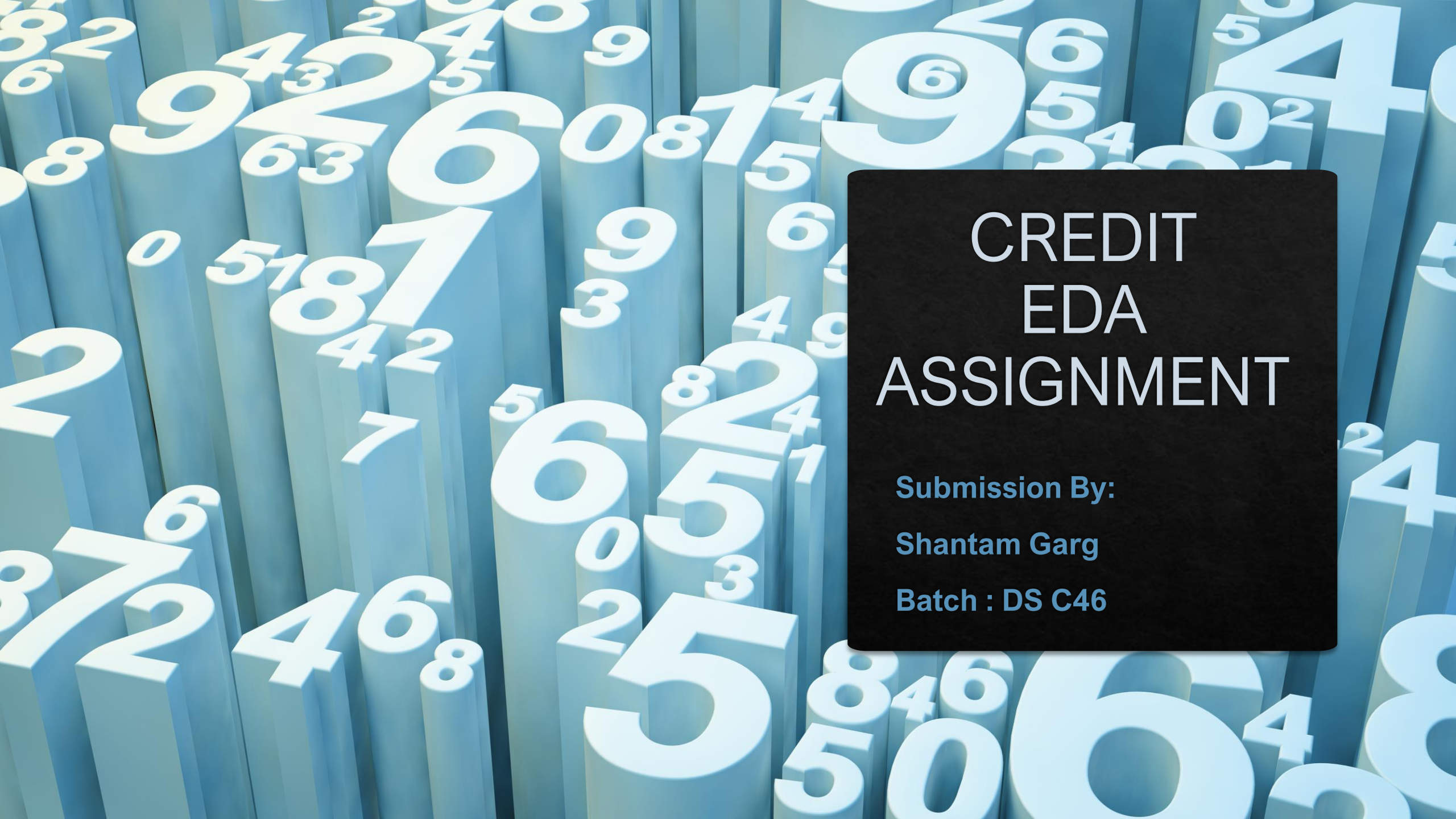


CREDIT EDA ASSIGNMENT

Submission By:

Shantam Garg

Batch : DS C46

Introduction

- In this assignment we are applying some analytical and visualization techniques on a real business scenario. We will also develop basic understanding as to how data can be used to minimize the risk of losing money while lending to customers.

Business Understanding

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Assignment Objective

- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Steps to Approach

- Understanding of Given Data (Application Data)
- Data Cleaning :
  - Identification of missing values and Treatment
  - Outlier and Treatment
- Sanity Checks :
  - Checking if the data present in the column is correct and relevant
- Data Imbalance
- Binning
- Univariate Analysis & Univariate Segmented Analysis
- Bivariate Analysis
- Analysis of Previous Application Data set with above steps
- Merging of Data frame Application data and Previous Application & Analysis
- Conclusion

Data Understanding (Application Data)

- Checking Shape, Info, D-Types and Describe of the data to get quick understanding of the data.

```
#Checking the Shape of the Dataframe:
print("Shape of the Dataframe is:",df0.shape)
```

```
Shape of the Dataframe is: (307511, 122)
```

```
#Checking the info of the Dataframe:
print(df0.info(verbose = True, show_counts = True))
```

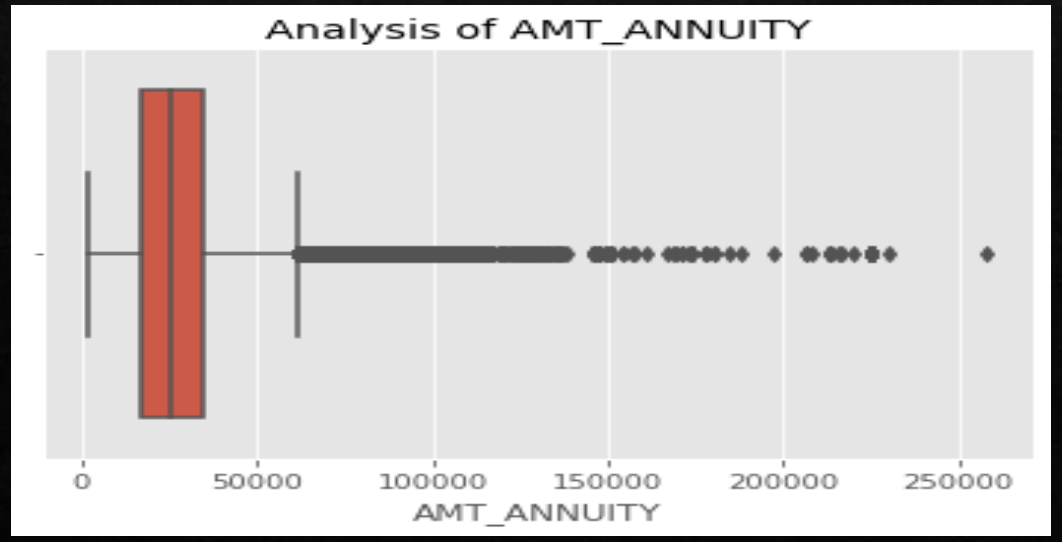
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 122 columns):
```

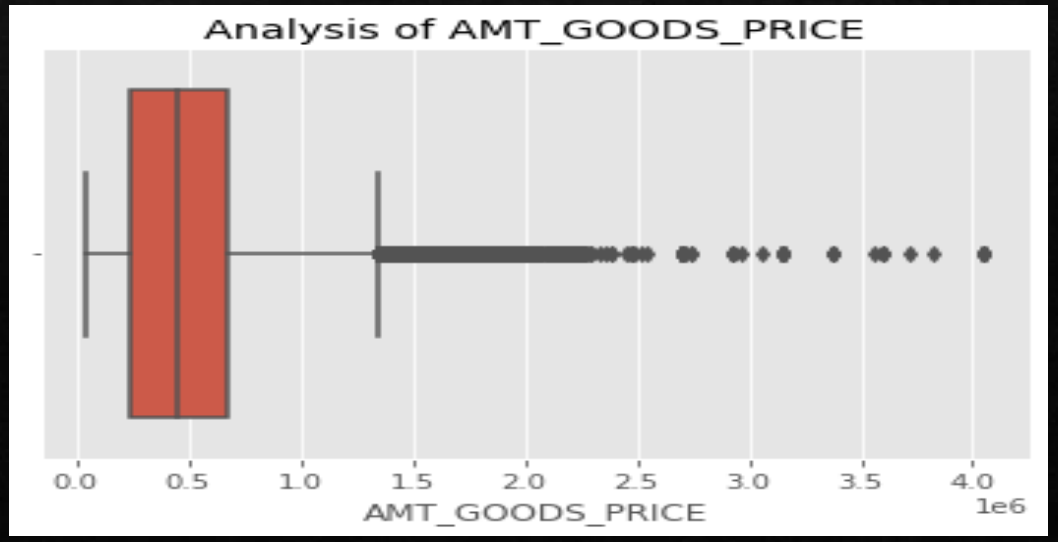
```
#Lets Check data type for all the columns:
df0.dtypes
```

```
SK_ID_CURR                    int64
TARGET                        int64
NAME_CONTRACT_TYPE           object
CODE_GENDER                  object
FLAG_OWN_CAR                 object
FLAG_OWN_REALTY              object
CNT_CHILDREN                  int64
AMT_INCOME_TOTAL            float64
AMT_CREDIT                  float64
AMT_ANNUITY                 float64
AMT_GOODS_PRICE             float64
```

Data Cleaning

- Identification of Missing Values and Treatment :
  - There were several Columns with missing value percentage greater than 40% so we dropped them.
  - Remaining columns with missing values we imputed them with mean/median/mode as required.
- Outliers:
  - There were several columns which have outliers present in them.





Sanity Checks

- From observing the data we saw that some column contains values which are negative. we converted it into positive for easier analysis.
- Columns with Negative value are DAYS_BIRTH , DAYS_EMPLOYED , DAYS_REGISTRATION , DAYS_ID_PUBLISH , DAYS_LAST_PHONE_CHANGE.

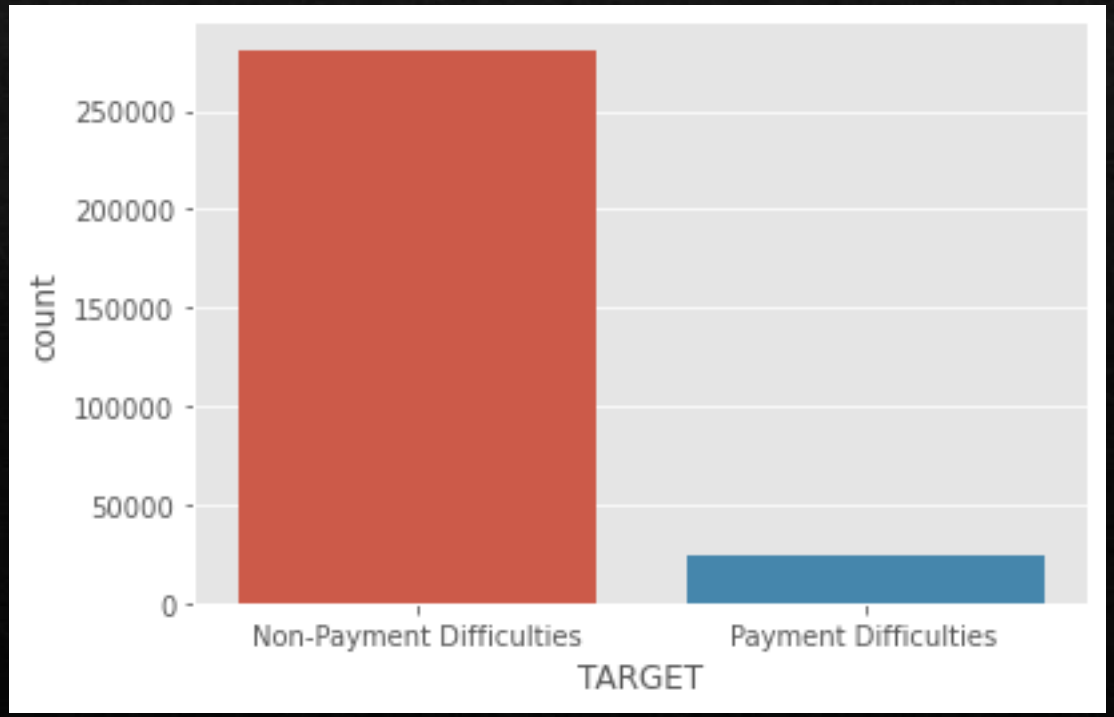
```
#Converting these values to (+)ve values:
df0["DAYS_BIRTH"] = df0["DAYS_BIRTH"].abs()
df0["DAYS_EMPLOYED"] = df0["DAYS_EMPLOYED"].abs()
df0["DAYS_REGISTRATION"] = df0["DAYS_REGISTRATION"].abs()
df0["DAYS_ID_PUBLISH"] = df0["DAYS_ID_PUBLISH"].abs()
df0["DAYS_LAST_PHONE_CHANGE"] = df0["DAYS_LAST_PHONE_CHANGE"].abs()

#Checking if the values are properly converted:
df0[["DAYS_BIRTH", "DAYS_EMPLOYED", "DAYS_REGISTRATION", "DAYS_ID_PUBLISH","DAYS_LAST_PHONE_CHANGE"]]
```

	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
0	9461	637	3648.0	2120	1134.0
1	16765	1188	1186.0	291	828.0
2	19046	225	4260.0	2531	815.0
3	19005	3039	9833.0	2437	617.0
4	19932	3038	4311.0	3458	1106.0

Data Imbalance

- When we observed the data set it was highly imbalanced almost 91.9% clients were Non-Payment Difficulties and about 8.1% for Payment Difficulties.
- From the observations we can say that the data is skewed towards the Non-Payment Difficulties & Data Imbalance Ratio is : 11.36%



Binning

- For more insightful understanding we binned the following columns:
  - DAYS_BIRTH
  - AMT_INCOME_TOTAL

```
#Let's convert age of client in years for more easier to understand format:
df0["Age_yrs"] = df0.DAYS_BIRTH.apply(lambda x: round((x/365),0))
```

```
#Let's assign an age group to the each client:
df0['AGE_GROUP'] = pd.cut(df0.Age_yrs,bins=np.arange(20,71,10))
df0.AGE_GROUP.value_counts()
```

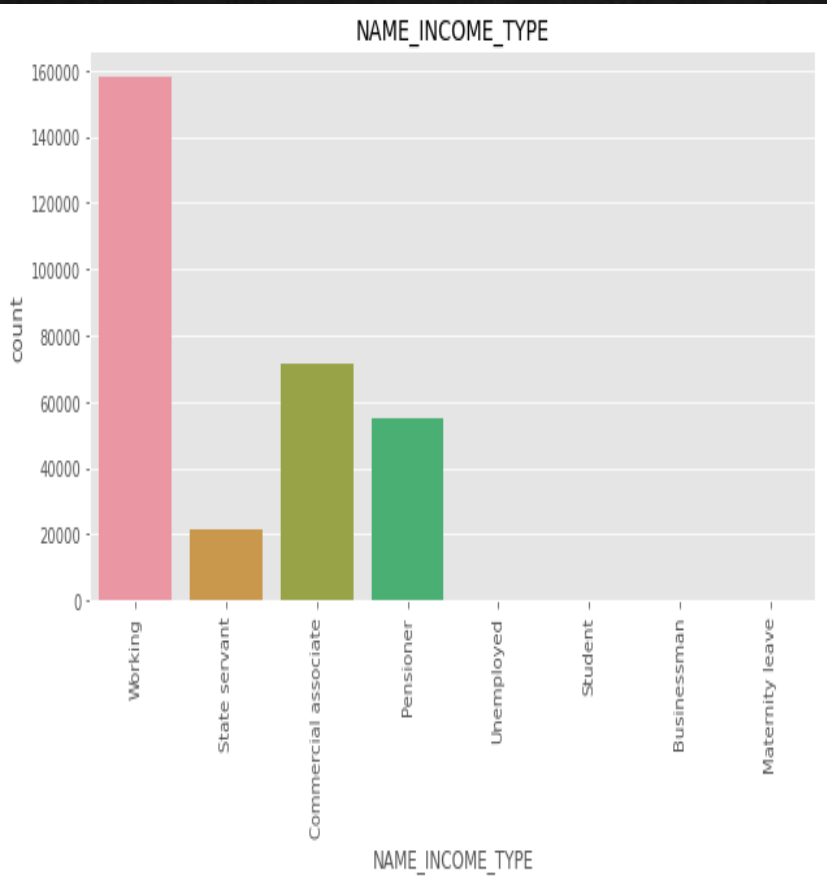
```
(30, 40]    82367
(40, 50]    75112
(50, 60]    67558
(20, 30]    48575
(60, 70]    32220
Name: AGE_GROUP, dtype: int64
```

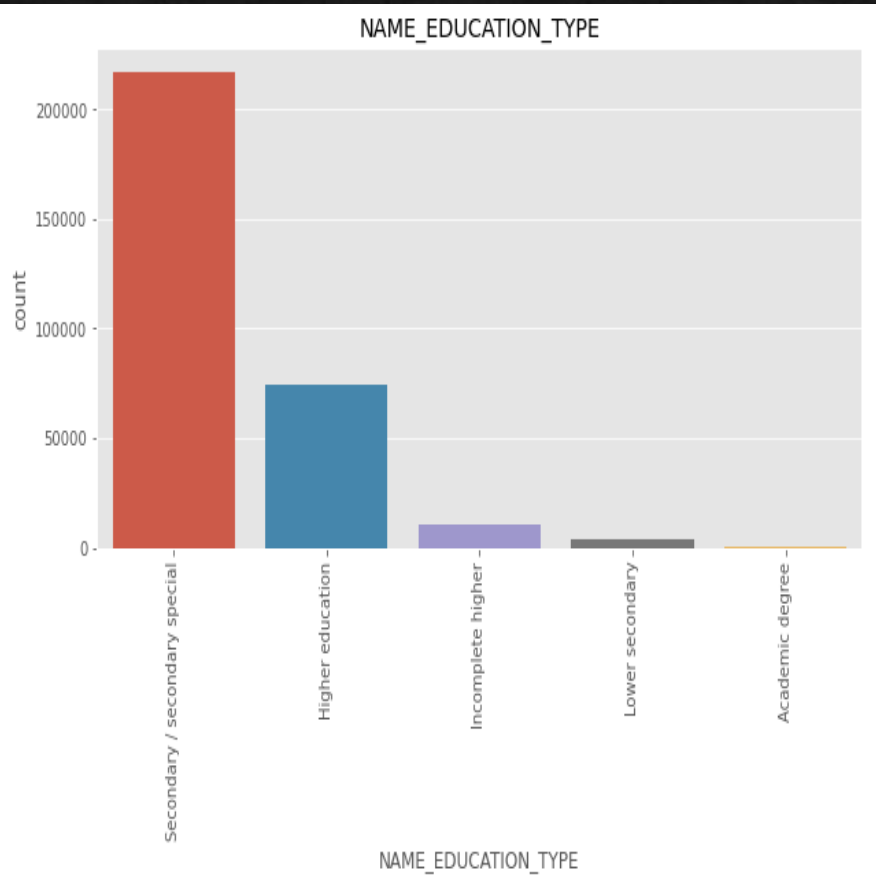
```
#Now let's assign the Income Groups to the each client's income:
df0["INCOME_GROUPS"]=pd.qcut(df0.AMT_INCOME_TOTAL,q=[0,0.2,0.4,0.6,0.8,1],labels=['Very Low','Low','Medium','High','Very High'])
df0.INCOME_GROUPS.value_counts()
```

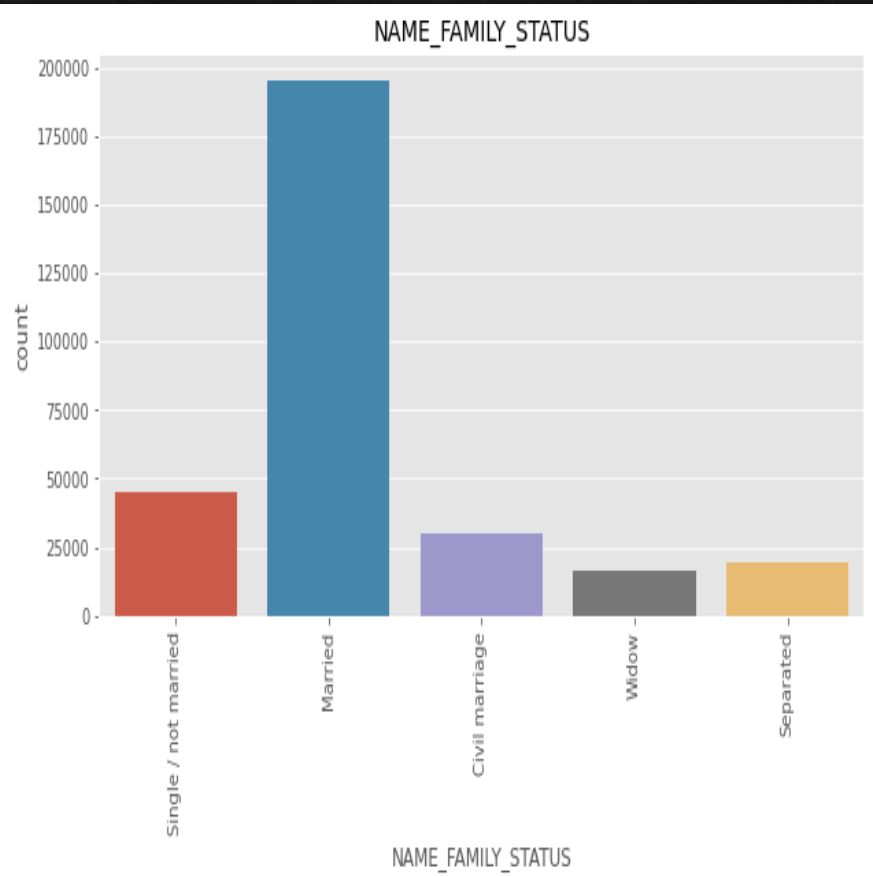
```
Low          85292
High         75141
Very Low     63314
Very High    46784
Medium       35301
Name: INCOME_GROUPS, dtype: int64
```

Univariate Analysis

- *Inference* : We can say that most client are working, have secondary education and are married.

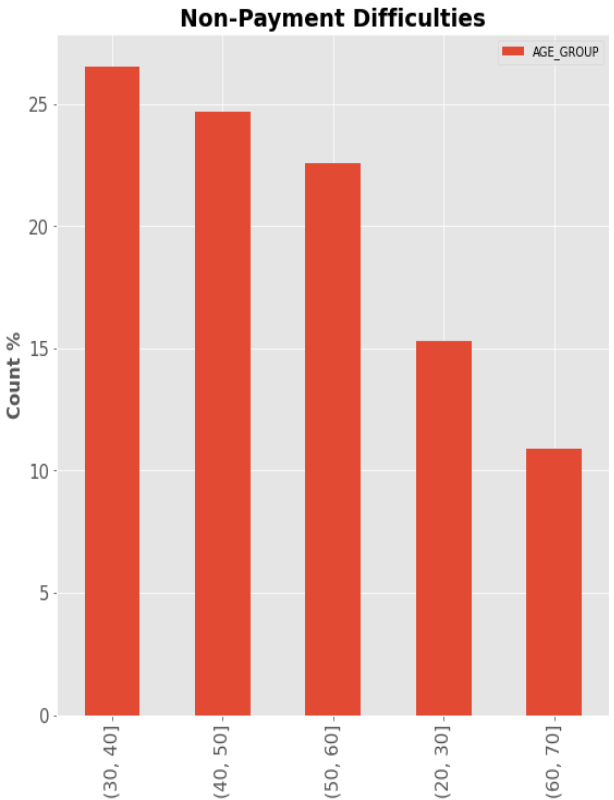


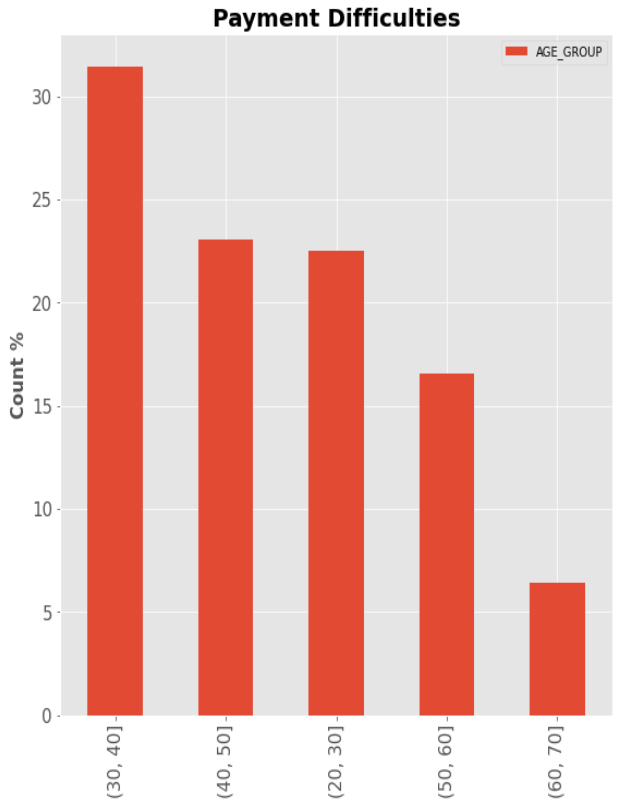


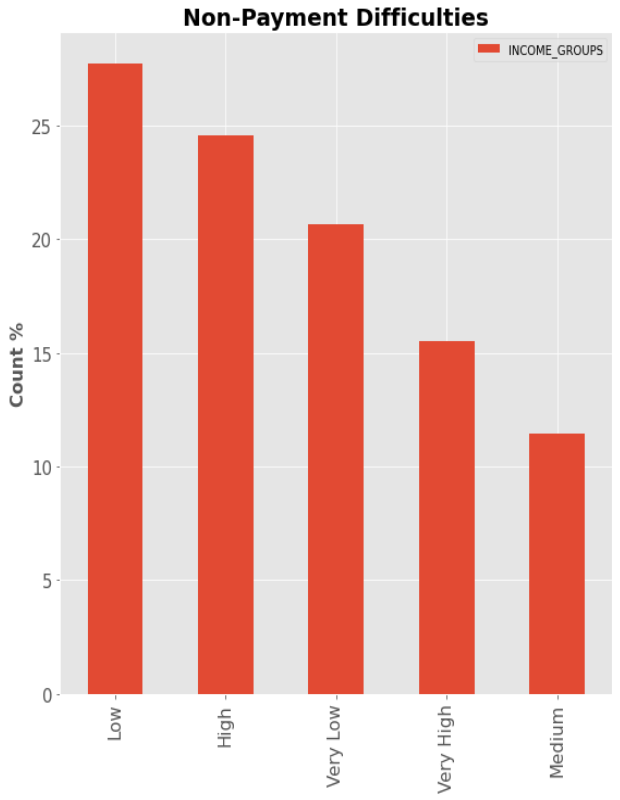


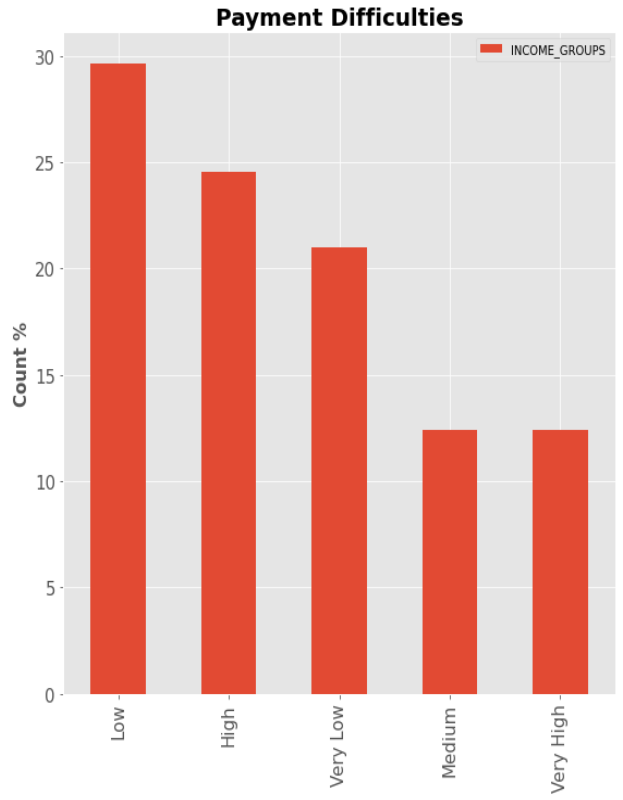
Univariate Segmented Analysis

- *AGE_GROUP* : (30-40) Age group people have a bit higher difficulty in paying loan while (60-70) age group people have lesser difficulty in paying their loans.
- *INCOME_GROUPS* : People with very high salary have lesser difficulties in paying their loan but most number of people who go for loan are people with low salary.



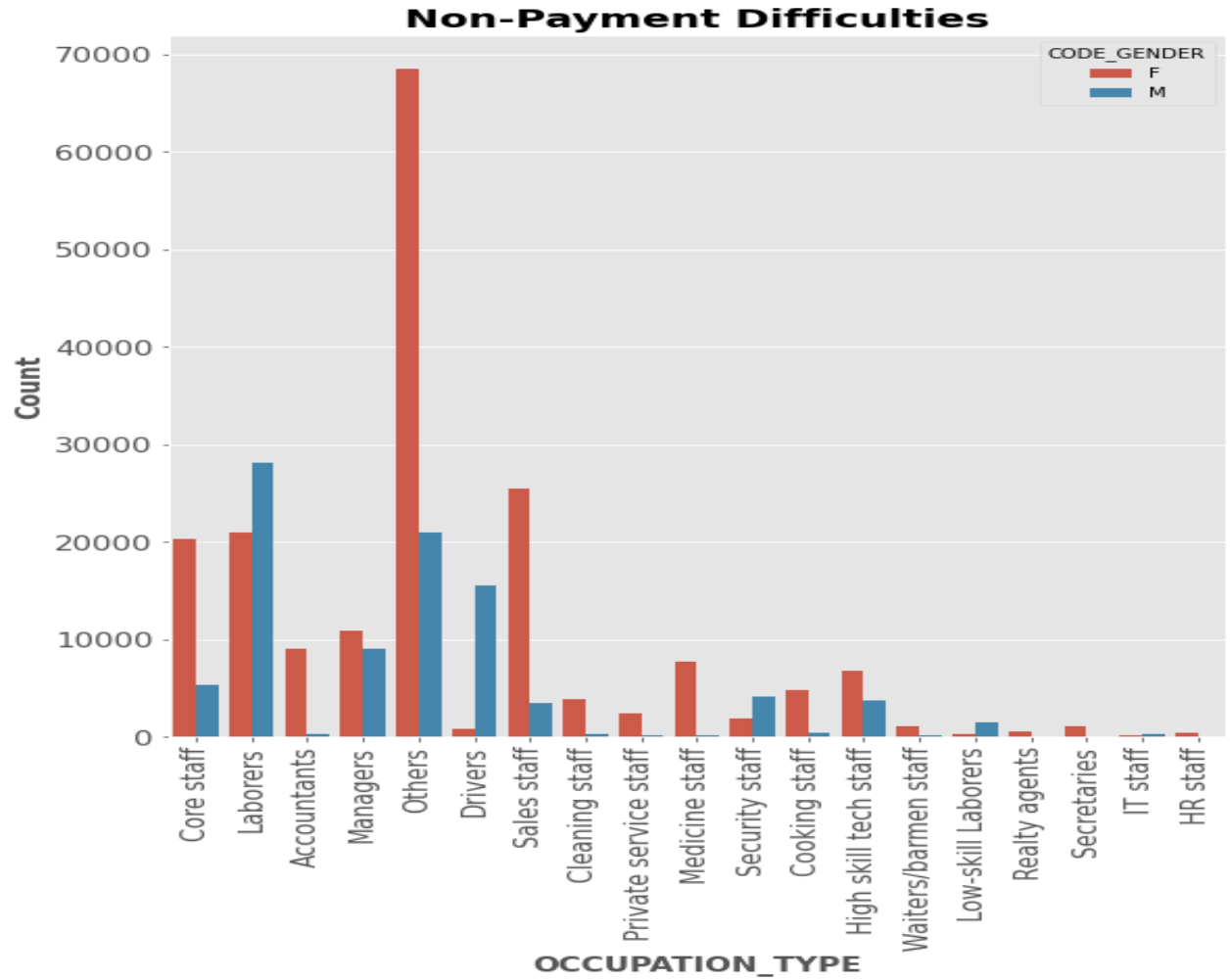


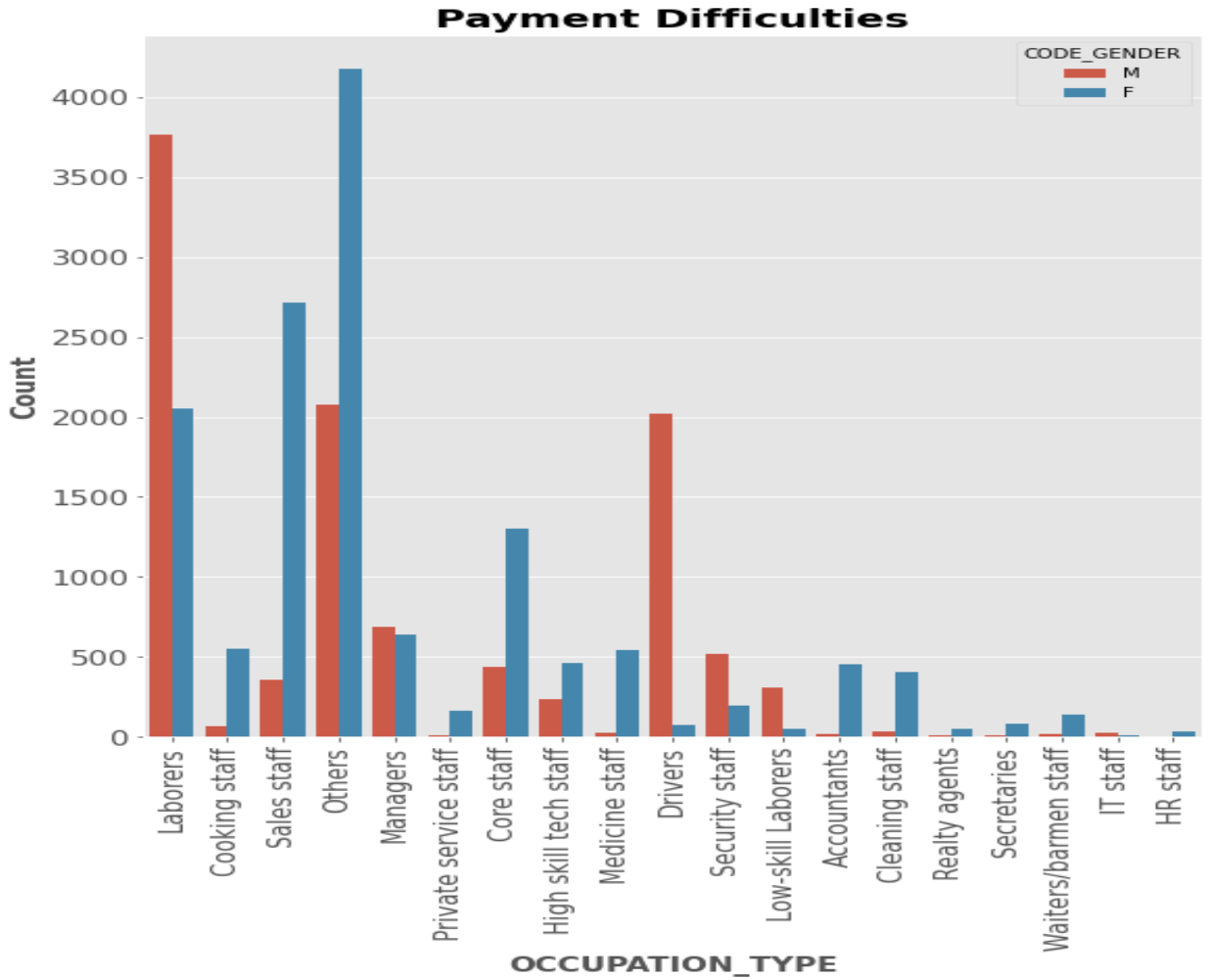




Bivariate Analysis

- *Inference* : We can see the sharp increase in laborers category in payment default category and also male laborer have most difficulties in paying loan.

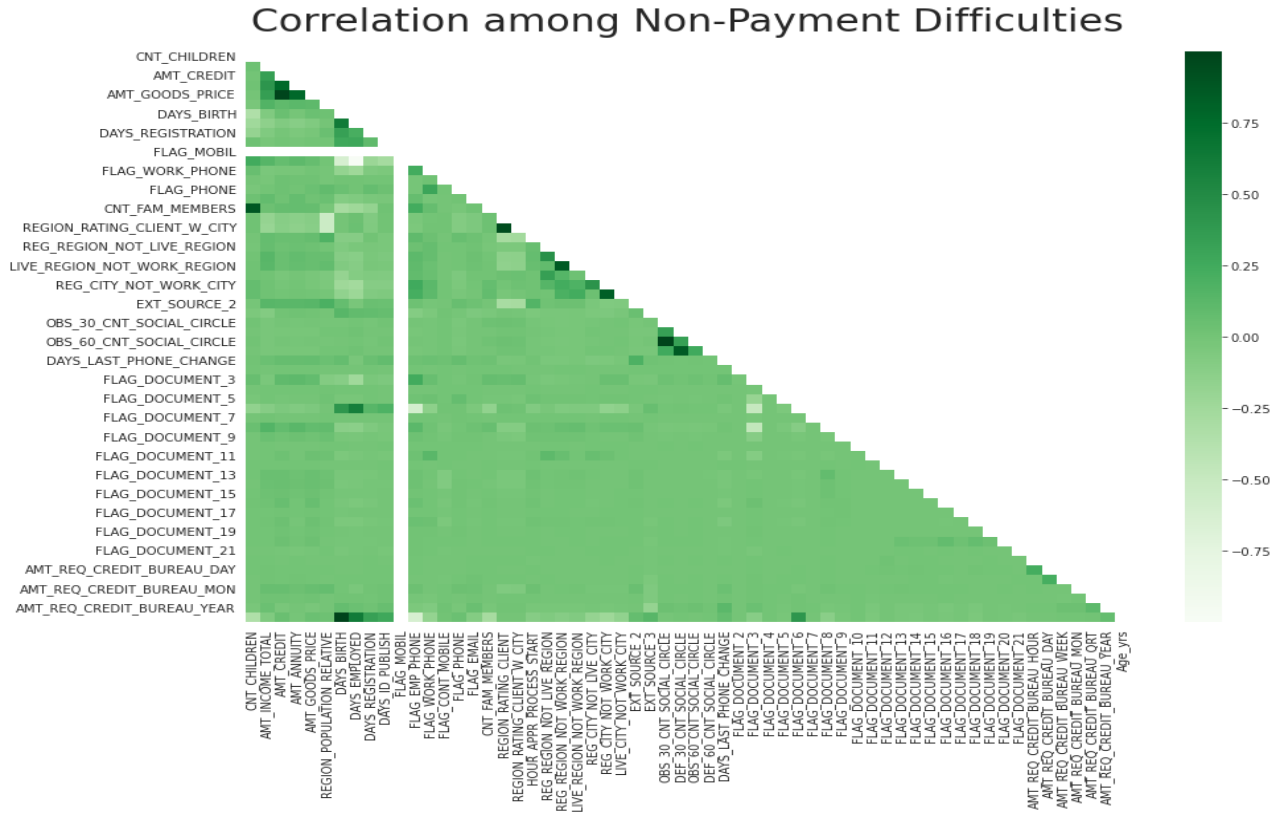


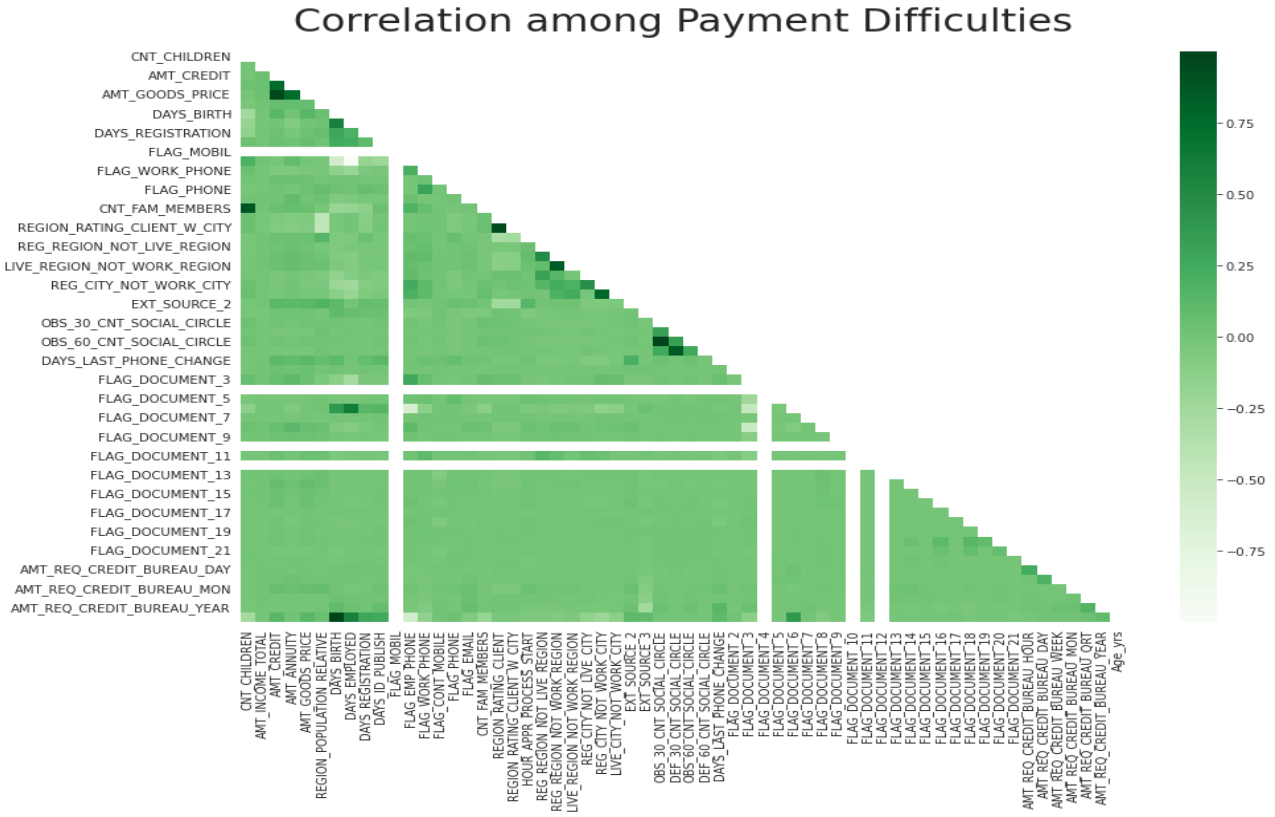


Heatmap for Correlation matrix

- Following Variables have high correlation in both Non-payment difficulties and Payment Difficulties:

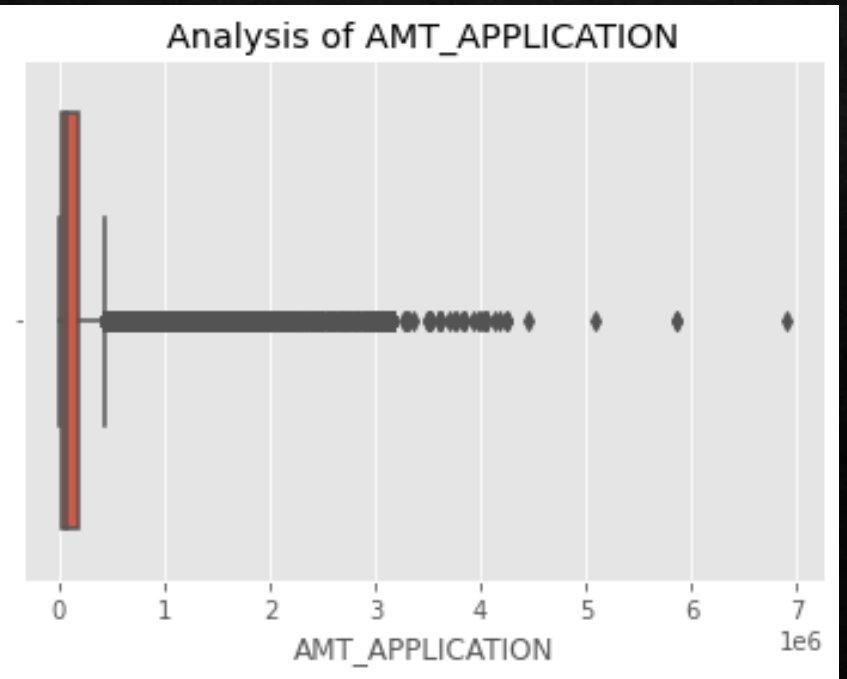
  - AMT_GOODS_PRICE and AMT_CREDIT
  - CNT_FAM_MEMBERS and CNT_CHILDREN
  - AMT_ANNUITY and AMT_CREDIT
  - AMT_ANNUITY and AMT_GOODS_PRICE

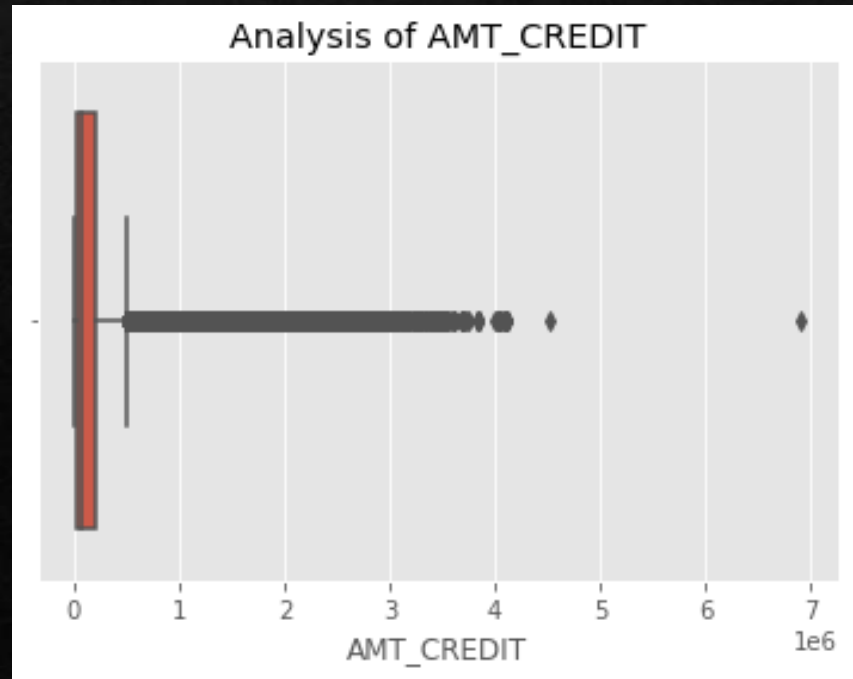




Data Understanding (Previous Application Data)

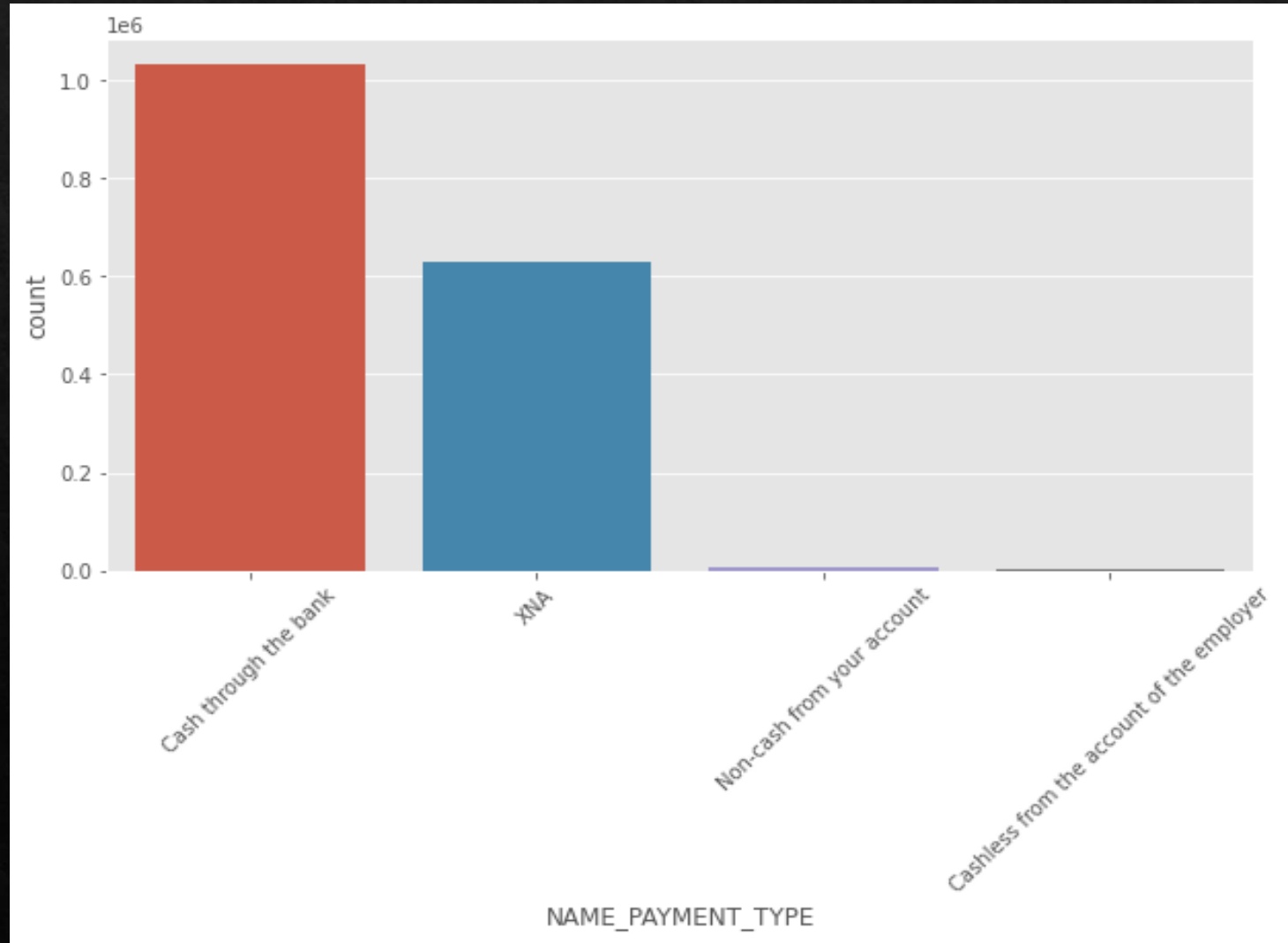
- Checking head, info, describe and D-Types of the data to get a basic understanding of the data.
- **Data Cleaning** : After checking the missing values percentage we dropped the columns with missing value greater than 40%. And we can impute the rest of the missing value with mean/median/mode as required.
- There were several columns with outliers present in them.





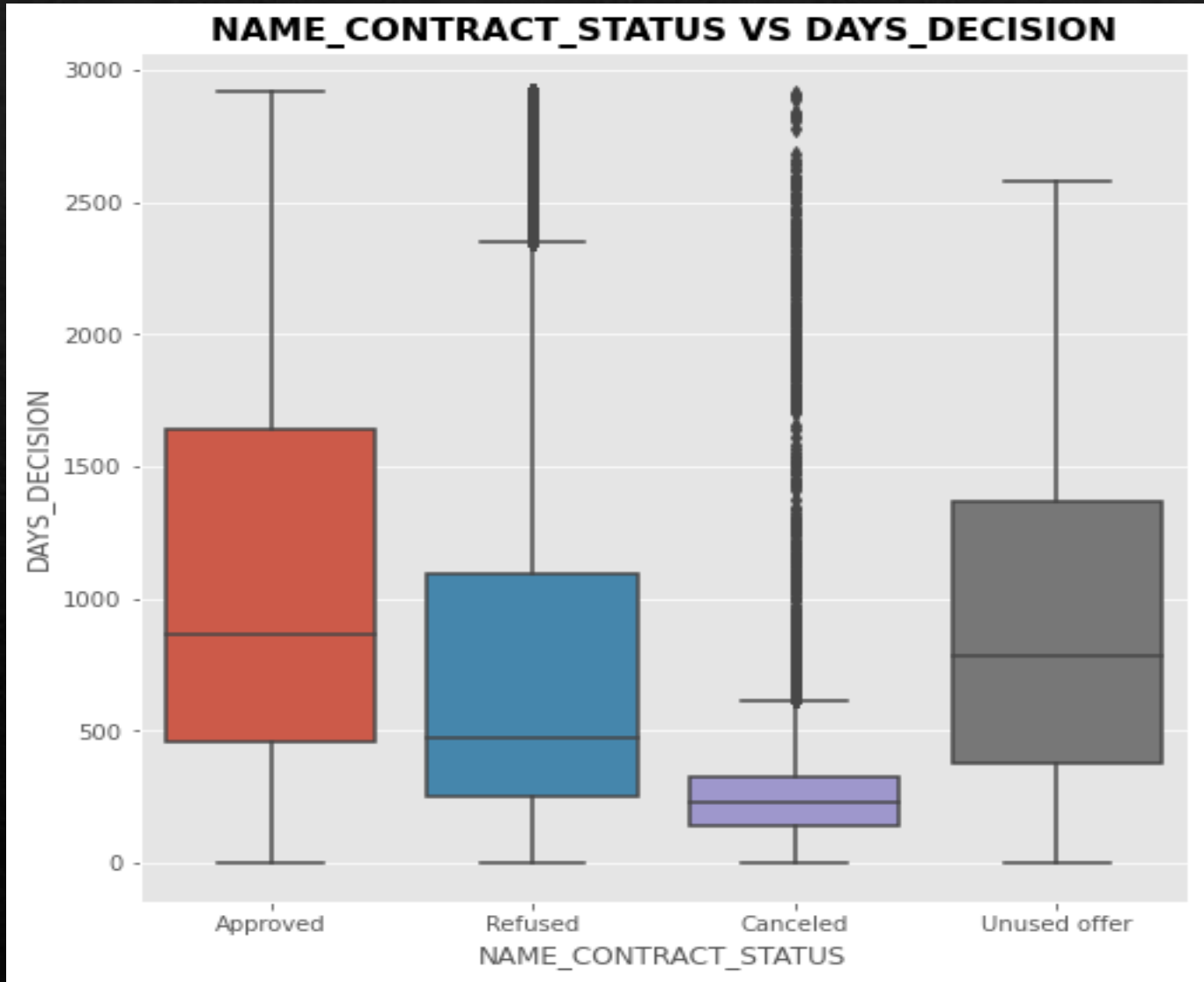
Univariate Analysis (Previous Application Data)

- Most of the client used Cash through the bank mode, while cashless from the account is the least favorable.



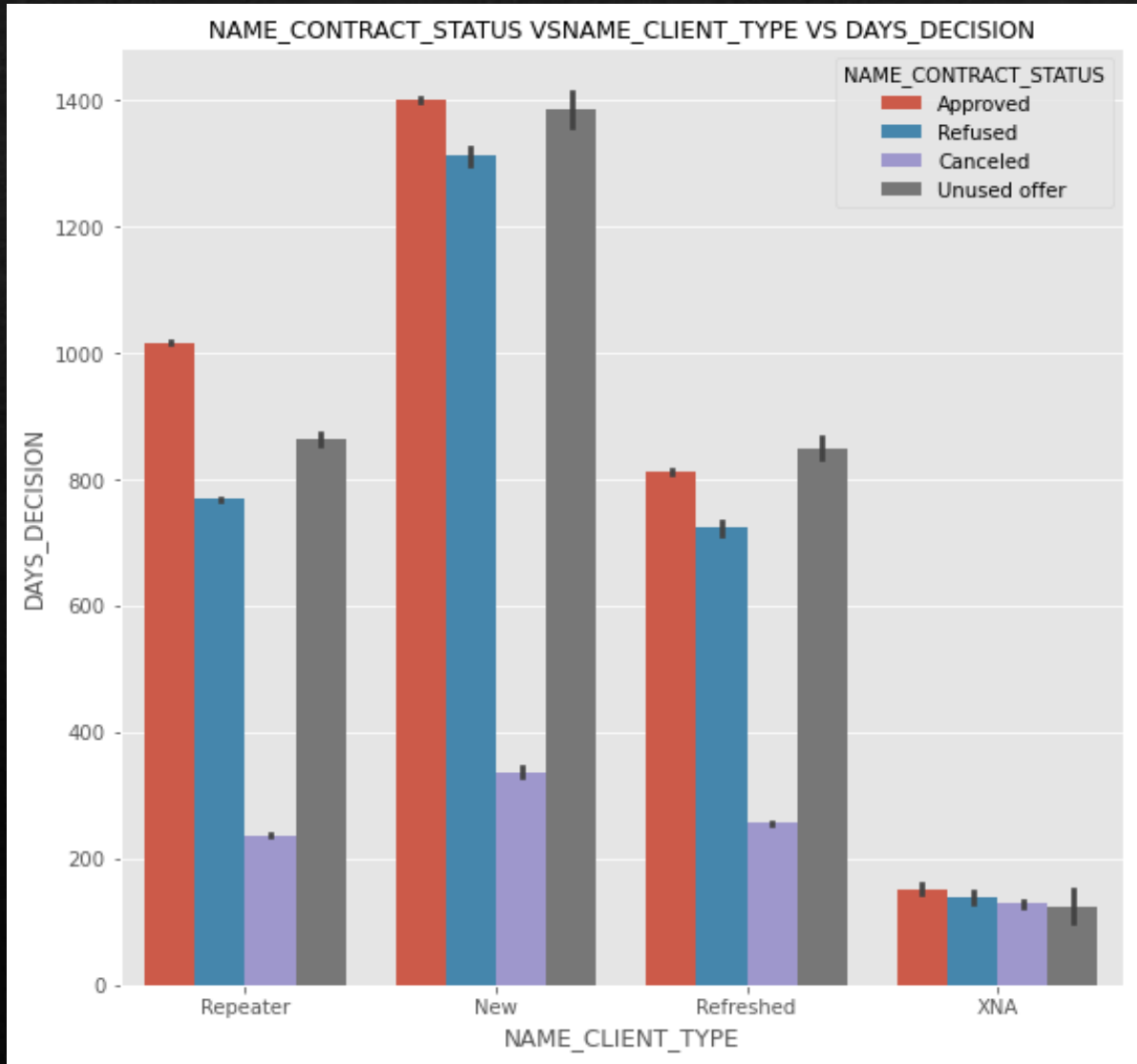
Bivariate Analysis (Previous Application Data)

- Bank takes large number of days to approve a loan in comparison with to cancelling or refusing a loan.



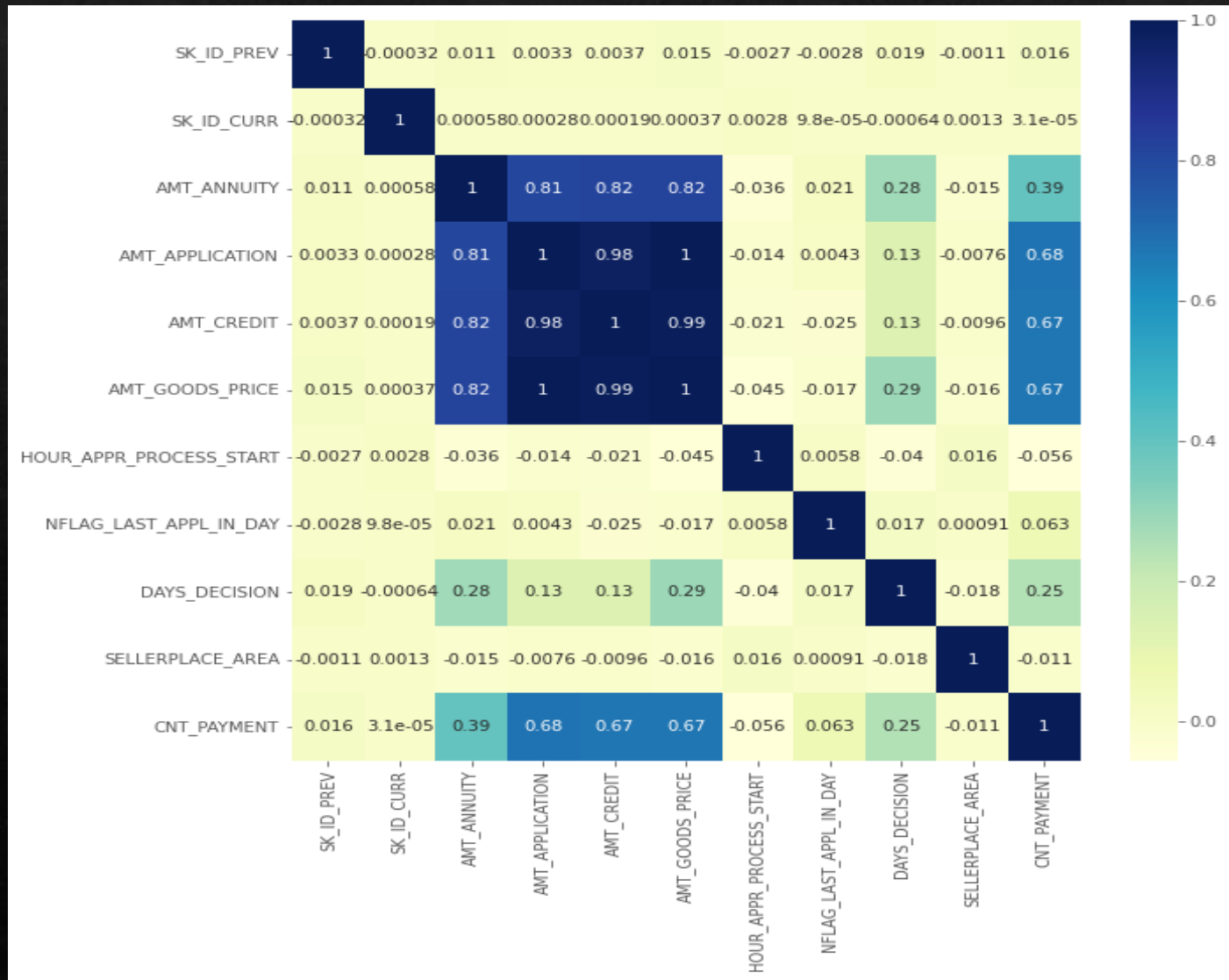
Bivariate Analysis (Previous Application Data)

- Banks take most amount of time for new clients to make a decision and in case of Repeater client banks take more no of days to give their approval while less in refusing and cancelling.



Heatmap of Correlation matrix (Previous Application Data)

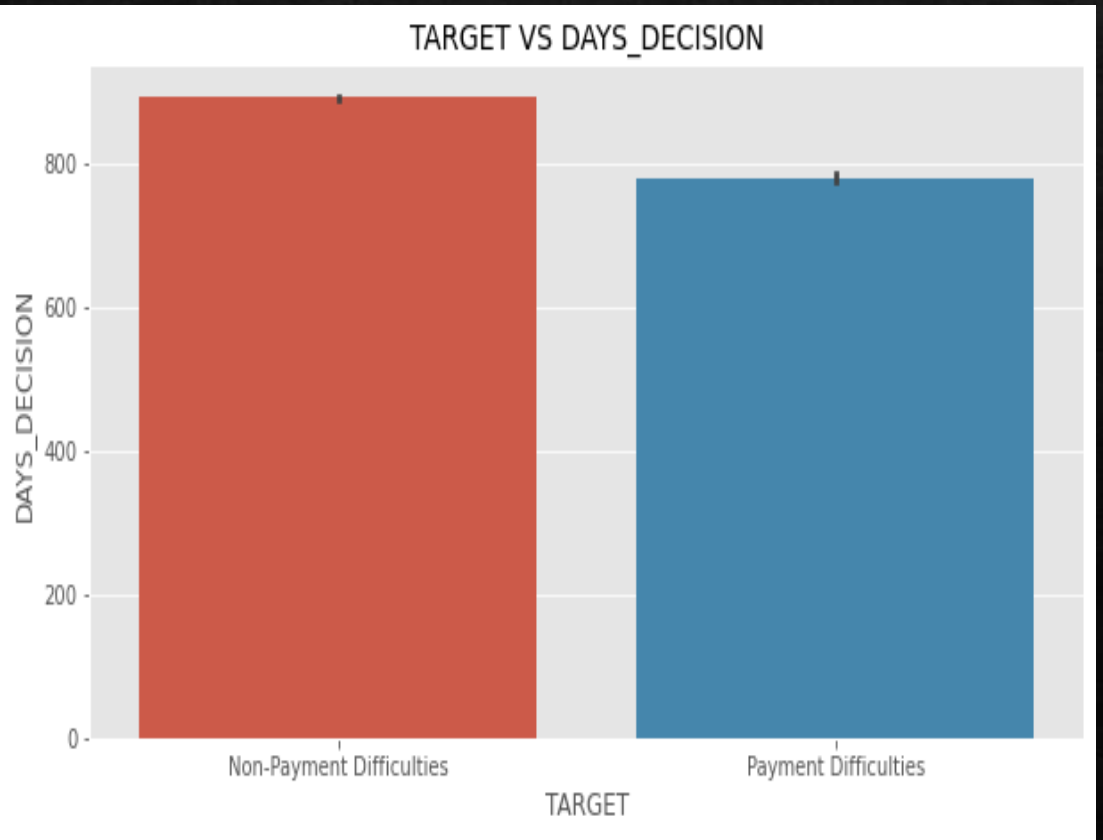
- AMT_GOODS_PRICE have a directly proportional relation with AMT_APPLICATION and AMT_CREDIT.
- Credit amount client asked on the previous application and Goods price of goods that clients mentioned on previous application.

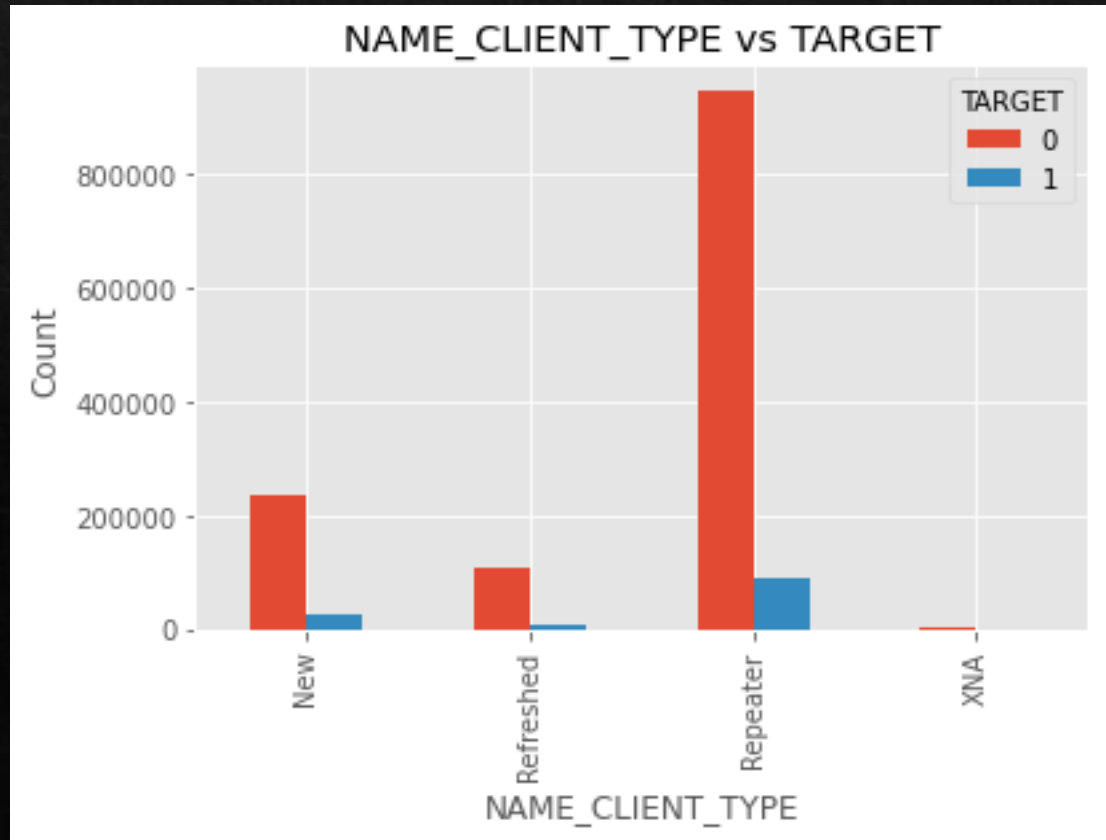


Merging of Data Frame

(Application Data and Previous Application Data)

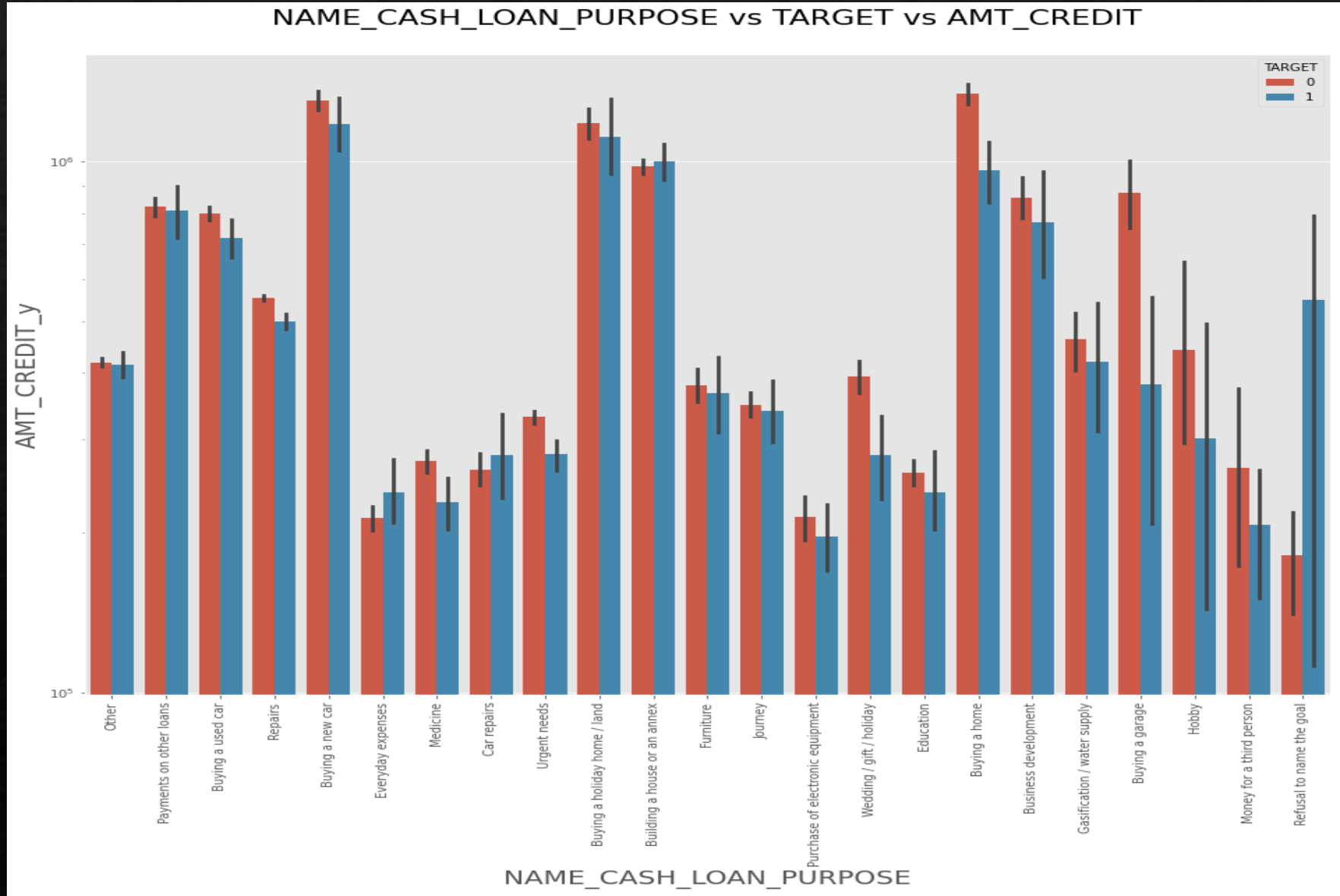
- Analysis of Merged Data :
  - Average time taken for approval of loan application is lesser for Defaulters.
  - Most of the defaulter are from the Repeater clients.





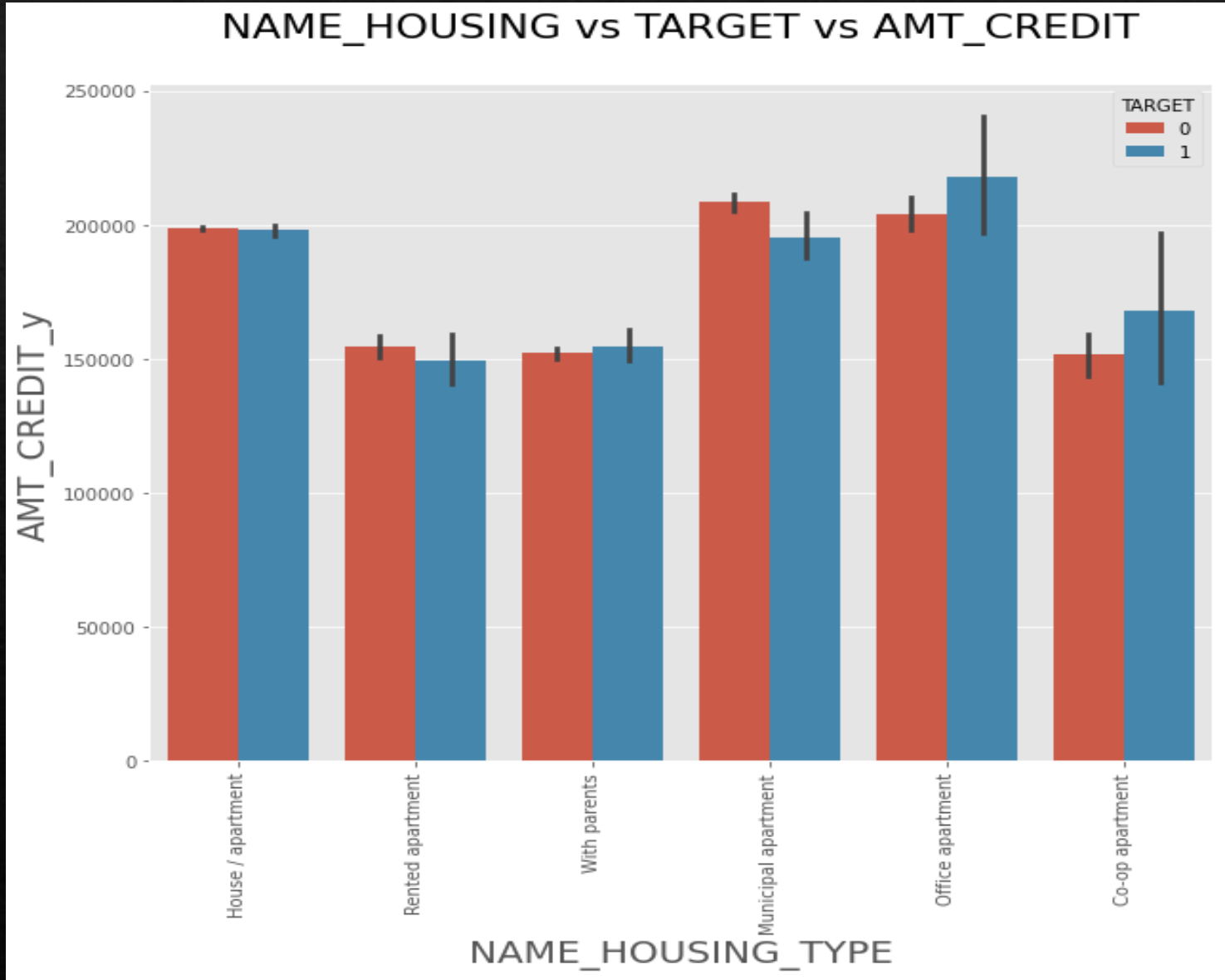
Analysis of Merged Data

- Client who refused to name the goal have high credit amount and also these are the one high defaulters, buying a garage and buying a home are goods categories to give loan as they are among non defaulters and have high loan amount.



Analysis of Merged Data

- Client who refused to name the goal have high credit amount and also these are the one high defaulters, buying a garage and buying a home are goods categories to give loan as they are among non defaulters and have high loan amount.



Conclusion

- Data is Highly Imbalanced 91.9% is Client with Non-Payment Difficulties and 8.1% is client with Payment Difficulties. Imbalance Ratio is 11.36%.
- Cash Loans are more popular than Revolving loans. Females have higher chance of applying for a loan
- Majority of clients who applied for loan are from Working class. Client with Secondary Education are more likely to apply for the loan.
- Count of married people applying for a loan is higher than the rest. Majority of the clients have House/Apartment.
- Laborers are more frequent customer of bank. Most of the client who applied for loan are Unaccompanied.
- (30-40) Age group people have a bit higher difficulty in paying loan while (60-70) age group people have lesser difficulty in paying their loans.
- People with very high salary have lesser difficulties in paying their loan but most number of people who go for loan are people with low salary.
- Female with the Secondary Education are the most defaulter.
- Client with married marital status and secondary education is at most risk of being a defaulter.
- Higher education client is least like to default on their loan.
- Consumer loans have the highest no of counts and also their approval rating is also the highest.
- Repeater clients are the one who have the highest frequency among the name_client_type.
- Bank takes large number of days to approve a loan in comparison with to cancelling or refusing a loan.
- Banks take most amount of time for new clients to make a decision.
- Average time taken for approval of loan application is lesser for Defaulters.
- Most of the defaulter are from the Repeater clients.
- Client with housing type as Office apartment and Co-op Apartment are the one with high Average Credit amount among defaulters category.

Recommendations

- Male laborers, female sales staff and male drivers have the highest payment difficulties and so the banks should put a capping on the loan amount.
- For Undisclosed goal category, banks should increase the interest rate for them.
- Non Defaulters have more cancellation and refusal as compared to the defaulters which ideally, should not be the case.
- Clients who have co-op apartment accommodation or office-apartment, Banks should put a capping on amount on their loan as they have high defaulters.
- Bank should target client with higher education as they have lesser number of non defaulters.
- Excise caution when giving loan to 30-40 age group people and carefully examine their loan application as they have large number of defaulters.