Wrangling Data

Dominguez Center for Data Science Workshop

2025-03-21

Questions Round-Up

How can I provide comments with my code?

- Outside of a code chunk, # = Create Section header, ## = Create sub-section header, …
- Instead of a code chunk, # = Don't run whatever is after this pound symbol

```
# Load the needed packages
library(tidyverse)

# 2 + 2
```

Any other questions?

Plan for today

- Recap data wrangling functions we covered last time.
- Learn about chaining data wrangling operations together with the pipe (`%>%` or `|>`).
- Go through the "more_wrangling_key.qmd" Quarto document.
- Remember that you can either fill in the "more_wrangling.qmd" file or follow along with the "more_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`.

```r
library(tidyverse)
```

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 the data.

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

Recap: Data wrangling from last session

Data wrangling = any transformations done on the data.

Some Thoughts on Wrangling

- Data are messy. Be prepared to wrangle.

  "Tidy datasets are all alike, but every messy dataset is messy in its own way." – Hadley Wickham

- 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.

Main Data Wrangling Operations in dplyr

summarize(): Summarize variable(s)

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

```
colleges_summary <- summarize(colleges,
                              mean_admit = mean(ADM_RATE, na.rm = TRUE),
                              lowest_admit = min(ADM_RATE, na.rm = TRUE) )
colleges_summary
```

```
# A tibble: 1 x 2
  mean_admit lowest_admit
       <dbl>        <dbl>
1      0.601       0.0693
```

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?

```
count(colleges, HBCU)
```

```
# A tibble: 2 x 2
   HBCU     n
  <dbl> <int>
1     0   203
2     1    17
```

```
count(colleges, HBCU, HIGHDEG)
```

```
# A tibble: 4 x 3
   HBCU HIGHDEG     n
  <dbl>   <dbl> <int>
1     0       3   105
```

```
2      0        4    98
3      1        3    10
4      1        4     7
```

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

Three examples below:

- Adding a new variable called `Location`: indicates if a school is in PA or not.
- Creating `HIGHDEG_CAT`: Which takes the numerical varaible `HIGHDEG` and creates a categorical version.
- Fixing `DEBT_MDN` so that `R` stores it as a numerical variable, not a categorical variable (which `R` calls a character vector).

```r
colleges <- mutate(colleges,
                   Location = if_else(STABBR == "PA", "PA", "NOT PA"),
                   HIGHDEG_CAT = case_match(HIGHDEG,
                                            3 ~ "Bachelor's degree",
                                            4 ~ "Graduate degree"),
                   DEBT_MDN = as.numeric(DEBT_MDN))
#Check work
glimpse(colleges)
```

```
Rows: 220
Columns: 26
$ 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       <dbl> 16000, 21500, 10699, 19500, 15000, 11948, 19500, 18667~
$ 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~
$ HIGHDEG_CAT    <chr> "Bachelor's degree", "Graduate degree", "Bachelor's de~
```

```
count(colleges, Location)
```

```
# A tibble: 2 x 2
  Location     n
  <chr>    <int>
1 NOT PA     199
2 PA          21
```

```
count(colleges, HIGHDEG_CAT)
```

```
# A tibble: 2 x 2
  HIGHDEG_CAT           n
  <chr>             <int>
1 Bachelor's degree   115
2 Graduate degree     105
```

`select()`: Extract variables

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

```
colleges2 <- select(colleges, INSTNM, 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

```
colleges3 <- filter(colleges,
                    STABBR %in% c("PA", "NJ", "VA", "MD", "DE", "WV", "DC"),
       UGDS > 1000,
       HIGHDEG != 4)
```

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.

```
colleges_adm_rate_complete <- drop_na(colleges, ADM_RATE)
```

More wrangling functions

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?

How can I combine two datasets?

- Often the data is stored across several datasets and you want to combine them into one in a principled way.
- Need a key that links the two datasets.

```
# Load data from the Opportunity Insights lab
opportunity_insights <- read_csv("data/opportunity_insights.csv")
```

Suppose we want to add the upward mobility information from the Opportunity Insights dataset to our colleges dataset. Opportunity Insight is a research initiative based at Harvard University and led by Raj Chetty, John Friedman, and Nathaniel Hendren, with the goal of improving upward mobility in the United States by studying barriers to economic opportunity and translating findings into policy change. They defined a college's mobility rate as the percentage of students with parents in the bottom income quintile who ended up in the top x% (in their mid-30s). The variables `mr_kq5_pq1` and `mr_ktop1_pq1` and refer to the percentage of students with parents in the bottom income quintile who ended up in the top 20% and top 1%, respectively.

Let's first look at smaller datasets so we can explore the different types of data joins. What are the key variables?

```
# Create smaller versions
colleges_nyc <- colleges %>%
  select(INSTNM, CITY, STABBR, ADM_RATE) %>%
  filter(CITY == "New York")
colleges_nyc
```

```
# A tibble: 3 x 4
  INSTNM                      CITY     STABBR ADM_RATE
  <chr>                       <chr>    <chr>     <dbl>
1 Barnard College             New York NY       0.0879
2 Marymount Manhattan College New York NY       0.721
3 The King's College          New York NY       0.453
```

```
opp_ny <- opportunity_insights %>%
  filter(state == "NY", tier_name == "Selective private")
opp_ny
```

```
# A tibble: 48 x 5
   name                                  state tier_name mr_kq5_pq1 mr_ktop1_pq1
   <chr>                                 <chr> <chr>         <dbl>        <dbl>
 1 Adelphi University                    NY    Selectiv~     0.0326   0.00261
 2 Alfred University                     NY    Selectiv~     0.0148   0.0000507
 3 Boricua College                       NY    Selectiv~     0.0364   0.000132
 4 Canisius College                      NY    Selectiv~     0.0236   0.00205
 5 Cazenovia College                     NY    Selectiv~     0.0126   0.000142
 6 Clarkson University                   NY    Selectiv~     0.0297   0.000624
 7 College Of Mount Saint Vincent And M~ NY    Selectiv~     0.0578   0.00173
 8 College Of New Rochelle               NY    Selectiv~     0.0287   0.00000964
 9 College Of Saint Rose                 NY    Selectiv~     0.0173   0.000686
10 D'Youville College                    NY    Selectiv~     0.0397   0.000104
# i 38 more rows
```

Three common types of joins:

```
# The inner join


# The full join


# The left join
```

Which join should we use if we want to add the upward mobility information to our colleges dataset?

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 and d.

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. Create a dataset of just the HBCUs and add the Opportunity Insights variables to that dataset. How many of the HBCUs in our dataset are in the Opportunity Insights dataset?

f. 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