[redacted]

[redacted]

by

Scott Greenberg

[redacted]

Date: [redacted]

[redacted]

[redacted]

[redacted]

[redacted] [redacted]

[redacted]

[redacted]

[redacted]

(i) [redacted]

Keywords such as “[redacted]”, based on the “Development, Relief, and Education for Alien Minors” (DREAM) act, and “DACA”, in full “Deferred Action for Childhood Arrivals,” made headline news in 2017 and 2018 and even resulted in a Government shutdown in January 2018. On January 20, 2021, they made headlines again in one of President Biden’s action plans, see `https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/20/preserving-and-fortifying-deferred-action-for-childhood-arrivals-daca/`.

Where do these DREAMERS live? The web site `https://qz.com/1069844/who-are-the-dreamers-and-where-do-they-live/` provides an initial answer to this question. The graph (a simple bar chart) can be found at `https://www.theatlas.com/charts/SysTw83tb`. I downloaded the data file called

`hw02_DACA_recipients_by_state.csv`

from this web site on 2/15/2018. Work with this version of the data file for this HW and do not download any updated version to avoid any confusion when it comes to interpreting these data.

(a) (2 Points) Load all R packages you need for this question. Read in the data. Make sure it is a data frame or transfer it into a data frame if it is not. Omit the data for the District of Columbia. Show the first 6 rows of the data frame and verify that there are 50 states remaining. Show your R code.

Answer:

```
library(httr)
```

```
## Warning:  package ’httr’ was built under R version 4.0.4
```

```
library(XML)
```

```
## Warning:  package ’XML’ was built under R version 4.0.3
```

```
library(maps)
```

```
## Warning:  package ’maps’ was built under R version 4.0.4
```

```
library(mapdata)
```

```
## Warning:  package ’mapdata’ was built under R version 4.0.4
```

```
library(RColorBrewer)
```

```
## Warning:  package ’RColorBrewer’ was built under R version 4.0.3
```

```
d.r.s <- read.csv("hw02_DACA_recipients_by_state.csv")
d.r.s.o <- d.r.s[d.r.s[, 1] != "District of Columbia", ]
head(d.r.s.o)
```

```
##        state  pct
## 1 California 0.57
## 2      Texas 0.45
## 3     Nevada 0.44
## 4    Arizona 0.40
## 5   Illinois 0.33
## 6 New Mexico 0.33
```

(b) (20 Points) Create two choropleth maps for the “pct” column of this data set. For one of these choropleth maps, split the data into quantiles of about 10 observations each. For the other one, use five equally wide intervals that range from 0% to 0.60%. As the data are rounded to just two decimals, it may not be possible to have exactly 10 observations in each class.

Work with the *maps* R package, but recall that *map(“state”)* only produces a map of the 48 lower states. So you need to add Alaska and Hawaii to the upper left and lower left corners, respectively. This can be done by creating a meaningful layout with the *layout* function. You also need the *mapdata* R package and then you can create maps for these two states via `map("world2Hires", "USA:Alaska")` and `map("world2Hires", "Hawaii")`. Include meaningful titles, labels, and legends. Use intervals that are closed on the **right** side. The first interval also should be closed on the left side. Use the 5–color “Blue” color scheme from the *RColorBrewer* R package. Be consistent in your coloring, i.e., make sure that Alaska and Hawaii get colors that match the 48 lower states. Copy–and–paste mistakes are easy to make in such similar plots — so check your results carefully, in particular for the second map you create!

As in *“Statistical Visualization I”*, create your graphs step–by–step and make refinements as needed. However, no need to include any of your initial or intermediate graphs unless I ask for those. Just include your final graphs. Always include your R code.

Answer:

```
layout.matrix <- matrix(1, nrow = 2, ncol = 3)
layout.matrix[1, 1] <- 2
layout.matrix[2, 1] <- 3
layout.matrix

##      [,1] [,2] [,3]
## [1,]    2    1    1
## [2,]    3    1    1

breaks <- c(0, .12, .24, .36, .48, .6)
p.break.class <- cut(d.r.s.o[, 2], breaks, include.lowest = TRUE)
p.break.col <- brewer.pal(5, "Blues")[p.break.class]
map.p.break.col <- p.break.col[order(d.r.s.o[, 1])][match.map("state",
                                        state.name)]

quant <- c(d.r.s.o[50, 2], d.r.s.o[40, 2], d.r.s.o[30, 2],
           d.r.s.o[20, 2], d.r.s.o[10, 2], d.r.s.o[1, 2])
p.quant.class <- cut(d.r.s.o[, 2], quant, include.lowest = TRUE)
```

```
p.quant.col <- brewer.pal(5, "Blues")[p.quant.class]
map.p.quant.col <- p.quant.col[order(d.r.s.o[, 1])][match.map("state",
                                         state.name)]
pdf("hw02_sol_dreamers_equal.pdf")
layout(layout.matrix)
par(mar = c(0, 0, 16, 0))
map("state", fill = TRUE, col = map.p.break.col)
title("Percent Daca Recepients by State w/ equally wide intervals")
par(mar = c(0, 0, 0, 0))
map("world2Hires", "USA:Alaska", fill = TRUE,
    col = cbind(d.r.s.o, p.break.col)[d.r.s.o[, 1] == "Alaska", 3])
par(mar = c(0, 0, 0, 0))
map("world2Hires", "Hawaii", fill = TRUE,
    col = cbind(d.r.s.o, p.break.col)[d.r.s.o[, 1] == "Hawaii", 3])
legend("bottomleft", legend = levels(p.break.class), title.adj = 0.12,
       title = "Equally spaced intervals of percents", cex = .6,
       fill = brewer.pal(5, "Blues"))
dev.off()

## pdf
##   2
```

```
pdf("hw02_sol_dreamers_quantile.pdf")
layout(layout.matrix)
par(mar = c(0, 0, 16, 0))
map("state", fill = TRUE, col = map.p.quant.col)
title("Percent Daca Recepients by State w/ intervals by quantile")
par(mar = c(0, 0, 0, 0))
map("world2Hires", "USA:Alaska", fill = TRUE,
    col = cbind(d.r.s.o, p.quant.col)[d.r.s.o[, 1] == "Alaska", 3])
par(mar = c(0, 0, 0, 0))
map("world2Hires", "Hawaii", fill = TRUE,
    col = cbind(d.r.s.o, p.quant.col)[d.r.s.o[, 1] == "Hawaii", 3])
legend("bottomleft", legend = levels(p.quant.class), title.adj = 0.12,
       title = "Percent intervals spaced by quantiles", cex = .6,
       fill = brewer.pal(5, "Blues"))

dev.off()

## pdf
##   2
```

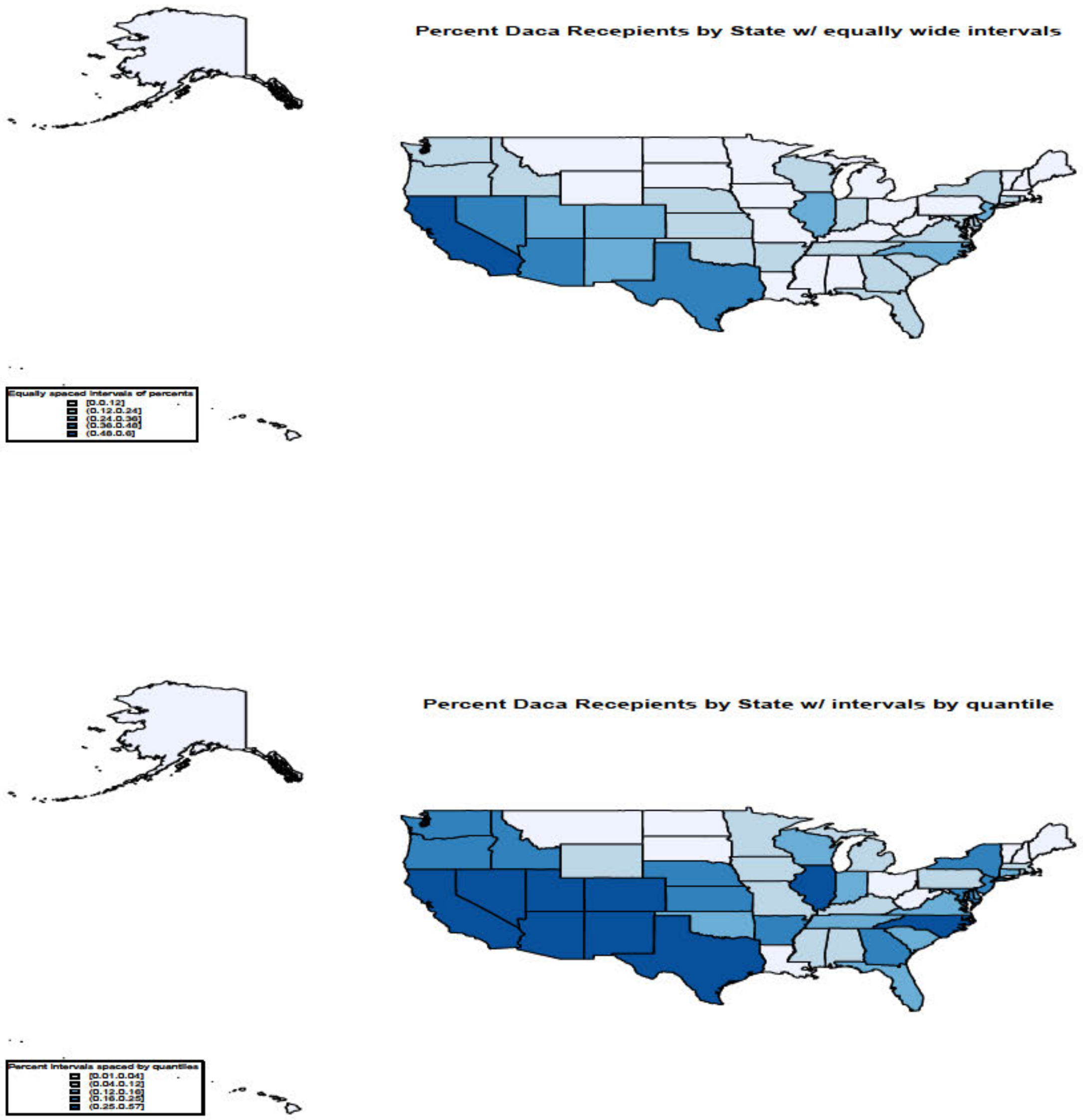
Percent Daca Recepients by State w/ equally wide intervals
Equally spaced intervals of percents
[0.0.12]
(0.12.0.24]
(0.24.0.36]
(0.36.0.48]
(0.48.0.6]
Percent Daca Recepients by State w/ intervals by quantile
Percent intervals spaced by quantiles
[0.01.0.04]
(0.04.0.12]
(0.12.0.16]
(0.16.0.25]
(0.25.0.57]

Figure 1: Two choropleth maps showing the percentage of DACA recipients by state. The map at the top uses equally wide intervals. The map at the bottom uses quantiles with about the same number of states in each class.

(c) (8 Points) Describe the geographic pattern (if any) you see in your maps. Do both of these choropleth maps convey the same messages? Or, do they lead to different interpretations? Be specific and explain any major differences.

Explicitly mention Alaska and Hawaii in your description. This will hopefully enforce that you check that they are colored correctly in both maps.

<u>Answer:</u>

When we compare the two choropleth maps we see that California has a disproportionately large percent of DACA reciepients. The value is so large that it is the only entry in its interval when looking at the top choropleth map (with equally wide intervals). While this information is lost when we look at the bottom choropleth map – as the southwest becomes a grouping of the largest amount and the northwest a grouping of the second largest amount (Hawaii and Alaska respectively excluded as they are both always in the group for the smallest amount) – we can distinguish better between groups with a small amount, and a very small amount for percent DACA recipients by state.

(ii) [redacted] (35 Points):

Alabama was the first state after the 2016 Presidential election where a special election took place on December 12, 2017, to fill in a United States Senate seat. Doug Jones, a Democrat, won against Roy S. Moore, a Republican, making Jones the first Democrat to win a Senate seat in Alabama in over 25 years. Results at the county level can be obtained from `https://www.nytimes.com/elections/results/alabama-senate-special-election-roy-moore-doug-jones`.

I am using the R code below to scrape these data from the web:

```r
library(httr)
library(XML)

page <- GET("https://www.nytimes.com/elections/results/alabama-senate-special-election-roy-moore-doug-jones")
summary(page)

# extract the nodes related to the tables
pagehtml <- htmlParse(page)
summary(pagehtml)
nodes <- getNodeSet(pagehtml, "//table")
nodes
class(nodes)
summary(nodes)
head(nodes[[2]])

# extract the election table from the page
etable <- readHTMLTable(nodes[[2]])
head(etable)

colnames(etable) <- gsub(" |\n|[-]|[.]", "", colnames(etable))
etable$JonesNum <- as.integer(gsub(",", "", etable$Jones))
etable$MooreNum <- as.integer(gsub(",", "", etable$Moore))
etable$WriteInsNum <- as.integer(gsub(",", "", etable$WriteIns))
head(etable)
```

(a) (2 Points) Load all additional R packages you need for this question. Calculate the percentages for Jones and Moore in each of the 67 counties, based on all votes (including write–ins) cast in these counties. Determine the winning percentages for Jones and Moore in each county and set as NA for the candidate who lost that county. Show the first 6 rows of the resulting data frame. Show your R code.

Answer:

```
library(data.table)

## Warning:  package 'data.table' was built under R version 4.0.3

maxs <- as.vector(unlist(lapply(data.frame(t(etable[, 6:8])), max)))
percents <- (maxs/rowSums(etable[, 6:8]))*100
winner <- max.col(etable[, 6:8])
na.df <- data.frame(matrix(nrow = 67, ncol = 2))
na.df[winner == 1, 1] <- percents[winner == 1]
na.df[winner == 2, 2] <- percents[winner == 2]
names(na.df) <- c("JonesWinPct", "MooreWinPct")
p.table <- cbind(etable, na.df)
head(p.table)

##       County   Jones  Moore WriteIns  Rpt JonesNum MooreNum WriteInsNum
## 1  Jefferson 149,522 66,309    3,710 100%   149522    66309        3710
## 2    Madison  65,664 46,313    3,446 100%    65664    46313        3446
## 3     Mobile  62,253 46,725    1,539 100%    62253    46725        1539
## 4 Montgomery  48,186 17,705      743 100%    48186    17705         743
## 5     Shelby  27,251 36,424    1,718 100%    27251    36424        1718
## 6    Baldwin  22,131 38,445    1,699 100%    22131    38445        1699
##   JonesWinPct MooreWinPct
## 1    68.10664          NA
## 2    56.88987          NA
## 3    56.32889          NA
## 4    72.31443          NA
## 5          NA    55.70015
## 6          NA    61.73424
```

(b) (5 Points) Create side–by–side boxplots for the winning percentages for the 2 candidates in the 67 counties. What do you observe? There is one interesting feature for one of the counties won by Moore. Which? Keep in mind that we have write–ins! Look very carefully or use additional summary statistics. Answer this part in 3 or 4 sentences and show your R code.

Answer:

```
pdf("hw02_sol_alabama_boxplots.pdf")
boxplot(p.table[, 9:10], ylab = "Winning percents",
        xlab = "Winner", main = paste0("Winning percents for ",
        "counties in Alabama's 2017 Special Senate Election"))
dev.off()

## pdf
##   2
```

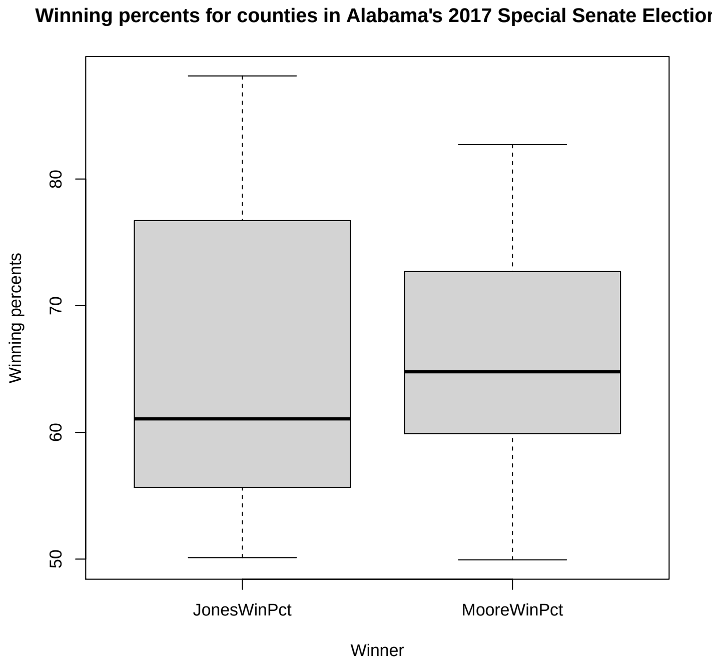

Figure 2: Boxplots for the 67 Alabama Counties of the Winning Percentages of the 2017 Alabama Special Election in the Doug Jones vs. Roy Moore Senate Race.

Comment:
What do you observe?

I observe Moore's boxplot being more narrow (having a larger 1st quartile and median but not 3rd quartile and maximum) than Jones' boxplot.

There is one interesting feature for one of the counties won by Moore. Which?

In Monroe county, Moore didn't win by recieving a majority of the vote, but rather by receiving a plurality of the vote. (Levy 2021) In other words he won the county but had less than 50 percent of the county's vote.

(c) (20 Points) Create a map using the *maps* R package, similar to the one shown at `https://www.nytimes.com/elections/results/alabama-senate-special-election-roy-moore-doug-jones`.
Use darker blue tones for higher winning percentages of Jones and darker red tones for higher winning percentages of Moore. Use a 10–class red–blue divergent color scheme that can be obtained as `brewer.pal(10, "RdBu")` from the *RColorBrewer* R package. Use similar intervals as used by the New York Times, but further improve their intervals to cover the entire range of the winning percentages. Do not forget to create a meaningful legend. Also display the names of the five cities on the map that are shown in the New York Times map. There exists a *map.cities()* R function in the *maps* R package. See the help page and do some experimentation such that only these five cities are listed by name on the map. Show your R code.

Answer:

```r
pdf("hw02_sol_alabama.pdf")
a.breaks <- c(49, 57, 65, 73, 81, 89)
a.break.class.d <- cut(p.table[, 9], a.breaks)
a.break.class.r <- cut(p.table[, 10], a.breaks)
d.col <- brewer.pal(10, "RdBu")[c(6, 7, 8, 9, 10)][a.break.class.d]
r.col <- brewer.pal(10, "RdBu")[c(5, 4, 3, 2, 1)][a.break.class.r]
a.col <-fcoalesce(d.col, r.col)
map.a.col <- a.col[order(gsub("^St", "S", etable[, 1]))]
par(xpd = FALSE)
map("county", "Alabama", fill = TRUE, border = "White",
    col = map.a.col)
map.cities(minpop = 80000)
title(paste0("Winning percents of counties in Alabama's 2017 ",
             "Special Senate Election"))
par(xpd = TRUE)
legend("bottom", legend = levels(a.break.class.d), title.adj = 0.12,
       title = " (D) Jones' Win% Interval", cex = .5,
       fill = brewer.pal(10, "RdBu")[c(6, 7, 8, 9, 10)])
legend("bottomright", legend = levels(a.break.class.r),
       title.adj = 0.12, title = " (R) Moore's Win% Interval",
       cex = .5, fill = brewer.pal(10, "RdBu")[c(5, 4, 3, 2, 1)])
dev.off()

## pdf
##   2
```

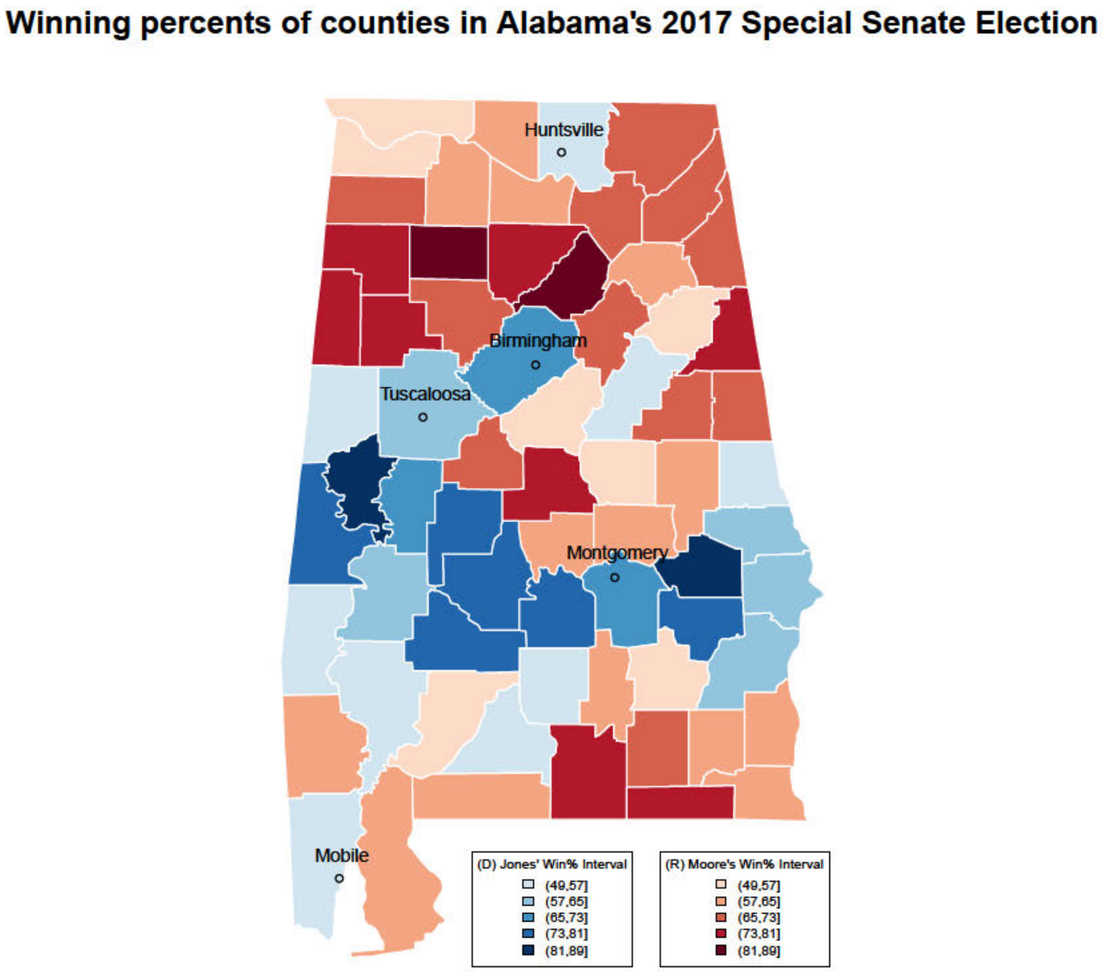

Figure 3: County–by–county Outcome of the 2017 Alabama Special Election in the Doug Jones (Dem) vs. Roy Moore (Rep) Senate Race. The map shows the winning percentage (in %) in each of the 67 Alabama counties.

(d) (8 Points) **Carefully compare your map with the map posted at** `https://www.nytimes.com/elections/results/alabama-senate-special-election-roy-moore-doug-jones`.

If you took *Data Technologies* in a previous semester, you should remember what might go wrong with geographic data and names from multiple sources. Compare spellings, the order how subregions appear in the different sources, etc. Once you are convinced that everything is correct in your map, describe the geographic pattern in the map. Where did Jones win and where did Moore win? Provide some city or county names and provide some interesting quantitative results here.

Clearly, if Moore won over Jones in Birmingham or Montgomery, something must have gone totally wrong in your map! It may also be a good idea to look at a real map of Alabama to answer this part.

Answer:

Doug Jones won most of the counties in the center of the state, in addition to the counties of the top 5 cities. However, the counties containing the cities of Huntsville and Mobile (Madison and Mobile), were won by a small margin, presumably due to being in the northernmost and southernmost parts of the state respectively. The most partisan counties for Moore, Blount and Winston, have a smaller win percentage (81.8 and 82.7 percent respectively) than the most partisan counties for Jones, Macon and Greene (88.1 and 87.6 percent respectively). Interestingly, all of these most partisan counties have a small amount of total voters, possibly indicating that they are all rural counties. In addition, with the exception of Blount County bordering Jefferson County (containing Birmingham), these partisan counties aren't near counties with large percentages of their opposition.

General Instructions

(i) Create a single pdf document, using R Markdown, Sweave, or knitr. When you take this course at the 6000–level, you have to use LaTeX in combination with Sweave or knitr. You only have to submit this one document to Canvas.

(ii) Include a title page that contains your name, your A–number, the number of the assignment, the submission date, and any other relevant information.

(iii) Start your answers to each main question on a new page (continuing with the next part of a question on the same page is fine). Clearly label each question and question part. Your answer to question (i) should start on page 2!

(iv) Show your R code and resulting graph(s) [if any] for each question part!

(v) Before you submit your homework, check that you follow all recommendations from Google's R Style Guide (see `http://web.stanford.edu/class/cs109l/unrestricted/resources/google-style.html`). Moreover, make sure that your R code is consistent, i.e., that you use the same type of assignments and the same type of quotes throughout your entire homework.

(vi) Give credit to external sources, such as stackoverflow or help pages. Be specific and include the full URL where you found the help (or from which help page you got the information). Consider R code from such sources as "legacy code or third–party code" that does not have to be adjusted to Google's R Style (even though it would be nice, in particular if you only used a brief code segment).

(vii) **Not following the general instructions outlined above will result in point deductions!**

(viii) For general questions related to this homework, please use the corresponding discussion board in Canvas! I will try to reply as quickly as possible. Moreover, if one of you knows an answer, please post it. It is fine to refer to web pages and R commands, but do not provide the exact R command with all required arguments or which of the suggestions from a stackoverflow web page eventually worked for you! This will be the task for each individual student!

(ix) Submit your single pdf file via Canvas by the submission deadline. Late submissions will result in point deductions as outlined on the syllabus.

References

Levy, L. (2021), ‘”majority” vs. ”plurality”: What their differences mean for this election’, *Dictionary.com* .
URL: *https://www.dictionary.com/e/majority-vs-plurality/*