# Skills Problem Set 1

Fernanda Sobrino

3/14/2021

Due Thursday April 8, 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: **___**

Add names of anyone you discussed this problem set with: **___**

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

Name your submission files `skills_ps_1.Rmd` and `skills_ps_1.pdf`. (10 points)

# 1 Setup

## 1.1 Installation (10 points)

1. If you do not have R and RStudio installed:watch and follow the video on how to install them
2. If you do not have a github account, set up one now
3. Download Github Desktop here. If you are familiar with using `git` through the command line you are welcome to do so.
4. Initialize your ps1 repository and download the pset_template here. Please read the README file which is visible on the repo's homepage.
5. Make sure you already installed these packages in R: `tidyverse`, `markdown`, and `dslabs`
6. Run a line of code which tests the packages are installed using Stackoverflow instructions posted here. Put the output in your problem set. This lets us know which packages successfully installed and which ones didn't
7. What is your github id?
8. Add and commit your code. Push it to github with commit message "start-up completed"
9. Now we'll practice reverting.
    a. Add the following text to you homework: "Why did the code on Github delete tindr?"
    b. Now push the code to Github.
    c. Now revert to the previous state of the code. (Now that the code is uncommitted, maybe it'll join tindr again.)

# 2 R for Data Science Exercises

## 2.1 First Steps (10 points)

Load `tidyverse` and `dslabs`

We will be using `polls_us_election_2016` data that is in the `dslabs` package

1. How many rows are there in `polls_us_election_2016`? How many columns? What do the rows represent? How about the columns
2. Make a scatter-plot of startdate vs rawpoll_clinton
3. What does the variable grade describes? Use the help '?polls_us_election_2016 to find out.
4. What happens if you make a scatter-plot of population and grade? Why is the plot not useful?

## 2.2 Grammar of graphics: mapping data to aesthetics (20 points)

1. Run `?polls_us_election_2016` to see the documentation for the data set. Run `head(polls_us_election_2016)` to see the first 6 rows of this data frame. Run `colnames(polls_us_election_2016)` to inspect which variables we have for each poll
2. Compare the following scatter-plots. Why are the two graphs different? Which graph is better representation of the data? (*You do not need to graph then to answer these questions.*)

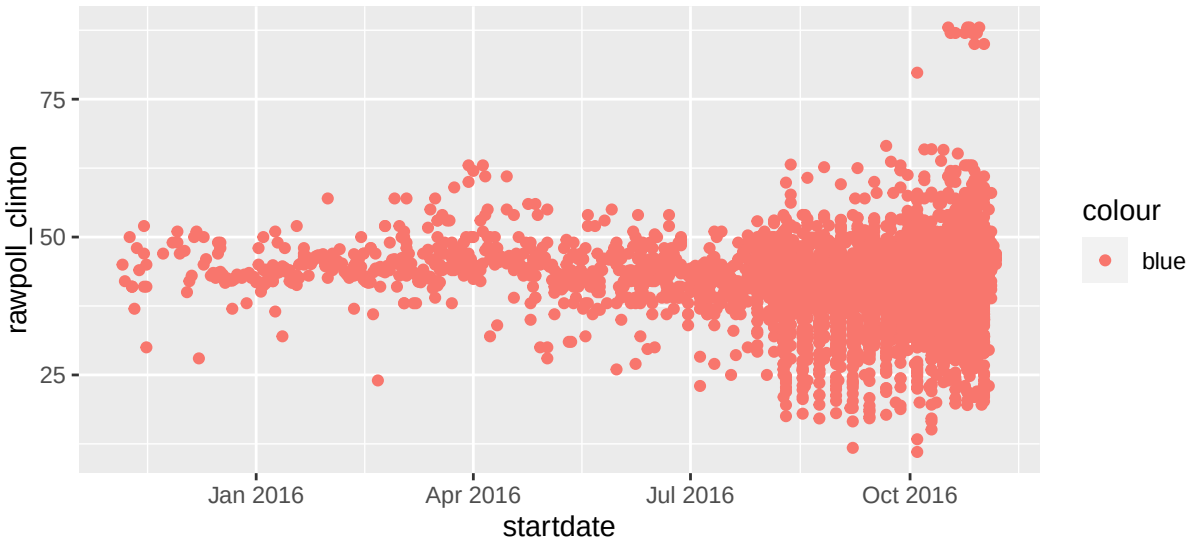
```
# Graph 1
ggplot(data = polls_us_election_2016) +
  geom_point(mapping = aes(x = startdate,
                           y = samplesize,
                           color = rawpoll_johnson))

# Graph 2
ggplot(data = polls_us_election_2016) +
  geom_point(mapping = aes(x = startdate,
                           y = samplesize,
                           color = as.character(rawpoll_johnson)))
```

3. What happens if you map an aesthetic to something other than a variable name, like `aes(color = samplesize >= 500)`

```
ggplot(data = polls_us_election_2016) +
  geom_point(mapping = aes(x = startdate,
                           y = rawpoll_clinton,
                           color = samplesize >= 500))
```

4. Common bugs: What's gone wrong with this code? Fix the code so the points are blue.

```
ggplot(data = polls_us_election_2016) +
  geom_point(mapping = aes(x = startdate,
                           y = rawpoll_clinton,
                           color = "blue"))
```

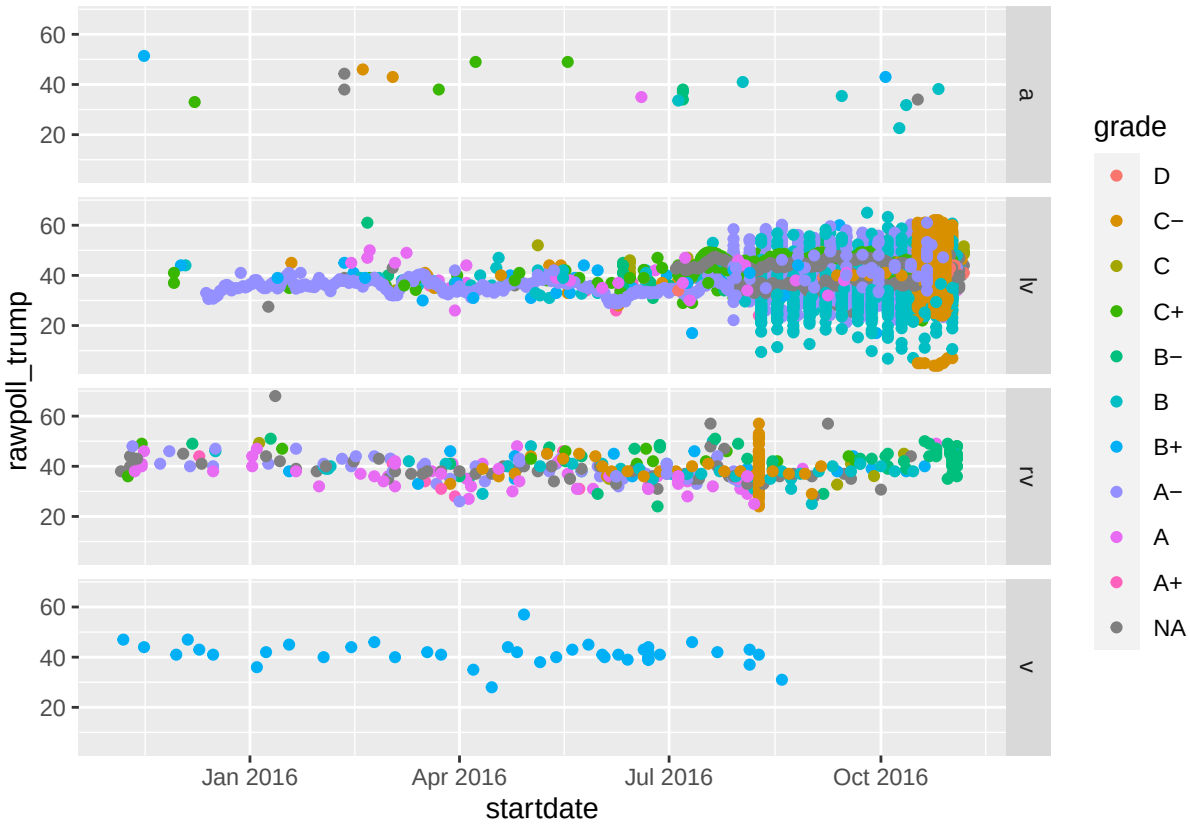



## 2.3 grammar of graphics: Facets (20 points)

1. Make the following plots. How does `facet_grid()` decide the layout of the grid?

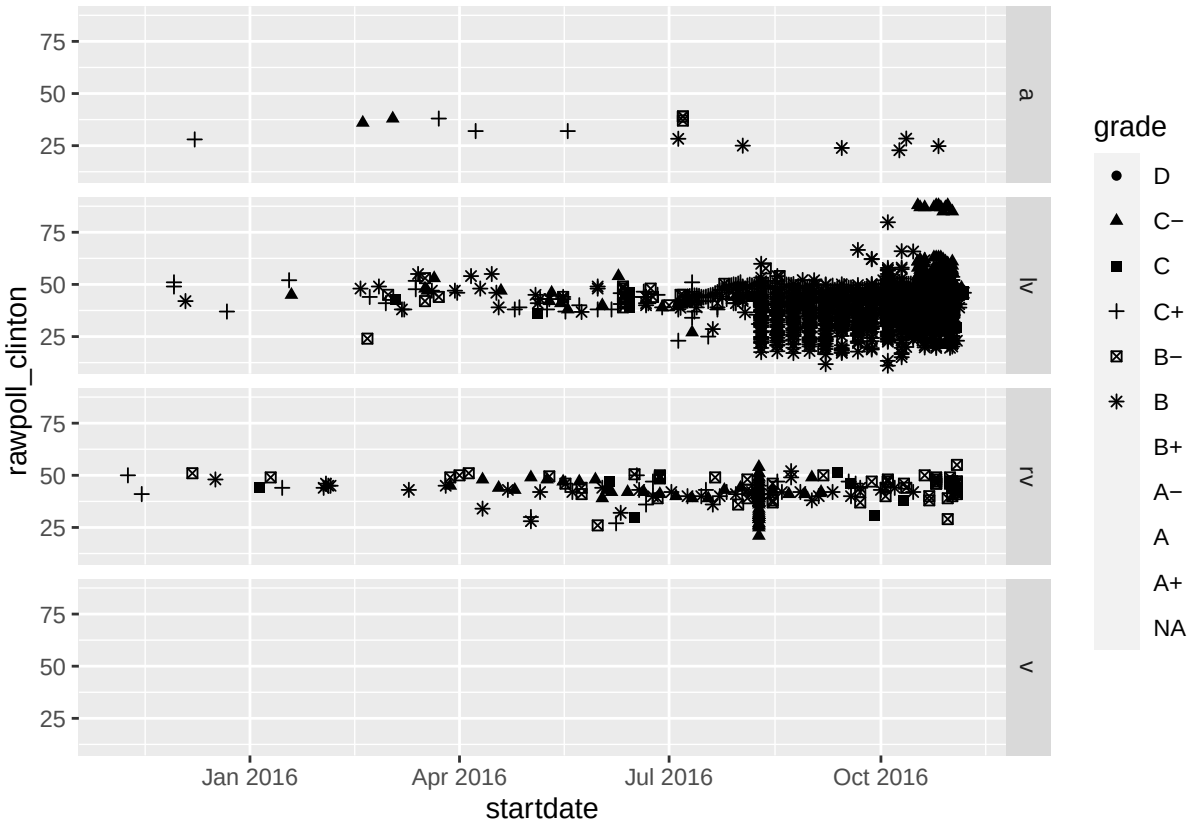
```
ggplot(data = polls_us_election_2016) +
  geom_point(mapping = aes(x = startdate,
                           y = rawpoll_clinton)) +
  facet_grid(cols = vars(population))

ggplot(data = polls_us_election_2016) +
  geom_point(mapping = aes(x = startdate,
                           y = rawpoll_clinton)) +
  facet_grid(rows = vars(grade))

ggplot(data = polls_us_election_2016) +
  geom_point(mapping = aes(x = startdate,
                           y = rawpoll_clinton)) +
  facet_grid(rows = vars(grade), cols = vars(population))
```

2. What happens if you facet a continuous variable? Provide an example
3. Reproduce the following graph




4. Rotate 45 degrees the `startdate` labels from the previous plot. You can use Google, include `ggplot` in your search to get more relevant answers. Remember to cite any code you gather from the internet.

5. Reproduce the following graph. Why are there grades missing?

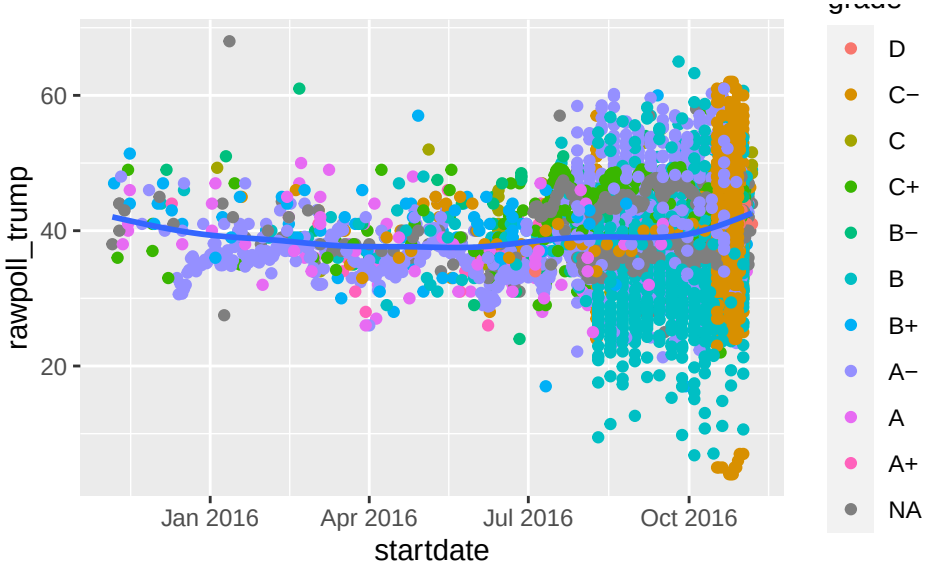



## 2.4 Grammar of graphics: geoms (10 pts)

1. What geom would you use to draw a line chart? A boxplot? A histogram? An area chart?
2. Will these two graphs look different? Why/why not?

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

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

3. You are trying to figure out if there is a relationship between Trump poll numbers and the quality of the polls. Write code to make this graph.




4. Make some changes to this graph:
   - make line red
   - make the x- and y-axes labels more informative using `+ labs()`
   - use an informative title
   - remove the legend (Google might be helpful to learn how) Are all four changes improvements? Which change made the plot worse and why?

#### 2.4.1 grammar of graphics: Statistical transformations (10 pts)

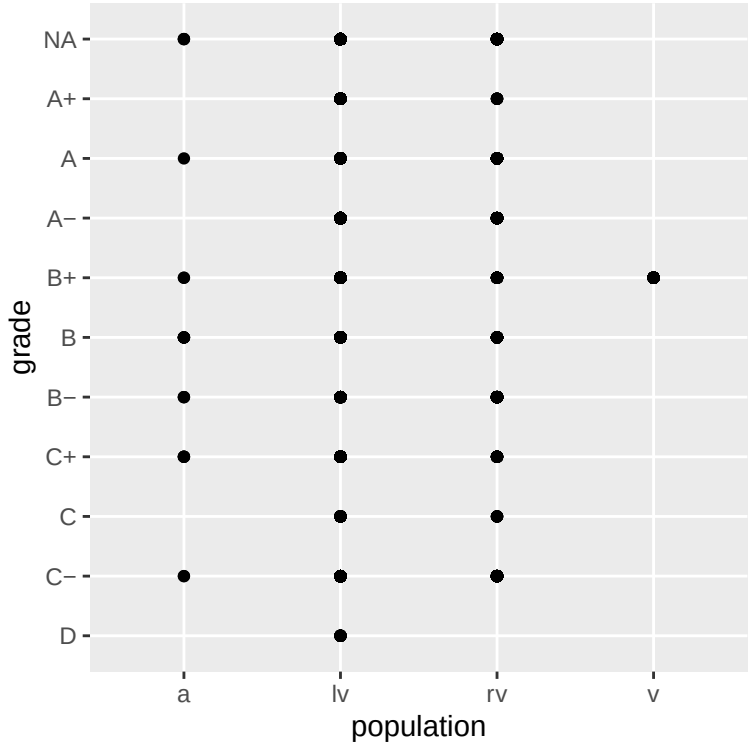
1. What does `geom_col()` do? How is it different from `geom_bar()`?
2. Plot `ggplot(data=polls_us_election_2016, aes(x=grade)) + geom_bar()`. Replace the `geom` with a `stat` to make the same graph.
3. Which 4 variables does `stat_smooth()` compute? How are these variables displayed on a graph made with `geom_smooth()`? What parameters (i.e. inputs to the function) control its behavior?
4. What is wrong with the following graph? Do we need to add `group = 1` to it? What denominator is ggplot using to determine proportions?

```
ggplot(data = polls_us_election_2016) +
  geom_bar(mapping = aes(x = grade, y = ..prop..))
```

### 2.5 grammar of graphics: Positional adjustments (5 pts)

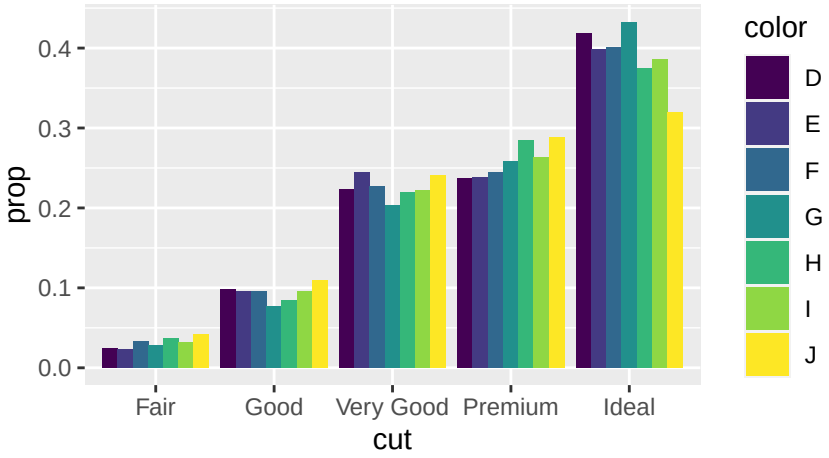
1. What is the problem with this plot? How could you improve it? (You already made this plot in this Pset)

```
ggplot(data = polls_us_election_2016,
       mapping = aes(x = population, y = grade)) +
  geom_point()
```




2. Compare and contrast `geom_jitter()` with `geom_count()`. Use vocabulary from the "grammar of graphics" to support your argument.

3. What's the default position adjustment for `geom_bar()`? What did we add to the code to change the default behavior of geom_bar? Here we are using the diamonds data set again




## 2.6 grammar of graphics: Coordinate systems (5 pts)

1. What happens when you use `coord_flip()`?

2. What is this plot telling us? What does `geom_abline()` do? Why is `coord_fixed()` important?

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

```
  geom_abline() +
  coord_fixed()
```