Wrangling Data

Dominguez Center for Data Science Workshop

2025-03-05

Recap from last time

- Does anyone want to share a graph they made?
- Does anyone have questions?

Plan for today

- Cover a little more `ggplot2`.
- Start going through data wrangling in `R`.
- Remember that you can either fill in the "wrangling.qmd" file or follow along with the "wrangling_key.qmd" file.

Load Packages

The packages we need for our explorations today (`readr` for reading in data, `ggplot2` for graphing data, and `dplyr` for wrangling/summarizing the data) are part of a popular suite of packages called the `tidyverse`.

```
library(tidyverse)
library(ggrepel)
```

Data Background

We will return to the same dataset we saw last time. Here's the background and description of the variables.

In 2013, the government decided to make data about colleges more accessible so that students and parents could more easily compare schools. These data are called the "College Scorecard" data and the 2024 dataset contains 3,305 variables on 6,484 universities in the US!

I have filtered that 2024 dataset to only include schools which confer majority baccalaureate degrees and where the majority of those degrees are in the arts and sciences based on the Carnegie Classification system. In other words, I filtered the data down to the schools which are "similar" to Bucknell (including Bucknell itself) and picked out some variables for us to explore.

Data Dictionary

Below are the code names and descriptions of the variables in our dataset.

- `UNITID`: Unique identifier
- `INSTNM`: Name of institution
- `CITY`: City
- `STABBR`: State
- `HIGHDEG`: Highest degree awarded (0 = Non-degree grants, 1 = Certificate degree, 2 = Associate degree, 3 = Bachelor's degree, 4 = Graduate degree)
- `PREDDEG`: Predominant undergraduate degree awarded (0 = Not classified, 1 = Predominantly certificate-degree granting, 2 = Predominantly associate's-degree granting, 3 = Predominantly bachelor's-degree granting, 4 = Entirely graduate-degree granting)
- `CONTROL`: Ownership (1 = Public, 2 = Private non-profit, 3 = Private for-profit)
- `HBCU`: Flag for Historically Black College and University
- `TUITFTE`: Net tuition revenue per full-time equivalent student
- `AVGFACSAL`: Average faculty salary
- `ADM_RATE`: Admission rate
- `SATVR75`: 75th percentile of SAT scores at the institution (critical reading)
- `SATMT75`: 75th percentile of SAT scores at the institution (math)
- `ACTCM75`: 75th percentile of the ACT cumulative score

- `COSTT4_A`: The average annual total cost of attendance, including tuition and fees, books and supplies, and living expenses for all full-time, first-time, degree/certificate-seeking undergraduates who receive Title IV aid.
- `NPT4_PRIV`: The average annual total cost of attendance, including tuition and fees, books and supplies, and living expenses, minus the average grant/scholarship aid
- `UGDS`: Enrollment of undergraduate certificate/degree-seeking students
- `UG25ABV`: Percentage of undergraduates aged 25 and above
- `PCTFLOAN_DCS`: Percentage of degree/certificate-seeking undergraduate students awarded a federal loan
- `PCTPELL_DCS`: Percentage of degree/certificate-seeking undergraduate students awarded a Pell Grant
- `DEBT_MDN`: The median original amount of the loan principal upon entering repayment
- `C100_4`: Completion rate for first-time, full-time students at four-year institutions (100% of expected time to completion)
- `RET_FT4`: First-time, full-time student retention rate at four-year institutions
- `MD_EARN_WNE_5YR`: Median earnings of graduates working and not enrolled 5 years after completing

Load the Data

Run the following code to load and inspect the data.

```
# Load the data
colleges <- read_csv("data/ccbasic21.csv")


# Wrangling data (Will talk through soon!)
colleges <- colleges %>%
  mutate(Location = if_else(STABBR == "PA", "PA", "Not PA"),
         CONTROL_CAT = case_match(CONTROL,
                                  1 ~ "Public",
                                  2 ~ "Private non-profit"),
         HIGHDEG_CAT = case_match(HIGHDEG,
                                  3 ~ "Bachelor's degree",
                                  4 ~ "Graduate degree"))
```

Graphs Recap

Guiding Principle: We will map variables from the **data** to the **aes**thetic attributes (e.g. location, size, shape, color) of **geom**etric objects (e.g. points, lines, bars).

```
ggplot(data = ---, mapping = aes(---)) +
  geom_---(---)
```

What is data wrangling??

Data wrangling = any transformations done on the data.

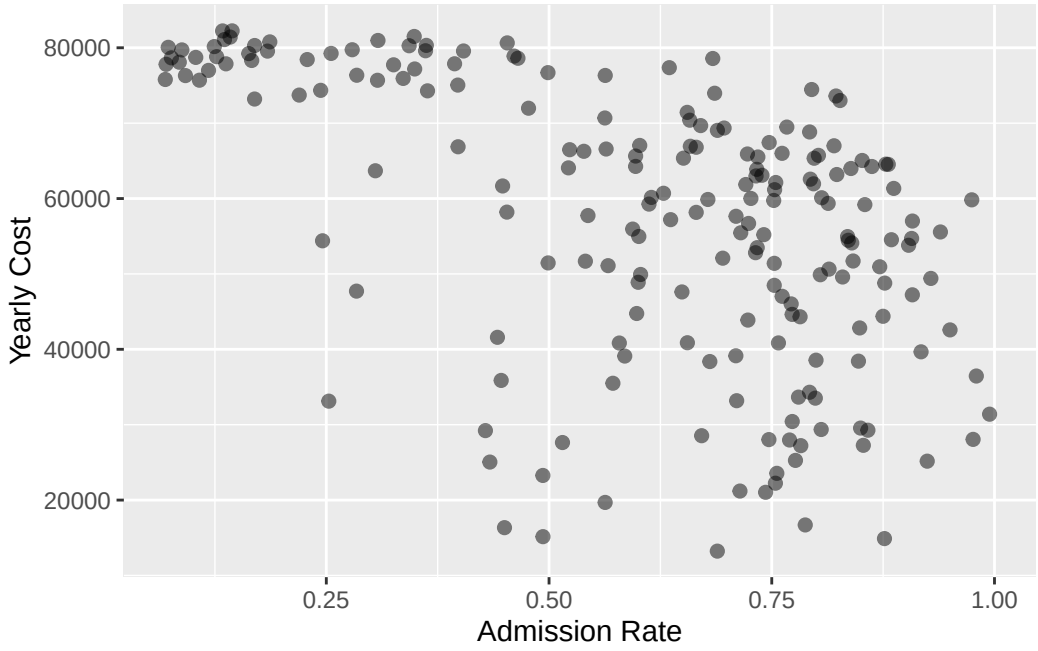
Examples:

- Summarizing the data by computing the mean of a variable.
- Counting the number of observations in the categories of a set of variables. (Excel users: Think pivot tables.)
- Dropping rows of the dataset that have missing values.
- Filtering down to just a subset of the data.
- Collapsing a categorical variable into fewer categories.
- Fixing how `R` stores a variable.
- Sorting the data by one of the variables.
- Joining multiple datasets together.
    - Won't see joins today but can learn about them here

Motivating Examples: Sprucing up our Graphs

How do I add Bucknell to my graph?

```
ggplot(data = colleges,
       mapping = aes(x = ADM_RATE,
                     y = COSTT4_A)) +
  geom_point(size = 2, alpha = .5) +
  labs(x = "Admission Rate", y = "Yearly Cost")
```



Option 1: Label all the schools.

Option 2: Create a new dataset that contains only Bucknell and then add that to the graph.

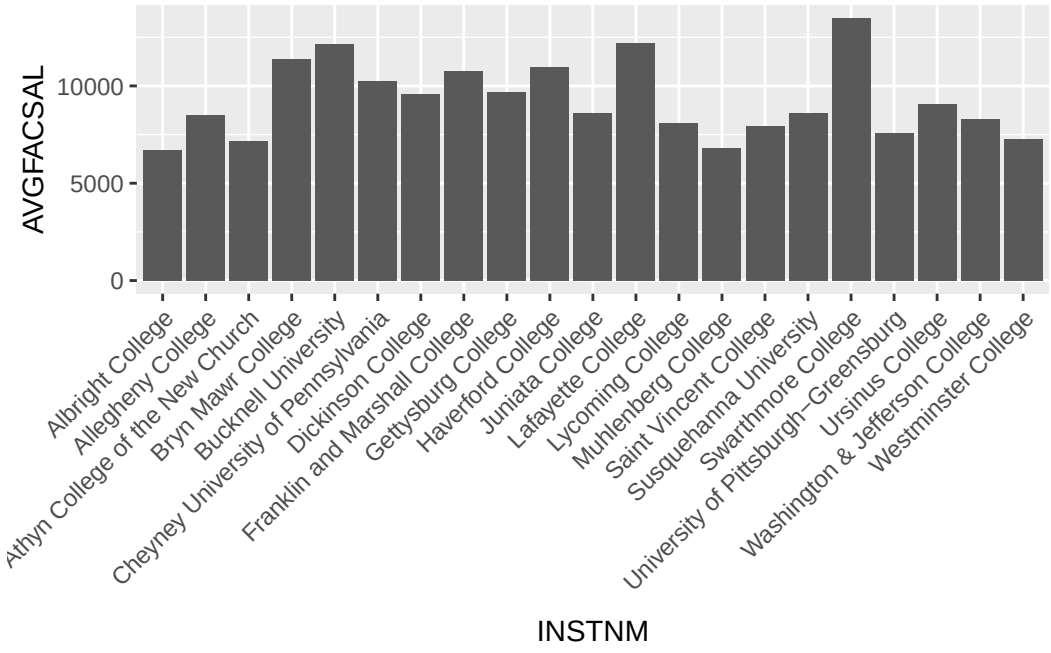
```
# Create a dataset that just has Bucknell's info

# Just label Bucknell
```

How do I reorder the bars of my bar graph?

```
# Let's focus on PA schools
pa_schools <- filter(colleges, STABBR == "PA")

# Look at average faculty salaries by school
ggplot(data = pa_schools, mapping = aes(x = INSTNM,
                                        y = AVGFACSAL)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



What is the current order for `INSTNM`? What is a better order for the bars?

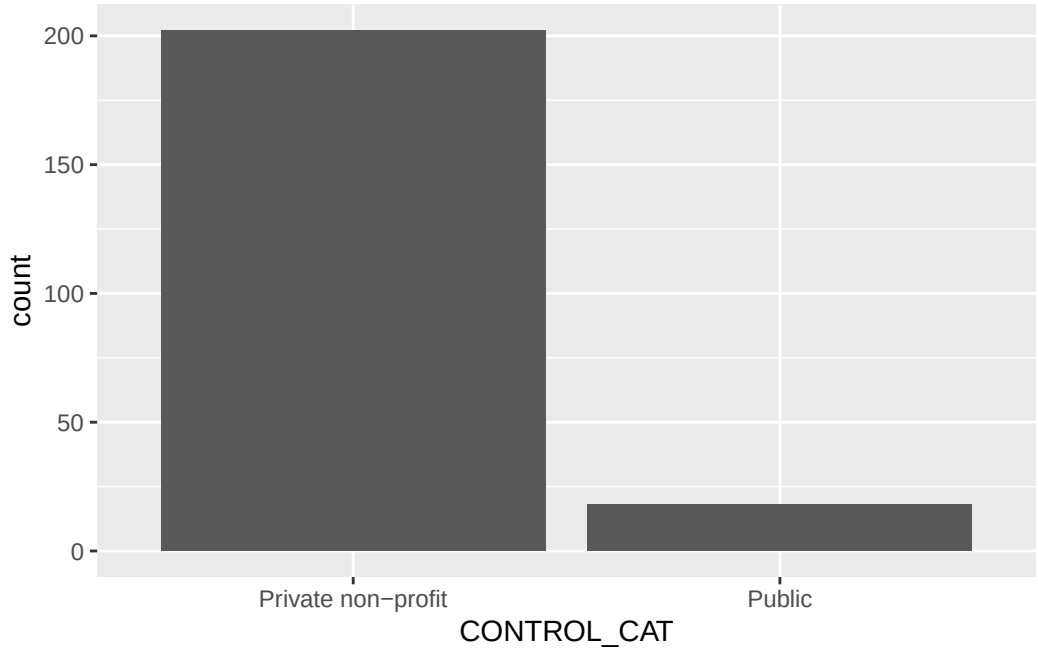
```
# Reorder institution name by the average faculty salaries


# Look at average faculty salaries by school
```

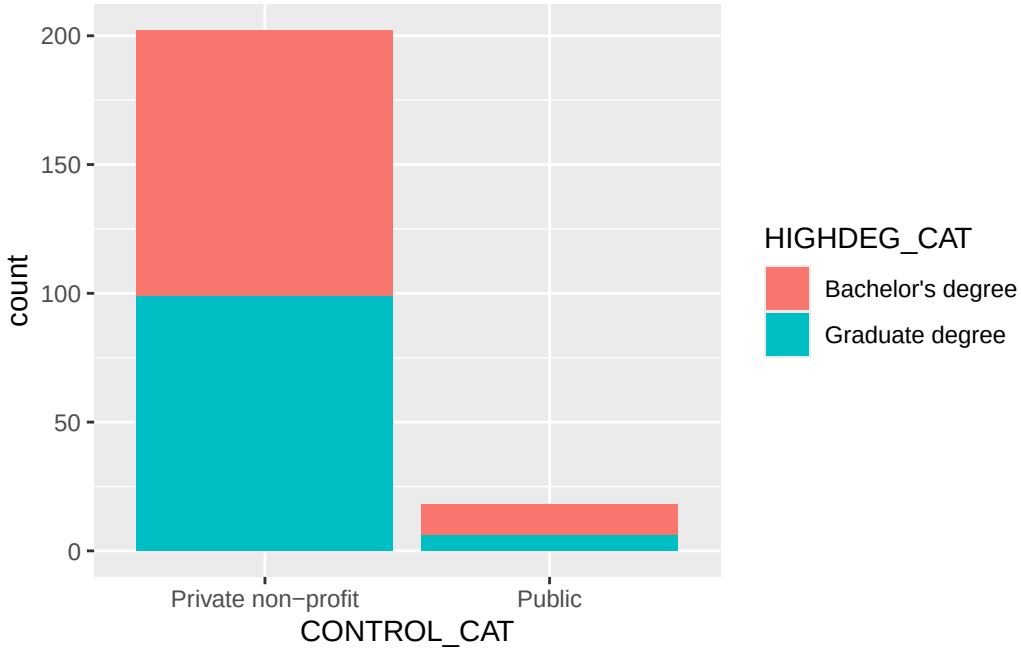
One more `ggplot` thing: `geom_bar()` versus `geom_col()`

Let's create bar graphs that compare the number of public and private schools by highest degree awarded.

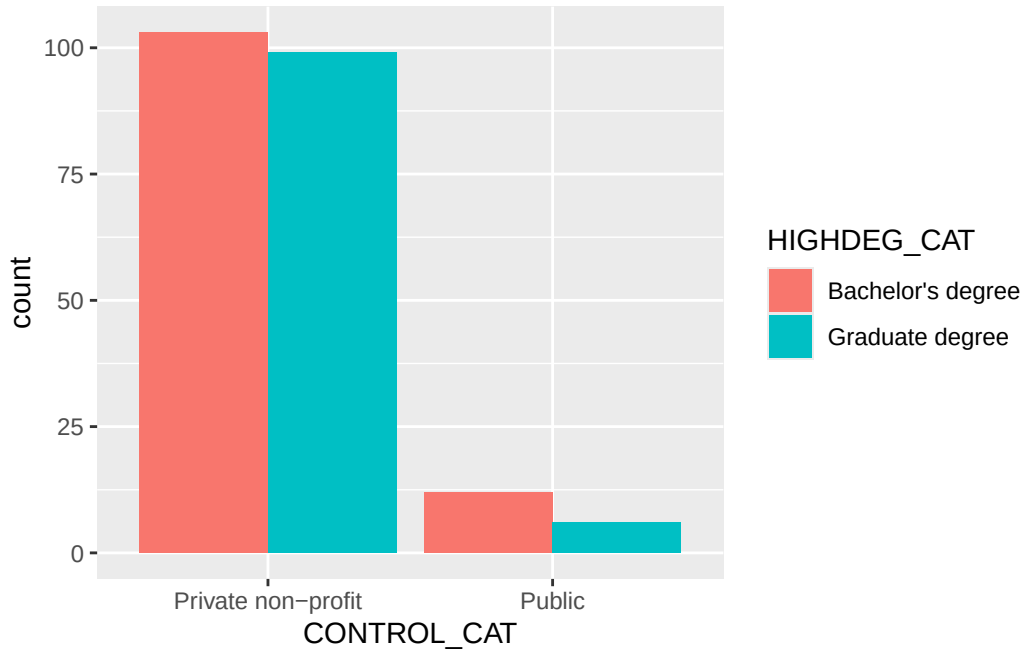
```
# Graph of counts (single variable)
ggplot(data = colleges, mapping = aes(x = CONTROL_CAT)) +
  geom_bar()
```



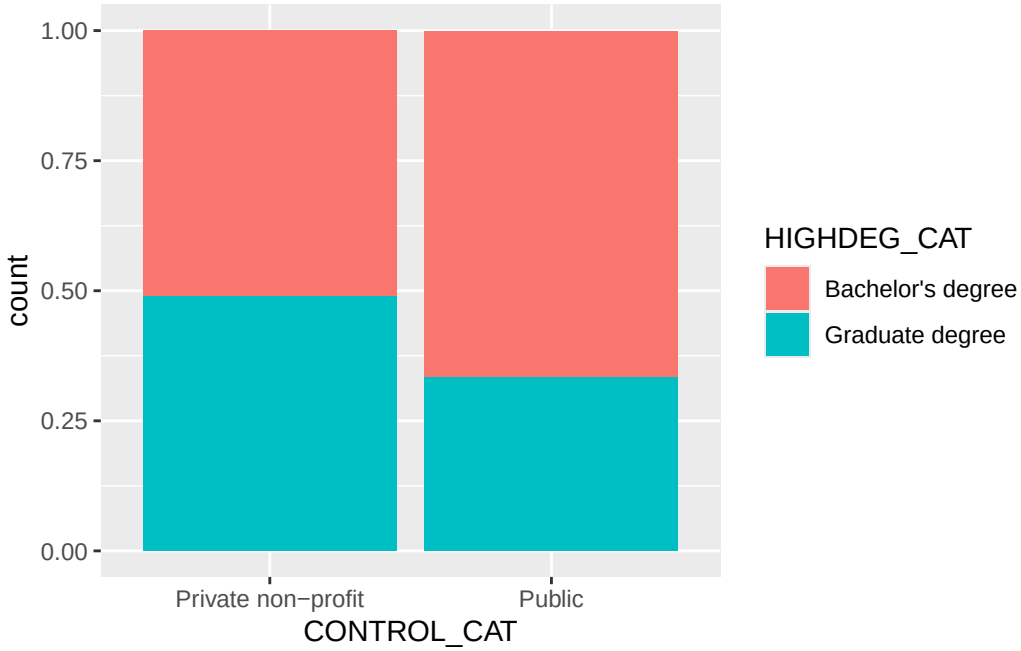

```
# Graph of counts (two variables)
ggplot(data = colleges, mapping = aes(x = CONTROL_CAT,
                                      fill = HIGHDEG_CAT)) +
  geom_bar()
```




```
# Graph of dodged counts  (two variables)
ggplot(data = colleges, mapping = aes(x = CONTROL_CAT,
                                      fill = HIGHDEG_CAT)) +
  geom_bar(position = "dodge")
```




```
# Graph of proportions  (two variables)
ggplot(data = colleges, mapping = aes(x = CONTROL_CAT,
                                      fill = HIGHDEG_CAT)) +
  geom_bar(position = "fill")
```



Main Data Wrangling Operations in `dplyr`

`summarize()`: Summarize variable(s)

What is the average admission rate? What is the lowest admission rate?

```
# Summarize


# Save summary to new dataset
```

`count()`: Add up number of rows for each category

How many historically black colleges and universities are in the dataset? Of those, how many award graduate degrees?

`mutate()`: Modify an existing variable or add new variables

Let's re-create the `Location` variable that indicates whether or not a college is in PA. What happened to the dimensions of `colleges` once we made this change?

```
# Check work with count()
```

Let's fix the class of `DEBT_MDN` and `HIGHDEG`. You can use `glimpse()` to see the classes of each variable. What happened to the dimensions of `colleges` once we made these changes?

```
# Check work with glimpse()
glimpse(colleges)
```

```
Rows: 220
Columns: 27
$ UNITID          <dbl> 100937, 101912, 106342, 107080, 107512, 112260, 115409~
$ INSTNM          <chr> "Birmingham-Southern College", "Oakwood University", "~
$ CITY            <chr> "Birmingham", "Huntsville", "Batesville", "Conway", "A~
$ STABBR          <chr> "AL", "AL", "AR", "AR", "AR", "CA", "CA", "CA", "CA", ~
$ HIGHDEG         <dbl> 3, 4, 3, 4, 4, 4, 3, 4, 3, 3, 3, 4, 4, 3, 3, 4, 4, 4, ~
$ PREDDEG         <dbl> 3, 3, 3, 3, 3, 3, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ~
$ CONTROL         <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, ~
$ HBCU            <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ TUITFTE         <dbl> 10340, 12279, 10188, 9685, 10749, 35957, 36122, 26331,~
$ AVGFACSAL       <dbl> 7029, 4842, 5817, 7889, 6735, 15333, 14478, 11309, 117~
$ ADM_RATE        <dbl> 0.5717, 0.6805, 0.5984, 0.6028, 0.7232, 0.1035, 0.1336~
$ SATVR75         <dbl> 670, NA, NA, 680, 610, 760, 770, NA, 750, NA, 770, 760~
$ SATMT75         <dbl> 610, NA, NA, 648, 590, 790, 790, NA, 760, NA, 790, 750~
$ ACTCM75         <dbl> 29, NA, NA, 30, 28, 35, 36, NA, 34, NA, 35, 34, NA, 32~
$ COSTT4_A        <dbl> 35495, 38377, 44749, 49928, 43878, 78723, 82236, NA, 7~
$ NPT4_PRIV       <dbl> 19723, 19686, 25183, 22780, 23086, 19489, 39671, NA, 3~
$ UGDS            <dbl> 968, 1378, 489, 1127, 1587, 1383, 906, 15, 1935, 1212,~
$ UG25ABV         <dbl> 0.0170, 0.1284, 0.0276, 0.0054, 0.0140, 0.0021, 0.0011~
$ PCTFLOAN_DCS    <dbl> 0.6452, 0.6477, 0.5934, 0.4483, 0.6109, 0.1627, 0.3646~
$ PCTPELL_DCS     <dbl> 0.2277, 0.4906, 0.3702, 0.2543, 0.2486, 0.2008, 0.1293~
$ DEBT_MDN        <chr> "16000", "21500", "10699", "19500", "15000", "11948", ~
$ C100_4          <dbl> 0.5854, 0.3351, 0.3085, 0.6743, 0.6174, 0.8318, 0.8826~
$ RET_FT4         <dbl> 0.7746, 0.7706, 0.5072, 0.7905, 0.7897, 0.9579, 0.9733~
$ MD_EARN_WNE_5YR <dbl> 56625, 51429, 45744, 49579, 48168, 108186, 154095, 418~
$ Location        <chr> "Not PA", "Not PA", "Not PA", "Not PA", "Not PA", "Not~
$ CONTROL_CAT     <chr> "Private non-profit", "Private non-profit", "Private n~
$ HIGHDEG_CAT     <chr> "Bachelor's degree", "Graduate degree", "Bachelor's de~
```

`select()`: Extract variables

Let's create a new dataset that only has the school name and location.

`filter()`: Extract cases

Let's filter down to schools that are:

- In the mid-atlantic: PA, NJ, VA, MD, DE, WV, DC
- Have undergraduate enrollments over 1000 students
- Don't have grad students

Let's filter down to just Bucknell.

```
bucknell <- filter(colleges, INSTNM == "Bucknell University")
```

`drop_na()`: Remove rows that have missing values for certain variables

Let's remove rows that are missing an admissions rate.

`arrange()`: Sort the cases

Let's sort rows by their admissions rate. Which schools has the lowest admissions rate? Which has the highest?

The pipe: %>% or |> for chaining together multiple wranglings

If you want to do multiple operations at once, you should use the pipe.

Suppose we want to look at `INSTNM`, `Location`, `ADM_RATE`, `UGDS`, `RET_FT4`, and `MD_EARN_WNE_5YR` for schools in PA that reported an admissions rate and we want to arrange the schools from largest undergraduate class to smallest undergraduate class.

`group_by()`: Perform actions by certain groups

For each of the Mid-Atlantic states, what is the average admission rate and how many schools are in each state?

Closing Thoughts on Wrangling

- Data are messy. Be prepared to wrangle.
- Before you start writing code ask yourself, what do I expect the wrangled data to look like? How many rows do I expect? How many columns?
- Don't try to wrangle all at once.
  - Write one line of code. Run it. And then keep going.
- Give the wrangled dataset a new name if you are removing rows or changing the structure drastically.

Your Optional Homework

If using your own data, do some wrangling that help answer questions of interest to you.

For the provided data, try to complete the following tasks.

a. How many schools are in each of the categories of `PREDDEG`?

b. Create a dataset that only contains schools that are predominantly bachelor's degree granting. Use this dataset for the following questions.

c. Compute the minimum, maximum, and median values of the median earnings of graduates working and not enrolled 5 years after completing. Useful `R` functions here are: `min()`, `max()`, `median()`.

d. Repeat part (c) but this time compute the summary statistics for both HBCUs and non-HBCUs.

e. Ask some of your own questions of the data and then wrangle the data in order to answer them.

Resources for Learning More about Data Wrangling with `dplyr`

- Modern Dive's chapter on Data Wrangling
- R for Data Science's chapter on Data Transformation
- `dplyr` cheatsheet: https://raw.githubusercontent.com/rstudio/cheatsheets/main/data-transformation.pdf