Practical Data Science with R

@MatthewRenze

#devup

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

Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

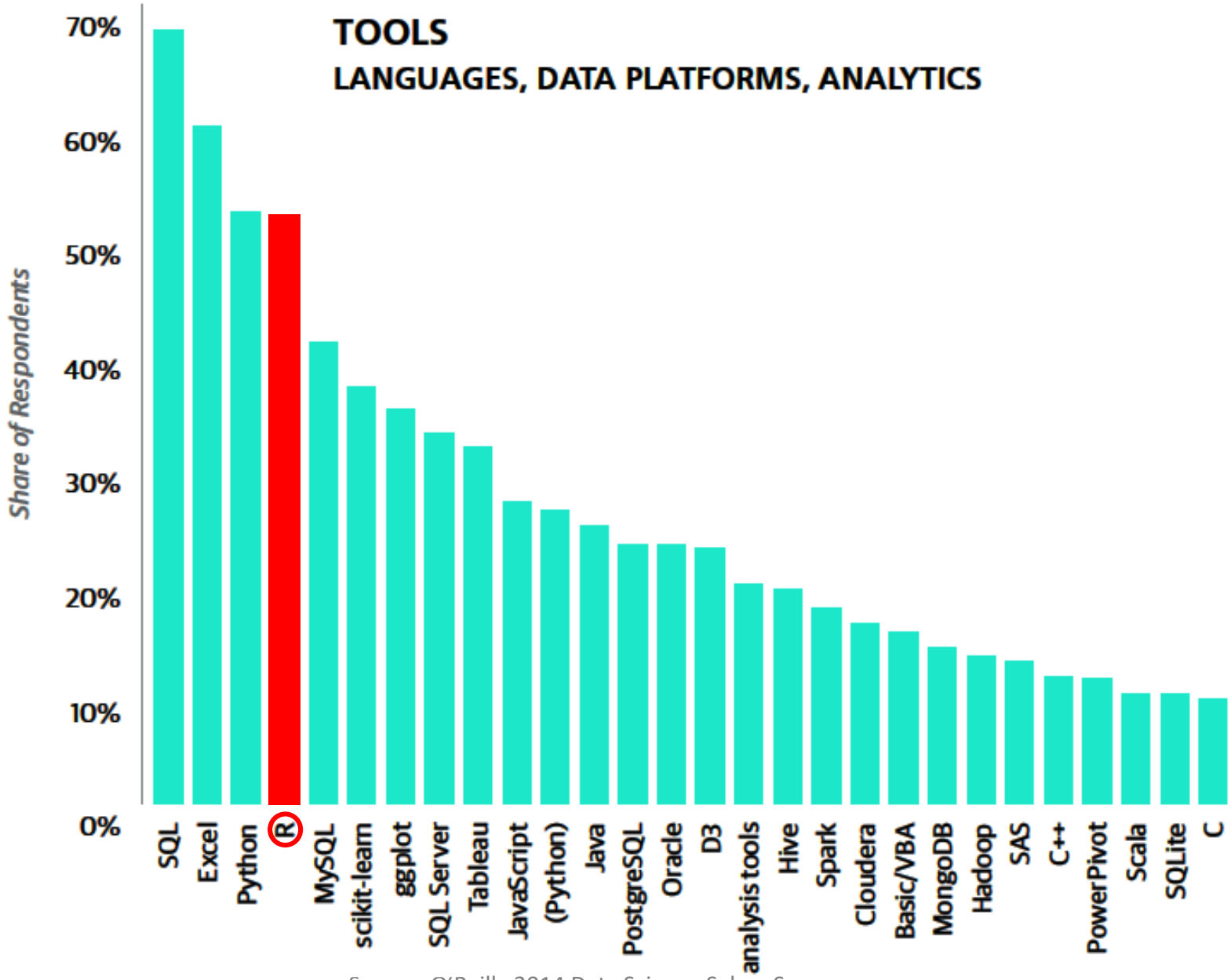
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Source: Indeed.com

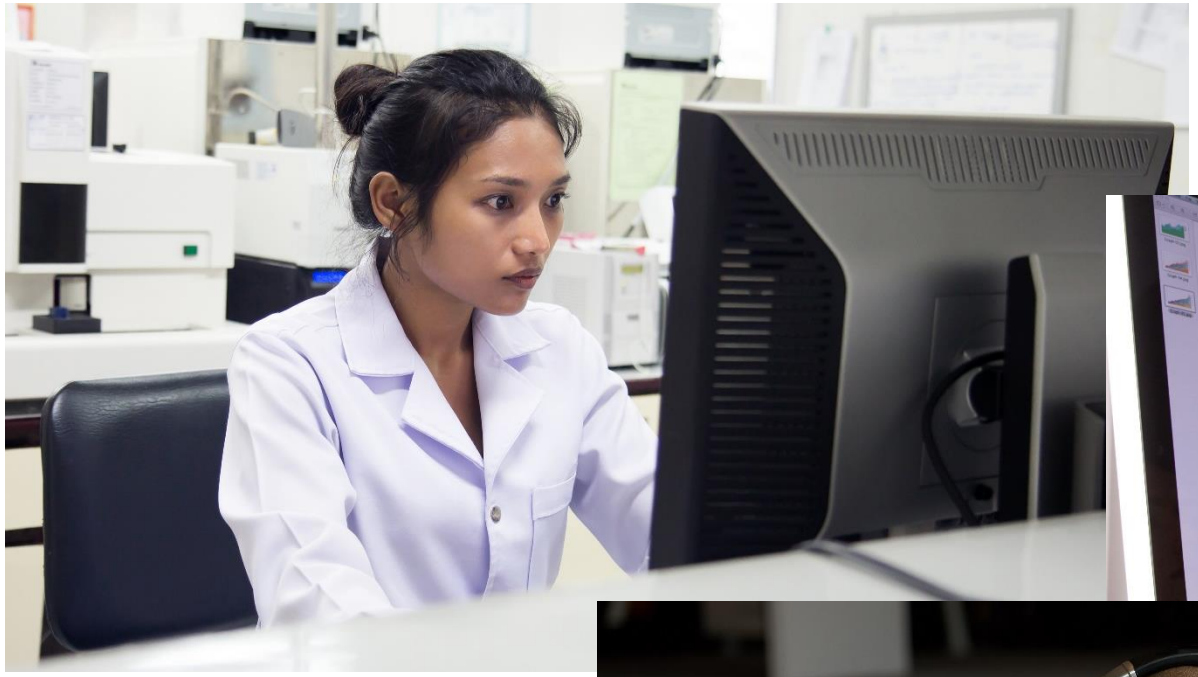
AVERAGE SALARY FOR High Paying Skills and Experience

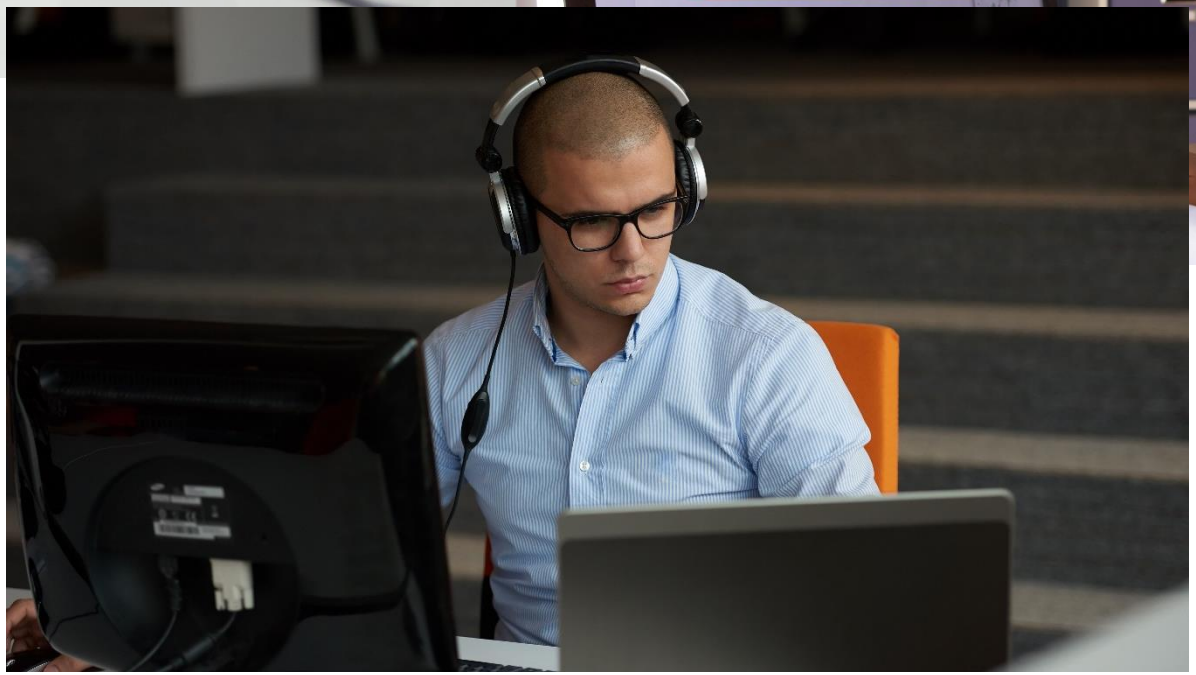
SKILL	2013	YR/YR CHANGE
R	$ 115,531	n/a
NoSQL	$ 114,796	1.6%
MapReduce	$ 114,396	n/a
PMBok	$ 112,382	1.3%
Cassandra	$ 112,382	n/a
Omnigraffle	$ 111,039	0.3%
Pig	$ 109,561	n/a
SOA (Service Oriented Architecture)	$ 108,997	-0.5%
Hadoop	$ 108,669	-5.6%
Mongo DB	$ 107,825	-0.4%
SOX (Sarbanes-Oxley)	$ 107,697	4.8%
Jetty	$ 107,406	0.4%
UML (Unified Modeling Language)	$ 107,387	4.7%

Source: Dice Salary Survey 2014



Source: O'Reilly 2014 Data Science Salary Survey





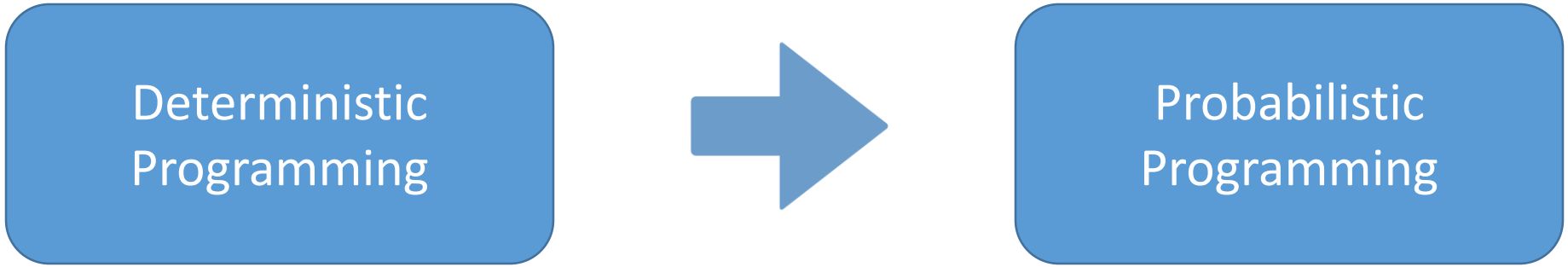
Overview

1. Introduction
2. Working with 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?

- ☑ 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?



If x and y then z

If P(x) > y then z

About Me

Data Science Consultant

Education

- B.S. in Computer Science
- B.A. in Philosophy

Community

- Public Speaker
- Pluralsight Author
- Microsoft MVP
- ASPInsider
- Open-Source Software

About You

What's your name?

What do you do?

Why did you attend?

Favorite super power?

Source: www.thatsmyface.com


thank you dev up Conference 2016 Sponsors !
PLATINUM
SYLLOGISTEKS
POLARIS SOLUTIONS
Byrne Software TECHNOLOGIES, INC.
OAKWOOD Powering Transformation. Together.
GOLD
World Wide Technology asynchrony labs
CONCERO TECHNOLOGY GROUP
ArchitectNow
VantageLinks
WORKFORCE CENTER SAINT LOUIS UNIVERSITY
KEYHOLE SOFTWARE
Daugherty BUSINESS SOLUTIONS
oasis digital
USI Ungerboeck Software INTERNATIONAL
Covenant TECHNOLOGY PARTNERS
VALOREM CONSULTING
EQUIFAX
UNISYS
ADAPTIVE SOLUTIONS GROUP
norton staffing the power of personal
rh Robert Half
SWANK MOTION PICTURES, INC. Tim Swank, Chairman
Microsoft
redgate
SILVER
LogicNP Software The power of Components

Schedule

Lectures (15 min)

Demos (10 min)

Labs (15 min)

Breaks (5 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

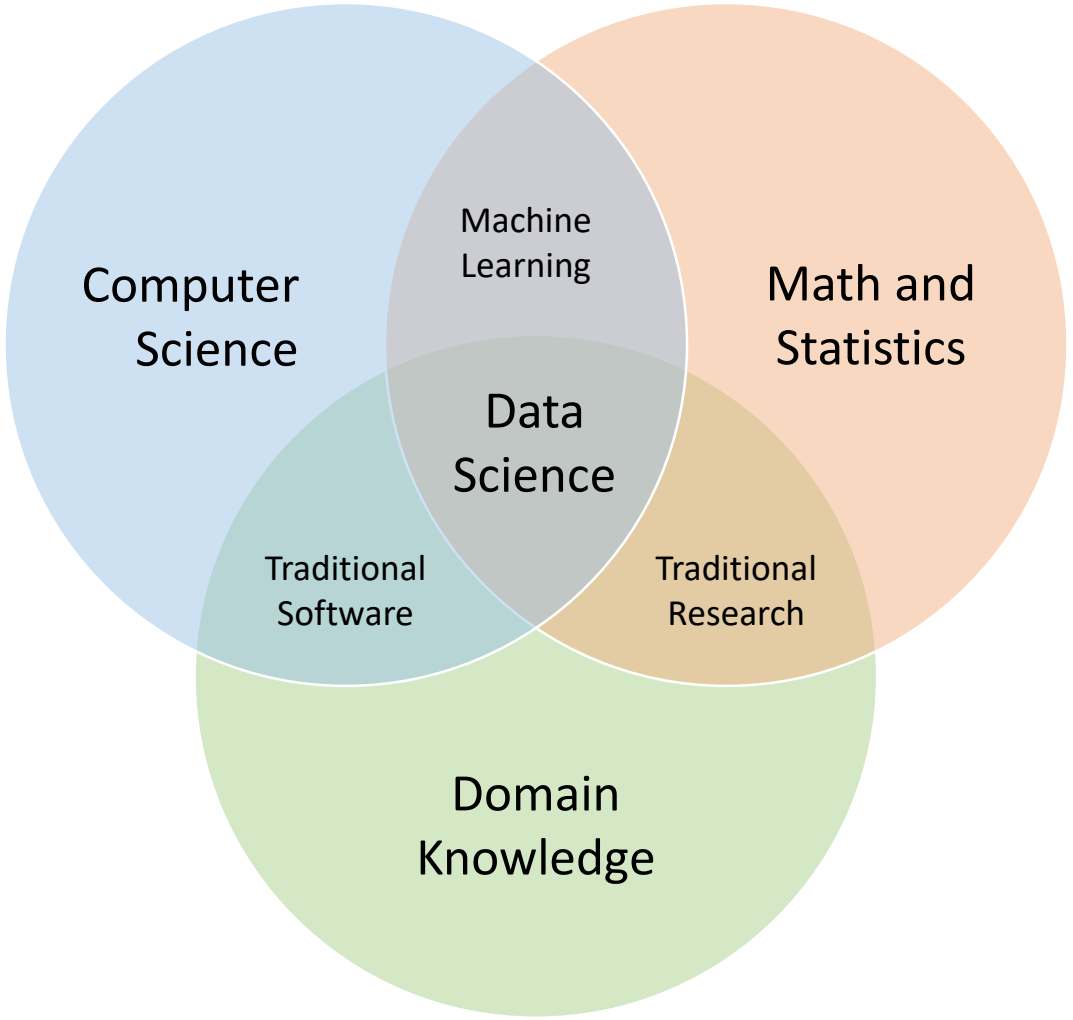
Advanced labs

Workshop URL

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

Introduction to Data Science

What is Data Science?



What is a Data Scientist?

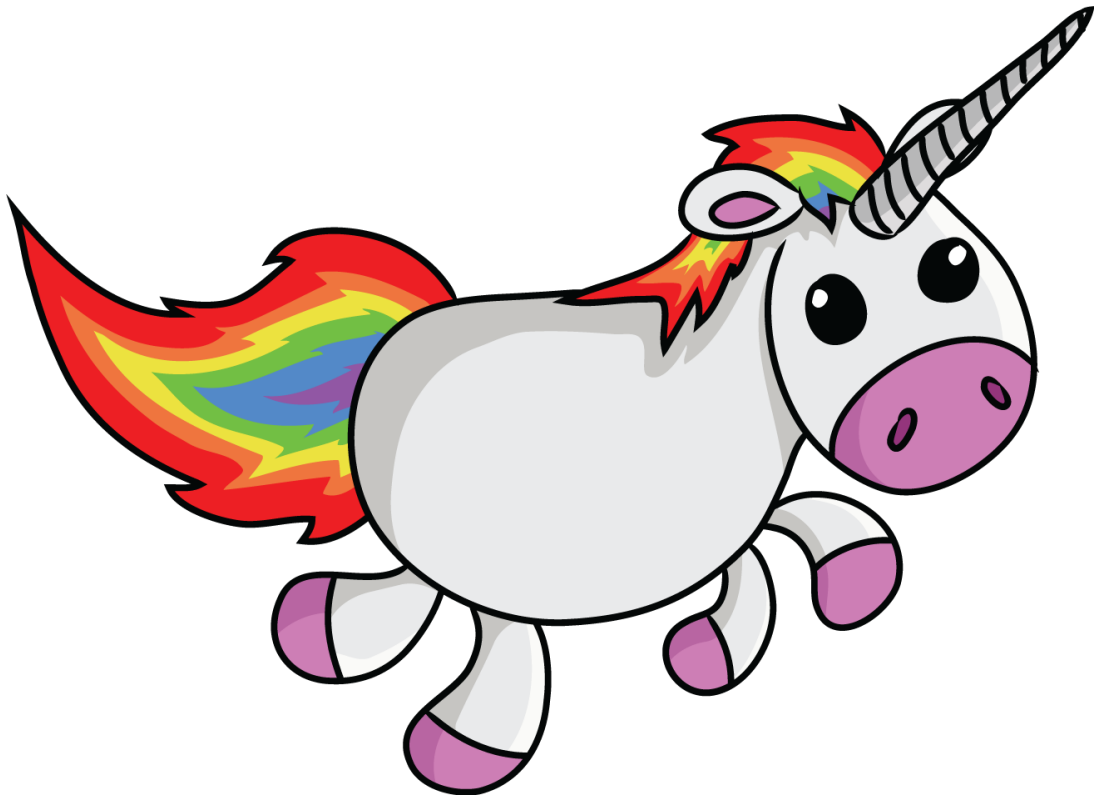
Performs data science

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"
Percentage of Matching Job Postings
0.04
0.02
0
Jan '06
Jan '07
Jan '08
Jan '09
Jan '10
Jan '11
Jan '12
Jan '13
Jan '14
Jan '15
indeed
one search. all jobs.

The Data Science Toolkit

Programming

Data manipulation

Descriptive statistics

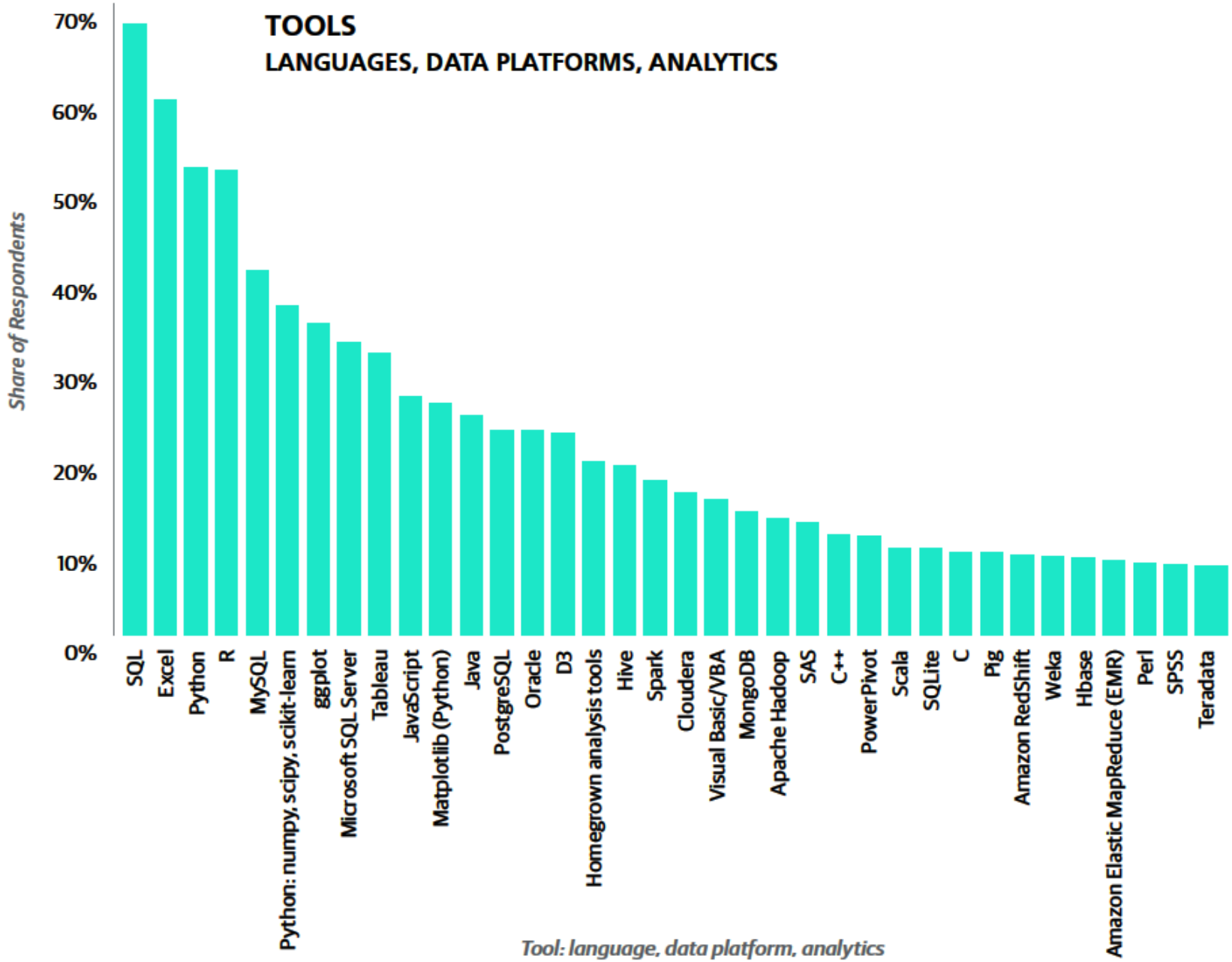
Data visualization

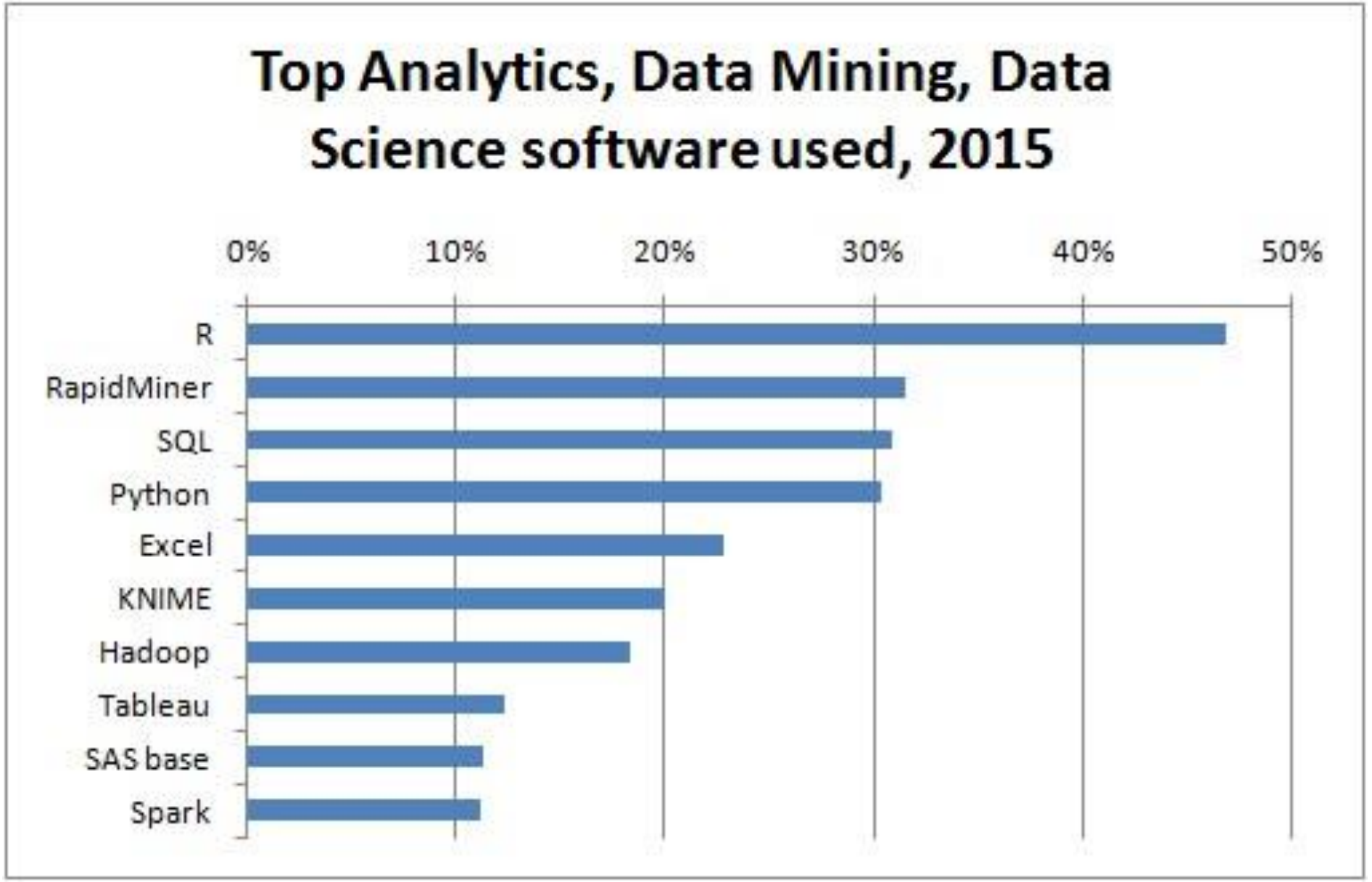
Statistical modeling

Big Data

Machine learning

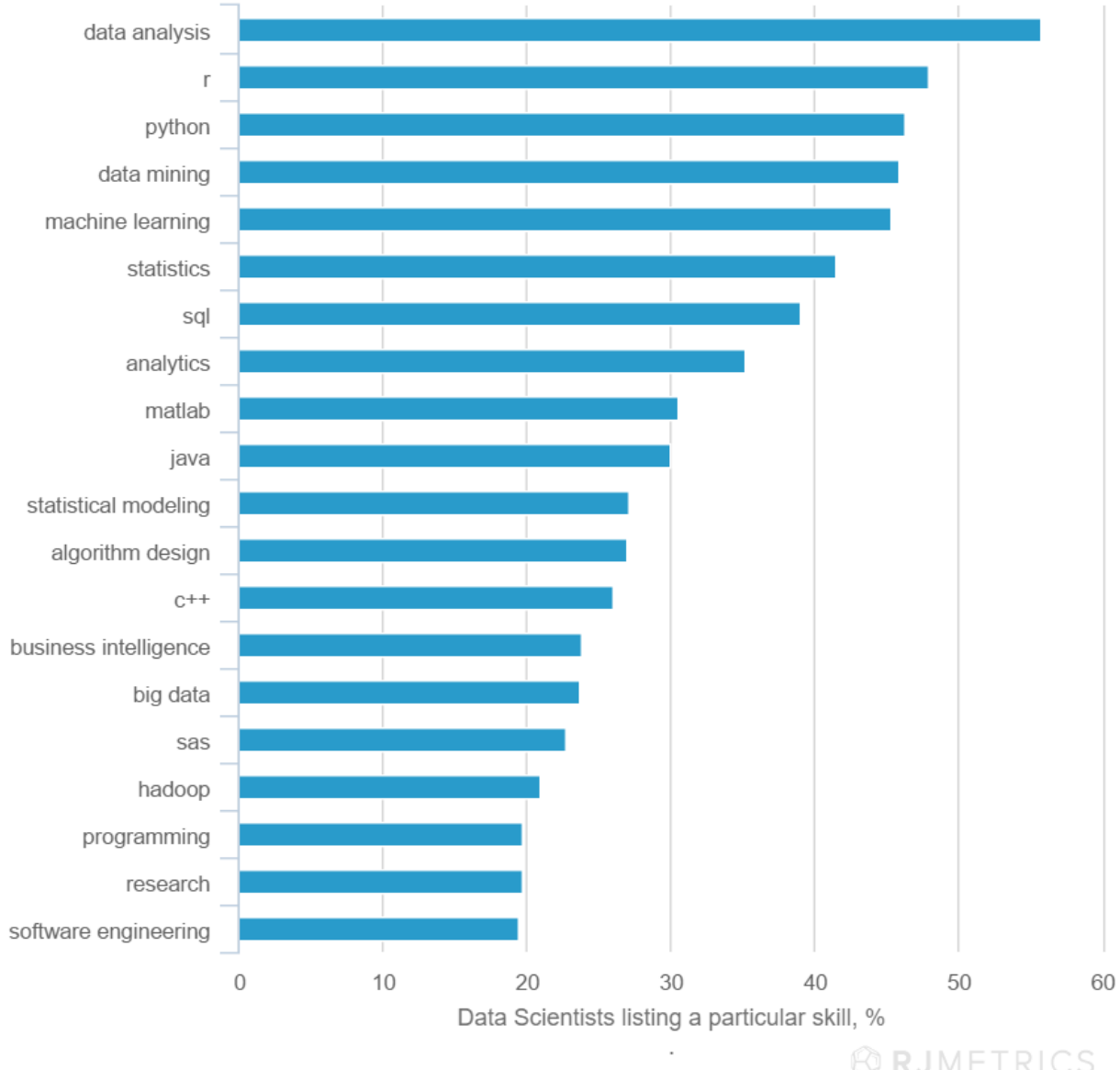
Deploying into production


TOOLS
LANGUAGES, DATA PLATFORMS, ANALYTICS
Share of Respondents
70%
60%
50%
40%
30%
20%
10%
0%
SQL
Excel
Python
R
MySQL
Python: numpy, scipy, scikit-learn
ggplot
Microsoft SQL Server
Tableau
JavaScript
Matplotlib (Python)
Java
PostgreSQL
Oracle
D3
Homegrown analysis tools
Hive
Spark
Cloudera
Visual Basic/VBA
MongoDB
Apache Hadoop
SAS
C++
PowerPivot
Scala
SQLite
C
Pig
Amazon RedShift
Weka
Hbase
Amazon Elastic MapReduce (EMR)
Perl
SPSS
Teradata
Tool: language, data platform, analytics



http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html

TOP 20 SKILLS OF A DATA SCIENTIST



Why is Data Science Important?

Internet of Things

Big Data

Machine learning

Fully autonomous systems

Why is Data Science Important?

Driven by economics

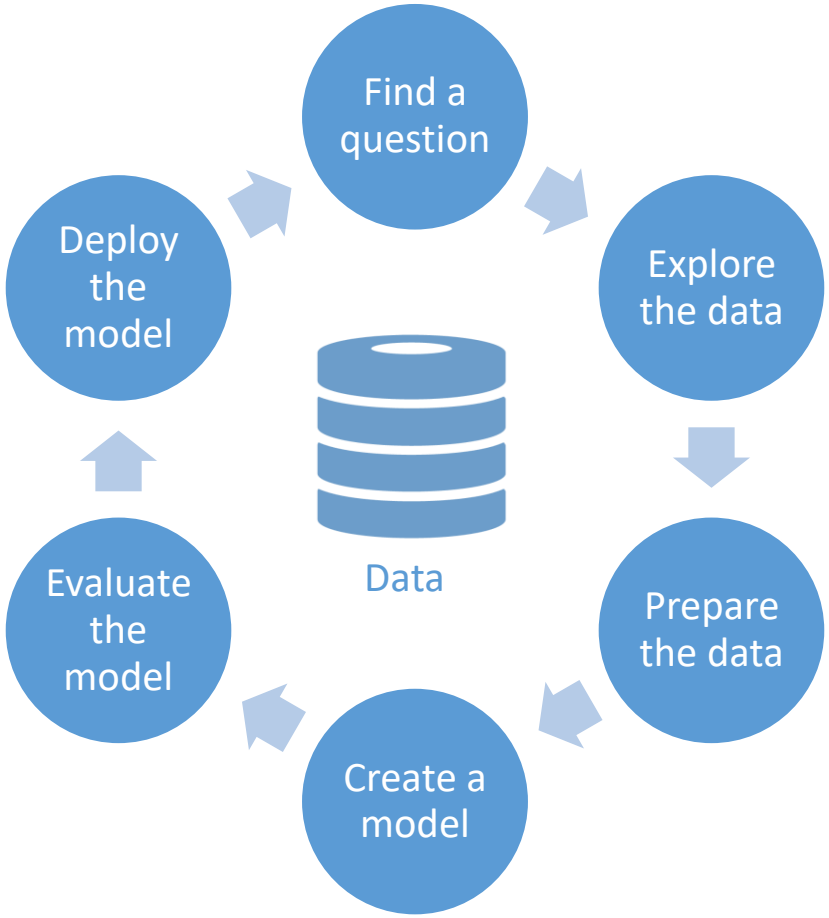
Possible by technology

Cost is decreasing

Value is increasing

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



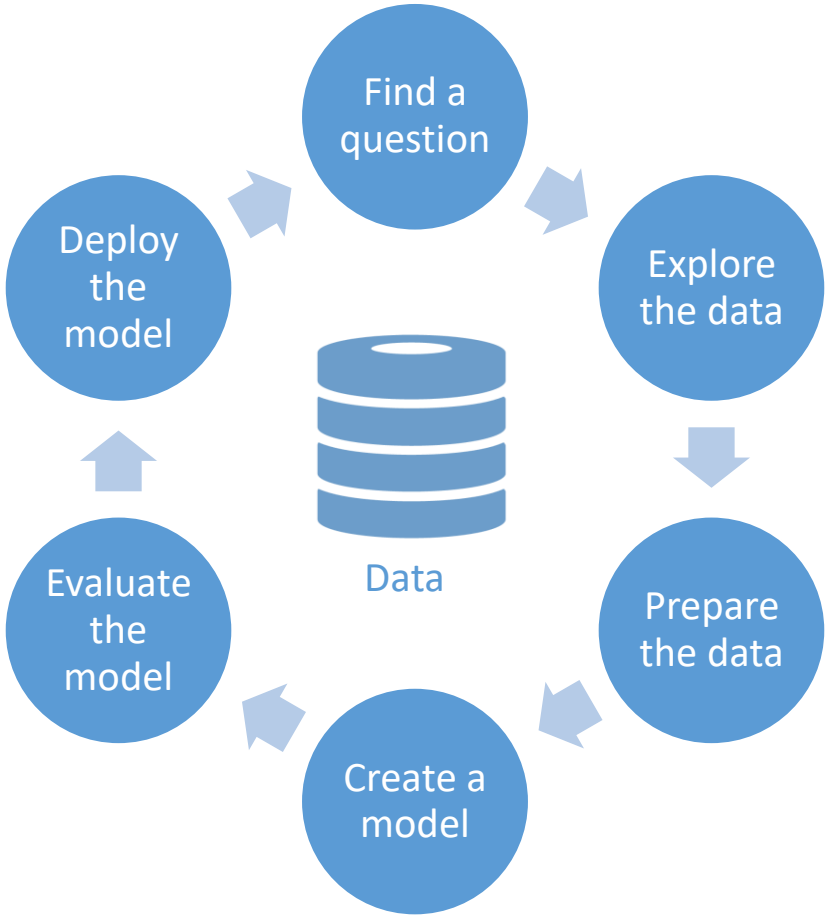
The Data Science Process

Iterative process

Non-sequential

Early termination

Established processes



Introduction to R

What is R?

Open source

Language and environment

Numerical and graphical analysis

Cross platform

What is R?

Active development

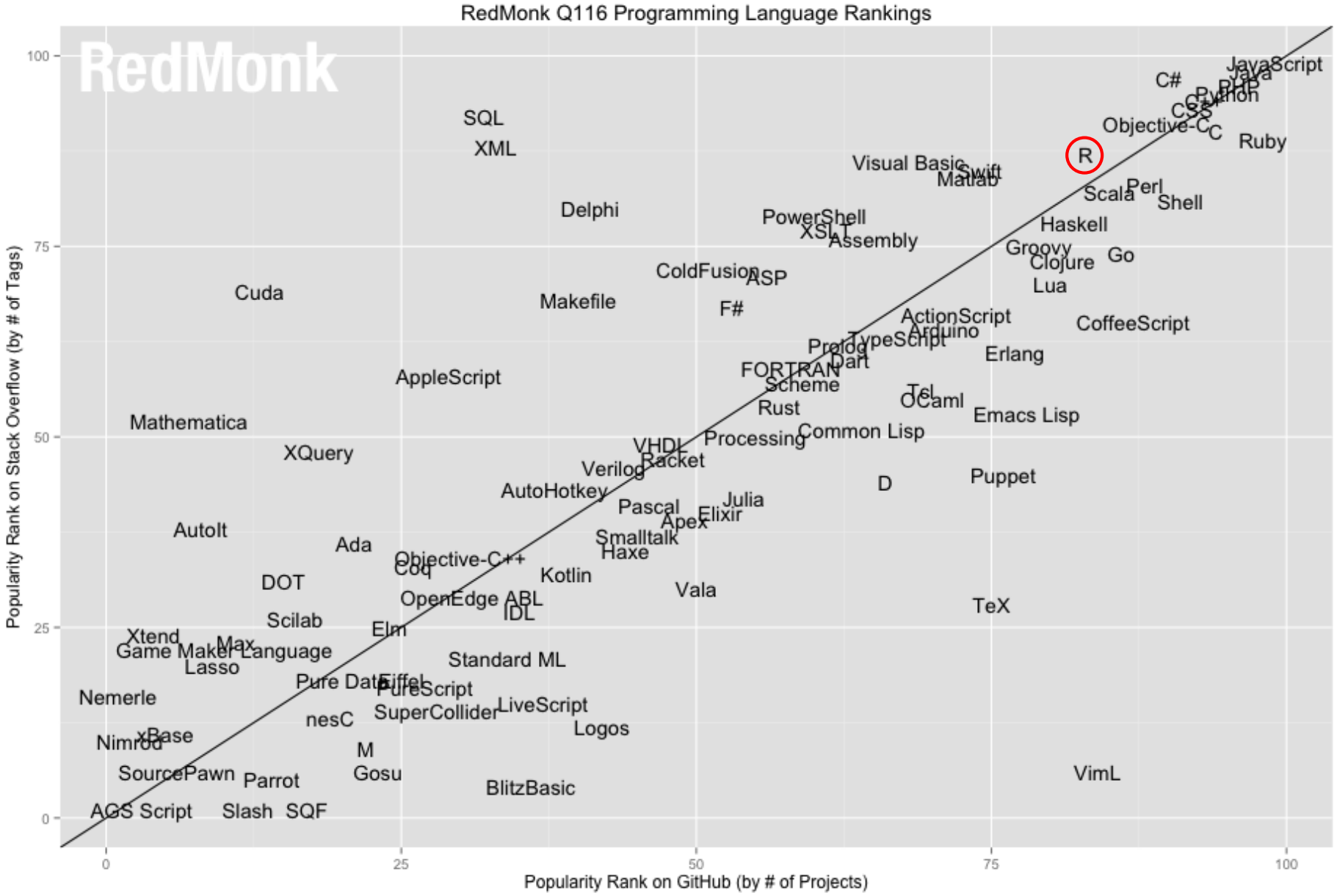
Large user community

Modular and extensible

9000+ extensions


FREE


FREE



Source: http://redmonk.com/sogrady/2016/07/20/language-rankings-6-16/

RGui (64-bit)

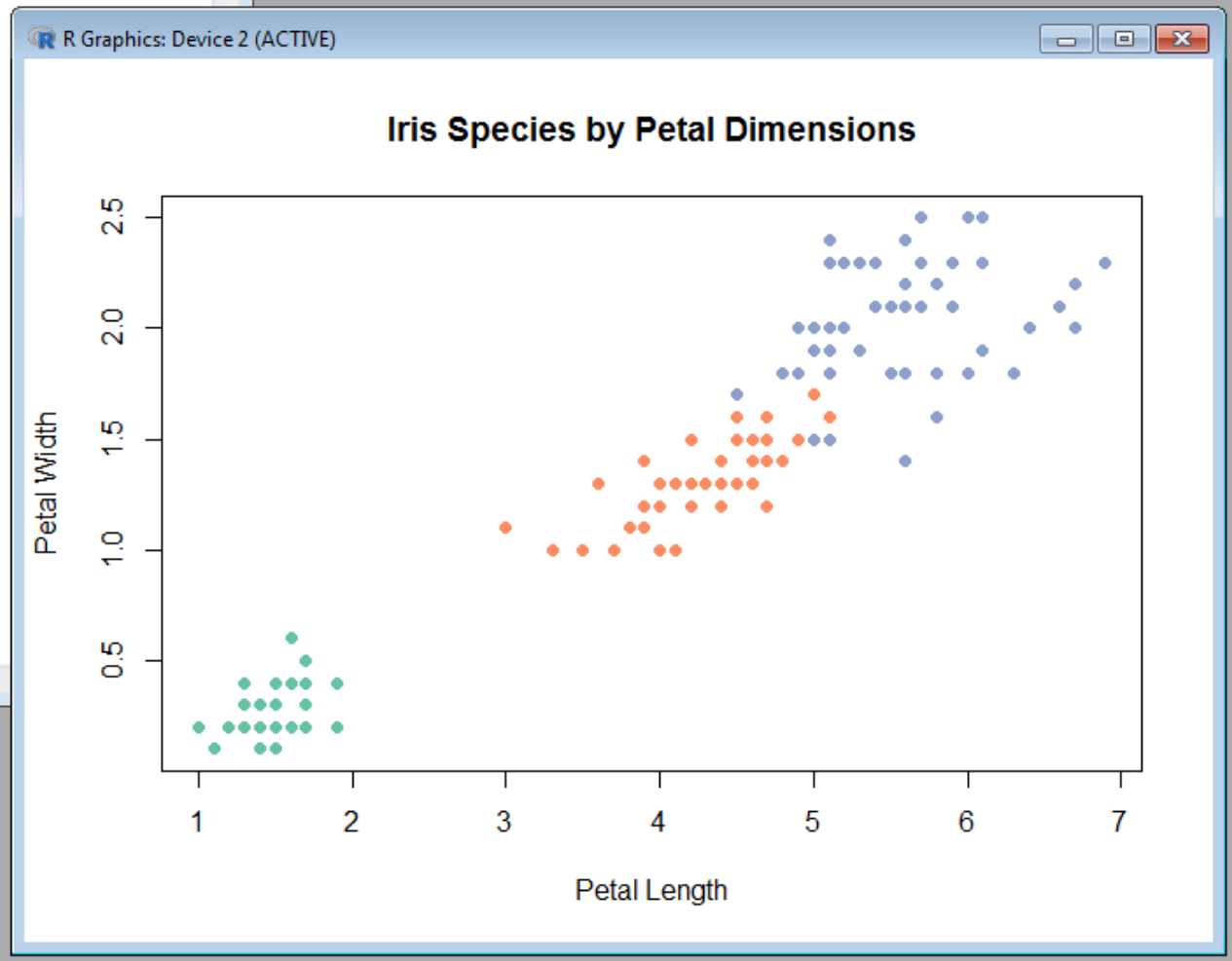
File History Resize Windows

R Console

```
> # Create a plot of species by dimension
> plot(
+     x = iris$Petal.Length,
+     y = iris$Petal.Width,
+     pch = 19,
+     col = palette[as.numeric(iris$Species)],
+     main = "Iris Species by Petal Dimensions",
+     xlab = "Petal Length",
+     ylab = "Petal Width")
>
> # Create a frequency table of species
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
>
> # Get the average petal length
> mean(iris$Petal.Length)
[1] 3.758
>
> # Get the correlation coefficient
> cor(
+     x = iris$Petal.Length,
+     y = iris$Petal.Width)
[1] 0.9628654
```

R Graphics: Device 2 (ACTIVE)



RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function | Addins | Project: (None)

Script.R

Source on Save | Run | Source

```
16 # Create a frequency table of species
17 table(iris$Species)
18
19 # Get the average petal length
20 mean(iris$Petal.Length)
21
22 # Get the correlation coefficient
23 cor(
24     x = iris$Petal.Length,
25     y = iris$Petal.Width)
```

21:1 (Top Level) R Script

Console ~/

```
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
> # Get the average petal length
> mean(iris$Petal.Length)
[1] 3.758
> # Get the correlation coefficient
> cor(
+ x = iris$Petal.Length,
+ y = iris$Petal.Width)
[1] 0.9628654
>
```

Environment History

Import Dataset | List

Global Environment

Data

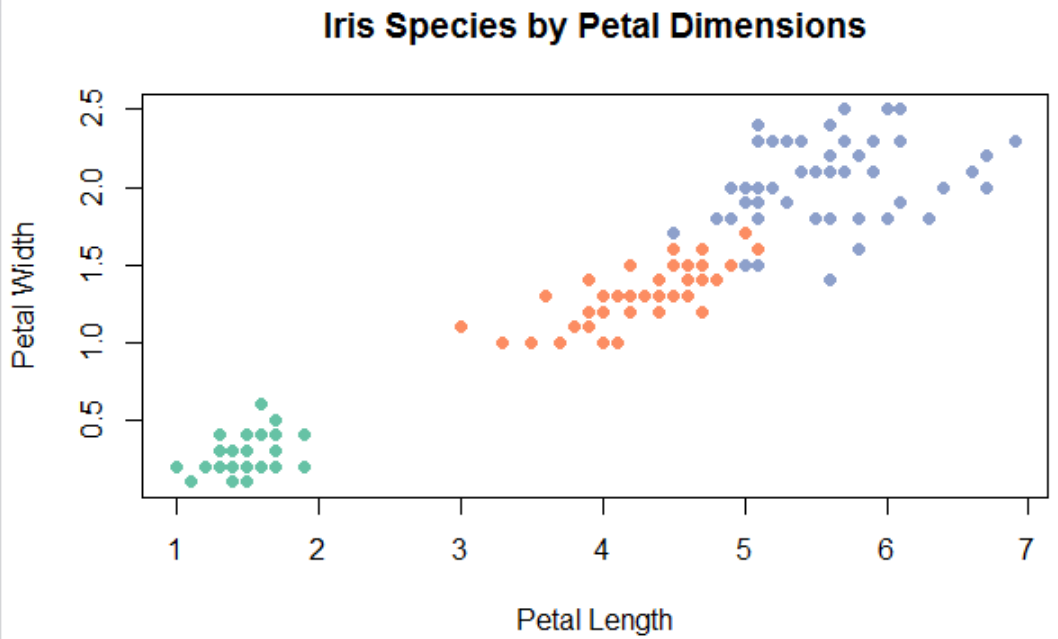
iris	150 obs. of 5 variables

Values

palette	chr [1:3] "#66C2A5" "#FC8D62" "#8DA0C…

Files Plots Packages Help Viewer

Zoom | Export | Publish



Script.R - Microsoft Visual Studio

File Edit View NCrunch Project Debug Team Tools Architecture Test ReSharper R Tools Analyze Window Help

Matthew Renze

Quick Launch (Ctrl+Q)

Attach...

Script.R

```
        main = "Iris Species by Petal Dimensions",
        xlab = "Petal Length",
        ylab = "Petal Width")

# Create a frequency table of species
table(iris$Species)

# Get the average petal length
mean(iris$Petal.Length)

# Get the correlation coefficient
cor(
    x = iris$Petal.Length,
    y = iris$Petal.Width)
```

100 %

R Interactive

Attach Debugger

```
> # Create a frequency table of species
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
> # Get the average petal length
> mean(iris$Petal.Length)
[1] 3.758
> # Get the correlation coefficient
> cor(
+     x = iris$Petal.Length,
+     y = iris$Petal.Width)
[1] 0.9628654
>
```

100 %

Variable Explorer

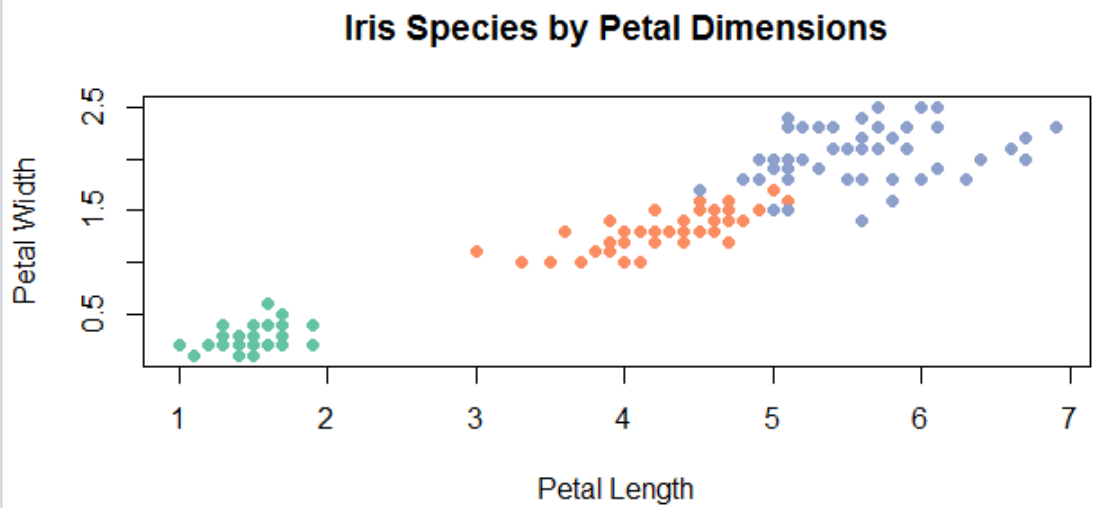
Search

.GlobalEnv

Name	Value	Class	Type
iris	150 obs. of 5 variables	data.frame	list
palette	chr [1:3] "#66C2A5" "#FC8D62" "#8DA0CB"	character	character

Variable Explorer | R History

R Plot



Solution Explorer | R Plot | R Package Manager | R Help

Error List | Output | Azure App Service Activity

Ready | Ln 30 | Col 1 | Ch 1 | INS | 7 | 0 | Root | master

Code Demo

Lab 1
R Programming Basics

Working with Data

Working with Data

Import

Clean

Transform

Export

Loading Data in R

File-based data

Web-based data

Databases

Statistical data

Cleaning Data

Reshape data

Rename columns

Convert data types

Ensure proper encoding

Ensure internal consistency

Handle errors and outliers

Handle missing values

Cleaning Data

base

tidyr

reshape2

stringr

lubridate

validate

Transforming Data

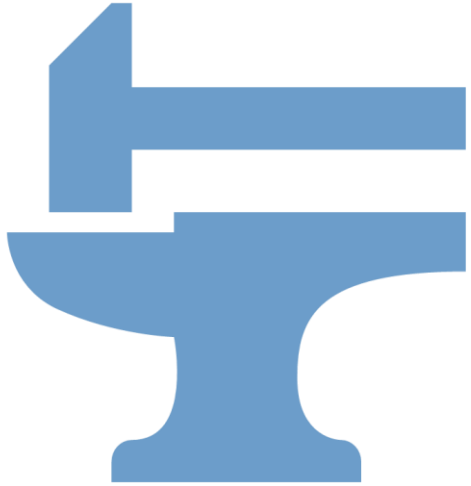
Select columns

Select rows

Group rows

Order rows

Merging data sets



Transforming Data

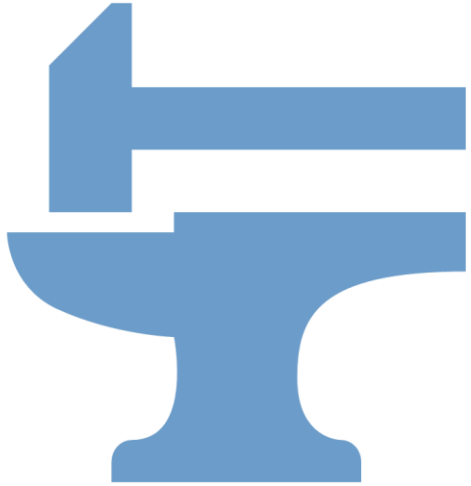
base

plyr

dplyr

data.table

sqldf



Exporting Data

File-based data

Web-based data

Databases

Statistical data

Using dplyr

Select

Filter

Mutate

Summarize

Arrange

Joins

Advice for Working with Data

Data munging is difficult
and time consuming

TIP: Record all steps

Open Movies Database

Movies						
Title	Year	Rating	Runtime (minutes)	Genre	Critic Score	Box Office
The Whole Nine Yards	2000	R	98	Comedy	45%	$57.3M
Cirque du Soleil	2000	G	39	Family	45%	$13.4M
Gladiator	2000	R	155	Action	76%	$187.3M
Dinosaur	2000	PG	82	Family	65%	$135.6M
Big Momma's House	2000	PG-13	99	Comedy	30%	$0.5M


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
1st Quartile	93
Median	101
Mean	104
3rd Quartile	113
Maximum	219

Statistical Terms

Observations

Variables

Categorical variables

Numeric variables

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

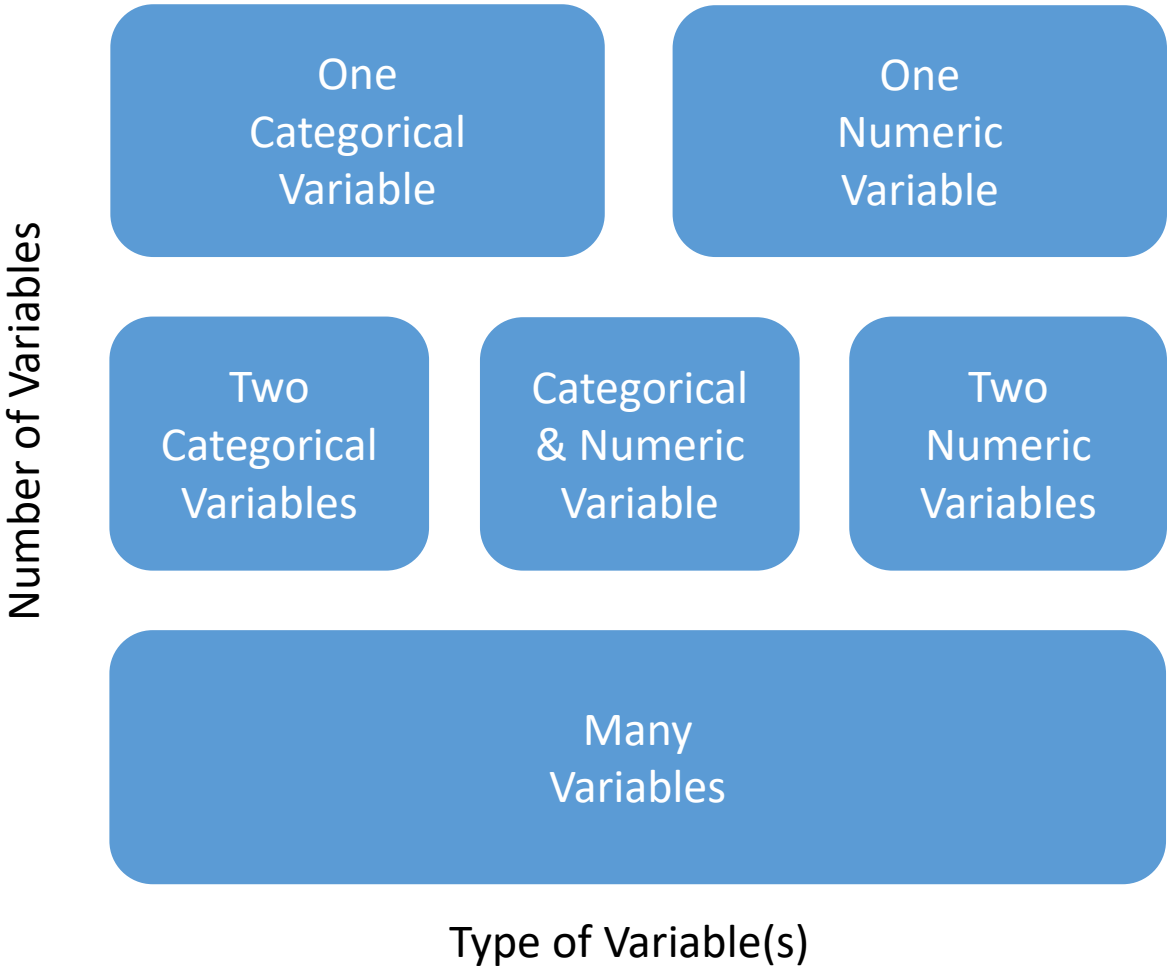
Types of Analysis

Number of variables

- 1 - Univariate
- 2 - Bivariate
- n - Multivariate

Type of variables

- Categorical
- Numeric



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%
…	…	…

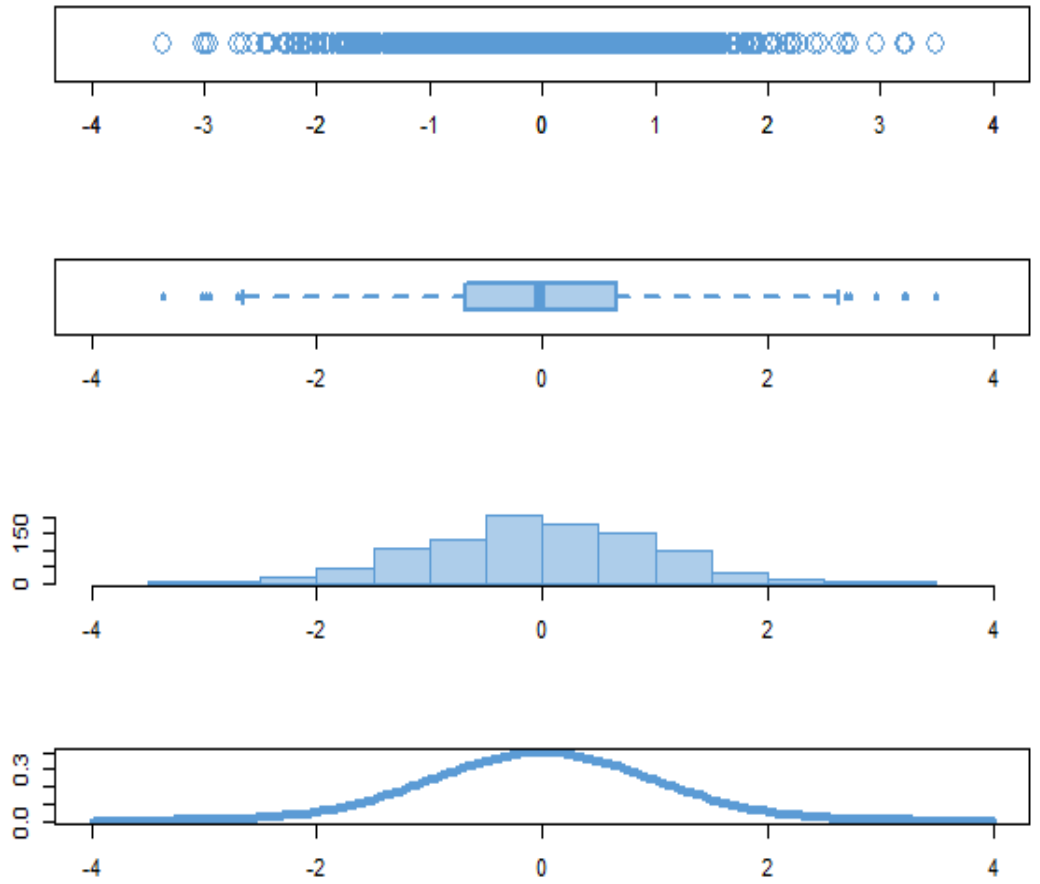
Analyzing One Numeric Variable

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

Analyzing 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

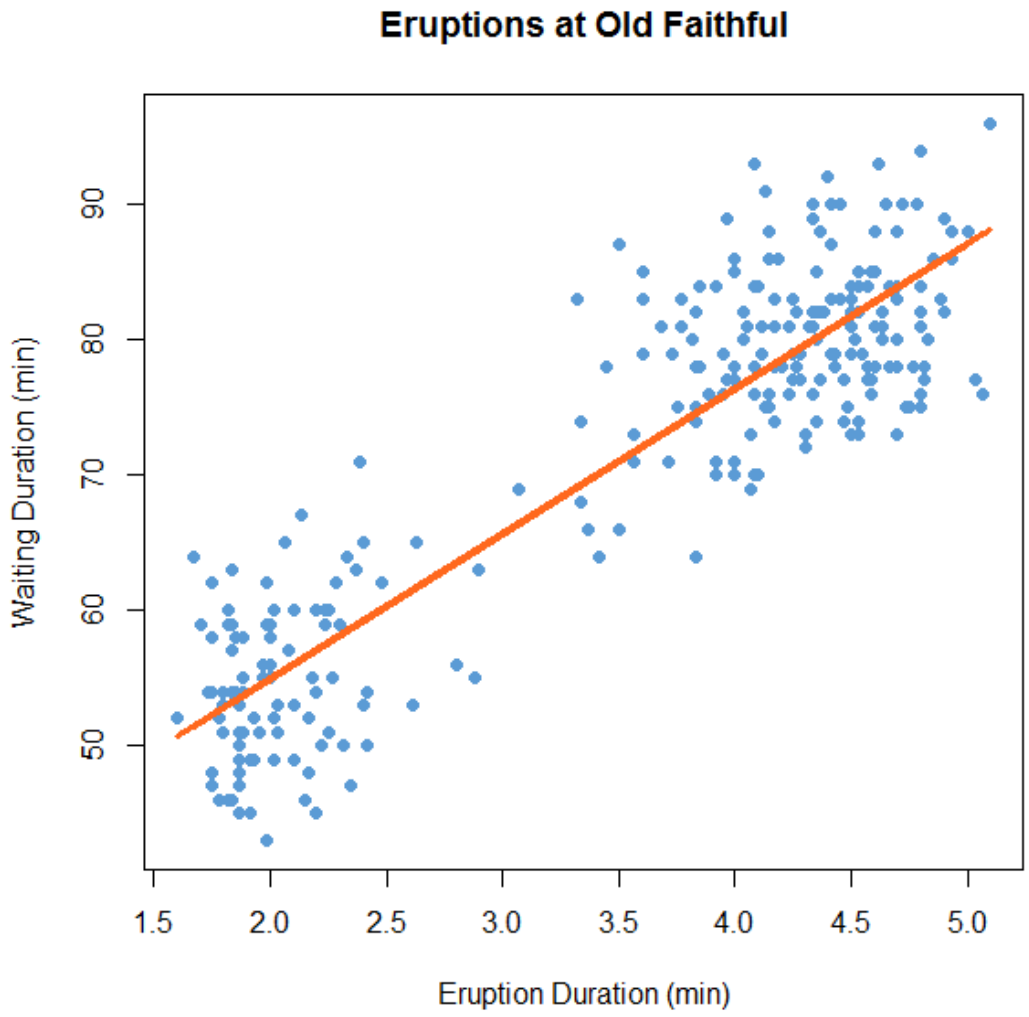
Analyzing Two Numeric Variables

Quantitative bivariate analysis

Explanitory vs. outcome

Covariance

Correlation

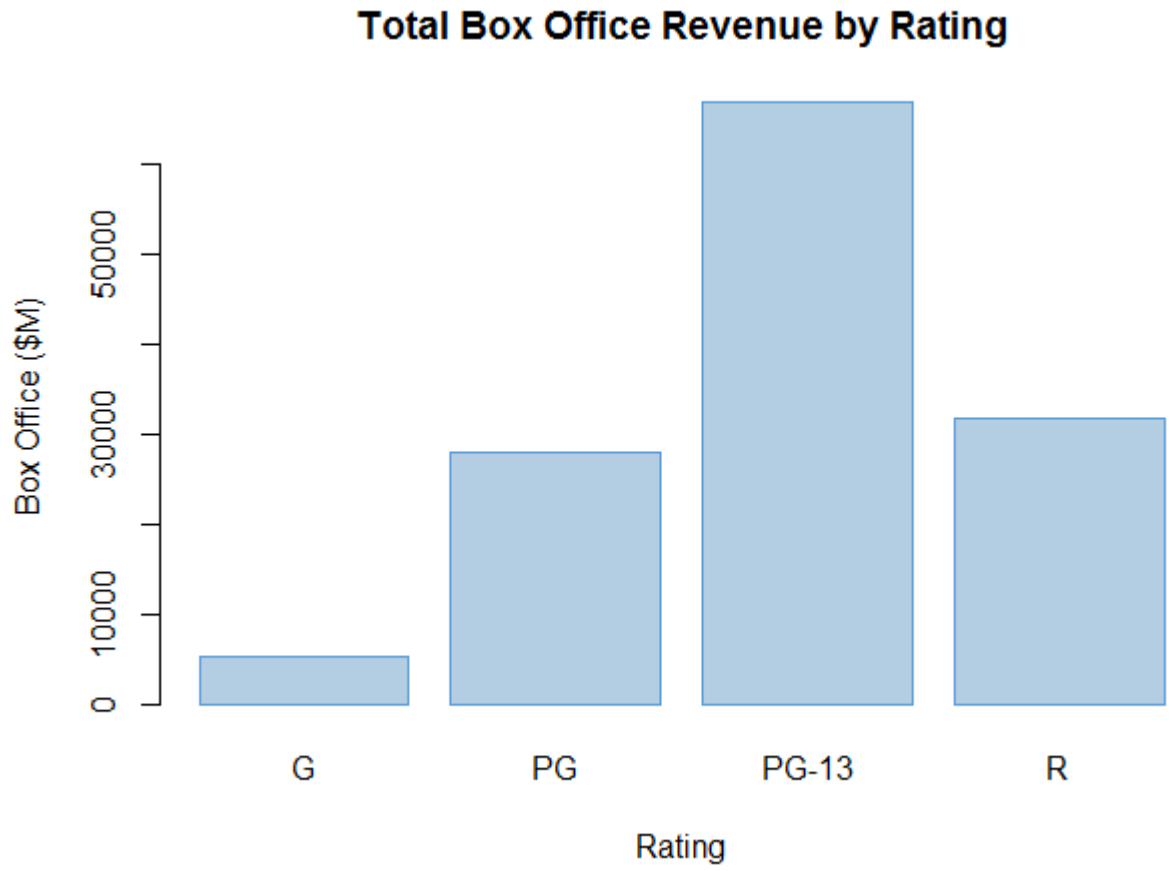


Analyzing a Numeric Variable Grouped by a Categorical Variable

One categorical variable

One numeric variable

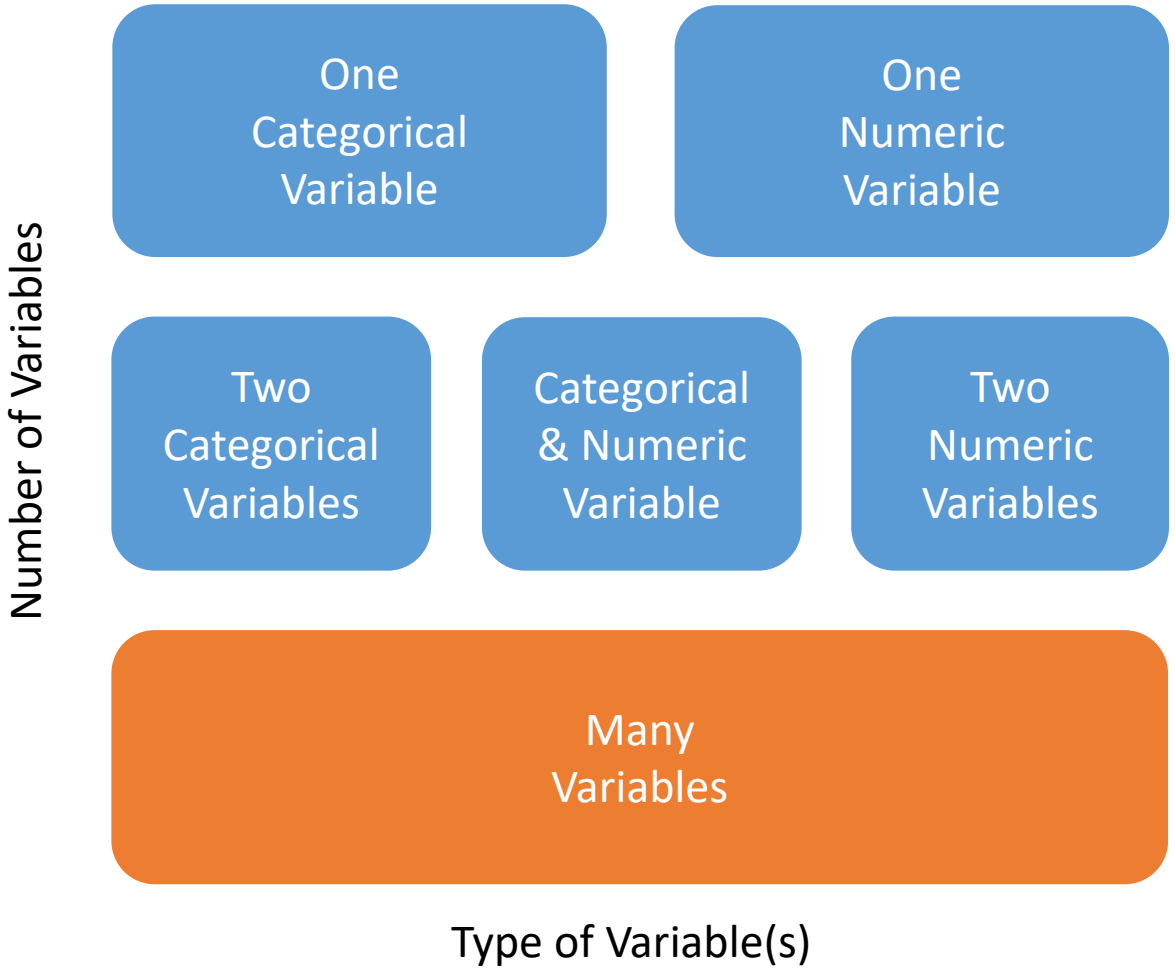
Aggregate measures



Analyzing Many Variables

Multivariate analysis

Specific meaning in statistics




COWBOYS &
Space Invaders:
The Musical
Extended Edition
R

Code Demo

Lab 3
Descriptive Statistics


COWBOYS &
Space Invaders:
The Musical
Extended Edition
R


COWBOYS &
Extended Edition
R

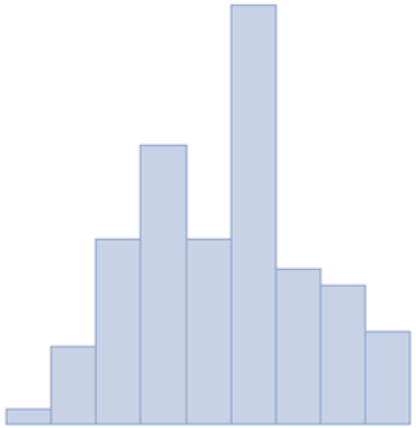
Data Visualization

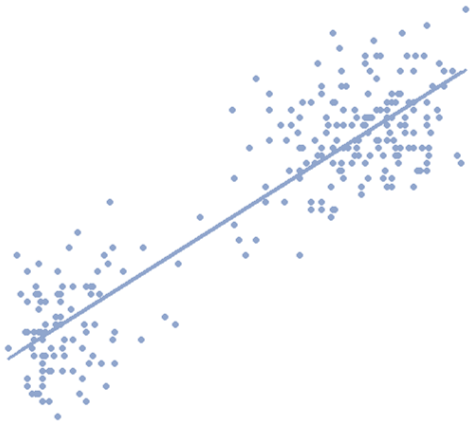
Data Visualization

Visual data representation

For human pattern recognition

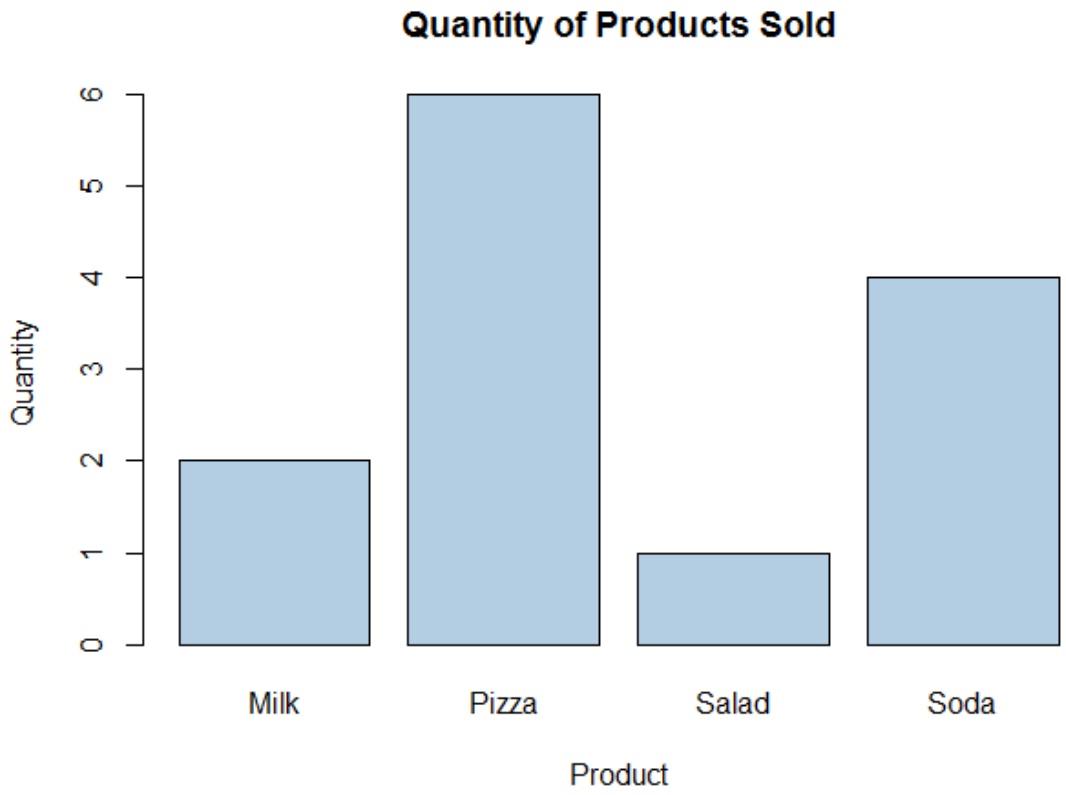
Map dimensions to visual





Data Visualization

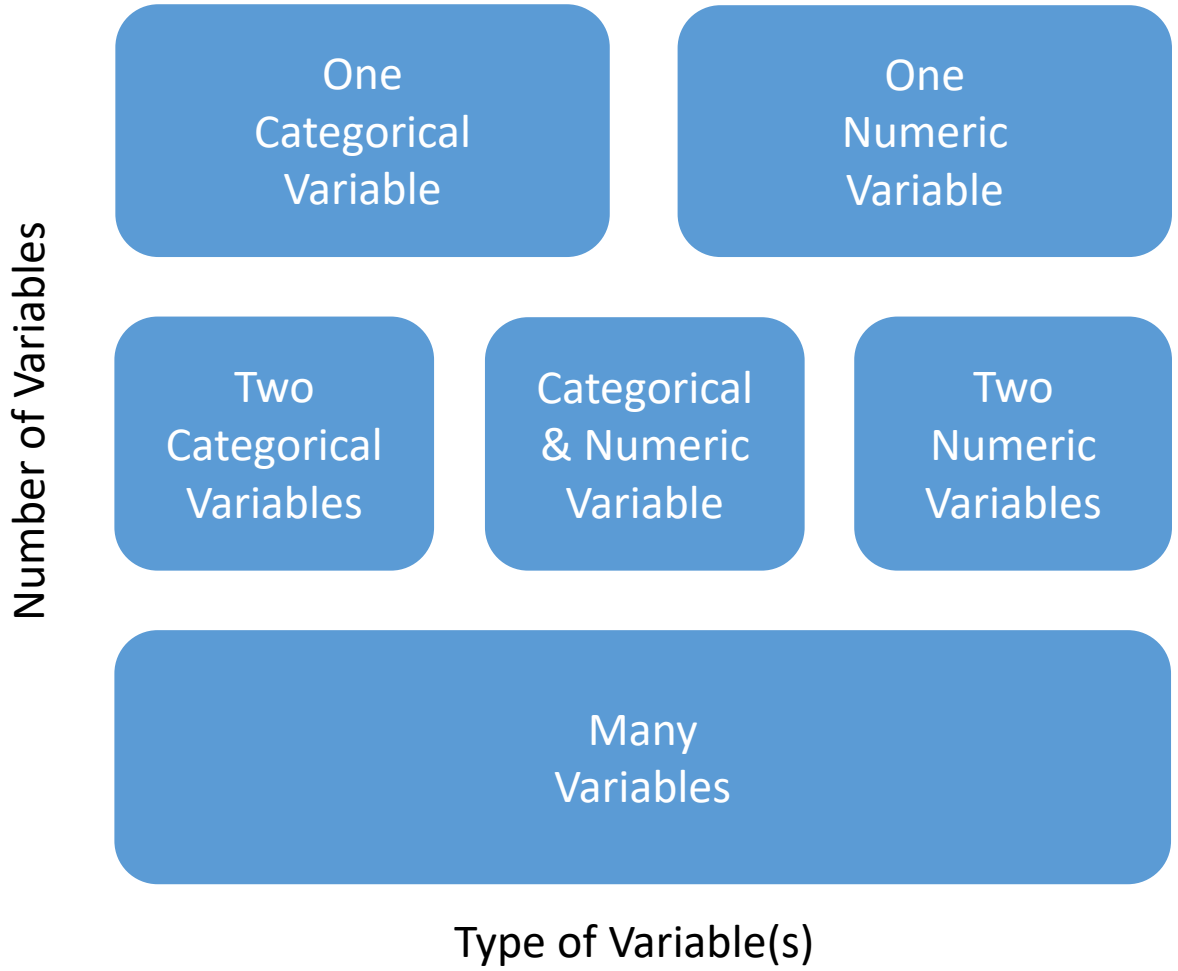
ID	Date	Customer	Product	Quantity
1	2015-08-27	John	Pizza	2
2	2015-08-27	John	Soda	2
3	2015-08-27	Jill	Salad	1
4	2015-08-27	Jill	Milk	1
5	2015-08-28	Miko	Pizza	3
6	2015-08-28	Miko	Soda	2
7	2015-08-28	Sam	Pizza	1
8	2015-08-28	Sam	Milk	1



Types of Data Visualizations

Number of variables

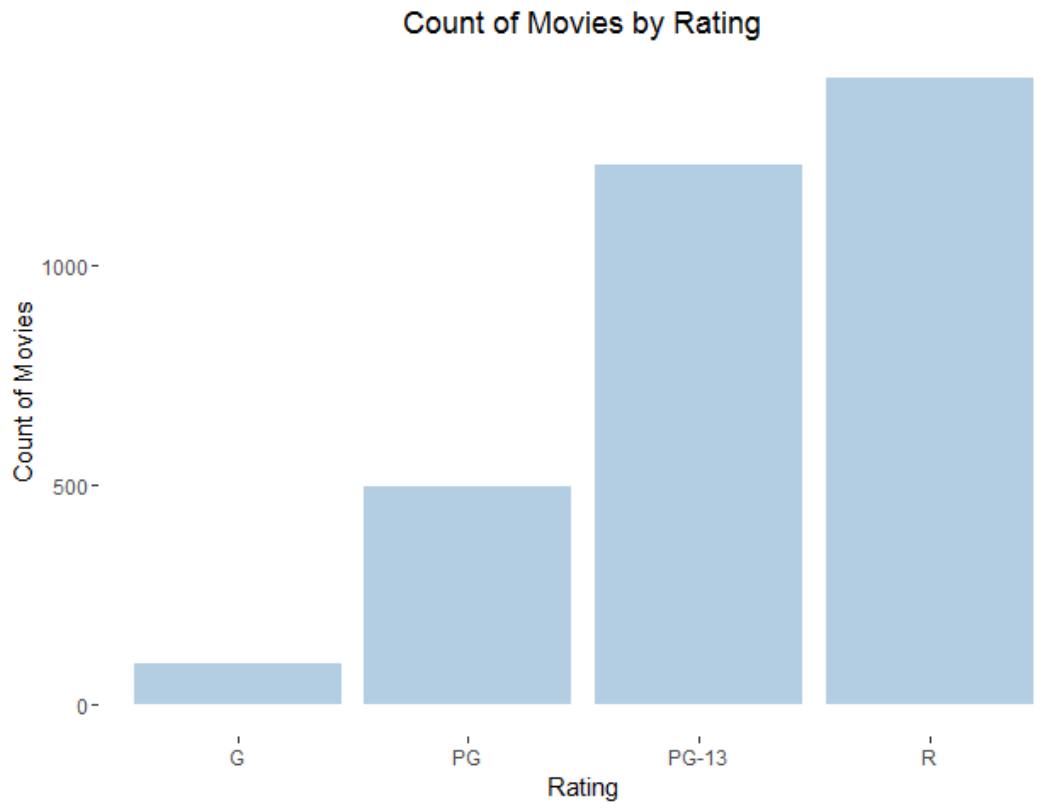
Type of variable(s)



Visualizing One Categorical Variable

Frequency

Proportion

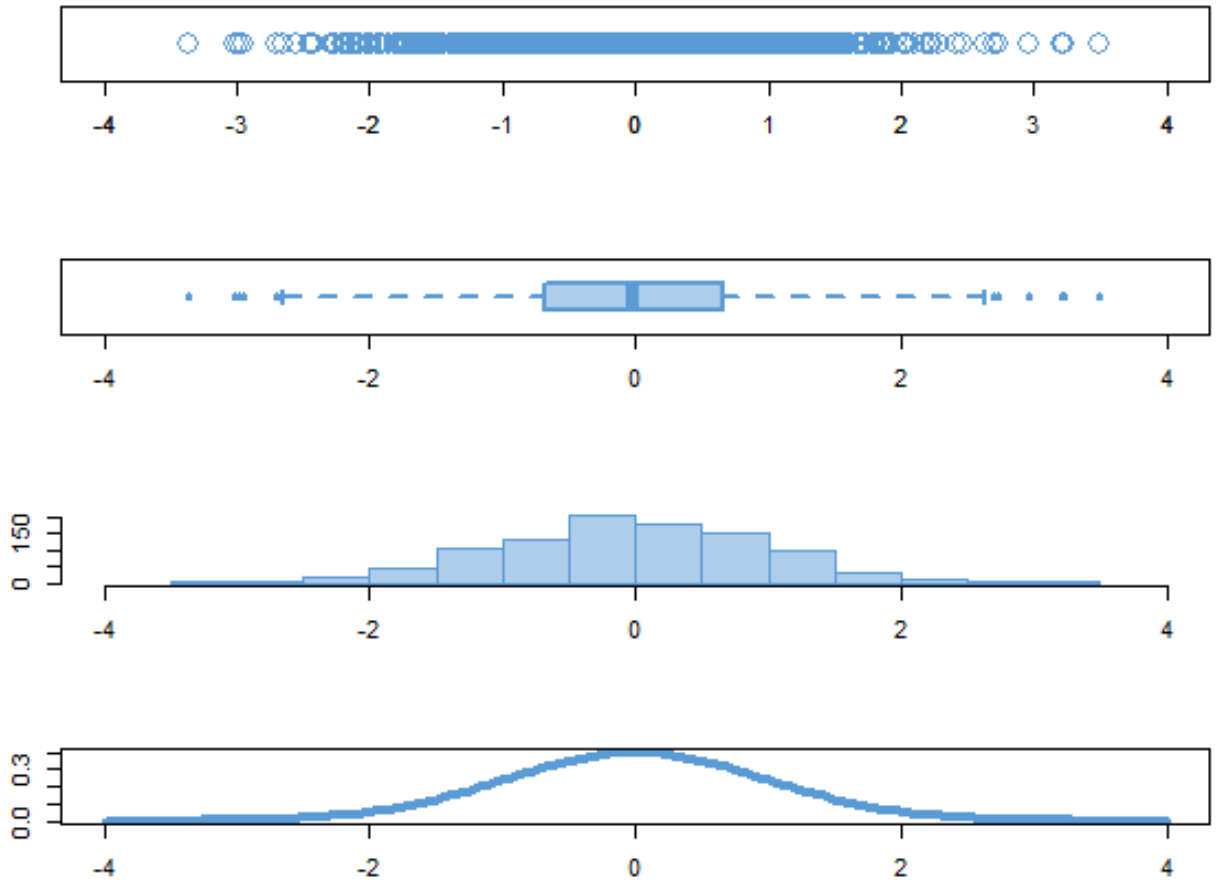


Visualizing One Numeric Variable

Location

Spread

Shape

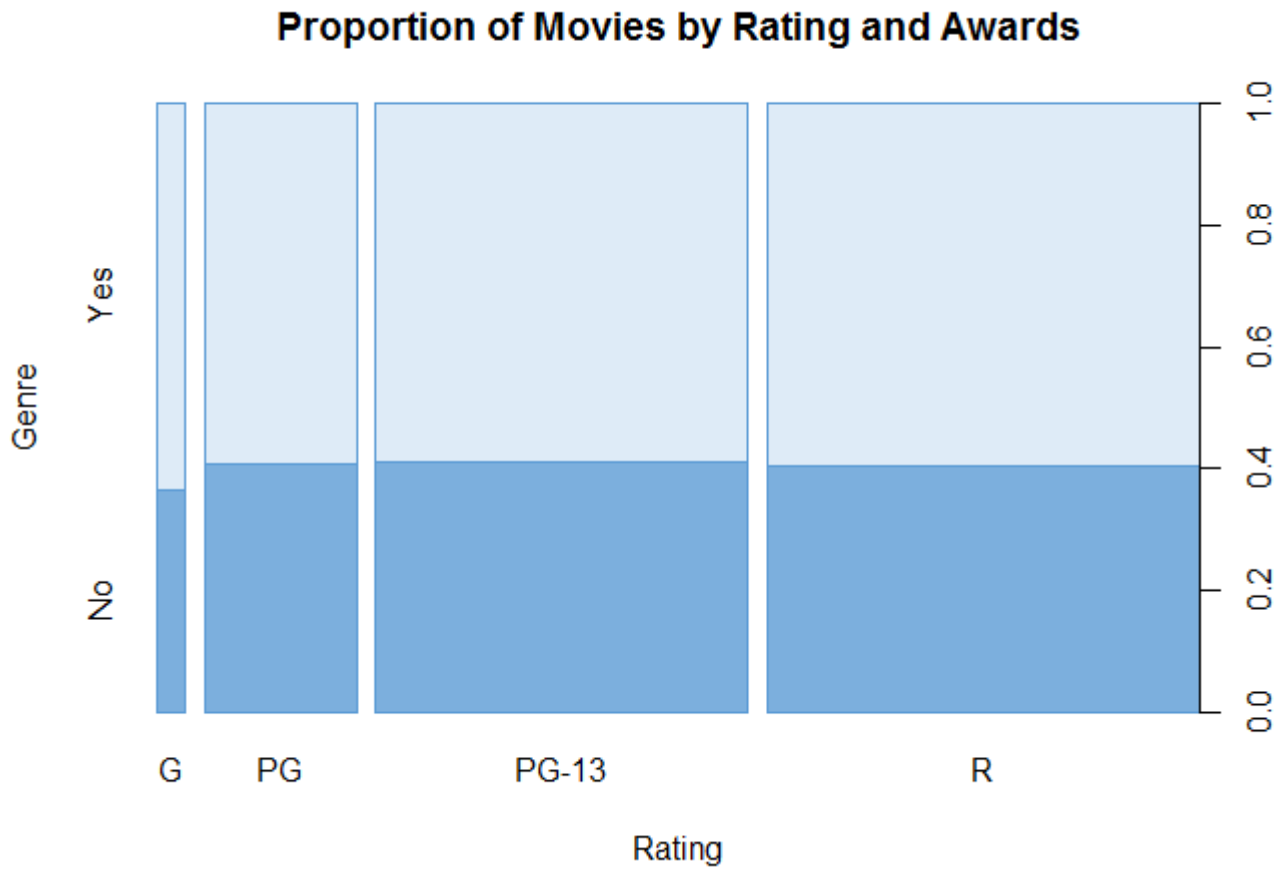


Visualizing Two Categorical Variables

Joint frequency

Marginal frequency

Relative frequency

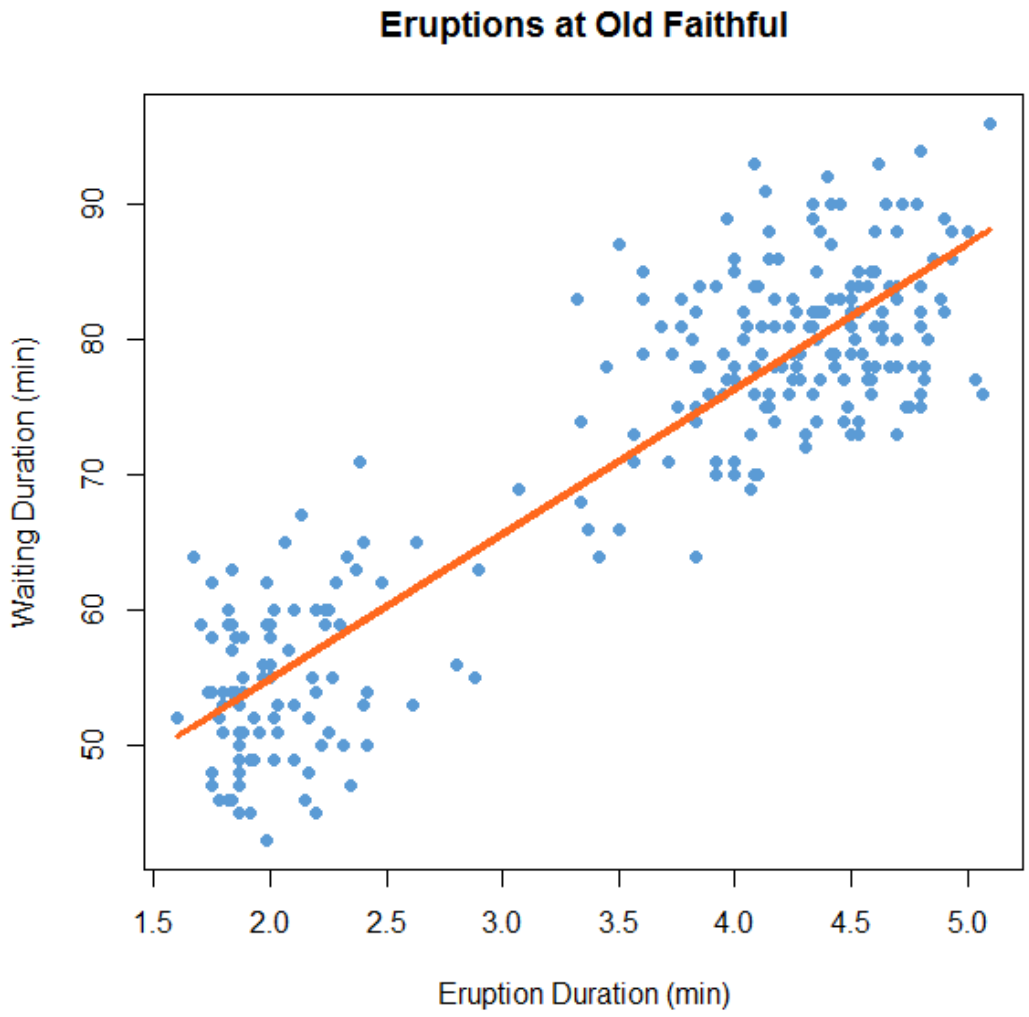


Visualizing Two Numeric Variables

Relationship

Correlation

Distribution

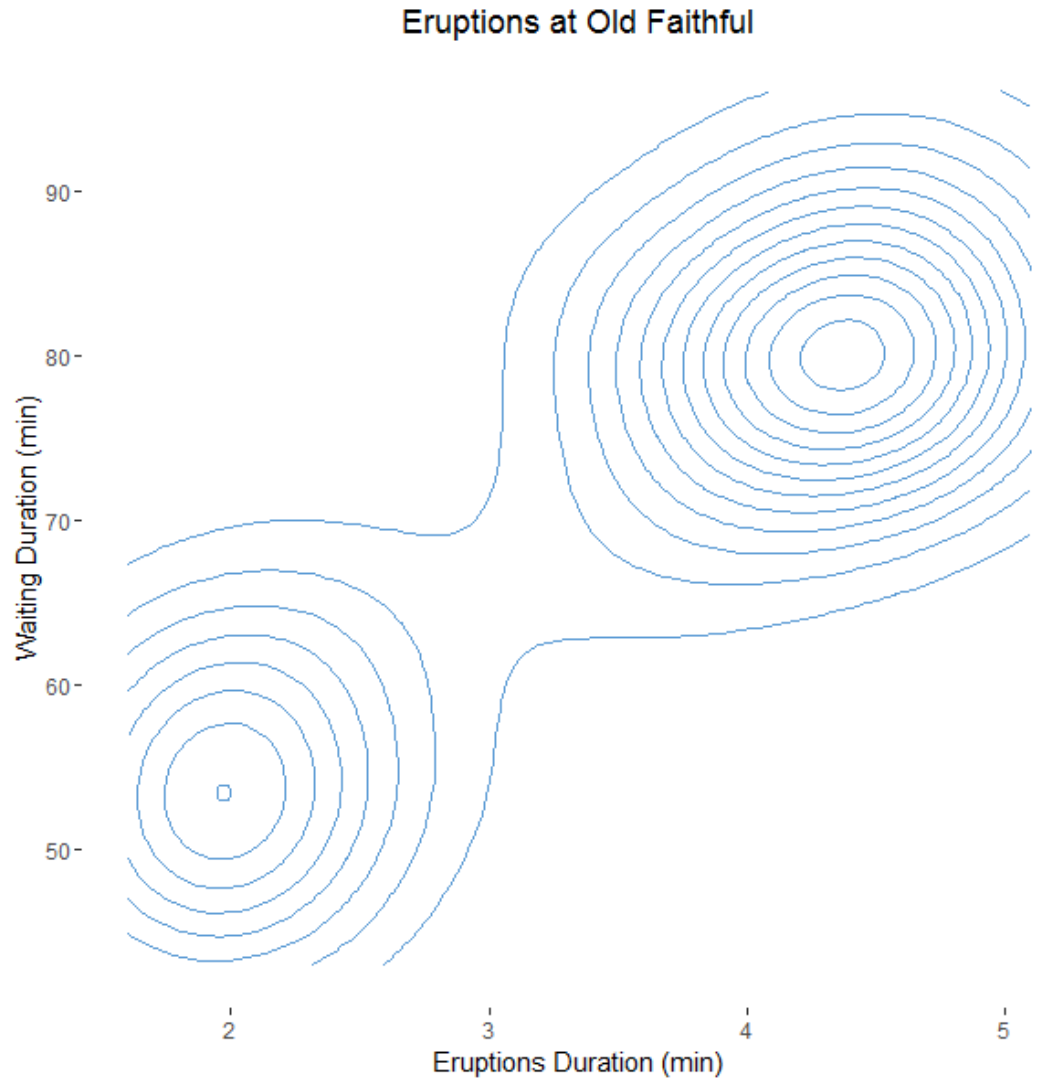


Visualizing Two Numeric Variables

Relationship

Correlation

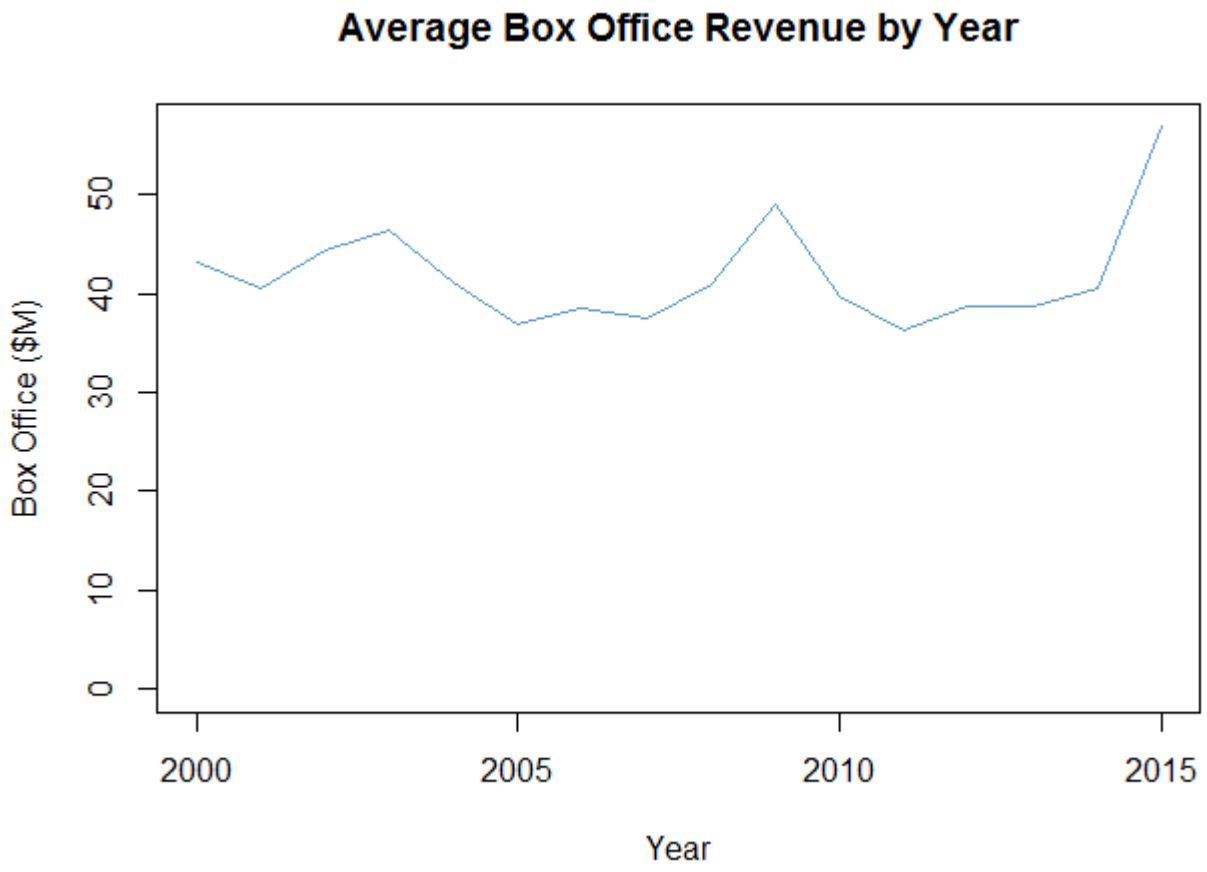
Distribution



Visualizing Time Series Data

Values over time

Rate of change

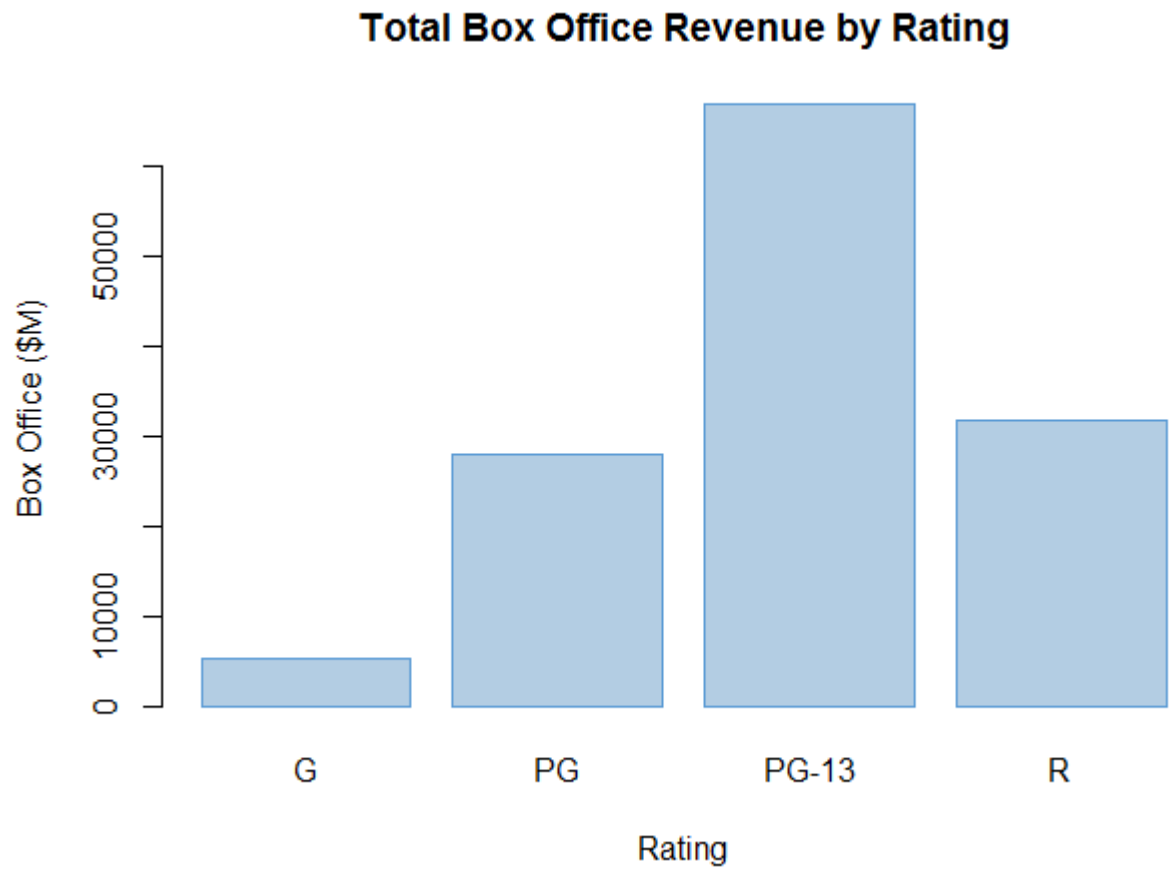


Visualizing a Numeric Variable Grouped by a Categorical Variable

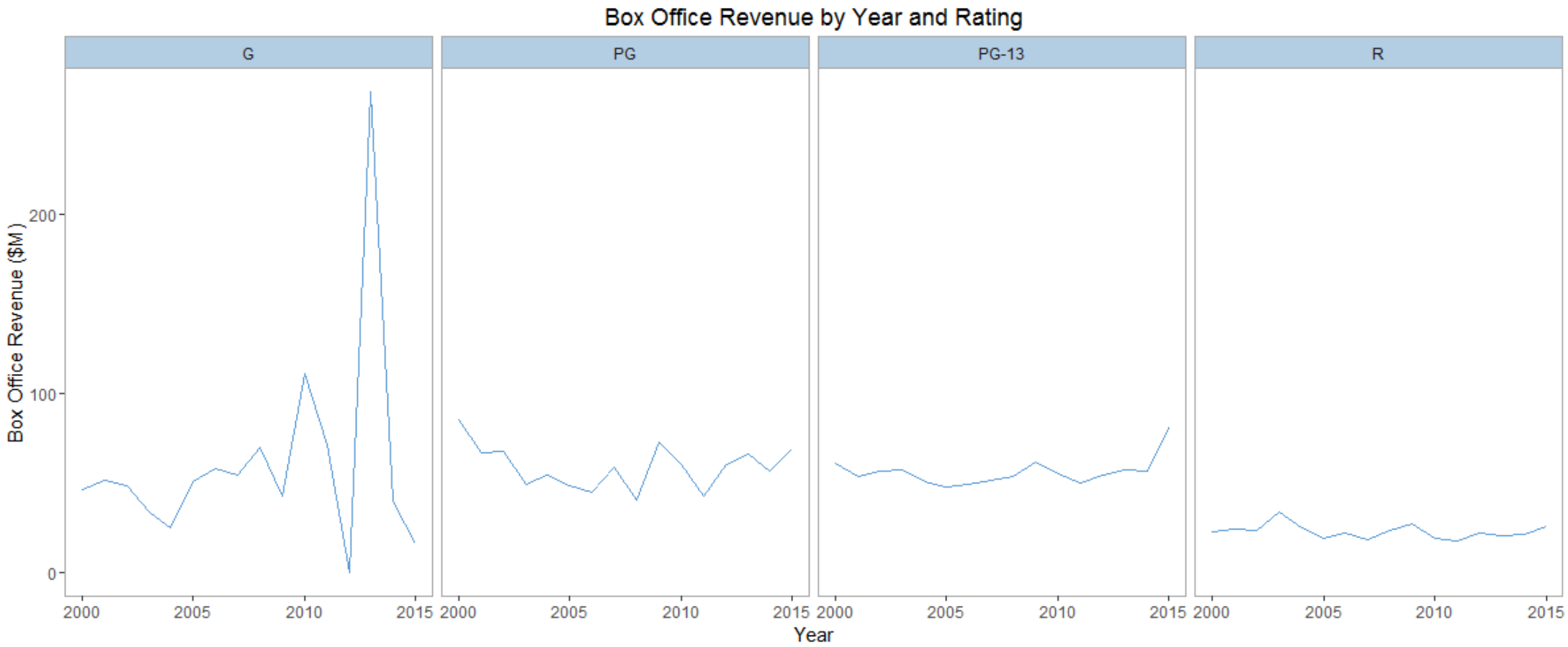
Aggregate

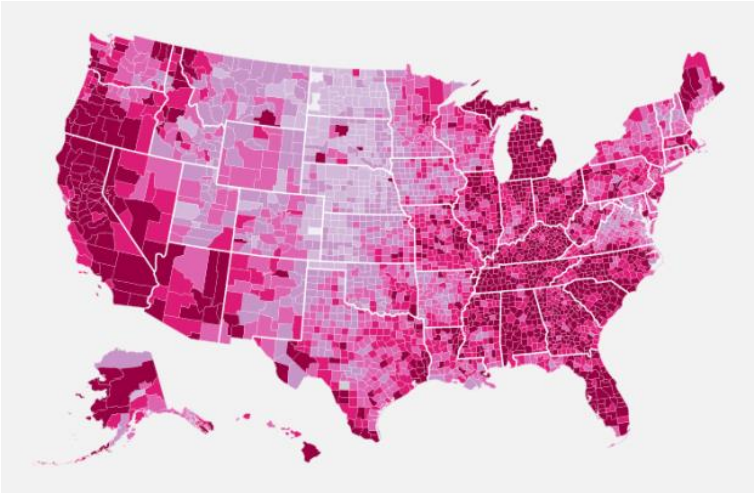
Grouped

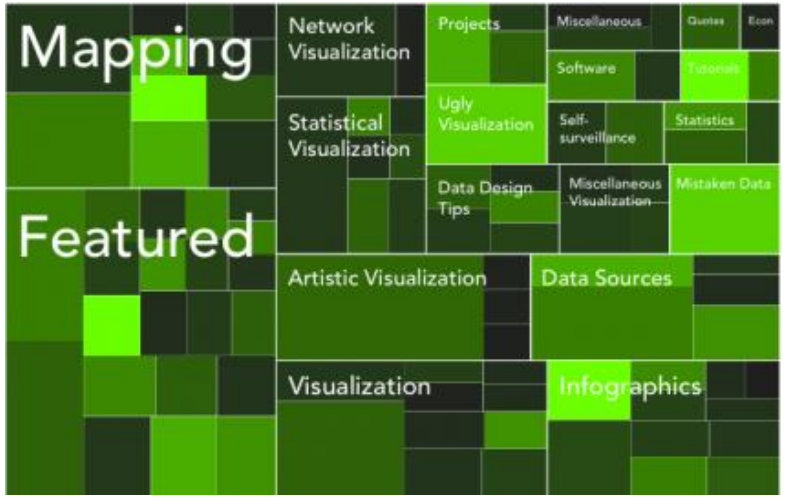
Comparison

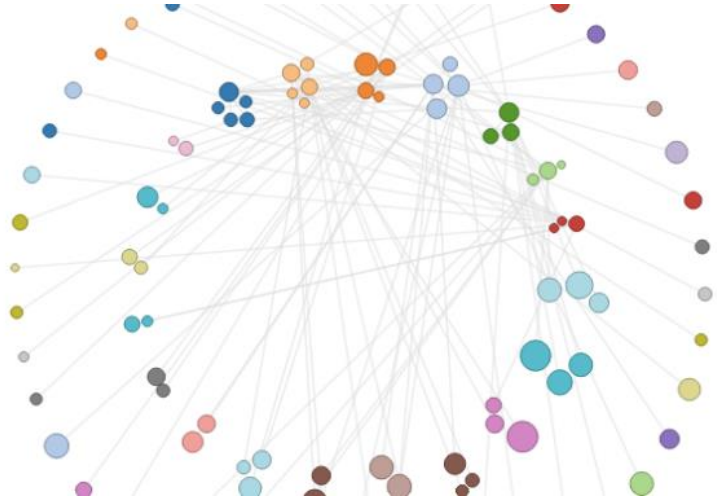


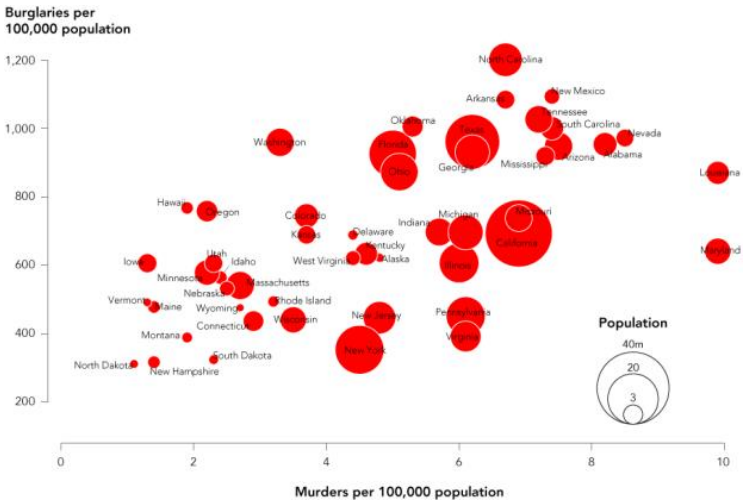
Visualizing Many Variables

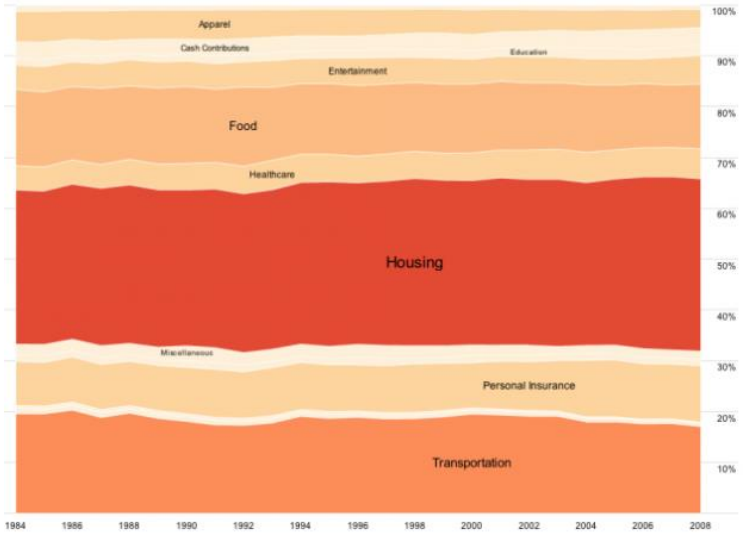


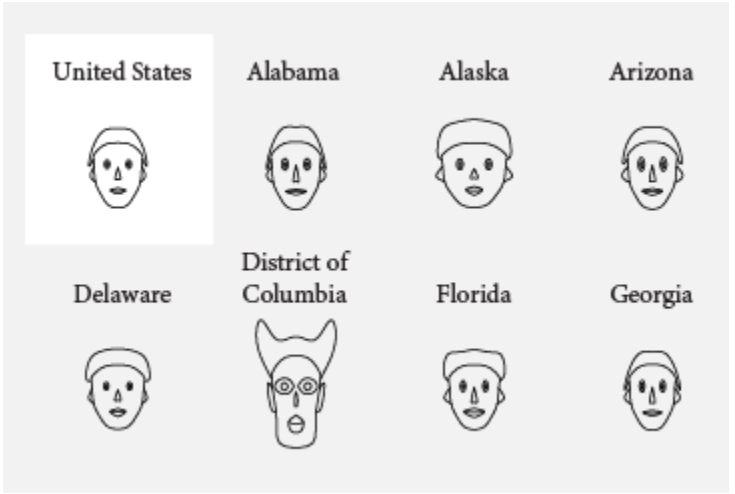




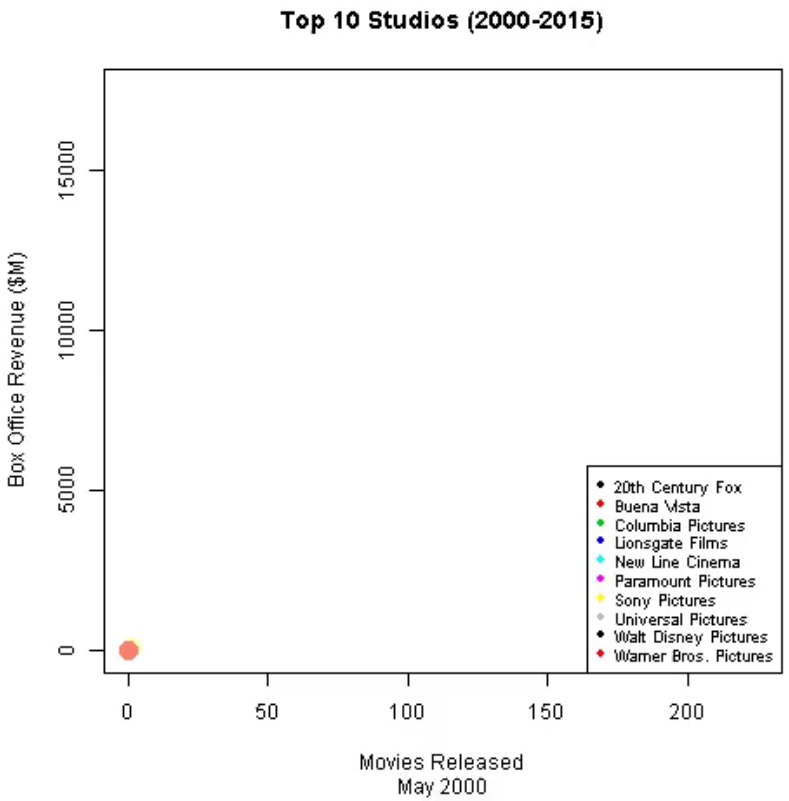


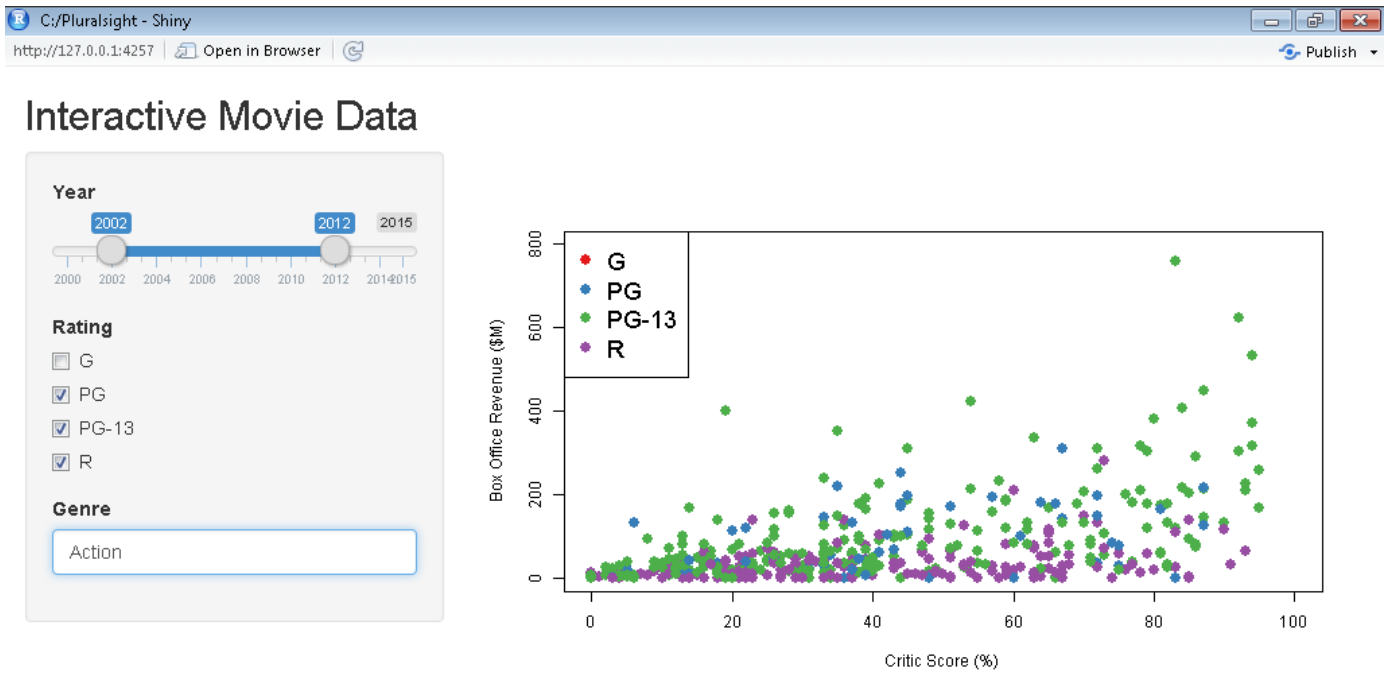




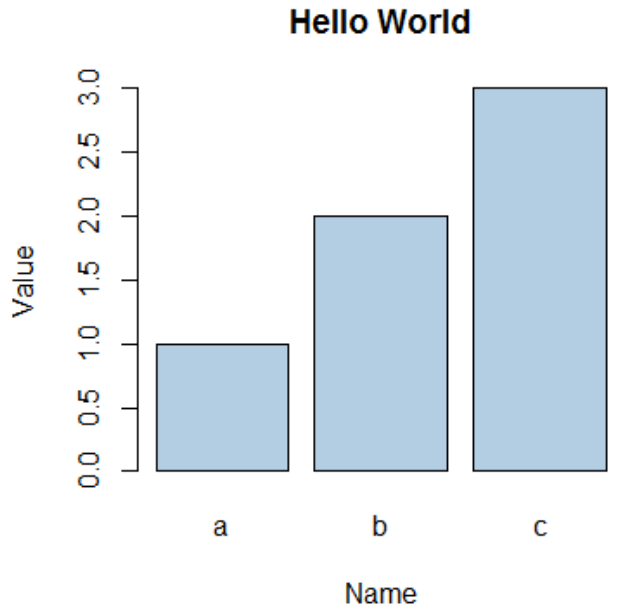


Source: Nathan Yau (www.flowingdata.com)

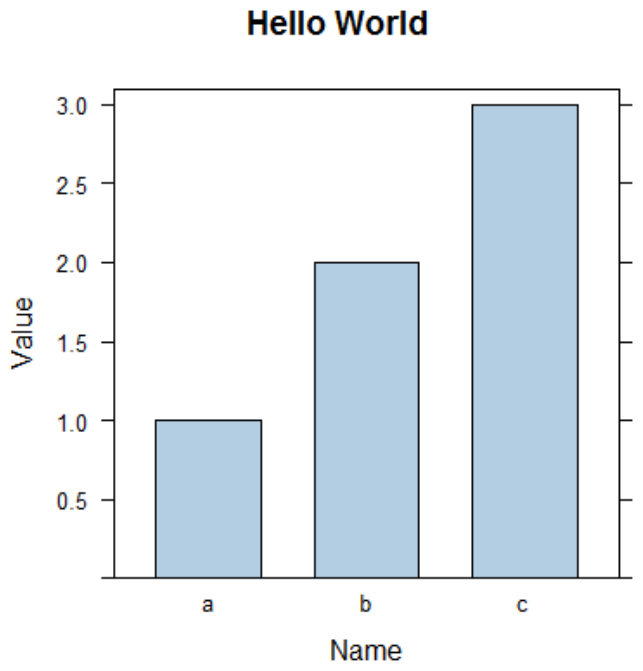

Top 10 Studios (2000-2015)
Box Office Revenue ($M)
0
5000
10000
15000
0
50
100
150
200
Movies Released
May 2000
20th Century Fox
Buena Vista
Columbia Pictures
Lionsgate Films
New Line Cinema
Paramount Pictures
Sony Pictures
Universal Pictures
Walt Disney Pictures
Warner Bros. Pictures


C:/Pluralsight - Shiny
http://127.0.0.1:4257
Open in Browser
Publish
Interactive Movie Data
Year
2002
2012
2015
2000 2002 2004 2006 2008 2010 2012 2014 2015
Rating
G
PG
PG-13
R
Genre
Action
Box Office Revenue ($M)
0
200
400
600
800
0
20
40
60
80
100
Critic Score (%)
G
PG
PG-13
R

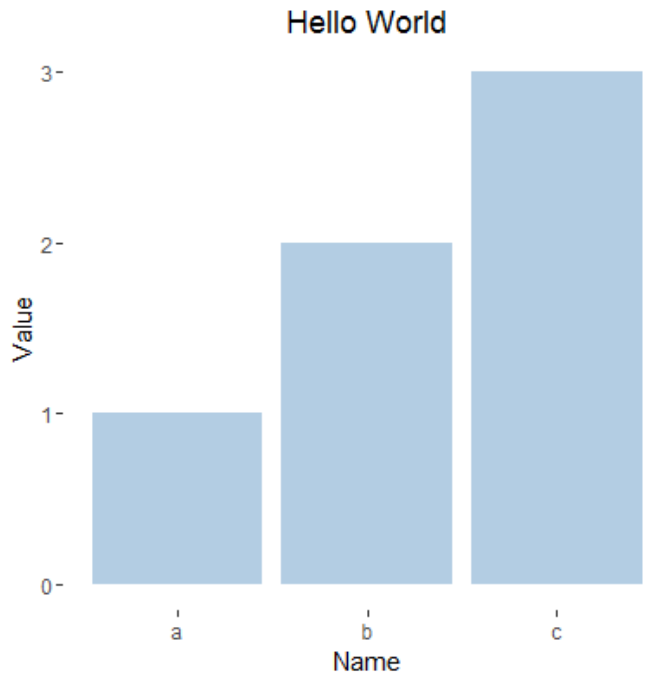
Plotting Systems in R



Base



Lattice



ggplot2


COWBOYS &
Space Invaders:
The Musical
Extended Edition
R

Code Demo

Lab 4
Data Visualization

The
Warlord
of the
Rings :
An Unexpected Adventure

Feature Length

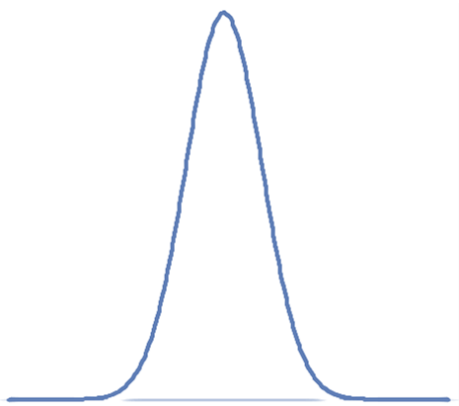
PG

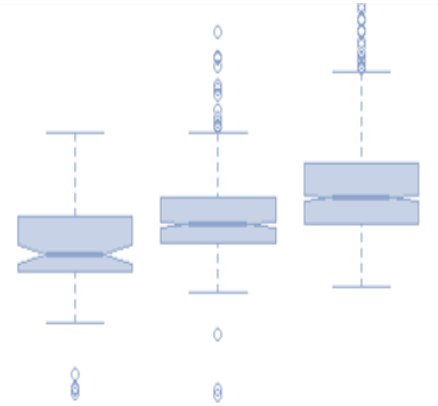
Statistical Modeling

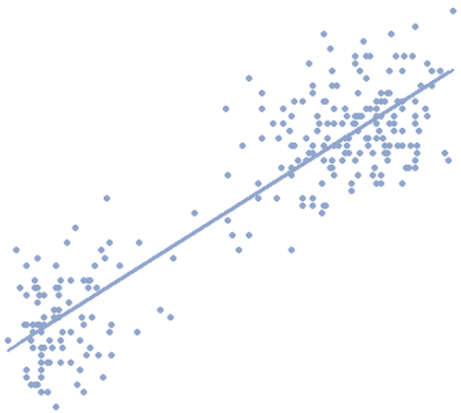
Statistical Model

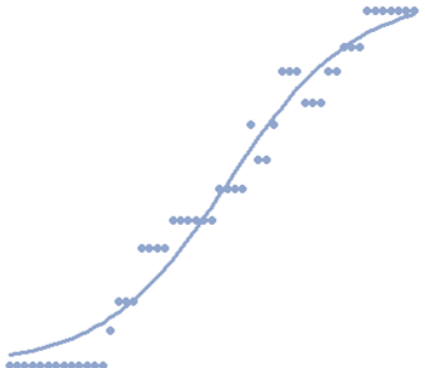
Mathematical equations

Approximation of reality









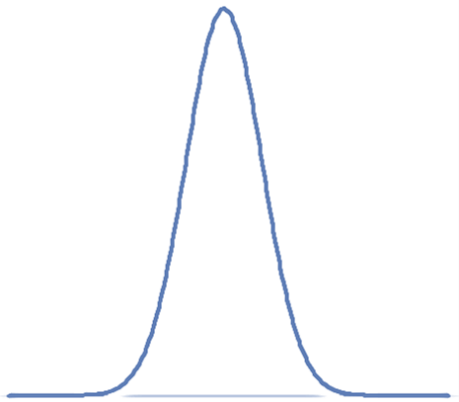
Statistical Model

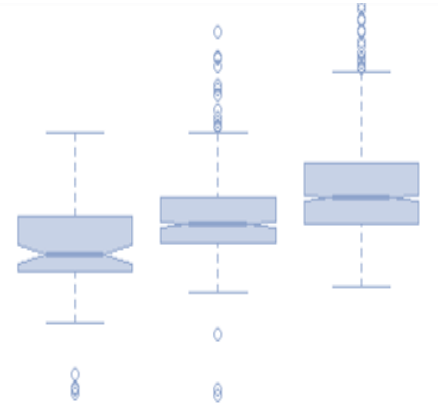
Description

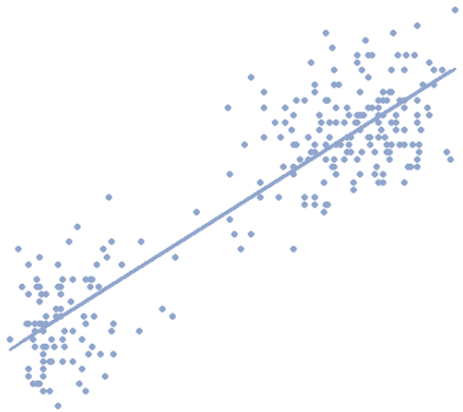
Inference

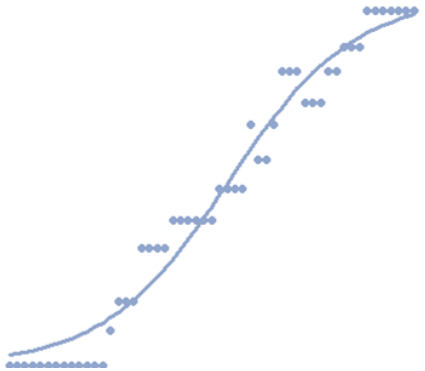
Comparison

Prediction





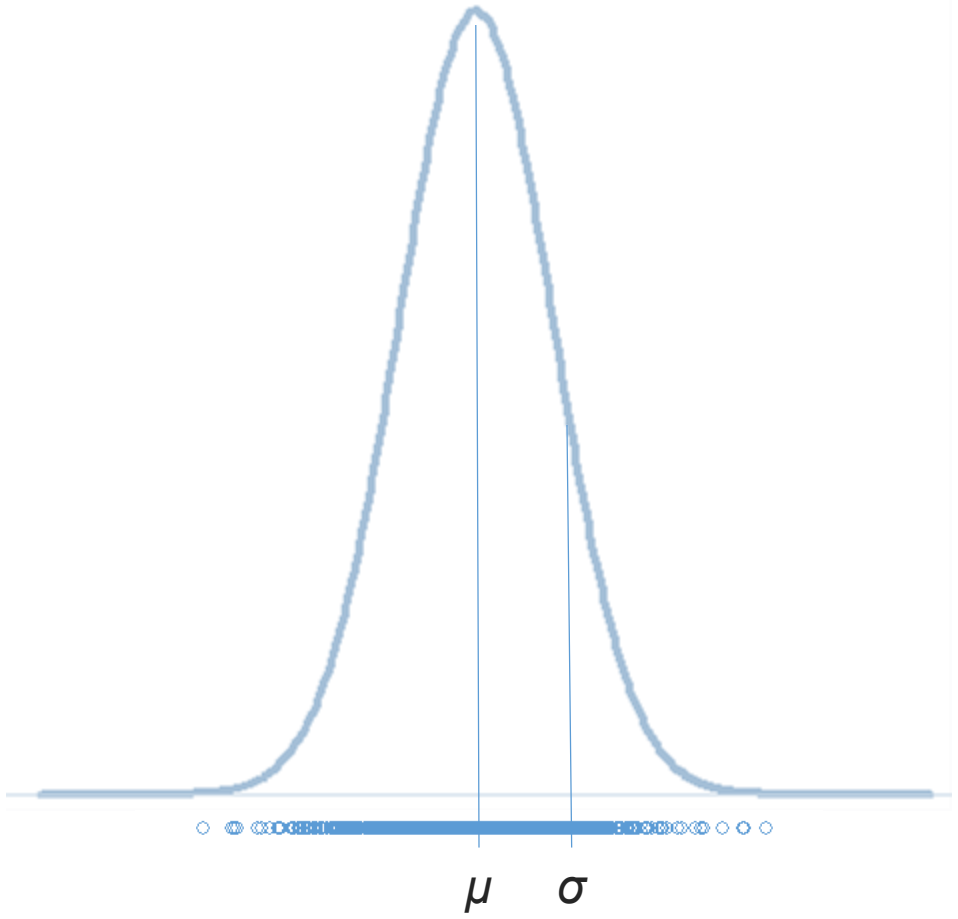




Statistical Model

Parametric

Non-parameteric



Types of Statistical Models

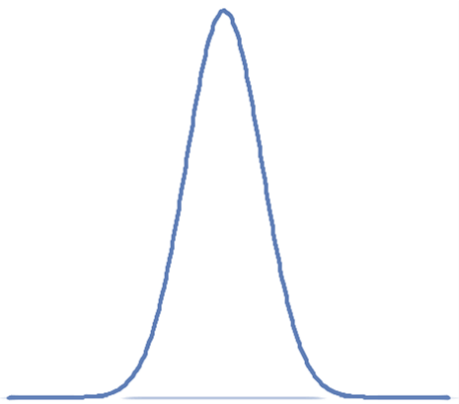
Probability distribution

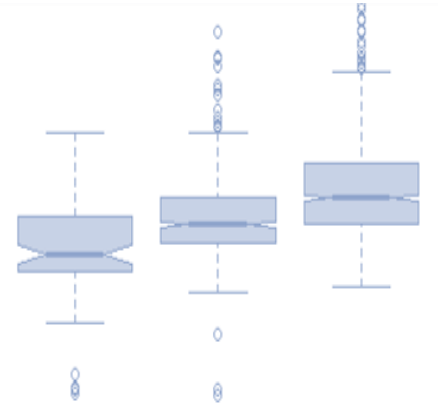
Analysis of variance (ANOVA)

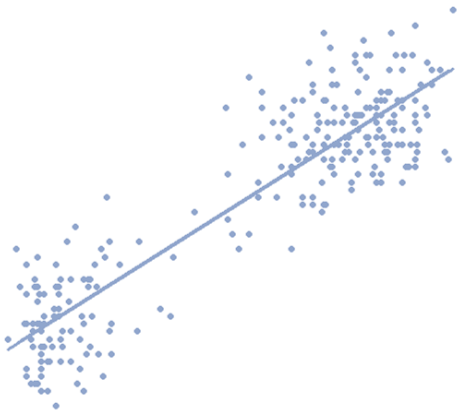
Linear regression

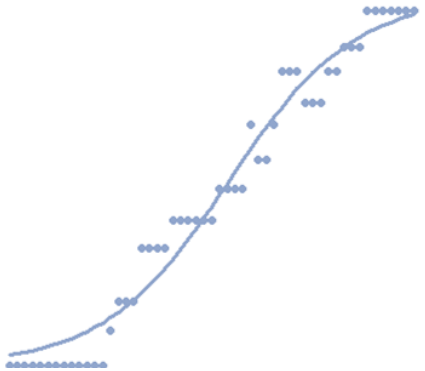
Non-linear regression

Bayesian network









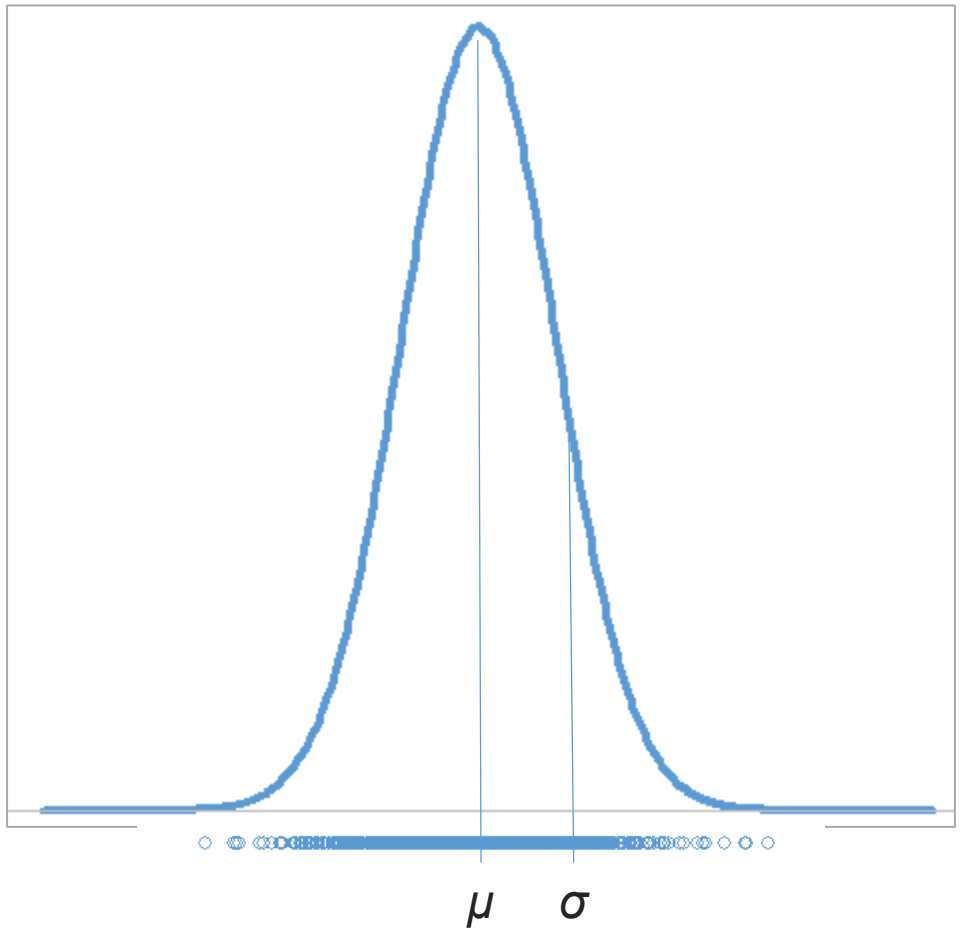
Gaussian Distribution

Probability distribution

Parametric model

- Mean (μ)
- Standard deviation (σ)

Generative model



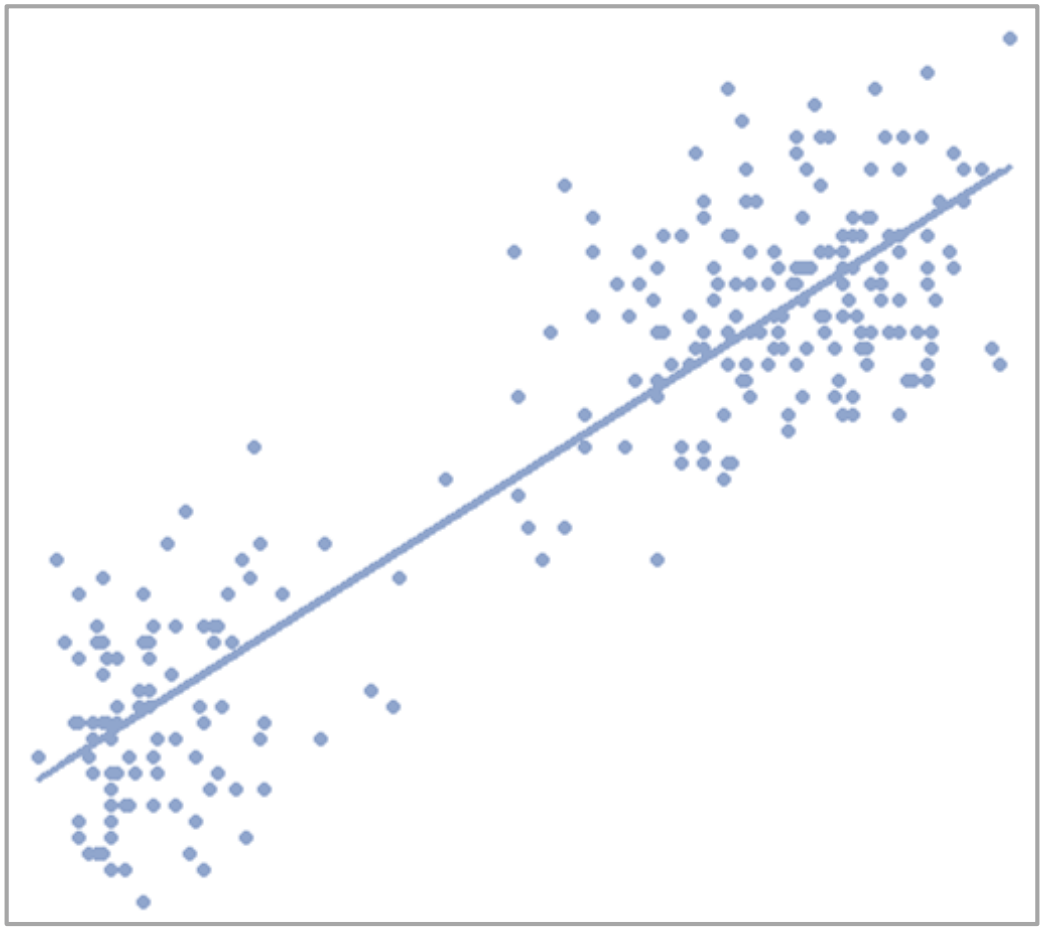
Simple Linear Regression

Relationship

Linear model

Explanatory variable

Outcome variable



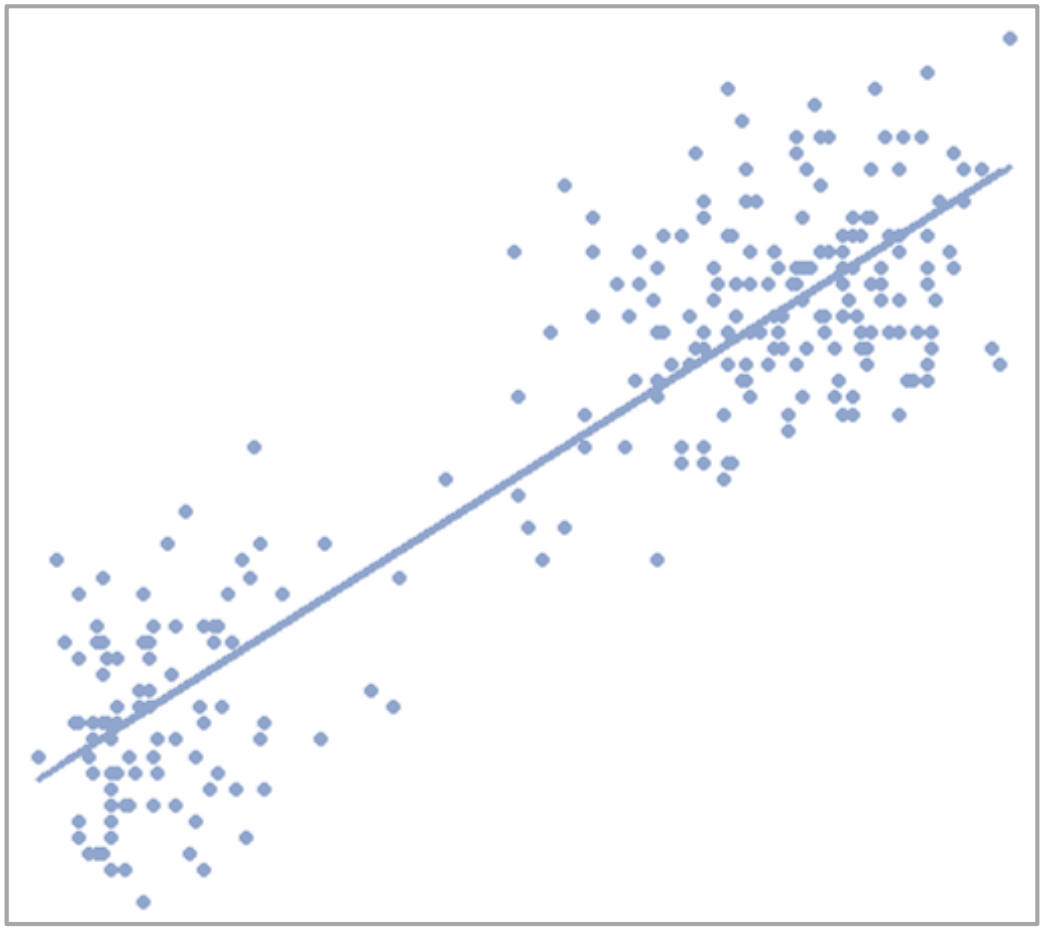
Simple Linear Regression

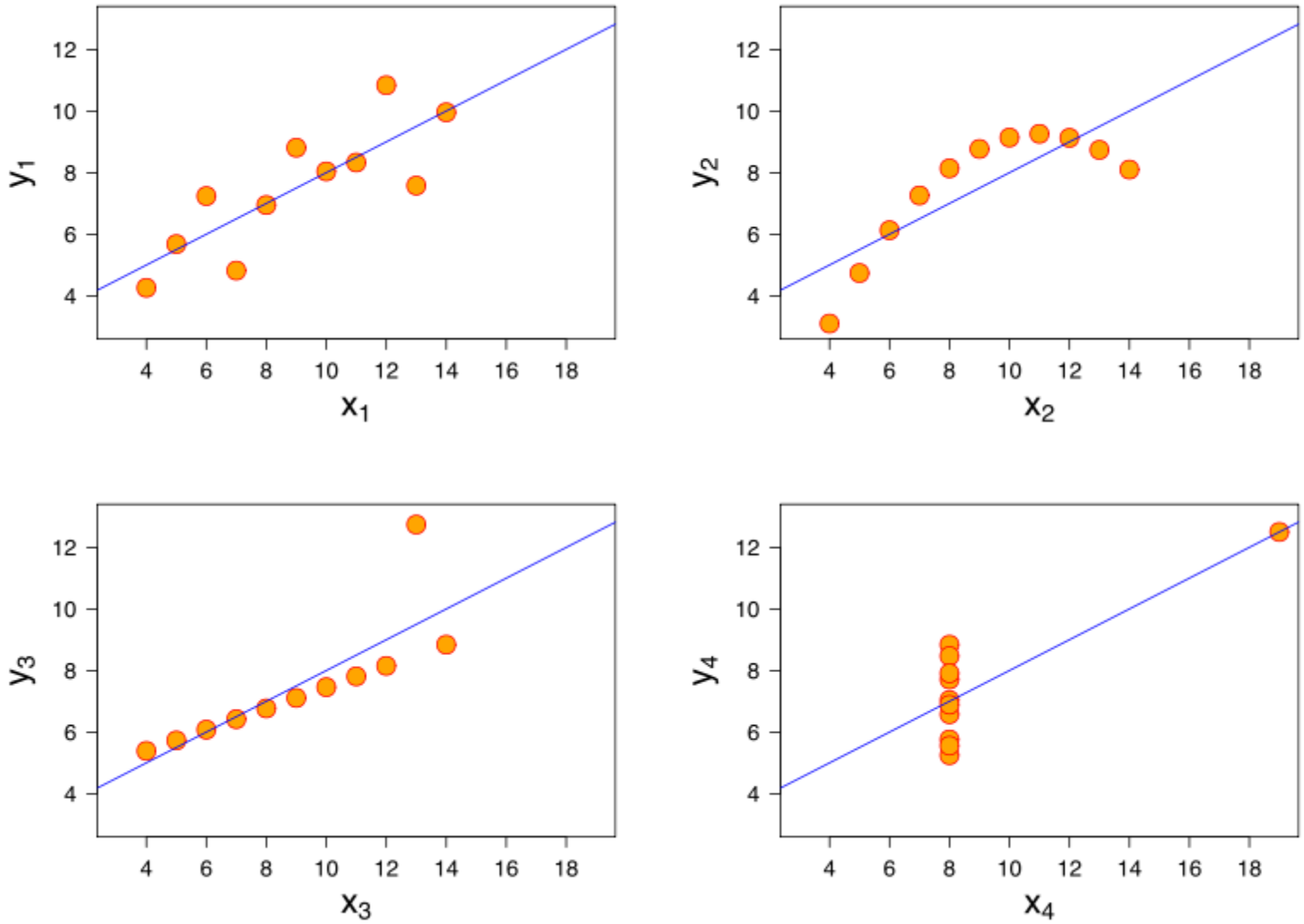
Linear predictor function

$y = m \cdot x + b$

Parameters estimated

Relies on assumptions





Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Photo by Danielle Langlois

Iris Data Set

Fisher's Iris Data				
Species	Petal Length	Petal Width	Sepal Length	Sepal Width
setosa	1.1	0.1	4.3	3
setosa	1.4	0.2	4.4	2.9
setosa	1.3	0.2	4.4	3
setosa	1.3	0.2	4.4	3.2
setosa	1.3	0.3	4.5	2.3
…		…	…	…

Iris Data Set

Iris Setosa

Iris Versicolor

Iris Virginica

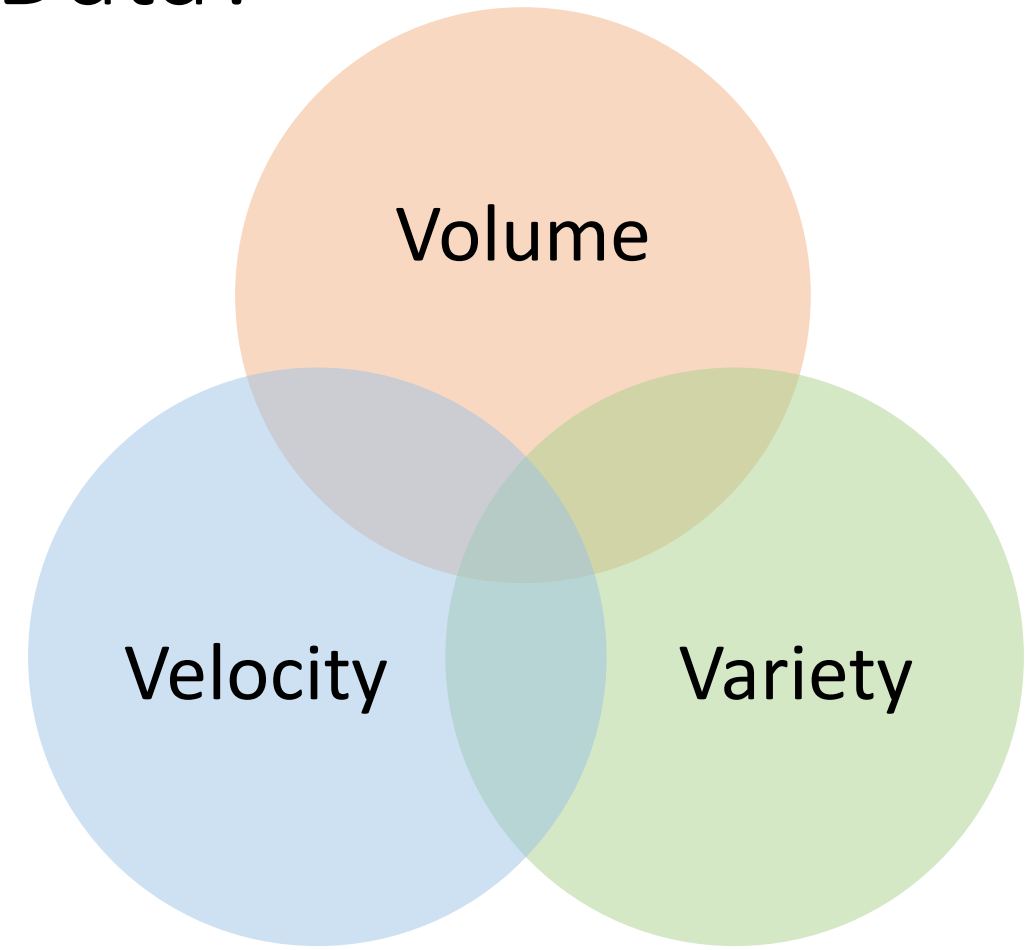
Code Demo

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

Big Data Decision Table				
Class	Rows	Size	Storage	Management
Small	Millions	Gigabytes	Memory	R with desktop computer
Medium	Billions	Terabytes	Hard disk	R with medium-data extensions
Big	Trillions	Petabytes	Clusters	R with big-data extensions

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

More hardware

Subsetting

Sampling

Microsoft R Open

Medium data packages

Big data packages

Microsoft R Server

More Hardware

More CPU, memory, disk

Cost vs. benefit

Cloud Virtual Machine

Sampling

Randomly selected

Subset of original data

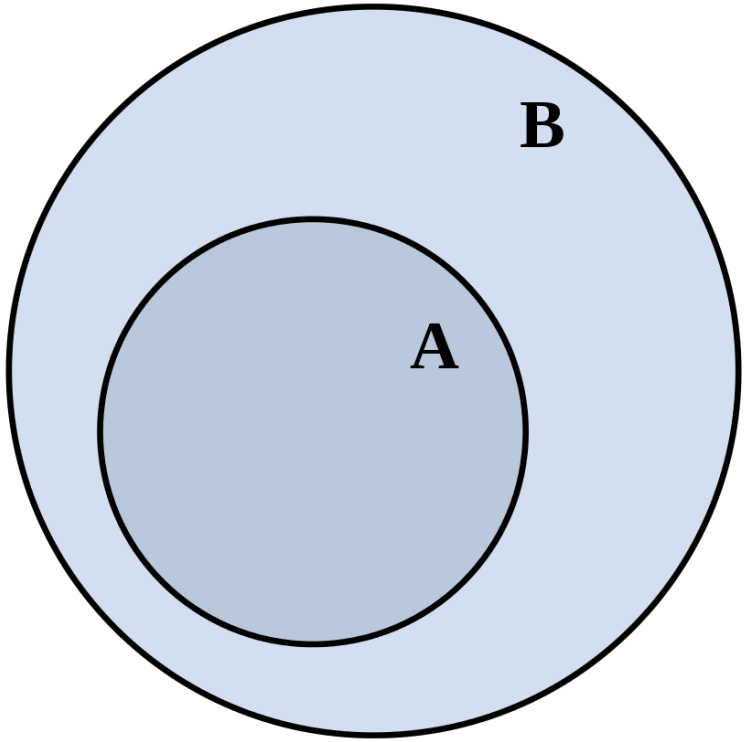
Low-cost solution

Subsetting

Select subset of data

Use dplyr with src

Various databases supported



Microsoft R Open

Enhance 64-bit R distribution

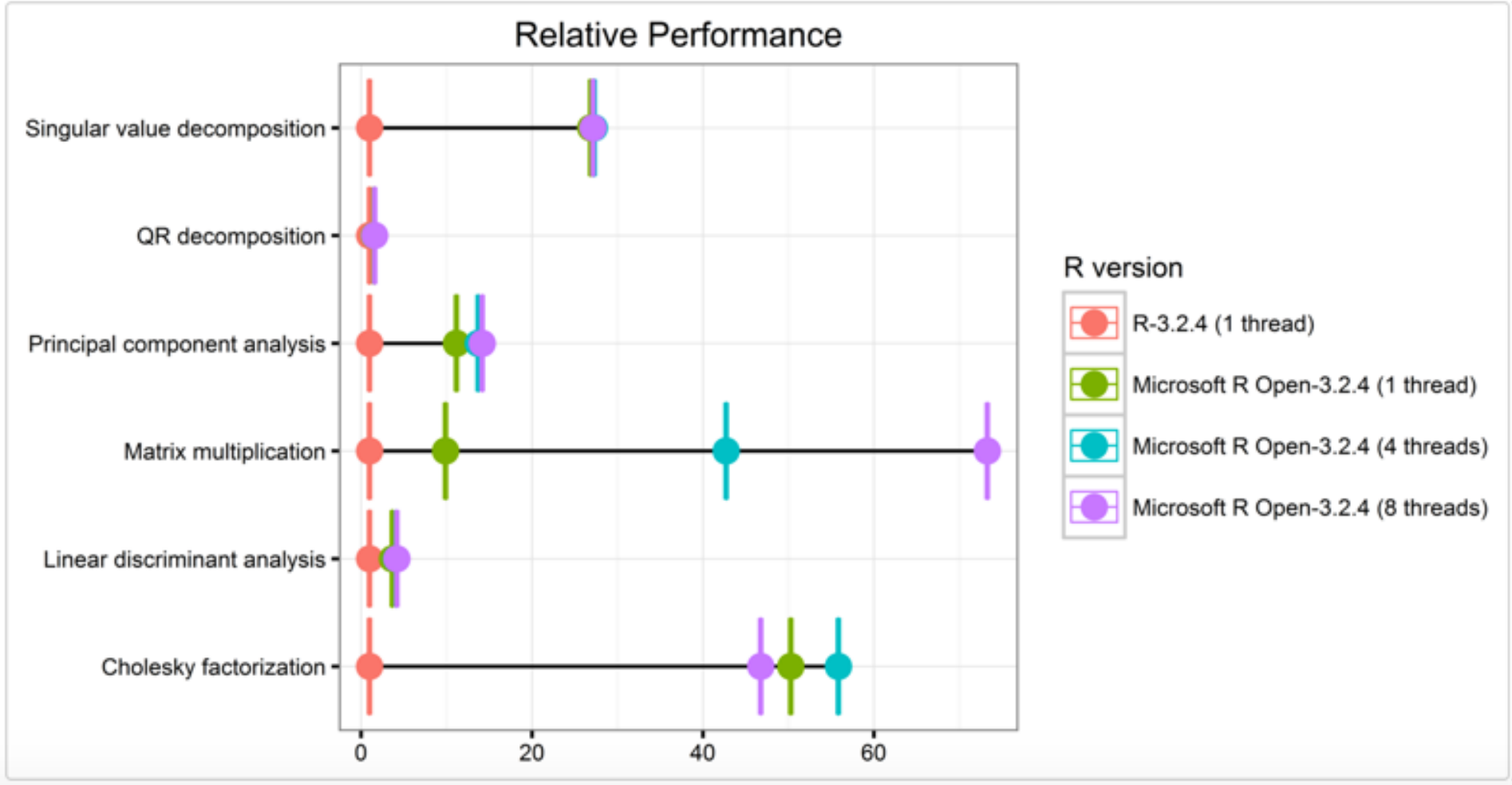
Multithreaded performance

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

3rd-Party Extension Packages

Big Memory

ff

3rd-Party Extension Packages

pbdR

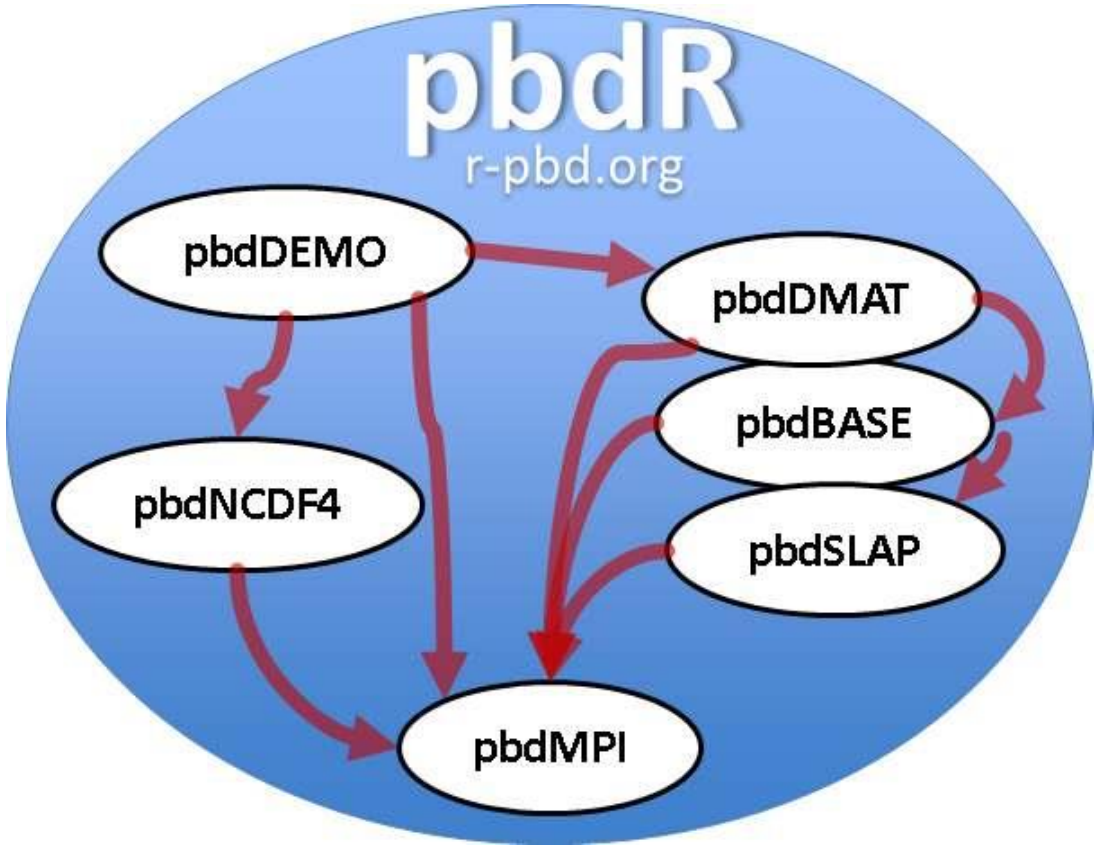
Rhipe

Rhive

Rhbase

Rhdfs,

Rmr



Source: Wikipedia

Microsoft R Server

Windows

SQL Server 2016

SUSE Linux

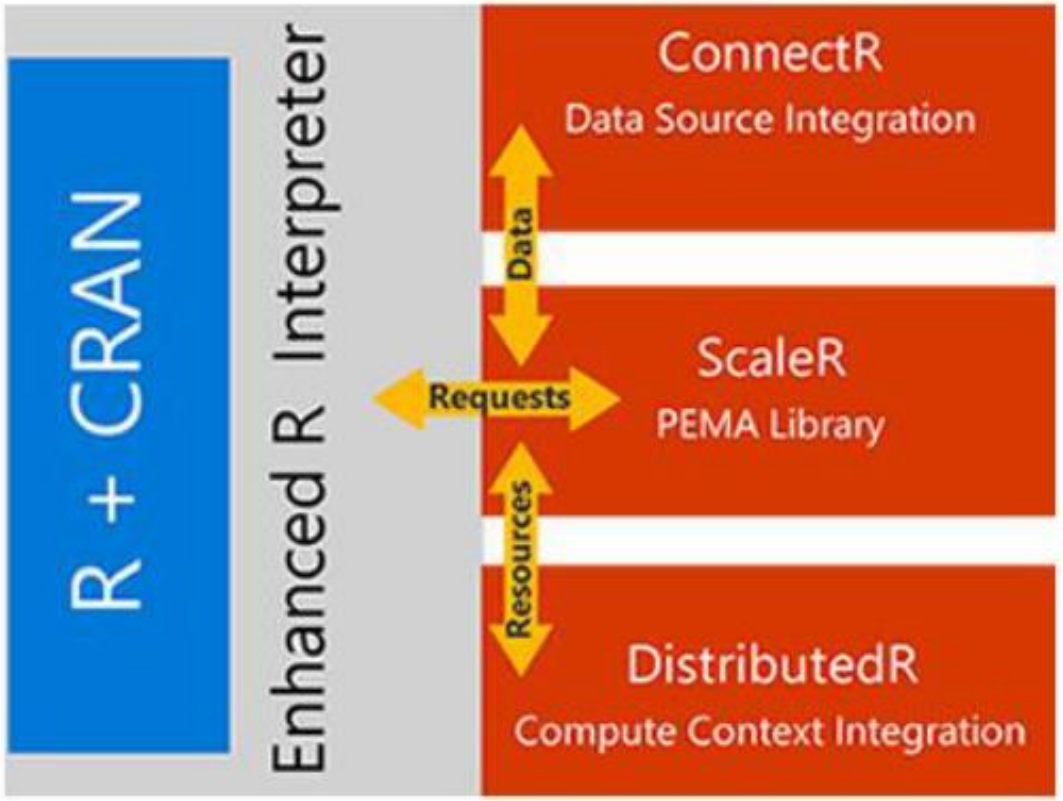
Redhat Linux

Teradata

Hadoop

HD Insight

Microsoft R Server



Source: Microsoft R Server

Code Demo

Lab 6
Handling Big Data

Machine Learning

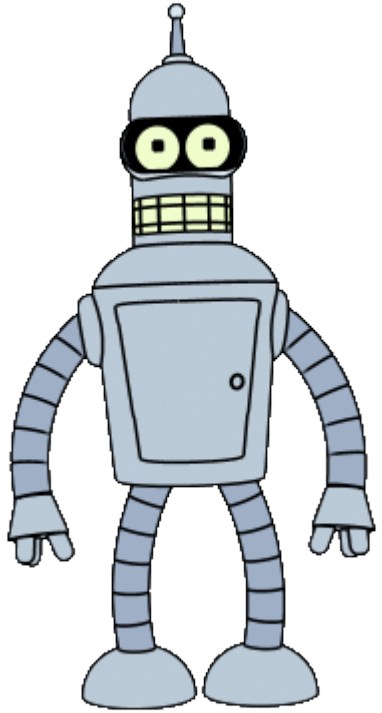
What is Machine Learning?

Subtype of artificial intelligence

Learning without being programmed

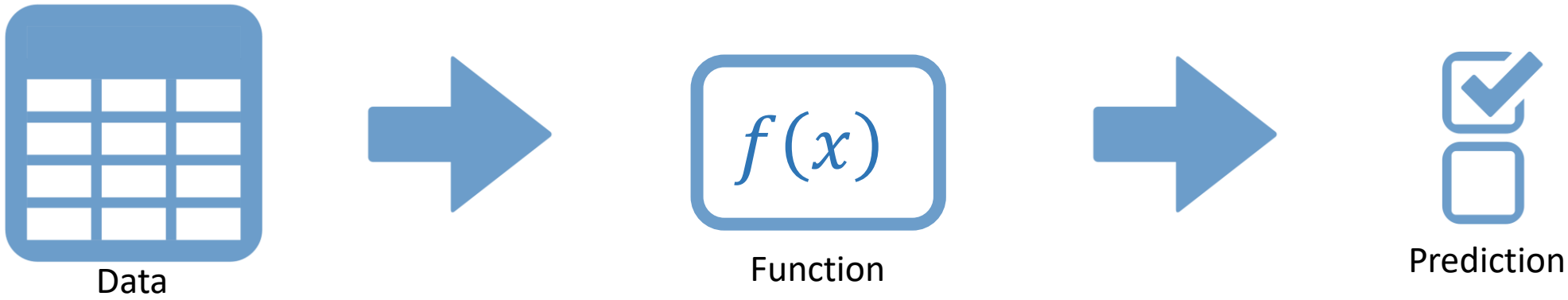
Similar to statistical modeling

Similar to data mining



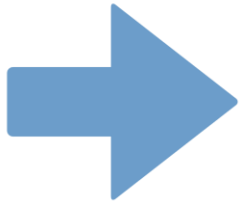
Source: Futurama

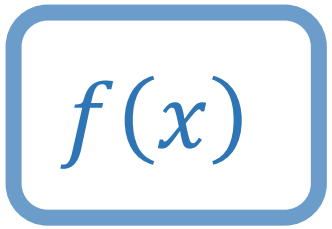
What is Machine Learning?



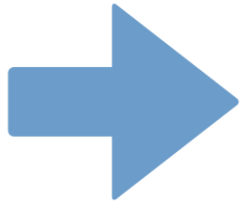
What is Machine Learning?

Data





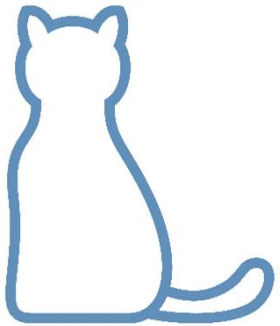
Function



Prediction

Cat

Dog



Is cat?

Yes

How does Machine Learning Work?

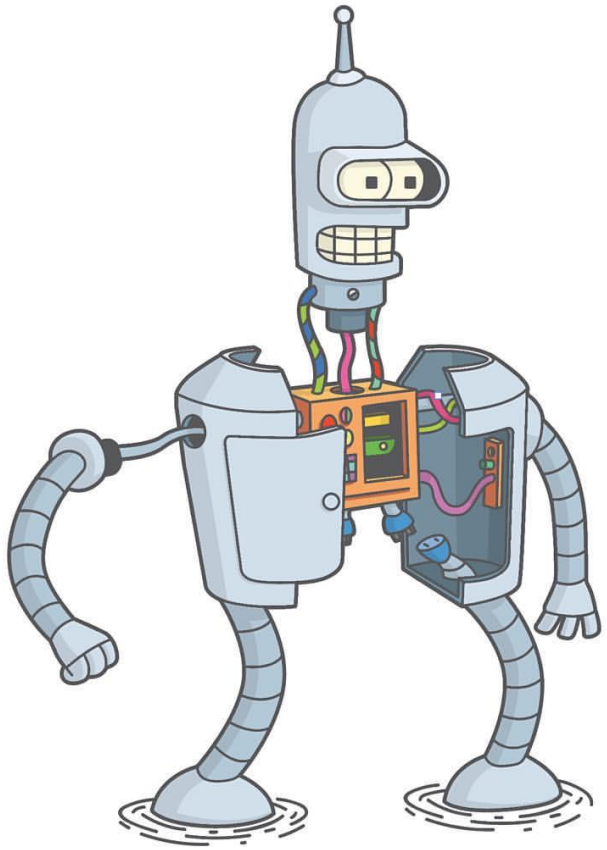
Uses statistical models

Model's parameters are trained

Trained with algorithm and data

Model is used to predict output

Prediction vs. explanation



Source: Futurama

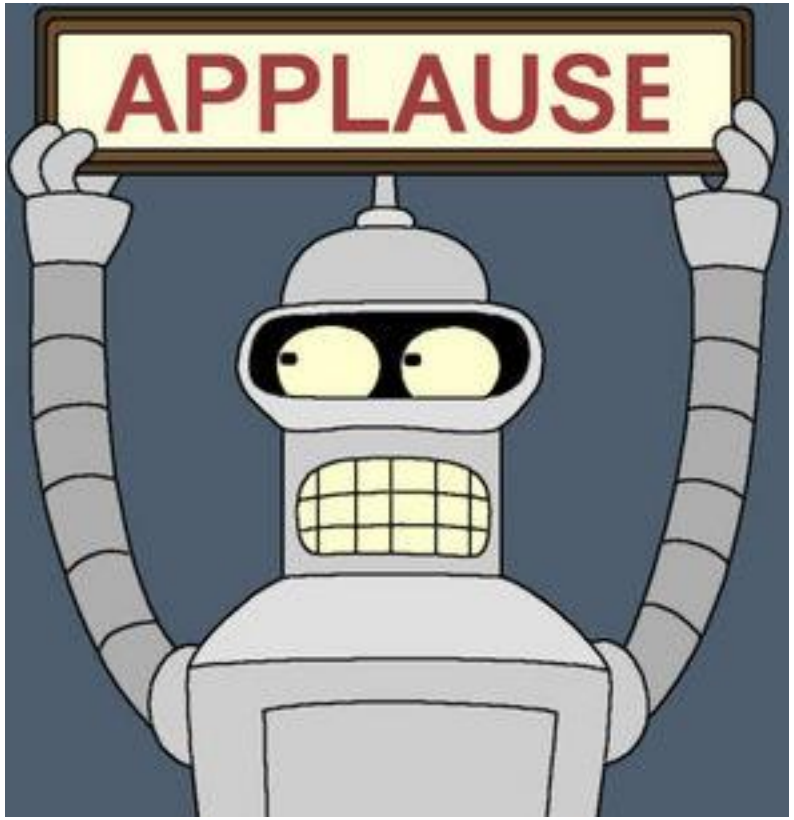
What Can Machine Learning Do?

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

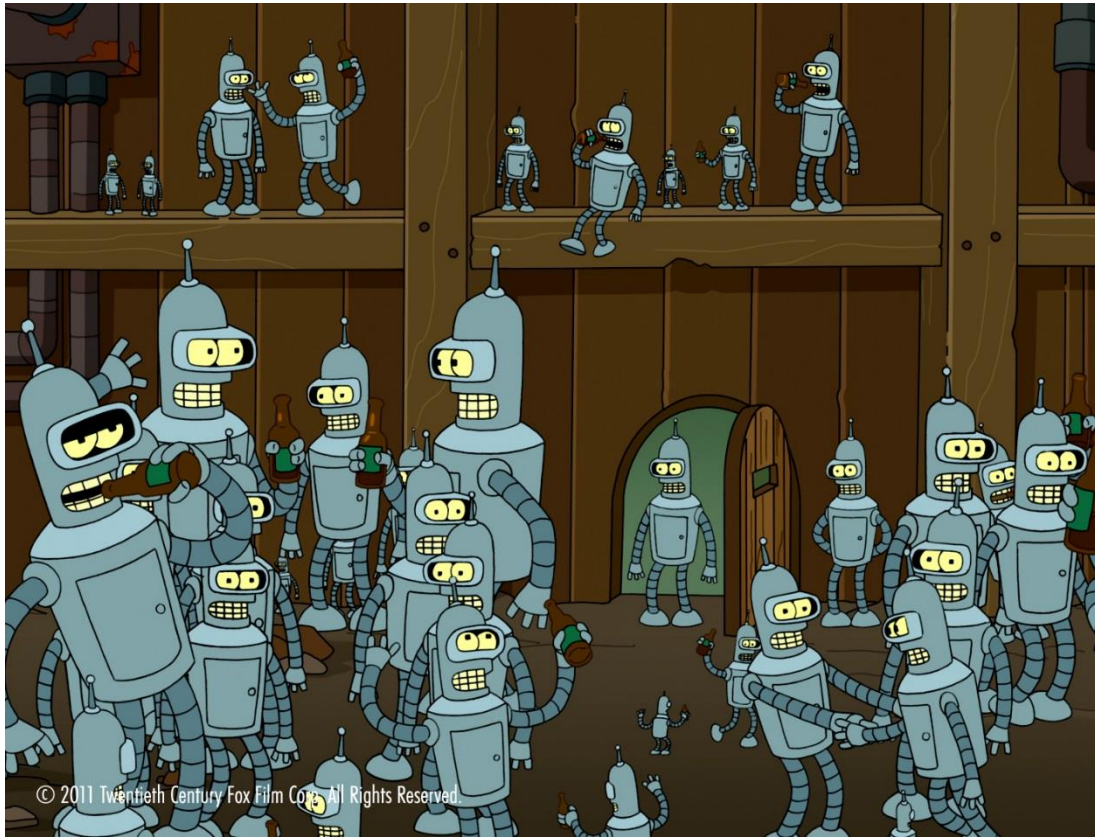
Linear Regression

Support Vector Machines

Neural Networks

K-Means Clustering

Ensemble Learning



Source: Futurama

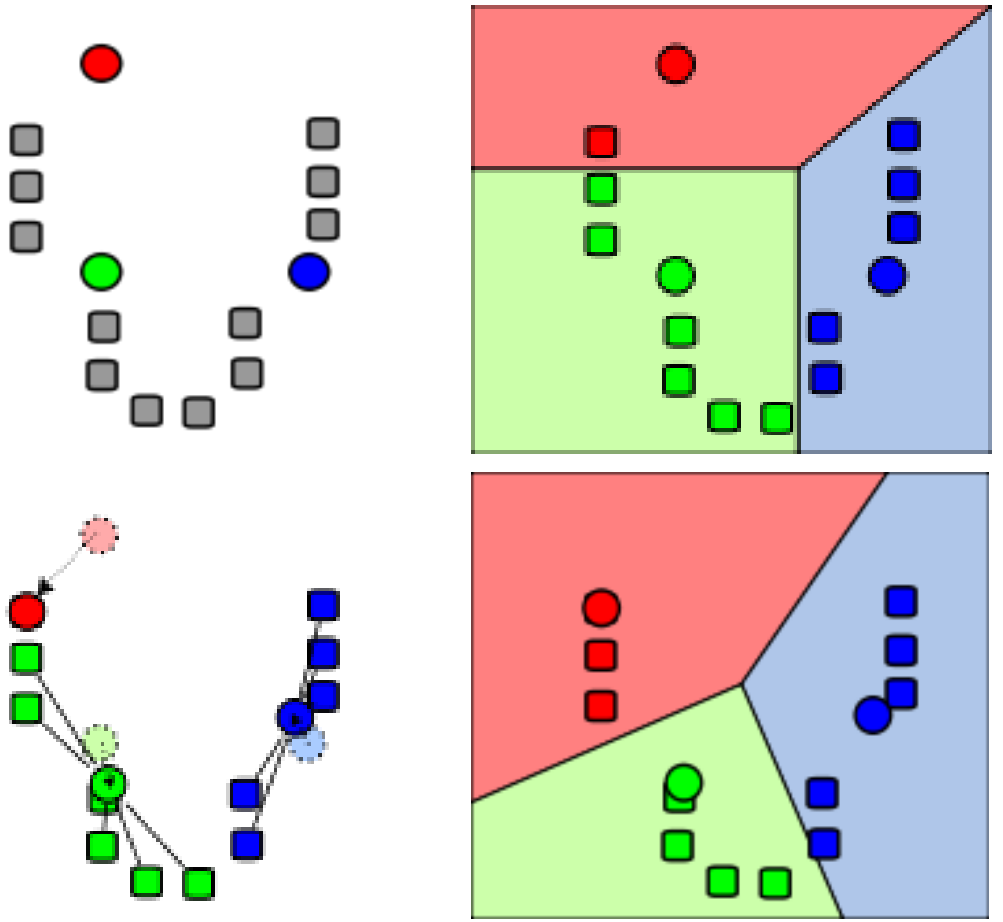
k-Means Clustering

Unsupervised learning

Specify k (# of clusters)

Algorithm finds centers

Random restarts



Source: Wikipedia

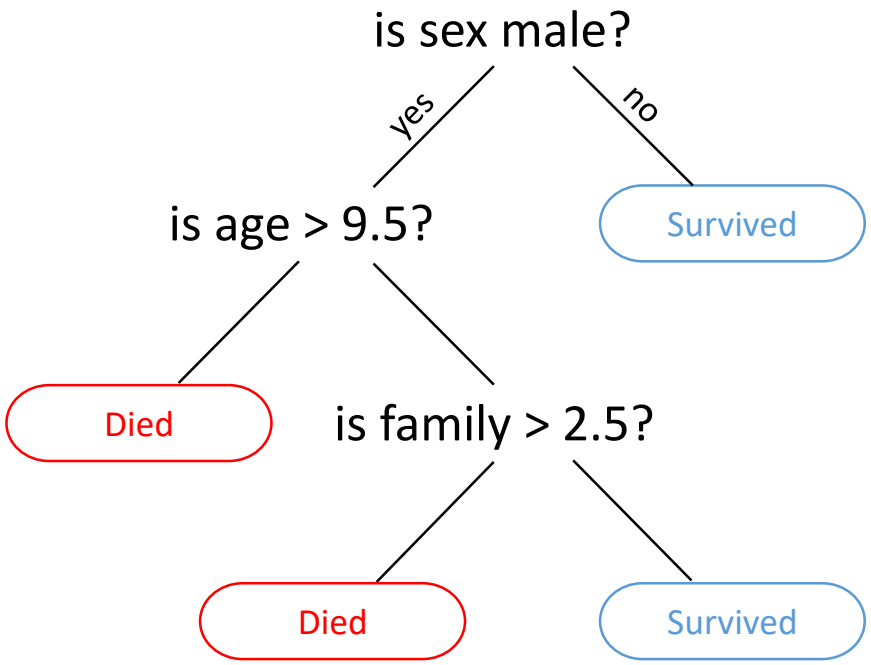
Decision Tree Classifier

Supervised learning

Tree of decisions

Easy to understand

Transparent



Naïve Bayes Classifier

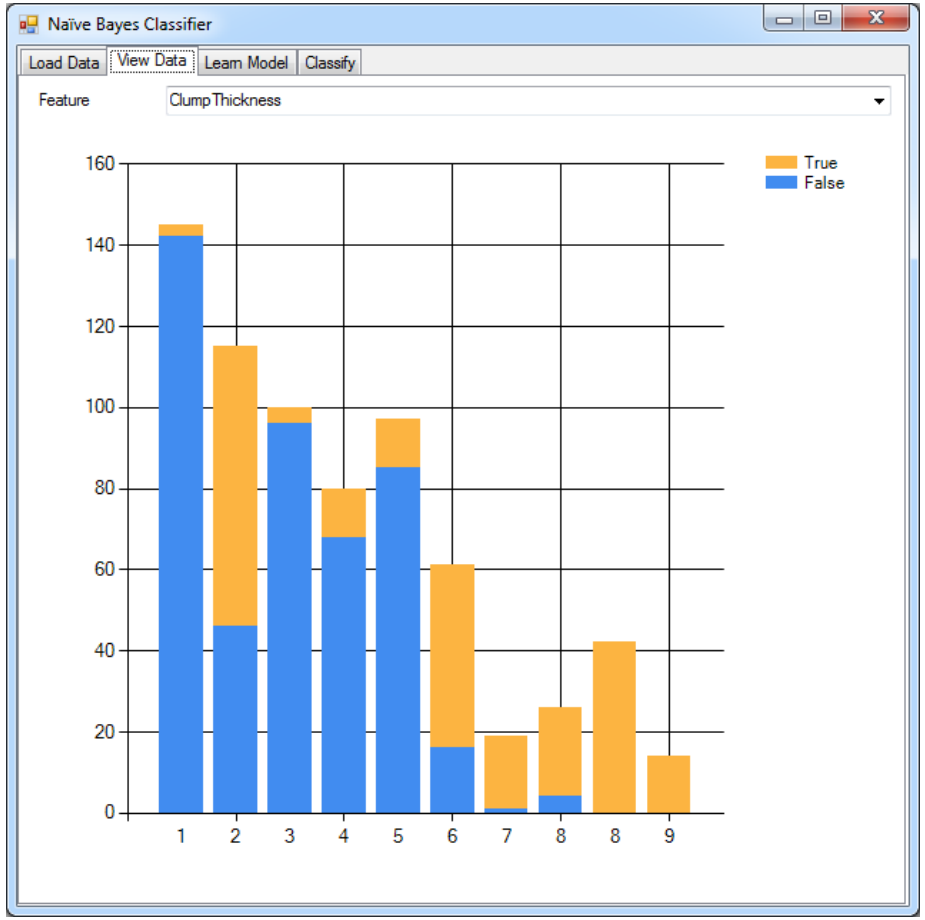
Supervised learning

Simple Bayesian classifier

Independence assumption

Relatively easy to understand

Transparent



Neural Network Classifier

Supervised learning

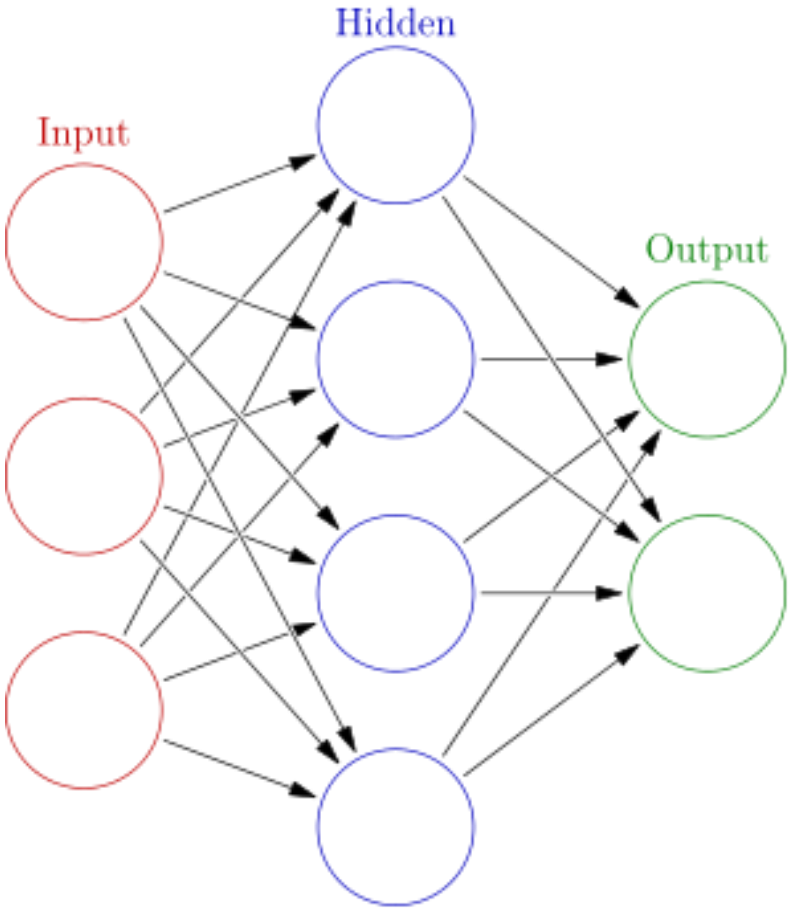
Model of neurons in brain

Layers of neurons

Backpropagation

Complex

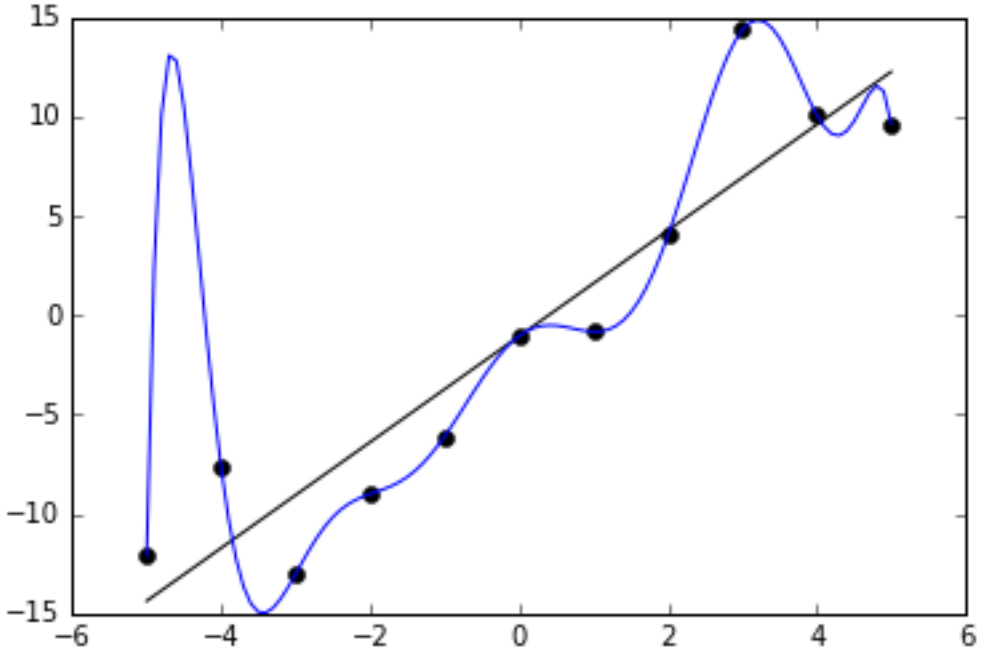
Not transparent



Source: Wikipedia

Overfitting and Regularization

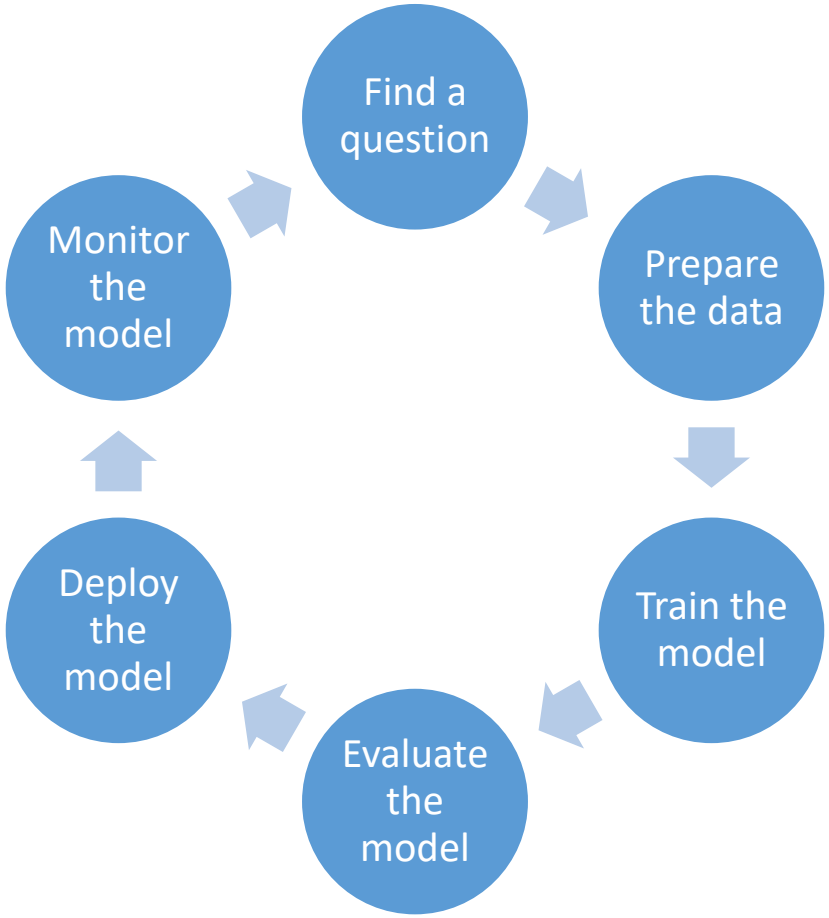
- Overfit – too specialized
- Underfit – too generalized
- Regularization techniques
  - Early stopping
  - Pruning (trees)
  - Adding noise
  - Parameter tuning



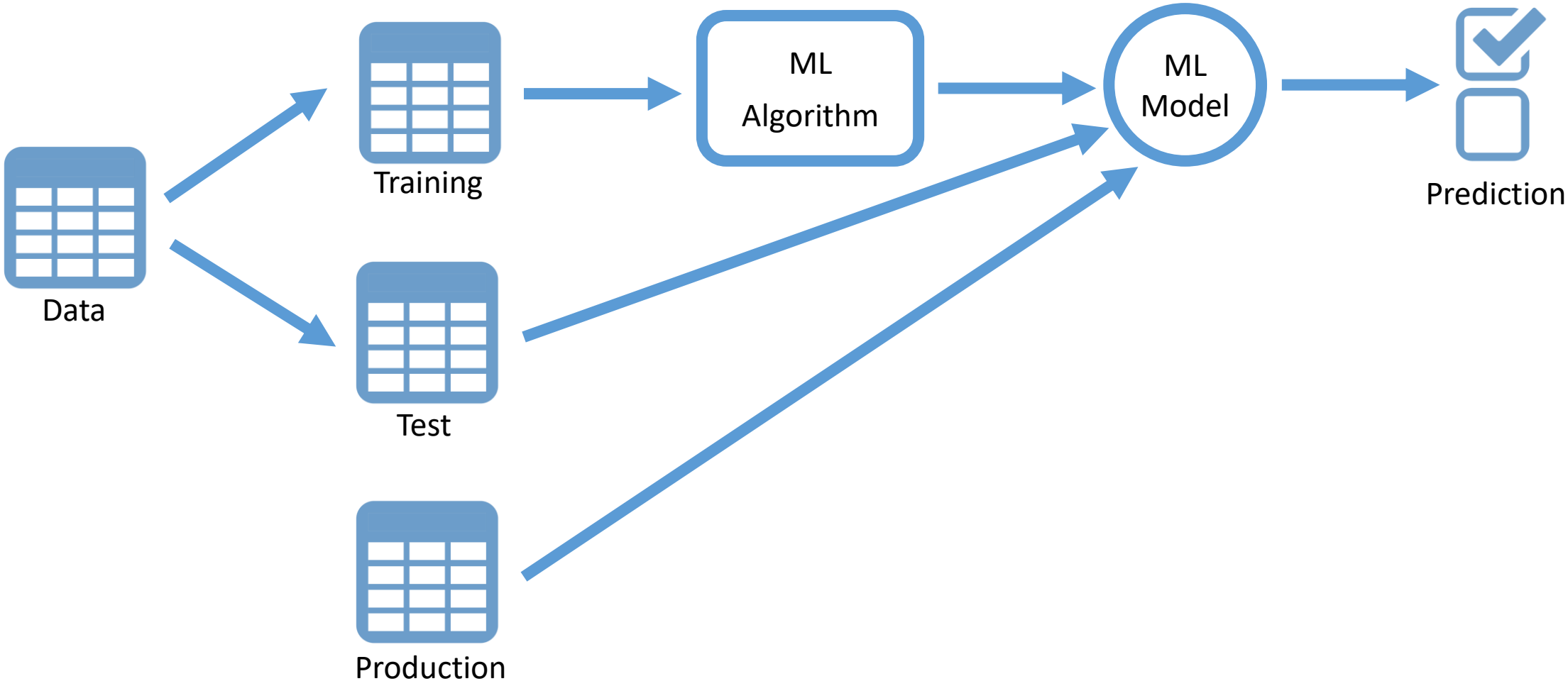
Source: Wikipedia

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



The Machine Learning Process



Photos by Radomił Binek,
Danielle Langlois, and Frank Mayfield

Iris Data Set

Iris Setosa

Iris Versicolor

Iris Virginica

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

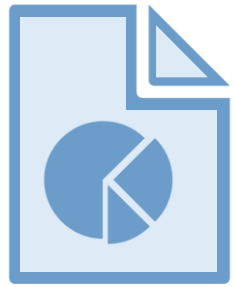
Export charts (Rstudio)

Create documents (Markdown)

Create interactive reports (Shiny)

Deploy to Server (R Server)

Deploy to Cloud (Azure ML)



C:/Workshop/Data - Shiny

http://127.0.0.1:4017 | Open in Browser | Publish

Iris Species Predictor

Petal Length (cm)

1 4 7

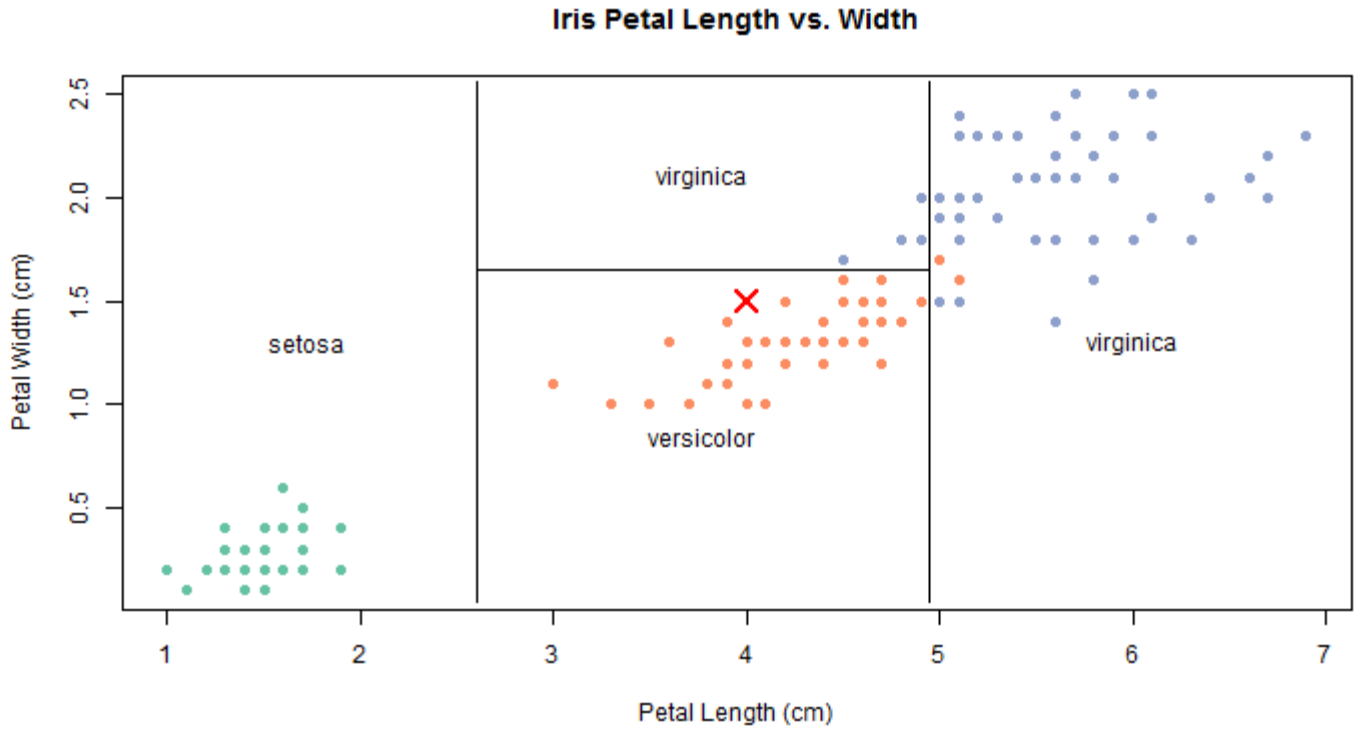
1 2 3 4 5 6 7

Petal Width (cm)

0 1.5 2.5

0 0.5 1 1.5 2 2.5

The predicted species is versicolor

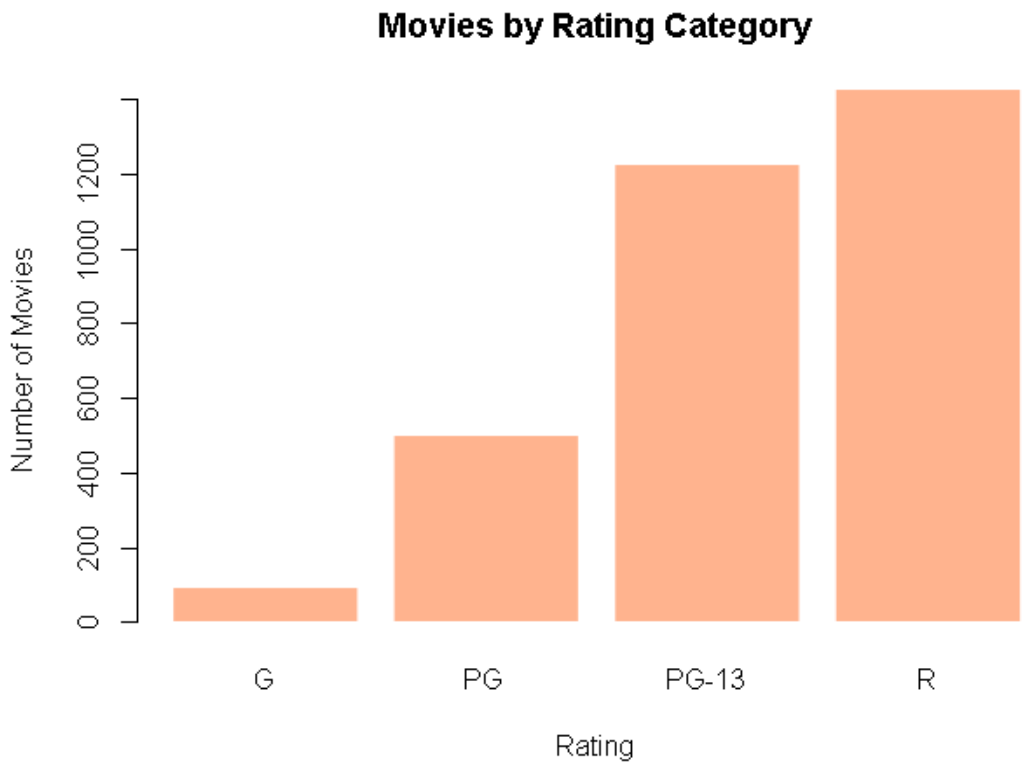


Code Demo

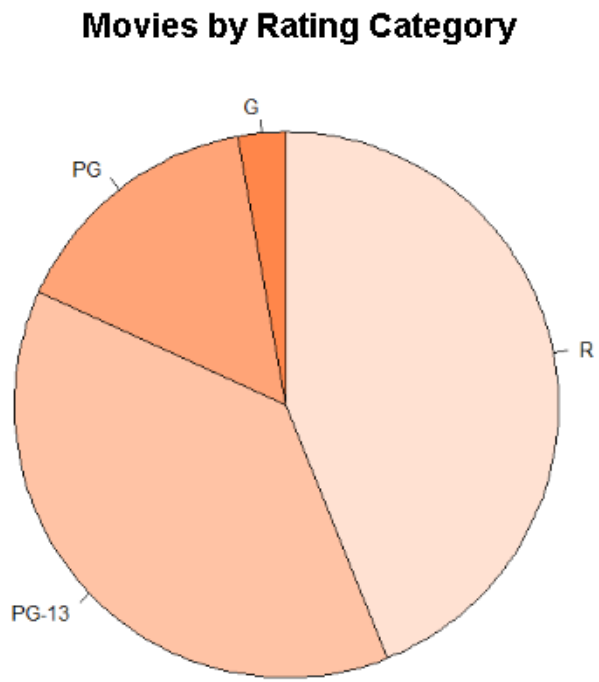

ADVICE
HELP
TIPS
SUPPORT
GUIDAN

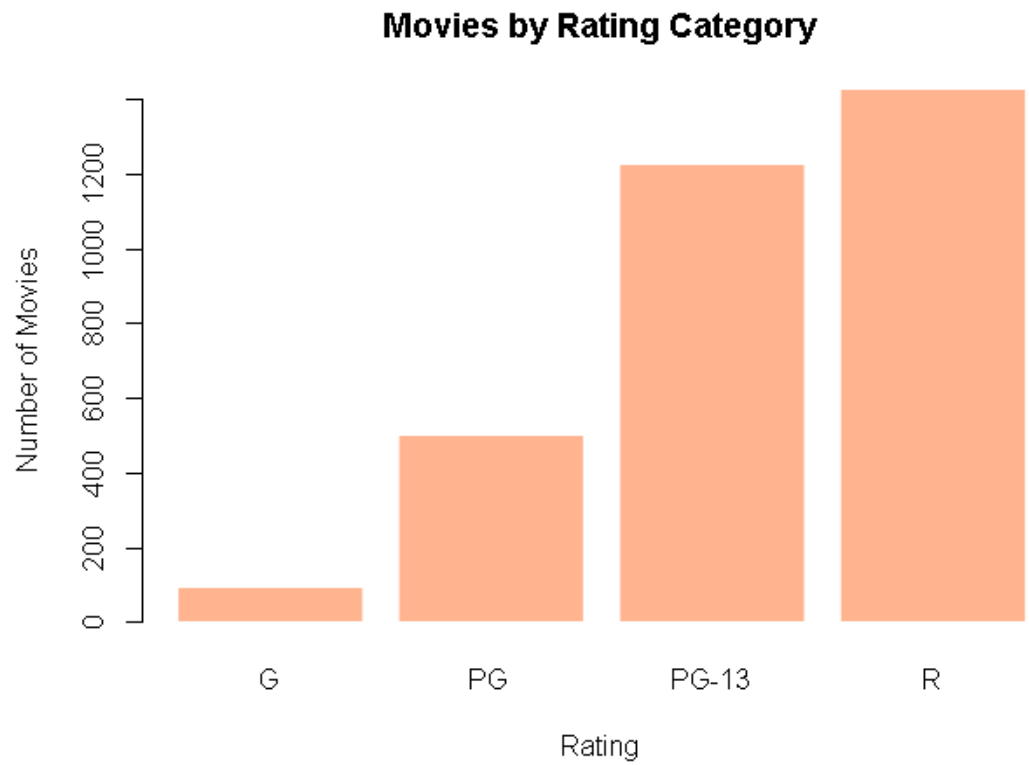
Start with a Question

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

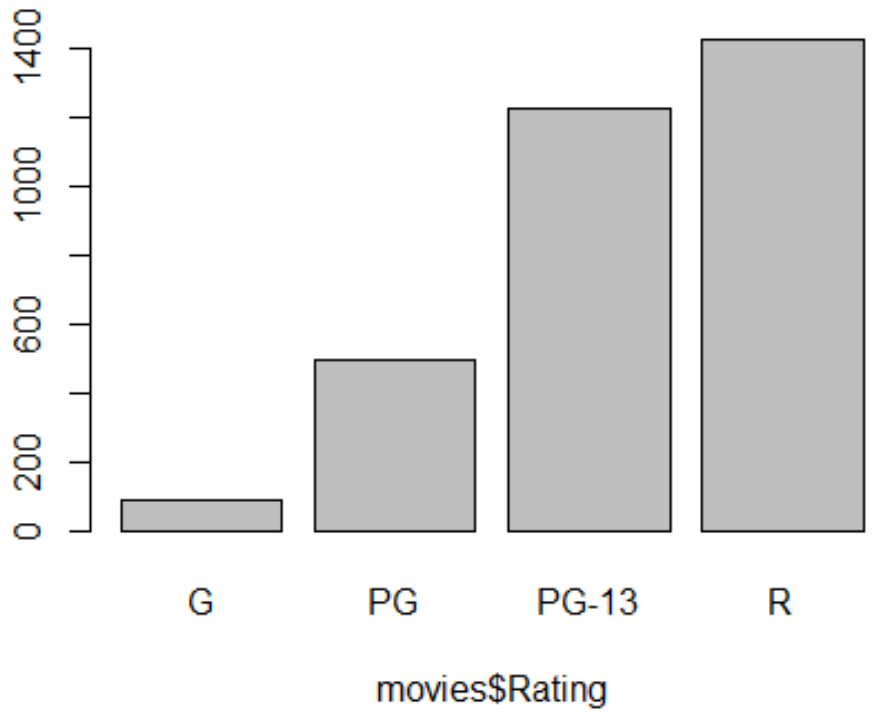


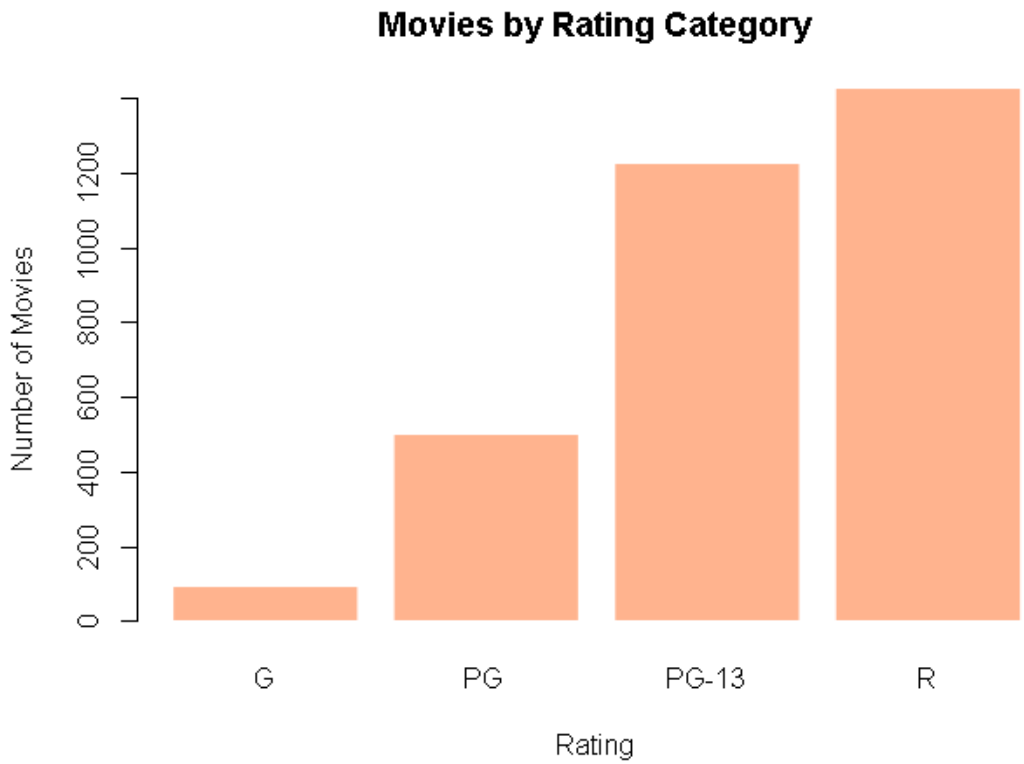
Use the Right Tool for the Job



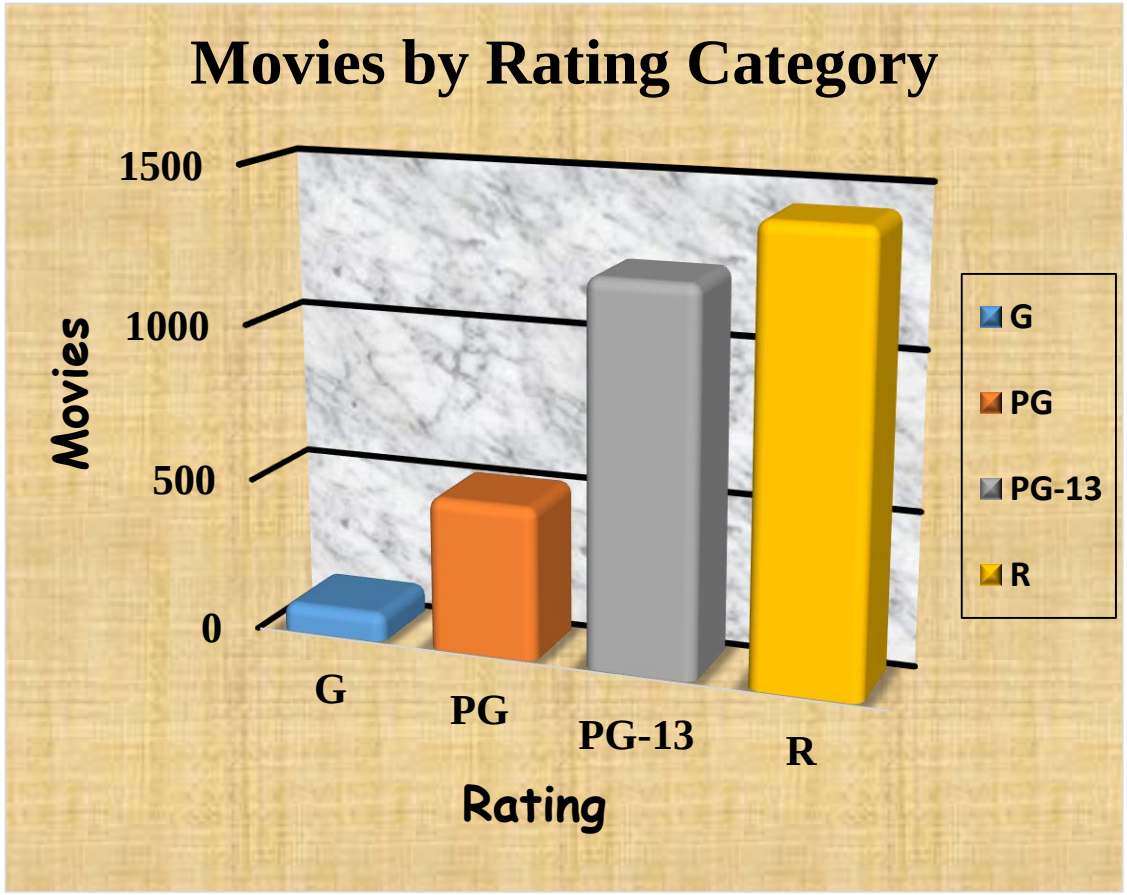


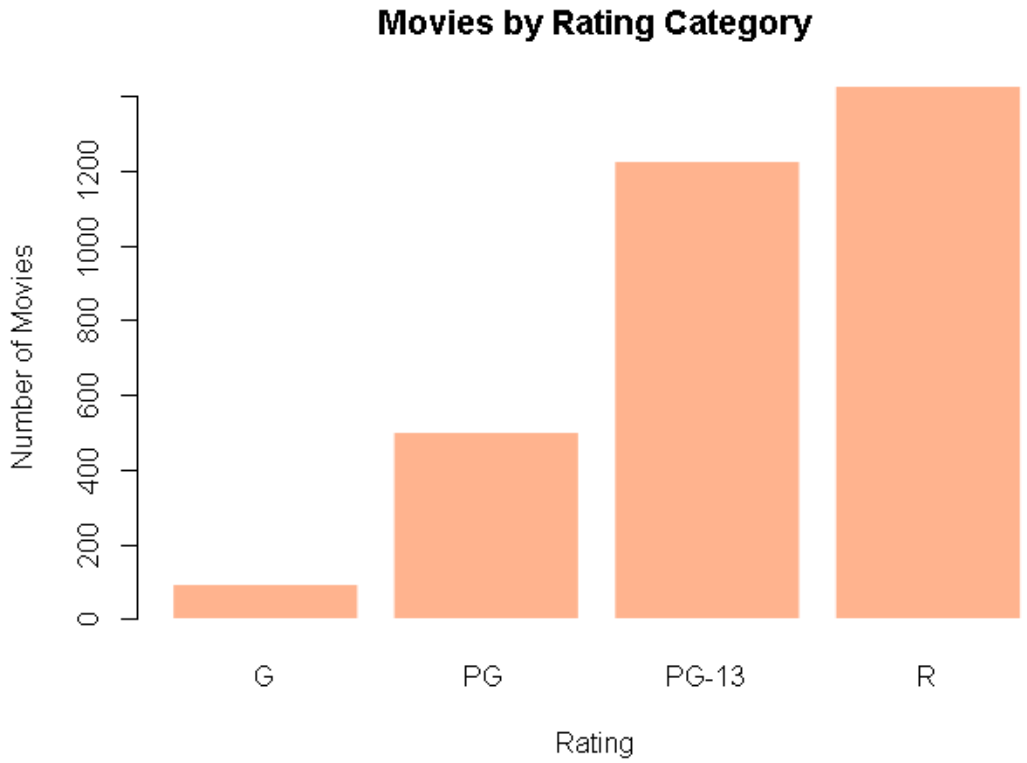
Know Your Audience



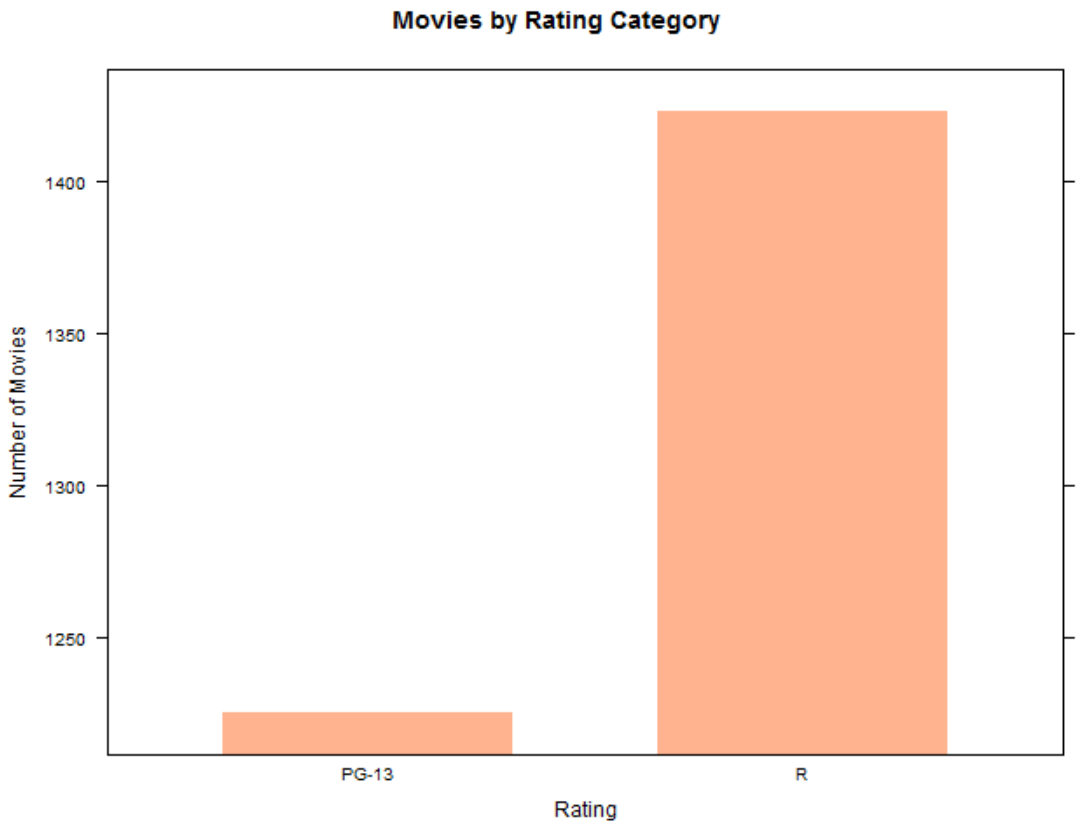


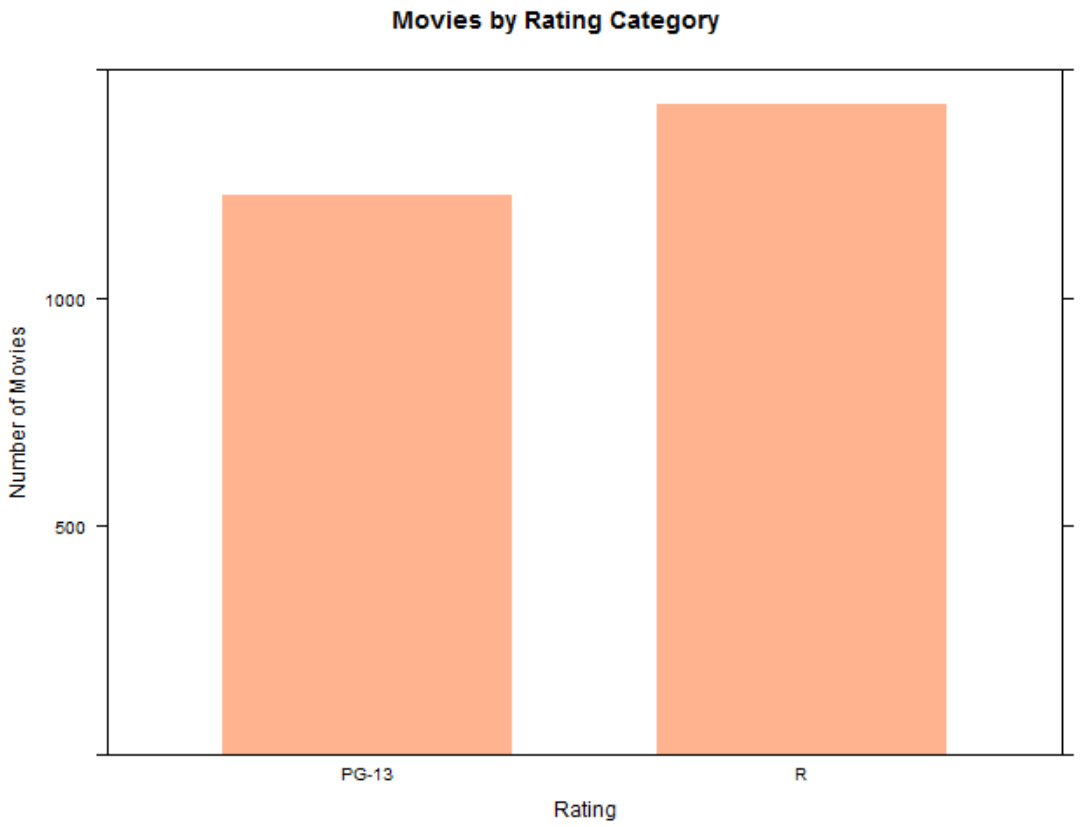
Create Clean Data Analyses





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…

Pluralsight: https://www.pluralsight.com

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

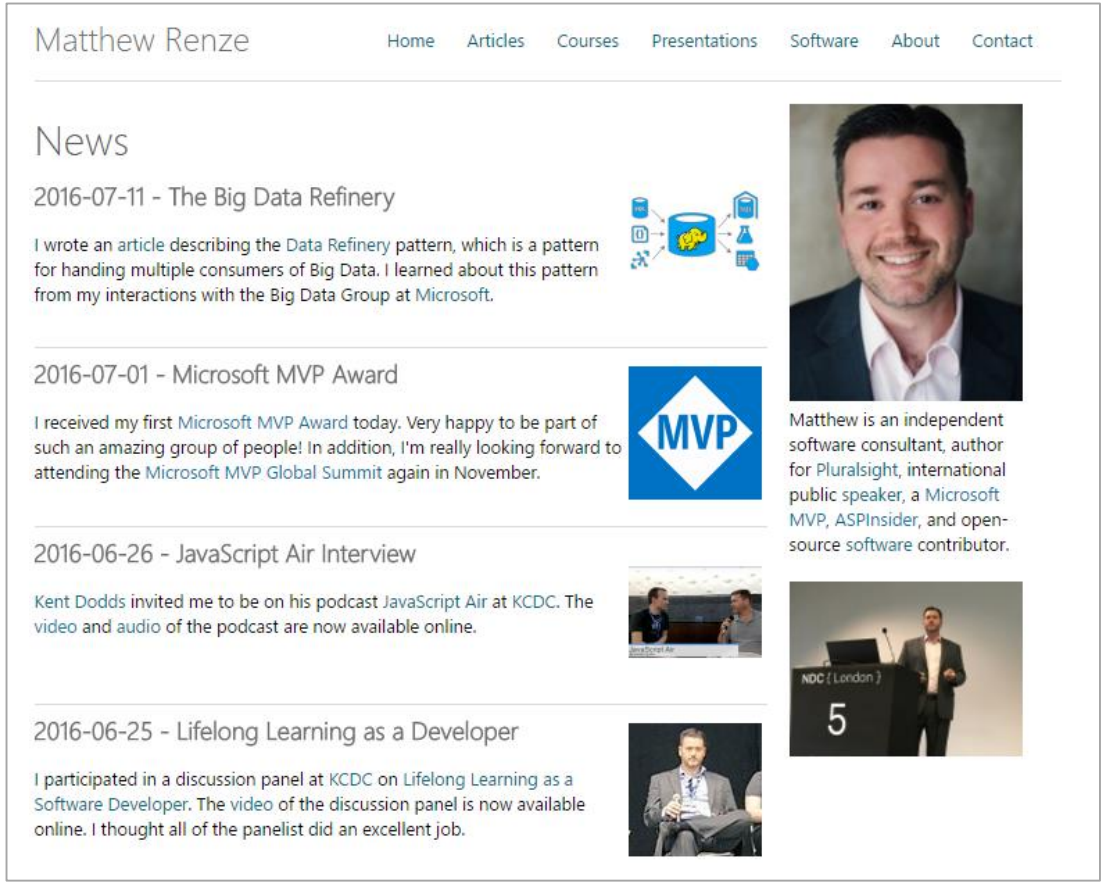
Articles

Presentations

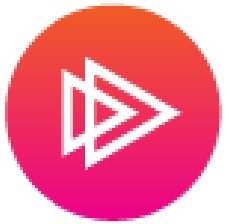
Source Code

Videos

Workshops



www.matthewrenze.com



PLURALSIGHT

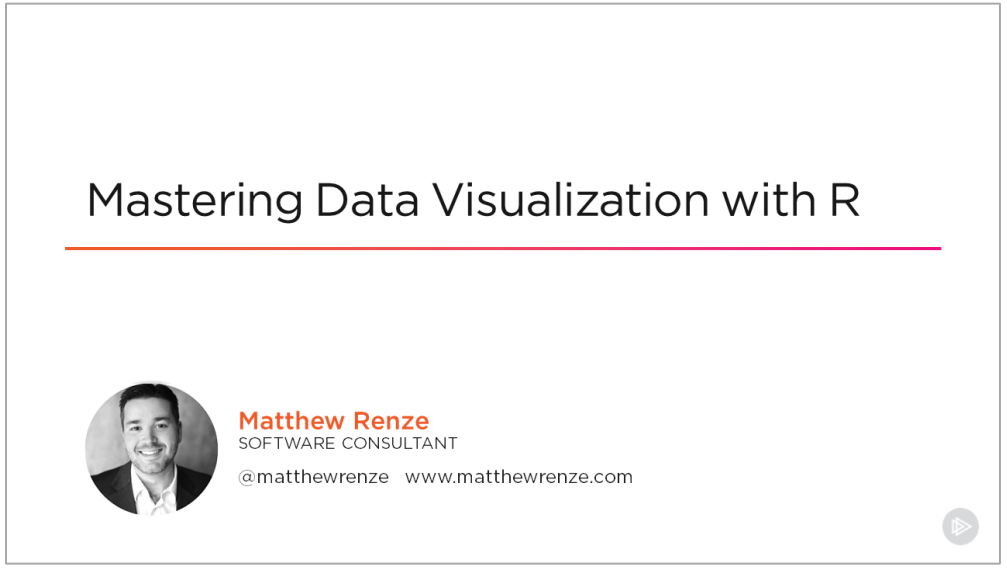
Exploratory Data Analysis with R

Beginning Data Visualization with R

Multivariate Data Visualization with R

Mastering Data Visualization with R

Data Science with R (coming soon)



www.pluralsight.com/authors/matthew-renze

Conclusion

Conclusion

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


dev up Conference 2016 Attendee Party
Ryse Nightclub
October 21st Friday 5:15 – 10:15
Food, Drinks, Games, and Fun!

Feedback

Feedback is very important to me!

What did you like?

What could I improve?

Contact Info

Matthew Renze

Data Science Consultant

Renze Consulting

Twitter: @matthewrenze

Email: matthew@matthewrenze.com

Website: www.matthewrenze.com

Thank You! :)