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.

# 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

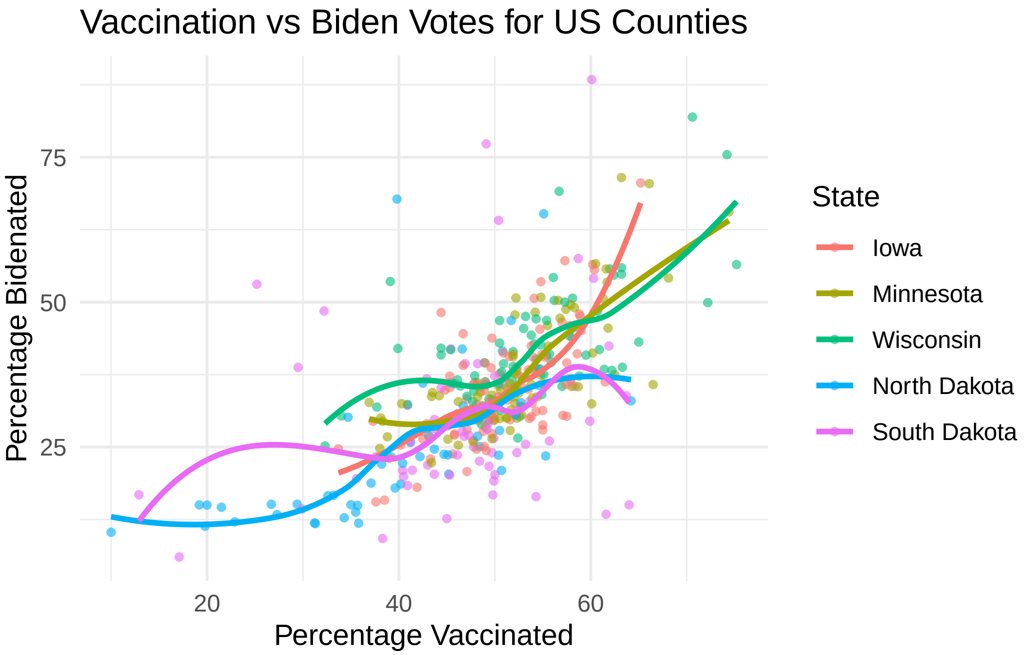
1. Consider only Minnesota and its surrounding states (Iowa, Wisconsin, North Dakota, and South Dakota). We want to examine the relationship between the percentage who voted for Biden and the percentage of complete vaccinations by state. Generate two plots to examine this relationship:

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","Wisconsin","Iowa","North Dakota","South Dakota")) |>
  mutate(state = fct_reorder2(state, perc_complete_vac, perc_Biden)) |>

  ggplot(aes(x = perc_complete_vac, y = perc_Biden, color = state)) +
    geom_point(size = 1, alpha = .6) +
    geom_smooth(method = "loess", se = FALSE) +
    theme_minimal() +
    labs(
      title = "Vaccination vs Biden Votes for US Counties",
      x = "Percentage Vaccinated",
      y = "Percentage Bidenated",
      color = "State"
    )#+
```

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



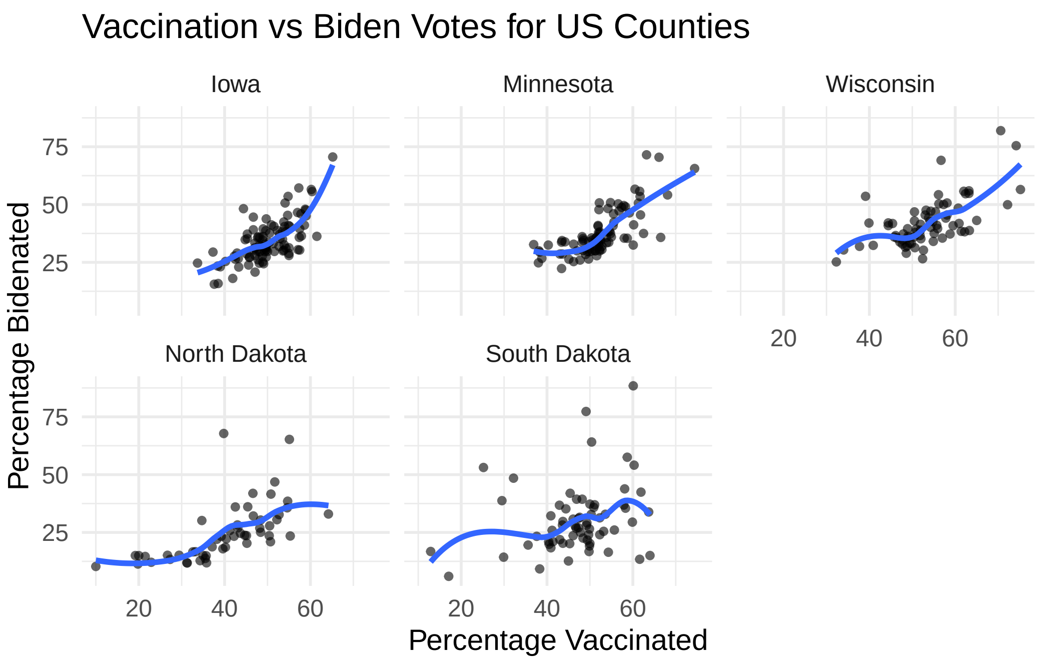
```
  #scale_color_viridis_d(option="magma")
```

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

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

  ggplot(aes(x = perc_complete_vac, y = perc_Biden)) +
    geom_point(size = 1, alpha = .6) +
    geom_smooth(method = "loess", se = FALSE) +
    facet_wrap(~state) +
    theme_minimal() +
    labs(
      title = "Vaccination vs Biden Votes for US Counties",
      x = "Percentage Vaccinated",
      y = "Percentage Bidenated",
      color = "State"
    )
```

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



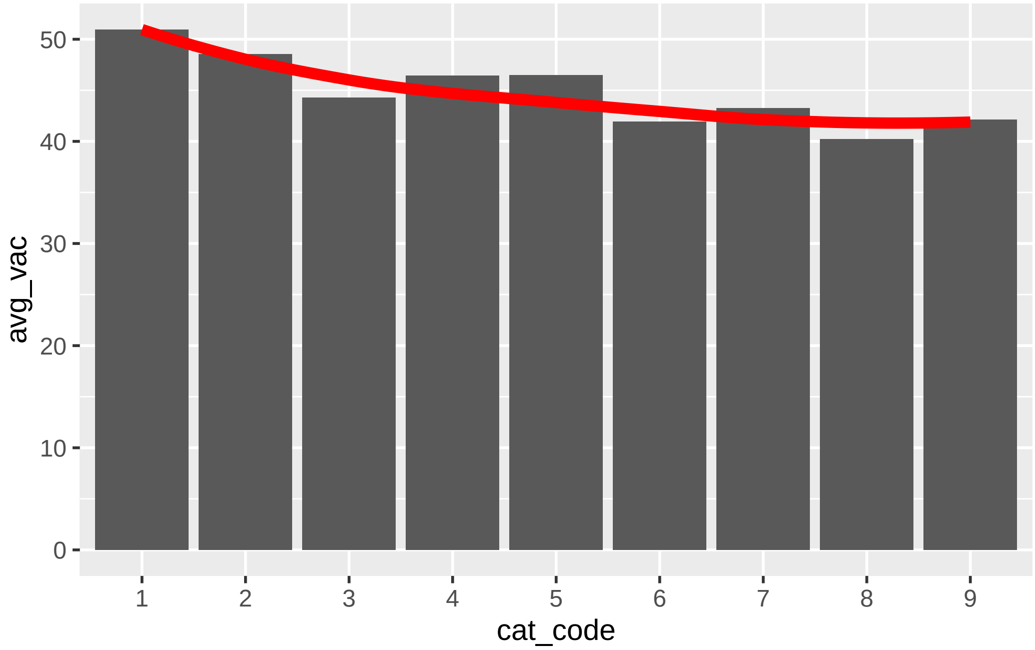
Describe which plot you prefer and why. What can you learn from your preferred plot?

I prefer the faceted plot because, without separating the information, things are too cluttered and busy to understand trends on a state-by-state level. The main thing I learn is that sometimes it may take more than one plot to convey all the relevant information in a digestible way.

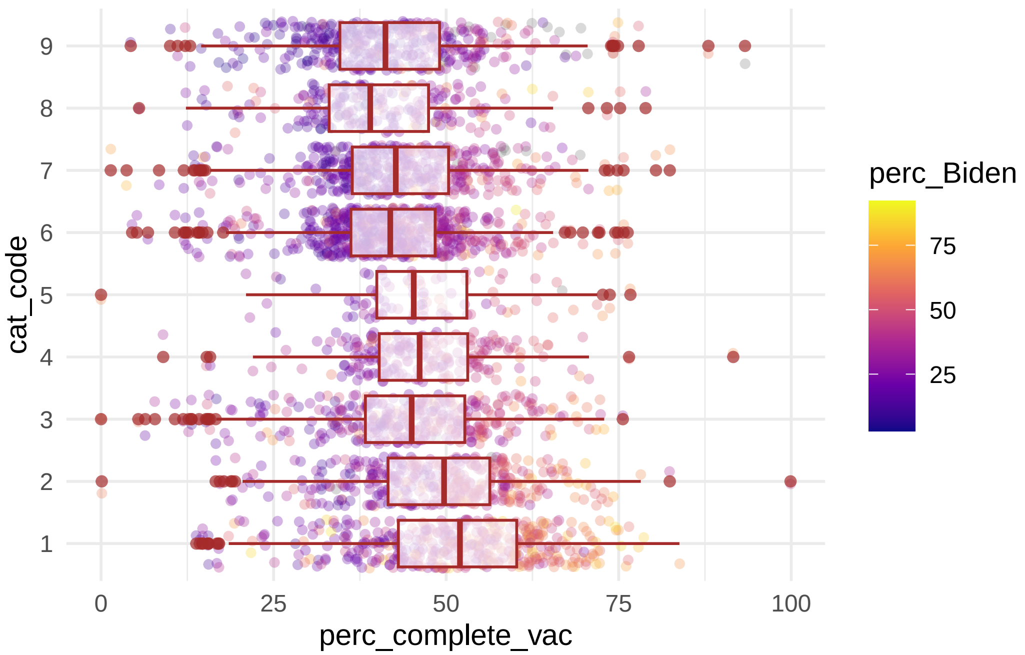
4. Produce 3 different plots for illustrating the relationship between the rural_urban_code and percent vaccinated. Hint: you can sometimes turn numeric variables into categorical variables for plotting purposes (e.g. `as.factor()`, `ifelse()`).

```
# Plot 1
vaccine_data |>
  mutate(cat_code = as.factor(rural_urban_code)) |>
  group_by(cat_code) |>
  summarize(avg_vac = mean(perc_complete_vac)) |>
  ggplot() +
    geom_col(aes(x = cat_code, y = avg_vac)) +
    geom_smooth(data = vaccine_data, aes(x = rural_urban_code, y = perc_complete_vac), metho
```

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

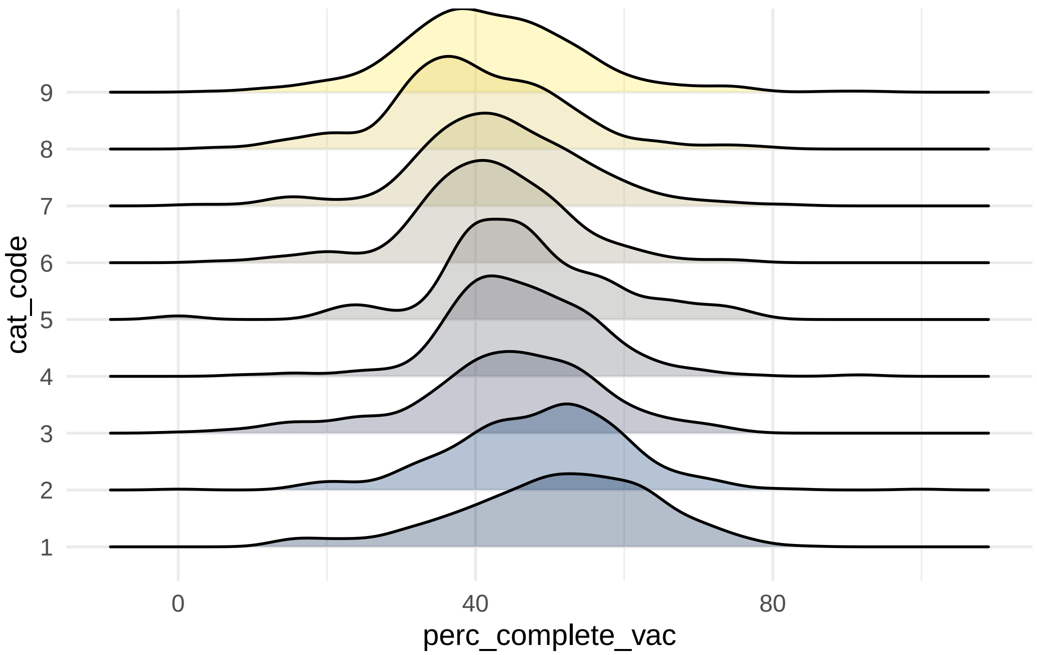


```
# Plot 2
                                                                                        ②
vaccine_data |>
  mutate(cat_code = as.factor(rural_urban_code)) |>
  ggplot(aes(y = cat_code, x = perc_complete_vac, color = perc_Biden)) +
    geom_jitter(size = 1.2, alpha = .3) +
    geom_boxplot(alpha = .7, color = "brown") +
    theme_minimal() +
    scale_color_viridis_c(option="C")
```



```
# Plot 3
library(ggridges)
vaccine_data |>
  mutate(cat_code = as.factor(rural_urban_code)) |>
  ggplot(aes(x = perc_complete_vac, y = cat_code, fill = cat_code)) +
    geom_density_ridges(alpha = .3) +
    theme_minimal() +
    scale_fill_viridis_d(option="E") +
    theme(legend.position = "none")
```

```
Picking joint bandwidth of 3.04
```



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 plot of these three is plot number 2. I like it because I think that, although it is a little harder to read initially, it is capable of conveying a lot more information without being outragously confusing. Part of this is that it is able to express a third variable via coloring, but the more important part is illustrating variability; this is one relationship where it is important to avoid deficit or over-generalized thinking, so I like that the second plot is able to show the high amount of variation within each urban code category. Along that line of thought, one takeaway is that layering a summarized visualization ABOVE a point-by-point visualization can be useful for fair expressions of the data.

Alt text: Here is a layered scatterplot and box plot. Each data point is a county in the United States. The x axis of this plot expresses the COVID vaccination percentage within each county, ranging from 0 to 100 percent, while the y axis is ordered in nine discrete categories partaining to the nine urban rural codes, with 1 being the most urban and 9 being the most rural. The jittered scatter plot shows a high degree of variability within every rural urban category, while the box plots show a small positive relationship between amount of urban-ness and percentage vaccinated. Furthermore, each county is colored according to the percentage which voted for Biden, and it is visually evident that there is a positive relationship between percentage vaccinated and percentage of votes for Biden.

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))
```

I would say that there are not a lot of useful conclusions to be extracted from this figure. Most of the ten states with the highest number of counties have a lower IQR, and thus lower variability, in Biden votes than the IQR of the states lumped into the "Other" category, but it's hard to say what that really means. The number of counties a state has is not necessarily relevant to its size or its population (for example, California is not among the top ten, and neither is New York), and the variability in Biden voting per county doesn't tell us much either.

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.

- 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

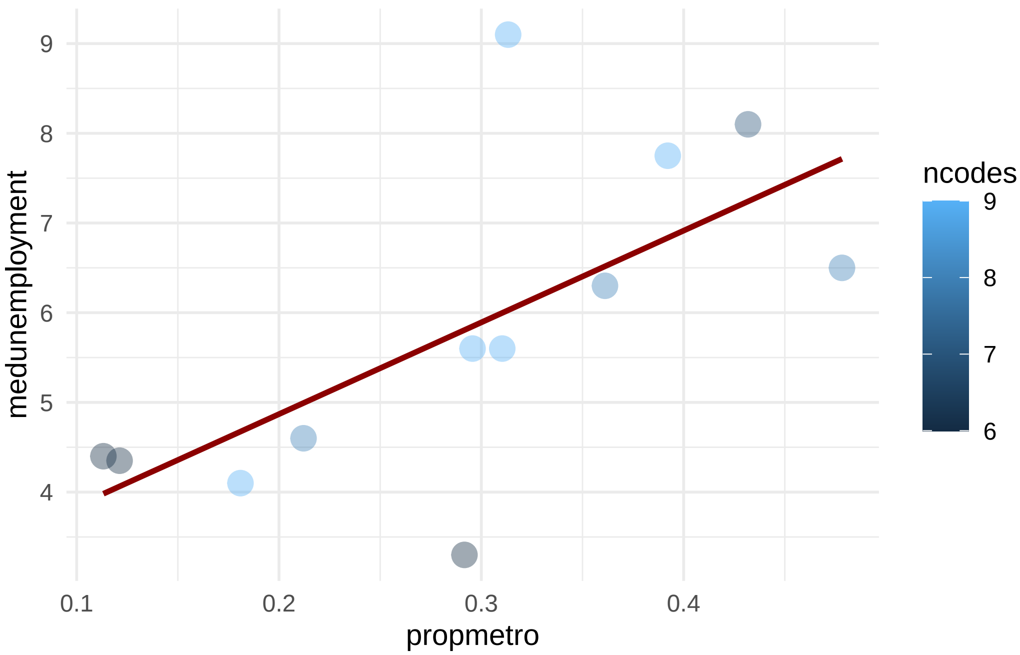
```
(midwests <- vaccine_data |>
  filter(region == "Midwest") |>
  group_by(state) |>
  summarize(
    ncodes = n_distinct(rural_urban_code),
    pop = sum(tot_pop),
    propmetro = sum(metro_status == "Metro") / n(),
    medunemployment = median(unemployment_rate_2020)
  ) |>
   arrange(pop)) |>
  print()
```

```
# A tibble: 12 x 5
   state        ncodes      pop propmetro medunemployment
   <chr>         <int>    <dbl>     <dbl>           <dbl>
 1 North Dakota      6   762062     0.113            4.4
 2 South Dakota      6   884659     0.121            4.35
 3 Nebraska          6  1261262     0.292            3.3
 4 Kansas            9  2913314     0.181            4.1
 5 Iowa              8  3155070     0.212            4.6
 6 Minnesota         9  5639632     0.310            5.6
 7 Wisconsin         8  5822434     0.361            6.3
 8 Missouri          9  6137428     0.296            5.6
 9 Indiana           8  6732219     0.478            6.5
10 Michigan          9  9986857     0.313            9.1
11 Ohio              7 11689100     0.432            8.1
12 Illinois          9 12671821     0.392            7.75
```

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_codes` in a state, but a single linear trend should be fit to all points. What can you conclude from the plot?

```
midwests |>
  ggplot(aes(x = propmetro, y = medunemployment)) +
    geom_point(aes(color = ncodes), size = 4, alpha = .4) +
    geom_smooth(method="lm", se=FALSE, color = "darkred") +
    theme_minimal()
```

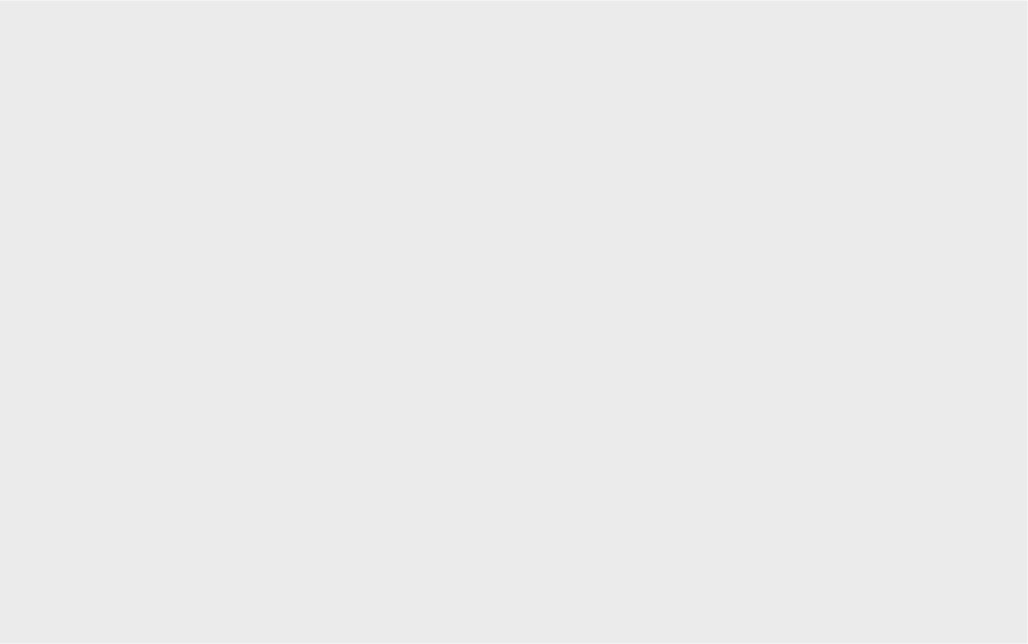
```
`geom_smooth()` using formula = 'y ~ x'
```



From this plot I can conclude that there may be a positive linear relationship between the proportion of metro counties in a state and the median unemployment rate between its counties. There does not appear to be a relationship between the diversity of rural urban codes and the other variables. There are also a few unusual points, albeit without a lot of leverage.

7. Generate an appropriate plot to compare vaccination rates between two subregions of the US: New England (which contains the states Maine, Vermont, New Hampshire, Massachusetts, Connecticut, Rhode Island) and the Upper Midwest (which, according to the USGS, contains the states Minnesota, Wisconsin, Michigan, Illinois, Indiana, and Iowa). What can you conclude from your plot?

```
vaccine_data |>
  mutate(neoregion = ifelse(
    state %in% c("Maine", "Vermont", "New Hampshire", "Massachusetts", "Connecticut", "Rhode
      state %in% c("Minnesota", "Wisconsin", "Michigan", "Illinois", "Indiana", "Iowa"), "Up
    ))
  ) |>
  filter(neoregion != "X") |>
  ggplot(aes(
    # it was at this moment that Dan realized he did not need to do this problem.
  ))
```



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:

```r
# 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?

The new tibble will have four columns: MD_group, a category representing the amount of people per MD in a county; mean_perc_vac, the average percentage of people vaccinated in the counties of a particular MD_group, except with anything greater than 95% removed; and mean_white, the average percentage of the white demographic within counties of a particular MD_group. It will have three rows, one for each of the MD_groups.

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)`?

Then only those counties with vaccination percentages greater than 95% would be considered, rather than excluded.

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

Then if there were any counties with an NA value for `perc_white`, the value for mean_white would also become NA.

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

Then we would get one row with the count and averages for the whole dataset.

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?

Because the geom_smooth() layer has not inherited any mapping, so it has no idea what to do.

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

It would have three rows and two columns. The rows would represent the three parties with the most representation across all the counties. One column would be the party name, the other the number of counties with that party as a representative.

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

Then the order of the boxplots would likely appear somewhat random; the way the code is currently written, they will be in order of their median perc_over_65 value, making it easier to see a relationship across the plots.

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?

Assuming that the four states with the most counties have all nine urban rural codes, the dataset would have 45 rows, each one representing a rural urban code for one of the four most countied states or the other category. The columns would be three: state/other, code, and average people per doctor.

b) Carefully describe the plot created above.

The plot would have colored lines for each of the four states and also the other category. The lines would show the relationship between urbanness and people per doctor for each of the state categories.

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

If there were any NAs in people_per_MD, then mean_people_per_MD would be summarized as NA.

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

Then the states in the legend would be in a seemingly random order, rather than being ordered according to where the lines for each one end up.