Exploring the housing crisis

with ggplot2 and plyr

Hadley Wickham

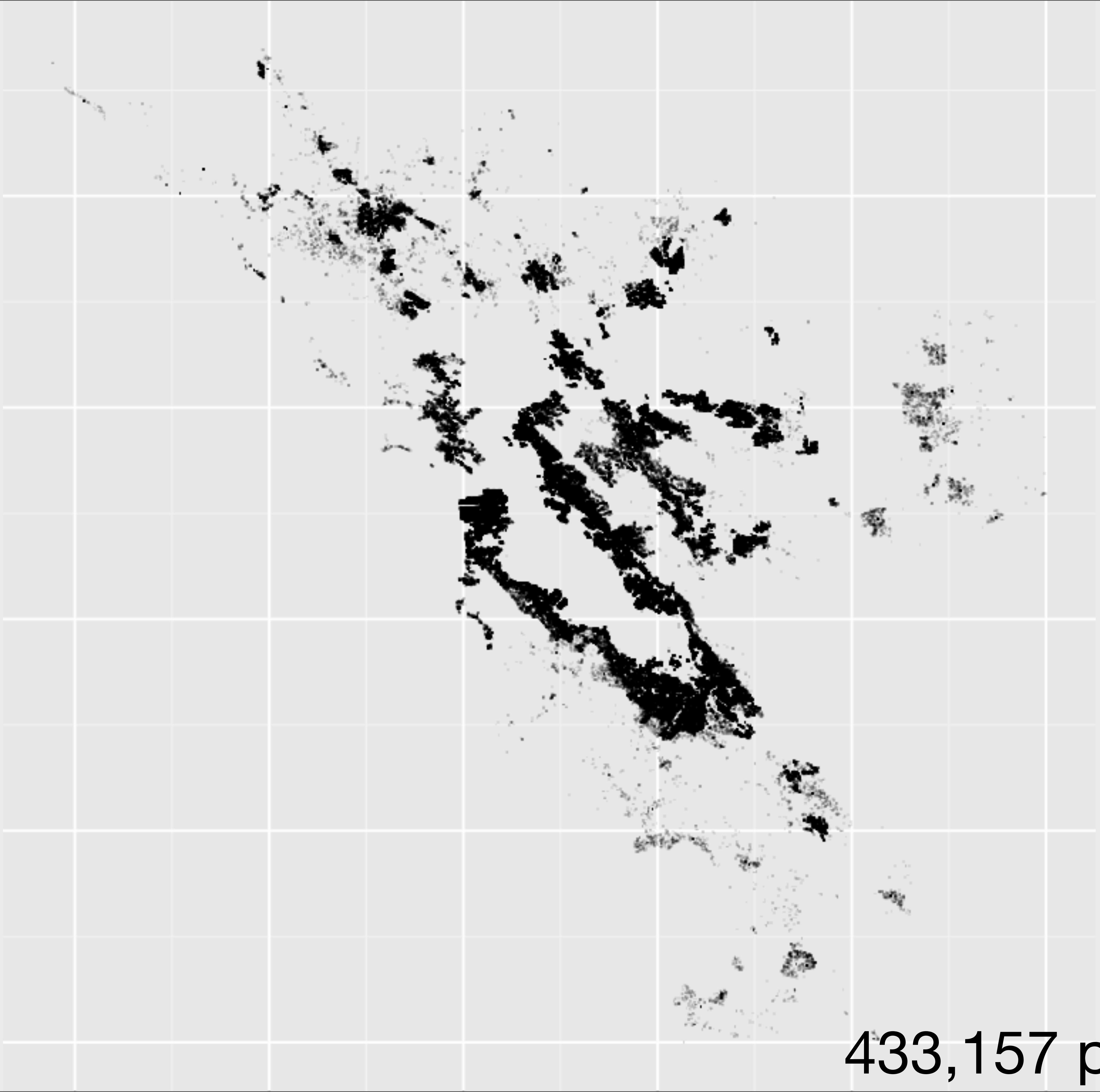
1. Introduction to the data
2. Basic graphics
   (histograms & scatterplots)
3. Adding extra variables
   (aesthetics and facetting)
4. Time series
   (using plyr to manipulate data)
5. Resources

About the data

521,495 house sales scraped from SF Chronicle, March 2003—November 2008.

Addresses geocoded using USC WebGIS. (83% to interpolated city block or better)

Variables: longitude, latitude, date, price, bedrooms, area

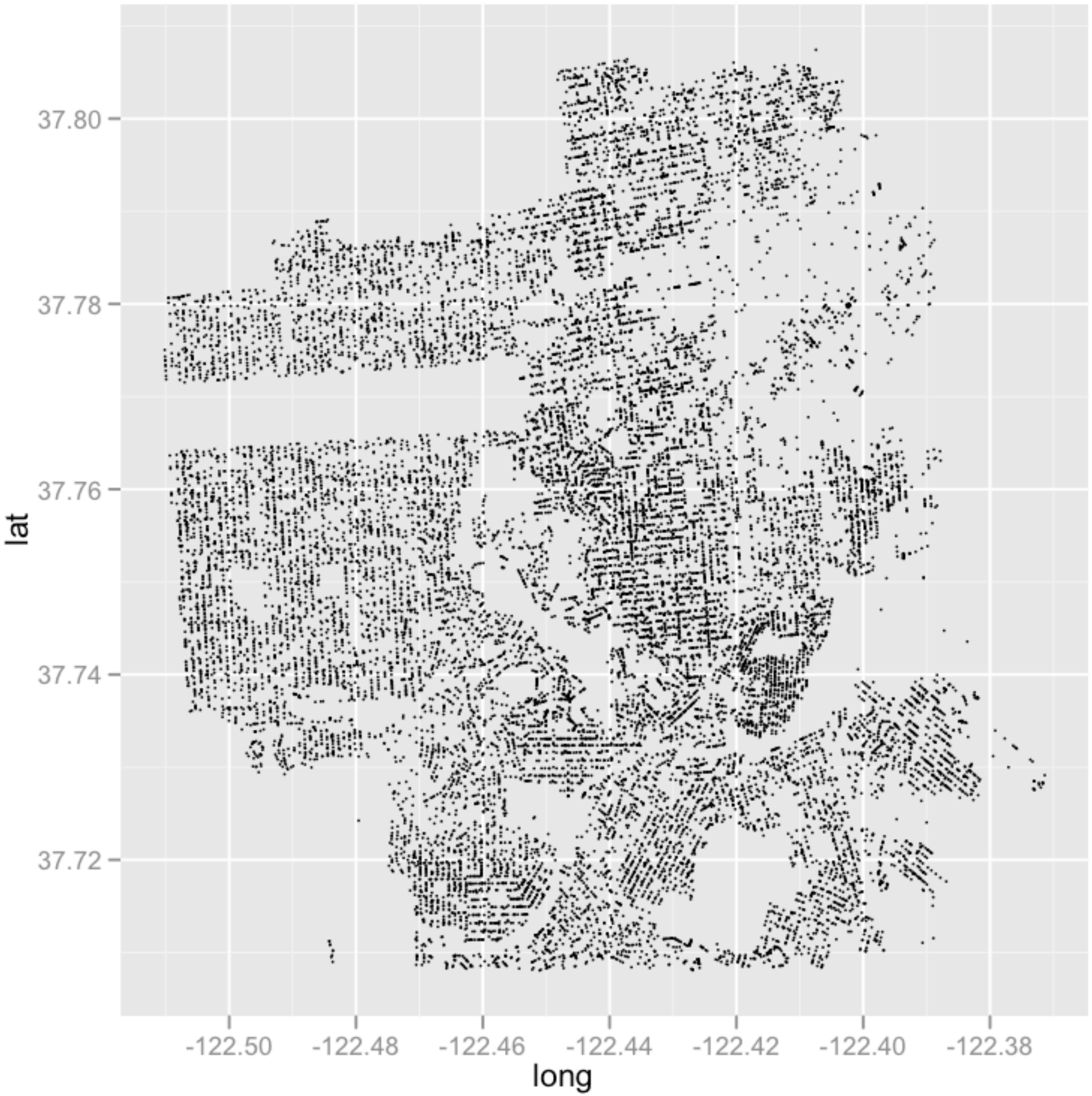


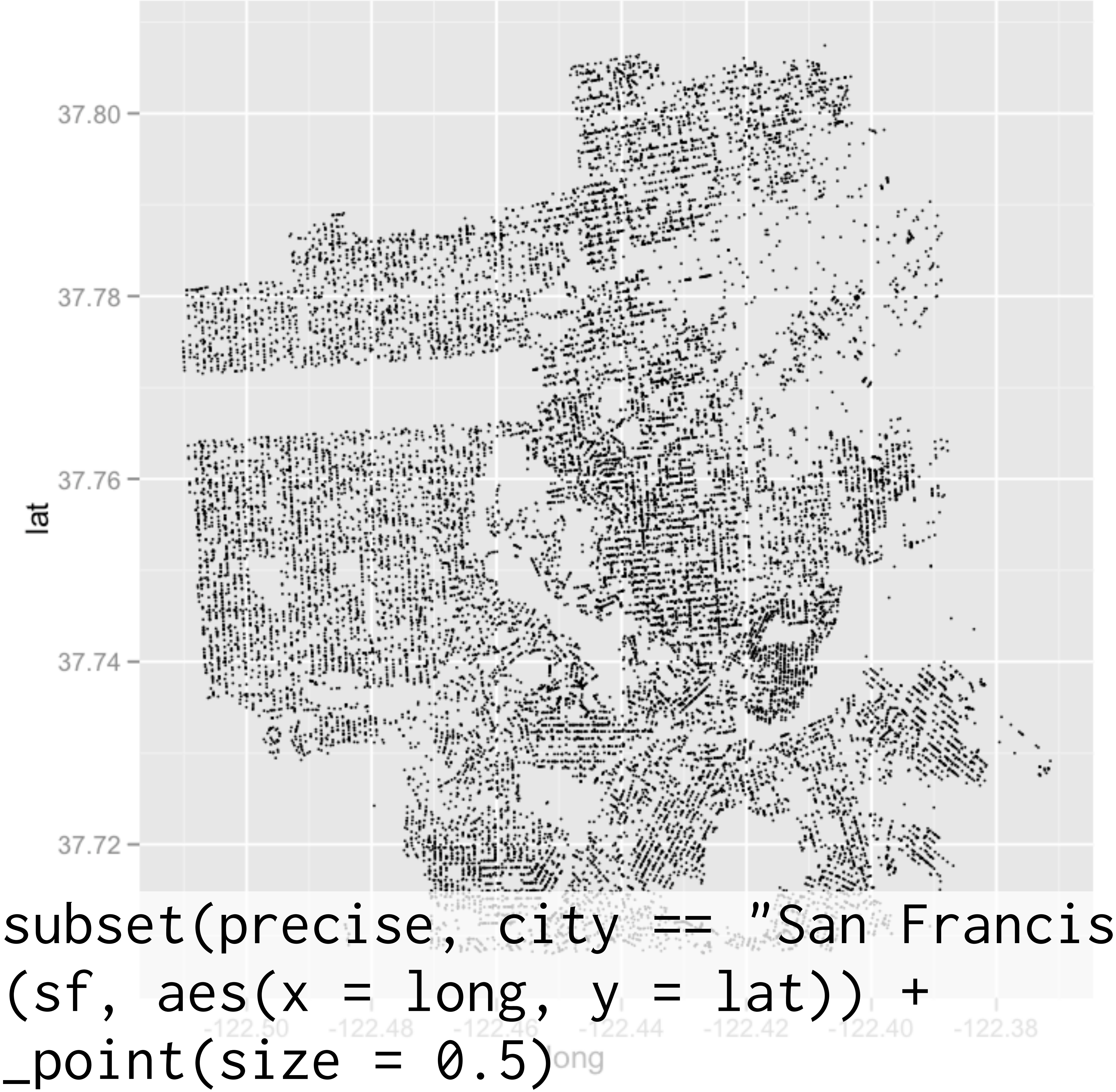
433,157 points!

Basic graphics:

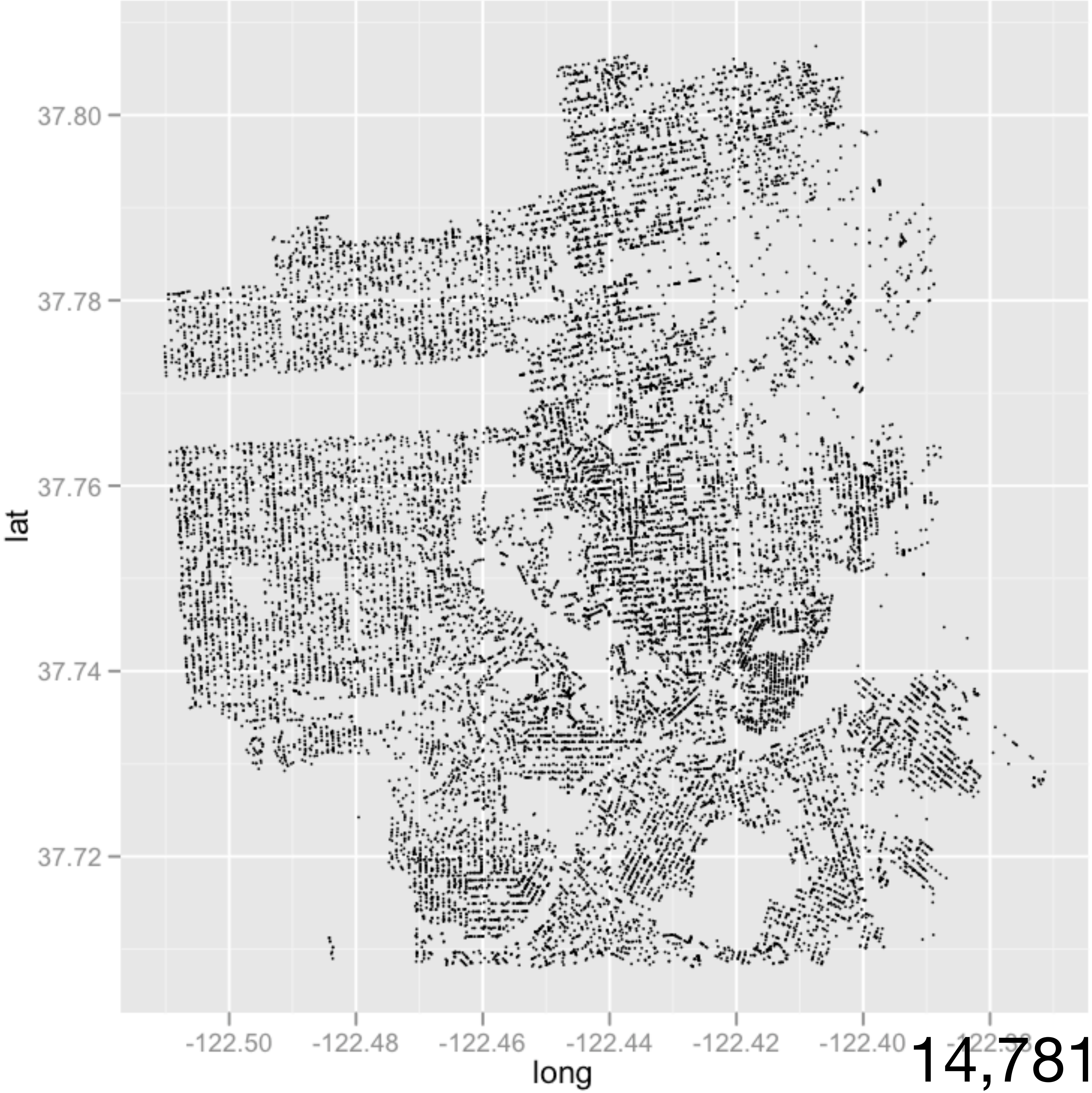
Scatterplots, bar charts and histograms

Where are the houses?

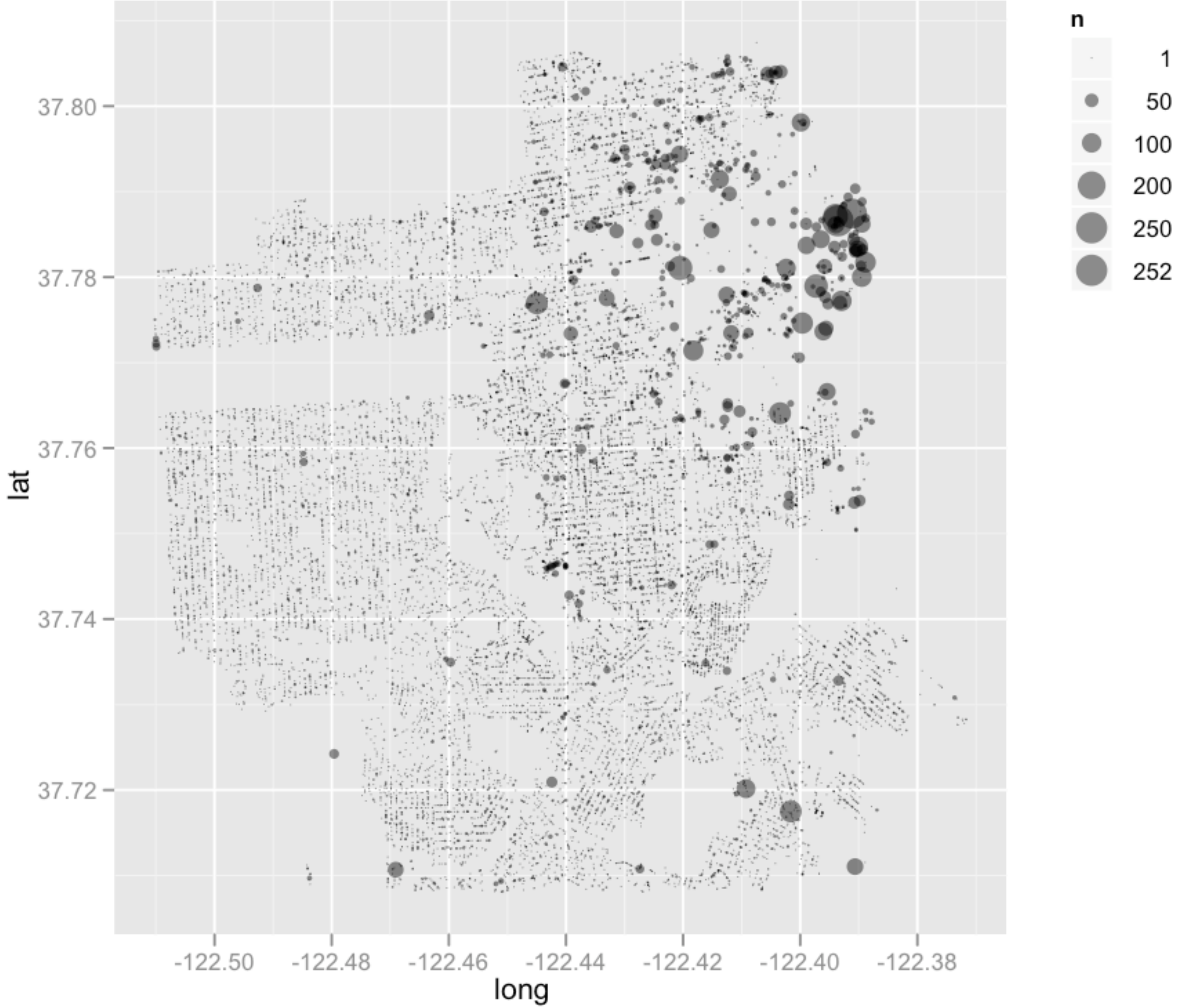

37.80
37.78
37.76
37.74
37.72
lat
-122.50
-122.48
-122.46
-122.44
-122.42
-122.40
-122.38
long

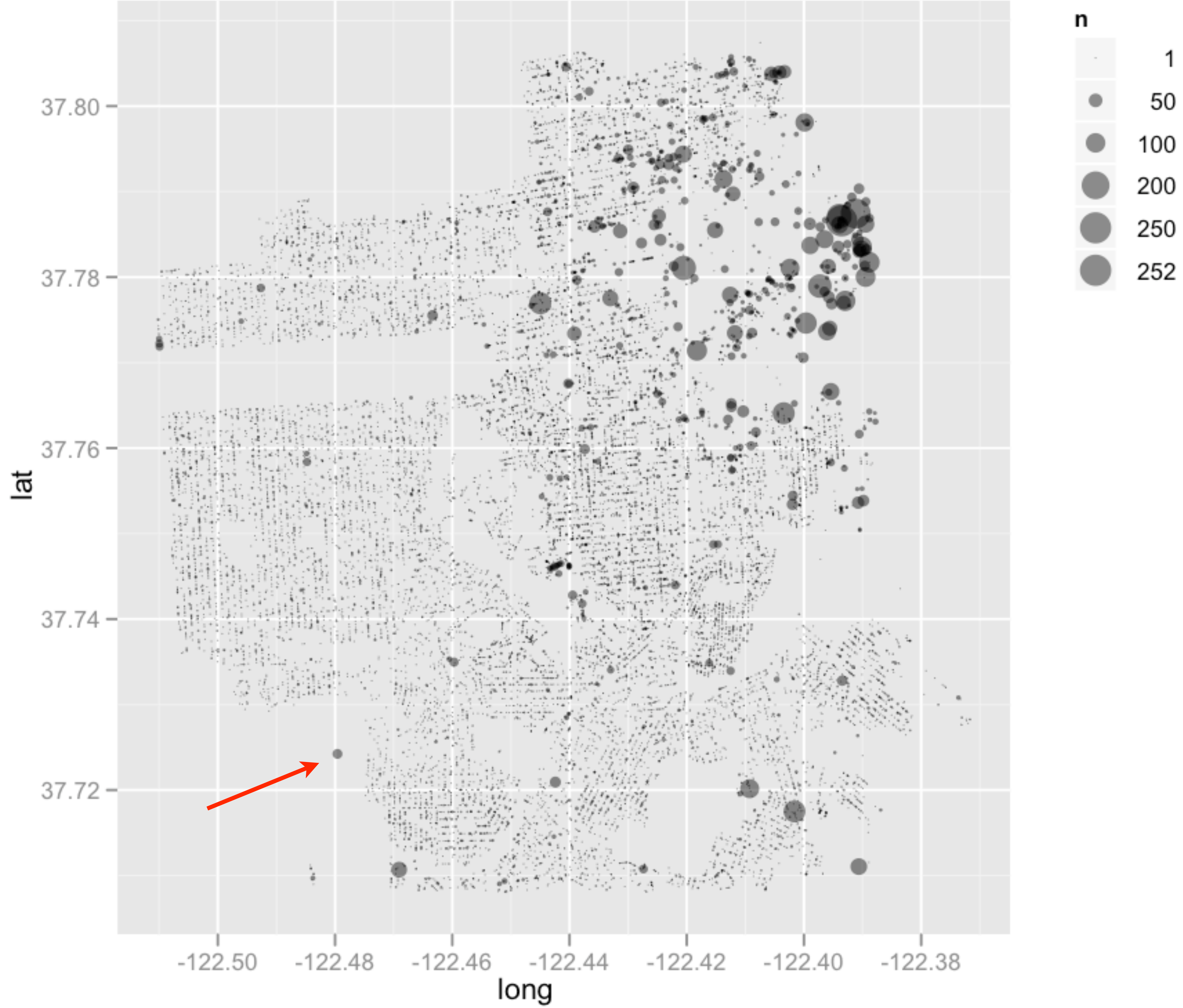


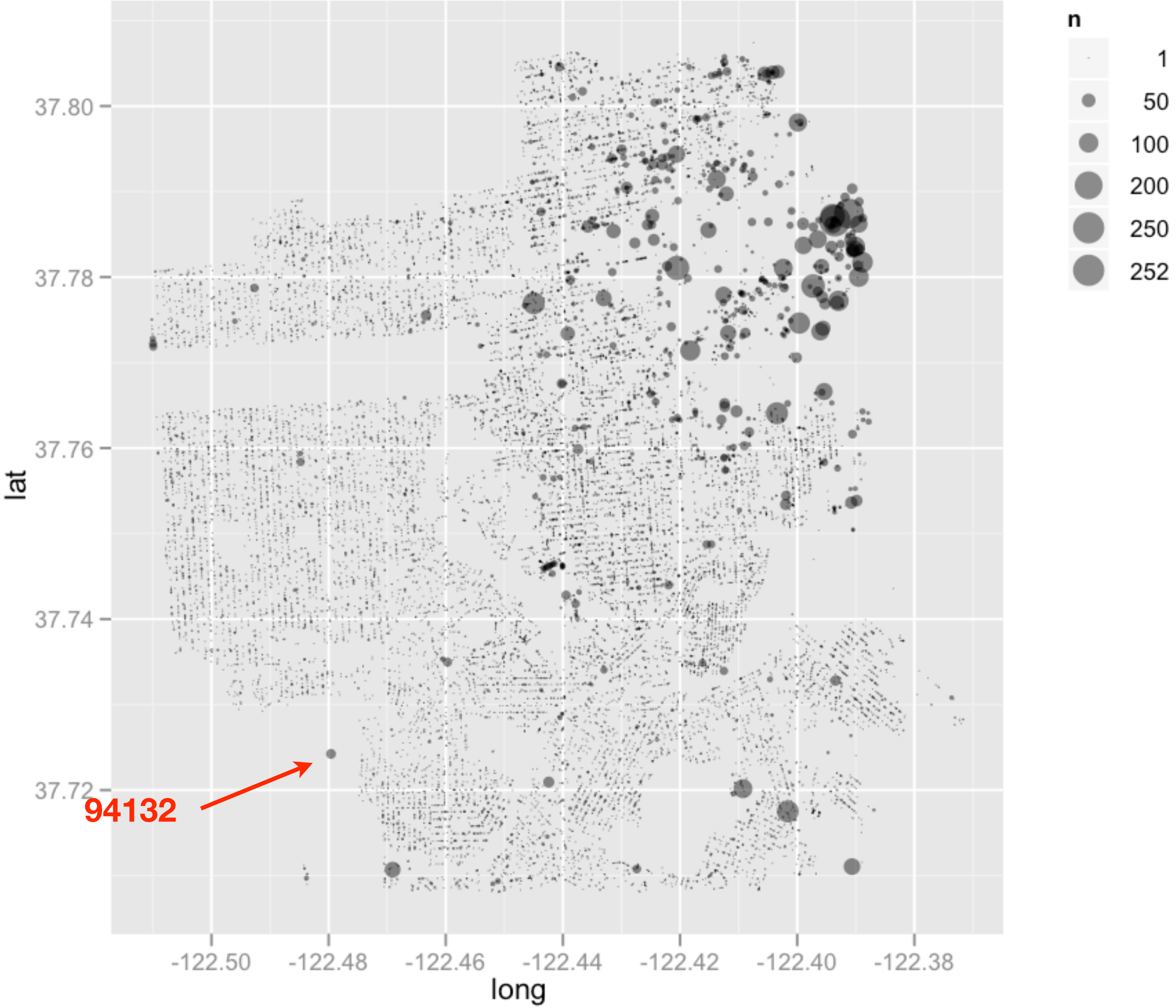
```
sf <- subset(precise, city == "San Francisco")
ggplot(sf, aes(x = long, y = lat)) +
  geom_point(size = 0.5)
```



14,781 points!

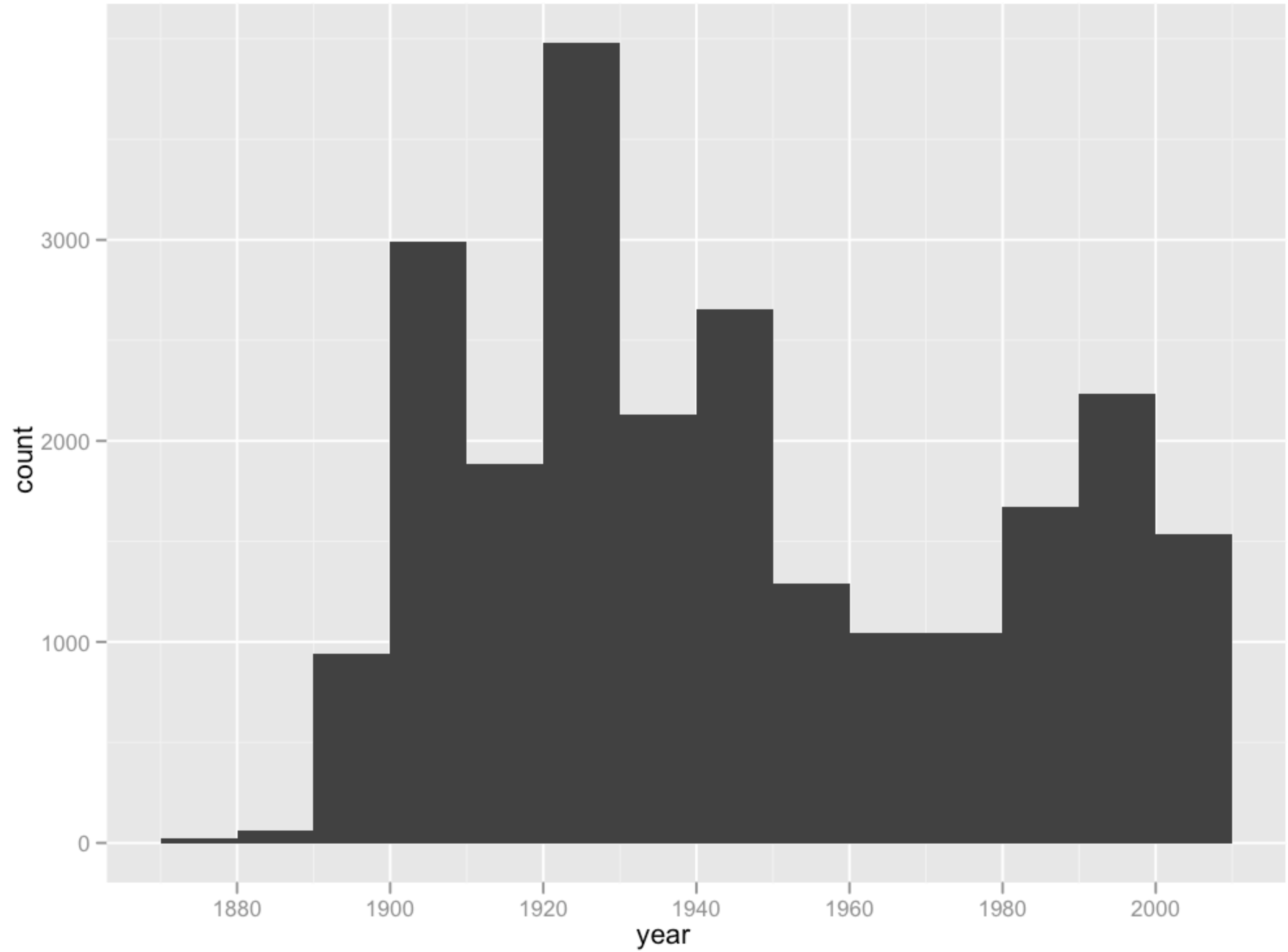

n
1
50
100
200
250
252
lat
37.80
37.78
37.76
37.74
37.72
long
-122.50
-122.48
-122.46
-122.44
-122.42
-122.40
-122.38

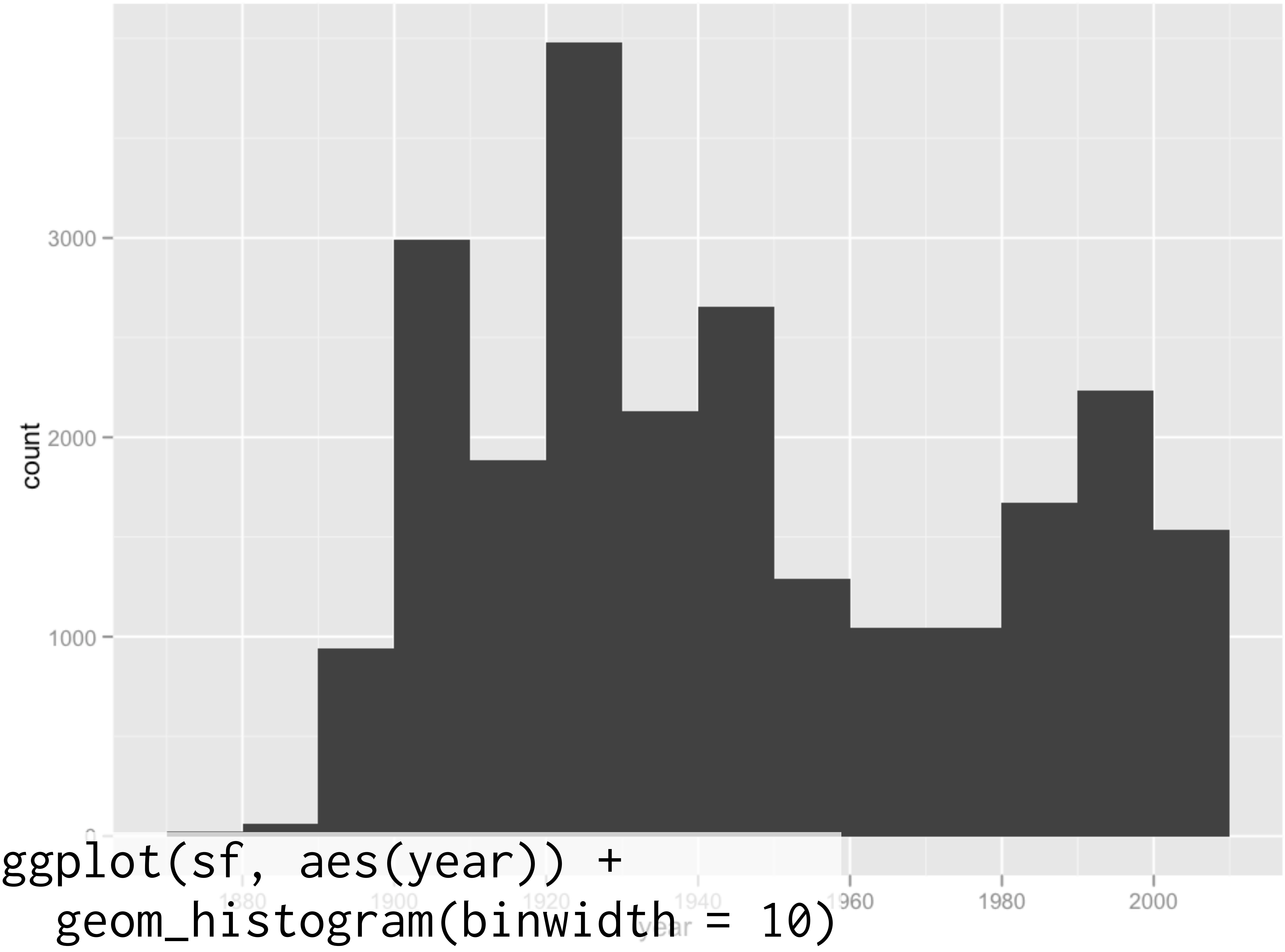

n
1
50
100
200
250
252
37.80
37.78
37.76
37.74
37.72
lat
-122.50
-122.48
-122.46
-122.44
-122.42
-122.40
-122.38
long


n
1
50
100
200
250
252
37.80
37.78
37.76
37.74
37.72
lat
-122.50
-122.48
-122.46
-122.44
-122.42
-122.40
-122.38
long
94132

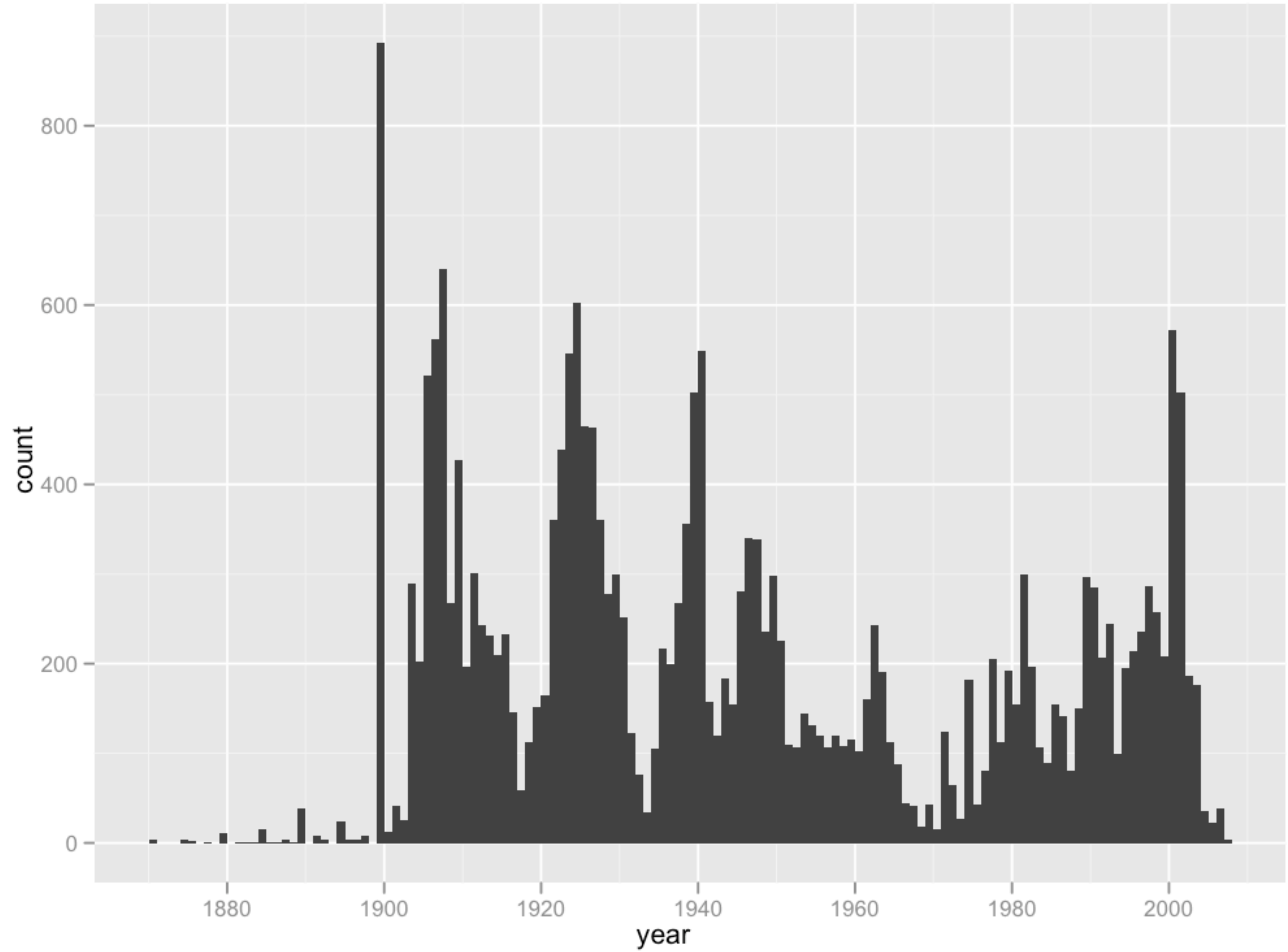
```
# In principle:
ggplot(sf, aes(long, lat)) +
  geom_point(stat = "sum")

# In practice:  (takes 42 seconds)
sfsum <- ddply(sf, c("lat", "long"), summarise,
  n = length(lat),
  avg_year = mean(year, na.rm = TRUE),
  .progress = "text"
)

ggplot(sfsum, aes(long, lat, size = n)) +
  geom_point(alpha = 1/2) +
  scale_area(to = c(0.3, 6), breaks = c(1, 50, 100, 200, 252))
```

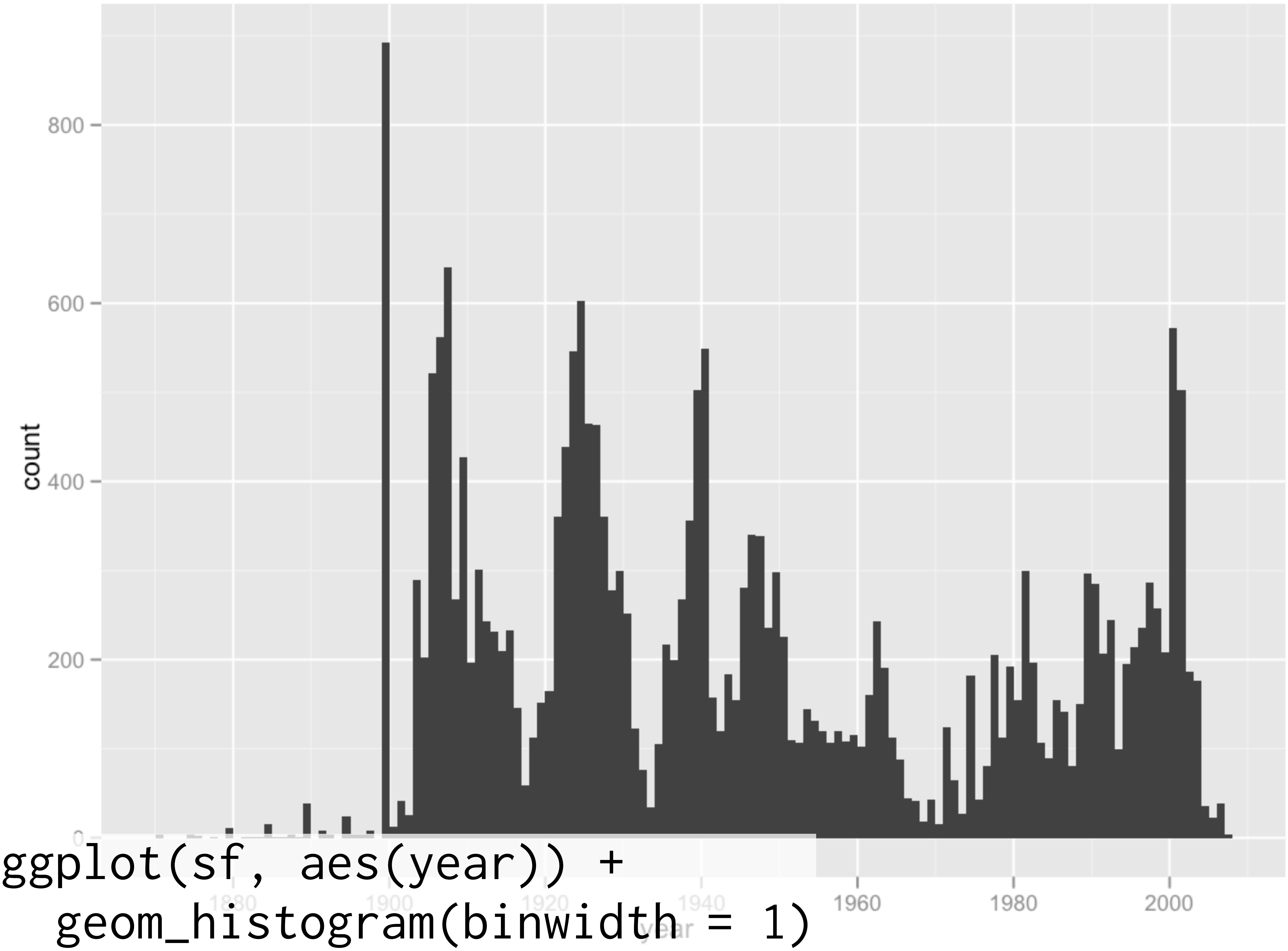
When were they build?


count
3000
2000
1000
0
1880
1900
1920
1940
1960
1980
2000
year

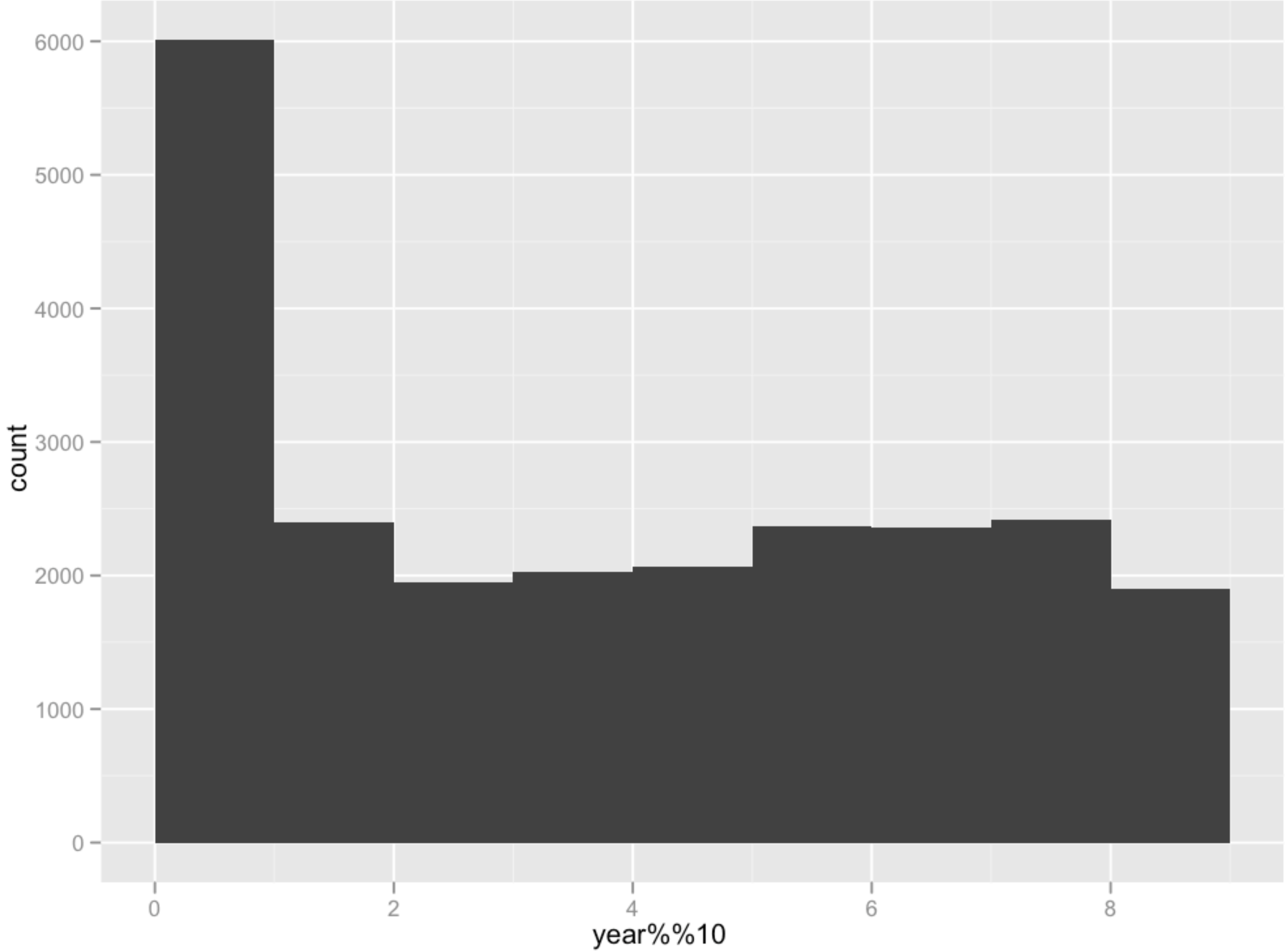


```
ggplot(sf, aes(year)) +
  geom_histogram(binwidth = 10)
```


count
800
600
400
200
0
1880
1900
1920
1940
1960
1980
2000
year



```
ggplot(sf, aes(year)) +
  geom_histogram(binwidth = 1)
```


6000
5000
4000
3000
2000
1000
0
count
0
2
4
6
8
year%%10

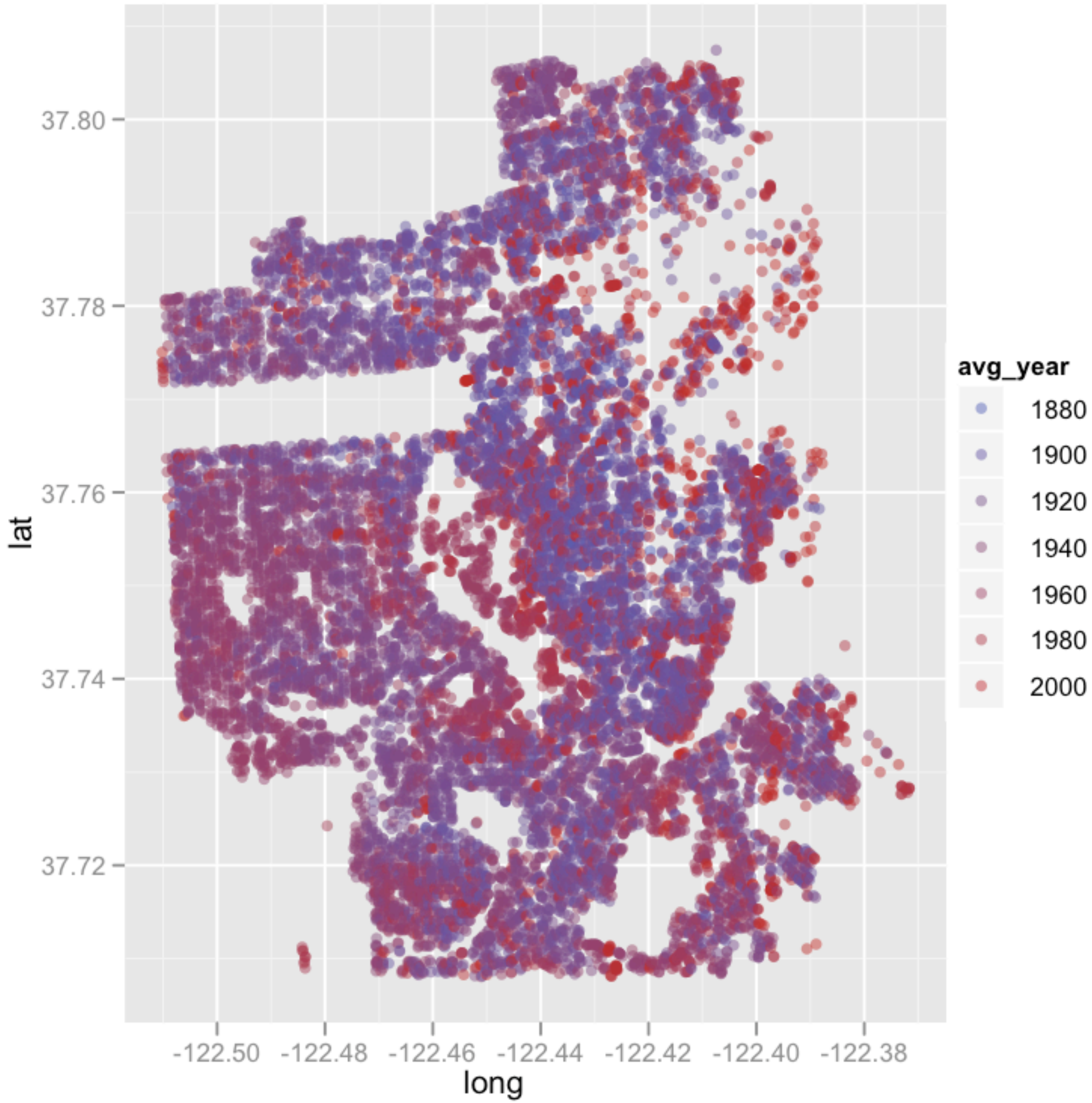
How does location vary with age?

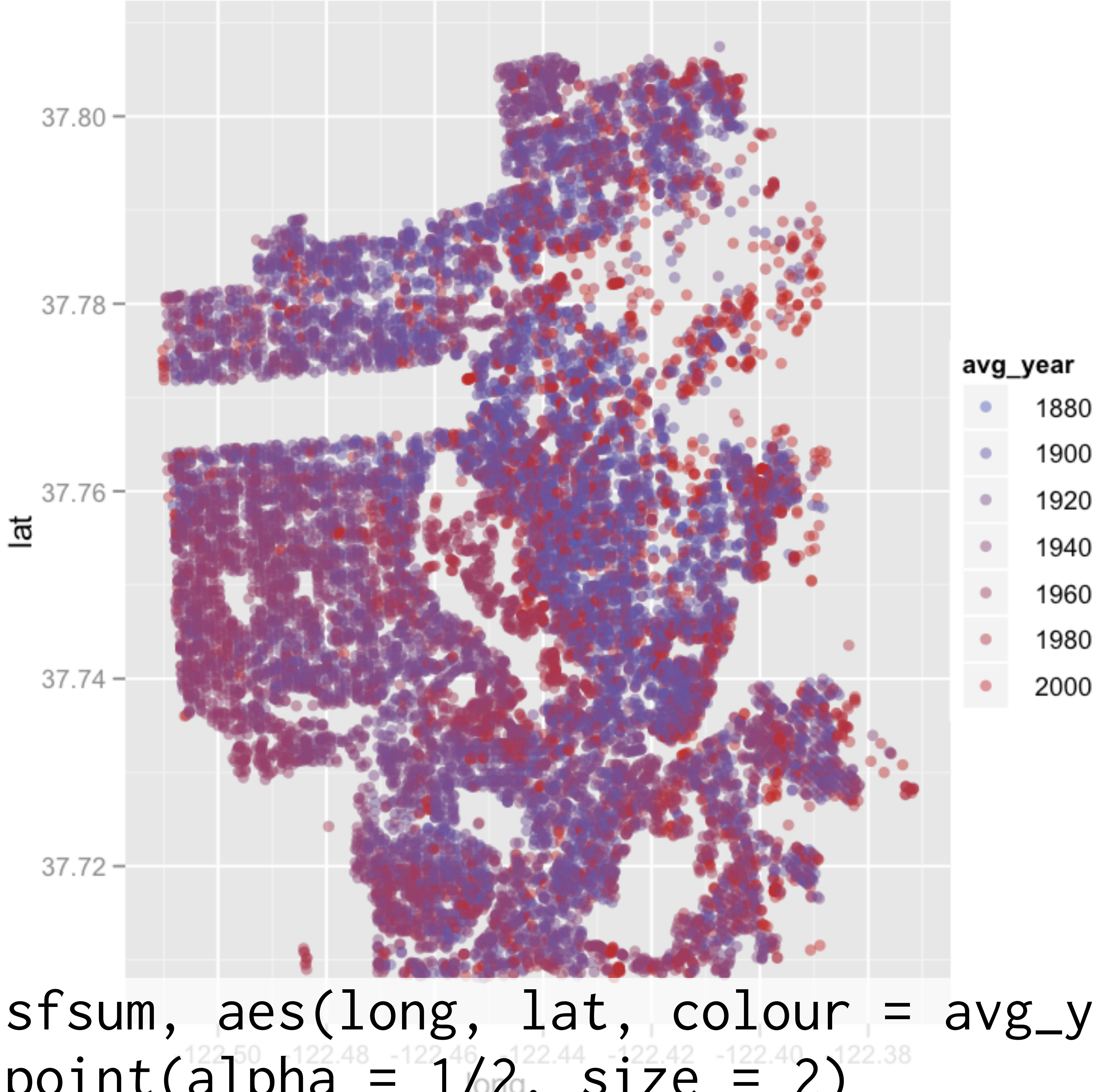
Additional variables

Can supplement any existing plot with additional variables in two ways: adding **aesthetics**, or creating **facets**.

Saw example of using size. Other aesthetics are: colour, fill, shape, linetype, alpha.

Facetting creates small multiples of subsets of the data.


37.80
37.78
37.76
37.74
37.72
lat
-122.50
-122.48
-122.46
-122.44
-122.42
-122.40
-122.38
long
avg_year
● 1880
● 1900
● 1920
● 1940
● 1960
● 1980
● 2000



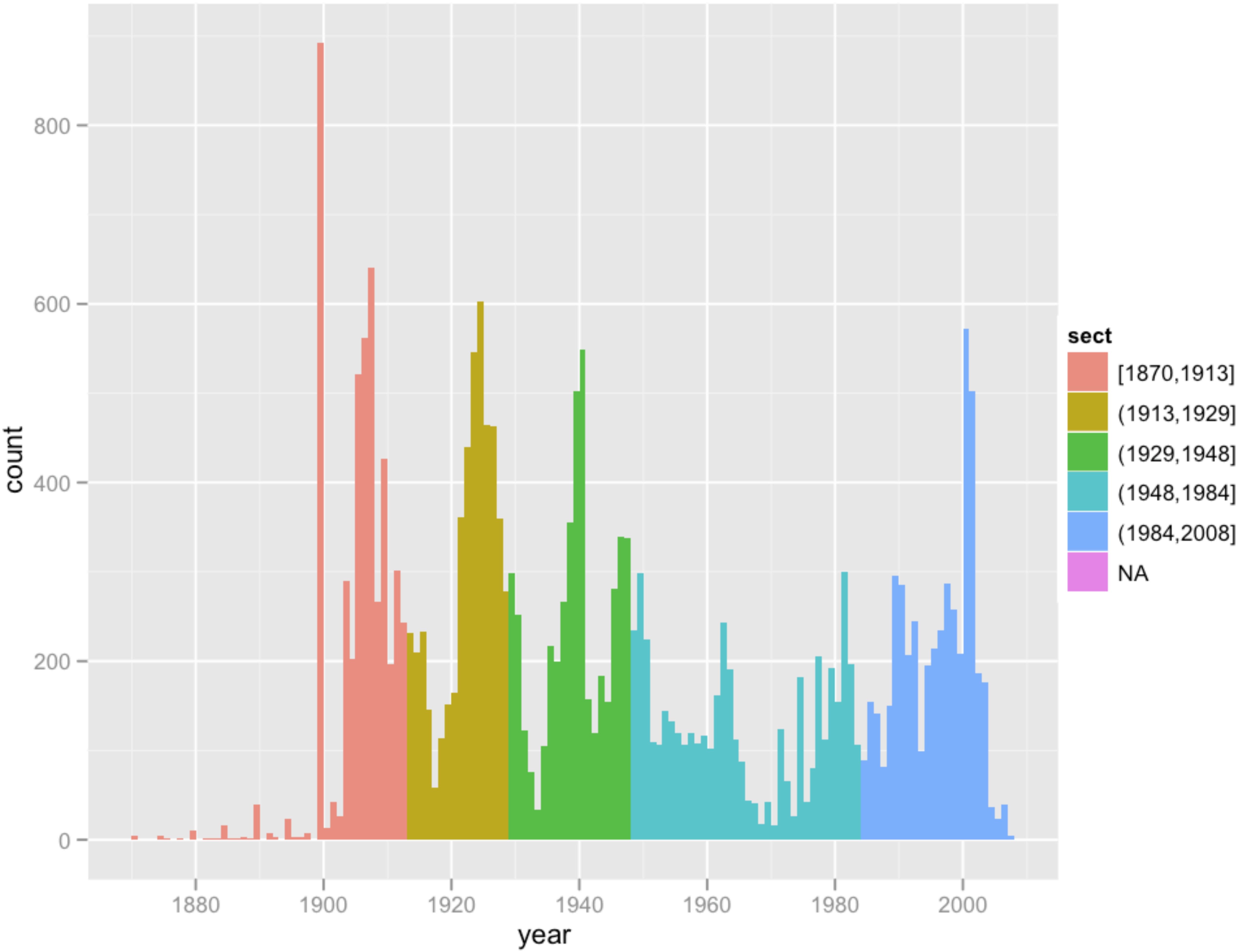
```
ggplot(sfsum, aes(long, lat, colour = avg_year)) +
  geom_point(alpha = 1/2, size = 2)
```

Interactive demo

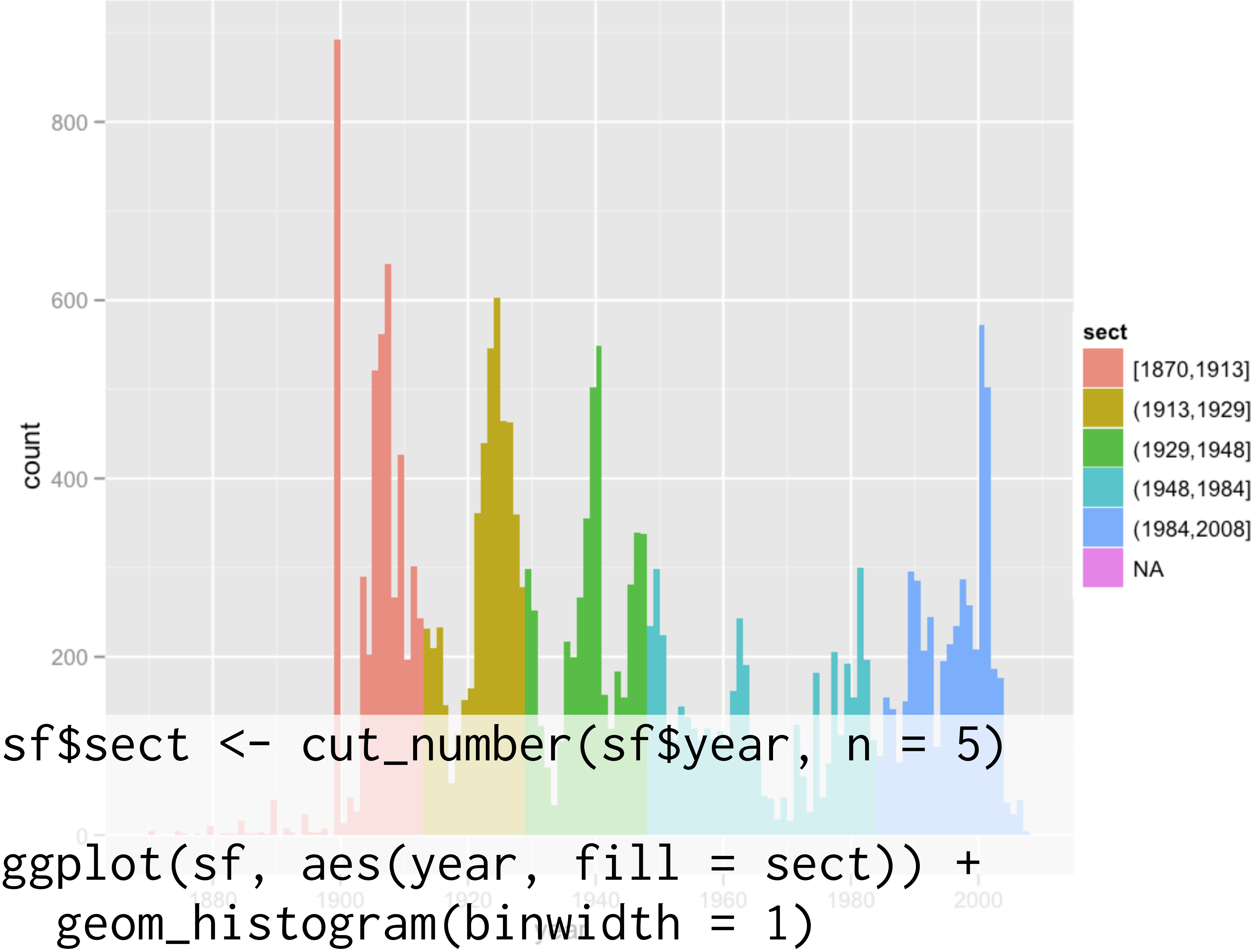
(a taste of future directions)

Let’s try facetting

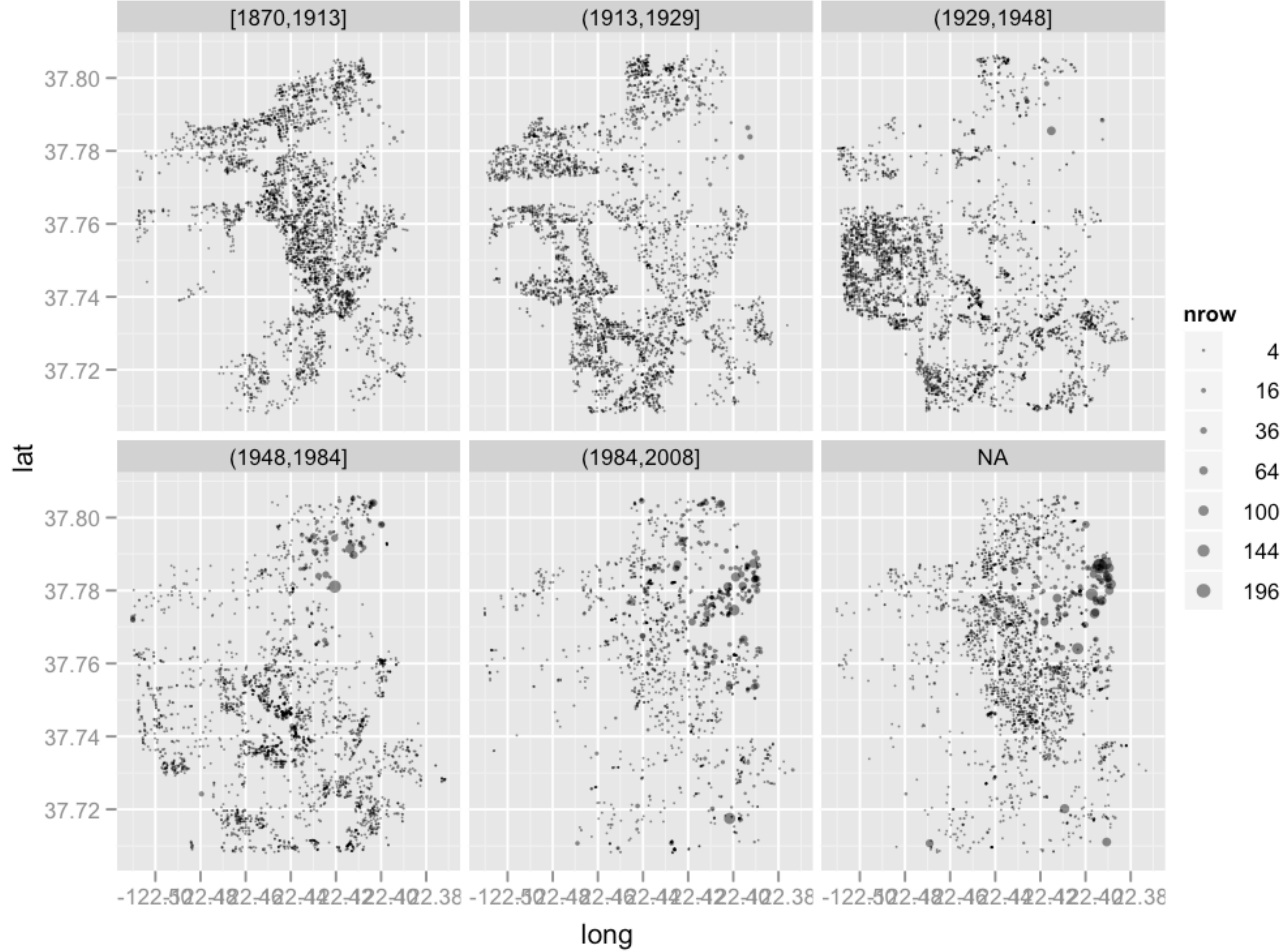
First, need a categorical variable to facet by



800
600
400
200
0
count
1880
1900
1920
1940
1960
1980
2000
year
sect
[1870,1913]
(1913,1929]
(1929,1948]
(1948,1984]
(1984,2008]
NA



```
sf$sect <- cut_number(sf$year, n = 5)

ggplot(sf, aes(year, fill = sect)) +
  geom_histogram(binwidth = 1)
```



[1870,1913]
(1913,1929]
(1929,1948]
(1948,1984]
(1984,2008]
NA
lat
37.80
37.78
37.76
37.74
37.72
long
-122.50
-122.48
-122.46
-122.44
-122.42
-122.40
-122.38
nrow
4
16
36
64
100
144
196

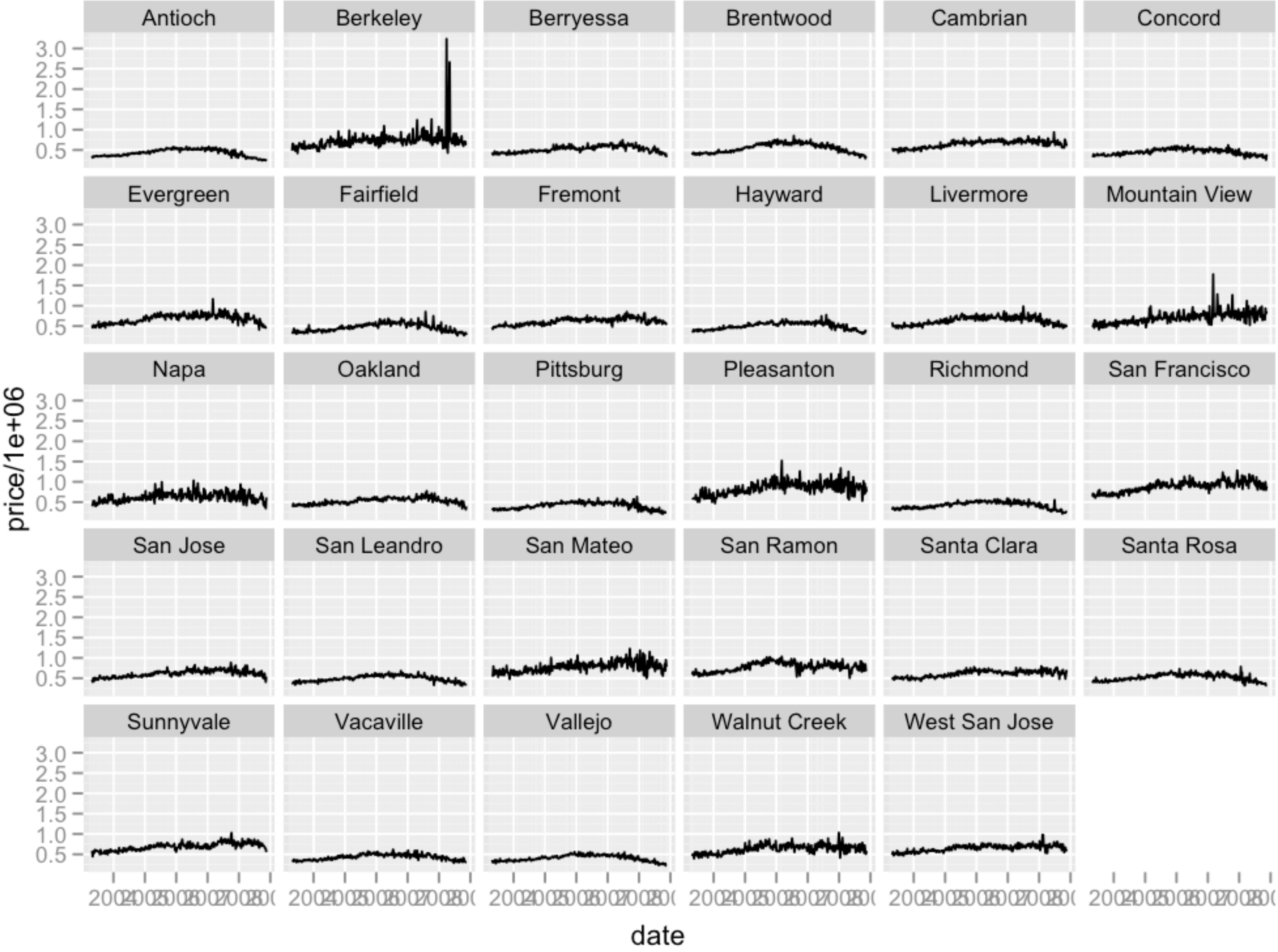
```
# Recount number at each location
sfsumsect <- ddply(sf, c("lat", "long" , "sect"),
  "nrow", .progress = "text")

# Display spatial distribution of each cut
ggplot(sfsumsect, aes(long, lat, size = nrow)) +
  geom_point(alpha = 1/2) +
  scale_area(to = c(0.5, 3)) +
  facet_wrap(~ sect)
```

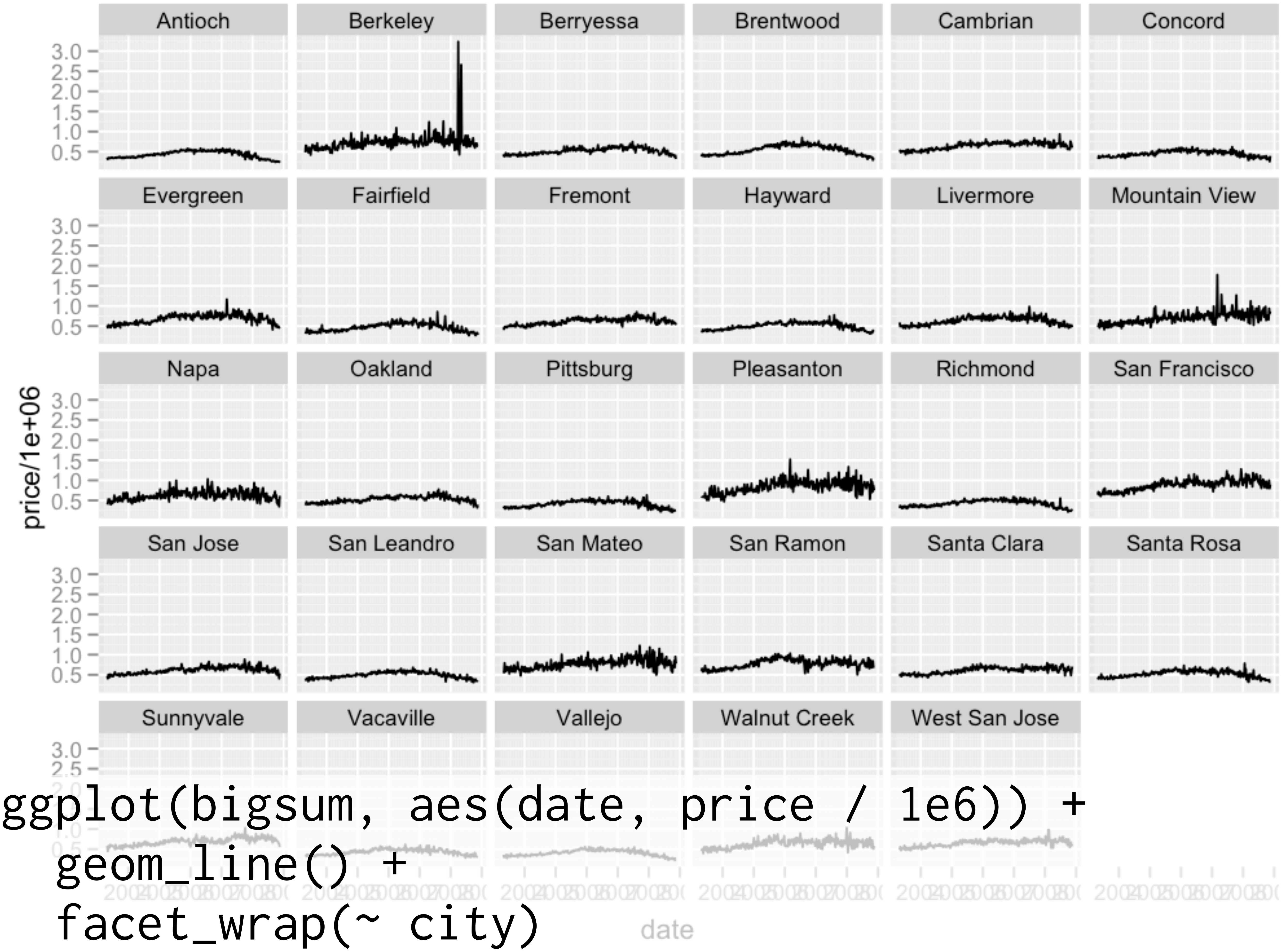
Time series

using plyr to manipulate data

Zoom out to the Bay area.
What’s going on in different cities?

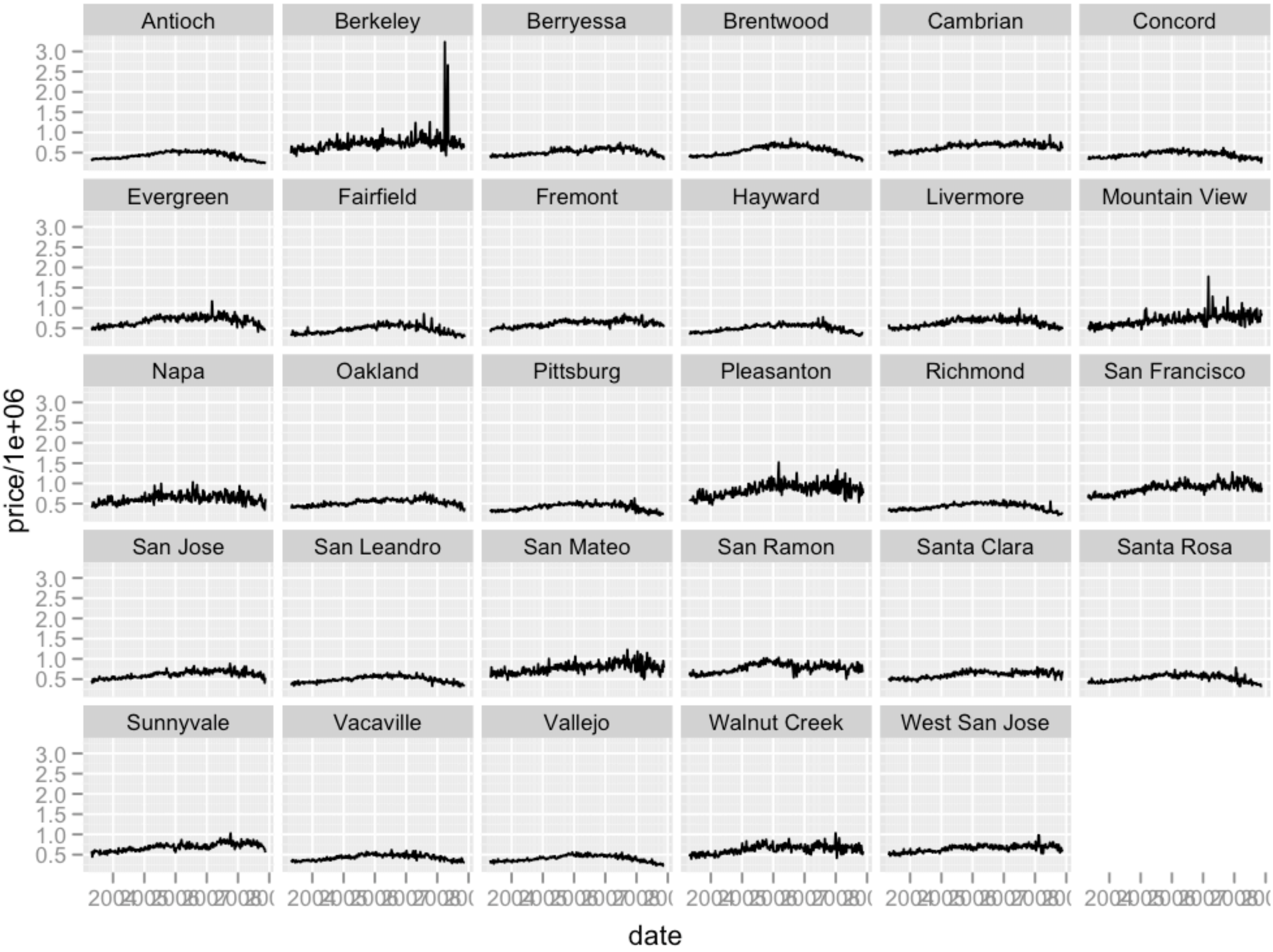
Focus on average sale price, **summarising** patterns, then **investigating** possible explanatory variables.

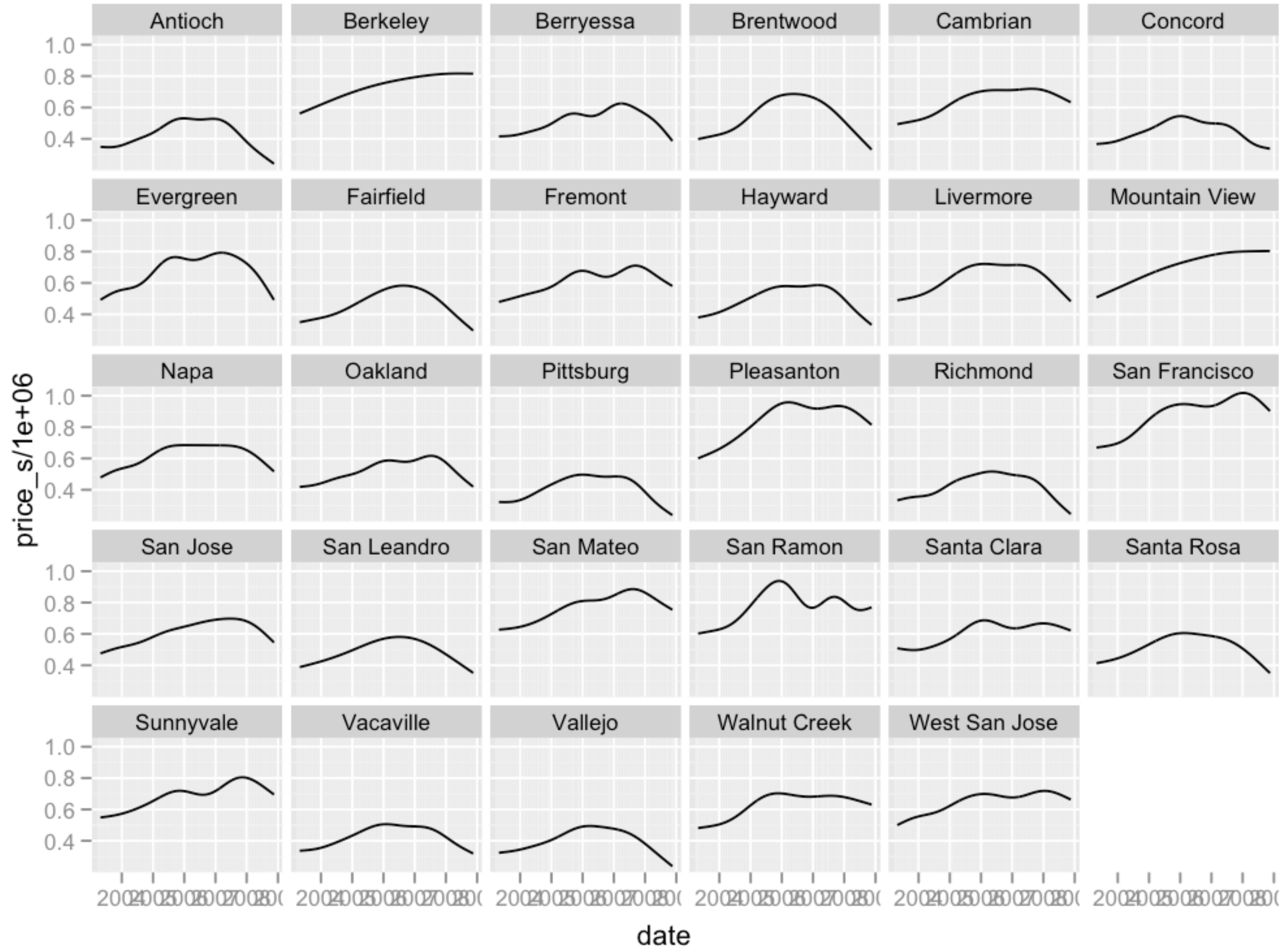

Antioch
Berkeley
Berryessa
Brentwood
Cambrian
Concord
Evergreen
Fairfield
Fremont
Hayward
Livermore
Mountain View
Napa
Oakland
Pittsburg
Pleasanton
Richmond
San Francisco
San Jose
San Leandro
San Mateo
San Ramon
Santa Clara
Santa Rosa
Sunnyvale
Vacaville
Vallejo
Walnut Creek
West San Jose
3.0
2.5
2.0
1.5
1.0
0.5
price/1e+06
2004 2005 2006 2007 2008
date



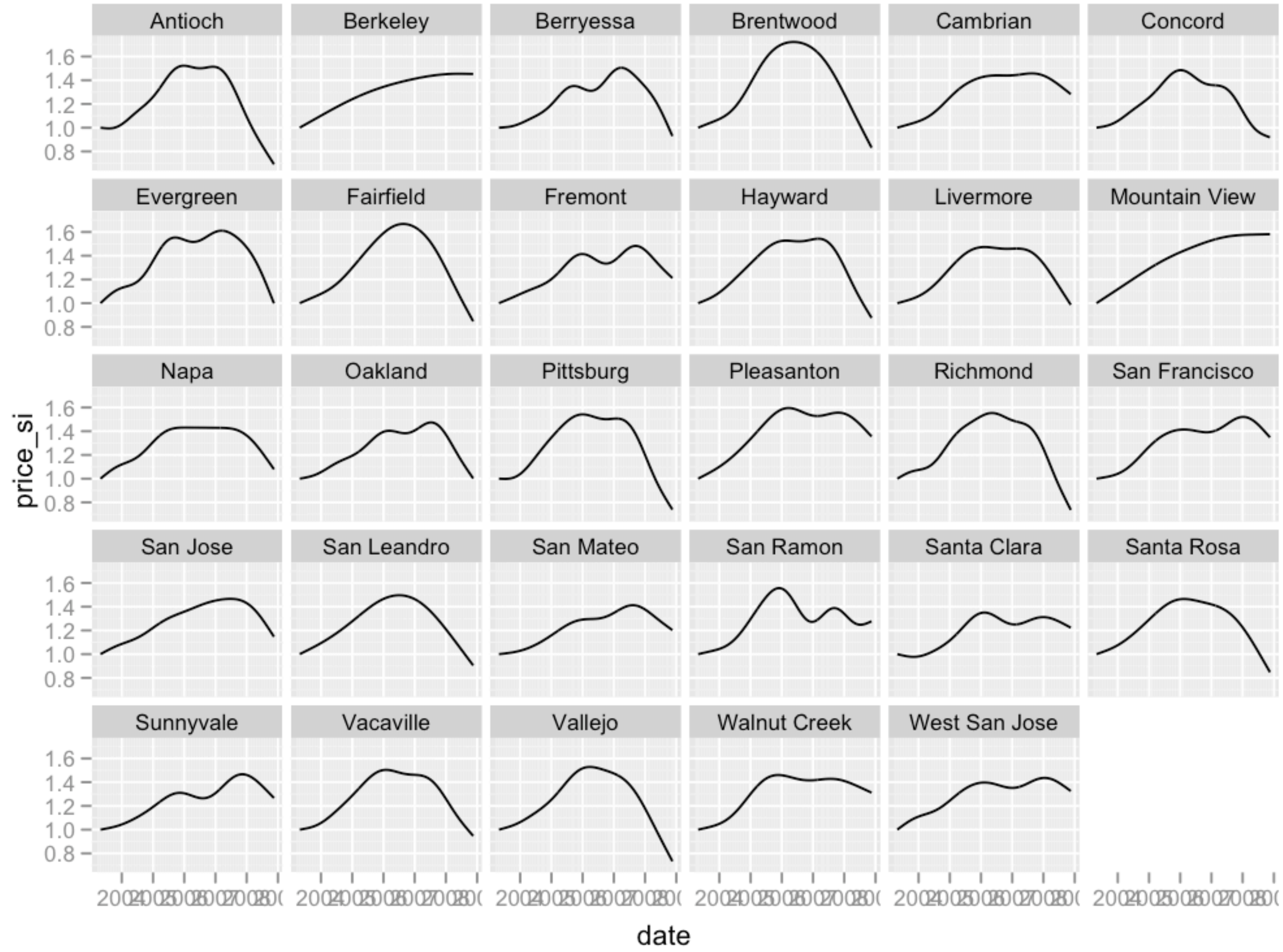
```
ggplot(bigsum, aes(date, price / 1e6)) +
  geom_line() +
  facet_wrap(~ city)
```

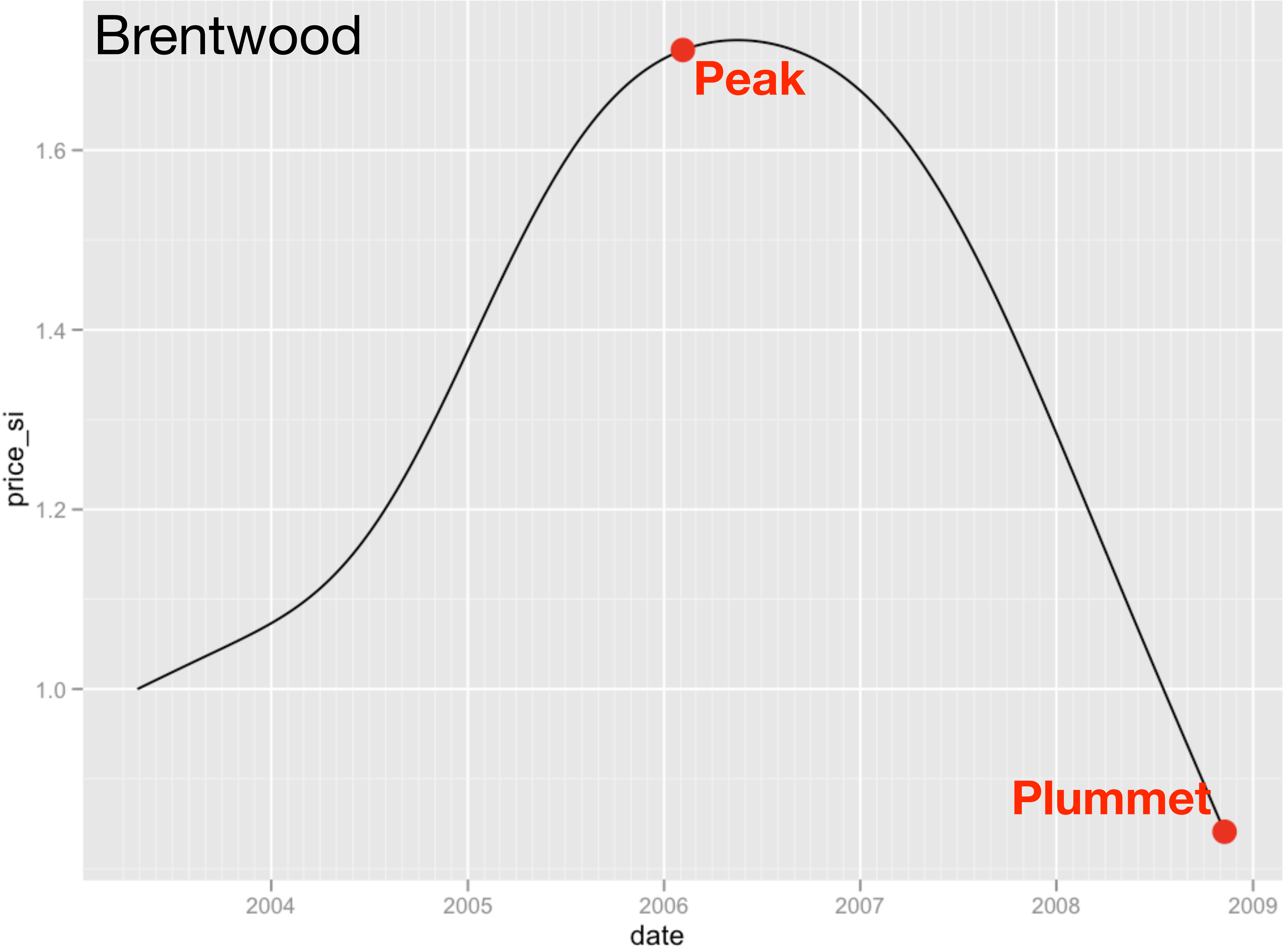
```
library(mgcv)
smooth <- function(y, x) {
  as.numeric(predict(gam(y ~ s(x)))
}

bigsum <- ddply(bigsum, ”city”, transform,
  price_s = smooth(price, as.numeric(date)))
```


Antioch
Berkeley
Berryessa
Brentwood
Cambrian
Concord
Evergreen
Fairfield
Fremont
Hayward
Livermore
Mountain View
Napa
Oakland
Pittsburg
Pleasanton
Richmond
San Francisco
San Jose
San Leandro
San Mateo
San Ramon
Santa Clara
Santa Rosa
Sunnyvale
Vacaville
Vallejo
Walnut Creek
West San Jose
3.0
2.5
2.0
1.5
1.0
0.5
price/1e+06
2004
2005
2006
2007
2008
date


Antioch
Berkeley
Berryessa
Brentwood
Cambrian
Concord
Evergreen
Fairfield
Fremont
Hayward
Livermore
Mountain View
Napa
Oakland
Pittsburg
Pleasanton
Richmond
San Francisco
San Jose
San Leandro
San Mateo
San Ramon
Santa Clara
Santa Rosa
Sunnyvale
Vacaville
Vallejo
Walnut Creek
West San Jose
1.0
0.8
0.6
0.4
2004
2005
2006
2007
2008
price_s/1e+06
date

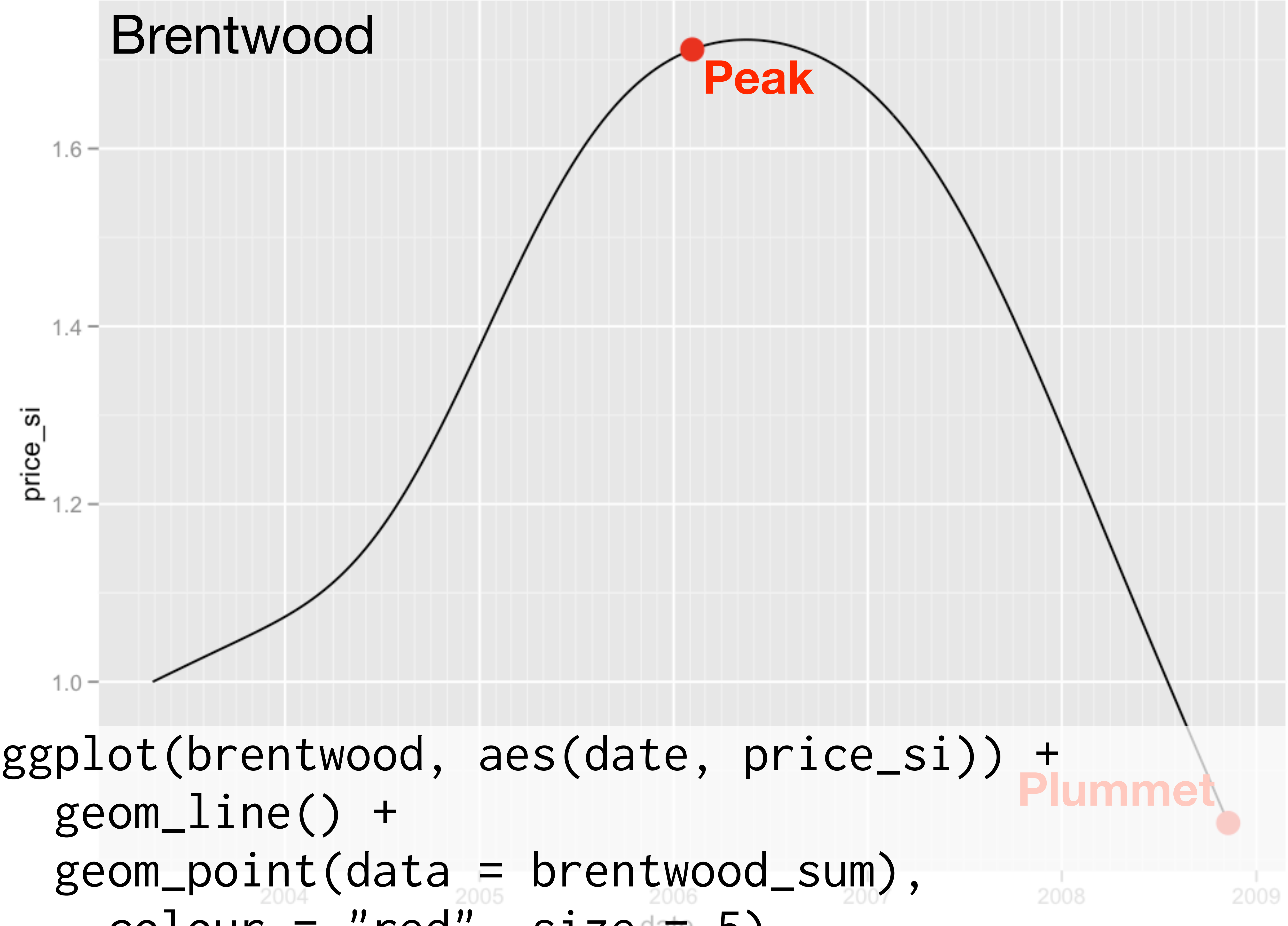
```
index <- function(y, x) {
  y / y[order(x)[1]]
}

bigsum <- ddply(bigsum, ”city”, transform,
  price_si = index(price_s, date))
```


Antioch
Berkeley
Berryessa
Brentwood
Cambrian
Concord
Evergreen
Fairfield
Fremont
Hayward
Livermore
Mountain View
Napa
Oakland
Pittsburg
Pleasanton
Richmond
San Francisco
San Jose
San Leandro
San Mateo
San Ramon
Santa Clara
Santa Rosa
Sunnyvale
Vacaville
Vallejo
Walnut Creek
West San Jose
price_si
1.6
1.4
1.2
1.0
0.8
2004
2005
2006
2007
2008
date

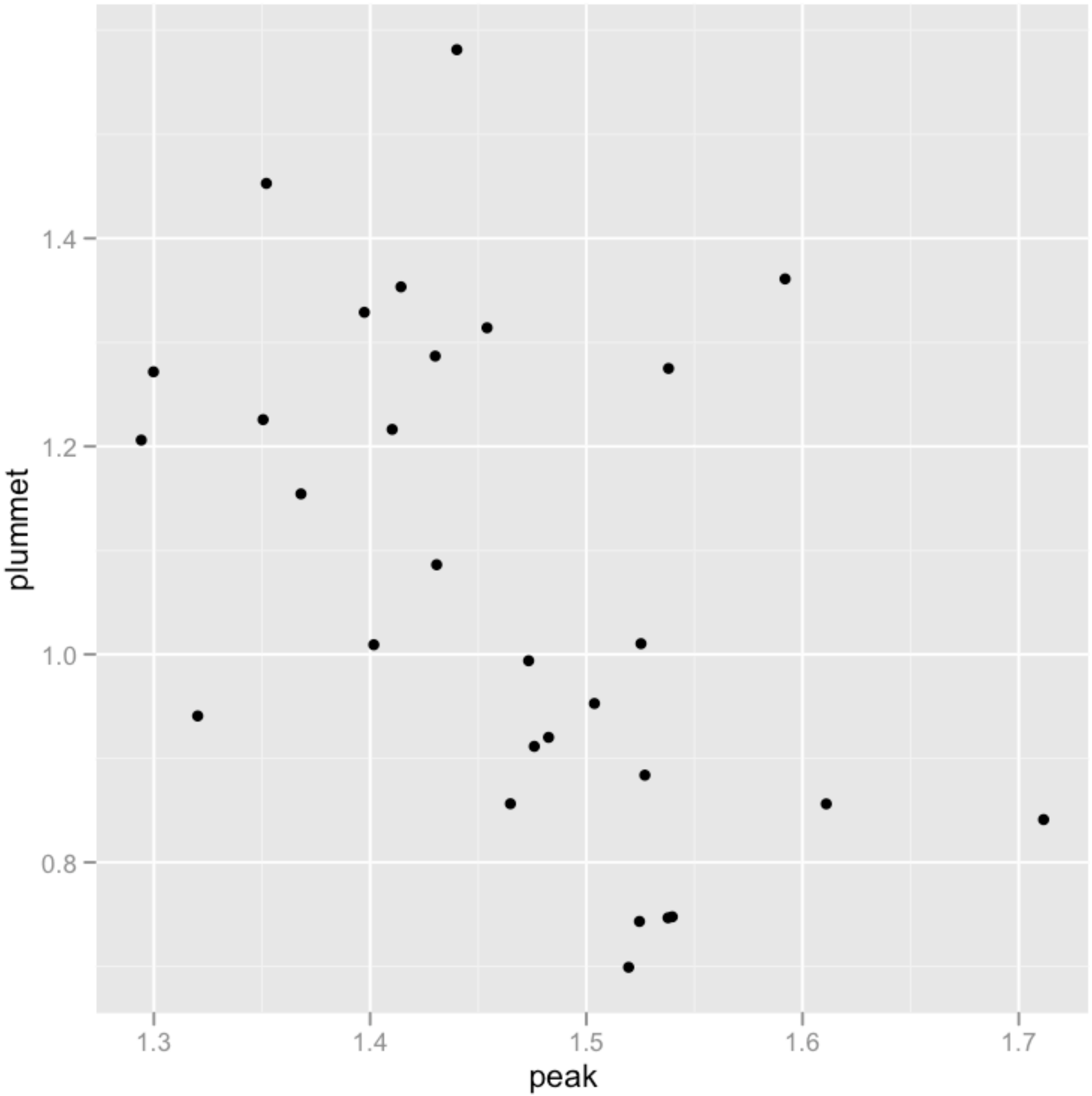


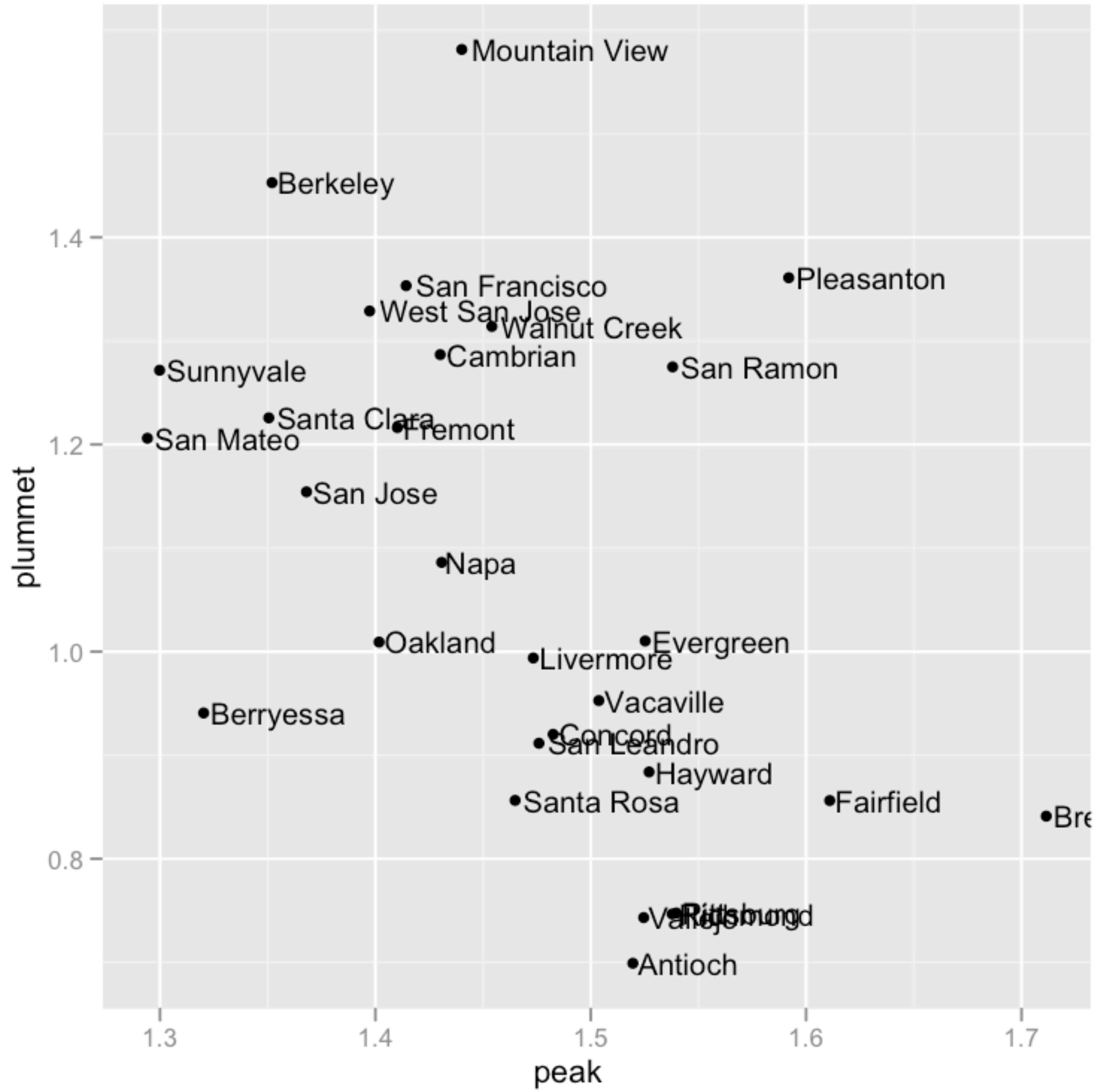
Peak

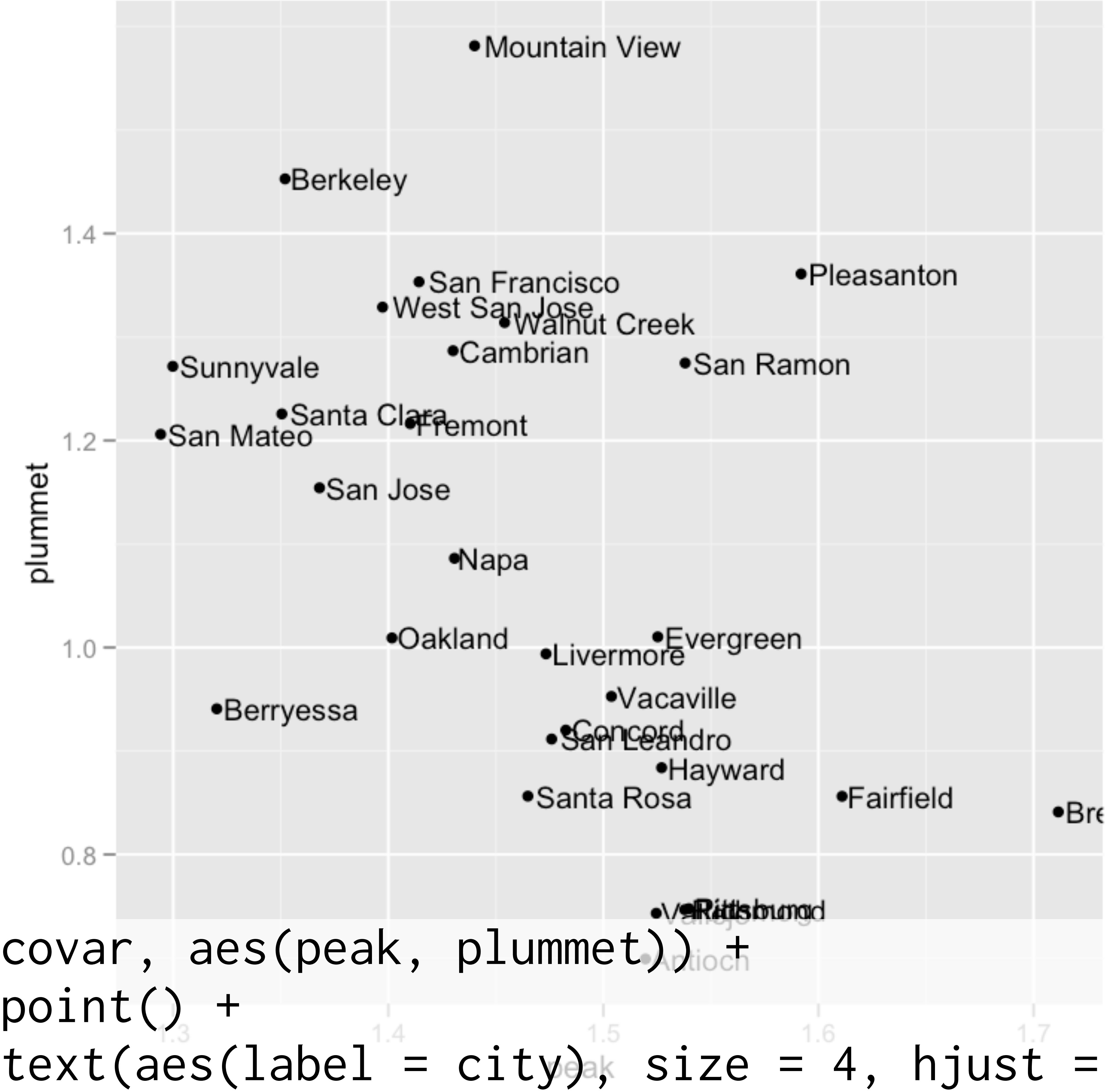
Plummet



```
ggplot(brentwood, aes(date, price_si)) +
  geom_line() +
  geom_point(data = brentwood_sum),
    colour = "red", size = 5)
```

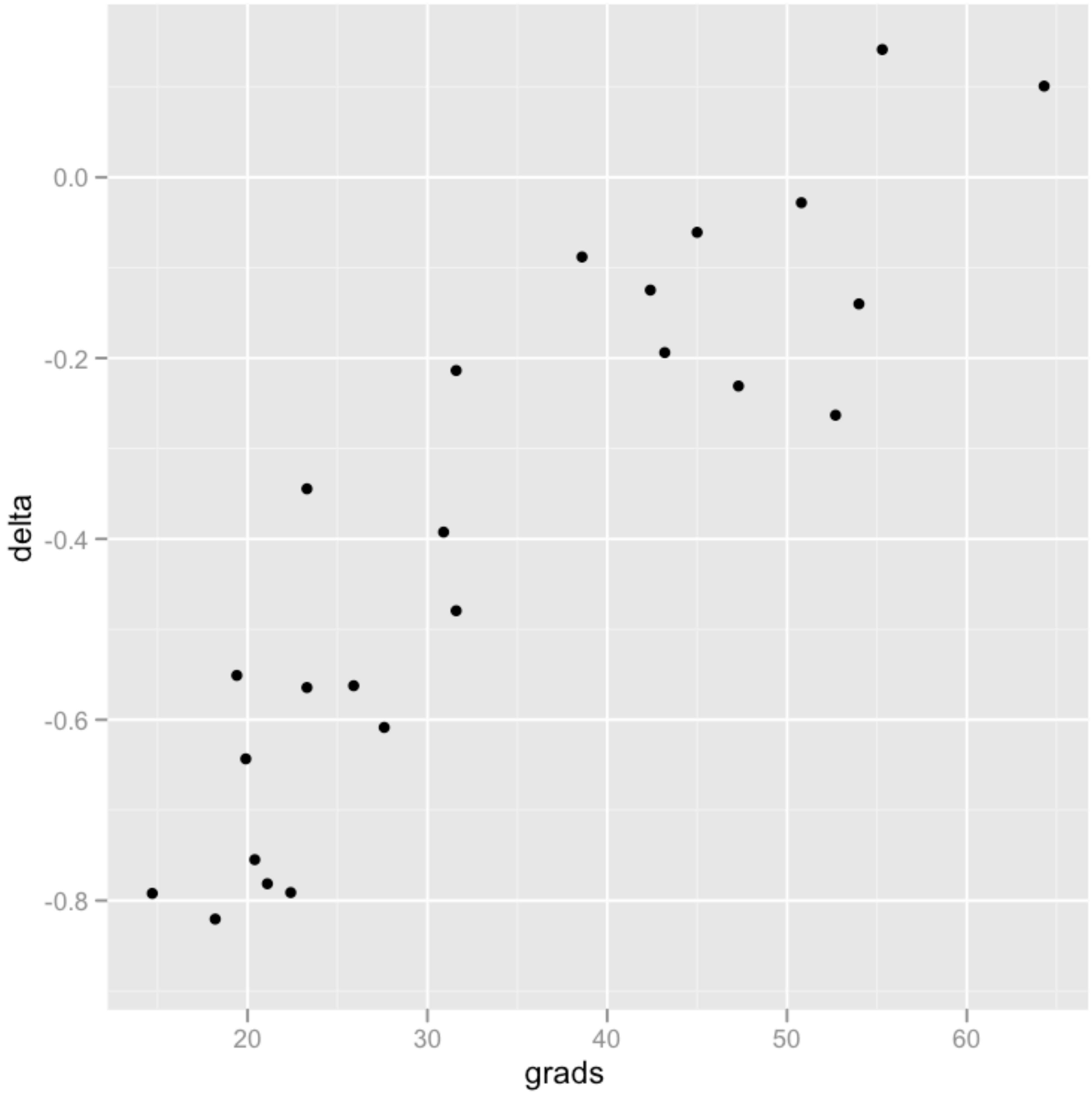

1.4
1.2
1.0
0.8
plummet
1.3
1.4
1.5
1.6
1.7
peak

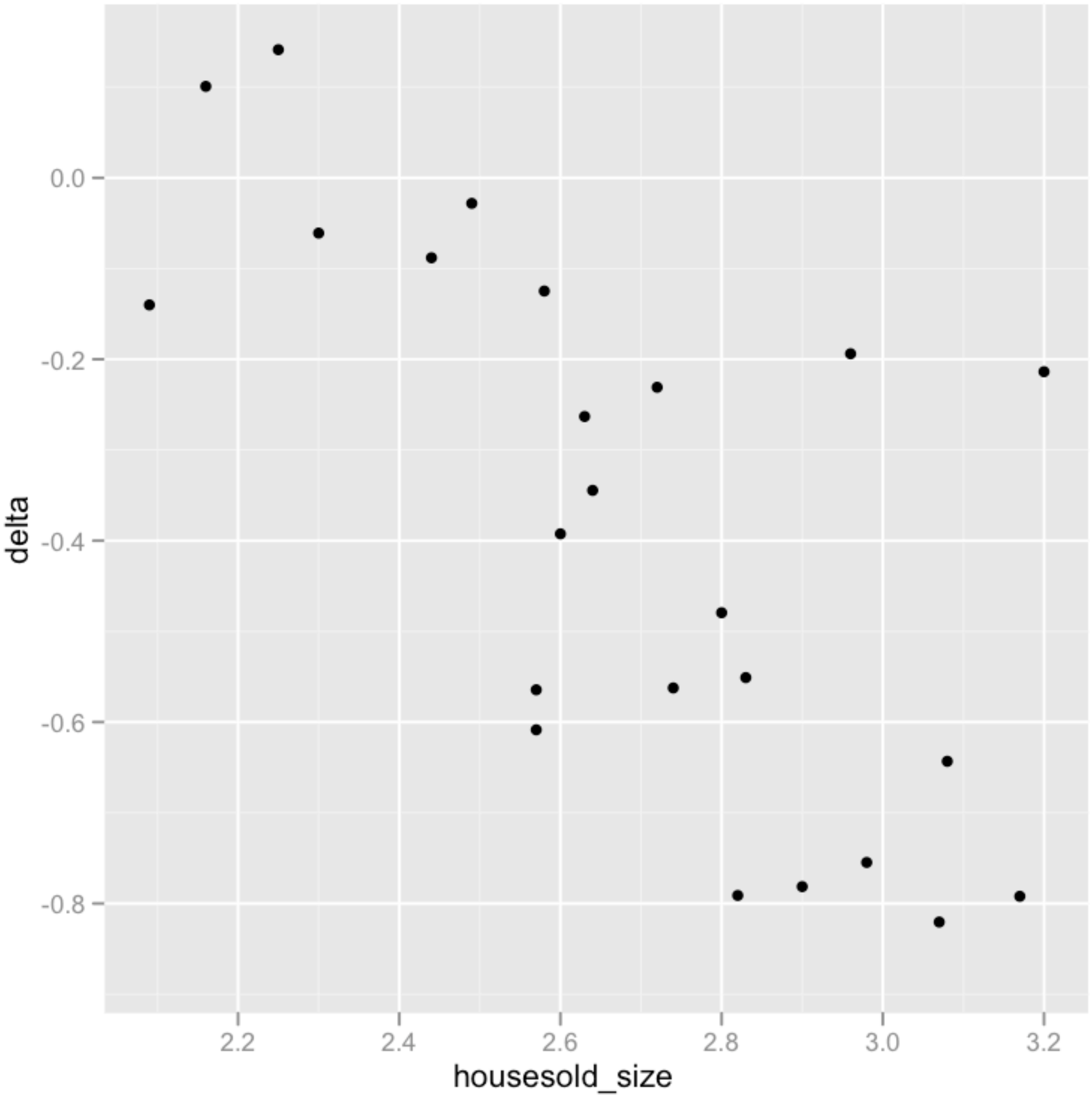

Mountain View
Berkeley
1.4
San Francisco
Pleasanton
West San Jose
Walnut Creek
Cambrian
San Ramon
Sunnyvale
Santa Clara
Fremont
San Mateo
1.2
San Jose
plummet
Napa
Oakland
Evergreen
1.0
Livermore
Vacaville
Berryessa
Concord
San Leandro
Hayward
Santa Rosa
Fairfield
Bre
0.8
Pittsburg
Richmond
Vallejo
Antioch
1.3
1.4
1.5
1.6
1.7
peak



```
ggplot(covar, aes(peak, plummet)) +
  geom_point() +
  geom_text(aes(label = city), size = 4, hjust = -0.05)
```

```
covar$delta <- with(covar, plummet - peak)

census <- read.csv(”census-city.csv”)
covar <- join(covar, census, by = ”city”)
```

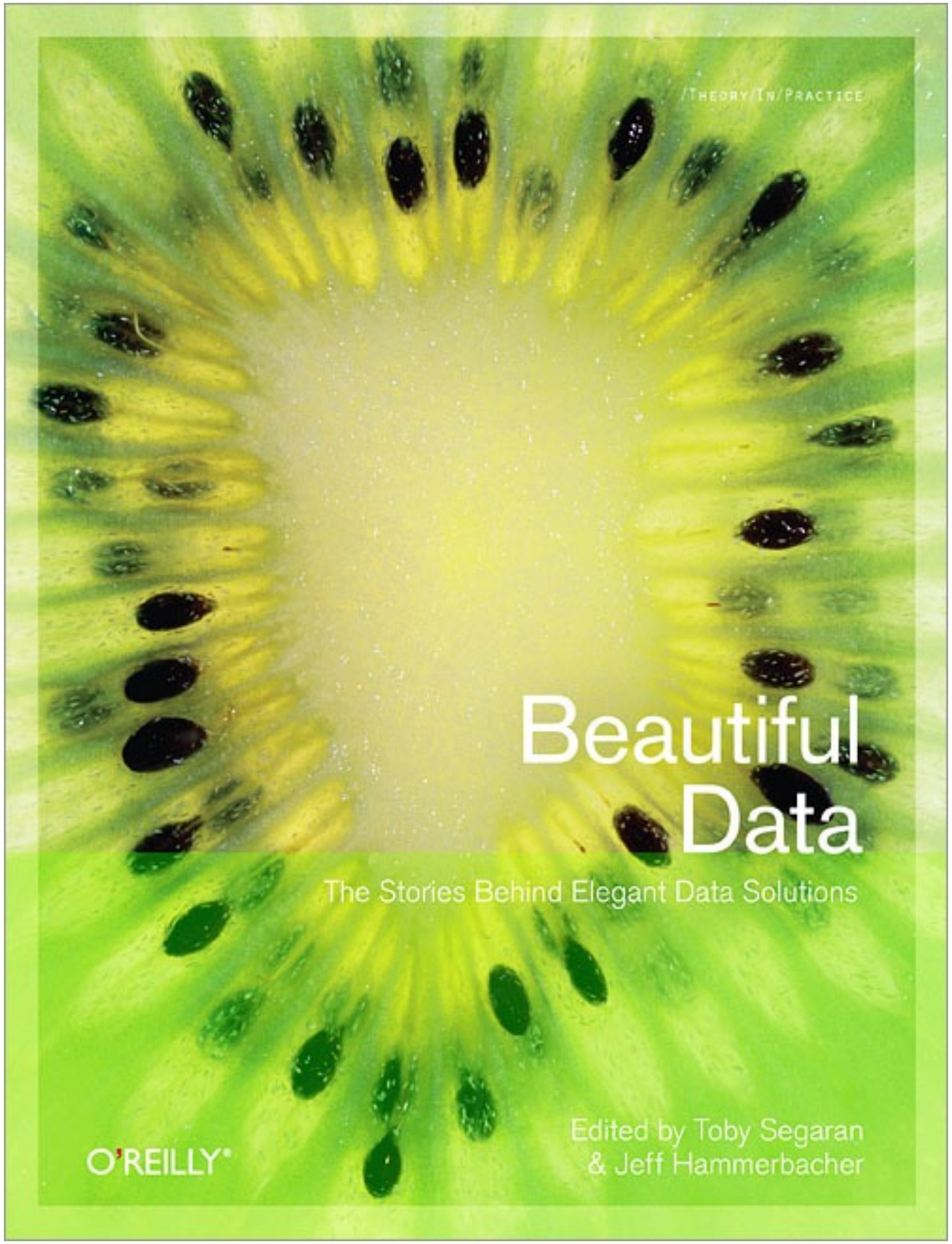

0.0
-0.2
-0.4
-0.6
-0.8
delta
20
30
40
50
60
grads


0.0
-0.2
-0.4
-0.6
-0.8
delta
2.2
2.4
2.6
2.8
3.0
3.2
housesold_size

Learn more

ggplot2
http://had.co.nz/ggplot2
http://groups.google.com/group/ggplot2

plyr
http://had.co.nz/plyr
http://groups.google.com/group/manipulatr



Chapter 18

All code and data openly licensed: http://github.com/hadley/sfhousing

Beautiful data is **reproducible**!

Thank you!