Bank Loan Case Study

View Slides

1. PROJECT DESCRIPTION
2. APPROACH
3. TECH-STACK USED
4. INSIGHTS
5. RESULT

By : Arvindh Kumar V

1) Project Description

Company receives loan applications and they have to decide to approve or not based on the applicant's profile.

2 types of risk:

1)Loss of Business

2)Financial Loss

So, I am provided with application_data, previous_application, data sets tables from which i must derive certain insights out of it and answer the questions. so it will be easy for me to handle it using **Jupyter Notebook and provide a detailed report**

2) Approach

Using columns_description I understood definitions of each column terms in application_data and previous_application.

Then loaded and read the dataset.

Using jupyter notebook I have inspected it and carried out data cleaning process.

Then handled null values, negative values, imputing values. This process helps for analysis purpose.

3) Tech-Stack Used

- I have used **Jupyter Notebook** web application.
- **Jupyter Notebook** provides an interactive computational environment. It produces documents (notebooks) that combine both inputs (code) and outputs into a single file
- It can analyze data, calculate statistics, and represent data as **charts or graphs using python code with certain libraries (like matplotlib, seaborn etc)**

4) Insights

First load dataset in juypter notebook

Present the overall approach of the **analysis**. Mention the problem statement and the analysis approach briefly

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Identification of applicants who are capable of repaying the loan using EDA is the aim of this case study

Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)
Hint: *Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.*

Percentage of null values for each column is calculated and columns are drop based on your requirement of percentage of null values >50% .

I have replaced some missing values (except in OCCUPATION_TYPE) with mode value for numeric columns and median value for continuous numeric columns and Days columns contained negative values so I have replaced with positive value using abs() function.

While observing flag own car, flag own realty it contain y and n value so I have replaced them with 1 and 0 with help of where condition.

In CODE_GENDER column we have 4 XNA (means null value) so I have imputed them with F because count of F is more.

In ORGANIZATION_TYPE column we have XNA and OCCUPATION_TYPE has null value so I have imputed with Pensioner because when we compare ORGANIZATION_TYPE and NAME_INCOME_TYPE for most of the XNA value we see NAME_INCOME_TYPE has Pensioner and almost count of XNA and Pensioner is almost equal.

I have applied qcut() function on AMT_INCOME_TOTAL and AMT_CREDIT with q=[0,0.2,0.4,0.6,0.8,1] (quartile) 5 categories very low, low, medium, high, very high and imputed these values in new column AMT_INCOME_TYPE and AMT_CREDIT_TYPE.

DAYS_BIRTH has days so I have converted values to years then I have applied cut() function on DAYS_BIRTH with bins=[19,25,35,60,100] 4 categories very young, young, middle age, senior citizen and imputed these values in column AGE_GROUP.

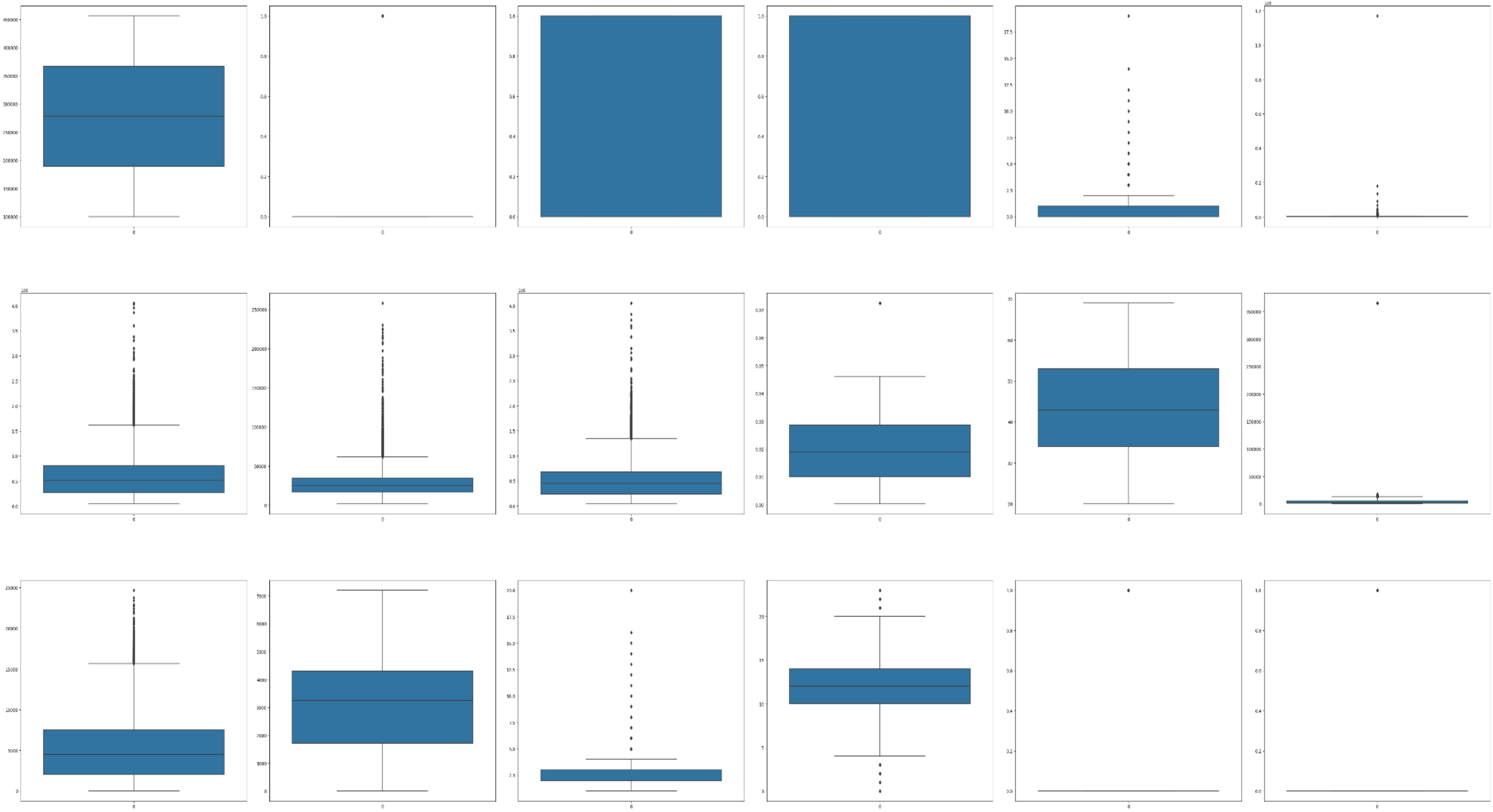
After checking datatypes of columns and I have changed object to category type.

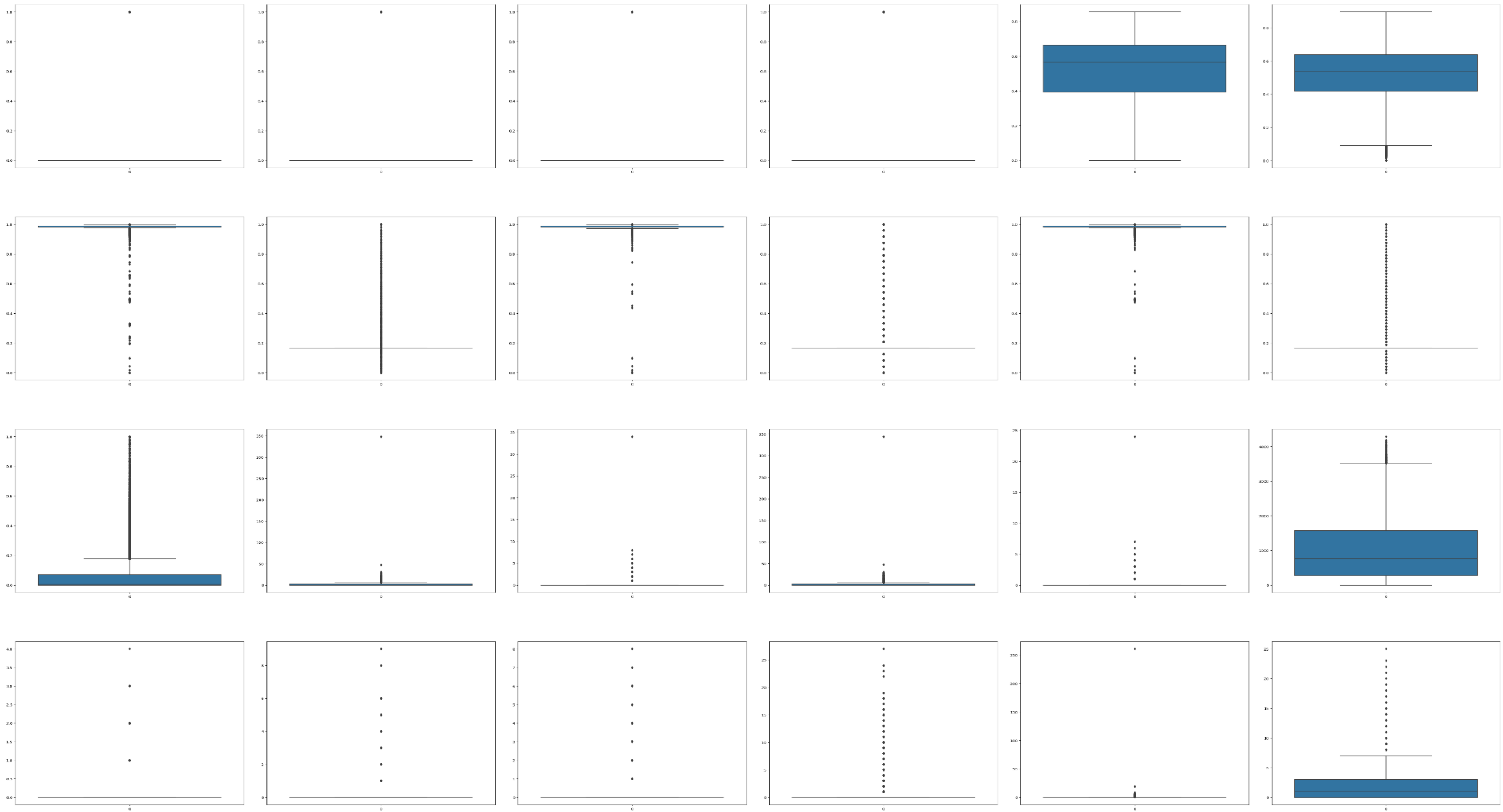
After gothroughing the dataframe I have drop some columns which are not necessary for analysis.

Identify if there are **outliers** in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.

I have used Boxplot on each column which has numerical values to determine outliers. As mentioned in question outliers are not removed.

After noticing Boxplot of each columns I have found out few insights.





Insights

SK_ID_CURR, DAYS_BIRTH, DAYS_ID_PUBLISH and EXT_SOURCE_2, EXT_SOURCE_3 don't have any outliers.

CNT_CHILDREN have outlier values having children more than 2.5

FLAG_OWN_CAR have no First and Third quantile and values lies within IQR, So most of the clients own a car

FLAG_OWN_REALTY have no First and Third quantile and values lies within IQR, So most of the clients own a House/Flat

DAYS_EMPLOYED, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, 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 and AMT_REQ_CREDIT_BUREAU_YEAR has very slim Boxplot and have a large number of outliers.

Identify if there is data imbalance in the data. Find the ratio of data imbalance.
Hint: *Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights*

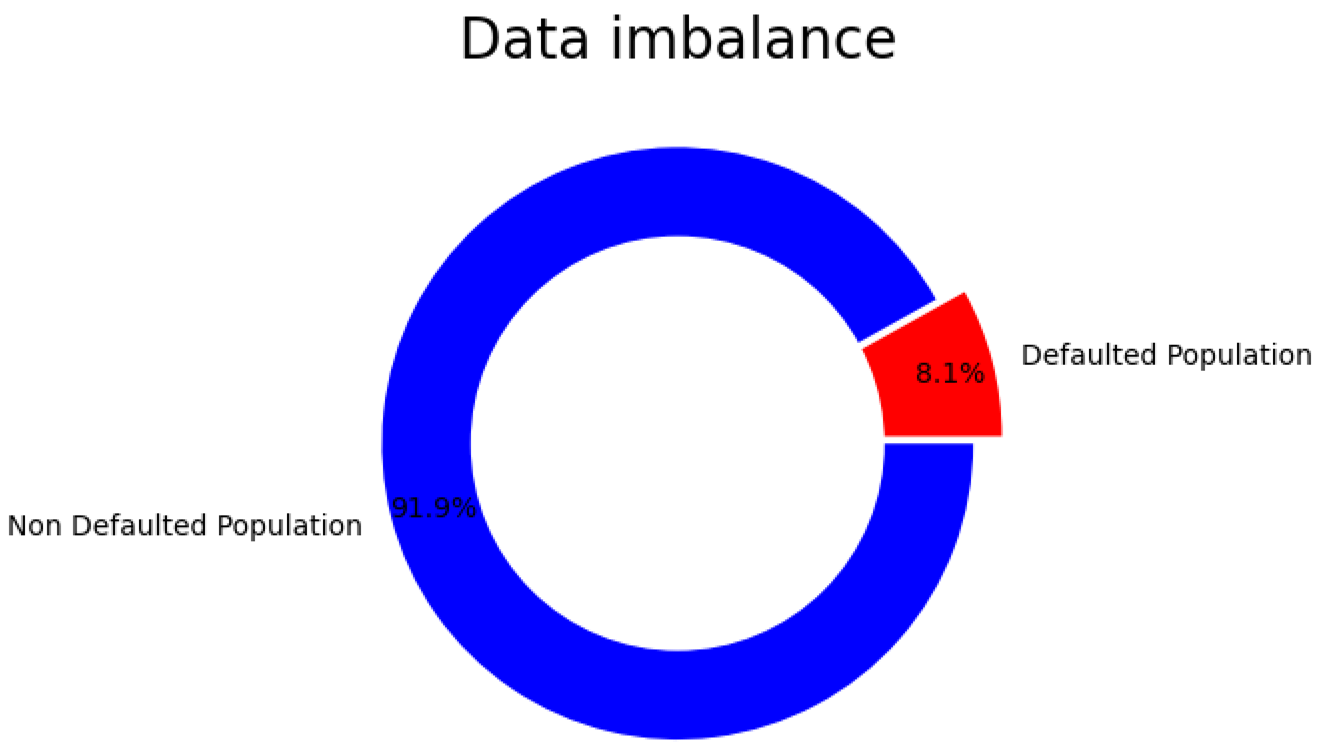
I have determined data imbalance using TARGET column.

Made new dataframe t0, t1 where TARGET VALUE is 0 and 1 respectively.

using len() function I determined length of both t0, t1 and divided them to find Imbalance ratio.

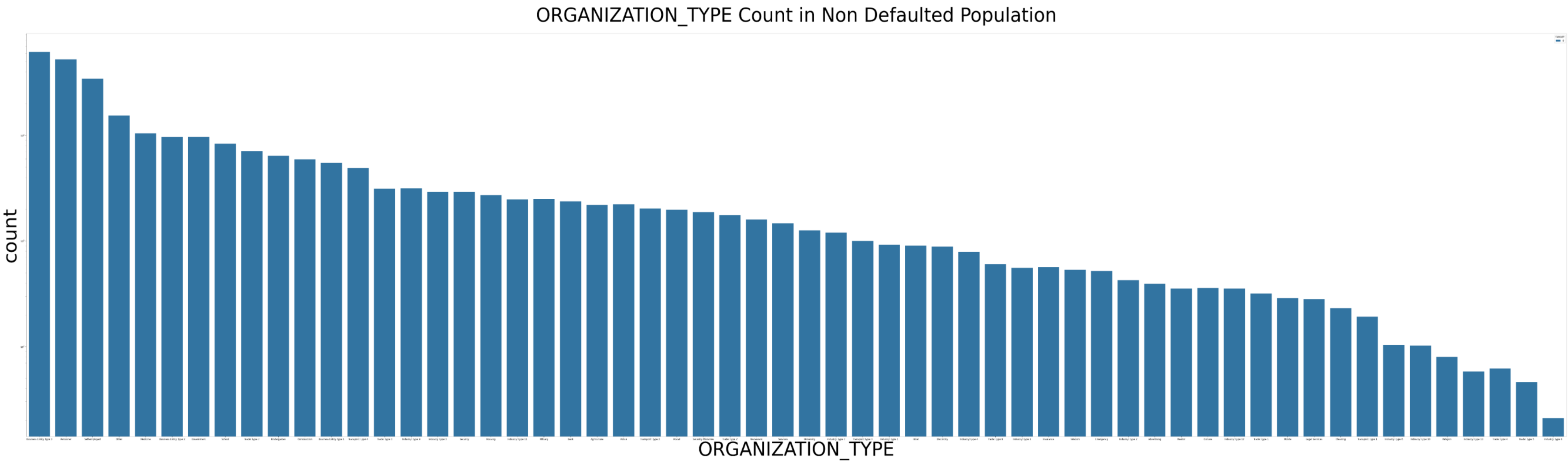
I have counted the no.of values in t0 and t1 and plotted a donut chart.

```
Count of Non Defaulted Population(0) : 91.92711805431351
Count of Defaulted Population(1) : 8.072881945686495
```

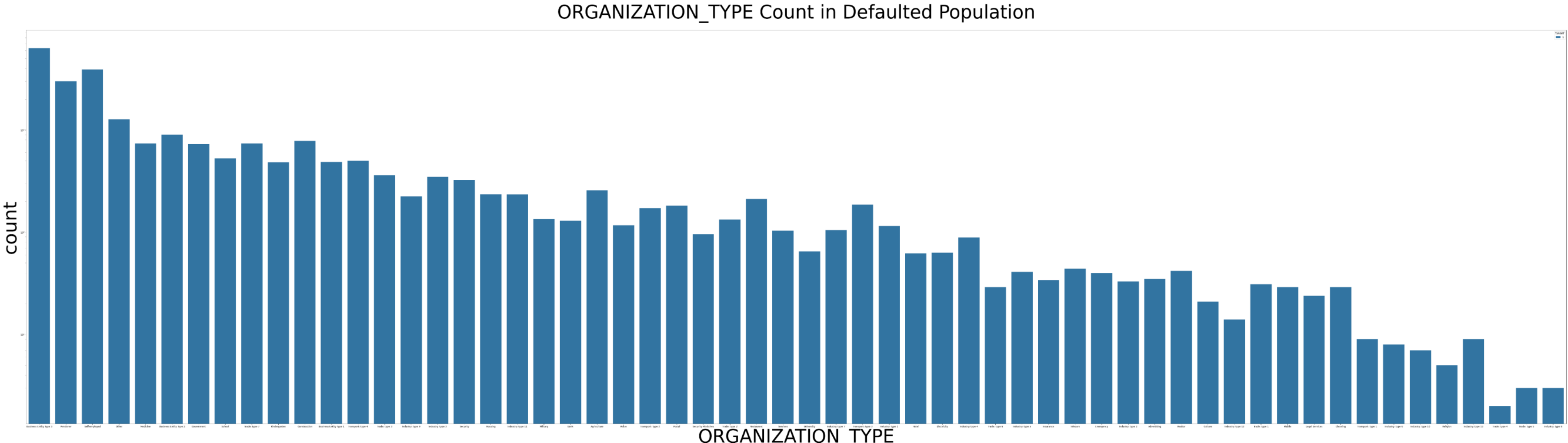


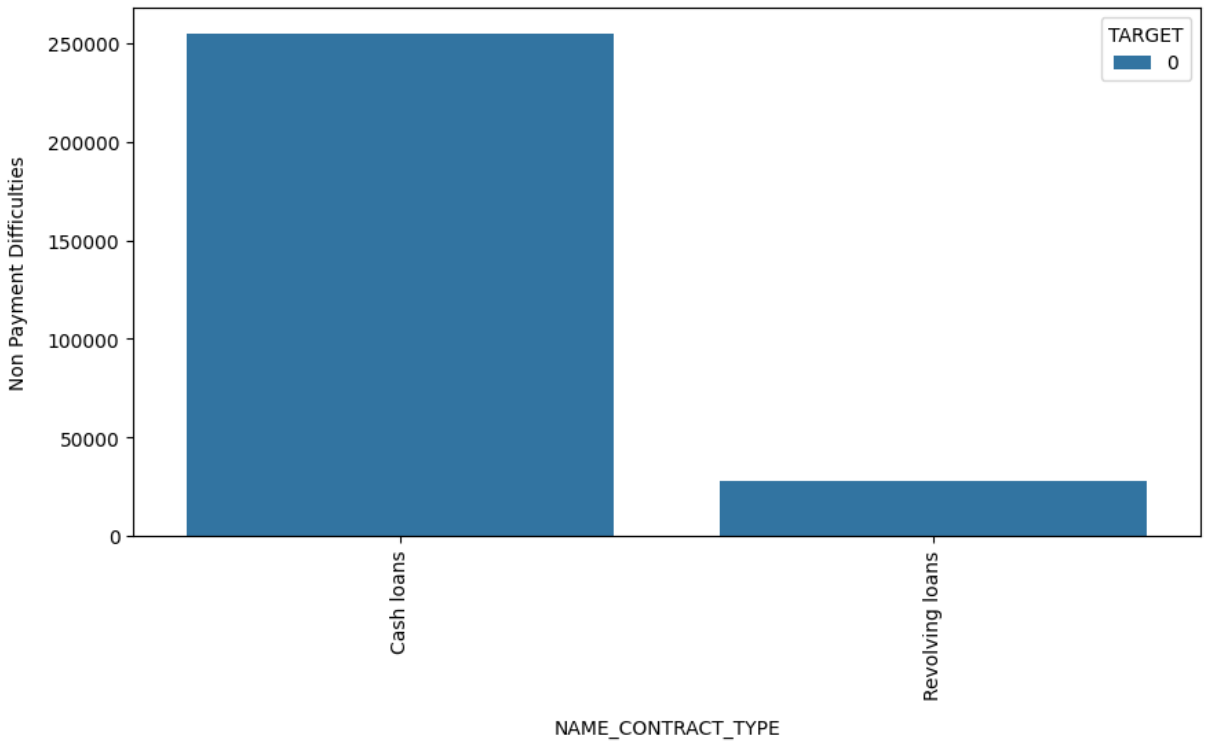
Include visualizations and **summarize** the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.

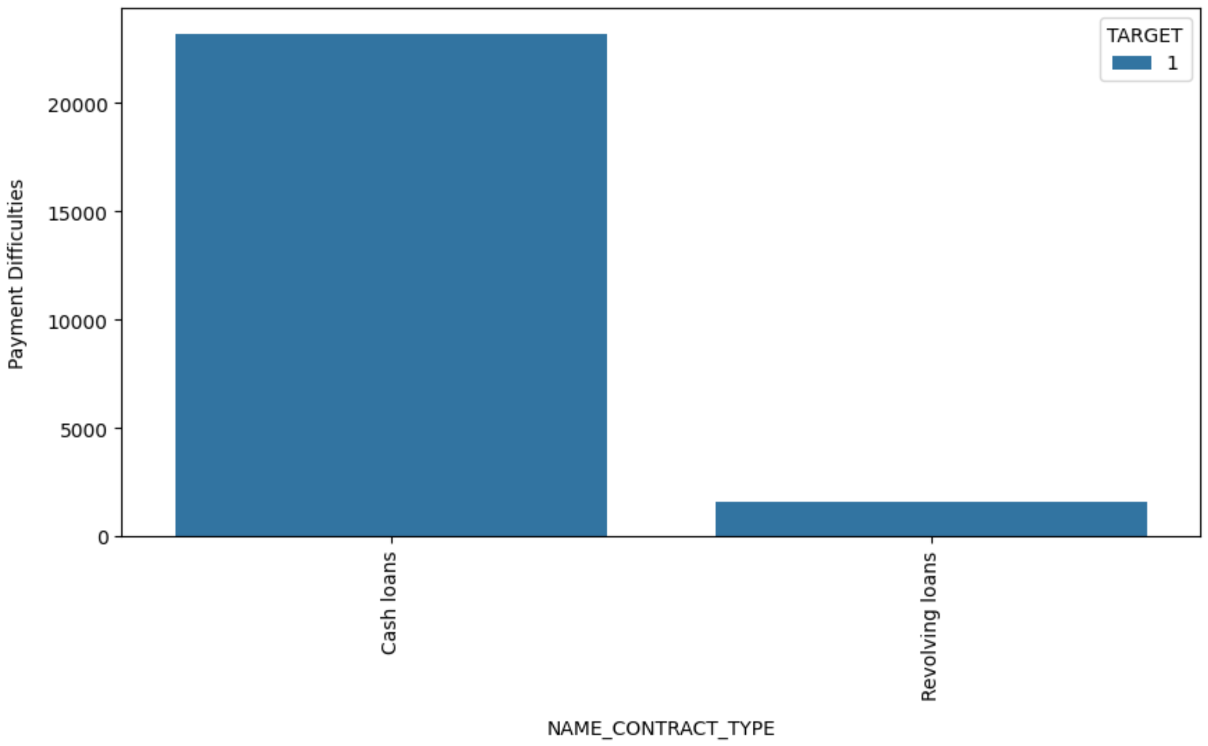
Non Defaulted consumers of ORGANIZATION_TYPE Business Entity Type 3, Pensioner, Self-employed, Other, Medicine, Business Entity Type 2, Government has applied for loan the most and Industry: type 8, Trade: type 5, Trade: type 4, Industry: type 13, Religion, Industry: type 10, Industry: type 6 has applied for loan the least.



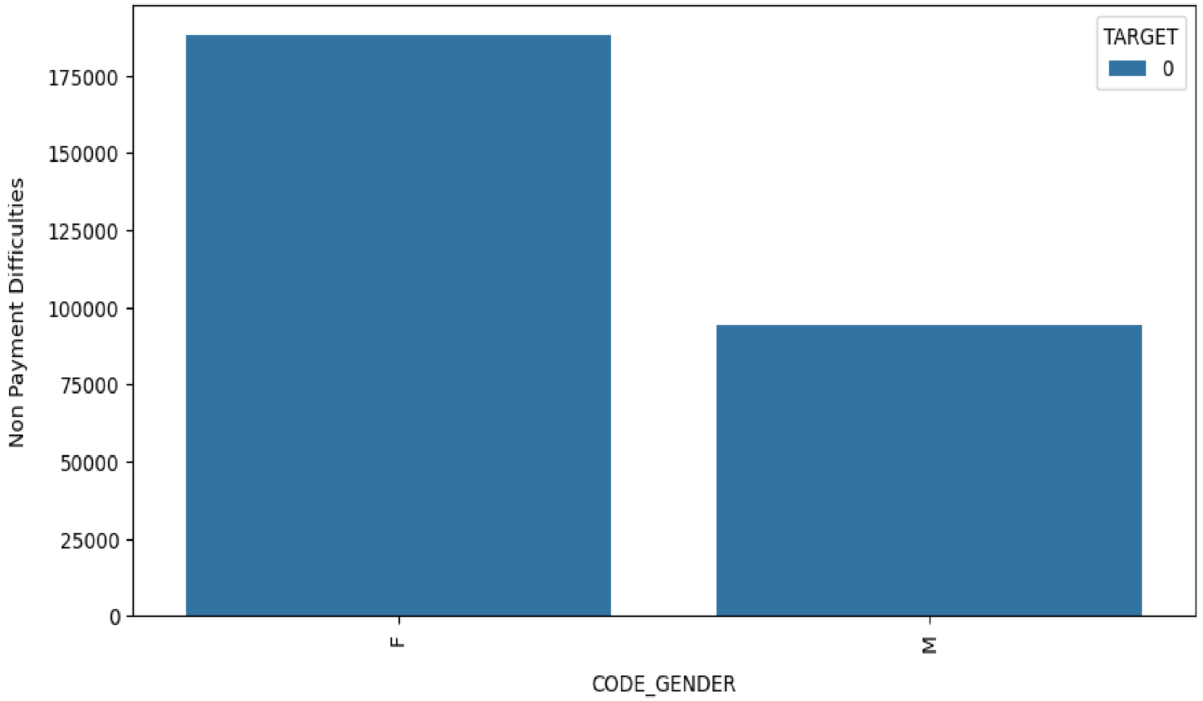
Defaulted consumers of ORGANIZATION_TYPE Business Entity Type 3, Self-employed , Pensioner, Other, Business Entity Type 2, Medicine, Government has applied for loan the most and Trade: type 4, Industry: type 8, Trade: type 5, Religion, Industry: type 10, Industry: type 6, Transport: type 1 has applied for loan the least.

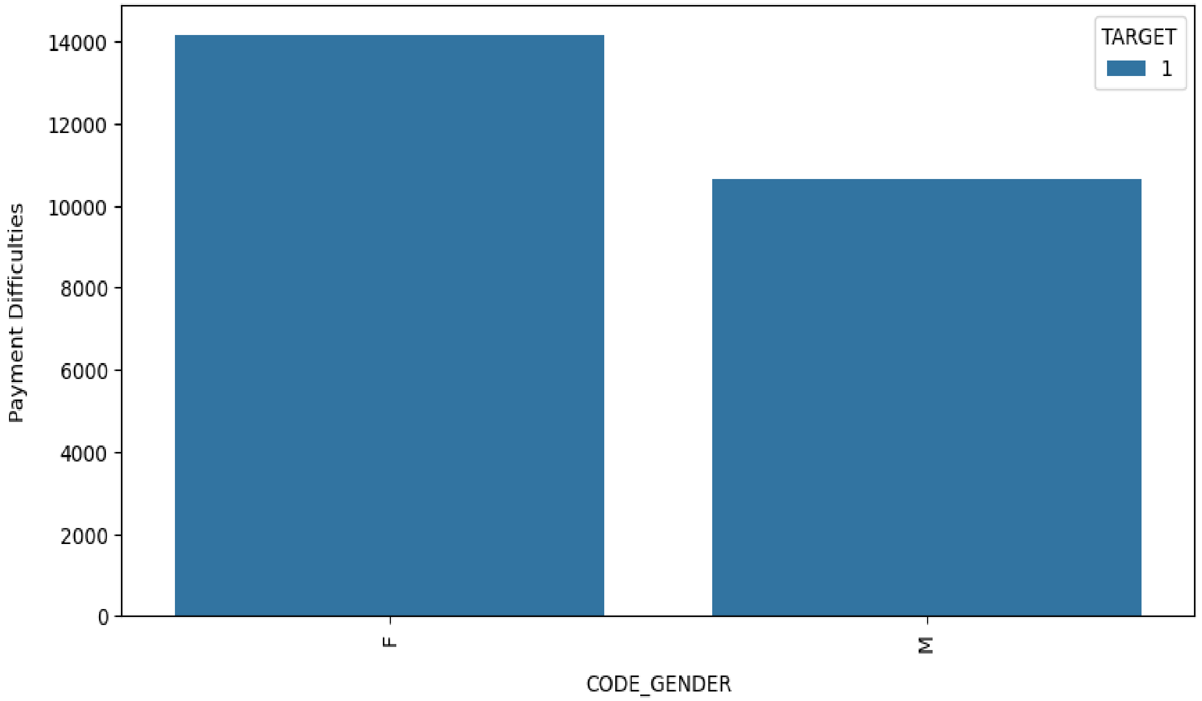




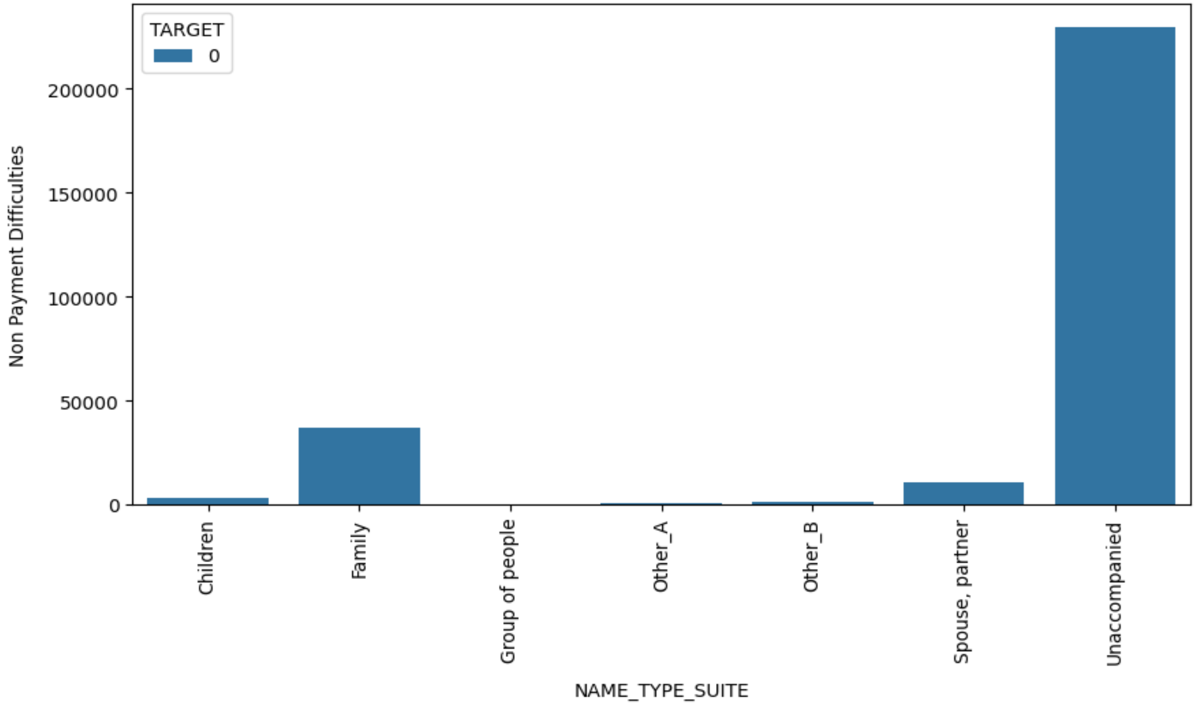


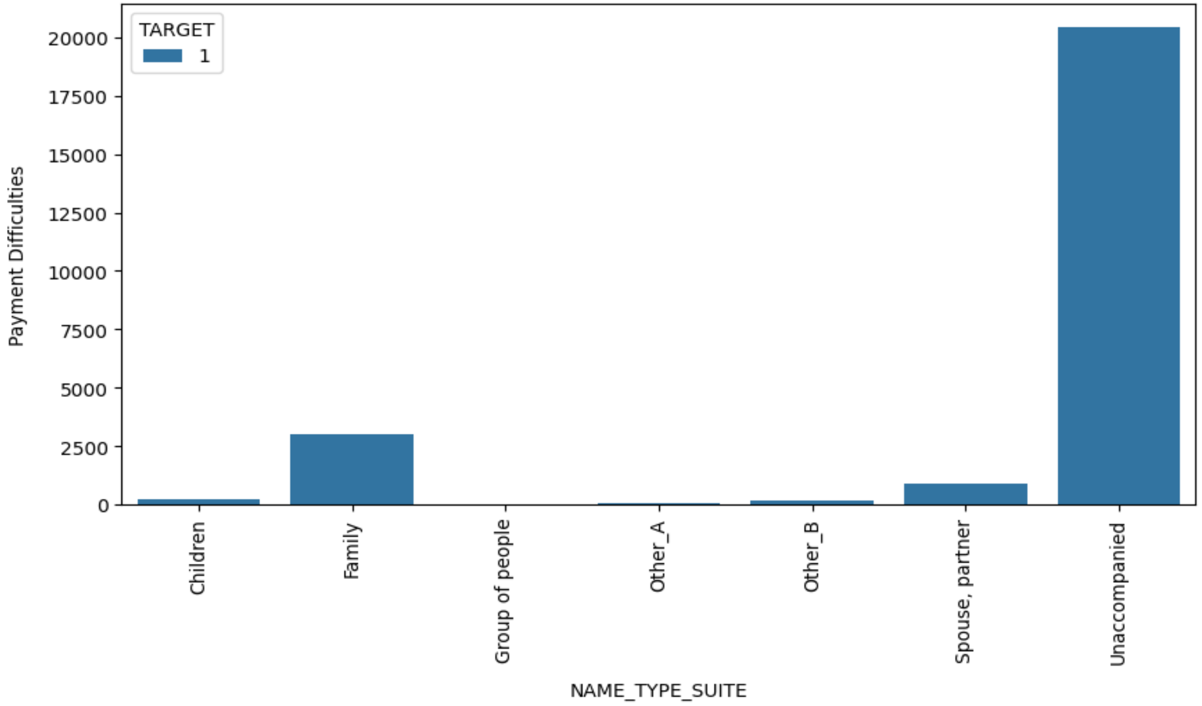
Most of Defaulter and Non defaulter clients has applied for Cash loans and only few have applied for Revolving loan



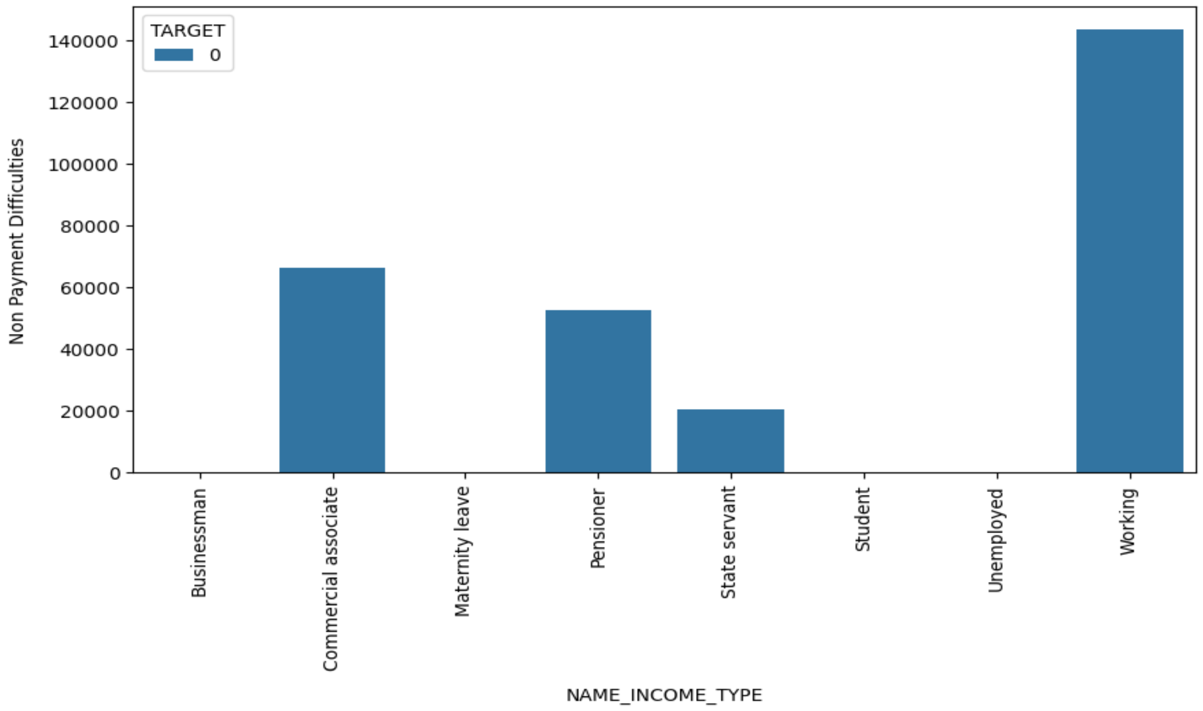


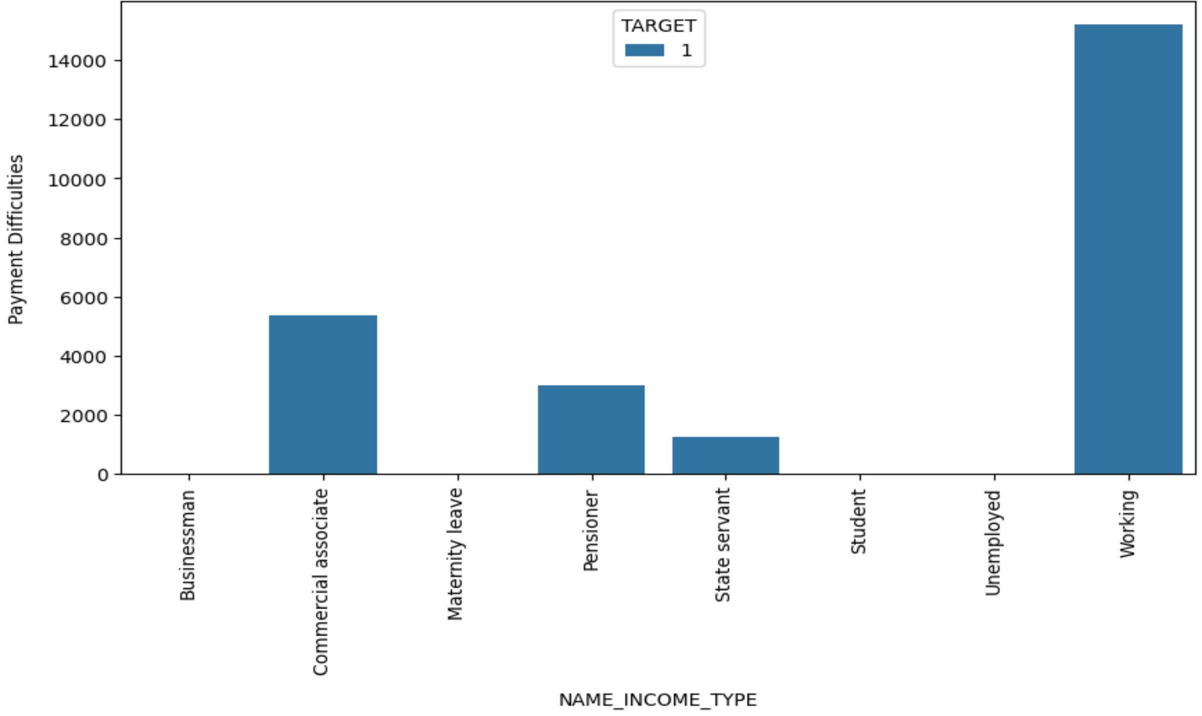
Most of Defaulter and Non defaulter clients has applied for loans are Female more than Male.



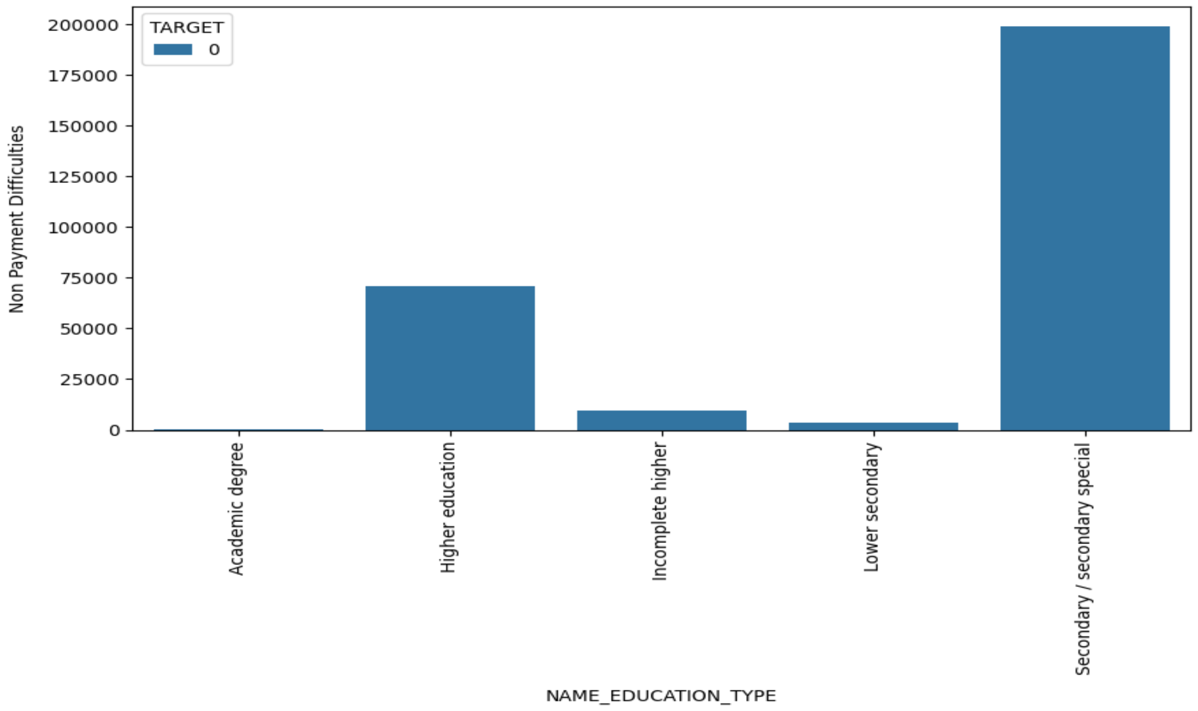


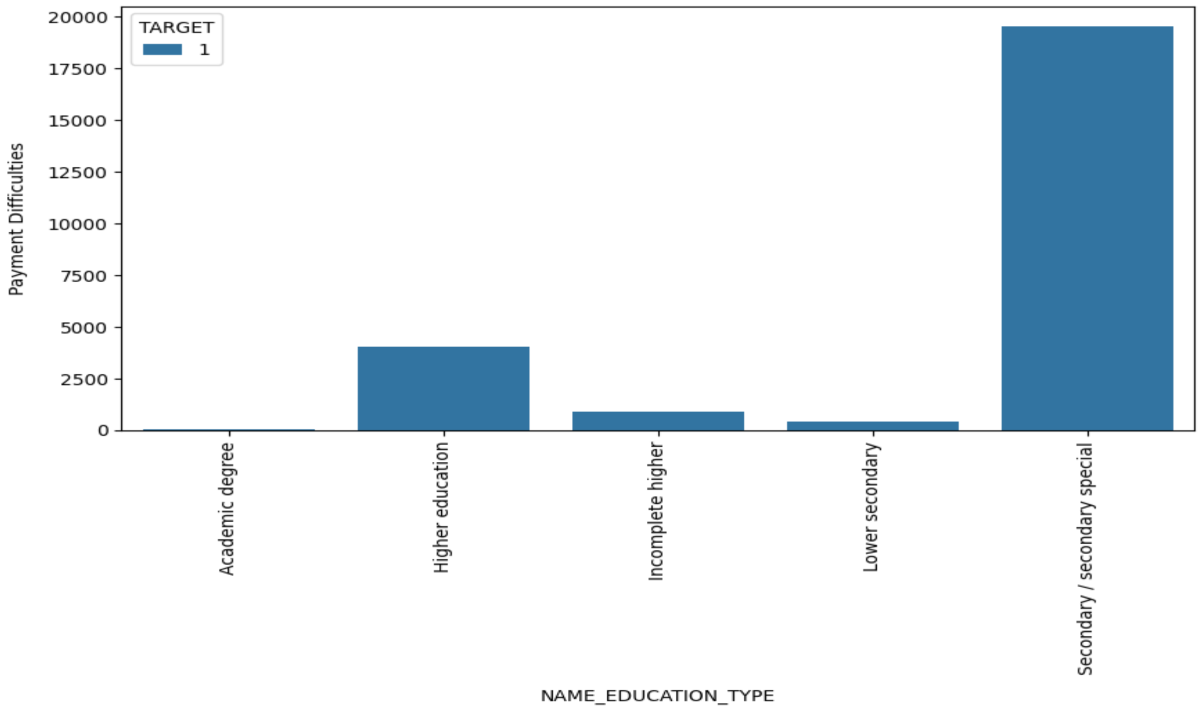
Most of Defaulter and Non defaulter clients were unaccompanied while when applying for loans and only few have clients were accompanied by family.



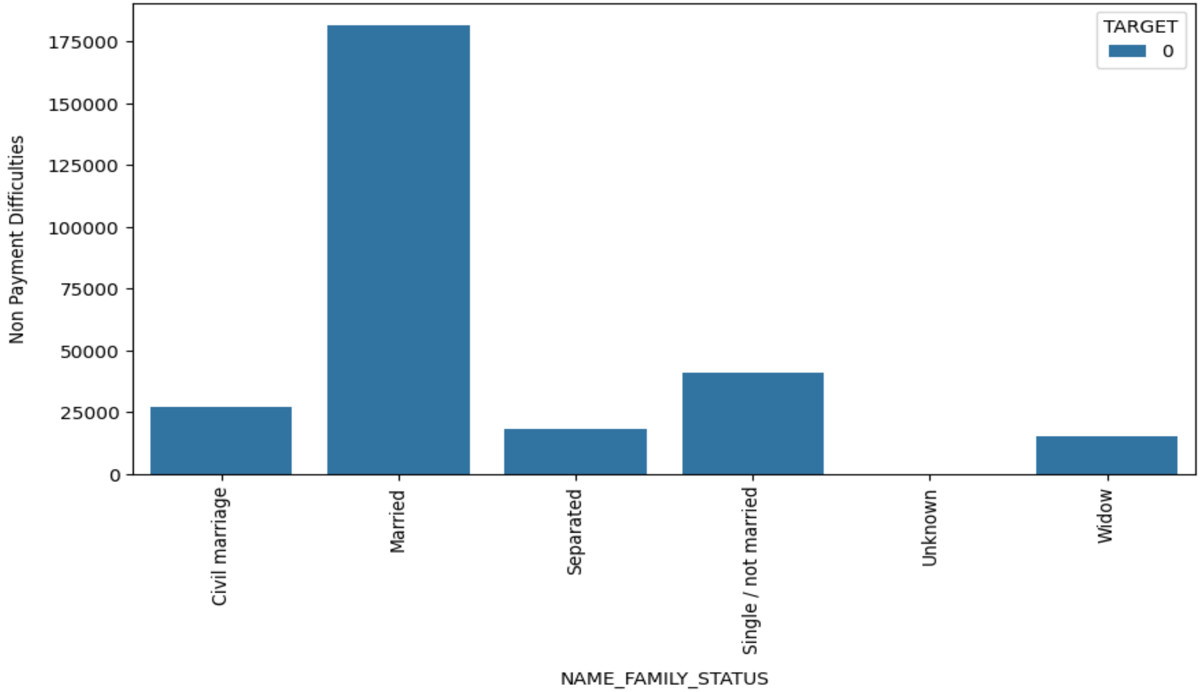


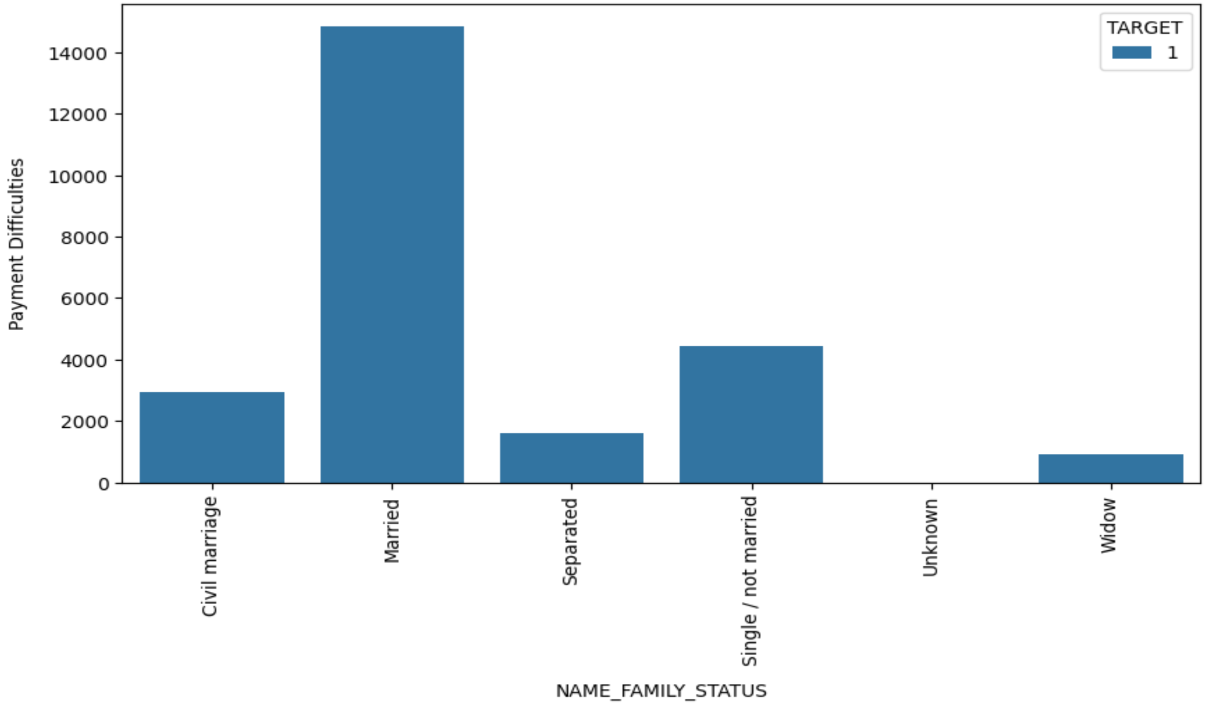
1. Most of Defaulter and Non defaulter clients who have applied for loans were Working, Commercial associate, Pensioner and State Servant.
2. Working have high risk
3. State Servant has Minimal risk



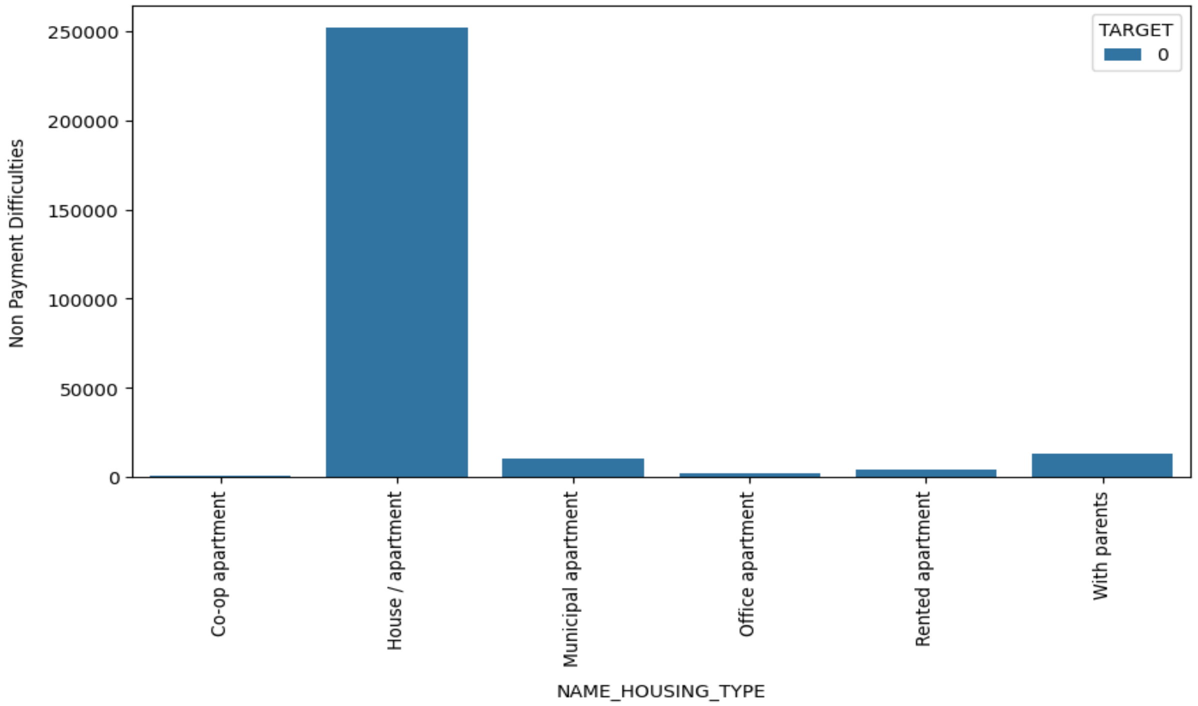


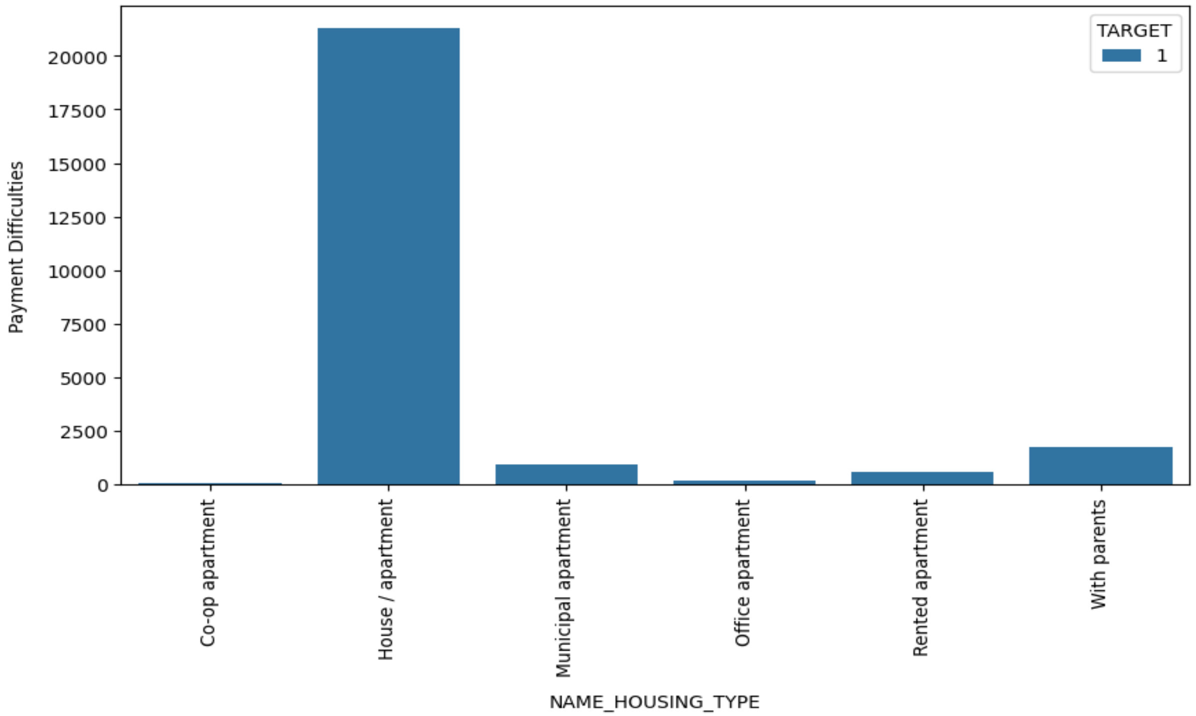
1. Most of Defaulter and Non defaulter clients has applied for loans have Secondary or Secondary Special education and next highest is Higher education.
2. Secondary or Secondary Special education have high risk.
3. Others have low risk.



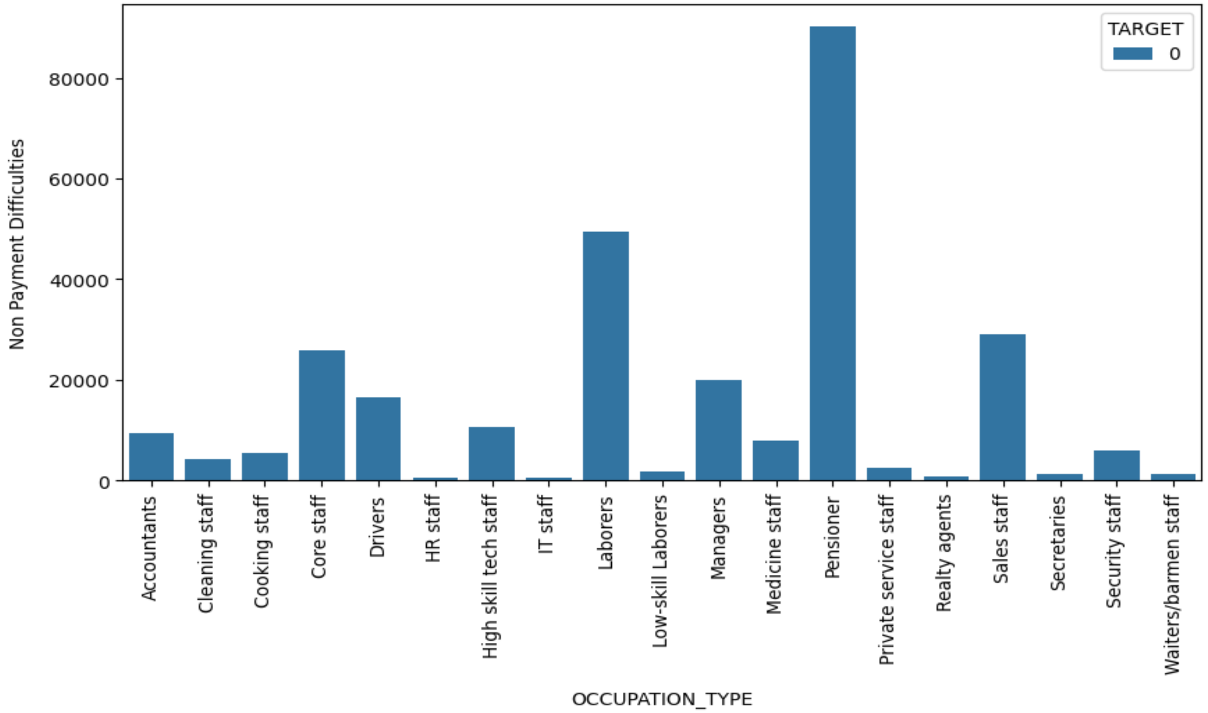


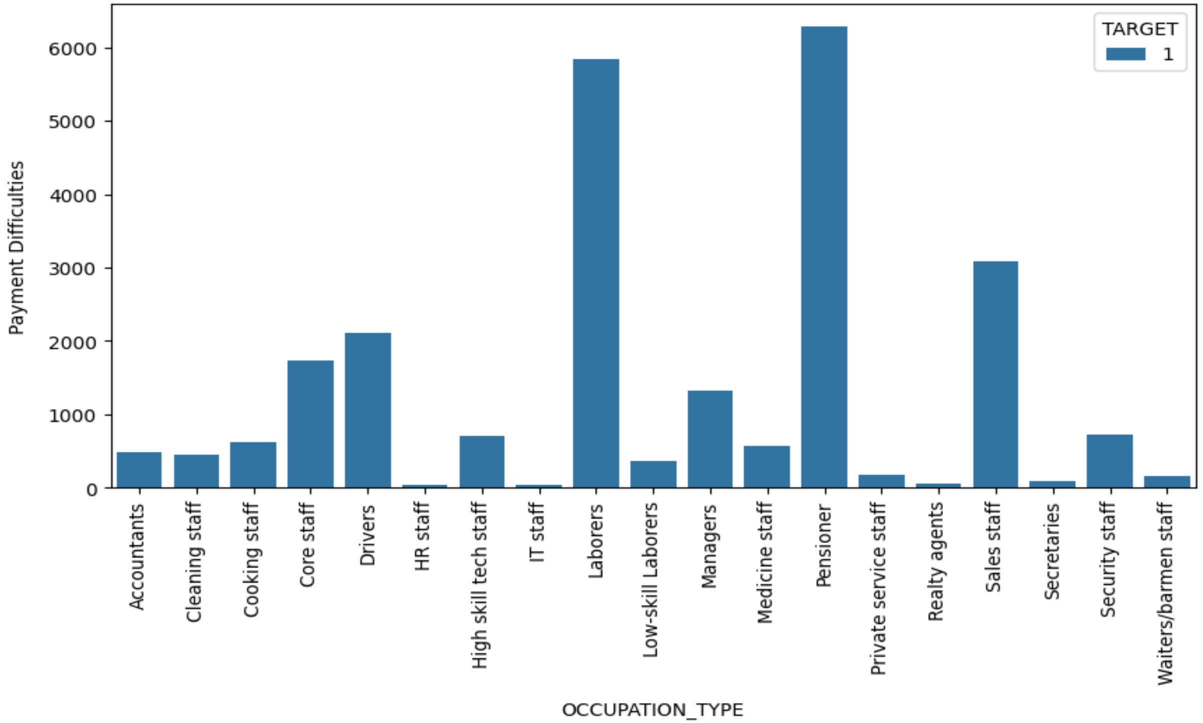
1. Most of Defaulter and Non defaulter clients has applied for loans are Married.(high risk)
2. Widows (in both Defaulter and Non defaulter) has less count for applying loan.(less risk)



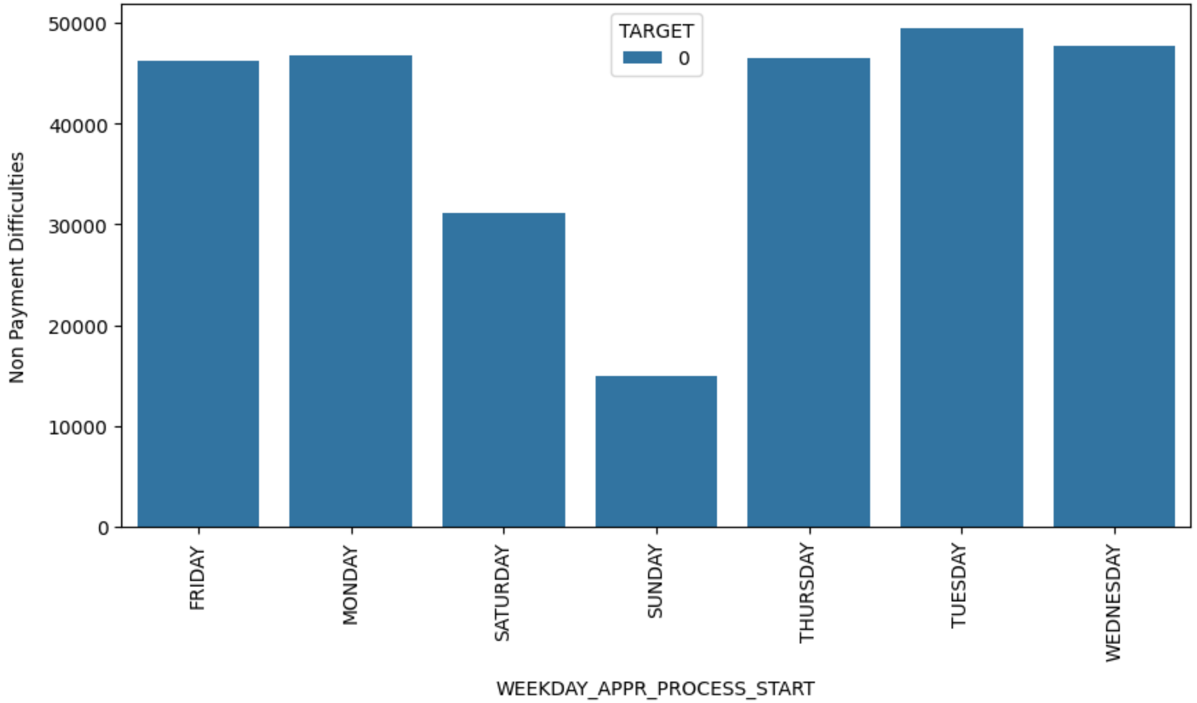


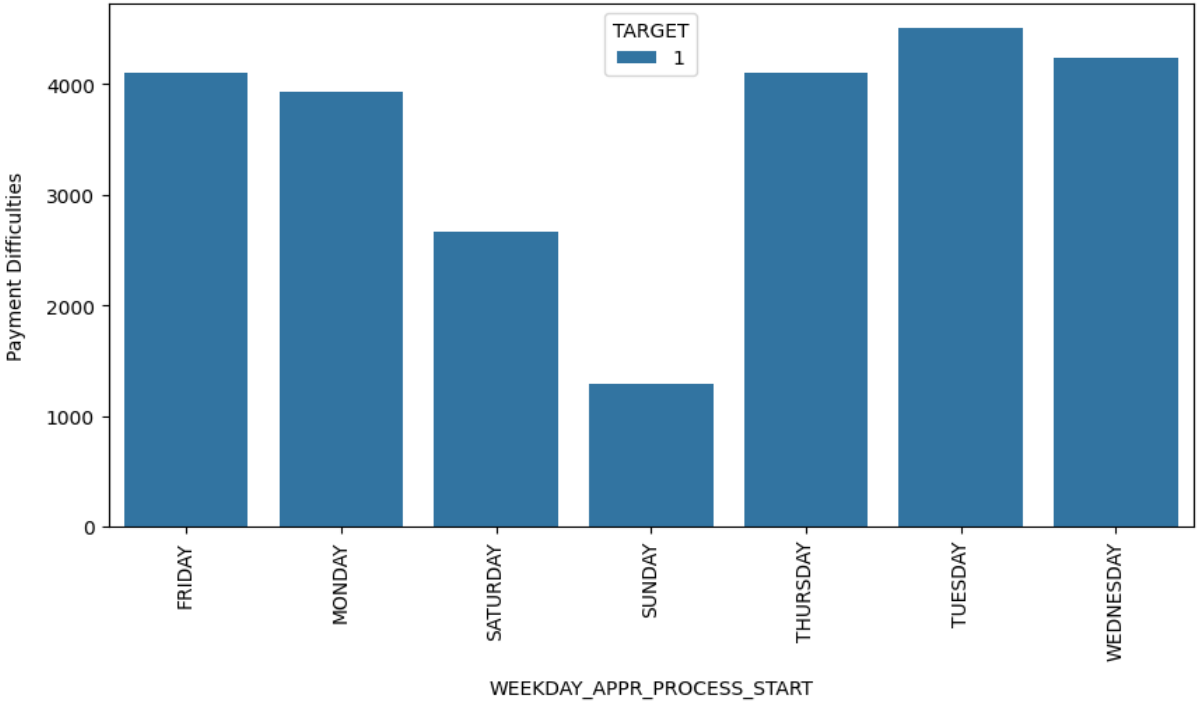
1. Most of Defaulter and Non defaulter clients has applied for loans are living in Home / apartment.
2. Approval of loan from these clients have high risk.



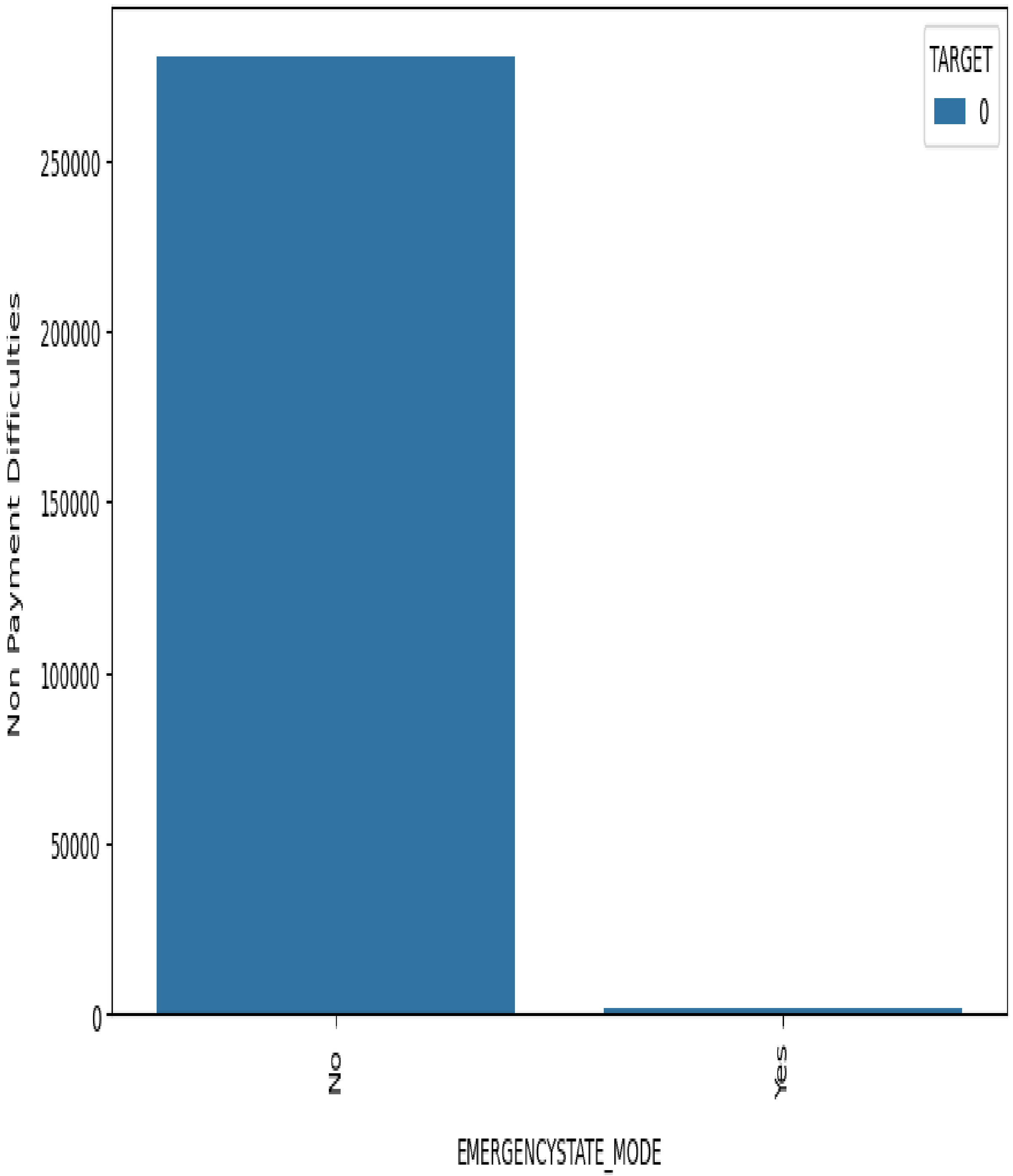


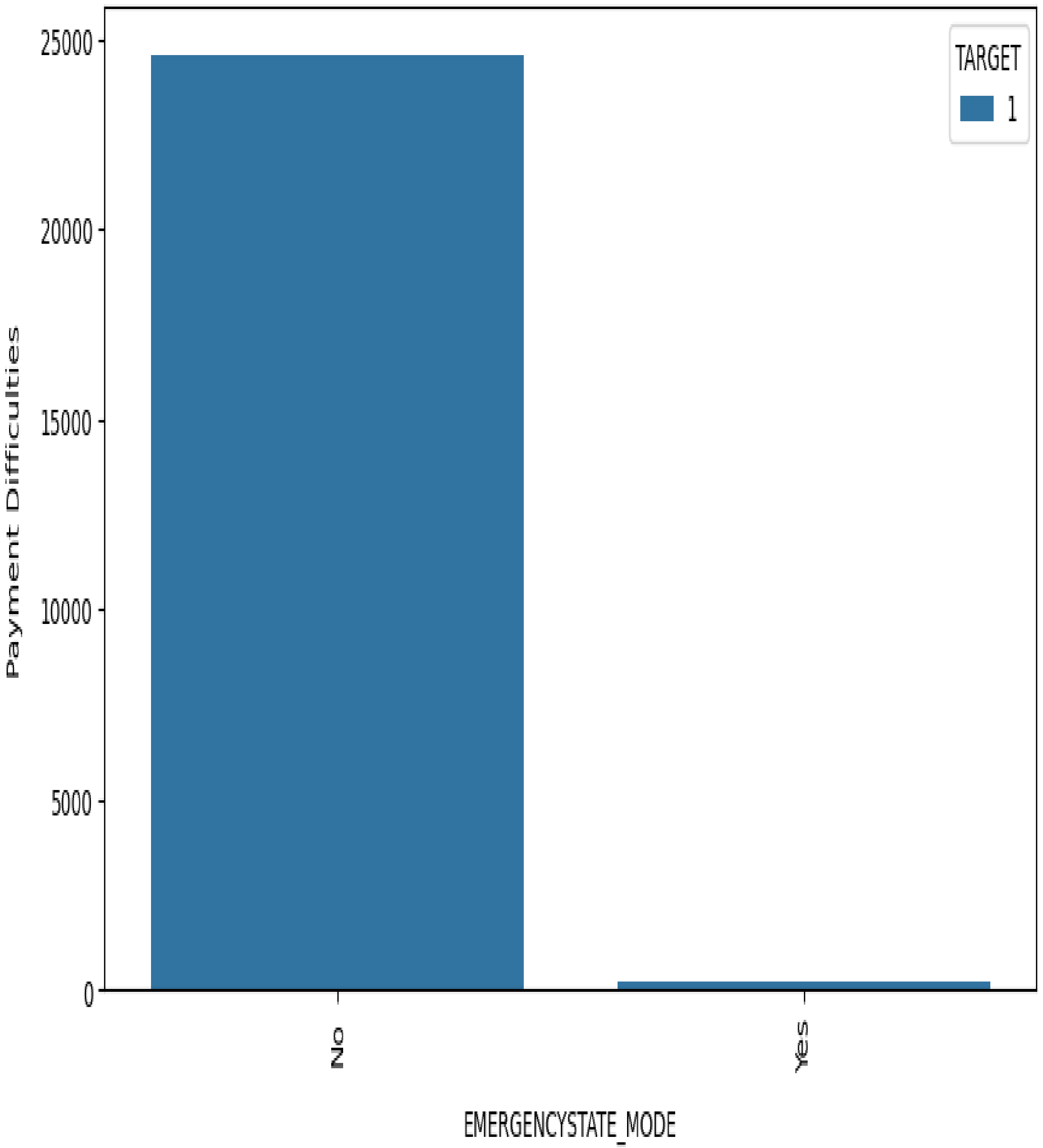
1. Most of Defaulter and Non defaulter clients has applied for loans are Pensioners, Laborers and Sales Staff.
2. Approval of loan for Pensioners and Laborers have high risk.

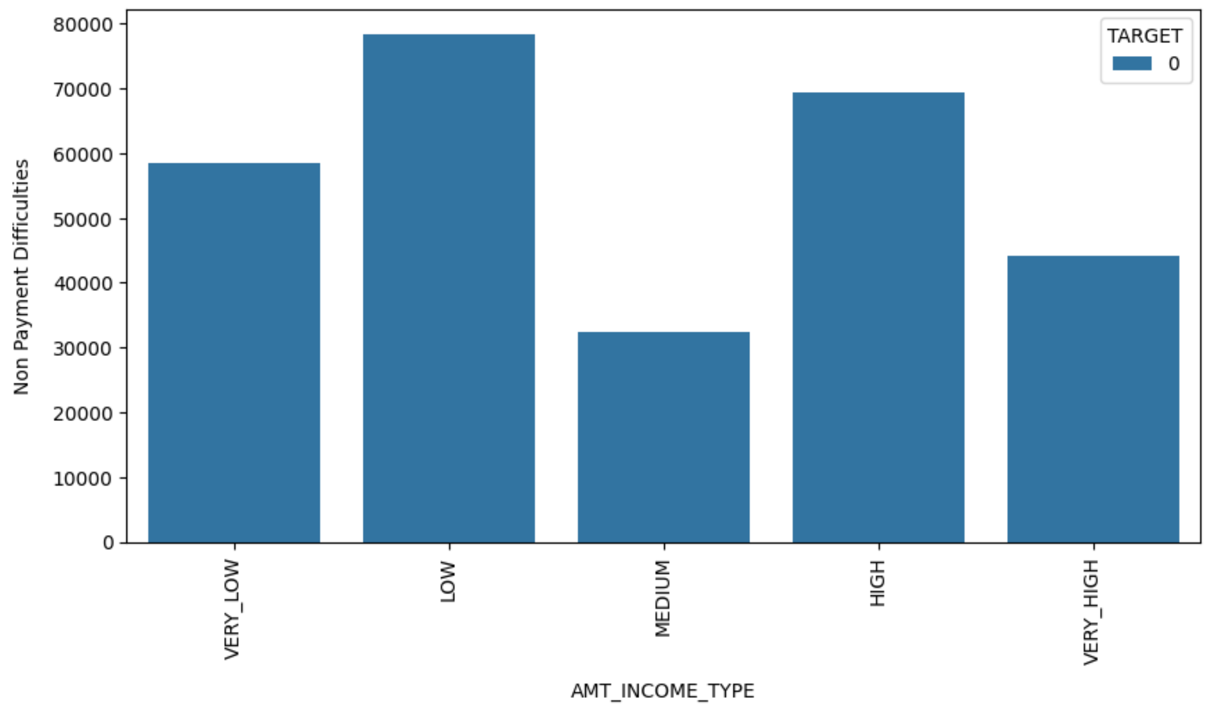


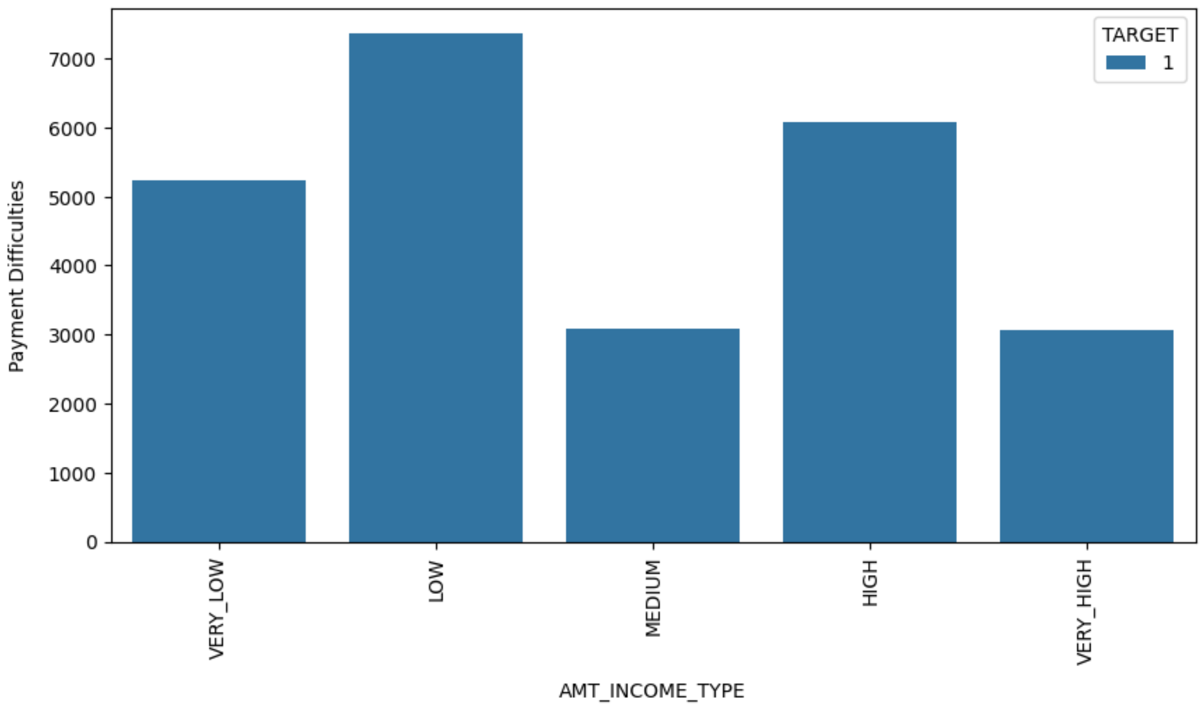


Most of Defaulter and Non defaulter clients has applied for loans on TUESDAY.

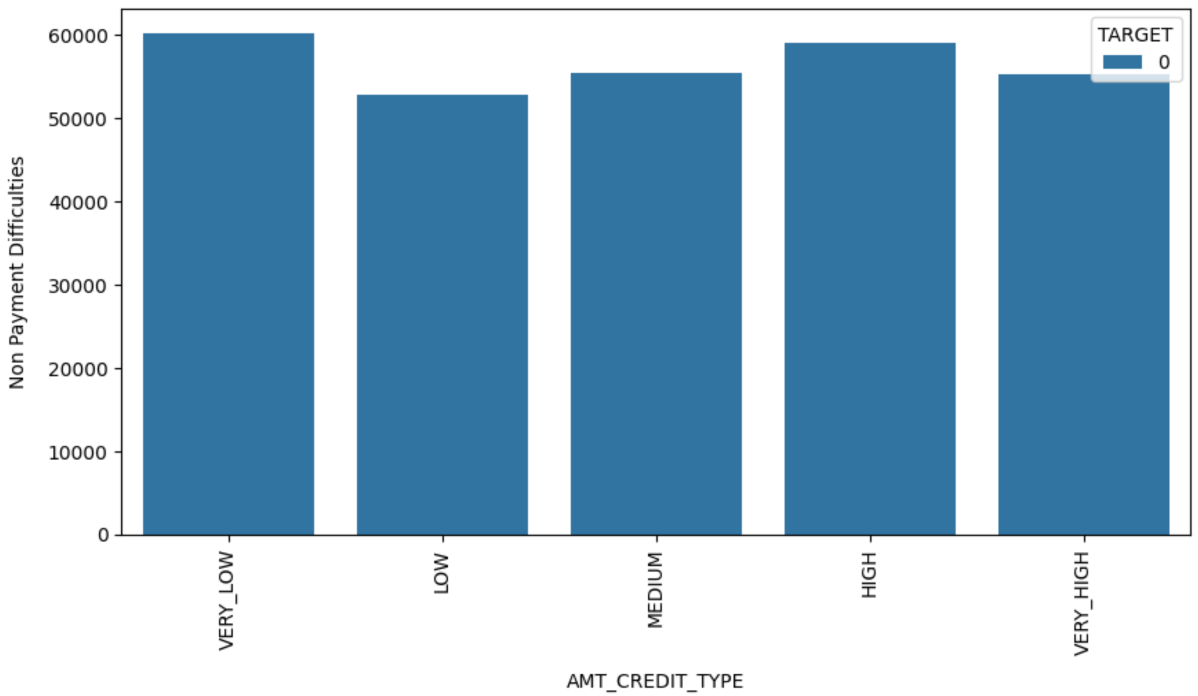

TARGET
0
Non Payment Difficulties
250000
200000
150000
100000
50000
0
No
Yes
EMERGENCYSTATE_MODE

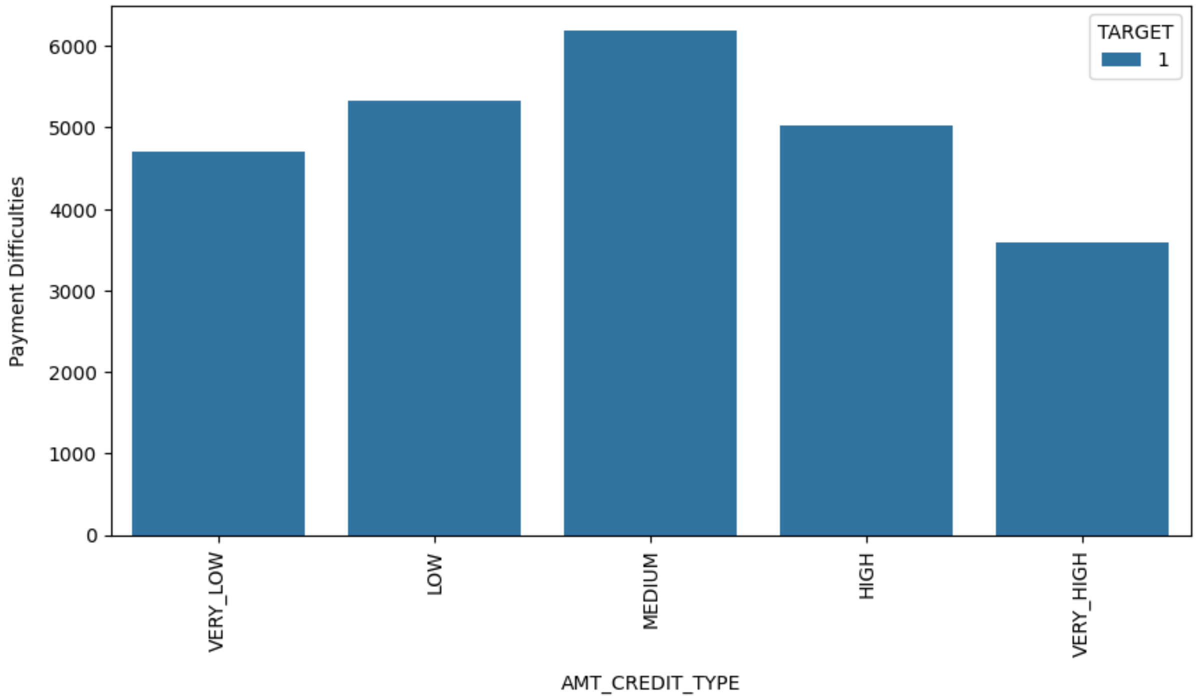

TARGET
1
Payment Difficulties
25000
20000
15000
10000
5000
0
No
Yes
EMERGENCYSTATE_MODE





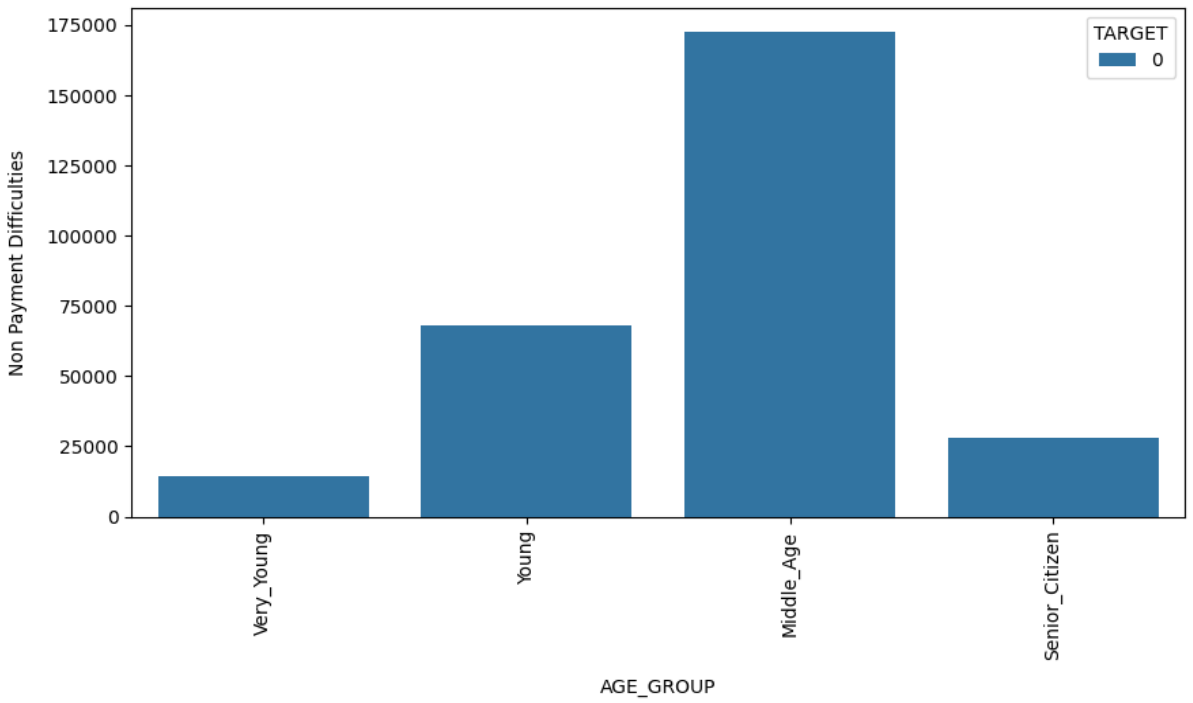
1. Most of Defaulter and Non defaulter clients has applied for loan are Low and High Salary range.
2. Approval of loans for these clients low and high may cause high risk.

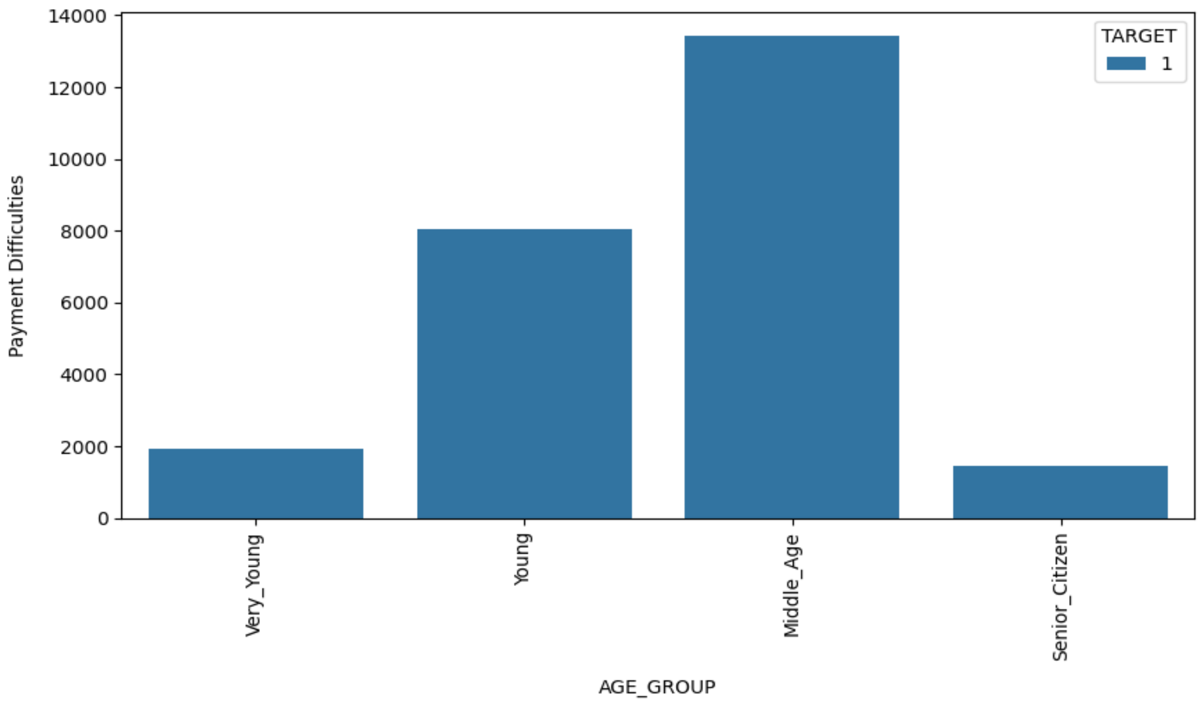




Most of Non Defaulter clients has applied for Very_LOW and HIGH Credit amount of loan.

Most of Defaulter clients has applied for Very_LOW and HIGH Credit amount of loan.





1. Most of Defaulter and Non defaulter clients has applied for loans are Middle_age (35 to 60) and young (25 to 35).
2. Approval of loan for these clients may result in high result
3. Very_young and Senior citizen has less paying defficulities so less risk.

Explain the **results of univariate, segmented univariate, bivariate analysis, etc.** in business terms.

Univariate analysis

It is the technique of comparing and analyzing the dependency of a single predictor and a response variable.

Segmented Univariate analysis

It is the technique used to find summary of a single data variable in form of segments.

Bivariate analysis

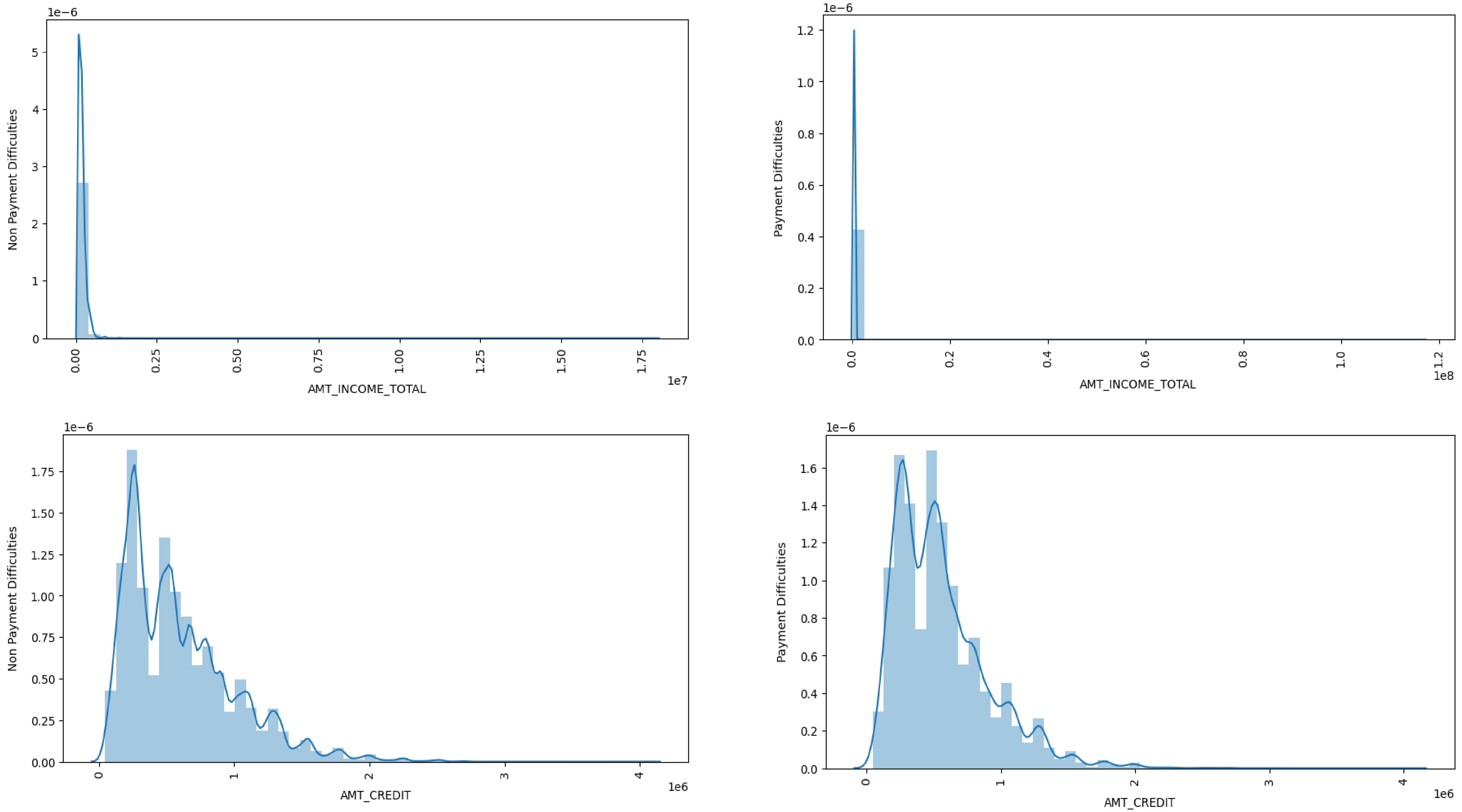
It is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. Here we analyze the changes occurred between the two variables and to what extent.

I have applied Univariate Analysis on Numerical columns with respect to TARGET column.

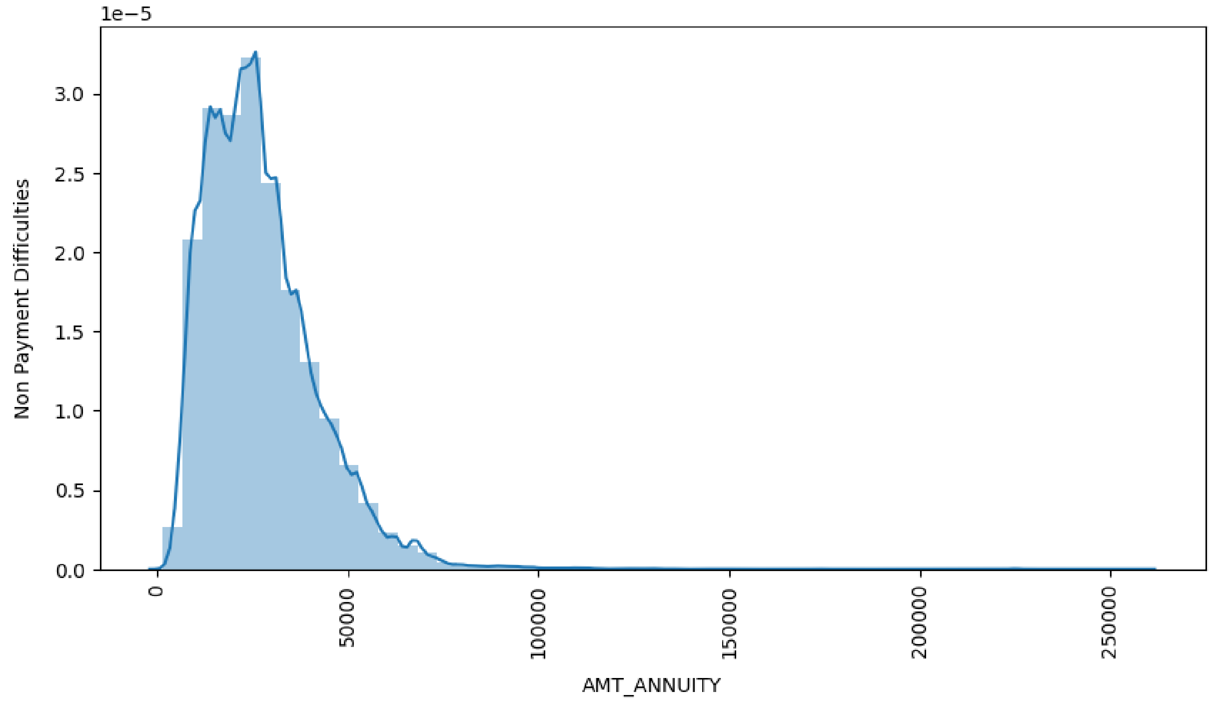
uni=["AMT_INCOME_TOTAL","AMT_CREDIT","AMT_ANNUITY","AMT_GOODS_PRICE"]

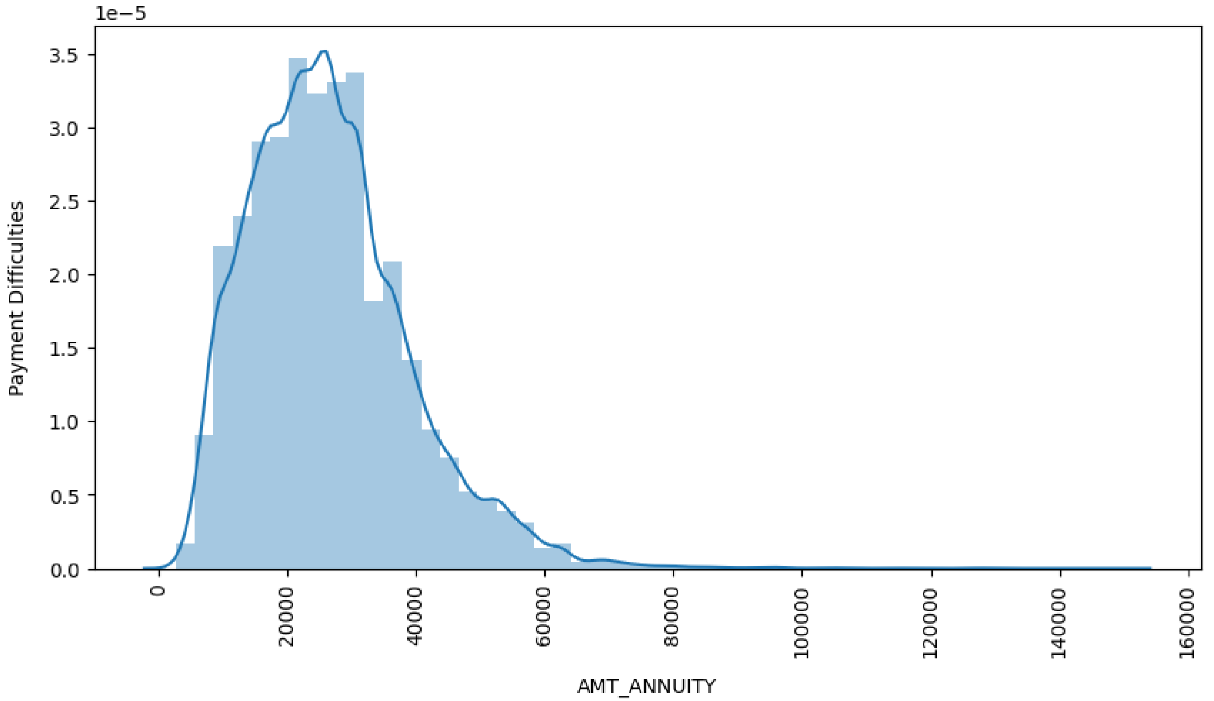
Insights

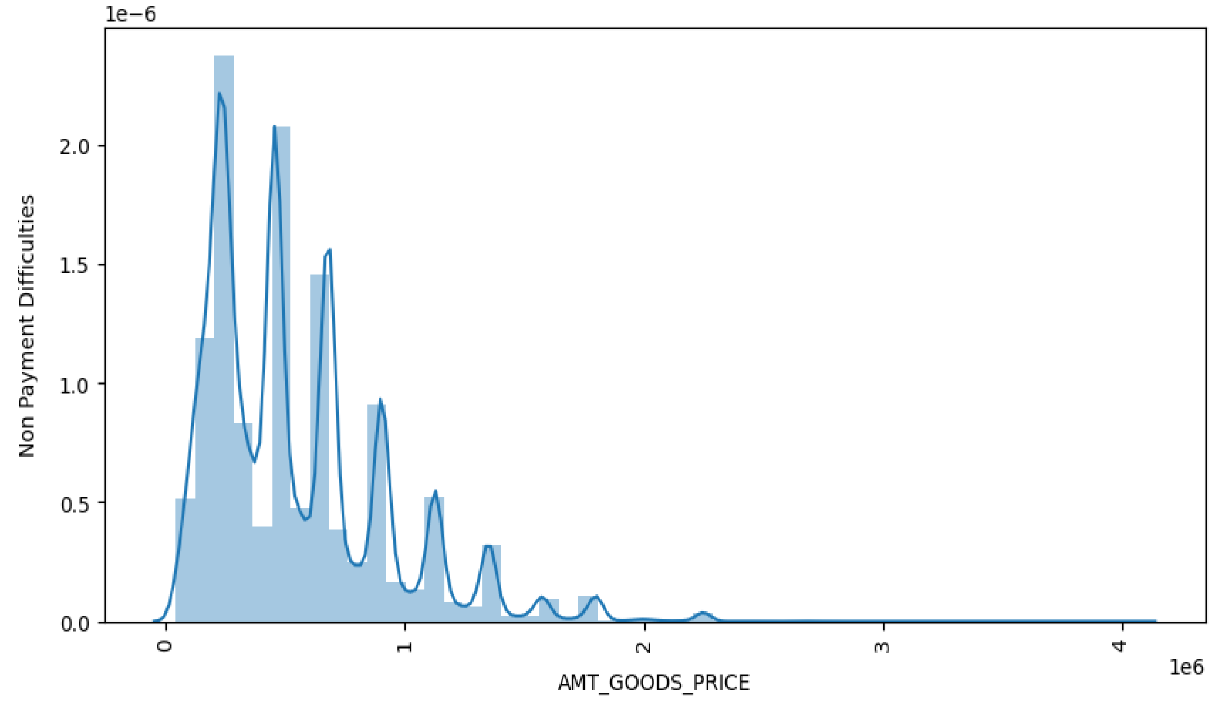
1. Insights are determined using the below mentioned Distribution plots*
2. Non Defaulter client has staggered income as compared to Default consumers. Distplot shows that the shape in Income total, Annuity, Credit and Good Price is similar for Non Defaulter and similar for Defaulter clients.
3. These plots also represents clients who have difficulty in paying back loans with respect to their income, loan amount, price of goods against which loan is procured and Annuity.

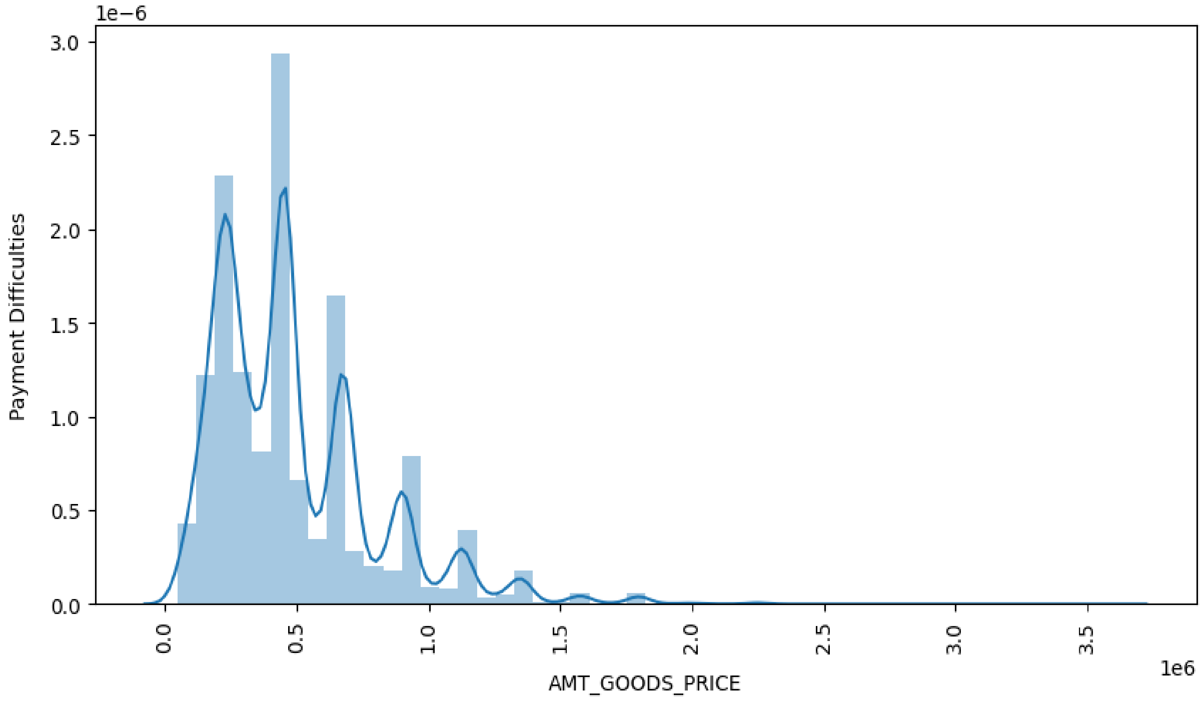


1e−6
5
4
3
2
1
0
Non Payment Difficulties
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e7
AMT_INCOME_TOTAL
1e−6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Payment Difficulties
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e8
AMT_INCOME_TOTAL
1e−6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Non Payment Difficulties
0
1
2
3
4
1e6
AMT_CREDIT
1e−6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Payment Difficulties
0
1
2
3
4
1e6
AMT_CREDIT


1e−5
Non Payment Difficulties
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
50000
100000
150000
200000
250000
AMT_ANNUITY

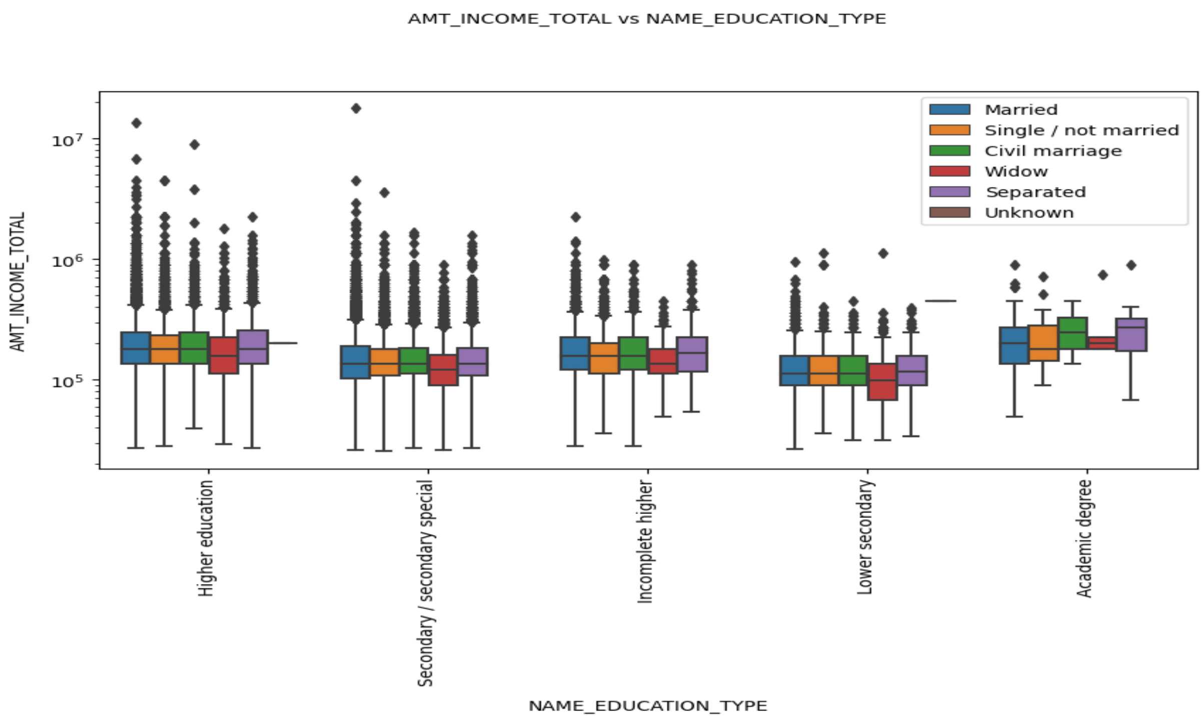

1e−5
Payment Difficulties
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
20000
40000
60000
80000
100000
120000
140000
160000
AMT_ANNUITY

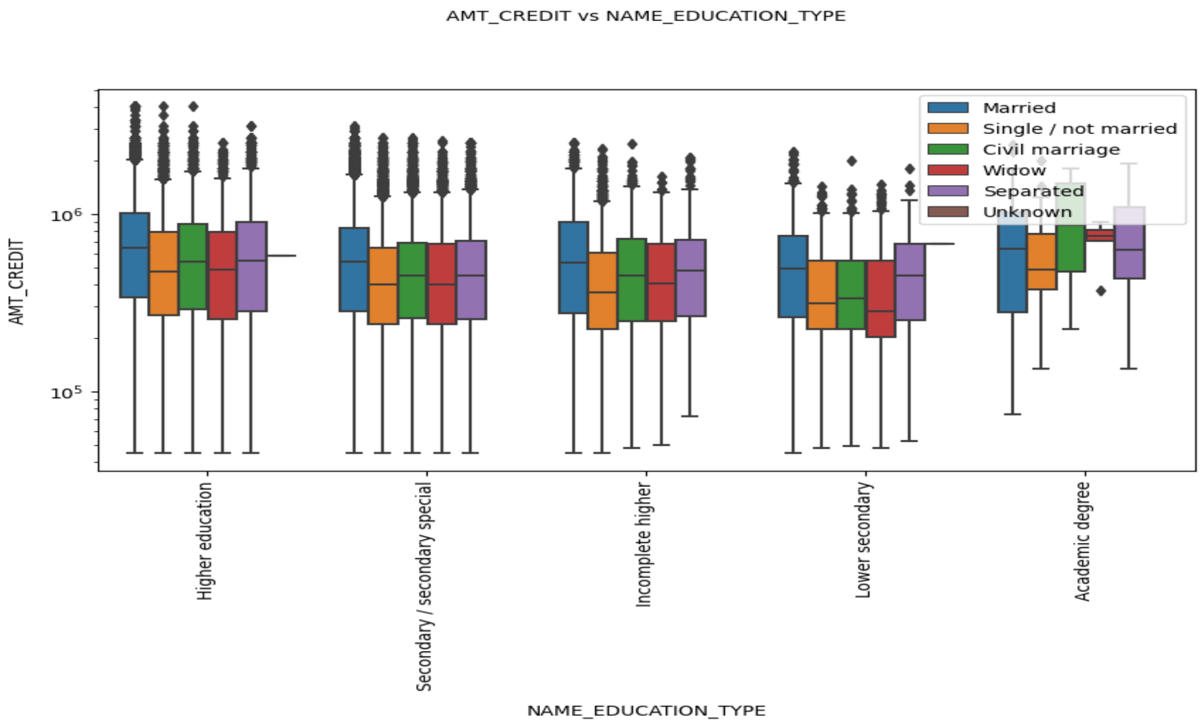

1e−6
Non Payment Difficulties
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
1e6
AMT_GOODS_PRICE


1e−6
Payment Difficulties
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e6
AMT_GOODS_PRICE

I have applied Bivariate Analysis on Numerical column with respect to Target column.

Insights are put down below the boxplot*



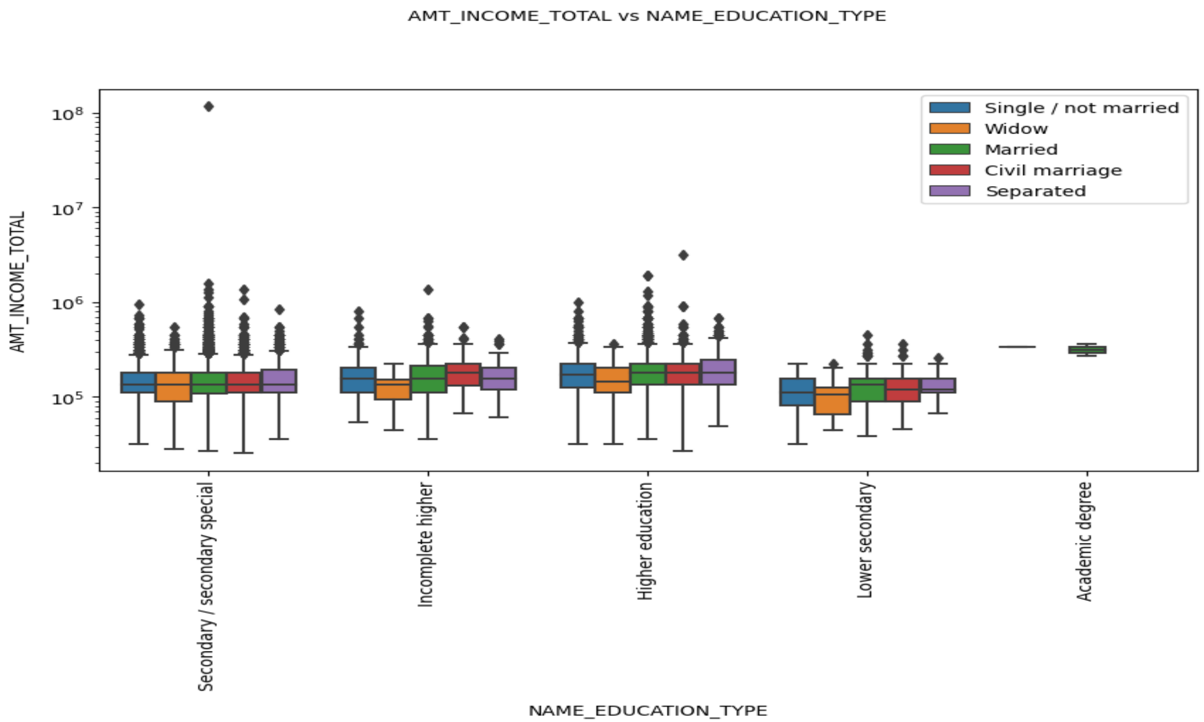


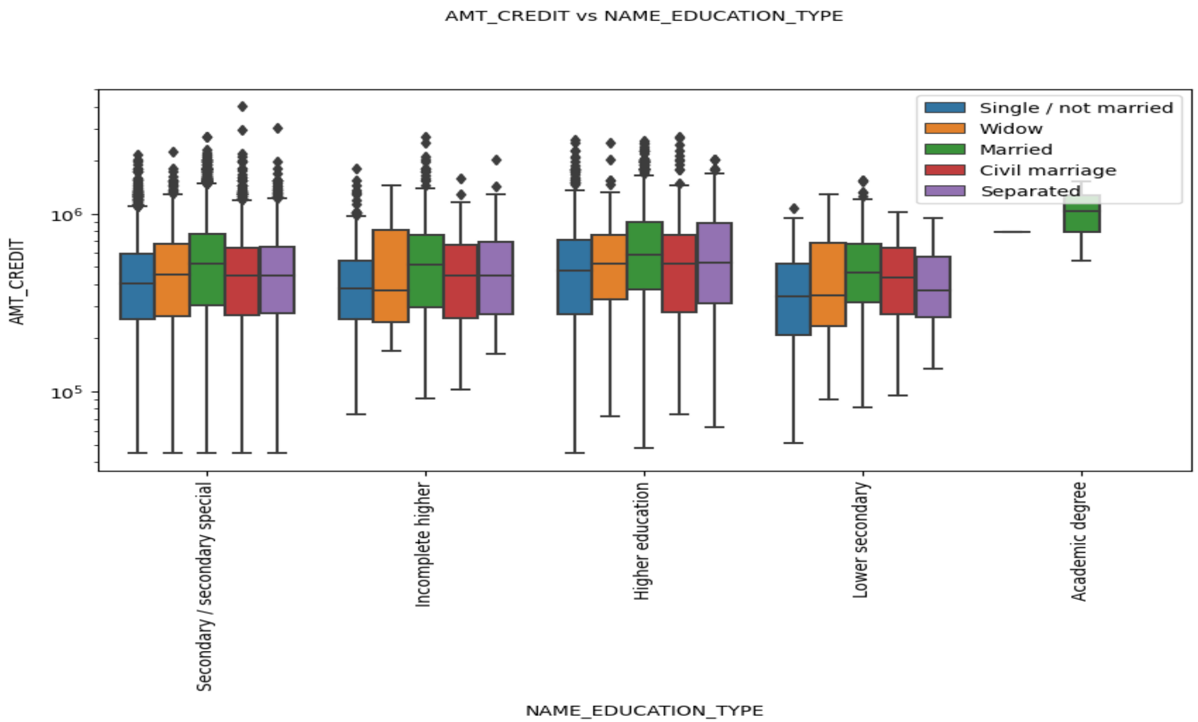
Insights

1. Some Non Defaulter clients having Higher Education has the highest income compared to others.
2. Incomplete Higher Education has higher incomes
3. Some of the clients having Secondary/Secondary Special Education has higher incomes.
4. Clients having Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special has more number of outliers
5. Non Defaulter Clients having academic degrees with all types of family statuses has very less outliers as compared to other types of education.

Insights

1. Some Non Defaulter clients having Higher Education, Secondary/Secondary Special Education, Incomplete Higher Education and Lower Secondary Education has high amount of credit.
2. Non Defaulter Clients with different **Education** types except **Academic degrees** have a large number of outliers.
3. Clients with an **Academic degree** and who is a Civil has **higher** credit loan.





Insights

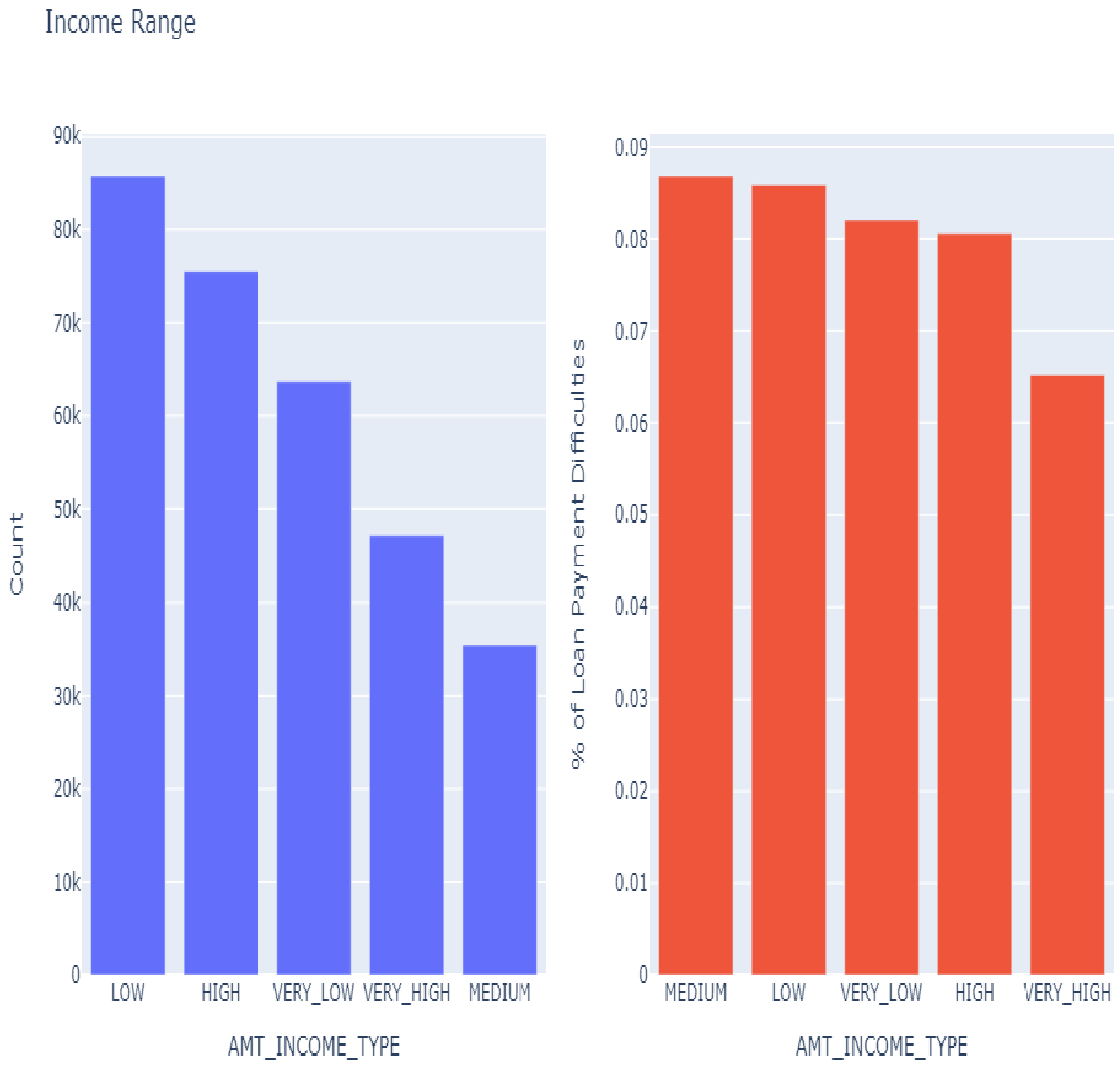
1. Defaulter Married clients with an academic degree has income amount which is much lesser as compared to others.
2. Defaulter Clients has less income as compared to Non Defaulter Clients

Insights

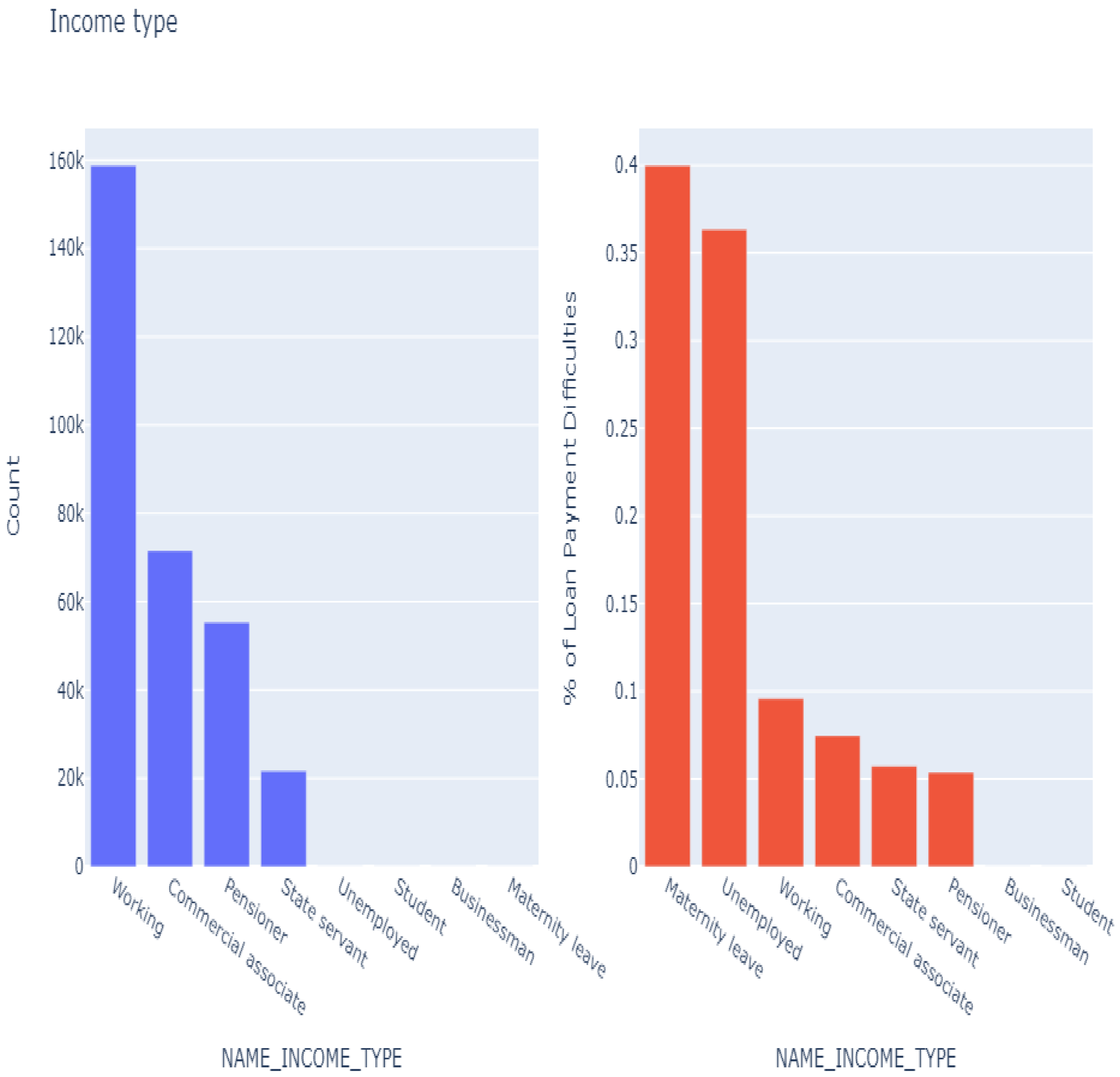
1. Some of the clients with Secondary/Secondary Special Education, Incomplete Higher Education, Higher Education, Lower Secondary Education has **high amount of credit loans**
2. Defaulter Married clients with an academic degree has higher credit loan
3. Single clients with **academic degrees have a very slim boxplot with no outliers**

I have applied Bivariate Analysis Categorical and plotted Barplots.

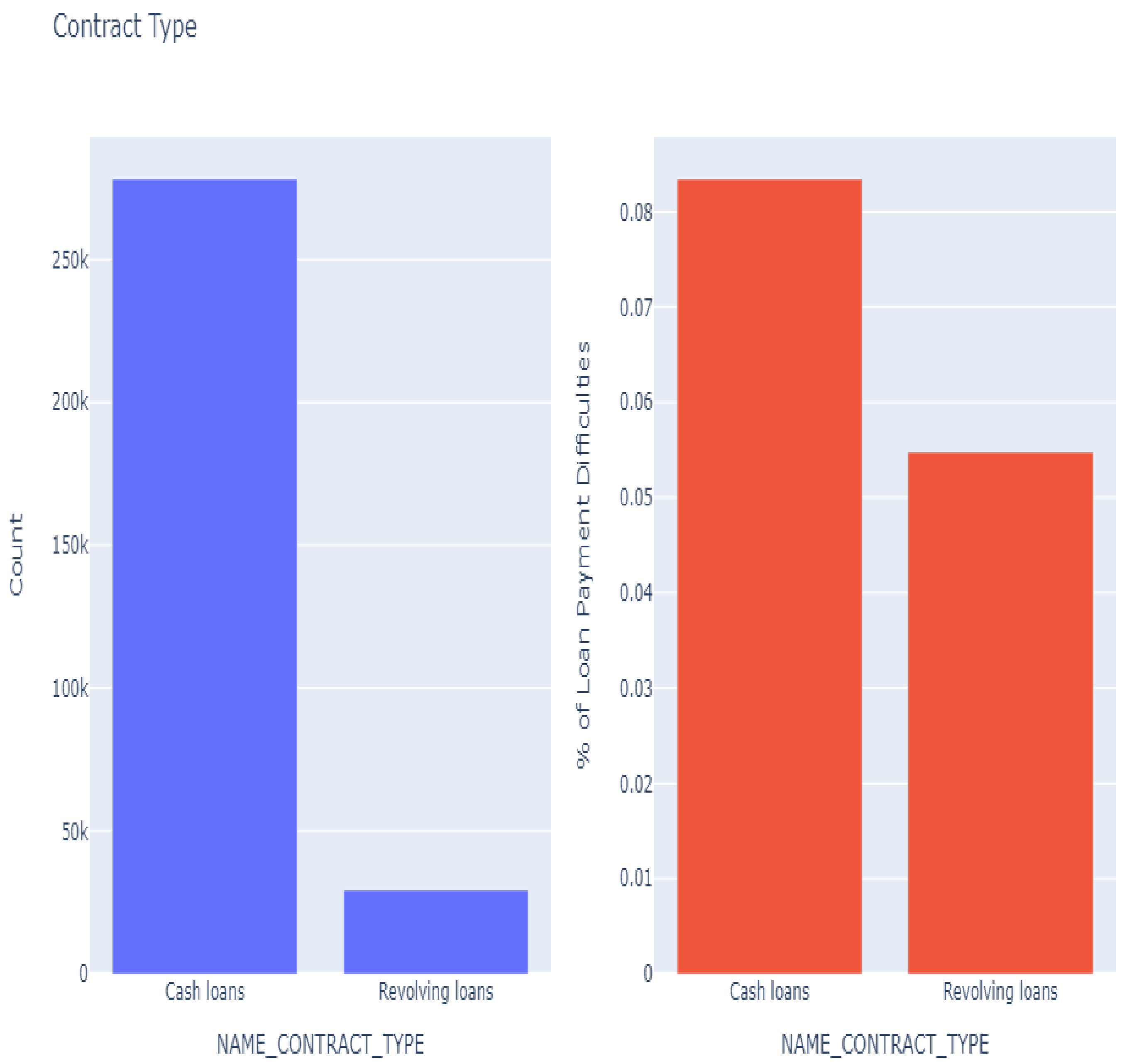
Distribution of Amount Income Range and the category with maximum % Loan-Payment Difficulties



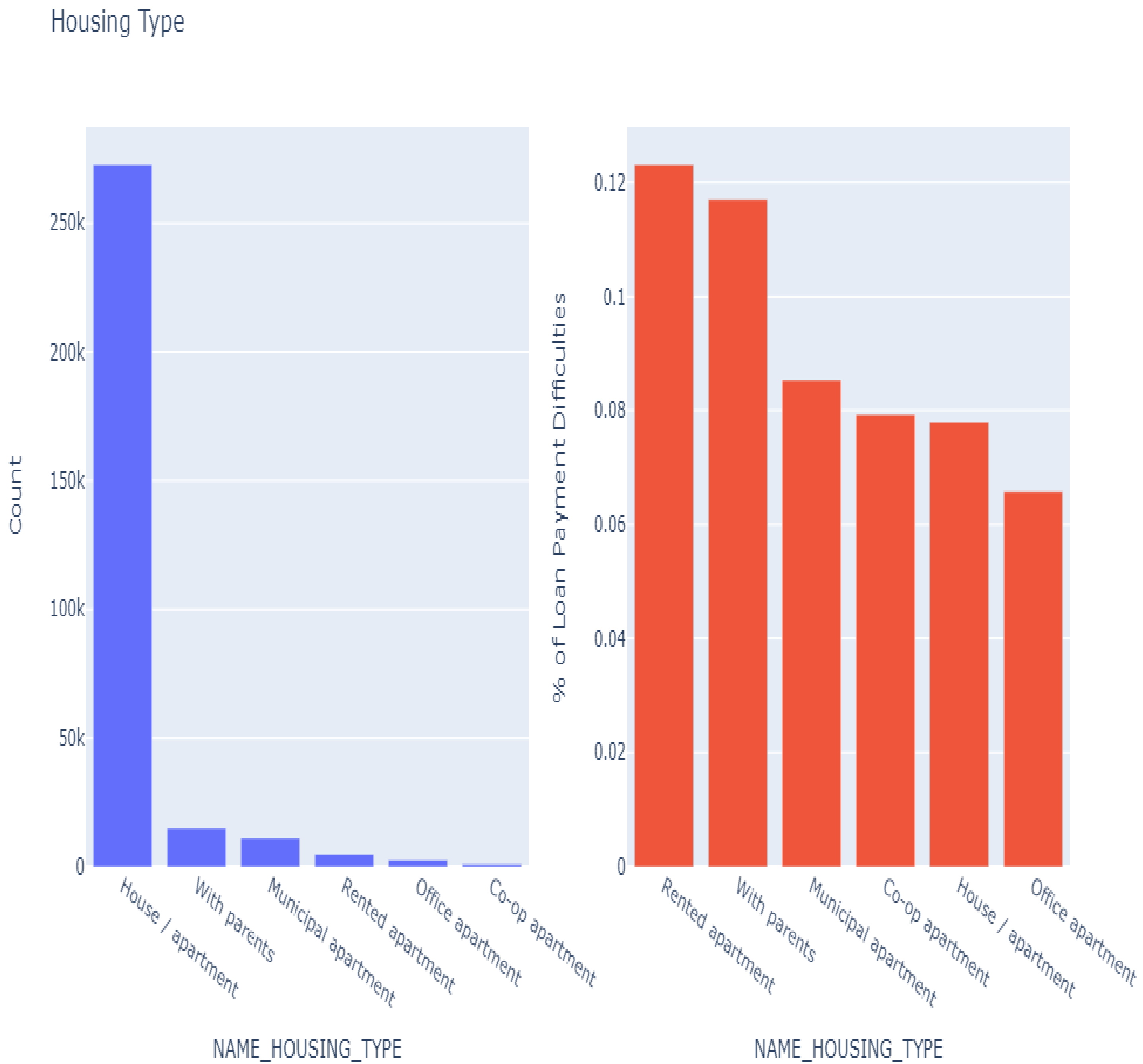
Distribution of Type of Income and the category with maximum Loan-Payment Difficulties



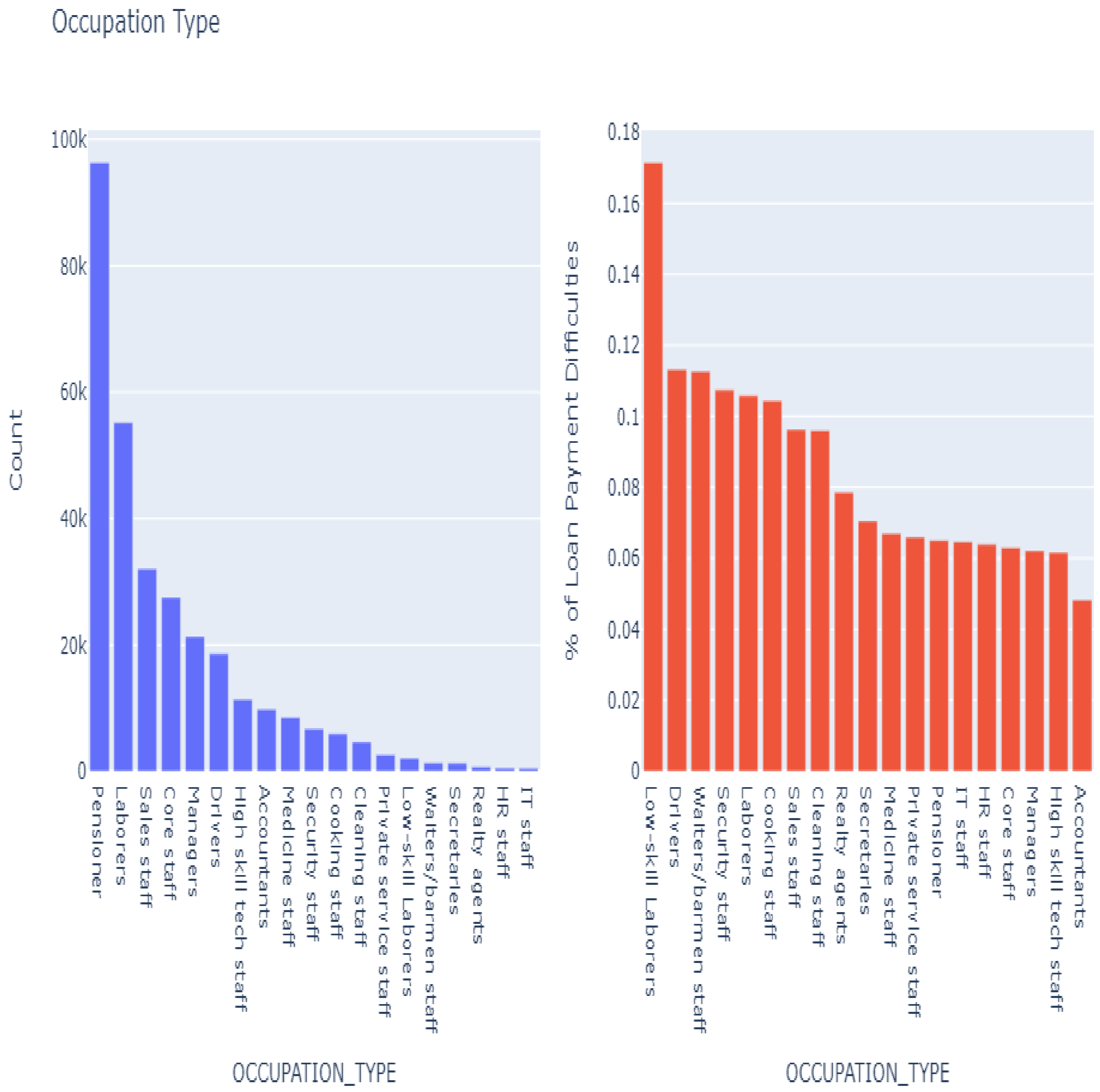
Distribution of Contract Type and the category with maximum Loan-Payment Difficulties



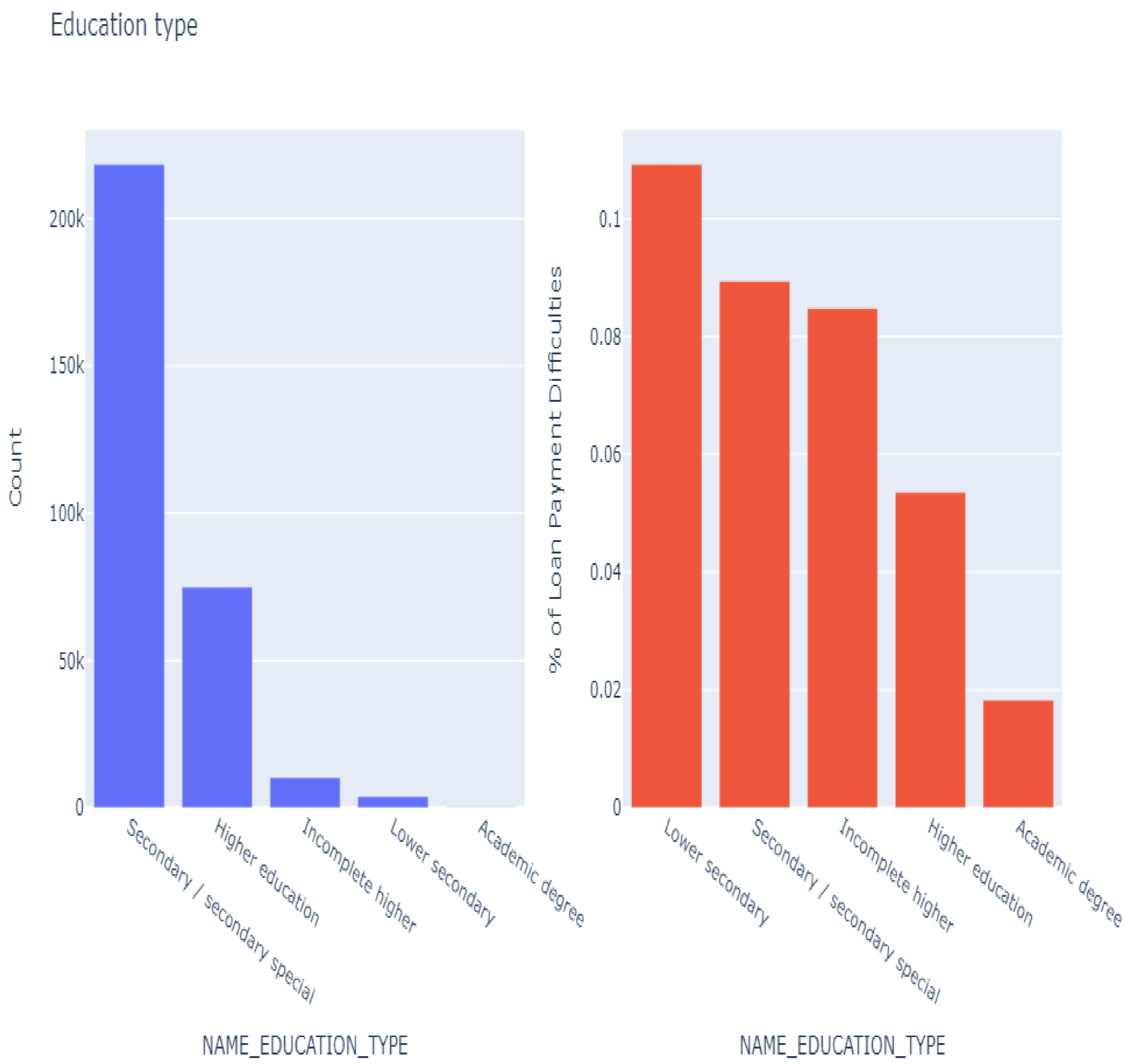
Distribution of Housing Type and the category with maximum Loan-Payment Difficulties



Distribution of Occupation Type and the category with maximum Loan-Payment Difficulties



Distribution of Education Type and the category with maximum Loan-Payment Difficulties



Find the top 10 **correlation** for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.

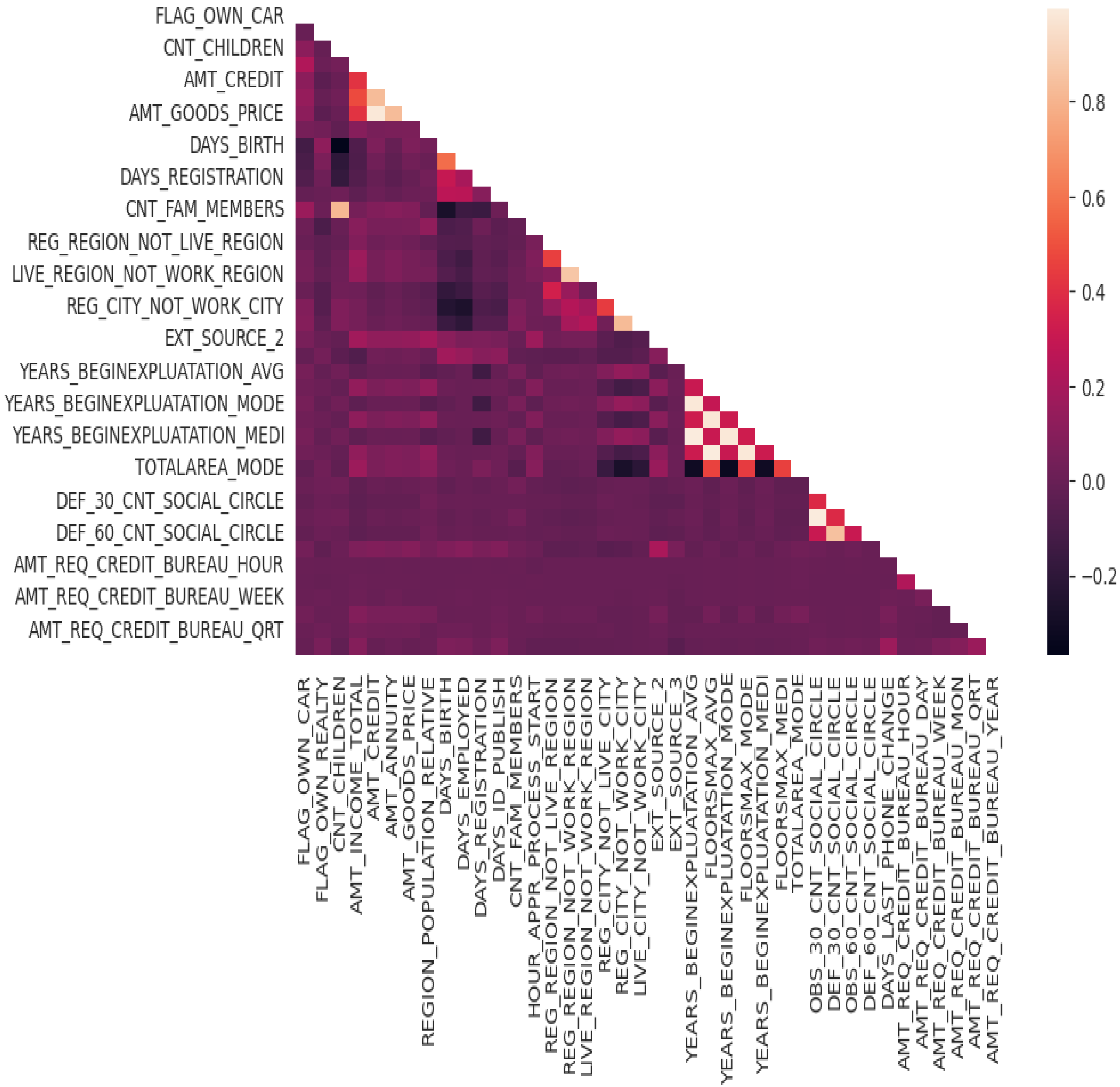
Correlations between numerical variables using Heatmap.

1. First I have found corelations distribution by separating the data into 2
2. Then plotted the correlation using heatmap.

Heatmap for Non Defaulter clients

Insights

1. AMT_CREDIT is higher in a densely populated area.
2. AMT_CREDIT is inversely proportional to the CNT_CHILDREN and DAYS_BIRTH
3. AMT_INCOME_TOTAL is also higher in a densely populated area.
4. AMT_INCOME_TOTAL is inversely proportional to the CNT_CHILDREN

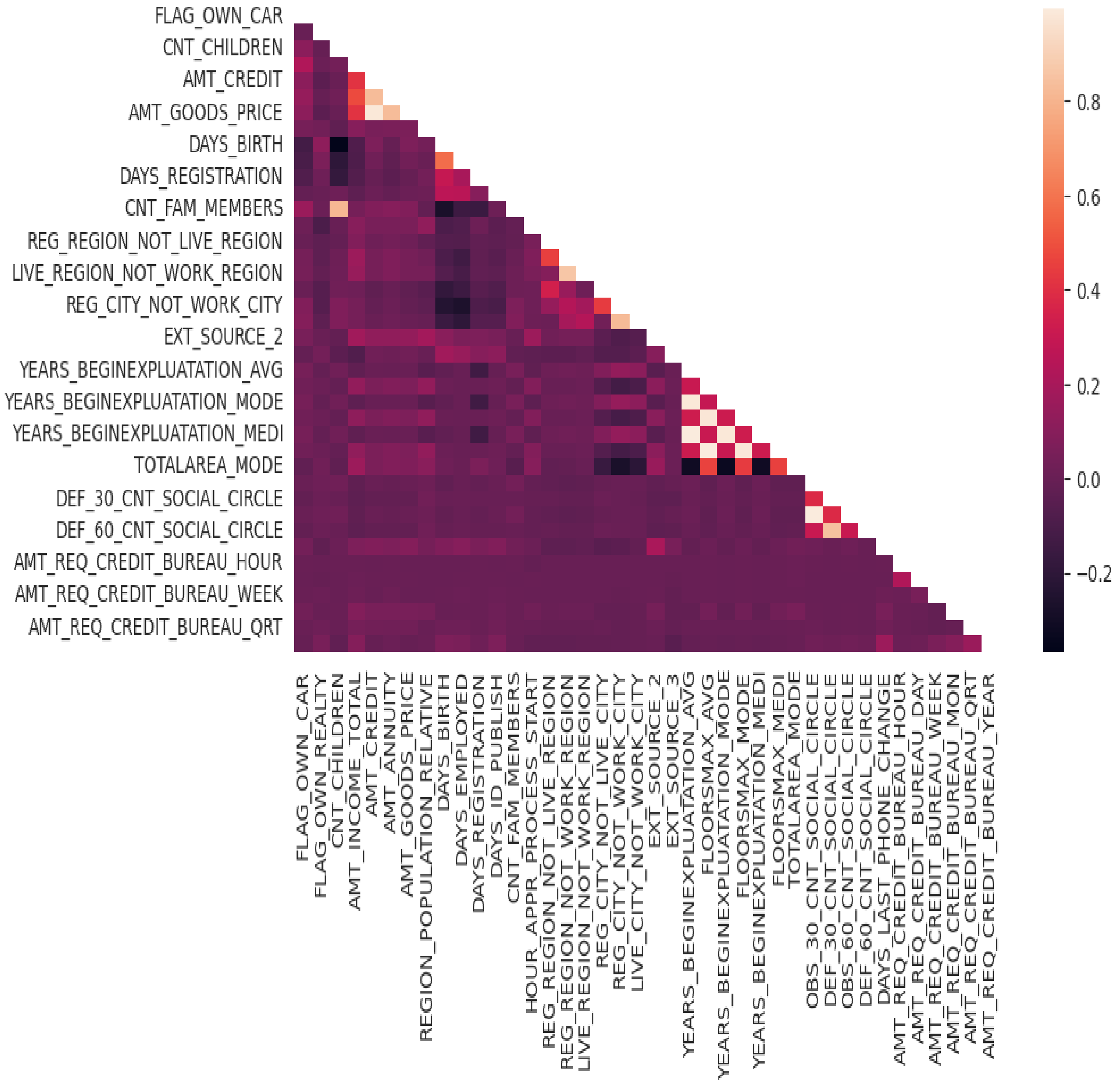


Heatmap for Defaulter clients

Insights

1. This heat map for Defaulter clients is almost similar to Non Defaulter clients . With few difference.

.



Top 10 Correlations for Non Defaulter and Defaulter Client

Insight

Correlation in both Non Defaulter and Defaulter Client are almost same.

Non Defaulter Client

In [57]:
```
Columns=t0.columns
corr=t0[Columns].corr(method = "pearson")
corr=corr.where(np.triu(np.ones(corr.shape),k=1).astype(np.bool))
top10_corr0=corr.unstack().reset_index()
```

In [58]:
```
top10_corr0.columns = ["VAR1","VAR2","CORRELATION"]
top10_corr0.dropna(subset=["CORRELATION"],inplace=True)
top10_corr0["CORR_ABS"]=top10_corr0["CORRELATION"].abs()
top10_corr0.sort_values("CORR_ABS", ascending=False).head(10)
```

Out[58]:

	VAR1	VAR2	CORRELATION	CORR_ABS
1417	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998510	0.998510
1243	FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997253	0.997253
1200	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.993656	0.993656
1245	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.988955	0.988955
342	AMT_GOODS_PRICE	AMT_CREDIT	0.987250	0.987250
1159	FLOORSMAX_MODE	FLOORSMAX_AVG	0.986569	0.986569
1116	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.971366	0.971366
1202	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.962498	0.962498
592	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571	0.878571
773	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861	0.861861

Defaulter Client

In [59]:
```
Columns=t1.columns
corr1=t1[Columns].corr(method = "pearson")
corr1=corr1.where(np.triu(np.ones(corr1.shape),k=1).astype(np.bool))
top10_corr1=corr1.unstack().reset_index()
```

In [60]:
```
top10_corr1.columns = ["VAR1","VAR2","CORRELATION"]
top10_corr1.dropna(subset=["CORRELATION"],inplace=True)
top10_corr1["CORR_ABS"]=top10_corr1["CORRELATION"].abs()
top10_corr1.sort_values("CORR_ABS", ascending=False).head(10)
```

Out[60]:

	VAR1	VAR2	CORRELATION	CORR_ABS
1417	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998270	0.998270
1243	FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997295	0.997295
1200	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.996181	0.996181
1245	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.989472	0.989472
1159	FLOORSMAX_MODE	FLOORSMAX_AVG	0.986935	0.986935
342	AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
1116	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.980758	0.980758
1202	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.978399	0.978399
592	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
1460	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869016	0.869016

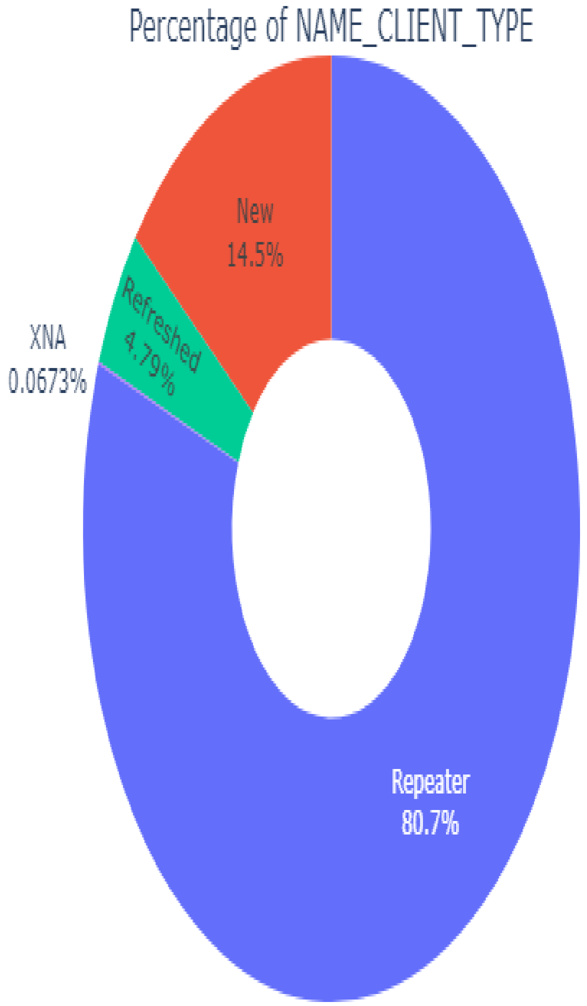
The above analysis is done on application_data.csv

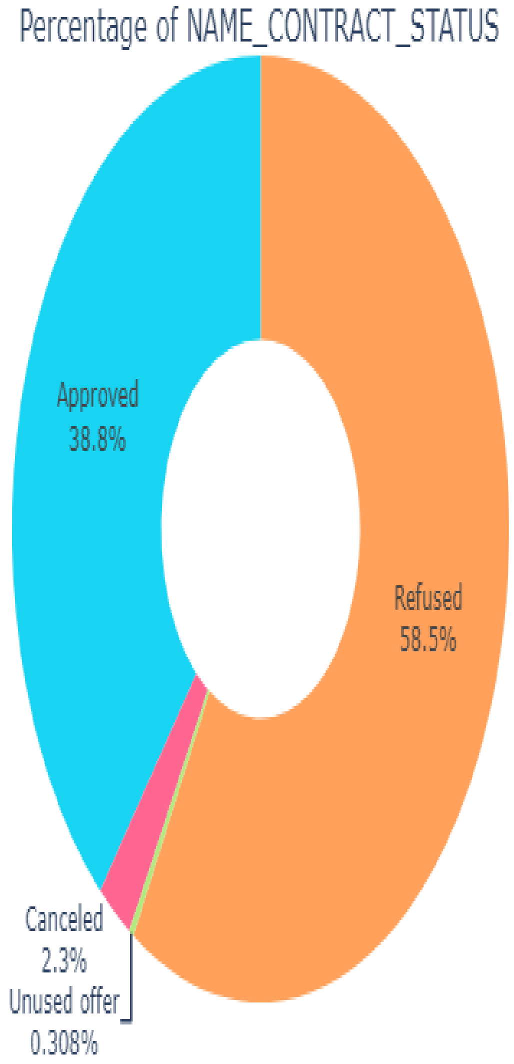
For previous_application.csv

- First load dataset in juypter notebook
- After loading into dataframe I have inspected it and carried out data cleaning process.
- Percentage of null values for each column is calculated and columns are drop based on your requirement of percentage of null values >40% .
- I have handled null value by imputing columns having null values ≤ 24% with Mode values for numeric columns except for continuous numeric columns we imputed with Median value
- I have removed column values which has XNA and XAP value.
- Then I have merged the application_data with previous_application dataset by creating a new dataframe.
- After merging I have renamed the column names

Loan Distributions and Purposes

1) Percentage count of NAME_CLIENT_TYPE and NAME_CONTRAST_STATUS





Insights

NAME_CLIENT_TYPE

80.7% clients are repeaters for applying loan

14.5% clients are new for applying loan

NAME_CONTRAST_STATUS

Approved: 38.8%

Canceled : 2.3%

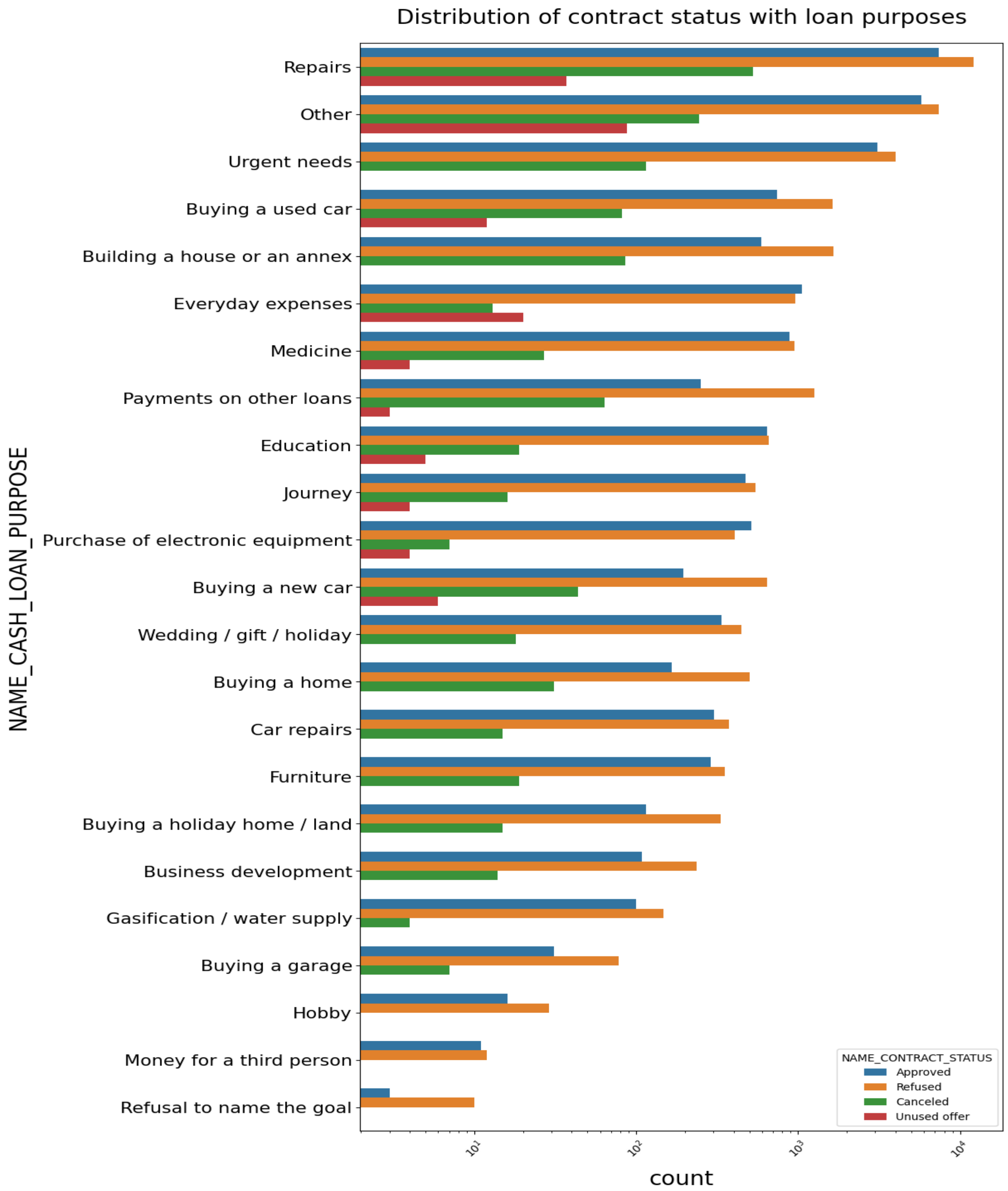
Unused offer : 0.308%

Refused : 58.5%

NAME_CONTRACT_STATUS with NAME_CASH_LOAN_PURPOSE based on log scale for easy representation

Insights

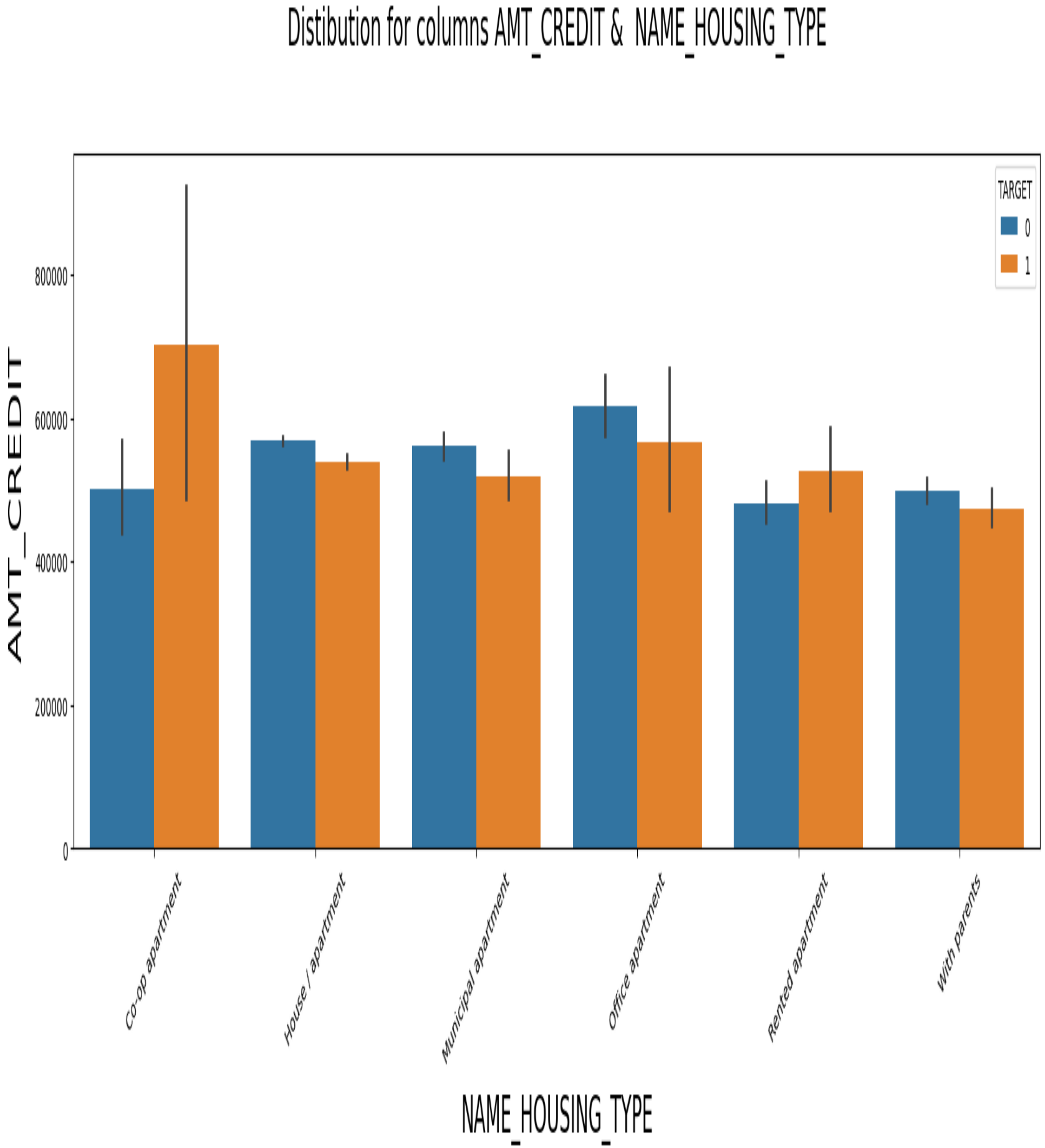
- Most approval and rejection of loans are from Repairs
- Car repair and furniture has equal number of approves and rejection
- Payments on other loans, Buying a new car, Buying a Holiday home/land, Buying a garage has more number of rejections than approves.



Barplot Distribution for column AMT_CREDIT and NAME_HOUSING_TYPE

Insights

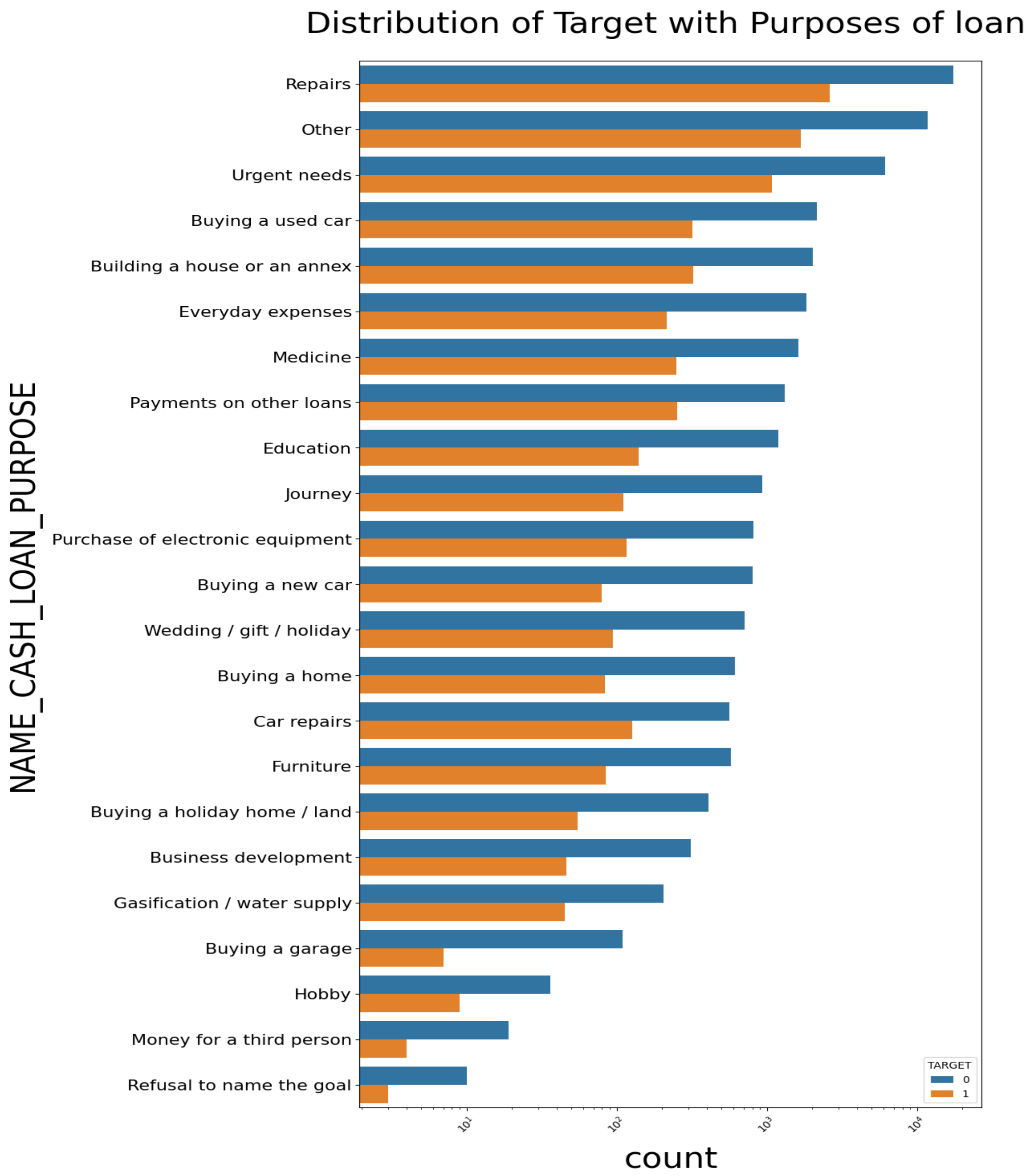
- office apartment is having higher credit of Non Defaulter clients and co-op apartment is having higher credit of Defaulter clients. So bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House\apartment or municipal apartment for successful payments.



TARGET with NAME_CASH_LOAN_PURPOSE based on log scale for the ease of representation

Insights

- Repairs has high variation for both clients.
- Car repair and furniture has equal ratio in payment difficulties



5) RESULT

After performing all these analysis process we can able identify patterns which indicate if a client has difficulty paying their installments 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.

Finally after performing all these analysis we can reduce the risks associated with the bank.