LOAN DEFAULTER Case Study

Done By

ANUPAMA RAJEEV

SUMITHA T

Problem Statement

Challenge:

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.

Consider a scenario where Mr. X is applying for loan in a bank .There are two possible errors that the bank can commit.

1. Credit loss**:** Bank will approve loan of Mr. X who could be a possible defaulter there by incurring loss of entire credit amount.
2. Interest loss : Bank rejects loan request of Mr. X who can make timely payments there by losing a good customer and interest on loan amount.

Solution:

This case study aims to identify patterns using EDA, from existing data available about the clients and past loan history of the clients. These patterns will provide insights, if client could be a possible loan defaulter and helps bank to take actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc

Analysis Approach

We have two different datasets.

1.***Application Dataset*** - The current information of the customer.

2.***Previous Application Dataset*** – The information about the previous loan data.

The datasets are subjected to below EDA processes for analysis.

- ✓ **Understanding the Dataset -** Reading the dataset, doing basic checks like datatypes of each column, shape, statistical description of numerical columns.
- ✓ **Data Cleaning -** Involves checking null/missing values ,fixing incorrect datatypes, analyzing the outliers and binning of continuous variables.
- ✓ **Univariate Analysis -** Distribution of continuous variables and count of categorical variables.
- ✓ **Bivariate Analysis -** Discovering patterns between two variables which involve:
    1. Categorical and categorical – e.g. Count plot
    2. Continuous to Continuous – e.g. Scatter plot
    3. Categorical to Continuous – e.g. Box plot
- ✓ **Multivariate Analysis -** Using more than two variables to discover trends between features. E.g. Heatmap
- ✓ **Deriving conclusions -** Getting insights.

Application Data set

- Below shows the first few records and columns of 'Application' data set.
- File is read and saved to data frame name 'appdata'.
- Data frame has 307511 rows and 122 columns.
- Column name 'TARGET' is the target variable with values 1 and 0. 1 being a defaulter and 0 being non-defaulter.

	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	202500.0000	406597.5
1	100003	0	Cash loans	F	N	N	0	270000.0000	1293502.5
2	100004	0	Revolving loans	M	Y	Y	0	67500.0000	135000.0
3	100006	0	Cash loans	F	N	Y	0	135000.0000	312682.5
4	100007	0	Cash loans	M	N	Y	0	121500.0000	513000.0

```
#checking for number of rows and columns in the dataframe.

appdata.shape
```

```
(307511, 122)
```

DATA CLEANING: Application Data set

- Below shows the list of some columns that has null values greater than or equal to 50%. Such columns won't be of much use for analysis. So dropped them.
- There are other columns too, which are not useful for this point of analysis. So made a list, and dropped them as well.
- ***Final data frame, has 307511 rows and 41 columns***

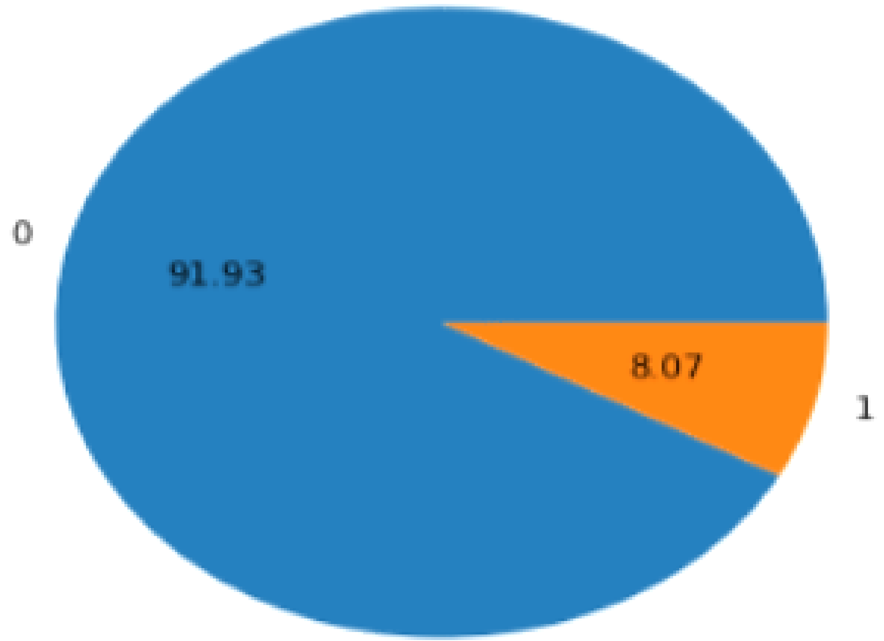
OWN_CAR_AGE	65.9908
EXT_SOURCE_1	56.3811
APARTMENTS_AVG	50.7497
BASEMENTAREA_AVG	58.5160
YEARS_BUILD_AVG	66.4978
COMMONAREA_AVG	69.8723
ELEVATORS_AVG	53.2960
ENTRANCES_AVG	50.3488
FLOORSMIN_AVG	67.8486
LANDAREA_AVG	59.3767
LIVINGAPARTMENTS_AVG	68.3550
LIVINGAREA_AVG	50.1933
NONLIVINGAPARTMENTS_AVG	69.4330
NONLIVINGAREA_AVG	55.1792
APARTMENTS_MODE	50.7497
BASEMENTAREA_MODE	58.5160
YEARS_BUILD_MODE	66.4978
COMMONAREA_MODE	69.8723
ELEVATORS_MODE	53.2960
ENTRANCES_MODE	50.3488
FLOORSMIN_MODE	67.8486

```
#list of unwanted columns

unwanted_cols = ['SK_ID_CURR','FLAG_MOBIL','FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE',
       'FLAG_EMAIL','REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY','YEARS_BEGINEXPLUATATION_AVG', 'FLOORSMAX_AVG',
        'YEARS_BEGINEXPLUATATION_MODE', 'FLOORSMAX_MODE','YEARS_BEGINEXPLUATATION_MEDI', 'FLOORSMAX_MEDI', 'TOTALAREA_MODE',
        'EMERGENCYSTATE_MODE', 'FLAG_DOCUMENT_2','FLAG_DOCUMENT_3', 'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5','FLAG_DOCUMENT_6',
        'FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8','FLAG_DOCUMENT_9', 'FLAG_DOCUMENT_10', 'FLAG_DOCUMENT_11','FLAG_DOCUMENT_12',
        'FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14','FLAG_DOCUMENT_15', 'FLAG_DOCUMENT_16', 'FLAG_DOCUMENT_17','FLAG_DOCUMENT_18',
        'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20','FLAG_DOCUMENT_21','REG_REGION_NOT_WORK_REGION','REG_CITY_NOT_WORK_CITY',
        'REGION_POPULATION_RELATIVE']
```

```
#removing all unwanted columns

appdata.drop(labels=unwanted_cols, axis=1, inplace=True)
```

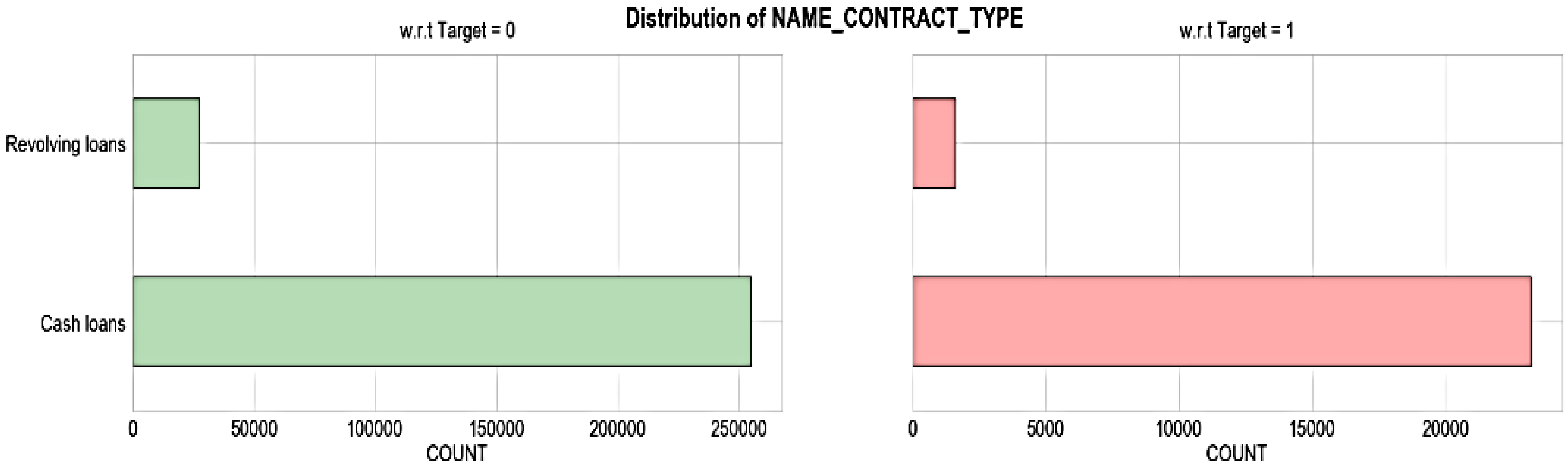
Data Imbalance

- As illustrated in the graph, the application data set has a very visible data imbalance in terms of defaulters and non- defaulters.,
  - 8.07% individuals are defaulters.
  - 91.93% are non-defaulters.
- The approach followed to deal with the data imbalance is divide and analyze the dataset based on defaulters and non-defaulters.
- In the following sections Target=1 stands for defaulters and Target =0 stands for non-defaulters.



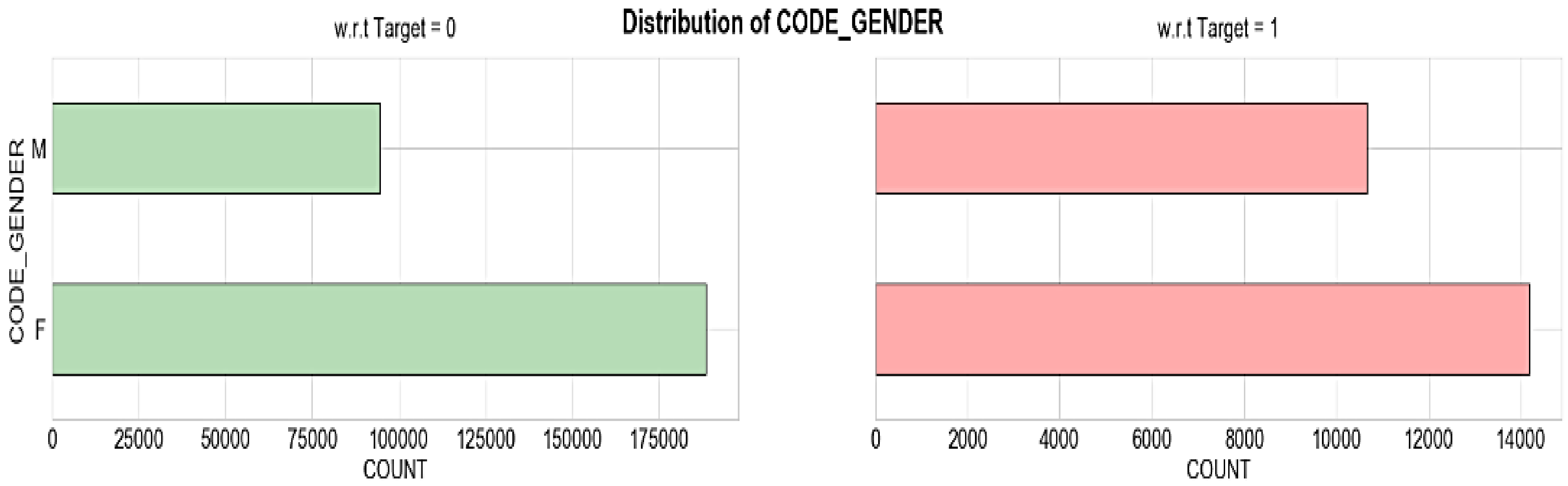
Univariate Analysis

- **Categorical Variable :** Distribution of Contract Type.
- **Conclusions :** Among defaulters and non-defaulters 'Cash Loans' category count is more compared to 'Revolving Loans.



Univariate Analysis

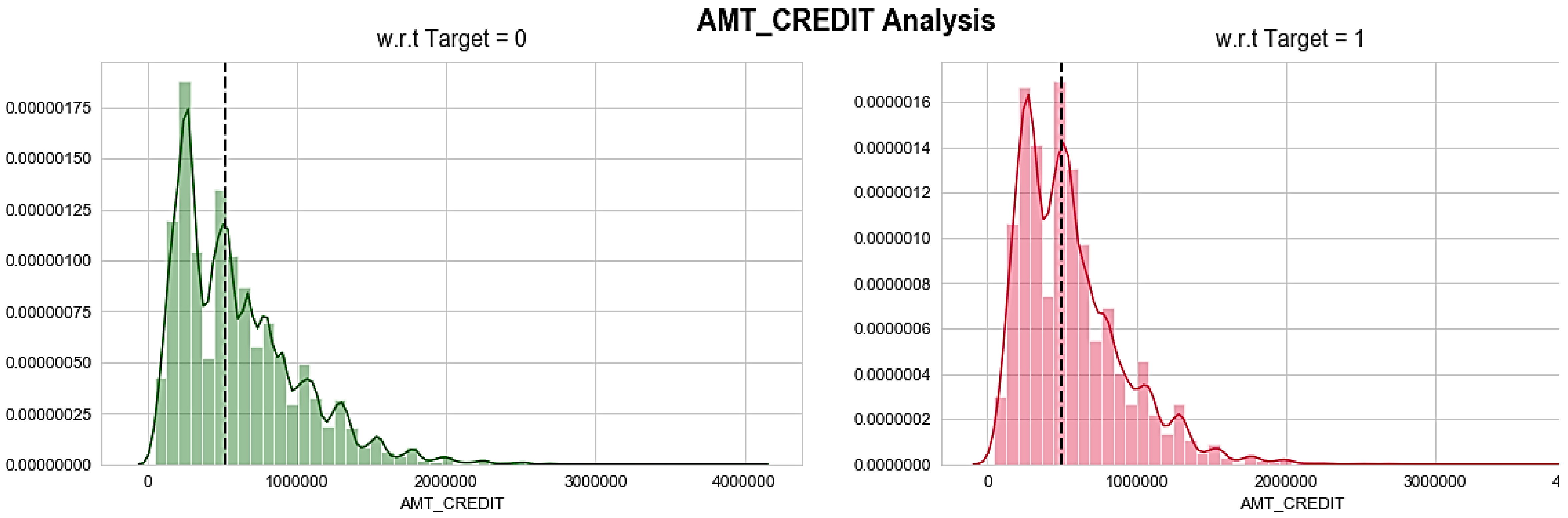
- **Categorical Variable :** Distribution of Gender type.
- **Conclusions :** Among defaulters and non-defaulters 'Female' gender category count is more compared to males.



Univariate Analysis

- **Continuous Variable :** Distribution of Credit Amount
- **Conclusions :** Most of the data falls to the right of the graph peak, it shows the presence of outliers.

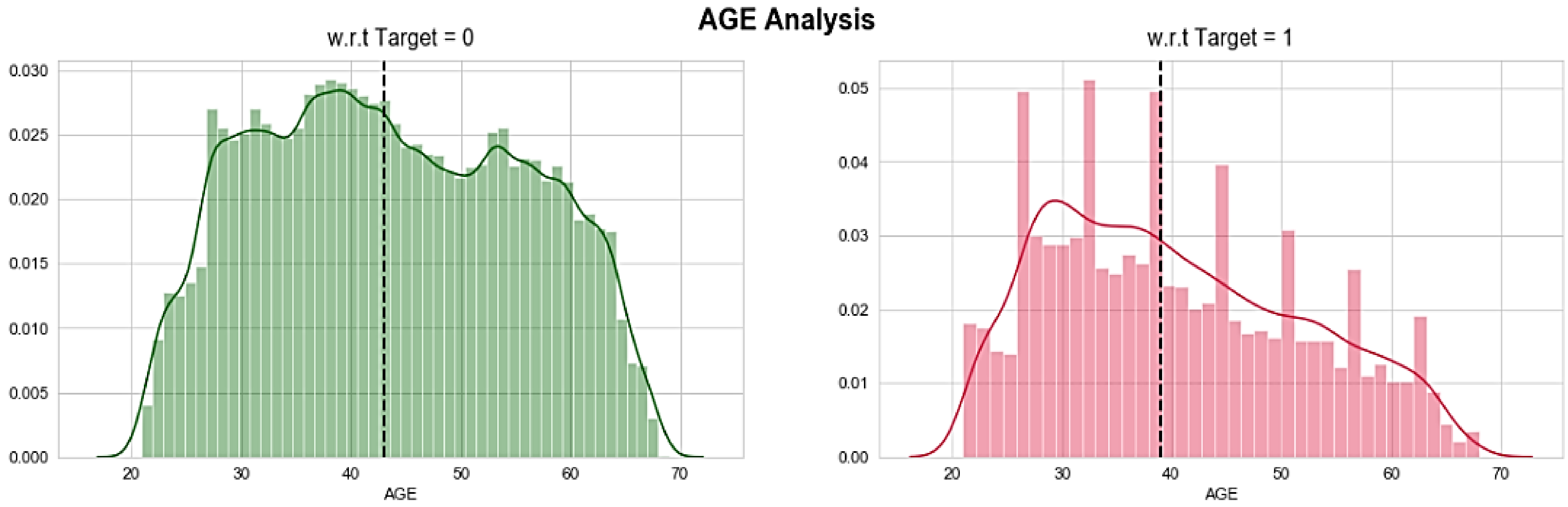
Black dotted lines in the graph represents the median.



Univariate Analysis

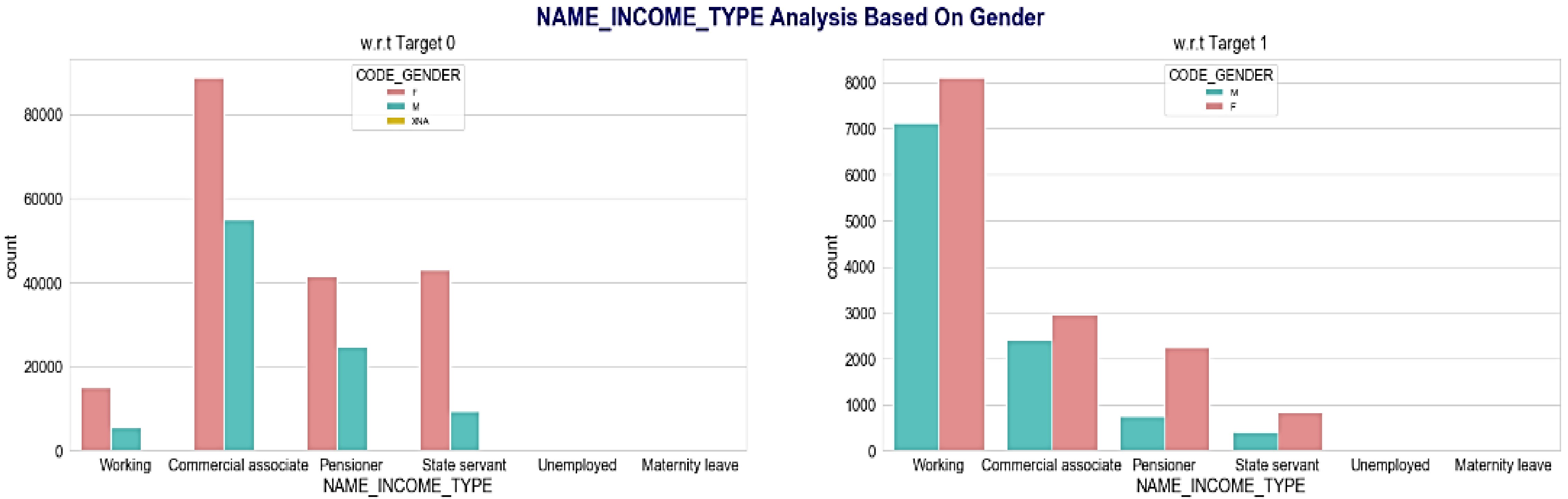
- **Continuous Variable :** Distribution of Age
- **Conclusions :** In target = 0 data set, age column has median value 43. In target = 1 data set, age column has median value 39. In both data set, there is no presence of outliers.

Black dotted lines in the graph represents the median.



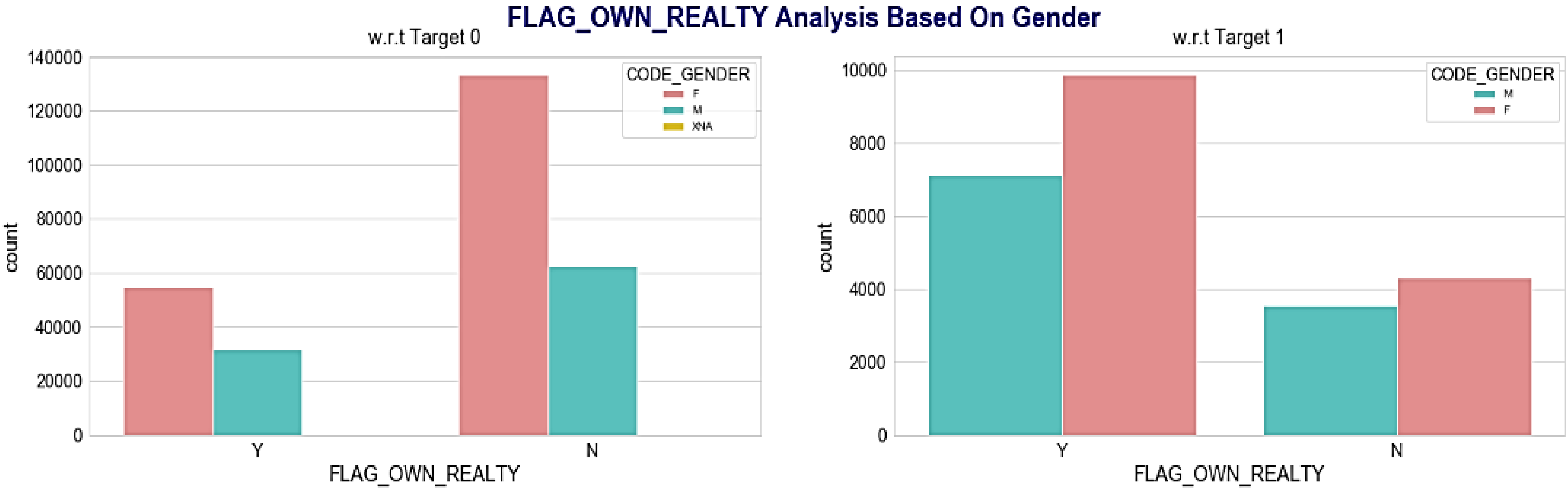
Bivariate Analysis

- **Categorical - Categorical:** NAME_INCOME_TYPE vs CODE_GENDER
- **Conclusions :**
- In target=0, Commercial associates are the highest category.
- In target = 1, Working category has the highest number of defaulters.



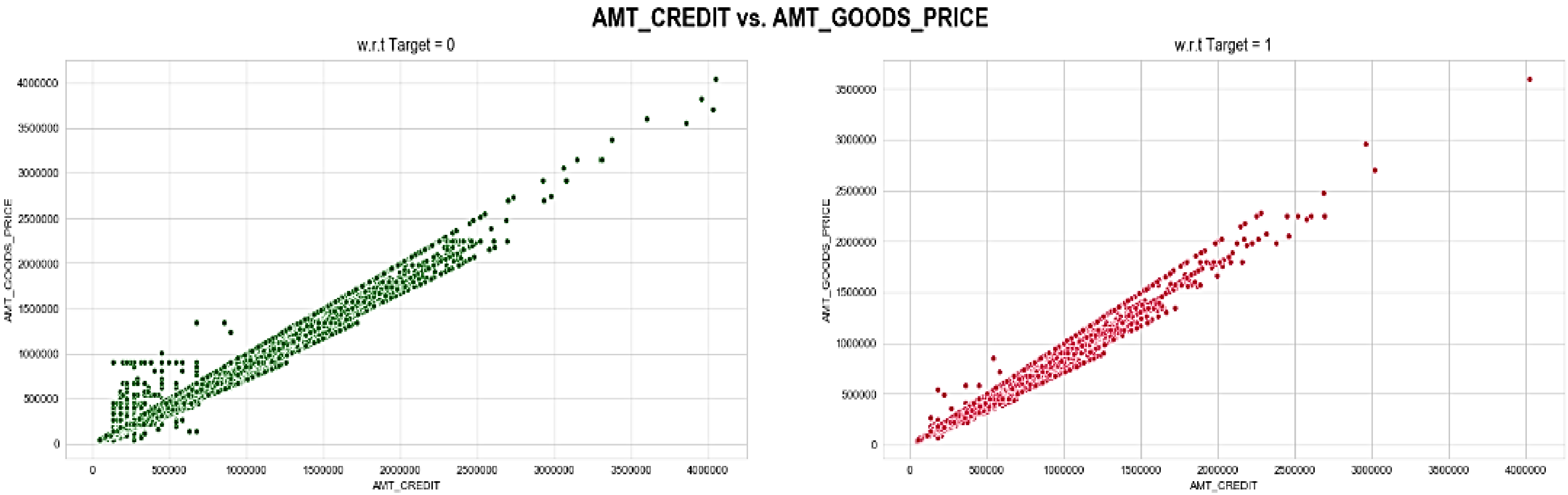
Bivariate Analysis

- **Categorical - Categorical:** FLAG_OWN_RELATY vs CODE_GENDER
- **Conclusions :** Customers who own a realty tend to default more as shown in target=1.



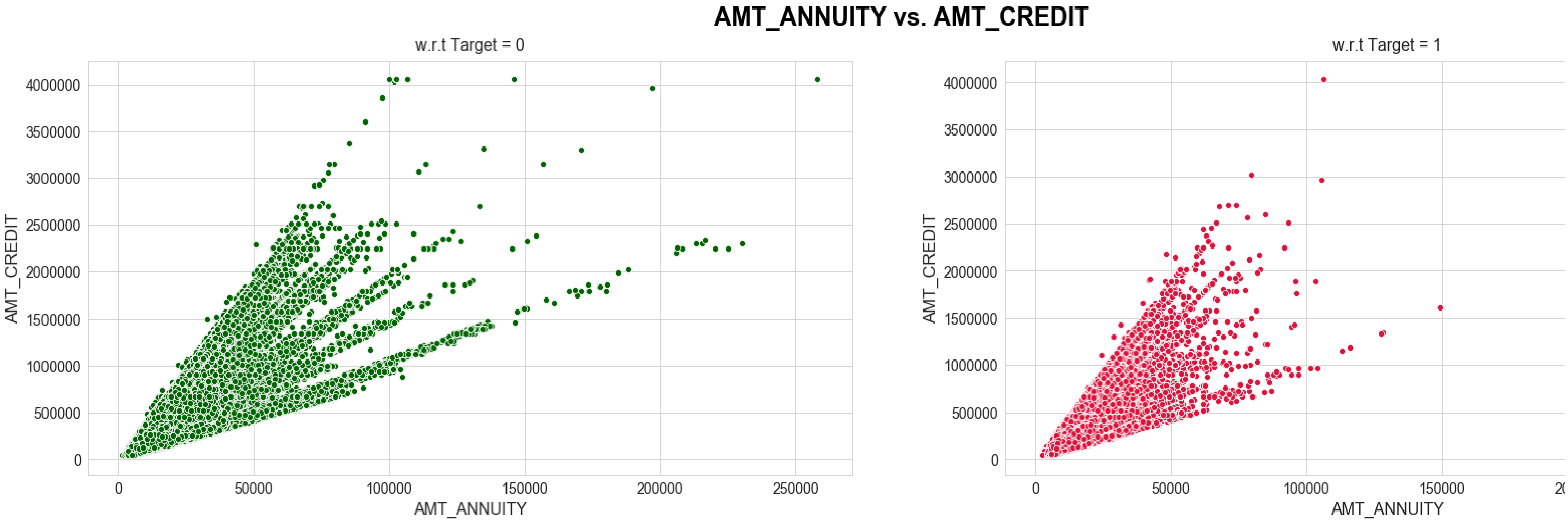
Bivariate Analysis

- **Continuous - Continuous** : AMT_CREDIT vs AMT_GOODS_PRICE
- **Conclusions :**
- AMT_GOODS_PRICE and AMT_CREDIT has a positive correlation.
- As the credit amount increases the Goods price also increases.



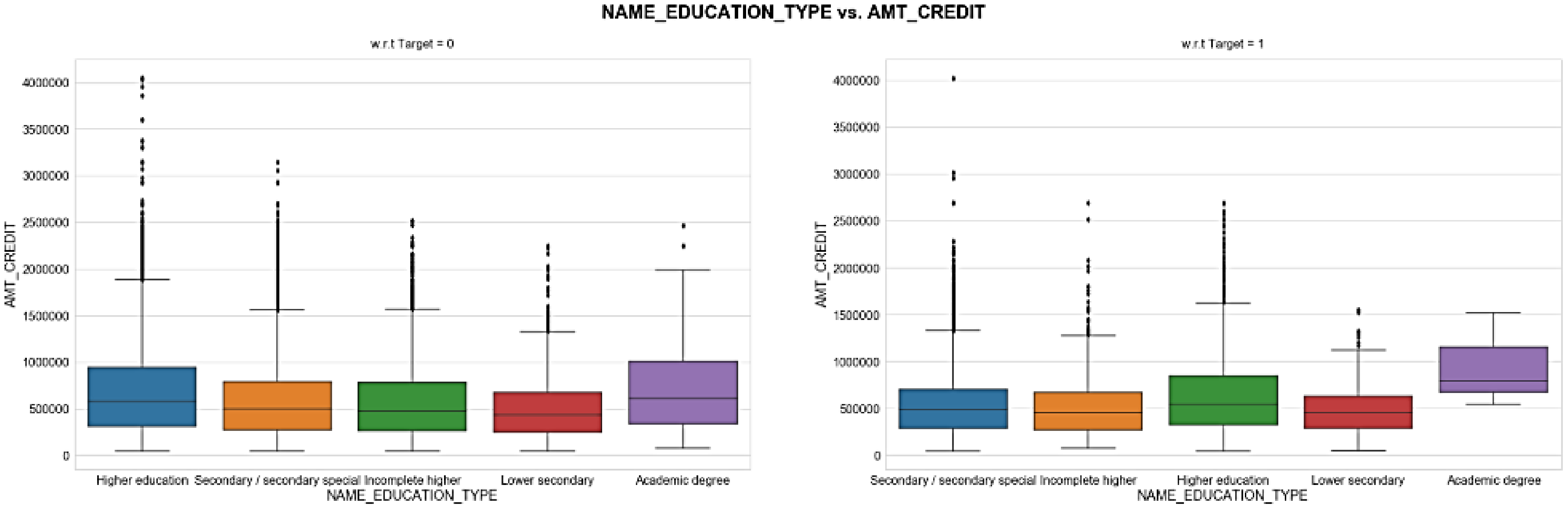
Bivariate Analysis

- **Continuous - Continuous** : AMT_ANNUITY vs AMT_CREDIT
- **Conclusions :**
- AMT_ANNUITY and AMT_CREDIT has a positive correlation.
- As the credit amount increases the annuity also increases.



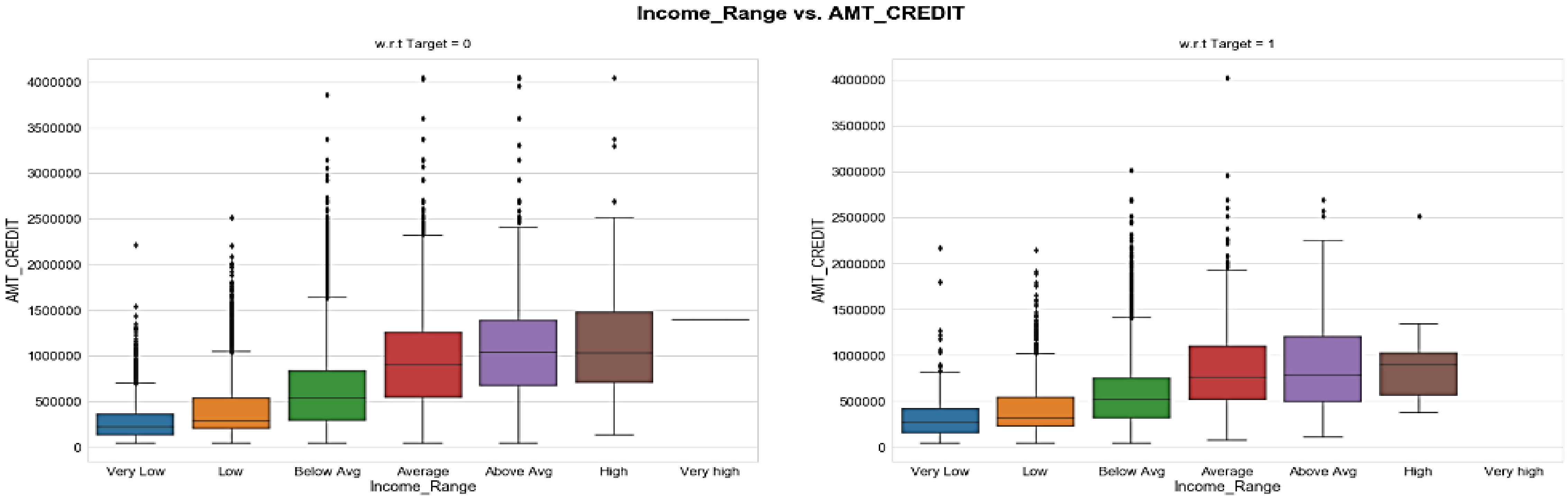
Bivariate Analysis

- **Categorical - Continuous:** NAME_EDUCATION_TYPE vs AMT_CREDIT
- **Conclusions :** The median of 'Academic Degree' category is higher in both categories. Also, 'Academic Degree' has lesser number of outliers compared to other categories.

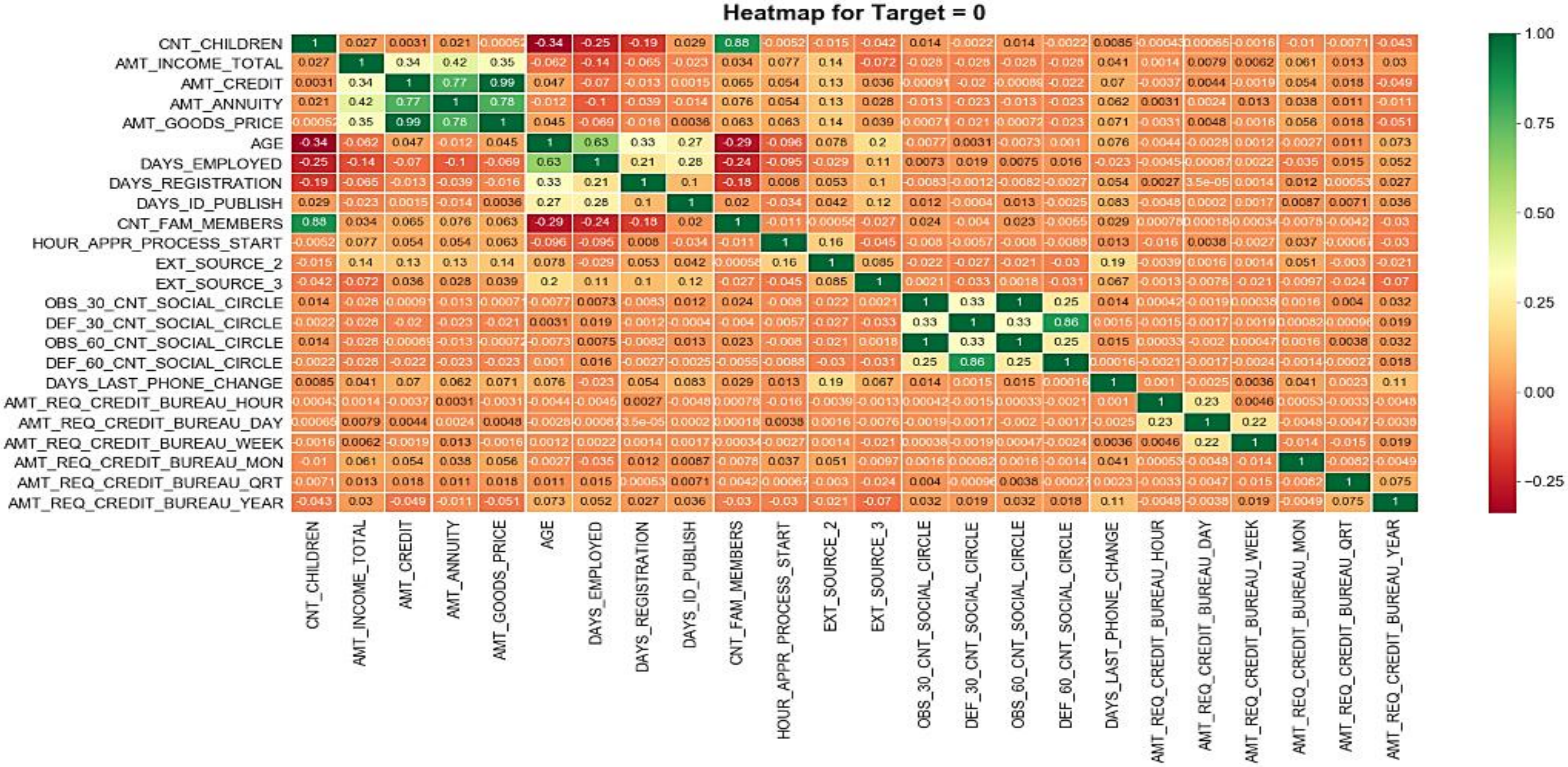


Bivariate Analysis

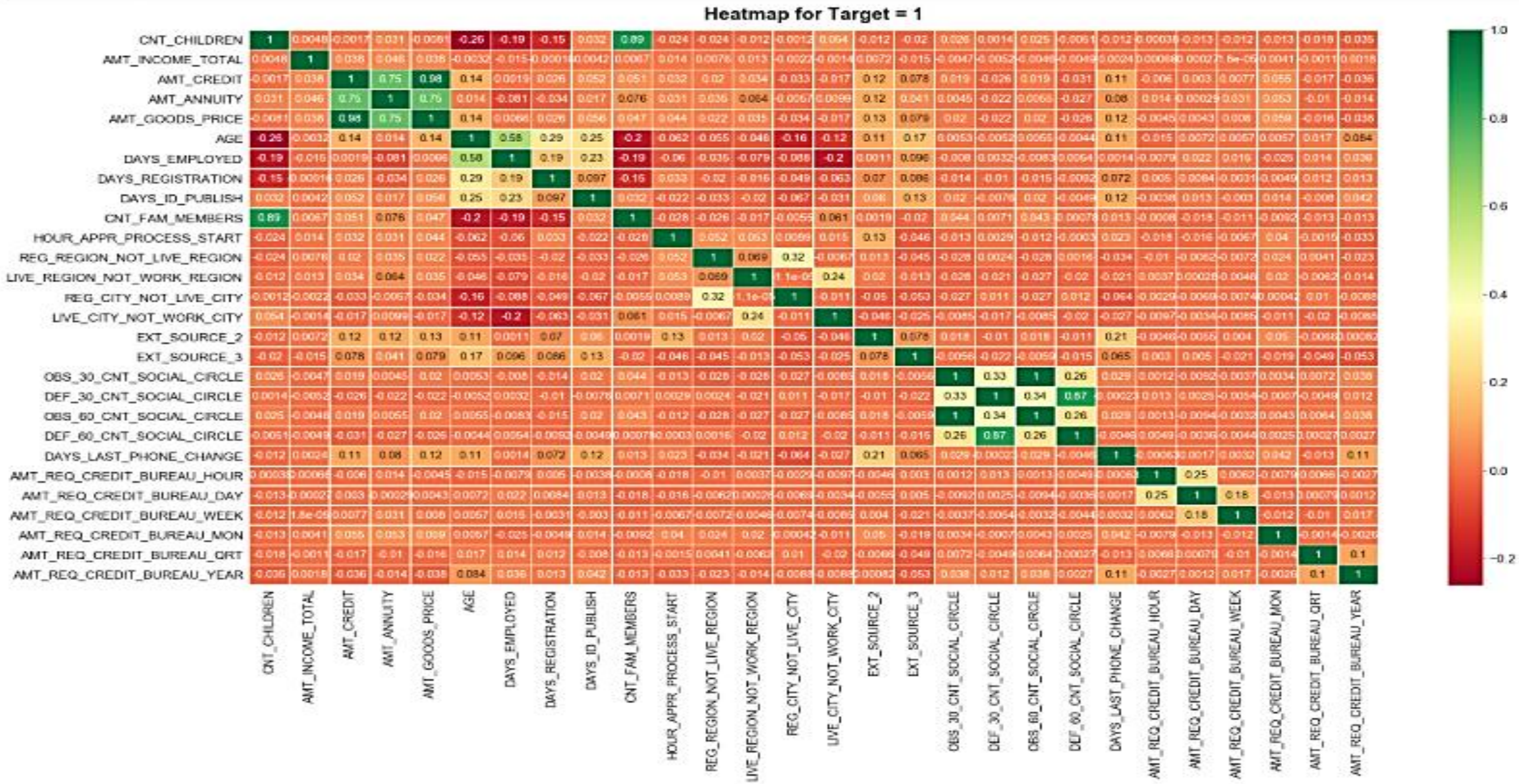
- **Categorical - Continuous:** Income_Range vs AMT_CREDIT
- **Conclusions :** In both dataset, it can be observed that, as the income range increases, amount applied for credit also increases



Multivariate analysis for Non-Defaulters [target = 0]



Multivariate analysis for Defaulters [target = 1]



Major Insights from the Application Data Set

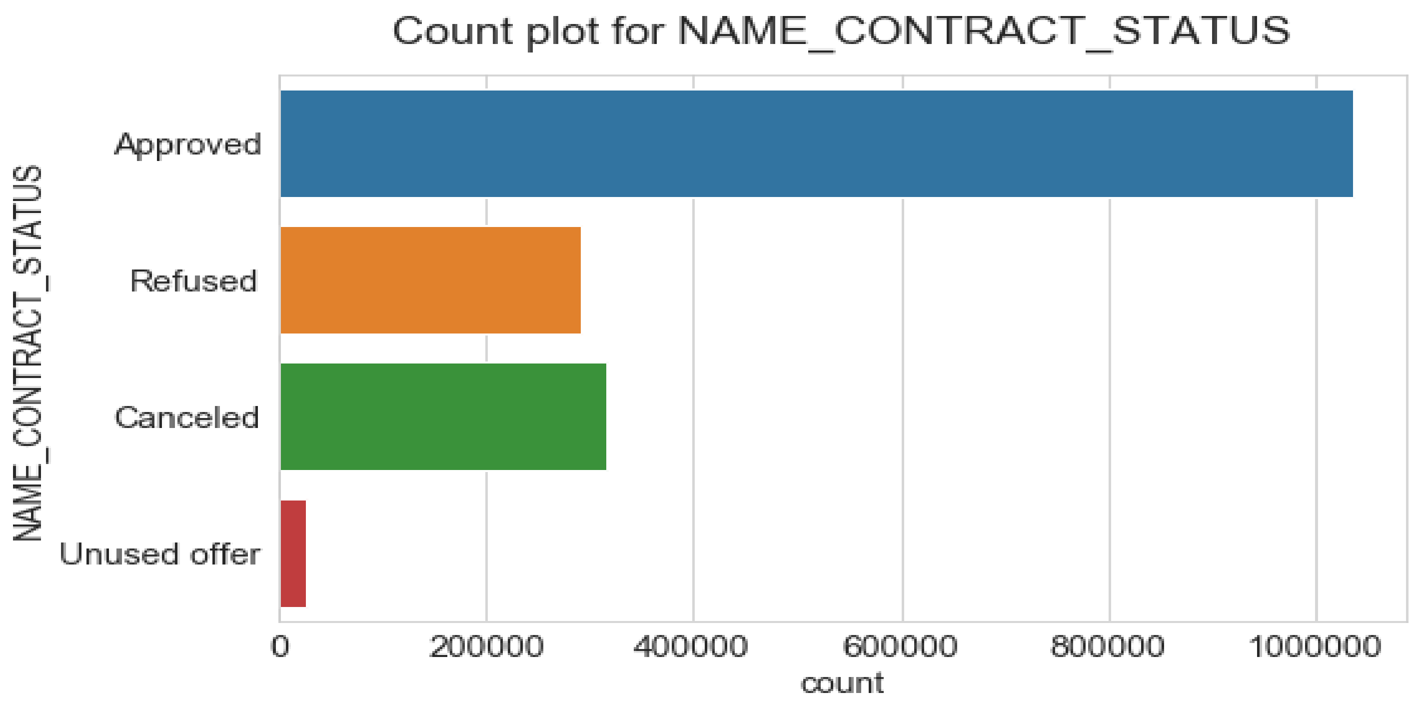
- 51 - 60 age group are highest in number for non-defaulters.
- 3L to 5.5L credit range have defaulted more whereas 1 - 3L range has the highest count in non-defaulters.
- More defaulters belong to the 'Female' category
- Males in 'Average` income type category have defaulted more compared to the females in the same category.
- Defaulters lies more in Secondary / Secondary special Education Type'. 'Incomplete higher' category group is the highest in non-defaulters
- Customers who own realty tend to default more

Top 5 correlation variables from heat map presented in previous slides (top 10 mentioned in notebook).

- OBS_60_CNT_SOCIAL_CIRCLE--OBS_30_CNT_SOCIAL_CIRCLE
- AMT_GOODS_PRICE--AMT_CREDIT
- CNT_FAM_MEMBERS--CNT_CHILDREN
- DEF_60_CNT_SOCIAL_CIRCLE--DEF_30_CNT_SOCIAL_CIRCLE
- AMT_ANNUITY--AMT_GOODS_PRICE

Segment 2 : Previous Application Data set

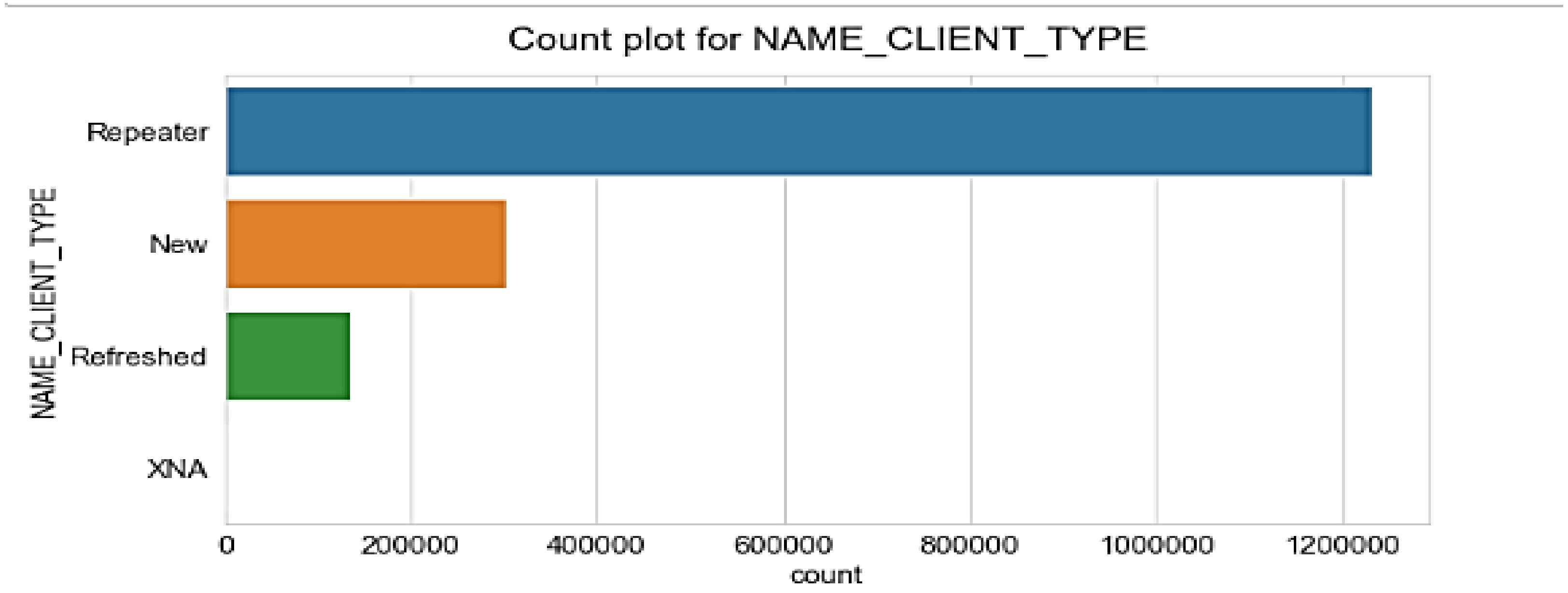
NAME_CONTRACT_STATUS is the target variable in this data set.




```
Normalized counts:
Approved        0.6207
Canceled        0.1894
Refused         0.1740
Unused offer    0.0158
Name: NAME_CONTRACT_STATUS, dtype: float64
```

Univariate Analysis

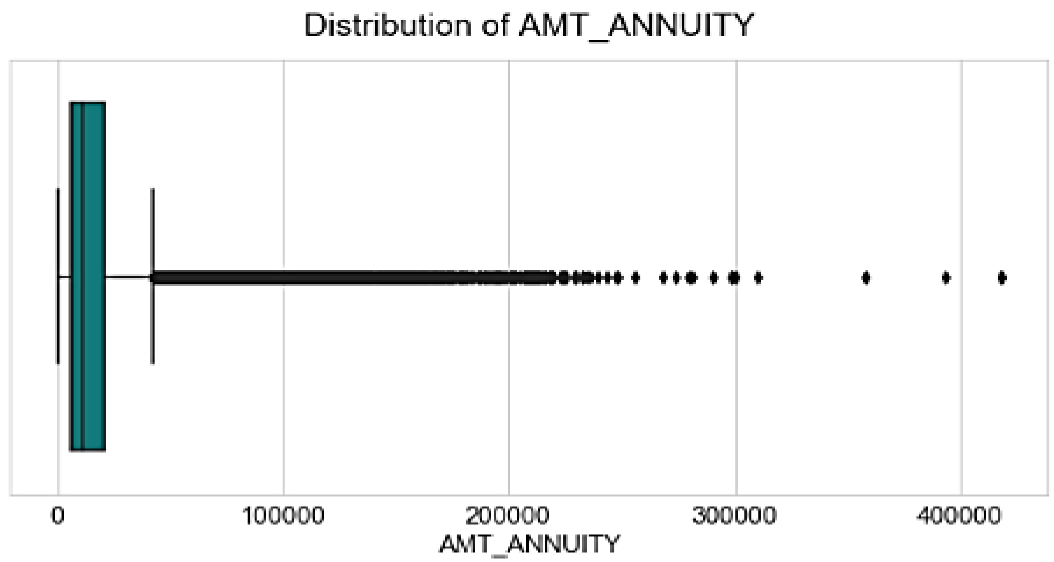
- **Categorical Variable :** NAME_CLIENT_TYPE
- **Conclusions :** 'Repeater' category client type is the highest.



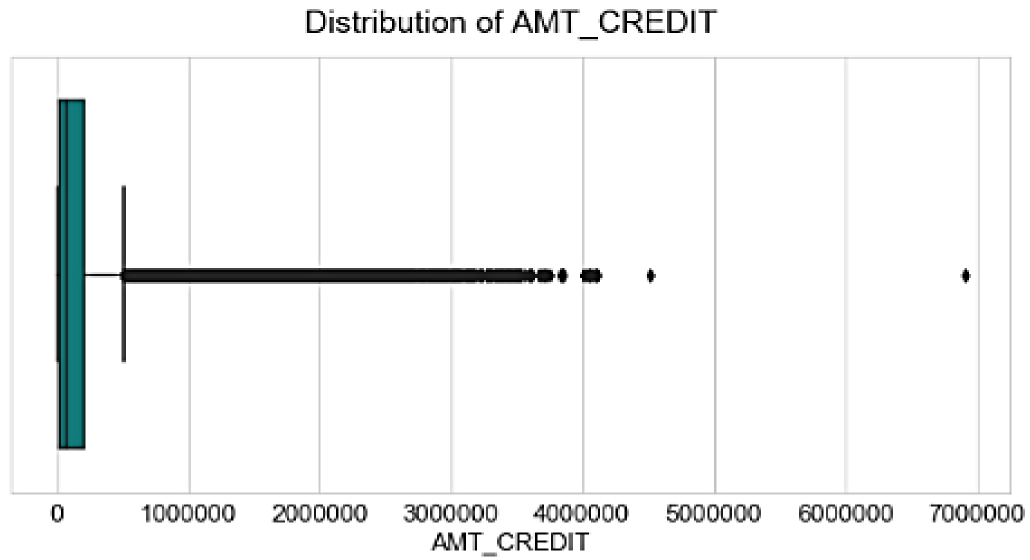

```
Normalized counts:
Repeater     0.7372
New          0.1804
Refreshed    0.0812
XNA          0.0012
Name: NAME_CLIENT_TYPE, dtype: float64
```

Univariate Analysis

- **Continuous Variable:** AMT_CREDIT and AMT_ANNUITY
- **Conclusions :** Both variable have presence of outliers.



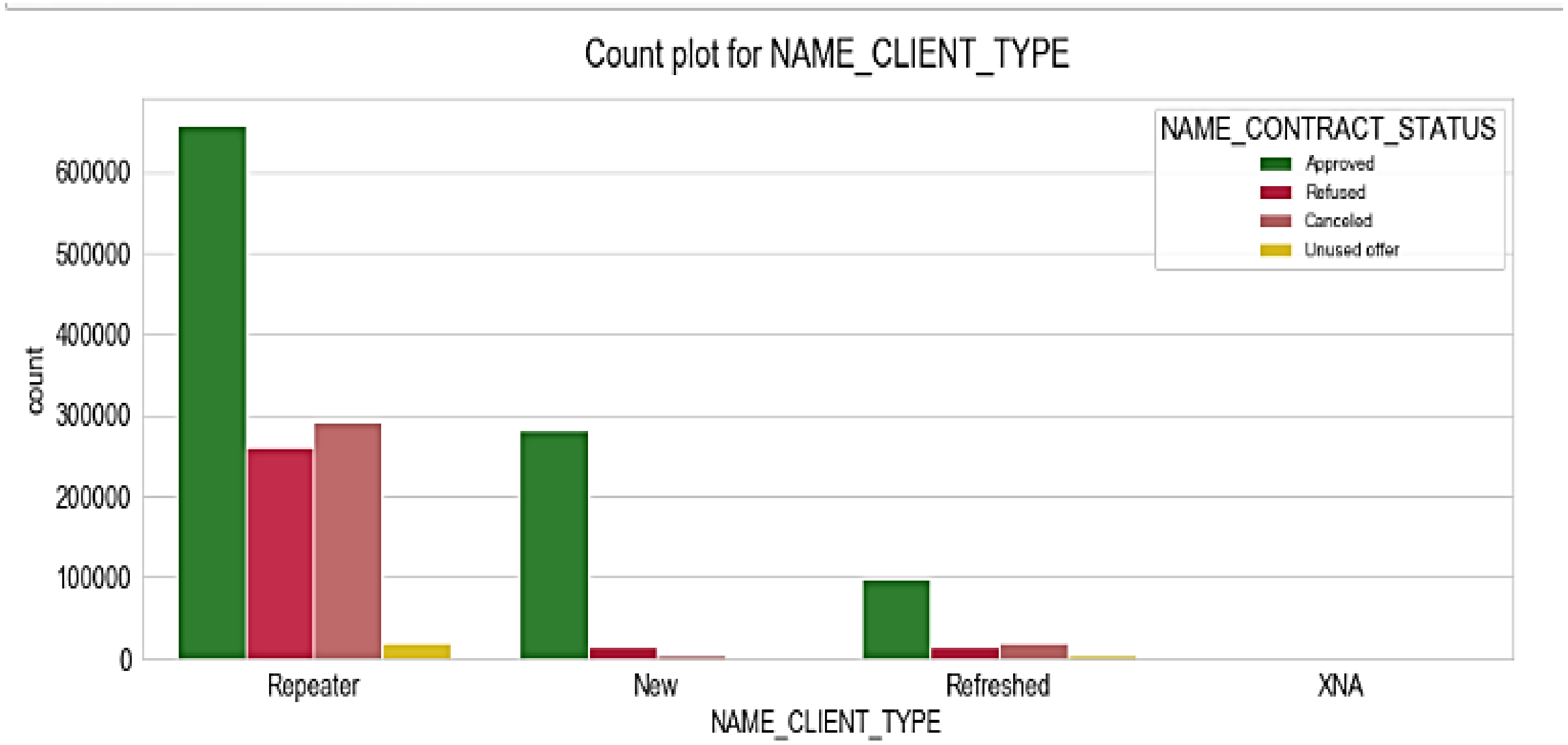

```
Statistical Description:
count   1297979.0000
mean      15955.1207
std       14782.1373
min           0.0000
25%        6321.7800
50%       11250.0000
75%       20658.4200
max      418058.1450
Name: AMT_ANNUITY, dtype: float64
```




```
Statistical Description:
count   1670213.0000
mean     196114.0212
std      318574.6165
min           0.0000
25%       24160.5000
50%       80541.0000
75%      216418.5000
max     6905160.0000
Name: AMT_CREDIT, dtype: float64
```

Bivariate Analysis

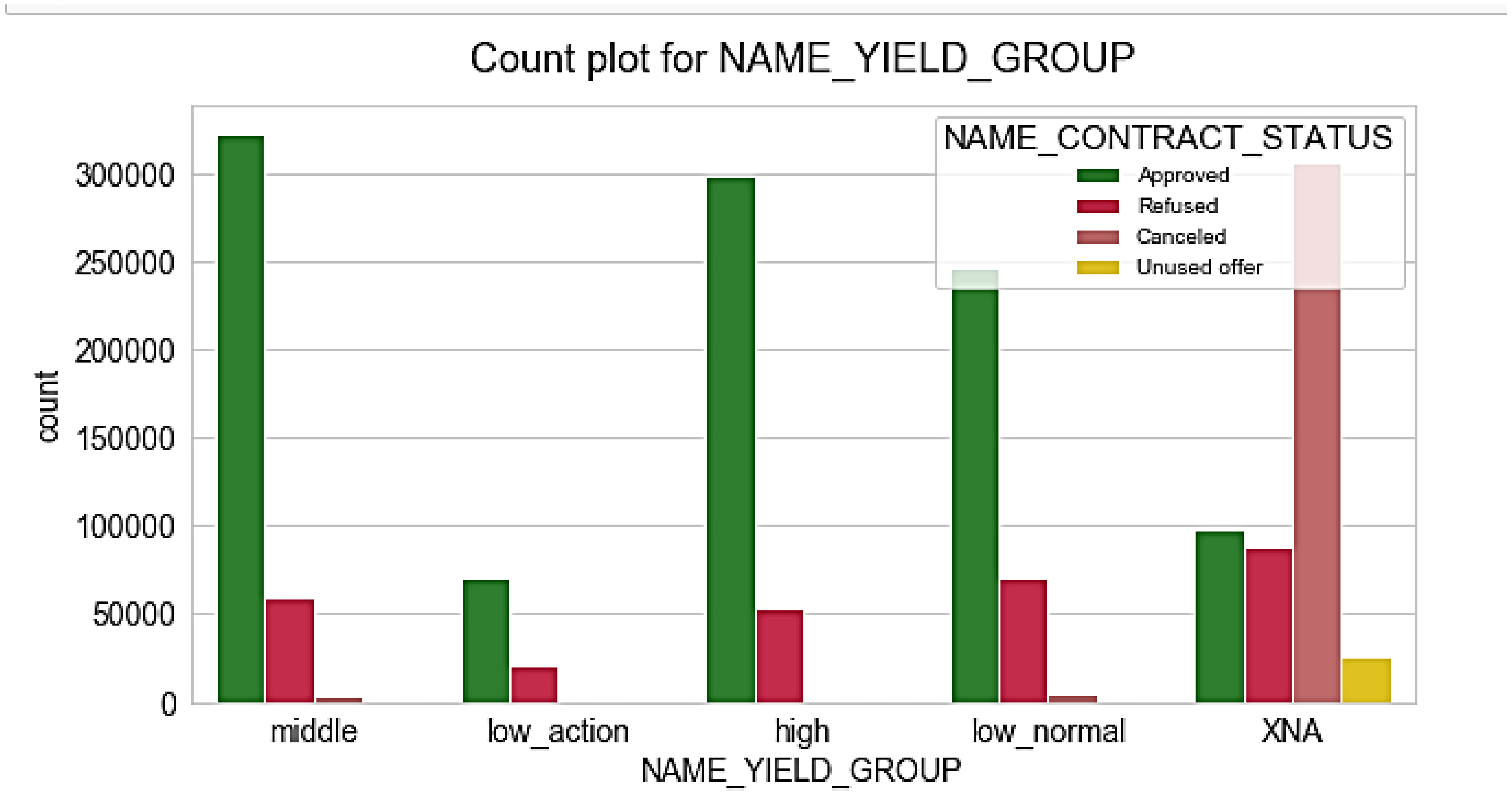
- **Categorical Variable :** NAME_CLIENT_TYPE *grouped by* NAME_CONTRACT_STATUS
- **Conclusions :** 'Repeater' category client type has the highest number of approved loans and Rejected loans.




```
NAME_CONTRACT_STATUS  NAME_CLIENT_TYPE
Approved              Repeater            0.6345
                      New                 0.2713
                      Refreshed           0.0937
                      XNA                 0.0006
Canceled              Repeater            0.9239
                      Refreshed           0.0618
                      New                 0.0112
                      XNA                 0.0031
Refused               Repeater            0.8974
                      Refreshed           0.0517
                      New                 0.0496
                      XNA                 0.0012
Unused offer          Repeater            0.7688
                      Refreshed           0.1495
                      New                 0.0804
                      XNA                 0.0012
Name: NAME_CLIENT_TYPE, dtype: float64
```

Bivariate Analysis

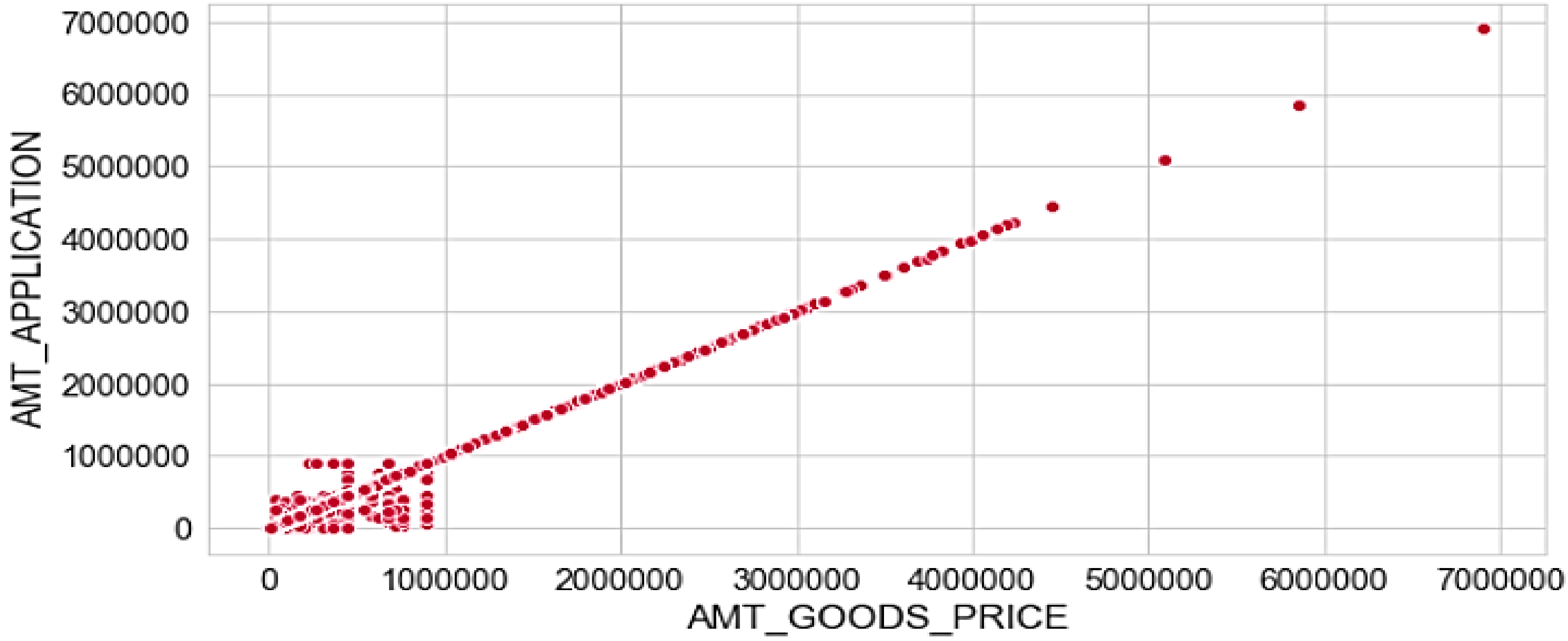
- **Categorical Variable :** NAME_YIELD_GROUP *grouped by* NAME_CONTRACT_STATUS
- **Conclusions :** middle' category client type has the highest number of approved loans; ‘XNA’ type has the highest Canceled and rejected loans.




```
NAME_CONTRACT_STATUS  NAME_YIELD_GROUP
Approved              middle              0.3116
                      high                0.2884
                      low_normal          0.2373
                      XNA                 0.0943
                      low_action          0.0684
Canceled              XNA                 0.9683
                      low_normal          0.0153
                      middle              0.0103
                      high                0.0032
                      low_action          0.0030
Refused               XNA                 0.3015
                      low_normal          0.2427
                      middle              0.2032
                      high                0.1832
                      low_action          0.0695
Unused offer          XNA                 0.9656
                      low_normal          0.0247
                      middle              0.0070
                      high                0.0027
Name: NAME_YIELD_GROUP, dtype: float64
```

Bivariate Analysis

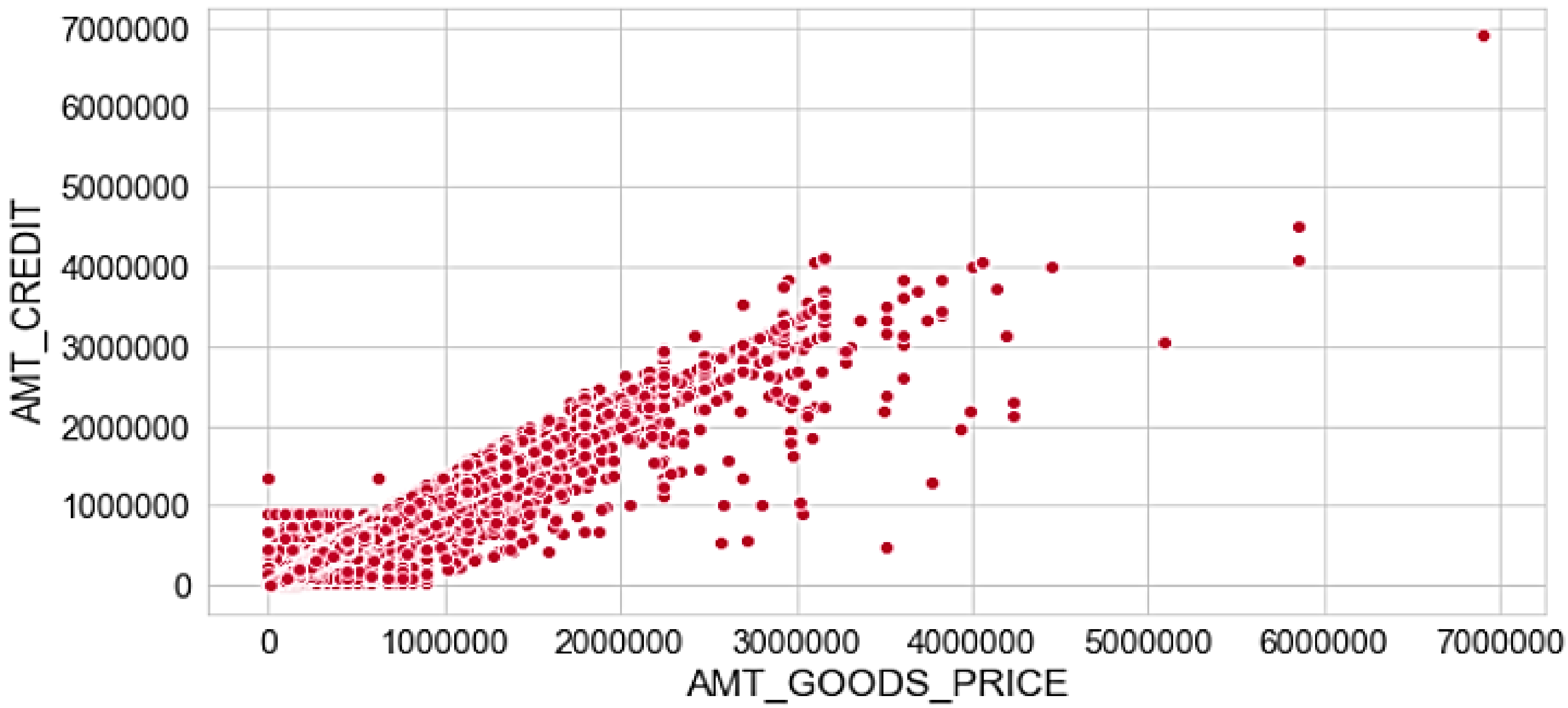
- **Continuous Variable :** AMT_GOODS_PRICE *vs. AMT_APPLICATION*
- **Conclusions :** There is a high correlation between the two variables. This means that, as one increases, the other increases as well.



Correlation Value = 0.9998837157835986

Bivariate Analysis

- **Continuous Variable :** AMT_GOODS_PRICE *vs. AMT_CREDIT*
- **Conclusions :** There is a high correlation between the two variables. This means that, as one increases, the other increases as well.



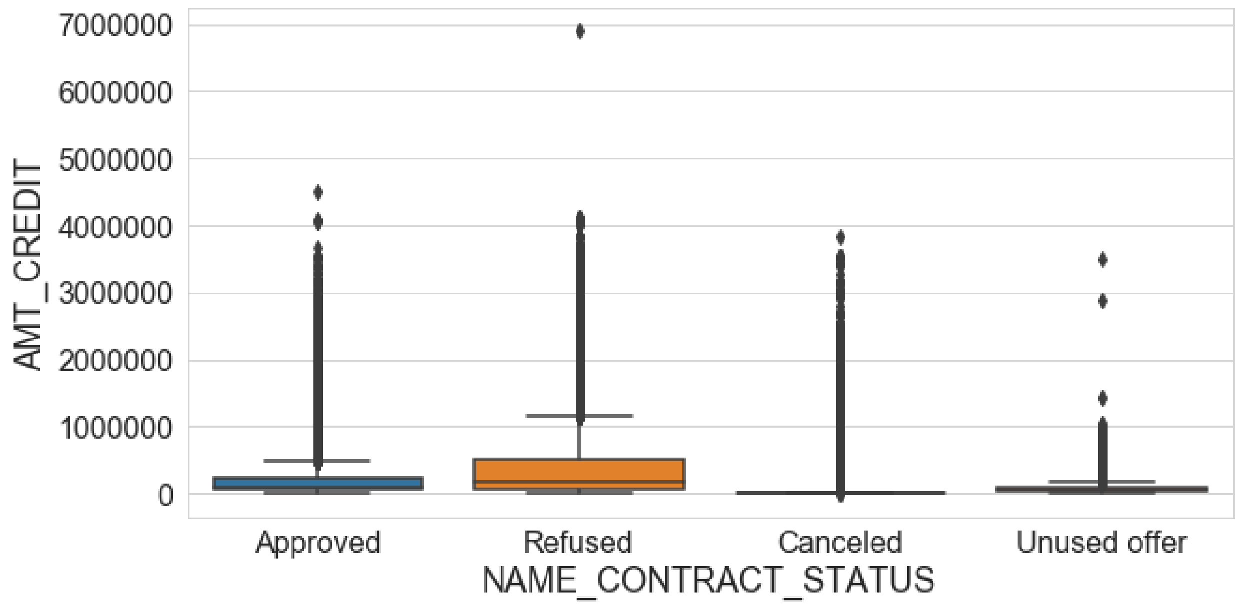

```
Correlation Value =  0.9930870506319731
```

Bivariate Analysis

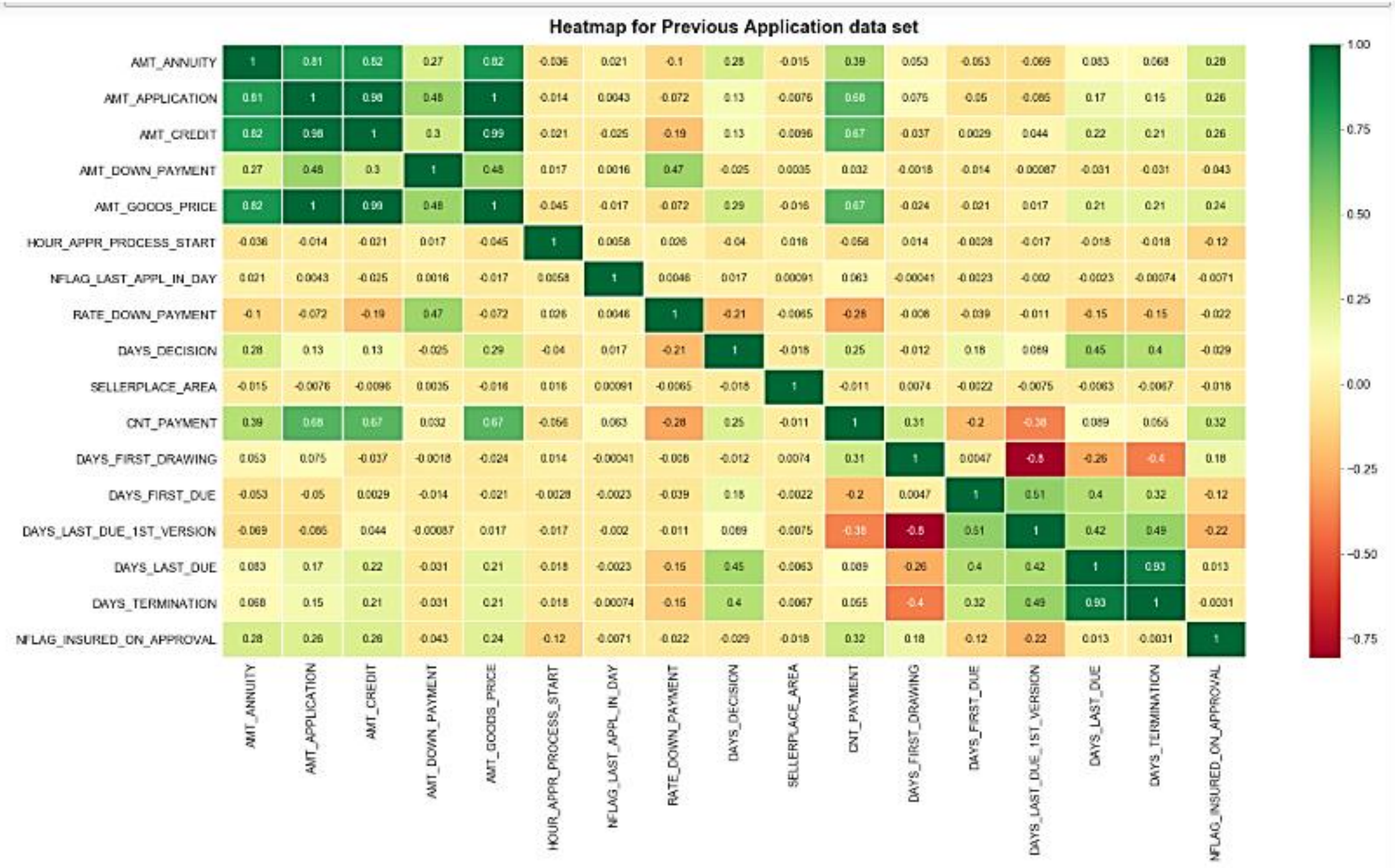
- **Continuous - Categorical Variable:** NAME_CONTRAT_STATUS *vs. AMT_CREDIT*
- **Conclusions :** Even though the count of approved loans is highest, Refused loans has the highest median. This is the result of a value in Rejected category which is around 69L

Statistical Description:

NAME_CONTRACT_STATUS	count	mean	std	min	25%	50%	75%	max
Approved	1036780.0000	202564.1821	275302.6663	0.0000	47970.0000	102208.5000	225000.0000	4509688.5000
Canceled	316319.0000	24187.0571	162451.7509	0.0000	0.0000	0.0000	0.0000	3847104.0000
Refused	290678.0000	371689.8412	468119.1805	0.0000	60138.0000	182956.5000	497520.0000	6905160.0000
Unused offer	26436.0000	69783.9908	64248.0629	0.0000	34378.8750	57960.0000	89955.0000	3511305.0000



Multivariate Analysis



Top 10 correlations:

1. AMT_APPLICATION and AMT_GOODS_PRICE 0.9999
2. AMT_CREDIT and AMT_GOODS_PRICE 0.9931
3. AMT_APPLICATION and AMT_CREDIT 0.9758
4. DAYS_LAST_DUE and DAYS_TERMINATION 0.9280
5. AMT_ANNUITY and AMT_GOODS_PRICE 0.8209
6. AMT_CREDIT and AMT_ANNUITY 0.8164
7. AMT_APPLICATION and AMT_ANNUITY 0.8089
8. CNT_PAYMENT and AMT_APPLICATION 0.6806
9. AMT_CREDIT and CNT_PAYMENT 0.6743
10. CNT_PAYMENT and AMT_GOODS_PRICE 0.6721

Major Insights from the Previous Data Set

- 'Repeater' client type category have higher count of 'Approved' loans whereas 'Refreshed' category have the least.
- 'XAP' and 'XNA' categories have the highest 'Cancelled' loans across many variables.
- 'Middle' category in NAME_YIELD_GROUP have highest approved loans.
- 'Low action' group has the least count of 'Approved' as well as 'Refused' loans
- 'Mobile' category has the highest count for the good category that the client has applied followed by consumer electronics.
- 'HC' code reject reason is the most prominent reason for rejected loans.

Top 5 correlation variables from heat map presented in previous slide (top 10 mentioned in notebook).

- AMT_APPLICATION -- AMT_GOODS_PRICE
- AMT_CREDIT-- AMT_GOODS_PRICE
- AMT_APPLICATION--AMT_CREDIT
- DAYS_LAST_DUE--DAYS_TERMINATION
- AMT_ANNUITY--AMT_GOODS_PRICE