**Data Wrangling II Session Notes**

* Hadley Wickham
  + Statistician
  + Chief scientist at Posit, which is an open-source data science company that I think owns the R software, you may have downloaded R and RStudio from Posit
  + Known for his work on R open-source software
* He came up with this concept of “Tidy data”
  + The basic idea of tidy data is essentially there’s one variable per column, one observation per row
  + Something like 80% of data science and the type of quantitative work that we do is cleaning, tidying, wrangling data
  + Wickham did research on how to make data wrangling and cleaning as easy and effective as possible
* His whole point is that once data is in a tidy format, it’s much easier to manipulate, model and visualize
  + He basically wanted to come up with a standardized data format so that tools and functions could be developed that could be applied to any “tidy” dataset
* So, he developed tidyverse, which is a package that includes libraries for data wrangling and manipulation

**Other random notes**

* Once I learned Tidyverse, I never looked back to base R
* It allows for really logical and organized code

**Pipes and filters**

* Tidyverse makes breaking your data wrangling and manipulation into different, distinct steps really easy
* Of course you can do this in base R, but you can create different versions of your dataframe and this way, it’s easy to go back and change things or access earlier steps of
* Meghan mentioned she uses the head function a lot
* I work with large medical claims datasets, so one function that I use a ton when I’m tidying, manipulation, cleaning my data is dim, just to check the number of rows and columns that I have in my dataset as I’m filtering and subsetting
* So if I filter something out and I’m only expecting like 100 rows to be filtered out, but I use dim and it turns out 10,000 rows have been filtered out, something went wrong and I can go back and check my filtering

**Generate plots (long format)**

* Tidyverse is a really easy, quick way to manipulate or create a new dataframe that is in the format necessary to plot
* It’s always a good idea to plot your data too, as you’re doing exploratory data analysis and data manipulation, it can help make sure that everything looks correct and is working as expected.
* We didn’t manipulate the diastolic and systolic blood pressure columns, but if we had, we could use this scatter plot to ask ourselves if the values look correct. So a normal blood pressure is less than 120 for systolic and less than 80 for diastolic, so these data look correct.

**Summary table**

* This can be helpful to evaluate variables if you want to stratify by a variable in an analysis
* And just to reiterate what Meghan mentioned this morning, it’s really easy to kind of mess something up in the data when you’re cleaning and manipulating it, so it’s good to get in the habit of checking your data frame dimensions and what it looks like, and also get in the habit of constantly summarizing variables and making sure that everything makes sense. Once you’re in a lab, you’ll probably be working with one type of data, so for example, I work with claims data, and as you get more familiar with it, you’ll start to just know what’s “normal” in the data. And the dim(), head() and summary() functions can help you quickly identify if something looks off. Or, even if we as coders didn’t mistype something, these things can help us quickly identify abnormalities in the data or outliers, and things like that. Catching these things early can just makes downstream analysis much easier.