Credit EDA Case Study

TRIPURA RAJAVARAPU

SWATHI KANTIPUDI

Problem Statement

Business Understanding:

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

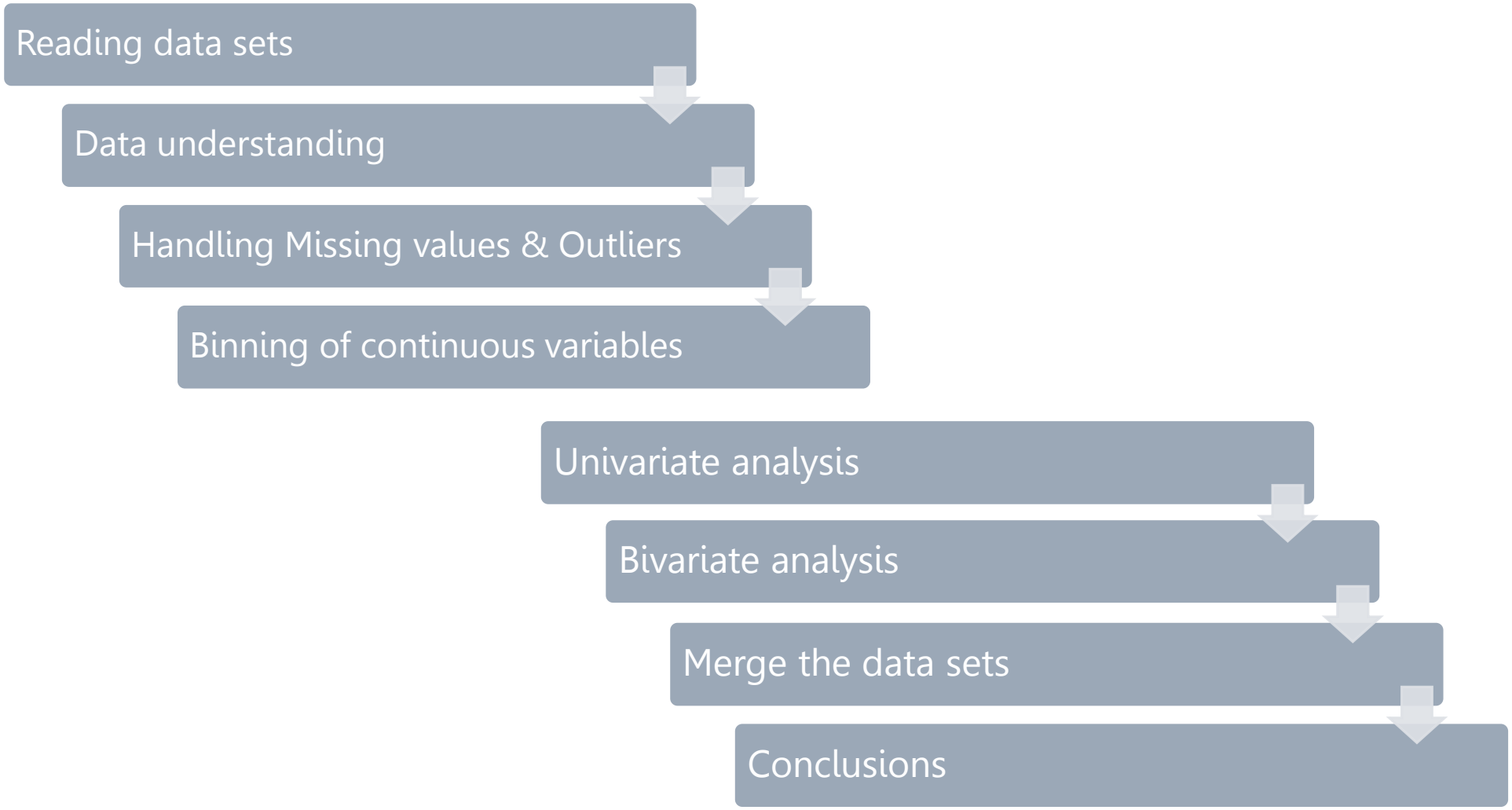
Because of that, some consumers use it as their advantage by becoming a defaulter.

By understanding this problem we will now analyses the datasets and find the necessary observations

Agenda:

- To analyze the patterns present in the data that will ensure that the applicants capable of repaying the loan are not rejected.
- To understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

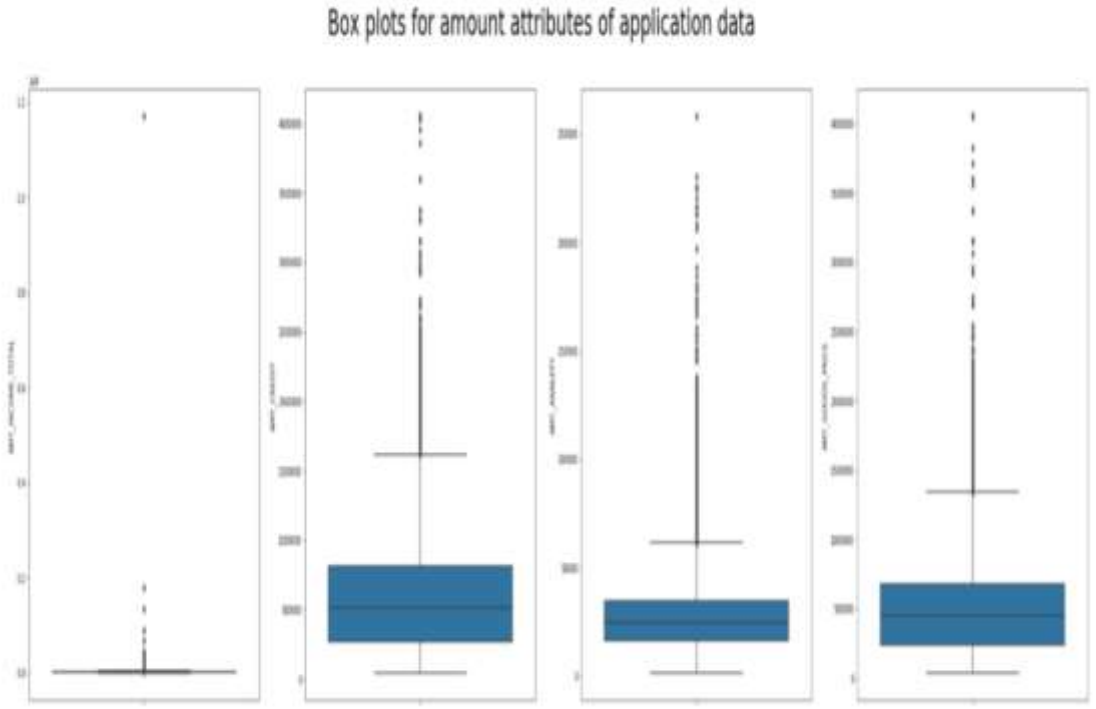
Approach for Analysis



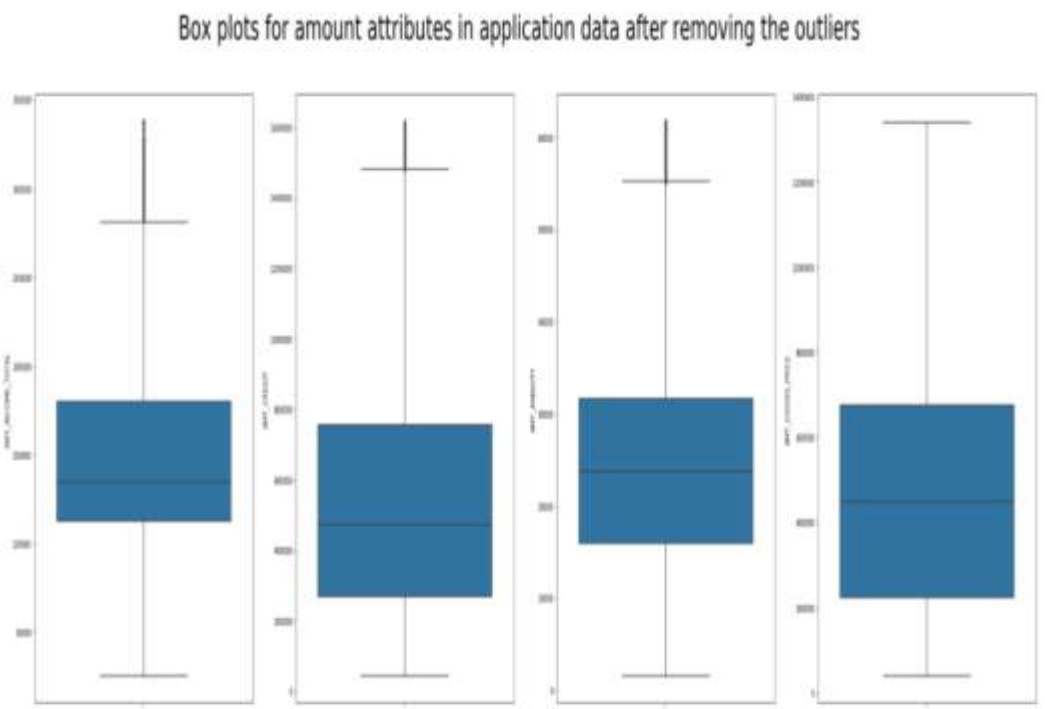
Analysis on Application Data

Outliers Handling(Interquartile range method)

BEFORE REMOVING OUTLIERS



AFTER REMOVING OUTLIERS



Outlier Treatment

Interquartile range method

In the before slide , on the left, we can see that they are outliers present in the data.

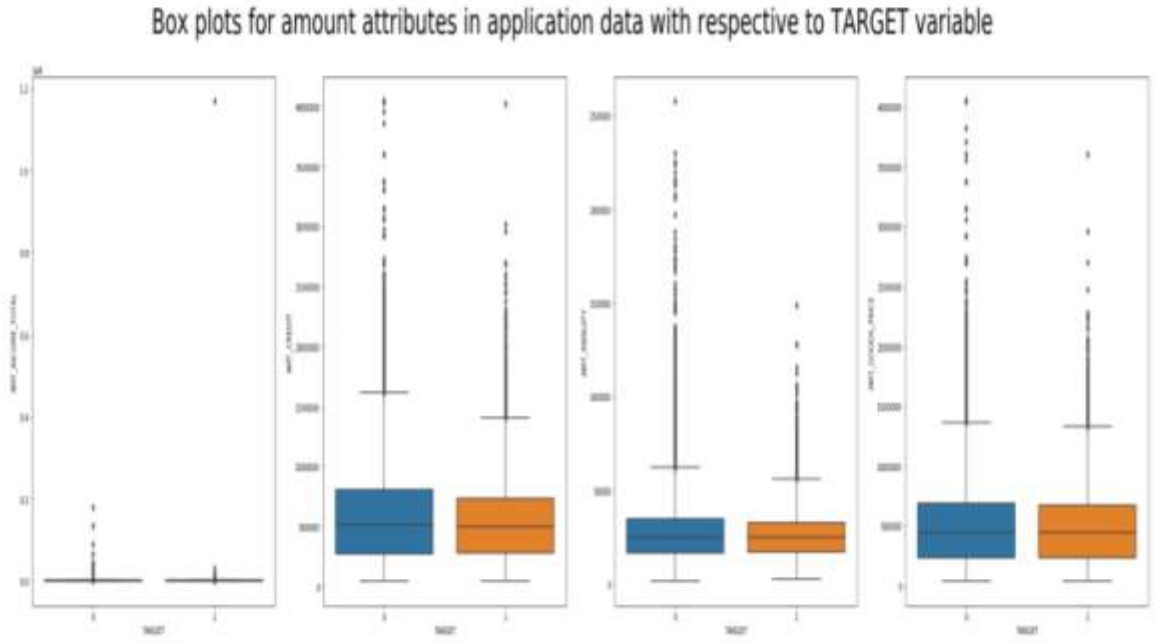
So, in order to overcome this problem, we calculated q1(lower quantile) and q2(upper quantile) for 25th and 75th percentile and IQR i.e., interquartile range

Then, all the rows having value greater than l=(q1-1.5*IQR) and less than h=(q1+1.5*IQR) are considered as the outliers and are dropped from the data frame.

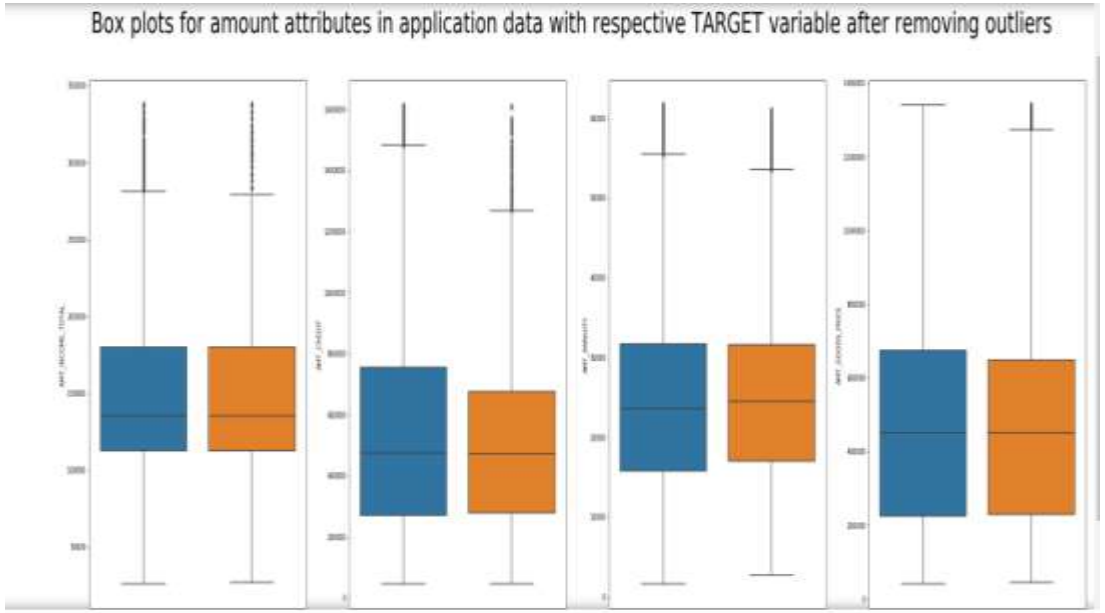
We can see the plot without outliers on the right side.

Outliers Handling w.r.t Target Variable

BEFORE REMOVING OUTLIERS



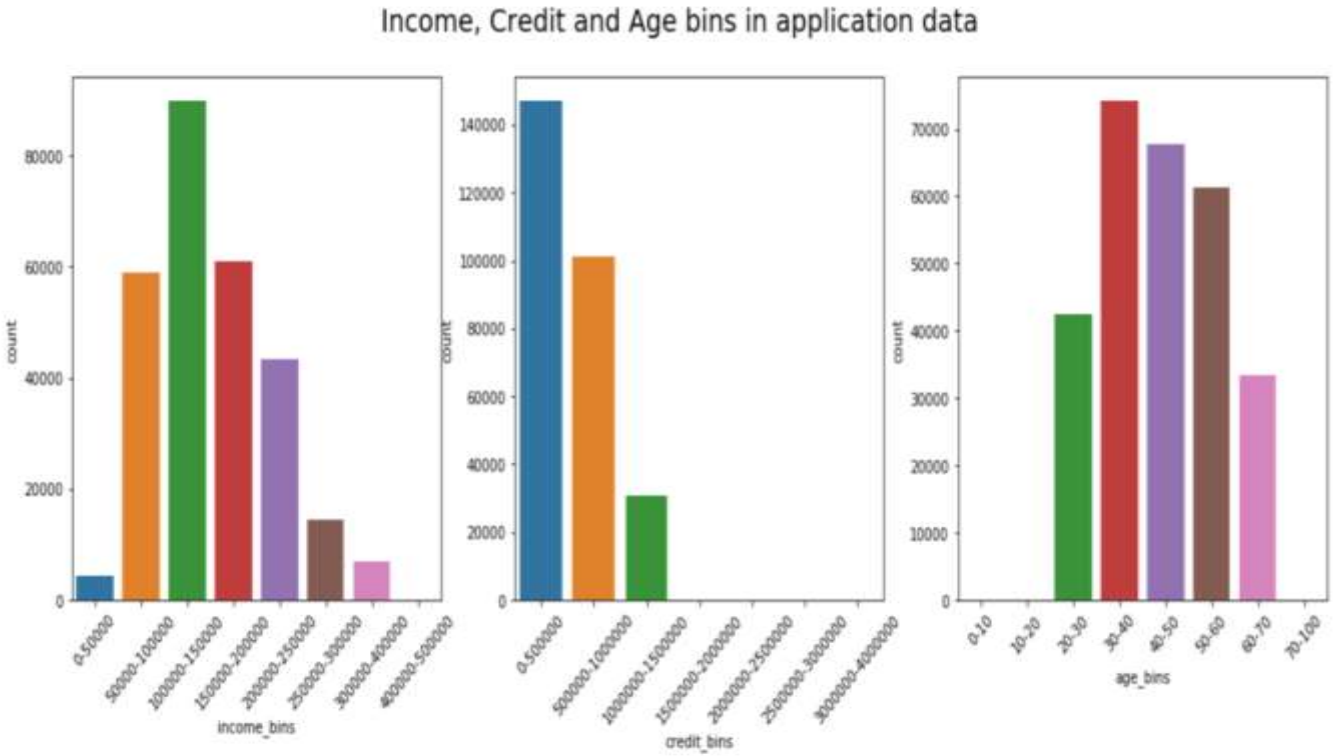
AFTER REMOVING OUTLIERS



Binning of Continuous variables

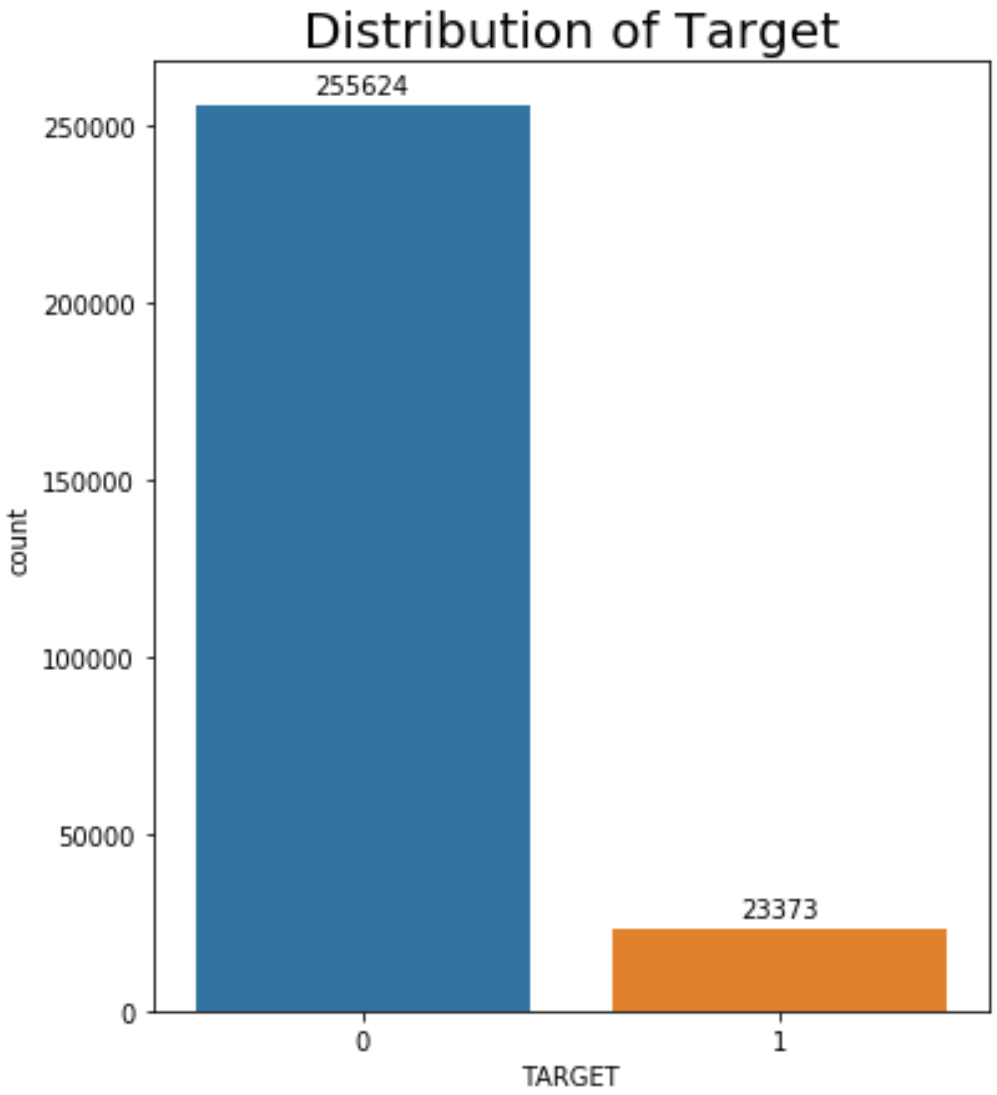
Below are the insights from the binning of the continuous variables:

- ✓ Customers with income between 100000 - 150000 are more than any other income group and least number of customers have the income ranging from 0-50000
- ✓ People in the age group of 30 - 40 have applied for loan more than any other age group.
- ✓ Most of the customers have credit amount in the range of 0 - 500000



Data Imbalance in Target Variable

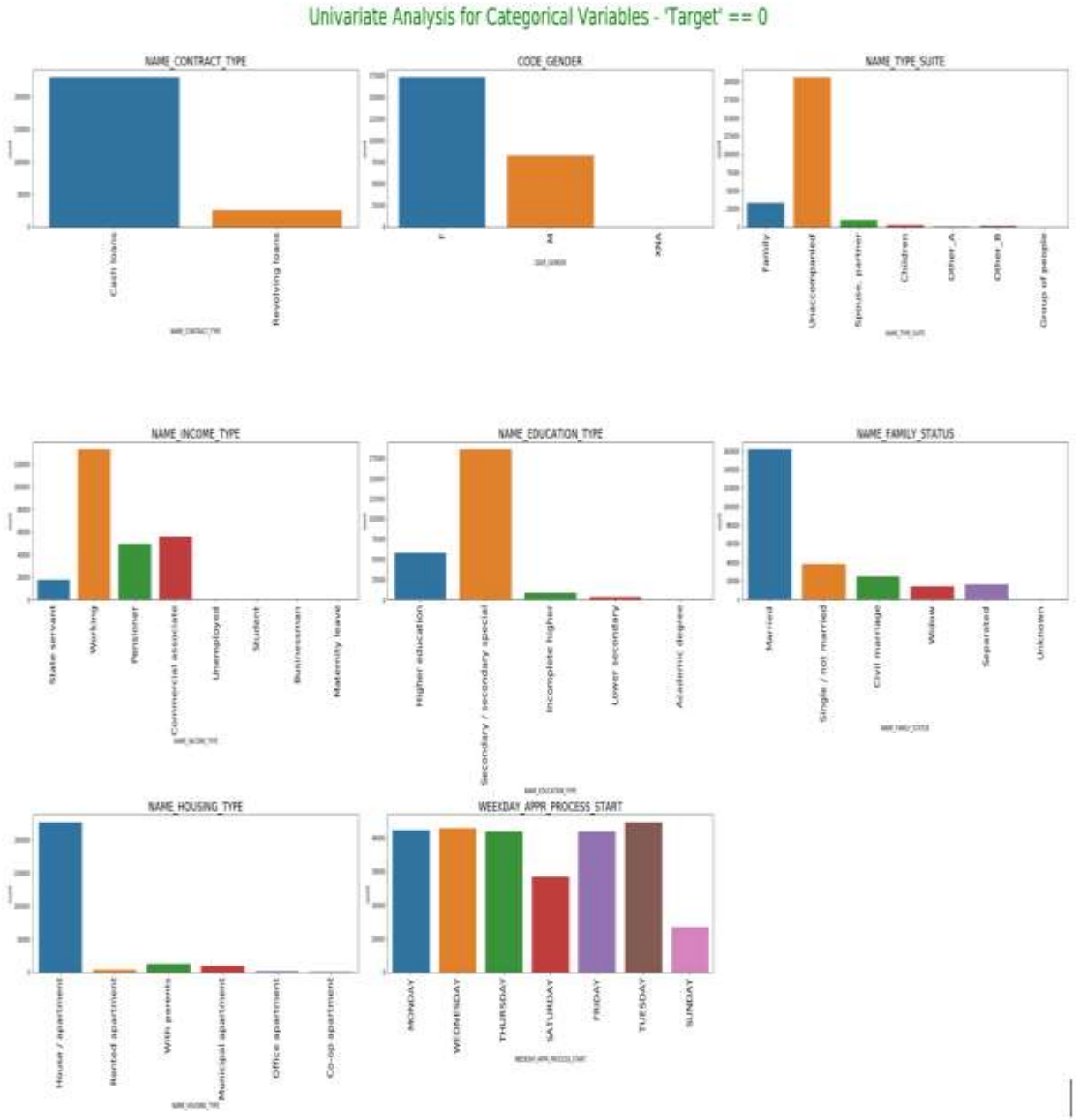
- ✓ There are 255624 rows with Target value '0' and 23373 with Target value '1' i.e. 91.62 % of the total rows determine the payment difficulties of the customers with target 0 and 8.38 % of that for Target 1.
- ✓ There is a huge data imbalance in the 'TARGET' variable between the values '0' and '1'



Univariate Analysis categorical variables Target 0

The plot depicts how each category is distributed among its subcategories for Target 0. Few observations depicted from the above graphs are mentioned below.

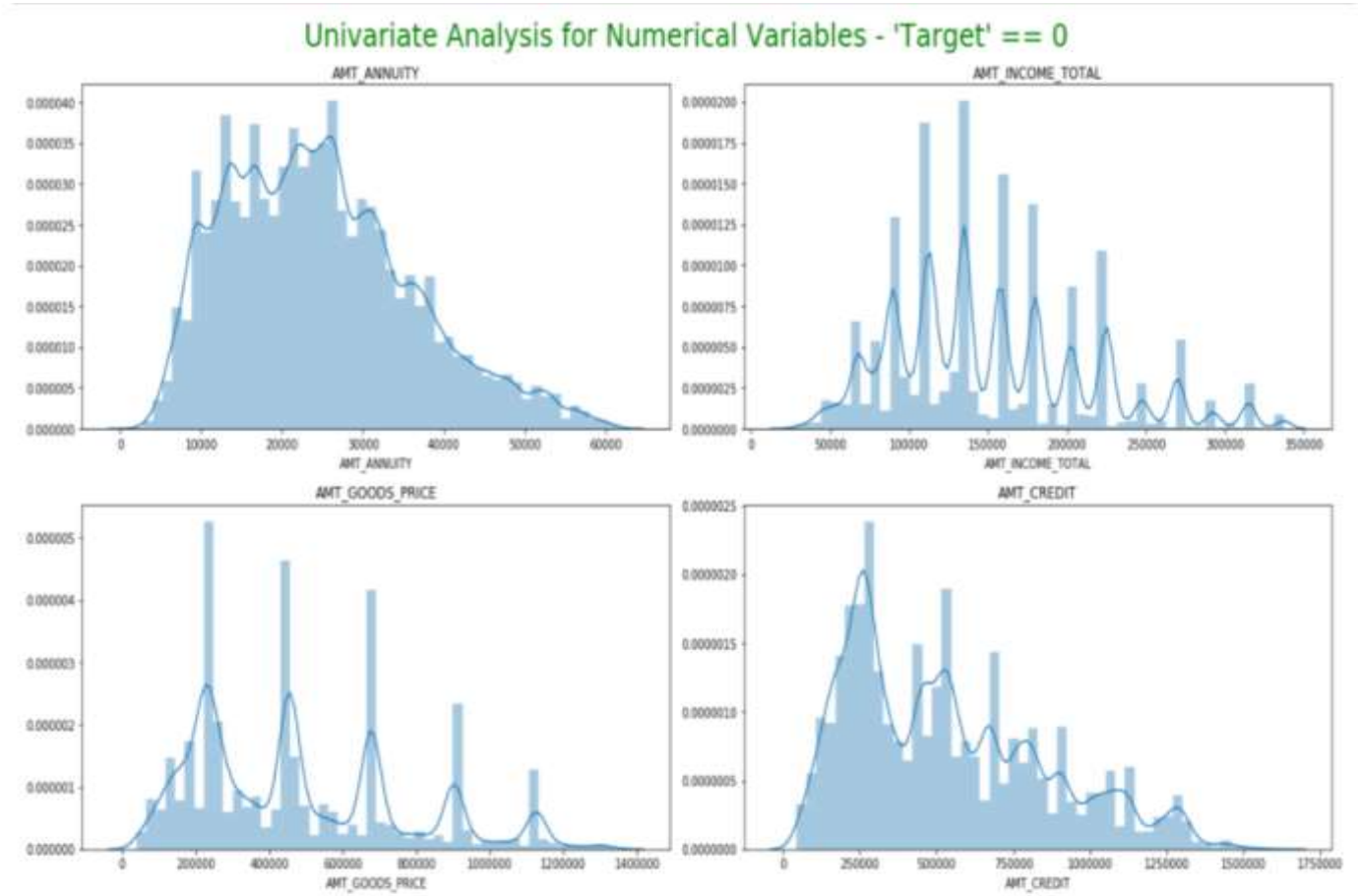
- ✓ Customers have applied for more Cash loans when compared to revolving loans
- ✓ Female customers are more in number than that of Male customers.
- ✓ Most of the customers are unaccompanied while applying for the loan.
- ✓ Based on the Income type, Working professionals are more in number when compared to the other subcategories.
- ✓ Most of the customers fall under the category of secondary education
- ✓ Most of the clients are married.
- ✓ Many of them live in their houses/apartments.
- ✓ Most of the clients applied Loan on Tuesday.



Univariate Analysis numerical variables Target 0

Analysis

- ✓ High values of 'AMT_ANNUITY' are concentrated between 10000 – 30000.
- ✓ High values of 'AMT_INCOME_TOTAL' are concentrated between 120000 - 150000 approx.
- ✓ Highest value of 'AMT_GOODS_PRICE' is found between 200000-300000
- ✓ Decreasing trend is observed in 'AMT_CREDIT' attribute.



Correlation w.r.t Target 0

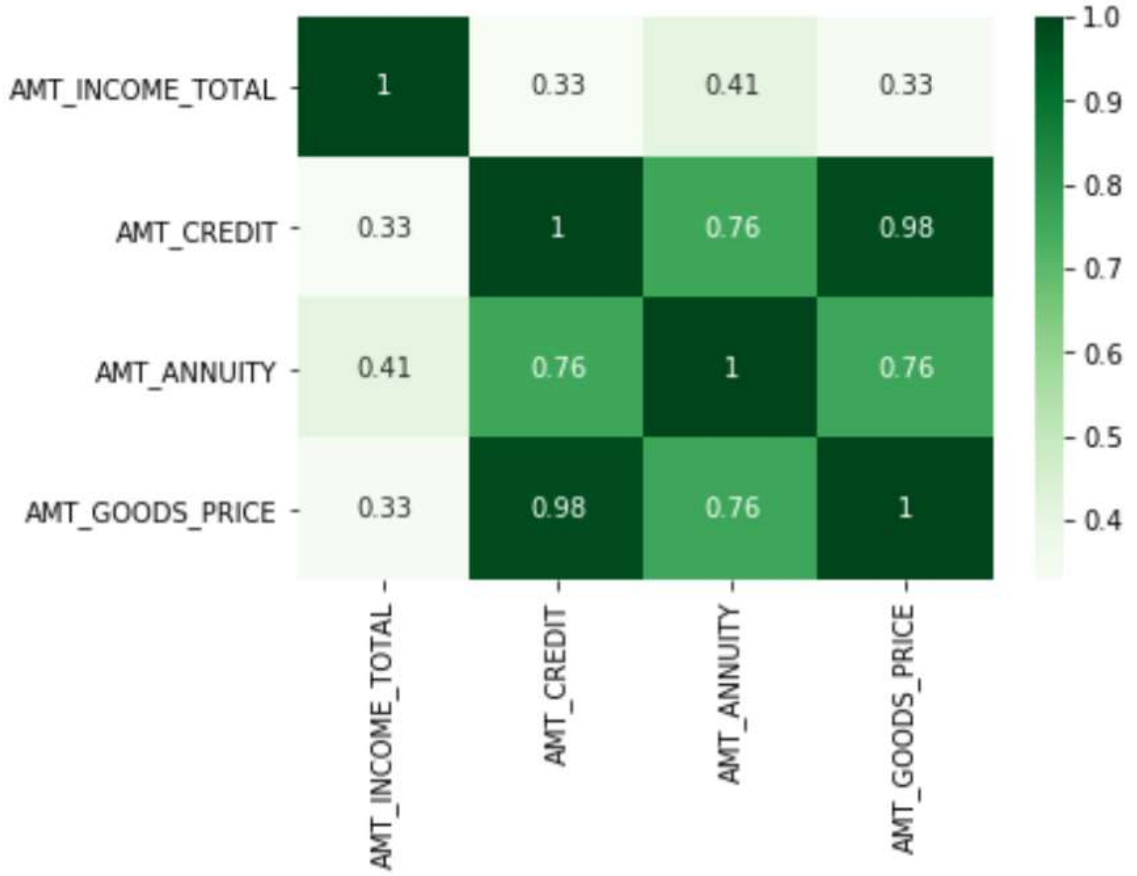
Analysis

The heatmap clearly depicts that the correlation between AMT_GOODS_PRICE and AMT_CREDIT is highest which is 0.98 which means that both the attributes are following a very good linear relationship.

Similarly, from the heatmap, the correlated coefficients for other feature combinations are listed below

- ✓ AMT_CREDIT & AMT_GOODS_PRICE - 0.98
- ✓ AMT_CREDIT & AMT_ANNUITY - 0.76
- ✓ AMT_ANNUITY & AMT_GOODS_PRICE - 0.76
- ✓ AMT_ANNUITY & AMT_INCOME_TOTAL - 0.41
- ✓ AMT_INCOME_TOTAL & AMT_GOODS_PRICE - 0.33
- ✓ AMT_INCOME_TOTAL & AMT_CREDIT - 0.33

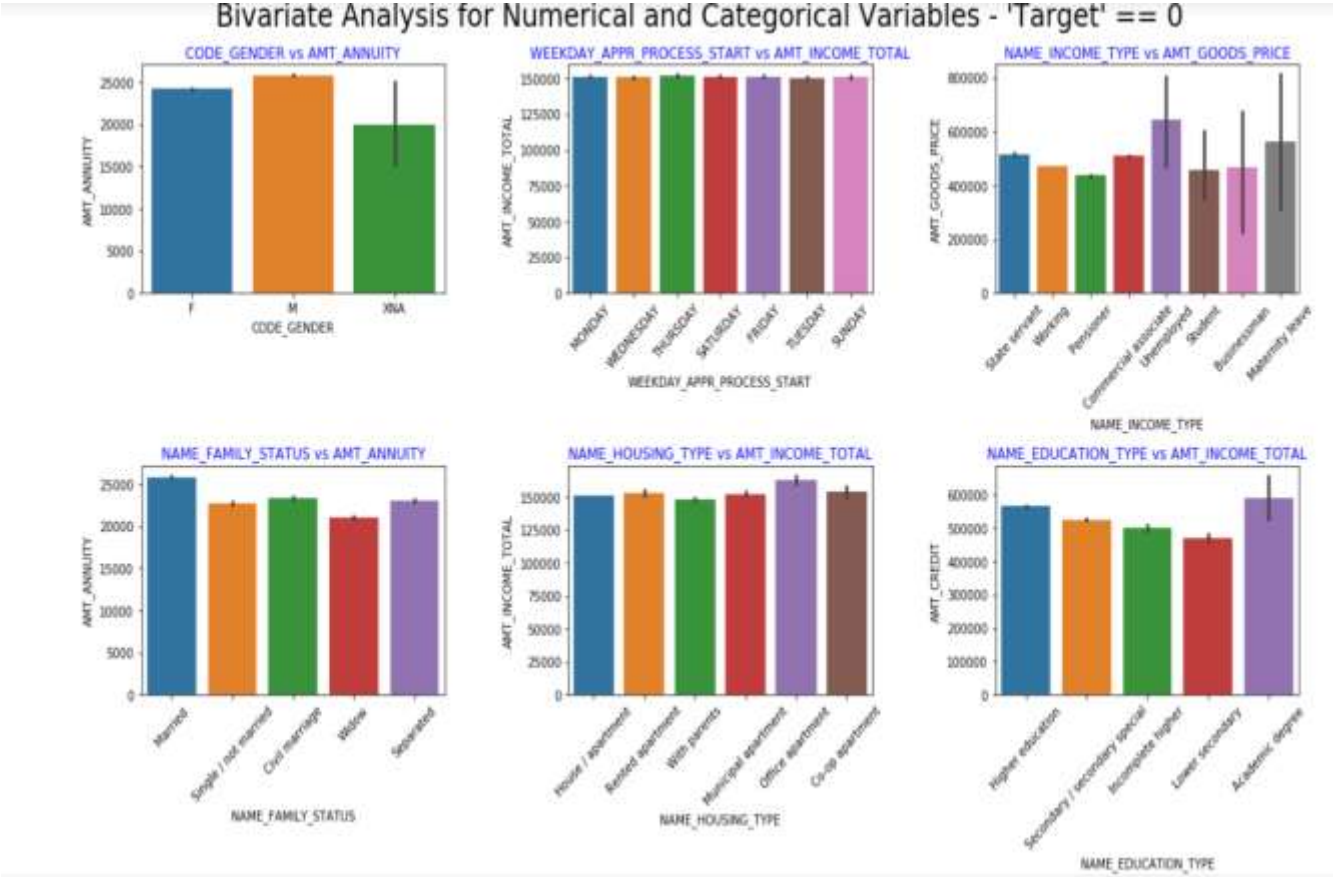
All of these correlation coefficients lies between 0-1 and hence the correlation is positive.



Bivariate Analysis for Numerical & Categorical variables Target-0

The above plot shows how the categorical features are distributed with the numerical variables. The below analysis are limited only for target variable 0

- ✓ The average loan annuity is higher for males when compared to other genders.
- ✓ The average loan annuity is higher for Married people when compared to people of other family status.
- ✓ The average price of the goods for which the loan is given is highest for unemployed people.
- ✓ The income of the clients with a housing situation of office apartment is more when compared to others.
- ✓ Clients with academic degree are having more income amount compared to others

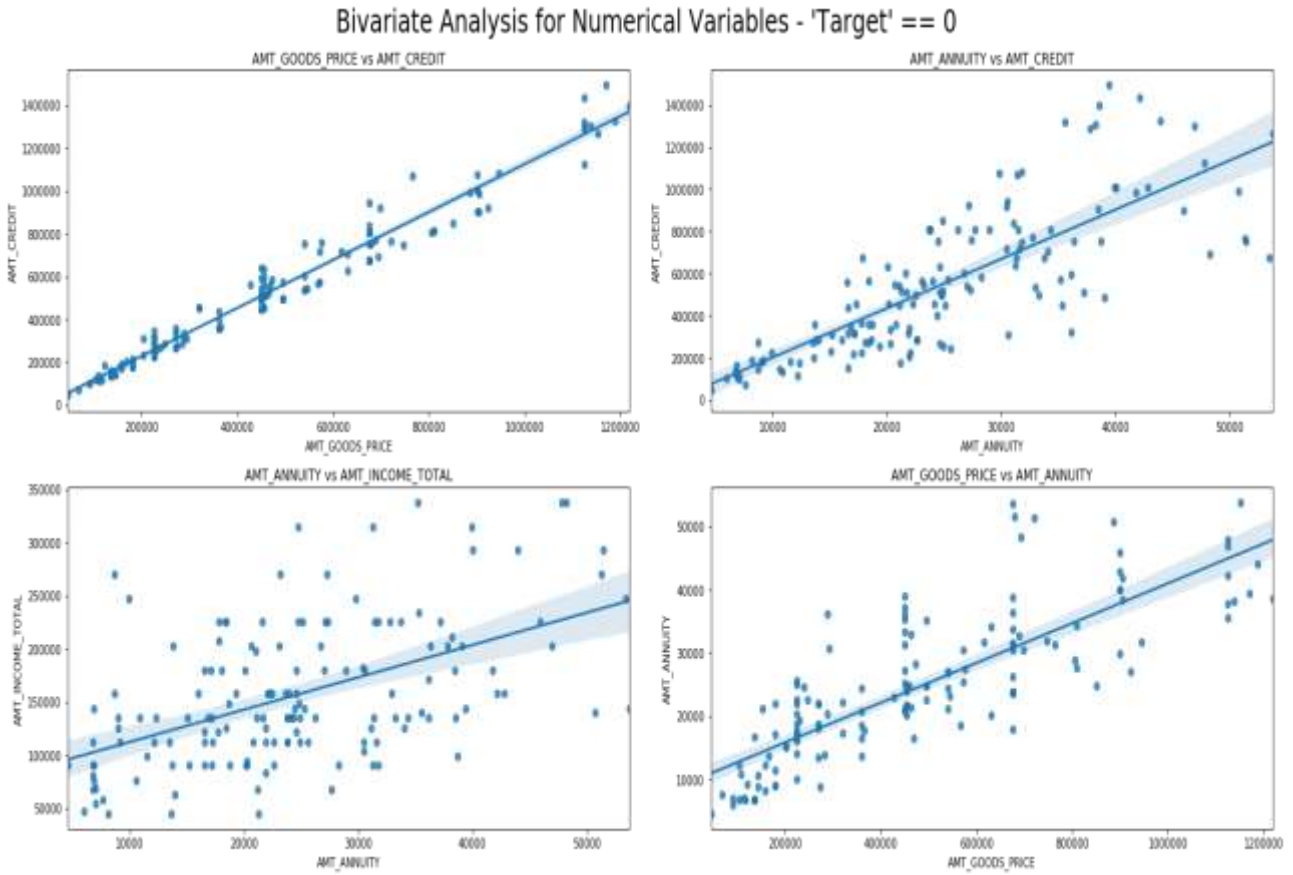


Bivariate Analysis for Numerical variables Target-0

Depending on the correlation coefficient, the graphs are distributed as above.

The best fit line describes the relationship between the attributes more precisely.

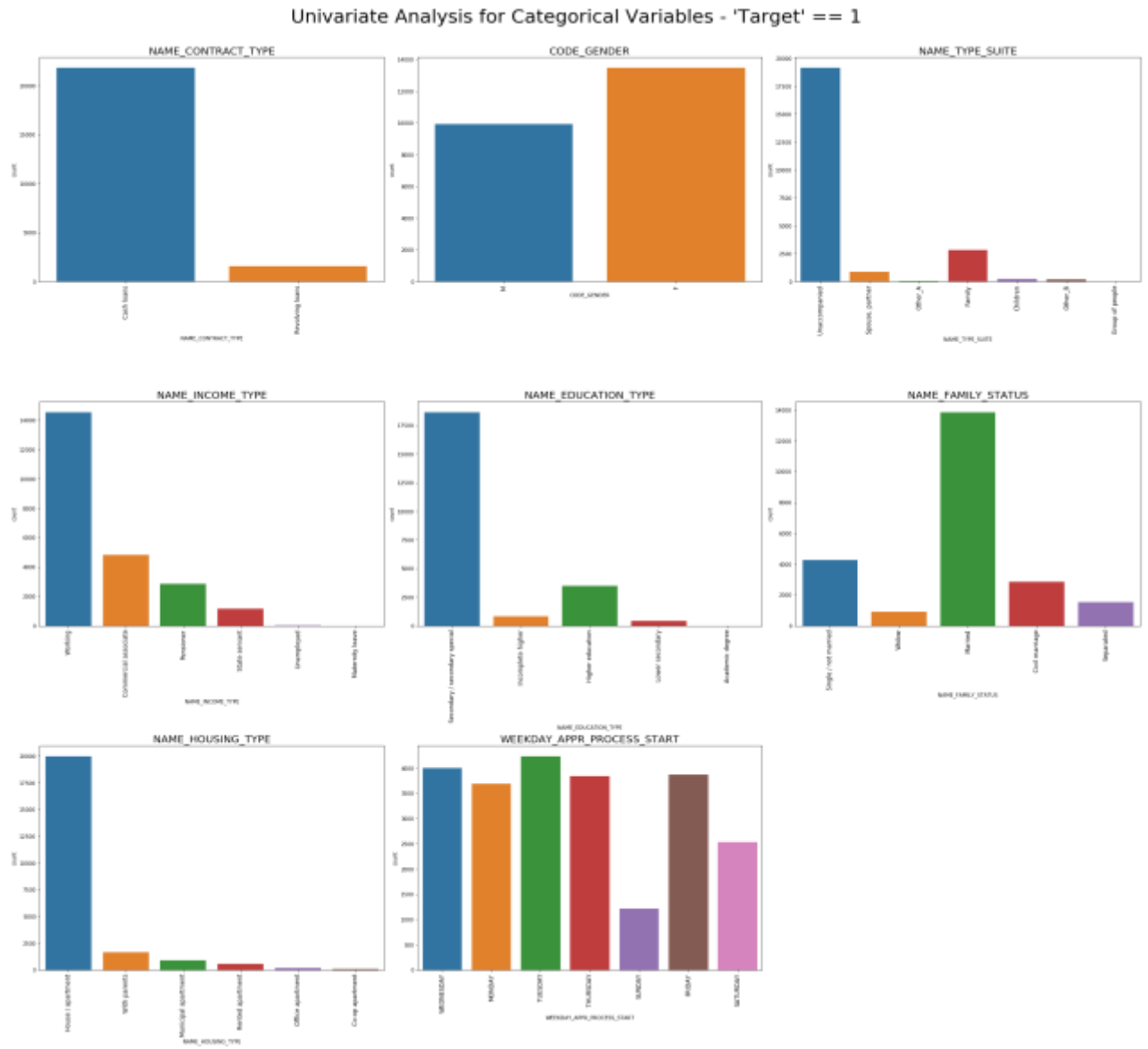
As the correlation between goods price and Credit amount is 0.98 which is very close to 1, here in the graph we can see the linear relationship between the two attributes very clearly compared to other cases.



Univariate Analysis for Categorical Variables Target -1

Below inferences are made from the graphs plotted above which are limited to the customers with payment difficulties corresponding to Target 1 only:

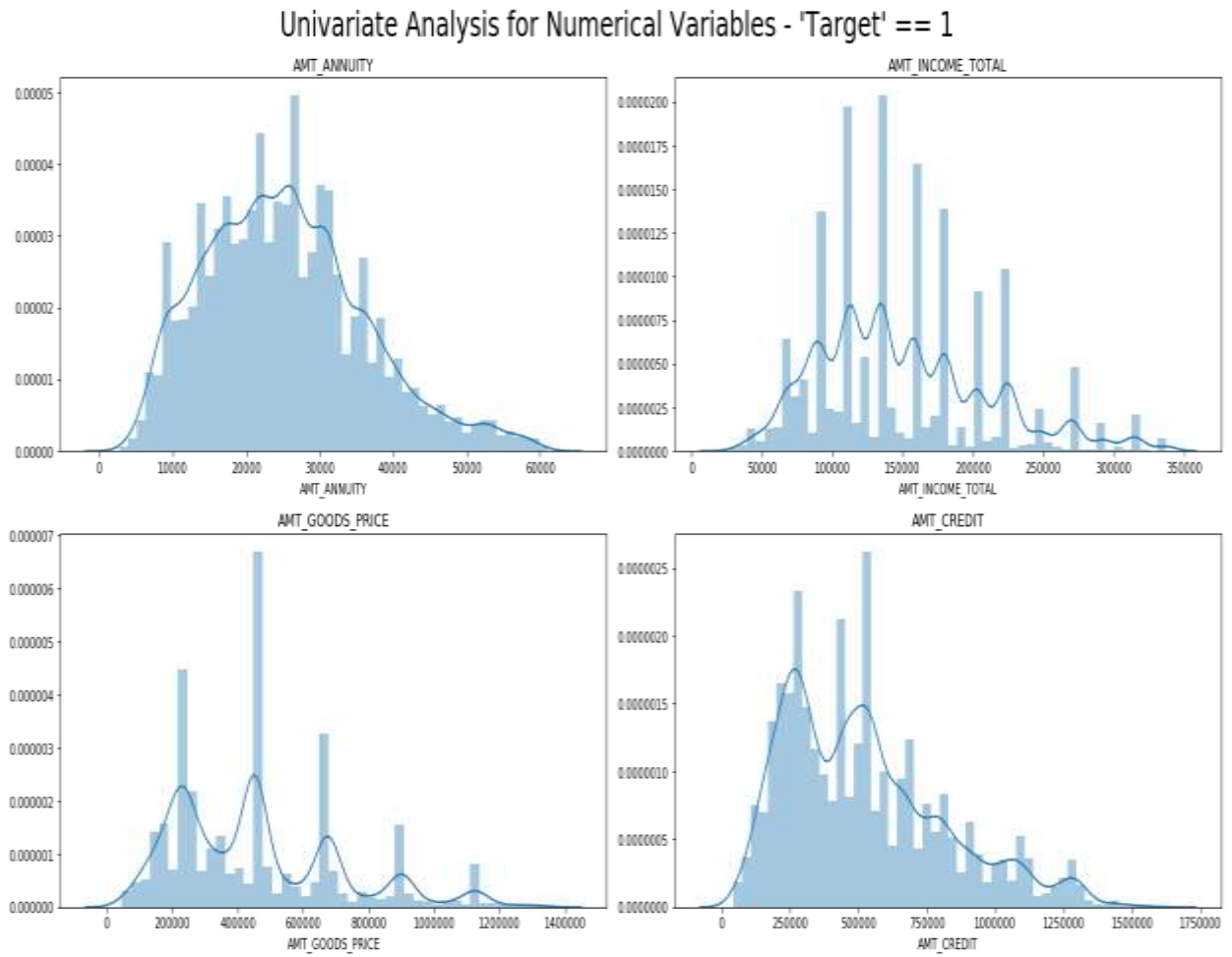
- ✓ Customers are applied for more Cash loans than Revolving loans.
- ✓ Female customers are more in number compared to Male.
- ✓ Most of the customers are unaccompanied while applying for the loan.
- ✓ Most of the people who are having payment difficulties are working professionals.
- ✓ Married couples are facing more payment difficulties when compared to people of other family statuses.
- ✓ Here, most of the customers fall under the category of secondary education
- ✓ The housing situation of the people with payment difficulties is House / apartment in majority
- ✓ Many of them applied loan on Tuesday compared to any other day.



Univariate Analysis for Numerical Variables Target -1

Analysis

- ✓ High values of 'AMT_ANNUITY' are concentrated between 20000 - 30000
- ✓ High values of 'AMT_INCOME_TOTAL' are approximately concentrated between 120000 - 150000
- ✓ Decreasing trend is seen in 'AMT_CREDIT'.

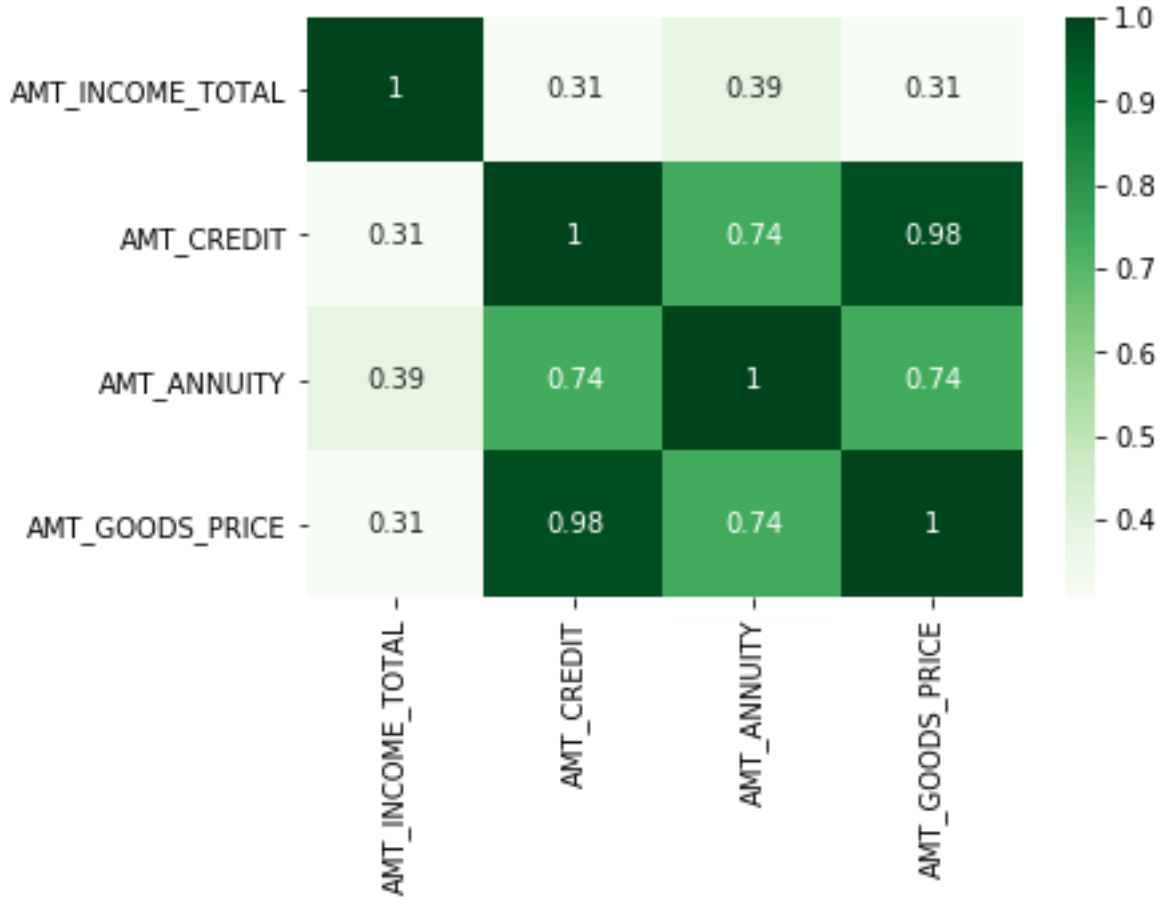


Correlation w.r.t Target -1

From the heatmap, we can determine that the correlation between AMT_GOODS_PRICE and AMT_CREDIT is the highest which is 0.98 which depicts that both of these features are having a very good linear relationship.

Similarly, from the heatmap, below are the features which have correlated nicely.

- ✓ AMT_CREDIT and AMT_GOODS_PRICE - 0.98
- ✓ AMT_CREDIT and AMT_ANNUITY - 0.74
- ✓ AMT_ANNUITY and AMT_GOODS_PRICE - 0.74
- ✓ AMT_ANNUITY and AMT_INCOME_TOTAL - 0.39
- ✓ AMT_INCOME_TOTAL and AMT_GOODS_PRICE - 0.31
- ✓ AMT_INCOME_TOTAL and AMT_CREDIT - 0.31



Top 10 Correlations

TARGET 0(OTHER CASES)

	Column1	Column2	Correlation
1114	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
214	AMT_GOODS_PRICE	AMT_CREDIT	0.98
719	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.95
632	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
863	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86
1150	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
971	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.83
215	AMT_GOODS_PRICE	AMT_ANNUITY	0.76
179	AMT_ANNUITY	AMT_CREDIT	0.76
463	FLAG_EMP_PHONE	DAYS_BIRTH	0.63

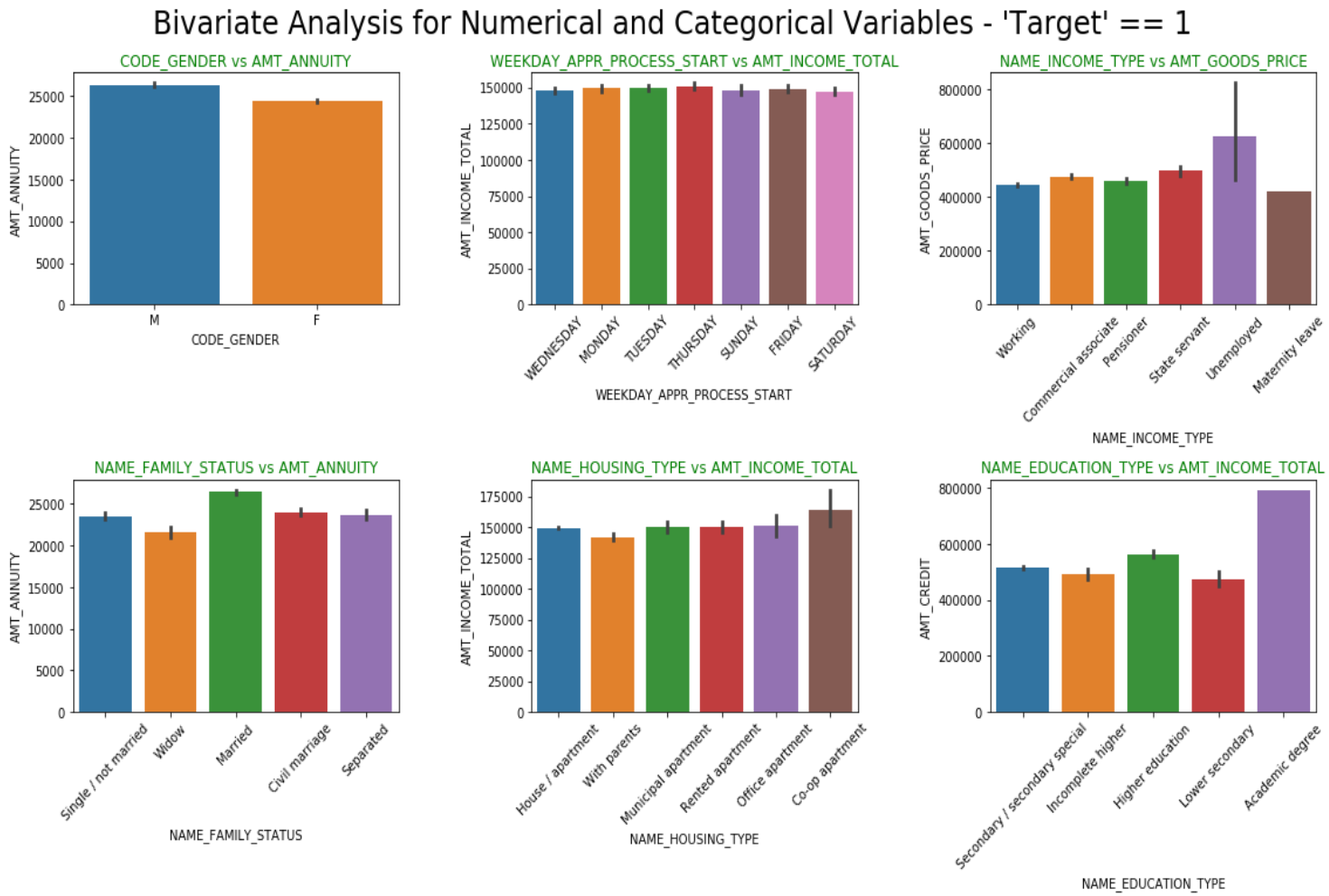
TARGET 1(CLIENT WITH DIFFICULTIES)

	Column1	Column2	Correlation
1114	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
214	AMT_GOODS_PRICE	AMT_CREDIT	0.98
719	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.96
632	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
1150	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
863	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.85
971	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.78
179	AMT_ANNUITY	AMT_CREDIT	0.74
215	AMT_GOODS_PRICE	AMT_ANNUITY	0.74
463	FLAG_EMP_PHONE	DAYS_BIRTH	0.59

Bivariate Analysis for Numerical & Categorical variables Target-1

The plot shows how the categorical features are distributed with the numerical variables. The below points are limited only for target value 1

- ✓ The average loan annuity is higher for males when compared to other genders.
- ✓ The average price of the goods for which the loan is given is highest for unemployed people.
- ✓ The average loan annuity is higher for Married people when compared to others.
- ✓ People with housing type Co-op apartment has the highest income compared to other housing types
- ✓ Income amount is more in case of people with Academic degree.

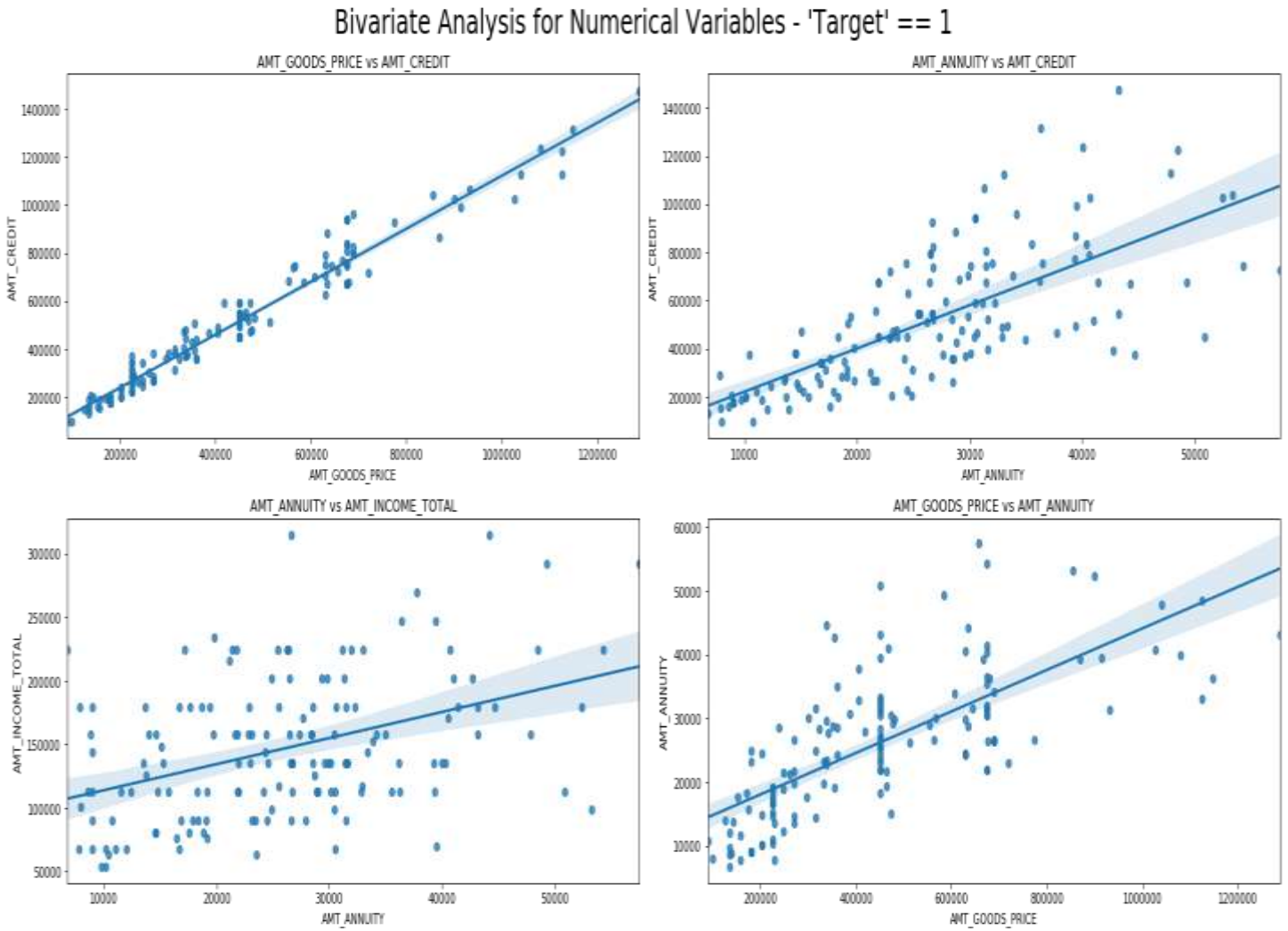


Bivariate Analysis for Numerical variables Target-1

Depending on the correlation coefficient, the graphs are distributed as shown in the figure.

The best fit line describes the relationship between the attributes more precisely.

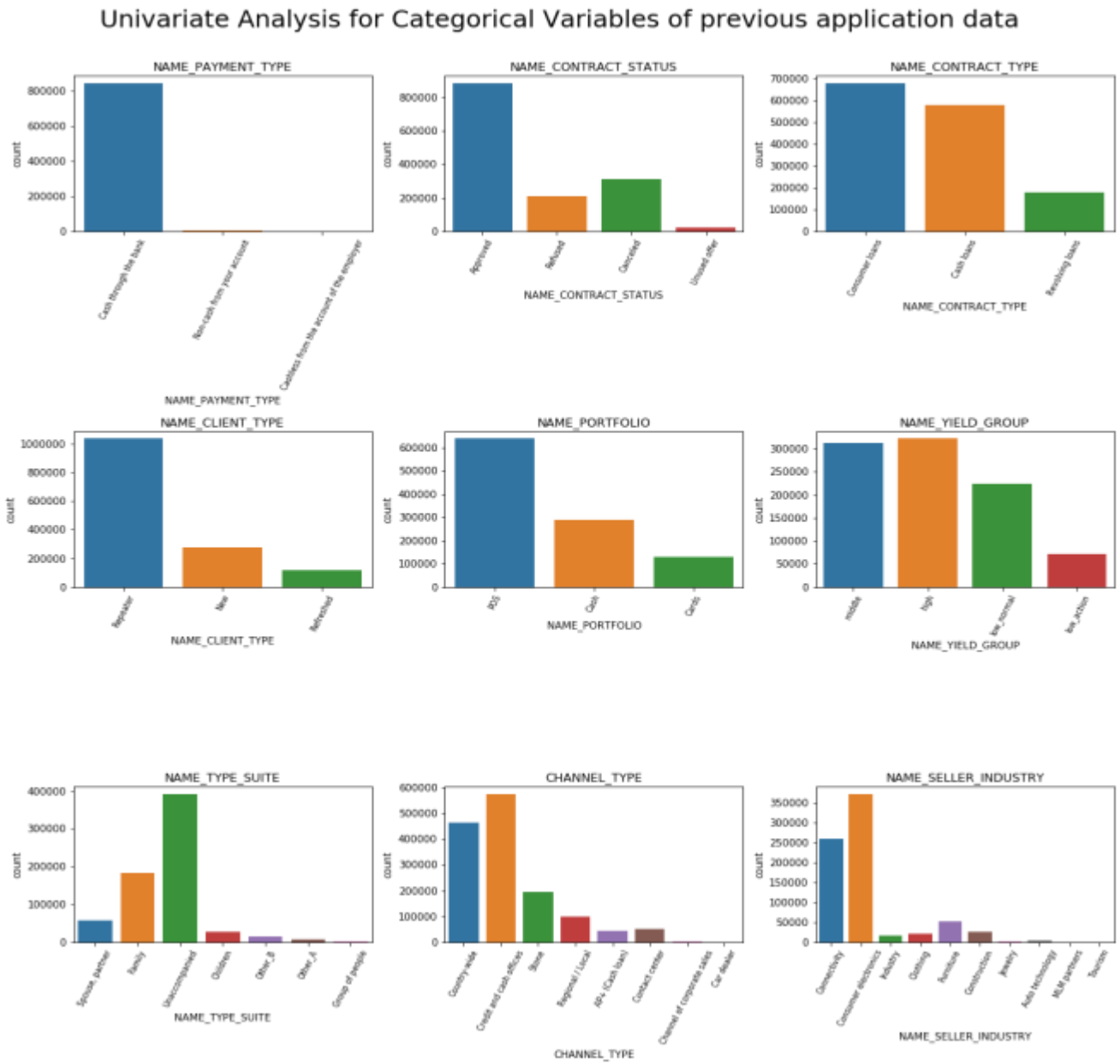
As the correlation between goods price and credit amount is 0.98 which is very close to 1, here in the above graph we can see the linear relationship between the two attributes very clearly compared to other cases.



Analysis on Previous Application Data

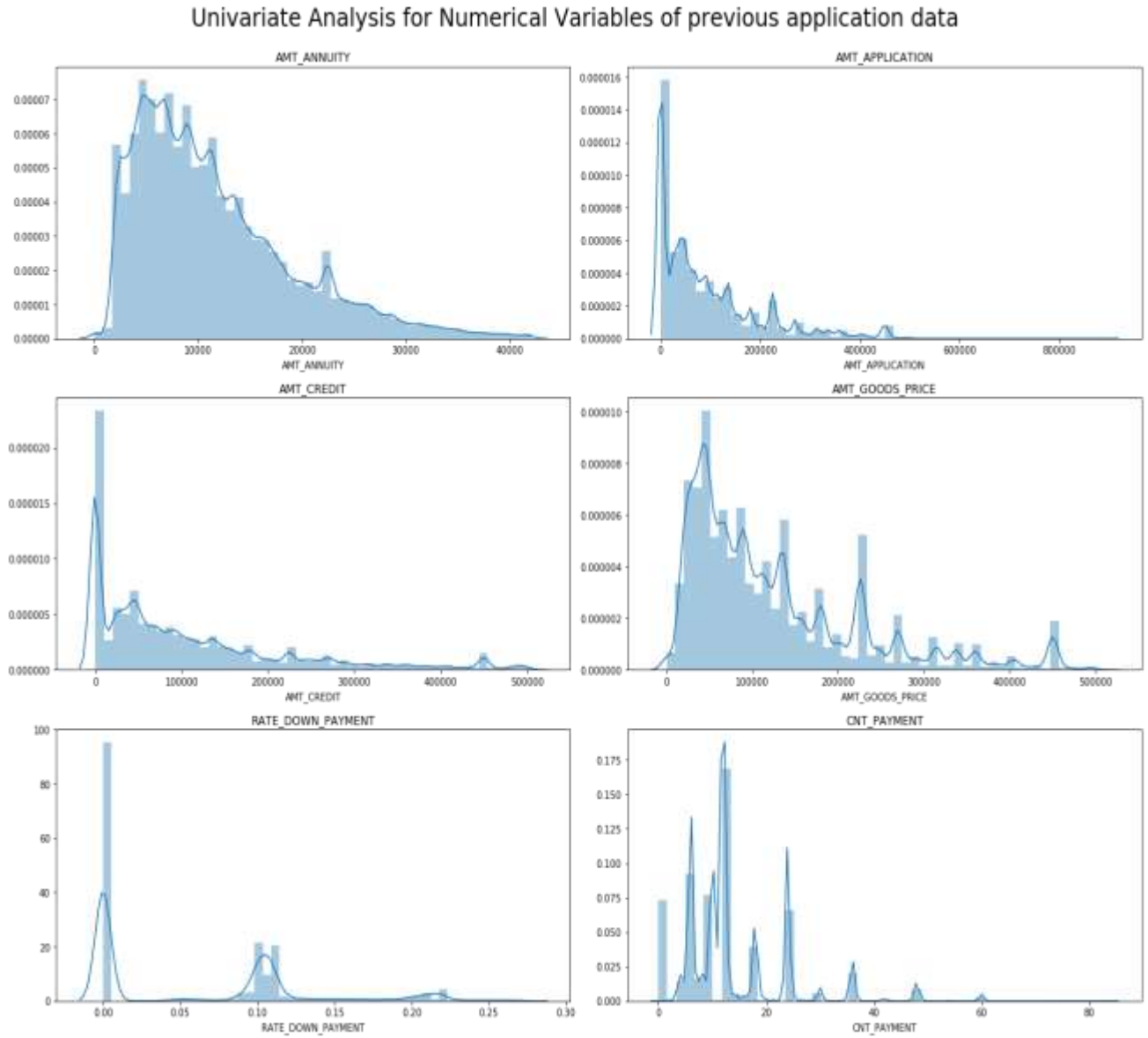
Univariate Analysis for Categorical Variables

- ✓ We see that maximum of the loans were sanctioned for cash through the bank payment type.
- ✓ Majority of the loans are approved with few rejected, unused and unused.
- ✓ High number of Consumer loans were sanctioned by the bank than cash and revolving loans.
- ✓ There were high number of repeater customers in the bank loan history.
- ✓ The number of loans approved for POS is high when compared to cash and cards.
  High number of loans were approved for high yield group followed by middle and low_normal.
- ✓ Majority of the loans are approved for unaccompanied in NAME_TYPE_SUITE.
- ✓ More loans were approved for credit and cash followed by country-wide in CHANNEL_TYPE.
- ✓ We see that max loans were approved for consumer electronics type followed by connectivity in NAME_SELLER_INDUSTRY.



Univariate Analysis for Numerical Variables

- ✓ High values of 'AMT_ANNUITY' and 'AMT_GOODS_PRICE' are concentrated below 10000
- ✓ 'AMT_APPLICATION', 'AMT_CREDIT' and 'RATE_DOWN_PAYMENT' have higher amounts of '0' value
- ✓ High Values of 'CNT_PAYMENT' is concentrated in the range of 10-20

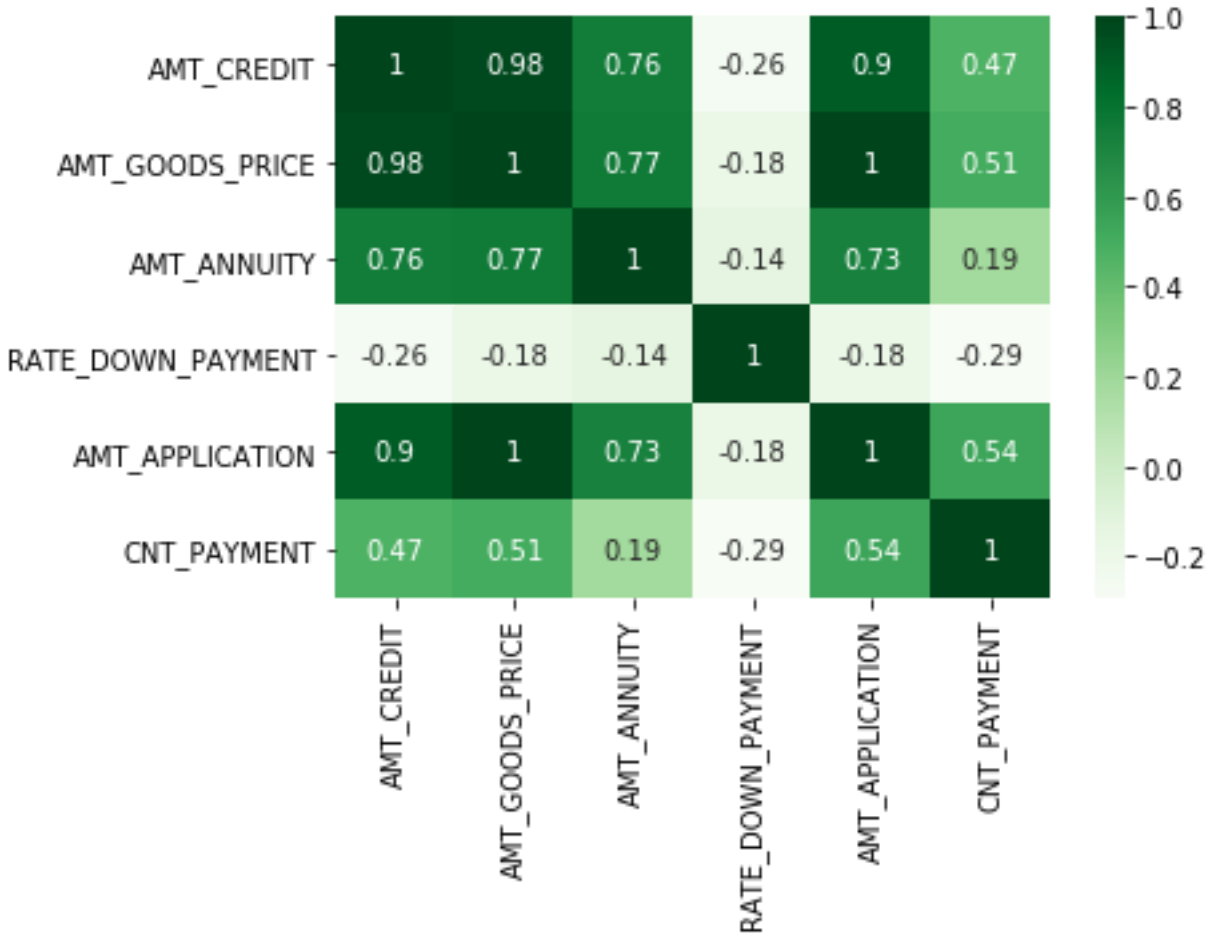


Correlation matrix for Numerical Variables

From the heatmap, the correlation between AMT_GOODS_PRICE and AMT_CREDIT is 0.98 which shows a very good linear relationship.

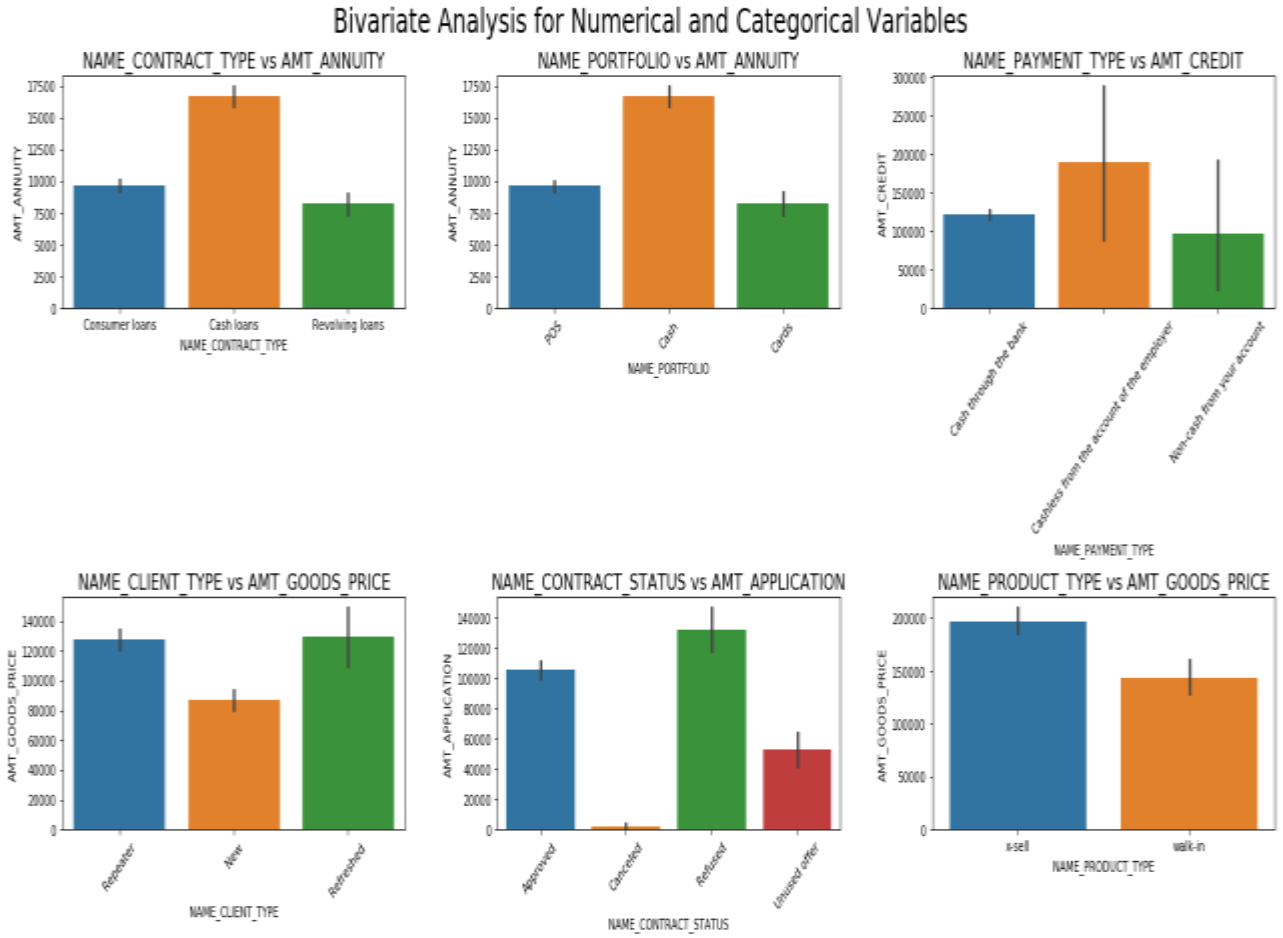
Similarly, below are the observations from the graph

- ✓ AMT_GOODS_PRICE and AMT_CREDIT - 0.98
- ✓ AMT_CREDIT and AMT_APPLICATION - 0.9
- ✓ AMT_GOODS_PRICE and AMT_ANNUITY - 0.77
- ✓ AMT_CREDIT and AMT_ANNUITY - 0.76
- ✓ AMT_ANNUITY and AMT_APPLICATION - 0.73
- ✓ CNT_PAYMENT and AMT_APPLICATION - 0.54
- ✓ RATE_DOWN_PAYMENT is having negative correlation with other variables.



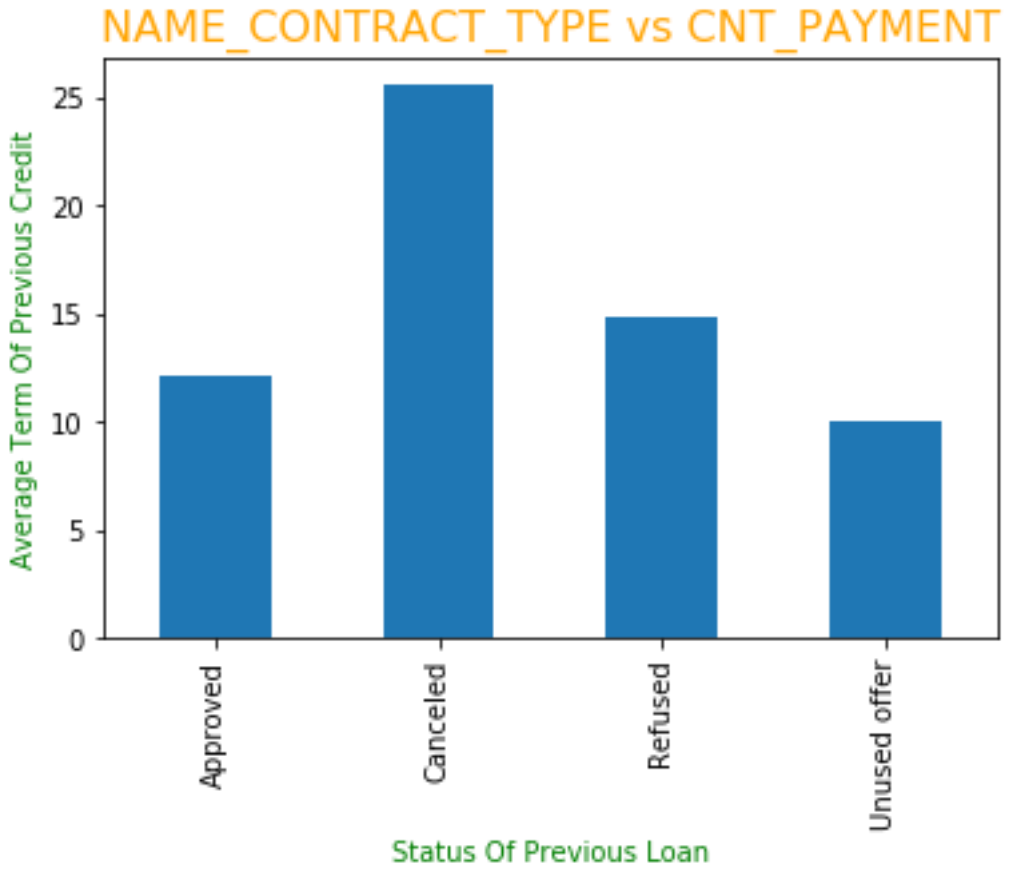
Bivariate Analysis for Numerical and Categorical Variables

- ✓ The annuity amount is very high for cash loans when compared to Consumer and Revolving loans
- ✓ Application Amount is more for rejected loans than other types of loans.
- ✓ Amount credit is more for cashless payment type when compared to other payments
- ✓ The annuity amount of the Cash Portfolio is high compared to POS and cards.



Bivariate Analysis for CNT_PAYMENT and NAME_CONTRACT_STATUS

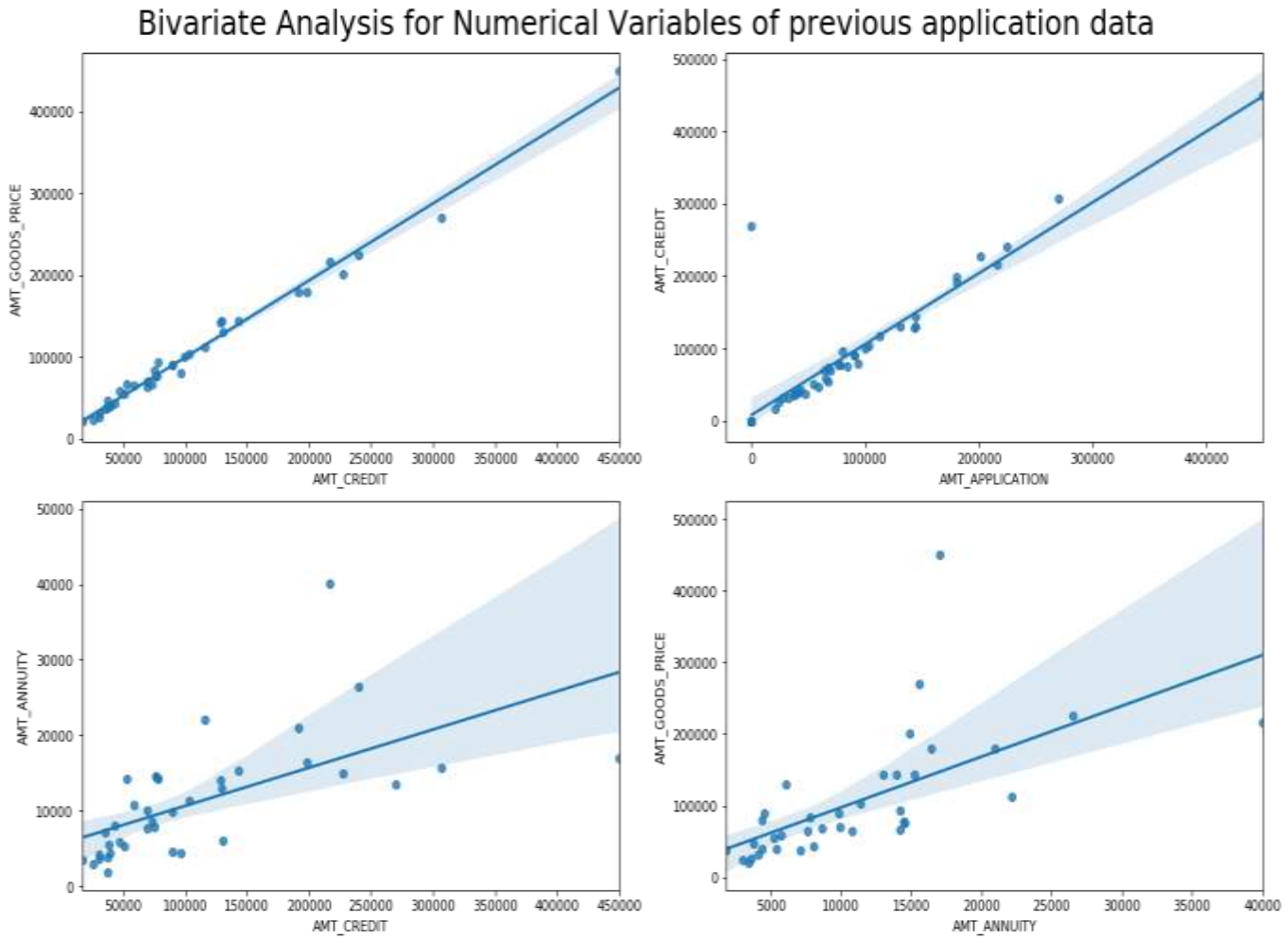
The graph shows that the average term of previous credit on previous loans being cancelled are more than being approved.



Bivariate Analysis for Numerical Variables

Analysis from the graph

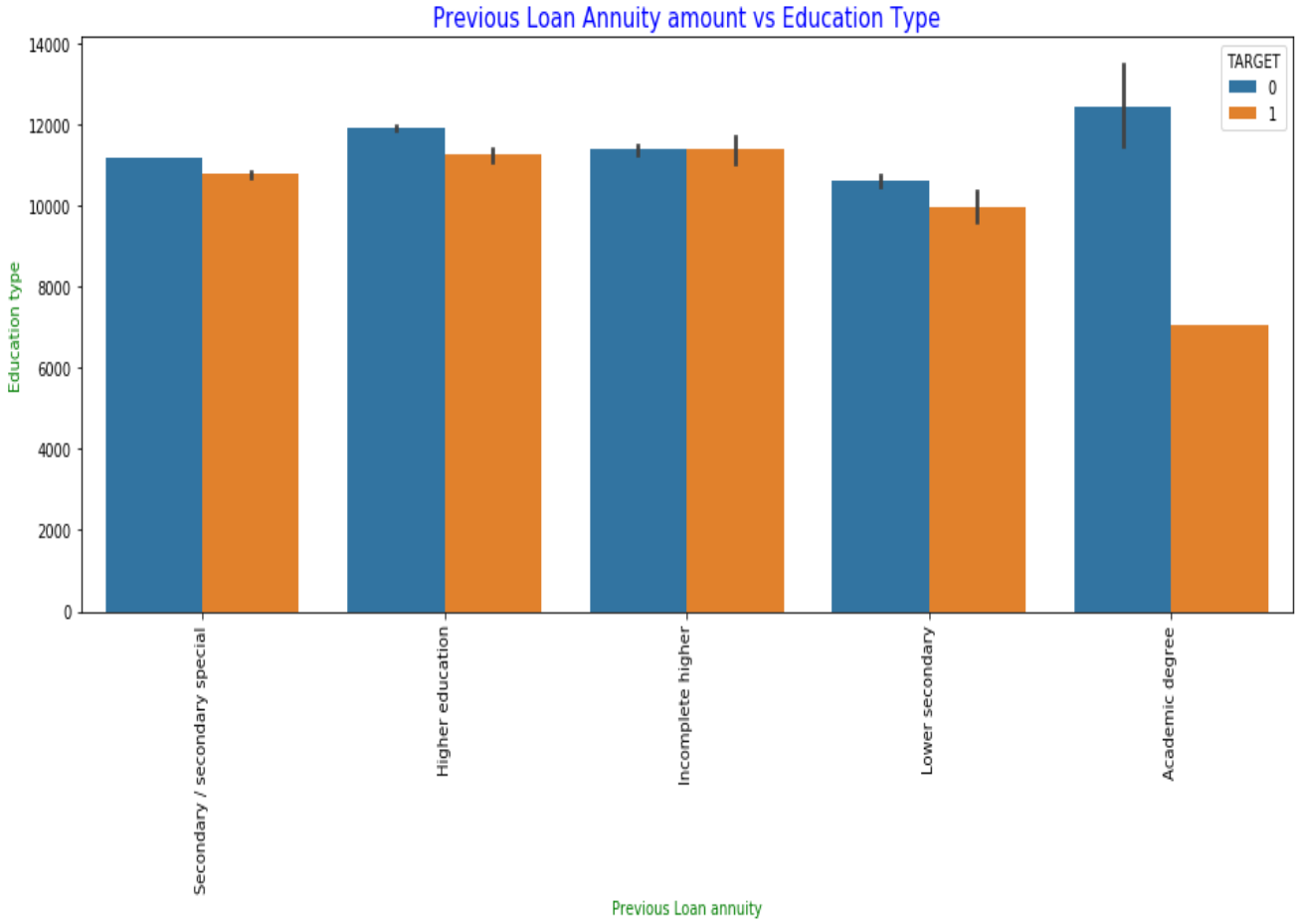
- ✓ The correlation between goods price and Credit amount is 0.98 which is very close to 1.
- ✓ Also the correlation between AMT_CREDIT and AMT_APPLICATION is 0.9
- ✓ Their linear relationship is clearly depicted in the graph.



Conclusions

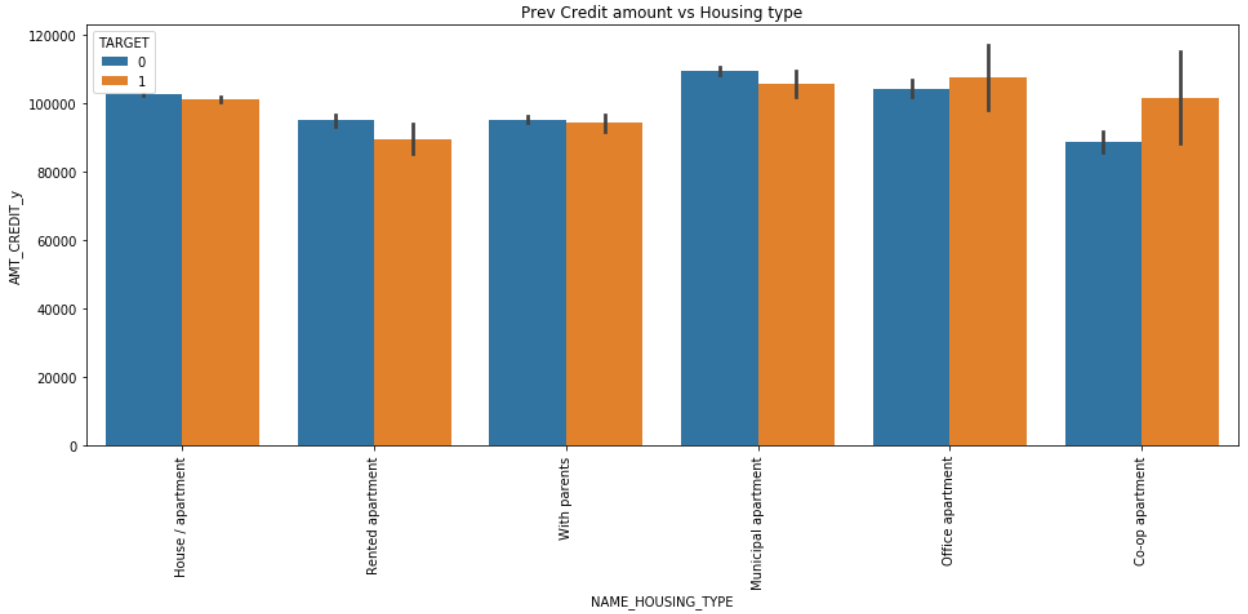
Observation 1:

From the graph, we can conclude that bank can give loans to clients with academic degree as they have low payment difficulties when compared to other education types and they are less likely to default.



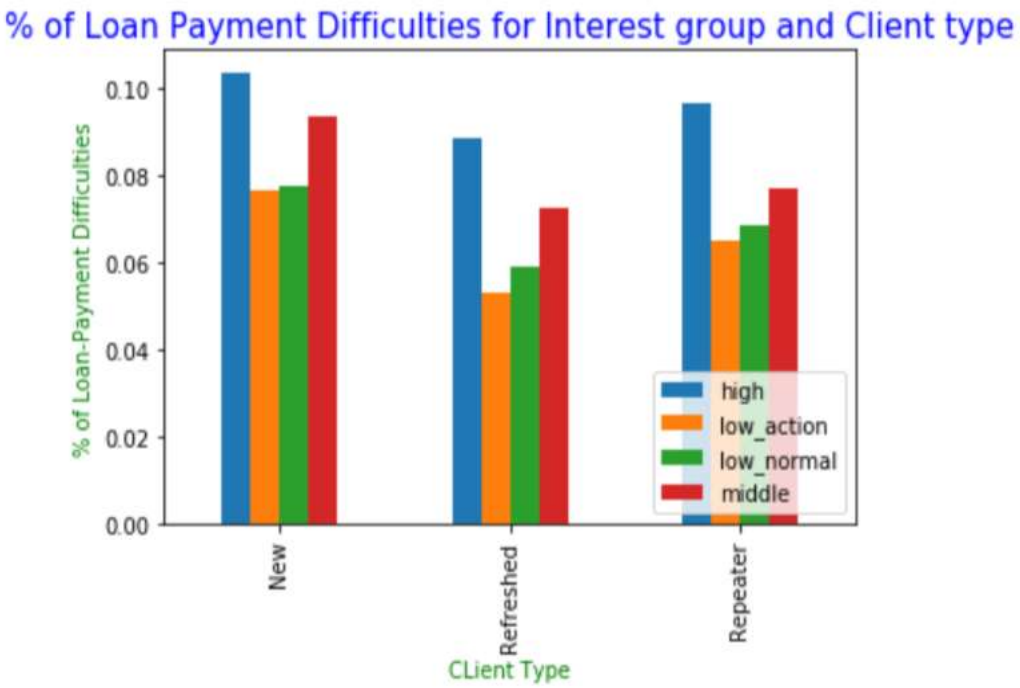
Observation 2 :

We can conclude that bank should avoid giving loans to the housing type of office and co-op apartment as they are having difficulties in payment. Bank can focus mostly on category of clients in rented apartments, house/apartments and with parents for successful payments



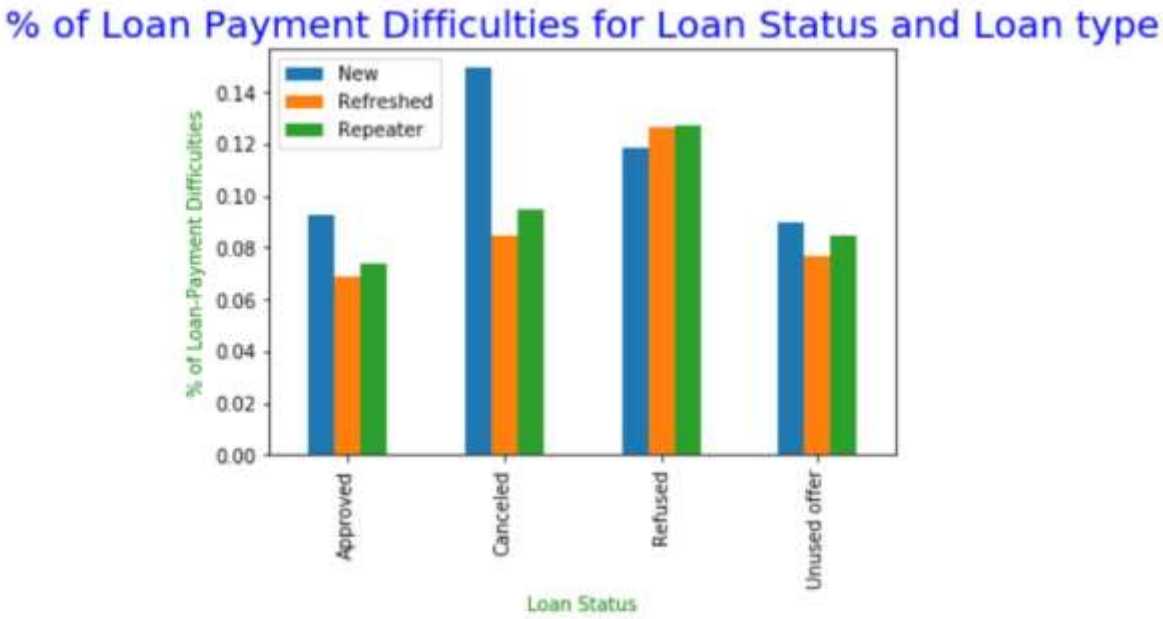
Observation 3 :

We can conclude from the above graph that the bank should focus more on Client type 'Refreshed' and 'Repeater' with yield types 'low_action', 'low-normal' and 'middile' as they have low payment difficulties when compared to other cases.



Observation 4 :

It can be observed from the graph that the **New** clients under **Cancelled** status have more payment difficulties compared to other cases. Bank should focus more on clients with **Approved** and **Unused offer** status as they are less likely to default



Observation 5 :

The above graph depicts the number of **APPROVED** loans for both the Target variables. Clients with 1 and 2 approved loans are having more payment difficulties. But, for customers with atleast 3 approved loans approved, the number of default cases are less when compared to clients with 1 and 2 approved loans(default cases). Therefore, **approved loans equal to 3 and above** from the previous applications represent a low default case.

THANK YOU