EDA Assignment made by

- *Pooja Jantwal*

Introduction

This assignment aims to give an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that we have learned in the EDA module, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

The loan-providing companies find it hard to give loans to 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 that specializes in lending various types of loans to urban customers. You have to use EDA to analyze 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 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):

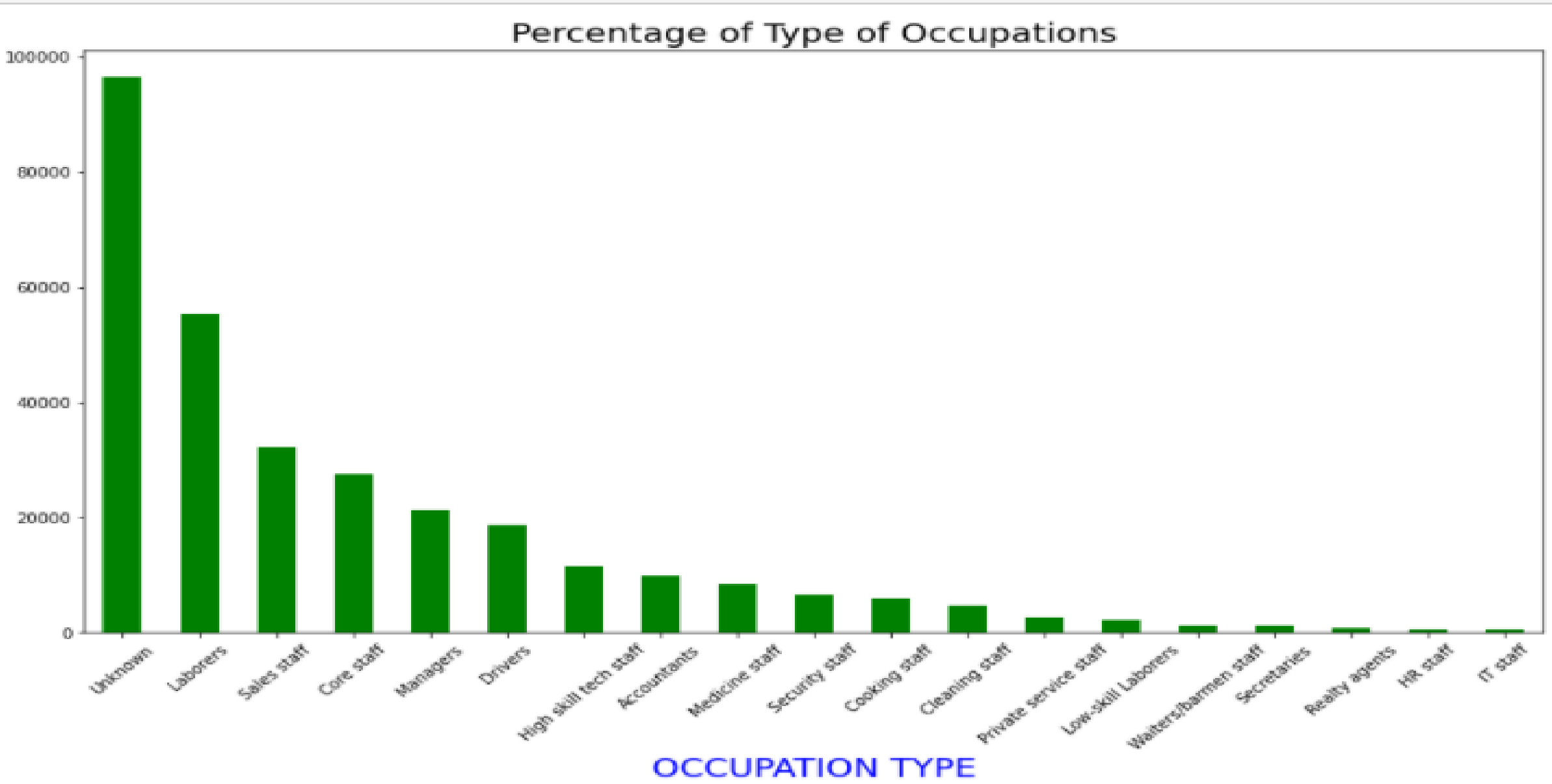
- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** The loan has been cancelled by the client but on different stages of the process.

This dataset has 3 files as explained below:

- *1. 'application_data.csv'* contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties.**
- *2. 'previous_application.csv'* contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused, or Unused offer.**
- *3. 'columns_description.csv'* is a data dictionary that describes the meaning of the variables.

Application Data

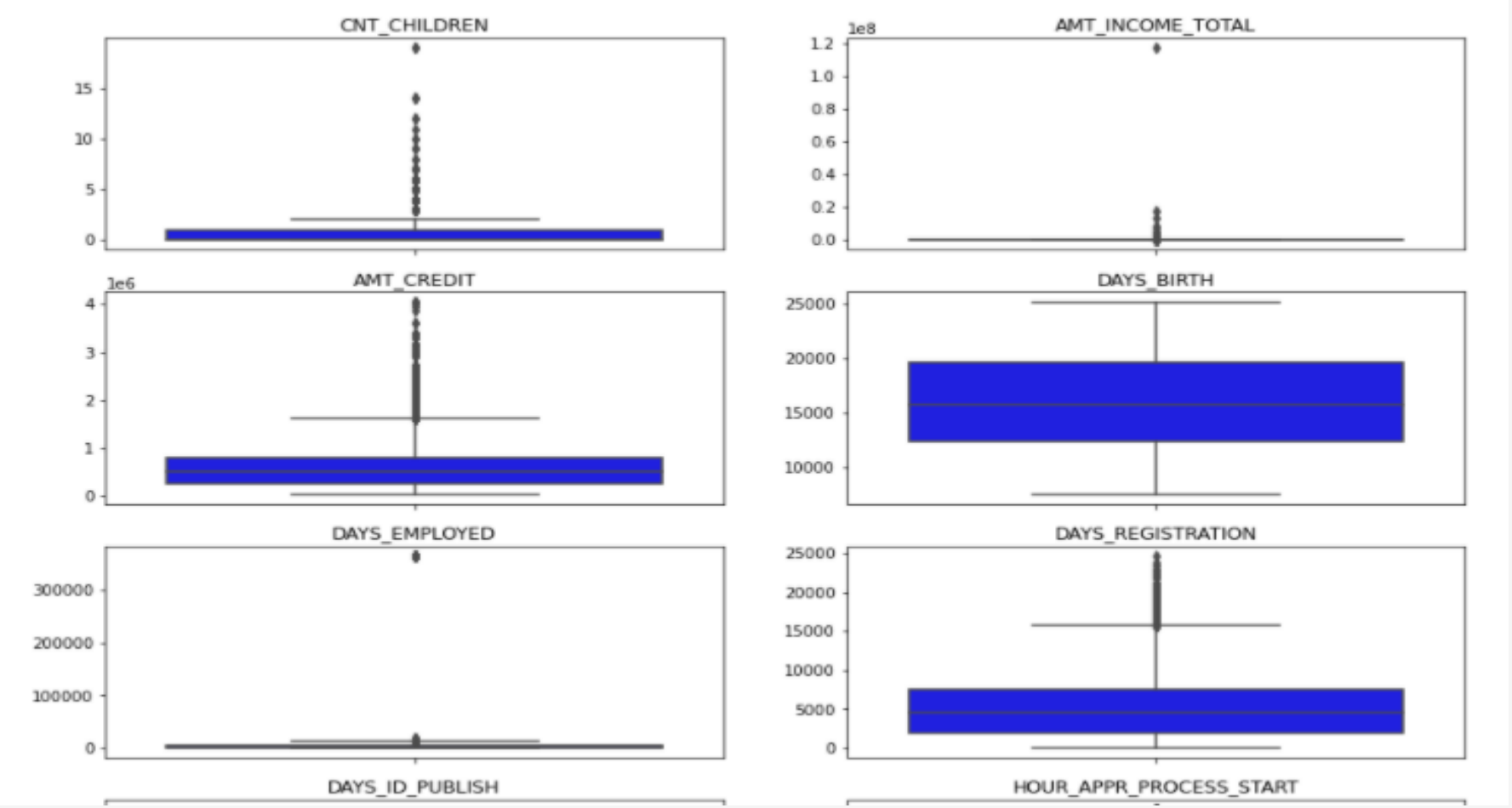
- Application data has 307511 rows and 122 columns.
- After checking for missing values, we took 40% as a threshold because greater than 40% data is not much relevant. However, data checking for greater than 13 % has relevant then we impute the value as per the requirement i.e. Mode, Median and Mean.
- Application data after fixing missing values and imputing the values has 307511 rows and 73 columns.
- OCCUPATION_TYPE has relevance but has missing values, we impute that with “Unkown” values. This is a MNAR situation - Missing Not At Random , Unknown occupation is showing relation with "Name_Contract_Type"- which means *Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application*, Hence applicant felt not feeling like to fill occupation column while taking cash loan - example instant loan, personal loan etc. which bank provides very quick and faster with minimal documentation.**

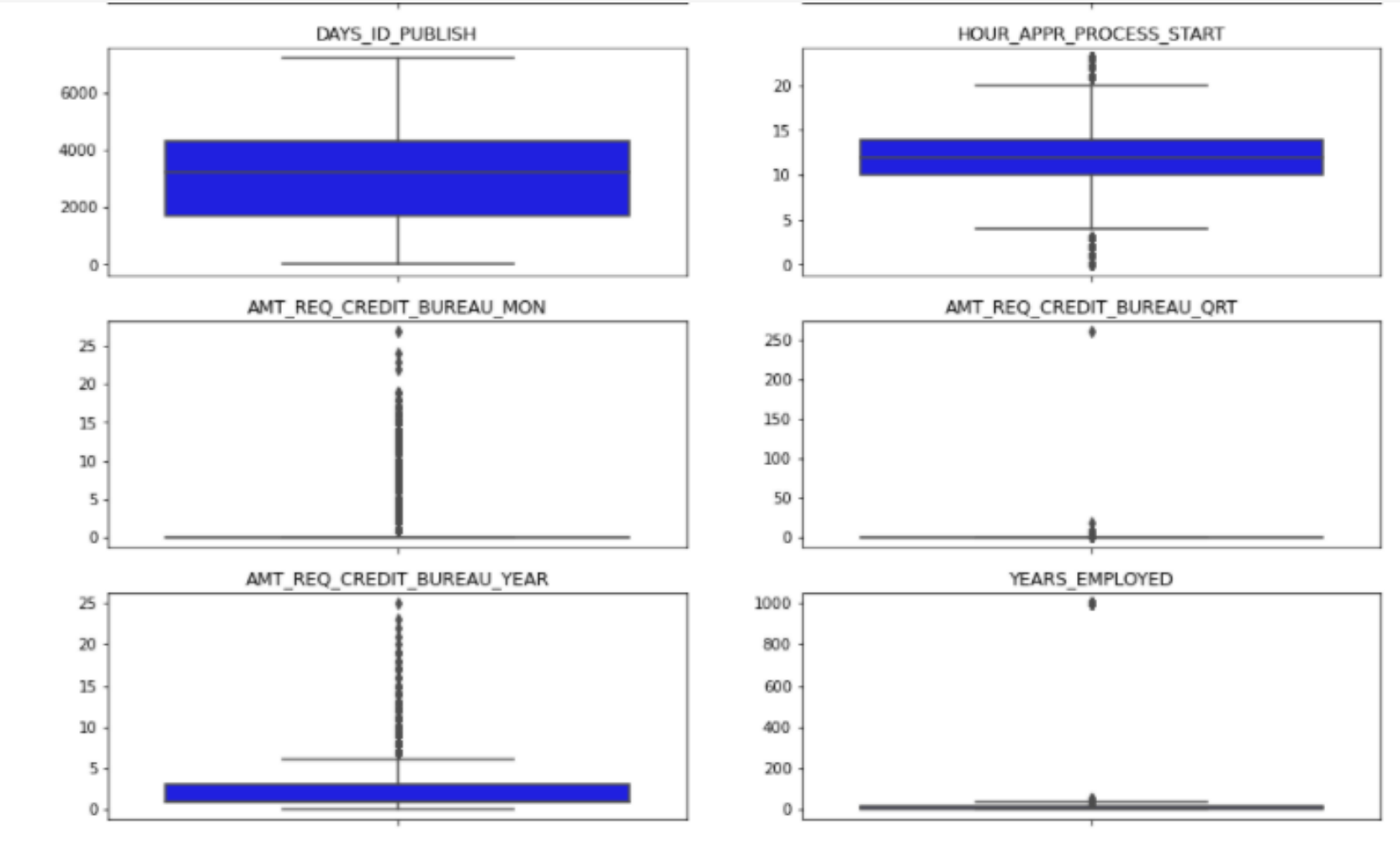


- Bar plot showing the highest value under Unknown named bar in Occupation Type

Checking Outliers for columns

'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH' , 'HOUR_APPR_PROCESS_START', 'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT' , 'AMT_REQ_CREDIT_BUREAU_YEAR' , 'YEARS_EMPLOYED'




DAYS_ID_PUBLISH
6000
4000
2000
0
HOUR_APPR_PROCESS_START
20
15
10
5
0
AMT_REQ_CREDIT_BUREAU_MON
25
20
15
10
5
0
AMT_REQ_CREDIT_BUREAU_QRT
250
200
150
100
50
0
AMT_REQ_CREDIT_BUREAU_YEAR
25
20
15
10
5
0
YEARS_EMPLOYED
1000
800
600
400
200
0

Observation

- CNT_CHILDREN- has outlier till 19 which looks like a wrong entry
- AMT_INCOME_TOTAL - has outlier till 12, which is the exception
- AMT_CREDIT - has many outliers
- DAYS_BIRTH - has no outlier and this column is very reliable
- DAYS_EMPLOYED - has outlier at 350000 which comes as 958 years, hence it is due to wrong entry
- DAYS_REGISTRATION - has many outliers
- DAYS_ID_PUBLISH - has no outlier and this column is very reliable
- HOUR_APPR_PROCESS_START- has both side outliers
- AMT_REQ_CREDIT_BUREAU_MON - has many outliers
- AMT_REQ_CREDIT_BUREAU_QRT - has extreme outliers
- AMT_REQ_CREDIT_BUREAU_YEAR - has many outliers
- YEARS_EMPLOYED - has exception outlier and ranging at 1000 years which is not possible

Previous Application Data

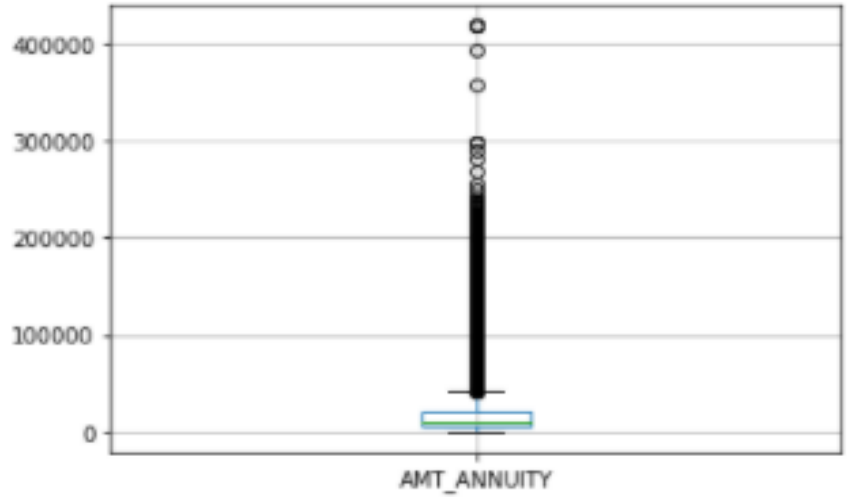
- Previous Application data has 1048575 rows and 37 columns.

- After checking for missing values, we took 50% as a threshold because greater than 40% data is not much relevant. However, data checking for greater than 15 % has relevant then we impute the value as per the requirement i.e. Mode, Median and Mean.

Observation

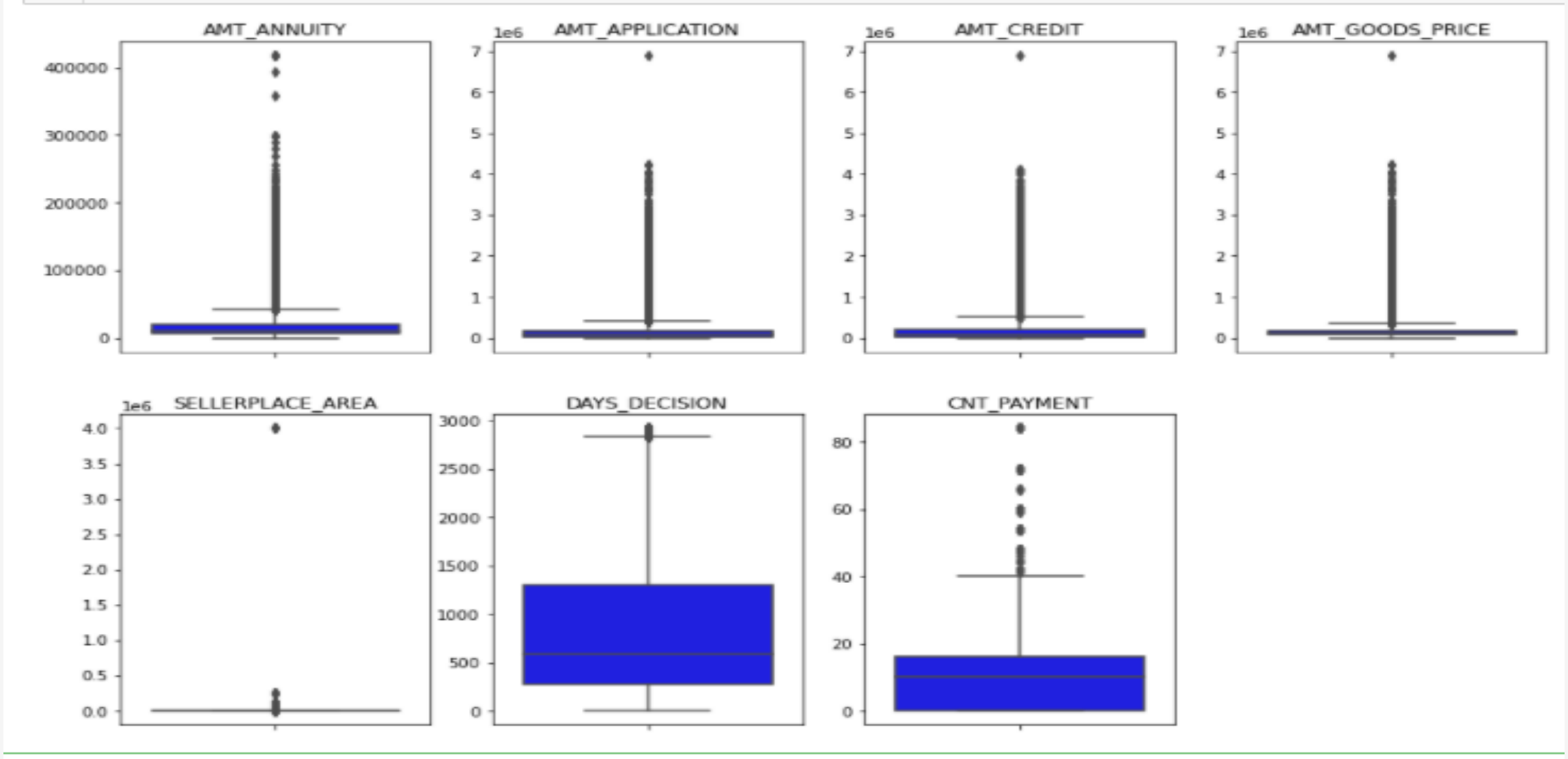
Showing so many outliers in box plot and hence imputing with mean would not be the right approach and hence imputing with the median.

```
count    815566.000000
mean      15891.265151
std       14745.557438
min           0.000000
25%        6301.350000
50%       11250.000000
75%       20523.003750
max      418058.145000
Name: AMT_ANNUITY, dtype: float64
```



Checking Outliers for columns

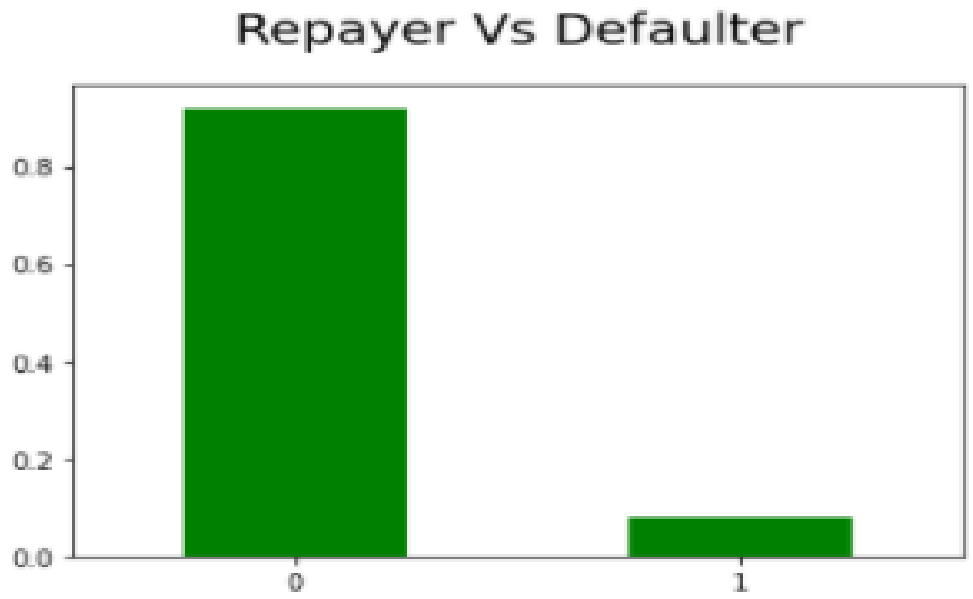
'AMT_ANNUITY','AMT_APPLICATION','AMT_CREDIT','AMT_GOODS_PRICE', 'SELLERPLACE_AREA','DAYS_DECISION','CNT_PAYMENT'



Data Analysis with the help of Univariate, Bivariate, and Multivariate

Univariate

```
0    91.927118
1     8.072882
Name: TARGET, dtype: float64
```

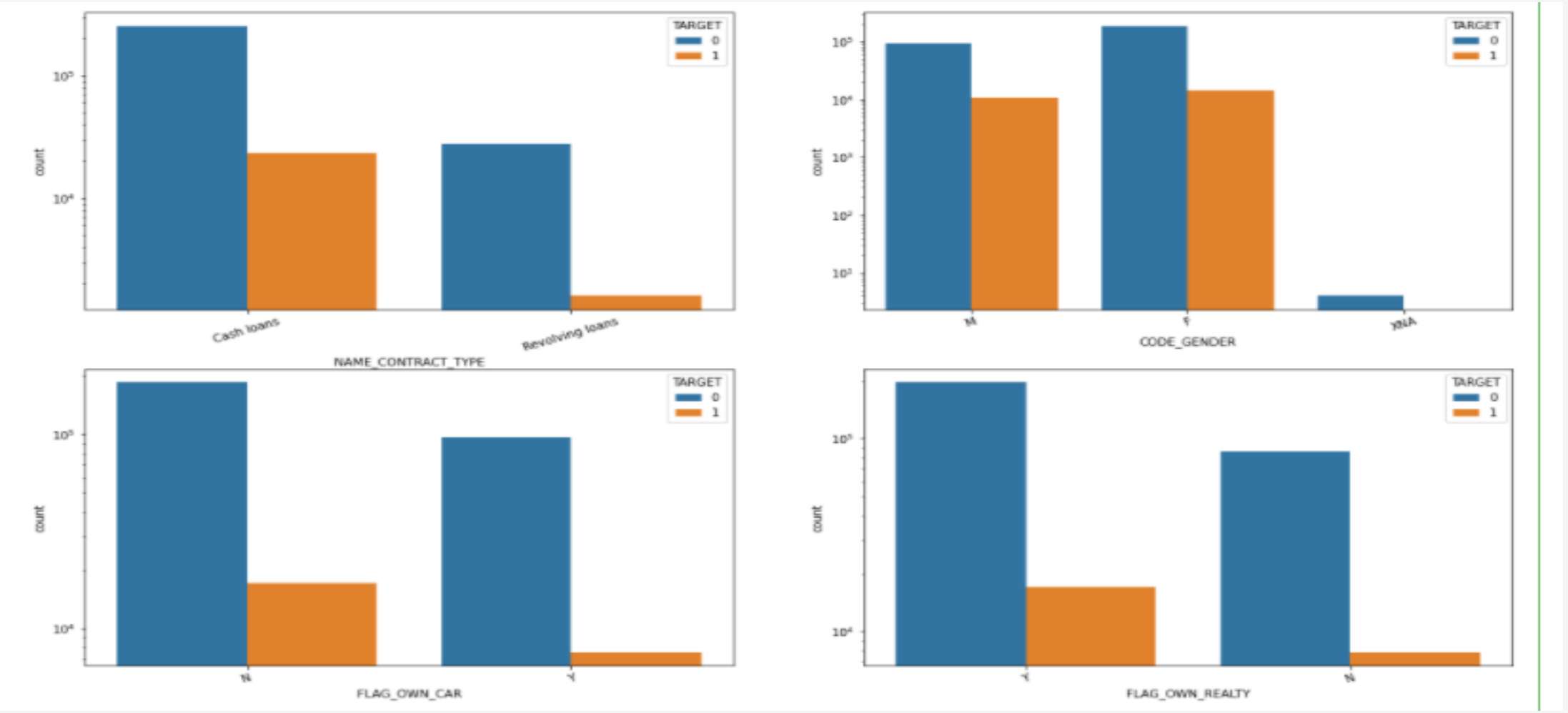


- In application_data there exists 91.927118% of "non-defaulter" and 8.072882% of "defaulter" customers.

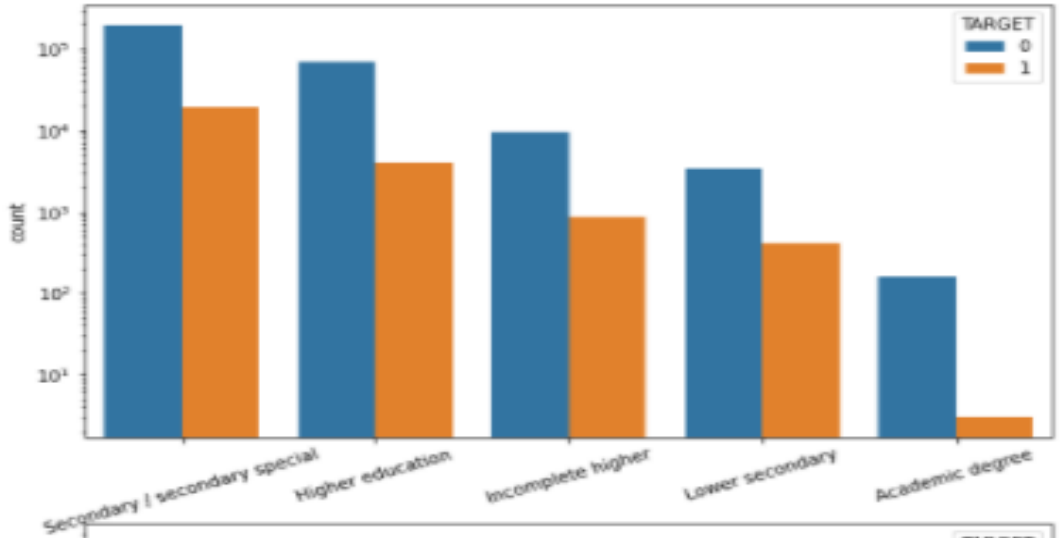
- Plot showing that secondary/special educated people are applying loans in high in number.
- Secondary/special educated people are applying loans high in number and Academic degree educated people are less applying loan.

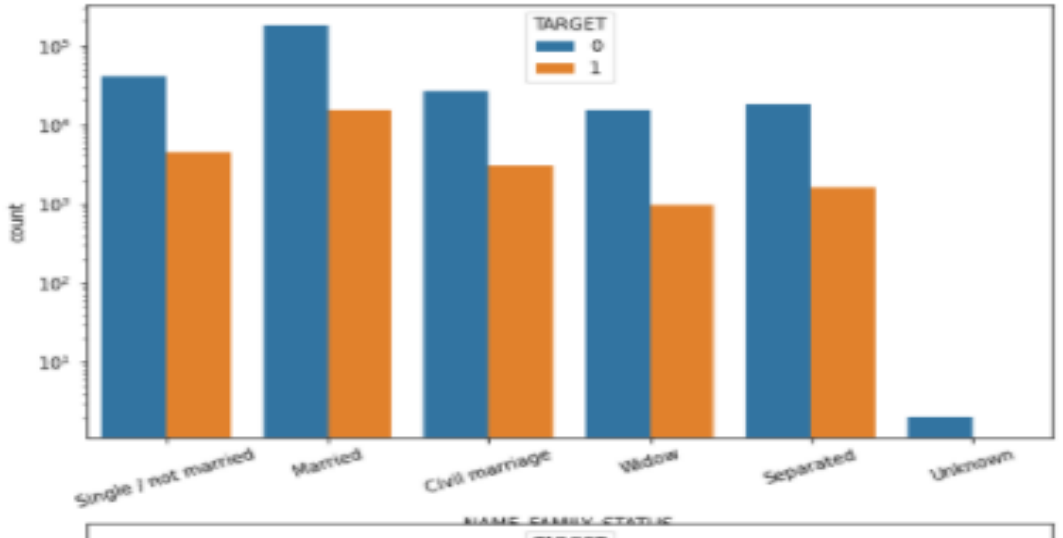
Data Analysis with the help of Univariate, Bivariate, and Multivariate

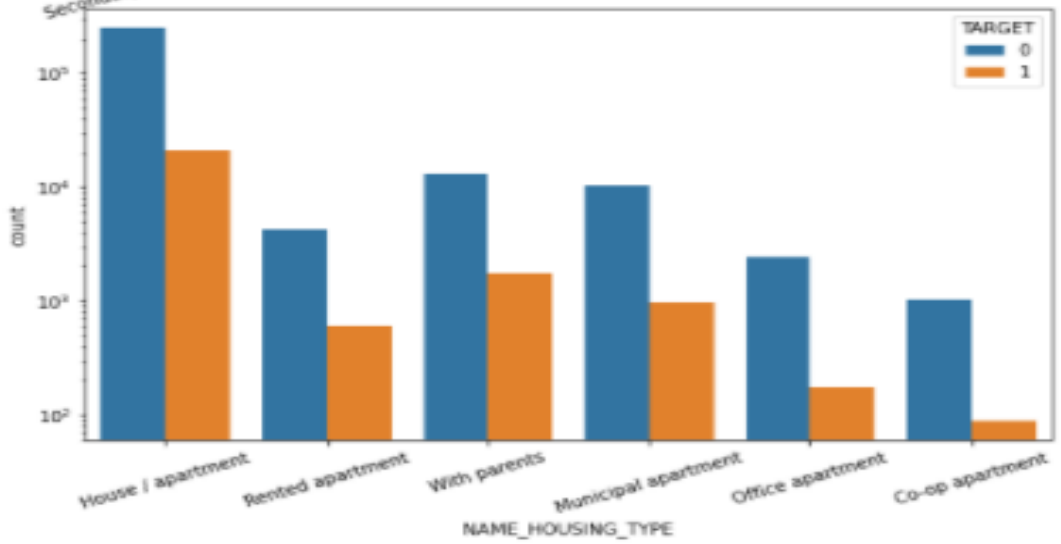
Bivariate – Target **versus** 'NAME_CONTRACT_TYPE','CODE_GENDER','FLAG_OWN_CAR','FLAG_OWN_REALTY', 'NAME_EDUCATION_TYPE','NAME_FAMILY_STATUS','NAME_HOUSING_TYPE','WEEKDAY_APPR_PROCESS_START'

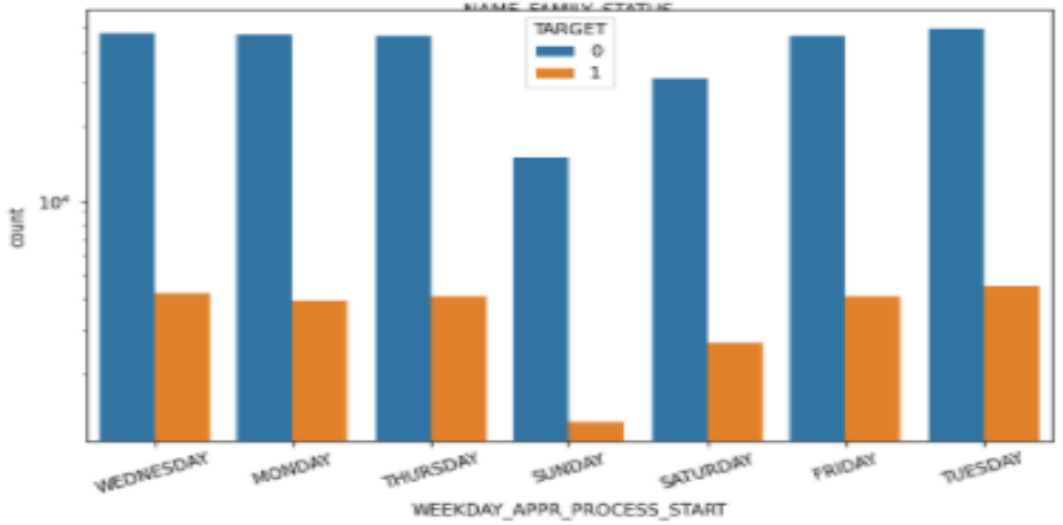


Bivariate – Target **versus** 'NAME_CONTRACT_TYPE','CODE_GENDER','FLAG_OWN_CAR','FLAG_OWN_REALTY', 'NAME_EDUCATION_TYPE','NAME_FAMILY_STATUS','NAME_HOUSING_TYPE','WEEKDAY_APPR_PROCESS_START'







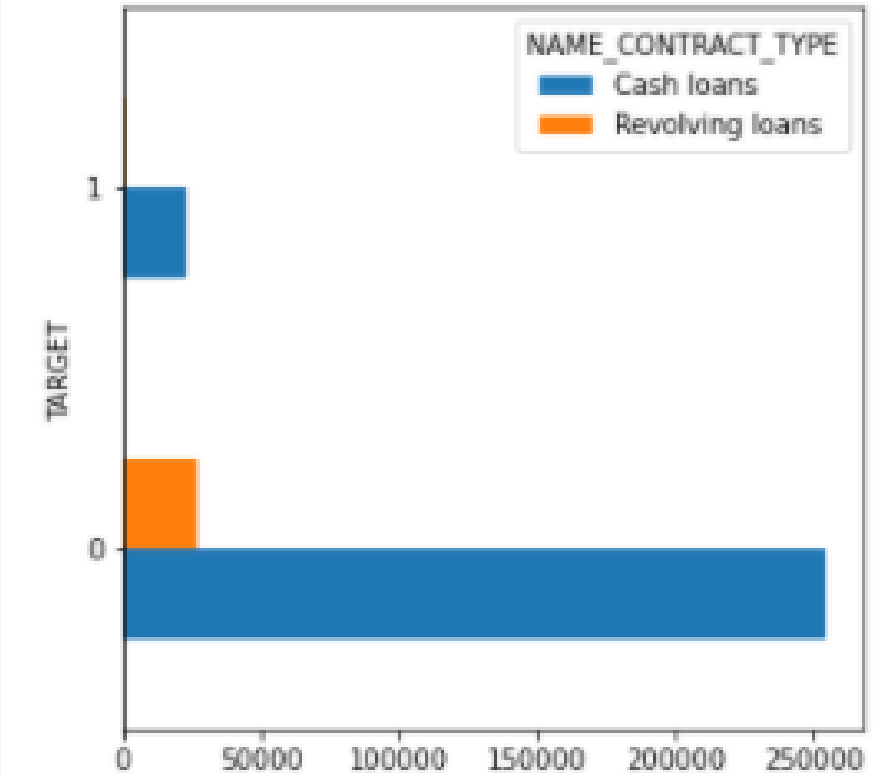


Observation

- * People tend to take more cash loans, and the default percentage of revolving loans is less
- * Female tends to take more loans
- * People who don’t own a car tends to take more loans
- * People with real estate tends to take more loans
- * we can conclude that secondary/special educated people are applying for loans in a high in number
- * We can say more married people tend to take more Loan as compared to other categories
- * People with houses or apartments tend to take more loans
- * People who started the application process on Sunday are less likely to default
- * Saturday and Sunday are less busy for banks in terms of loan applications

Correlation between TARGET and NAME_CONTRACT_TYPE variables

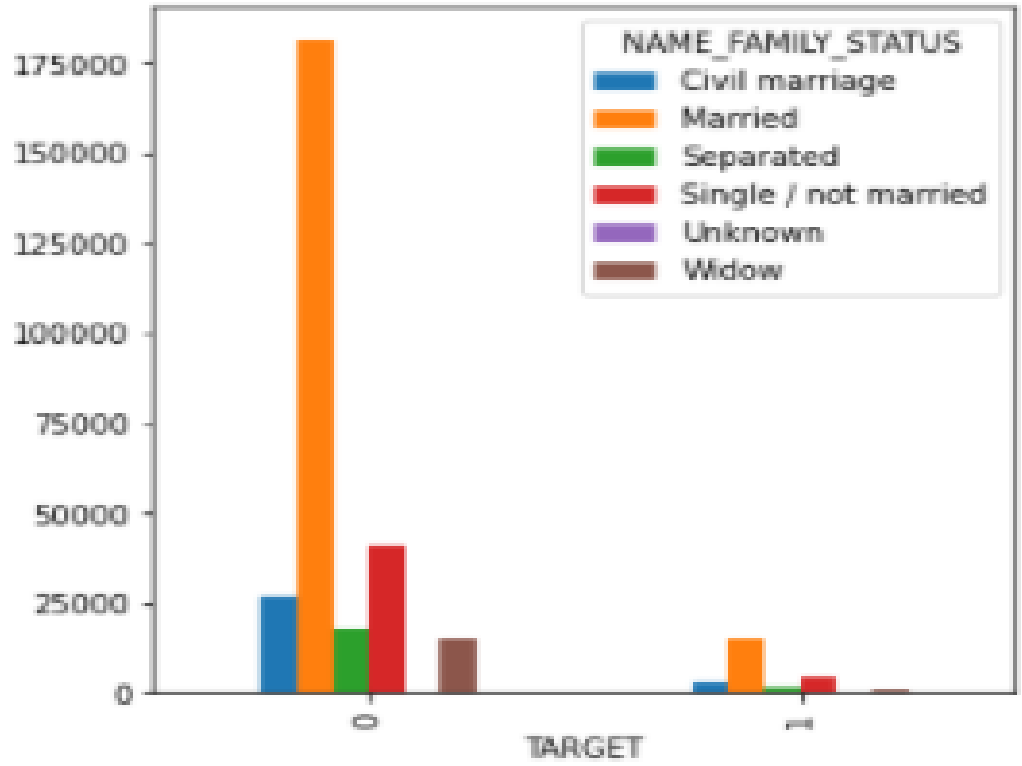
NAME_CONTRACT_TYPE	Cash loans	Revolving loans
TARGET		
0	255011	27675
1	23221	1604



The plot shows that Cash Loans applied more for loans and has fewer defaulters

Correlation between TARGET and NAME_FAMILY_STATUS variables

NAME_FAMILY_STATUS	Unknown	Widow
TARGET		
0	2	15151
1	0	937

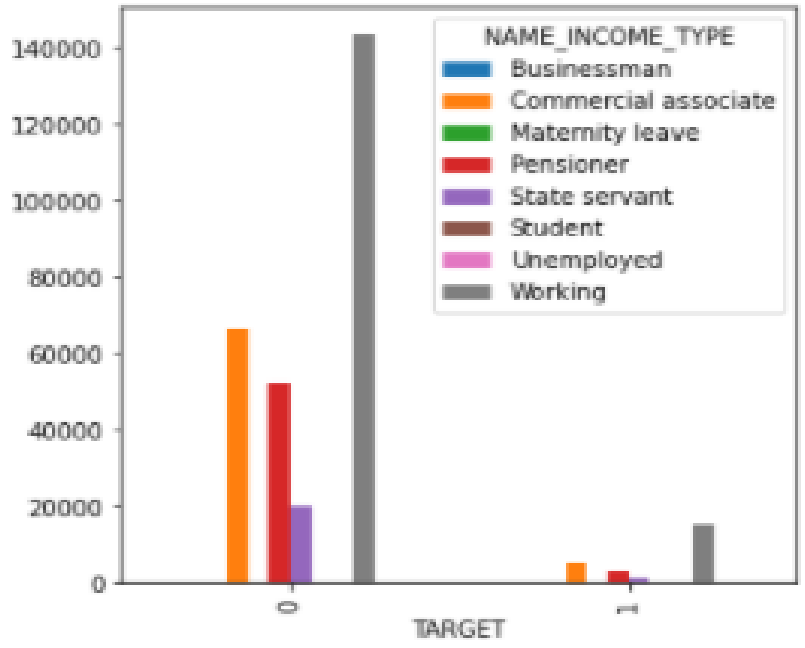


The plot shows that Married people apply more for loans and has fewer defaulters

Correlation between TARGET and NAME_INCOME_TYPE variables

NAME_INCOME_TYPE	Businessman	Commercial associate	Maternity leave
TARGET			
0	10	66257	3
1	0	5360	2

NAME_INCOME_TYPE	Pensioner	State servant	Student	Unemployed	Working
TARGET					
0	52380	20454	18	14	143550
1	2982	1249	0	8	15224

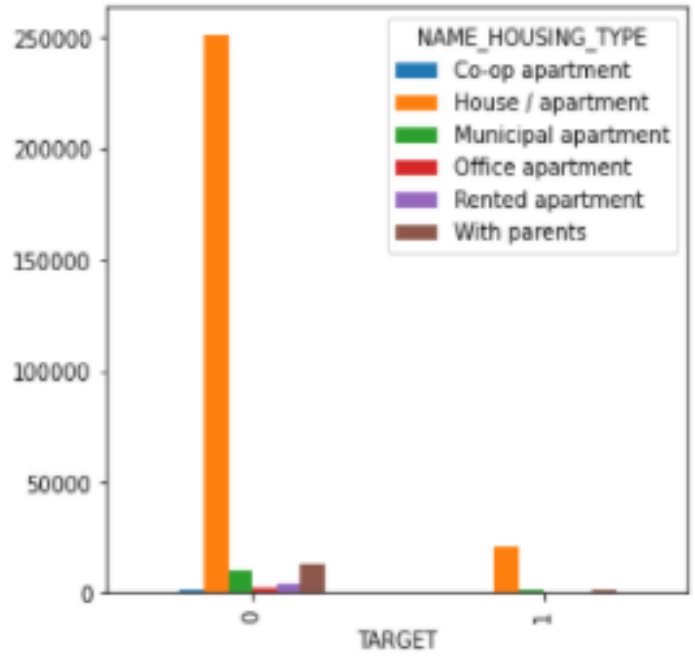


The plot shows that working people apply more for loans and has fewer defaulters

Correlation between TARGET and NAME_HOUSING_TYPE variables

NAME_HOUSING_TYPE	Co-op apartment	House / apartment	Municipal apartment
TARGET			
0	1033	251596	10228
1	89	21272	955

NAME_HOUSING_TYPE	Office apartment	Rented apartment	With parents
TARGET			
0	2445	4280	13104
1	172	601	1736



The plot shows that House/apartment owners apply more for loans and has fewer defaulters

Final Conclusion

- The majority of the previously cancelled client repaid the loan.
- Clients who were refused by the bank for a loan earlier now turned into repaying clients,

and these clients could be contacted for loans now

Top 10 major variables for predicting the loan risk :

1 NAME_EDUCATION_TYPE - higher degree has fewer chances of defaults
2. AMT_INCOME_TOTAL - higher the income fewer chances to default
3. DAYS_BIRTH – people with a high age group have fewer chances to default
4. AMT_CREDIT - people with moderate credit amounts have fewer chances to default
5. DAYS_EMPLOYED – higher the working experience fewer chances to defaults
6. AMT_ANNUITY – moderate loan has comparatively fewer chances of defaults
7. NAME_INCOME_TYPE – Students and Businessmen have no default
8. CODE_GENDER – Females apply for a loan in a high percentage
9. NAME_HOUSING_TYPE - home/ apartments has fewer defaults
10. NAME_FAMILY_STATUS - married people are fewer defaulters

THANK YOU !!