

# Practical Data Science with R

Matthew Renze

CFT Training for AIG

# 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

Editorial: The war over...  
Misgoverning Argentina  
The economic shift from West to East  
Geographically modified crops: biowar  
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...



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



Source: Wikipedia

Sink

or

Swim

# Overview

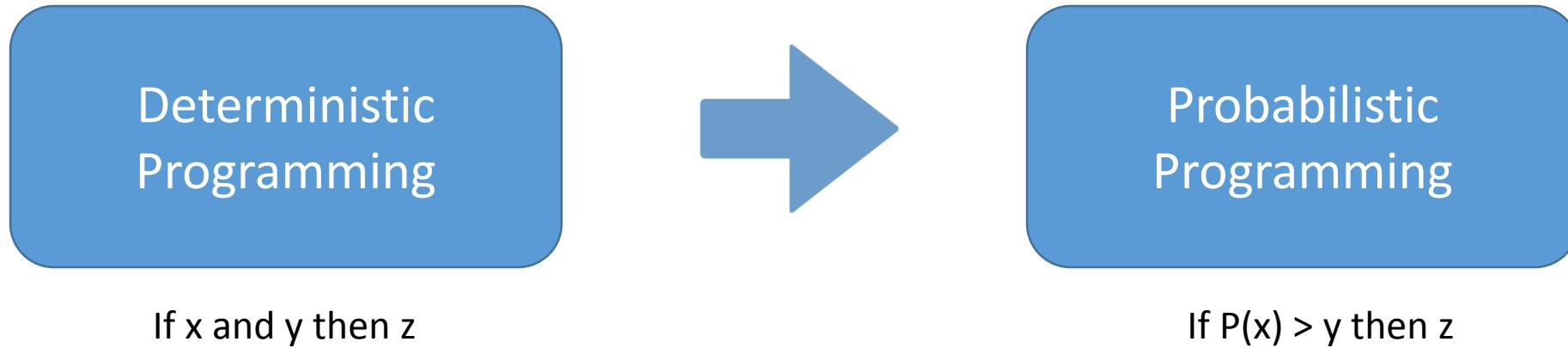
1. Introduction
2. Transforming Data
3. Descriptive Statistics
4. Data Visualization
5. Statistical Modeling
6. Handling Big Data
7. Machine Learning
8. R in Practice



# 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

# How Does This Apply to Me?



# About Me

Independent software consultant

Education

B.S. in Computer Science

B.A. in Philosophy

Community

Pluralsight Author

Public Speaker

Microsoft MVP

ASPIInsider

Open-Source Software

IOWA STATE  
UNIVERSITY



# Agenda

- 1. Introduction
  - 2. Transforming Data
  - 3. Descriptive Statistics
  - 4. Data Visualization
  - 5. Statistical Modeling
  - 6. Handling Big Data
  - 7. Machine Learning
  - 8. R in Practice
- 
- 1. Transforming Data Project
  - 2. Descriptive Statistics Project
  - 3. Data Visualization Project
  - 4. Statistical Modeling Project
  - 5. Machine Learning Project

# Schedule

Lectures (15 min)

Demos (10 min)

Labs (20 min)

Breaks (10 min)

# Logistics

Pairing for labs is optional

Ask questions if needed

Come and go as needed

Feedback forms at the end

# Lab Options

Type all code

Copy and paste

Run demo scripts

# Workshop URL

<http://www.matthewrenze.com/workshops/practical-data-science-with-r-aig/>

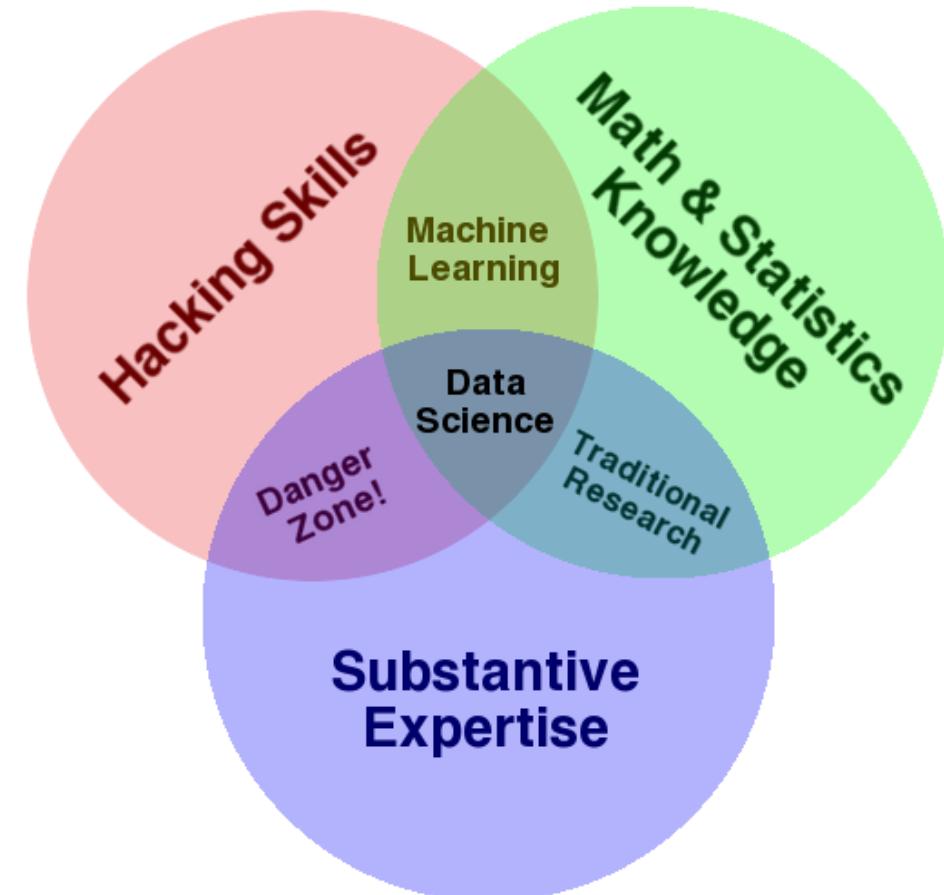
# Introduction to Data Science

# What is Data Science

Interdisciplinary field

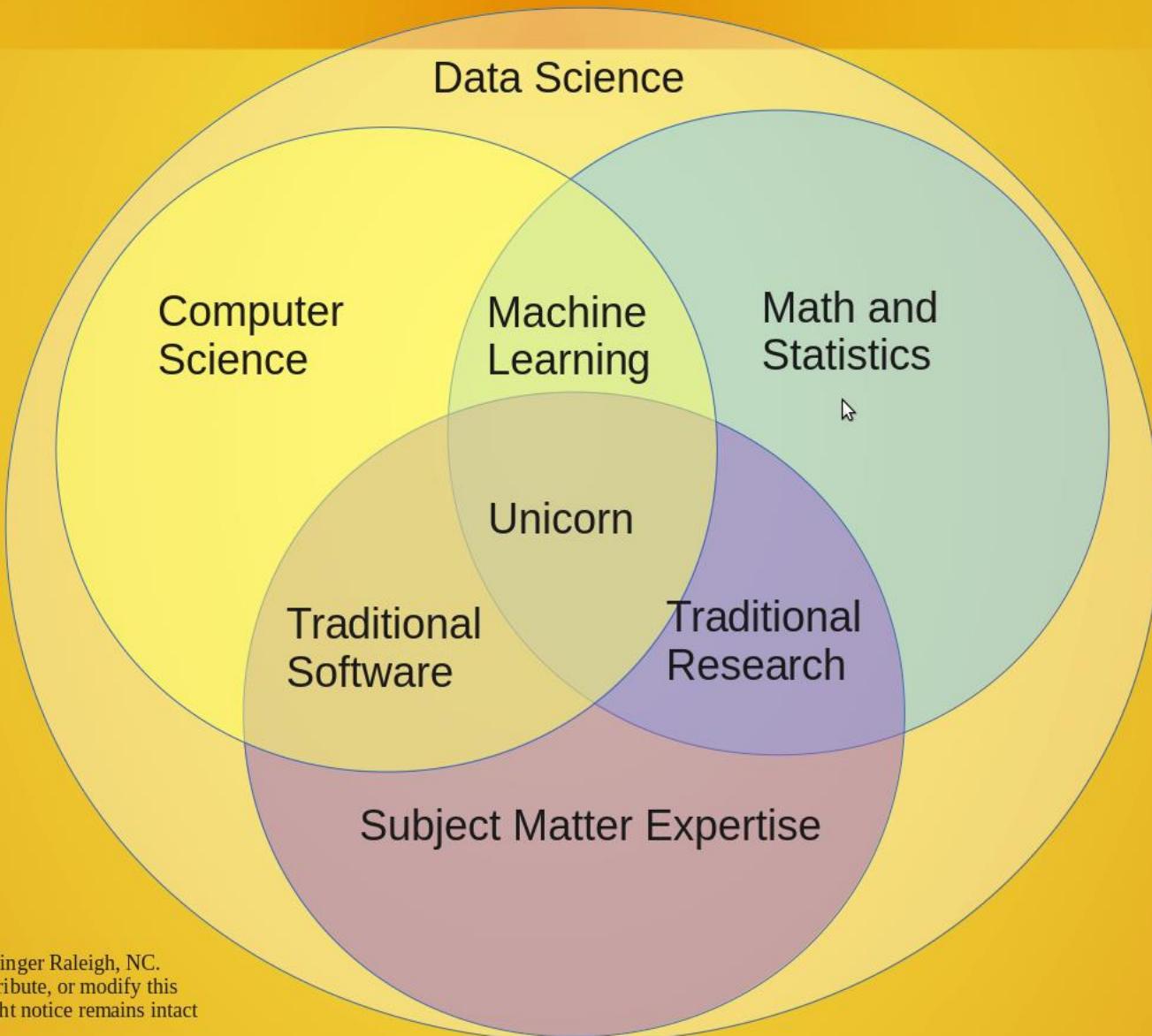
Derive insight from data

Transform data into knowledge



Source: [www.drewconway.com](http://www.drewconway.com)

# Data Science Venn Diagram v2.0



# What is a Data Scientist?

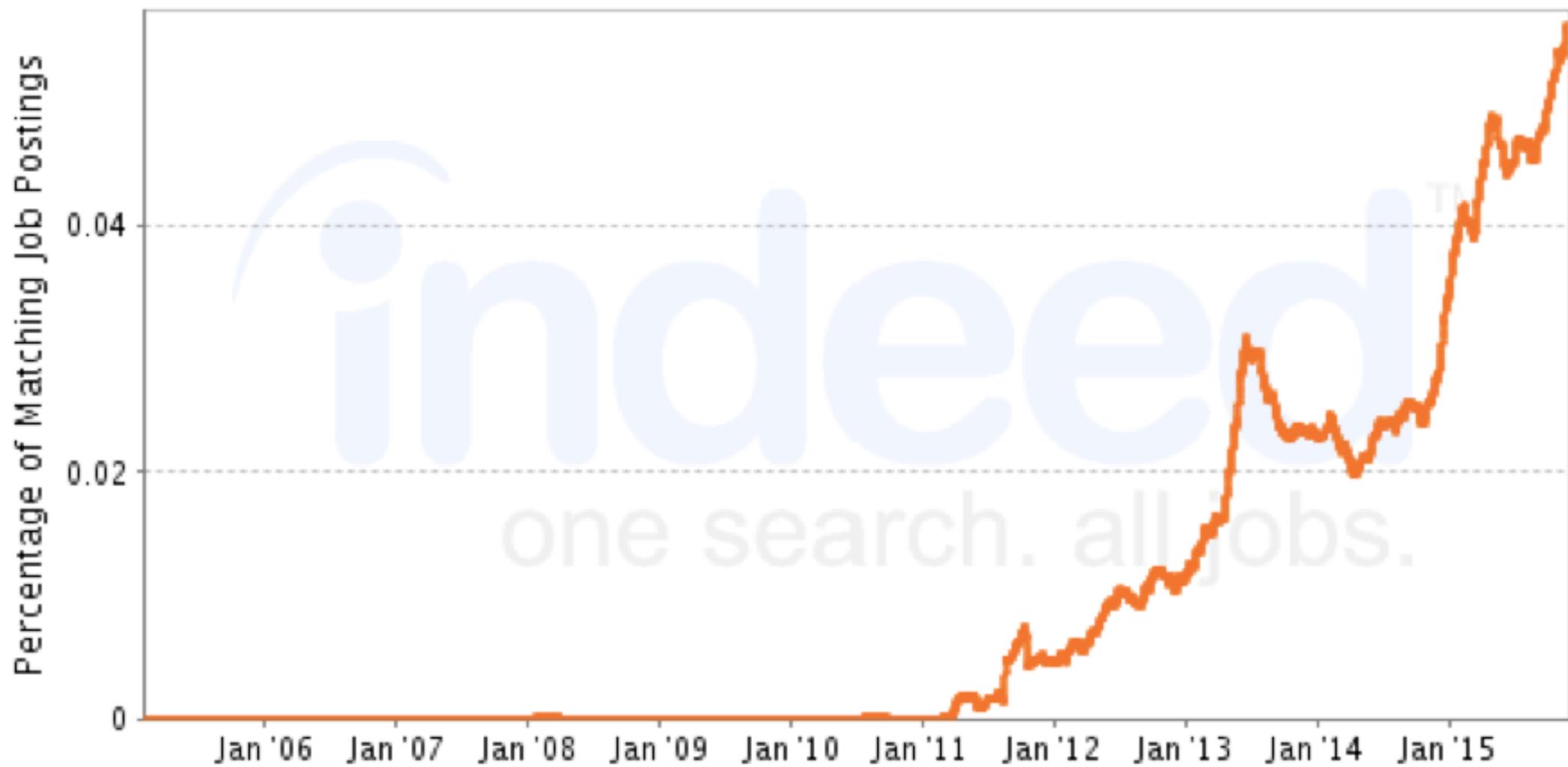
- Performs data science
- Proper accreditations
- More than a scientist
- More than an analyst
- More than a developer



Source: <http://www.clipartpanda.com>

## Job Trends from Indeed.com

— "Data Scientist"



# The Data Science Toolkit

Programming

Data transformation

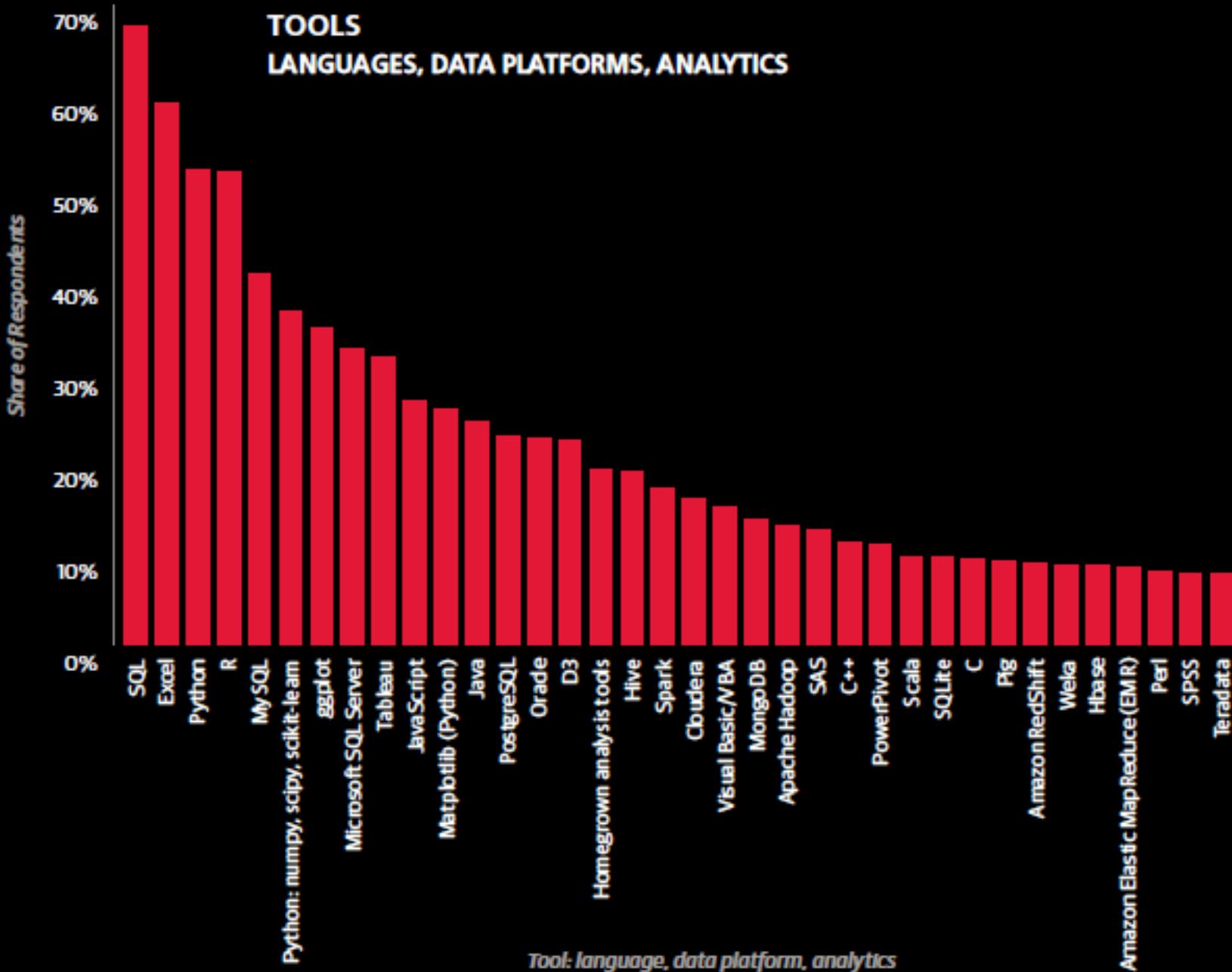
Descriptive statistics

Data visualization

Statistical modeling

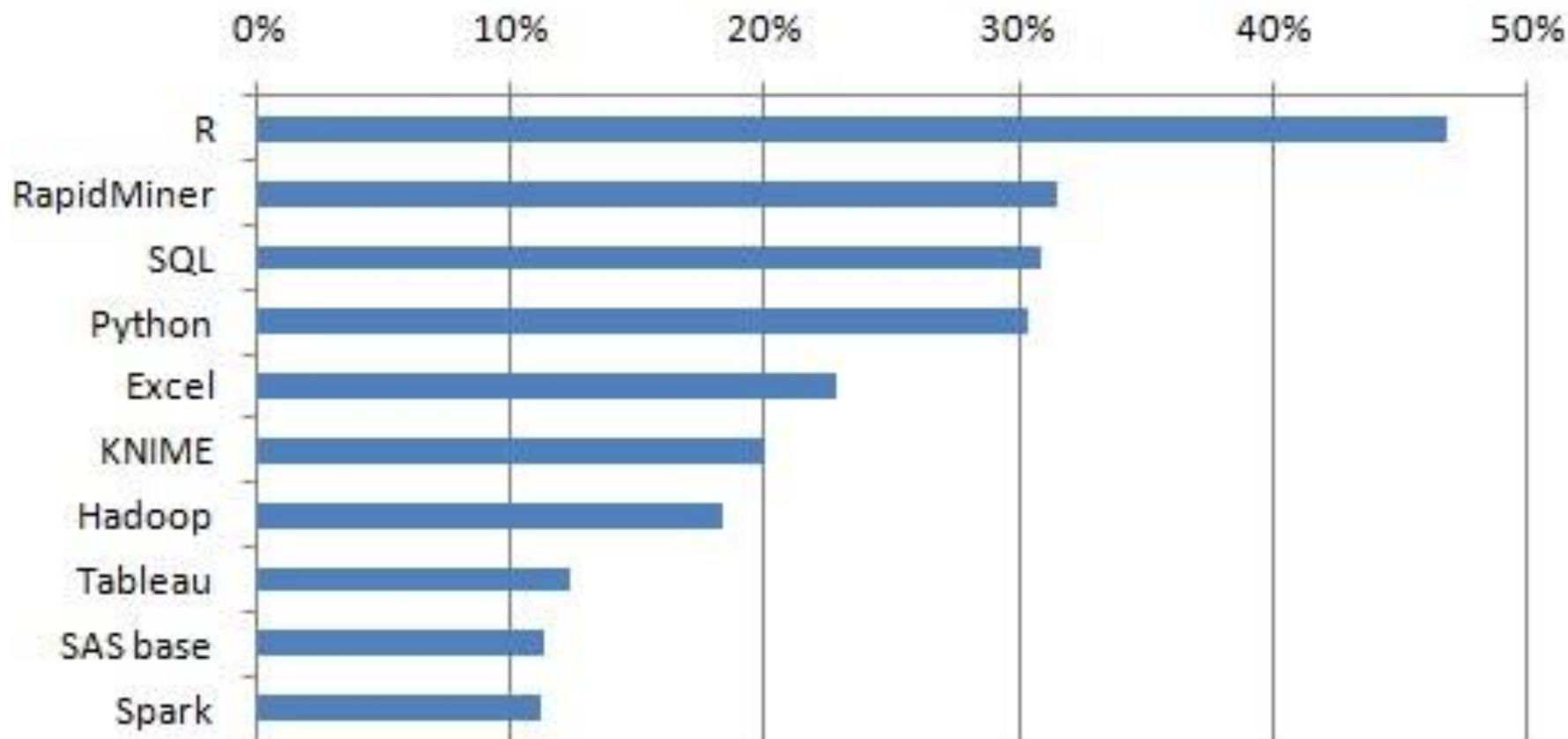
Big Data

Machine learning

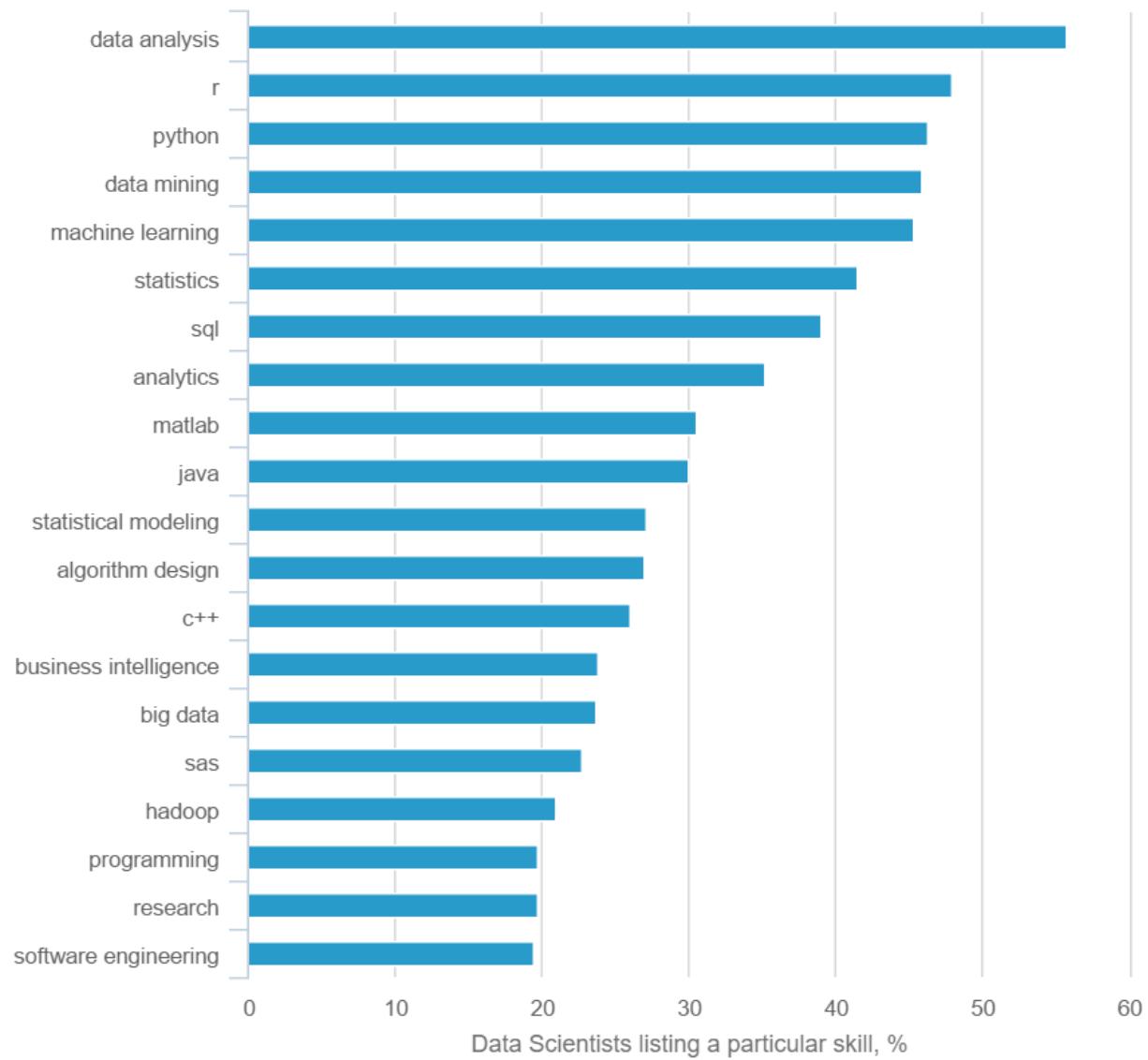


Source: O'Reilly 2015 Data Science Salary Survey

# Top Analytics, Data Mining, Data Science software used, 2015



## TOP 20 SKILLS OF A DATA SCIENTIST



# Why is Data Science Important?

Lots of data

Fast computers

Powerful analysis tools

Prediction vs. explanation

# The Data Science Process

1. Ask a question
2. Explore the data
3. Prepare the data
4. Create a model
5. Evaluate the model
6. Deploy the model / results



# The Data Science Process

Iterative process  
Non-sequential  
Early termination  
Established process (CRISP-DM)



# Introduction to R

# What is R?

Open source

Language and environment

Numerical and graphical analysis

Cross platform



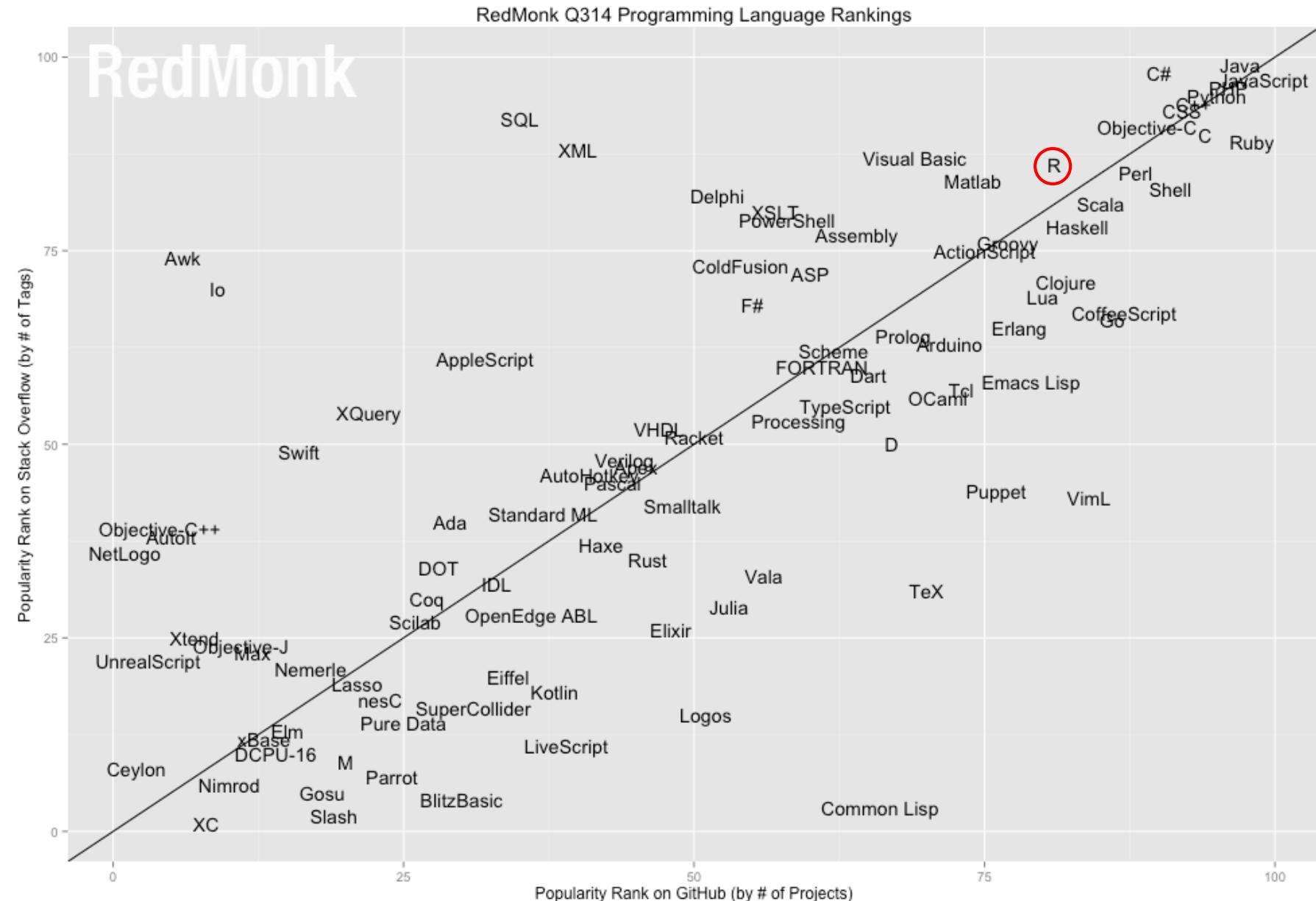
# What is R?

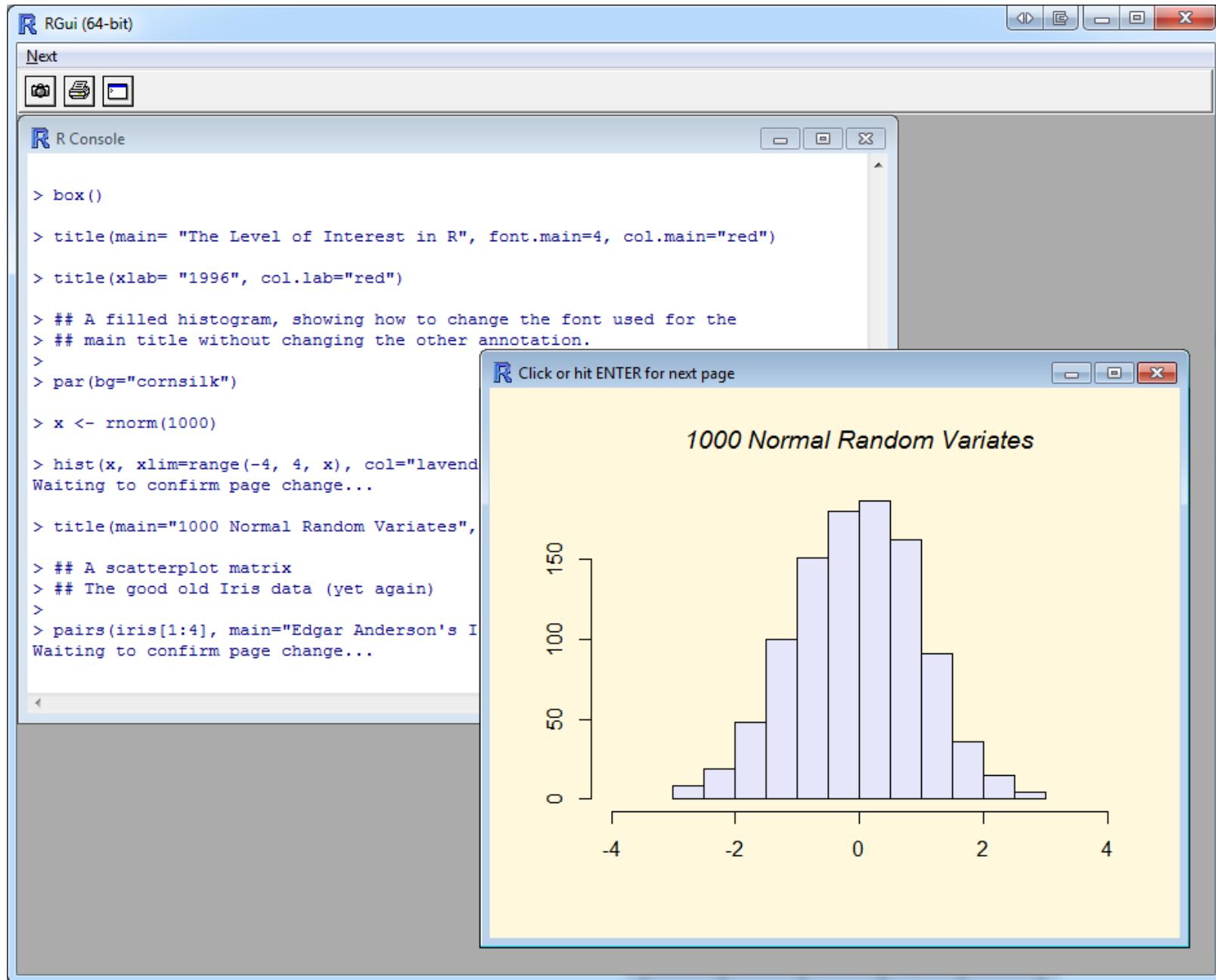
Active development  
Large user community  
Modular and extensible  
6700+ extensions

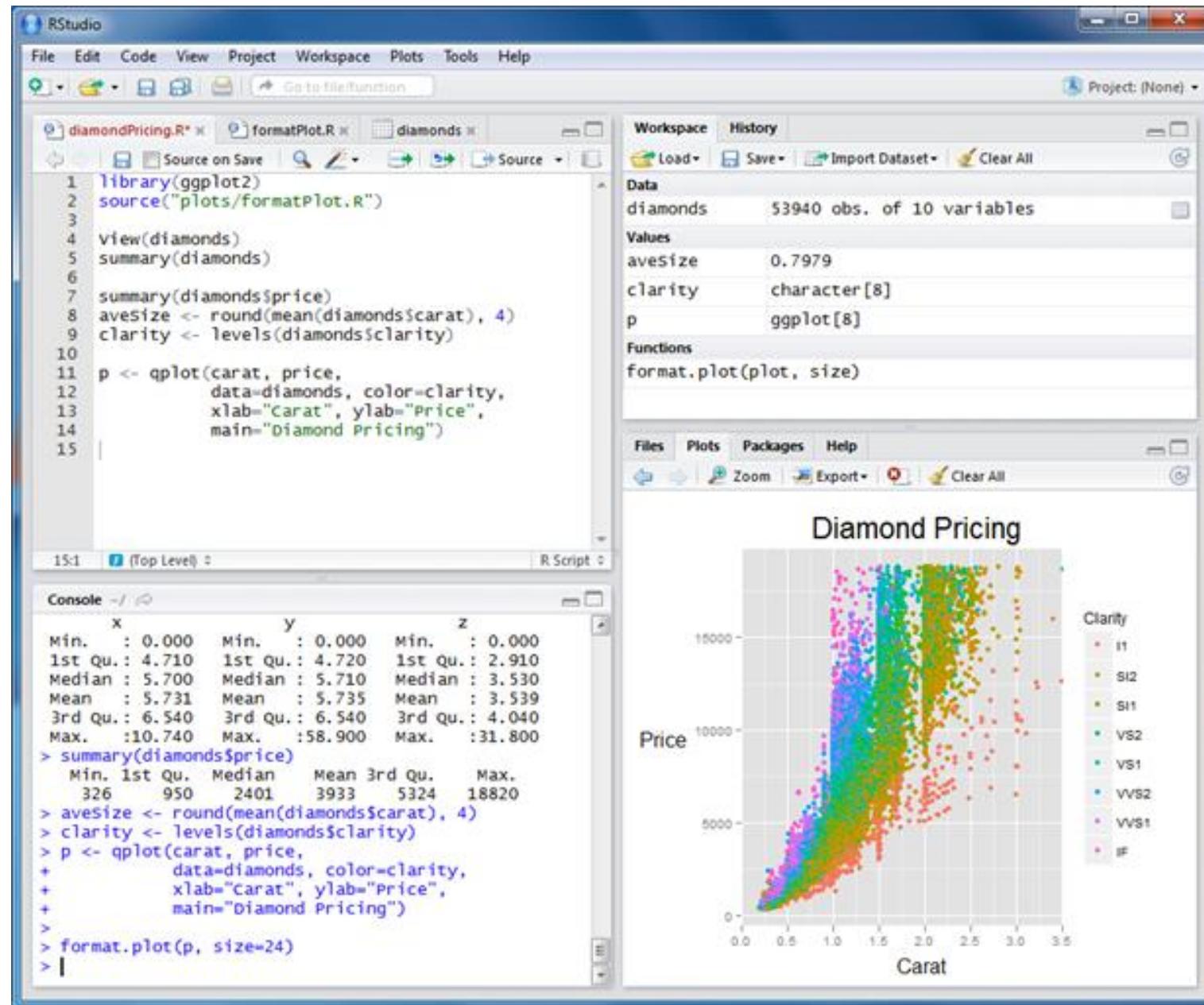


# FREE!









1-1 - Introduction (Base).R - Microsoft Visual Studio

File Edit View NCrunch Project Debug Team Data Lake Tools Architecture Test ReSharper R Tools Analyze Window Help Matthew Renze

1-1 - Introduction (Base).R Script1.R\*

```
# Create a bar chart
barplot(
  names = df$name,
  height = df$value,
  col = "skyblue",
  main = "Hello World",
  xlab = "Name",
  ylab = "Value")

# View the help files
?plot

?barplot
```

R Interactive

```
> # Plot using default parameter order
> plot(df$name, df$value)
> # Plot using named parameters
> plot(
+   x = df$name,
+   y = df$value)
> # Create a bar chart
> barplot(
+   names = df$name,
+   height = df$value,
+   col = "skyblue",
+   main = "Hello World",
+   xlab = "Name",
+   ylab = "Value")
>
```

Variable Explorer

Global Environment

Name	Value	Class	Type
df	3 obs. of 2 variables	data.frame	list
@.Data	List of 2	list	list
@names	chr [1:2] "Name" "Value"	character	character
@row.names	int [1:3] 1 2 3	integer	integer
@.S3Class	"data.frame"	character	character
Name	Factor w/ 3 levels "a","b","c": 1 2 3	factor	integer
Value	num [1:3] 1 2 3	numeric	double

Variable Explorer R History

R Plot

Hello World

Value

a b c

Name

Solution Explorer R Plot R Help

Ready Ln 29 Col 1 Ch 1 INS Cloud develop ✓

# Code Demo

# Lab 1

## R Programming Basics

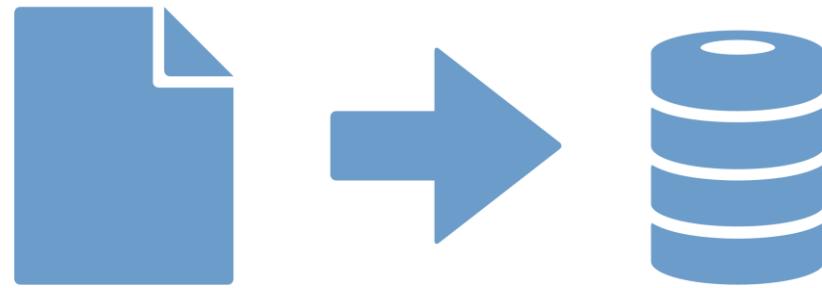
# Transforming Data

# 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

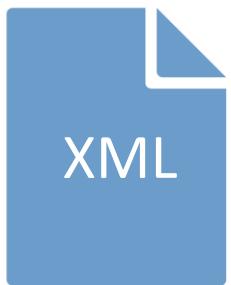
Transforming data

Handling missing data

Handling anomalous values

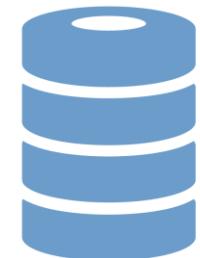
# Loading Data in R

File-based data

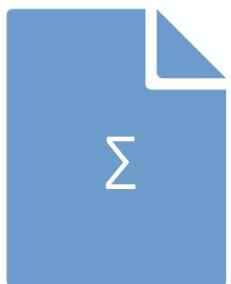


Web-based data

Databases



Statistical data



And many more...

# Cleaning Data

This step is often the  
Most difficult  
Most time consuming

TIP: Record all steps



# Data Munging Tools in R

`plyr`

`dplyr`

`reshape2`

`tidyR`

`stringr`

`lubridate`



# dplyr

## Single Data Frame Verbs

Select

Filter

Mutate

Summarize

Arrange

## Multiple Data Frame Verbs

Inner join

Left join

Semi join

Anti join

Union

Intersect

Set difference

# Open Movies Database

Movies								
Title	Year	Rating	Runtime (minutes)	Genre	Critic Score	Box Office	Awards	Inter-national
The Whole Nine Yards	2000	R	98	Comedy	45%	\$57.3M	No	No
Cirque du Soleil	2000	G	39	Family	45%	\$13.4M	Yes	No
Gladiator	2000	R	155	Action	76%	\$187.3M	Yes	Yes
Dinosaur	2000	PG	82	Family	65%	\$135.6M	Yes	No
Big Momma's House	2000	PG-13	99	Comedy	30%	\$0.5M	Yes	Yes

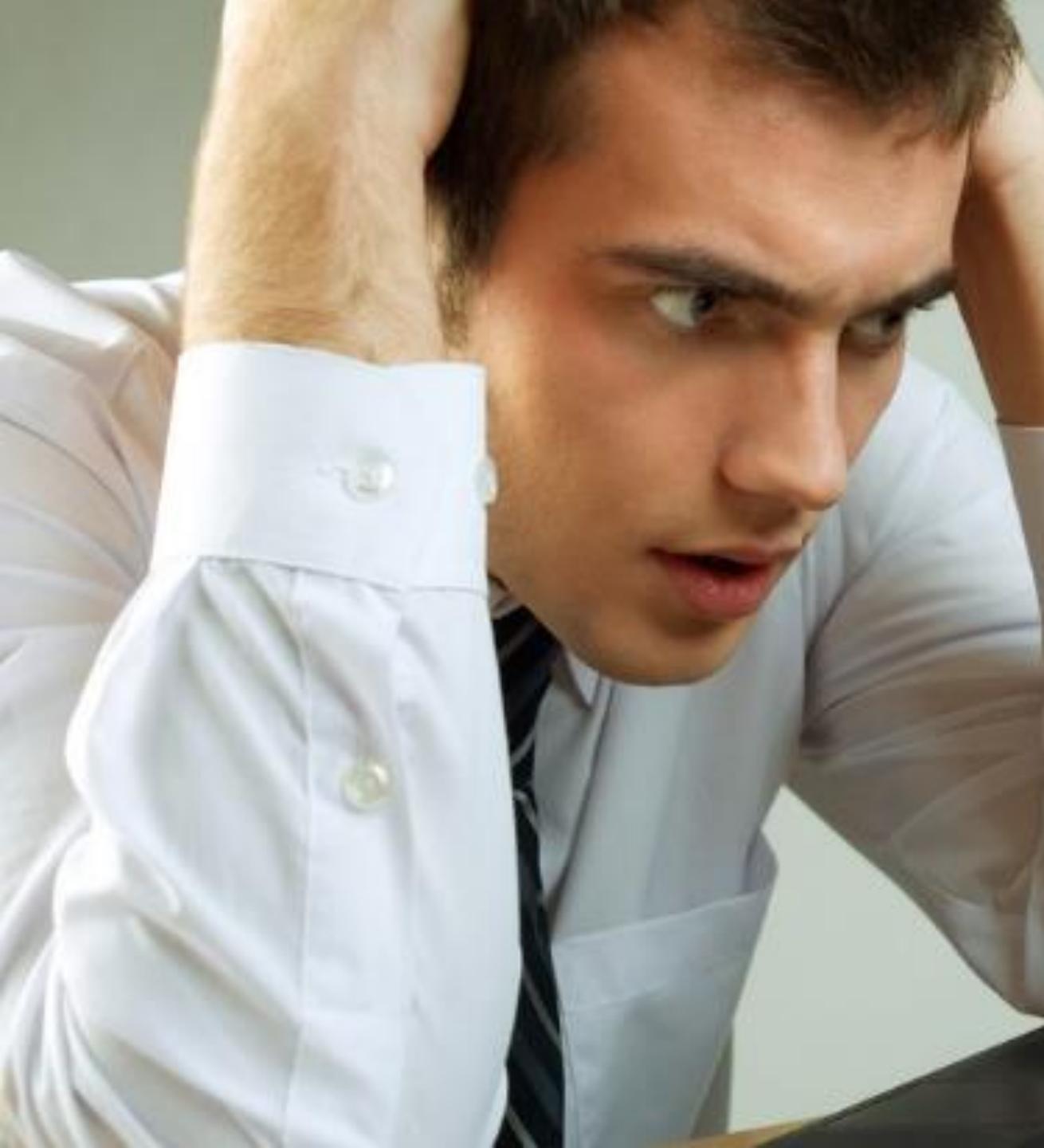


PROD. NO.  
SCENE

TAKE

ROLL





1. Column with wrong name
2. Rows with missing values
3. Runtime column has units
4. Revenue in multiple scales
5. Wrong file format

# Code Demo

# Lab 2

## Transforming Data



# Descriptive Statistics

# Descriptive Statistics

Describe data

Provides a summary

aka: Summary statistics

Movie Runtime	
Statistic	Value (minutes)
Minimum	38
1 <sup>st</sup> Quartile	93
Median	101
Mean	104
3 <sup>rd</sup> 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

1 - Univariate

2 - Bivariate

n - Multivariate

## Type of variables

Categorical - Qualitative

Numeric - Quantitative

Number of Variables

One  
Categorical  
Variable

One  
Numeric  
Variable

Two  
Categorical  
Variables

Categorical  
& Numeric  
Variable

Two  
Numeric  
Variables

Many  
Variables

Type of Variable(s)

# Analyzing One Categorical Variable

Qualitative univariate analysis

Frequency of observations

Movies by Genre		
Genre	Frequency	Percentage
Action	612	9%
Adventure	496	7%
Animation	168	2%
Comedy	1281	18%
Drama	1570	22%
Horror	269	4%
...	...	...

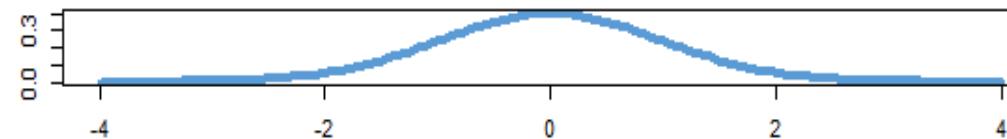
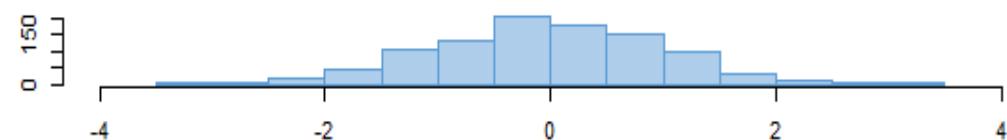
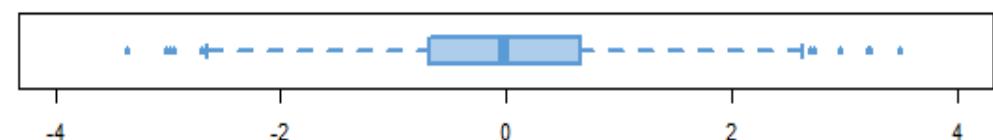
# Analyzing One Numeric Variable

Quantitative univariate analysis

Central tendency

Dispersion

Shape



# Visualizing Two Categorical Variables

Qualitative bivariate analysis

Joint frequency

Marginal frequency

Relative frequency

Contingency table

Movies by Genre and Rating					
Genre	G	PG	PG-13	R	Total
Action	2	70	311	229	612
Adventure	44	179	209	64	496
Animation	43	111	8	6	168
Comedy	45	258	472	506	1218
Drama	12	136	586	836	1570
Family	38	181	10	1	230
...	...	...	...	...	...
Total	230	1207	2686	3058	7181

# Visualizing Two Categorical Variables

Qualitative bivariate analysis

Joint frequency

Marginal frequency

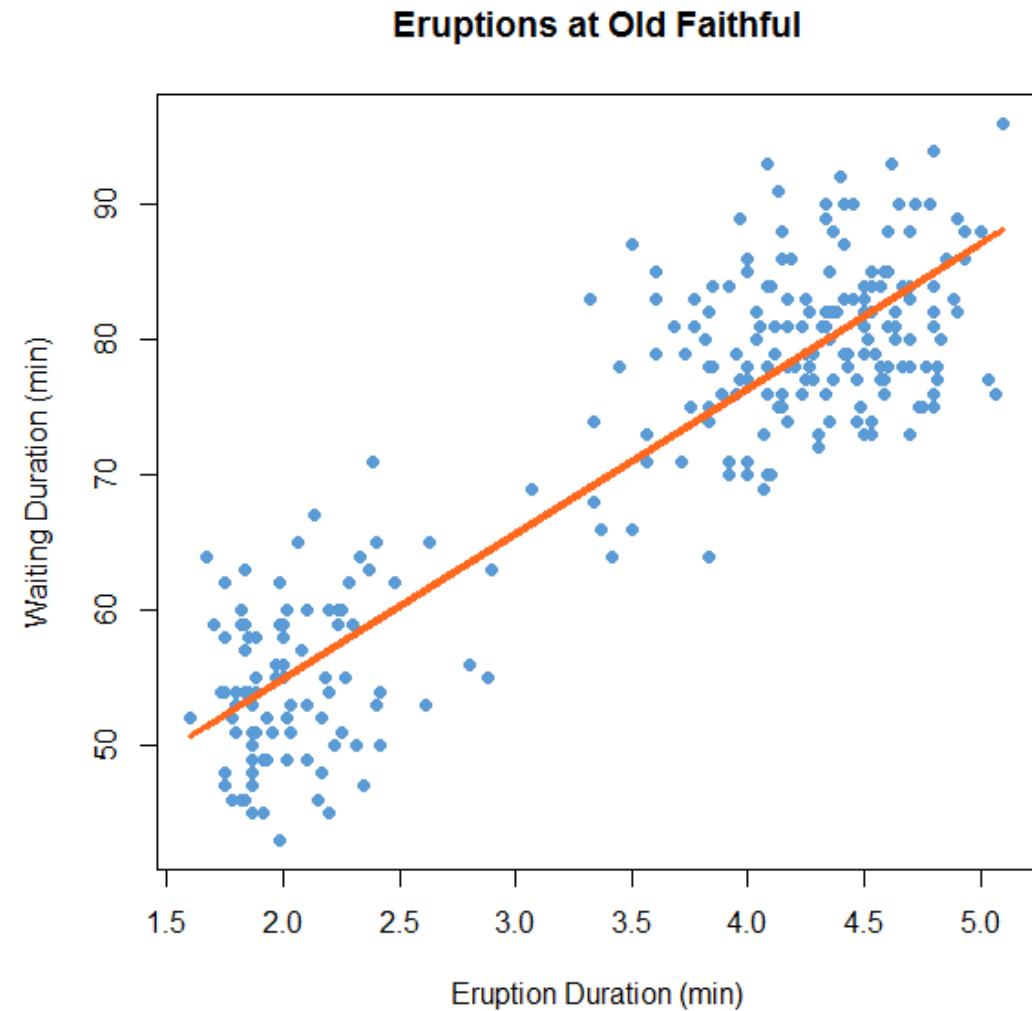
Relative frequency

Contingency table

Movies by Genre and Rating					
Genre	G	PG	PG-13	R	Total
Action	0.001	0.010	0.043	0.032	0.086
Adventure	0.006	0.025	0.029	0.009	0.069
Animation	0.006	0.015	0.001	0.001	0.023
Comedy	0.006	0.036	0.066	0.070	0.170
Drama	0.002	0.019	0.082	0.116	0.219
Family	0.005	0.025	0.001	0.001	0.033
...	...	...	...	...	...
Total	0.032	0.168	0.374	0.426	1.000

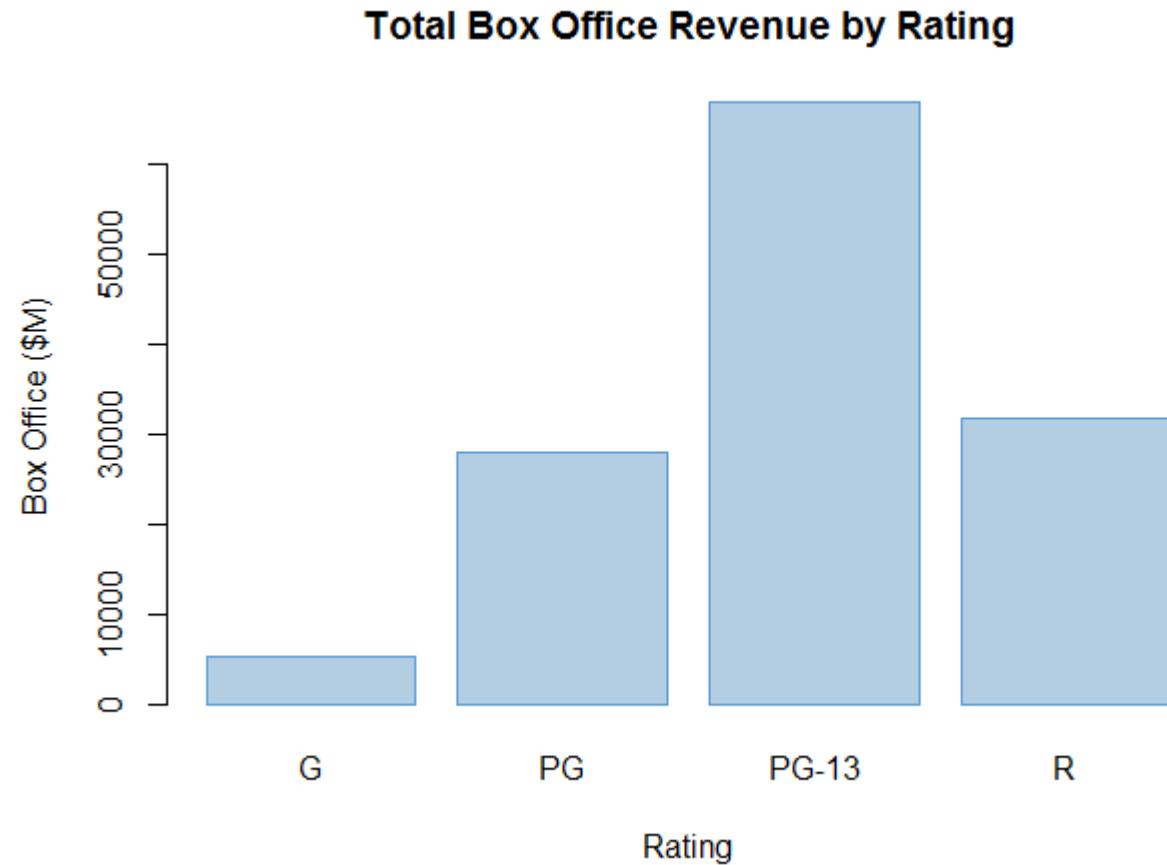
# Analyzing Two Numeric Variables

Quantitative bivariate analysis  
Predictor vs. outcome  
Covariance  
Correlation



# Analyzing a Numeric Variable Grouped By a Categorical Variable

One categorical variable  
One numeric variable  
Aggregate measures



# Analyzing Many Variables

Multivariate analysis  
Specific meaning in statistics

Number of Variables

One  
Categorical  
Variable

One  
Numeric  
Variable

Two  
Categorical  
Variables

Categorical  
& Numeric  
Variable

Two  
Numeric  
Variables

Many  
Variables

Type of Variable(s)





# COWBOYS & Space Invaders: The Musical



Extended Edition



# Code Demo

# Lab 3

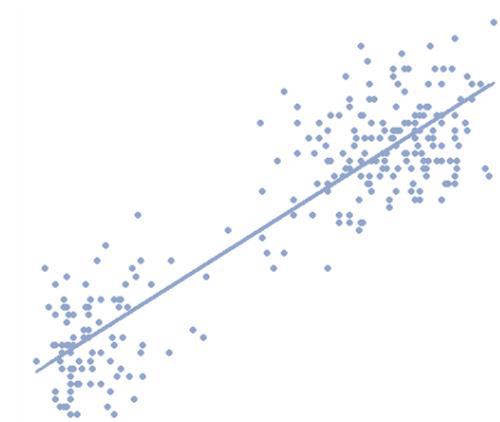
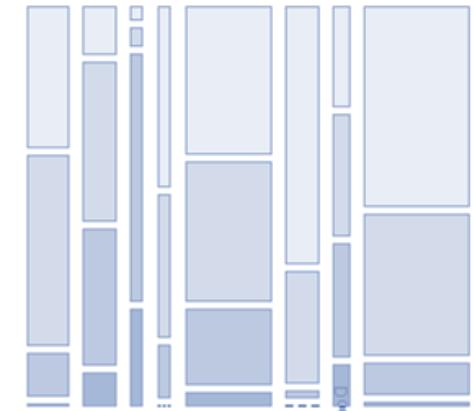
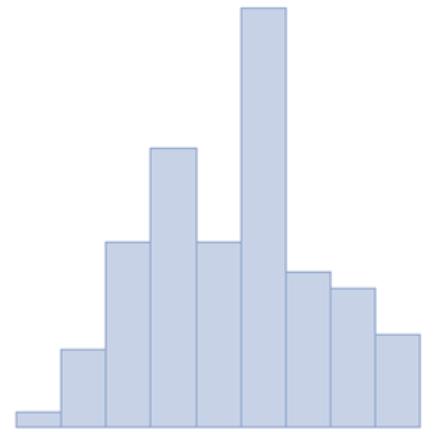
## Descriptive Statistics



# Data Visualization

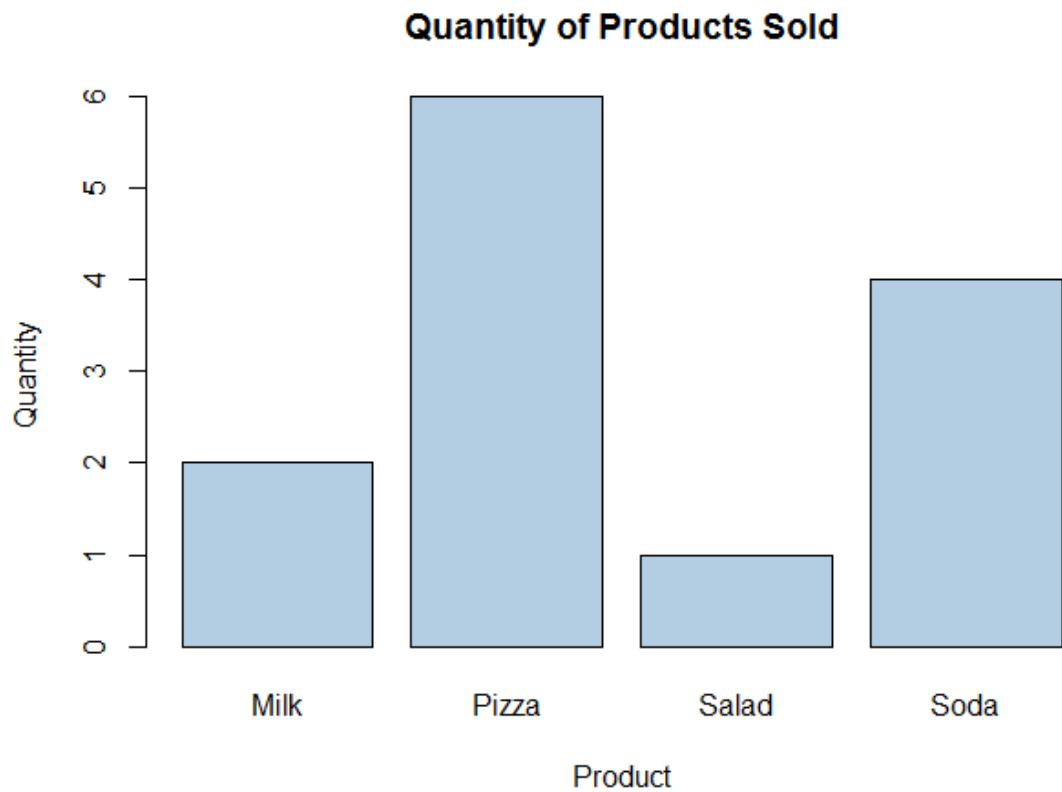
# Data Visualization

Visual data representation  
For human pattern recognition  
Map dimensions to visual



# 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  
Type of variable(s)

Number of Variables

One  
Categorical  
Variable

One  
Numeric  
Variable

Two  
Categorical  
Variables

Categorical  
& Numeric  
Variable

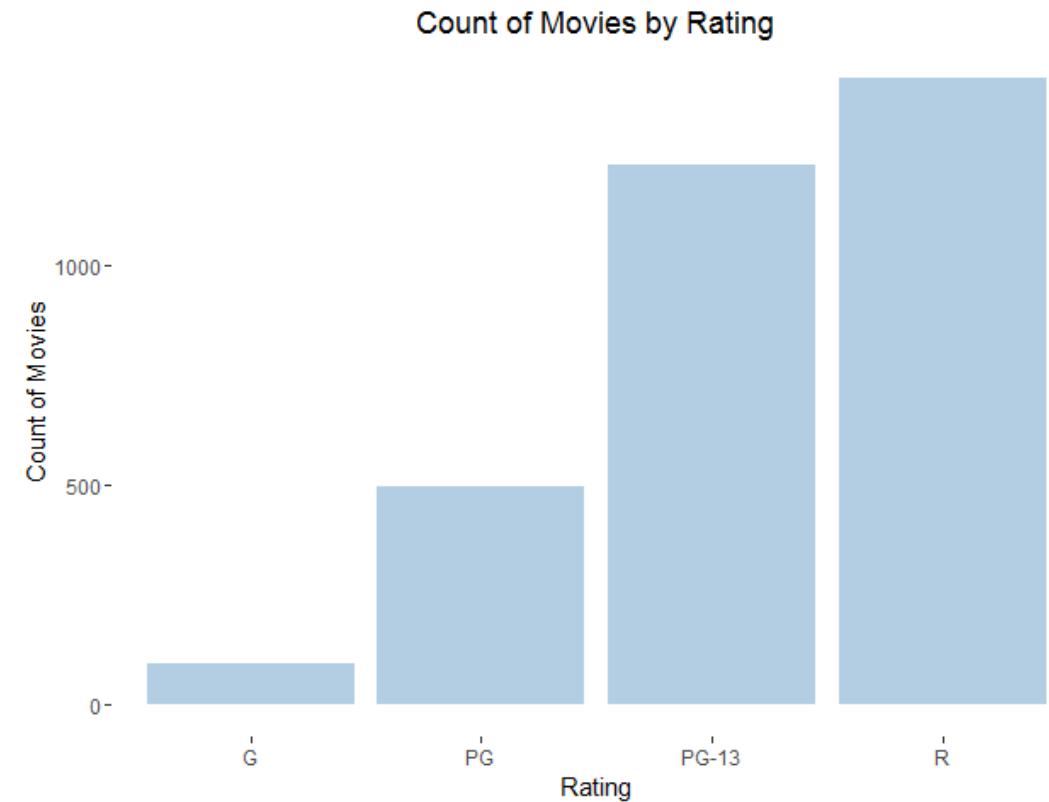
Two  
Numeric  
Variables

Many  
Variables

Type of Variable(s)

# Visualizing One Categorical Variable

Frequency  
Proportion

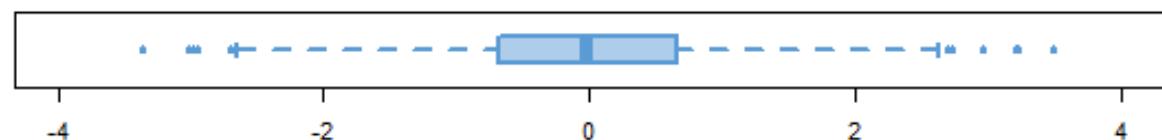


# Visualizing One Numeric Variable

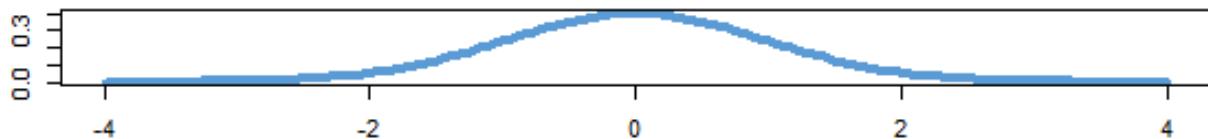
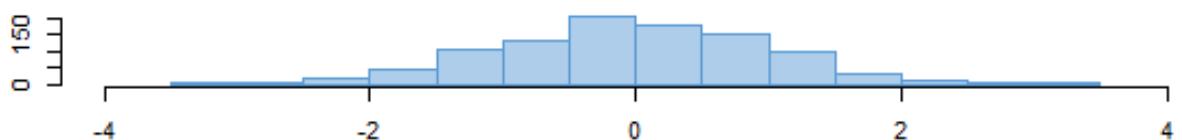
Location



Spread



Shape

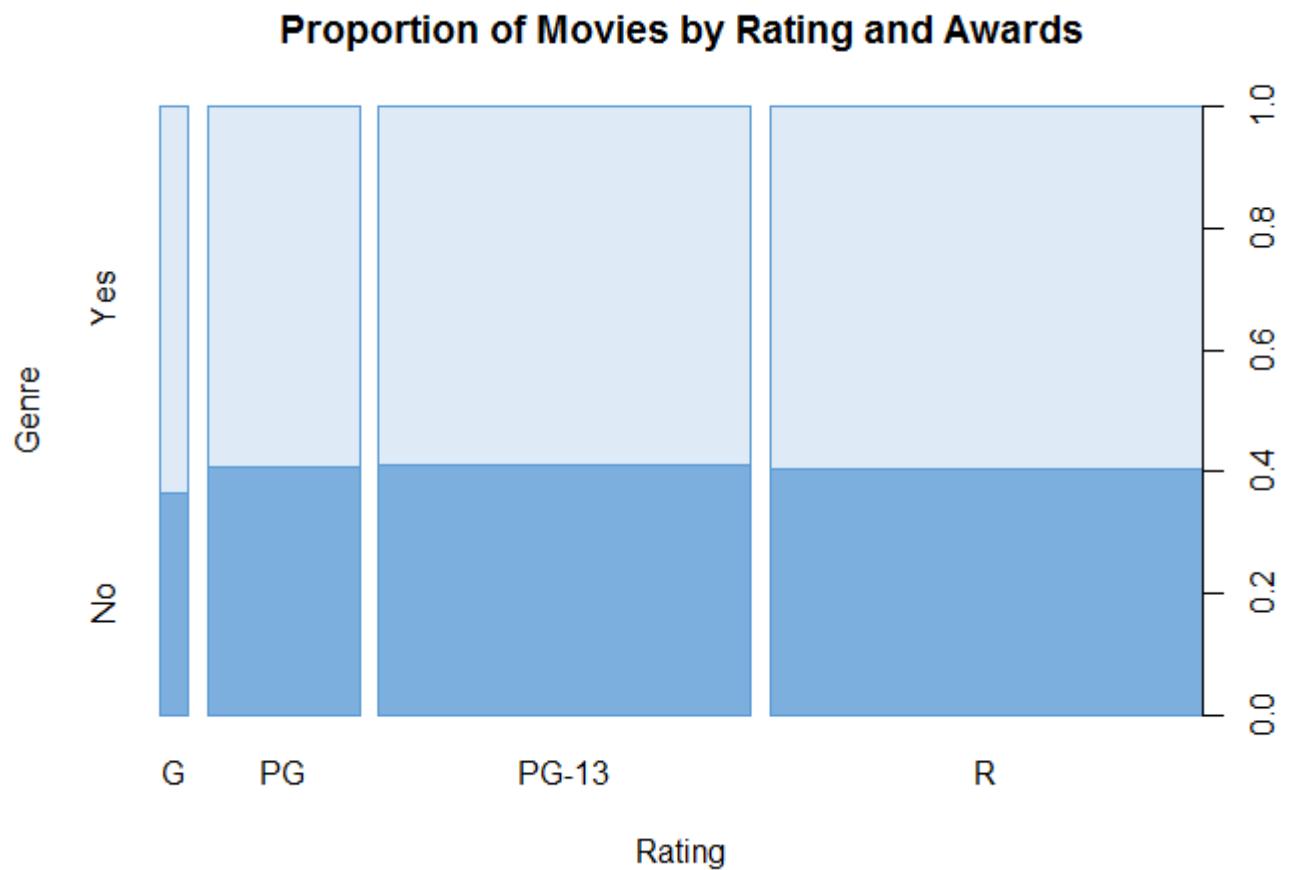


# Visualizing Two Categorical Variables

Joint frequency

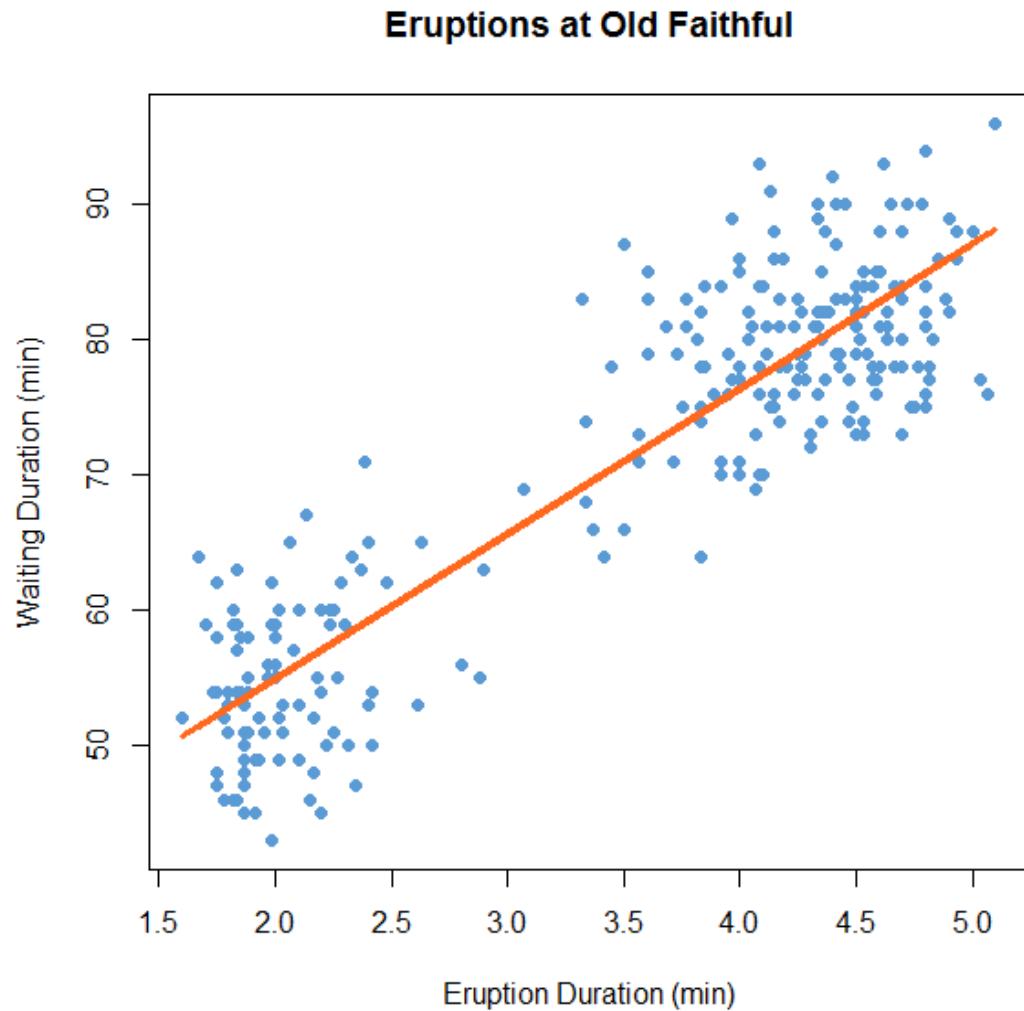
Marginal frequency

Relative frequency



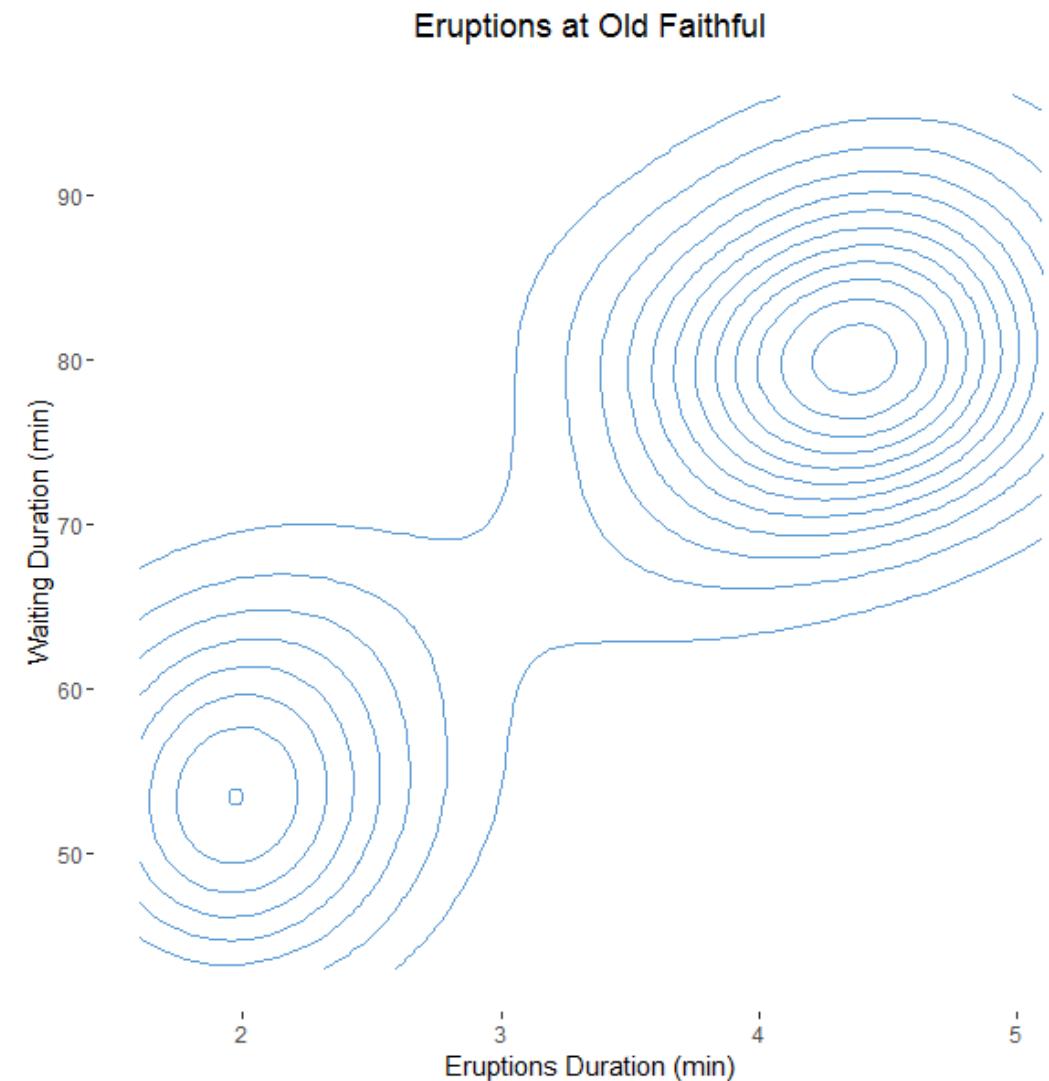
# Visualizing Two Numeric Variables

Relationship  
Correlation  
Distribution



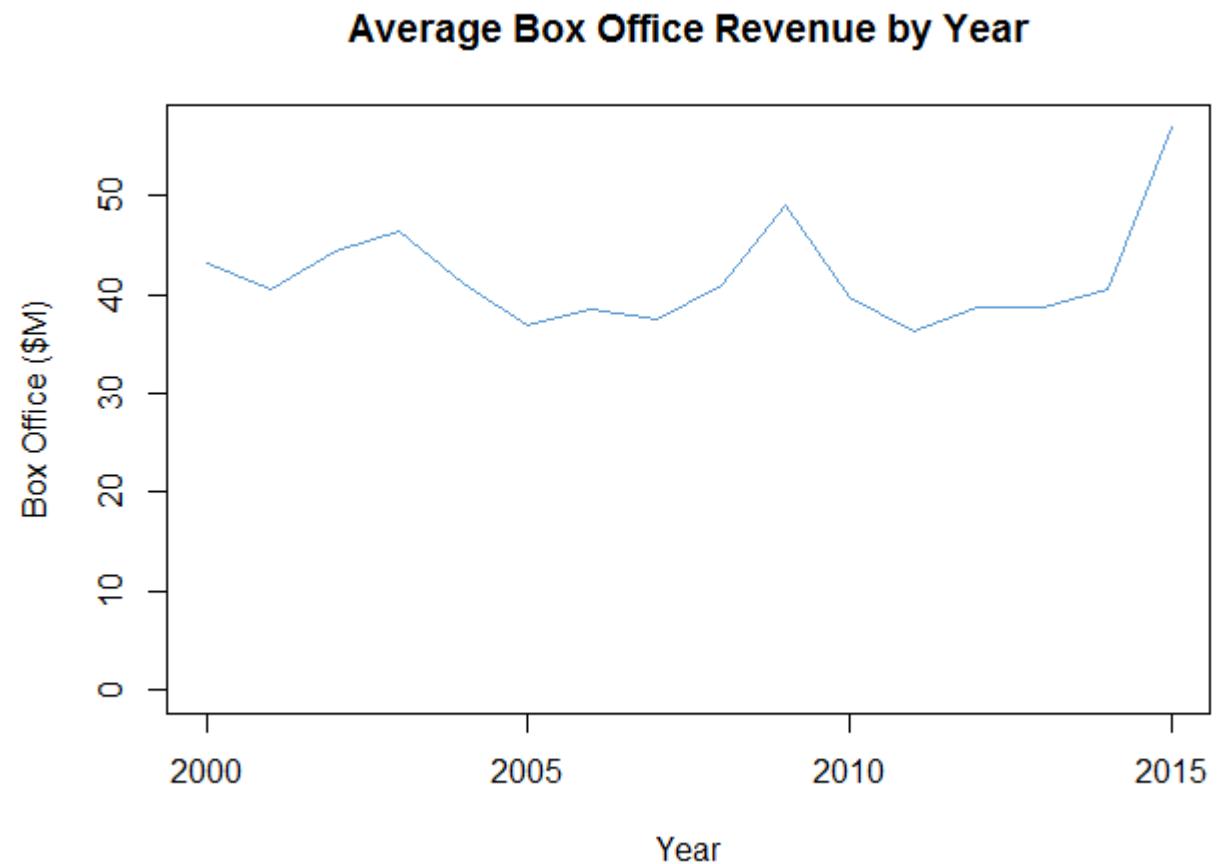
# Visualizing Two Numeric Variables

Relationship  
Correlation  
Distribution



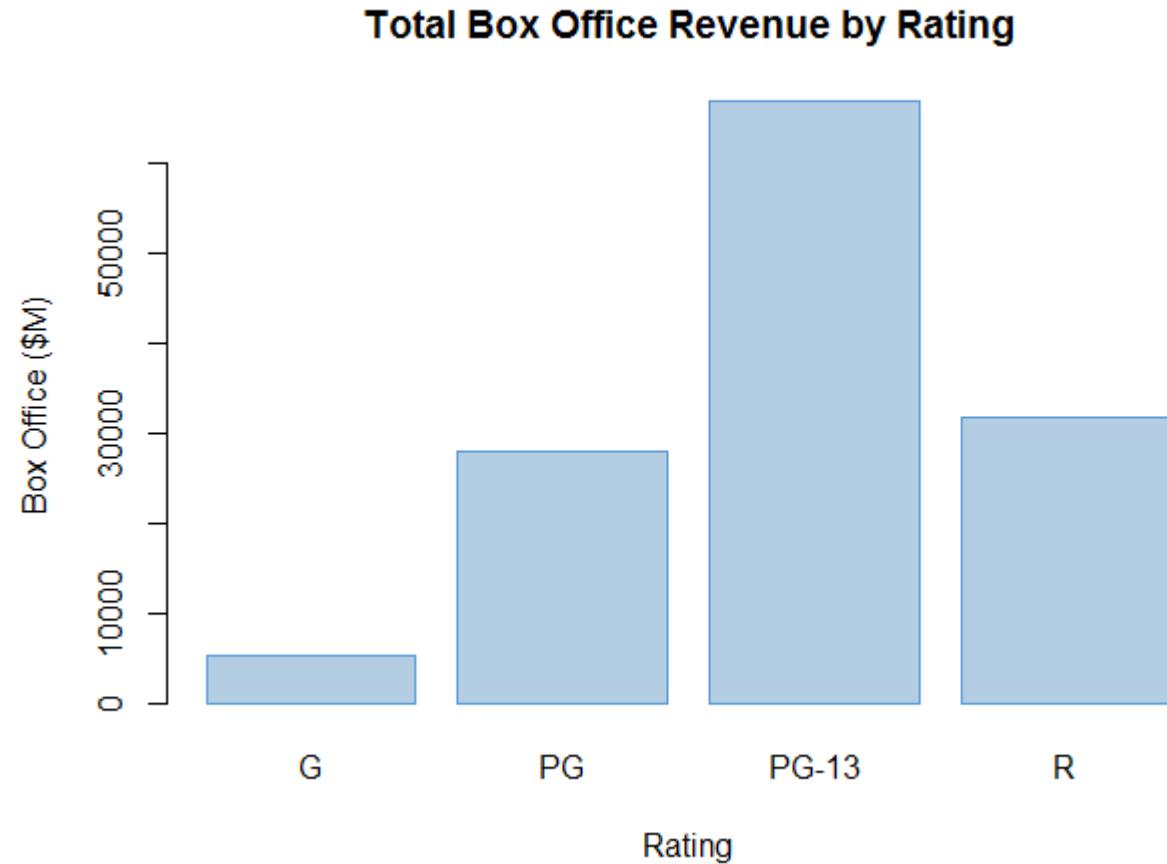
# Visualizing Time Series Data

Values over time  
Rate of change

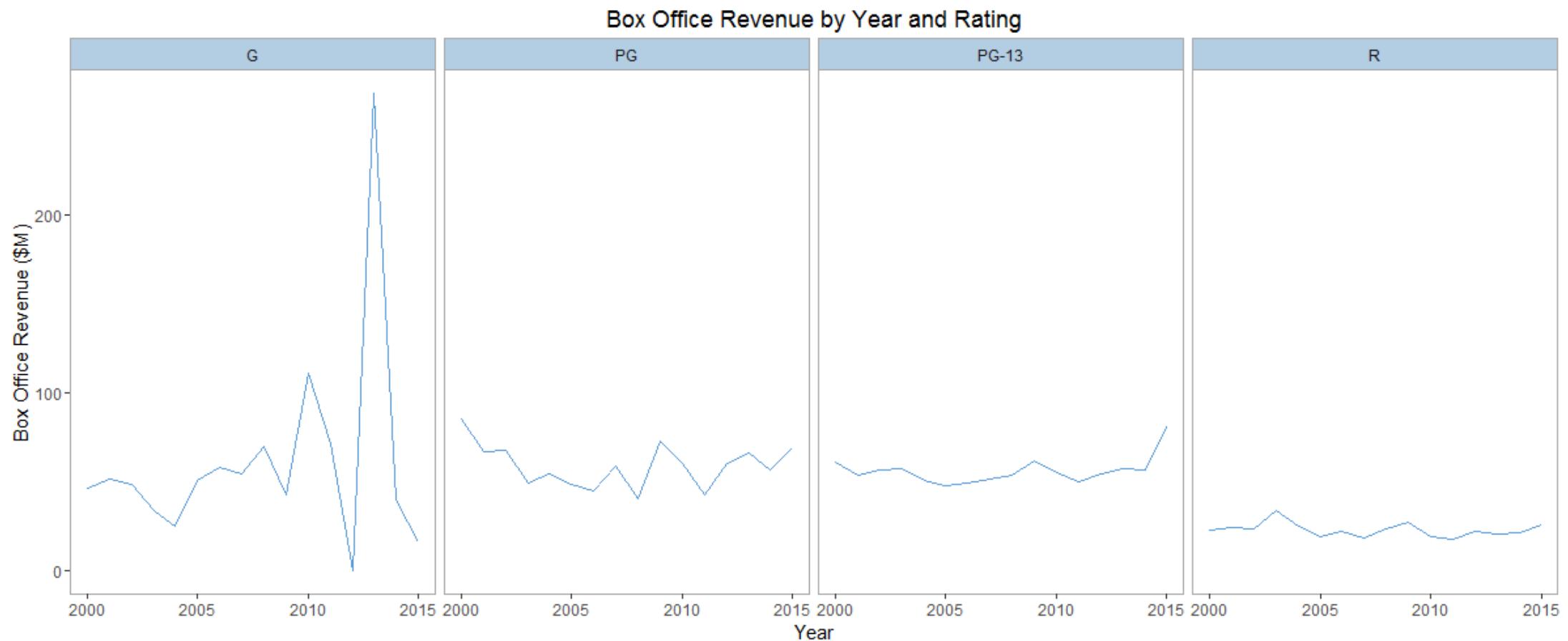


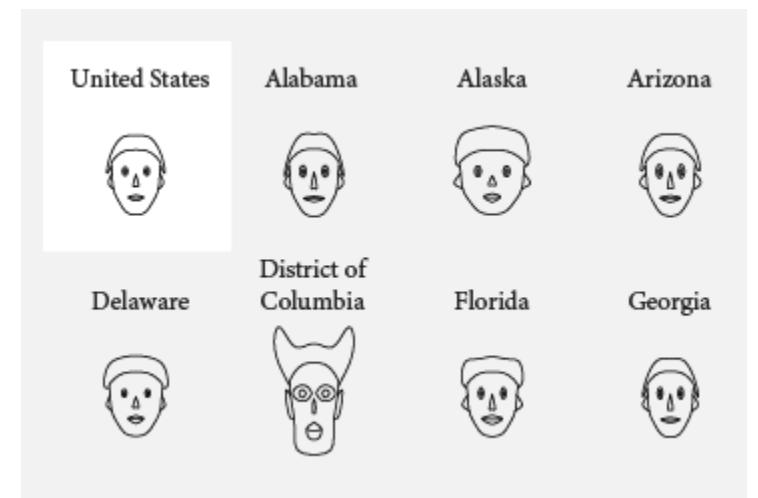
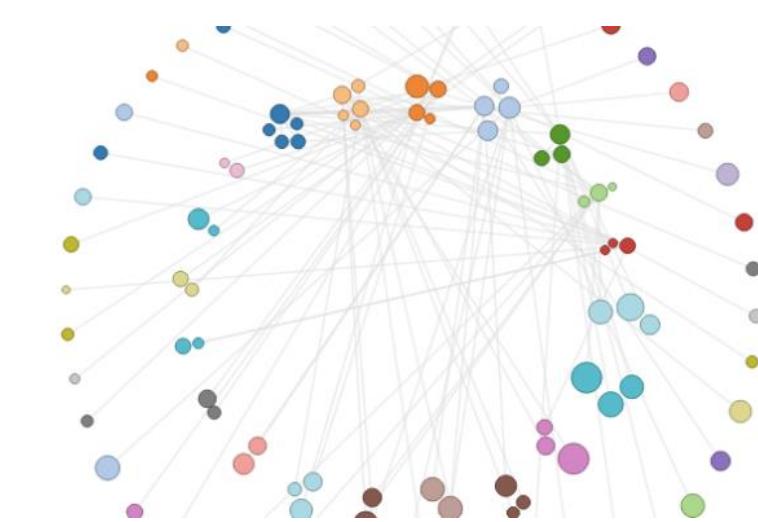
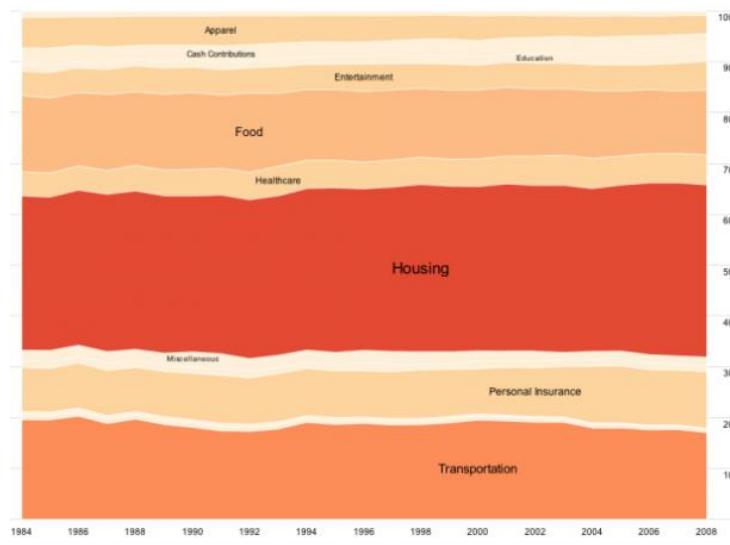
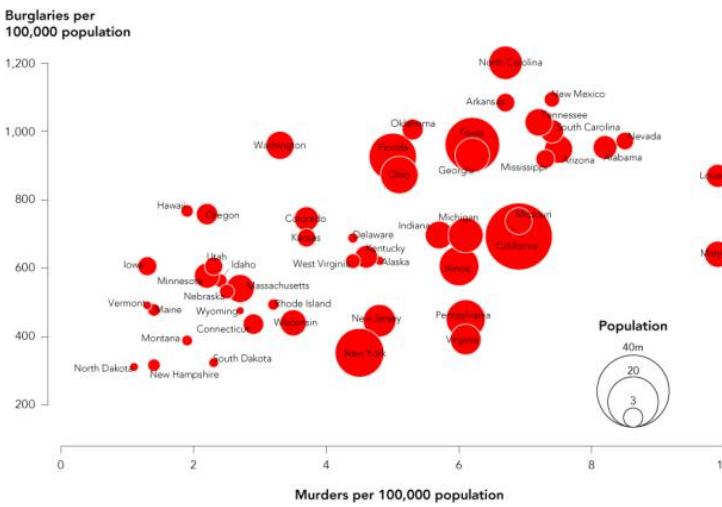
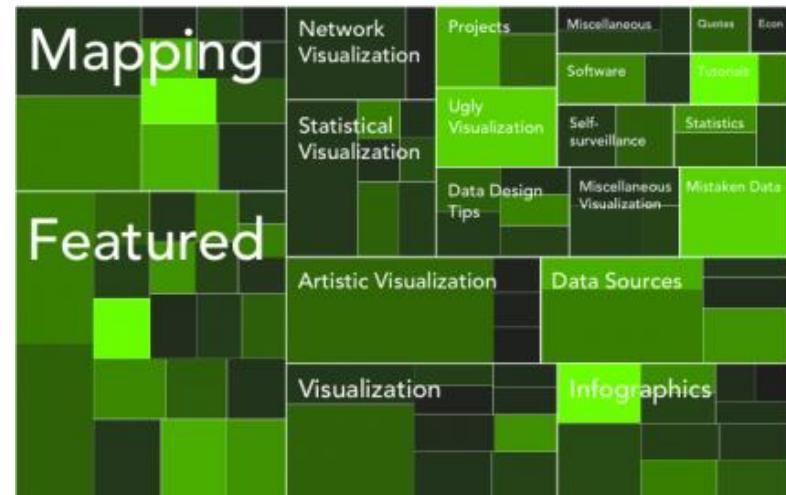
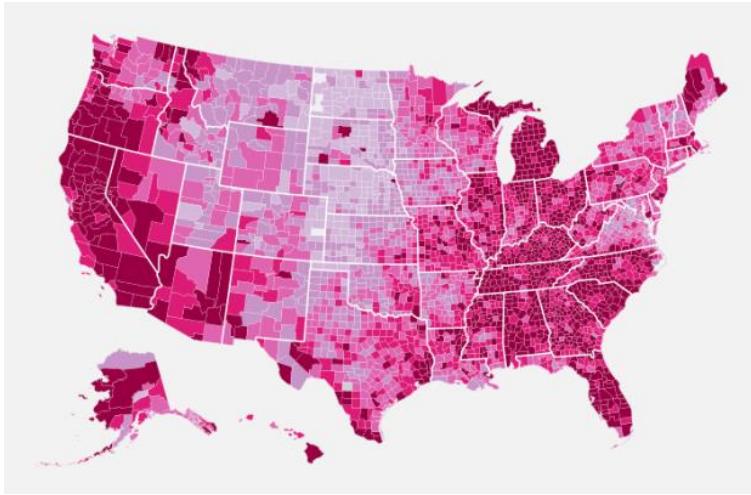
# Visualizing a Numeric Variable Grouped by a Categorical Variable

Aggregate  
Grouped  
Comparison



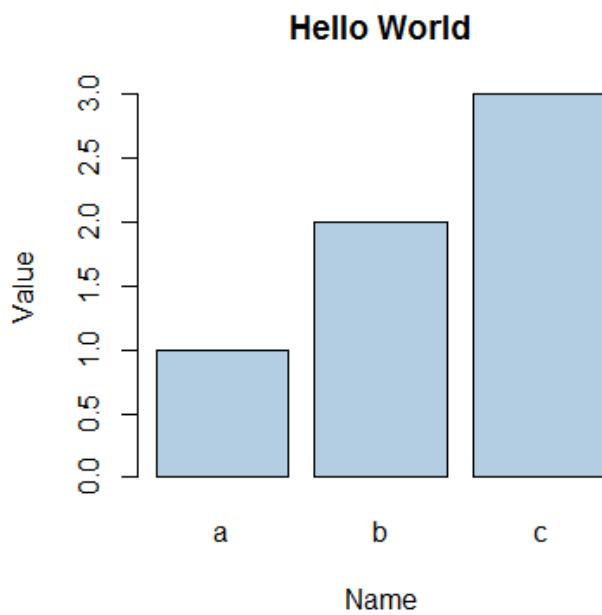
# Visualizing Many Variables



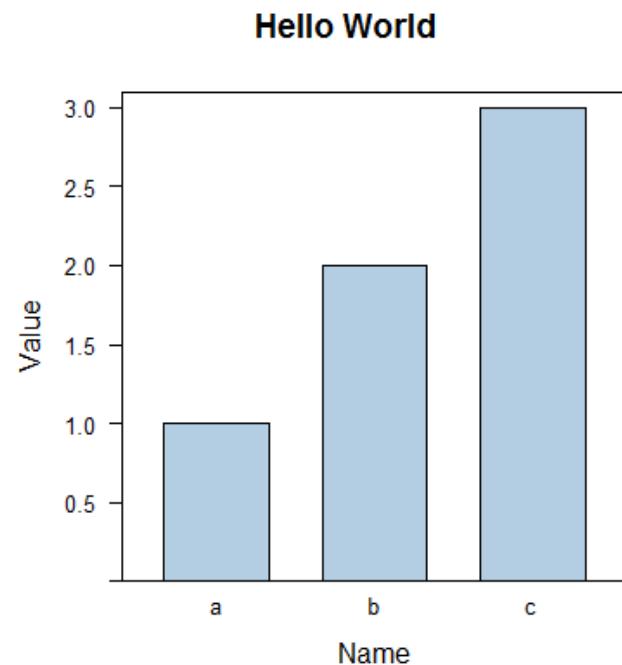


Source: Nathan Yau ([www.flowingdata.com](http://www.flowingdata.com))

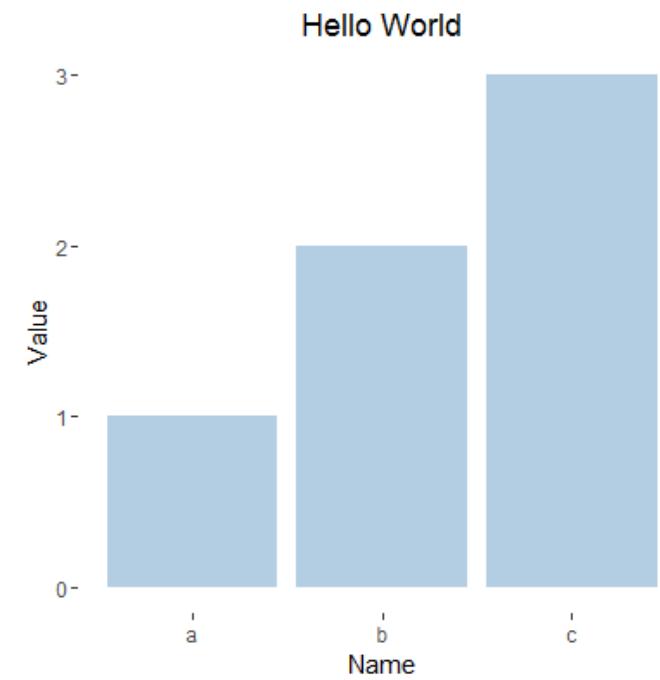
# Plotting Systems in R



Base



Lattice



ggplot2



# COWBOYS & Space Invaders: The Musical



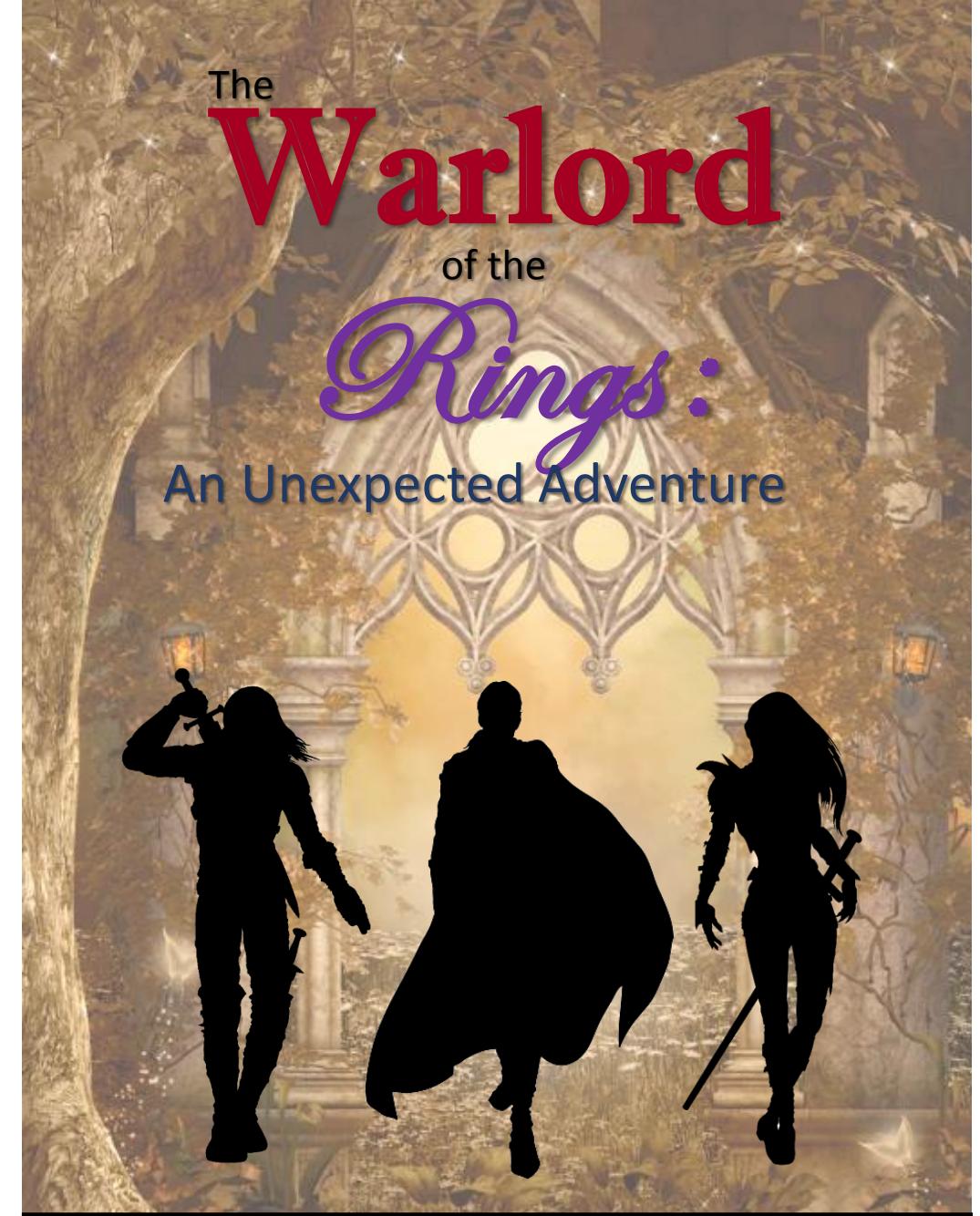
Extended Edition



# Code Demo

# Lab 4

## Data Visualization



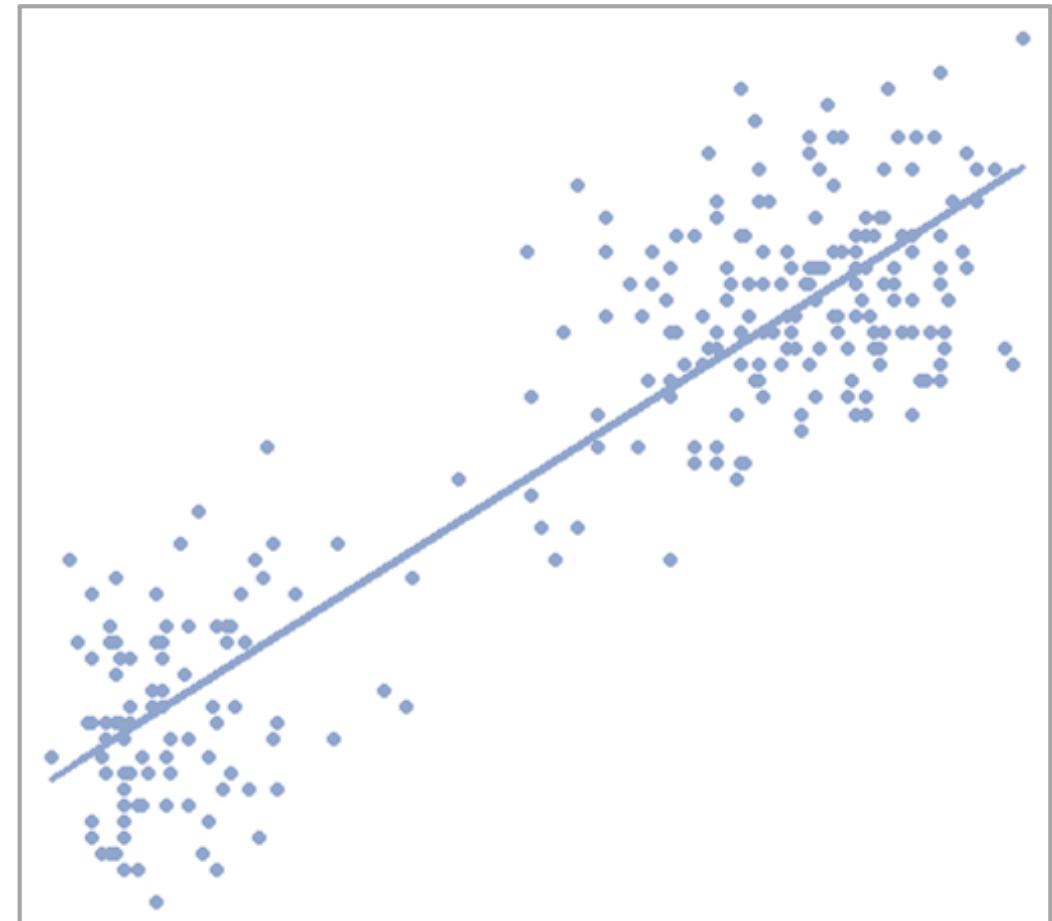
Feature Length

PG

# Statistical Modeling

# Statistical Models

Representation of reality  
Mathematical function  
Parametric models  
Non-parametric models



# Types of Statistical Models

Probability distribution

ANOVA

Linear regression

Non-linear regression

Bayesian network

# Gaussian Distribution

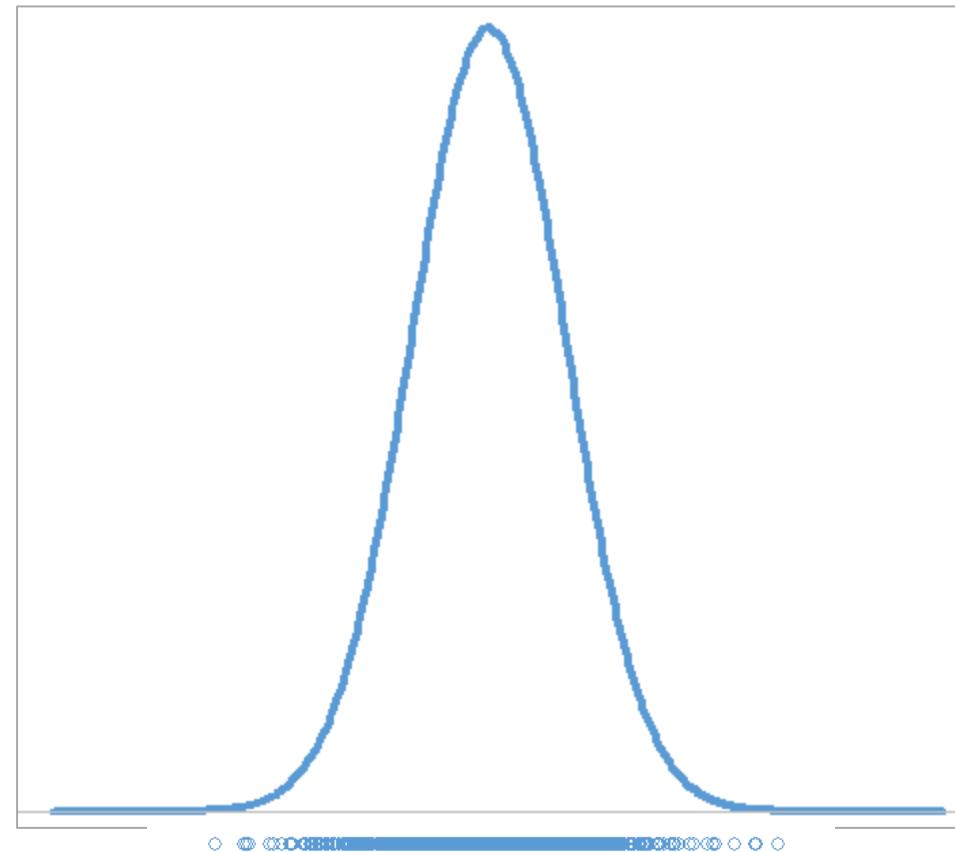
Probability distribution

Parametric model

Mean

Standard deviation

Generative model



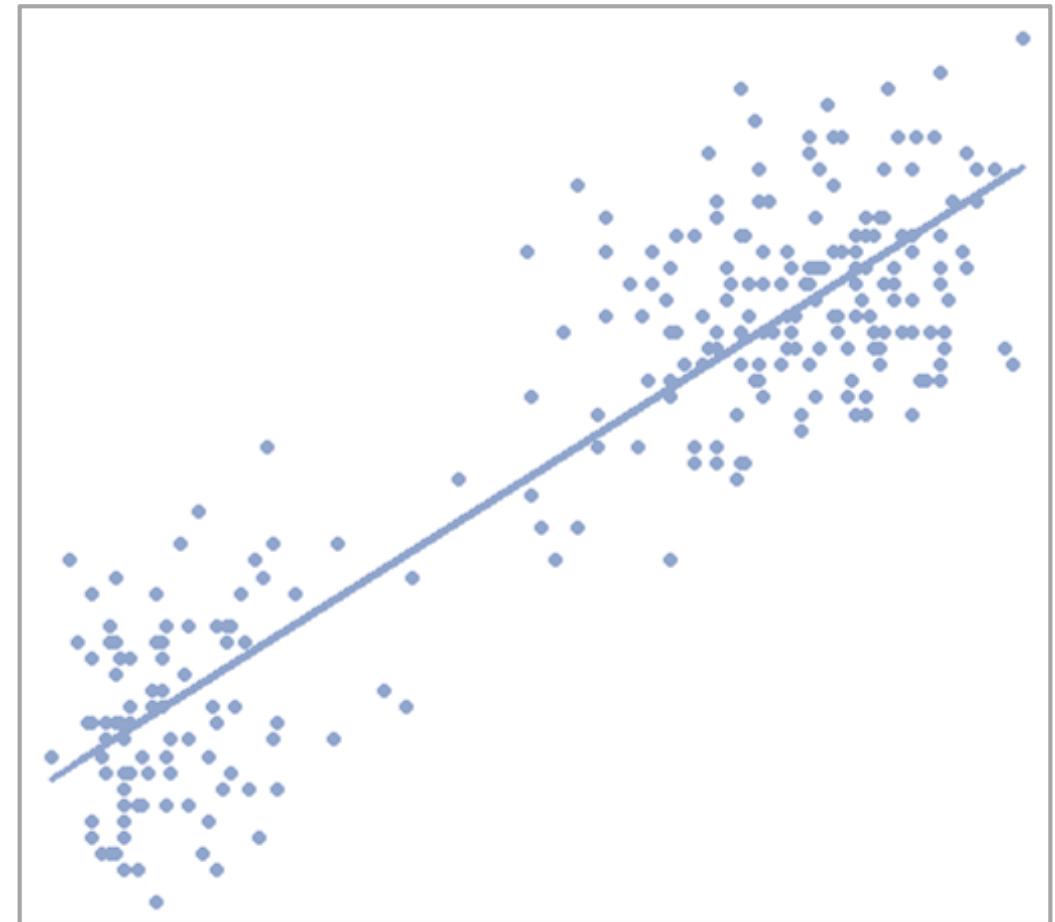
# Simple Linear Regression

Linear model

Relationship between variables

Continuous explanatory variable

Single scalar dependent variable



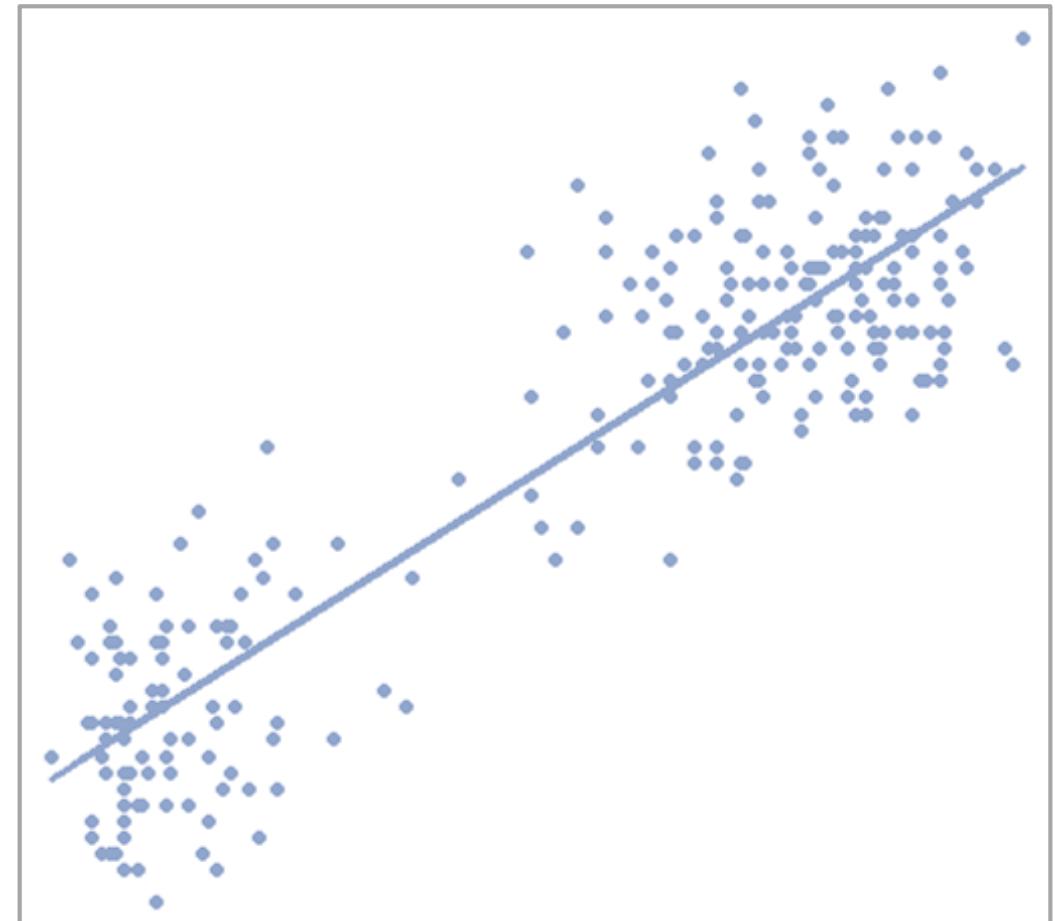
# Simple Linear Regression

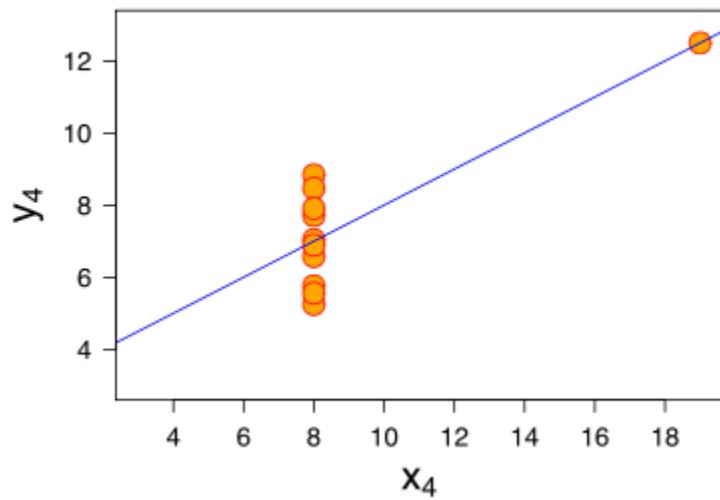
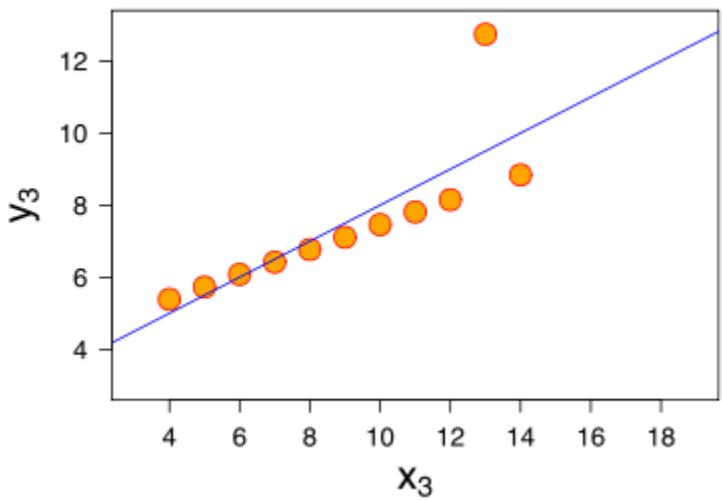
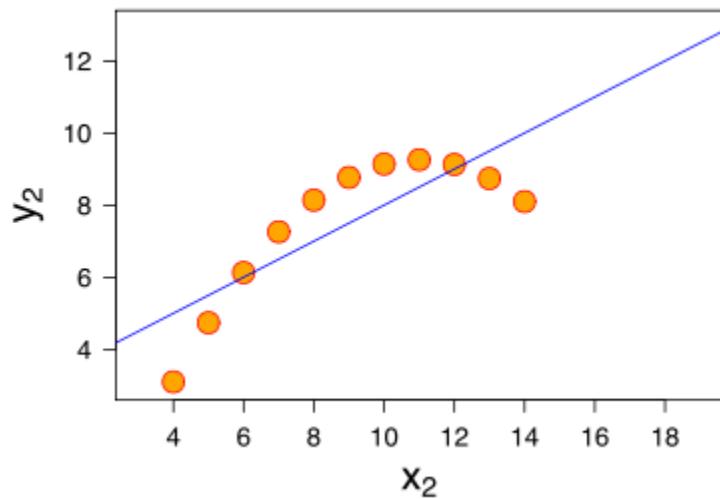
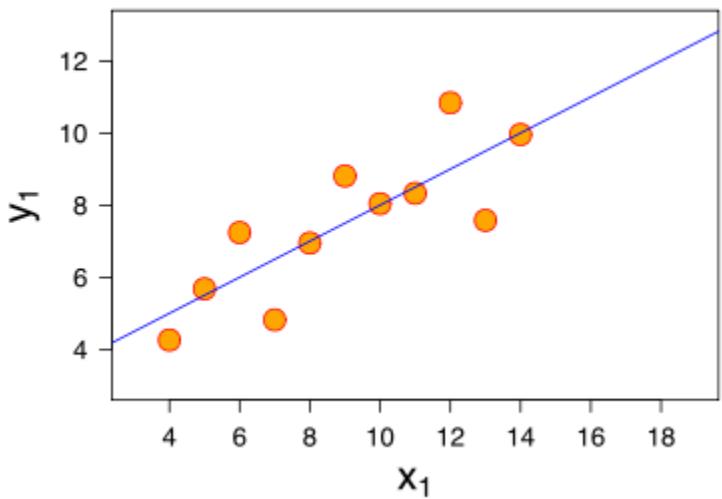
Linear predictor function

$$y = m \cdot x + b$$

Parameters estimated from data

Makes certain assumptions





Source: [https://en.wikipedia.org/wiki/Anscombe%27s\\_quartet](https://en.wikipedia.org/wiki/Anscombe%27s_quartet)





Photo by Danielle Langlois

# Code Demo

# Lab 5

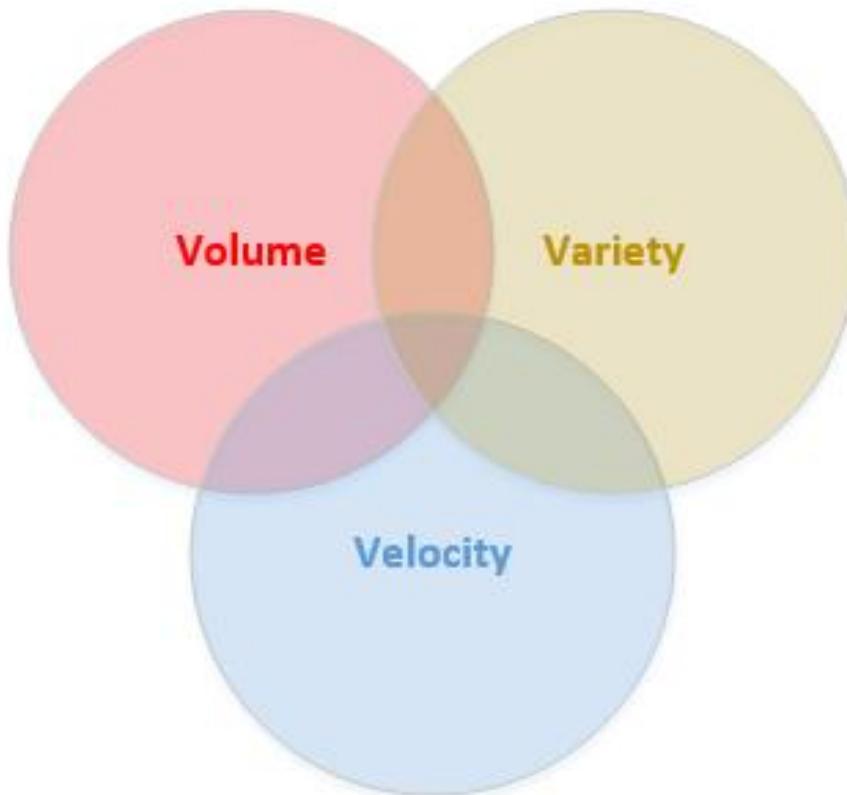
## Statistical Models



Photos by Radomił Binek,  
Danielle Langlois, and Frank Mayfield

# Handling Big Data

# What is Big Data?



# Big Data is a Moving Target

What is Big Data today

Will not be Big Data tomorrow



# Do I Have Big Data?

Can it fit in memory?

Can it fit on your hard drive?

Can you process it in a day?

Does it fit into tables?



# Big Data Decision Table

class	size	manage with	how it fits	examples
<b>small</b>	< 10 GB	Excel, R	fits in one machine's memory	thousands of sales figures
<b>medium</b>	10GB-1TB	indexed files, monolithic DB	fits on one machine's disk	millions of web pages
<b>Big</b>	> 1TB	Hadoop, distributed DBs	stored across many machines	billions of web clicks

Note: Last updated in 2010

Source: Michael E Driscoll

# Decision Table for Big Data in R

Class	Rows	Manage with
Small	Less than 1 million	R on desktop computer
Medium	1 million to 1 billion	R with 3 <sup>rd</sup> party support
Big	More than 1 billion	R with big data support

Source: Jan Wijffels at useR!-Conference (2013)

If you don't have a big data problem,  
but you think you have a big data problem,  
then you've just created a big data problem!

# How to Handle Big Data in R

Sampling

Microsoft R Open

3<sup>rd</sup>-party medium data packages

3<sup>rd</sup>-party big data packages

Microsoft R Server

# Sampling

Randomly selected  
Subset of original data  
Low-cost solution

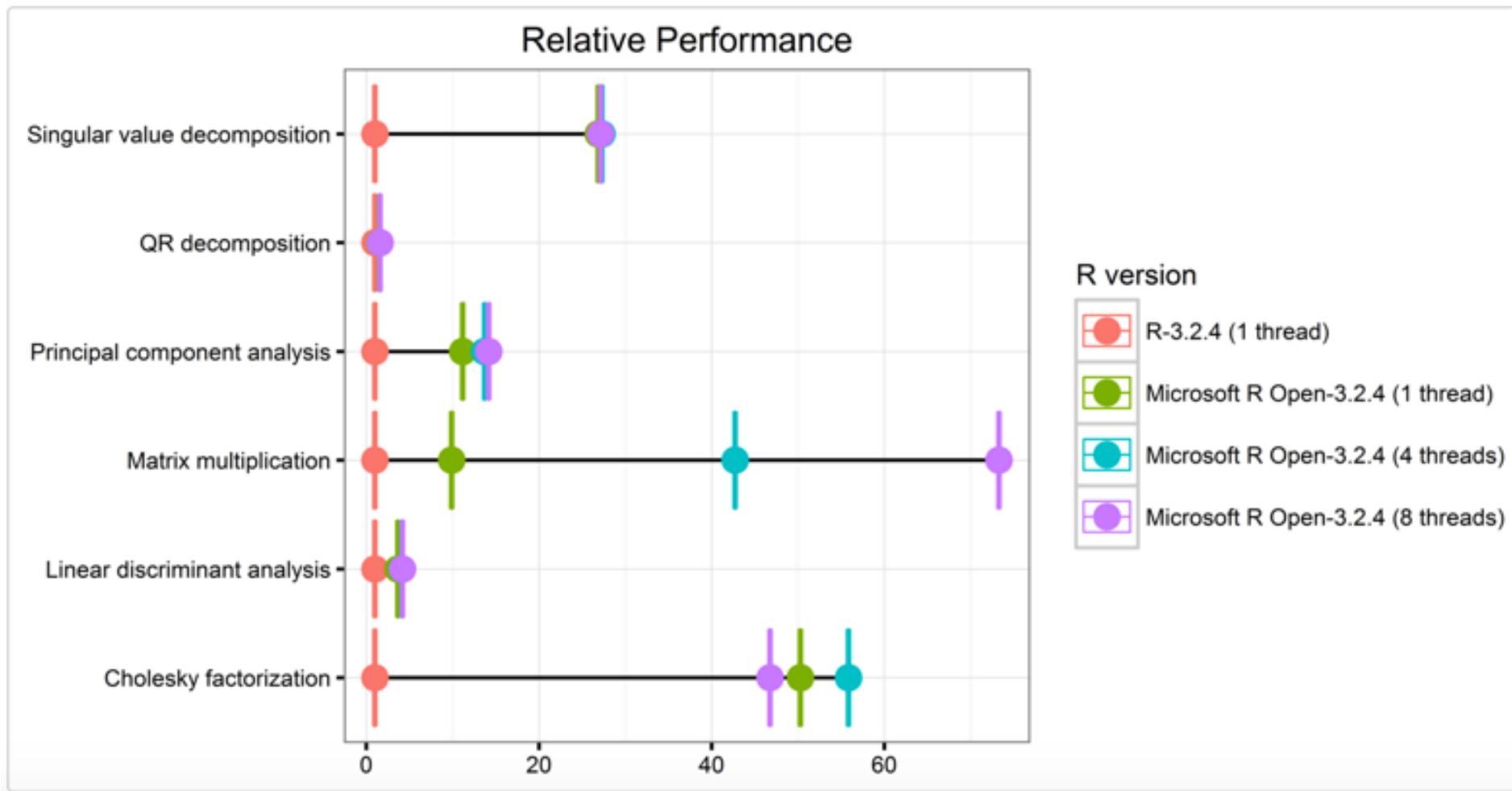
# Microsoft R Open

- Enhance 64-bit R distribution
- Multithreaded performance
- Only helps with CPU constraint
- Package repository snapshots



Source: Microsoft R Open

# Microsoft R Open Performance



Source: <https://mran.microsoft.com/documents/rro/multithread/#mt-bench>

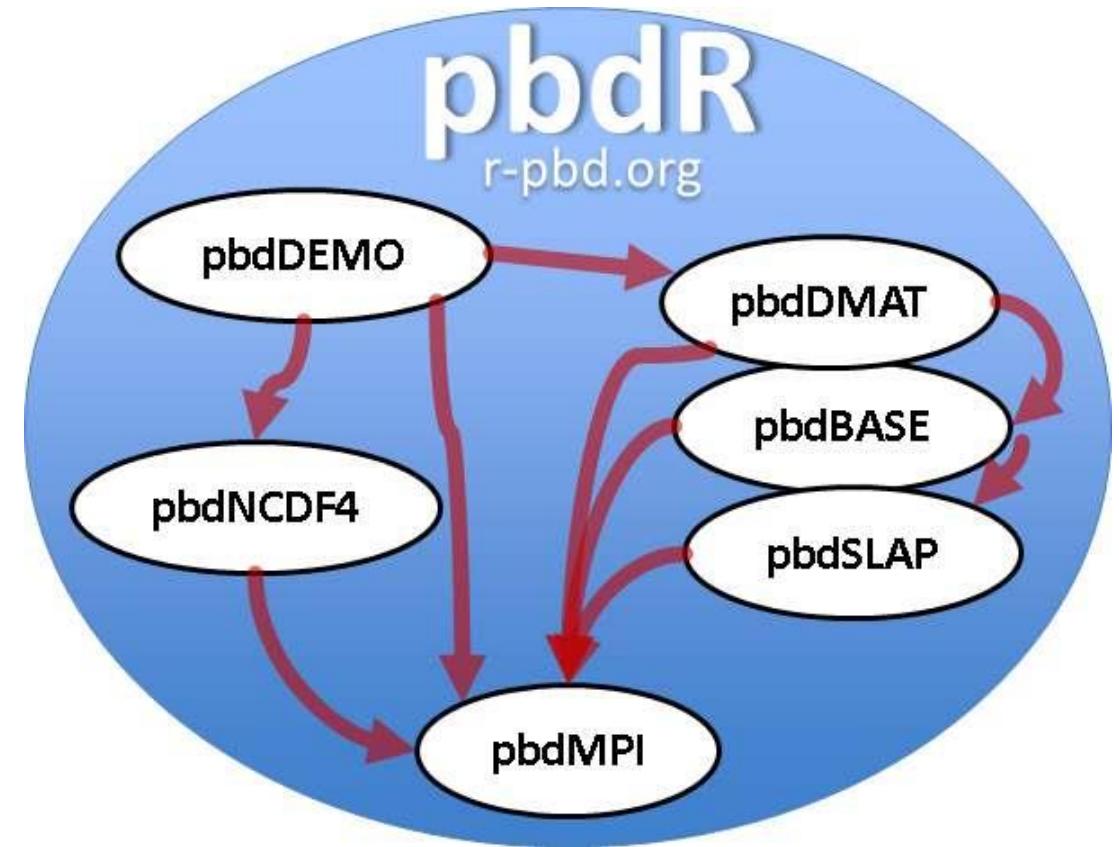
# 3<sup>rd</sup>-Party Extension Packages

Big Memory

ff

# 3<sup>rd</sup>-Party Extension Packages

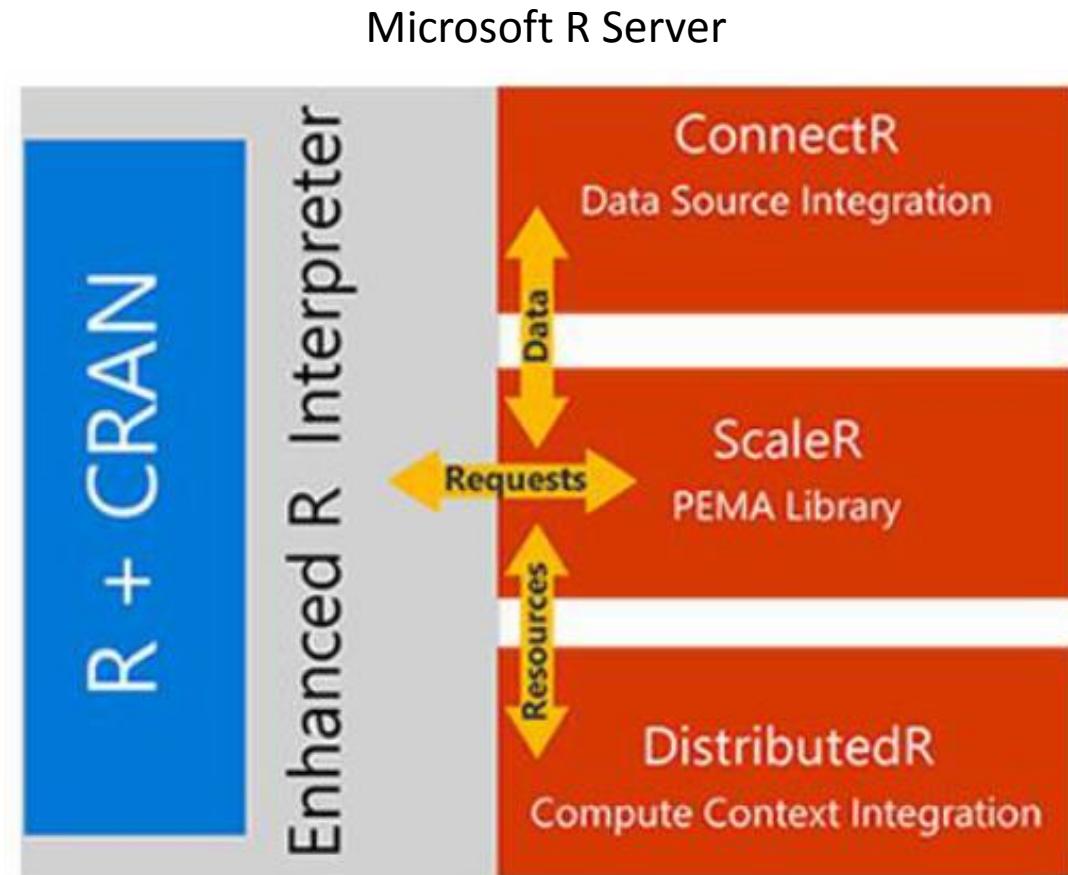
pbdR  
Rhipe  
Rhive  
Rbase  
Rhdfs,  
Rmr



Source: Wikipedia

# Microsoft R Server

Windows  
SQL Server 2016  
SUSE Linux  
Redhat Linux  
Teradata  
Hadoop  
HD Insight



Source: Microsoft R Server

No Demo  
or Lab 6

# Machine Learning

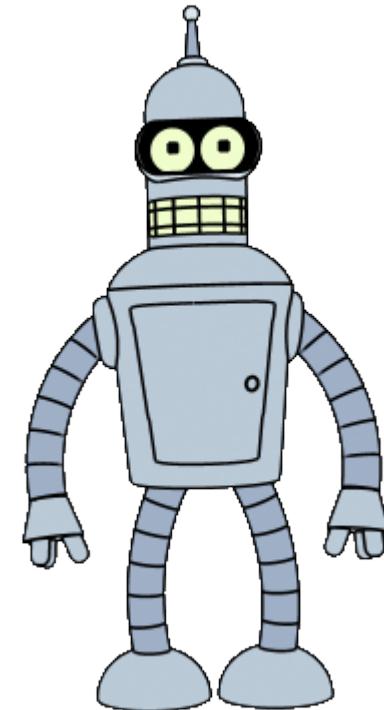
# What is Machine Learning?

Subtype of artificial intelligence

Learning without being programmed

Similar to statistical modeling

Similar to data mining



Source: Futurama

# How does Machine Learning Work?

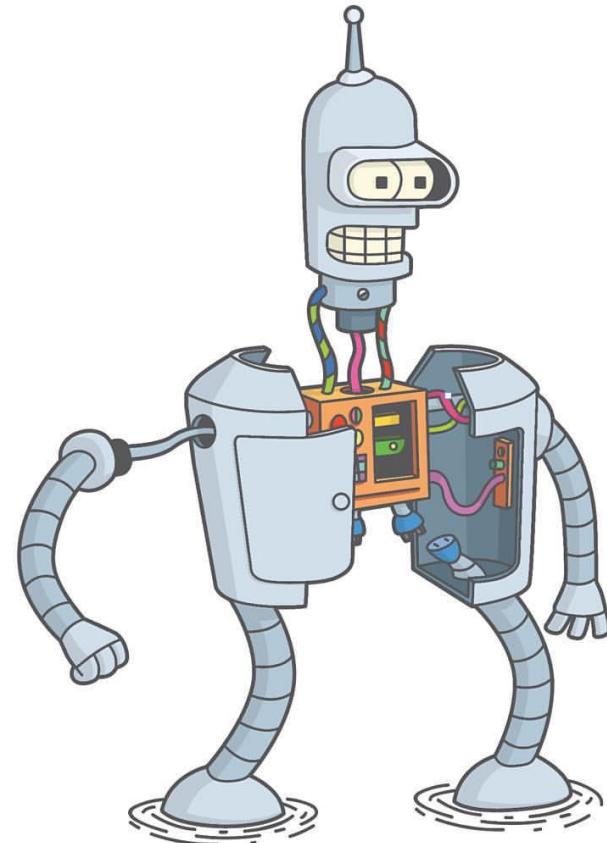
Uses statistical models

Model's parameters are trained

Trained with algorithm and data

Model is used to predict output

Prediction vs. explanation



Source: Futurama

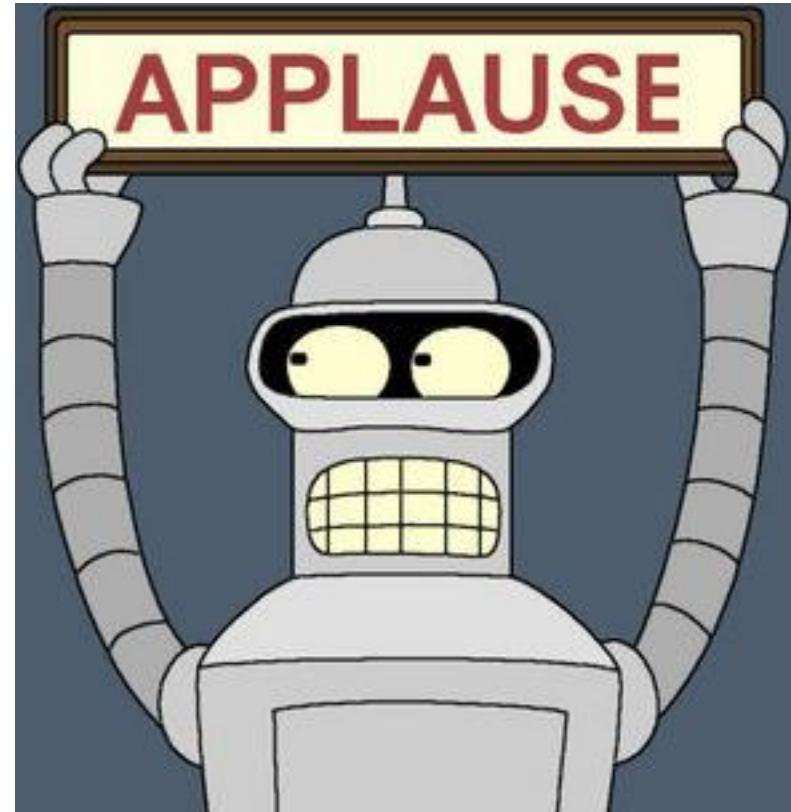
# What Can Machine Learning Do?

Classification

Regression

Clustering

Anomaly detection



Source: Futurama

# Types of Machine Learning

Supervised  
Unsupervised  
Reinforcement



Source: Futurama

# Types of ML Algorithms

Decision Trees

Naïve Bayes Classifier

Linear Regression

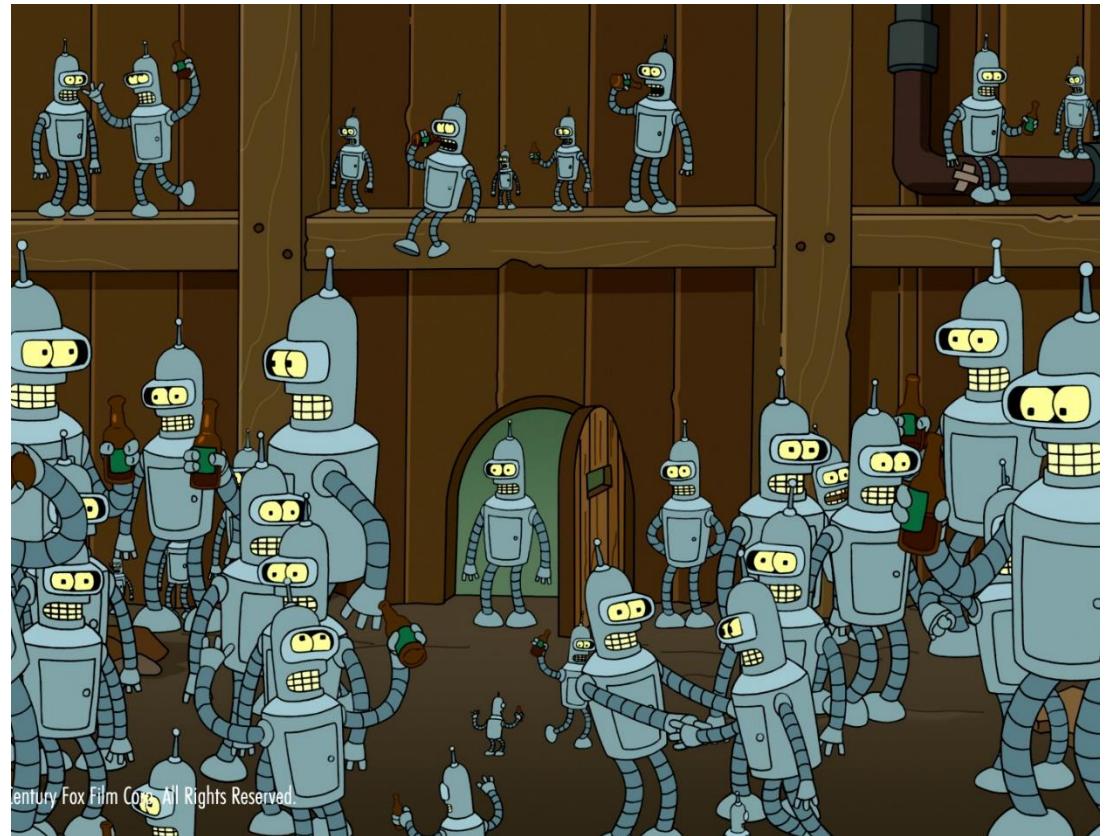
Logistic Regression

Support Vector Machines

Neural Networks

K-Means Clustering

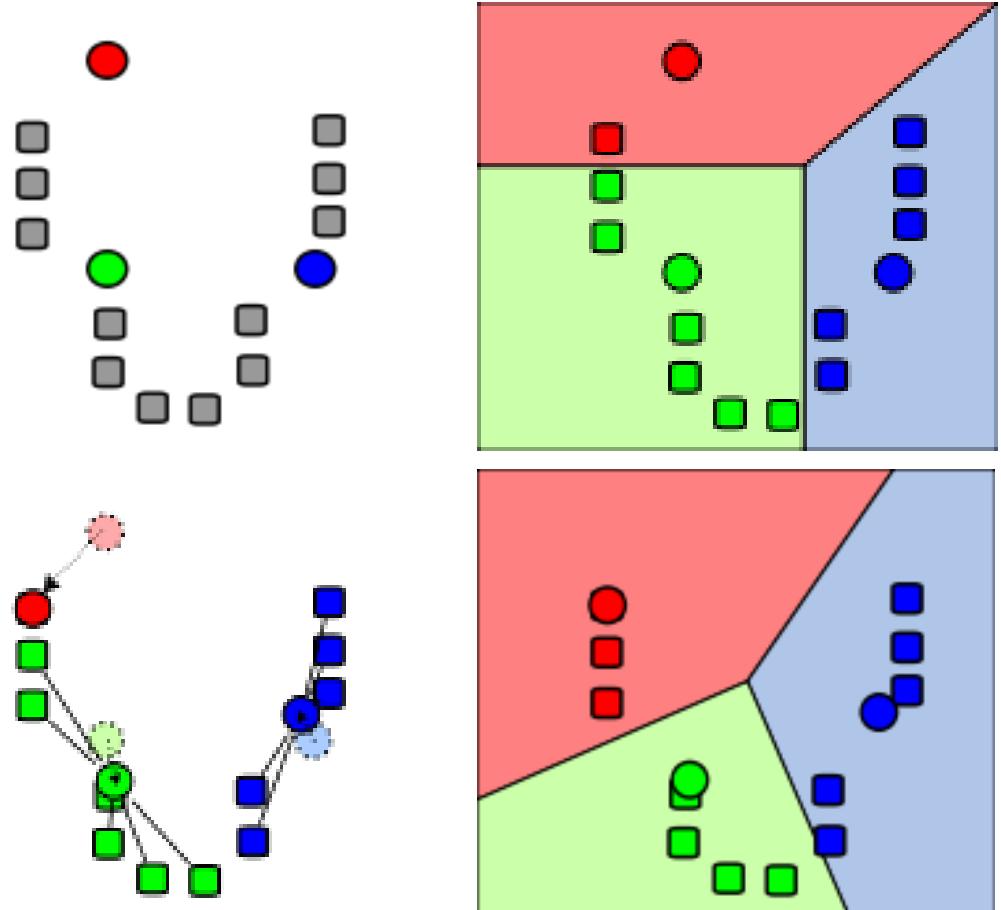
Ensemble Learning



Source: Futurama

# k-Means Clustering

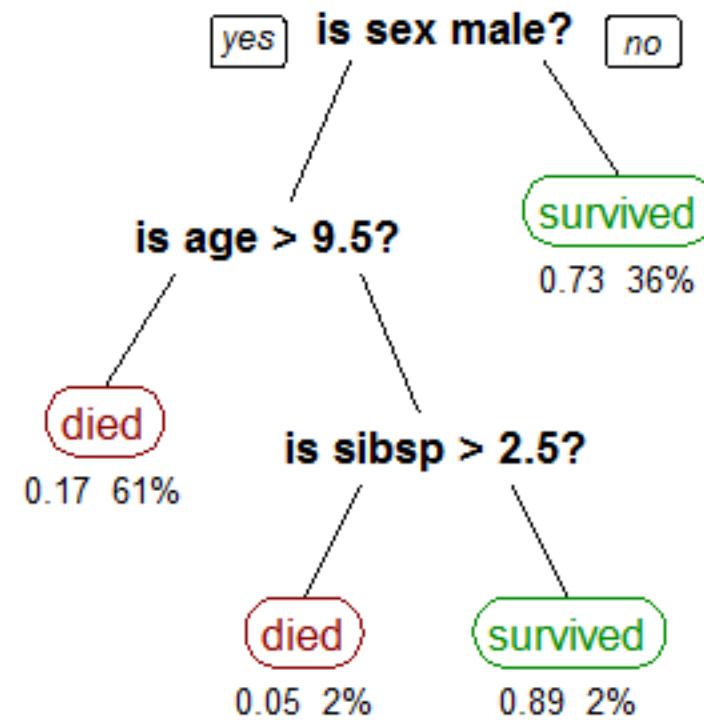
- Unsupervised learning
- Specify k (# of clusters)
- Algorithm finds centers
- Random restarts



Source: Wikipedia

# Decision Tree Classifier

Supervised learning  
Tree of decisions  
Easy to understand  
Transparent



Source: Wikipedia

# Naïve Bayes Classifier

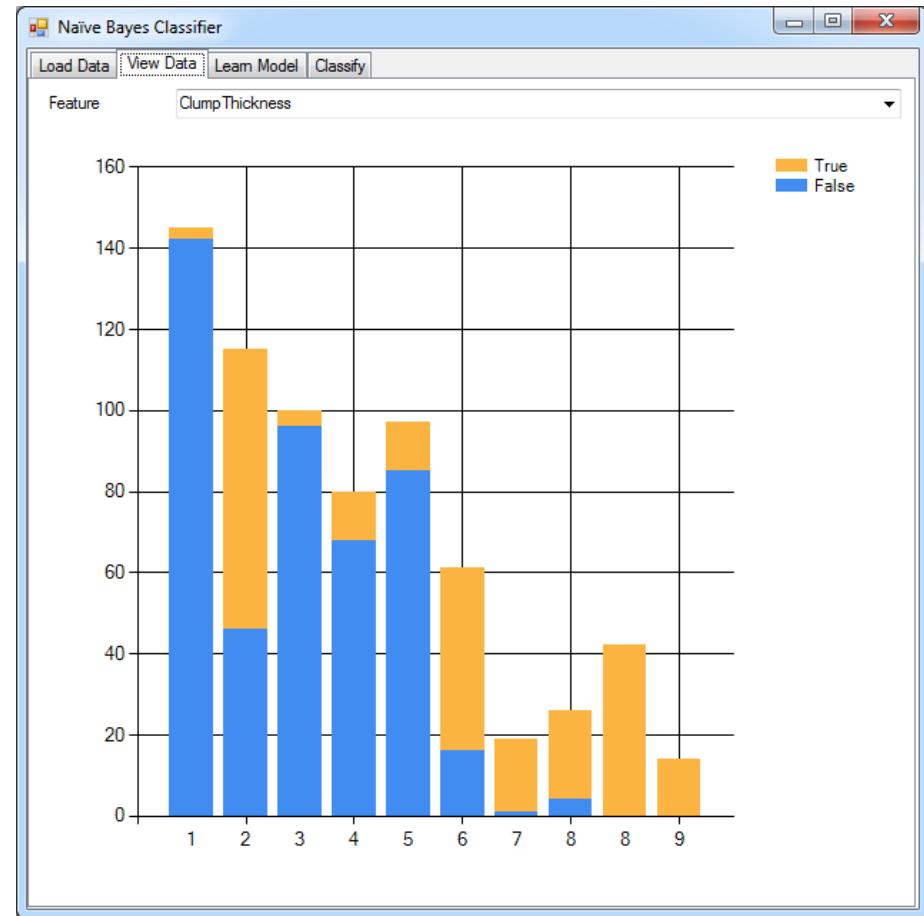
Supervised learning

Simple Bayesian classifier

Independence assumption

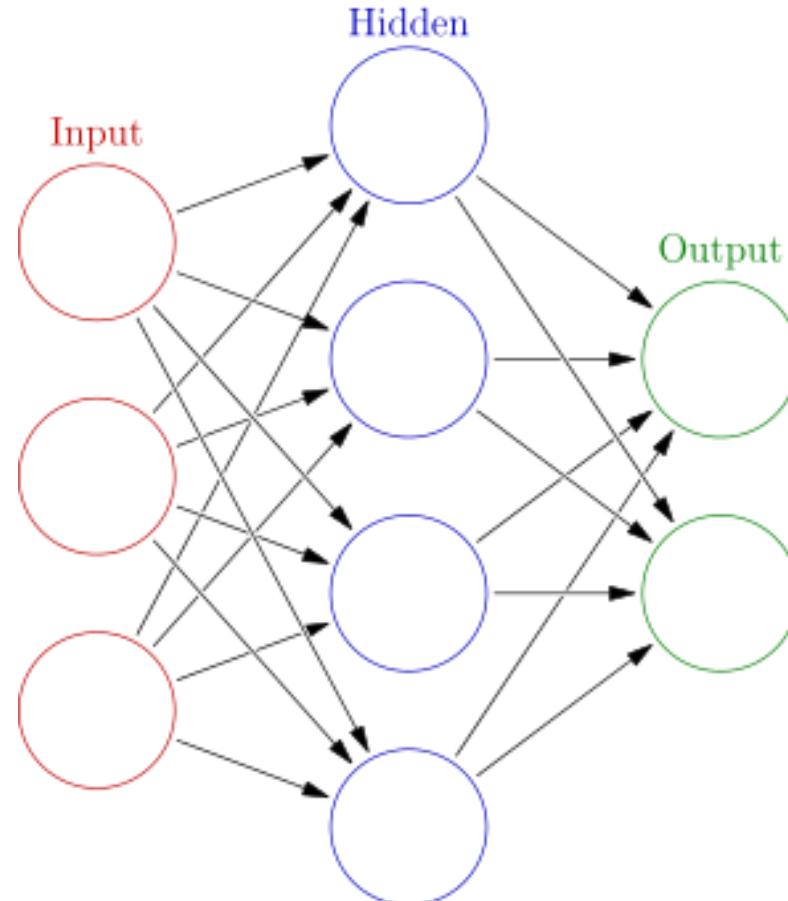
Relatively easy to understand

Transparent



# Neural Network Classifier

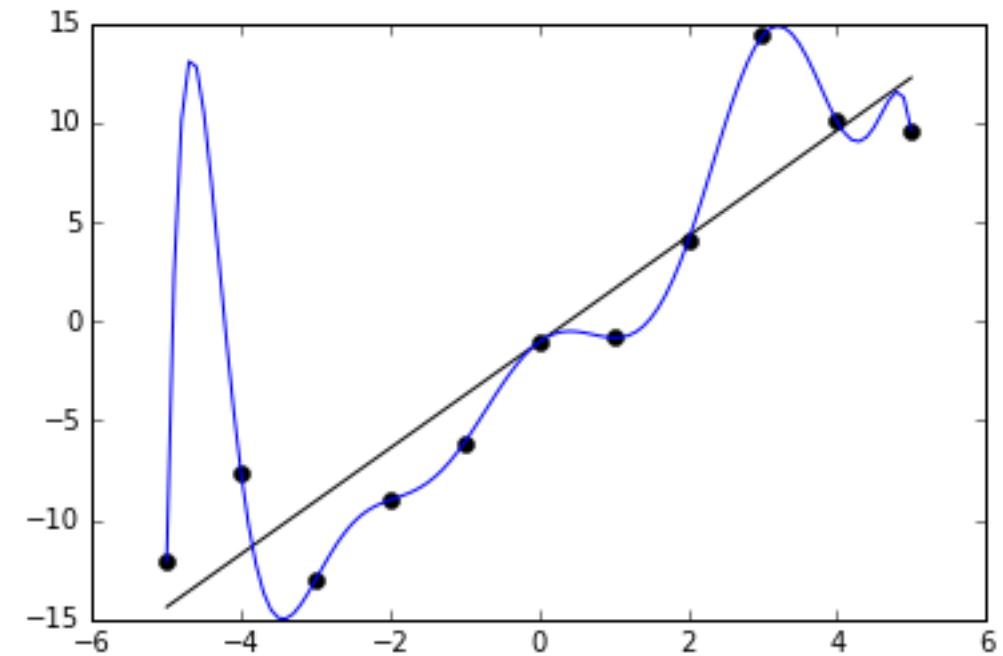
Supervised learning  
Model of neurons in brain  
Layers of neurons  
Backpropagation  
Complex  
Not transparent



Source: Wikipedia

# Overfitting and Regularization

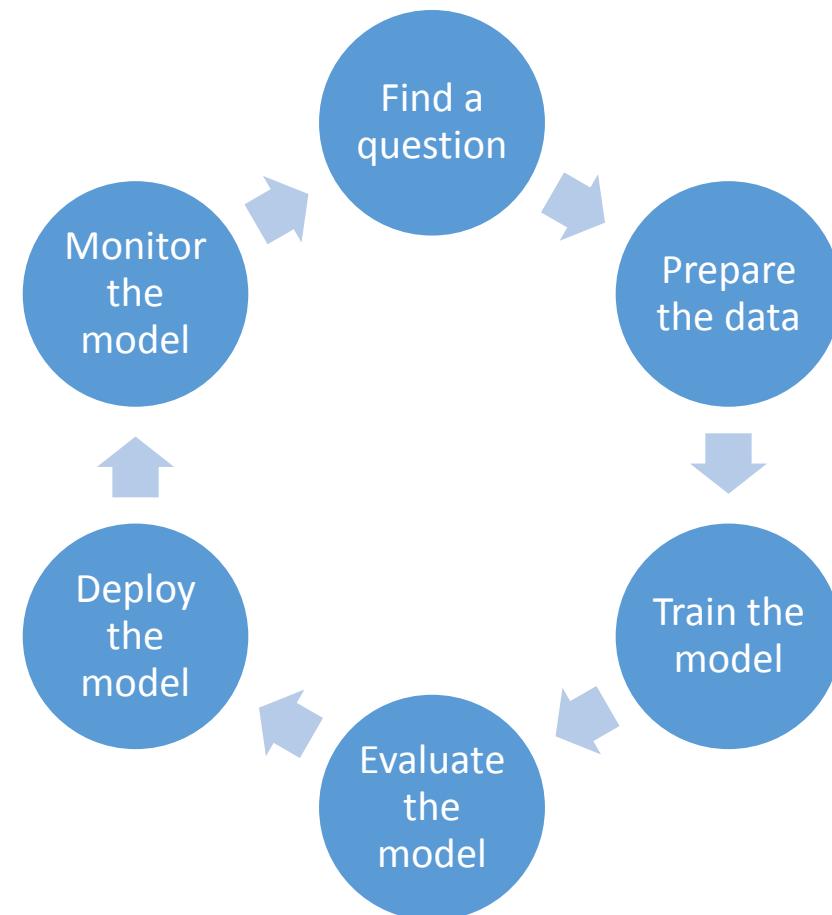
- Overfit – too specialized
- Underfit – too generalized
- Regularization techniques
  - Early stopping
  - Pruning (trees)
  - Adding noise
  - Parameter tuning



Source: Wikipedia

# The Machine Learning Process

1. Find a question
2. Prepare the data
3. Train the model
4. Evaluate the model
5. Deploy the model
6. Monitor the model





Photos by Radomił Binek,  
Danielle Langlois, and Frank Mayfield

# Code Demo

# Lab 7

# Machine Learning



# R in Practice

# How to Use R in Practice?

1. Deploying R to production
2. Best practices
3. Creating reproducible research

# How to Deploy to Production

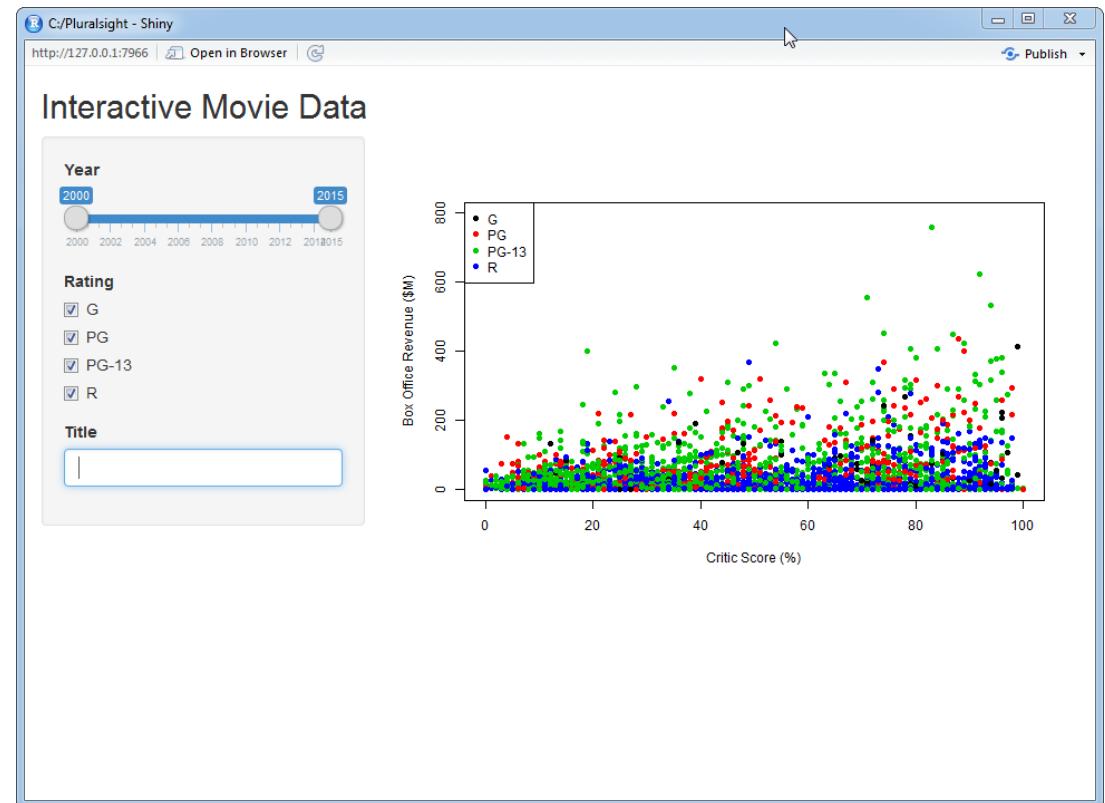
Export charts (Rstudio)

Create documents (Markdown)

Create interactive reports (Shiny)

Deploy to Server (R Server)

Deploy to Cloud (Azure ML)



# Code Demo



ADVICE

TIPS

GUIDANCE

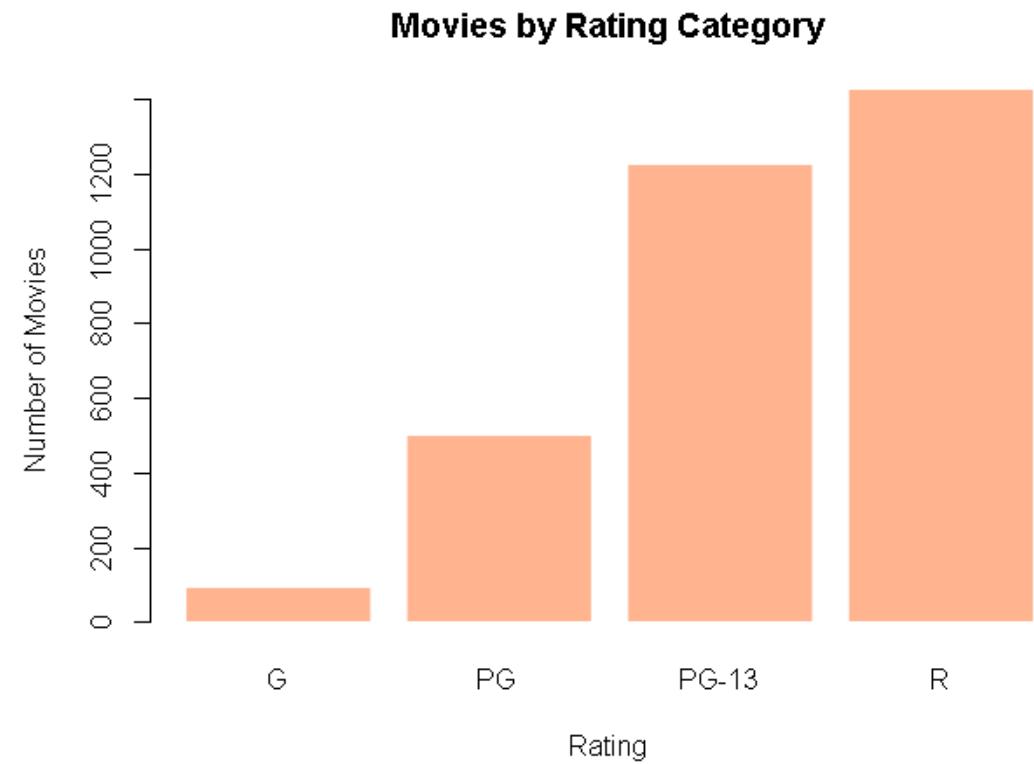
HELP

SUPPORT

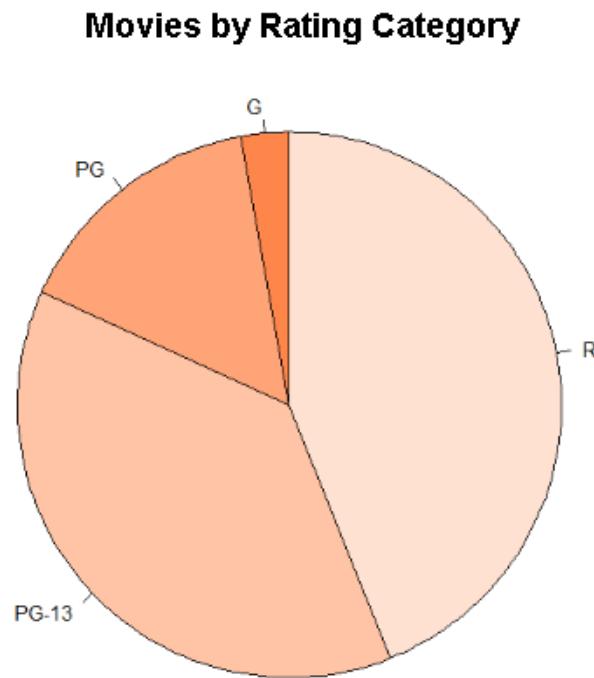
GUIDANCE

# Start with a Question

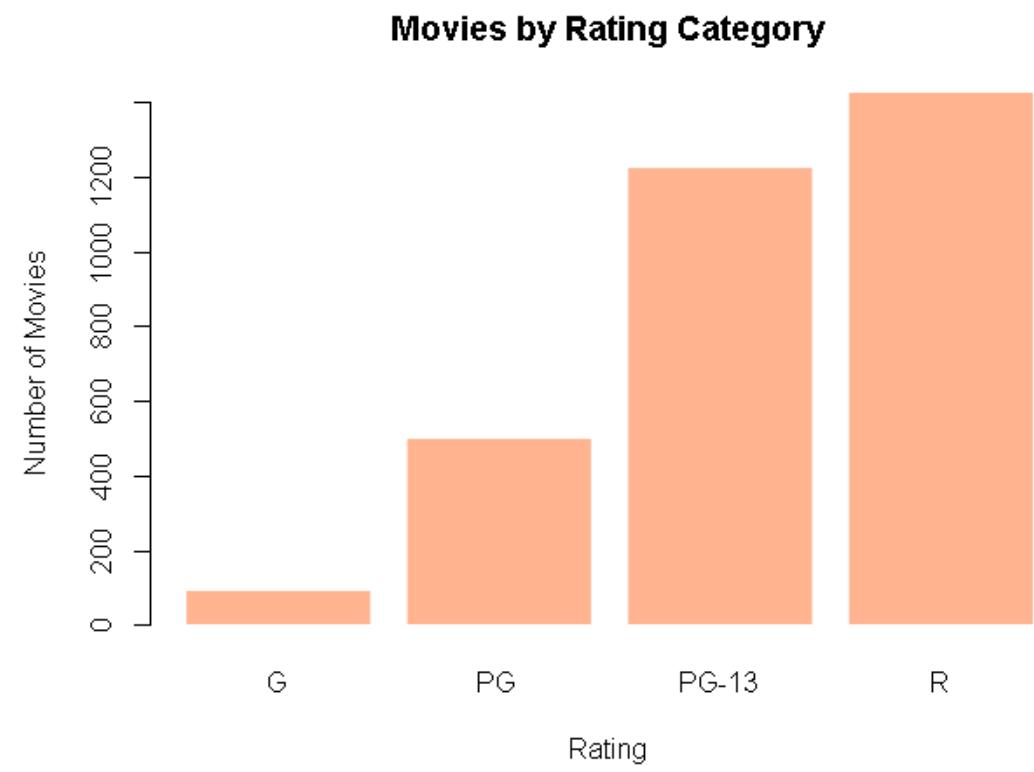
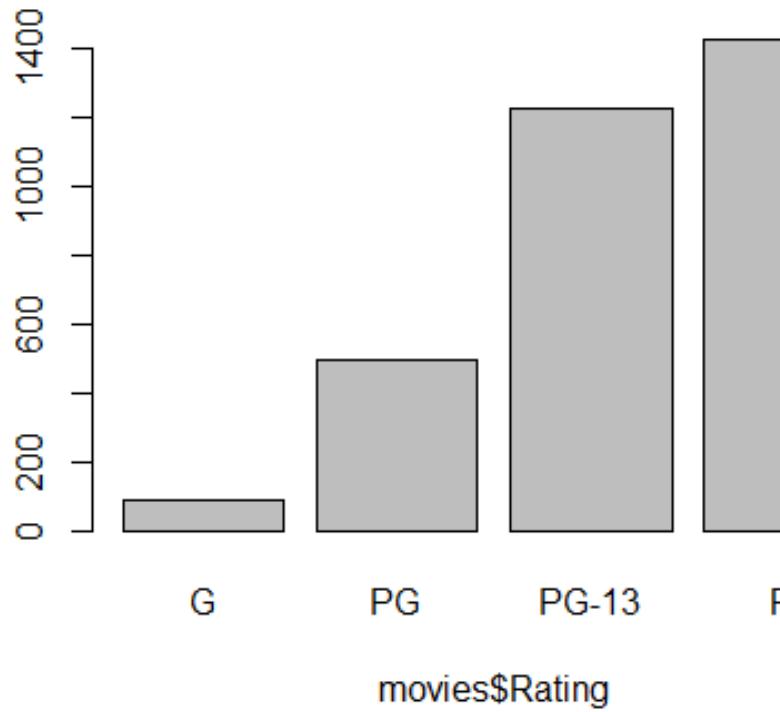
How many movies  
were released in  
each rating category  
from 2000 to 2015?



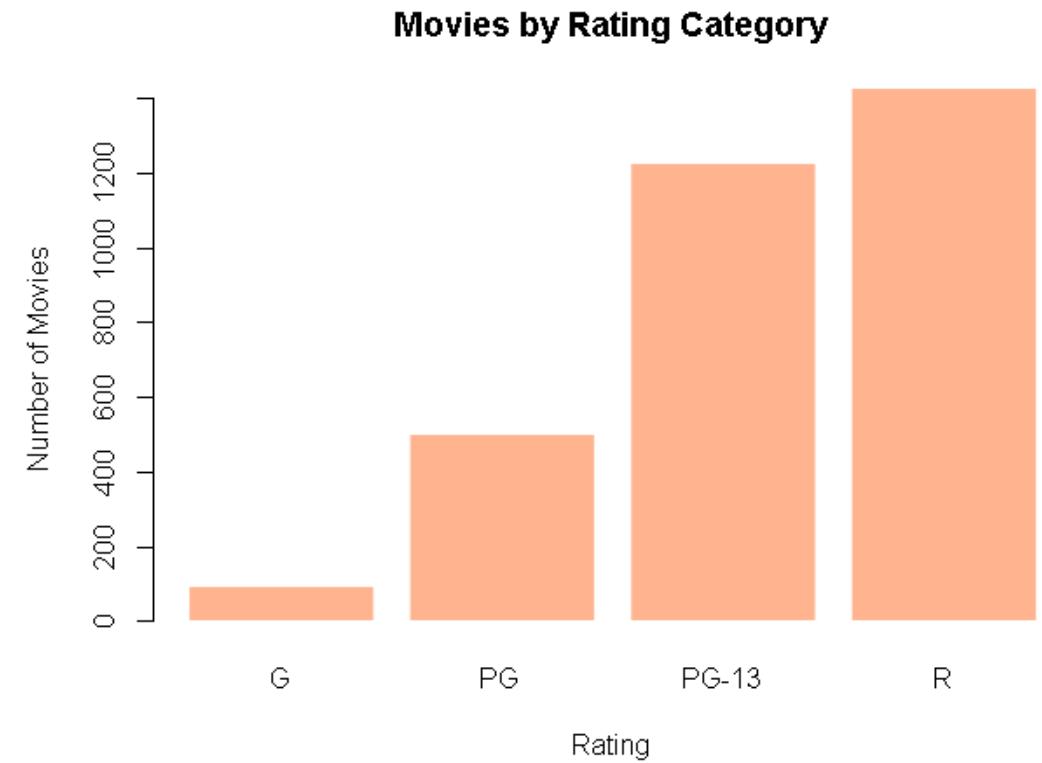
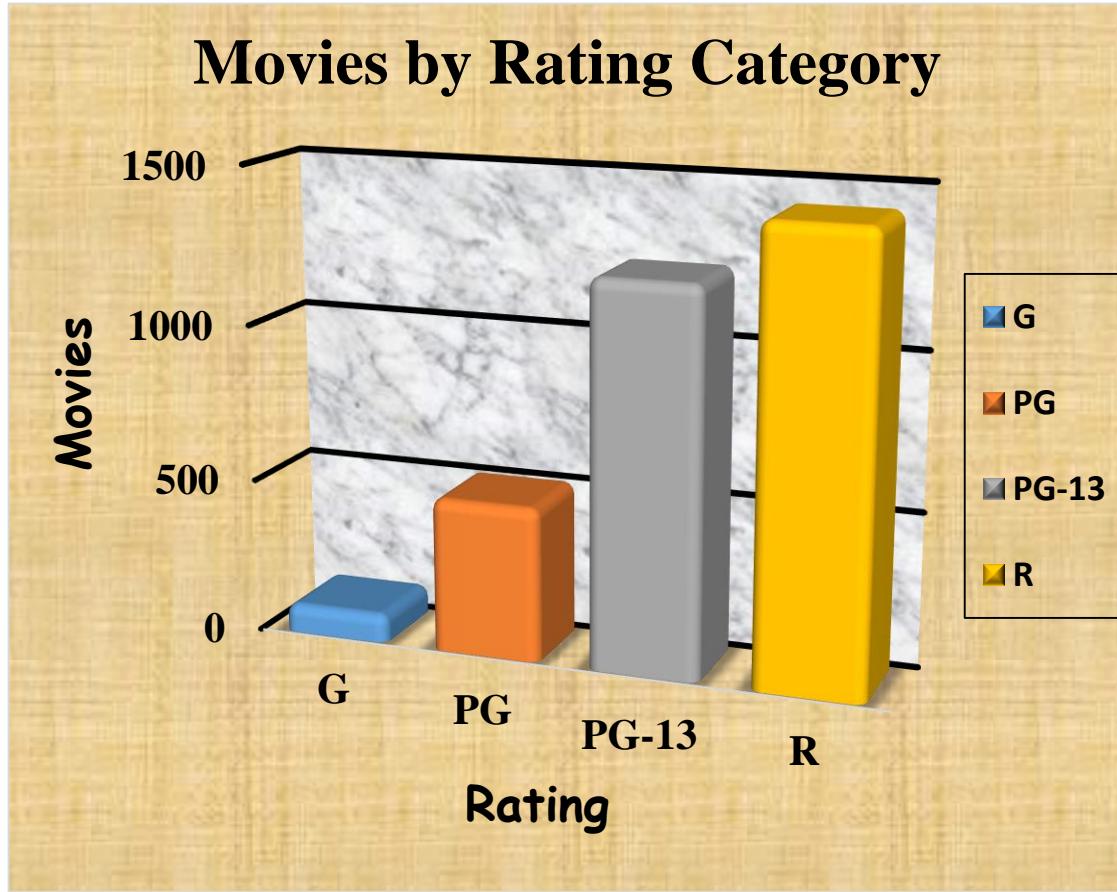
# Use the Right Tool for the Job



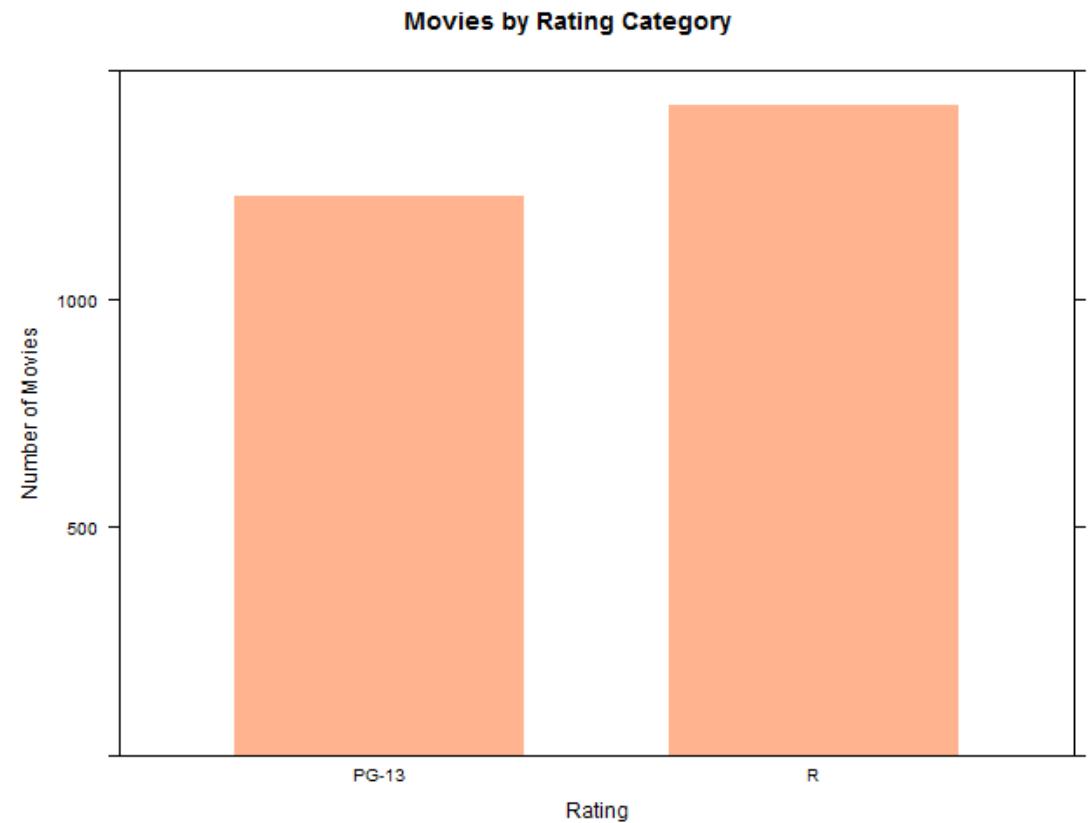
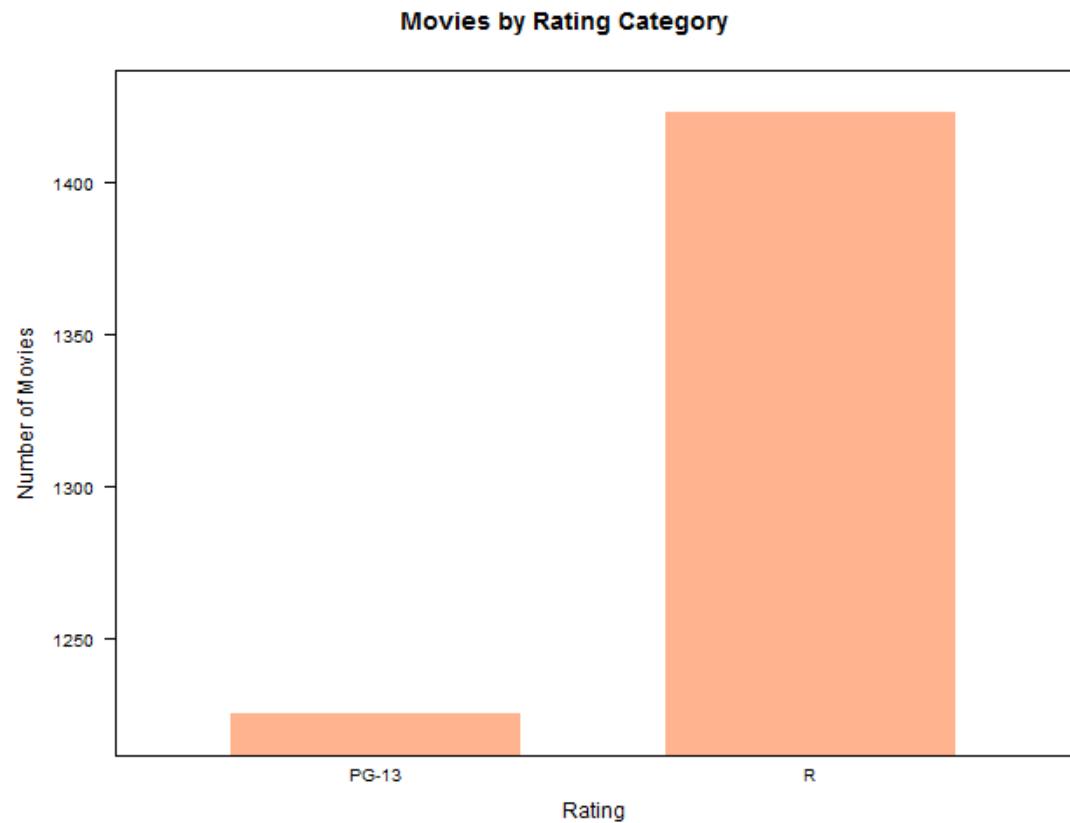
# Know Your Audience



# Create Clean Data Analyses



# Avoid Biases and Information Distortions



# Create Reproducible Research

Replication is hallmark of science  
Big issue in science right now  
Allows other to verify findings  
Creates transparency  
Allows other to build upon work



Source: <https://blog.mendeley.com>

# How Do We Create Reproducible Research?

- Provide raw data and code
- Script all analysis steps
- Use source control
- State all assumptions
- Use markdown



Source: <https://blog.mendeley.com>

# Where to Go Next...

Coursera: <https://www.coursera.org/specializations/jhu-data-science>

Revolutions: <http://blog.revolutionanalytics.com>

Flowing Data: <http://flowingdata.com>

R-Blogger: <http://www.r-bloggers.com>

R-Seek: <http://rseek.org>

# My Website

Articles

Presentations

Source Code

Videos

Workshops

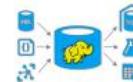
Matthew Renze

Home Articles Courses Presentations Software About Contact

## News

2016-07-11 - The Big Data Refinery

I wrote an article describing the Data Refinery pattern, which is a pattern for handing multiple consumers of Big Data. I learned about this pattern from my interactions with the Big Data Group at Microsoft.



2016-07-01 - Microsoft MVP Award

I received my first Microsoft MVP Award today. Very happy to be part of such an amazing group of people! In addition, I'm really looking forward to attending the Microsoft MVP Global Summit again in November.



Matthew is an independent software consultant, author for Pluralsight, international public speaker, a Microsoft MVP, ASPIndustry, and open-source software contributor.

2016-06-26 - JavaScript Air Interview

Kent Dodds invited me to be on his podcast JavaScript Air at KCDC. The video and audio of the podcast are now available online.



2016-06-25 - Lifelong Learning as a Developer

I participated in a discussion panel at KCDC on Lifelong Learning as a Software Developer. The video of the discussion panel is now available online. I thought all of the panelist did an excellent job.



[www.matthewrenze.com](http://www.matthewrenze.com)

# Conclusion

# Conclusion

1. Introduction
2. Transforming Data
3. Descriptive Statistics
4. Data Visualization
5. Statistical Modeling
6. Handling Big Data
7. Machine Learning
8. R in Practice

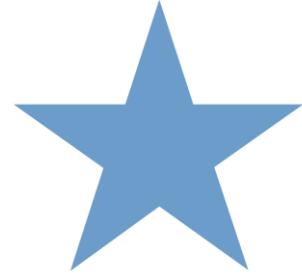


# Feedback

Feedback is very important to me!

One thing you liked?

One thing I could improve?



# Contact Info

Matthew Renze

Data Science Consultant

Renze Consulting

Twitter: [@matthewrenze](https://twitter.com/matthewrenze)

Email: [matthew@matthewrenze.com](mailto:matthew@matthewrenze.com)

Website: [www.matthewrenze.com](http://www.matthewrenze.com)



Thank You! : )