EDA CREDIT ASSIGNMENT

By

Sannan Dabir

Ph. No: 8169683931.

Email id: sannandabir26@gmail.com.

PROBLEM STATEMENT

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected. 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.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,

All other cases: All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

Approved: The Company has approved loan Application

Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.

Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

Unused offer: Loan has been cancelled by the client but at different stages of the process. In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

Business Objectives

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.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

STEPS FOLLOWED FOR EDA:

1. Reading the data and checking basic information.
2. Removing the unwanted Columns.
3. Imputing the missing values of Columns.
4. Fixing the invalid values and Binning the columns.
5. Outliers Identification.
6. Performing the Uni-variate Analysis.
7. Performing the Bi-variate and Multi-variate analysis.

Analysis for application data

Step 1: Reading the data and checking the basic information.

- The application data is stored in a data frame called data1 and it contains **3,07,511 rows** and **122 columns** initially.

Step 2: Removing the unwanted columns.

Step2A: Removing the columns having a higher number of Null Values:

- All the columns having null value percentages **higher than 35%** have been dropped from the data frame.

The following columns have been dropped as they have a large number of null values.

```
In [12]: Null_Col_35.index

Out[12]: Index(['COMMONAREA_MEDI', 'COMMONAREA_AVG', 'COMMONAREA_MODE',
               'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAPARTMENTS_AVG',
               'NONLIVINGAPARTMENTS_MEDI', 'FONDKAPREMONT_MODE',
               'LIVINGAPARTMENTS_MODE', 'LIVINGAPARTMENTS_AVG',
               'LIVINGAPARTMENTS_MEDI', 'FLOORSMIN_AVG', 'FLOORSMIN_MODE',
               'FLOORSMIN_MEDI', 'YEARS_BUILD_MEDI', 'YEARS_BUILD_MODE',
               'YEARS_BUILD_AVG', 'OWN_CAR_AGE', 'LANDAREA_MEDI', 'LANDAREA_MODE',
               'LANDAREA_AVG', 'BASEMENTAREA_MEDI', 'BASEMENTAREA_AVG',
               'BASEMENTAREA_MODE', 'EXT_SOURCE_1', 'NONLIVINGAREA_MODE',
               'NONLIVINGAREA_AVG', 'NONLIVINGAREA_MEDI', 'ELEVATORS_MEDI',
               'ELEVATORS_AVG', 'ELEVATORS_MODE', 'WALLSMATERIAL_MODE',
               'APARTMENTS_MEDI', 'APARTMENTS_AVG', 'APARTMENTS_MODE',
               'ENTRANCES_MEDI', 'ENTRANCES_AVG', 'ENTRANCES_MODE', 'LIVINGAREA_AVG',
               'LIVINGAREA_MODE', 'LIVINGAREA_MEDI', 'HOUSETYPE_MODE',
               'FLOORSMAX_MODE', 'FLOORSMAX_MEDI', 'FLOORSMAX_AVG',
               'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BEGINEXPLUATATION_MEDI',
               'YEARS_BEGINEXPLUATATION_AVG', 'TOTALAREA_MODE', 'EMERGENCYSTATE_MODE'],
              dtype='object')
```

Step 2B: Removing the Unnecessary Columns.

- Removing the columns which are unnecessary for the analysis. Find below the list of columns that have been removed.

```
In [17]: Flag_Doc_Columns=data1[['FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_3',
          'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5', 'FLAG_DOCUMENT_6',
                'FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8', 'FLAG_DOCUMENT_9',
                'FLAG_DOCUMENT_10', 'FLAG_DOCUMENT_11', 'FLAG_DOCUMENT_12',
                'FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14', 'FLAG_DOCUMENT_15',
                'FLAG_DOCUMENT_16', 'FLAG_DOCUMENT_17', 'FLAG_DOCUMENT_18',
                'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20', 'FLAG_DOCUMENT_21','FLAG_MOBIL',
                'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE',
                'FLAG_EMAIL']]
```

- After removing the null value columns and unnecessary columns, the data frame contains a total of 47 columns.

<u>Step 3: Imputing the missing values.</u>

- **18 Columns** contain missing values of which the Occupation Type column contains the most missing values.
- As there is a very large number of null values in Occupation Type, they are not imputed as imputing them may hamper the analyses and may give biased data.
- Instead, the null values are replaced by the **"Missing"** value.

OCCUPATION_TYPE	31.345545
EXT_SOURCE_3	19.825307
AMT_REQ_CREDIT_BUREAU_YEAR	13.501631
AMT_REQ_CREDIT_BUREAU_QRT	13.501631
AMT_REQ_CREDIT_BUREAU_MON	13.501631
AMT_REQ_CREDIT_BUREAU_WEEK	13.501631
AMT_REQ_CREDIT_BUREAU_DAY	13.501631
AMT_REQ_CREDIT_BUREAU_HOUR	13.501631
NAME_TYPE_SUITE	0.420148
OBS_30_CNT_SOCIAL_CIRCLE	0.332021
OBS_60_CNT_SOCIAL_CIRCLE	0.332021
DEF_30_CNT_SOCIAL_CIRCLE	0.332021
DEF_60_CNT_SOCIAL_CIRCLE	0.332021
EXT_SOURCE_2	0.214626
AMT_GOODS_PRICE	0.090403
AMT_ANNUITY	0.003902
CNT_FAM_MEMBERS	0.000650
DAYS_LAST_PHONE_CHANGE	0.000325

- The columns *AMT_REQ_CREDIT_BUREAU_YEAR, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, NAME_TYPE_SUITE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, OBS_30_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE* having the null values have been imputed with the **mode value** of the columns.
- The mode value has been used to impute as the columns are not continuous numerical columns.
- The columns *AMT_GOODS_PRICE, AMT_ANNUITY, CNT_FAM_MEMBERS, DAYS_LAST_PHONE_CHANGE, EXT_SOURCE_3, and EXT_SOURCE_2* have been imputed by the **median value** of the respective columns.
- After imputing the values, no column of the application_data data frame has null values.

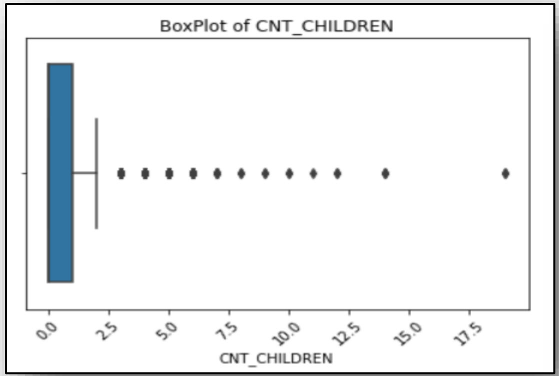
Step 4: Fixing the invalid values and Binning the columns.

- DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION and DAYS_ID_PUBLISH columns had **negative values** which are converted to absolute values.
- CODE_GENDER column had 4 XNA values which are replaced by the **mode value** as it is a categorical column.
- Also, the 55,374 XNA values in the ORGANIZATION_TYPE column have been replaced by the '**Others**' value as imputing such a large number of values may hamper the analysis.
- FLAG_OWN_CAR and FLAG_OWN_REALTY have **Y/N values** which have been converted to **1 and 0** to ease the analysis process.

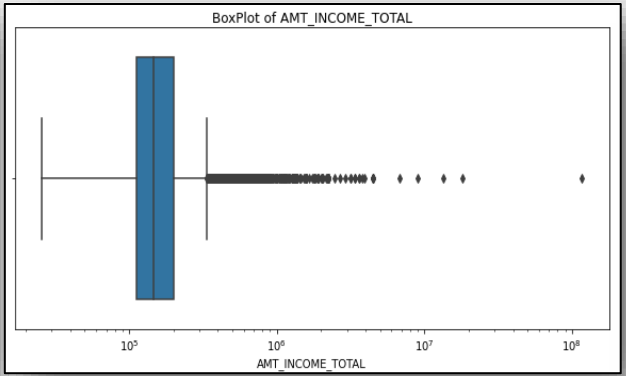
- The AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE have been binned to new columns having the **Very_Low, Low, Medium, High and Very_High values** based on **0, 0.2, 0.4, 0.6, 0.8, 1 quantile** using the qcut.
- Similarly, the DAYS_BIRTH has also been converted to AGE CATEGORY having the categorical values of **Young, Young_Adult, Middle_Age, and Senior** values based on the **bins value of 18, 24, 35, 60, 100.**
- The following new columns have been added by binning the data
  - AMT_INCOME_CAT
  - AMT_CREDIT_CAT
  - AMT_ANNUITY_CAT
  - AMT_GOODS_PRICE_CAT
  - AGE_CAT

Step 5: Handling Outliers.

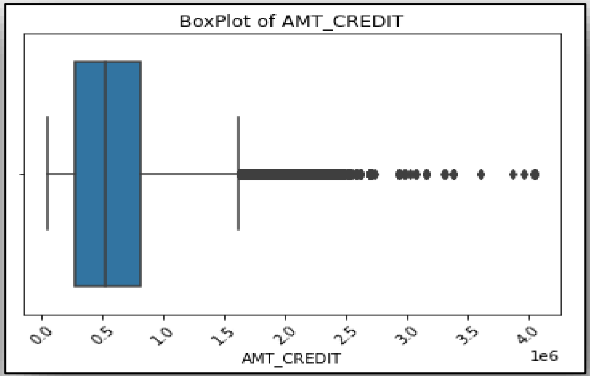
1. CNT_CHILDREN has some outliers as there are some values outside of the upper fence.



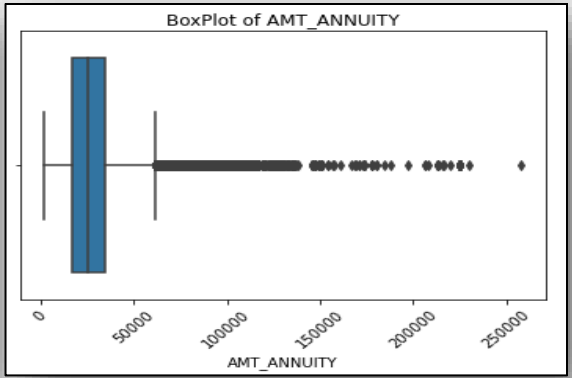
2. AMT_INCOME_TOTAL has large number of outliers as there are values outside of the upper fence.



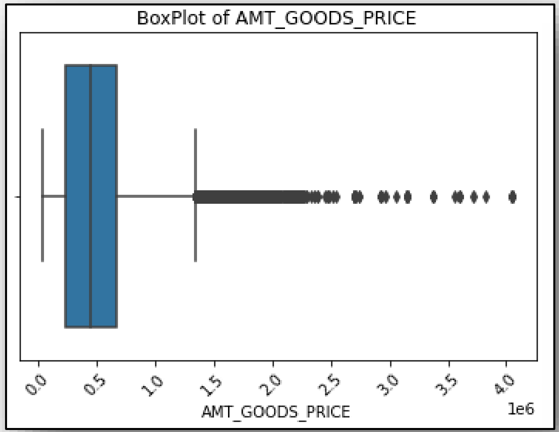
3. AMT_CREDIT has large number of outliers as there are many values outside of the upper fence.



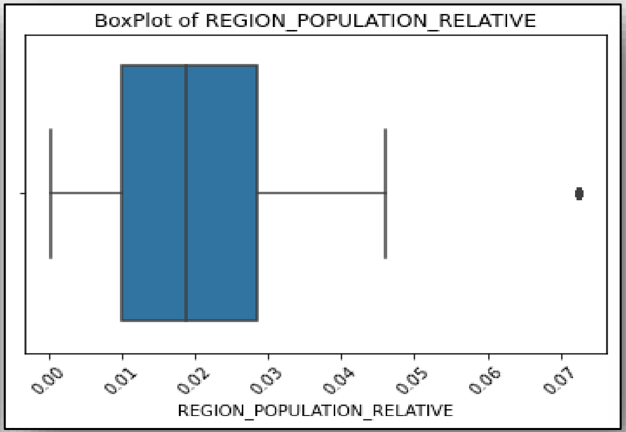
4. AMT_ANNUITY column also has a large number of outliers as many values are above the upper fence of boxplot.



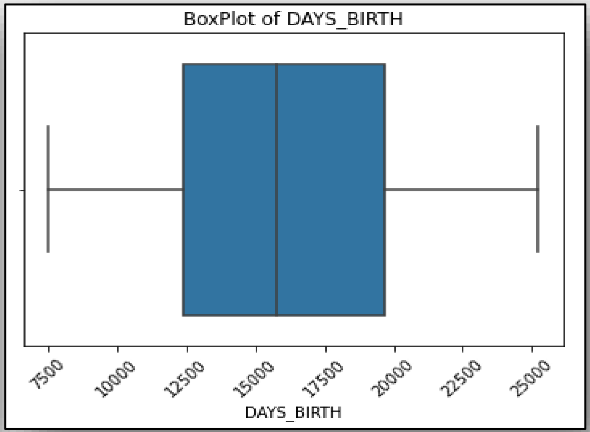
5. AMT_GOODS_PRICE also has many outliers as there are large number of values outide upper fence.



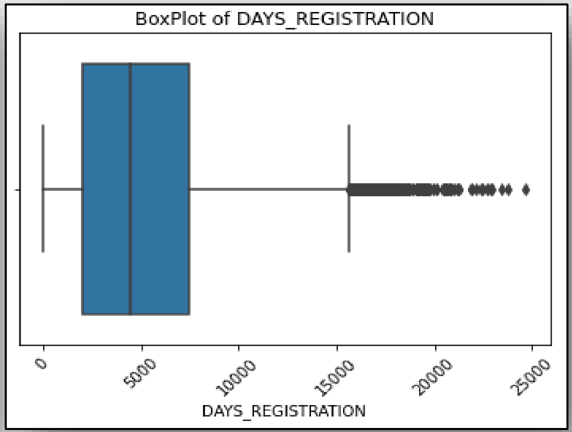
6. REGION_POPULATION_RELATIVE has very very low outliers



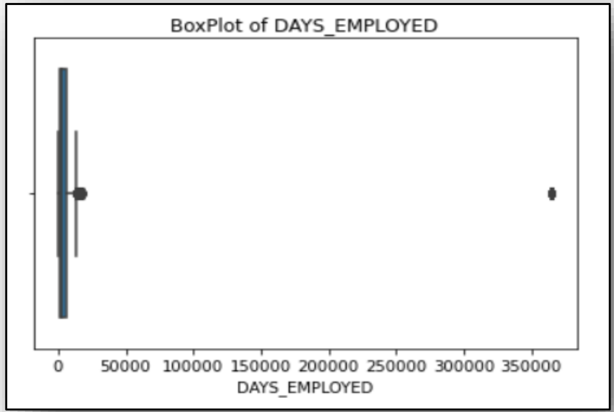
7. DAYS_BIRTH does not has any outlier



8. DAYS_REGISTRATION also contains many outliers as seen by the values outside the upper fence.



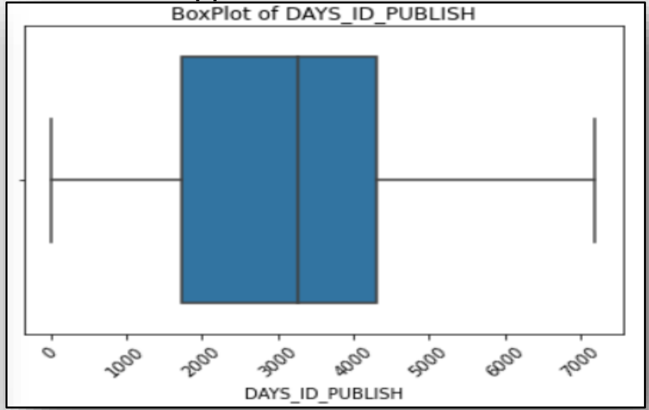
9. DAYS_EMPLOYED has some outliers as there are many values outside of the upper fence.
Also, there are **55,374** values of **3,65,243** which is illogical.



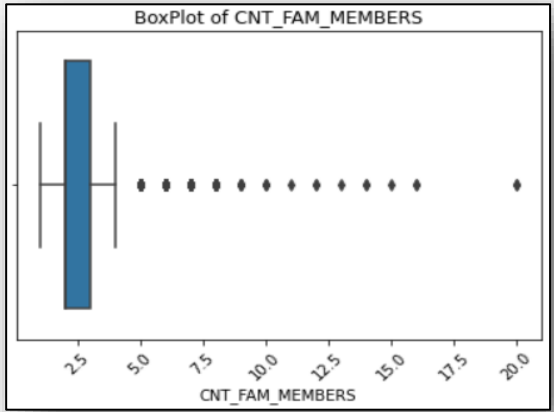

```
In [284]: data1['DAYS_EMPLOYED'].value_counts().sort_values(ascending=False)

Out[284]: 365243    55374
          200         156
          224         152
          230         151
          199         151
                     ...
          9281          1
          9556          1
          14743         1
          16266         1
          8694          1
          Name: DAYS_EMPLOYED, Length: 12574, dtype: int64
```

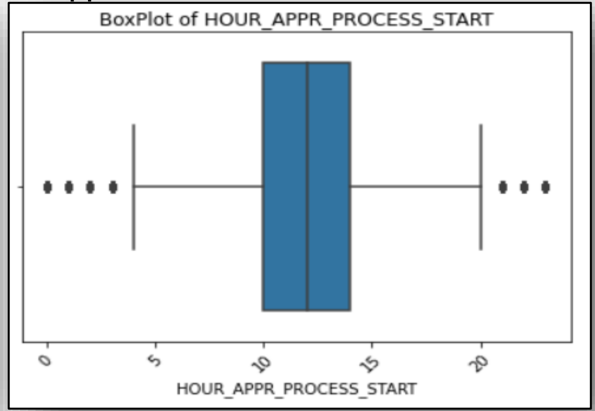
10. DAYS_ID_PUBLISH does not has any outlier as no values are outside the lower and upper fence.



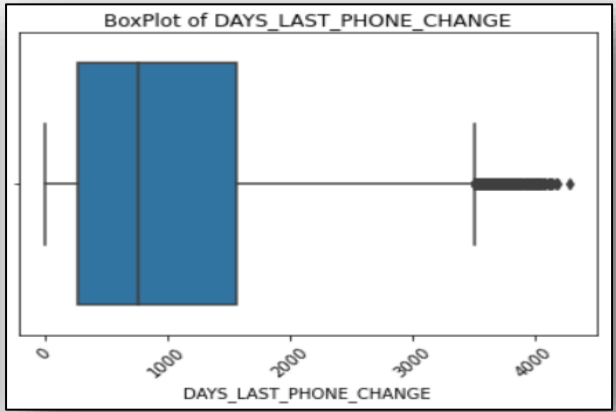
11. CNT_FAM_MEMBERS has very few outliers and most of the applicants have around 4 family members.



12. HOUR_APPR_PROCESS_START also has very few outliers as there are some values outside the lower and upper fence.

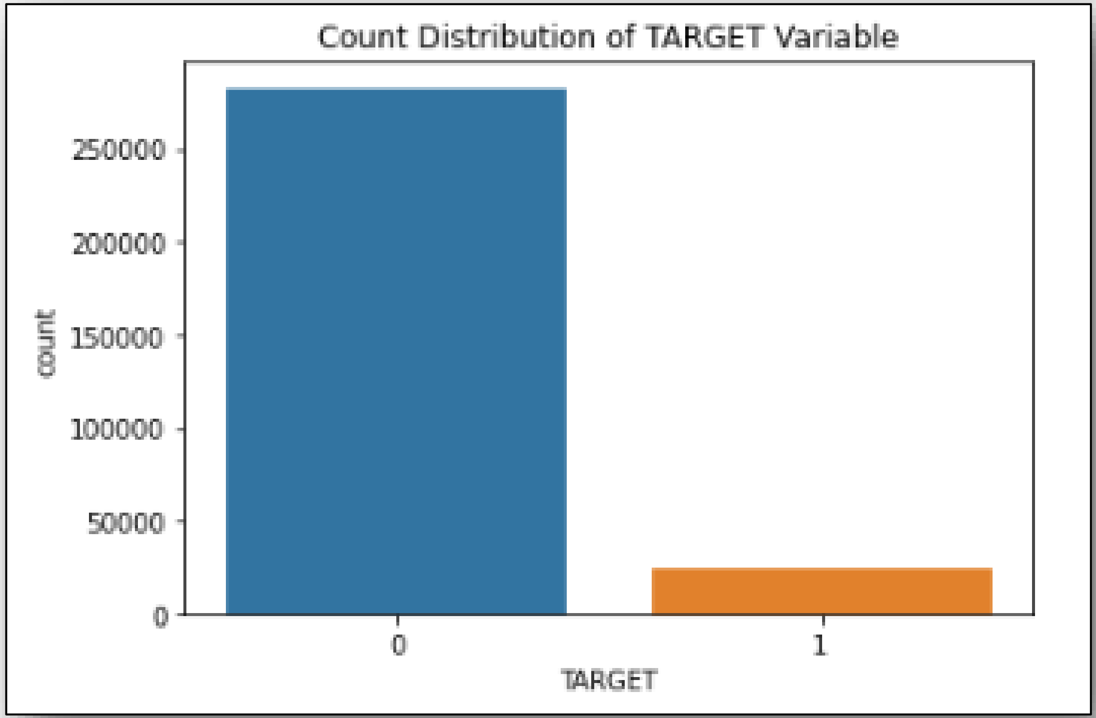


13. DAYS_LAST_PHONE_CHANGE also has outliers as there are data points outside the upper fence.

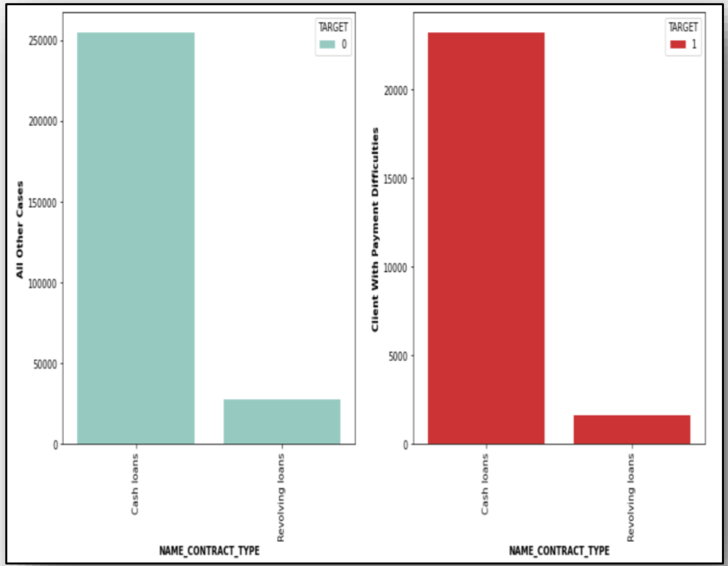


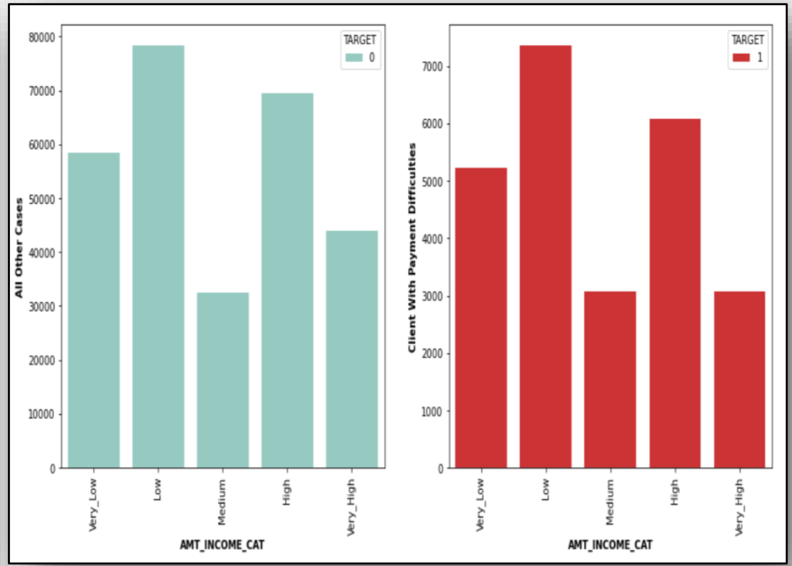
DATA IMBALANCE OF TARGET VARIABLE

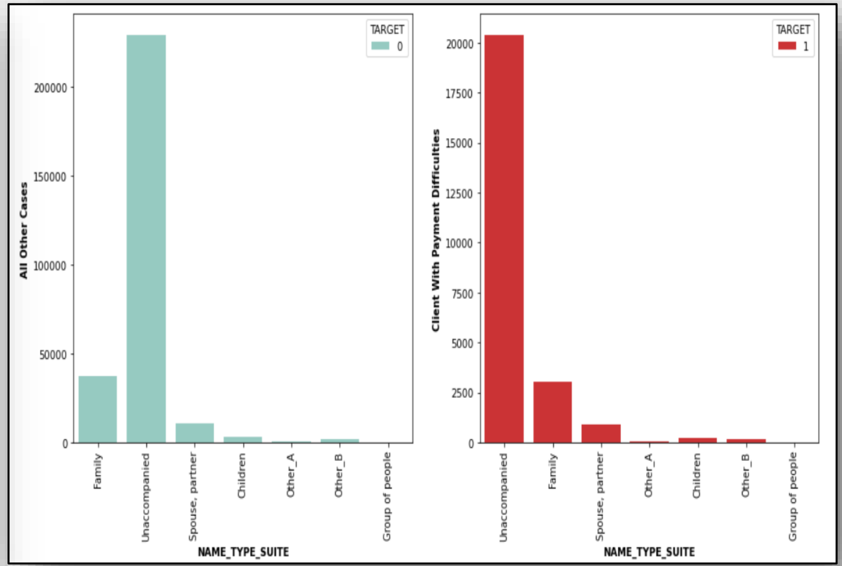
- In the target variable, **Target 0 has 91.93** percent of data whereas **Target 1 has 8.07** percent of data.
- The data imbalance is the ratio of the number of values of Target 0 to a number of values of Target 1.
- So the imbalance ratio is **11.39.**

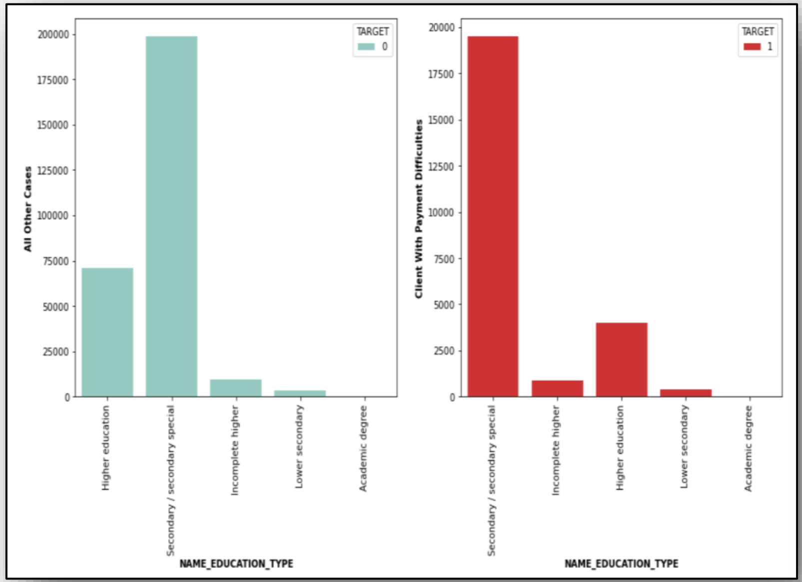


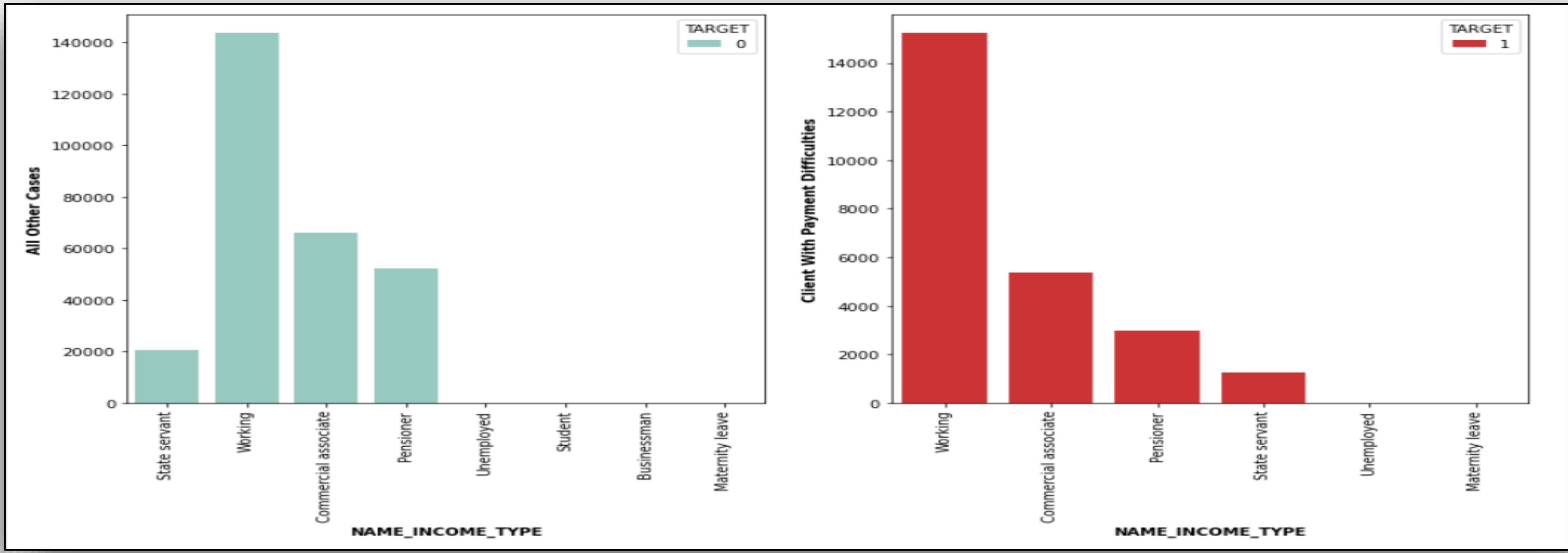
Step 6: Performing the Univariate Analysis

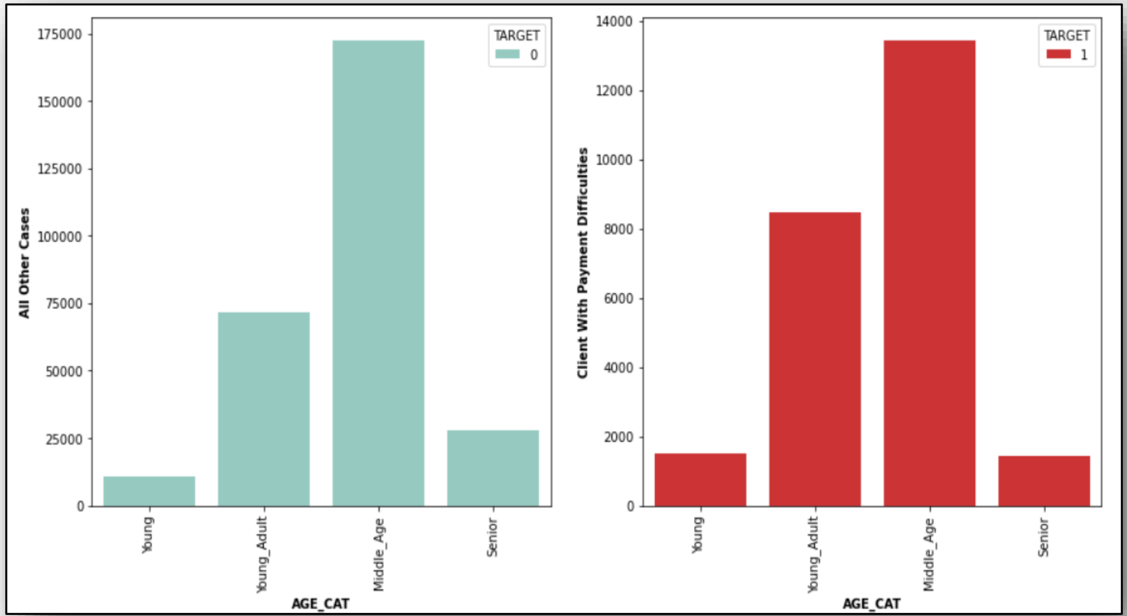


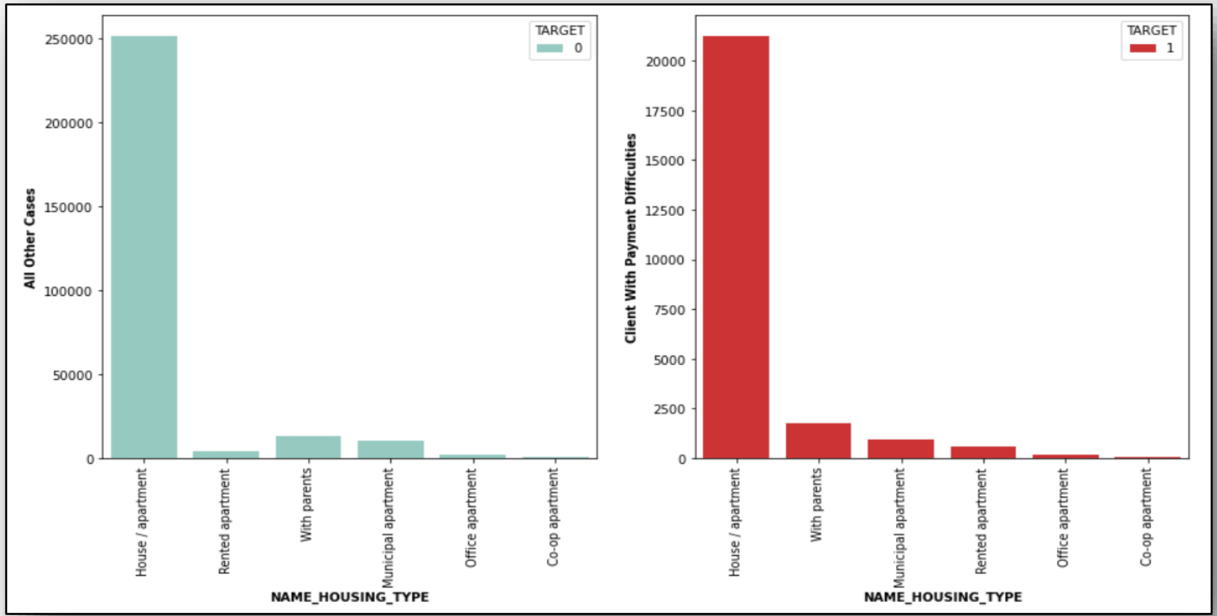


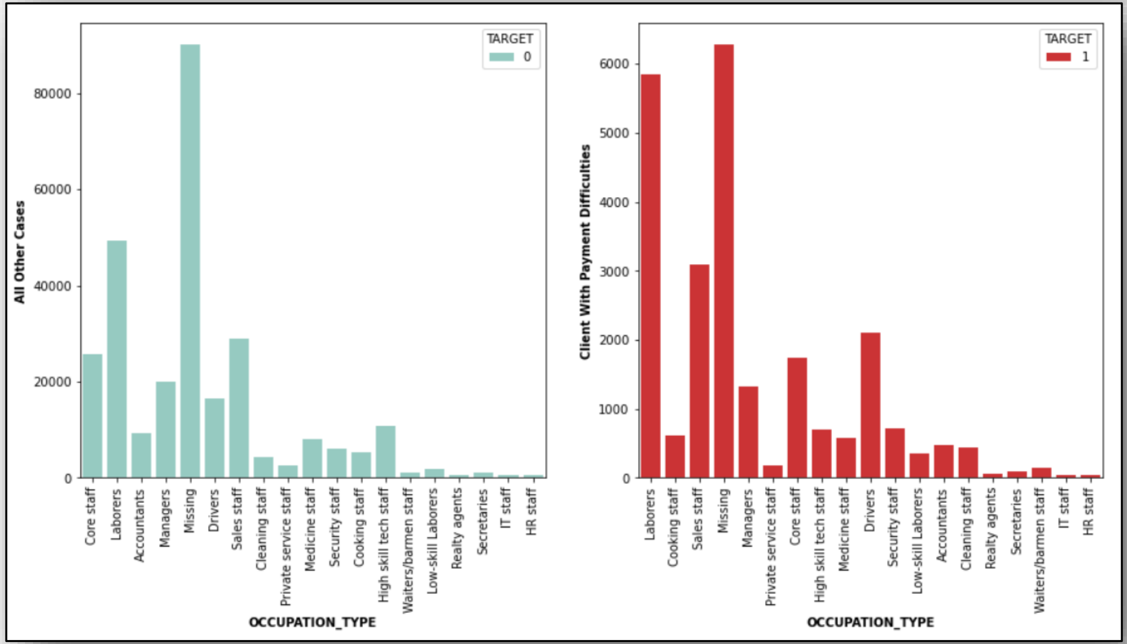


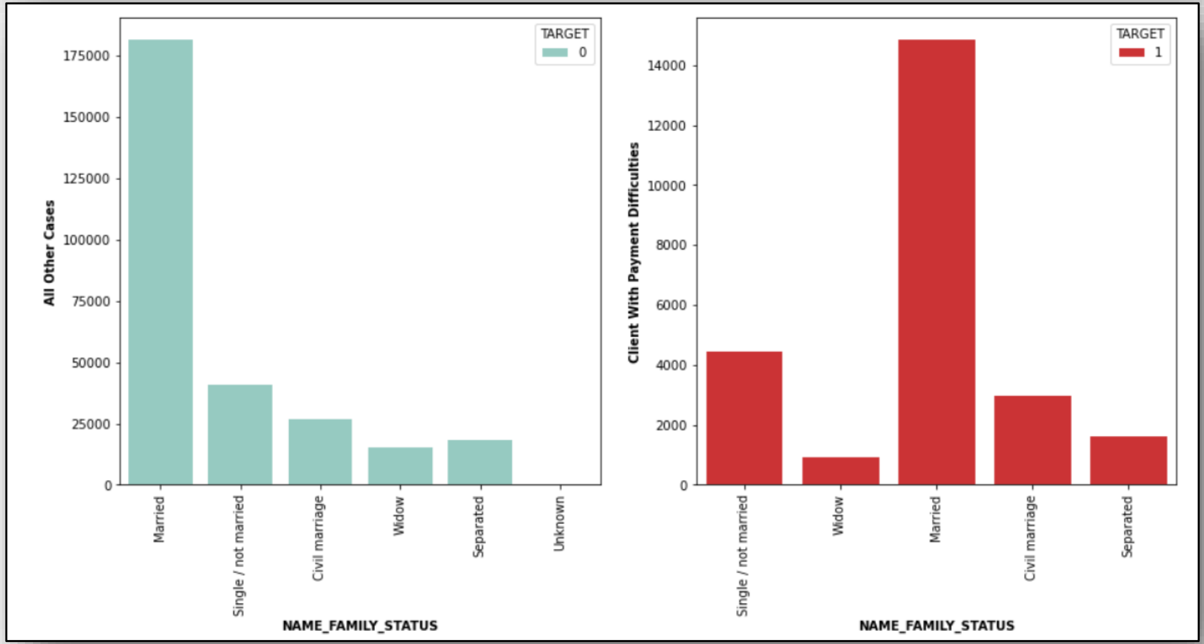





All Other Cases
Client With Payment Difficulties
TARGET
0
1
Young
Young_Adult
Middle_Age
Senior
AGE_CAT


All Other Cases
Client With Payment Difficulties
TARGET
0
1
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE


All Other Cases
Client With Payment Difficulties
TARGET
0
1
Core staff
Laborers
Accountants
Managers
Missing
Drivers
Sales staff
Cleaning staff
Private service staff
Medicine staff
Security staff
Cooking staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE


All Other Cases
Client With Payment Difficulties
TARGET
0
1
Married
Single / not married
Civil marriage
Widow
Separated
Unknown
NAME_FAMILY_STATUS

Insights for the Analysis

1. **NAME_CONTRACT_TYPE**

   For both Clients with Payment Difficulties and All Other Cases, most of the applicants have applied for Cash Loans and very less applicants have applied for Revolving Loans.

2. **CODE_GENDER**

   1. Around 65% of Females and 35% of Males do not have Payment Difficulties.
   2. Around 60 % of Females and 40% of Males have Payment Difficulties.

3. **NAME_TYPE_SUITE**

   1. Most of the applicants for both categories were unaccompanied by anyone.
   2. For those applicants who were accompanied, most of them were accompanied by their Families for both the defaulters and non-defaulters.
   3. Both the defaulters and non-defaulter have similar proportions and this does not affect the clients while applying for loans.

4. **NAME_INCOME_TYPE**

   1. Working Class was having the most difficulties in loan payment followed by Commercial Associates and Pensioner. State Servants were least likely to default.
   2. For those with no difficulties in payments, the Working class is the most who applied for loans followed by Commercial Associates and Pensioners.
   3. The Unemployed, Students, and businessmen have applied the least for loans.

5. **NAME_EDUCATION_TYPE**

   1. Client with Secondary/Secondary Special education is mostly applying for loans followed by Higher Education.
   2. Similarly, Applicants with Secondary/Secondary Special education are also most likely to have payment difficulties. Other education types have less risk of default.

6. **NAME_FAMILY_STATUS**

   1. Married employee have applied for the loan most followed by Singles and Married employees are also the most likely to default as well.
   2. Widows have the least risk of defaults as compared to others.

7. **NAME_HOUSING_TYPE**

   Most of the applicants were having a house or living in an apartment for both the defaulters as well as non-defaulters.

8. **OCCUPATION_TYPE**

   1. Laborer, Sales Staff and Core Staff applicants have applied the most for loans respectively. IT Staff and HR Staff have the least number of loan applications.
   2. Also, Laborers, Sales Staff, and Drivers are most likely to default on the payments.

9. **WEEKDAY_APPR_PROCESS_START**

   There is a fairly equal distribution of process application start days if you consider weekdays. Sunday was the day when the application processes were least started for both the defaulters and non-defaulters.

10. **AMT_INCOME_CATEGORY**

    1. Low, Very Low, and High-Income applicants have applied the most for loans.
    2. Low-income applicants are most likely to default whereas Very High-Income applicants are least likely to default.

11. **AGE_CAT**

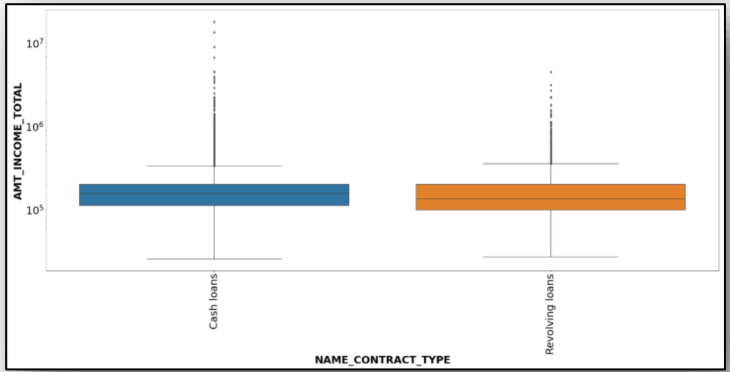
    1. Middle Age applicants have applied for the loans the most followed by Young Adults. Young applicants have the least loan applications.
    2. Middle Age applicants have the highest risk of Defaulting.

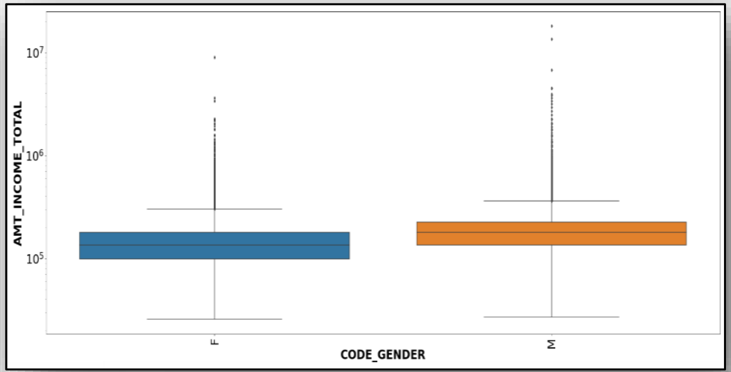
Step 7: Performing the Bivariate Analysis and Multivariate Analysis

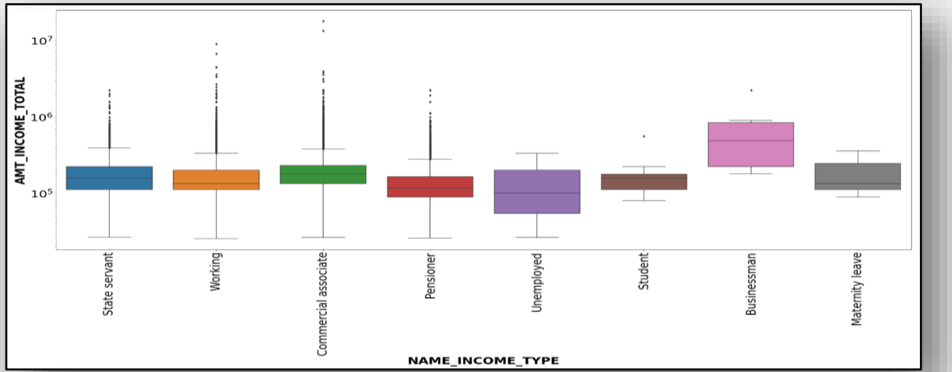
For other cases (Target 0) , the following points have been observed:

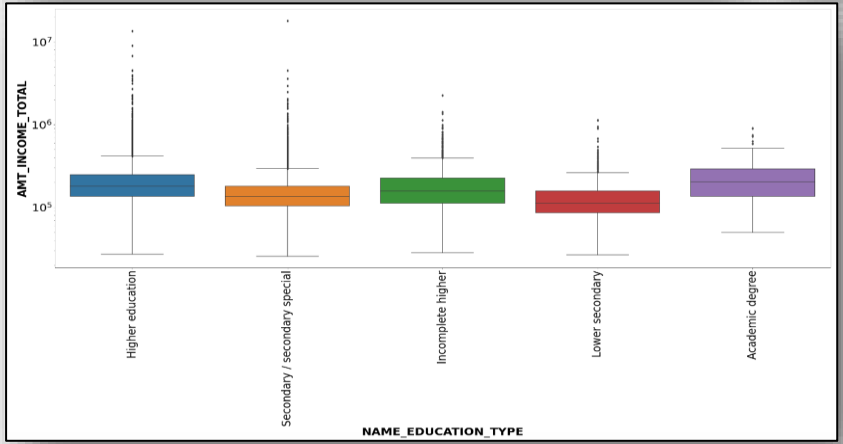
1. There is not much difference in income for those applying for revolving loans and Cash loans.
2. Males have a higher Income IQR Range as compared to females.
3. Businessmen have the highest income.
4. Those having Academic Degrees and Higher Education have more income as compared to others
5. Managers have the highest income followed by Private Service Staff, Highly Skilled Tech Staff, and IT Staff.
6. Amount Credit is higher for cash loans as compared to revolving loans.
7. There is not much difference in the amount credited for males and females.
8. Amount Credited for Businessman is on the higher side as compared to others.
9. Those having Academic Degrees and Higher Education have been credited higher amount.
10. Managers, Realty Agents, Highly skilled Tech Staff, Accountants, and HR Staff have also been credited higher amounts as compared to others
11. Cash Loans have higher Annuity Amounts as compared to revolving loans.
12. Businessmen have higher loan annuity.
13. Academic Degree and Higher Education employees have higher loan annuity amount

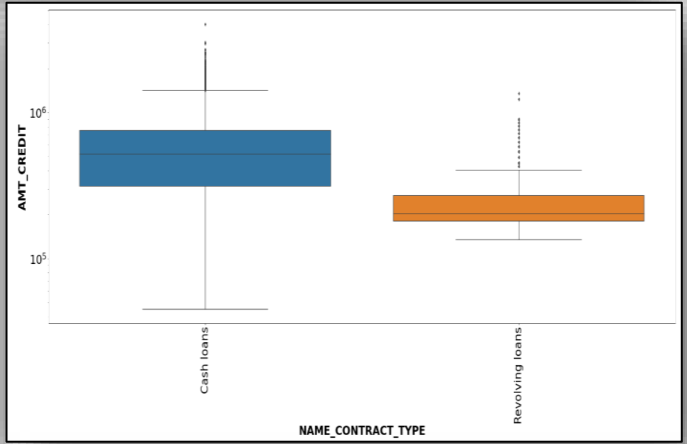
The next slide has the plots from the above analysis.

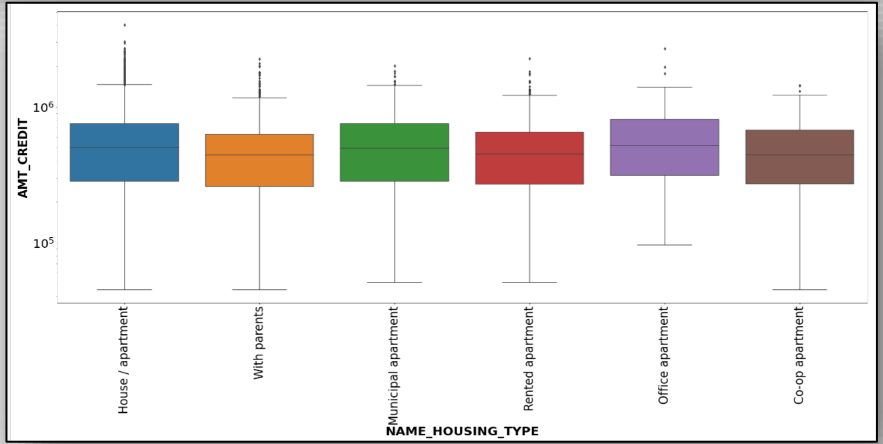

AMT_INCOME_TOTAL
10^7
10^6
10^5
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

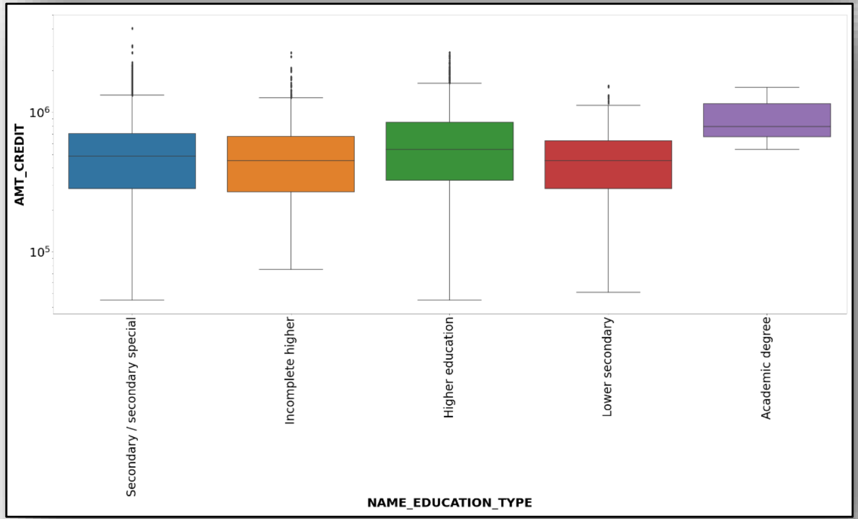

AMT_INCOME_TOTAL
10^7
10^6
10^5
F
M
CODE_GENDER

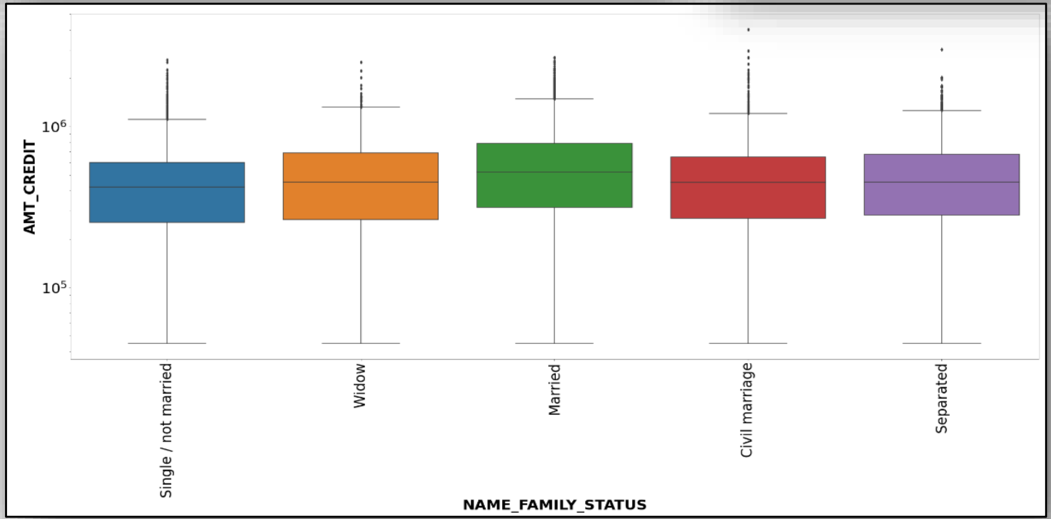

AMT_INCOME_TOTAL
10^7
10^6
10^5
State servant
Working
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE


AMT_INCOME_TOTAL
10^7
10^6
10^5
Higher education
Secondary / secondary special
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE


AMT_CREDIT
10^6
10^5
Cash loans
Revolving loans
NAME_CONTRACT_TYPE


AMT_CREDIT
10^6
10^5
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE


AMT_CREDIT
10^6
10^5
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

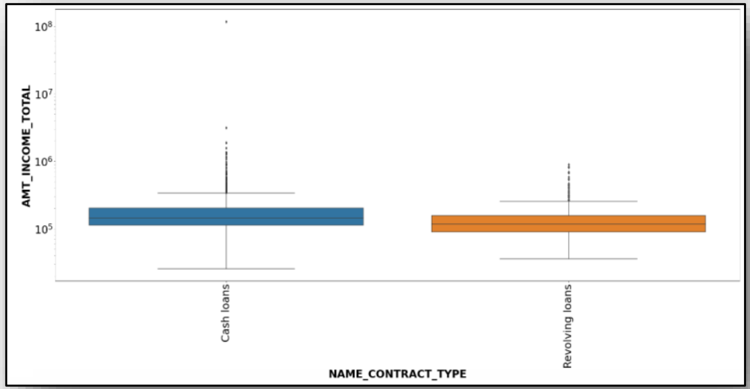

AMT_CREDIT
10^6
10^5
Single / not married
Widow
Married
Civil marriage
Separated
NAME_FAMILY_STATUS

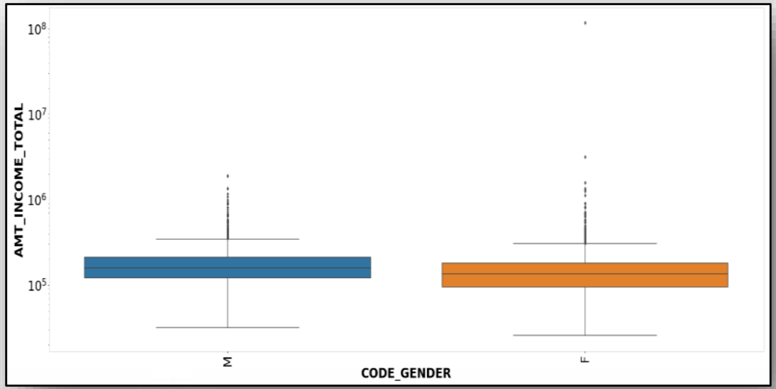
Step 7: Performing the Bivariate Analysis and Multivariate Analysis

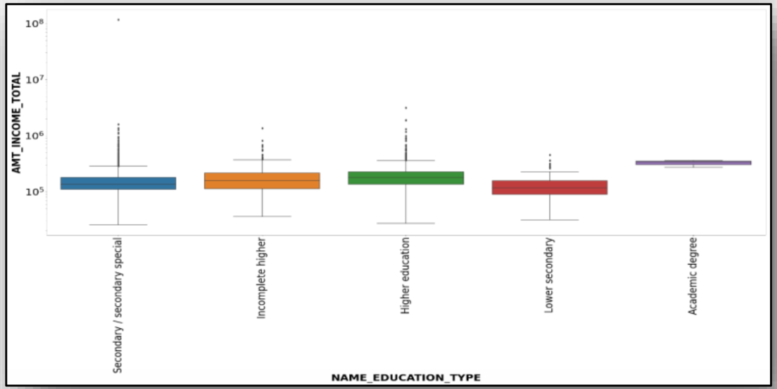
For those having payment difficulties(Target 1), the following points have been observed:

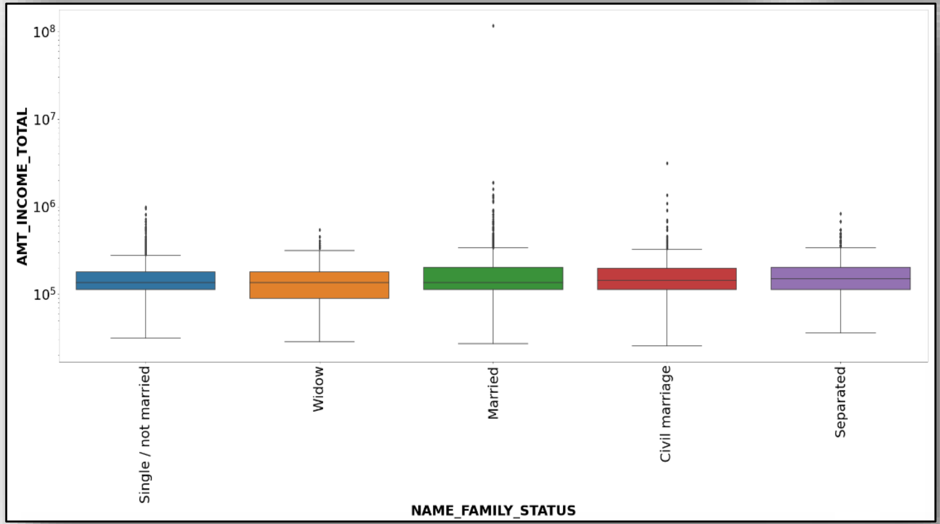
1. Client having lower income has applied for revolving loans whereas higher income clients have applied for cash loans.
2. Females have slightly lesser income as compared to males.
3. Widowers have most of the income below the median. Married applicants have income on the higher side of the median.
4. Academic Degree Holders have the highest income followed by clients having Higher Education.
5. Managers have the Highest Income whereas cooking staff, waiters, and cleaning staff have the lowest income.
6. Credit Amount and Annuity Amount is very low for revolving loans as compared to Cash Loans.
7. There is not much difference in the Credit Amount and the Annuity Amount for males and females.
8. Managers and Accountants have the highest credit amount as compared to others.
9. Academic Degree Holders have the highest credit amount as compared to other education types.

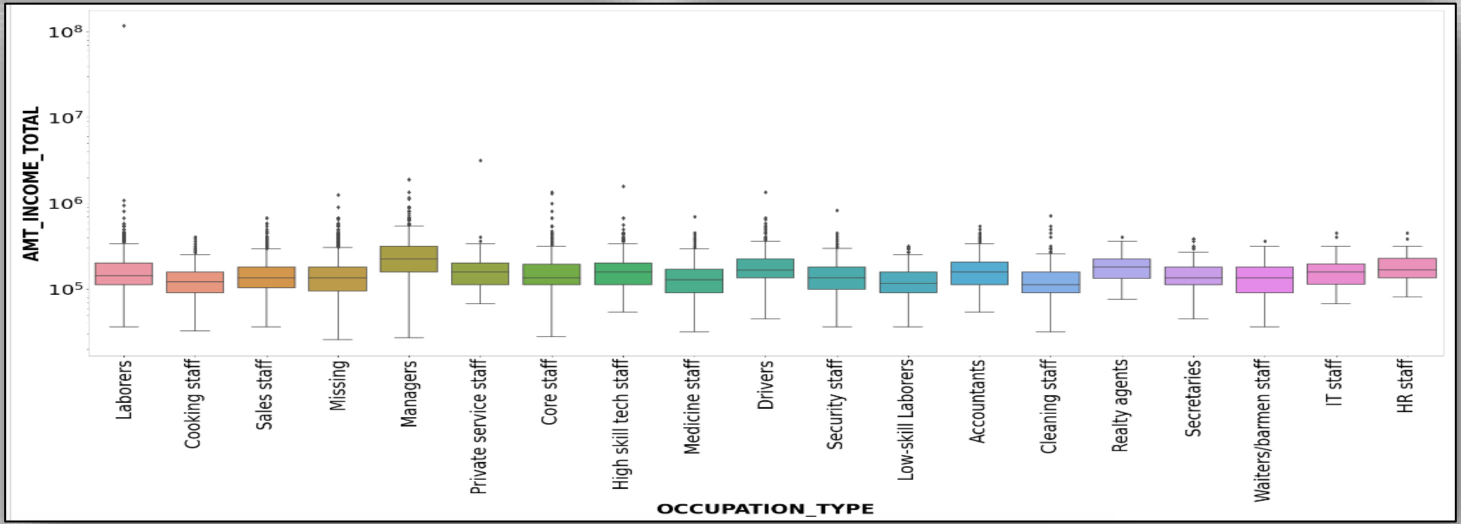
The next slide has the plots from the above analysis.

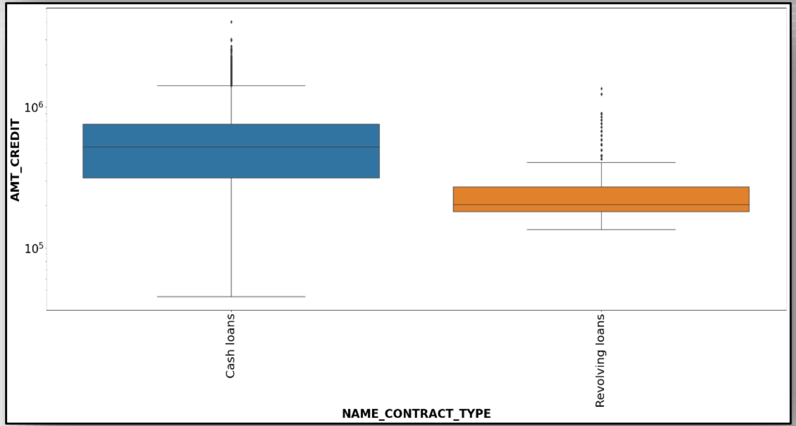

AMT_INCOME_TOTAL
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

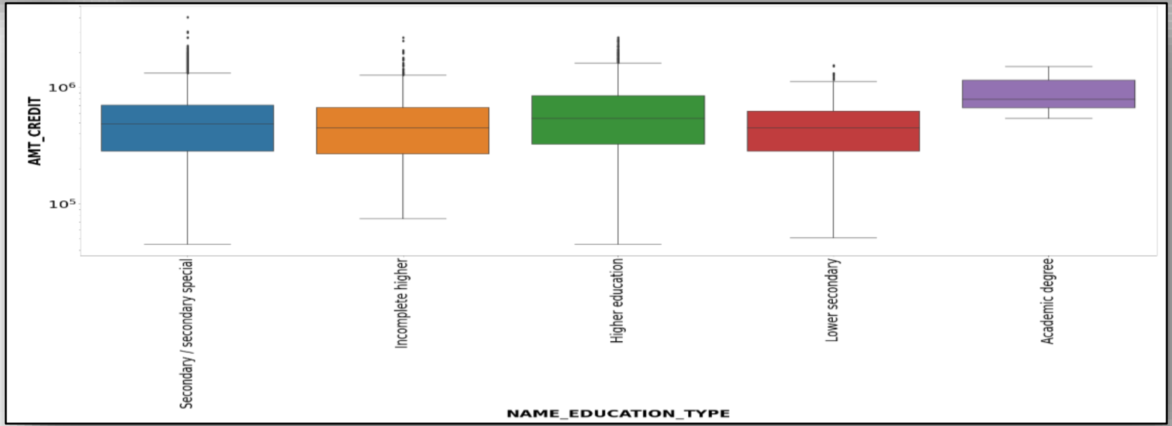

AMT_INCOME_TOTAL
M
F
CODE_GENDER


AMT_INCOME_TOTAL
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

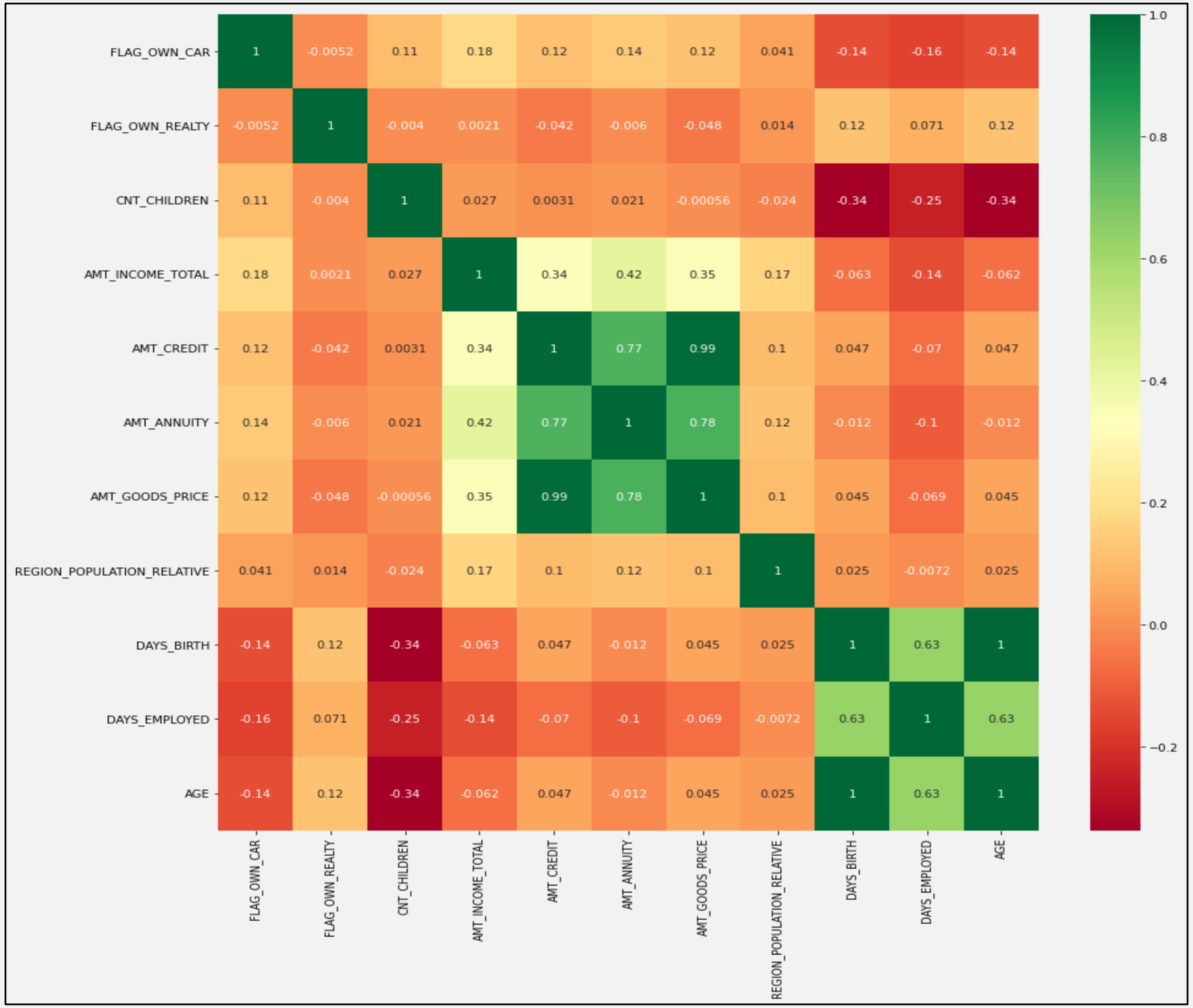

AMT_INCOME_TOTAL
Single / not married
Widow
Married
Civil marriage
Separated
NAME_FAMILY_STATUS


AMT_INCOME_TOTAL
Laborers
Cooking staff
Sales staff
Missing
Managers
Private service staff
Core staff
High skill tech staff
Medicine staff
Drivers
Security staff
Low-skill Laborers
Accountants
Cleaning staff
Realty agents
Secretaries
Waiters/barmen staff
IT staff
HR staff
OCCUPATION_TYPE


AMT_CREDIT
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

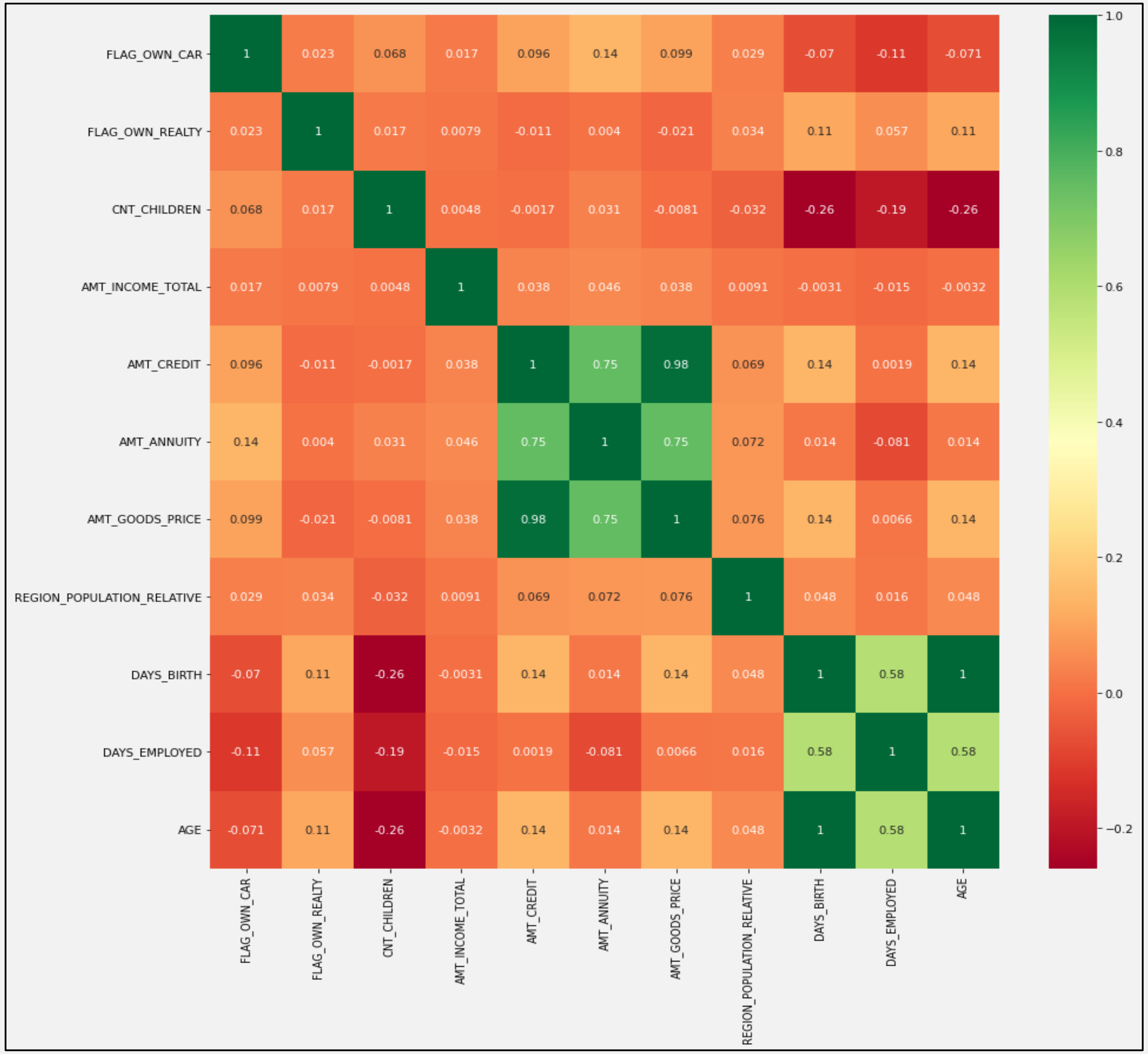

AMT_CREDIT
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

Heatmap for Target0



1.AMT_CREDIT and AMT_GOODS_PRICE have a very high positive correlation which is expected as the credit amount depends on the goods for which the loan is given.
2.AMT_CREDIT also has a high correlation with AMT_ANNUITY which indicates that the annuity amount increase with the credit amount.
3.AMT_CREDIT has a weak positive correlation with the AMT_INCOME_TOTAL.
4.AGE has a high positive correlation with the day's employees which shows that older clients have been employed for a higher amount of time.

Heatmap for Target1



1.AMT_CREDIT and AMT_GOODS_PRICE have a very high positive correlation which is expected as the credit amount depends on the goods for which the loan is given.
2.AMT_CREDIT also has a high correlation with AMT_ANNUITY which indicates that the annuity amount increase with the credit amount.
3.AMT_CREDIT has no correlation with the AMT_INCOME_TOTAL.
4.AGE has a medium positive correlation with the day's employees which shows that older clients have been employed for a higher amount of time.

ANALYSIS OF PREVIOUS APPLICATION DATA

Step 1: Reading the Data: The Previous Application data is stored in data2 data frame and it initially has 16,70,214 rows and 37 columns.

Step 2: Removing the unwanted columns: Removing the columns having a number of null values count more than 40% of the total value.

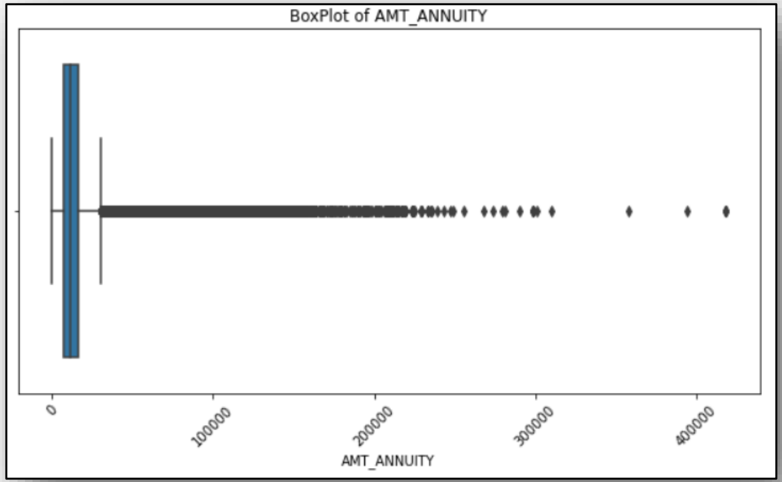
There are 11 columns having null values more than 40% of the total values. So after these columns, the data frame has 26 columns.

Step 3: Imputing the Null Values:

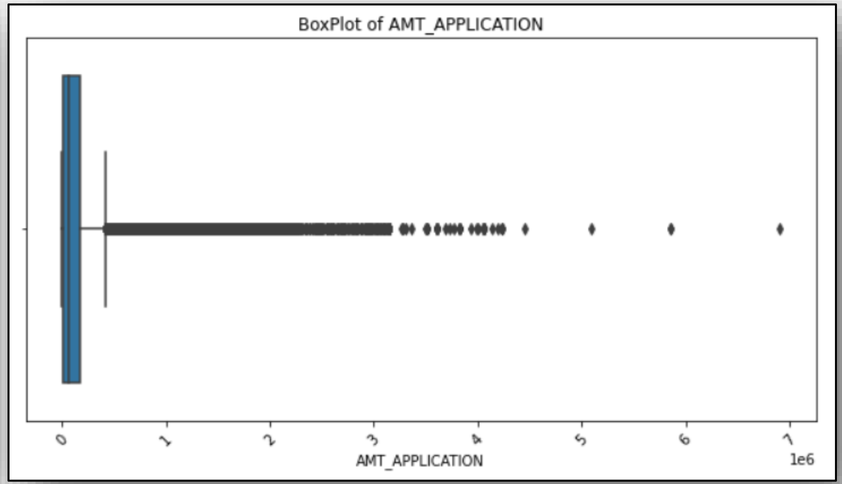
1. The AMT_GOODS_PRICE, AMT_ANNUITY, CNT_PAYMENT, AMT_CREDIT and PRODUCT_COMBINATION have null values.
2. Replacing the null values in AMT_GOODS_PRICE, AMT_ANNUITY, CNT_PAYMENT, and AMT_CREDIT with the median values of the variables and null values of PRODUCT_COMBINATION by the mode value of the variable as it is a categorical column.
3. It is also observed that 10 columns had XNA values which are replaced by np.nan and the count of null values is checked again. Six columns have null values greater than 30% and have been dropped from the data frame.
4. Remaining four columns i.e. NAME_PORTFOLIO, CODE_REJECT_REASON, NAME_CLIENT_TYPE, AND NAME_CONTRACT_TYPE having null values are replaced by the mode of the column as all the four variables are categorical variables.
5. Now the data frame has 20 columns.

Step 4: Outliers Identification:

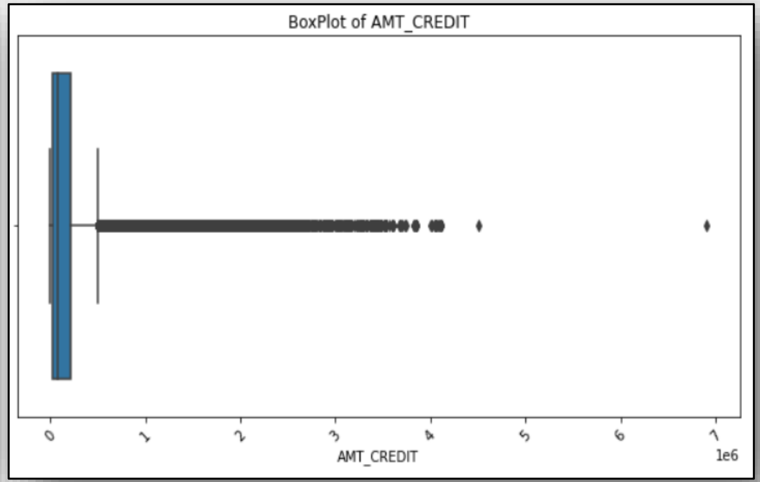
1. AMT_ANNUITY variable has a large number of outliers as many values are present outside the upper fence.



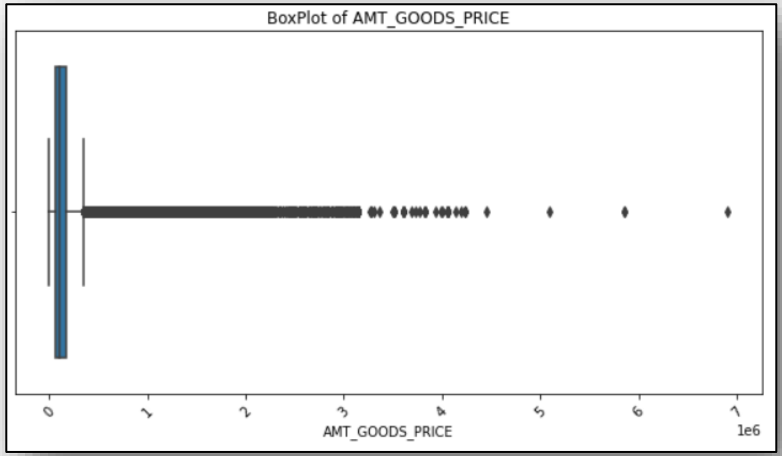
2. AMT_APPLICATION variable also has a large number of outliers as a large number of values are outside the upper fence.



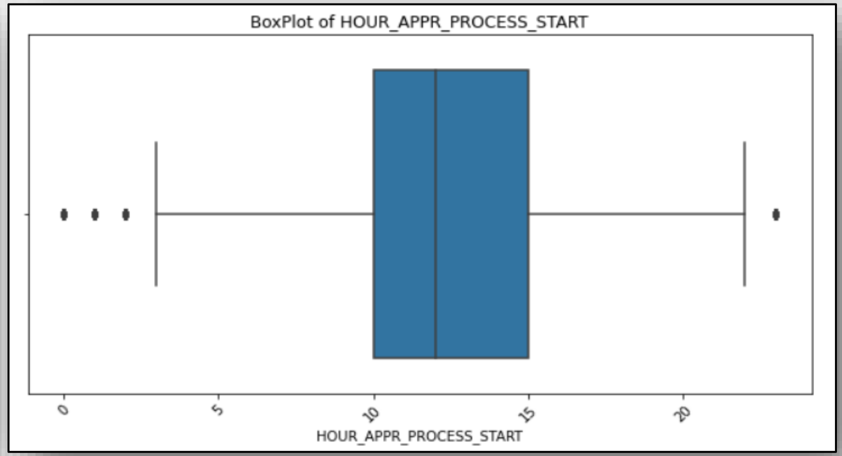
3. AMT_CREDIT variable contains many outliers as many data points are outside the upper fence.



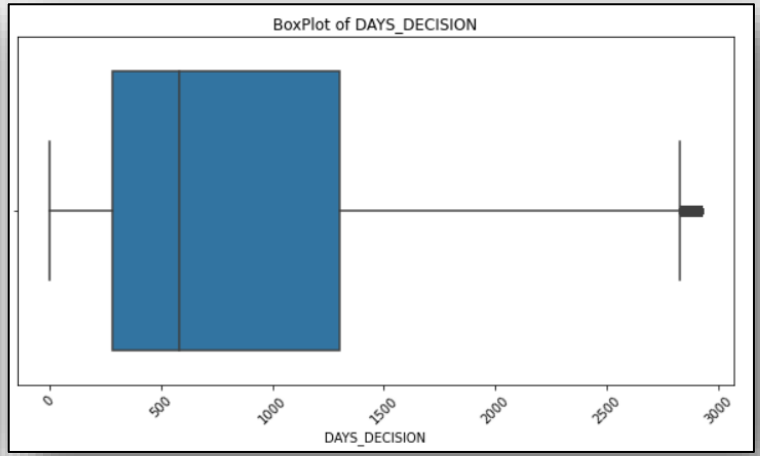
4. AMT_GOODS_PRICE variable also has large number of outliers as many values are outside of the upper fence.



5. HOUR_AOORR_PROCESS_START variable has very less values lying outside the lower and upper fence. Hence it has very few outliers.

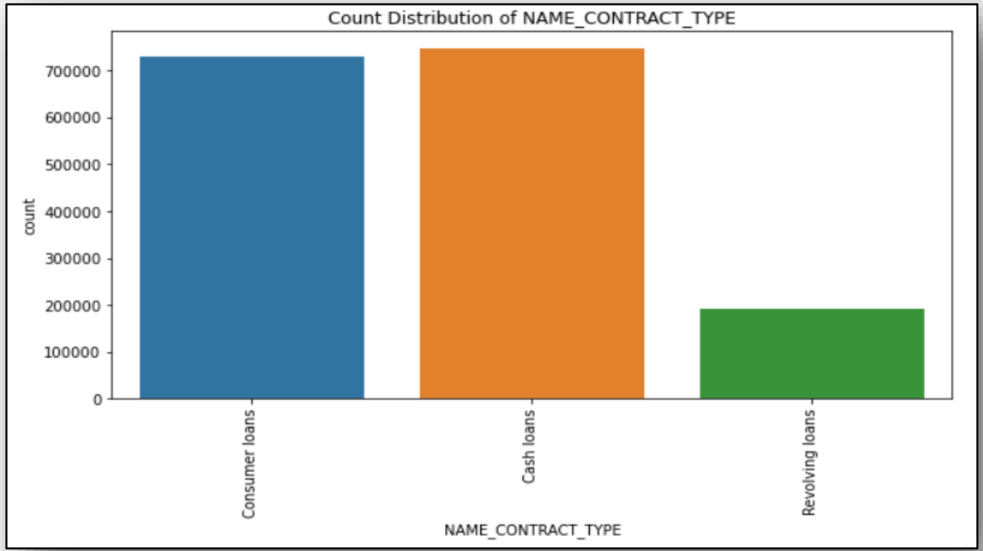


6. DAYS_DECISION variable has a few outliers outside the upper fence.

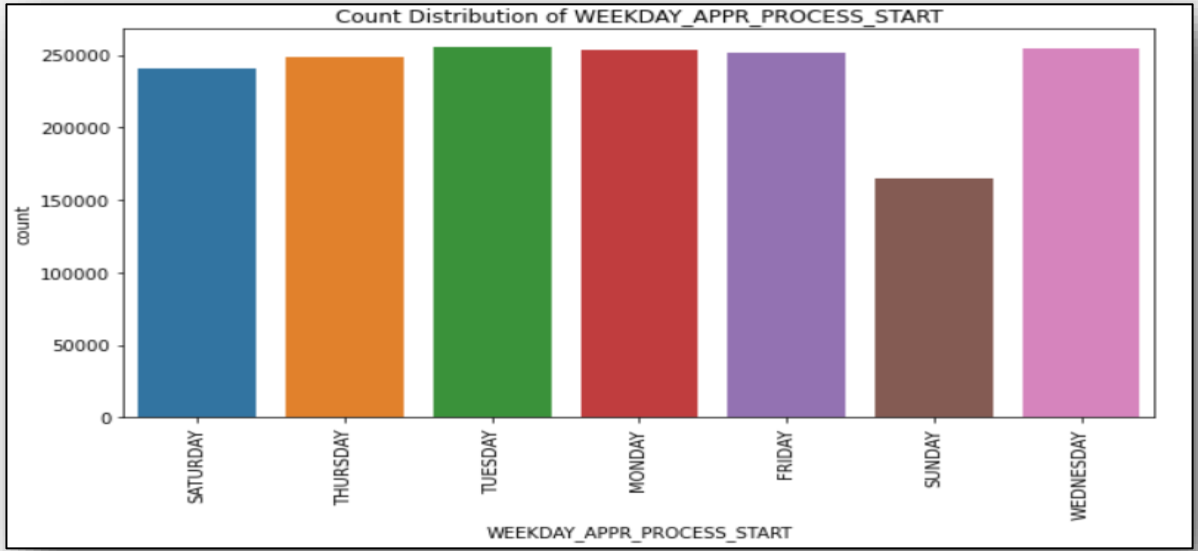


Step 5: Univariate Analysis Insights:

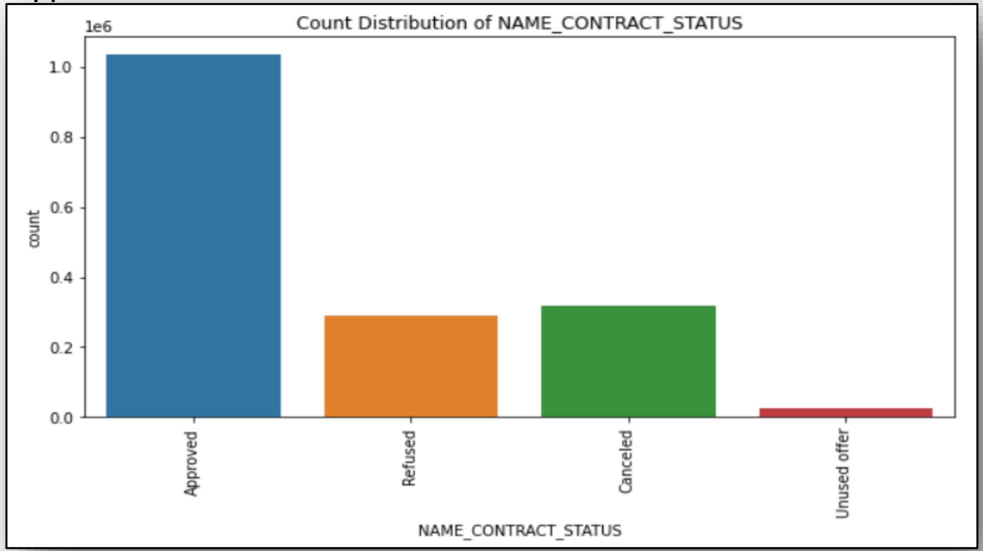
1. Most of the contract types were Consumer Loans and Cash Loans and their number of applications was almost the same. Revolving Loans received the least applications.



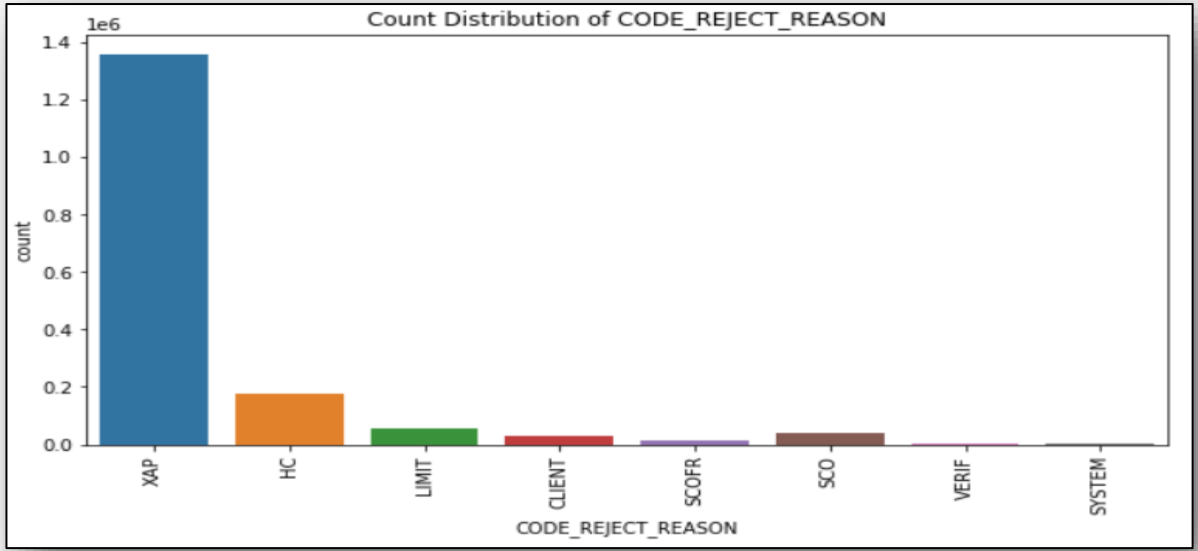
2. There was not much difference between the days on which the application process started. Although, on Sunday, there were least applications applied.



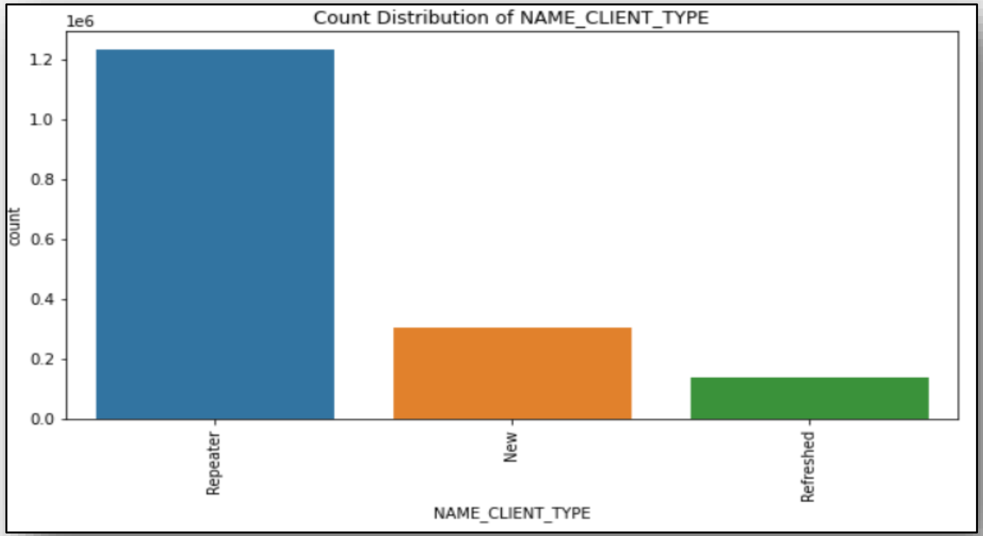
3. Most of the contracts were approved whereas very less applications left as Unused offer.



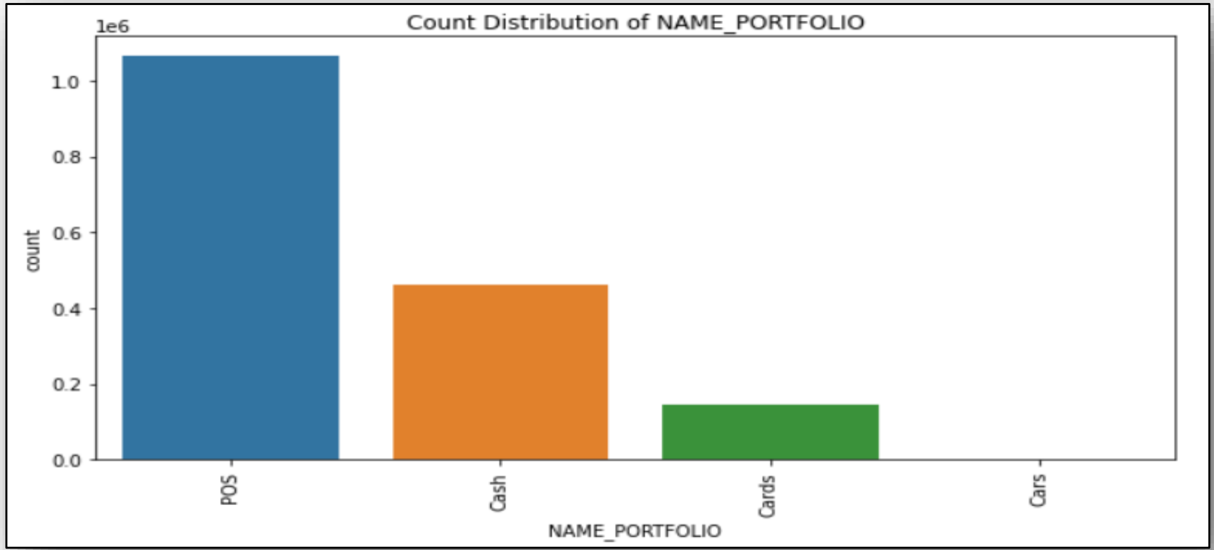
4. XAP was the code rejection reason for very high number of applications which were rejected with HC being the distant second rejection reason.



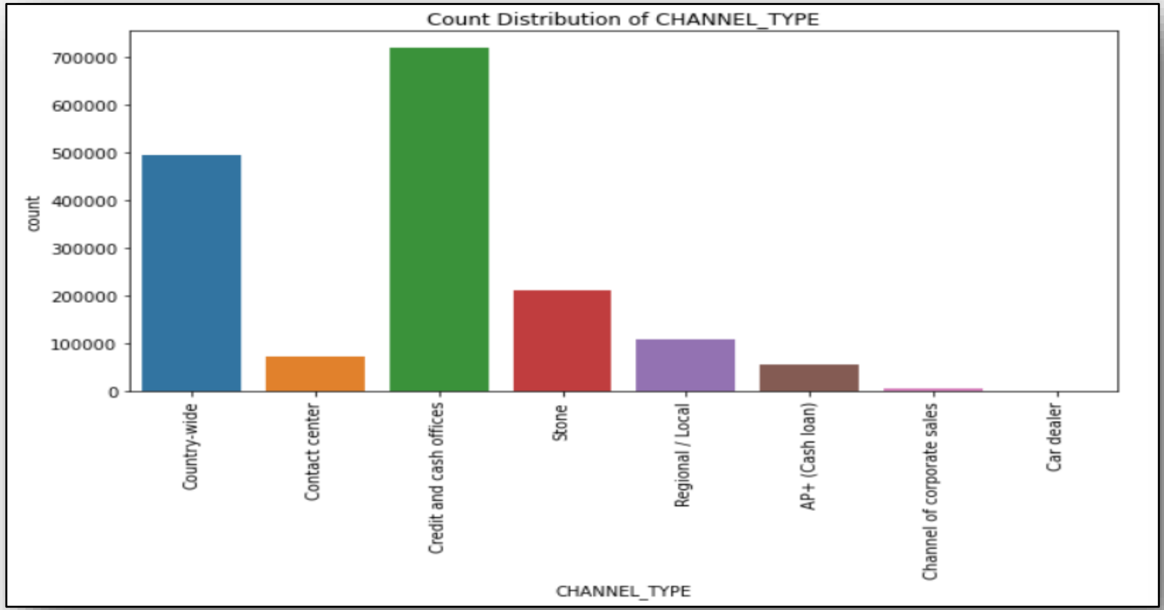
5. Very large number of applications were raised by repeaters whereas Refreshed clients raised the least number of applications.



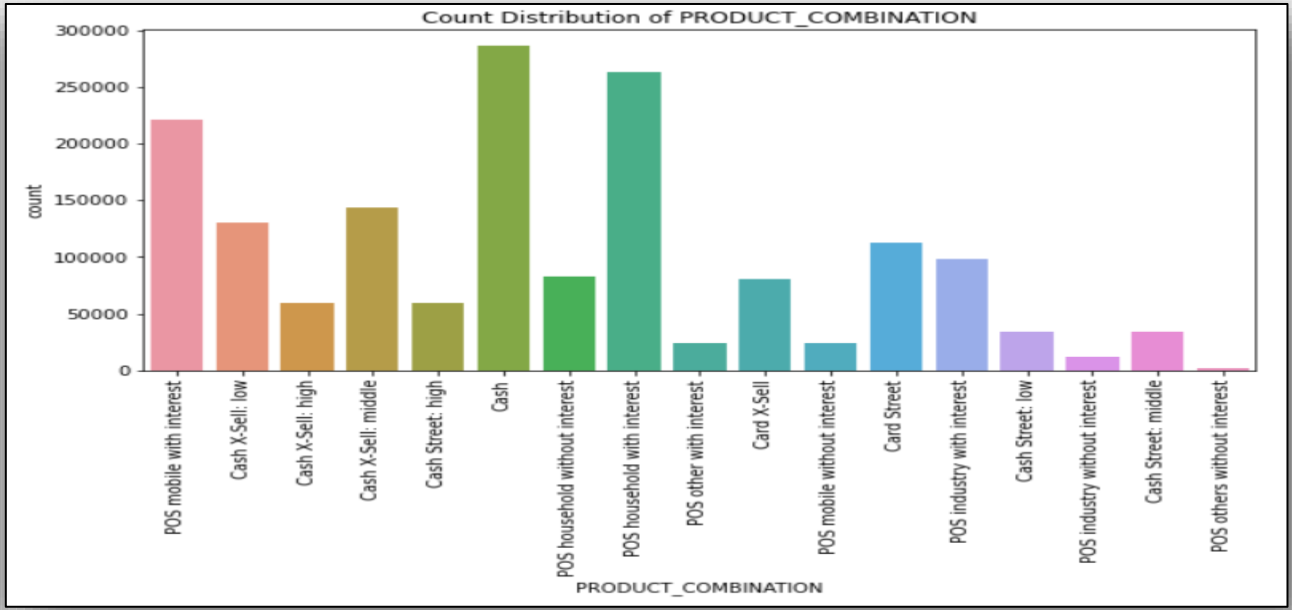
6. Most of the previous applications were for POS whereas for CARS, least amount of applications were received.



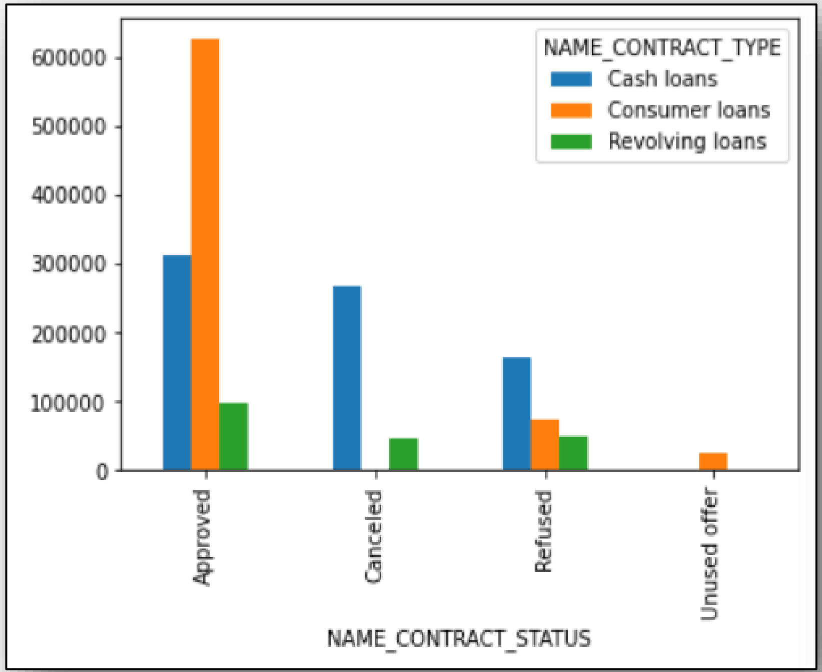
7. Most of the clients were acquired through Credit and Cash Offices followed by Country Wide and Stone.

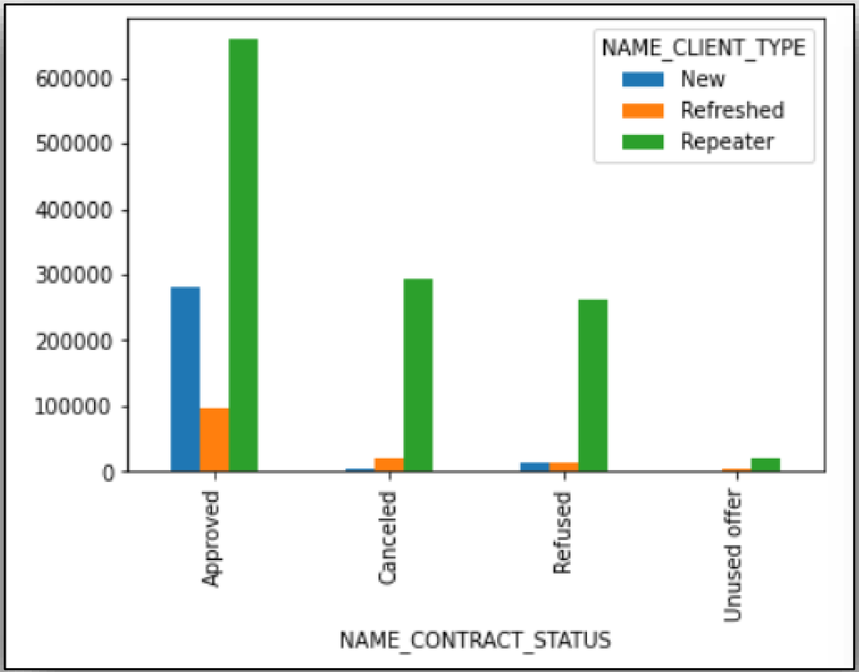


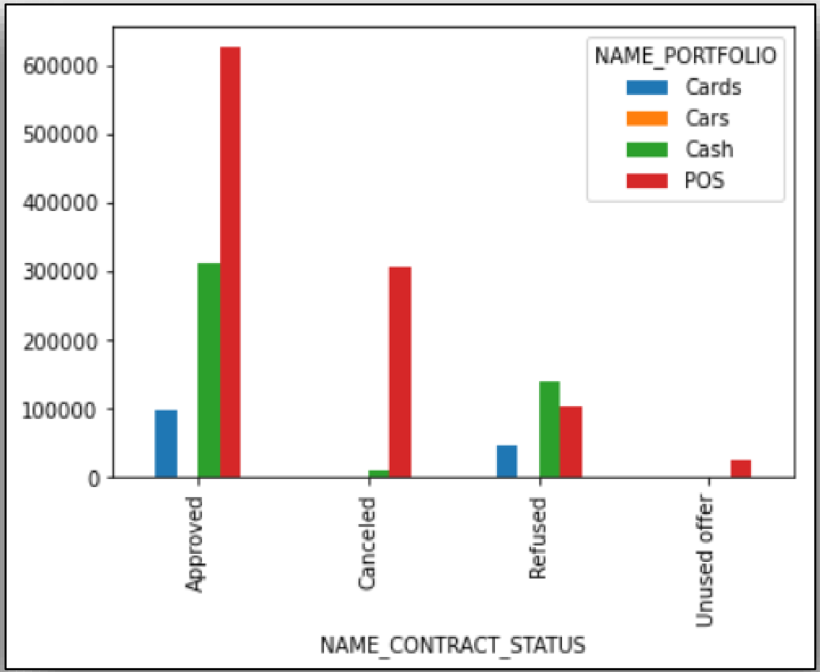
8. Cash, POS mobile with interest, and POS household with interest are the most popular product combinations.



Step 5: Bi-Variate Analysis Insights:

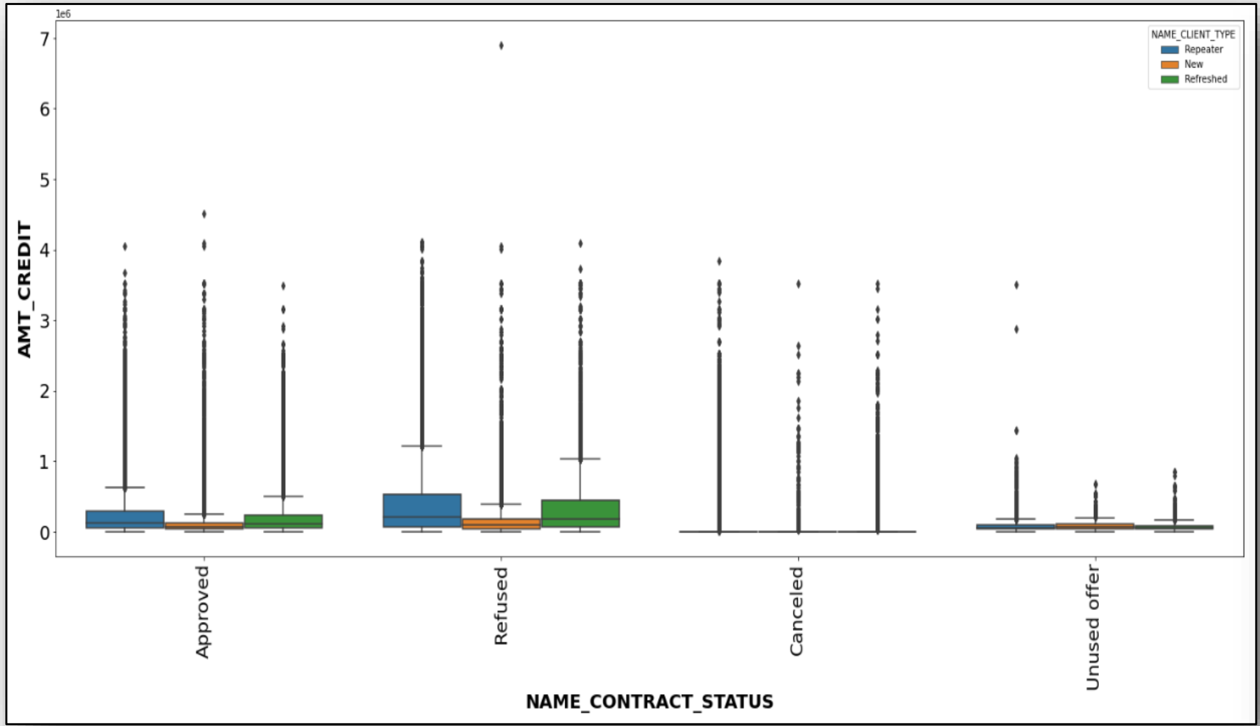


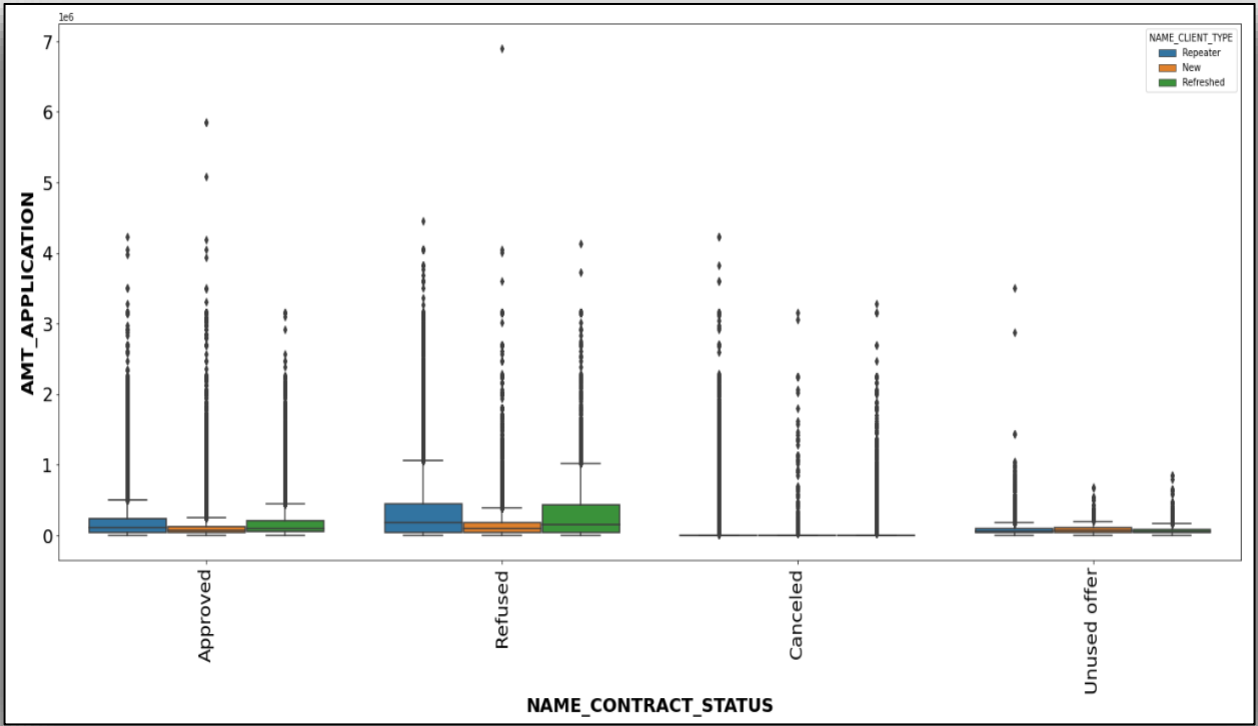




1. Most of the approved loans were Consumer loans whereas most of the Refused loans are Cash loans. Also, most of the Canceled loans are Cash loans as well.
2. Repeaters have applied the most for loans. Although, Repeaters client's loans were approved the most, followed by New clients and Refreshed client's loans were approved the least. Repeater clients also have a higher rate of Refusal as well as Cancellation, although this may be due to the fact that repeaters have applied for the loan the most as well.
3. POS loans were the most approved type of loan followed by Cash and Cards. Cash loans have the highest rate of Refusal.

Step 5: Bi-Variate Analysis Insights:

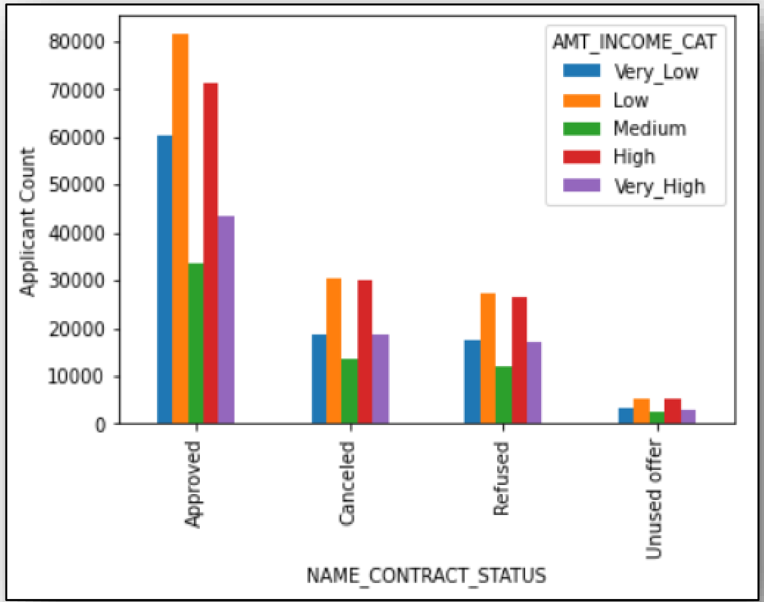




From the above plots, we can see that for the applications which were refused, the credit amount, as well as the application amount, were on the higher side as compared to the approved applications.
Also, all of the CONTRACT_STATUS variables have outliers as well.

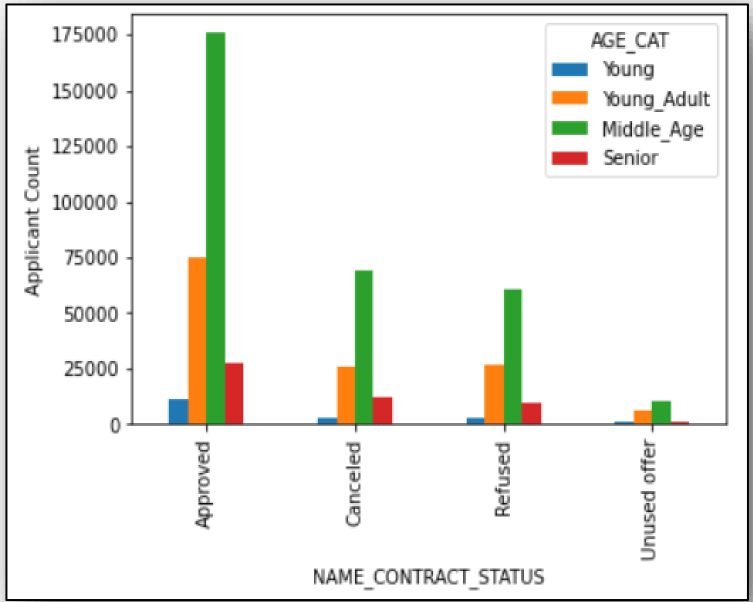
ANALYSIS OF MERGED DATA

The Application data and the Previous application data are merged in a new data frame called as data3 using an inner join. The merged data has 14,13,701 rows and 72 columns.



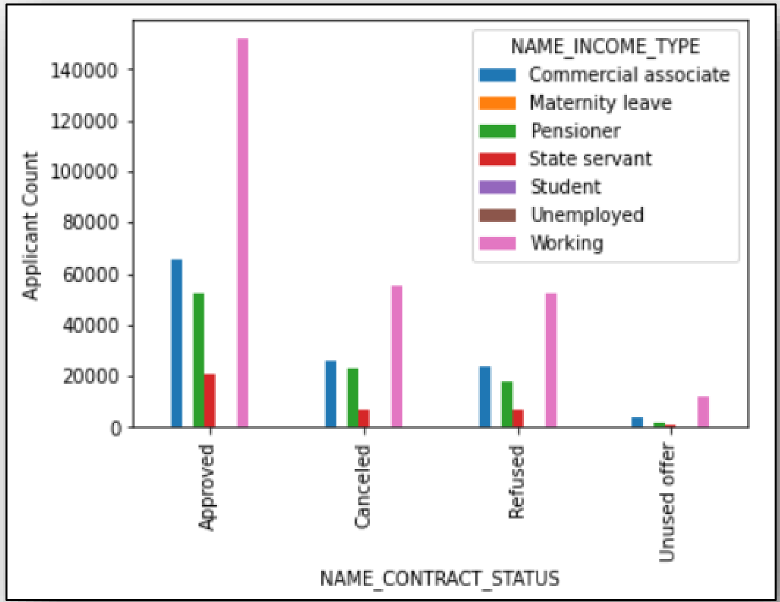
Most loan applications are applied by Low Income and High-Income applicants.

Also, the approval, refusal and cancellation rate are higher for Low and High Income applicants.



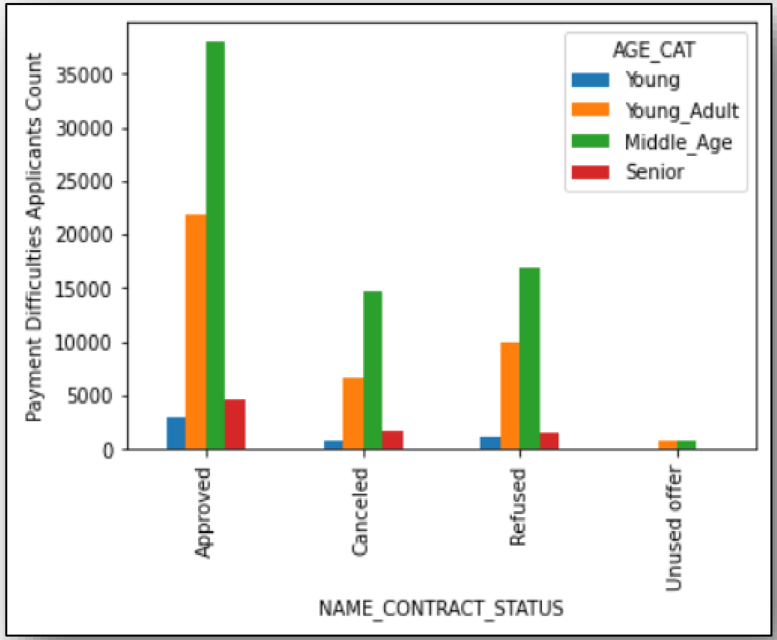
Most of the loan applications are applied by the Middle Age clients followed by Young Adults.

Middle Age clients also have the highest approval, cancellation or refusal rate.



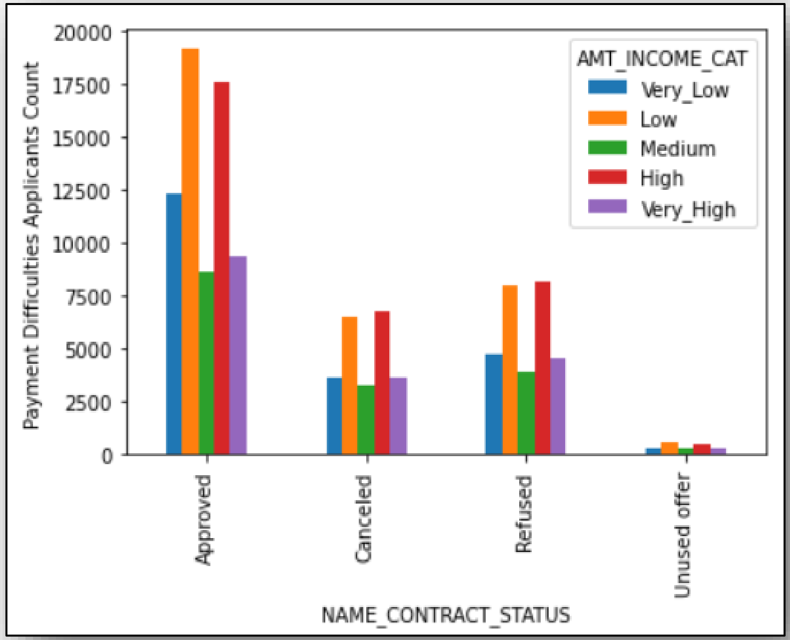
Working Employees have applied for loans the most followed by Commercial Associates.

Similarly, the Approval and Refusal rate is also the highest for working employees as well which may be due to the fact that they apply for loans the most.

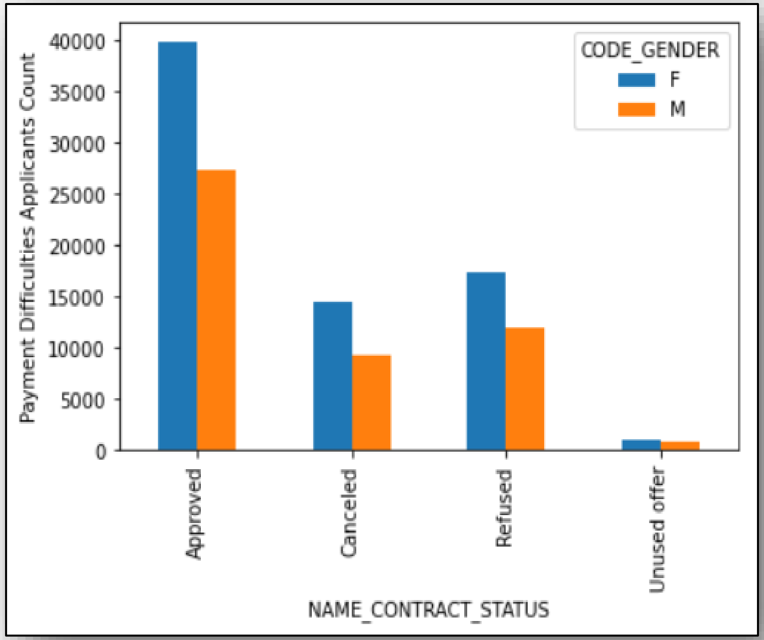


For all the categories, the middle age clients have higher difficulties in payment as compared to other age groups.

This is also because the loan applications were applied by the middle age clients the most.



Low-Income Category applicants have a higher rate of Payment difficulties as compared to other income categories.



Females have higher approval, cancellations, and Refusal rates as compared to males for those applicants which are having difficulties with payment.

CONCLUSION

- Banks should provide loans to Commercial Associates as they have defaulted less and should try to focus on how to make sure that Working-Class clients do not default as much as they are currently defaulting.
- Those having a Secondary education are also defaulting on payments the most as compared to other education types. Also, banks should try to onboard more clients having Higher Education for loans as they have defaulted less.
- Banks should also focus on clients which are not married to provide loans as they have defaulted less. Married clients have payment difficulties more as compared to other clients.
- Laborer's, Sales Staff, and Driver have defaulted the most as compared to the other Occupation types.
- Low Income Category clients have higher rate of payment difficulties as compared to other income types.
- Most of the previous applications which were refused, the application amount and the credit amount was on a higher side as compared to the previous applications which were approved.
- Most of the previous loans that were approved were Consumer Loans of POS type.
- Most of the previous loans that were refused were of Cash type.