mine@stat.duke.edu
@minebocek
mine-cetinkaya-rundel


bit.ly/intro-stat-ds

2016 GAISE

1. Teach statistical thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and a purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

In addition to these six recommendations, which remain central, we suggest two new emphases for the first recommendation (teach statistical thinking) that reflect modern practice and take advantage of widely available technologies:

a. **Teach statistics as an investigative process of problem-solving and decision-making**. Students should not leave their introductory statistics course with the mistaken impression that statistics consists of an unrelated collection of formulas and methods. Rather, students should understand that statistics is a problem-solving and decision-making *process* that is fundamental to scientific inquiry and essential for making sound decisions.

b. **Give students experience with multivariable thinking.** We live in a complex world in which the answer to a question often depends on many factors. Students will encounter such situations within their own fields of study and everyday lives. We must prepare our students to answer challenging questions that require them to investigate and explore relationships among many variables. Doing so will help them to appreciate the value of statistical thinking and methods.

http://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf

1. Teach statistical thinking.

   a. **Teach statistics as an investigative process of problem-solving and decision-making.** Students should not leave their introductory statistics course with the mistaken impression that statistics consists of an unrelated collection of formulas and methods. Rather, students should understand that statistics is a problem-solving and decision-making *process* that is fundamental to scientific inquiry and essential for making sound decisions.

   b. **Give students experience with multivariable thinking.** We live in a complex world in which the answer to a question often depends on many factors. Students will encounter such situations within their own fields of study and everyday lives. We must prepare our students to answer challenging questions that require them to investigate and explore relationships among many variables. Doing so will help them to appreciate the value of statistical thinking and methods.

2. Focus on conceptual understanding.
3. Integrate real data with a context and a purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

(1) NOT a commonly used subset of tests and intervals and produce them with hand calculations

(2) Multivariate analysis requires the use of computing

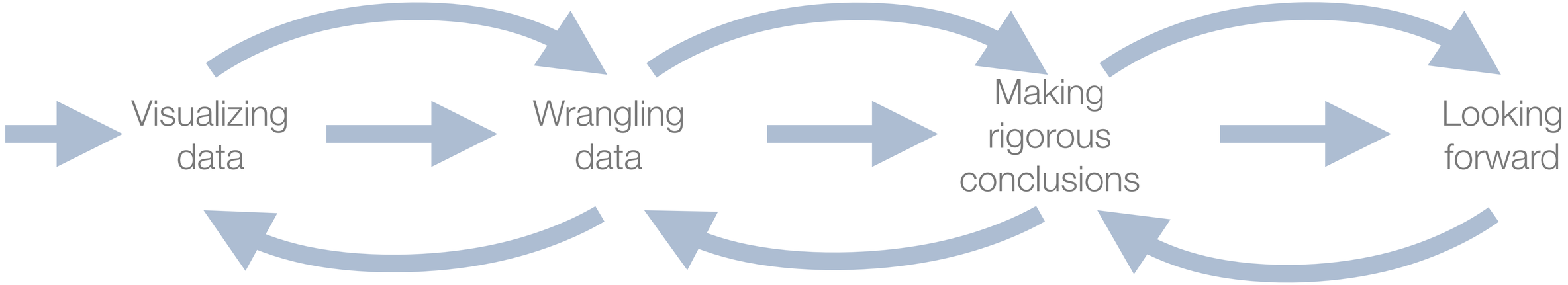
(3) NOT use technology that is only applicable in the intro course or that doesn't follow good science principles

(4) Data analysis isn't just inference and modeling, it's also data importing, cleaning, preparation, exploration, and visualization

So, what does this mean?

- A course that satisfies these four points is looking more like today's intro data science courses than (most) intro stats courses
- But this is not because intro stats is inherently "bad for you"
- Instead it is because it's time to visit intro stats in light of emergence of data science

An intro data science & statistical thinking curriculum



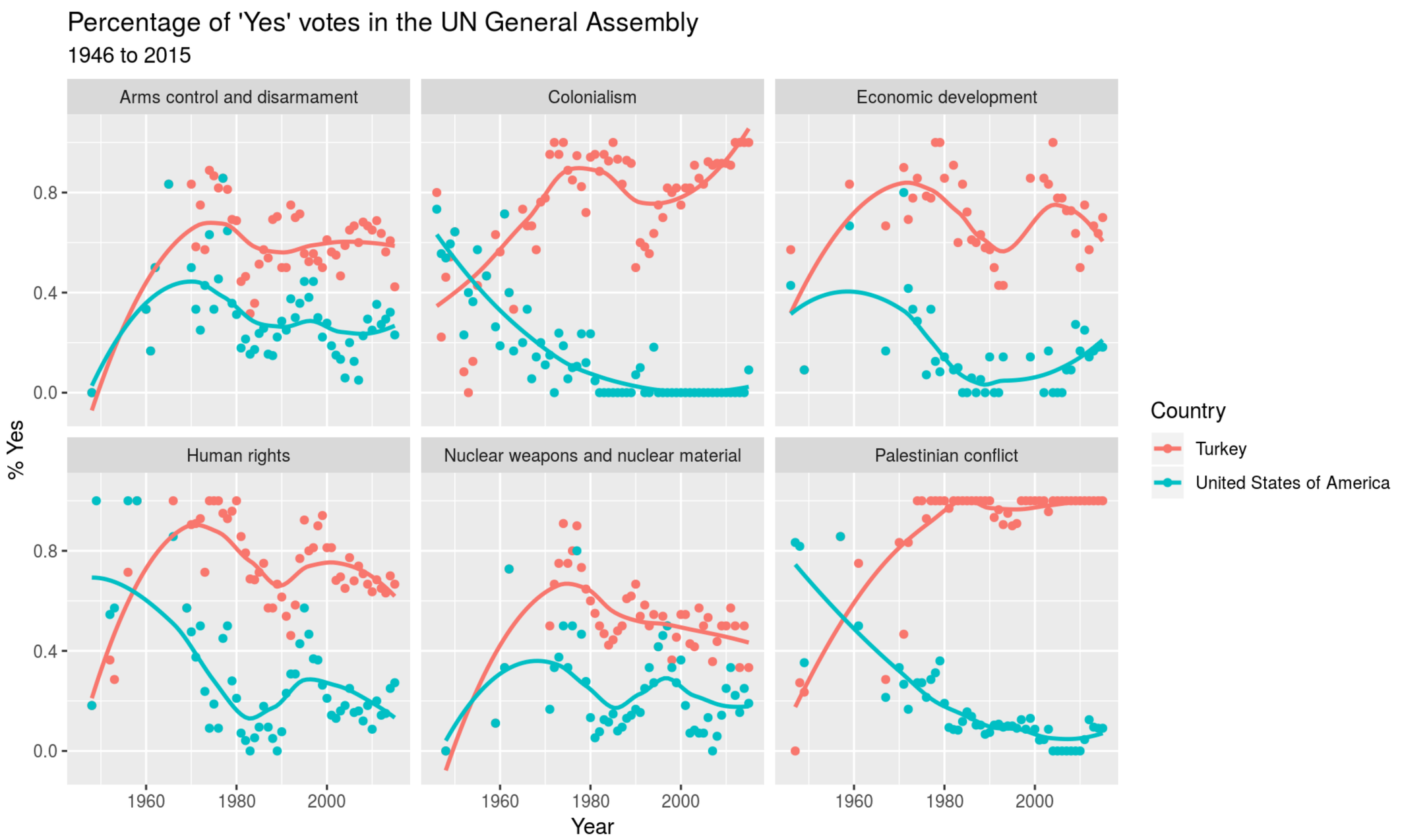
Fundamentals of data & data viz, confounding variables, Simpson's paradox (R + RStudio + R Markdown + git/GitHub)

Tidy data, data frames vs. summary tables, recoding and transforming variables, web scraping and iteration

Building and selecting models, visualizing interactions, prediction & model validation, inference via simulation

Interactive viz & reporting with Shiny, text analysis, Bayesian inference, ???

Ex 1. UN Votes

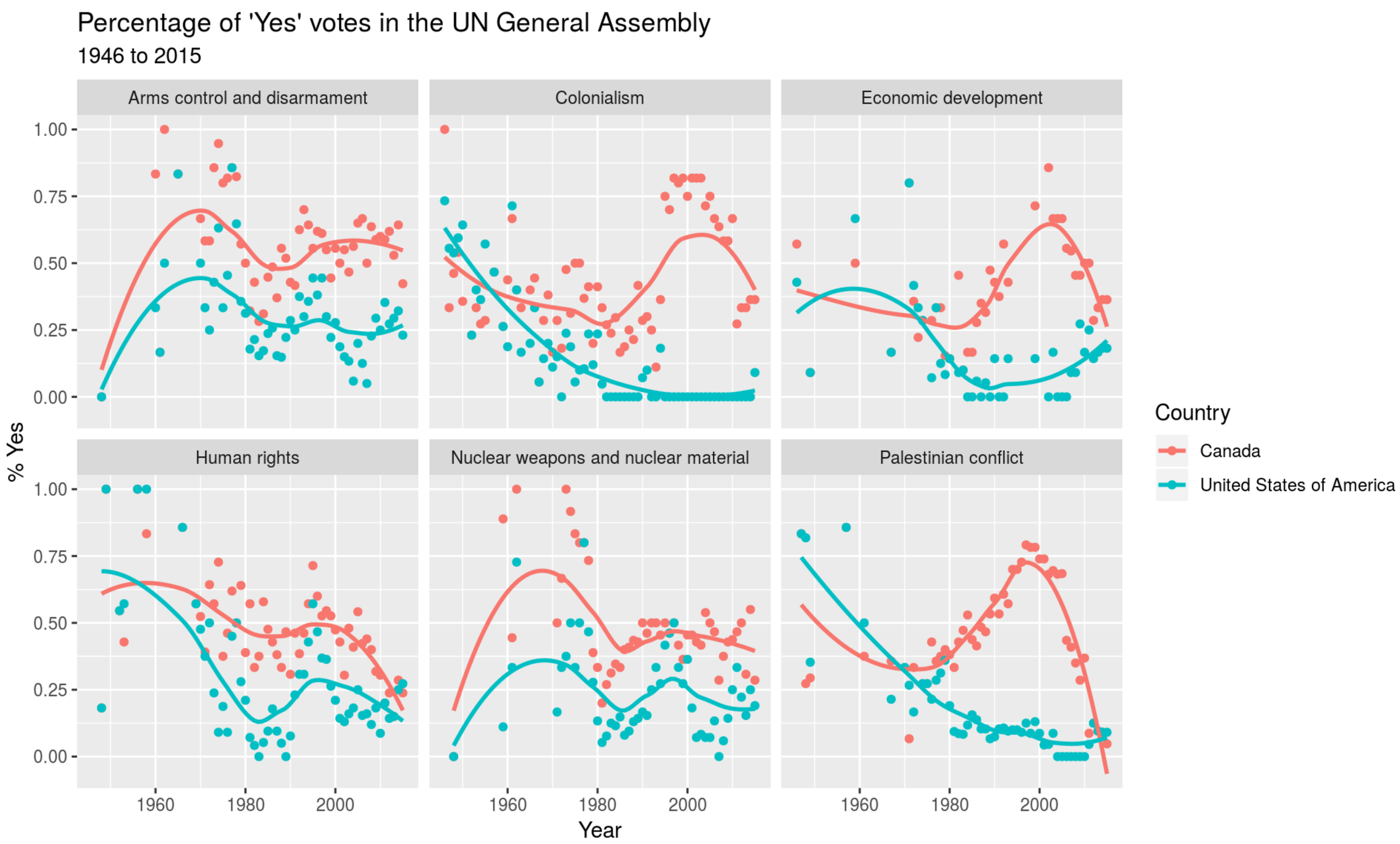


`unvotes`: Erik Voeten "Data and Analyses of Voting in the UN General Assembly" Routledge Handbook of International Organization, edited by Bob Reinalda (published May 27, 2013)

```r
un_votes %>%
  filter(country %in% c("United States of America", "Turkey")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(country, year = year(date), issue) %>%
  summarize(
    votes = n(),
    percent_yes = mean(vote == "yes")
    ) %>%
  filter(votes > 5) %>%  # only use records where there are more than 5 votes
  ggplot(mapping = aes(x = year, y = percent_yes, color = country)) +
    geom_point() +
    geom_smooth(method = "loess", se = FALSE) +
    facet_wrap(~ issue) +
    labs(
      title = "Percentage of 'Yes' votes in the UN General Assembly",
      subtitle = "1946 to 2015",
      y = "% Yes",
      x = "Year",
      color = "Country"
    )
```

```
un_votes %>%
  filter(country %in% c("United States of America", "Turkey")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(country, year = year(date), issue) %>%
  summarize(
    votes = n(),
    percent_yes = mean(vote == "yes")
    ) %>%
  filter(votes > 5) %>%  # only use records where there are more than 5 votes
  ggplot(mapping = aes(x = year, y = percent_yes, color = country)) +
    geom_point() +
    geom_smooth(method = "loess", se = FALSE) +
    facet_wrap(~ issue) +
    labs(
      title = "Percentage of 'Yes' votes in the UN General Assembly",
      subtitle = "1946 to 2015",
      y = "% Yes",
      x = "Year",
      color = "Country"
    )
```

```
un_votes %>%
  filter(country %in% c("United States of America", "Canada")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(country, year = year(date), issue) %>%
  summarize(
    votes = n(),
    percent_yes = mean(vote == "yes")
    ) %>%
  filter(votes > 5) %>%  # only use records where there are more than 5 votes
  ggplot(mapping = aes(x = year, y = percent_yes, color = country)) +
    geom_point() +
    geom_smooth(method = "loess", se = FALSE) +
    facet_wrap(~ issue) +
    labs(
      title = "Percentage of 'Yes' votes in the UN General Assembly",
      subtitle = "1946 to 2015",
      y = "% Yes",
      x = "Year",
      color = "Country"
    )
```

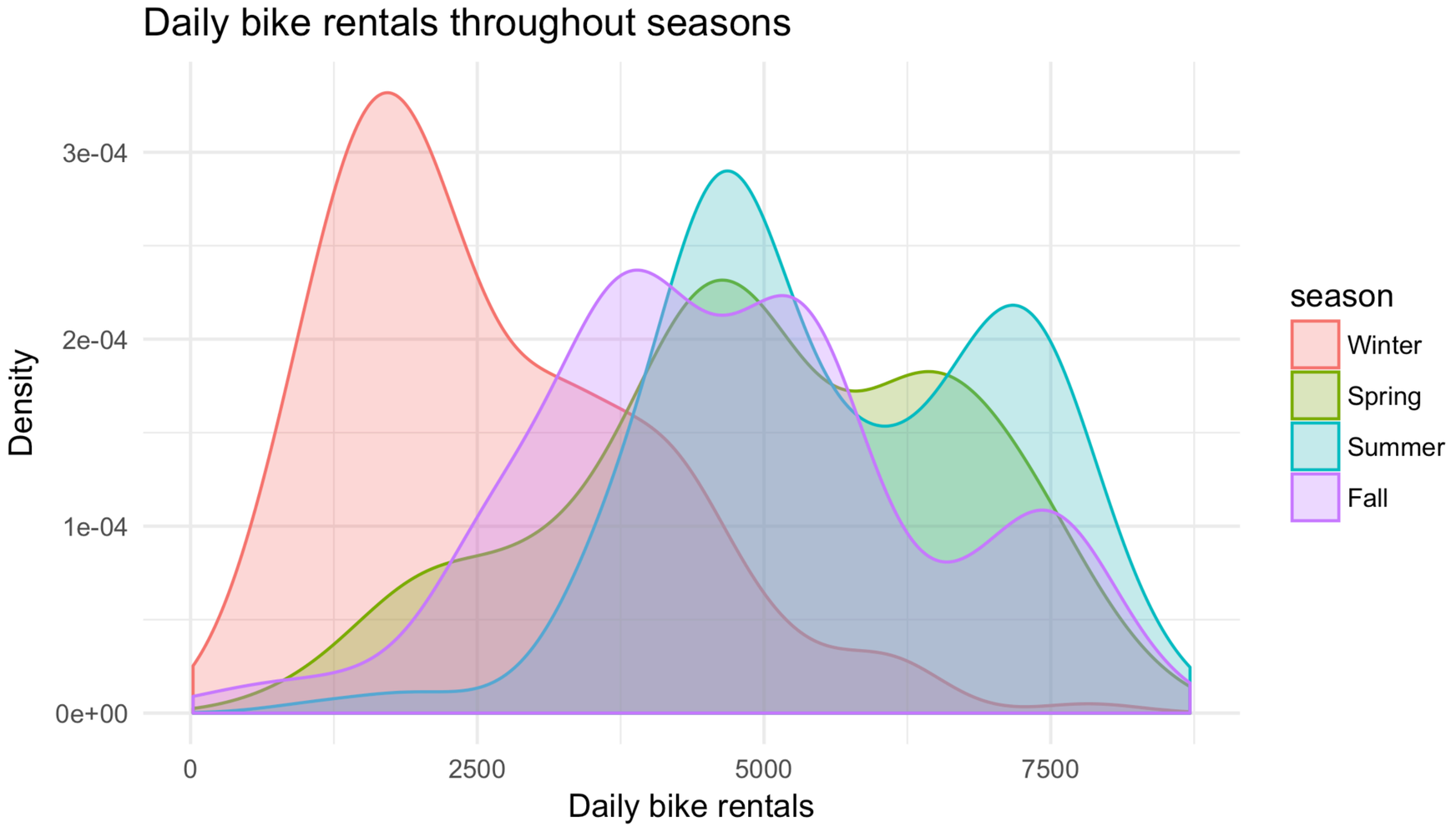

Percentage of 'Yes' votes in the UN General Assembly
1946 to 2015
Arms control and disarmament
Colonialism
Economic development
Human rights
Nuclear weapons and nuclear material
Palestinian conflict
% Yes
1.00
0.75
0.50
0.25
0.00
1960
1980
2000
Year
Country
Canada
United States of America

Learning goals

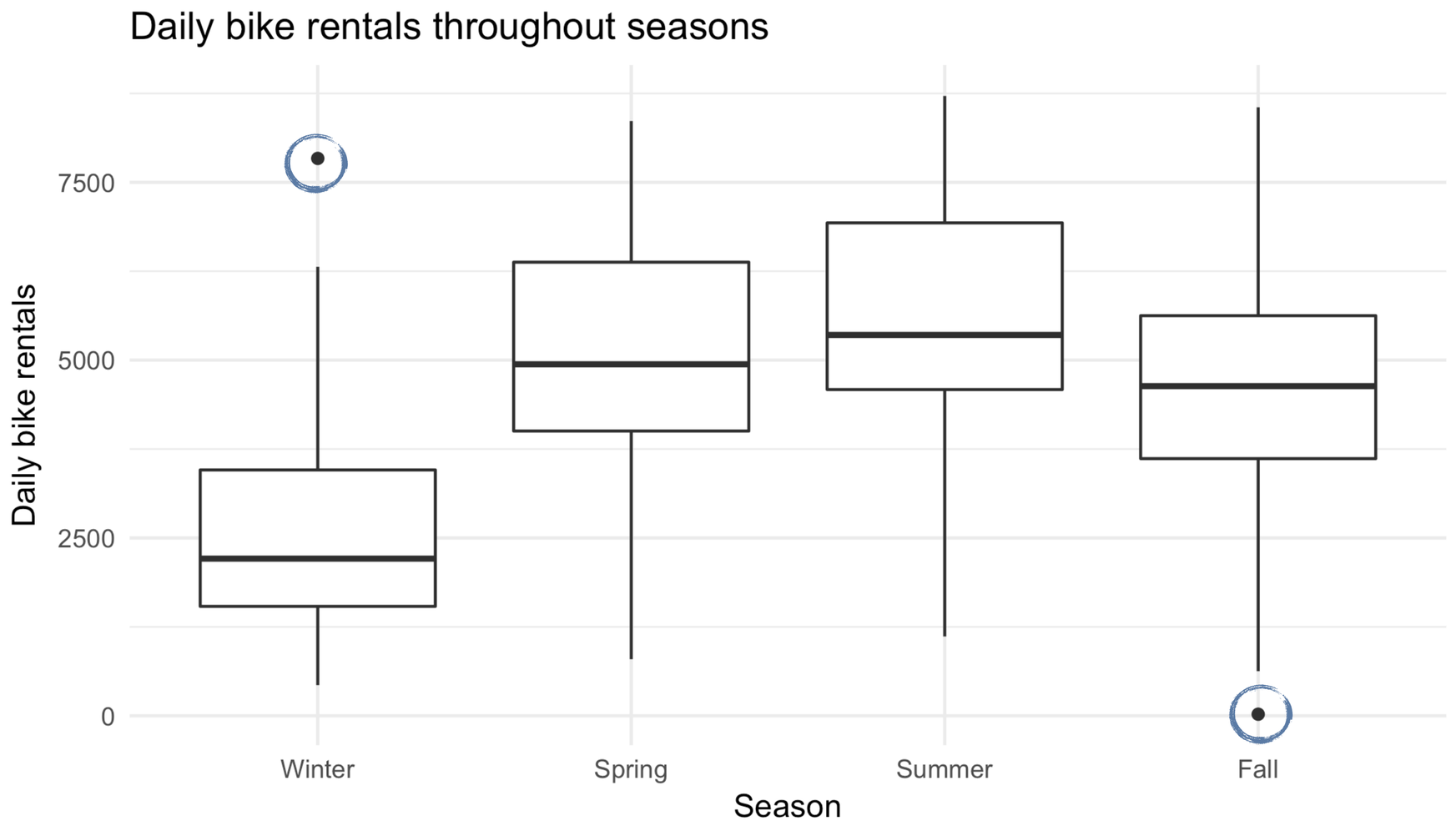
- **Main:** Multivariate data visualization on day one of class
- **Get for free:** Your first experience writing code on day one of class

Ex 2. DC bike rentals

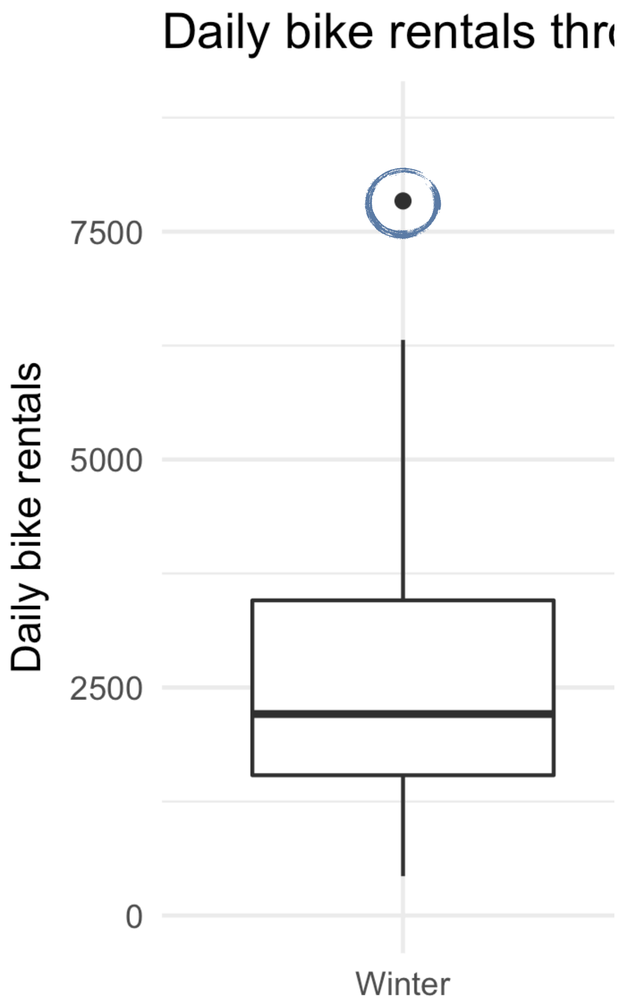
Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.



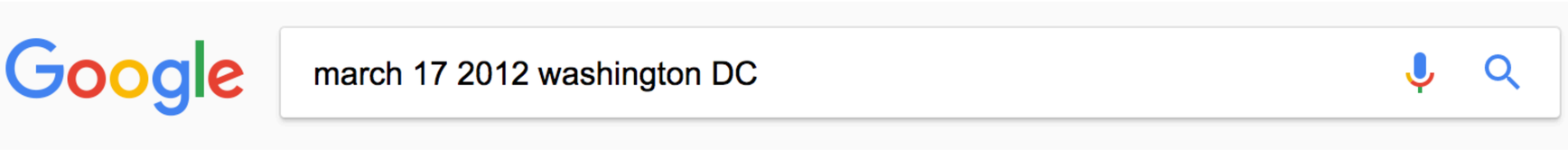
Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.



Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.

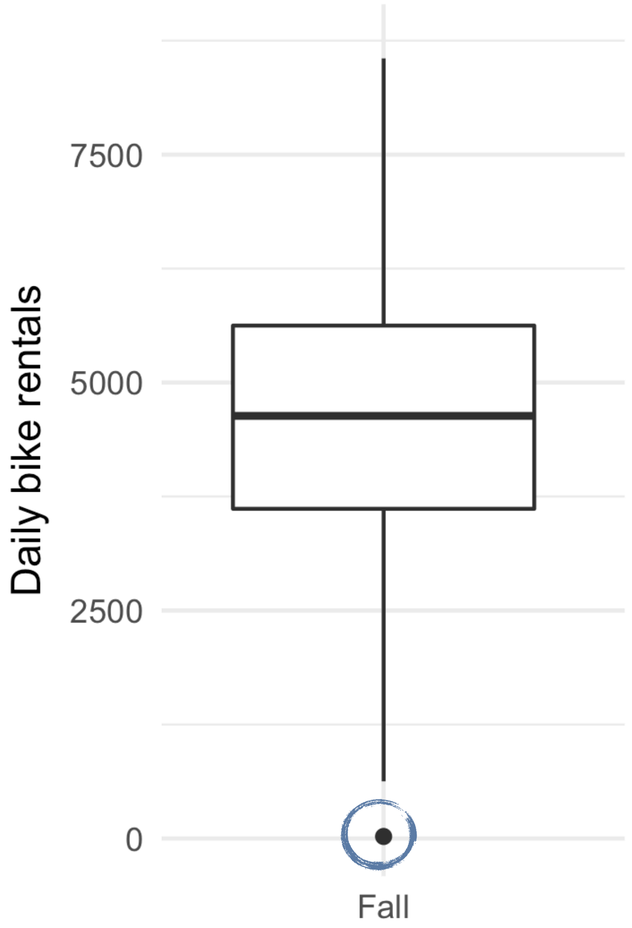



```
bike %>%
  filter(season == "Winter") %>%
  summarise(max = max(cnt), day_min = dteday[which.max(cnt)])
## # A tibble: 1 x 2
##     min day_min
##   <dbl> <date>
## 1  7836 2012-03-17
```

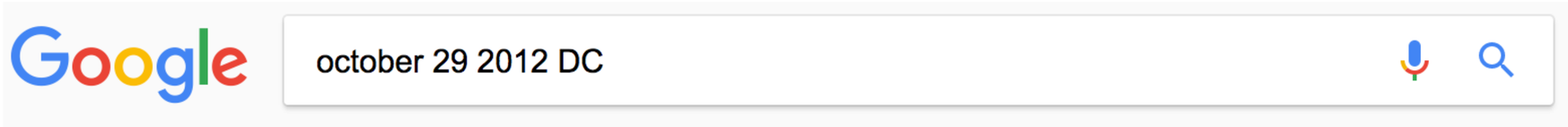


President Obama at the Dubliner on St. Patrick's Day | whitehouse.gov
https://obamawhitehouse.archives.gov/.../2012/.../17/president-obama-dubliner-st-patr...
17 Mar 2012 - President Barack Obama is reflected in a mirror at the Dubliner, an Irish pub in **Washington, D.C.**, with his Irish cousin, Henry Healy, and Ollie Hayes, a pub owner in Moneygall, Ireland, on St. Patrick's Day, **Saturday, March 17, 2012**. (Official White House Photo by Pete Souza). President Obama Greets the ...

Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.




```
bike %>%
  filter(season == "Fall") %>%
  summarise(min = min(cnt), day_min = dteday[which.min(cnt)])
## # A tibble: 1 x 2
##     min day_min
##   <dbl> <date>
## 1  22.0 2012-10-29
```



Washington DC shuts down in preparation for hurricane Sandy | US ...
https://www.theguardian.com/.../2012/oct/29/washington-dc-shutdown-hurricane-sand...
28 Oct 2012 - Panic buying and fears over capability of power companies as first light rains reach city.

Hurricane Sandy: high winds and flooding hit US east coast – Monday ...
https://www.theguardian.com/.../us.../2012/oct/29/hurricane-sandy-new-york-live-blo...
30 Oct 2012 - Announces plans for a Storm Recovery Support Call Center operated by @ Delaware_DHSS will open tomorrow at noon. Gov. Markell: The message for tonight is to hunker down. Weather Underground (@wunderground). 470,000 without power in 13 states and **DC**. **October 29**,

Learning goals

- **Main:** Prediction and model selection
- **Get for free:** Use of outside data

Ex 3. Paris paintings

89 Deux tableaux très riches de compoſition, d'une belle exécution, & dont le mérite eſt très remarquable, chacun de 17 pouces 3 lignes de haut, ſur 23 pouces de large; le premier, peint ſur bois, vient du Cabinet de Madame la Comteſſe de Verrue; il repréſente un départ pour la chaſſe : on y voit ſur le devant un

10660

Le ſecond tableau, qui eſt ſur toile, fait voir un terrein d'une grande étendue, près la mer qui eſt à gauche, & ſur laquelle ſont des vaiſſeaux : on y voit auſſi des bagages que l'on décharge d'un charriot, des hommes, des femmes, des enfants, deux chevaux qui mangent, & des mulets chargés de bagages.

Two paintings very rich in composition, of a beautiful execution, and whose merit is very remarkable, each 17 inches 3 lines high, 23 inches wide; the first, painted on wood, comes from the Cabinet of Madame la Comtesse de Verrue; it represents a departure for the hunt: it shows in the front a child on a white horse, a man who gives the horn to gather the dogs, a falconer and other figures nicely distributed across the width of the painting; two horses drinking from a fountain; on the right in the corner a lovely country house topped by a terrace, on which people are at the table, others who play instruments; trees and fabriques pleasantly enrich the background.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	name	sale	lot	dealer	year	origin_author	origin_cat	school_pntg	diff_origin	price	count	subject	authorstandard	artistliving	authorstyle	author	winnir
2517	R1777-86	R1777	86	R	1777	D/FL	D/FL	D/FL	0	620.0	1	2 femmes, enfants, paysage vu à travers une arcade	Bega, Cornelis Pieterszoon	0	n/a	Corneille Bega	Lebrun
2518	R1777-87	R1777	87	R	1777	D/FL	D/FL	D/FL	0	12,000.0	1	Course du hareng	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Donjeu
2519	R1777-88	R1777	88	R	1777	D/FL	D/FL	D/FL	0	8,000.0	1	Paysage sablonneux	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Lambe
2520	R1777-89a	R1777	89	R	1777	D/FL	D/FL	D/FL	0	5,300.0	1	Départ pour la chasse	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Langlie

	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	winningbidder	winningbiddertype	endbuyer	Interm	type_interm ed	Height_in	Width_in	Surface_Rect	Diam_in	Surface_Rnd	Shape	Surface	material	mat	quantity	nfigures	engraved
2516	Feuillet	D	D	0		16	20	320			squ_rect	320	toile	t	1	0	0
2517	Lebrun, Jean-Baptiste-Pierre	D	D	0		13.25	11	145.75			squ_rect	145.75	bois	b	1	0	0
2518	Donjeux, Vincent	D	D	0		23	29.25	672.75			squ_rect	672.75	toile	t	1	50	0
2519	Lambert, John (Chevalier Lambert)	C	C	0		23	30	690			squ_rect	690	toile	t	1	0	1
2520	Langlier, Jacques for Poullain, Antoine	DC	C	1	D	17.25	23	396.75			squ_rect	396.75	bois	b	1	0	0

	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF
1	nfigures	engraved	original	prevcoll	othartist	paired	figures	finished	lrgfont	relig	landsALL	lands_sc	lands_figs	lands_ment	arch	mytho	peasant	othgenre	singlefig	portrait	still_life	discauth	history	allegory	pastorale	other
2516	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2517	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
2518	50	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
2519	0	1	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2520	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

- **mat** - category of material (a=silver, al=alabaster, ar=slate, b=wood, bc=wood and copper, br=bronze frames, bt=canvas on wood, c=copper, ca=cardboard, co=cloth, e=wax, g=grissaille technique, h=oil technique, m=marble, mi=miniature technique, o=other, p=paper, pa=pastel, t=canvas, ta=canvas?, v=glass, n/a=NA, (blanks)=NA)
- **Shape** - shape of painting

```
pp <- pp %>%
  mutate(
    Shape = fct_collapse(Shape, oval = c("oval", "ovale"),
                                round = c("round", "ronde"),
                                squ_rect = "squ_rect",
                                other = c("octogon", "octagon", "miniature")),
    mat = fct_collapse(mat, metal = c("a", "br", "c"),
                            canvas = c("co", "t", "ta"),
                            paper = c("p", "ca"),
                            wood = "b",
                            other = c("e", "g", "h", "mi", "o", "pa", "v", "al", "ar", "m"))
  )
```

Learning goals

- **Main:** data provenance + modelling diagnostic, log transformation
- **Get for free:** iterative data cleanup informed by analysis results + experience working with #otherpeoplesdata

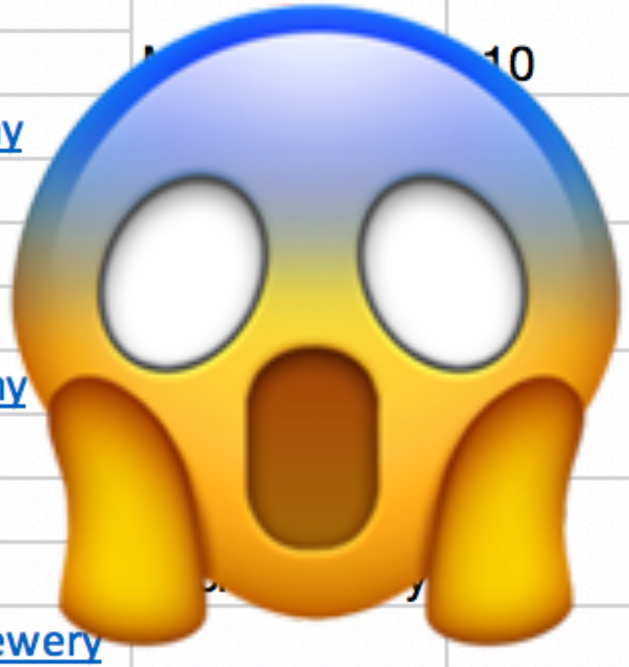
Ex 4. Breweries

North Carolina Breweries

241 active | 25 closed

Name	Type	Beer Count	My Count	Est.
1718 Brewing Ocracoke Ocracoke	Brewpub	3	-	2018
217 Brew Works Wilson	Microbrewery	10	-	2017
3rd Rock Brewing Company Trenton	Microbrewery	13	-	2016
7 Clans Brewing Cherokee	Client Brewer	1	-	2018
Andrews Brewing Company Andrews	Microbrewery	18	-	2014
Angry Troll Brewing Elkin	Microbrewery	8	-	2017
Appalachian Mountain Brewery Boone	Microbrewery	79	-	2013
Archetype Brewing Asheville	Microbrewery	24	-	2017

	A	B	C	D	E
1	**Name**	**Type**	**Beer Count**	**My Count**	**Est.**
2	1718 Brewing Ocracoke				
3	Ocracoke	Brewpub	3	-	2018
4	217 Brew Works				
5	Wilson		10	-	2017
6	3rd Rock Brewing Company				
7	Trenton			-	2016
8	7 Clans Brewing				
9	Cherokee			-	2018
10	Andrews Brewing Company				
11	Andrews			-	2014
12	Angry Troll Brewing				
13	Elkin			-	2017
14	Appalachian Mountain Brewery				
15	Boone	Microbrewery	79	-	2013
16	Archetype Brewing				
17	Asheville	Microbrewery	24	-	2017



```
library(tidyverse)
library(rvest)

page <- read_html("https://www.ratebeer.com/breweries/north%20carolina/33/213/")

names <- page %>%
  html_nodes("#brewerTable a:nth-child(1)") %>%
  html_text() %>%
  str_trim()

active_cities <- page %>%
  html_nodes(".filter") %>%
  html_text()

closed_cities <- page %>%
  html_nodes("#brewerTable span") %>%
  html_text()

cities <- c(active_cities, closed_cities)

…

ncbreweries <- tibble(
  name = names,
  city = cities,
  …
)

write_csv(ncbreweries, path = "data/ncbreweries.csv")
```

	name	city	type	beercount	est	status	url
1	1718 Brewing Ocracoke	Ocracoke	Brewpub	3	2018	Active	https://www.ratebeer.com//brewers/171
2	217 Brew Works	Wilson	Microbrewery	10	2017	Active	https://www.ratebeer.com//brewers/217
3	3rd Rock Brewing Company	Trenton	Microbrewery	13	2016	Active	https://www.ratebeer.com//brewers/3rd-
4	7 Clans Brewing	Cherokee	Client Brewer	1	2018	Active	https://www.ratebeer.com//brewers/7-cl
5	Andrews Brewing Company	Andrews	Microbrewery	18	2014	Active	https://www.ratebeer.com//brewers/and
6	Angry Troll Brewing	Elkin	Microbrewery	8	2017	Active	https://www.ratebeer.com//brewers/ang
7	Appalachian Mountain Brewery	Boone	Microbrewery	79	2013	Active	https://www.ratebeer.com//brewers/app
8	Archetype Brewing	Asheville	Microbrewery	24	2017	Active	https://www.ratebeer.com//brewers/arch
9	Asheville Brewing Company	Asheville	Brewpub	88	2003	Active	https://www.ratebeer.com//brewers/ashe
10	Ass Clown Brewing Company	Cornelius	Microbrewery	108	2011	Active	https://www.ratebeer.com//brewers/ass-
11	Aviator Brewing Company	Fuquay Varina	Microbrewery	60	2008	Active	https://www.ratebeer.com//brewers/avia
12	Balsam Falls Brewing Company	Sylva	Microbrewery	25	2018	Active	https://www.ratebeer.com//brewers/bals
13	Barking Duck Brewing Company	Mint Hill	Microbrewery	16	2014	Active	https://www.ratebeer.com//brewers/bark
14	Barrel Culture Brewing and Blending	Durham	Microbrewery	40	2017	Active	https://www.ratebeer.com//brewers/barr

Learning goals

- **Main:** data harvesting
- **Get for free:** working with text data + iteration

Myths

1. Students aren’t interested in learning programming

2. It’s not possible to teach statistical concepts and programming in just one course

3. Teaching programming takes up valuable time that can otherwise be used towards teaching important statistical concepts

So, do we need both

… intro data science and intro stats?

- Yes, and no
- No need to frame data science as a technical field that only students with certain (computational) interest and experience are interested in
- Also no need to think of the intro stats course as the course where students who don’t fall in that bucket go into

Goal

Learn from both courses to come up with a course

- that addresses current guidelines
- is modern and current
- and with sufficient resources to help faculty who are new to it teach it

bit.ly/dsbox-web

bit.ly/dsbox-repo

Search...

Hello #dsbox

Course content

Technology stack

Pedagogy

Built with ♥ and blogdown

Data Science in a Box

How can we effectively and efficiently teach data science to students with little to no background in computing and statistical thinking? How can we equip them with the skills and tools for reasoning with various types of data and leave them wanting to learn more?

This introductory data science course that is our (working) answer to this question. The core content of the course focuses on data acquisition and wrangling, exploratory data analysis, data visualization, inference, modeling, and effective communication of results. Time permitting, the course also introduces additional concepts and tools like interactive visualization and reporting Bayesian inference. A heavy emphasis is placed on a consitent syntax (with tools from the tidyverse), reproducibility (with R Markdown) and version control and collaboration (with git/GitHub). In addition, out-of-class learning is supplemented with interactive tutorials. The goal of the course is to bring students from zero to being able to work in a team to complete a fully reproducible data analysis project on a dataset of their choice and answering questions they care about.

Data Science in a Box contains the materials required to teach (or learn from) the course described above, all of which are freely-available and open-source. They include course materials such as slide decks, homework assignments, guided labs, sample exams, a final project assignment, as

So, everyone goes into the same course?

It depends…

- How many students are you serving, and will you need to split them into separate sections anyway?
  - Suggestion: Split based on those planning on taking only 1-2 stats courses anyway vs. those planning on a quantitative major
  - Students can change their mind, but this will serve most well.
- Do the courses serving these two audiences differ?
  - Potentially…
  - Think about what is essential for students to be exposed to in the first (maybe only) course vs. what can wait till a second course?
  - E.g. Computing and reproducibility is non-negotiable, but could version control wait?

Intro

Data Science

Stats

Do we need both?

bit.ly/intro-stat-ds


mine@stat.duke.edu


@minebocek


mine-cetinkaya-rundel