Skills Problem Set 2

Fernanda Sobrino

3/18/2021

Due Thursday April 15, midnight Central Time.

Upload your pdf to canvas.

Push your code to your repo on Github Classroom.

This submission is my work alone and complies with the 30535 integrity policy.

Add your initials to indicate your agreement: **ES**

Add names of anyone you discussed this problem set with:

Guillermo Antonio Trefogli Wong

Late coins used this pset: 0. Late coins left after submission: 9.

Name your submission files `skills_ps_2.Rmd` and `skills_ps_2.pdf`. (10 points)

1 1 git concepts (10 points)

1. git is software for distributed version control. List 4 benefits of distributed version control.

```
# Changes to files are tracked between computers,  so all the developers could
# keep track of what was happening to files at any given time.

# Branching and merging can happen automatically and quickly.

# Developers can work on code offline. This increases productivity

# Multiple copies of the software eliminate reliance on a single backup file.
```

For the next questions, we will reference "Learn git concepts, not commands". Read sections from "Overview" through "Branching". It is written with git command line in mind, but github desktop has all these features as well. Focus on the concepts (ie understand the pictures). Save some changes to your homework and make sure its being tracked by git with github deskop.

i. What is the remote repository for this homework? Be as specific as possible.

```
# For my specific homework, my remote repository is linked below:
# https://github.com/datasci-harris/problem-set-skills-2-guccimane457

# answer source: https://github.com/UnseenWizzard/git_training
# Github Classroom is the remote repository for this homework.
# The remote repository is where you send your changes when you want to share
# them with other people, and where you get their changes from.
```

ii. How do you add a file to staging in github desktop? (This is subtle, because it happens automatically).

```
# After creating a Github account, downloading Github desktop application,
# logging into your account...

# then accepting the assignment and your assignment repository has been
# created in Github...

# We "add" a file to staging in github desktop by working on our assignment
# using Rstudio. The assignment files are in pdf and rmd.

# The staging area can hold changes from any number of files that you want to
# commit as a single snapshot. All your changes now appear in GitHub Desktop.
# Decide whether they go together in one commit, or need separate commits, and
# use the blue bars and ticky boxes and unstage or restage the lines.

# answer source: https://jcszamosi.github.io/mcmaster_swc_git_gui/03-create-changes/#:~:text=The%20stag
```

iii. How do you commit an issue to the local repository? (This is not subtle).

```
# We commit issues to the local respository with the "commit to main" command,
# which is also a blue button in the lower left hand corner of Github Desktop.
# Note that a commit is not automatically transferred to the remote server.
# If you want to exchange commits with others or share them, use the "Push"
# command (Ctrl + P). This will push your code to the remote repository.

# answer source: https://www.git-tower.com/learn/git/commands/git-commit/
```

iv. How does github desktop decide what part of your code to show in the main part of the window?

```
# Github desktop decides what part of your code to show in the main part of
# the window based on your "Current Branch". Branches are used by Github to
# organize your code as you progress and commit versions of your code. As you
# or others make changes, you can create a new branch with a unique name to
# separate your or everyones unique changes.

# answer source: https://thenewstack.io/dont-mess-with-the-master-working-with-branches-in-git-and-gith
```

v. What branch are you on right now? Why?

```
# I am on the main branch, where all changes eventually merge back to, and is
# the official working version of my project. The main branch is the default
# branch, and is the one your code will commit to if you have not yet created
# any new branches (such is the case for me with this pset)
```

vi. If you were to click on “current branch”, type a name and click the “New Branch” button, you would create a new branch.

a. What would happen to the files in your working directory?

```
# The changes locally overwrite the state of the working directory, of the
# new branch you have just switched to. Your master branch state is unchanged
# and can be restored by git checkout master.
```

b. What would happens in the remote repo?

```
# The remote repo also is not affected.

# If you create a "new branch", it shows up in the remote repo

# Git branches are designed to be a fail-safe mechanism for integrating code
# and sharing changes between repositories.
```

c. What changes, if anything?

```
# the branch name (for your new branch)

# You have the option to bring your in-progress work over to the newly created
# branch. You can also stash your in-progress work on the main branch, leaving
# it unaffected.
```

d. Why would you want to work on a different branch?

```
# As you or others make changes, you can create a different or new branch with
# a unique name to separate your or everyones unique changes.

# Git's branching functionality lets you create new branches of a project to
# test ideas, isolate new features, or experiment without impacting the main
# project.

# answer source: https://thenewstack.io/dont-mess-with-the-master-working-with-branches-in-git-and-gith
```

vii. If you created a new branch in the previous step, nice! Experimentation develops your skills and understanding. Now, make sure you are on master as you continue your homework.

2 2 Fun with dplyr

2.1 2.1 Debugging mindset (5 points)

1. Why does this code doesn't work?

```
# the code worked for me after respelling 'my_varıable' to 'my_variable'

my_variable <- 10
my_variable
```

```
## [1] 10
## Error in eval(expr, envir, enclos): object 'my_variable' not found
```

2. Fix the following code so it works

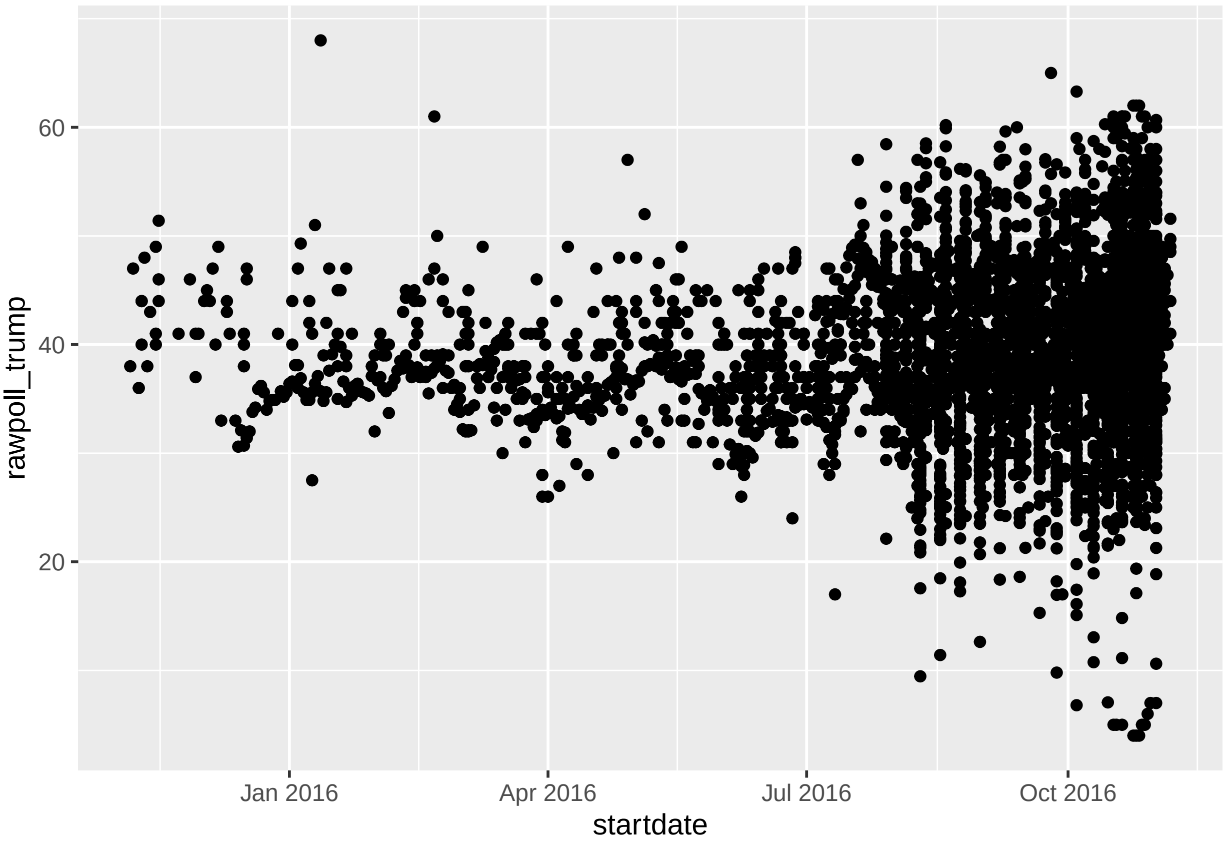
```
# library(dsblabs)

# ggplot(dota = polls_us_election_2016) +
# geom_point(mapping = aes(x = startdate,
# y = rawpoll_trum))
# fliter(polls_us_election_2016, cyl = "Florida")
# filter(diamond, carat > 3)

# FIXED CODE IS BELOW:

library(dslabs)
view(polls_us_election_2016)
view(diamonds)

ggplot(data = polls_us_election_2016) +
  geom_point(mapping = aes(x = startdate,y = rawpoll_trump))
```



```
filter(polls_us_election_2016, state == "Florida") %>% head(10)
```

```
##       state  startdate    enddate                    pollster grade samplesize
## 1  Florida 2016-11-03 2016-11-06      Quinnipiac University    A-         884
## 2  Florida 2016-11-01 2016-11-02                 Remington  <NA>        2352
## 3  Florida 2016-11-02 2016-11-04                    YouGov     B        1188
## 4  Florida 2016-10-20 2016-10-24                 SurveyUSA     A        1251
## 5  Florida 2016-11-01 2016-11-07              SurveyMonkey    C-        4092
## 6  Florida 2016-10-27 2016-11-01 CNN/Opinion Research Corp.    A-         773
## 7  Florida 2016-11-01 2016-11-02          Gravis Marketing    B-        1220
## 8  Florida 2016-11-06 2016-11-06           Trafalgar Group     C        1100
## 9  Florida 2016-10-25 2016-10-27             Siena College     A         815
## 10 Florida 2016-10-25 2016-10-26            Marist College     A         779
##    population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1          lv           46.00         45.00            2.00               NA
## 2          lv           45.00         48.00              NA               NA
## 3          rv           45.00         45.00              NA               NA
## 4          lv           48.00         45.00            2.00               NA
## 5          lv           47.00         45.00            4.00               NA
## 6          lv           49.00         47.00            3.00               NA
## 7          rv           46.00         45.00            4.00               NA
## 8          lv           46.13         49.72            2.43               NA
## 9          lv           42.00         46.00            4.00               NA
## 10         lv           45.00         44.00            5.00               NA
##    adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1         46.44315      43.93999        2.098310               NA
## 2         44.85722      46.49677              NA               NA
## 3         47.07455      46.99468              NA               NA
## 4         46.74555      45.86589        1.520730               NA
## 5         45.59190      44.32744        1.692430               NA
## 6         48.35252      45.23579        2.469063               NA
## 7         46.02363      44.52199        4.647916               NA
## 8         45.75904      46.82230        3.495849               NA
## 9         42.40145      48.60084        2.457160               NA
## 10        43.82500      45.76098        2.980521               NA
```

```
filter(diamonds, carat > 3) %>% head(1)
```

```
## # A tibble: 1 x 10
##   carat cut     color clarity depth table price     x     y     z
##   <dbl> <ord>   <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  3.01 Premium I     I1       62.7    58  8040   9.1  8.97  5.67
```

3. Press Alt (Option) + Shift + K. What happens? How can you get to the same place using the menus?

```
# Pressing Alt (Option) + Shift + K in RStudio will pull up the Keyboard
# Shortcut quick reference popup.

# You can also pull up the keyboard shortcut quick reference popup using the
# menus by clicking 'Tools' tab, then selecting 'Keyboard Shortcuts Help'
```

2.2 2.2 Filter (15 points)

1. Using the polls_us_election_2016 data frame in the dslabspackage find the following
2. Polls for the states of Hawaii and Alaska

```r
# we use the "|" sign to include multiple criteria together in our filter
filter(polls_us_election_2016, state == "Hawaii" | state == "Alaska") %>%
  head(10)
```

```
##     state  startdate    enddate                pollster grade samplesize
## 1  Alaska 2016-11-03 2016-11-06        Gravis Marketing    B-        617
## 2  Hawaii 2016-11-01 2016-11-07            SurveyMonkey    C-        426
## 3  Alaska 2016-11-01 2016-11-07            SurveyMonkey    C-        409
## 4  Alaska 2016-10-25 2016-10-27 Google Consumer Surveys     B        446
## 5  Alaska 2016-10-21 2016-10-26        Craciun Research  <NA>        400
## 6  Alaska 2016-10-11 2016-10-13  Lake Research Partners    B+        500
## 7  Alaska 2016-10-05 2016-10-06       Moore Information     B        500
## 8  Hawaii 2016-10-30 2016-11-06            SurveyMonkey    C-        426
## 9  Alaska 2016-10-30 2016-11-06            SurveyMonkey    C-        382
## 10 Hawaii 2016-10-04 2016-11-06                  YouGov     B        289
##    population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1          rv            41.0          44.0             3.0               NA
## 2          lv            52.0          28.0             9.0               NA
## 3          lv            31.0          48.0            12.0               NA
## 4          lv            38.0          39.0            11.0               NA
## 5          lv            47.0          43.0             7.0               NA
## 6          lv            36.0          37.0             7.0               NA
## 7          lv            34.0          37.0            10.0               NA
## 8          lv            52.0          29.0             9.0               NA
## 9          lv            31.0          47.0            13.0               NA
## 10         lv            50.3          27.9             4.1               NA
##    adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1         40.84795      43.33498        3.726098               NA
## 2         50.61398      27.33769        6.692430               NA
## 3         29.60705      47.33447        9.692430               NA
## 4         43.43333      46.13872        9.828628               NA
## 5         44.75680      44.77065        6.616653               NA
## 6         36.93134      42.37487        5.751367               NA
## 7         36.22323      41.11790        8.260216               NA
## 8         50.68395      28.39450        6.677361               NA
## 9         29.67449      46.38923       10.677360               NA
## 10        51.16296      30.58176        3.811562               NA
```

2. Polls with sample sizes bigger than 500 people

```r
filter(polls_us_election_2016, samplesize > 500) %>% head(10)
```

```
##         state  startdate    enddate
## 1        U.S. 2016-11-03 2016-11-06
## 2        U.S. 2016-11-01 2016-11-07
## 3        U.S. 2016-11-02 2016-11-06
## 4        U.S. 2016-11-04 2016-11-07
## 5        U.S. 2016-11-03 2016-11-06
## 6        U.S. 2016-11-03 2016-11-06
## 7        U.S. 2016-11-02 2016-11-06
## 8        U.S. 2016-11-03 2016-11-05
## 9  New Mexico 2016-11-06 2016-11-06
## 10       U.S. 2016-11-04 2016-11-07
##                                                       pollster grade samplesize
## 1                                     ABC News/Washington Post    A+       2220
## 2                                      Google Consumer Surveys     B      26574
## 3                                                        Ipsos    A-       2195
## 4                                                       YouGov     B       3677
## 5                                             Gravis Marketing    B-      16639
## 6  Fox News/Anderson Robbins Research/Shaw & Company Research     A       1295
## 7                                      CBS News/New York Times    A-       1426
## 8                                 NBC News/Wall Street Journal    A-       1282
## 9                                                     Zia Poll  <NA>       8439
## 10                                                    IBD/TIPP    A-       1107
##    population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1          lv           47.00         43.00            4.00               NA
## 2          lv           38.03         35.69            5.46               NA
## 3          lv           42.00         39.00            6.00               NA
## 4          lv           45.00         41.00            5.00               NA
## 5          rv           47.00         43.00            3.00               NA
## 6          lv           48.00         44.00            3.00               NA
## 7          lv           45.00         41.00            5.00               NA
## 8          lv           44.00         40.00            6.00               NA
## 9          lv           46.00         44.00            6.00               NA
## 10         lv           41.20         42.70            7.10               NA
##    adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1         45.20163      41.72430        4.626221               NA
## 2         43.34557      41.21439        5.175792               NA
## 3         42.02638      38.81620        6.844734               NA
## 4         45.65676      40.92004        6.069454               NA
## 5         46.84089      42.33184        3.726098               NA
## 6         49.02208      43.95631        3.057876               NA
## 7         45.11649      40.92722        4.341786               NA
## 8         43.58576      40.77325        5.365788               NA
## 9         44.82594      41.59978        7.870127               NA
## 10        42.92745      42.23545        6.316175               NA
```

3. Polls run by YouGov, Google Consumer Surveys or SurveyMonkey

```r
# library(dplyr)
# target <- c("Ticker1", "Ticker2", "Ticker3")
# filter(df, Ticker %in% target)

# code source: https://stackoverflow.com/questions/50687466/filtering-column-by-multiple-values

filter_polls <- c("YouGov", "Google Consumer Surveys", "SurveyMonkey")
filter(polls_us_election_2016, pollster %in% filter_polls ) %>% head(10)
```

```
##           state  startdate    enddate                pollster grade samplesize
## 1          U.S. 2016-11-01 2016-11-07 Google Consumer Surveys     B      26574
## 2          U.S. 2016-11-04 2016-11-07                  YouGov     B       3677
## 3          Ohio 2016-11-02 2016-11-04                  YouGov     B       1189
## 4       Georgia 2016-11-03 2016-11-05                  YouGov     B        995
## 5  Pennsylvania 2016-11-03 2016-11-05                  YouGov     B        931
## 6       Florida 2016-11-02 2016-11-04                  YouGov     B       1188
## 7          U.S. 2016-10-31 2016-11-06            SurveyMonkey    C-      70194
## 8    California 2016-10-25 2016-10-31                  YouGov     B       1498
## 9       Florida 2016-11-01 2016-11-07            SurveyMonkey    C-       4092
## 10         Utah 2016-11-03 2016-11-05                  YouGov     B        762
##    population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1          lv           38.03         35.69            5.46               NA
## 2          lv           45.00         41.00            5.00               NA
## 3          lv           45.00         46.00              NA               NA
## 4          lv           43.00         49.00            4.00               NA
## 5          lv           45.00         43.00            4.00               NA
## 6          rv           45.00         45.00              NA               NA
## 7          lv           47.00         41.00            6.00               NA
## 8          lv           53.00         33.00            4.00               NA
## 9          lv           47.00         45.00            4.00               NA
## 10         lv           23.00         40.00            7.00               24
##    adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1         43.34557      41.21439        5.175792               NA
## 2         45.65676      40.92004        6.069454               NA
## 3         44.93624      45.09646              NA               NA
## 4         43.80799      48.99024        5.169494               NA
## 5         45.82043      42.99602        5.169494               NA
## 6         47.07455      46.99468              NA               NA
## 7         45.65592      40.37888        3.677361               NA
## 8         53.93975      34.06845        4.511946               NA
## 9         45.59190      44.32744        1.692430               NA
## 10        23.83396      40.00230        8.169495               24
```

4. Polls with A+ grade and percentage for Trump less than 30

```
filter(polls_us_election_2016, grade == "A+", rawpoll_trump < 30)
```

```
##        state  startdate    enddate                                pollster
## 1   Maryland 2016-09-27 2016-09-30                ABC News/Washington Post
## 2 Washington 2016-08-09 2016-08-13                          Elway Research
## 3 California 2016-06-08 2016-07-02 Field Research Corporation (Field Poll)
## 4   Maryland 2016-03-30 2016-04-03                ABC News/Washington Post
##   grade samplesize population rawpoll_clinton rawpoll_trump rawpoll_johnson
## 1    A+        706         lv              63            27               4
## 2    A+        350         lv              45            24              NA
## 3    A+        495         lv              50            26              10
## 4    A+        752         rv              63            28              NA
##   rawpoll_mcmullin adjpoll_clinton adjpoll_trump adjpoll_johnson
## 1               NA        62.70862      28.73903        1.469119
## 2               NA        44.68377      27.79350              NA
## 3               NA        52.61833      29.79511        5.842896
## 4               NA        60.84333      30.40937              NA
##   adjpoll_mcmullin
## 1               NA
## 2               NA
## 3               NA
## 4               NA
```

5. Polls where the adjusted percentage for Clinton is between 40 and 60 percent (inclusive)

```r
filter(polls_us_election_2016, adjpoll_clinton < 60 & adjpoll_clinton > 40) %>%
  head(10)
```

```
##         state  startdate    enddate
## 1        U.S. 2016-11-03 2016-11-06
## 2        U.S. 2016-11-01 2016-11-07
## 3        U.S. 2016-11-02 2016-11-06
## 4        U.S. 2016-11-04 2016-11-07
## 5        U.S. 2016-11-03 2016-11-06
## 6        U.S. 2016-11-03 2016-11-06
## 7        U.S. 2016-11-02 2016-11-06
## 8        U.S. 2016-11-03 2016-11-05
## 9  New Mexico 2016-11-06 2016-11-06
## 10       U.S. 2016-11-04 2016-11-07
##                                                       pollster grade samplesize
## 1                                     ABC News/Washington Post    A+       2220
## 2                                      Google Consumer Surveys     B      26574
## 3                                                        Ipsos    A-       2195
## 4                                                       YouGov     B       3677
## 5                                             Gravis Marketing    B-      16639
## 6  Fox News/Anderson Robbins Research/Shaw & Company Research     A       1295
## 7                                      CBS News/New York Times    A-       1426
## 8                                 NBC News/Wall Street Journal    A-       1282
## 9                                                     Zia Poll  <NA>       8439
## 10                                                    IBD/TIPP    A-       1107
##    population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1          lv           47.00         43.00            4.00               NA
## 2          lv           38.03         35.69            5.46               NA
## 3          lv           42.00         39.00            6.00               NA
## 4          lv           45.00         41.00            5.00               NA
## 5          rv           47.00         43.00            3.00               NA
## 6          lv           48.00         44.00            3.00               NA
## 7          lv           45.00         41.00            5.00               NA
## 8          lv           44.00         40.00            6.00               NA
## 9          lv           46.00         44.00            6.00               NA
## 10         lv           41.20         42.70            7.10               NA
##    adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1         45.20163      41.72430        4.626221               NA
## 2         43.34557      41.21439        5.175792               NA
## 3         42.02638      38.81620        6.844734               NA
## 4         45.65676      40.92004        6.069454               NA
## 5         46.84089      42.33184        3.726098               NA
## 6         49.02208      43.95631        3.057876               NA
## 7         45.11649      40.92722        4.341786               NA
## 8         43.58576      40.77325        5.365788               NA
## 9         44.82594      41.59978        7.870127               NA
## 10        42.92745      42.23545        6.316175               NA
```

6. Polls where Trump raw winning margin over Clinton is bigger than 10%

```r
filter(polls_us_election_2016, (rawpoll_trump - rawpoll_clinton) > 10) %>%
  head(10)
```

```
##       state  startdate    enddate              pollster grade samplesize
## 1  Missouri 2016-10-31 2016-11-01 Public Policy Polling    B+       1083
## 2  Missouri 2016-11-01 2016-11-02 Clarity Campaign Labs     B       1036
## 3  Missouri 2016-10-31 2016-11-01             Remington  <NA>       1722
## 4      Utah 2016-11-03 2016-11-05                YouGov     B        762
## 5      Utah 2016-10-31 2016-11-02       Emerson College     B       1000
## 6      Utah 2016-11-03 2016-11-05       Trafalgar Group     C       1350
## 7   Indiana 2016-11-01 2016-11-07          SurveyMonkey    C-       1700
## 8  Missouri 2016-10-27 2016-10-28         BK Strategies  <NA>       1698
## 9    Kansas 2016-10-26 2016-10-30             SurveyUSA     A        624
## 10 Kentucky 2016-11-01 2016-11-07          SurveyMonkey    C-       1315
##    population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1          lv           37.00         50.00            4.00               NA
## 2          lv           38.00         54.00              NA               NA
## 3          lv           39.00         52.00            4.00               NA
## 4          lv           23.00         40.00            7.00            24.00
## 5          lv           19.60         39.80            3.00            27.60
## 6          lv           29.52         39.95            3.89            24.52
## 7          lv           35.00         52.00           10.00               NA
## 8          lv           39.00         53.00              NA               NA
## 9          lv           38.00         49.00            7.00               NA
## 10         lv           35.00         54.00            6.00               NA
##    adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1         36.55124      49.59908        5.028656               NA
## 2         37.65693      52.77196              NA               NA
## 3         39.02343      50.77483        5.124639               NA
## 4         23.83396      40.00230        8.169495         24.00000
## 5         19.75684      38.01440        3.075563         27.70142
## 6         29.32624      37.13456        5.055889         24.52000
## 7         33.58643      51.32490        7.692430               NA
## 8         38.35085      51.20917              NA               NA
## 9         36.85588      48.73752        7.213034               NA
## 10        33.58942      53.32629        3.692430               NA
```

7. Polls where McMullin percentage is more than 5 %.

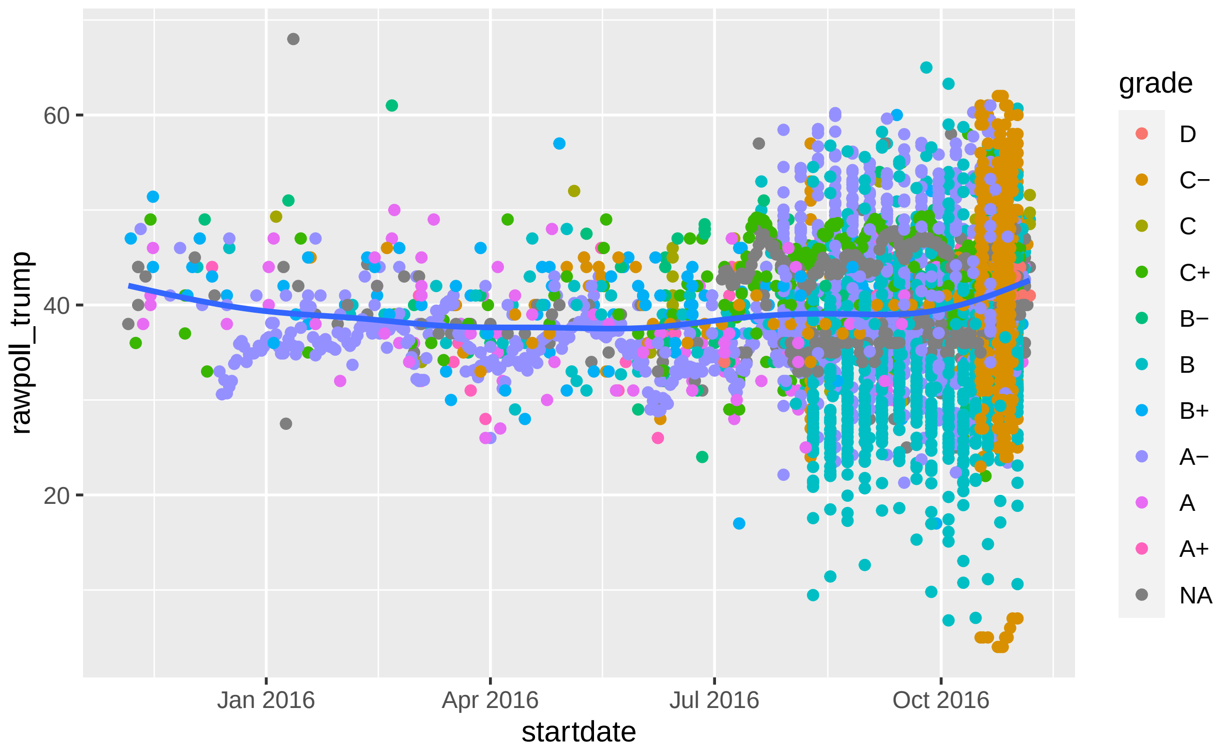
```
filter(polls_us_election_2016, rawpoll_mcmullin > 5, adjpoll_mcmullin > 5) %>% head(10)
```

```
##    state  startdate    enddate                                 pollster grade
## 1   Utah 2016-11-03 2016-11-05                                   YouGov     B
## 2   Utah 2016-10-31 2016-11-02                          Emerson College     B
## 3   Utah 2016-10-30 2016-10-31                         Gravis Marketing    B-
## 4   Utah 2016-11-03 2016-11-05                          Trafalgar Group     C
## 5   Utah 2016-11-01 2016-11-07                             SurveyMonkey    C-
## 6   Utah 2016-10-30 2016-11-02                      Monmouth University    A+
## 7   Utah 2016-10-29 2016-10-31 Rasmussen Reports/Pulse Opinion Research    C+
## 8   Utah 2016-11-01 2016-11-03                             Y2 Analytics    C+
## 9   Utah 2016-10-20 2016-10-27                   Dan Jones & Associates    C+
## 10  Utah 2016-10-10 2016-10-12                      Monmouth University    A+
##    samplesize population rawpoll_clinton rawpoll_trump rawpoll_johnson
## 1         762         lv           23.00         40.00            7.00
## 2        1000         lv           19.60         39.80            3.00
## 3        1424         rv           29.00         35.00            3.00
## 4        1350         lv           29.52         39.95            3.89
## 5        1479         lv           31.00         34.00            7.00
## 6         402         lv           31.00         37.00            4.00
## 7         750         lv           31.00         42.00            3.00
## 8         500         lv           24.00         33.00            5.00
## 9         823         lv           24.00         32.00            4.00
## 10        403         lv           28.00         34.00            9.00
##    rawpoll_mcmullin adjpoll_clinton adjpoll_trump adjpoll_johnson
## 1             24.00        23.83396      40.00230        8.169495
## 2             27.60        19.75684      38.01440        3.075563
## 3             24.00        29.04086      34.78405        3.422875
## 4             24.52        29.32624      37.13456        5.055889
## 5             25.00        29.59989      33.33115        4.692430
## 6             24.00        30.06568      36.70382        4.644697
## 7             21.00        31.67657      42.74799        2.965430
## 8             28.00        25.34813      35.57386        4.569504
## 9             30.00        23.49263      33.38083        2.880316
## 10            20.00        27.24268      36.62770        7.310543
##    adjpoll_mcmullin
## 1          24.00000
## 2          27.70142
## 3          24.13522
## 4          24.52000
## 5          25.00000
## 6          24.10142
## 7          21.16903
## 8          28.06761
## 9          30.37186
## 10         20.81133
```

2. Remember this graph from last problem set?

```
ggplot(data = polls_us_election_2016,
       mapping = aes(x = startdate,
                     y = rawpoll_trump)) +
    geom_point(aes(color = grade)) +
    geom_smooth(se = FALSE)
```

```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



There is a poll after January where Trump percentage is above 60 with no grade. What poll is that one?

```
# code source found in 16.1.1 Poll Data: https://rafalab.github.io/dsbook/models.html
polls_us_election_2016 %>%

filter(startdate > 2016-01-31, rawpoll_trump > 60, is.na(grade))
```

```
##     state  startdate    enddate          pollster grade samplesize population
## 1 Alabama 2016-01-12 2016-01-12 Strategy Research  <NA>       2700         rv
##   rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1              32            68              NA               NA
##   adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1        30.92318      67.31449              NA               NA
```

1. Common bugs: You want to see missing values in a dataframe. You run the following code and get no results. Why is wrong?

```
filter(polls_us_election_2016, grade == NA)
```

```
##  [1] state            startdate        enddate          pollster
##  [5] grade            samplesize       population       rawpoll_clinton
##  [9] rawpoll_trump    rawpoll_johnson  rawpoll_mcmullin adjpoll_clinton
## [13] adjpoll_trump    adjpoll_johnson  adjpoll_mcmullin
## <0 rows> (or 0-length row.names)
```

```
# we should use 'is.na' to check whether expression evaluates to NA
# such as the code below
filter(polls_us_election_2016, is.na(grade)) %>%
  head(5)
```

```
##        state  startdate    enddate                 pollster grade samplesize
## 1 New Mexico 2016-11-06 2016-11-06                 Zia Poll  <NA>       8439
## 2       U.S. 2016-11-05 2016-11-07 The Times-Picayune/Lucid  <NA>       2521
## 3       U.S. 2016-11-01 2016-11-07    USC Dornsife/LA Times  <NA>       2972
## 4   Virginia 2016-11-01 2016-11-02                Remington  <NA>       3076
## 5  Wisconsin 2016-11-01 2016-11-02                Remington  <NA>       2720
##   population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1         lv           46.00         44.00               6               NA
## 2         lv           45.00         40.00               5               NA
## 3         lv           43.61         46.84              NA               NA
## 4         lv           46.00         44.00              NA               NA
## 5         lv           49.00         41.00              NA               NA
##   adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1        44.82594      41.59978        7.870127               NA
## 2        45.13966      42.26495        3.679914               NA
## 3        45.32156      43.38579              NA               NA
## 4        45.27399      41.91459              NA               NA
## 5        48.22713      38.86464              NA               NA
```

2. How many polls have missing rawpoll_mcmullin? Why do you think this is happening?

```
filter(polls_us_election_2016, is.na(rawpoll_mcmullin)) %>%
  head(10)
```

```
##         state  startdate    enddate
## 1        U.S. 2016-11-03 2016-11-06
## 2        U.S. 2016-11-01 2016-11-07
## 3        U.S. 2016-11-02 2016-11-06
## 4        U.S. 2016-11-04 2016-11-07
## 5        U.S. 2016-11-03 2016-11-06
## 6        U.S. 2016-11-03 2016-11-06
## 7        U.S. 2016-11-02 2016-11-06
## 8        U.S. 2016-11-03 2016-11-05
## 9  New Mexico 2016-11-06 2016-11-06
## 10       U.S. 2016-11-04 2016-11-07
##                                                      pollster grade samplesize
## 1                                    ABC News/Washington Post    A+       2220
## 2                                     Google Consumer Surveys     B      26574
## 3                                                       Ipsos    A-       2195
## 4                                                      YouGov     B       3677
## 5                                            Gravis Marketing    B-      16639
## 6  Fox News/Anderson Robbins Research/Shaw & Company Research     A       1295
## 7                                     CBS News/New York Times    A-       1426
## 8                                NBC News/Wall Street Journal    A-       1282
## 9                                                    Zia Poll  <NA>       8439
## 10                                                   IBD/TIPP    A-       1107
##    population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1          lv           47.00         43.00            4.00               NA
## 2          lv           38.03         35.69            5.46               NA
## 3          lv           42.00         39.00            6.00               NA
## 4          lv           45.00         41.00            5.00               NA
## 5          rv           47.00         43.00            3.00               NA
## 6          lv           48.00         44.00            3.00               NA
## 7          lv           45.00         41.00            5.00               NA
## 8          lv           44.00         40.00            6.00               NA
## 9          lv           46.00         44.00            6.00               NA
## 10         lv           41.20         42.70            7.10               NA
##    adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1         45.20163      41.72430        4.626221               NA
## 2         43.34557      41.21439        5.175792               NA
## 3         42.02638      38.81620        6.844734               NA
## 4         45.65676      40.92004        6.069454               NA
## 5         46.84089      42.33184        3.726098               NA
## 6         49.02208      43.95631        3.057876               NA
## 7         45.11649      40.92722        4.341786               NA
## 8         43.58576      40.77325        5.365788               NA
## 9         44.82594      41.59978        7.870127               NA
## 10        42.92745      42.23545        6.316175               NA
```

```
# there are 4,178 observations with polls missing rawpoll_mcmullin.
# This large number of polls having missing mcmullin could be because he is an
# independent party candidate, and minor party candidates are often left out of
# poll questions.
# This could also be because mcmullin was not projected to be included on very
# many state ballots.
```

3. What happens to observations with missing values in grade when you filter by grade == “A”? Why? (Hint: compare “C minus” == “A”, “A” == “A”, and NA == “A”).

```
# observations with missing values in grade will be excluded when we apply
# filter by grade == "A". This is because when we apply filter, we create a
# subset based on a specific criteria. Any observations within grade that is
# not equal to "A" is excluded, missing values and all.

filter(polls_us_election_2016, grade == "A") %>% head(10)
```

```
##              state  startdate    enddate
## 1             U.S. 2016-11-03 2016-11-06
## 2             U.S. 2016-11-01 2016-11-03
## 3        Wisconsin 2016-10-26 2016-10-31
## 4   North Carolina 2016-11-04 2016-11-06
## 5          Florida 2016-10-20 2016-10-24
## 6         New York 2016-11-03 2016-11-04
## 7          Arizona 2016-10-30 2016-11-01
## 8       Washington 2016-10-31 2016-11-02
## 9          Georgia 2016-10-30 2016-11-01
## 10      California 2016-10-28 2016-10-31
##                                                       pollster grade samplesize
## 1  Fox News/Anderson Robbins Research/Shaw & Company Research     A       1295
## 2                                               Marist College     A        940
## 3                                         Marquette University     A       1255
## 4                                                Siena College     A        800
## 5                                                    SurveyUSA     A       1251
## 6                                                Siena College     A        617
## 7                                               Marist College     A        719
## 8                                                    SurveyUSA     A        681
## 9                                              Marist College     A        707
## 10                                                  SurveyUSA     A        747
##    population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1          lv              48            44               3               NA
## 2          lv              44            43               6               NA
## 3          lv              46            40               4               NA
## 4          lv              44            44               3               NA
## 5          lv              48            45               2               NA
## 6          lv              51            34               5               NA
## 7          lv              40            45               9               NA
## 8          lv              50            38               4               NA
## 9          lv              44            45               8               NA
## 10         lv              56            35               4               NA
##    adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1         49.02208      43.95631        3.057876               NA
## 2         42.83406      43.43819        4.780429               NA
## 3         46.10344      40.97982        2.897062               NA
## 4         44.21875      45.08290        2.335250               NA
## 5         46.74555      45.86589        1.520730               NA
## 6         51.30942      35.12664        4.344325               NA
## 7         38.85863      45.72376        7.587354               NA
## 8         48.87802      36.95385        4.691579               NA
## 9         42.81871      45.65459        6.587354               NA
## 10        54.86451      34.27885        4.457467               NA
```

4. What does NA | TRUE evaluate to? Why?

```
NA | TRUE
```

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

```
# evaluates as TRUE because the outcome is not ambiguous. using | it looks to
# determine a result of NA (ambiguous) OR TRUE (not ambiguous). We know it will
# evaluate as TRUE since NA is a valid logical object. Where a component of x
# or y is NA, the result will be NA if the outcome is ambiguous.
```

5. What does FALSE & NA evaluate to? Why?

```
FALSE & NA
```

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

```
# evaluates as FALSE also because wherever a component of x or y is NA, the
# result will be NA if the outcome is ambiguous. Here the outcome is not
# ambigous (FALSE). It reads as FALSE AND NA.
```

2.3 2.3 Select (10 points)

1. What happens if you include the name of a variable multiple times in a select() call?

```
# code found in Exercise 5.4.2: https://jrnold.github.io/r4ds-exercise-solutions/transform.html

# The select() call ignores the duplication. Any duplicated variables are only
# included once, in the first location they appear. The select() function does
# not raise an error or warning or print any message if there are duplicated
# variables.
library("tidyverse")

select(polls_us_election_2016, grade, pollster, grade, pollster, startdate) %>%
  head(10)
```

2. Typically, R is case sensitive, but select helpers ignore case by default. Change the default to return an empty tibble.

```
# select(polls_us_election_2016, contains("RAW")))

# we include ignore.case = FALSE To change the default behavior
select(polls_us_election_2016, contains("RAW", ignore.case = FALSE))
```

```
## data frame with 0 columns and 4208 rows
```

```
# code source from Exercise 5.4.4: https://jrnold.github.io/r4ds-exercise-solutions/transform.html
```

3. Brainstorm as many distinct ways as possible to select rawpoll_clinton, rawpoll_trump, adjpoll_clinton, and adjpoll_trump

```
# select() then using contains() within it
select(polls_us_election_2016, contains("clinton") | contains("trump")) %>%
  head(10)
```

```
##    rawpoll_clinton adjpoll_clinton rawpoll_trump adjpoll_trump
## 1            47.00        45.20163         43.00      41.72430
## 2            38.03        43.34557         35.69      41.21439
## 3            42.00        42.02638         39.00      38.81620
## 4            45.00        45.65676         41.00      40.92004
## 5            47.00        46.84089         43.00      42.33184
## 6            48.00        49.02208         44.00      43.95631
## 7            45.00        45.11649         41.00      40.92722
## 8            44.00        43.58576         40.00      40.77325
## 9            46.00        44.82594         44.00      41.59978
## 10           41.20        42.92745         42.70      42.23545
```

```
# specifically selecting variables with 'clinton' or 'trump'
select(polls_us_election_2016, rawpoll_clinton, rawpoll_trump, adjpoll_clinton,
       adjpoll_trump) %>% head(10)
```

```
##    rawpoll_clinton rawpoll_trump adjpoll_clinton adjpoll_trump
## 1            47.00         43.00        45.20163      41.72430
## 2            38.03         35.69        43.34557      41.21439
## 3            42.00         39.00        42.02638      38.81620
## 4              45.00         41.00        45.65676      40.92004
## 5              47.00         43.00        46.84089      42.33184
## 6              48.00         44.00        49.02208      43.95631
## 7              45.00         41.00        45.11649      40.92722
## 8              44.00         40.00        43.58576      40.77325
## 9              46.00         44.00        44.82594      41.59978
## 10             41.20         42.70        42.92745      42.23545
```

```
# using select() then using ends_with() in it
select(polls_us_election_2016, ends_with("clinton") | ends_with("trump")) %>%
  head(10)
```

```
##    rawpoll_clinton adjpoll_clinton rawpoll_trump adjpoll_trump
## 1            47.00        45.20163         43.00      41.72430
## 2            38.03        43.34557         35.69      41.21439
## 3            42.00        42.02638         39.00      38.81620
## 4            45.00        45.65676         41.00      40.92004
## 5            47.00        46.84089         43.00      42.33184
## 6            48.00        49.02208         44.00      43.95631
## 7            45.00        45.11649         41.00      40.92722
## 8            44.00        43.58576         40.00      40.77325
## 9            46.00        44.82594         44.00      41.59978
## 10           41.20        42.92745         42.70      42.23545
```

3 Arrange (5 pts)

1. Sort polls to find the ones where Clintons percentage is the highest. Use %>% head(1) to print just one row.

```
arrange(polls_us_election_2016, desc(rawpoll_clinton)) %>% head(1)
```

```
##                  state  startdate    enddate     pollster grade samplesize
## 1 District of Columbia 2016-10-30 2016-11-06 SurveyMonkey    C-        315
##   population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1         lv              88             7               2               NA
##   adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1        86.70544      6.406481      -0.3226397               NA
```

```
# we could similarly complete the same sort for highest % in adjpoll_clinton.
# arrange(polls_us_election_2016, desc(adjpoll_clinton))

# code source from Exercise 5.3.2: https://jrnold.github.io/r4ds-exercise-solutions/transform.html#arra
```

2. Find the 5 polls that interview more people. Only show pollster, samplesize, population, and grade.

```
Five_polls <- arrange(polls_us_election_2016, desc(samplesize)) %>% head(5)

select(Five_polls, pollster, samplesize, population,
       grade) %>% head(10)
```

```
##       pollster samplesize population grade
## 1       YouGov      84292         lv     B
## 2 SurveyMonkey      70194         lv    C-
## 3 SurveyMonkey      40816         lv    C-
## 4 SurveyMonkey      32226         rv    C-
## 5 SurveyMonkey      32225         lv    C-
```

```
# code source from Exercise 5.3.4: https://jrnold.github.io/r4ds-exercise-solutions/transform.html#arra
```

3. How could you use arrange() to sort all missing values to the start? (Hint use is.na(), you can use any variable with missing values here)

```
arrange(polls_us_election_2016, desc(is.na(samplesize)), samplesize
        ) %>% head(10)
```

```
##            state  startdate    enddate                pollster grade samplesize
## 1       Illinois 2016-07-11 2016-07-12       Basswood Research    C+         NA
## 2        Wyoming 2016-10-04 2016-10-09 Google Consumer Surveys     B         35
## 3          Maine 2016-10-04 2016-10-09 Google Consumer Surveys     B         37
## 4  New Hampshire 2016-09-21 2016-09-26 Google Consumer Surveys     B         39
## 5         Hawaii 2016-09-14 2016-09-20 Google Consumer Surveys     B         42
## 6        Wyoming 2016-09-27 2016-10-03 Google Consumer Surveys     B         43
## 7   Rhode Island 2016-10-10 2016-10-14 Google Consumer Surveys     B         45
## 8        Vermont 2016-10-04 2016-10-09 Google Consumer Surveys     B         47
## 9   North Dakota 2016-09-27 2016-10-03 Google Consumer Surveys     B         49
## 10  Rhode Island 2016-09-14 2016-09-20 Google Consumer Surveys     B         50
##    population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1          lv           46.40         32.50            5.20               NA
## 2          lv           11.04         52.74           14.56               NA
## 3          lv           42.82         41.51            4.64               NA
## 4          lv           40.36         38.63            4.68               NA
## 5          lv           43.20         31.59            2.79               NA
## 6          lv           18.80         35.07            9.70               NA
## 7          lv           57.33         13.05            2.26               NA
## 8          lv           57.02         16.11           10.26               NA
## 9          lv           22.78         45.48           12.00               NA
## 10         lv           46.40         30.90            8.84               NA
##    adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin
## 1         49.62926      36.34003       2.1205840               NA
## 2         17.06495      61.68032      11.7002400               NA
## 3         48.92751      50.77821       1.7802390               NA
## 4         48.03972      47.65478       1.0121100               NA
## 5         51.06284      40.50726      -1.1183940               NA
## 6         25.57183      43.87120       6.3307300               NA
## 7         63.03978      22.53819      -0.3066937               NA
## 8         63.13193      25.39574       7.4002390               NA
## 9         29.64619      54.47583       8.6307310               NA
## 10        54.48433      40.10019       4.9316060               NA
```

```
# interestingly enough Illinois Basswood Research is the only pollster with NA
# for samplesize.
# To put NA values first, we can add an indicator of whether the column has a
# missing value. Then we sort by the missing indicator column and the column of
# interest. For example, to sort the data frame by departure time (dep_time) in
# ascending order but NA values first, run the following.

# code source: Exercise 5.3.1: https://jrnold.github.io/r4ds-exercise-solutions/transform.html#arrange-
```

4 4 Mutate (15 pts)

1. Currently the variables rawpoll_candidate tells the percentage for that candidate. Convert each of them in number of people that pick that candidate. Store them in variables raw_candidate.

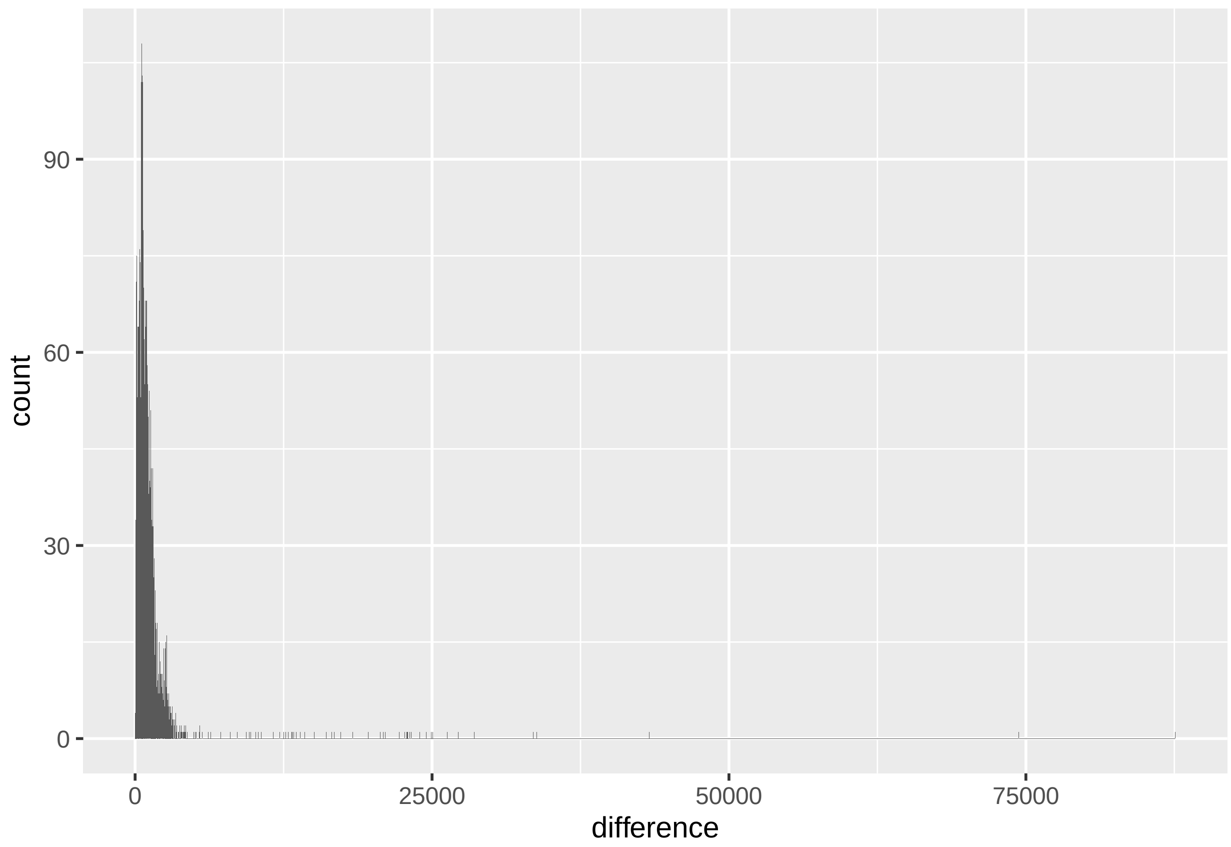
```
df_raw <- mutate(polls_us_election_2016,
       raw_clinton = round((rawpoll_clinton/100)*samplesize),
       raw_trump = round((rawpoll_trump/100)*samplesize),
       raw_johnson = round((rawpoll_johnson/100)*samplesize),
       raw_mcmullin = round((rawpoll_mcmullin/100)*samplesize)
       )
```

2. Make a plot to compare samplesize and the sum of your recently created variables for Trump and Clinton.

```
df_raw <- df_raw %>% mutate(difference = samplesize - (raw_trump - raw_clinton))
view(df_raw)

ggplot(data = df_raw, mapping = aes(x = difference)) +
  geom_histogram(binwidth = 25)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



3. We expect that samplesize - (raw_trump + raw_clinton) will be near zero for all polls. Why is this not happening?

```
# its not happening because there are many missing values in the data.
# also, this is not happening in some polls because there are voters who chose
# note to vote at all or chose either johnson or mcmullin in the samplesize.
```

4. How can you fix the problem in the last graph so that samplesize - (raw_trump - raw_clinton) is closer to zero?

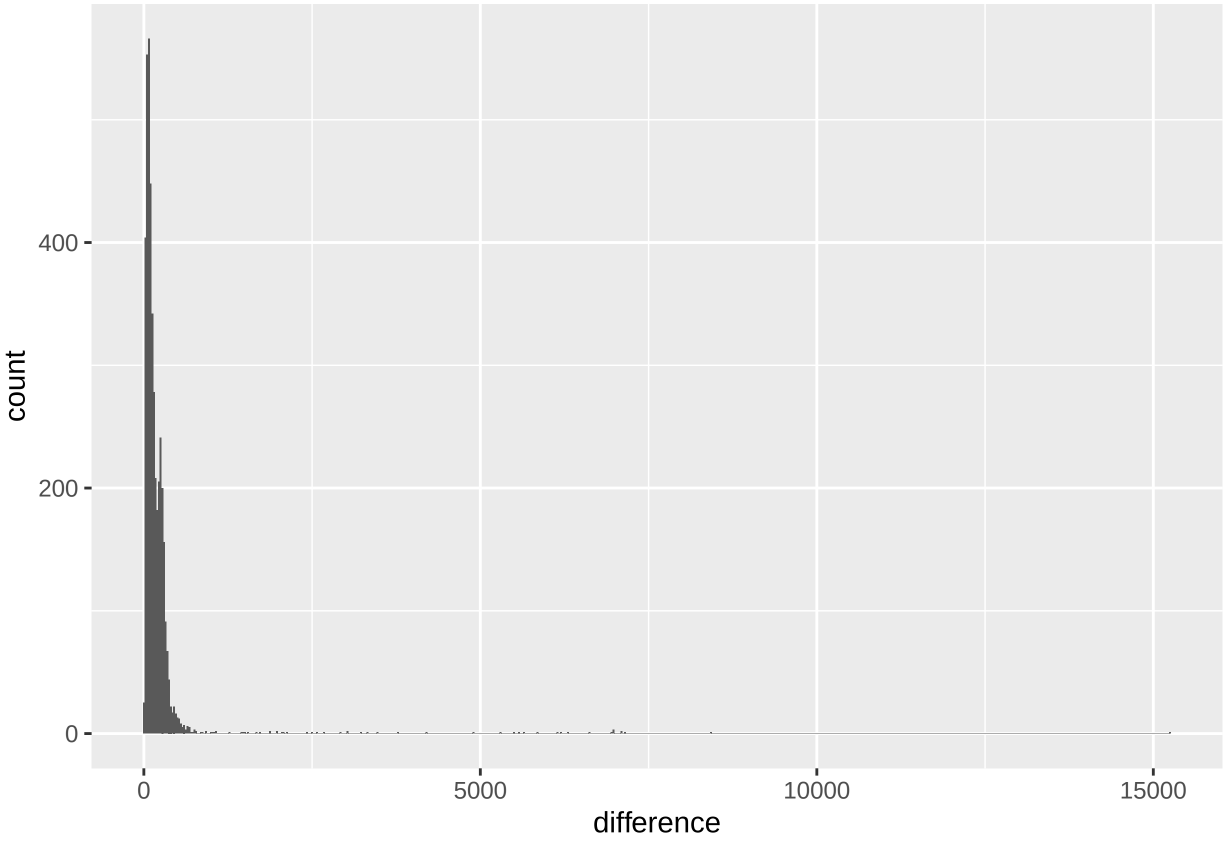
```
# subtract all other candidates, johnson and mcmullin OR
# switch equation for difference to be samplesize - (raw_clinton - raw_trump) OR
# switch equation for difference to be samplesize - (raw_clinton + raw_trump)

df_raw2 <- mutate(df_raw,
       raw_clinton = round((rawpoll_clinton/100)*samplesize),
       raw_trump = round((rawpoll_trump/100)*samplesize),
       raw_johnson = round((rawpoll_johnson/100)*samplesize),
       raw_mcmullin = round((rawpoll_mcmullin/100)*samplesize),
       difference = samplesize - (raw_clinton + raw_trump)
       )

view(df_raw2)
```

```r
ggplot(data = df_raw2, mapping = aes(x = difference)) +
  geom_histogram(binwidth = 25)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



5. Does your fix from the last question solve the problem? If not, discuss what might be happening and why we still see big differences between the sample size and the sum of all the quantities.

```r
# My fix from the last question slightly adjusts the graph where the difference
# is closer to zero on average, but there seem to be higher counts of non-zero
# values.

# Essentially the value samplesize - (raw_trump - raw_clinton) would be positive
# for any case when voters choose a candidate other than clinton or trump or
# are represented as part of the samplesize but not in any of the raw scores
# (raw_clinton, raw_trump, raw_johnson, raw_mcmullin)
```

6. Find the 7 polls where Johnson percentage is the highest using the ranking function. How do you want to handle ties? Carefully read the documentation for min_rank()

```r
polls_us_election_2016 %>%
  mutate(rank_order = min_rank(rawpoll_johnson)) %>%
  arrange(desc(rank_order)) %>%
  head(7)
```

```
##        state  startdate    enddate                 pollster grade samplesize
## 1 New Mexico 2016-08-09 2016-09-01             SurveyMonkey    C-       1788
## 2 New Mexico 2016-09-27 2016-09-29 Research & Polling, Inc.     A        501
## 3       Utah 2016-08-09 2016-09-01             SurveyMonkey    C-        722
## 4 New Mexico 2016-10-28 2016-11-01 The Times-Picayune/Lucid  <NA>        567
## 5 New Mexico 2016-10-17 2016-10-25             SurveyMonkey    C-       1095
## 6 New Mexico 2016-08-10 2016-08-16  Google Consumer Surveys     B        181
## 7 New Mexico 2016-10-18 2016-10-26             SurveyMonkey    C-       1175
##   population rawpoll_clinton rawpoll_trump rawpoll_johnson rawpoll_mcmullin
## 1         rv           37.00         29.00           25.00               NA
## 2         lv           35.00         31.00           24.00               NA
## 3         rv           27.00         34.00           23.00               NA
## 4         lv           39.00         31.00           22.00               NA
## 5         lv           41.00         34.00           21.00               NA
## 6         lv           37.35         25.97           20.47               NA
## 7         lv           42.00         34.00           20.00               NA
##   adjpoll_clinton adjpoll_trump adjpoll_johnson adjpoll_mcmullin rank_order
## 1        39.24093      34.76038        17.98822               NA       2799
## 2        37.93401      34.56285        19.36413               NA       2798
## 3        29.24107      39.76059        15.98822               NA       2797
## 4        39.44445      33.99644        20.36684               NA       2796
## 5        39.61370      35.80706        17.23450               NA       2795
## 6        44.56049      36.36966        16.27112               NA       2794
## 7        40.60988      35.65900        16.34258               NA       2793
```

5 5 Summarize (15 pts)

1. How many polls each company is running? Come up with at least two different ways to get to the same result

```
pollster_companies <- polls_us_election_2016 %>%
  group_by(pollster, grade) %>%
  summarise(n_polls = n()) %>% head(10)
```

```
## ‘summarise()‘ has grouped output by ’pollster’. You can override using the ‘.groups‘ argument.
```

```
polls_us_election_2016 %>%
  count(pollster) %>% head(10)
```

```
##                             pollster  n
## 1           ABC News/Washington Post 28
## 2            American Research Group  9
## 3                American Strategies  1
## 4                  Angus Reid Global  1
## 5      Anzalone Liszt Grove Research  2
## 6           Arizona State University  2
## 7   Associated Industries of Florida  3
## 8         Baldwin Wallace University  2
## 9              Ball State University  1
## 10                    Baruch College  1
```

```
pollster_companies
```

```
## # A tibble: 10 x 3
## # Groups:   pollster [10]
##    pollster                         grade n_polls
##    <fct>                            <fct>   <int>
##  1 ABC News/Washington Post         A+         28
##  2 American Research Group          C+          9
##  3 American Strategies              <NA>        1
##  4 Angus Reid Global                A-          1
##  5 Anzalone Liszt Grove Research    C           2
##  6 Arizona State University         C+          2
##  7 Associated Industries of Florida <NA>        3
##  8 Baldwin Wallace University       <NA>        2
##  9 Ball State University            <NA>        1
## 10 Baruch College                   B-          1
```

2. Calculate average poll size by grade for the state of Florida.

```
Florida_polls <- polls_us_election_2016 %>%
  filter(state == "Florida") %>%
  group_by(grade) %>%
  summarise(sample_mean = mean(samplesize))

Florida_polls
```

```
## # A tibble: 10 x 2
##    grade sample_mean
##    <fct>       <dbl>
##  1 C-          2290.
##  2 C            780.
##  3 C+           714.
##  4 B-          1322.
##  5 B           1285.
##  6 B+           754.
##  7 A-           825.
##  8 A           1221.
##  9 A+           536.
## 10 <NA>        1000.
```

3. Do all the polls by the same company have the same grade?

```
# Yes, all polls by the same company have the same grade. For example, pollster
# "Google Consumer Surveys" are all exclusively given grade "B".
```

4. Which state has the higher average vote for other (Mcmullin) candidates? Why do you think this the case?

```
# Utah has the highest average vote for mcmullin as a presidential candidate.
# This result could be partly due to Republican backlash against Trump following
# release of a controversial 2005 video showing Trump bragging about obscene
```

```r
# sexual conduct with women (source: http://www.deseretnews.com/article/865664606/Poll-Trump-falls-into

# mcmullin spent a significant amount of time campaigning to Utah voters
# he was positively accepted by voters and fit in with the political landscape
# in Utah in 2016 as a viable conservative candidate alternative to Donald Trump

mcmullin_state <- polls_us_election_2016 %>%
  group_by(state, population) %>%
  summarise(poll_mcmullin = mean(rawpoll_mcmullin, na.rm = TRUE)) %>%
  filter(state != "U.S")
```

```
## 'summarise()' has grouped output by 'state'. You can override using the '.groups' argument.
```

```r
arrange(mcmullin_state, desc(poll_mcmullin)) %>% head(1)
```

```
## # A tibble: 1 x 3
## # Groups:   state [1]
##   state population poll_mcmullin
##   <fct> <chr>              <dbl>
## 1 Utah  lv                  24.2
```

```r
mcmullin_state
```

```
## # A tibble: 137 x 3
## # Groups:   state [57]
##    state      population poll_mcmullin
##    <fct>      <chr>              <dbl>
##  1 Alabama    lv                   NaN
##  2 Alabama    rv                   NaN
##  3 Alaska     lv                   NaN
##  4 Alaska     rv                   NaN
##  5 Arizona    lv                   NaN
##  6 Arizona    rv                   NaN
##  7 Arizona    v                    NaN
##  8 Arkansas   lv                   NaN
##  9 Arkansas   rv                   NaN
## 10 California lv                   NaN
## # ... with 127 more rows
```

5. Find all the states were more than 10 different companies ran polls. Order the results from most polling companies to fewest.

```r
state_pollsters <- polls_us_election_2016 %>%
  group_by(state) %>%
  summarise(n_polls = n(),
            n_companies = n_distinct(pollster)) %>%
  filter(state != "U.S.", n_companies >= 10) %>%
  arrange(desc(n_companies))

state_pollsters
```

```
## # A tibble: 28 x 3
##    state          n_polls n_companies
##    <fct>            <int>       <int>
##  1 Florida            148          35
##  2 Pennsylvania       125          30
##  3 Ohio               115          28
##  4 North Carolina     125          27
##  5 Nevada              93          25
##  6 New Hampshire      112          24
##  7 Arizona             79          22
##  8 Colorado            80          21
##  9 Virginia            91          21
## 10 Michigan            86          20
## # ... with 18 more rows
```

6. Calculate the number of poll by grade and type of population. Add your results the max() and min() percentage for Trump that each type of poll has. Report your results so that better grades are at the top of your table.

```
polls_us_election_2016 %>%
  group_by(grade, population) %>%
  summarise(n_grade = n(),
            n_population = n(),
            max_trump = max(rawpoll_trump),
            min_trump = min(rawpoll_trump)) %>%
  arrange(desc(grade)) %>%
  head(10)
```

```
## `summarise()` has grouped output by 'grade'. You can override using the `.groups` argument.
```

```
## # A tibble: 10 x 6
## # Groups:   grade [4]
##    grade population n_grade n_population max_trump min_trump
##    <fct> <chr>        <int>        <int>     <dbl>     <dbl>
##  1 A+    lv              77           77      52        24
##  2 A+    rv               7            7      46        28
##  3 A     a                1            1      35        35
##  4 A     lv              96           96      51        26
##  5 A     rv              62           62      49        25
##  6 A-    lv            1025         1025      61        21.3
##  7 A-    rv              60           60      48        26
##  8 B+    a                2            2      51.4      43
##  9 B+    lv             139          139      60        17
## 10 B+    rv              21           21      46        33
```

7. How many poll companies are running polls in Alabama and Arkansas? Which of these companies are ONLY running polls in Alabama or Arkansas and how many polls are they running?

```
# There are 7 different poll companies running polls in Arkansas
# There are 5 different poll companies running polls in Alabama
# University of Arkansas
# how to show all the different pollsters in arkansas and alabama?

alabama_arkansas <- polls_us_election_2016 %>%
  group_by(state) %>%
  summarise(n_polls = n(),
            n_companies = n_distinct(pollster)) %>%
  filter(state == "Alabama" | state == "Arkansas") %>%
  arrange(desc(n_companies))

alabama_arkansas2 <- polls_us_election_2016 %>%
  filter(state == "Alabama" | state == "Arkansas") %>%
  select(state, pollster)

example <- polls_us_election_2016 %>%
  group_by(pollster) %>%
  summarise(n_state = n(),
            n_companies = n_distinct(state)) %>%
  arrange(desc(n_companies))

filter(polls_us_election_2016, state == "Alabama" | state == "Arkansas",
       rawpoll_trump < 30)
```

```
##  [1] state            startdate        enddate          pollster
##  [5] grade            samplesize       population       rawpoll_clinton
##  [9] rawpoll_trump    rawpoll_johnson  rawpoll_mcmullin adjpoll_clinton
## [13] adjpoll_trump    adjpoll_johnson  adjpoll_mcmullin
## <0 rows> (or 0-length row.names)
```

6 6 Practical Application (15 pts)

1. We are interested in how good different pollsters were at predicting the actual election results. The popular vote result was 48.2% for Clinton and 46.1% for Trump across the US. Let's see which pollsters got closer results. We are interested in how close the spread (the difference between the proportion of the two candidates 0.482 - 0.461 = 0.021) of each poll was to the real one. We will be using polls for the whole country that ended on or after October 31 (enddate >= "2016-10-31")

a. Calculate the spread for each poll and show the mean spread by pollster

```
# we include national polls conducted during the week of the election
polls <- polls_us_election_2016 %>%
  filter(state == "U.S." & enddate >= "2016-10-31")

# polls <- polls_us_election_2016 %>% mutate(spread = abs(rawpoll_clinton - rawpoll_trump))

polls <- polls_us_election_2016 %>%
  mutate(spread = rawpoll_clinton/100 - rawpoll_trump/100)

polls %>% pull(spread) %>% mean()
```

```
## [1] 0.02162151
```

b. Assume that there are only two parties so that spread = 2 * p - 1. Construct a 95% confidence interval between the two main candidates on election day.

c. Calculate the variable p

```
# spread = 2 * p - 1
# 2 * p = spread - 1
# p = (spread + 1)/2

polls <- polls %>% mutate(p = (spread + 1)/2)
```

b. Calculate the standard deviation of p. Remember sd = 2 * sqrt(p*(1-p)/n). Why is this formula true in this particular case?

```
# spread = 2 * p - 1
# 2 * p = spread - 1
# p = (spread + 1)/2

polls <- polls %>% mutate(sd = 2 * sqrt(p*(1-p)/samplesize))
```

c. Finally calculate the lower and upper confidence interval for the spread. Remember ci = spread +- qnorm(0.975)*sd(spread). Why is this true?

```
polls <- polls %>% mutate(lower_ci = spread - qnorm(0.975) * sd(spread))

polls <- polls %>% mutate(upper_ci = spread + qnorm(0.975) * sd(spread))
```

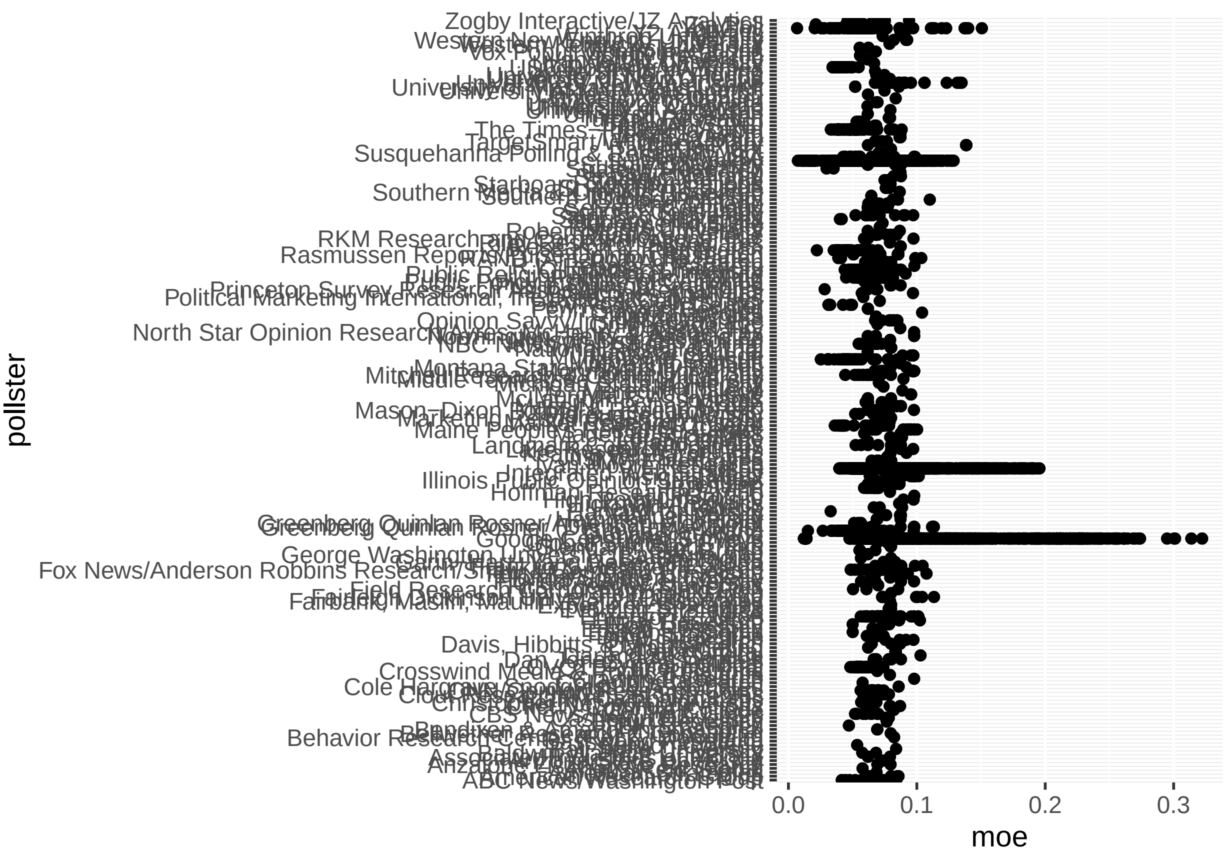
c. Calculate an error variable, the difference between the poll spread and the actual spread from the election. Plot this error by pollster. Flip the pollster names, otherwise your graph will be impossible to read. You already did something similar in the last pset. From this graph you can see which pollster under, over or overall predicted the election night spread.

```
polls <- polls %>% mutate(moe = 1.96 * 2 * sqrt(p * (1 - p) /(polls$samplesize)))

view(polls)

ggplot(data = polls, aes(x = pollster,
                         y = moe)) +
      geom_point() +
      coord_flip()
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```


pollster
Zogby Interactive/JZ Analytics
Susquehanna Polling & Research
Southern Media & Opinion Research
RKM Research and Communications
Rasmussen Reports/Pulse Opinion Research
Princeton Survey Research Associates International
Political Marketing International
Opinion Savvy/InsiderAdvantage
North Star Opinion Research
Montana State University Billings
Mason-Dixon Polling & Research
Maine People's Resource Center
Illinois Public Opinion Strategies
Hoffman Research Group
Greenberg Quinlan Rosner
George Washington University
Fox News/Anderson Robbins Research/Shaw & Company Research
Field Research Corporation
Fairbank, Maslin, Maullin, Metz & Associates
Davis, Hibbitts & Midghall
Crosswind Media & Public Relations
Cole Hargrave Snodgrass & Associates
CBS News/New York Times
Behavior Research Center
ABC News/Washington Post
0.0
0.1
0.2
0.3
moe