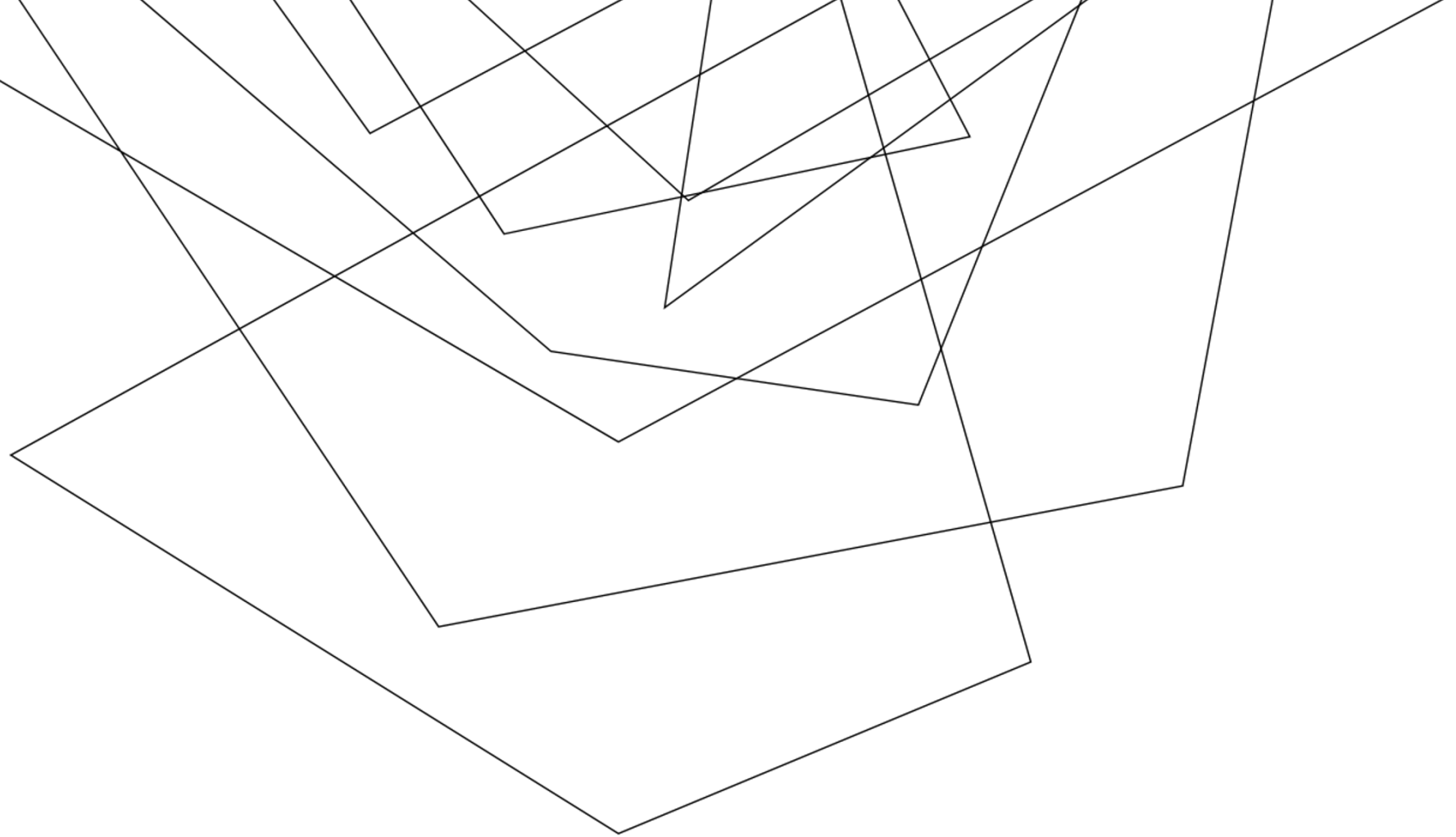


CREDIT EDA CASE STUDY

~ Jinesh Puglia

PROBLEM STATEMENT - 1

INTRODUCTION

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

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.

BUSINESS UNDERSTANDING

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.
To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

PROBLEM STATEMENT - II

Results Expected by Learners

Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.

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.

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.

Identify if there is data imbalance in the data. Find the ratio of data imbalance.

Hint: How will you analyse the data in case of data imbalance? You can plot more than one type of plot to analyse the different aspects due to data imbalance. For example, you can choose your own scale for the graphs, i.e. one can plot in terms of percentage or absolute value. Do this analysis for the **'Target variable'** in the dataset (**clients with payment difficulties** and **all other cases**). Use a mix of univariate and bivariate analysis etc.

Hint: Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights.
Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
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.
Include visualisations and summarise 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.**

You need to submit one/two Ipython notebook which clearly explains the thought process behind your analysis (either in comments of markdown text), code and relevant plots. The presentation file needs to be in PDF format and should contain the points discussed above with the necessary visualisations. Also, all the visualisations and plots must be done in Python(should be present in the Ipython notebook), though they may be recreated in Tableau for better aesthetics in the PPT file.

IMPORT REQUIRED LIBRARIES

Import Numpy
Import Pandas
Import Matplot.pyplot

READING AND UNDERSTANDING THE DATASET

Read CSV
Show head
Read shape
See Info
Describe
Show columns

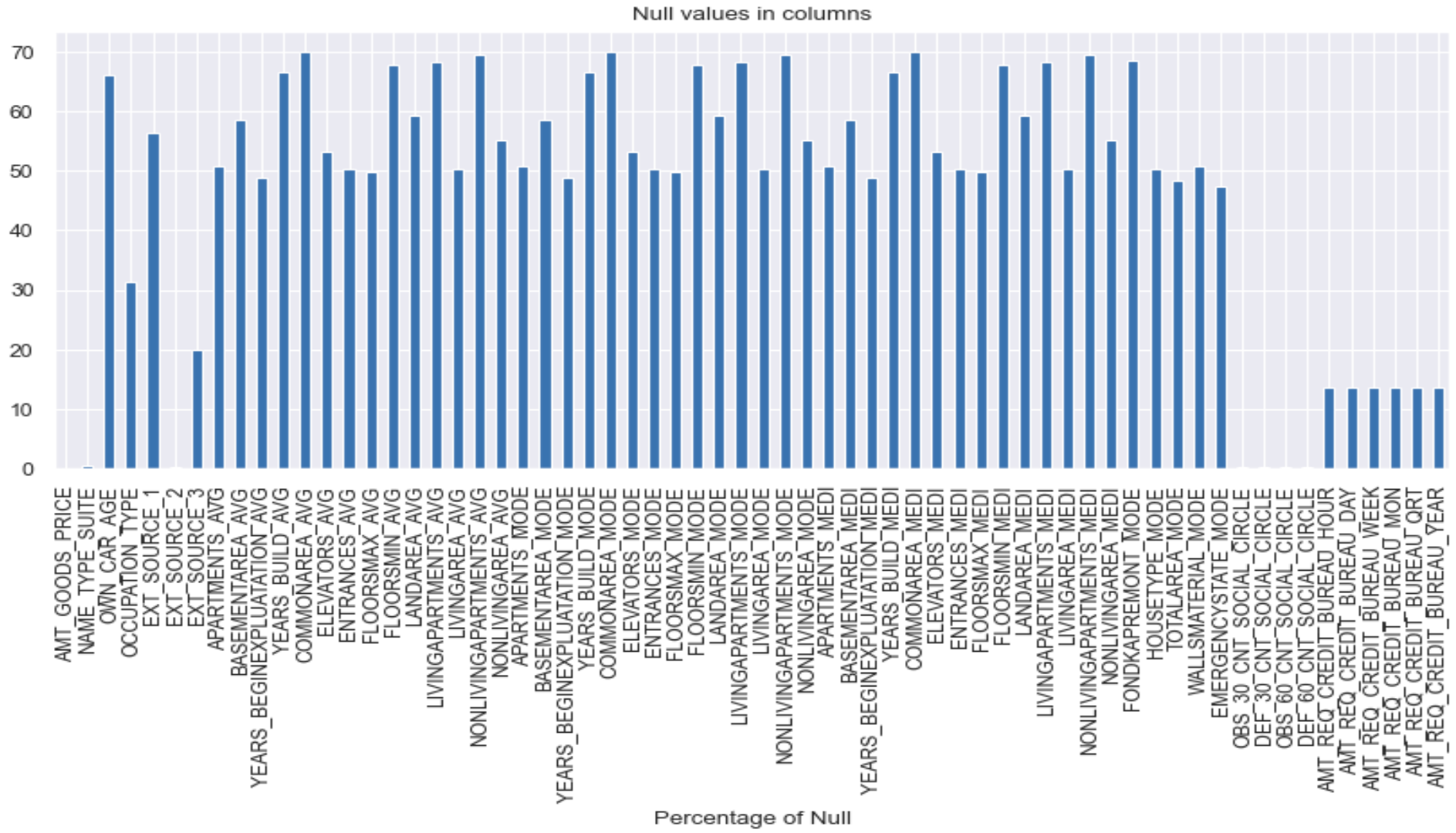
DATA CHECKING AND MISSING VALUES

Make a function to get null values
Missing values of all columns
Finding out columns with only null values
Visualizing null values of columns in Graph

VISUALIZING NULL VALUES OF COLUMNS IN GRAPH

Make a function to get null values
Missing values of all columns
Finding out columns with only null values
Visualizing null values of columns in Graph

VISUALIZING NULL VALUES OF COLUMNS IN GRAPH



TAKING OUT COLUMNS WITH >50%

```
Number of columns with null value > 50% : 41
OWN_CAR_AGE                    65.99
EXT_SOURCE_1                   56.38
APARTMENTS_AVG                 50.75
BASEMENTAREA_AVG               58.52
YEARS_BUILD_AVG                66.50
COMMONAREA_AVG                 69.87
ELEVATORS_AVG                  53.30
ENTRANCES_AVG                  50.35
FLOORSMIN_AVG                  67.85
LANDAREA_AVG                   59.38
LIVINGAPARTMENTS_AVG           68.35
LIVINGAREA_AVG                 50.19
NONLIVINGAPARTMENTS_AVG        69.43
NONLIVINGAREA_AVG              55.18
APARTMENTS_MODE                50.75
BASEMENTAREA_MODE              58.52
YEARS_BUILD_MODE               66.50
COMMONAREA_MODE                69.87
ELEVATORS_MODE                 53.30
ENTRANCES_MODE                 50.35
FLOORSMIN_MODE                 67.85
LANDAREA_MODE                  59.38
LIVINGAPARTMENTS_MODE          68.35
LIVINGAREA_MODE                50.19
NONLIVINGAPARTMENTS_MODE       69.43
NONLIVINGAREA_MODE             55.18
APARTMENTS_MEDI                50.75
BASEMENTAREA_MEDI              58.52
YEARS_BUILD_MEDI               66.50
COMMONAREA_MEDI                69.87
ELEVATORS_MEDI                 53.30
ENTRANCES_MEDI                 50.35
FLOORSMIN_MEDI                 67.85
LANDAREA_MEDI                  59.38
LIVINGAPARTMENTS_MEDI          68.35
```

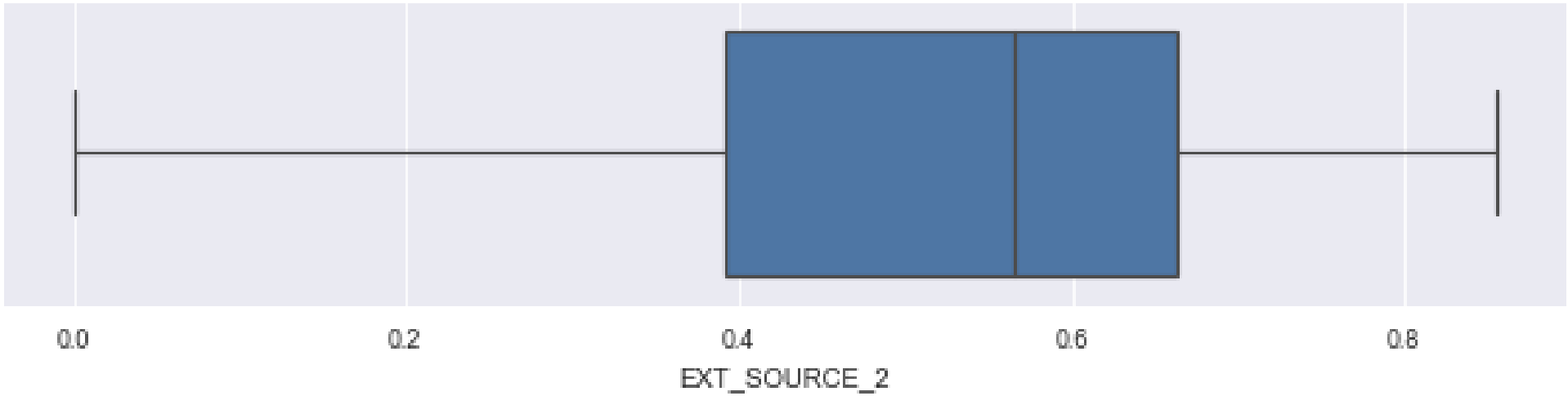
COLUMNS WITH NULL VALUES <15%

```
Number of columns with null value < 15% : 13
AMT_GOODS_PRICE                0.09
NAME_TYPE_SUITE                0.42
EXT_SOURCE_2                   0.21
OBS_30_CNT_SOCIAL_CIRCLE       0.33
DEF_30_CNT_SOCIAL_CIRCLE       0.33
OBS_60_CNT_SOCIAL_CIRCLE       0.33
DEF_60_CNT_SOCIAL_CIRCLE       0.33
AMT_REQ_CREDIT_BUREAU_HOUR    13.50
AMT_REQ_CREDIT_BUREAU_DAY     13.50
AMT_REQ_CREDIT_BUREAU_WEEK    13.50
AMT_REQ_CREDIT_BUREAU_MON     13.50
AMT_REQ_CREDIT_BUREAU_QRT     13.50
AMT_REQ_CREDIT_BUREAU_YEAR    13.50
dtype: float64
```

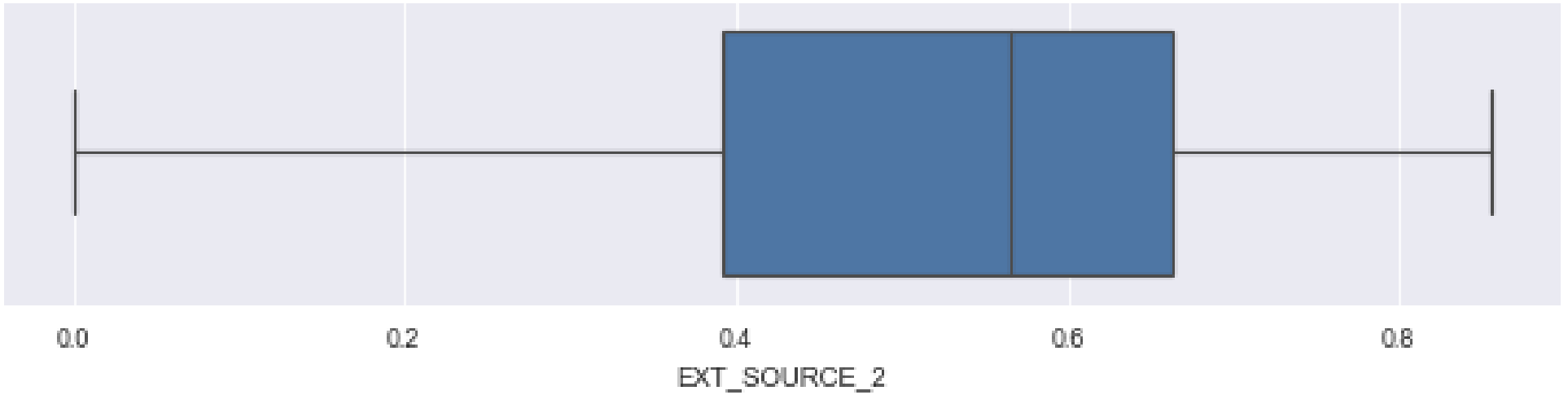
IDENTIFYING UNIQUE VALUES WITH COLUMNS <15%

```
EXT_SOURCE_2                  119831
AMT_GOODS_PRICE                 1002
OBS_30_CNT_SOCIAL_CIRCLE          33
OBS_60_CNT_SOCIAL_CIRCLE          33
AMT_REQ_CREDIT_BUREAU_YEAR        25
AMT_REQ_CREDIT_BUREAU_MON         24
AMT_REQ_CREDIT_BUREAU_QRT         11
DEF_30_CNT_SOCIAL_CIRCLE          10
DEF_60_CNT_SOCIAL_CIRCLE           9
AMT_REQ_CREDIT_BUREAU_DAY          9
AMT_REQ_CREDIT_BUREAU_WEEK         9
NAME_TYPE_SUITE                    7
AMT_REQ_CREDIT_BUREAU_HOUR         5
dtype: int64
```

CONTINOUS VARIBALE



CONTINOUS VARIBALE



Observation from Boxplots:

For 'EXT_SOURCE_2' no outliers present. So data is rightly present.

For 'AMT_GOODS_PRICE' outlier present in the data. so need to impute with median value: 4

Now removing the columns from the data set which are unused for better analysis

- Drop
- Head
- Shape

Imputing the value'XNA' which means not available for the column 'CODE_GENDER‘

- Locating
- Value counts

checking the CODE_GENDER

- Info

CASTING VARIABLE INTO NUMERIC IN THE DATASET

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREI
0	100002	1	Cash loans	M	N	Y	0	202500.0	40659
1	100003	0	Cash loans	F	N	N	0	270000.0	129350
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	13500
3	100006	0	Cash loans	F	N	Y	0	135000.0	31268
4	100007	0	Cash loans	M	N	Y	0	121500.0	51300
5	100008	0	Cash loans	M	N	Y	0	99000.0	49049
6	100009	0	Cash loans	F	Y	Y	1	171000.0	156072
7	100010	0	Cash loans	M	Y	Y	0	360000.0	153000
8	100011	0	Cash loans	F	N	Y	0	112500.0	101961
9	100012	0	Revolving loans	M	N	Y	0	135000.0	40500

10 rows × 83 columns

'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH‘

Age/Days columns are in -ve which needs to be converted to +ve value

- Head
- Tail

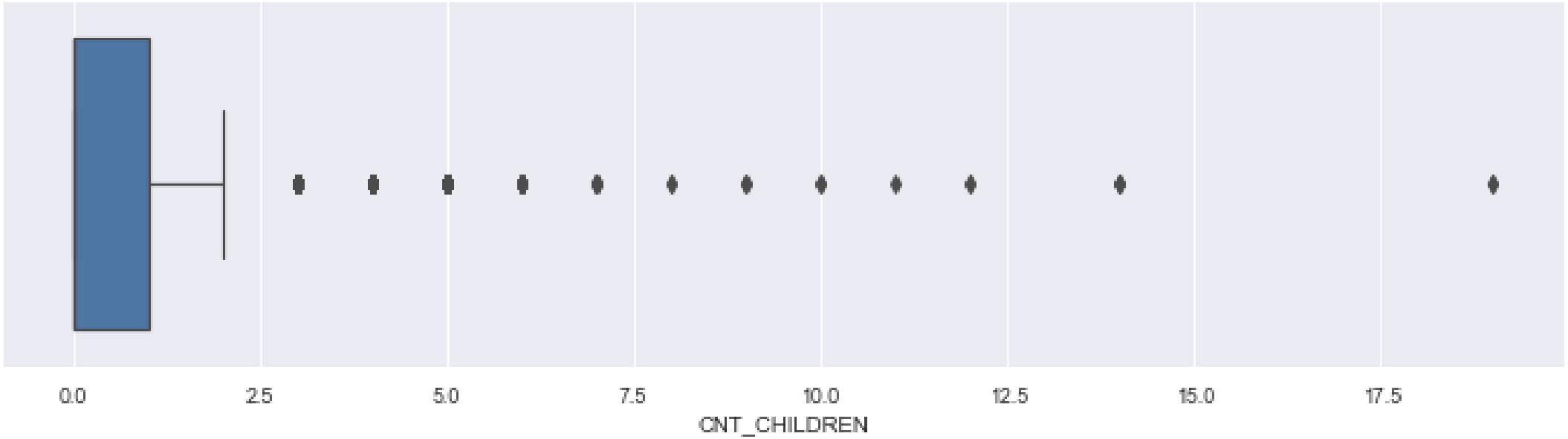
Checking outliers of numerical_column

- Describe
- Head

Now lets check box plot for 'CNT_CHILDREN', 'AMT_INCOME_TOTAL','AMT_CREDIT', 'AMT_ANNUITY', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION‘

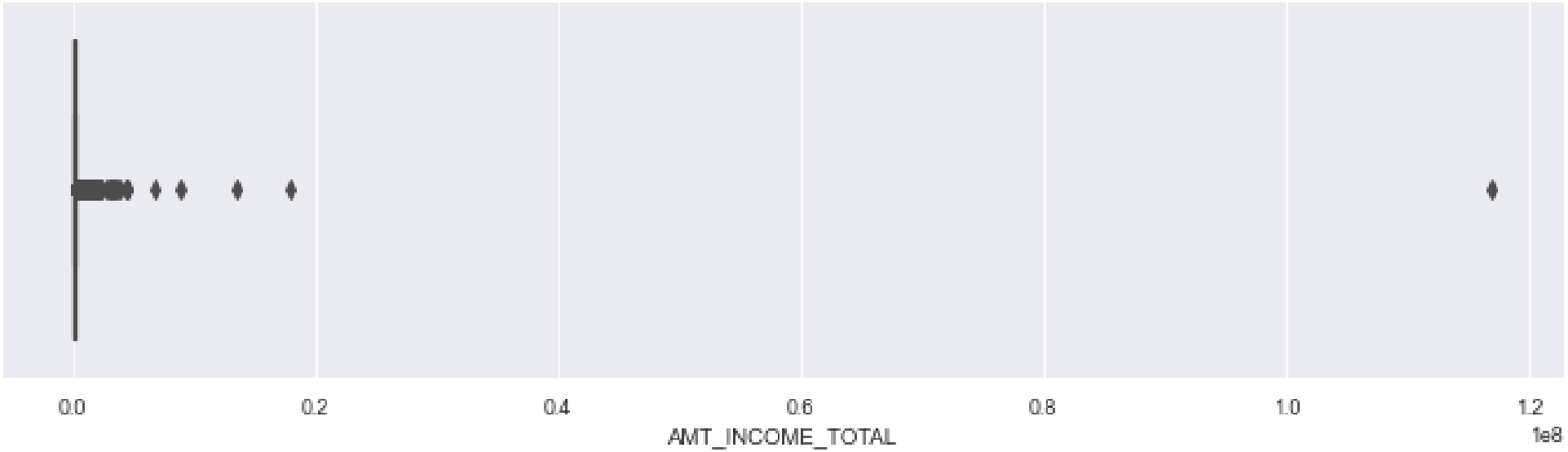
- Figure
- Boxplot
- Show

CNT_CHILDREN



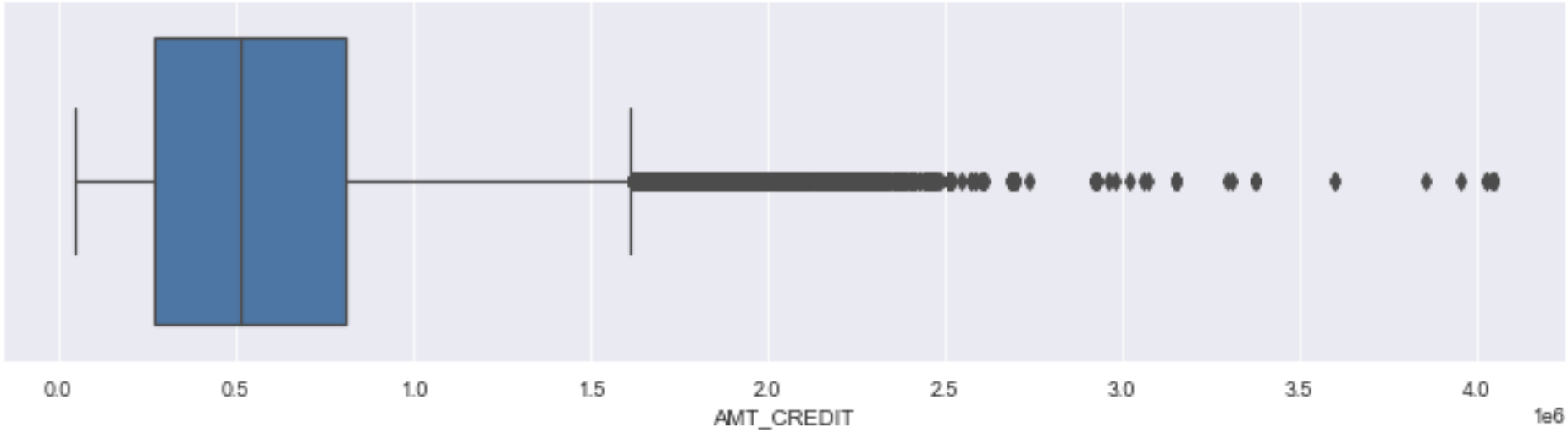
AMT_INCOME_TOTAL

- Figure
- Boxplot
- Show



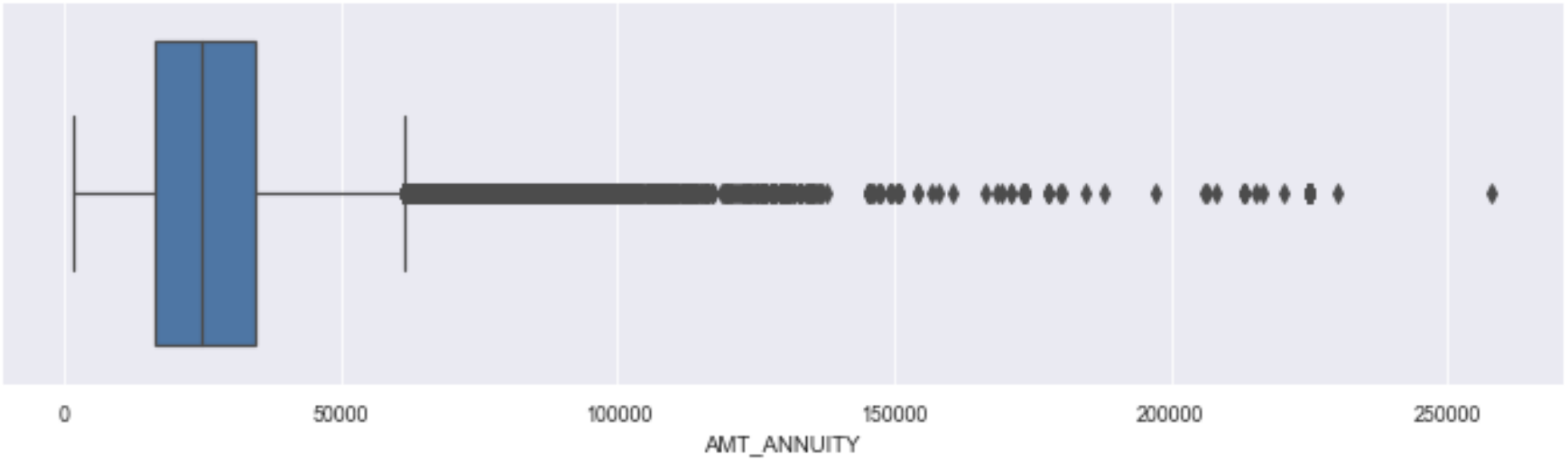
AMT_CREDIT

- Figure
- Boxplot
- Show



AMT_ANNUITY

- Figure
- Boxplot
- Show

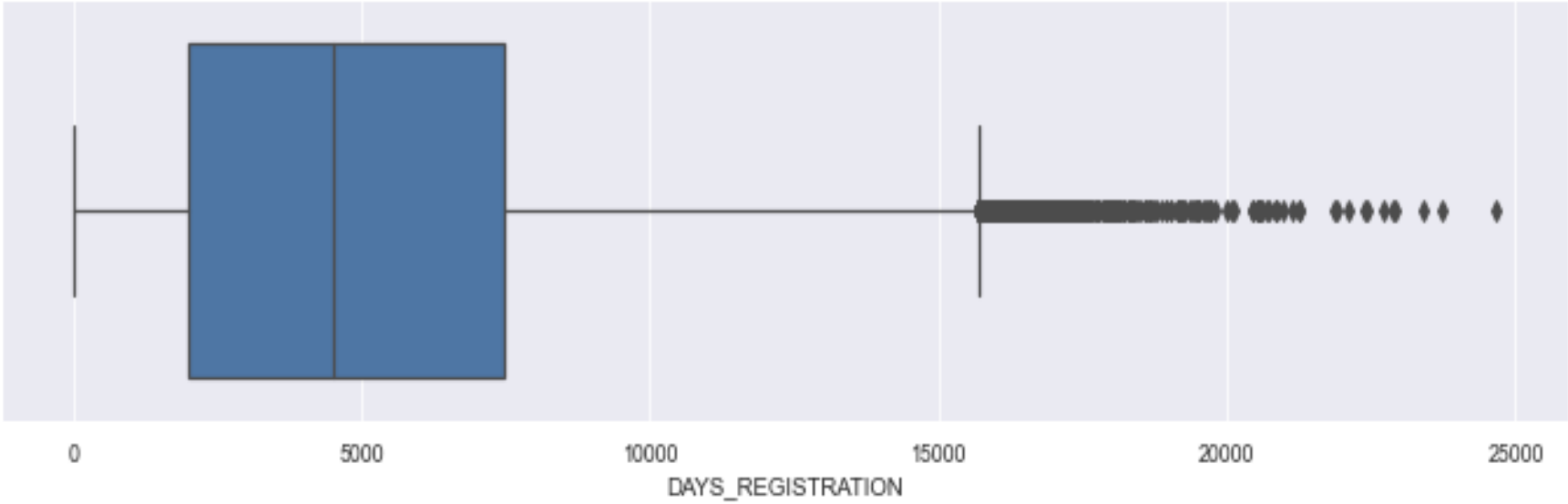


DAYS_EMPLOYED

- Figure
- Boxplot
- Show

DAYS_REGISTRATION

- Figure
- Boxplot
- Show



Same with this too 1st quartiles and 3rd quartile for DAYS_EMPLOYED is stays first quartile.

- From above box plots we found that numeric columns have outliers

Creating bins for continous variable categories column 'AMT_INCOME_TOTAL', 'AMT_GOODS_PRICE' and 'AMT_CREDIT‘

We will do for all three same

bins = [0,100000,200000,300000,400000,500000,10000000000]

slots = ['<100000', '100000-200000','200000-300000','300000-400000','400000-500000', '500000 and above']

ANALYSIS

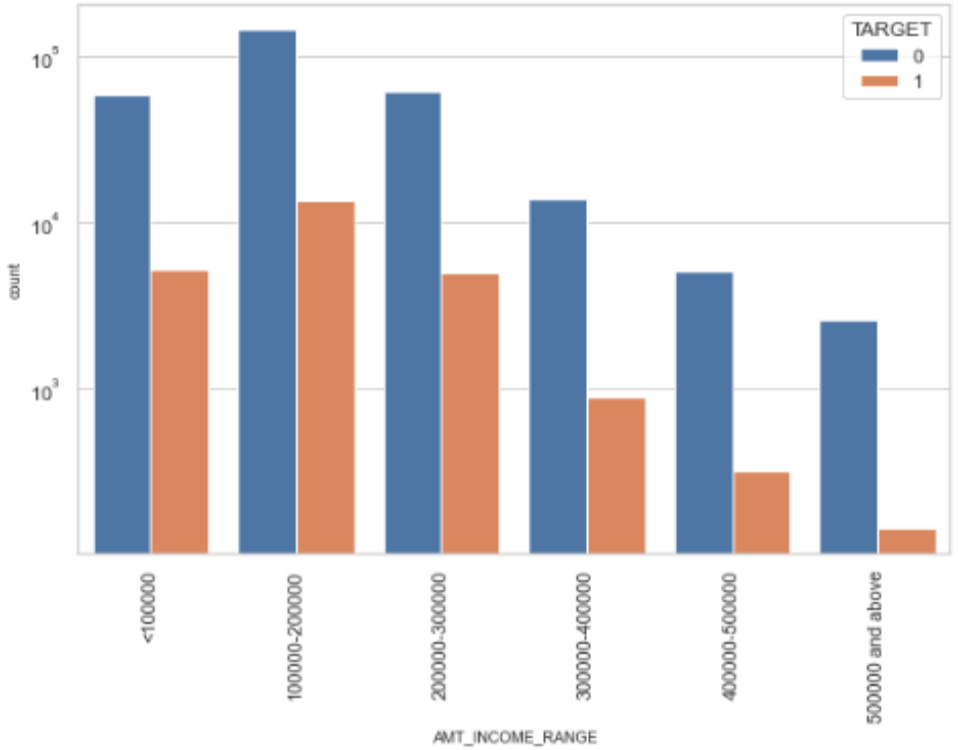
Dividing the dataset into two dataset of Target=1(client with payment difficulties) and Target=0(all other)

Calculating Imbalance percentage

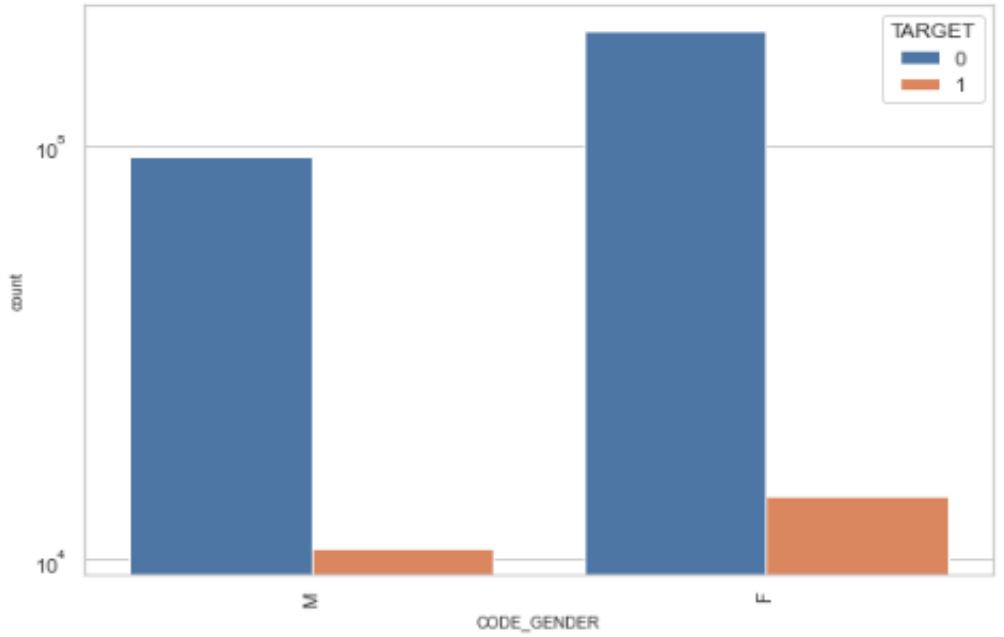
Since the majority is target0 and minority is target1

UNIVARIATE ANALYSIS

AMT_INCOME_RANGE

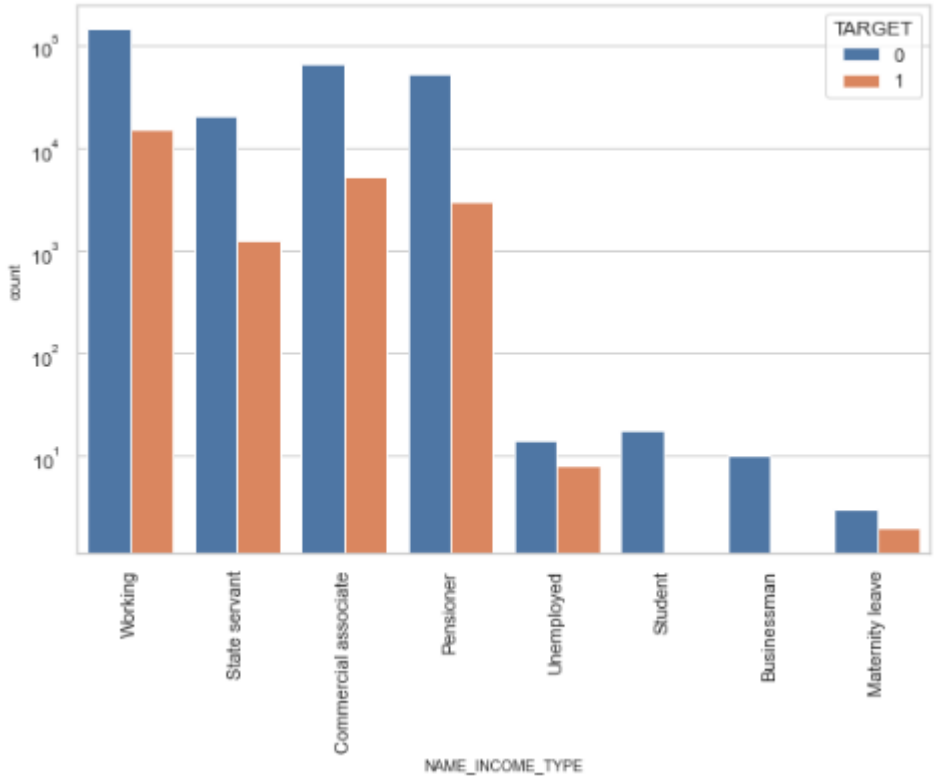


CODE_GENDER

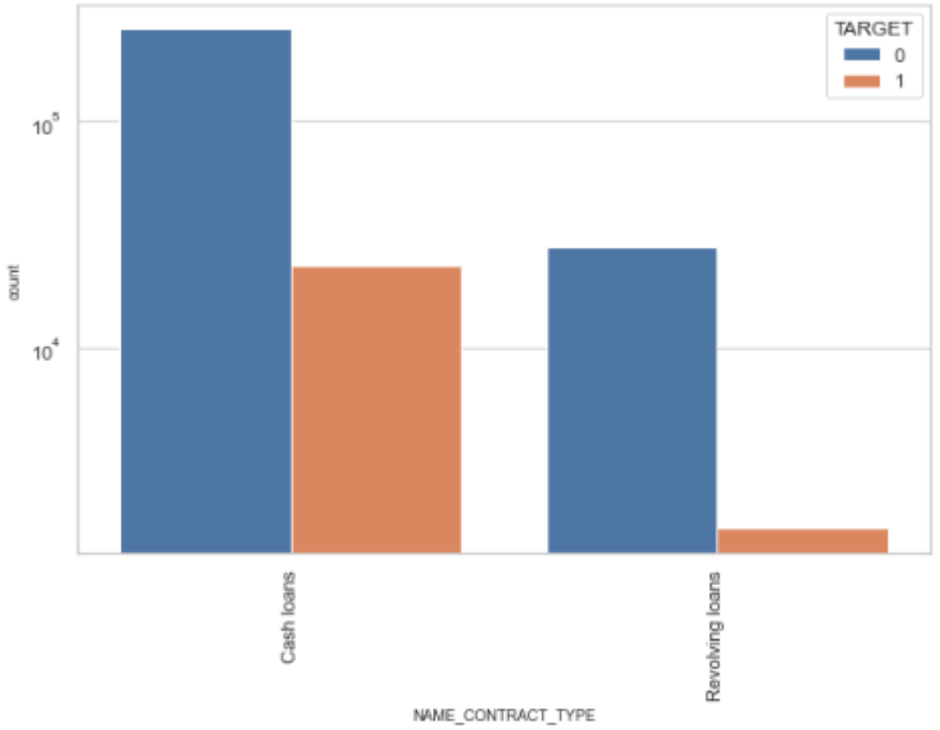


UNIVARIATE ANALYSIS

NAME_INCOME_TYPE



NAME_CONTRACT_TYPE



OBSERVATIONS

AMT_INCOME_RANGE :

- People in range 100000-200000 have high number of loan and also have high in defaulter
- Income segment >500000 has less defaulter.

CODE_GENDER:

- The % of defaulters are more in Male than Female

NAME_INCOME_TYPE:

- Student and business are higher in percentage of loan repayment.
- Working, State servent and Commercial associates are higher in default percentage.
- Maternity category is significantly higher problem in repayment.

NAME_CONTRACT_TYPE:

- For contract type ‘Cash loans’ are high in number of credits than ‘Revolving loans’ contract type.
- By above graph 'Revolving loans' is small amount compared to ‘Cash loans’

CORELATION

Top 10 corelated variables: Target0 dataaframe

	Var1	Var2	Correlation
3192	ENTRANCES_MEDI	ENTRANCES_AVG	1.0
3817	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.0
3125	ELEVATORS_MEDI	ELEVATORS_AVG	1.0
3058	COMMONAREA_MEDI	COMMONAREA_AVG	1.0
3259	FLOORSMIN_MEDI	FLOORSMIN_AVG	1.0

Top 10 correlated variables: Target1 dataaframe

	Var1	Var2	Correlation
3058	COMMONAREA_MEDI	COMMONAREA_AVG	1.0
3259	FLOORSMIN_MEDI	FLOORSMIN_AVG	1.0
3192	ENTRANCES_MEDI	ENTRANCES_AVG	1.0
3817	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.0
3460	LIVINGAREA_MEDI	LIVINGAREA_AVG	1.0

READING PREVIOUS_APPLICATION DATASET

Reading the previous_application csv file

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEK
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0	
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0	
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0	
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0	
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0	

5 rows × 37 columns

Describing the previous application data frame

	SK_ID_PREV	SK_ID_CURR	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	HOUR_APPR_PROCESS_STA
count	1.670214e+06	1.670214e+06	1.297979e+06	1.670214e+06	1.670213e+06	7.743700e+05	1.284699e+06	1.670214e+
mean	1.923089e+06	2.783572e+05	1.595512e+04	1.752339e+05	1.961140e+05	6.697402e+03	2.278473e+05	1.248418e+
std	5.325980e+05	1.028148e+05	1.478214e+04	2.927798e+05	3.185746e+05	2.092150e+04	3.153966e+05	3.334028e+
min	1.000001e+06	1.000010e+05	0.000000e+00	0.000000e+00	0.000000e+00	-9.000000e-01	0.000000e+00	0.000000e+
25%	1.461857e+06	1.893290e+05	6.321780e+03	1.872000e+04	2.416050e+04	0.000000e+00	5.084100e+04	1.000000e+
50%	1.923110e+06	2.787145e+05	1.125000e+04	7.104600e+04	8.054100e+04	1.638000e+03	1.123200e+05	1.200000e+
75%	2.384280e+06	3.675140e+05	2.065842e+04	1.803600e+05	2.164185e+05	7.740000e+03	2.340000e+05	1.500000e+
max	2.845382e+06	4.562550e+05	4.180581e+05	6.905160e+06	6.905160e+06	3.060045e+06	6.905160e+06	2.300000e+

8 rows × 21 columns

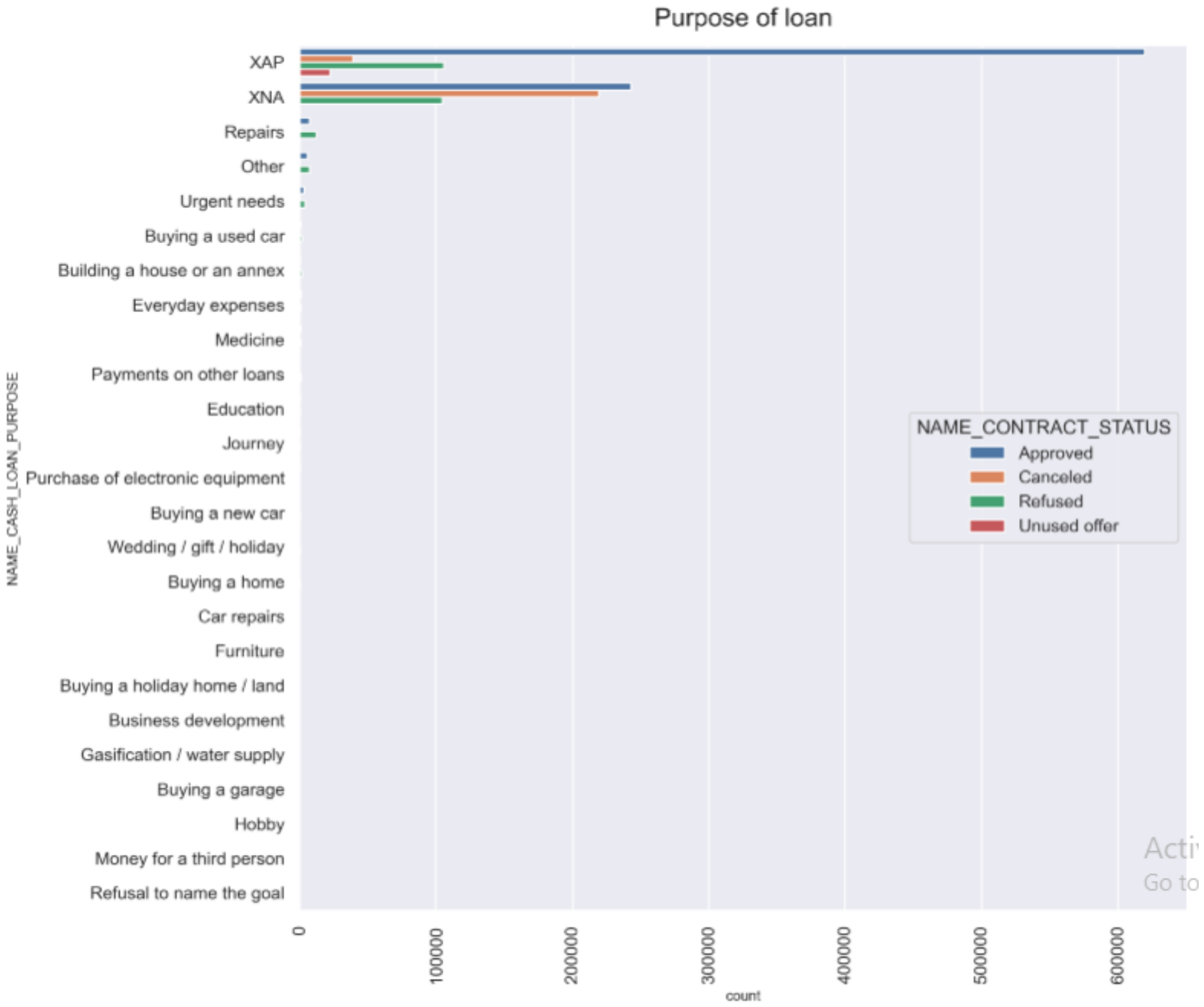
Merging both the dataframes (application_data, previous_application)

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_C
0	100002	1	Cash loans	M	N	Y	0	202500.0	
1	100003	0	Cash loans	F	N	N	0	270000.0	500
2	100003	0	Cash loans	F	N	N	0	270000.0	500
3	100003	0	Cash loans	F	N	N	0	270000.0	500
4	100004	0	Revolving loans	M	Y	Y	0	67500.0	

5 rows × 120 columns

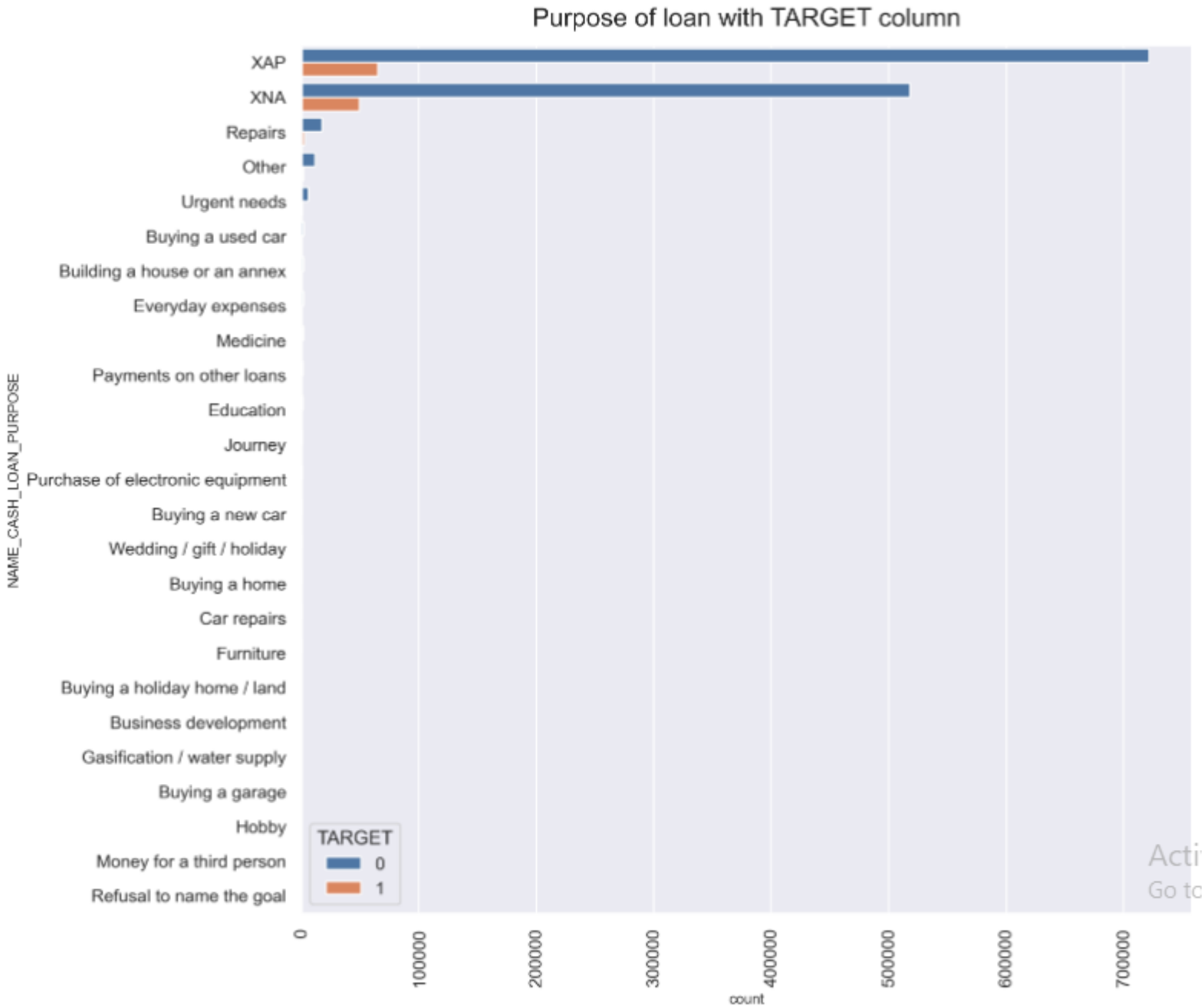
PERFORMING UNIVARITE ANALYSIS

Purpose of loan



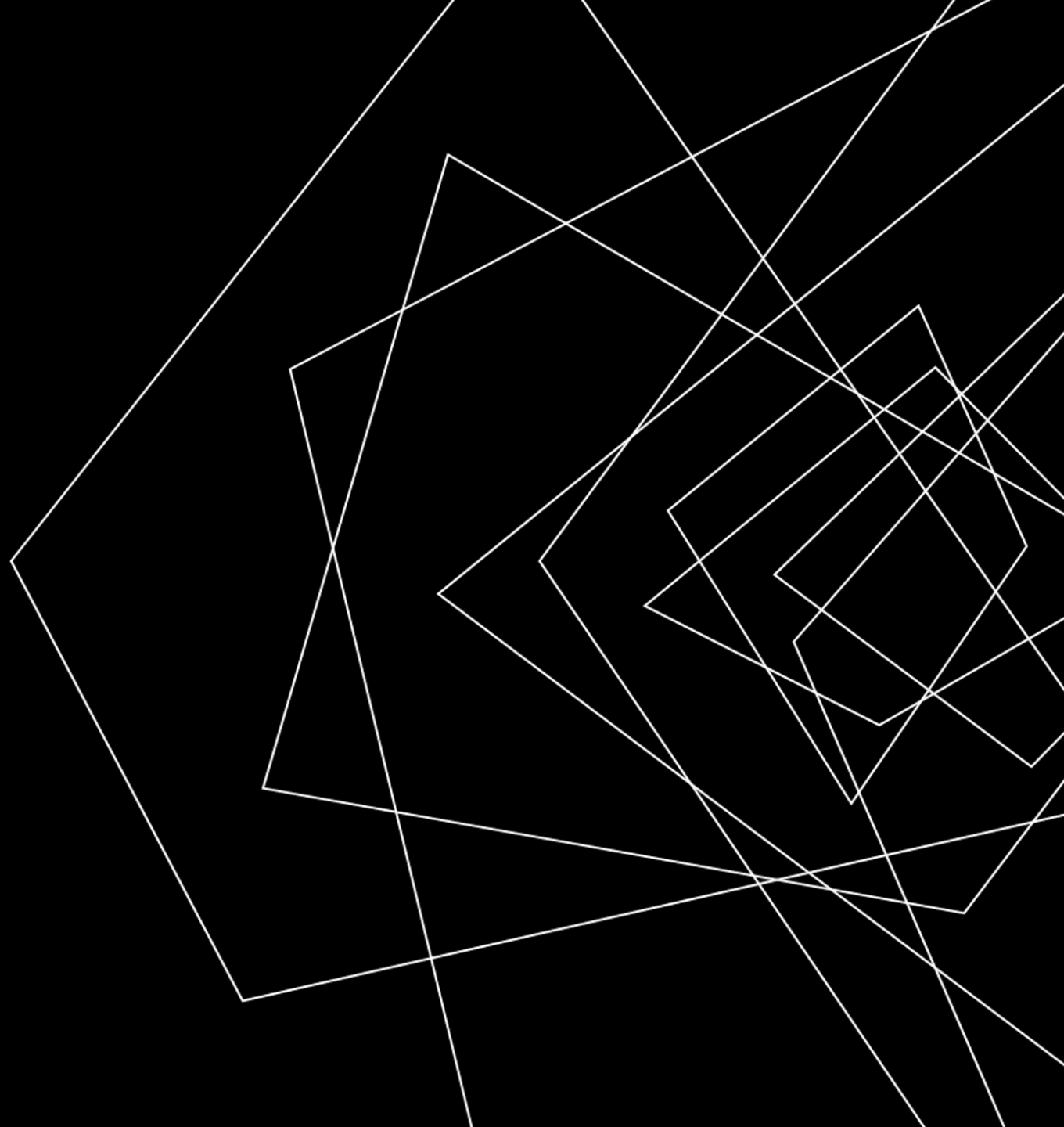
PERFORMING UNIVARITE ANALYSIS

Purpose of loan with TARGET column



OBSERVATION

Most of loan rejection was from 'repairs'



CONCLUSION FROM THE ANALYSIS

1. Banks must target more on contract type ‘Student’ , ’Pensioner’ and ‘Businessman’ for profitable business

2. Banks must focus less on income type ‘Working’ as it is has most number of unsuccessful payments in oreder to get rid of financial loss for the organization