Project 3: Analyzing Real Estate Data | Data Analysis with R

- Author: Chi-Yuan Cheng (cyuancheng AT gmail DOT com)

- updated: June 25, 2015

---

1. Introduction

In this project, I analyzed residential real estate data of USA to explore the historical house prices by state and by region. I also digged a bit deeper to look at which variable in the Census data impacted the house price in America between 2008 and 2012.

2. Data Manipulation

(a) Load and reshape data

The data I used are:

- Housing data of USA, 1996-2015, from Zillow (http://www.zillow.com/research/data/).
    - Median Listing Price per Square Feet
    - Median Sold Price per Square Feet
- US American Community Survey 5 year estimates, 2008-2012 (data (http://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t))
    - I obtained the Census data using an API from US Census Bureau (http://api.census.gov/data/key_signup.html).
- Violent crime rates (per 100,000) by state in USA, 1960-2010, from rMaps (https://github.com/ramnathv/rMaps/tree/master/data).

The data structure of house data set:

```
## 'data.frame':    1780680 obs. of  6 variables:
##  $ RegionName: Factor w/ 6290 levels "Aberdeen","Abingdon",..: 3618 2980 891 4114
4119 2811 4650 1180 4657 2508 ...
##  $ state     : Factor w/ 43 levels "AL","AR","AZ",..: 30 4 13 34 3 29 4 38 4 9 ...
##  $ Metro     : Factor w/ 558 levels "","Aberdeen",..: 349 285 90 389 390 270 438 1
16 440 240 ...
##  $ CountyName: Factor w/ 927 levels "Adair","Adams",..: 691 481 190 649 507 165 75
2 216 761 255 ...
##  $ monthyear : Date, format: "1996-04-01" "1996-04-01" ...
##  $ price     : num  121.4 112.9 90.7 42.3 62.5 ...
```

This date set contains the following features: RegionName, state, Metro, CountyName, monthyear, and price. The features of interest in this project are *state*, *Metro*, *monthyear*, and *price*.

Besides house data, I attempted to use Census data to understand how economy and demographics affect house price. I think these data will help support my investigation and data exploratory of house price data.

I obtained the Census data of US American Community Survey 5 year estimates (2008-2012) from an API (http://api.census.gov/data/key_signup.html).

The features of interest are below:

no.	table.number	feature	dataframe
1	B25026	Total population in occupied housing units by tenure	“df_pop_state”
2	B01002	Median age by sex	“df_median.age”
3	B25119	Median household income by tenure	“df_median.income”
4	B25105	Median Monthly housing costs	“df_median.housecost”
5	B25097	Mortgage status by median value	“df_median.housevalue”
6	B25103	Mortgage status by median real estate taxes paid	“df_housetax”
7	B25119	Median year structure built	“df_houseyear”
8	B25039	Median year householder moved into unit by tenure	“df_moveyear”

Besides Census data, I also explored *violent crime rates* data of USA from the rMaps package.

(b) Clean data

I removed outliers from the house data.

Fig. 2-1a: Original data set

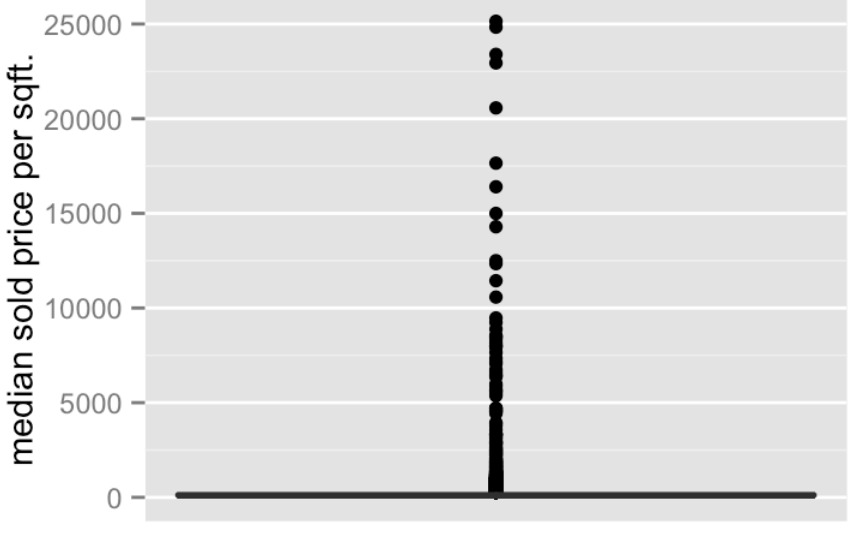
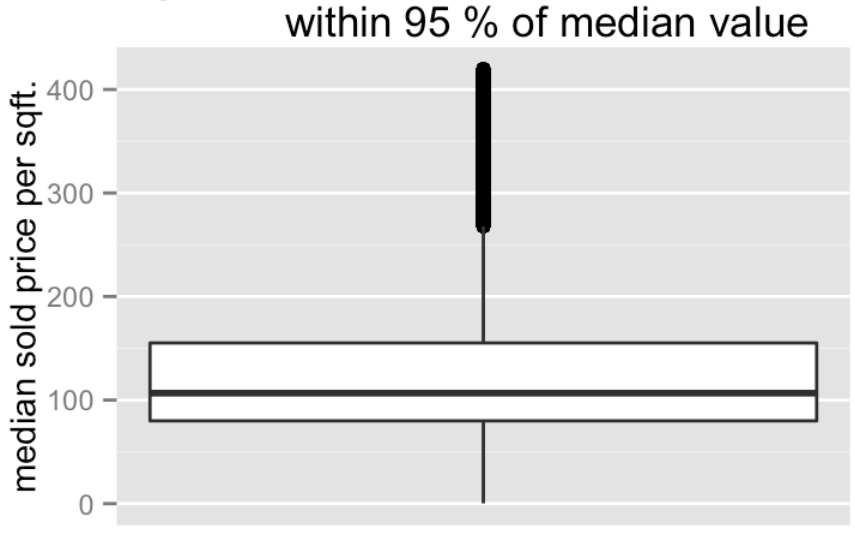


Fig. 2-1b: Remove outliers that are not within 95 % of median value



The median sold price in the house data set contains lots of outliers (Fig 2-1a). I removed the outliers and only included the prices that are within 95 % of median price value. The box plot after removing outliers is shown in Fig 2-1b.

(c) Data overview

I summarized the data contained the data of house price, Census data, and crime rates between 2008 and 2012:

```
##   state    region median_sold_price     n rn      pop med.age
## 1    AL     South           88.7611  1767  1  4777326    37.8
## 2    AR     South           73.7366  2172  4  2916372    37.4
## 3    AZ      West          100.7179  2768  3  6410979    36.0
## 4    CA      West          199.2475 17672  5 37325068    35.2
## 5    CO      West          130.6381  3516  6  5042853    36.1
## 6    CT Northeast          155.0388  3229  7  3572213    40.0
##   med.housecost med.housevalue med.income housetax houseyear house.year
## 1           739         122300      54032      575      1980         32
## 2           667         106300      51169      715      1981         31
## 3          1009         175900      61647     1460      1988         24
## 4          1441         383900      82899     3379      1973         39
## 5          1154         236800      74794     1529      1980         32
## 6          1421         285900      89691     5059      1963         49
##   yearmove move.year median_crime
## 1     2002        10        450.1
## 2     2003         9        505.3
## 3     2004         8        426.5
## 4     2003         9        473.3
## 5     2004         8        338.8
## 6     2001        11        300.9
```

The observations of Census data:

```
## [1] 51
```

The observations of house data:

```
## [1] 42
```

It turns out that 9 states (AK, ID, KS, MT, ND, NM, SD, UT, and WY) are missing in the house data set of Zillow.

To get a better sense of the variables I will explore, I used univariate analysis to obtain the distribution of house price, time, and state in the house data set

Fig. 2-2a: Distribution of house price in the data set

Fig. 2-2b: Distribution of time in the data set

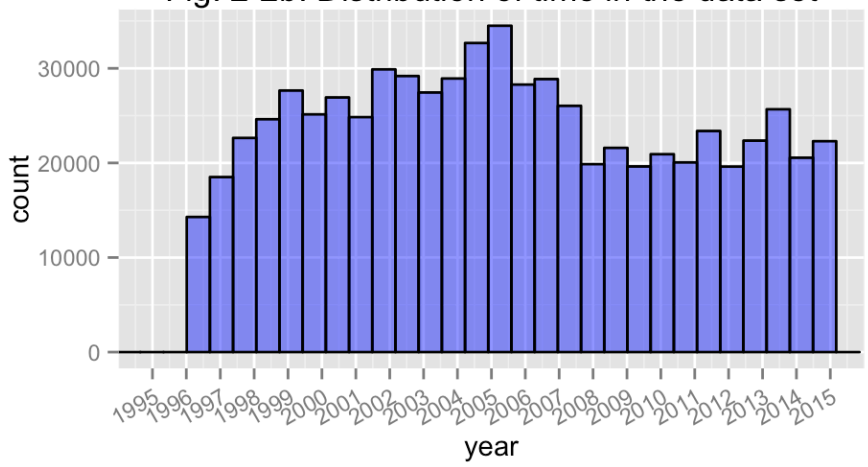


Fig. 2-2c: Distribution of state in the data set

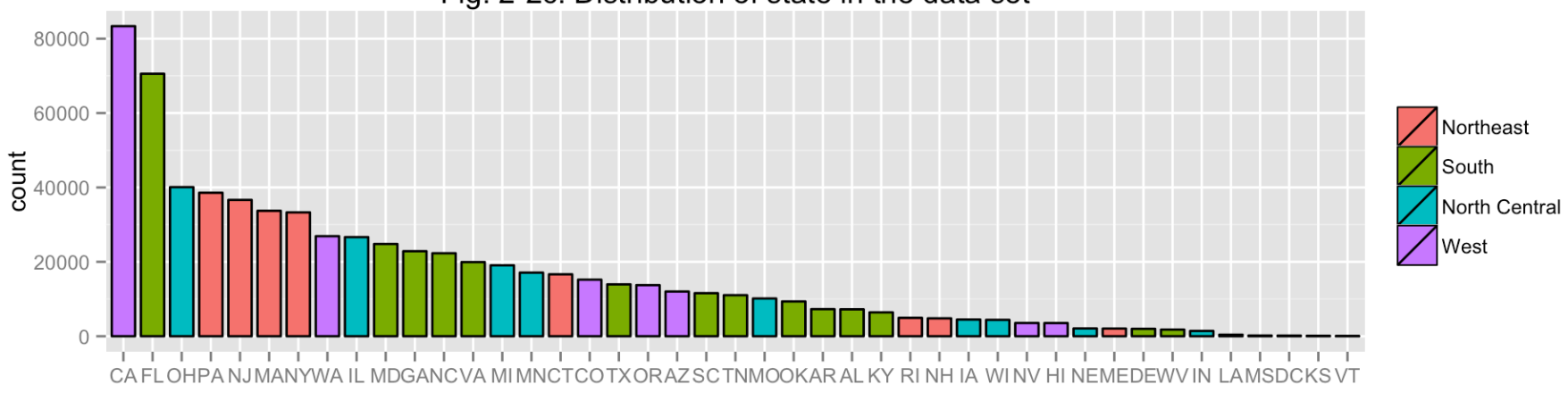


Fig 2-2a shows that the house price exhibits a positively skewed distribution, which has the mean to the right of the median value. According to Fig 2-2b, the house data is recorded between 1996 and 2015. The maximal count of data was accumulated between 2004 and 2006. Fig 2-2c shows that the count of data for each state has a great distribution. CA has the largest data pool, whereas VT has the smallest data pool.

3. Data Analysis and Exploration

A. House data

(a) Comparison of sold house price and listed house price in USA (2009-2014)

The idea here is to see how the listed house price different from the sold house price. Firstly, let's compare the house sold price and listed price between 2009 and 2014.

Summary of sold house price (2009-2014):

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.84   85.19  114.90  134.90  163.80  419.80
```

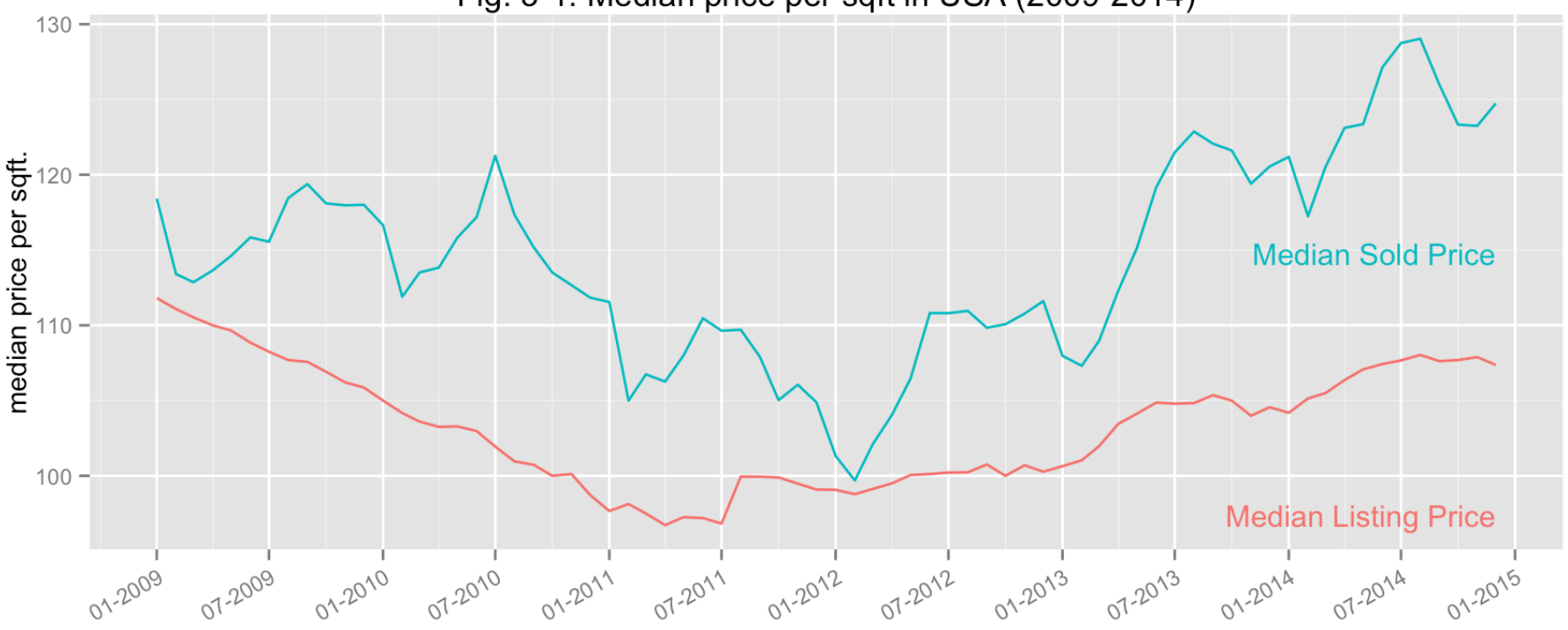
Summary of listed house price (2009-2014):

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.051  78.420 103.100 117.900 141.200 375.500
```

The sold house price is on average slightly higher than the listed house price between 2009 and 2014. It would be interesting to compare the prices before 2009. But unfortunately, the data of median listed price from Zillow is only available after October 2008.

Next, we can further compare the median listed price vs. median sold price with time.

Fig. 3-1: Median price per sqft in USA (2009-2014)



Again, the median sold price is overall higher than the median listing price between 2009 and 2014. The minimal sold price occurred in January 2012. Since than, the sold price drastically raises to about 30 %.

Because the Census data we will explore is between 2008 and 2012, it would be interesting to know what is the most expensive and cheapest home sold during this time period.

(b) The 5 most expensive and less expensive home sold (2008-2012)

The 5 most expensive home sold (2008-2012):

```
##        state      RegionName       Metro  monthyear    price    region
## 10730     CA  West Hollywood Los Angeles 2012-12-01 419.8367      West
## 31160     DC      Washington  Washington 2011-01-01 419.8348      <NA>
## 53775     HI         Kaneohe    Honolulu 2010-12-01 419.7760      West
## 104688    NY      Great Neck    New York 2008-08-01 419.7698 Northeast
## 100811    NV Incline Village        Reno 2008-02-01 419.7434      West
```

The 5 cheapest home sold (2008-2012):

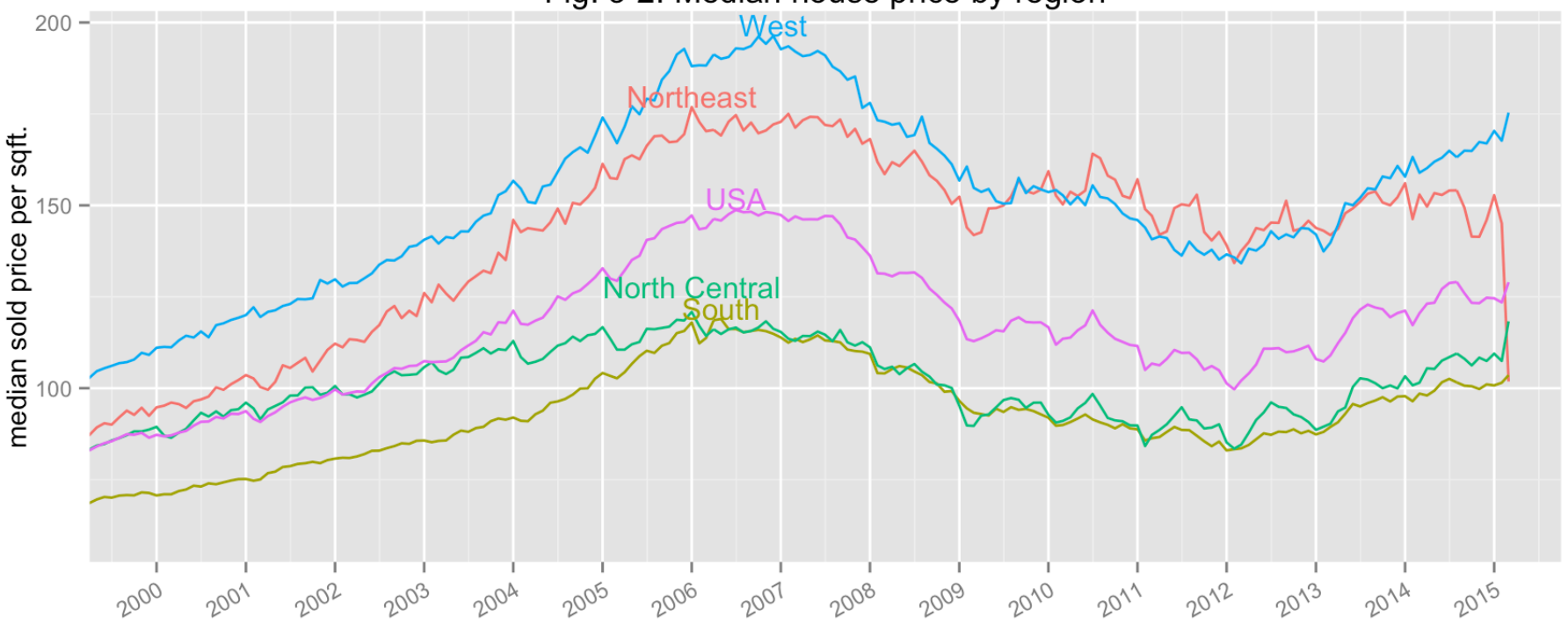
```
##        state       RegionName        Metro  monthyear    price
## 44477     FL          Palatka      Palatka 2012-03-01 10.84184
## 127777    PA Chester Township Philadelphia 2011-03-01 11.51398
## 121628    PA     Mahanoy City   Pottsville 2009-06-01 11.64080
## 129332    PA       McKeesport   Pittsburgh 2012-11-01 14.14720
## 61237     IN             Gary      Chicago 2011-01-01 15.04630
##              region
## 44477         South
## 127777     Northeast
## 121628     Northeast
## 129332     Northeast
## 61237  North Central
```

Between 2008 and 2012, the most expensive house was sold in Los Angeles, CA, whereas the cheapest house was sold in Palatka, FL. Notably, three of the top 5 most expensive home were sold in the West.

(c) Comparison of house sold price by region (2000-2014)

The house price largely depends on the location. Here, let's take a look at the house price changes with time at different regions.

Fig. 3-2: Median house price by region

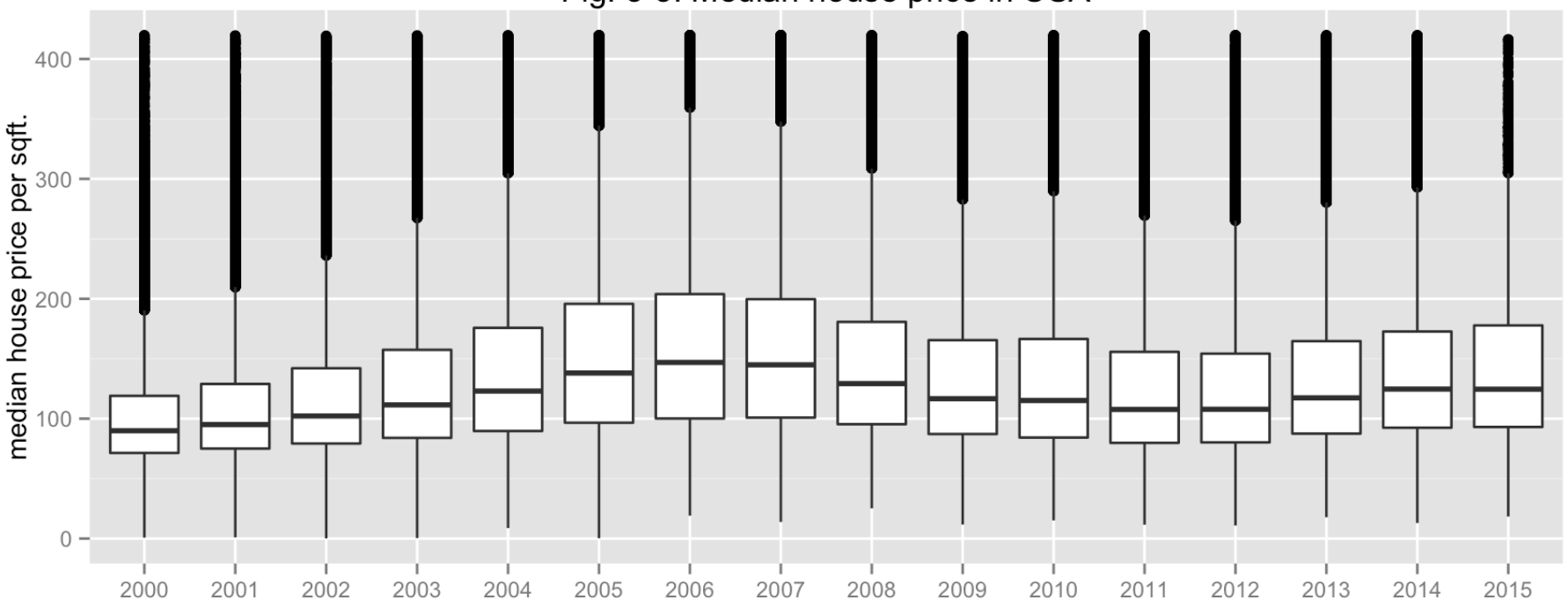


The house prices in the West and Northeast are above national median price, whereas the prices in the North Central and South are below the national median price between 2000 and 2014. Generally, the maximal price appears in 2007, following by a graduate decrease in price until 2012. The price then increases after 2012.

Interestingly, the house price in the Northeast drastically drops in the beginning of 2015. The price is even below the national median price and price in the North Central. This decline of house price might be caused by the several snowstorms (http://www.boston.com/news/local/massachusetts/2014/01/03/the-first-major-snow-storm-underway/v7af5SOnFX09Oj3pjIECeK/story.html) in the Northeast in the end of last year.

To make sure Fig 3-2 is not over-plotting, I used a box plot to show the variance in the data with time.

Fig. 3-3: Median house price in USA

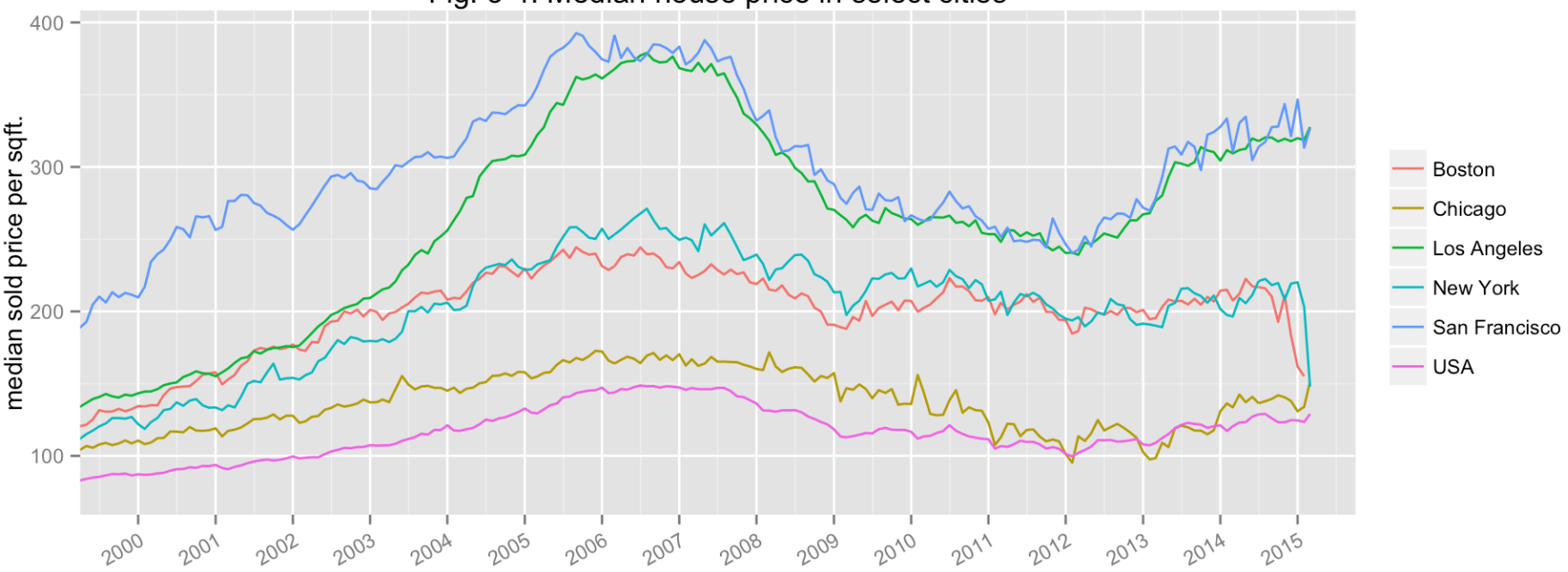


The above figure shows that almost all the data follows the same trend as in Fig 3-2, suggesting that Fig 3-2 is not over-plotting.

(d) Comparison of house sold price in select cities (2000-2014)

Secondly, we can compare the house price in select cities.

Fig. 3-4: Median house price in select cities



The trend of sold price in select cities are very similar with the trend in their regions shown in Fig 3-2. The maximal house price is between 2006 and 2007, following by a graduate decline. The minimal price reaches in 2012, and price increases again until 2015. The house prices in New York and Boston are dropped in the end of 2014. Overall, the house prices of these big cities are higher than the national price. The house prices in San Francisco and Los Angeles are about a factor of 2-3 higher than the national median price.

(e) Comparison of sold price in Santa Barabra (1996-2014)

I live in Santa Barbara, CA for 6 years. I'm very interested in the house price in this area, so I did a data exploration on the house price of cities in Santa Barbara county, as following.

```
## Source: local data frame [10 x 3]
##
##       RegionName median_price   n
## 1         Goleta     292.3177 122
## 2     Los Alamos     281.6901   1
## 3  Santa Barbara     270.0513  82
## 4    Carpinteria     236.5199  51
## 5        Solvang     225.8456  66
## 6      Guadalupe     204.6679   1
## 7       Buellton     200.8393  35
## 8     Santa Ynez     154.4892   7
## 9    Santa Maria     154.1066 228
## 10        Lompoc     151.1586 224
```

The house price in local region of Santa Barbara shows a large distribution. The difference between the highest and lowest house price in this area is about a factor of 2.

Next, the question we can ask is when is the best month to buy a house in Santa Barabara. To answer this question, I explored the house price changes by month in Santa Barabara, and compared it with the house price changes in USA.

Fig. 3-5a: Changes of median sold price by month in Santa Barbara (1996-2014)

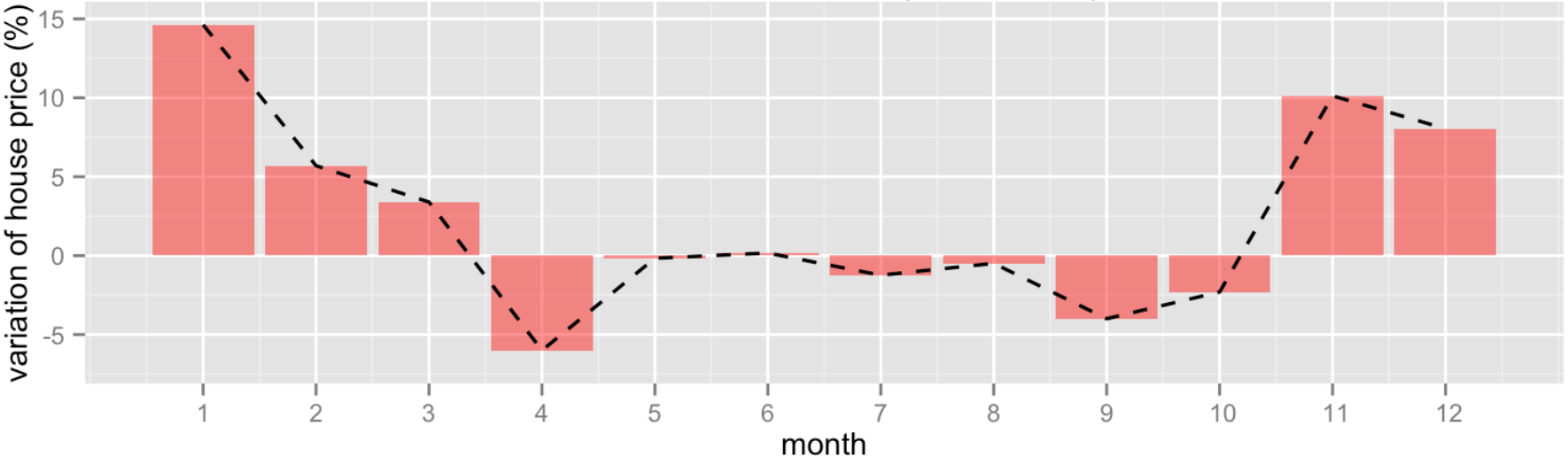
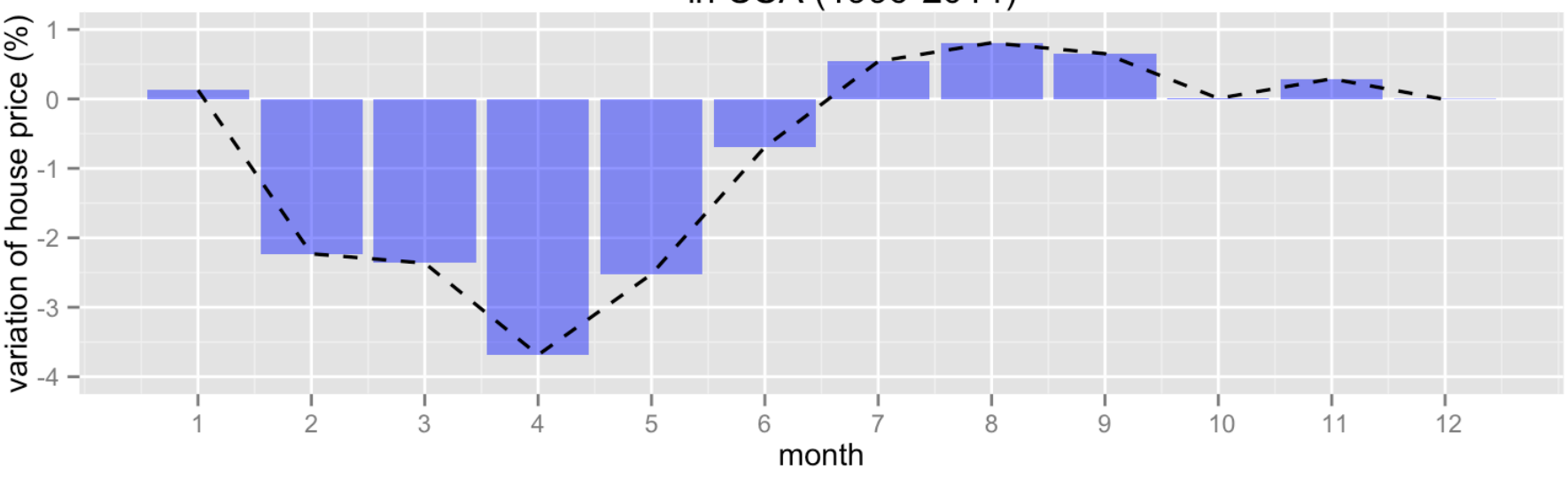


Fig. 3-5b: Changes of median sold price by month in USA (1996-2014)



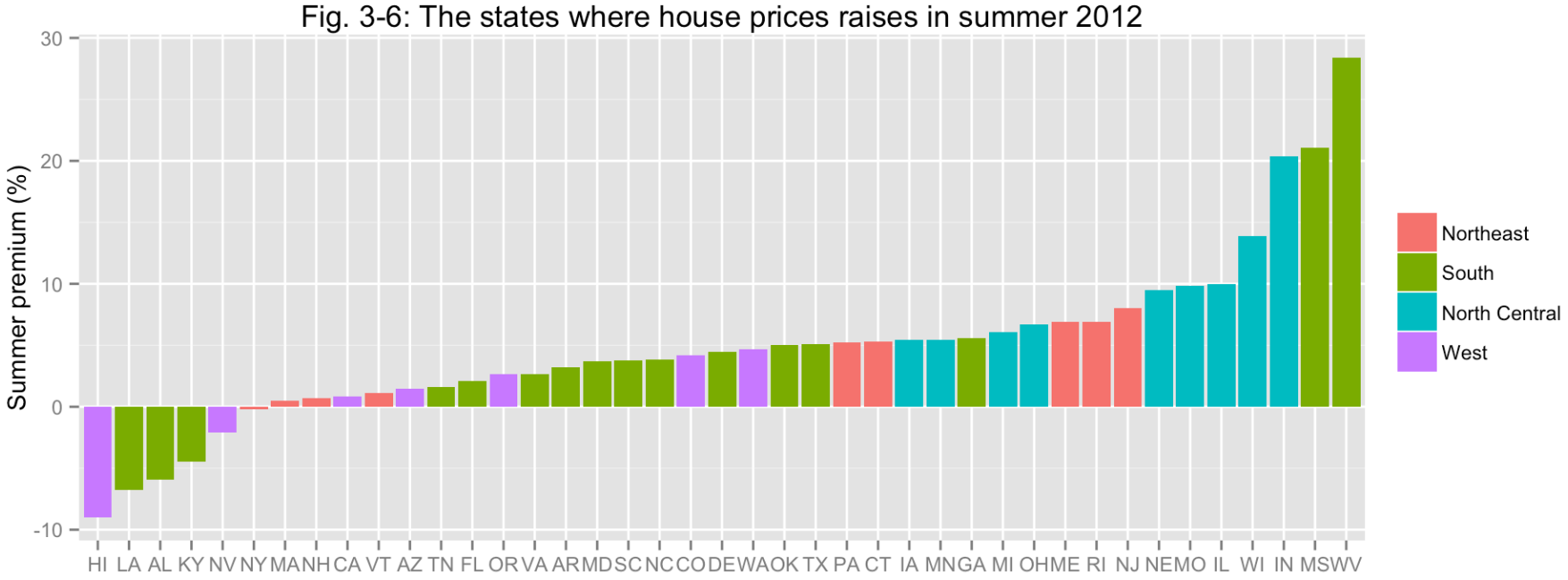
Comparing the monthly median price in Santa Barbara city (Fig 3-5a) and USA (Fig 3-5b), I found that Santa Barbara historically had higher house price in January, February, March, November, and December. Therefore, it may not be a good idea to buy house in Santa Barbara during these months. The best month to buy a house in Santa Barbara might be in April, because the house price drops about 5 % in this month.

Unlike the monthly house price in Santa Barbara, It seems like the house price in USA has a season pattern, with lower prices in spring and higher prices in summer. Specifically, the national median price in spring is about 2-3.5 % lower than its median value, but the median national price in summer is about 1 % higher than its median value. Moreover, I found that the variation of house price in Santa Barbara is much greater than that in USA.

(f) House price change depending on the season

We know from Fig 3-5b that the national house price gets raised in summer. It would be interesting to look at the states where their house price get raised in summer months.

Fig. 3-6: The states where house prices raises in summer 2012



The house prices of most of the states in the North Central and the Northeast raise in the summer than in the winter. Interestingly, HI, LA, AL, KY and NV have the opposite pricing trend. The house prices in these states are much cheaper in the summer than in the winter. Again, the house price in CA seems to be insensitive to the season, which is consistent to our early finding in Fig 3-5a.

B. Census data (2008-2012)

In the following, I explored Census data between 2008 and 2012. Among the variables that I considered to explore are below.

(a) The population estimate

Let's first compare the population by state and by region.

Fig. 3-7a: Median population by state above national median value

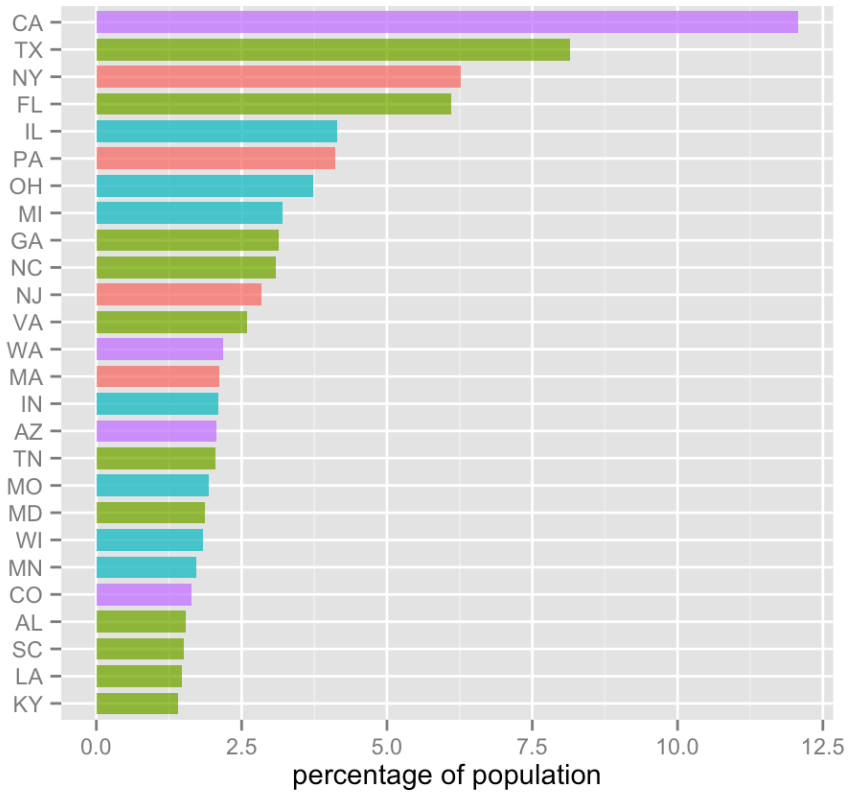
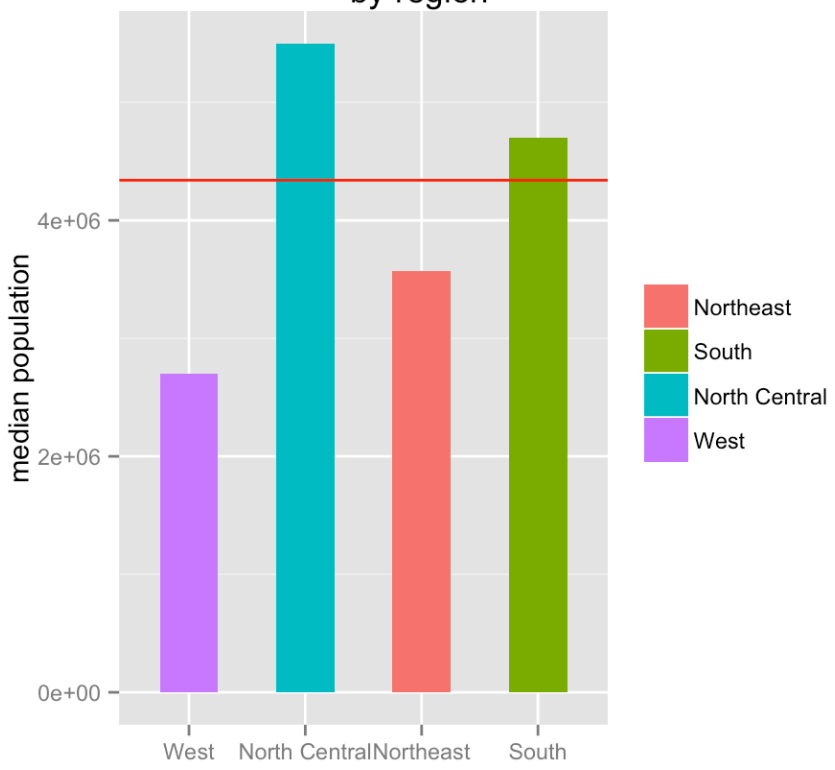


Fig. 3-7b: Median population by region



CA has the largest population in USA (12 %). The most populous region in USA is the North Central, which is above the median national value (red line). The West has the lowest population in USA.

(b) Median age

I'm interested to know whether the median age of people in state affects the house price. For example, the house price for a state with a high percentage of senior residents might be quire different that the state with a high population of younger residents and family. I explored the distribution of aged population by state and by region.

Fig. 3-8a: Median age by state above national median value

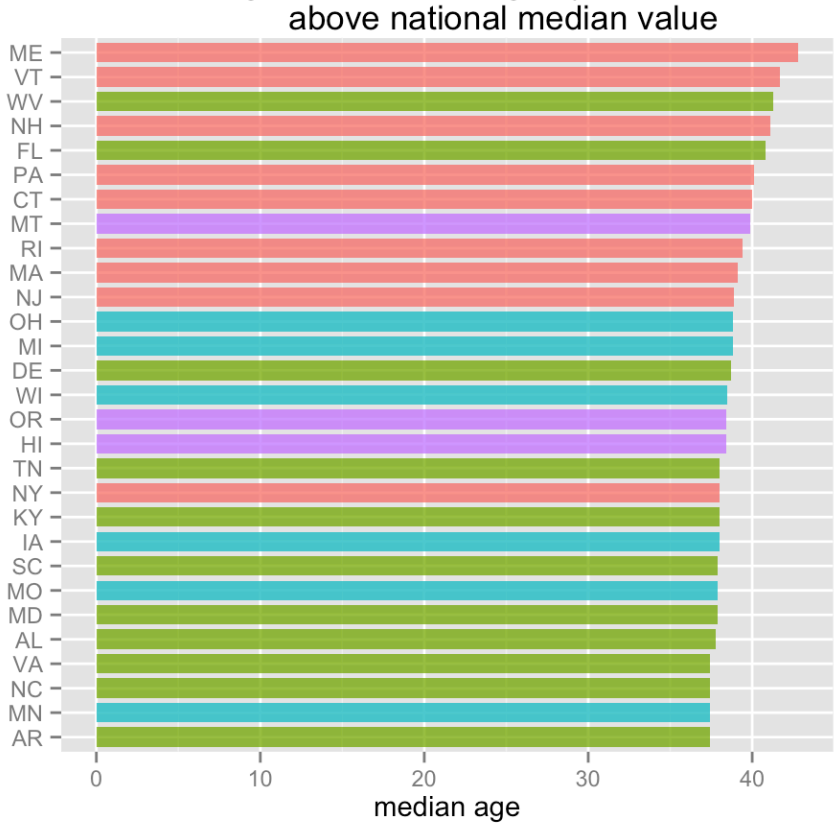
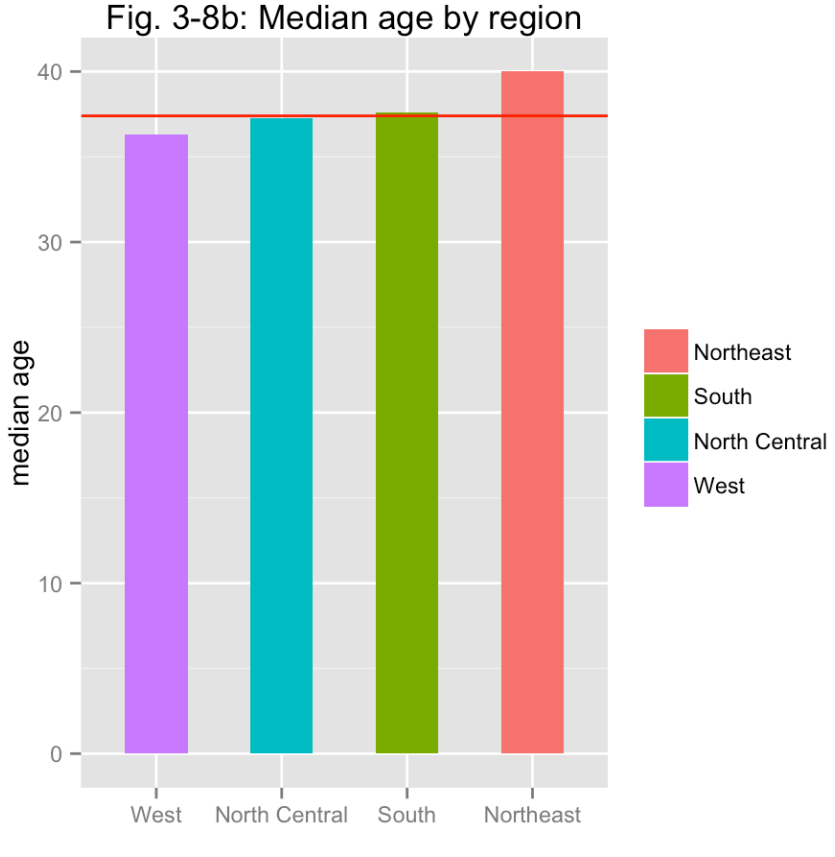


Fig. 3-8b: Median age by region



ME has the oldest median age, whit a value of 42.8. That's more than five years above the national median age of 37.2. Interestingly, senior residents tend to live in the Northeast, whereas younger residents tend to live in the West.

(c) Median household income

Obviously, household income can affect the house price. Therefore, it is important to know the household income by state and by region.

Fig. 3-9a: Median household income by state above national median value

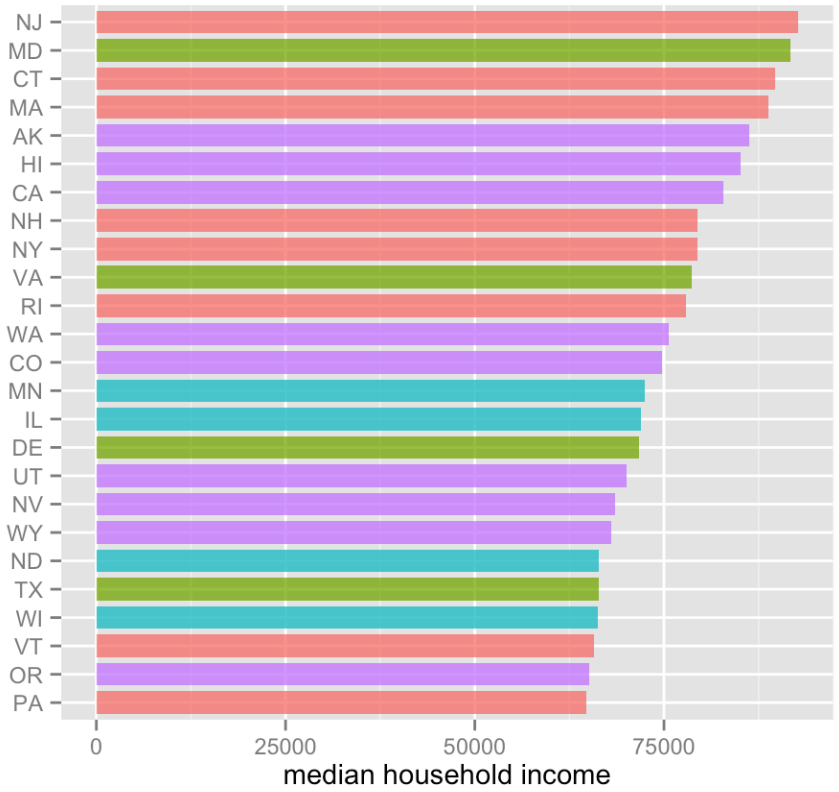
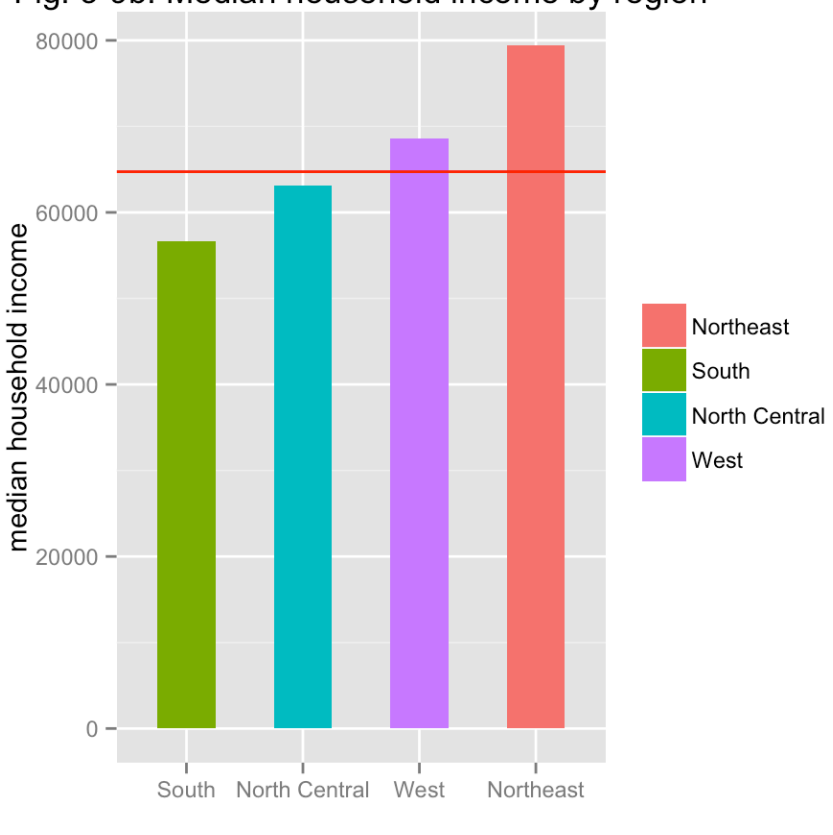


Fig. 3-9b: Median household income by region



People in NJ has the highest household income (92,722), which is about 40 % higher then the national median income. Although the house price in CA is very high, people lives in CA does not earn the highest income among other states. In general, residents in the Northeast and the West earn more household income, with their values above national median value.

(d) Homeowership

I'm interested in analyzing what portion of homeowners' income pays for housing costs. Using Census data, I compared the median homeowner's housing costs to the median homeowner's income for each state.

Fig. 3-10a:Ratio between homeowners' income and housing costs above national value

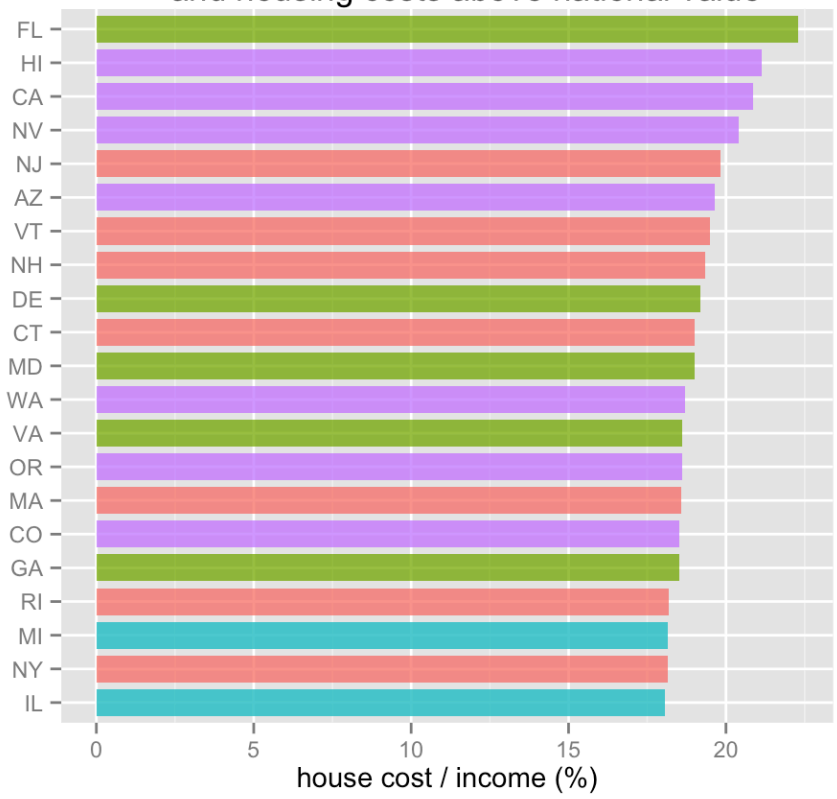
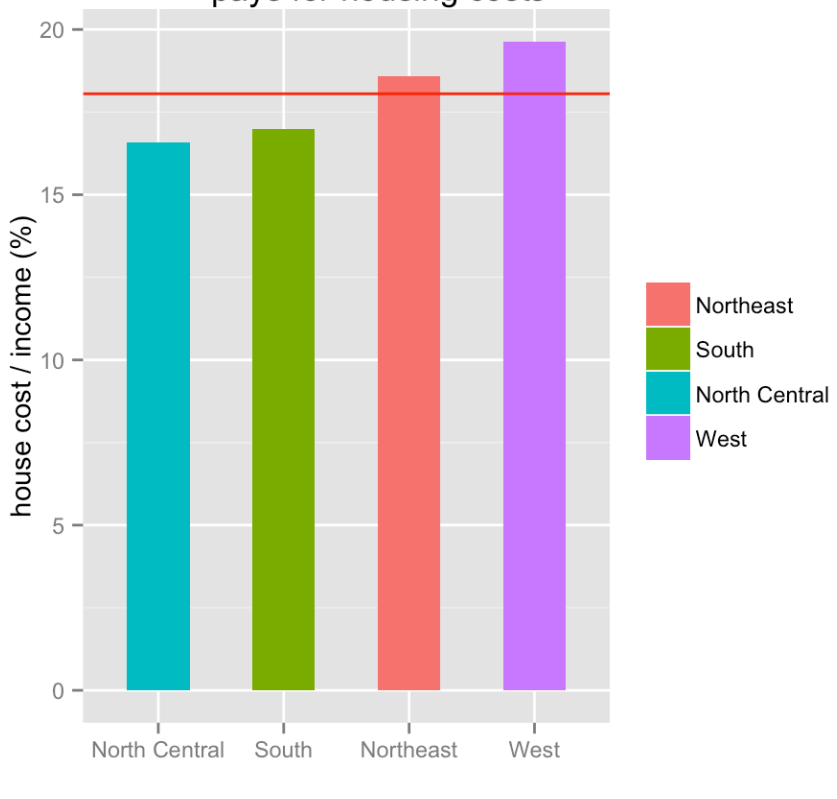


Fig. 3-10b: Percentage of homeowners' income pays for housing costs

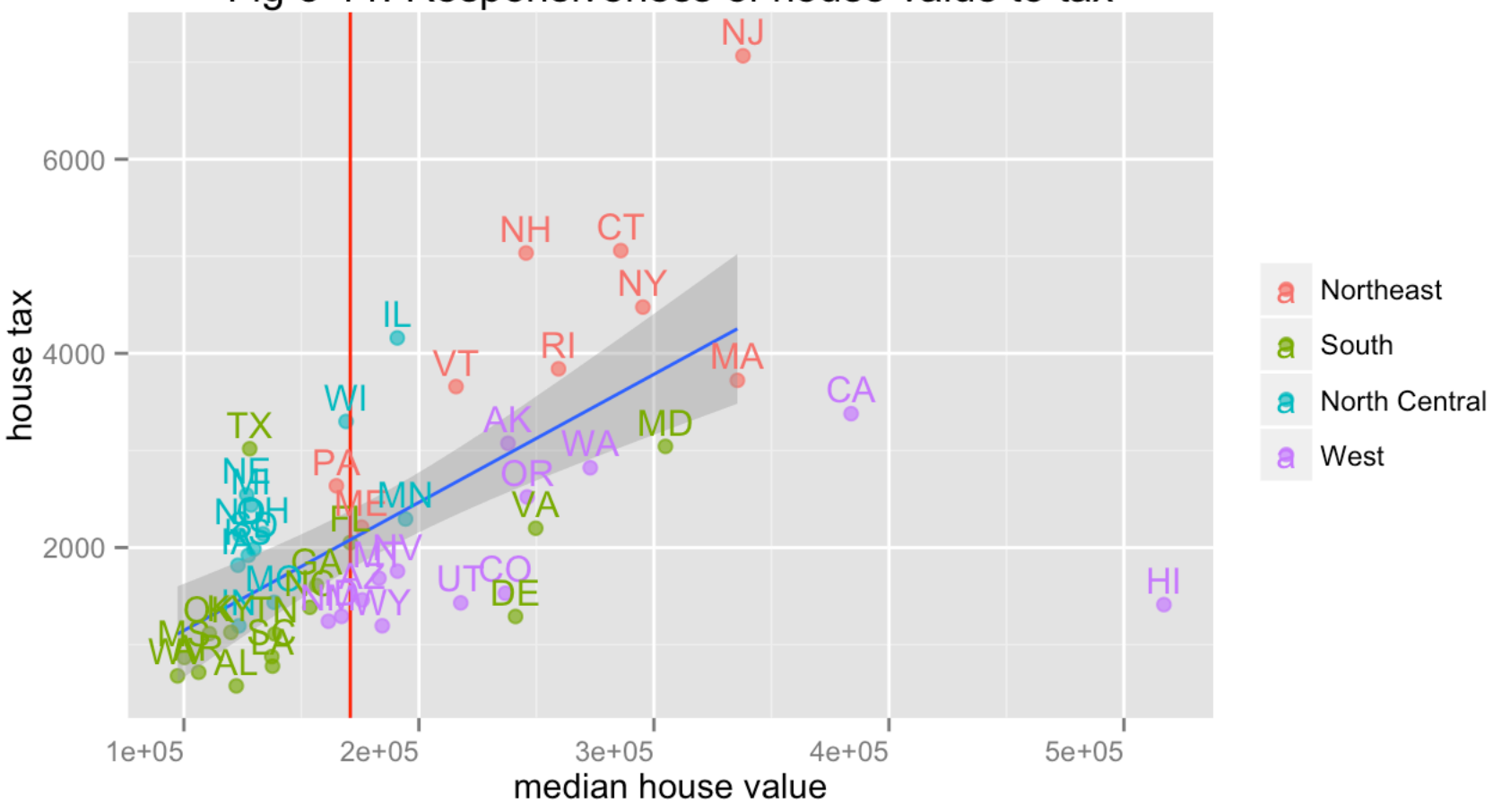


Surprisingly, the most expensive state for homeowners is FL, where homeowners pay 22.3 % of their income in housing costs. Homeowners in CA pays 20.9 % of their income in housing costs. In general, homeowners in the West and Northeast pay most from their income in the housing costs.

(e) House value and tax paid

House value and tax paid by homeowners are two key factors to affect the house price. I'm interested to know is there any relation between house tax and house value.

Fig 3-11: Responsiveness of house value to tax



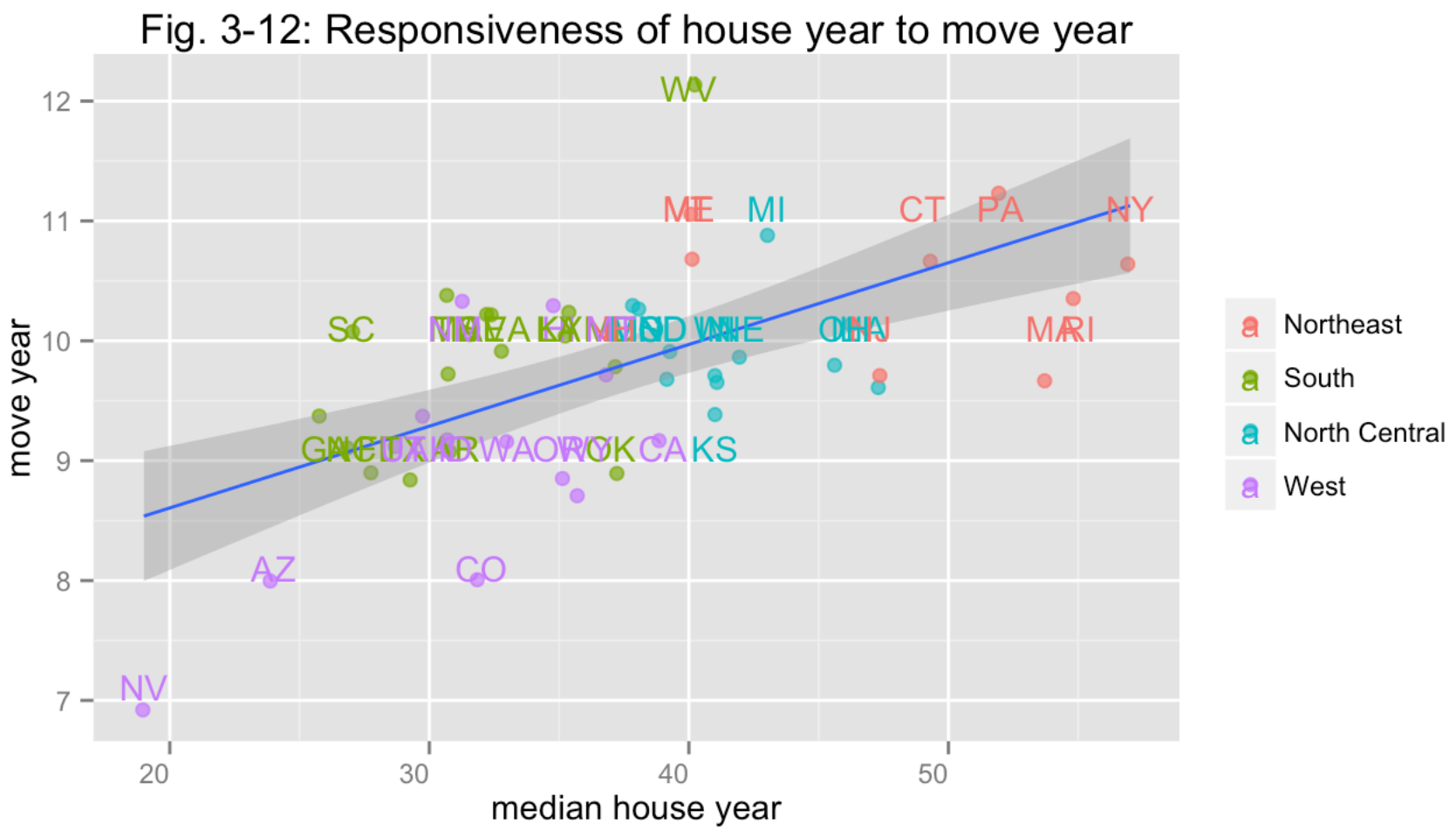
House tax is approximately linearly proportional to house value (r^2 = 0.44). Although the house price is very high in HI, the homeowners in HI do not pay higher house taxes. In contrast, homeowners in NJ pay the highest house tax. Another interesting finding was that the Northeast and the West has higher house values than the national median value (red line).

r^2 value: (excluded CA, HI, NJ)

```
## [1] 0.4415058
```

(f) House age and years for homeowner moved into unit

People tends to live in a new house. It would be intriguing to know whether the house age relates to the time period for homeoweners to stay in their house.



The relation between house age and the year for homeowner to stay in the unit has a poor linear correlation (r^2 = 0.2). NY has the oldest median house age of 62, and homeowners in NY stay in their house for about 11 years. Generally, people in the Northeast and North Central tend to live in their house in a longer period of time (> 9 years). In contrast, people in the West, especially in NY anc AZ, moves more frequently.

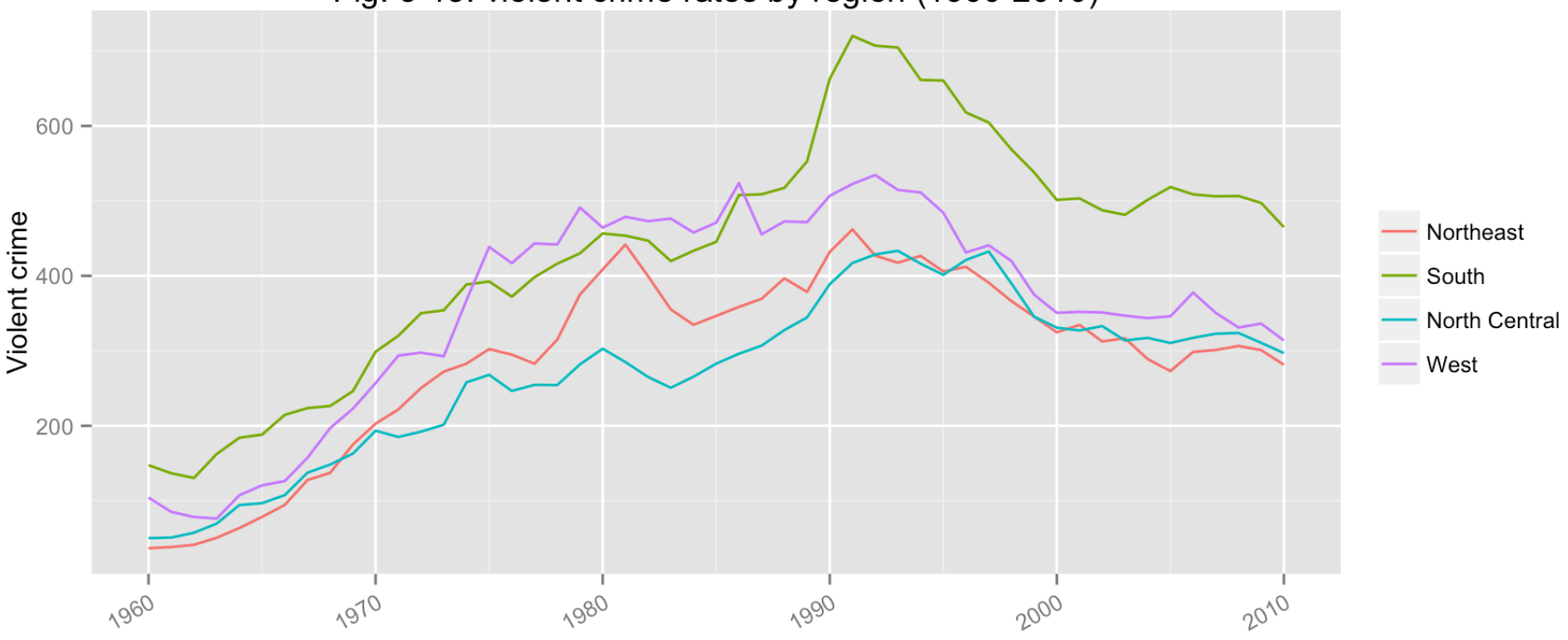
r^2 value:

```
## [1] 0.2047416
```

(g) Violent crime rates

It is obviously that crime rates has a great impact on the house price. Firstly, let's take a look at how the crime rate changes with time in different regions.

Fig. 3-13: violent crime rates by region (1960-2010)



Four regions in USA show the same pattern for crime rate with time. The violent crime rate increases linearly since 1960, and it reaches to the highest value in 1992. After 1992, the crime rate decreases gradually. Noted that the South has the highest crime rate in USA after 1990.

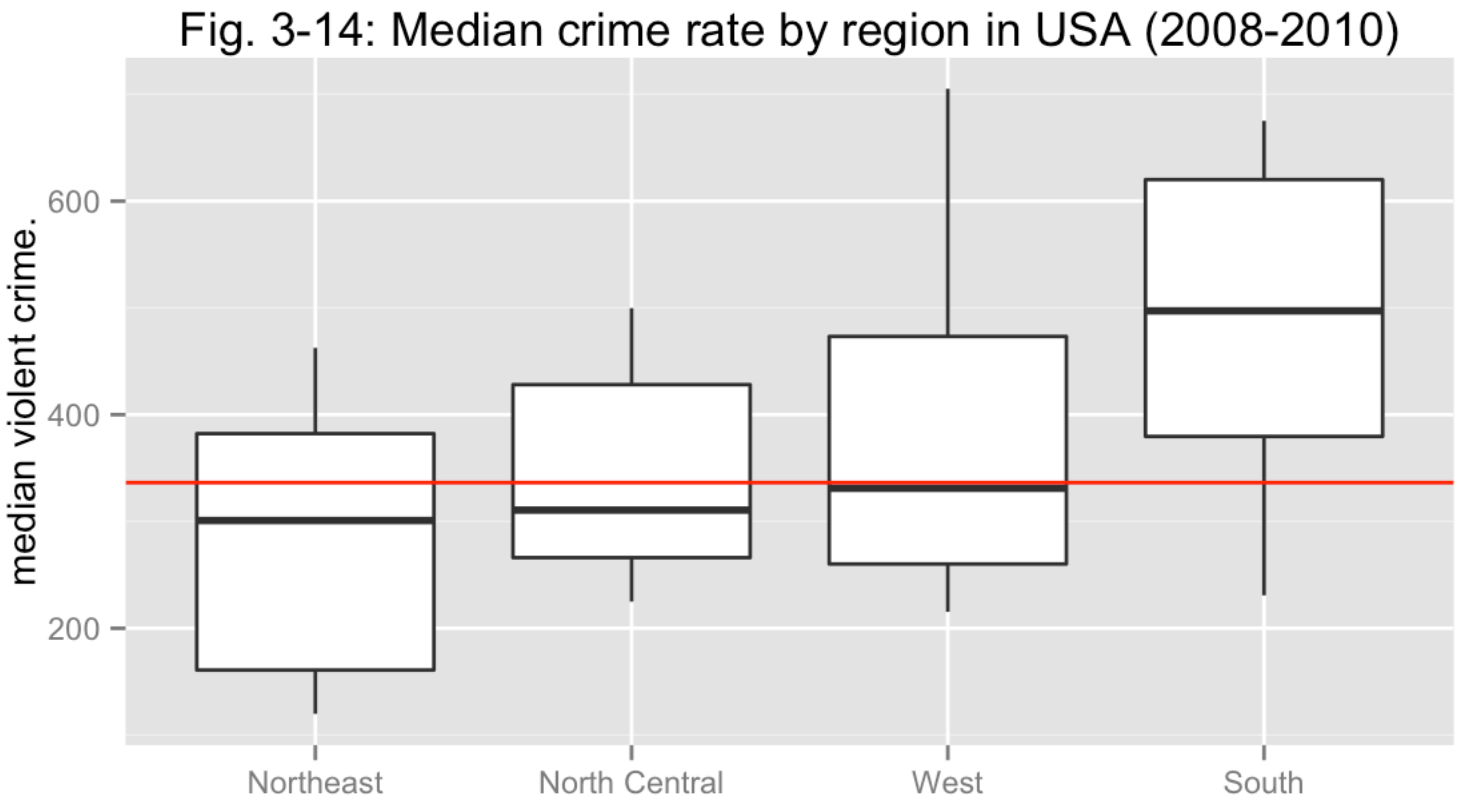
Next, let's take a look at the crime data between 2008 and 2010.

The top 5 crime states are:

```
##    state median_crime region
## 33    NV         705.2   West
## 40    SC         675.2  South
## 42    TN         666.0  South
## 8     DE         645.4  South
## 18    LA         642.9  South
```

NV is the state with the highest crime rate.

It would be interesting to compare the crime rates by region.



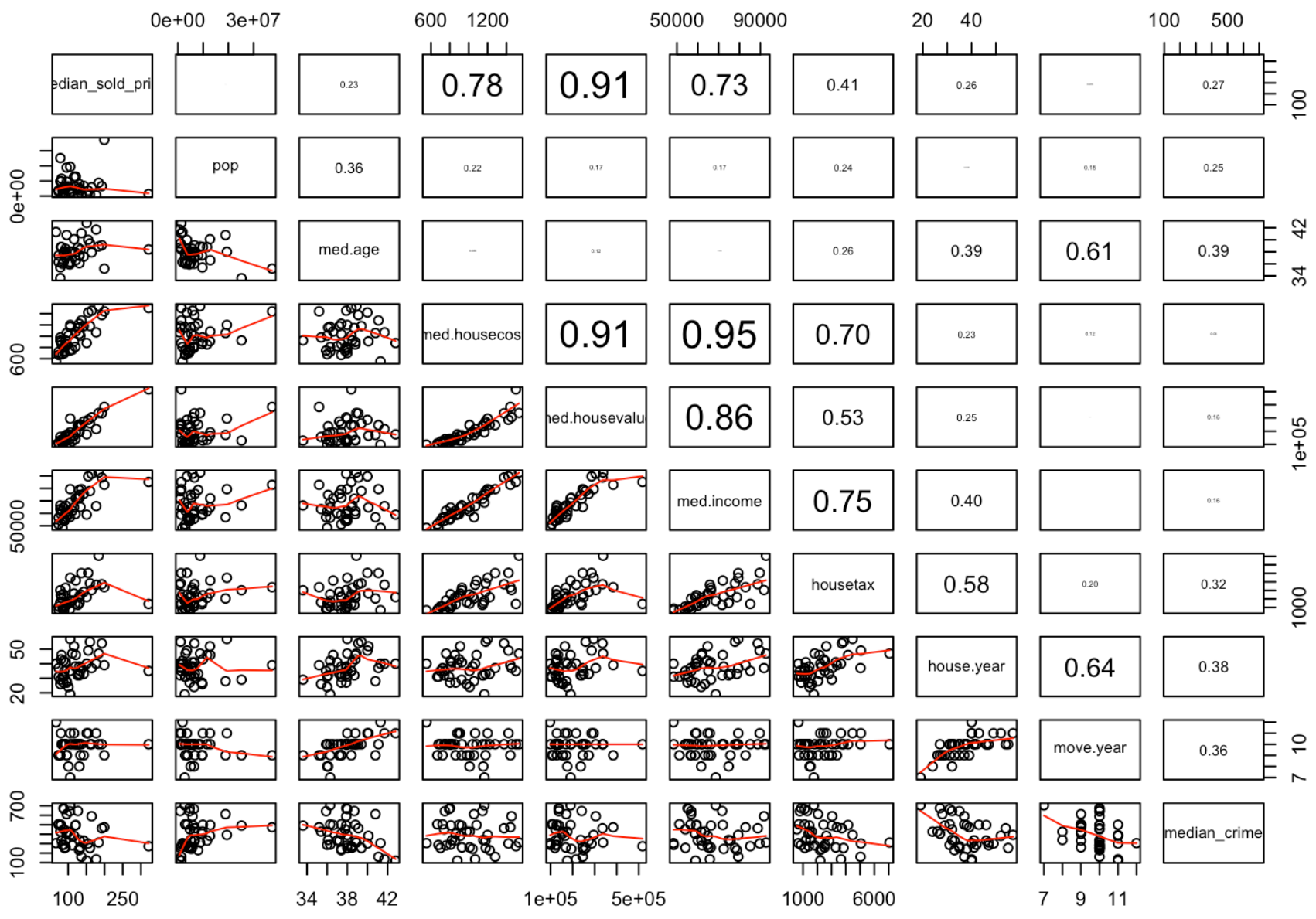
The South has the highest violent crime rate of 497.2, which is 47.8 % higher than the national median value of 336.4. In contrast, the Northeast has the lowest violent crime rate of 300.9 in USA between 2008 and 2010.

C. Census v.s. house data

Here, I want to explore the relationship between the Census data and house data.

(a) Correlation between house data and Census data (2008-2012)

Fig. 3-15: Comparision of house price and Census data in USA (2008-2012)



We can summarize our observation based on Fig 3-15:

- **Affordability**: House price has a strong linear correlation with house cost, house value, and household income, indicating that people who has higher income can afford more expensive house.
- **Stability**: Median age of people in state has a roughly linear correlation with the years that homeowners stay in their house ($r^2 = 0.61$). This makes sense as older people tends to settle in one place for a longer period of time.
- **Tax**: House with higher value or cost will need to pay more tax. Homeowners has higher income will also pay more house tax.
- **Crime**: It was surprised that crime rate does not strongly correlate to the house price. I think the crime rate may become important when we explore house price in the city level.

(b) Attempt to predict house price from Census data

I want to see if I can build a linear model based on Census data to predict the house price.

```
##
## Calls:
## m1: lm(formula = log(I(median_sold_price)) ~ I(med.age), data = merged_df2)
## m2: lm(formula = log(I(median_sold_price)) ~ I(med.age) + med.housecost,
```

```
##     data = merged_df2)
## m3: lm(formula = log(I(median_sold_price)) ~ I(med.age) + med.housecost +
##     med.housevalue, data = merged_df2)
## m4: lm(formula = log(I(median_sold_price)) ~ I(med.age) + med.housecost +
##     med.housevalue + med.income, data = merged_df2)
## m5: lm(formula = log(I(median_sold_price)) ~ I(med.age) + med.housecost +
##     med.housevalue + med.income + housetax, data = merged_df2)
## m6: lm(formula = log(I(median_sold_price)) ~ I(med.age) + med.housecost +
##     med.housevalue + med.income + housetax + house.year, data = merged_df2)
## m7: lm(formula = log(I(median_sold_price)) ~ I(med.age) + med.housecost +
##     med.housevalue + med.income + housetax + house.year + move.year,
##     data = merged_df2)
## m8: lm(formula = log(I(median_sold_price)) ~ I(med.age) + med.housecost +
##     med.housevalue + med.income + housetax + house.year + move.year +
##     median_crime, data = merged_df2)
##
## ================================================================================
=============
##                        m1         m2         m3         m4         m5         m6         m7
m8
## --------------------------------------------------------------------------------
-------------
## (Intercept)       2.986**    2.401***   2.999***   2.890***   2.687***   2.687***   2.680
***   3.085***
##                  (1.036)    (0.587)    (0.465)    (0.501)    (0.625)    (0.652)    (0.657
)    (0.694)
## I(med.age)        0.046      0.034*     0.028*     0.029*     0.032*     0.032*     0.039
*     0.029
##                  (0.027)    (0.015)    (0.012)    (0.012)    (0.013)    (0.015)    (0.018
)    (0.019)
## med.housecost               0.001***   0.000     -0.000     -0.000     -0.000     -0.000
0.000
##                            (0.000)    (0.000)    (0.000)    (0.000)    (0.001)    (0.001
)    (0.001)
## med.housevalue                         0.000***   0.000***   0.000***   0.000***   0.000
***   0.000**
##                                       (0.000)    (0.000)    (0.000)    (0.000)    (0.000
)    (0.000)
## med.income                                        0.000      0.000      0.000      0.000
0.000
##                                                  (0.000)    (0.000)    (0.000)    (0.000
)    (0.000)
## housetax                                                    -0.000     -0.000     -0.000
-0.000
##                                                             (0.000)    (0.000)    (0.000
)    (0.000)
## house.year                                                             -0.000      0.001
0.002
##                                                                        (0.005)    (0.005
)    (0.005)
```

```
## move.year                                                                -0.028
-0.022
##                                                                          (0.043
)   (0.042)
## median_crime
-0.000
##
(0.000)
## ------------------------------------------------------------------------------
-------------
## R-squared         0.068     0.712     0.835     0.836     0.838     0.838     0.84
0     0.851
## N                41        41        41        41        41        41        41
41
## ==============================================================================
=============
```

The models take into accounts the following features: *median age*, *house cost*, *house value*, *household income*, *house tax*, *house year*, *move year*, and *crime rates* in state. Finally, the r^2 is 0.848.

Next, I used the median national values of the variables between 2008 and 2012 as input values to predict the house price, and calculate the accuracy of the prediction.

```
## [1] 0.9927913
```

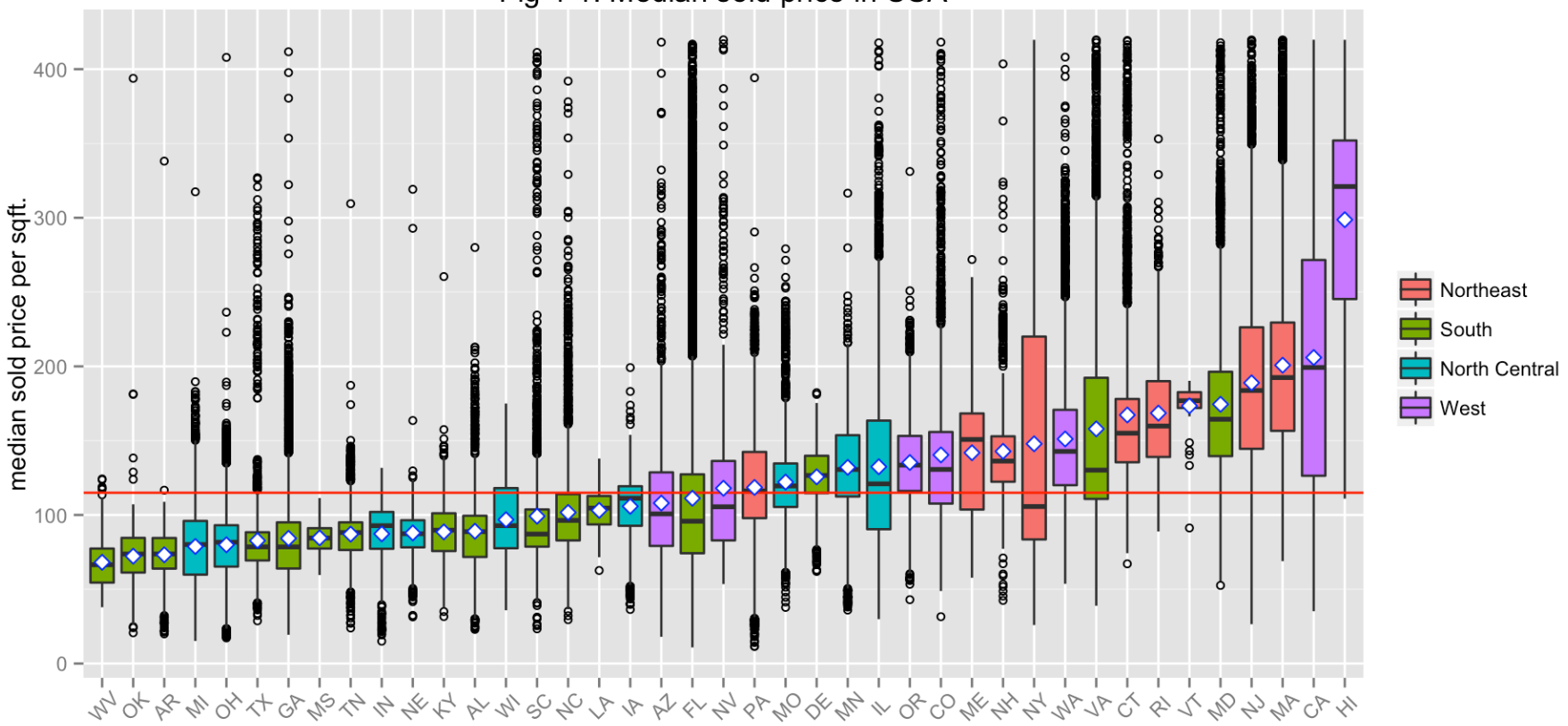
The accuracy of prediction for median house price is 93.4 %.

4. Final Plots and Summary

In the following, I will summarize the data exploration of the house data.

(a) Median sold price by state in USA

Fig 4-1: Median sold price in USA



The median sold price in HI is the highest in USA, which is about a factor of 3 higher than the national median value (shown in red line). The median sold price in CA is the second highest in the country, but it exhibits a very wide distribution among other states. Moreover, the sold price in CA is about a factor 2 higher than the national sold price.

Interestingly, the house price in the New York city is thought to be very expensive, but the median sold price of the entire NY state is slightly below the national median price. Overall, We can clearly see that house prices of most states in the Northeast and the West are above the national value, whereas the prices of most South states are below the national value. This results are consistent with our previous findings.

(b) Median sold price by month (2008-2014)

Fig 3-5b suggests that house price in USA may has a season pattern. Here, I explore the median sold price by month in USA to see if the house price has a seasonal trend.

Fig 4-2: Median sold price per sqft by month

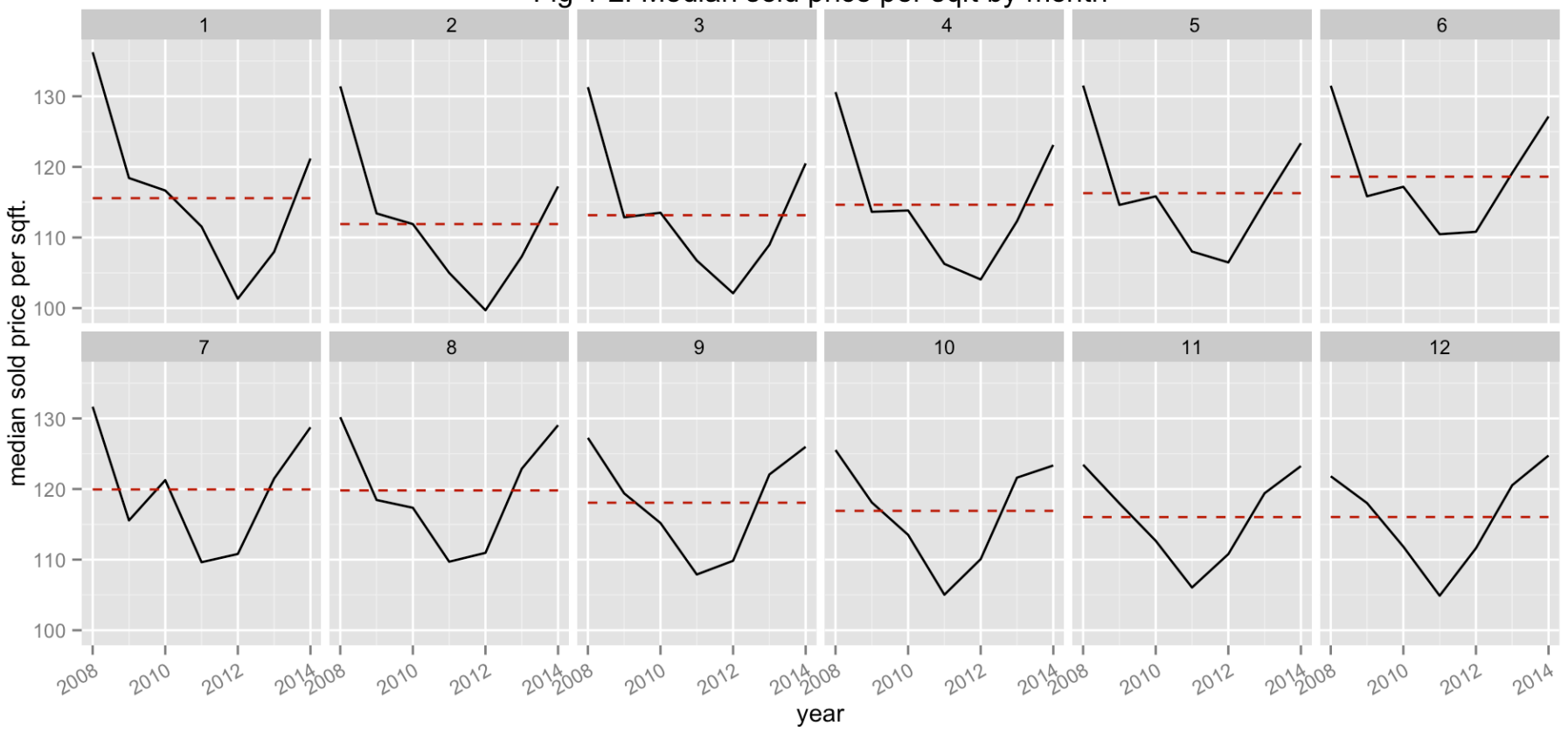


Fig 4-2 shows the median house price by month between 2008 and 2014. The median house prices (red dashed line) in June, July, and August are the highest among other months. Therefore, it confirms our early exploration that there's a clear seasonal trend in house prices, with a peak sometime in the summer. This finding indicates that the season does affect house prices in USA.

(c) Explore the relationship between Census data and house price (2008-2012)

Finally, I want to compare three key factors, house cost/income ratio, house age, and crime rate, to the house price.

Fig 4-3a: Responsiveness of house cost/income to price

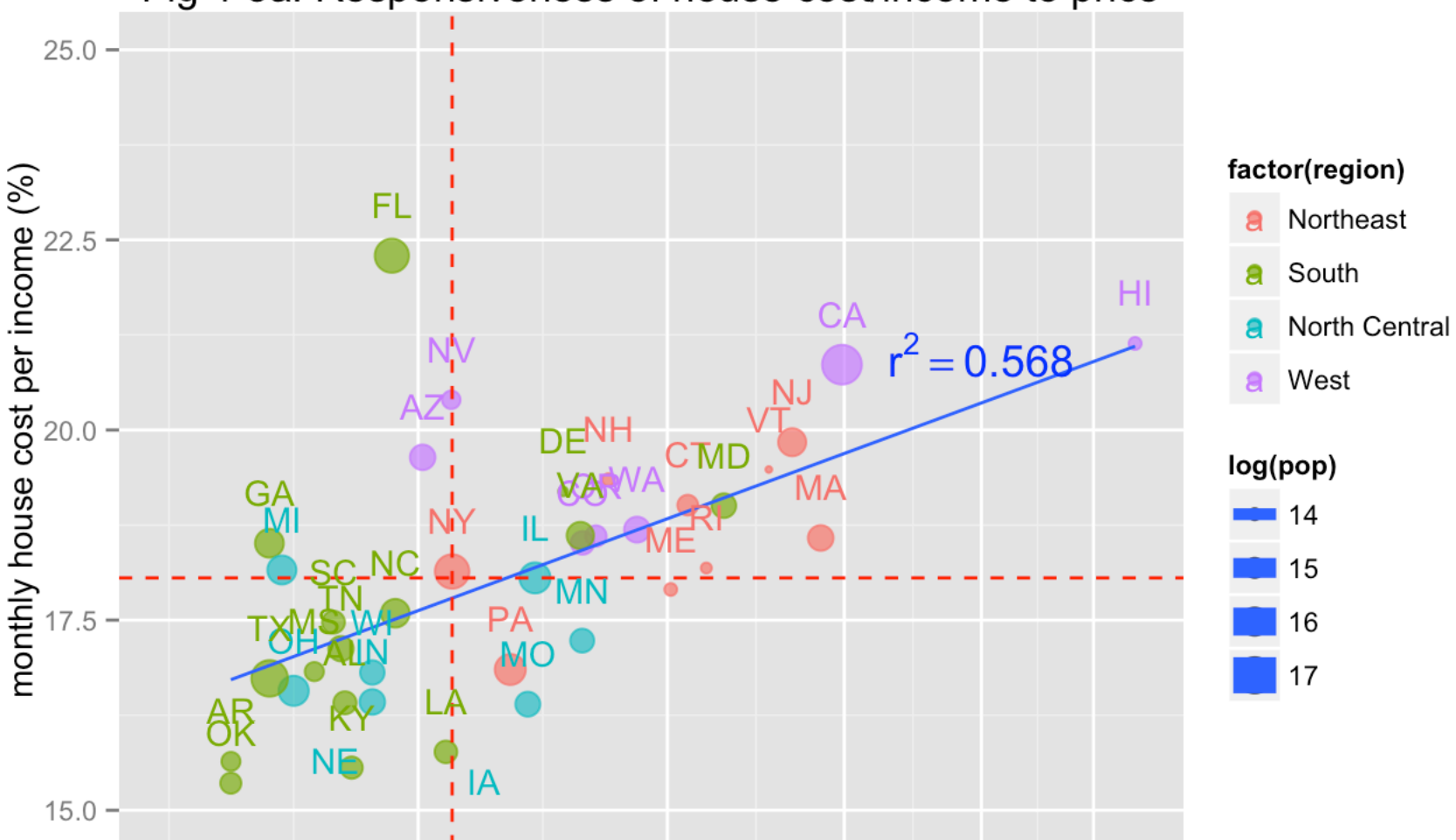
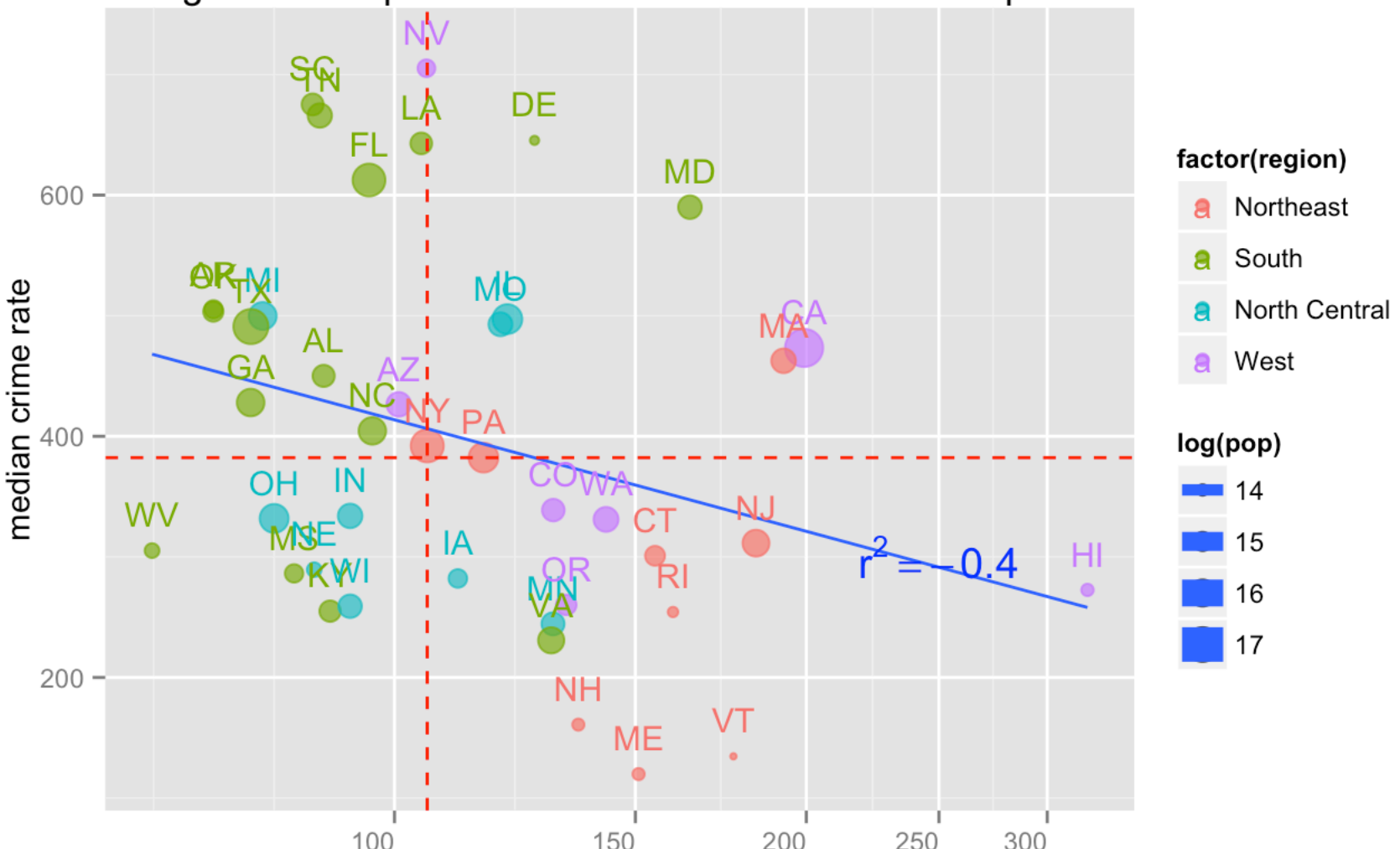


Fig 4-3b: Responsiveness of house year to price

100
150
200
250
300
median house price per sqft.

Fig 4-3c: Responsiveness of crime rate to house price



100
150
200
250
300

median house price per sqft.

- **Homeowership vs. house price**

The most expensive state for homeowners is FL, where they pay 22.3 % of their income in housing costs. Also, the homeowners in the West and Northeast need to pay most of homeowner's income in the housing costs. In most cases, the higher the house price, the higher the cost/income ratio.

- **House year vs. house price**

I expected to see new house has a higher sold price, but it seems to be not the case. The general trend was that older house gets higher sold price. Specifically, the Northeast has more old house, but their house prices are much higher than other regions. In contract, the West has more new house, but the house price in most West states are below the national median price.

- **Crime rate vs. house price**

As expected, the house price is in general inversely proportional to the crime rate. The South states has higher crime rate, while their house prices are much cheaper.

- **Summary**

What's interesting here is that the real estate markets have the trend with demographic features. We can use the median value of house price and the demographic features to found that there are four quadrants in the Fig 4-3.

Generally, the markets in the Northeast and West state are in the top-right corner in Fig 4-3a (or the bottom-right corner in Fig 4-3c). Most of these markets are in "hot economic" states, with high household income and low crime rate. In contrast, most of the South states are in the opposite corner, which has low income and high crime rate. These house markets are relatively cheap. What catches my eye here are the house in these markets are relatively new, so it might be a good idea for future investment.

5. Reflection

Struggles

One of the challenges I had with the data set was to deal with the format from different source. For example, I had a problem with grouping two dataframes by state with different sizes. Also, I found many values that were missing in the house data set from Zillow. Another challenge was to find useful information from data set to compare housing price. I attempted to find the school rating in state, but I can only found the data at the city level.

Successes

I was able to extract house price at select cities and states, and monitored the price changes with time. I was especially interested in the comparison of house sold price by month, and found a clear seasonal pattern for house prices. I also was enable to use a linear model to predict the house sold price. Although there are many other factors that can affect house price, the analysis in this project could be a starting point to better understand the real estate price in future.

Future Exploration

The real estate price largely depends on neighborhood. It would be interesting to analyze the data at city level. The data for future exploration could be unemployment rate, school rating, distance to major employment centres, accessibility to highways, and etc. However, I found these data sets are not easy to access. If one could obtain these data set, it would give clues of how these features affect house price. This information could be very useful for future investments and exploration.