Final Project Progress Report

1. Project Summary

Implement an infrastructure to analyze U.S. mortgage data and home price data for rating analysts. So far we have started exploring samples of our data sources to identify the cleanliness of the data and specify the ancillary sources that will be useful to add. We have also begun building the infrastructure to store and process the data on a large scale. We have not yet run into major roadblocks.

2. Project Requirements

a. *Core Requirements*
   i. Perform exploratory data analysis on raw data
   ii. Download a subset of data from data sources into AWS
   iii. Load raw data into HDFS
   iv. Create Hive metadata for underlying data
   v. Perform exploratory analysis using Tableau and PySpark

b. *Additional Requirements*
   i. Write a shell script to download data from websites in a batch
   ii. Try to add more historical data to analyze scaling capacity
   iii. Incorporate streaming home price data using Zillow API
   iv. Perform a simple logistic regression for loan default
   v. Build a dashboard in Tableau
   vi. Think about how this project can evolve in future

3. Tasks Completed

a. *Exploration of Ancillary Data Sources*
   i. Explored data sets available from Federal Housing Finance Agency (FHFA), US Federal Reserve Economic Data (FRED) including compilations from Bureau of Labor Statistics and Census, and Zillow
   ii. Identified data sets that are most useful in contents, ability to match to primary data sets, and ease of accessibility. Summary of these data sets is included here, with more detail in the appendix
      1. FHFA - house price index by quarter by three-digit zip code
      2. Census - median or average price of new single-family homes by quarter by region
      3. Census - new single-family houses for sale at end of month by region
      4. Census - new single-family houses sold each month by region
      5. Zillow - % of homes decreasing in value by state by month
      6. Zillow - % of homes increasing in value by state by month

b. *Downloading a subset of data from data sources into AWS*
   i. Created accounts for Fannie Mae and Freddie Mac websites.

ii. Downloaded the latest quarter data (txt format) from these websites. Both agencies separate data into loan acquisition data and loan performance data.
iii. Transferred the downloaded data to AWS EC2 instance using “scp” command.

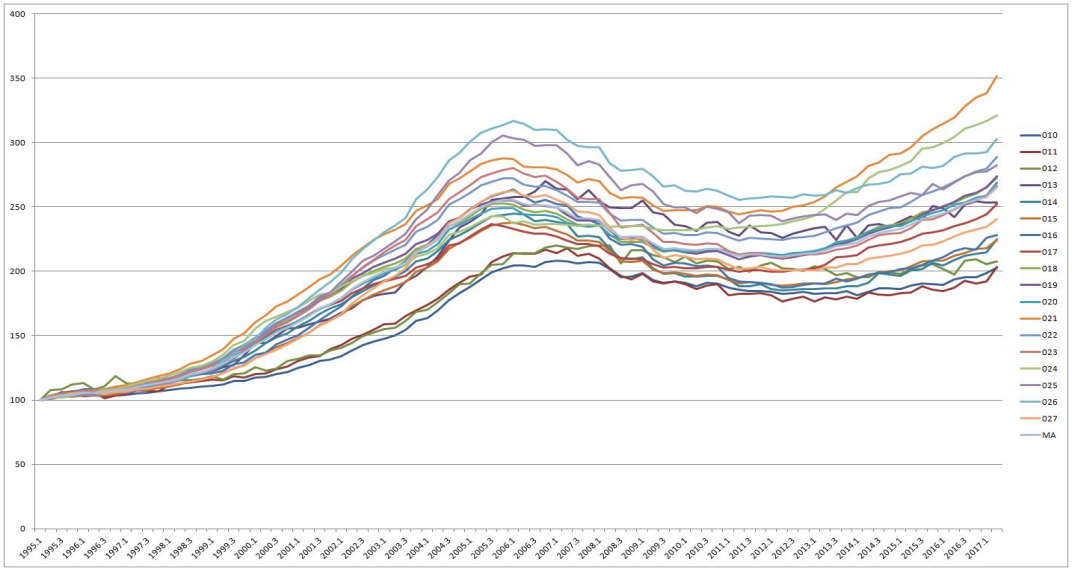
c. *Loading raw data into HDFS*
   i. Created 4 HDFS folders: Fannie Mae acquisition data, Fannie Mae performance data, Freddie Mac acquisition data, and Freddie Mac performance data.
   ii. Loaded raw txt files into HDFS.
   iii. https://github.com/kr900910/W205/tree/master/mortgage-data-analysis/loading_and_modeling

d. *Creating Hive metadata for underlying data*
   i. Combined Fannie Mae acquisition data and Freddie Mac acquisition data into a single loan acquisition data.
   ii. Combined Fannie Mae performance data and Freddie Mac performance data into a single loan performance data.
   iii. Since these two data sources use different notations for same field, we need to further investigate how we will standardize those fields.
   iv. https://github.com/kr900910/W205/tree/master/mortgage-data-analysis/transforming

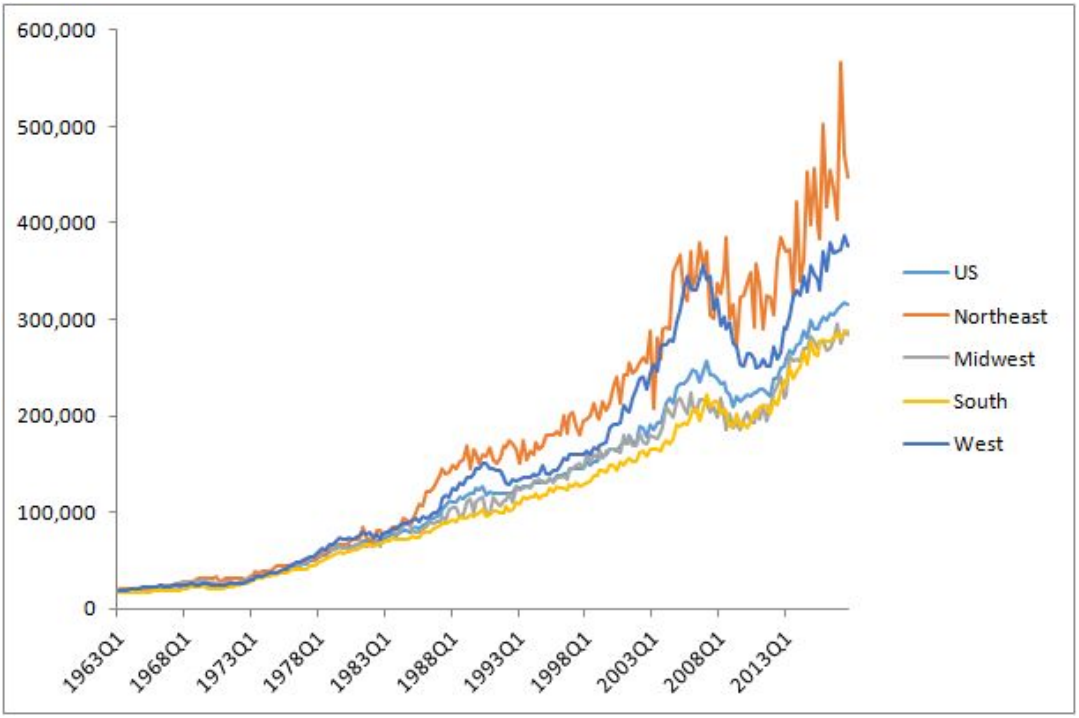
Appendix: Details on Data Exploration

1. *Exploration of Federal Housing Finance Agency House Price Index Data*
   a. Downloaded data and examined it to determine which data set is most useful and how it connects to our primary data sets
   b. Note that the entire dataset including historical is updated in full each quarter. When a data refresh is needed for our database, we will need to pull the entire dataset and replace the one in our infrastructure
   c. Description of index: HPI is a weighted, repeat-sales index, meaning that it measures average price changes in repeat sales or refinancings on the same properties. This information is obtained by reviewing repeat mortgage transactions on single-family properties whose mortgages have been purchased or securitized by Fannie Mae or Freddie Mac since January 1975. The methodology is a modified version of the Case-Shiller geometric weighted repeat-sales procedure
   d. Concluded that we will use three-digit zip code data. This data is:
      i. Available at a quarterly level from 1995-2017. Dataset is small with only 79,380 rows
      ii. Not seasonally adjusted
      iii. Includes all transactions - purchases as well as refinancing appraisals
      iv. For zip codes without sufficient data, it uses aggregated data and thus has no missing values. 66% is zip-code level, 13% is MSA level, and 21% is state level
      v. Can be downloaded from https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx#mpo at ‘Quarterly Data: All-Transaction Indexes: Three-Digit Zip Codes’
   e. Reasons for this choice:
      i. Matches Fannie Mae/Freddie Mac acquisition data three-digit zip field
      ii. Provides more granular information than MSA or state level

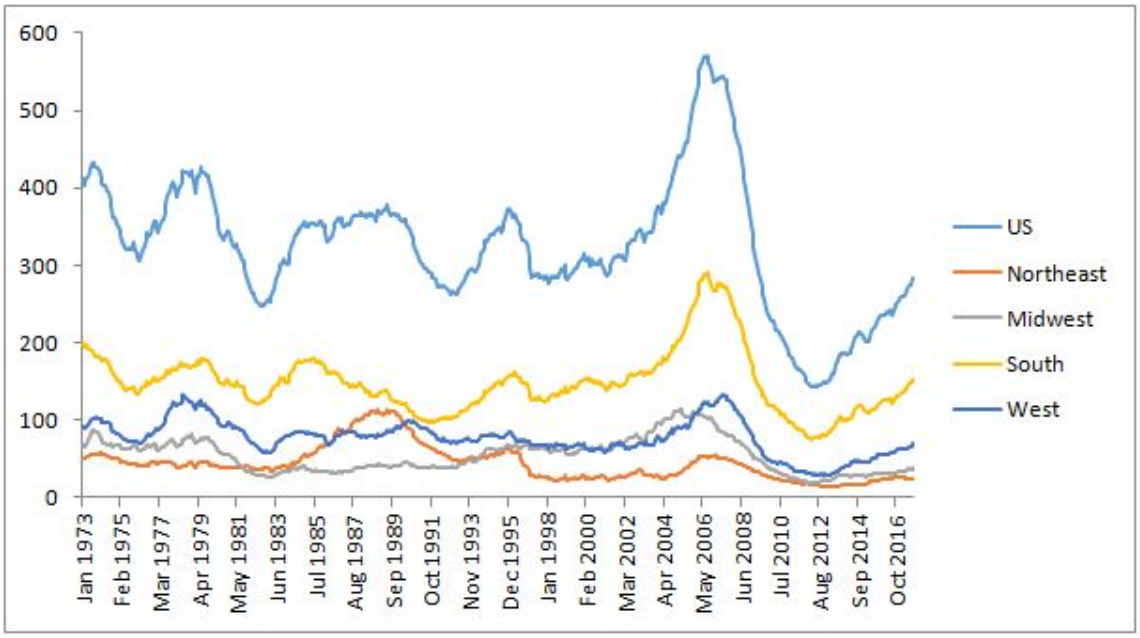
iii. While high-level trends do not vary much in a given state, there are substantial differences by state (e.g., see the graph of all three digit zip codes in MA)



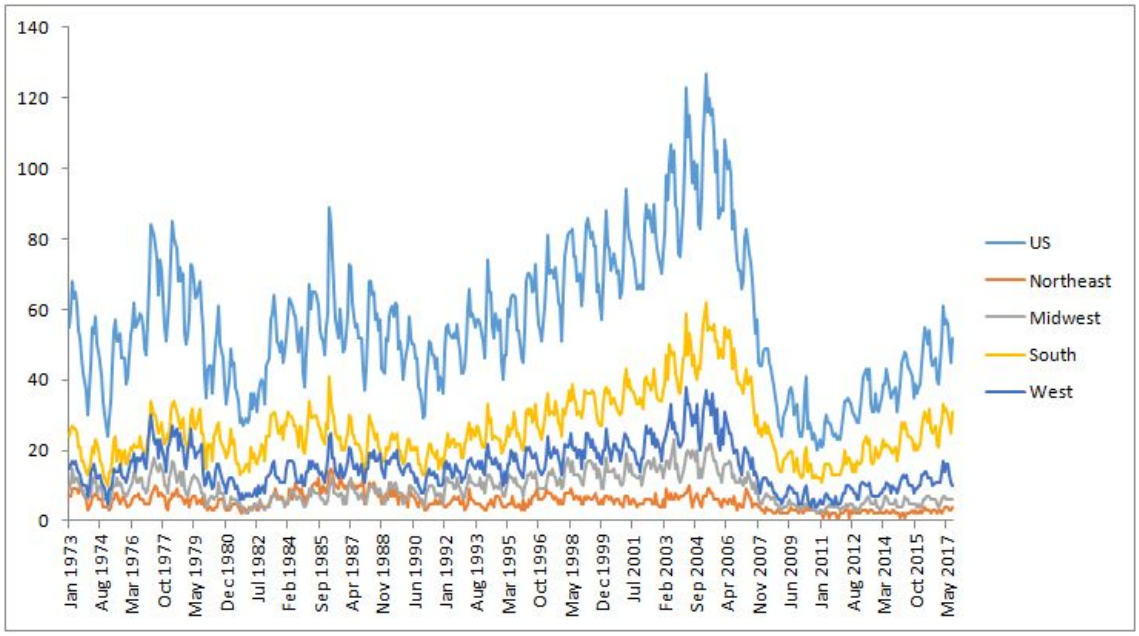
2. *Exploration of US FRED data sets (including Bureau of Labor and Census)*
    a. Collates data sources from numerous other government sources including Bureau of Labor Statistics and Census
    b. Identified numerous data sets concerning employment, price indices, housing vacancy rates, rental amounts, and housing prices/sales. However most of the data is not easily accessible for a time series across locations. Generally if state level data exists, it needs to be downloaded by state and then merged. Since these are ancillary data sets, we have decided to instead focus on our primary data and additional sources that are more easily accessible
    c. However, we did find a couple potentially useful data sets from the US Census that were at the region level (Northeast, Midwest, South, and West) going back to around 1975
        i. In order to use these, we will have to develop a mapping of regions to states in order to tie into Fannie Mae and Freddie Mac data sets
        ii. Data sets are about new single-family homes and are available from https://www.census.gov/construction/nrs/historical_data/index.html
        iii. Median and average price by quarter titled ‘Median and Average Sales Price of Houses Sold by Region’



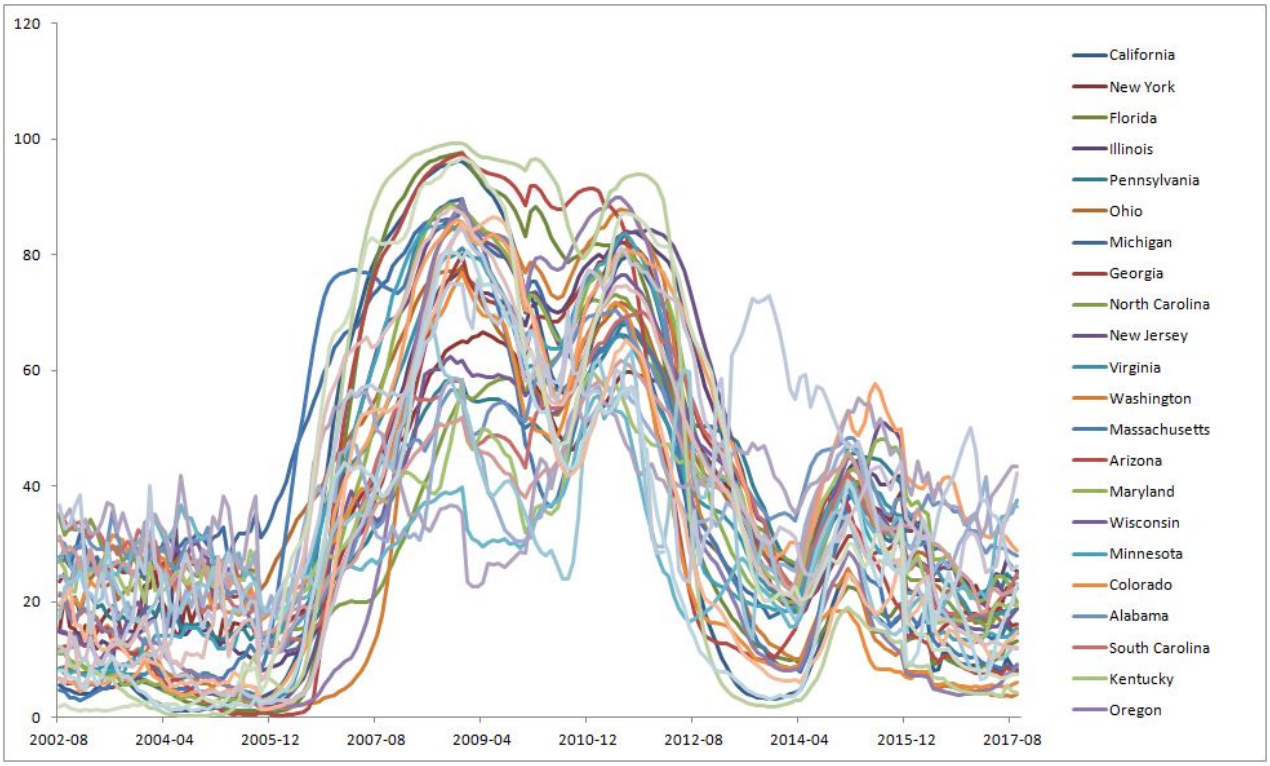
iv. Houses for sale at end of period by month titled 'Houses for Sale'



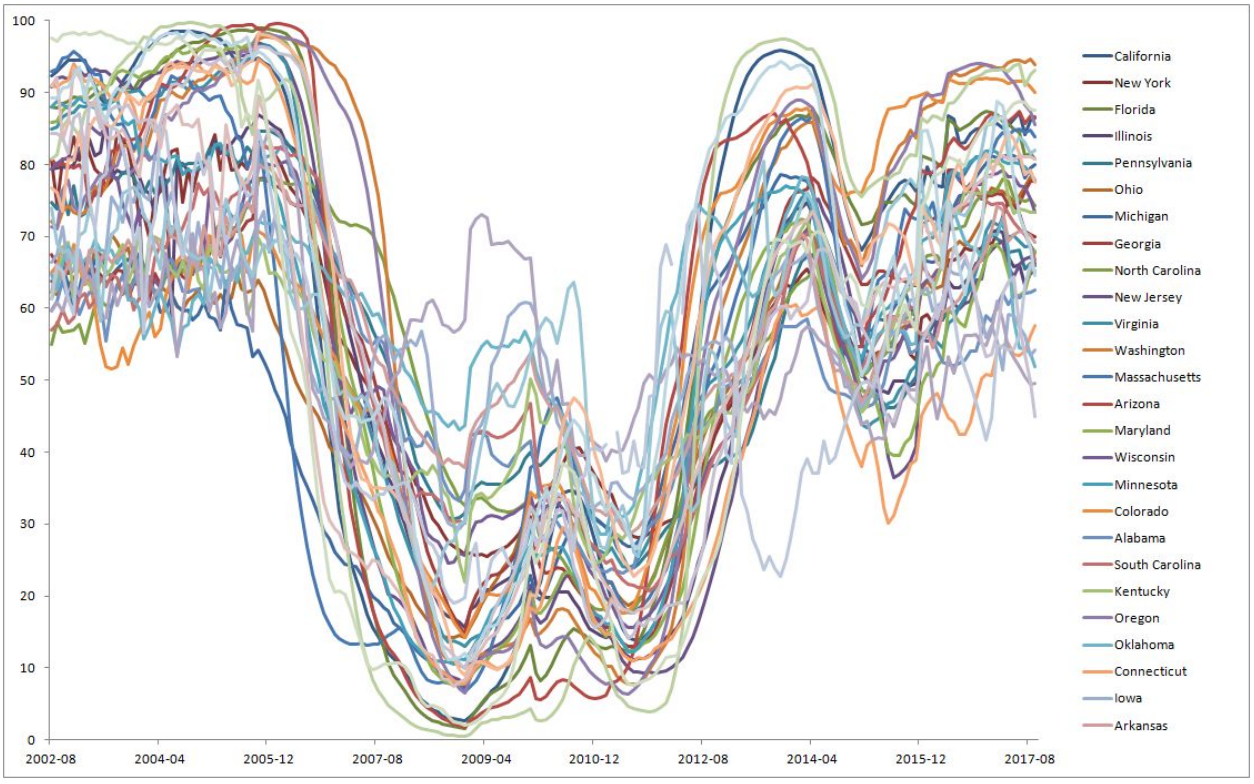
v. House sold during period by month titled ‘Houses Sold’



3. *Exploration of Zillow data sets*
    a. Zillow API allows access to look up current information on individual properties, with access limited to 1000 calls/day. This is not as useful of a dataset for a broad time series analysis, and would also require substantial effort to implement. Thus, we have decided to focus on our primary datasets and ancillary data that has more value for effort
    b. Static data sets available at state, metro, county, city, zip code, or neighborhood level. Recommend state level due to ease of matching to Fannie Mae and Freddie Mac data sets (zip code is too granular to match to three-digit zip in primary data sets)
    c. Static data sets available from: https://www.zillow.com/research/data/
    d. Static data sets recommended for use:
        i. % of homes decreasing in value by state by month, with data available for all states from 2002-08 onwards with only a few missing values for Maine that can be interpolated (listed as ‘Decreasing Values’)



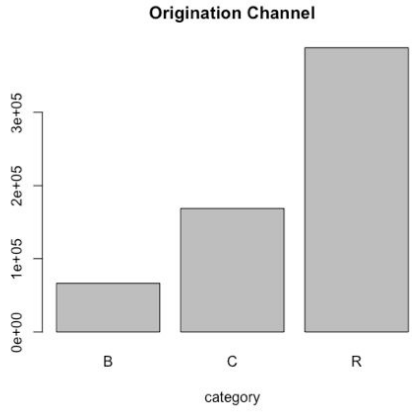
ii. % of home increasing in value by state by month, with data available for all states from 2002-08 onwards with only a few missing values for Maine that can be interpolated (listed as 'Increasing Values')



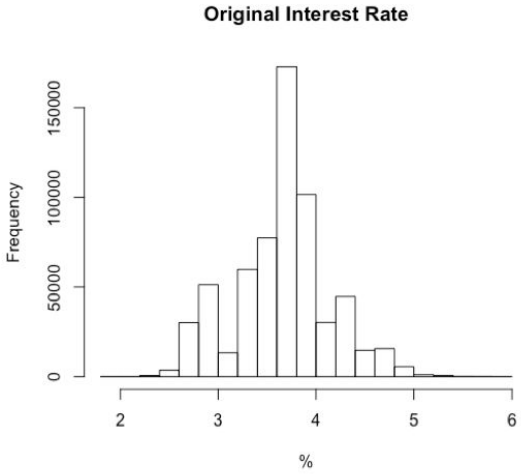
e. Additional interesting data sets exist, but do not go back very far in time or have missing values for several states. These could be brought in for specific purposes, but only after other recommended data sets have been set up
    i. Zillow Rental Index (ZRI) for all homes by state by month, with data available for all states starting in 2011-10 (listed as ZRI Time Series: SFR, Condo/Co-op)
    ii. Zillow Home Value Index (ZHVI) for all homes by state by month, with data available for all states starting in 2009-08 (listed as ZHVI All Homes (SFR, Condo/Co-op) Time Series)
    iii. Price to rent ratio by state by month, with data available for all states starting in 2011-08 (listed as Price-to-Rent Ratio)
    iv. Median listing price by state by month, with data available for all states starting in 2013-11 (listed as Median List Price)
    v. Median rental list price by state by month, with data available for all states starting in 2015-06 (listed as Median Rent List Price ($), SFR, Condo/Co-op)

4. *Exploration of Fannie Mae Acquisitions and Performance Data*

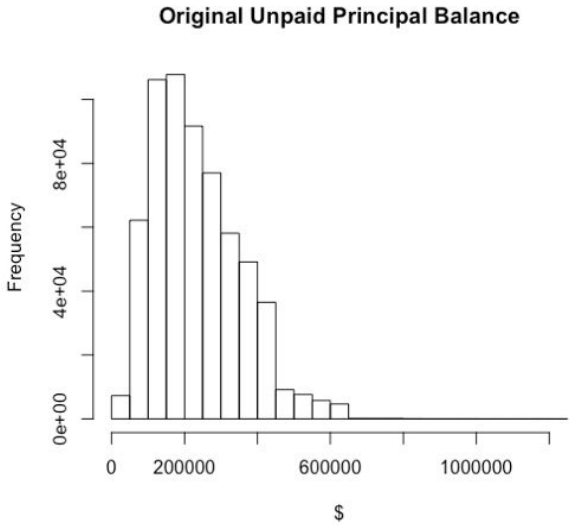
a. 2016Q3 Acquisitions Data: 623,514 observations of 25 variables
    1) Loan Identifier (Number)
    2) Origination Channel (Factor with 3 levels; B (Broker), R (Retail) and C (Correspondent))



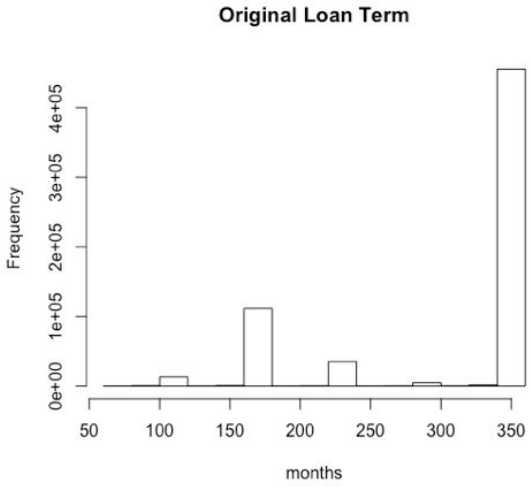
    3) Seller Name (Factor with 21 levels; the overwhelmingly dominant category is “Other”)
    4) Original Interest Rate (Number)



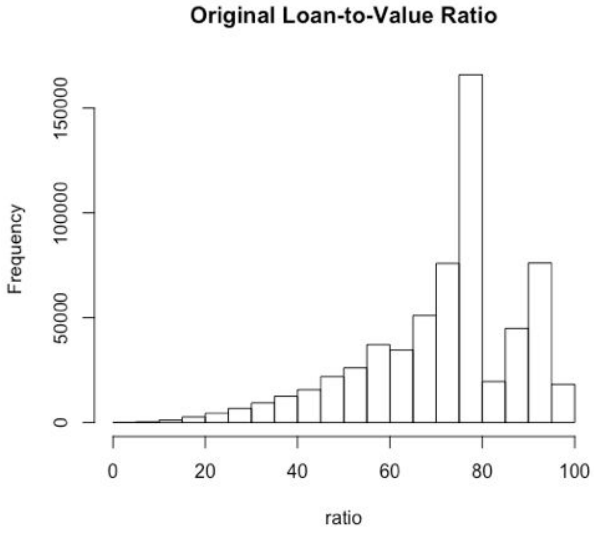
5) Original Unpaid Principal Balance (Number)



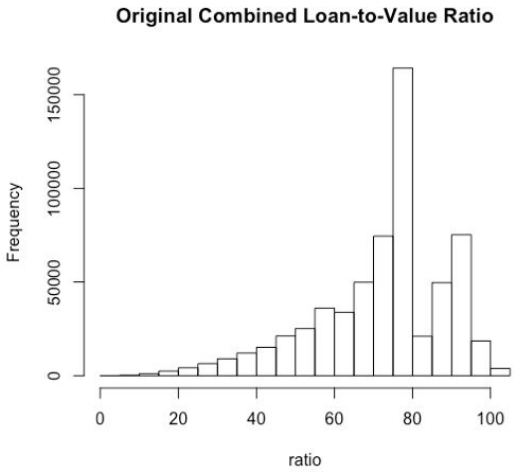
6) Original Loan Term (int number)



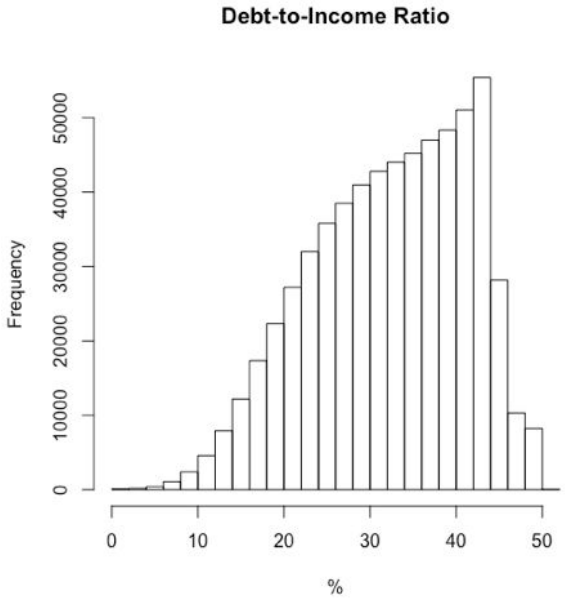
7) Origination date (date; factor with 25 levels)
8) First Payment Date (date; factor with 25 levels; first payments generally happen a month after origination date)
9) Original Loan-to-Value Ratio (LTV) (int number)



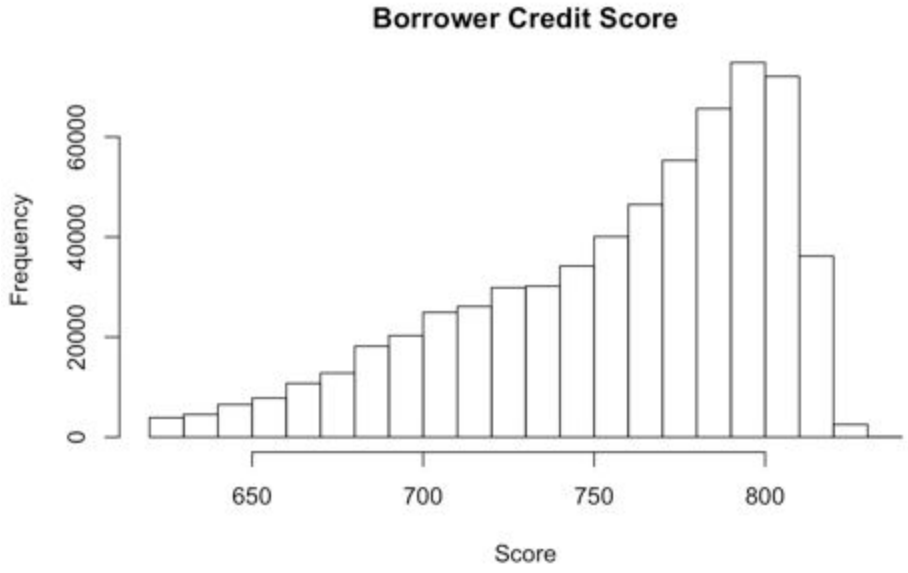
10) Original Combined Loan-to-Value Ratio (int number):“reflects the loan-to-value ratio inclusive of all loans secured by a mortgaged property on the origination date of the underlying mortgage loan.”



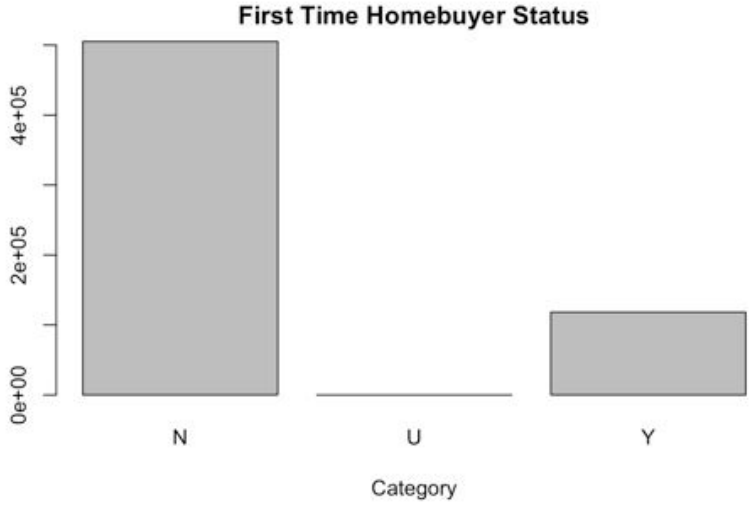
11) Number of Borrowers (integer, roughly 50% 1 and 50% 2)
12) Debt-to-income Ratio (integer)



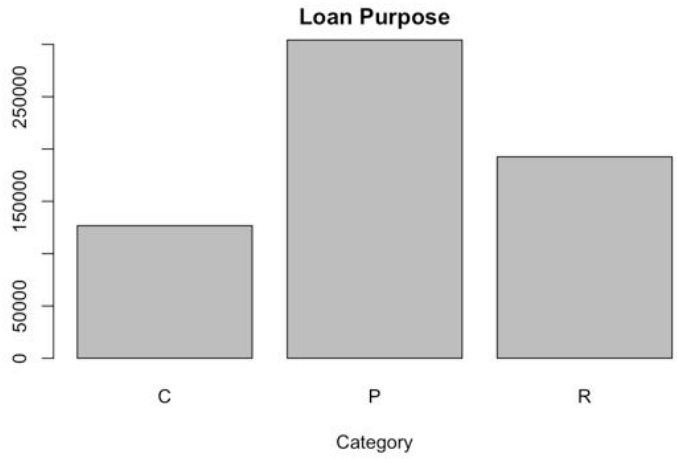
13) Borrower Credit Score (integer)



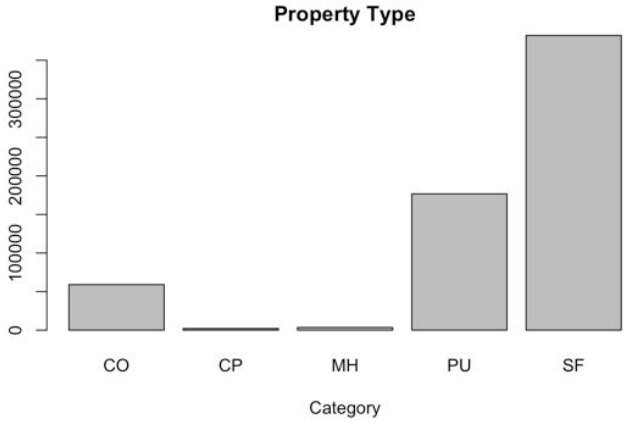
14) First-time Home Buyer Indicator (3-level factor; Y="Yes", N="No", U="Unknown")



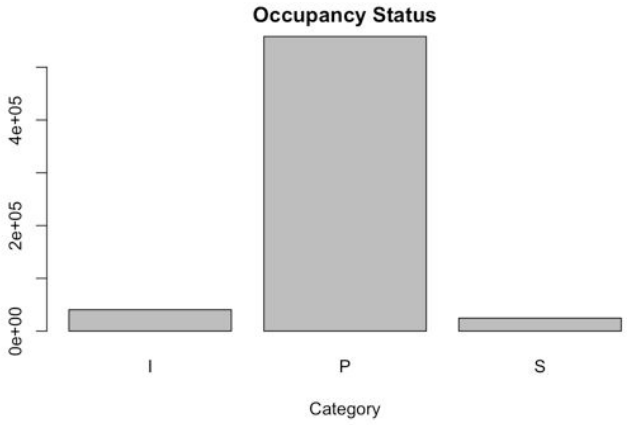
15) Loan Purpose (3-level factor; C = "Cash-out", P = "Purchase", R = "No Cash-out Refinance")



16) Property Type (5-level factor; SF = “Single Family”, CO = “Condo”, CP = “Co-Op”, MH = “Manufactured Housing”, PU = “PUD”)

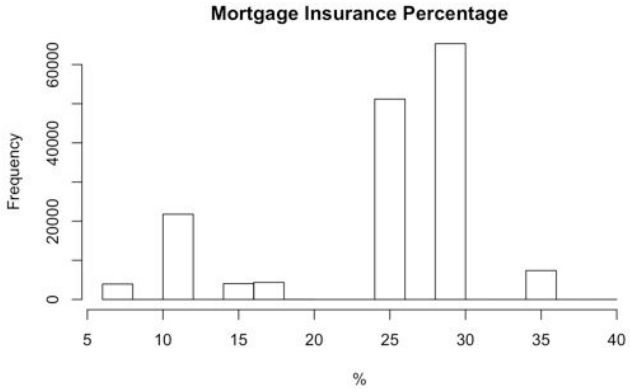


17) Number of Units (integer, almost 100% = 1)
18) Occupancy Status (3-level factor; I = “Investor”, P = “Principal”, S = “Second”)

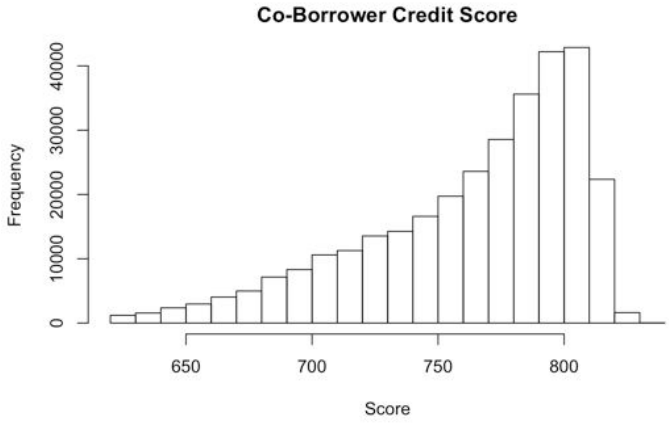


19) Property State (54-level factor, two-letter state codes
20) Zip (3-digit, integer)

21) Mortgage Insurance Percentage (Many NAs = no insurance?)

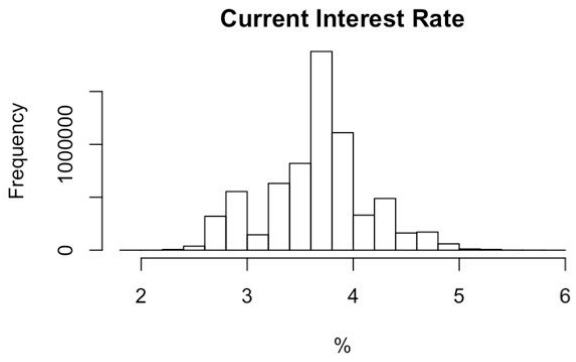


22) Product Type (Factor with only one level, FRM = “Fixed Rate Mortgage Loan”; there are no adjustable rate mortgages)
23) Co-Borrower Credit Score (integer; many NAs = no co-borrower?)

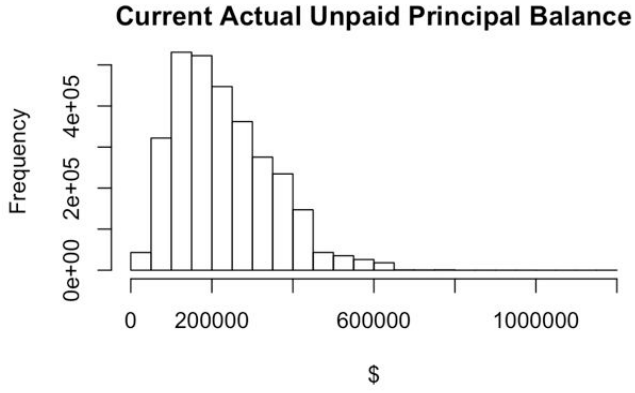


24) Mortgage Insurance Type (source of the Mortgage Insurance premium payment; integer, 1 = Borrower Paid, 2 = Lender Paid, 3 = Investor Paid, many NAs (none), mostly 1s)
25) Relocation Mortgage Indicator (Factor with 2 levels; N = “no”, Y = “yes”, almost 100% = N)

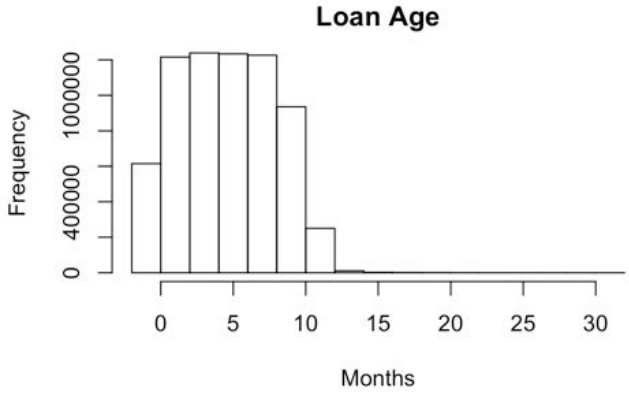
b. 2016Q3 Performance Data: 6,730,844 observations of 31 variables
    1) Loan Identifier (Number)
    2) Monthly Reporting Period (date; factor with 12 levels)
    3) Servicer Name (factor with 22 levels, mostly blank, dominant category is Other)
    4) Current Interest Rate (number)



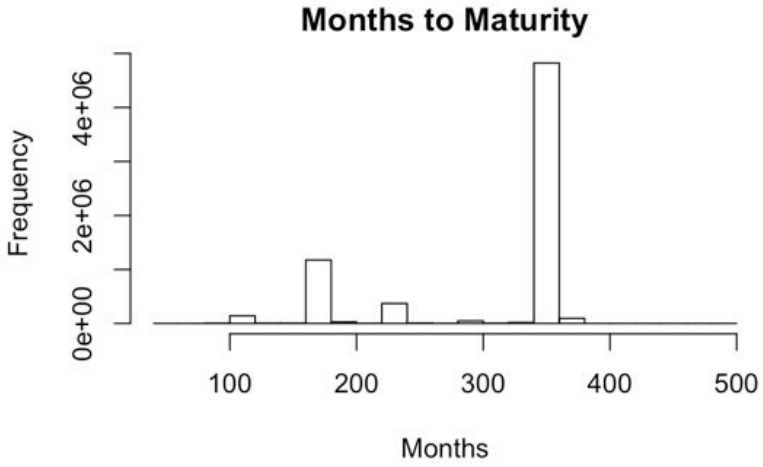
5) Current Actual Unpaid Principal Balance (UPB) (number; many NAs)



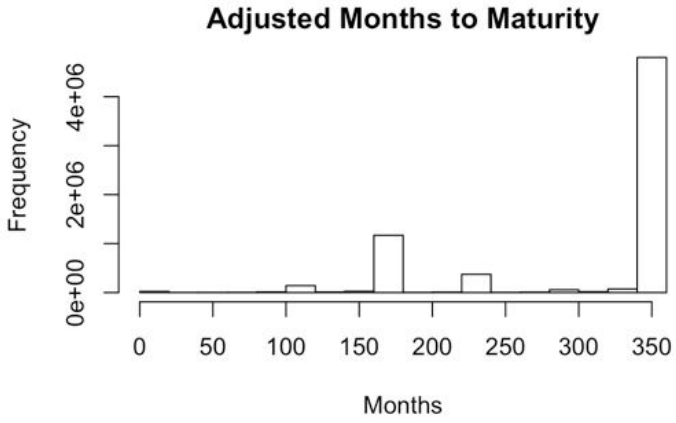
6) Loan Age (integer)



7) Remaining Months to Legal Maturity (integer)

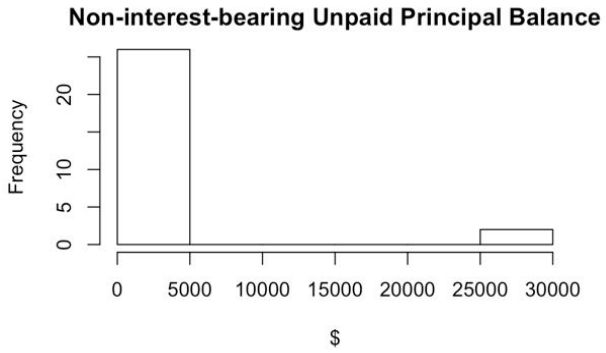


8) Adjusted Remaining Months to Maturity (integer)



9) Maturity Date (date; factor with 242 levels)
10) Metropolitan Statistical Area (MSA) (integer = 5 digit MSA code)
11) Current Loan Delinquency Status (factor with 14 levels 0 = Current, or less than 30 days past due; 1 = 30-59 days; 2 = 60-89 days; 3 = 90-199 days. Mostly at 0)
12) Modification Flag (factor with 2 levels, N = “no”, Y = “yes”, almost 100% = N)
13) Zero Balance Code (code indicating reason mortgage loan balance reduced to zero, mostly NAs; 01 = Prepaid or Matured; 02 = Third Party Sale; 03 = Short Sale; 06 = Repurchased; 09 = Deed-in-Lieu; REO Disposition; 15 = Note Sale; 16 = Reperforming Loan Sale)
14) Zero Balance Effective Date (date; factor with 12 levels)
15) Last Paid Installment Date (blank)
16) Foreclosure Date (blank)

17) Disposition Date (blank)
18) Foreclosure Costs (blank)
19) Property Preservation and Repair Costs (blank)
20) Asset Recovery Costs (blank)
21) Miscellaneous Holding Expenses and Credits (blank)
22) Associated Taxes for Holding Property (blank)
23) Net Sale Proceeds (blank)
24) Credit Enhancement Proceeds (blank)
25) Repurchase Make Whole Proceeds (blank)
26) Other Foreclosure Proceeds (blank)
27) Non Interest Bearing UPB (number, mostly NAs)



28) Principal Forgiveness UPB (blank)
29) Repurchase Make Whole Proceeds Flag (Factor with 3 levels; almost entirely blank “”)
30) Foreclosure Principal Write-off Amount (blank)
31) Servicing Activity Indicator (Factor with 3 levels; “”, N and Y, almost 100% = N)