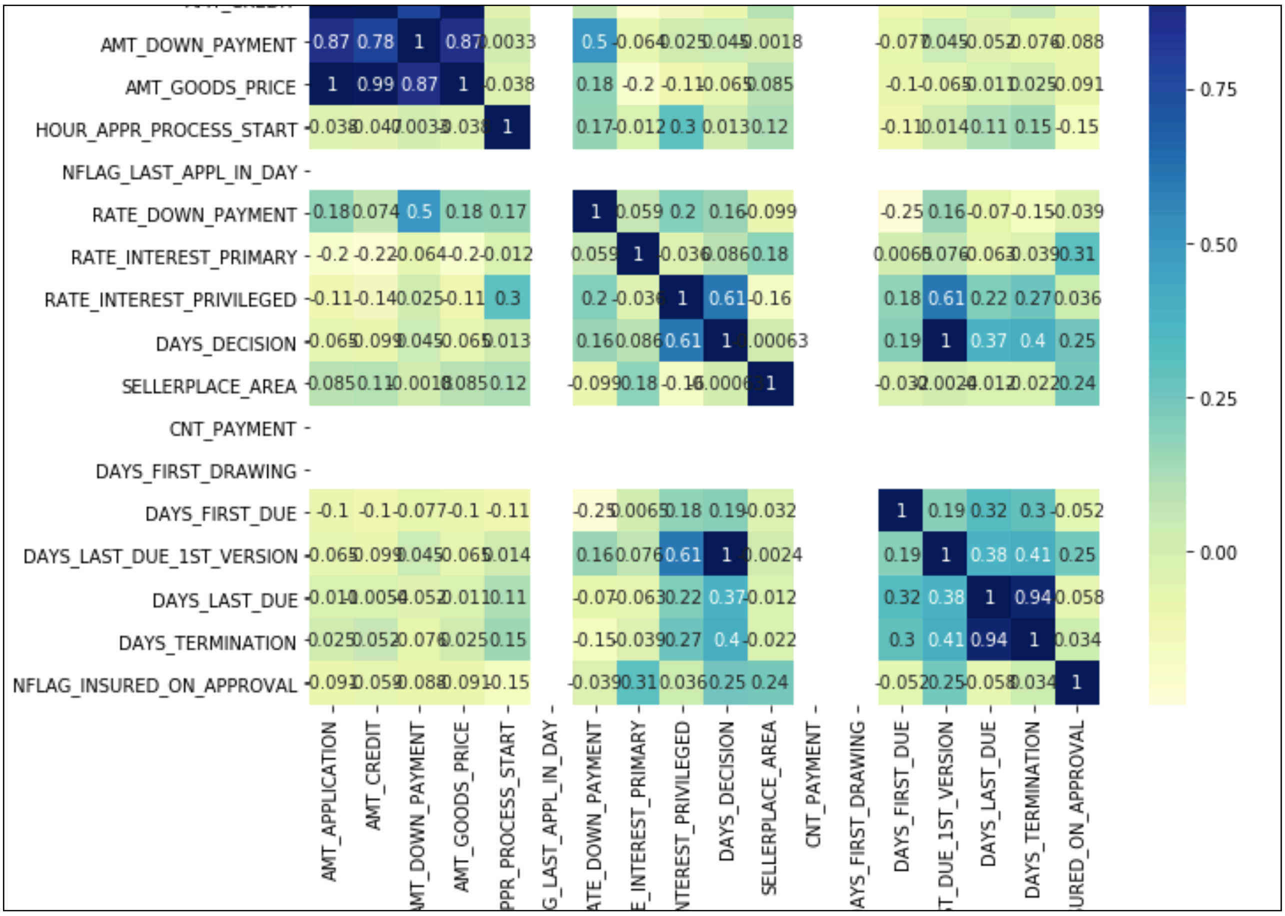
EDA Case Study

10 June 2019



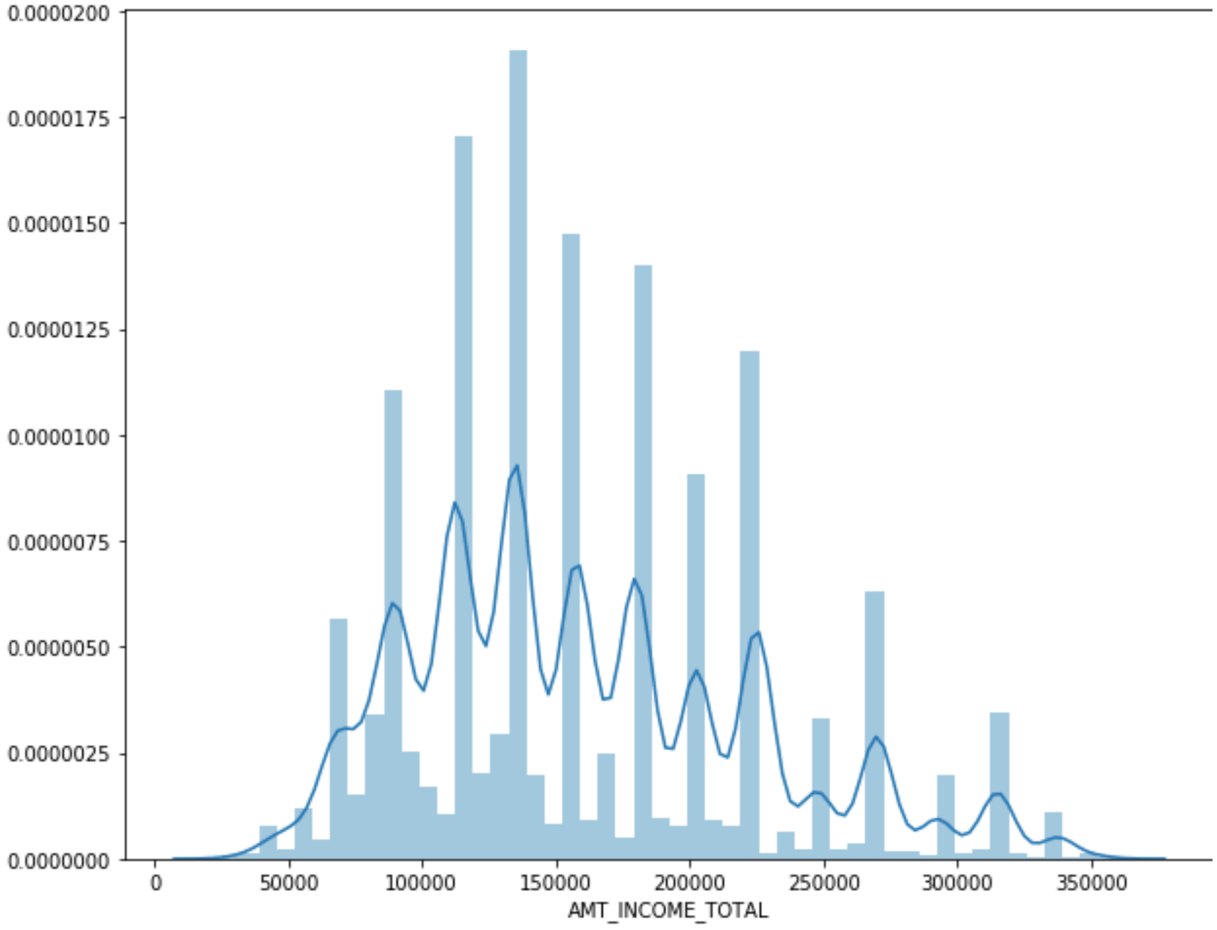
Application Data

We have application data to wherein we already know which client has defaulted on payments. Here, we are trying to find out the parameters which can tell us if an applicant will default on payment or not.

Univariate Analysis

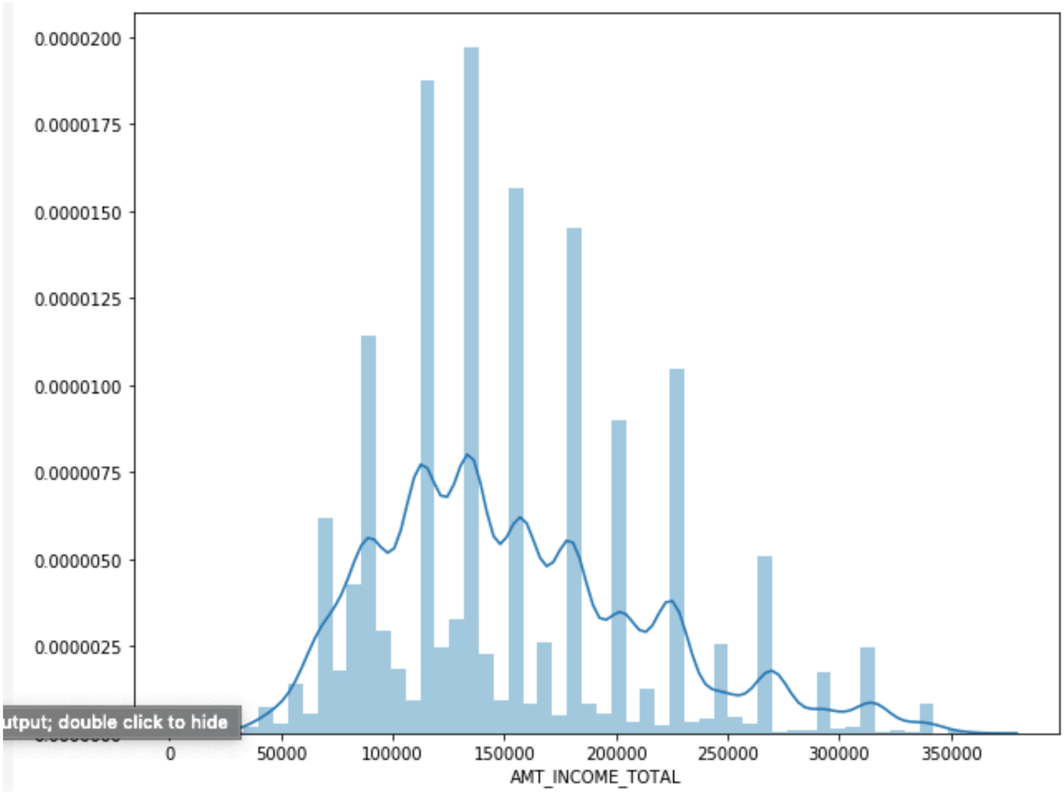
At first we try to analyse numerical variable to see if we can get some insights from the same.

Total Income



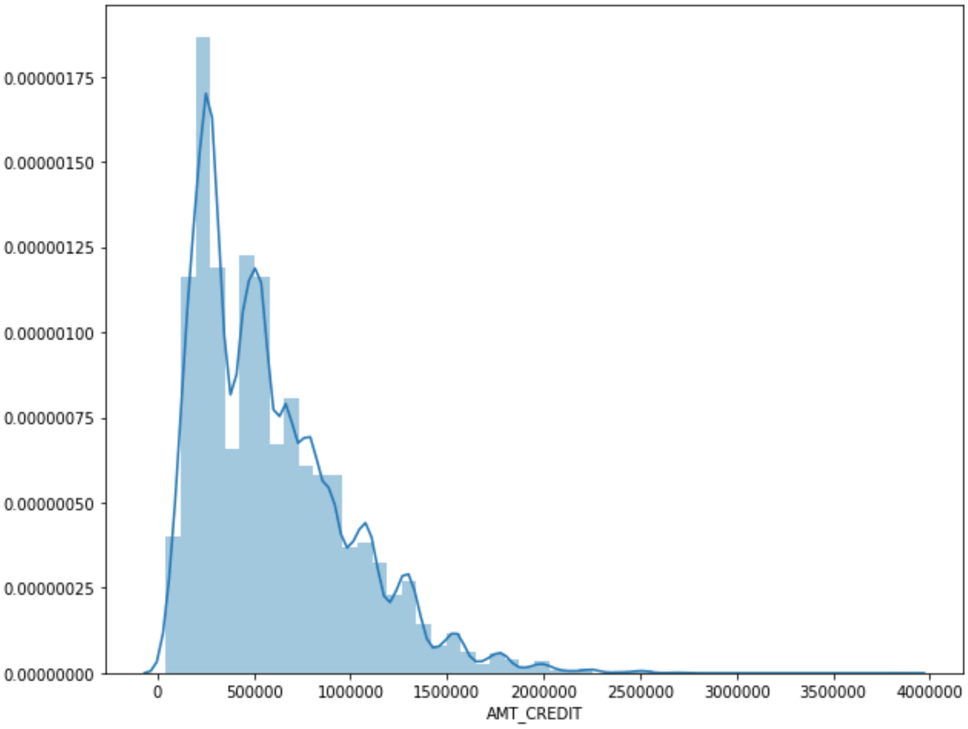
From total income we see that most of the non-defaulters have net earnings between one lakh to one lakh sixty thousand.

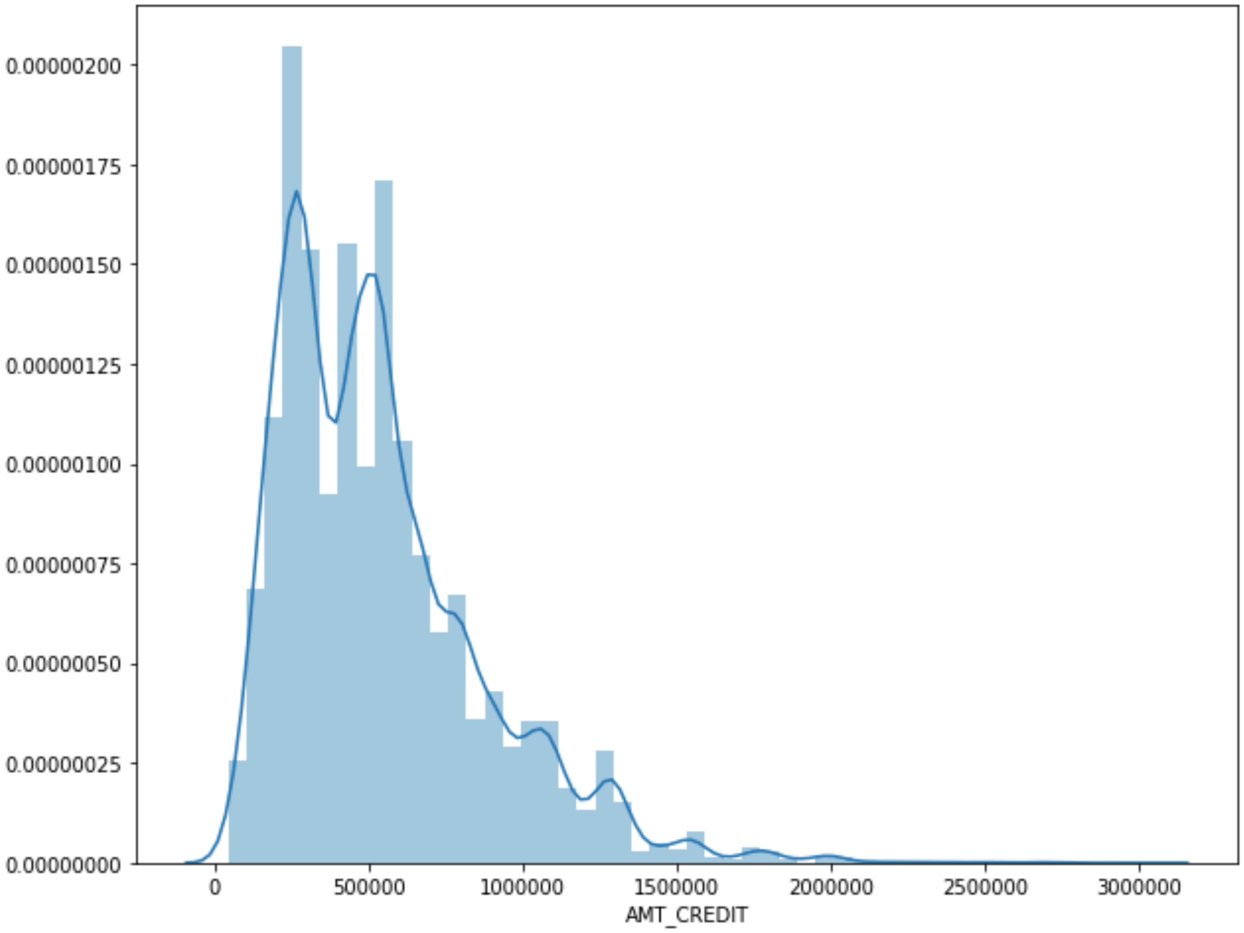
However defaulter have lesser net income as compared to non-defaulters.



Loan Amount

In this variable we see that loan amount is smaller of non-defaulters.

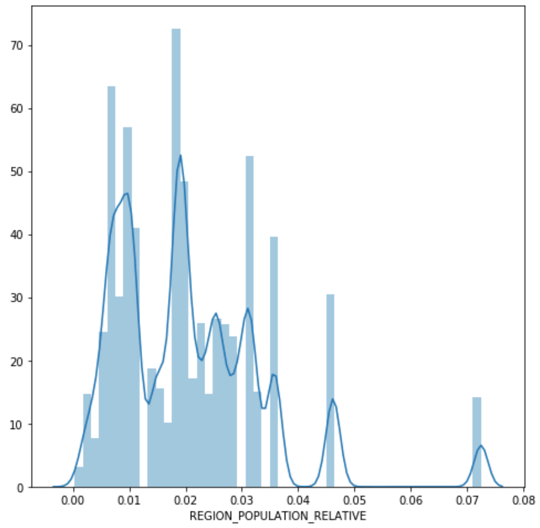


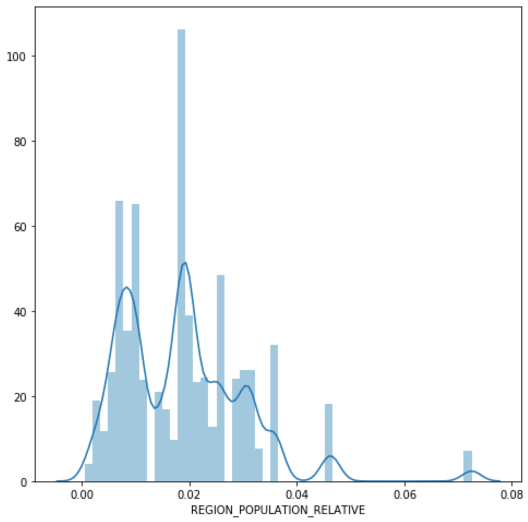


However when sam process is repeated for defaulters then, we get to see that applicants have applied for bigger amounts as loan.

Region Population

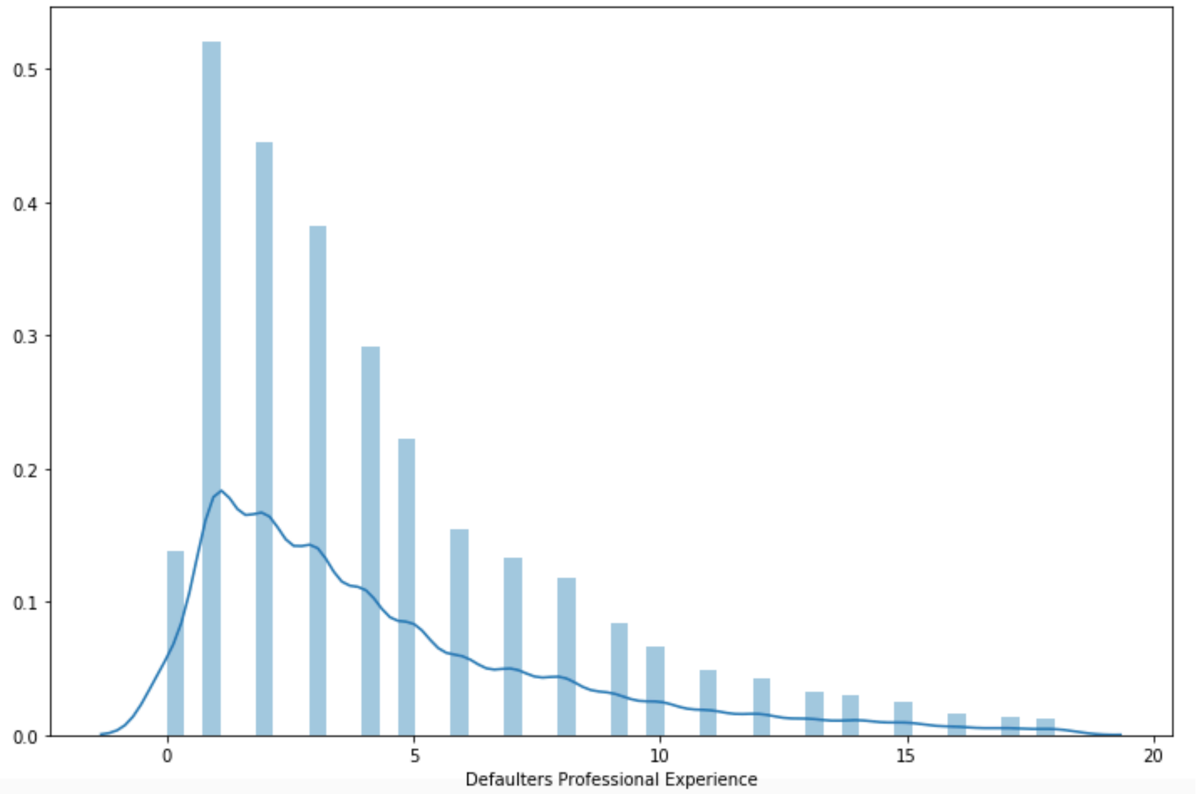
After analysis of applicants residential region's population density, we see that it does not have any variance. Both the groups, defaulters & non-defaulter have less population density.





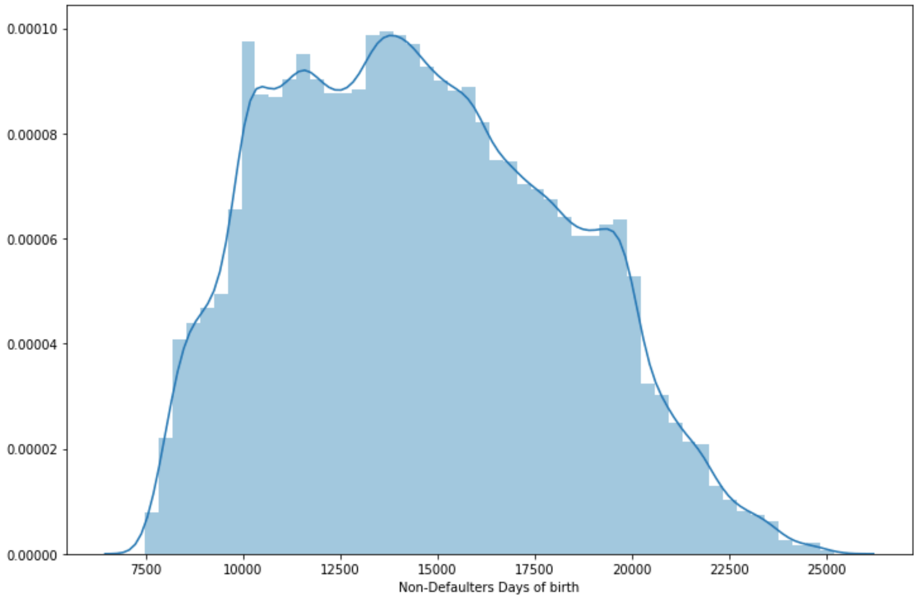
Employment Years

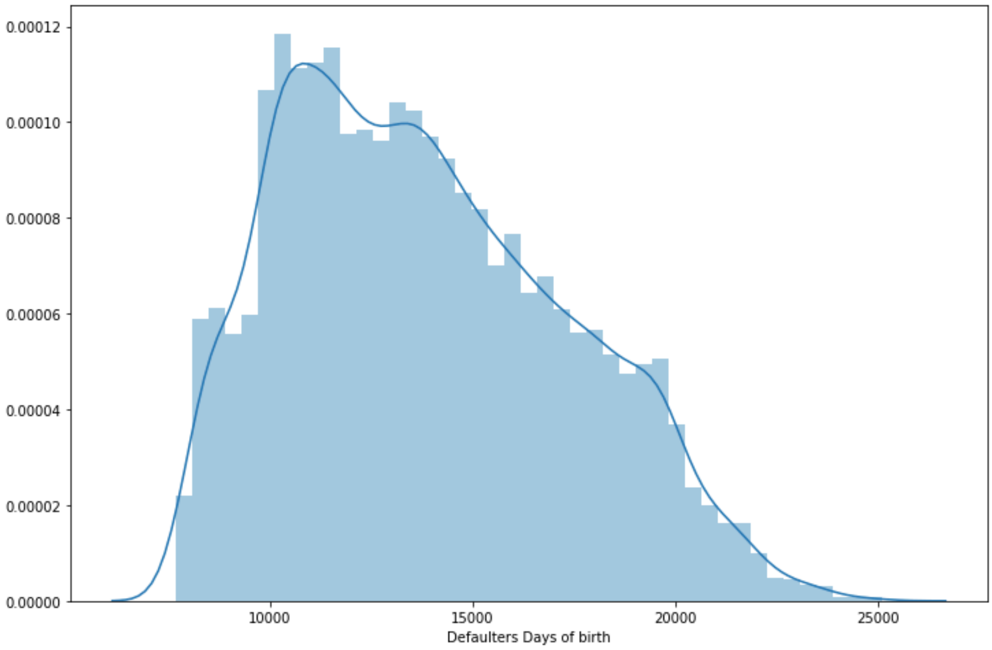
After converting employed days to years, we see that majority of defaulters are below five years of experience.



Age

Just by plotting total days from birth, we get to see that most of the defaulters are young in age.





Univariate Analysis

Now we analyse categorical variables.

Imbalance

First of all, we found the imbalance percentage and ratio, which came out as 10.16% and 0.1:1 respectively.

Then we analysed few categorical variables - gender, has own car, has own realty, highest education of applicant. In here, we found that, females take more loans as compared to males. However they tend to default less than males. In our analysis, we found that 8.1% of females have defaulted in comparison of 10.99% to males.

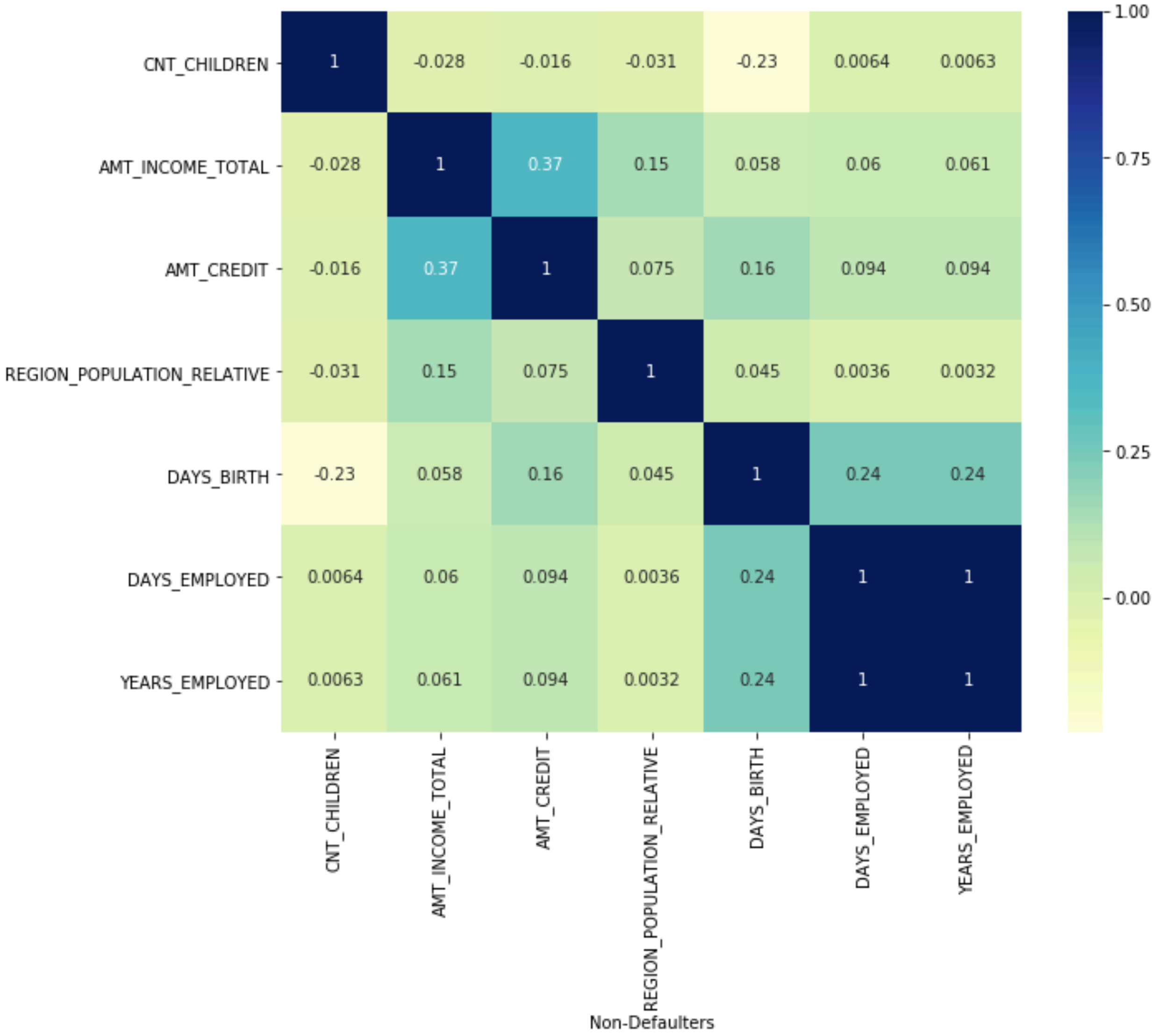
Same analysis cam out for applicants having a car or not. Those applicants who owned a car defaulted less than those who didn't. In our comparison we found 7.96% of defaulters had a car whereas 9.97% didn't.

However when we compared realty, it was constant for both the groups who had realty at 9.19% and who didn't at 9.28% with a very marginal difference.

Education was expected to give different result though. Applicants with highest education defaulted less and it increased when education level of applicant dipped. So we can say that higher the educational qualifications of applicant, the lesser chance we have of defaulting.

Correlation

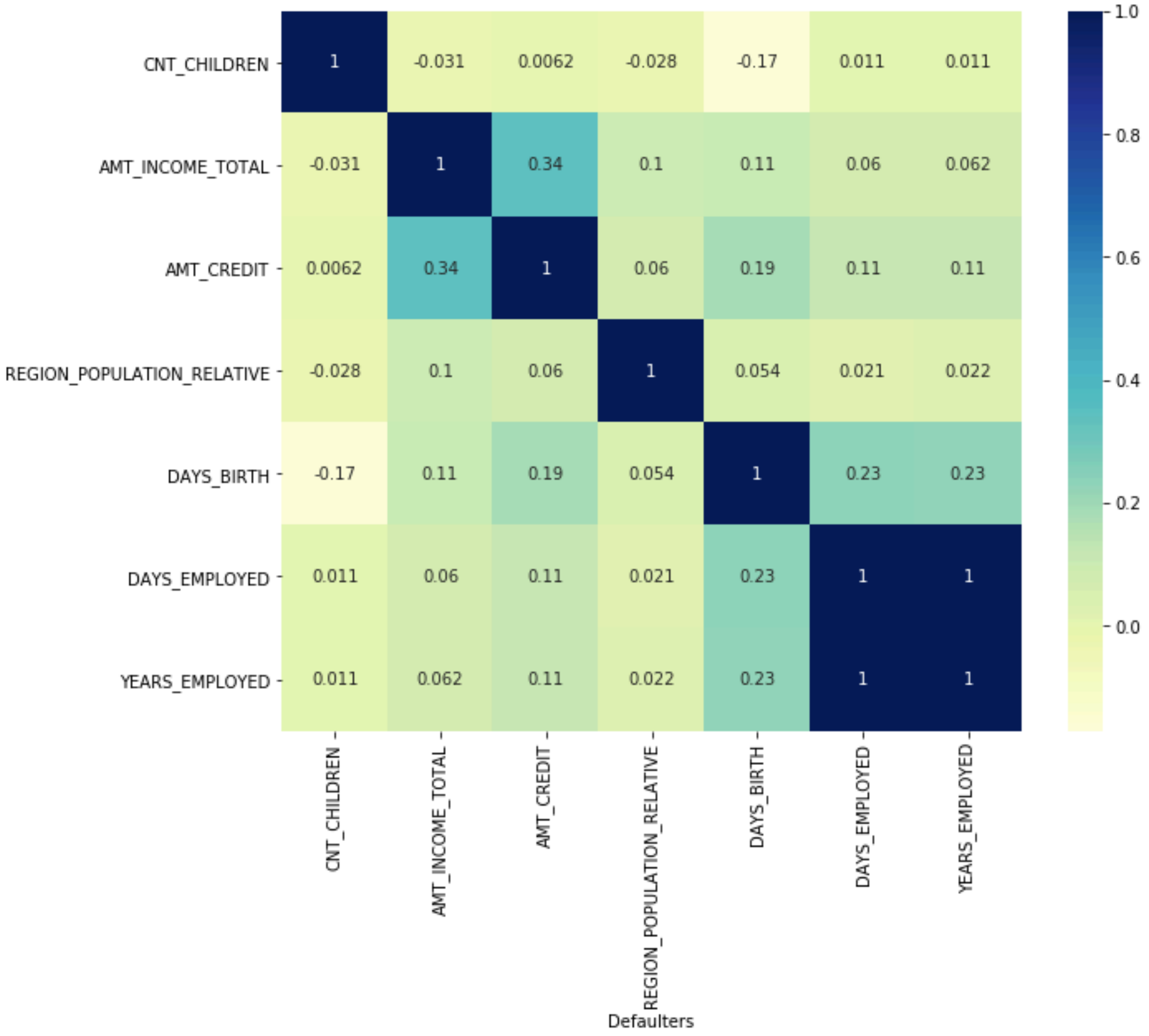
Now we found out the correlation for both the groups - defaulters & non-defaulters.



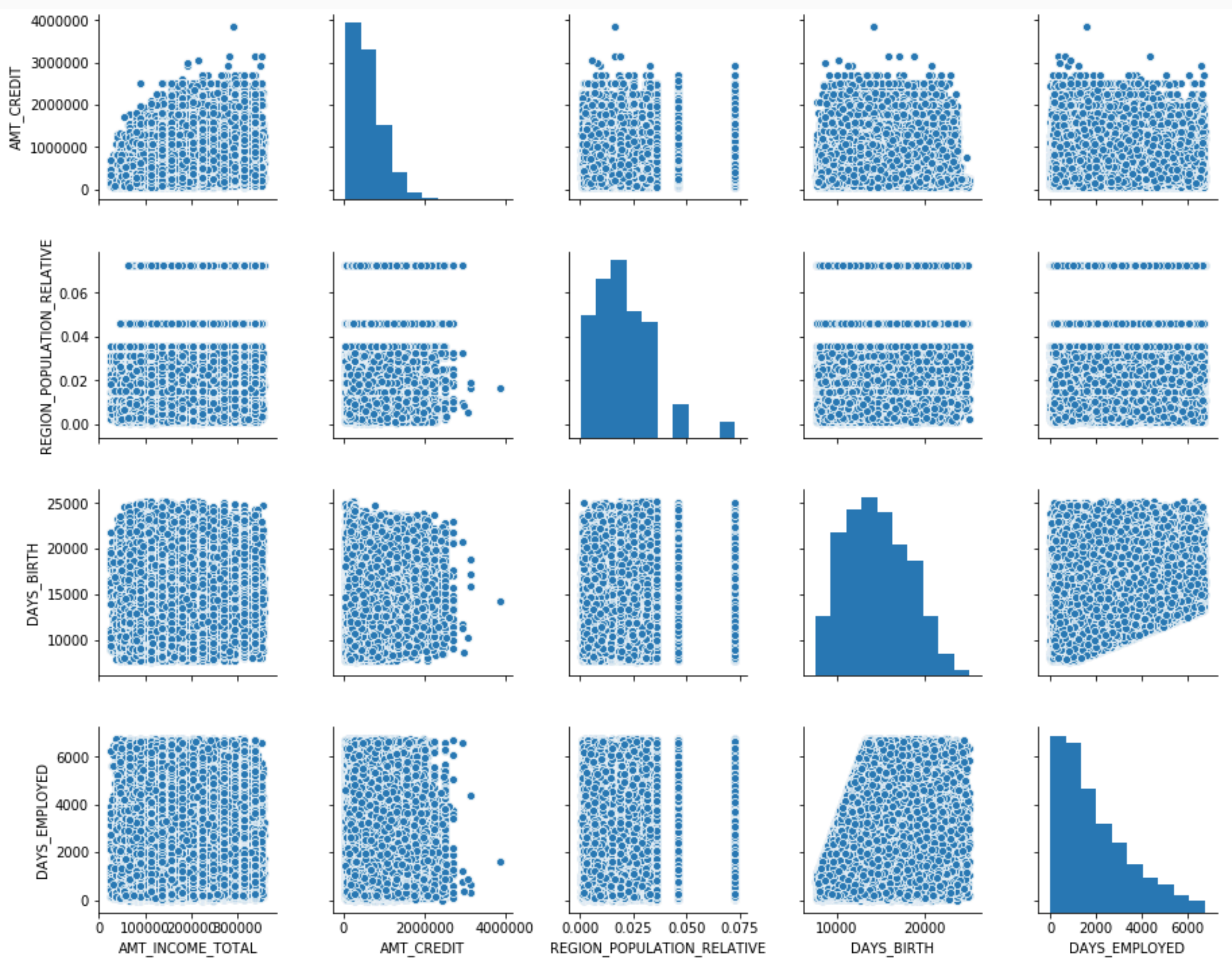
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	YEARS_EMPLOYED
CNT_CHILDREN	1	-0.028	-0.016	-0.031	-0.23	0.0064	0.0063
AMT_INCOME_TOTAL	-0.028	1	0.37	0.15	0.058	0.06	0.061
AMT_CREDIT	-0.016	0.37	1	0.075	0.16	0.094	0.094
REGION_POPULATION_RELATIVE	-0.031	0.15	0.075	1	0.045	0.0036	0.0032
DAYS_BIRTH	-0.23	0.058	0.16	0.045	1	0.24	0.24
DAYS_EMPLOYED	0.0064	0.06	0.094	0.0036	0.24	1	1
YEARS_EMPLOYED	0.0063	0.061	0.094	0.0032	0.24	1	1

From the above correlation for non-defaulters, we observe that credit amount & total income are highly correlated at 0.37 which is followed by professional experience & age at 0.24. Hence these are the critical parameters to examine before giving a loan.

When same exercise was repeated for defaulters, to our expectation the same parameters were highly correlated. However there was a marginal difference, where loan amount & total income was at 0.34 and professional experience & age was at 0.23.



Bivariate Analysis

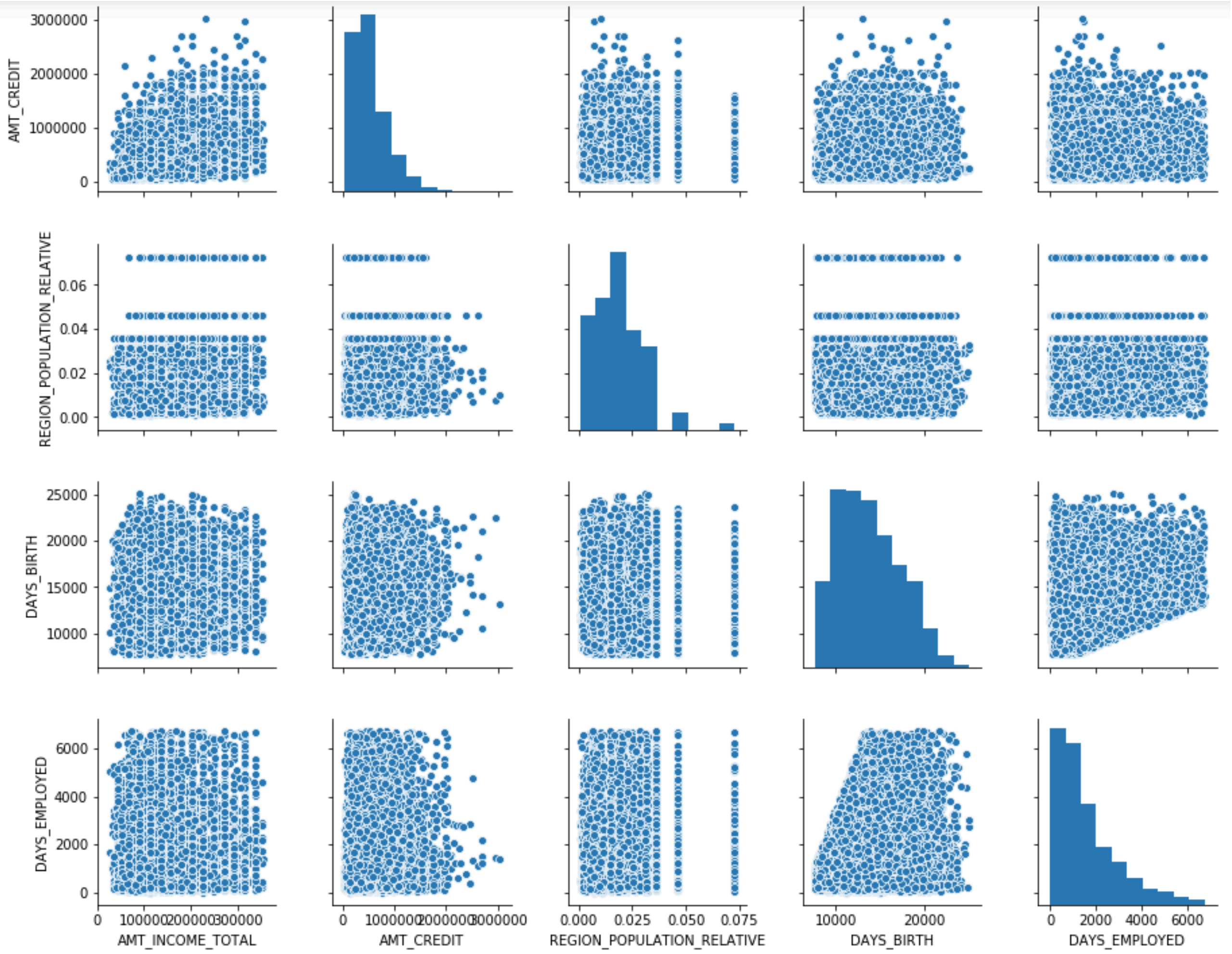


With bivariate analysis of non-defaulters, we observed-

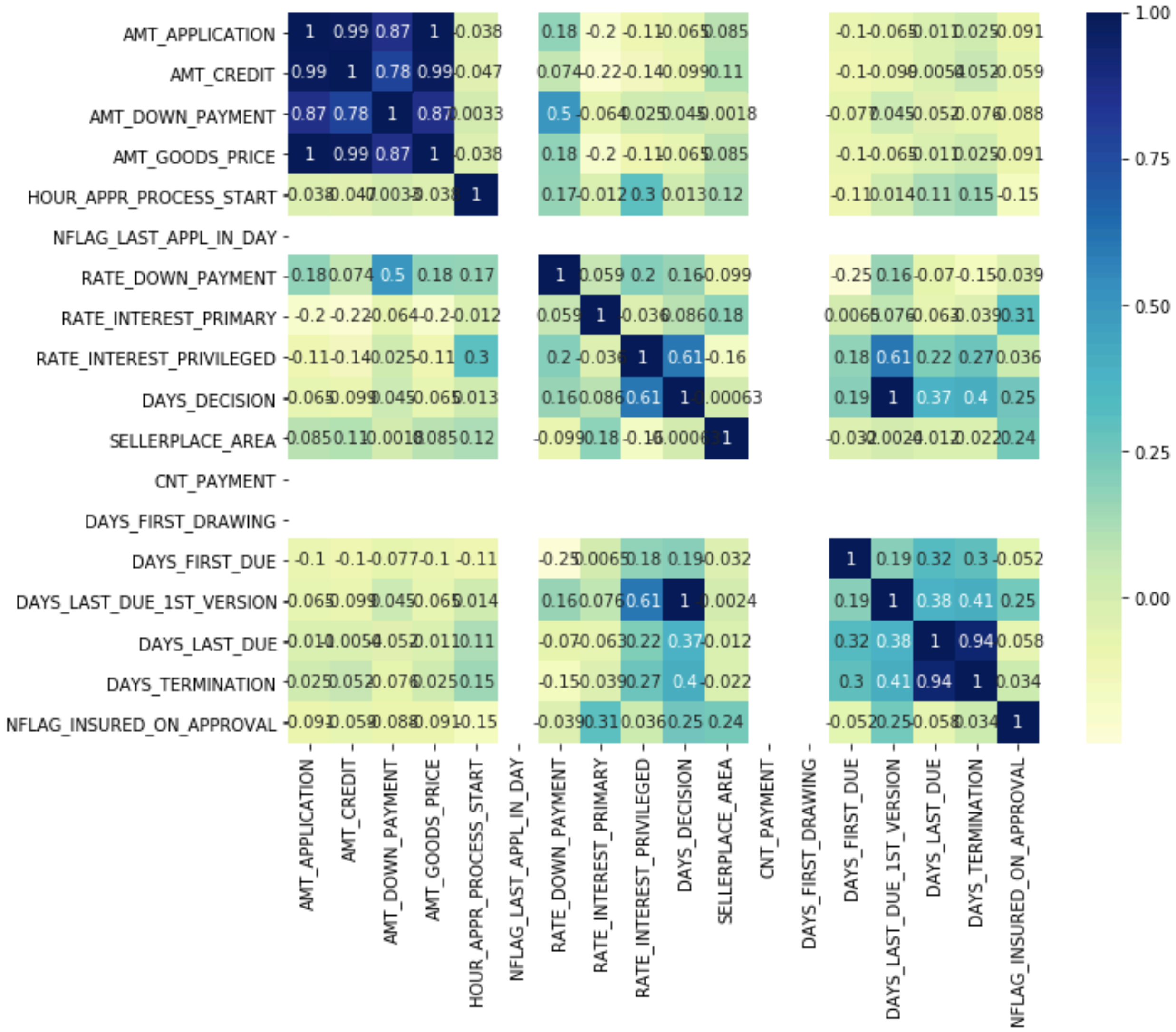
- Loan amount increases with the increase of net income.
- Most applicants are from lesser populated areas.
- We have equal distribution of applicants from all age group and professional experience.

However when same exercise was repeated for defaulters, we found that-

- Loan amount is equal or higher than net income.
- Applicants still belonged to lesser populated areas.
- Applicants were young in age.
- Applicants had less professional experience, less than five years.



Correlation



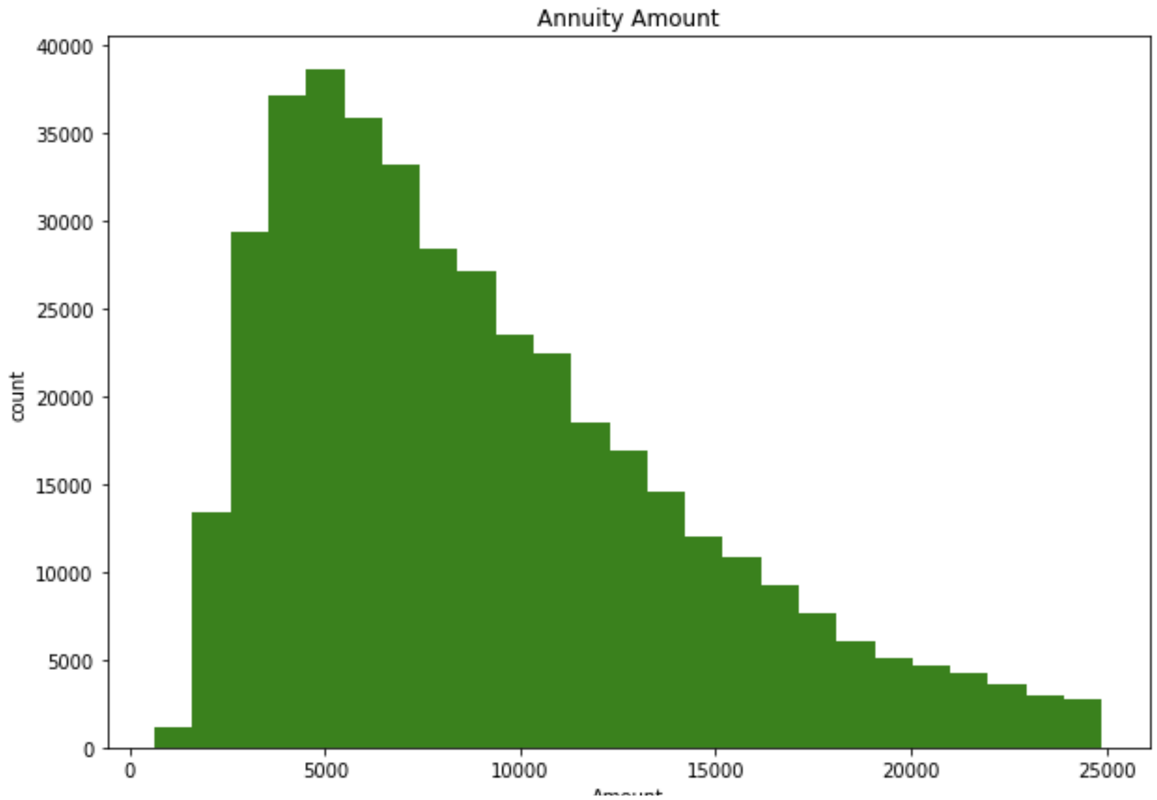
When we compared the correlations of defaulters, non-defaulters and applicants from previous data, we found-

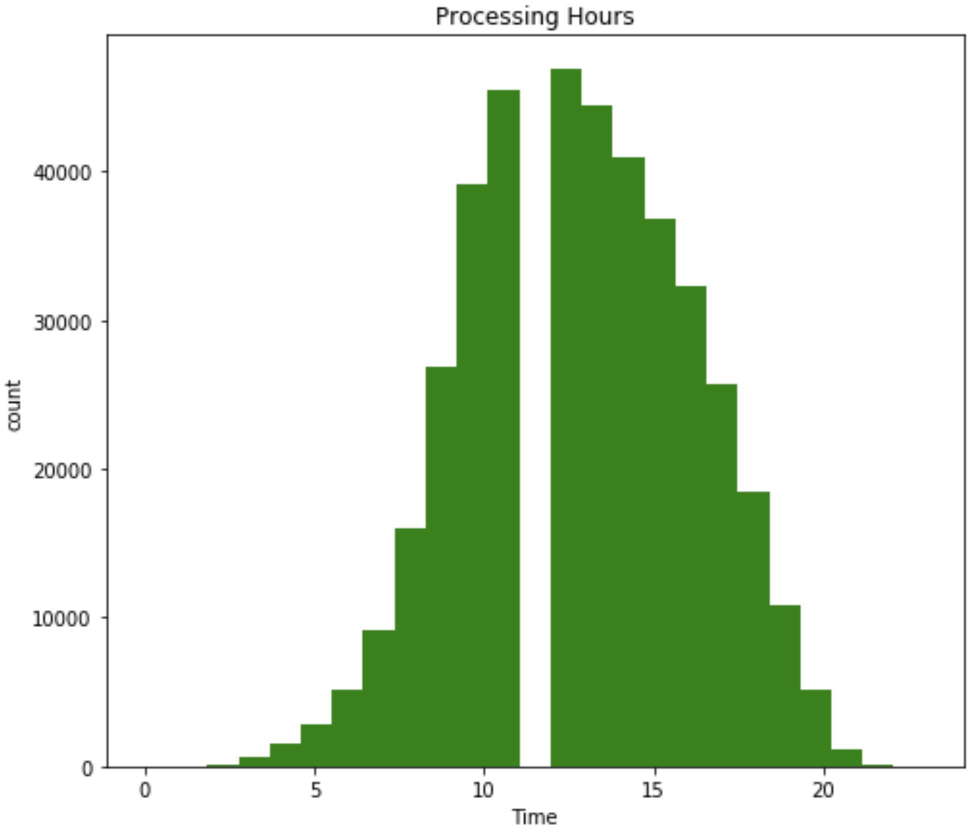
- Credit amount & net income is the highly correlated columns and it is the same for previous data.
- DAYS_LAST_DUE & DAYS_TERMINATION are also highly correlated.

Since the other considered columns are not available, so we couldn't find more relatable correlations.

Previous Data

We also performed univariate analysis on previous data, which stated that most of the population tends to make an annuity amount of less than ten thousand per month.





Through processing hours, we identified that client generally applies for loan in the afternoon, as the plot was skewed on the left.