Case Study

Exploratory Data Analysis

Sindhu N
Aniket Verma

Objective: 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. In other words, to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Given Data:

1) Application Data: It contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
2) Previous Application: It contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Let us first consider the application dataset. We will divide the complete data set into two categories based on the target variable.

Target variable = 1: client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample.

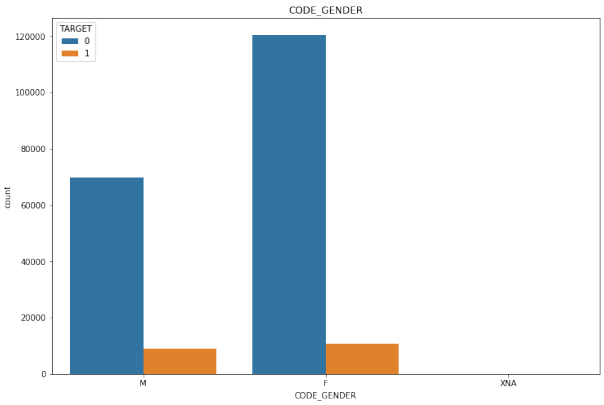
Target variable = 0: all other cases)

Let us try to identify the effects of various parameters on probability of an applicant defaulting on a loan installment.

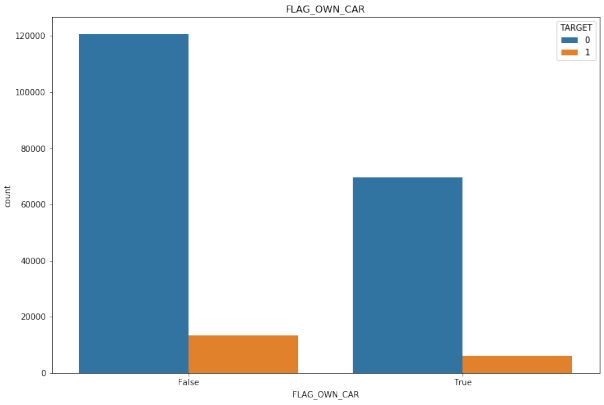
UNIVARIATE DISTRIBUTIONS:

CATEGORICAL DATA:

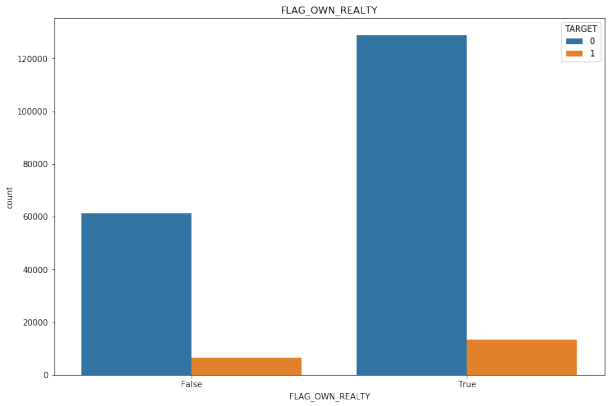
- GENDER
- *Observation: The probability of women defaulting on a loan installment is much smaller than that of men. Moreover, the number of women applicants are more than the number of men applicants.*
- *Inference: Granting a loan to women is much safer as they are more proactive in timely payments.*



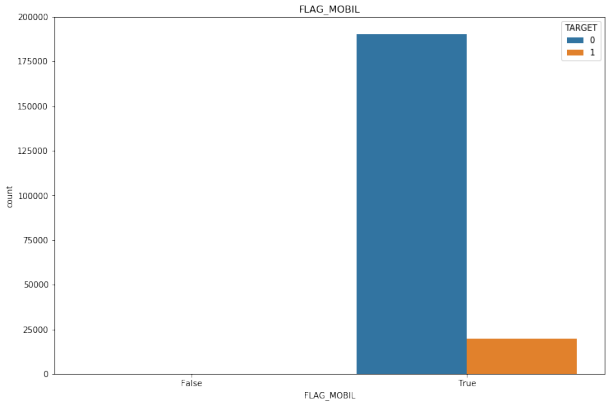
- VEHICLE OWNERSHIP
- *Observation: The probability of people who don't own a vehicle defaulting on loan installment is much lower than that of people who own a vehicle.*
- *Inference: Most people who own a vehicle might have a running CAR LOAN EMI, which forms a significant chunk of their monthly salary and thereby causing occasional loan payment misses due to lack of finance.*



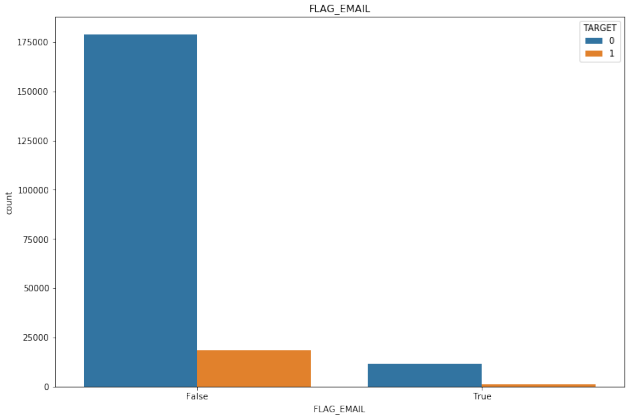
- REAL ESTATE OWNERSHIP
- *Observation: The probability of people who own a property defaulting on a loan is much smaller than those who do not own a property.*
- *Inference: This result is contrary to the previous result of vehicle ownership. The most apt explanation of the trend is that while a vehicle is a depreciating asset and is solely used for individual comfort, a real estate property might be an investment in terms of land value or even if the property has been rented out which would help in cover home loan EMI.*



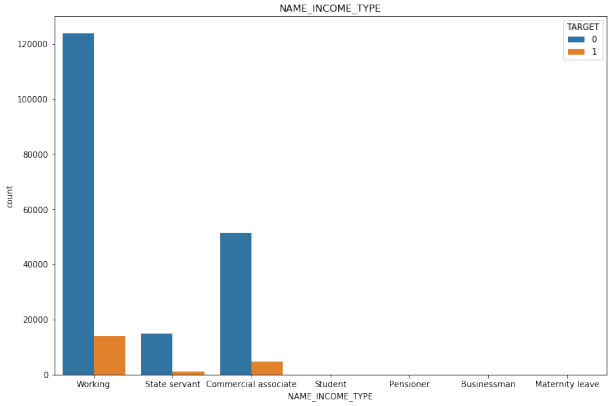
- MOBILE PHONE OWNERSHIP
- *Observation: In today's world there is hardly anyone who doesn't own a cellphone. This is pretty evident from the below representation.*
- *Inference: A small chunk of people who own a cellphone default on their payment cycles. This can be attributed to the fact that though everyone has a cellphone, not everyone is aware of mobile banking services in order to make their payment easy. Hence, we cannot take the mobile ownership as a parameter to decide whether the person will default the loan or not because almost everyone owns a mobile phone now a days, this is not good parameter to decide.*



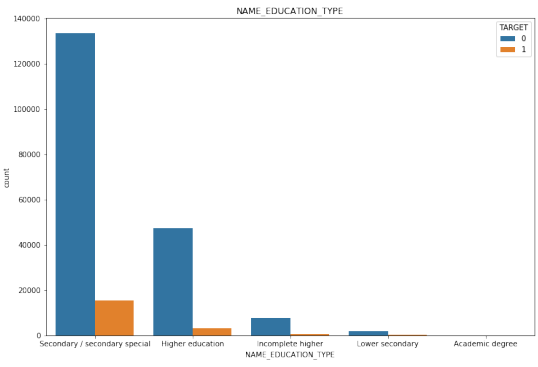
- HAS VALID EMAIL
- *Observation: This is a very unusual trend where the number of people who do not own an email ID is way more than those who do.*
- *Inference: Not much can be said from this data. However, an valid email ID should be a parameter as it makes two way communication easier.*



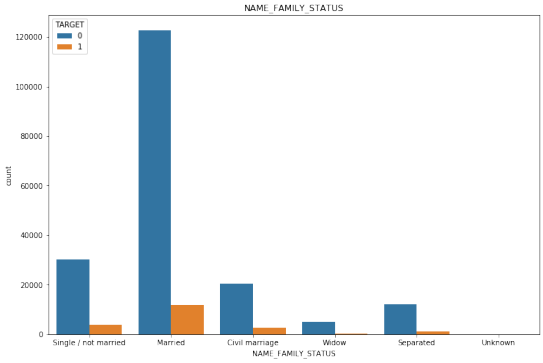
- INCOME TYPE
- *Observation: Working class has the majority number of loan applicants followed by commercial associates and state servants. Students, pensioners do not constitute loan applicants because they do not have a significant income and are in survival phase. Businessmen have good capital gains while expectant mothers are off work with no regular income.*
- *Inference: Working class people are the usually the safest to provide loans while strong checks need to be done for minimal income groups before loan disbursal.*



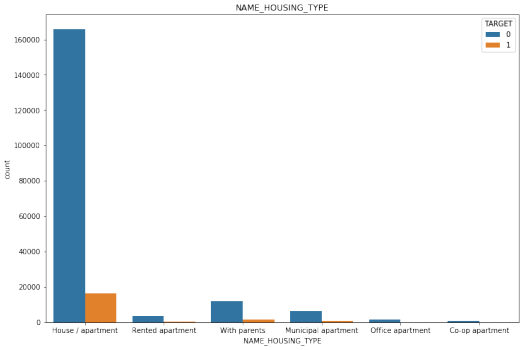
- LEVEL OF EDUCATION
- *Observation: Without a doubt the majority class of people applying for loan is amongst secondary education followed by higher education.*
- *Inference: People with lesser education might have a proportionate income and may not be eligible for loan. The people with secondary education usually have lesser pay than those with higher education and are more likely to have a loan application.*



- FAMILY STATUS
- *Observation: Married people have the highest number of applicants as they have to purchase more liabilities and assets like home, vehicle, children education etc.*
- *Inference: Though the count of defaulters is higher in the case of married couples, the relative proportion doesn't vary significantly across categories.*

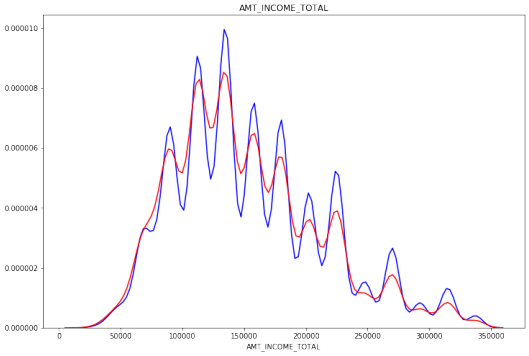


- TYPE OF ACCOMODATION
- *Observation: People which own house (purchased) form the biggest pool of loan applicants. Rest all categories do not have home loan EMI.*
- *Inference: Again the relative proportions remain the same.*

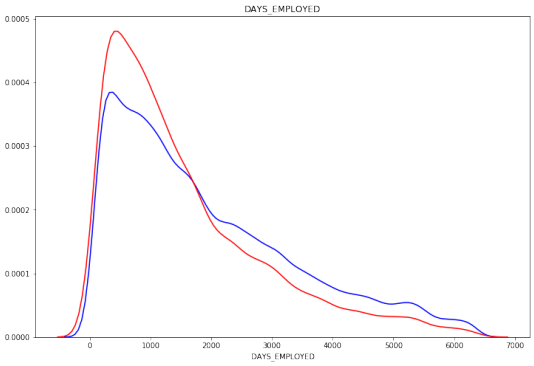


CONTINUOUS DATA:

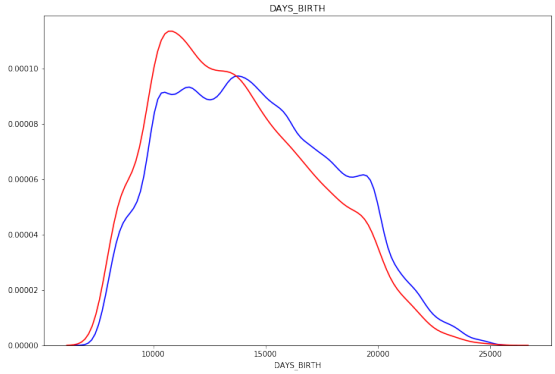
- TOTAL INCOME
- *Observation: For people with higher income (roughly > 250000) the probability of defaulting becomes lower.*
- *Inference: As we move towards the higher income group overall the number of people who avail loans decreases and so does the probability of defaulting.*



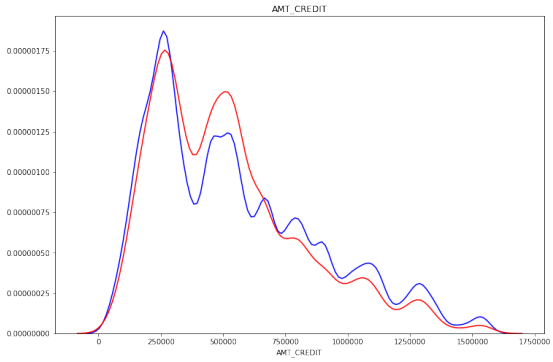
- NUMBER OF DAYS EMPLOYED
- *Observation: Up till roughly 1600 days (5 years approx), the probability of defaulting is higher which decreases beyond that point.*
- *Inference: People who have started their career or are in the initial phases are more susceptible to defaulting as compared to mid-level or senior employees. This is due to various reasons like lifestyle, responsibility, financial planning etc.*



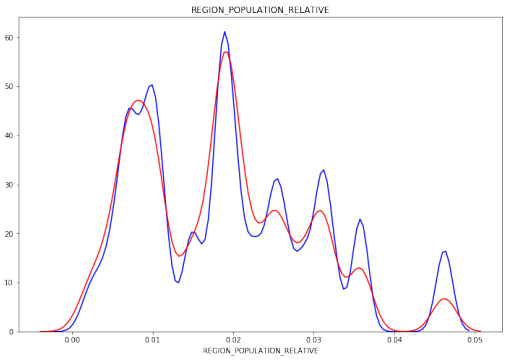
- AGE (DAYS SINCE BIRTH)
- *Observation: For people with less age the probability of defaulting is higher.*
- *Inference: This goes in perfect sync with our previous observation.*



- LOAN AMOUNT CREDITED
- *Observation: Using the trend we observe that there are two intersections at roughly 250000 and 650000. Let us call these points A and B. Let region before A is a steady region with not much to work with. Between A and B, higher credit amount is disbursed but the probability of defaulting increases. Vice versa for the region beyond B.*
- *Inference: This shows that probability of defaulting changes with credit amount and higher credit amounts attract more defaulting.*



- RELATIVE POPULATION REGION
- *Observation: Observing the numerous intersections in the graph, not much can be said about this variable's impact.*



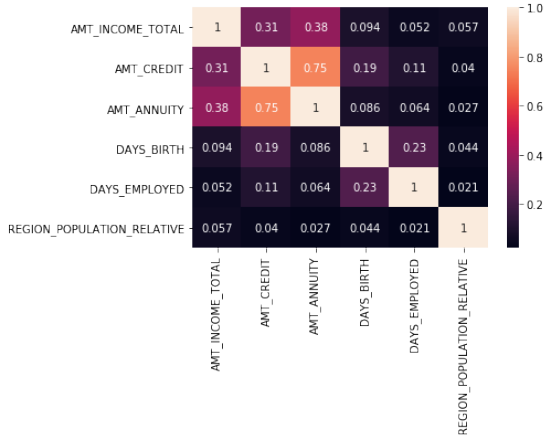
CORRELATION OF NUMERICAL DATA:

Now let us determine the correlation between the numerical (continuous) data of defaulter and non-defaulters.

Target = 0:

	VAR1	VAR2	CORR
13	AMT_ANNUITY	AMT_CREDIT	0.763883
12	AMT_ANNUITY	AMT_INCOME_TOTAL	0.396813
6	AMT_CREDIT	AMT_INCOME_TOTAL	0.325310
27	DAYS_EMPLOYED	DAYS_BIRTH	0.239472
19	DAYS_BIRTH	AMT_CREDIT	0.156734
20	DAYS_BIRTH	AMT_ANNUITY	0.089828
25	DAYS_EMPLOYED	AMT_CREDIT	0.087420
30	REGION_POPULATION_RELATIVE	AMT_INCOME_TOTAL	0.081095
26	DAYS_EMPLOYED	AMT_ANNUITY	0.062220
24	DAYS_EMPLOYED	AMT_INCOME_TOTAL	0.053175
33	REGION_POPULATION_RELATIVE	DAYS_BIRTH	0.042837
18	DAYS_BIRTH	AMT_INCOME_TOTAL	0.042641
32	REGION_POPULATION_RELATIVE	AMT_ANNUITY	0.041480
31	REGION_POPULATION_RELATIVE	AMT_CREDIT	0.038188
34	REGION_POPULATION_RELATIVE	DAYS_EMPLOYED	0.005775

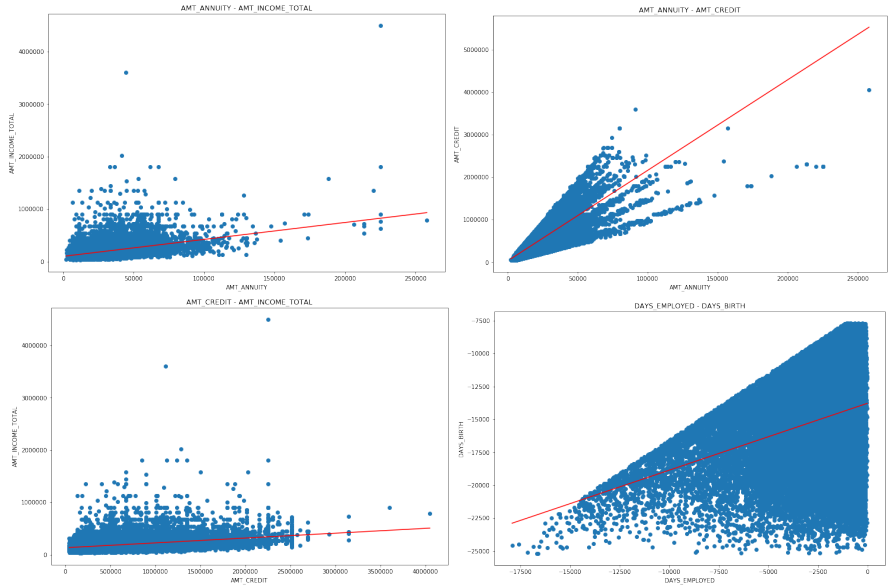
Target = 1:

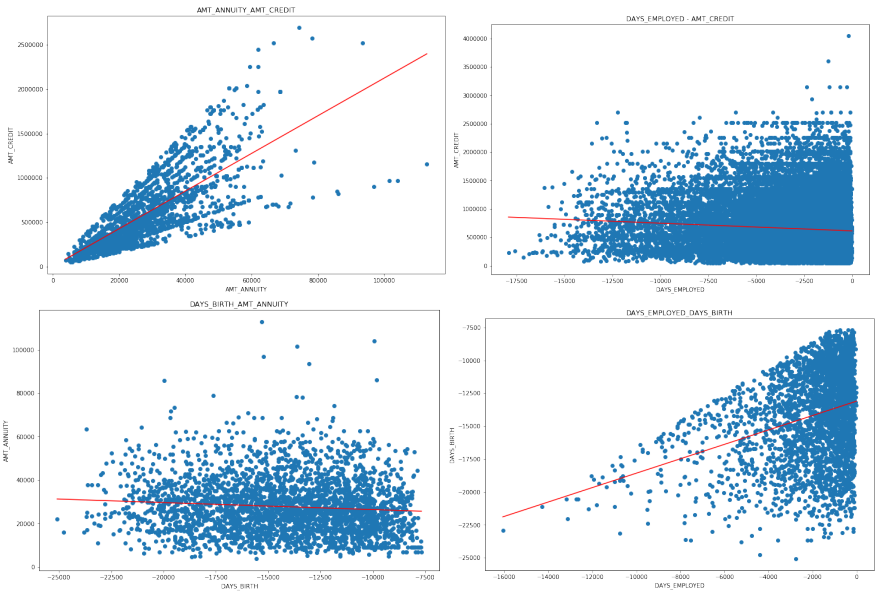


	VAR1	VAR2	CORR
13	AMT_ANNUITY	AMT_CREDIT	0.746852
12	AMT_ANNUITY	AMT_INCOME_TOTAL	0.378480
6	AMT_CREDIT	AMT_INCOME_TOTAL	0.311648
27	DAYS_EMPLOYED	DAYS_BIRTH	0.229761
19	DAYS_BIRTH	AMT_CREDIT	0.185906
25	DAYS_EMPLOYED	AMT_CREDIT	0.106430
18	DAYS_BIRTH	AMT_INCOME_TOTAL	0.093949
20	DAYS_BIRTH	AMT_ANNUITY	0.086205
26	DAYS_EMPLOYED	AMT_ANNUITY	0.063707
30	REGION_POPULATION_RELATIVE	AMT_INCOME_TOTAL	0.056996
24	DAYS_EMPLOYED	AMT_INCOME_TOTAL	0.051738
33	REGION_POPULATION_RELATIVE	DAYS_BIRTH	0.044094
31	REGION_POPULATION_RELATIVE	AMT_CREDIT	0.039900
32	REGION_POPULATION_RELATIVE	AMT_ANNUITY	0.027300
34	REGION_POPULATION_RELATIVE	DAYS_EMPLOYED	0.021424

- *Inferences: The maximum correlation is between loan credit amount and loan annuity which is pretty obvious. But the key insights here are:*
  - *Age (DAYS_BIRTH) has a better correlation to loan credit amount (AMT_CREDIT) than work experience (DAYS_EMPLOYED). This means that a person who is more in age will be preferred over work experience though the difference is not high.*
  - *Region Relative Population doesn't have much impact on other parameters.*
- *Clearly the Top 10 correlations for Target = 0 and Target = 1 slightly vary.*
  - *The Top 5 correlations are same while the order of the rest slightly differ.*

BIVARIATE DISTRIBUTIONS:





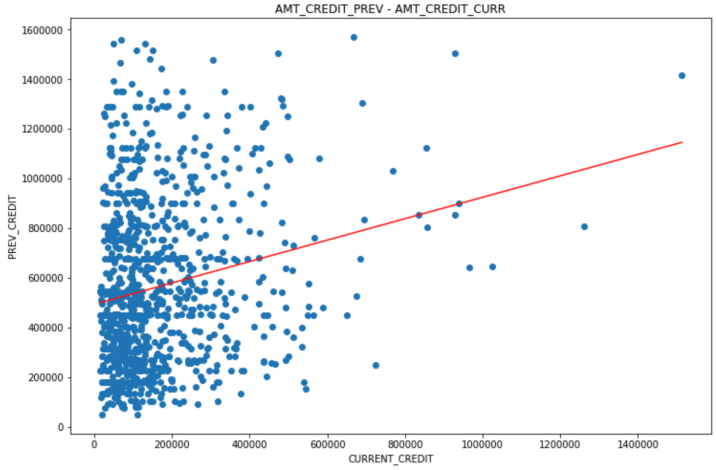
The above plots are for the bivariate analysis of some numerical columns. The results have been described in detail using correlation.

Merging the previous application data with the current applications:

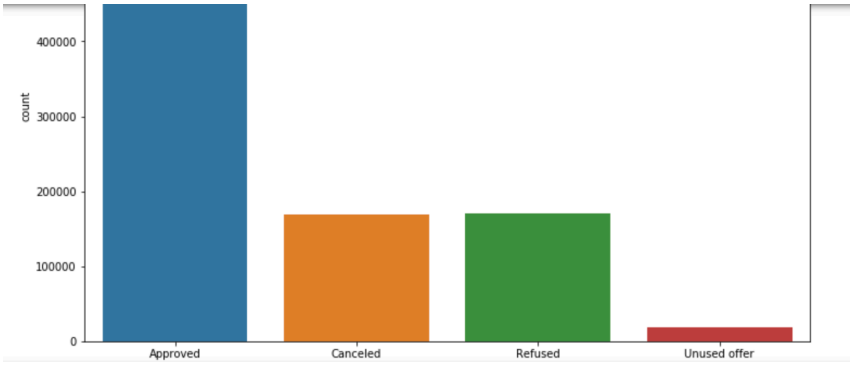
Now let us see what insights can we get from the merged dataset.

Key Insights:

- Let us compare the current loan credit amount with the average of previous loan credit amounts.



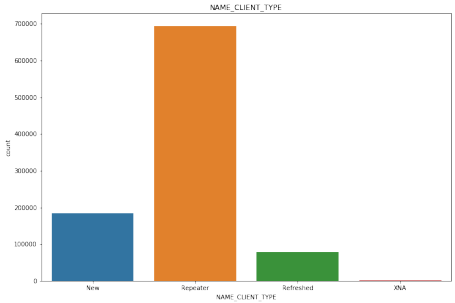
  Inference: The current credit amount is increasing at a steeper rate than the previous credit amount. This means that on average more loan is disbursed.
- The status of loans



From the above data we can derive two very important conclusions:

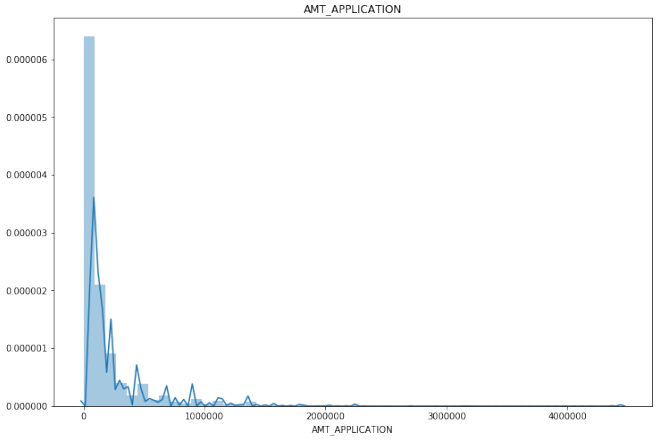
- The maximum number of approvals for a person has been 24.
- The maximum number of rejections for a person has been a whooping 68!!!

- The loan application statuses are as follows:



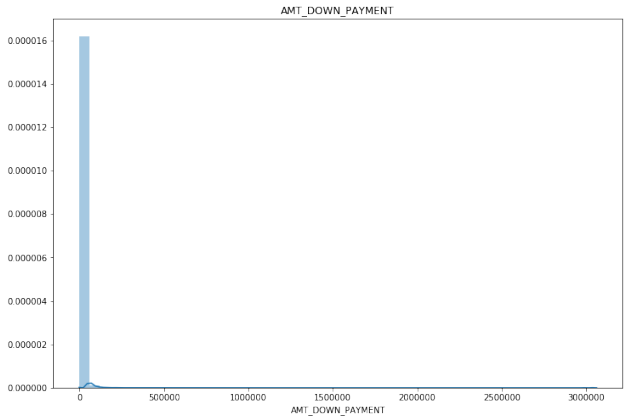
This means that most of the people are already holding a loan credit before they applied the current one.

- The credit amount that was asked on the previous loan was as follows:



As we can see here that the credit amount of previous loans was always on the lower side.

- Now let us look at the down payment of previous loans.



There was hardly any down payment made which resonates with the low credit amount.

- Now let us finally look at the loan status vs the rejection reasons.

NAME_CONTRACT_STATUS	CODE_REJECT_REASON	
Approved	XAP	599201
	XNA	8
Canceled	XAP	169216
Refused	HC	94682
	LIMIT	35478
	SCO	24532
	SCOFR	9859
	SYSTEM	475
	VERIF	2144
	XAP	1
	XNA	3060
Unused offer	CLIENT	17970

  - Out of the total refused loans, most were due to HC and LIMIT.
  - 17970 offers were still pending.
  - Nearly 6 lakh applications were approved.
