Responsys®

A Gentle Introduction to R

and its Applications in Business Intelligence

Michael Driscoll, Principal, Dataspora, Inc.
Jim Porzak, Sr. Director of Analytics, Responsys, Inc.

SDForum, Business Intelligence SIG
15 July, 2008. Palo Alto, California

Outline

Responsys

- Evolution of R
- R as a BI Tool
- Jim’s Case Studies
- Mike’s Case Studies
- Getting Started with R
- Wrap


R

Pronouncing “R”

Responsys®

First there was S

Responsys

- R is the free (GNU), open source, version of S
  - S developed by John Chambers *et al* while at Bell Labs in 80's
  - For "data analysis and graphics" (with statistics emphasis)
  - Ver.4 defined by the "Green Book" *Programming with Data*, 1998
  - "S-Plus" now owned by Insightful Corp., Seattle, WA
- R was initially written in early 1990's
  - by *Robert* Gentleman and *Ross* Ihaka
  - Statistics Department of the University of Auckland
  - GNU GPL release in 1995
  - "R" is before "S", as in "HAL" is before "IBM"
- Since 1997 a core group of ± 20 developers
  - Initial V1.0 released in February, 2000
  - Continually developed with a new 0.1 level release ~ 6 months

Current State of R

Responsys

As of October 2004

- V2.0 Released October, 2004
- Windows, Mac OS, Linux & Unix ports
- Over 400 submitted packages from "abind" to "zoo"
- 12th newsletter (Volume 4/2) published September 2004
- The first useR! – R User Conference held in Vienna May 2004
- ~400 R-help messages per week
- ~ Dozen texts specifically on R or with R examples and code
- R language generally accepted to be more powerful than S-Plus
- Some interesting GUI work in progress

As of July 2008

- V2.7.1 Released June, 2008
- Windows, Mac OS, Linux & Unix ports (including Vista)
- 1450+ packages; "aaMI" to "zoo" (+45 Omegahat, +260 Bioconductor)
- 22nd newsletter (Volume 8/1) published May 2008
- The fourth useR! conference next month in Dortmund, Germany
- ~700 R-help messages per week
- 65 texts specifically on R or with R examples and code
- R ~ universally taught.
- Commercial support (REvolution, …)
- JGR, Rattle, RCmdr, …
- *Web based applications now easier*

www.r-project.org

Responsys

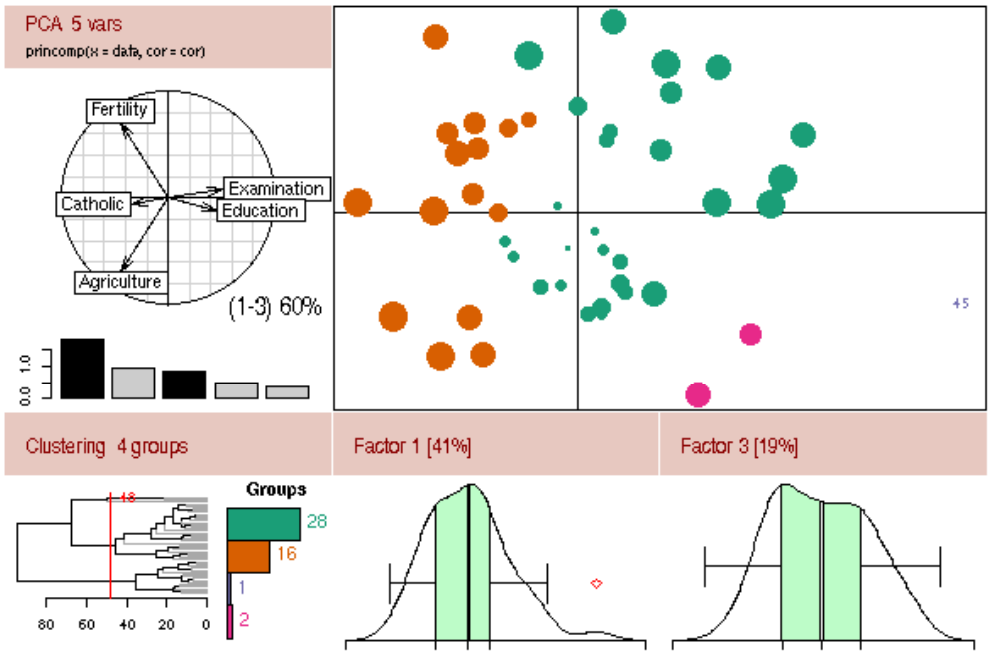
About R
- What is R?
- Contributors
- Screenshots
- What's new?

Download
- CRAN

R Project
- Foundation
- Members & Donors
- Mailing Lists
- Bug Tracking
- Developer Page
- Conferences
- Search

Documentation
- Manuals
- FAQs
- Newsletter
- Wiki
- Books
- Certification

The R Project for Statistical Computing



Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed
Newsletter

The Comprehensive R Archive Network

Frequently used pages

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Linux
- MacOS X
- Windows

Source Code for all Platforms

Windows and Mac users most likely want the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- **The latest release** (2008-06-23): R-2.7.1.tar.gz (read what's new in the latest version).
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

Task Views 1st step in finding relevant packages

Responsys

cran.cnr.berkeley.edu/web/views/

CRAN

Mirrors
What's new?
Task Views
Search

About R

R Homepage

Software

R Sources
R Binaries
Packages
Other

Documentation

Manuals
FAQs
Contributed
Newsletter

CRAN Task Views

Bayesian	Bayesian Inference
Cluster	Cluster Analysis & Finite Mixture Models
Econometrics	Computational Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
gR	gRaphical Models in R
MachineLearning	Machine Learning & Statistical Learning
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Psychometrics	Psychometric Models and Methods
Robust	Robust Statistical Methods
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
Survival	Survival Analysis

R Help – Best Support!

Responsys

Remove label "R-Help" | Report Spam | Delete | More actions... | Refresh | 1 - 50 of 10513 Older › Oldest »

Select: All, None, Read, Unread, Starred, Unstarred

march	Inbox **[R] gbn with jumps** - Hi everybody I'd like to simulate a Generalized Wiener Process with jumps. Any sugge	7:03 am
Vladimir Eremeev	Inbox **[R] simpler solution (untested)** - axis says that this function has the logical parameter outer "indicating whe	6:56 am
march	Inbox **[R] gbm with jumps** - Hi everybody I'd like to simulate a Generalized Wiener Process with jumps. Any sugge	6:48 am
Rafael, **Peter**, **Vladimir** (3)	Inbox **[R] Three horizontal axes OR Two axes on same side?** - Dear list: I need to reproduce a plot with three d	6:44 am
Bram Kuijper	Inbox **[R] levelplot not adjusting colors** - Hi all, I try to make a levelplot from the Trellis graphics package of count	6:41 am
Marta Rufino	Inbox **[R] warning in GAM** - Hello, I have a problem when doing gam (from gam library; I am using R 2.4.0, windows	5:48 am
Antje, **Peter** (4)	Inbox **[R] Error in plot.new() : Figure margins too large** - Hello, was could be the reason for such an error messa	5:33 am
Indermaur, **Ken**, **Prof** (3)	Inbox **[R] batch job GLM calculations** - Hello I want to batch job the calculation of many GLM-models, extract sor	1:19 am
Adrian .. **Prof**, **Adrian** (9)	Inbox **[R] a question of substitute** - The 'Right Thing' is for oneway.test() to allow a variable for the first argument, a	12:42 am
David, **Marc** (2)	Inbox **[R] zero margin / marginless plots** - Hi, I'd like to produce a marginless or zero margin plot so that the pixe	7:37 pm
Walter, **Torsten**, **Richard** (3)	Inbox **[R] posthoc tests with ANCOVA** - The WoodEnergy example in package HH (available on CRAN) is similar #	Jan 10
karl.sommer	Inbox **[R] axis date format in lattice** - Hello list, plotting the following example 1 in lattice only labels the x-axis wi	Jan 10
Tong .. Prof, **François** (9)	Inbox **[R] A question about R environment** - Philippe Grosjean] >Please, don't reinvent the wheel: putting functior	Jan 10
Michael, **Peter** (2)	Inbox **[R] TCL/TK and R documentation?** - I am hoping something has changed since I last asked about this. Is tl	Jan 10
Simon, Setzer.Wood.., **Ken** (3)	Inbox **[R] problems with optim, "for"-loops and machine precision** - Two possibilities for why your 7 parameter	Jan 10
Darren Weber	Inbox **[R] axis labels at subset of tick marks** - For example, this works: x = seq(-100, 1000, 25) y = x * x plot(x,y,	Jan 10
Colleen.Ross .. **Thomas** (3)	Inbox [R] SAS and R code hazard ratios - On Wed, 10 Jan 2007, Colleen.Ross@kp.org wrote: > I am new to R and	Jan 10
Thomas, **Duncan**, Peter (3)	Inbox [R] "go" or "goto" command - Thomas L Jones wrote: > Some computer languages, including C, have a "go" c	Jan 10
Feng, David, **Feng** (3)	Inbox **[R] logistic regression packages** - Hi David: Thanks for you information. 2 further questions: 1. I found out th	Jan 10
David	Inbox [R] Installation problem with package mixtools - I am trying to install mixtools on Debian Etch and get the follc	Jan 10
Tord, **Roger** (2)	Inbox [R] map data.frame() data after having linked them to a read.shape() object - On Wed, 10 Jan 2007, Tord Snäl	Jan 10
Stephen, chao (2)	Inbox [R] using DBI - The way MySQL works, I use RMySQL to contact, which in turn uses DBI. There is a library R	Jan 10
Paul Mathews	Inbox [R] Meeting announcement: An Introduction to Data Analysis Using R - An Introduction to Data Analysis Usin	Jan 10
Kati, roger (2)	Inbox [R] 2 problems with latex.table (quantreg package) - reproducible - The usual R-help etiquette recommends: 1	Jan 10
John .. Jeffrey, Brian (12)	Inbox [R] scripts with littler - Brian Ripley wrote: > Exactly as documented. The argument is named 'new' and not ...	Jan 10
Jenny, Zoltan (3)	Inbox [R] correlation value and map - Hi Zoltan, Right, I have 30x32=960 data points per year (It is actually the mear	Jan 10

Core Developers!

Significant New R Books

Responsys

Springer

Use R

12 results

Applied Econometrics with R
Kleiber, C., Zeileis, A., ISBN 978-0-387-77316-2, 2008, Softcover
... More

Bioconductor Case Studies
Hahne, F., Huber, W. (et al.), ISBN 978-0-387-77239-4, 2008, Softcover
... More

Analysis of Integrated and Co-integrated Time Series with R
Pfaff, B., R-code for examples in the book, ISBN 978-0-387-75966-1, 2008, Softcover
... More

Morphometrics with R
Claude, J., ISBN 978-0-387-77789-4, 2008, Softcover
... More

Applied Spatial Data Analysis with R
Bivand, R.S., Pebesma, E.J. (et al.), ISBN 978-0-387-78170-9, 2008, Softcover
... More

Wavelet Methods in Statistics with R
Nason, G.P., ISBN 978-0-387-75960-9, 2008, Softcover
... More

Statistical Methods for Environmental Epidemiology with R
Peng, R.D., Dominici, F., ISBN 978-0-387-78166-2, 2008, Softcover
... More

Data Manipulation with R
Spector, P., ISBN 978-0-387-74730-9, 2008, Softcover
... More

Lattice · Multivariate Data Visualization with R
Sarkar, D., ISBN 978-0-387-75968-5, 2008, Softcover
... More

Interactive and Dynamic Graphics for Data Analysis
Cook, D., Swayne, D.F., ISBN 978-0-387-71761-6, 2007, Softcover
... More

Bayesian Computation with R
Albert, J., ISBN 978-0-387-71384-7, 2007, Softcover
... More

Analysis of Phylogenetics and Evolution with R
Paradis, E., ISBN 978-0-387-32914-7, 2006, Softcover
... More

Software for Data Analysis
Programming with R
Series: Statistics and Computing
Chambers, John M.
2008, Approx. 510 p., Hardcover
ISBN: 978-0-387-75935-7

Not yet published. Available: July 18, 2008

Introductory Statistics with R
Series: Statistics and Computing
Dalgaard, Peter
2nd ed., 2008, XVI, 364 p., Softcover
ISBN: 978-0-387-79053-4

Not yet published. Available: July 25, 2008

useR! 2008

Responsys

www.statistik.uni-dortmund.de/useR-2008/

The R User Conference 2008

August 12-14, Technische Universität Dortmund, Germany

Organizer: Fakultät Statistik, Technische Universität Dortmund
Co-Organizer: Austrian Association for Statistical Computing
Sponsors: R Foundation for Statistical Computing

AASC Austrian Association for Statistical Computing
CRC Press Taylor & Francis Group CHAPMAN & HALL
REvolution computing
Technische Universität Dortmund SFB 475: Reduction of Complexity in Multivariate Data Structures
Google
Springer
technische universität dortmund
mangosolutions data analysis that delivers
Lufthansa A STAR ALLIANCE MEMBER
MERCK & CO., INC. Whitehouse Station, N.J., U.S.A.
WILEY-BLACKWELL
CAMBRIDGE UNIVERSITY PRESS
statistics.com
ZEMENTIS ADAPA Predictive Analytics Edition
StatSoft

Conference

About the Conference
Date & Location
Important Dates
Call for Papers
Download: Logo, Flyer, Poster
Funding
Participants

Program

Conference Program
Invited Lectures
Presentations
Tutorials
Social Program
Program Committee
Online Registration

Dortmund

Accommodation
About Dortmund
Travel information

About the Conference

useR! 2008, the R user conference, takes place at the Fakultät Statistik, Technische Universität Dortmund, Germany from 2008-08-12 to 2008-08-14. Pre-conference tutorials will take place on August 11.
The conference is organized by the Fakultät Statistik, Technische Universität Dortmund and the Austrian Association for Statistical Computing (AASC). It is funded by the R Foundation for Statistical Computing.

Date & Location

August 12-14, 2008 (iCalendar file)

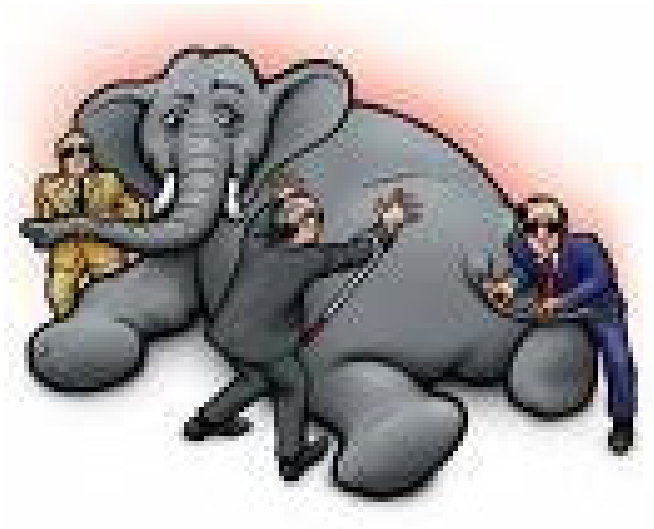
Fakultät Statistik
Technische Universität Dortmund

Responsys

R as a Business Intelligence Tool

So just what is Business Intelligence?

Responsys



From Wikipedia:

In 1989 Howard Dresner, later a Gartner Group analyst, popularized BI as an umbrella term to describe "concepts and methods to improve business decision making by using fact-based support systems."

In modern businesses the use of standards, automation and specialized software, including analytical tools, allows large volumes of data to be extracted, transformed, loaded and warehoused to greatly increase the speed at which information becomes available for decision-making.

Business Intelligence Tools

Responsys

Again From Wikipedia:

The key general categories of business intelligence tools are:

- Spreadsheets
- Reporting and querying software
- OLAP
- Digital Dashboards
- Data mining
- Process mining
- Business performance management

We play here.

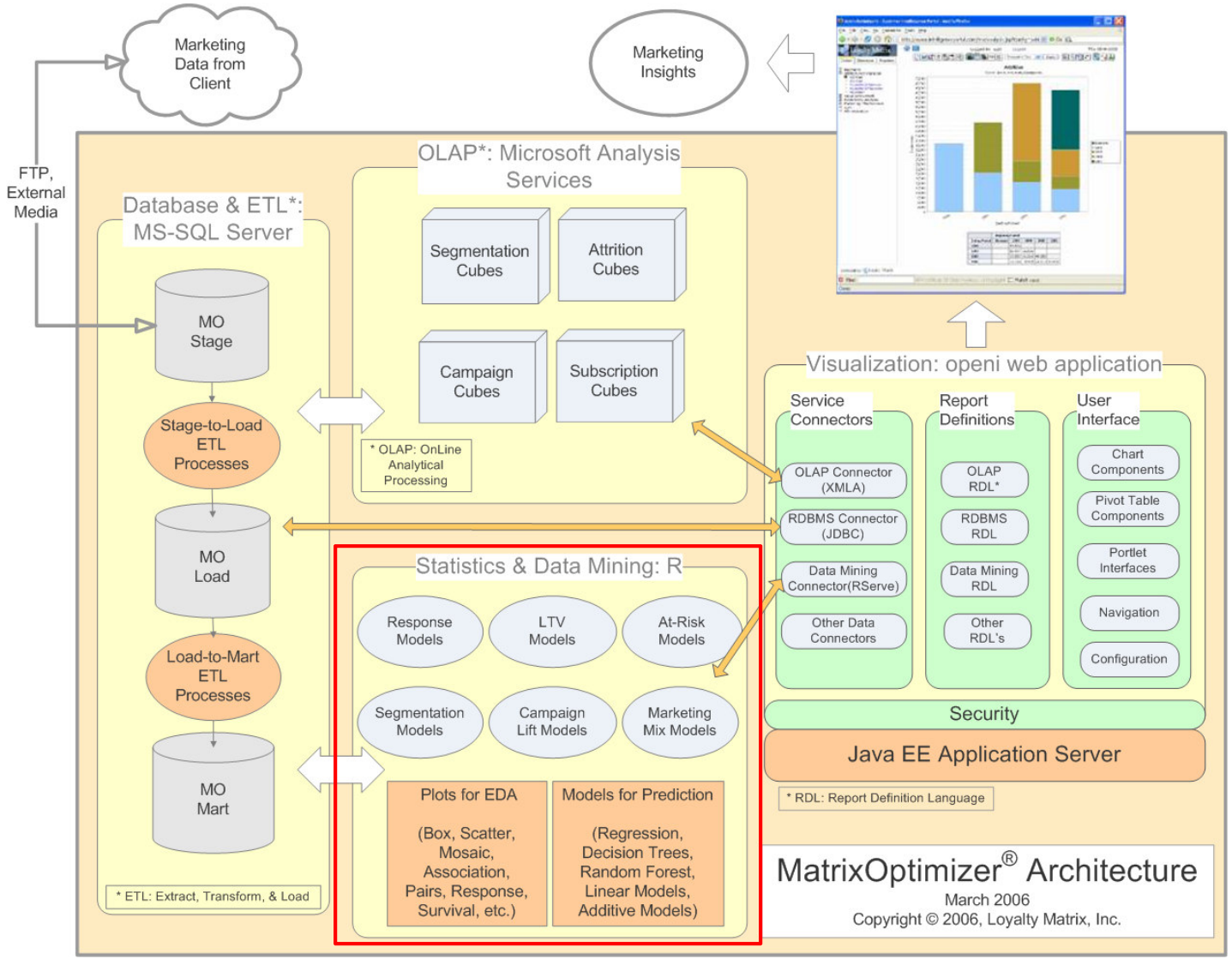
Positioning R against Traditional BI

Responsys®

Characteristic	Traditional BI	R & Friends
Cutting Edge Methods	- / +	+ + +
Naive Interactive Use	+ + +	- - - (some GUIs help)
Reproducible Results	- /+	+ + +
Massive Data Handling	+ +	- - - (stay tuned)
Data Base Reporting	+ + +	- - - (N/A)
Visualization	+ +	+ + +
Predictive Analytics	+	+ + +
Verifiable Methods	-	+ + +
Data Mining	- / ++	+ + +

Leverage R's Strengths in Combination w/ Classical BI

Responsys



Jim's R Examples

- R Help Message Counts
- Data Profiling
- Reproducible Reporting
- Customer Segmentation

Responsys®

R Help Message Count & JGR Demo

R Help Message Counts (JGR Demo)

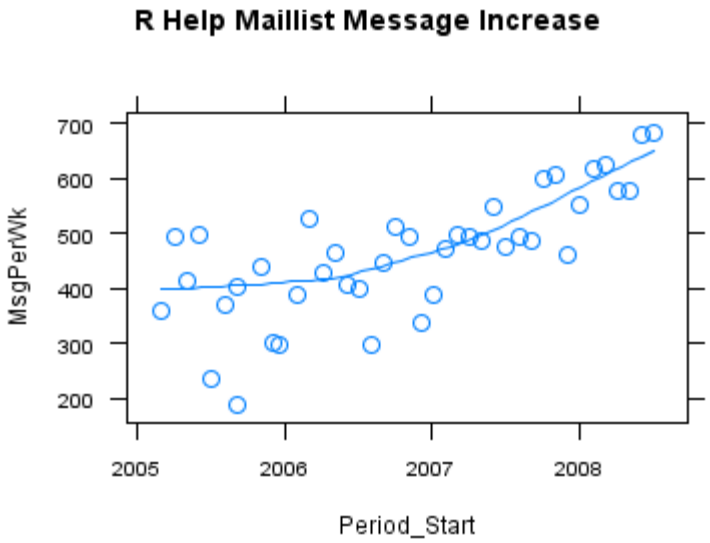
Responsys

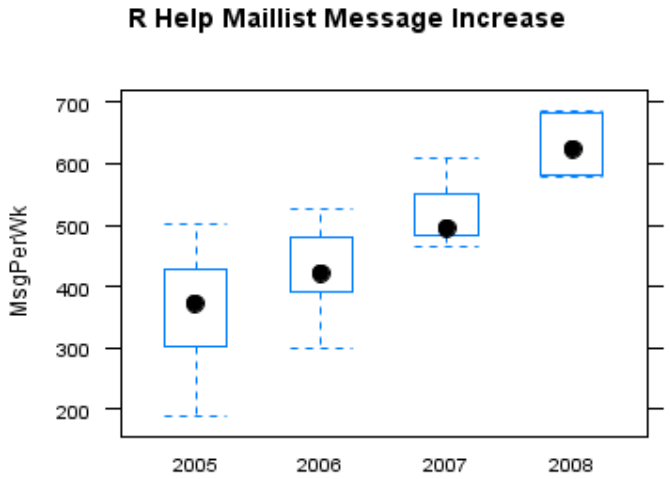
From raw data:

period	*re-sorted*	*messages*
31 Dec 2007 to 31 Jan 2008	by thread by subject by author attachment	2451
31 Jan 2008 to 29 Feb 2008	by thread by subject by author attachment	2565
29 Feb 2008 to 31 Mar 2008	by thread by subject by author attachment	2781
31 Mar 2008 to 30 Apr 2008	by thread by subject by author attachment	2486
30 Apr 2008 to 30 May 2008	by thread by subject by author attachment	2483
1 Jun 2008 to 30 Jun 2008	by thread by subject by author attachment	2824
30 Jun 2008 to 15 Jul 2008	by thread by subject by author attachment	1417

**See Appendix for data & Code*

To insights:





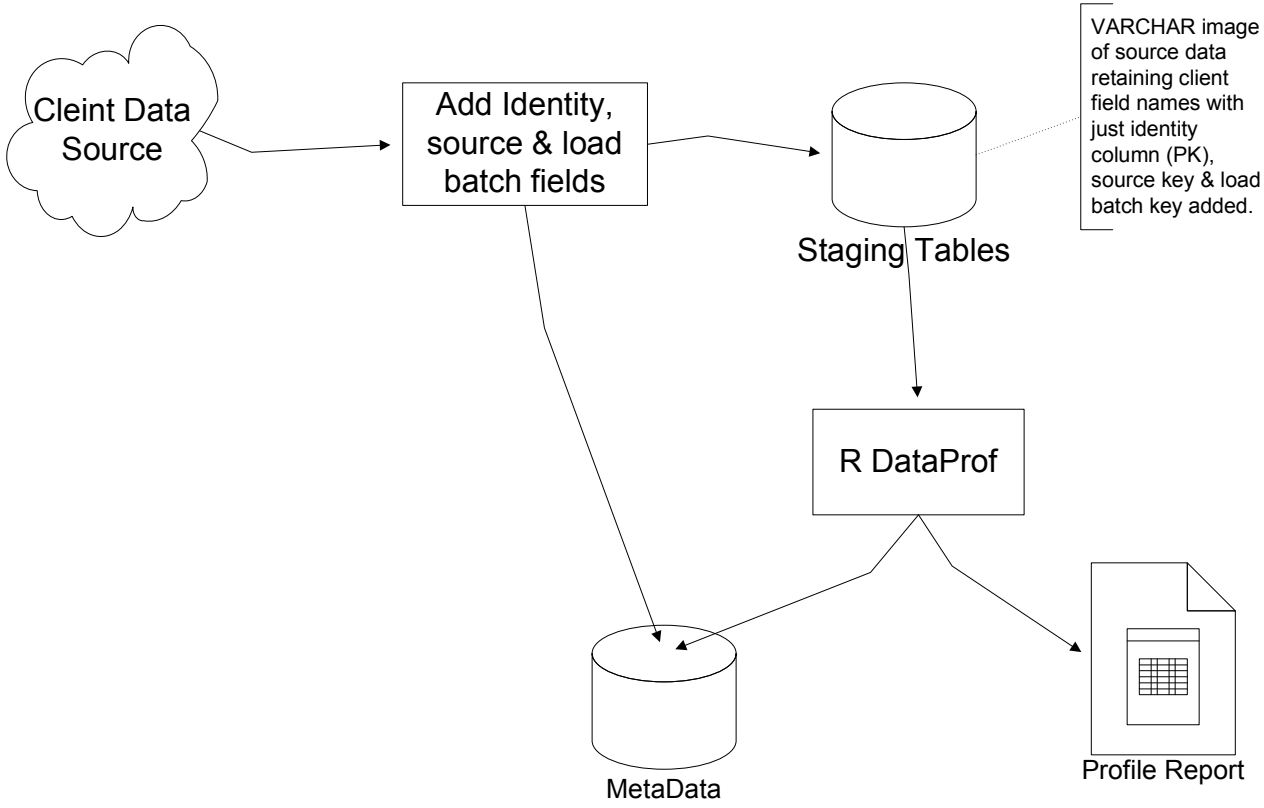
Responsys

Data Profiling with R

Data Profiling in R – Basics

Responsys

- Reference: *Data Quality – The Accuracy Dimension* by Jack E. Olson
- Where we profile:



Data Profiling in R – Profiler Column Output

Responsys

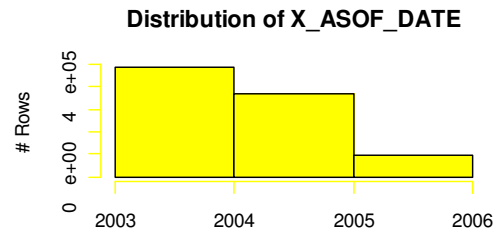
Header: Database details for field

Summary Counts & %'s
Empty, Numeric & Date only for character strings

AMA_Stage . RECEIVABLE_TXN . X_ASOF_DATE 19 varchar(8000)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	3,861,249	2,908,244	28,564	0	0	953,005
%	100.00	75.32	0.74	0.00	0.00	100.00

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2003-01-14	2003-08-12	2003-12-29	2004-02-12	2004-09-08	2005-10-17



Head: NA|NA|NA|NA|NA|NA

Summary Stats if numeric or date

Appropriate plot type

Footer: Notes about field

Data Profiling in R – Examples (1)

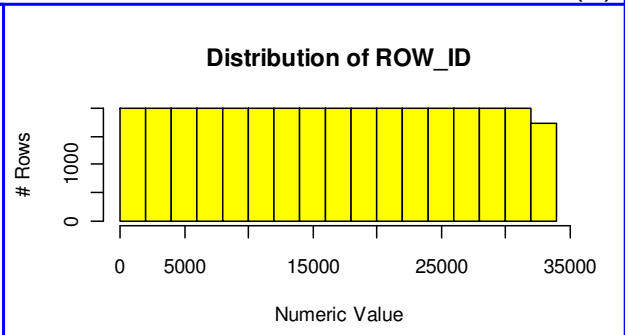
Responsys

- A Surrogate Key

DataProfTest . LOAD_WELCOME_KIT . ROW_ID 1 int(4)

	Rows	Nulls	Distinct
#	33,733	0	33,733
%	100.00	0.00	100.00

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	8,430	16,900	16,900	25,300	33,700



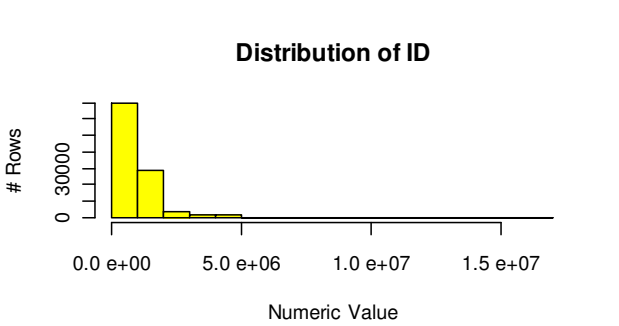
Head: 1|2|3|4|5|6

- Probable Business Key

TriRaw . raw_accounts . ID 1 varchar(8)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	104,021	0	104,021	0	104,021	2,707
%	100.00	0.00	100.00	0.00	100.00	2.60

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
90,700	301,000	803,000	929,000	1,440,000	16,900,000



Head: 13222820|13208821|13012891|13007600|13003661|13003041

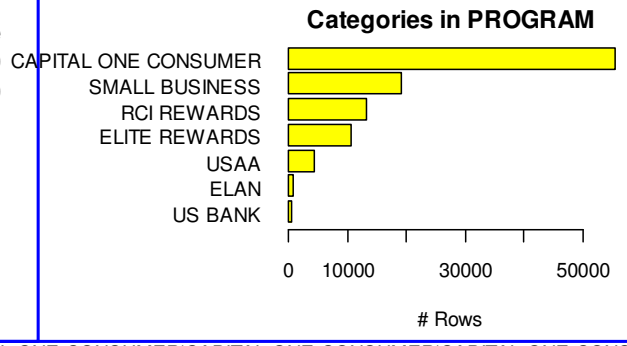
Data Profiling in R – Examples (2)

Responsys

- A Few Categories

TriRaw . raw_accounts . PROGRAM 3 varchar(20)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	104,021	0	7	0	0	0
%	100.00	0.00	0.007	0.00	0.00	0.00

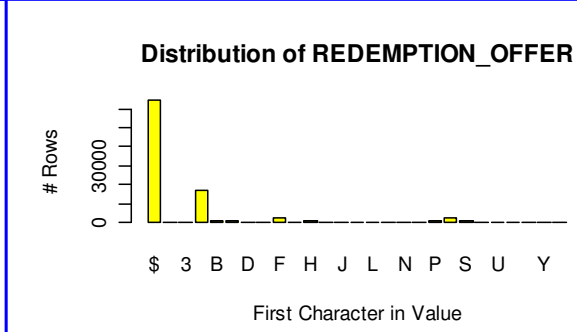


Head: CAPITAL ONE CONSUMER|CAPITAL ONE CONSUMER|CAPITAL ONE CONSUMER|CAPITAL ONE CONSUMER|CAPITAL ONE CONSUMER|CAPITAL ONE CONSUM

- Many Categories

TriRaw . raw_redemptions . REDEMPTION_OFFER 3 varchar(30)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	90,651	0	866	0	0	0
%	100.00	0.00	0.955	0.00	0.00	0.00



Head: Weber Portable Gas Grill|Weber Portable Gas Grill|Weber Portable Gas Grill|Sony AM/FM Sport Walkman|Sony AM/FM Sport Walkman|Sony AM/FM Sport Walkman

Data Profiling in R – Examples (3)

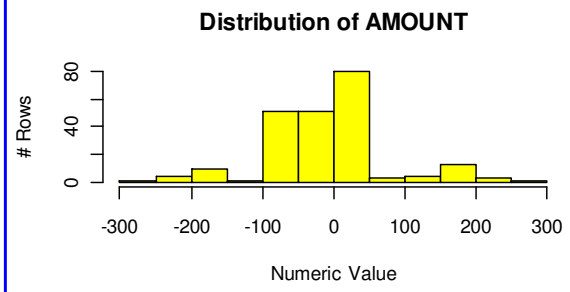
Responsys

- Numeric Value

AMA_Stage . RECEIVABLE_TXN_DETAIL . AMOUNT 3 varchar(8000)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	8,103,330	0	18,600	0	8,103,330	3,820
%	100.00	0.00	0.23	0.00	100.00	0.047

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-295	-50	-25	0	50	295



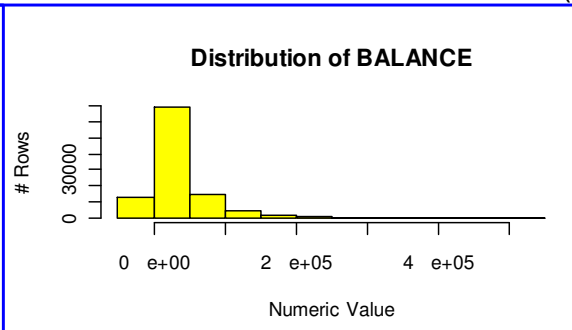
Head: 30|-30|247|-247|30|-30

- Another Numeric Value

TriRaw . raw_accounts . BALANCE 5 varchar(6)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	104,021	0	53,663	0	104,021	21,118
%	100.00	0.00	51.59	0.00	100.00	20.30

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-47,700	4,160	17,300	31,600	43,100	516,000



Head: 6416|5116|10195|10000|10000|10000

Data Profiling in R – Examples (4)

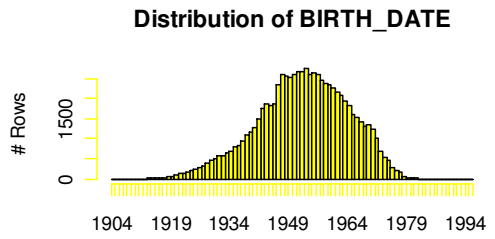
Responsys

- Reasonable Dates

TriRaw . raw_accounts . BIRTH_DATE 8 varchar(10)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	104,021	25,695	17,750	0	0	78,326
%	100.00	24.70	17.06	0.00	0.00	100.00

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1904-07-02	1945-07-06	1953-04-12	1952-10-29	1961-03-08	1995-01-29



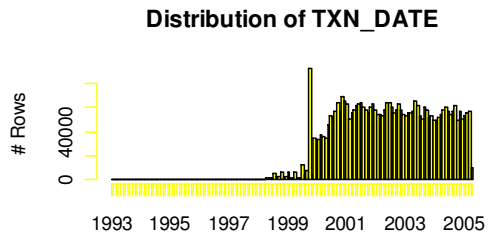
Head: 3/30/1954|3/17/1969|4/5/1971|5/11/1939|4/14/1952|2/10/1960

- Unusual Dates

AMA_Stage . RECEIVABLE_TXN . TXN_DATE 8 varchar(8000)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	3,861,249	0	1,773	0	0	3,861,190
%	100.00	0.00	0.046	0.00	0.00	100.00

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1993-01-29	2001-07-31	2002-11-30	2002-12-02	2004-04-30	2005-10-17



Head: 1/27/2000 0:00:00|1/27/2000 0:00:00|1/27/2000 0:00:00|1/27/2000 0:00:00|1/27/2000 0:00:00|1/27/2000 0:00:00

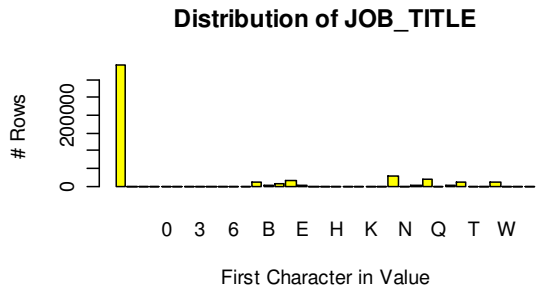
Data Profiling in R – Examples (5)

Responsys

- Customer Job Title not too useful

AMA_Stage . CUSTOMER . JOB_TITLE 7 varchar(8000)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	471,400	340,240	34,242	127	5	0
%	100.00	72.18	7.26	0.027	0.004	0.00

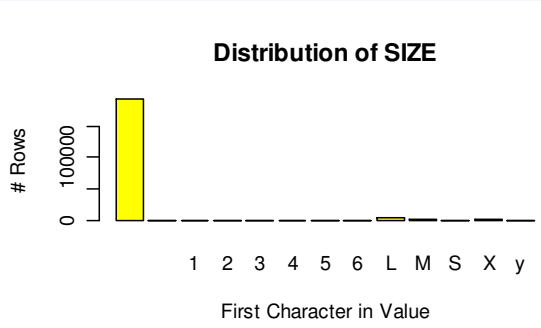


Head: NA|NA|NA|NA|NA|NA

- T-shirt Size also not reliable

OliviaMatrix . RS_CONTACT_INFO . SIZE 11 varchar(16)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	204,051	192,277	29	1	0	0
%	100.00	94.23	0.014	0.00	0.00	0.00



Head: NA|NA|Large|NA|XXL|NA

Responsys®

Reproducible Reporting

What's Reproducible Reporting?

Responsys

- Use methods from "Reproducible Research"
  - a published paper should include data and analysis code to produce tables, charts, and conclusions
  - See www.stat.washington.edu/jaw/jaw.research.reproducible.html

- Using R for "Reproducible Reporting" leverages:
  - R's connectivity
  - R's advanced analysis & visualization
  - odfWeave to produce OpenOffice text document
    - Business analysts can edit
    - Export to PDF or .doc
    - Based on Sweave (LaTeX) work by Fritz Leisch .

- Following example is actual data quality assurance report we run every month as part of a large data warehouse update

Actual Output Report

Responsys

MO QA 2.0 Counts for Load As Of 2008-07-04

Summary

Note missing week & very unusual Upgrade ramp-up highlighted below.

Missing Data:

WeekOfSatThe_	NumberNew
2008-06-21	0

Unusual Trends:

Data Set	p.value	AboveBelow	Severity
Upgrades	0.001268	-------+++++	HIGHLY SUSPECT
SessionsConsumed	0.053724	++++++-++---	Unusual

Where AboveBelow shows weeks with counts above, +, or below, -, mean of all weeks.

Membership

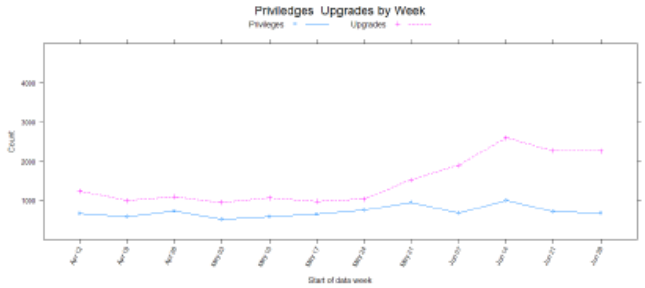
Notes:

Service Requests and Membership Events

Notes:

Miscellaneous Events

Notes:



odfWeave “source”

Responsys

Responsys

MO QA 2.0 Counts for Load As Of \Sexpr{mo_AsOfDate}

Summary

Missing Data:

```
<<MissingTable, echo = FALSE, results = xml>>=
if(!is.null(nrow(Missing))) {
  odfTable(Missing, useRowNames = FALSE, colnames = colnames(Missing))
} else {
  odfCat("No completely missing data elements for any weeks.")
}
@
```

Unusual Trends:

```
<<UnusualTable, echo = FALSE, results = xml>>=
if(!is.null(nrow(Suspects))) {
  #op <- options()
  #options(digits = 3)
  odfTable(Suspects, useRowNames = TRUE,
           colnames = c("Data Set", colnames(Suspects)))
  #options(op)
} else {
  odfCat("No unusual trends observed.")
}

if(!is.null(nrow(Suspects))) {
  odfCat("Where AboveBelow shows weeks with counts above, +, or below, -,
  mean of all weeks.")
}
@
```

Membership

Notes:

```
<<MembCounts, echo = FALSE, fig = TRUE>>=
sID(6.5, 3, 200)
print(
xyplot(NumberNew + NumberLost ~ WeekOfSatThe_, CountsByWeek,
       type = "b", ylim = c(0, 50000),
       auto.key = list(title = "New and Lost Members by Week", space =
"top",
                       columns = 2, lines = TRUE),
       xlab = "Start of data week", ylab = "Count",
       scales = list(x = list(at = CountsByWeek$WeekOfSatThe_, rot = 60)))
)
@
```

In-line “sweave” expression

Conditional table

Runs Test Exceptions in a table

Lattice X-Y Plot

See appendix for full width snap shots.

Unsupervised Clustering

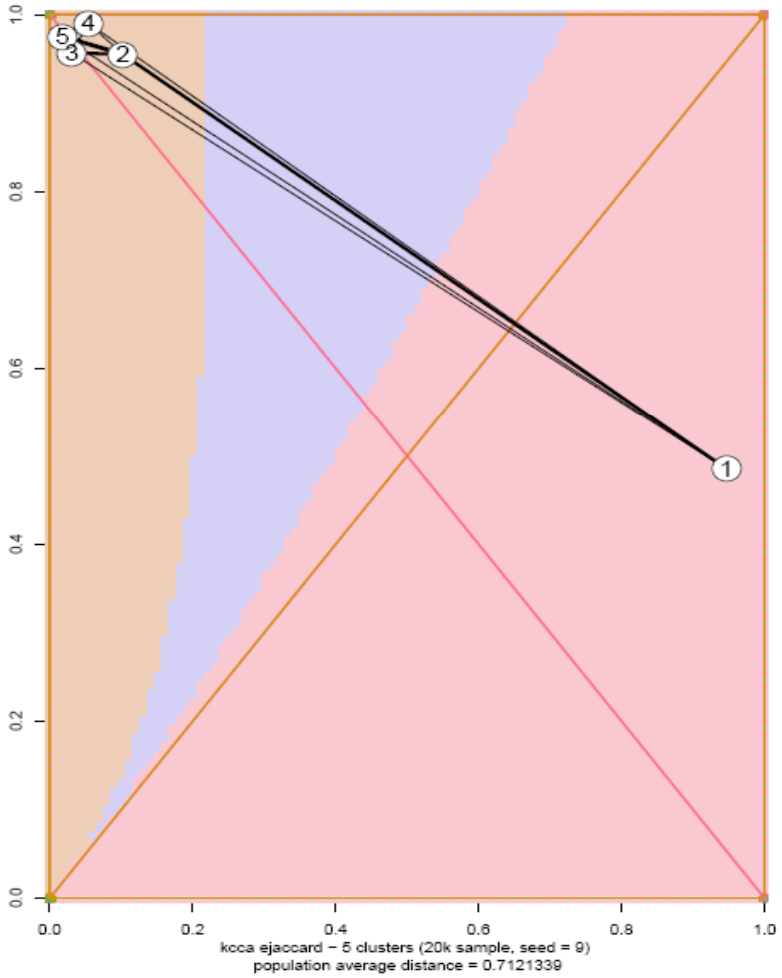
Prospect Segmentation – Problem Statement

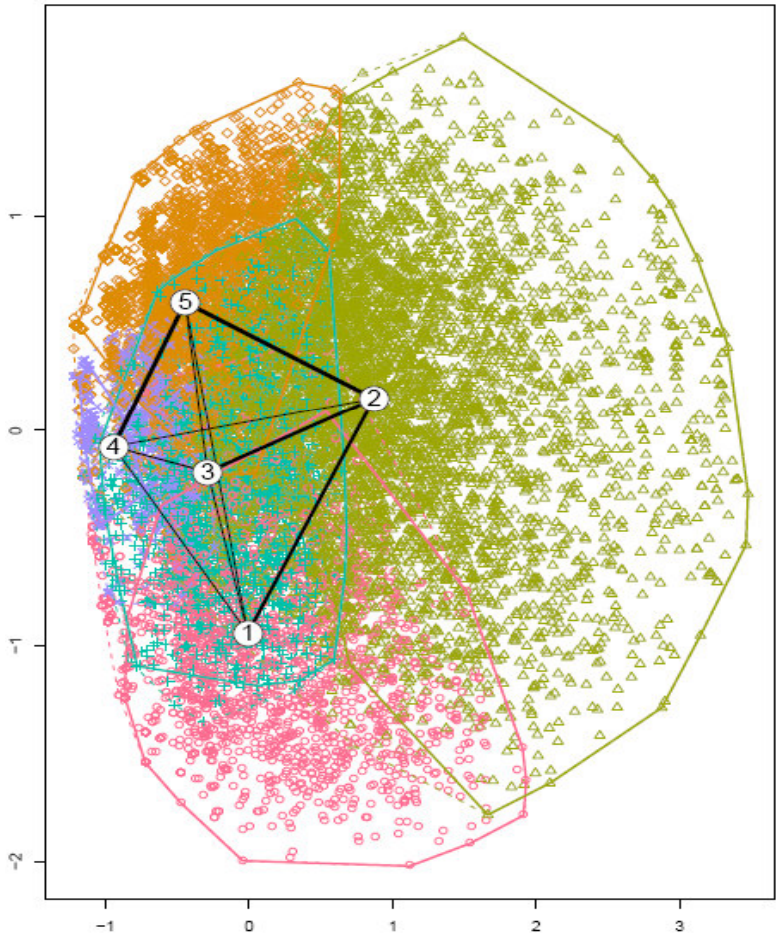
Responsys®

- A tech company surveying prospects
  - ~ 20 k respondents
  - ~ 35 check box type questions covering
    - Respondent's role
    - Hardware type
    - Interests
    - Applications
- Use Fritz Leisch's flexclust package
  - *K*-Centroids Cluster Analysis (KCCA)
  - Jaccard distance best suited for "preference" survey
  - do 4 runs with # clusters = 3 through 8
    - look for stable & meaningful clusters
  - see: www.stat.uni-munchen.de/~leisch
- Goal – assign future respondents to actionable segment
  - marketing action
  - marketing message

Customer Segmentation – Separation Plots

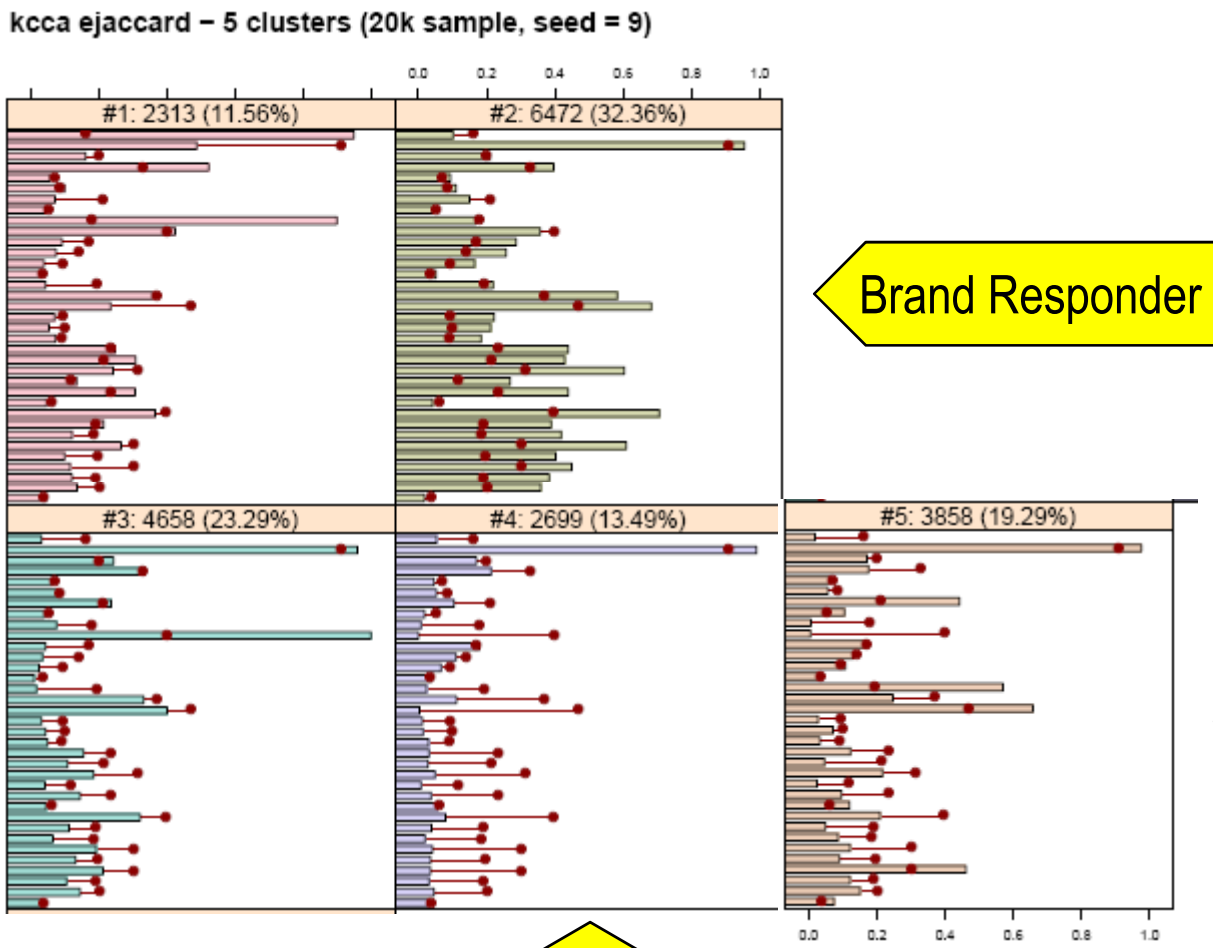
Responsys®





Customer Segmentation - Segments

Responsys



Loyalist A

Brand Responder

Loyalist B

Student

Brand Non-responder

Mike's R Examples

What is a good model?

Responsys

"All models are wrong. Some models are useful."

– George E.P. Box

A good statistical model is

- **Intuitive**
- **Estimable**
- ***Actionable***

Beyond providing insight, a good model suggests how to improve a system: "buy more of *x* and less of *y*."

A good model drives decisions.

What is a good model for a baseball hitter?

Responsys®

What is a good model?

Responsys

*"a hitter should be measured by his success in that what he is trying to do, ... **create runs.** It is startling, when you think about it, how much confusion there is about this."*

-Bill James, quoted in *Moneyball* (p.76)

Runs scored per game ~ At Bats, Walks, Hits (Singles, Doubles, Triples, HRs), Sac Flies, Hit by Pitch

What is a good model for a baseball hitter?

Responsys

Components of our baseball hitter model:

Runs scored per game (per team)
At Bats, Walks, Hits (Singles, Doubles, Triples, HRs),
Sac Flies, Hit by Pitch (per team)

Data source:

baseball-databank.org

Actionable Result:

Identify the most valuable hitters in the league.

Our tools:

R, RMySQL

What is a good model for a baseball hitter?

Responsys

Query the database to populate our R "data frame"

```
library(RMySQL)

con <- dbConnect(dbDriver('MySQL'),
            user='mdriscol',
            password = 'mypass',
            host = 'localhost',
            dbname = 'bbdb')

resultSet <- dbSendQuery(con,
               "select AB,BB,H,2B,3B,HR,SF,HBP,G,R
                  from teams
                  where yearID between 2000 and 2005")

teamStats <- fetch(resultSet, n=-1)

attach(teamStats)
```

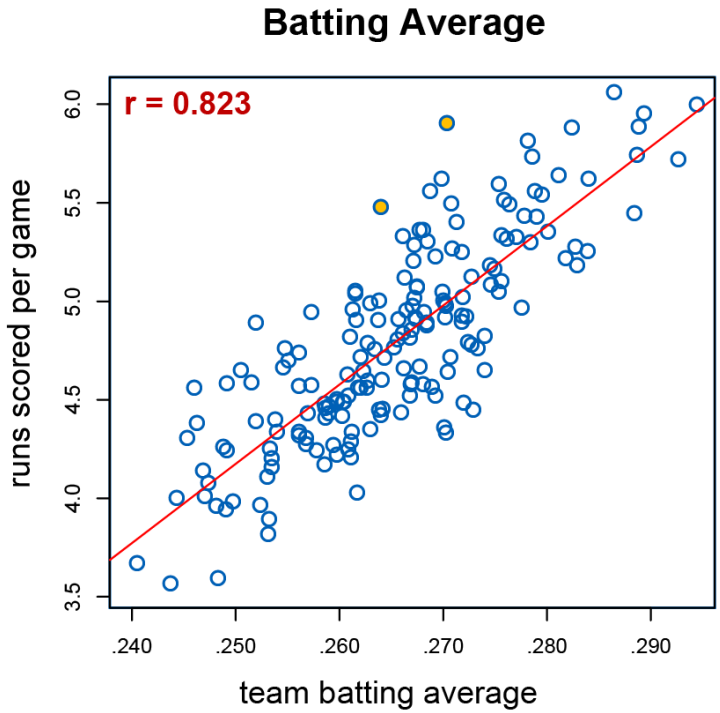
What is a good model for a baseball hitter?

Responsys

MODEL 1

Runs per game ~ Batting Average

Batting Average = Hits / At Bats




```
rpg <- R/G
bavg <- H/AB
plot(rpg ~ bavg)
```

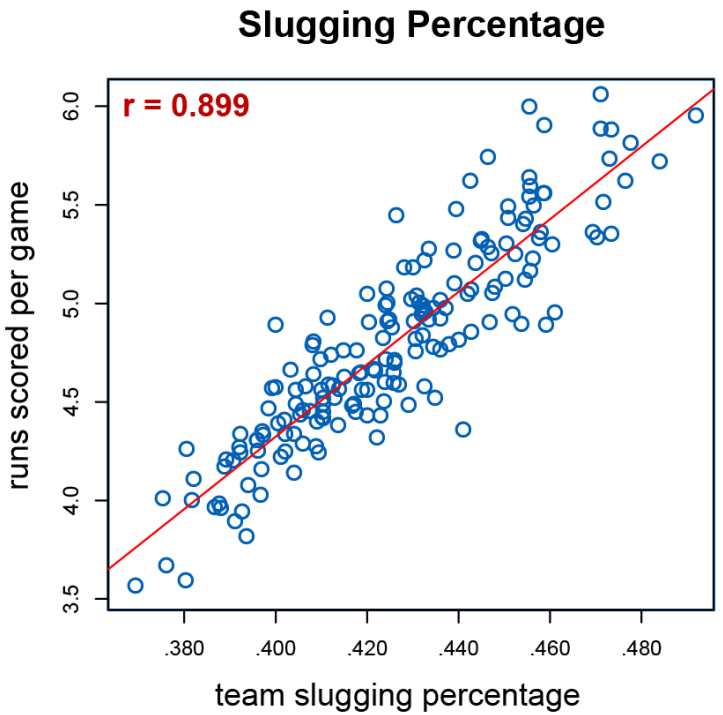
What is a good model for a baseball hitter?

Responsys

MODEL 2

Runs per game ~ Slugging %

Slugging % = Total Bases / At Bats




```
rpg <- R/G
slug <- (H + 2B + 3B*2 + HR*3)/AB
plot(rpg ~ slug)
```

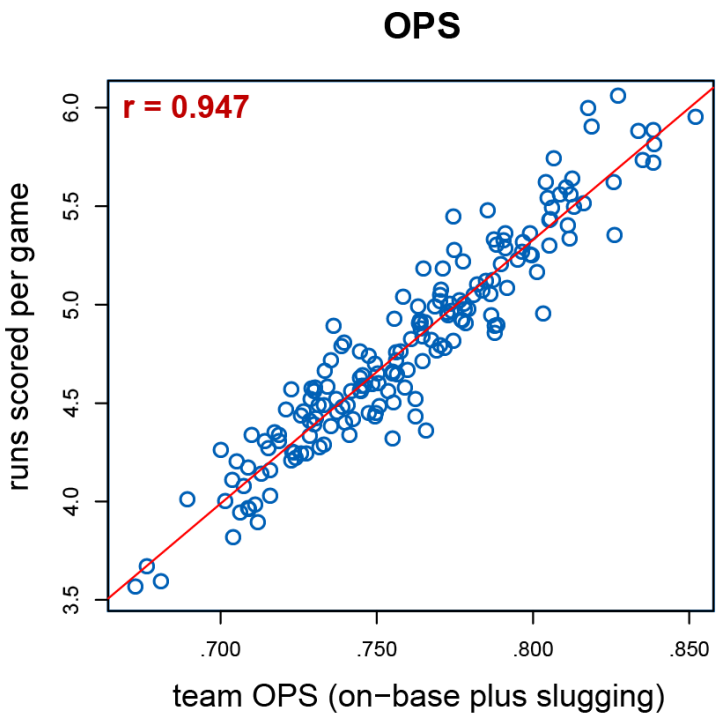
What is a good model for a baseball hitter?

Responsys

MODEL 3

Runs per game ~ OPS

OPS = **O**n Base **P**lus **S**lugging

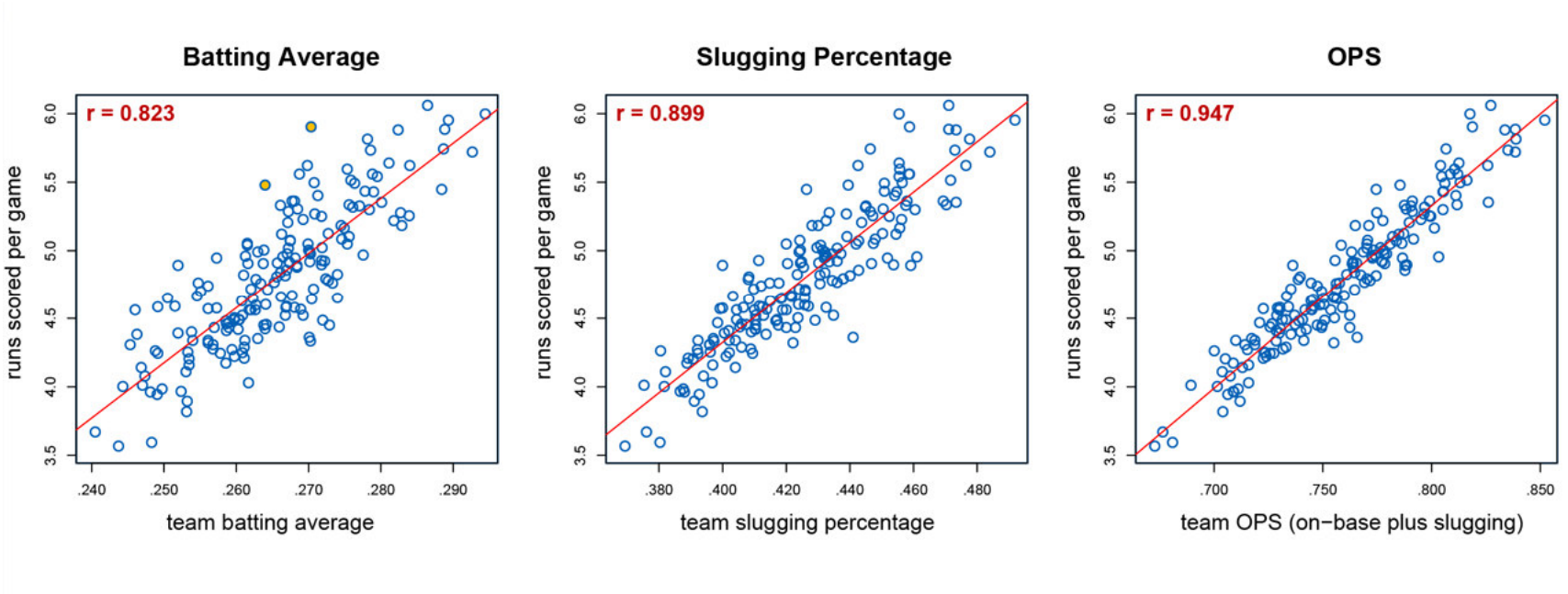



```
onbase <- (H+BB+HBP)/(AB+BB+SF)
OPS <- onbase + slug
plot(rpg ~ OPS)
```

What is a good model for a baseball hitter?

Responsys

Conclusion: **OPS** is the best predictor of **runs**



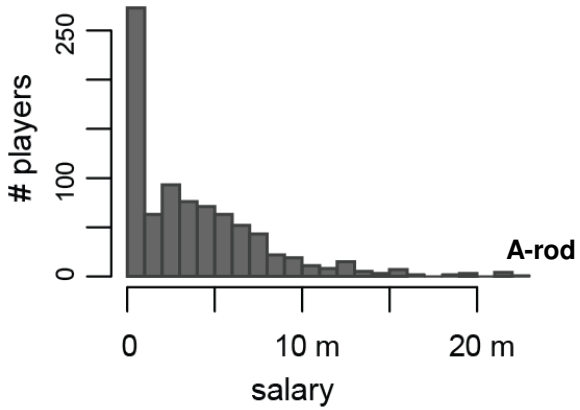
Hitters with high **OPS** are the most valuable hitters

Does OPS predict a player's salary?

Responsys

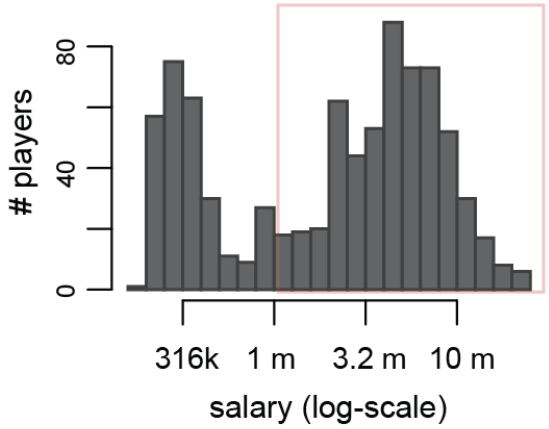
We take 836 data points for players between 2000 and 2005.

```
batters <- sql.fetch.all(con,
'select yearID,
salary,
b.*
from salaries s
 join master m using (idxLahman)
 join batting b using (idxLahman)
 join teams t on (t.idxTeams = b.idxTeams)
 where s.idxTeams = t.idxTeams
and yearID between 2000 and 2005
group by yearID, playerID
having AB > 300')
```



As is common with income data, non-normally distributed

log10(salary)

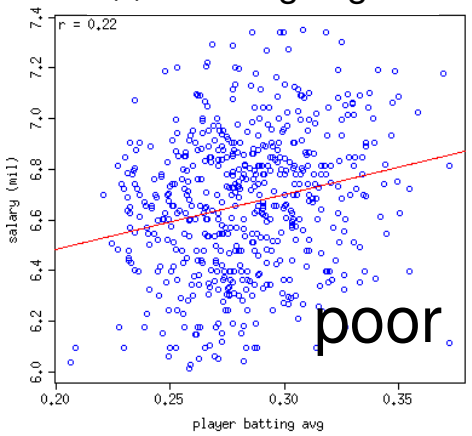


Log-normalizing the data reveals a bi-modal distribution; we'll restrict our analysis to players who earn > $1m annually

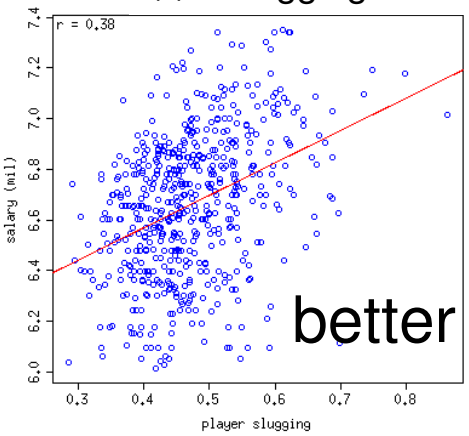
Does OPS predict a hitter's salary?

Responsys

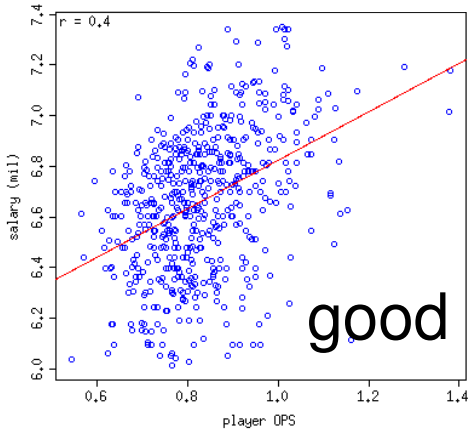
$$ *v* batting avg



$$ *v* slugging

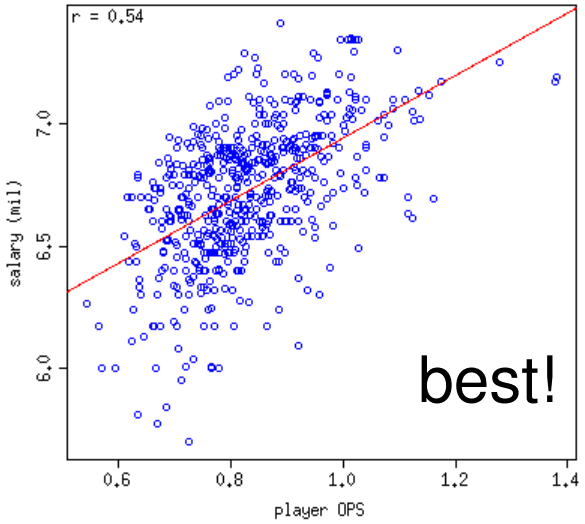


$$ *v* **OPS**



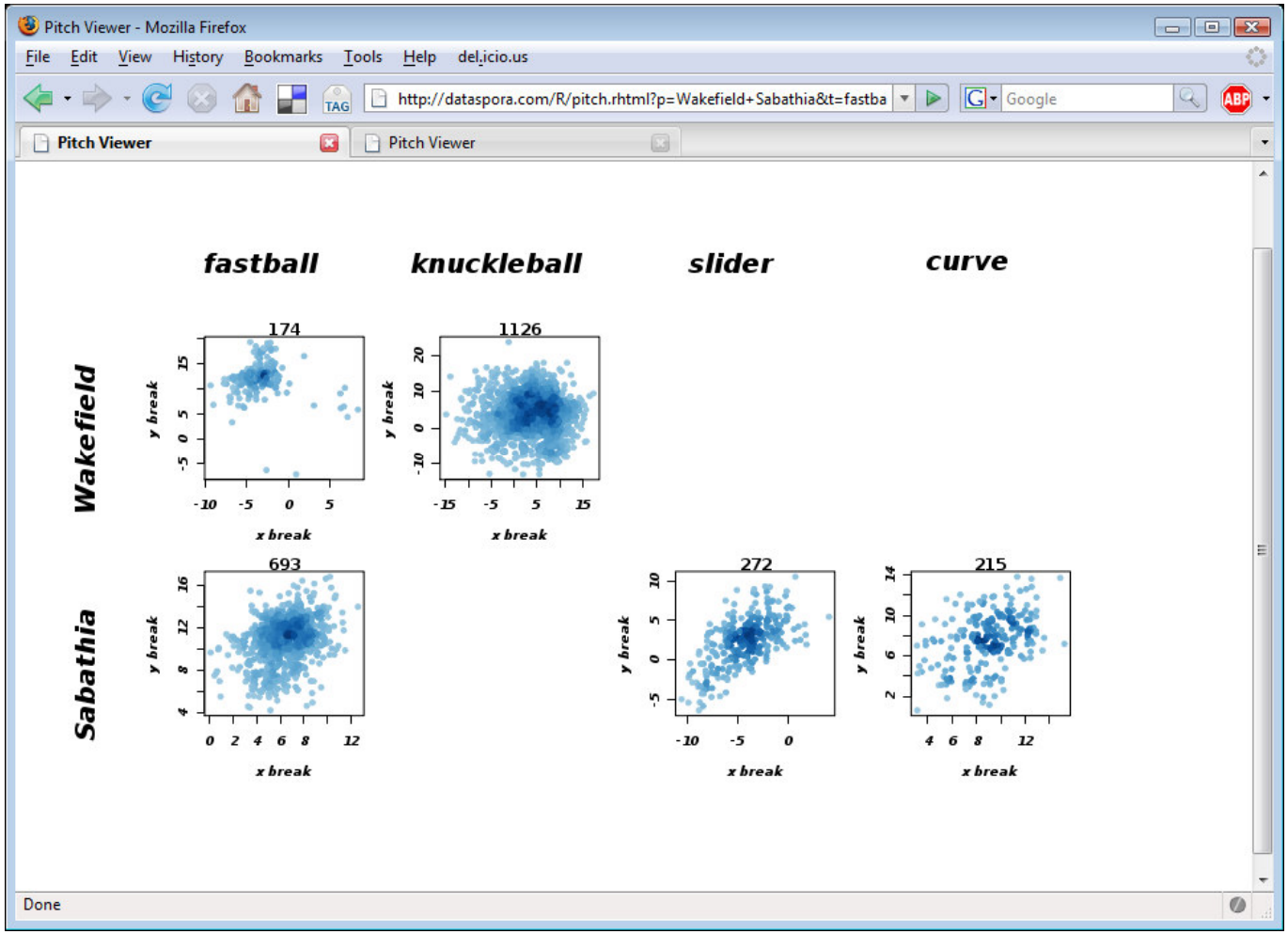
A hitter's previous year's **OPS** predicts his salary better than any other batting statistic

$$ *v* ***last year's*** **OPS**



Web Dashboards with Rapache

Responsys



http://www.dataspora.com/R

Getting Started with R

R Links

Responsys

- R Homepage: www.r-project.org
  - The official site of R
- R Foundation: www.r-project.org/foundation
  - Central reference point for R development community
  - Holds copyright of R software and documentation
- Local CRAN:
  - Mirror site
    - We use: cran.cnr.berkeley.edu
    - Find yours at: cran.r-project.org/mirrors.html
  - Current Binaries
  - Current Documentation & FAQs
  - Links to related projects and sites
- JGR Site: jgr.markushelbig.org/JGR.html

R Basics – Learning More

Responsys

Wikipedia

http://en.wikipedia.org/wiki/R_(programming_language)

An Introduction to R

http://cran.cnr.berkeley.edu/doc/manuals/R-intro.html

Links to all “official” manuals (html & pdf)

http://cran.cnr.berkeley.edu/manuals.html

R Graph Gallery

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

R Wiki

http://wiki.r-project.org/rwiki/doku.php

Introductory Statistics with R
Series: Statistics and Computing
Dalgaard, Peter
2nd ed., 2008, XVI, 364 p., Softcover
ISBN: 978-0-387-79053-4

Not yet published. Available: July 25, 2008

Questions? Comments?

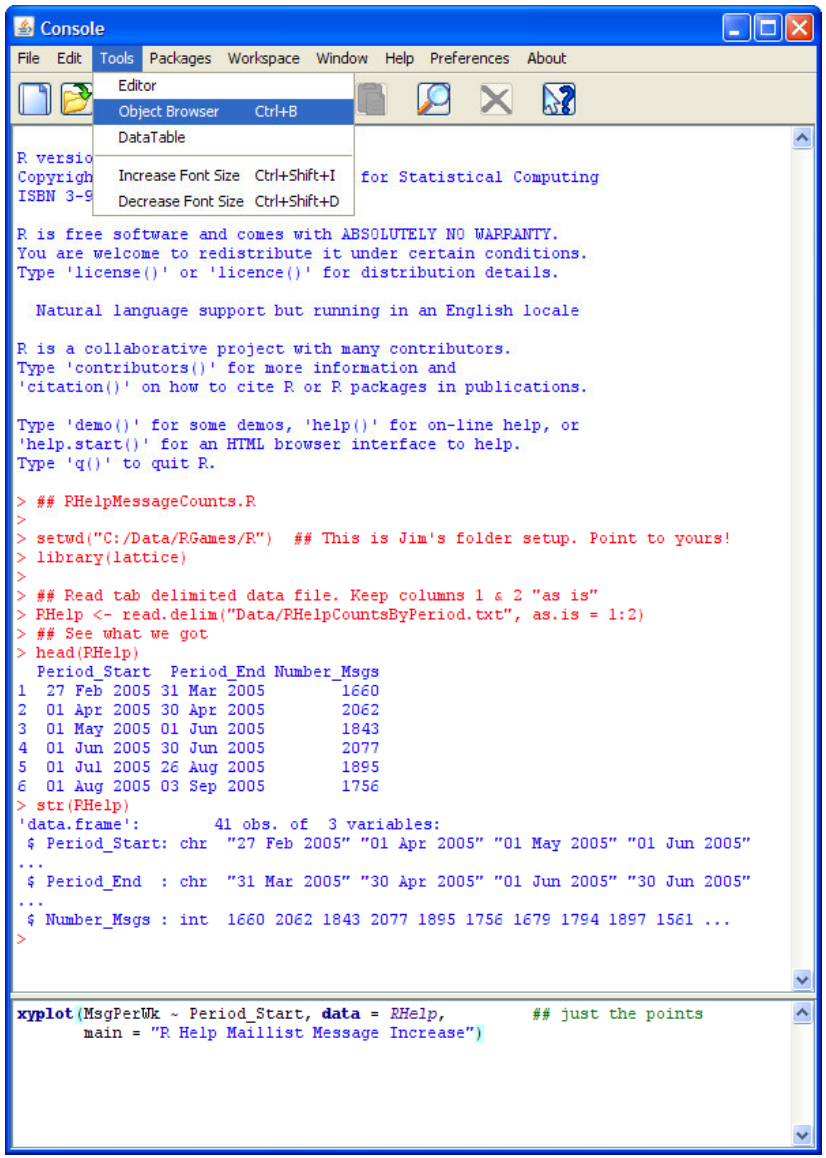
Responsys

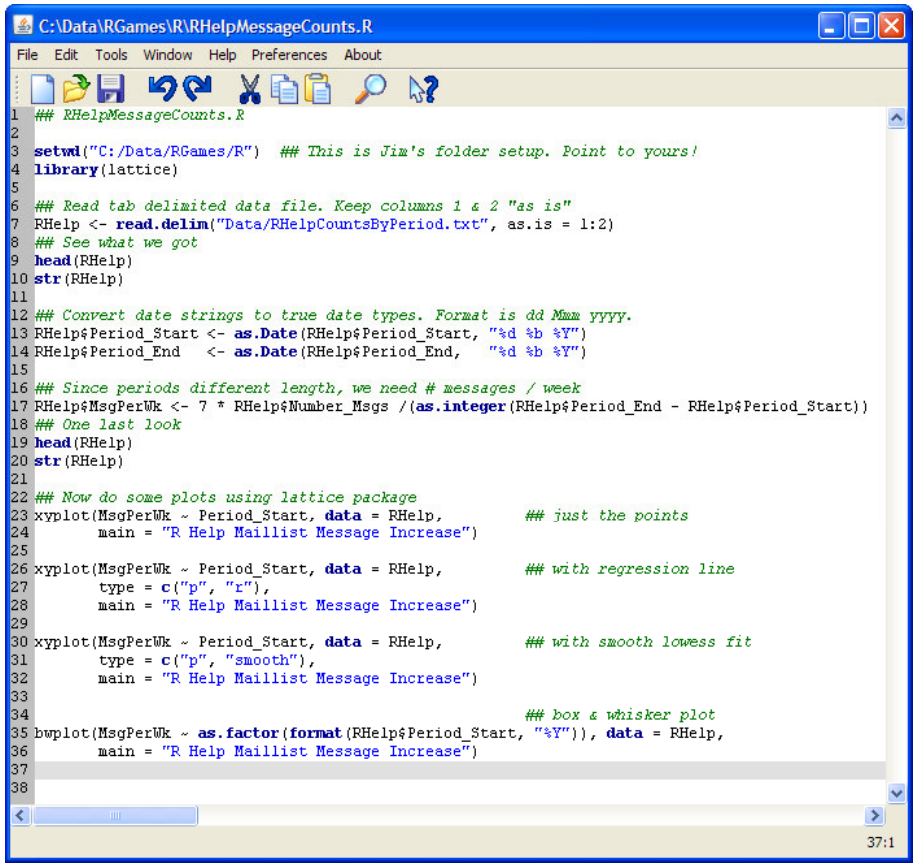
- Now would be the time!
- Keep in contact!
  - Mike's email: mike@dataspora.com
  - Jim's email: JPorzak@Responsys.com
- Jim's past presentations: www.porzak.com/JimArchive

Appendix

JGR Demo Details

Responsys





Download data & R code from
www.porzak.com/JimArchive/RHelp.zip

Expanded odfWeave source file (1)

Responsys

Responsys

MO QA 2.0 Counts for Load As Of \Sexpr{mo_AsOfDate}

Summary

Missing Data:

```
<<MissingTable, echo = FALSE, results = xml>>=
if(!is.null(nrow(Missing))) {
  odfTable(Missing, useRowNames = FALSE, colnames = colnames(Missing))
} else {
  odfCat("No completely missing data elements for any weeks.")
}
@
```

Unusual Trends:

```
<<UnusualTable, echo = FALSE, results = xml>>=
if(!is.null(nrow(Suspects))) {
  #op <- options()
  #options(digits = 3)
  odfTable(Suspects, useRowNames = TRUE,
           colnames = c("Data Set", colnames(Suspects)))
  #options(op)
} else {
  odfCat("No unusual trends observed.")
}

if(!is.null(nrow(Suspects))) {
  odfCat("Where AboveBelow shows weeks with counts above, +, or below, -,
  mean of all weeks.")
}
@
```

Expanded odfWeave source file (2)

Responsys

Membership

Notes:

```
<<MembCounts, echo = FALSE, fig = TRUE>>=
siD(6.5, 3, 200)
print(
xyplot(NumberNew + NumberLost ~ WeekOfSatThe_, CountsByWeek,
       type = "b", ylim = c(0, 50000),
       auto.key = list(title = "New and Lost Members by Week", space =
"top",
                       columns = 2, lines = TRUE),
       xlab = "Start of data week", ylab = "Count",
       scales = list(x = list(at = CountsByWeek$WeekOfSatThe_, rot = 60)))
)
@
```