Venkata Ramana Puppala (DS C- 039)

Credit EDA Assignment

Credit EDA Assignment - Contents

1. Understanding the Domain

2. Loading the data

3. Checking Metadata of the data

4. Missing value Check and Treatment

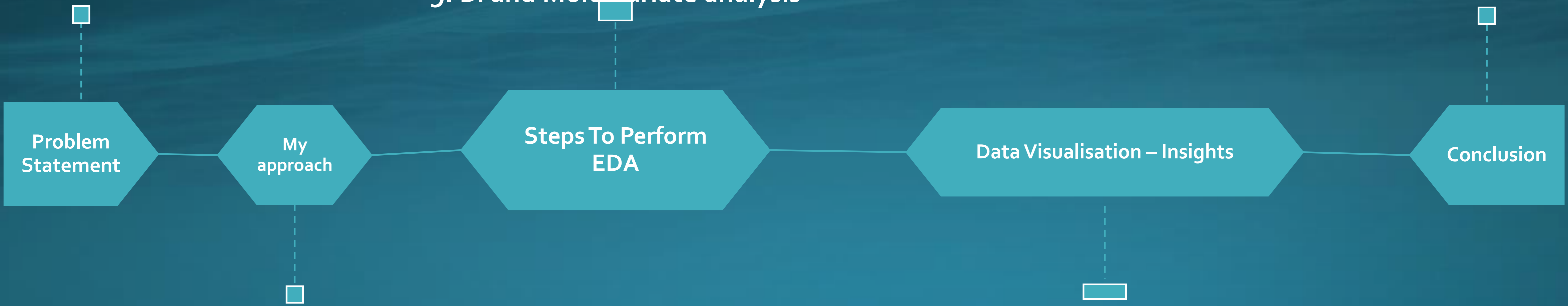
5. Outliers check and Treatment

6.Data Imbalance Check

7. Univariate analysis

8. Segmented Univariate analysis

9. Bi and Multi variate analysis



Problem Statement

My approach

Steps To Perform EDA

Data Visualisation – Insights

Conclusion

1. Box plot, 2. Count plot, 3. Bar plot, 4.Pie plot, 5. Distribution plots, 6. Scatter plot , 7. Heat map

1. Problem Statement:

- As the Fintech Companies, it is difficult to provide the loans to those customers because unavailability of Credit history with them. So, There will be two possibilities of loss making to the companies . They are:

  1. Giving the loan to the customers who are likely to default.
  2. Not giving the loan to the customers who are likely to repay.

- So, Exploratory Data analysis (EDA) is used to find the patterns in the data which will help in deciding the whether to give loan or not. To bring some insights, proper analysis need to be done on the target attributes (Univariate analysis) and Effect of other attributes on the target variables (Bi and Multi variate analysis).

2. My Approach:

1. **Understanding the Business Objective and Problem Statement**
2. **Understanding the Attributes of data set given**
3. **Understanding the meta data and performing Data wrangling Techniques and**
4. **To draw insights with help of Data Analysis.**

3. Steps to perform in EDA:

I. Understanding the Domain

II. Load the Data

III. Checking the Metadata of the Data\

IV. Missing values Check and Treatment

V. Outliers Check and Treatment

VI. Data Imbalance Check

VII. Univariate Analysis

VIII. Segmented Univariate Analysis

IX. Bi and Multi Variate analysis

3.1 Data Cleaning:

- Data Cleaning includes Checking the null values , Dropping the Columns, Dropping the Duplicate rows and Imputing the null values appropriately depends upon the type of the variable (Categorical or Numerical Variable).
- **Dropping the Columns:**

  Dropping the columns which are having the 40% of missing values.
- **Imputing the Rows:**
  - Treating the Occupation type:

    Creating the new category called “Others” where values are missing as the percent of missing is 31 , I prefer adding new category instead of replacing with “mode”.
  - Treating the “EXT_SOURCE_3:
    - Imputing the missing value with mean as there are no outliers and difference between mean and median is also less.
  - Treating the AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_HOUR :
    - Imputing the missing values with Median since data has more outliers and distributed unevenly.

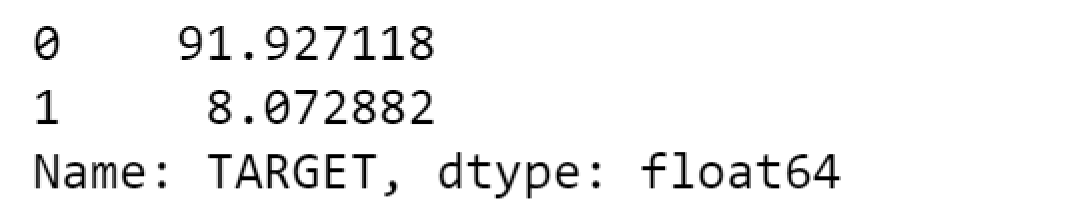
3.1 Data Cleaning:

- Treating the OBS_30_CNT_SOCIAL_CIRCLE,DEF_30_CNT_SOCIAL_CIRCLE,OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE:
    - Imputing the missing values with Median since data has more outliers.
- Null values are checked on all columns to make sure data is processed perfectly.

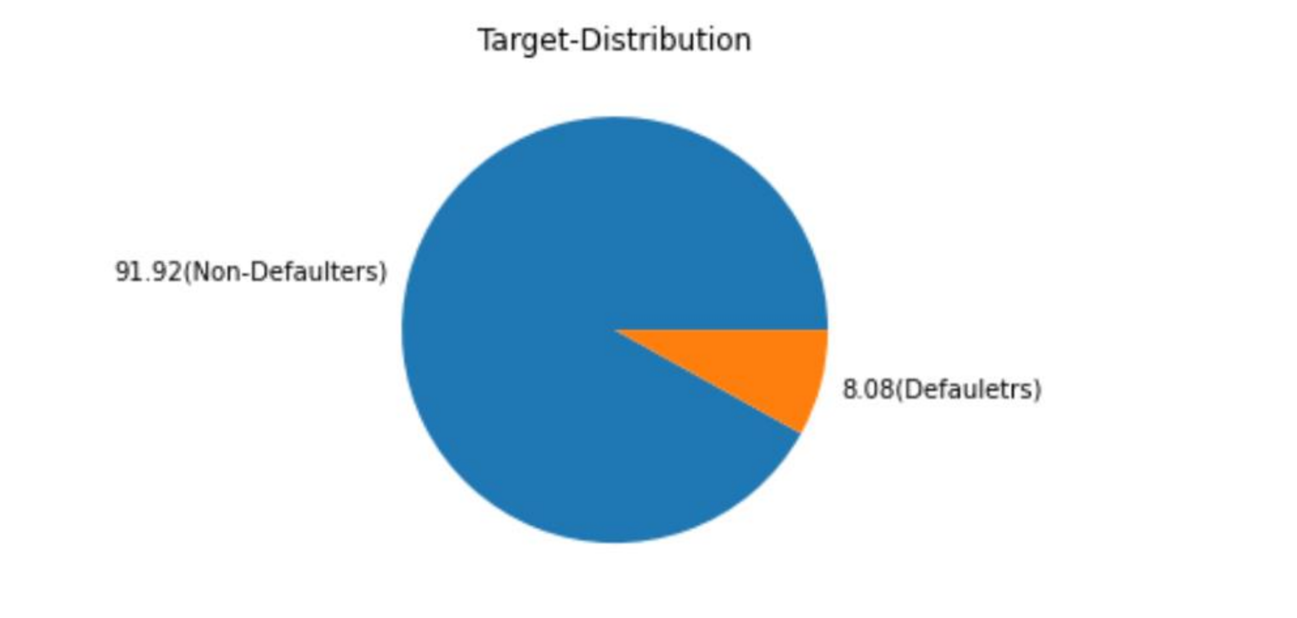
- **Outliers Check and Treatment:**
    - Numerical Columns like AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, FLAG_DOCUMENT_18, FLAG_DOCUMENT_19, FLAG_DOCUMENT_20, FLAG_DOCUMENT_21, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR are capped and floored with the help of **for loop.**
    - Box plots are applied on the different numerical columns to check the Outliers data in dataframe.

3.2 Data Imbalance:

- Data imbalance is checking the Proportion of Categorical variable distributed. In this assignment Target Column is Checked for Data imbalance.

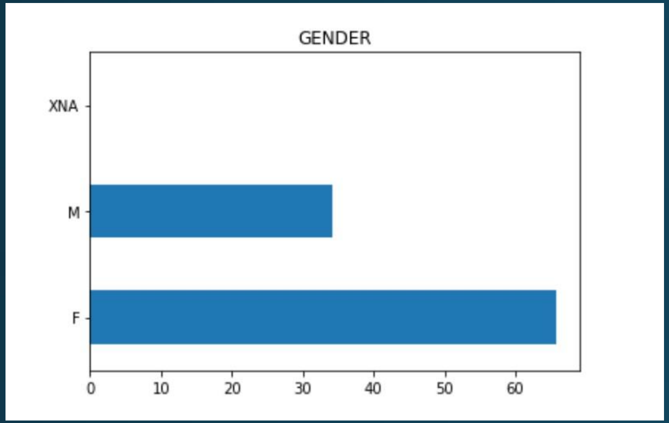


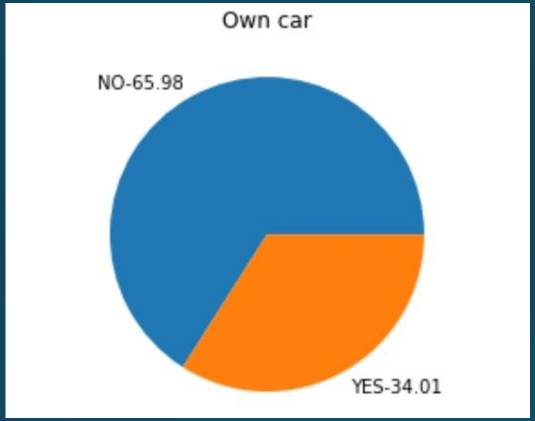
```
0    91.927118
1     8.072882
Name: TARGET, dtype: float64
```

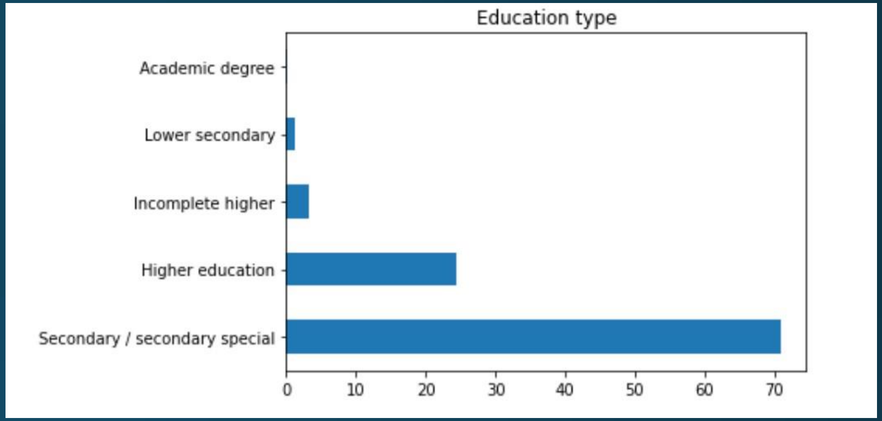


- In the target column, 0 - indicates the Non Default, 1 – indicates Default.
- 91.92 % of non defaulters and 8.08% of defaulters were found in the data, which usual seeing 8 % of defaulters overall.

3.3 Univariate plots- Categorical Variables:





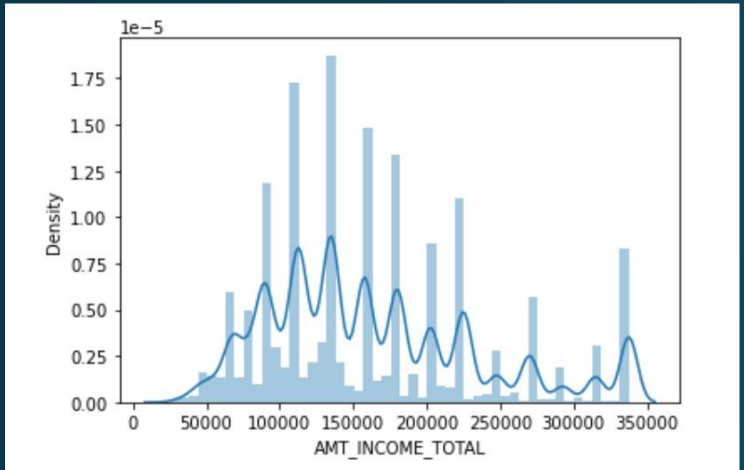


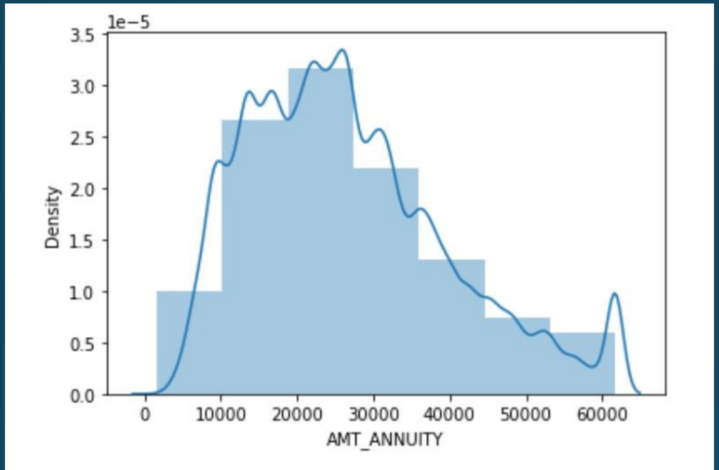
- The percentage of Female sample data is almost double compared to the Male Sample data.

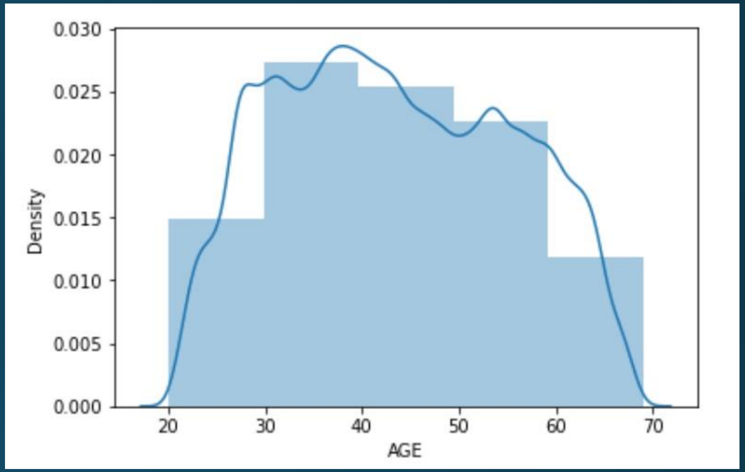
- The percentage of people having a car is lesser in the total dataset .

- The Category- Secondary/Secondary special id the Dominating Education Category among the other categories.

3.4 Univariate plots- Numerical Variables:





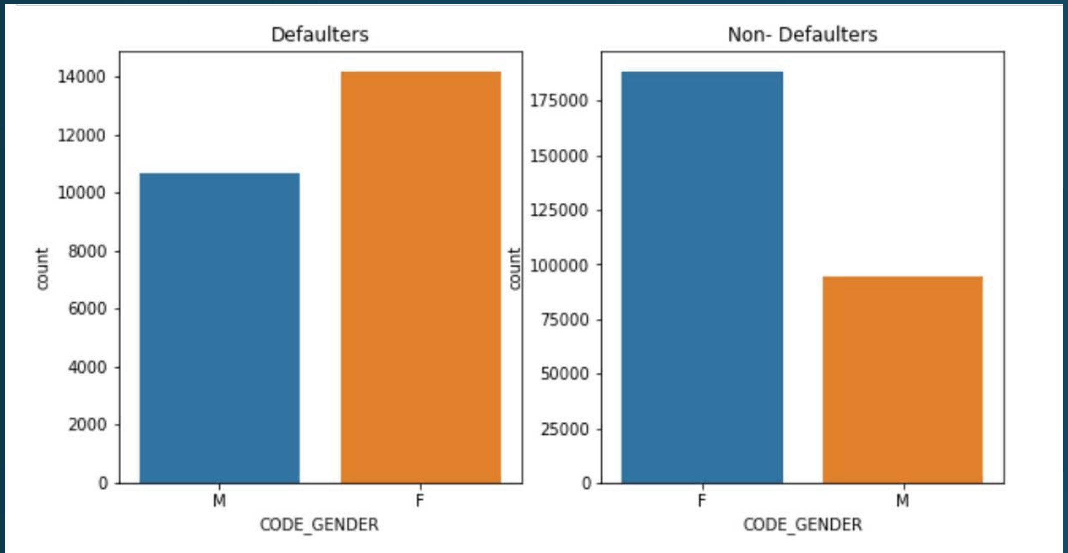


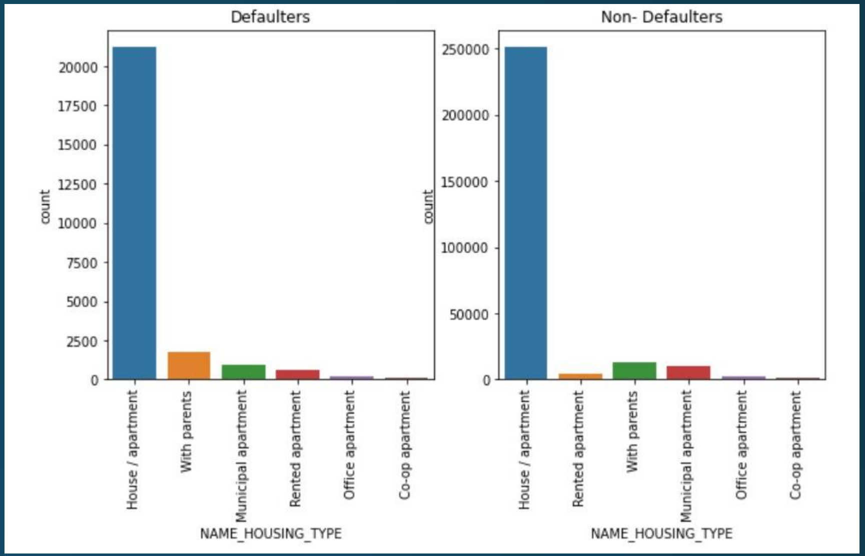
- The income is peaked at between intervals of 100000 to 150000

- The Amount of annuity highly distributed in the intervals of 20000 to 30000.

- The Age group between 30 to 40 years are more compared to the other age groups.

3.5 Segmented Univariate Analysis- Categorical Variables:

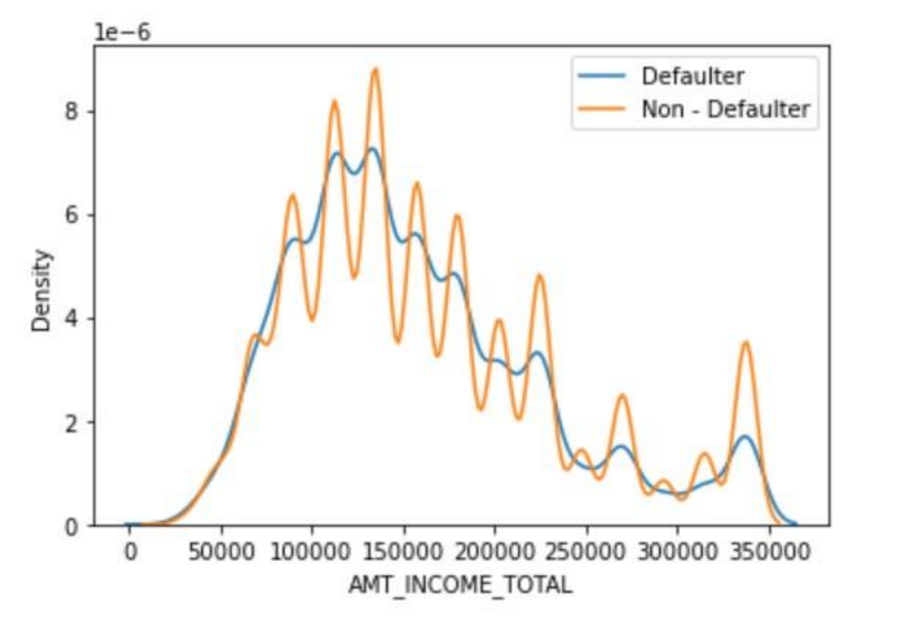


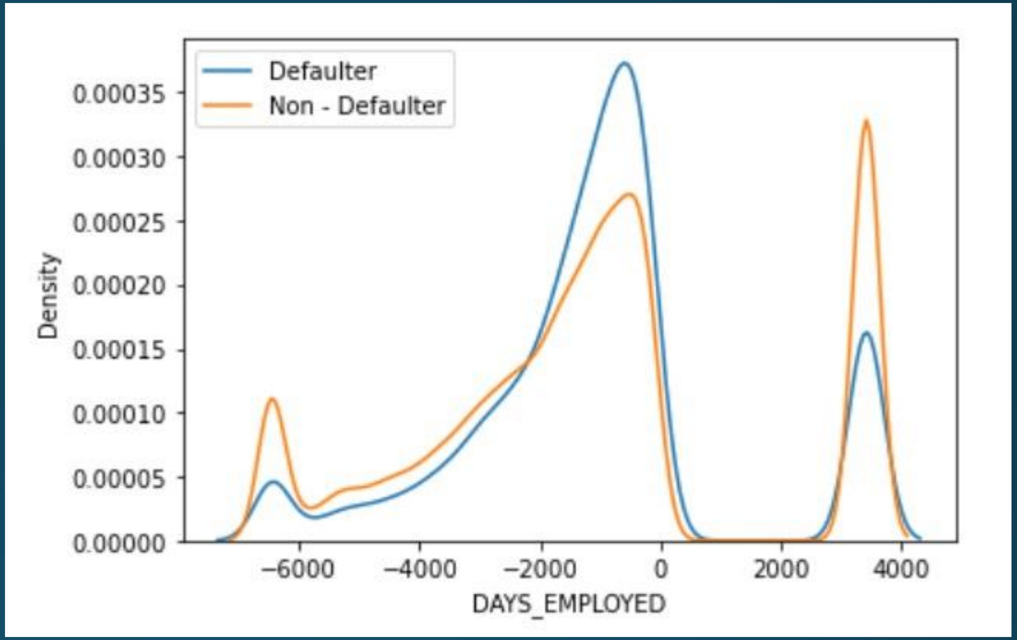


- The sample data contains more female records compared to male records. There is no significant change in the Gender attribute when compared between Defaulters and Non-defaulters.

- **When comparing defaulters and non-defaulters with Housing type, it is found that majority living in House/ Apartment. The descending order of value counts remains same.**

3.6 Segmented Univariate Analysis- Numerical Variables:

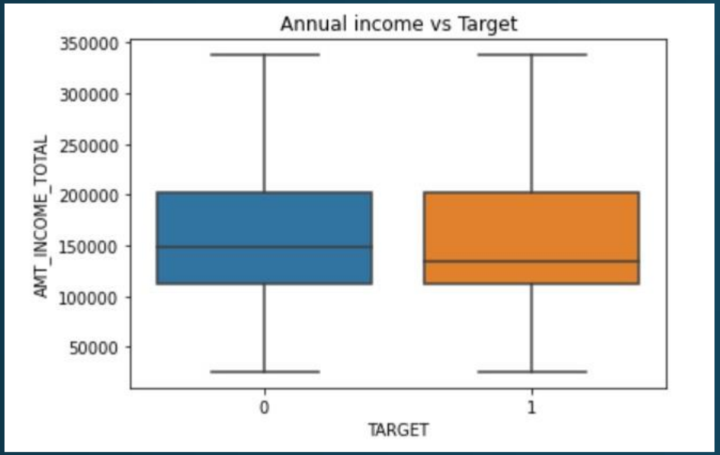


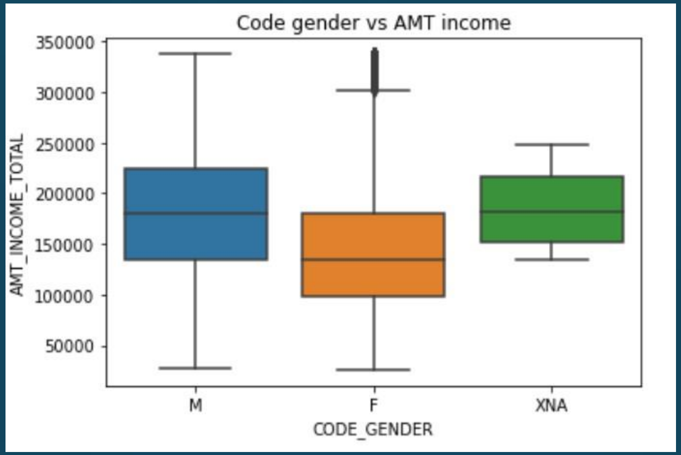


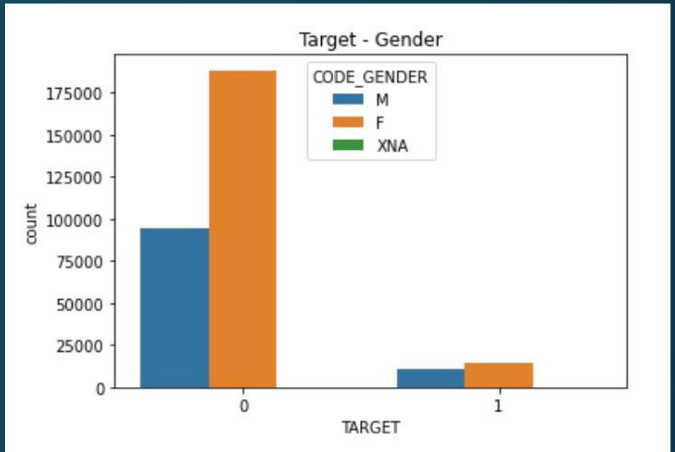
- The Amount of income – Non defaulters is dominating at high distribution places compared to defaulters income.

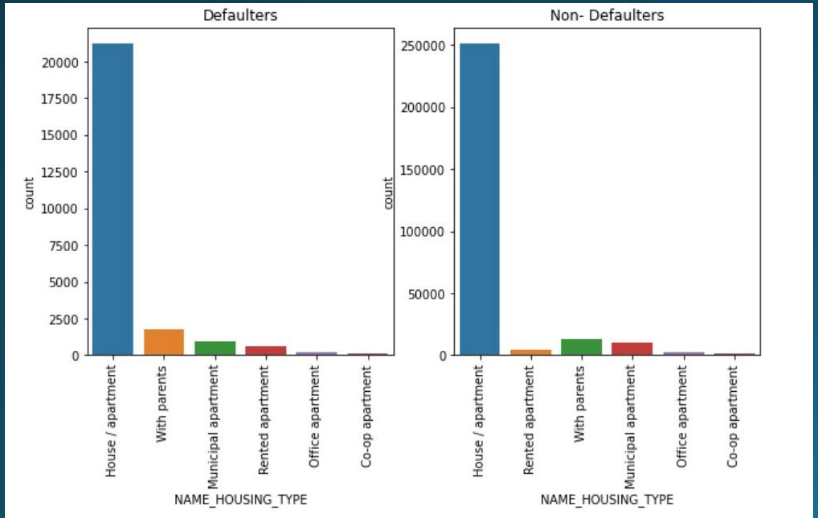
Non defaulters having high peaks in the positive days employed where as defaulters having high peaks in the –Negative days employed.

3.7Bi- Variate Plots: (on Application data)


Annual income vs Target


Code gender vs AMT income

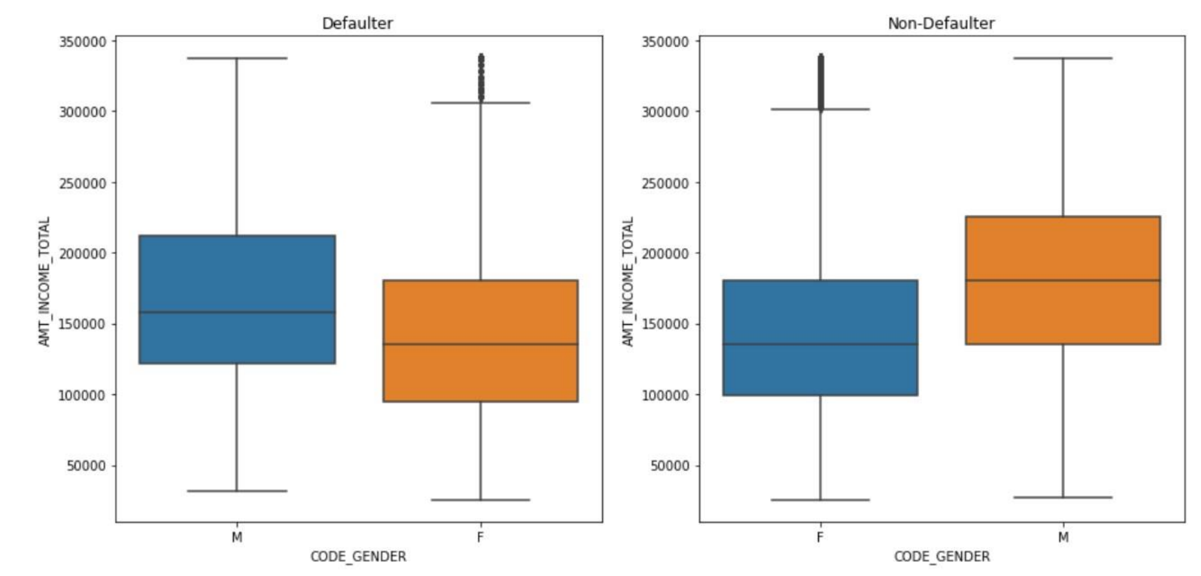

Target - Gender


Defaulters

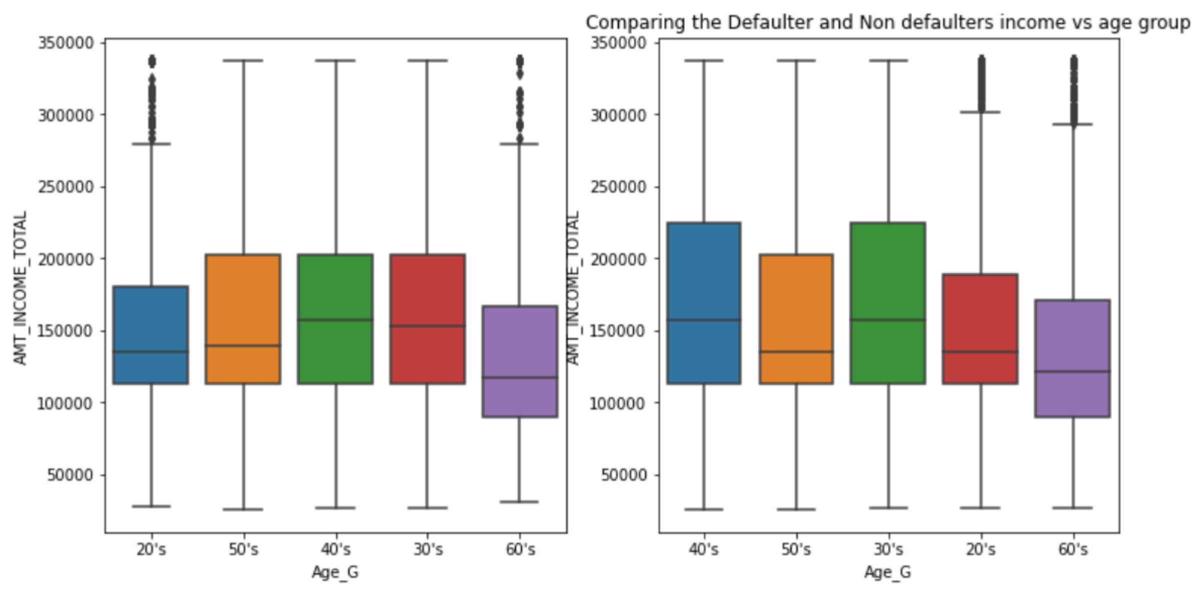
Non- Defaulters

- The Median Annual income of Non- Defaulters is higher compared to the defaulters Annual Income.
- The Median Annual income of Male proportionately higher than Female.
- Housing type of both Defaulters and Non- defaulter almost showing the same pattern.

3.8 Bi- Variate Plots on segmented data frames:

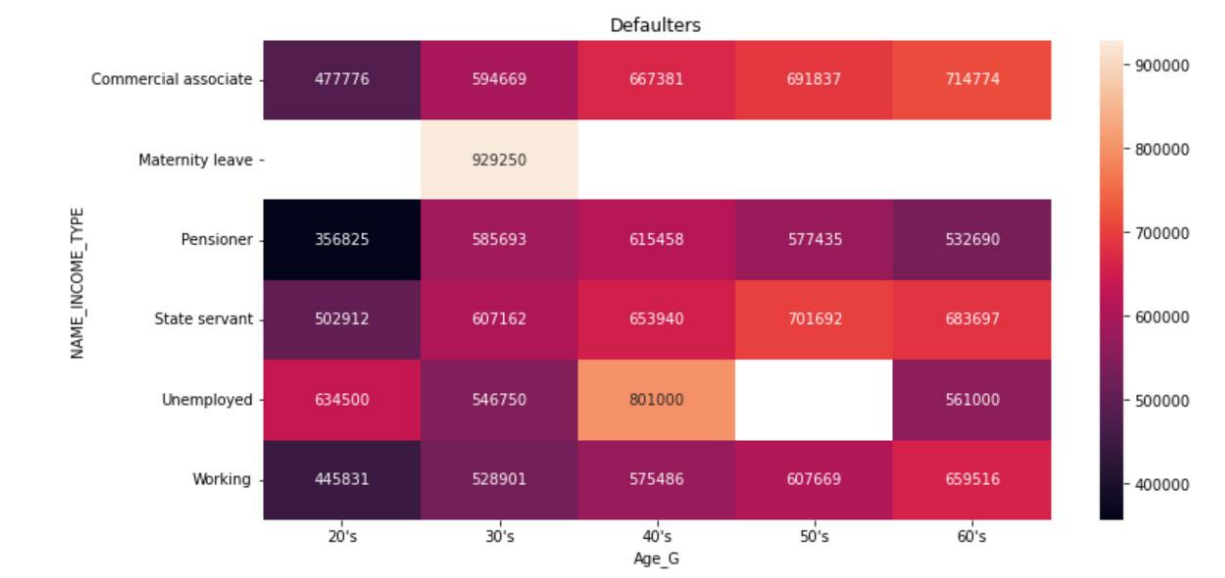


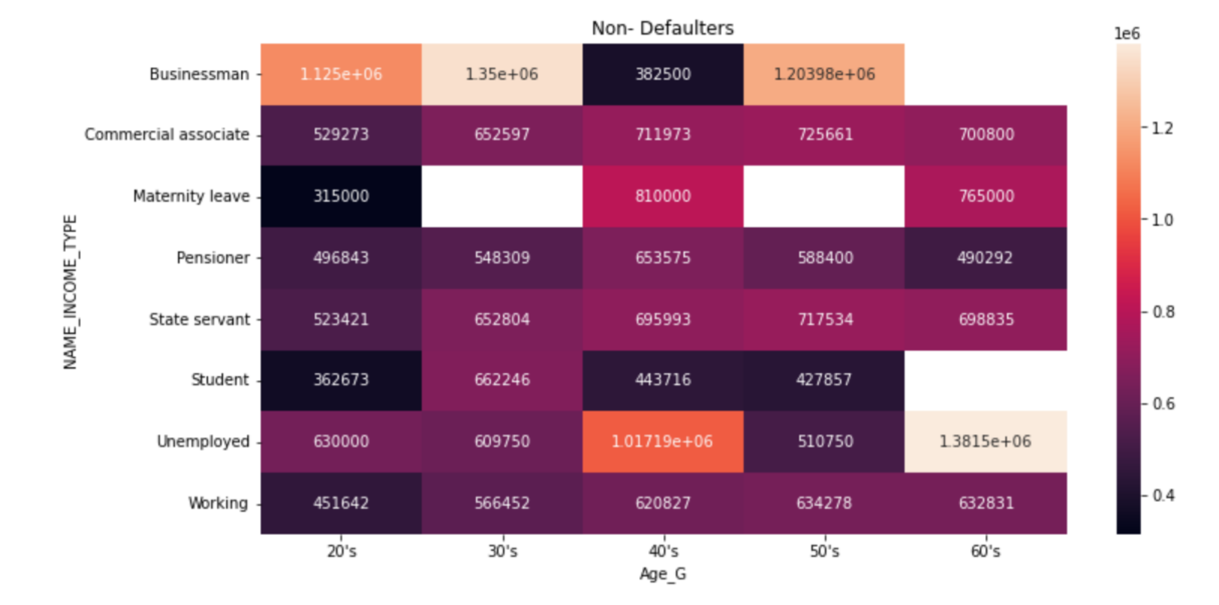
- The Median Annual income of Male proportionately higher than Female which also showing same pattern in both Default and Non-Default Categories.



- The Median Annual income 40 to 50 age group is higher in Default Category where as Median Annual income 40 to 50 & 30 to 40 similar in Non default Category.

3.9 Multi-Variate Analysis (Heat map):



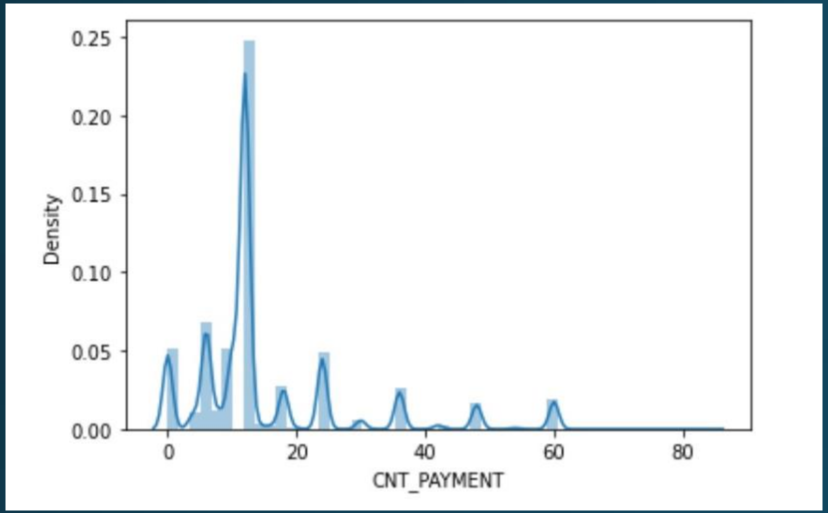


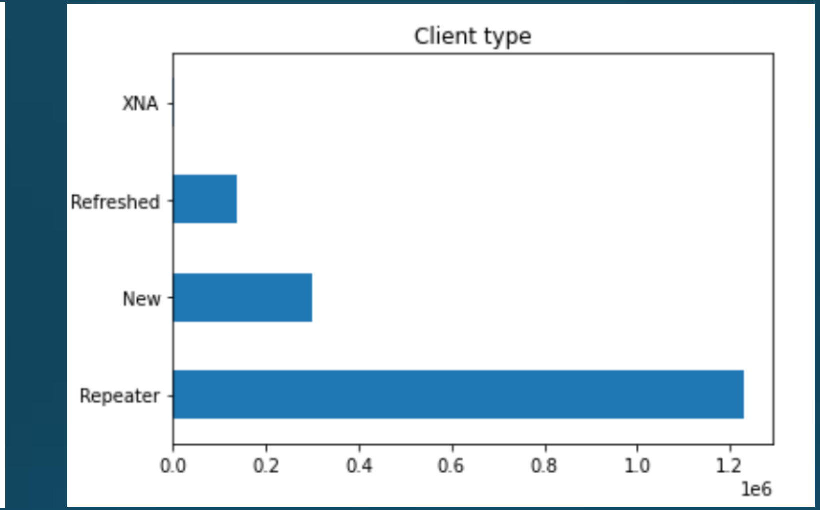
- Correlation between Amount of the income and Age is existing

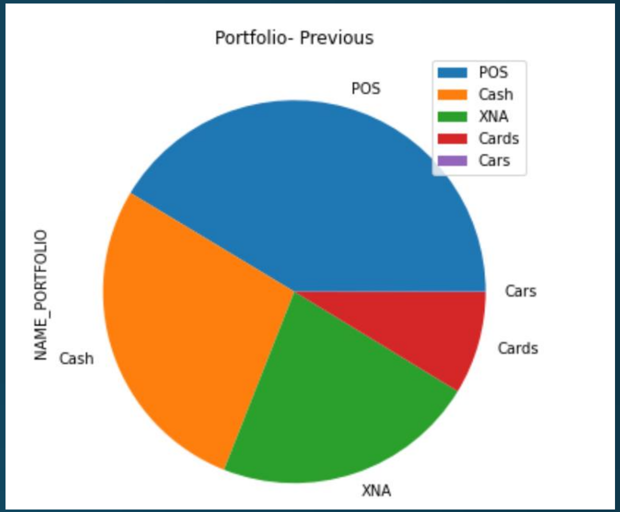
4. Data Wrangling on Previous Application data:

- Dropping the columns which are having missing values percentage more than 30.
- Dropping the Duplicate rows.
- Imputing the Columns which are having missing values. If columns are categorical then imputing with mode and if columns continuous then imputing with median.
- Checking the numerical variables outlier Treatment. Values are capped and floored Accordingly.
- **EDA Steps:**
  1. **Understanding the metadata of the data**
  2. **Understanding the Columns**
  3. **Missing value check and Treatment**
  4. **Outlier check and Treatment**
  5. **Univariate Analysis**
  6. **Bi and Multi Variate analysis**

4.1 Univariate plots:





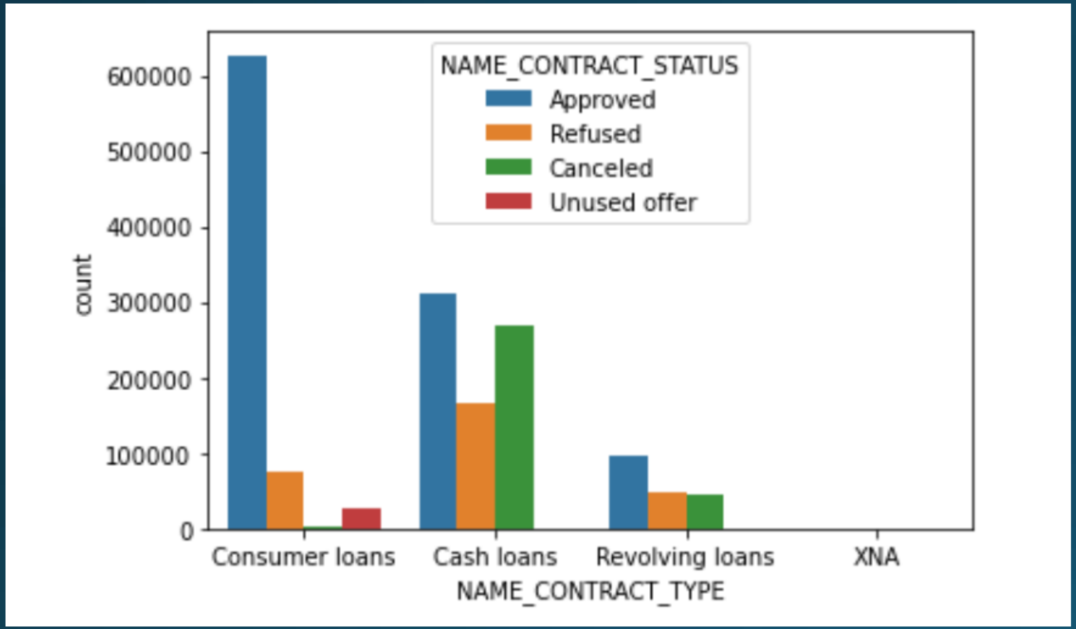


- Data is mainly distributed from 0 to 20.

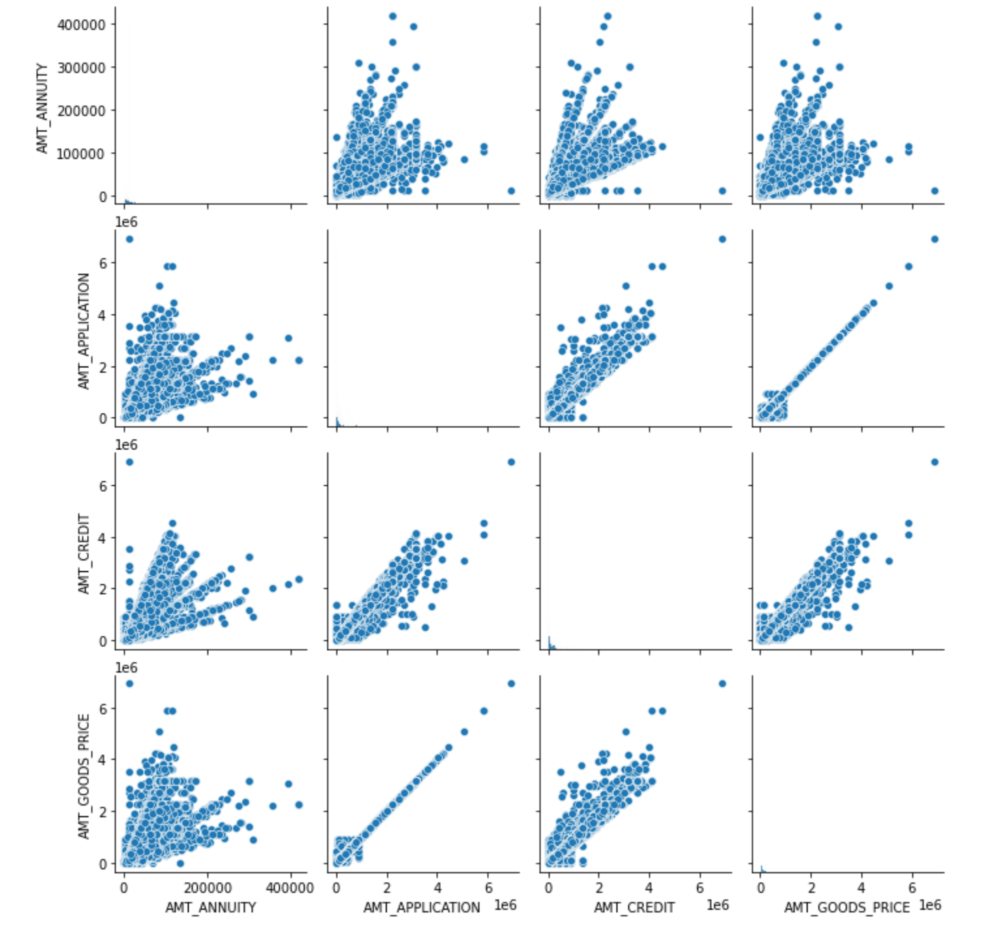
- From Previous Application data, it is clear that most of the customers are repeating. The Majority of repeats is very significant in number.

From previous data, it is also Evident that POS and Cash Occupies major portion in the Portfolio section.

4.3 Bi- Variate Plots:

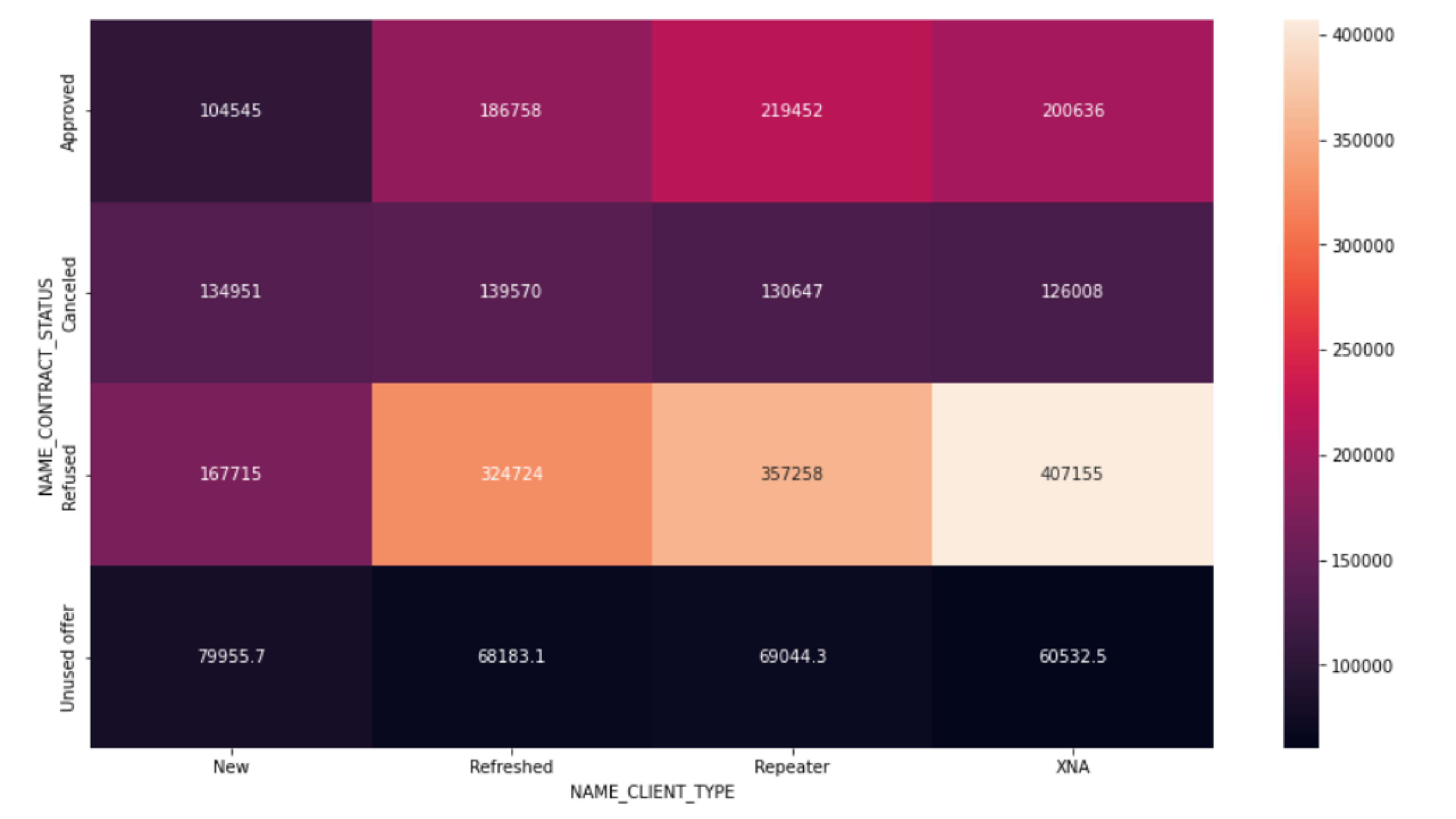


In the Cash loan sub category, Cancelled Applications are more compared to the Refused Applications which is weird comparing with Other categories in the Contract type.



Annuity, Application and Credit and Goods Price amount are highly related. Also Goods purchased is directly proportional to Loan amount.

4.4 Multi-Variate analysis (Heat plot):



It is observed that categories like Refused and Unused offer have more goods value compared to other Categories.

5. Summary:

- The percentage of Female sample data is almost double compared to the Male Sample data.
- The Median Annual income of Non- Defaulters is higher compared to the defaulters Annual Income.
- The Median Annual income of Male proportionately higher than Female.
- Variables like Age, Income, Income type, Employment and gender effecting the Target variable.
- The Median Annual income 40 to 50 age group is higher in Default Category where as Median Annual income 40 to 50 & 30 to 40 similar in Non default Category.
- From Previous Application data, it is clear that most of the customers are repeating. The Majority of repeats is very significant in number and it is also Evident that POS and Cash Occupies major portion in the Portfolio section.
- From previous application data , Annuity, Application and Credit and Goods Price amount are highly related. Also Goods purchased is directly proportional to Loan amount.
- In the Cash loan sub category, Cancelled Applications are more compared to the Refused Applications which is weird comparing with Other categories in the Contract type and also observed that categories like Refused and Unused offer have more goods value compared to other Categories.
- 30 - 40 Age group, medium income, unemployed, (labourers, Drivers) are some of the categories need to be taken care while giving loans because there high possibility for defaulting.