CREDIT EDA CASE STUDY

ASHUTOSH NAYAK

MAHESH PRASAD MISHRA

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 are capable of repaying the loan are not rejected.

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.

SOLUTION APPROACH

We have followed following EDA approach for this solution.
Data Understanding :

- Sampling of data to find out data definitions.
- Analyze the data types of each of the columns and if needed try to modify the data types suitable for our analysis.
- Try to understand the all the columns that are available and try to indentify the variables for our univariate and bivariate analysis.

Data Cleaning :

- As mentioned in the requirement we have removed the columns which has null value percentage more than 45% and we have umpute the variables with zero which has null value more than 14% and again the variables which has less than 14% null value we have imputed with median for numerical variable.
- For categorical variables the null value should be imputed with mode value of the respective column but we did not find any null value in the categorical variable.

SOLUTION APPROACH

Data Cleaning :

- We have found out the outliers for some of the numeric columns and imputed with median value for the outliers.
- We have filtered out the data based on the target variables and have analyzed the data for data imbalance.

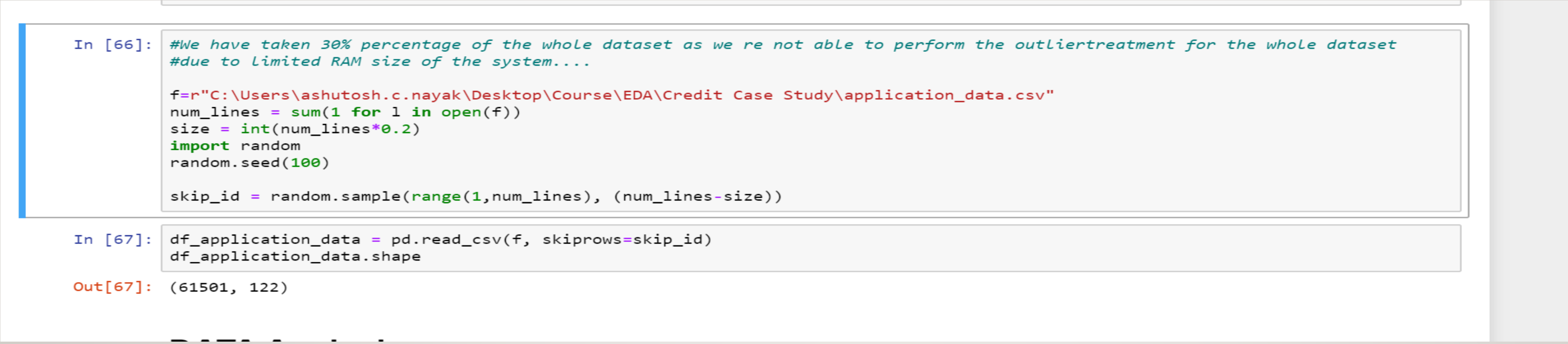
Data Analysis:

- We have identify continuous variables and categorical variables for each Target category for univariate analysis.
- For each Target variables we have done Bivariate analysis for continuous-continuous variables and continuous-categorical variables.
- Find the Correlation between all these variables to find out clients with payment difficulties
- Using different visualization technique present the analysis.

DATA UNDERSTANDING

- Sampling of data to find out data definitions
- Understanding the various Features of data

Sampling of data to find out data definitions



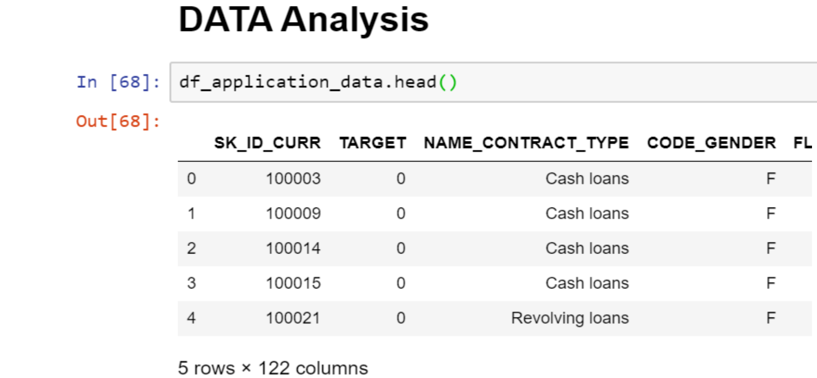
In [66]:
```
#We have taken 30% percentage of the whole dataset as we re not able to perform the outliertreatment for the whole dataset
#due to limited RAM size of the system....

f=r"C:\Users\ashutosh.c.nayak\Desktop\Course\EDA\Credit Case Study\application_data.csv"
num_lines = sum(1 for l in open(f))
size = int(num_lines*0.2)
import random
random.seed(100)

skip_id = random.sample(range(1,num_lines), (num_lines-size))
```

In [67]:
```
df_application_data = pd.read_csv(f, skiprows=skip_id)
df_application_data.shape
```

Out[67]:
```
(61501, 122)
```

As given data set is quite large for Analysis. We have taken only 30% of the data for our analysis. This will help us to speed up the data analysis with minimal memory consumption.

- Application_data.csv has all the data related the Loan Applicant



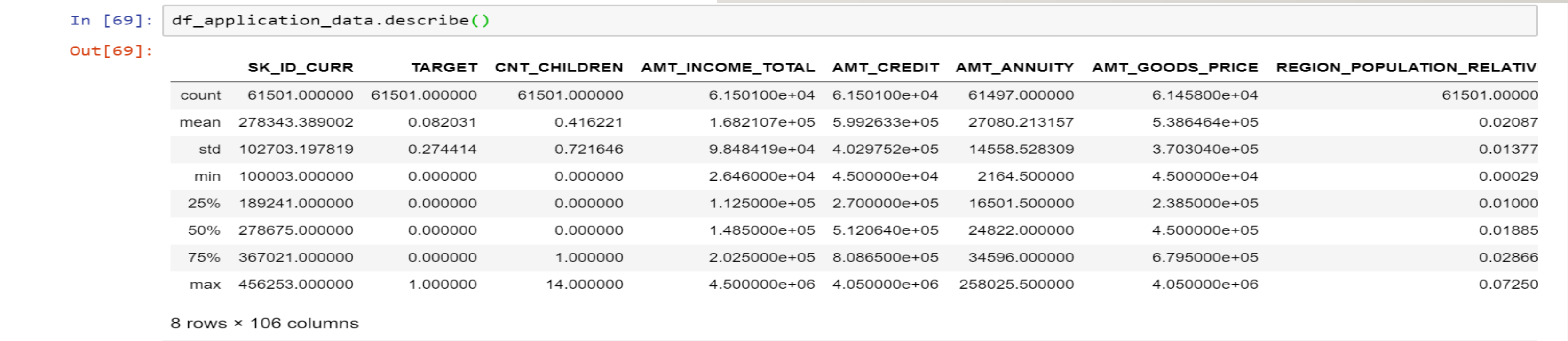
DATA Analysis

In [68]:
```
df_application_data.head()
```

Out[68]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FL
0	100003	0	Cash loans	F	
1	100009	0	Cash loans	F	
2	100014	0	Cash loans	F	
3	100015	0	Cash loans	F	
4	100021	0	Revolving loans	F	

5 rows × 122 columns



In [69]:
```
df_application_data.describe()
```

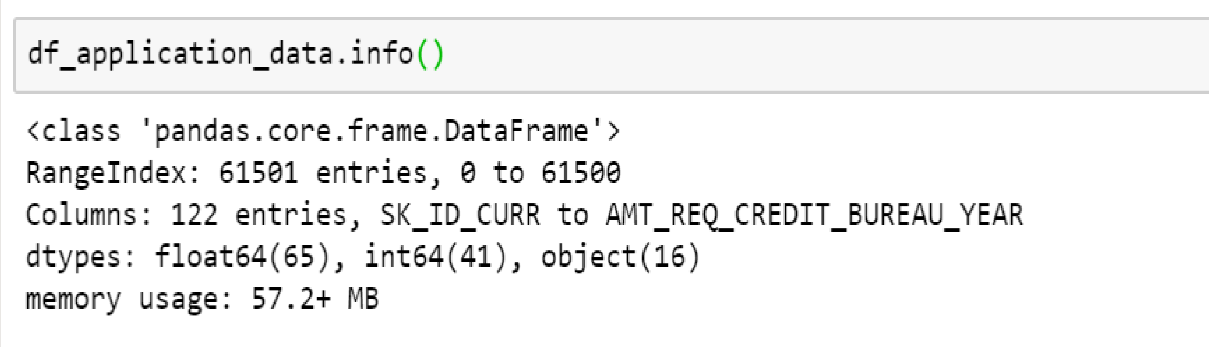
Out[69]:

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIV
count	61501.000000	61501.000000	61501.000000	6.150100e+04	6.150100e+04	61497.000000	6.145800e+04	61501.00000
mean	278343.389002	0.082031	0.416221	1.682107e+05	5.992633e+05	27080.213157	5.386464e+05	0.02087
std	102703.197819	0.274414	0.721646	9.848419e+04	4.029752e+05	14558.528309	3.703040e+05	0.01377
min	100003.000000	0.000000	0.000000	2.646000e+04	4.500000e+04	2164.500000	4.500000e+04	0.00029
25%	189241.000000	0.000000	0.000000	1.125000e+05	2.700000e+05	16501.500000	2.385000e+05	0.01000
50%	278675.000000	0.000000	0.000000	1.485000e+05	5.120640e+05	24822.000000	4.500000e+05	0.01885
75%	367021.000000	0.000000	1.000000	2.025000e+05	8.086500e+05	34596.000000	6.795000e+05	0.02866
max	456253.000000	1.000000	14.000000	4.500000e+06	4.050000e+06	258025.500000	4.050000e+06	0.07250

8 rows × 106 columns

Understanding the various Features of data

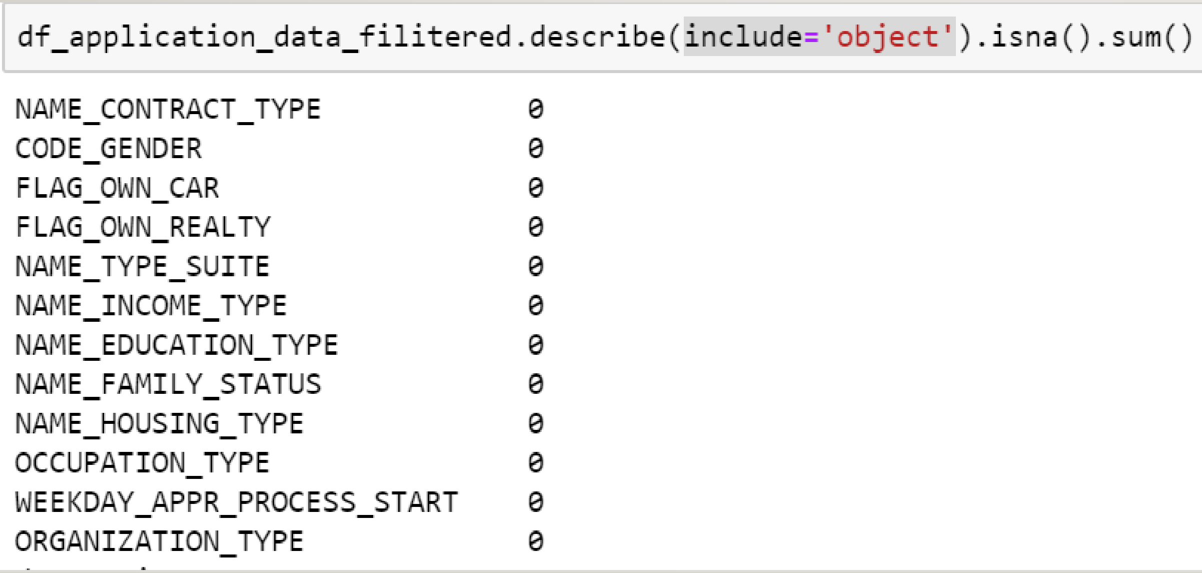
- Analyzing the data types of al the columns.



```
df_application_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61501 entries, 0 to 61500
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 57.2+ MB
```

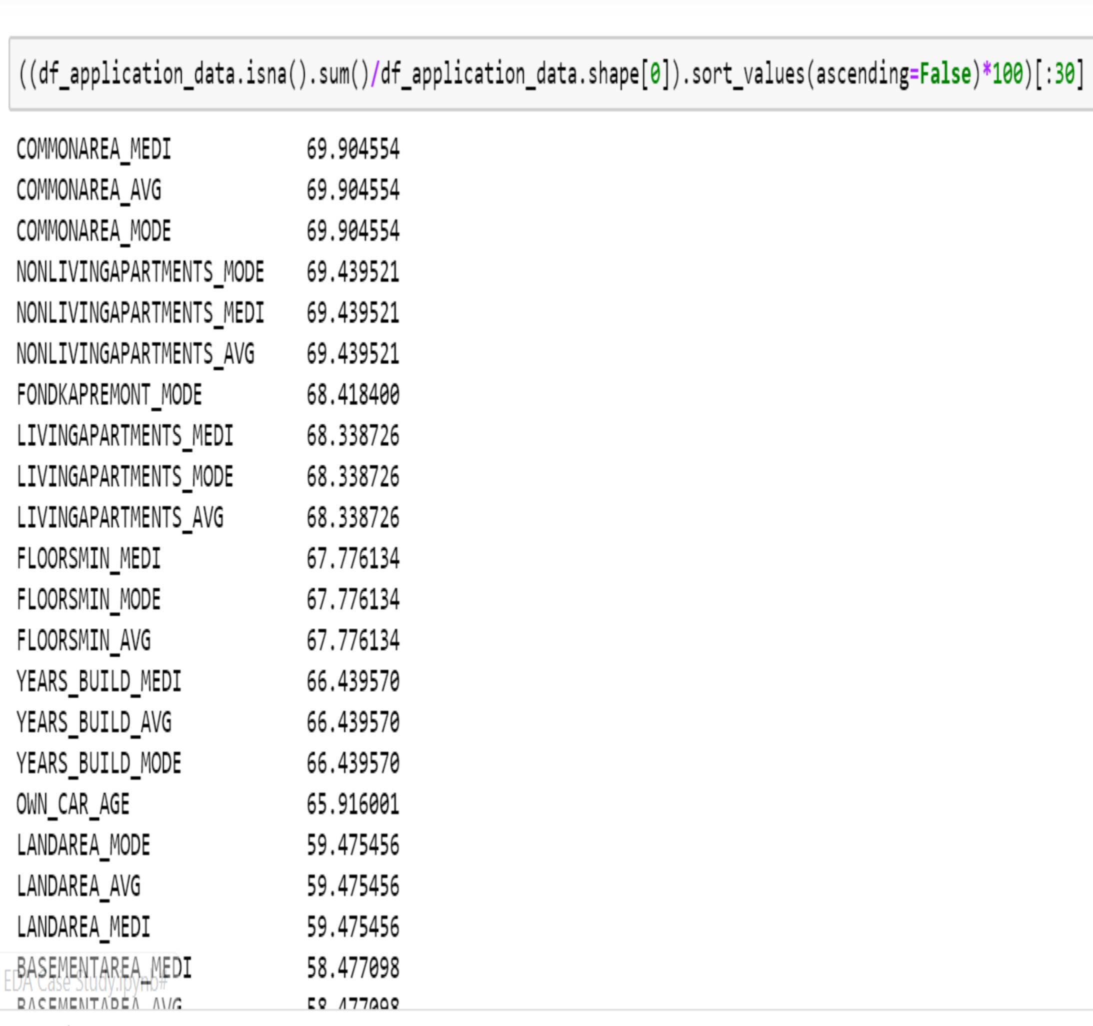
- Finding NULL values For categorical variable



```
df_application_data_filitered.describe(include='object').isna().sum()
```

```
NAME_CONTRACT_TYPE            0
CODE_GENDER                   0
FLAG_OWN_CAR                  0
FLAG_OWN_REALTY               0
NAME_TYPE_SUITE               0
NAME_INCOME_TYPE              0
NAME_EDUCATION_TYPE           0
NAME_FAMILY_STATUS            0
NAME_HOUSING_TYPE             0
OCCUPATION_TYPE               0
WEEKDAY_APPR_PROCESS_START    0
ORGANIZATION_TYPE             0
```

- Finding NULL values in each Feature

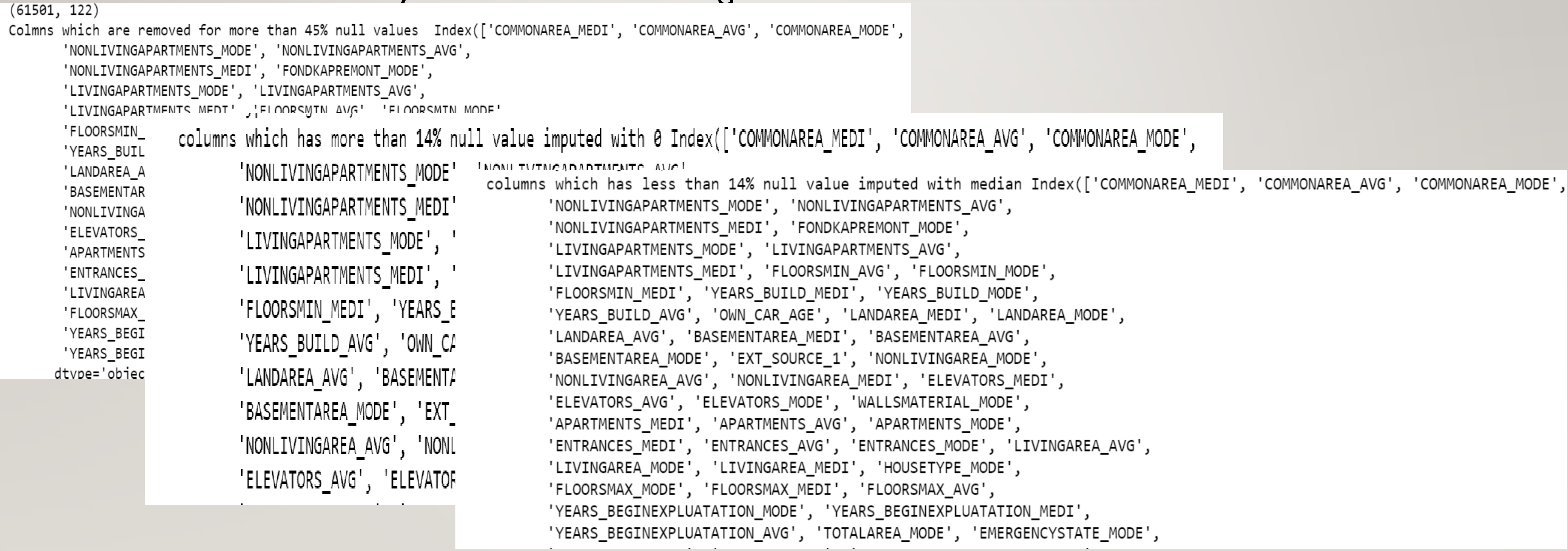


```
((df_application_data.isna().sum()/df_application_data.shape[0]).sort_values(ascending=False)*100)[:30]
```

```
COMMONAREA_MEDI             69.904554
COMMONAREA_AVG              69.904554
COMMONAREA_MODE             69.904554
NONLIVINGAPARTMENTS_MODE    69.439521
NONLIVINGAPARTMENTS_MEDI    69.439521
NONLIVINGAPARTMENTS_AVG     69.439521
FONDKAPREMONT_MODE          68.418400
LIVINGAPARTMENTS_MEDI       68.338726
LIVINGAPARTMENTS_MODE       68.338726
LIVINGAPARTMENTS_AVG        68.338726
FLOORSMIN_MEDI              67.776134
FLOORSMIN_MODE              67.776134
FLOORSMIN_AVG               67.776134
YEARS_BUILD_MEDI            66.439570
YEARS_BUILD_AVG             66.439570
YEARS_BUILD_MODE            66.439570
OWN_CAR_AGE                 65.916001
LANDAREA_MODE               59.475456
LANDAREA_AVG                59.475456
LANDAREA_MEDI               59.475456
BASEMENTAREA_MEDI           58.477098
BASEMENTAREA_AVG            58.477098
```

Data Cleaning :

- As mentioned in the requirement we have removed the columns which has null value percentage more than 45% and we have umpute the variables with zero which has null value more than 14% and again the variables which has less than 14% null value we have imputed with median for numerical variable.
- For categorical variables the null value should be imputed with mode value of the respective column but we did not find any null value in the categorical variable.



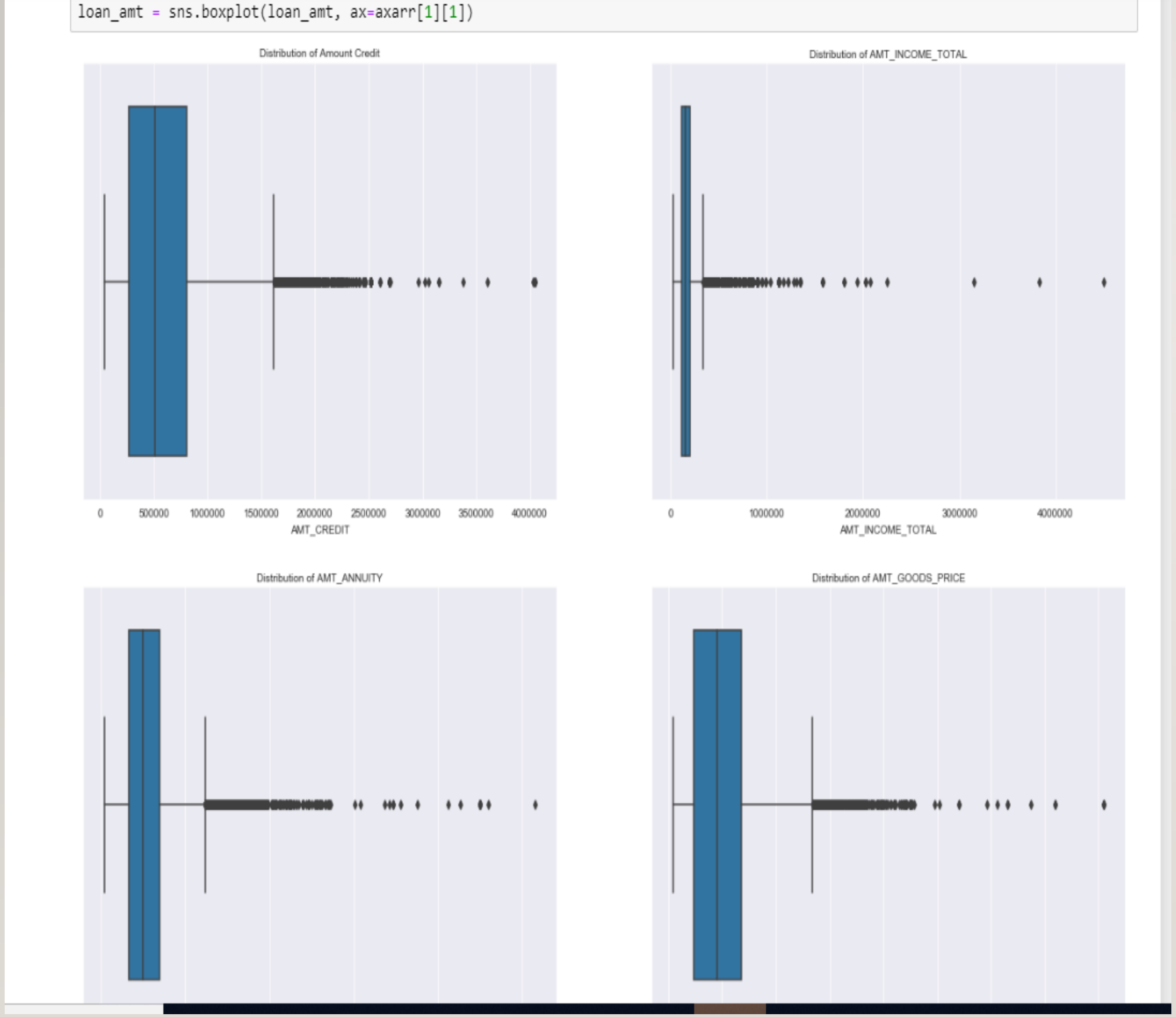
```
(61501, 122)
Colmns which are removed for more than 45% null values  Index(['COMMONAREA_MEDI', 'COMMONAREA_AVG', 'COMMONAREA_MODE',
       'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAPARTMENTS_AVG',
       'NONLIVINGAPARTMENTS_MEDI', 'FONDKAPREMONT_MODE',
       'LIVINGAPARTMENTS_MODE', 'LIVINGAPARTMENTS_AVG',
       'LIVINGAPARTMENTS_MEDI', 'FLOORSMIN_AVG', 'FLOORSMIN_MODE'
       'FLOORSMIN_
       'YEARS_BUIL
       'LANDAREA_A
       'BASEMENTAR
       'NONLIVINGA
       'ELEVATORS_
       'APARTMENTS
       'ENTRANCES_
       'LIVINGAREA
       'FLOORSMAX_
       'YEARS_BEGI
       'YEARS_BEGI
      dtype='objec
```

```
columns which has more than 14% null value imputed with 0 Index(['COMMONAREA_MEDI', 'COMMONAREA_AVG', 'COMMONAREA_MODE',
       'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAPARTMENTS_AVG',
       'NONLIVINGAPARTMENTS_MEDI'
       'LIVINGAPARTMENTS_MODE', '
       'LIVINGAPARTMENTS_MEDI', '
       'FLOORSMIN_MEDI', 'YEARS_E
       'YEARS_BUILD_AVG', 'OWN_CA
       'LANDAREA_AVG', 'BASEMENTA
       'BASEMENTAREA_MODE', 'EXT_
       'NONLIVINGAREA_AVG', 'NONL
       'ELEVATORS_AVG', 'ELEVATOR
```

```
columns which has less than 14% null value imputed with median Index(['COMMONAREA_MEDI', 'COMMONAREA_AVG', 'COMMONAREA_MODE',
       'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAPARTMENTS_AVG',
       'NONLIVINGAPARTMENTS_MEDI', 'FONDKAPREMONT_MODE',
       'LIVINGAPARTMENTS_MODE', 'LIVINGAPARTMENTS_AVG',
       'LIVINGAPARTMENTS_MEDI', 'FLOORSMIN_AVG', 'FLOORSMIN_MODE',
       'FLOORSMIN_MEDI', 'YEARS_BUILD_MEDI', 'YEARS_BUILD_MODE',
       'YEARS_BUILD_AVG', 'OWN_CAR_AGE', 'LANDAREA_MEDI', 'LANDAREA_MODE',
       'LANDAREA_AVG', 'BASEMENTAREA_MEDI', 'BASEMENTAREA_AVG',
       'BASEMENTAREA_MODE', 'EXT_SOURCE_1', 'NONLIVINGAREA_MODE',
       'NONLIVINGAREA_AVG', 'NONLIVINGAREA_MEDI', 'ELEVATORS_MEDI',
       'ELEVATORS_AVG', 'ELEVATORS_MODE', 'WALLSMATERIAL_MODE',
       'APARTMENTS_MEDI', 'APARTMENTS_AVG', 'APARTMENTS_MODE',
       'ENTRANCES_MEDI', 'ENTRANCES_AVG', 'ENTRANCES_MODE', 'LIVINGAREA_AVG',
       'LIVINGAREA_MODE', 'LIVINGAREA_MEDI', 'HOUSETYPE_MODE',
       'FLOORSMAX_MODE', 'FLOORSMAX_MEDI', 'FLOORSMAX_AVG',
       'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BEGINEXPLUATATION_MEDI',
       'YEARS_BEGINEXPLUATATION_AVG', 'TOTALAREA_MODE', 'EMERGENCYSTATE_MODE',
```

Univariate Analysis : Outlier Analysis

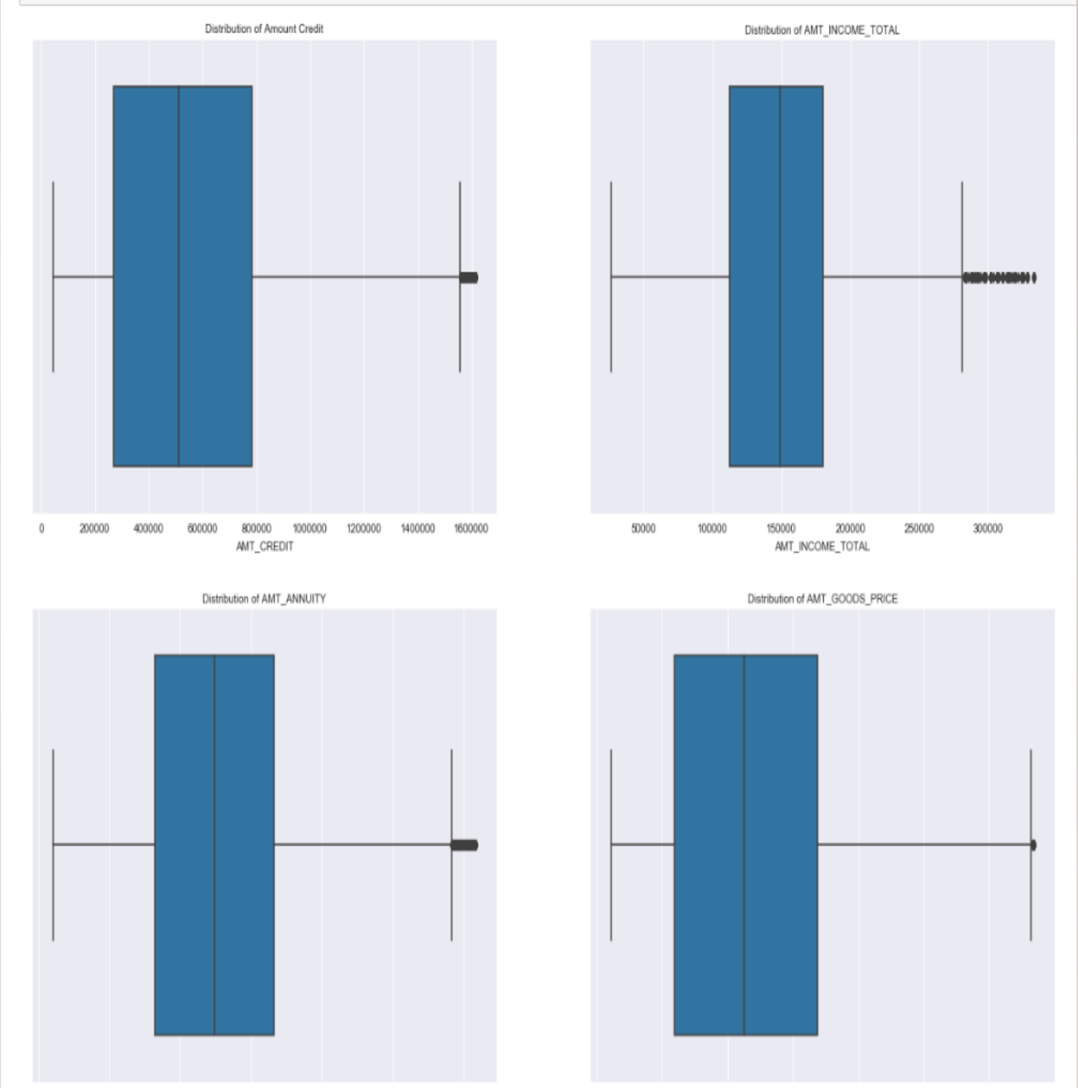
Finding out the Outliers in Univariate variables



```
loan_amt = sns.boxplot(loan_amt, ax=axarr[1][1])
```

Distribution of Amount Credit
0
500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
AMT_CREDIT
Distribution of AMT_INCOME_TOTAL
0
1000000
2000000
3000000
4000000
AMT_INCOME_TOTAL
Distribution of AMT_ANNUITY
Distribution of AMT_GOODS_PRICE

After imputing of Outliers from Univariate variables



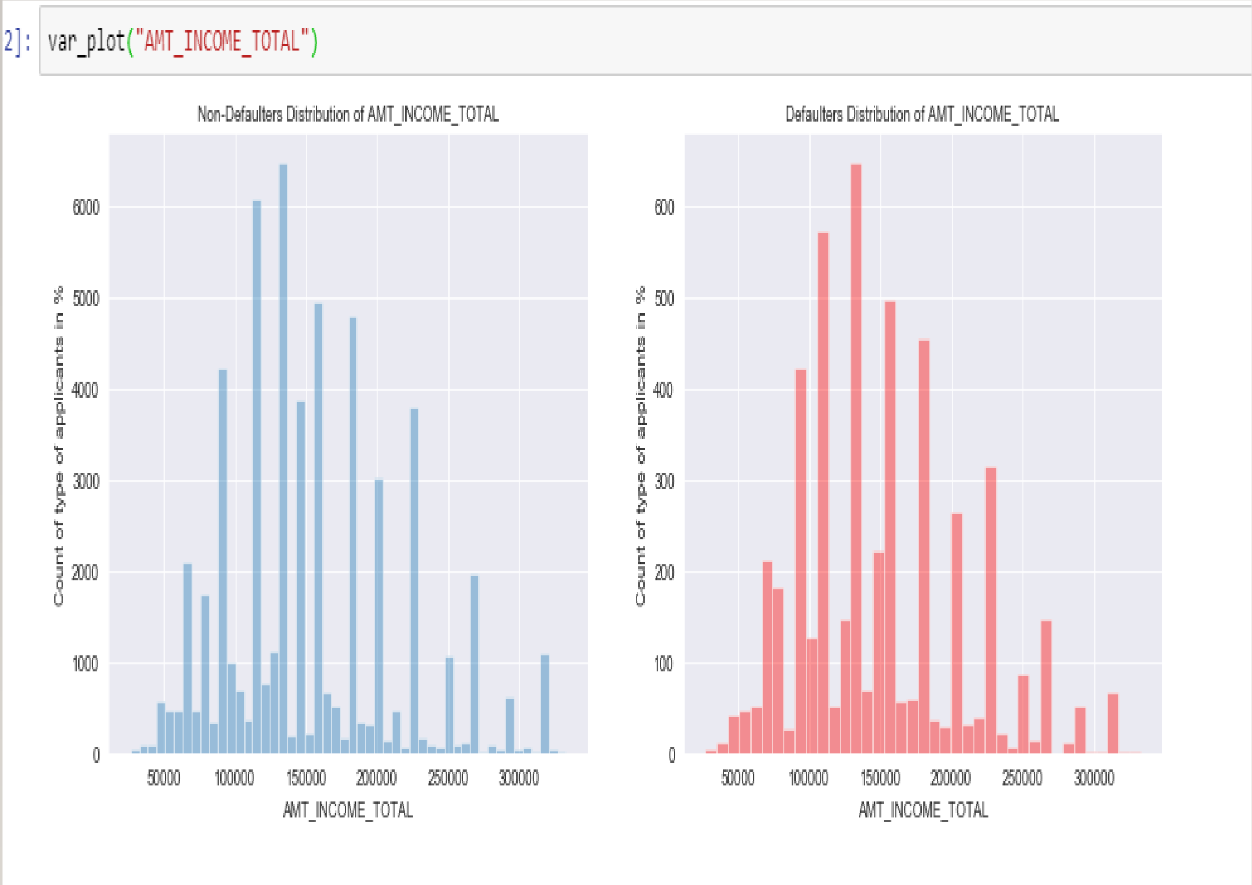
Distribution of Amount Credit
0
200000
400000
600000
800000
1000000
1200000
1400000
1600000
AMT_CREDIT
Distribution of AMT_INCOME_TOTAL
50000
100000
150000
200000
250000
300000
AMT_INCOME_TOTAL
Distribution of AMT_ANNUITY
Distribution of AMT_GOODS_PRICE

Univariate Analysis : Continuous Variables

We have identified following numerical variable Univariate Continuous Variable analysis

- AMT_CREDIT
- AMT_INCOME_TOTAL
- AMT_ANNUITY
- AMT_GOODS_PRICE

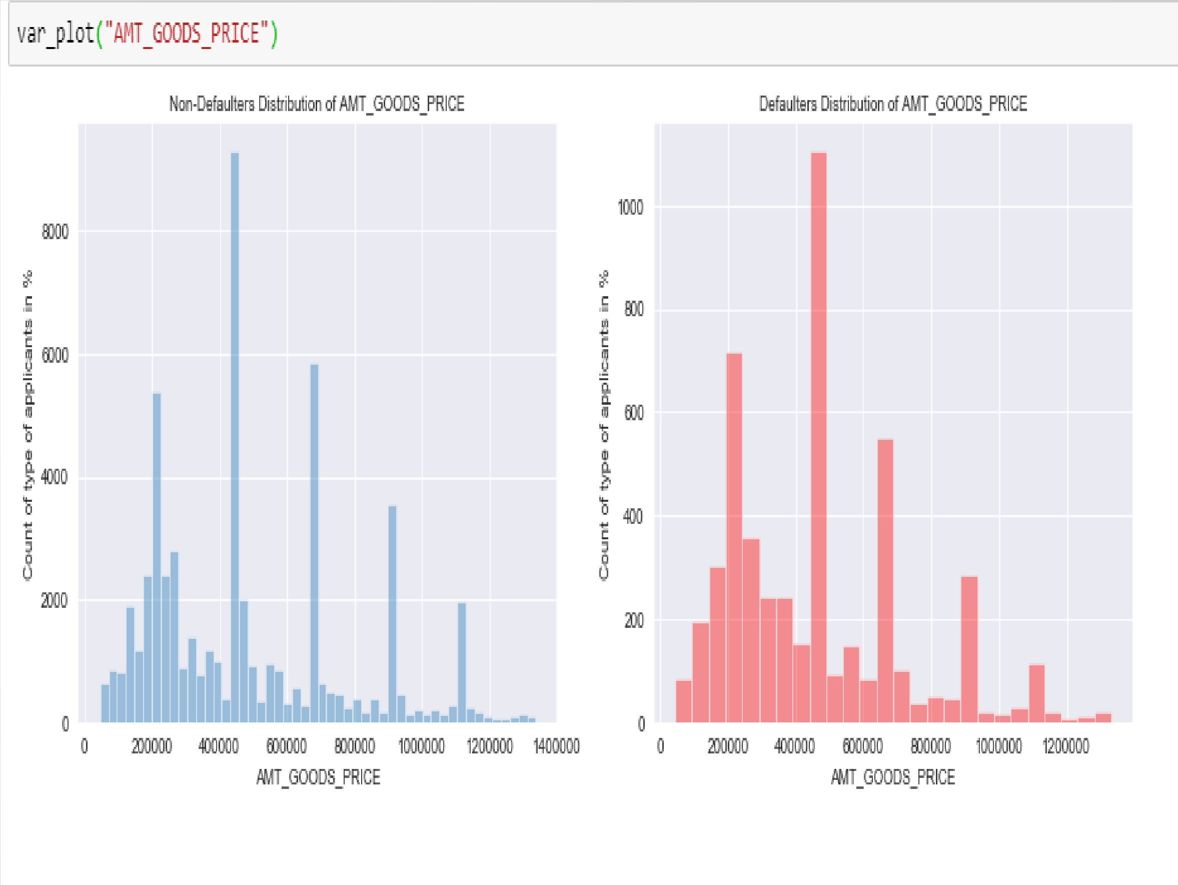
Univariate Analysis : Amount Income Total



```
2]: var_plot("AMT_INCOME_TOTAL")
```

Non-Defaulters Distribution of AMT_INCOME_TOTAL
Count of type of applicants in %
0
1000
2000
3000
4000
5000
6000
50000
100000
150000
200000
250000
300000
AMT_INCOME_TOTAL
Defaulters Distribution of AMT_INCOME_TOTAL
Count of type of applicants in %
0
100
200
300
400
500
600
50000
100000
150000
200000
250000
300000
AMT_INCOME_TOTAL

Univariate Analysis : Amount Goods Price



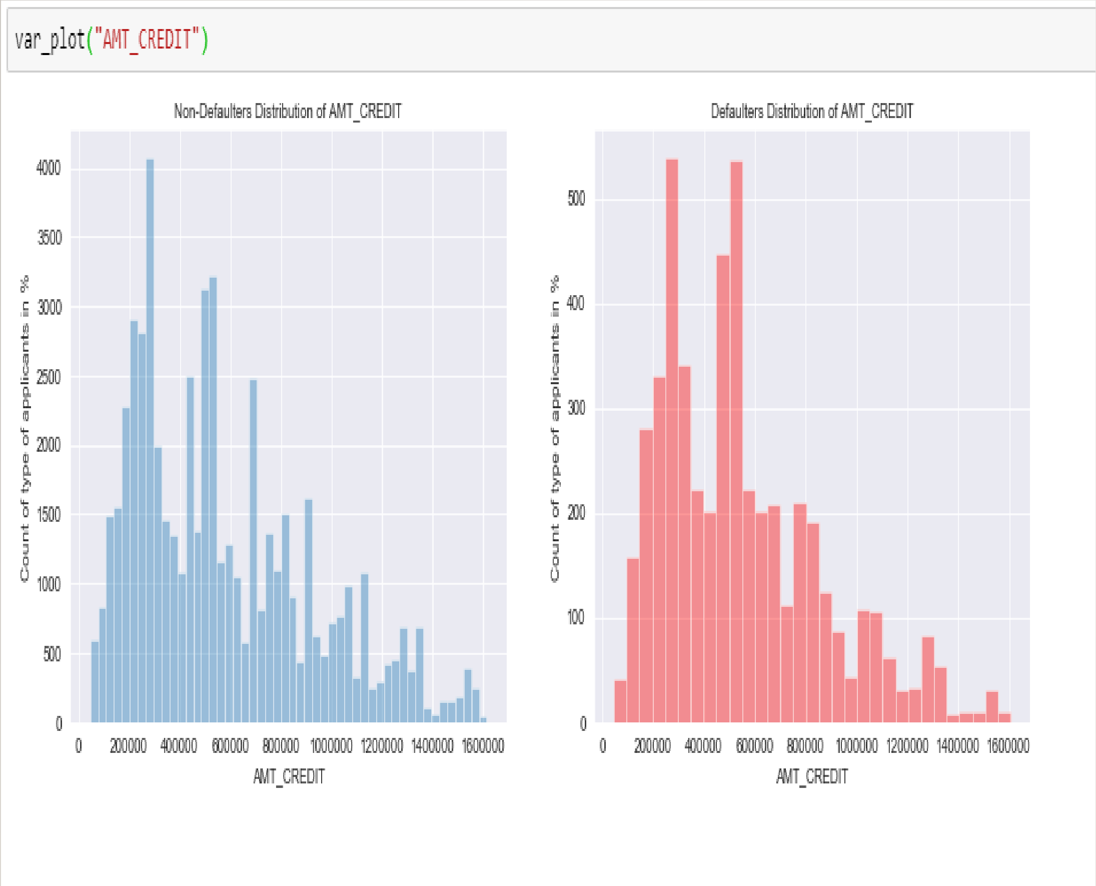
```
var_plot("AMT_GOODS_PRICE")
```

Non-Defaulters Distribution of AMT_GOODS_PRICE
Count of type of applicants in %
0
2000
4000
6000
8000
0
200000
400000
600000
800000
1000000
1200000
1400000
AMT_GOODS_PRICE
Defaulters Distribution of AMT_GOODS_PRICE
Count of type of applicants in %
0
200
400
600
800
1000
0
200000
400000
600000
800000
1000000
1200000
AMT_GOODS_PRICE

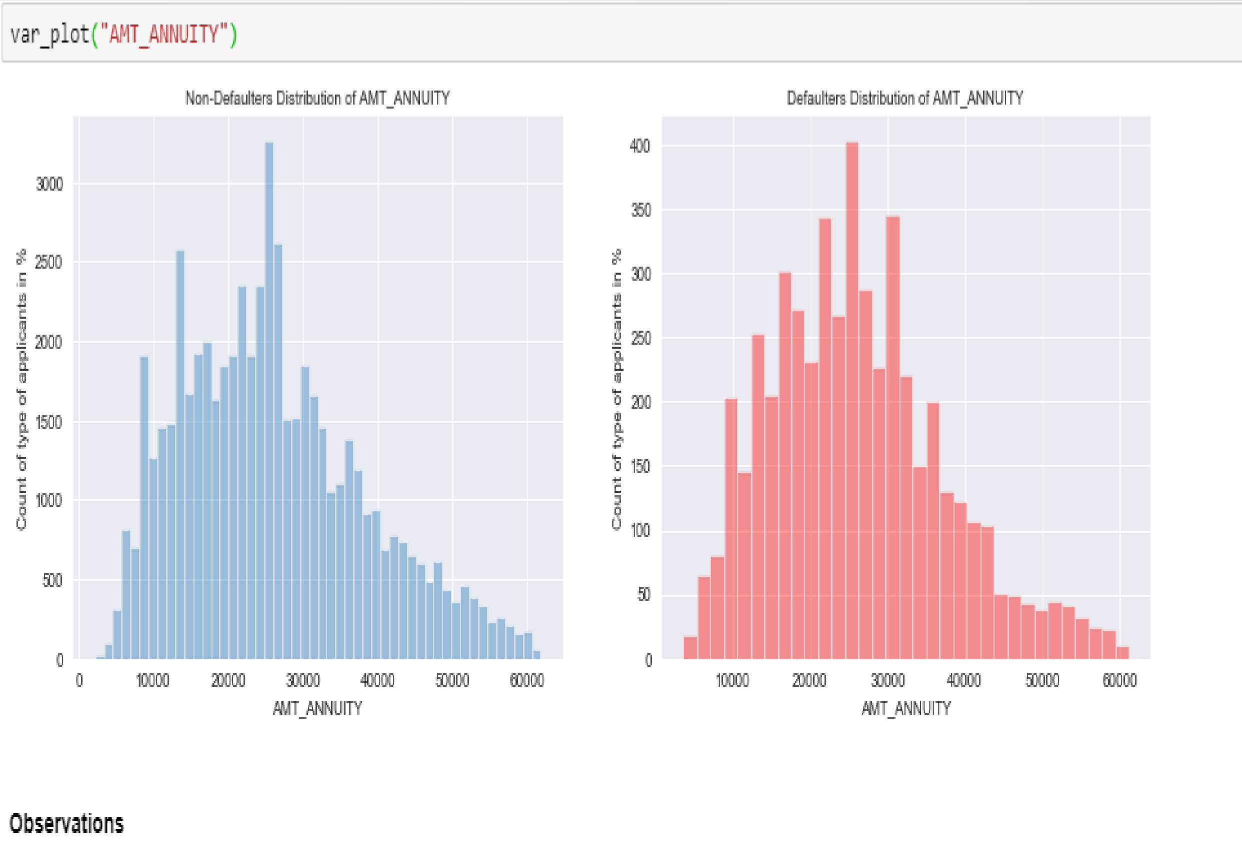
People with high income are repaying the loans and people who have taken high amount loan are also repaying the loan

Univariate Analysis : Amount Credit

Univariate Analysis : Amount Annuity



```
var_plot("AMT_CREDIT")
```



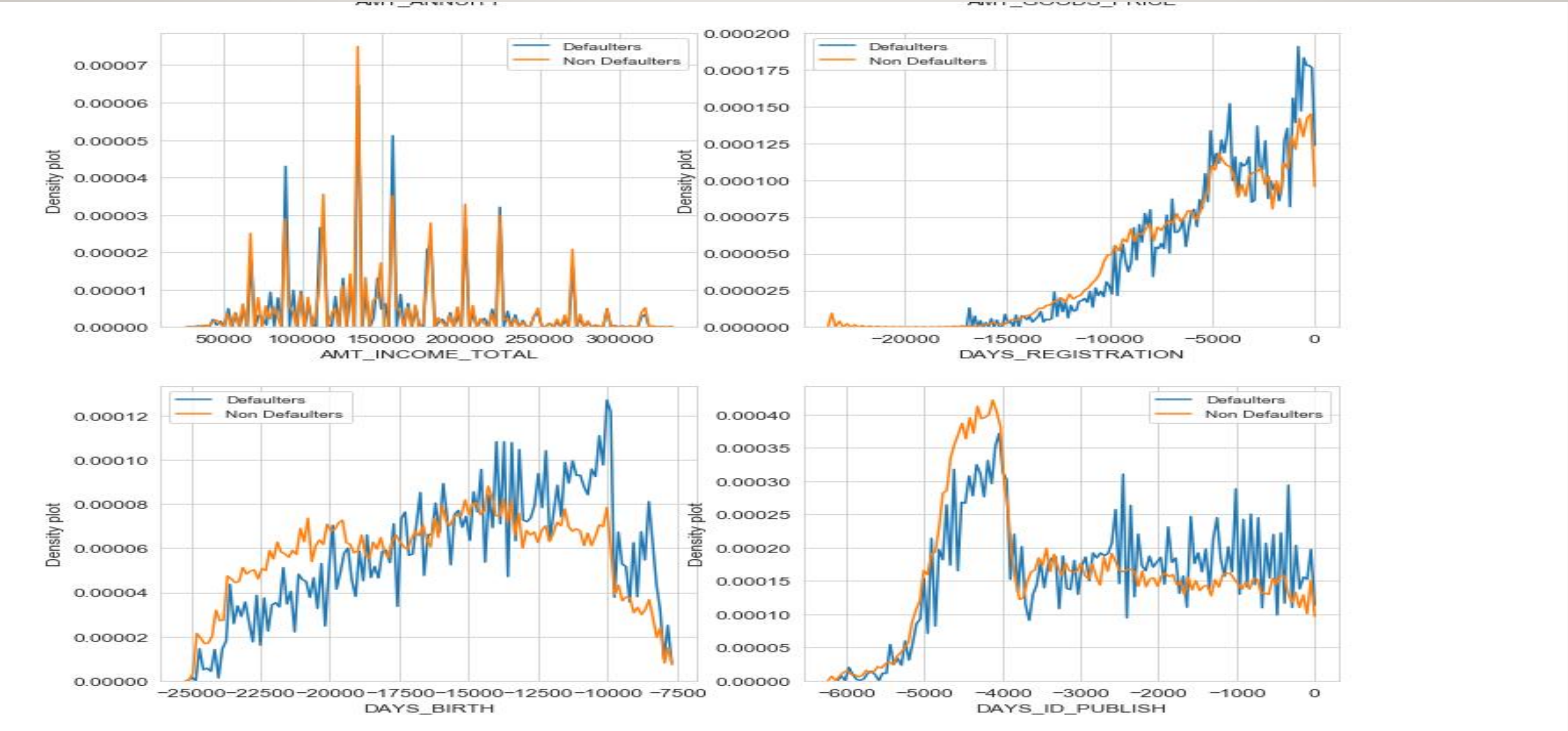
```
var_plot("AMT_ANNUITY")
```

Non-Defaulters Distribution of AMT_ANNUITY
Count of type of applicants in %
0
500
1000
1500
2000
2500
3000
0
10000
20000
30000
40000
50000
60000
AMT_ANNUITY
Defaulters Distribution of AMT_ANNUITY
Count of type of applicants in %
0
50
100
150
200
250
300
350
400
10000
20000
30000
40000
50000
60000
AMT_ANNUITY

Observations

People with amount credit more than 1600000 are repaying the loans in most of the cases

- Persons with age between 30 years and 50 years has high number of defaulter.
- person with id changed in last 1000 days of application has high number of defaulters



Categorical Variable Analysis

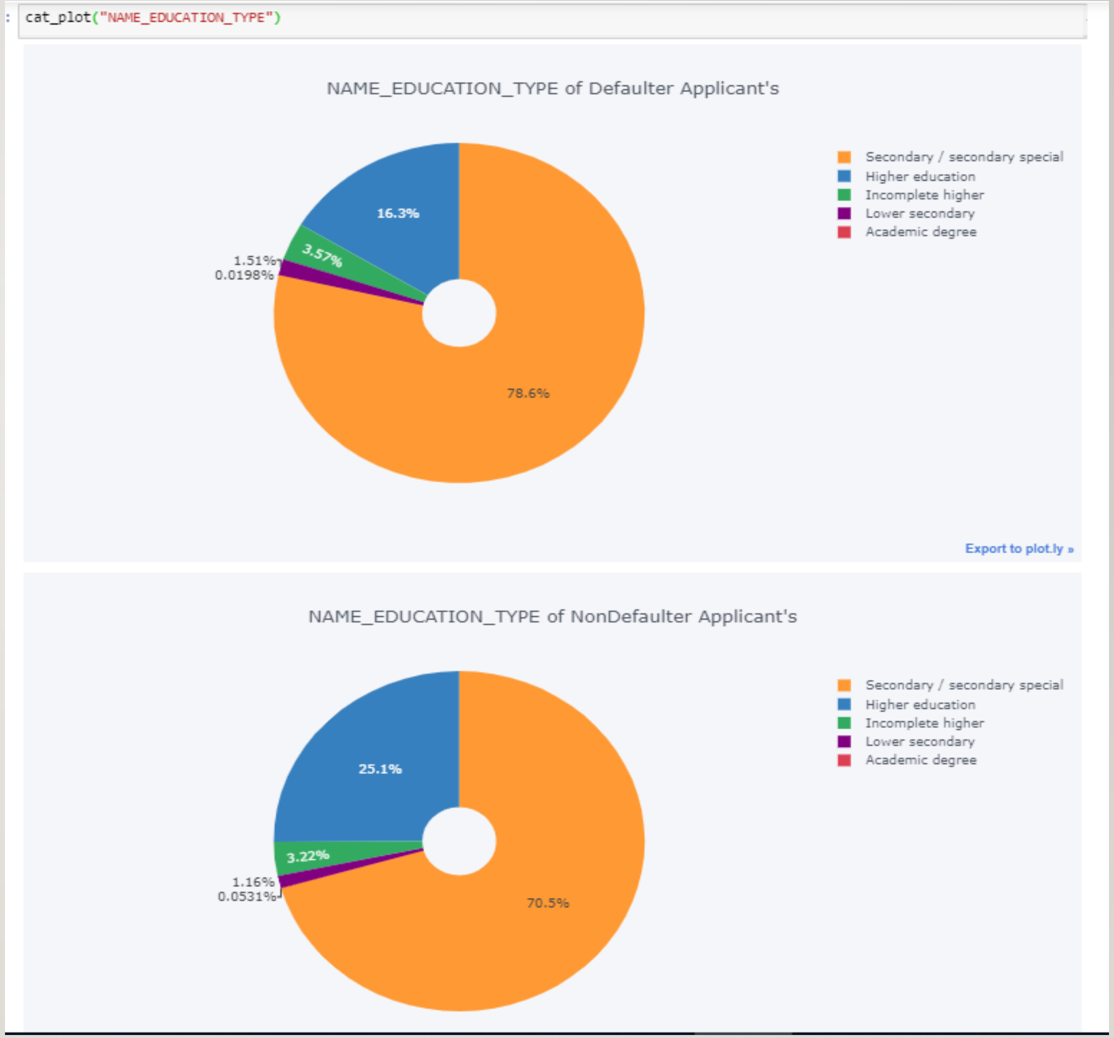
We have identified following variable as Categorical Variable for analysis:

- NAME_EDUCATION_TYPE
- NAME_INCOME_TYPE
- NAME_CONTRACT_STATUS
- NAME_FAMILY_STATUS

CATEGORICAL VARIABLE: EDUCATION TYPE

- With Education Type categorical variable use we can identify what are % of different Education types are Defaulters or Non-Defaulters.

- Close to 16% out of total defaulters are Higher educated, however most of the defaulter are lies in Secondary Education.

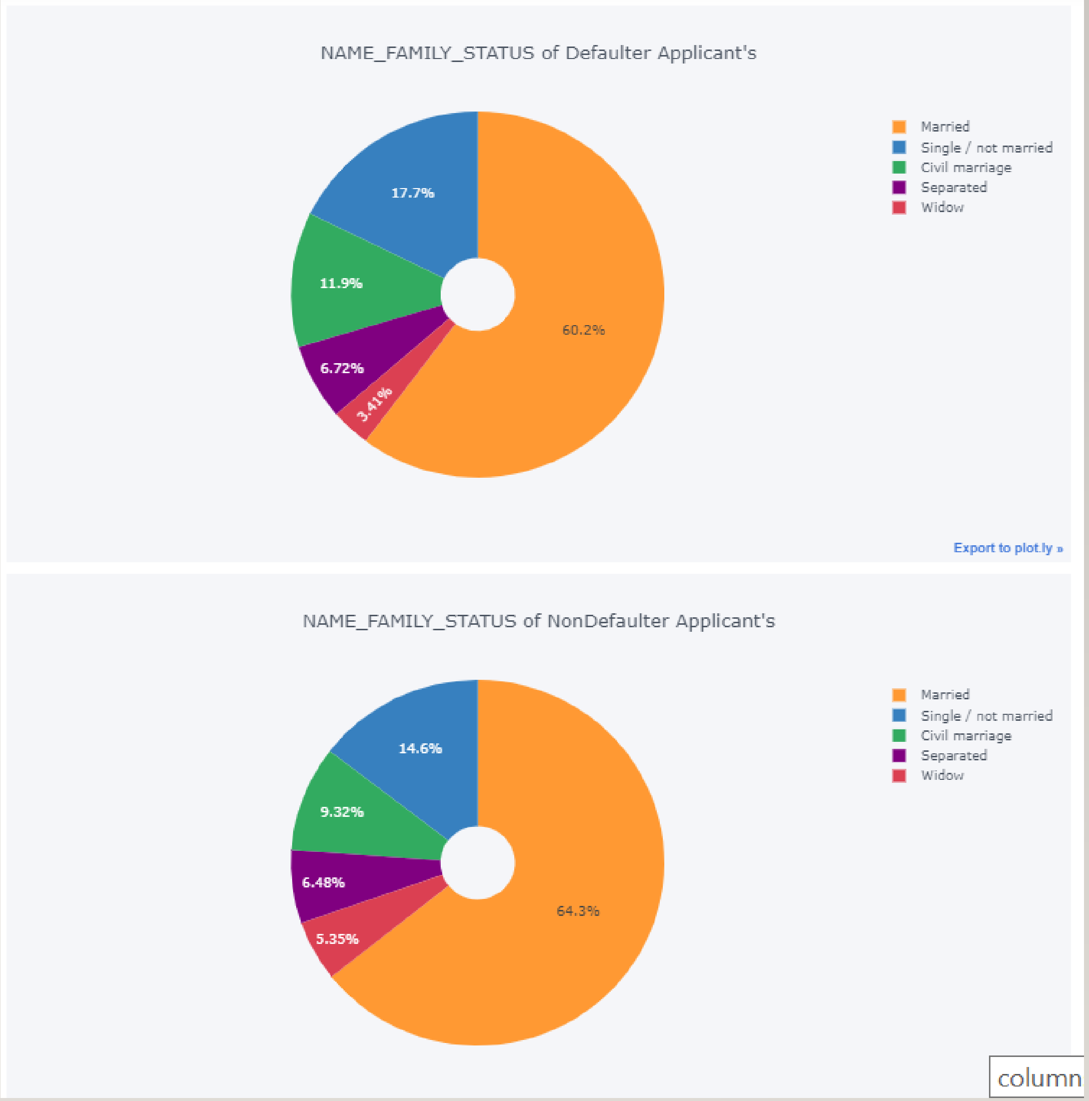


```
cat_plot("NAME_EDUCATION_TYPE")
```

CATEGORICAL VARIABLE: FAMILY STATUS

- With Family status categorical variable use we can identify what are % of different family status are Defaulters or Non-Defaulters.

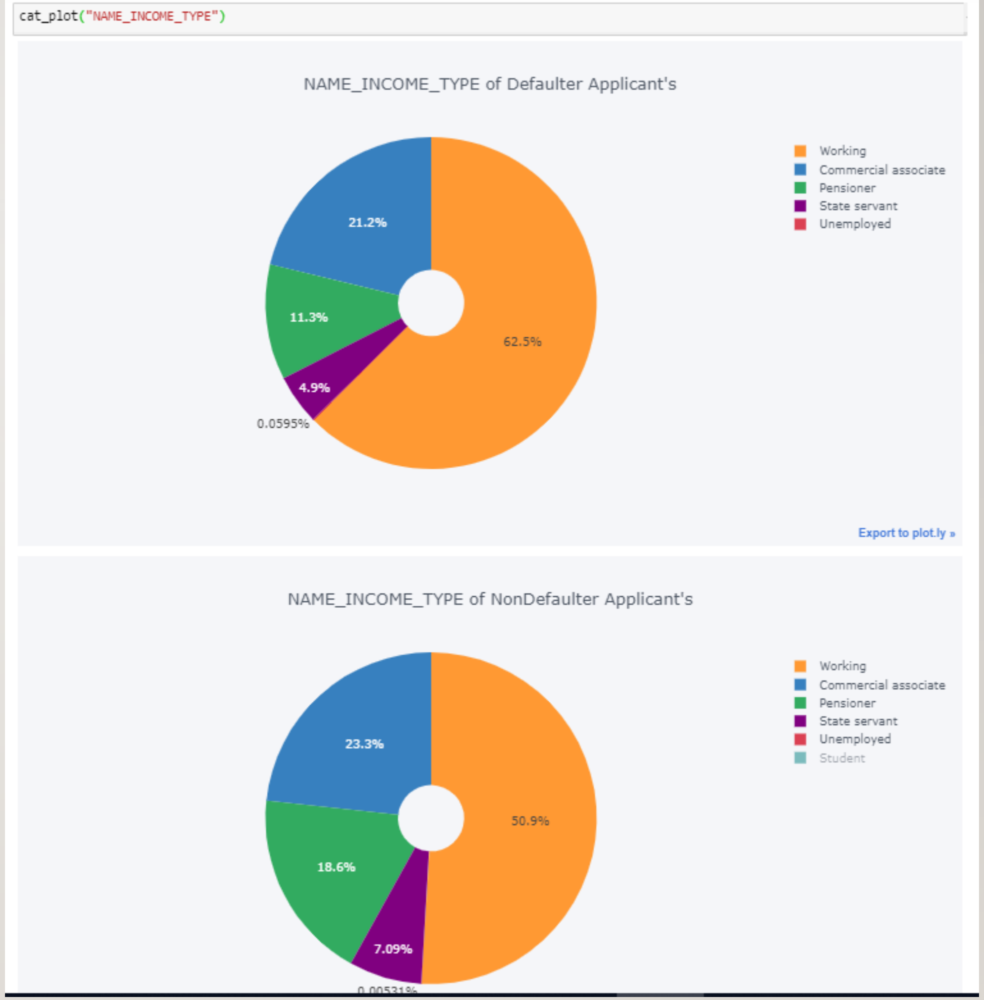
- Widows are more likely to repay the loan when compared to appliants with the other family statuses



CATEGORICAL VARIABLE: INCOME TYPE

- With income type categorical variable use we can identify what are % of different income type are Defaulters or Non-Defaulters.

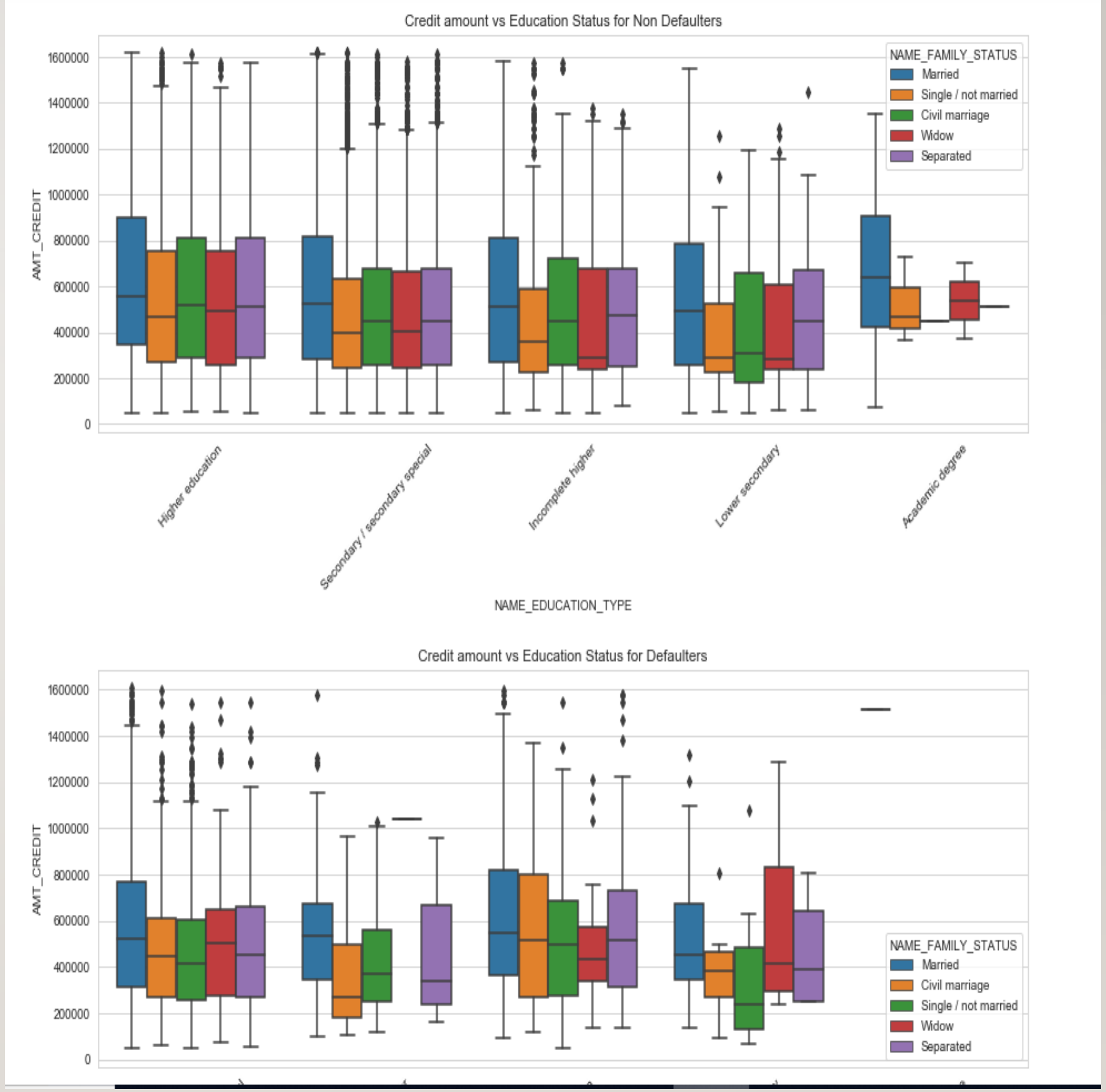
- From the above plot we can conclude that All the Students and Businessman are repaying loan.



```
cat_plot("NAME_INCOME_TYPE")
```

BIVARIATE VARIABLE ANALYSIS CREDIT AMOUNT VS EDUCATION STATUS

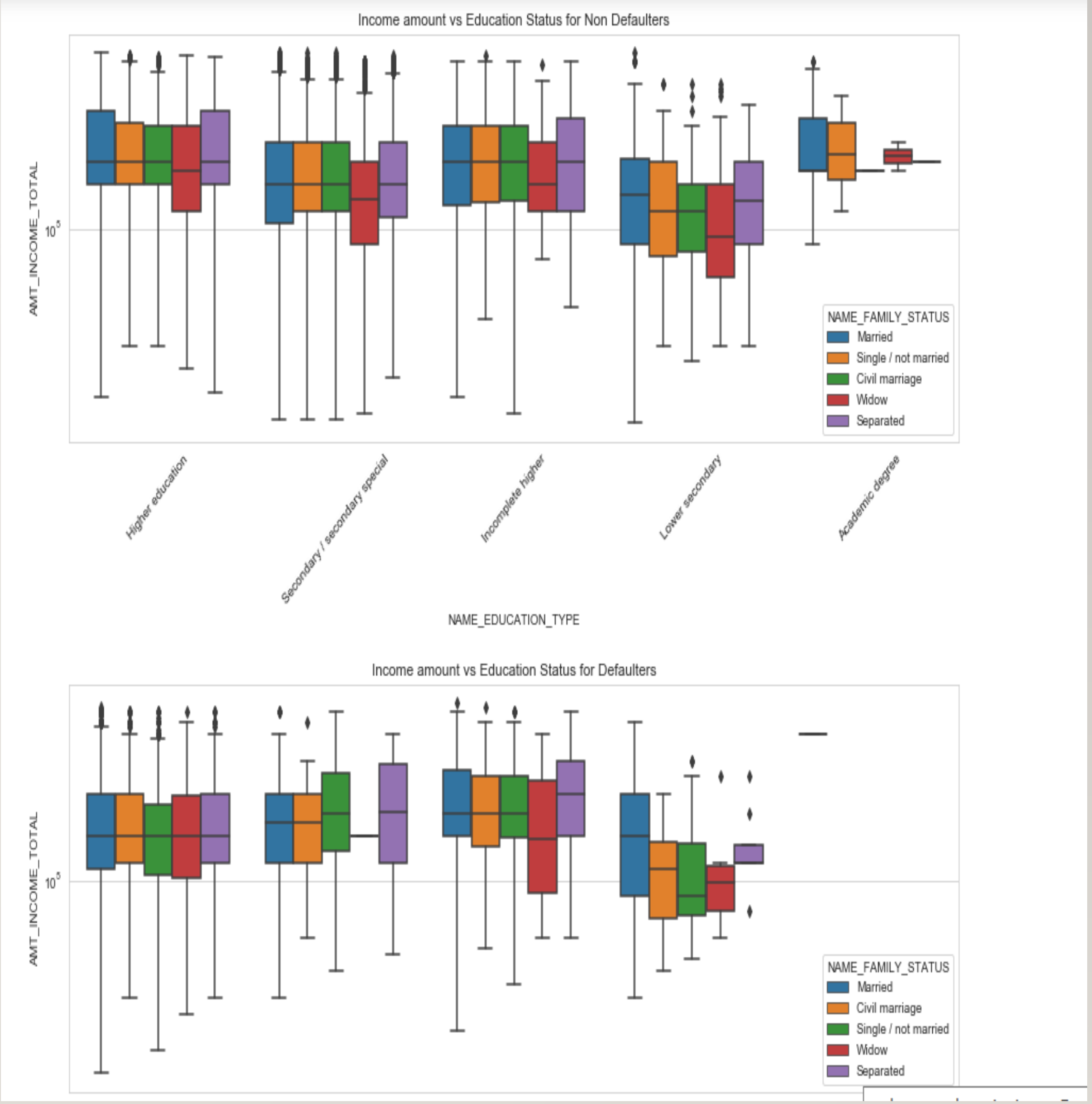
- for non defaulters, Academic degree education with Family status of 'civil marriage', 'marriage' and 'separated' of are having higher number of credits than others.
- Also, higher education with family status of 'marriage', 'single' and 'civil marriage' are having more outliers.
- For defaulters having higher education with Family status of 'civil marriage', 'marriage' and 'separated' are having higher number of credits than others.



BIVARIATE VARIABLE ANALYSIS

INCOME TOTAL VS EDUCATION STATUS

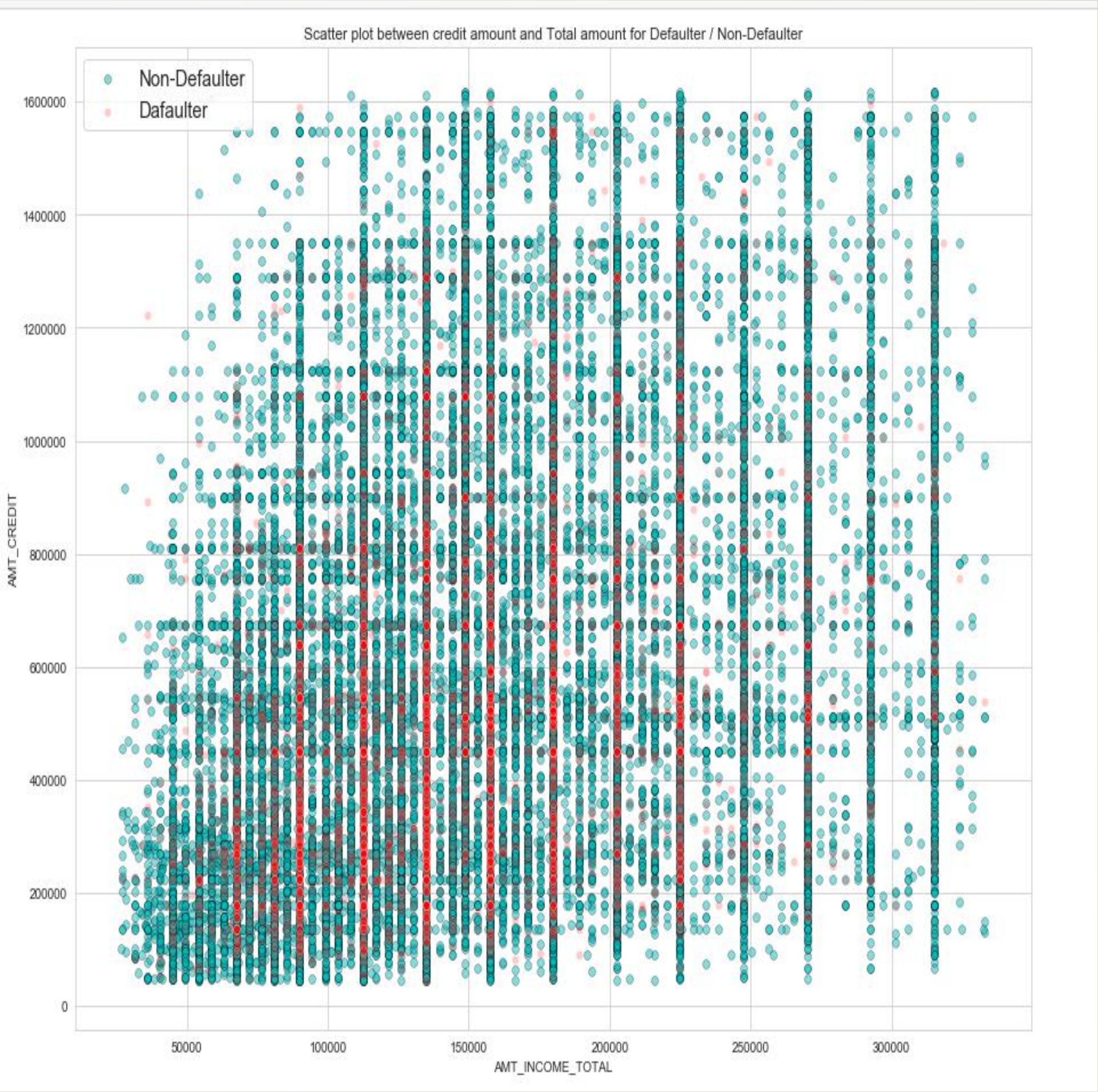
- Persons with 'Higher education', the income amount is mostly equal for all the family status.
- Lower secondary are having less income amount than others.



BIVARIATE VARIABLE ANALYSIS

INCOME TOTAL VS AMOUNT CREDIT

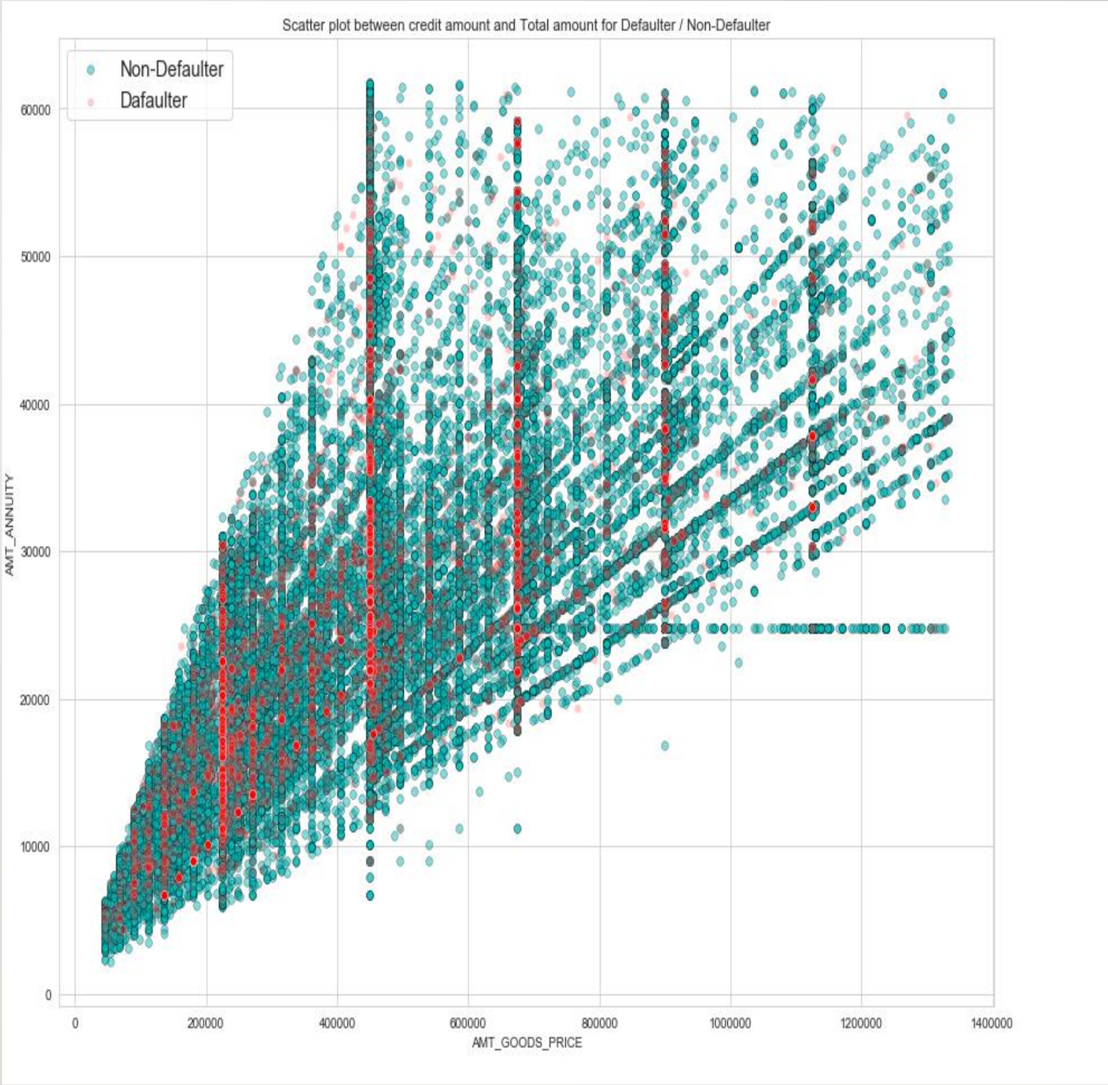
- loan amount between 20000 and 100000 and total income between 10000 to 25000 has most of the defaulter list



BIVARIATE VARIABLE ANALYSIS

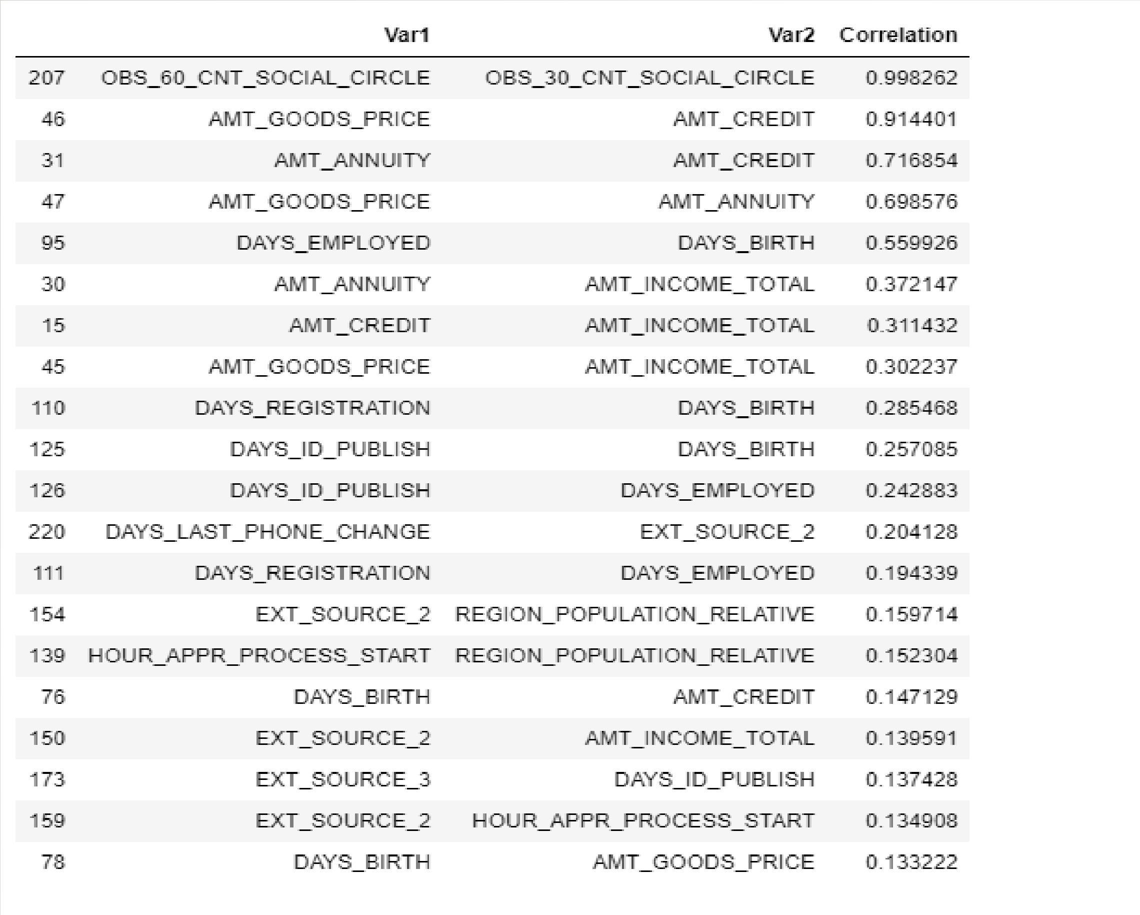
AMOUNT GOODS PRICE VS AMOUNT ANNUITY

- Amount goods price between 400000 and 800000 and amount annuity between 10000 and 40000 we have max no of defaulters



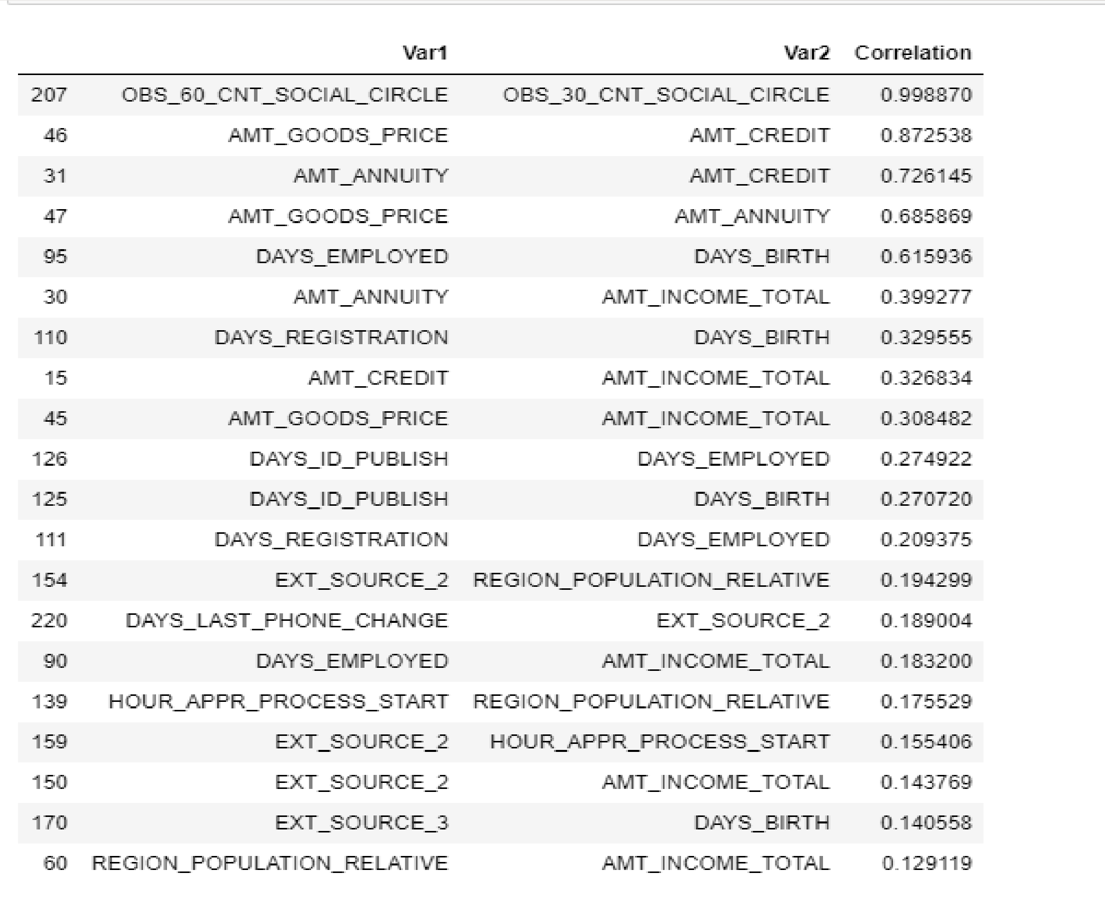
Finding the correlation between continuous variables

correlation between continuous variables For defaulters



	Var1	Var2	Correlation
207	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998262
46	AMT_GOODS_PRICE	AMT_CREDIT	0.914401
31	AMT_ANNUITY	AMT_CREDIT	0.716854
47	AMT_GOODS_PRICE	AMT_ANNUITY	0.698576
95	DAYS_EMPLOYED	DAYS_BIRTH	0.559926
30	AMT_ANNUITY	AMT_INCOME_TOTAL	0.372147
15	AMT_CREDIT	AMT_INCOME_TOTAL	0.311432
45	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.302237
110	DAYS_REGISTRATION	DAYS_BIRTH	0.285468
125	DAYS_ID_PUBLISH	DAYS_BIRTH	0.257085
126	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.242883
220	DAYS_LAST_PHONE_CHANGE	EXT_SOURCE_2	0.204128
111	DAYS_REGISTRATION	DAYS_EMPLOYED	0.194339
154	EXT_SOURCE_2	REGION_POPULATION_RELATIVE	0.159714
139	HOUR_APPR_PROCESS_START	REGION_POPULATION_RELATIVE	0.152304
76	DAYS_BIRTH	AMT_CREDIT	0.147129
150	EXT_SOURCE_2	AMT_INCOME_TOTAL	0.139591
173	EXT_SOURCE_3	DAYS_ID_PUBLISH	0.137428
159	EXT_SOURCE_2	HOUR_APPR_PROCESS_START	0.134908
78	DAYS_BIRTH	AMT_GOODS_PRICE	0.133222

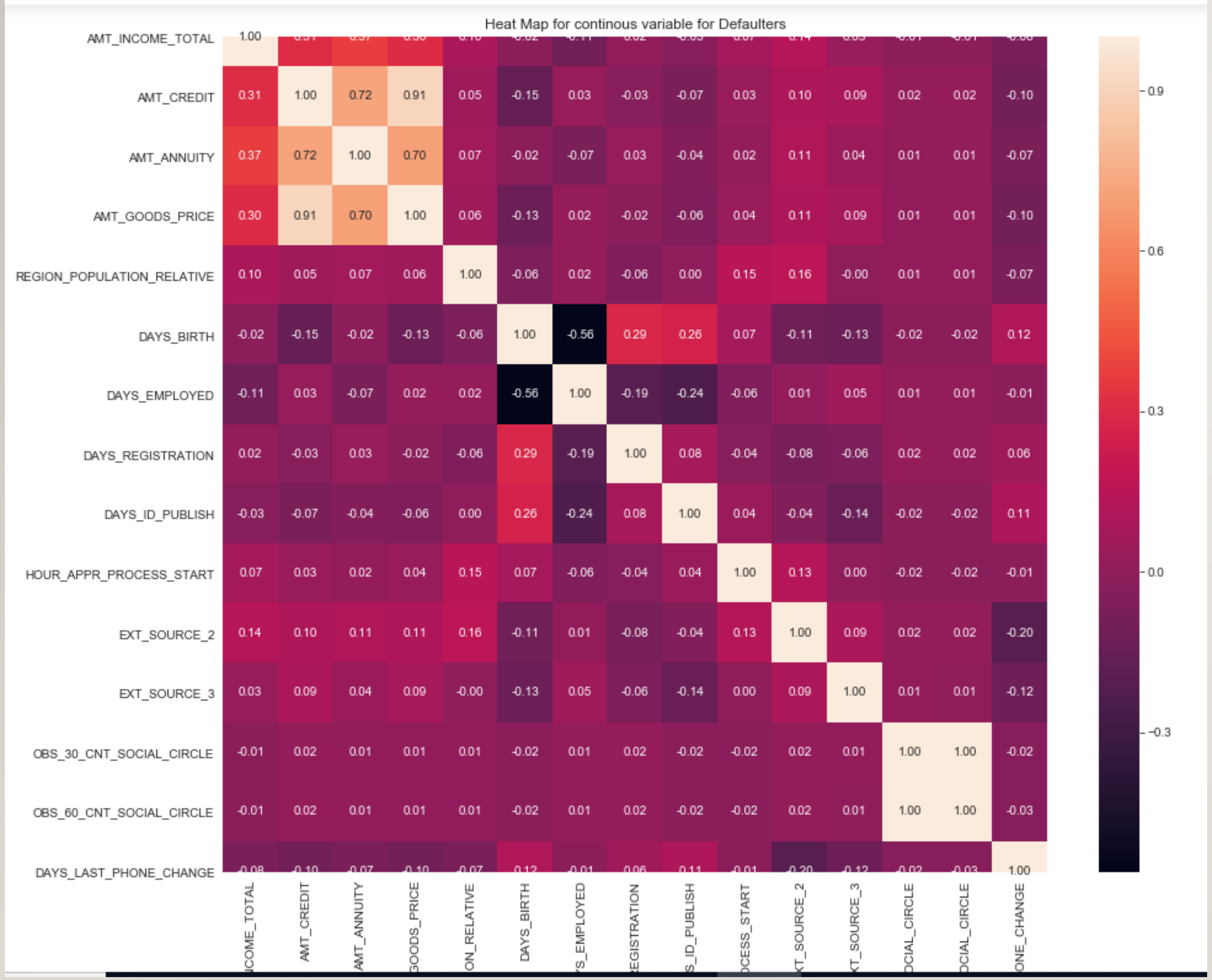
correlation between continuous variables For Non defaulters



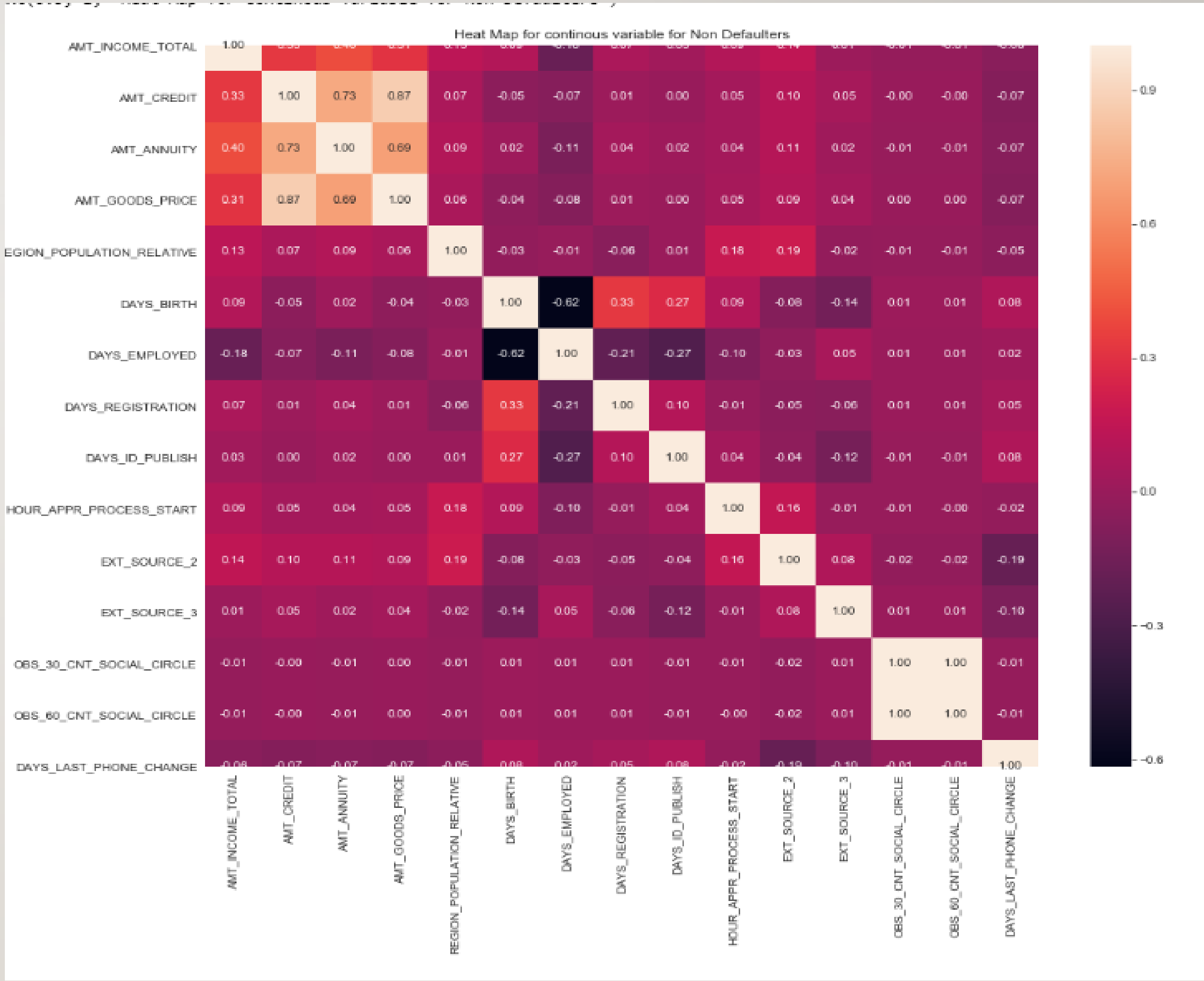
	Var1	Var2	Correlation
207	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998870
46	AMT_GOODS_PRICE	AMT_CREDIT	0.872538
31	AMT_ANNUITY	AMT_CREDIT	0.726145
47	AMT_GOODS_PRICE	AMT_ANNUITY	0.685869
95	DAYS_EMPLOYED	DAYS_BIRTH	0.615936
30	AMT_ANNUITY	AMT_INCOME_TOTAL	0.399277
110	DAYS_REGISTRATION	DAYS_BIRTH	0.329555
15	AMT_CREDIT	AMT_INCOME_TOTAL	0.326834
45	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.308482
126	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.274922
125	DAYS_ID_PUBLISH	DAYS_BIRTH	0.270720
111	DAYS_REGISTRATION	DAYS_EMPLOYED	0.209375
154	EXT_SOURCE_2	REGION_POPULATION_RELATIVE	0.194299
220	DAYS_LAST_PHONE_CHANGE	EXT_SOURCE_2	0.189004
90	DAYS_EMPLOYED	AMT_INCOME_TOTAL	0.183200
139	HOUR_APPR_PROCESS_START	REGION_POPULATION_RELATIVE	0.175529
159	EXT_SOURCE_2	HOUR_APPR_PROCESS_START	0.155406
150	EXT_SOURCE_2	AMT_INCOME_TOTAL	0.143769
170	EXT_SOURCE_3	DAYS_BIRTH	0.140558
60	REGION_POPULATION_RELATIVE	AMT_INCOME_TOTAL	0.129119

Finding the correlation between continuous variables

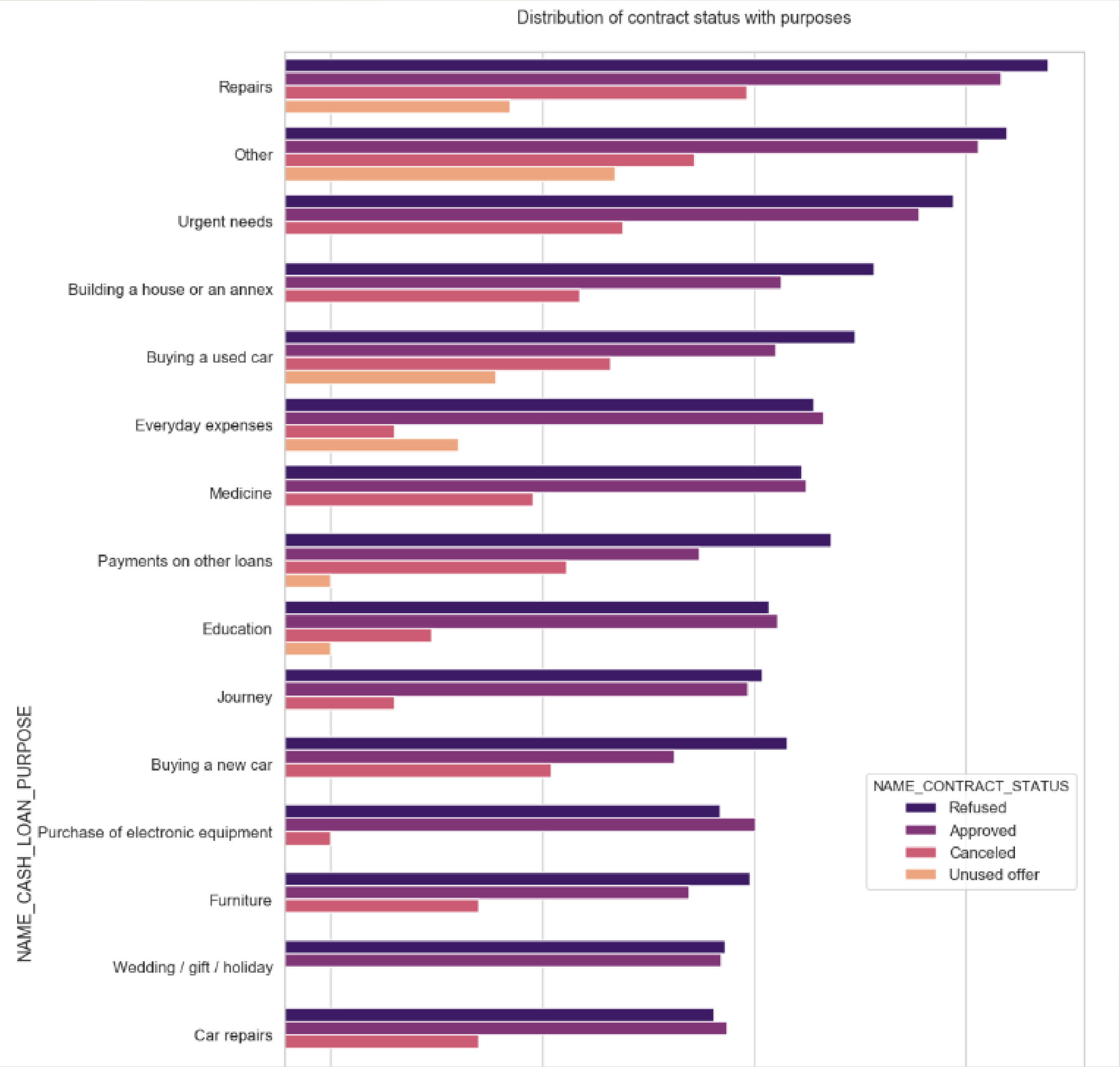
correlation between continuous variables For defaulters



correlation between continuous variables For Non defaulters



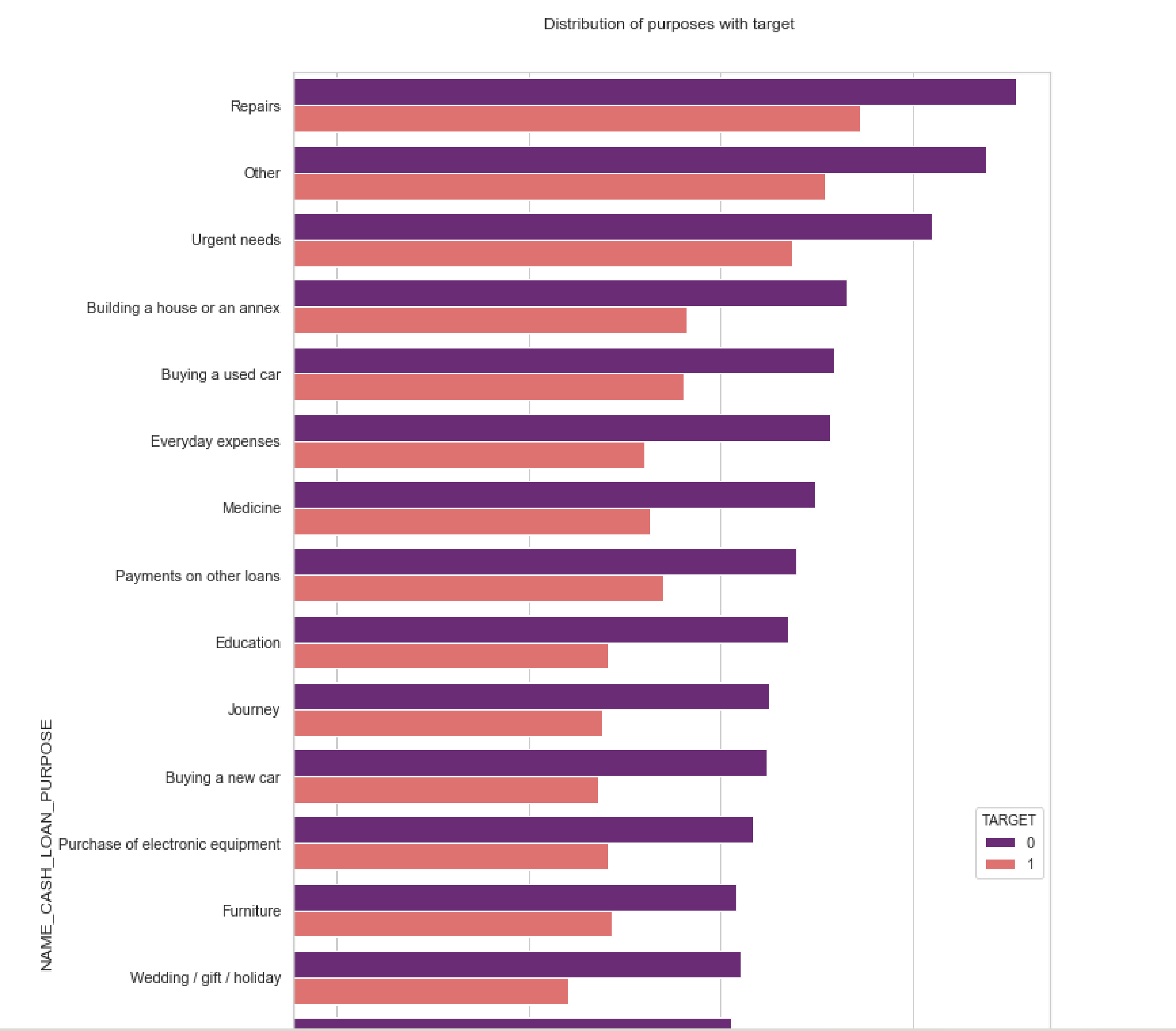
contract status with purposes in logarithmic scale



CATEGORICAL VARIABLE: PREVIOUS LOAN APPLICATION STATUS

- Loan taken fr purpose of repair has highest no of rejection
- For education purposes we have almost equal number of approves and rejection
- loan taken for the purpose of "Payign other loans" and "buying a new car" is having significant higher rejection than approves

Distribution of contract status with Taget



CATEGORICAL VARIABLE: PREVIOUS LOAN APPLICATION STATUS

- Loan taken for the purpose of 'Repairs' are most of te defaulters.
- Loan taken for the purpoese of 'Buying a garage', 'Business developemt', 'Buying land','Buying a new car' and 'Education' are most of the defaulters.
- Hence we can focus on these purposes for which the client is having for minimal payment difficulties.

FINDINGS FROM THE ANALYSIS

1. People who are taking large amount of loan are likely to repay the loan
2. Persons with age between 30 years and 50 years has high number of defaulter.
3. person with id changed in 1000 days of application has high number of defaulters.
4. From the above plot we can conclude that All the Students and Businessman are repaying loan.
5. Widows are more likely to repay the loan when compared to appliants with the other family statuses.
6. People with Academic Degree are more likely to repay the loan only 0.0198% have not repayed the loan.
7. The focus of the bank should be more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.
8. The focus of the bank should be less on income type ‘Working’ as they are having most number of unsuccessful payments.
9. Loan taken for the purpose ‘Repair’ is having higher number of unsuccessful payments on time.