Credit EDA Case Study

-- Submitted by: Vibhu Adithya, Seema S B

Note: In the PDF, as the attachments get lost, they have been added into the zip file. Wherever there is reference to attachments, please refer to the PNG files in the zip file along with this.

Data Set: application_data.csv

1. **Data Preparation:**

   Assumption: Any specific variable/column that has more than 13% of its data as null values are not good enough to be used for analysis.

   a. **Handling columns with Null Values**
      i. Find out the percentage of null values in each of the columns of the dataframe
      ii. Find out the columns having more than 13% of its values as null
      iii. Drop those columns from the dataframe **->** 57 columns were dropped as it would not make any insightful meaning to use them
      iv. Identify columns which still have null values. 10 columns had null values as below:

```
AMT_ANNUITY
AMT_GOODS_PRICE
NAME_TYPE_SUITE
CNT_FAM_MEMBERS
EXT_SOURCE_2
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
```

   b. Imputation Metrics
      i. Before removing/imputing the null values, describe function clearly shows that the mean and median are best metrics to be imputed for the column EXT_SOURCE_2
      ii. Till the 99th quantile, there isn't any significant impact and hence for those data the null values can be imputed using the mean after treating the outliers for the columns OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE.
      iii. The null values can be imputed with the value '0' for columns DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE as most of the values are '0'.
      iv. For NAME_TYPE_SUITE column, it makes sense to impute it with the value 'Unaccompanied' which has the highest frequency
      v. For the column AMT_GOODS_PRICE, it would be best for the right tail data to be clipped with the 95th quantile value(not dropped as clipping would make better sense since the data would exist but will not skew the results as it had done before)
      vi. The null values can be imputed with the value '0' for column CNT_FAM_MEMBERS as number of null values is very few.

   c. Conversion of column data types
      i. The datatype of the column CNT_FAM_MEMBERS is converted from float to int type as the number of members in a family can't be float.

   d. Column Derivation
      i. AGE column was created by dividing the absolute value of DAYS_BIRTH by 365.25
      ii. LOAN_PERIOD column was created by dividing AMT_CREDIT column with AMT_ANNUITY

e. Outlier Treatment

i. The outliers in the below columns are treated using the IQR method:

1. AMT_ANNUITY
2. AMT_INCOME_TOTAL
3. AMT_CREDIT

ii. As mentioned in earlier step, the column AMT_GOODS_PRICE is clipped at its 95th quantile

f. Binning of Continuous variables

i. Bins were created for the age column with values from 21-30, 31-40 so on up to 61-70

ii. Bins were created for the LOAN_PERIOD column with values from 6-10, 11-15 so on up to 51-55

2. Data Analysis

a. Balance Check and Data split

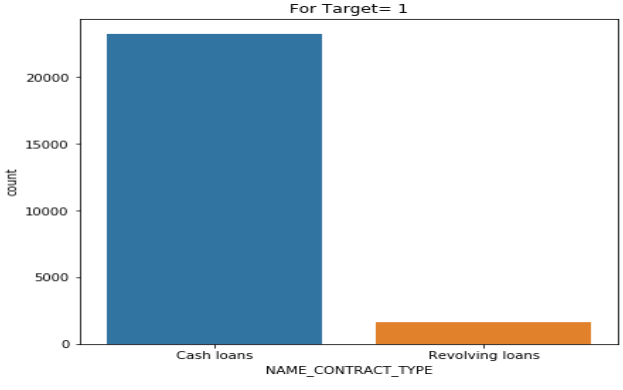
i. The dataset is checked for imbalance percentage with respect to TARGET i.e., 0 or 1

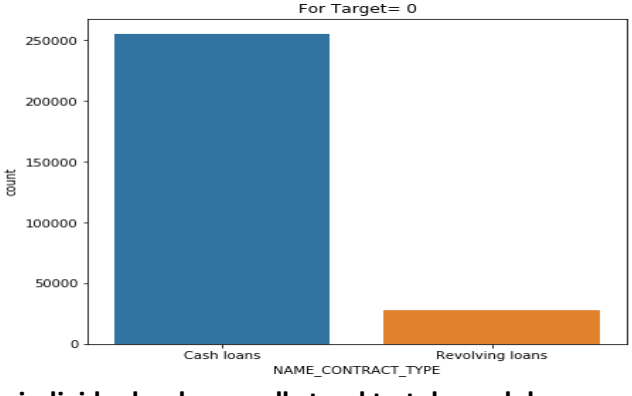
ii. For analysis purpose, the data set was divided into two data frames based on the TARGET column

1. app_data_tgt_1denoted the data set where clients had payment difficulties
2. app_data_tgt_0 denoted otherwise

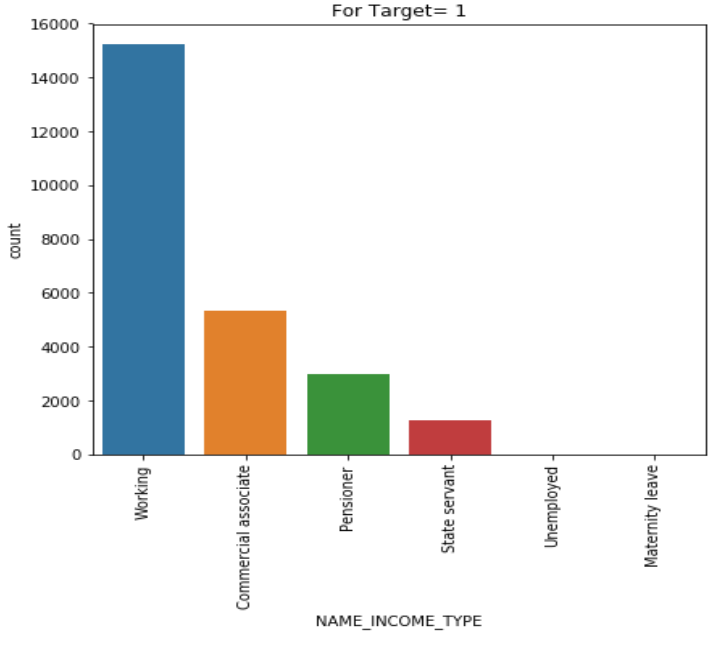
24,825 applicants had payment difficulties and 282,686 did not

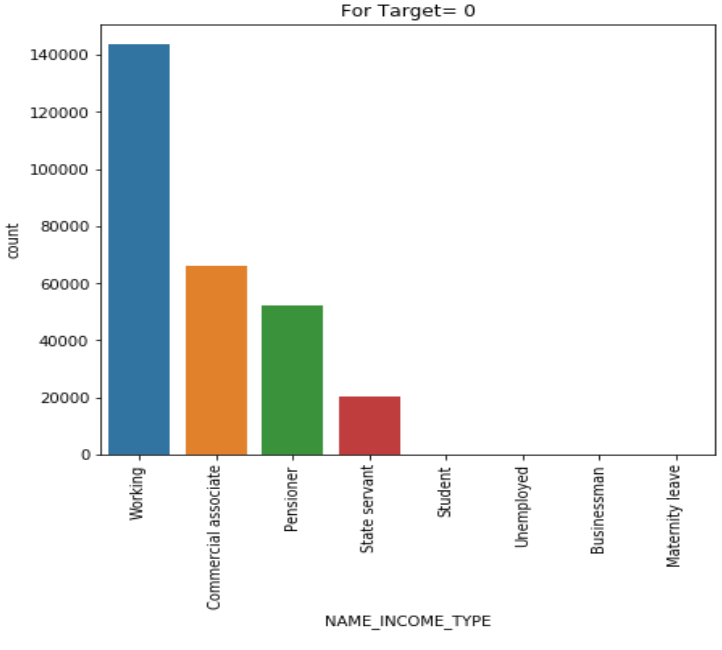
b. Univariate Analysis:



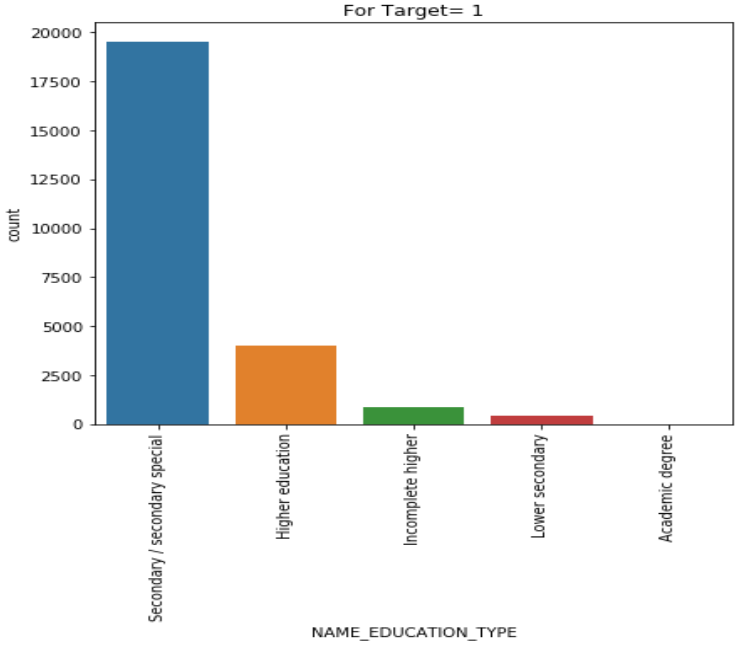


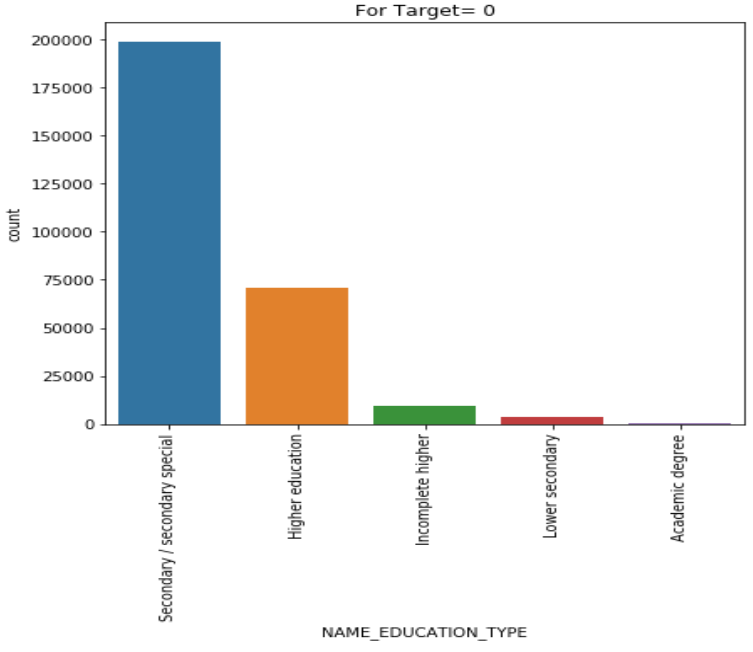
From the above, we can see that loans were taken primarily by individuals who usually tend to take cash loans and not by those who tend to take revolving loans



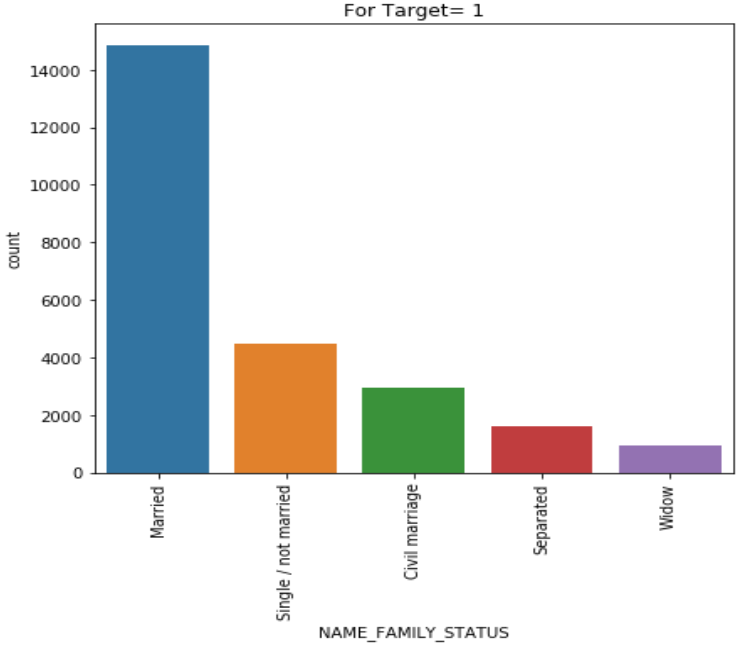


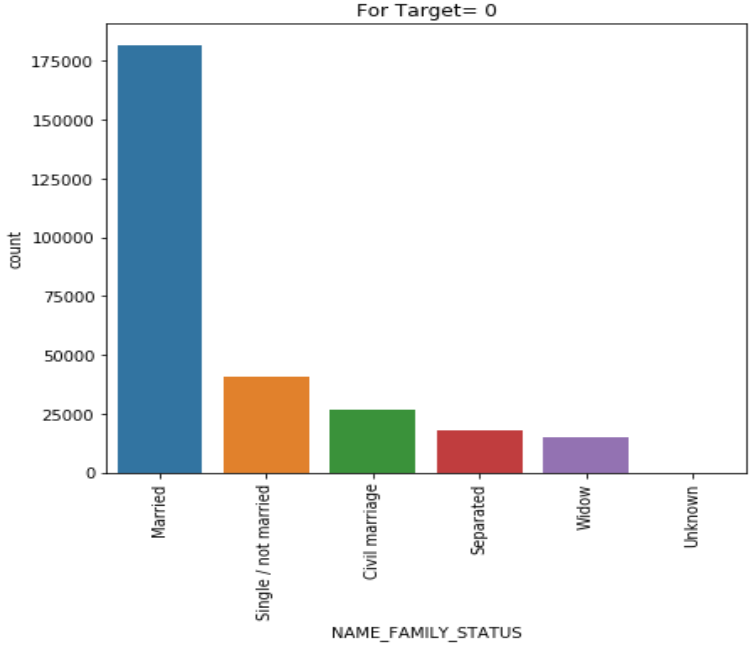
From the above, we can see that irrespective of the target, working class applied the highest for loan followed by commercial associate. One possible assumption from this is that people who earned more had more confidence in the repaying capacity and hence were willing to take loan more than the others



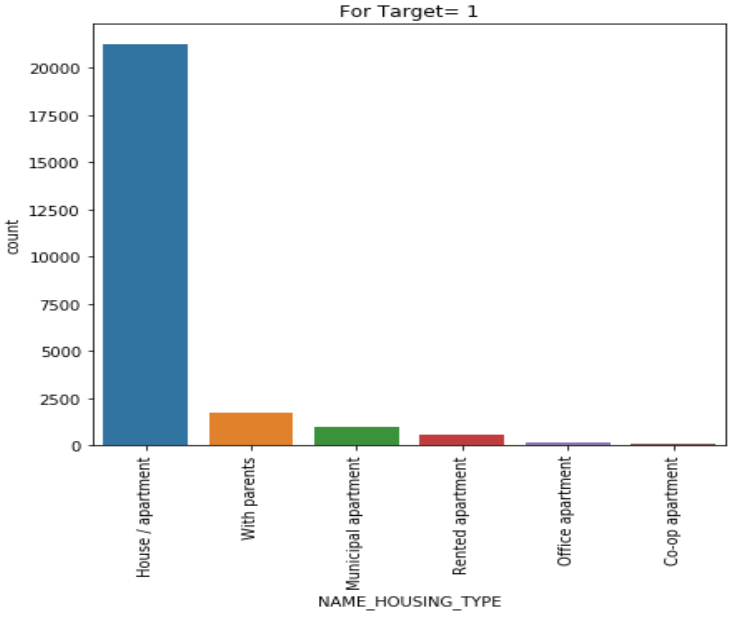


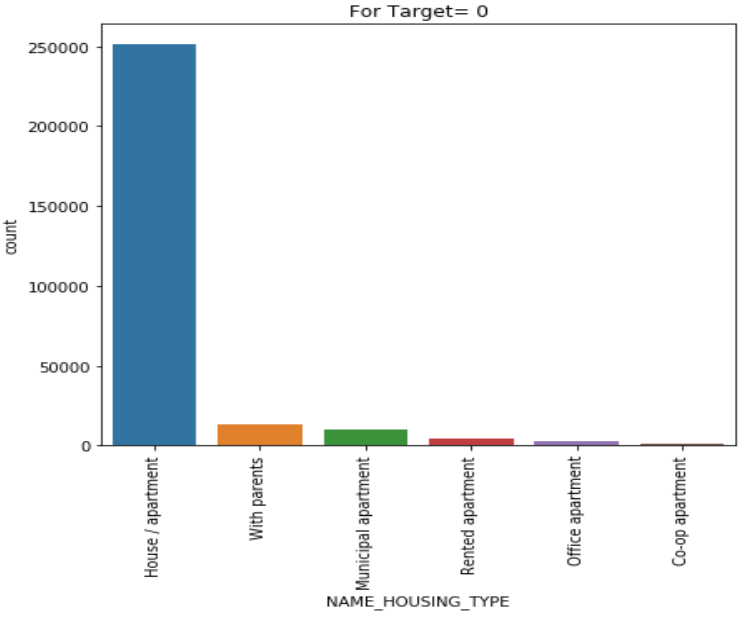
From the above, we can see that irrespective of the target, the applicants with Secondary education have the higher likelihood of applying for loan.



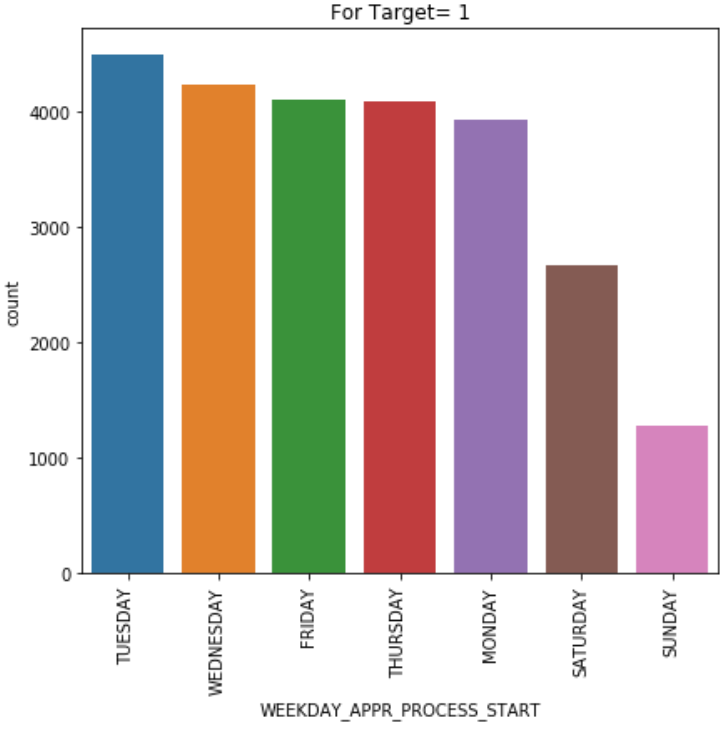


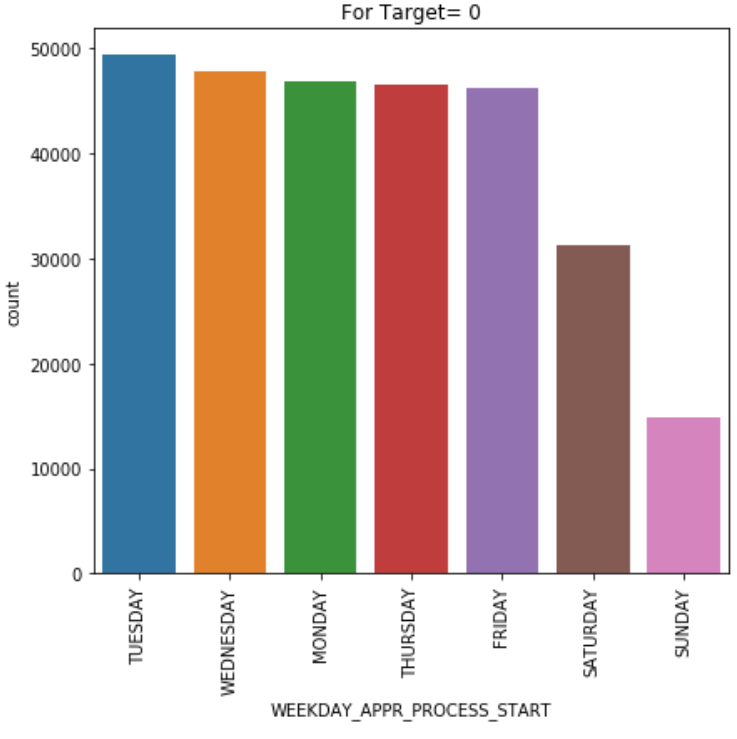
From the above, we can see that irrespective of the target, married individuals had applied for loans much more often than single or the other types. This might mean that married applicants have more commitments in life where loan would be a requirement, or it can just mean that there is more married population than the rest



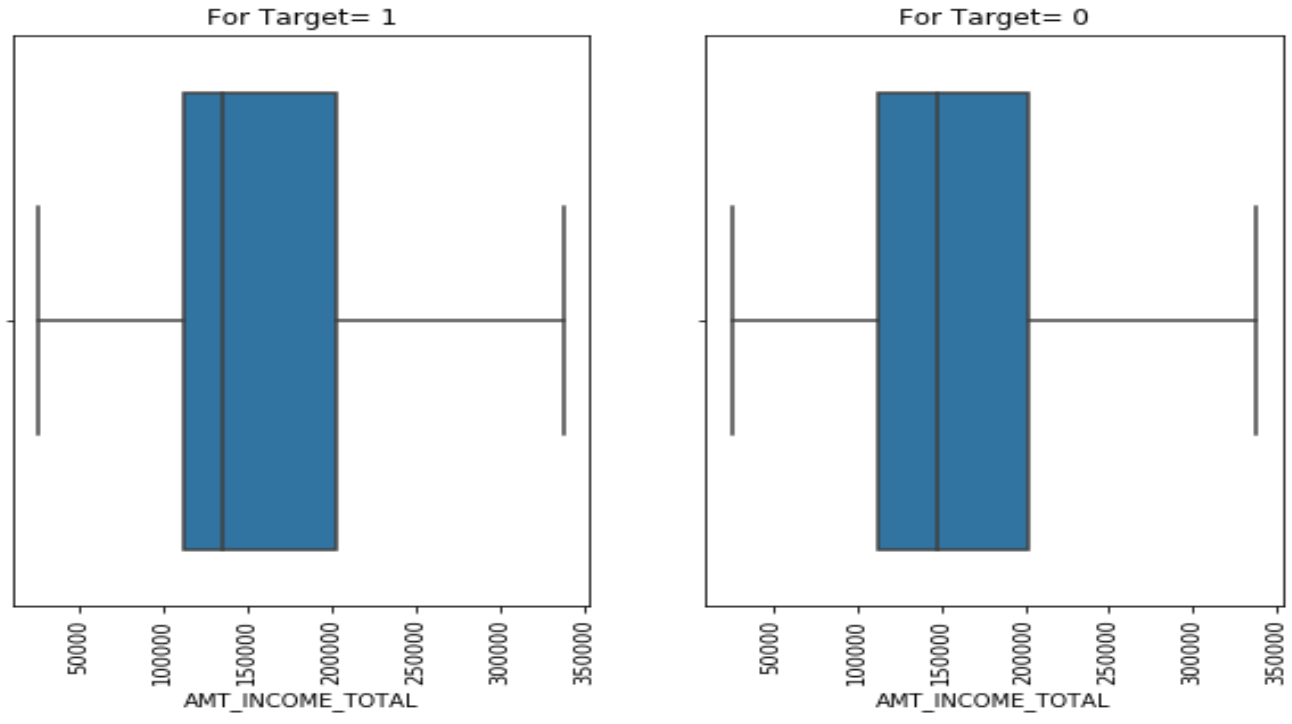


From the above, it is very clear those applicants who live in House/Apartment are exponentially higher than the other types. This might mean that the data set had the applicant information primarily of housing loans and this data might have been collected in urban areas where Apartments are highly prevalent

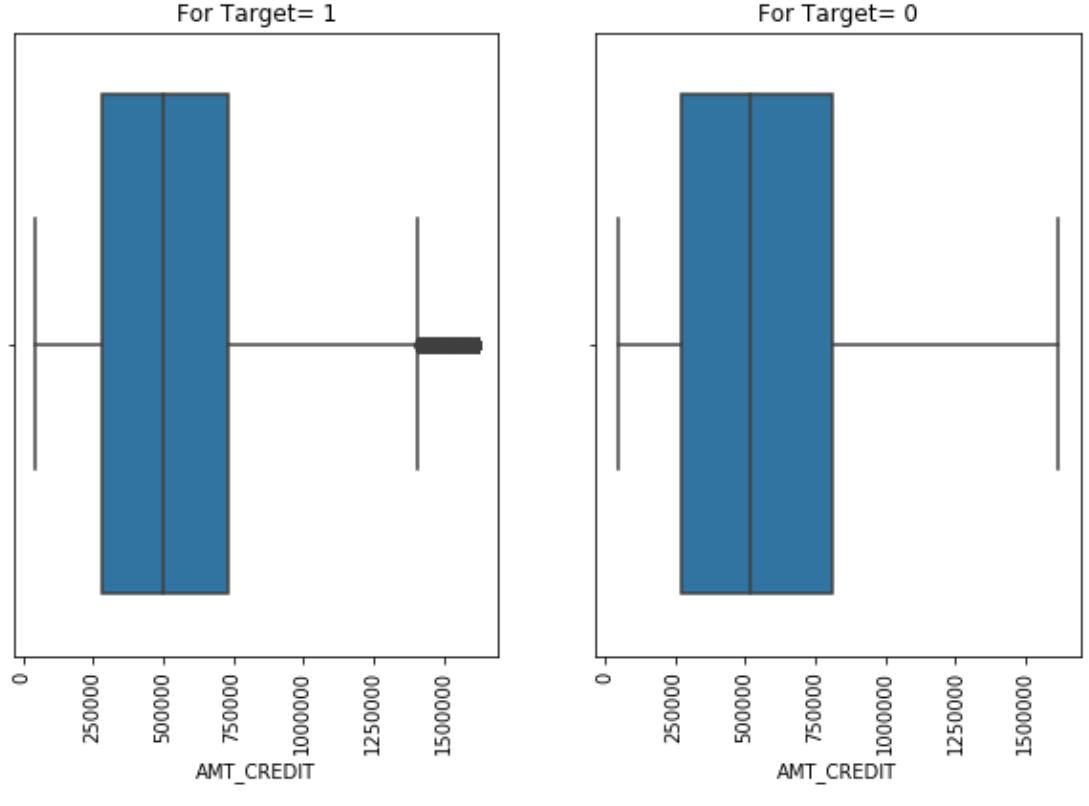




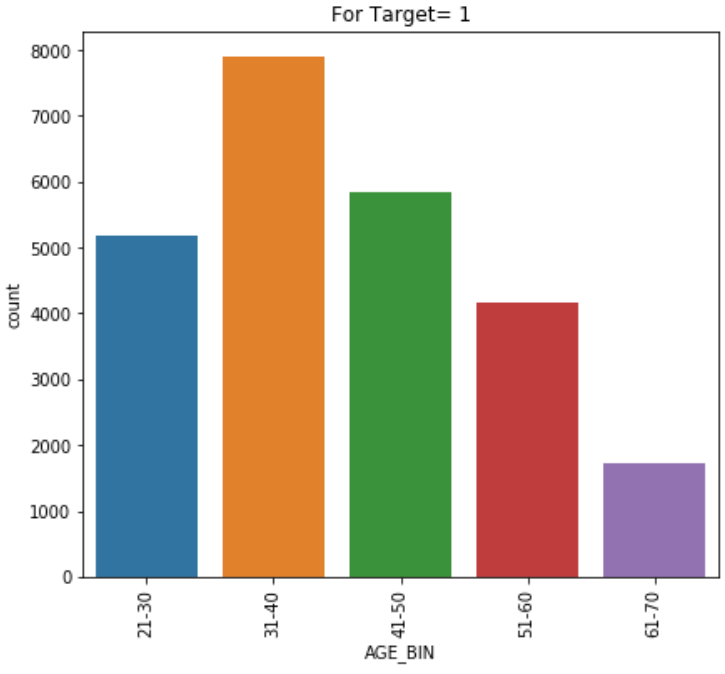
Clients seem to have applied for loans primarily on the weekdays, partially on Saturday and much lesser on Sundays. This may be due to the fact that banks don't function on Sundays and on alternate Saturdays. The applications that were submitted on Sundays might be through agents/brokers who work on Sundays as well for incentives

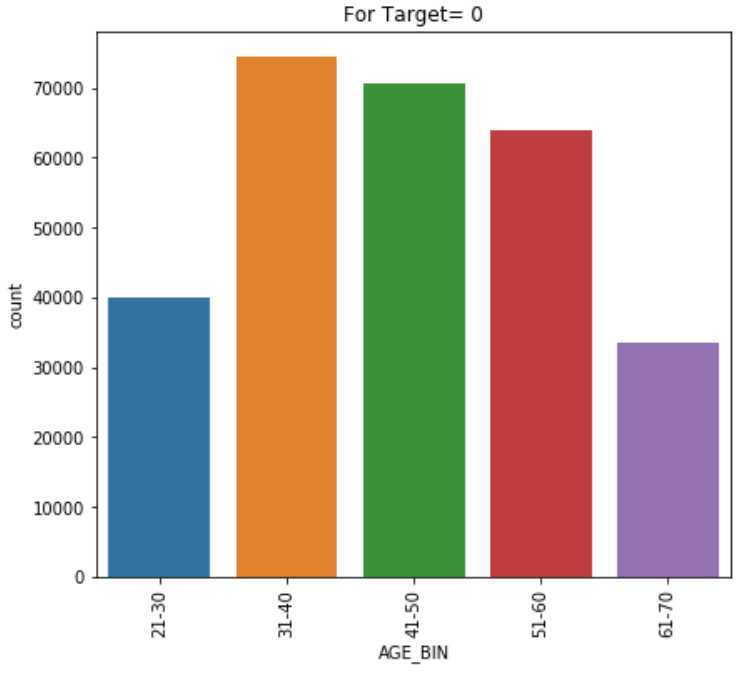


From the above, we can see that irrespective of the target, the total income of the applicants lies in the range of 100000 to 200000

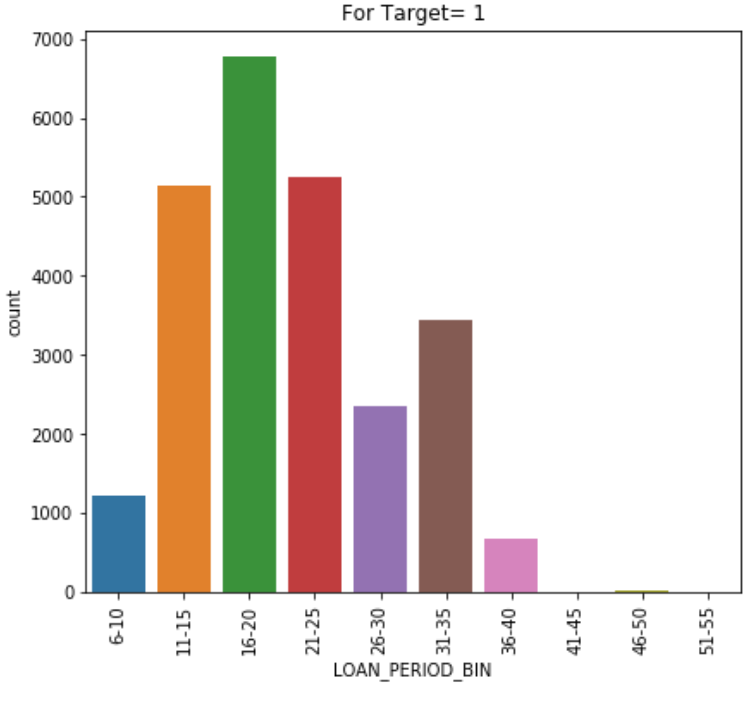


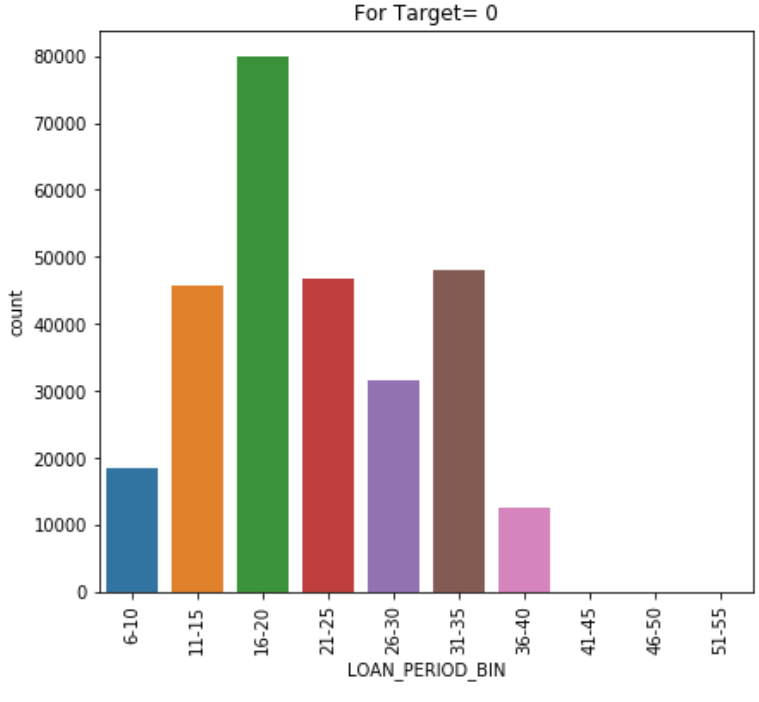
From the above, we can see that irrespective of the target, the credit amount of applicants are similar and lies in the range of 250000 to 750000





From the above, when the applicant is a little younger (say 21-40 years of age), they seem to have payment difficulties but in case when the applicant is little older than the other case, they don't seem to have payment difficulties.





From the above, we can see the pattern of applicants having payment difficulties (or defaulters) and non-defaulters are similar.

c. **Correlation matrix**

i. **From the below two attached images (zoom to see the values), it is clearly visible that the correlation pattern is very much similar for the defaulters and non-defaulters.**

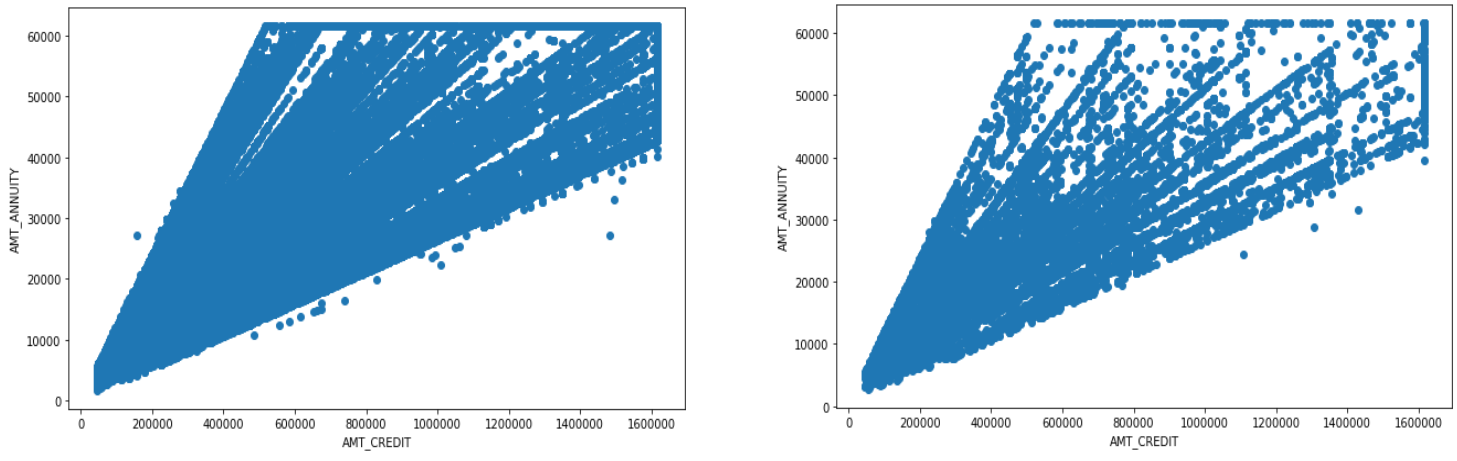
corr_tgt_0.png corr_tgt_1.png

ii. **We can also see that in both the data sets, AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE are highly correlated to each other.**

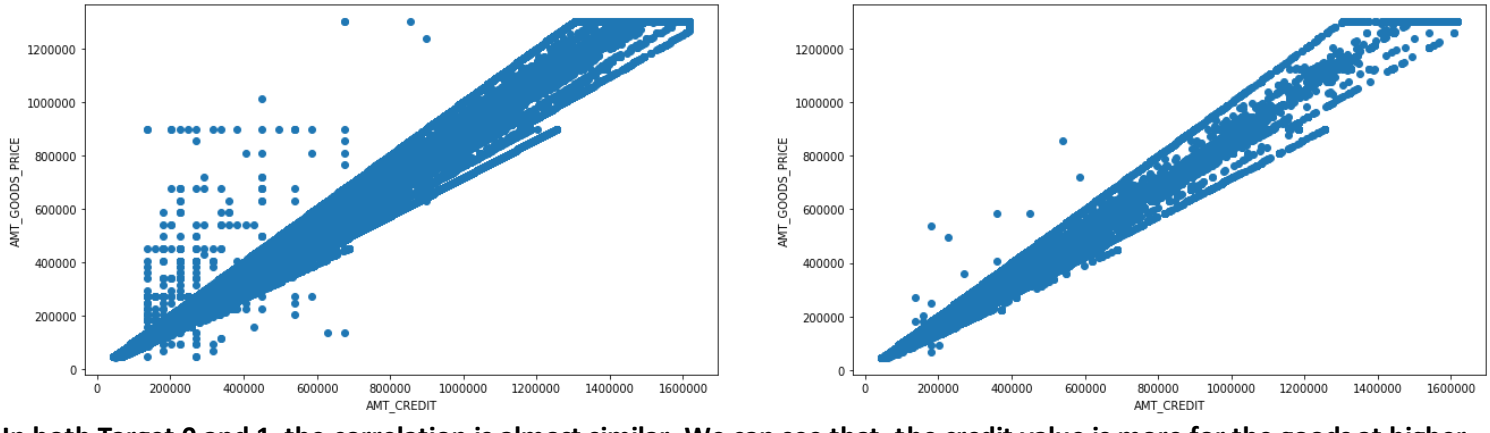
iii. **Also LOAN_PERIOD, AMT_CREDIT and AMT_GOODS_PRICE are significantly correlated to each other.**

d. **Bivariate Analysis:**

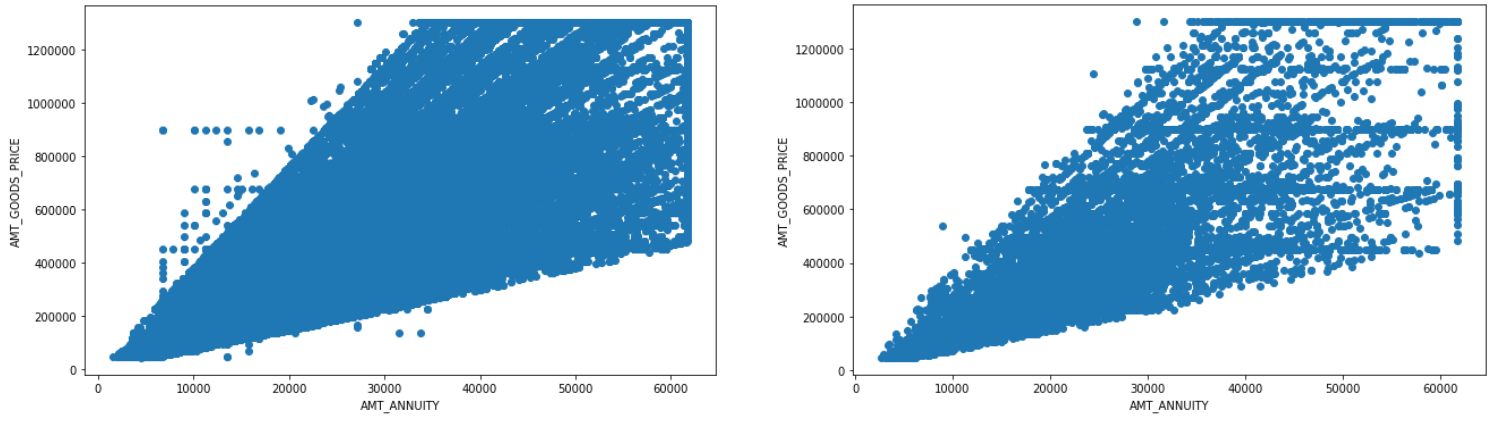
[Left plot corresponds to Target 0 and right plot to Target 1]



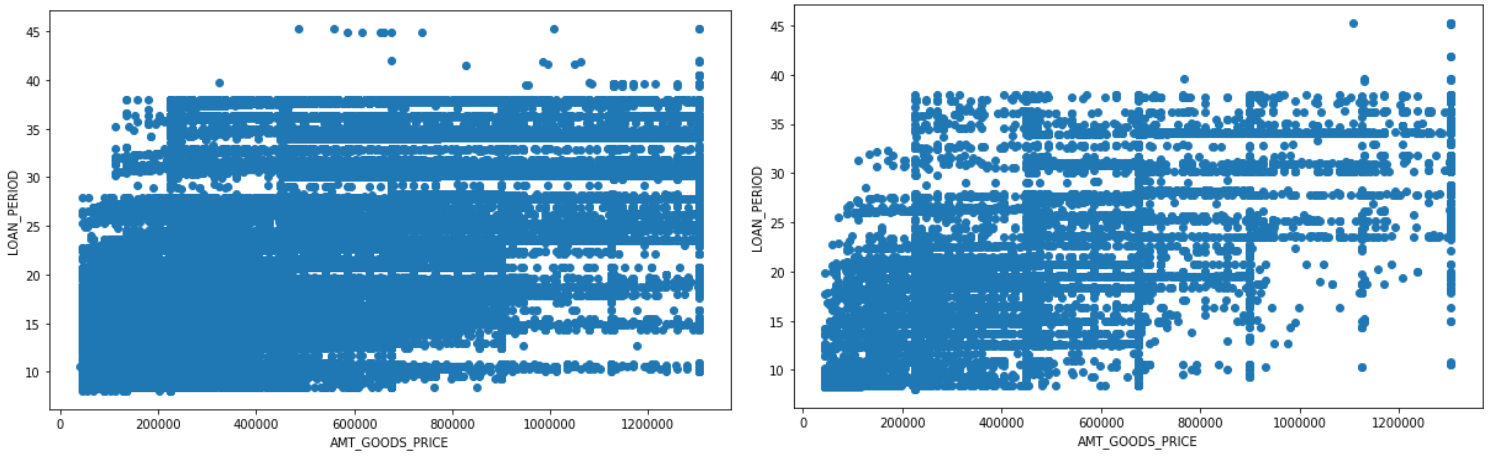
In both Target 0 and 1, the correlation is almost similar. Higher the credit value, higher will be the amount to be paid annually.



In both Target 0 and 1, the correlation is almost similar. We can see that, the credit value is more for the goods at higher price.

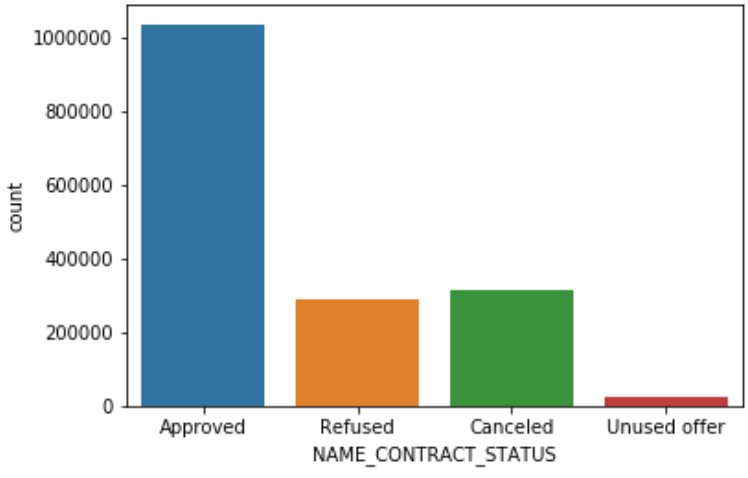


In both Target 0 and 1, the correlation is almost similar. The annuity amount is high for higher goods price which is because, higher the good price there will be higher credit amount.

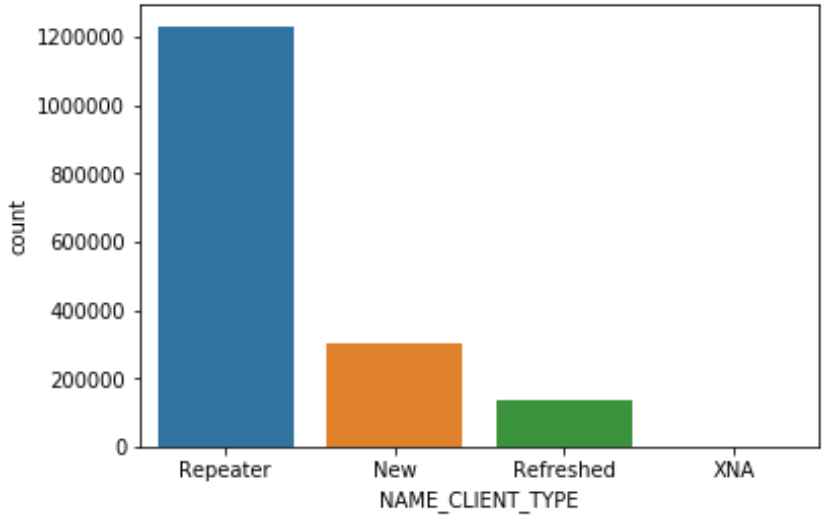


In both Target 0 and 1, the correlation is almost similar. The applicants had payment difficulties more at loan period of 20 years but with increase in price amount or the loan period they are sparsely populated, which could be because the loan was not approved such high good price or for greater tenure of loan.

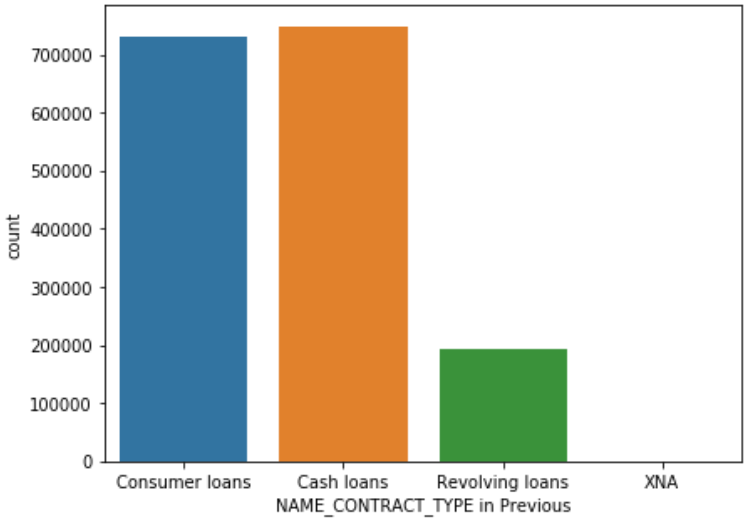
3. **Analysis on previous and current application data:**
    a. **Previous and current dataset**
        i. Merge both the previous and current data
    b. **Univariate Analysis**

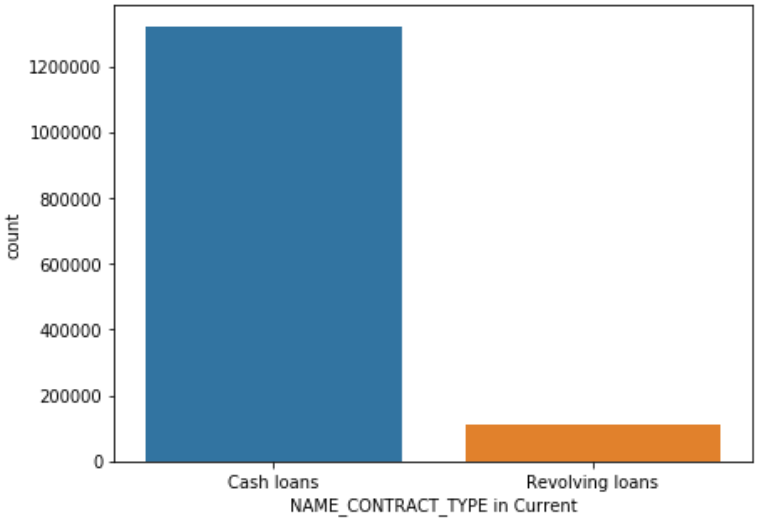


From the above we can see that most of the previous applications are of 'Approved' status, which means the rejection rate of applications by the bank was very low for the previous application.



From the above, we can say that most of these 'Approved' applications are 'Repeaters' which could result in the inference that since the rejection rate is less, there are more applications who likes to avail loan again or the bank wants to hold the customers who have been previously sanctioned.





From the above, we can see that cash loans were more predominant in both the previous and current data than revolving loans. Also, the applicants who availed consumer loans and XNA type loans previously have preferred Cash loans in the current application.

c. Correlation

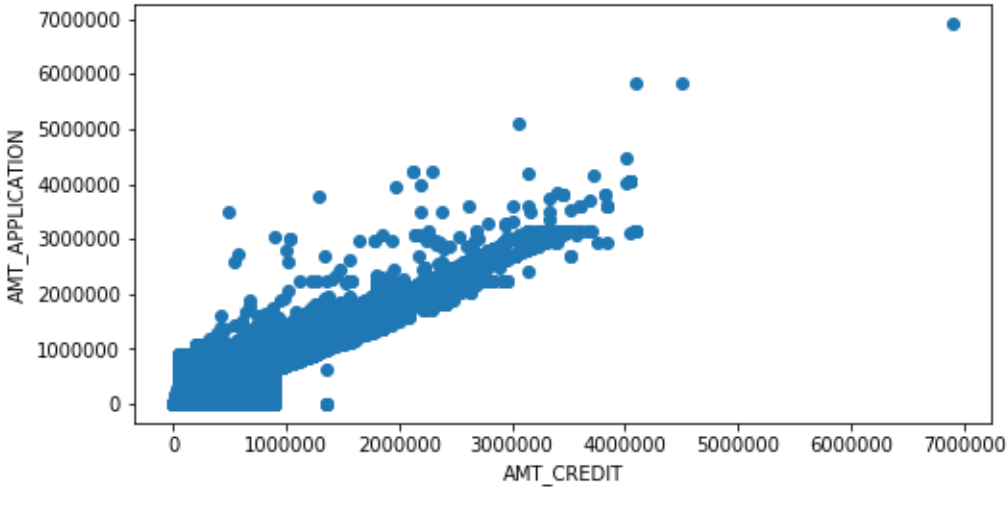
i. From the below two attached images (zoom to see the values), it is clearly visible that the correlation pattern is very much similar for the ones where the applicant has payment difficulties and otherwise.

corr_merge.png

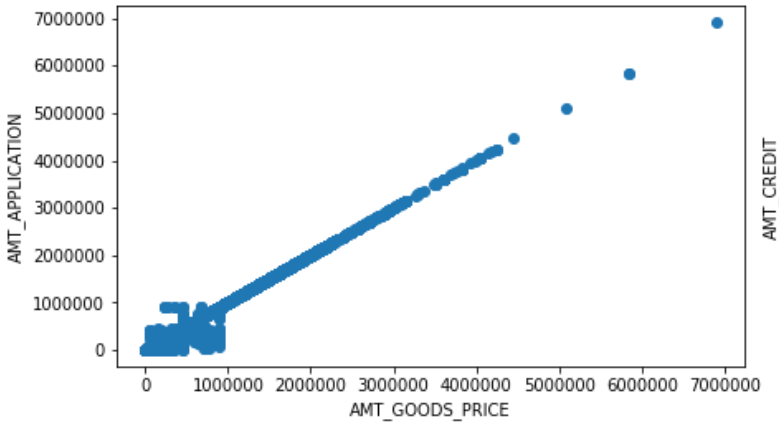
ii. Similar to the previous data set, we can also see that in the merged data sets, the below columns are highly correlated:
   1. DAYS_LAST_DUE and DAYS_TERMINATION
   2. AMT_ANNUITY_x, AMT_APPLICATION, AMT_CREDIT_x and AMT_GOODS_PRICE_x
   3. AMT_ANNUITY_y, AMT_CREDIT_y and AMT_GOODS_PRICE_y

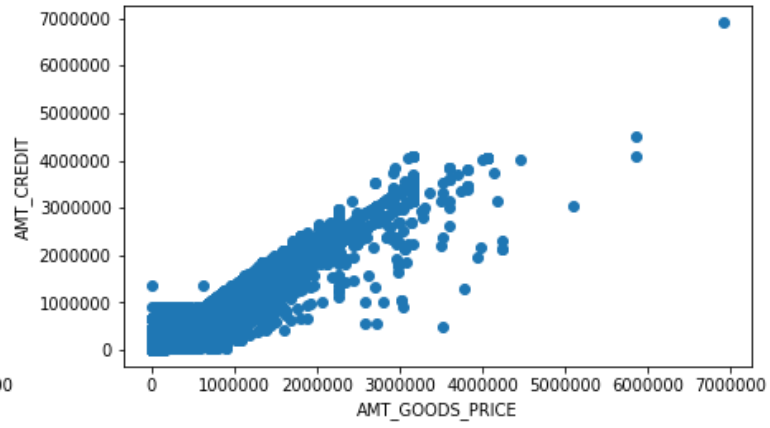
iii. Similarly, based on the correlation matrix, lot of other meaningful insights can be derived.

d. Bivariate Analysis

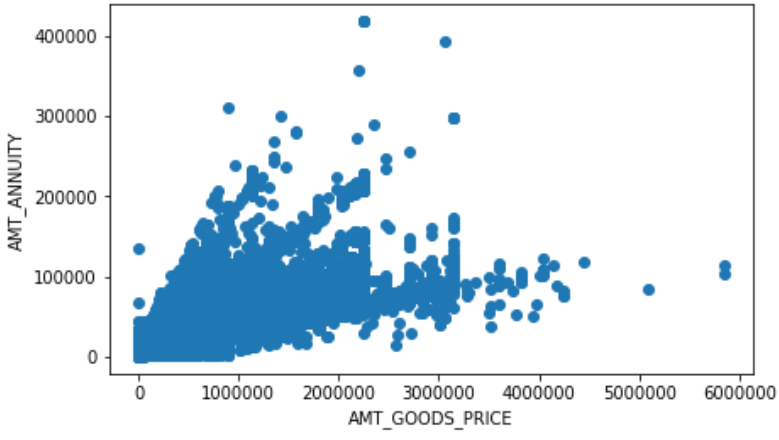


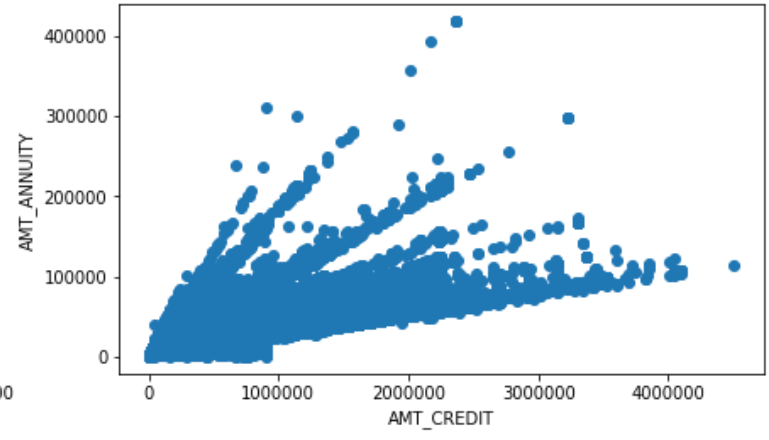
From the above we can see that not always the expected loan amount has been approved. There are scenarios where in the applicants got the credit as per their applied value or sometimes lower or sometimes higher too.



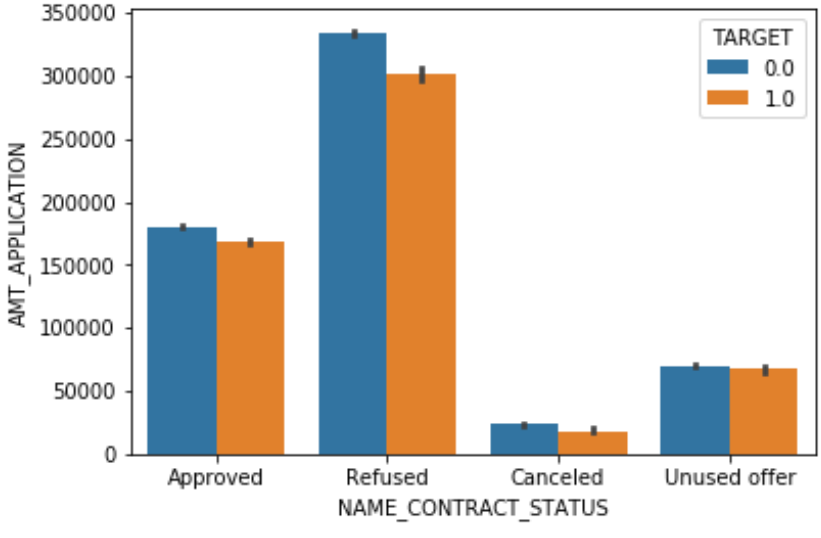


From the before mentioned steps, higher the goods price higher will be the credit and also not all the time the credit will meet the expected application amount, hence there can be some deviations at such points.





As mentioned earlier, the higher good price will result to higher credit amount and hence higher annuity. But this can also vary based on various factors like total income of individual or loan period too. For e.g., the annuity will be more for applicants with higher income total for a credit amount and can be much lesser for applicants with lower income total for the same credit amount.



From the above we can clearly see that the bank has approved for those who had payment difficulties and refused for those who were non-defaulters.

So based on the above explained insights, the bank can restructure its plans so that it can result in more profit and gains by approving loans to non-defaulting clients and making sure less credit loans are approved to defaulters or defaulters loan application are refused.