8380 R Portfolio 1

Caitlin Jagla

3/25/2022

A) Load & Explore Data

This dataset describes measles vaccination rates across schools for the 2018-2019 school year.

```
# load data
df <- read_csv("measles.csv")

# load table describing variables (I made this csv based on the table in the pdf)
desc <- read_csv("measles_desc.csv")
```

```
# view description of variables contained in the dataset
desc
```

```
## # A tibble: 13 x 2
##    variable description
##    <chr>    <chr>
##  1 index    index ID
##  2 state    school state
##  3 year     academic year
##  4 name     school name
##  5 type     school type (public, private, or charter)
##  6 city     city
##  7 county   county
##  8 enroll   number of students enrolled
##  9 mmr      MMR (measles, mumps, rubella) vaccination rate
## 10 overall  overall vaccination rate
## 11 xrel     percent of students exempt from vaccination for religious reasons
## 12 xmed     percent of students exempt from vaccination for medical reasons
## 13 xper     percent of students exempt from vaccination for personal reasons
```

```
# preview dataset
glimpse(df)
```

```
## Rows: 66,113
## Columns: 13
## $ index   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 11, 12, 13, 14, 15, 15, 16,~
## $ state   <chr> "Arizona", "Arizona", "Arizona", "Arizona", "Arizona", "Arizon~
## $ year    <chr> "2018-19", "2018-19", "2018-19", "2018-19", "2018-19", "2018-1~
## $ name    <chr> "A J Mitchell Elementary", "Academy Del Sol", "Academy Del Sol~
```

```
## $ type    <chr> "Public", "Charter", "Charter", "Charter", "Charter", "Public"~
## $ city    <chr> "Nogales", "Tucson", "Tucson", "Phoenix", "Phoenix", "Phoenix"~
## $ county  <chr> "Santa Cruz", "Pima", "Pima", "Maricopa", "Maricopa", "Maricop~
## $ enroll  <dbl> 51, 22, 85, 60, 43, 36, 24, 22, 26, 78, 78, 35, 54, 54, 34, 57~
## $ mmr     <dbl> 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 10~
## $ overall <dbl> -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1~
## $ xrel    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ xmed    <dbl> NA, NA, NA, NA, 2.33, NA, NA, NA, NA, NA, NA, 2.86, NA, 7.41, ~
## $ xper    <dbl> NA, NA, NA, NA, 2.33, NA, 4.17, NA, NA, NA, NA, NA, NA, NA, NA~
```

B) Data Tidying

1. Many rows appear to be repeat observations. The dataset has 66113 observations, but only 46411 are distinct.
2. Missing values are represented differently in different variables (`NA` OR `-1` OR `null`). They should be standardized across all columns.
3. The majority of datapoints (0.7271641%) are from the 2018-2019 academic year, so I will focus on these and remove all the others to avoid repeated sampling from the same school. This also has the effect of tidying the `type` variable so that it only includes `public`, `private`, `charter`, or `NA`.
4. I am interested in vaccination rates, so I will remove any rows that are missing both the MMR and overall vaccination rate values.

```
df <- df |>
  distinct() |> # remove non-distinct rows
  mutate(across(where(is.character), ~ na_if(.x, "null")),
         across(where(is.numeric), ~ na_if(.x, -1))) |> # standardize representation of missing values
        filter(year == "2018-19") |> # keep only 2018-2019 academic year
        filter(!(is.na(mmr) & is.na(overall))) # remove datapoints that don't report either rate
```

C) Generate an Aggregated Vaccination Rate

Many states report only one type of vaccination rate (either `mmr` or `overall`, not both). This makes it challenging to compare vaccination rates across the country, as essentially the `vaccination rate type` is confounded with the `state` variable. I wondered if it would be possible to integrate the two variables into a single aggregated vaccination rate.

This strategy relies on a few assumptions:

1. In most US communities in 2018-2019, vaccination rates would have been roughly equivalent across all standard childhood vaccinations. Therefore, the MMR rate should generally reflect the overall vaccination rate, and vice versa.
2. Extreme differences between overall vaccination rate and MMR vaccination rate within the same school are likely related to vaccination schedules or other reporting issues. In order to gauge a school community's general vaccine uptake, in cases where both rates are reported and there is a very large difference, it seems reasonable to take whichever rate is higher. A potential issue with this assumption is that MMR vaccines specifically were falsely reported to be linked to autism, so hesitancy for these vaccines may not be representative of overall vaccine hesitancy. If this were the case, I would expect to see a pattern of MMR vaccination rates being lower than overall vaccination rates reported by the same schools. This question is investigated below, alongside additional validation of this aggregated rate variable approach.

```r
# generate an `aggregate vaccination rate` column


# if/else mutate to aggregate the two vaccination rates into a single column
    # if mmr exists and overall does not, use mmr
    # if overall exists and mmr does not, use overall
    # if both exist, use whichever is higher
    # if both exist and are equal, use overall
df <- df |> mutate(vacc_rate =
                       case_when(
                         !is.na(mmr) & is.na(overall) ~ mmr,
                         !is.na(overall) & is.na(mmr) ~ overall,
                         !is.na(mmr) & !is.na(overall) & overall == mmr ~ overall,
                         !is.na(mmr) & !is.na(overall) & overall > mmr ~ overall,
                         !is.na(mmr) & !is.na(overall) & mmr > overall ~ mmr))
```

```r
# generate variables to investigate rate differences

# if/else mutate to describe source of aggregate vacc_rate
df <- df |> mutate(rate_source =
                       case_when(
                         !is.na(mmr) & is.na(overall) ~ "mmr",
                         !is.na(overall) & is.na(mmr) ~ "overall",
                         !is.na(mmr) & !is.na(overall) ~ "both"))

# if/else mutate to calculate differences between the two vaccination rates
    # if mmr exists and overall does not, use -9999 as placeholder
    # if overall exists and mmr does not, use -9999 as placeholder
    # if both exist, subtract mmr rate from overall rate
df <- df |> mutate(diff =
                       case_when(
                         !is.na(mmr) & is.na(overall) ~ -9999,
                         !is.na(overall) & is.na(mmr) ~ -9999,
                         !is.na(mmr) & !is.na(overall) ~ overall - mmr))
```

C2) Investigation of differences in MMR & overall vaccination rates reported by the same schools:

- schools where only one rate was reported: `n` = 18735
- schools where both rates are reported & those rates are exactly equal: `n` = 4892
- schools where both rates are reported & `overall` > `mmr`: `n` = 3
- schools where both rates are reported & `mmr` > `overall`: `n` = 4674
- Fig C2a) MMR & overall vaccination rates reported by the same school generally differ by 10% or less
- Fig C2b & C2c) MMR & overall vaccination rates reported by the same school are generally unaffected by school type or state
- Fig C2d) Aggregate vaccination rate is generally unaffected by rate source

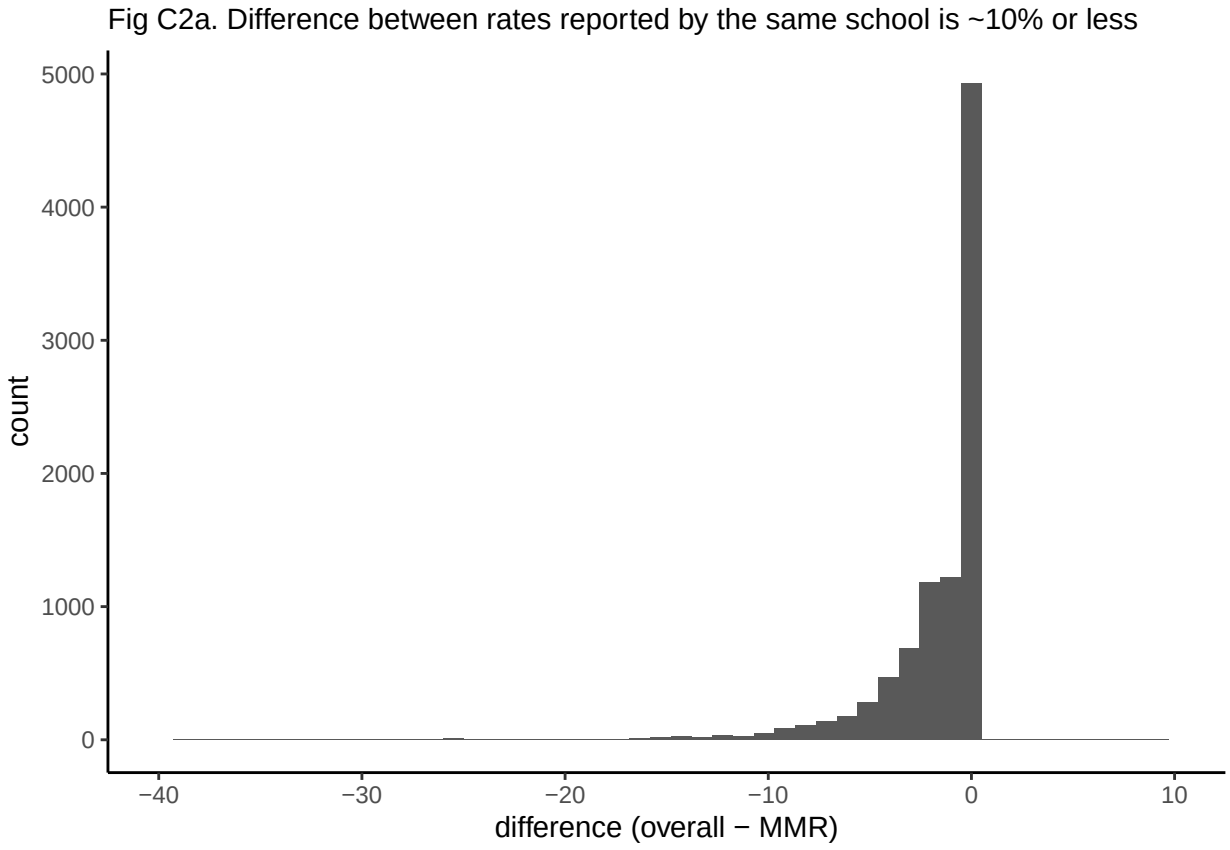
```r
# compare differences in mmr & overall vaccination rates reported by the same school


# calculate mean & stdev for differences between the two rates reported by the same school
df |> filter(diff != -9999) |> summarize(diff_mean = mean(diff),
                                         diff_sd = sd(diff))
```

```
## # A tibble: 1 x 2
##   diff_mean diff_sd
##       <dbl>   <dbl>
## 1     -1.81    3.49
```

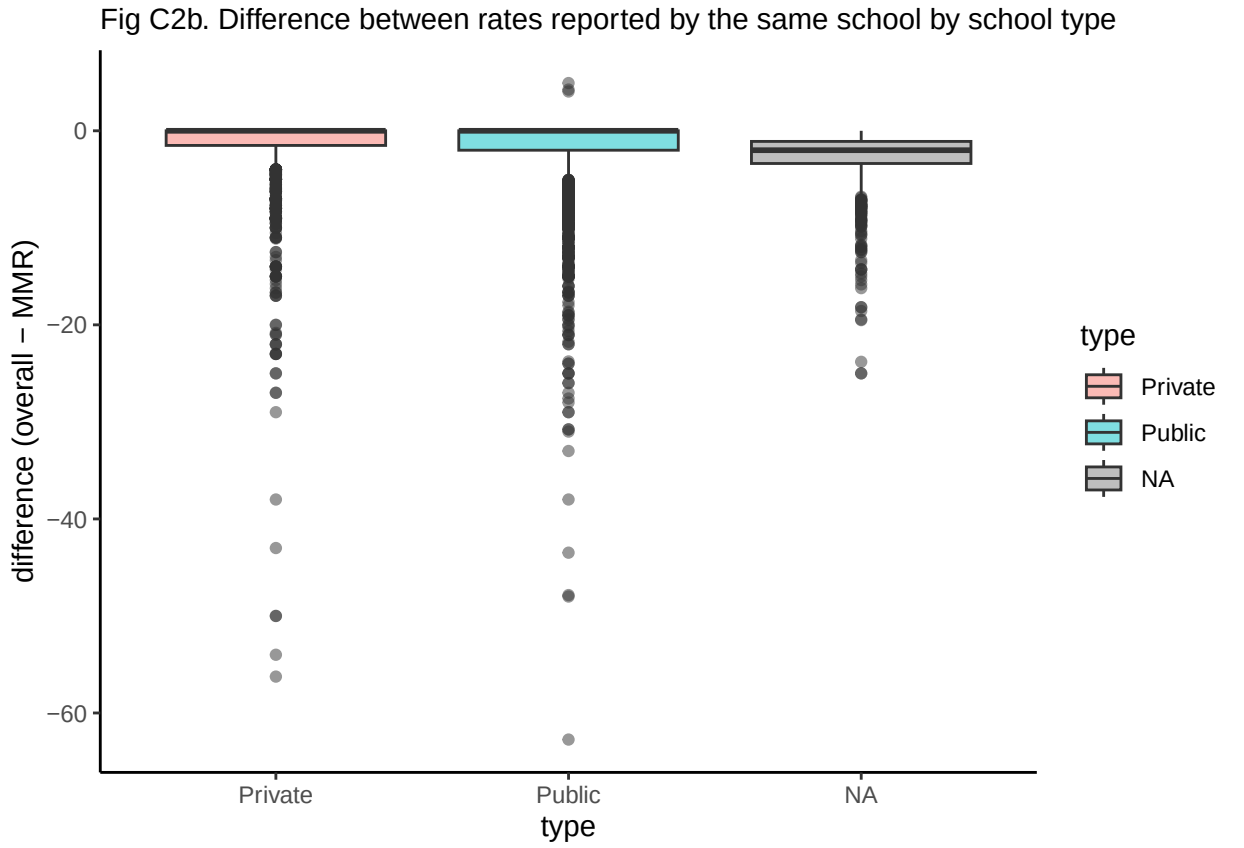
```r
df |> filter(diff != -9999) |>
  ggplot(aes(x = diff)) +
  geom_histogram(bins = 50) + xlim(-40,10) +
  labs(subtitle = "Fig C2a. Difference between rates reported by the same school is ~10% or less",
       x = "difference (overall - MMR)") +
  theme_classic()
```

```
## Warning: Removed 9 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

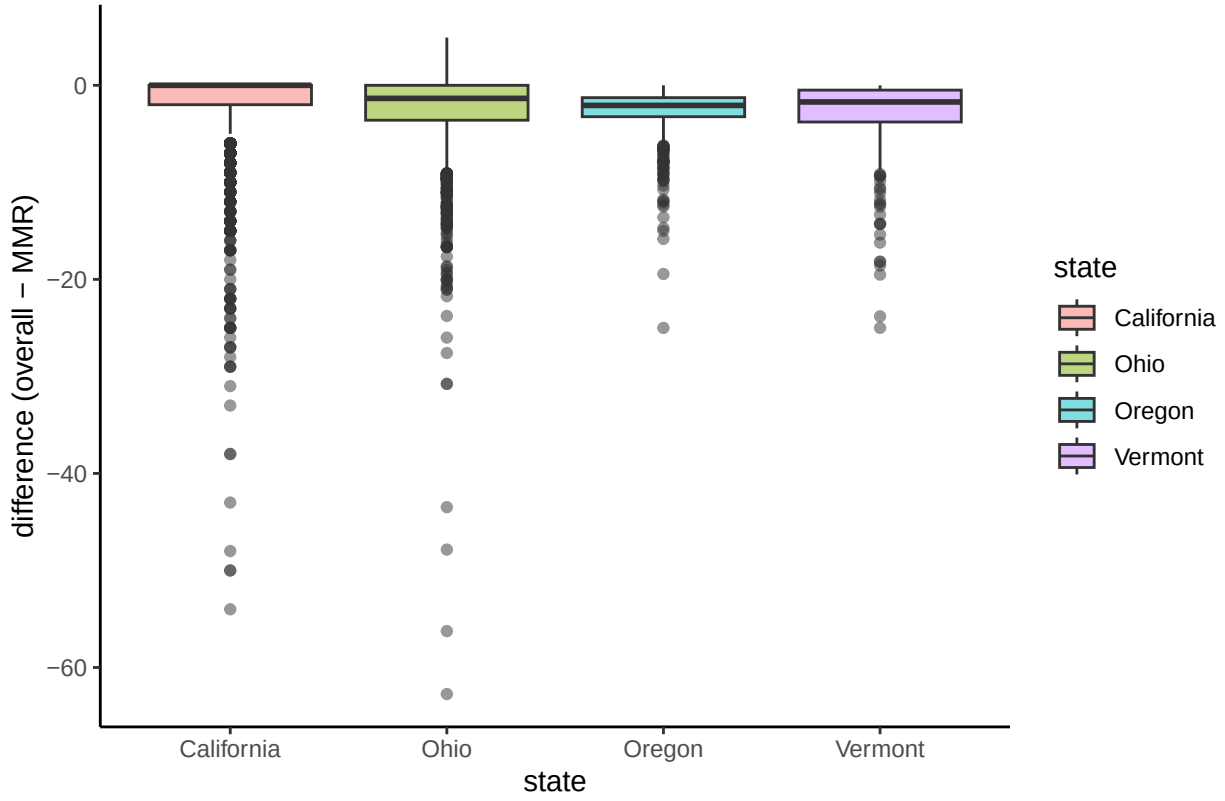


```
df |> filter(diff != -9999) |>
  ggplot(aes(y = diff, x = type, fill = type)) +
  geom_boxplot(alpha = 0.5) +
  labs(subtitle = "Fig C2b. Difference between rates reported by the same school by school type",
       y = "difference (overall - MMR)") +
  theme_classic()
```




```
df |> filter(diff != -9999) |>
  ggplot(aes(y = diff, x = state, fill = state)) +
  geom_boxplot(alpha = 0.5) +
 labs(subtitle = "Fig C2c. Difference between rates reported by the same school by state",
       y = "difference (overall - MMR)") +
  theme_classic()
```

Fig C2c. Difference between rates reported by the same school by state

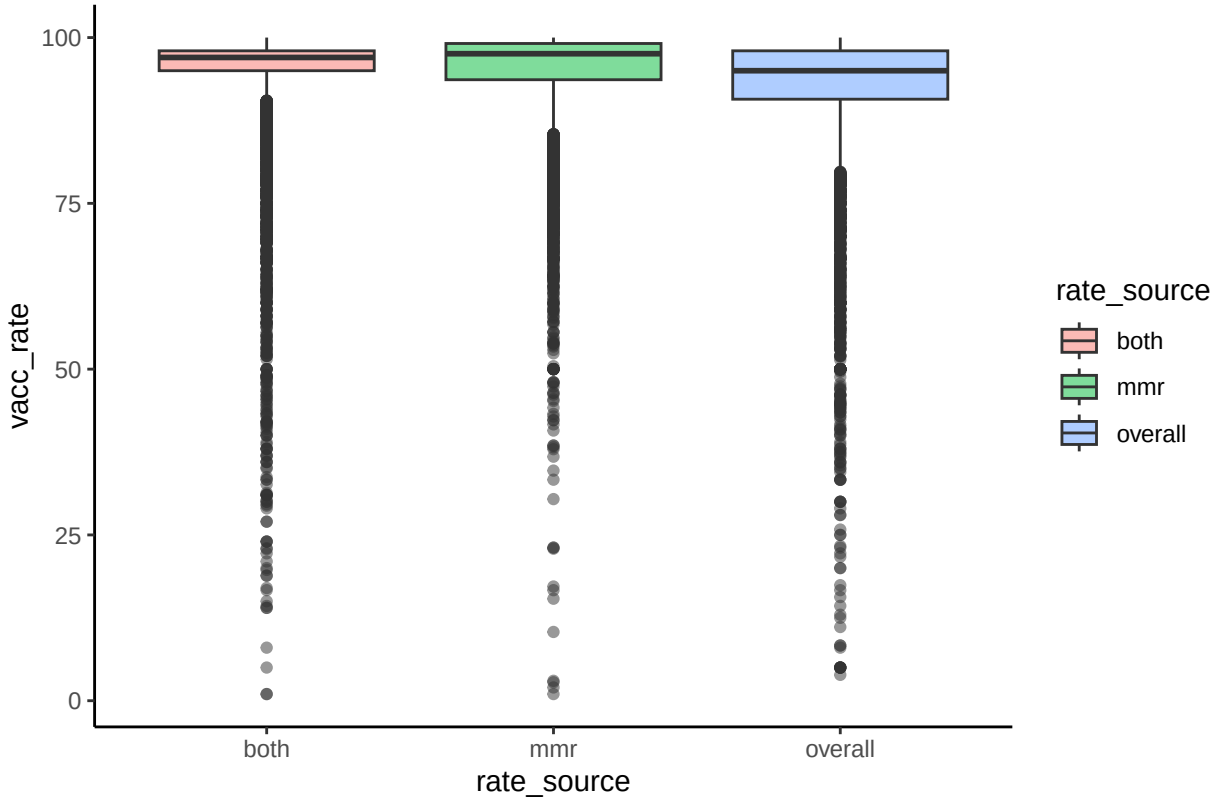



```
# summarize by source for aggregated rate
df |> group_by(rate_source) |> summarize(vaccrate_mean = mean(vacc_rate),
                                         vaccrate_sd = sd(vacc_rate))
```

```
## # A tibble: 3 x 3
##   rate_source vaccrate_mean vaccrate_sd
##   <chr>               <dbl>       <dbl>
## 1 both                 94.4        8.79
## 2 mmr                  94.4        8.69
## 3 overall              92.5        9.45
```

```
#visually verify that rate source for the aggregated `vacc_rate` variable does not confound
df |>
  ggplot(aes(y = vacc_rate, x = rate_source, fill = rate_source)) +
  geom_boxplot(alpha = 0.5) +
  labs(subtitle = "Fig C2d.  Aggregate vaccination rate is generally unaffected by rate source") +
  theme_classic()
```

Fig C2d. Aggregate vaccination rate is generally unaffected by rate source



Based on these results, I believe it is reasonable to aggregate the two different vaccination rate types into a single `vacc_rate` variable in order to facilitate across-state comparisons. This data wrangling will make it easier to analyze vaccination trends nationwide.

D) Investigation of Vaccination Rates By State

Comparing vaccination rates across all the states in the tidied dataset, it is clear that there a few states with substantially lower vaccination levels among students than the nationwide mean of 93.633652. In contrast, 13 of 20 states remaining in the filtered dataset have median vaccination rates greater than or equal to 95%.

```
# summarize by state for aggregated rate
df |>
  group_by(state) |>
  summarize(vaccrate_med = median(vacc_rate),
            vaccrate_mean = mean(vacc_rate),
            vaccrate_sd = sd(vacc_rate)) |>
  arrange(vaccrate_med)
```

```
## # A tibble: 20 x 4
##    state          vaccrate_med vaccrate_mean vaccrate_sd
##    <chr>                 <dbl>         <dbl>       <dbl>
##  1 Arkansas               82.2          80.5        8.38
##  2 Idaho                  86.7          82.4       15.6
```

```
##  3 Wisconsin              90           85.8       12.4
##  4 Michigan               93.5         92.1        6.81
##  5 Florida                94.4         92.5        8.67
##  6 Ohio                   94.6         90.4       12.7
##  7 Maine                  94.7         92.6        8.97
##  8 Arizona                95           92.7        8.70
##  9 Oregon                 95.5         93.9        6.34
## 10 Tennessee              95.8         94.9        4.44
## 11 Virginia               95.9         93.8        8.16
## 12 Missouri               96           91.1       12.8
## 13 Iowa                   96.9         95.8        4.73
## 14 Vermont                97.2         94.6        8.64
## 15 Rhode Island           97.4         94.9        9.02
## 16 Texas                  97.9         93.1       10.7
## 17 California             98           95.6        7.29
## 18 Illinois               98.2         97.1        4.64
## 19 Massachusetts          99           96.9        6.30
## 20 North Carolina        100           96.8        7.52
```

```r
# plot vaccination rate by state
df |>
  ggplot(aes(x = vacc_rate, y = reorder(state, vacc_rate, median), fill = state)) +
  geom_boxplot(alpha = 0.5, show.legend = FALSE) +
  labs(title = "Fig D1) Aggregate vaccination rates by state") +
  theme_classic()
```