Implementing R

in a Small IR Office

Gary R. Moser
Institutional Research Analyst
Heald College
garyrmoser@gmail.com

Heald College
EST. 1863

Goals for this Presentation

1. Introduce R and its uses for Institutional Researchers

2. Demonstrate how I've used R

3. Provide guidance to facilitate your own implementation

My Reasons for Learning R…

1. Extensibility (SPSS DevCentral, VBA, Python…)
2. Expansion of skill set to include more powerful tools
3. Zero cost – Low risk
4. Portability
5. Fun!

What is R?

- R is a Language *AND* an Environment for statistical computing and graphics
- An open-source version of the S language developed at Bell Laboratories by John Chambers and colleagues.
- R is **free** software under the terms of the Free Software Foundation's GNU General Public License

http://www.r-project.org/

What is R Useful For?

- Data Acquisition
  - Web, Excel, ODBC/SQL, .CSV, SAS, SPSS, Stata, etc…
- Data Manipulation
  - Simple to very complex
- Statistical Analysis
  - Wide array of statistical methods and procedures
- Recurring and Ad-Hoc Reporting and Analysis
- Graphics!

http://www.r-project.org/

Pros:

- Vast‡ number of analytical methods
- Publication-quality graphics
- Very active and supportive user community
- Powerful language - no need for separate macro or scripting languages‡
- Useful for analyzing aggregated‡ data
- Many add-on packages, easy to install
  (e.g. googleVis, ggplot2, plyr, wordcloud…)
- **FREE!**

Heald COLLEGE EST. 1863

*partially adapted from Andrew Robinson's icebreakR, available at http://cran.r-project.org/other-docs.html

Cons:

- Steep initial learning curve – important to find good learning resources
- Conceptually and functionally different than SAS and SPSS
- Pretty tables (currently) require work
- Limited GUI support
- Memory limited to RAM
  (rarely an issue - many ways to handle)
- Not well-suited to brute force solutions
- “R will not hold your hand”

*partially adapted from Andrew Robinson’s icebreakR, available at http://cran.r-project.org/other-docs.html

Is R Accurate?

Yes (as accurate as SPSS and SAS)*

*References:

Comparative study of the reliability of nine statistical software packages:
http://www.sciencedirect.com/science/journal/01679473/51/8

Software Comparison:
http://finzi.psych.upenn.edu/R/Rhelp02a/archive/97802.html

Wikibooks Comparison of Statistical Software:
http://en.wikibooks.org/wiki/Statistics/Numerical_Methods/Numerical_Comparison_of_Statistical_Software

The UserR Community:

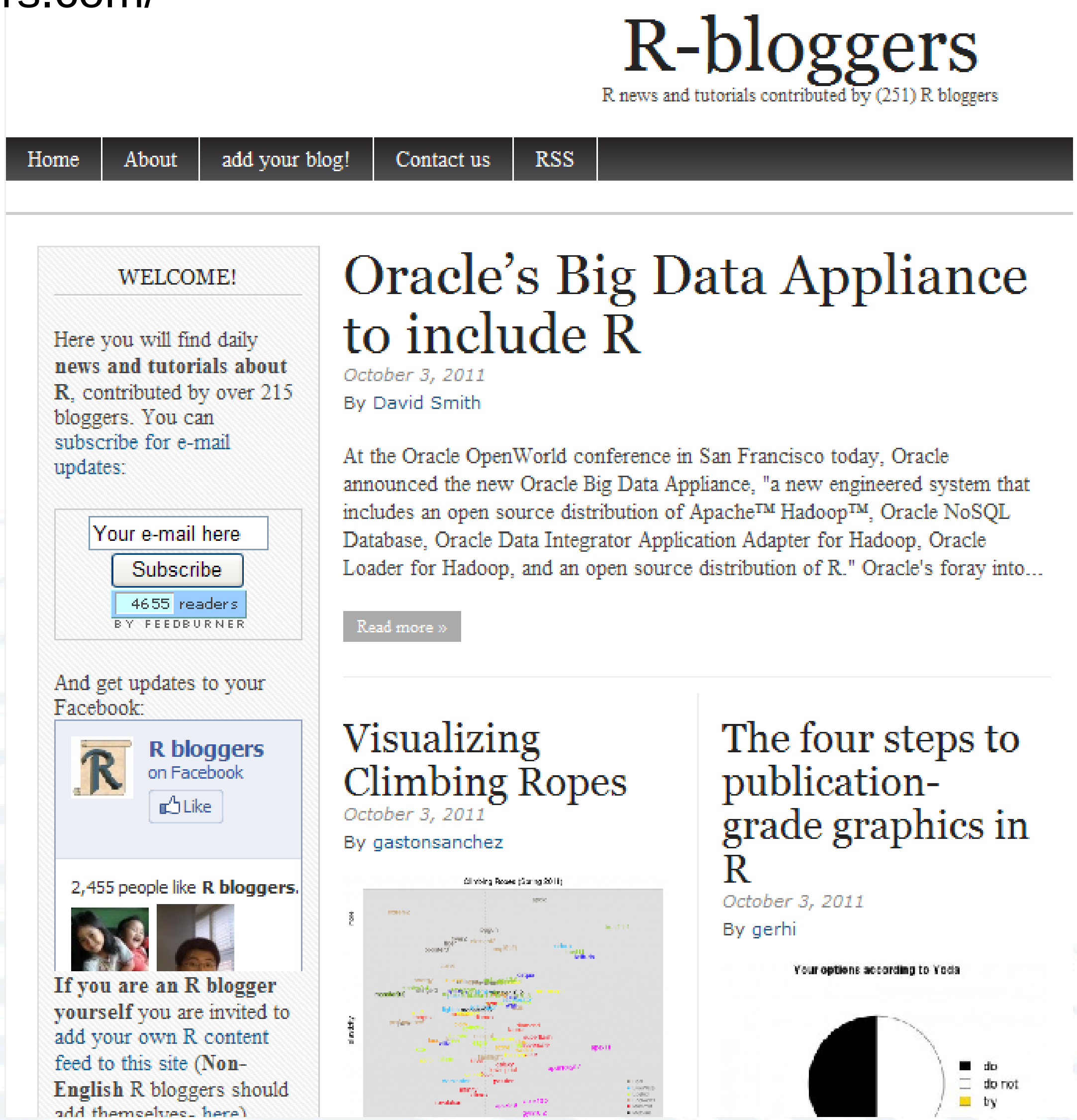
R for SPSS, SAS, and Stata Users (very good):
http://r4stats.com/

R Listserve
r-help@r-project.org
Subscribe: https://stat.ethz.ch/mailman/listinfo/r-help

R Blog Aggregator
http://www.r-bloggers.com/

Bay Area (& others!) UseR Meetup Group
http://www.meetup.com/R-Users/

http://www.r-bloggers.com/



R-bloggers

R news and tutorials contributed by (251) R bloggers

Home | About | add your blog! | Contact us | RSS

WELCOME!

Here you will find daily **news and tutorials about R**, contributed by over 215 bloggers. You can subscribe for e-mail updates:

Your e-mail here

Subscribe

4655 readers BY FEEDBURNER

And get updates to your Facebook:

R bloggers on Facebook

Like

2,455 people like **R bloggers**.

If you are an R blogger yourself you are invited to add your own R content feed to this site (**Non-English** R bloggers should add themselves- here)

Oracle's Big Data Appliance to include R

October 3, 2011

By David Smith

At the Oracle OpenWorld conference in San Francisco today, Oracle announced the new Oracle Big Data Appliance, "a new engineered system that includes an open source distribution of Apache™ Hadoop™, Oracle NoSQL Database, Oracle Data Integrator Application Adapter for Hadoop, Oracle Loader for Hadoop, and an open source distribution of R." Oracle's foray into...

Read more »

Visualizing Climbing Ropes

October 3, 2011

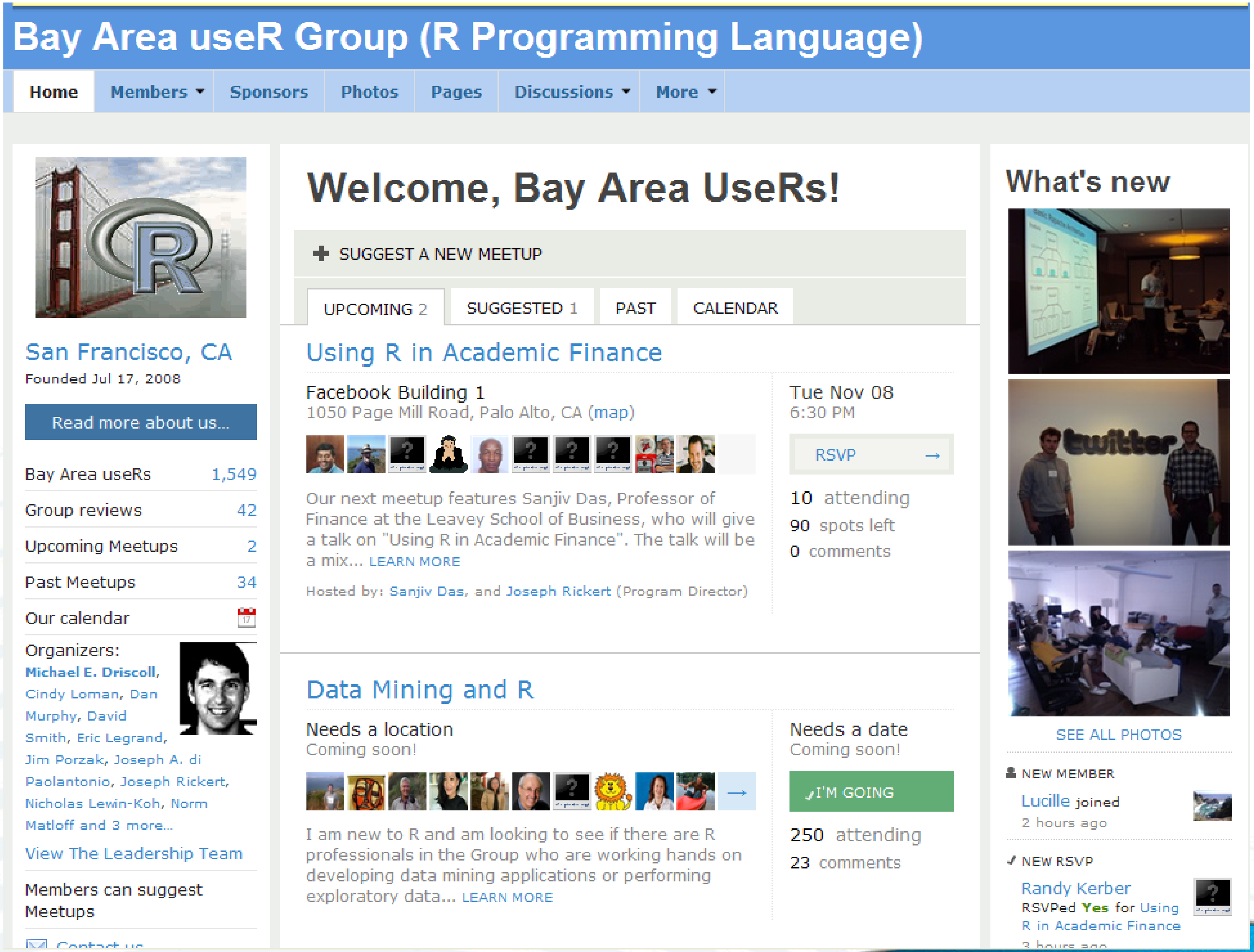
By gastonsanchez

The four steps to publication-grade graphics in R

October 3, 2011

By gerhi

http://www.meetup.com/R-Users/



Bay Area useR Group (R Programming Language)

San Francisco, CA

Founded Jul 17, 2008

Read more about us...

Bay Area useRs	1,549
Group reviews	42
Upcoming Meetups	2
Past Meetups	34
Our calendar	

Organizers: **Michael E. Driscoll**, Cindy Loman, Dan Murphy, David Smith, Eric Legrand, Jim Porzak, Joseph A. di Paolantonio, Joseph Rickert, Nicholas Lewin-Koh, Norm Matloff and 3 more...

View The Leadership Team

Members can suggest Meetups

Contact us

Welcome, Bay Area UseRs!

✚ SUGGEST A NEW MEETUP

Using R in Academic Finance

Facebook Building 1
1050 Page Mill Road, Palo Alto, CA (map)

Tue Nov 08
6:30 PM

RSVP →

Our next meetup features Sanjiv Das, Professor of Finance at the Leavey School of Business, who will give a talk on "Using R in Academic Finance". The talk will be a mix... LEARN MORE

Hosted by: Sanjiv Das, and Joseph Rickert (Program Director)

10 attending
90 spots left
0 comments

Data Mining and R

Needs a location
Coming soon!

Needs a date
Coming soon!

✓I'M GOING

I am new to R and am looking to see if there are R professionals in the Group who are working hands on developing data mining applications or performing exploratory data... LEARN MORE

250 attending
23 comments

What's new

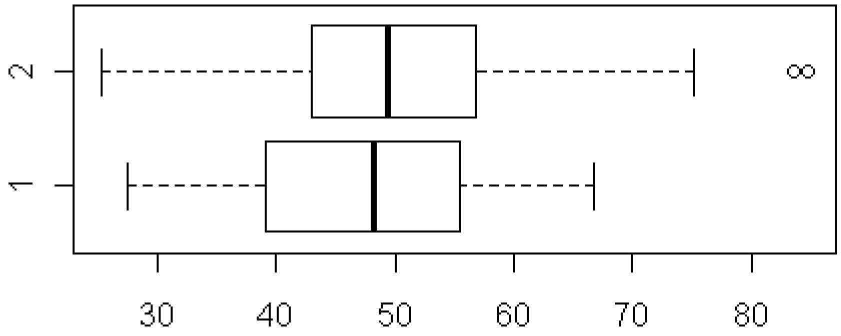
SEE ALL PHOTOS

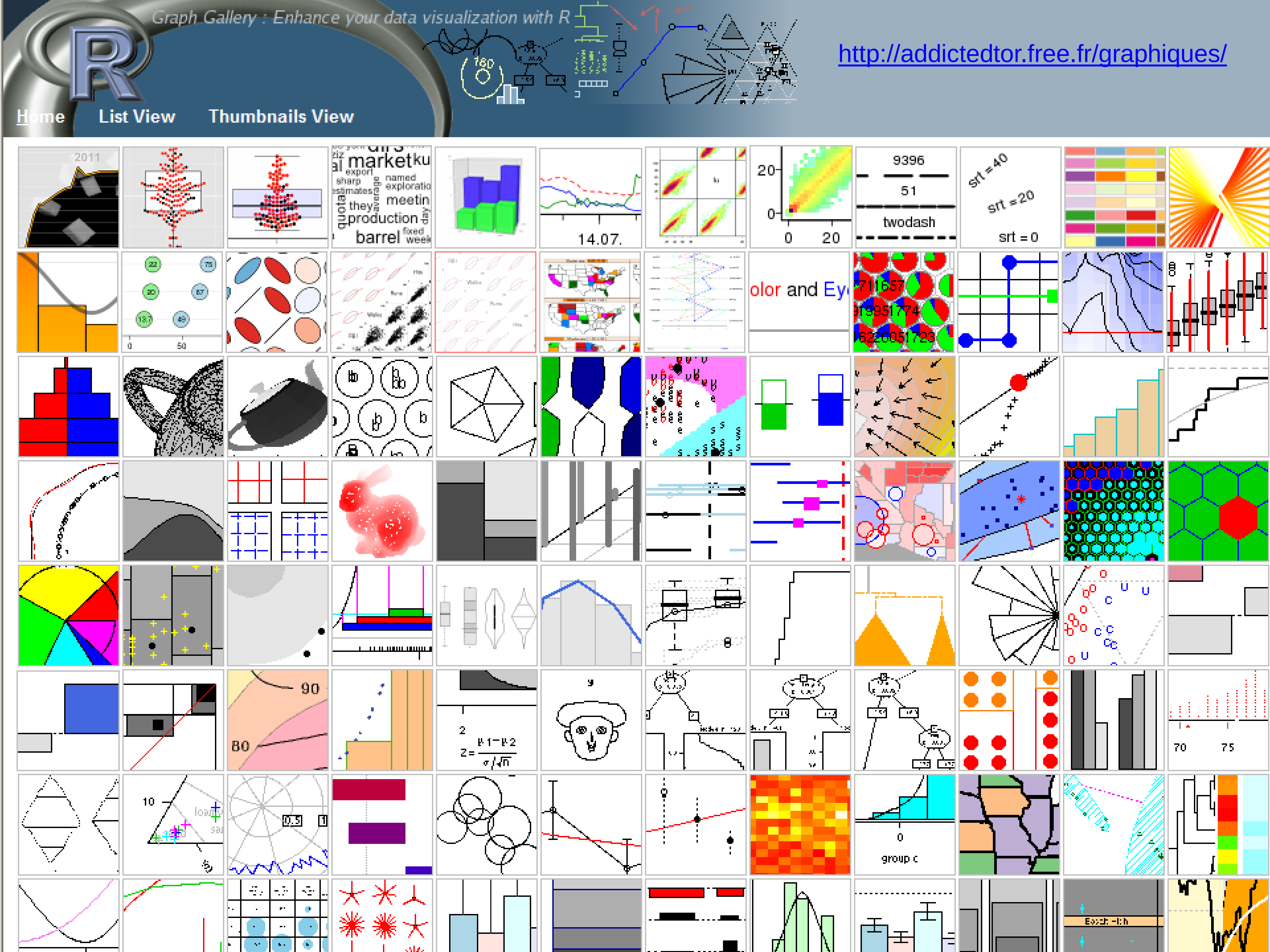
NEW MEMBER
Lucille joined
2 hours ago

✓ NEW RSVP
Randy Kerber
RSVPed Yes for Using R in Academic Finance
3 hours ago

Heald COLLEGE
EST. 1863

Graphics!



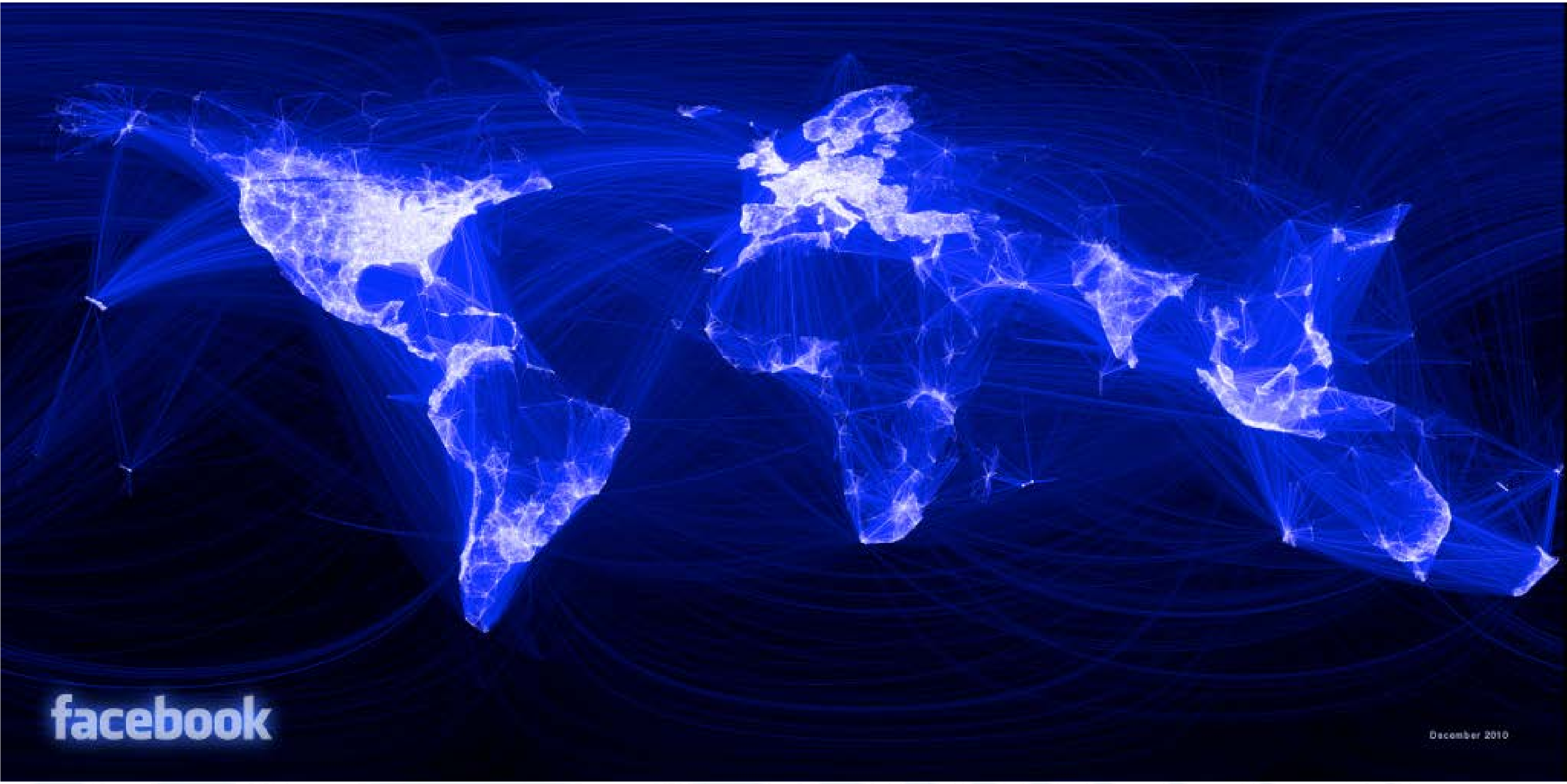


Graph Gallery : Enhance your data visualization with R

http://addictedtor.free.fr/graphiques/

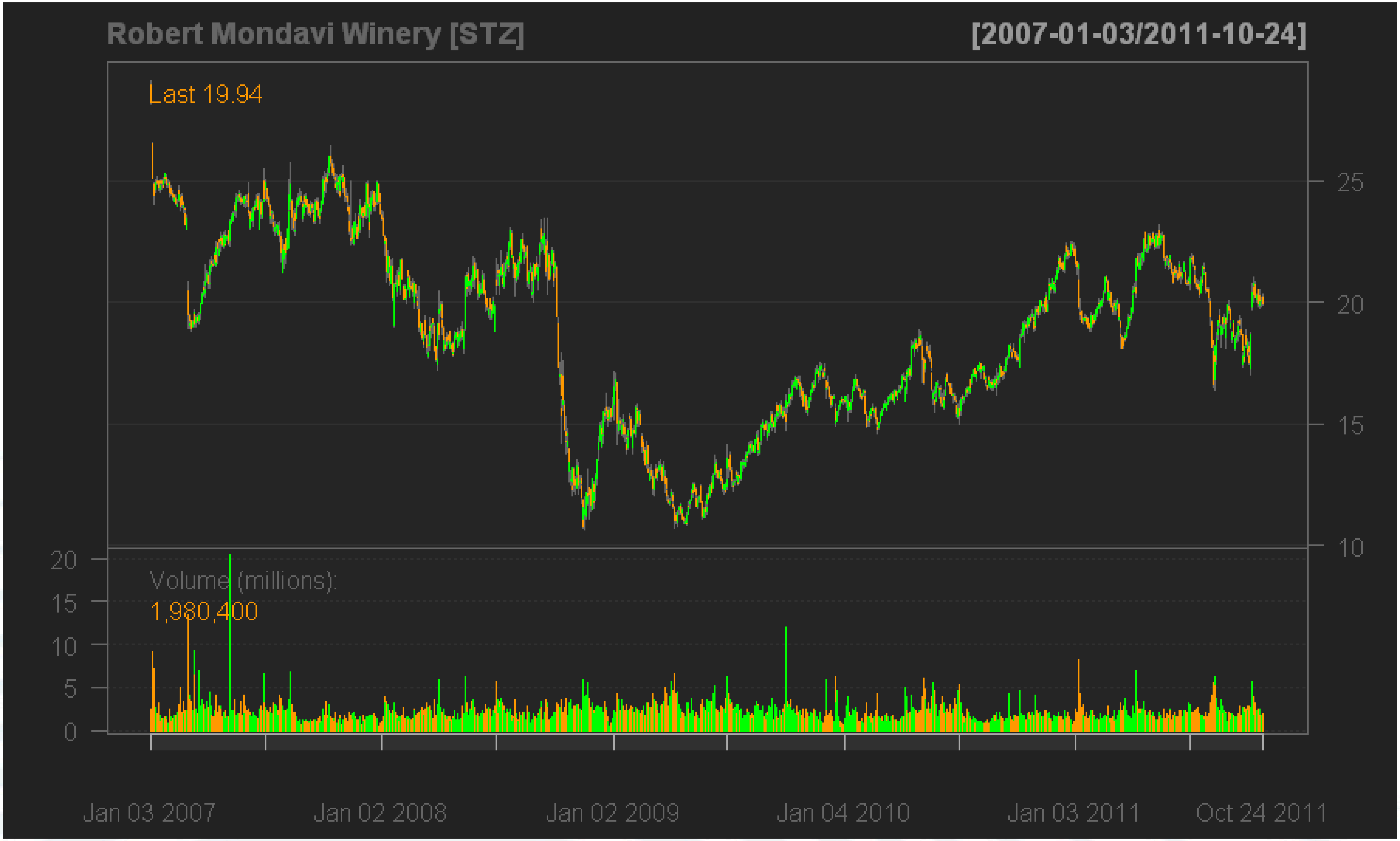
Home List View Thumbnails View

http://paulbutler.org/archives/visualizing-facebook-friends/

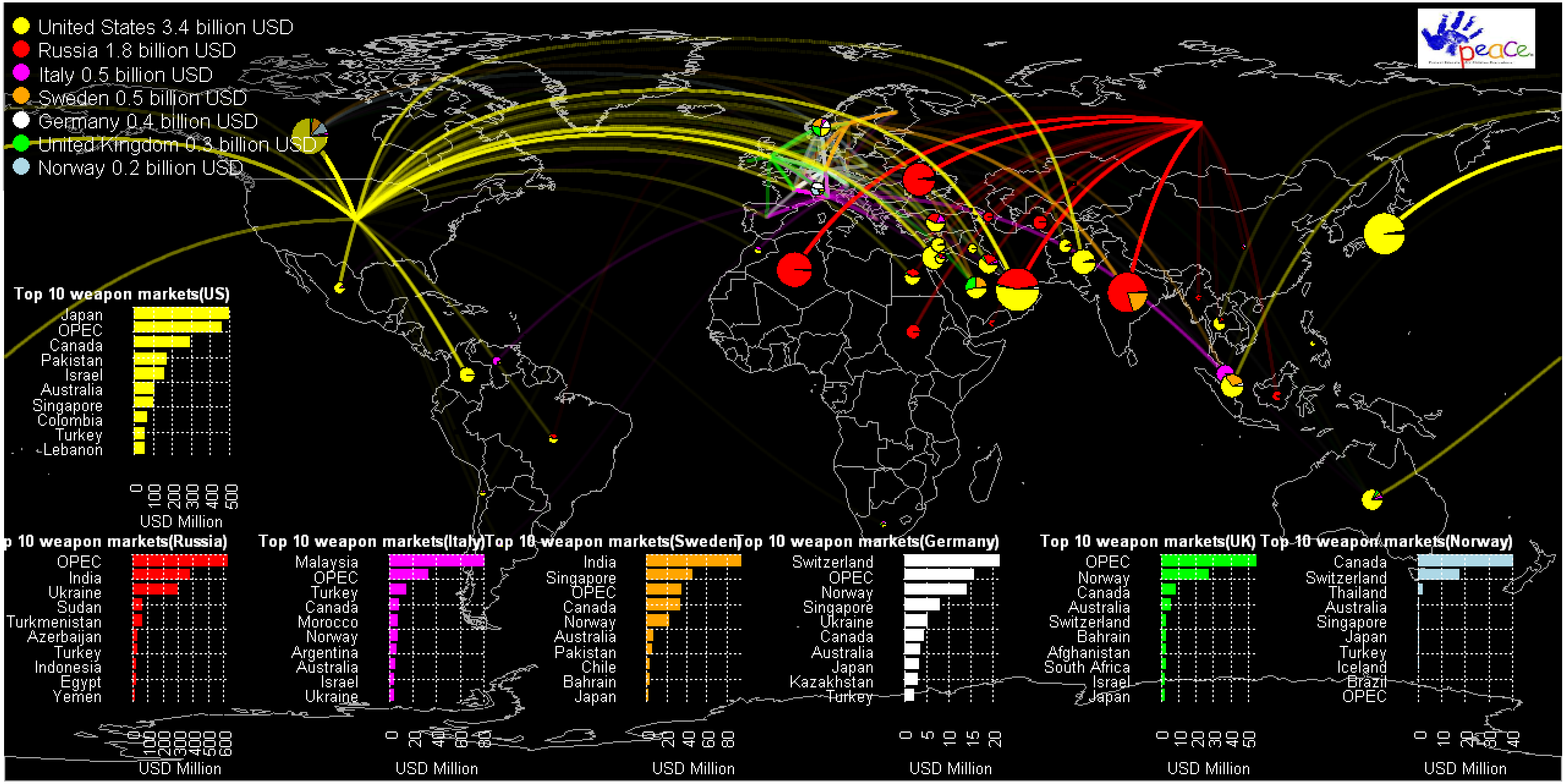


"…aside from the addition of the logo and date text, the image was produced entirely with about 150 lines of R code with no external dependencies."

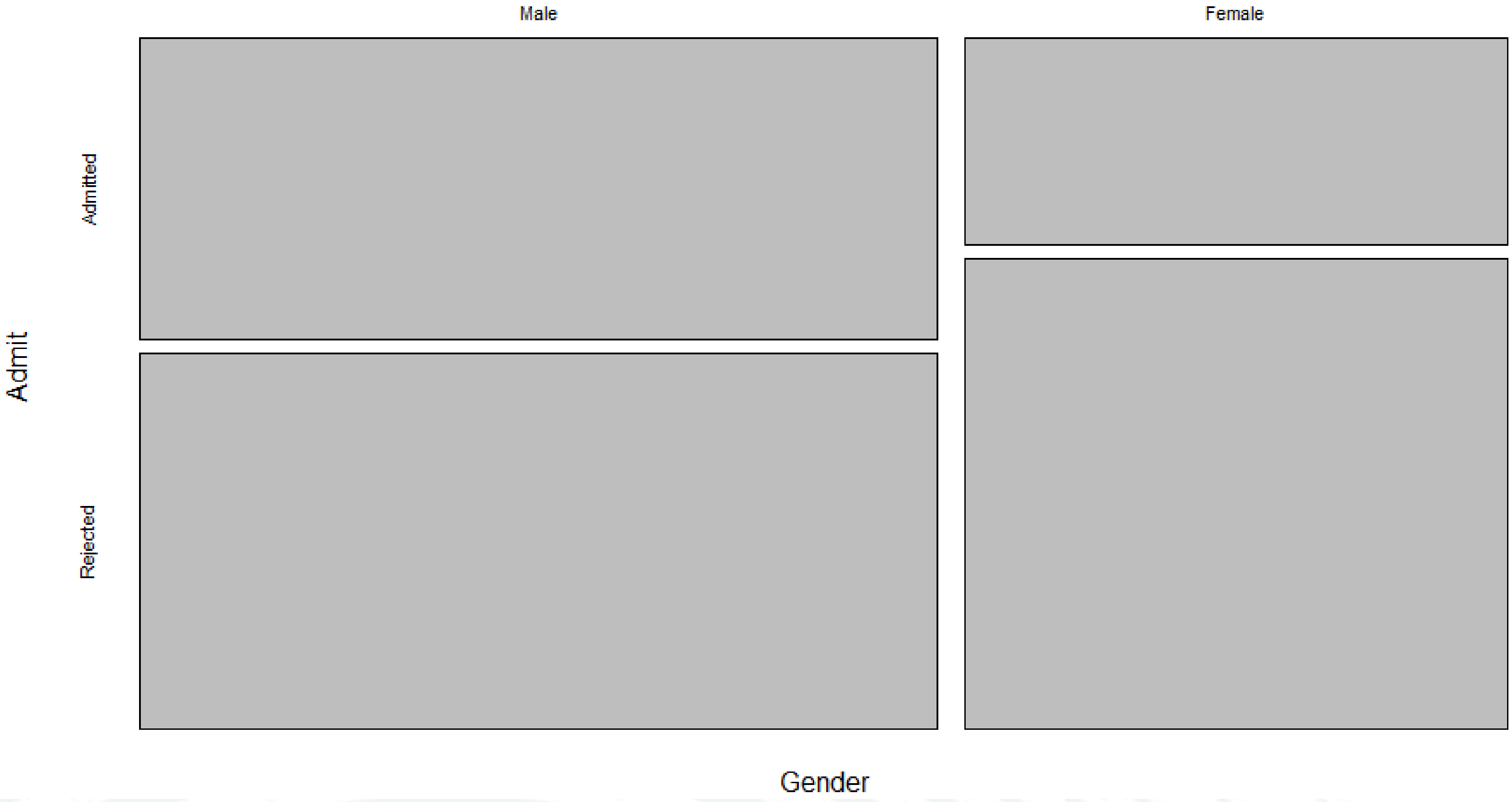
Created using the “quantmod” package



International Weapon Exports in 2010, from http://www.intracen.org/exporters/Stat_export_country_product/



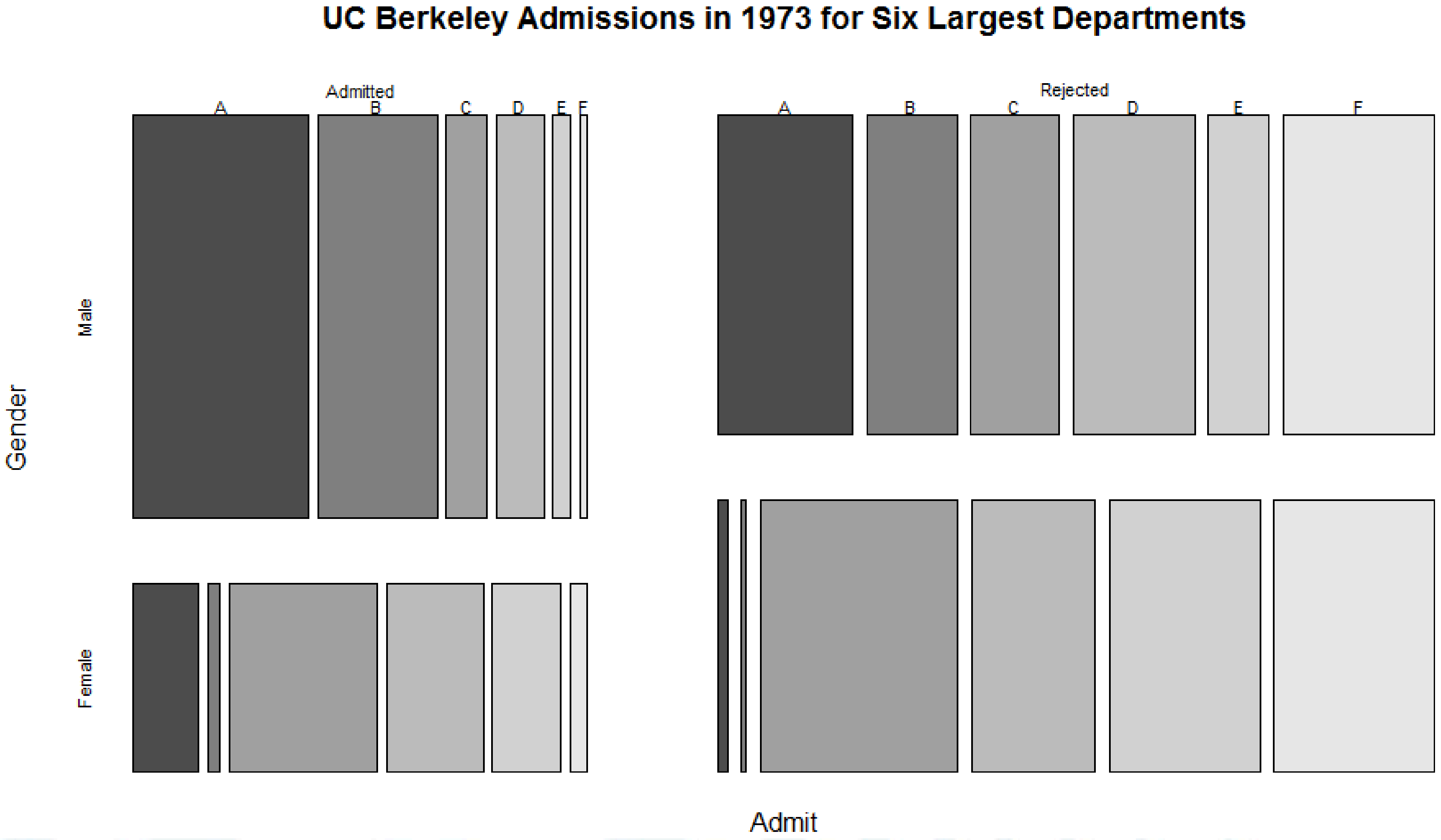
UC Berkeley Admissions in 1973 - Admits by Gender



Function call:

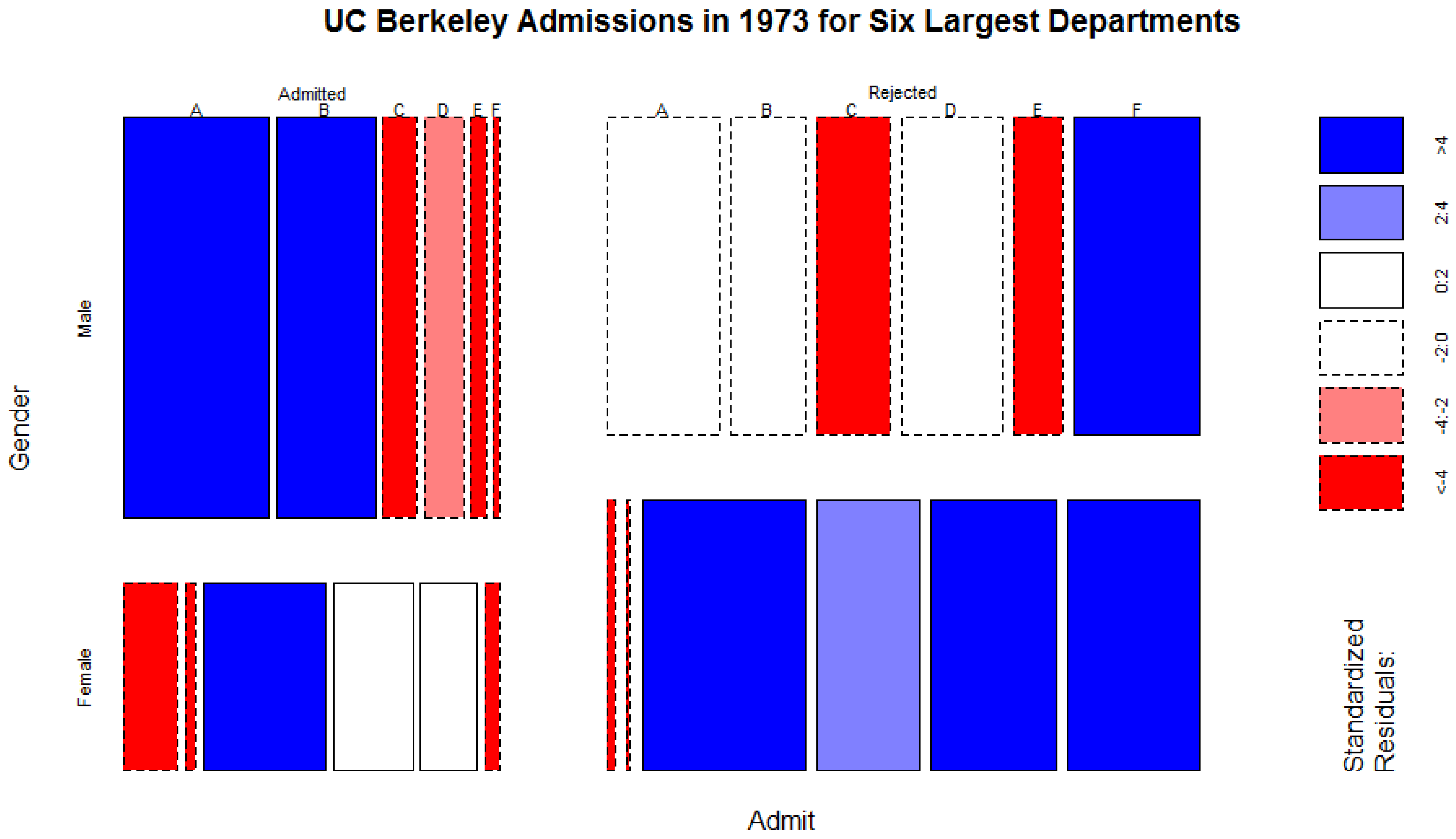
```
mosaicplot(~ Gender + Admit, colour=TRUE, data=UCBAdmissions,
           main="UC Berkeley Admissions in 1973 - Admits by Gender")
```

P.J. Bickel, E.A. Hammel and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data From Berkeley". *Science* **187** (4175): 398–404. doi:10.1126/science.187.4175.398



```
Function call:
mosaicplot(UCBAdmissions, colour=TRUE, main="UC Berkeley Admissions in 1973 for Six Largest Departments")
```

P.J. Bickel, E.A. Hammel and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data From Berkeley". *Science* **187** (4175): 398–404. doi:10.1126/science.187.4175.398



```
mosaicplot(UCBAdmissions, shade=TRUE, main="UC Berkeley Admissions in 1973 for Six Largest Departments")
```

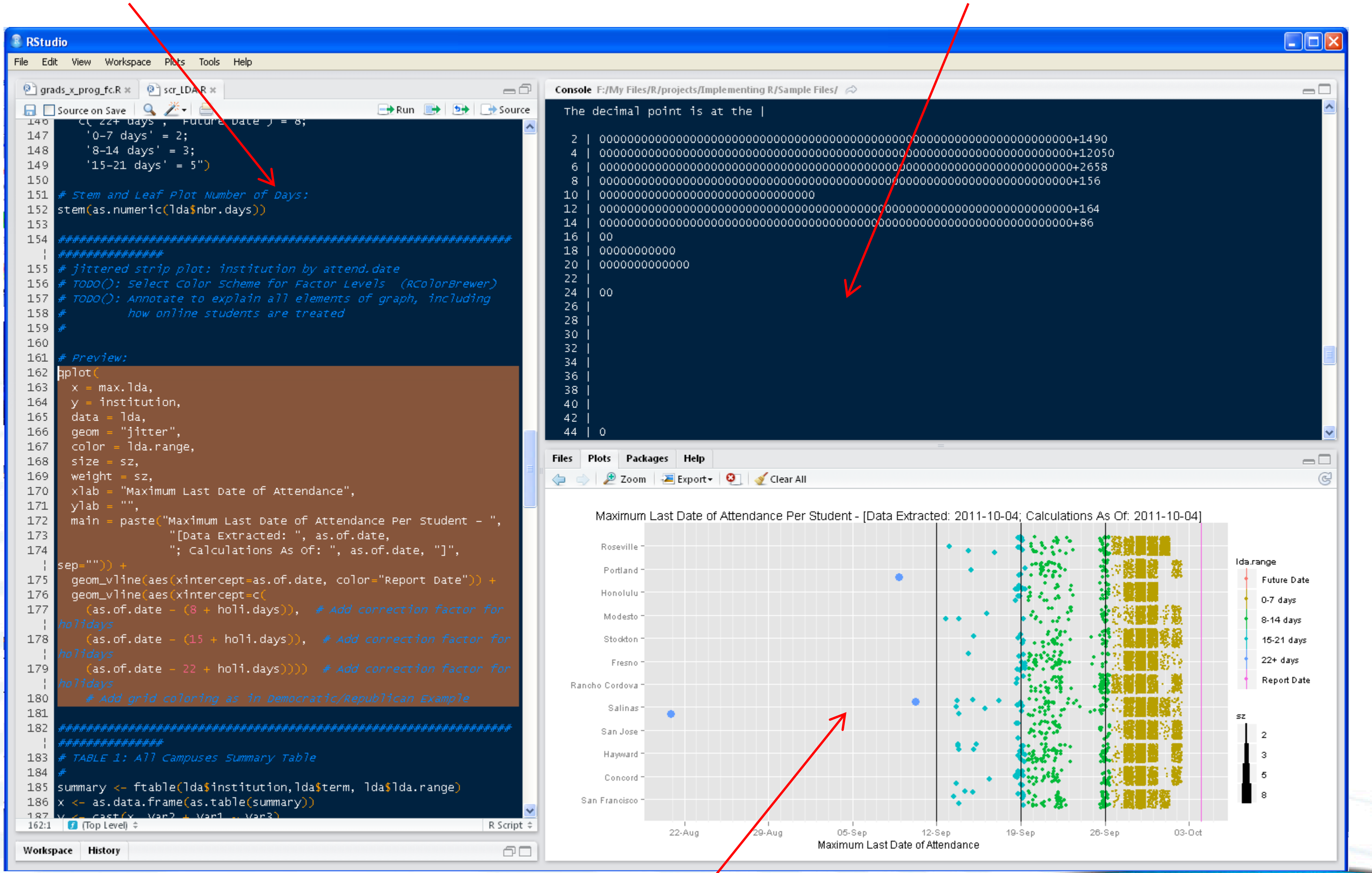
P.J. Bickel, E.A. Hammel and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data From Berkeley". *Science* **187** (4175): 398–404. doi:10.1126/science.187.4175.398

R User Interface

Overview of R (via R Studio)

Script Pane

Console [text output, scratch work]



RStudio
File Edit View Workspace Plots Tools Help
grads_x_prog_fc.R
scr_LDA.R
Source on Save
Run
Source
162:1
(Top Level)
R Script
Workspace
History
Console F:/My Files/R/projects/Implementing R/Sample Files/
Files Plots Packages Help
Zoom
Export
Clear All

```
   c('22+ days', 'Future Date') = 8;
    '0-7 days' = 2;
    '8-14 days' = 3;
    '15-21 days' = 5")

# Stem and Leaf Plot Number of Days:
stem(as.numeric(lda$nbr.days))

###############################################################################
# jittered strip plot: institution by attend.date
# TODO(): Select Color Scheme for Factor Levels  (RColorBrewer)
# TODO(): Annotate to explain all elements of graph, including
#          how online students are treated
#

# Preview:
qplot(
  x = max.lda,
  y = institution,
  data = lda,
  geom = "jitter",
  color = lda.range,
  size = sz,
  weight = sz,
  xlab = "Maximum Last Date of Attendance",
  ylab = "",
  main = paste("Maximum Last Date of Attendance Per Student - ",
               "[Data Extracted: ", as.of.date,
               "; Calculations As Of: ", as.of.date, "]", sep="")) +
  geom_vline(aes(xintercept=as.of.date, color="Report Date")) +
  geom_vline(aes(xintercept=c(
    (as.of.date - (8 + holi.days)),  # Add correction factor for holidays
    (as.of.date - (15 + holi.days)),  # Add correction factor for holidays
    (as.of.date - 22 + holi.days)))  # Add correction factor for holidays
    # Add grid coloring as in Democratic/Republican Example

###############################################################################
# TABLE 1: All Campuses Summary Table
#
summary <- ftable(lda$institution,lda$term, lda$lda.range)
x <- as.data.frame(as.table(summary))
y <- cast(x, Var2 + Var1 ~ Var3)
```

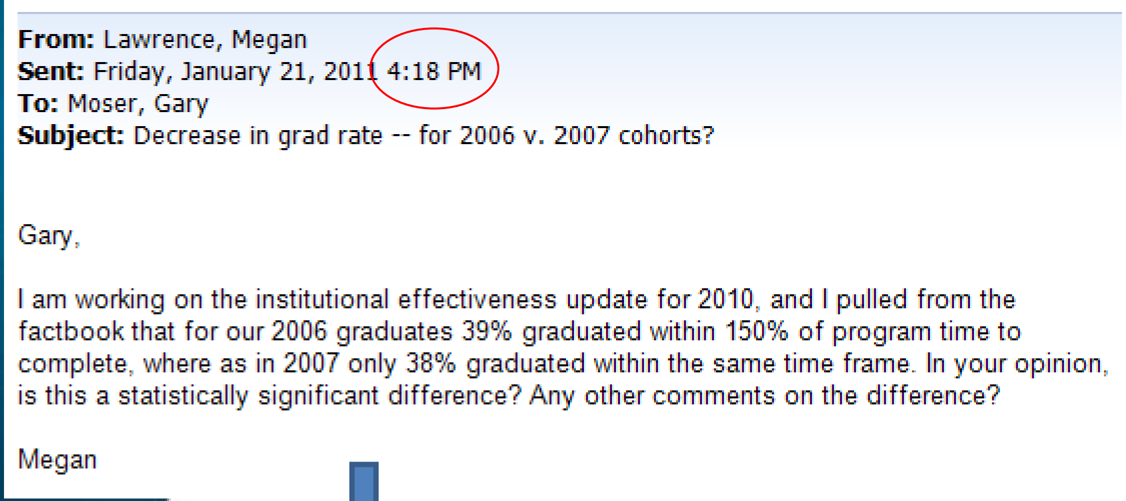
```
 The decimal point is at the |

  2 | 000000000000000000000000000000000000000000000000000000000000000000000+1490
  4 | 000000000000000000000000000000000000000000000000000000000000000000000+12050
  6 | 000000000000000000000000000000000000000000000000000000000000000000000+2658
  8 | 000000000000000000000000000000000000000000000000000000000000000000000+156
 10 | 0000000000000000000000000000000
 12 | 000000000000000000000000000000000000000000000000000000000000000000000+164
 14 | 000000000000000000000000000000000000000000000000000000000000000000000+86
 16 | 00
 18 | 00000000000
 20 | 0000000000000
 22 |
 24 | 00
 26 |
 28 |
 30 |
 32 |
 34 |
 36 |
 38 |
 40 |
 42 |
 44 | 0
```

Graphics/Packages/Help Pane

Applications to IR

Example:

Determining Whether Change in Graduation Rate is Statistically Significant



From: Lawrence, Megan
Sent: Friday, January 21, 2011 4:18 PM
To: Moser, Gary
Subject: Decrease in grad rate -- for 2006 v. 2007 cohorts?

Gary,

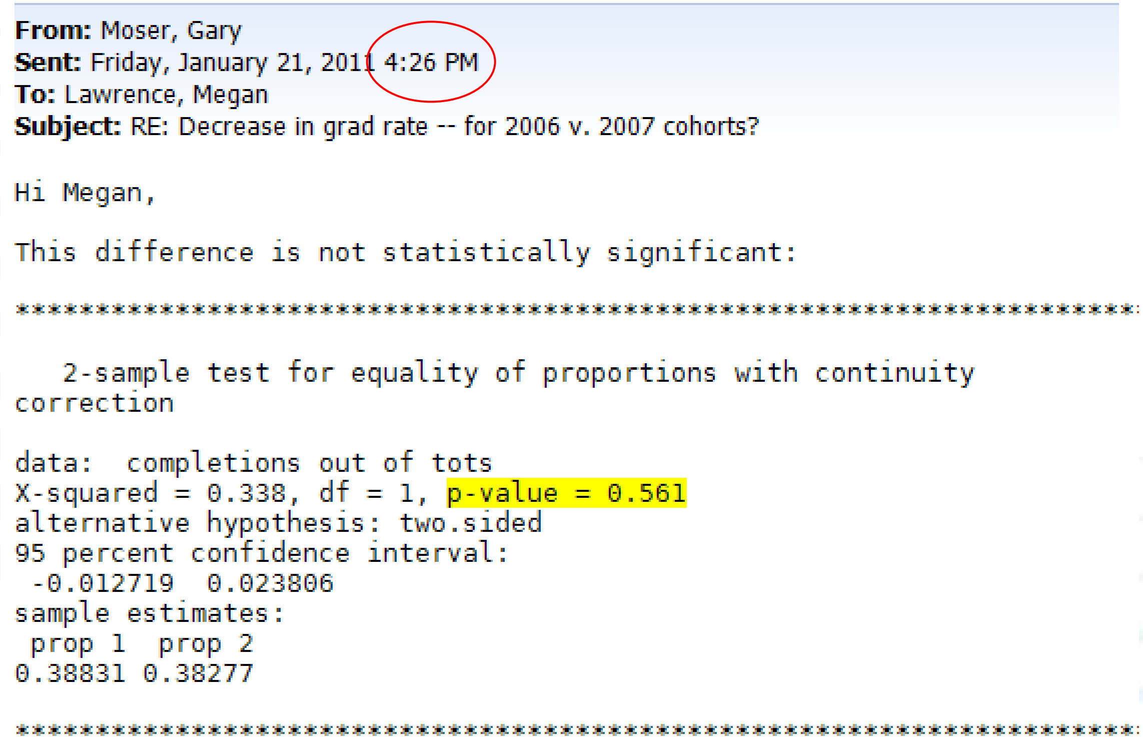
I am working on the institutional effectiveness update for 2010, and I pulled from the factbook that for our 2006 graduates 39% graduated within 150% of program time to complete, where as in 2007 only 38% graduated within the same time frame. In your opinion, is this a statistically significant difference? Any other comments on the difference?

Megan

…from the Factbook, number of grads within 150% of program length for F06 and F07 [2,206, 2088] and total cohort counts [5681, 5455]

Function Call:

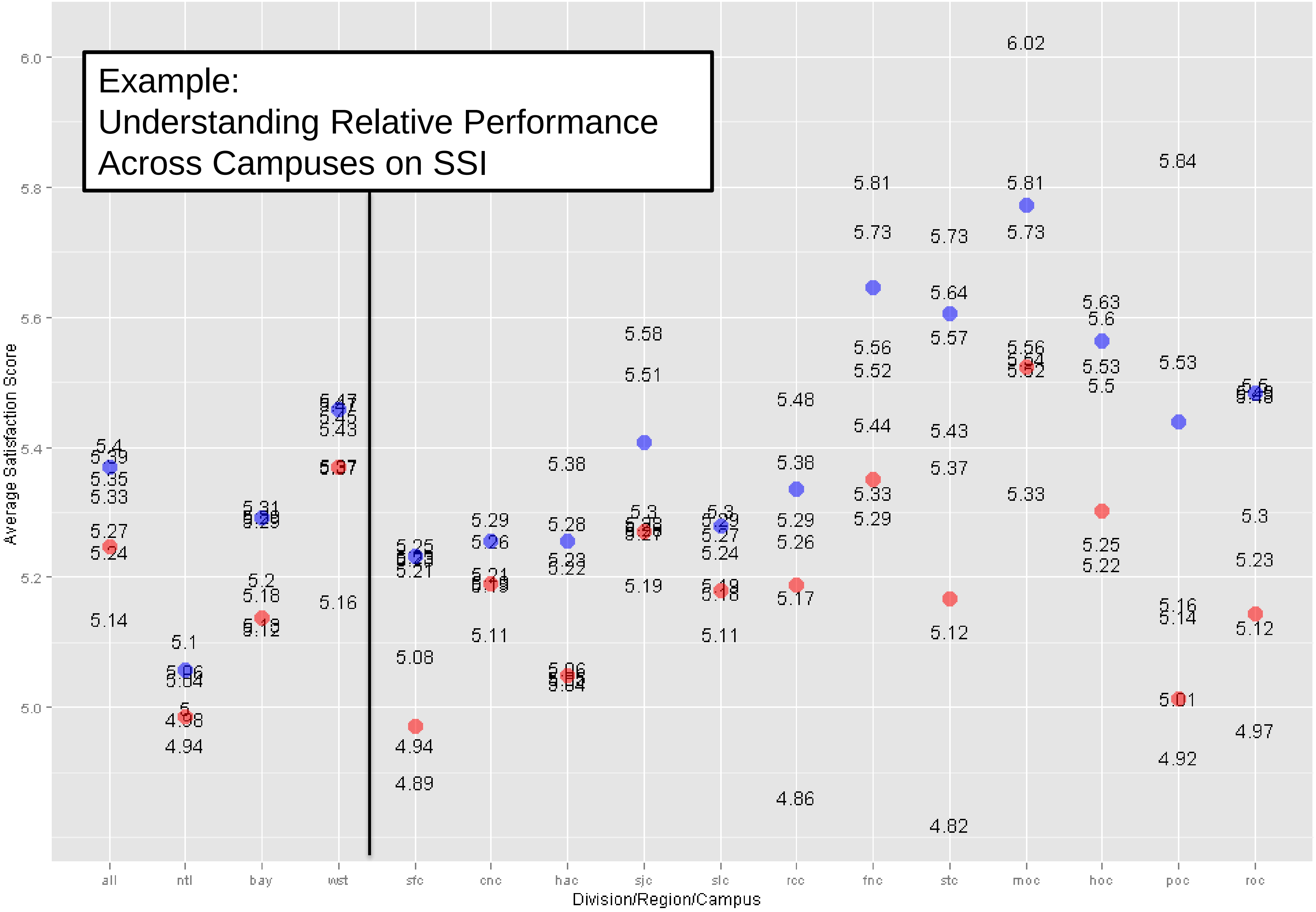
```
prop.test(
  x=c(2206, 2088),
  n=c(5681, 5455))
```



From: Moser, Gary
Sent: Friday, January 21, 2011 4:26 PM
To: Lawrence, Megan
Subject: RE: Decrease in grad rate -- for 2006 v. 2007 cohorts?

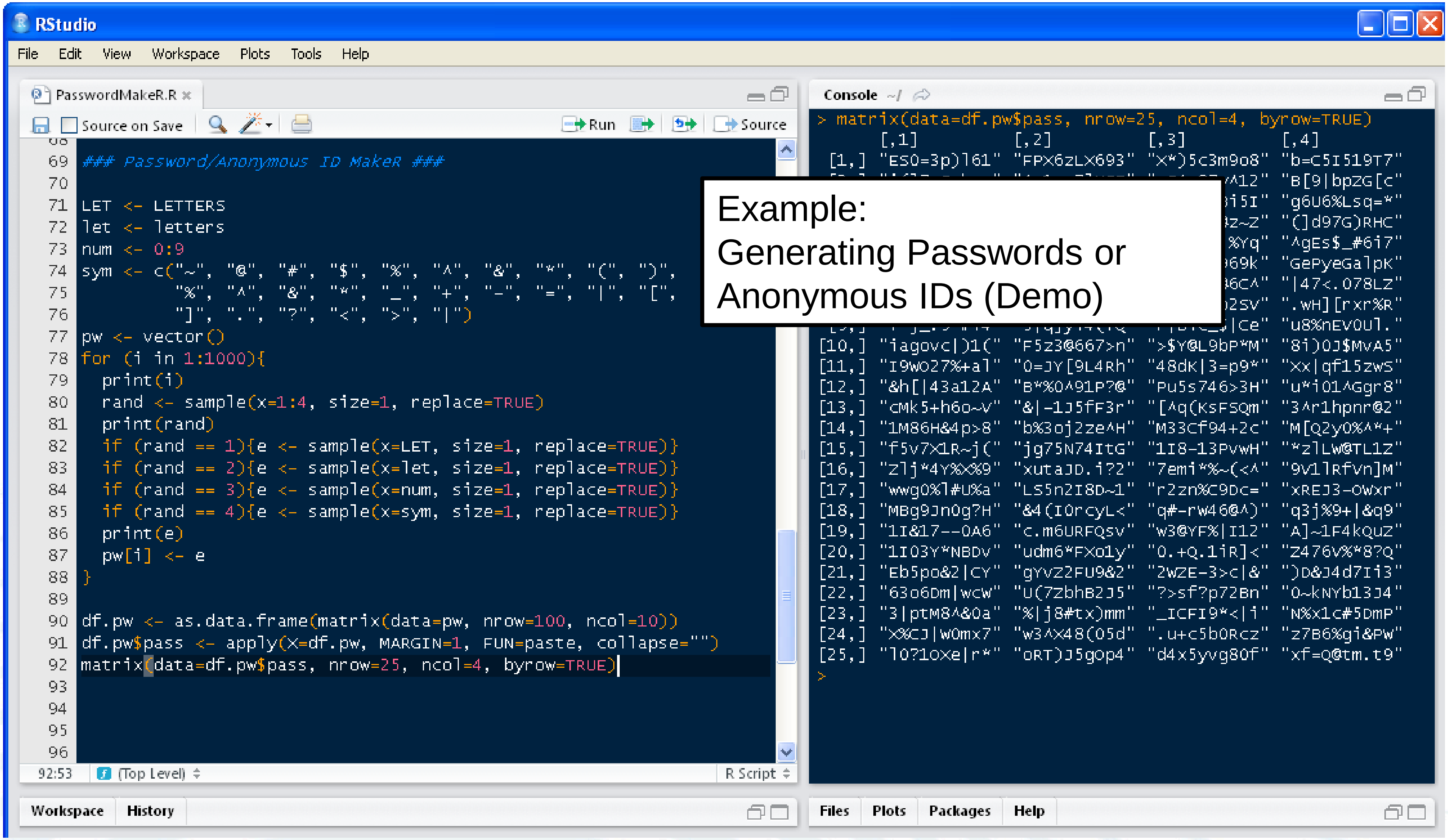
```
Hi Megan,

This difference is not statistically significant:

*******************************************************************************

   2-sample test for equality of proportions with continuity
correction

data:  completions out of tots
X-squared = 0.338, df = 1, p-value = 0.561
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.012719  0.023806
sample estimates:
 prop 1   prop 2
0.38831 0.38277

*******************************************************************************
```

SSI Avg Sat Scores by Campus - 2005 to 2011
Blue Indicator at 75th Percentile
Red Indicator at 20th Percentile



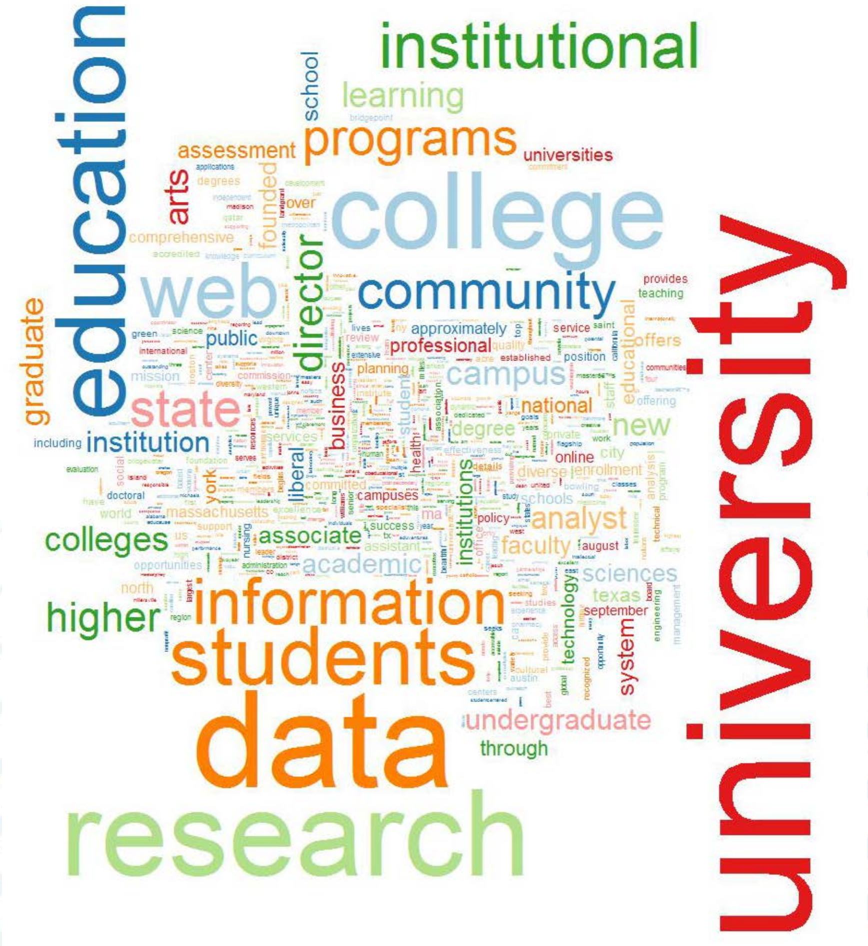
Example:
Understanding Relative Performance
Across Campuses on SSI

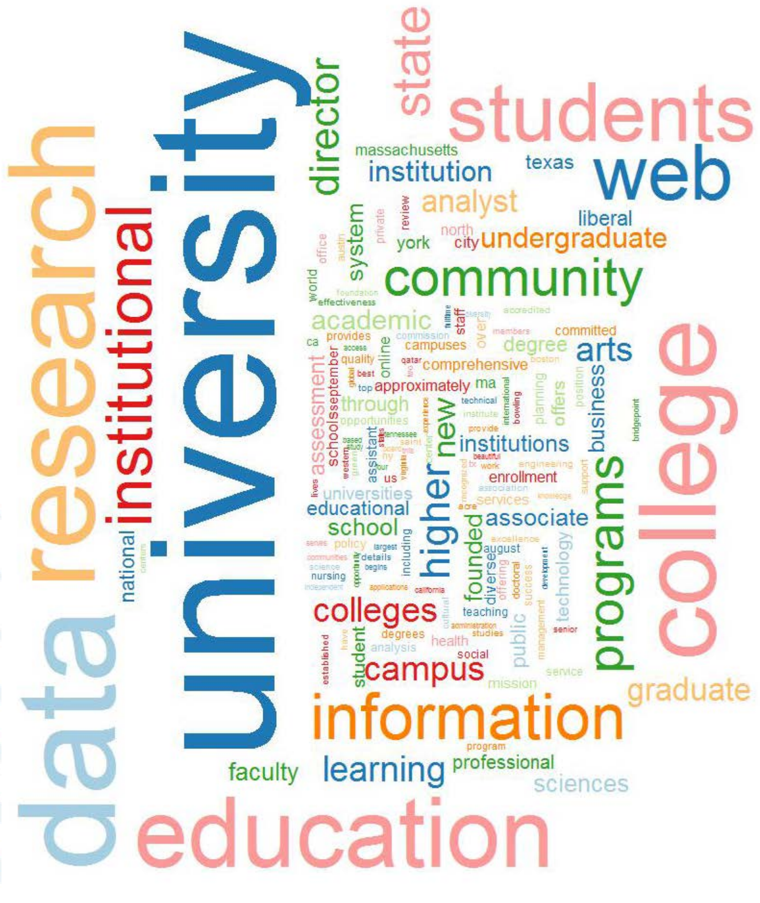
Example: Generating Passwords or Anonymous IDs (Demo)



```r
### Password/Anonymous ID MakeR ###

LET <- LETTERS
let <- letters
num <- 0:9
sym <- c("~", "@", "#", "$", "%", "^", "&", "*", "(", ")",
         "%", "^", "&", "*", "_", "+", "-", "=", "|", "[",
         "]", ".", "?", "<", ">", "|")
pw <- vector()
for (i in 1:1000){
  print(i)
  rand <- sample(x=1:4, size=1, replace=TRUE)
  print(rand)
  if (rand == 1){e <- sample(x=LET, size=1, replace=TRUE)}
  if (rand == 2){e <- sample(x=let, size=1, replace=TRUE)}
  if (rand == 3){e <- sample(x=num, size=1, replace=TRUE)}
  if (rand == 4){e <- sample(x=sym, size=1, replace=TRUE)}
  print(e)
  pw[i] <- e
}

df.pw <- as.data.frame(matrix(data=pw, nrow=100, ncol=10))
df.pw$pass <- apply(X=df.pw, MARGIN=1, FUN=paste, collapse="")
matrix(data=df.pw$pass, nrow=25, ncol=4, byrow=TRUE)
```

```
> matrix(data=df.pw$pass, nrow=25, ncol=4, byrow=TRUE)
      [,1]          [,2]          [,3]          [,4]
 [1,] "ESO=3p)l61"  "FPX6zLX693"  "X*)5c3m9o8"  "b=C5I519T7"
                                       ^12"  "B[9|bpzG[c"
                                       i5I"  "g6U6%Lsq=*"
                                       z~Z"  "(]d97G)RHC"
                                       %Yq"  "^gEs$_#6i7"
                                       69k"  "GePyeGalpK"
                                       6C^"  "|47<.078LZ"
                                       2Sv"  ".wH][rxr%R"
                                      |Ce"  "u8%nEVOUl."
[10,] "iagovc|)1("  "F5z3@667>n"  ">$Y@L9bP*M"  "8i)0J$MvA5"
[11,] "I9WO27%+al"  "O=JY[9L4Rh"  "48dK|3=p9*"  "Xx|qf15zwS"
[12,] "&h[|43a12A"  "B*%0^91P?@"  "Pu5s746>3H"  "u*i01^Ggr8"
[13,] "cMk5+h6o~V"  "&|-1J5fF3r"  "[^q(KsFSQm"  "3^r1hpnr@2"
[14,] "1M86H&4p>8"  "b%3oj2ze^H"  "M33Cf94+2c"  "M[Q2y0%^*+"
[15,] "f5v7X1R~j("  "jg75N74ItG"  "1I8-13PvwH"  "*zlLW@TL1Z"
[16,] "Zlj*4Y%X%9"  "xutaJD.i?2"  "7emi*%~(<^"  "9V1lRfvn]M"
[17,] "wwg0%l#U%a"  "LS5n2I8D~1"  "r2zn%C9Dc="  "xREJ3-OWxr"
[18,] "MBg9Jn0g?H"  "&4(I0rcyL<"  "q#-rW46@^)"  "q3j%9+|&q9"
[19,] "1I&17--0A6"  "c.m6URFQsv"  "w3@YF%|I12"  "A]~1F4kQuZ"
[20,] "1I03Y*NBDv"  "udm6*FXo1y"  "0.+Q.1iR]<"  "Z476V%*8?Q"
[21,] "Eb5po&2|CY"  "gYvZ2FU9&2"  "2WZE-3>c|&"  ")D&J4d7Ii3"
[22,] "63o6Dm|wcW"  "U(7ZbhB2J5"  "?>sf?p72Bn"  "0~kNYb13J4"
[23,] "3|ptM8^&0a"  "%|j8#tx)mm"  "_ICFI9*<|i"  "N%x1c#5DmP"
[24,] "X%CJ|W0mx7"  "w3^X48(05d"  ".u+c5b0Rcz"  "z7B6%gi&PW"
[25,] "l0?1oXe|r*"  "oRT)J5gOp4"  "d4x5yvg80f"  "xf=Q@tm.t9"
>
```





Heald
COLLEGE
EST. 1863

Essential Books:

1.) R for SAS and SPSS Users (Robert Muenchen)

2.) Introductory Statistics with R (Peter Dalgaard)

3.) Data Manipulation with R (Phil Spector)

4.) ggplot2 (Hadley Wickham)

Questions / Comments