cognostics metrics for data visualization

Barret Schloerke Statistics PhD Candidate Purdue University

About myself

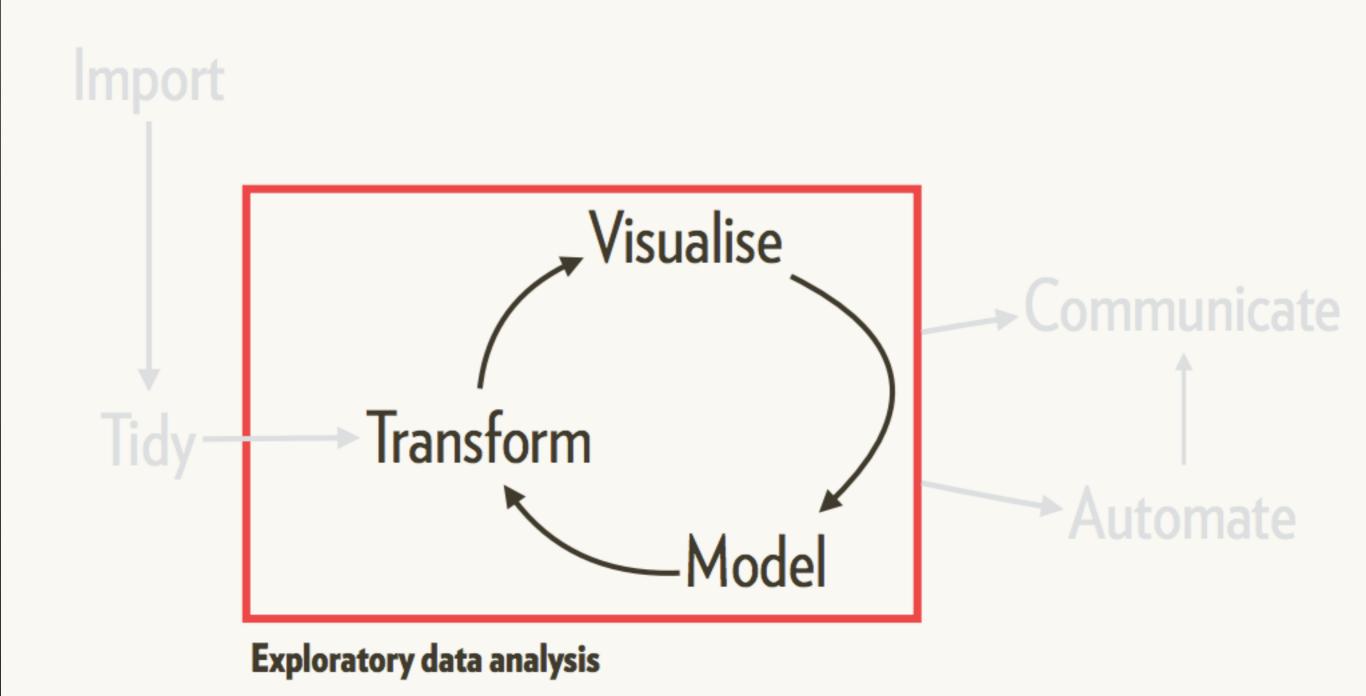
Purdue University

- 4th Year PhD Candidate in Statistics
- Research in large data visualization using R http://deltarho.org
 - 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



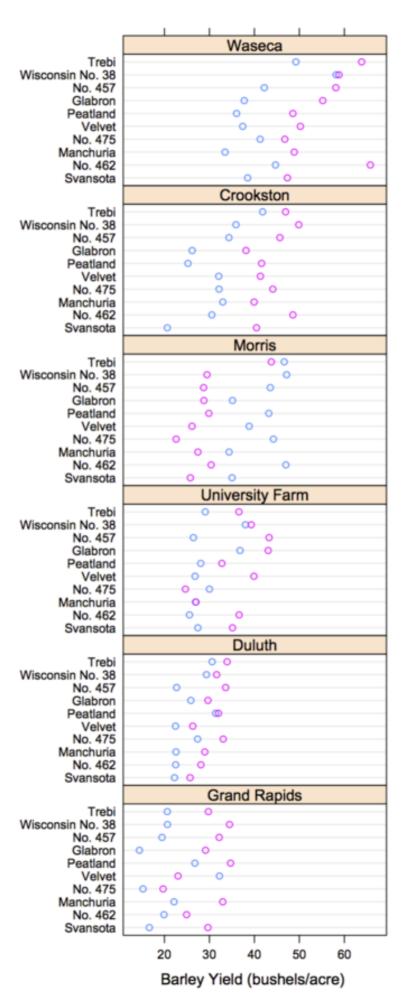
Visualization of Large Data

- 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

View Options

Panel Layout

Panel Labels

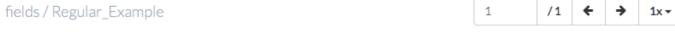
Related Displays

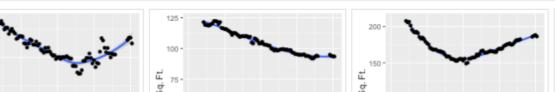
Cognostics

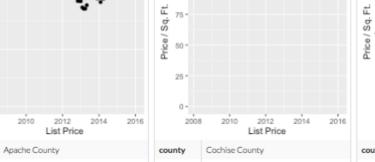
<u>▼</u> Table Sort / Filter

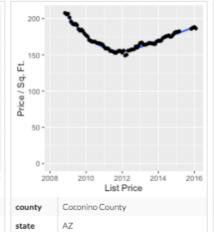
<u>III</u> Univariate Filter

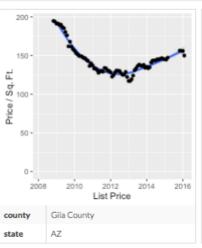
Active Cognostics

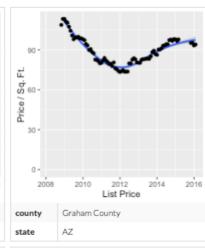


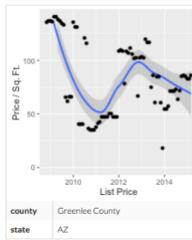




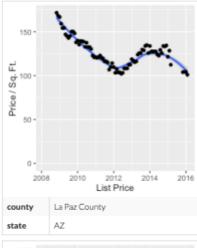


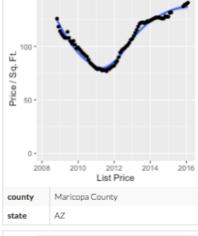


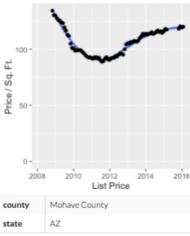


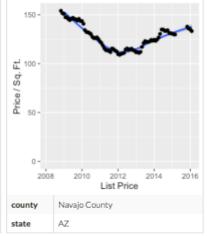


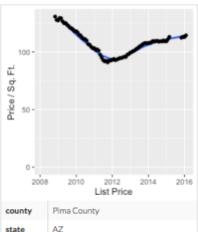
Price / Sq. Ft.

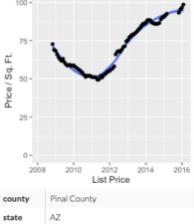


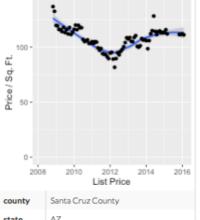


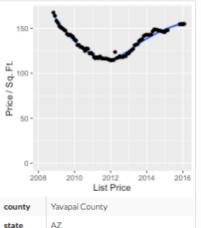


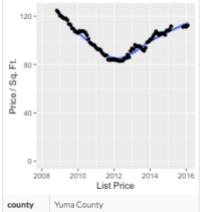












Georgia



Display Information

View Options

Panel Layout

Panel Labels

Cognostics

<u>I</u> Table Sort / Filter

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

Active Cognostics

fields / Regular_Example 1 /1 ← → 1x →																0			
ο 150 150 220000000 2200000000 List Prio	Dice / 200024 List Price	00 60 200024 00 200024 01 List Price	0 100 - 100	00 100 00 20000000000000000000000000000000000	0 75 : ▼ 2000000000000000000000000000000000000	b 100 ▼ 20 20000000000000000000000000000000000	00 48	00 75 - 1 22000040	0 220000000000000000000000000000000000	200004 List Price	S 75 2000040 2000040 List Pric	0 75	o lost Pri	0 75 - 200000000000000000000000000000000000	0 88 200 0 200 200 List Price	00 € 100 20000040 List Price	200024 List Pric	22000000000000000000000000000000000000	5 75 2200044 OL List Price
county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county
S 88 ▼	S 190	S 75	20 - 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Duce / Solution / Solu	S 88 200004	e/Sq.	S 199 S	20000000000000000000000000000000000000	S 48 - V	0 60 - 50 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	S 120 >	. S / 28	2000004 List Pric	S 199	20000010 List Price	S / 93	Sq.	S 100 - 50	b 1200 🖢
List Price	د List Pri	200024 List Pric			△ List Pric	22000000000000000000000000000000000000			2 200000000000000000000000000000000000		List Pri	22000000000000000000000000000000000000					200004		ŭ List Pri
state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state	state
50 100 - 50 - 100	2 220000000000000000000000000000000000	5 190 3 5 28 20220090 CList Prior	0 60	0 10 10 10 10 10 10 10 10 10 10 10 10 10	0 2200000000 List Pric	0 88 → 2200000000 List Price	50 - 1	0 190 3 190	2 220000000000000000000000000000000000	0 88 2000 2000 2000 2000 2000 2000 2000	0 90 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	5 190 28 2222089 C. List Pri	S 150	S 200000000	S BB: V	S 126 2000 2000 2000 2000 2000 2000 2000	0 222000240 List Pric	S 100 - 50 - 50 - 22222111	S 100 - 50 - 100 - 50 - 100 - 50 - 100 - 50 - 100 - 50 - 100 - 50 - 100 - 50 - 100 - 50 - 5
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
2000118 List Pri	58 - 222200000000000000000000000000000000	20000000000000000000000000000000000000	05 190 :	200 200 200 200 200 200 200 200 200 200	0 100 1 0 200 1 0 200 1 0 200 1 0 100 1 0 200 1 0 100 1 0 200 1 0 100 1 0	bs / epi-	S 126 € S 2000 2000 2000 2000 2000 2000 2000 20	05 158 : 158 : 2000000000000000000000000000000000000	00 20000000000000000000000000000000000	S / S / S / S / S / S / S / S / S / S /	0 100 - 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2200004 6. List Price	20000000000000000000000000000000000000	20000000000000000000000000000000000000	200024 List Pric	28888024 List Pric	20000000000000000000000000000000000000	2200004 List Pric	0 48 1 20 20 20 20 20 20 20 20 20 20 20 20 20
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
200024 List Pric	00 88 100 200024 02 200024 04 List Pric	22000000000000000000000000000000000000	22000000000000000000000000000000000000	2020000 List Pri	20000000000000000000000000000000000000	0 190 190 190 190 190 190 190 190 190 19	S 28 2000 2000 2000 2000 2000 2000 2000	0 100 - 0 0 22000004 List Pri	00 2000m 2000m List Pri	2000 2000 2000 2000 2000 2000 2000 200	05 100 - 0 50 - 0 22220080	28: 22:09004 6. List Pric	22000000000000000000000000000000000000	S / 50 190 190 20000004 List Pri	00 75	05 150 200000000000000000000000000000000	22000000000000000000000000000000000000	20000000 List Pri	05 75 1 1 2220024 2220024 List Pric
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
S 190 S S 200 S List Pri	00 88 200024 200024 List Pric	0 150 - 0 150	22000000000000000000000000000000000000	20000000000000000000000000000000000000	20000000000000000000000000000000000000	0 100 - 100	25 2000000000000000000000000000000000	00 75 . ₩ 20 20000000000000000000000000000000000	00 2020004 List Pric	0 100 - 50 - 100 -	2009044 List Pric	00 200 200 200 200 200 200 200 200 200	20000000000000000000000000000000000000	50 50 50 50 50 50 50 50 50 50 50 50 50 5	05 158 : 0 58 : Zazandan List Pri	05 150	00 150 3 √ 00 2000000000000000000000000000000000	05 150	00
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
bS/⊕ 2200000000000000000000000000000000000	0 75 20 20 20 20 20 20 20 20 20 20 20 20 20	22000040 List Pric	0 100 - 50 0 100 - 50 0 100 - 50 227771111	20000000000000000000000000000000000000	Duck Price	0 58 3 4 4 6 List Pric	22000000 List Pric		0 48 - → 0 28	2000040 List Pric	0 200004 List Pric	S 150 22000004 List Pri		00 22000000000000000000000000000000000	2200004 List Pric	22000040 List Pric	S 150 2000 2000 2000 List Pri	200024 List Pric	20000000000000000000000000000000000000
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
22000040 List Pric	2200024 List Pric	0 100 - 50 - 0 7 2 2000000000000000000000000000000	22000000000000000000000000000000000000	22000000000000000000000000000000000000	S 75 : S 25 : S	28000000000000000000000000000000000000	200024 List Pric	2200004 List Pric	25	200024 List Pric	0 100 - ₩ 50 - 100 - ₩ 22221131 List Pri	20000000000000000000000000000000000000	0 48 - 10 20 20 20 20 4	2200004 List Pric	200024 List Pric	25 25 20 22 22 22 22 22 22 22 22 22 22 22 22			
county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county	county

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?

http://deltarho.org github.com/delta-rho github.com/schloerke