Exploratory Data Analysis with R

Matthew Renze
@matthewrenze
#hdc15

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

The New York Times

For Today's Graduate, Just One Word: Statistics

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

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

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

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]

Introduction to R

What is R?

- Open source
- Implementation of S
- Language and environment
- Numerical and graphical analysis
- Cross platform

Source: www.r-project.org

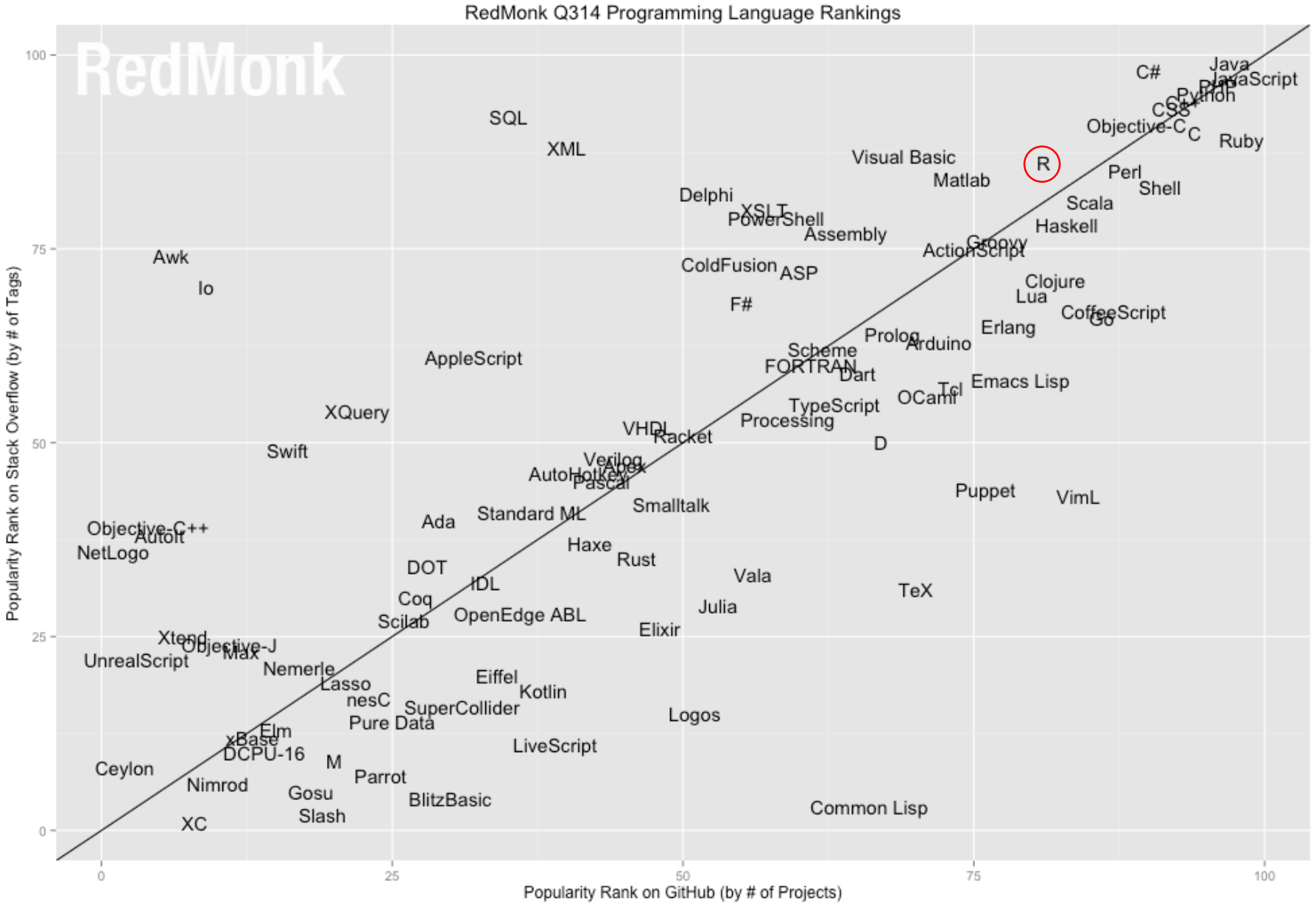
What is R?

- Active development
- Large user community
- Modular and extensible
- 6700+ extensions

and best of all…


FREE


FREE



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

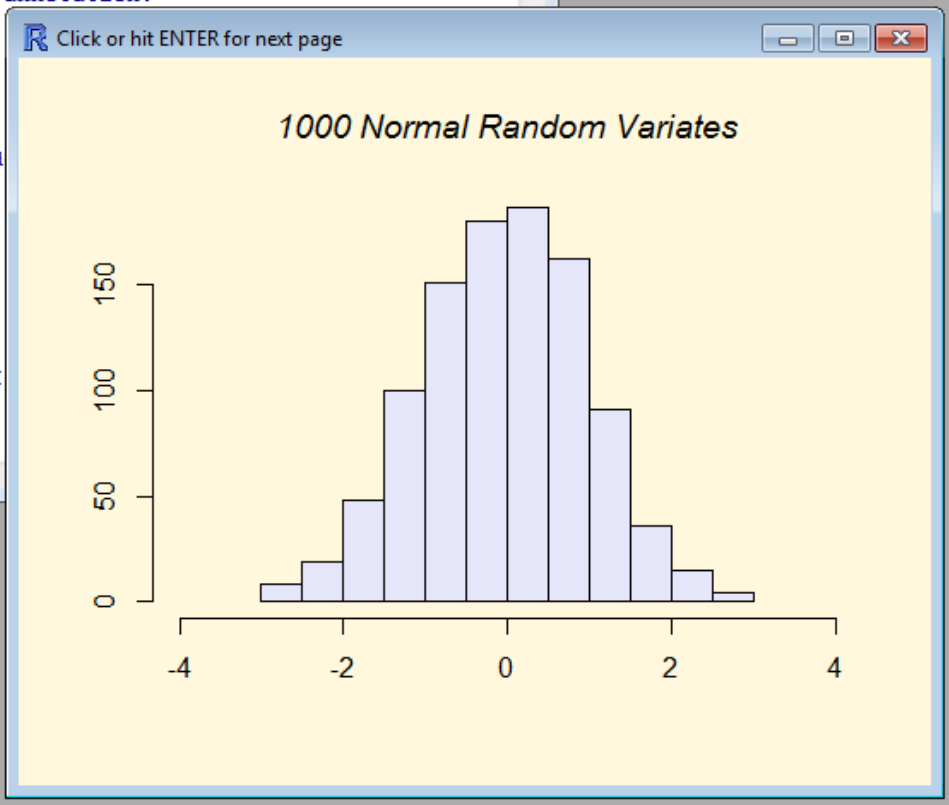
RGui (64-bit)

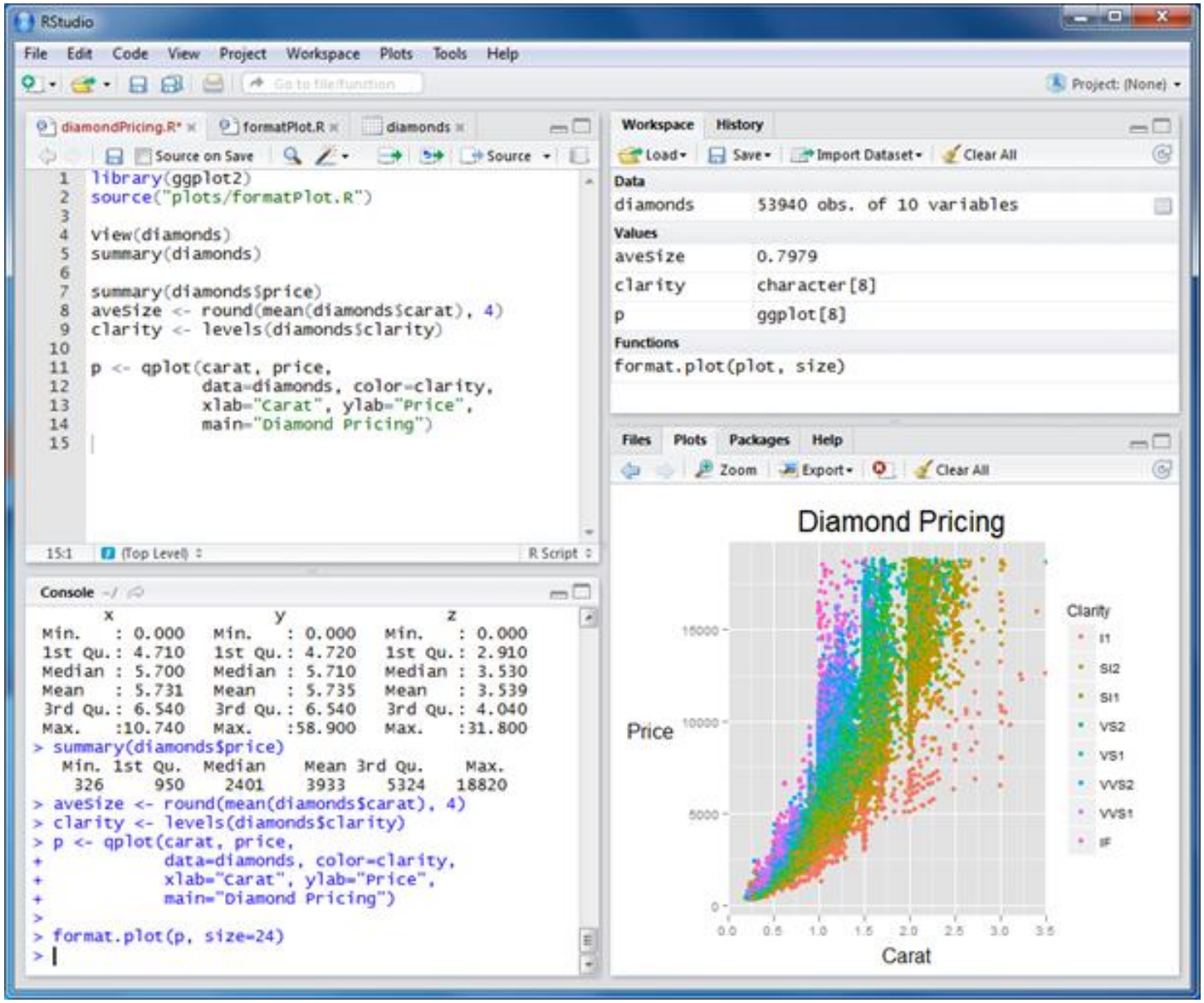
Next

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

Data Munging

Data Munging

- Transforming data
- Raw data to usable data
- Data must be 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

- File-based data
- Web-based data
- Databases
- Statistical data
- And many more…

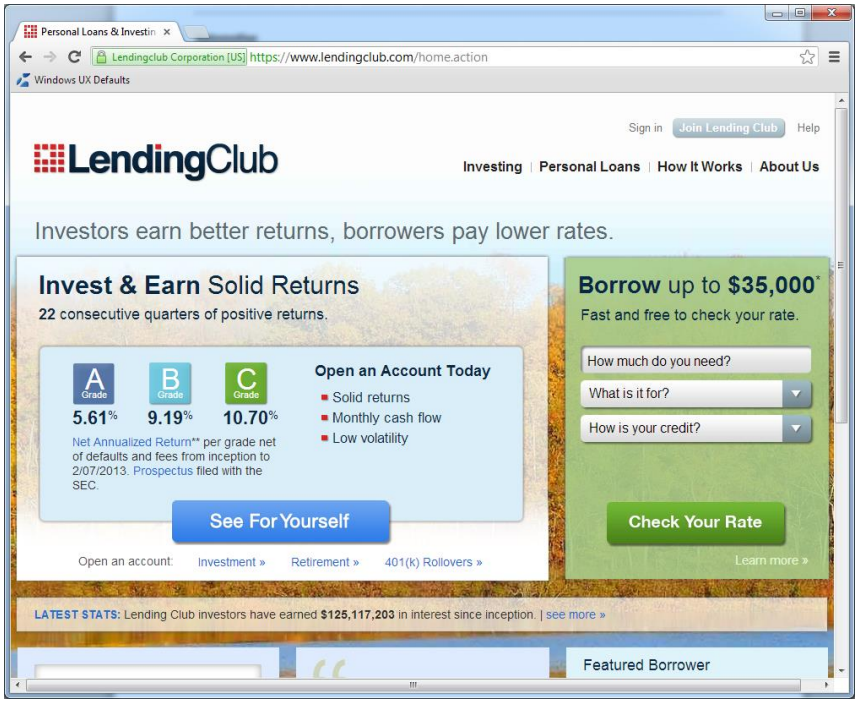
Cleaning Data

- This step is often the:
  - Most difficult
  - Most time consuming
- TIP: Record all steps

Source: Wikimedia

Code Demo: Lending Club Dataset

- Peer-to-peer loans
- *Problem:* Data are not ready for analysis
- *Goal:* Prepare the data for analysis



Source: www.lendingclub.com

Code Demo

Descriptive Statistics

Descriptive Statistics

- Describe data
- Provides a summary
- 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

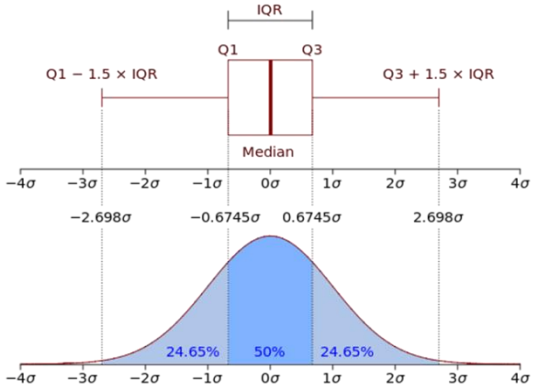
Statistical Terms

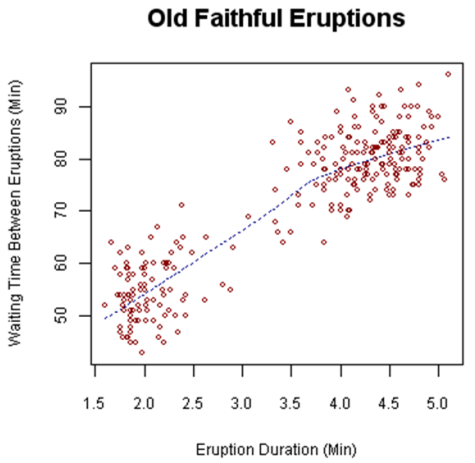
- Observations
- Variables
- Qualitative variable
- Quantitative variable

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

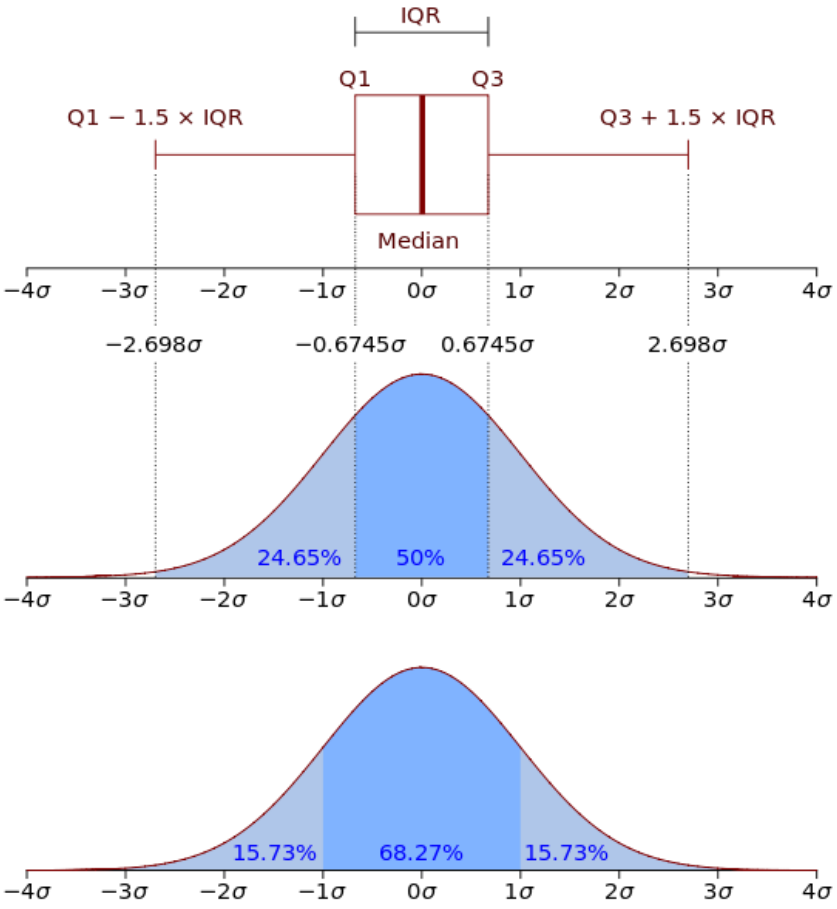
- Type of variables
  - Qualitative
  - Quantitative
- Number of variables
  - Univariate
  - Bivariate
  - Multivariate





Univariate Analysis

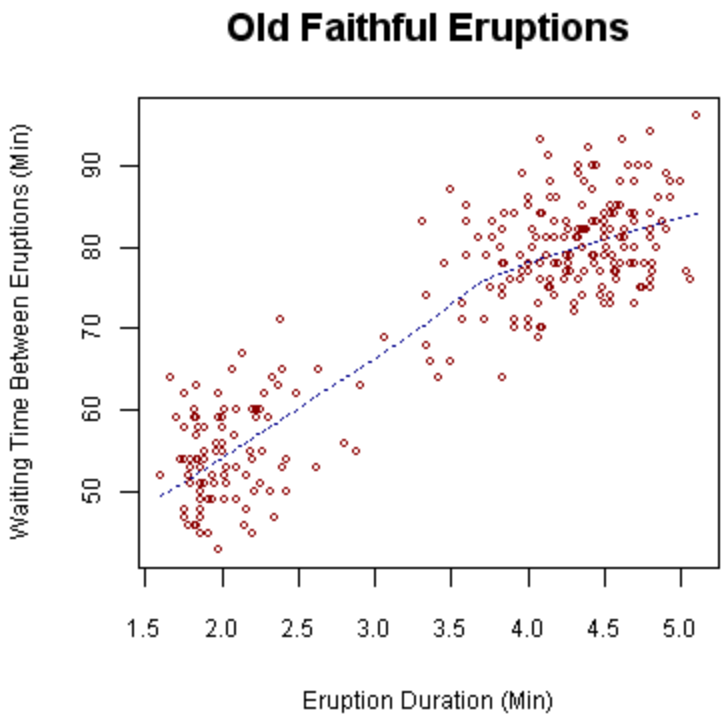
- One variable
- Measures include:
  - Central tendency
  - Dispersion



Source: Wikipedia

Bivariate Analysis

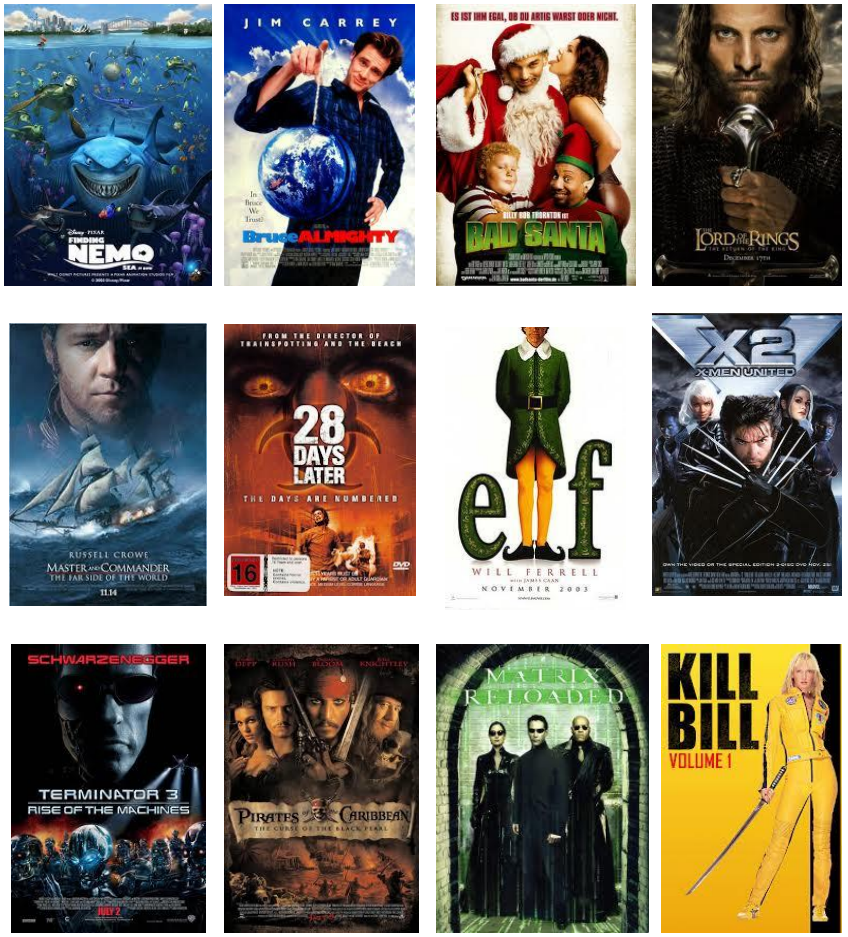
- Two variables
  - Predictor
  - Outcome
- Measures include
  - Covariance
  - Correlation



Source: Wikipedia

Code Demo: Movies Data Set

- Movies from 2003
- *Goal:* What movies made the most money



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

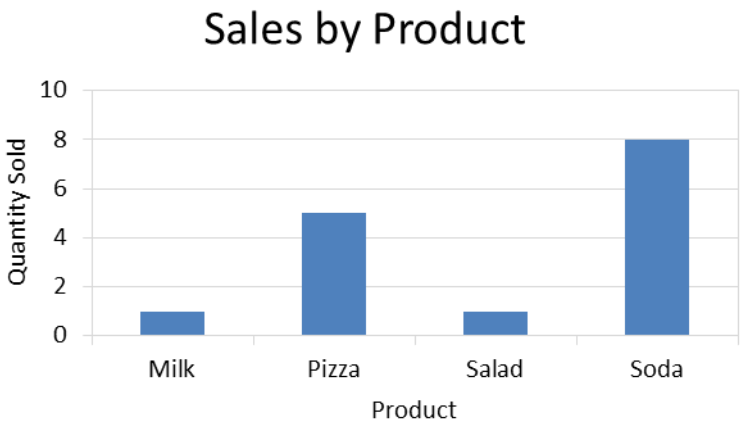
Code Demo

Data Visualization

Data Visualization

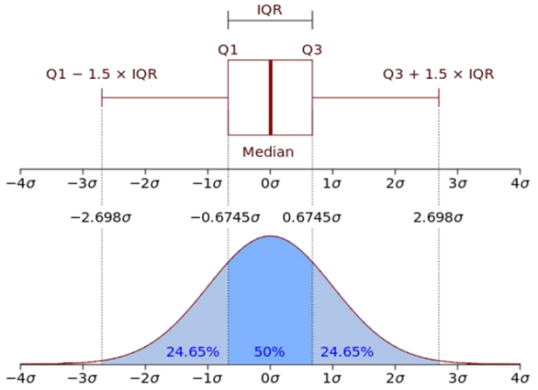
- Visual data representation
- For human pattern recognition
- Map dimensions to visual characteristics

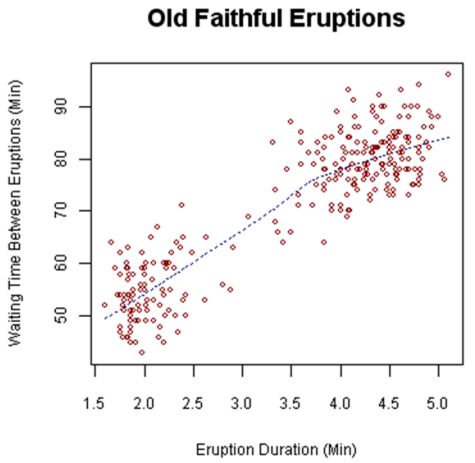
ID	Date	Customer	Product	Quantity
1	2012-10-27	John	Pizza	2
2	2012-10-27	John	Soda	2
3	2012-10-27	Jill	Salad	1
4	2012-10-27	Bob	Milk	1
5	2012-10-28	Sue	Soda	3
6	2012-10-28	Bob	Pizza	2
7	2012-10-28	Jill	Pizza	1
8	2012-10-28	Jill	Soda	3



Types of Data Visualizations

- Type of variable(s)
  - Qualitative
  - Quantitative
- Number of variables
  - Univariate
  - Bivariate
  - Multivariate

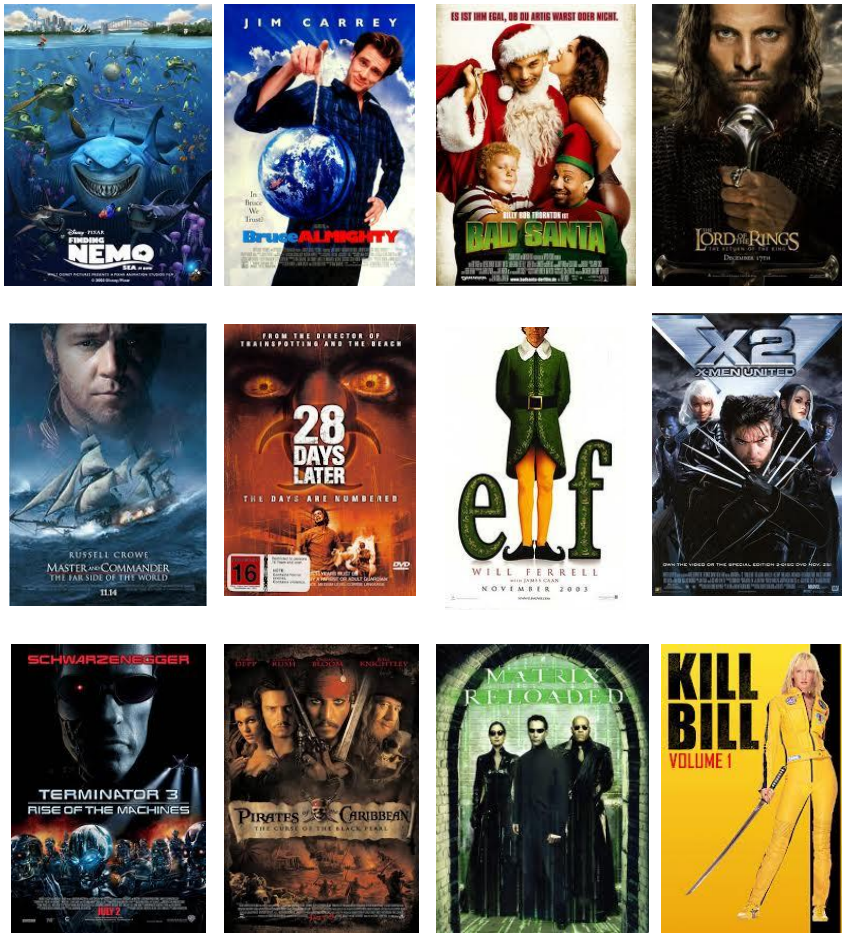




Source: Wikipedia

Code Demo: Movies Data Set

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



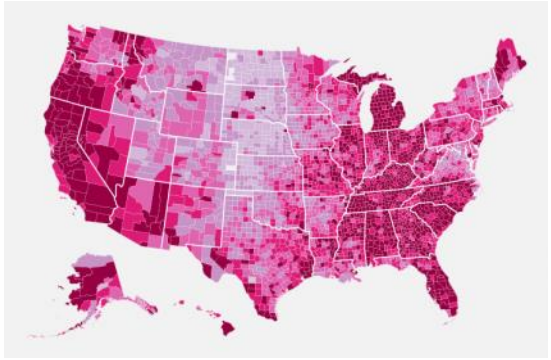
Source: http://www.rossmanchance.com/iscam2/files.html

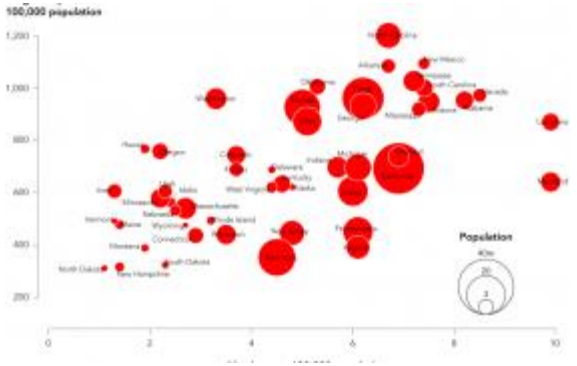
Code Demo

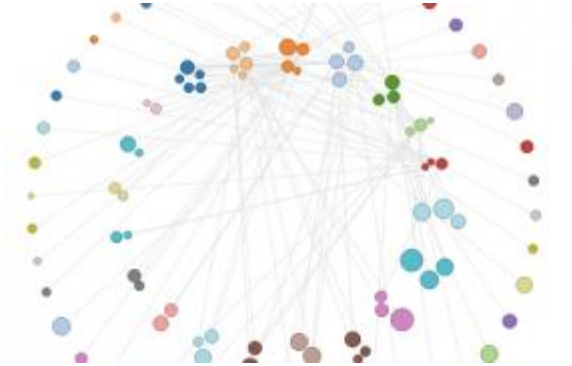
Beyond R and EDA

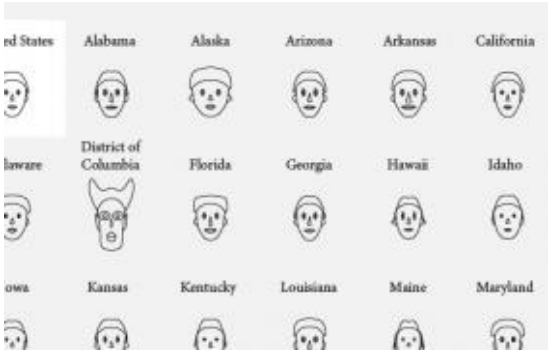

This is just the tip of the iceberg!

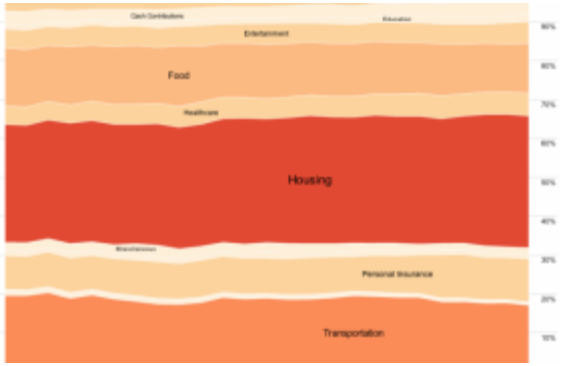
Advanced Visualizations with R







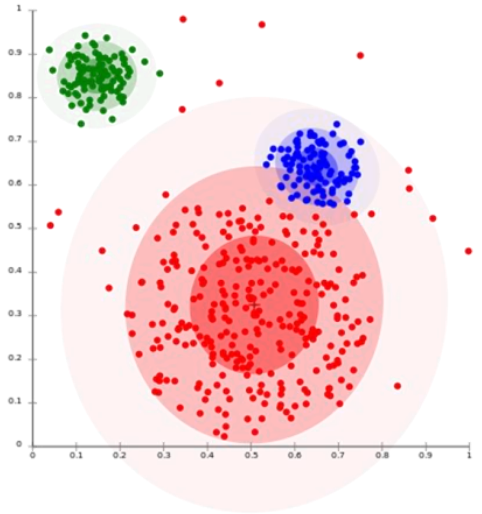


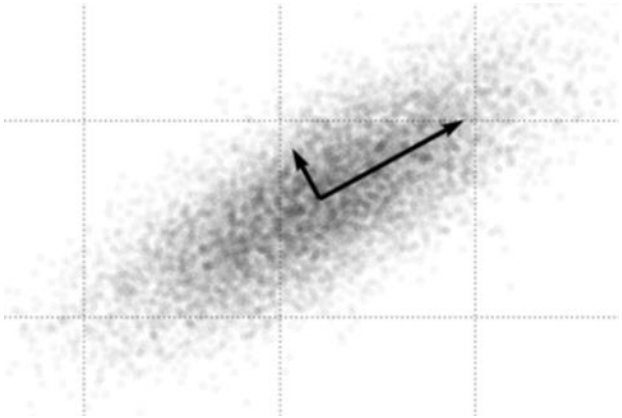


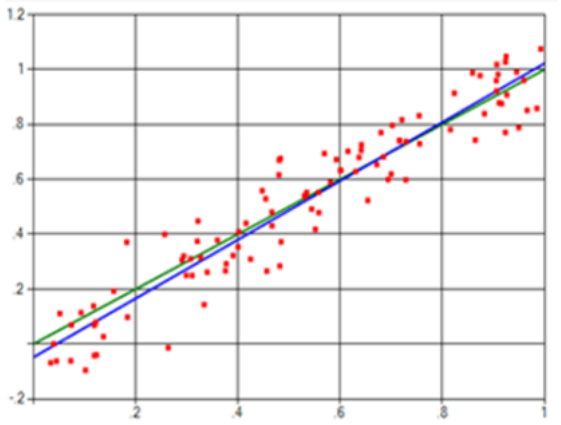
Source: Flowing Data

Advanced Data Analysis with R

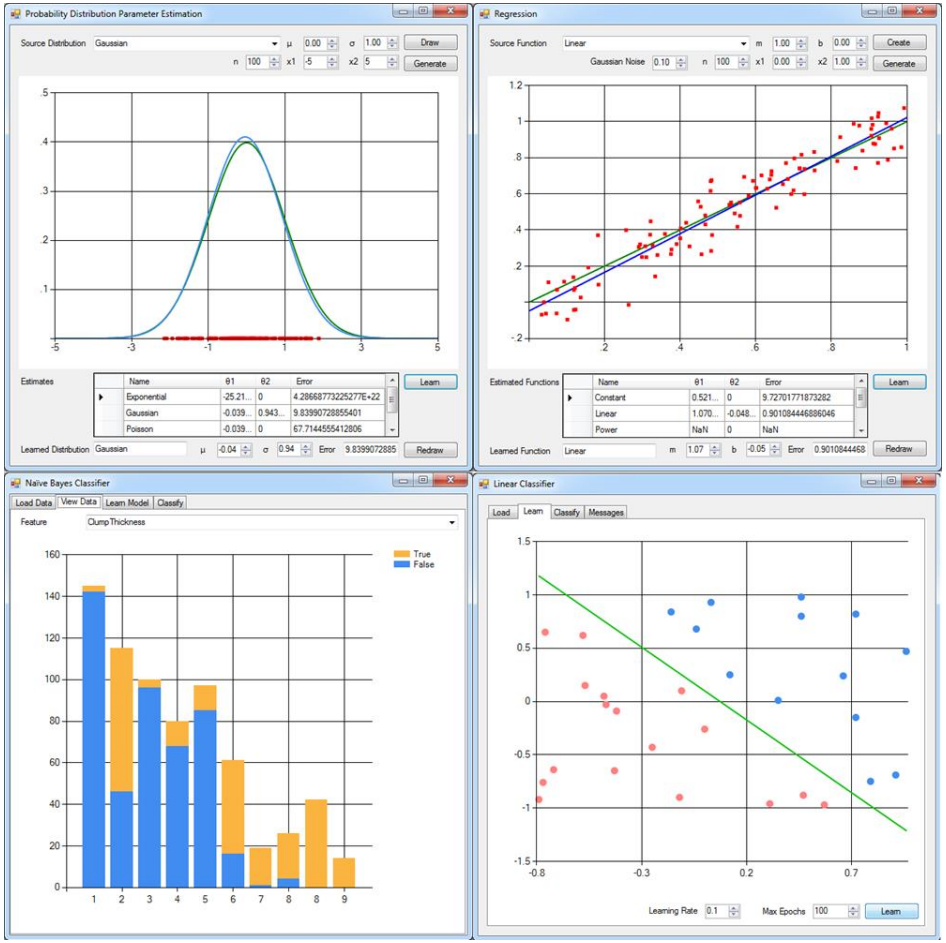
- Cluster Analysis
- Statistical Modeling
- Dimensionality Reduction
- Analysis of Variance (ANOVA)



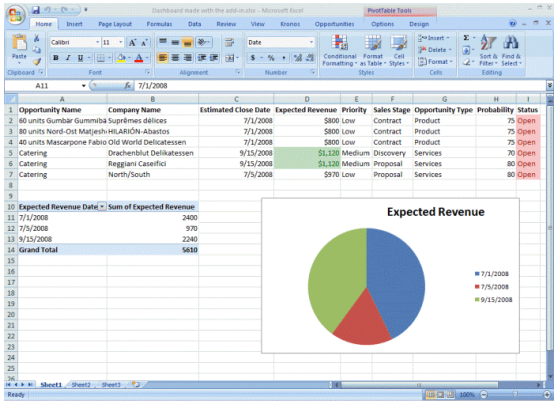


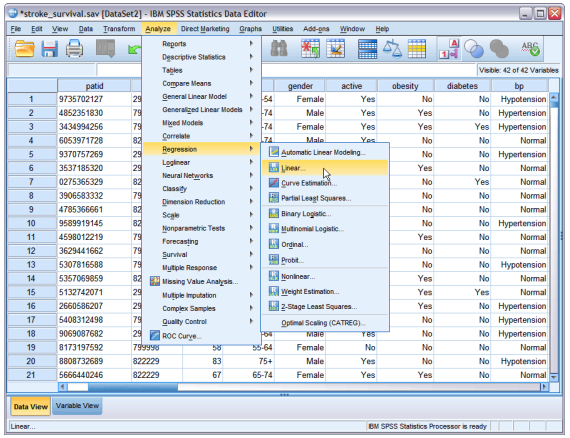


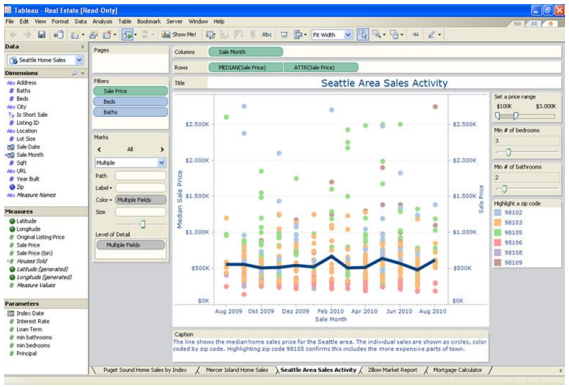
Data Mining and Machine Learning with R

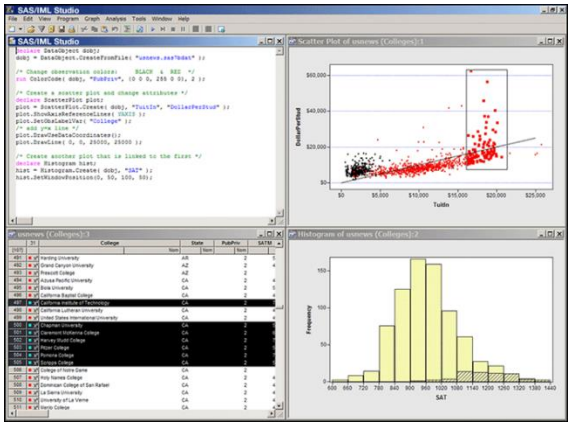


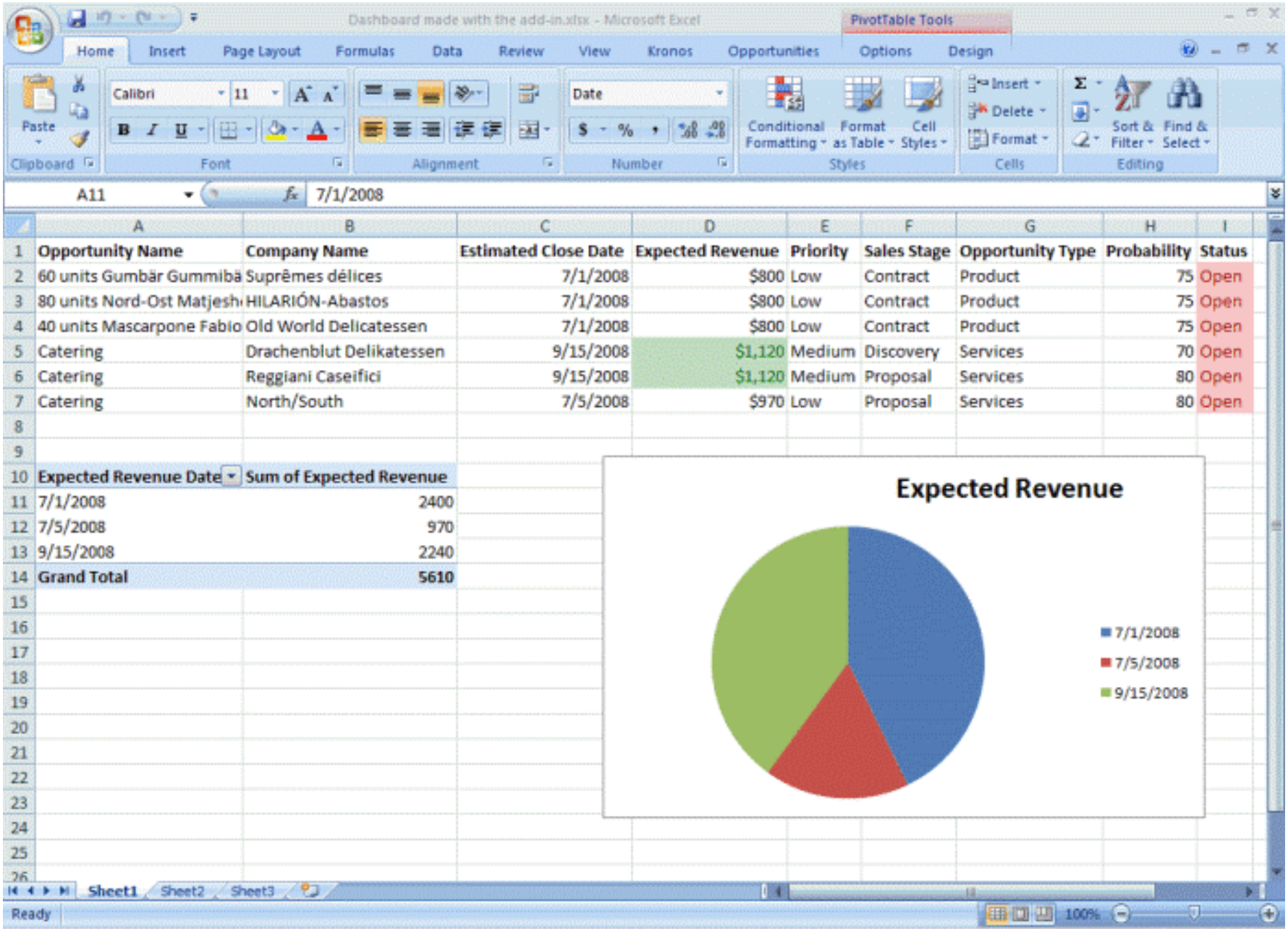
Alternatives to R for EDA



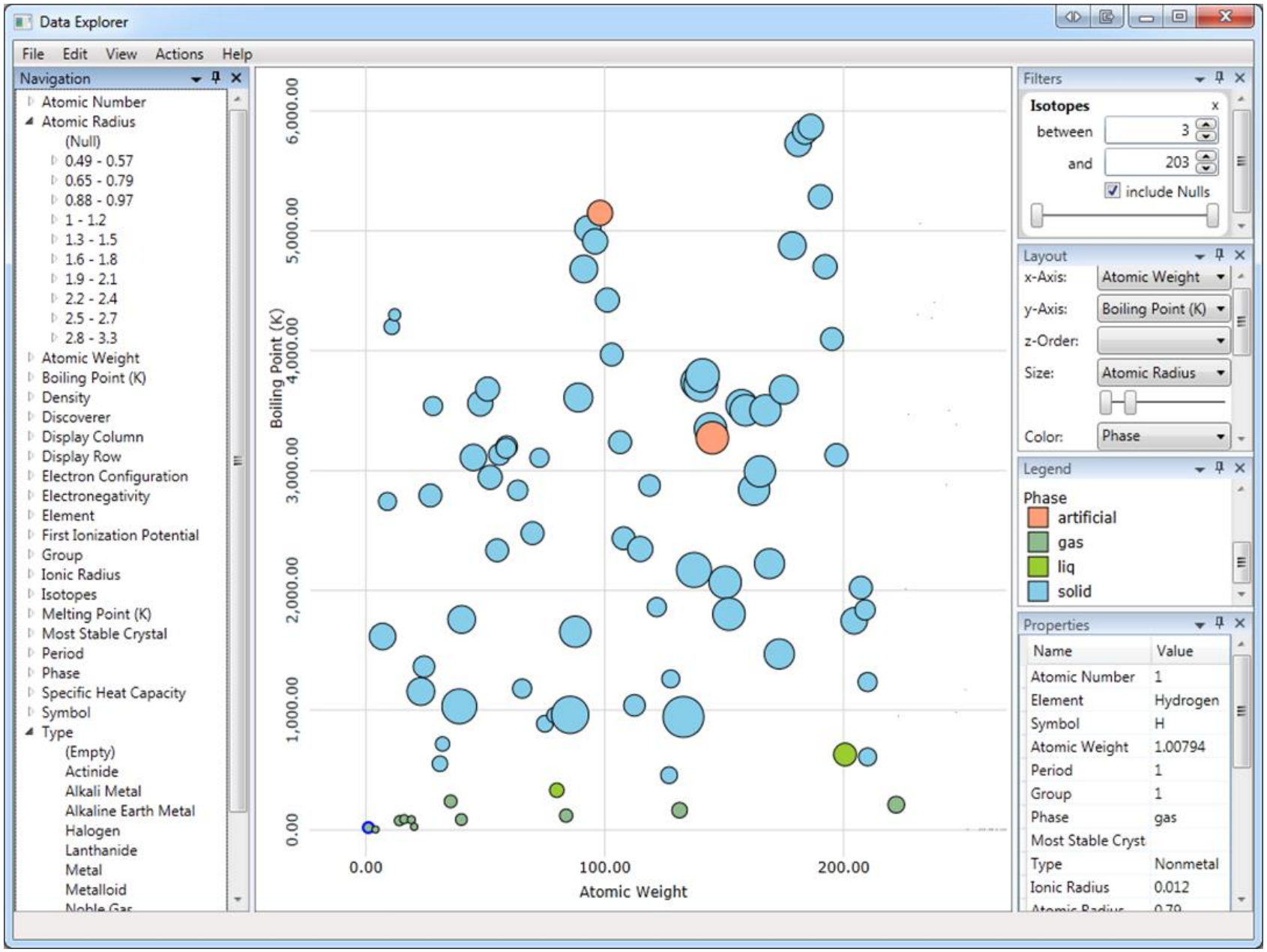


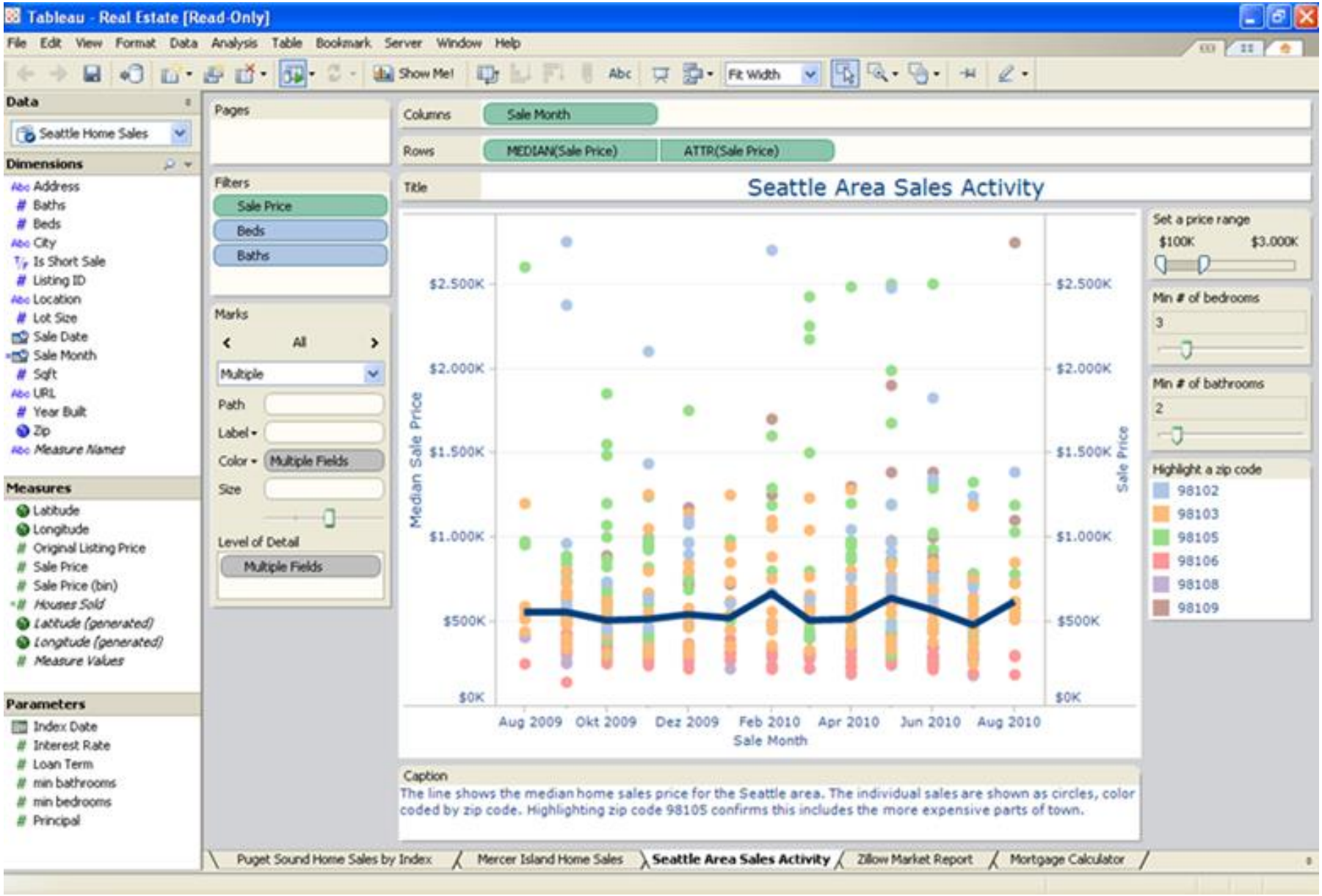




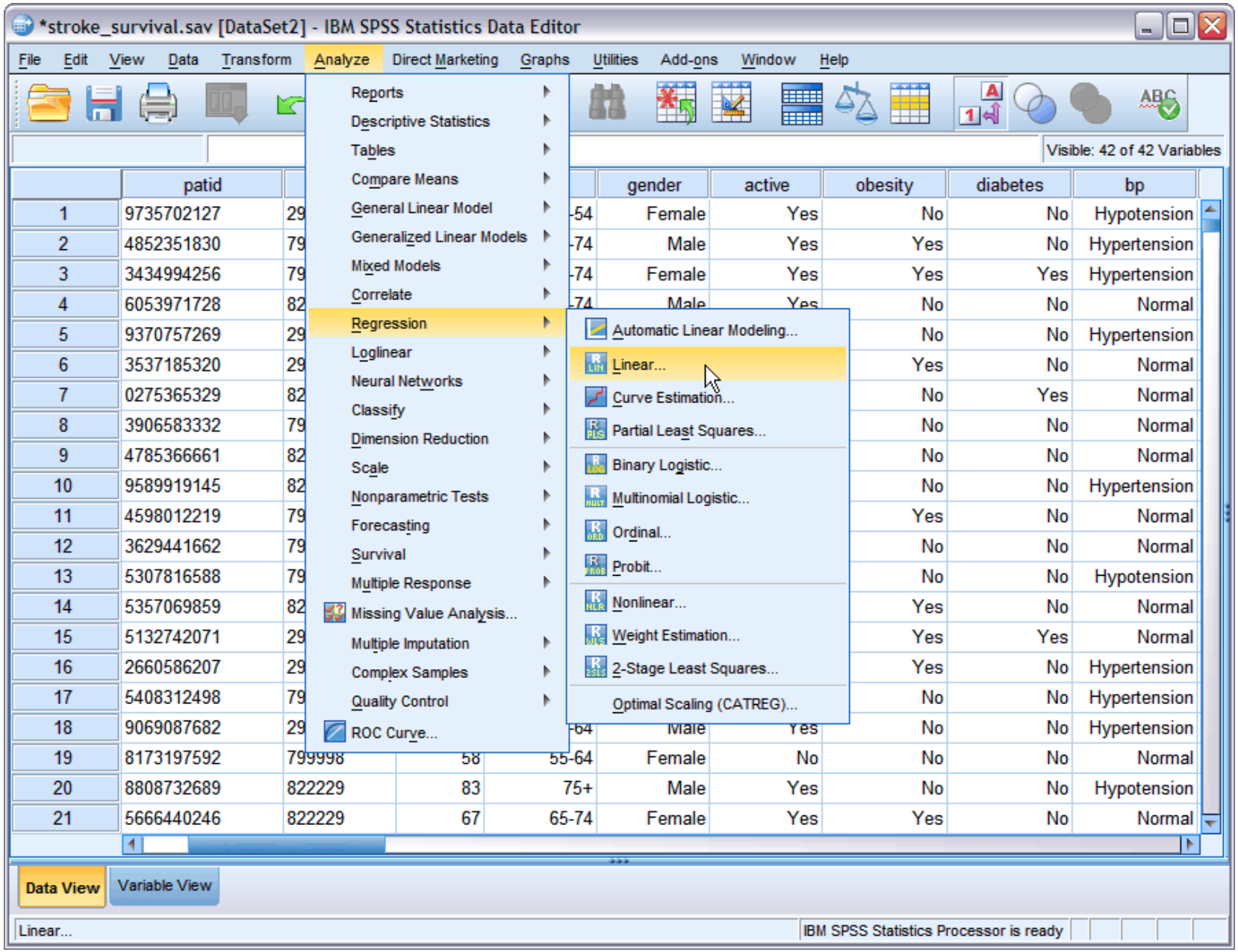


Source: Microsoft

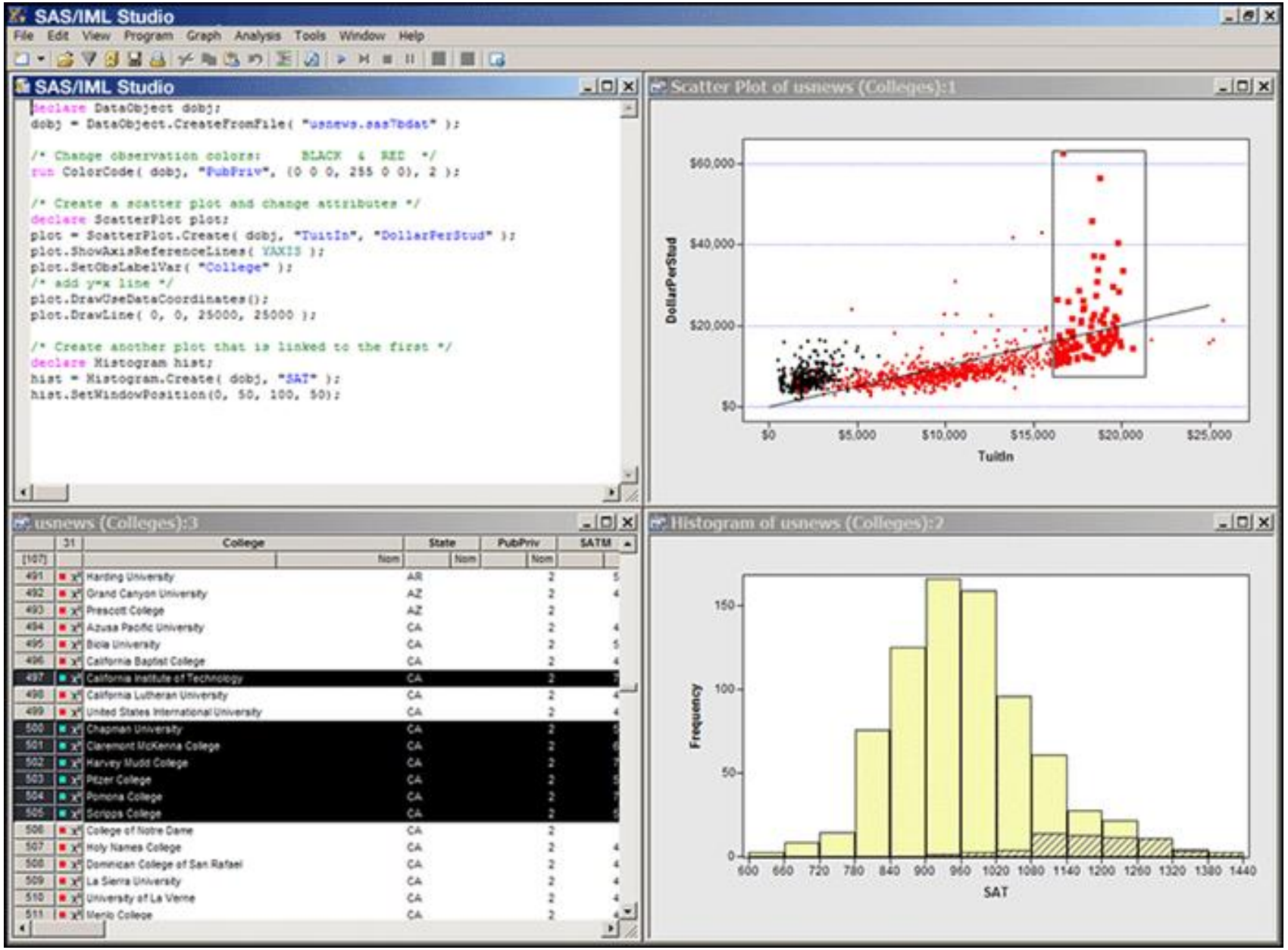

Data Explorer
File
Edit
View
Actions
Help
Navigation
Atomic Number
Atomic Radius
(Null)
0.49 - 0.57
0.65 - 0.79
0.88 - 0.97
1 - 1.2
1.3 - 1.5
1.6 - 1.8
1.9 - 2.1
2.2 - 2.4
2.5 - 2.7
2.8 - 3.3
Atomic Weight
Boiling Point (K)
Density
Discoverer
Display Column
Display Row
Electron Configuration
Electronegativity
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



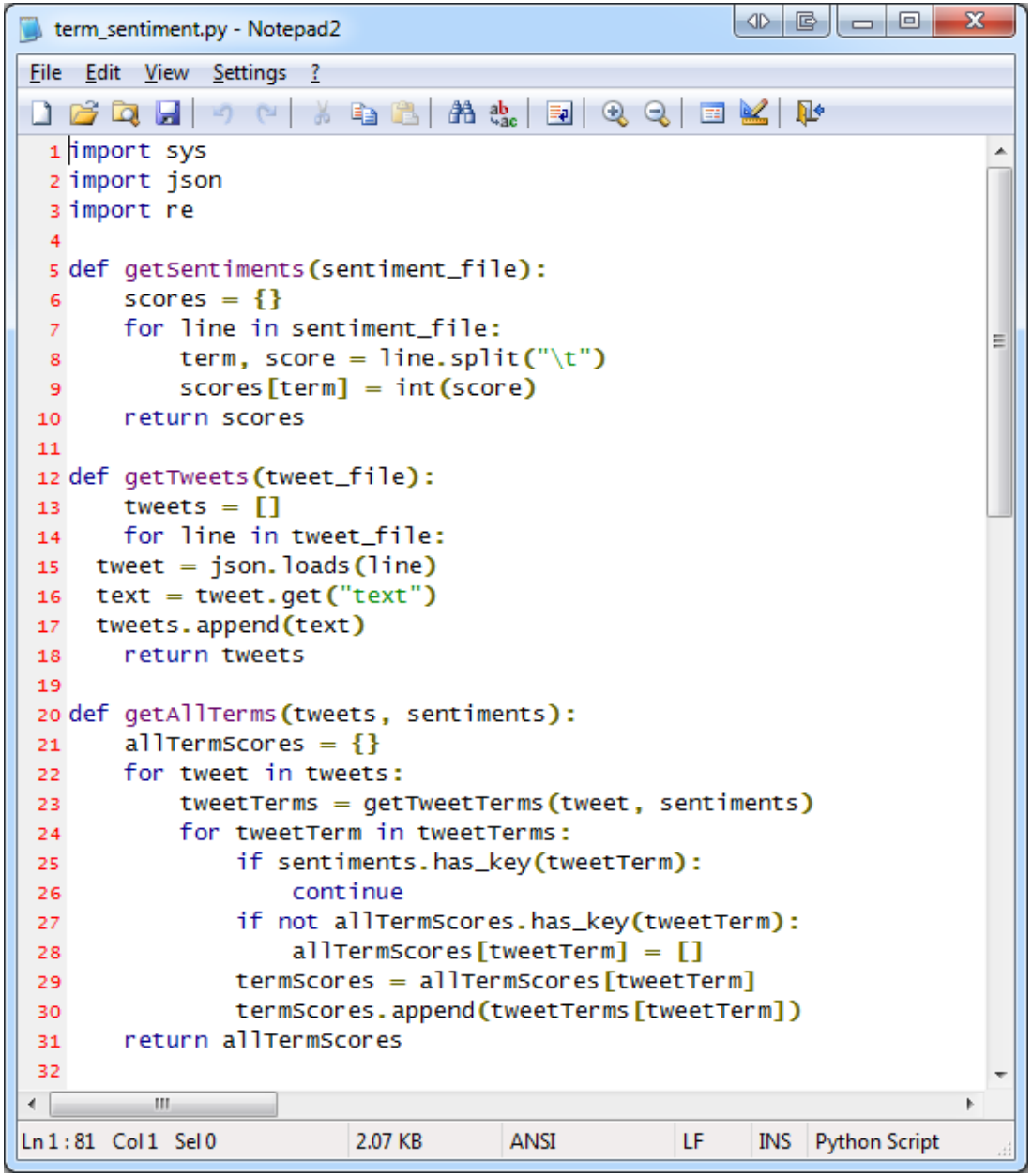
Source: Tableau



Source: IBM SPSS


SAS/IML Studio
File Edit View Program Graph Analysis Tools Window Help
SAS/IML Studio
declare DataObject dobj;
dobj = DataObject.CreateFromFile("usnews.sas7bdat");
/* Change observation colors: BLACK & RED */
run ColorCode(dobj, "PubPriv", {0 0 0, 255 0 0}, 2);
/* Create a scatter plot and change attributes */
declare ScatterPlot plot;
plot = ScatterPlot.Create(dobj, "TuitIn", "DollarPerStud");
plot.ShowAxisReferenceLines(YAXIS);
plot.SetObsLabelVar("College");
/* add y=x line */
plot.DrawUseDataCoordinates();
plot.DrawLine(0, 0, 25000, 25000);
/* Create another plot that is linked to the first */
declare Histogram hist;
hist = Histogram.Create(dobj, "SAT");
hist.SetWindowPosition(0, 50, 100, 50);
Scatter Plot of usnews (Colleges):1
$60,000
$40,000
$20,000
$0
DollarPerStud
$0
$5,000
$10,000
$15,000
$20,000
$25,000
TuitIn
usnews (Colleges):3
31
College
State
PubPriv
SATM
[107]
Nom
491 Harding University AR 2 5
492 Grand Canyon University AZ 2 4
493 Prescott College AZ 2
494 Azusa Pacific University CA 2 4
495 Biola University CA 2 5
496 California Baptist College CA 2 4
497 California Institute of Technology CA 2
498 California Lutheran University CA 2 4
499 United States International University CA 2 4
500 Chapman University CA 2
501 Claremont McKenna College CA 2
502 Harvey Mudd College CA 2
503 Pitzer College CA 2
504 Pomona College CA 2
505 Scripps College CA 2
506 College of Notre Dame CA 2
507 Holy Names College CA 2 4
508 Dominican College of San Rafael CA 2 4
509 La Sierra University CA 2 4
510 University of La Verne CA 2 4
511 Menlo College CA 2 4
Histogram of usnews (Colleges):2
150
100
50
0
Frequency
600 660 720 780 840 900 960 1020 1080 1140 1200 1260 1320 1380 1440
SAT

Source: SAS


term_sentiment.py - Notepad2
File Edit View Settings ?
import sys
import json
import re
def getSentiments(sentiment_file):
scores = {}
for line in sentiment_file:
term, score = line.split("\t")
scores[term] = int(score)
return scores
def getTweets(tweet_file):
tweets = []
for line in tweet_file:
tweet = json.loads(line)
text = tweet.get("text")
tweets.append(text)
return tweets
def getAllTerms(tweets, sentiments):
allTermScores = {}
for tweet in tweets:
tweetTerms = getTweetTerms(tweet, sentiments)
for tweetTerm in tweetTerms:
if sentiments.has_key(tweetTerm):
continue
if not allTermScores.has_key(tweetTerm):
allTermScores[tweetTerm] = []
termScores = allTermScores[tweetTerm]
termScores.append(tweetTerms[tweetTerm])
return allTermScores
Ln 1 : 81 Col 1 Sel 0
2.07 KB
ANSI
LF
INS
Python Script

Code Demo

Where to Go Next…

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

Conclusion

Conclusion

- Introduction to R
- Data munging
- Descriptive statistics
- Data visualization
- Beyond R & EDA

Feedback

- Feedback is very important to me
- One thing you liked?
- One thing I could improve?

Contact Info

Matthew Renze
@matthewrenze
matthew@renzeconsulting.com

Renze Consulting
www.renzeconsulting.com

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