Building visualisations

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair

Department of Statistics / Rice University

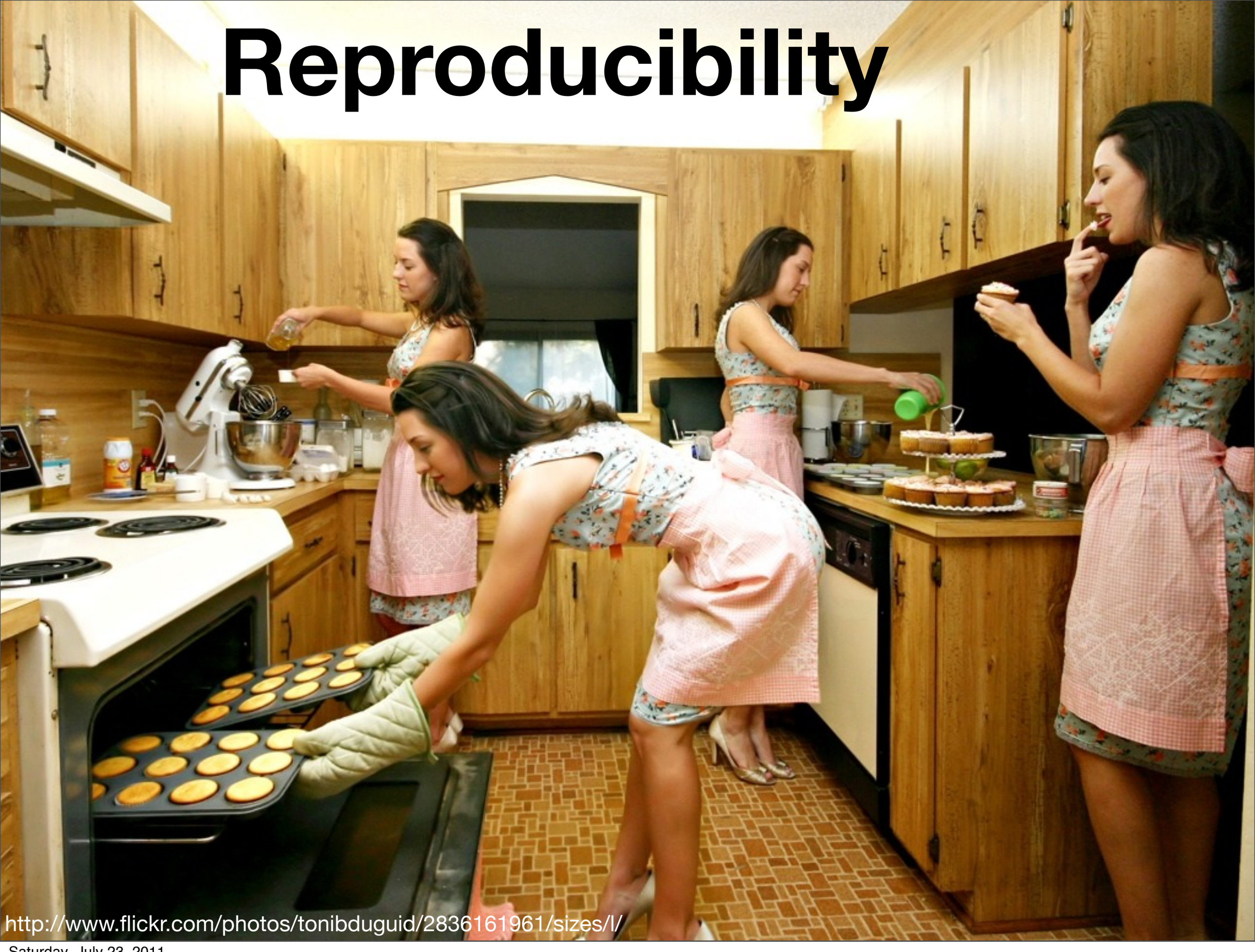
March 2010

Use R and ggplot2

1. Why use a programming language?
2. Why use R?
3. Why use ggplot2?
4. Two case studies

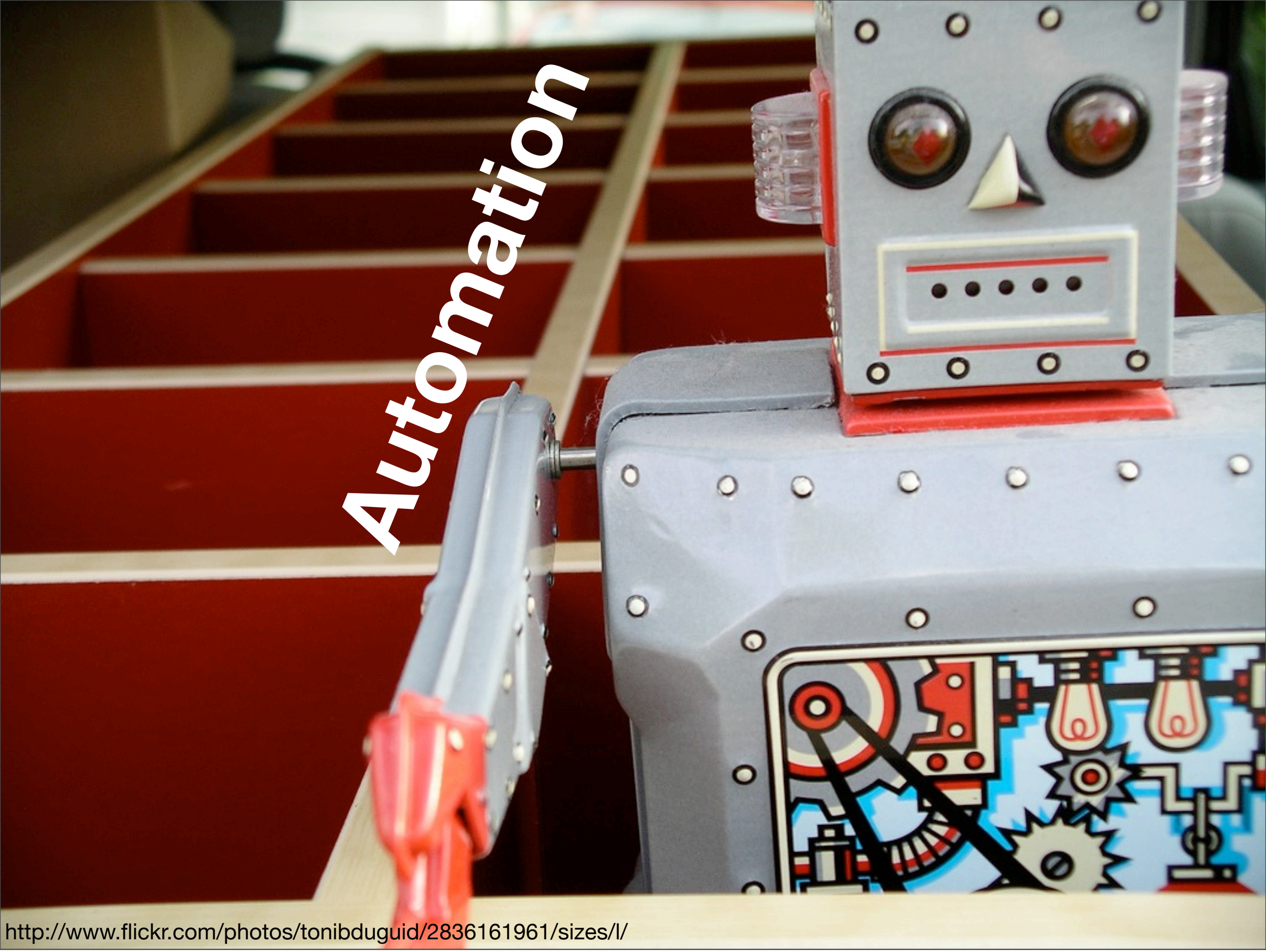
Why use a programming language?

Reproducibility



http://www.flickr.com/photos/tonibduguid/2836161961/sizes/l/

Automation



http://www.flickr.com/photos/tonibduguid/2836161961/sizes/l/

Just text

```
# Load data and create smaller subsets
tb <- read.csv("tb.csv")
tb2008 <- subset(tb, year == 2008)

# Choropleth map ------------------------------------------------------------
borders <- read.csv("world-borders.csv")
choro <- merge(tb2008, borders, by = "iso2")
choro <- choro[order(choro$order), ]

qplot(long, lat, data = choro, fill = cut_number(rate, 5), geom = "polygon", group =
group) + scale_fill_brewer("Rate", pal = "Blues")

# Bubble maps ---------------------------------------------------------------
centres <- read.csv("world-centres.csv")
bubble <- merge(centres, tb2008, by = "iso2")

world_coord <- coord_map(xlim = c(-180, 180), ylim = c(-50, 70))

# This is basically what a choropleth is showing us
qplot(long, lat, data = bubble, size = area, colour = rate) +
  scale_area(to = c(2, 25), legend = FALSE) +
  world_coord

# More traditional options
qplot(long, lat, data = bubble, size = rate) + world_coord
qplot(long, lat, data = bubble, size = log10(pop), colour = rate) +
  world_coord

# Even better if we add world boundaries
ggplot(bubble, aes(long, lat)) +
  geom_polygon(data = borders, aes(group = group)) +
  geom_point(aes(colour = rate)) +
  coord_map()
ggsave("world-4.png", width = 8, height = 6, dpi = 128)

# Works better if we tweak aesthetics
ggplot(bubble, aes(long, lat)) +
  geom_polygon(data = borders, aes(group = group), colour = "grey70",
```

Communication

http://www.flickr.com/photos/altemark/337248947/sizes/l/


Learning curve

Why R?

```
SEXP applyClosure(SEXP call, SEXP op, SEXP arglist, SEXP rho, SEXP suppliedenv)
{
    SEXP body, formals, actuals, savedrho;
    volatile  SEXP newrho;
    SEXP f, a, tmp;
    RCNTXT cntxt;

    /* formals = list of formal parameters */
    /* actuals = values to be bound to formals */
    /* arglist = the tagged list of arguments */

    formals = FORMALS(op);
    body = BODY(op);
    savedrho = CLOENV(op);

    /*  Set up a context with the call in it so error has access to it */

    begincontext(&cntxt, CTXT_RETURN, call, savedrho, rho, arglist, op);

    /*  Build a list which matches the actual (unevaluated) arguments
        to the formal paramters.  Build a new environment which
        contains the matched pairs.  Ideally this environment sould be
        hashed.  */

    PROTECT(actuals = matchArgs(formals, arglist, call));
    PROTECT(newrho = NewEnvironment(formals, actuals, savedrho));

    /*  Use the default code for unbound formals.  FIXME: It looks like
        this code should preceed the building of the environment so that
        this will also go into the hash table.  */

    /* This piece of code is destructively modifying the actuals list,
       which is now also the list of bindings in the frame of newrho.
       This is one place where internal structure of environment
       bindings leaks out of envir.c.  It should be rewritten
       eventually so as not to break encapsulation of the internal
       environment layout.  We can live with it for now since it only
       happens immediately after the environment creation.  LT */
```

Open source

Freedom

http://www.flickr.com/photos/amagill/3367543296/sizes/l/

Community

http://www.flickr.com/photos/ianlayzellphotographs/3977042044


Prickly
http://www.flickr.com/photos/meantux/367751359

Runs anywhere

http://www.flickr.com/photos/jonlucas/204213732

Build it yourself

http://www.flickr.com/photos/wwworks/2473052504

http://www.flickr.com/photos/54945394@N00/2987214939

Connectivity

http://www.flickr.com/photos/billy64/2226377312

Why ggplot2?

"If any number of magnitudes are each the same multiple of the same number of other magnitudes, then the sum is that multiple of the sum."

Euclid, ~300 BC

“If any number of magnitudes are each the same multiple of the same number of other magnitudes, then the sum is that multiple of the sum.”

Euclid, ~300 BC

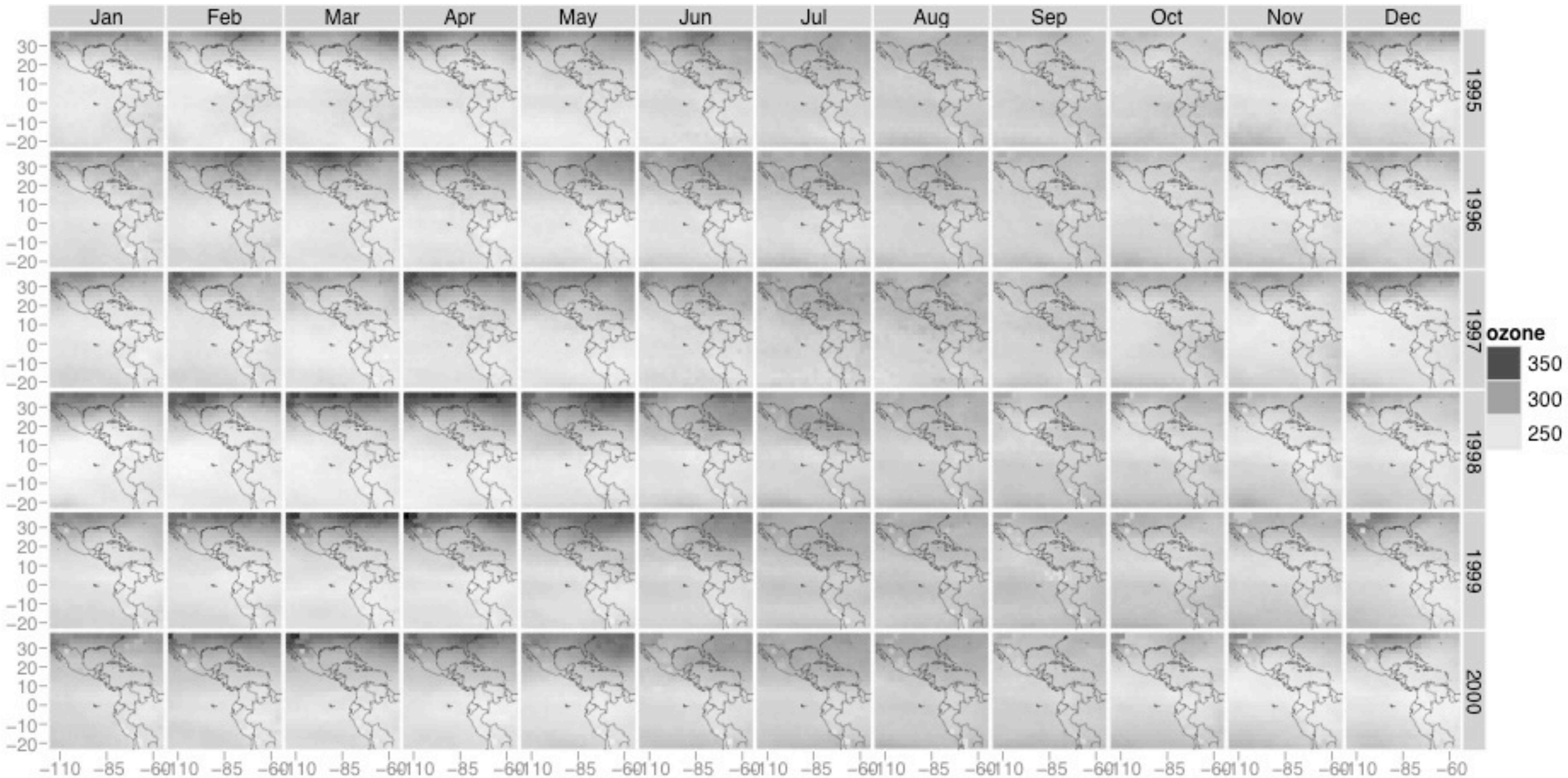
$$ab + ac = a(b + c)$$

The grammar of graphics

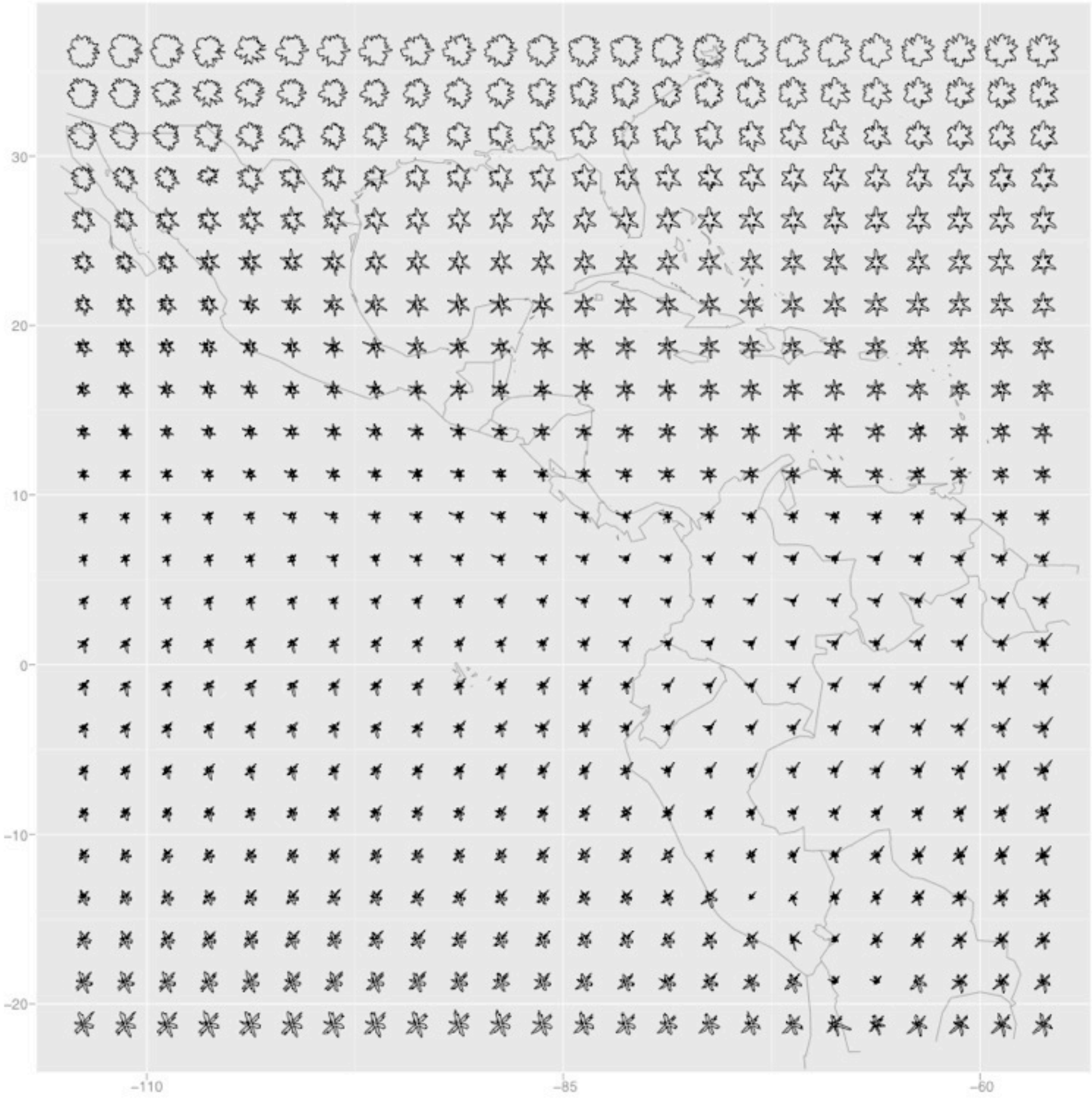
Like English grammar, defines the components that make up a statistical graphic and specifies how they can be arranged.

An abstraction which makes thinking, reasoning and communicating graphics easier

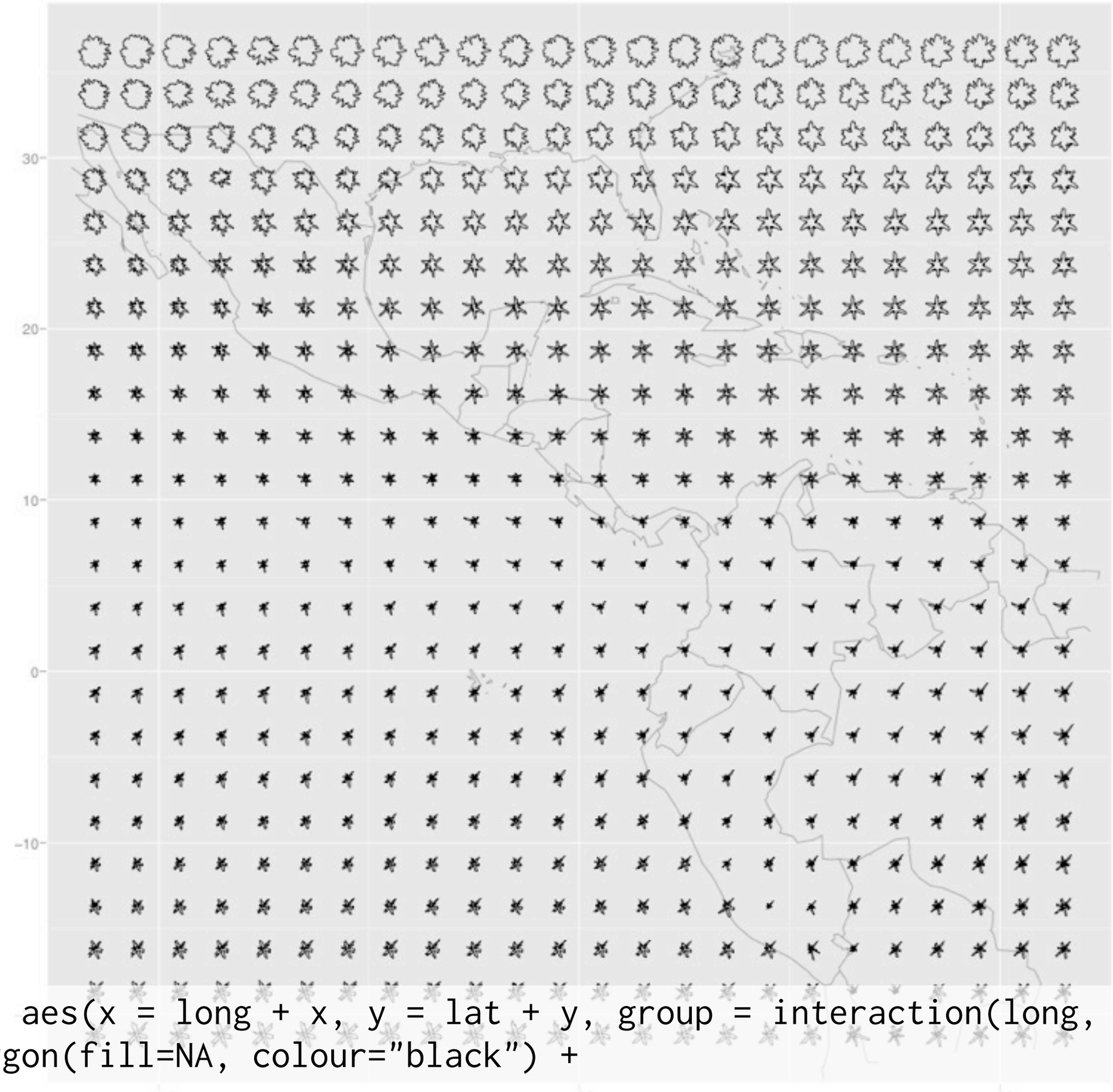
Developed by Leland Wilkinson, particularly in “The Grammar of Graphics” 1999/2005




```
ggplot(expo, aes(long, lat, fill = ozone)),
  geom_tile() +
  facet_grid(year ~ month) +
  scale_fill_gradient(low="white", high="black") +
  map
```



Saturday, July 23, 2011




```
ggplot(df, aes(x = long + x, y = lat + y, group = interaction(long, lat))) +
 geom_polygon(fill=NA, colour="black") +
 map
```

Practically

Looks good, and takes care of fiddly details like legends.

Allows you to create new plots as well as reuse old plots.

Makes doing the right thing easy, while keeping harder things possible.

Continuum of expertise.

Case study 1

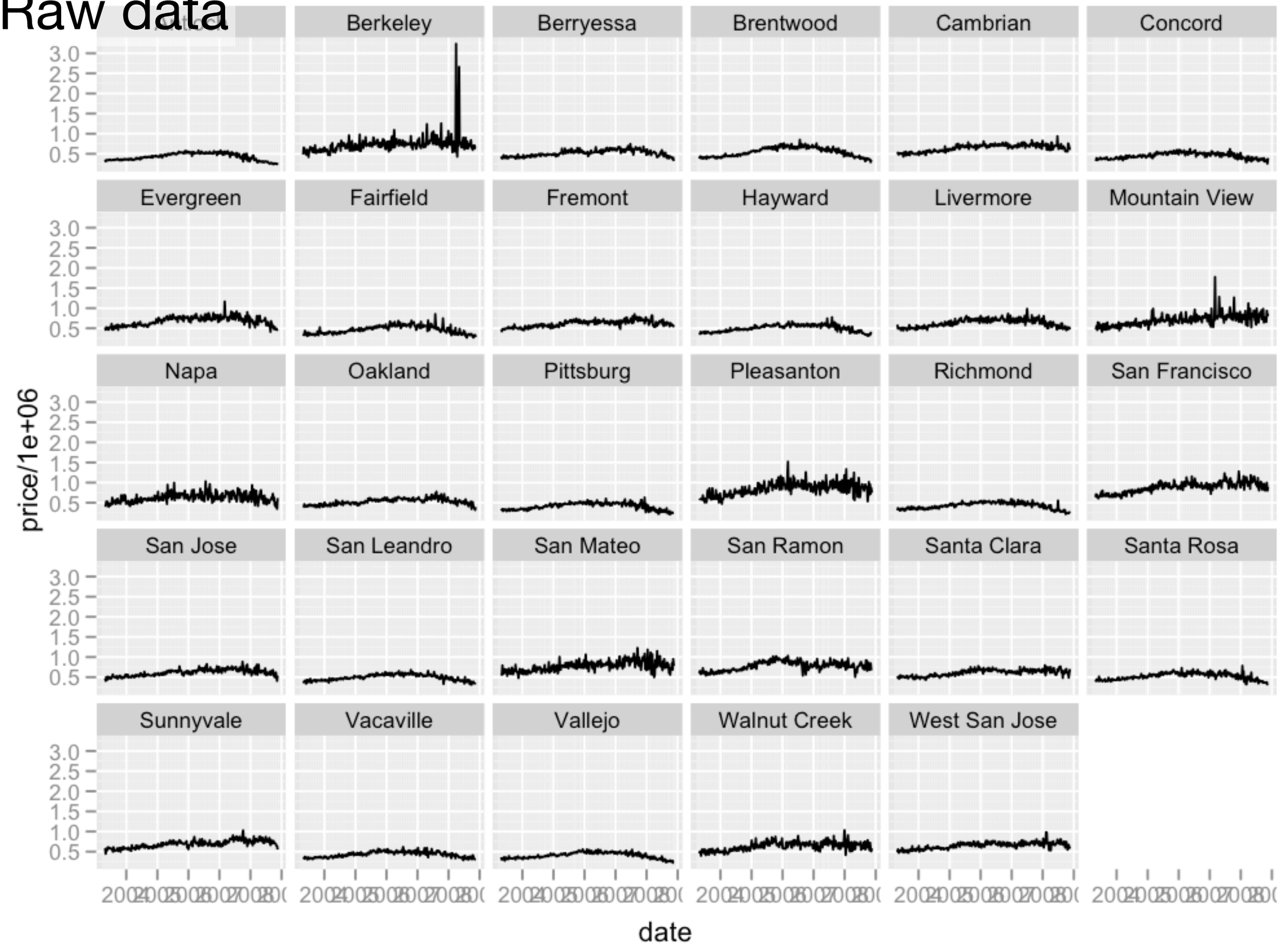
Data

About 250,000 house sales in the Bay Area (around San Francisco).

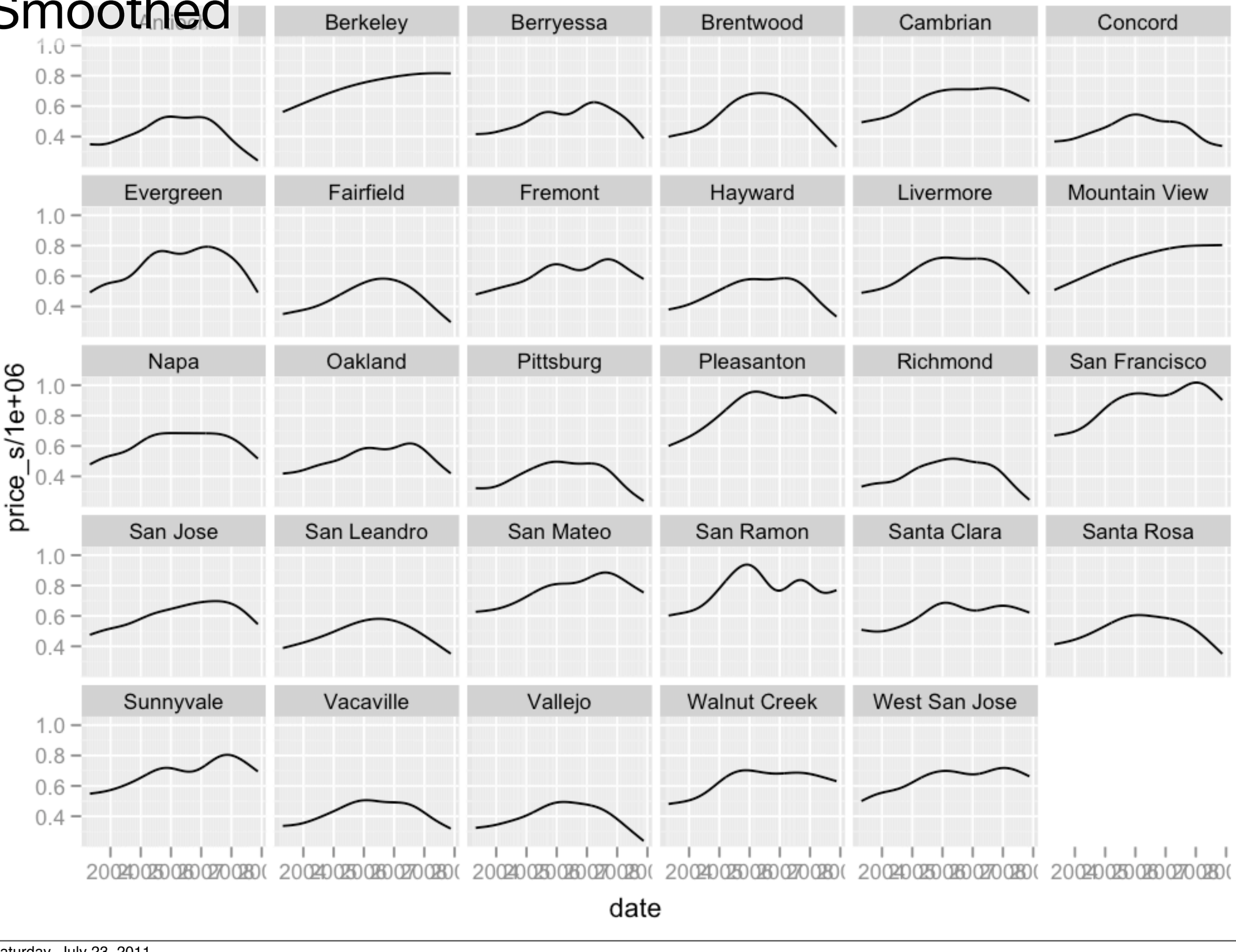
Addresses, date & sale price.

Illustrates combination of graphics and modelling that makes R so powerful.

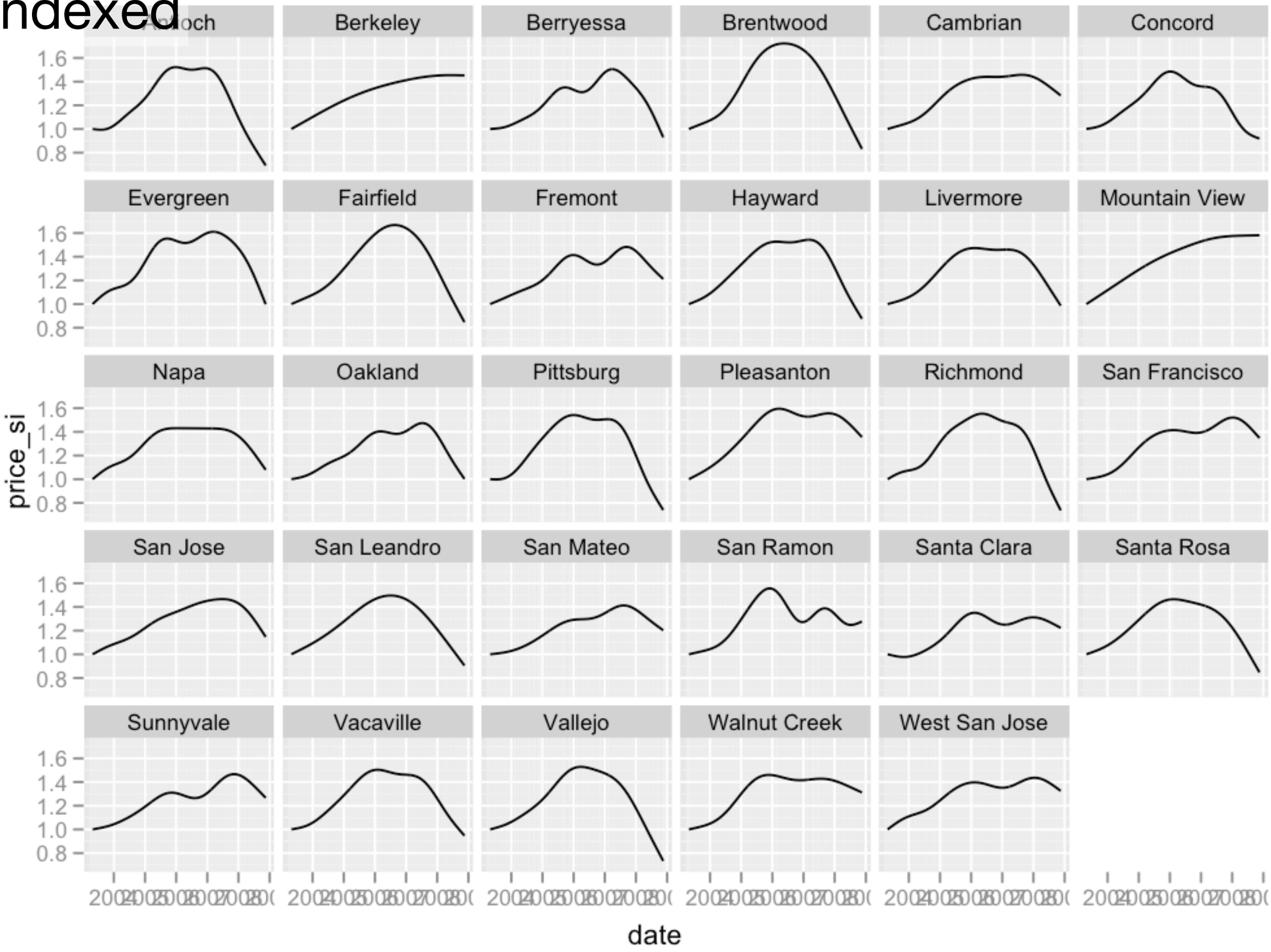
Raw data

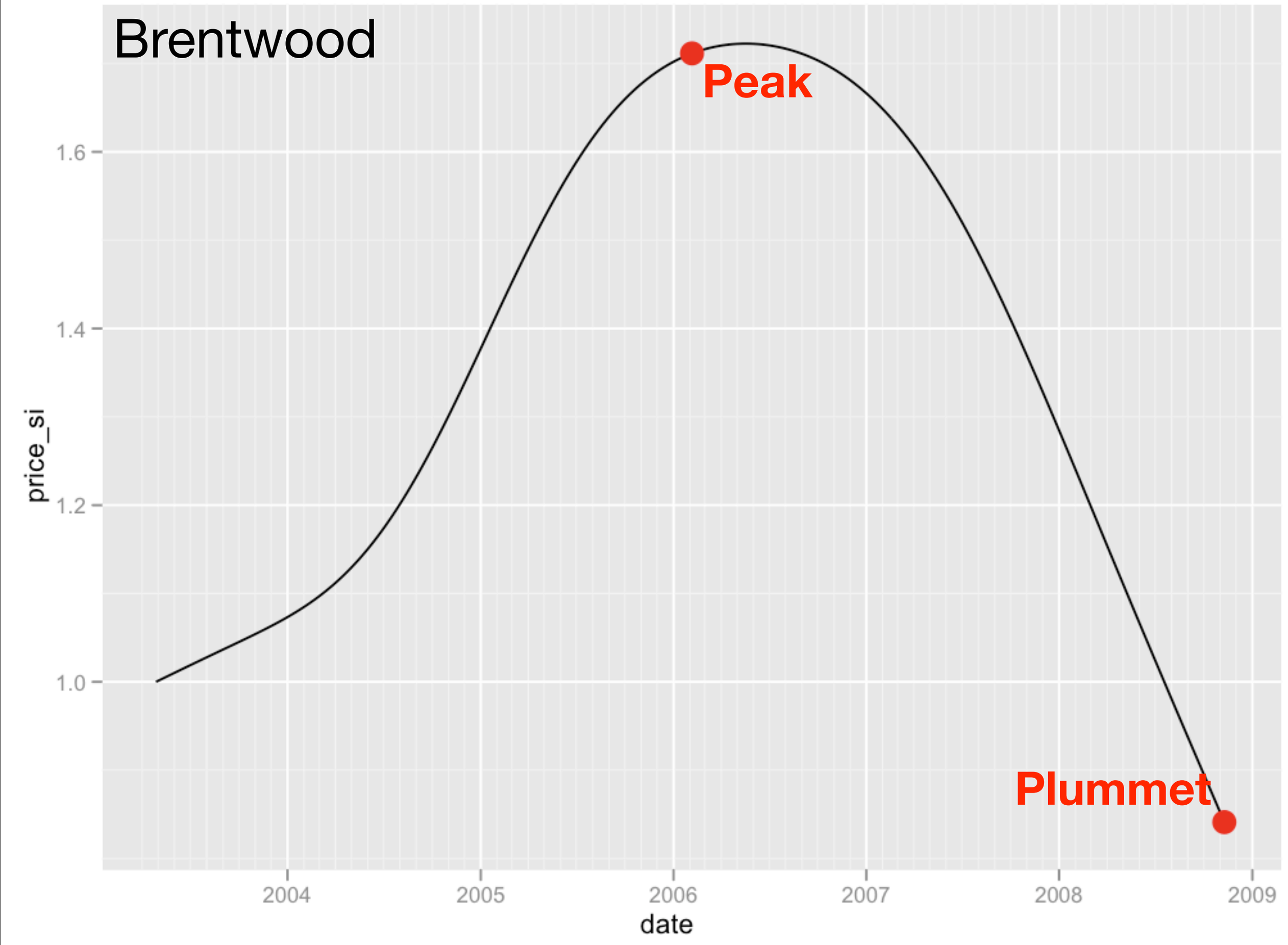

Antioch
Berkeley
Berryessa
Brentwood
Cambrian
Concord
Evergreen
Fairfield
Fremont
Hayward
Livermore
Mountain View
Napa
Oakland
Pittsburg
Pleasanton
Richmond
San Francisco
San Jose
San Leandro
San Mateo
San Ramon
Santa Clara
Santa Rosa
Sunnyvale
Vacaville
Vallejo
Walnut Creek
West San Jose
price/1e+06
3.0
2.5
2.0
1.5
1.0
0.5
2002
04
05
06
07
08
date

Smoothed

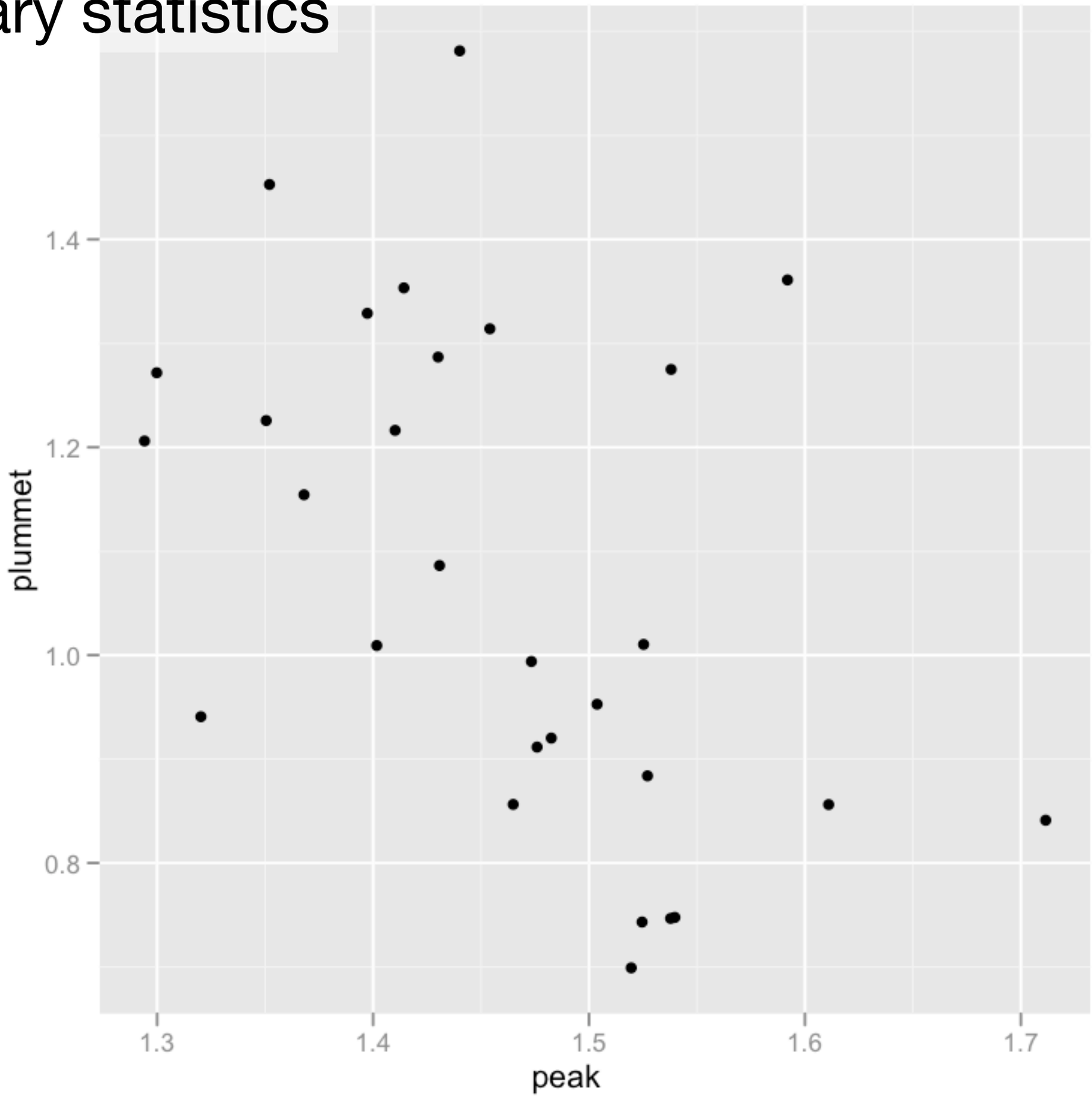

Antioch
Berkeley
Berryessa
Brentwood
Cambrian
Concord
Evergreen
Fairfield
Fremont
Hayward
Livermore
Mountain View
Napa
Oakland
Pittsburg
Pleasanton
Richmond
San Francisco
San Jose
San Leandro
San Mateo
San Ramon
Santa Clara
Santa Rosa
Sunnyvale
Vacaville
Vallejo
Walnut Creek
West San Jose
price_s/1e+06
1.0
0.8
0.6
0.4
2002
2004
2005
2006
2007
2008
date

Indexed

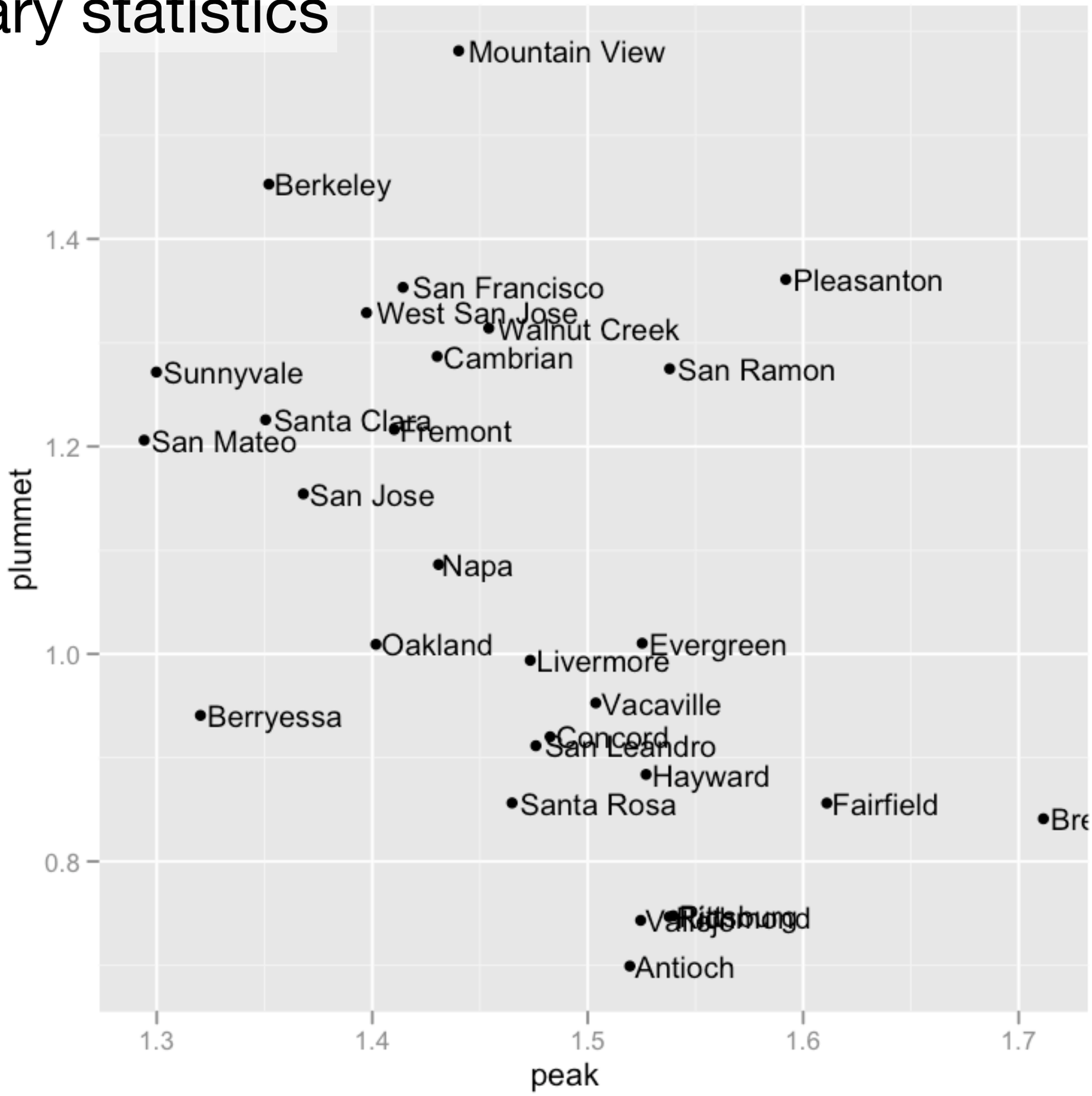

Antioch
Berkeley
Berryessa
Brentwood
Cambrian
Concord
Evergreen
Fairfield
Fremont
Hayward
Livermore
Mountain View
Napa
Oakland
Pittsburg
Pleasanton
Richmond
San Francisco
San Jose
San Leandro
San Mateo
San Ramon
Santa Clara
Santa Rosa
Sunnyvale
Vacaville
Vallejo
Walnut Creek
West San Jose
price_si
1.6
1.4
1.2
1.0
0.8
2002
04
05
06
07
08
date


Brentwood
Peak
Plummet
1.6
1.4
1.2
1.0
price_si
2004
2005
2006
2007
2008
2009
date

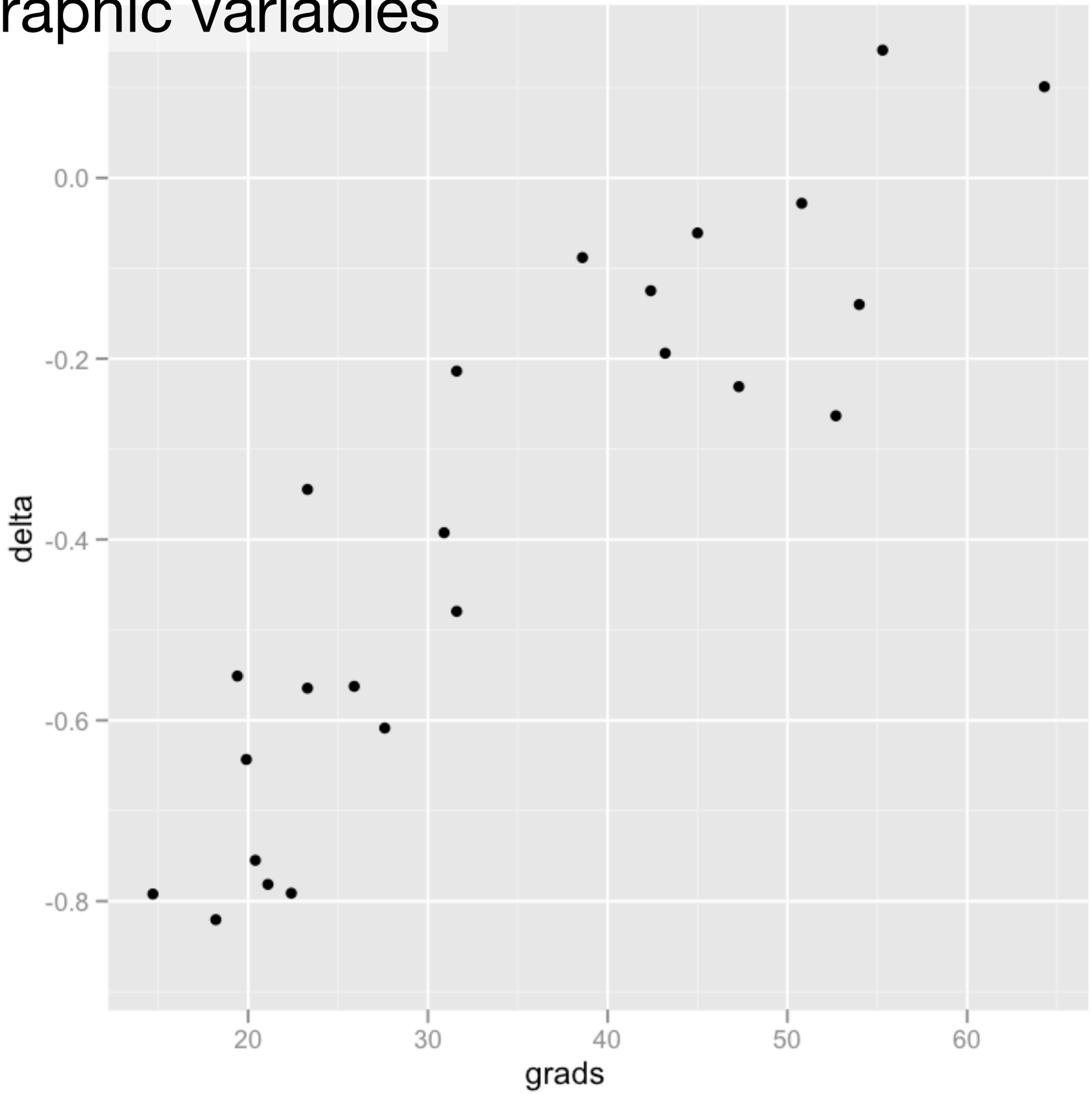
Summary statistics



Summary statistics



Demographic variables



Case study 2

Challenge

Compare per capita house sales for new vs. existing houses in 25 Texas cities.

Illustrates some common data manipulation challenges and how R can be used to overcome them.

Amarillo
Austin
Beaumont
Brownsville
College Station
Corpus Christi
Dallas
El Paso
Houston
Killeen
Laredo
Longview
Lubbock
McAllen
Midland
Odessa
San Angelo
San Antonio
Sherman
Tyler
Victoria
Waco
Wichita Falls

15
10
5

01
05
09

Houses (per 10000 people)

Date

Type
sold
built

4 data sources

Average **sale price** from multiple listing data provided by the Real Estate Center at A&M.

Average price of **new construction** from the census.

Population data, also from the census.

All connected by **metropolitan statistical area**.

Table 3u. New Privately Owned Housing Units Authorized
Unadjusted Units by Metropolitan Area

January 2000

	Total	1 Unit	2 Units	3 and 4 Units	5 Units or More	Num of Structures With 5 Units or More
Abilene* TX MSA	16	16	0	0	0	0
Albany* GA MSA	138	42	0	0	96	12
Albany-Schenectady-Troy* NY MSA	85	75	0	0	10	1
Albuquerque* NM MSA	371	337	0	4	30	2
Alexandria* LA MSA	29	29	0	0	0	0
Allentown-Bethlehem-Easton* PA MSA	98	70	0	4	24	2
Altoona* PA MSA	4	4	0	0	0	0

Table 3u. New Privately Owned Housing Units Authorized
Unadjusted Units by Metropolitan Area

January 2008

Monthly Coverage	Total	1 Unit	2 Units	3 & 4 Units	5 Units or more	Num of Structures With 5 Units or more	Percent
Abilene, TX	14	10	4	0	0	0	91
Akron, OH	93	46	0	3	44	7	69
Albany, GA	24	22	2	0	0	0	84
Albany-Schenectady-Troy, NY	39	39	0	0	0	0	59
Albuquerque, NM	204	163	0	0	41	2	100
Alexandria, LA	41	41	0	0	0	0	97
Allentown-Bethlehem-Easton, PA-NJ	118	113	0	0	5	1	100
Altoona, PA	3	3	0	0	0	0	7

Table 3u. New Privately Owned Housing Units Authorized
Unadjusted Units by Metropolitan Area

January 2008

Different headers

Different column widths

Monthly Coverage	Total	1 Unit	2 Units	3 & 4 Units	5 Units or more	Num of Struc-tures With 5 Units
Percent						
Abilene, TX	14	10	4	0	0	0	91
Akron, OH	93	46	0	3	44	7	69
Albany, GA	24	22	2	0	0	0	84
Albany-Schenectady-Troy, NY	39	39	0	0	0	0	59
Albuquerque, NM	204	163	0	0	41	2	100
Alexandria, LA	41	41	0	0	0	0	97
Allentown-Bethlehem-Easton, PA-NJ	118	113	0	0	5	1	100
Altoona, PA	3	3	0	0	0	0	7

Different wrapping conventions

Different variables

MLS data:
Houston

Construction data:
Houston-Galveston-Brazoria
Houston-Baytown-Sugar Land, TX
Houston-Sugar Land-Baytown, TX

Population data:
Houston-Sugar Land-Baytown TX

Days of work...

Amarillo
Austin
Beaumont
Brownsville
College Station
Corpus Christi
Dallas
El Paso
Houston
Killeen
Laredo
Longview
Lubbock
McAllen
Midland
Odessa
San Angelo
San Antonio
Sherman
Tyler
Victoria
Waco
Wichita Falls

15
10
5

01
05
09

Houses (per 10000 people)

Date

Type
sold
built

Conclusions

A programming language gives you: reproducibility, automation, communication, but has a learning curve.

R gives you: freedom, a community, connectivity, building blocks, but the community can be prickly and it is (relative to other languages) slow.

ggplot2 gives you a way to succinctly describe visualisations, and practically makes it easy to create plots.