**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

All the features missing values, spaces and NaN have been replaced with suitable substitutions. Below are a few features and their substitutions:

MSA: Replace missing values with “*Neither MSA nor MD/Unknown*”

ORIGINAL\_DEBT\_TO\_INCOME\_RATIO: Spaces replaced *with > 65*

ORIGINAL\_INTEREST\_RATE: Replace missing values with *mean*

SUPER\_CONFORMING\_FLAG: Replace missing value with “*No*”

PRODUCT\_TYPE: Replace missing value with “*FRM*”

**Analysis**

The dataset contains several interesting features which can be used to derived additional information about the borrower.

**Use case 1:**

1. Categorized user’s based on their credit score.



We considered 650 as the threshold, all the user’s below 650 credit score were labelled as “Good Credit” and others were labelled as “Moderate Credit”

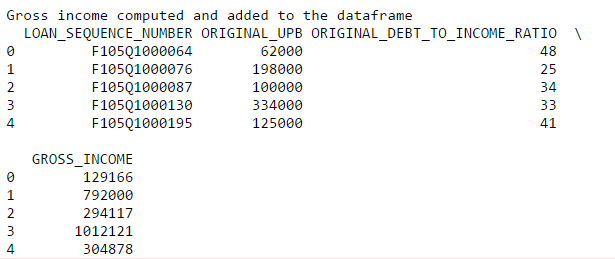
1. Computed the borrower’s gross income. Since the gross income is a key aspect in determining the loan amount and interest rate, it was crucial to compute this field. We calculated this using the formula:

GROSS\_INCOME= UNPAID PREPAID BALANCE

------------------------------------------------------ x 100

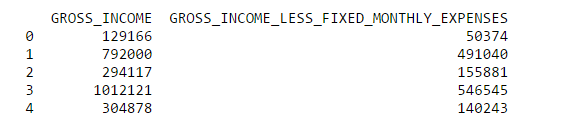
ORIGINAL\_DEBT\_TO\_INCOME\_RATIO

i.e sanctioned loan amount / total debt amount \*100



1. To compute the gross\_income\_less\_fixed\_monthly\_expenses we deducted the GROSS\_INCOME and the UPB and we also deducted additional expenses of the borrower. Note: As per Bureau of Statistics the average expenditure of a married couple with children is 13% of their gross income.

gross\_income\_less\_fixed\_monthly\_expenses = GROSS\_INCOME-UPB- (GROSS\_INCOME \* 0.13)



1. From the new derived information, we calculated the “Mortgage Qualification Amount” i.e the loan amount that should be sanctioned as per our study.

Note: Referred the computation done by <http://tcalc.timevalue.com/all-financial-calculators/mortgage-calculators/mortgage-qualification-calculator.aspx>

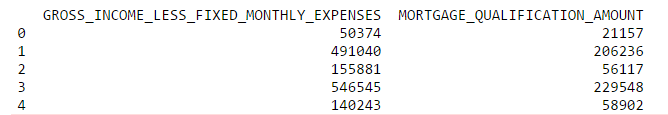
Condition:

if borrower is in "Medium Credit" standing:

take minimum of (28% of gross income,36% of gross income less fixed monthly expenses)

else if user is in "Good Credit" standing:

take minimum of (36% of gross income,42% of gross income less fixed monthly expenses)

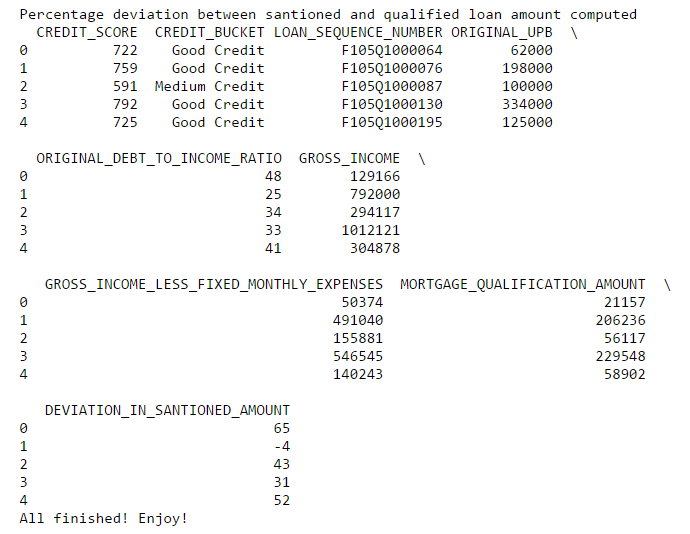


1. To analyze the deviation between the actual loan sanctioned amount (UPB) and the computed mortgage qualification amount we calculated the

DEVIATION = ORIGINAL\_UPB - Mortgage Qualification Amount

-------------------------------------------------------------------------------- x 100

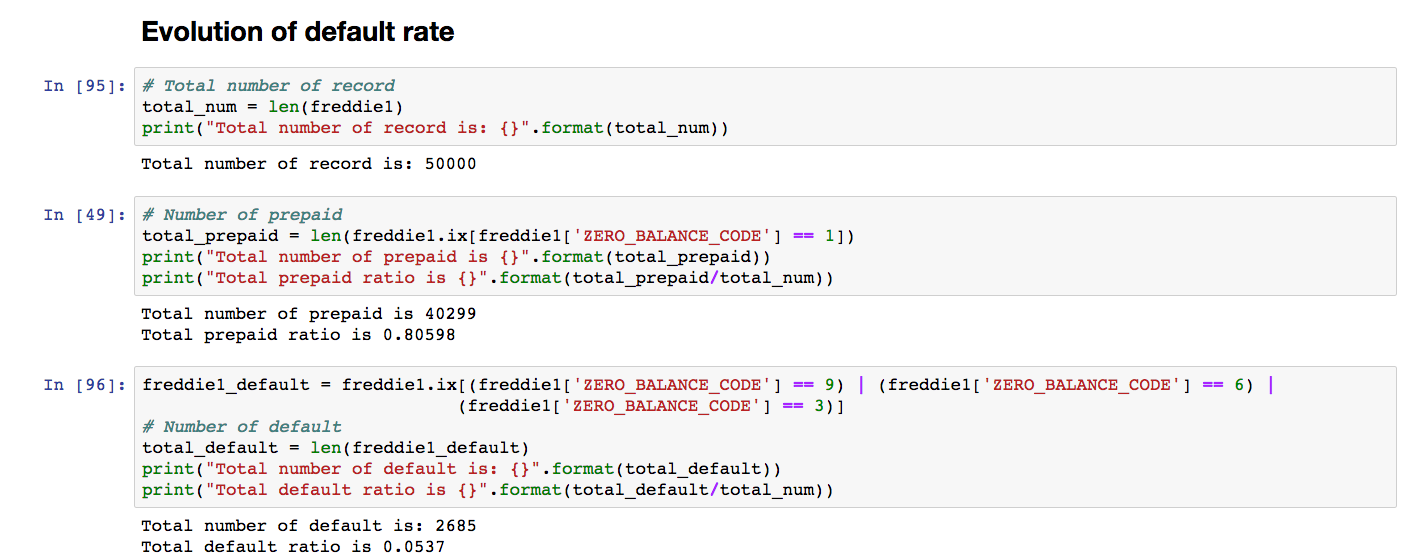
ORIGINAL\_UPB



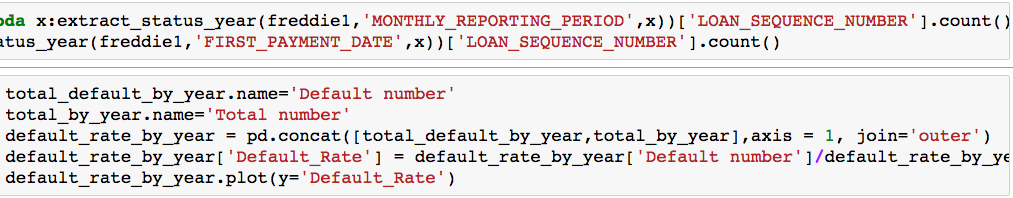
**Use case 1 conclusion:** By computing the deviation percentage in sanctioned loan amount, we identified the borrower’s for whom this percentage was very high. We then observed that most of these borrower’s were flagged as defaulters. Thus, this could be a good metric to identify customer’s who could be potential defaulters and take measures like increase their interest rate or apply a penalty fee or reduce the loan sanction amount.

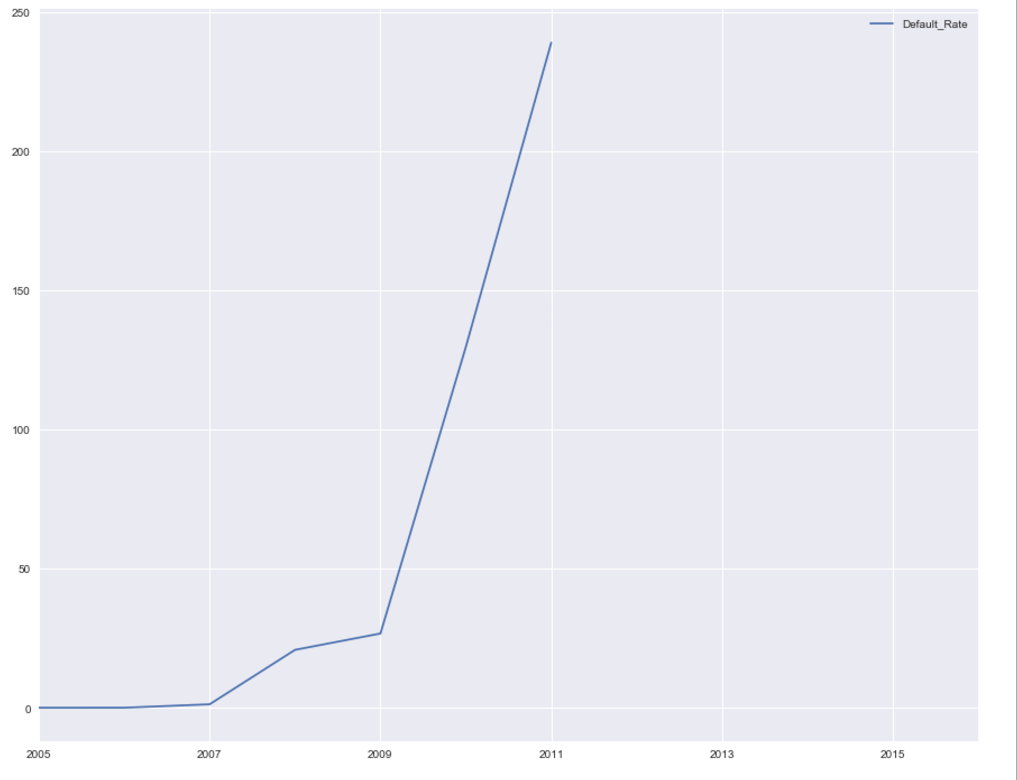
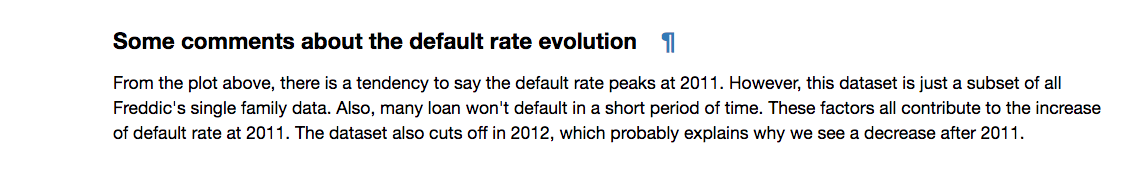
**Use case 2: Calculate defaulter rate over the years**

For any financial organization, the number of defaulters is a crucial metric. To analyze the defaulter rate over years, we calculated a count of defaulters and analyzed the evolution of defaulters.



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





**Use case 2 Conclusion:**

According to the graph there was a sudden increase in the defaulter rate in 2007 this is because 2007 was economic depression year and rate starts stabilizing towards 2009 as it was marked the beginning of economic boom years.

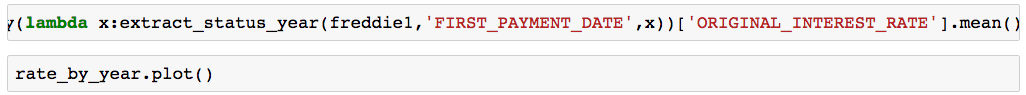
Note: The sudden increment in the year 2011 is because we do not have complete data for the year 2011.

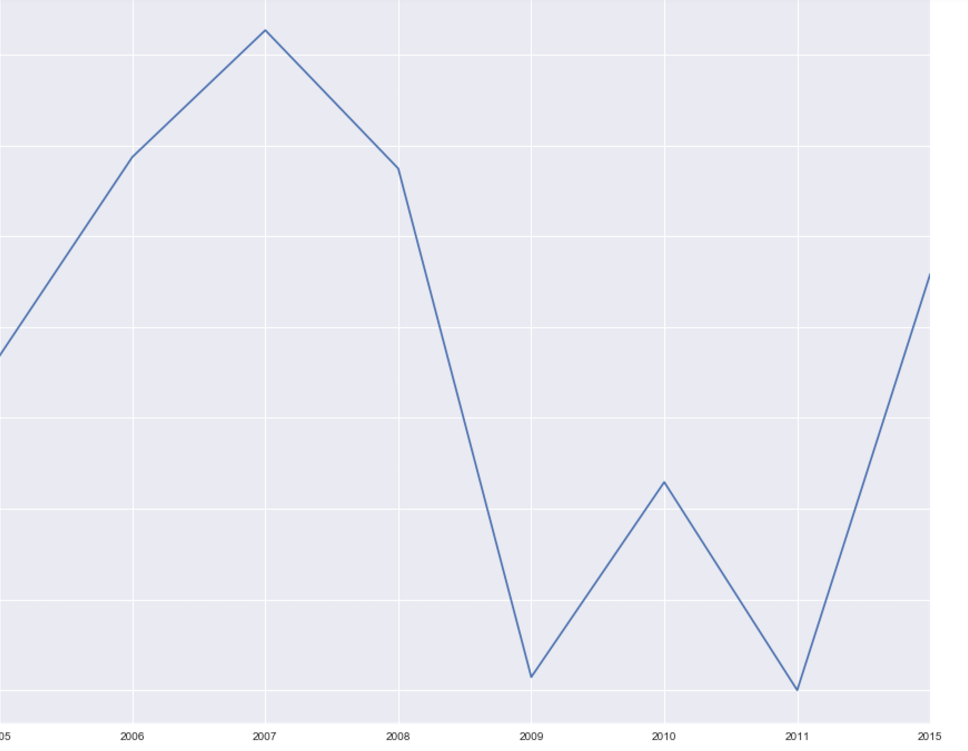
The ratio of prepaid loans in year 2005 was .805

The ratio of default loans in year 2005 was 0.0537

**Use case 3: Effect of interest rate over the years**

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

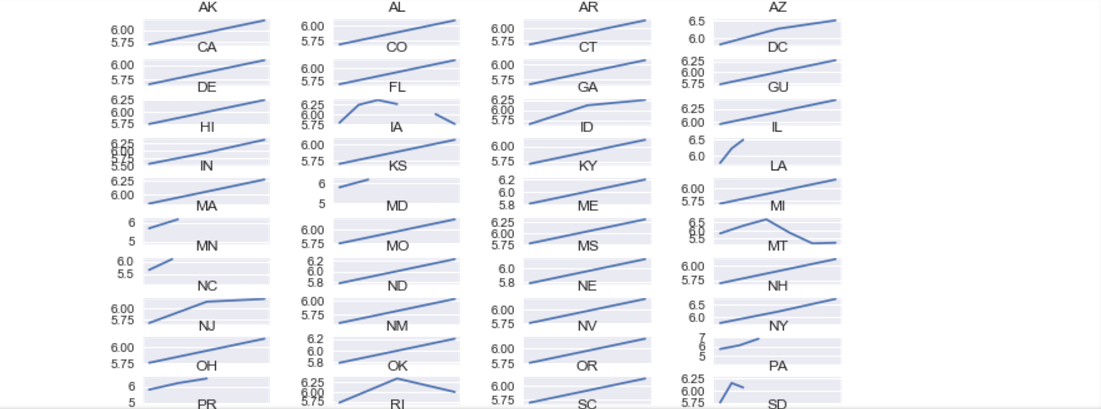




**Use case 3 conclusion:**

We observed that 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**

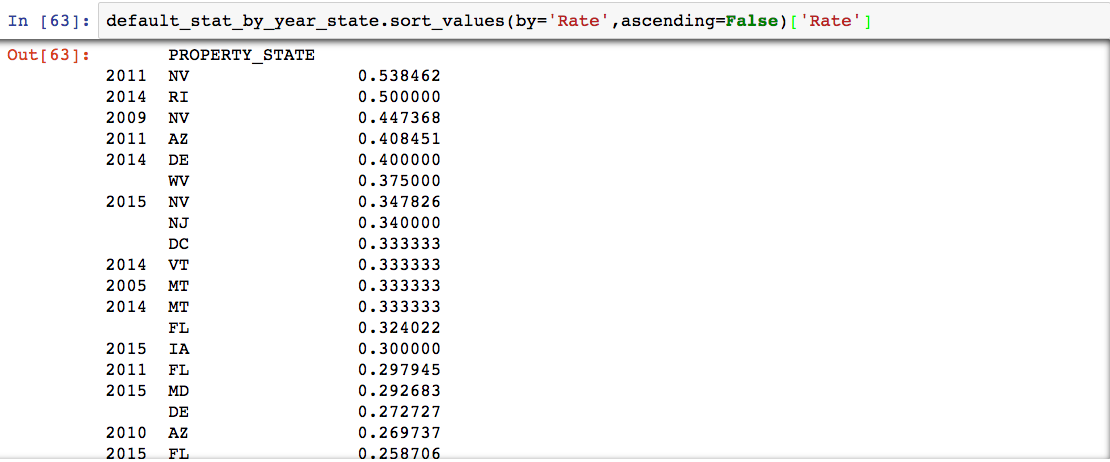


**Summary metrics:**

**Metric 1:**

**Defaulter rate in each state per year**

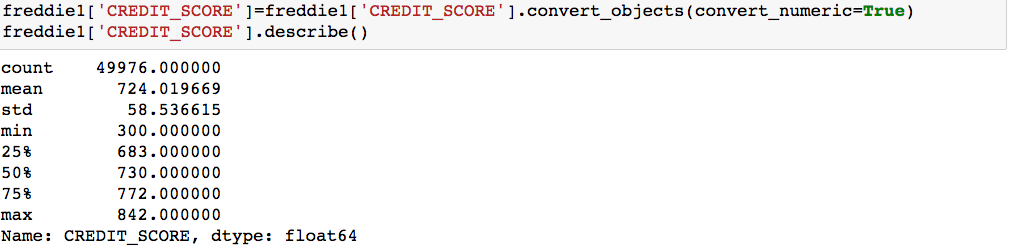




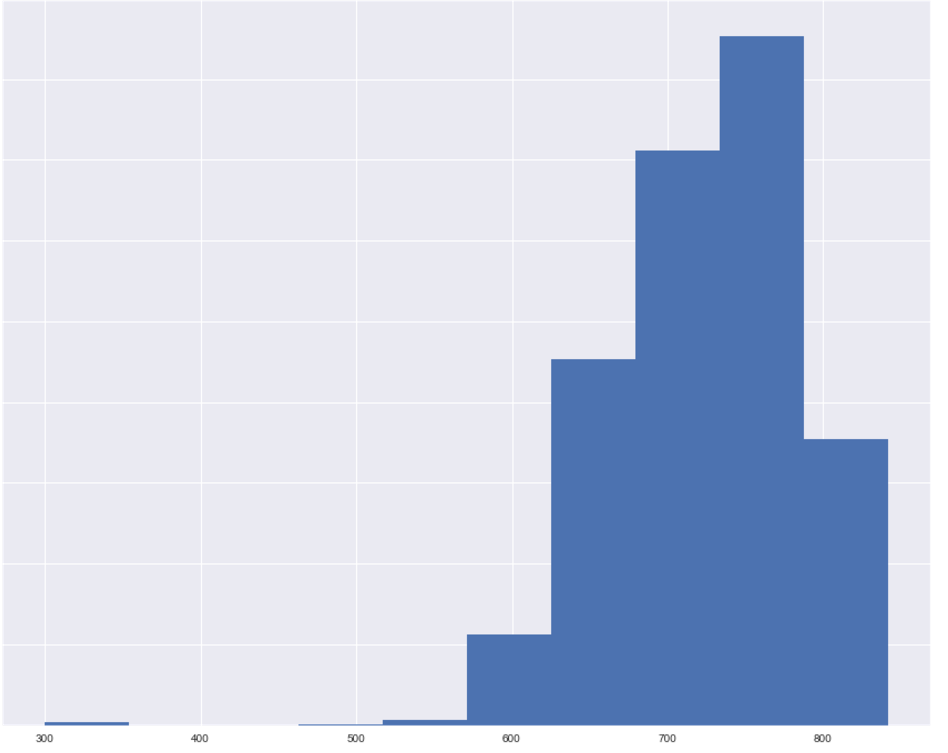
**Metric 2:**

**Defaulter cases by Credit Score**

The average credit score of all applications is 724. Surprising!!!. 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**



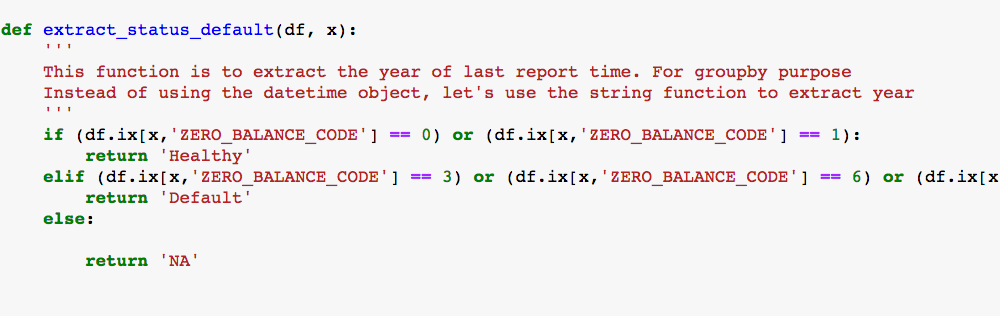
**Metric 3:**

**Analyzing zero balance code over the years**

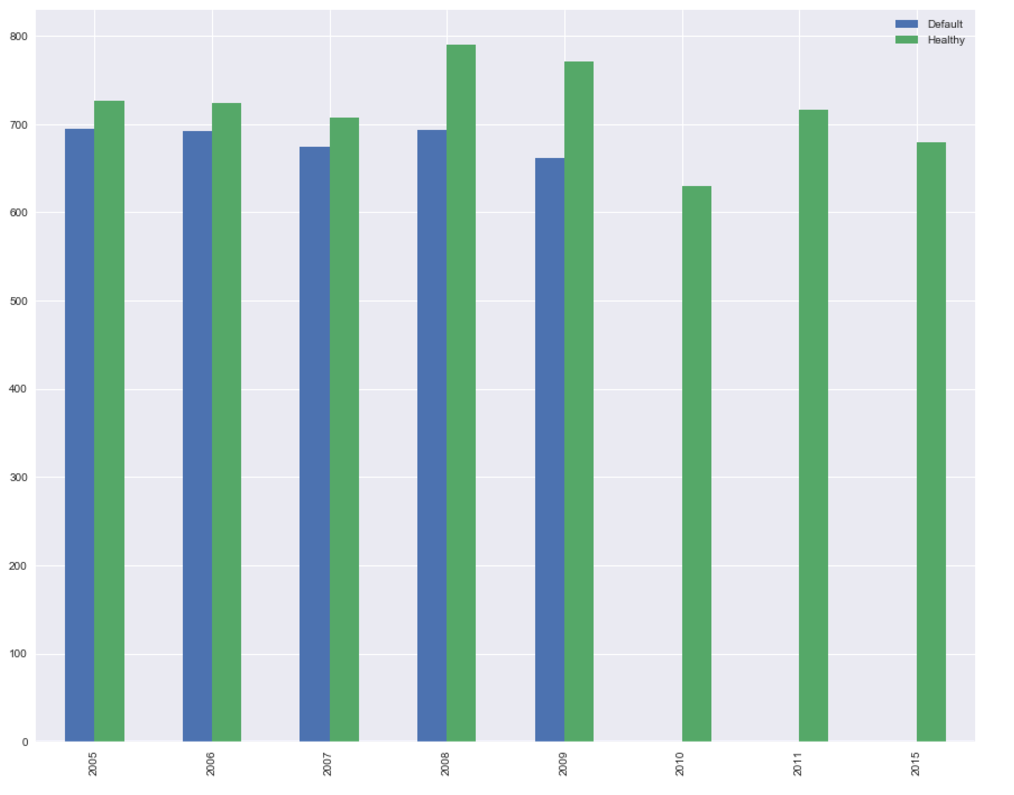
Creating set based on Zero\_balance\_code

Zero\_balance\_code with 0 and 1 are healthy

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



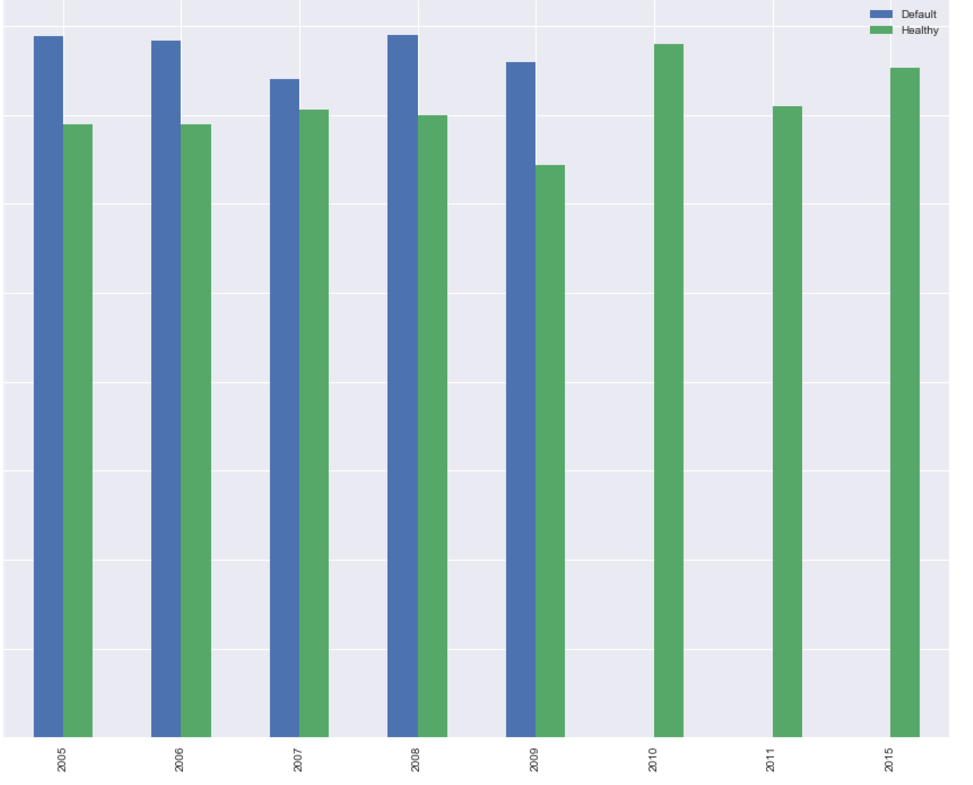
**Graph 1: Loan Status vs credit score**



This graph shows some interesting findings:

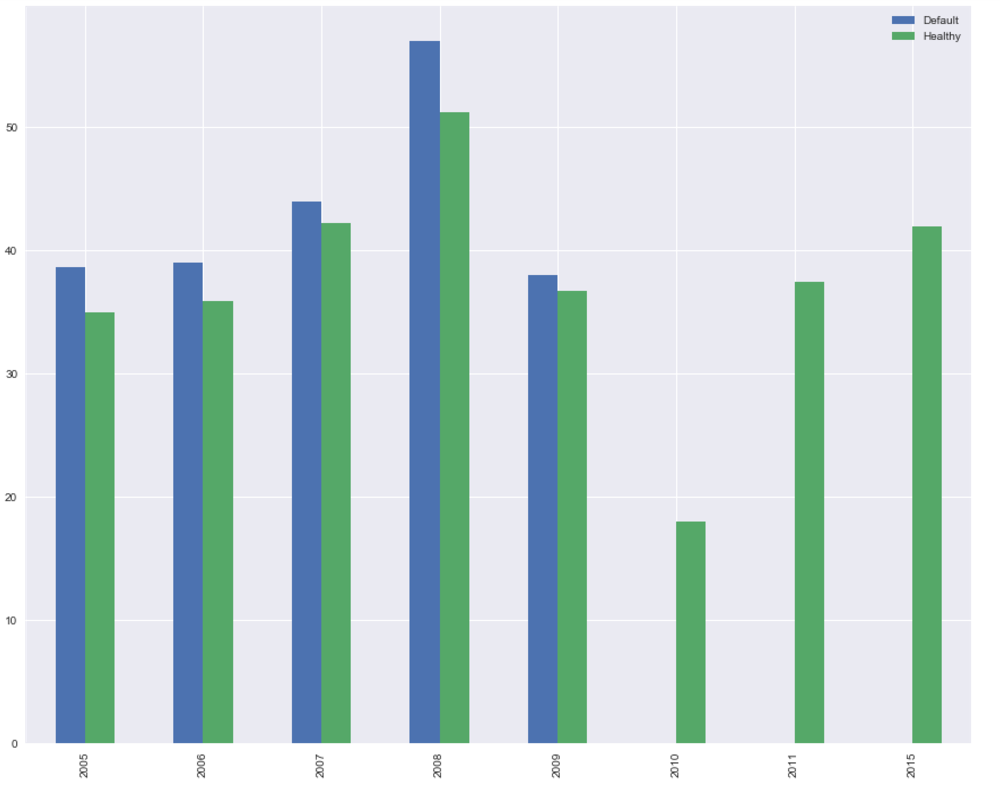
1. The overall trend of applicant's credit score is increasing by year.

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

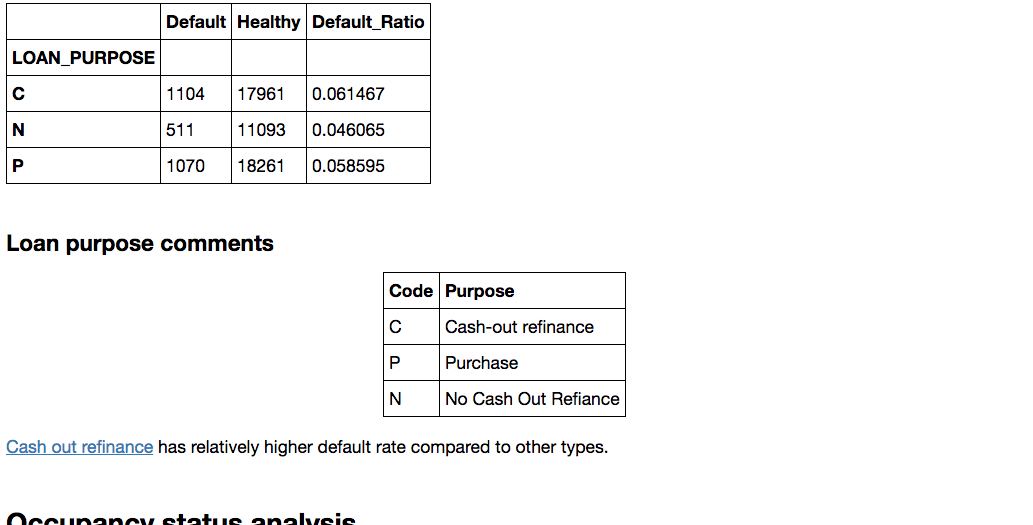
### 

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.

**Metric 4:**

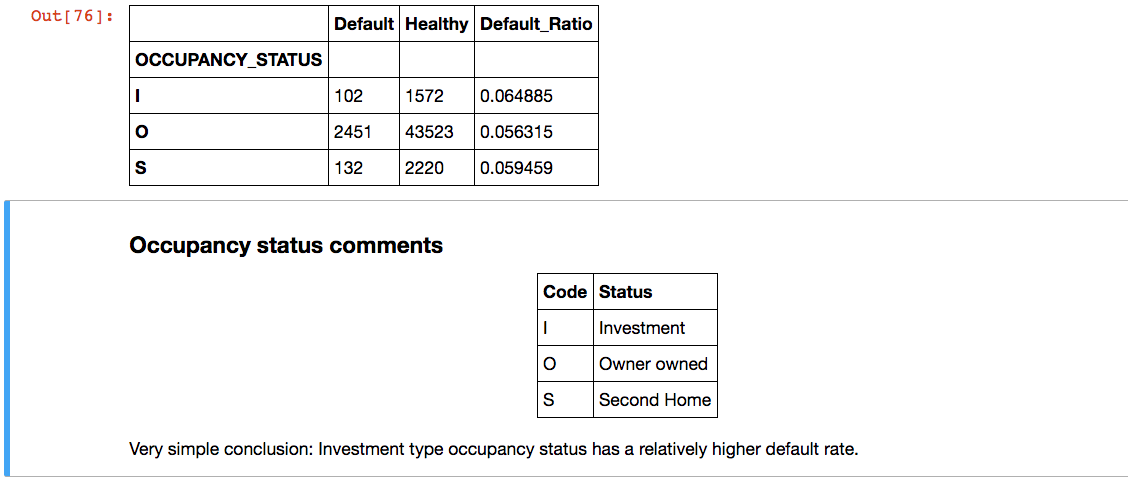
**Loan Purpose analysis**

Cash out finance has higher defaulter rate



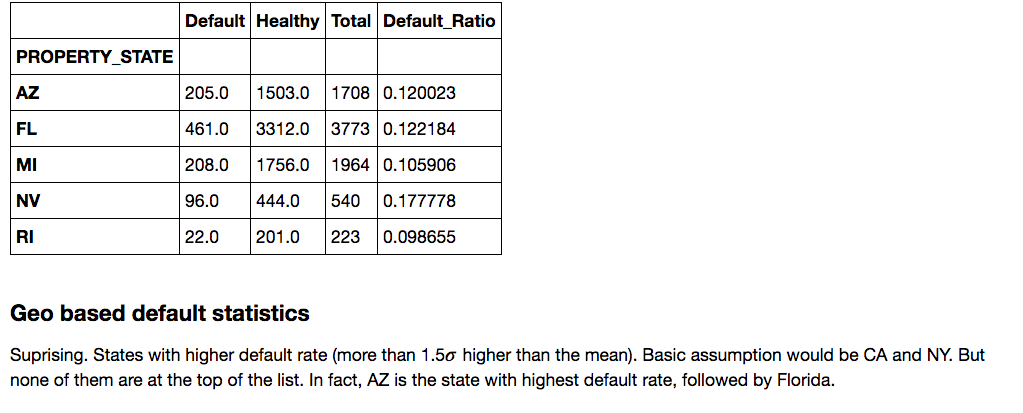
**Metric 5:**

**Occupancy status analysis**



**Metric 6:**

**Geo-Location analysis**

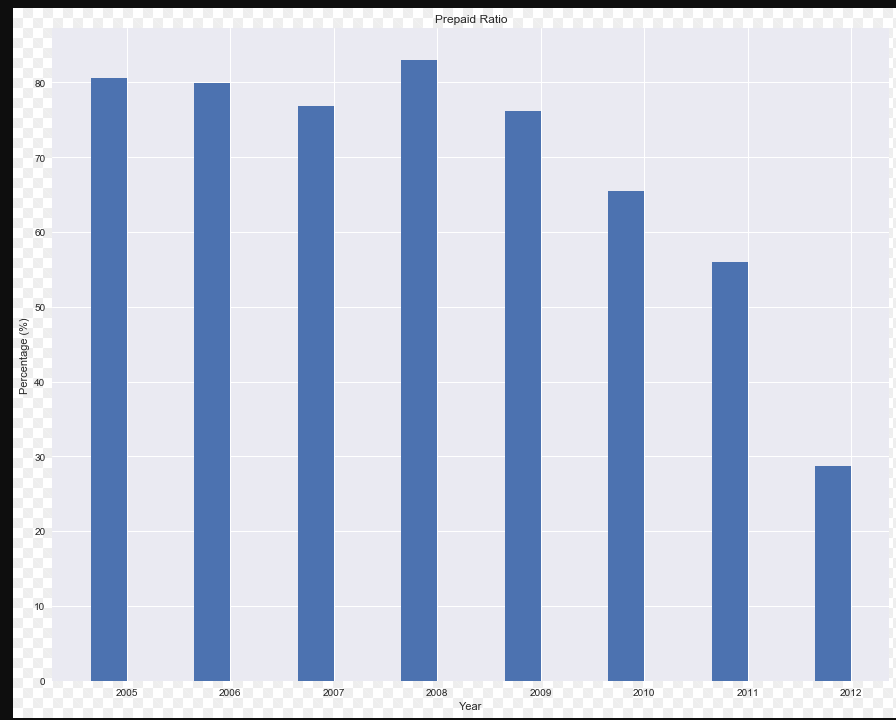


**Interesting findings:**

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

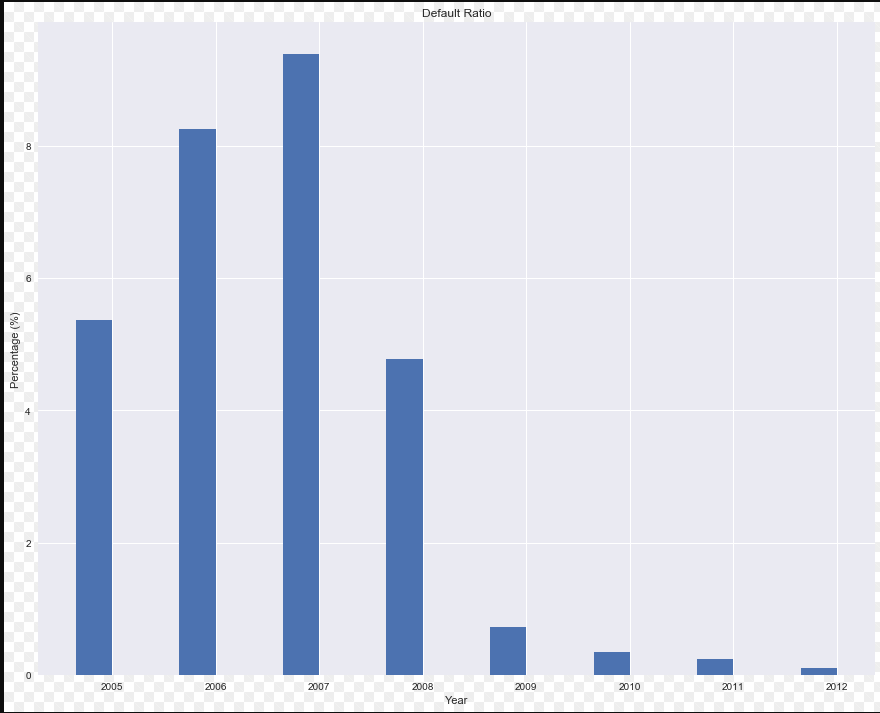
**How many mortgages have been prepaid in these years?**

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



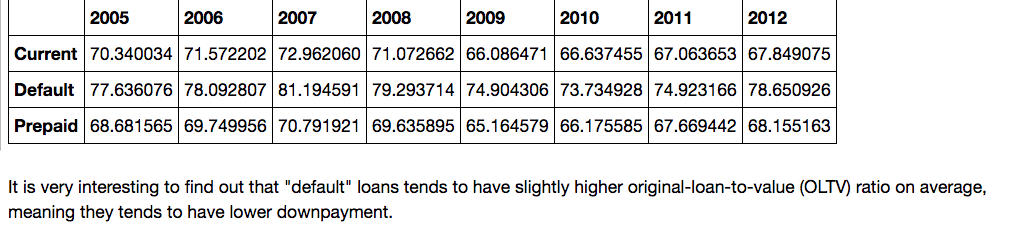
**Defaulter Rate**

Highest during depression then reduces as economy is doing well

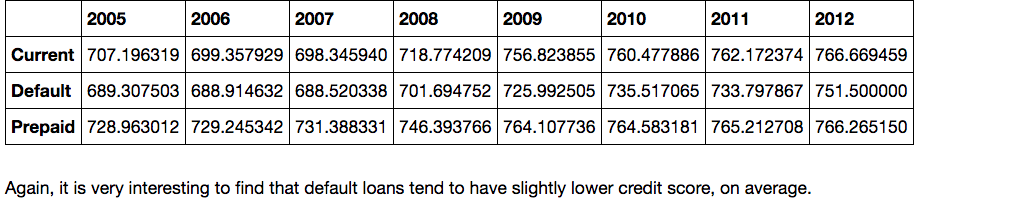


Default rate is very high in 2007 due to economic recession due to subprime lending. More details on <https://en.wikipedia.org/wiki/United_States_housing_bubble>

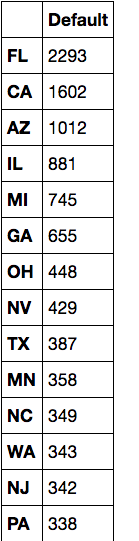
**OLTV for default loan**



**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**