Exploratory Data Analysis with R

@matthewrenze

#codemash

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

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...

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

Source: Wikipedia

Sink or Swim

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
  - B.A. in Philosophy
- Community
  - Pluralsight Author
  - ASPInsider
  - Public Speaker
  - Open-Source Software

IOWA STATE UNIVERSITY

PLURALSIGHT

Introduction to R

What is R?

- Open source
- Language and environment
- Numerical and graphical analysis
- Cross platform

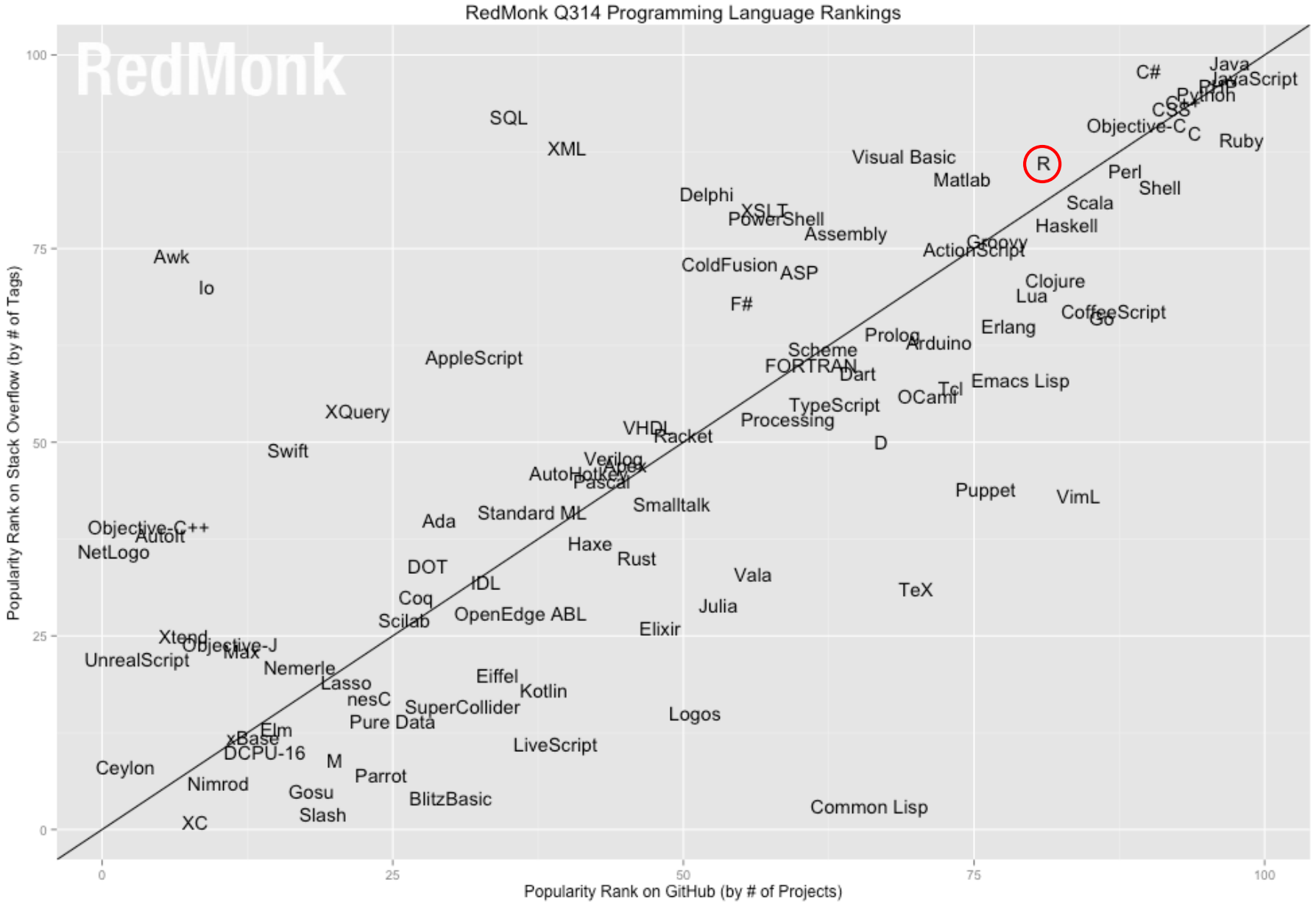
What is R?

- 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/

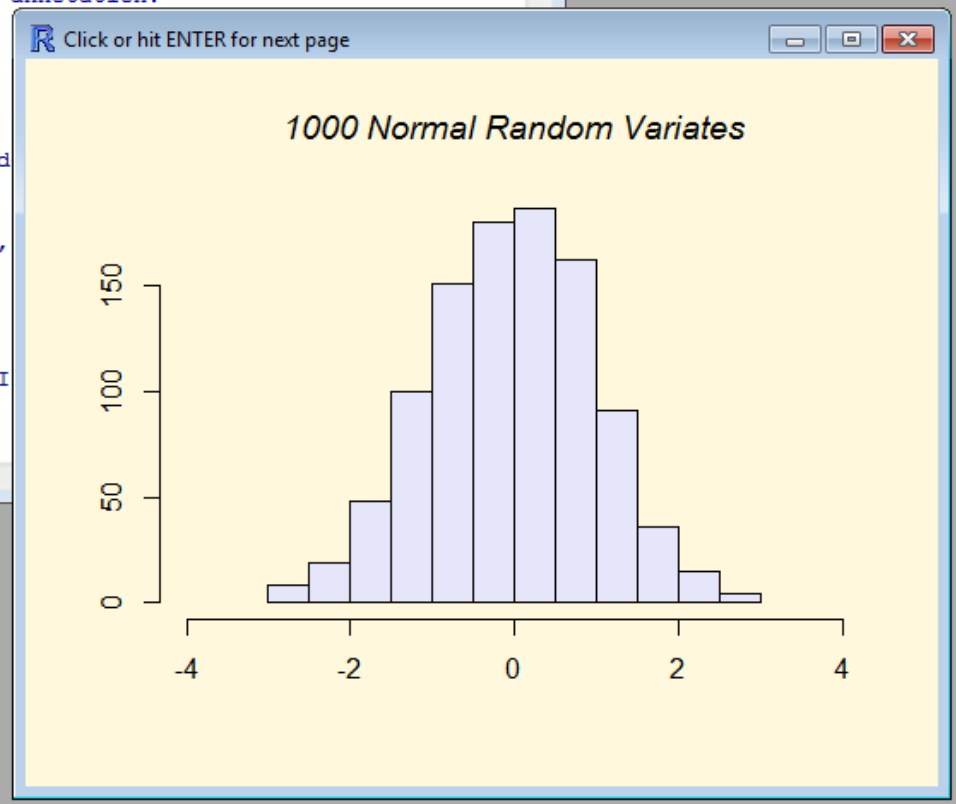
RGui (64-bit)

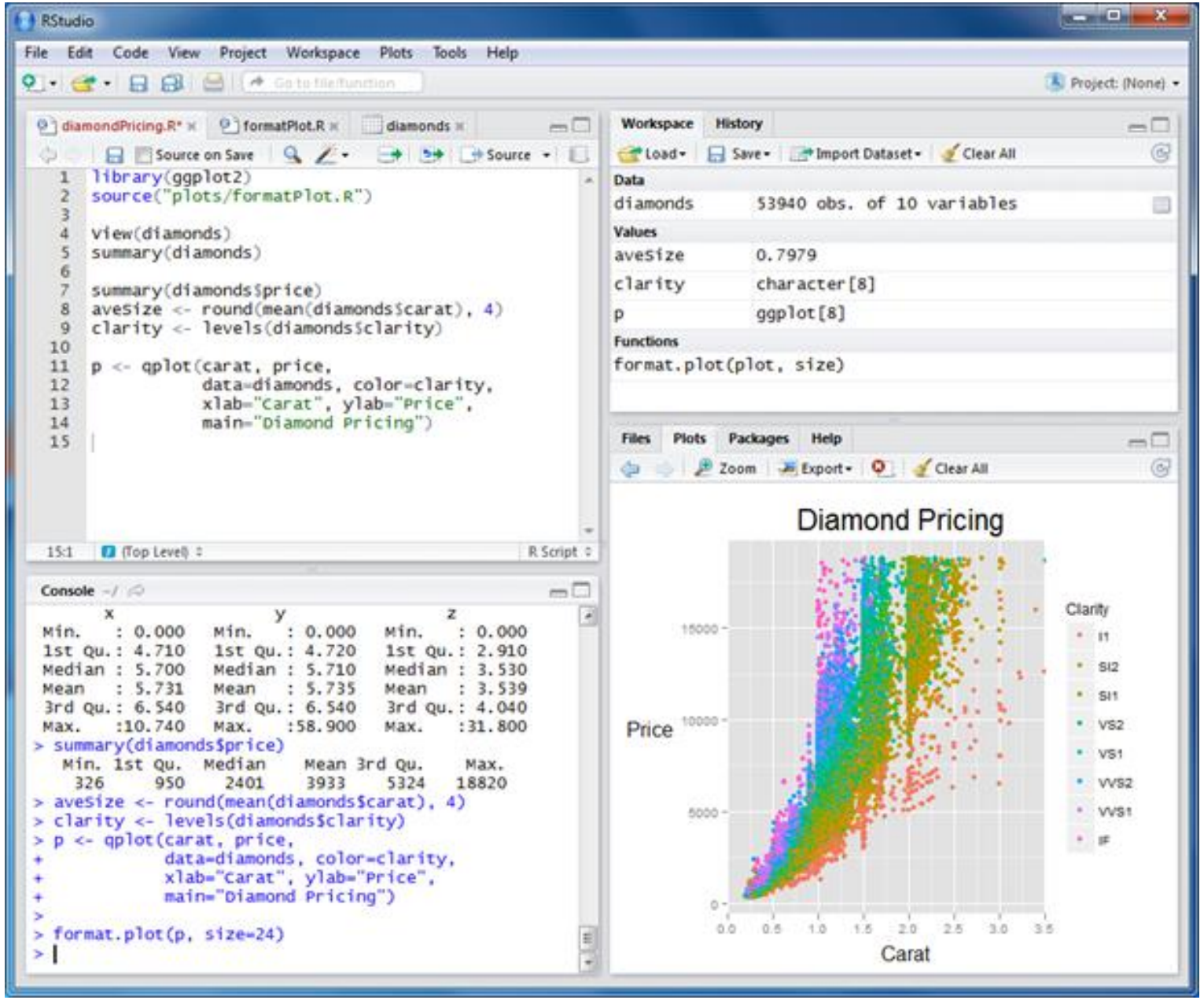
Next

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





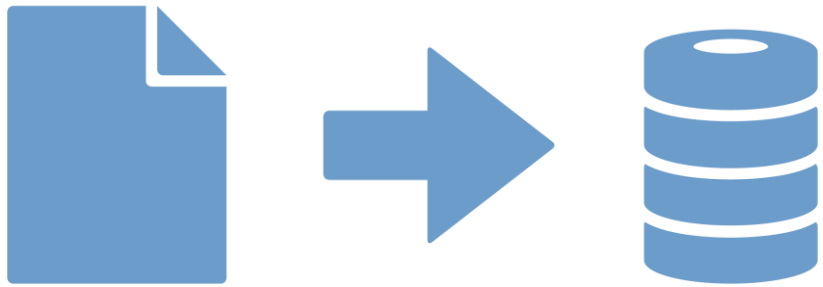
Source: www.rstudio.com/ide/

Code Demo

Data Munging

Data Munging

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



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


PROD. NO.
SCENE
TAKE
ROLL

- Column with wrong name
- Rows with missing values
- Runtime column has units
- Revenue in multiple scales
- Wrong file format

Code Demo

Descriptive Statistics

Descriptive Statistics

- Describe data
- Provides a summary
- aka: Summary statistics

Movie Runtime	
Statistic	**Value (minutes)**
Minimum	38
1st Quartile	93
Median	101
Mean	104
3rd Quartile	113
Maximum	219

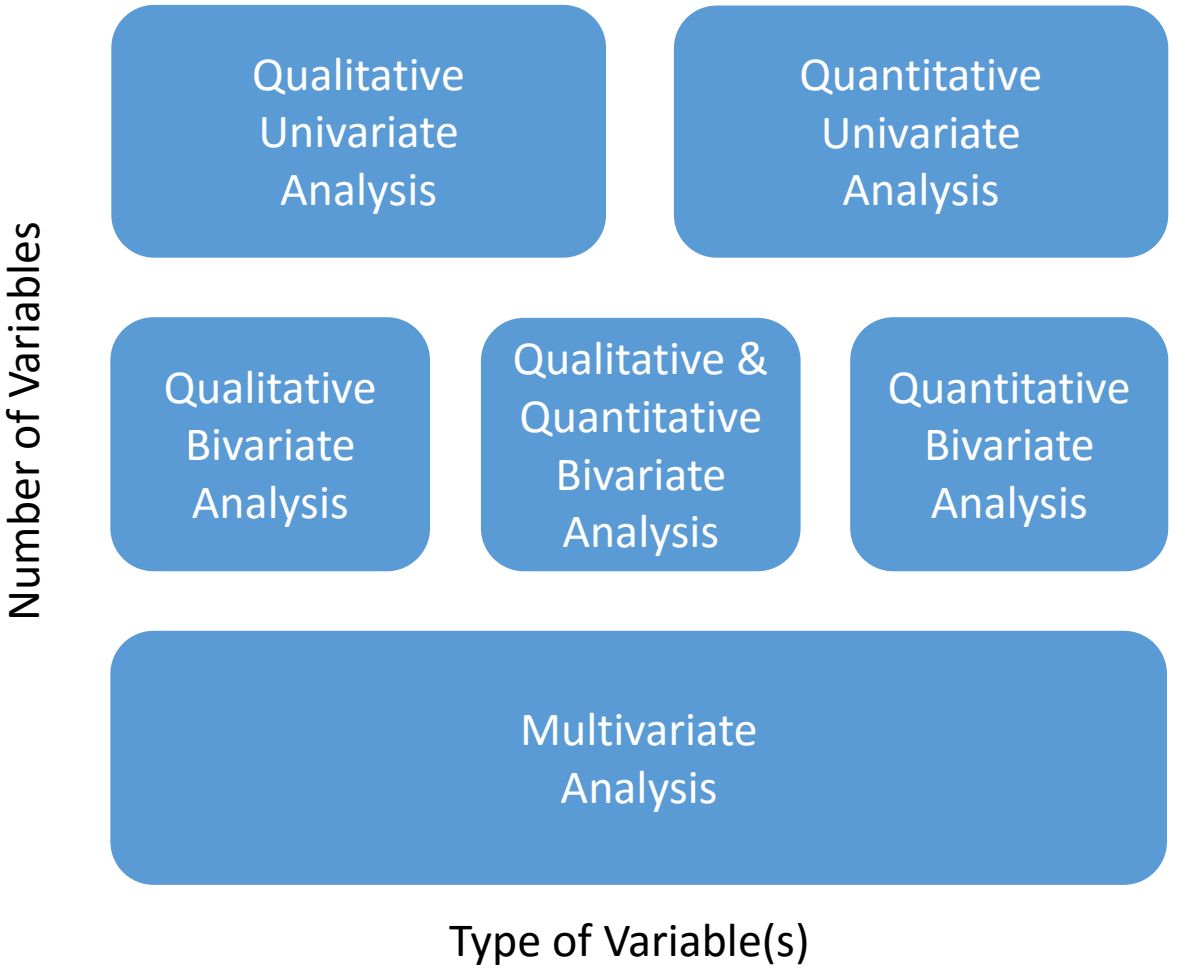
Statistical Terms

- Observations
- Variables
- Qualitative variable
- Quantitative variable

ID	Date	Customer	Product	Quantity
1	2015-08-27	John	Pizza	2
2	2015-08-27	John	Soda	2
3	2015-08-27	Jill	Salad	1
4	2015-08-27	Jill	Milk	1
5	2015-08-28	Miko	Pizza	3
6	2015-08-28	Miko	Soda	2
7	2015-08-28	Sam	Pizza	1
8	2015-08-28	Sam	Milk	1

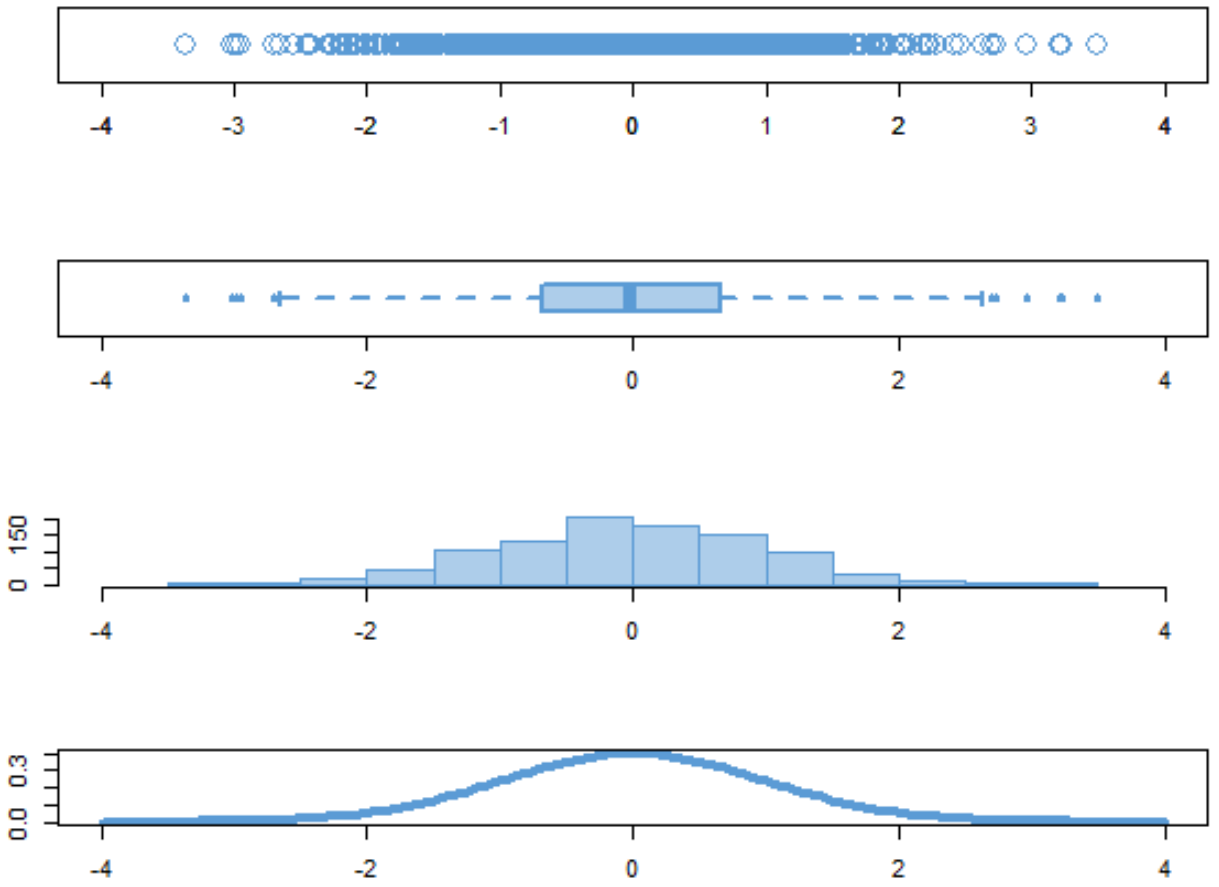
Types of Analysis

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



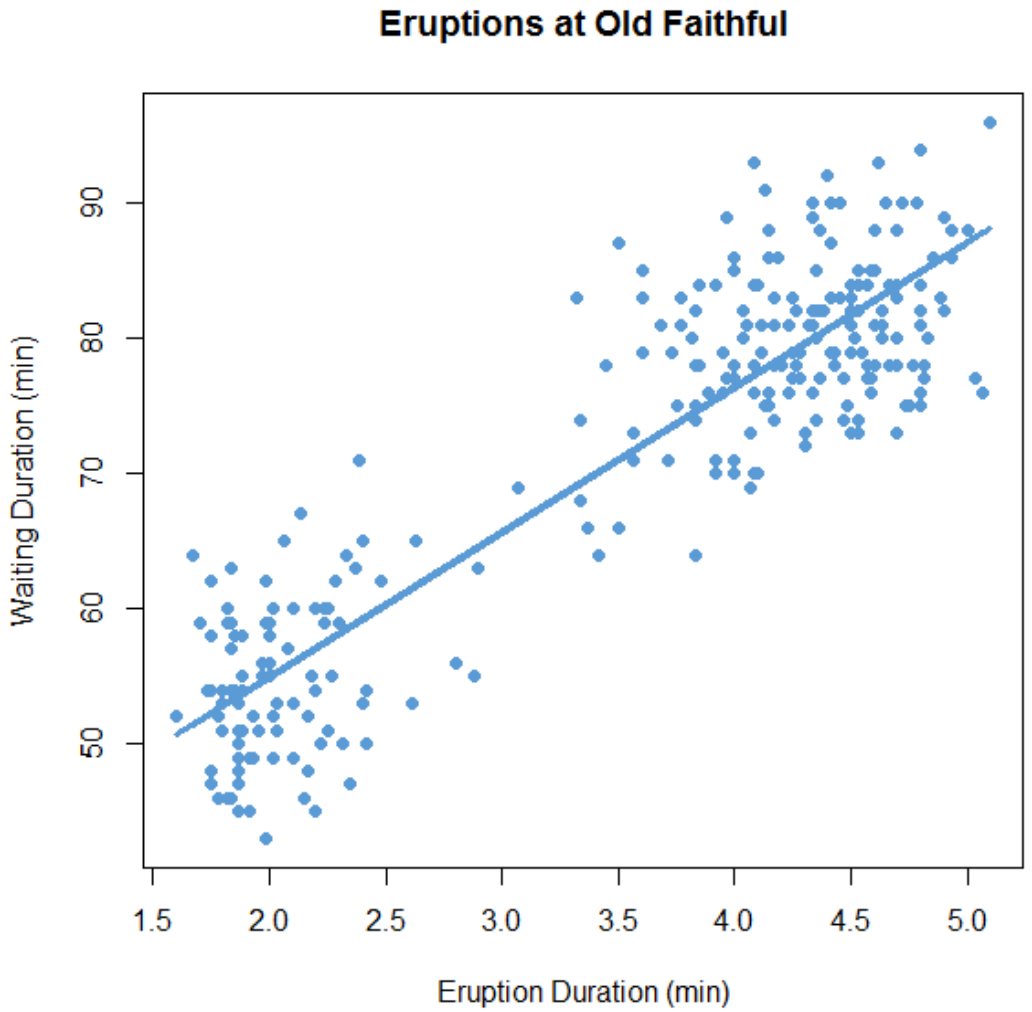
Univariate Analysis

- One variable
- Qualitative
  - Frequency
- Quantitative
  - Central tendency
  - Dispersion



Bivariate Analysis

- Qualitative
  - Joint frequency
- Quantitative
  - Two variables
    - Predictor
    - Outcome
  - Measures
    - Covariance
    - Correlation




COWBOYS &
Space Invaders:
The Musical
Extended Edition
R

Code Demo

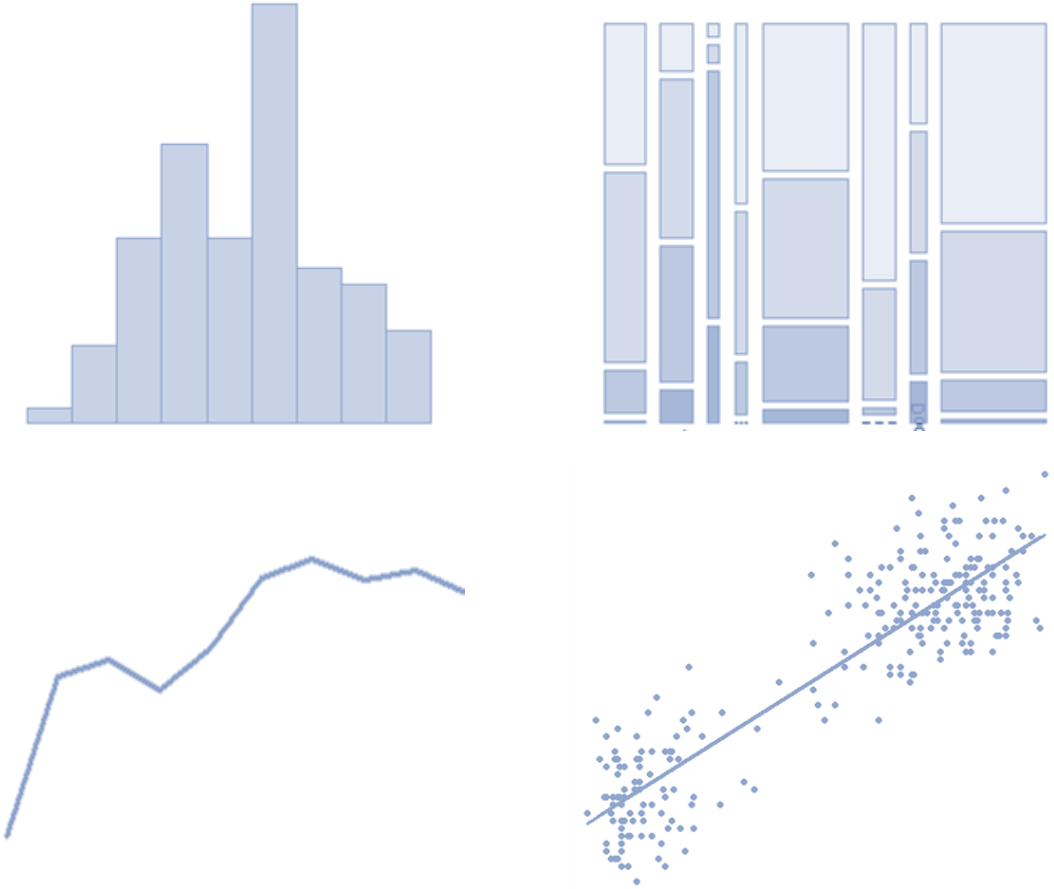

COWBOYS &
Space Invaders:
The Musical
Extended Edition
R


COWBOYS &
Extended Edition
R

Data Visualization

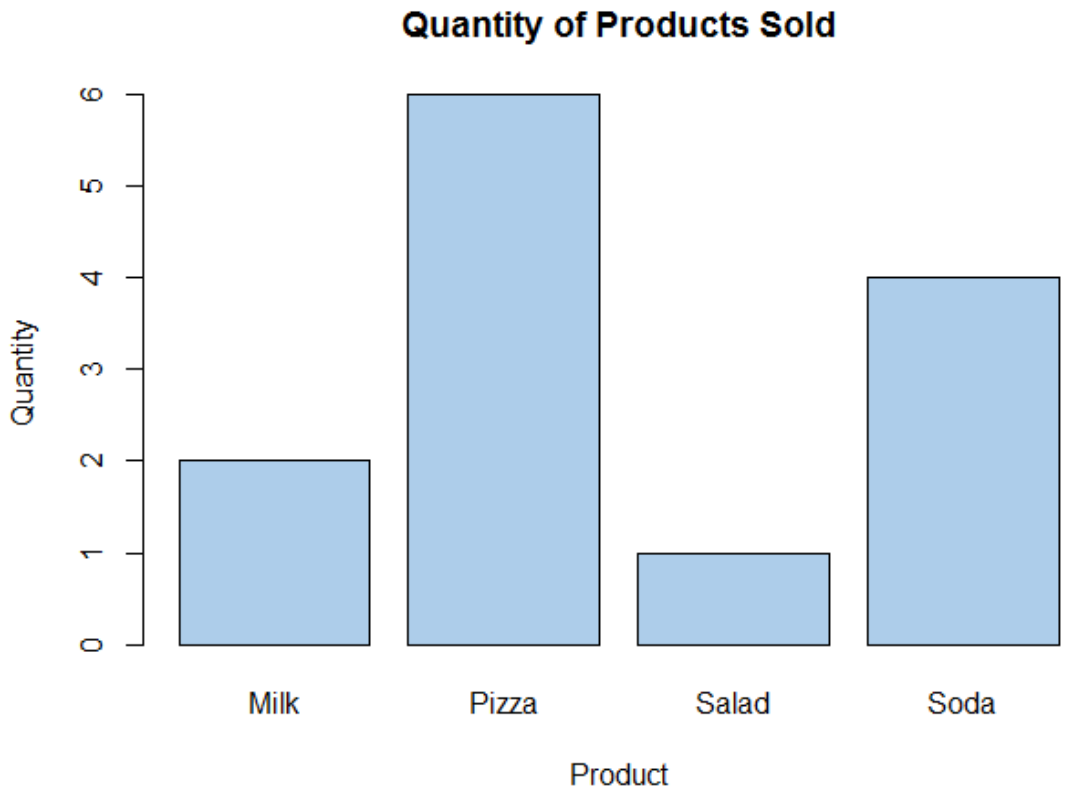
Data Visualization

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



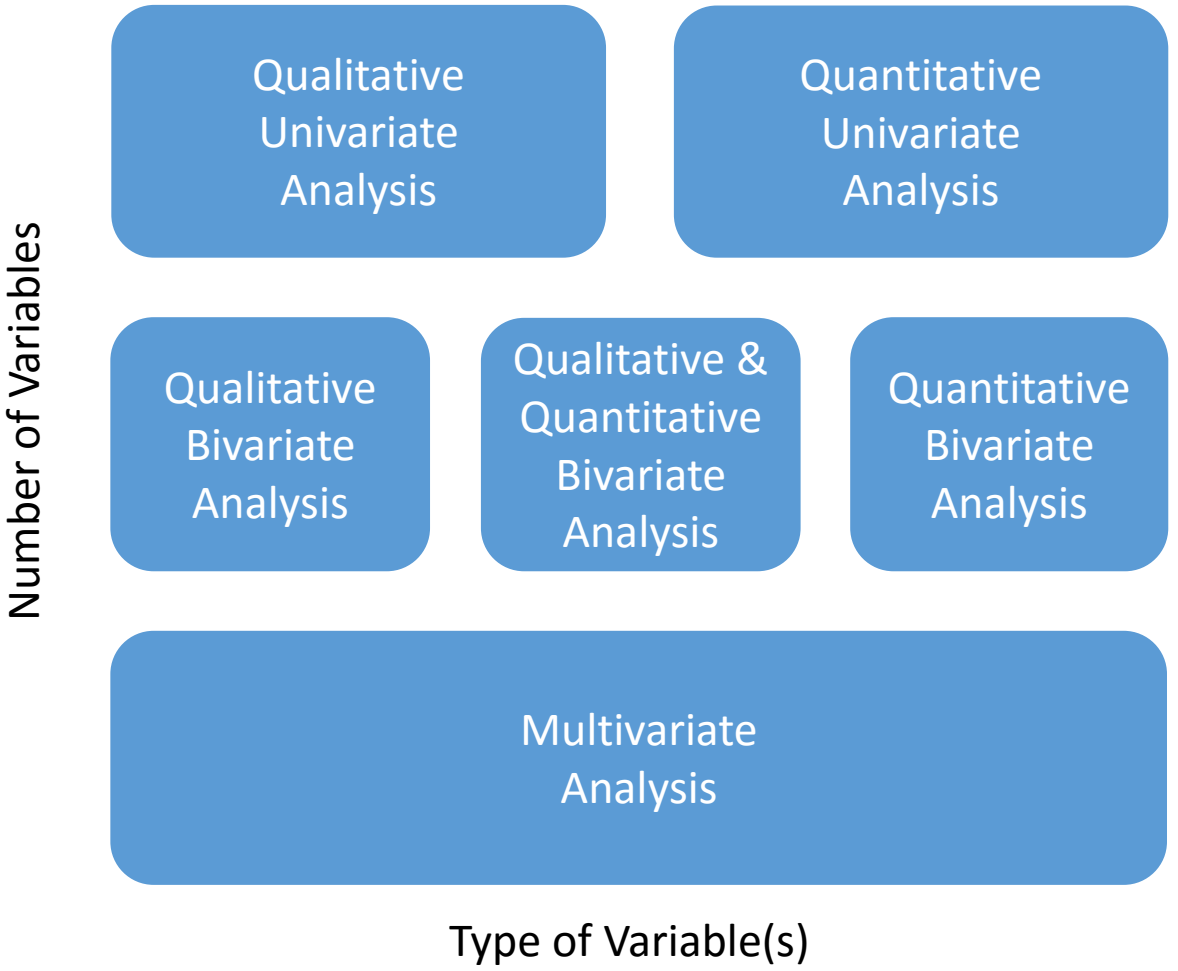
Data Visualization

ID	Date	Customer	Product	Quantity
1	2015-08-27	John	Pizza	2
2	2015-08-27	John	Soda	2
3	2015-08-27	Jill	Salad	1
4	2015-08-27	Jill	Milk	1
5	2015-08-28	Miko	Pizza	3
6	2015-08-28	Miko	Soda	2
7	2015-08-28	Sam	Pizza	1
8	2015-08-28	Sam	Milk	1



Types of Data Visualizations

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




COWBOYS &
Space Invaders:
The Musical
Extended Edition
R

Code Demo

The
Warlord
of the
Rings :
An Unexpected Adventure

Feature Length

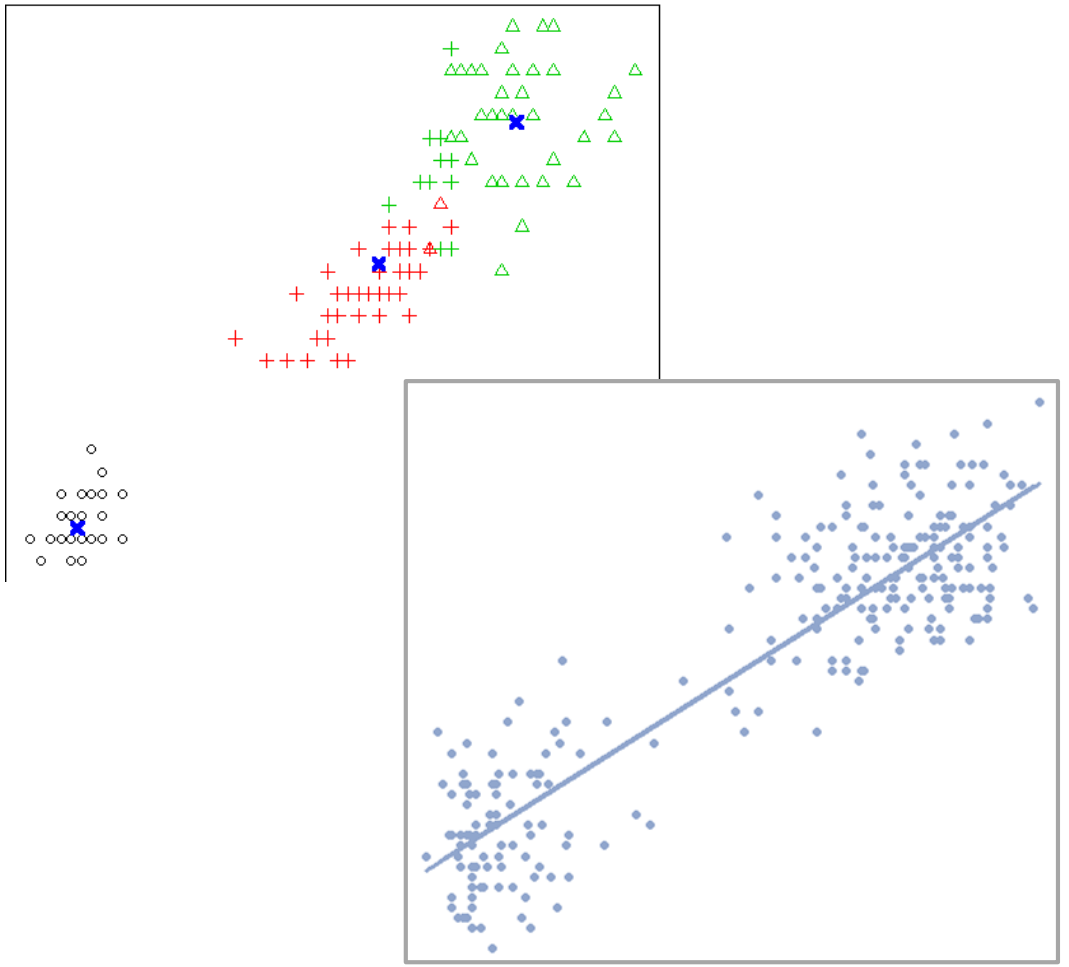
PG

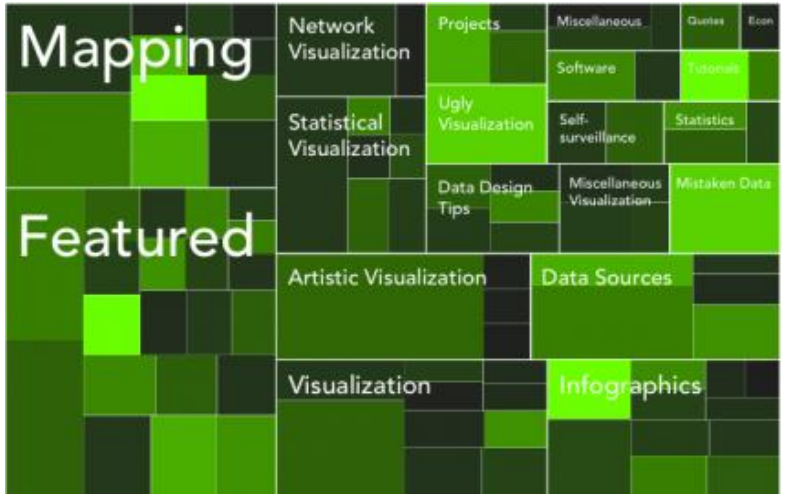
Beyond R and EDA

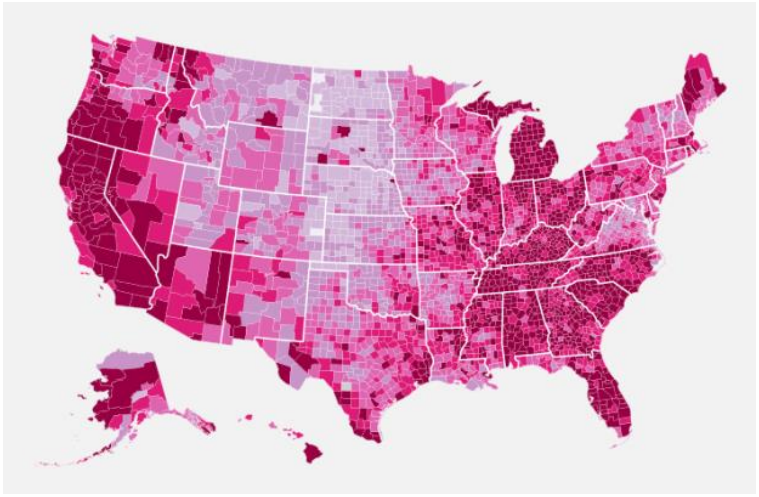

This is just the tip of the iceberg!

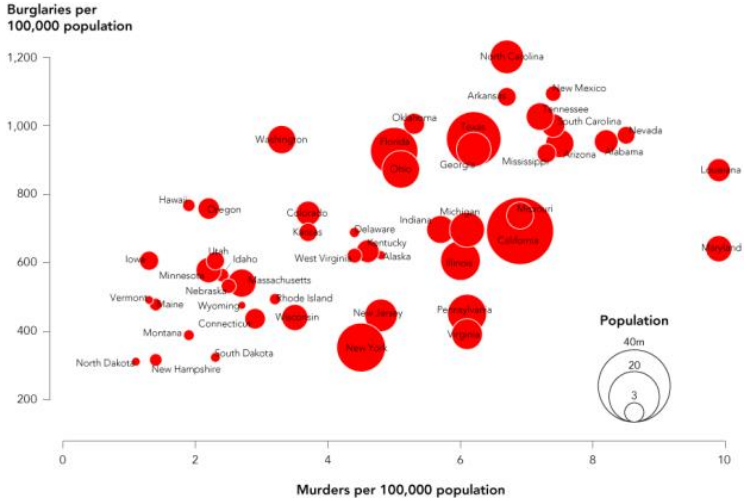
Advanced Data Analysis with R

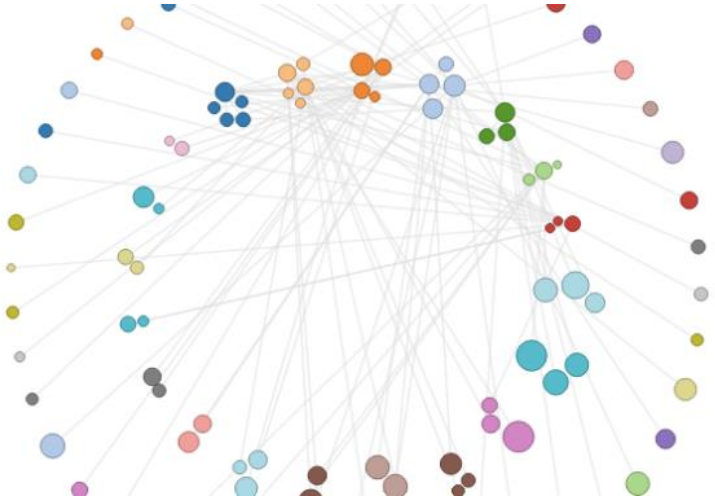
- Cluster Analysis
- Statistical Modeling
- Dimensionality Reduction
- Analysis of Variance (ANOVA)

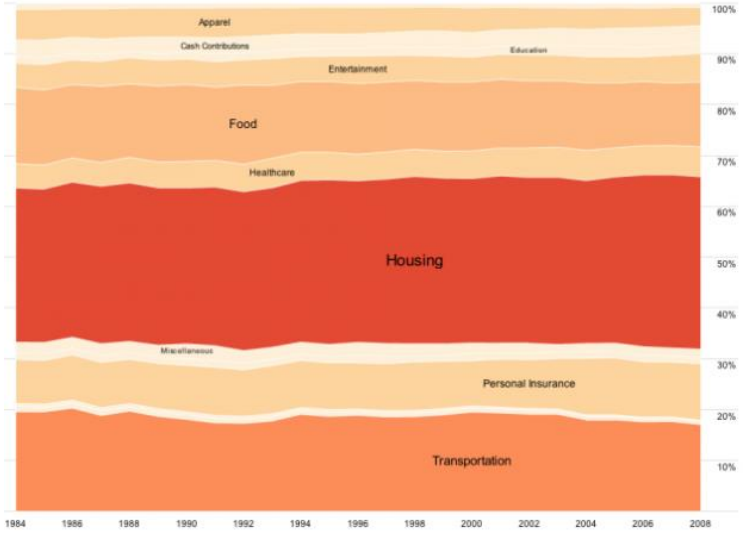


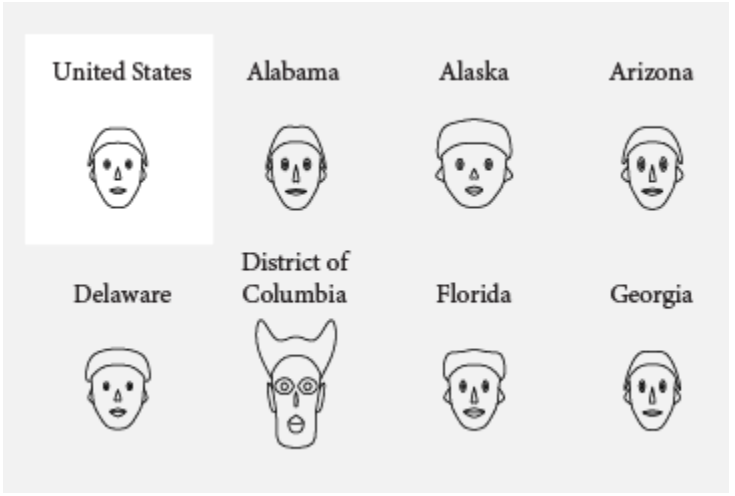






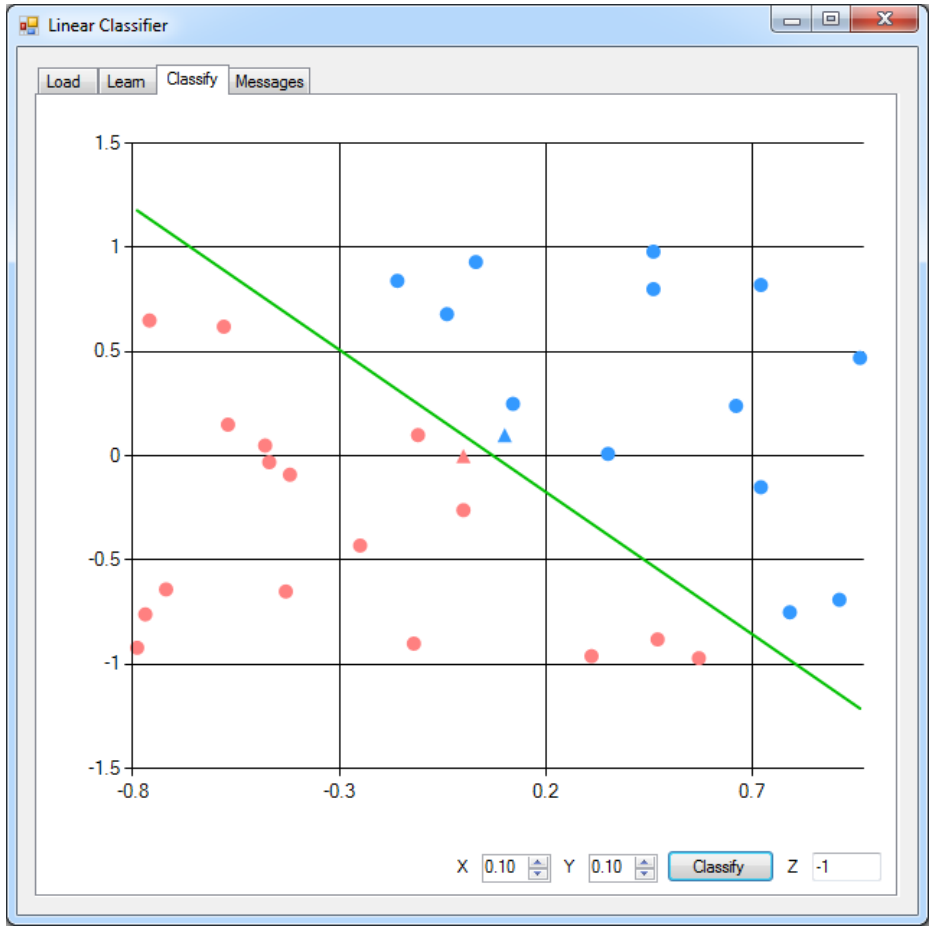


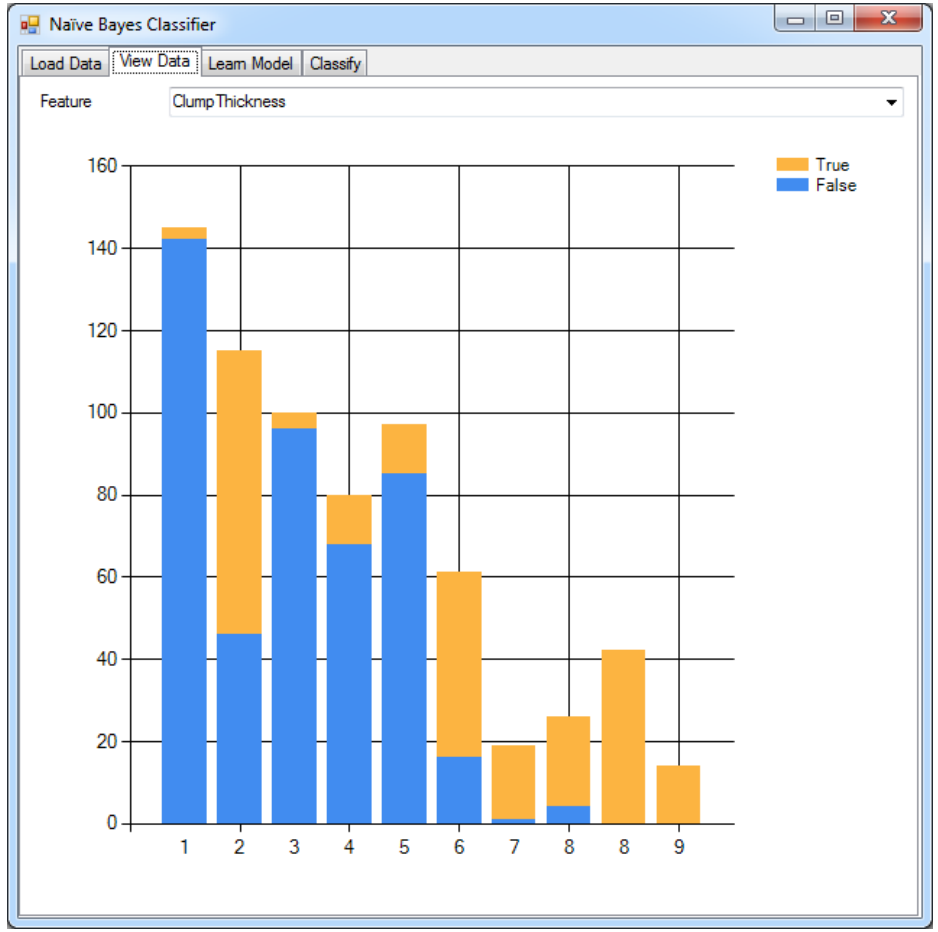




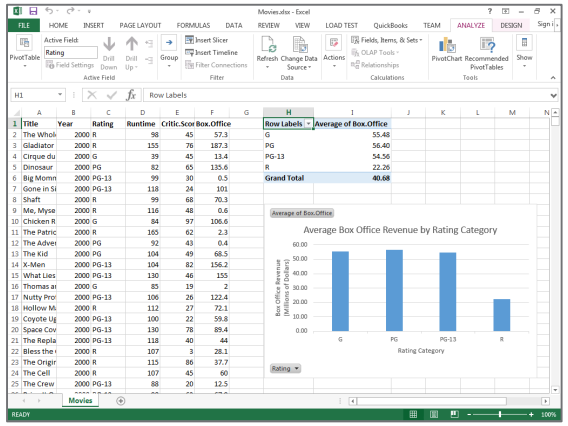
Source: Nathan Yau (www.flowingdata.com)

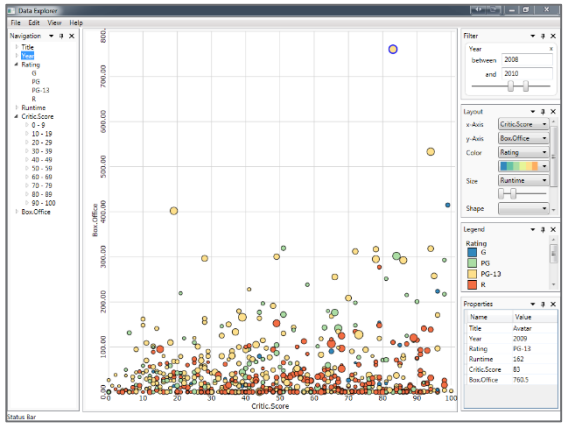
Data Mining and Machine Learning with R

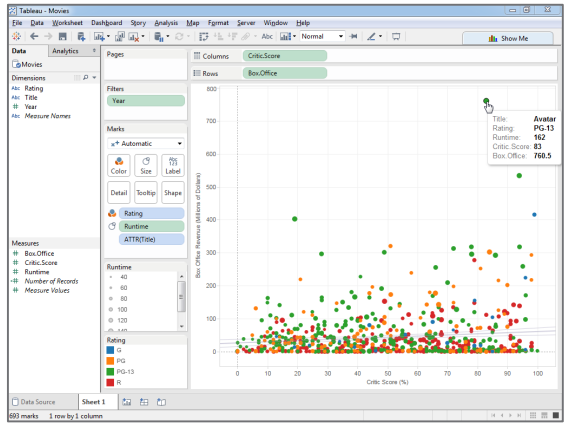


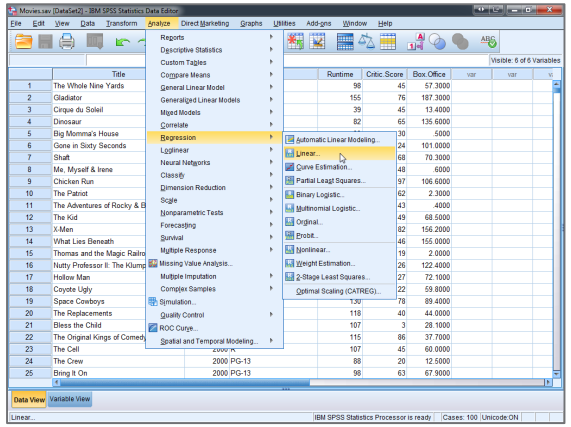


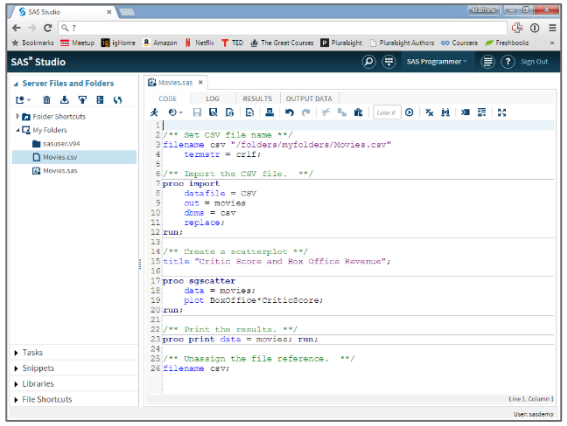
Alternatives to R for EDA

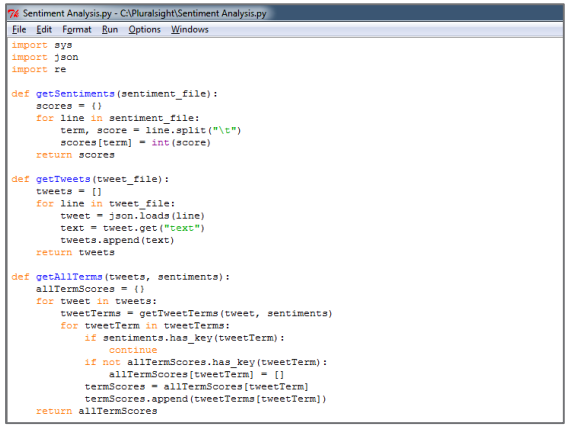


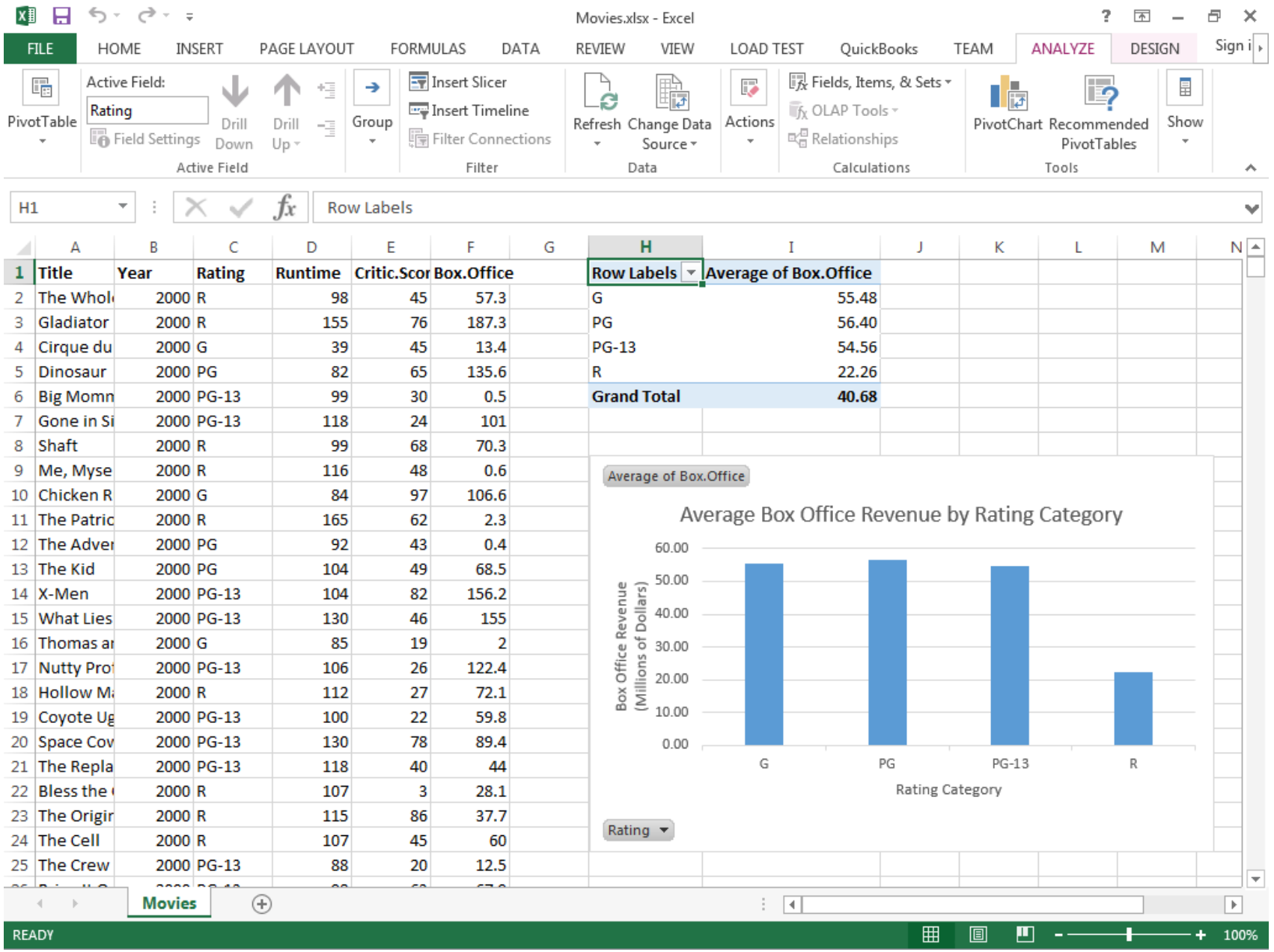






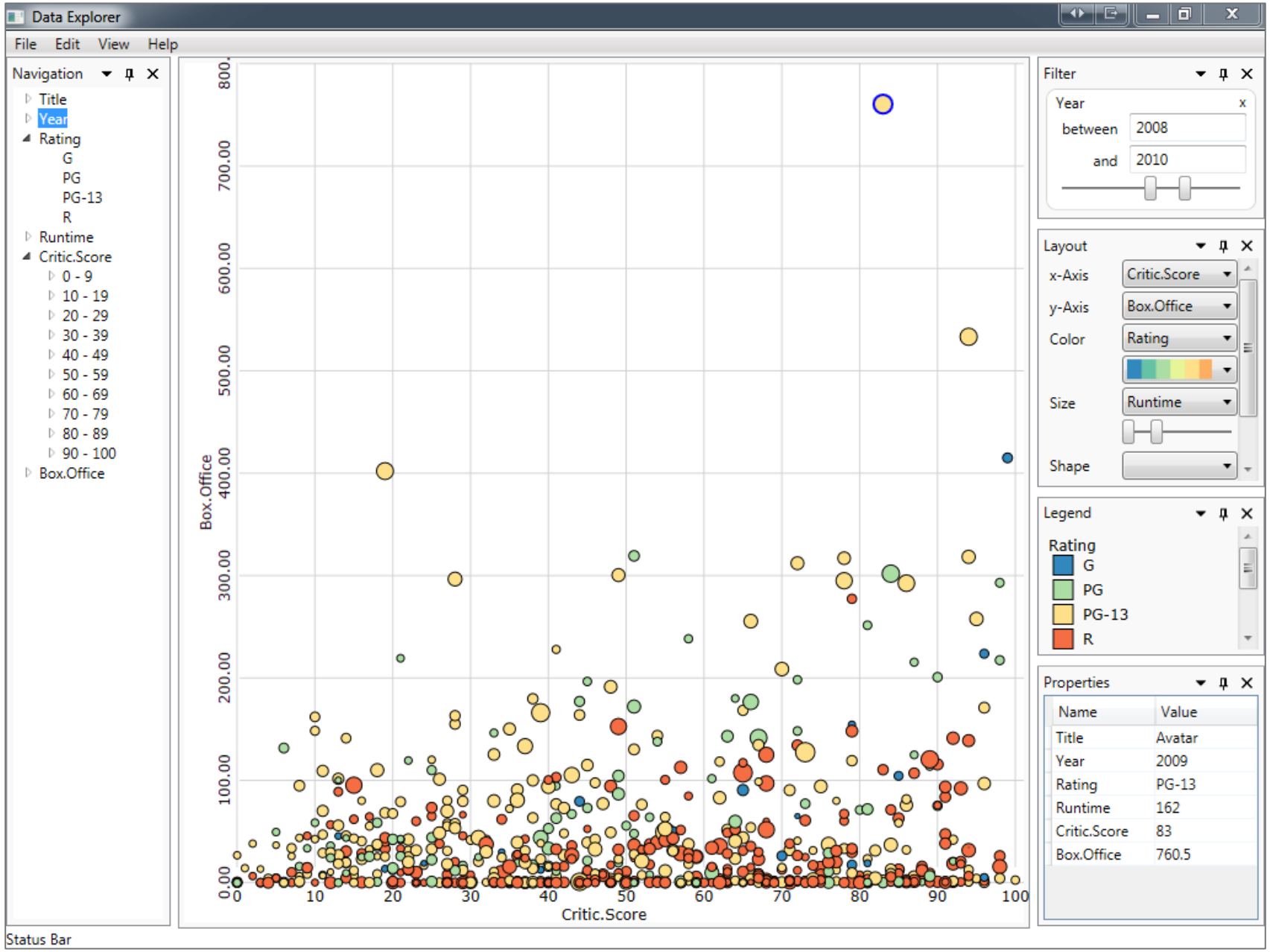






	Title	Year	Rating	Runtime	Critic.Scor	Box.Office
2	The Whol	2000	R	98	45	57.3
3	Gladiator	2000	R	155	76	187.3
4	Cirque du	2000	G	39	45	13.4
5	Dinosaur	2000	PG	82	65	135.6
6	Big Momn	2000	PG-13	99	30	0.5
7	Gone in Si	2000	PG-13	118	24	101
8	Shaft	2000	R	99	68	70.3
9	Me, Myse	2000	R	116	48	0.6
10	Chicken R	2000	G	84	97	106.6
11	The Patric	2000	R	165	62	2.3
12	The Adver	2000	PG	92	43	0.4
13	The Kid	2000	PG	104	49	68.5
14	X-Men	2000	PG-13	104	82	156.2
15	What Lies	2000	PG-13	130	46	155
16	Thomas ar	2000	G	85	19	2
17	Nutty Prof	2000	PG-13	106	26	122.4
18	Hollow Ma	2000	R	112	27	72.1
19	Coyote Ug	2000	PG-13	100	22	59.8
20	Space Cov	2000	PG-13	130	78	89.4
21	The Repla	2000	PG-13	118	40	44
22	Bless the	2000	R	107	3	28.1
23	The Origir	2000	R	115	86	37.7
24	The Cell	2000	R	107	45	60
25	The Crew	2000	PG-13	88	20	12.5

Source: Microsoft


Data Explorer
File
Edit
View
Help
Navigation
Title
Year
Rating
G
PG
PG-13
R
Runtime
Critic.Score
0 - 9
10 - 19
20 - 29
30 - 39
40 - 49
50 - 59
60 - 69
70 - 79
80 - 89
90 - 100
Box.Office
Box.Office
0.00
100.00
200.00
300.00
400.00
500.00
600.00
700.00
800.00
0
10
20
30
40
50
60
70
80
90
100
Critic.Score
Filter
Year
between
2008
and
2010
Layout
x-Axis
Critic.Score
y-Axis
Box.Office
Color
Rating
Size
Runtime
Shape
Legend
Rating
G
PG
PG-13
R
Properties
Name
Value
Title
Avatar
Year
2009
Rating
PG-13
Runtime
162
Critic.Score
83
Box.Office
760.5
Status Bar

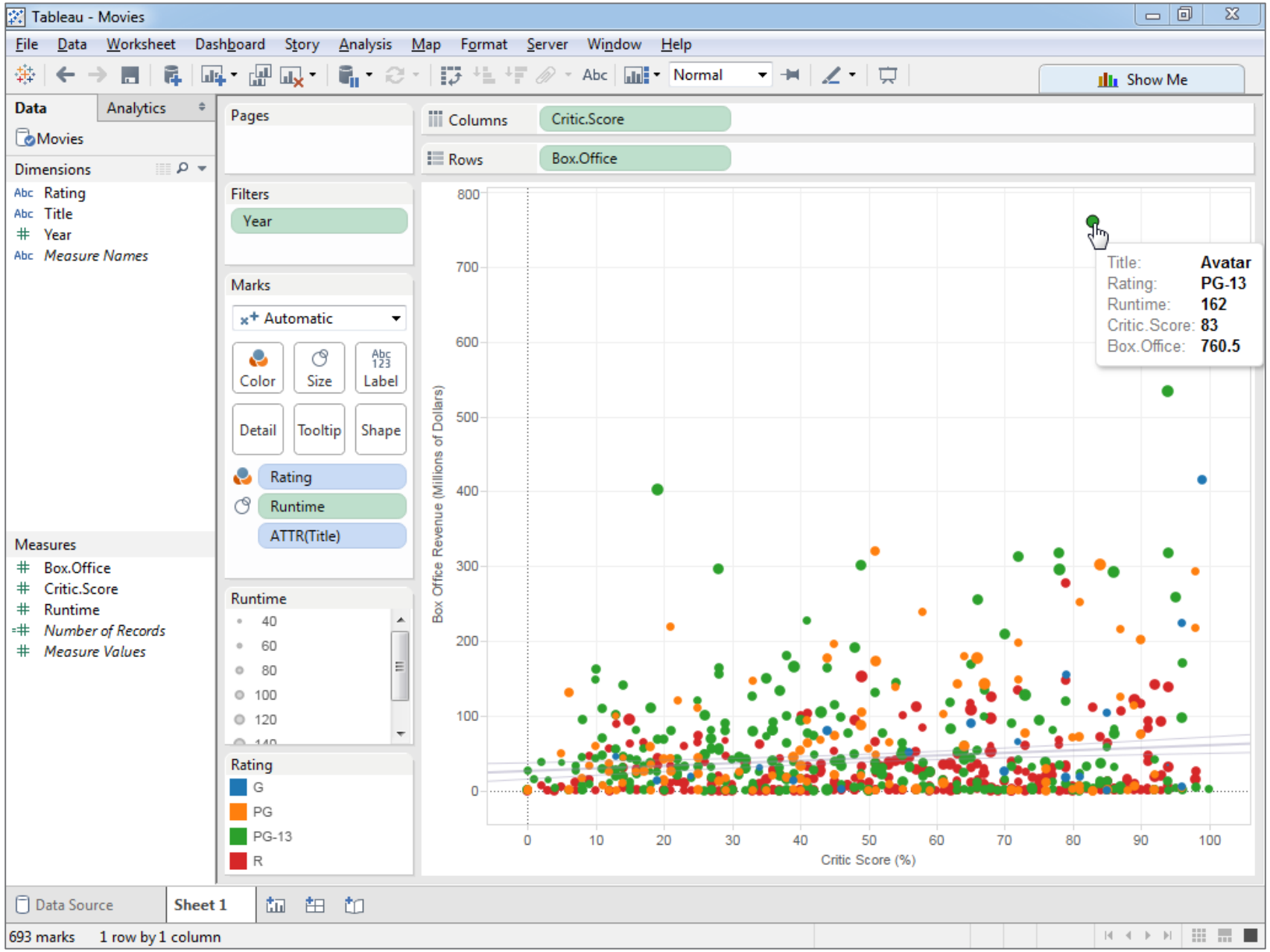
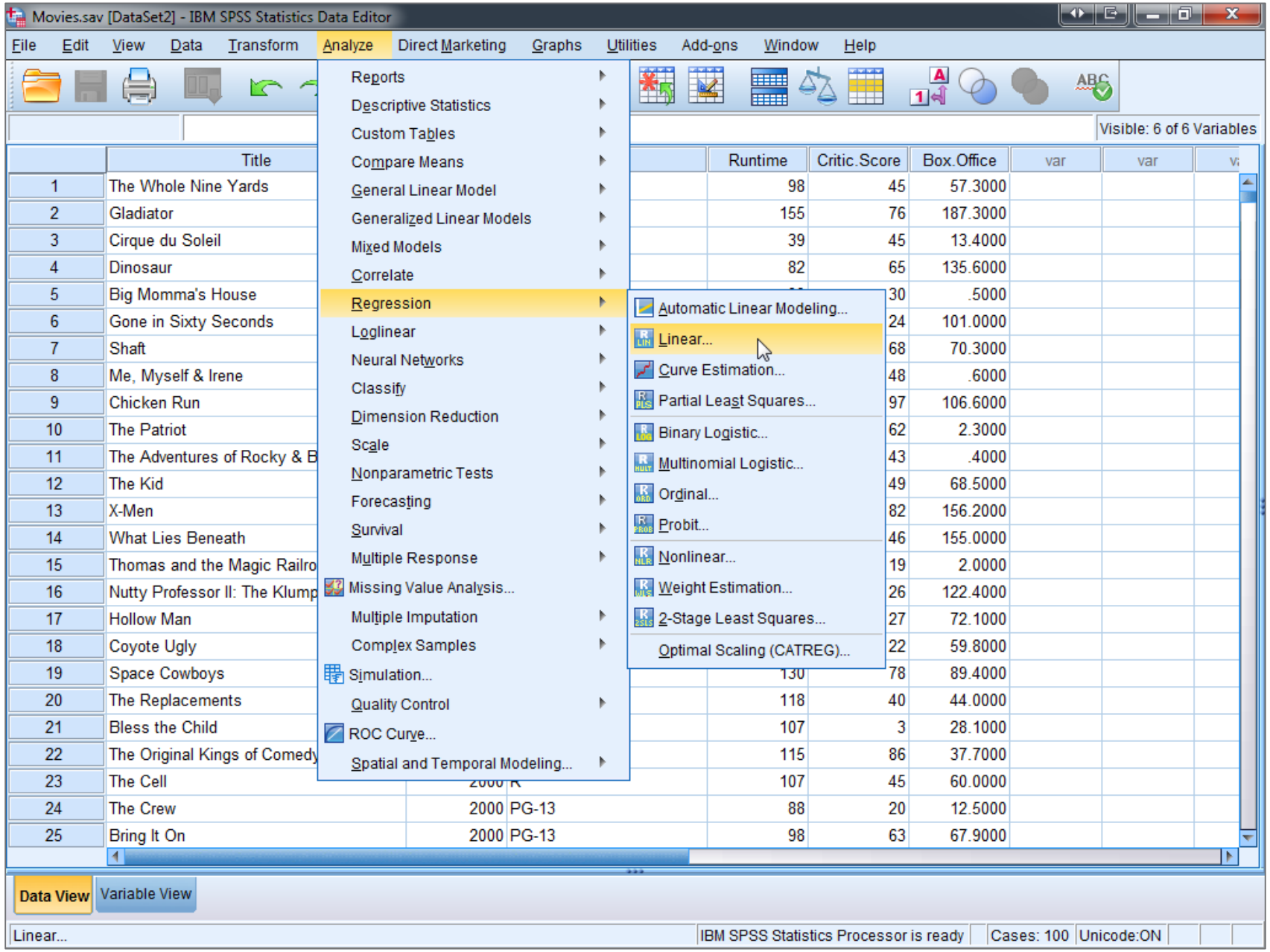
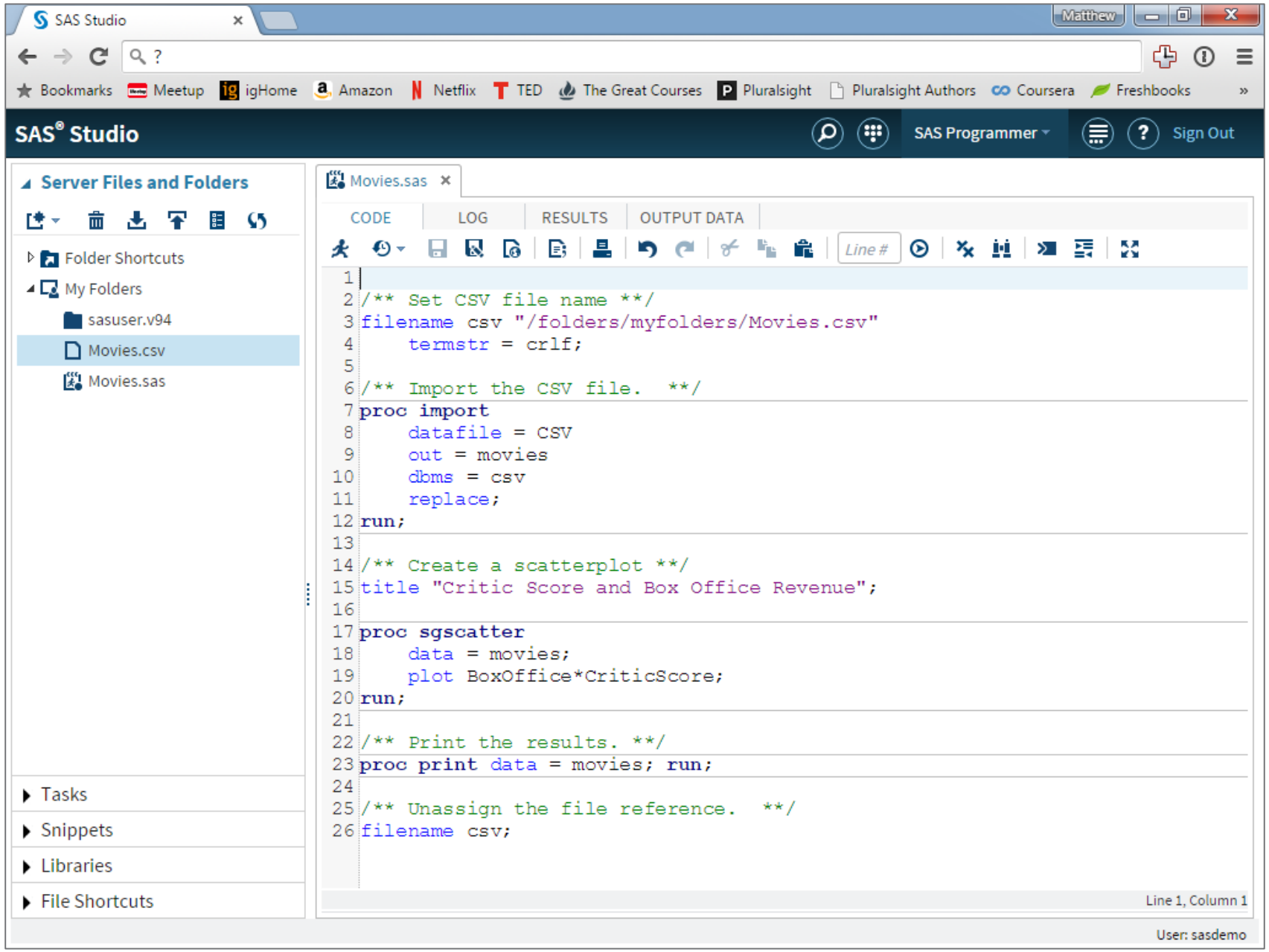


Tableau - Movies
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Normal
Show Me
Data
Analytics
Movies
Dimensions
Abc Rating
Abc Title
Year
Abc Measure Names
Measures
Box.Office
Critic.Score
Runtime
Number of Records
Measure Values
Pages
Filters
Year
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Shape
Rating
Runtime
ATTR(Title)
Runtime
40
60
80
100
120
Rating
G
PG
PG-13
R
Columns
Critic.Score
Rows
Box.Office
Title: Avatar
Rating: PG-13
Runtime: 162
Critic.Score: 83
Box.Office: 760.5
Box Office Revenue (Millions of Dollars)
800
700
600
500
400
300
200
100
0
0 10 20 30 40 50 60 70 80 90 100
Critic Score (%)
Data Source
Sheet 1
693 marks
1 row by 1 column


Movies.sav [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Custom Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Spatial and Temporal Modeling...
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling (CATREG)...
Visible: 6 of 6 Variables
Title
Runtime
Critic.Score
Box.Office
var
var
1 The Whole Nine Yards 98 45 57.3000
2 Gladiator 155 76 187.3000
3 Cirque du Soleil 39 45 13.4000
4 Dinosaur 82 65 135.6000
5 Big Momma's House 30 .5000
6 Gone in Sixty Seconds 24 101.0000
7 Shaft 68 70.3000
8 Me, Myself & Irene 48 .6000
9 Chicken Run 97 106.6000
10 The Patriot 62 2.3000
11 The Adventures of Rocky & B 43 .4000
12 The Kid 49 68.5000
13 X-Men 82 156.2000
14 What Lies Beneath 46 155.0000
15 Thomas and the Magic Railro 19 2.0000
16 Nutty Professor II: The Klump 26 122.4000
17 Hollow Man 27 72.1000
18 Coyote Ugly 22 59.8000
19 Space Cowboys 130 78 89.4000
20 The Replacements 118 40 44.0000
21 Bless the Child 107 3 28.1000
22 The Original Kings of Comedy 115 86 37.7000
23 The Cell 107 45 60.0000
24 The Crew 2000 PG-13 88 20 12.5000
25 Bring It On 2000 PG-13 98 63 67.9000
Data View
Variable View
Linear...
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON


SAS Studio
Matthew
Bookmarks
Meetup
igHome
Amazon
Netflix
TED
The Great Courses
Pluralsight
Pluralsight Authors
Coursera
Freshbooks
SAS® Studio
SAS Programmer
Sign Out
Server Files and Folders
Folder Shortcuts
My Folders
sasuser.v94
Movies.csv
Movies.sas
Tasks
Snippets
Libraries
File Shortcuts
Movies.sas
CODE
LOG
RESULTS
OUTPUT DATA
Line #
1
2 /** Set CSV file name **/
3 filename csv "/folders/myfolders/Movies.csv"
4 termstr = crlf;
5
6 /** Import the CSV file. **/
7 proc import
8 datafile = CSV
9 out = movies
10 dbms = csv
11 replace;
12 run;
13
14 /** Create a scatterplot **/
15 title "Critic Score and Box Office Revenue";
16
17 proc sgscatter
18 data = movies;
19 plot BoxOffice*CriticScore;
20 run;
21
22 /** Print the results. **/
23 proc print data = movies; run;
24
25 /** Unassign the file reference. **/
26 filename csv;
Line 1, Column 1
User: sasdemo

Sentiment Analysis.py - C:\Pluralsight\Sentiment Analysis.py

File Edit Format Run Options Windows

```
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
```

Where to Go Next…

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

Exploratory Data Analysis with R

Matthew Renze

@matthewrenze | www.matthewrenze.com

PLURALSIGHT

www.pluralsight.com/courses/r-data-analysis

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

- Twitter: @matthewrenze
- Email: matthew@renzeconsulting.com
- Website: www.matthewrenze.com