Exploratory Data Analysis with R

Matthew Renze

DAMA Iowa Chapter

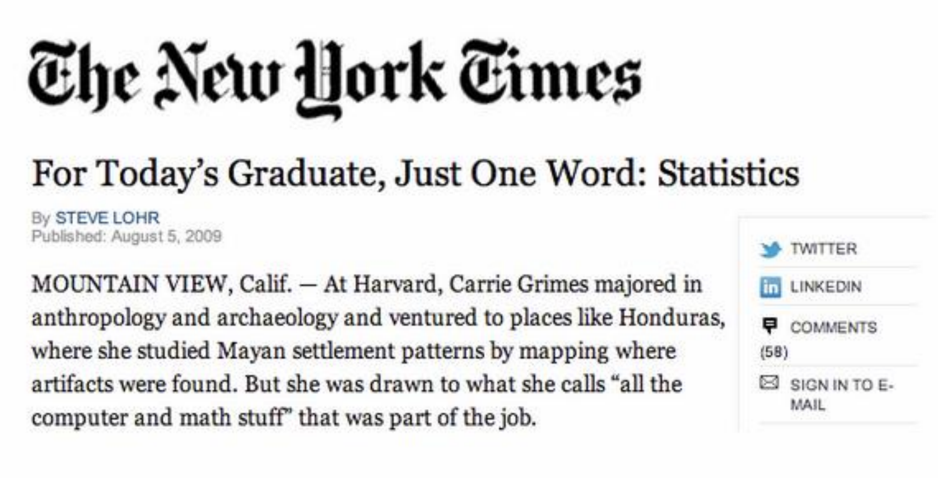
Motivation

The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, … because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

Motivation

The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT



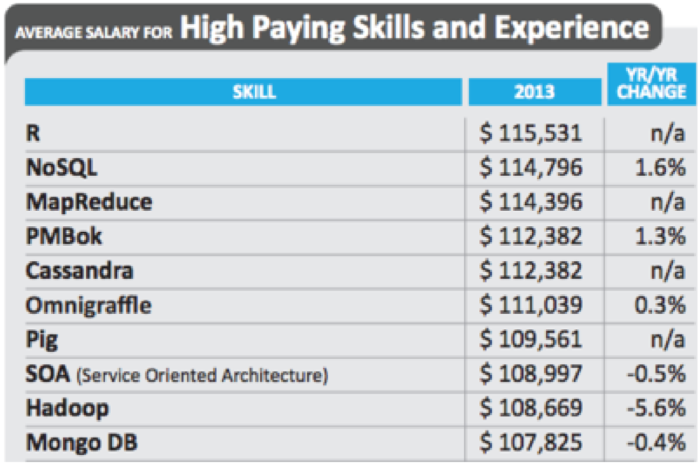
The New York Times

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

TWITTER
LINKEDIN
COMMENTS (58)
SIGN IN TO E-MAIL



AVERAGE SALARY FOR **High Paying Skills and Experience**

SKILL	2013	YR/YR CHANGE
R	$ 115,531	n/a
NoSQL	$ 114,796	1.6%
MapReduce	$ 114,396	n/a
PMBok	$ 112,382	1.3%
Cassandra	$ 112,382	n/a
Omnigraffle	$ 111,039	0.3%
Pig	$ 109,561	n/a
SOA (Service Oriented Architecture)	$ 108,997	-0.5%
Hadoop	$ 108,669	-5.6%
Mongo DB	$ 107,825	-0.4%

Source: Dice 2014 Tech Salary Survey Results

How Does This Apply to Me?

- As a software developer, I often:
  - ☑ Perform log file analysis
  - ☑ Analyze software performance
  - ☑ Analyze code metrics for code quality
  - ☑ Detect anomalies in source data
  - ☑ Transform or clean data files to make them usable
  - ☑ Help decision makers make decisions based on data

A Flood of Data is Coming…

Sink or Swim

Source: http://www.dot.gov.nt.ca/

Source: Wikipedia

Overview

- Introduction to R
- Data Munging
- Descriptive Statistics
- Data Visualization
- Beyond R and EDA

About Me

- Independent software consultant
- Education
  - B.S. in Computer Science (ISU)
  - B.A. in Philosophy (ISU)
- Training
  - Kimball Group – Data Warehousing
  - ESRI - ArcGIS, ArcSDE, and ArcGIS Server
  - Data Science Specialization (Johns Hopkins) [In progress]

IOWA STATE UNIVERSITY

KIMBALL GROUP

esri

JOHNS HOPKINS UNIVERSITY

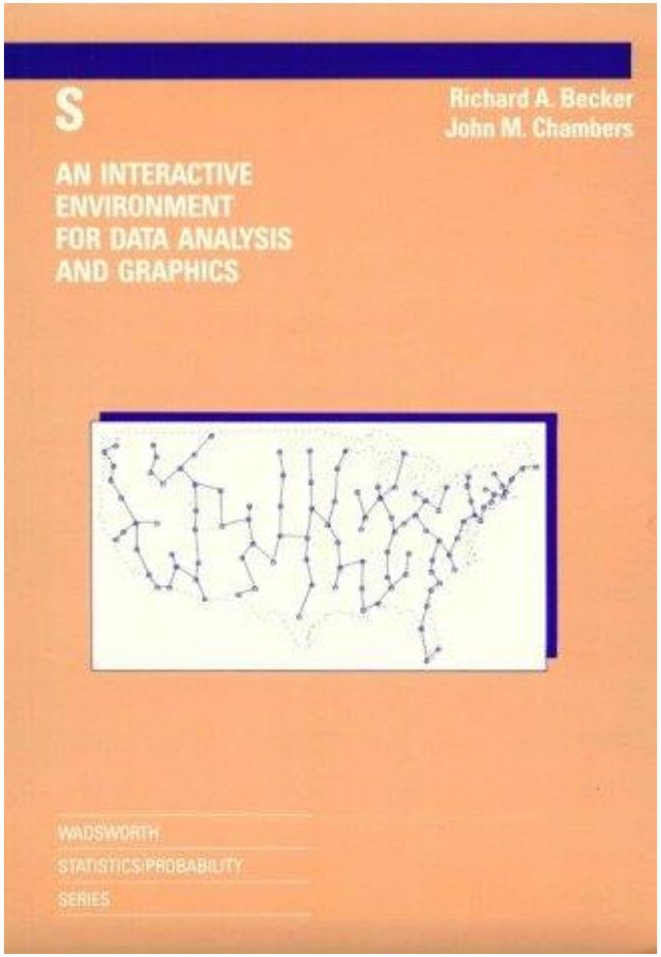
Introduction to R

What is R?

- R is an open source implementation of S

What is S?

- Statistical programming language
- Developed at Bell Labs in 1976
- Currently owned by TIBCO Software



S

Richard A. Becker
John M. Chambers

AN INTERACTIVE
ENVIRONMENT
FOR DATA ANALYSIS
AND GRAPHICS

WADSWORTH
STATISTICS/PROBABILITY
SERIES

A Brief History of R

- 1991 - R is developed by:
  - Ross Ihaka
  - Robert Gentleman
- 1995 - R became open source
- 2000 - R v.1.0 was released
- Today, R is at v.3.1.1

Source: https://www.stat.auckland.ac.nz/~ihaka/downloads/the-r-project.pdf

Source: www.auklandlifestyle.com

What is R?

R is:

- an open source implementation of S
- a language and an environment
- provides methods for both statistical and graphical data analysis
- runs on Windows, Mac, and Unix systems

Source: www.r-project.org

What is R?

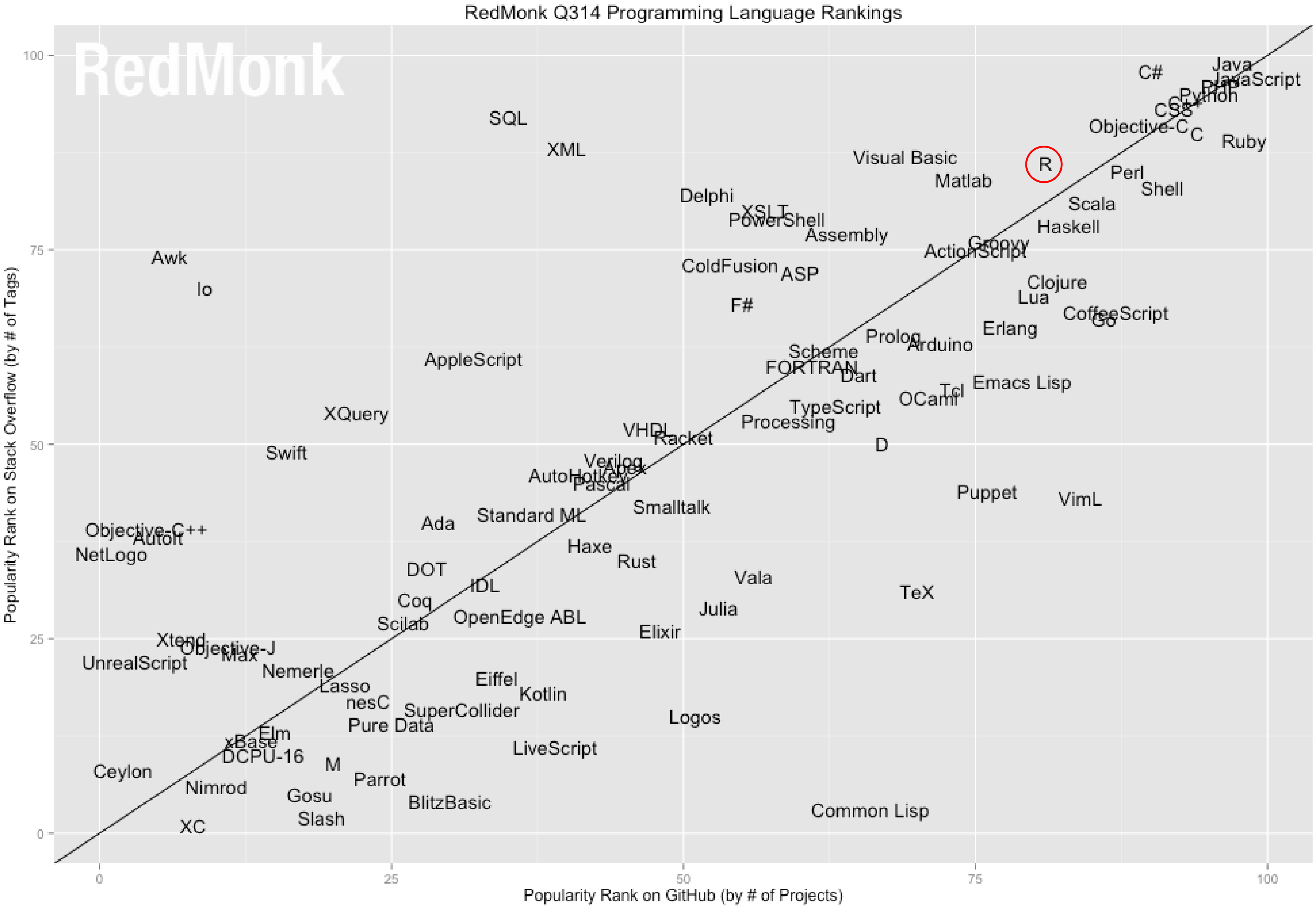
R is also:

- actively under development
- has a large user community
- is very modular and extensible
- has over 4000 extension packages

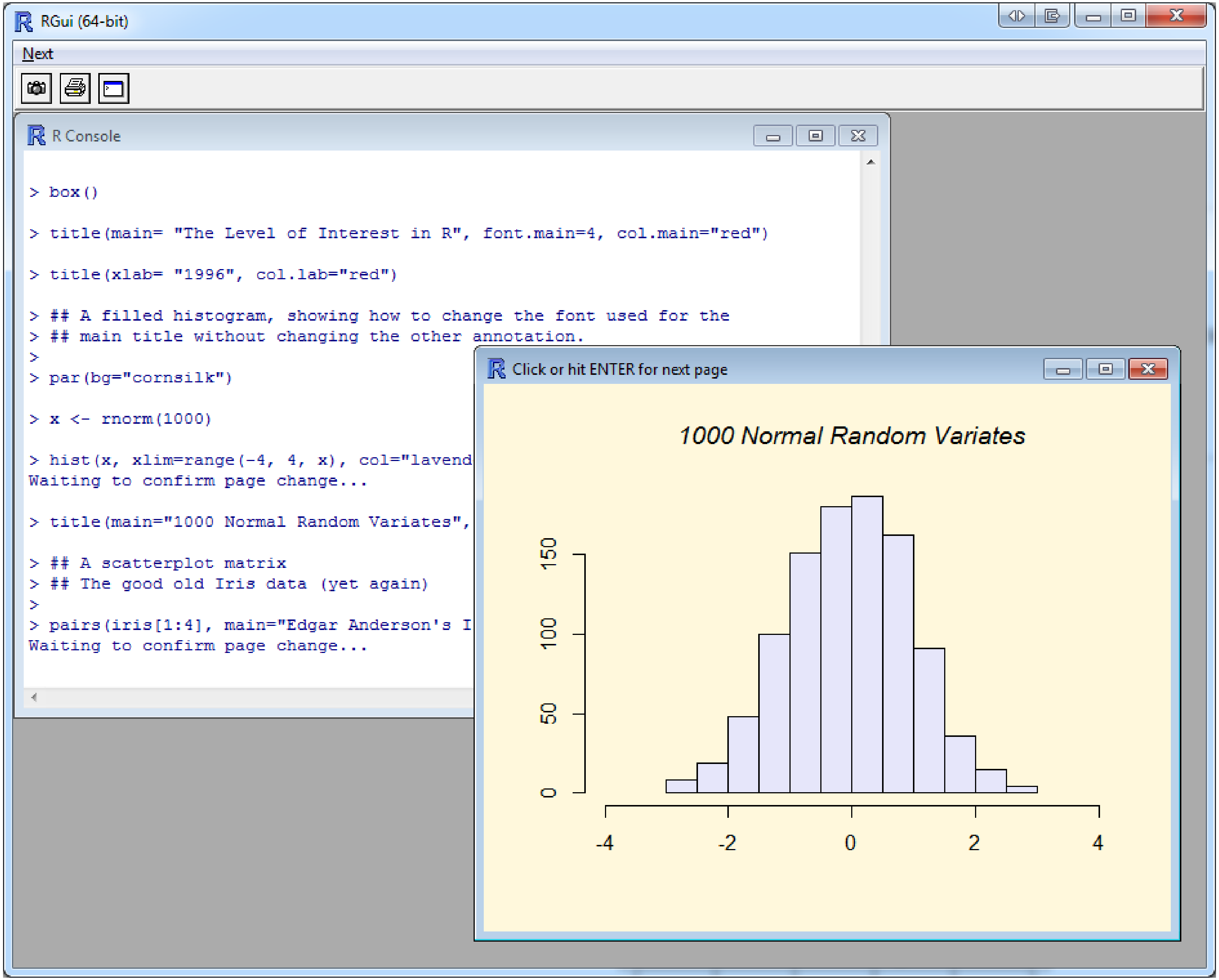
and best of all…

FREE

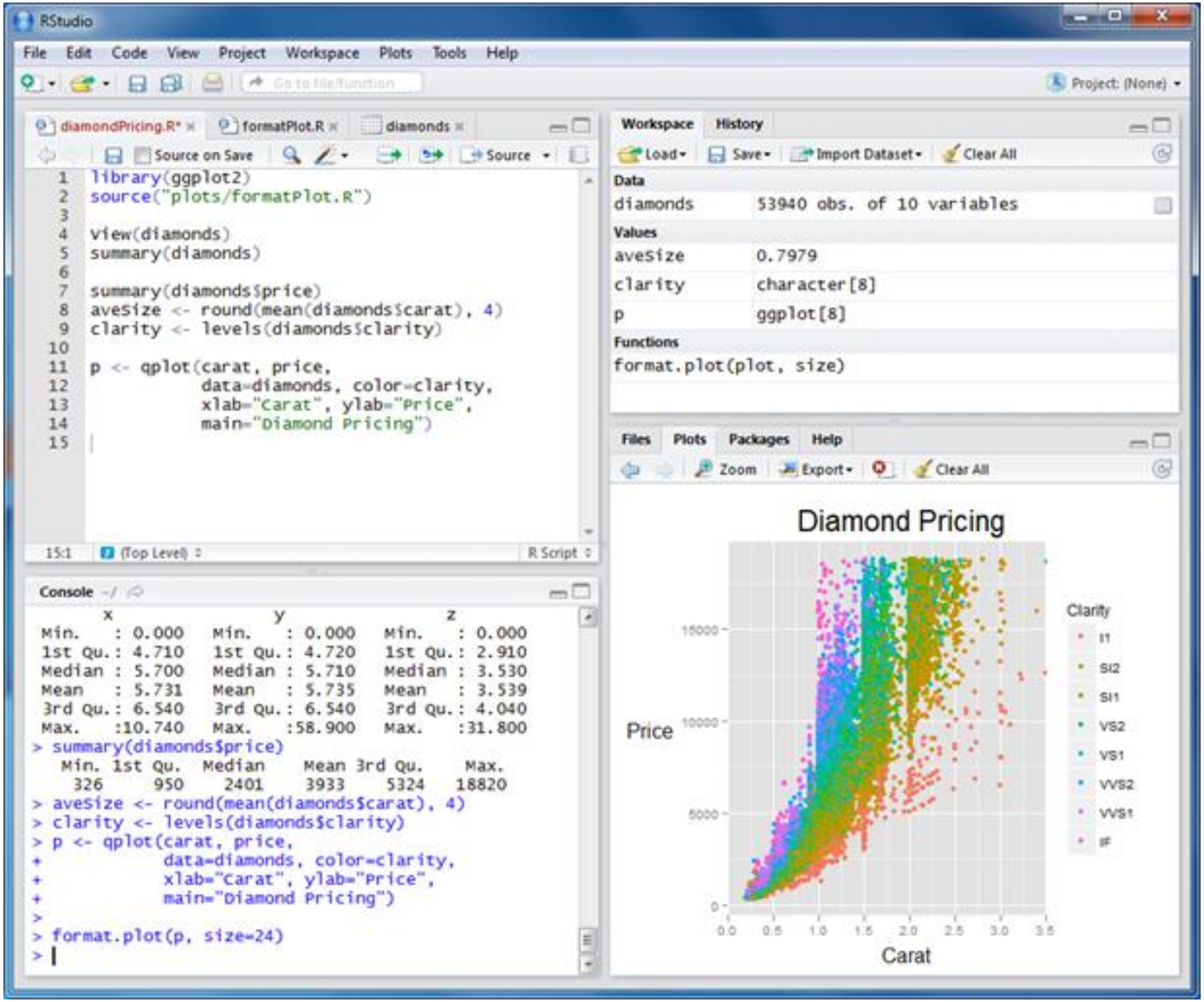
FREE



Source: http://redmonk.com/sogrady/2014/06/13/language-rankings-6-14/



```
> box()

> title(main= "The Level of Interest in R", font.main=4, col.main="red")

> title(xlab= "1996", col.lab="red")

> ## A filled histogram, showing how to change the font used for the
> ## main title without changing the other annotation.
>
> par(bg="cornsilk")

> x <- rnorm(1000)

> hist(x, xlim=range(-4, 4, x), col="lavend
Waiting to confirm page change...

> title(main="1000 Normal Random Variates",

> ## A scatterplot matrix
> ## The good old Iris data (yet again)
>
> pairs(iris[1:4], main="Edgar Anderson's I
Waiting to confirm page change...
```



```
library(ggplot2)
source("plots/formatPlot.R")

View(diamonds)
summary(diamonds)

summary(diamonds$price)
aveSize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)

p <- qplot(carat, price,
           data=diamonds, color=clarity,
           xlab="Carat", ylab="Price",
           main="Diamond Pricing")
```

Data

diamonds	53940 obs. of 10 variables

Values

aveSize	0.7979
clarity	character[8]
p	ggplot[8]

Functions

format.plot(plot, size)

```
       x                y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
>
```

Source: www.rstudio.com/ide/

Code Demo

Data Munging

Data Munging

- Transforming data from a raw form to a usable form
- Many data sets are not initially ready for analysis
- Data must be transformed or cleaned first

Source: Wikimedia

Data Munging Tasks

- Renaming variables
- Data type conversion
- Encoding, decoding, or recoding data
- Merging data sets
- Transforming data
- Handling missing data (imputing)
- Handling anomalous values

Loading Data in R

- R supports a wide variety of data sources
  - File-based data
    - CSV, TAB, Excel, etc.
  - Web-based data
    - XML, HTML, JSON, etc.
  - Databases
    - JDBC, ODBC, SQL Server, Oracle, MySQL, Access, etc.
  - Statistical data
    - SAS, SPSS, Stata
  - And many more…

Cleaning Data

- This step is often the:
  - Most difficult
  - Most time consuming
- TIP: Record all steps using a script so you can reapply the steps whenever they are needed

Source: Wikimedia

Code Demo: Lending Club Dataset

- Sample of 2,500 peer-to-peer loans
- *Problem:* The data are not in a digestible format
- *Goal:* Prepare the data for analysis

Investors earn better returns, borrowers pay lower rates.

Invest & Earn Solid Returns

22 consecutive quarters of positive returns.

Net Annualized Return** per grade net of defaults and fees from inception to 2/07/2013. Prospectus filed with the SEC.

Borrow up to $35,000*

Fast and free to check your rate.

LATEST STATS: Lending Club investors have earned $125,117,203 in interest since inception. | see more »

Source: www.lendingclub.com

Code Demo

Descriptive Statistics

Descriptive Statistics

- Describe data in quantitative or qualitative ways
- Provides a summary of the shape of the data
- aka: Summary statistics

Interest Rate	
Statistic	Value
Minimum	5.42
1st Quartile	10.16
Median	13.11
Mean	13.07
3rd Quartile	15.80
Maximum	24.89
Variance	17.45
Standard Deviation	4.17

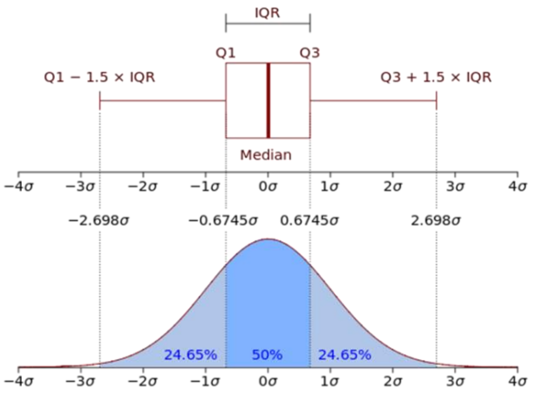
Statistical Terms

- Observations
  - Rows in the table
- Variables
  - Columns in the table
- Qualitative variable
  - Categorical values
- Quantitative variable
  - Numeric values

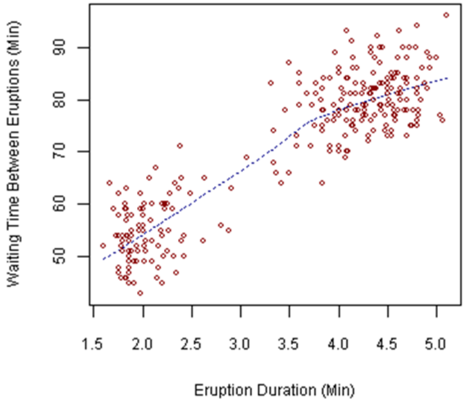
ID	Date	Customer	Product	Quantity
1	2012-10-27	John	Pizza	2
2	2012-10-27	John	Soda	2
3	2012-10-27	Jill	Salad	1
4	2012-10-27	Bob	Milk	1
5	2012-10-28	Sue	Soda	3
6	2012-10-28	Bob	Pizza	2
7	2012-10-28	Jill	Pizza	1
8	2012-10-28	Jill	Soda	3

Types of Numerical Analysis

- Several types based on:
  - Type of variable(s)
    - Qualitative (Categorical)
    - Quantitative (Numerical)
  - Number of variables
    - Univariate (One)
    - Bivariate (Two)
    - Multivariate (Many)

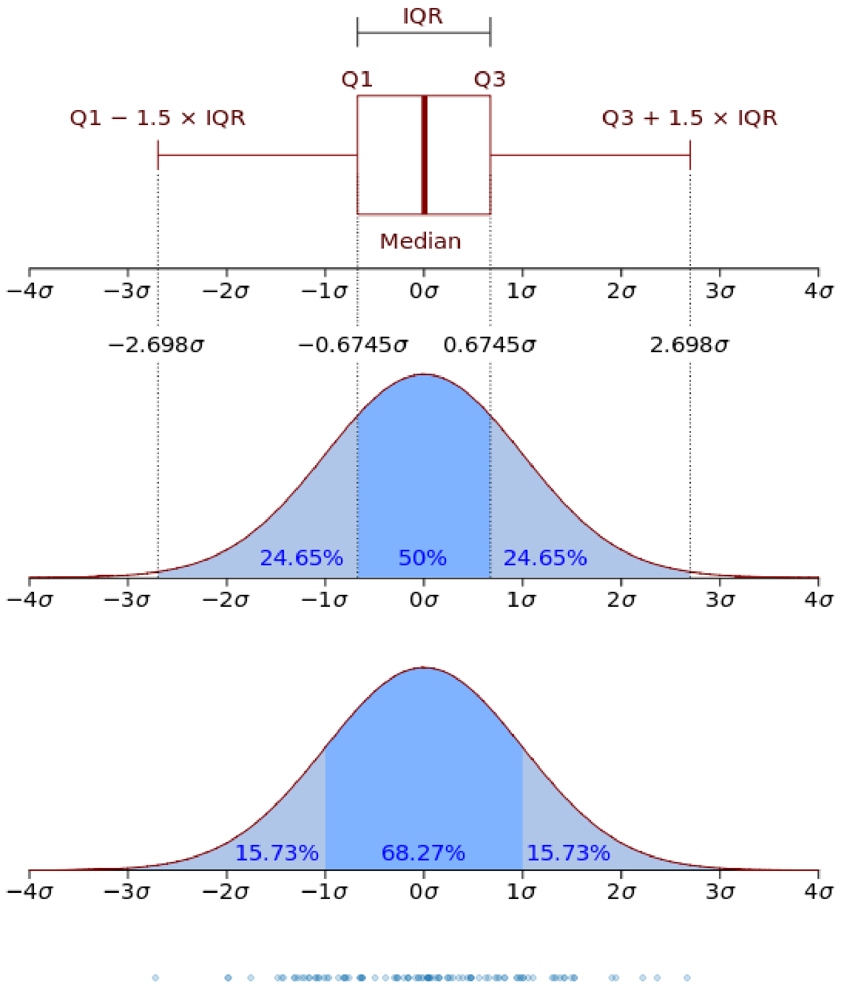


Old Faithful Eruptions



Univariate Analysis

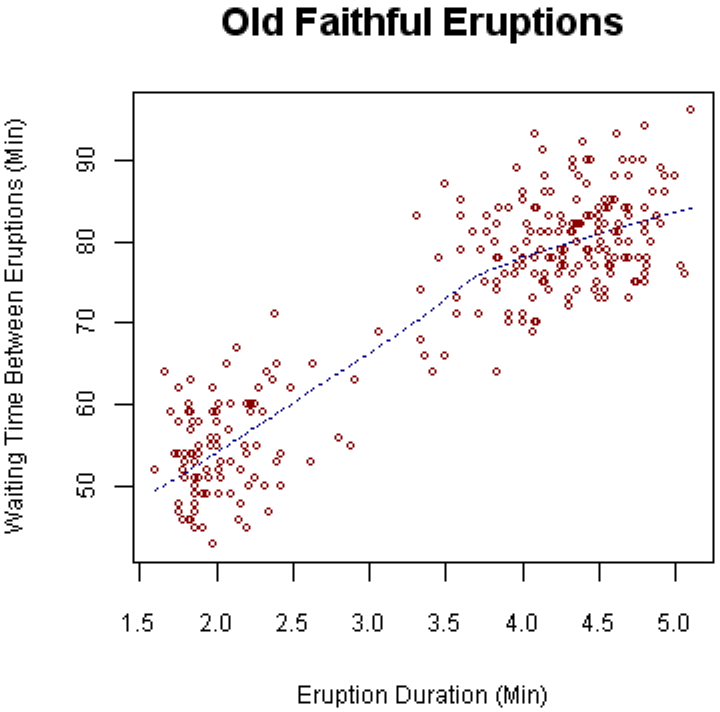
- Analysis of a single variable
- Measures include:
  - Central tendency
    - Mean
    - Median
    - Mode
  - Dispersion
    - Min
    - Max
    - Range
    - Quartiles
    - Variance
    - Standard deviation



Source: Wikipedia

Bivariate Analysis

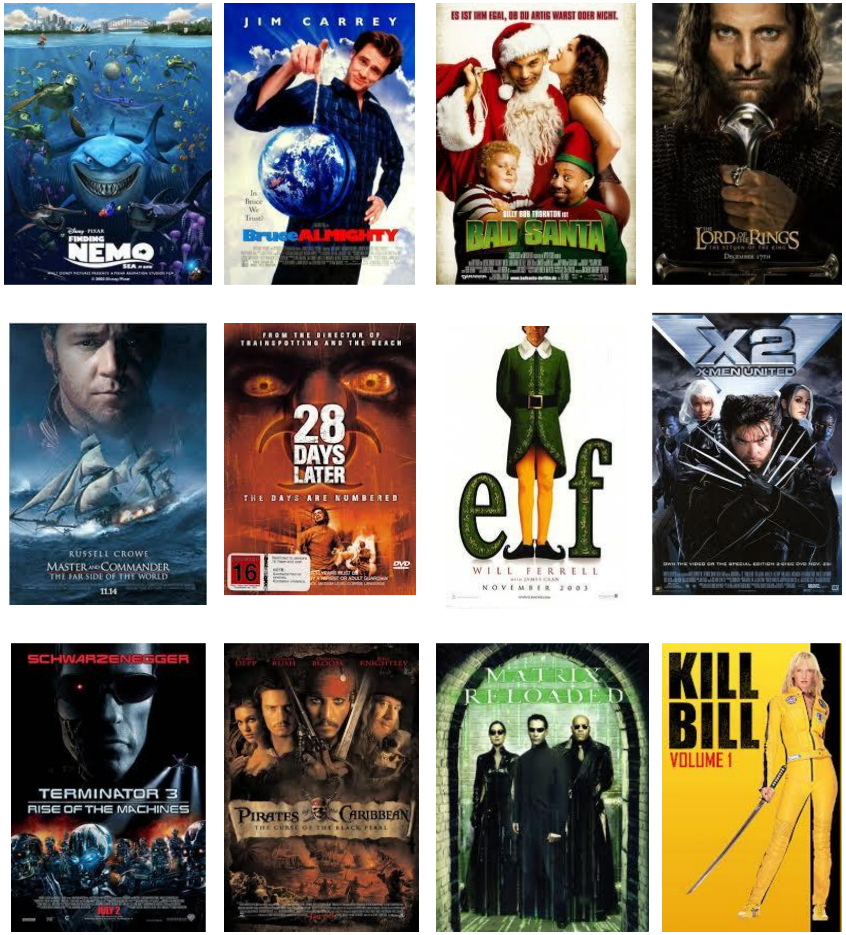
- Analysis of the relationship between two variables
  - Predictor
  - Outcome
- Measures include:
  - Covariance
  - Correlation



Source: Wikipedia

Code Demo: Movies Data Set

- Collection of movies from 2003
- Measurements include:
  - Movie Name
  - Rating (e.g., G, PG, R)
  - Genre (e.g., Action)
  - Running Length (min)
  - Critic Score (%)
  - Box Office Revenue ($)
- *Goal:* Determine what types of movies made the most money



Source: http://www.rossmanchance.com/iscam2/files.html

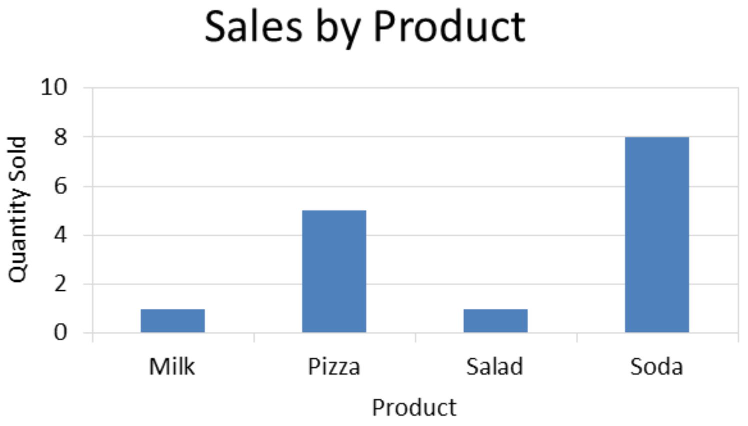
Code Demo

Data Visualization

Data Visualization

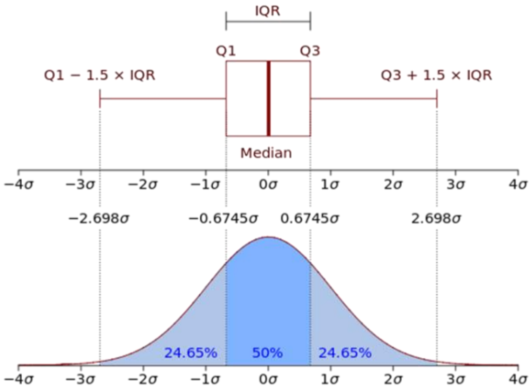
- Representation of data via visual means
- Human brain is exceptionally good at visual pattern recognition
- Map dimensions of data to visual characteristics:
  - Location
  - Size
  - Color
  - Shape

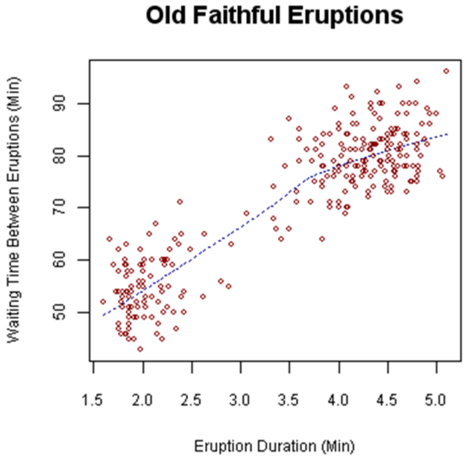
ID	Date	Customer	Product	Quantity
1	2012-10-27	John	Pizza	2
2	2012-10-27	John	Soda	2
3	2012-10-27	Jill	Salad	1
4	2012-10-27	Bob	Milk	1
5	2012-10-28	Sue	Soda	3
6	2012-10-28	Bob	Pizza	2
7	2012-10-28	Jill	Pizza	1
8	2012-10-28	Jill	Soda	3



Types of Data Visualizations

- Several types based on:
  - Type of variable(s)
    - Qualitative (Categorical)
    - Quantitative (Numerical)
  - Number of variables
    - Univariate
    - Bivariate
    - Multivariate

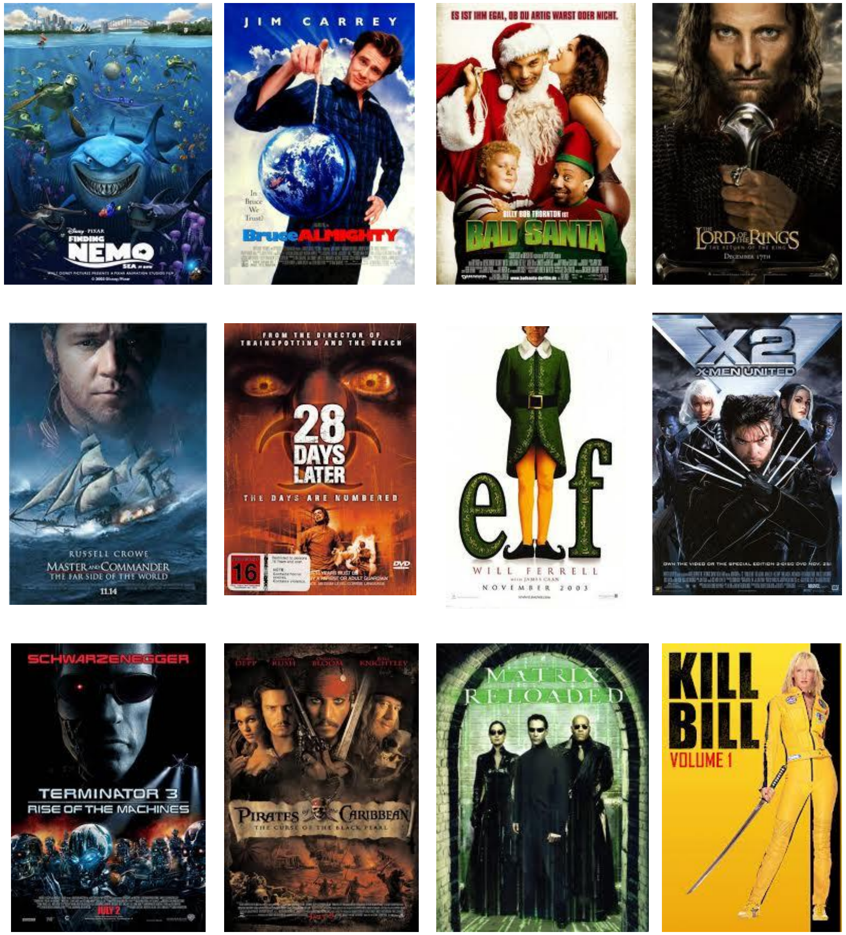




Source: Wikipedia

Code Demo: Movies Data Set

- *Goal:* Visualize what types of movies make the most money

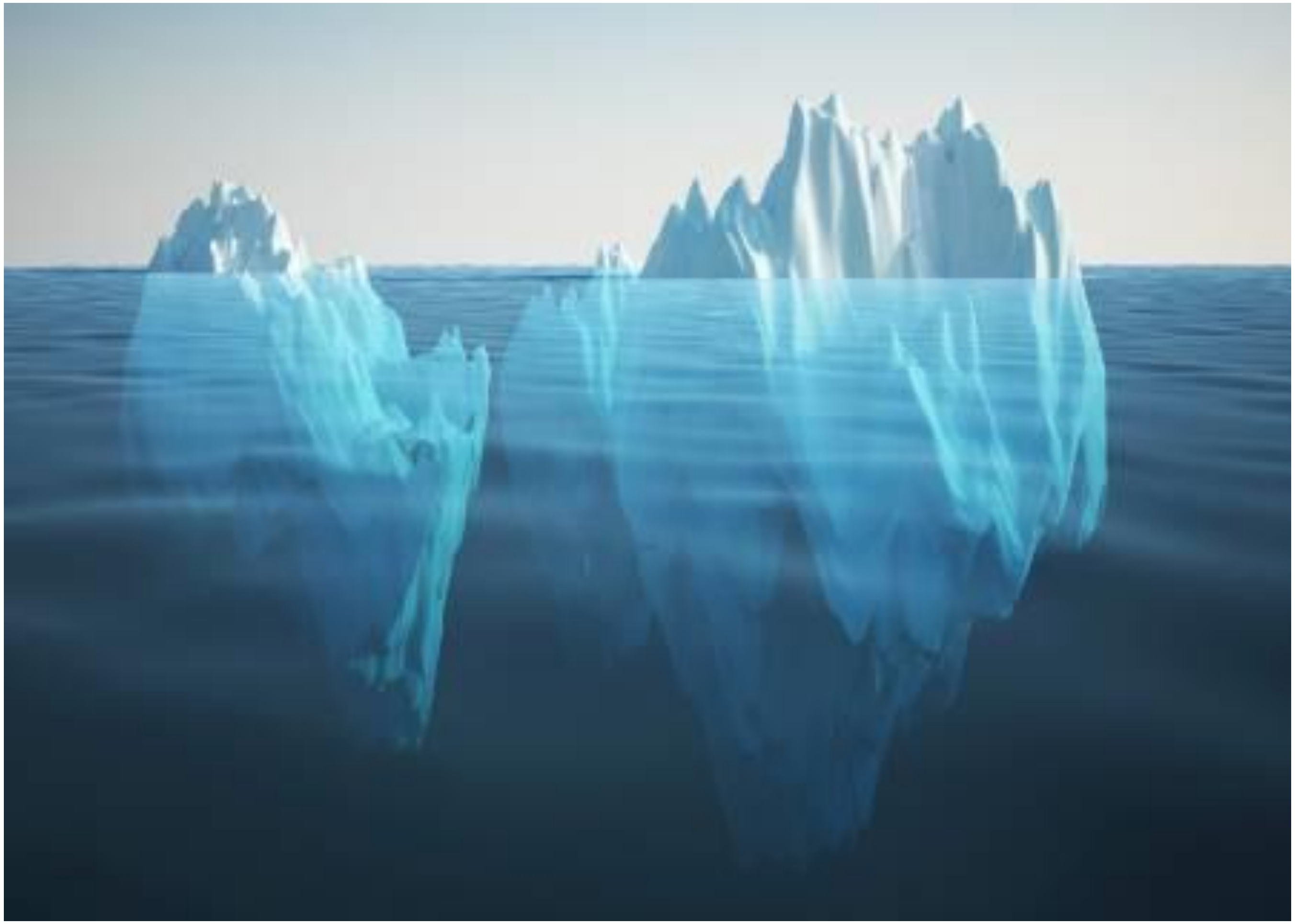


Source: http://www.rossmanchance.com/iscam2/files.html

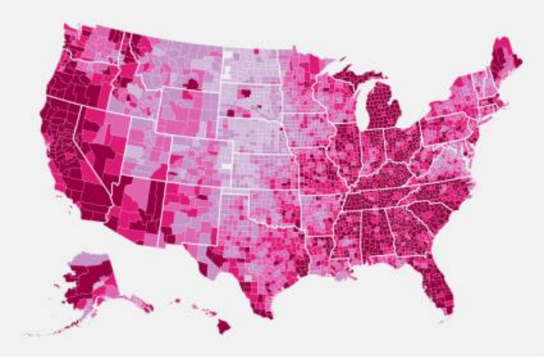
Code Demo

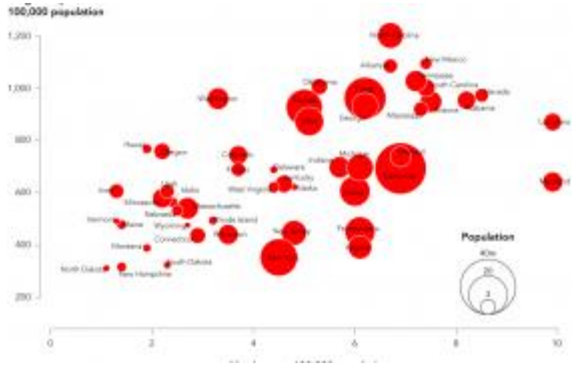
Beyond R and EDA

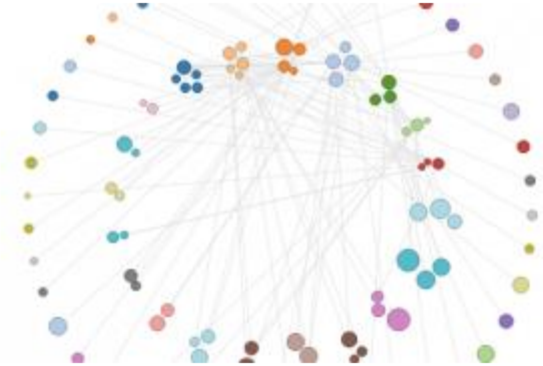
This is just the tip of the iceberg!

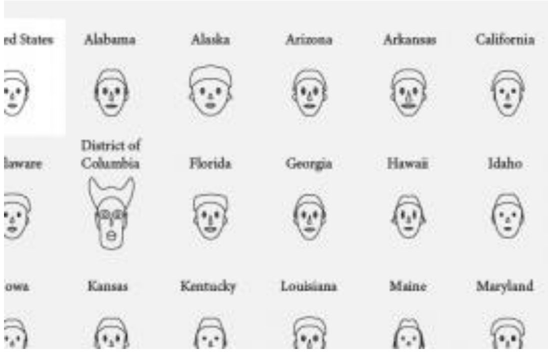


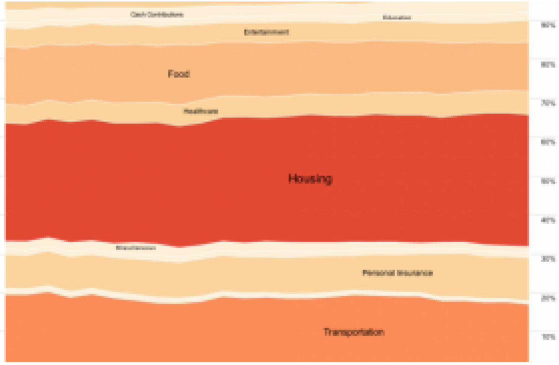
Advanced Visualizations with R







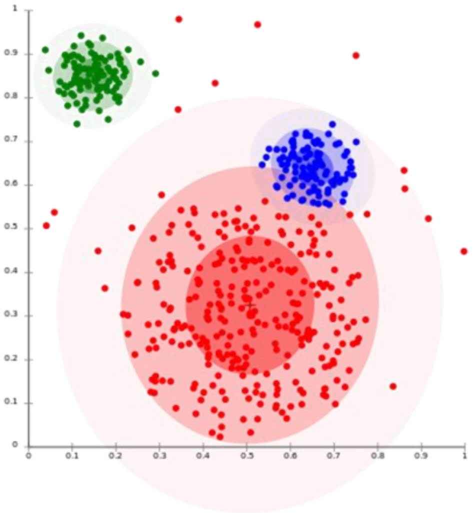


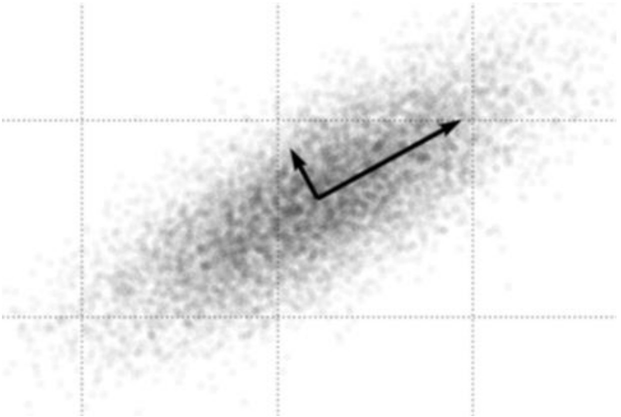


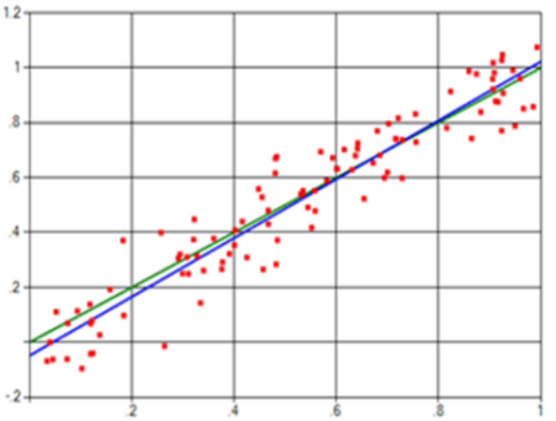
Source: Flowing Data

Advanced Data Analysis with R

- Cluster Analysis
- Statistical Modeling
- Dimensionality Reduction
- Analysis of Variance (ANOVA)

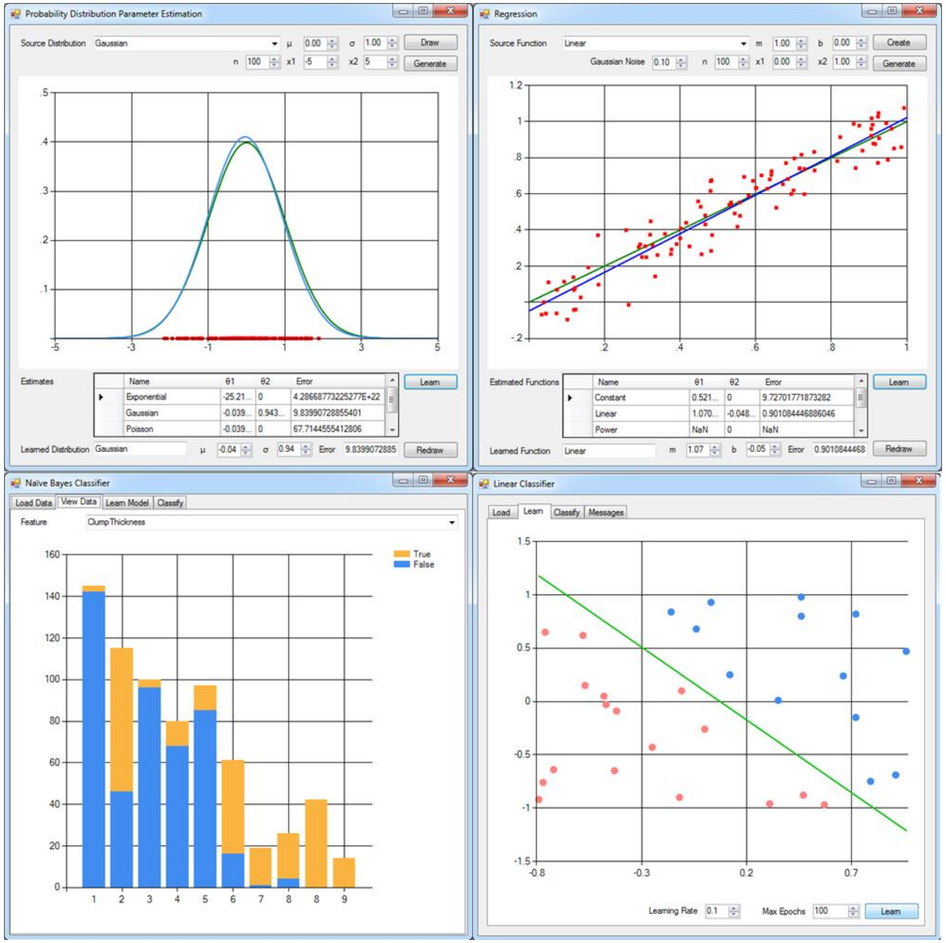






Data Mining and Machine Learning with R

- EDA uses human for pattern recognition
- Doesn’t scale well for higher dimensional data
- Need to use machines for pattern recognition
  - Data Mining
  - Machine Learning



Probability Distribution Parameter Estimation

Source Distribution: Gaussian; μ 0.00; σ 1.00; n 100; x1 -5; x2 5

Estimates

	Name	θ1	θ2	Error
▸	Exponential	-25.21...	0	4.28668773225277E+22
	Gaussian	-0.039...	0.943...	9.83990728855401
	Poisson	-0.039...	0	67.7144555412806

Learned Distribution: Gaussian; μ -0.04; σ 0.94; Error 9.8399072885

Regression

Source Function: Linear; m 1.00; b 0.00; Gaussian Noise 0.10; n 100; x1 0.00; x2 1.00

Estimated Functions

	Name	θ1	θ2	Error
▸	Constant	0.521...	0	9.72701771873282
	Linear	1.070...	-0.048...	0.901084446886046
	Power	NaN	0	NaN

Learned Function: Linear; m 1.07; b -0.05; Error 0.9010844468

Naive Bayes Classifier

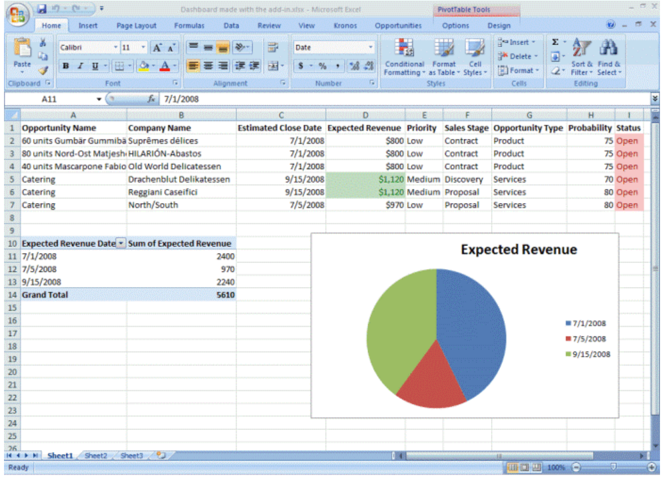
Feature: Clump Thickness

Linear Classifier

Learning Rate 0.1; Max Epochs 100

Alternatives to R for EDA

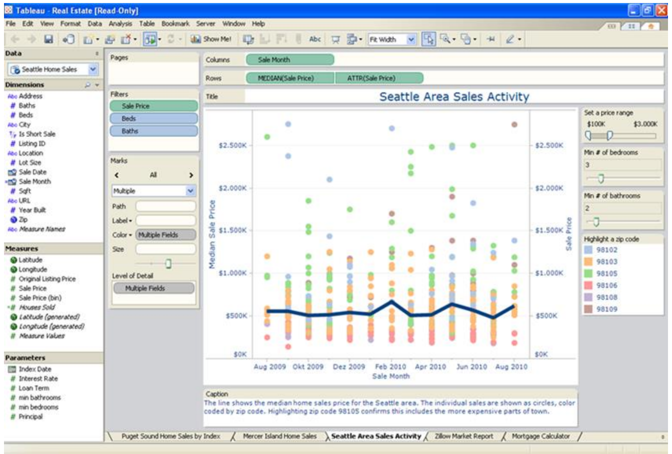
- Spreadsheets
- Interactive Data Visualization Tools
- Statistical Analysis Software
- Other Statistical Programming Languages
- General-Purpose Programming Languages



	A	B	C	D	E	F	G	H	I
1	Opportunity Name	Company Name	Estimated Close Date	Expected Revenue	Priority	Sales Stage	Opportunity Type	Probability	Status
2	60 units Gumbär Gummibä	Suprêmes délices	7/1/2008	$800	Low	Contract	Product	75	Open
3	80 units Nord-Ost Matjesh	HILARIÓN-Abastos	7/1/2008	$800	Low	Contract	Product	75	Open
4	40 units Mascarpone Fabio	Old World Delicatessen	7/1/2008	$800	Low	Contract	Product	75	Open
5	Catering	Drachenblut Delikatessen	9/15/2008	$1,120	Medium	Discovery	Services	70	Open
6	Catering	Reggiani Caseifici	9/15/2008	$1,120	Medium	Proposal	Services	80	Open
7	Catering	North/South	7/5/2008	$970	Low	Proposal	Services	80	Open

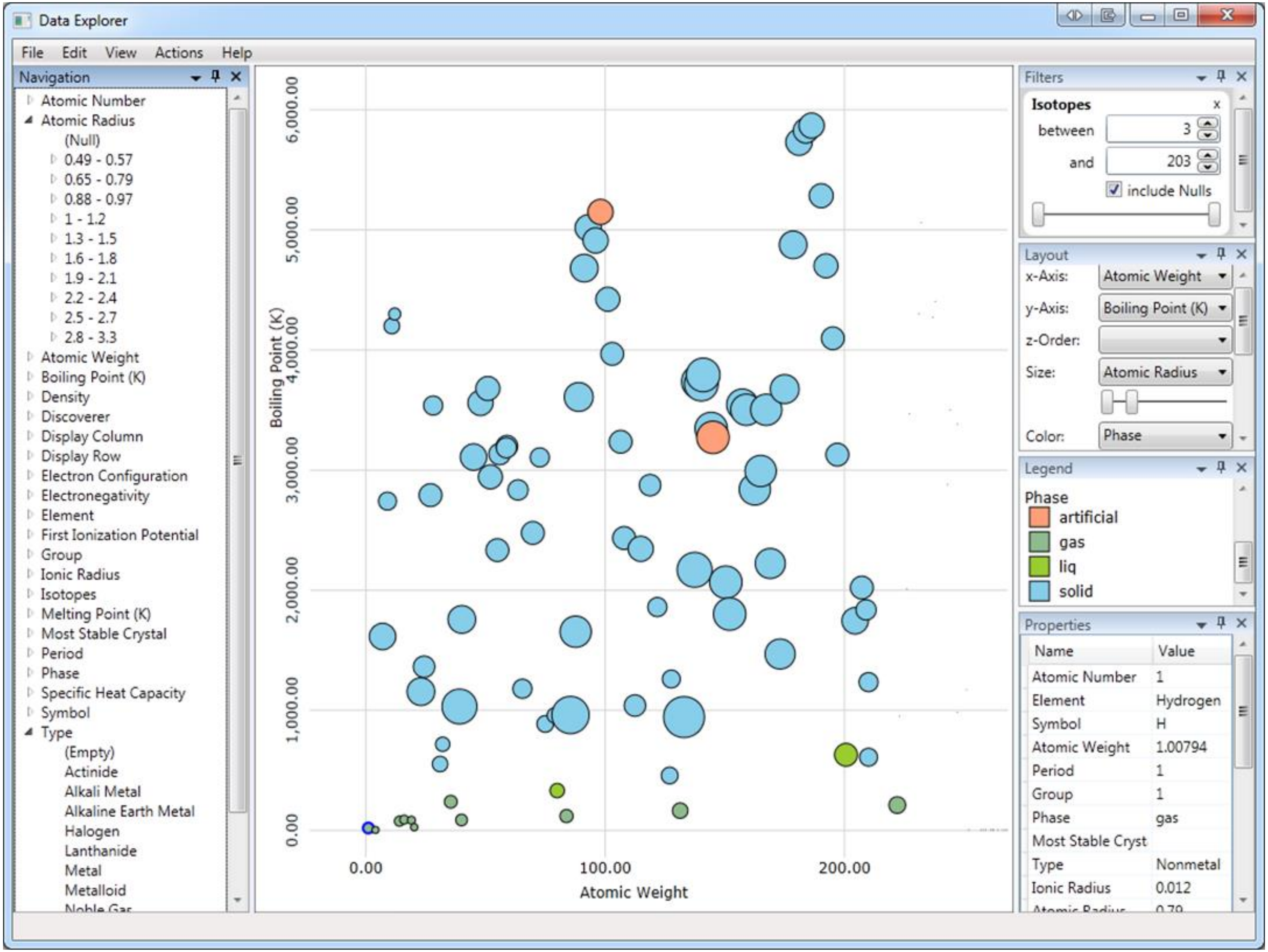
	Expected Revenue Date	Sum of Expected Revenue
11	7/1/2008	2400
12	7/5/2008	970
13	9/15/2008	2240
14	Grand Total	5610

Source: Microsoft



Caption
The line shows the median home sales price for the Seattle area. The individual sales are shown as circles, color coded by zip code. Highlighting zip code 98105 confirms this includes the more expensive parts of town.

Source: Tableau



Filters

Isotopes x

between 3

and 203

☑ include Nulls

Layout

x-Axis: Atomic Weight

y-Axis: Boiling Point (K)

z-Order:

Size: Atomic Radius

Color: Phase

Legend

Phase

🟧 artificial

🟩 gas

🟩 liq

🟦 solid

Properties

Name	Value
Atomic Number	1
Element	Hydrogen
Symbol	H
Atomic Weight	1.00794
Period	1
Group	1
Phase	gas
Most Stable Cryst	
Type	Nonmetal
Ionic Radius	0.012
Atomic Radius	0.79

Code Demo

Where to Go Next…

- R website: http://www.cran.r-project.org
- R Studio: http://www.rstudio.com
- Coursera: https://www.coursera.org/
- Revolutions: http://blog.revolutionanalytics.com/
- Flowing Data: http://flowingdata.com
- R-Blogger: http://www.r-bloggers.com/

- R Quick Reference Card: http://cran.r-project.org/doc/contrib/Short-refcard.pdf

Conclusion

Conclusion

- R is a very popular language for data analysis
- EDA can provide rapid understanding of data
- R + EDA = Powerful insight into your data!

Feedback

- Feedback is very important to me
- Specific feedback I’m looking for:
  - One thing you liked about the presentation
  - One thing you think I could improve on

Contact Info

Matthew Renze
matthew@renzeconsulting.com

Renze Consulting
www.renzeconsulting.com

Data Explorer
http://www.data-explorer.com