Credit EDA Assignment

Exploratory Data Analysis

- SIDDHANT NAIK

DSC59

PROBLEM STATEMENT

BUSINESS UNDERSTANDING:

** A loan provider company is finding it hard to provide loans to people due to there credit history being non-existent and insufficient, due to which some
consumers are taking advantage by being defaulter. With the help of EDA concepts I had make sure applicants who are capable to replay loan there
application did not get rejected .**

*** Decision that can be taken by company while receiving a loan are :**

1. **Approved**: Company approves the loan.
2. **Cancelled**: Client cancelled the application within the process due to further financial issue .
3. **Refused**: Company rejected the loan application.
4. **Unused Offer**: Loan got cancelled by the client after approved by the bank .

*** Above decisions can be broadly described into two types of risks associated with bank’s decision:**

1. Applicant likely to repay the loan but bank doesn’t approves it
2. Applicant likely to not repay the loan but bank approves it

PROBLEM STATEMENT

BUSSINESS OBJECTIVE:

The aim is to get the knowledge of the factors behind loan default(driver variables) which company could utilize for its portfolio and risk assessment. Understanding the difficulty in payment for taking actions regarding the applicant such as :

1. **Denying the loan**
2. **Reducing the amount of loan**
3. **Lending loan at higher interest etc.**

This will ensure capable applicants are not rejected through bank’s end and advantage taking defaulters should not be given loan facility further .

UNDERSTANDING DATA

The dataset consists of 3 files which are explained as follows:

1. **‘application_data.csv’ – Information of the applicant whether client has payment difficulty.**

2. **‘previous_application.csv’- Information of client’s previous loan application containing whether previous application was rejected , approved, cancelled or unused offer.**

3. **‘columns_description.csv’- library of description for all the variables in above two csv files**

ADD YOUR TEXT HERE

enter your text.Click here to enter your

here to enter

your text.Click here to enter your text.Click here to

enter your text.

ADD YOUR TEXT HERE

Click here to enter your text.Click here to enter your text.Click here to enter your text.Click here to enter your text.Click here to enter your text.Click here to enter your text.

IMPORTING LIBRARIES

1. import warnings warnings.filterwarnings('ignore’)

2. import pandas as pd

3. import numpy as np

4. import seaborn as sns

5. import matplotlib.pyplot as plt

6. from plotly.subplots import make_subplots

7. import plotly.graph_objects as go

Used two data sets for analysis as:

1. application_data.csv as df

2. previous_application.csv as pre_app

APPROACH USED IN df

1. Inspected data_set using shape.info() and describe() function to get the insight for rows and columns , data type of different variables and stastical information of numerical variables respectively .

2. Proceeded , with checking presence of null values using isnull().sum() function in the data set.

3. Once , got the null value percentage for every variables then assumed to eliminate the null values having percentage more than 40% , as theortically

25 to 30% or more can be used to analysis but more than 50% are strictly prohibited to use so just assumed to go with more than 40% null percentage
value.

4. After these all checked for 0 values presence in numerical data type variables and replaced them with mean and mode values of that columns
respectively .

5. After that checked for XNA values presence in object data type and performed imputation over them and converted them to categorical variable.

6. Then dropped the not usefull variables from the datasets and confirmed the null presence remove by checking isnull().sum() function
.
7. Performed Outlier Analysis

8. Fetched the data imbalance value for TARGET 0 and TARGET 1
.
9. Performed Univariate and BiVariate Analysis further and found some insights

APPROACH USED IN PRE_APP

1. Inspected data_set using shape , info() and describe() function to get the insight for rows and columns , data type of different variables and stastical
information of numerical variables respectively .

2. Proceeded , with checking presence of null values using isnull().sum() function in the data set.

3. Once , got the null value percentage for every variables then assumed to eliminate the null values having percentage more than 22% , as theortically 25 to
30% or more can be used to analysis but more than 50% are strictly prohibited to use so just assumed to go with more than 40% null percentage value.

4. After these all checked for 0 values presence in numerical data type variables and replaced them with mean and mode values of that columns respectively .

5. After that checked for XNA values presence in object data type and performed imputation over them and converted them to categorical variable.

6. Then dropped the not usefull variables from the datasets and confirmed the null presence remove by checking isnull().sum() function.

7. Performed Outlier analysis and removed if required

NOTES : FURTHUR PROCEED TO MERGE THE APP_DATA AND PRE_DATA DATASET INTO MERGE DATA SET

APPROACH USED IN MERGED DATASET :

1. Performed merging of two data_sets as follows merge=pd.merge(left=app_data, right=pre_data,how='inner', on='SK_ID_CURR',suffixes=['_PREV','_CURR'])

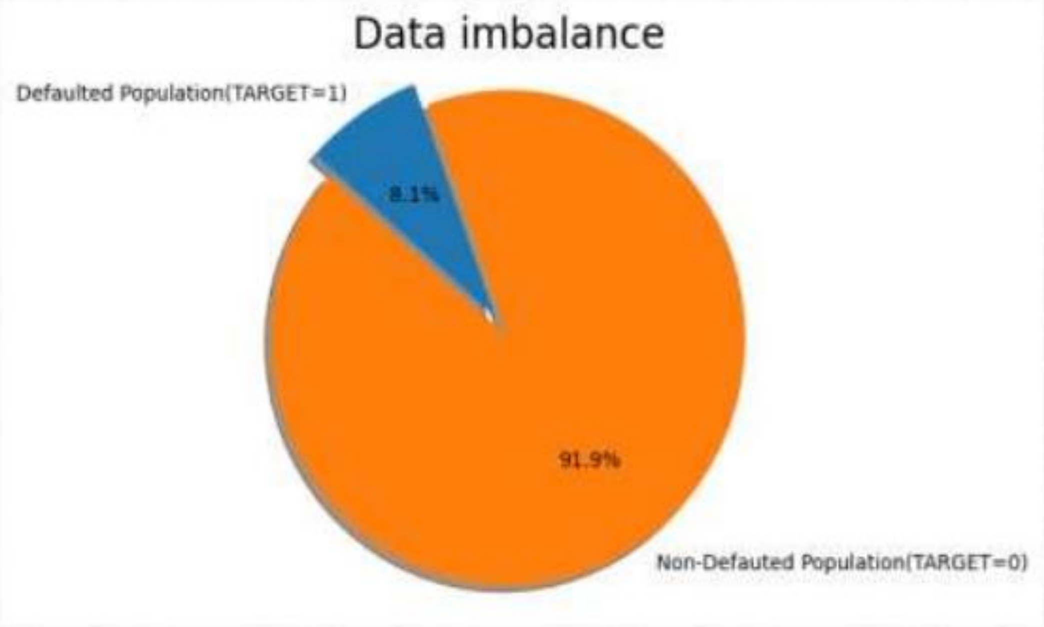
2. Inspected data_set using shape , info() and describe() function to get the insight for rows and columns , data type of different variables and stastical information of numerical variables respectively .

3. Then , on the basis of the TARGET variable divided the dataset into two datasets as follows:
a. me0 as data set containing TARGET variable value as 0, then inspected its structure through shape
b. m1 as data set containing TARGET variable value as 1, then inspected its structure through shape

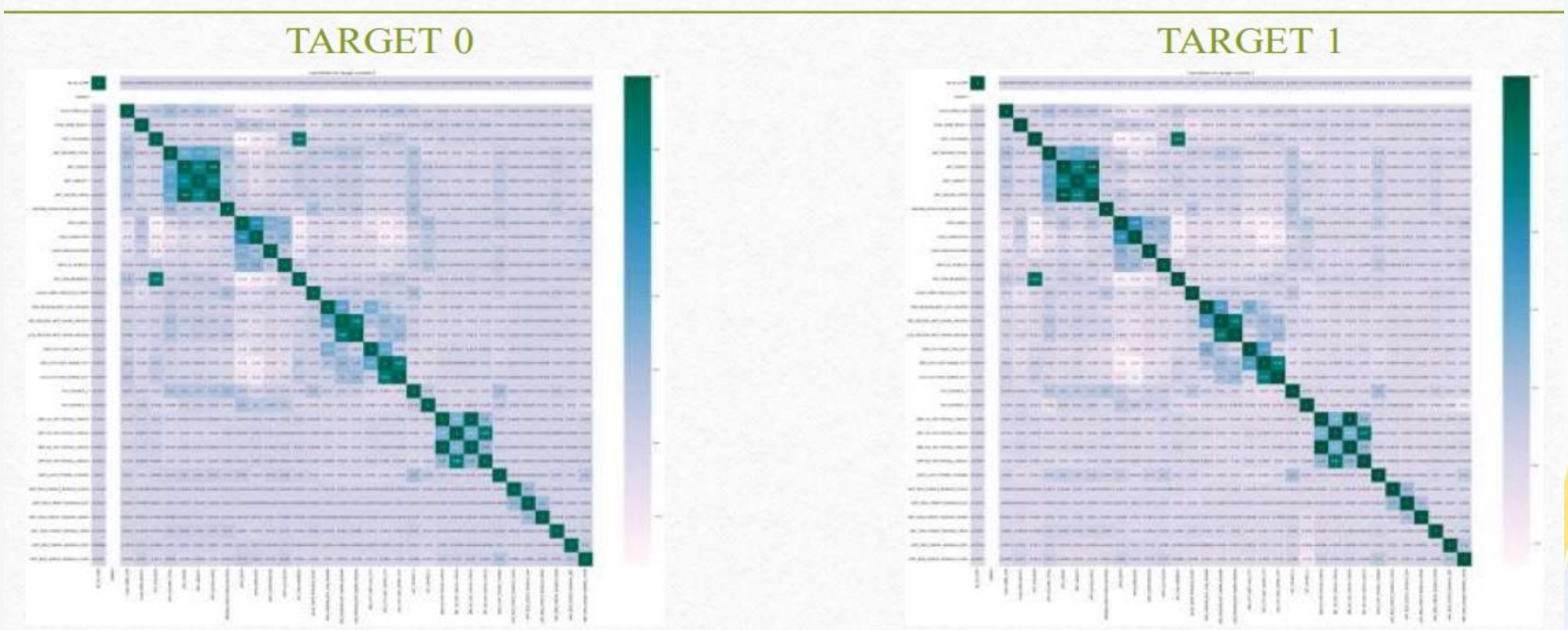
4. Performed multivariate analysis on me0 and m1 with respective to NAME_CONTRACT_STATUS and NAME_CLIENT_TYPE with different categorical variables once and obtained the insights for the EDA

DATA IMBALANCE

- application data data_set of application data having variable df is really imbalanced. TARGET 1 population is around 8.1% and TARGET 0 population is around 91.9%
- Ratio achieved of data imbalance is 11.3



CORRELATION OF TARGET VALUES WITH DIFFERENT
NUMERICAL VALUES WITHIN APPLICATION DATA SET

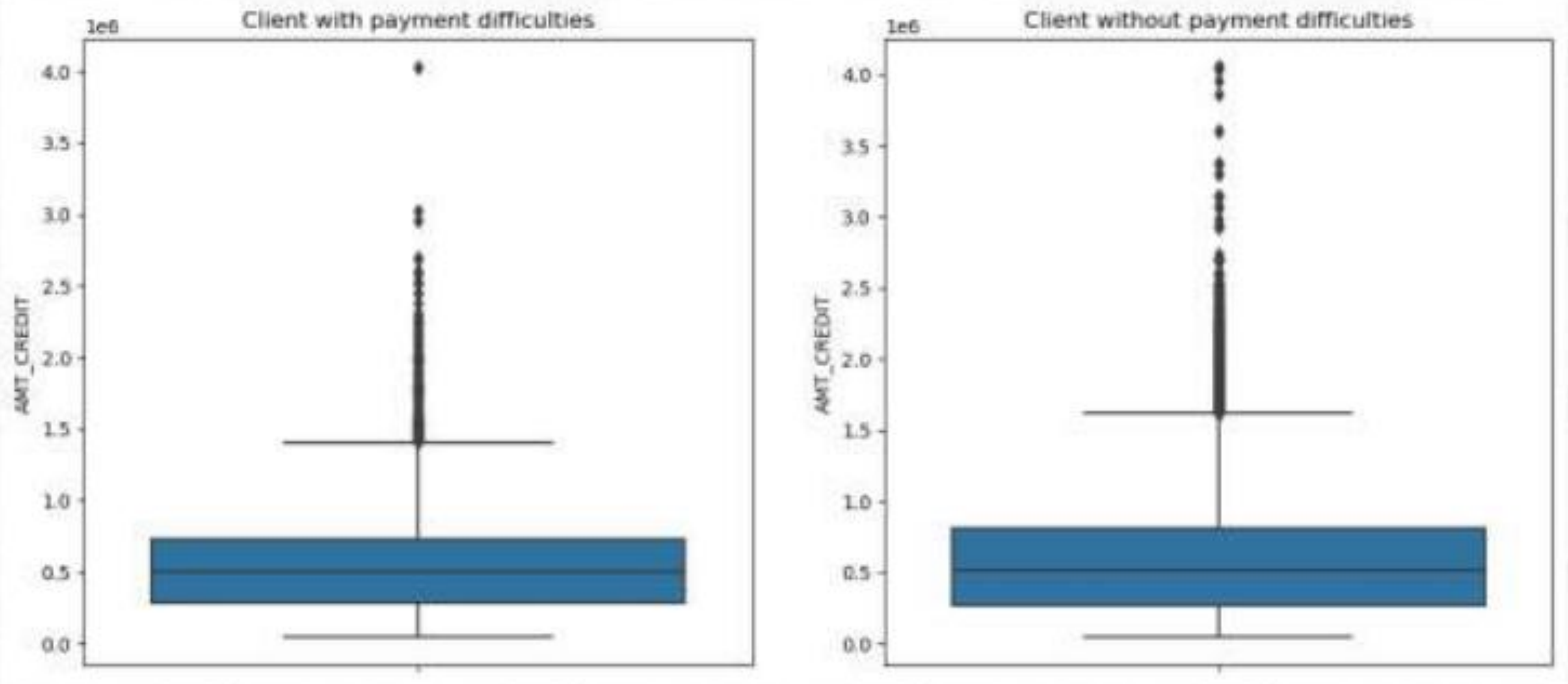


TARGET 0

TARGET 1

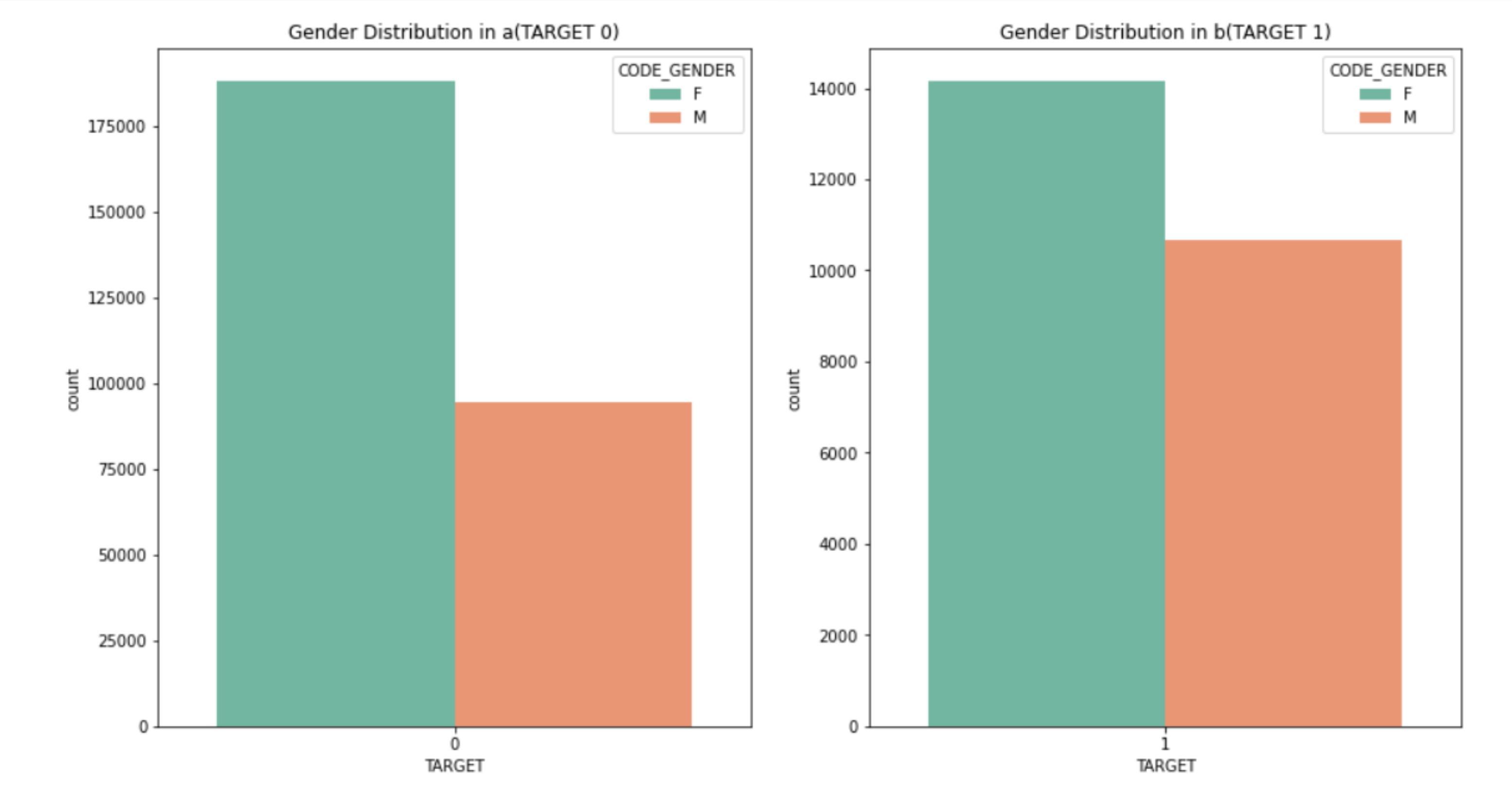
UNIVARIATE ANALYSIS AMT_CREDIT BOXPLOT

Greater quartile for client without payment difficulty
More outlier for client with payment difficulites.



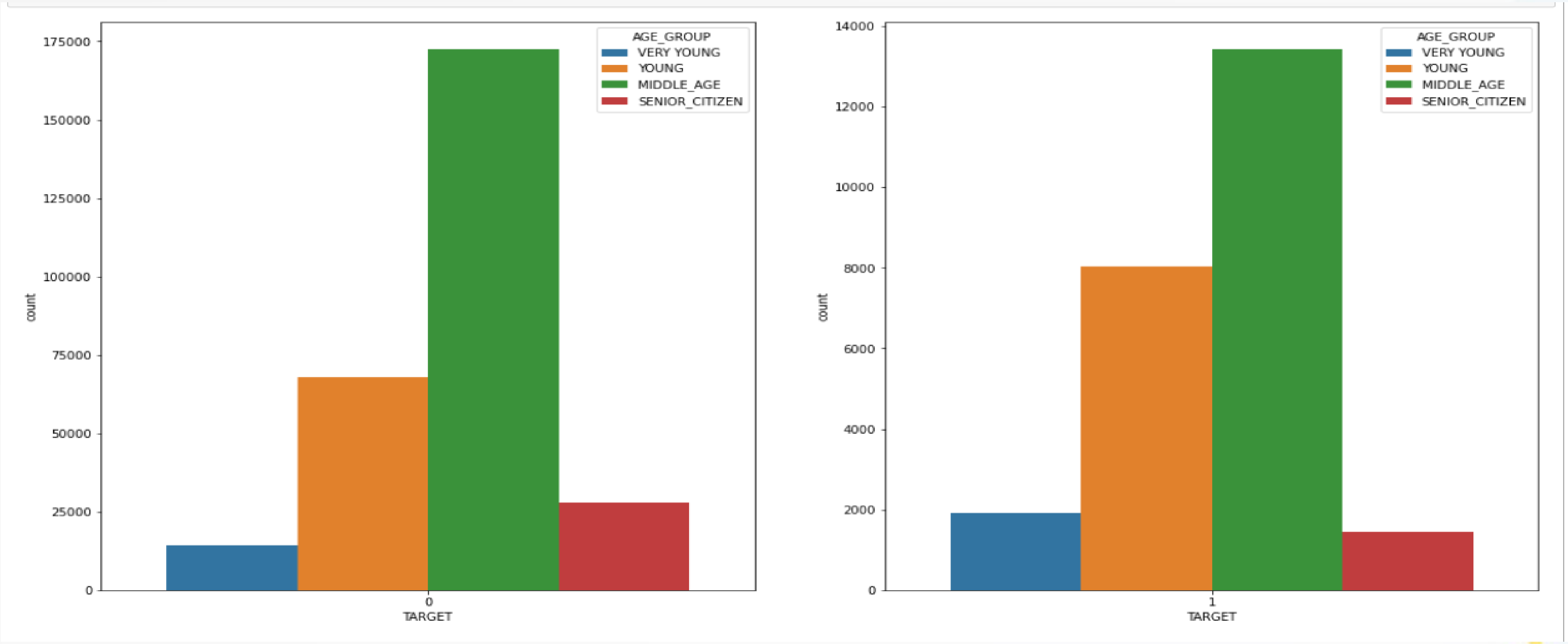
GENDER DISTRIBUTION

It shows Female clients applied higher than male clients for loan



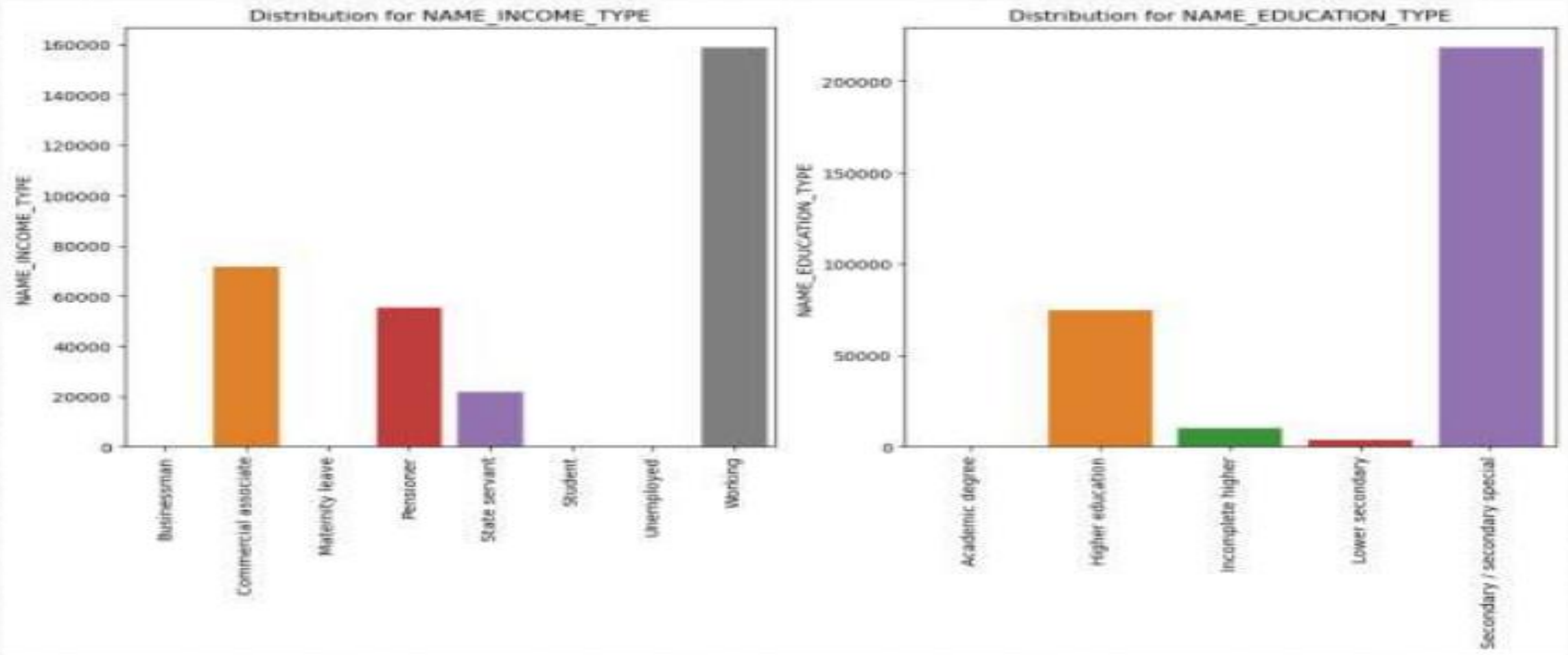
Age Distribution on Target 0 and 1

Middle Age(35-60) the group seems to applied higher than any other age group for loans in the case of Defaulters as well as Non-defaulters.
Also, Middle Age group facing paying difficulties the most.
While Senior Citizens(60-100) and Very young(19-25) age group facing paying difficulties less as compared to other age groups.



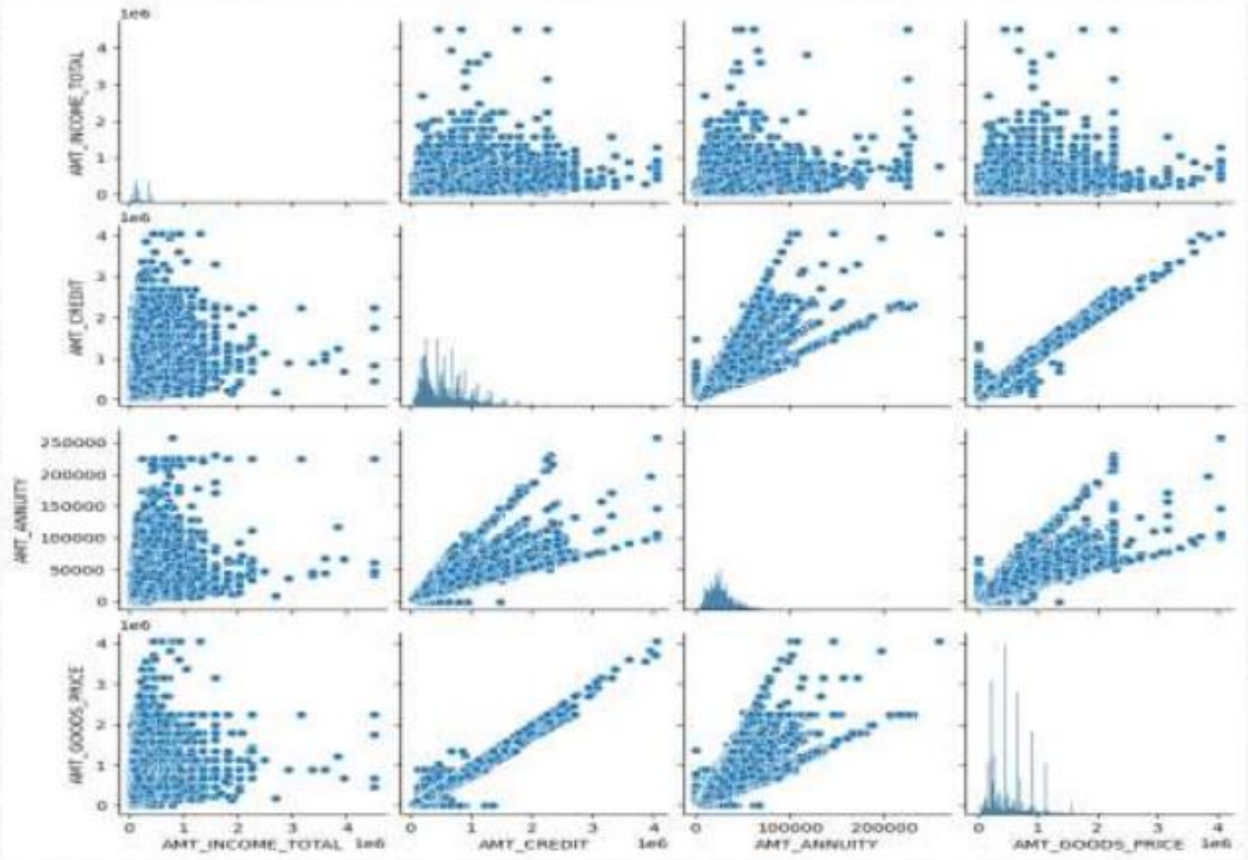
UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLES

Clients with income type category Businessman, Maternity leave , Pentioner and student is very less.
Clients with income type Working is highest in the data
Few clients are present with education type Academic degree and lower Secondary
Clients with education type secondary / secondary special is highest in data



BIVARIATE ANALYSIS FOR NUMERICAL VS NUMERICAL

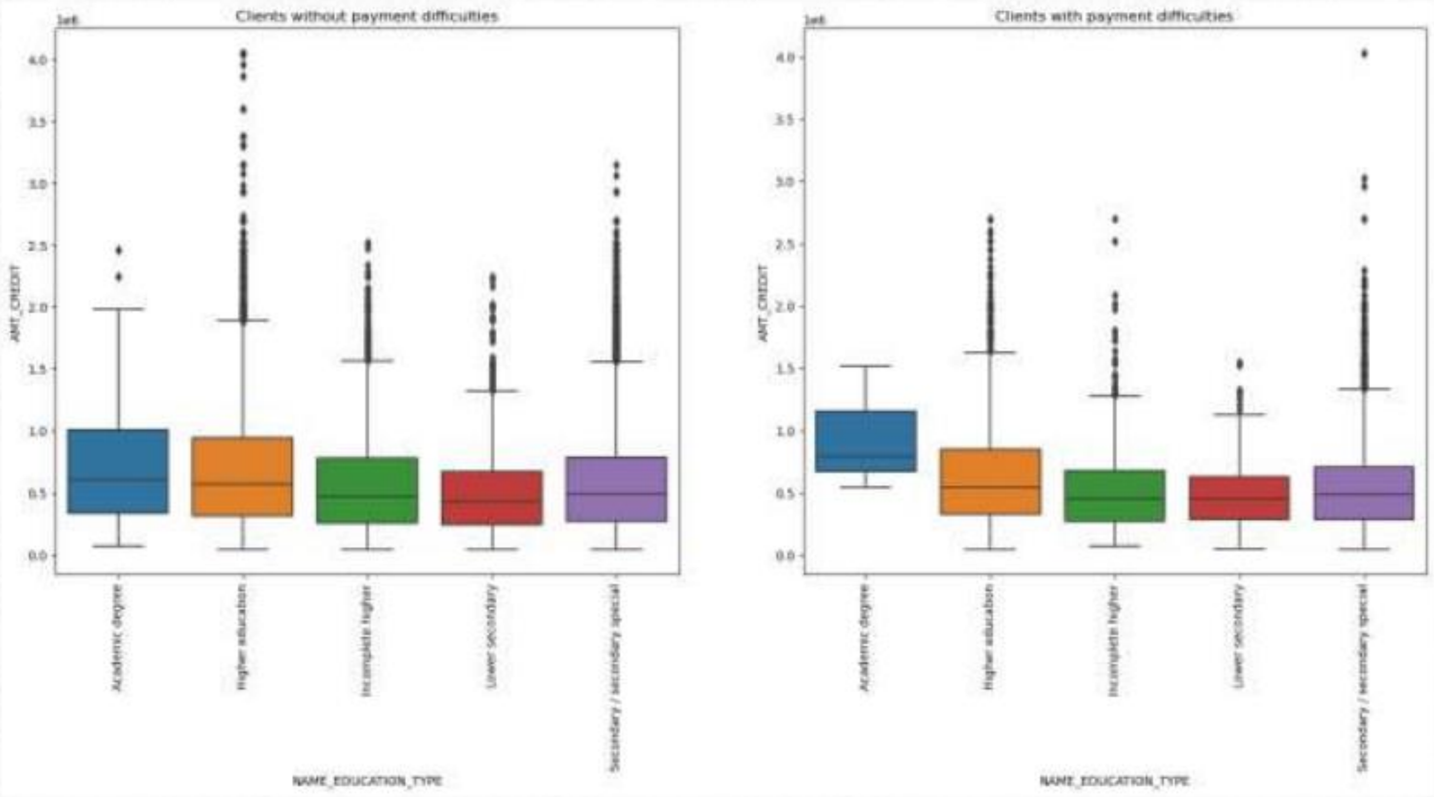
Linear CORRELATION between AMT_GOODS_PRICE vs AMT_CREDI



BIVARIATE ANALYSIS FOR NUMERICAL VS CATEGORY

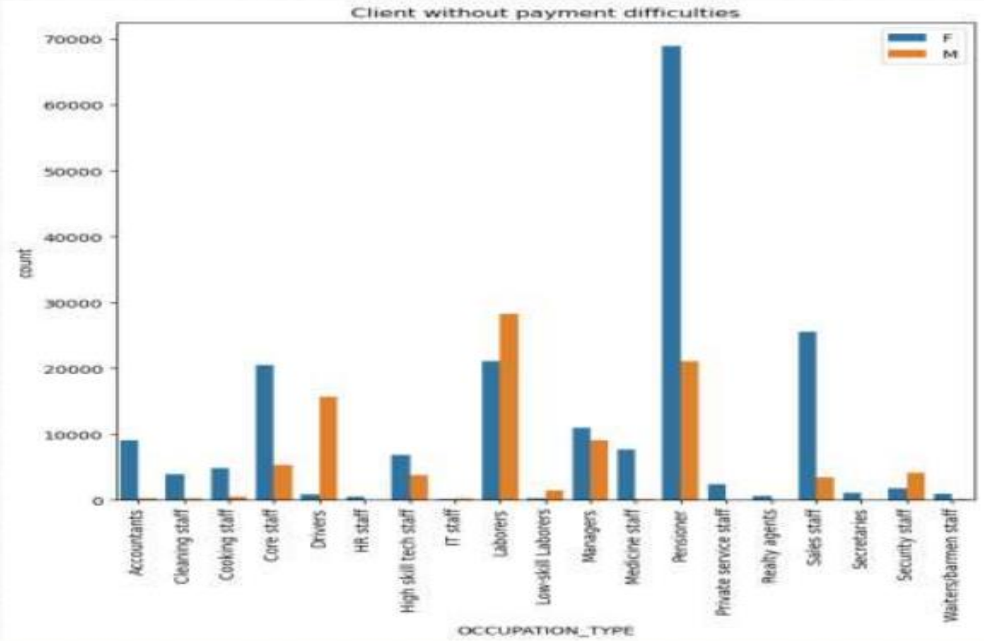
Outliers are high for customer without payment difficulties Less outliers for Academic degreee for both targets variables

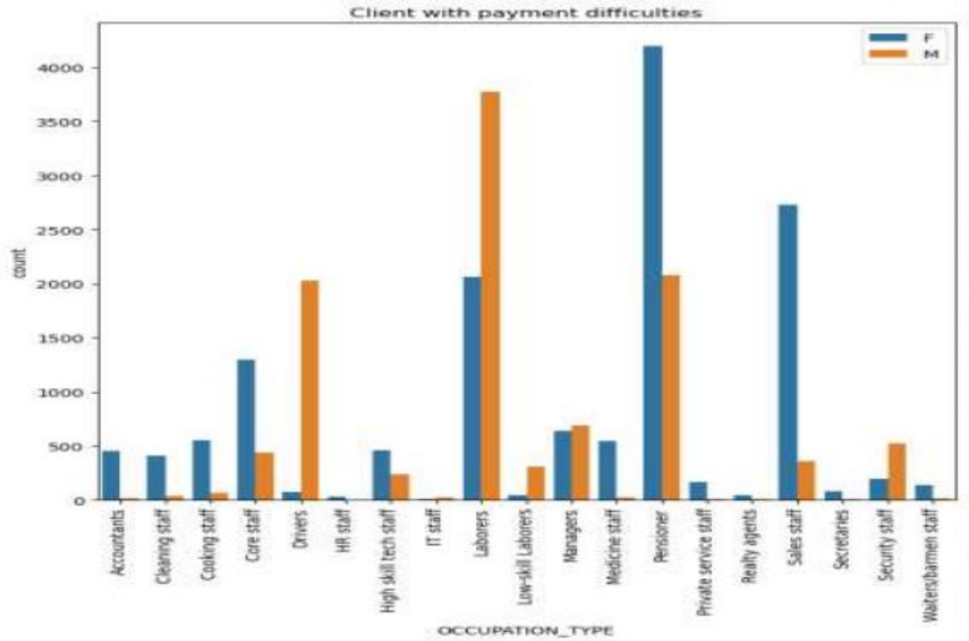
Mean for academic degree is highest and incomplete higher is low



BIVARIATE ANALYSIS CATEGORICAL VS CATAGORICAL VALUES

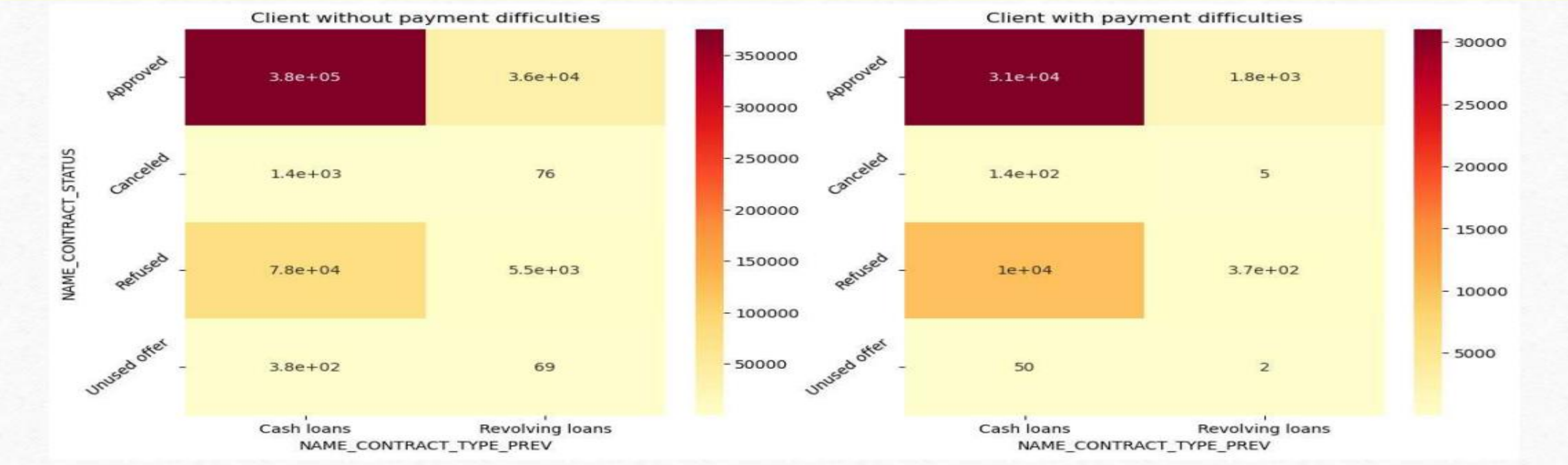
- Very less male clients present in Reality agents, Private service staff, Medicine staff, Waiters/barmen staff
- Number of male cilents is more in Drivers, Security staff and Laborers
- Very few Female and male clints are present having occupation type HR staff and IT staff for both Customer without payment
- difficulties and Customer with payment difficulties





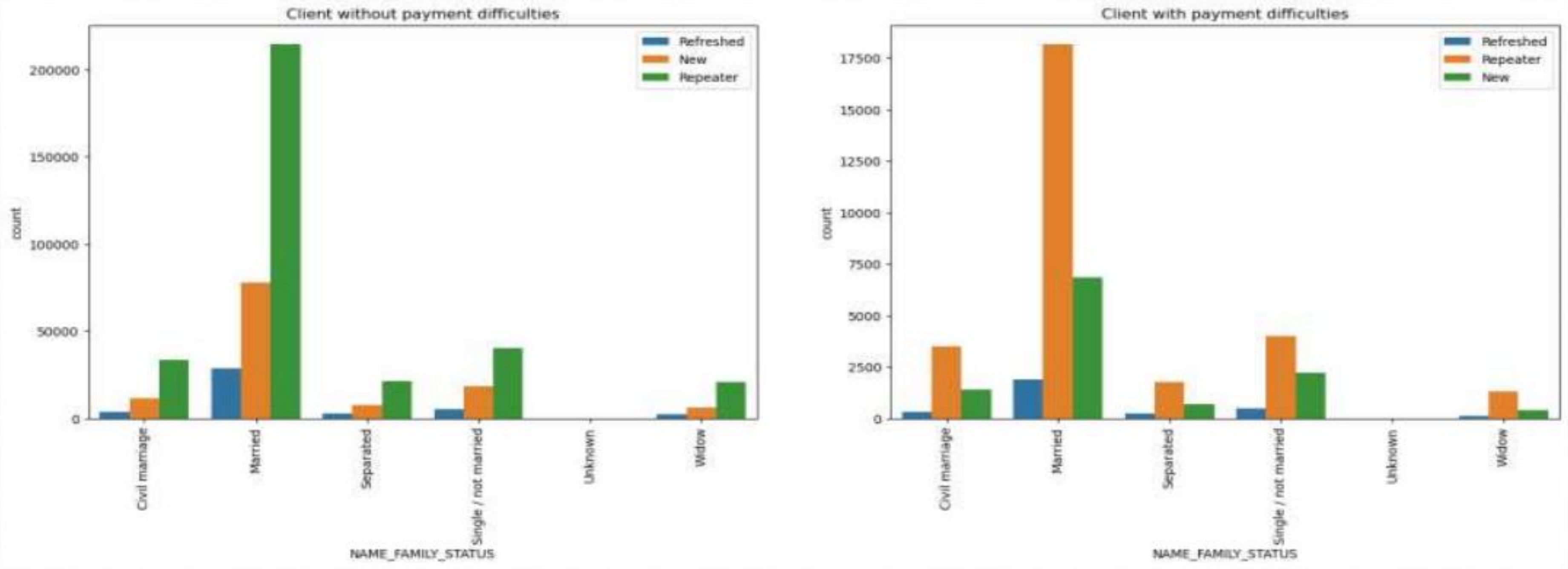
DISTIBUTIONS OF MERGED DATASESET NAME_CONTRACT_STATUS VS NAME_CONTRACT_TYPE_PREV for with and without payment difficulties .

Large number of clients approved for cash loans for both types of clients who are having difficulty with payments or not .
Very less number of resolving loans were cancelled for clients with payment difficulties



DISTIBUTIONS OF MERGED DATASESET

- NAME_CLIENT_TYPE' with 'NAME_FAMILY_STATUS' for with and without payment difficulties
- Married Clients are highest and Widow clients are lowest in both categories
- New clients are more without having payment difficultie



So here are few points what we came to know from the EDA process is that

1. In general post chances to get defaulted fast is more as cash loans were more distributed than resolving loans

2. Chances of getting loan amount back is not with the working income type rather than category such as bussinessman maternaity leave and pensioner should be focussed more for reppaying the loan

3. Repair category is having higher number of unsuccesfull repayment of the loan . So try not to focus more on that side

4. Middle Age(35-60) the group seems to applied higher than any other age group for loans in the case of Defaulters as well as Non-defaulters.
Also, Middle Age group facing paying difficulties the most.
While Senior Citizens(60-100) and Very young(19-25) age group facing paying difficulties less as compared to other age groups.

THANKS