**Pre-Processing Steps :**

Script Name:Loan-Data-to-Data-Frame.ipynb

1. The sample data from Single family dataset is downloaded and unzipped.
2. Script reads data year wise and converts sample\_orig\_{year} and sample\_svcg\_{year} into Data frames with appropriate data types based on field type. A join is performed on both files by LOAN\_SEQUENCE\_NUMBER.
3. The final merged data-frame is written in processed file folder and saved as summary\_{year} csv file.

Post-Processing Data Analysis year wise

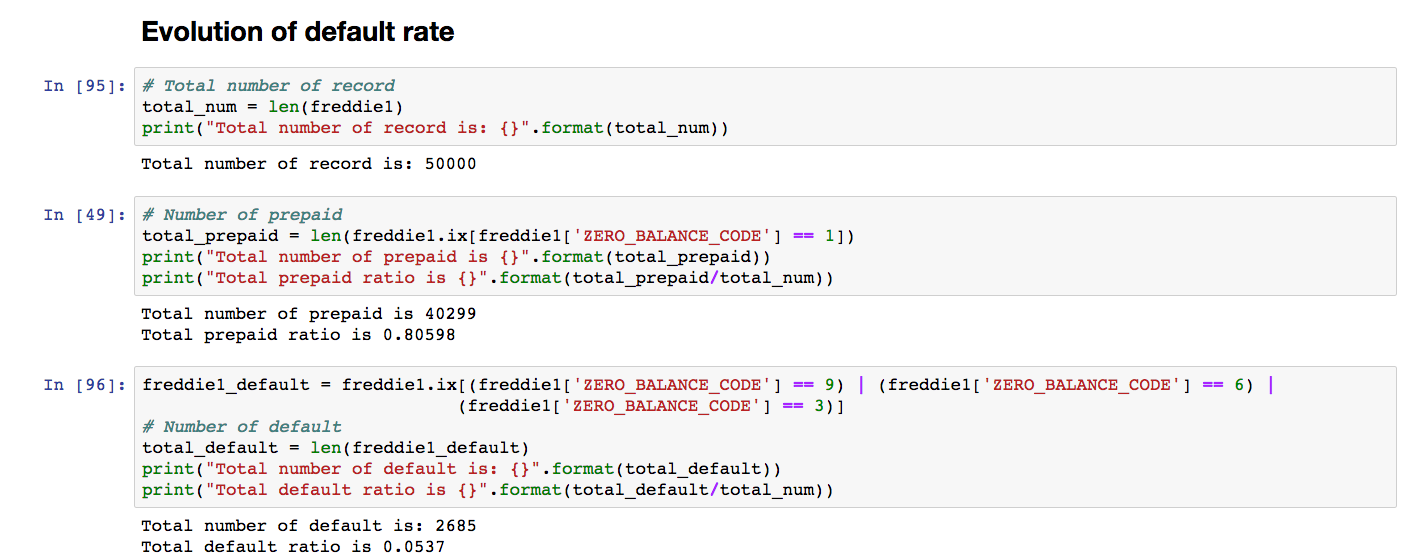
Script Name:Exploratory\_Data\_Analysis\_Loan\_Dataset.ipynb

1. The script takes year as input and performs summarization on various fields.
2. For illustration purpose I am including findings of year 2005

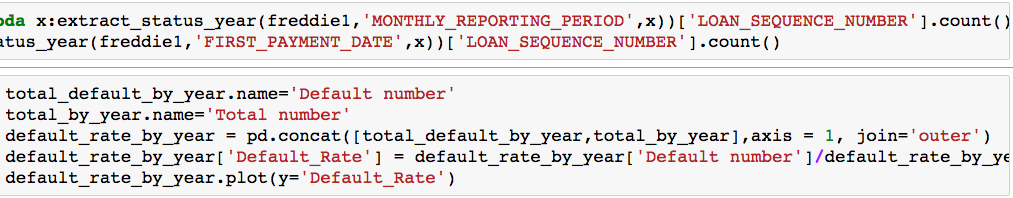
**Analysis**

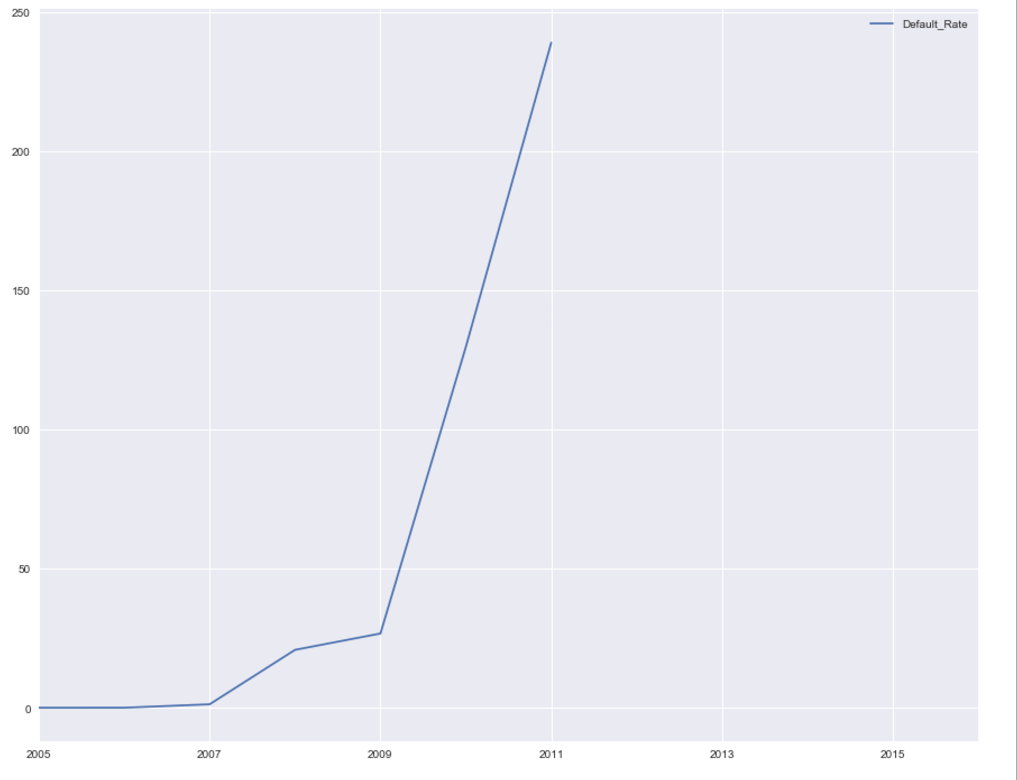
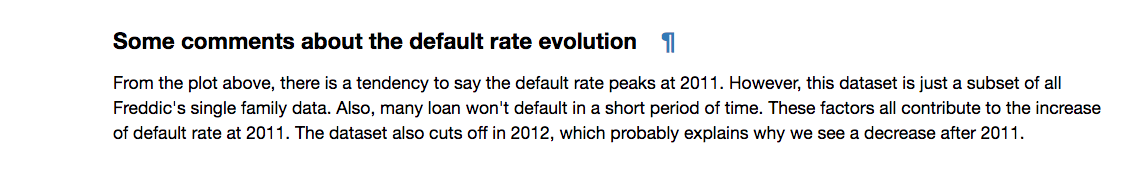
**The ratio of prepaid loans in year 2005 was .805**

**The ratio of default loans in year 2005 was 0.0537**

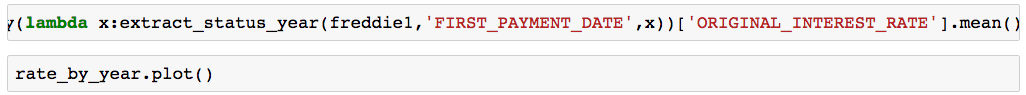


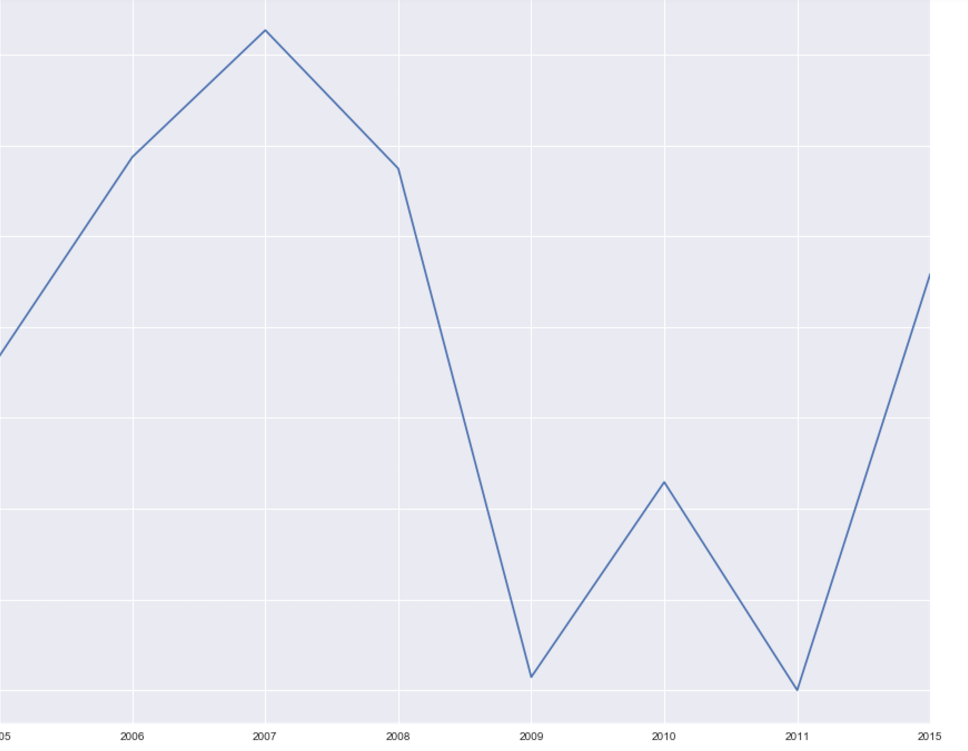
**Default rate evolution over years (computed using monthly reporting period and loan origination)**





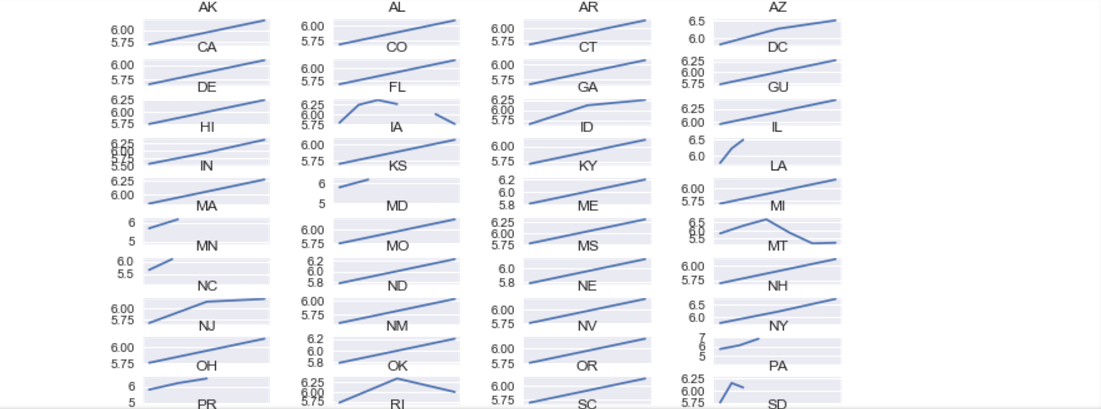
**The evolution of rate of interest by the year (Computed by first payment date and original interest rate)**





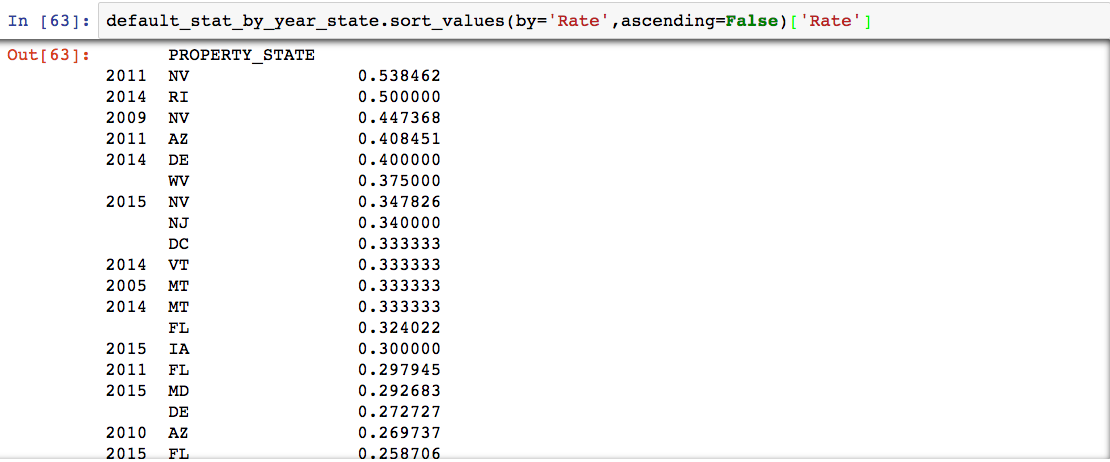
**The interest rate spiked during economic depression (2007-2008) due to subprime credit crisis and reduced over the years.**

**Evolution of interest rate state wise**



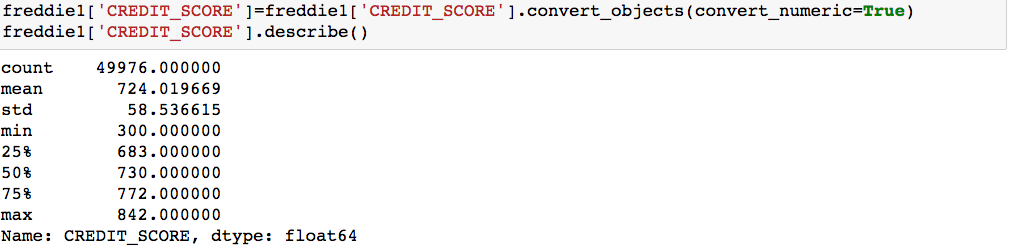
**Default cases in each state each year**



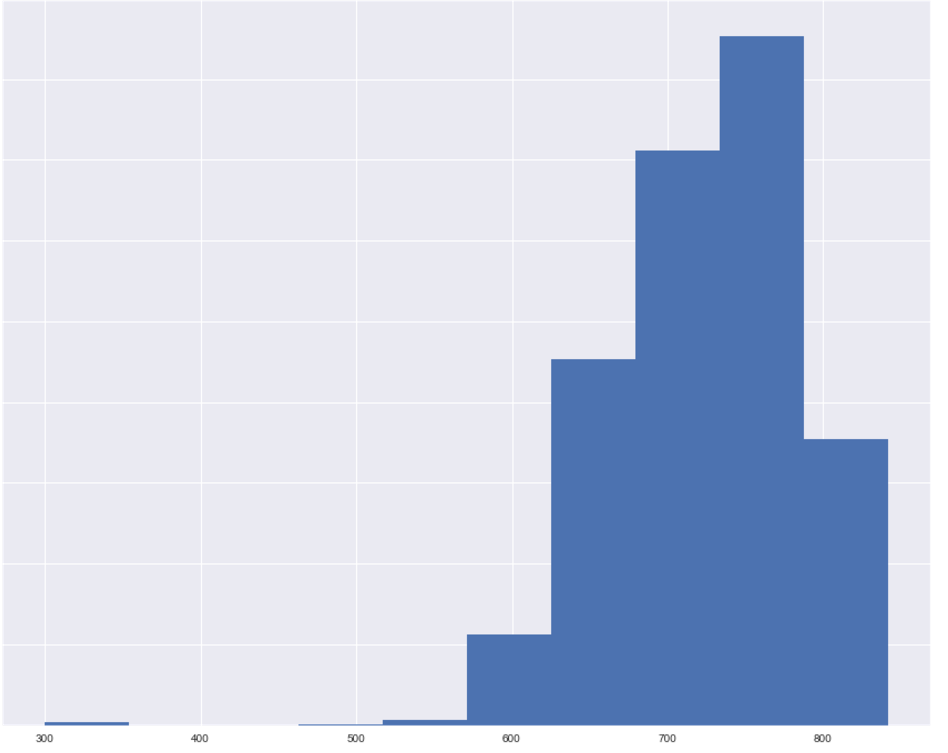


**Default cases by Credit Score**

**The average credit score of all applications is 724. Suprising. Then average potential homeowner has a \*\*excellent credit score\*\* .If you have credit score less than 700 then you are already at only 25% percentile!**



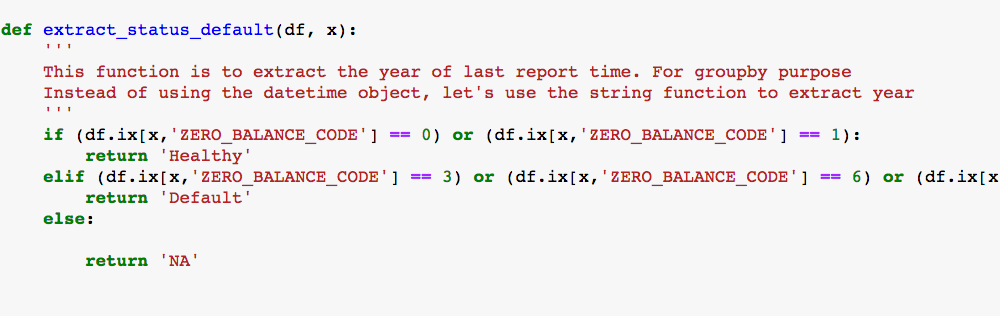
**Histogram of credit score**

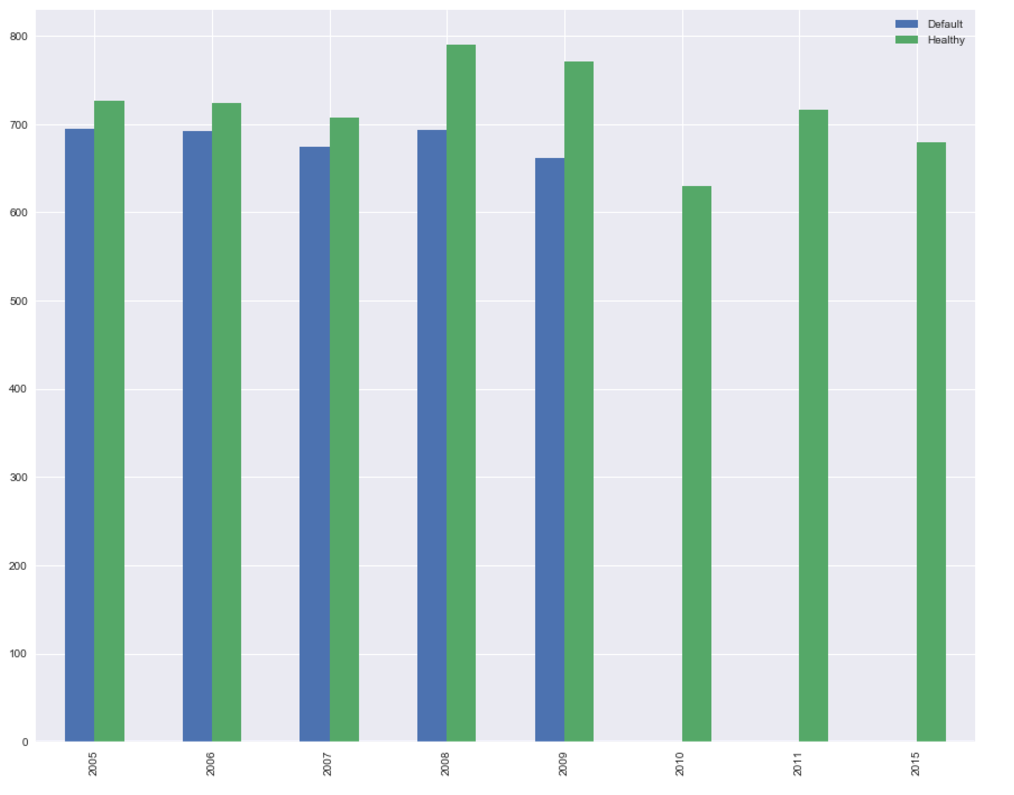


**Creating set based on Zero\_balance\_code**

**Zero\_balance\_code with 0 and 1 are healthy**

**Zero\_balance\_code with 3,6 and 9 are default**



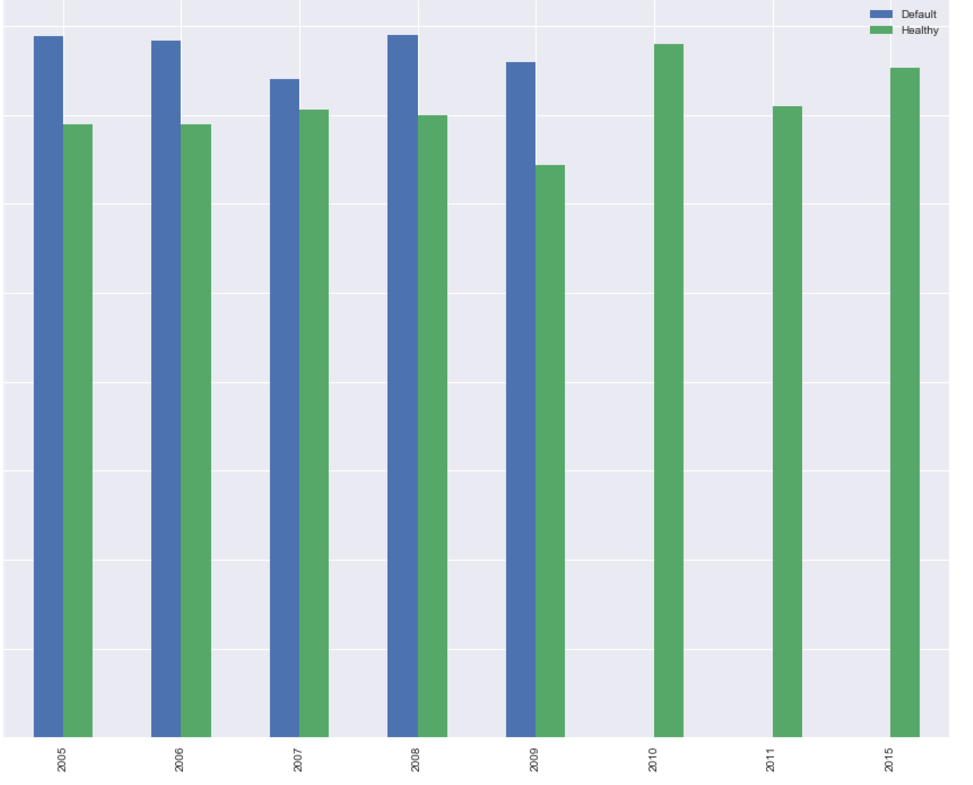


**Loan Status and Its credit score**

This graph shows some interesting findings:

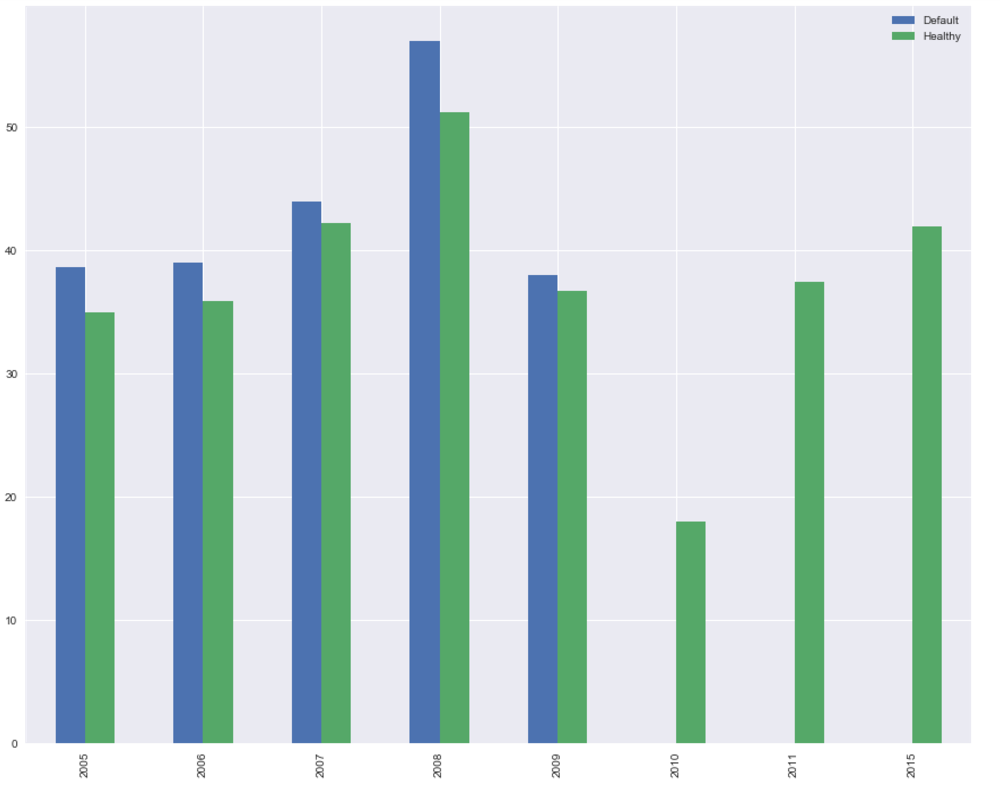
1. The overall trend of applicant's credit score is increasing by year.

2. Default cases have lower average credit score compared to non-default cases. Therefore, credit score should be considered as one feature in the following machine learning model**.**

**Initial Loan-to-Value ratio by status**

### **Original Loan to value ratio (OLTV)**

**Another interesting plot. Clearly, healthy loan has OLTVs around 70%, while default cases are at 80%. This makes sense too. As higher OLTV means you borrow more money to buy the house, which makes the mortgage riskier.**

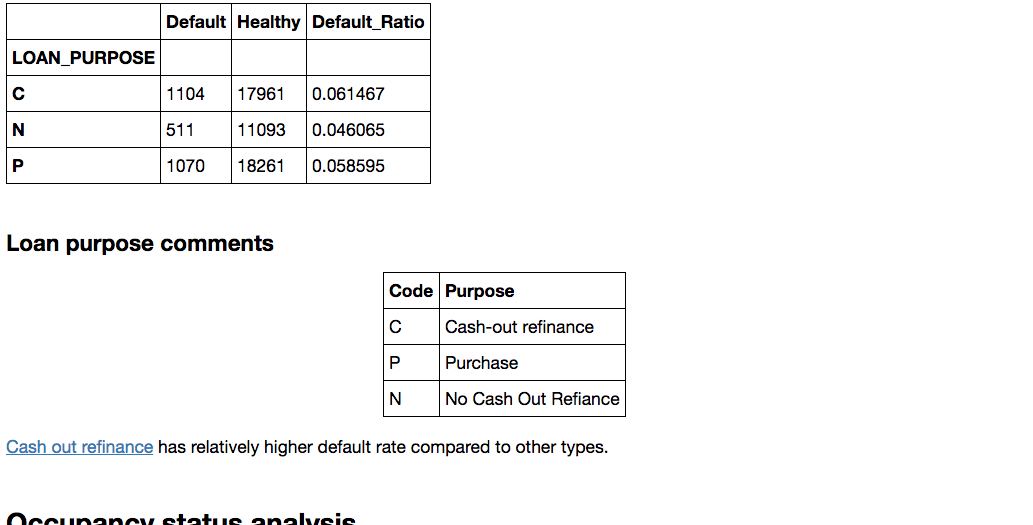
**Debt-to-Income ratio**

### 

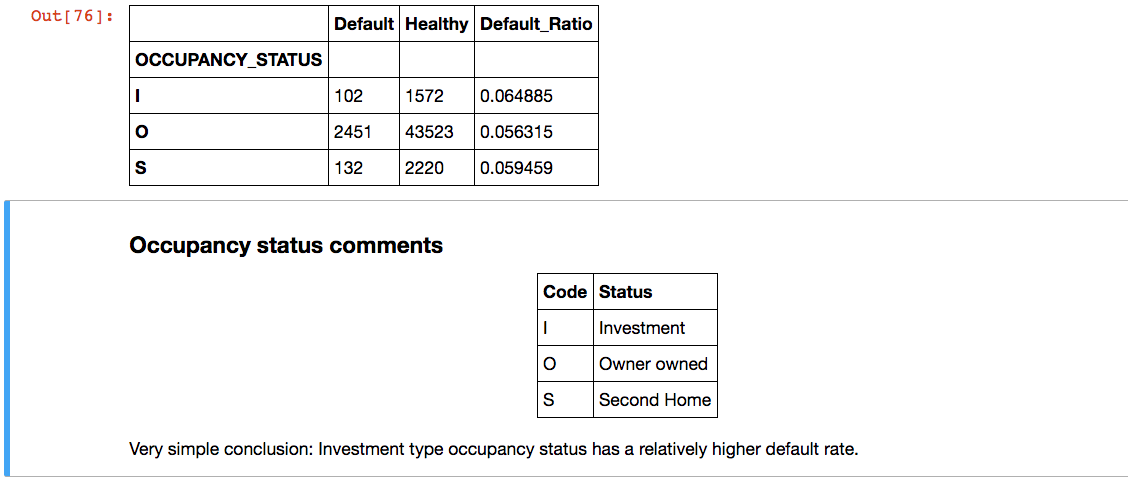
There are a lot of websites out there to tell you how much you can afford to buy a house. Like [this one](http://www.realtor.com/mortgage/tools/affordability-calculator/). The conservative approach uses ~37% as DTI and the aggressive uses 41%. Analysis of Freddic's data kind of supports this theory. Default cases in general has higher DTI compared to those non-default cases. Also, before the sub-prime crisis, you can get a loan with DTI higher than 40%. Now, the general trend of DTI becomes more and more conservative. DTI rates are now between 30-35%. So next time when you use the online house affordability calculator.

**Loan Purpose analysis**

**Cash out finance has higher default rate**

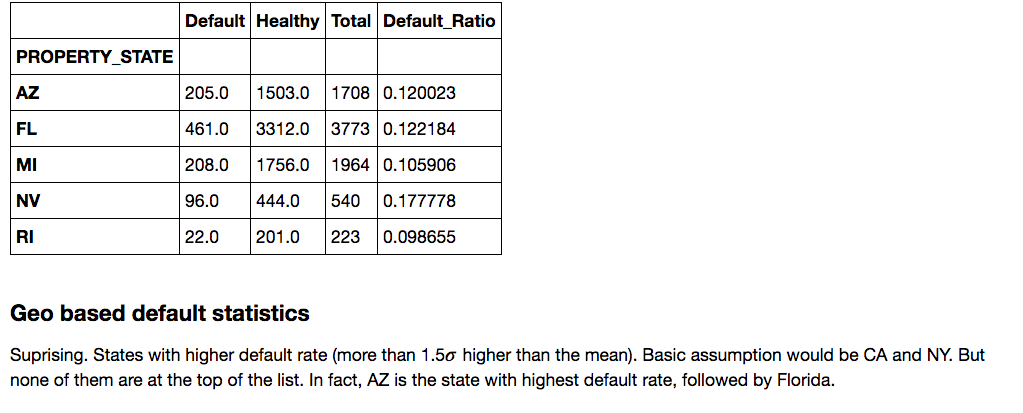


**Occupancy status analysis**



**Geo-Location analysis**

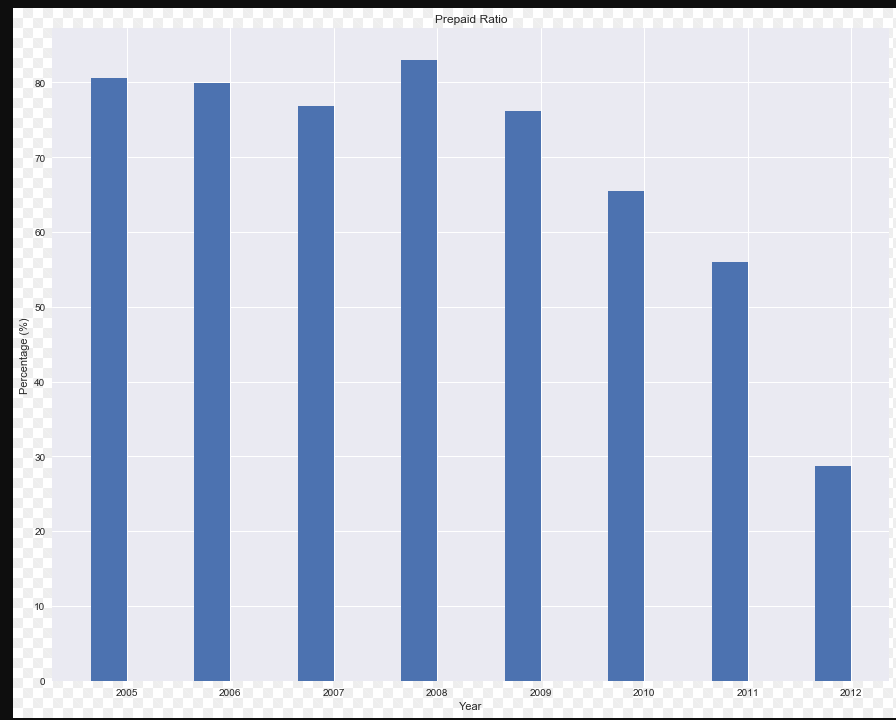
**### States have higher than mean + 1.5 sigma default rate?**



**Aggregated Loan Data Analysis over years 2005-2012**

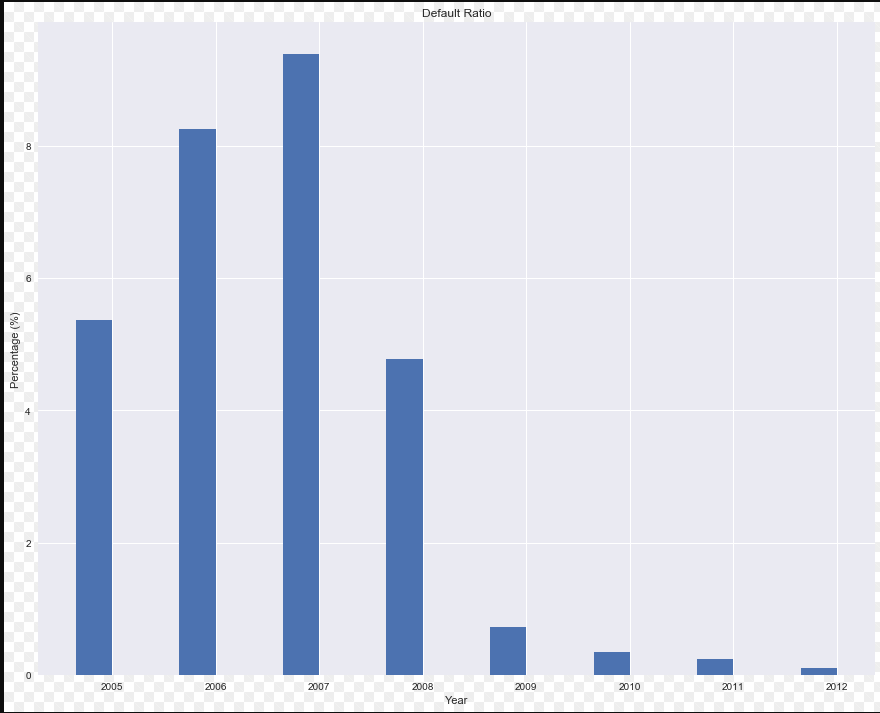
**### How many mortgages have been prepaid in these years?**

#The prepaid ratio is lower in 2007 due to economic depression. #https://en.wikipedia.org/wiki/Financial\_crisis\_of\_2007%E2%80%932008



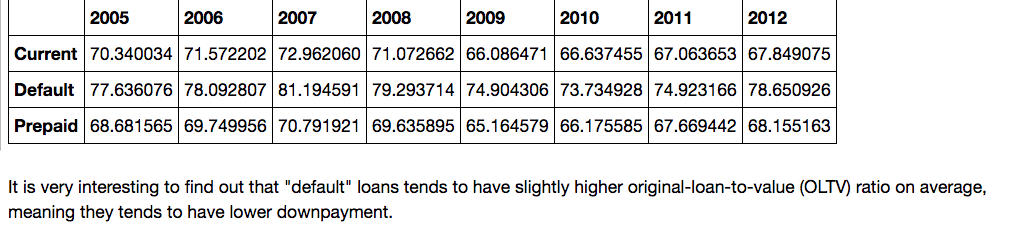
**Default Rate**

Highest during depression then reduces as economy is doing well

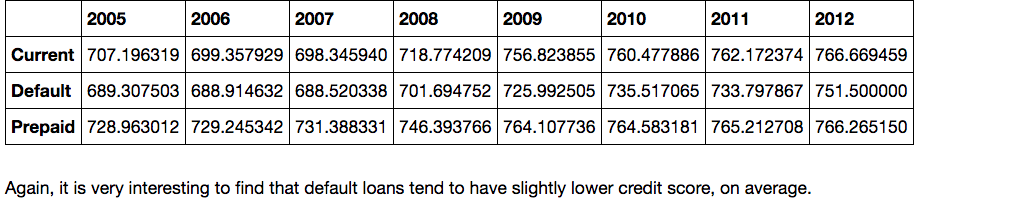


Default rate is very high in 2007 due to economic recession due to subprime lending. More details on <https://en.wikipedia.org/wiki/United_States_housing_bubble>

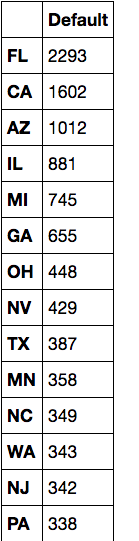
**OLTV for default loan**



**Debt to income ratio for Default loans**



**Default over years in states**



**#FL and CA tend to have more default cases over all these years**