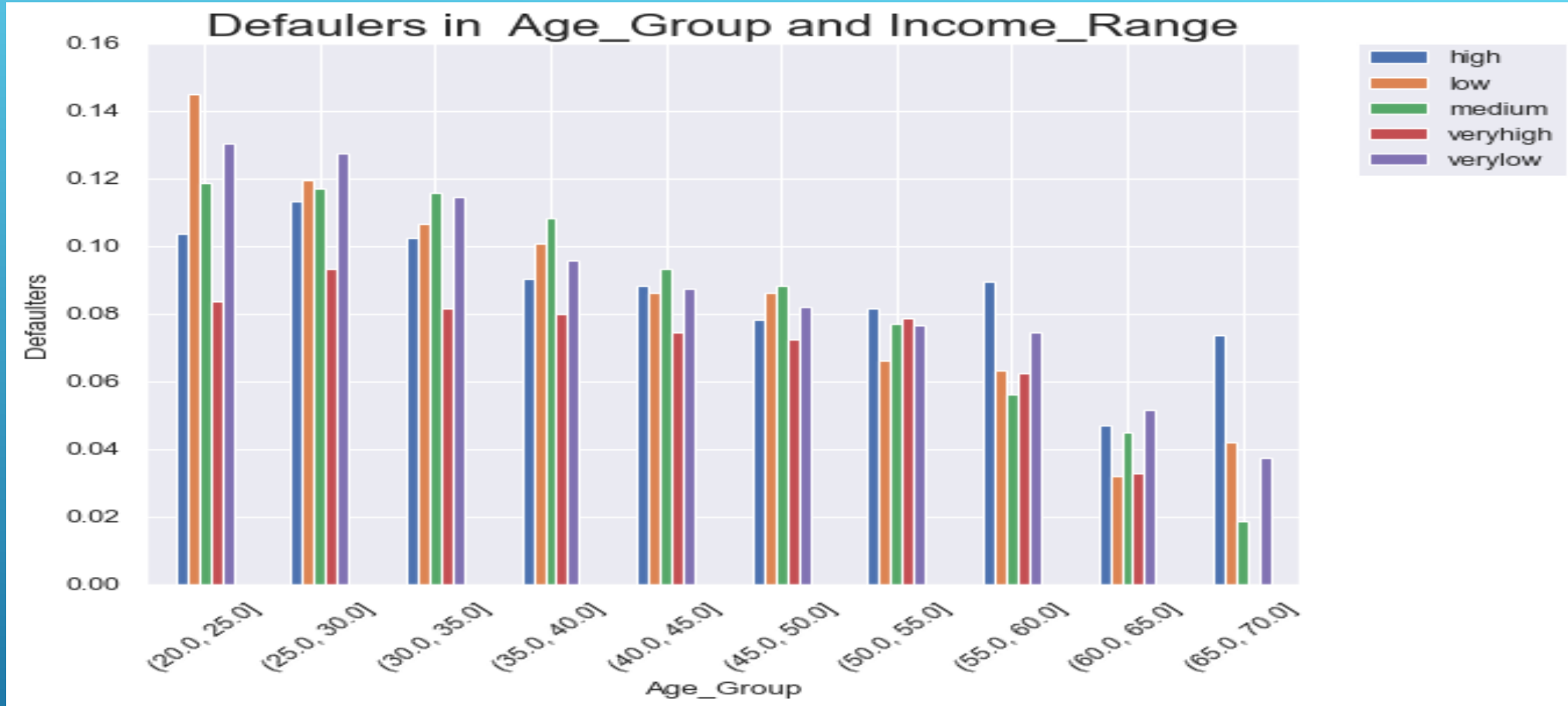
- *Customers aging between 20 and 30 years have the maximum percentage of Defaulters when compared to other Age groups.*
- *Customers aging between 55 and 70 years have the minimum percentage of Defaulters when compared to other Age groups.*
- *Customers above the age of 55 years with very high Income are less likely to Default.*

Conclusions

- *Bank should give out more Revolving loans, as the customers with revolving loans are less likely to Default.*
- *Most of the Consumer loans are approved by the banks, but there's a majority of customers who are either cancelling the loan application or it is left Unused.*
- *Females are less likely to Default than that of Males.*
- *The education level of a customer is a good indicator of their ability to pay back. Higher education level customers are more reliable than others.*
- *A customer's age is also an important factor. Although Older customers are less likely to default, that does not mean that Young customers should be excluded; they must be educated on the entire loan cycle and the consequences of defaulting.*
- *Single customers and Civil Married customers are more likely to Default, so either their loans should be refused or approve the loan for a higher interest rate.*
- *Customers with higher incomes are less likely to Default.*
- *Customers with higher work experience are less likely to Default.*