Problem Statement - I

Introduction

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, 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 as 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 on different stages of the process. In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of 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).

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 data dictionary which describes the meaning of the variables.

Importing Required Libraries

```
# Import libraries import numpy as np
import pandas as pd import
matplotlib.pyplot as plt import seaborn as
sns import os import warnings
warnings.filterwarnings('ignore')
```

Reading and Understanding the Dataset

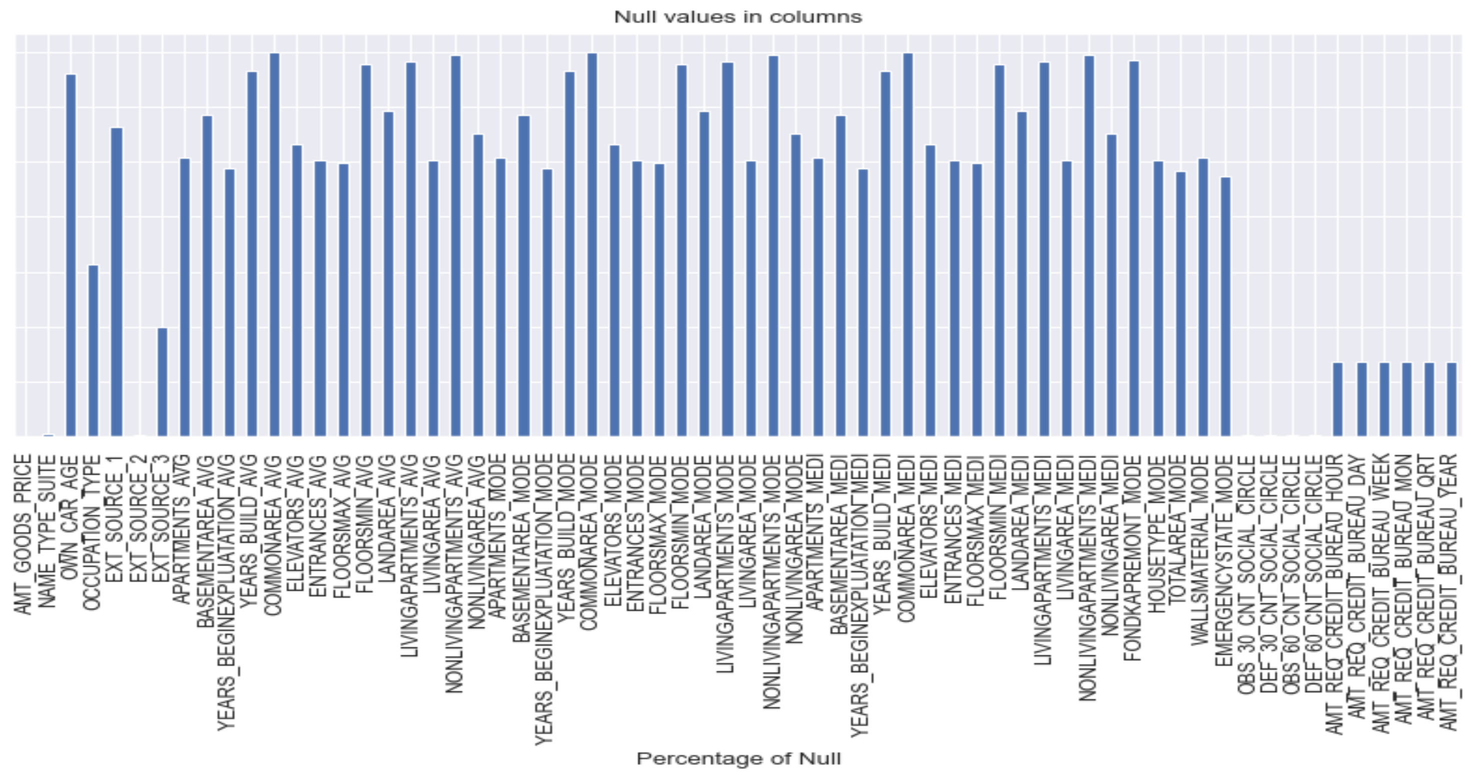
```
PP = pd.read_csv("application_data.csv")
PP.head()
```

Data checking and Missing values

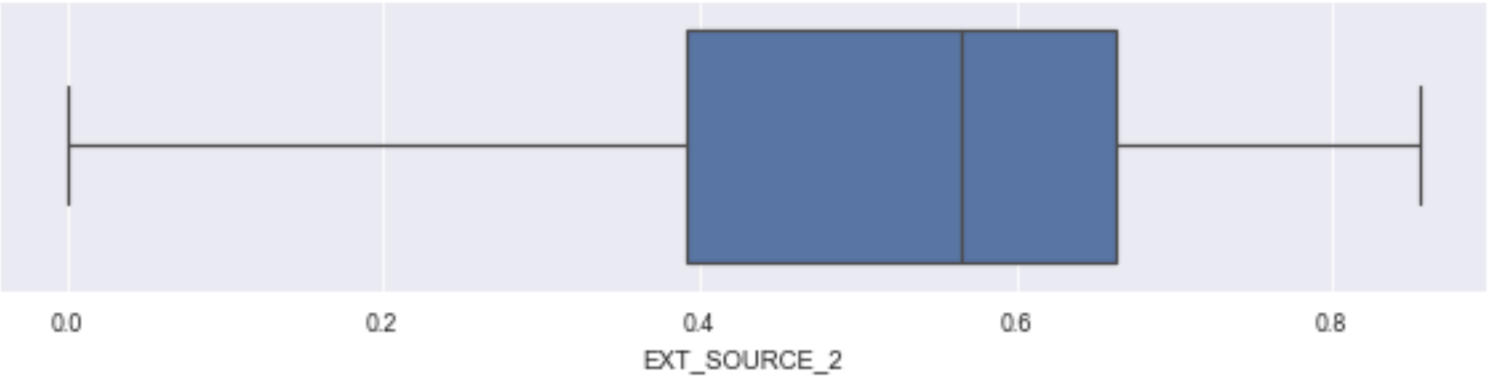
```
# Function to get null value
# Missing values of all columns
# Finding out columns with only null values
# Visualizing Null values of columns in graph
# Taking out columns with >50%
# Columns with null values <15%
# Identifying unique values with columns <15%
```

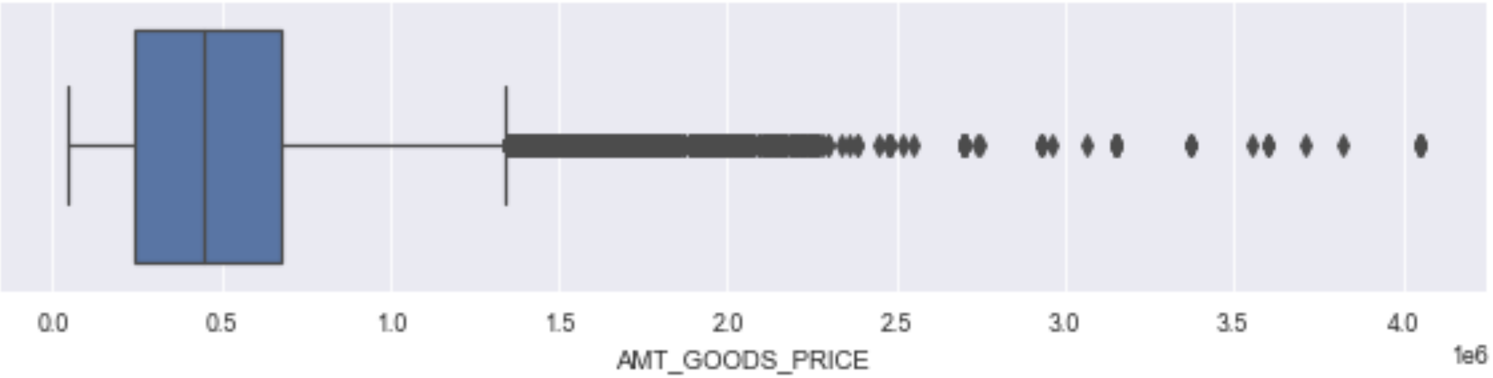
From the above we can see that first two (EXT_SOURCE_2, AMT_GOODS_PRICE) are continuous variables and remaining are categorical variables

# Visualizing Null values of columns in graph



# 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
```

```
# Imputing the value'XNA' which means not available for the column 'CODE_GENDER’
```

XNA values are very low and Female is the majority. So lets replace XNA with gender 'F'

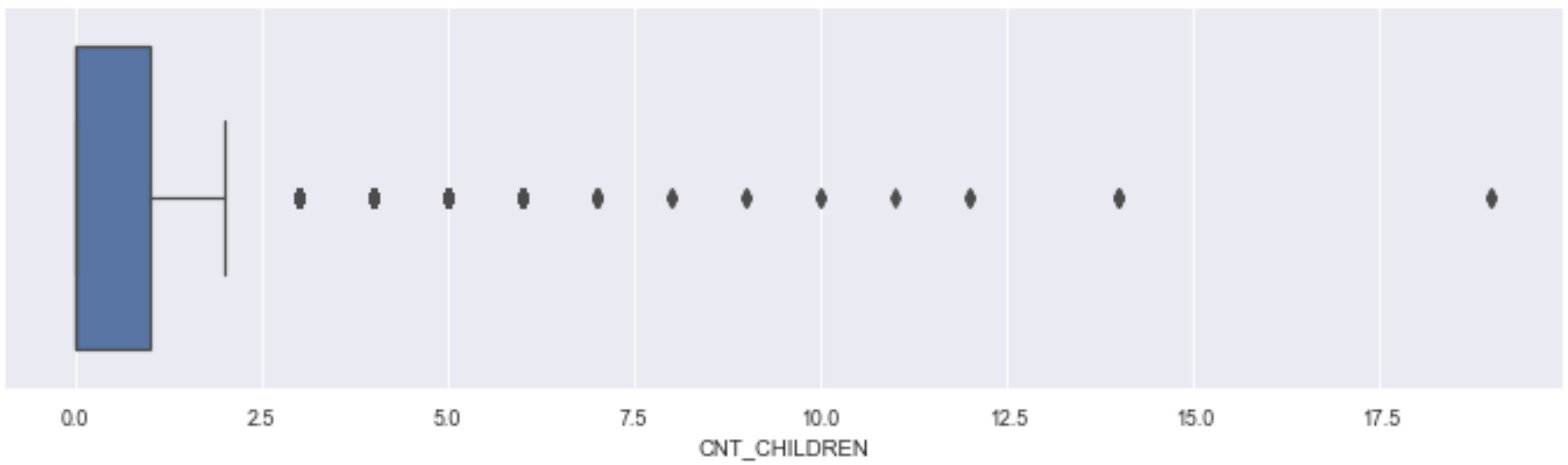
```
# Casting variable into numeric in the dataset
# Age/Days columns are in -ve which needs to be converted to +ve value
# Checking outliers of numerical_column
# Now lets check box plot for 'CNT_CHILDREN',
'AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','DAYS_EMPLOYED',
'DAYS_REGISTRATION’
```

•1st quartile is missing for CNT_CHILDREN which means most of the data are present in the 1st quartile.

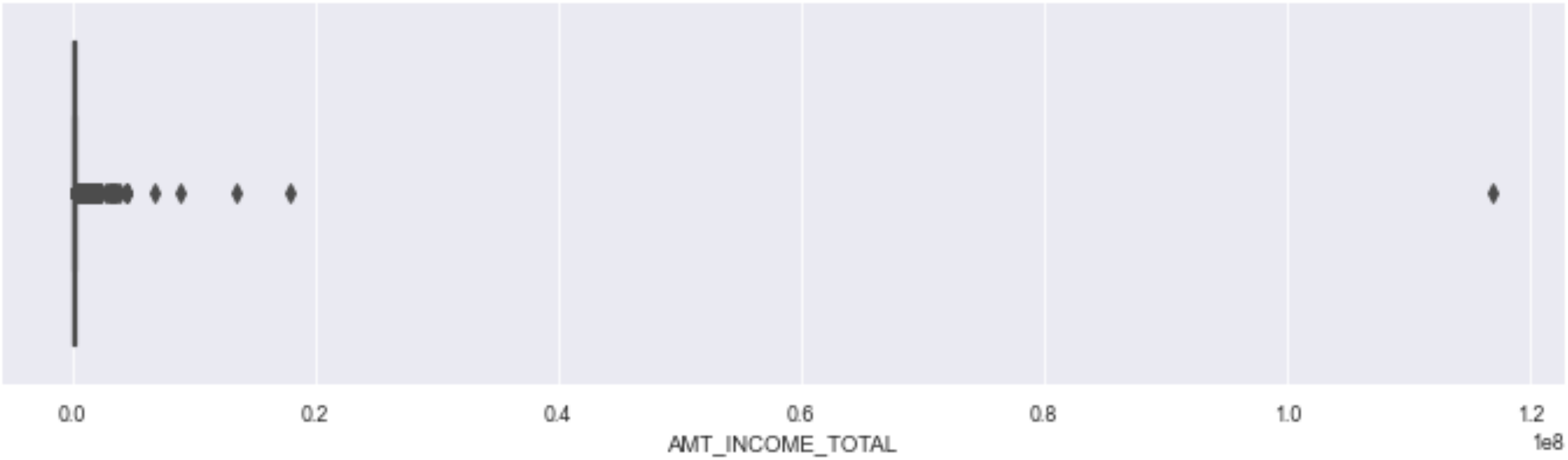
•In AMT_INCOME_TOTAL only single high value data point is present as outlier

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

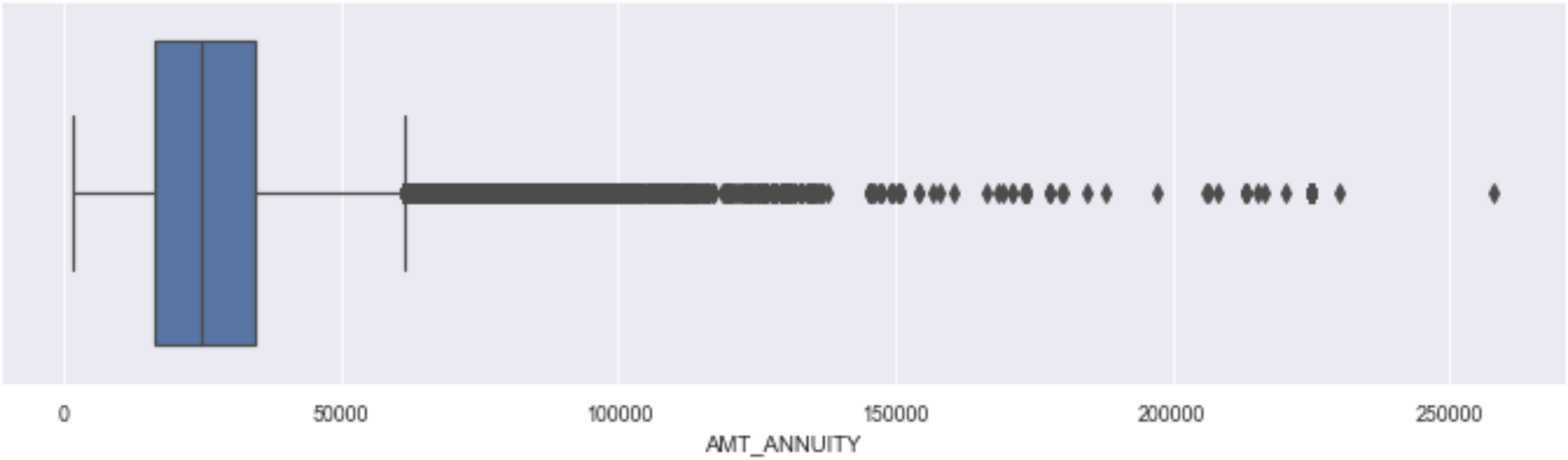
- **1st quartile is missing for CNT_CHILDREN which means most of the data are present in the 1st quartile.**



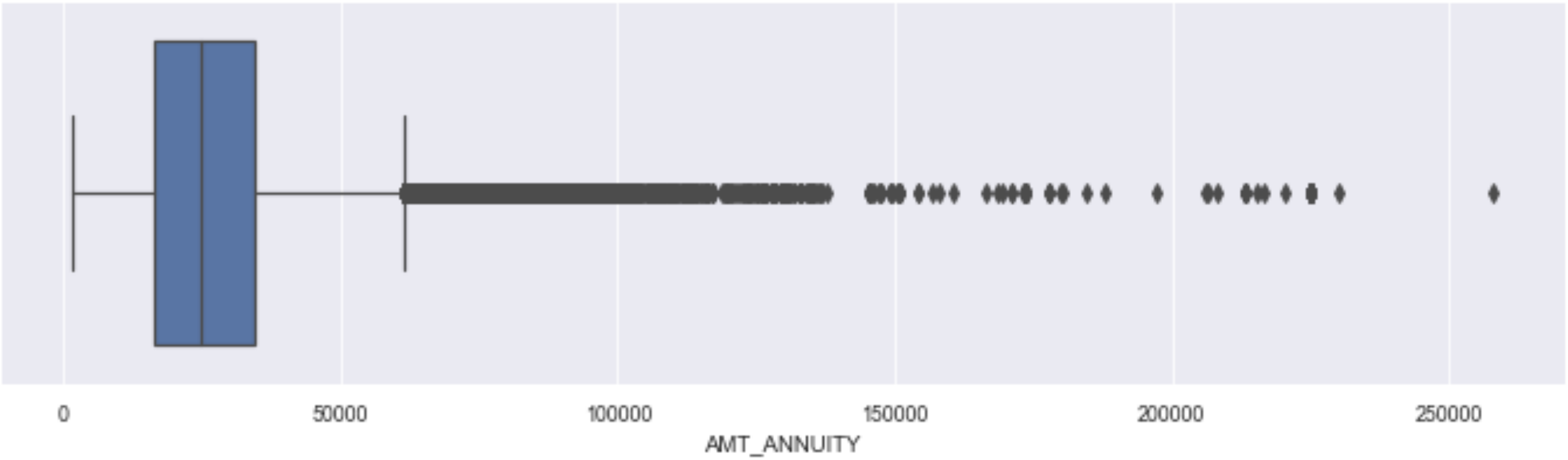
- In AMT_INCOME_TOTAL only single high value data point is present as outlier



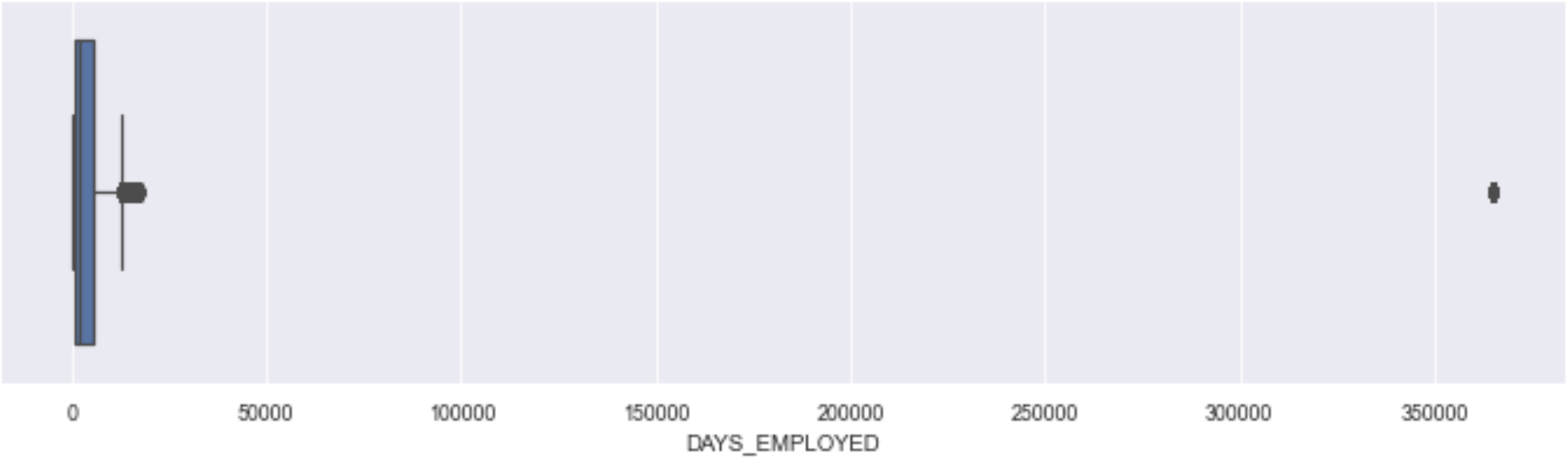
- AMT_CREDIT has little bit more outliers



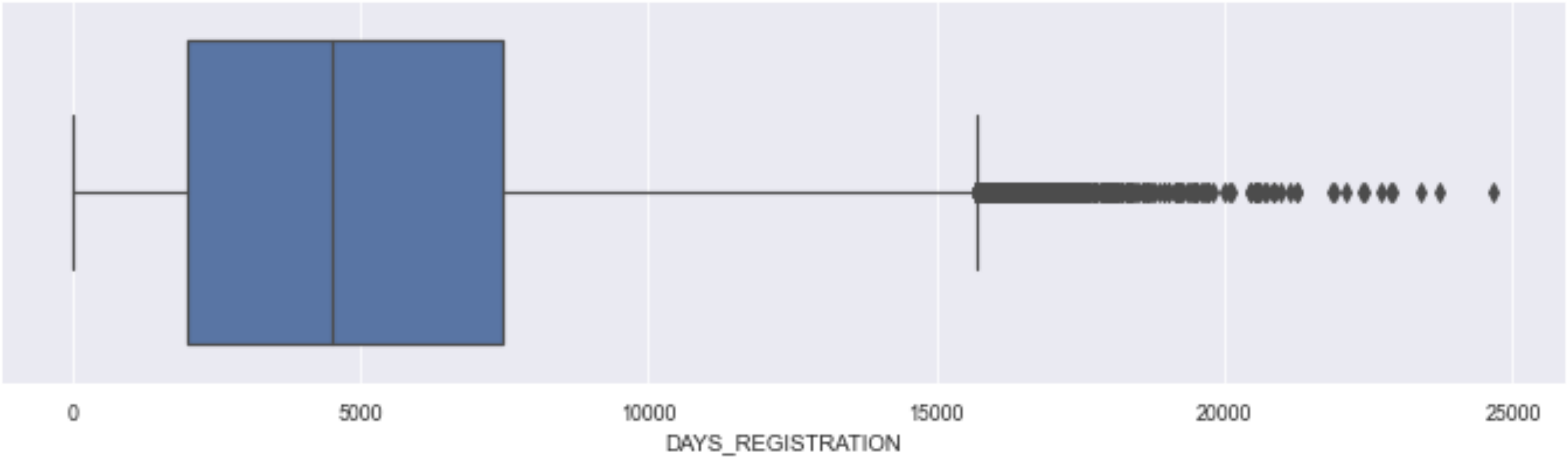
- 1st quartiles and 3rd quartile for AMT_ANNUITY is moved towards first quartile.



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



- **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**



- In AMT_INCOME_TOTAL only single high value data point is present as outlier
- AMT_CREDIT has little bit more outliers
- 1st quartiles and 3rd quartile for AMT_ANNUITY is moved towards first quartile.
- Same with this too 1st quartiles and 3rd quartile for DAYS_EMPLOYED is stays first quartile.

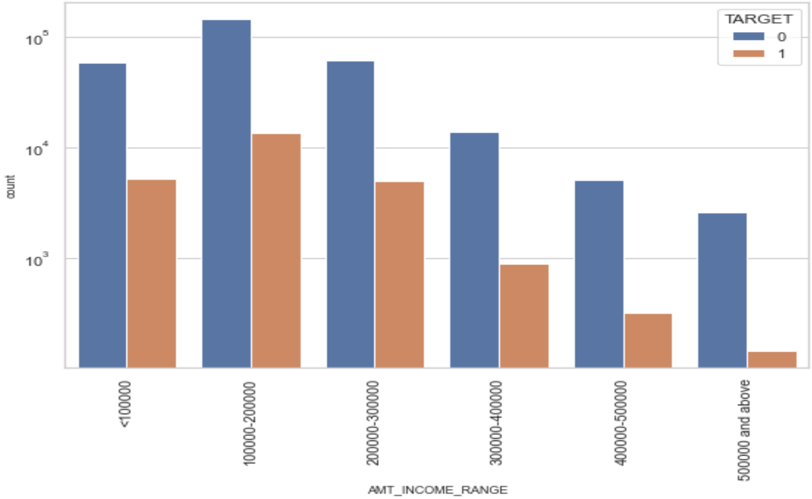
- Same with this too 1st quartiles and 3rd quartile for DAYS_EMPLOYED is stays first quartile.

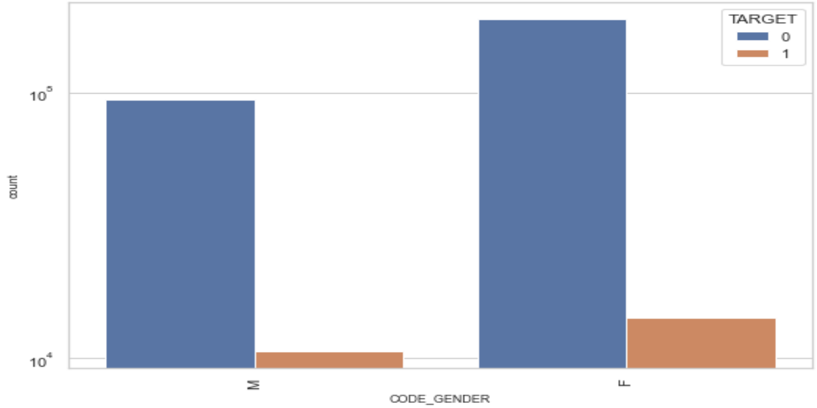
- From above box plots we found that numeric columns have outliers

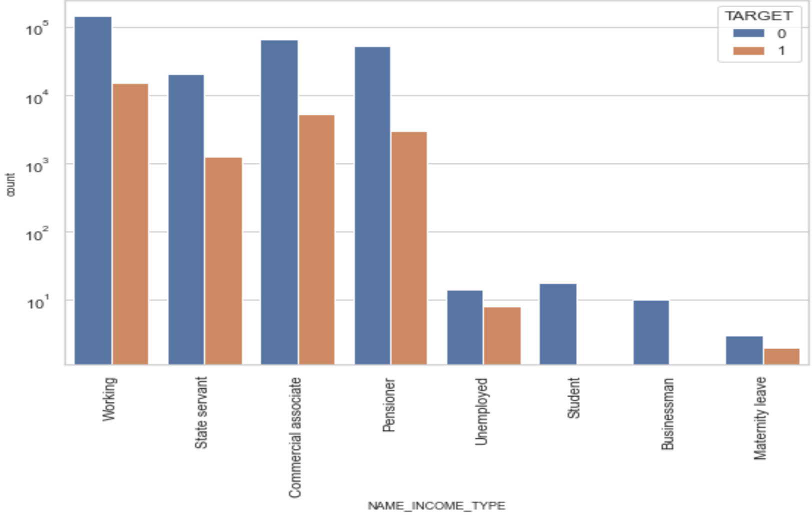
Analysis

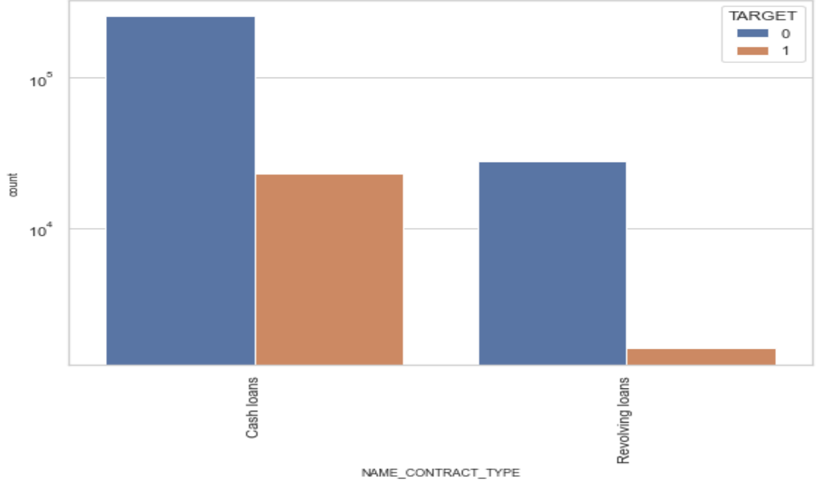
```
# Dividing the dataset into two dataset of
Target=1(client with payment difficulties) and
Target=0(all other)
```

# Univariate Analysis









# Univariate Analysis

#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 servant 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

❑ **#Observation:**

❑ From the above corelation analysis we can say that the highest corelation (1.0) is between (ENTRANCES_MEDI with ENTRANCES_AVG)

Reading Previous_application Dataset

```
# Reading the previous_application csv file

preapp = pd.read_csv('previous_application.csv’)
preapp.head()

# Number of rows and columns in previous application data frame
preapp.shape

 # Knowing the previous application data frame info
 preapp.info()

 # describing the previous application data frame
 preapp.describe()

 # Finding out null values
 Nu_col = null_percentage(preapp)
 Nu_col.head()
```

```
# Removing null values >0

# Now removing null values <50
```

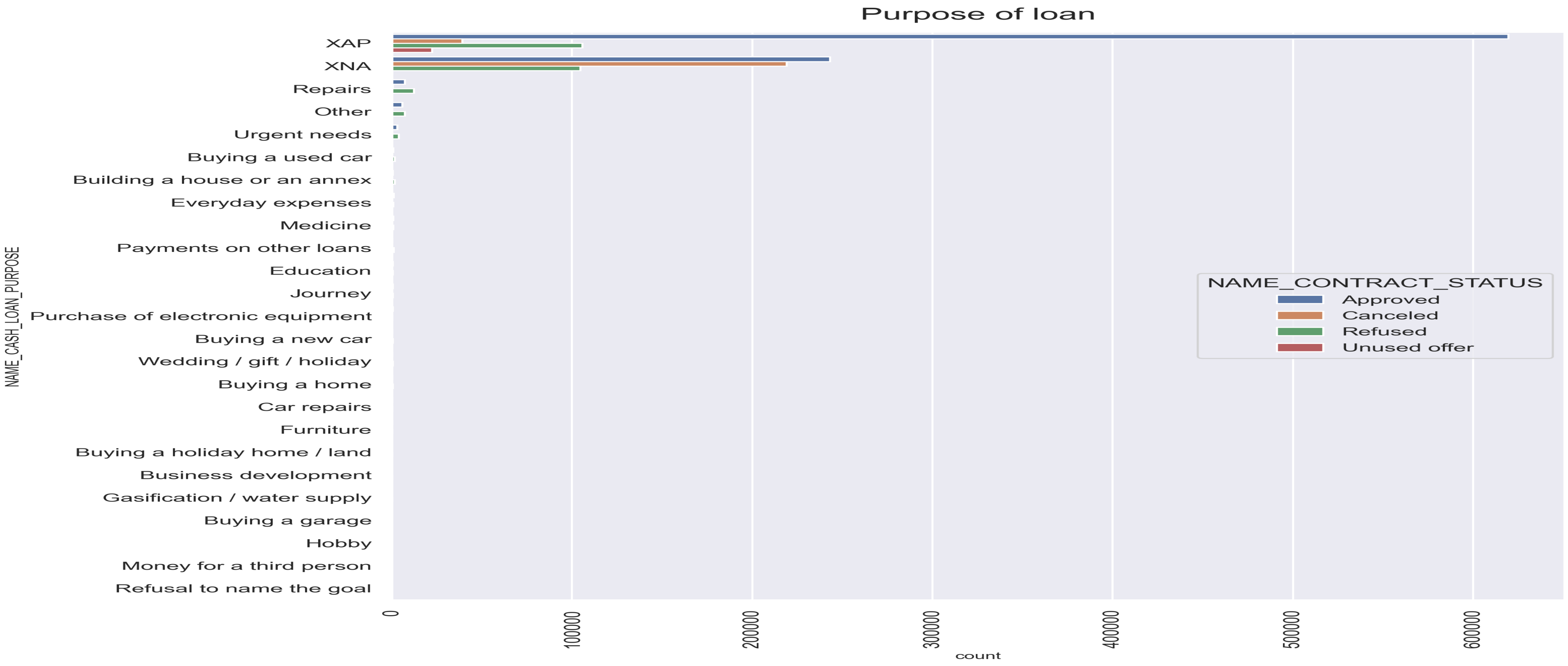
```
# Merging both the dataframes (application_data, previous_application)
```

```
# Performing univariate analysis
```

# Purpose of loan

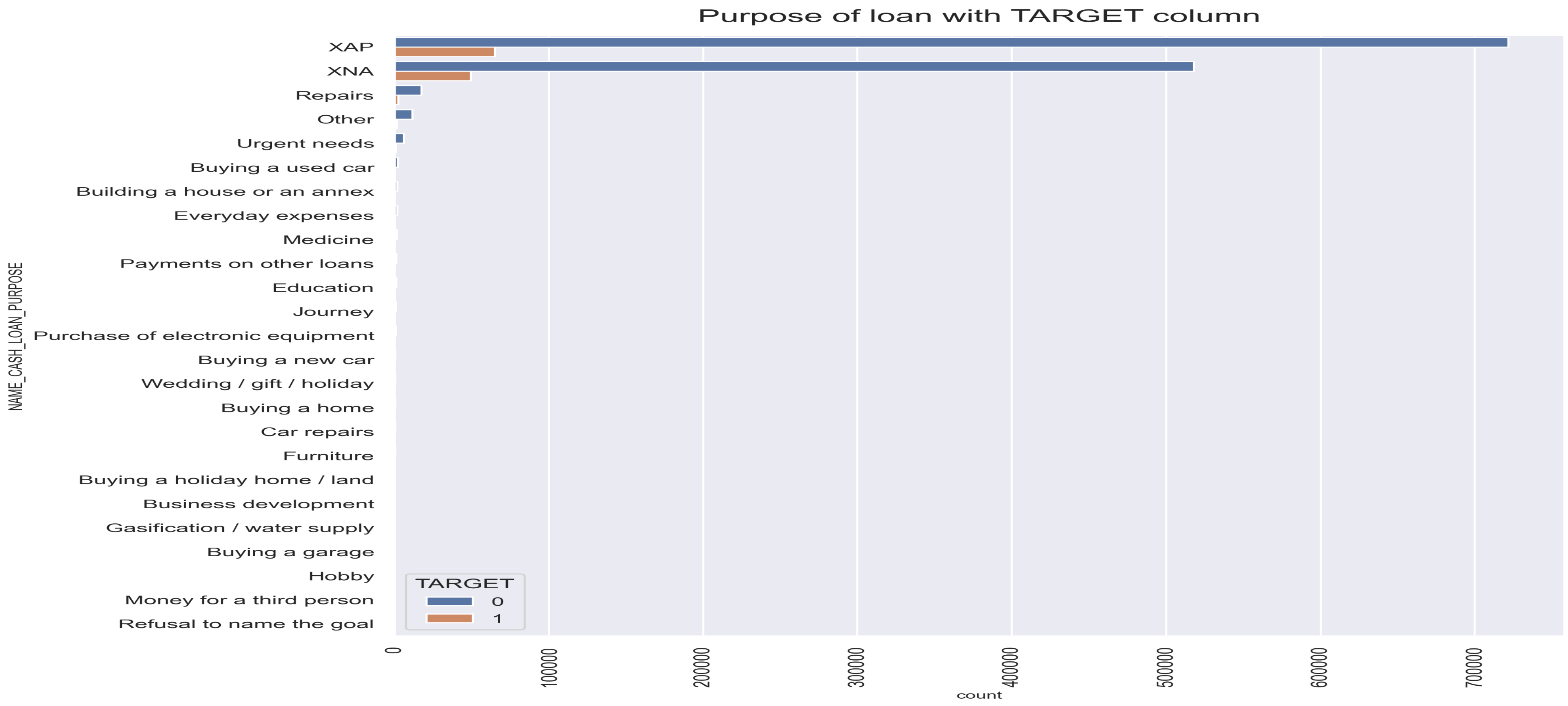
❑ Observation

❑ - Most of loan rejection was from 'repairs'



# Purpose of loan with TARGET column

- ❑ Observation
- ❑ - Most of loan rejection was from 'repairs'



Conclusion from the Analysis

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

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