Bank Loan Default Risk Analysis

Introduction

This case study aims to give an idea about applying EDA in real business scenario. In this case study we will 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.

Abstract of Risk factors associated with bank decision

The project Bank Loan Default Risk Analysis is about predicting how much amount of loan should be disbursed based on the attributes of person's income, property (if he is holding any).

Apart from the above said the main factors that impact are will the person be **capable of repaying the loan amount or if not how much will be able to repay based on income, or will he default.**

These factors which can be utilised this knowledge for its portfolio and risk assessment

Algorithm

An algorithm is an step by step sequential procedure which creates a structure of handling tasks or projects given

- Getting Jupyter Ready
- Reading and Understanding Data
- Data Cleaning and Manipulation
- Data Analysis
- Conclusions

Getting jupyter Ready : Importing Libraries

```
In [24]: #importing required packages

         import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         import matplotlib.style as style
         import seaborn as sns
         import itertools
         %matplotlib inline

         #setting up plot style
         style.use('seaborn-poster')
         style.use('fivethirtyeight')
```

Reading and Understanding data

Importing data and reading using pandas, understanding metrics like shape, informations, objects, if there any null values

Selecting relevant attributes from data and performing basic operations

Reading and Understanding data

Reading and Understanding Data

In [27]:
```
#Importing the input files

import csv
import os
for dirname, _, filenames in os.walk('/Desktop/IT'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

In [28]:
```
Application_Data = pd.read_csv(r'C:\Users\Venkat RC\Desktop\IT\Naresh IT\resu
Application_Data
```

In [28]:
```
Application_Data = pd.read_csv(r'C:\Users\Venkat RC\Desktop\IT\Naresh IT\resu
Application_Data
```

Out[28]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	F
0	100002	1	Cash loans	M	N	
1	100003	0	Cash loans	F	N	
2	100004	0	Revolving loans	M	Y	
3	100006	0	Cash loans	F	N	
4	100007	0	Cash loans	M	N	
...	...	...	...	...	...	
307506	456251	0	Cash loans	M	N	
307507	456252	0	Cash loans	F	N	
307508	456253	0	Cash loans	F	N	
307509	456254	1	Cash loans	F	N	
307510	456255	0	Cash loans	F	N	

307511 rows × 122 columns

```
Database size - Application_Data          : 37516342
```

In [31]:
```python
#Database column types
Application_Data.info(verbose=True)
```

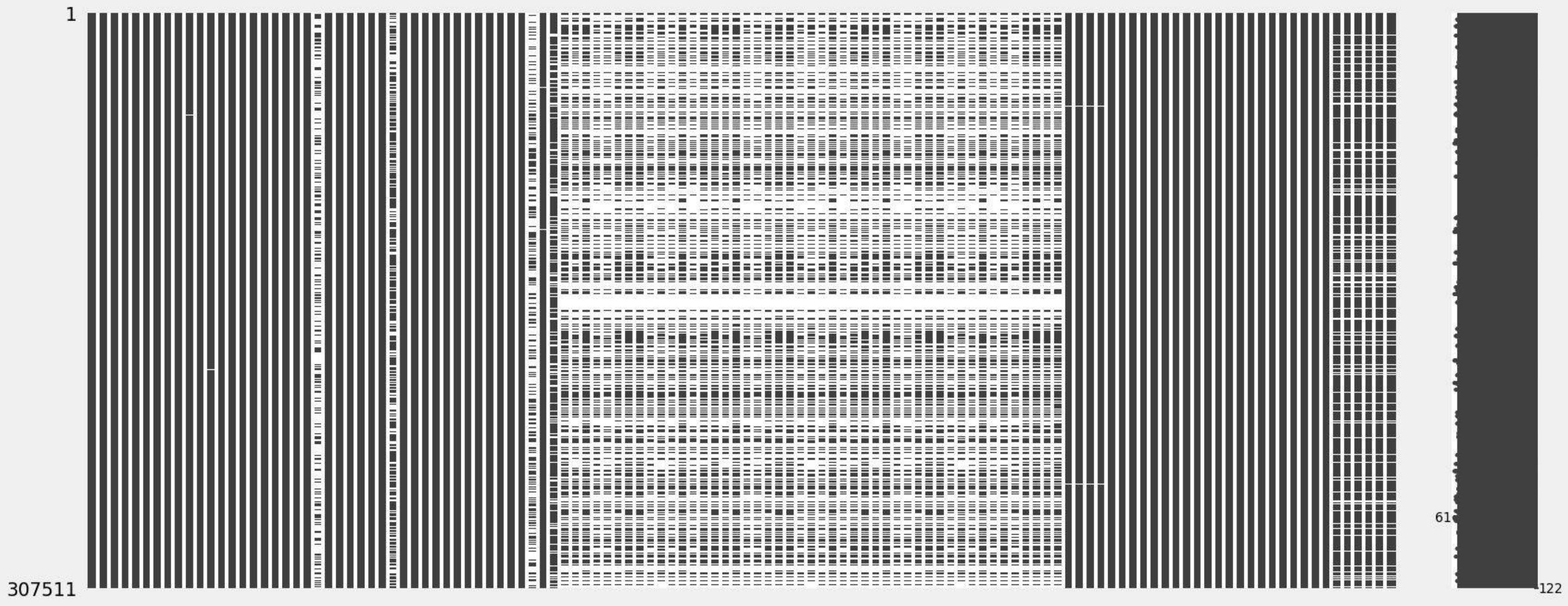
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 122 columns):
 #    Column                       Dtype
---   ------                       -----
 0    SK_ID_CURR                   int64
 1    TARGET                       int64
 2    NAME_CONTRACT_TYPE           object
 3    CODE_GENDER                  object
 4    FLAG_OWN_CAR                 object
 5    FLAG_OWN_REALTY              object
 6    CNT_CHILDREN                 int64
 7    AMT_INCOME_TOTAL             float64
 8    AMT_CREDIT                   float64
 9    AMT_ANNUITY                  float64
 10   AMT_GOODS_PRICE              float64
 11   NAME_TYPE_SUITE              object
 12   NAME_INCOME_TYPE             object
 13   NAME_EDUCATION_TYPE          object
```

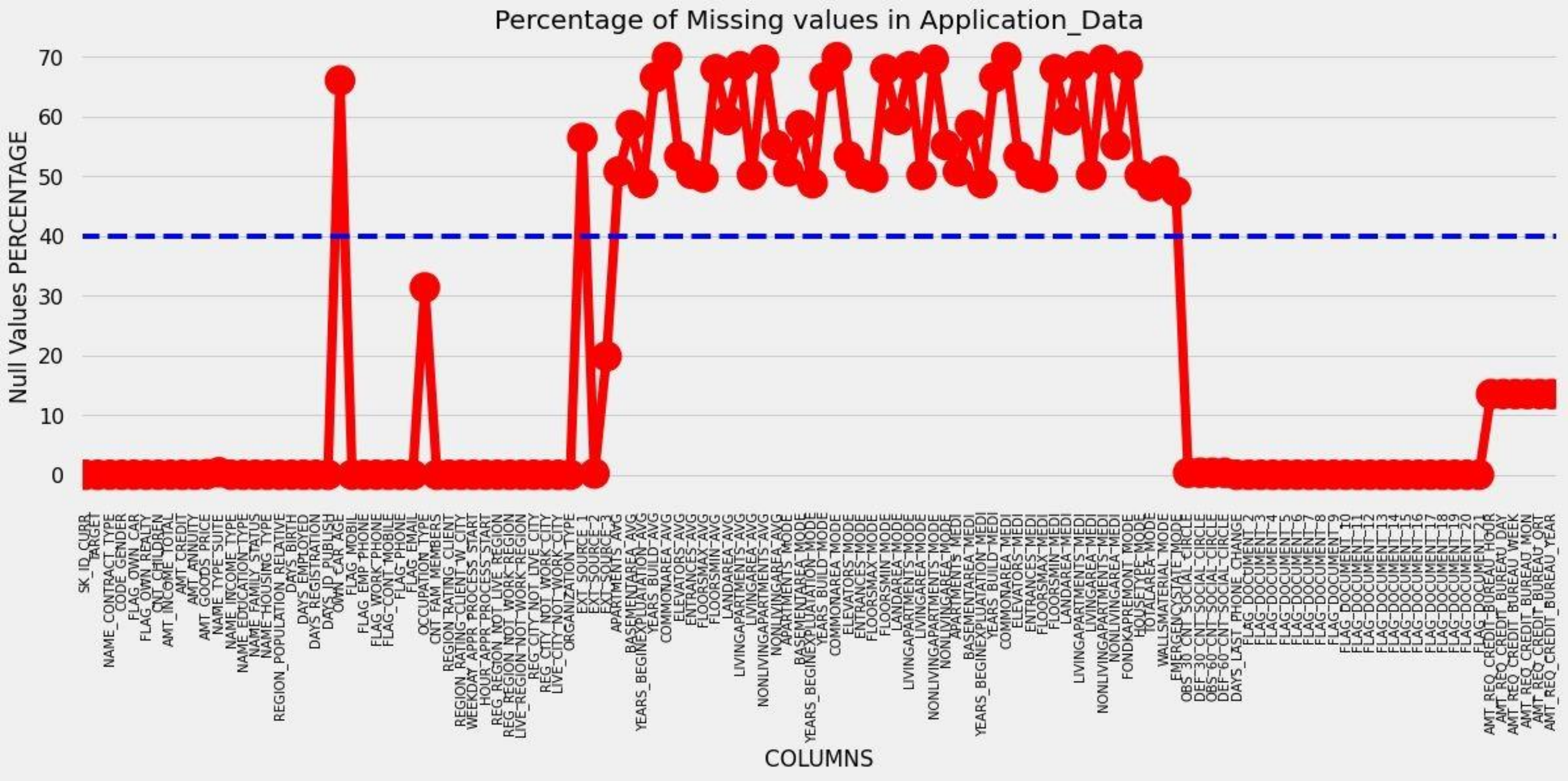
In [32]:
```python
#Checking the numeric variables of the dataframes
Application_Data.describe()
```

Out[32]:

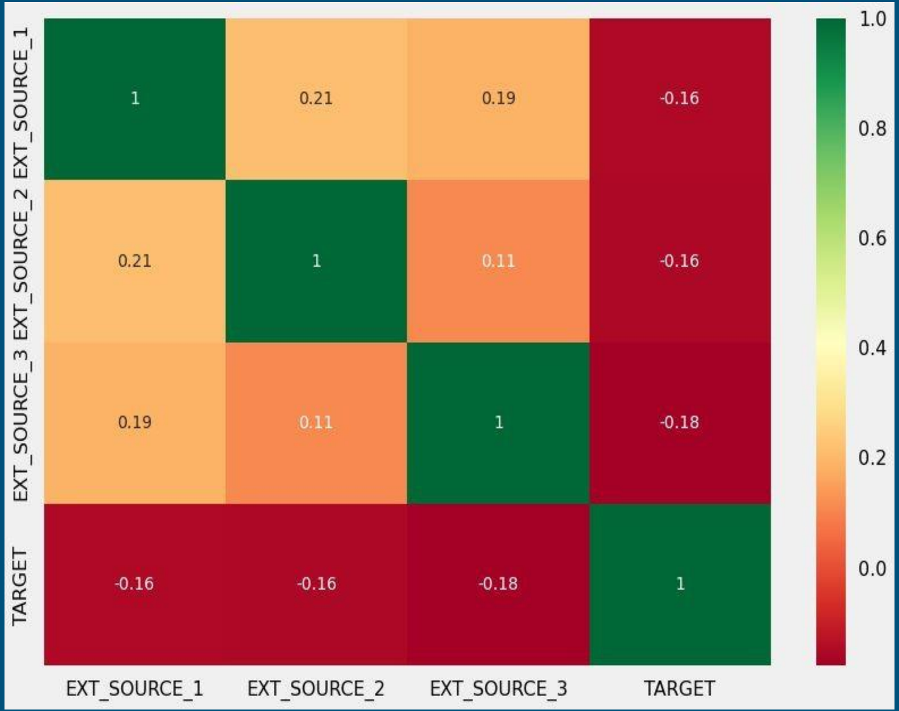
	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AM
count	307511.000000	307511.000000	307511.000000	3.075110e+05	3.075110e+05	30
mean	278180.518577	0.080729	0.417052	1.687979e+05	5.990260e+05	2
std	102790.175348	0.272419	0.722121	2.371231e+05	4.024908e+05	1
min	100002.000000	0.000000	0.000000	2.565000e+04	4.500000e+04	
25%	189145.500000	0.000000	0.000000	1.125000e+05	2.700000e+05	1
50%	278202.000000	0.000000	0.000000	1.471500e+05	5.135310e+05	2
75%	367142.500000	0.000000	1.000000	2.025000e+05	8.086500e+05	3
max	456255.000000	1.000000	19.000000	1.170000e+08	4.050000e+06	25

Data Cleaning and Manipulation

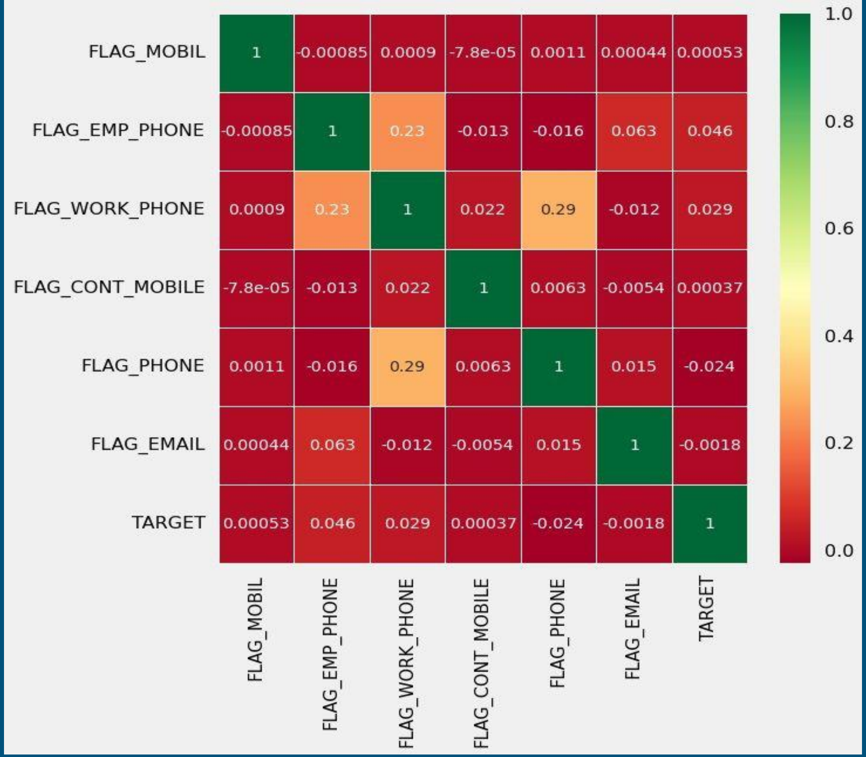



Percentage of Missing values in Application_Data
Null Values PERCENTAGE
70
60
50
40
30
20
10
0
COLUMNS

Analyse and Delete Unnecessary columns



Correlation

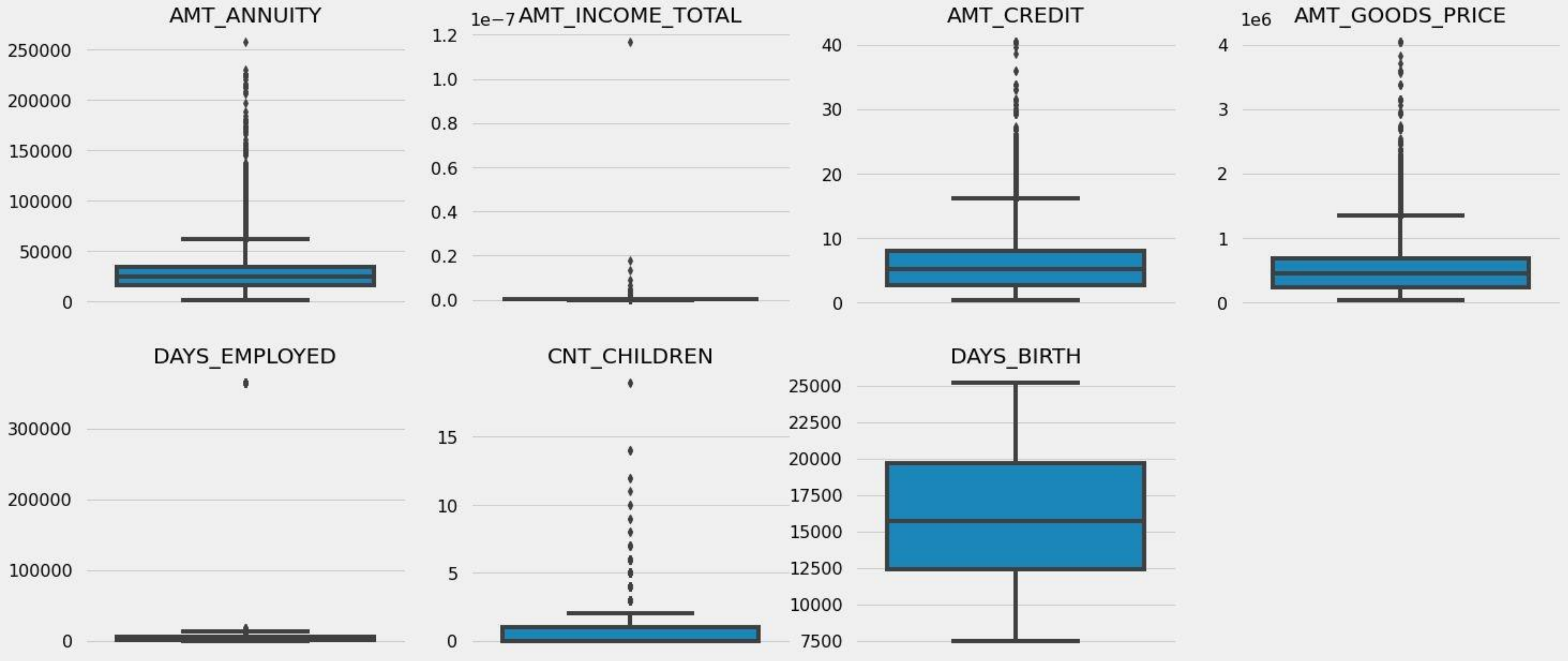


Insights:

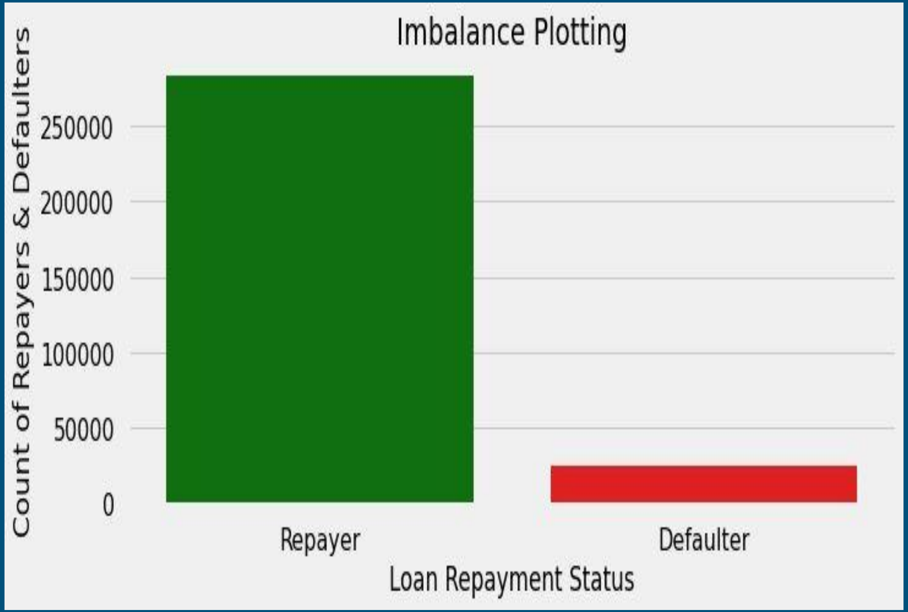
In previous slide: Based on the heatmap, we can say that EXT_SOURCE_1 has 56% null values, where as EXT_SOURCE_3 has to close to 20% null values

There is no correlation between flags of mobile, email etc with loan repayment thus these columns can be deleted.

Outlier detection



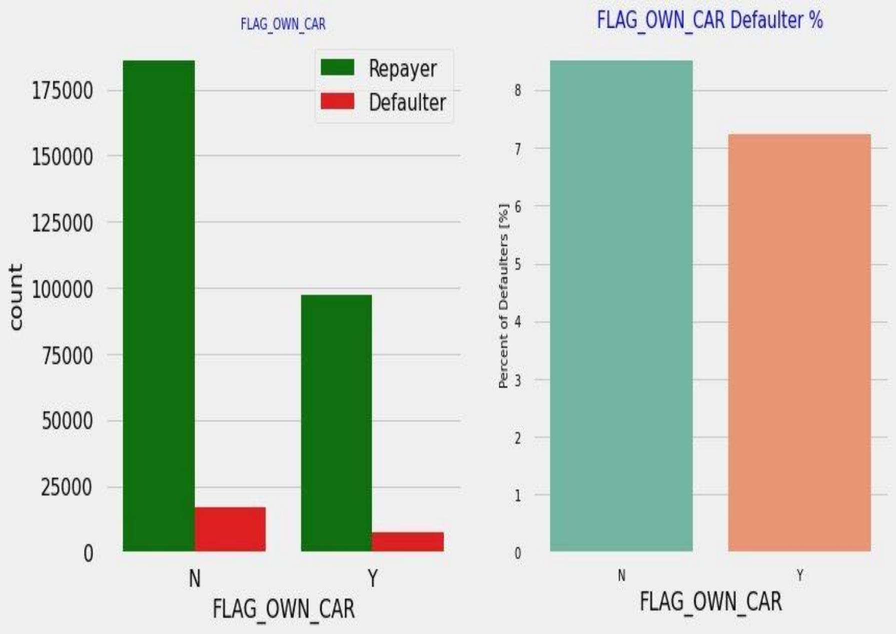
Data Analysis



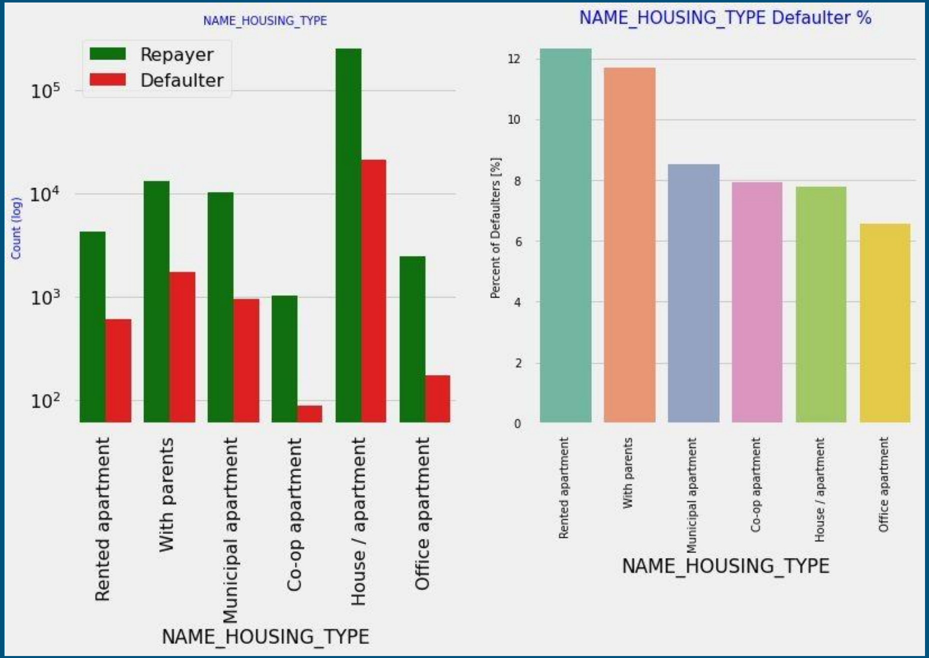
Inference:

Contract type: revolving loans are just a small fraction (10%) from the total number of leans in the same time a larger amount of revolving loans, comparing with their frequency, are not repaid.

Categorical variable analysis

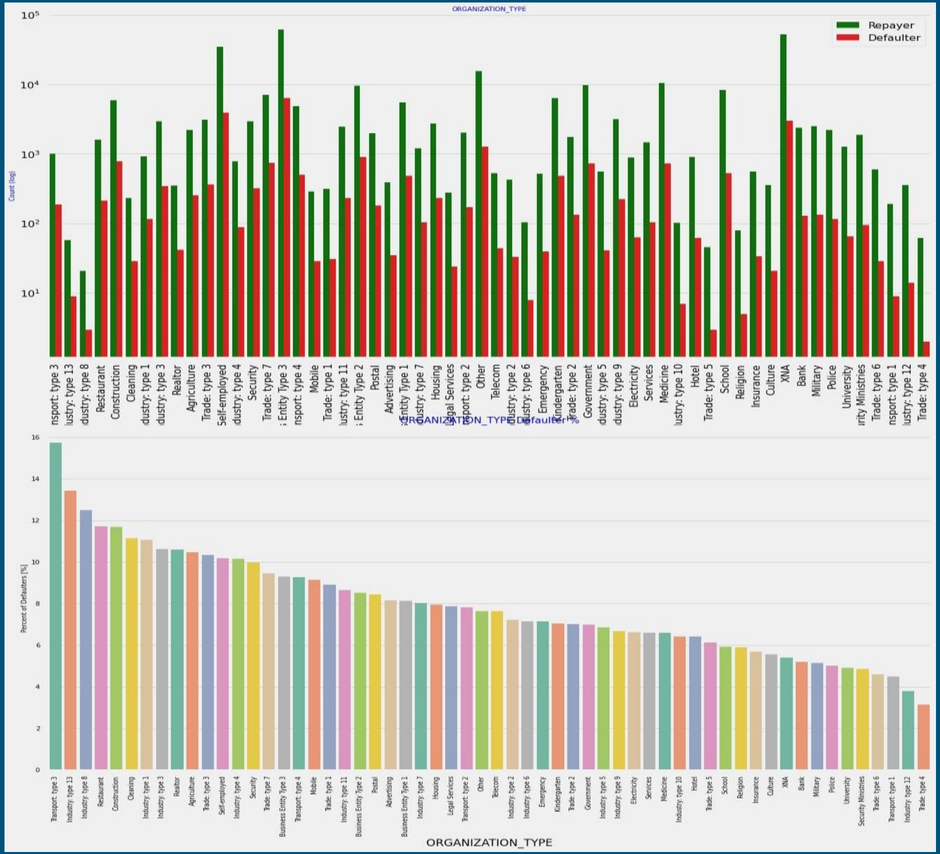


The number of female clients is almost double the number of male clients, based on the percentage of defaulted credits, males have a higher chance of not returning their loans (~10%), comparing with women (~7%)



Viewpoints:

Majority of people live in House/Apartment , whereas people living in office apartments have lowest default rate while people living with parents (~11.5%) and living in rented apartments (>12%) have higher probability of defaulting

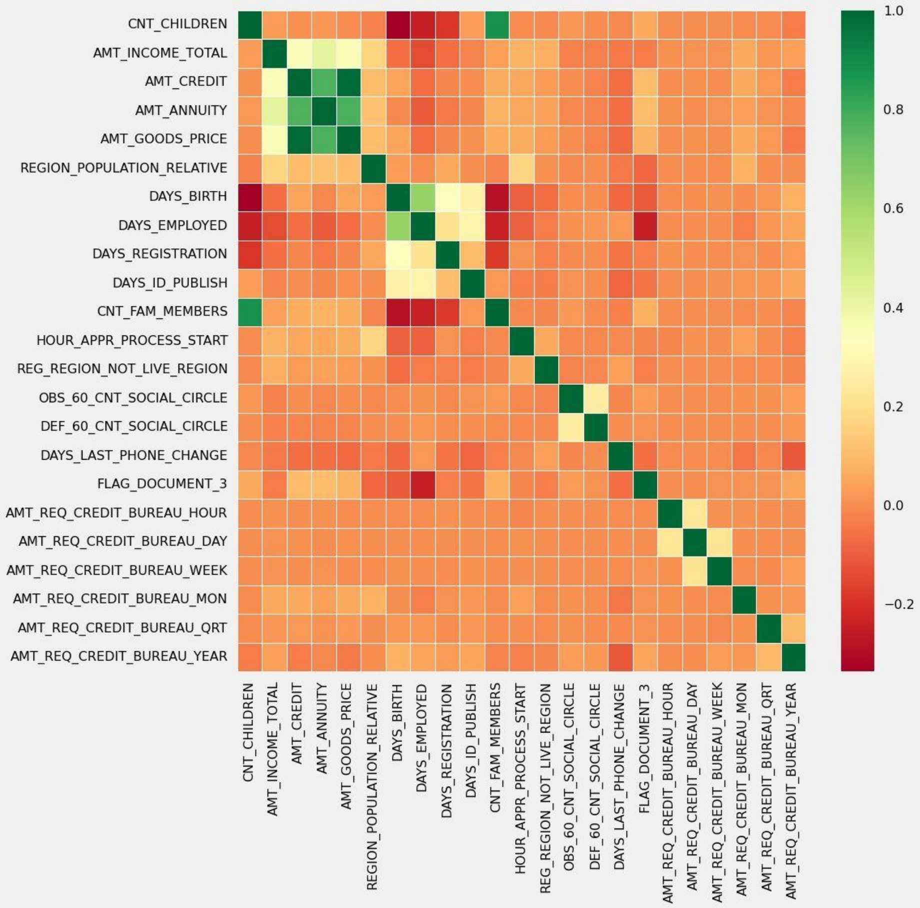


View Points:

The companies with the largest percentage of loans that have not been repaid fall into the following categories: Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed individuals have a comparatively high default rate, so they should be avoided when applying for loans or given loans with higher interest rates to reduce the chance of default.

The majority of those that apply for loans are from Business Entity Type 3

Organisation type information is unavailable (XNA) for a very large number of applications.



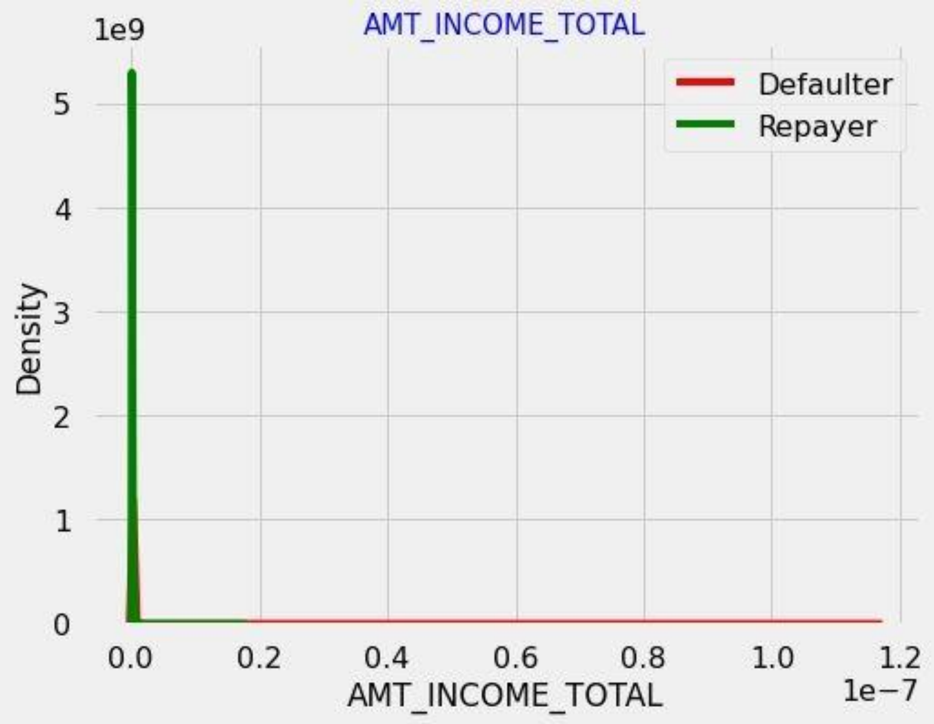
View points:

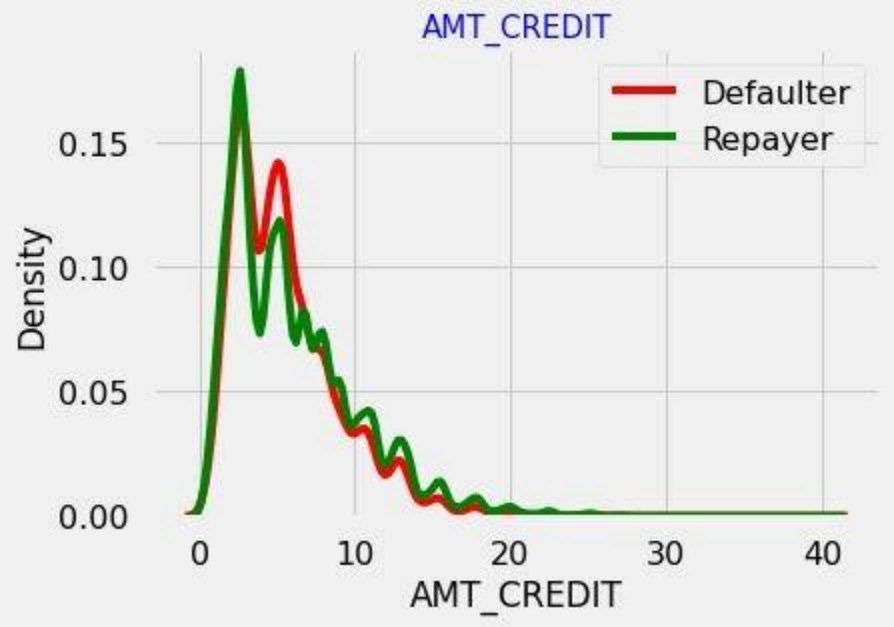
Correlating factors amongst repayers:

Credit amount is highly correlated with amount of goods price, loan annuity, total income

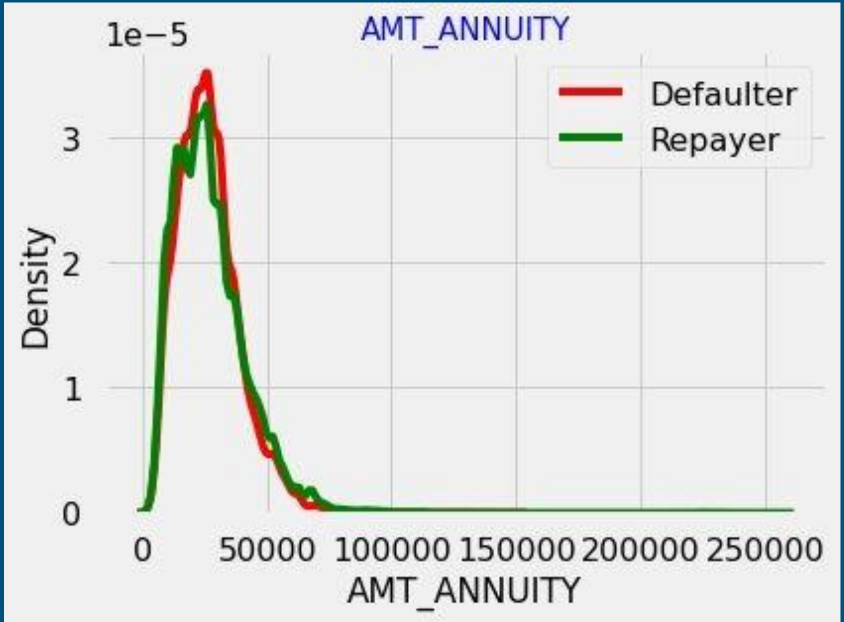
We can also see that repayers have high correlation in number of days employed.

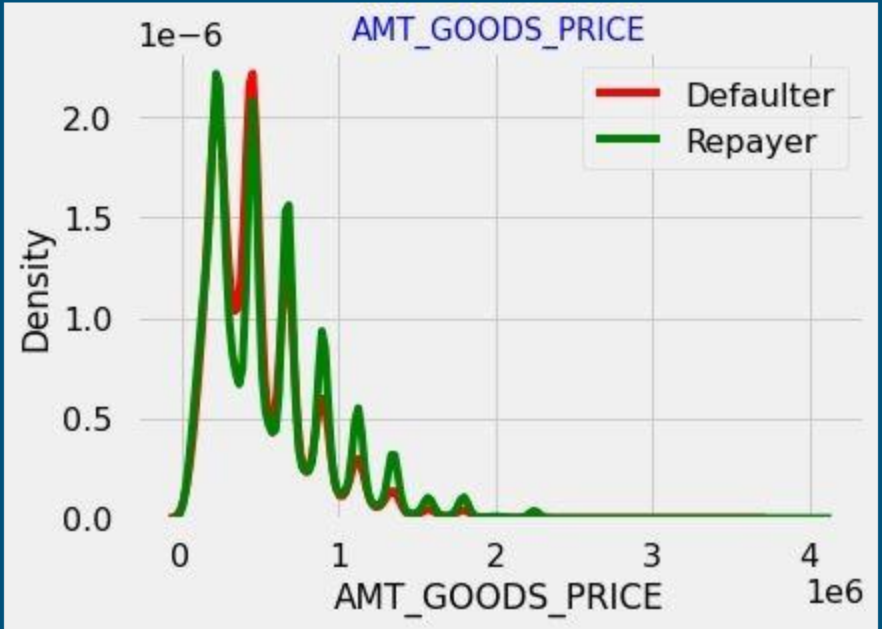
Numerical Univariate Analysis



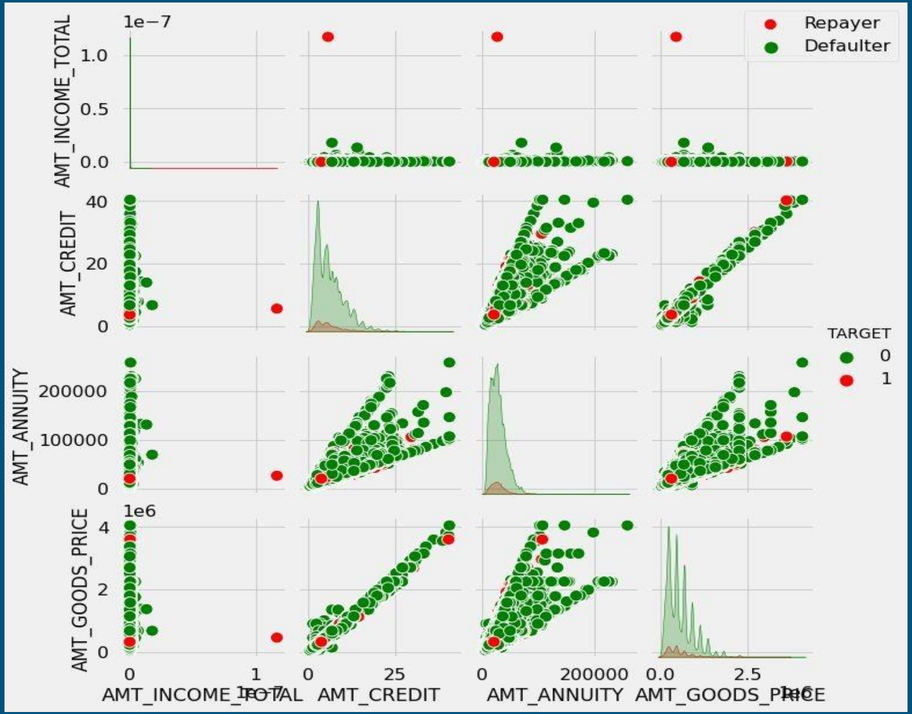


Numerical Univariate analysis





Pair plot



View points

When amy_annuinty is greater than 15000 amt_goods_price > 3M, there is a lesser chance of defaulters

AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line

There are very less defaulters for AMT_CREDIT > 3M

Suggestions

After analyzing datasets, there are few attributes of a client with which the bank would be able to identify if they will repay the loan or not.

90% of the previously cancelled client have actually replayed the loan.Keep track of the cause so that the bank can later decide and negotiate conditions with those who are paying their debts in order to expand its commercial opportunities.

88% of the clients whose loan requests had previously been denied by the bank are now reimbursing clients. Documenting the cause for rejection could therefore help to reduce the company loss and allow for additional loan requests from these clients.

Thank You

Durgesh Kumar Nandan