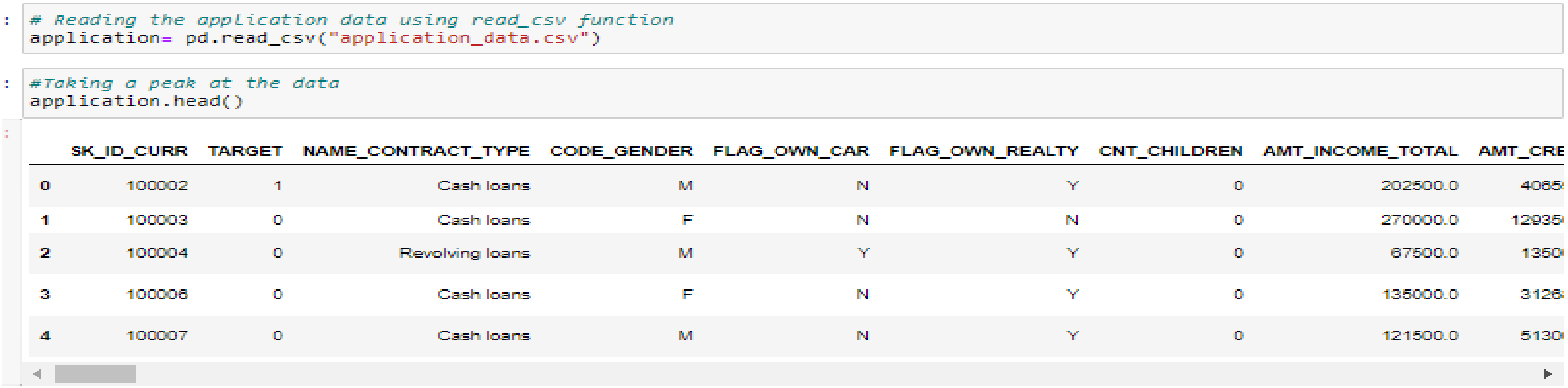
upGrad

Credit EDA Case Study

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 we have learnt in the EDA module, 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.

Reading the Data

- Reading the application data using read_csv function

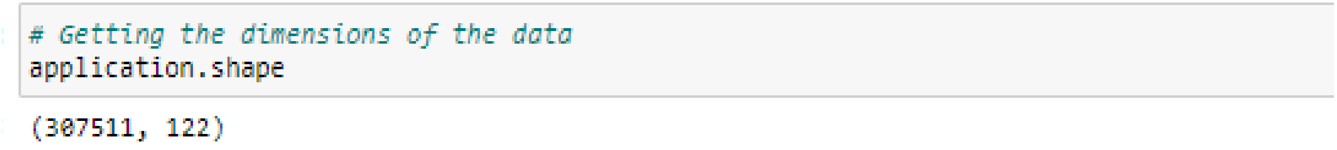


```
# Reading the application data using read_csv function
application= pd.read_csv("application_data.csv")
```

```
#Taking a peak at the data
application.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CRE
0	100002	1	Cash loans	M	N	Y	0	202500.0	4065
1	100003	0	Cash loans	F	N	N	0	270000.0	12935
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	1350
3	100006	0	Cash loans	F	N	Y	0	135000.0	3126
4	100007	0	Cash loans	M	N	Y	0	121500.0	5130

- Inspection of the application Data



```
# Getting the dimensions of the data
application.shape
```

```
(307511, 122)
```

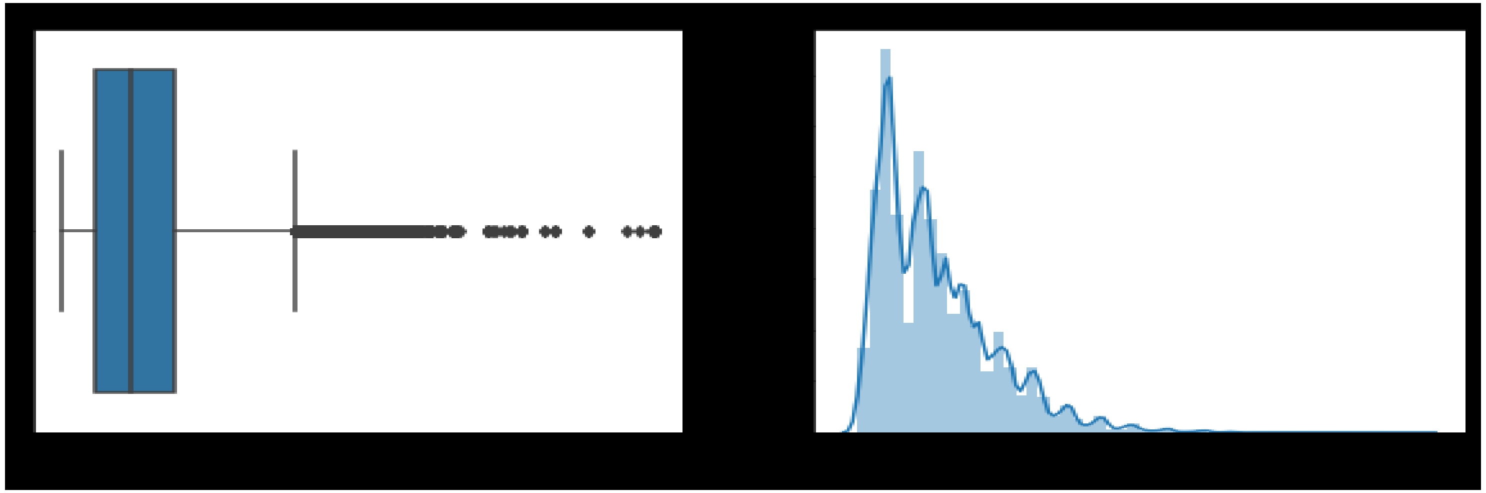
Dealing With Missing Data

- We had many dcolumns with % Null values more thann 47%
- We took the decision of dropping those columns.
- Yet some columns had missing data with null %more than 13.
- To those remaining 13% we have given suggestions regarding on how to fill the data. Some with median, some with 0 and some with mean.

Treating columns with wrong Data Type or Data

- Some columns were filled with “XNA” values as strings. Although they were null values. So we changed the data of such columns
- Columns such as DAYS_REGSTERATION, AMT_REQ_CREDIT_BUREAU and many more had wrong data types. They were changed.

Outlier Detection



The above graph is for column:AMT_CREDIT.

- As there are few outliers in the data, we can go ahead and deal it by dropping those rows. Rows with Values more than 16,00,000 rupees can dropped.

Similarly many other conclusions can be drawn from other columns. Few are given below.

- There were many outliers in AMT_INCOME_TOTAL .The IQR clculated is 90000. If we would want to fix them , we can go ahead by dropping the values more than 4,00,000 rupees. As only 2% of the data is more than that value.
- :In AMT_ANNUITY column we can observe that there are few outliers. Values more than 60,000 rupees are outliers. Since only 2% of the data is more than that, we can go ahead in dropping them.

Analysis

Our target variable is defined as:

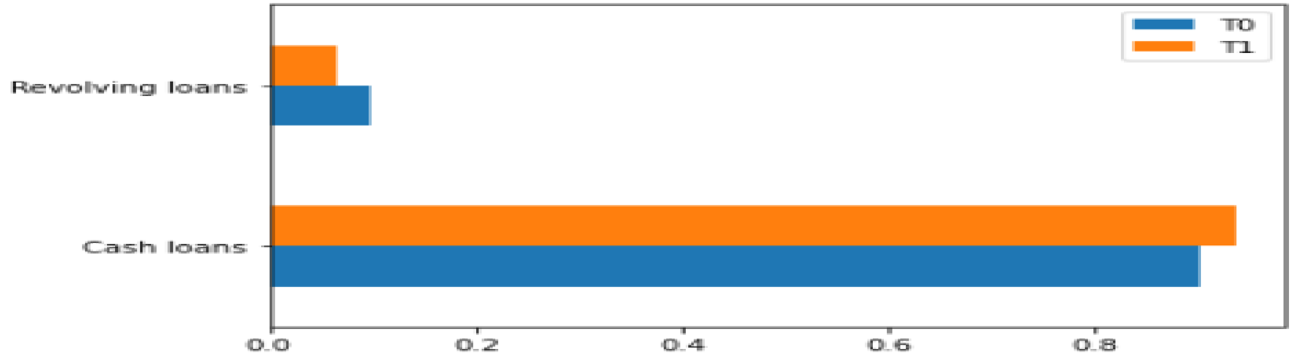
- 1: If the client faces difficulty in paying any installment of the loan. (Only 10% of people in the data are not able to repay the loan on time)
- 0: If the client is able to pay all the installments. (90% of the people in the data are able pay the loan on time)

We divided the data into two parts,

- T1: in this dataframe, target variable will be 1 i.e data of all the defaulters
- T0: in this dataframe, target variable wil be 0 i.e data of all the non-defaulters.

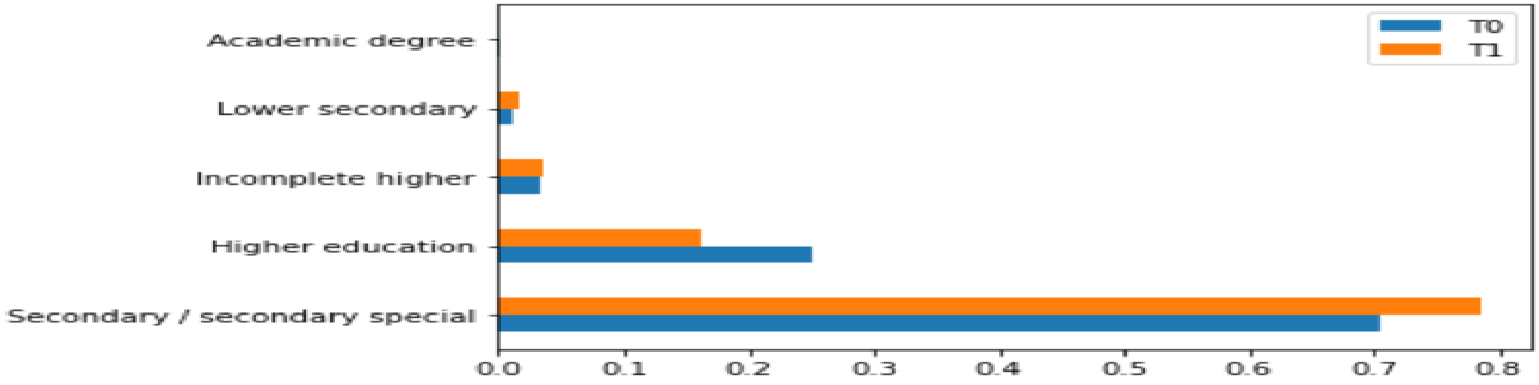
Univariate Analysis

NAME_CONTRACT_TYPE



- 93% of non-Defaulters could repay Cash loan and 7% of them could repay Revolving loan
- We can see that 90% of defaulted people defaulted in Cash loan and 10% defaulted in Revolving loan

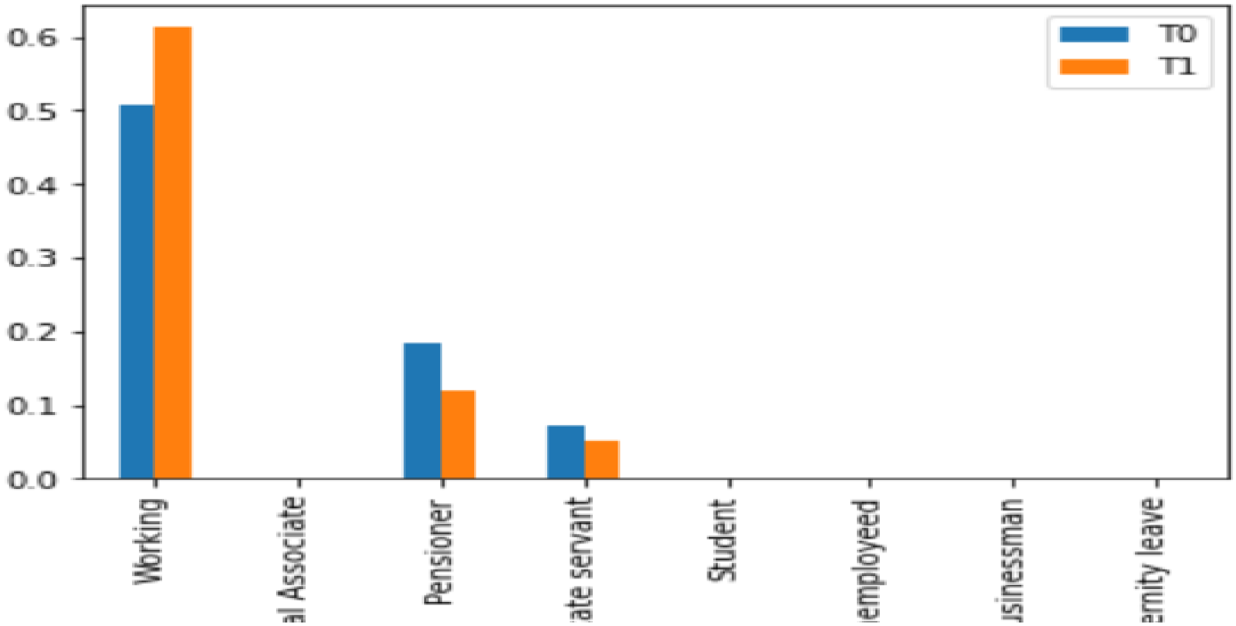
NAME_EDUCATION_TYPE



- One can see that the proportion of people failing to pay the loan is lower than the proportion of people who are able to with higher education.

upGrad

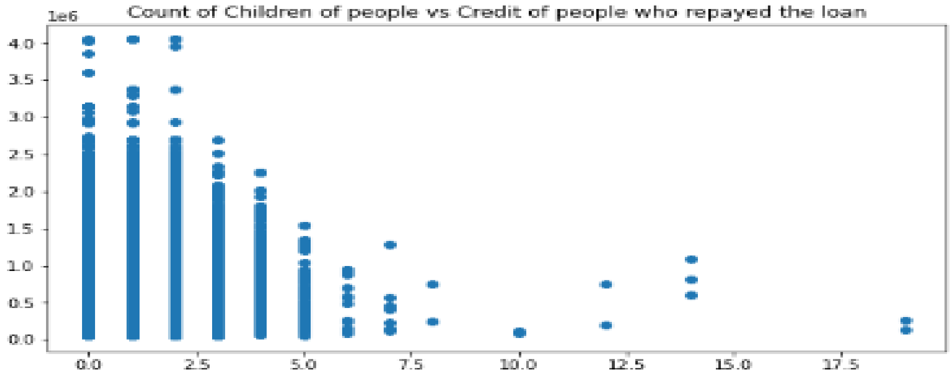
NAME_INCOME_TYPE

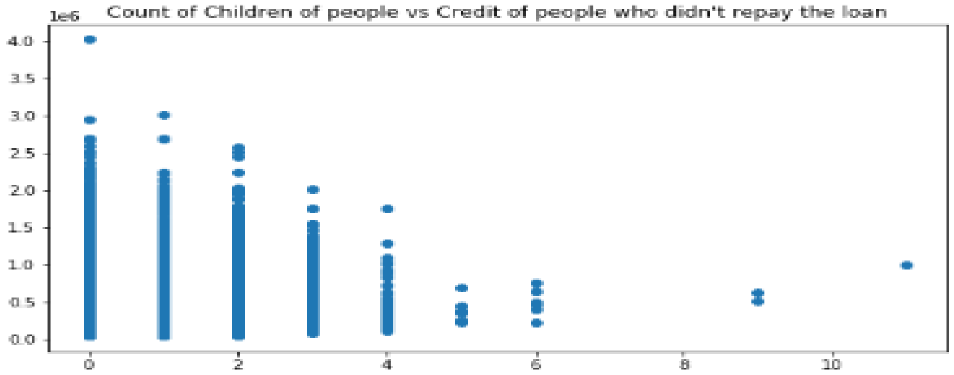


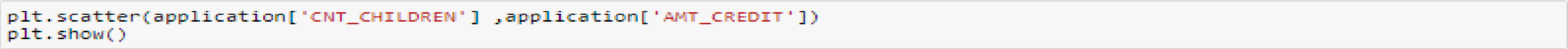
- 61% of defaulters are from working category,12% are from pensioner category ,21% are from commercial associate.
- 50% of non defaulters are from Working category, 18% from pensioner category 23% are from commercial associate category.

Bivariate Analysis

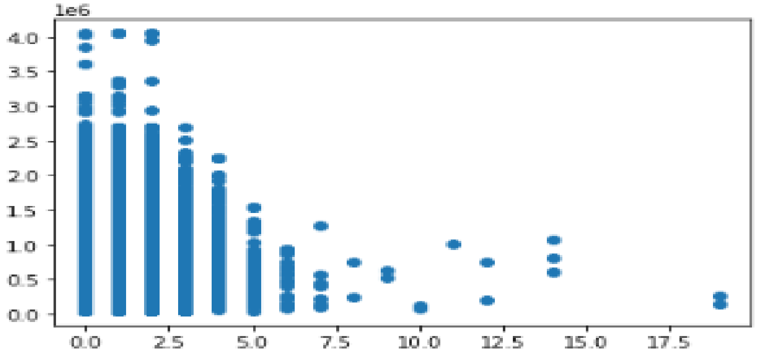
AMT_CREDIT vs CNT_CHLDREN





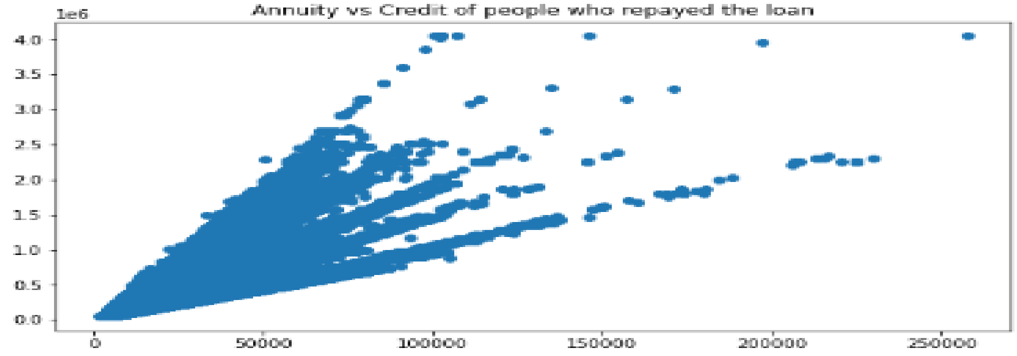


```
plt.scatter(application['CNT_CHILDREN'] ,application['AMT_CREDIT'])
plt.show()
```

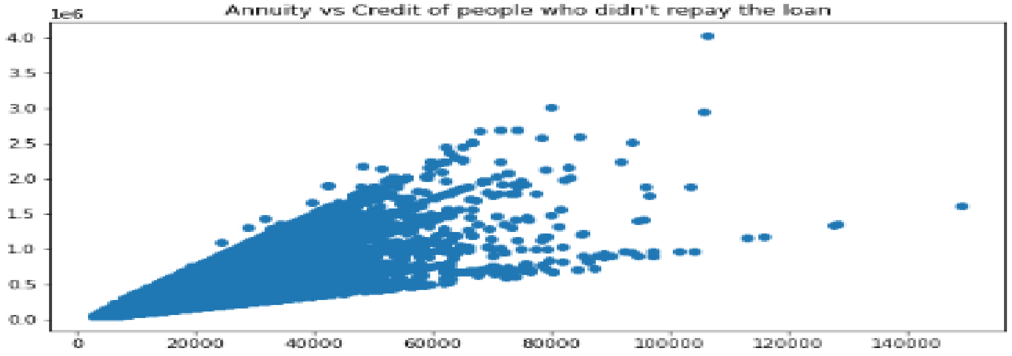


- 1. We can Infer from the graph that people with higher numerber of children do not generally tend to go for the loan.
- 2. Also there isn't much difference in the pattern of having children and the credit they have for both defaulters and non-defaulters.

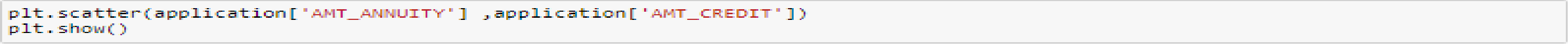
AMT_CREDIT vs AMT_ANNUITY`



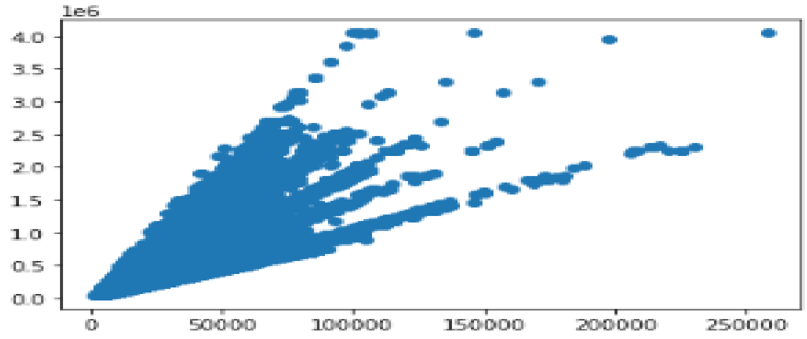
Annuity vs Credit of people who repayed the loan



Annuity vs Credit of people who didn't repay the loan



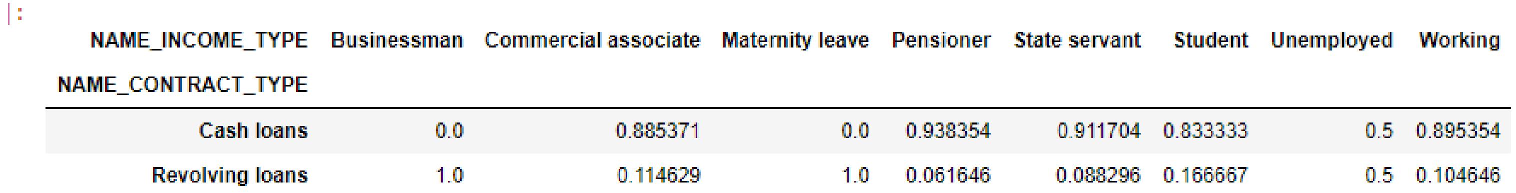
```
plt.scatter(application['AMT_ANNUITY'] ,application['AMT_CREDIT'])
plt.show()
```



- We can see that Annuity is proportional to Credit. Higher the credit on the person, higher is the annuity.
- This was seen for both people who defaulted and people who didn't

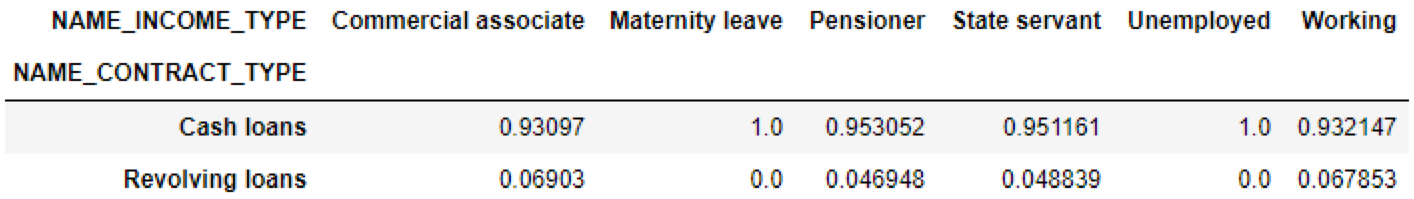
NAME_INCOME_TYPE vs NAME_CONTRACT_TYPE

Non-Defaulters Crosstab



NAME_INCOME_TYPE	Businessman	Commercial associate	Maternity leave	Pensioner	State servant	Student	Unemployed	Working
NAME_CONTRACT_TYPE								
Cash loans	0.0	0.885371	0.0	0.938354	0.911704	0.833333	0.5	0.895354
Revolving loans	1.0	0.114629	1.0	0.061646	0.088296	0.166667	0.5	0.104646

Defaulters Crosstab

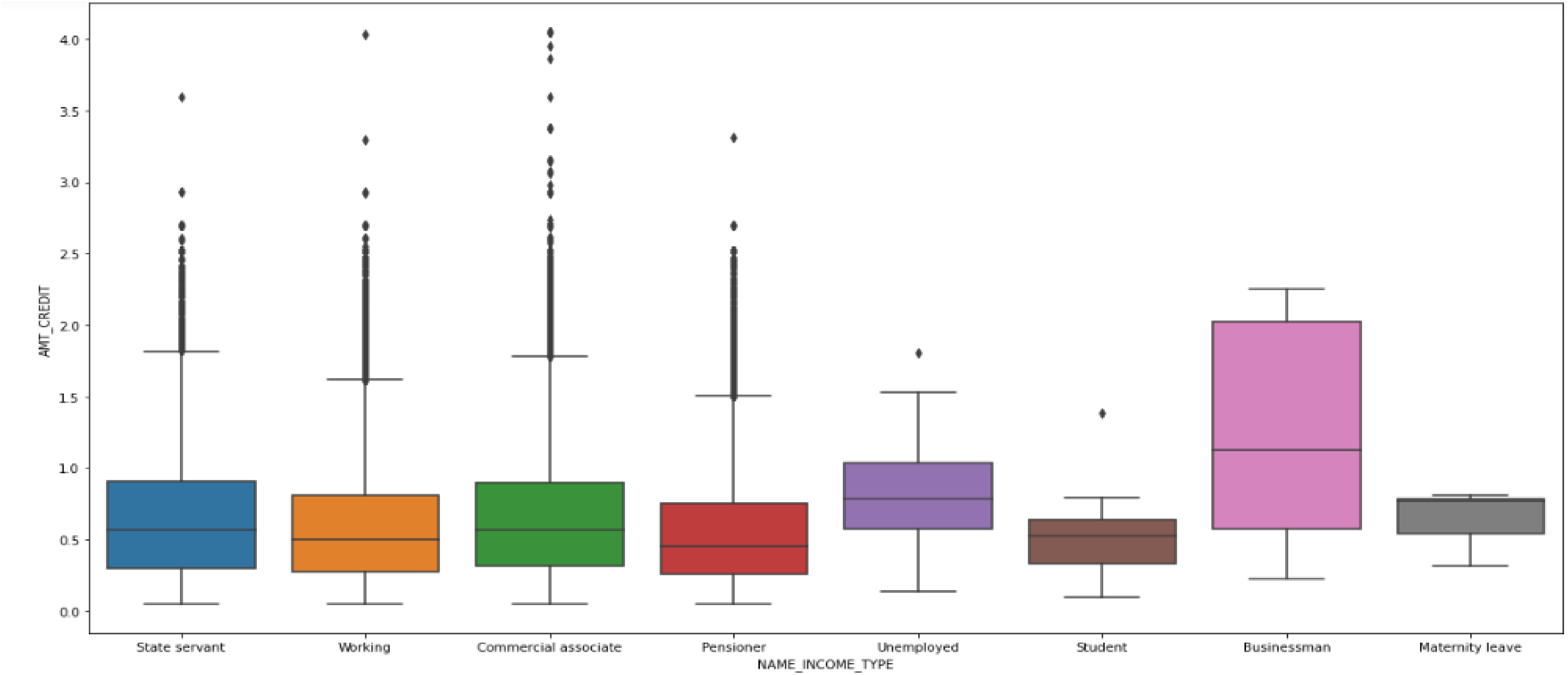


NAME_INCOME_TYPE	Commercial associate	Maternity leave	Pensioner	State servant	Unemployed	Working
NAME_CONTRACT_TYPE						
Cash loans	0.93097	1.0	0.953052	0.951161	1.0	0.932147
Revolving loans	0.06903	0.0	0.046948	0.048839	0.0	0.067853

- Of those who are not a defaulter and are commercial associate, 88% repayed cash loan and 12% repayed Revolving loans
- Of those who are defaulter and are commercial associate,93% failed in repaying cash loans and 7% failed in repaying revolving loans

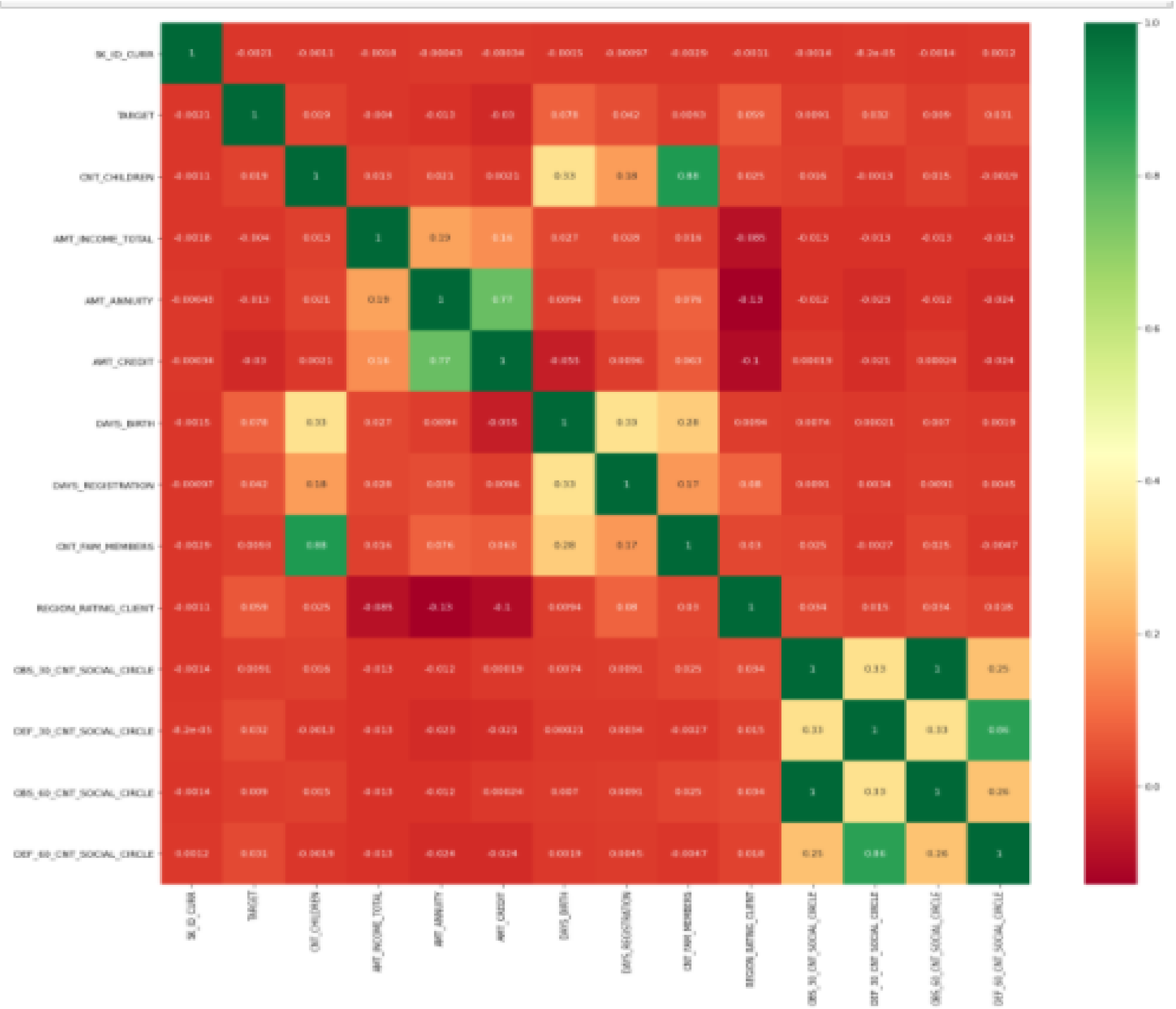
NAME_INCOME_TYPE vs AMT_CREDIT

Non-Defaulters:



- For non-Defaulters, the median of Buissmen's AMT_CREDIT is higher than the rest of the income type.
- Interquartile range is also highest for businessman.

Highly Correlated Columns



We Drew a correlation heatmap to check which columns were highly correlated to our target variable.

We found the fllowing columns highly related compared to others:

- DAYS_BIRTH
- REGION_ATING_CLIENT
- DAYS_REGISTRATION
- DEF_30_CNT_SOCIAL_CIRCLE
- DEF_60_CNT_SOCIAL_CIRCLE