Exploratory Data Analysis with R

Matthew Renze

@matthewrenze

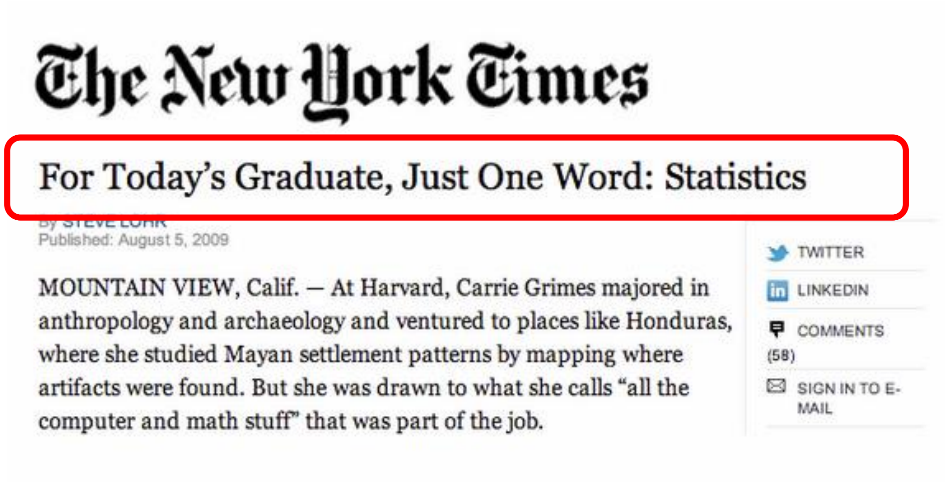
#KCDC15

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 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

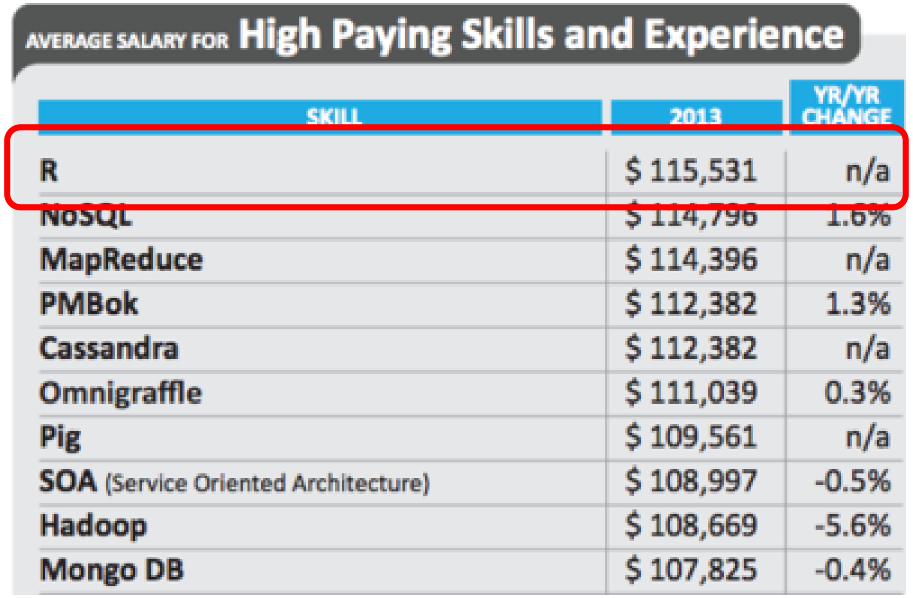

The New York Times

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

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]

IOWA STATE UNIVERSITY

KIMBALL GROUP

esri

JOHNS HOPKINS UNIVERSITY

Titanium Sponsors

DATA BANK
powered by actifio

epiq
SYSTEMS

VALOREM
CONSULTING
WWW.VALOREMCONSULTING.COM
REGISTER TO WIN AN XBOX!

PAIGE
TECHNOLOGIES
INTELLIGENT PAIRING. PERPETUAL SUCCESS.

Platinum Sponsors

perceptivesoftware
from Lexmark

ADAPTIVE™
SOLUTIONS GROUP

KU
EDWARDS CAMPUS
The University of Kansas

pinsight®
MEDIA+

VinSolutions
Make every connection count.

MS MULTI SERVICE®
INNOVATION WHERE IT MATTERS.

Gold Sponsors

Balance® Innovations

DST®
SYSTEMS

LRS
Consulting Services

NETCHEMIA™
Transforming the way education works

ADVANTAGE
TECH

KEYHOLE
SOFTWARE

OAKWOOD

stackify

GARMIN.

NEW DIRECTIONS™
TOGETHER IS THE WAY FORWARD

Bradford & Galt
CONSULTING SERVICES

dsi.

FitBark

Cerner

jack henry
& ASSOCIATES INC.®

UNITEDLEX

Commerce Bank

TRIPLE-I™

BATS
Making Markets Better®

LIFERAY.

imodules

eccoselect
PEOPLE | PROJECTS | PERFORMANCE

CENTRIQ
TRAINING

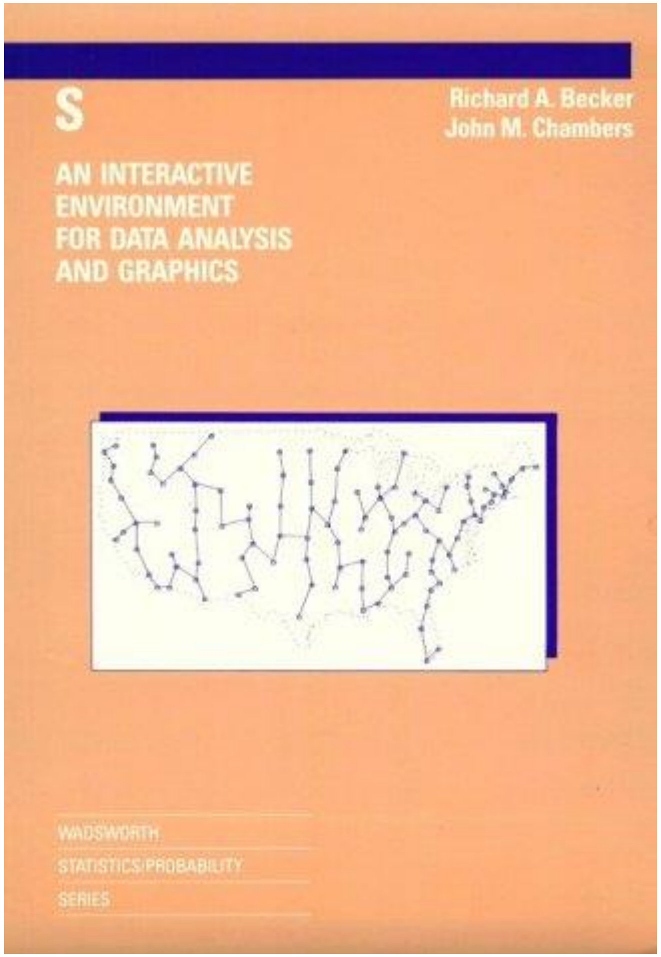
Introduction to R

What is R?

- R is an open source implementation of S

What is S?

- Statistical language
- Bell Labs in 1976
- Owned by TIBCO



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 - developed by:
  - Ross Ihaka
  - Robert Gentleman
- 1995 - open source
- 2000 - v.1.0 released
- Today, R is at v.3.2.0

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

Source: www.auklandlifestyle.com

What is R?

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

Source: www.r-project.org

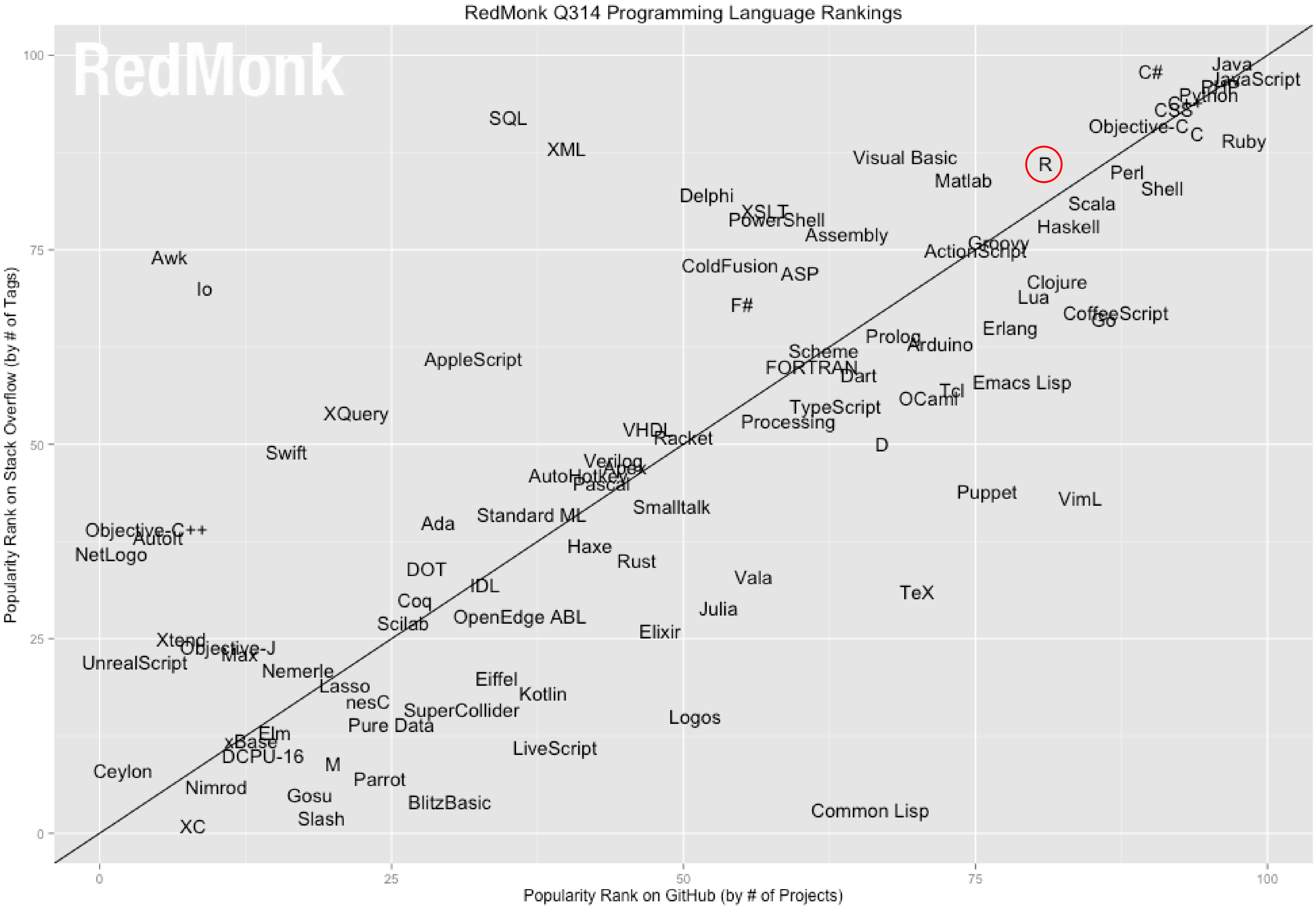
What is R?

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

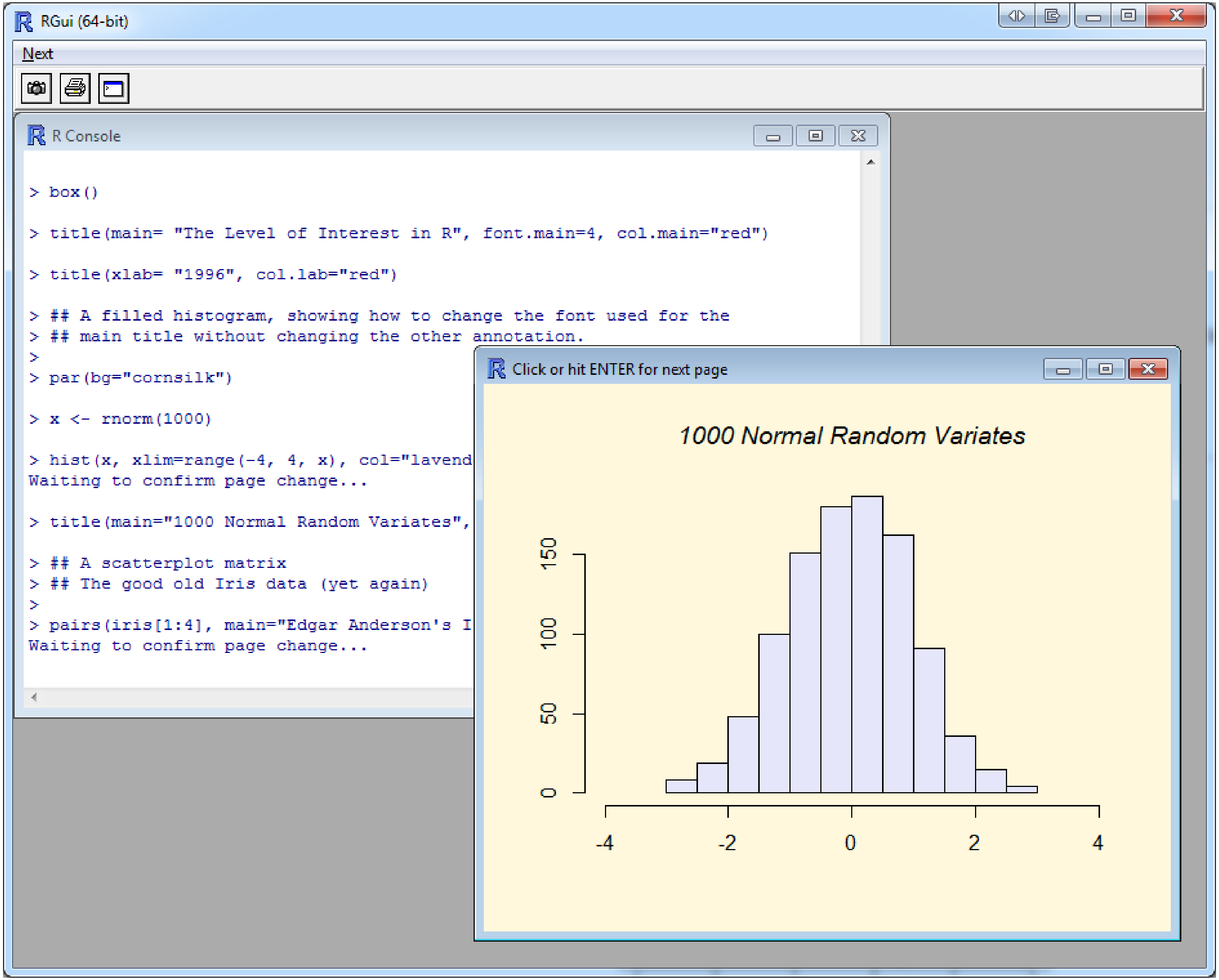
and best of all…

FREE

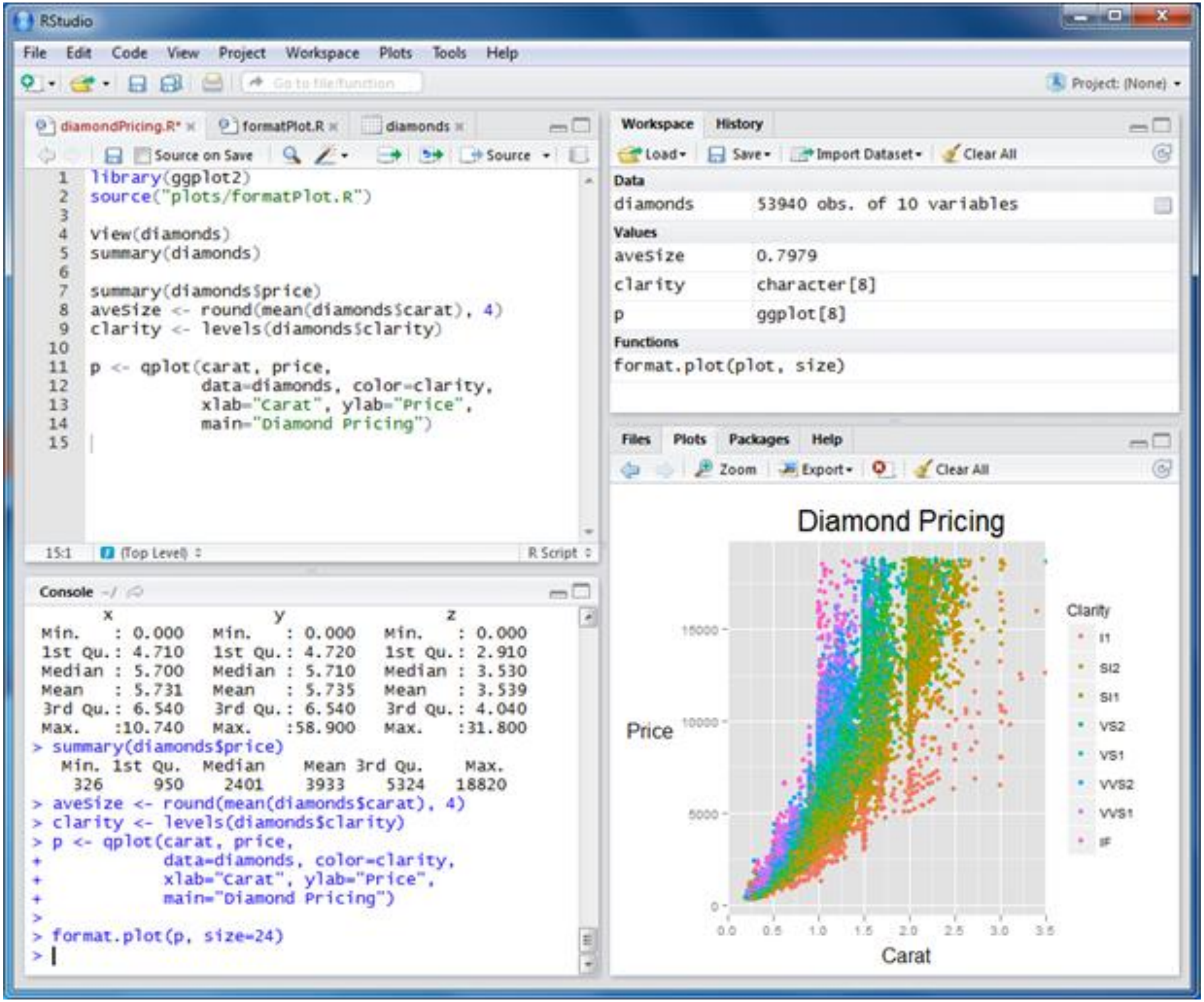
FREE



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



```
> 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...
```



```r
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
- 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

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.

Open an Account Today

- Solid returns
- Monthly cash flow
- Low volatility

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
- 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

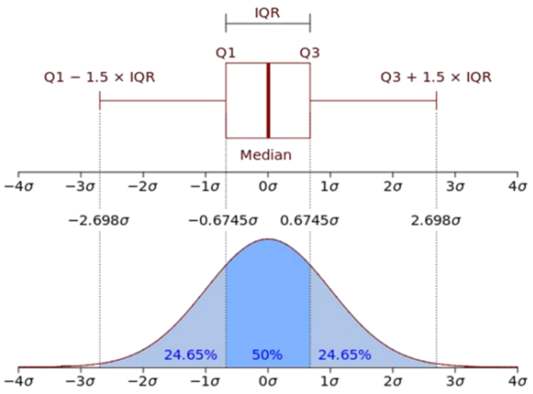
Statistical Terms

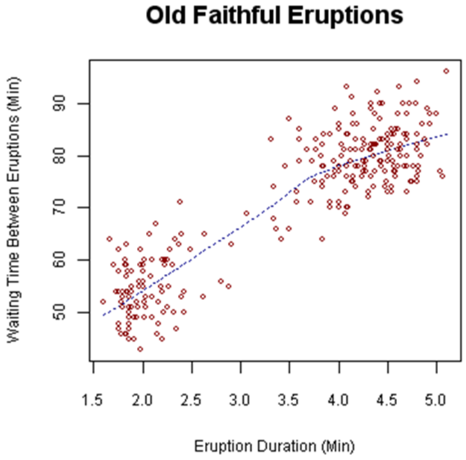
- Observations
- Variables
- Qualitative variable
- Quantitative variable

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

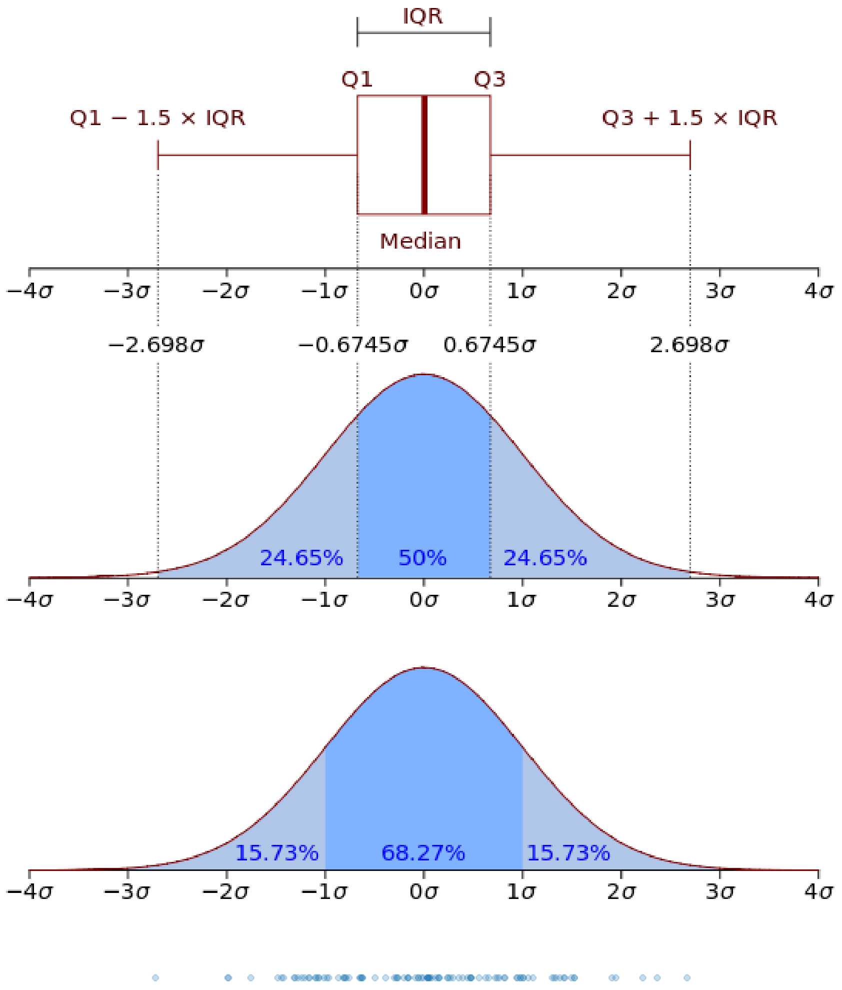
- Type of variables
  - Qualitative
  - Quantitative
- Number of variables
  - Univariate
  - Bivariate
  - Multivariate





Univariate Analysis

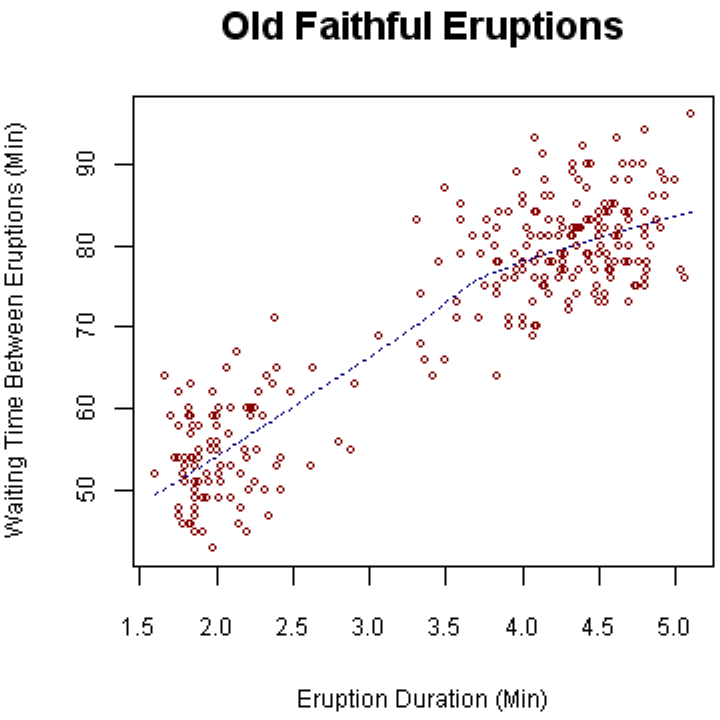
- One variable
- Measures include:
  - Central tendency
  - Dispersion



Source: Wikipedia

Bivariate Analysis

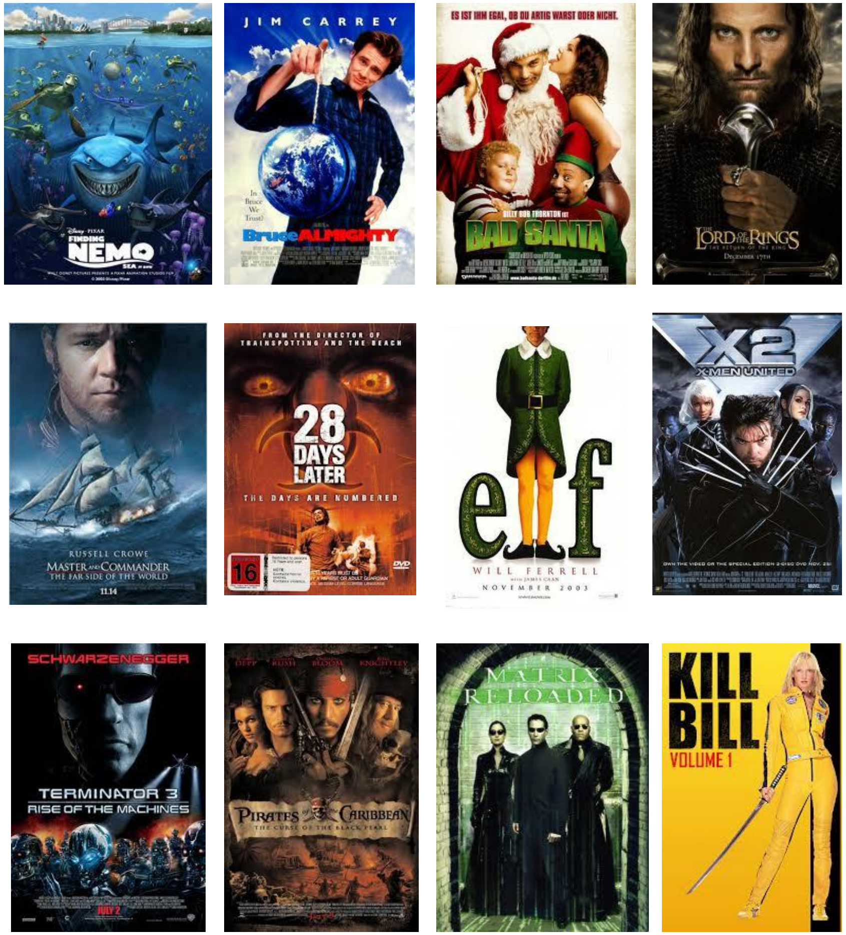
- Two variables
  - Predictor
  - Outcome
- Measures include
  - Covariance
  - Correlation



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

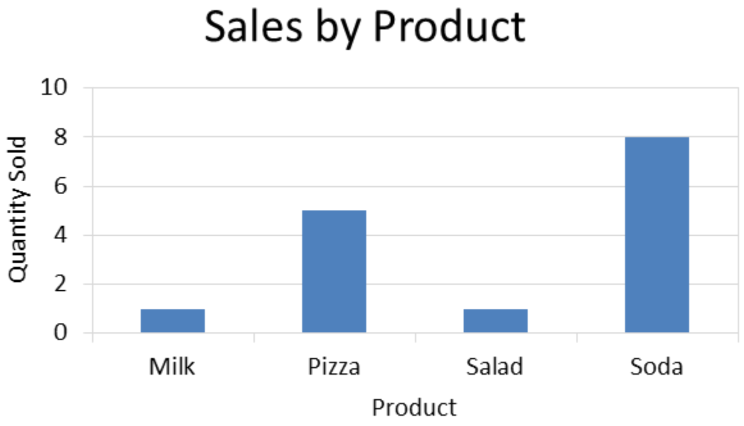
Code Demo

Data Visualization

Data Visualization

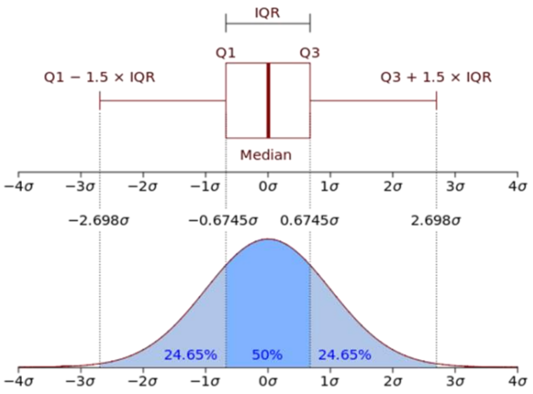
- Visual data representation
- For human pattern recognition
- Map dimensions to visual characteristics

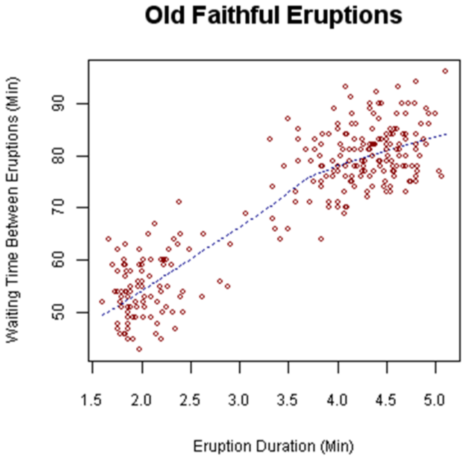
ID	Date	Customer	Product	Quantity
1	2012-10-27	John	Pizza	2
2	2012-10-27	John	Soda	2
3	2012-10-27	Jill	Salad	1
4	2012-10-27	Bob	Milk	1
5	2012-10-28	Sue	Soda	3
6	2012-10-28	Bob	Pizza	2
7	2012-10-28	Jill	Pizza	1
8	2012-10-28	Jill	Soda	3



Types of Data Visualizations

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

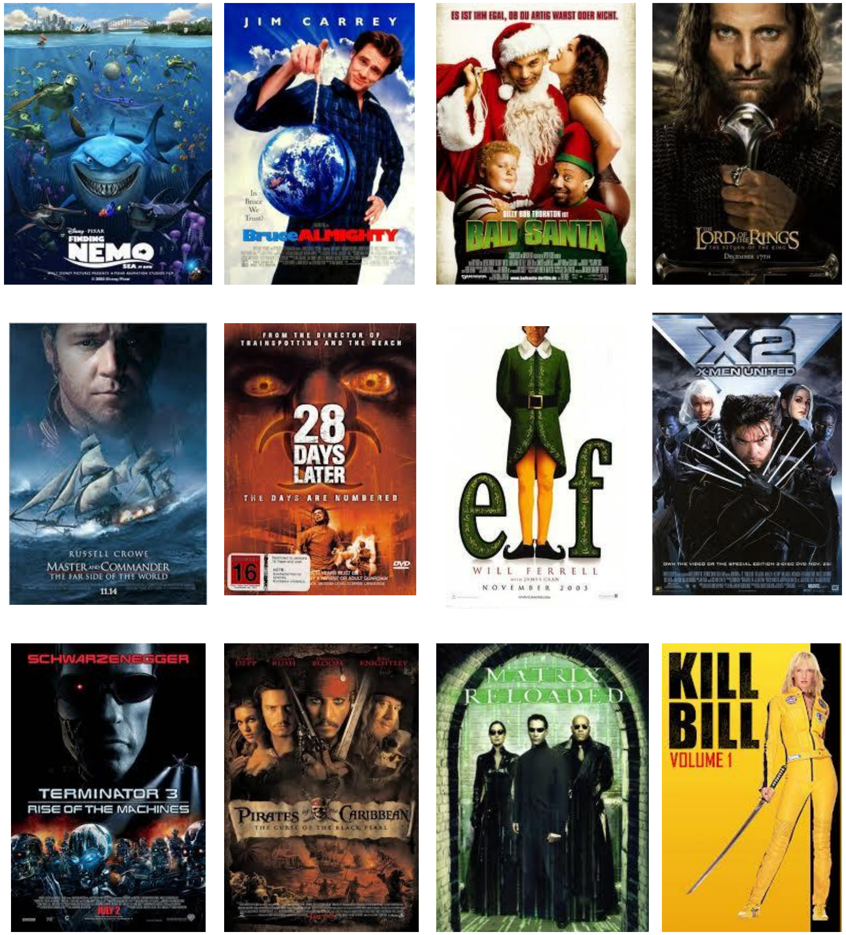




Source: Wikipedia

Code Demo: Movies Data Set

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



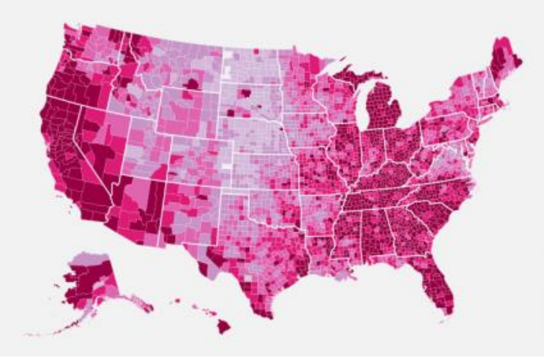
Source: http://www.rossmanchance.com/iscam2/files.html

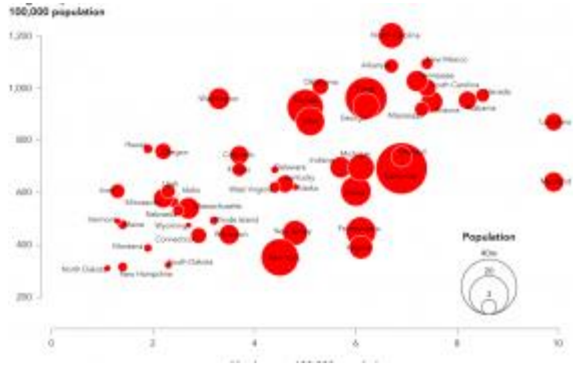
Code Demo

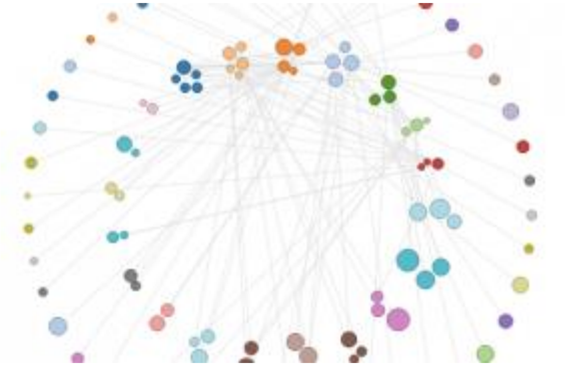
Beyond R and EDA

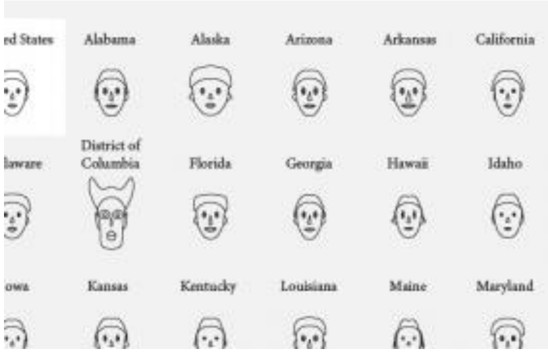
This is just the tip of the iceberg!

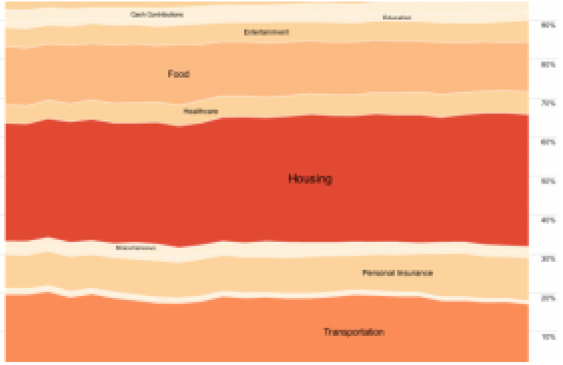
Advanced Visualizations with R







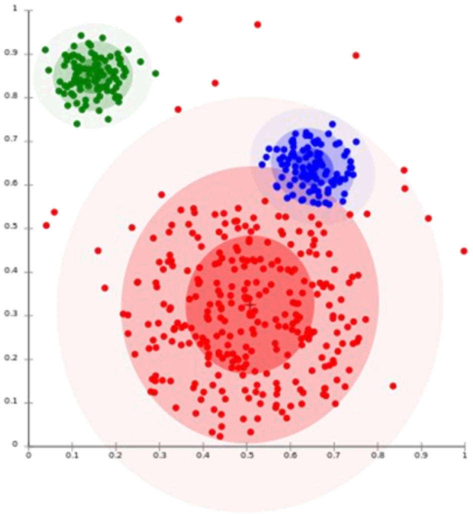


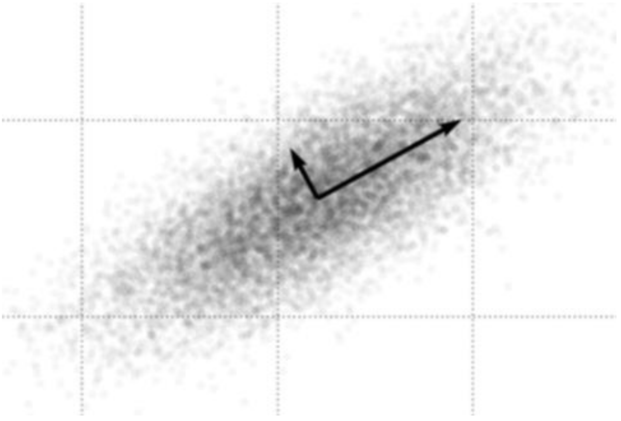


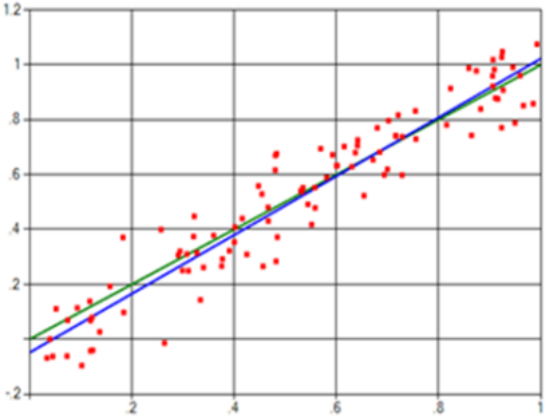
Source: Flowing Data

Advanced Data Analysis with R

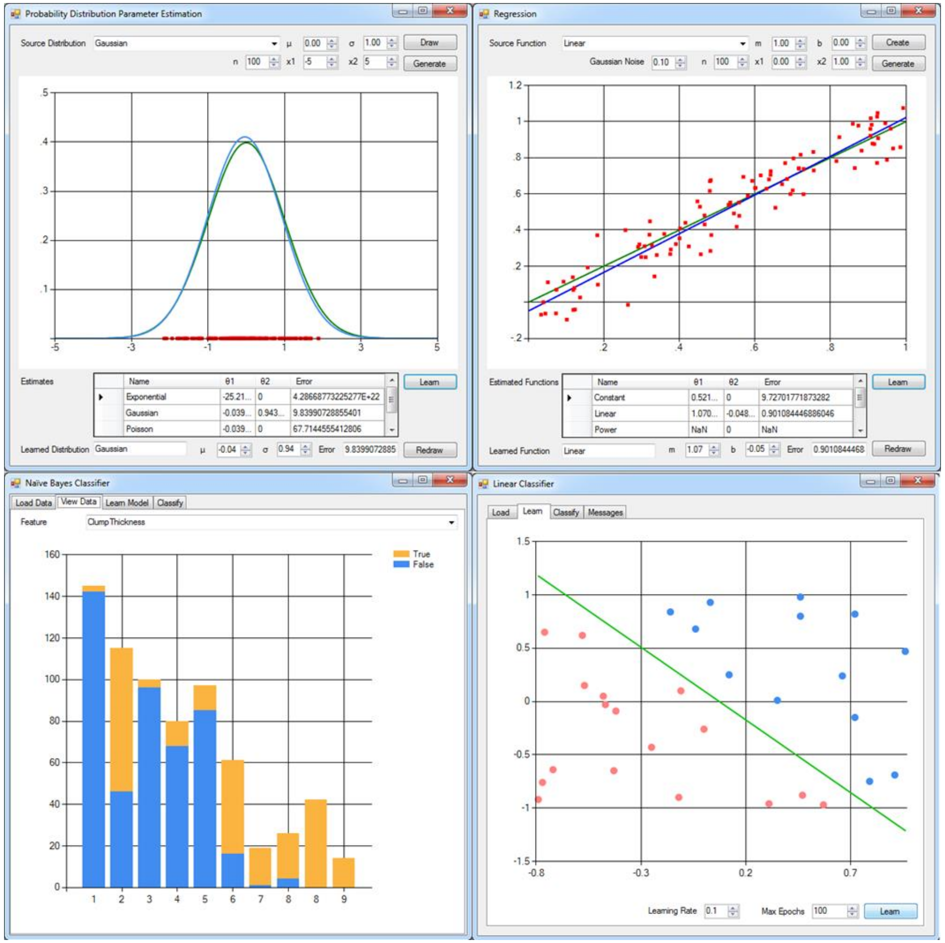
- Cluster Analysis
- Statistical Modeling
- Dimensionality Reduction
- Analysis of Variance (ANOVA)







Data Mining and Machine Learning with R



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

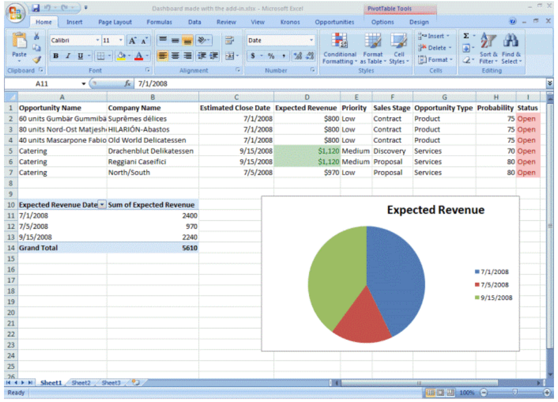
Naïve Bayes Classifier

Feature: ClumpThickness

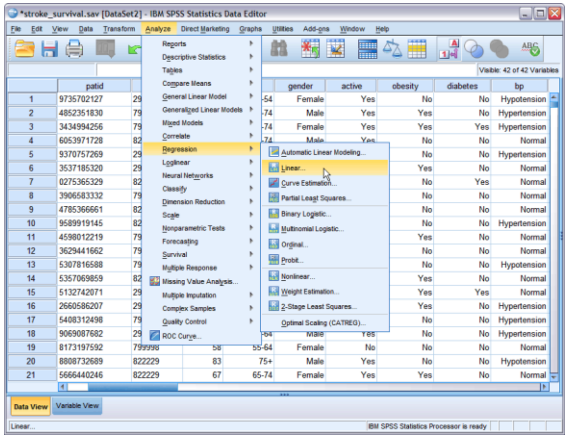
Linear Classifier

Learning Rate 0.1; Max Epochs 100

Alternatives to R for EDA



	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
10	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							



*stroke_survival.sav [DataSet2] - IBM SPSS Statistics Data Editor

	patid	gender	active	obesity	diabetes	bp
1	9735702127	Female	Yes	No	No	Hypotension
2	4852351830	Male	Yes	Yes	No	Hypertension
3	3434994256	Female	Yes	Yes	Yes	Hypertension
4	6053971728	Male	Yes	No	No	Normal
5	9370757269			No	No	Hypertension
6	3537185320			Yes	No	Normal
7	0275365329			No	Yes	Normal
8	3906583332			No	No	Normal
9	4785366661			No	No	Normal
10	9589919145			No	No	Hypertension
11	4598012219			Yes	No	Normal
12	3629441662			No	No	Normal
13	5307816588			No	No	Hypotension
14	5357069859			Yes	No	Normal
15	5132742071			Yes	Yes	Normal
16	2660586207			Yes	No	Hypertension
17	5408312498			No	No	Hypertension
18	9069087682	Male	Yes	No	No	Hypertension
19	8173197592	Female	No	No	No	Normal
20	8808732689	Male	Yes	No	No	Hypotension
21	5666440246	Female	Yes	Yes	No	Normal

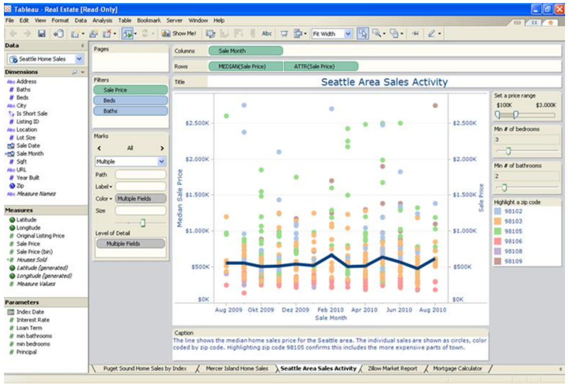
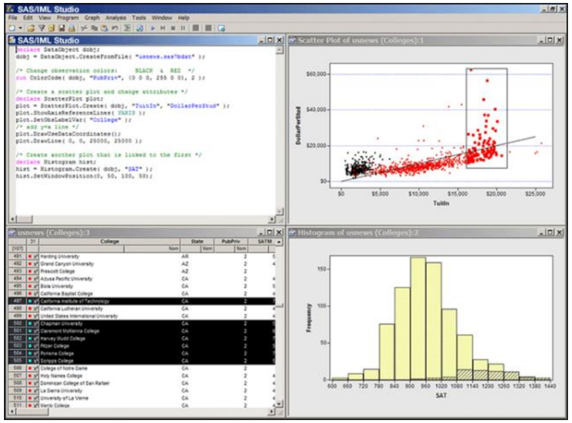


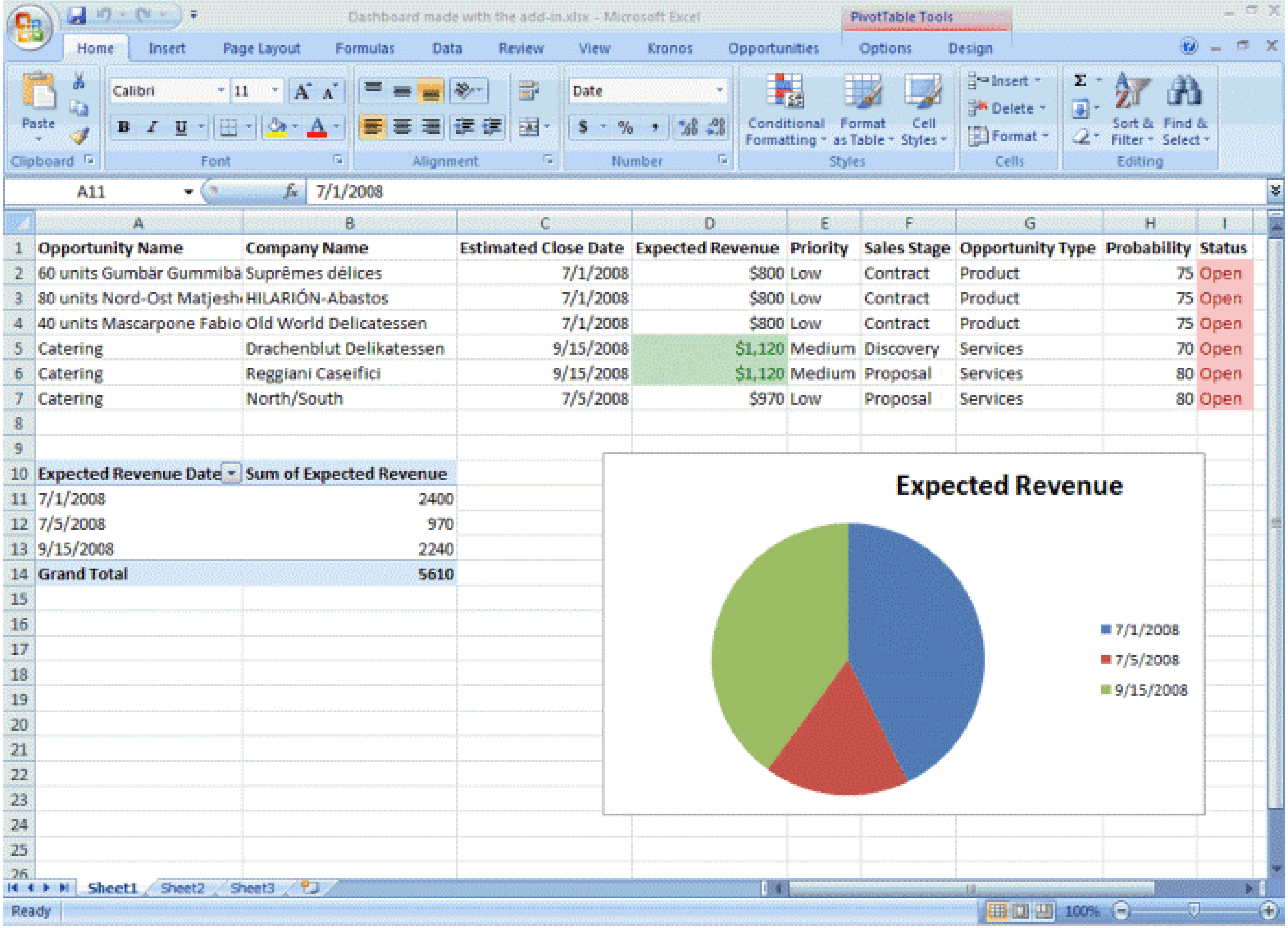
Tableau - Real Estate [Read-Only]

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.



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);
```

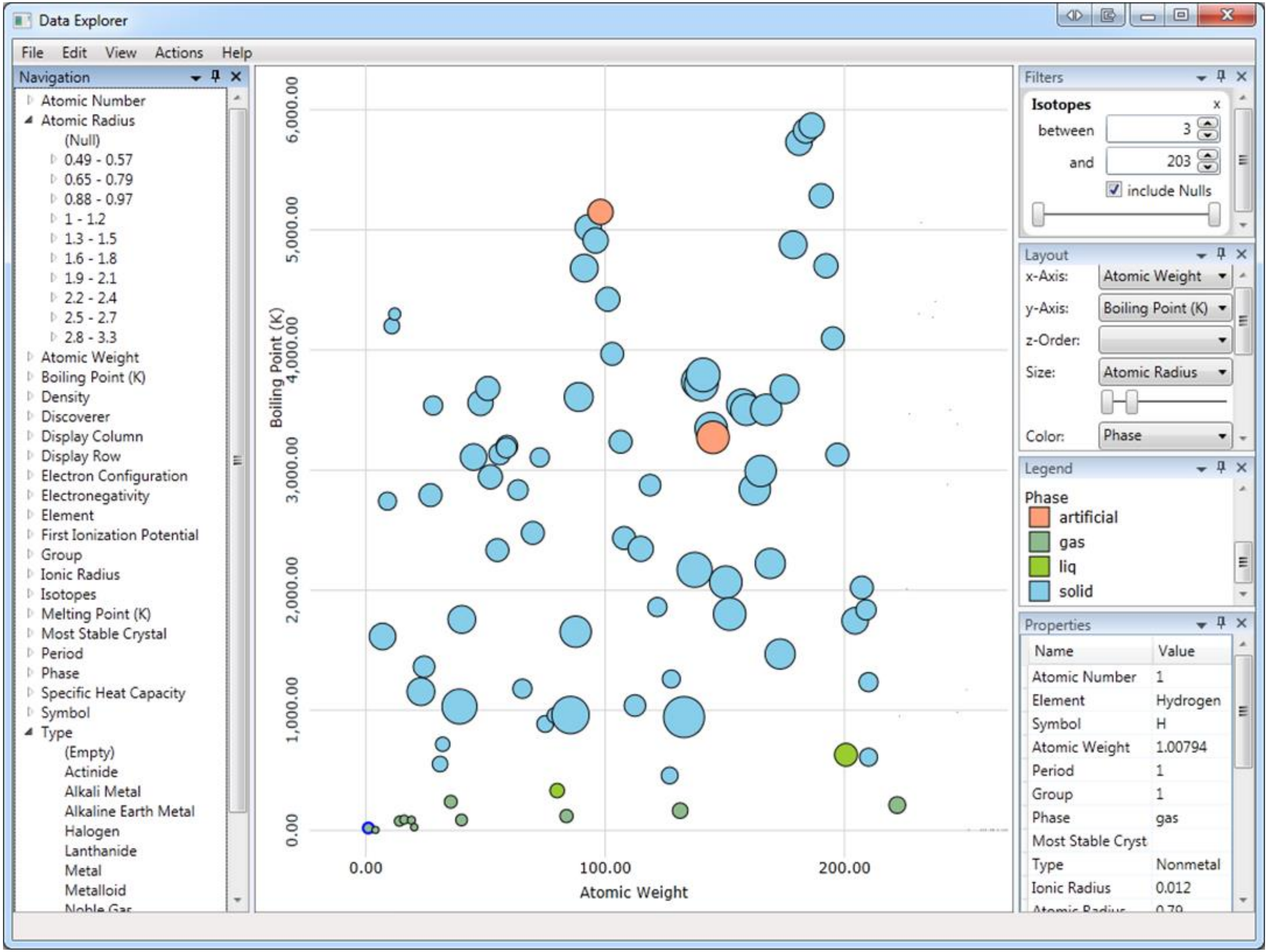


A11 · 7/1/2008

	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
7/1/2008	2400
7/5/2008	970
9/15/2008	2240
Grand Total	5610

Source: Microsoft



Isotopes x

between 3

and 203

☑ include Nulls

x-Axis: Atomic Weight

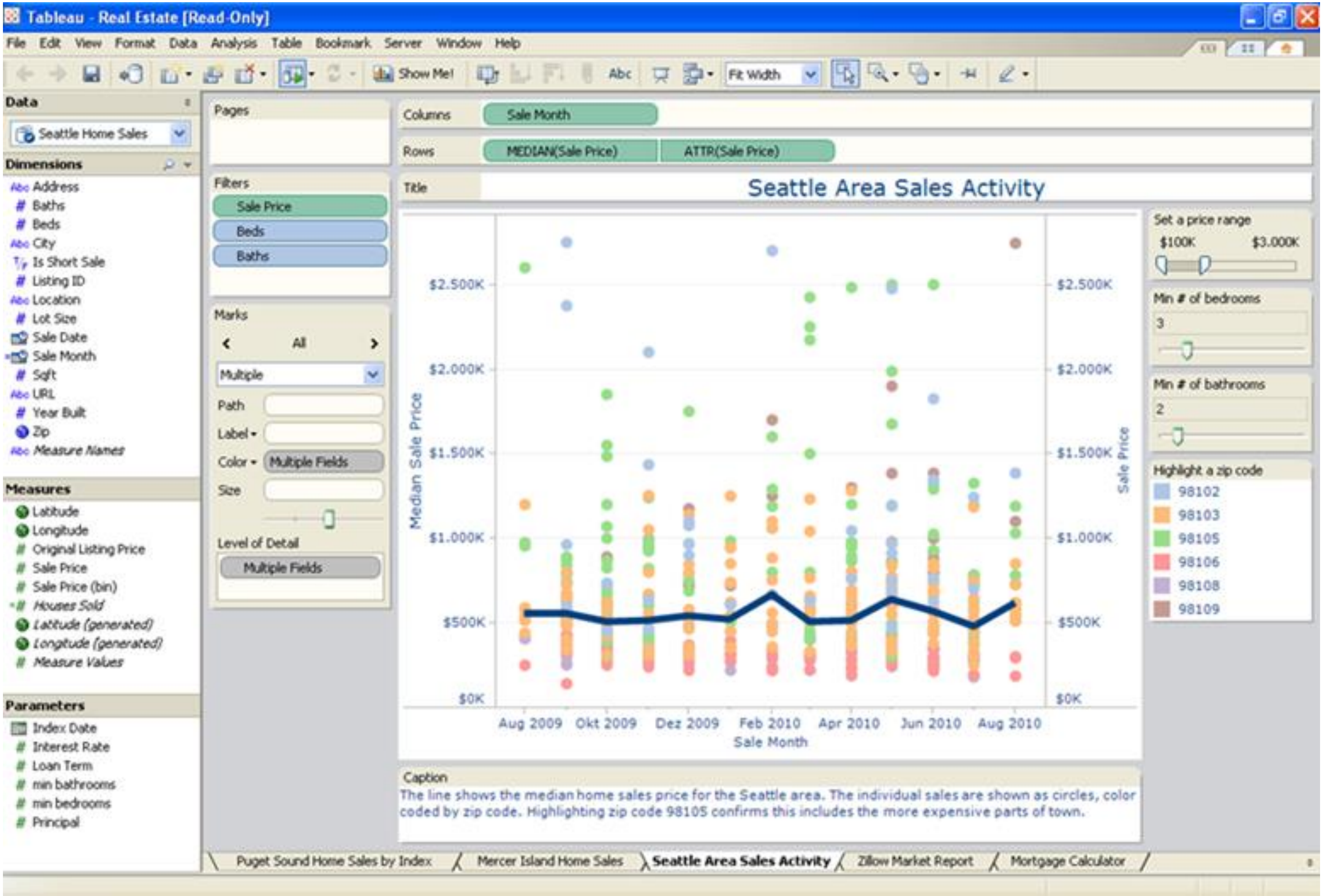
y-Axis: Boiling Point (K)

z-Order:

Size: Atomic Radius

Color: Phase

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



Columns: Sale Month

Rows: MEDIAN(Sale Price) ATTR(Sale Price)

Title: Seattle Area Sales Activity

Marks: Multiple

Color: Multiple Fields

Level of Detail: Multiple Fields

Set a price range: $100K $3.000K

Min # of bedrooms: 3

Min # of bathrooms: 2

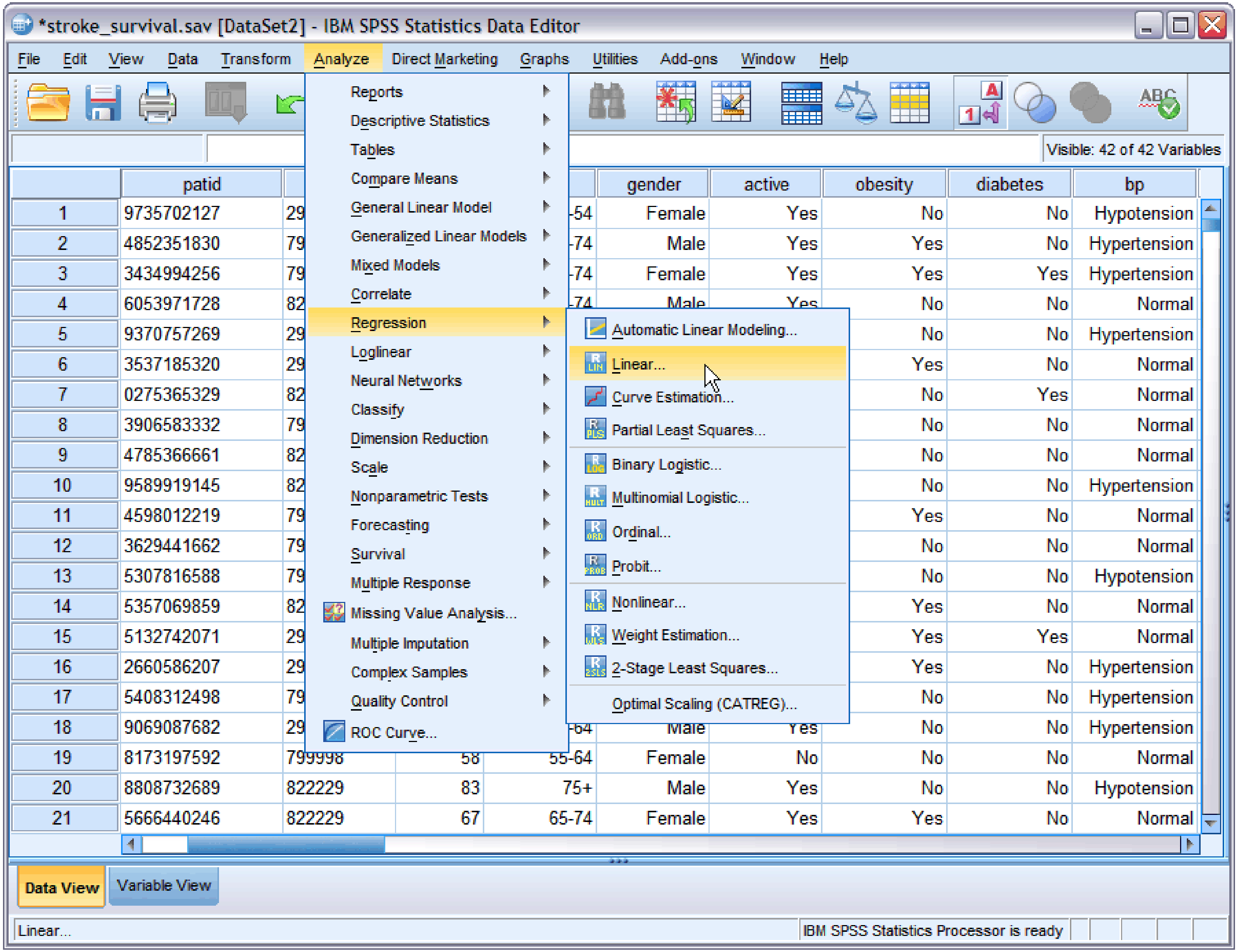
Highlight a zip code

🟦	98102
🟧	98103
🟩	98105
🟥	98106
🟪	98108
🟫	98109

Caption

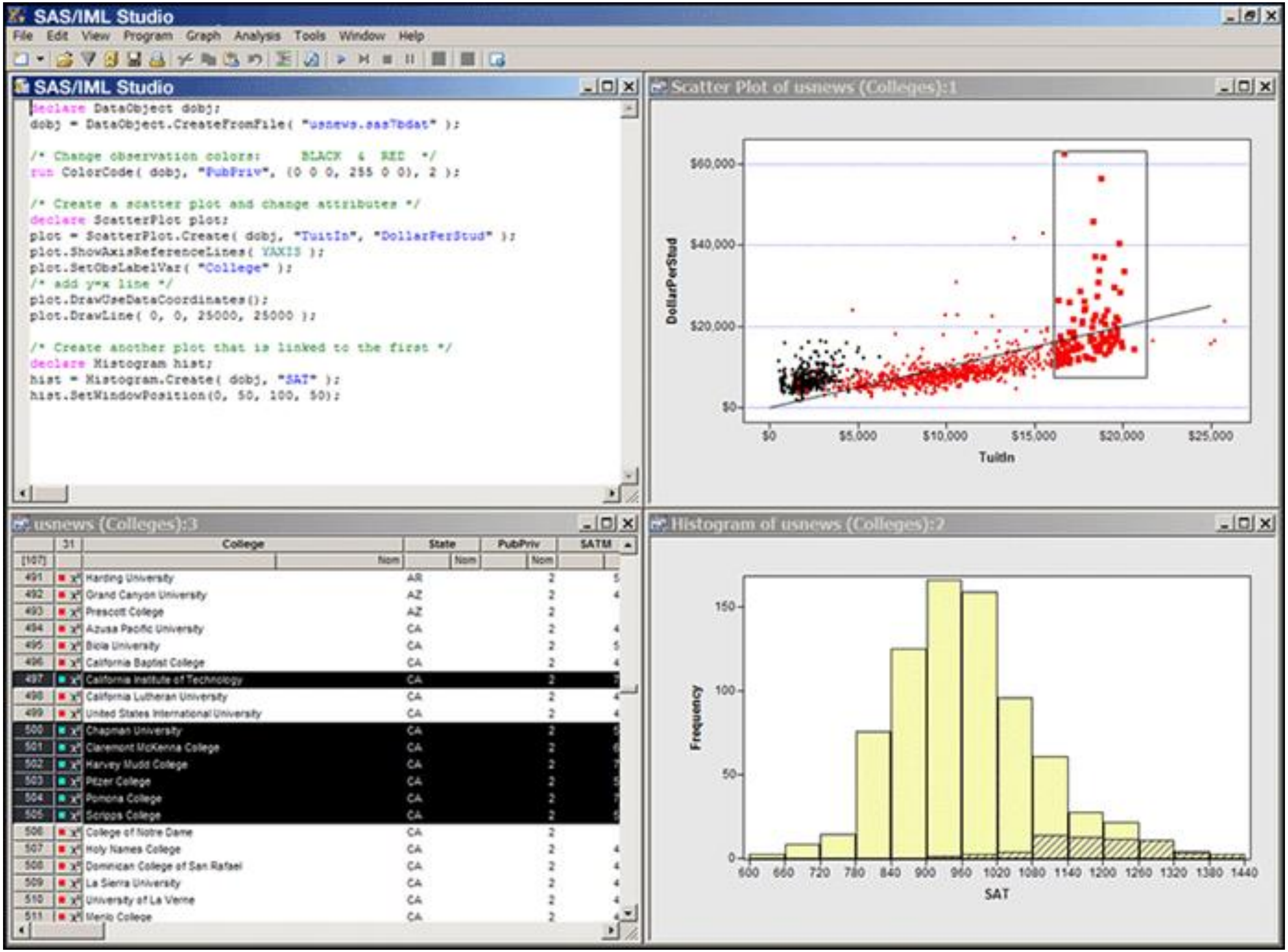
The line shows the median home sales price for the Seattle area. The individual sales are shown as circles, color coded by zip code. Highlighting zip code 98105 confirms this includes the more expensive parts of town.

Source: Tableau



	patid				gender	active	obesity	diabetes	bp
1	9735702127	29		-54	Female	Yes	No	No	Hypotension
2	4852351830	79		-74	Male	Yes	Yes	No	Hypertension
3	3434994256	79		-74	Female	Yes	Yes	Yes	Hypertension
4	6053971728	82		-74	Male	Yes	No	No	Normal
5	9370757269	29					No	No	Hypertension
6	3537185320	29					Yes	No	Normal
7	0275365329	82					No	Yes	Normal
8	3906583332	79					No	No	Normal
9	4785366661	82					No	No	Normal
10	9589919145	82					No	No	Hypertension
11	4598012219	79					Yes	No	Normal
12	3629441662	79					No	No	Normal
13	5307816588	79					No	No	Hypotension
14	5357069859	82					Yes	No	Normal
15	5132742071	29					Yes	Yes	Normal
16	2660586207	29					Yes	No	Hypertension
17	5408312498	79					No	No	Hypertension
18	9069087682	29		-64	Male	Yes	No	No	Hypertension
19	8173197592	799998	58	55-64	Female	No	No	No	Normal
20	8808732689	822229	83	75+	Male	Yes	No	No	Hypotension
21	5666440246	822229	67	65-74	Female	Yes	Yes	No	Normal

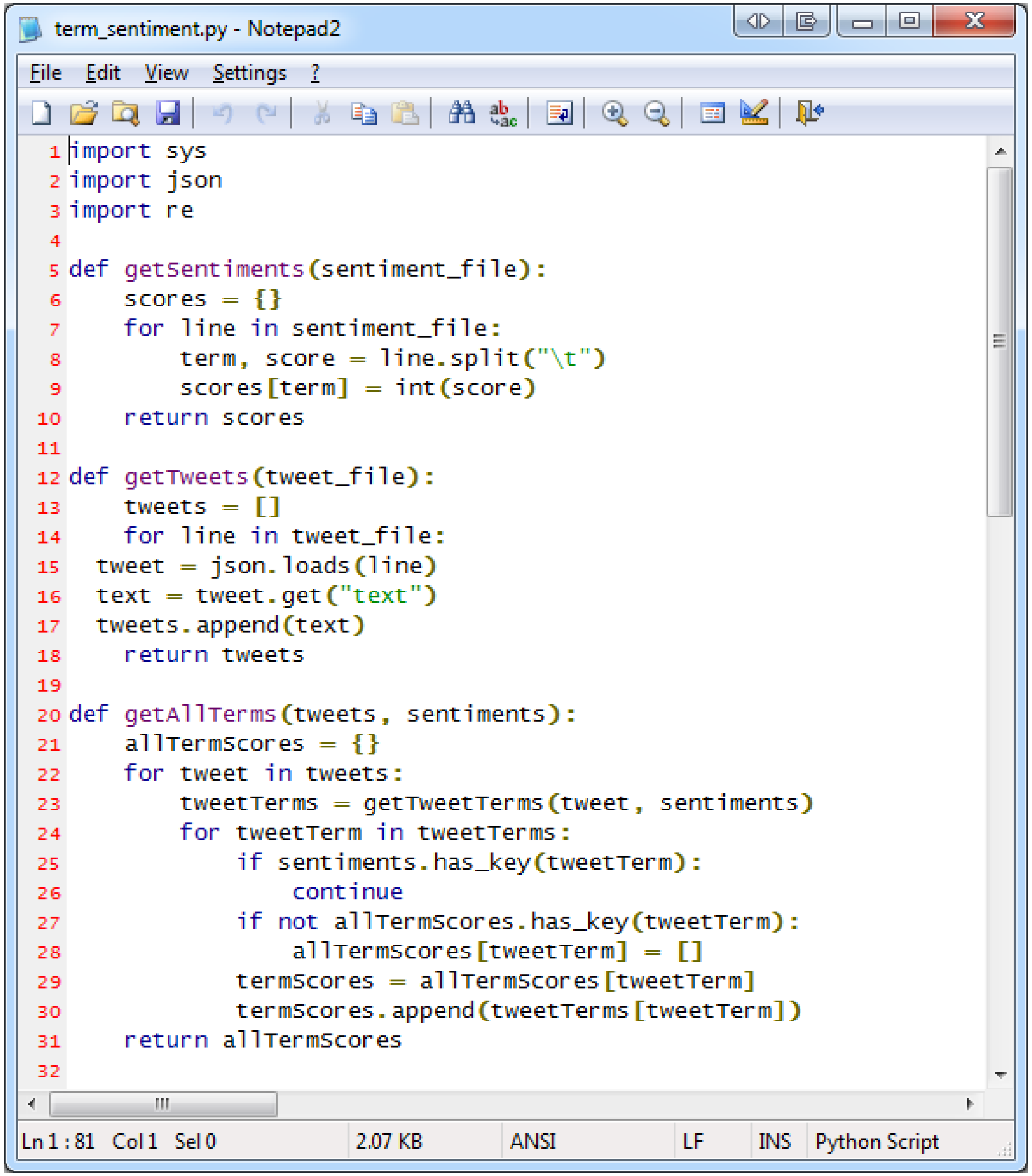
Source: IBM SPSS



```
declare DataObject dobj;
dobj = DataObject.CreateFromFile( "usnews.sasTbdat" );

/* 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);
```

	31	College	State	PubPriv	SATM
[107]		Nom	Nom	Nom	
491	🟥 x	Harding University	AR	2	5
492	🟥 x	Grand Canyon University	AZ	2	4
493	🟥 x	Prescott College	AZ	2	
494	🟥 x	Azusa Pacific University	CA	2	4
495	🟥 x	Biola University	CA	2	5
496	🟥 x	California Baptist College	CA	2	4
497	🟩 x	California Institute of Technology	CA	2	7
498	🟥 x	California Lutheran University	CA	2	4
499	🟥 x	United States International University	CA	2	4
500	🟩 x	Chapman University	CA	2	5
501	🟩 x	Claremont McKenna College	CA	2	6
502	🟩 x	Harvey Mudd College	CA	2	7
503	🟩 x	Pitzer College	CA	2	5
504	🟩 x	Pomona College	CA	2	7
505	🟩 x	Scripps College	CA	2	5
506	🟥 x	College of Notre Dame	CA	2	
507	🟥 x	Holy Names College	CA	2	4
508	🟥 x	Dominican College of San Rafael	CA	2	4
509	🟥 x	La Sierra University	CA	2	4
510	🟥 x	University of La Verne	CA	2	4
511	🟥 x	Menlo College	CA	2	4

Source: SAS



```
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

```

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