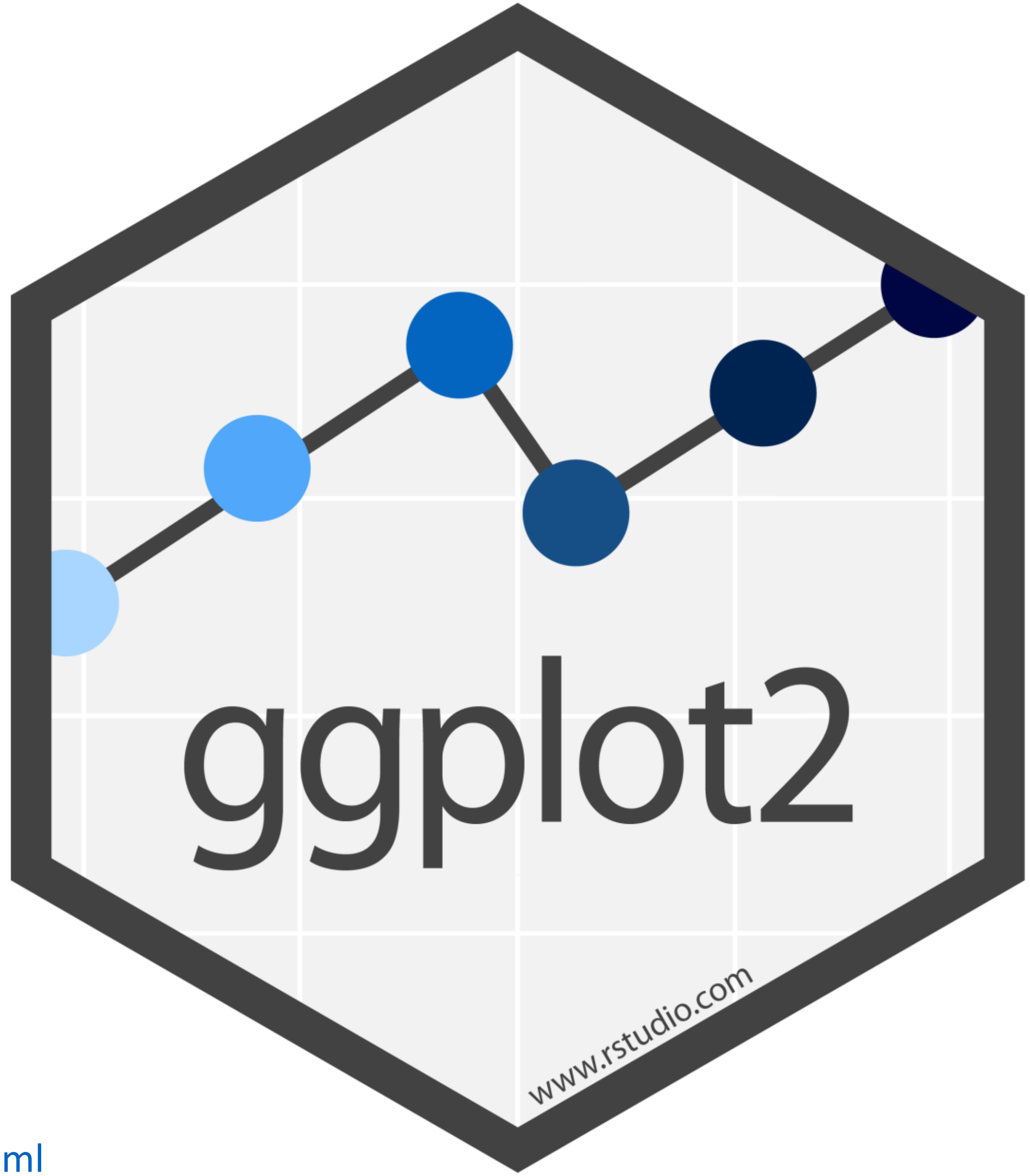
Intro to ggplot2

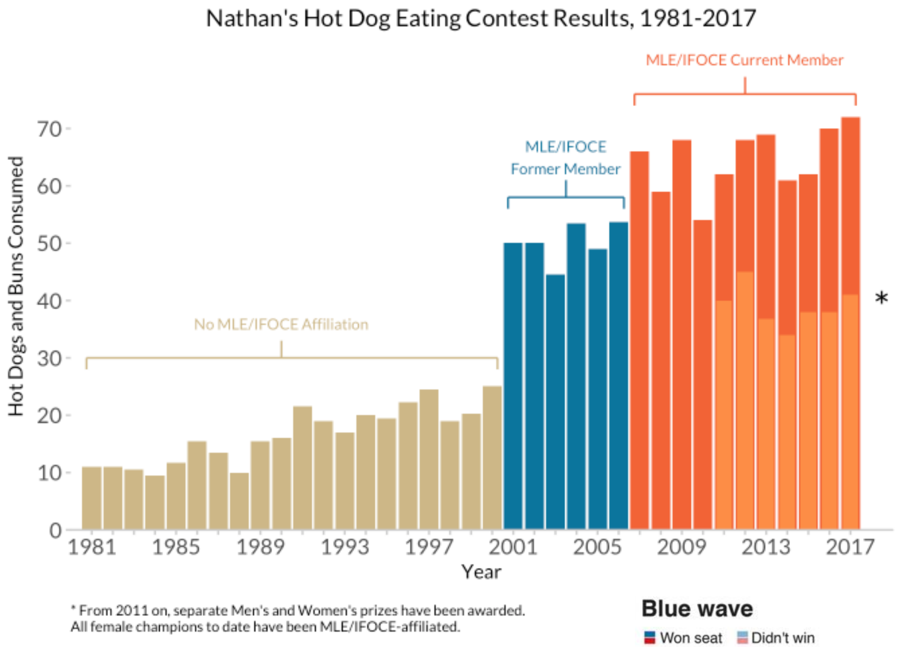
January 2020



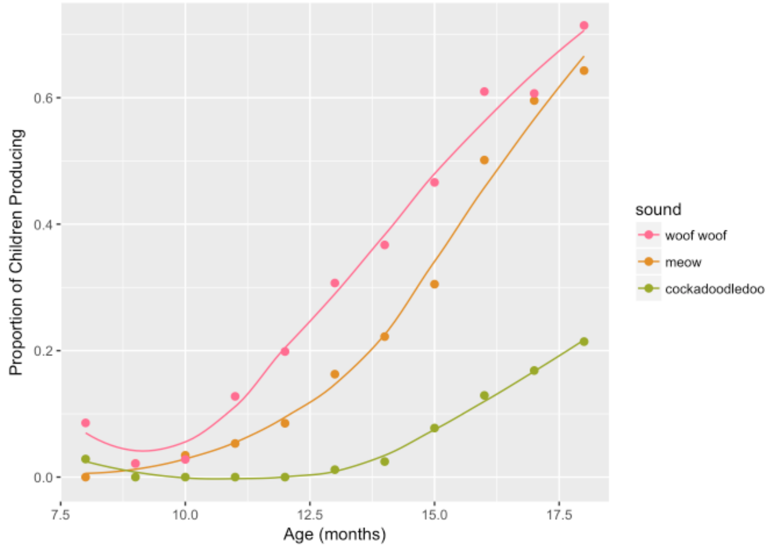
https://psrc.github.io/intro-ggplot2/

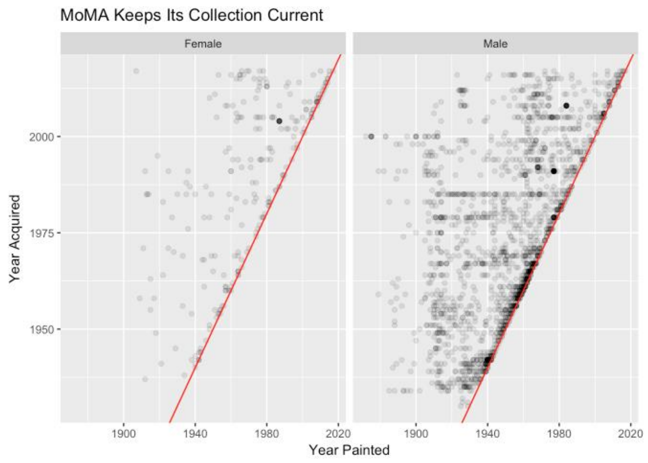
Follow along with the class outline:
https://psrc.github.io/intro-ggplot2/content/class_outline.html

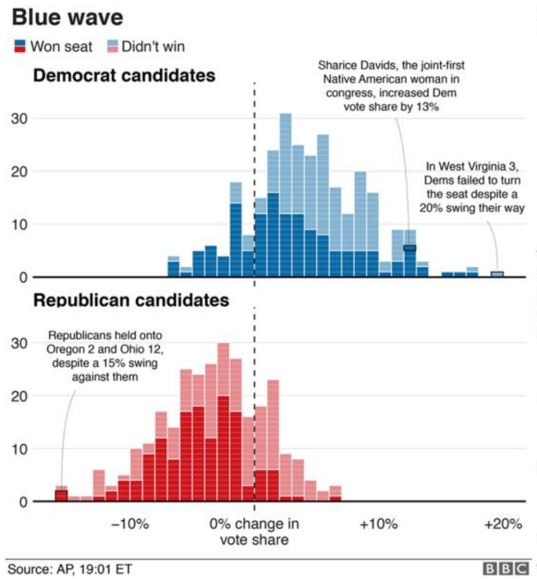
ggplot2


Nathan's Hot Dog Eating Contest Results, 1981-2017

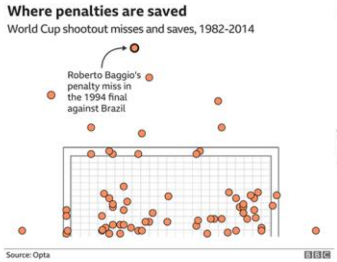
* From 2011 on, separate Men's and Women's prizes have been awarded. All female champions to date have been MLE/IFOCE-affiliated.




MoMA Keeps Its Collection Current

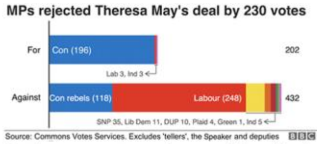

Blue wave

Source: AP, 19:01 ET

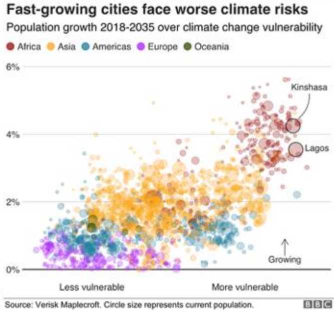

Where penalties are saved

World Cup shootout misses and saves, 1982-2014

Source: Opta

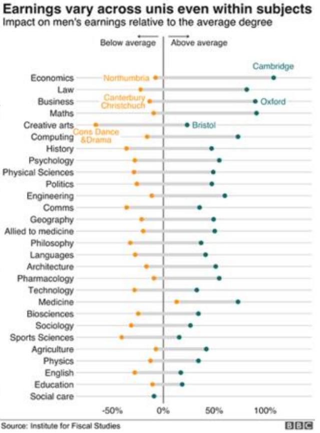

MPs rejected Theresa May's deal by 230 votes

Source: Commons Votes Services. Excludes 'tellers', the Speaker and deputies


Fast-growing cities face worse climate risks

Population growth 2018-2035 over climate change vulnerability

Source: Verisk Maplecroft. Circle size represents current population.


Earnings vary across unis even within subjects

Impact on men's earnings relative to the average degree

Source: Institute for Fiscal Studies

A versatile library

- Stand alone image
- Render in Rmarkdown reports
- Render in Shiny applications
- Integrate with other packages

Goals

- Getting started with simple graphs
- Glimpse of all the options and tools available

Agenda

1. Code along
   1. Bar Graph
   2. Other Graph Types
   3. Facets
2. Extensions
3. My favorite resources

Example ggplots from https://apreshill.github.io/data-vis-labs-2018/, https://github.com/bbc/bbplot

Why ‘ggplot2’? Where does the 2 come from ???

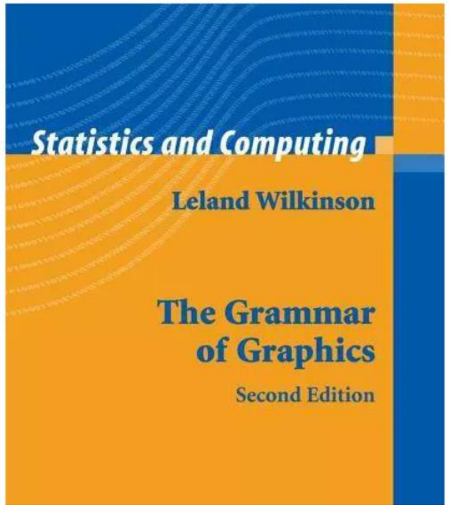
ggplot == ggplot2

About ggplot2

- Created by Hadley Wickham in 2005
- Based on *The Grammar of Graphics* (1999, Leland Wilkinson)

Benefits of ggplot2

- Customize and edit parts easily because everything is broken down into individual components
- Can store plots in variables
- Flexibility

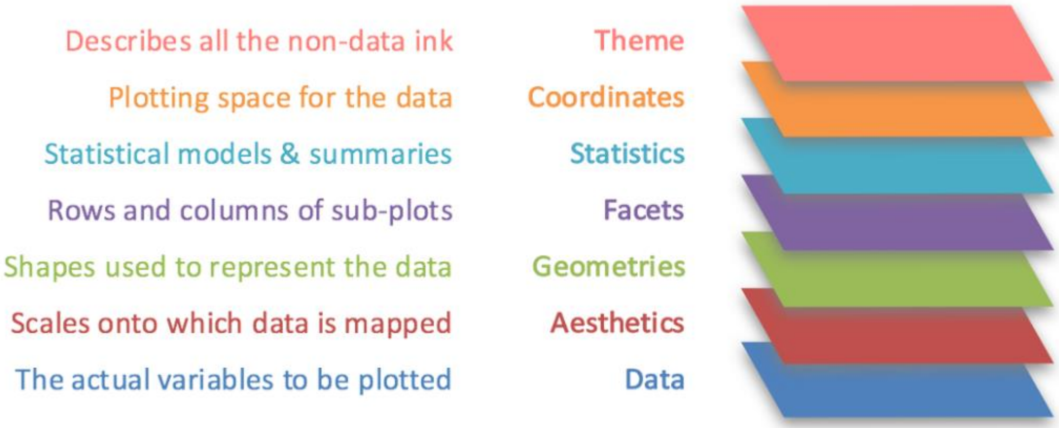


Statistics and Computing

Leland Wilkinson

The Grammar of Graphics

Second Edition



Describes all the non-data ink	Theme
Plotting space for the data	Coordinates
Statistical models & summaries	Statistics
Rows and columns of sub-plots	Facets
Shapes used to represent the data	Geometries
Scales onto which data is mapped	Aesthetics
The actual variables to be plotted	Data

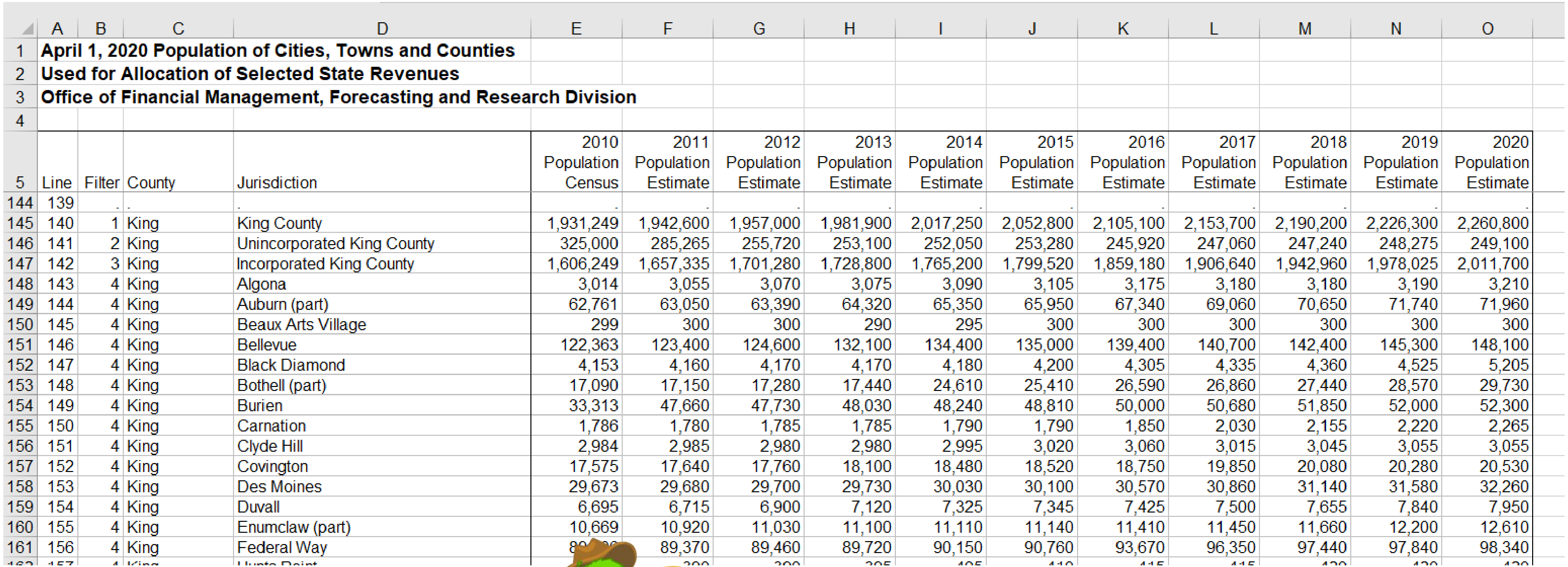
Grammar of Graphics: A layered approach to elegant visuals

- Member of the Tidyverse (RStudio)
- As long as you have a data frame you can use ggplot, with or without other Tidyverse packages

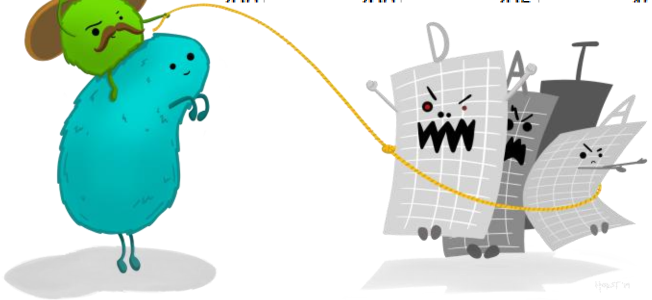
The tidyverse is an opinionated collection of R packages designed for data science. All package share an underlying design philosophy, grammar, and data structures.

-- tidyverse.org

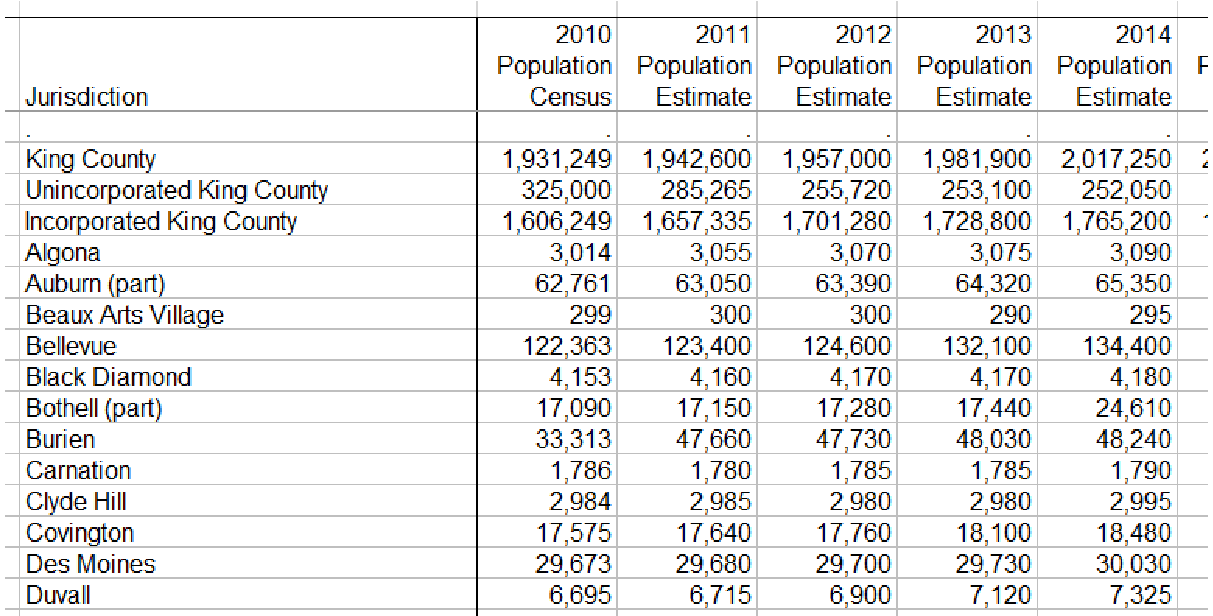
Original Data format



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**April 1, 2020 Population of Cities, Towns and Counties**														
2	**Used for Allocation of Selected State Revenues**														
3	**Office of Financial Management, Forecasting and Research Division**														
4															
5	Line	Filter	County	Jurisdiction	2010 Population Census	2011 Population Estimate	2012 Population Estimate	2013 Population Estimate	2014 Population Estimate	2015 Population Estimate	2016 Population Estimate	2017 Population Estimate	2018 Population Estimate	2019 Population Estimate	2020 Population Estimate
144	139	.	.	.	.	.	.	.	.	.	.	.	.	.	.
145	140	1	King	King County	1,931,249	1,942,600	1,957,000	1,981,900	2,017,250	2,052,800	2,105,100	2,153,700	2,190,200	2,226,300	2,260,800
146	141	2	King	Unincorporated King County	325,000	285,265	255,720	253,100	252,050	253,280	245,920	247,060	247,240	248,275	249,100
147	142	3	King	Incorporated King County	1,606,249	1,657,335	1,701,280	1,728,800	1,765,200	1,799,520	1,859,180	1,906,640	1,942,960	1,978,025	2,011,700
148	143	4	King	Algona	3,014	3,055	3,070	3,075	3,090	3,105	3,175	3,180	3,180	3,190	3,210
149	144	4	King	Auburn (part)	62,761	63,050	63,390	64,320	65,350	65,950	67,340	69,060	70,650	71,740	71,960
150	145	4	King	Beaux Arts Village	299	300	300	290	295	300	300	300	300	300	300
151	146	4	King	Bellevue	122,363	123,400	124,600	132,100	134,400	135,000	139,400	140,700	142,400	145,300	148,100
152	147	4	King	Black Diamond	4,153	4,160	4,170	4,170	4,180	4,200	4,305	4,335	4,360	4,525	5,205
153	148	4	King	Bothell (part)	17,090	17,150	17,280	17,440	24,610	25,410	26,590	26,860	27,440	28,570	29,730
154	149	4	King	Burien	33,313	47,660	47,730	48,030	48,240	48,810	50,000	50,680	51,850	52,000	52,300
155	150	4	King	Carnation	1,786	1,780	1,785	1,785	1,790	1,790	1,850	2,030	2,155	2,220	2,265
156	151	4	King	Clyde Hill	2,984	2,985	2,980	2,980	2,995	3,020	3,060	3,015	3,045	3,055	3,055
157	152	4	King	Covington	17,575	17,640	17,760	18,100	18,480	18,520	18,750	19,850	20,080	20,280	20,530
158	153	4	King	Des Moines	29,673	29,680	29,700	29,730	30,030	30,100	30,570	30,860	31,140	31,580	32,260
159	154	4	King	Duvall	6,695	6,715	6,900	7,120	7,325	7,345	7,425	7,500	7,655	7,840	7,950
160	155	4	King	Enumclaw (part)	10,669	10,920	11,030	11,100	11,110	11,140	11,410	11,450	11,660	12,200	12,610
161	156	4	King	Federal Way	[illegible]	89,370	89,460	89,720	90,150	90,760	93,670	96,350	97,440	97,840	98,340



Artwork by @allison_horst

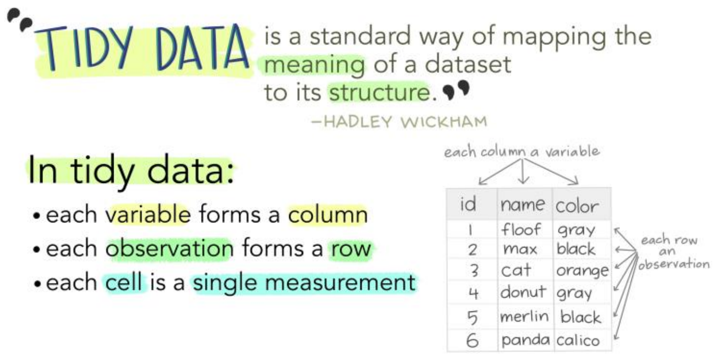


Jurisdiction	2010 Population Census	2011 Population Estimate	2012 Population Estimate	2013 Population Estimate	2014 Population Estimate	F
.	-	-	-	-	-	
King County	1,931,249	1,942,600	1,957,000	1,981,900	2,017,250	2
Unincorporated King County	325,000	285,265	255,720	253,100	252,050	
Incorporated King County	1,606,249	1,657,335	1,701,280	1,728,800	1,765,200	1
Algona	3,014	3,055	3,070	3,075	3,090	
Auburn (part)	62,761	63,050	63,390	64,320	65,350	
Beaux Arts Village	299	300	300	290	295	
Bellevue	122,363	123,400	124,600	132,100	134,400	
Black Diamond	4,153	4,160	4,170	4,170	4,180	
Bothell (part)	17,090	17,150	17,280	17,440	24,610	
Burien	33,313	47,660	47,730	48,030	48,240	
Carnation	1,786	1,780	1,785	1,785	1,790	
Clyde Hill	2,984	2,985	2,980	2,980	2,995	
Covington	17,575	17,640	17,760	18,100	18,480	
Des Moines	29,673	29,680	29,700	29,730	30,030	
Duvall	6,695	6,715	6,900	7,120	7,325	
Enumclaw (part)	10,669	10,920	11,03			
Federal Way	89,306	89,370	89,46			

Tidy data

Melt or 'pivot longer' to move values away from column headers

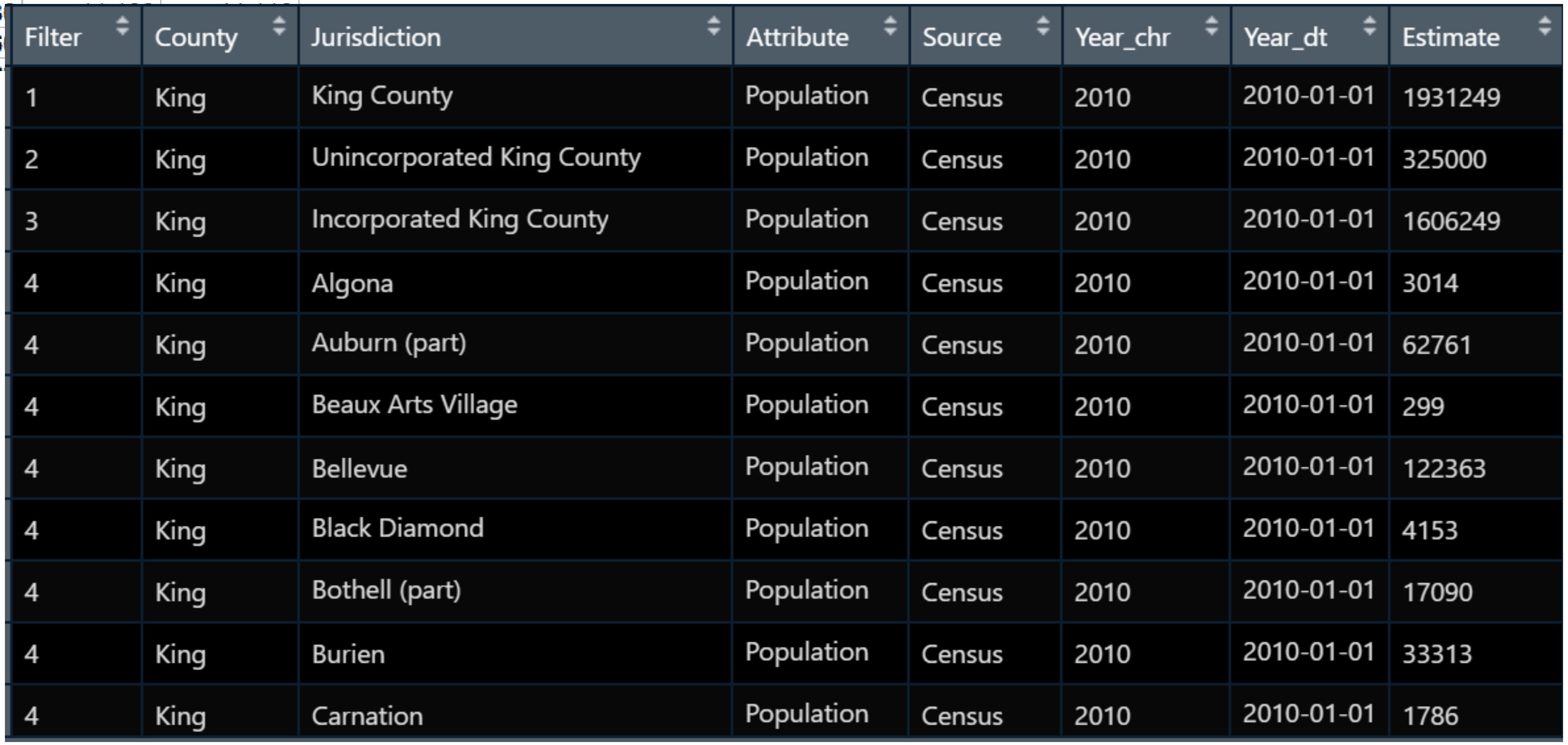
Using the melt() function from the reshape2 package


TIDY DATA is a standard way of mapping the meaning of a dataset to its structure.
—HADLEY WICKHAM
In tidy data:
• each variable forms a column
• each observation forms a row
• each cell is a single measurement
each column a variable
id name color
1 floof gray
2 max black
3 cat orange
4 donut gray
5 merlin black
6 panda calico
each row an observation

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

Artwork by @allison_horst

About tidy data



Filter	County	Jurisdiction	Attribute	Source	Year_chr	Year_dt	Estimate
1	King	King County	Population	Census	2010	2010-01-01	1931249
2	King	Unincorporated King County	Population	Census	2010	2010-01-01	325000
3	King	Incorporated King County	Population	Census	2010	2010-01-01	1606249
4	King	Algona	Population	Census	2010	2010-01-01	3014
4	King	Auburn (part)	Population	Census	2010	2010-01-01	62761
4	King	Beaux Arts Village	Population	Census	2010	2010-01-01	299
4	King	Bellevue	Population	Census	2010	2010-01-01	122363
4	King	Black Diamond	Population	Census	2010	2010-01-01	4153
4	King	Bothell (part)	Population	Census	2010	2010-01-01	17090
4	King	Burien	Population	Census	2010	2010-01-01	33313
4	King	Carnation	Population	Census	2010	2010-01-01	1786

geom_ = geometric objects drawn to represent the data

aes(…) aka aesthetics = "Mapping the variable"
what variable (aka column) do we want to be the basis for ________

- X axis
- Y axis
- color
- fill
- shape
- size
- ….

- ➢ The aes() only takes mappings from the data onto the geom.
- ➢ For something fixed, set it inside the geom function but outside of aes()
- ➢ Options in the aes() is dependent on the type of geom_

To find specific aes() arguments for a geom, ask for help in the console

- ?<name of geom function>

scale_ = changing scale limits or change the **way** our data are mapped onto our geom_

Data values → Visual values of an aesthetic

Most scale functions follow the format: `scale_{aesthetic}_{method}` or `scale_{aesthetic}_{datatype}` where

- aesthetic are our aesthetic mappings such as color, fill, shape
- method is how the colors, fill colors, and shapes are chosen
- Datatype is the datatype of the variable being mapped

Scales can also be used to …
Change scale

- Linear → Log10

Change our axes

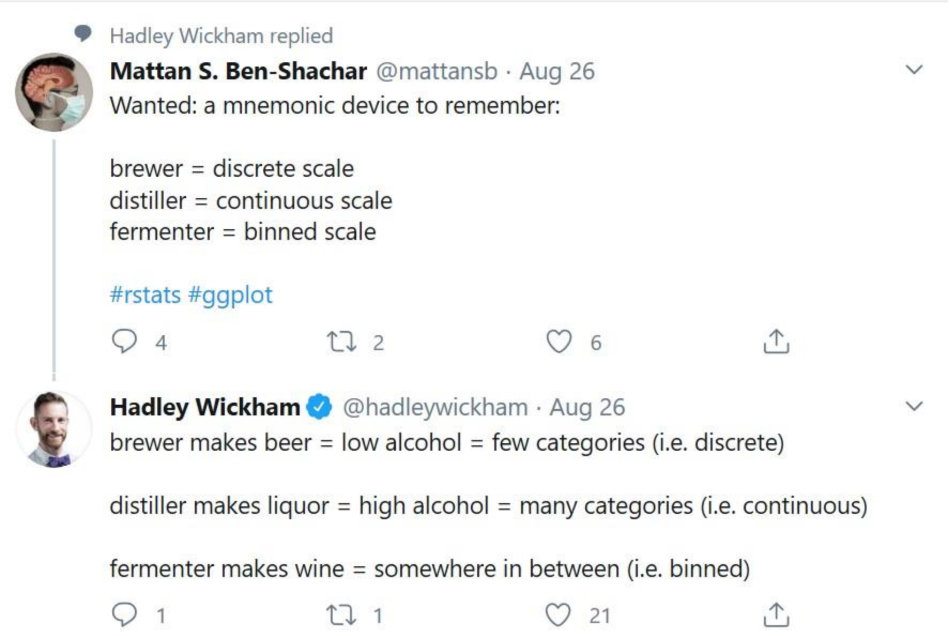
- override the labels
- change the breaks

https://nceas.github.io/oss-lessons/dataviz-and-interactive-tools/module-1-ggplot2.html

Your data influences which function(s) you can use

Which should I use? scale_x_discrete or scale_x_continuous? scale_{aesthetic}_brewer or scale_{aesthetic}_distiller?

CONTINUOUS
measured data, can have ∞ values within possible range.
I am 3.1" tall
I weigh 34.16 grams
DISCRETE
observations can only exist at limited values, often counts.
I have 8 legs and 4 spots!
@allison_horst


Hadley Wickham replied

Mattan S. Ben-Shachar @mattansb · Aug 26

Wanted: a mnemonic device to remember:

brewer = discrete scale
distiller = continuous scale
fermenter = binned scale

#rstats #ggplot

4 2 6

Hadley Wickham @hadleywickham · Aug 26

brewer makes beer = low alcohol = few categories (i.e. discrete)

distiller makes liquor = high alcohol = many categories (i.e. continuous)

fermenter makes wine = somewhere in between (i.e. binned)

1 1 21

theme() = control the non-data part of your plot (titles, labels, fonts, background, gridlines, and legends)

There are a ton of keyword arguments you could use in theme():
https://ggplot2.tidyverse.org/reference/theme.html

Depending on what you're changing, you'll have to wrap your arguments with one of the following functions:

- element_line(): modify the line elements of the theme
- element_text(): to modify the text elements
- element_rect(): to modify the rectangle elements
- element_blank(): to remove the element

There are also preset themes in ggplot2

- theme_bw()
- theme_minimal()
- theme_classic()
- theme_dark()

Factors the categorical datatype

Each unique value can be represented by a label and a level which determines its place when sorted

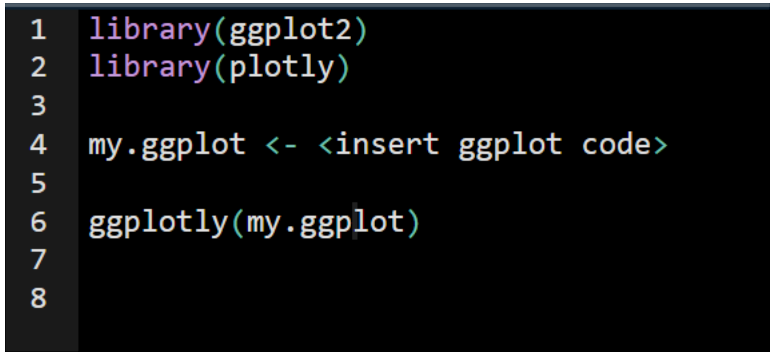
Character datatypes in ggplot2 will display in alphabetical order. If you want to customize the order of the values, convert that column into a factor datatype.

For example, displaying character values in non-alphabetical order (e.g. month names, Starbucks cup sizes, education attainment)

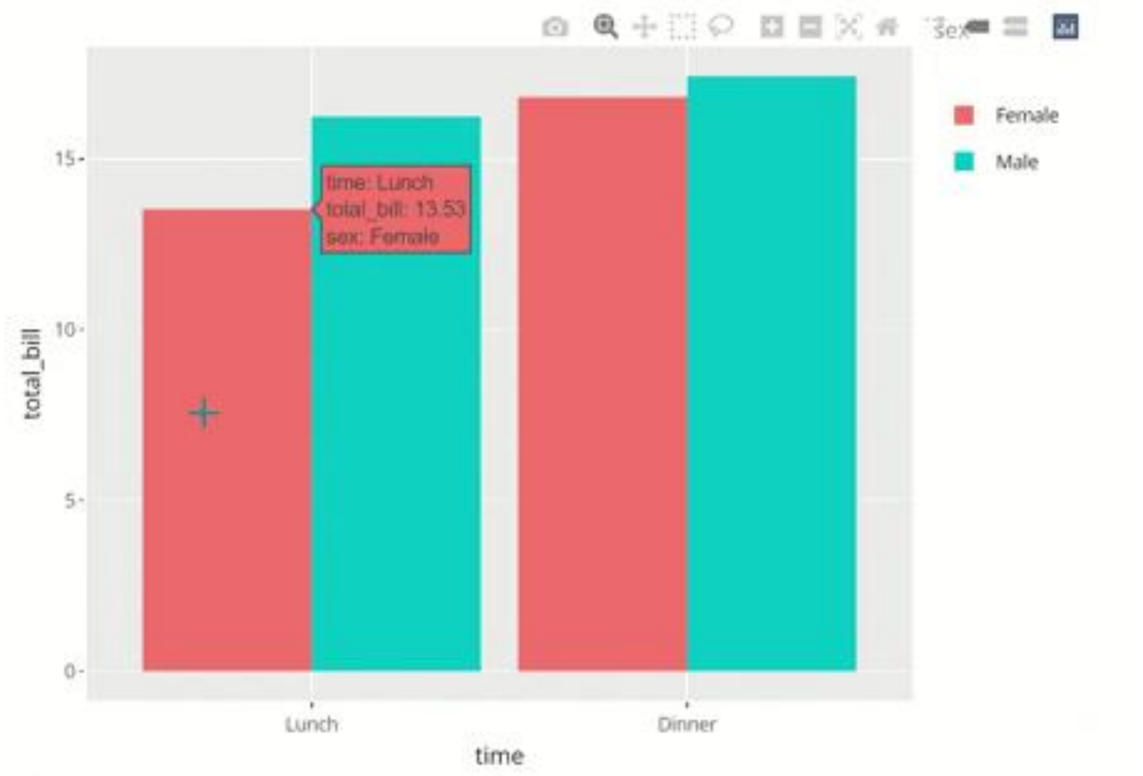
plotly

https://plotly.com/r/
https://plotly.com/ggplot2/

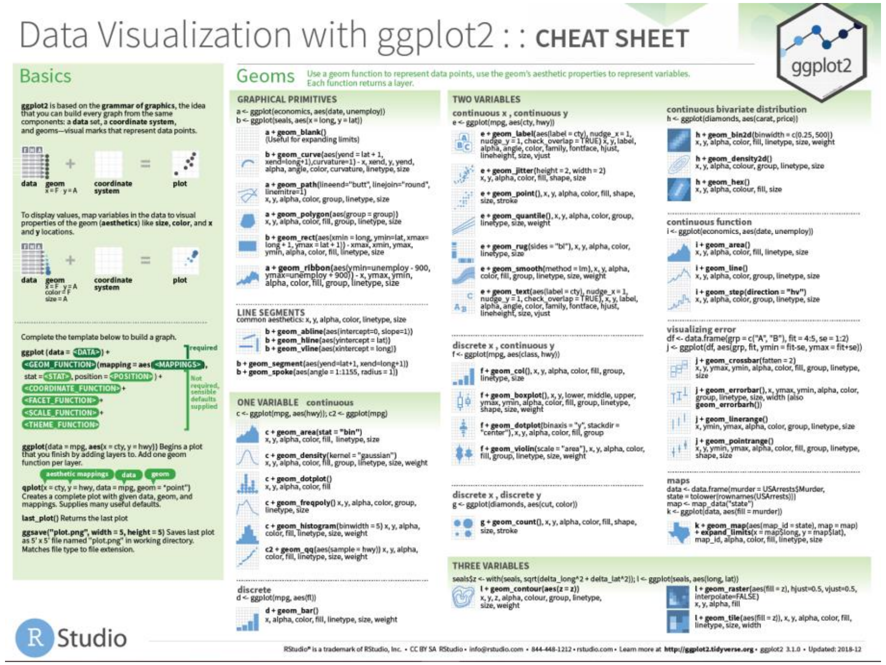
- Another statistical graphing library that by default provides interactive visuals
- Has ggplot2 integration



```
library(ggplot2)
library(plotly)

my.ggplot <- <insert ggplot code>

ggplotly(my.ggplot)


```



Resources



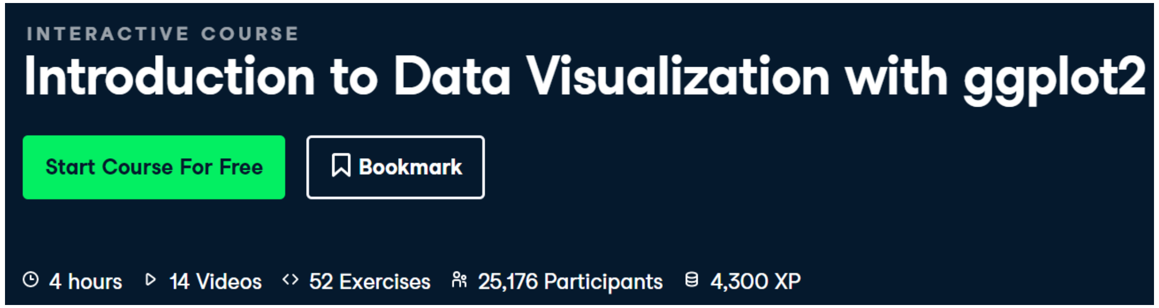
Data Visualization with ggplot2 : : CHEAT SHEET

Basics

Geoms

https://github.com/rstudio/cheatsheets/blob/master/data-visualization-2.1.pdf

https://learn.datacamp.com/courses/introduction-to-data-visualization-with-ggplot2

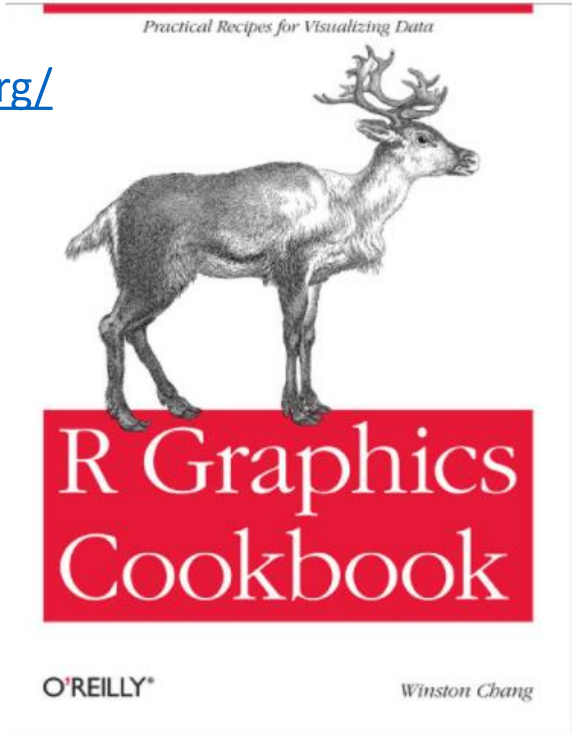


INTERACTIVE COURSE

Introduction to Data Visualization with ggplot2

4 hours · 14 Videos · 52 Exercises · 25,176 Participants · 4,300 XP

https://r-graphics.org/

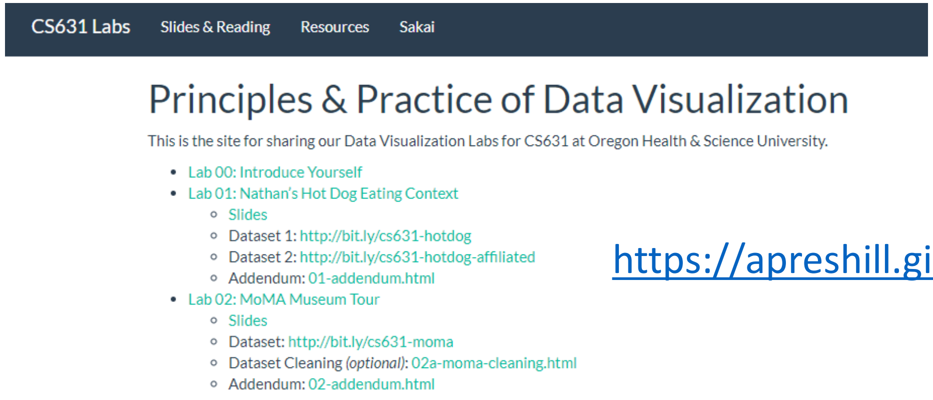


Practical Recipes for Visualizing Data

R Graphics Cookbook

O'REILLY

Winston Chang



Principles & Practice of Data Visualization

This is the site for sharing our Data Visualization Labs for CS631 at Oregon Health & Science University.

- Lab 00: Introduce Yourself
- Lab 01: Nathan's Hot Dog Eating Context
  - Slides
  - Dataset 1: http://bit.ly/cs631-hotdog
  - Dataset 2: http://bit.ly/cs631-hotdog-affiliated
  - Addendum: 01-addendum.html
- Lab 02: MoMA Museum Tour
  - Slides
  - Dataset: http://bit.ly/cs631-moma
  - Dataset Cleaning *(optional)*: 02a-moma-cleaning.html
  - Addendum: 02-addendum.html

https://apreshill.github.io/data-vis-labs-2018/