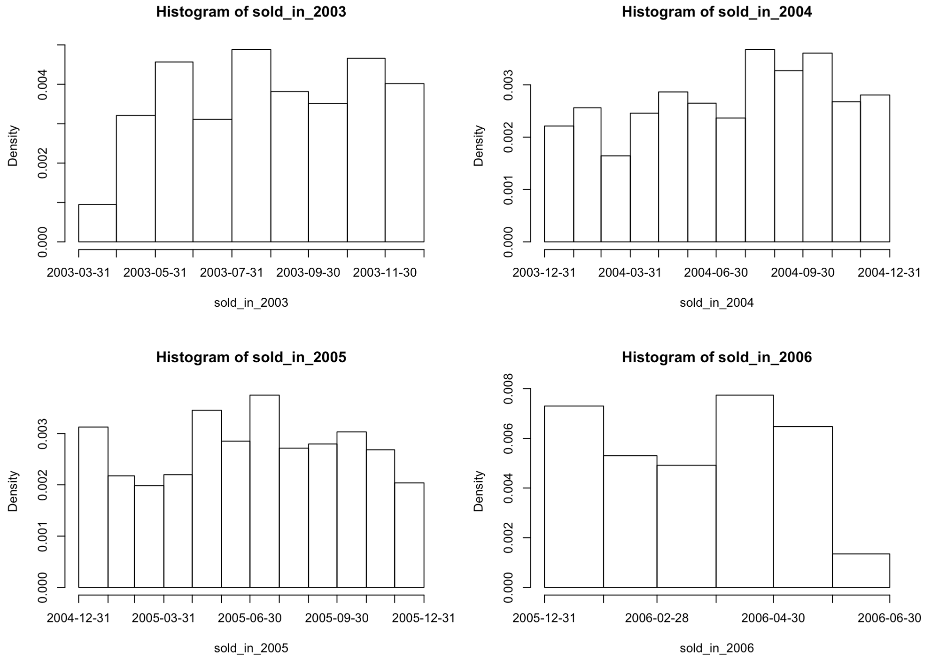
Name: Hao Luo
ID: 912423597

Qn2:
Timespan for housing sales is 1134 days. Timespan of the construction dates is 120 years.

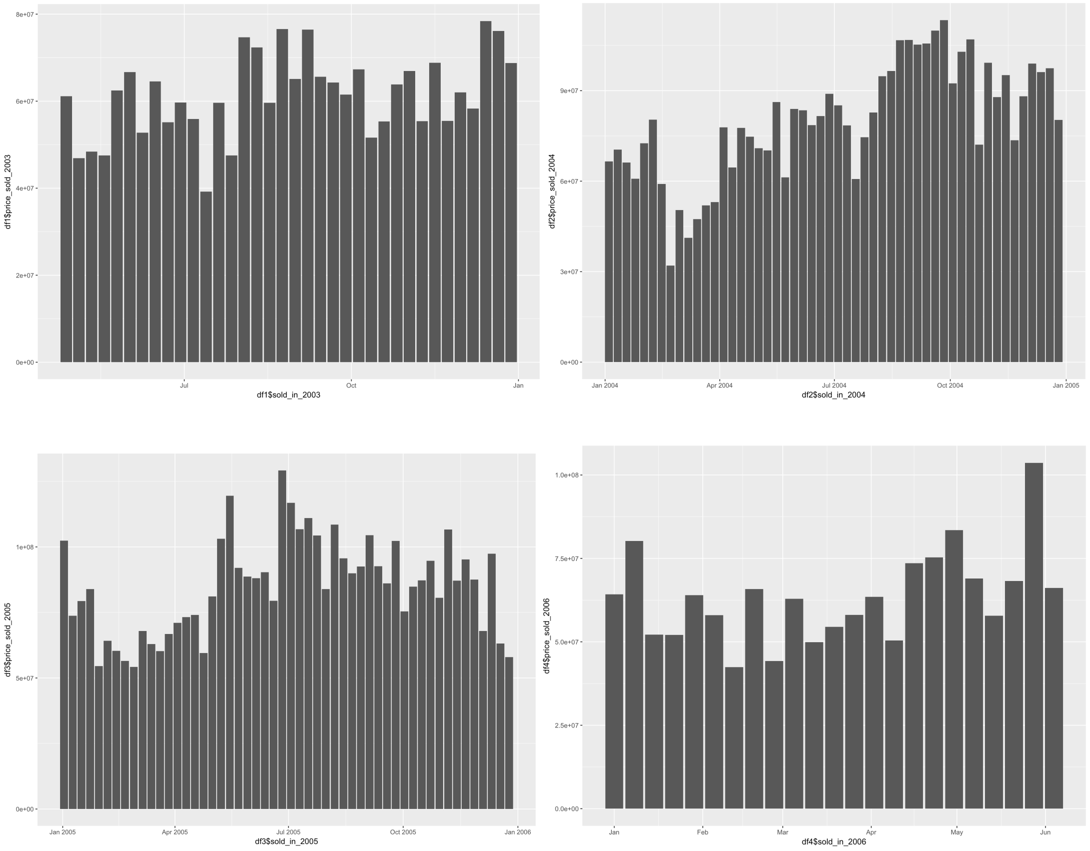
Qn3:
Numbers of sales over time.



In histogram 2006, on the Y-axis, numbers of sales is highest of all.

Comparing overall four graphs from 2004 to 2006. From June to September, numbers of sales are the highest over the year. Data suggest that, numbers of sales is likely to be highest from June to September in the future.

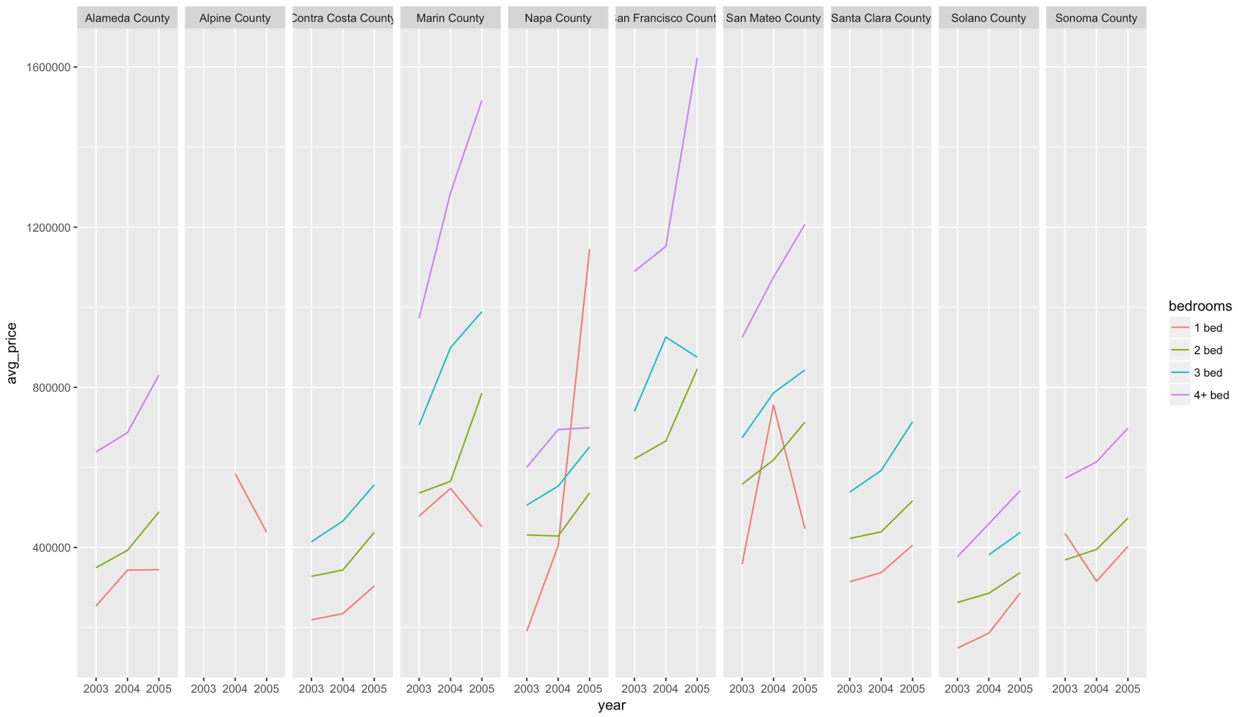
Average house price over time.



From the graphs, data indicates that the overpricing is increasing over the years.

Data also suggest economics booms over the years in housing markets.

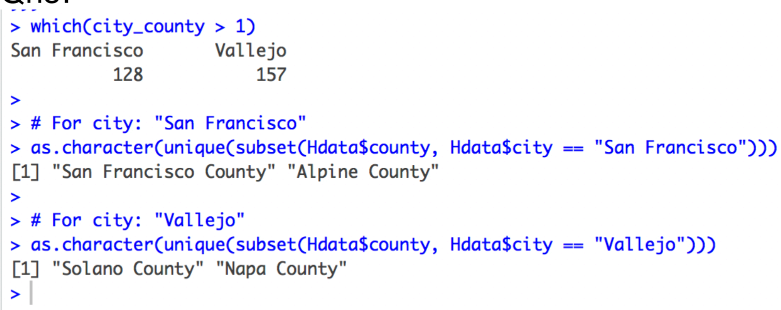
Qn4:



Marin County , San Francisco County, and San Mateo County are the most expesive county of all. In addition, the sales of 4+ bed housing are dramatically increasing as pricing going up. Alpine County only has one bedroom sales over the year. Overall, the prive of housing in all countys are going up for the most part. Strangely, the single bedroom in San Mateo Count started decreasing after 2014.

In Napa County, most single bedrooms houses are most popular and in demend. Due to the fact that Napa is well known for red wines. It might attract many tourists and therefore increases demands for singgle bedrooms.

Qn5:

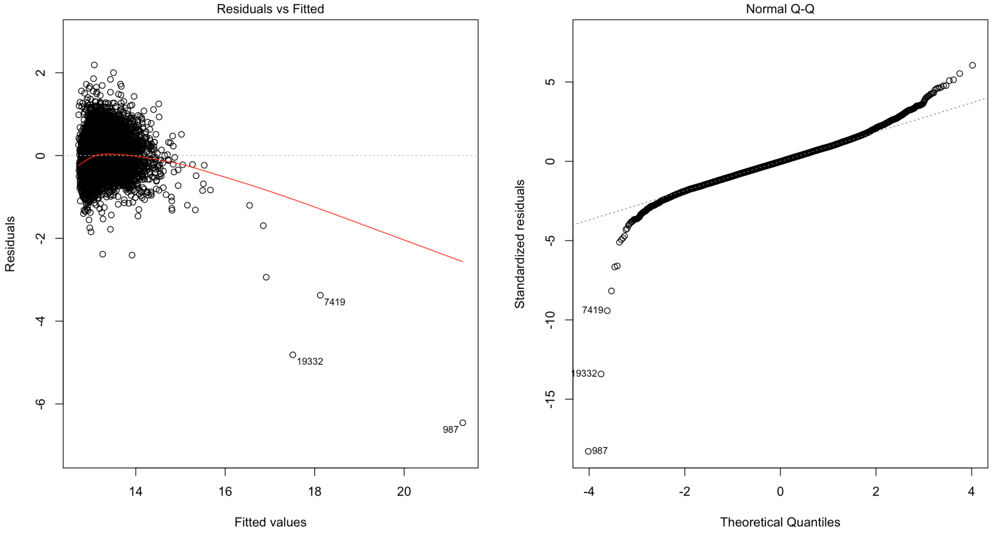


```
> which(city_county > 1)
San Francisco       Vallejo
          128           157
>
> # For city: "San Francisco"
> as.character(unique(subset(Hdata$county, Hdata$city == "San Francisco")))
[1] "San Francisco County" "Alpine County"
>
> # For city: "Vallejo"
> as.character(unique(subset(Hdata$county, Hdata$city == "Vallejo")))
[1] "Solano County" "Napa County"
>
```

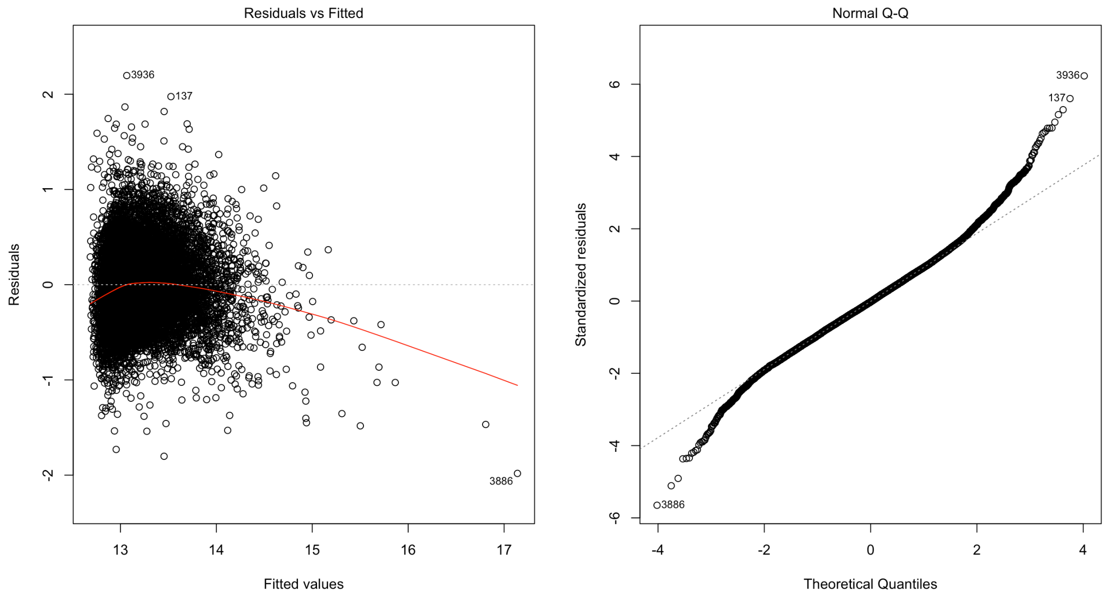
San Francisco and Vallejo have more than one counties dues to the size of the cities. They both have more than one cities.

Qn6:

Before removing outliners:

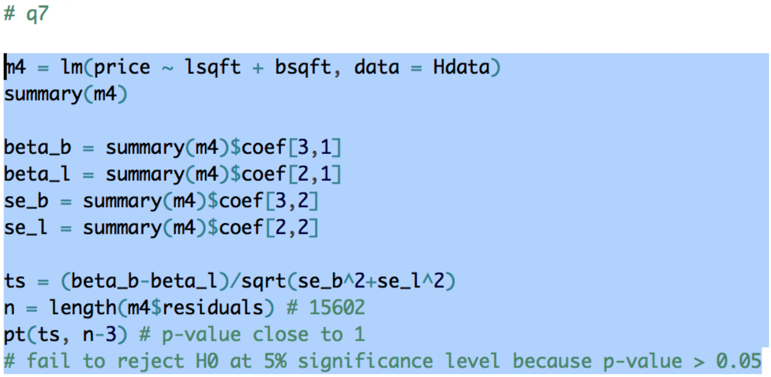


After removing outliers:



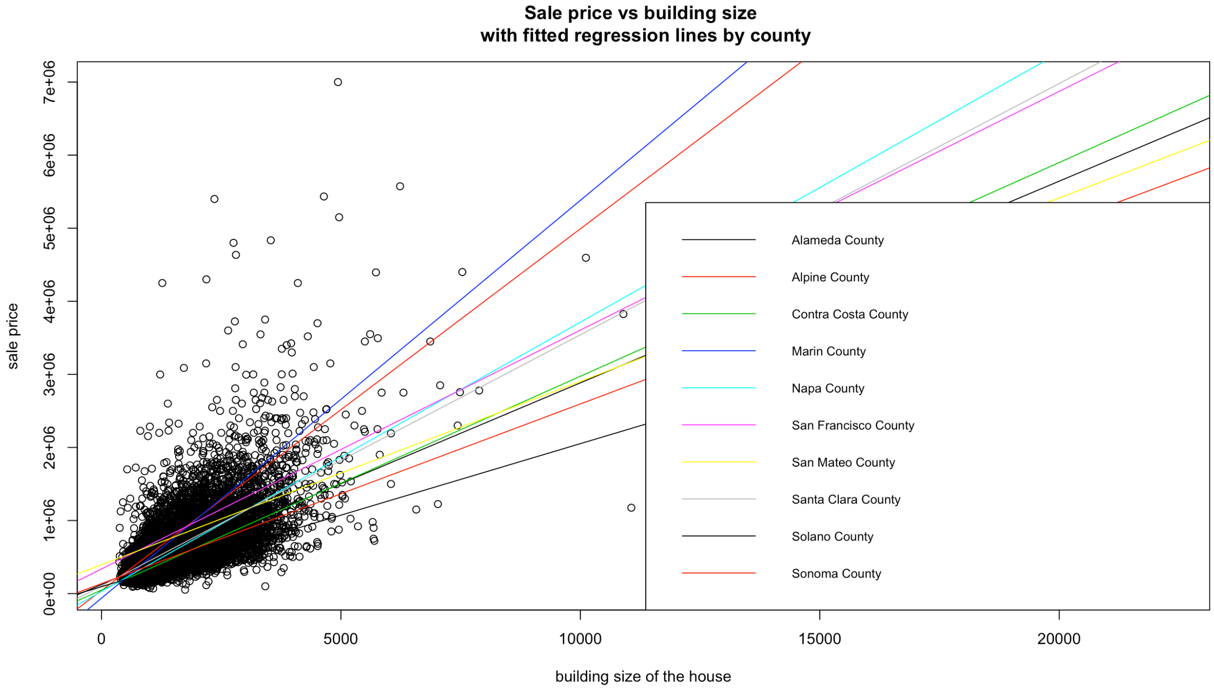
The data is still not normal aftter removing outliers.

Qn7:



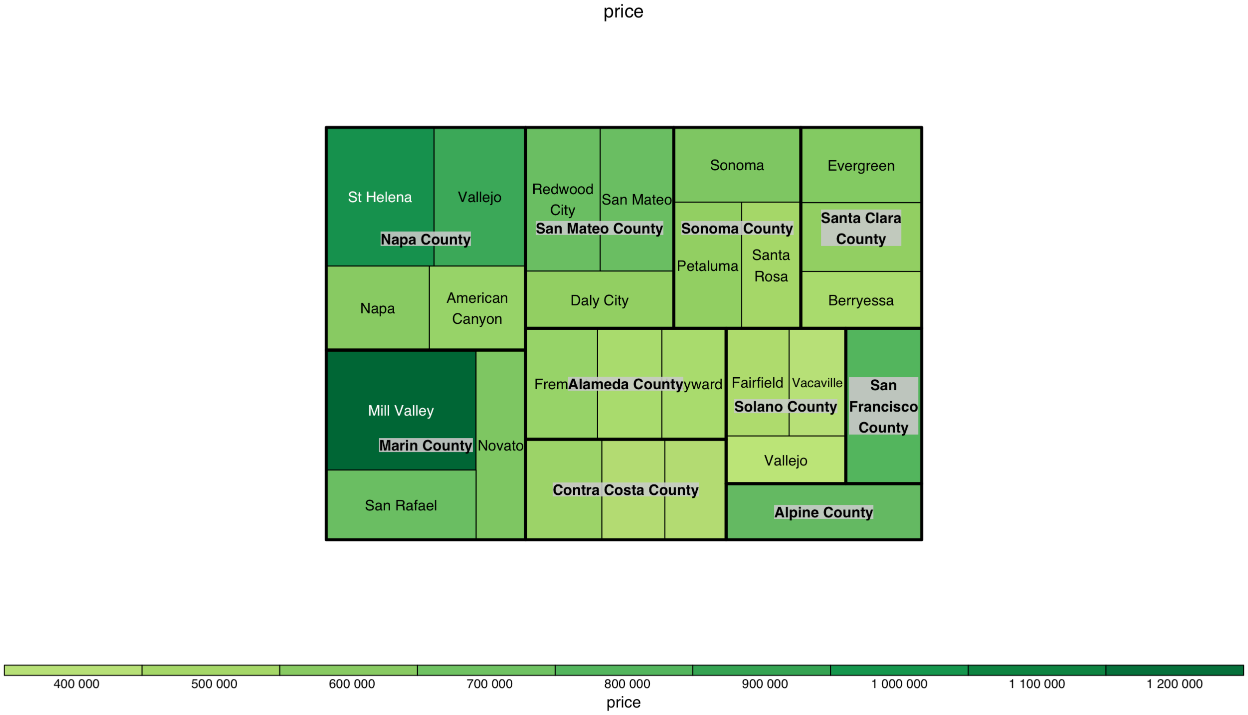
```
# q7

m4 = lm(price ~ lsqft + bsqft, data = Hdata)
summary(m4)

beta_b = summary(m4)$coef[3,1]
beta_l = summary(m4)$coef[2,1]
se_b = summary(m4)$coef[3,2]
se_l = summary(m4)$coef[2,2]

ts = (beta_b-beta_l)/sqrt(se_b^2+se_l^2)
n = length(m4$residuals) # 15602
pt(ts, n-3) # p-value close to 1
# fail to reject H0 at 5% significance level because p-value > 0.05
```

Qn8:



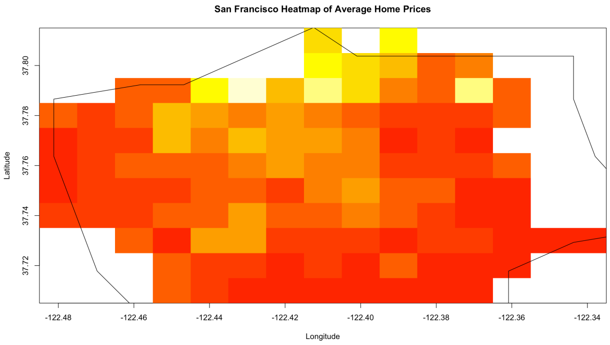
They are not parallel. Data suggesting, they are not depend on the county.

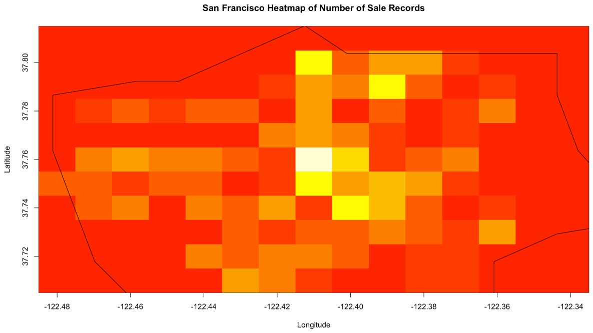
Qn9:



Marin County is the most expensive county of all. Solano County has the lowest prices of all.

Qn10:


San Francisco Heatmap of Average Home Prices


San Francisco Heatmap of Number of Sale Records

On Heat map, data suggest Marin county has most sales made and the frequency is the highest.

The yellow square indicates higher frequency, whereas the red indicates no sales.