Denver RUG (DRUG) Inaugural Meeting: Adventures with ggplot2 and MLB Data

Ryan Elmore

rtelmore@gmail.com

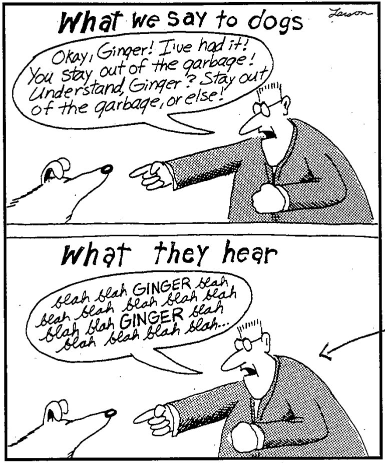
@rtelmore

National Renewable Energy Lab

14 September 2010

Organizers

- Scott Sibbel (@ScottSibbel), Epidemiology Research Fellow, CSPH
- Melissa Santos (@ansate), Ph.D. Student, UC Denver
- Matt Parker (@mmparker), Statistical Research Specialist, Denver Metropolitan TB Clinic
- Me


What we say to dogs
Larson
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!
What they hear
blah blah GINGER blah blah blah blah blah blah blah blah GINGER blah blah blah blah blah blah...

The Far Side®
•2007•
AUGUST
14
Tuesday

Math/Stats/CS Professor during seminar

Audience during seminar

Question: Do we need a Denver/Front Range RUG? Answer: Yes!

Why? Two primary reasons:

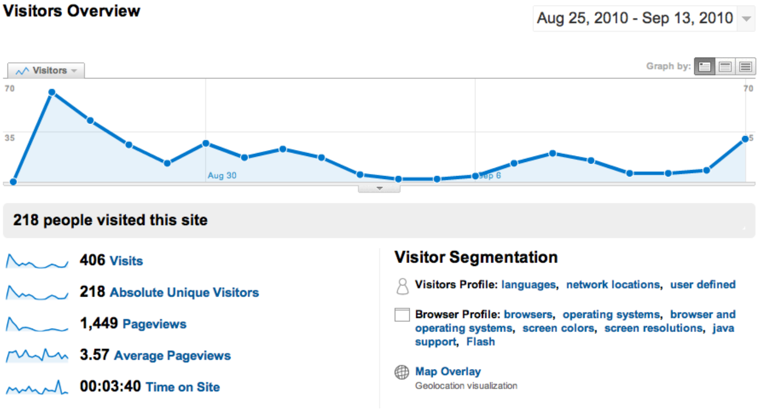
- To connect users in the Front Range.
- To facilitate the learning of R.

Think about what you would want with respect to types of talks (tutorial vs. analysis), a presence outside of meetup.com, etc.

A not-at-all-random sampling of comments ...

- "Finding out types of **analysis** people do with R."
- "I hope to **meet** new people and **learn** more about R."
- "**Contacts** that share an interest/enthusiasm in the language and are willing to **help** one another along the road of learning."
- "Become a **Jedi** with R!"

From Google Analytics



Visitors Overview

Aug 25, 2010 - Sep 13, 2010

218 people visited this site

406 Visits

218 Absolute Unique Visitors

1,449 Pageviews

3.57 Average Pageviews

00:03:40 Time on Site

Visitor Segmentation

Visitors Profile: languages, network locations, user defined

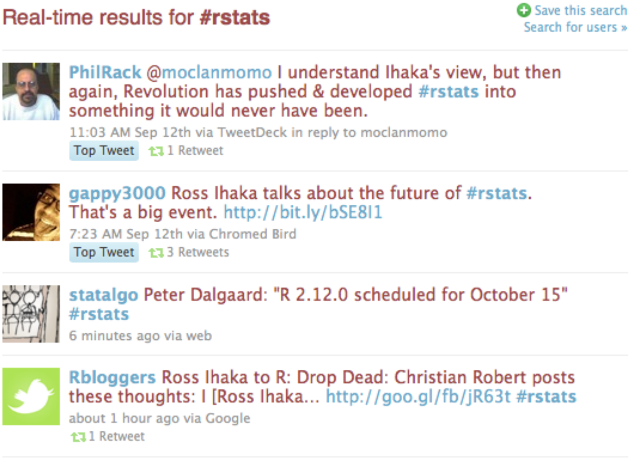
Browser Profile: browsers, operating systems, browser and operating systems, screen colors, screen resolutions, java support, Flash

Map Overlay
Geolocation visualization

R Resources on the Web

- CRAN (`http://streaming.stat.iastate.edu/CRAN/`)
- R-bloggers (`http://www.r-bloggers.com/`)
- Stack Overflow (`http://stackoverflow.com/tags/r/`)
- ggplot2 (`http://had.co.nz/ggplot2/`)
- R-forge (`http://r-forge.r-project.org/`)
- Twitter with rstats hash tag
- Revolution (`http://blog.revolutionanalytics.com/`)

Twitter



Real-time results for **#rstats**

Save this search
Search for users »

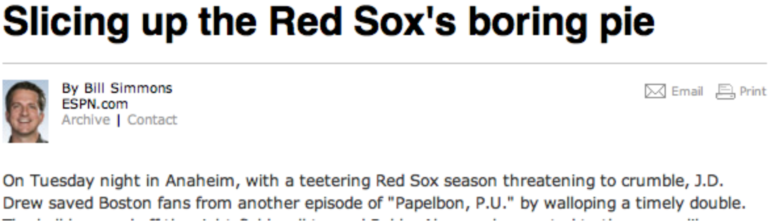
PhilRack @moclanmomo I understand Ihaka's view, but then again, Revolution has pushed & developed #rstats into something it would never have been.
11:03 AM Sep 12th via TweetDeck in reply to moclanmomo
Top Tweet 1 Retweet

gappy3000 Ross Ihaka talks about the future of #rstats. That's a big event. http://bit.ly/bSE8I1
7:23 AM Sep 12th via Chromed Bird
Top Tweet 3 Retweets

statalgo Peter Dalgaard: "R 2.12.0 scheduled for October 15" #rstats
6 minutes ago via web

Rbloggers Ross Ihaka to R: Drop Dead: Christian Robert posts these thoughts: I [Ross Ihaka... http://goo.gl/fb/jR63t #rstats
about 1 hour ago via Google
1 Retweet

It all started at DIA ...



Slicing up the Red Sox's boring pie

By Bill Simmons
ESPN.com
Archive | Contact

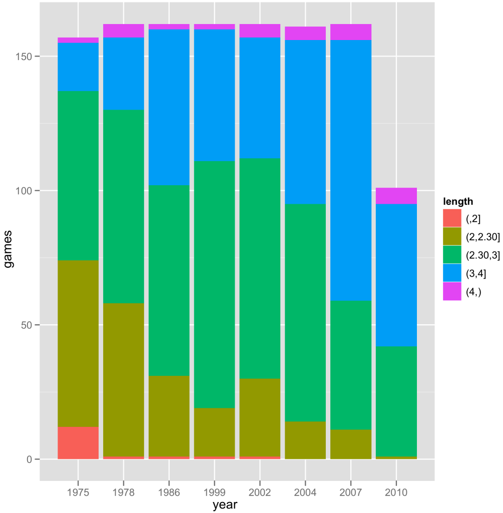
Email Print

On Tuesday night in Anaheim, with a teetering Red Sox season threatening to crumble, J.D. Drew saved Boston fans from another episode of "Papelbon, P.U." by walloping a timely double.

Q: Why are people losing interest in Red Sox games?

A: "**THE TIME OF THE GAMES: 55 PERCENT**
The biggie. The hammer. The killer."

Simmons' Red Sox Data



The Summarized Data

```
> head(sim.dat)
  Year Cat Count
1 1975   A    12
2 1975   B    62
3 1975   C    63
4 1975   D    18
5 1975   E     2
6 1978   A     1
```

R Code

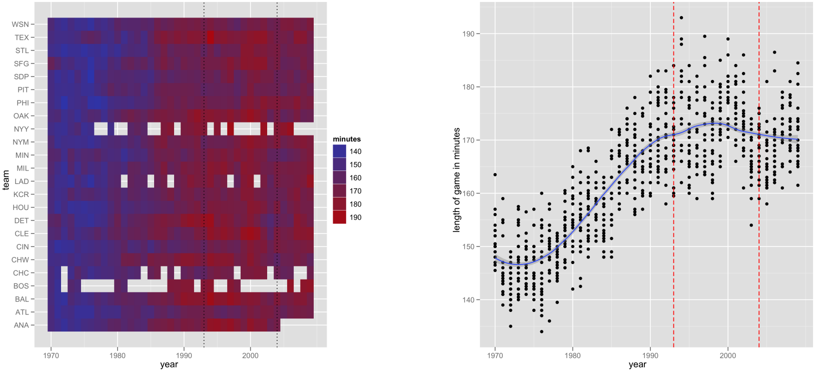
```r
ggplot(data = sim.dat, aes(x = as.factor(Year),
   y = Count, fill = Cat)) +
  geom_bar(stat = ’identity’) +
  scale_fill_discrete(name = "length",
  breaks = unique(sim.dat$Cat)) +
  scale_x_discrete("year") +
  scale_y_continuous("games")

ggsave("~/Sports/Simmons/mlb_length_1.png",
   hei = 7, wid = 7)
```

A Few Thoughts ...

- The underlying “unit” for the data is minutes. So what?
- We are only looking at a subset of years here. Why these particular years?
- How do the Red Sox compare to all of the other teams in Major League Baseball?
- Ah crap, I have to download a lot of data...and my flight boards in ten minutes!

More Figures



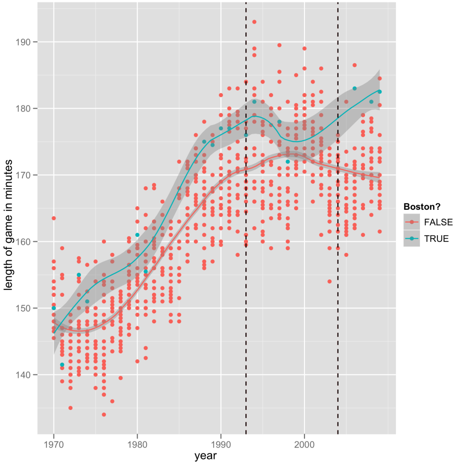
Examples of geom_tile() and geom_point() with smooth layer.

More R Code

```
ggplot(data = rs.dat, aes(x = year, y = team,
   fill = med_len)) +
  geom_tile() +
  scale_fill_continuous("minutes")
last_plot() + geom_vline(x=c(1993,2004),lty=3)

qplot(x = year, y = med_len, data = rs.dat,
  geom = c("point","smooth"), span = .5,
  ylab = "length of game in minutes")
last_plot() + geom_vline(x = c(1993,2004),
   lty = 2, col = "red")
```

Boston vs !Boston



More R Code

```
rs.dat$bs <- rs.dat$team == ’BOS’

qplot(x = year, y = med_len, data = rs.dat,
  geom = c("point","smooth"), span = .5,
  colour = bs,
  ylab = "length of game in minutes")
last_plot() + scale_colour_discrete(name = "Boston?")
last_plot() + geom_vline(x = c(1993,2004),
  lty = 2, col = "black")
ggsave("~/Sports/Simmons/mlb_length_5.png",
  hei = 7, wid = 7)
```

Final thoughts

- Anecdotal evidence towards an increase in game length; however, a slight decrease in recent years.
- Share your code when you can! The code for this project is posted at github.
- Give back to the R community if possible.
- Blog, tweet, etc. about what you find!

Where do you go from here?

Check out

- `thelogcabin.wordpress.com`
- `http://github.com/rtelmore`

and of course

- `http://www.meetup.com/DenverRUG/`

Where do we go from here?

- Types of talks?
- Email lists for R questions?
- Anything else?
- Network!