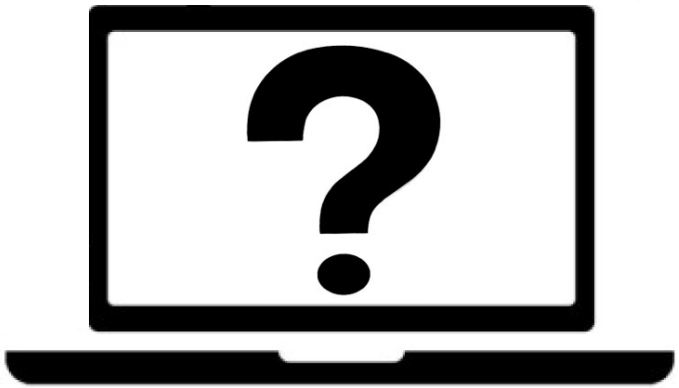
Channeling Curiosity into R Data Visualization Projects



Alex Albright, Harvard University

Aspiring Economist & Data Visualization Enthusiast

R Ladies

February 20, 2018

Agenda

- Introduce myself and how I got into 
- Demonstrate how questions can lead to R projects
- Present data visualizations and code for select projects
  - Teaser: will discuss *Sports Illustrated*, science degrees, Pixar, the Senate, and *Friends*!
- Talk about workflow: generating questions, using notebooks, writing code, sharing product/code

Aspiring Economist

I am a 2nd year PhD student.

Table 4: Impact of Award assigned at Year t on Math Olympiad (MO) Performance at Year $t+1$ (Classmates)

Panel B. Classmates	Participated in MO exam at Year t+1			
	(1)	(2)	(3)	(4)
Award	0.0014***	0.0014***	0.0013***	0.0011**
	(0.0005)	(0.0005)	(0.0004)	(0.0004)
Students (obs.)	3,540,290	3,540,290	5,114,922	4,899,094
Classrooms (Clusters)	117,882	117,882	170,331	163,337
Dep. variable control mean	.024	.024	.024	.024
Bdw selection	h= .62 (Min)	h= .62 (Min)	h= .82 (Avg)	h= .79 (optimal)
Controls	No	Yes	Yes	Yes
Panel B. Classmates	MO score at Year t+1			
	(1)	(2)	(3)	(4)
Award	0.0362**	0.0409***	0.0293**	0.0367**
	(0.0166)	(0.0147)	(0.0124)	(0.0152)
Students (obs.)	89,872	89,872	127,331	84,033
Classrooms (Clusters)	54,213	54,213	77,191	50,554
Dep. variable control s.d.	.82	.82	.81	.82
Bdw selection	h= .62 (Min)	h= .62 (Min)	h= .82 (Avg.)	h= .57 (optimal)
Controls	No	Yes	Yes	Yes

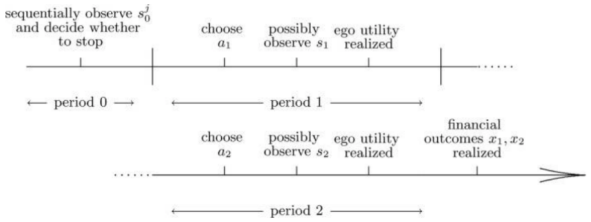
The expectation on the right hand side of (2) is found by integrating over the range of Q, which splits into two parts, depending on which of the two expressions yields the max. Therefore

$$\begin{aligned} V_m &= \int_0^{Q_m} \beta V_m \, dq + \int_{Q_m}^1 (q + \beta V_{m-1}) \, dq \\ &= \beta V_m Q_m + \tfrac{1}{2}\left[1 - (Q_m)^2\right] + \beta V_{m-1}(1 - Q_m) \\ &= \beta (V_m - V_{m-1}) Q_m + \tfrac{1}{2} - \tfrac{1}{2}(Q_m)^2 + \beta V_{m-1} \\ &= (Q_m)^2 + \tfrac{1}{2} - \tfrac{1}{2}(Q_m)^2 + \beta V_{m-1} \\ &= \tfrac{1}{2} + \tfrac{1}{2}(Q_m)^2 + \beta V_{m-1} \\ &= \tfrac{1}{2} + \tfrac{1}{2}\beta^2 (V_m - V_{m-1})^2 + \beta V_{m-1}. \end{aligned}$$

Note that every choice function is a completion of itself. Moreover, it is generally the case that nontrivial completions of a hospital's choice function exist. For an example, consider the choice function C^{h} of hospital h induced by the preference relation given in the Sherlock–Watson example in the Introduction,

$$\succ_{\mathsf{h}} : \{\mathsf{S}^{\mathsf{r}}, \mathsf{W}^{\mathsf{c}}\} \succ \{\mathsf{S}^{\mathsf{c}}\} \succ \{\mathsf{W}^{\mathsf{c}}\} \succ \{\mathsf{S}^{\mathsf{r}}\} \succ \varnothing.$$

Note that the choice function C^{h} induced by the preference relation $\succ_{\mathsf{h}}$ is not substitutable, as $\mathsf{S}^{\mathsf{r}} \notin \{\mathsf{S}^{\mathsf{c}}\} = C^{\mathsf{h}}(\{\mathsf{S}^{\mathsf{r}}, \mathsf{S}^{\mathsf{c}}\})$, while $\mathsf{S}^{\mathsf{r}} \in \{\mathsf{S}^{\mathsf{r}}, \mathsf{W}^{\mathsf{c}}\} = C^{\mathsf{h}}(\{\mathsf{S}^{\mathsf{r}}, \mathsf{S}^{\mathsf{c}}, \mathsf{W}^{\mathsf{c}}\})$.



1 T2 pure & T1 mix?

This is not possible since the indifference condition fails to hold [both if T2 plays W and if T2 plays S]:

$$u_1(W,W) = 4 > 2 = u_1(S,W)$$
$$u_1(S,S) = 2 > -2 = u_1(W,S)$$

2 T1 pure & T2 mix?

This is not possible since the indifference condition fails to hold [both if T1 plays W and if T1 plays S]:

$$u_2(W,S) = 2 > -2 = u_2(W,W)$$
$$u_2(S,W) = -2 > -4 = u_2(S,S)$$

3 Both mix?

We know that since we have a finite game, there must exist a Nash equilibrium. Since we have shown that neither team can play pure strategies, it must be the case that both mix.

T1 must be indifferent between playing W and S. Say T1 plays W with probability α and S with probability $1-\alpha$, then we set

$$4\alpha - 2(1-\alpha) = 2\alpha + 2(1-\alpha)$$
$$\alpha = \frac{2}{3}$$

Moreover, T2 must also be indifferent between playing W and S. Say T2 plays W with probability β and S with probability $1-\beta$, then we set

$$-2\beta - 2(1-\beta) = 2\beta - 4(1-\beta)$$
$$\beta = \frac{1}{3}$$

Thus, the mixed Nash equilibrium is: $(\frac{2}{3}W \oplus \frac{1}{3}S, \frac{1}{3}W \oplus \frac{2}{3}S)$

Table 3: Treatment Impacts on Perceptions about Exam Performance

	Abs(Gap)	SD Beliefs	Gap	Abs(Gap)
	(1)	(2)	(3)	(4)
Treatment	-6.596***	-2.626***	-1.967**	-10.421***
	(0.642)	(0.420)	(0.787)	(1.249)
TreatX(Prior-Mock Exam Score)			-0.368***	
			(0.036)	
(Prior-Mock Exam Score)			0.612***	
			(0.029)	
TreatX(SD Prior)				0.215***
				(0.053)
SD Prior				-0.205***
				(0.039)
Mean Dep. Var. in Control	19.59	17.45	15.96	19.59
Number of Obs	2293	2293	2293	2293
R-squared	0.290	0.083	0.539	0.299
Number of Clusters	90	90	90	90

NOTE: * significant at 10%; ** significant at 5%; *** significant at 1%.
Strata dummies included in all specifications, standard errors clustered at the school level. Sample of ninth graders in schools who belong to the treated and the control group. All specifications include the following set of pre-determined individual and school characteristics (see Table 2 for details): gender (male), previous mock-test, previous mock-test with results, attendance to preparatory course, morning shift, both parents in the household, parents with higher education, SES index (above median), currently working, plan to attend college, and a dummy for whether one of the above has a missing value.

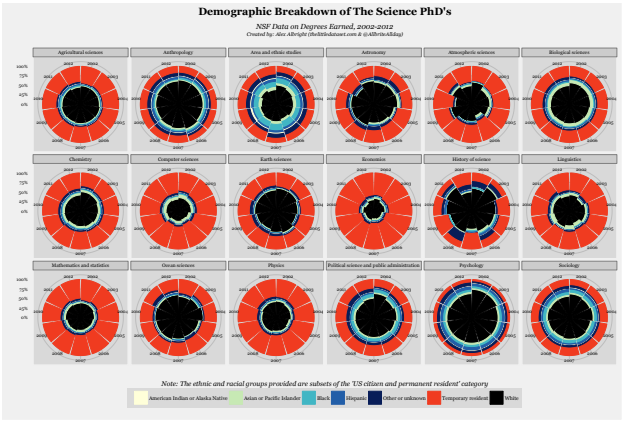
A pair of beliefs for employers about the two groups will be *self-confirming* if, by choosing standards optimal for those beliefs, employers induce workers from the two groups to become qualified at precisely the rate postulated by the beliefs. Thus, we can define equilibrium as follows.

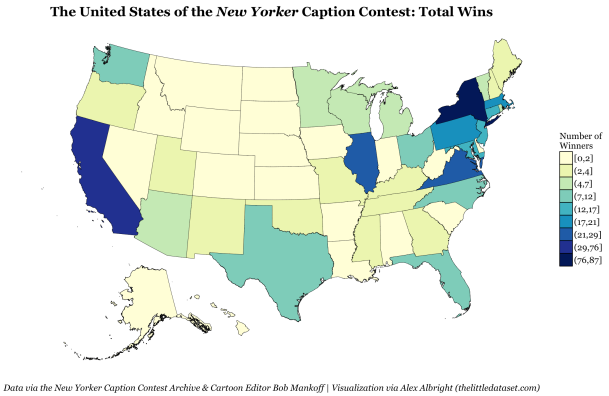
Definition 1: An *equilibrium* is a pair of beliefs $(\pi_{\mathrm{b}}, \pi_{\mathrm{w}})$ satisfying[9]

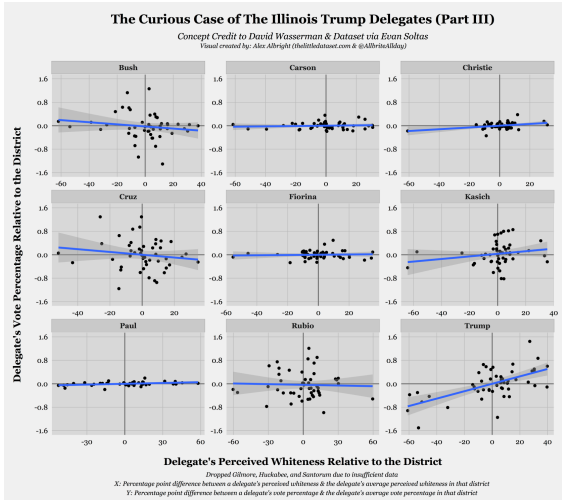
$$(3) \qquad \pi_i = G(\beta(s^*(\pi_i))) \qquad i = \mathrm{b}, \mathrm{w}.$$

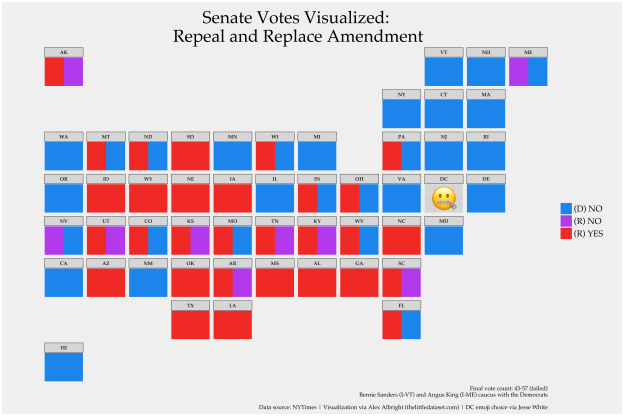
Data Visualization Enthusiast

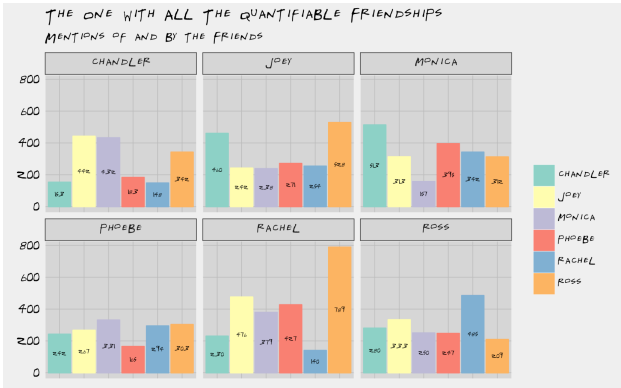
I share my work at `thelittledataset.com`.

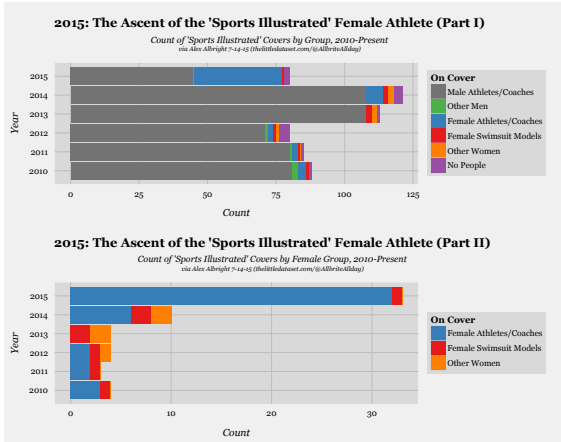






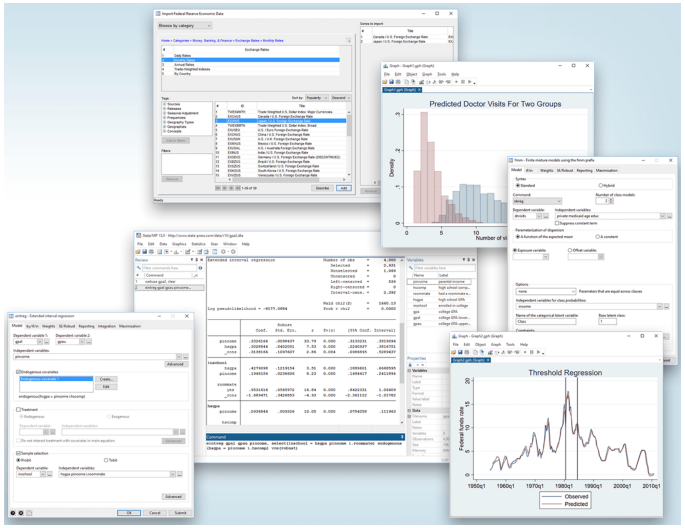






Why R?

Well, economists actually mainly use STATA®



- So, I started coding in Stata
- Eventually realized I couldn't easily share my work
- Stata is not free!
- Limited to those with institutional or industry access
  - Isolates academic research: makes it less publicly accessible & more black box-y

R Ladies

Why R?

In contrast to Stata...  is free and open source!

And...

- it can be faster too
- it has a vibrant/active/helpful community
  - check out `#rstats` twitter
  - R-ladies organizations
- data visualization options are beautiful!
  - thanks `ggplot2`
  - more on this soon...

Now what?

There are tons of great resources for learning R as well as tons of great ways to use R.

I wanted to learn to **build data visualizations to answer questions visually**.

By making data visualizations, you practice data storytelling. You learn how to:

- formulate interesting questions
- find/collect datasets
- clean and shape data
- identify the heart of data-driven results $\approx$ be conceptually succinct
- practice intriguing and friendly design
- use packages and customize plots

Illustrating by example

I will show how curiosity and questions can fuel R data visualization projects:

- How often are women on the cover of *Sports Illustrated* (and not in swimsuits)?
- How do sciences compare in their %s of women and ethnic/racial groups?
- How has Pixar movie success changed over time?
- How did all US Senators vote on healthcare repeal?
- Which characters are closest on the TV show *Friends*?

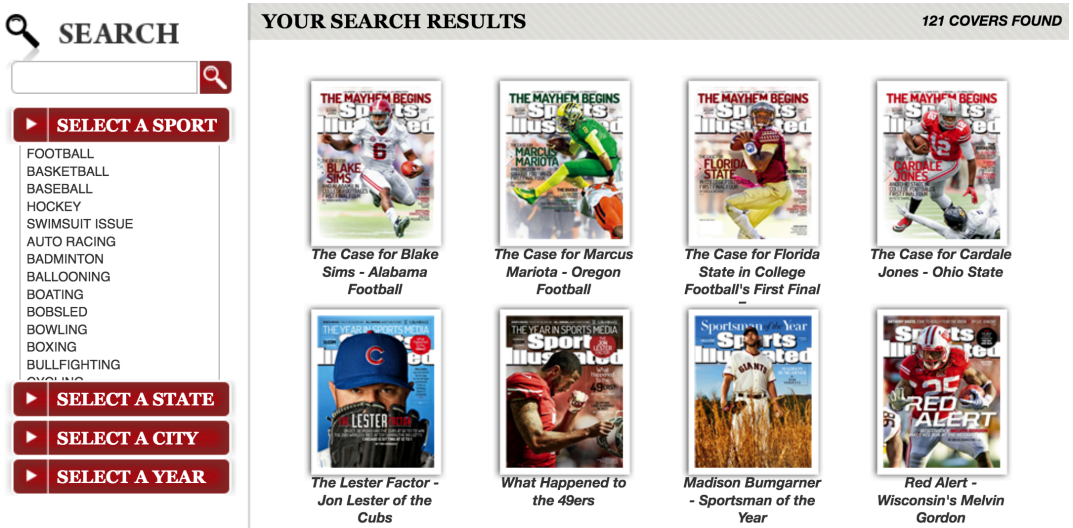
For each question:

- Outline motivation/question
- Present "visual answer"
- Present key code snippets (workhorse package for almost everything = `ggplot2`!)

How often are women on the cover of *Sports Illustrated* (and not in swimsuits)?

How often are women on the cover of *Sports Illustrated* (and not in swimsuits)?

- Comedian John Oliver did a "How Is This Still a Thing?" piece on the swimsuit issue
- Became curious about gender representation in the magazine
- Collected data from the online *Sports Illustrated* Covers Archive for years 2010-2014



How often are women on the cover of *Sports Illustrated* (and not in swimsuits)?

- Five-year time span and 487 covers (often multiple covers for same week)
- Counted the number of covers featuring:
  - male athletes/coaches (448)
  - other miscellaneous men (4)
  - female athletes/coaches (13)
  - female swimsuit models (7)
  - other miscellaneous women (5)
  - covers featuring no individual or group of individuals in particular (10)
- For the few dozen with men and women on the cover, count as female covers unless obviously male athlete at center

How often are women on the cover of *Sports Illustrated* (and not in swimsuits)?

If anything, I overestimate women's representation.

A cover featuring a male athlete (woman in the background)

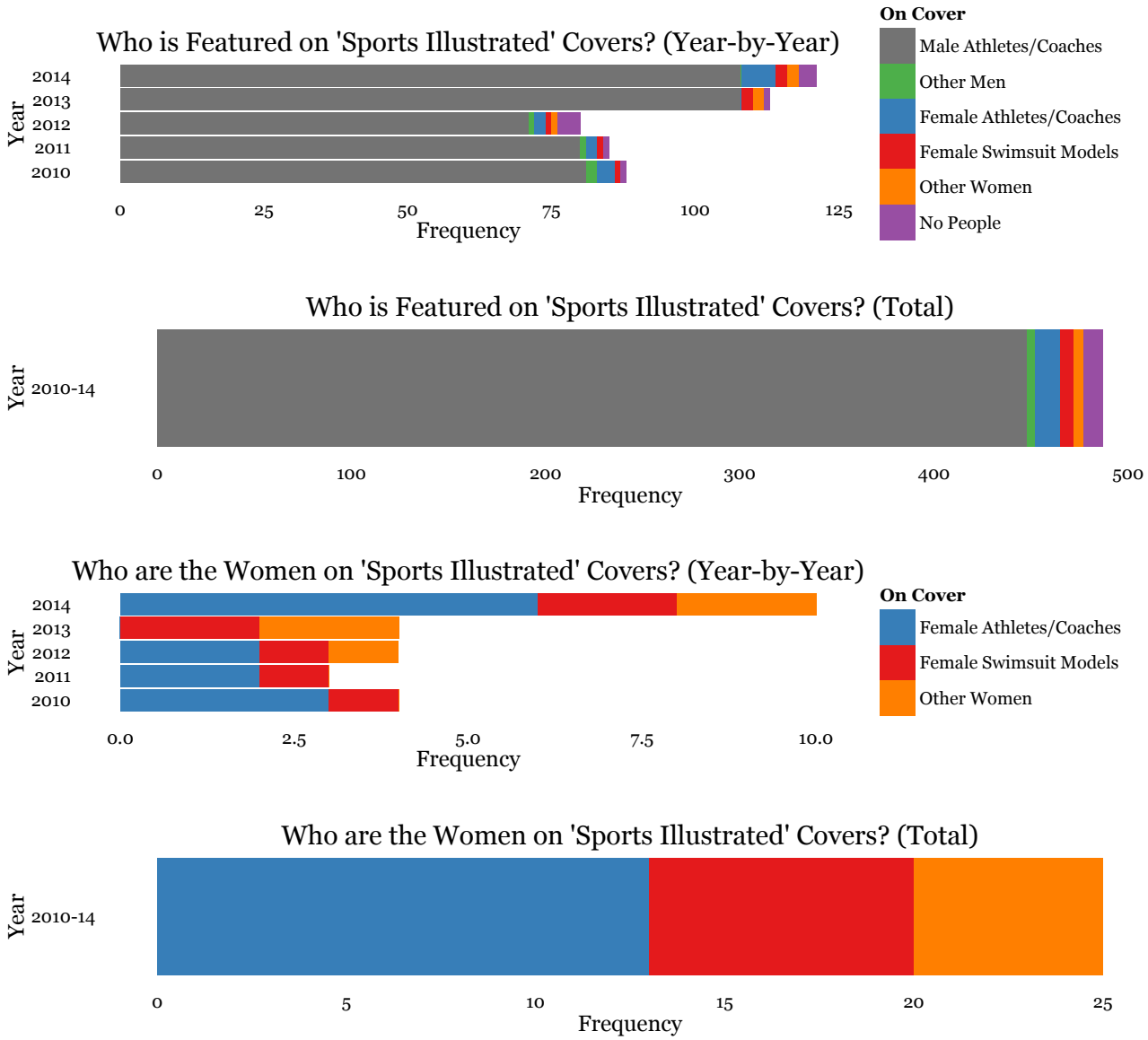
A cover featuring miscellaneous women

How often are women on the cover of *Sports Illustrated* (and not in swimsuits)?

How to visualize?

- Have cover counts and associated years
- Interested in showing covers distribution over each year as well as full 5-year distribution
- Want to show distribution for the subset of female-focused covers

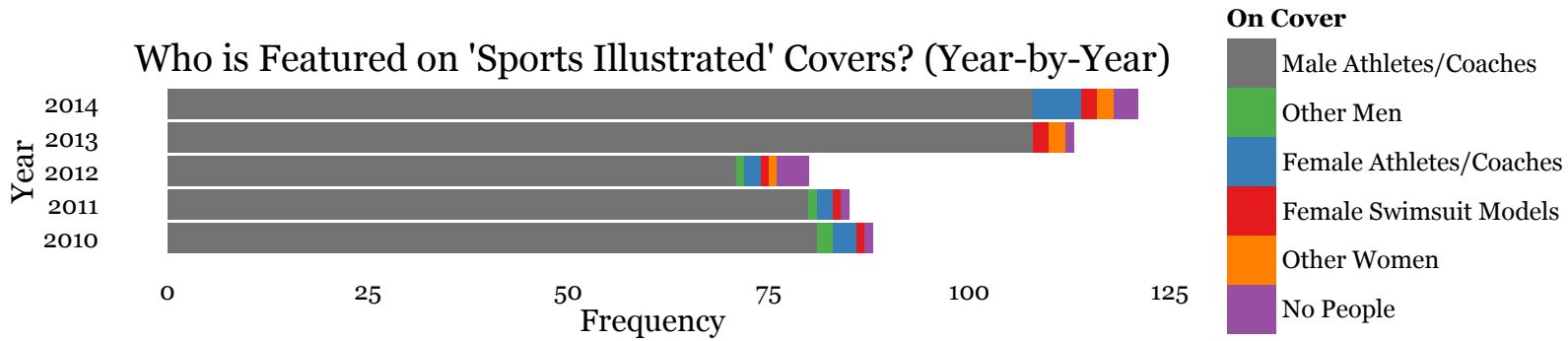
How often are women on the cover of *Sports Illustrated* (and not in swimsuits)?



- Only 2.67% of 2010-2014 covers feature female athletes or coaches
- 52% female covers feature athletes or coaches while 99.12% of male covers feature athletes or coaches
- Kate Upton has more covers (4) than all WOC athletes (3)

How often are women on the cover of *Sports Illustrated* (and not in swimsuits)?

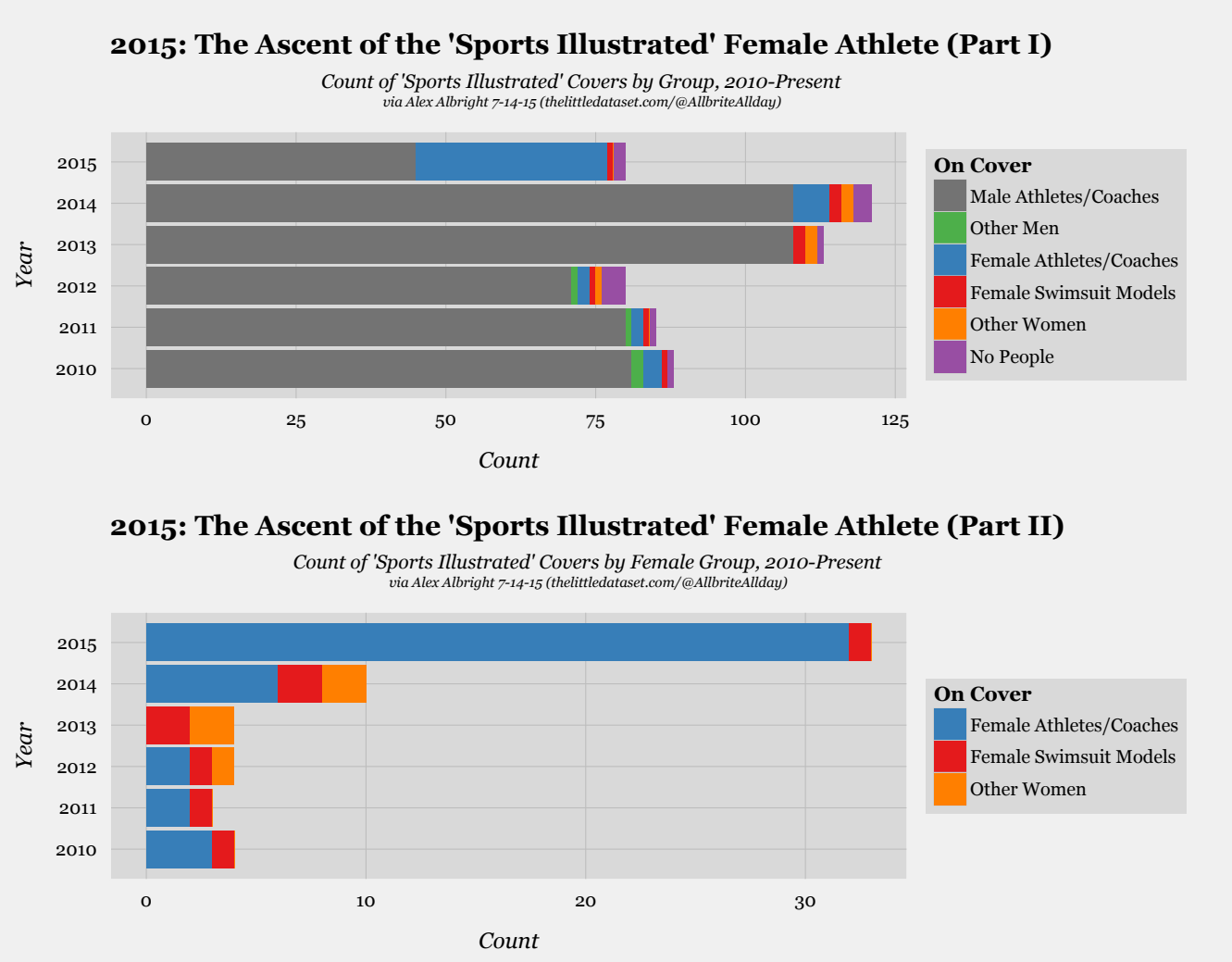
Generate four plots and then use a `multiplot()` function to combine them (could also use `grid.arrange()`)




```
a<-ggplot(data = si.data, aes(x = Year, y=Frequency, fill = On.Cover)) +
      geom_bar(stat="identity")+ coord_flip()+ scale_fill_manual(values = pal2, guide_legend(title="On Cover")) +
      labs(title = "Who is Featured on 'Sports Illustrated' Covers? (Year-by-Year)") +
      theme_tufte(ticks=FALSE, base_family="Georgia")
```

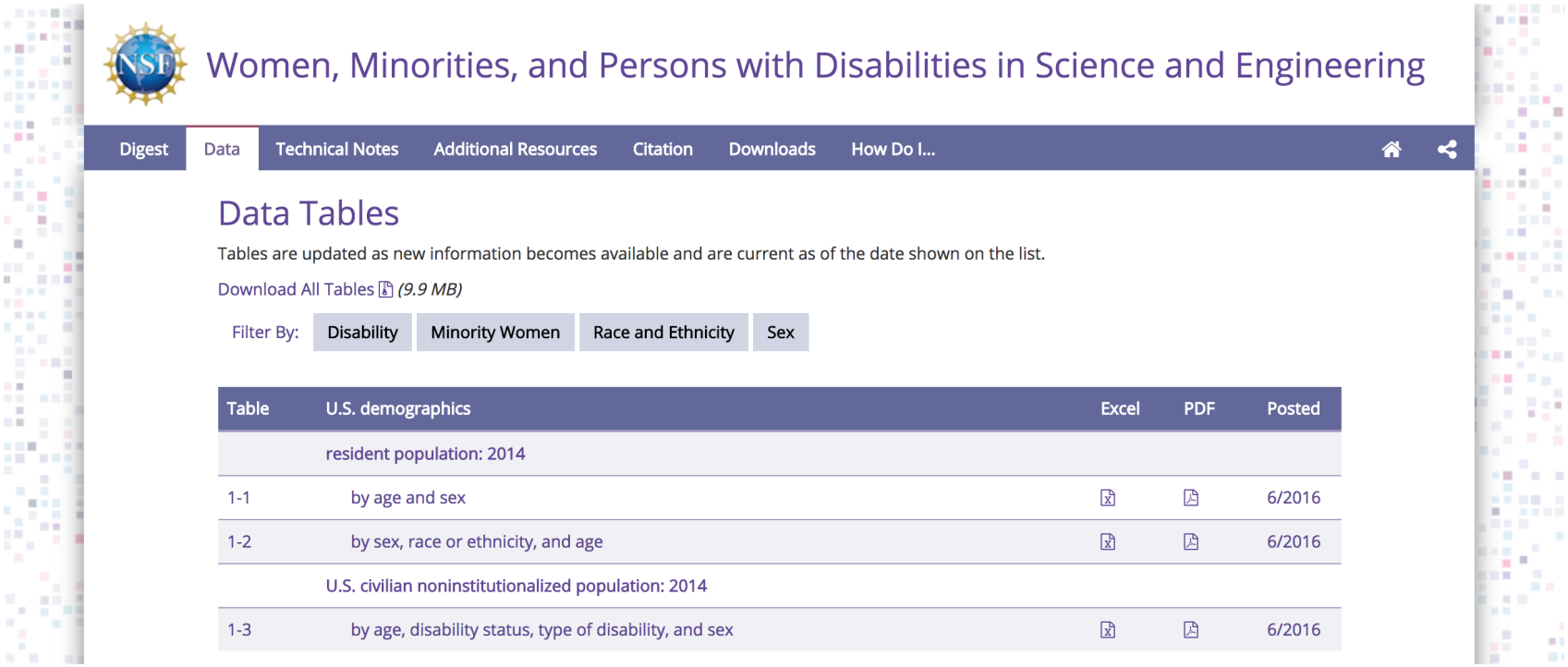
- Use simple `geom_bar()` with `fill = On.Cover` as an aesthetic mapping/`aes()`
- `On.Cover` is the collected variable that defines the type of cover

How often are women on the cover of *Sports Illustrated* (and not in swimsuits)?



- July 2015, *Sports Illustrated* released 25 different covers to celebrate the World Cup-Winning US Women's Soccer Team
- Big impact on previous graph!
- Use my own customized theme here instead of `theme_tufte()`

How do sciences compare in their %s of women and ethnic/racial groups?



Women, Minorities, and Persons with Disabilities in Science and Engineering

Digest | Data | Technical Notes | Additional Resources | Citation | Downloads | How Do I...

Data Tables

Tables are updated as new information becomes available and are current as of the date shown on the list.

Download All Tables *(9.9 MB)*

Filter By: Disability | Minority Women | Race and Ethnicity | Sex

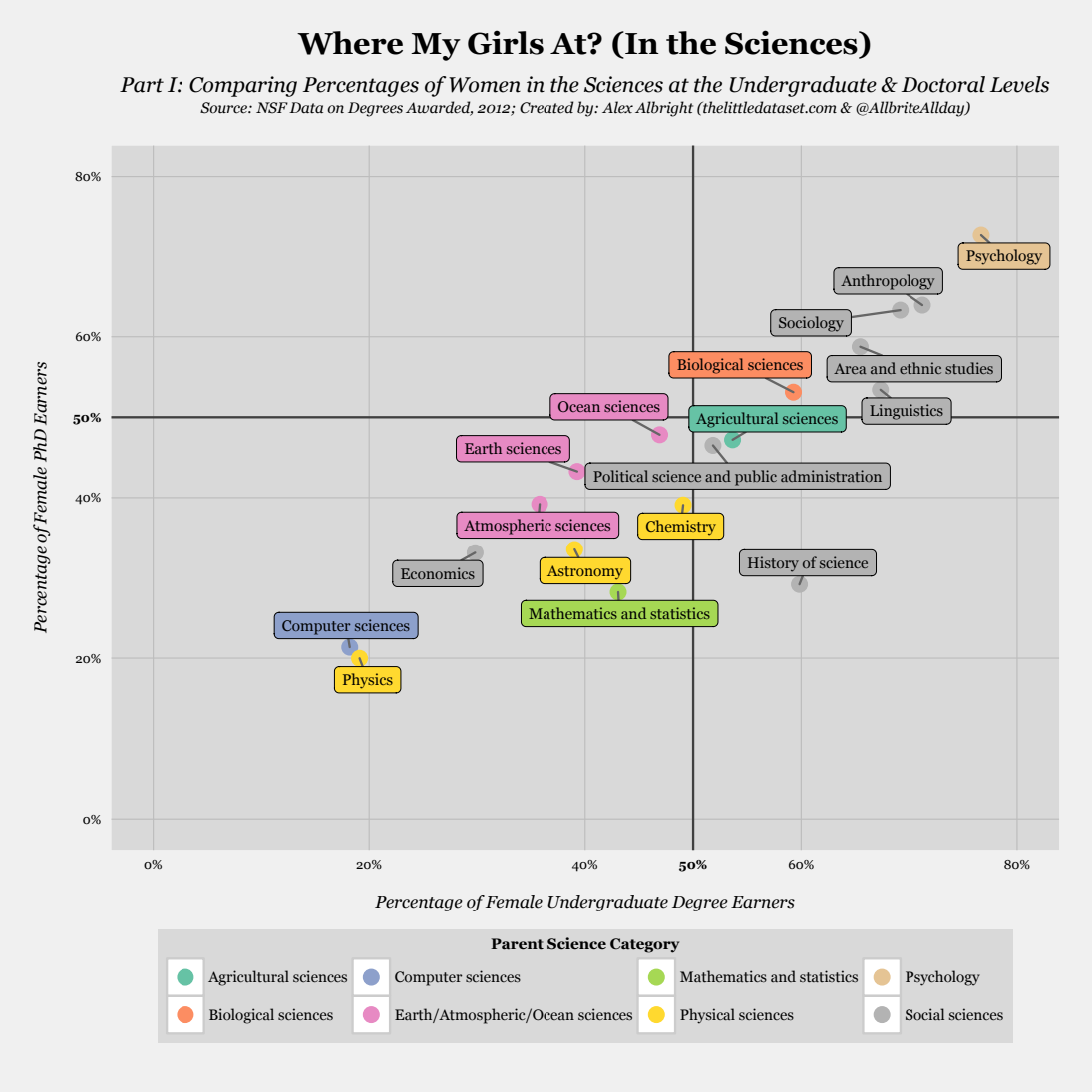
Table	U.S. demographics	Excel	PDF	Posted
	resident population: 2014			
1-1	by age and sex			6/2016
1-2	by sex, race or ethnicity, and age			6/2016
	U.S. civilian noninstitutionalized population: 2014			
1-3	by age, disability status, type of disability, and sex			6/2016

How do sciences compare in their %s of women and ethnic/racial groups?

- Became curious about this when applying for graduate school
- "The sciences" often treated as a homogeneous group – they are not in terms of representation
- Looked into the NSF Open Data Portal
- Found data for 2002-2012

First, I ask: how do sciences compare in female representation at undergraduate and doctoral levels?
(Use only 2012 data)

How do sciences compare in their %s of women and ethnic/racial groups?



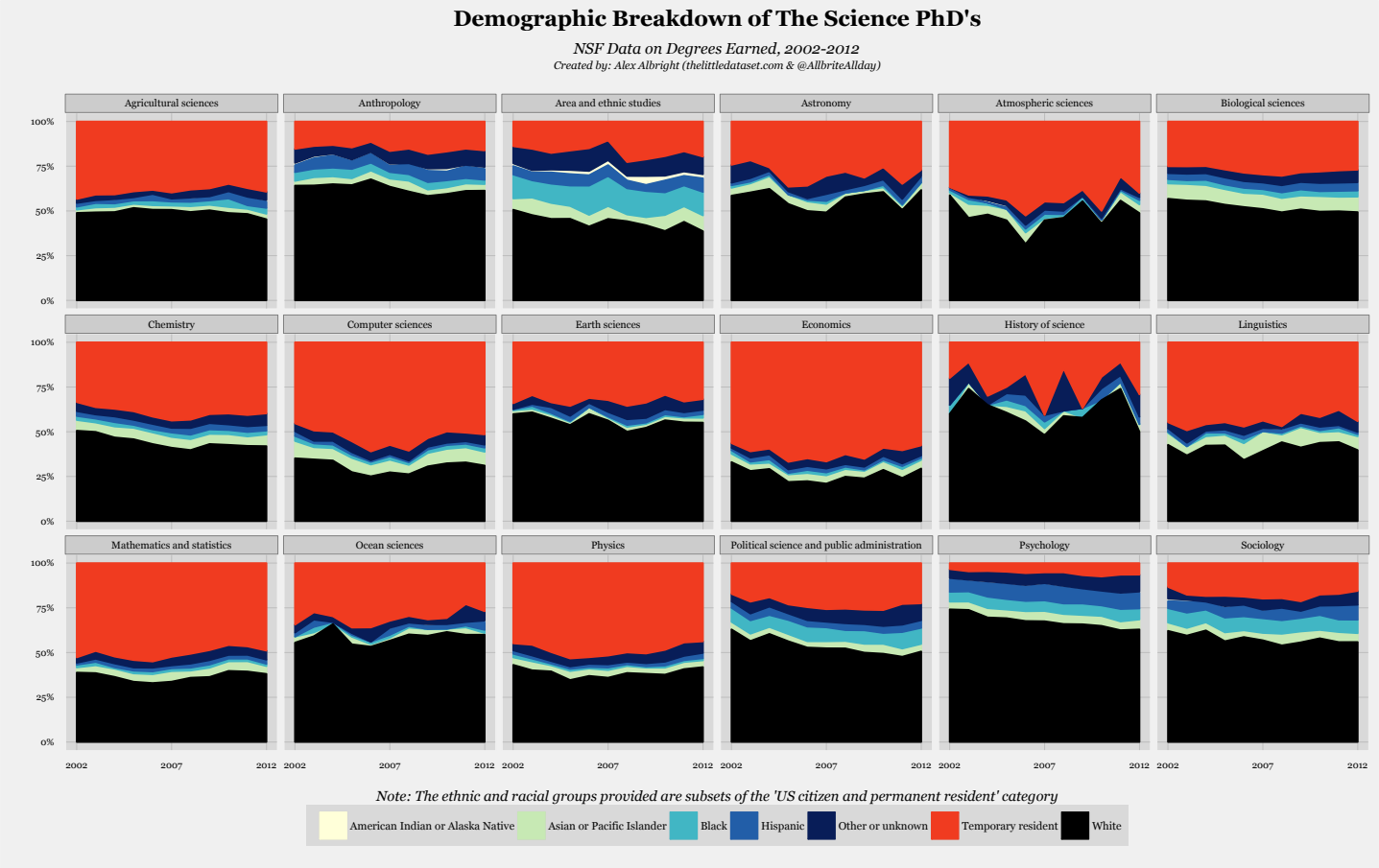
- Visually identify each discipline and the parent science group
- Use `ggplot()` with `geom_point()`
- Set color within `geom_point()` based on parent category of science
- Use `geom_label_repel()` to label the points and make sure that they don't overlap
- Note variation within parent categories

How do sciences compare in their %s of women and ethnic/racial groups?

- Now, ethnic/racial representation for the sciences 2002-2012
- Want to look over time
- Can use stacked area graphs, nightingale graphs, and line charts
- Use `facet_wrap(~Type, ncol=6)` to create a graphic with distinct plots for each field (`Type`=type of science)

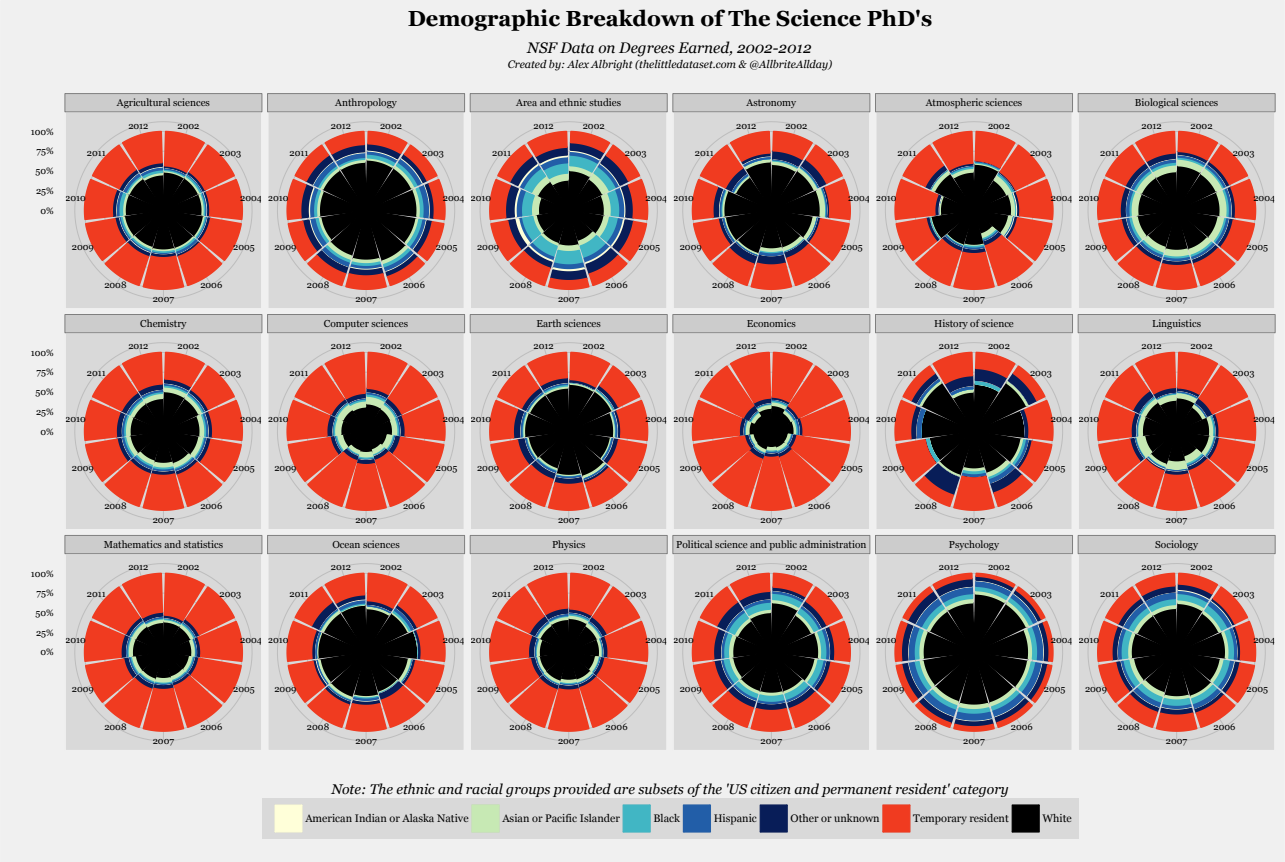
How do sciences compare in their %s of women and ethnic/racial groups?

Using `geom_area(aes(fill=Group), position='stack')`:



How do sciences compare in their %s of women and ethnic/racial groups?

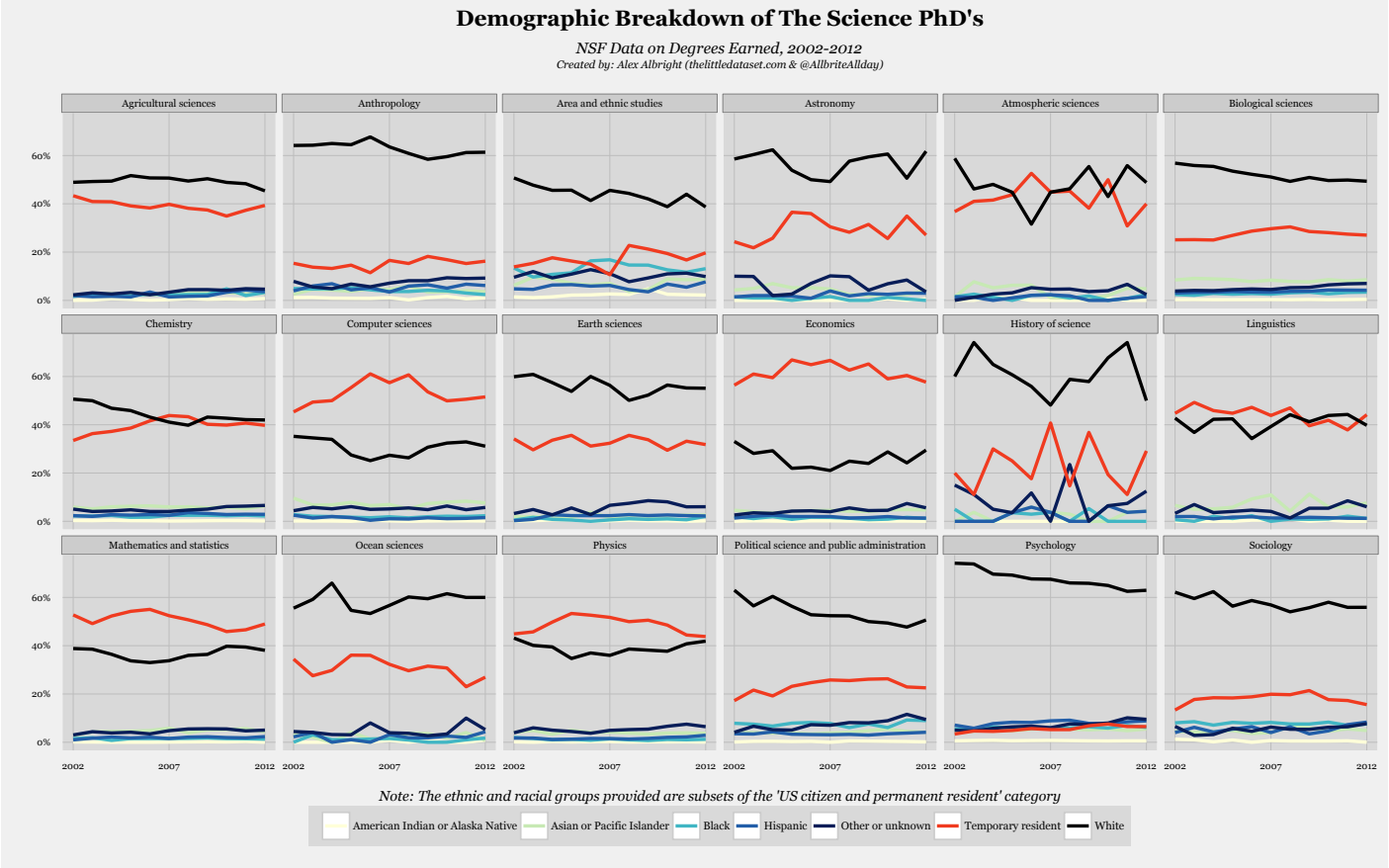
Using `geom_bar(stat="identity", aes(fill=Group), position='stack')+ coord_polar()`:



Visually clear for comparing red and black areas...
less so for the smaller categories

How do sciences compare in their %s of women and ethnic/racial groups?

Using `geom_line(size=1)`:



Clearest for discussing time trends in any specific field

How do sciences compare in their %s of women and ethnic/racial groups?

Part of the code for creating these three visuals:

```
p<-ggplot(data=sci_moreall, aes(x=year, y=perc, group=Group, fill=Group, color=Group)) +
  scale_fill_manual(values = pal1) +
  scale_color_manual(values = pal1) +
  my_theme()+
  facet_wrap(~Type, ncol=6)+
  labs(title="", x="Note: The ethnic and racial groups provided are subsets of the 'US citizen and permanent resident' category", y="")+
  scale_y_continuous(labels = percent_format())+
  ggtitle(expression(atop(bold("Demographic Breakdown of The Science PhD's"),
                          atop(italic("NSF Data on Degrees Earned, 2002-2012"),
                               atop(italic("Created by: Alex Albright (thelittledataset.com & @AllbriteAllday)")),""))))
#Stacked area graph
p+geom_area(aes(fill=Group), position='stack')+ scale_x_discrete(breaks=c("2002","2007","2012"))
#Nightingale graphs
p+geom_bar(stat="identity", aes(fill=Group), position='stack')+ coord_polar()+ facet_wrap(~Type, ncol=6)
#Line graphs
p+geom_line(size=1)+facet_wrap(~Type, ncol=6) + scale_x_discrete(breaks=c("2002","2007","2012"))
```

How has Pixar movie success changed over time?

How has Pixar movie success changed over time?

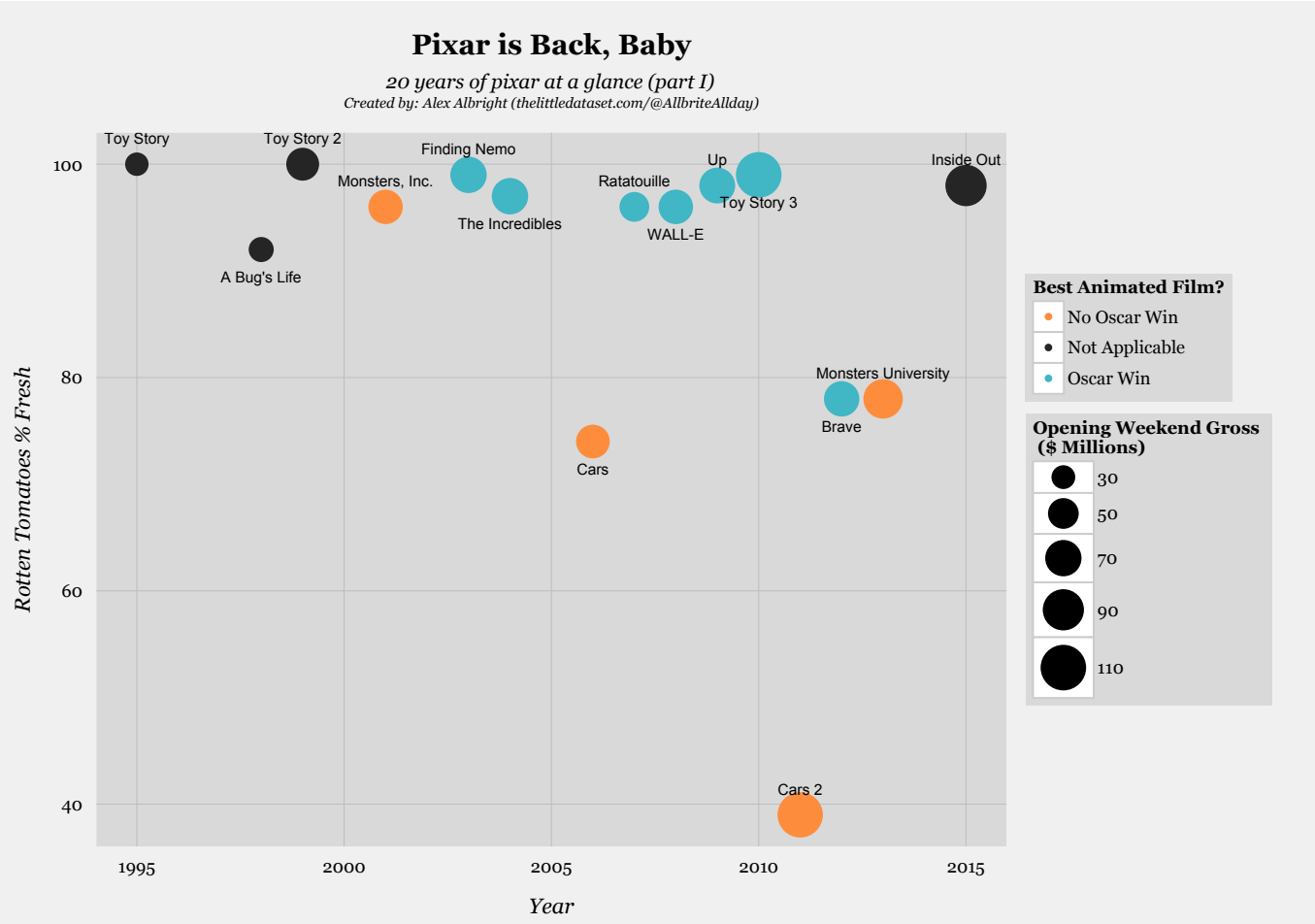
Dimensions of success:

- reviews/public perception
- awards/critical acclaim
- $$$

Focus on some of the most salient related metrics:
Oscar awards, Rotten Tomatoes score, and opening weekend gross

How can one illustrate all three dimensions simultaneously?

How has Pixar movie success changed over time?

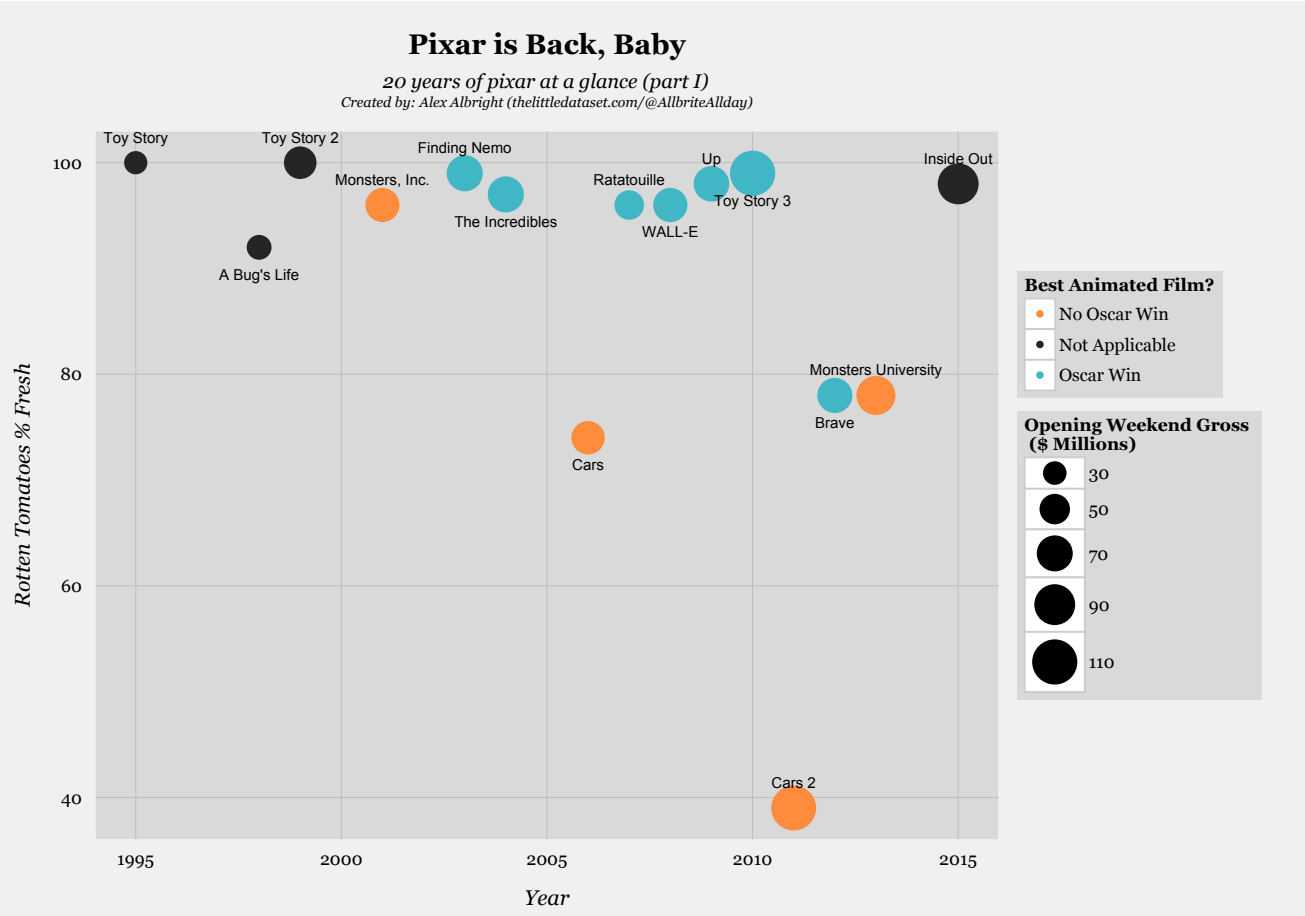


Bubble chart!

- scatterplot to show % fresh over time
- bubble size is revenue
- color is Oscar status

Inside Out was a return to BC2 (Before Cars 2) review levels!

How has Pixar movie success changed over time?



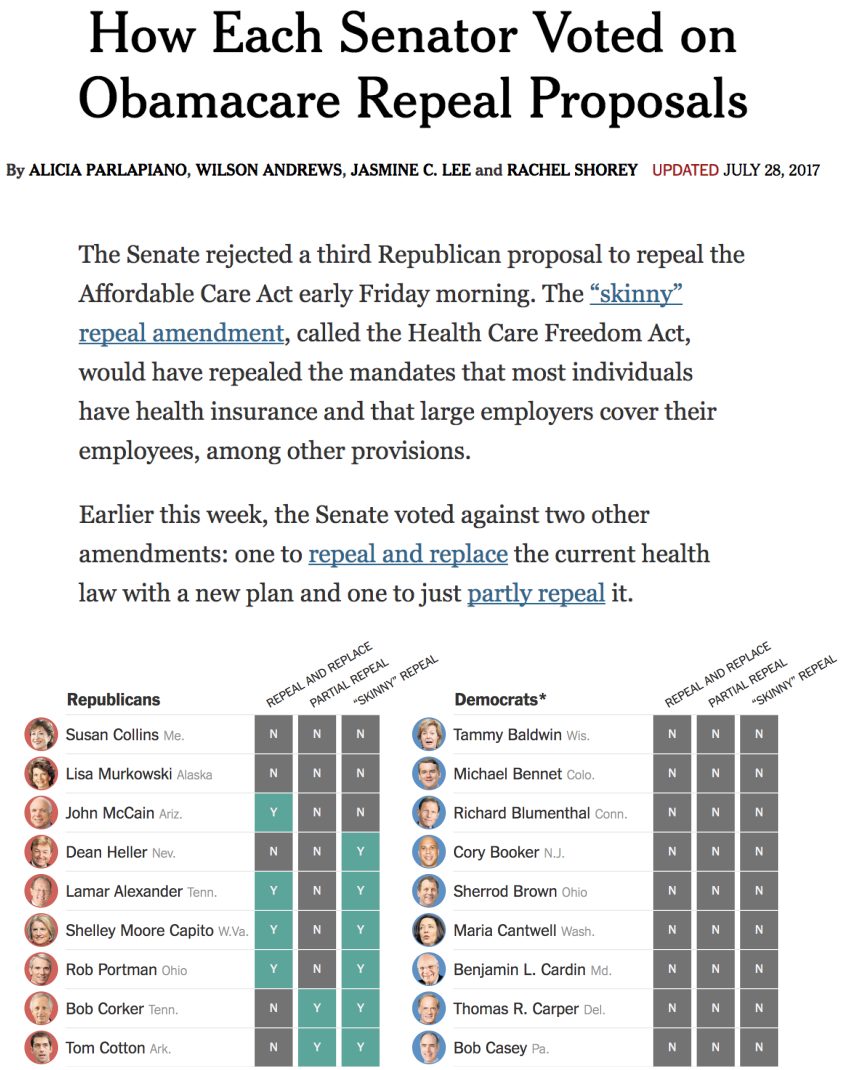
Details:

- trunctate y-axis to emphasize quality changes
- scale gross to area of the bubbles (not radii)

Relevant code snippets:

```
-geom_point(aes(size=weekend, color=factor(Oscars)))
-scale_size_area()
```

How did all US Senators vote on healthcare repeal?



How Each Senator Voted on Obamacare Repeal Proposals

By ALICIA PARLAPIANO, WILSON ANDREWS, JASMINE C. LEE and RACHEL SHOREY UPDATED JULY 28, 2017

The Senate rejected a third Republican proposal to repeal the Affordable Care Act early Friday morning. The "skinny" repeal amendment, called the Health Care Freedom Act, would have repealed the mandates that most individuals have health insurance and that large employers cover their employees, among other provisions.

Earlier this week, the Senate voted against two other amendments: one to repeal and replace the current health law with a new plan and one to just partly repeal it.

Republicans	Repeal and replace	Partial repeal	"Skinny" repeal
Susan Collins Me.	N	N	N
Lisa Murkowski Alaska	N	N	N
John McCain Ariz.	Y	N	N
Dean Heller Nev.	N	N	Y
Lamar Alexander Tenn.	Y	N	Y
Shelley Moore Capito W.Va.	Y	N	Y
Rob Portman Ohio	Y	N	Y
Bob Corker Tenn.	N	Y	Y
Tom Cotton Ark.	N	Y	Y

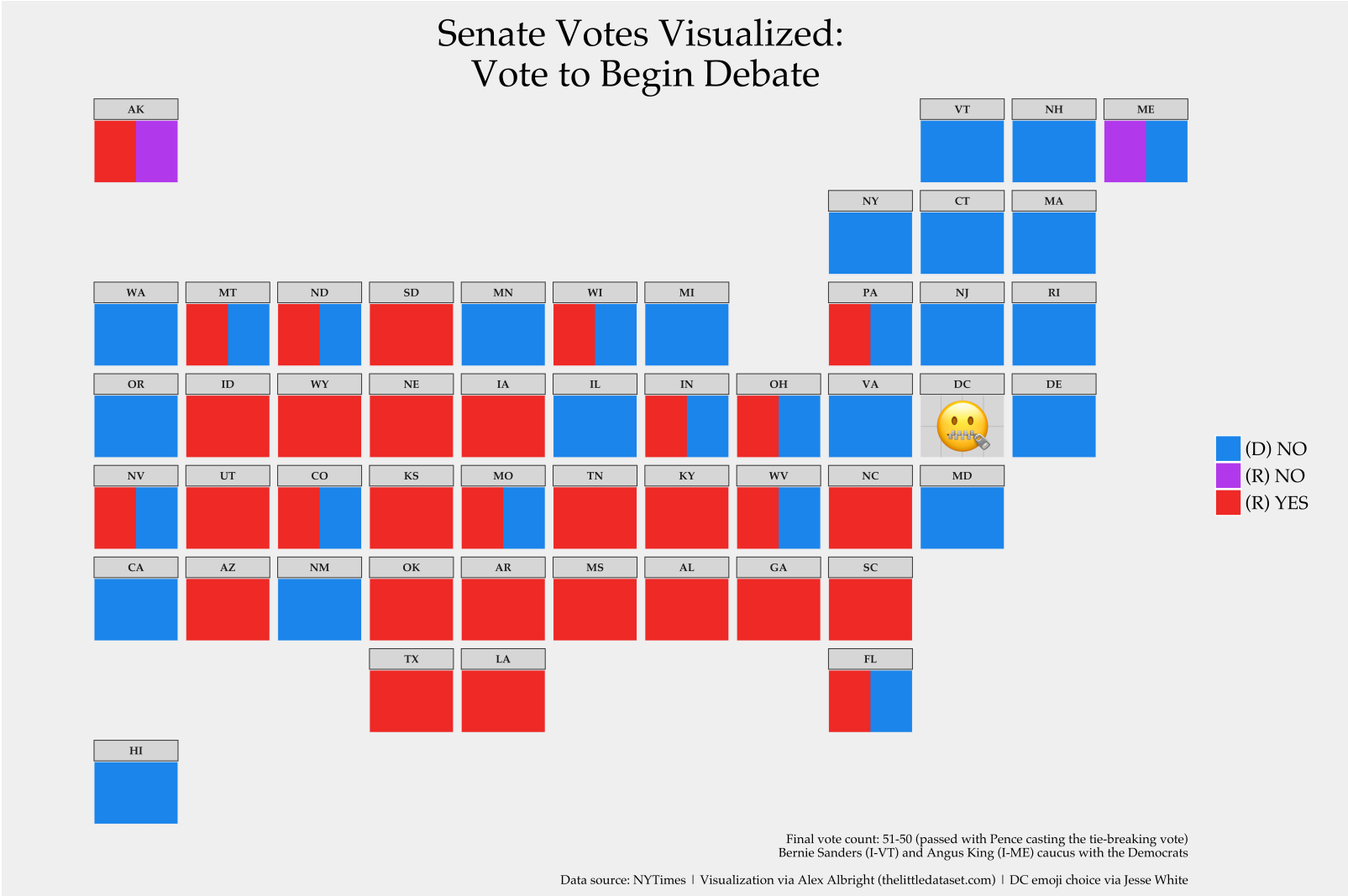
Democrats*	Repeal and replace	Partial repeal	"Skinny" repeal
Tammy Baldwin Wis.	N	N	N
Michael Bennet Colo.	N	N	N
Richard Blumenthal Conn.	N	N	N
Cory Booker N.J.	N	N	N
Sherrod Brown Ohio	N	N	N
Maria Cantwell Wash.	N	N	N
Benjamin L. Cardin Md.	N	N	N
Thomas R. Carper Del.	N	N	N
Bob Casey Pa.	N	N	N

There were three Obamacare repeal proposals this summer (after vote to begin debate)

- Each failed, but in a different way
- News outlets (such as the *NYTimes*) reported how each Senator voted for all the proposals
- Still was hard to see the full picture of votes

Used vote data to geographically illustrate Senator healthcare repeal votes

How did all US Senators vote on healthcare repeal?



Tuesday, July 25th, 2017

How did all US Senators vote on healthcare repeal?

Useful exercise for using `geofacet` package!

```
deb<-ggplot(healthcare_debate_vote, aes("", votes, fill = senator_group)) +
  geom_col(alpha = 1, width = 1) +
  my_theme()+
  coord_flip()+
  scale_fill_manual(values = c("dodgerblue2", "darkorchid2", "firebrick2"), breaks=c("Dem", "R_vote_against_debat
e", "R_vote_for_debate"), labels=c("(D) NO     ", "(R) NO     ", "(R) YES     ")) +
  facet_geo(~ state, grid = "us_state_grid2", label="code") +
  scale_y_continuous(expand = c(0, 0)) +
  ggtitle("Senate Votes Visualized:\n Vote to Begin Debate")+
  labs(caption = "Final vote count: 51-50 (passed with Pence casting the tie-breaking vote)\nBernie Sanders (I-V
T) and Angus King (I-ME) caucus with the Democrats\n\nData source: NYTimes | Visualization via Alex Albright (the
littledataset.com) | DC emoji choice via Jesse White") +
  theme(
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    strip.text.x = element_text(size = 7))+
  ggsave("deb.png", width = 12, height = 8, dpi = 800)
```

Made tile grid maps

- Code up using `geom_col()` so that each state consists of two columns (one for each senator) and columns are coded based on party and vote
- Use `facet_geo()` to arrange these 2 column plots into the spaces associated with the appropriate states

How did all US Senators vote on healthcare repeal?

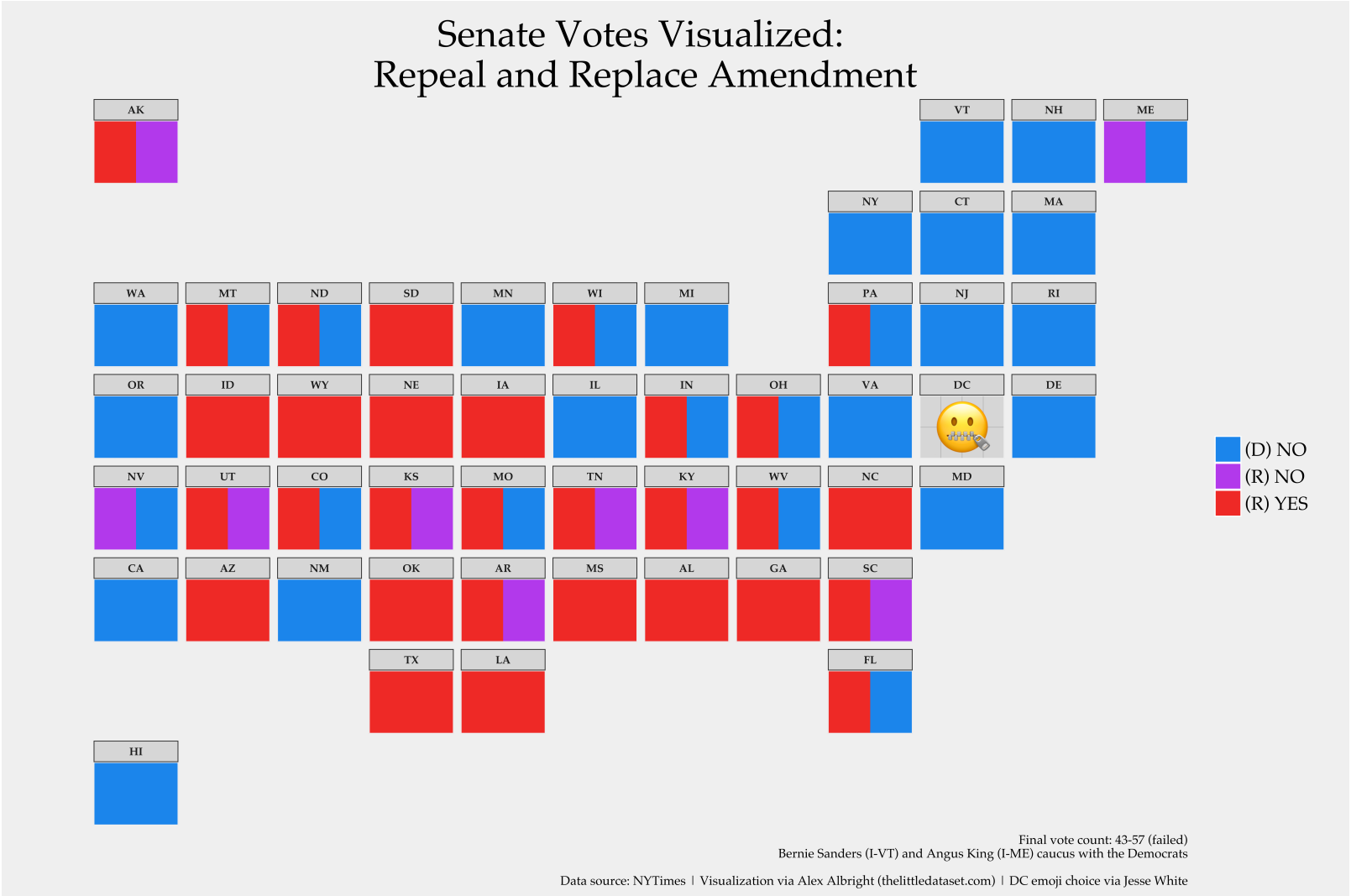
Also practice using the `magick` package!

```
# Now call back the plot
background <- image_read("deb.png")
# And bring in a zipper emoji
zipper_raw <- image_read("zipper.png")
zipper <- zipper_raw %>%
  image_scale("400")
new <- image_composite(background, zipper, offset = "+6650+2850")
image_write(new, "deb_final.png", flatten = F)
```

Insert emoji image onto DC space (since they have no Senators)

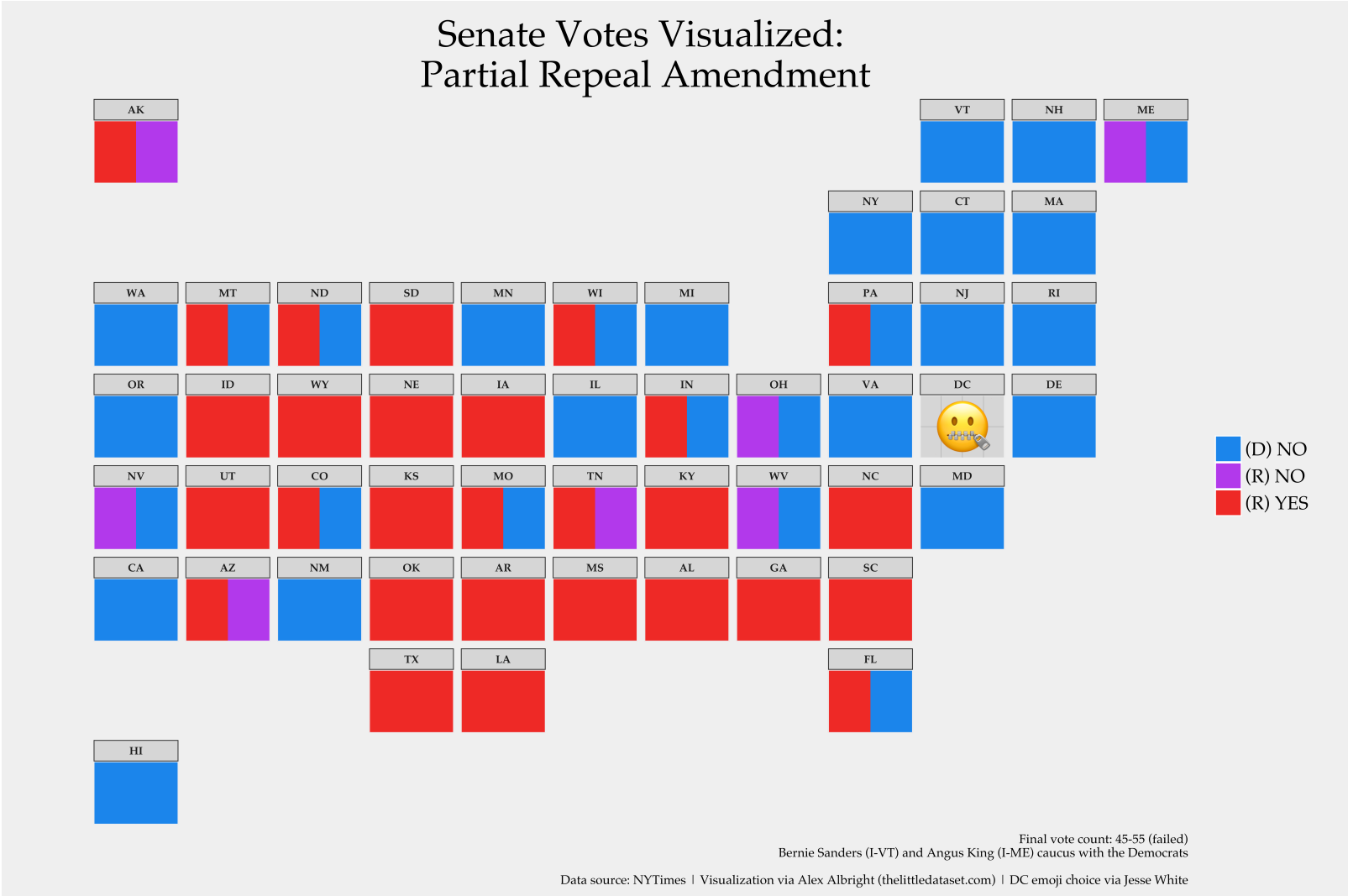
- Use `image_read()` to call existing `png` images
- Make new image with `image_composite()` and place emoji on map using `offset` coordinates

How did all US Senators vote on healthcare repeal?



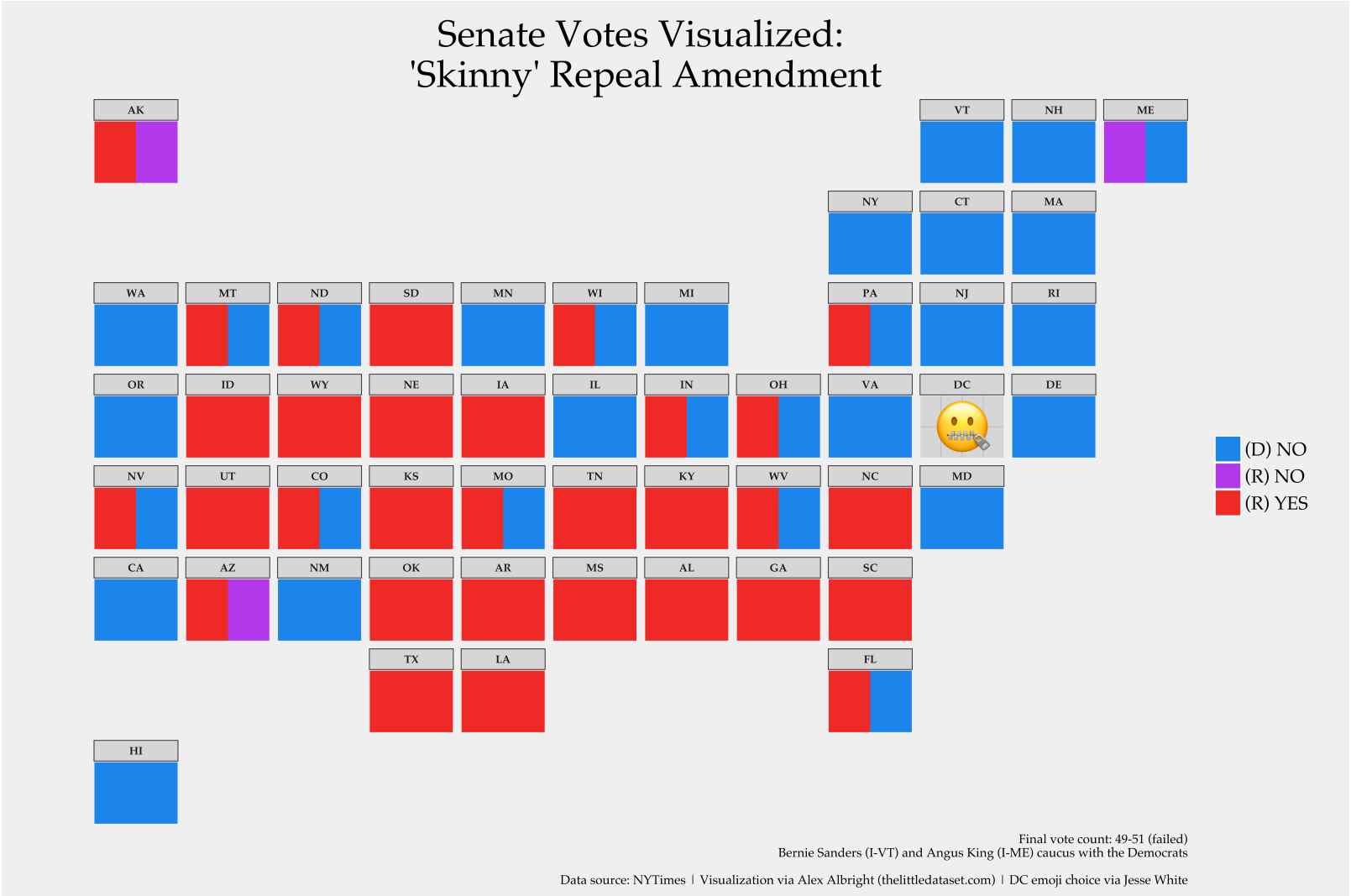
Still Tuesday, July 25th, 2017

How did all US Senators vote on healthcare repeal?



Wednesday, July 26th, 2017

How did all US Senators vote on healthcare repeal?



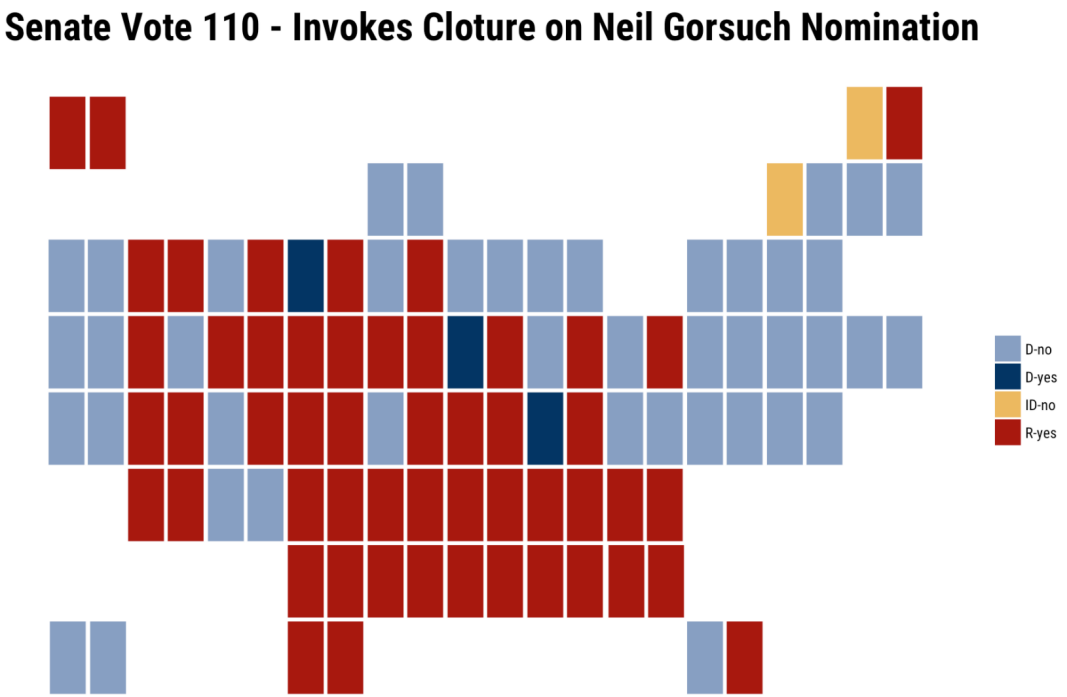
Friday, July 28th, 2017

How did all US Senators vote on healthcare repeal?

Another option: use `voteogram` package

- Easy to retrieve voting data (`sen <- roll_call("senate", 115, 1, 110)`) & make vote maps:

```
senate_carto(sen) +
  labs(title="Senate Vote 110 - Invokes Cloture on Neil Gorsuch Nomination") +
  theme_ipsum_rc(plot_title_size = 24) +
  theme_voteogram()
```



Which characters are closest on the TV show *Friends*?

January 2015: Friends on Netflix!

"In re-watching the show, I remembered that certain pairs of characters were closer (friendship-wise) than others–and I began to wonder whether one could illustrate the closeness (or lack thereof) between certain characters using quantitative data from the 236 episodes of the show..."

This was a project that started before I was familiar with R but I revisited a few times with more data visualization knowledge over time...

Which characters are closest on the TV show *Friends*?

What kind of data is needed to answer this question?

- Lots of options
- Ideally, shared screen time
  - Impossible to collect without making it your life for years
- First pass: settled on collecting data on which characters had shared plotlines
- Collected subjectively by reviewing episodes on Netflix and through Wikipedia summaries
- Coded all the episodes by plot dynamics

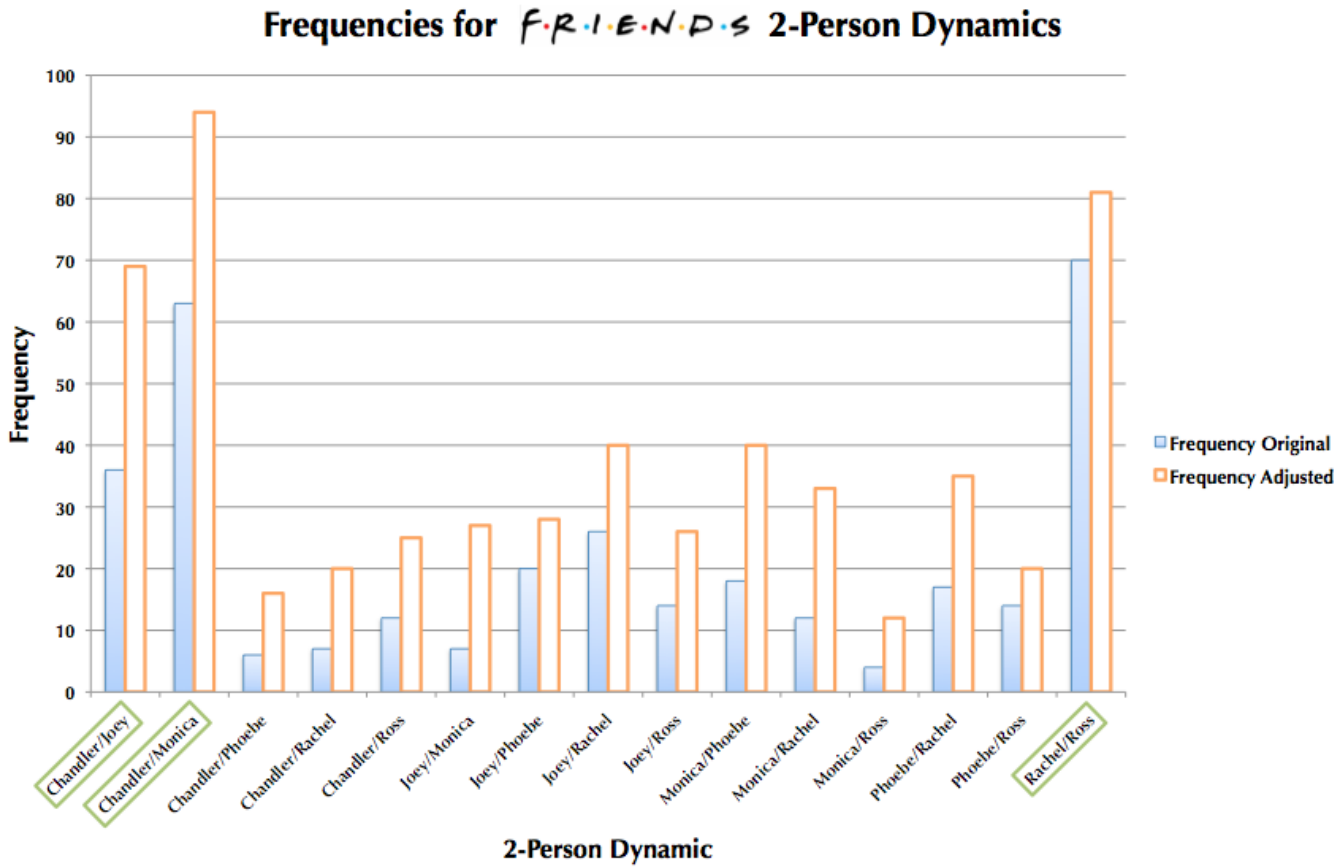
Which characters are closest on the TV show *Friends*?

Let me formalize this very silly endeavor...

Let us consider the question of character groupings in basic mathematical terms. There are six friends, each an element of the overarching "group," defined as the set F={1,2,3,4,5,6} where 1, 2, 3, 4, 5, 6 represent Chandler, Joey, Monica, Phoebe, Rachel, and Ross respectively (the listing method is alphabetical). Each episode features character groupings in the form of shared plots, which in turn correspond to subsets of F. For example, "The One With George Stephanopoulos" would be represented by the set TOWGS={{1,2,6},{3,4,5}} since the plotline with the guys at the hockey game, {1,2,6} (⊆F), is an element of TOWGS as is the plotline with the girls getting pizza/watching George drop his towel, {3,4,5} (⊆F). There are 64 possible subsets of set F, including both the empty set and F itself (64=2^6).

Which characters are closest on the TV show *Friends*?

Armed with plotline data, January-2015-Alex only knew how to make graphs in Excel... like this:



- Frequency Adjusted = all shared plotlines (not just those exclusive to the two individuals of interest)

Which characters are closest on the TV show *Friends*?

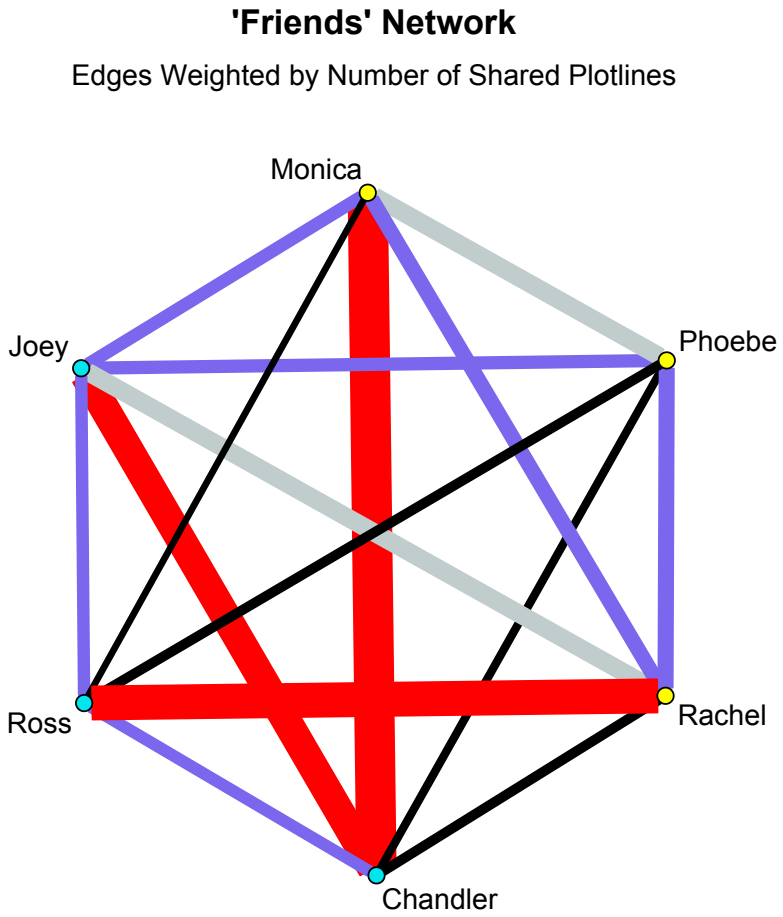
Findings:

- Most frequent two-person dynamics (marked in green) are: Chandler/Monica, Chandler/Joey, and Rachel/Ross
- Interesting: Rachel and Ross share more exclusively 2-person plots than do Monica and Chandler (70 to 63) despite the fact that latter duo shares more plots overall than the former (94 to 81)
  - Hypothesis: Could be due to the fact that Rachel and Ross, an on-again-off-again couple, had a complicated romantic history that could have inhibited them from regularly interacting in larger group plots while Monica and Chandler were friends consistently until dating and then marriage

Which characters are closest on the TV show *Friends*?

After learning some R, decided to make a network visualization for this same dataset

- Use `network` package in R
- Code a basic adjacency matrix for the 6 characters and then a weights matrix based on shared plotlines count
- Call on these matrices `plot.networks()` command to generate the network visualization



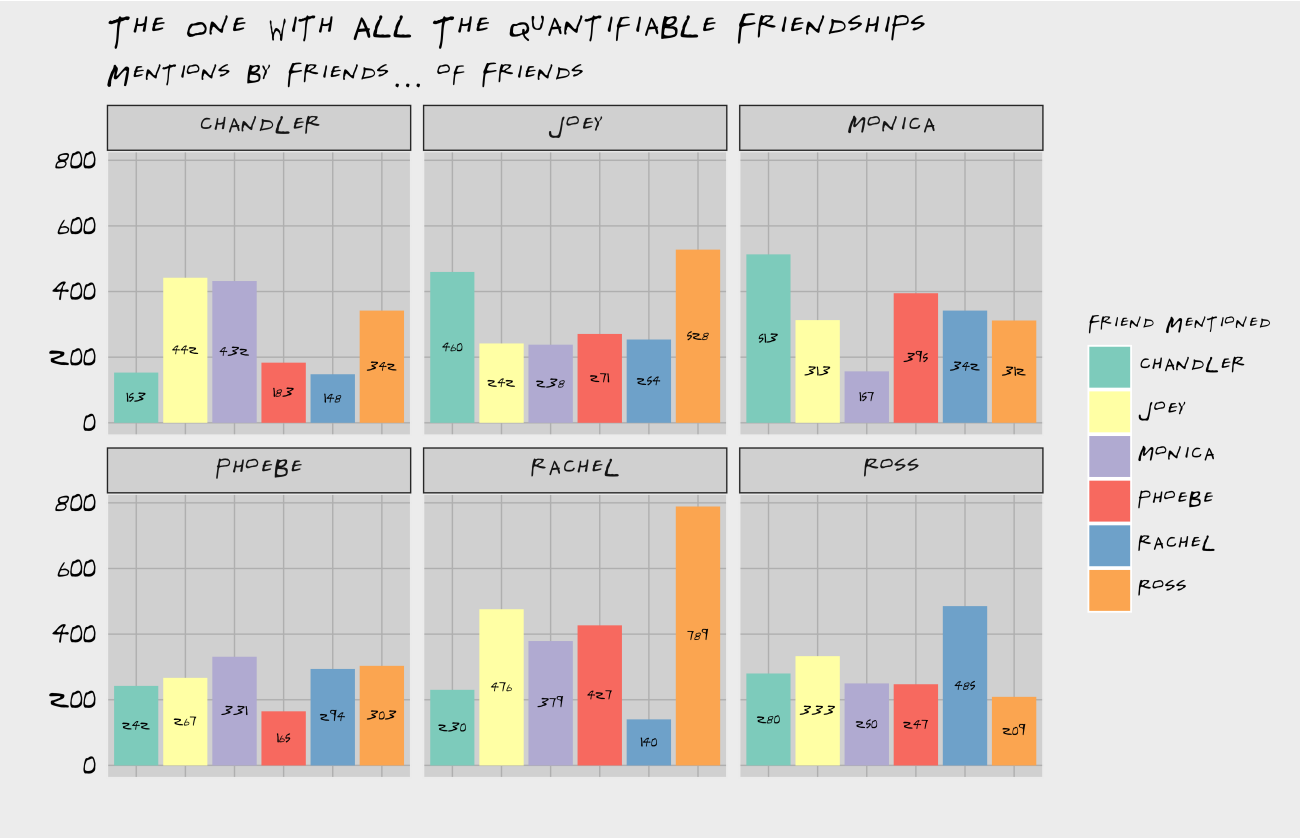
Edge Coloring: Black [0,20]; Purple (20,35]; Gray (35,50]; Red (50,94]
Vertex Coloring: Yellow for Female Characters; Blue for Male Characters

Which characters are closest on the TV show *Friends*?

This past summer I used another dataset for this same question!

- Giora Simchoni wrote a blog post on scraping *Friends* scripts using `rvest` and then formatting the data with `purrr` and `stringr`
- Q: What new metric of closeness could be taken from scripts?
  A: How often characters say each others' names
- Sidenote: I include nicknames ("Mon", "Rach", "Pheebs", and "Joe")
  - "Pheebs" is undoubtably the stickiest of the group
  - Characters say "Pheebs" 370 times, which has a comfortable cushion over the second-place nickname "Mon" (used 73 times)
  - Characters differ in their usage of each others' nicknames: e.g., while Joey calls Phoebe "Pheebs" 38.3% of the time, Monica calls her by this nickname only 4.6% of the time

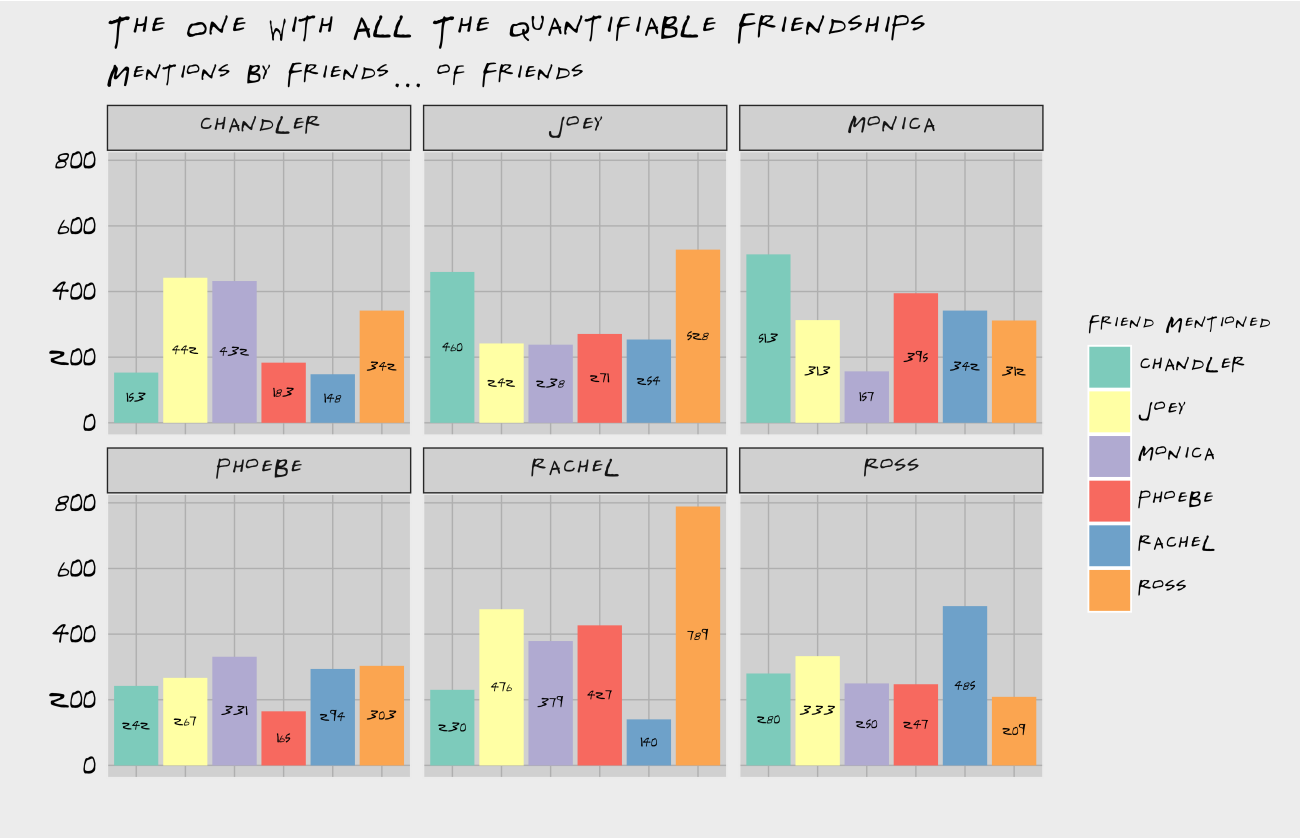
Which characters are closest on the TV show *Friends*?



- Rachel says Ross's name the most! (789 times! OK, we get it, Rachel, you're in love.)
- Joey-Chandler, Monica-Chandler, Ross-Rachel still at the forefront (Note: Ross sticks out notably for Joey)

- Could be fun to calculate how reciprocated mentions are for each character coupling...

Which characters are closest on the TV show *Friends*?

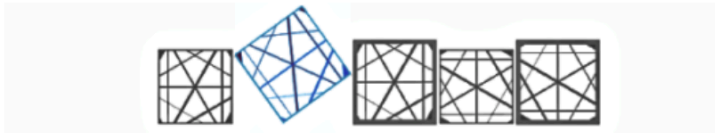


- With `ggplot()`, use `geom_bar()` and `facet_wrap` to create 6 plots – one for each character
- Use `geom_text()` to label the bars with the values

- Most important: use `theme_bw(base_size=9, base_family="Friends")` to use the *Friends* font!

Want access to code and datasets?

Check out `thelittledataset.com`:



THE LITTLE DATASET THAT COULD

theories and observations from a young economist

ARCHIVE ABOUT RESEARCH CV GITHUB TWITTER RPUBS CONTACT COLLABS & PRESS

Repos for all projects on `GitHub` and R notebooks published on `RPubs` page

Workflow

Ideation

- Write down random questions that pop into your head, think about how they could be answered with data
  - Always test ideas on your friends (do they want to know the answer?)
- Play with fun datasets – *Data Is Plural* newsletter is a great source of interesting datasets
- Read the news, think about stories that could be succinctly visually communicated

R practices

- Use notebooks with RStudio
- Can narrate what you are doing and allow your future self to recall why you did what
- Encourages good communication between you and others as well as between you and your future selves

Notebooks

Snippets of notebook used for Senate vote maps:

Senate Votes Visualized

Code

Alex Albright

8-1-17

Last week there were four widely-covered Senate votes about healthcare. The results of the votes were often broken down by party (R)/(D) but I wanted a geographic visual summary of the results. I use this notebook to visualize the recent results in the Senate with the machinery of `facet_geo` in the `geofacet` package.

First, let's generate/clean/format our data.

I load required packages.

Hide

```
library(geofacet);library(dplyr);library(ggplot2);library(ggrepel); library(extrafont); li
brary(ggthemes);library(reshape);library(grid);
library(scales);library(RColorBrewer);library(gridExtra);
library(magrittr);library(magick)
```

Now, I then constructed a `csv` file that contained structured information on the vote to start debate and the three following proposal votes. The NYTimes consistently provided information on how all Senators voted, so I used the following two articles to construct my data set: article on the vote to start debate & article on the three following votes.

I then read in the `csv` file I created from the NYTimes data on the senate votes.

Hide

```
healthcare <- read.csv('hcare_votes.csv')
```

For the purposes of `ggplot2`, we have to reshape everything to be long format. Let's do that with a column for the categories of senators. Then we will subset for each of the 4 votes of interest (starting debate, repeal and replace, partial repeal, and 'skinny' repeal) to create plots for each.

Hide

1. Vote to Start Debate

This passed 51-50 on Tuesday, July 25, 2017. (Pence cast the tie-brekaing vote)

We subset to the columns relevant to these results.

Hide

```
healthcare_debate_vote<-healthcare[which(healthcare$senator_group=="R_vote_for_debate" | h
ealthcare$senator_group=="R_vote_against_debate"| healthcare$senator_group == "Dem"),]
```

Plot and save as `png`:

Hide

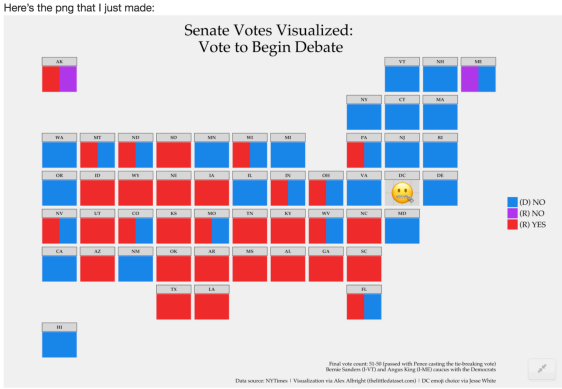
```
deb<-ggplot(healthcare_debate_vote, aes("", votes, fill = senator_group)) +
  geom_col(alpha = 1, width = 1) +
  my_theme()+
  coord_flip()+
  scale_fill_manual(values = c("dodgerblue2", "darkorchid2", "firebrick2"),
breaks=c("Dem", "R_vote_against_debate", "R_vote_for_debate"), labels=c("(D) NO     ", "
(R) NO     ", "(R) YES     ")) +
  facet_geo(~ state, grid = "us_state_grid2", label="code") +
  scale_y_continuous(expand = c(0, 0)) +
  ggtitle("Senate Votes Visualized:\n Vote to Begin Debate")+
  labs(caption = "Final vote count: 51-50 (passed with Pence casting the tie-breaking vot
e)\nBernie Sanders (I-VT) and Angus King (I-ME) caucus with the Democrats\n\nData source:
 NYTimes | Visualization via Alex Albright (thelittledataset.com) | DC emoji choice via Je
sse White") +
  theme(
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank(),
    strip.text.x = element_text(size = 7))+
  ggsave("deb.png", width = 12, height = 8, dpi = 800)
```

Now, I want to add a zipper emoji to DC's plot space, as DC has no senators but appears in the graph.

Hide

```
# Now call back the plot
background <- image_read("deb.png")
# And bring in a zipper emoji
zipper_raw <- image_read("zipper.png")
zipper <- zipper_raw %>%
  image_scale("400")
new <- image_composite(background, zipper, offset = "+6650+2850")
image_write(new, "deb_final.png", flatten = F)
```

Here's the png that I just made:



- Code is clearly separate from narration/comments
- Can make sections; chronology of thought process and workflow is clear
- Visualizations rendered alongside work process

Notebooks

Another plus of notebooks:

They can double as Valentine's Day cards.

Workflow Continued

Writing your code

- Use *Stack Overflow* as a resource (can learn a ton by reading question responses)
- Find code examples of approaches by other R users

Sharing

- Make your work reproducible and accessible *(plug for blogging!)*
- Encourages extensions and allows people to check your work
- Contributing resources means contributing to a culture of learning (important for encouraging a broader group of individuals to engage with R)
- It's satisfying!

The End

In short, channeling curiosity into data visualization projects is an exciting and effective way to get comfortable with R!

Any questions for me?