Review of Data Science 1

You can download this .qmd file from here. Just hit the Download Raw File button.

Determinants of COVID vaccination rates

First, a little detour to describe several alternatives for reading in data:

If you navigate to my Github account, and find the `264_spring_2025` repo, there is a Data folder inside. You can then click on `vacc_Mar21.csv` to see the data we want to download. This link should also get you there, but it's good to be able to navigate there yourself.

```
# Approach 1
vaccine_data <- read_csv("Data/vaccinations_2021.csv")                    ①

# Approach 2
vaccine_data <- read_csv("~/264_spring_2025/Data/vaccinations_2021.csv")   ②

# Approach 3
vaccine_data <- read_csv("https://joeroith.github.io/264_spring_2025/Data/vaccinations_2021.c

# Approach 4
vaccine_data <- read_csv("https://raw.githubusercontent.com/joeroith/264_spring_2025/refs/hea
```

① Approach 1: create a Data folder in the same location where this .qmd file resides, and then store vaccinations_2021.csv in that Data folder
② Approach 2: give R the complete path to the location of vaccinations_2021.csv, starting with Home (~)
③ Approach 3: link to our course webpage, and then know we have a Data folder containing all our csvs
④ Approach 4: navigate to the data in GitHub, hit the Raw button, and copy that link

A recent Stat 272 project examined determinants of covid vaccination rates at the county level. Our data set contains 3053 rows (1 for each county in the US) and 14 columns; here is a quick description of the variables we'll be using:

- `state` = state the county is located in
- `county` = name of the county
- `region` = region the state is located in
- `metro_status` = Is the county considered "Metro" or "Non-metro"?
- `rural_urban_code` = from 1 (most urban) to 9 (most rural)
- `perc_complete_vac` = percent of county completely vaccinated as of 11/9/21
- `tot_pop` = total population in the county
- `votes_Trump` = number of votes for Trump in the county in 2020
- `votes_Biden` = number of votes for Biden in the county in 2020
- `perc_Biden` = percent of votes for Biden in the county in 2020
- `ed_somecol_perc` = percent with some education beyond high school (but not a Bachelor's degree)
- `ed_bachormore_perc` = percent with a Bachelor's degree or more
- `unemployment_rate_2020` = county unemployment rate in 2020
- `median_HHincome_2019` = county's median household income in 2019

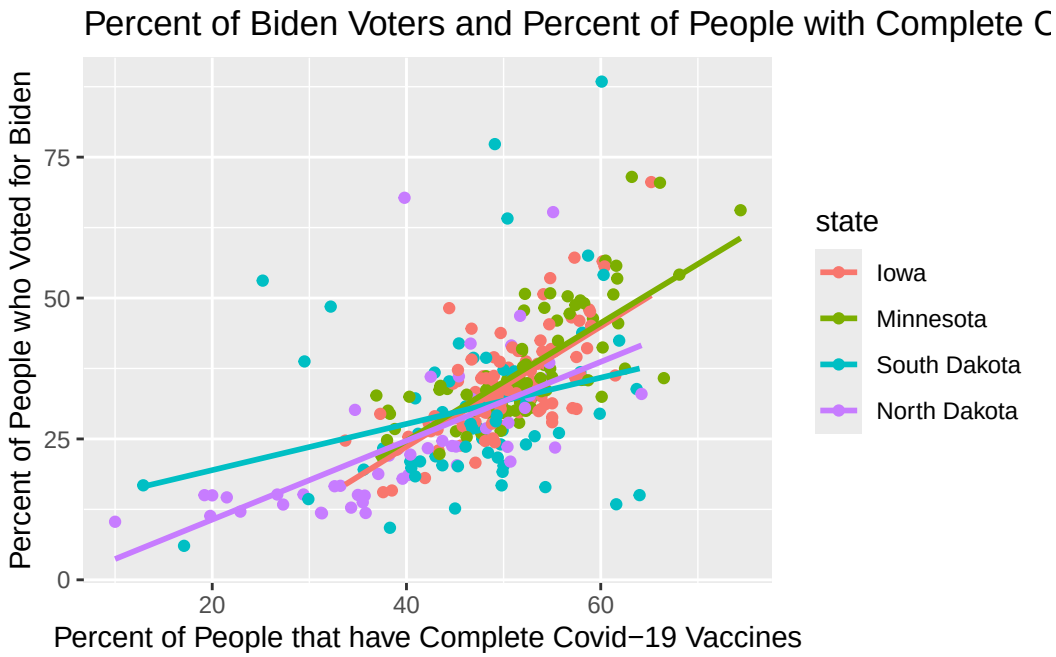
1. Consider only Minnesota and its surrounding states (Iowa, Wisconsin, North Dakota, and South Dakota). We want to examine the relationship between the percentage who voted for Biden and the percentage of complete vaccinations by state. Generate two plots to examine this relationship:

a) A scatterplot with points and smoothers colored by state. Make sure the legend is ordered in a meaningful way, and include good labels on your axes and your legend. Also leave off the error bars from your smoothers.

```
vaccine_data |>
  filter(state %in% c("Minnesota", "Iowa", "Wisonsin", "North Dakota", "South Dakota")) |>
  mutate(state = fct_reorder2(state, perc_Biden, perc_complete_vac)) |>

  ggplot(aes(x = perc_complete_vac, y = perc_Biden, color = state)) +
  geom_point() +
  geom_smooth(se = FALSE, method = lm) +
  labs(x = "Percent of People that have Complete Covid-19 Vaccines", y = "Percent of People
```

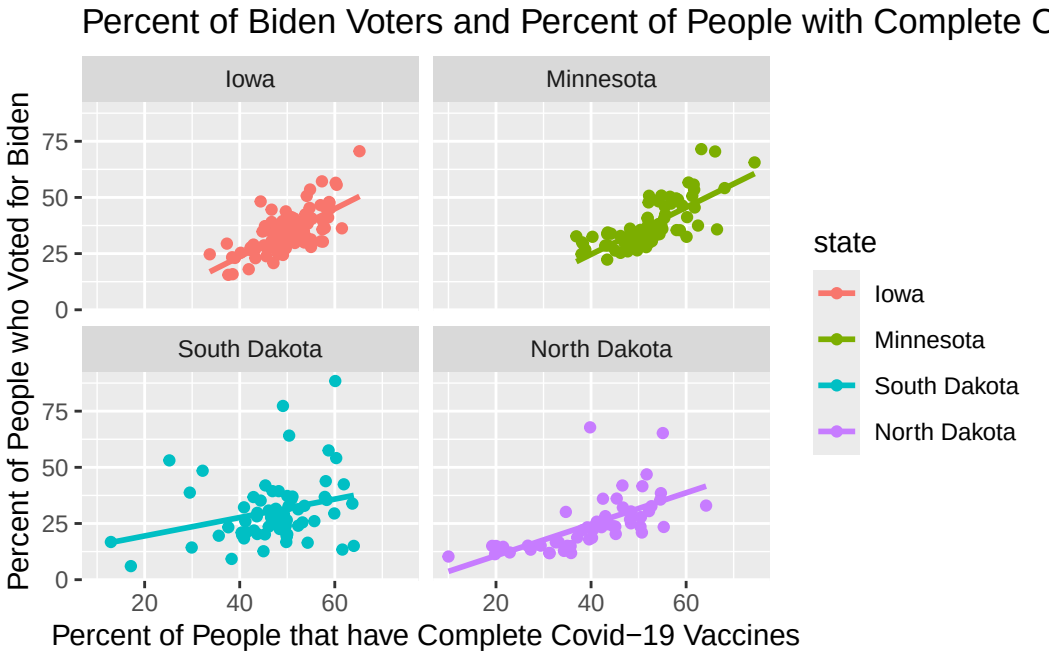
```
`geom_smooth()` using formula = 'y ~ x'
```



b) One plot per state containing a scatterplot and a smoother.

```
vaccine_data |>
  filter(state %in% c("Minnesota", "Iowa", "Wisonsin", "North Dakota", "South Dakota")) |>
  mutate(state = fct_reorder2(state,  perc_Biden, perc_complete_vac)) |>
  ggplot(aes(x = perc_complete_vac, y = perc_Biden, color = state)) +
  geom_point() +
  geom_smooth(se = FALSE, method = lm) +
  facet_wrap(~state) +
  labs(x = "Percent of People that have Complete Covid-19 Vaccines", y = "Percent of People w
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Describe which plot you prefer and why. What can you learn from your preferred plot?

I prefer the second plot because it is easier to see the trends in each of the states.

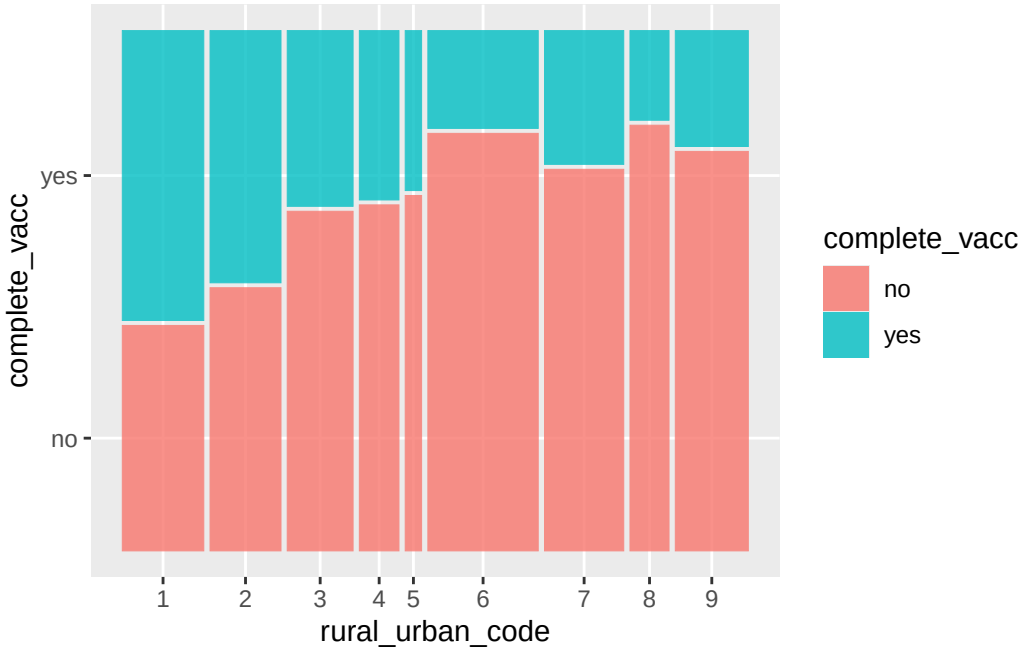
4. Produce 3 different plots for illustrating the relationship between the rural_urban_code and percent vaccinated. Hint: you can sometimes turn numeric variables into categorical variables for plotting purposes (e.g. `as.factor()`, `ifelse()`).

```
vaccine_data |>
  mutate(complete_vacc = ifelse(perc_complete_vac > 50, "yes", "no")) |>
  ggplot() +
  geom_mosaic(aes(x = product(rural_urban_code), fill = complete_vacc))
```

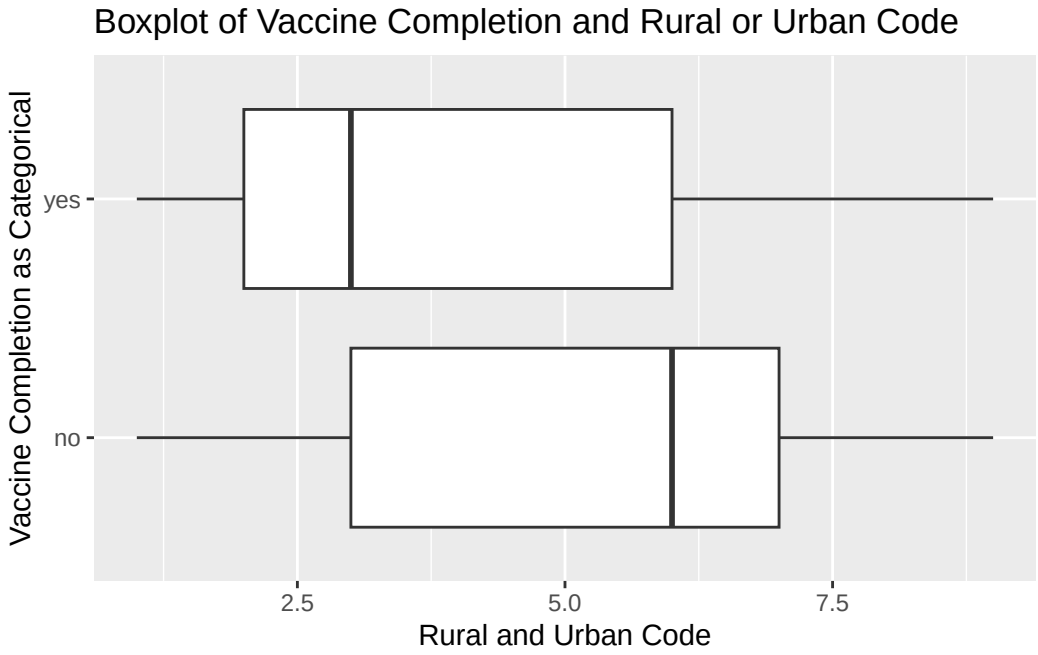
```
Warning: The `scale_name` argument of `continuous_scale()` is deprecated as of ggplot2
3.5.0.
```

```
Warning: The `trans` argument of `continuous_scale()` is deprecated as of ggplot2 3.5.0.
i Please use the `transform` argument instead.
```

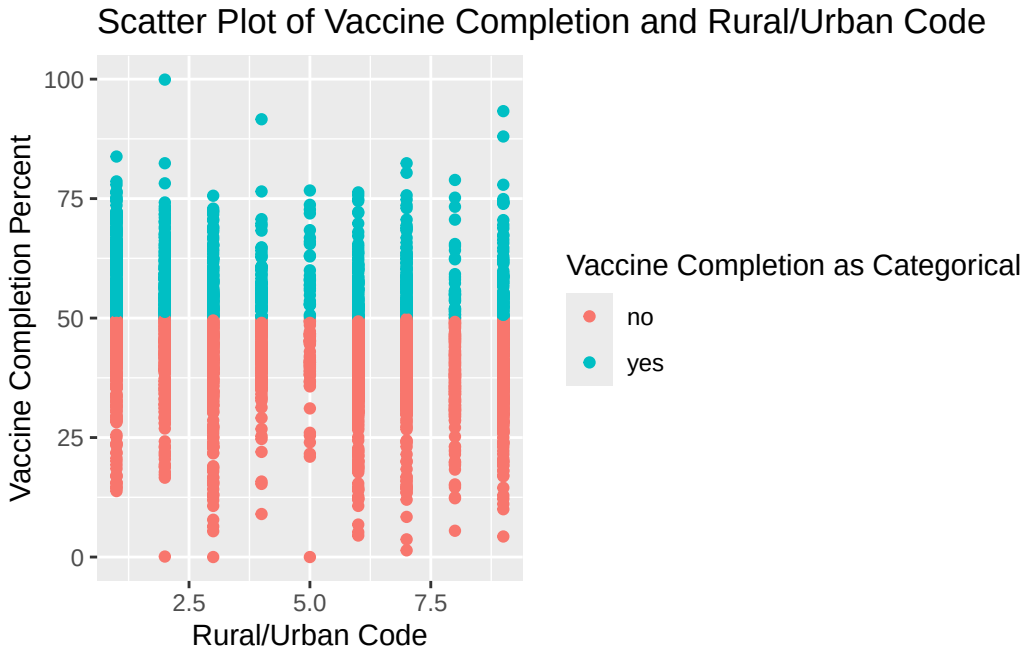
```
Warning: `unite_()` was deprecated in tidyr 1.2.0.
i Please use `unite()` instead.
i The deprecated feature was likely used in the ggmosaic package.
  Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
```




```
vaccine_data |>
  mutate(complete_vacc = ifelse(perc_complete_vac > 50, "yes", "no")) |>
  ggplot(aes(x = rural_urban_code, y = complete_vacc)) +
  geom_boxplot() +
  labs(x = "Rural and Urban Code", y= "Vaccine Completion as Categorical", title = "Boxplot
```




```
vaccine_data |>
  mutate(complete_vacc = ifelse(perc_complete_vac > 50, "yes", "no"))|>
  ggplot(aes(x = rural_urban_code, y = perc_complete_vac, color = complete_vacc)) +
  geom_point() +
  labs(title = "Scatter Plot of Vaccine Completion and Rural/Urban Code", x = "Rural/Urban Co
```



State your favorite plot, why you like it better than the other two, and what you can learn from your favorite plot. Create an alt text description of your favorite plot, using the Four Ingredient Model. See this link for reminders and references about alt text.

My favorite of these plots is the mosaic plot because it is the best at reprsenting the percentages of each rural/urban code and the vaccination status of the people in the area.

This is a mosaic plot that is representing the relationship between, rural_urban_code on the x-axis and complete_vacc on the y-axis. Complete vaccine is a categorical variable that has values of yes or no and rural_urban_code has values of 1-9, 1 being urban and 9 being rural. We can see that urban areas have more people that are fully vaccinated then rural areas.

5. BEFORE running the code below, sketch the plot that will be produced by R. AFTER running the code, describe what conclusion(s) can we draw from this plot?

```
vaccine_data |>
  filter(!is.na(perc_Biden)) |>
  mutate(big_states = fct_lump(state, n = 10)) |>
  group_by(big_states) |>
  summarize(IQR_Biden = IQR(perc_Biden)) |>
  mutate(big_states = fct_reorder(big_states, IQR_Biden)) |>
  ggplot() +
    geom_point(aes(x = IQR_Biden, y = big_states))
```

from this we can see that the IQR of Biden voters increases

6. In this question we will focus only on the 12 states in the Midwest (i.e. where region == "Midwest").

a) Create a tibble with the following information for each state. Order states from least to greatest state population.

```
midwest_vaccines <- vaccine_data |>
  filter(region == "Midwest") |>
  group_by(state) |>
  summarize(rural_urban = n_distinct(rural_urban_code),
            population = sum(tot_pop),
            prop_metro = mean(metro_status == "Metro"),
            unemployment_median = median(unemployment_rate_2020)) |>
  arrange(population)
```

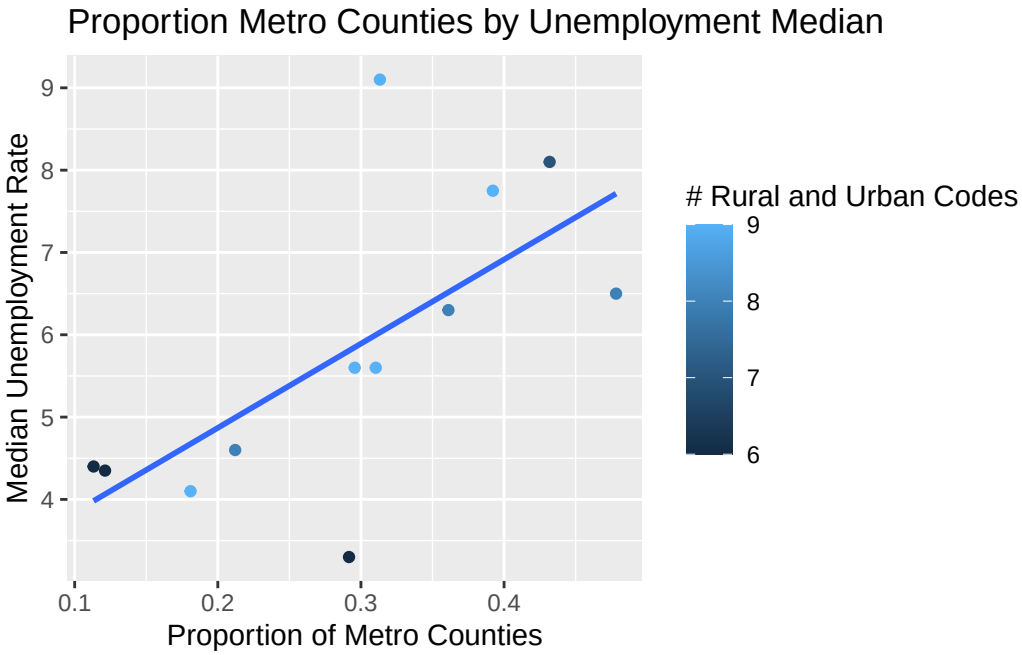
- number of different `rural_urban_code`s represented among the state's counties (there are 9 possible)
- total state population
- proportion of Metro counties
- median unemployment rate

b) Use your tibble in (a) to produce a plot of the relationship between proportion of Metro counties and median unemployment rate. Points should be colored by the number of different `rural_urban_code`s in a state, but a single linear trend should be fit to all points. What can you conclude from the plot?

```
midwest_vaccines |>
  ggplot(aes(x = prop_metro, y = unemployment_median, color = rural_urban)) +
  geom_point() +
  geom_smooth(aes(), se = FALSE, method = lm) +
  labs(x = "Proportion of Metro Counties", y = "Median Unemployment Rate", color = "# Rural a
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: The following aesthetics were dropped during statistical transformation:
colour.
i This can happen when ggplot fails to infer the correct grouping structure in
  the data.
i Did you forget to specify a `group` aesthetic or to convert a numerical
  variable into a factor?
```



We can see that there is weak positive linear trend with the proportion of metro counties having a higher unemployment rate.

In this next section, we consider a few variables that could have been included in our data set, but were NOT. Thus, you won't be able to write and test code, but you nevertheless should be able to use your knowledge of the tidyverse to answer these questions.

Here are the hypothetical variables:

- HR_party = party of that county's US Representative (Republican, Democrat, Independent, Green, or Libertarian)
- people_per_MD = number of residents per doctor (higher values = fewer doctors)
- perc_over_65 = percent of residents over 65 years old
- perc_white = percent of residents who identify as white

8. Hypothetical R chunk #1:

```
# Hypothetical R chunk 1
temp <- vaccine_data |>
  mutate(new_perc_vac = ifelse(perc_complete_vac > 95, NA, perc_complete_vac),
         MD_group = cut_number(people_per_MD, 3)) |>
  group_by(MD_group) |>
  summarise(n = n(),
            mean_perc_vac = mean(new_perc_vac, na.rm = TRUE),
            mean_white = mean(perc_white, na.rm = TRUE))
```

a) Describe the tibble `temp` created above. What would be the dimensions? What do rows and columns represent?

Temp is mutating the tibble to include 4 new variables that are describing the percentage of people with a covid vaccine that is at least 95 percent complete. Then adding another variable MD_group is making 3 groups with an equal number of observations and then mean_perc_vacc is taking the mean of people with 95% vaccinations status and mean_white is taking the mean of the percentatge of white people. So the tibble would have the dimensions 3053 x 22. The rows are each observation and the columns are each of the variables that we collected or produced by mutating the data.

b) What would happen if we replaced `new_perc_vac = ifelse(perc_complete_vac > 95, NA, perc_complete_vac)` with `new_perc_vac = ifelse(perc_complete_vac > 95, perc_complete_vac, NA)`?

If we replaced this it would give na for everyone that has the vaccination status of 95 or above and then give us the vaccination status for everyone with below a 95.

c) What would happen if we replaced `mean_white = mean(perc_white, na.rm = TRUE)` with `mean_white = mean(perc_white)`?

We would have a lot of na values because it is not getting rid of the people that don't identify as white.

d) What would happen if we removed `group_by(MD_group)`?

It would summarizes the entirety of the data set instead of each MD_group individually.

9. Hypothetical R chunk #2:

```
# Hypothetical R chunk 2
ggplot(data = vaccine_data) +
  geom_point(mapping = aes(x = perc_over_65, y = perc_complete_vac,
                           color = HR_party)) +
  geom_smooth()

temp <- vaccine_data |>
  group_by(HR_party) |>
  summarise(var1 = n()) |>
  arrange(desc(var1)) |>
  slice_head(n = 3)

vaccine_data |>
  ggplot(mapping = aes(x = fct_reorder(HR_party, perc_over_65, .fun = median),
```

```
                        y = perc_over_65)) +
    geom_boxplot()
```

a) Why would the first plot produce an error?

It would produce an error because the aes are in geom_point, geom_smooth has no aes to plot.

b) Describe the tibble `temp` created above. What would be the dimensions? What do rows and columns represent?

temp is creating a tibble that is 3 x 1 which is grouping by HR_party which there are 5 different observations and then it is counting var1 and arranging in descending order then only selecting the top 3 options.

c) What would happen if we replaced `fct_reorder(HR_party, perc_over_65, .fun = median)` with `HR_party`?

It would create a boxplot in the descending order of people over 65 and hr_party.

10. Hypothetical R chunk #3:

```
# Hypothetical R chunk 3
vaccine_data |>
  filter(!is.na(people_per_MD)) |>
  mutate(state_lump = fct_lump(state, n = 4)) |>
  group_by(state_lump, rural_urban_code) |>
  summarise(mean_people_per_MD = mean(people_per_MD)) |>
  ggplot(mapping = aes(x = rural_urban_code, y = mean_people_per_MD,
      colour = fct_reorder2(state_lump, rural_urban_code, mean_people_per_MD))) +
    geom_line()
```

a) Describe the tibble piped into the ggplot above. What would be the dimensions? What do rows and columns represent?

this would be a tibble of 4 x 3 with the mutate making a variable that is lumping the states into 4 rows then grouping by state_lump and rural_urban_code the it is creating a new variable tha tis the mean of people_per_md

b) Carefully describe the plot created above.

This is a line plot that is showing the rural_urban_codes on the x-axis and the mean_people_per_md on the y-axis it is then colored by the reordered plot with the states that have the highest state_lump then by ruralubrancode.

c) What would happen if we removed `filter(!is.na(people_per_MD))`?

it would produce an error because there would be nas in teh people per md and we can't take the mean of NA's.

d) What would happen if we replaced `fct_reorder2(state_lump, rural_urban_code, mean_people_per_MD)` with `state_lump`?

it would just color by the state_lump which would