IT/Dev CONNECTIONS

Exploratory Data Analysis with R

Matthew Renze
Data Science Consultant
Renze Consulting

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."

Job Postings for Data Scientists

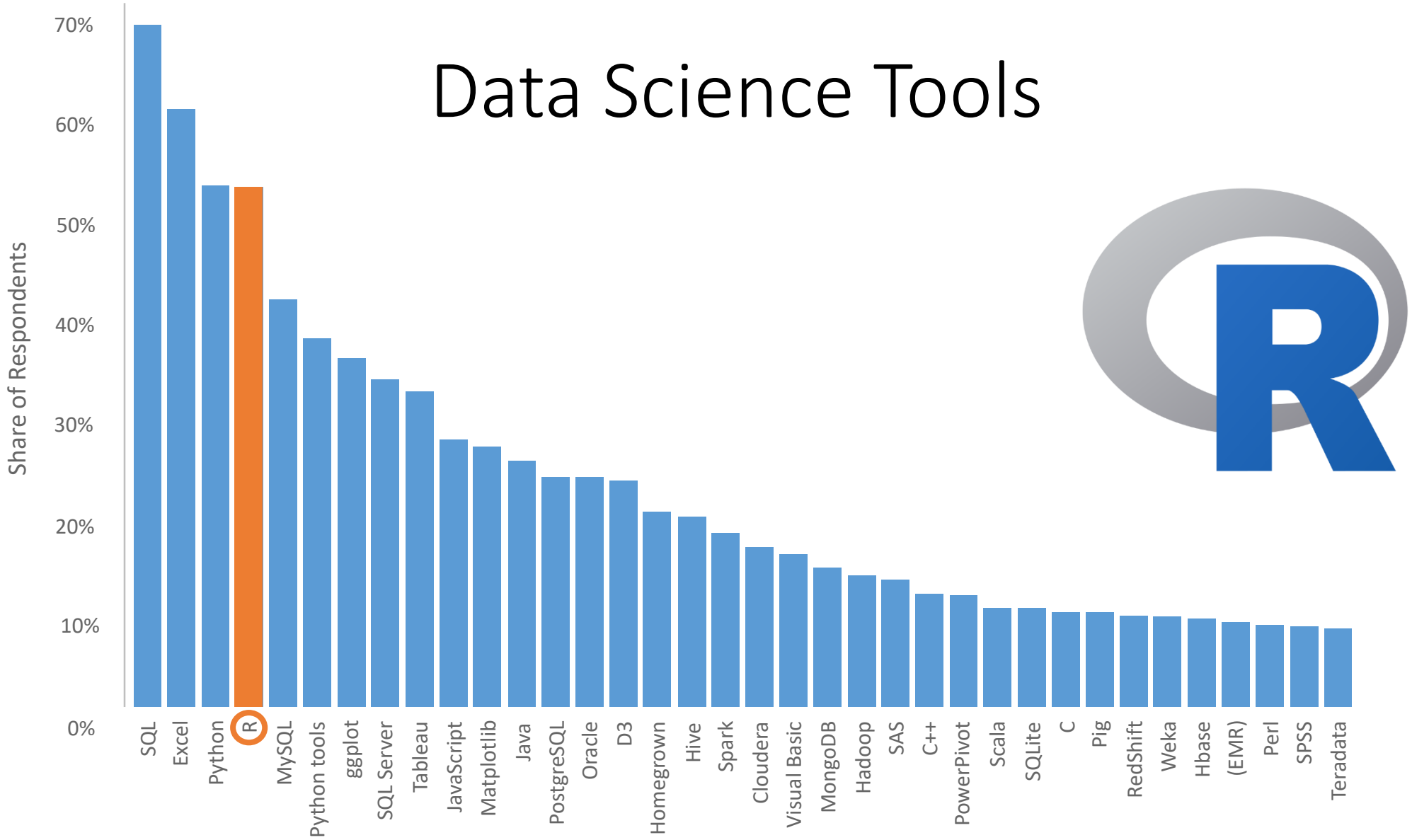
Top-paying Tech Skills

Skill	2016	Change
HANA (High Performance Analytical Application)	$ 128,958	-3.3%
MapReduce	$ 125,009	-0.3%
Cloud Foundry	$ 124,038	n/a
Hbase	$ 123,934	5.7%
Omnigraffle	$ 123,782	-1.9%
Cassandra	$ 123,459	2.2%
Apache Kafka	$ 122,728	n/a
SOA (Service Oriented Architecture)	$ 122,094	-1.9%
Ansible	$ 121,382	n/a
Jetty	$ 120,978	1.3%
PaaS (Platform as a Service)	$ 120,403	-4.4%
Elasticsearch	$ 120,002	n/a
ABAP (Advanced Business Application Programming)	$ 119,961	0.5%
NoSQL	$ 119,498	1.3%
CMMI (Capability Maturity Model Integration)	$ 119,466	-0.6%
Amazon Redshift	$ 119,197	n/a
Pig	$ 119,118	-4.2%
Solr	$ 119,032	0.1%
Cloudera	$ 118,896	-9.0%
Docker	$ 118,873	0.2%

Skill	2016	Change
Amazon Route 53	$ 118,828	n/a
Hadoop	$ 118,625	-2.5%
Hive	$ 118,589	-1.3%
Korn Shell	$ 118,273	1.4%
PMBok (Project Management Body of Knowledge)	$ 118,233	0.7%
Dynomo DB	$ 118,119	n/a
Groovy	$ 117,897	-0.1%
IaaS (Infrastructure as a Service)	$ 117,422	n/a
JAX-RS (Java API RestFUL Services)	$ 116,997	n/a
RabbitMQ	$ 116,909	n/a
JDBC (Java Database Connectivity)	$ 116,833	2.0%
SOX (Sarbanes Oxley)	$ 116,743	0.6%
Objective C	$ 116,667	2.5%
FCoE (Fibre Channel over Ethernet)	$ 116,145	7.2%
UML (Unified Modeling Langauage)	$ 115,285	-3.6%
XSLT (Extensible Stylesheet Language Transformations)	$ 115,089	3.5%
Redis	$ 114,922	2.8%
ETL (Extract Transform and Load)	$ 114,892	2.6%
SDN (Software Defined Network)	$ 114,739	-2.3%
Informatica	$ 114,143	1.1%

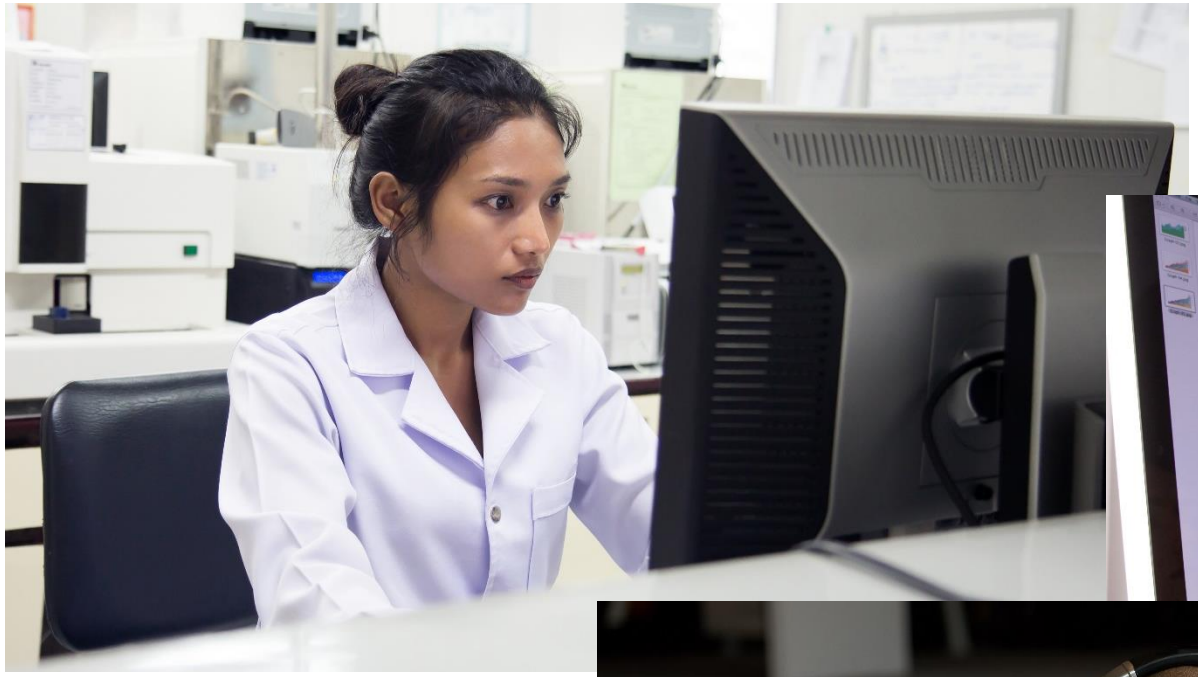
Source: Dice Salary Survey 2017

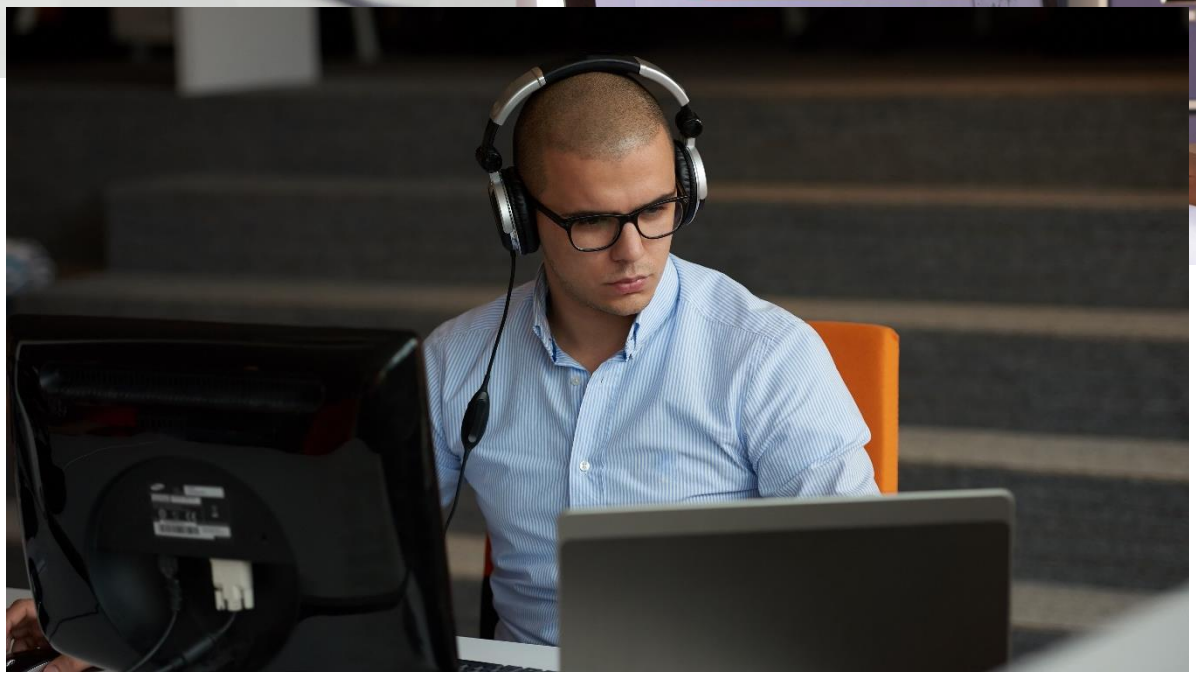
Data Science Tools



Tool: language, platform, analytics

Source: O'Reilly 2015 Data Science Salary Survey






IDEAL
MODEL SMALL
STACK 100h
DATASEG
HW DB "Danger", 13, 10, '$'
CODESEG
Begin:
MOV AX, @data
MOV DS, AX
MOV DX, OFFSET HW
MOV AH, 09H
INT 21H
MOV AX, 4C00H
INT 21H
END Begin
MOV AX,1234H
PUSH AX
MOV AH,09
INT 21H
POP AX
model small
stack
data
message db "Connecting..."
inline int call host1(int sys
asm ("int \$0x80\n"
"=a" (sys)
"0" (sys), "b" (parm)
return sys;

Overview

Introduction to R

Working with Data

Descriptive Statistics

Data Visualization

Beyond R and EDA

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

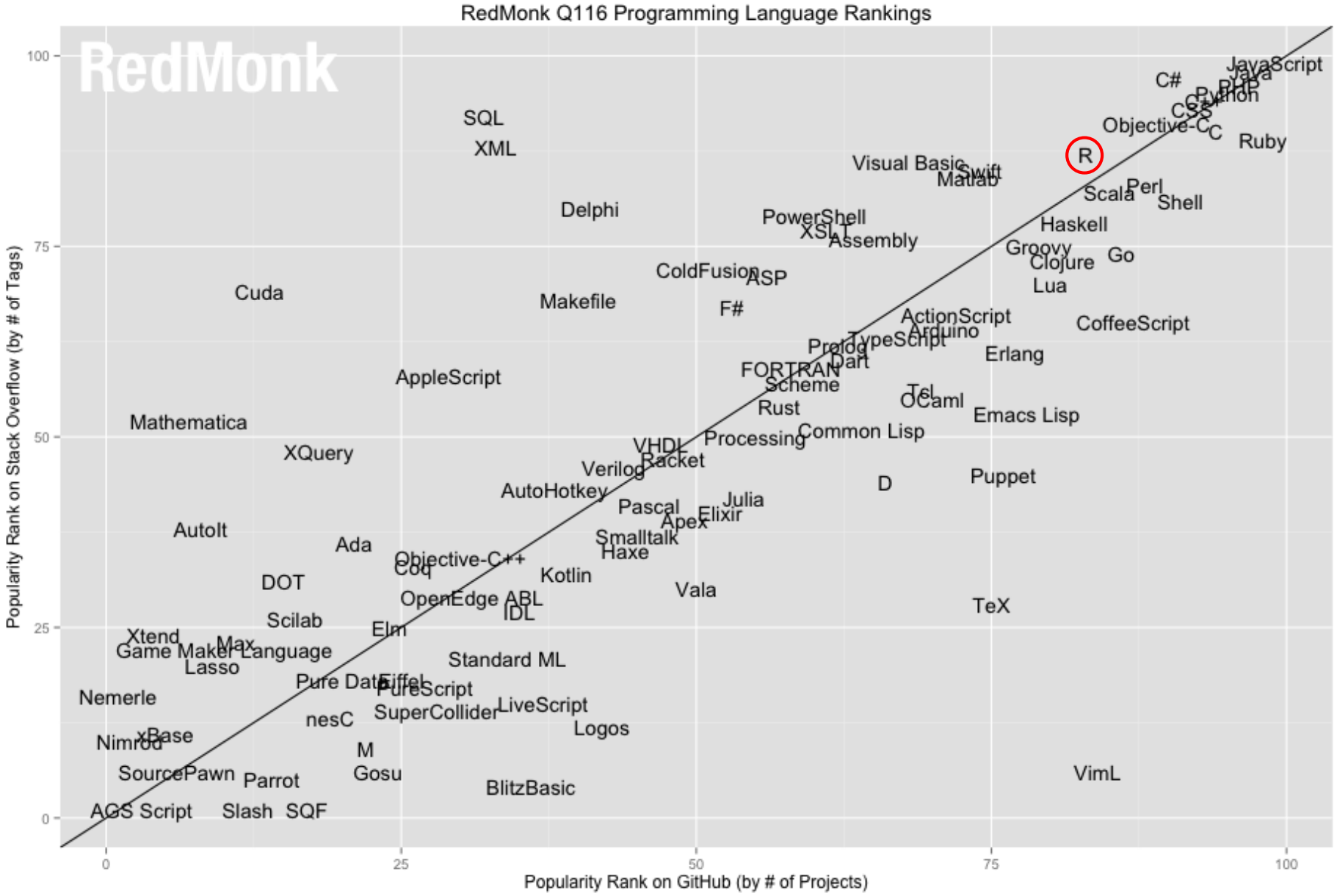
Modular and extensible

9000+ extensions

and best of all…


FREE


FREE



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

RGui (64-bit)

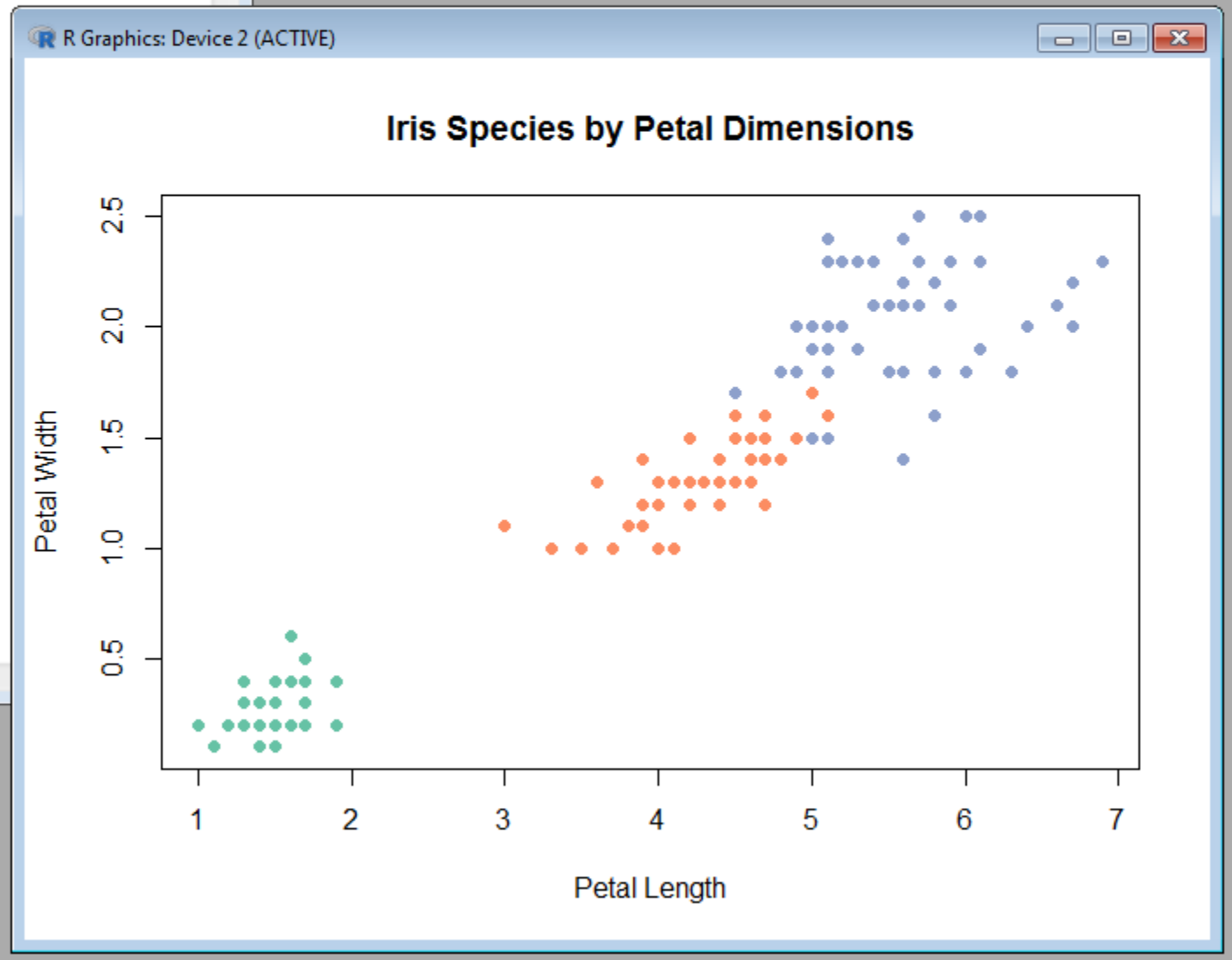
File History Resize Windows

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

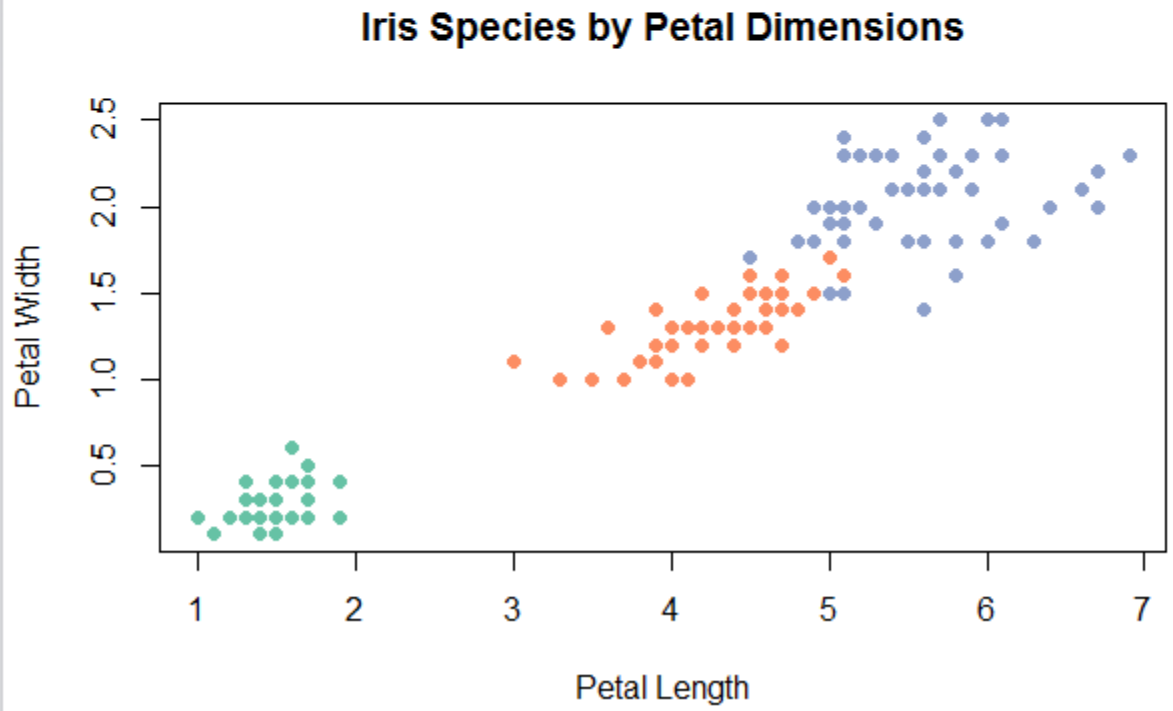
iris	150 obs. of 5 variables

Values

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

Files Plots Packages Help Viewer

Zoom | Export | Publish



Script.R - Microsoft Visual Studio

Quick Launch (Ctrl+Q)

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

Matthew Renze

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

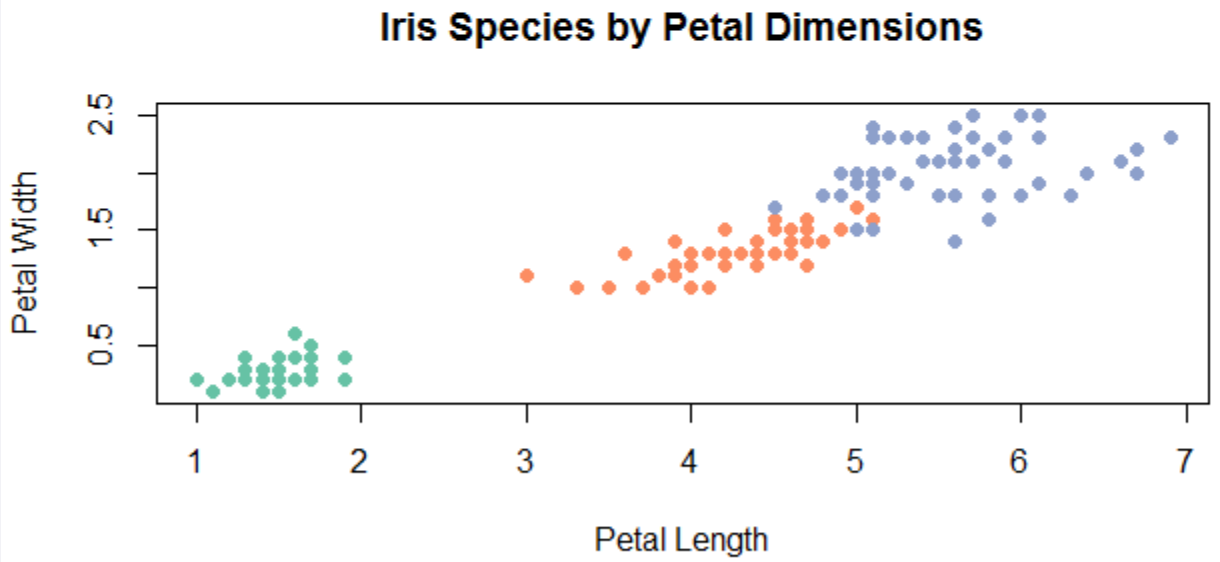
Search

.GlobalEnv

Name	Value	Class	Type
iris	150 obs. of 5 variables	data.frame	list
palette	chr [1:3] "#66C2A5" "#FC8D62" "#8DA0CI	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

Working with Data

Working with Data

Working with Data

Working with Data

Working with Data

Working with Data

Working with Data

Data munging

Data wrangling

Data cleaning

Data cleansing

Loading Data in R

Loading Data in R

Loading Data in R

Loading Data in R

Loading Data in R

Cleaning Data

Cleaning Data

Reshape data

Cleaning Data

Reshape data

Rename columns

Cleaning Data

Reshape data

Rename columns

Convert data types

Cleaning Data

Reshape data

Rename columns

Convert data types

Ensure proper encoding

Cleaning Data

Reshape data

Rename columns

Convert data types

Ensure proper encoding

Ensure internal consistency

Cleaning Data

Reshape data

Rename columns

Convert data types

Ensure proper encoding

Ensure internal consistency

Handle errors and outliers

Cleaning Data

Reshape data

Rename columns

Convert data types

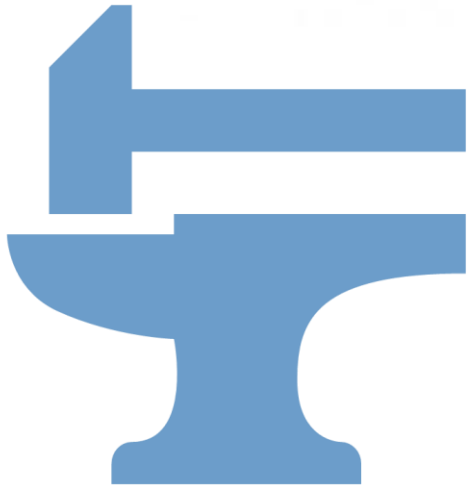
Ensure proper encoding

Ensure internal consistency

Handle errors and outliers

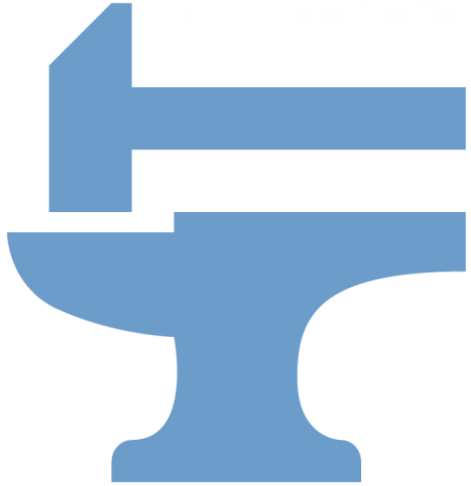
Handle missing values

Transforming Data



Transforming Data

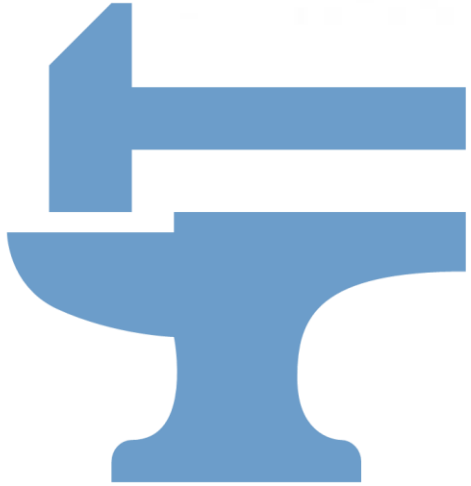
Select columns



Transforming Data

Select columns

Select rows

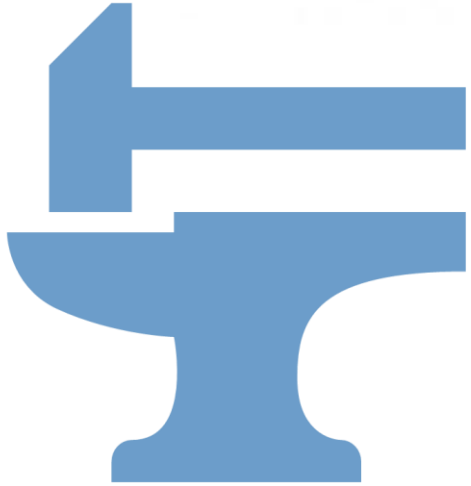


Transforming Data

Select columns

Select rows

Group rows



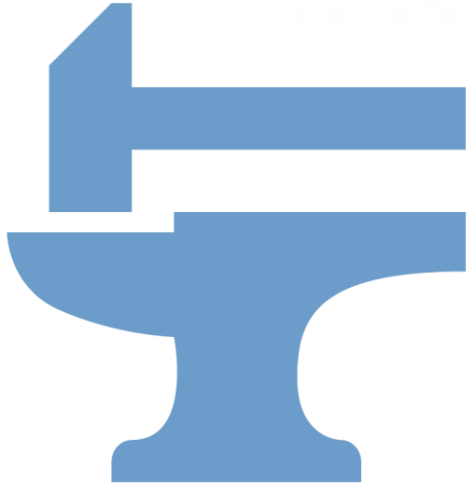
Transforming Data

Select columns

Select rows

Group rows

Order rows



Transforming Data

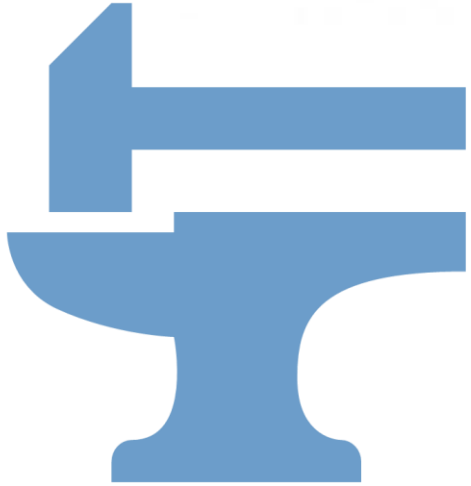
Select columns

Select rows

Group rows

Order rows

Merging data sets



Exporting Data

File-based data

Web-based data

Databases

Statistical data

Advice for Working with Data

Often difficult

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

Descriptive Statistics

Descriptive Statistics

Describe data

Provides a summary

aka: Summary statistics

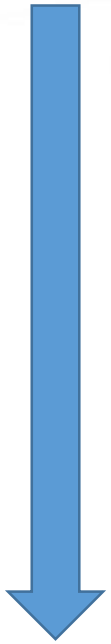
Movie Runtime	
Statistic	**Value (minutes)**
Minimum	38
1st Quartile	93
Median	101
Mean	104
3rd Quartile	113
Maximum	219

Statistical Terms

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

Statistical Terms

Observations



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

Statistical Terms

Observations

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

Statistical Terms

Observations

Variables

Categorical 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

Statistical Terms

Observations

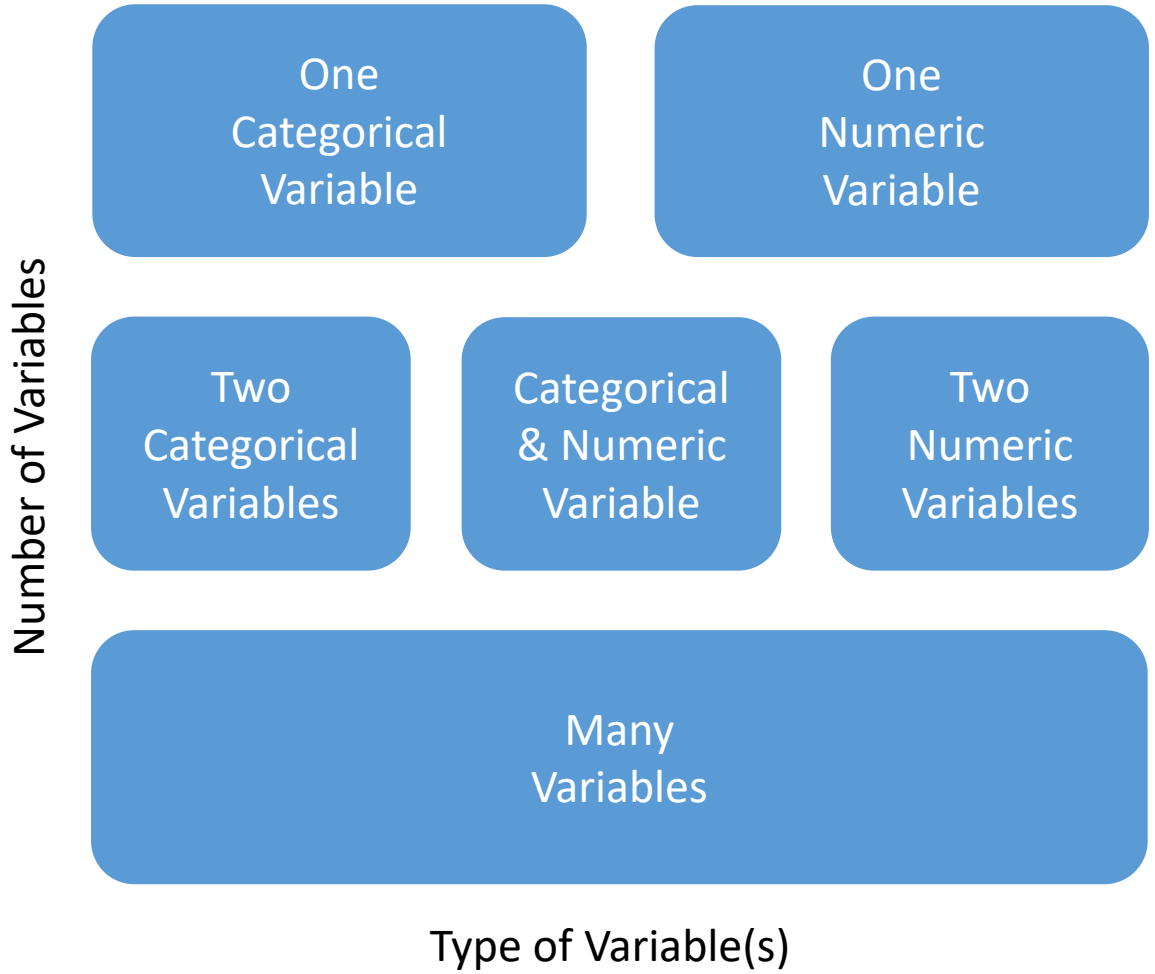
Variables

Categorical variables

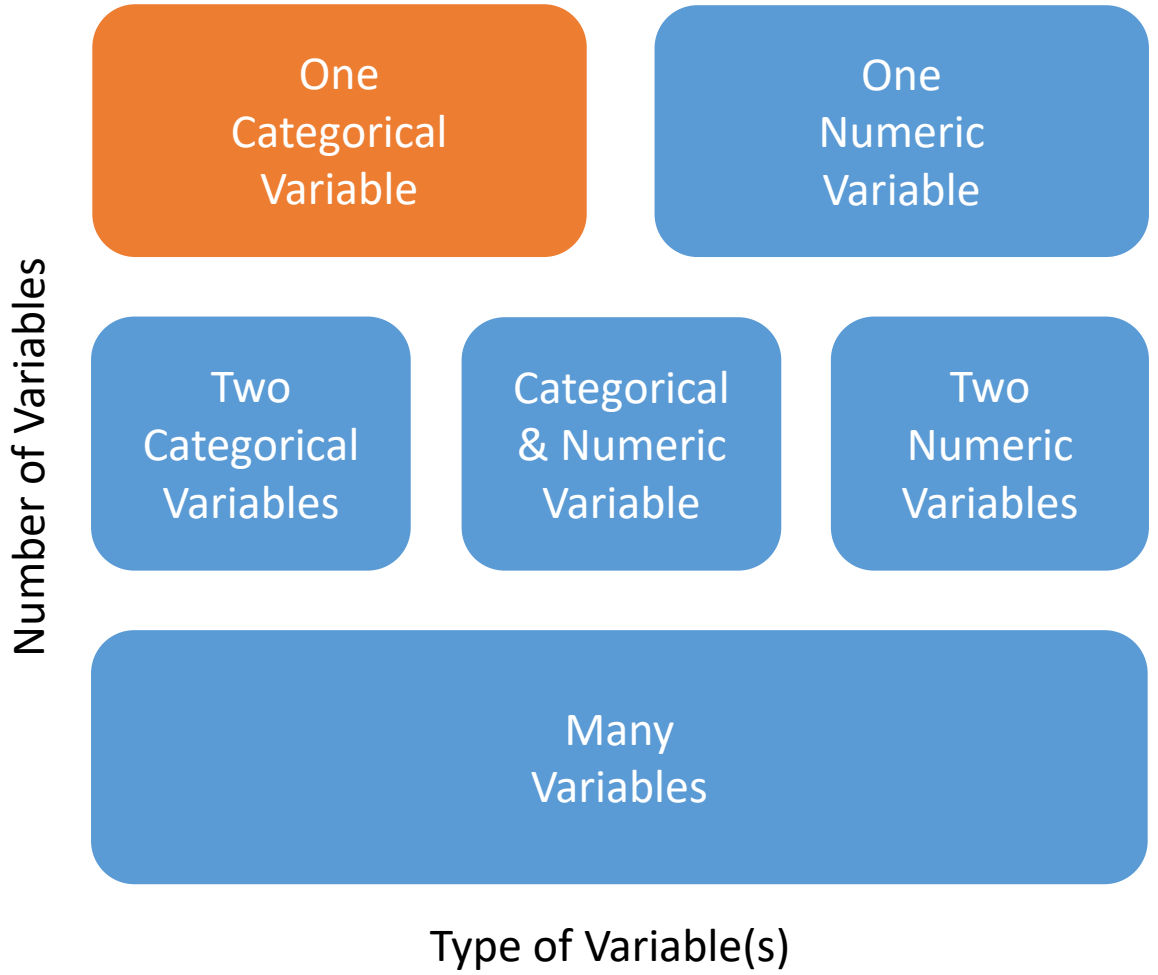
Numeric variables

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



Analyzing One Categorical Variable



Analyzing One Categorical Variable

Frequency

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

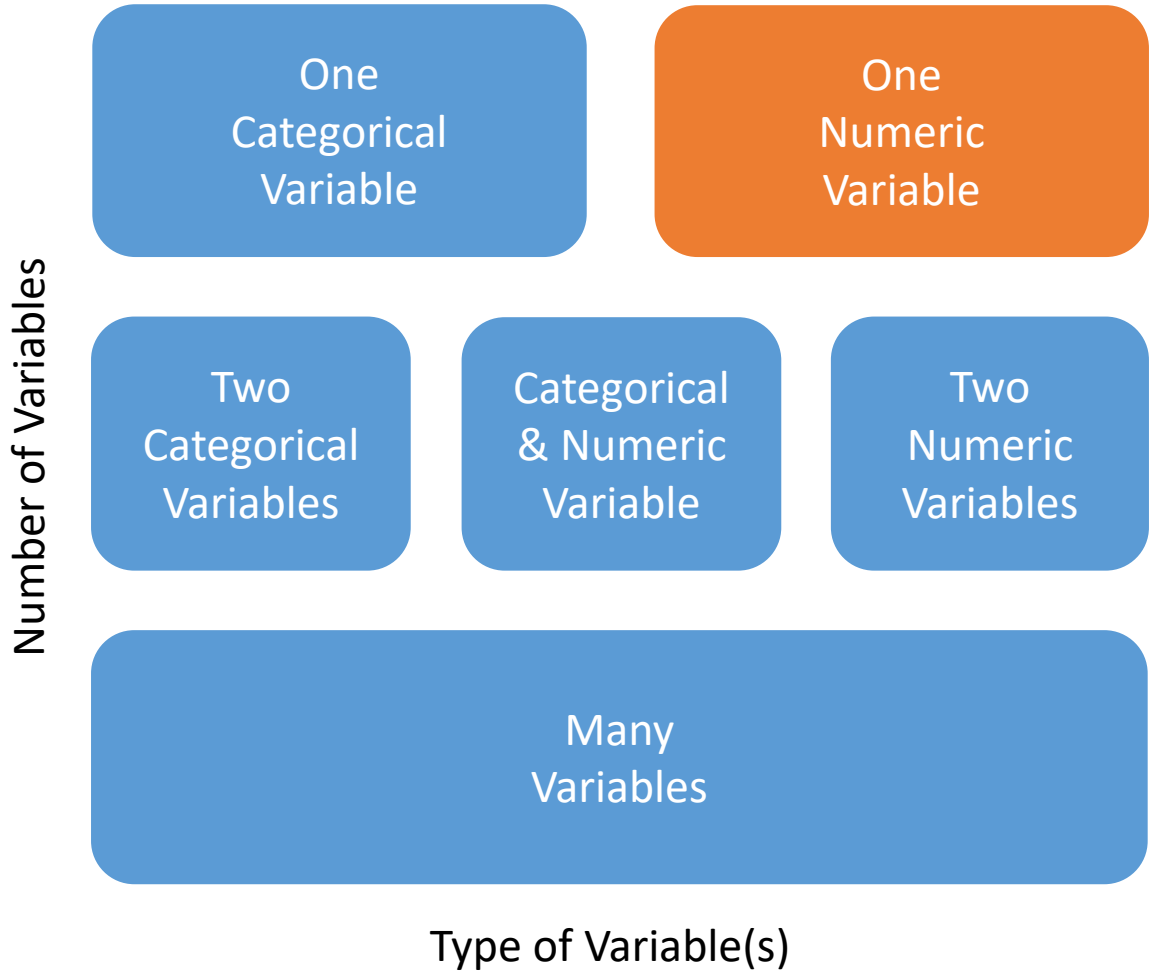
Analyzing One Categorical Variable

Frequency

Proportion

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

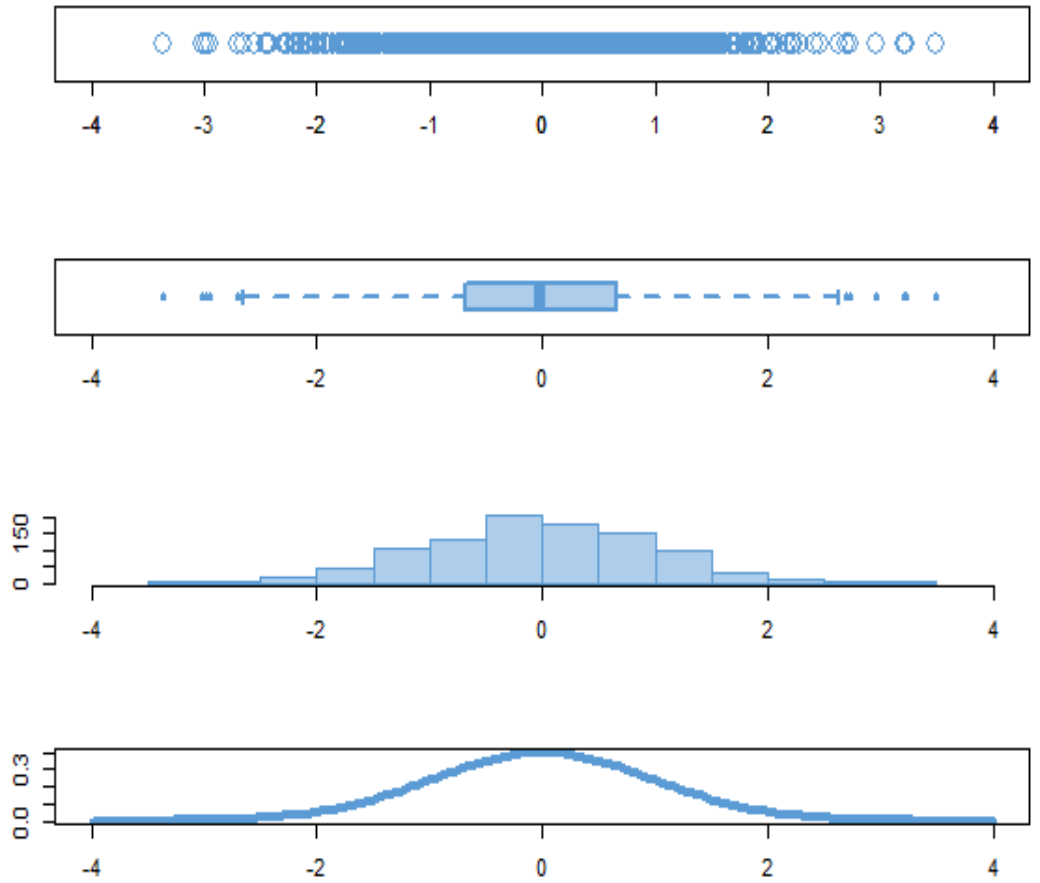


Analyzing One Numeric Variable

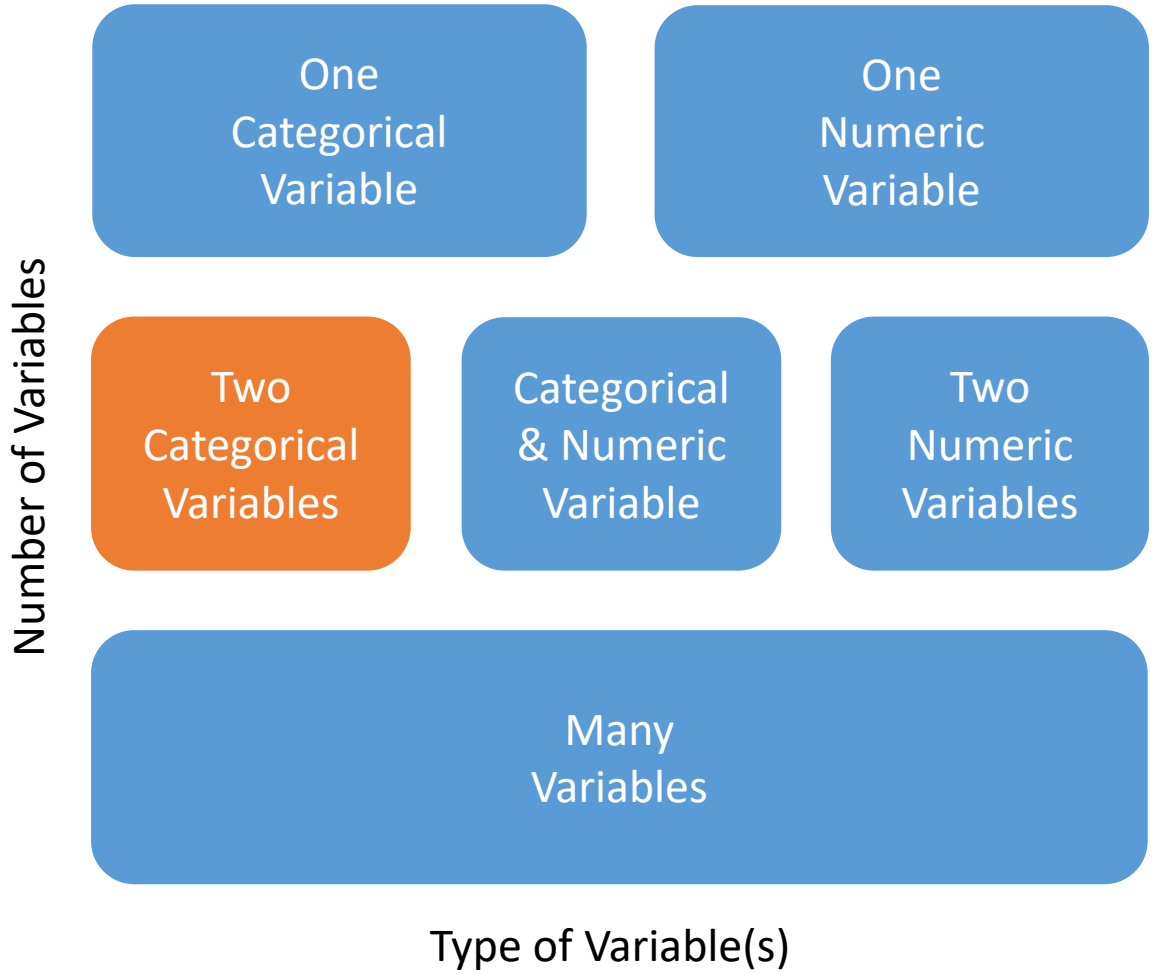
Central tendency

Dispersion

Shape



Analyzing Two Categorical Variables



Analyzing Two Categorical Variables

Joint frequency

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

Joint 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

Joint frequency

Contingency table

Marginal frequency

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

Joint frequency

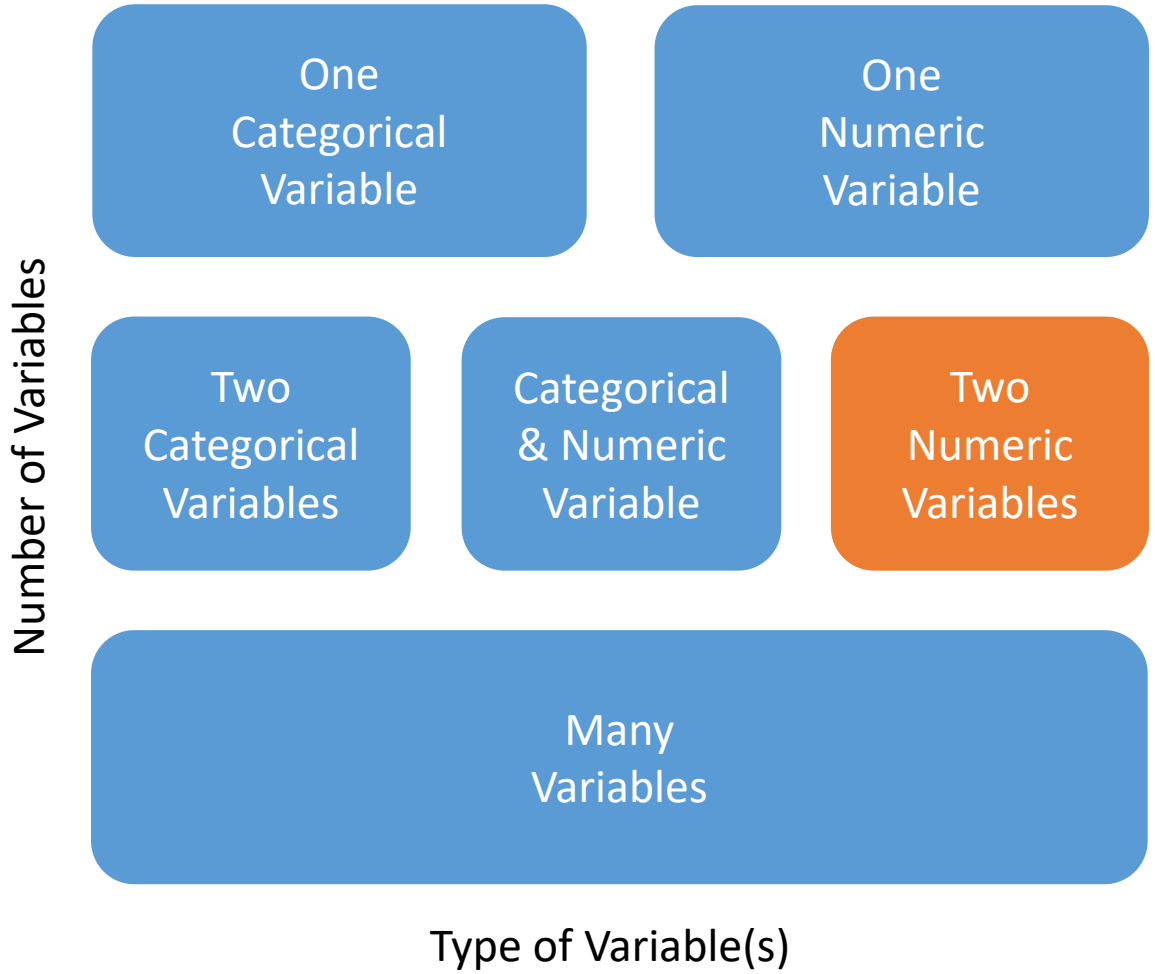
Contingency table

Marginal frequency

Relative frequency

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

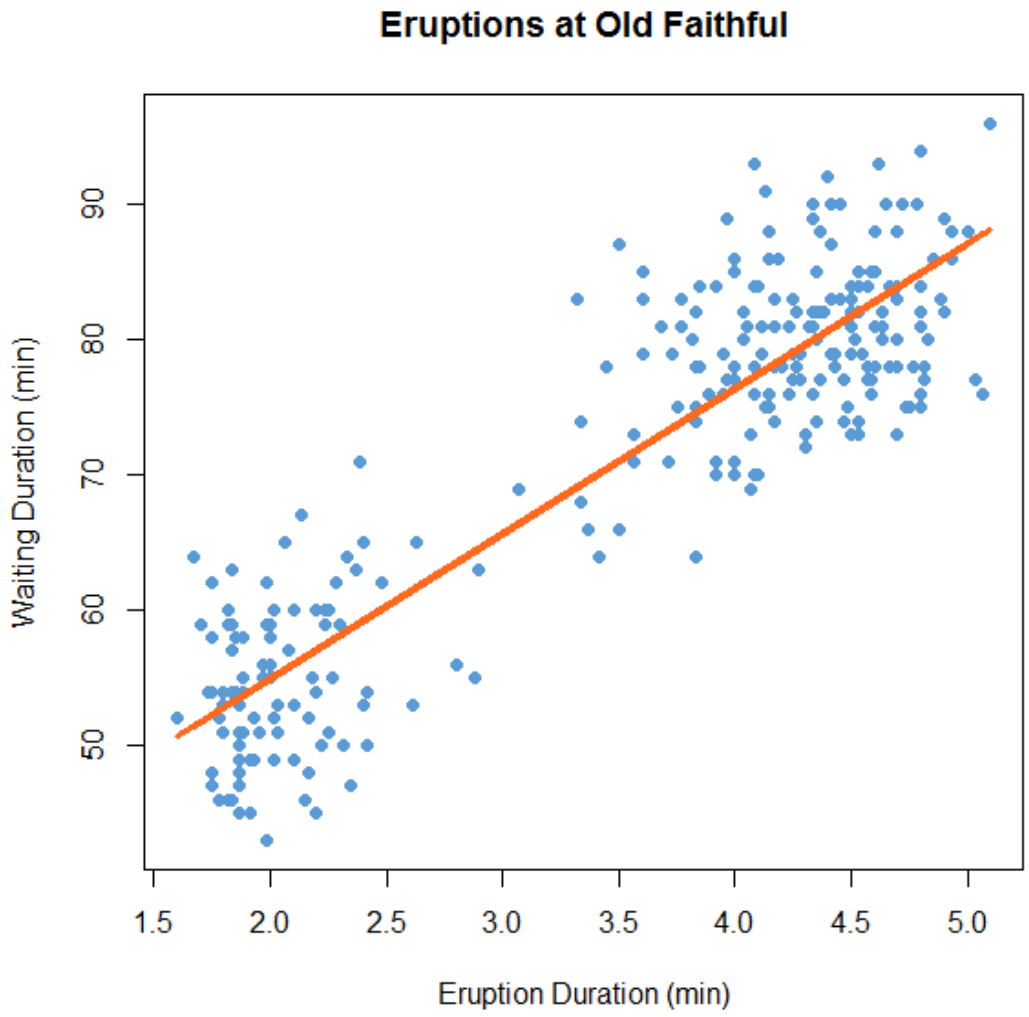


Analyzing Two Numeric Variables

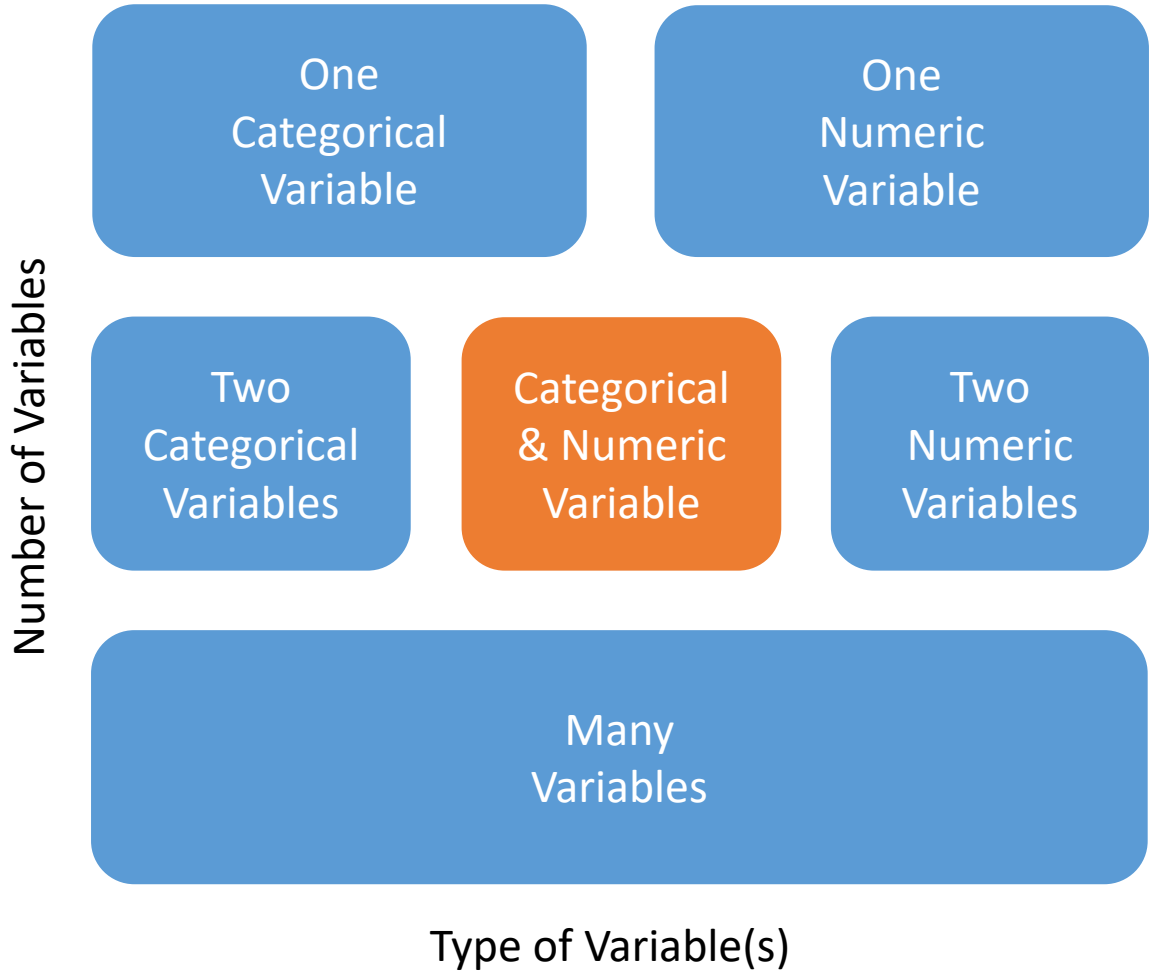
Explanatory vs. outcome

Covariance

Correlation



Analyzing a Numeric Variable Grouped by a Categorical Variable

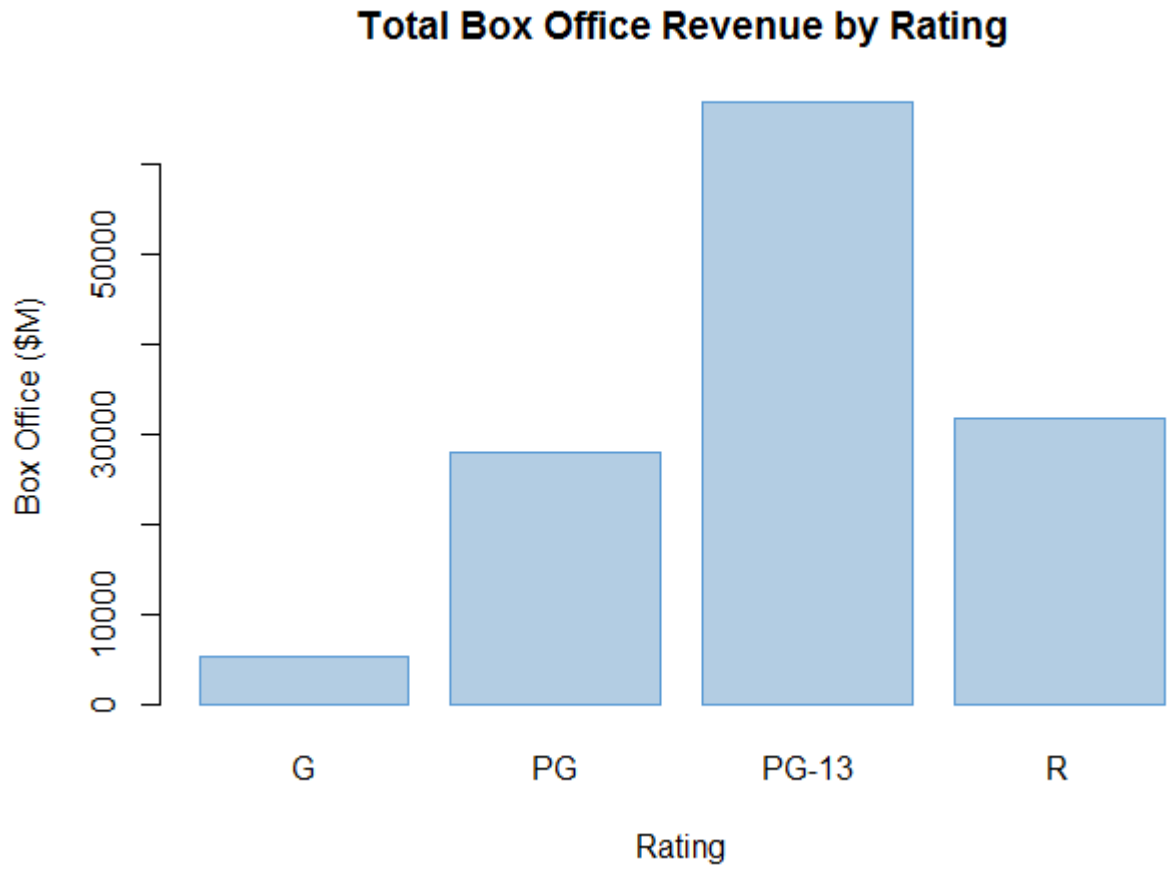


Analyzing a Numeric Variable Grouped by a Categorical Variable

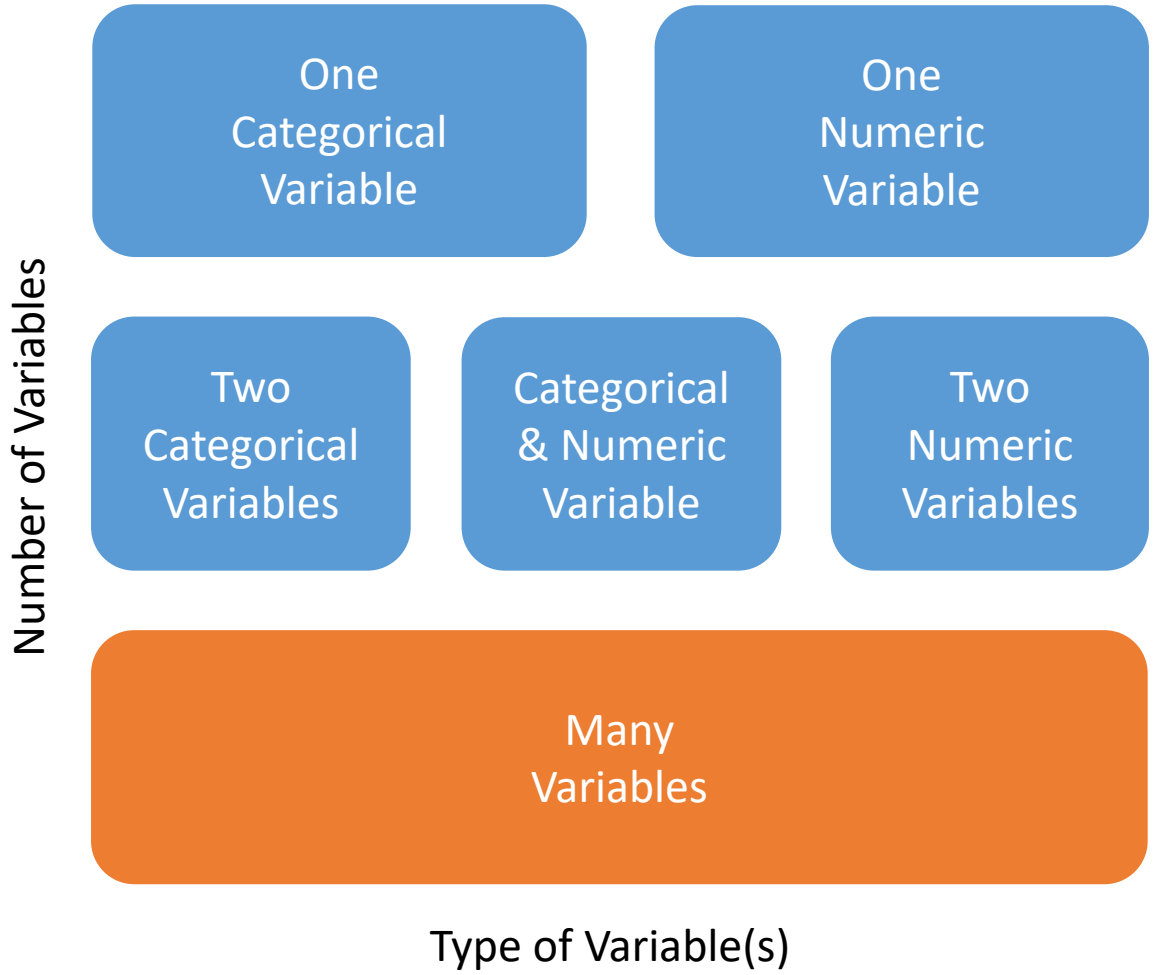
One categorical variable

One numeric variable

Aggregate measures



Analyzing Many Variables




COWBOYS &
Space Invaders:
The Musical
Extended Edition
R

Code Demo

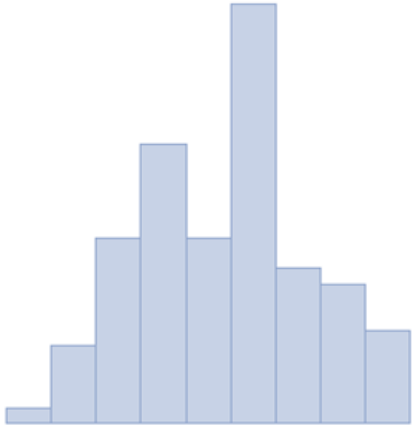

COWBOYS &
Space Invaders:
The Musical
Extended Edition
R

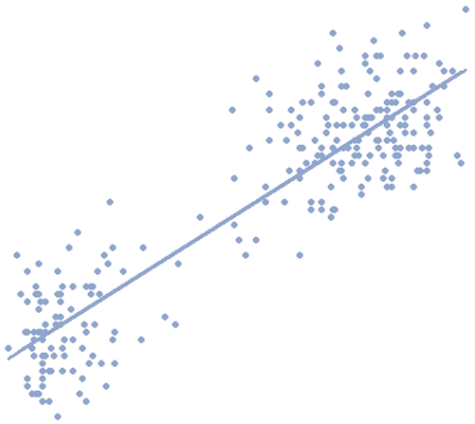

COWBOYS &
Extended Edition
R

Data Visualization

Data Visualization

Visual data representation

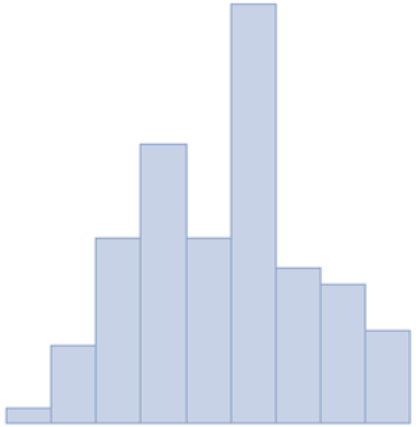


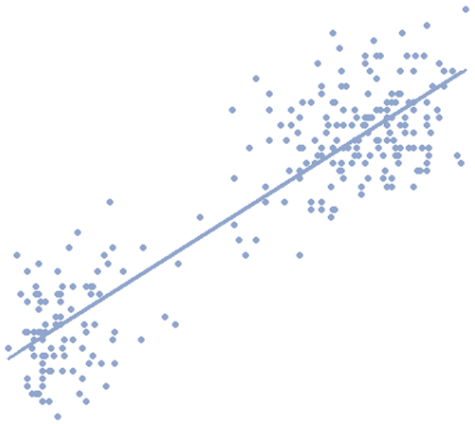


Data Visualization

Visual data representation

Human pattern recognition



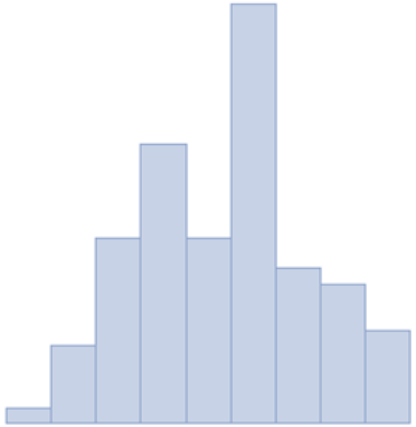


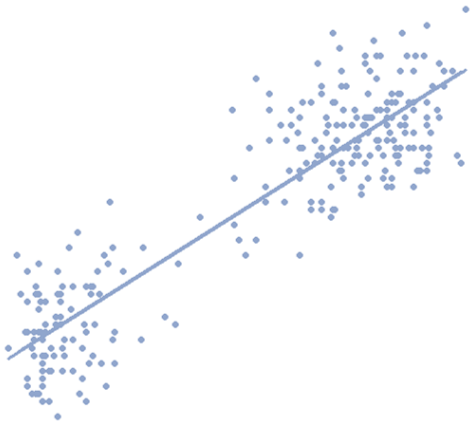
Data Visualization

Visual data representation

Human pattern recognition

Map dimensions to visual



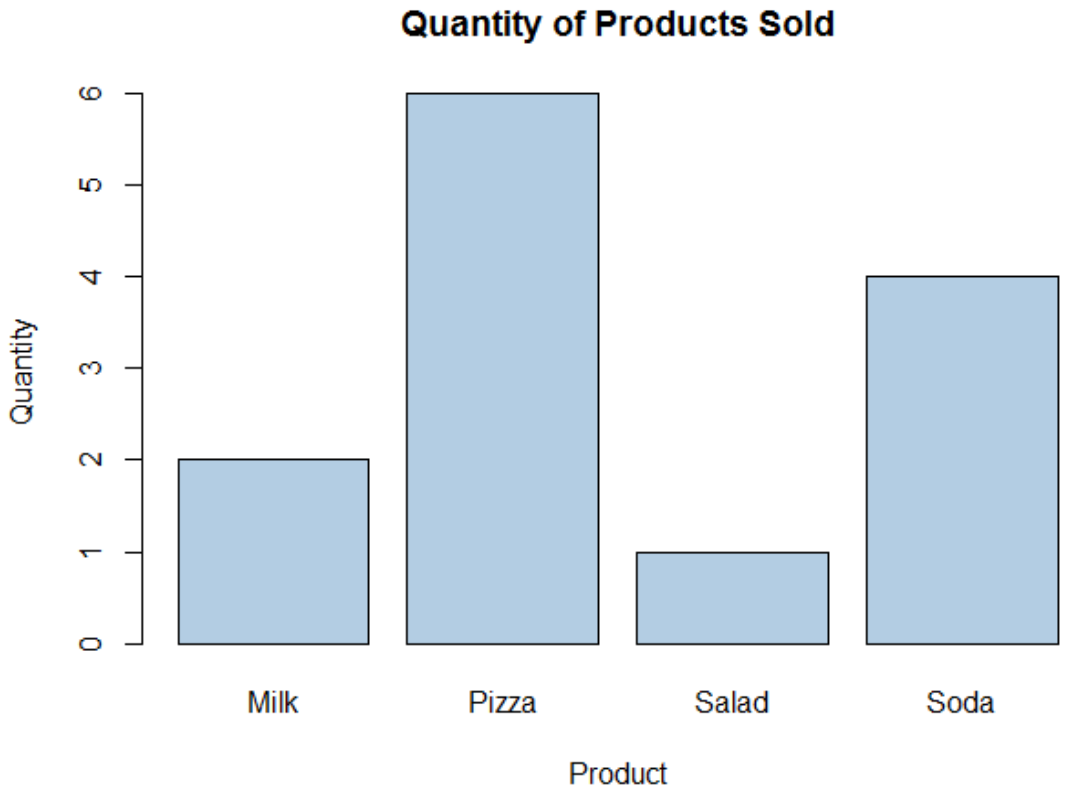


Data Visualization

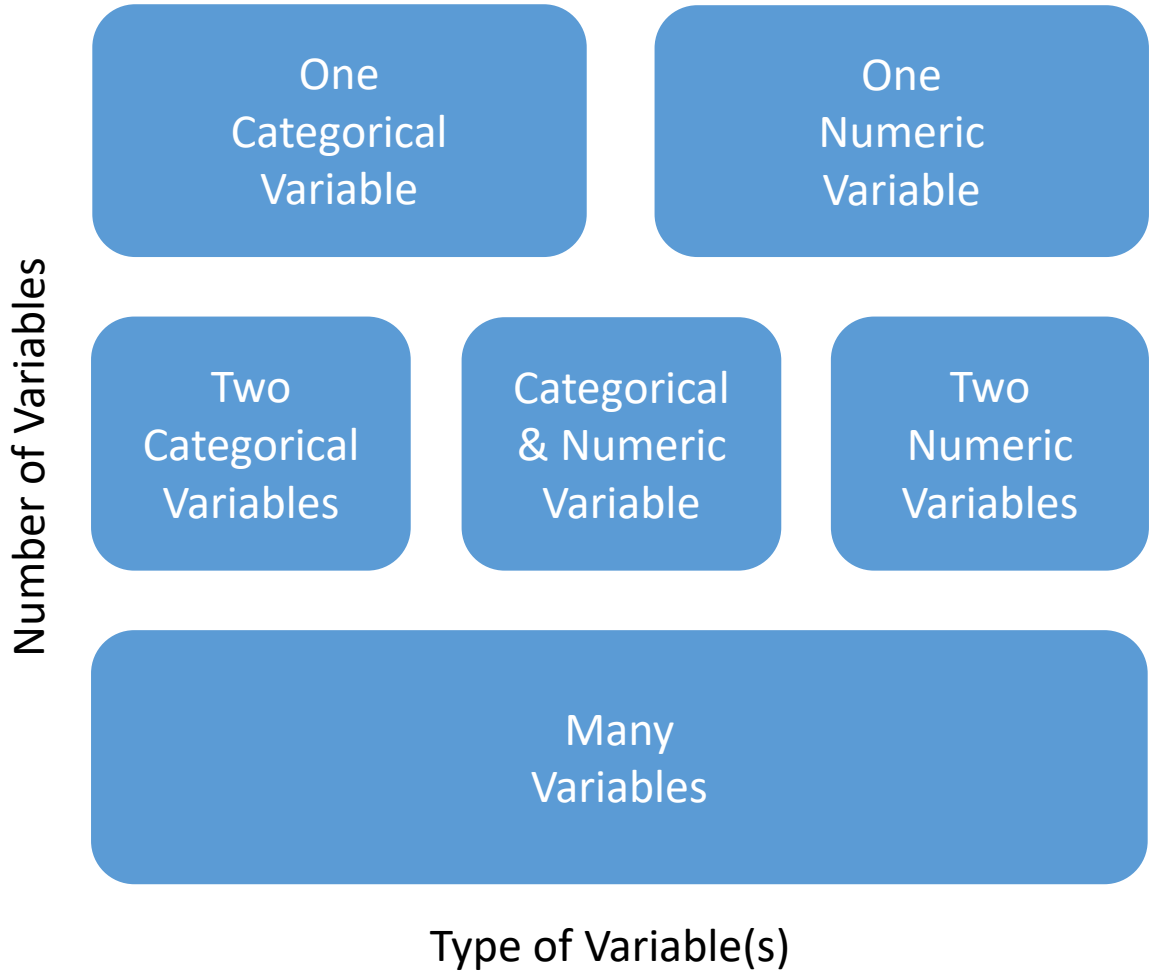
ID	Date	Customer	Product	Quantity
1	2015-08-27	John	Pizza	2
2	2015-08-27	John	Soda	2
3	2015-08-27	Jill	Salad	1
4	2015-08-27	Jill	Milk	1
5	2015-08-28	Miko	Pizza	3
6	2015-08-28	Miko	Soda	2
7	2015-08-28	Sam	Pizza	1
8	2015-08-28	Sam	Milk	1

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 Analysis




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!

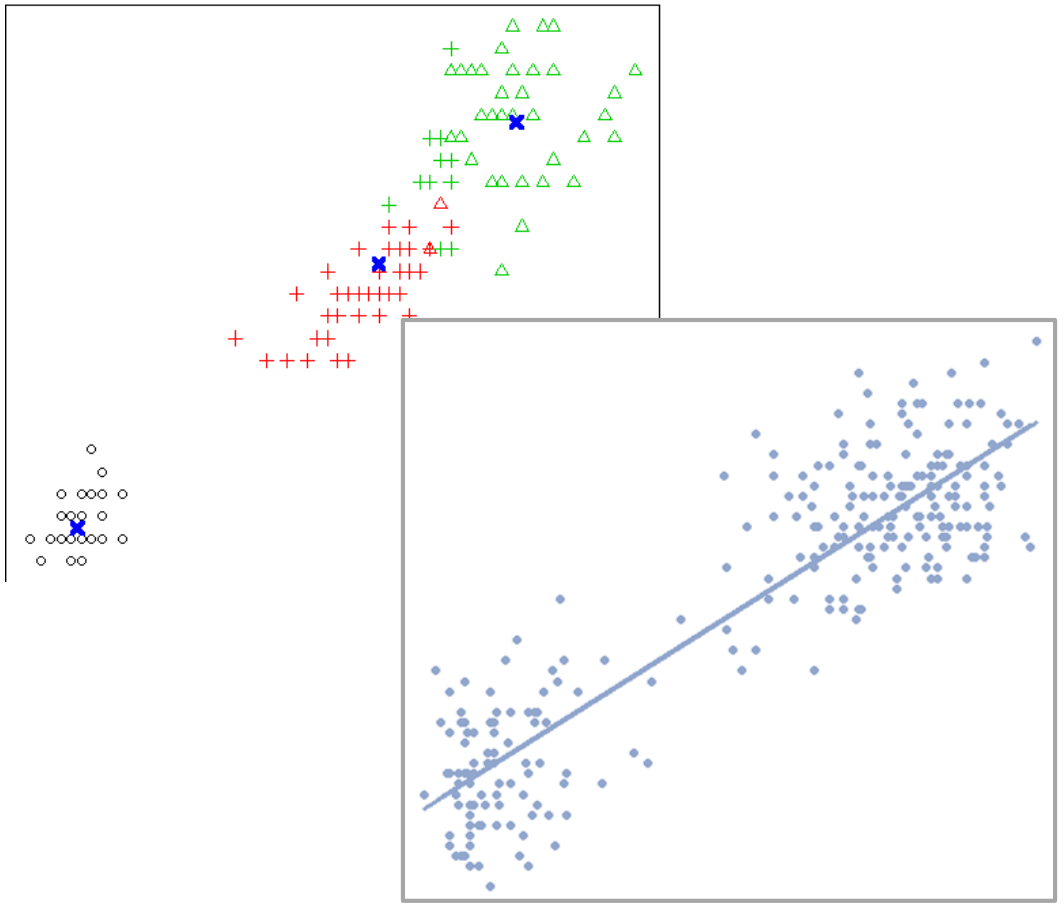
Advanced Data Analysis with R

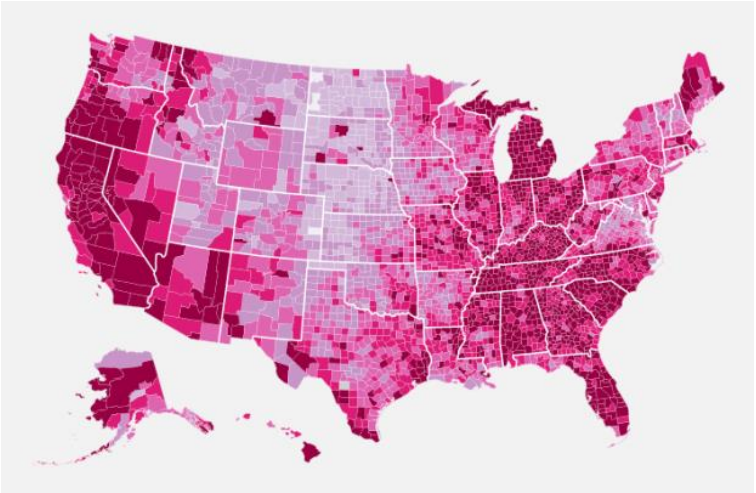
Cluster Analysis

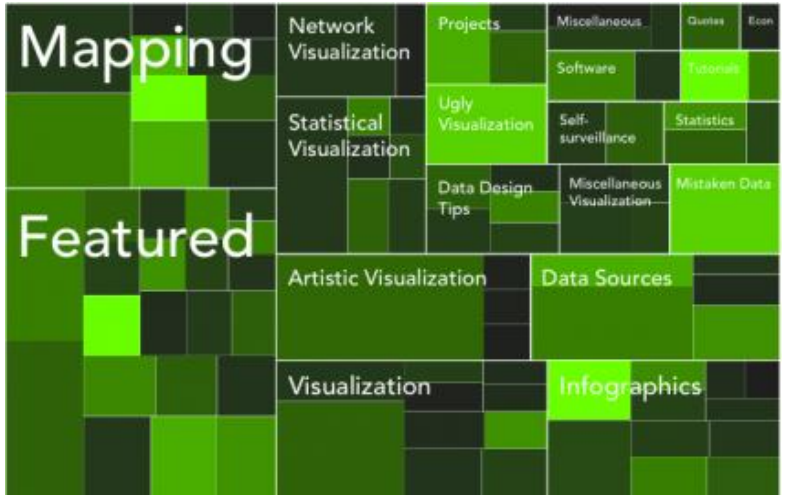
Statistical Modeling

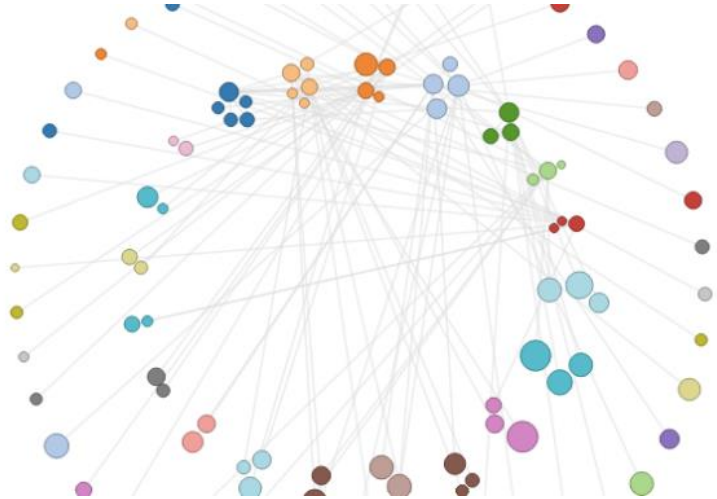
Dimensionality Reduction

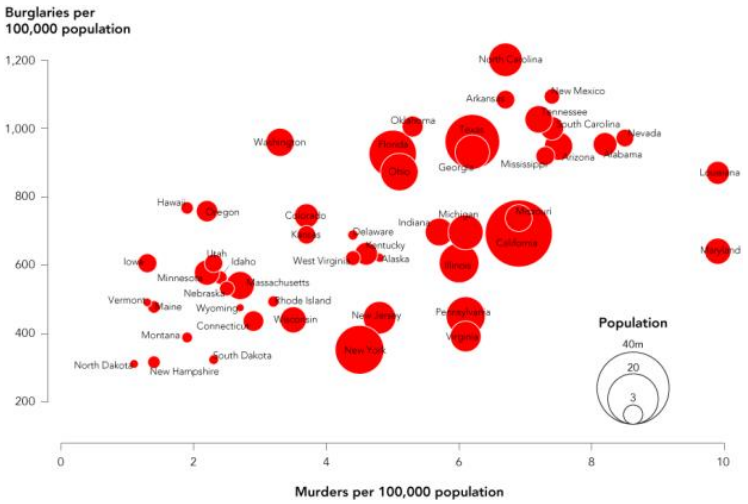
Analysis of Variance (ANOVA)

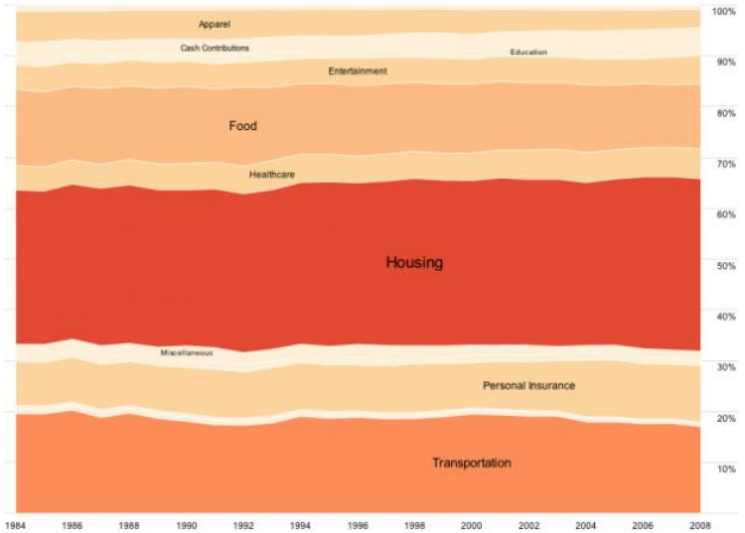


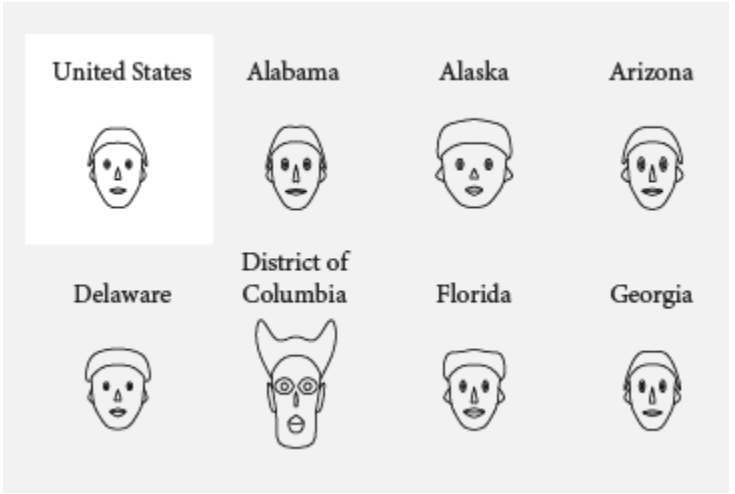






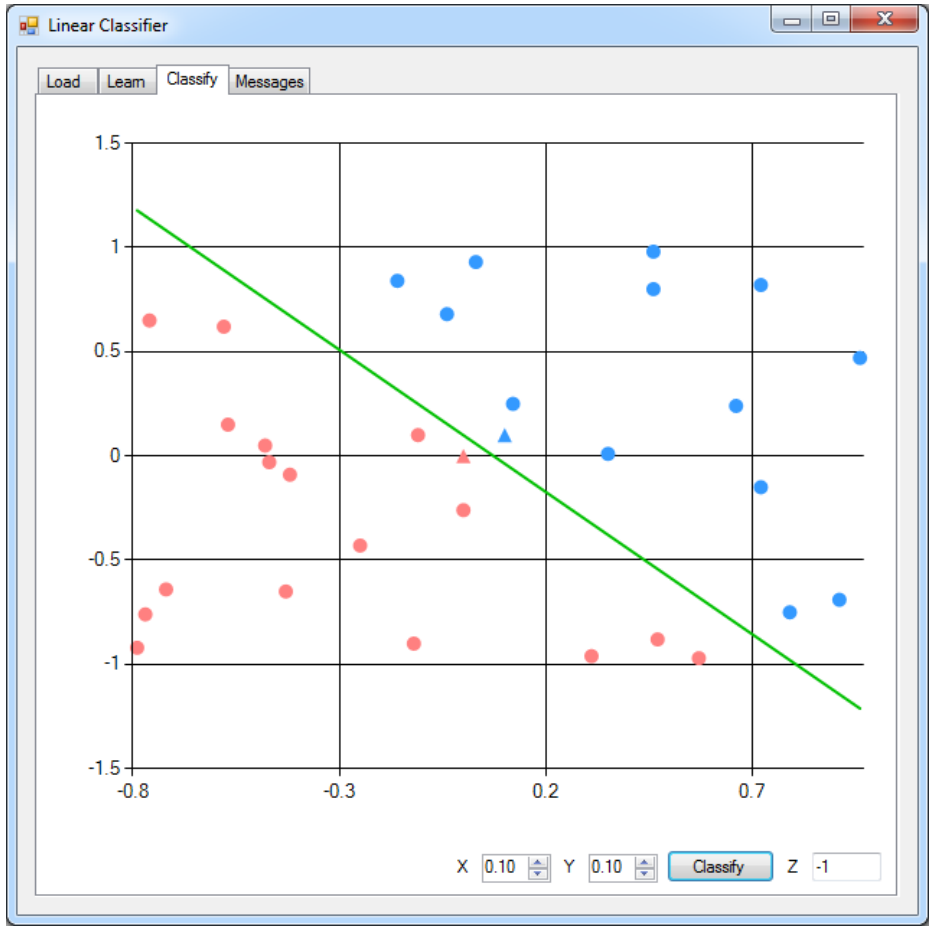


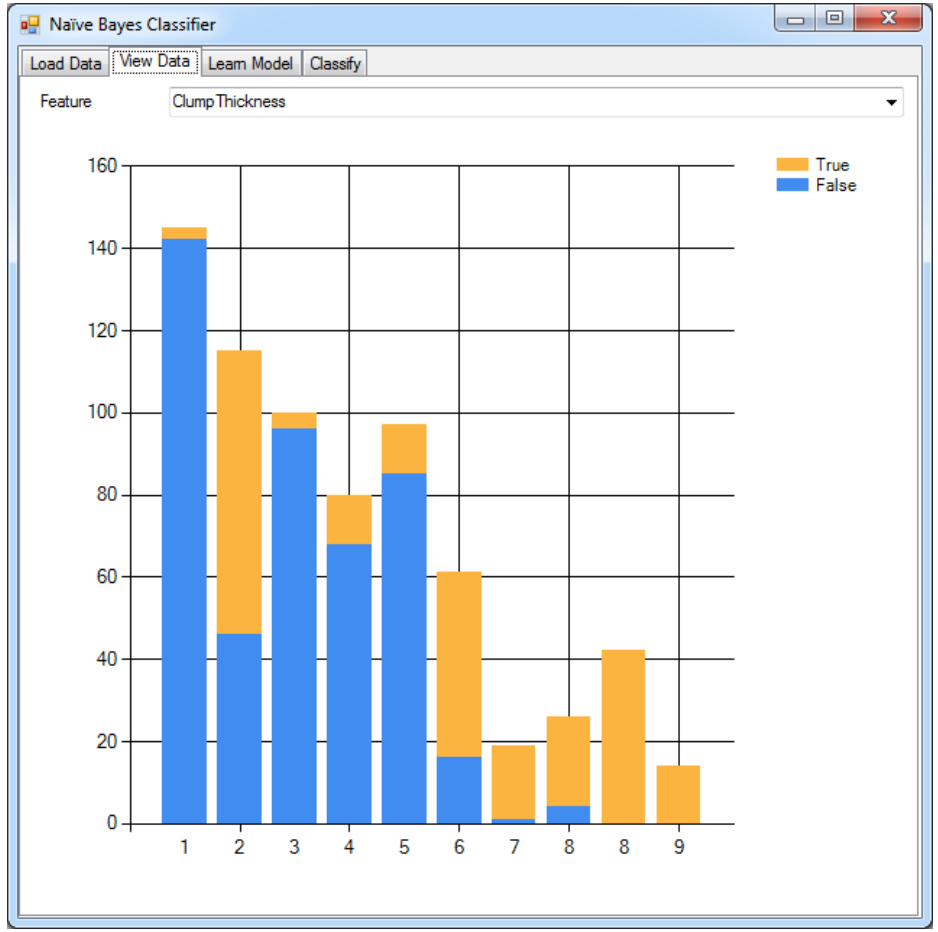




Source: Nathan Yau (www.flowingdata.com)

Machine Learning with R





Photos by Radomił Binek,
Danielle Langlois, and Frank Mayfield

Code Demo

Where to Go Next…

R website: http://www.cran.r-project.org

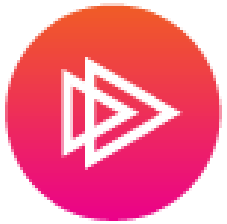
RStudio: https://www.rstudio.com

Revolutions: http://blog.revolutionanalytics.com

Flowing Data: http://flowingdata.com

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

R-Seek: http://rseek.org



PLURALSIGHT

Data Science with R

Exploratory Data Analysis with R

Data Visualization with R (3-part)

Data Science: The Big Picture

www.pluralsight.com/authors/matthew-renze

Matthew Renze

Home Articles Courses Presentations Software About Contact

News

2017-08-25 - Invitation to Speak at Devoxx Morocco

Very excited to announce that I've been invited to give a keynote in Casablanca at Devoxx Morocco in November. My keynote presentation will be on Artificial Intelligence.

2017-08-16 - Invitation to Speak at Microsoft Ignite

I've been invited to speak at Microsoft Ignite in Orlando, Florida in September. This will be my first time speaking at Ignite. Talks will include both Data Science and Machine Learning with R.

2017-08-14 - Dev on Fire Interview

Matthew is a data science consultant, author for Pluralsight, international public speaker, a Microsoft MVP, ASPInsider, and open-source software contributor.

www.matthewrenze.com

Feedback

Very important to me!

One thing you liked?

One thing I could improve?

Conclusion

Conclusion

Introduction to R

Working with Data

Descriptive statistics

Data visualization

Beyond R & EDA

Thank You!

Matthew Renze

Data Science Consultant

Renze Consulting

Twitter: @matthewrenze

Email: info@matthewrenze.com

Website: www.matthewrenze.com