TA Session 10: Data Visualization – Tidyverse

Harris Coding Camp

Summer 2024

General Guidelines

You may encounter some functions we did not cover in the lectures. This will give you some practice on how to use a new function for the first time. You can try following steps:

1. Start by typing `?new_function` in your Console to open up the help page.
2. Read the help page of this `new_function`. The description might be a bit technical for now. That's OK. Pay attention to the Usage and Arguments, especially the argument `x` or `x,y` (when two arguments are required).
3. At the bottom of the help page, there are a few examples. Run the first few lines to see how it works.
4. Apply it in your questions.

It is highly likely that you will encounter error messages while doing this exercise. Here are a few steps that might help get you through it:

1. Locate which line is causing this error first.
2. Check if you have a typo in the code. Sometimes your group members can spot a typo faster than you.
3. If you enter the code without any typo, try googling the error message. Scroll through the top few links see if any of them helps.
4. Try working on the next few questions while waiting for help by TAs.

Data and background

1. Load `tidyverse`, `haven`, and `readxl` in your Rmd/script.

```
library(tidyverse)
library(haven)
library(readxl)
```

2. We'll use midwestern demographic data which is at this link. The dataset includes county level data for a single year. We call data this type of data "cross-sectional" since it gives a point-in-time cross-section of the counties of the midwest. (The world inequality data is "time-series" data).

3. Again, we'll work with data sets from `recent_college_grads.dta`, which you can download here. This is a data on college majors and earnings, specifically the data behind the FiveThirtyEight story "The Economic Guide To Picking A College Major".

4. Again, we'll be working with the `diamonds` dataset, which contains the prices and other attributes of almost 54,000 diamonds. See more details here.

I. Manipulating College Data

How do the distributions of median income compare across major categories?

A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value below which 20% of the observations may be found.

We will be working with the data set `recent_college_grads.dta`. There are three types of incomes reported in this data frame: `p25th`, `median`, and `p75th`. These correspond to the 25th, 50th, and 75th percentiles of the income distribution of sampled individuals for a given major.

We need to do a few things to answer this question "How do the distributions of median income compare across major categories?". First, we need to group the data by `major_category`. Then, we need a way to summarize the distributions of median income within these groups. This decision will depend on the shapes of these distributions. So first, we need to visualize the data.

1. Let's first take a look at the distribution of all median incomes using `geom_histogram`, without considering the major categories.

```
ggplot(data = ____,
       mapping = aes(x = median)) +
  geom_histogram()
```

2. Try binwidths of 1000 and 5000 and choose one. Explain your reasoning for your choice.

```
ggplot(data = ___,
       mapping = aes(x = median)) +
  geom_histogram(binwidth = ___)
```

We can also calculate summary statistics for this distribution using the `summarize` function:

```
college_recent_grads %>%  # replace the data name to yours
  summarize(min = min(median), max = max(median),
            mean = mean(median), med = median(median),
            sd = sd(median),
            q1 = quantile(median, probs = 0.25),
            q3 = quantile(median, probs = 0.75))
```

```
## # A tibble: 1 x 7
##     min    max   mean   med     sd    q1    q3
##   <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 22000 110000 40151. 36000 11470. 33000 45000
```

3. Based on the shape of the histogram you created in the previous part, determine which of these summary statistics above (min, max, mean, med, sd, q1, q3) is/are useful for describing the distribution. Write up your description and include the summary statistic output as well. You can pick single/multiple statistics and briefly explain why you pick it/them.

4. Next, we facet the plot by major category. Plot the distribution of `median` income using a histogram, faceted by `major_category`. Use the binwidth you chose in part 2.

```
ggplot(data = ___,
       mapping = aes(x = median)) +
  geom_histogram(bindwidth = ___) +
  facet_wrap(.~major_category)
```

What types of majors do women tend to major in?

First, let's create a new vector called `stem_categories` that lists the major categories that are considered STEM fields.

```
stem_categories <- c("Biology & Life Science",
                     "Computers & Mathematics",
                     "Engineering",
                     "Physical Sciences")
```

5. Then, we can use this to create a new variable in our data frame indicating whether a major is STEM or not. Complete the code.

```
college_recent_grads <-
  college_recent_grads %>%
  mutate(major_type = ifelse(...))
```

6. Create a scatterplot of median income vs. proportion of women in that major, colored by whether the major is in a STEM field or not. Describe the association between these three variables.

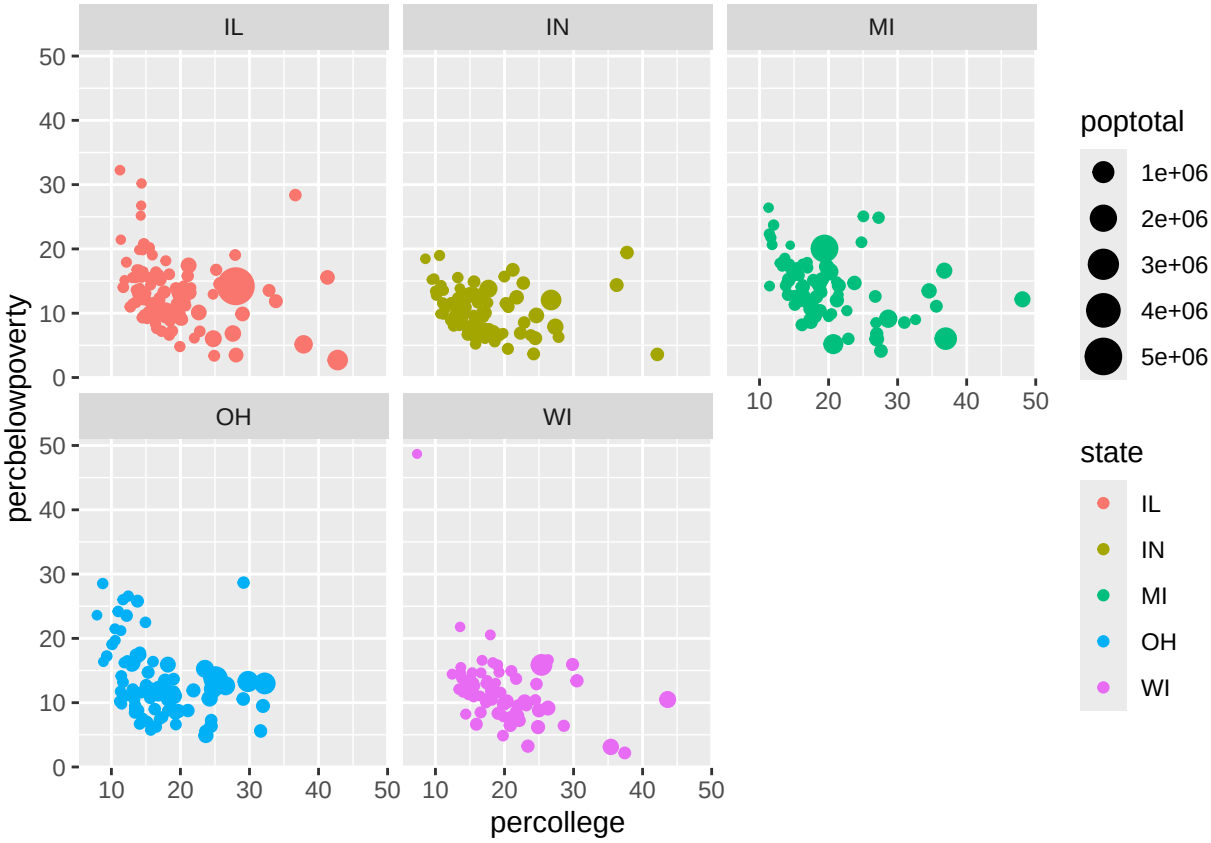
```
ggplot(data = ___,
       mapping = aes(x = ___,
                     y = ___,
       color = major_type)) +
  geom_point()
```

7. We can use the logical operators to also `filter` our data for STEM majors whose median earnings is less than median for all majors's median earnings, which we found to be $36,000 earlier. Your output should only show the major name and median, 25th percentile, and 75th percentile earning for that major and should be sorted such that the major with the lowest median earning is on top.

II. Manipulating Midwest Data

Recall `ggplot` works by mapping data to aesthetics and then telling `ggplot` how to visualize the aesthetic with `geoms`. Like so:

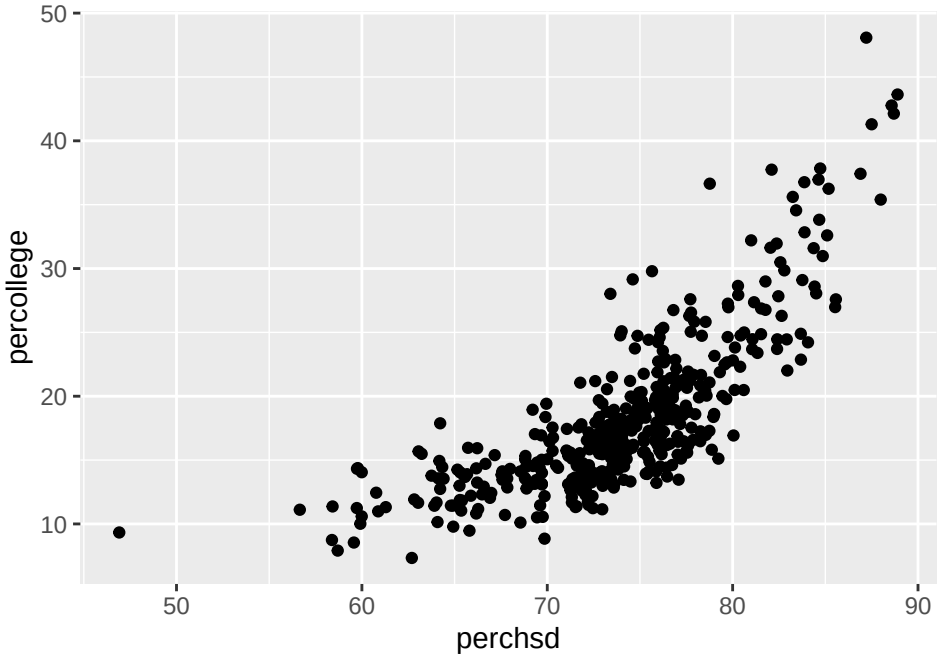
```
midwest %>%
  ggplot(aes(x = percollege,
             y = percbelowpoverty,
             color = state,
             size = poptotal)) +
    geom_point() +
    facet_wrap(vars(state))
```

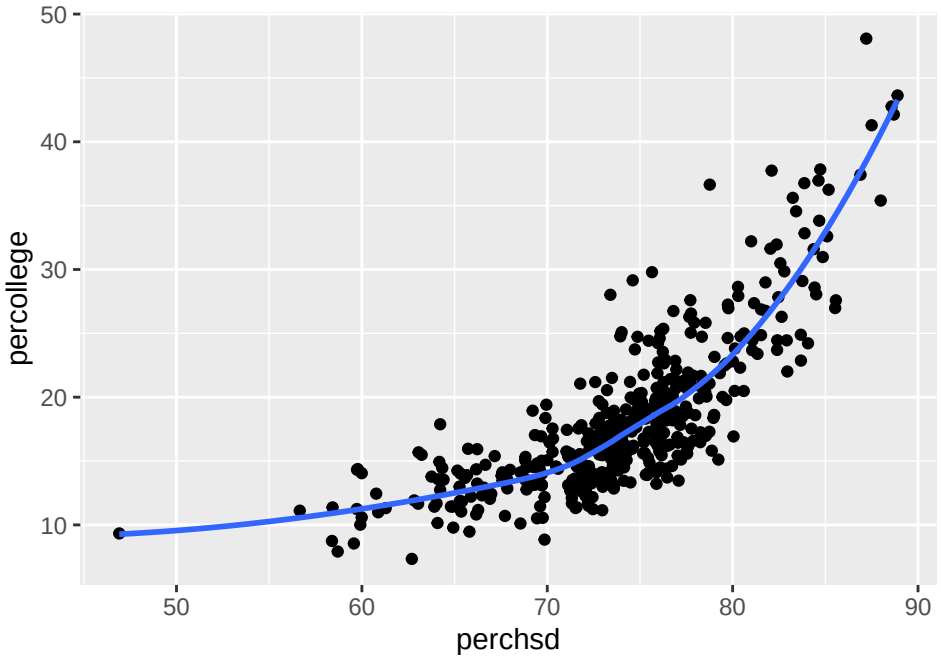


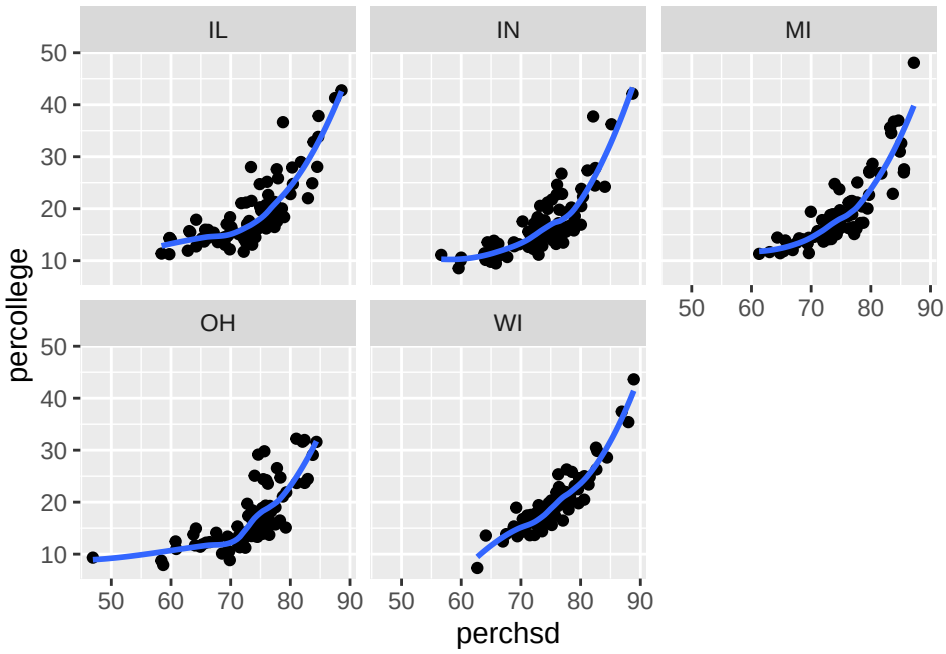
1. Which is more highly correlated with poverty at the county level, college completion rates or high school completion rates? Is it consistent across states? Change one line of code in the above graph.

`geoms`

For the following, write code to reproduce each plot using `midwest`:

1. 

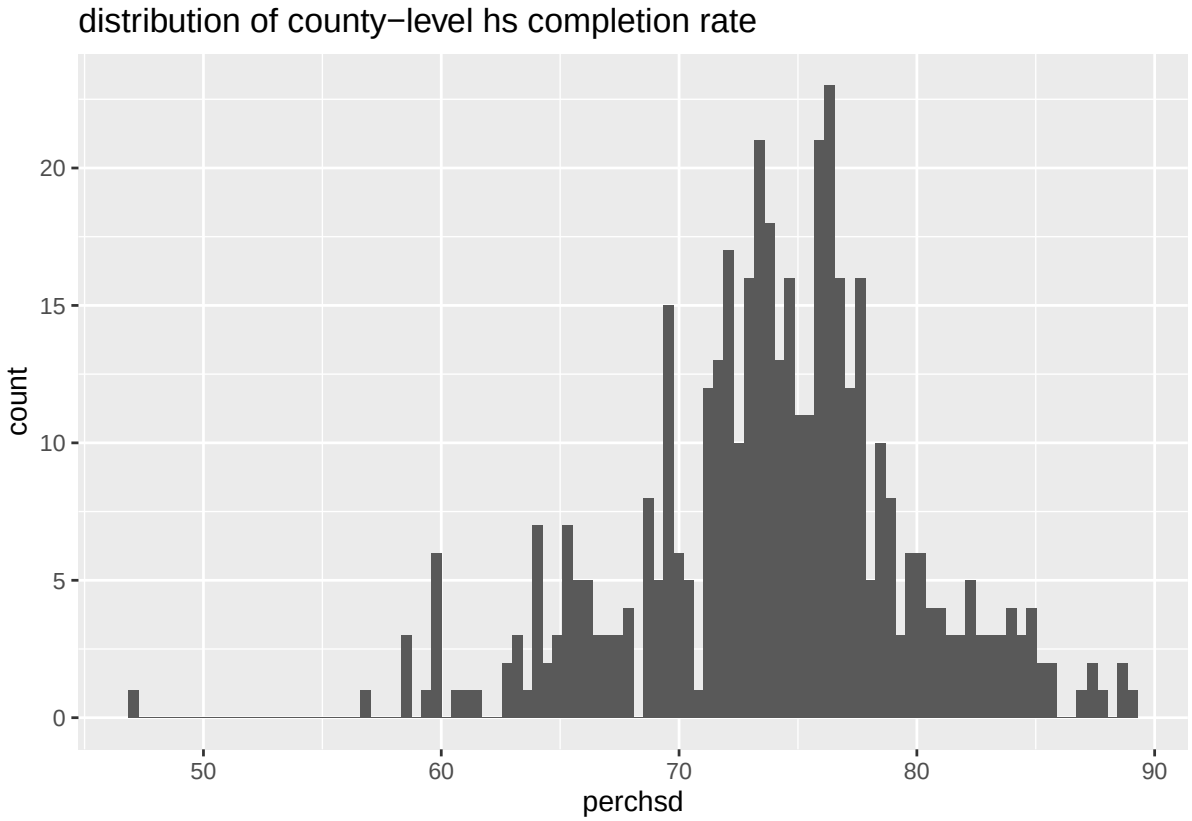
2. 

3. 

4. Use `geom_boxplot()` instead of `geom_point()` for "Asian population by metro status".

5. Histograms are used to visualize distributions. What happens when you change the `bins` argument? What happens if you leave the `bins` argument off?

```
midwest %>%
  ggplot(aes(x = perchsd)) +
    geom_histogram(bins = 100) +
    labs(title = "distribution of county-level hs completion rate")
```



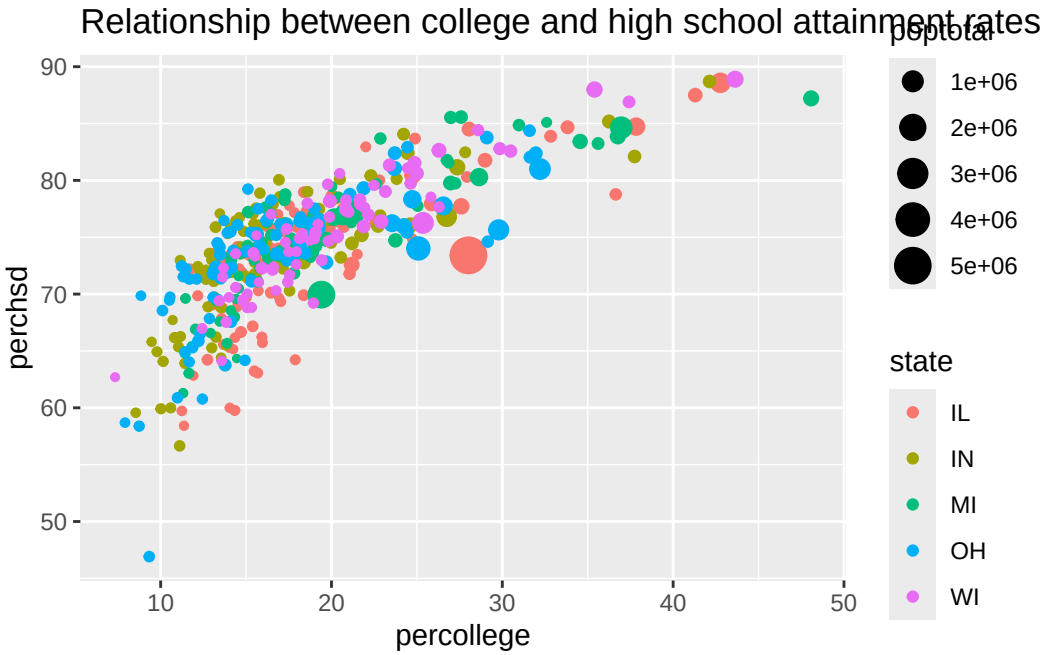
6. Remake "distribution of county-level hs completion rate" with `geom_density()` instead of `geom_histogram()`.

7. (Optional) Add a vertical line at the median `perchsd` using `geom_vline`. You can calculate the median directly in the ggplot code.

Aesthetics

For the following, write code to reproduce each plot using `midwest`.

1. Use `x`, `y`, `color` and `size`.



2. Add smooth lines. Get rid of the error around your smooth lines by adding the argument `se = F`.
3. Now try faceting with `facet_grid` and the code `facet_grid(col = vars(inmetro), rows = vars(state))` to your plot.
4. There's a `geom` called `geom_bar` that takes a dataset and calculates the count. Read the following code and compare it to the `geom_col` code above. Describe how `geom_bar()` is different than `geom_col`.

```
midwest %>%
 ggplot(aes(x = state, color = state)) +
   geom_bar()
```

Revisit the `diamonds` dataset

Just like TA session 6, we will again be working with the `diamonds` dataset, which contains the prices and other attributes of almost 54,000 diamonds.

- price: Price in US dollars.
- carat: Weight of the diamond.
- cut: Cut quality (ordered worst to best).
- color: Color of the diamond (ordered best to worst).
- clarity: Clarity of the diamond (ordered worst to best).
- x: Length in mm.
- y: Width in mm.
- z: Depth in mm.
- depth: Total depth percentage: 100 * z / mean(x, y)
- table: Width of the top of the diamond relative to the widest point.

1. Is there any relationship between `price` and `carat` of a diamond? What might explain that pattern? Run the code below and comment on the result.

```
ggplot(diamonds) +
  geom_point(aes(x = price, y = carat),
             color = "black", fill = "lightblue") +
  ggtitle("Diamonds Price vs Carat")
```

2. Redo part 1, separately for observations of each `cut` (Hint: add one line of code which includes `facet_grid()`). Is there any relationship between price and cut quality of a diamond? What might explain that pattern?

3. Redo part 2. Is there any relationship between `price` per `carat` (defined as `price` divided by `carat`) and `cut` of a diamond? Run the code below, and compare the result with the one from part 2.

```
ggplot(diamonds) +
  geom_point(aes(x = price, y = carat),
             color = "black", fill = "lightblue") +
  ggtitle("Diamonds Price vs Carat by Cut") +
  scale_x_log10() +
  facet_grid(. ~ cut)
```