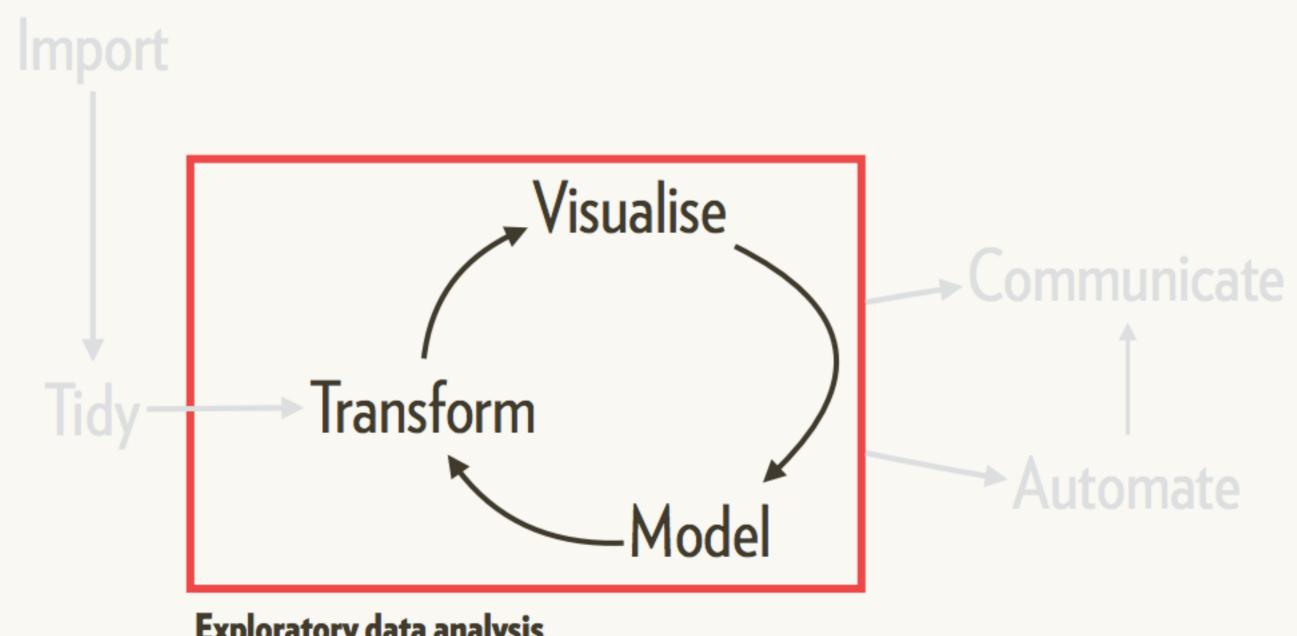
cognostics metrics for data visualization

Barret Schloerke Statistics PhD Candidate Purdue University

About myself

- Purdue University
 - 4th Year PhD Candidate in Statistics
 - Research in large data visualization using R <u>www.tessera.io</u>
 - Dr. William Cleveland and Dr. Ryan Hafen
- Metamarkets.com 1.5 years
 - Front end engineer node.js
- Iowa State University
 - B.S. in Computer Engineering
 - Research in statistical data visualization with R
 - Dr. Di Cook, Dr. Hadley Wickham, and Dr. Heike Hofmann

Exploratory Data Analysis



Exploratory data analysis

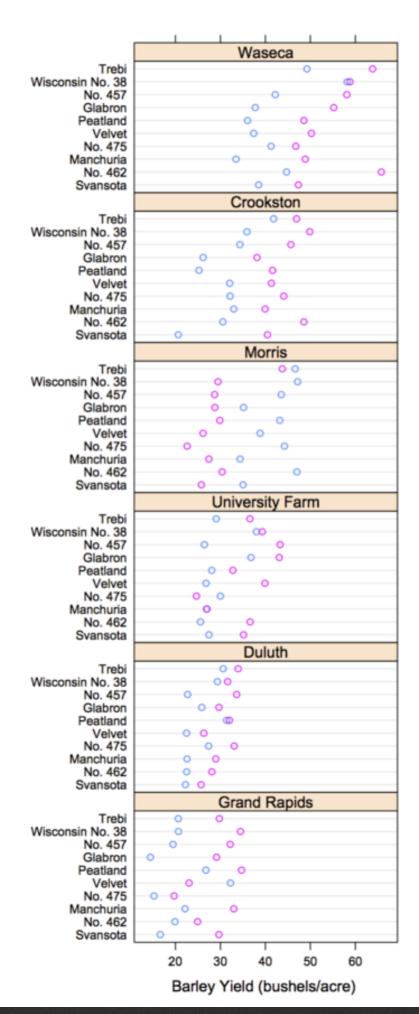
Visualization of Large Data

- Most large data visualization tools or approaches either
 - Summarize the large data to create a single plot
 - Are very specialized and heavily engineered for a particular domain
- Summaries are critical but can hide very interesting phenomena (e.g. Simpson's Paradox)
- Specialized tools can be useful but do not fit Exploratory Data Analysis paradigm (slow implementation)

We must be able to flexibly visualize complex data in detail even when the data is large!

Trellis Display

- Data is split into meaningful subsets, usually conditioning on variables of the dataset
- A visualization method is applied to each subset
- The image for each subset is called a "panel"
- Panels are arranged in an array of rows, columns, and pages, resembling a garden trellis
- facet()'ing in ggplot2



1932 1931

Why Trellis is Effective

- Flexible to create
 - Data complexity / dimensionality / size
 can be handled by splitting the data into subsets
 - Complete freedom with what is plotted in every panel
- Effective to consume
 - Understand one panel —> Understand every panel
 - Scanning across panels elicits comparisons to reveal repetition and change, pattern and surprise

Example / Data Description

- Monthly median home listing and number of units sold for 2,984 counties in the contiguous United States From 2008 to January 2016,
- Harvested from Quandl's Zillow

> housing %>% dplyr::group by(county, state)

```
Source: local data frame [247,082 x 7]
Groups: county, state [2,984]
                   county state time nSold medListPriceSqft
    fips
                     <fctr> <fctr>
  <fctr>
                                      <date>
                                              <dbl>
                                                               <dbl>
   06037 Los Angeles County
                               CA 2008-01-31 505900
                                                                 NA
   06037 Los Angeles County CA 2008-02-29 497100
                                                                 NA
   06037 Los Angeles County CA 2008-03-31 487300
                                                                 NA
   06037 Los Angeles County CA 2008-04-30 476400
                                                                 NA
   06037 Los Angeles County
                               CA 2008-05-31 465900
                                                                 NA
   06037 Los Angeles County
                               CA 2008-06-30 456000
                                                                 NA
   06037 Los Angeles County
                               CA 2008-07-31 445700
                                                                 NA
   06037 Los Angeles County
                               CA 2008-08-31 435300
                                                                 NA
   06037 Los Angeles County
                               CA 2008-09-30 426700
                                                                 NA
   06037 Los Angeles County
                                CA 2008-10-31 419800
                                                            273.3073
# ... with 247,072 more rows, and 1 more variables: medSoldPriceSqft <dbl>
```

Arizona



Display Information

Price / Sq. Ft.

Panel Layout

Panel Labels

⊕ Related Displays

<u>▼</u> Table Sort / Filter

III Univariate Filter

Active Cognostics

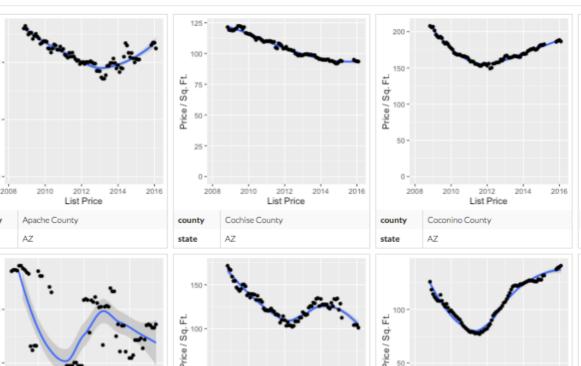
fields / Regular_Example /1 + **→**



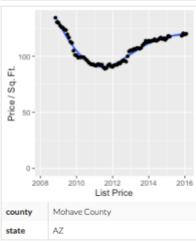








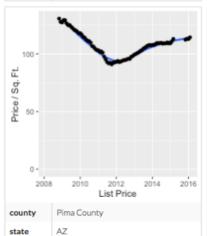




List Price

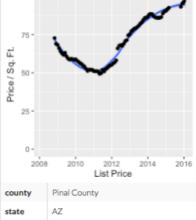
Gila County



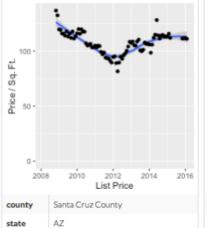


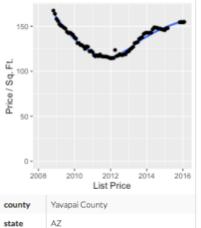
Greenlee County

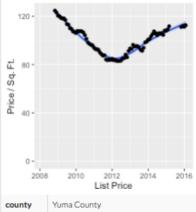
state



La Paz County







ΑZ

state

List Price

Georgia



Display Information

View Options

Panel Layout

Panel Labels

Cognostics

<u>III</u> <u>U</u>nivariate Filter

<u>₩</u> Bivariate Filter

Active Cognostics

fields/	Regular_	Example						1	/	1 +	→ 1x ·	-							0
S 158 List Pri	200024 List Pric	00 60 400 200024 List Pric county	O 100 - 50 - 50 - 50 - 50 - 50 - 50 - 50	b 198 7 200 200 200 200 200 200 200 200 200 2	D 75 - 200000000000000000000000000000000000	b 100 w 50 m	0 40 28 200004 List Pric	5 75 2208044 List Pric	S 68 - 200000000000000000000000000000000000	2208044 List Pric	50 75 2208044 List Pric	S 25 2208046 List Pric	D 105 - 105	0 75 - 28 2 2000000 List Pric	D BB : 2000 BB : 100	D 2208044 List Pric	S 220024 List Pric	D S S S S S S S S S S S S S S S S S S S	S 2200040 List Pric
22000000000000000000000000000000000000	b 100 y 200 200 200 200 200 200 200 200 200	0 75 1 200024 2 200024 List Pric	S 120 - S 120	List Pri	22000041 List Pric	2200000000 List Pric	List Pri	22000048 List Pri	22000000000000000000000000000000000000	00 60	D 120 → O 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 200000000000000000000000000000000000	0 75 2000044 List Pric	0 100 - 0 2000 130 County	22000044 List Pric	S 100 2000 1	22000041 List Pric	05 100 0 100	D 120 D 2200000
state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state
0 100 - 100	22000000000000000000000000000000000000	5 190 30 50 50 50 50 50 50 50 50 50 50 50 50 50	22000000000000000000000000000000000000	22000040 List Pric	S 2000000000000000000000000000000000000	2 200000000 List Pric	50 150 - 150	0 100 100 100 100 100 100 100 100 100 1	22000000000000000000000000000000000000	2 200000000000000000000000000000000000	22080240 List Pric	0 100 3 100 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	S 150 - S 2000000000000000000000000000000000	22000000000000000000000000000000000000	O BB C 2200000000000000000000000000000000	S 125 W	D 22380241	S 100 - S 100	0 100 - 50 - 100 S 50 - 100 S 20090189
state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state
S 100 - 100	b 158 :♥ / e 22220000 List Pri	0 75 - 0 200000000	0 190 : √ 0 200 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	S 200 200 200 200 200 200 200 200 200 20	0 100 . √ 50 . √ 20 200 List Pri	0 100 € 0 50 - 0 20	0 125 ₹ 20 125 ₹ 20 125 ₹ List Pri	Ø 158 ♣ / e 2220081 List Pri	Ø 80 € V 28 200000000000000000000000000000000000	O List Pri	Ø 100 -	0 75 - 0 2000040 List Price	Ø 80 = ▼ / 0 20000000000000000000000000000000000	0 75	0 40 1000 0 40 1000 0 200024 0 List Price	Ø 75 : ✓ 28 : ✓ 38 : 10	0 20 200000000000000000000000000000000	200904 List Pric	5 60 100 9 20 2008044 List Pric
county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county
00 48 - 10 200024 List Price	S 10 10 10 10 10 10 10 10 10 10 10 10 10	0 75 - 0 2000000000000000000000000000000000	0 75 - 200000000000000000000000000000000000	g 199	0 80	0 100 - 0 100	b 190 - 28 - 28 - 28 - 28 - 28 - 28 - 28 - 2	0 100 - 0 200000000000000000000000000000000	0 100 - 0 2000000	00 100 100 100 100 100 100 100 100 100	Ø 100 - ₩ / 50 2 2220090	05 68	0 75 - 10 2000000000000000000000000000000000	0 190 190 190 190 190 190 190 190 190 19	0 75 - 200000000 200000000000000000000000000	S 150 2000000	00 80 - 00 20000000000000000000000000000	0 100 TA 0 20 200 TA 0 200 TA 0 200 TA 1 100 TA 0 200 TA 1 100 TA 1 1	0 25 25 200004 List Pric
county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county
5 190	5 88 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	5150 V 100 V 50 V 5	575 - 200000000000000000000000000000000000		State St	State 5 100 - 50 - 50 - 50 - 50 - 50 - 50 - 50	5 75 200000000 List Pric	5 75 . ▼	State D	50 100 - ₩ 50 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 -	57.5	State D 100 P	575	50 60	State S 158 - S 2000 A 158 - S 2000	50 150 150 150 150 150 150 150 150 150 1	50 150 50 50 50 50 50 50 50 50 50 50 50 50 5	State D 158 1 2000024 D 2000024 List Pri	50
state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	county	state	state
2200000000 List Price	5 75 2000024 - 2000024 - List Pric	2000004 List Pric	S 100 - S 2228008	S 75 - ₩ 2000000000000000000000000000000000000	20000000000000000000000000000000000000	60 3 40 1 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2000001 List Pric	5 190 190 190 190 190 190 190 190 190 190	68 3 2 222222222222222222222222222222222	05 75	S E E E E E E E E E E E E E E E E E E E	05 158 1 220000000 List Pri	0 60	22000000000000000000000000000000000000	05 75	220000010 List Pric	05 158 2000 2000 A List Pri	200024 List Pric	0 75 0 25 2000000000000000000000000000000000
county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county
state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state
220904 List Prio	2008004 List Pric	0 100 - 50 - 0 0 2 200000000000000000000000		List Pric	20000000000000000000000000000000000000	22000000000000000000000000000000000000			0 75 200004 List Pric	200024 List Pric	List Pri	22000000000000000000000000000000000000	200024 List Pric	0 75	200024 List Pric	0 (8 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0			
state	state	county	county	county	county	county	state	county	county	county	state	state	county	county	state	county	county	county	state

Scaling Trellis

- Large data lends itself nicely to the idea of Trellis
 Display
 - Typically comprised of collections of smaller data from many subjects, sensors, locations, time periods, etc.
 - It is natural to break the data up based on these dimensions and make a plot for each subset
- Potentially thousands or millions of panels
 - Will never be able to (or want to) view all of them!

Scaling Trellis with Cognostics

- Scaling Trellis:
 - Data are split into meaningful subsets, usually conditioning on variables of the dataset
 - A visualization method is applied to each subset
 - A set of cognostics that measure attributes of interest for each subset is computed
 - Panels are arranged in an array of rows, columns, and pages, resembling a garden trellis, with the arrangement being specified through interactions with the cognostics

Scaling Trellis with Cognostics

- Scaling Trellis:
 - Data are split into meaningful subsets, usually conditioning on variables of the dataset
 - A visualization method is applied to each subset
 - A set of cognostics that measure attributes of interest for each subset is computed
 - Panels are arranged in an array of rows, columns, and pages, resembling a garden trellis, with the arrangement being specified through interactions with the cognostics
- Can be achieved with the Trelliscope package

Trelliscope Demo

Calculated Cognostics

```
advanced_cog <- function(x) {</pre>
  zillow_string <- gsub(" ", "-", do.call(paste, getSplitVars(x)))</pre>
 model <- loess(</pre>
    medListPriceSqft ~ as.numeric(time),
    data = subset(x, !is.na(medListPriceSqft))
  residuals <- model$residuals
 list(
    res_std_err = cog(model$s, desc = "residual standard error"),
    enp = cog(model$enp, desc = "effective number of parameters"),
    mean_list = cogMean(x$medListPriceSqft),
    n_obs_list = cog(
      length(which(!is.na(x$medListPriceSqft))),
      desc = "number of non-NA list prices"
    zillow_href = cogHref(
      sprintf("http://www.zillow.com/homes/%s_rb/", zillow_string),
      desc = "zillow link"
```

Automatic Cognostics

- Cumbersome to manually specify many cognostics for a Trelliscope display
- Should be able to automatically compute cognostics based on the context of what is being plotted
 - Help foster a scalable Trellis system
- Analyze the plot objects and choose "best" cognostics based on the plot specification
 - ggplot2
 - · rbokeh
 - · plotly

For example...

- For scatterplot layers:
 - Number of observations
 - Number of missing values
 - Summary statistics of x and y-axis variables
- For statistical layers (such as geom_smooth)
 - RMSE of fit
 - Number of outliers
- Etc. (much research going on / to be done here...)

Future work

- Continually adding more cognostics to be created for each plotting layer
- · Implement
 - Fully integrate into trelliscope
 - · Parse ggplot2, rbokeh, and plotly objects

Questions?

www.tessera.io github.com/tesseradata github.com/schloerke