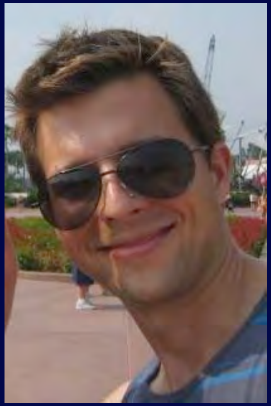
# Visualizing Large Data Sets

## The Housing Crisis

Statistics Department

# Who's who?

Dr. Hadley
Wikcham



Garrett
Grolemund

Dex
Gannon

Gabi
Quart

Barret
Schloerke

# Outline

- Topic: The Housing Crisis
- Collecting and Cleaning Data
- Exploration and Analysis
- Communication

# The Housing Crisis

- Real estate bubble
- Personal
- Little organized public information
- Government expenditures
- Still unfolding
- Affecting global economy

# What we hope to accomplish

- **What** is the housing crisis?
- **Where** has it hit the hardest?
- **When** did start? **When** will it end?
- **Who** does it affect?

# Challenges

- How do we retrieve useful information from large data sets?
- How do we communicate our results?

# Problems with Large Data

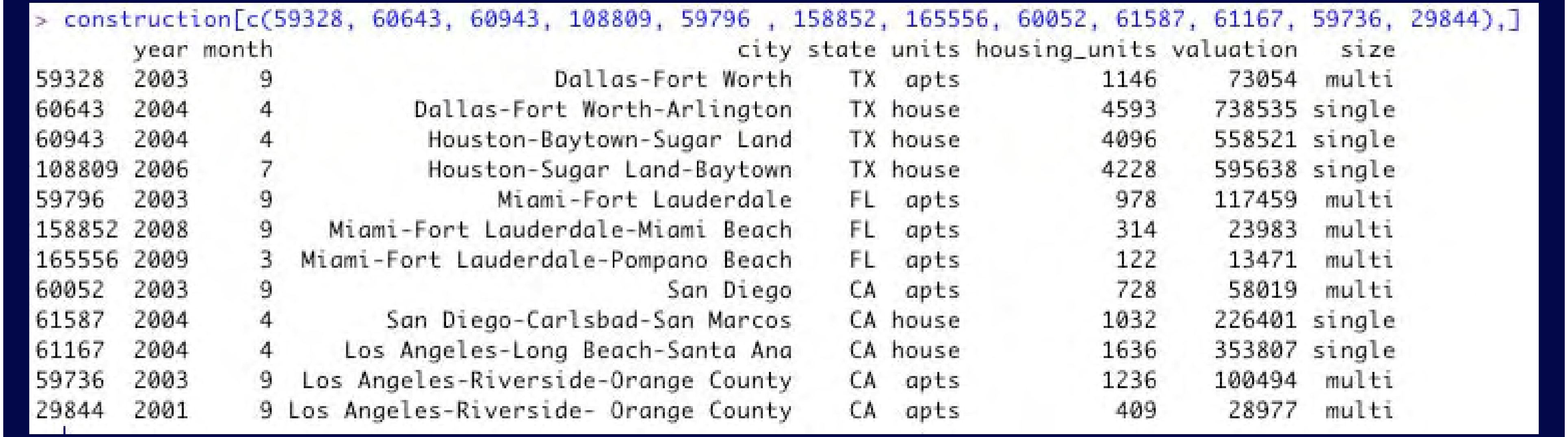
- Hard to find
- Costs money & Licenses
- Big and UGLY
- Dirty - what is clean?

# Clean data is:

- 4 C's
  - –Consistent
  - –Concise
  - –Complete
  - –Correct

Cleaning data sets is hard!

# Consistent



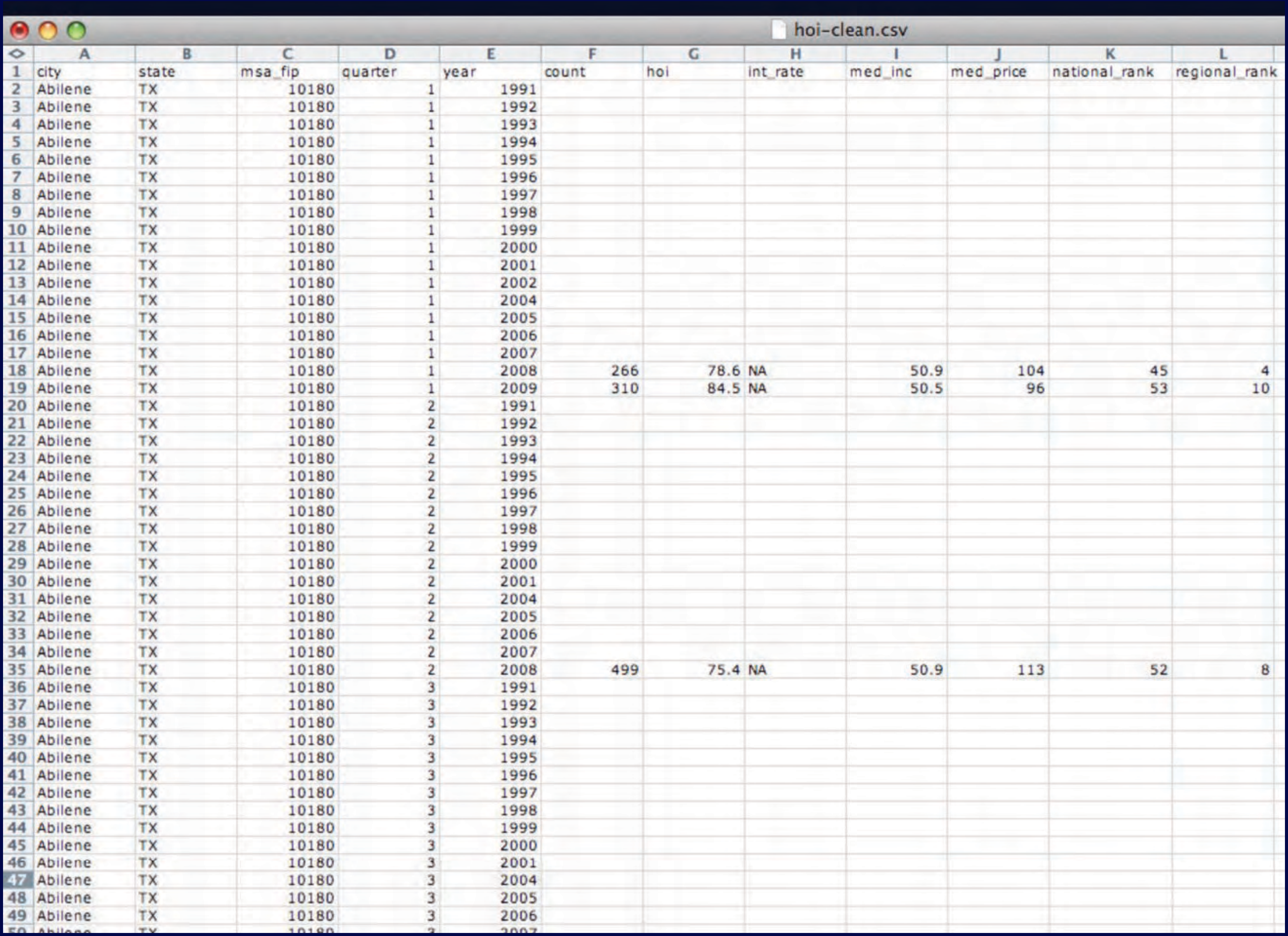
```
> construction[c(59328, 60643, 60943, 108809, 59796 , 158852, 165556, 60052, 61587, 61167, 59736, 29844),]
       year month                                 city state units housing_units valuation   size
59328  2003     9                    Dallas-Fort Worth    TX  apts          1146     73054  multi
60643  2004     4          Dallas-Fort Worth-Arlington    TX house          4593    738535 single
60943  2004     4           Houston-Baytown-Sugar Land    TX house          4096    558521 single
108809 2006     7           Houston-Sugar Land-Baytown    TX house          4228    595638 single
59796  2003     9                Miami-Fort Lauderdale    FL  apts           978    117459  multi
158852 2008     9    Miami-Fort Lauderdale-Miami Beach    FL  apts           314     23983  multi
165556 2009     3  Miami-Fort Lauderdale-Pompano Beach    FL  apts           122     13471  multi
60052  2003     9                            San Diego    CA  apts           728     58019  multi
61587  2004     4        San Diego-Carlsbad-San Marcos    CA house          1032    226401 single
61167  2004     4     Los Angeles-Long Beach-Santa Ana    CA house          1636    353807 single
59736  2003     9  Los Angeles-Riverside-Orange County    CA  apts          1236    100494  multi
29844  2001     9 Los Angeles-Riverside- Orange County    CA  apts           409     28977  multi
```

# Concise

```
20708701220101000000010100000000866740600000000100050001001000020000000200000002000000020000000200000002000000020000000200000002000000020000000100000
102000000020000000200000002000000020000000200000002000000000150100001500000000100720001053394100000000000000000000158105540000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
307087410100025250016430110800200110490110001010000103474000000000001053390036000000000000000000-010000640159101000000000000000003000000000000000010
002111101033021001100000523000000250220000164015901178115150232080011000101011151005000112000001005000100000101000020000002000000020000000200000000200
000200000200000000200000000000200000000000000000000222000000000000000000000200000000000000000001000102000000200000200000000020000020000002000000020000
000015010000150000000000101074222200000000120220000200004000000000000000000000000000000000000000000000000000000000000000000000000000000000001000000
000000000000000000000000000000000000000000000000000111100000000000000000500015810000005150191200000175910258100001054115100 2-1042-104005005705705 7
001-1-1-1-102000002000000002000000022200200000200000000000000000000000010000000000000000000000000000001000001110101111100000
10921200109012000001010170303101-1110111123000000000000001400020202000000000200000000200001001396020000000200000002000000020000000100033902000000020
000002000000020000000200000002000000010001982100007002000000020000000200000002000000020000000200000000020032000139600000607209122120020500000000954
840000000000000000000000000000000000000188606975072203099048010970000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
209212013201010000000101000000007990409000000010013960200000002000000020000000200000001000339020000000200000002000000020000000200000002000000010 0019
821000070020000000200000002000000020000000200000002000000000200320001396000000607209200009548441000000000262000000000155605540000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
309212410100026740026400110800200110490150001010000093137000000000009548400090200000000000000000-010000880186101000000000000000008000000000000000010
004111002045021201100000523000000240211840188018601178115160244130011000101011111001960111000001013960200000020000000200000002000000020000000200000 0103
390200000200000000200000000000200000000000000000000222000000000000000000000200000000000000000001019821000700200000200000000020000020000002000000020000
000020032000139600000607210941221200000001112100002000040000000000000000000000000000000000000000000000000000000000000000000000000000000000001000000
000000000000000000000000000000000000000000000000000111100026200000000005000172300000050801067000001958802946000011488151007210 42-1050008057057057
001-1-1-1-102000002000000002000000011200200000200000000000000000000000300000000200000000000000001200000000001110111111100000
109721001-1012000001010170303101-1110111123000000000000001400020202000000000200000000200001004600020000000200000002000000020000000200000002000000020
000002000000020000000200000002000000010000072200000000200000002000000020000000200000002000000020000000000460720004600000000721922212002120000000861
970000000000000000000000000000000000000002600041850706208310522083000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
209721013201010000000101000000008667414000000001004600020000000200000002000000020000000200000002000000020000000200000002000000020000000200000001000000
722000000020000000200000002000000020000000200000002000000000046072000460000000072192000086197410000000000000000000155605540000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
309721410100025160026430110800200110490110001010000085375000000000008619700560000000000000000000-010000831095101000000000000000004000000000000000010
002111101641020801100000523000000160220000183109501178115130240130011000101011151027000111000001046000200000020000000200000002000000020000000200
000200000200000000200000000000200000000000000000000222000000000000000000000200000000000000000001000722000000200000200000000020000020000002000000020000
```

208.9 MB of this … or in other terms, 69,209 printed pages

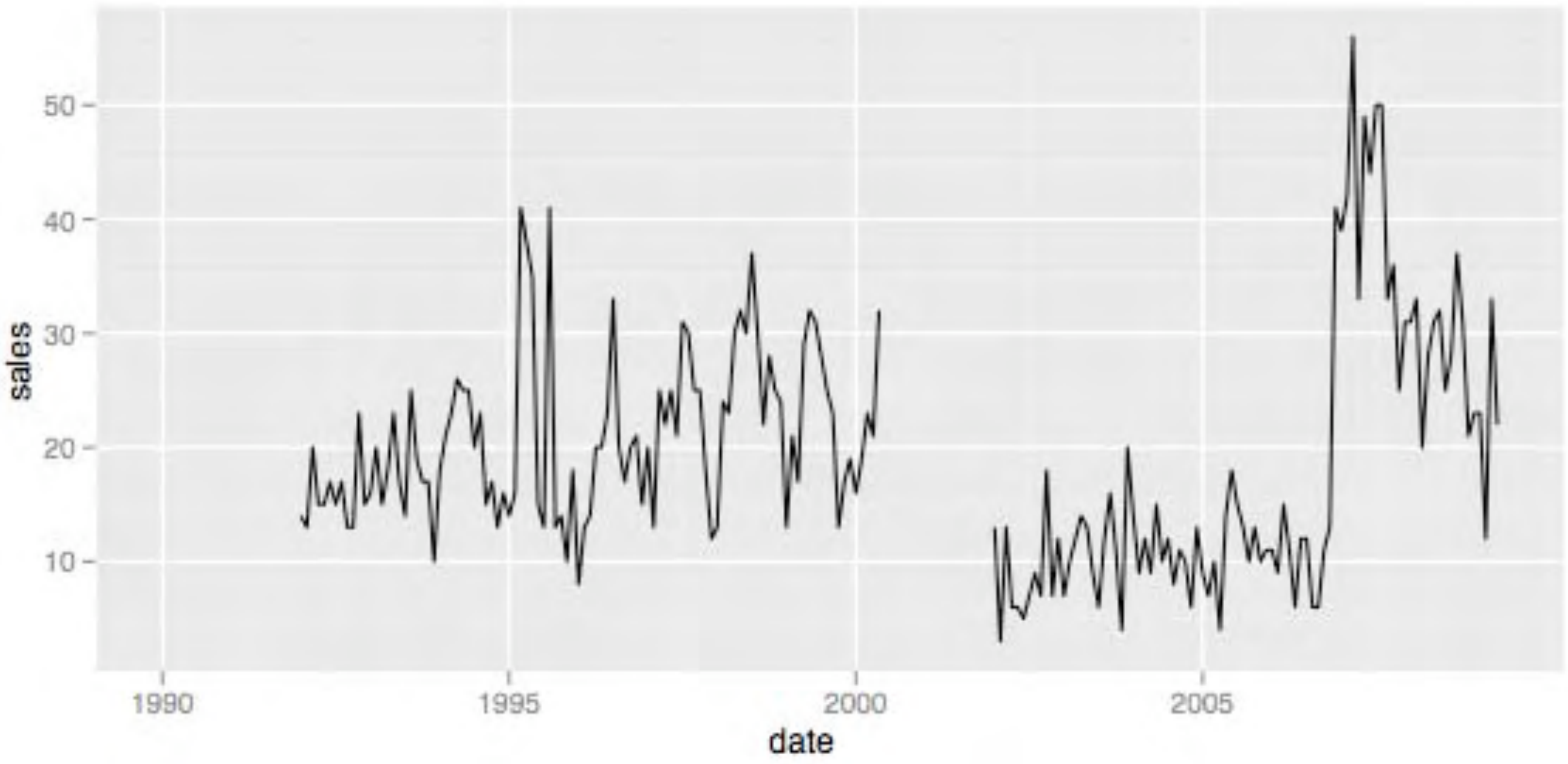
# Complete



hoi-clean.csv

| ◇ | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | city | state | msa_fip | quarter | year | count | hoi | int_rate | med_inc | med_price | national_rank | regional_rank |
| 2 | Abilene | TX | 10180 | 1 | 1991 | | | | | | | |
| 3 | Abilene | TX | 10180 | 1 | 1992 | | | | | | | |
| 4 | Abilene | TX | 10180 | 1 | 1993 | | | | | | | |
| 5 | Abilene | TX | 10180 | 1 | 1994 | | | | | | | |
| 6 | Abilene | TX | 10180 | 1 | 1995 | | | | | | | |
| 7 | Abilene | TX | 10180 | 1 | 1996 | | | | | | | |
| 8 | Abilene | TX | 10180 | 1 | 1997 | | | | | | | |
| 9 | Abilene | TX | 10180 | 1 | 1998 | | | | | | | |
| 10 | Abilene | TX | 10180 | 1 | 1999 | | | | | | | |
| 11 | Abilene | TX | 10180 | 1 | 2000 | | | | | | | |
| 12 | Abilene | TX | 10180 | 1 | 2001 | | | | | | | |
| 13 | Abilene | TX | 10180 | 1 | 2002 | | | | | | | |
| 14 | Abilene | TX | 10180 | 1 | 2004 | | | | | | | |
| 15 | Abilene | TX | 10180 | 1 | 2005 | | | | | | | |
| 16 | Abilene | TX | 10180 | 1 | 2006 | | | | | | | |
| 17 | Abilene | TX | 10180 | 1 | 2007 | | | | | | | |
| 18 | Abilene | TX | 10180 | 1 | 2008 | 266 | 78.6 | NA | 50.9 | 104 | 45 | 4 |
| 19 | Abilene | TX | 10180 | 1 | 2009 | 310 | 84.5 | NA | 50.5 | 96 | 53 | 10 |
| 20 | Abilene | TX | 10180 | 2 | 1991 | | | | | | | |
| 21 | Abilene | TX | 10180 | 2 | 1992 | | | | | | | |
| 22 | Abilene | TX | 10180 | 2 | 1993 | | | | | | | |
| 23 | Abilene | TX | 10180 | 2 | 1994 | | | | | | | |
| 24 | Abilene | TX | 10180 | 2 | 1995 | | | | | | | |
| 25 | Abilene | TX | 10180 | 2 | 1996 | | | | | | | |
| 26 | Abilene | TX | 10180 | 2 | 1997 | | | | | | | |
| 27 | Abilene | TX | 10180 | 2 | 1998 | | | | | | | |
| 28 | Abilene | TX | 10180 | 2 | 1999 | | | | | | | |
| 29 | Abilene | TX | 10180 | 2 | 2000 | | | | | | | |
| 30 | Abilene | TX | 10180 | 2 | 2001 | | | | | | | |
| 31 | Abilene | TX | 10180 | 2 | 2004 | | | | | | | |
| 32 | Abilene | TX | 10180 | 2 | 2005 | | | | | | | |
| 33 | Abilene | TX | 10180 | 2 | 2006 | | | | | | | |
| 34 | Abilene | TX | 10180 | 2 | 2007 | | | | | | | |
| 35 | Abilene | TX | 10180 | 2 | 2008 | 499 | 75.4 | NA | 50.9 | 113 | 52 | 8 |
| 36 | Abilene | TX | 10180 | 3 | 1991 | | | | | | | |
| 37 | Abilene | TX | 10180 | 3 | 1992 | | | | | | | |
| 38 | Abilene | TX | 10180 | 3 | 1993 | | | | | | | |
| 39 | Abilene | TX | 10180 | 3 | 1994 | | | | | | | |
| 40 | Abilene | TX | 10180 | 3 | 1995 | | | | | | | |
| 41 | Abilene | TX | 10180 | 3 | 1996 | | | | | | | |
| 42 | Abilene | TX | 10180 | 3 | 1997 | | | | | | | |
| 43 | Abilene | TX | 10180 | 3 | 1998 | | | | | | | |
| 44 | Abilene | TX | 10180 | 3 | 1999 | | | | | | | |
| 45 | Abilene | TX | 10180 | 3 | 2000 | | | | | | | |
| 46 | Abilene | TX | 10180 | 3 | 2001 | | | | | | | |
| 47 | Abilene | TX | 10180 | 3 | 2004 | | | | | | | |
| 48 | Abilene | TX | 10180 | 3 | 2005 | | | | | | | |
| 49 | Abilene | TX | 10180 | 3 | 2006 | | | | | | | |
| 50 | Abilene | TX | 10180 | 3 | 2007 | | | | | | | |

# Correct



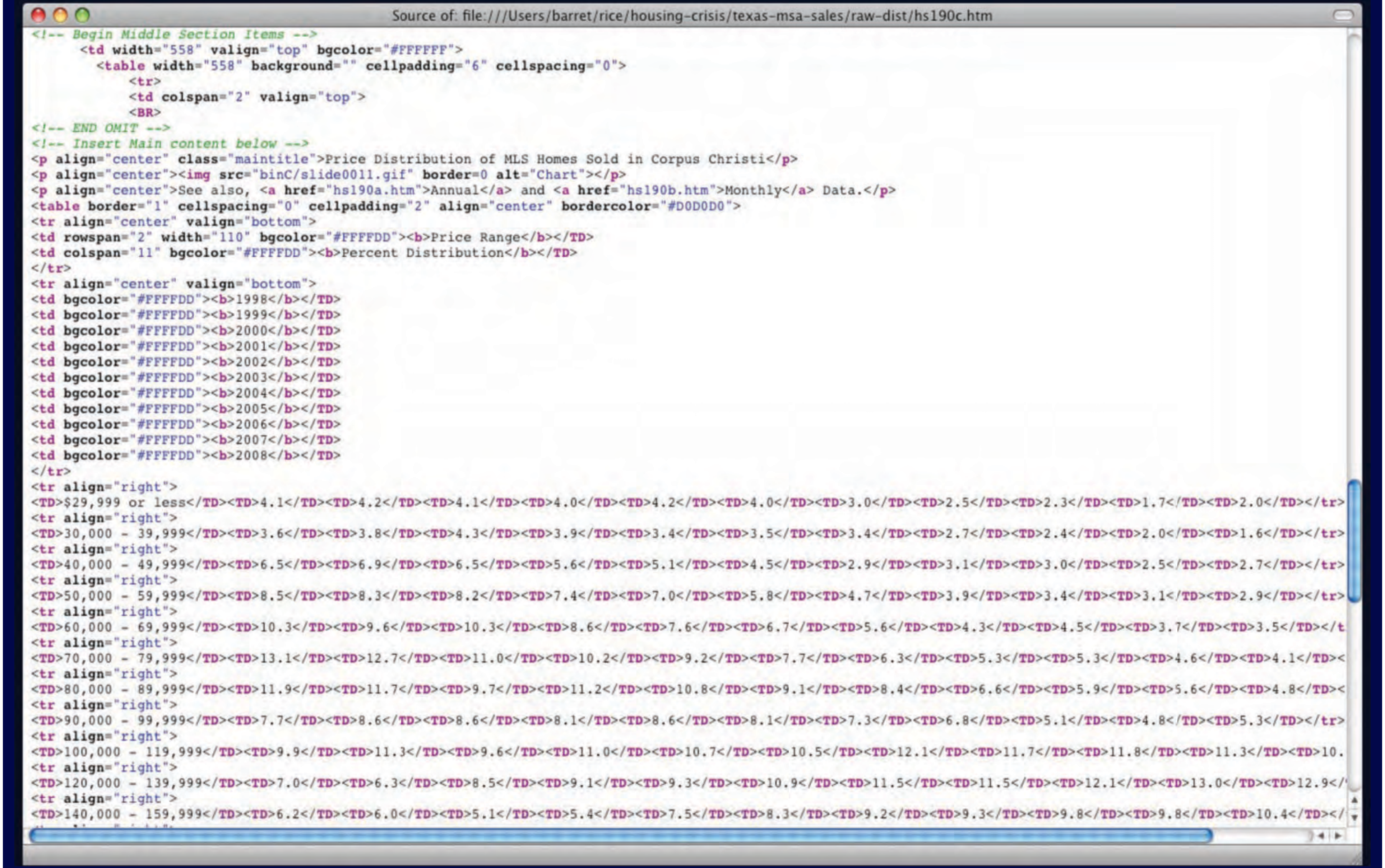
# R

- Programming language similar to Matlab used for statistical computing and graphics
- Used to “clean” data sets

# Why R?

- Statistical standard
- Great graphics
- Data cleaning capabilities
- Open source
  - Necessary for true reproducibility

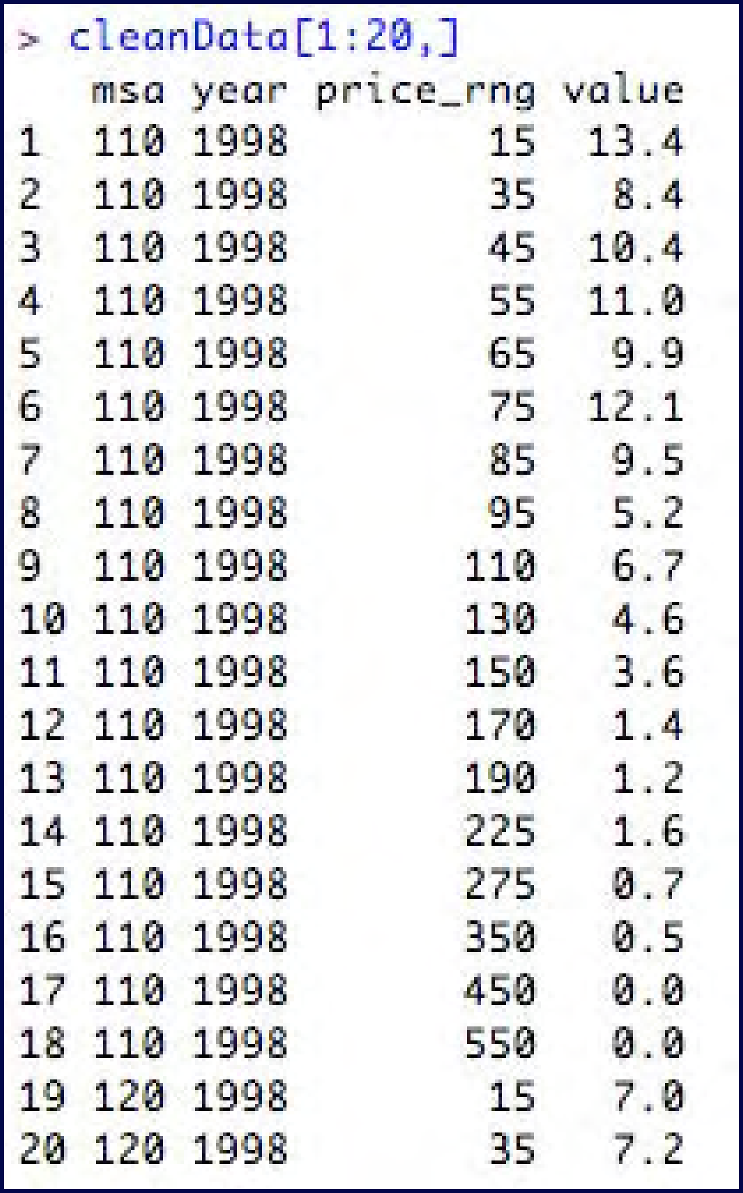
# Dirty …



Source of: file:///Users/barret/rice/housing-crisis/texas-msa-sales/raw-dist/hs190c.htm

```
<!-- Begin Middle Section Items -->
      <td width="558" valign="top" bgcolor="#FFFFFF">
        <table width="558" background="" cellpadding="6" cellspacing="0">
            <tr>
            <td colspan="2" valign="top">
            <BR>
<!-- END OMIT -->
<!-- Insert Main content below -->
<p align="center" class="maintitle">Price Distribution of MLS Homes Sold in Corpus Christi</p>
<p align="center"><img src="binC/slide0011.gif" border=0 alt="Chart"></p>
<p align="center">See also, <a href="hs190a.htm">Annual</a> and <a href="hs190b.htm">Monthly</a> Data.</p>
<table border="1" cellspacing="0" cellpadding="2" align="center" bordercolor="#D0D0D0">
<tr align="center" valign="bottom">
<td rowspan="2" width="110" bgcolor="#FFFFDD"><b>Price Range</b></TD>
<td colspan="11" bgcolor="#FFFFDD"><b>Percent Distribution</b></TD>
</tr>
<tr align="center" valign="bottom">
<td bgcolor="#FFFFDD"><b>1998</b></TD>
<td bgcolor="#FFFFDD"><b>1999</b></TD>
<td bgcolor="#FFFFDD"><b>2000</b></TD>
<td bgcolor="#FFFFDD"><b>2001</b></TD>
<td bgcolor="#FFFFDD"><b>2002</b></TD>
<td bgcolor="#FFFFDD"><b>2003</b></TD>
<td bgcolor="#FFFFDD"><b>2004</b></TD>
<td bgcolor="#FFFFDD"><b>2005</b></TD>
<td bgcolor="#FFFFDD"><b>2006</b></TD>
<td bgcolor="#FFFFDD"><b>2007</b></TD>
<td bgcolor="#FFFFDD"><b>2008</b></TD>
</tr>
<tr align="right">
<TD>$29,999 or less</TD><TD>4.1</TD><TD>4.2</TD><TD>4.1</TD><TD>4.0</TD><TD>4.2</TD><TD>4.0</TD><TD>3.0</TD><TD>2.5</TD><TD>2.3</TD><TD>1.7</TD><TD>2.0</TD></tr>
<tr align="right">
<TD>30,000 - 39,999</TD><TD>3.6</TD><TD>3.8</TD><TD>4.3</TD><TD>3.9</TD><TD>3.4</TD><TD>3.5</TD><TD>3.4</TD><TD>2.7</TD><TD>2.4</TD><TD>2.0</TD><TD>1.6</TD></tr>
<tr align="right">
<TD>40,000 - 49,999</TD><TD>6.5</TD><TD>6.9</TD><TD>6.5</TD><TD>5.6</TD><TD>5.1</TD><TD>4.5</TD><TD>2.9</TD><TD>3.1</TD><TD>3.0</TD><TD>2.5</TD><TD>2.7</TD></tr>
<tr align="right">
<TD>50,000 - 59,999</TD><TD>8.5</TD><TD>8.3</TD><TD>8.2</TD><TD>7.4</TD><TD>7.0</TD><TD>5.8</TD><TD>4.7</TD><TD>3.9</TD><TD>3.4</TD><TD>3.1</TD><TD>2.9</TD></tr>
<tr align="right">
<TD>60,000 - 69,999</TD><TD>10.3</TD><TD>9.6</TD><TD>10.3</TD><TD>8.6</TD><TD>7.6</TD><TD>6.7</TD><TD>5.6</TD><TD>4.3</TD><TD>4.5</TD><TD>3.7</TD><TD>3.5</TD></t
<tr align="right">
<TD>70,000 - 79,999</TD><TD>13.1</TD><TD>12.7</TD><TD>11.0</TD><TD>10.2</TD><TD>9.2</TD><TD>7.7</TD><TD>6.3</TD><TD>5.3</TD><TD>5.3</TD><TD>4.6</TD><TD>4.1</TD><
<tr align="right">
<TD>80,000 - 89,999</TD><TD>11.9</TD><TD>11.7</TD><TD>9.7</TD><TD>11.2</TD><TD>10.8</TD><TD>9.1</TD><TD>8.4</TD><TD>6.6</TD><TD>5.9</TD><TD>5.6</TD><TD>4.8</TD><
<tr align="right">
<TD>90,000 - 99,999</TD><TD>7.7</TD><TD>8.6</TD><TD>8.6</TD><TD>8.1</TD><TD>8.6</TD><TD>8.1</TD><TD>7.3</TD><TD>6.8</TD><TD>5.1</TD><TD>4.8</TD><TD>5.3</TD></tr>
<tr align="right">
<TD>100,000 - 119,999</TD><TD>9.9</TD><TD>11.3</TD><TD>9.6</TD><TD>11.0</TD><TD>10.7</TD><TD>10.5</TD><TD>12.1</TD><TD>11.7</TD><TD>11.8</TD><TD>11.3</TD><TD>10.
<tr align="right">
<TD>120,000 - 139,999</TD><TD>7.0</TD><TD>6.3</TD><TD>8.5</TD><TD>9.1</TD><TD>9.3</TD><TD>10.9</TD><TD>11.5</TD><TD>11.5</TD><TD>12.1</TD><TD>13.0</TD><TD>12.9</
<tr align="right">
<TD>140,000 - 159,999</TD><TD>6.2</TD><TD>6.0</TD><TD>5.1</TD><TD>5.4</TD><TD>7.5</TD><TD>8.3</TD><TD>9.2</TD><TD>9.3</TD><TD>9.8</TD><TD>9.8</TD><TD>10.4</TD></
```
]

# …to Clean



```
> cleanData[1:20,]
   msa year price_rng value
1  110 1998        15  13.4
2  110 1998        35   8.4
3  110 1998        45  10.4
4  110 1998        55  11.0
5  110 1998        65   9.9
6  110 1998        75  12.1
7  110 1998        85   9.5
8  110 1998        95   5.2
9  110 1998       110   6.7
10 110 1998       130   4.6
11 110 1998       150   3.6
12 110 1998       170   1.4
13 110 1998       190   1.2
14 110 1998       225   1.6
15 110 1998       275   0.7
16 110 1998       350   0.5
17 110 1998       450   0.0
18 110 1998       550   0.0
19 120 1998        15   7.0
20 120 1998        35   7.2
```

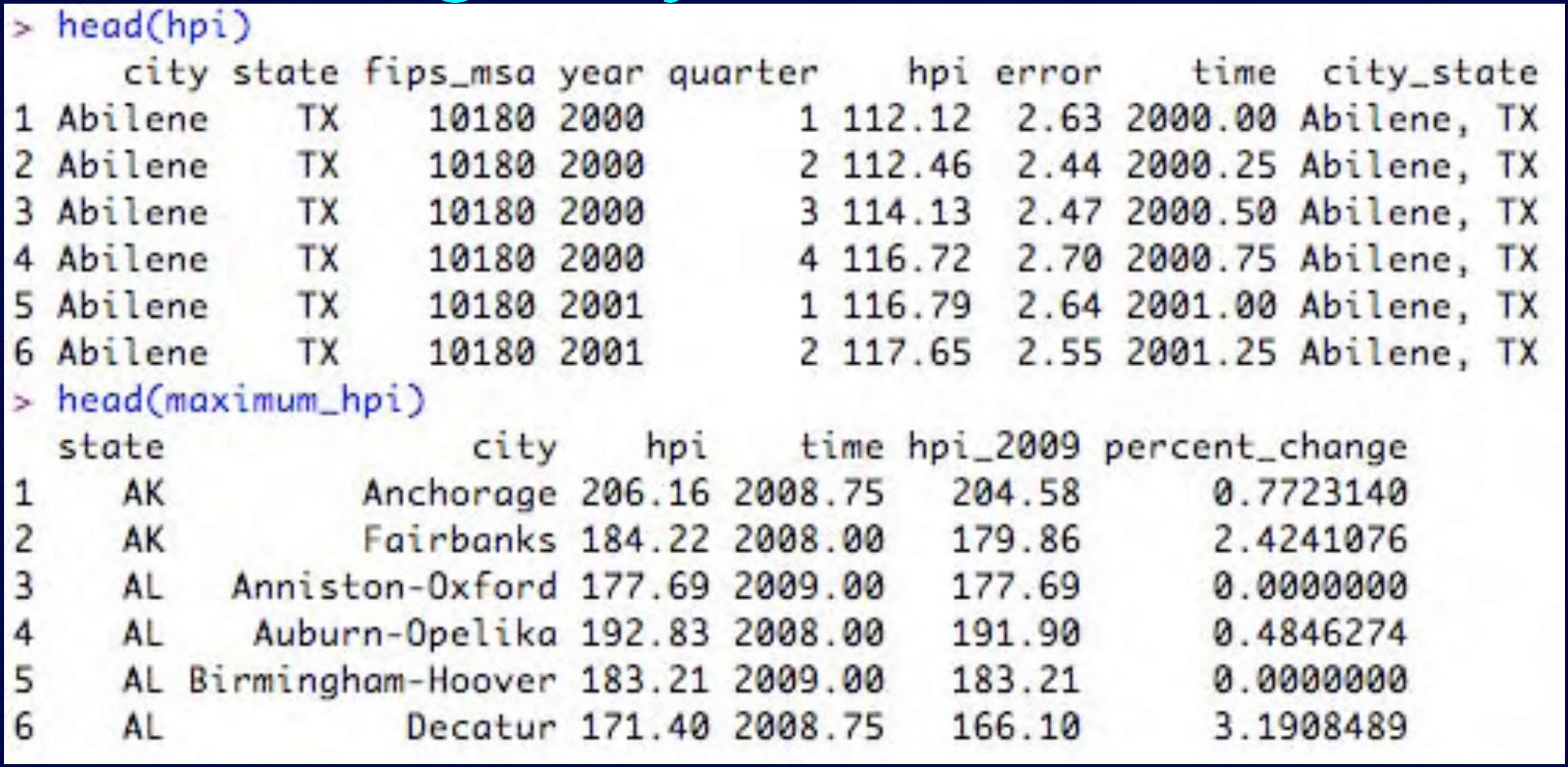
# Our Data

- Construction
- Housing price indexes (HPI)
- Vacancy
- GDP, Retirement, etc.
- Demographic information from the census

# What is a
# Housing Price Index

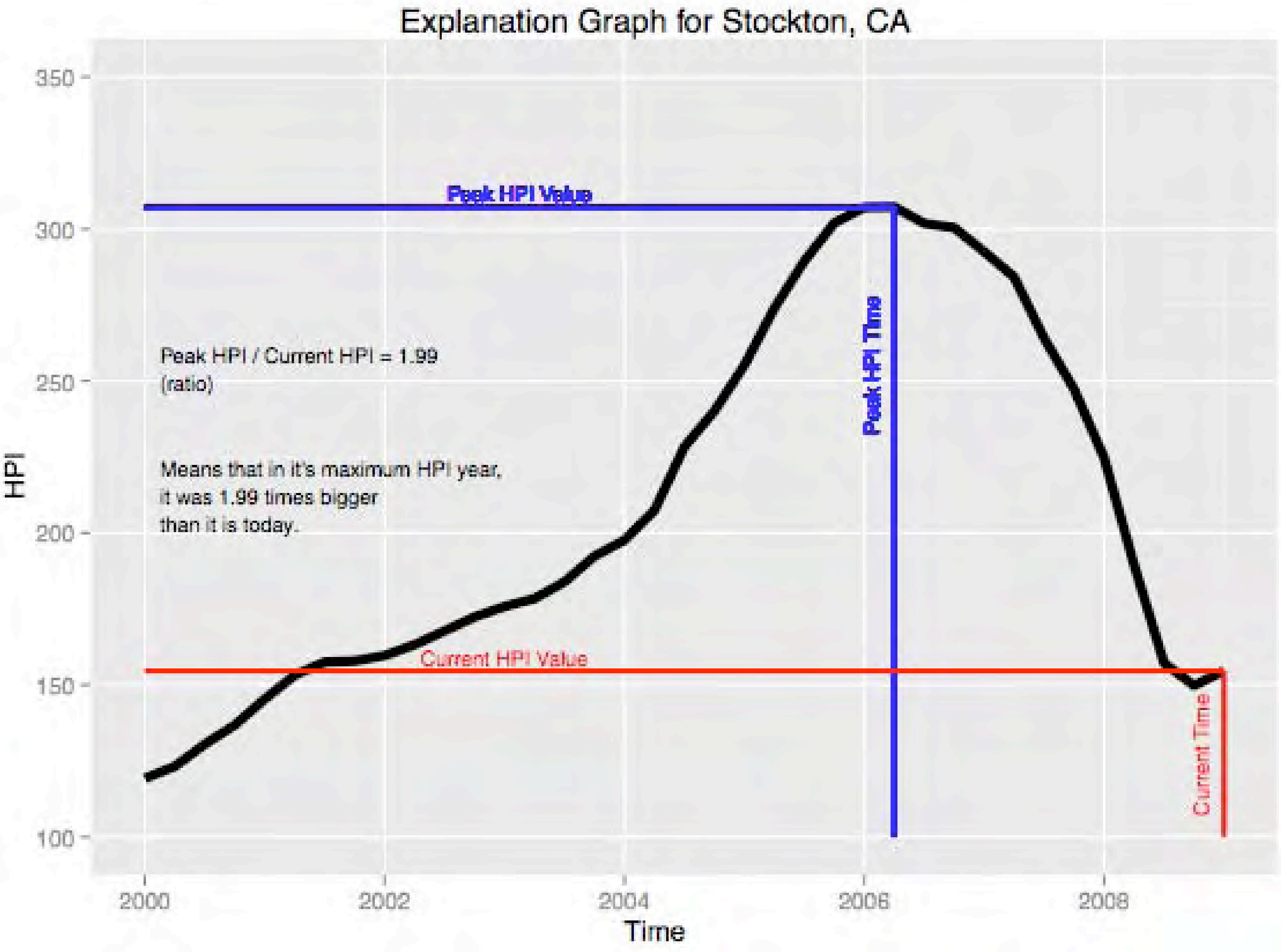
- Definition: Index- scale representing the average value of specified prices as compared with some reference figure
- (HPI Current / HPI index date) * 100
- The HPI is a broad measure of the movement of single-family house prices.

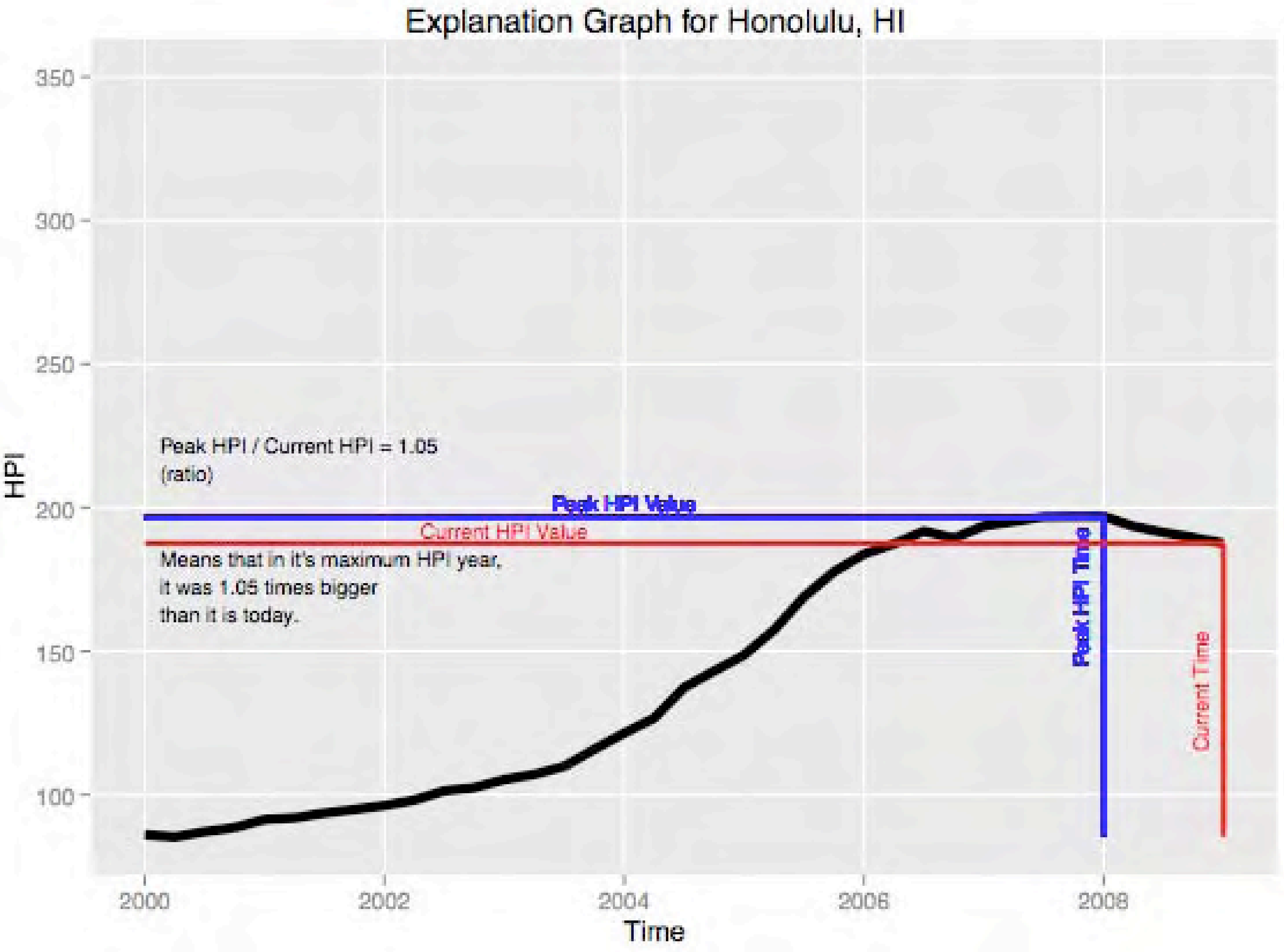
# Federal Housing Finance Agency HPI Data*



```
> head(hpi)
     city state fips_msa year quarter    hpi error    time  city_state
1 Abilene    TX    10180 2000       1 112.12  2.63 2000.00 Abilene, TX
2 Abilene    TX    10180 2000       2 112.46  2.44 2000.25 Abilene, TX
3 Abilene    TX    10180 2000       3 114.13  2.47 2000.50 Abilene, TX
4 Abilene    TX    10180 2000       4 116.72  2.70 2000.75 Abilene, TX
5 Abilene    TX    10180 2001       1 116.79  2.64 2001.00 Abilene, TX
6 Abilene    TX    10180 2001       2 117.65  2.55 2001.25 Abilene, TX
> head(maximum_hpi)
  state              city    hpi    time hpi_2009 percent_change
1    AK         Anchorage 206.16 2008.75   204.58      0.7723140
2    AK         Fairbanks 184.22 2008.00   179.86      2.4241076
3    AL   Anniston-Oxford 177.69 2009.00   177.69      0.0000000
4    AL    Auburn-Opelika 192.83 2008.00   191.90      0.4846274
5    AL Birmingham-Hoover 183.21 2009.00   183.21      0.0000000
6    AL           Decatur 171.40 2008.75   166.10      3.1908489
```

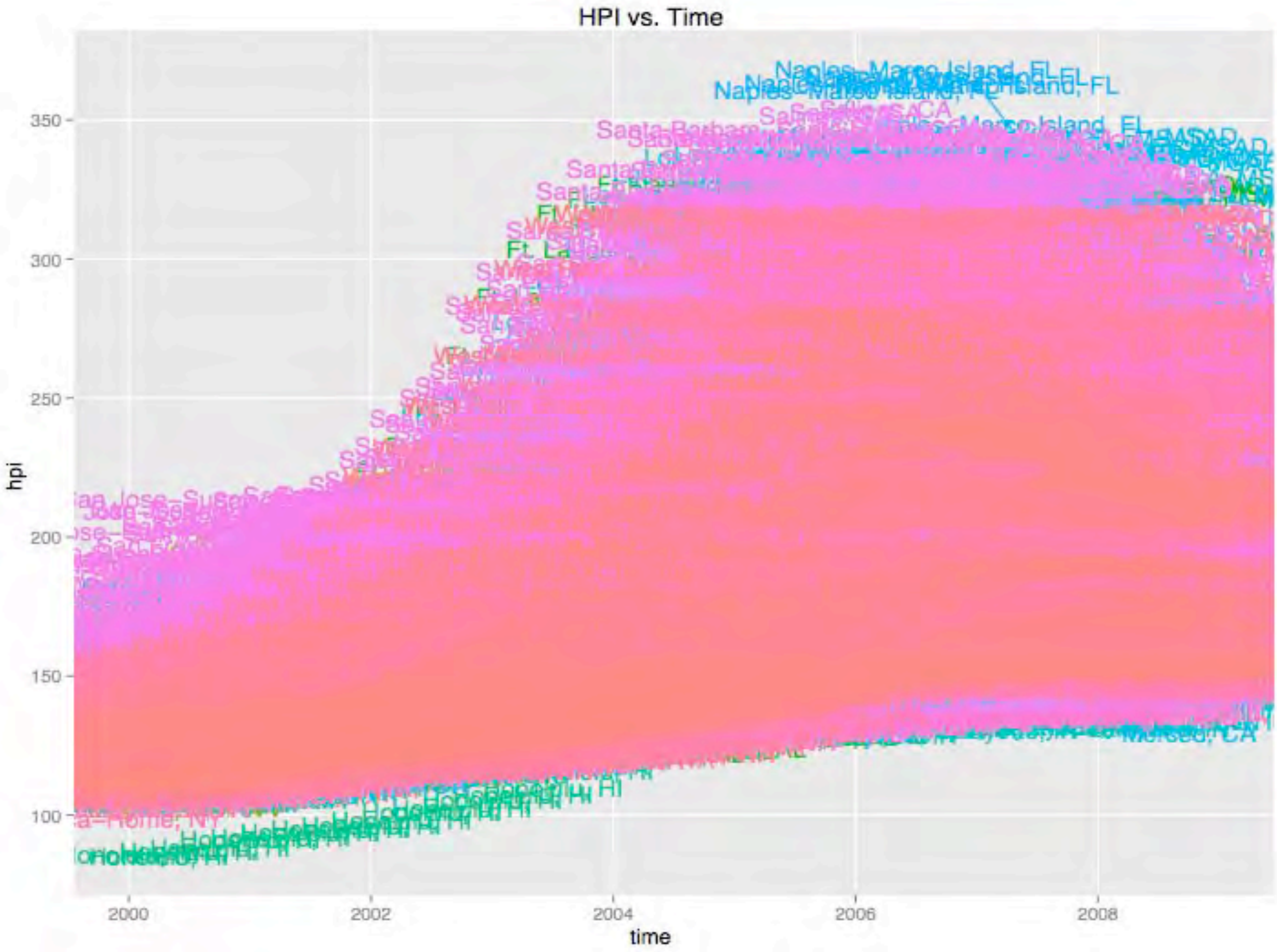
*This information is obtained by reviewing repeat mortgage transactions on single-family properties whose mortgages have been purchased or securitized by Fannie Mae or Freddie Mac.


Explanation Graph for Stockton, CA
Peak HPI Value
Peak HPI Time
Peak HPI / Current HPI = 1.99
(ratio)
Means that in it's maximum HPI year,
it was 1.99 times bigger
than it is today.
Current HPI Value
Current Time
HPI
350
300
250
200
150
100
2000
2002
2004
2006
2008
Time


Explanation Graph for Honolulu, HI
HPI
350
300
250
200
150
100
Time
2000
2002
2004
2006
2008
Peak HPI / Current HPI = 1.05
(ratio)
Peak HPI Value
Current HPI Value
Means that in it's maximum HPI year,
it was 1.05 times bigger
than it is today.
Peak HPI Time
Current Time

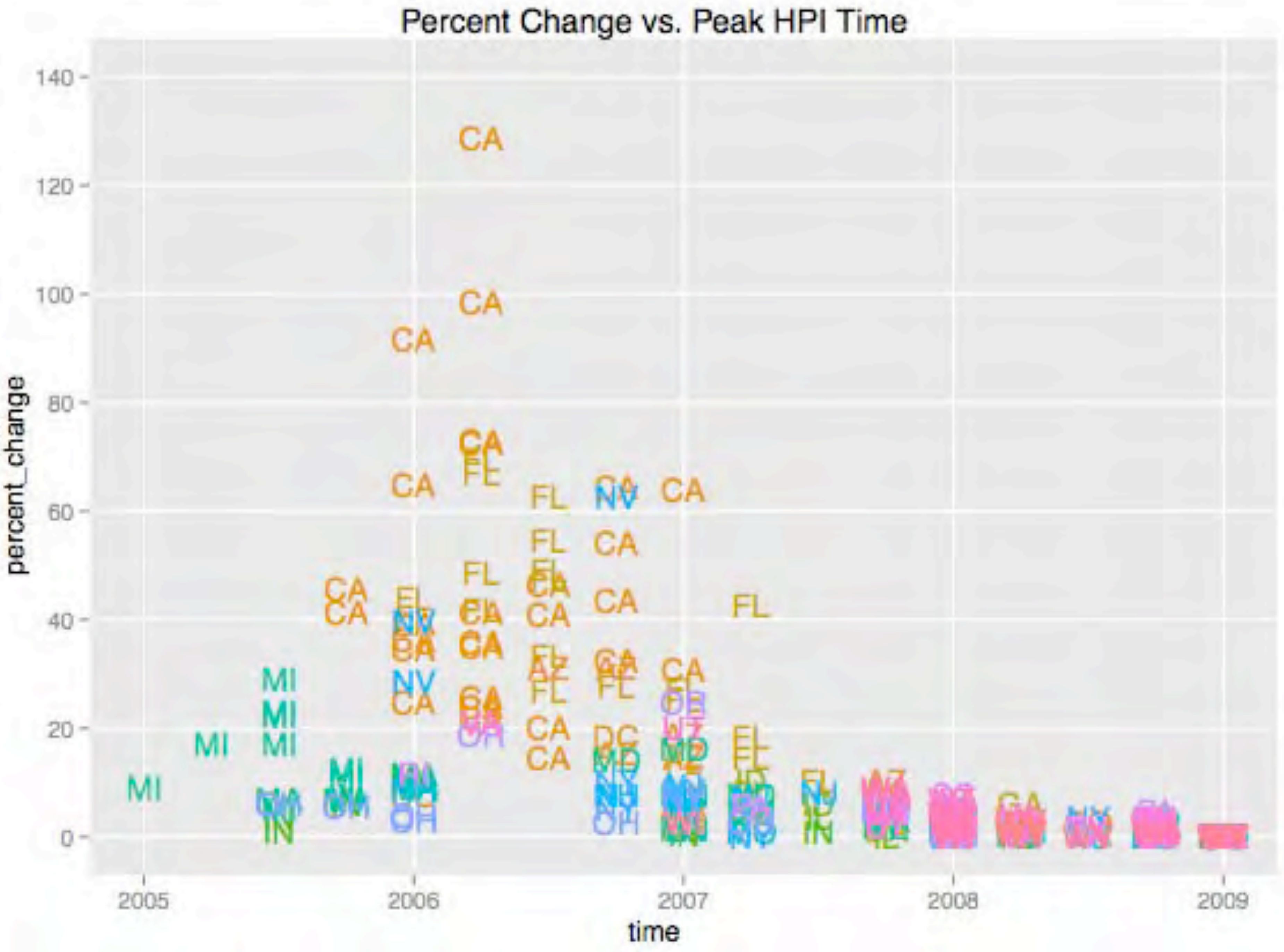
# Exploration and Analysis

- Few preconceived notions
  - Follow the data
  - Relate multiple data sets
- Size of data is overwhelming
- Start small!
  - Start with a city then build from there

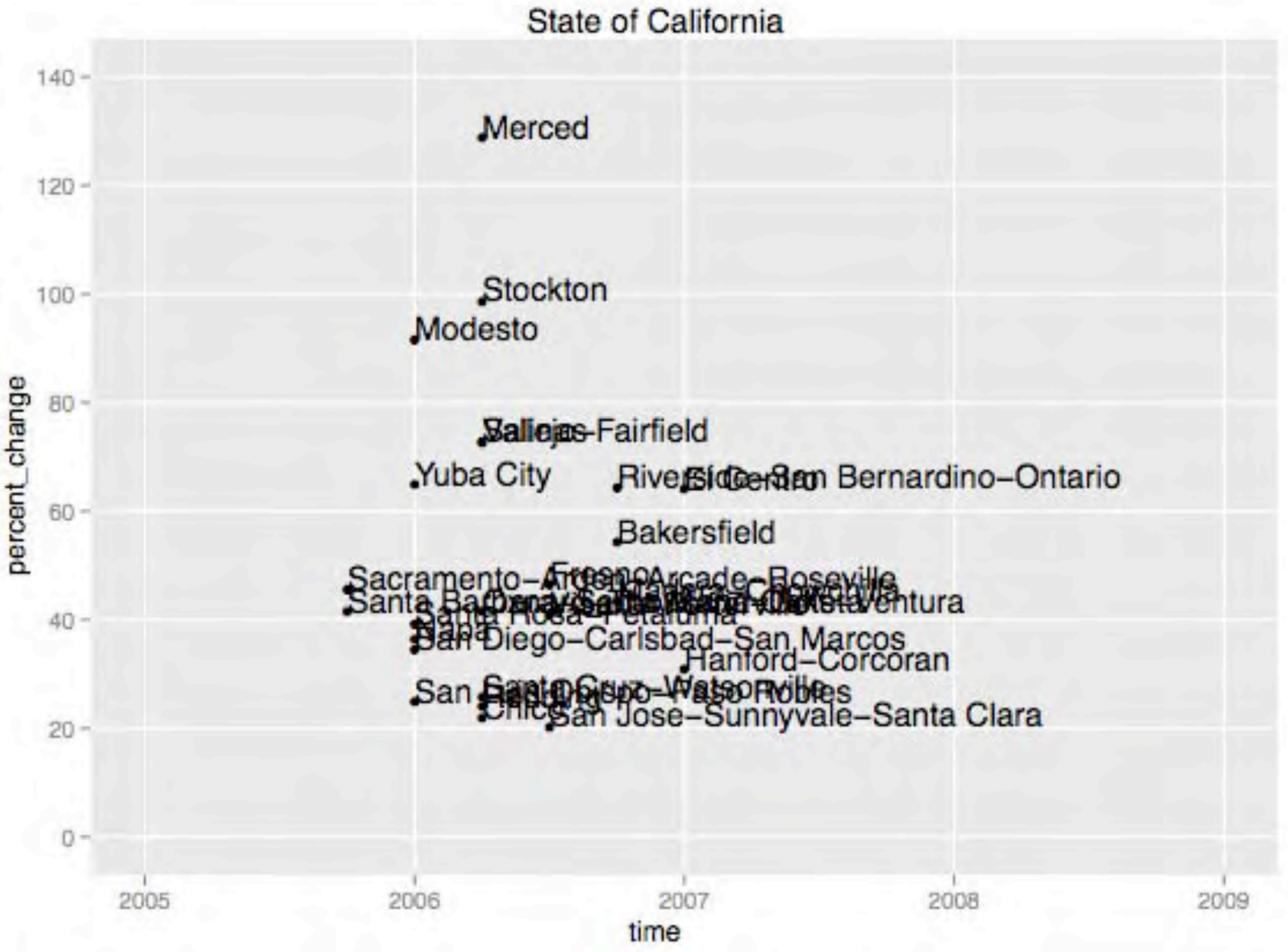

HPI vs. Time
hpi
350
300
250
200
150
100
time
2000
2002
2004
2006
2008
Naples-Marco Island, FL
Marco Island, FL
Santa Barbara
Merced, CA
Ft. La
Honolulu, HI
a-Rome, NY
San Jose-Sun

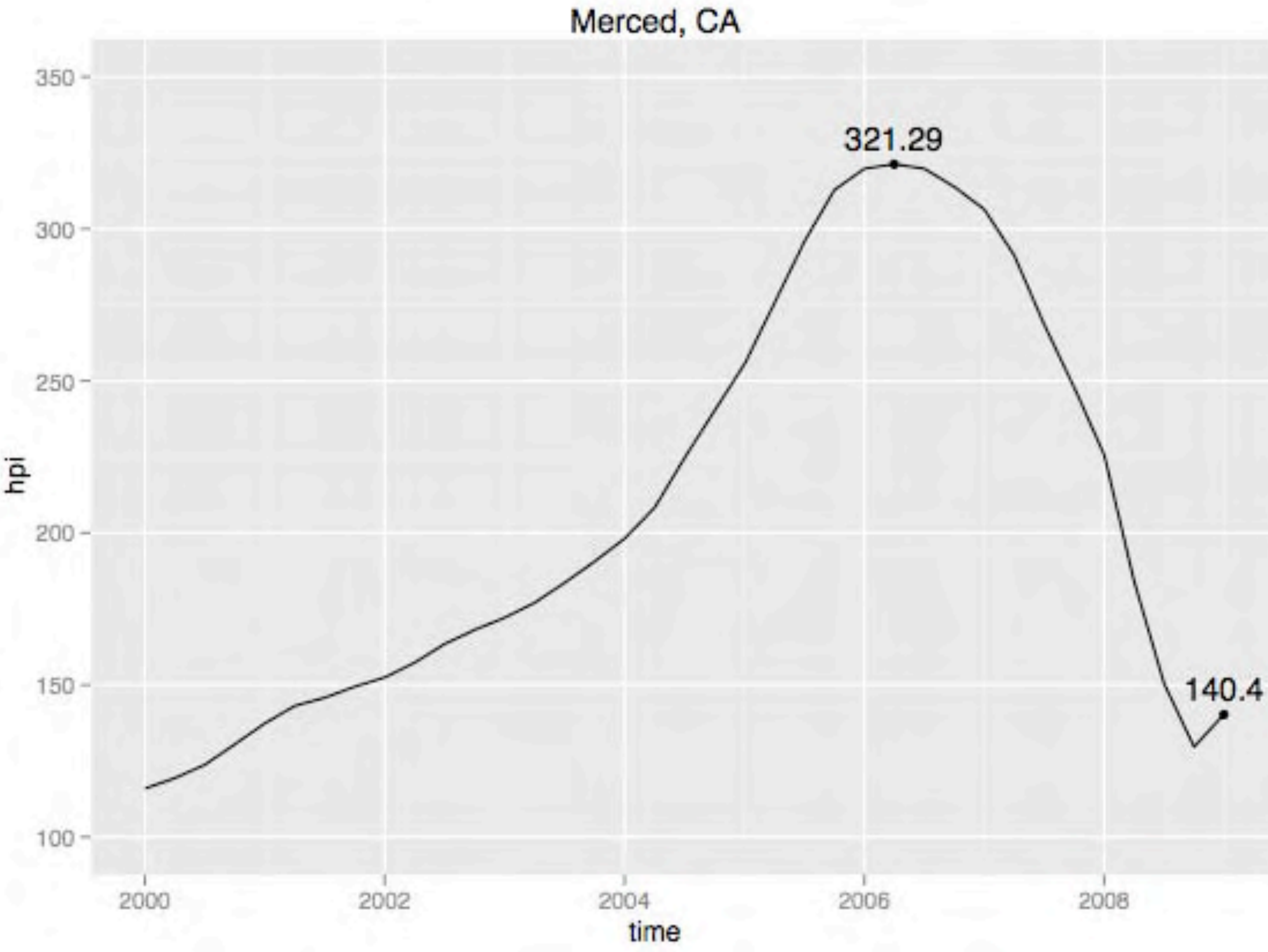
# Interesting Findings

- Data set: Housing Price Indexes from Federal Housing Finance Agency (FHFA)
- Merced, California



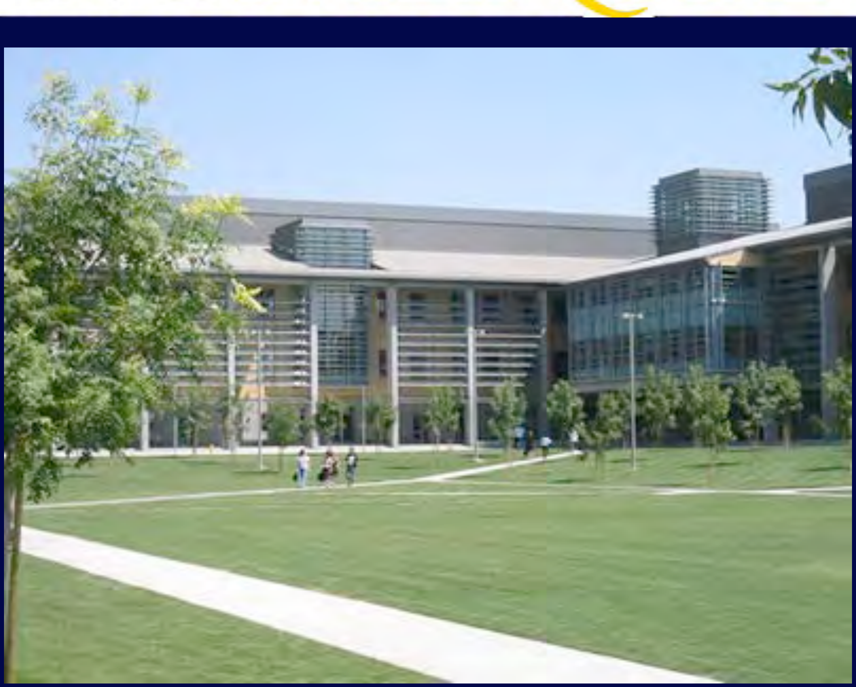
Percent Change vs. Peak HPI Time


State of California
percent_change
time
0
20
40
60
80
100
120
140
2005
2006
2007
2008
2009
Merced
Stockton
Modesto
Vallejo–Fairfield
Salinas
Yuba City
Riverside–San Bernardino–Ontario
El Centro
Bakersfield
Sacramento–Arden–Arcade–Roseville
Fresno
Santa Ba
Oxnard–Thousand Oaks–Ventura
Santa Rosa–Petaluma
Napa
San Diego–Carlsbad–San Marcos
Hanford–Corcoran
Santa Cruz–Watsonville
San Luis Obispo–Paso Robles
San
Chico
San Jose–Sunnyvale–Santa Clara

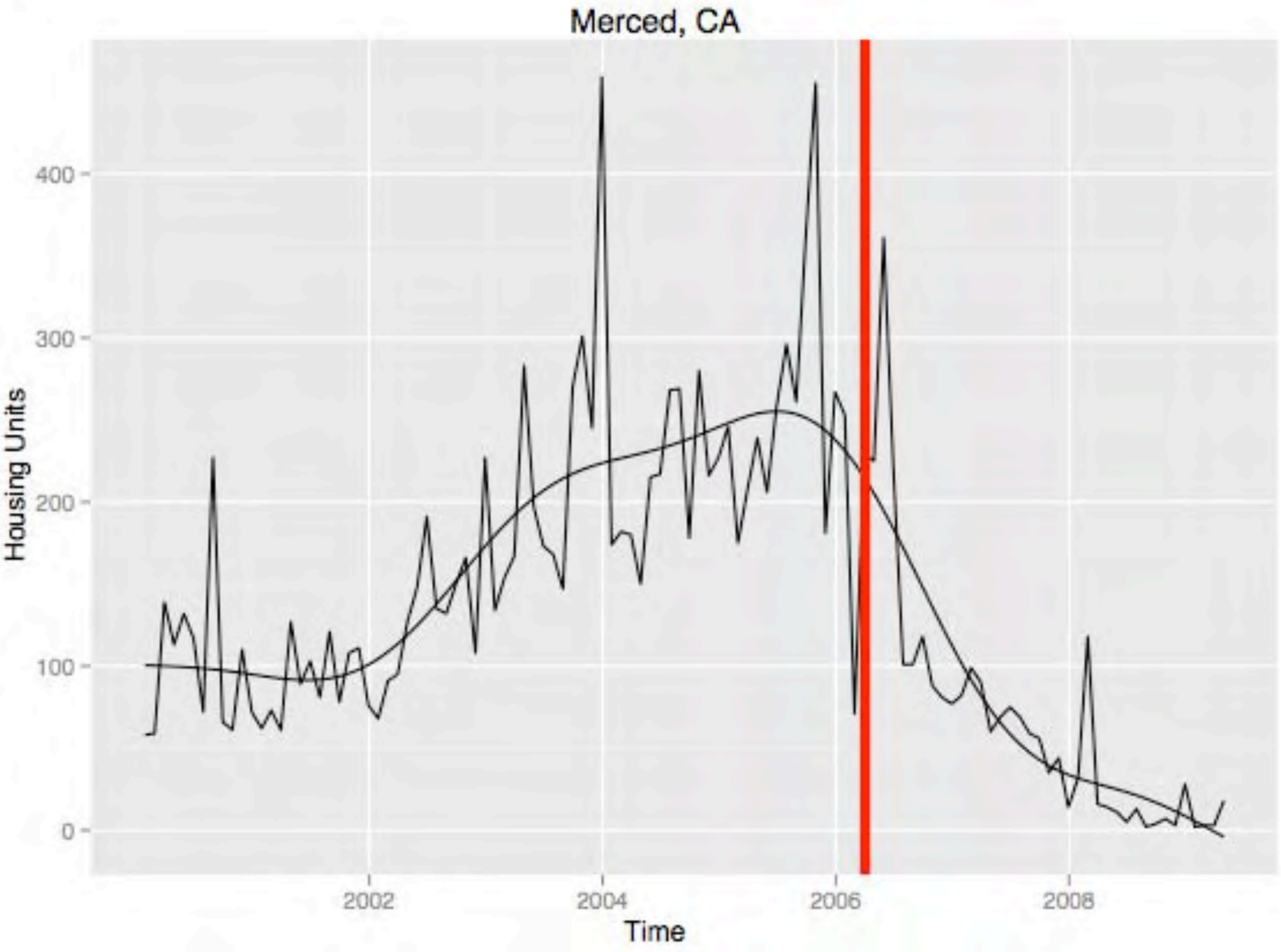

Merced, CA
321.29
140.4
350
300
250
200
150
100
hpi
2000
2002
2004
2006
2008
time

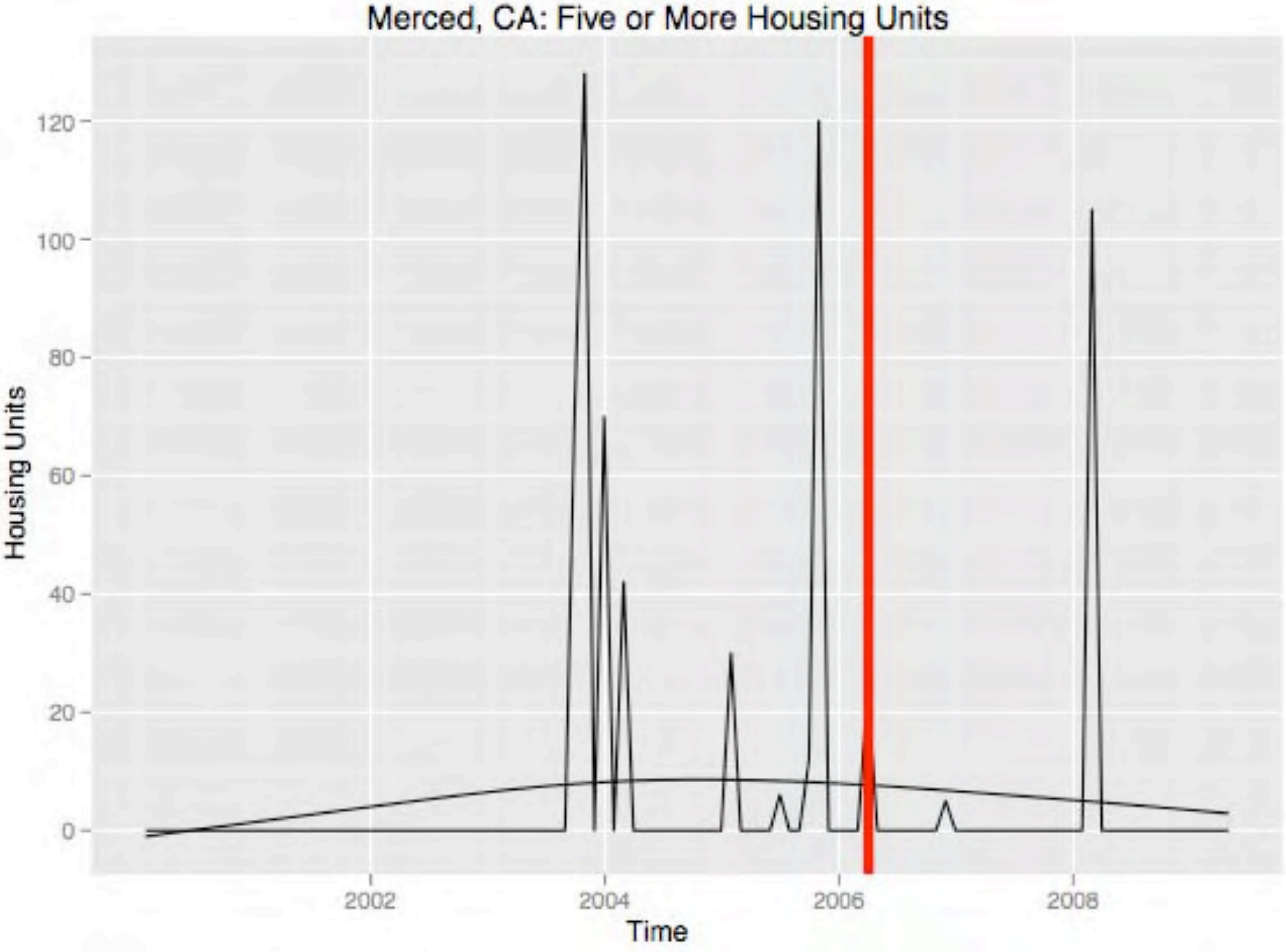
# Possible Causes for the Bubble

UNIVERSITY OF CALIFORNIA
UCMERCED



- In September 2005 University California, Merced finished construction
- More housing
  - Construction implies demand, causes increase in price


Merced, CA
Housing Units
400
300
200
100
0
2002
2004
2006
2008
Time


Merced, CA: Five or More Housing Units
Housing Units
120
100
80
60
40
20
0
2002
2004
2006
2008
Time

# Future Analysis of Merced

- Is this pattern consistent among other cities?
- Foreclosures
- Examine relationship between construction and house prices

# Other Explorations

- Vacation Spots: people who own second homes
- Where are people moving?
- Renting vs. owning

I CAN'T AFFORD MY MORTGAGE BECAUSE OF MY PAY CUT. THE BANK WILL TAKE MY HOUSE.
scottadams@aol.com
www.dilbert.com
I SAVED A BUNDLE BY BEING A RENTER. I SHOULD BUY YOUR HOUSE FOR NEXT TO NOTHING.
6-26-09 ©2009 Scott Adams, Inc./Dist. by UFS, Inc.
TOO SOON?

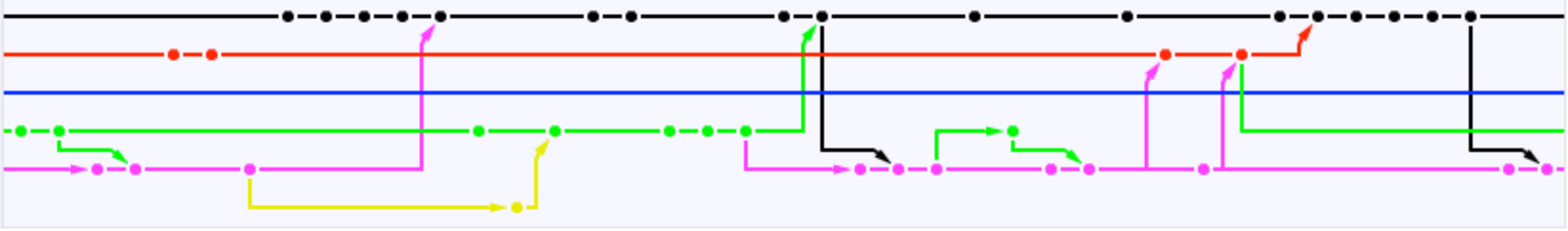
# Communicating Our Findings

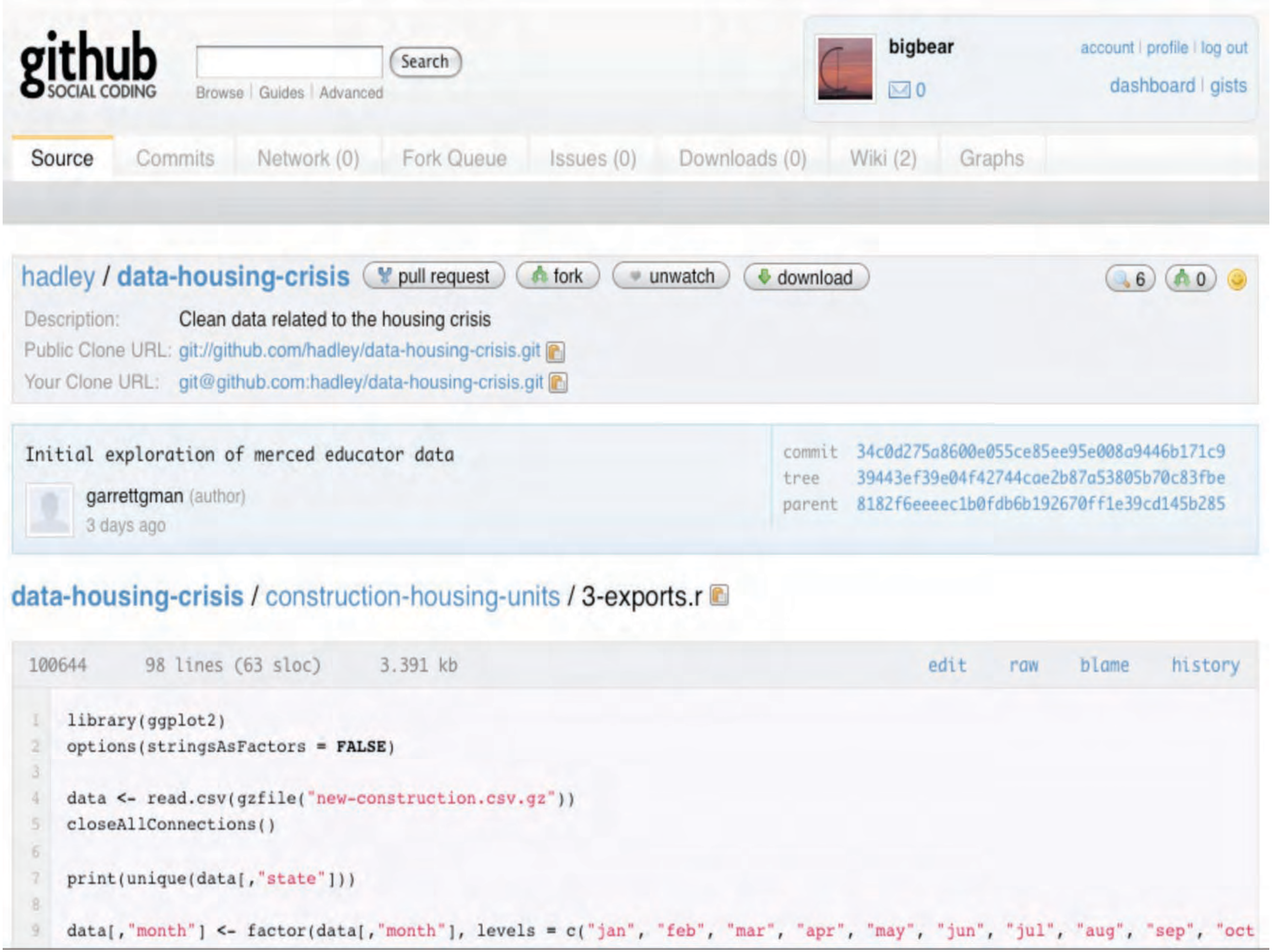
- Reproducibility
  - Requirement of good science
  - Complete record of the data and processes
  - Work should verifiable
  - Others can build upon previous work

- Data & R Code
  - Download
  - Clean
  - Exports

# Making it Available

- Github Website
  - Tracks and posts changes to data made from multiple individuals
  - Free to post and download
  - http://github.com/hadley/data-housing-crisis/tree/master





hadley / data-housing-crisis

Description: Clean data related to the housing crisis

Public Clone URL: git://github.com/hadley/data-housing-crisis.git

Your Clone URL: git@github.com:hadley/data-housing-crisis.git

Initial exploration of merced educator data

garrettgman (author)

3 days ago

commit 34c0d275a8600e055ce85ee95e008a9446b171c9

tree 39443ef39e04f42744cae2b87a53805b70c83fbe

parent 8182f6eeeec1b0fdb6b192670ff1e39cd145b285

data-housing-crisis / construction-housing-units / 3-exports.r

100644 98 lines (63 sloc) 3.391 kb

```
library(ggplot2)
options(stringsAsFactors = FALSE)

data <- read.csv(gzfile("new-construction.csv.gz"))
closeAllConnections()

print(unique(data[,"state"]))

data[,"month"] <- factor(data[,"month"], levels = c("jan", "feb", "mar", "apr", "may", "jun", "jul", "aug", "sep", "oct
```

# Communication

- Once we have interesting findings, how do we communicate them to the public?
- Interactive Website: http://money.cnn.com/news/storysupplement/economy/gapmap/index.htm
- Protovis

# Overview

- Good data is hard to find: not consistent, not concise, not complete, not correct
- Use R to clean
- Use R to analyze: discovered Merced CA, big effect of UC Merced
- Reproducibility crucial, other researchers can build on our work
- To do: communicate our findings