

I.I.I.T Bangalore & Upgrad

Post Graduate Diploma in Data Science (Jan-22 Cohort)

Credit EDA Assignment

By: Arohi Malviya

A. UNIVARIATE ANALYSIS

Univariate Analysis is performed by dividing the dataset into two data frames namely: Defaulters(Target=1) & Non-Defaulters(Target=0)

1. Categorical Ordered Univariate Analysis:

- ✓ Income
- ✓ Credit
- ✓ Age
- ✓ Region Rating
- ✓ Education Type
- ✓ Count of Family Members

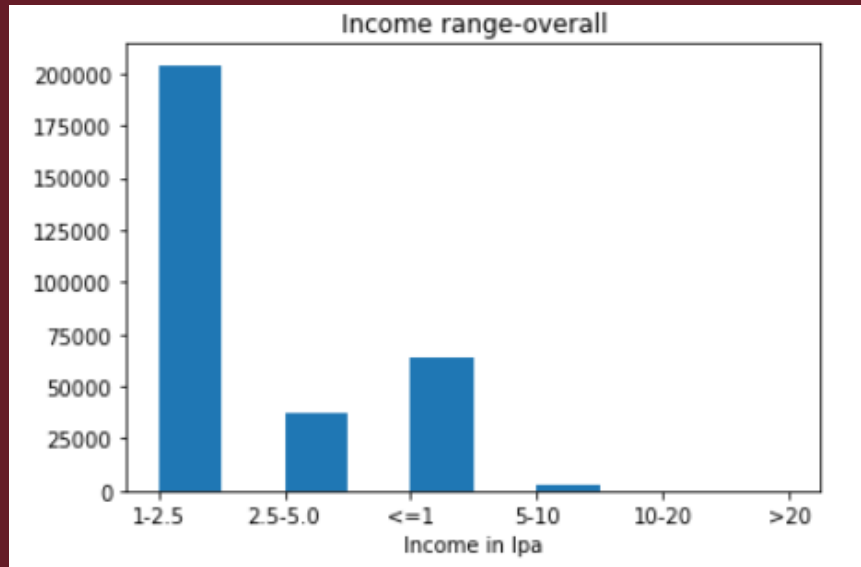
2. Categorical Unordered Univariate Analysis:

- ✓ Income Type
- ✓ Housing Type

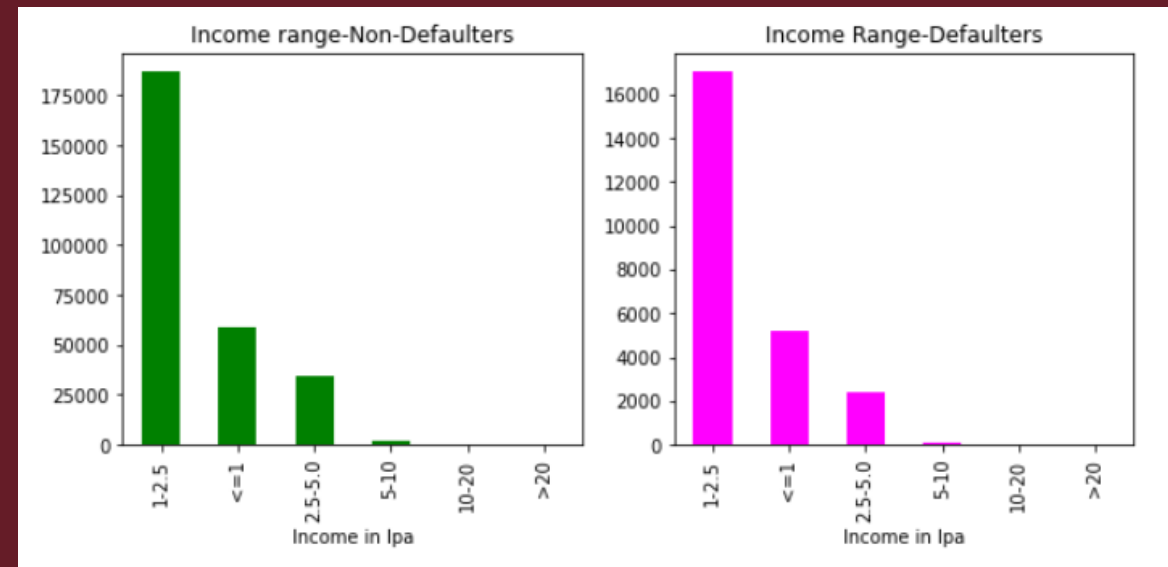
1. Categorical Ordered Univariate Analysis

1. Income

The Column is created by binning the AMT_INCOME_TOTAL column into groups of income.



A1 1.1 Income distribution for complete data set



A1 1.2 (a) Income distribution for non-defaulters

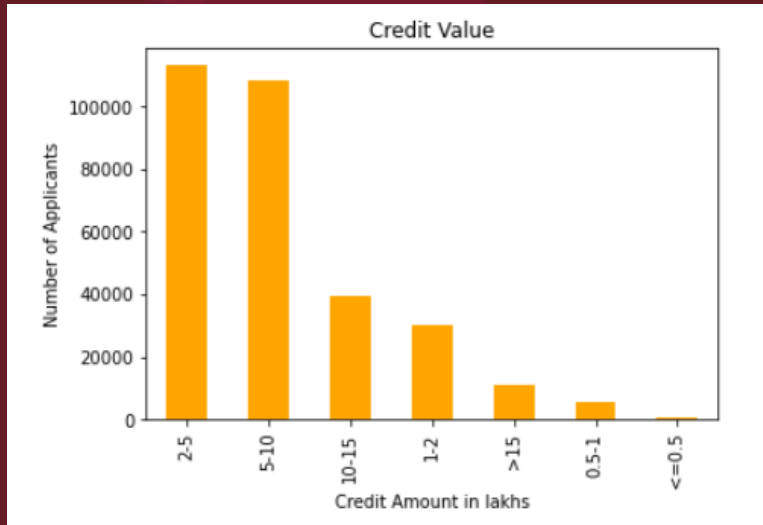
A1 1.2 (b) Income distribution for defaulters

Inference

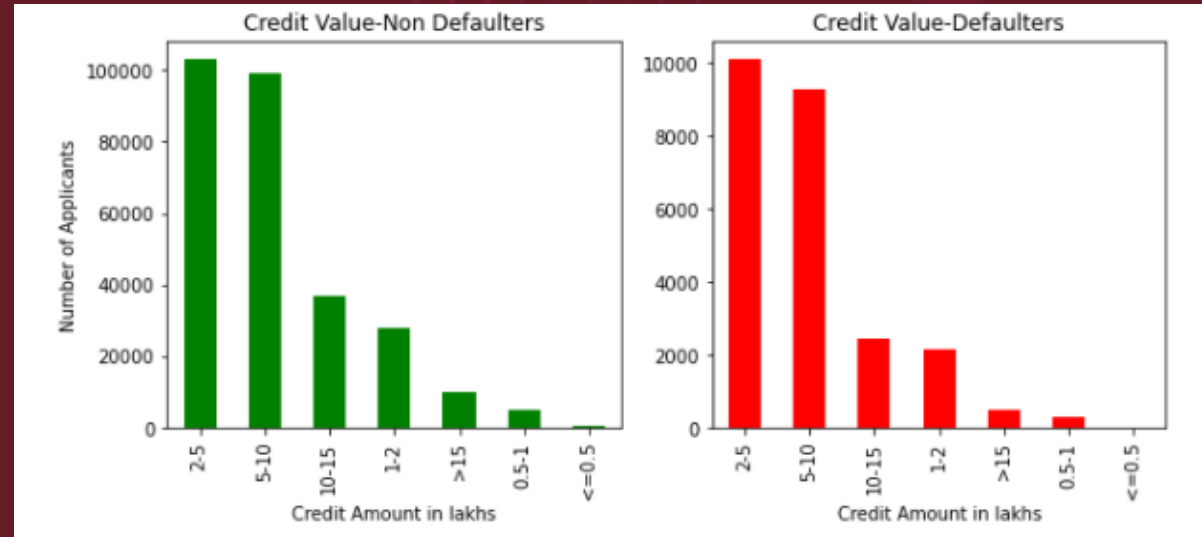
- The majority of applicants are in the Income range of 1-2.5L, followed by less than 1L
- This analysis suggests us that, in terms of absolute numbers, most of the defaulters are having annual income less than 5L.
- Top Defaulters are in the income group of 1-2.5L

2. Credit

The Column is created by binning the AMT_CREDIT column into groups of Credit Amount.



A1 2.1 Credit Amount distribution for complete data set



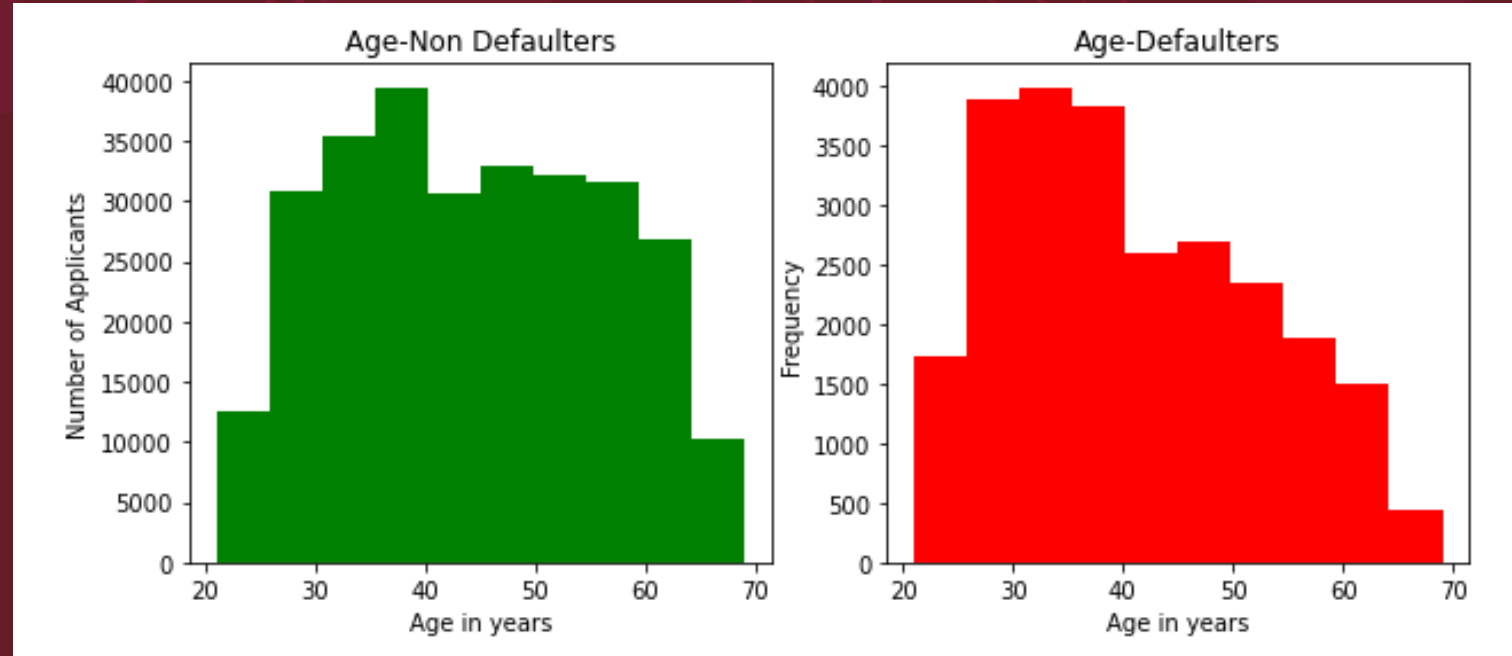
A1 2.2 (a) Credit amount distribution for non-defaulters

A1 2.2 (b) credit Amount distribution for defaulters

Inference

- More than 50% credit applications are between 2-15 lakhs
- In terms of credit value provided, the maximum defaulters lie in the range between 2-10 lakhs.

3. AGE



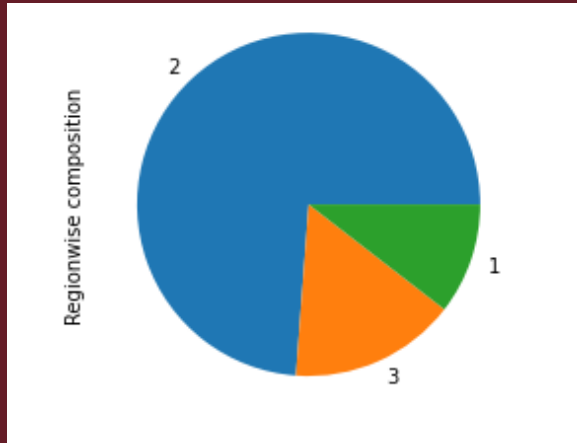
A1 3.1 (a) Age distribution for Non Defaulters

A1 3.1 (b) Age distribution for Defaulters

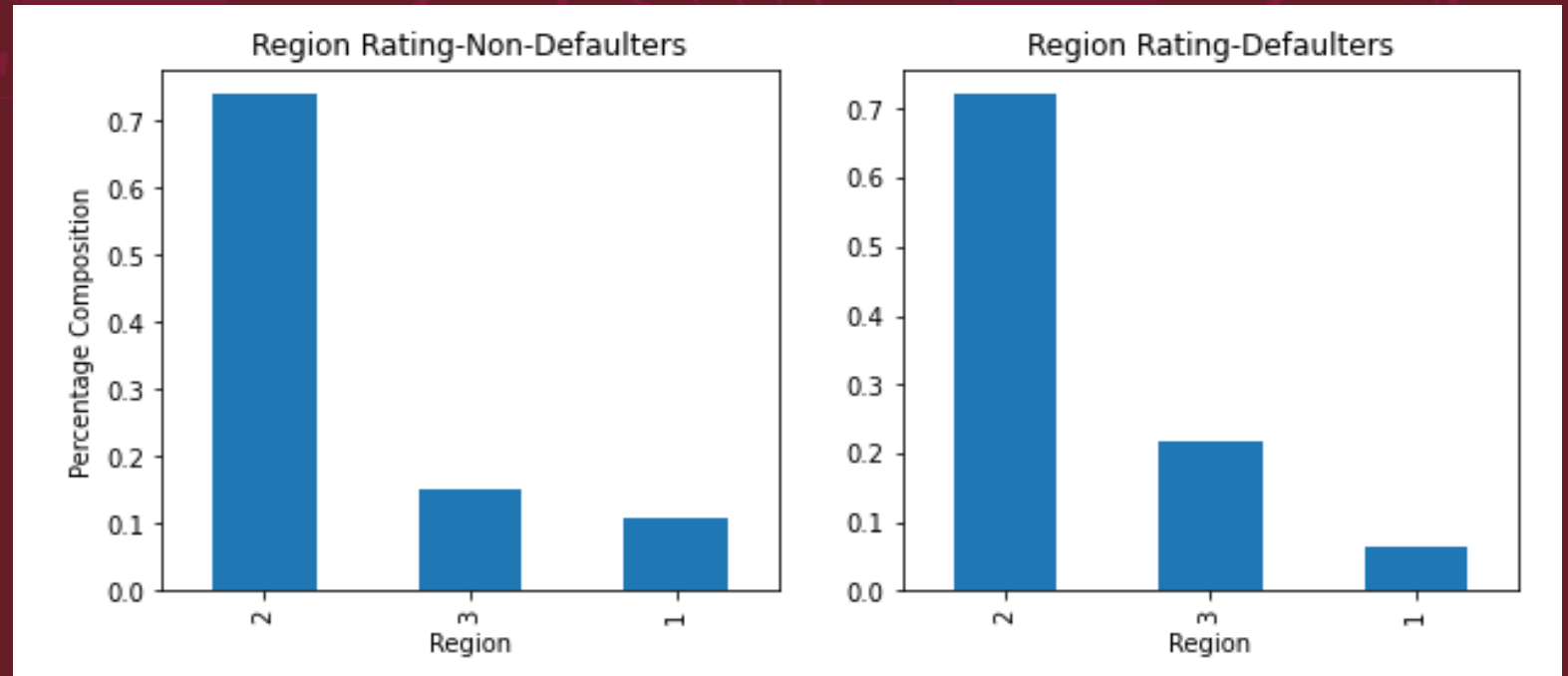
Inference

- The applications for credit are for age between 20 to 70
- We can conclude that people aged between 25-40 contains the majority of defaulters.

4. Region Rating



A 4.1 Region wise composition for overall Credit applicants



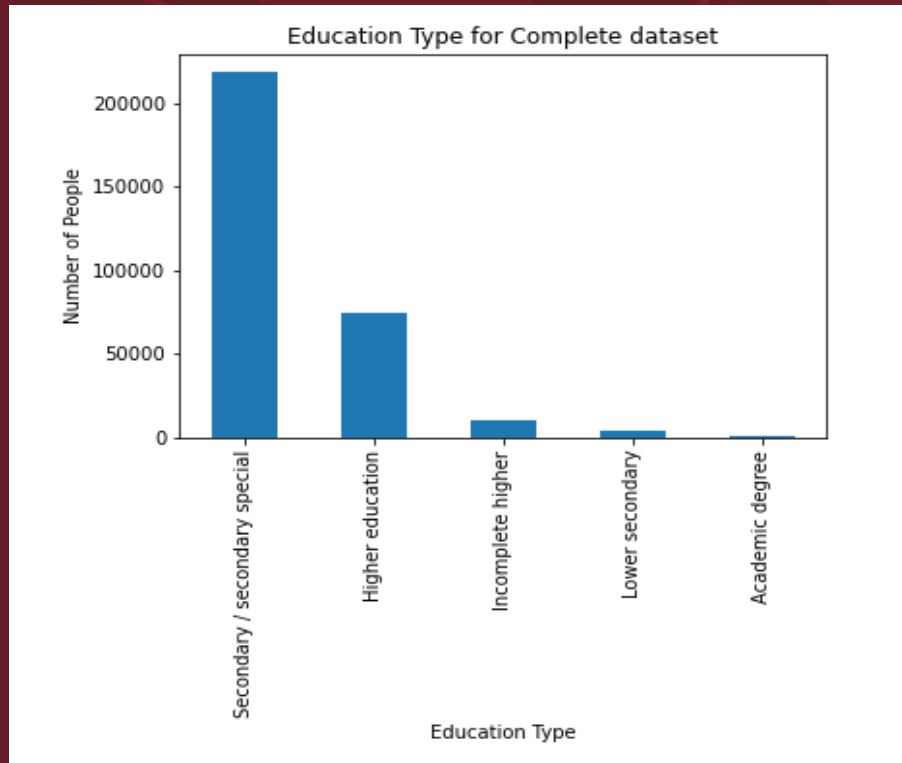
A 4.2 (a) Percentage distribution region wise for non-defaulters

A 4.2 (b) Percentage distribution region wise for defaulters

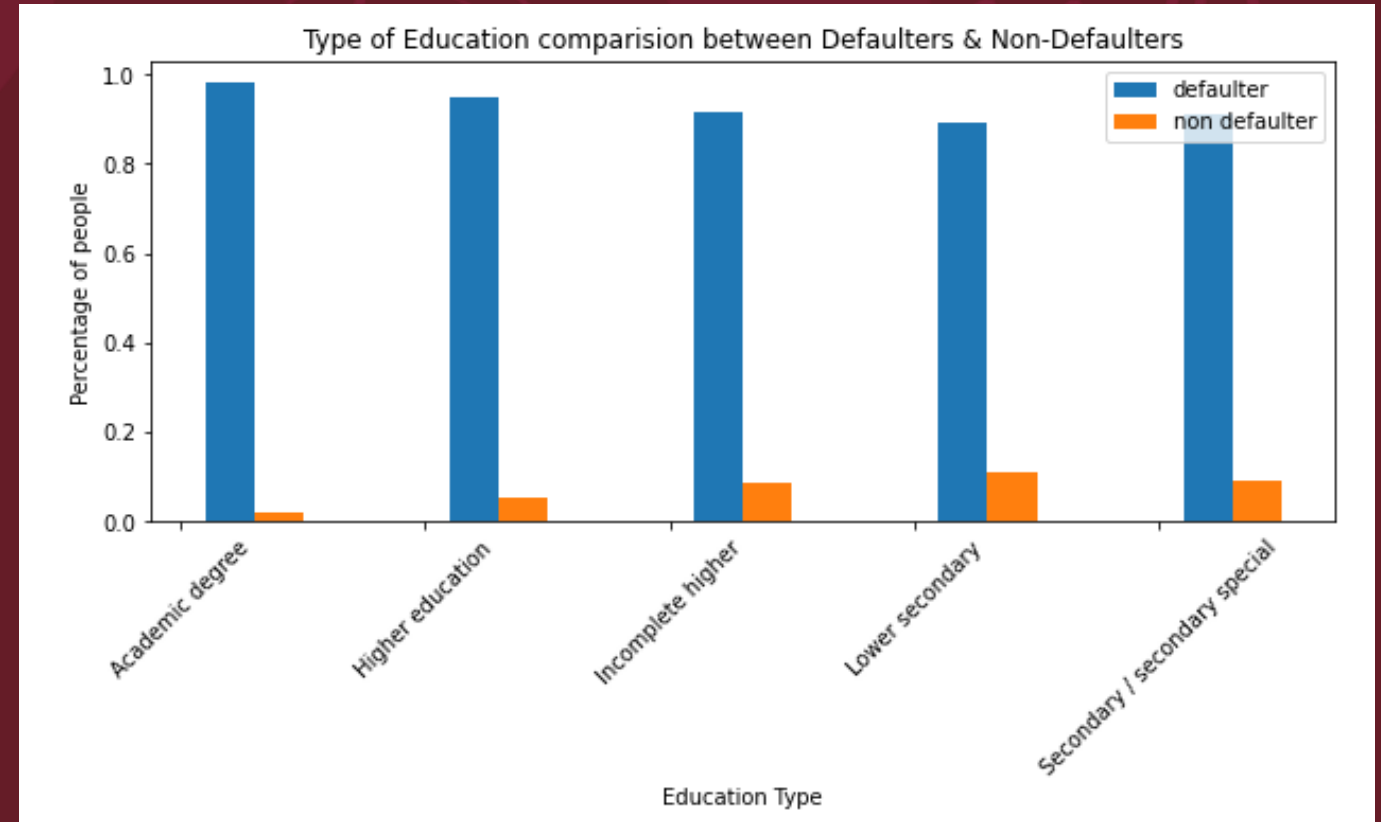
Inference

A quick observation can be made through the above plot that the percentage of defaulters to the total number is comparatively higher in case of region 2.

5. Education Type



A1 5.1 Education Type composition for overall Credit applicants



A1 5.2 (a) Individual percentage composition for Education Type for non-defaulters
A1 5.2 (b) Individual percentage composition for Education Type for defaulters

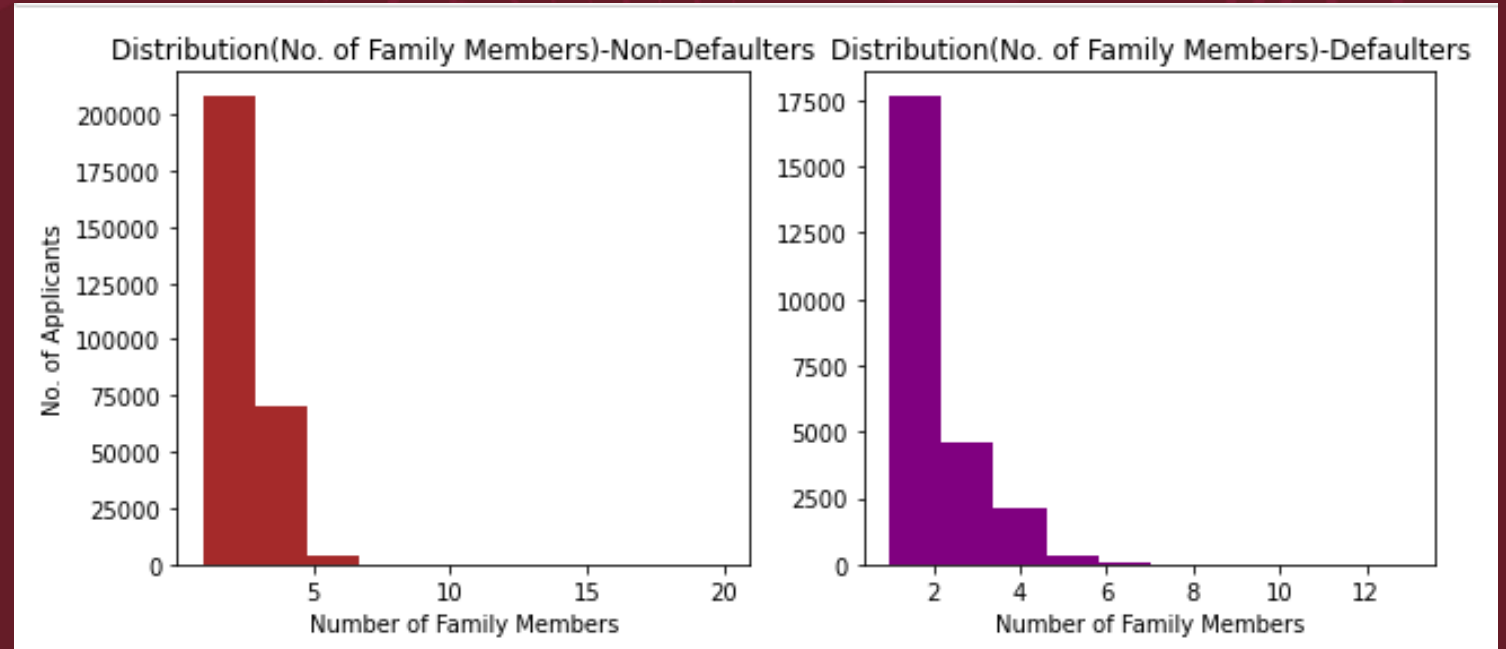
Inference

- Major applicants have Secondary level of education.
- People who either have not completed their higher or completed only secondary education have defaulted the most.

6. Family Members

Inference

- Major applicants are either single or have a family of 2.
- The number of defaulters are high for people with only 1 or 2 members as compared to Families having members between 3 and 5 as per the univariate analysis above.

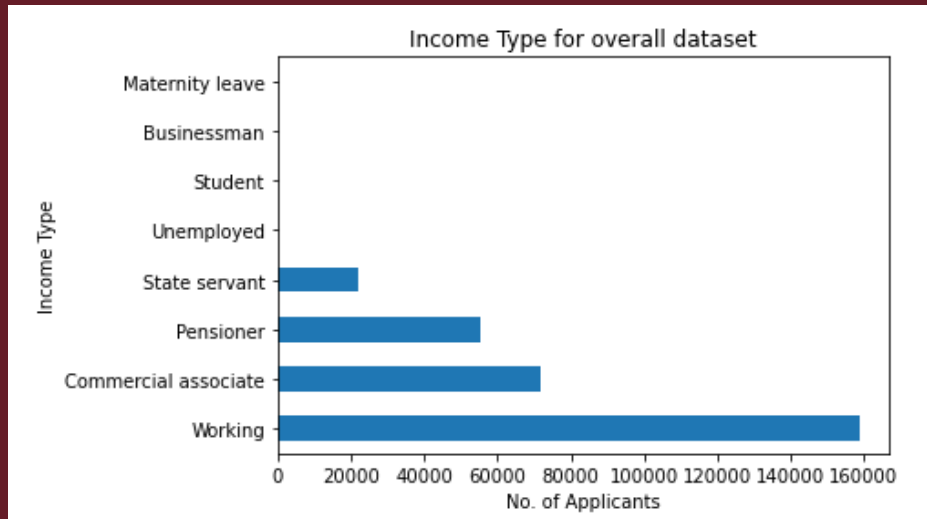


A1 6.1 (a) No. Of applicants vs number of family members for non-defaulters

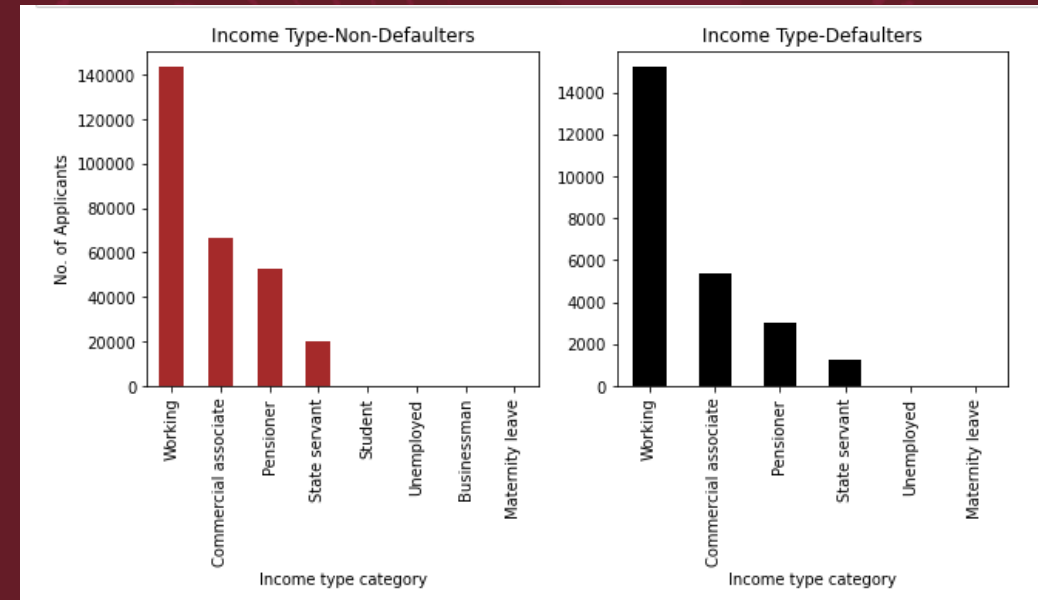
A1 6.1 (b) No. Of applicants vs number of family members for defaulters

2. Categorical Unordered Univariate Analysis

1. Income Type



A2 1.1 Income Type composition for overall Credit applicants



A2 1.2 (a) Income type composition for non-defaulters

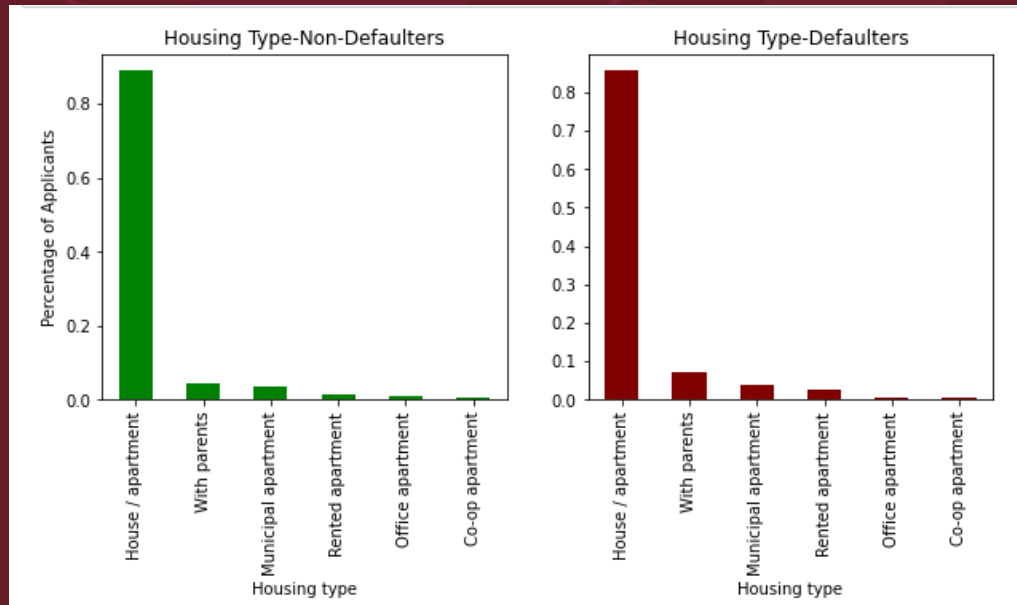
A2 1.2 (b) Income type composition for defaulters

Inference

- Working professionals, Commercial Associate, Pensioners are the top 3 applicants for loan
- Other professionals have quite less number.
- Working class has shown more defaults as compared to the other category of income sourced personals.
- The default pattern is proportional to the number of applicants

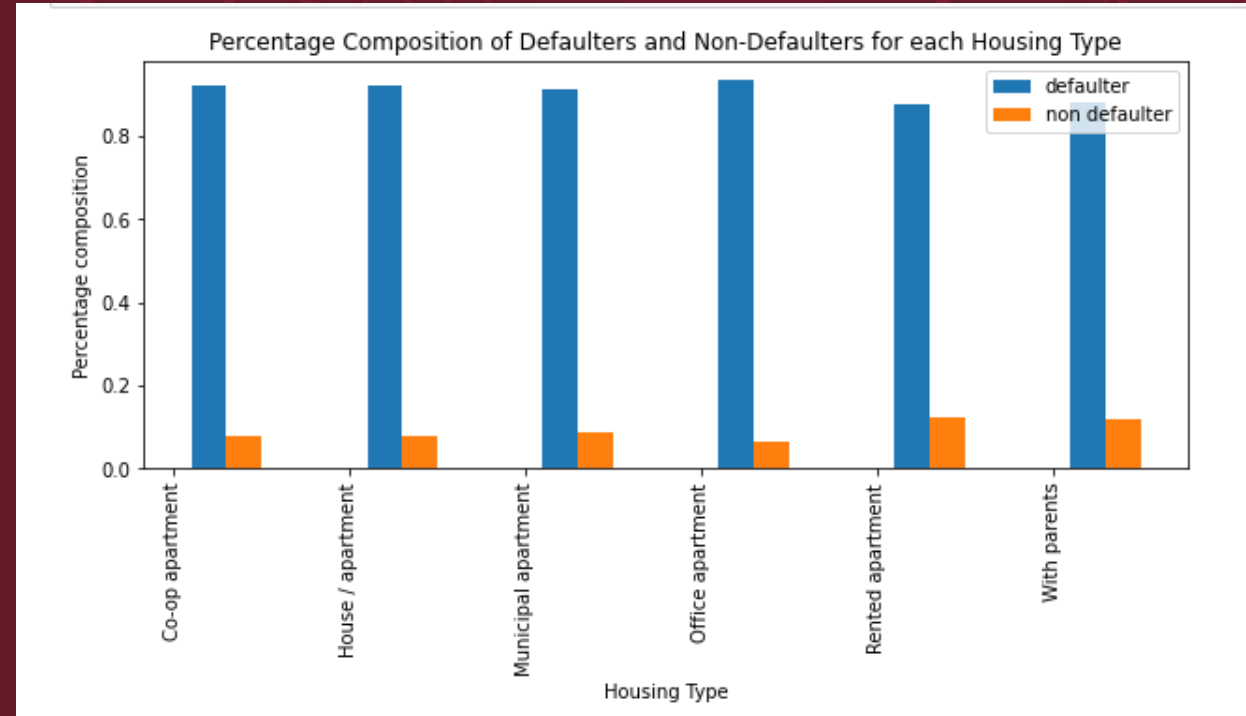
2. Categorical Unordered Univariate Analysis

2. Housing Type



A2 2.1 (a) Housing type percentage composition for non-defaulters

A2 2.1 (b) Housing type percentage composition for defaulters



A2 2.2 Percentage composition of each Housing type

Inference

Most of the people own a house/apartment apply for credit while People living in rented apartments have defaulted the most as compared to other housing facility residents.

B. BIVARIATE ANALYSIS

Bivariate Analysis is performed on the two data frames namely: Defaulters(Target=1) & Non-Defaulters(Target=0)

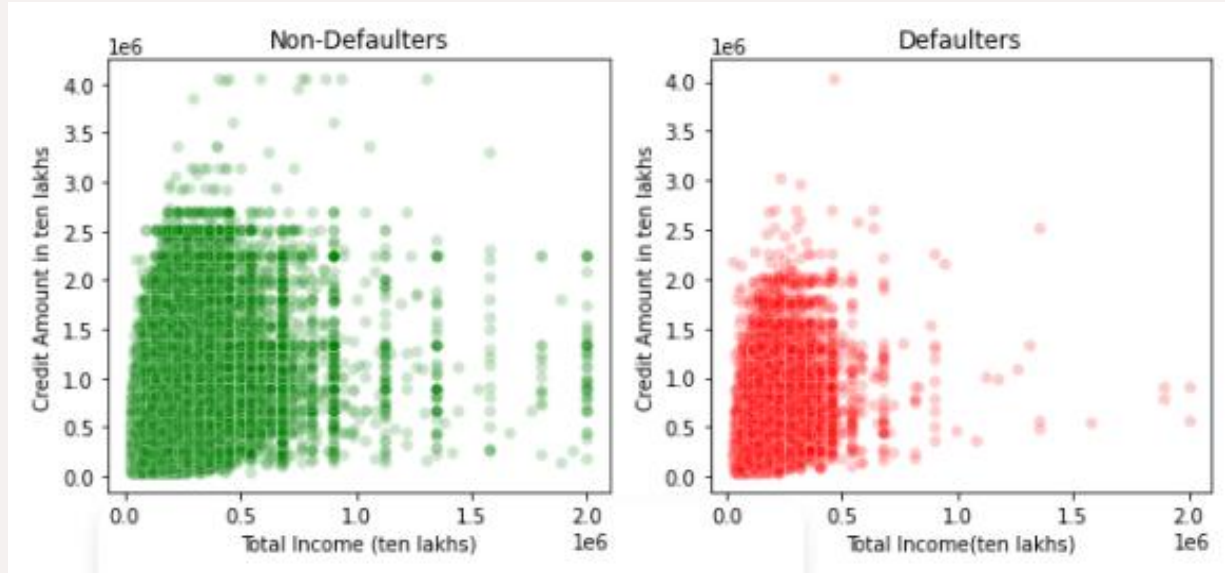
1. Numeric-Numeric Bivariate Analysis:

- ✓ Credit vs Income
- ✓ Annuity vs Income
- ✓ Credit vs Annuity
- ✓ Credit vs Goods Price
- ✓ Income vs Credit vs Annuity vs Goods Price (group plots)
- ✓ Income vs Credit vs Annuity vs Goods Price (Correlation Matrix -Heatmap)
- ✓ Income vs Credit vs Age (Correlation Matrix -Heatmap)

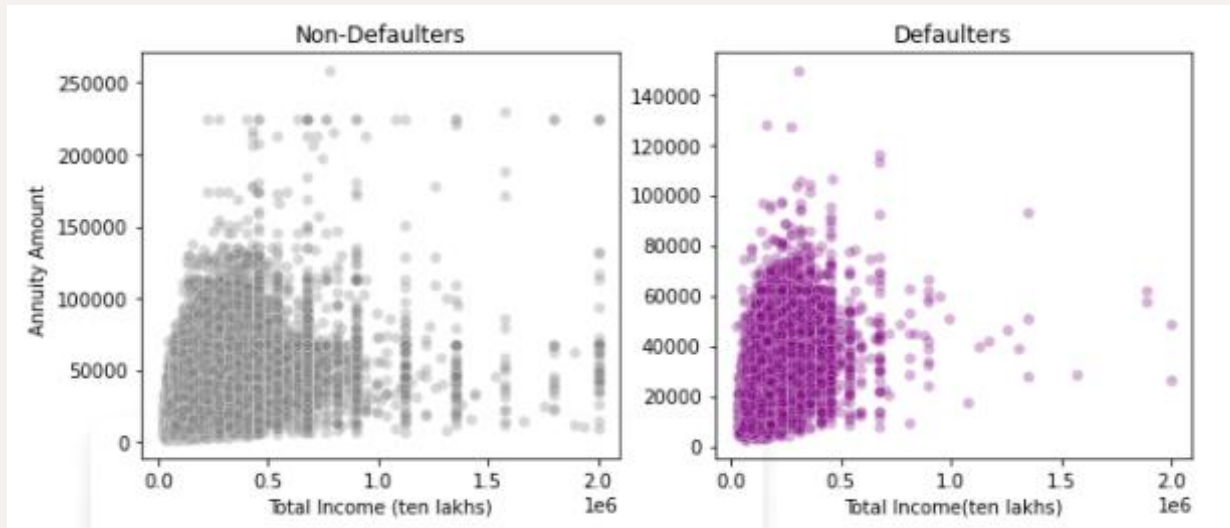
2. Numeric-Categorical Bivariate Analysis:

- ✓ Contract vs Total Income
- ✓ Education Type vs Income
- ✓ Income vs Income Type
- ✓ Credit vs Housing Type

Credit vs Income Amount



Annuity vs Income Amount



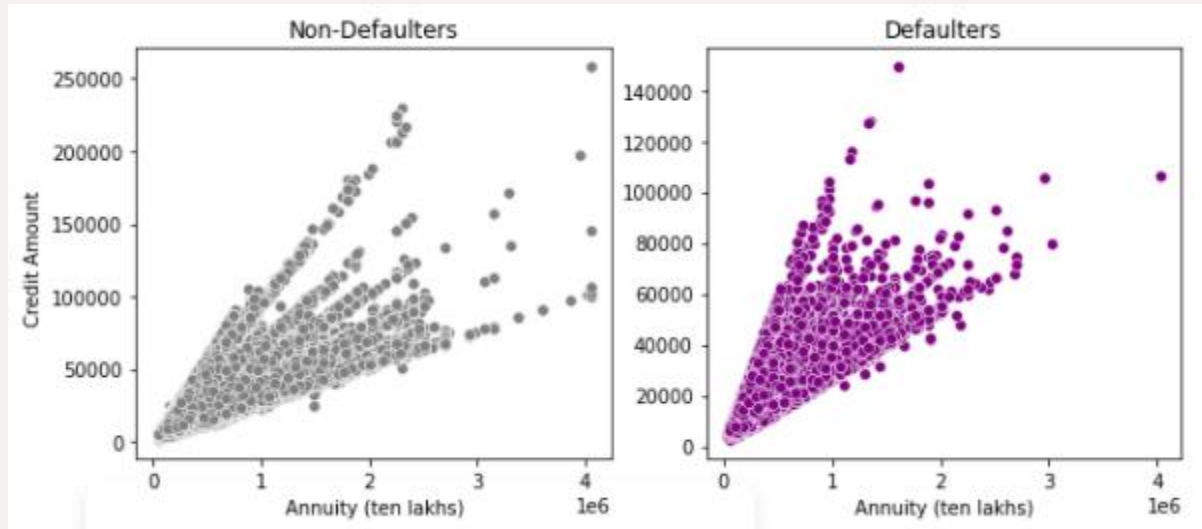
Numeric-Numeric Bivariate Analysis

Inferences

This analysis suggests us that, in terms of absolute numbers, most of the defaulters are having annual income less than 5 lakhs and they have taken credit less than 20 lakhs

With respect to annuity, within the same income brackets as above, customers having annuity less than 70000 have defaulted the most.

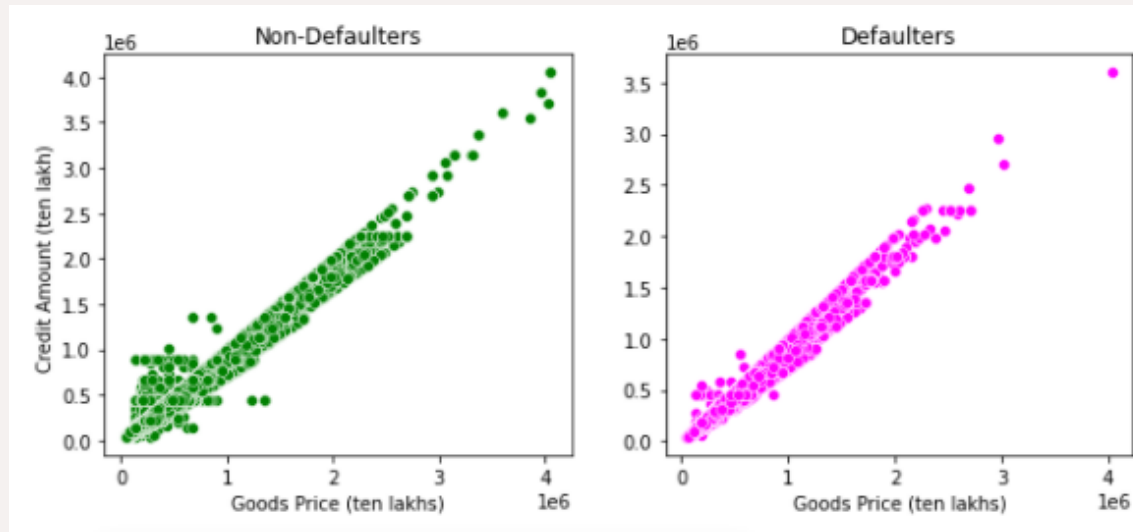
Credit vs Annuity Amount



Inferences

- The graph shows a strong correlation between Annuity & Credit amount
- A slight variation in Credit amount & Annuity between Non defaulters & defaulters can be seen.
- Defaulters have a comparatively lesser credit amount(1.4 lakh) and high credit have defaulted more as that of Non Defaulters (2.5lakh)
- A very strong correlation exist between Credit and Goods Price considering the fact that loan is approved based on the value of Goods Price.

Credit vs Goods Price

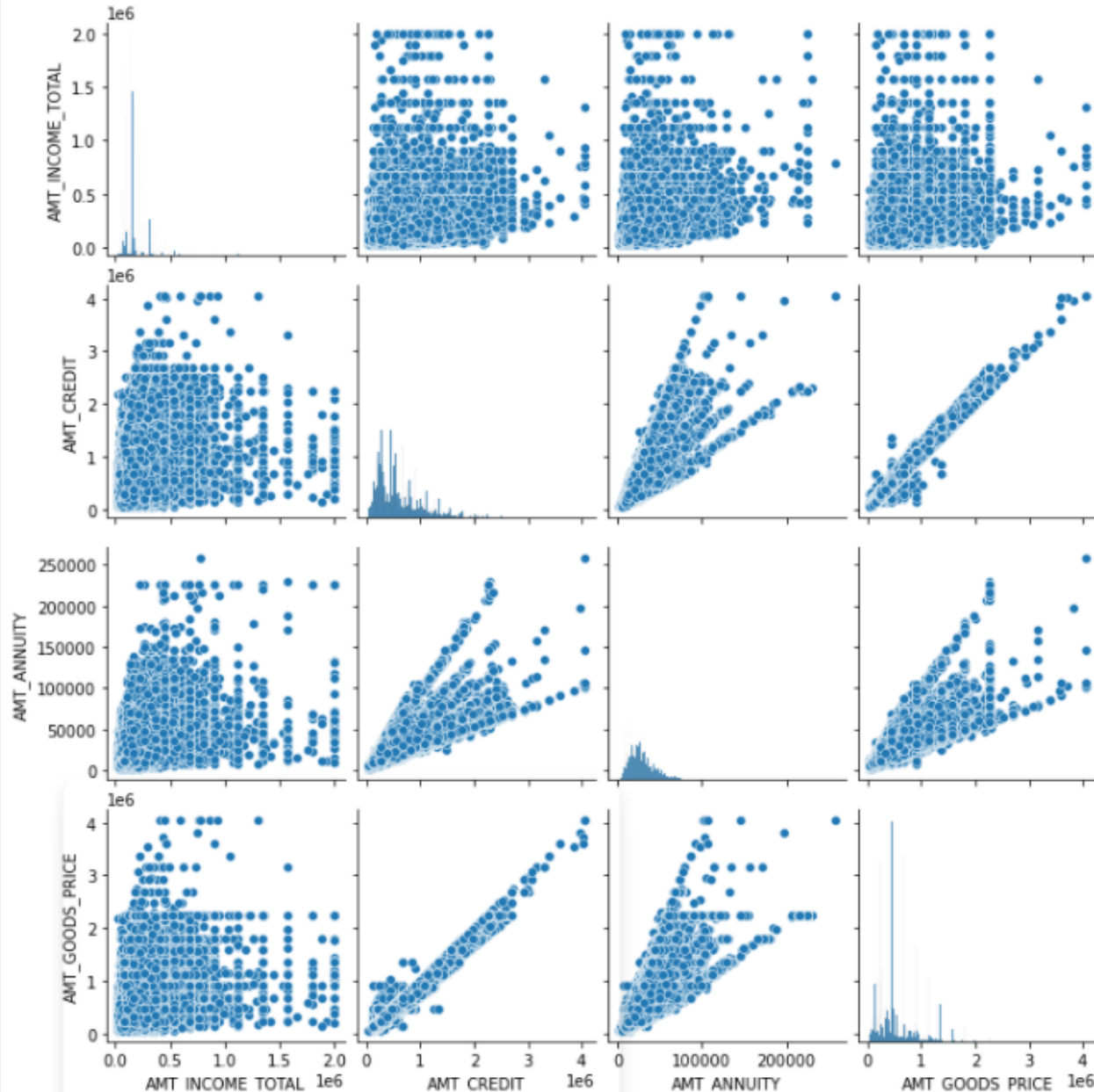


Income vs Credit vs Annuity vs Goods Price

Numeric-Numeric Bivariate Analysis

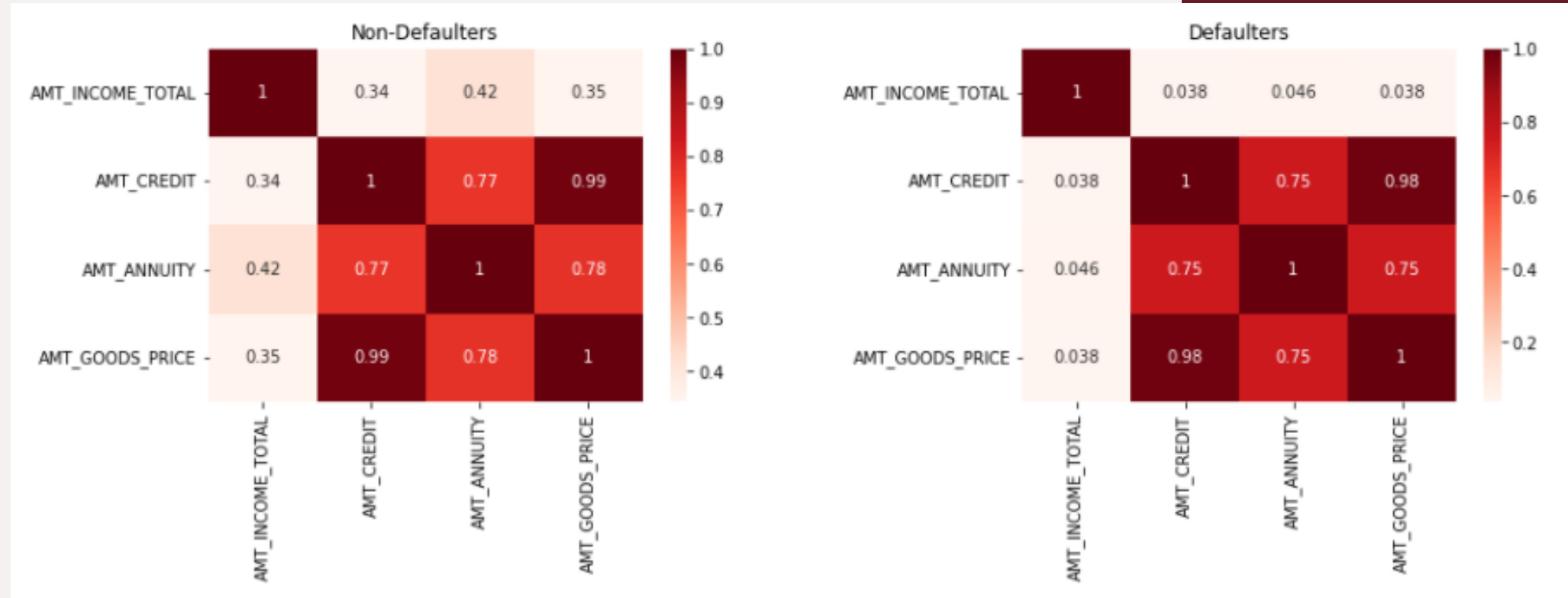
Inferences

- Good Price vs Income
The data point is spread across but densely populated for Income <5 lakh and Goods price below 20 lakhs
- The overall graphs suggest that major defaulters have income below 5 lakhs and credit less than 20 lakhs
- Majority of defaulters are those with Annuity less than 70000.
- For applicants with high income and high credit value, the number of defaulters are less but not zero.



Income vs Credit vs Annuity vs Goods Price

Correlation Matrix

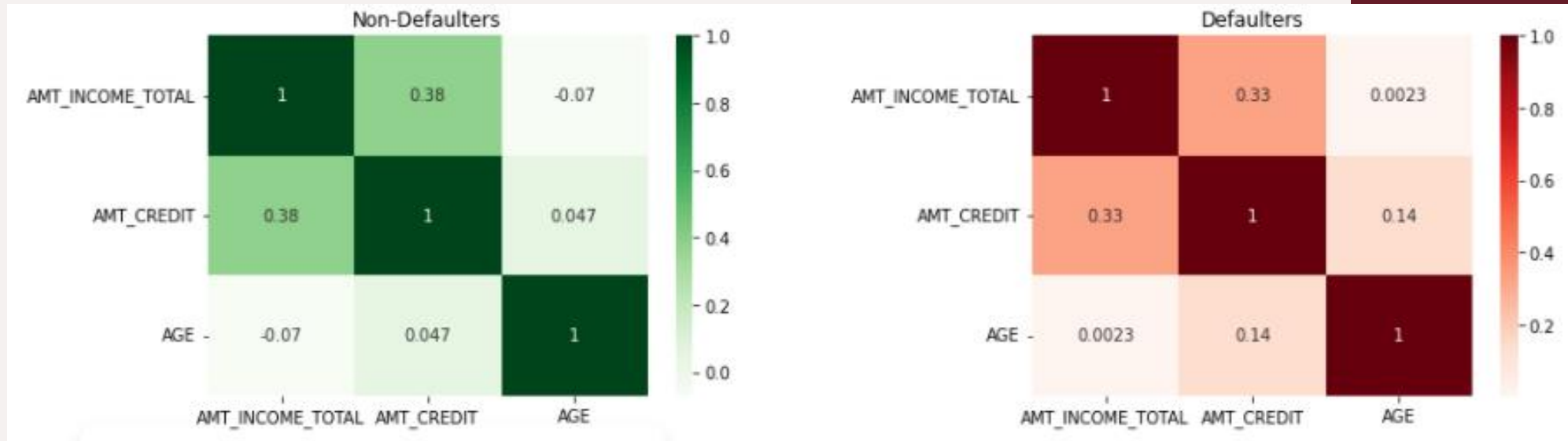


Inferences

- The Heatmap confirms our conclusions made in previous slide
- A very strong correlation value of 0.98 is seen between Credit and Goods Price
- A strong correlation 0.75 between Credit and Annuity & Goods Price and Annuity

Income vs Credit vs Age

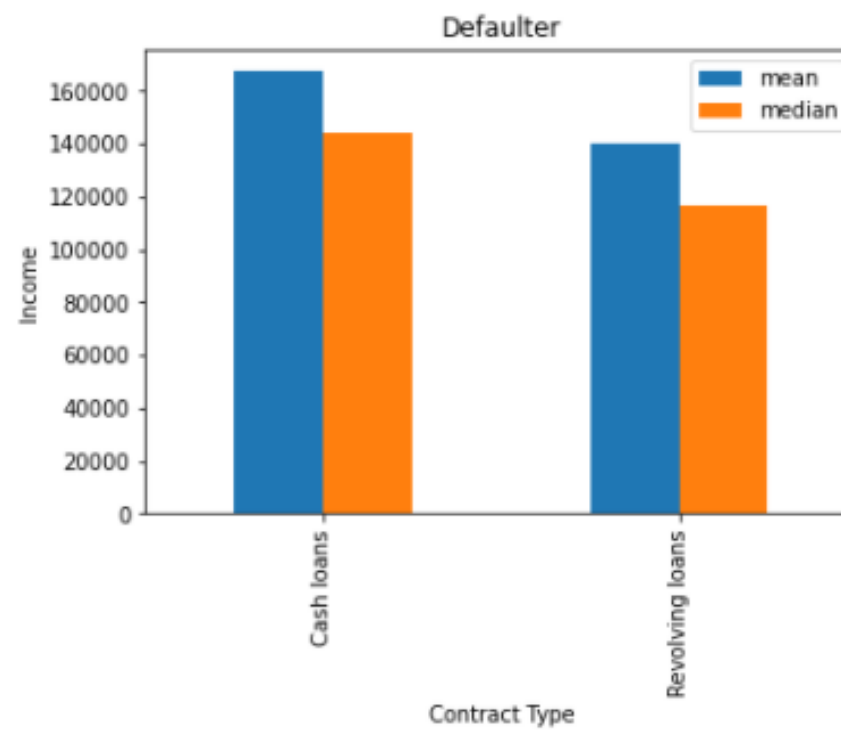
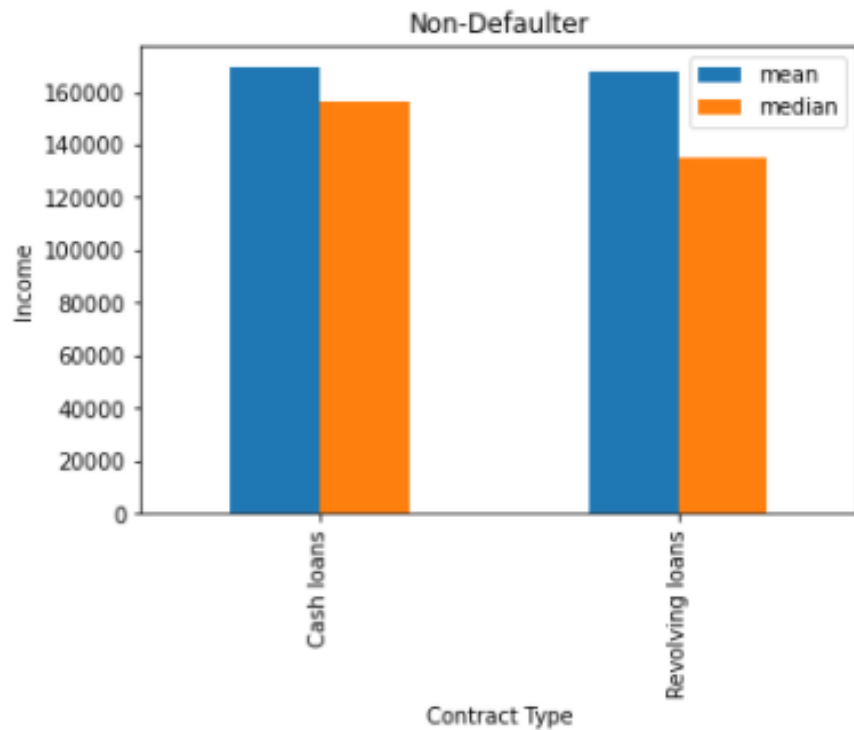
Correlation Matrix



Inferences

- The matrix shows that there is not much correlation between Income and Age or Credit and age for both defaulter / non defaulters
- In this case we cannot see any direct or linear pattern, there may or may not be any indirect pattern.

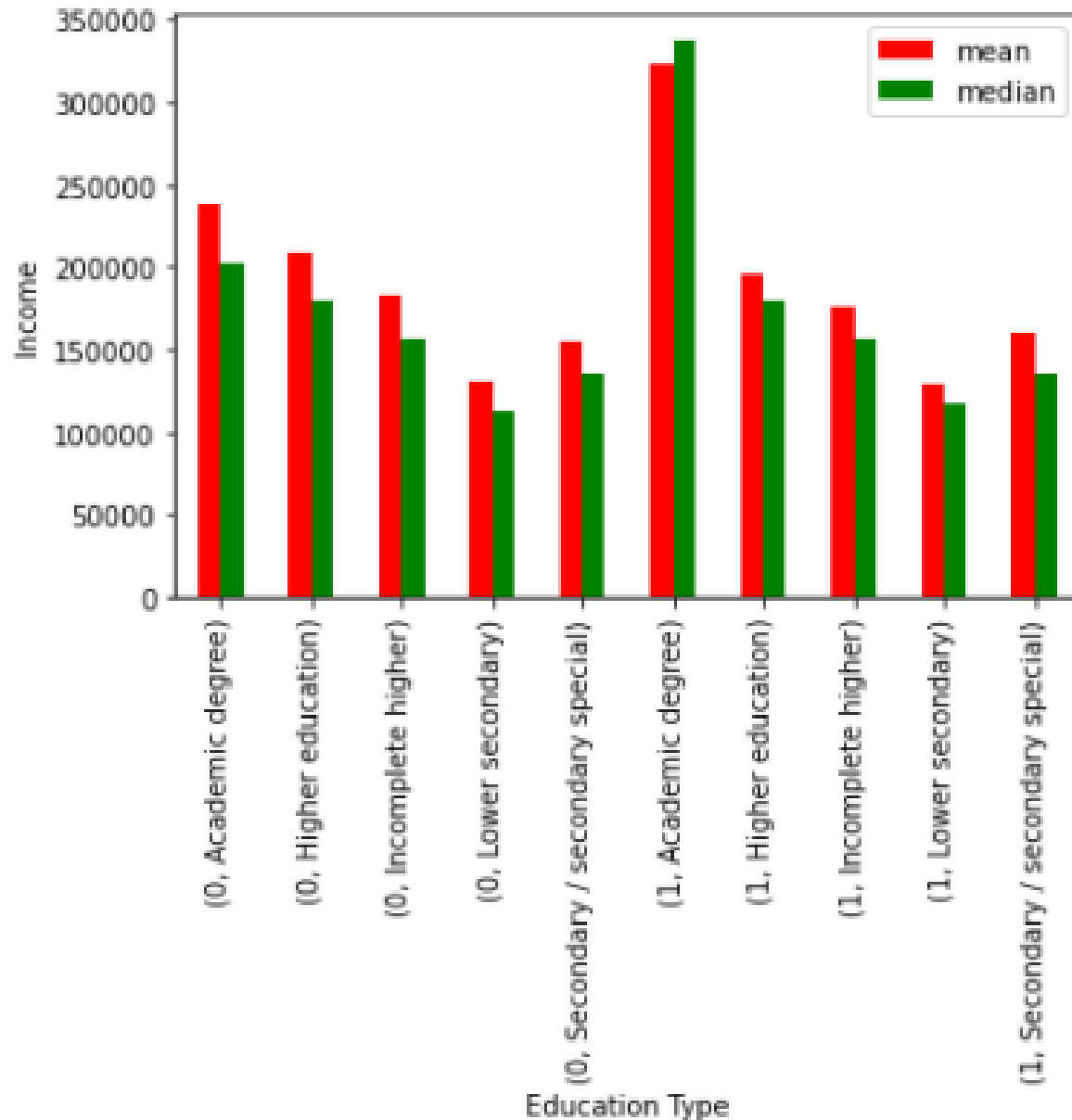
Contract Type vs Total Income



Inferences

- We can observe that the mean(1.7 lakh) & median(1.5 & 1.4 lakh respectively) for Cash loans is same in both cases
- But for Revolving loans the mean & median is comparatively less for defaulters(mean=1.4, median=1.2), non-defaulters (mean=1.7, median=1.4)
- Defaulters have Revolving loans of value between 1.7 & 1.1.

Education vs Income



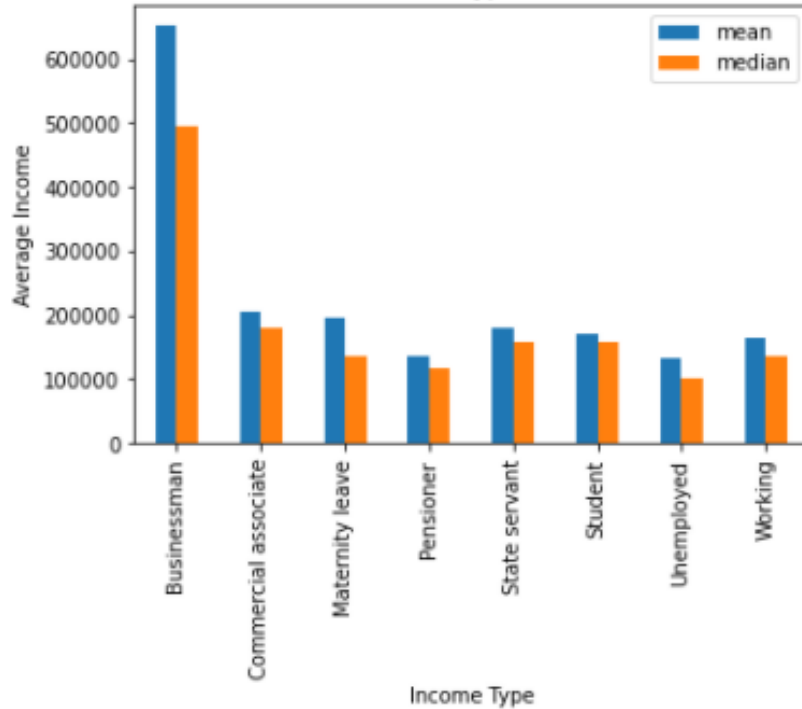
Numeric-Categorical Bivariate Analysis

Inferences

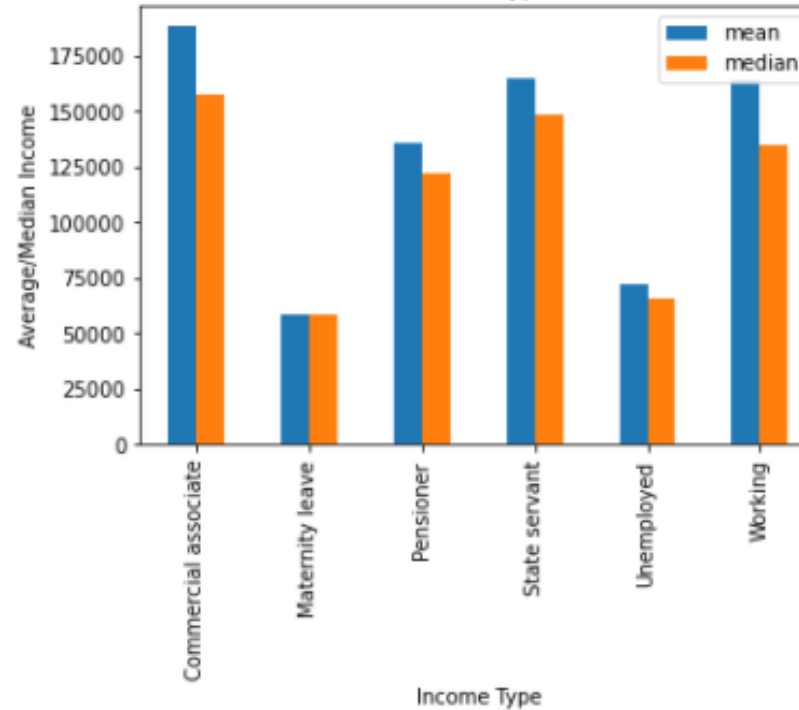
- For each class of Education, Academic degree holder are top defaulters with avg salary 2.4-3.2 lakhs
- The second highest defaulters are those with Higher education and avg salary 2 lakhs
- For applicants with education level below higher education avg salary between 1.75 & 2 lakhs ,face payment difficulties

Income vs Income Type

Income vs Income Type-Non-Defaulters



Income vs Income Type-Defaulters



Inferences

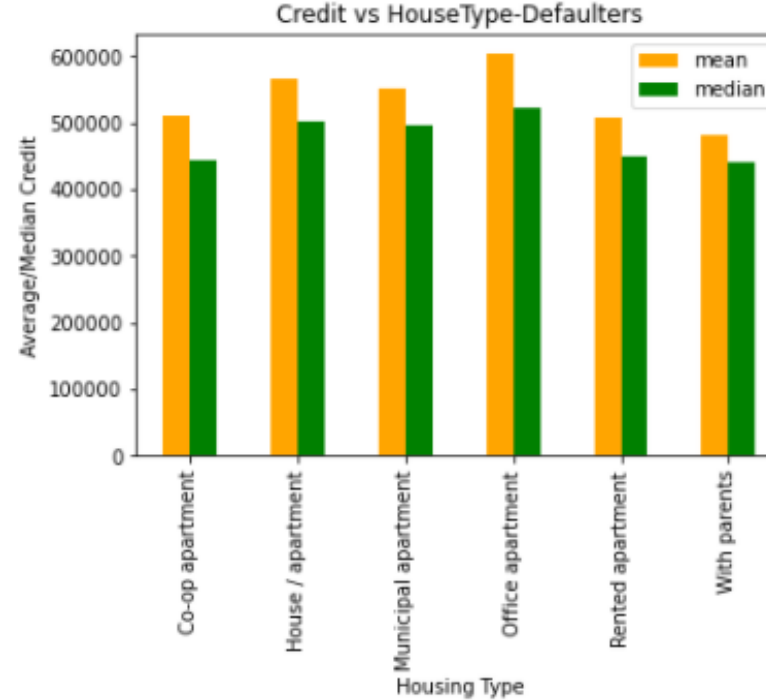
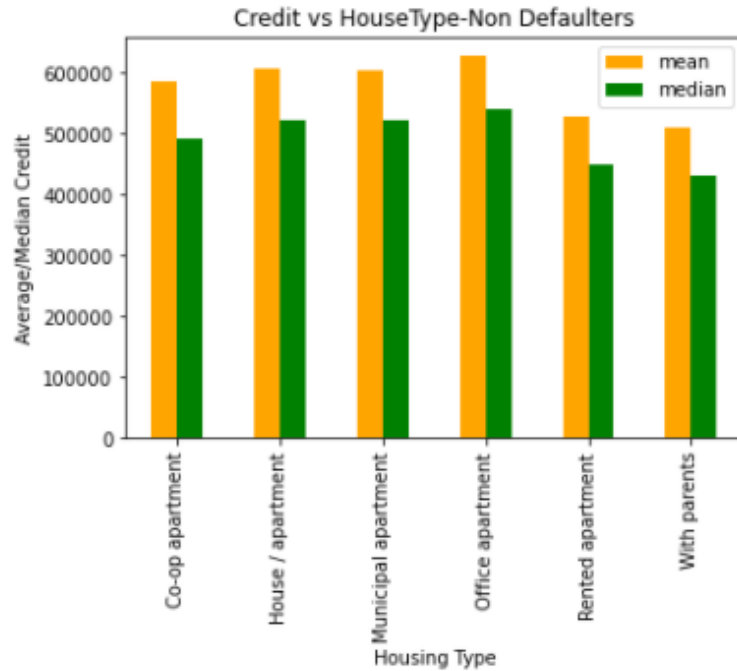
- Businessman have high avg Income and are not defaulters.
- Commercial Associates with 2lakh avg earnings are top defaulters.
- Working people & State servant with avg income near to 1.6 lakh are also major defaulters

Numeric-Categorical Bivariate Analysis

Inferences

- The defaulters are in a credit range of 4.5 to 6 lakhs
- Top defaulters having credit value of 5-6 lakhs and occupies office apartments, House/apartment, Municipal Apartments

Credit vs Housing Type



Numeric-Categorical Bivariate Analysis

Inferences

- The defaulters are in a credit range of 4.5 to 6 lakhs
- Top defaulters having credit value of 5-6 lakhs and occupies office apartments, House/apartment, Municipal Apartments

C. MULTIVARIATE ANALYSIS

Categorical-Categorical-Numerical Analysis is done using Heatmap for Multivariate Analysis

- ✓ Credit Group vs Income Group vs Target
- ✓ Family Status vs Income Group vs Target
- ✓ Housing Type vs Credit Group vs Target
- ✓ Region Rating vs Income Group vs Target
- ✓ Region Rating vs Credit Group vs Target

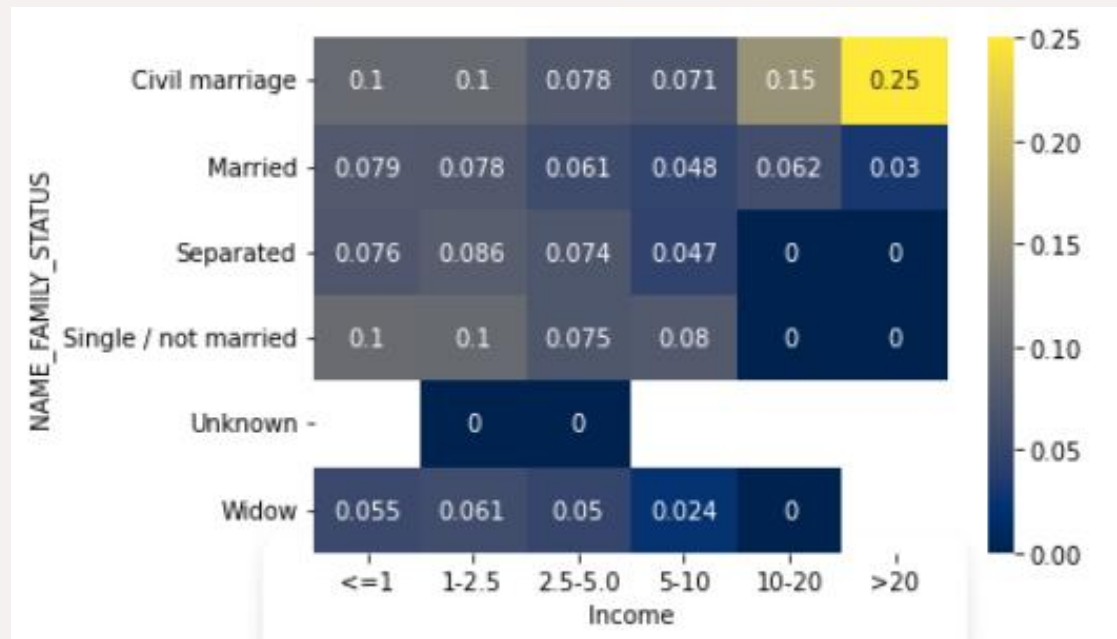
Income vs Credit vs Target



Inferences

- Customers having income less than 2.5L and have taken credit of more than 2L have defaulted the most
- Also, some customers who have an income of 5-10L and have taken credit of 1-5 have also defaulted but with deeper analysis, we come to know that the number is very low.

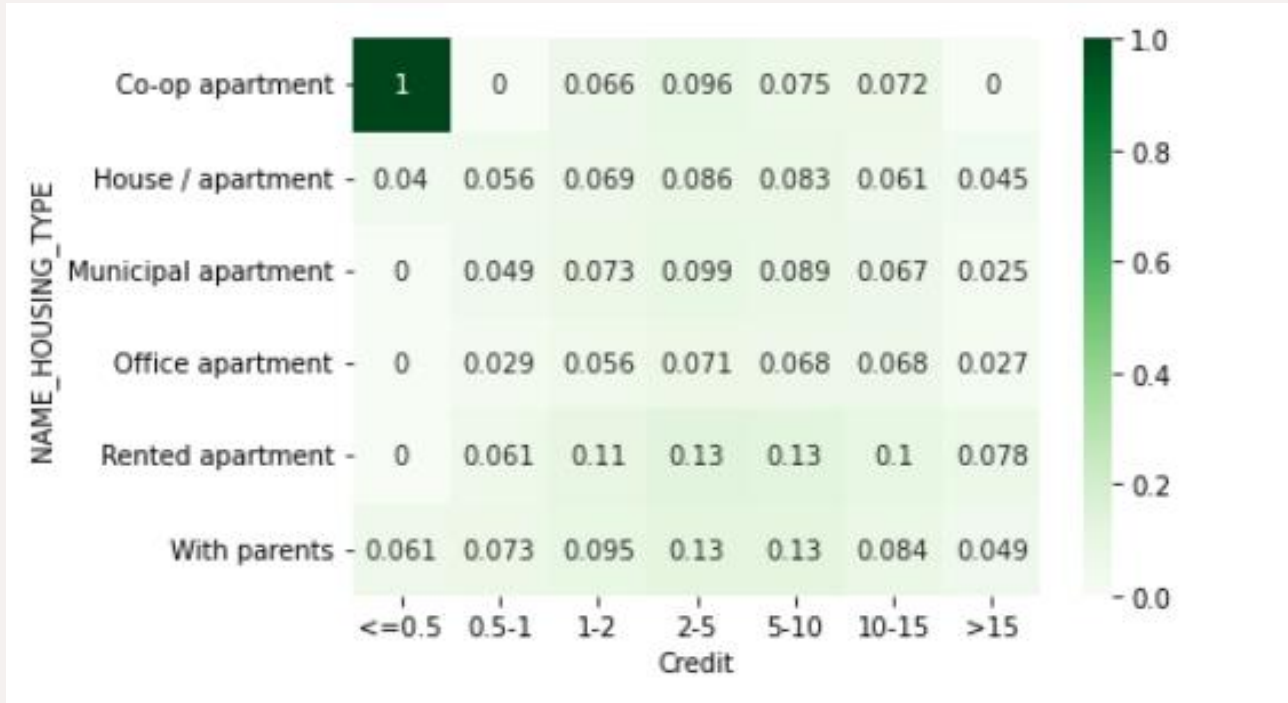
Family Status vs Income vs Target



Inferences

- Singles having income less than 2.5L are contributing most to the number of defaulters.
- Civil marriage customers having income less than 2.5L are also contributing to defaulters
- One strange point observed is that civil marriage customers having income >20L have the highest target value of 0.25, After digging deeper into numbers, we can see that there is only one such customer who is acting as an outlier.

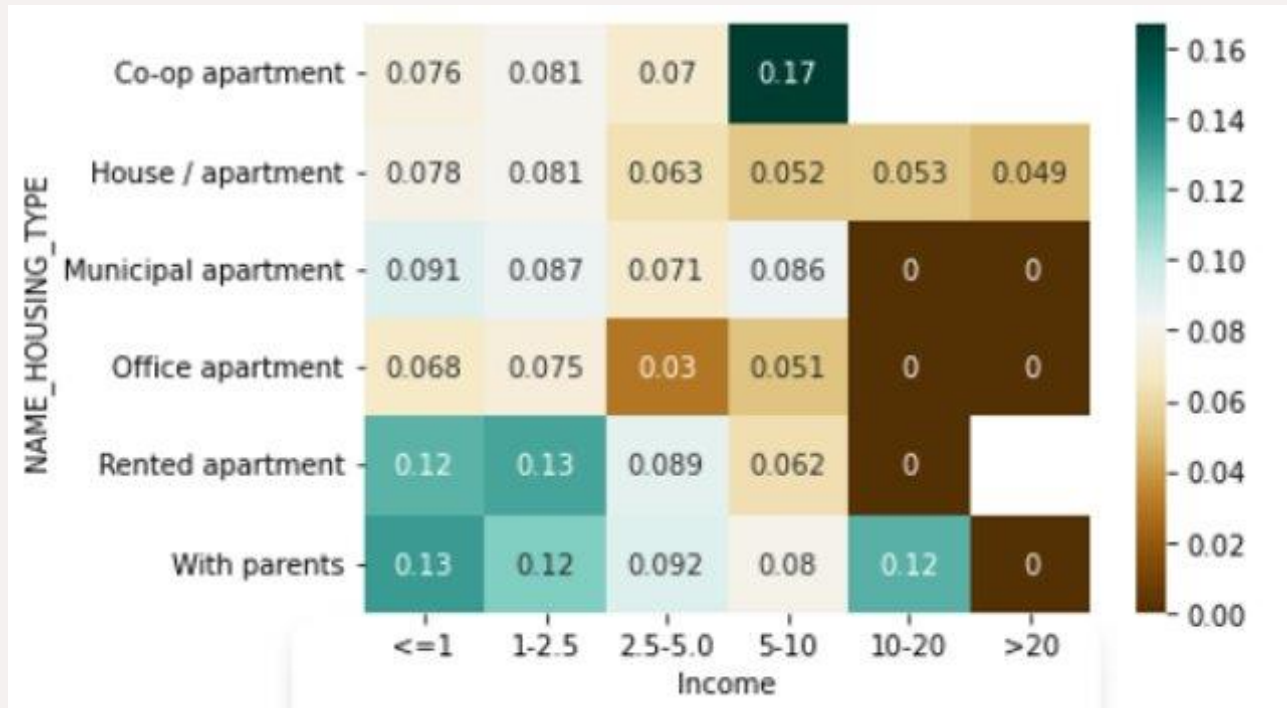
Housing Type vs Credit vs Target



Inferences

- All the customers having credit more than 1L and living in a rented apartment have defaulted the most
- Also, customers living in an Co-op apartment and having credit less than 0.5L have also defaulted

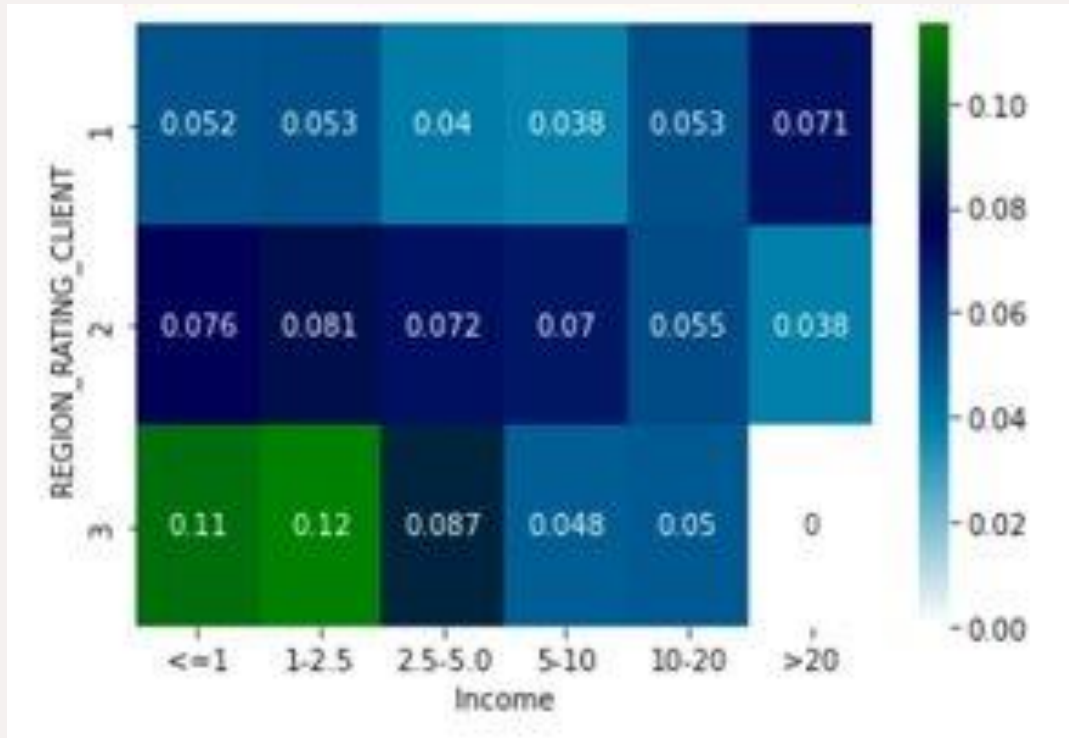
Housing Type vs Income vs Target



Inferences

- All the customers having income less than 5L and living in a rented apartment have defaulted the most
- Also, customers living in an Co-op apartment and having income more than 5L

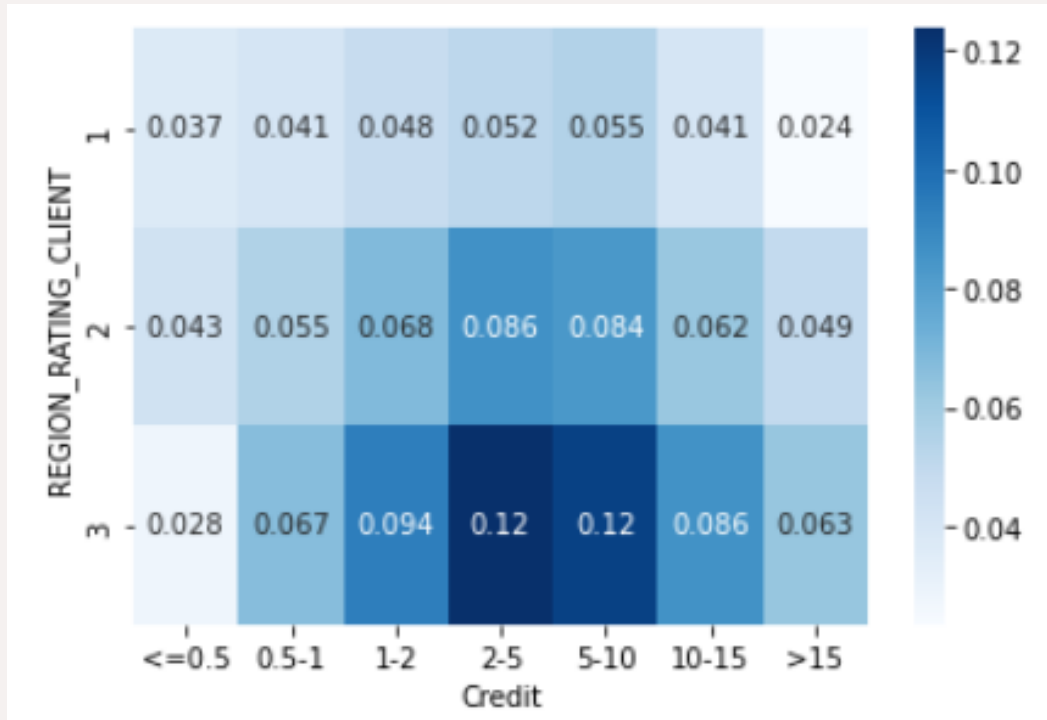
Region Rating vs Income vs Target



Inferences

- All the customers having income less than 5L and living in region 3 have defaulted the most
- Also, customers living in region 2 and having income less than 5L have also defaulted

Region Rating vs Credit vs Target



Inferences

- All the customers living in region 3 and taken the credit b/w 1 and 10L have defaulted the most
- Also, customers living in region 2 and having income b/w 2 and 10L have also defaulted

Conclusion

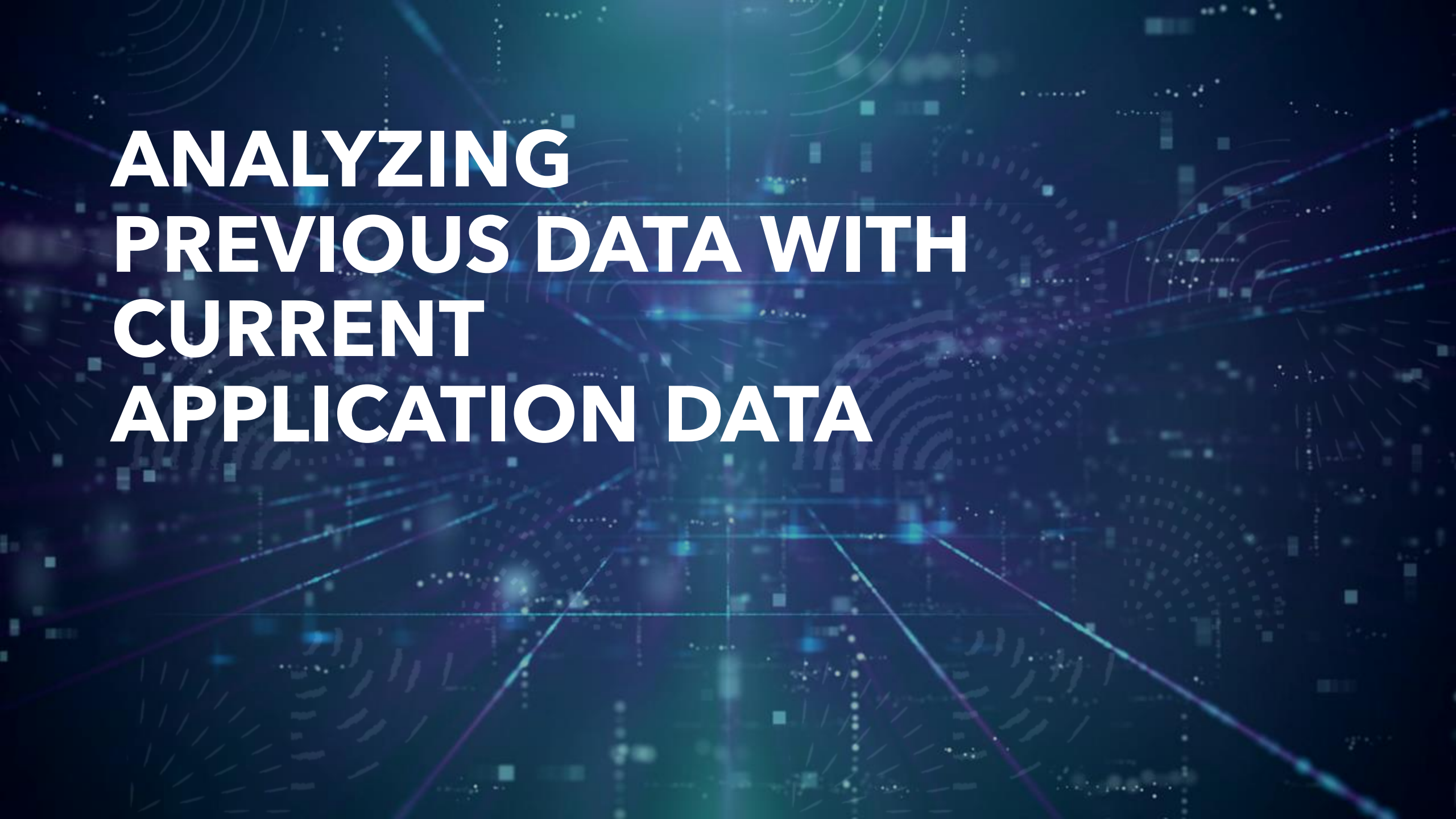
Through our univariate, bivariate and multivariate analysis, the following driving factors can be concluded:

1. In case of Univariate, Applicants with:

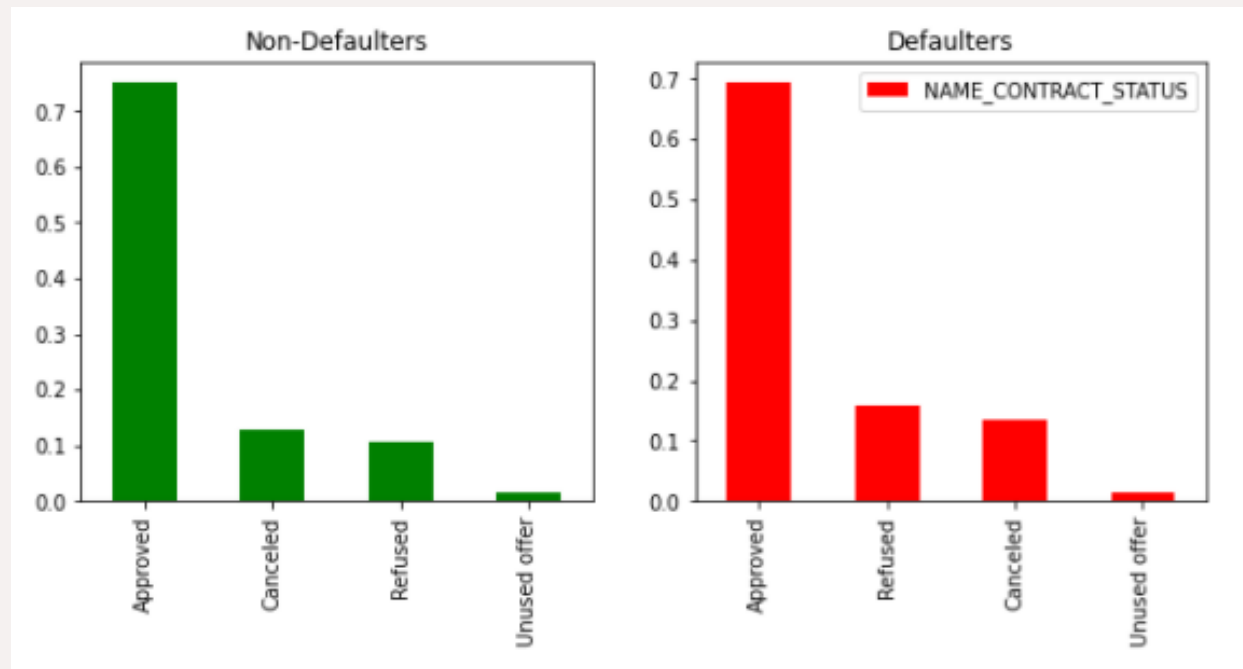
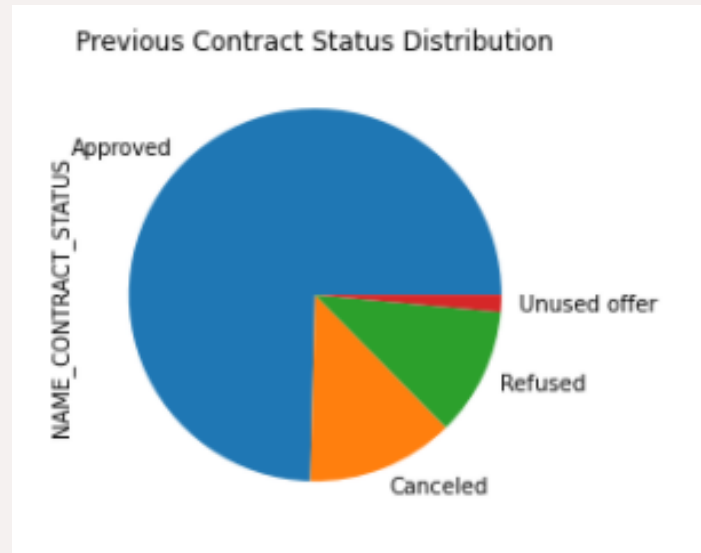
- ✓ Income less than 2.5L
- ✓ Credit b/w 2-10L
- ✓ Age b/w 25-40
- ✓ Region 2
- ✓ Secondary education
- ✓ Working class
- ✓ Rented Apartments

2. In case of Bi & Multivariate, Applicants with:

- ✓ Low to Avg. credit and high annuity
- ✓ Income less than 2.5L and credit more than 2L
- ✓ Singles and Civil marriages having income less than 2.5L
- ✓ Rented apartment and credit more than 1L and income less than 5
- ✓ Region 3 and Income less than 5L and credit b/w 2 to 10L

The background is a dark blue to black gradient, overlaid with a complex network of glowing lines and dots. The lines are primarily in shades of blue and purple, some straight and some curved, creating a sense of depth and movement. Small, bright blue and white dots are scattered throughout, resembling data points or stars. The overall effect is a futuristic, high-tech digital environment.

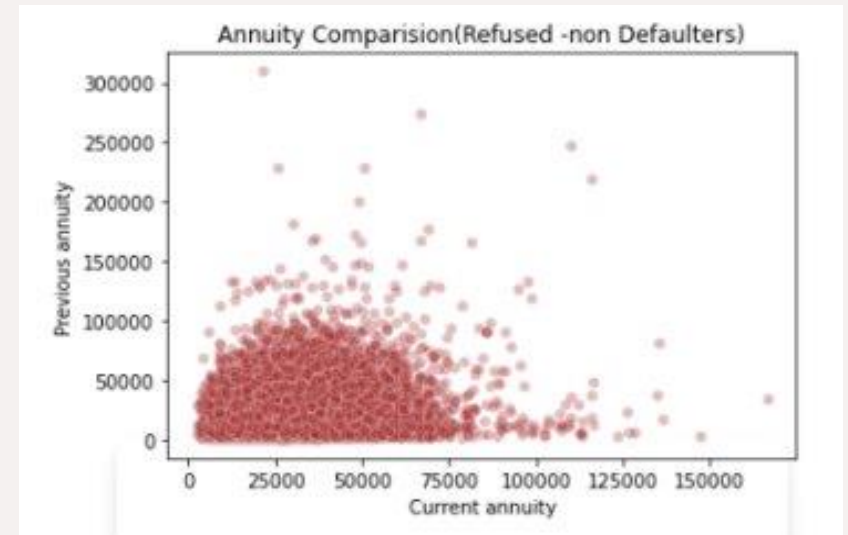
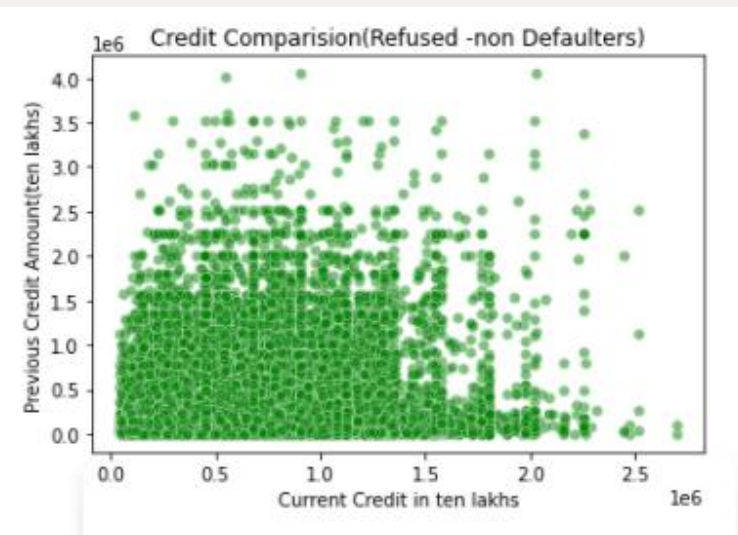
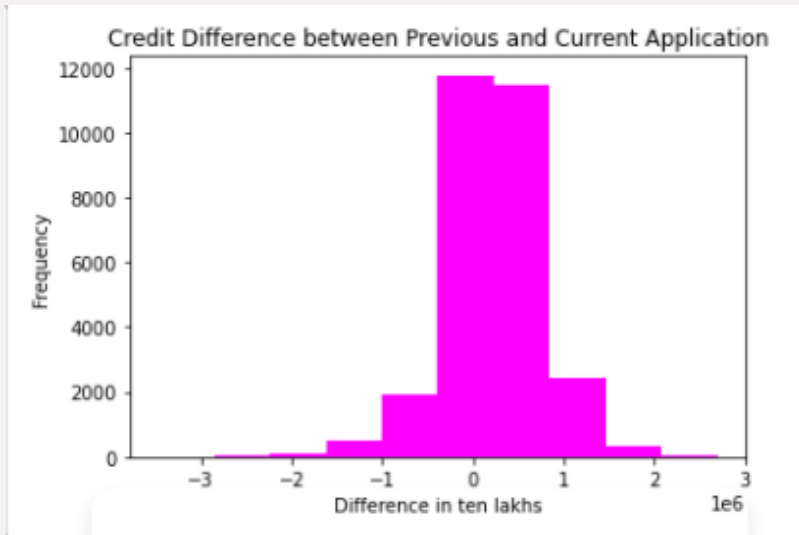
ANALYZING PREVIOUS DATA WITH CURRENT APPLICATION DATA



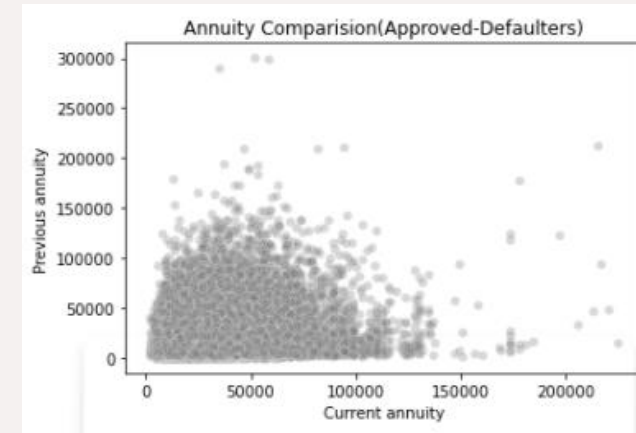
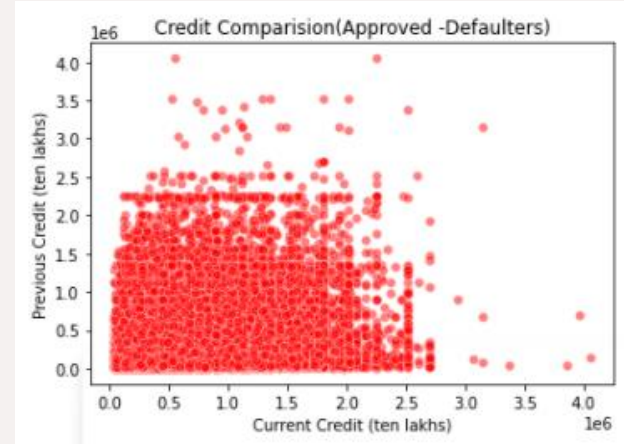
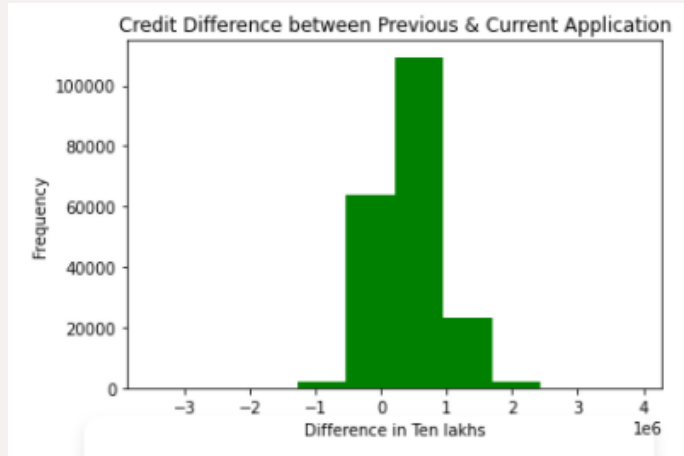
After merging the 2 dataset and analyzing the data for Applications which were previously refused/approved leading to loss in business.

Inferences

- Most of the applications were approved. Out of those around 11% of past Refused applications are non-defaulters while 69% previously Approved applications are now having payment difficulties



- There are around 29 thousand customers who were refused loans previously, It has resulted into a loss of 665 Crores as a difference of credit between previous application and current application
- Applicants with lower current annuity are in the non-defaulters category



- There are around 2 lakh customers who were approved loans previously, Out of these around 91% has a current loan with increased loan amount.
- On comparing the Credit difference for those 91% applicants, it has a difference of 9006 Crores
- For those 9% with Previous loan amount less than Current credit amount, leads to a difference of 507 Cr
- The applicants with higher annuity are in the defaulters category as compared to non-defaulters

Final Conclusion

Through our univariate, bivariate and multivariate analysis, the following driving factors can be concluded:

1. As described in the previous conclusion slide, through multi variate analysis, we could gather major driving factors which should be kept in mind while processing a loan:

- ✓ Income less than 2.5L and credit more than 2L
- ✓ Singles and Civil marriages having income less than 2.5L
- ✓ Rented apartment and credit more than 1L and income less than 5
- ✓ Region 3 and Income less than 5L and credit b/w 2 to 10L

Applicants flagged under such cases should either be given less credit, low annuity or a similar control measure.

2. Conclusion drawn from comparing data of current and previous applications:

- ✓ The defaulters with approved loans can become non defaulters is their annuity is decreased
- ✓ A lower credit amount will also reduce the number of defaulters
- ✓ Interest rate & Credit amount should be monitored with respect to the total income.