EDA CASE STUDY ON LOANS

Compiled By:-Pallavi Thakur

Objective

- The objective of this case study is to find the customers on the basis of various factors to check whether they are able to repay the loans or they are defaulters.This can be done using EDA.

- The company can understand the driving factors behind the loan defaulters, the variables/columns which are strong indicators of loan default.

Data Set

- **'application_data.csv'** contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

- **'previous_application.csv'** contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

- **'columns_description.csv'** is data dictionary which describes the meaning of the variables.

Flow of case study(APPROACH)

- Reading and Understanding of the data application_data
- Data Cleaning
- Find null value columns
- Imputation of Null columns
- dropping columns having null values m than 40%
- Standardization of columns to correct format
- Target column is divided into 2 columns (defaulters or repayer)
- Finding outliers in important columns(analysis)
- Univariate Analysis
- Bi-Variate Analysis
- Reading and Understanding of the previous_application dataset
- Data cleaning
- Analysis
- Merging of application_data and previous_application dataset
- Analysis of columns in merged dataset to get final result

Cleaning in Application_data

- (307511, 122) rows and columns in dataset
- 49 columns have more than 40% missing values and these columns are related to area which won't affect the analysis. Hence dropping is the good idea.
- Further 2 columns "OCCUPATION_TYPE AND EXT_SOURCE_3 having null values greater than 15%
- Again **EXT_SOURCE_3** is not required for analysis .Hence dropped
- **OCCUPATION_TYPE** column is **categorical** which we need for further analysis so using **mode** we filled the null values.
- 5 columns related to AMT_CREDIT_BUREAU having null values are again filled using **mode**
- **AMT_ANNUITY" and "AMT_GOODS_PRICE"** are numerical columns and important too so null values are replaced using **median**
- **"NAME_TYPE_SUITE"** is categorical column hence null values are replaced using **mode**
- Columns related to **Social _circle** having null values replaced using **median.**
- Finally we left with **3 columns** having very small percentage of null values and not so important. So we left them untouched.

Standardization in Application_data

- 5 columns started with **DAYS** prefix contains negative values that are converted into positive values.
- **DAYS_BIRTH,DAYS_EMPLOYED,DAYS_REGISTRATION,DAYS_ID_PUBLISH** conatins value in days . For better reading i converted it in to years.
- Handling **XNA** values in columns **CODE_GENDER** and other columns of dataframe.

Binning of columns in Application_data

- Binning the **DAYS_BIRTH** column in terms of years and observed that **maximum applicants are in the age range of 30-40.**
- Binning of continuous variables **AMT_INCOME_TOTAL,AMT_CREDIT** and **AMT_CREDIT_RANGE** in the category of 'VERY_LOW', 'LOW', "MEDIUM", 'HIGH', 'VERY_HIGH' for better reading.

Imbalanced columns

- In the dataset, the most important column is **TARGET**
  - Target variable 1 - client with payment difficulties
  - Target variable 0 - all other cases, ie no payment difficulties

- In application_data approx **91.927118** are non-defaulting and have no payment difficulties and **8.072882** are defaulting applicants.

- We can see this is not a balanced data set, and the imbalance between 2 is very high.

Analysis(Outliers)

Observations using boxplot :-

- we can observe that there is some value around **120M** which is an outlier in **AMT_INCOME_TOTAL** column.

- We did not find any outliers in the column **DAYS_BIRTH**(named APPLICANT_AGE) column.

- In the column **AMT_ANNUITY,** we have found one outlier which is greater than **250000.**

- **DAYS_EMPLOYED**(named as EMPLOYED_YRS) column tells us, number of days person started current employment before the application. we observe value greater than **1000 years** which is surely outlier.

- **DAYS_REGISTRATION**(named as registration_yrs) is a column which tells us how many days before the application, the client changed his registration. Here we have observed the value to be **70 years** which is an outlier.

- **CNT_CHILDREN** column means no of child client have outlier near to the value of **18.**

Analysis(Checking the distribution)

- Distribution of **OCCUPATION_TYPE** column

Observations:- Laborers have **highest** count followed by Sales Staff and Core Staff whereas **IT Staff** has the **lowest** count for application of loans.

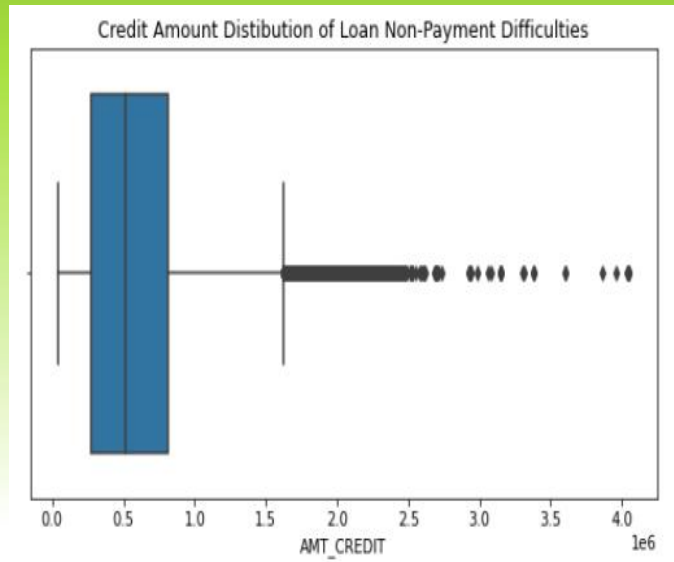
- Distribution of **ORGANIZATION_TYPE** column

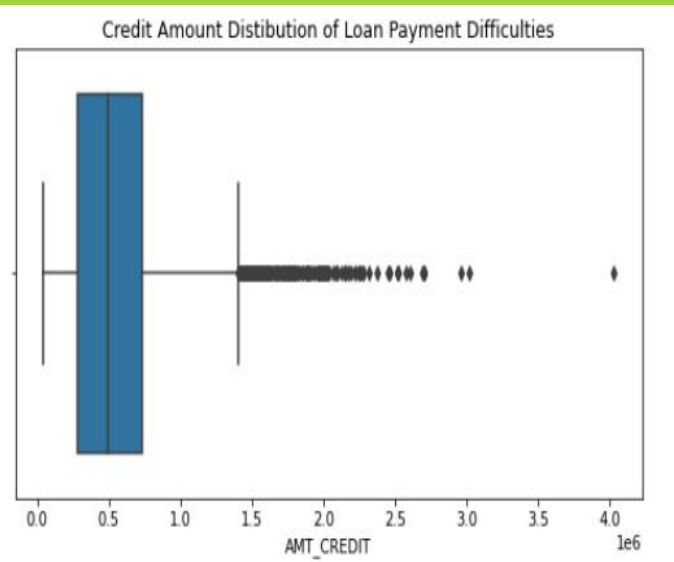
Observations:- Business Entity type 3 has highest count for applying for loan where as Industry type 13,trade type 4 ,trade type 5 and Industry type 8 are not at all interested

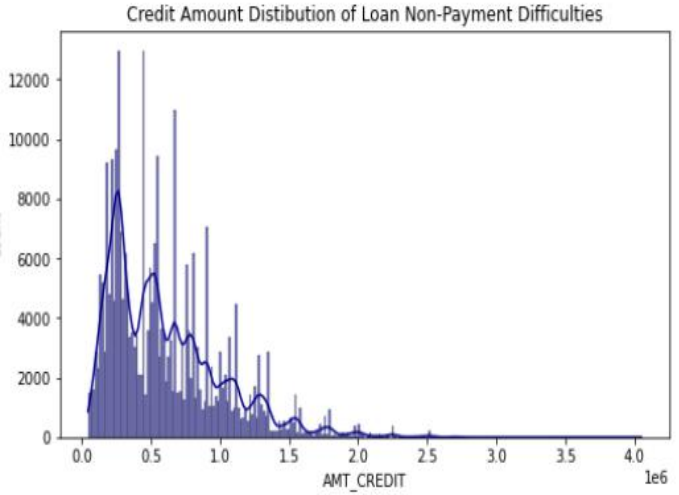
Univariate Analysis

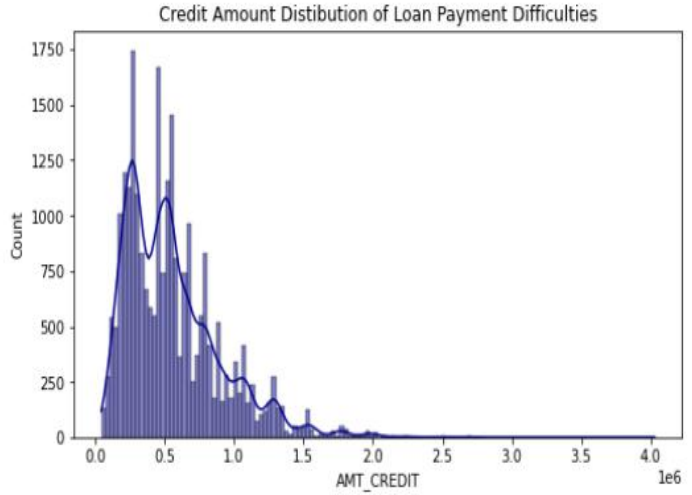
Observations:-

- We can observe that there's not much difference in defaulter and non-defaulter females but the males defaulter ratio is higher than non-defaulters in **Gender Distribution according to the target variable.**
- We can observe that there are few outliers in both the credit amount customers that is payment difficulty customers and payment non-difficulty customers.
- We can also see that the distribution does not appear to be bell or normal curve. The distribution is more inclined towards the first quartile.





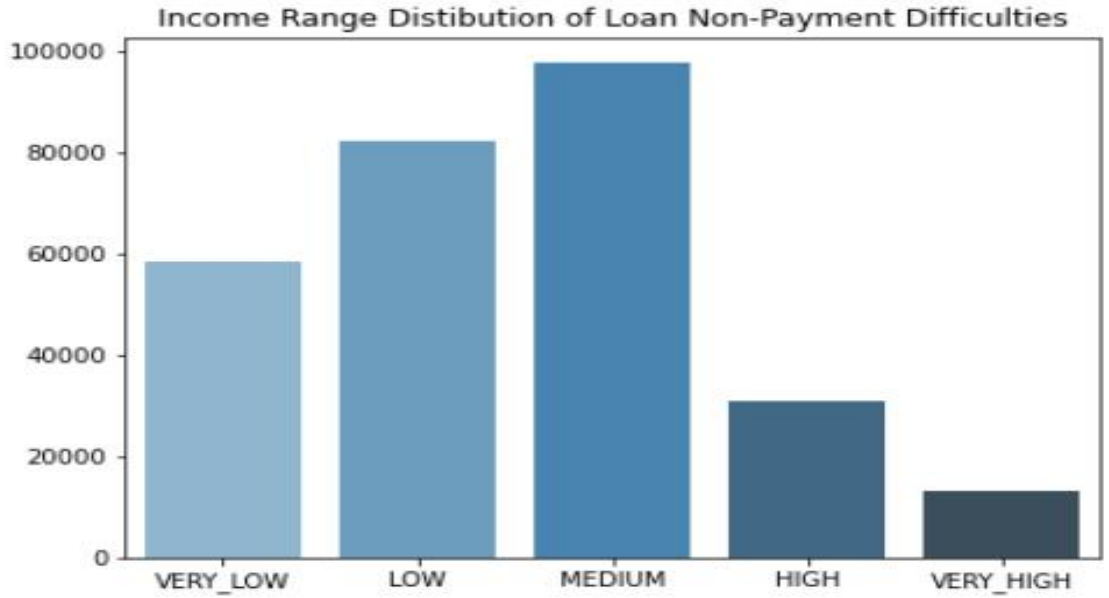


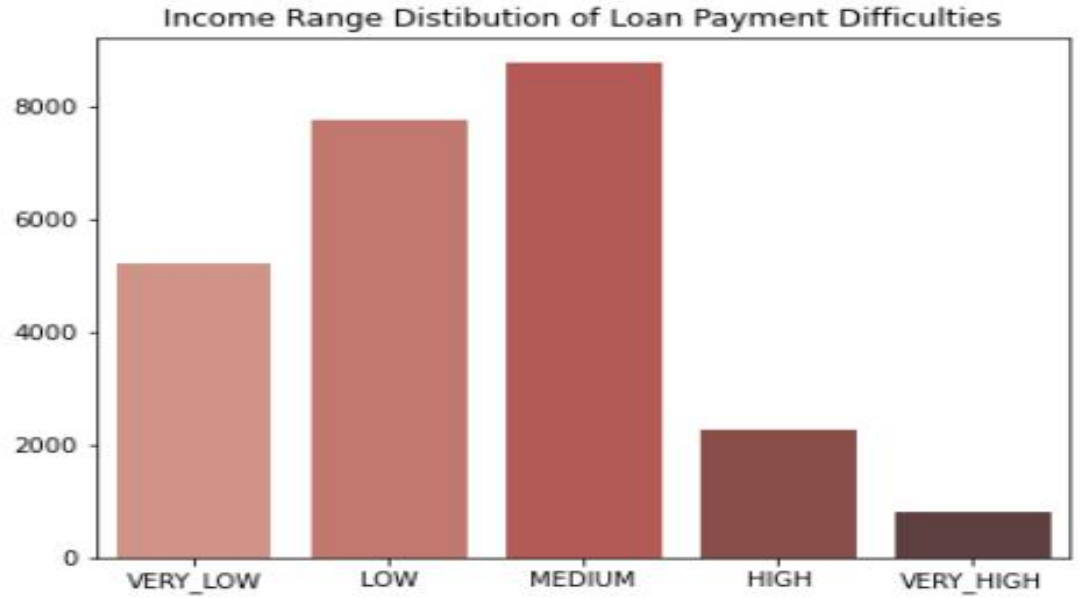


Univariate Analysis

Observations-

- We can observe that customers with very low and low credit range default more as compared to the others categories.
- There does not seem to be any such pattern for the number of children of defaulters and non-defaulters.
- We can observe that with growing income the defaulters seem to decrease that is people with lower income range tend to default more as compared to those with higher income ranges.

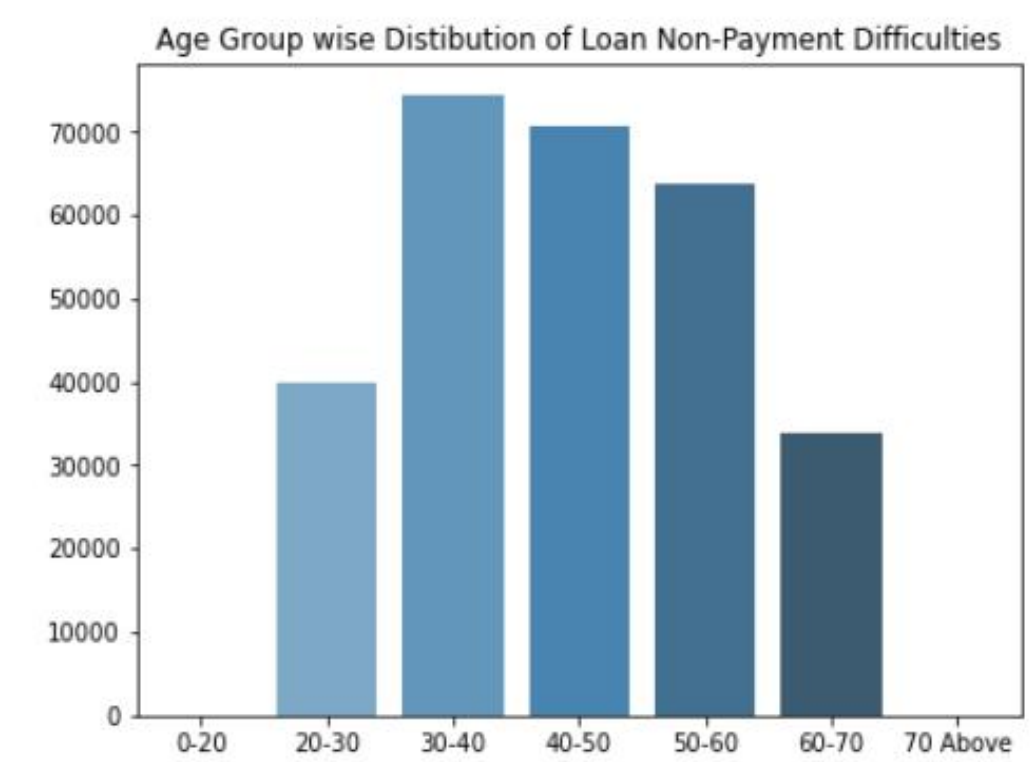


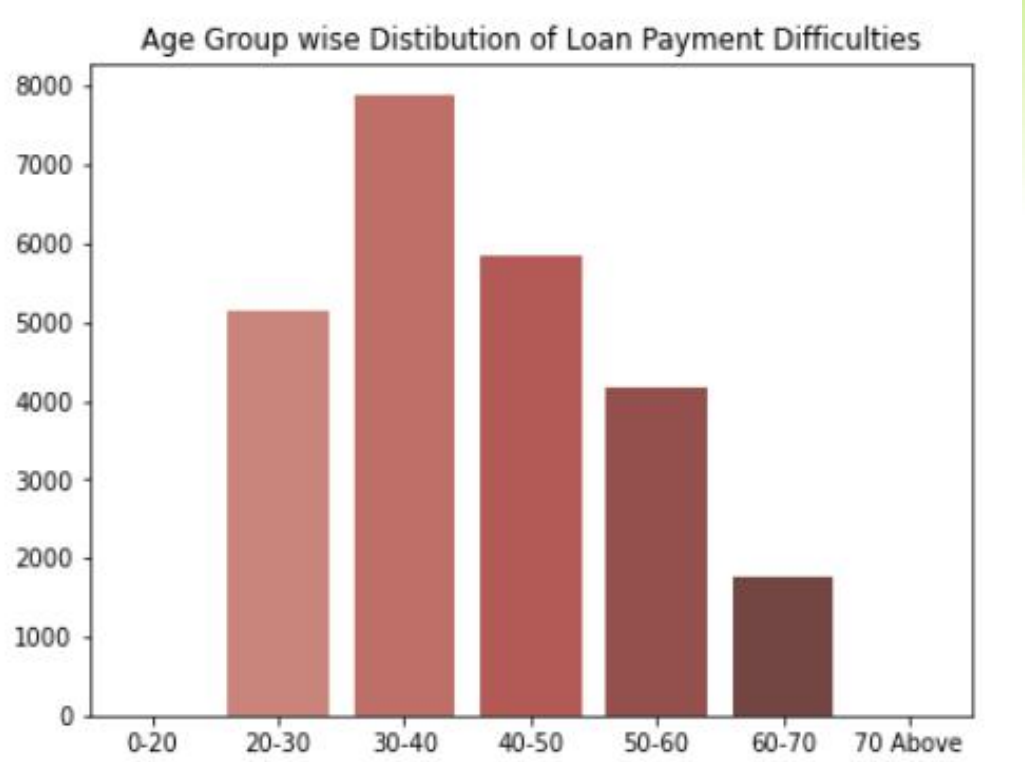


Univariate Analysis

Observations-

- We can observe that middle-aged people and young people tend to default more as compared to other age groups(younger and older).

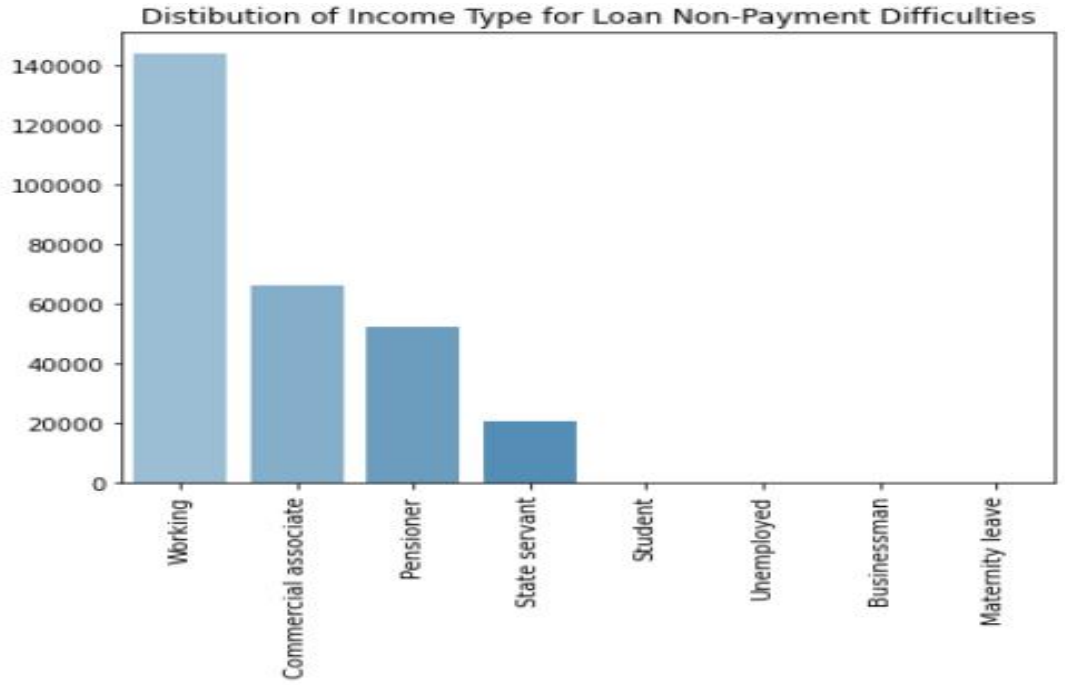


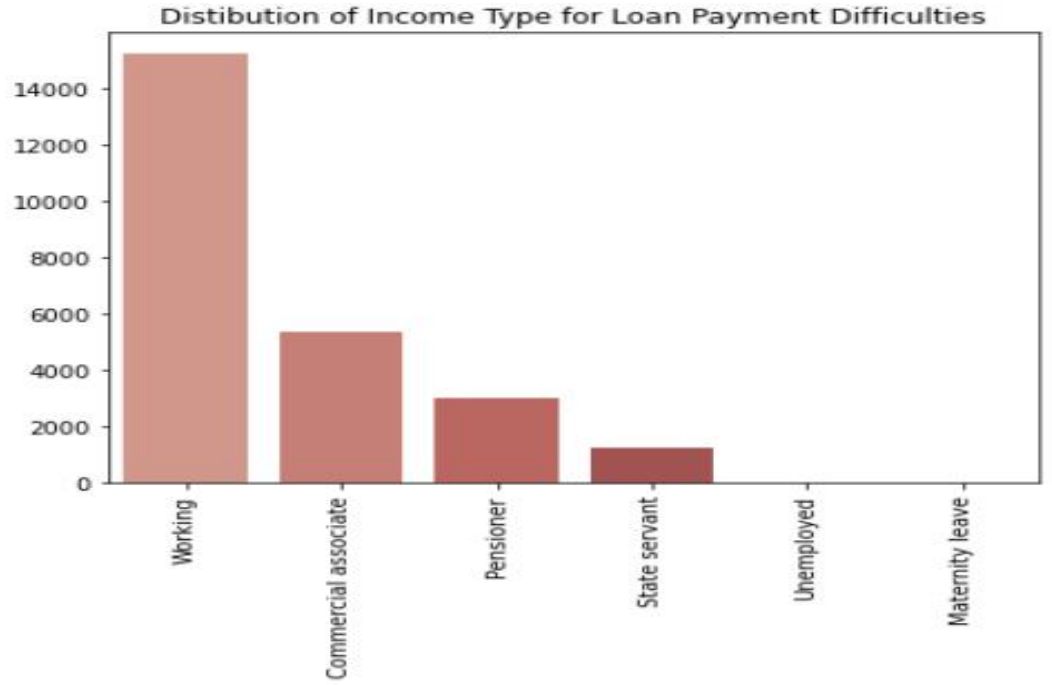


Univariate Analysis

Observations-

- We can infer that most of the defaulters tend to lie in the first quartile of the **GOODS PRICE DISTRIBUTION** column and the curve is not a normal or bell curve.
- We can see that Working people have majority in Non payment and payment difficulties
- No businessman and maternity leave employees and unemployed are there in Payment difficulty.





Univariate Analysis

Observations-

- There is decrease in the percentage of married and widowed with in Payment Difficulties
- There is Increase in the percentage of Single/Not married and Civil marriage in Loan Payment Difficulties.

- We can see that there exists people who own house/apartment are in both Loan Non-Payment Difficulties and Payment Difficulties.

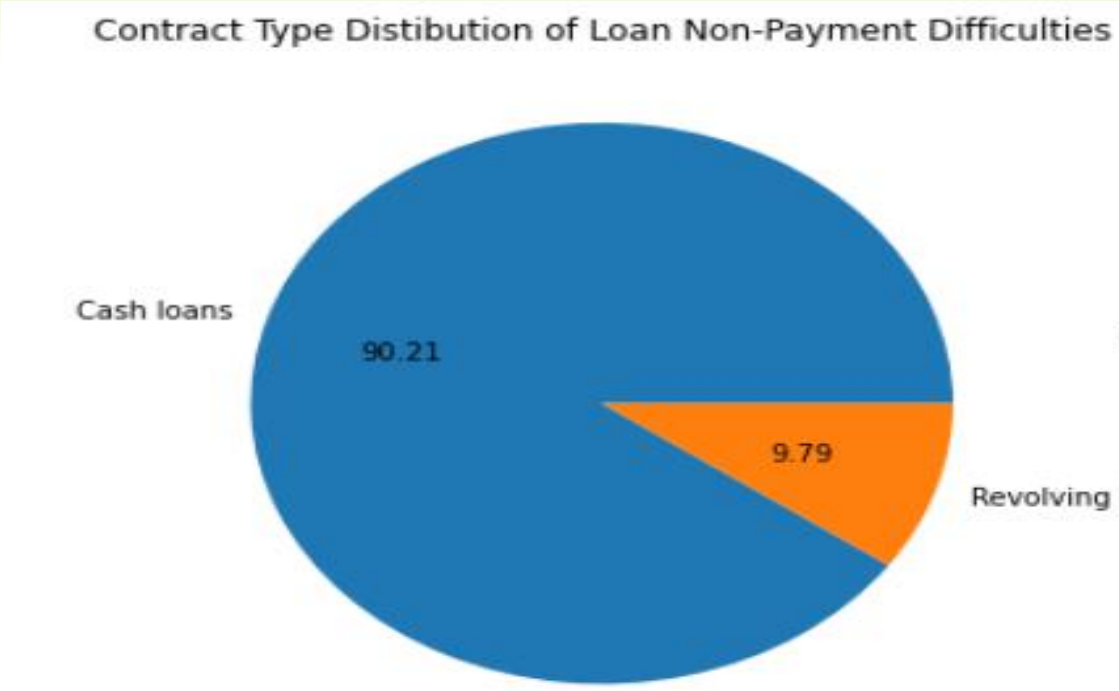
- We can conclude that secondary/secondary special educated people applying for loan have high percentage
- Academic degree people have very lower percentage in applying for loan.

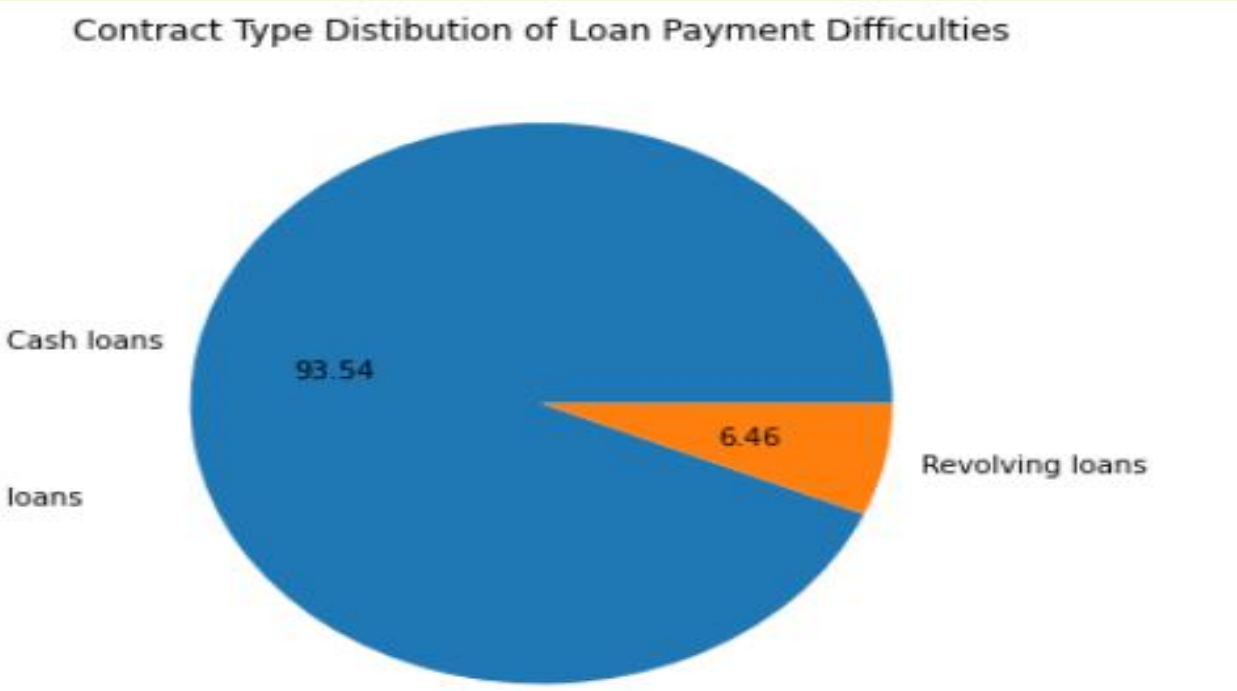
- Little changes are there in both graphs for people with payment difficulties and without payment difficulties in **NAME_TYPE_SUITE** column.

Univariate Analysis

Observations-

- We can Conculde that cash loans are preferred by both Loan Payment Difficulties and Loan-Non Payment Difficulties
- There is a decrease in the percentage of Payment Difficulties for revolving loans.





Bi-Variate Analysis

- **Bi-Variate Analysis of numerical columns:-**

- Using pairplot AMT_CREDIT', 'AMT_ANNUITY', 'AMT_INCOME_TOTAL', 'AMT_GOODS_PRICE', 'DAYS_BIRTH are analysed .
- **Obseravtions:-**

❖ We can observe that there is high positive co-relation between goods price and Amount credit. there also appears to be some corelation between goods price and amount annuity for Non payment diffculty.

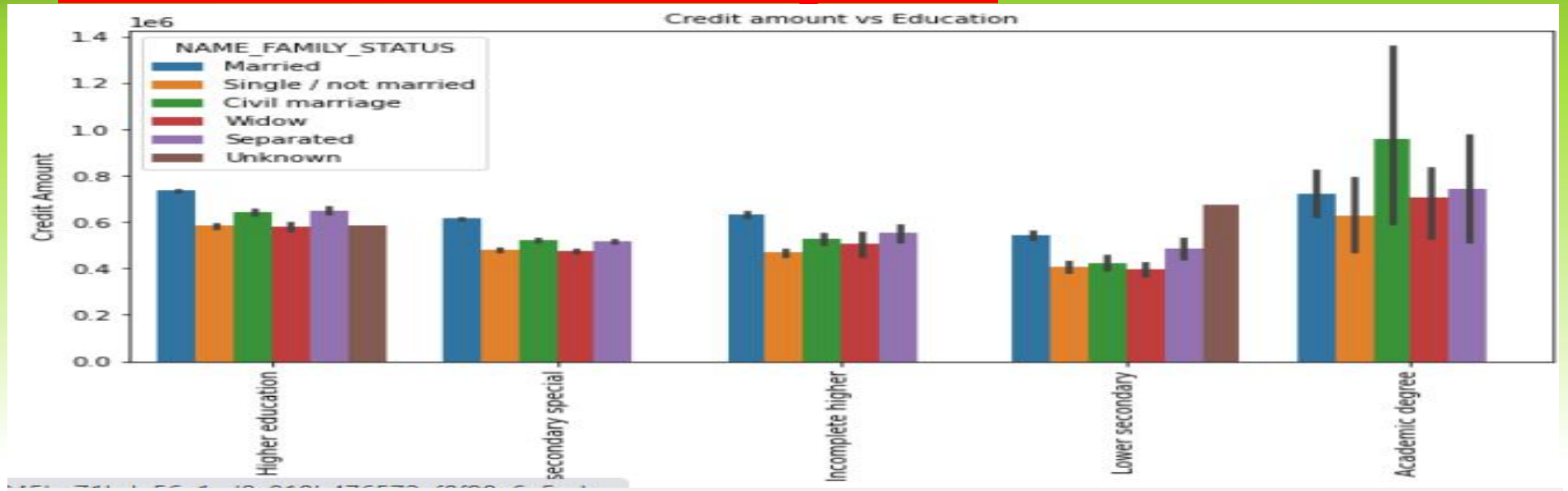
❖ We can observe that there is high co-relation between annuity amount and goods price for Payment difficulty customers.

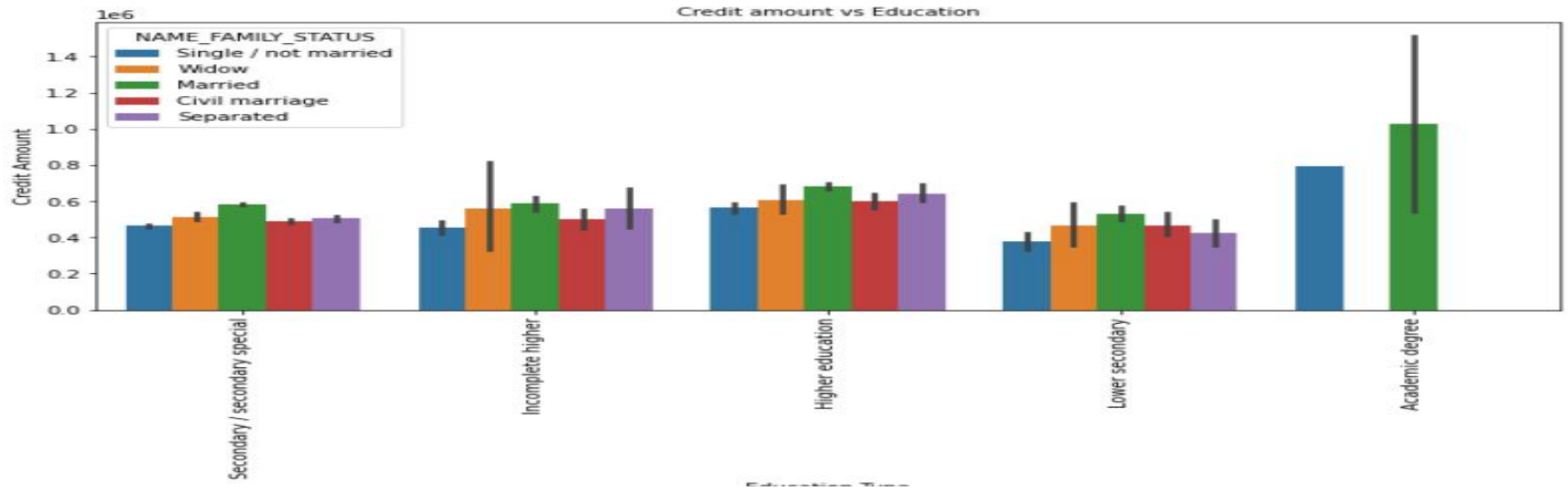
Bi-Variate Analysis

- **Bi-Variate Analysis of numerical columns:-**
- **Obseravtions:-**

❖ People with academic degree (Single & Married) have high credit amount for defaulters compared to rest of educated people.

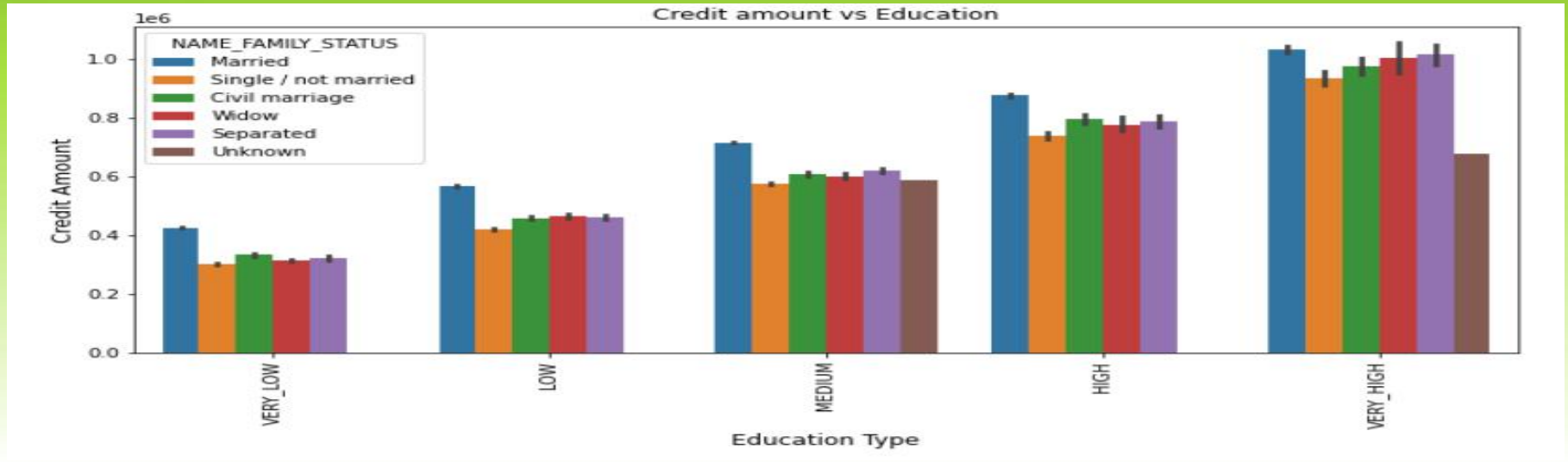
❖ We can conclude that people with academic degree (single, separated and widows) tend to default lesser compared to other categories.

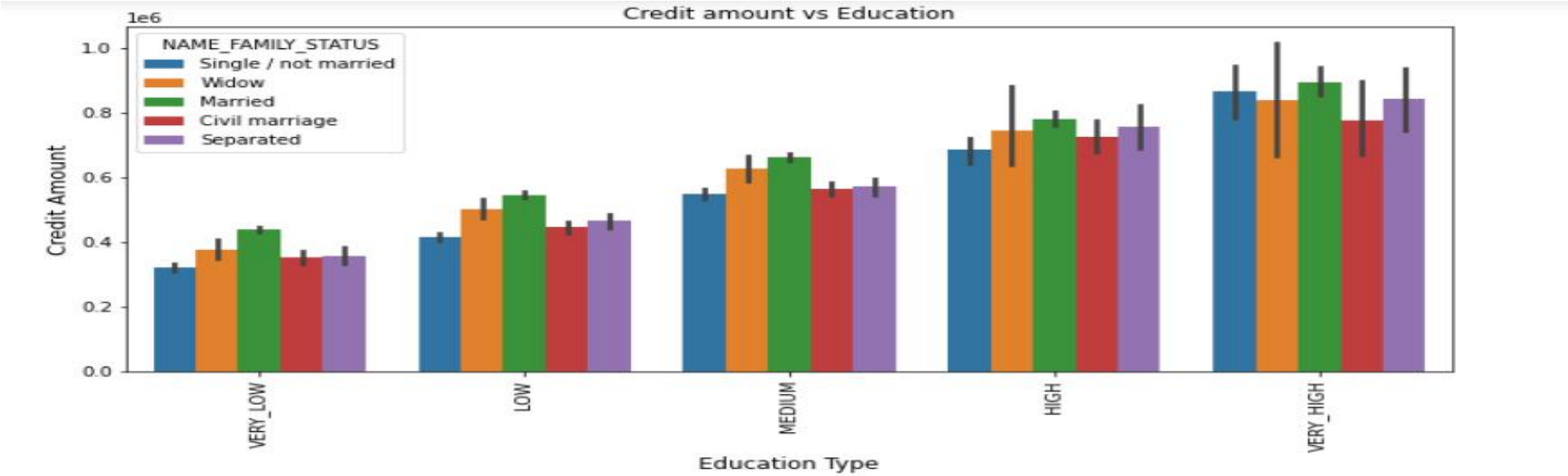




Bi-Variate Analysis

- **Bi-Variate Analysis of numerical columns:-**
- **Obseravtions:-**
- In Income range **“very high”,** family status of Married, single and seperated have high credit amount.





Previous Application DataFrame

- Univariate analysis

❖ We can observe that most of our consumer's contract product type lies either in Consumer Loan category or Cash Loan Category.

❖ We can observe that most of our consumer's contract have been approved (62.07%), Only a few were unused (1.58%).

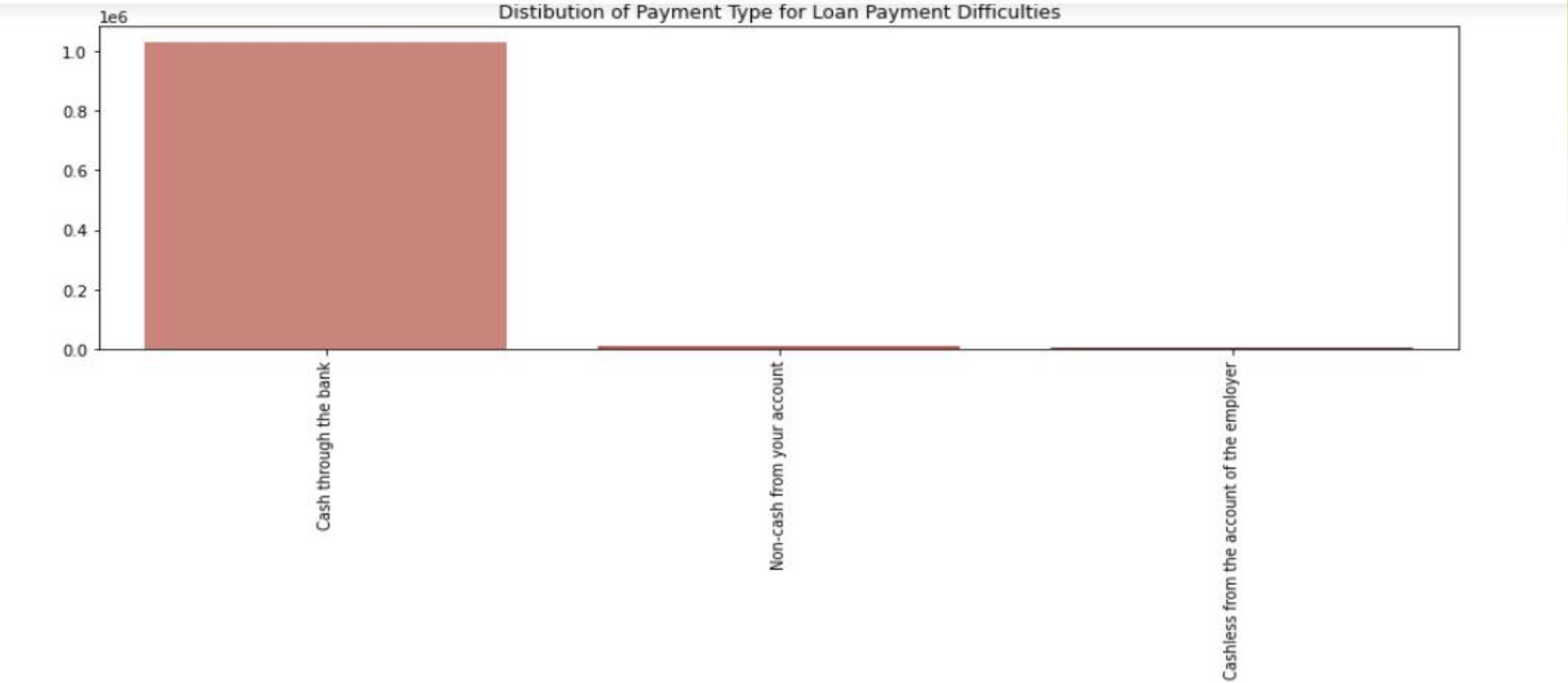
❖ Most of the customers will do payment by taking cash from the bank.

❖ Most clients who are applying are repeaters.

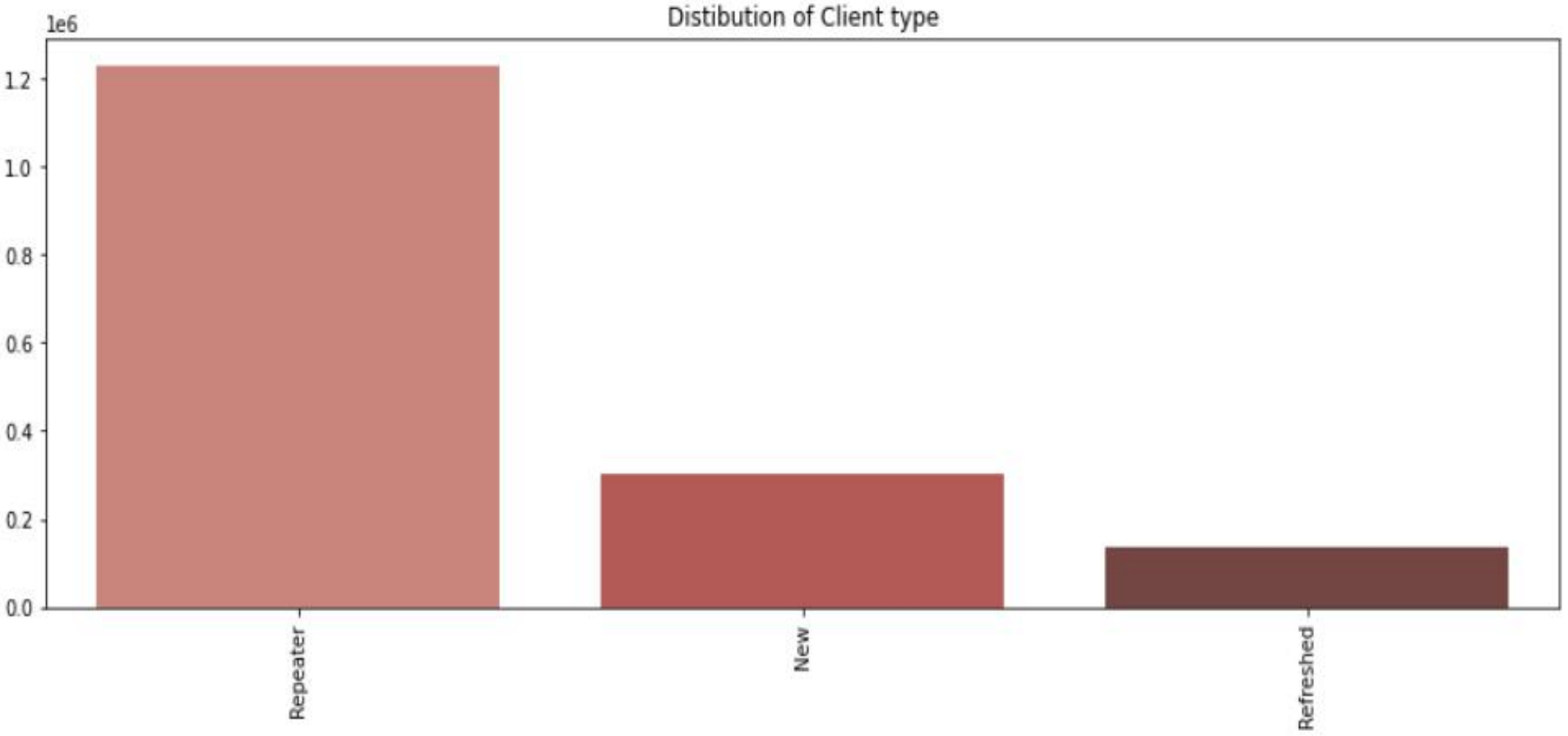
❖ **HC is the most common reason for rejection of previous loans.**

- # Univariate analysis

❖ Most of the customers will do payment by taking cash from the bank



- Univariate analysis
  - ❖ Most clients who are applying are repeaters.



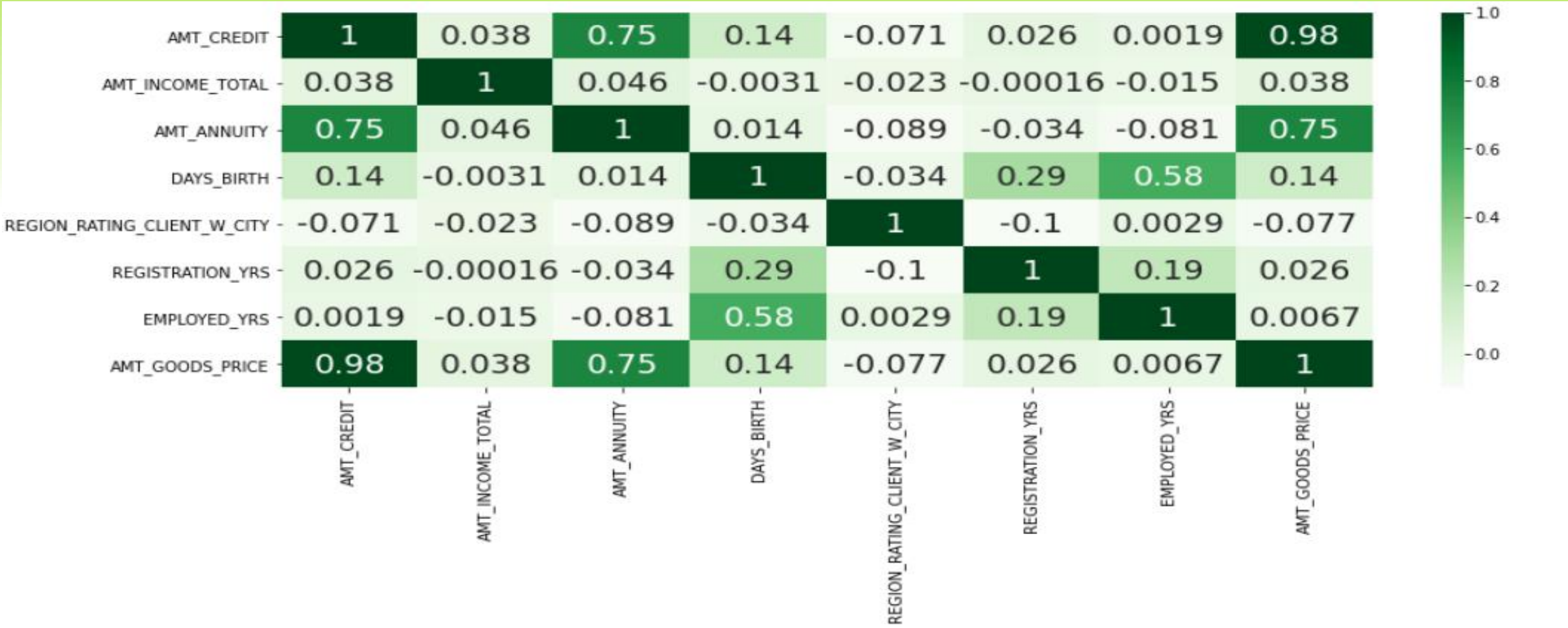
Merging DataFrames

- **Analysis**

TYPE_x STATUS	Cash loans	Revolving loans
• Approved	0.078105	0.049836
• Cancelled	0.094178	0.058751
• Refused	0.123735	0.069429
• Unused offer	0.084637	0.061972

HeatMap Analysis

As we can see there is a strong relationship between the credit amount and the amount goods price.



Conclusion:-

- We can observe that most of the **defaulting customers** have a approved(4.75%), previously refused (2.08%) or canceled (1.68%) applications and unused offers are (0.13%) .

- We can observe that most of the **non- defaulting customers** have a approved(57.92%), canceled (16.66%) ,Refused (15.27%) applications and unused offers are (1.47 %) .

THANKYOU