U.S. Mortgage Performance Before and After the Financial Crisis: Data Wrangling and Cleaning

Data Collection

Data was downloaded at http://www.creditriskanalytics.net/datasets-private.html. Provided freely as a companion to the Wiley text “Credit Risk Analytics” by Baesens, Rosch, and Scheule. Data was downloaded as a .rar file and unzipped into a csv file to be used for the analysis.

Data Cleaning

Since the data in question was used for a text, the authors had already thoroughly cleaned the dataset. However, to ensure this an exploratory analysis of the values was performed. To do this some data wrangling and visualization were used. Prior to any wrangling the dataset was inspected using df.info() to observe the data type and number of non-null observations for each column. The only column will null values is “LTV_time” which represents the loan-to-value ratio at the given point in time. Only 270 out of 622,489 rows (0.04%) have null values, a very low proportion, but maybe not surprising given that the data was already cleaned.Given that such a low proportion of the data is missing values, the missing values will be removed. However, upon inspection the missing values belong to 18 unique mortgages, 17 of which have consecutive mortgages IDs, meaning it’s very unlikely that these are missing completely at random. Possibly these mortgages all belonged to one lender who did not disclose the loan-to-value ratio. This problem is not investigated further, but if access to the original data was available it would be important to look into this more.

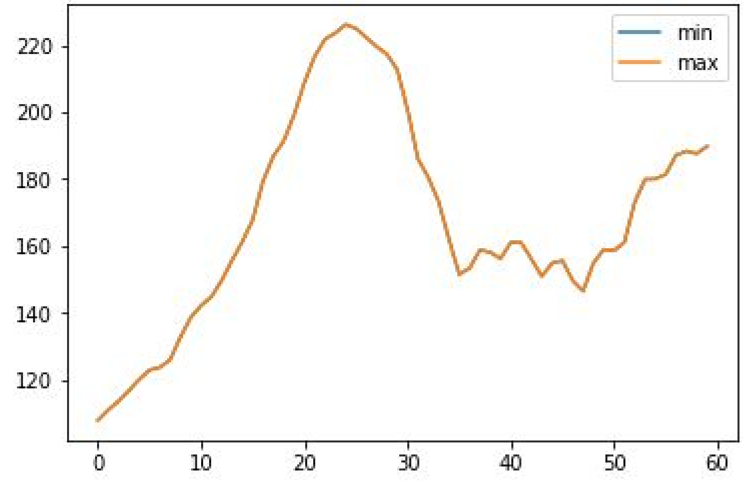
Data Wrangling/Verification

In its original form, the data is set up as a dataframe with 622,489 rows and 23 columns. Each row represents a one period snapshot of an individual mortgage. Nearly all 50,000 unique mortgages has multiple rows corresponding to them since they have more than one period observed, creating a time series component to the data.

To verify there are 50,000 unique mortgages, groupby aggregation using the pandas package was used. It is verified that there are 50,000 mortgages in the dataset prior to the removal of the 18 with missing loan-to-value ratios.

The dataset has a number of macroeconomic variables measured at each period of time, including: GDP, unemployment rate, and national house price index. To ensure data quality, it is verified that these values match each other across different mortgages (i.e. the unemployment rate reported for mortgage X at time period 50 equals the unemployment rate reported for mortgage Y at time period 50). To do this, the dataset is grouped by time period and a min and max are calculated for each macro variable. Then the min and max values are charted as time series using pandas/matplotlib to quickly visualize there are no discrepancies. As an example, the below chart is of the national house price index min and max over time. They are identical for every period in this case and also in the case of GDP and unemployment.

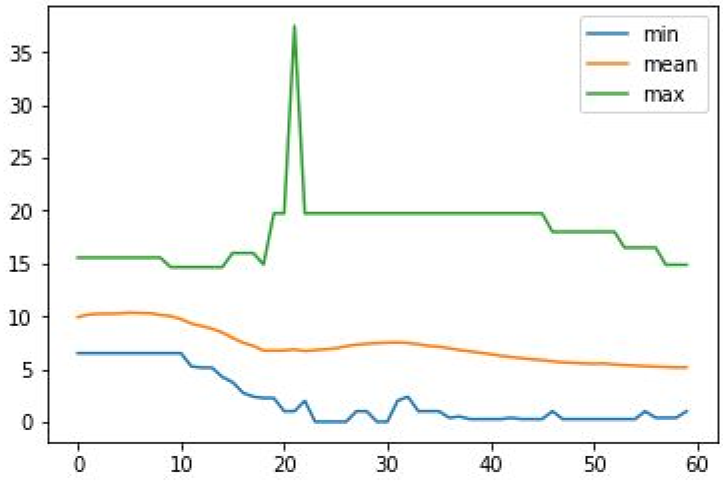
National House Price Index: Min/Max by Period



Since our dataset was collected over a finite period and represents time series we have to be concerned with censored data. Mortgages that were originated prior to the start of our observation window are “left-censored” while mortgages that are not paid off or defaulted on at the end of our observation are “right-censored”. For our purposes, right-censored mortgages are particularly problematic and will bias forecasts of payment/default if not dealt with properly. To count the number of left-censored and right-censored mortgages we use the “default_time” and “payoff_time” columns. These are binary variables provided for each row indicating whether the mortgage was completely paid off or was defaulted on for the given time period. To calculate the number of right-censored observations, the dataframe is grouped by the unique mortgage ID and the two columns are summed. Of the 50,000 mortgages, 15,158 went into default during the 60 period observation window and 26,589 were paid off. That leaves 8,253 mortgages as “right-censored” (50,000 - 15,158 - 26,589 = 8,253). So over 16% of the mortgages are right-censored, a relatively large proportion to drop in order to meet the assumptions for traditional regression analysis. Furthermore, it seems very unlikely that this 16% of the sample is random so removing it would be dangerous.

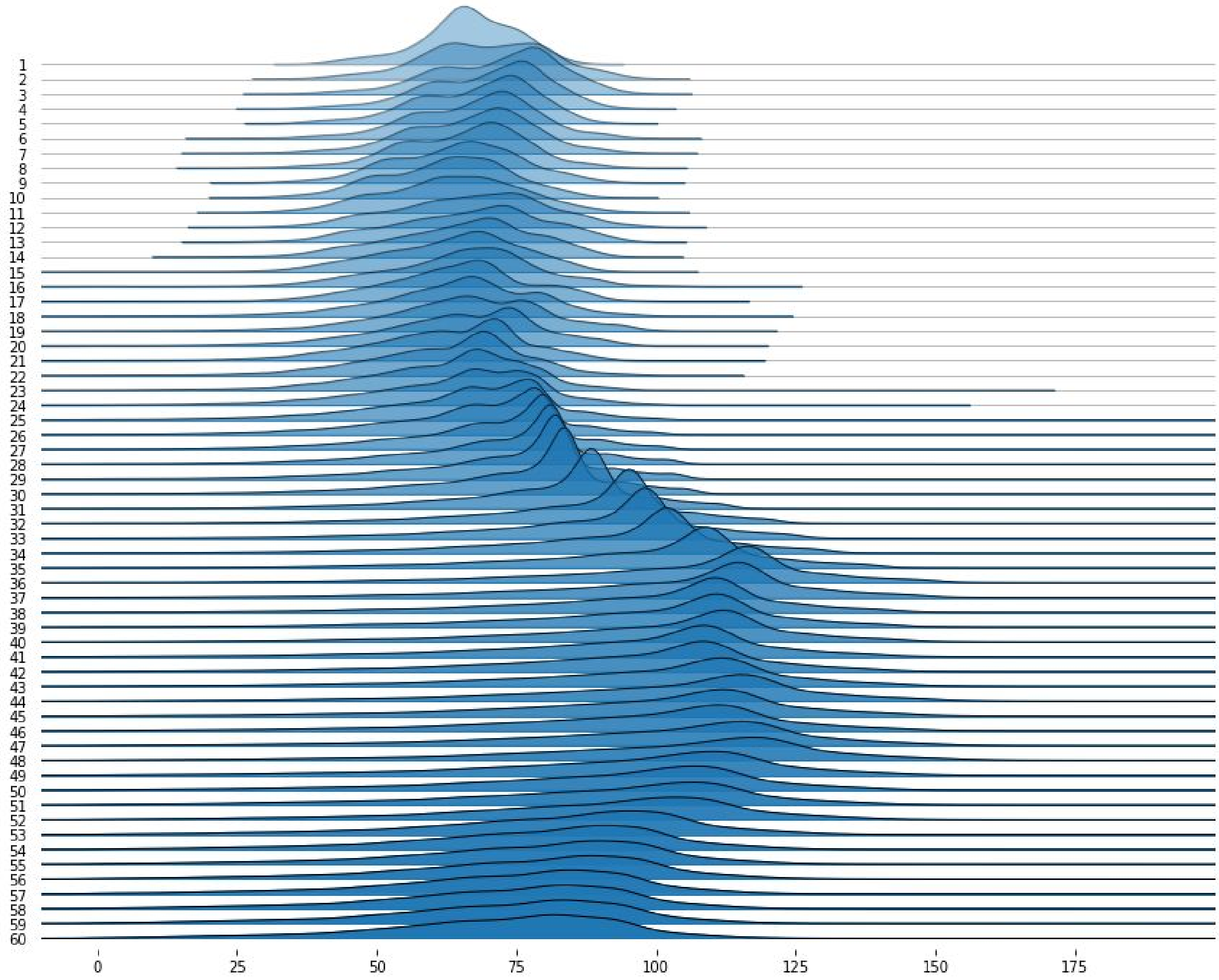
Finally, the variables that are non-binary are investigated for outliers. As an example, the minimum, maximum, and mean interest rate for all reported mortgages is calculated and charted below. Generally a mortgage’s interest rate is persistent across time periods because the loan will not be refinanced regularly due to the prohibitive costs. However, many mortgages do refinance during the lifetime of the loan and so the interest rate is not constant over time for each observed mortgage. There is an interesting dynamic in the dataset easily observed in the chart below where the minimum and mean interest rate trend lower from approximately period 10 through 20 and remain relatively flat thereafter. On the other hand, the maximum increases over the same time period and includes a large spike in period 22, when the maximum increases to a 37.5% interest rate, before falling back to 19.75% the next period. Surprisingly, this interest rate does not correspond to a single mortgage, but multiple all in the same time period. It does not appear to be an error but possibly a number of adjustable-rate mortgages, but this should be looked into further.

Min/Mean/Max Interest Rate Over Time



Another method used to explore the distributions of the data is the joypy package,used to create “joyplots” to visualize how distributions change over time. As an example, the distribution of the ratio of loan-to-value are visualized below. There is an easily observable shift in the distribution across time and it is clear that the distributions have fat tails, which is not surprising given this is a financial time series and will need to be kept in mind during further analysis.

Loan-to-Value Distributions Over Time


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
0
25
50
75
100
125
150
175