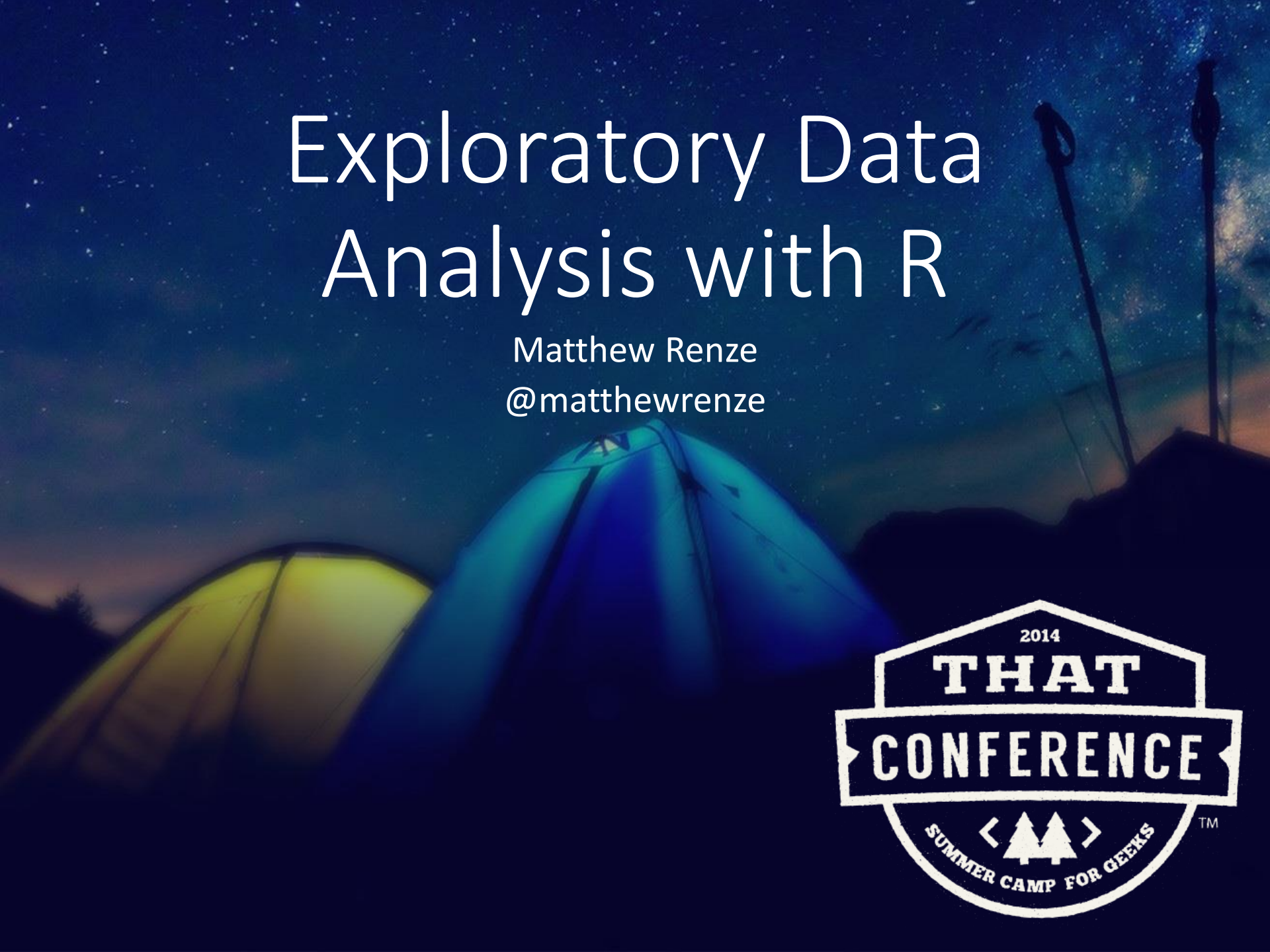
Exploratory Data Analysis with R

Matthew Renze
@matthewrenze



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 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
  - R
  - Exploratory Data Analysis (EDA)
- Exploratory Data Analysis with R
  - Data Munging
  - Descriptive Statistics
  - Data Visualization
- Beyond R and EDA
- Q & A

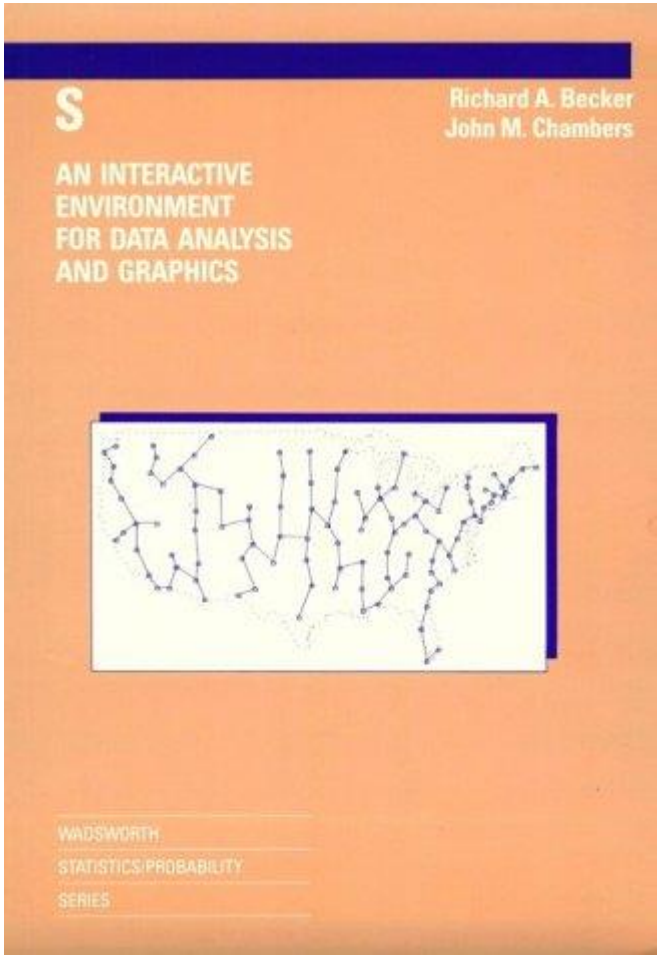
Introduction to R

What is R?

- R is an open source implementation of S

What is S?

- Statistical programming language
- Developed at Bell Labs in 1976
- Originally implemented in Fortran
- Later rewritten in C
- Currently owned by TIBCO Software



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:

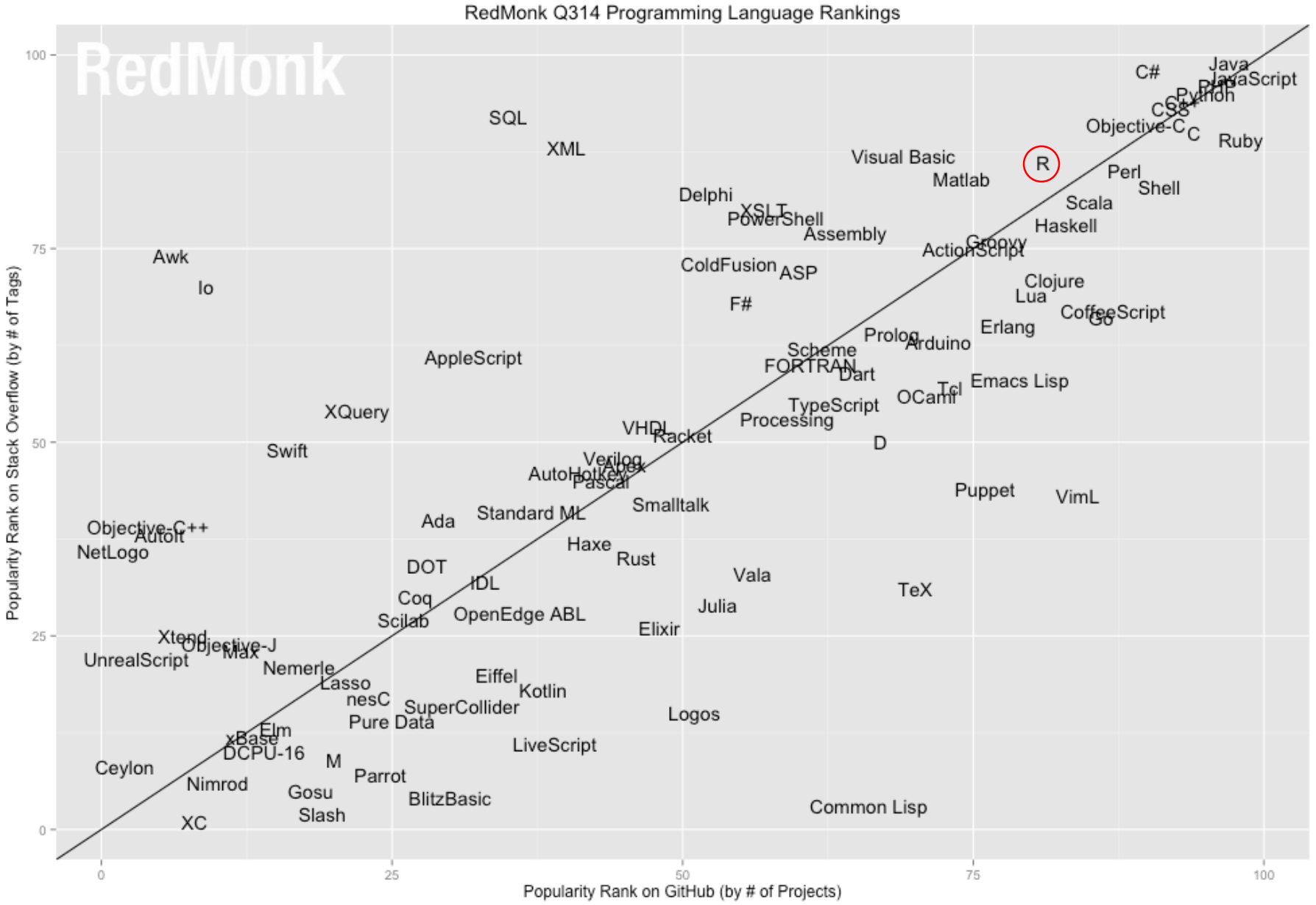
- an open source implementation of S
- a language and an environment
- provides methods for both statistical and graphical data analysis
- runs on Windows, Mac, and Unix systems

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
- is free (as in beer… and as in speech)



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

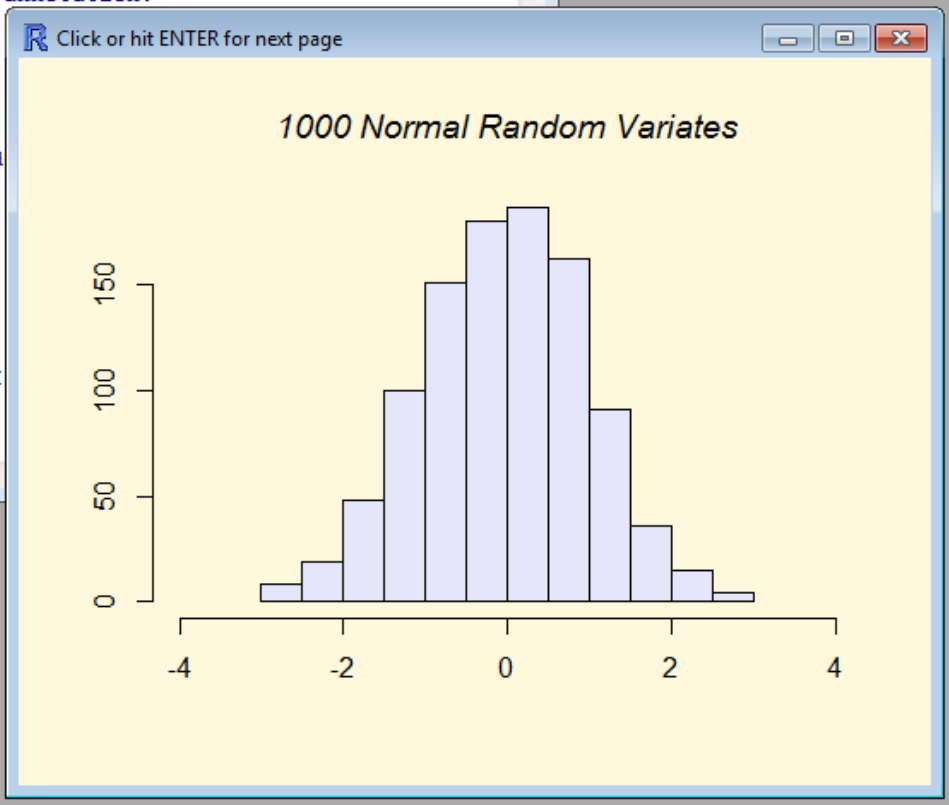
RGui (64-bit)

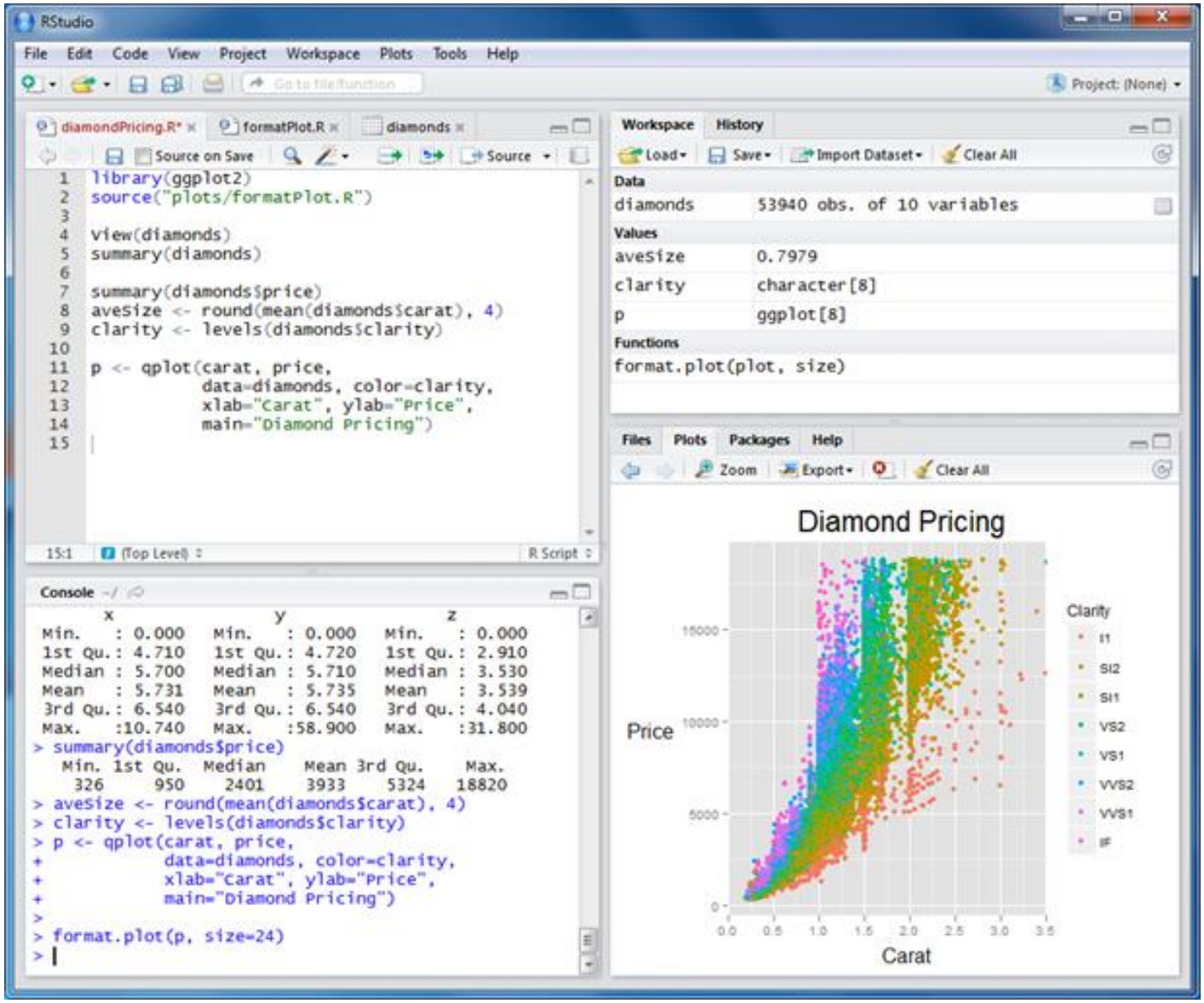
Next

R Console

```
> 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...
```

Click or hit ENTER for next page





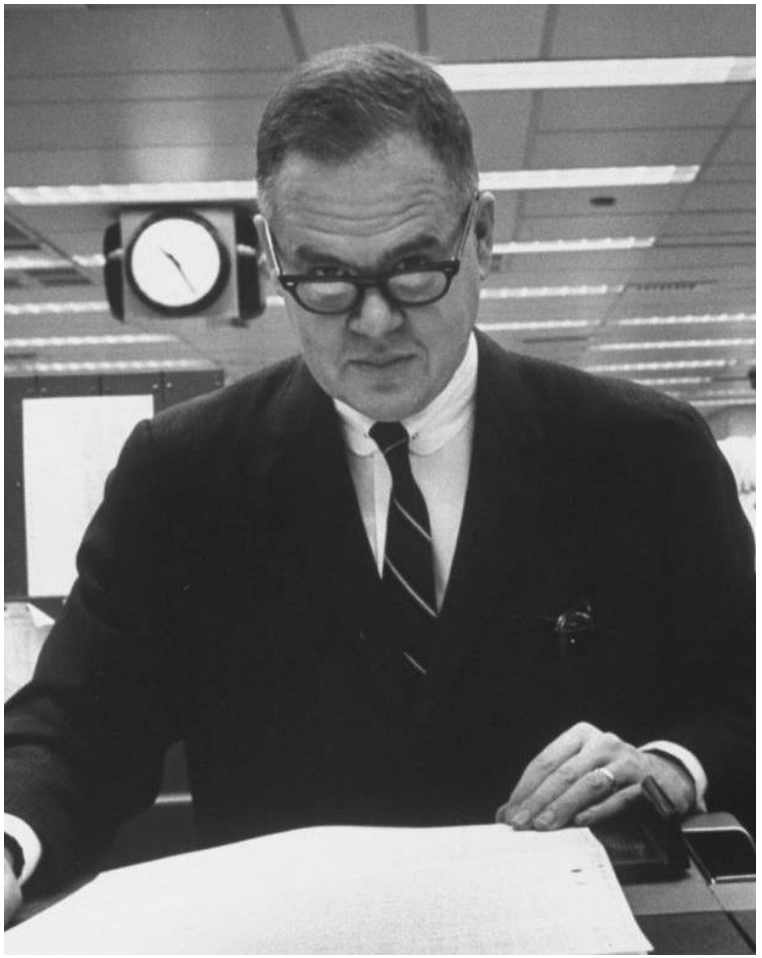
Source: www.rstudio.com/ide/

Code Demo

Exploratory Data Analysis

Exploratory Data Analysis (EDA)

- One of many approaches to data analysis
- Objectives:
  - Discover patterns
  - Identify anomalies
  - Suggest hypotheses
  - Check assumptions
- Promoted by John Tukey

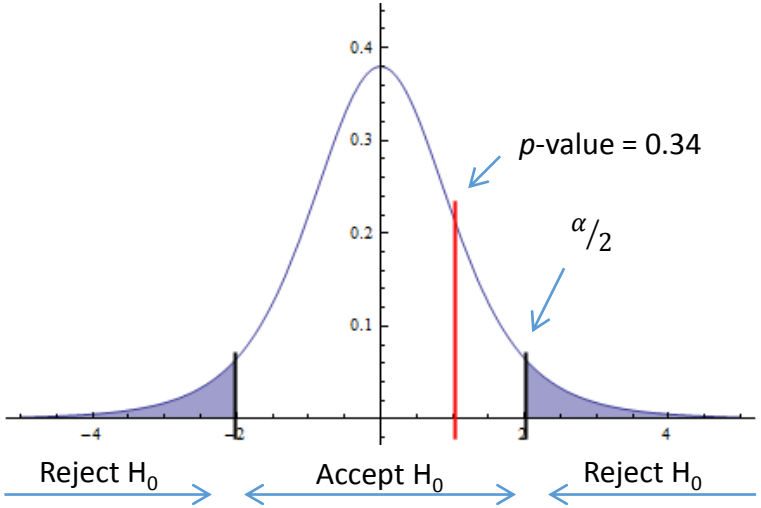


Source: Time Magazine

Confirmatory vs. Exploratory Data Analysis

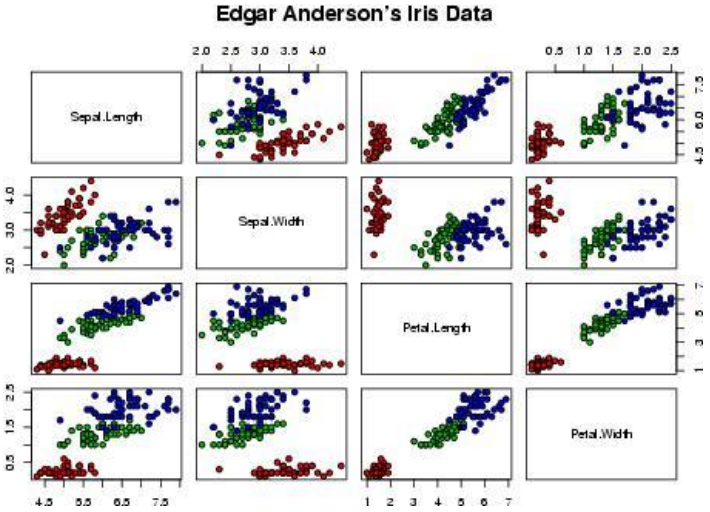
Confirmatory

- Start with hypothesis
- Test the null hypothesis
- Uses statistical models



Exploratory

- No hypothesis at first
- Generating hypothesis
- Uses graphical methods



Problems with EDA

- Not rigorous like formal statistical methods
  - Susceptible to specific types of statistical biases
- Not useful for inference or prediction
- Not efficient for massive data sets

Note: There are ways to avoid some of these issues

Don’t Practice Statistics without a License!

Data Munging

Data Munging

- Transforming data from a raw form to a usable form
- aka: Data cleaning or data wrangling
- Many data sets are not initially ready for data analysis
- Data must be transformed or cleaned first in order to be analyzed

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
- 14 measures include:
  - Amount Requested
  - Amount Funded
  - Interest Rate
  - Monthly Income
  - FICO Score
- *Problem:* The data are not in a digestible format
- *Goal:* Prepare the data for analysis

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

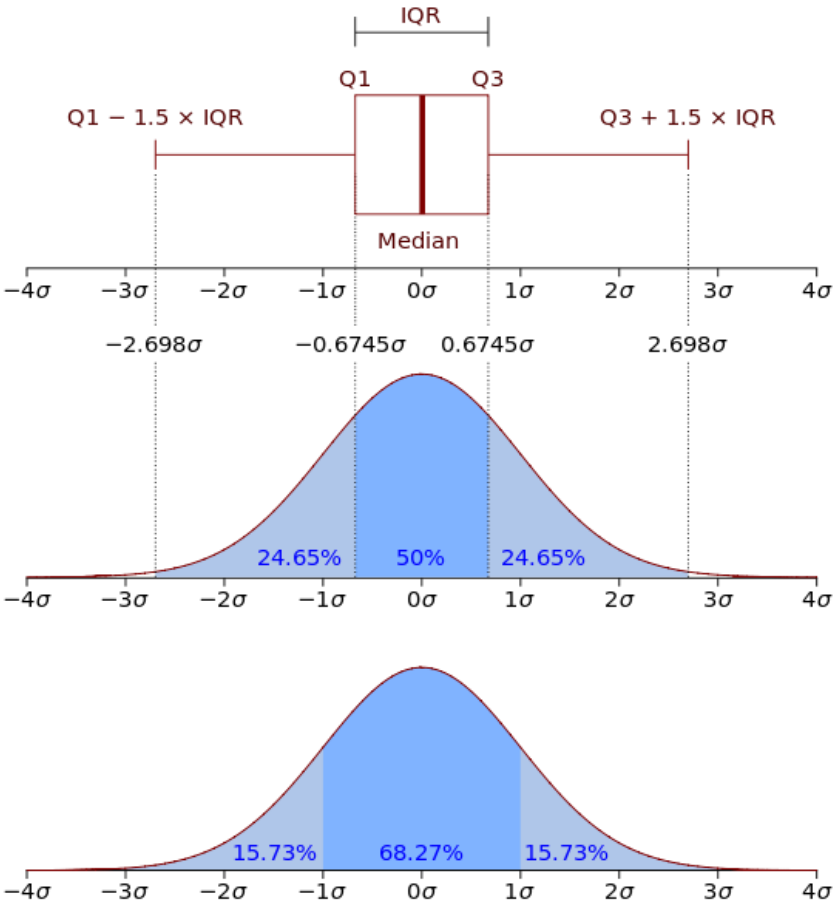
Statistical Terms

- 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

Univariate Analysis

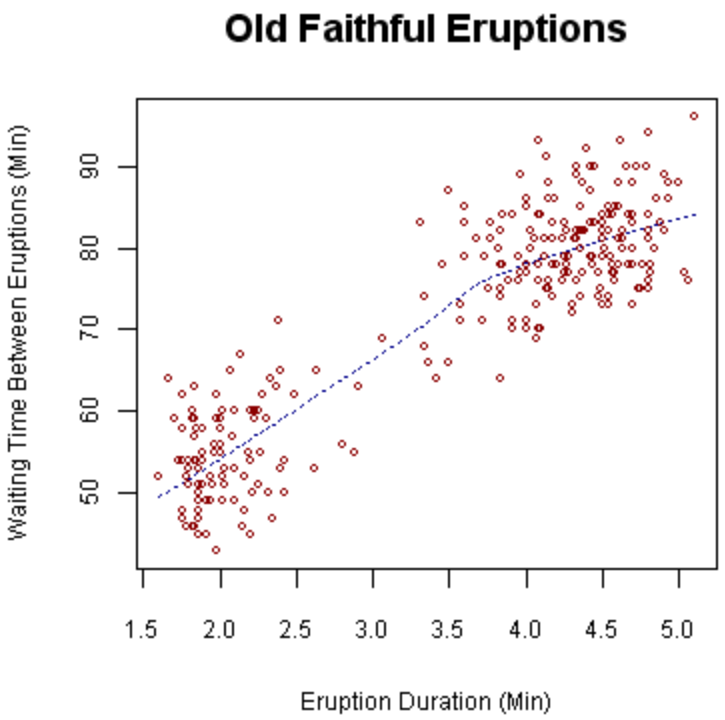
- Analysis of a single variable
- Measures include:
  - Central tendency
    - Mean
    - Median
    - Mode
  - Dispersion
    - Min
    - Max
    - Range
    - Quartiles
    - Variance
    - Standard deviation



Source: Wikipedia

Bivariate Analysis

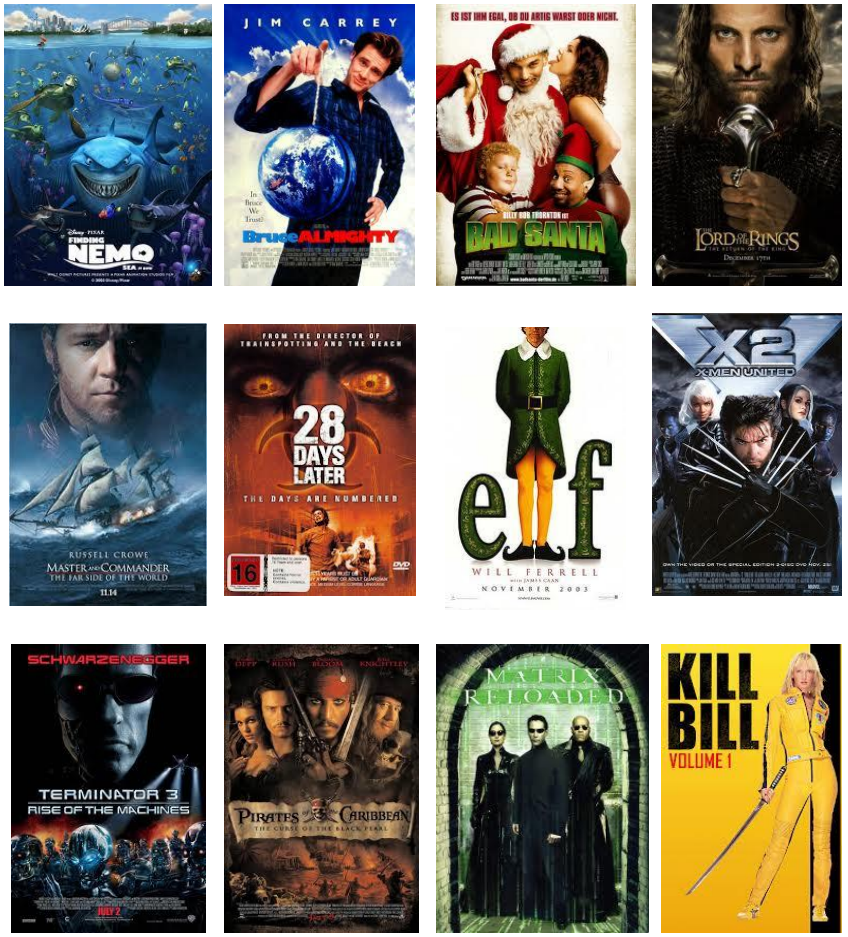
- Analysis of the relationship between two variables
  - Predictor
  - Outcome
- Measures include:
  - Covariance
  - Correlation



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)
  - Rotten Tomatoes Score
  - Box Office Revenue ($)
- *Goal:* Determine what types of movies made the most money



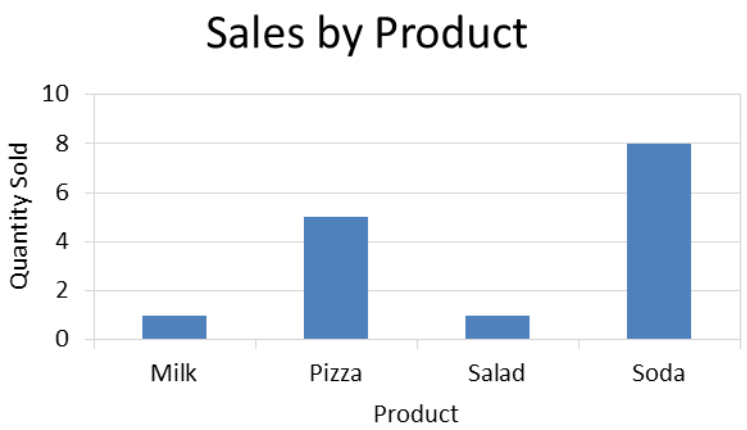
Source: http://www.rossmanchance.com/iscam2/files.html

Code Demo

Data Visualization

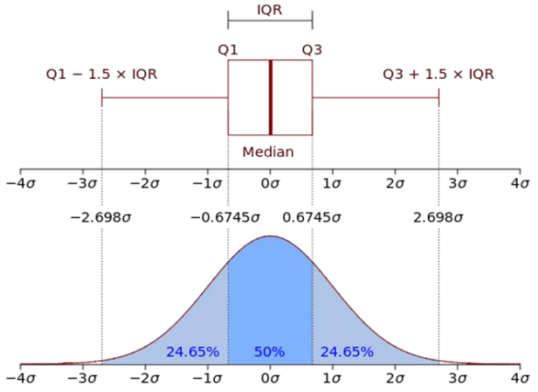
Data Visualization

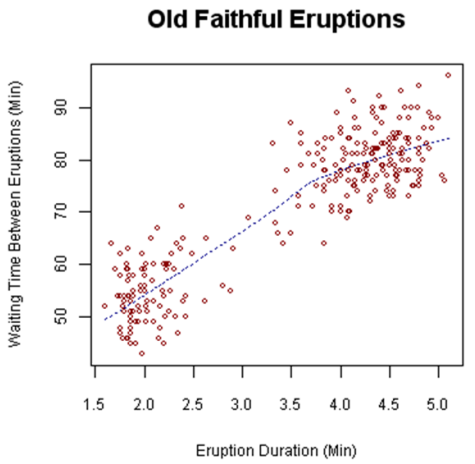
- Representation of data via visual means
- Human brain is exceptionally good at visual pattern recognition
- Map dimensions of data to visual characteristics:
  - Location
  - Size
  - Color
  - Shape



Types of Data Visualizations

- Several types of data visualizations based on:
  - Type of Variable(s)
    - Qualitative (Categorical)
    - Quantitative (Numerical)
  - Number of Variables
    - Univariate
    - Bivariate
    - Multivariate

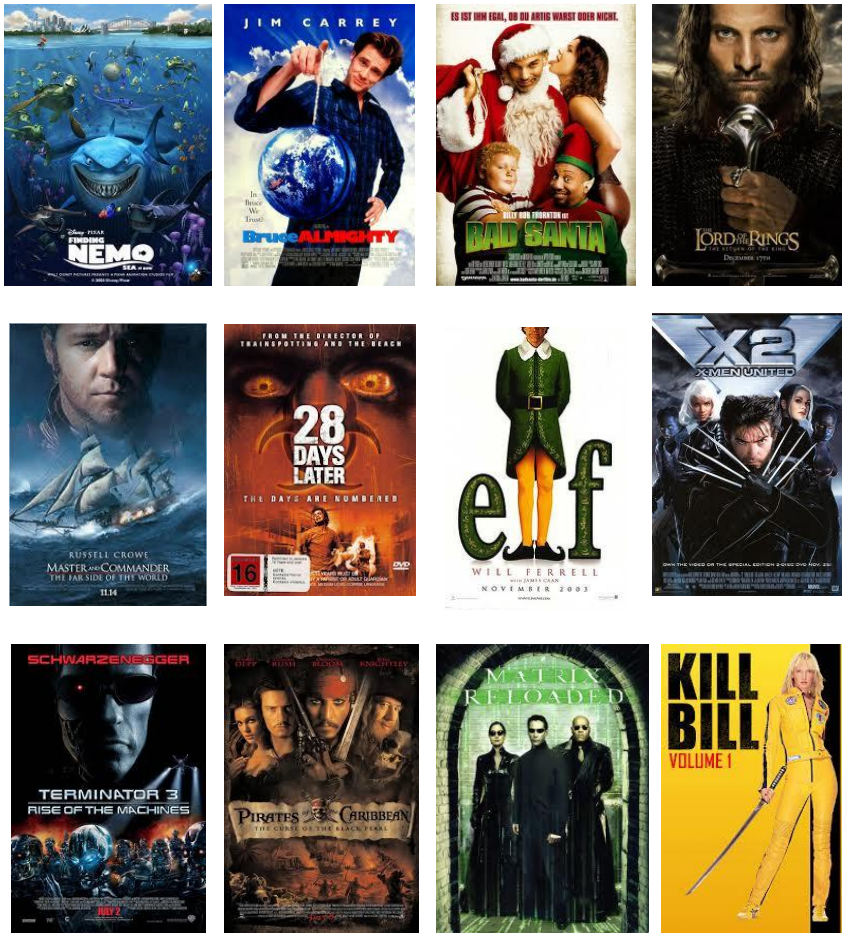




Source: Wikipedia

Code Demo: Movies Data Set

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

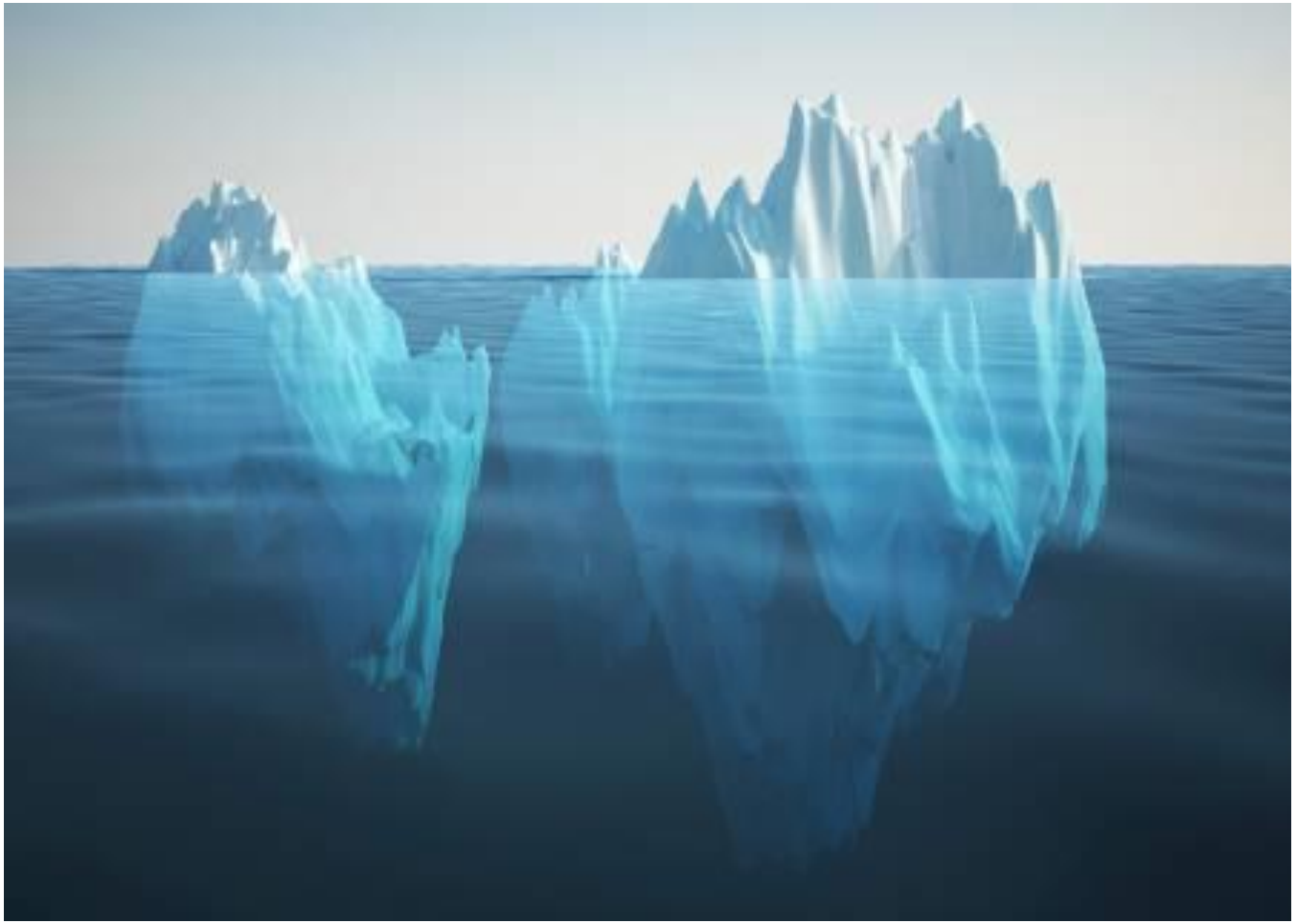


Source: http://www.rossmanchance.com/iscam2/files.html

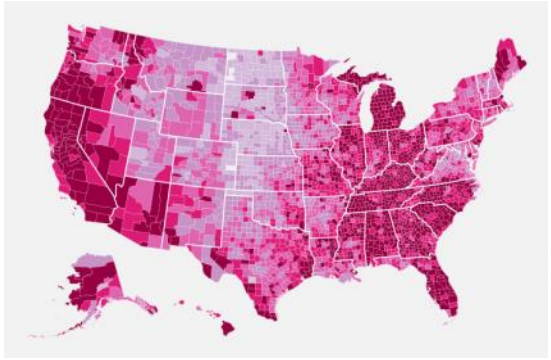
Code Demo

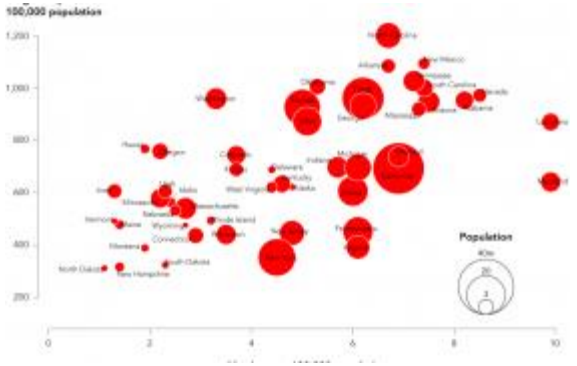
Beyond R and EDA

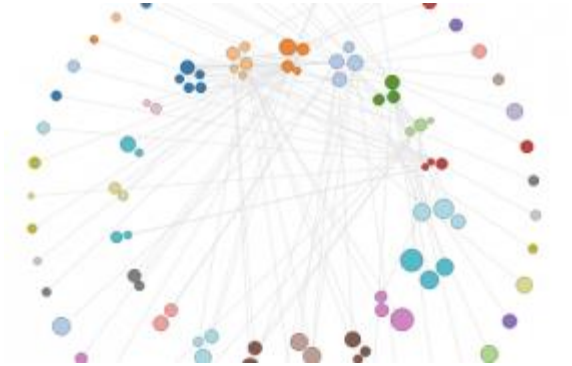
This is just the tip of the iceberg!

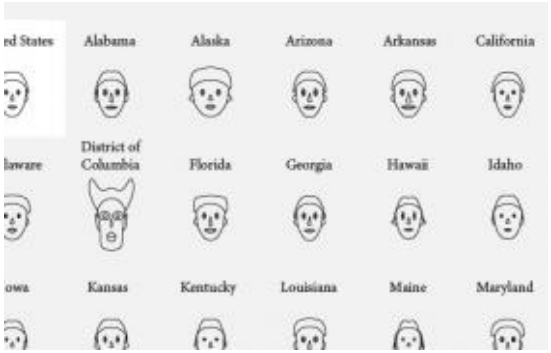


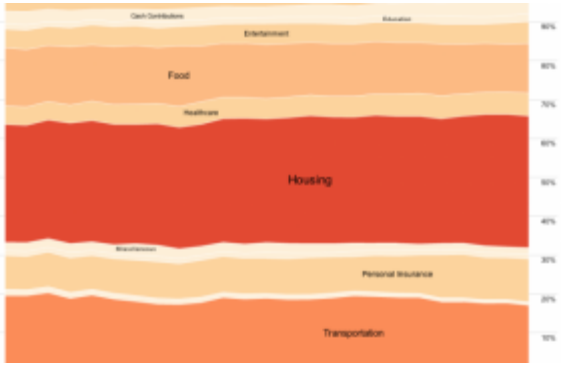
Advanced Visualizations with R







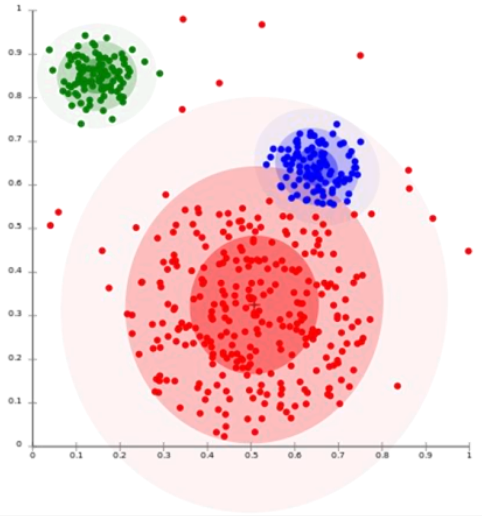


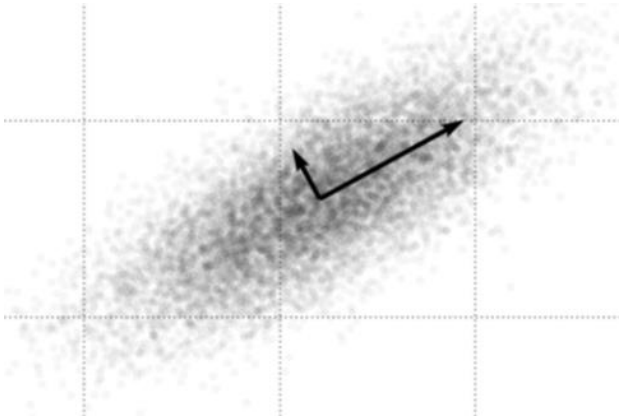


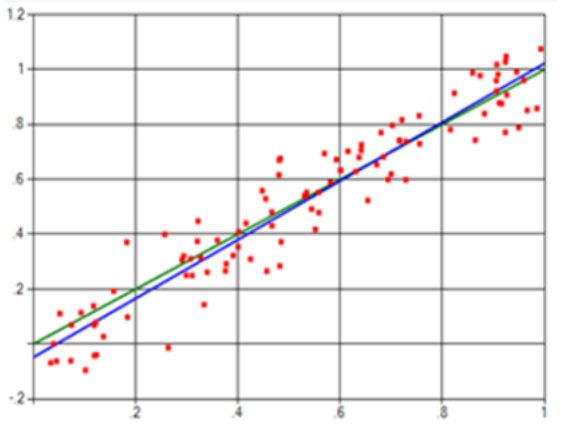
Source: Flowing Data

Advanced EDA with R

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

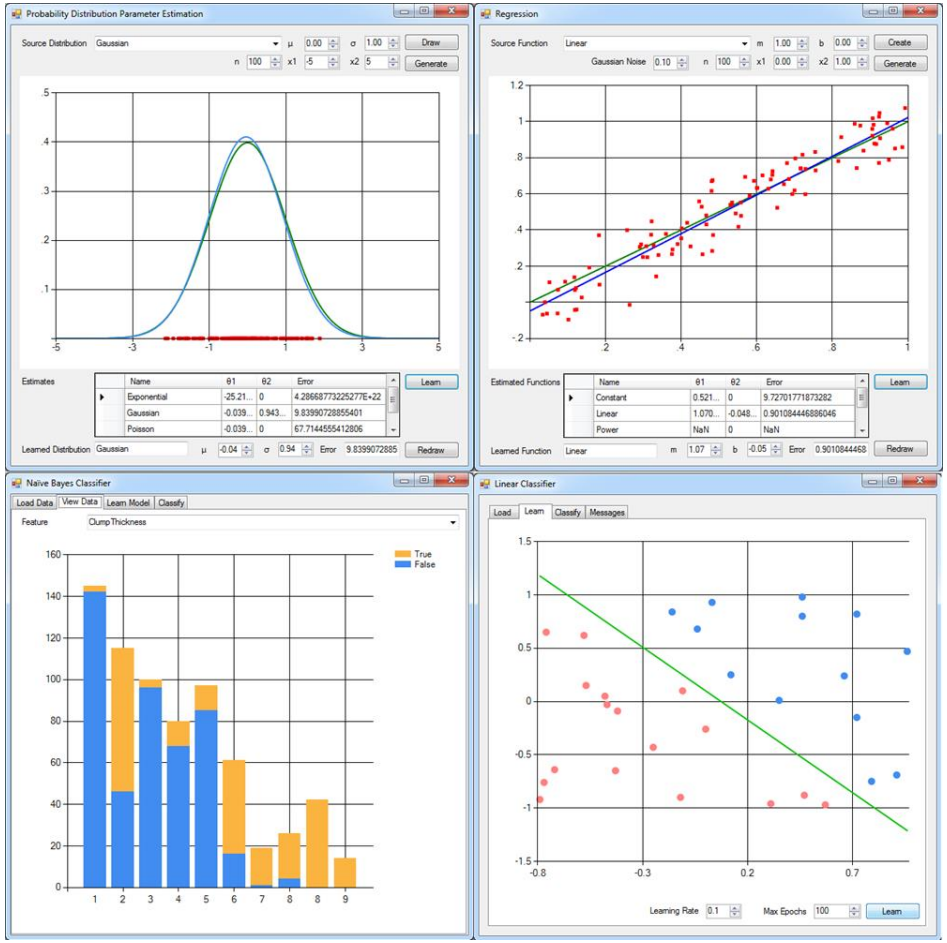






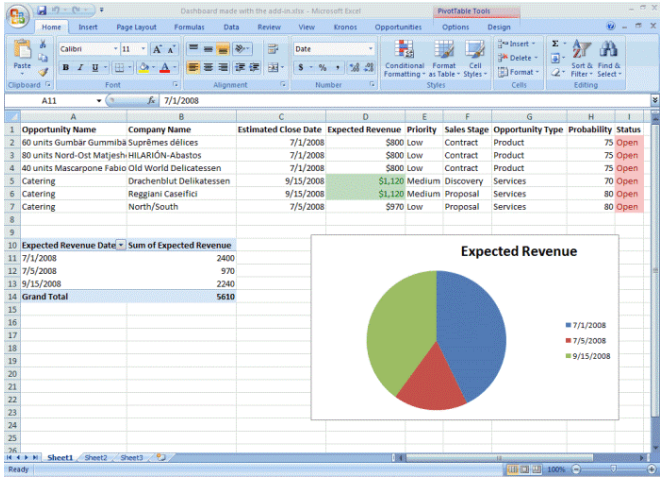
Machine-based EDA with R

- 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

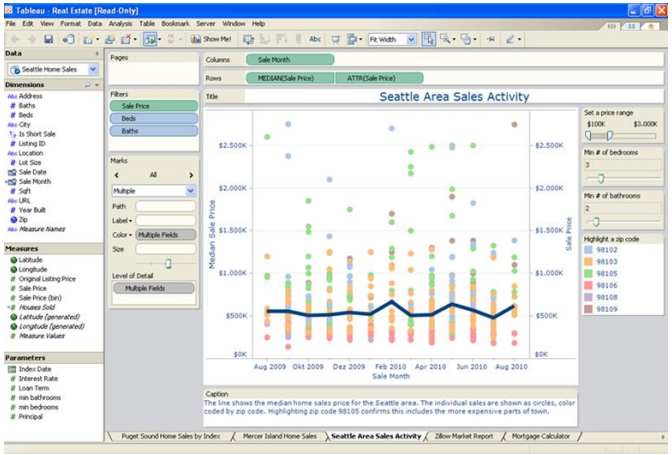


Alternatives to R for EDA

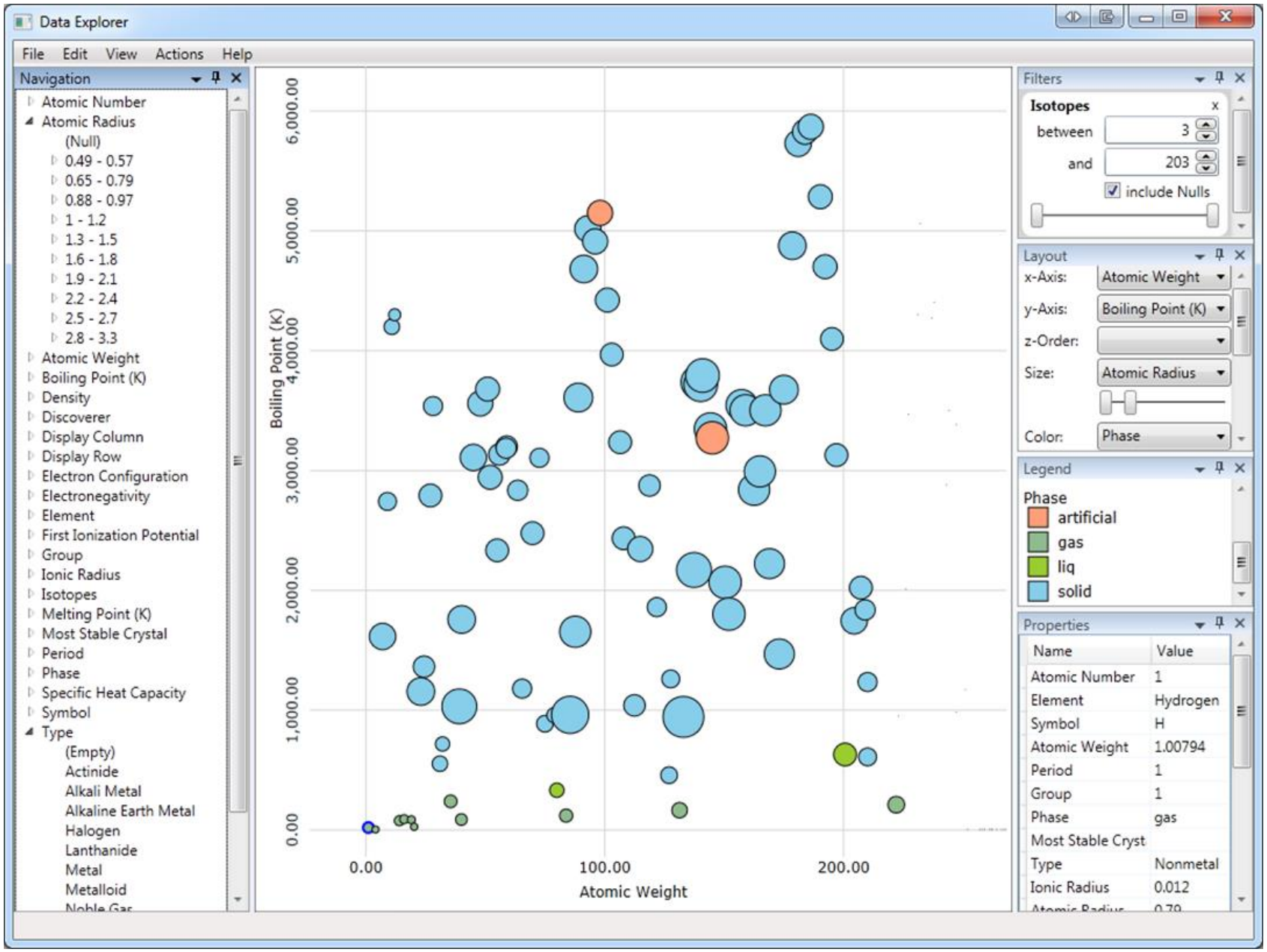
- Spreadsheets
- Interactive Data Visualization Tools
- Statistical Analysis Software
- Other Statistical Programming Languages
- General-Purpose Programming Languages



Source: Microsoft



Source: Tableau


Data Explorer
File
Edit
View
Actions
Help
Navigation
Atomic Number
Atomic Radius
(Null)
0.49 - 0.57
0.65 - 0.79
0.88 - 0.97
1 - 1.2
1.3 - 1.5
1.6 - 1.8
1.9 - 2.1
2.2 - 2.4
2.5 - 2.7
2.8 - 3.3
Atomic Weight
Boiling Point (K)
Density
Discoverer
Display Column
Display Row
Electron Configuration
Electronegativity
Element
First Ionization Potential
Group
Ionic Radius
Isotopes
Melting Point (K)
Most Stable Crystal
Period
Phase
Specific Heat Capacity
Symbol
Type
(Empty)
Actinide
Alkali Metal
Alkaline Earth Metal
Halogen
Lanthanide
Metal
Metalloid
Noble Gas
Boiling Point (K)
6,000.00
5,000.00
4,000.00
3,000.00
2,000.00
1,000.00
0.00
0.00
100.00
200.00
Atomic Weight
Filters
Isotopes
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

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!

Thank You to Our Sponsors!

inrule TECHNOLOGY®

API HEALTHCARE

WordPress.com

PayPal + Braintree Developer

SKYLINE TECHNOLOGIES

SafeNet

DevExpress

omni resources

corvisa cloud

TechSmith

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

@matthewrenze

Renze Consulting

www.renzeconsulting.com