**MIDTERM PROJECT**



**FREDDIE MAC SINGLE FAMILY LOAN DATASET**

http://www.freddiemac.com/research/datasets/sf\_loanlevel\_dataset.html

**TEAM 5**

**AKILAN RAJENDIRAN**

**DHRUV KANAKIA**

UNDER THE GUIDANCE OF:

Prof. SRIKANTH KRISHNAMURTHY

PRANJAL JAIN

Contents

[1.1. TREE DIAGRAM(WORKFLOW) 3](#_Toc488441099)

[2.1. DOCKER INSTRUCTIONS FOR EXECUTION: 4](#_Toc488441100)

[3.1. WEB SCRAPING and PRE-PROCESSING 5](#_Toc488441101)

# **TREE DIAGRAM(WORKFLOW)**

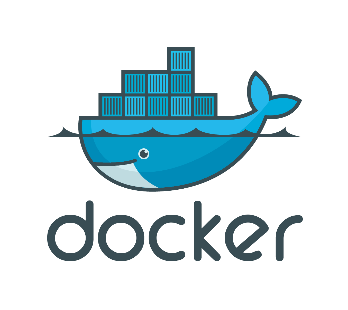
WEB SCRAPING and PRE-PROCESSING

EDA ON PRE-PROCESSED DATA

BUILDING and EVALUATING MODELS

PREDICTION

CLASSIFICATION

 IMAGE

## **2.1. DOCKER INSTRUCTIONS FOR EXECUTION:**

* Docker pull akil06/midterm:image1
* sudo docker run -ti -p 8888:8888 image id /bin/bash
* jupyter notebook --ip 0.0.0.0 --no-browser --allow-root
* Docker pull akil06/midterm:image2
* sudo docker run -ti -p 8888:8888 image id /bin/bash
* jupyter notebook --ip 0.0.0.0 --no-browser --allow-root
* Docker pull akil06/midterm:image3
* sudo docker run -ti -p 8888:8888 image id /bin/bash
* jupyter notebook --ip 0.0.0.0 --no-browser --allow-root

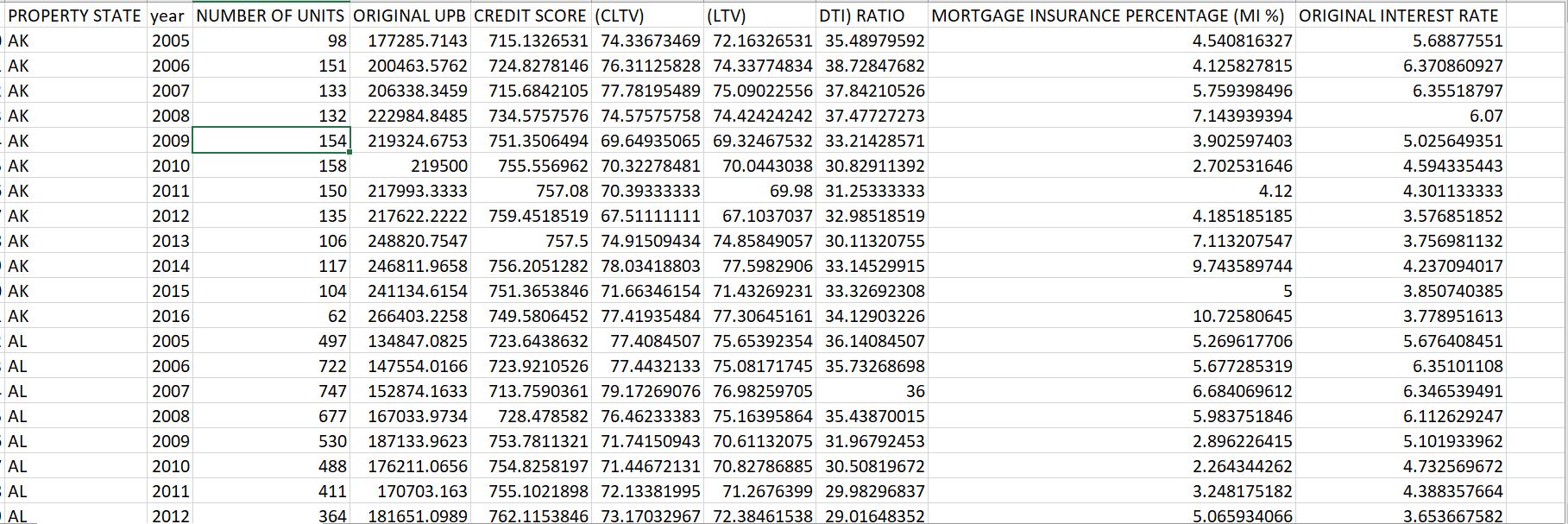
­­

### **3.1. WEB SCRAPING and PRE-PROCESSING**

* We used requests and lxml library to get pass the login page and download the data.
* The downloaded data is unzipped in the unzipped\_files folder and these files are read in the python notebook.
* We read the origination and performance files separately from the unzipped folder and created Summary files.
* Below are the screenshots of few summary files created.

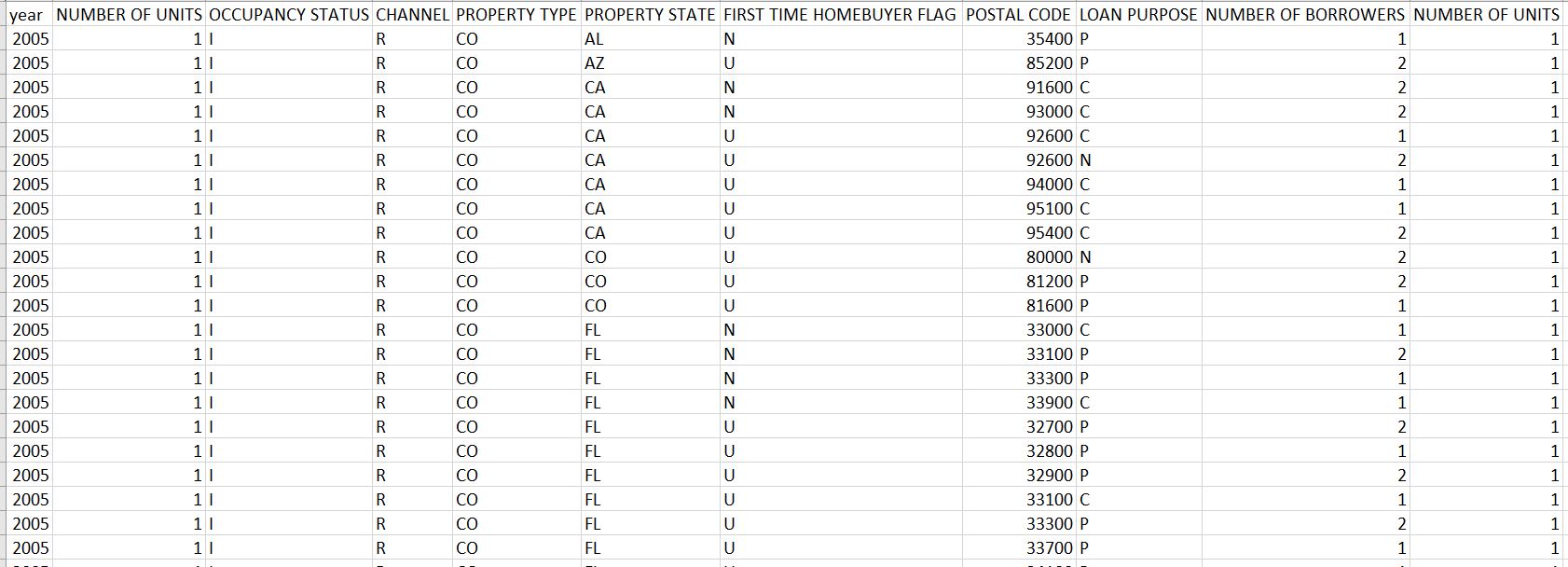
1. **ORIGINATION SUMMARIES:**
2. This summary file shows summary of all numerical measure in origination file that can be indexed on either ‘year’ or ‘State’ and perform aggregation function on remaining column. Basically, this is aggregation of numerical values.

This can be viewed in .csv file named ‘Orig\_Loan\_Details\_Over\_Year\_State.csv’



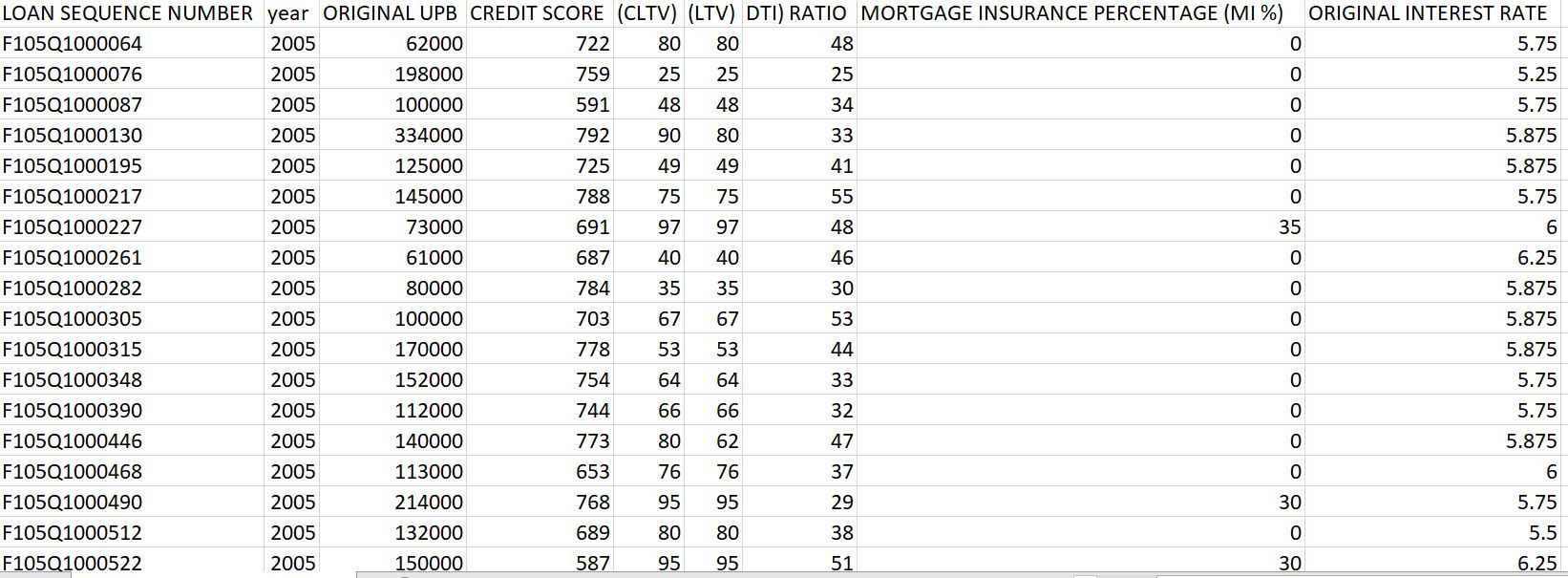
1. The second summary file shows the summaries of loan type i.e. what type of loan, what loan purpose, loan type, Number of borrowers, Unit count etc. were given over the years. Basically, this is summary of different parameters of loans.

This can be viewed in .csv file named ‘Loan\_Type\_Year\_State.csv’



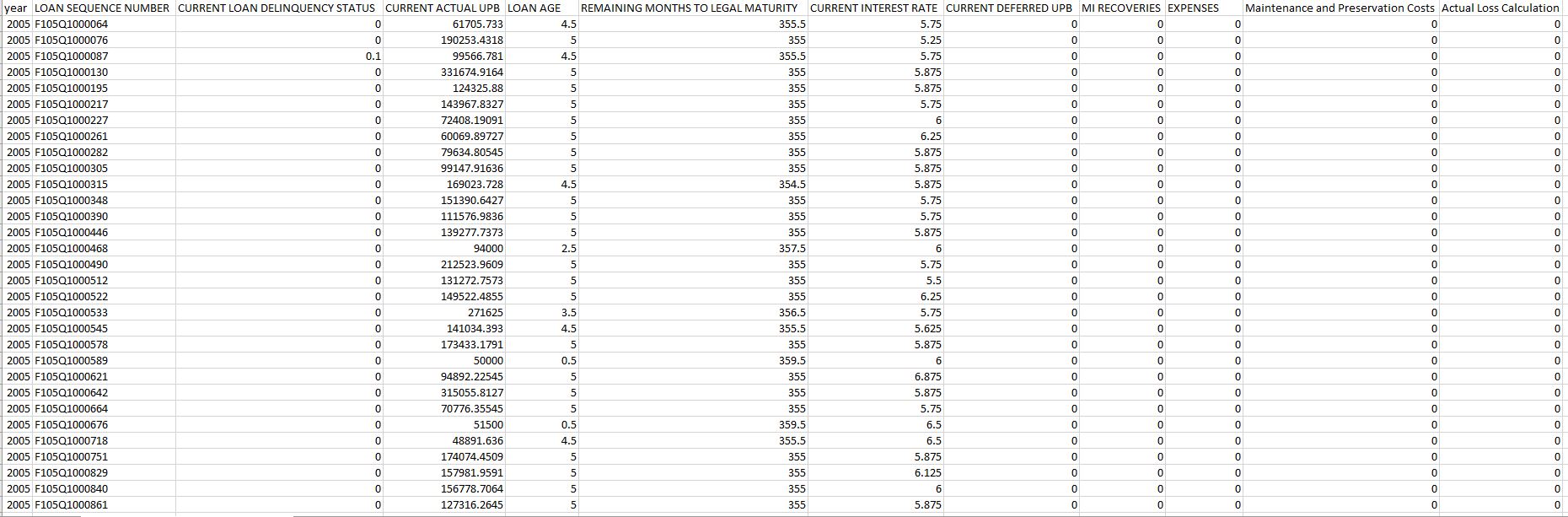
1. This summary was created so that we can inner join Performance file on LOAN SEQUENCE NUMBER and do detailed analysis.

This can be viewed in ‘Orig\_Loan\_Number\_Summary.csv’



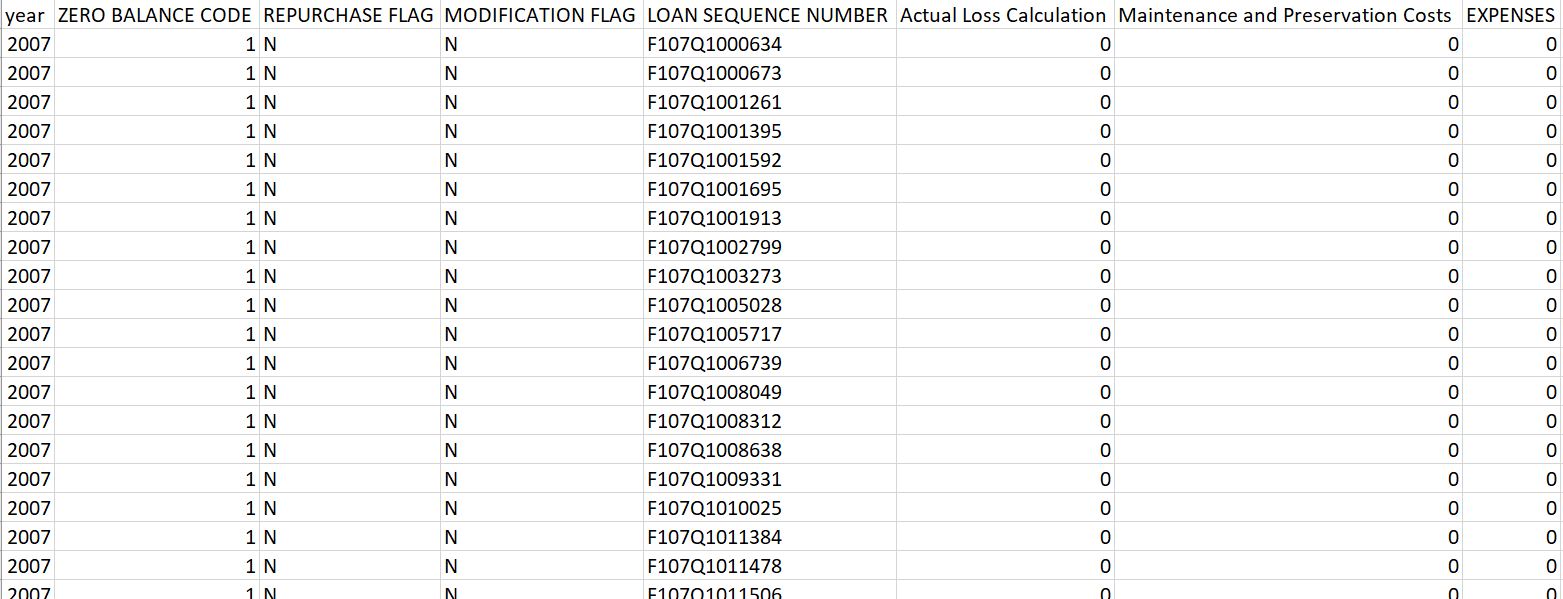
1. **PERFORMANCE SUMMARIEs:**
2. This summary file includes all the numerical attributes that can be indexed either by year or by Loan sequence number.

This can be viewed in Summary1\_Performance.csv



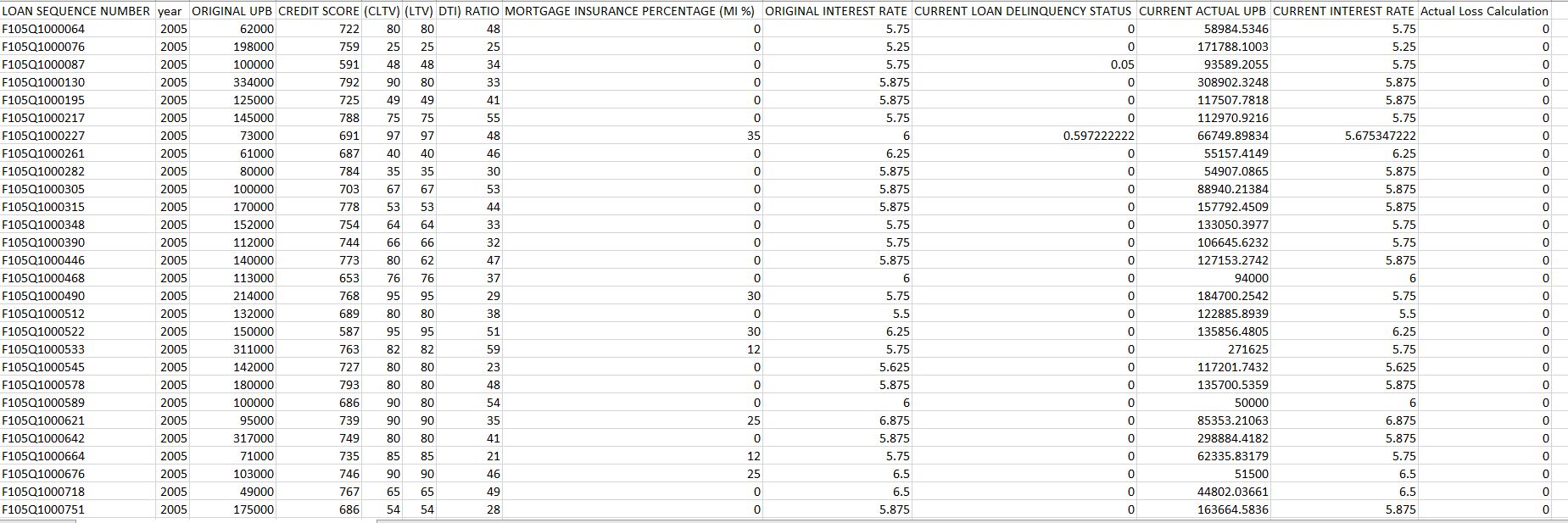
1. This summary includes detail about each loan over the year. The most important parameter here is actual loss calculation over the Zero balance code which we are using in EDA.

This file can be viewed in Summary2\_Performance.csv



1. **SUMMARIES CONTAINING BOTH PERFORMANCE AND ORIGINATION:**
2. This summary files include summaries that have been obtained by merging performance and origination keeping in mind few EDA to be performed on them.

The file can be viewed in ‘Performance\_Origination\_Summary1.csv’



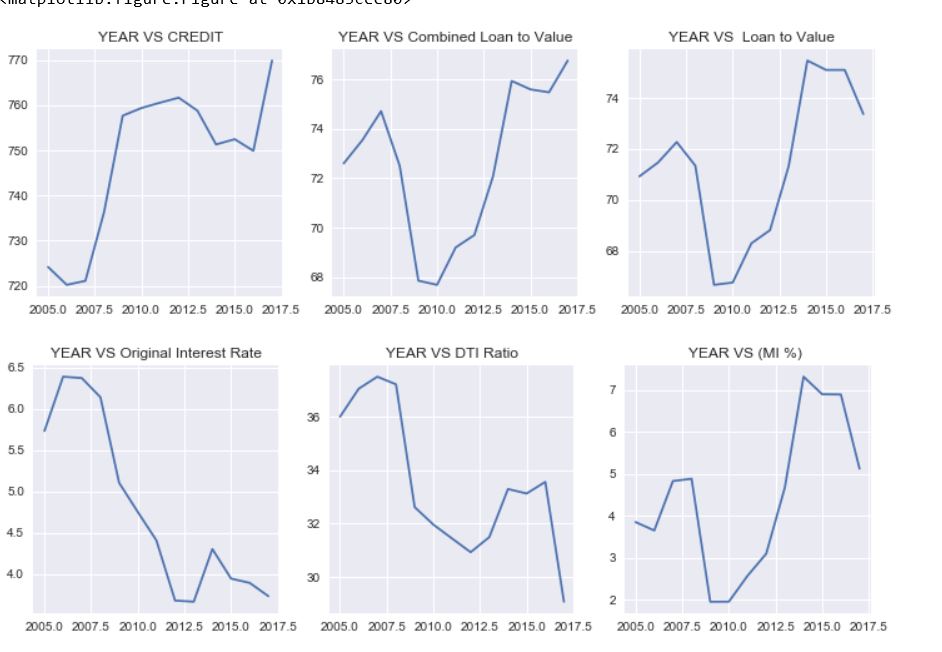
#### **4.1. Pre-Processing and Exploratory Data Analysis**

1. A screenshot on how we are handling the missing values based on the documentation.

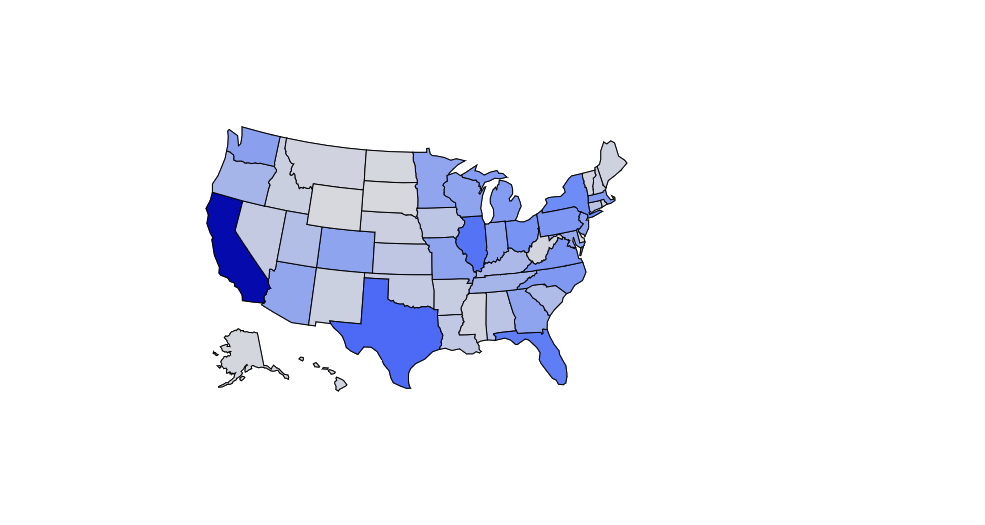


1. Based on the summary files created we have plotted the EDA.
   1. Yearly trend of all the major numerical columns analyzing at how they vary over the years

**CONCLUSION**: We can see that Loan to Value, COmbined Loan to Value and MI% are very much related to each other. We can also the decline in almost all the graphs starting from the year 2007 i.e. when the recession period started. As economy was improving we can see increase in almost all the plots tarting from year 2011. One more analysis that we can see is almost every plot reaching its peak in 2014 post recession. We will look more into it in further analysis

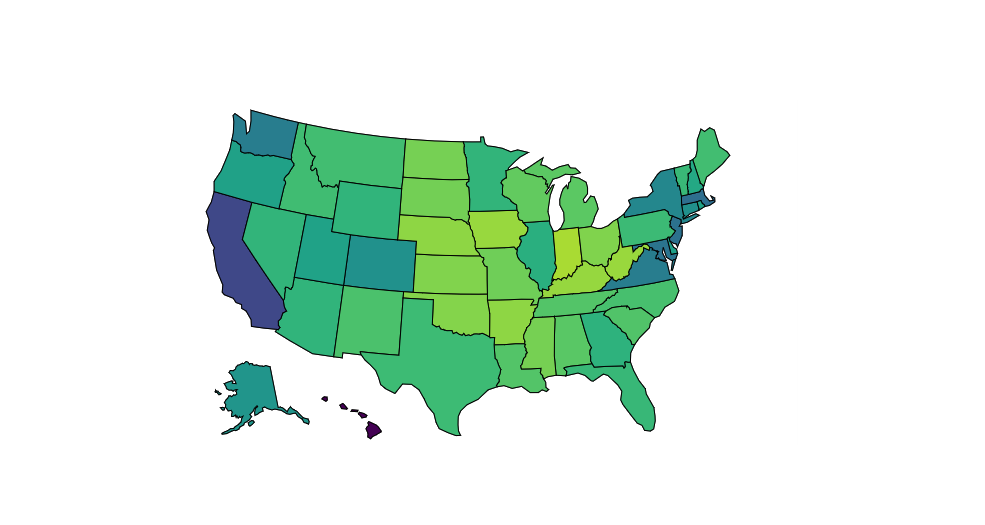


* 1. NUMBER OF LOANS OVER THE STATE:



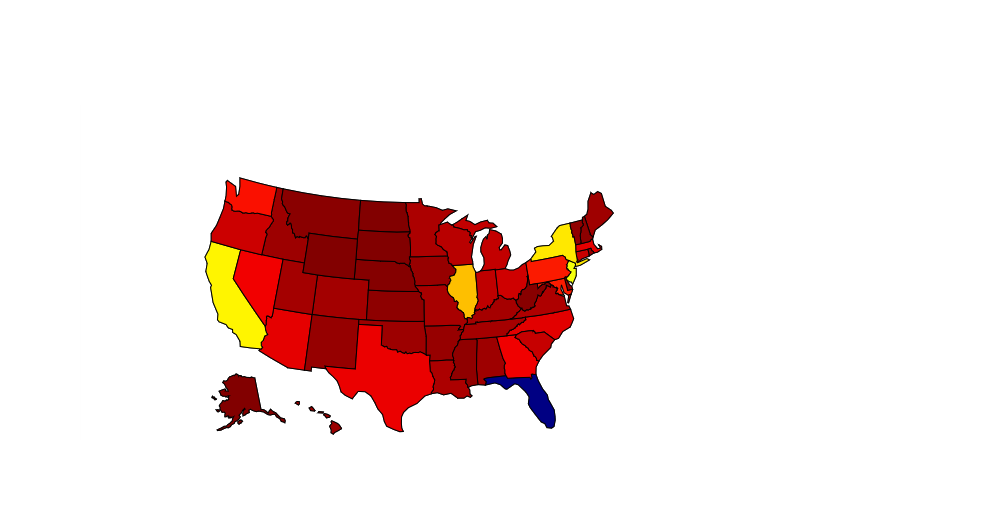
**CONCLUSION:** The top 5 states having most number of loans are CALIFORNIA, TEXAS, FLORIDA, ILLINOIS and NEW YORK.

* 1. **ORIGINAL UPB OVER THE STATE:**



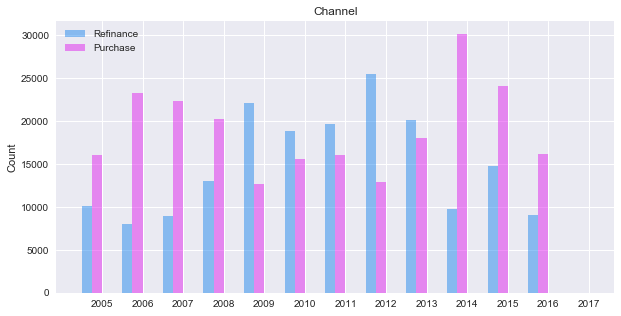
**CONCLUSION:** CALIFORNIA still leads the race when it comes for original UPB. Comparing from the above graph i.e number of loans given, FLORIDA and TEXAS has less number of Original UPB. While, states like Virginia, Colorado, Washington and Maryland have significantly more UPB

* 1. **CURRENT DELIQUENCY OVER THE STATE**



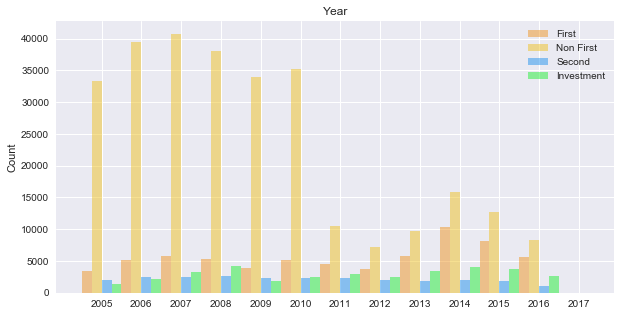
**CONCLUSION:**  Surprisingly , FLORIDA has the highest number of Delinquency rate even after less number of UPB. Other states followed by Florida having more Delinquency rates are CALIFORNIA, ILLIONOIS, NEW JERSEY, NEW YORK.

* 1. **CHANNEL TYPE COUNT OVER THE YEAR**

****

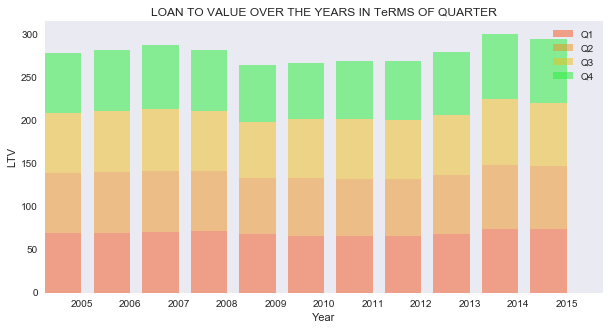
**CONCLUSION:** We can see that the purchasing count going downwards from 2007 onwards as the recession begins. In 2009, the refinance is way past purchasing. The purchase loan is then increasing as the economy recovers. LTV was 76.6% in 2014 - a new high since 2000, largely due to the increase of purchase volume over refinances.

* 1. **OCCUPANY vs YEAR vs COUNT**

****

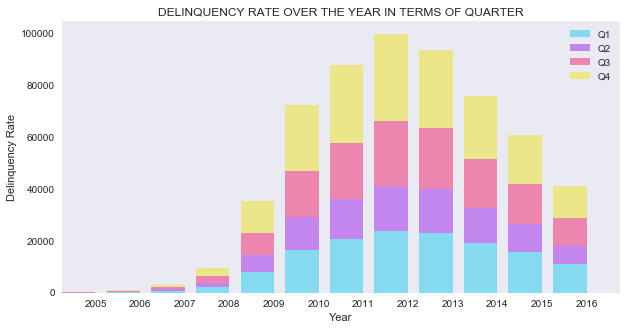
**CONCLUSION:** Since, interest rates went down drastically in 2013. You can see the rise in purchase in previous graph. While, in this graph we can clearly see that 2014 was the best year for homebuyers after recession as investors, seond time home buyers and even first time home buyers ratio went up. Investors and non first also reached its peak in 2014. The decline in interest rate internet says is because of a combination of economic weakness in Europe, concerns about Ebola and geopolitical turmoil around the world. It was also because of new lending scheme that was introduced during that time

* 1. **LOAN TO VALUE IN TERMS OF QUARTER:**

****

**CONCLUSION:** We can see LTV going down in the recession period. We can clearly see Quarter 4 is the quarter which is mostly affected over all the year's

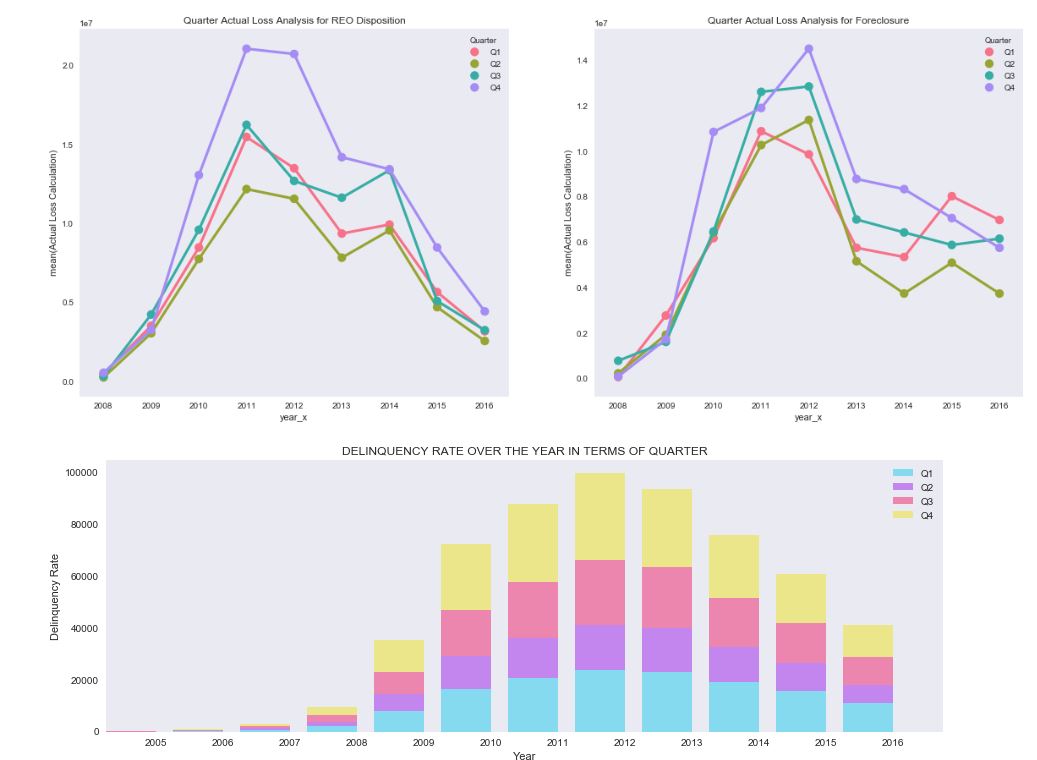
* 1. **DELINQUENCY RATE OVER THE QUARTER:**

****

**CONCLUSION:** As the recession hit the economy in 2007 we can clearly see increase in DELIQUENCY rates. With less jobs and less LTV the Deliquency rate was bound to increase and it reached its peak in 2012. Quarter 4 again has the highest number of Deliquency increase rates. With economy improving over the year's we can clearly see decline in Deliquency after 2012

* 1. **LOSS VS QUARTER COMPARING IT WITH DELIQUENCY:**

**CONCLUSION:** We can clearly see that REO is much more as compared to RECLOSURE. When we look at Deliquency and LOSS PLots we can clearly see the relation. With the increase in deliquency the Loss has also increased. Quarter 4 again has significantly higher rate in all the 3 plots.

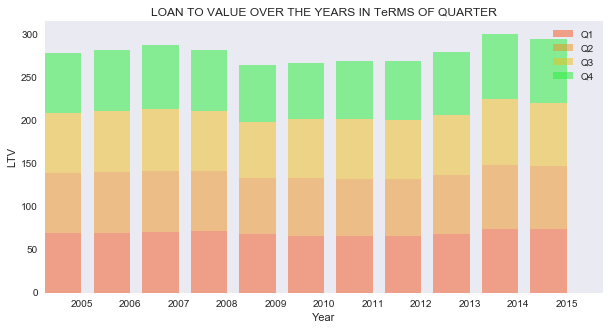
****

**FINAL CONCLUSION:** The conclusion points that we have come across after analyzing the data are:

Q4 has normally been the major declining quarter in the recession period.

The laws passed by OBAMA in 2009 has played a significant role to get the economy back on track. The laws passed by OBAMA that allowed more and more people to refinance their loan because of which was designed to help more responsible homeowners refinance their mortgages.

The law passed by OBAMA in October 2009, i.e. third quarter directly impacted the DELIQUENCY, LTV, DTI, INTEREST RATE. You can see below how the LTV graph Q3 color stabilized after 2009 and eventually went up.

****

So, the major factor getting the economy back on track is the **HOME AFFORDABLE REFINANCE PROGRAM** launched by OBAMA in 2009.