

# 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 Economist

FEBRUARY 27TH - MARCH 5TH 2010

Economist.com

Gordon Brown's pitch  
What went wrong at RBS  
Genetically modified crops blossom  
The EU woos Russia  
The right to eat cats and dogs

# The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

A man in a suit holds a large green and yellow umbrella over a small flower.

# The New York Times

## For Today's Graduate, Just One Word: Statistics

By STEVE LORIN  
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

**W**hen 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



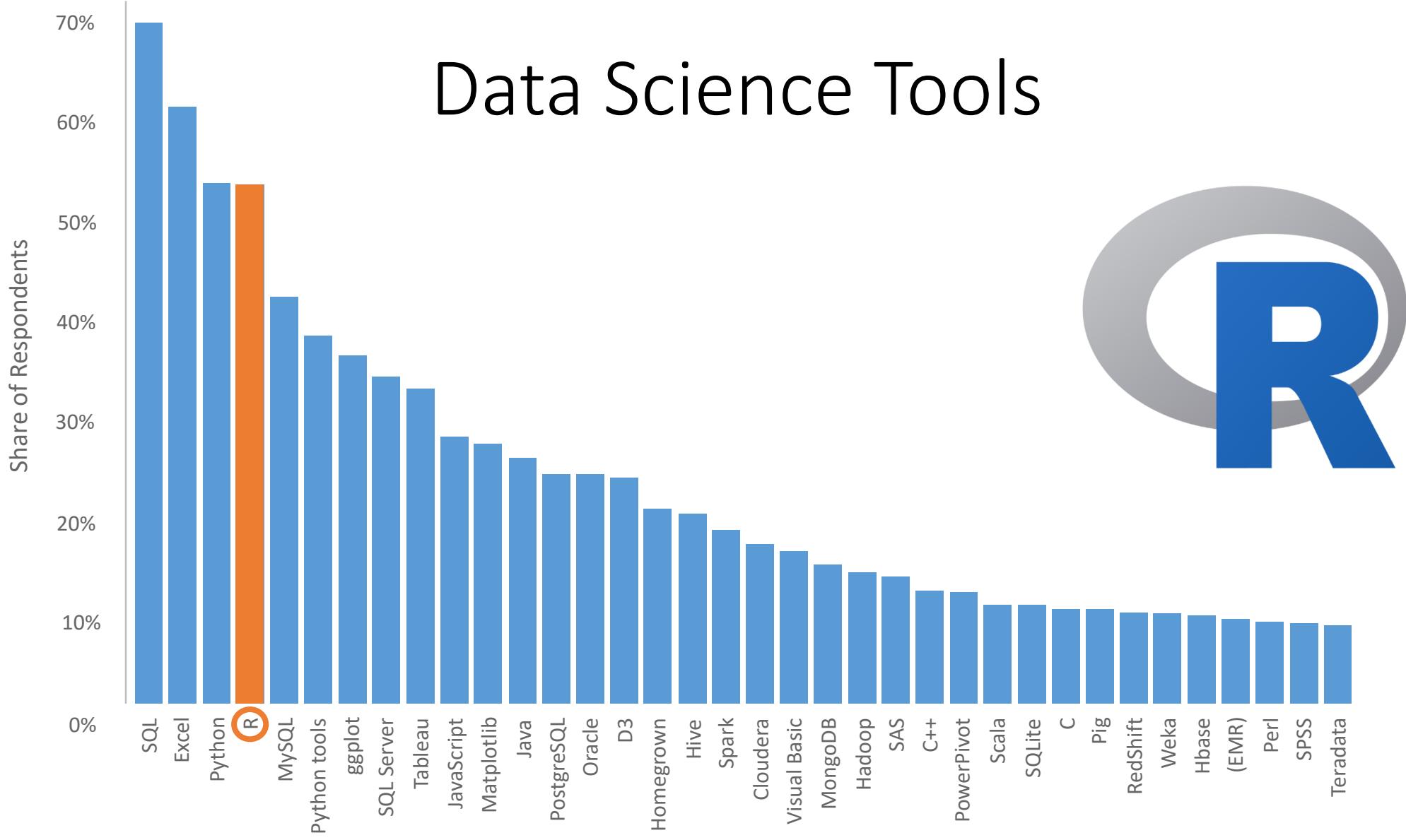
Source: Indeed.com

# Top-paying Tech Skills

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

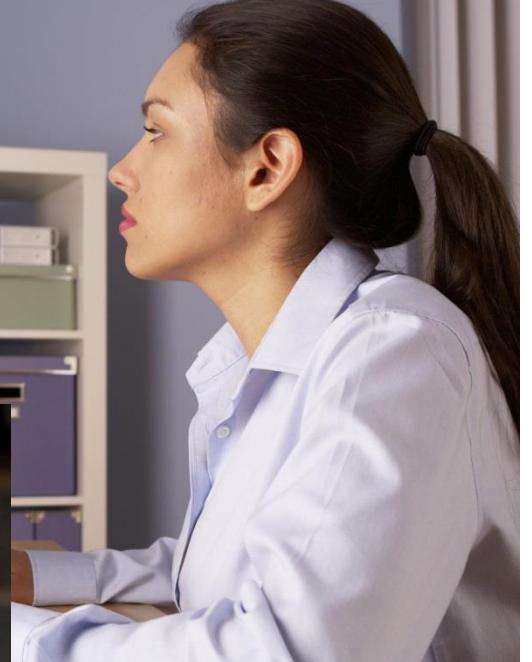
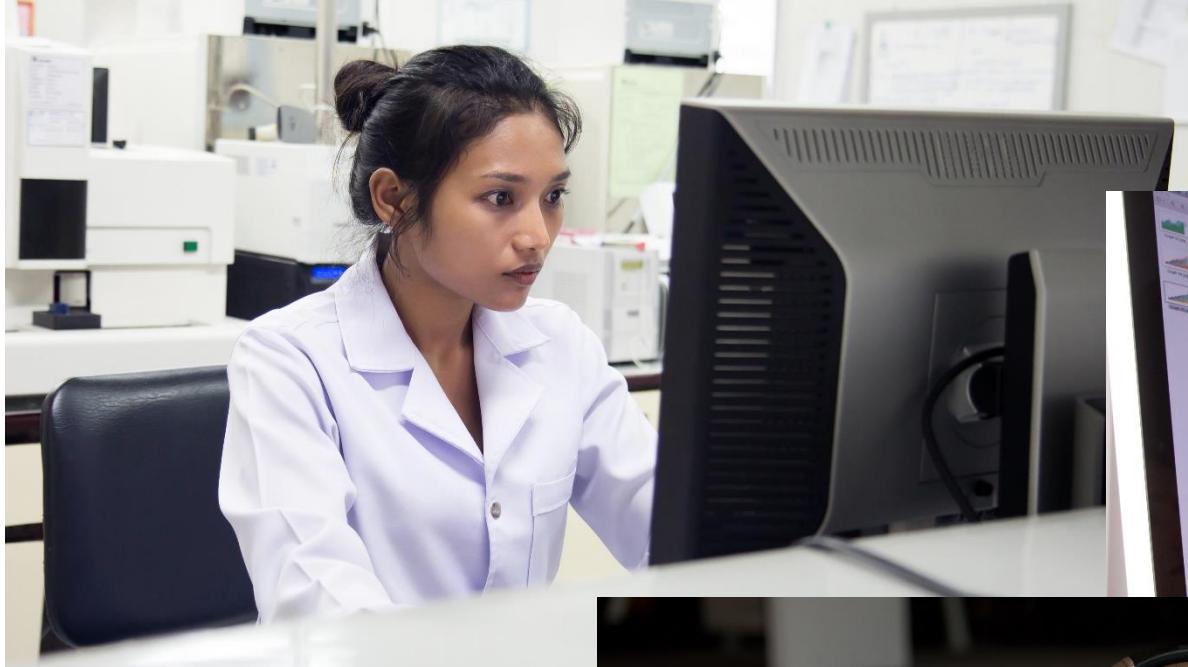
Skill	2016	Change
<b>Amazon Route 53</b>	\$ 118,828	n/a
<b>Hadoop</b>	\$ 118,625	-2.5%
<b>Hive</b>	\$ 118,589	-1.3%
<b>Korn Shell</b>	\$ 118,273	1.4%
<b>PMBok</b> (Project Management Body of Knowledge)	\$ 118,233	0.7%
<b>Dynamo DB</b>	\$ 118,119	n/a
<b>Groovy</b>	\$ 117,897	-0.1%
<b>IaaS</b> (Infrastructure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

Source: Dice Salary Survey 2017



Tool: language, platform, analytics

Source: O'Reilly 2015 Data Science Salary Survey







# 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...

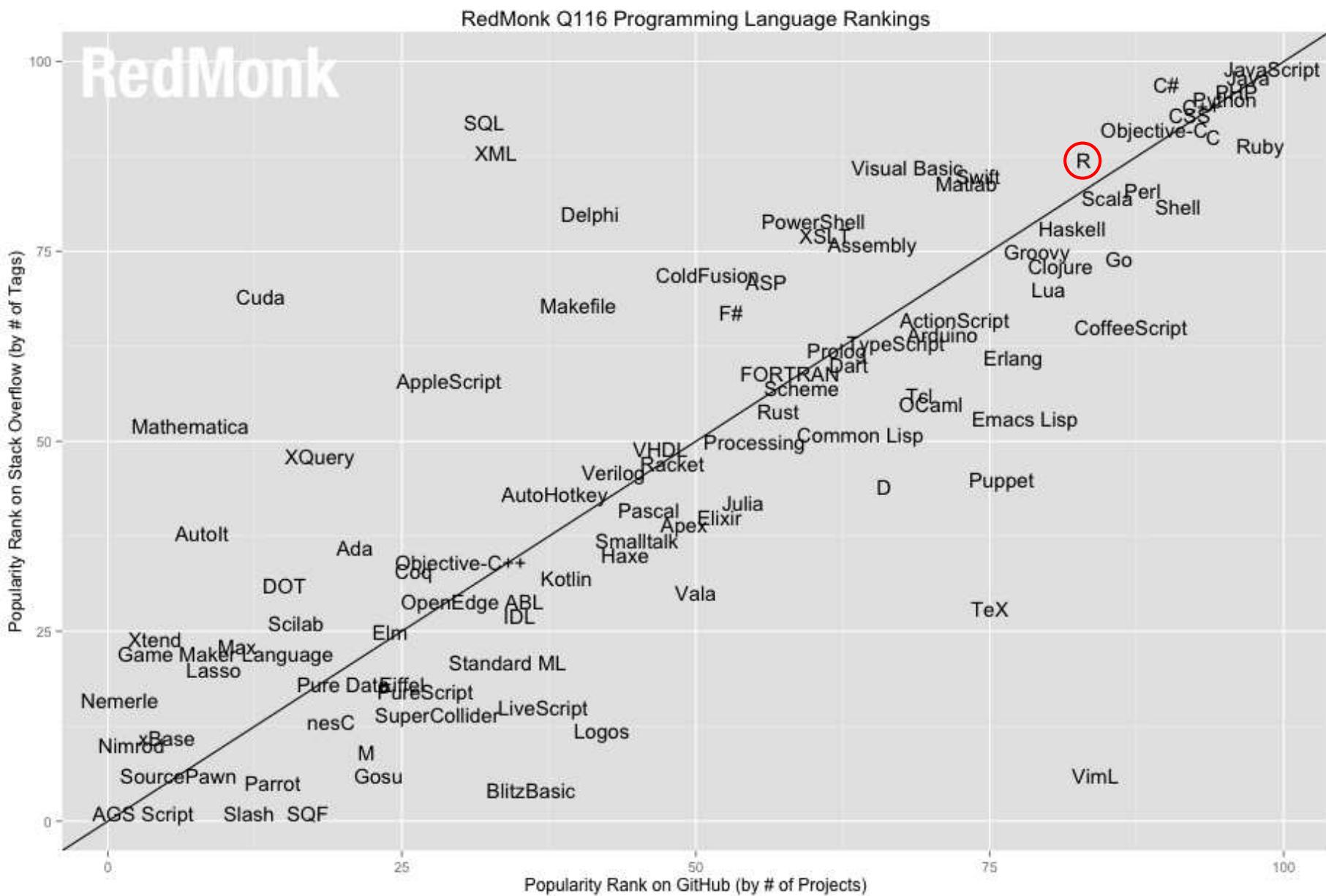


# FREE



A low-angle photograph of the Statue of Liberty against a clear blue sky. Her right arm is raised high, holding a torch aloft. Her left arm is bent, holding a tablet or smartphone that displays the word "FREE".

FREE



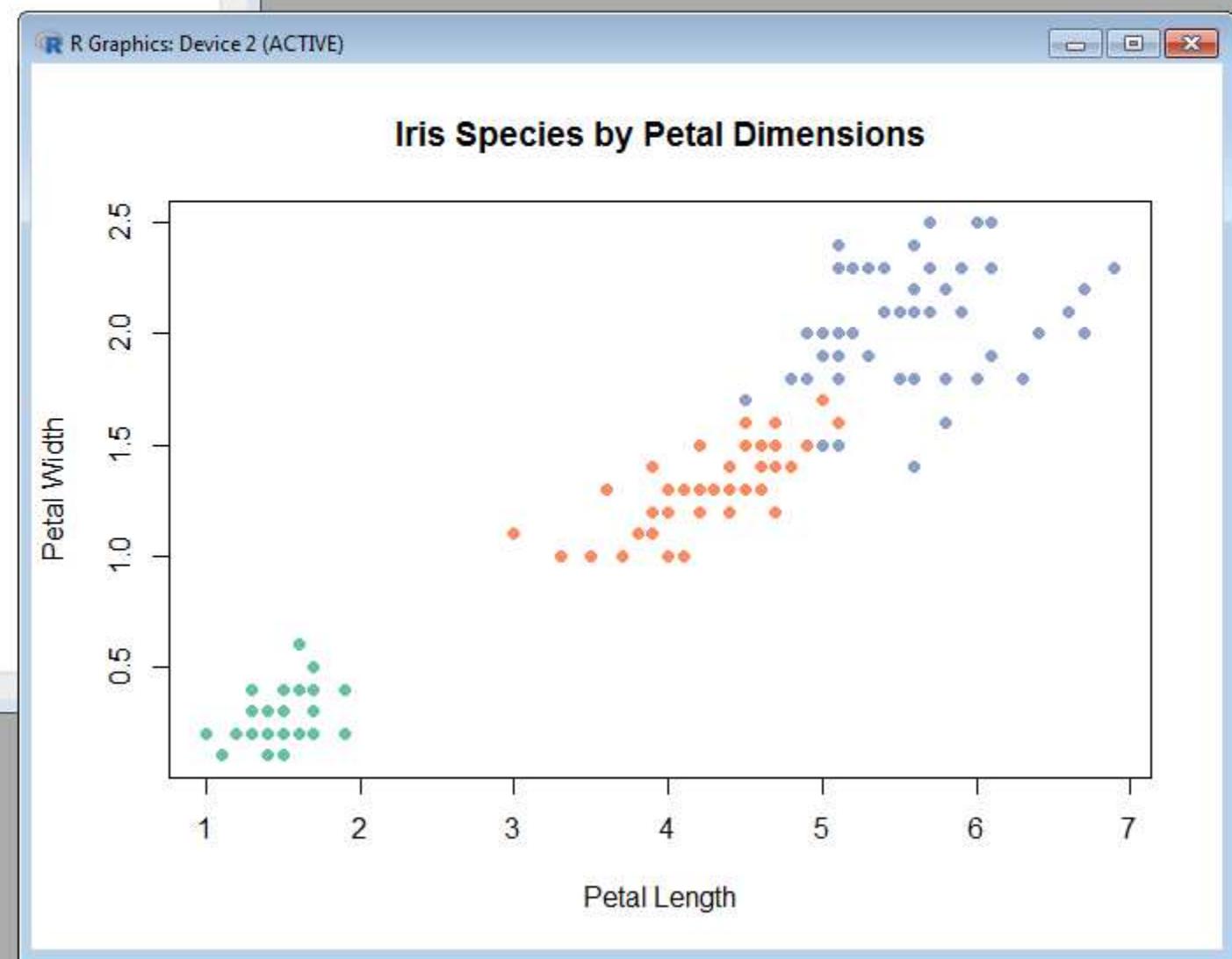


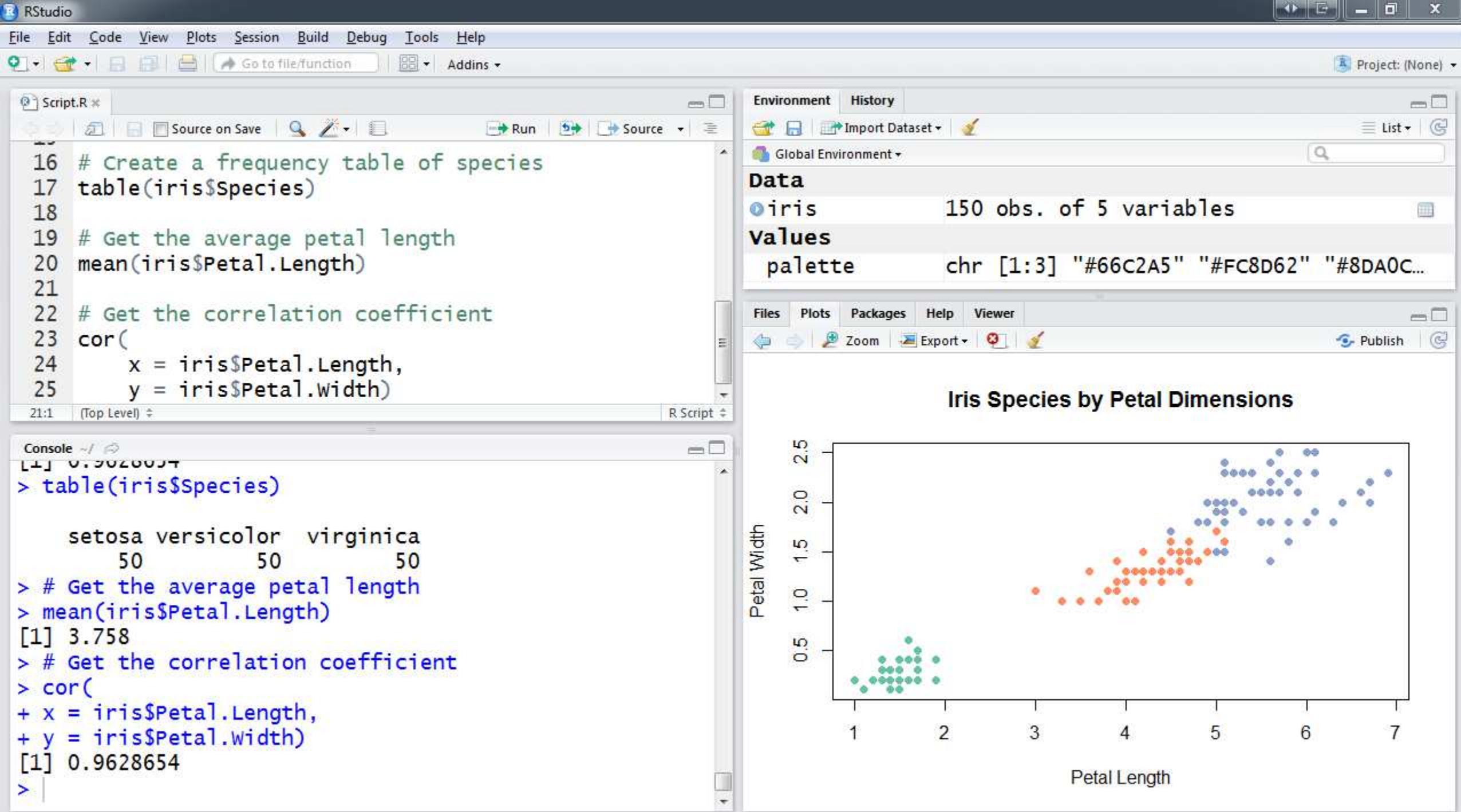
## R Console

```
> # 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
```





Script.R - Microsoft Visual Studio

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

Matthew Renze

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)
```

R Interactive

```
> # 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
>
```

Variable Explorer

Name	Value	Class	Type
iris	150 obs. of 5 variables	data.frame	list
palette	chr [1:3] "#66C2A5" "#FC8D62" "#8DA0CE"	character	character

Variable Explorer R History

R Plot

### Iris Species by Petal Dimensions

Petal Width

Petal Length

Solution Explorer R Plot R Package Manager R Help

Error List Output Azure App Service Activity

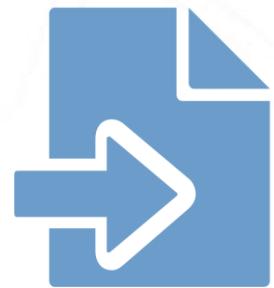
Ready Ln 30 Col1 Ch1 INS ↑ 7 ⌂ 0 ⌂ Root ⌂ master

# Code Demo

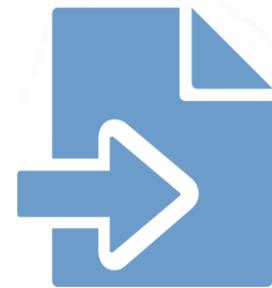
# Working with Data

# Working with Data

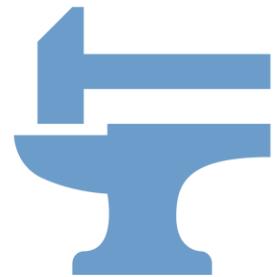
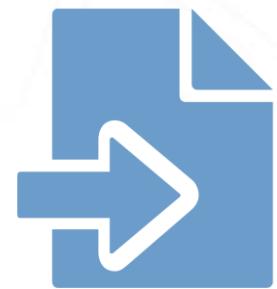
# Working with Data



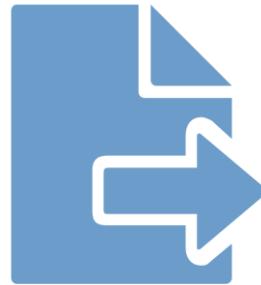
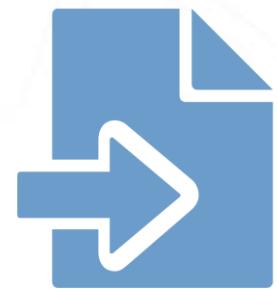
# Working with Data



# Working with Data



# Working with Data



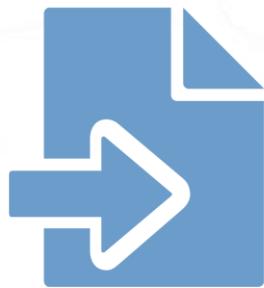
# Working with Data

Data munging

Data wrangling

Data cleaning

Data cleansing

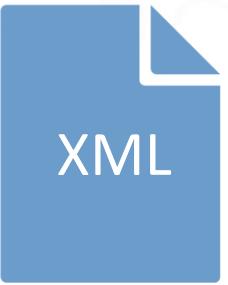


# Loading Data in R

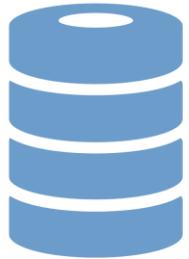
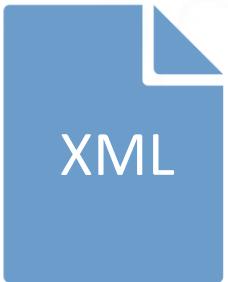
# Loading Data in R



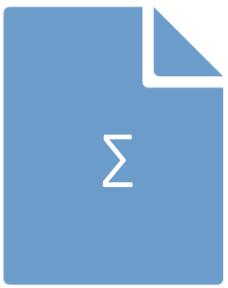
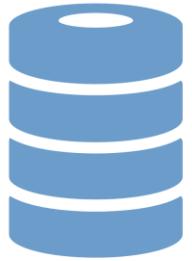
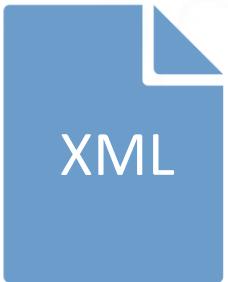
# Loading Data in R



# 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

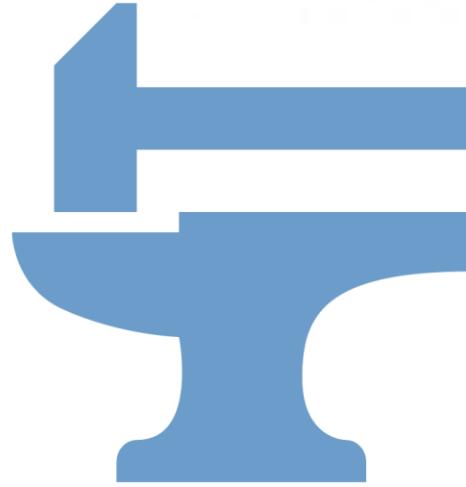
Ensure internal consistency

Handle errors and outliers

Handle missing values

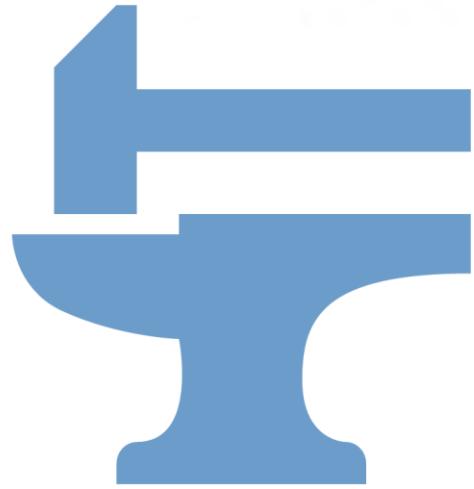


# Transforming Data



# Transforming Data

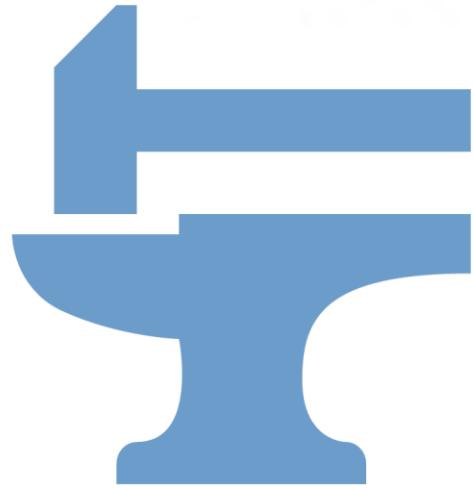
Select columns



# Transforming Data

Select columns

Select rows

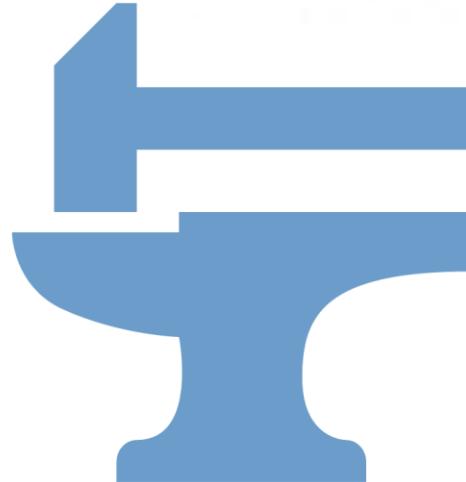


# Transforming Data

Select columns

Select rows

Group rows



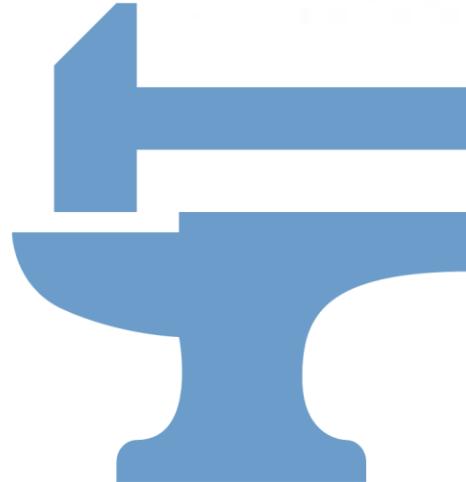
# Transforming Data

Select columns

Select rows

Group rows

Order rows



# Transforming Data

Select columns

Select rows

Group rows

Order rows

Merging data sets



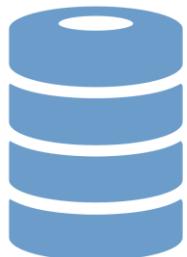
# Exporting Data

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

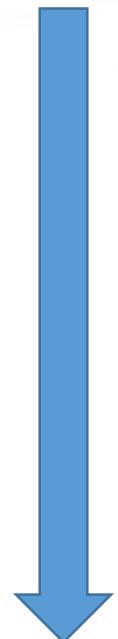
Movie Runtime	
Statistic	Value (minutes)
Minimum	38
1 <sup>st</sup> Quartile	93
Median	101
Mean	104
3 <sup>rd</sup> Quartile	113
Maximum	219

# 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

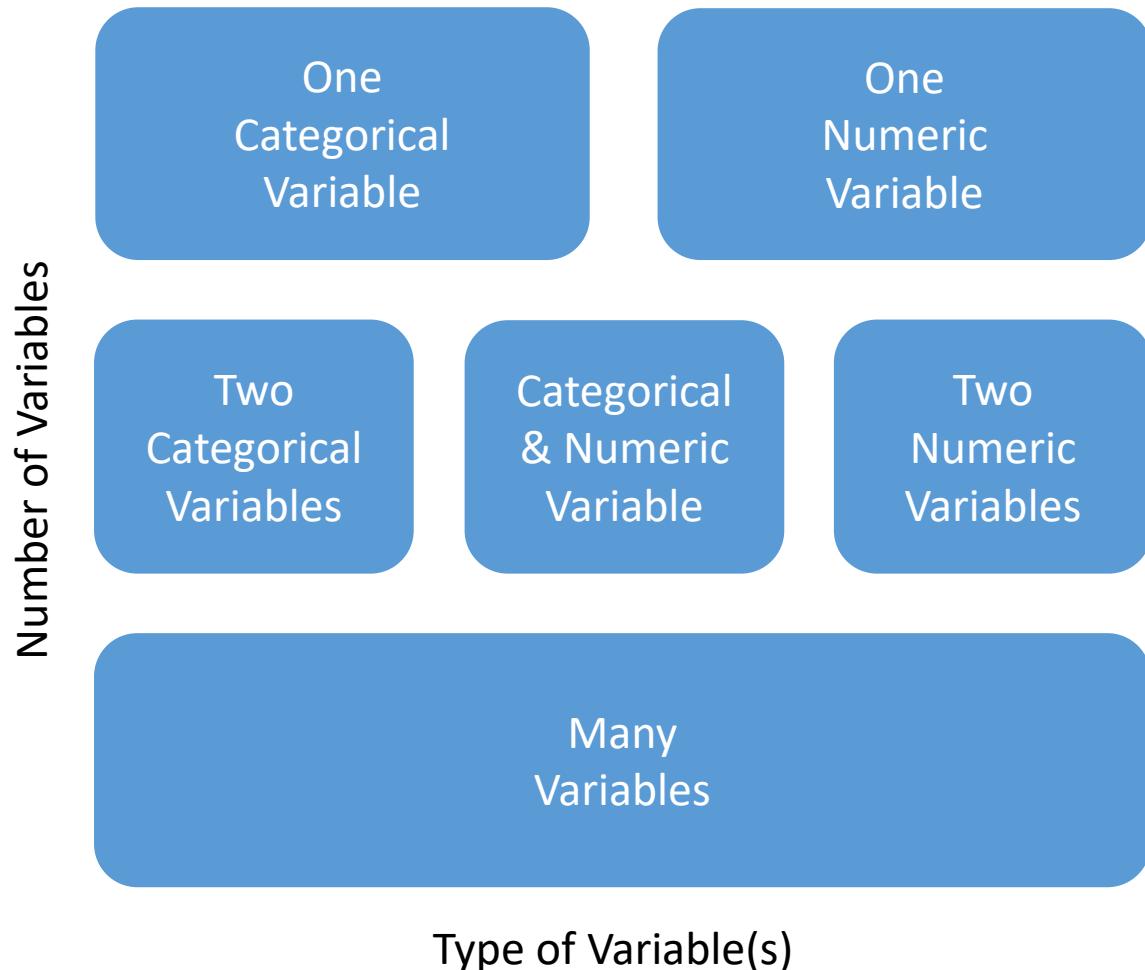
Variables

Categorical variables

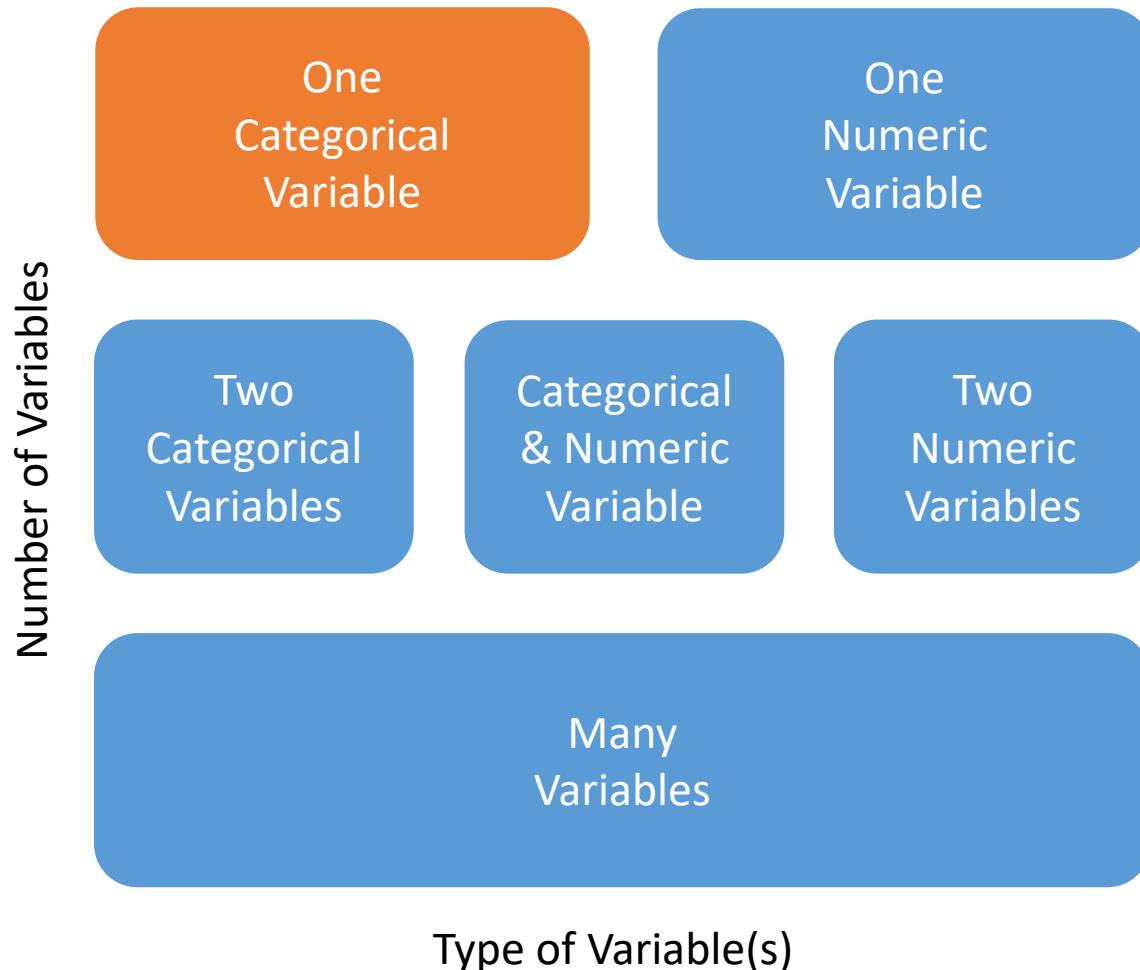
Numeric variables

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

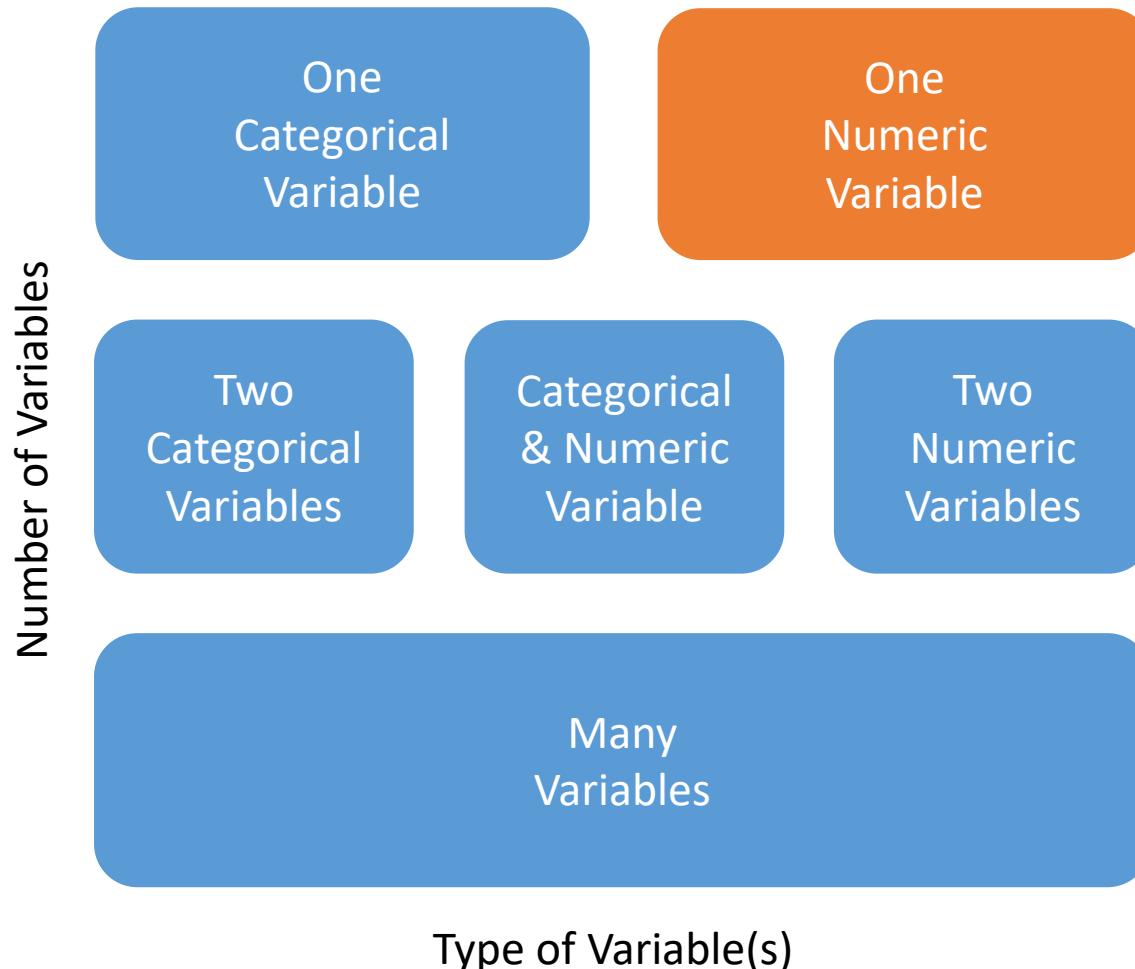
Movies by Genre		
Genre	Frequency	Percentage
Action	612	9%
Adventure	496	7%
Animation	168	2%
Comedy	1281	18%
Drama	1570	22%
Horror	269	4%
...	...	...

# 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

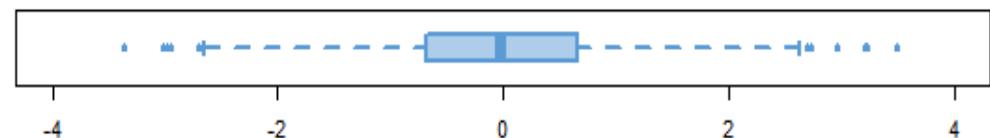


# Analyzing One Numeric Variable

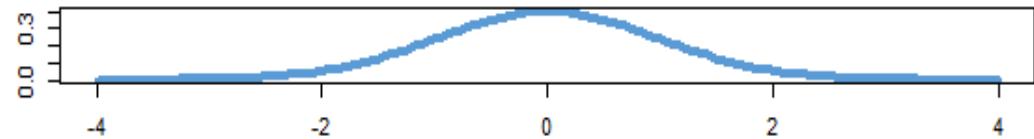
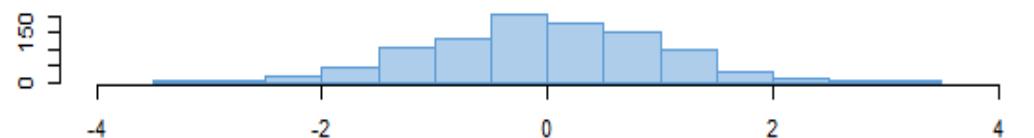
Central tendency



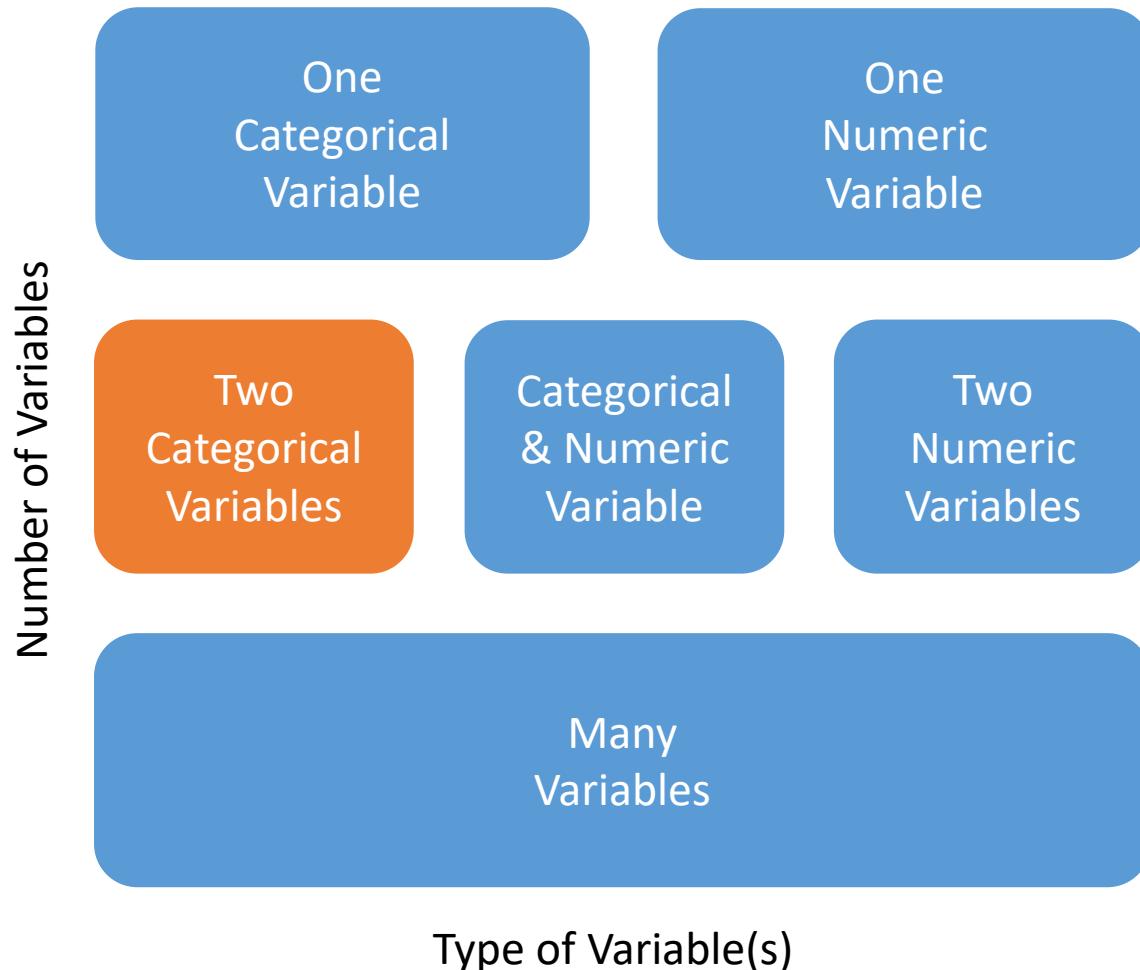
Dispersion



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

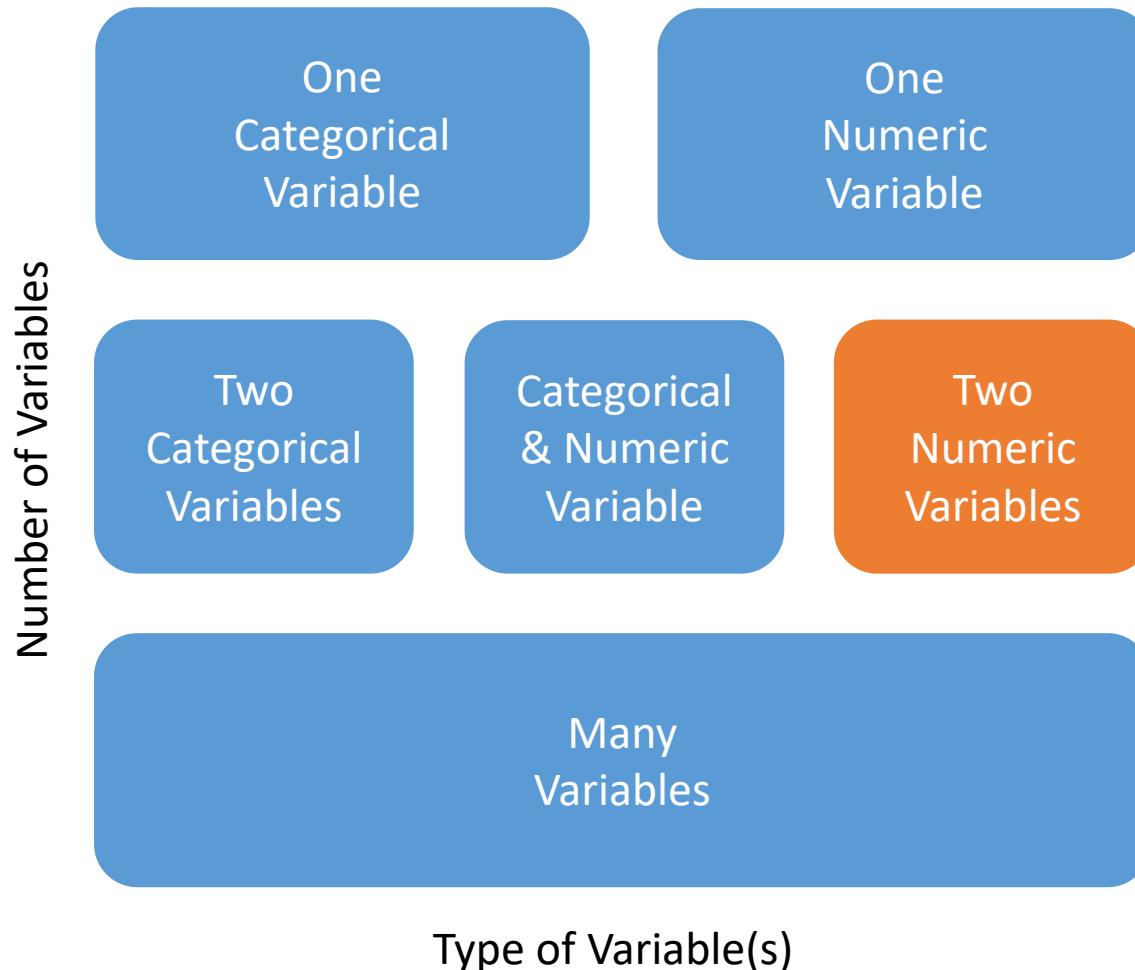
Contingency table

Marginal frequency

Relative frequency

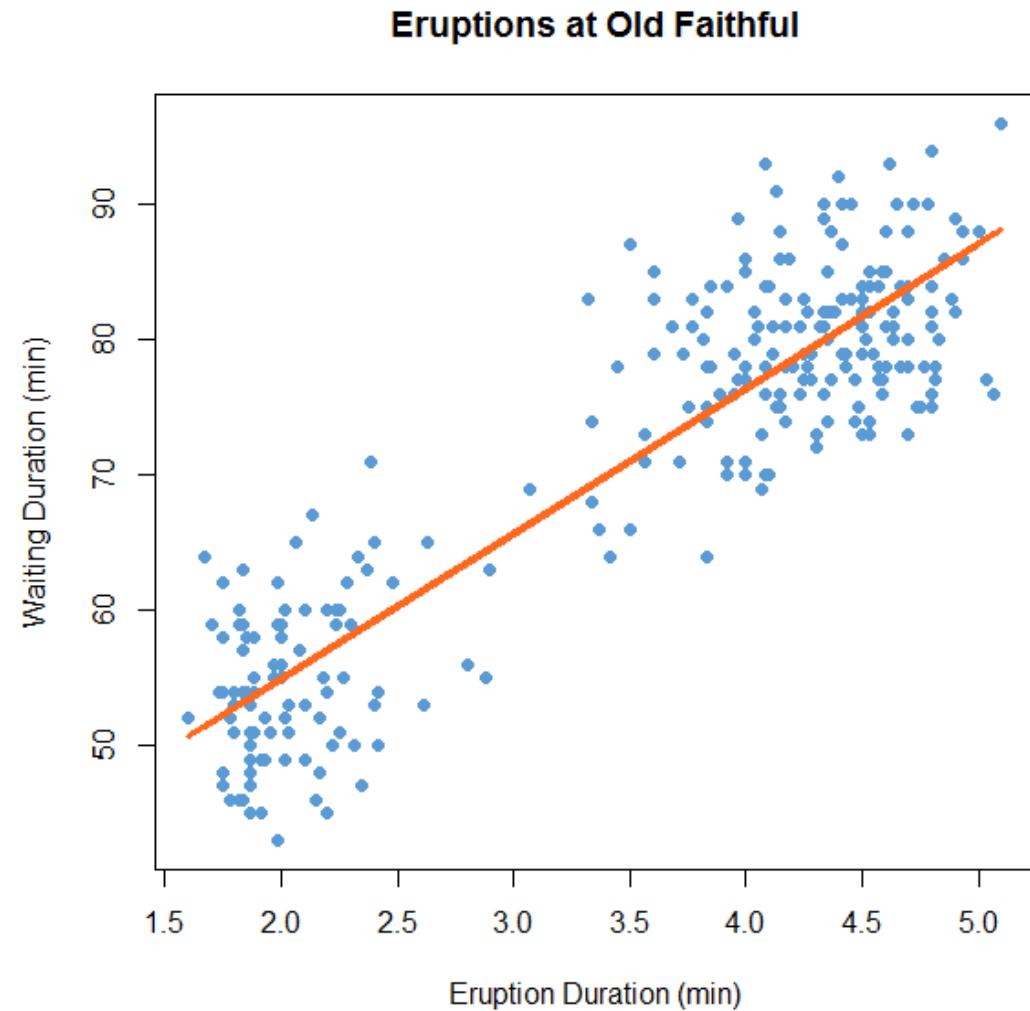
Movies by Genre and Rating					
Genre	G	PG	PG-13	R	Total
Action	0.001	0.010	0.043	0.032	0.086
Adventure	0.006	0.025	0.029	0.009	0.069
Animation	0.006	0.015	0.001	0.001	0.023
Comedy	0.006	0.036	0.066	0.070	0.170
Drama	0.002	0.019	0.082	0.116	0.219
Family	0.005	0.025	0.001	0.001	0.033
...	...	...	...	...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

# Analyzing Two Numeric Variables

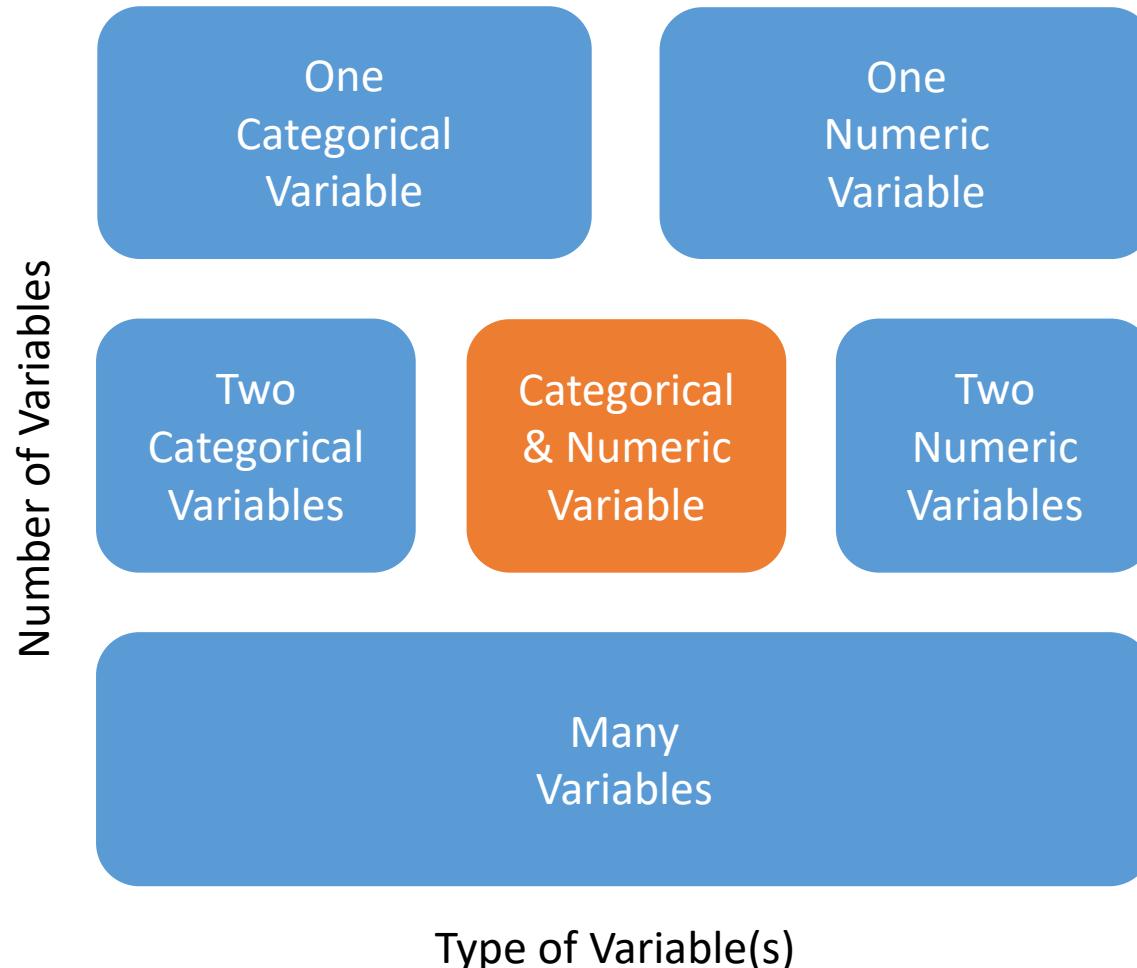


# Analyzing Two Numeric Variables

Explanatory vs. outcome  
Covariance  
Correlation

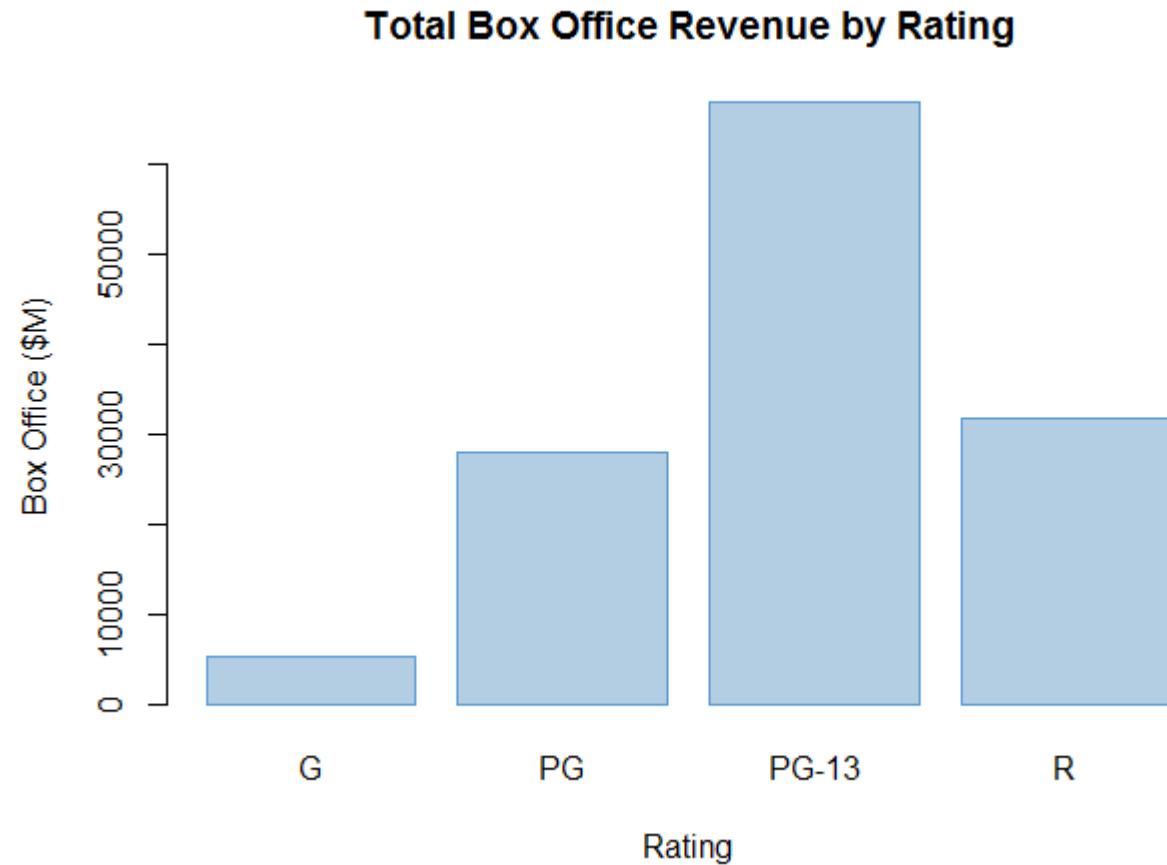


# Analyzing a Numeric Variable Grouped by a Categorical Variable

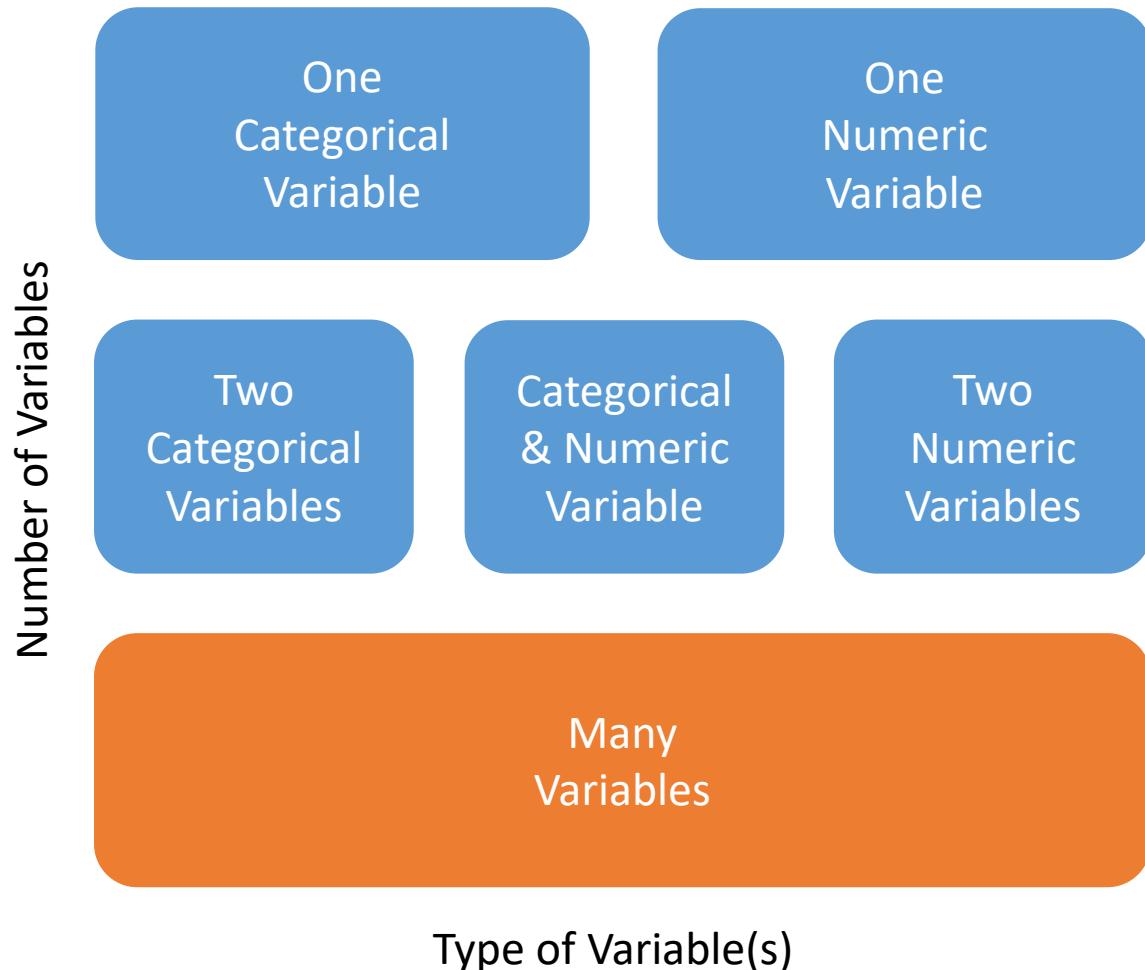


# Analyzing a Numeric Variable Grouped by a Categorical Variable

One categorical variable  
One numeric variable  
Aggregate measures



# Analyzing Many Variables







# COWBOYS & Space Invaders: The Musical



Extended Edition



# Code Demo



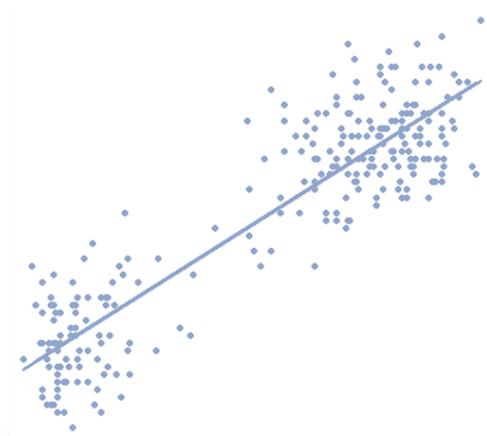
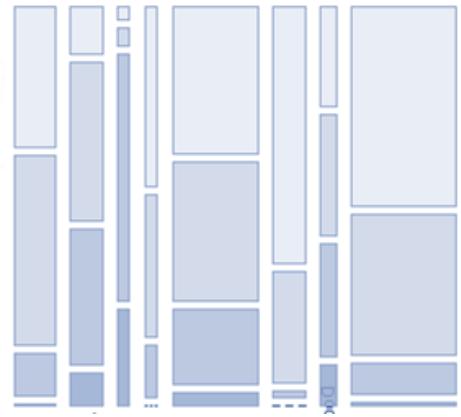
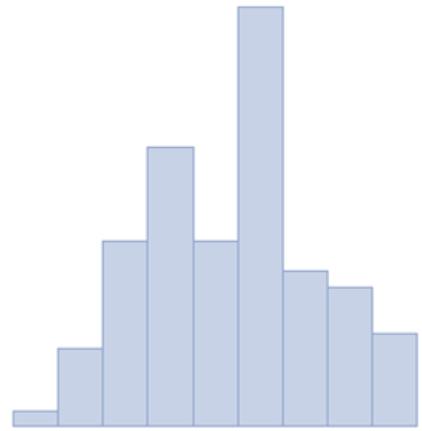
COWBOYS &  
Space Invaders:



# Data Visualization

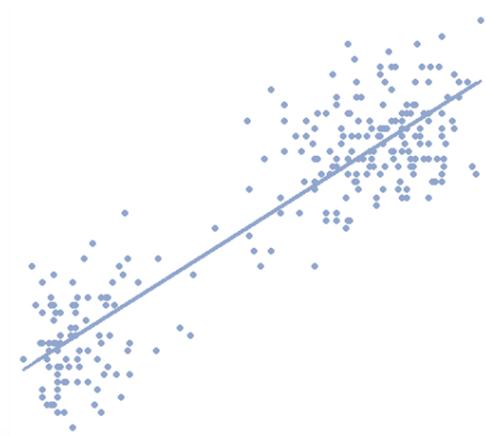
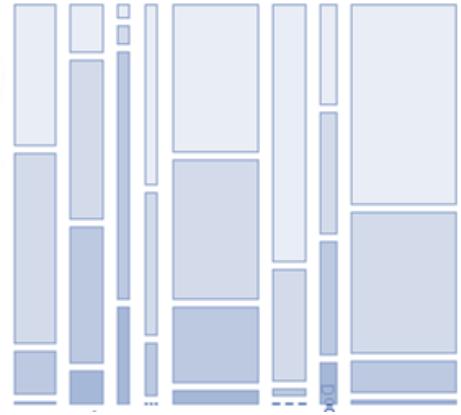
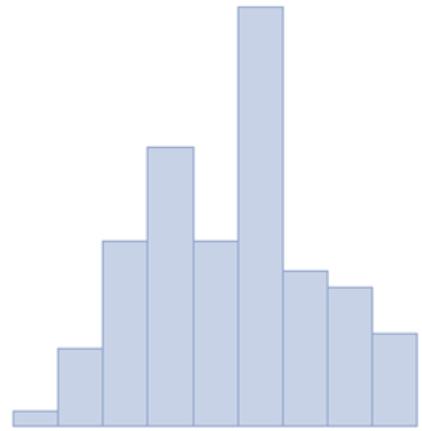
# Data Visualization

Visual data representation



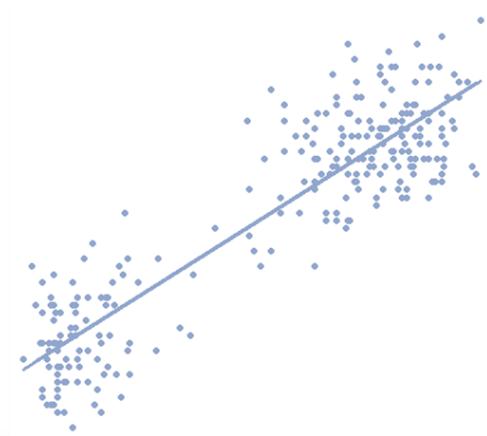
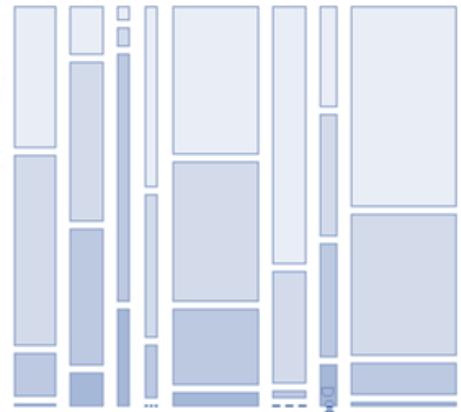
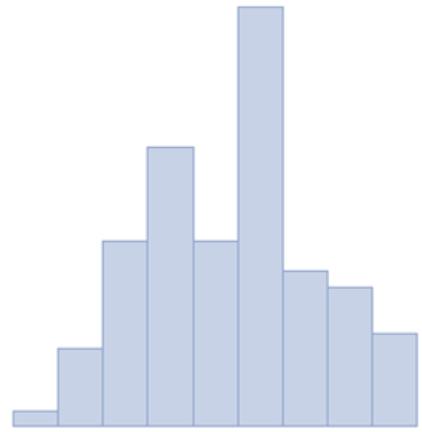
# Data Visualization

Visual data representation  
Human pattern recognition



# Data Visualization

Visual data representation  
Human pattern recognition  
Map dimensions to visual

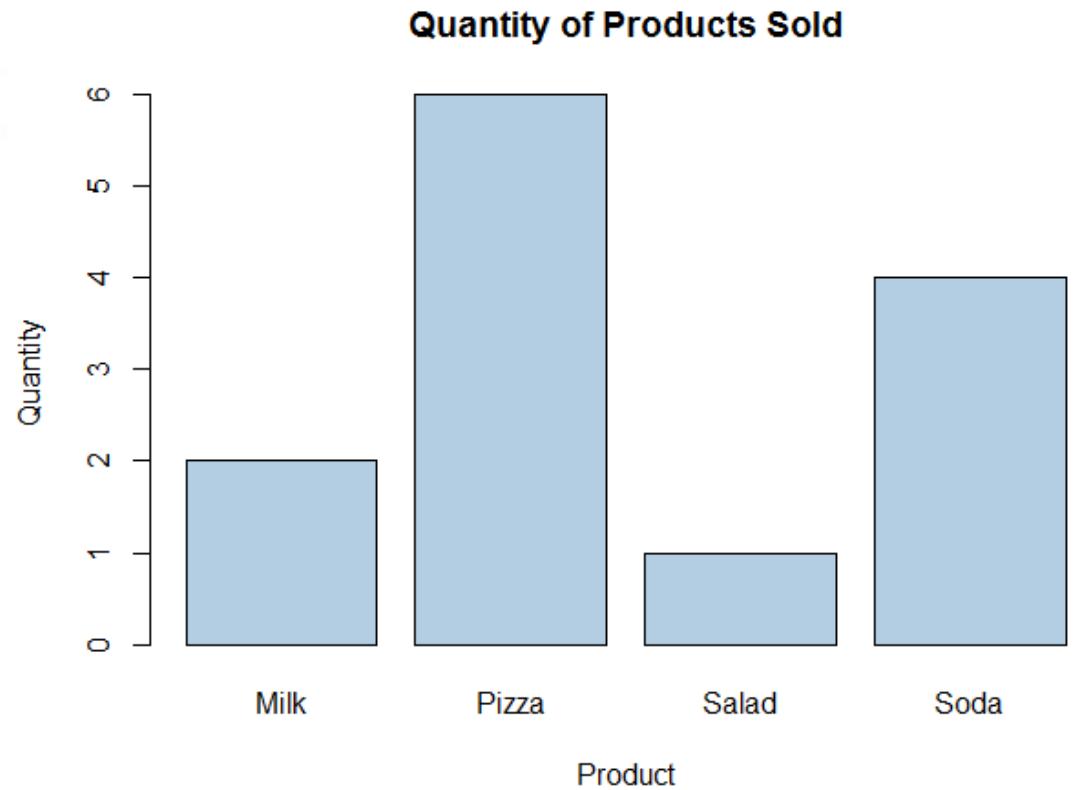


# Data Visualization

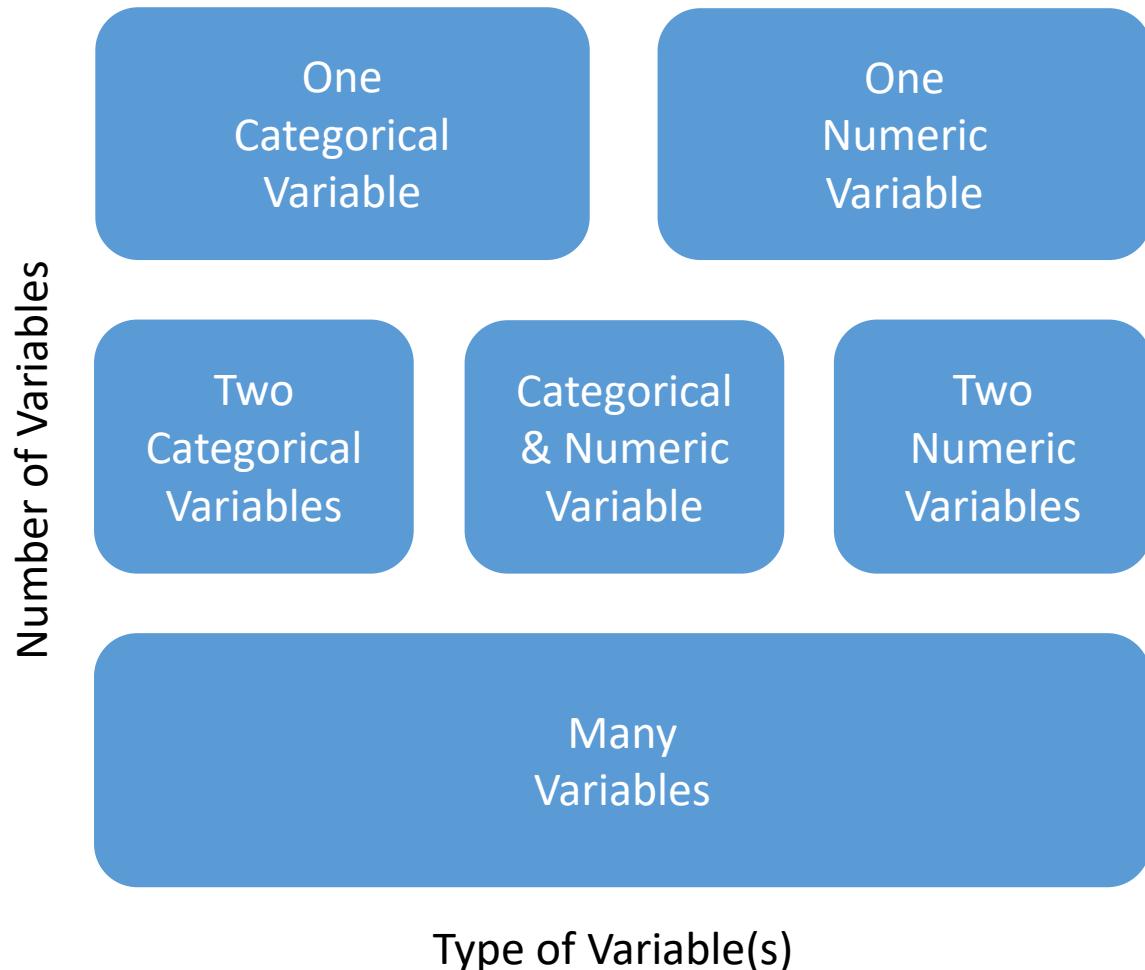
ID	Date	Customer	Product	Quantity
1	2015-08-27	John	Pizza	2
2	2015-08-27	John	Soda	2
3	2015-08-27	Jill	Salad	1
4	2015-08-27	Jill	Milk	1
5	2015-08-28	Miko	Pizza	3
6	2015-08-28	Miko	Soda	2
7	2015-08-28	Sam	Pizza	1
8	2015-08-28	Sam	Milk	1

# 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





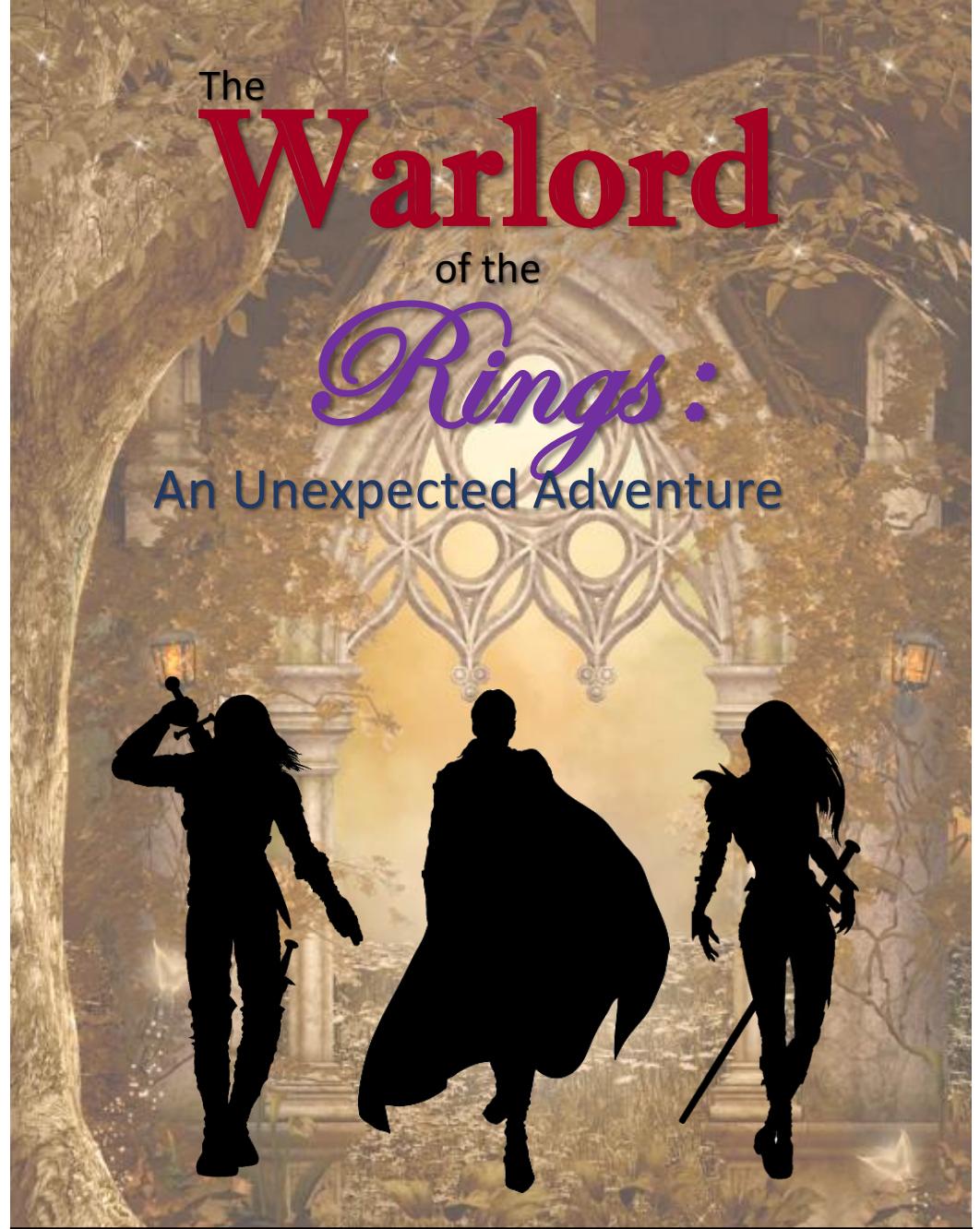
# COWBOYS & Space Invaders: The Musical



Extended Edition



# Code Demo



Feature Length

PG

# Beyond R and EDA



This is just the tip of the iceberg!

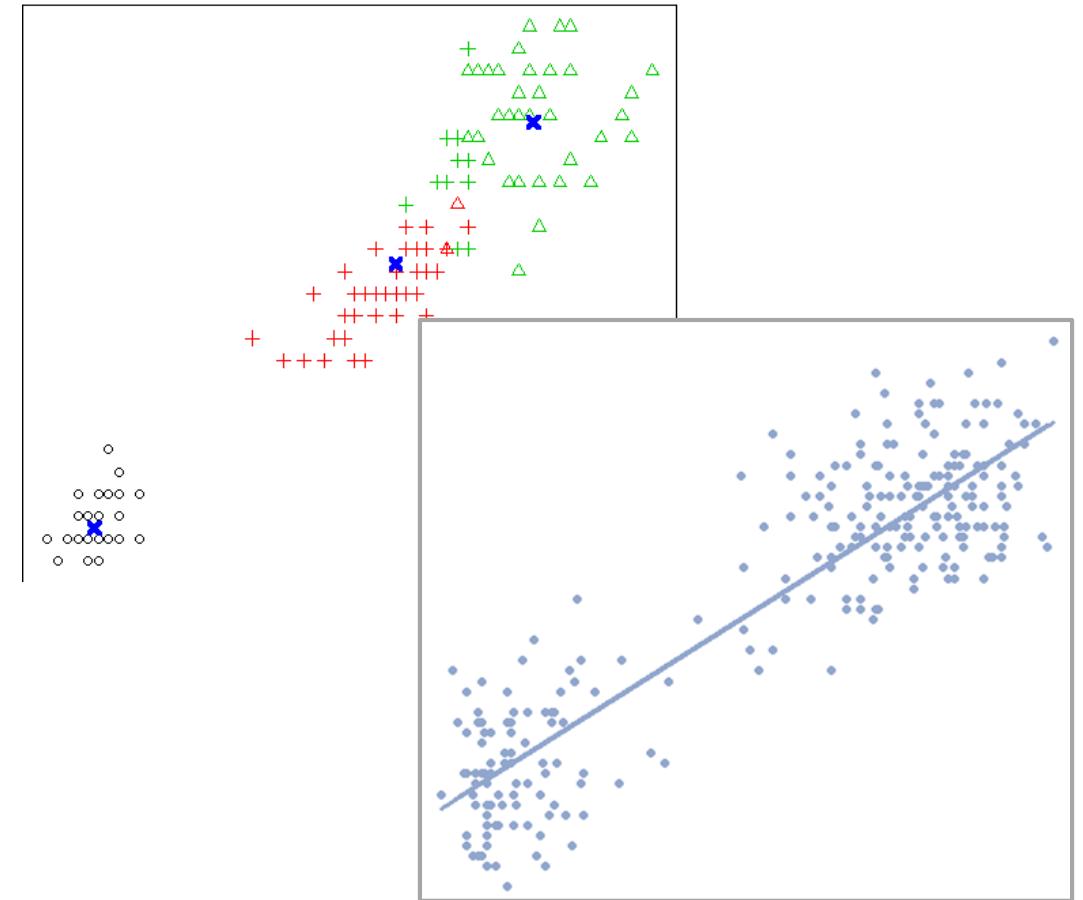
# Advanced Data Analysis with R

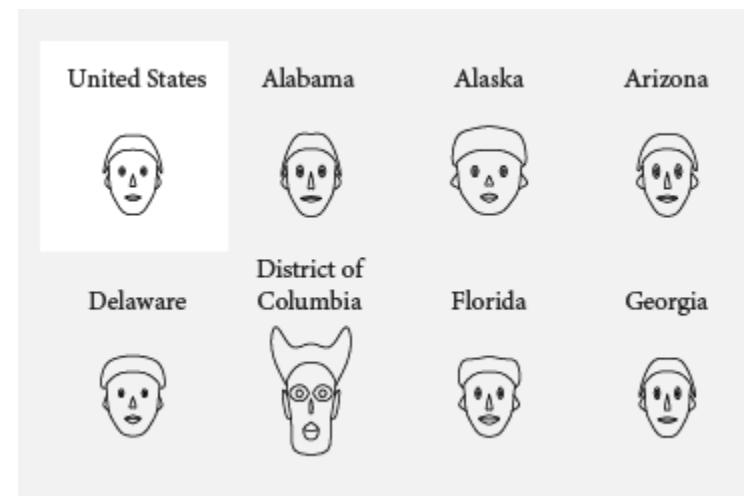
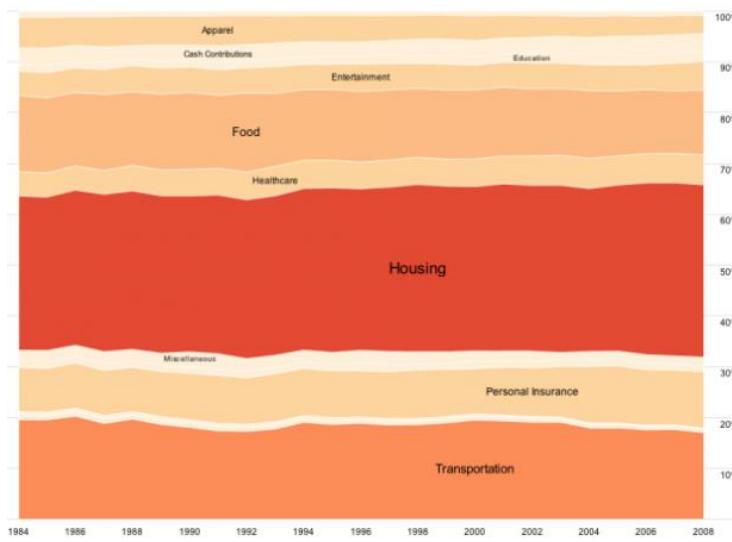
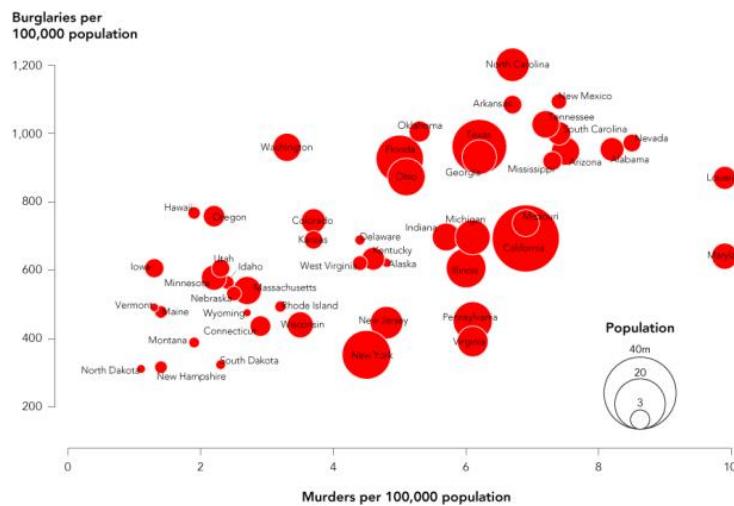
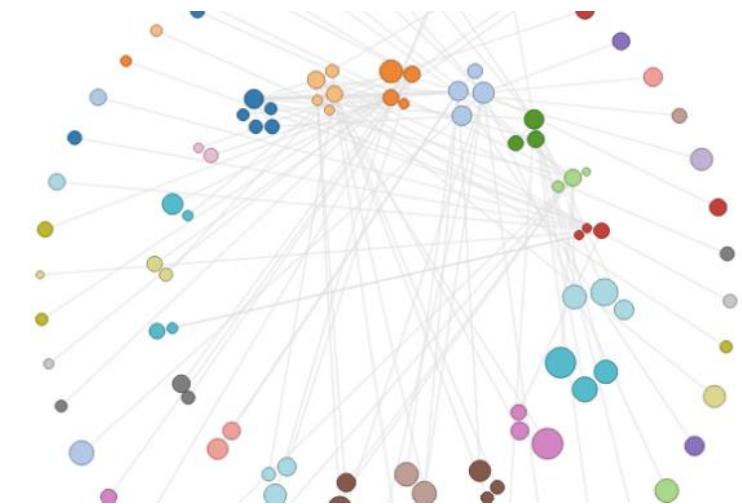
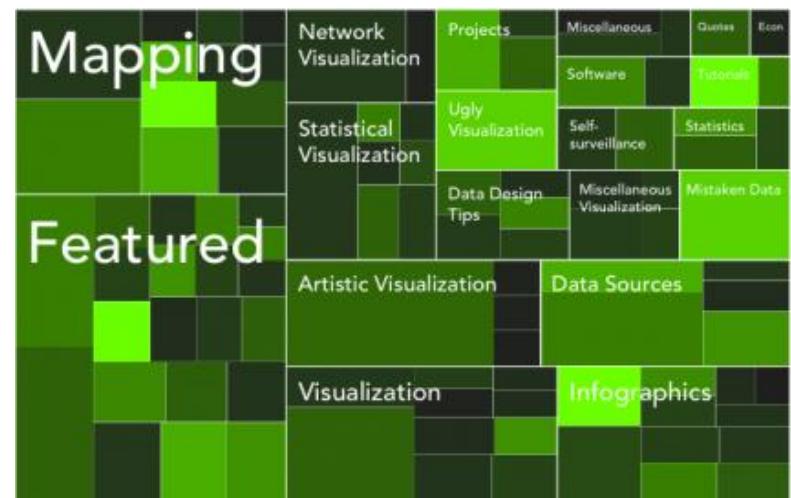
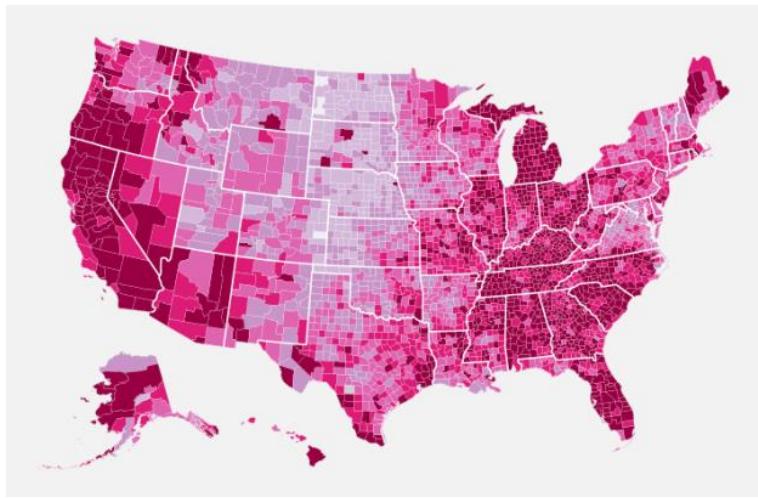
Cluster Analysis

Statistical Modeling

Dimensionality Reduction

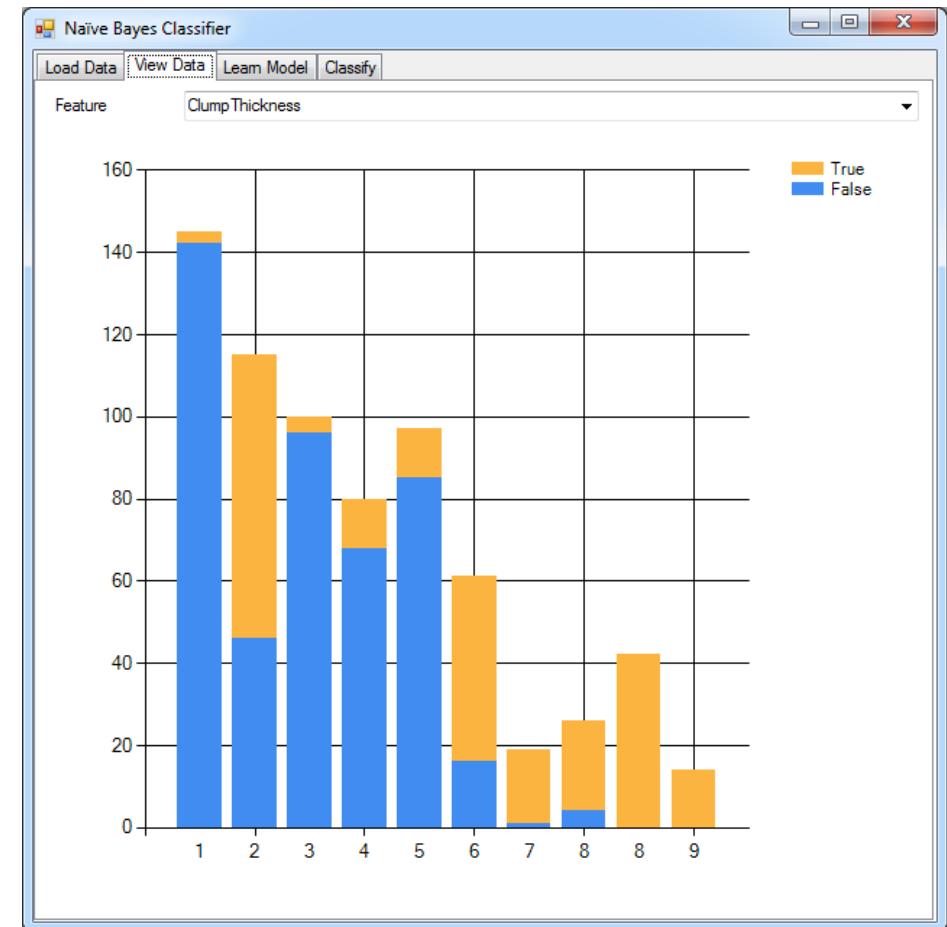
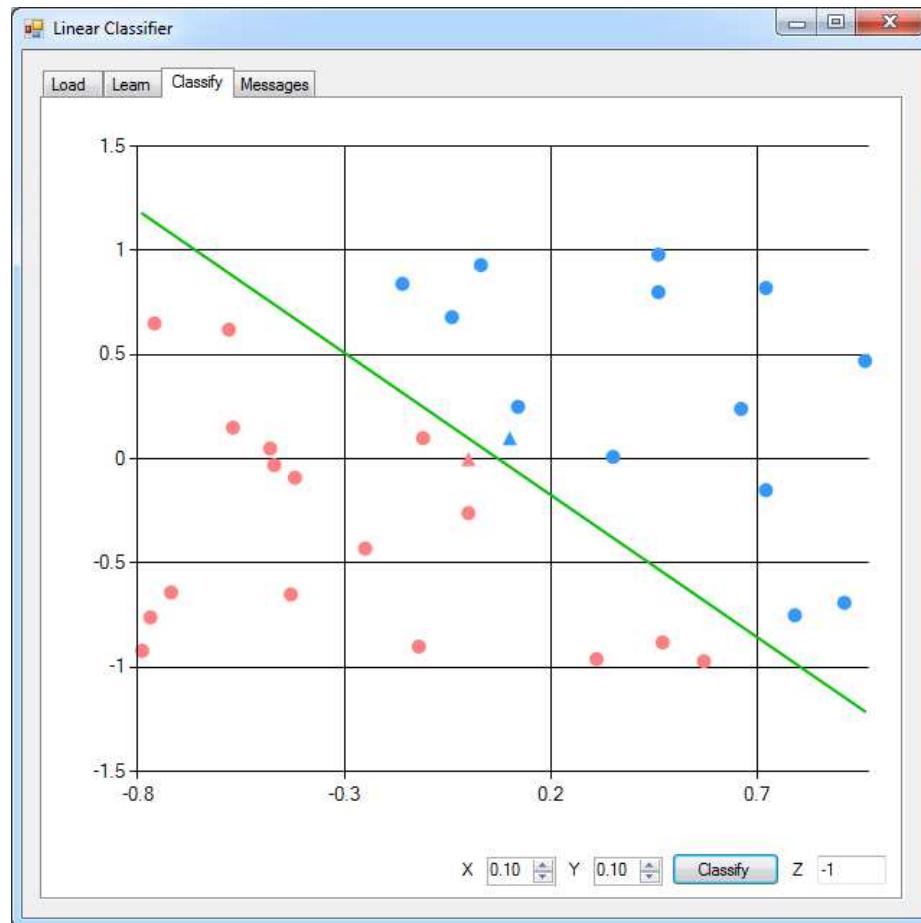
Analysis of Variance (ANOVA)





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

Exploratory Data Analysis with R



Matthew Renze

@matthewrenze | www.matthewrenze.com

pluralsight

[www.pluralsight.com/authors/matthew-renze](http://www.pluralsight.com/authors/matthew-renze)

## 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.



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

### 2017-08-14 - Dev on Fire Interview

# 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](https://twitter.com/matthewrenze)

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

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

