Welcome to DATA 151

I'm so glad you're here!

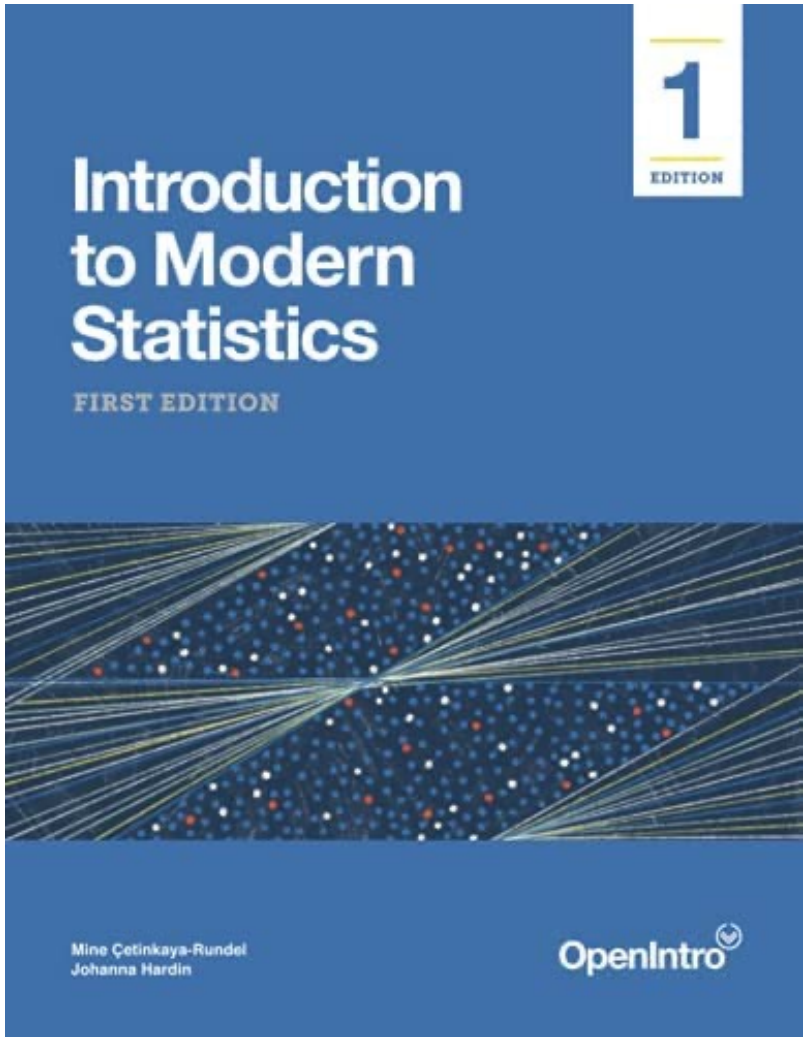
DATA 151: CLASS 8A
INTRODUCTION TO DATA SCIENCE (WITH R)

CATEGORICAL DATA ANALYSIS: TABLES AND BARS (PART 2)

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

ANNOUNCEMENTS

RELEVANT READING



Introduction to Data Science:

- Tuesday:
  - Introduction to Modern Statistics
  - Ch 4: Exploring Categorical Data
- Thursday:
  - Introduction to Modern Statistics
  - Ch 5: Exploring Numeric Data

HOMEWORK REMINDER

Due this week: (EXTENSION DUE 10/18)

- *HW #6: DC Introduction to Data Visualization in ggplot2*
  - ***No submission on WISE necessary, do on DataCamp***
- *Project Milestone #3: EDA Step 1*
  - Ask questions and form hypotheses

HOMEWORK REMINDER

Due this week: (DUE 10/20)

- *HW #7: DC Exploratory Data Analysis with Categorical Data*
  - *Just one chapter*
  - ***No submission on WISE necessary, do on DataCamp***
- *Project Milestone #4: EDA Step 2*
  - Create Tables and Bar Graphs

EXTRA CREDIT OPPORTUNITY

Data & Computing Tea

On the Research Experience for Undergraduates,

Thursday the 20th, 11:30 AM, Ford 201

Meelad Doroodchi, a major in the department who completed a REU (Research Experience for Undergraduates) this summer. Meelad will share some thoughts on the program and his experience, and then we will have time for open discussion.

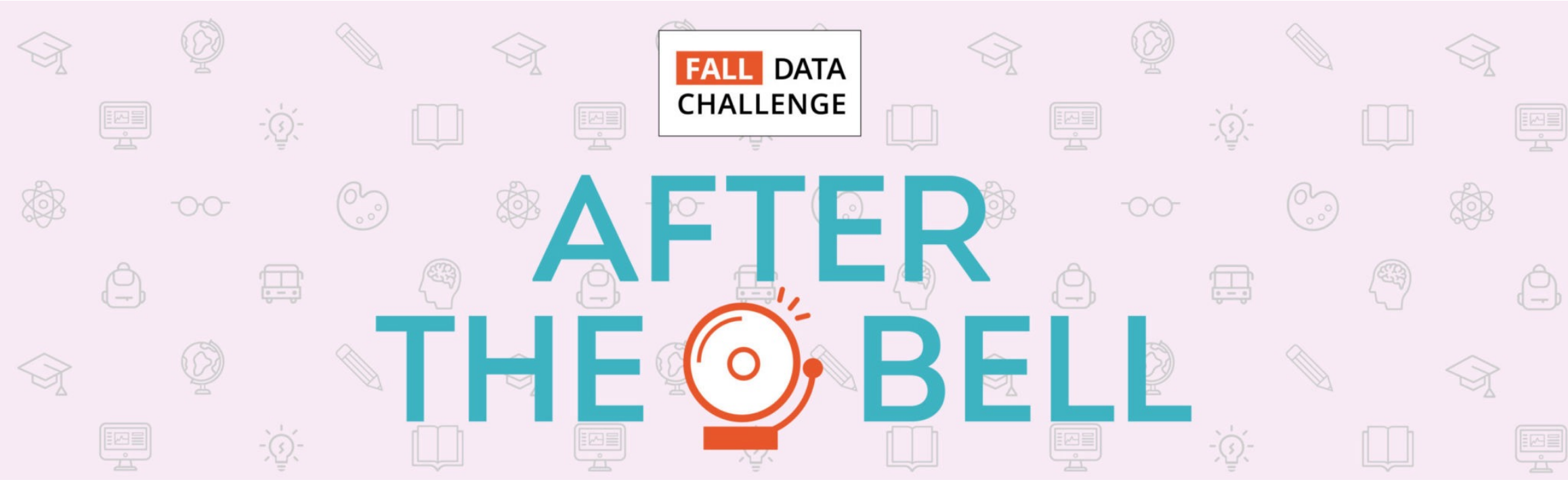
Pizza will be provided.

EXTRA CREDIT OPPORTUNITY

If you go to the presentation and do a 1-page write up about your take-aways and how this work relates to how we are learning data science in this class, I will give you **4 extra credit points** toward your

Midterm #1 grade!

ANOTHER EXTRA CREDIT OPPORTUNITY



Students will work in teams as they analyze data on the K-12 educational experience “After the Bell.” This year’s theme will require student teams to dive into data on the impacts of school choice and family engagement in school activities and homework. Teams will provide recommendations on factors that best optimize family involvement and support of K-12 students’ academic excellence.

Submissions will be accepted from October 17 to November 6, at 11:59pm EST.

BOARD OF TRUSTEES

OnThursday Oct 20

The Willamette Board of Trustees will be meeting in FORD 102 (our classroom space)

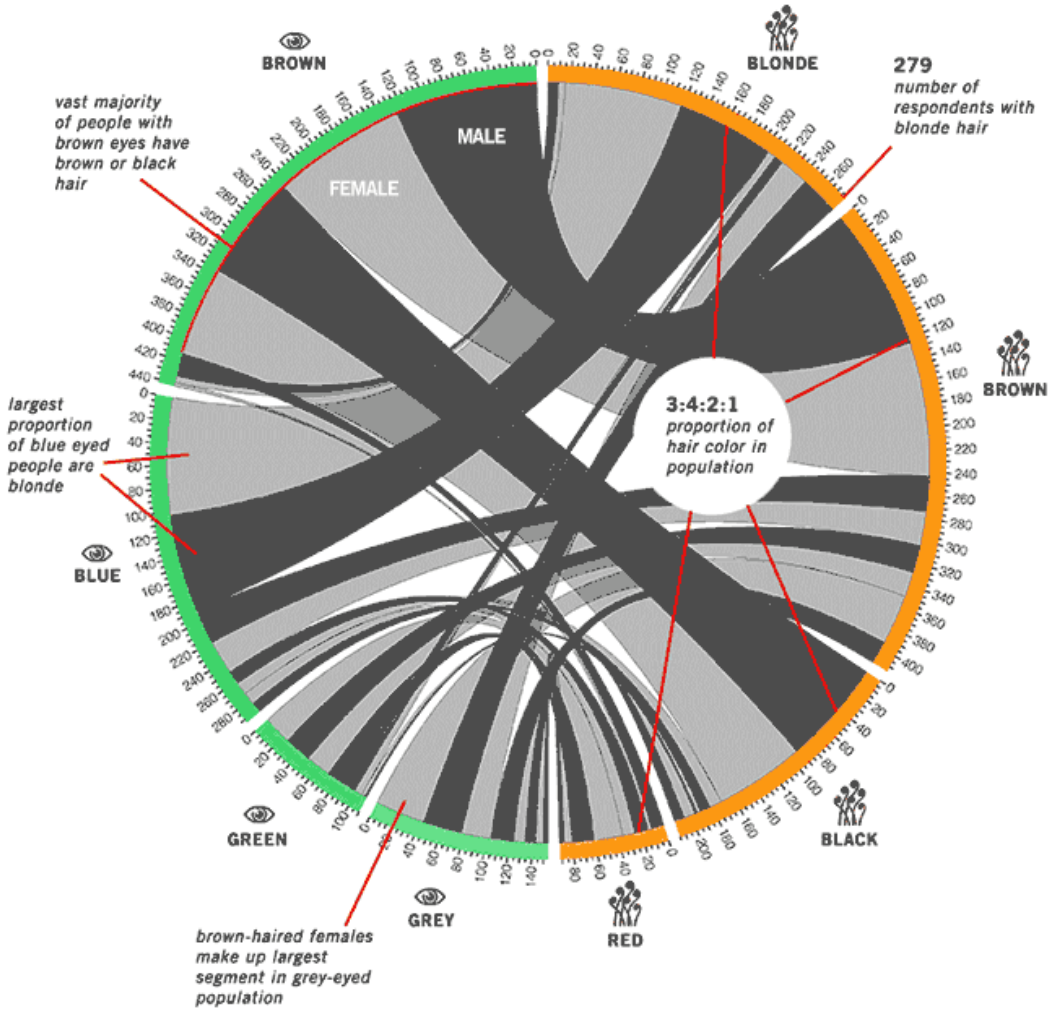
For this one day only we have been asked to relocate our class to Eaton 209

WORKING WITH CATEGORICAL DATA

REFRESH: TYPES OF DISTRIBUTIONS

- Joint Distribution
- Marginal Distribution
- Conditional Distribution

WARM UP EXAMPLE



WARM UP EXAMPLE

Consider the following survey of 100 students:

	Black	Brunette	Red	Blonde	Total
Black	11	20	4	1	**36**
Blue	3	14	3	16	**36**
Hazel	3	9	2	2	**16**
Green	2	5	2	3	**12**
Total	**19**	**48**	**11**	**22**	**100**

RECALL: TYPES OF DISTRIBUTIONS

What type of distribution is this?

	Black	Brunette	Red	Blonde	Total
Black	0.11	0.20	0.04	0.01	**0.36**
Blue	0.03	0.14	0.03	0.16	**0.36**
Hazel	0.03	0.09	0.02	0.02	**0.16**
Green	0.02	0.05	0.02	0.03	**0.12**
Total	**0.19**	**0.48**	**0.11**	**0.22**	**1.00**

RECALL: TYPES OF DISTRIBUTIONS

What type of distribution is this?

	Black	Brunette	Red	Blonde	Total
Black	0.11	0.20	0.04	0.01	**0.36**
Blue	0.03	0.14	0.03	0.16	**0.36**
Hazel	0.03	0.09	0.02	0.02	**0.16**
Green	0.02	0.05	0.02	0.03	**0.12**
Total	**0.19**	**0.48**	**0.11**	**0.22**	**1.00**

RECALL: TYPES OF DISTRIBUTIONS

What type of distribution is this?

	Black	Brunette	Red	Blonde	Total
Black	0.58	0.42	0.36	0.04	---
Blue	0.16	0.29	0.27	0.73	---
Hazel	0.16	0.19	0.18	0.09	---
Green	0.10	0.10	0.18	0.14	---
Total	**1.00**	**1.00**	**1.00***	**1.00**	---

* rounding error

RECALL: TYPES OF DISTRIBUTIONS

Where do these numbers come from?

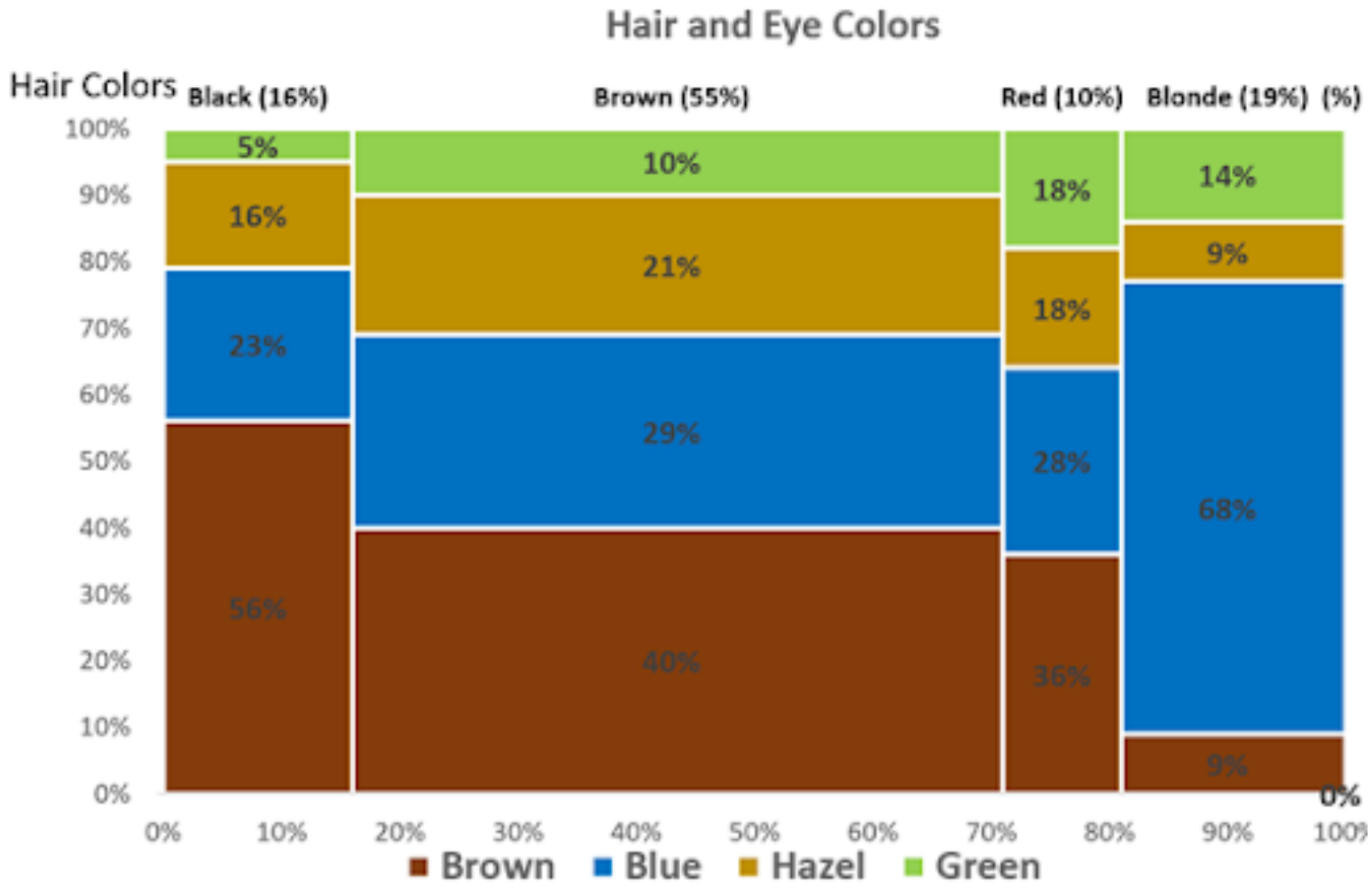
	Black
Black	11
Blue	3
Hazel	3
Green	2
Total	**19**

→

→

Conditional	Given Black
11/19	0.58
3/19	0.16
3/19	0.16
2/19	0.10
19/19	**1.00**

MOSAIC PLOTS!



EXAMPLE 2: IMMIGRATION POLICY

MOTIVATING EXAMPLE #2: IMMIGRATION POLICY

Views on immigration. Nine-hundred and ten (910) randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.[48]

QUESTIONS OF INTEREST

a. What percent of these Tampa, FL voters identify themselves as conservatives?

b. What percent of these Tampa, FL voters are in favor of the citizenship option?

c. What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

d. What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

e. Do political ideology and views on immigration appear to be associated? Explain your reasoning.

TRANSITION TO R STUDIO FOR OUR HANDS-ON ACTIVITY

GETTING STARTED

Step 0: Install the package

```
#install.packages("openintro")
library(openintro)
```

Step 1: Load the Data

```
data("immigration")
str(immigration)
```

```
## tibble [910 × 2] (S3: tbl_df/tbl/data.frame)
##  $ response : Factor w/ 4 levels "Apply for citizenship",..:
1 1 1 1 1 1 1 1 1 1 ...
##  $ political: Factor w/ 3 levels "conservative",..: 1 1 1 1 1
1 1 1 1 1 ...
```

STEP 2

Step 2: Re-level categories

By default R will order a variable alphabetically, but we might not want that.

```
immigration$political<-as.character(immigration$political)
immigration$political<-factor(immigration$political,
                              levels = c("conservative","moderat
e", "liberal"))
```

QUESTION 1

What percent of these Tampa, FL voters identify themselves as conservatives?

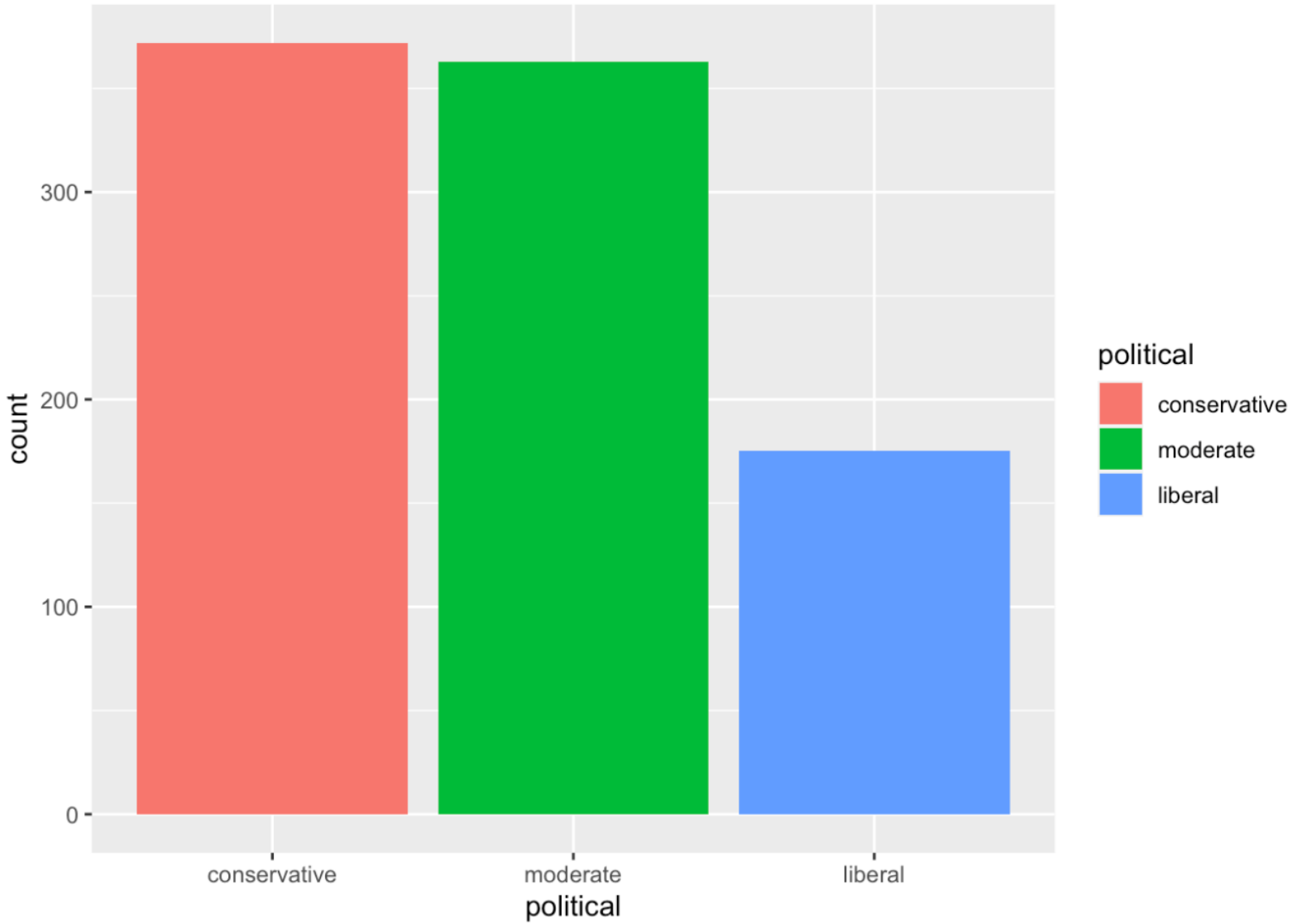
We will learn two new functions to work with individual level data:

- `table()`
- `prop.table()`

```r
# Table for Political affiliation
# use table() function
tabPol<-table(immigration$political)

# the prop.table() function must take a table object
prop.table(tabPol)
```

```
##
## conservative     moderate      liberal
##    0.4087912    0.3989011    0.1923077
```

```
# create a graph to display the distribution
ggplot(immigration, aes(x=political, fill=political))+
  geom_bar()
```



We can also use `kable` to make tables in R markdown:

```
library(knitr)
kable(tabPol, col.names = c('Party', 'Count'),
      caption = "Distribution of Political Indentities")
```

Distribution of Political Indentities

Party	Count
conservative	372
moderate	363
liberal	175

QUESTION 2

What percent of Tampa, FL voters are in favor of the citizenship option?

Use the functions

- `table()`
- `prop.table()`

```r
# Table for citizenship response
# use table() function
tabResp<-table(immigration$response)

# use prop.table()
prop.table(tabResp)
```

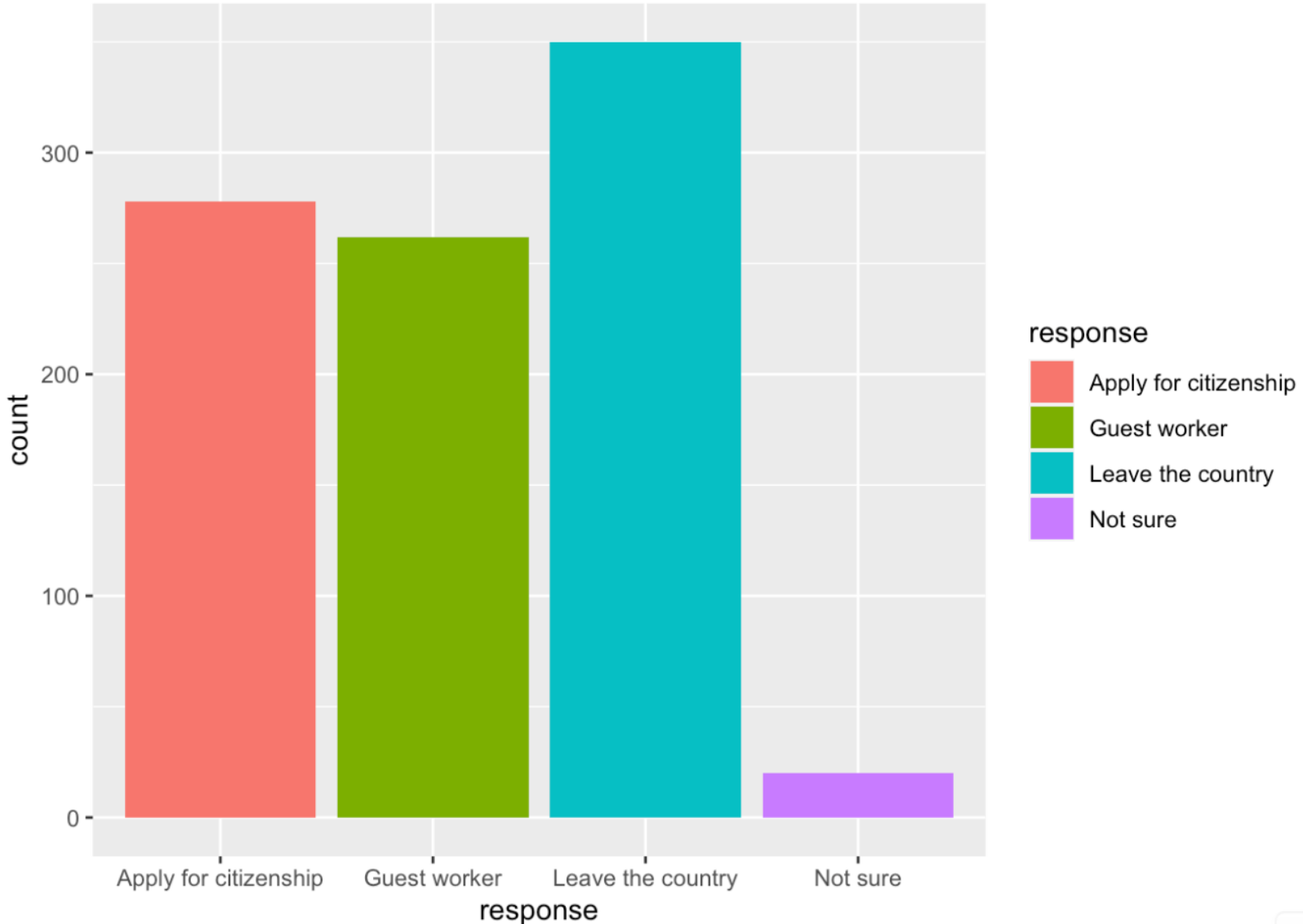
```
##
## Apply for citizenship          Guest worker     Leave the cou
ntry
##             0.30549451            0.28791209             0.3846
1538
##               Not sure
##             0.02197802
```

```
# kable
kable(tabResp, col.names = c('Response', 'Count'),
      caption = "Distribution of Response to Citizenship")
```

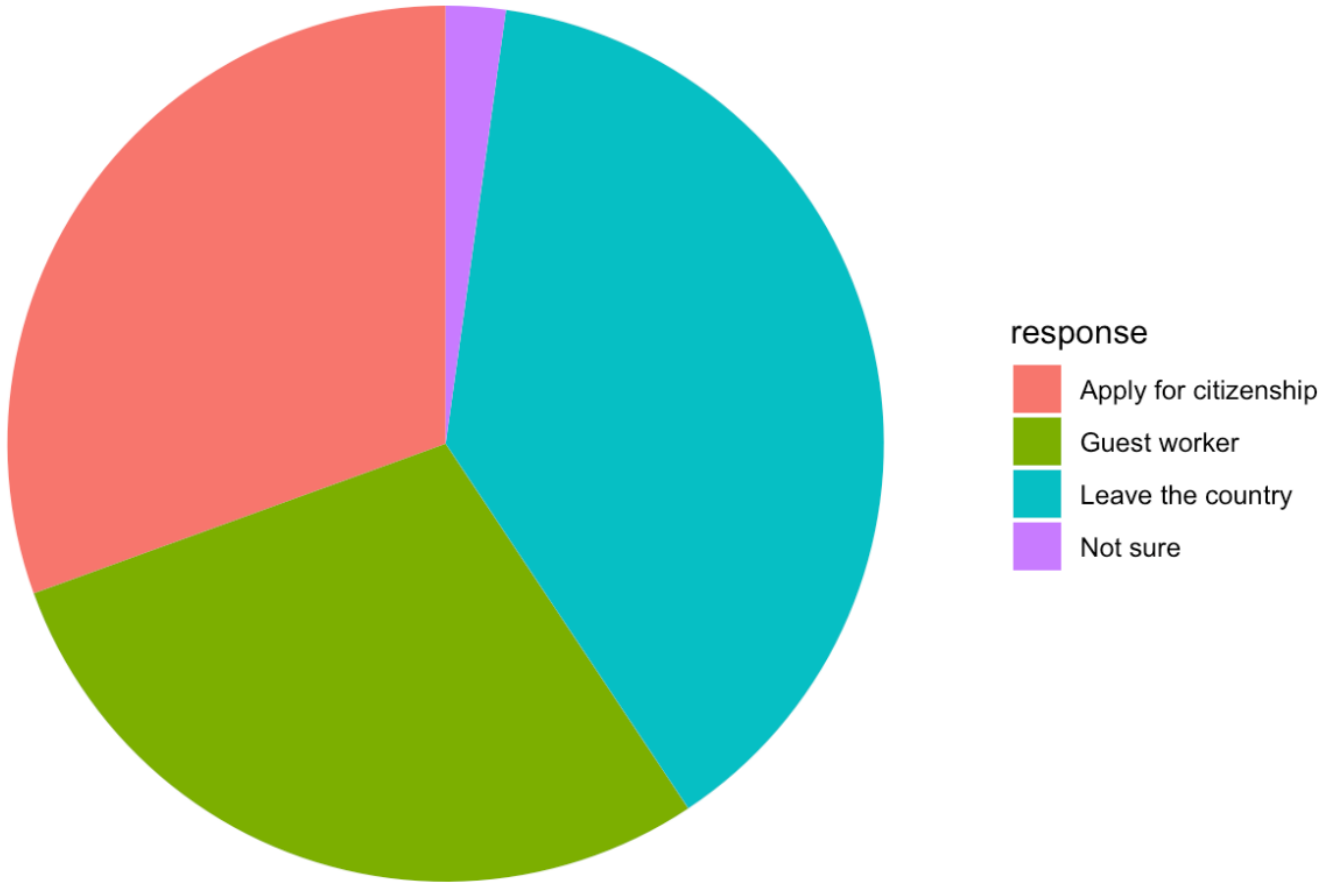
Distribution of Response to Citizenship

Response	Count
Apply for citizenship	278
Guest worker	262
Leave the country	350
Not sure	20

```
# create a graph to display the distribution
ggplot(immigration, aes(x=response, fill=response))+
  geom_bar()
```



```
# pie graph
ggplot(immigration, aes(x=1, fill=response))+
  geom_bar()+
  coord_polar("y", start=0)+
  theme_void()
```



QUESTION 3

What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

Use the functions

- `table()`
- `prop.table()`

```r
## conservative and citizen
# Row then col
tabPolResp<-table(immigration$political, immigration$response)
tabPolResp
```

```
##
##              Apply for citizenship Guest worker Leave the country Not sure
##   conservative                    57           121                179       15
##   moderate                       120           113                126        4
##   liberal                        101            28                 45        1
```

```
## kable
kable(tabPolResp)
```

	Apply for citizenship	Guest worker	Leave the country	Not sure
conservative	57	121	179	15
moderate	120	113	126	4
liberal	101	28	45	1

```
## joint
prop.table(tabPolResp)
```

```
##
##                Apply for citizenship Guest worker Leave the country    Not sure
##   conservative            0.062637363  0.132967033       0.196703297 0.016483516
##   moderate                0.131868132  0.124175824       0.138461538 0.004395604
##   liberal                 0.110989011  0.030769231       0.049450549 0.001098901
```

```
sum(prop.table(tabPolResp))
```

```
## [1] 1
```

```
## kable
kable(round(prop.table(tabPolResp),2))
```

	Apply for citizenship	Guest worker	Leave the country	Not sure
conservative	0.06	0.13	0.20	0.02
moderate	0.13	0.12	0.14	0.00
liberal	0.11	0.03	0.05	0.00

QUESTION 4

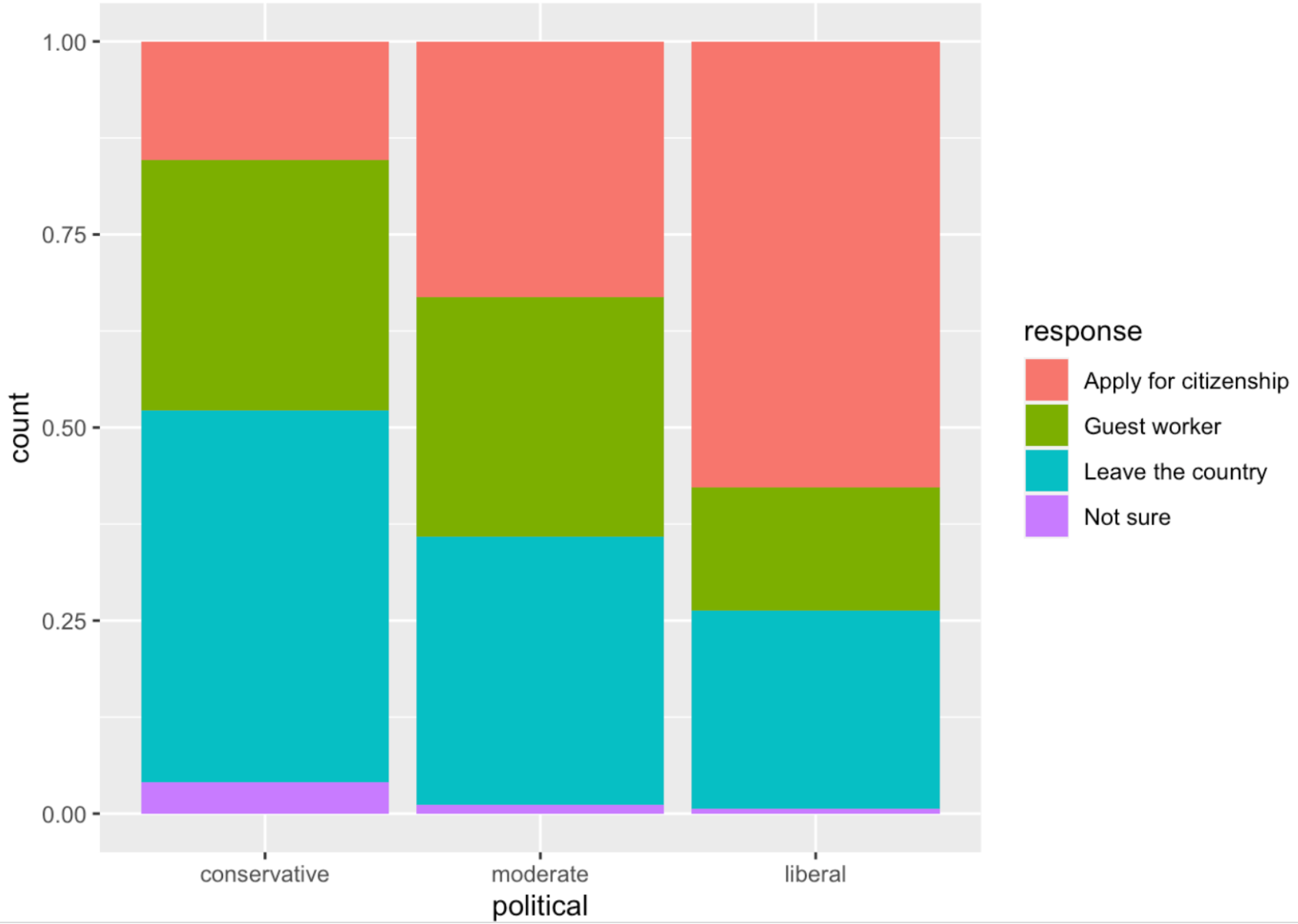
What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

```
## marginal prop
prop.table(tabPolResp, 1)
```

1 = marginal on the row dimension

```
##
##                Apply for citizenship Guest worker Leave the country     Not sure
##   conservative             0.153225806  0.325268817        0.481182796 0.040322581
##   moderate                 0.330578512  0.311294766        0.347107438 0.011019284
##   liberal                  0.577142857  0.160000000        0.257142857 0.005714286
```

```
ggplot(immigration, aes(x=political, fill=response))+
  geom_bar(position="fill")
```



EXAMPLE 3: SIMPSON'S PARADOX

SIMPSON'S PARADOX

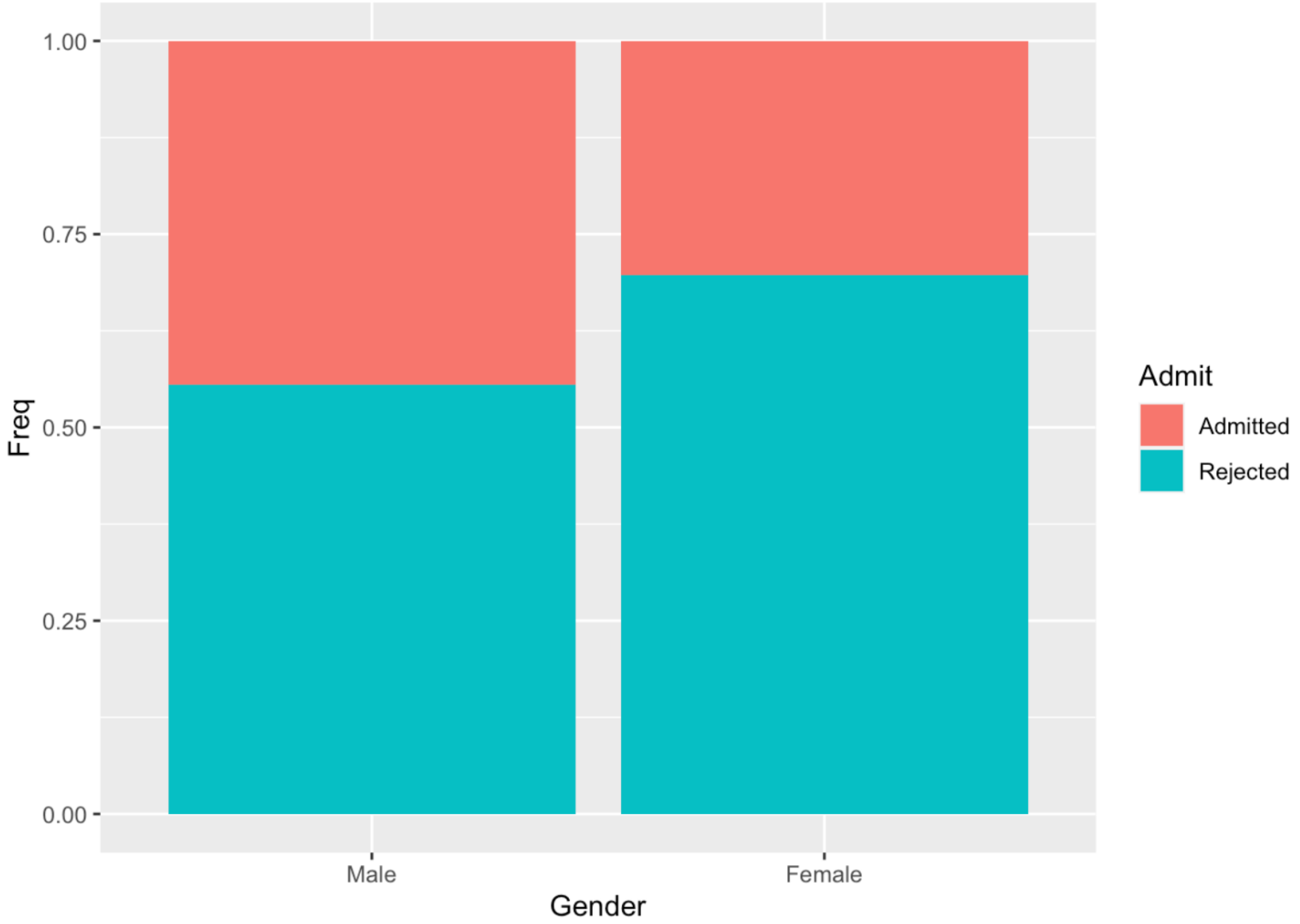
Simpson's Paradox (aka The Ecological Fallacy):
A phenomenon in which a trend appears in several different groups but disappears or reverses when the groups are combined.

EXAMPLE #3: SIMPSON'S PARADOX

- 1973 UC Berkeley Gender Bias in Admissions
- "One of the first universities to be sued for sexual discrimination" (with a statistically significant difference)

	Applicants Admitted	
Men	8442	44%
Women	4321	35%

EXAMPLE #3: SIMPSON'S PARADOX

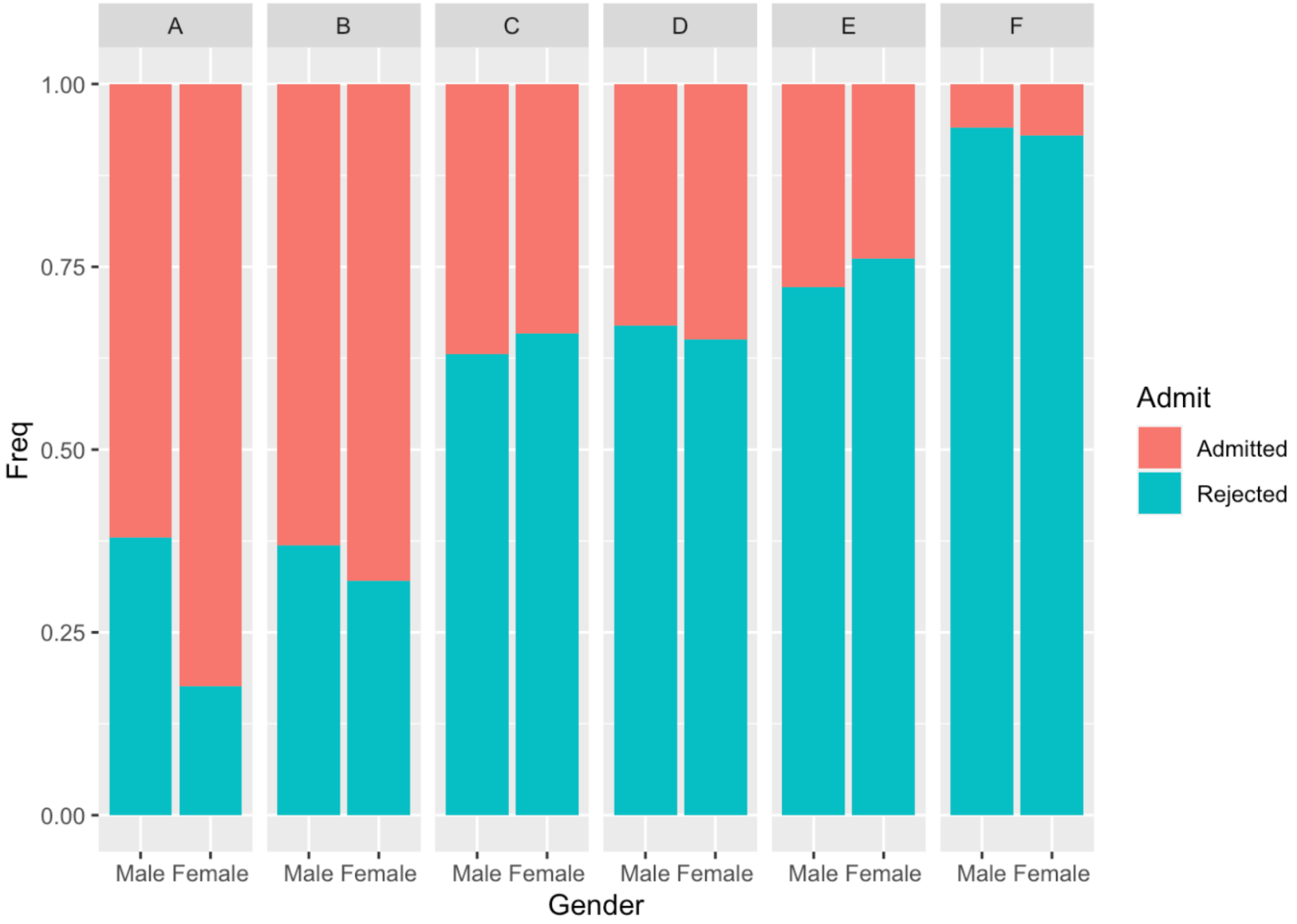


SIMPSON'S PARADOX

- But, when you dig into the data…

Department	# of Men	# of Women	Men Accepted	Women Accepted
A	**825**	**108**	62%	**82%**
B	560	25	63%	**68%**
C	325	593	37%	34%
D	417	375	33%	**35%**
E	191	393	28%	24%
F	**373**	**341**	6%	**7%**
Total	**8442**	**4321**		

EXAMPLE #3: SIMPSON'S PARADOX



SIMPSON'S PARADOX

How does this happen?

"The simple explanation is that ***women tended to apply to the departments that are the hardest to get into****, and men tended to apply to departments that were easier to get into. (Humanities departments tended to have less research funding to support graduate students, while science and engineer departments were awash with money.) So women were rejected more than men.* ***Presumably, the bias wasn't at Berkeley but earlier in women's education, when other biases led them to different fields of study than men****."*

TIME FOR GROUP WORK

MILESTONE #4

- **Due 10/20 - Milestone #4:** Exploratory Data Analysis Step #2
  - Create Tables and Bar Graphs
    - **Goal:** Work to answer at least one of your questions of interest for categorical variables of interest
    - Must have at least one one-way table, and at least one two-way table. Describe interesting distributions (joint, condition, marginal) and what they tell you about your categorical data
    - Must have at least one bar graph with one variable and at least one bar graph with two variables. Describe your graphical voices for how you are presenting your data.

- Please submit using Rmarkdown

ANOTHER ACTIVITY…
IF WE HAVE TIME

FIVETHIRTYEIGHT ACTIVITY

READ THE ARTICLE

JUN. 26, 2020, AT 7:00 AM

Voter Registrations Are Way, Way Down During The Pandemic

By Kaleigh Rogers and Nathaniel Rakich

Graphics by Elena Mejia Lutz

Filed under 2020 Election

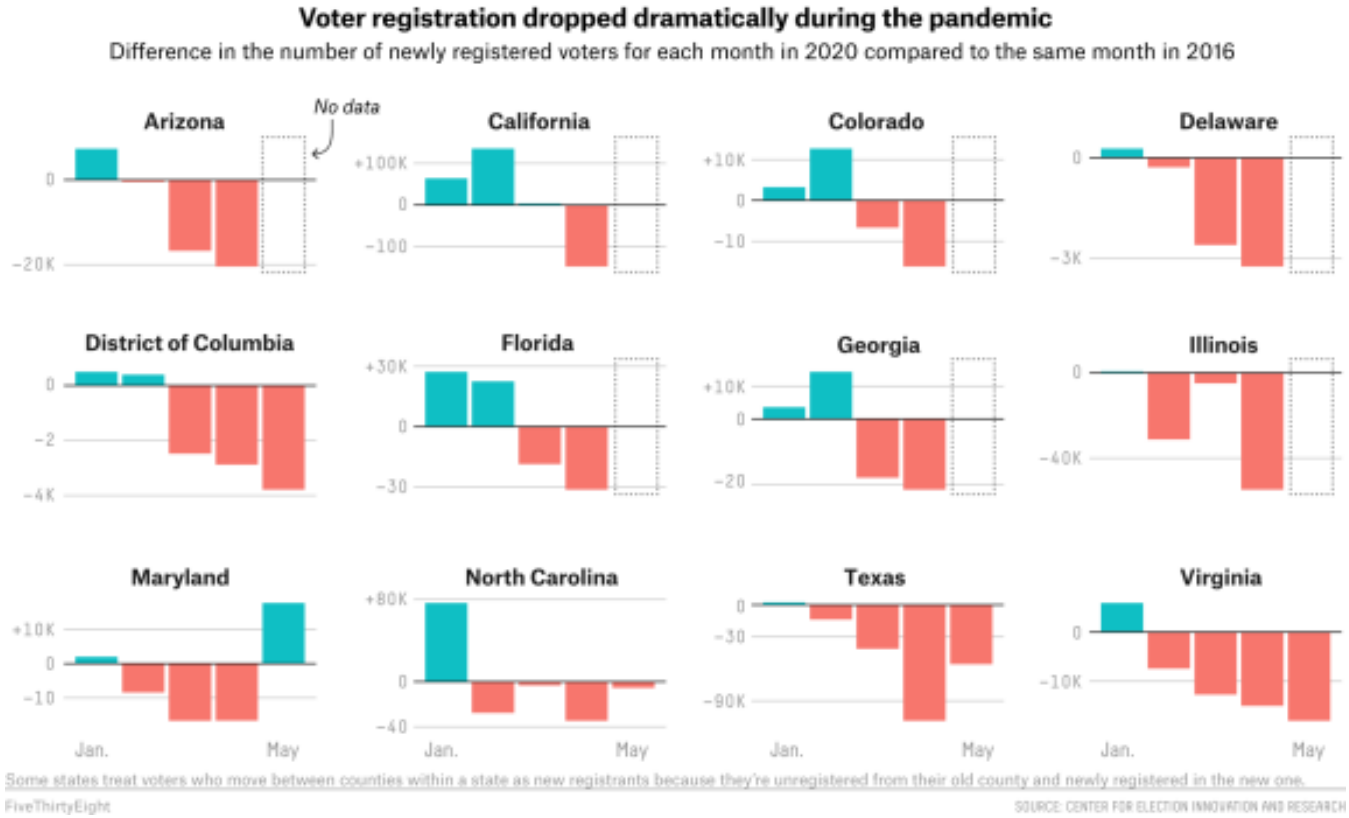
Get the data on GitHub

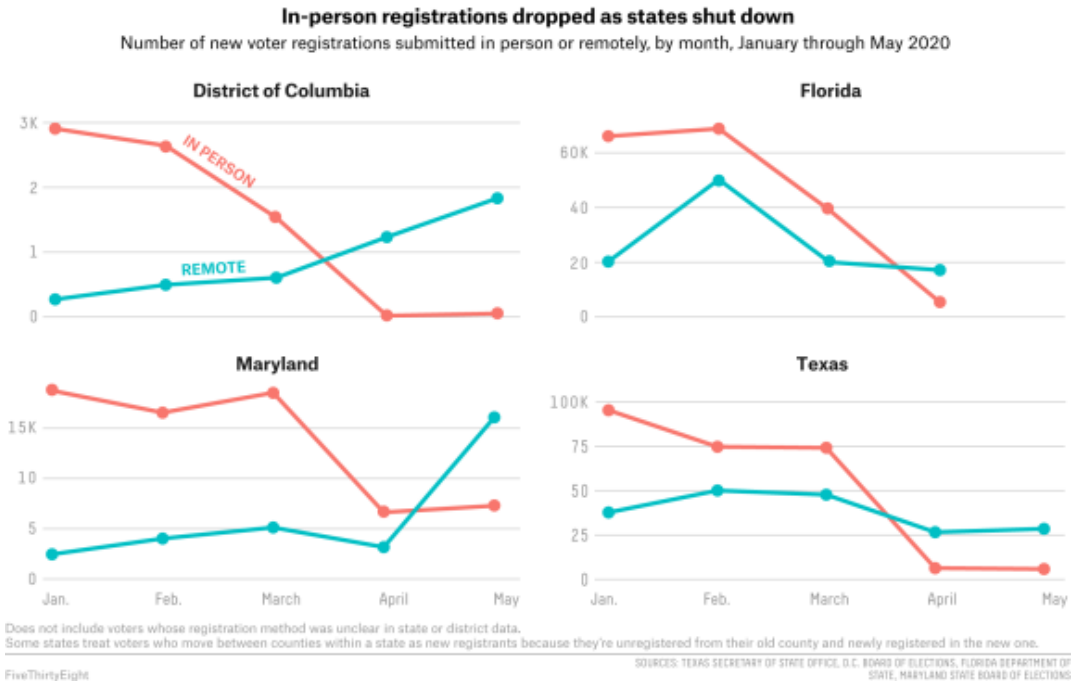
Volunteers instruct citizens how to register to vote during a protest march. ROBERT NICKELSBERG / GETTY IMAGES

Poll after poll showed a high level of enthusiasm for voting in the general

DISCUSS IN SMALL GROUPS

1. How are graphics used to tell the author's story?
2. What geometries are used?





WHAT DOES THE RAW DATA LOOK LIKE?

How to access the data:

```
# Load the tidyverse
library(tidyverse)

# Import data
vreg<-read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/maste
r/voter-registration/new-voter-registrations.csv",
                header=TRUE)
```

WE MIGHT WANT TO RELEVEL

1. **Relevel the data so that its in the right order:**

```
# Level the Month variable so that its in the right order (ie not alphabetical)
vreg$Month<-factor(vreg$Month,
                   levels=c("Jan", "Feb", "Mar", "Apr", "May"))
```

ARE WE GOING TO NEED TO TIDY THE DATA?

2. Tidy the data:

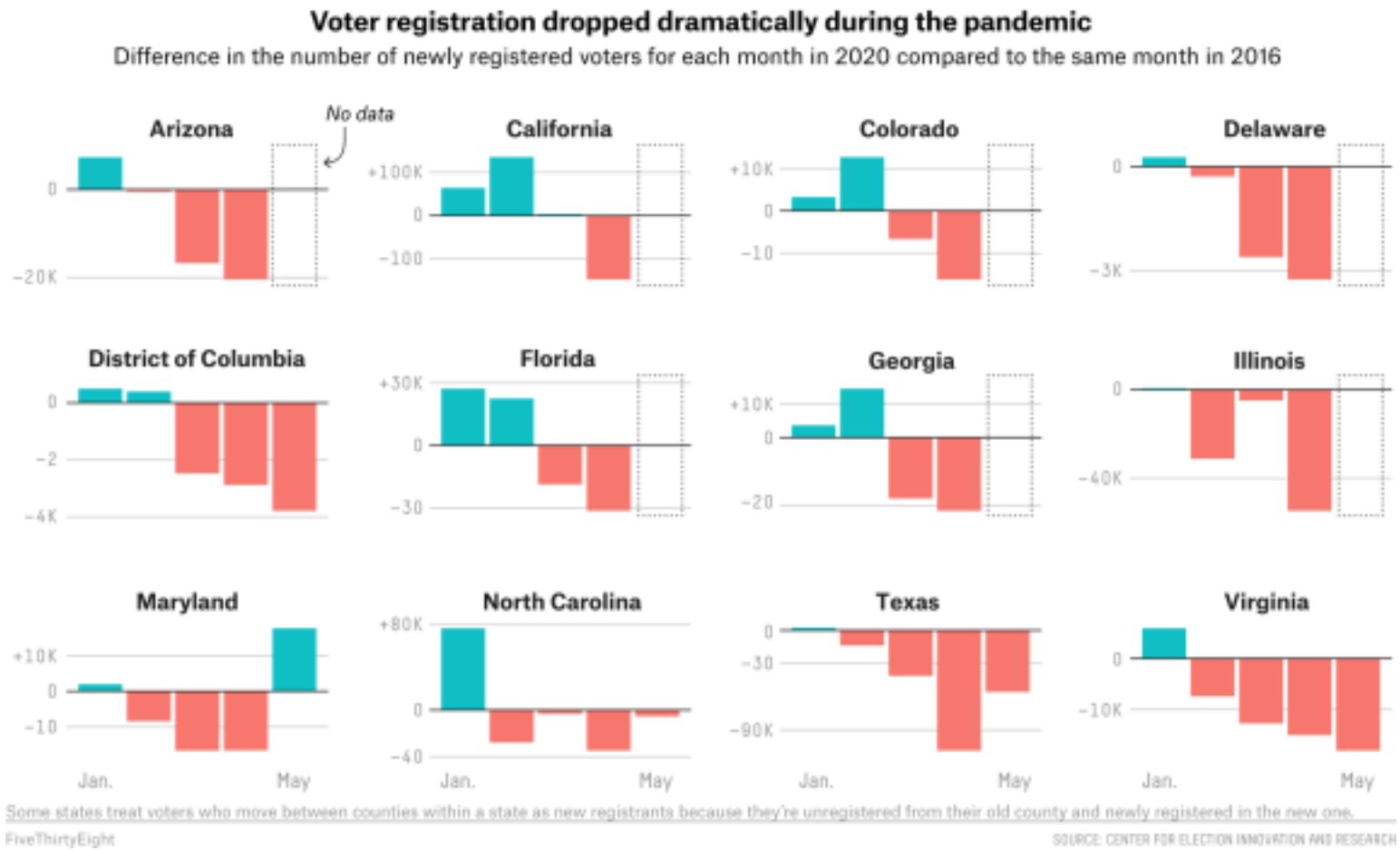
```
### USE spread() FROM tidyr
vregYear<-vreg%>%
  spread(Year, New.registered.voters)

### RENAME THE COLUMNS
colnames(vregYear)<-c("Jurisdiction", "Month", "Y2016", "Y2020")
```

MUTATE!

3. **Mutate to add the change:**

```
### mutate() FROM dplyr()
vregChange<-vregYear%>%
  mutate(change=Y2020-Y2016)
```

RECREATE THIS GRAPH IN SMALL GROUPS



Task: Using the tools we have covered so far, recreate this graph.