HW Week 5 – Introduction to Data Viz

Joshua Rosenberg, Alex Lishinski

February 11, 2021

Welcome!

Welcome to *week 5*!

Record the meeting

Learning a language is hard

- None of this is easy or simple
- You're programming (and also working with new concepts)
- And, learning how to navigate a new course

https://twitter.com/datalorax_/status/1361137115497603073

(So, good job! We recognize your difficulties and persistence.)

Breakout rooms!

Starting with whomever is most advanced in one's graduate program:

One question:

- What kinds of visualization do you find useful/interesting?

One reflection/discussion:

Below, what is the a) data, b) function(s), c) argument name(s), and d) argument(s)?

```
state_data_final <- state_data_merge %>%
  complete(state, year = 2011:2020) %>%
  group_by(state) %>%
  fill(adopted, year_month)
```

Review of last week's class

Last week we discussed wrangling and tidying data:

1. Reshaping data

- `pivot_wide()` and `pivot_long()`

1. Joining data

- `left_join()`, `inner_join()`, and others

1. Grouped data operations with dplyr

- `group_by()` and `summarize()`

Review of last week's class

Reading

- From R for Data Science: https://r4ds.had.co.nz/tidy-data.html
- tidy data:
- every variable has its own column
- every observation has its own row

- every value has its own cell

- tidy data makes it easier to use similar tools (even with very different datasets and types of data)

- tidy data works well with R

Review of last week's class

**TB cases*

- Where is the year variable represented?
- Where is the cases variable represented?
- How many observations does each row represent?

```
library(tidyverse)

table4a
```

```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

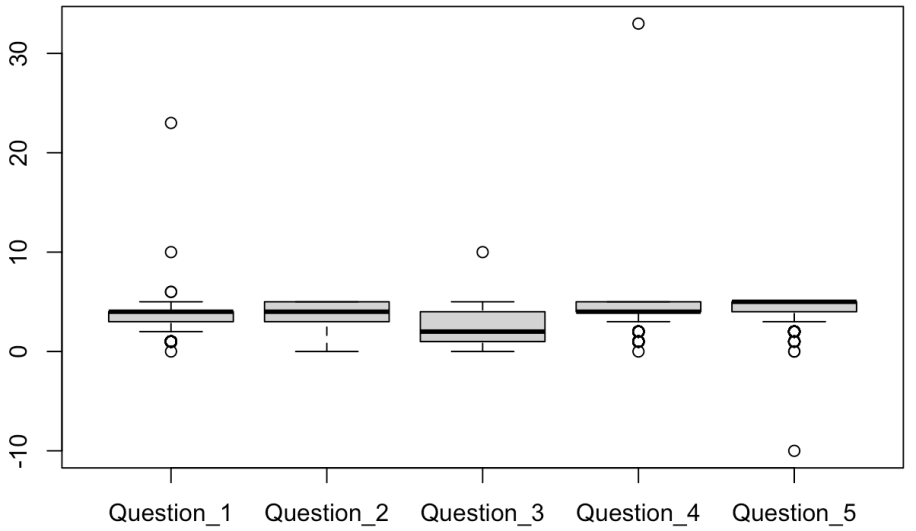
Review of last week's class

- tidy data:

Homework highlights

What do you notice? What do you wonder about?

```
boxplot(pivot_data[3:7])
```



This week's topics

Overview

1. Introduction to data viz
2. A bit more tidying data
3. Data ethics

We are by no means done with the data tidying functions we discussed last week!

Part 1/3: Introduction to data viz

1. Intro to Data Viz

Outline

A. Why visualize data? B. How can we visualize data in R? C. And, how can we make our visualizations aesthetically pleasing?

1A: Why visualize data?

One answer:

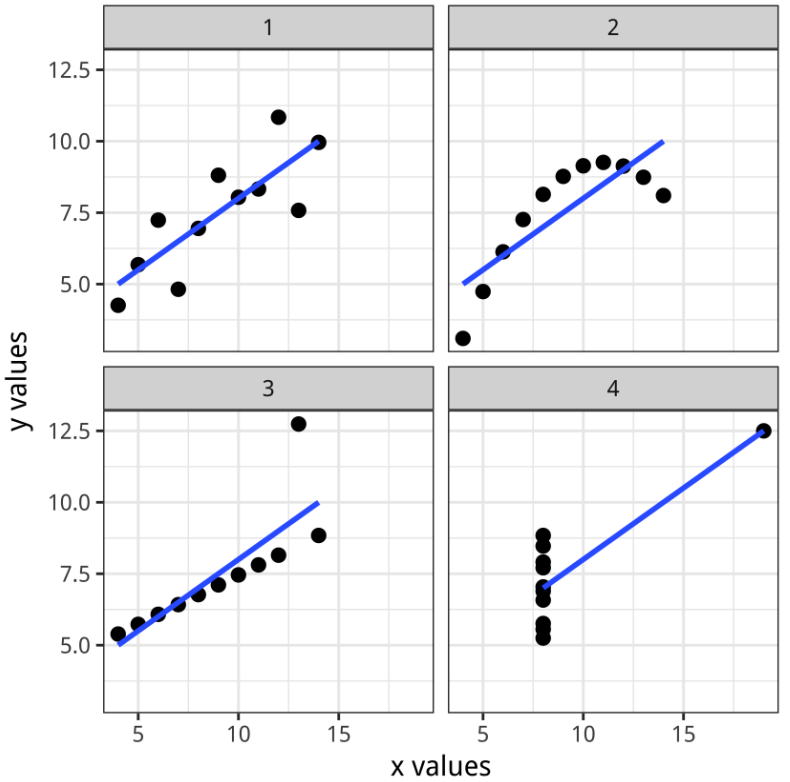
"You should look at your data." (Healy, 2018)

To elaborate on this:

- Visualizations allow to *understand the structure and nature of your data*, and to begin to understand what might relate to what else
- Just like we want to be constantly looking at our data in its spreadsheet/table/data frame format (e.g., `str()`, `glimpse()`, and `View()`), visualizing our data can help us to make sure our data contains what we think it does–and it can alert us to when it does not

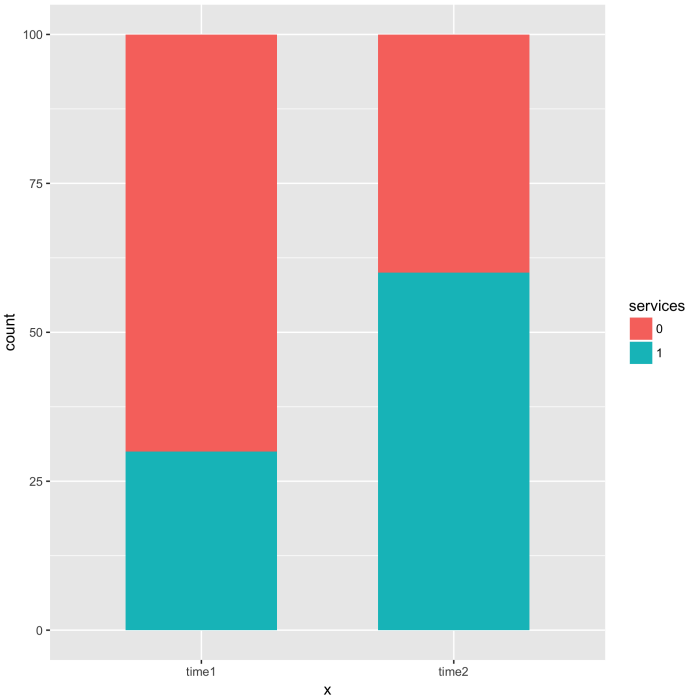
1A: Why visualize data?

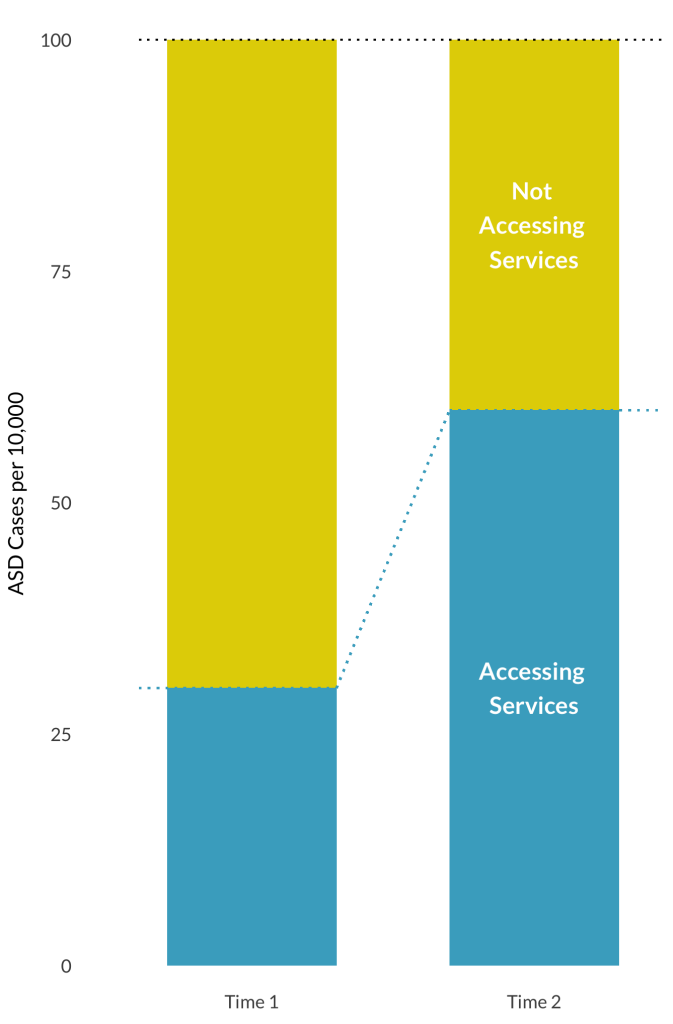
These four different data sets have the same correlation (type `anscombe` in R to view the data), but are very different



1A: Why visualize data

Another reason to visualize data is to *communicate with others*; you can use visualizations to communicate your findings or results. In example:





Estimates of prevalence based on population sampling will remain stable over time if true prevalence is stable.

Estimates of prevalence based on individuals accessing services can create an illusion of an increase in prevalence over time, yet still underestimate prevalence at both time points.

https://apreshill.github.io/ohsu-biodatavis/slides.html#33

1B: How to visualize data

One way to think about visualizing data is in terms of the *type* of visualization you create:

- Histogram
- Density plot
- Scatter plot
- Bar chart
- Pie chart (gasp!)
- Time series plot/line chart

1B: How to visualize data

Another way to think about visualizing data is in terms of the elements that make up a plot.

The *grammar of graphics* (Wickham, 2010, Wilkinson, 2012) has a particular answer to the question of what a plot includes:

Why a grammar of graphics?

- gain insight into complex figures
- reveal deeper relationships between what may appear to be unrelated visualizations
- more flexibly and creatively visualize data––including in ways that do not fit well into one type of plot
- suggest what makes a good figure

1B: How to visualize data?

One view of visualizations is that they consist of four components:

1. Data
2. One or more geometric objects (shape, point, line, etc.)
3. A mapping between variables in the data and the geometric objects and their characteristics (including their size and color)
4. A theme

1B: How to visualize data?

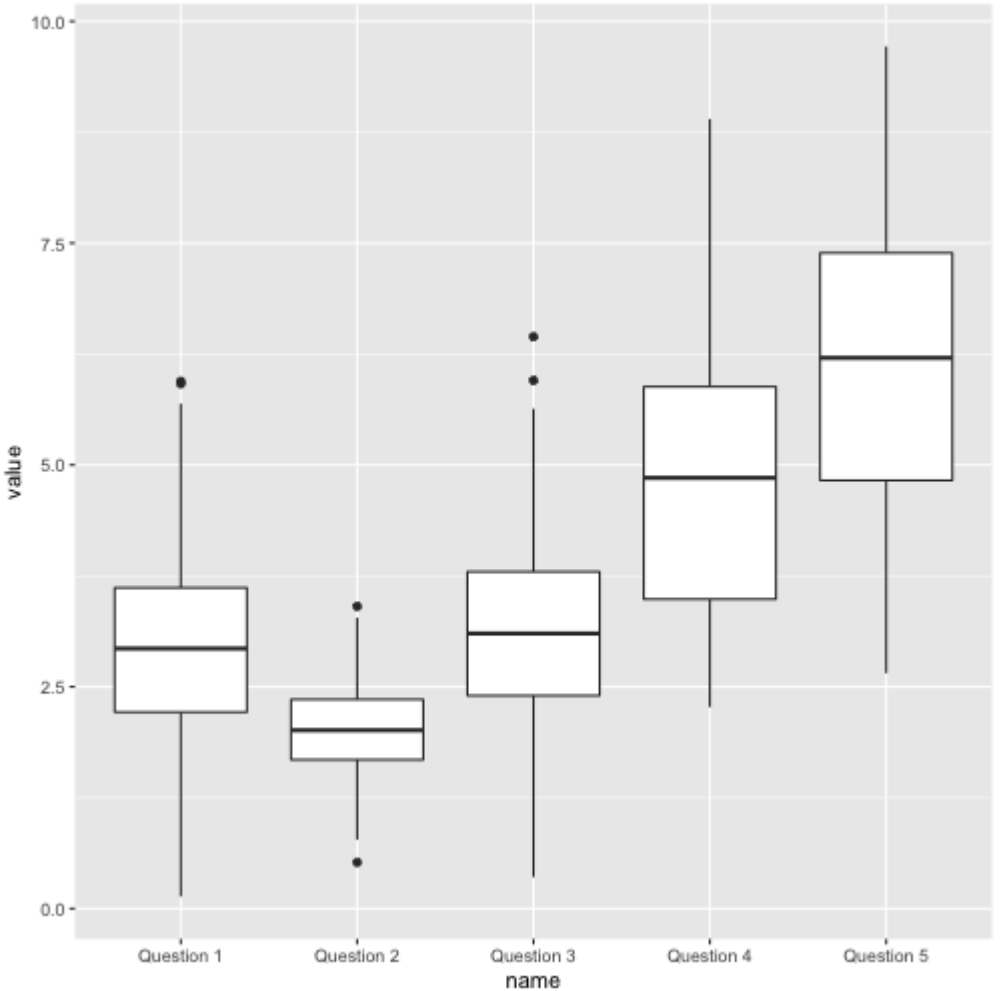
Let's see how this might appear:

```
data
```

```
## # A tibble: 1,618 x 2
##    name       value
##    <chr>      <dbl>
##  1 Question 1  3.46
##  2 Question 1  3.56
##  3 Question 1  1.92
##  4 Question 1  4.63
##  5 Question 1  3.55
##  6 Question 1  4.82
##  7 Question 1  2.69
##  8 Question 1  3.29
##  9 Question 1  2.91
## 10 Question 1  3.87
## # … with 1,608 more rows
```

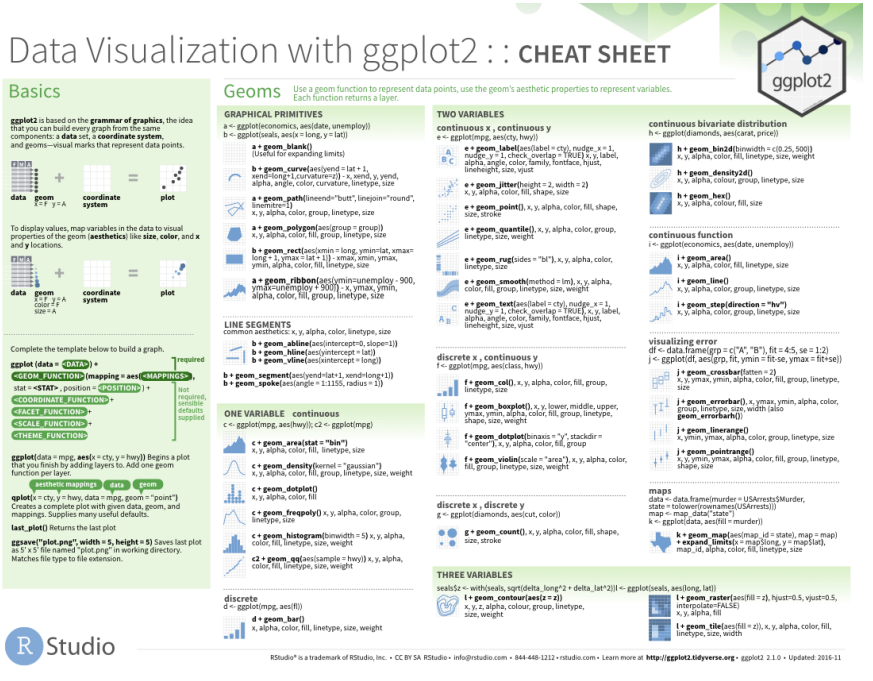
```
data %>%
  ggplot(aes(x = name, y = value)) +
  geom_boxplot()
```

1B: How to visualize data?



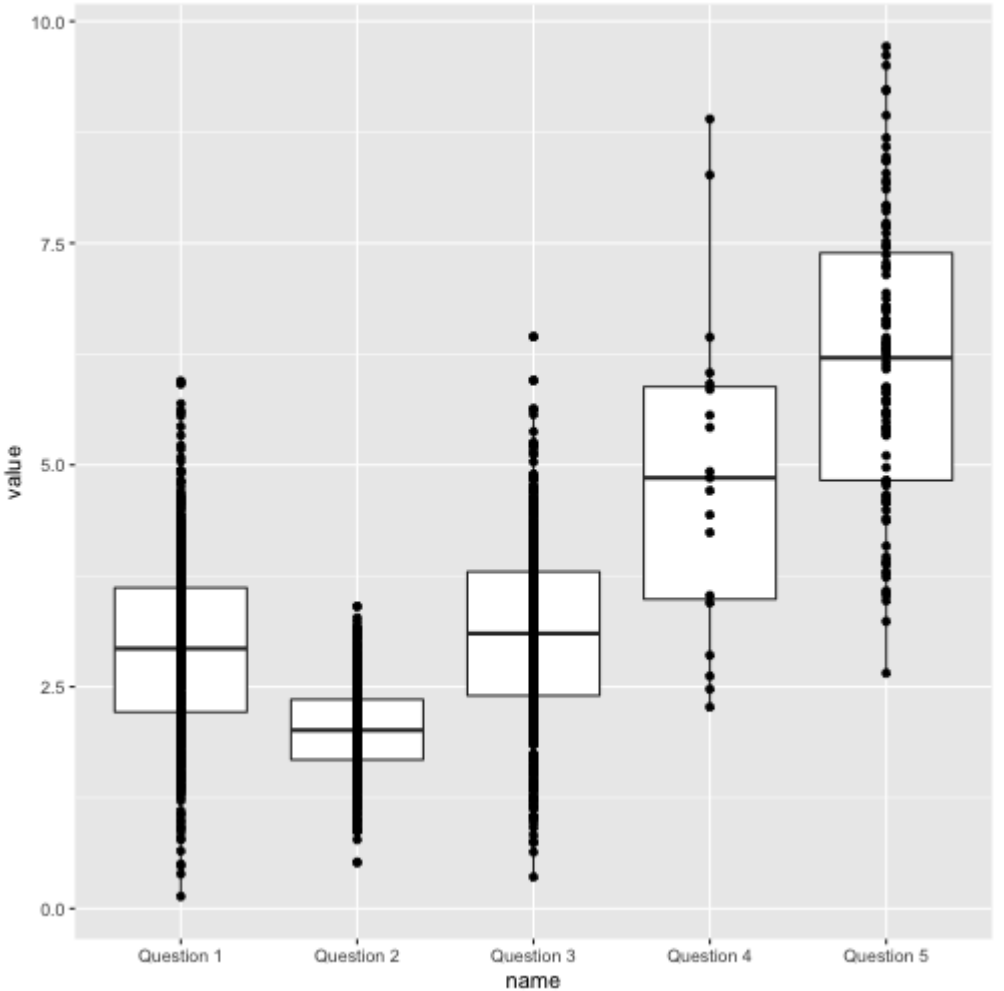
1B: How to visualize data

- The previous slide contained a u potentially *useful* plot
- However, we might be able to improve both its interpretability and its aesthetic

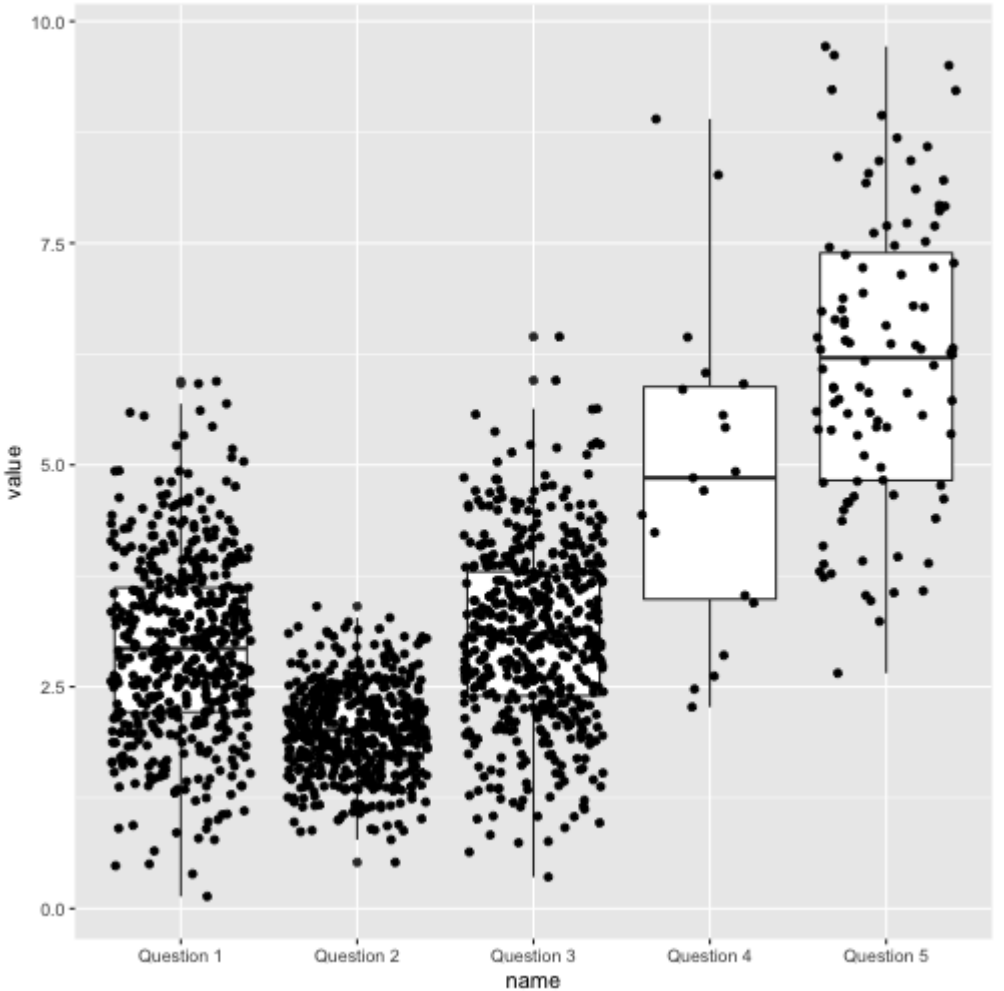


https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf

1B: How to visualize data?



1B: How to visualize data?

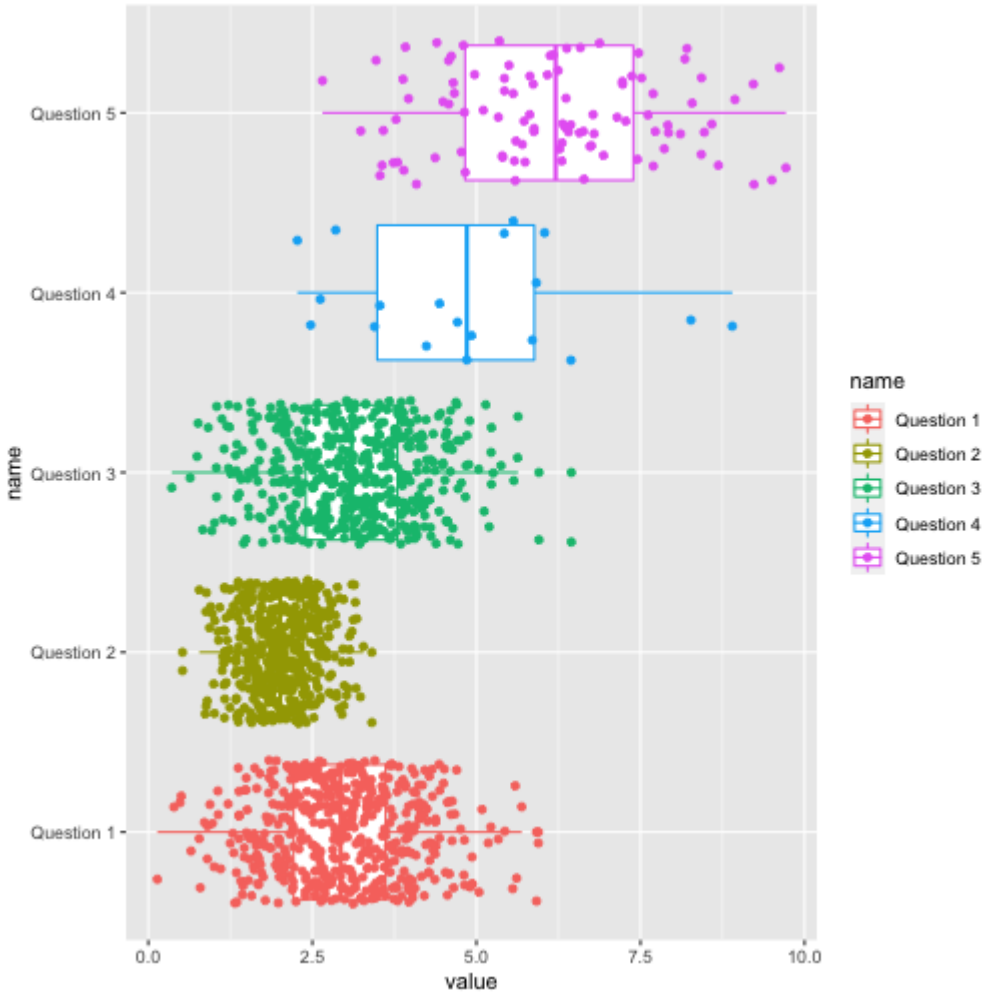


1B: How to visualize data

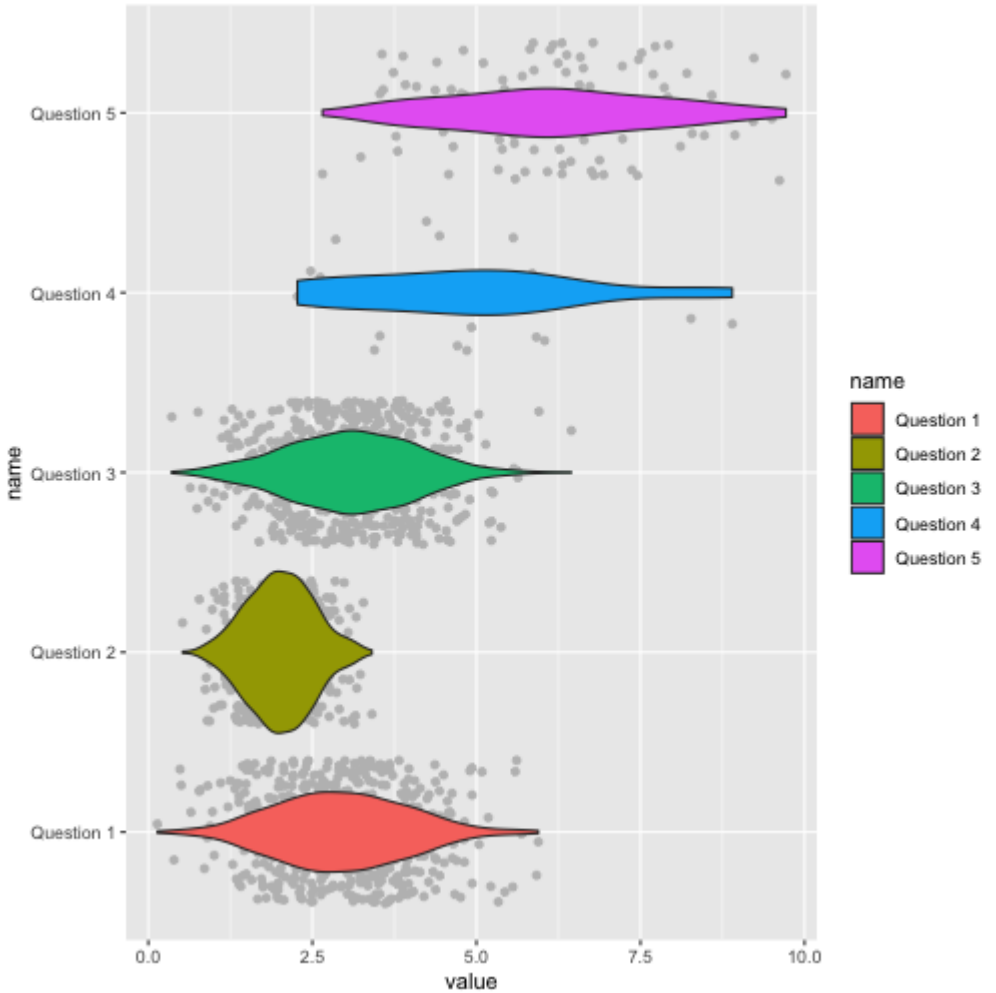
You can create different plots by:

- Changing the aesthetic *mapping* between variables in the data and geometric objects
- Changing the geometric objects

1B: How to visualize data?



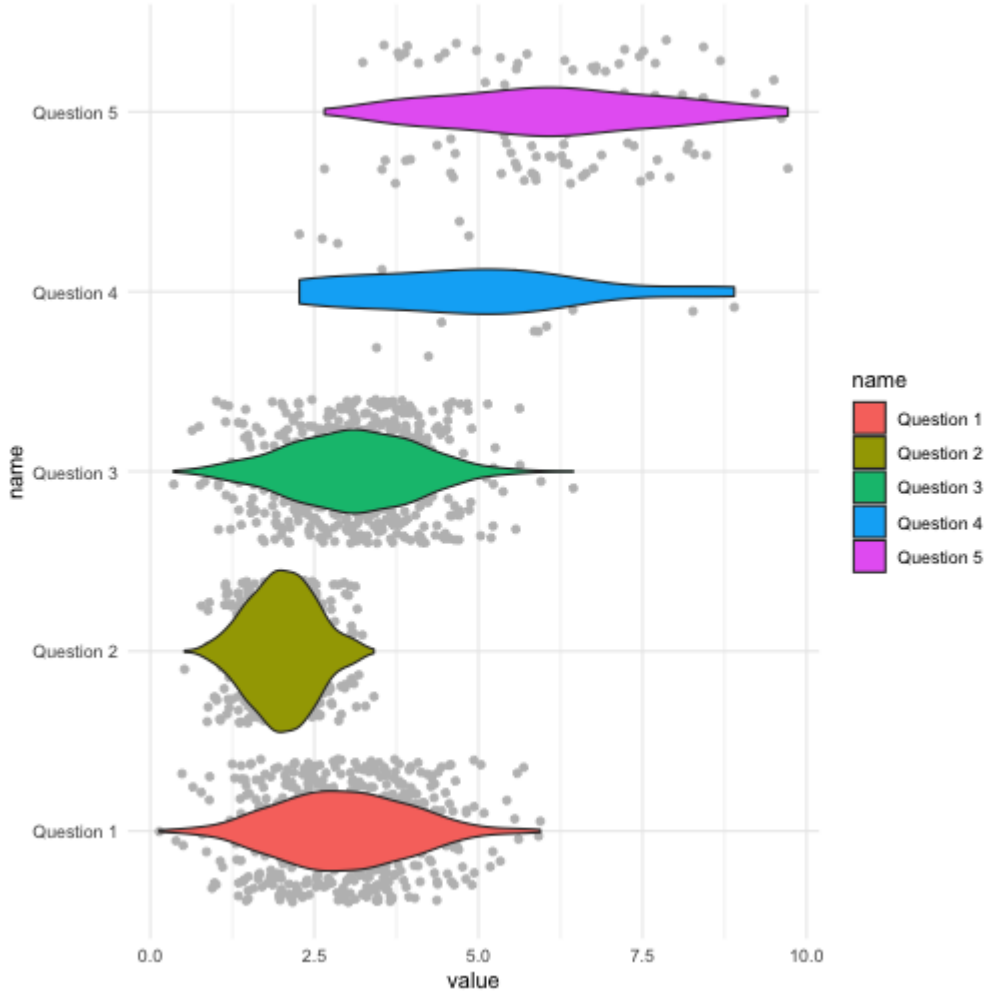
1B: How to visualize data?



1C: How to make visualizations aesthetically pleasing

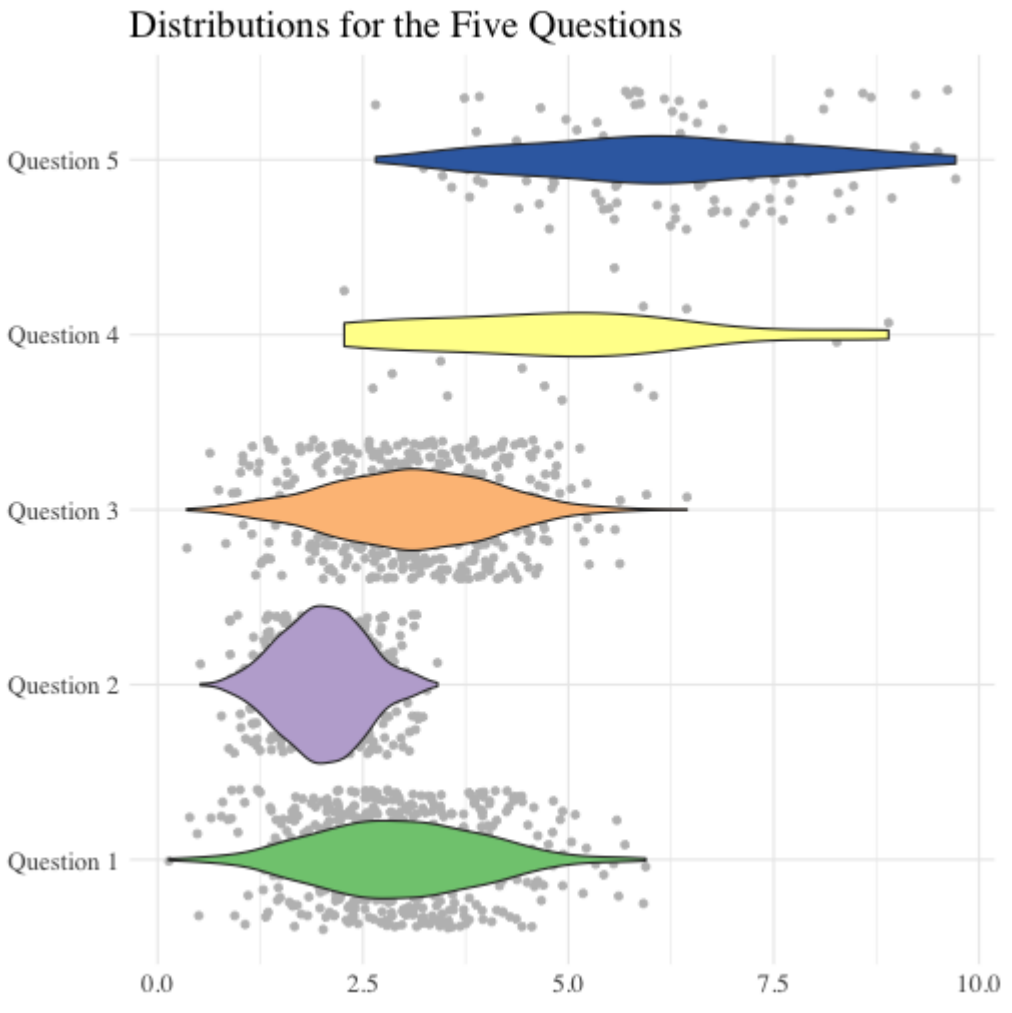
Theming and fine-tuning

```
data %>%
  ggplot(aes(x = value, y = name, fill = name)) +
  geom_jitter(color = "gray") +
  geom_violin() +
  theme_minimal()
```


Question 5
Question 4
Question 3
Question 2
Question 1
name
0.0
2.5
5.0
7.5
10.0
value
name
Question 1
Question 2
Question 3
Question 4
Question 5

Theming and fine-tuning

```
data %>%
  ggplot(aes(x = value, y = name, fill = name)) +
  geom_jitter(color = "gray") +
  geom_violin() +
  theme_minimal() +
  scale_fill_brewer("", type = "qual") +
  ylab(NULL) +
  xlab(NULL) +
  theme(text = element_text(size = 16, family = "Times"),
        legend.position = "none") +
  ggtitle("Distributions for the Five Questions")
```


Distributions for the Five Questions
Question 5
Question 4
Question 3
Question 2
Question 1
0.0
2.5
5.0
7.5
10.0

Part 2/3: A bit more tidying data

2: How does tidying data relate to data viz?

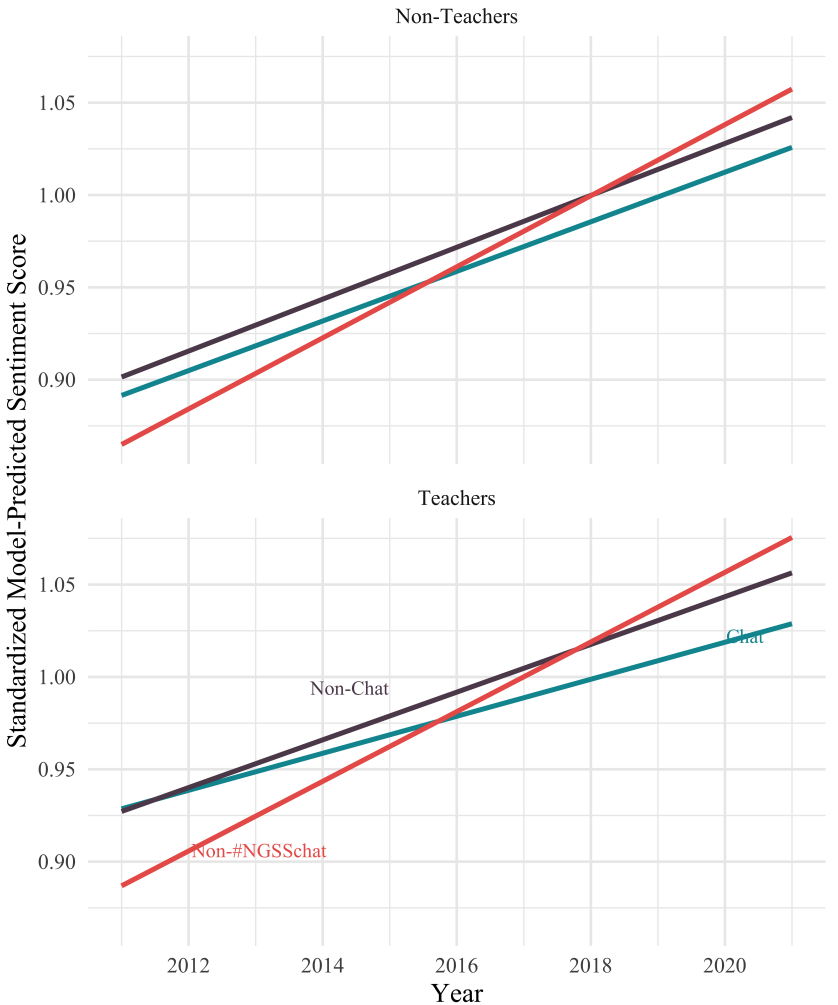
Often, we have to make changes to our data frame in order to create the visualization we would like to create.

2: How does tidying data relate to data viz?

Making a new variable prior to plotting the data

```
pred_frame %>%
  mutate(isTeacher = ifelse(isTeacher == 0, "Non-Teachers", "Teac
  ggplot(aes(year_of_post_centered + 2016, prediction)) +
  geom_line(aes(color = type_of_tweet),
            size = 1.3) +
  geom_text(aes(label = label, color = type_of_tweet),
            family = "Times New Roman",
            data = label_frame) +
  facet_wrap(~isTeacher, ncol = 1) +
  scale_x_continuous("Year", breaks = seq(2010, 2020, 2)) +
  labs(x = "Year",
       y = "Standardized Model-Predicted Sentiment Score")
```

2: How does tidying data relate to data viz?



2: How does tidying data relate to data viz?

Other data tidying steps we might take prior to visualizing data:

- **recoding** variables
- **creating a factor** (so that we can order elements of a plot as we wish for them to be ordered)
- **grouping** and **summarizing** to plot a summary statistic
- realizing that your data processing and tidying was not quite sufficient, so **returning to those stages** before finalizing your visualization
- **re–running our analysis** (`.Rmd` file) because we discovered an issue with our data

Part 3/3: Data ethics

3: Why data ethics?

With great data powers comes great responsibility!

- Ethics matter, especially when we are working with vulnerable populations (or data about them)
- And, ethical concerns may extned beyond what our Institutional Review Board considers

3: Why data ethics?

"... Surveillance photos were taken from the building on the upper right and captured images of more than 1.700 students, faculty members and other passers–by walking on the path near the West Lawn, the large grassy area on the left."

https://www.denverpost.com/2019/05/27/cu–colorado–springs–facial–recognition–research/

3: Examples of positive and negative data ethics

- Positive: Privacy threat modeling and plan mitigation strategies (Lundberg et al., 2019)
- Negative: Identifying individual participants through social media posts included in presentations and publications

Course Logistics

- Exam 1: Recap
- Homework 5: Due by Tuesday, 1/23
- Reading: Considerations for using social media data in learning design and technology research (Greenhalgh et al., 2020)

Random

- Analyzing educational data with open science best practices, R, and OSF
- Continue to express your work/challenges with R as precisely as possible, e.g.:
  - A data frame versus a data file
  - The name for a data frame
- How do I format code in Slack? Try enclosing code in back tick marks ` `

Wrapping up

In your base group's Slack channel:

- What is one thing you learned today?
- What is something you want to learn more about?
- *Also*, in GIF form (type `/giphy` in Slack, and then a random term), summarize how you are feeling about R