NAWQA GW Training Session
Starting a project and using the tidyverse
5/30/2018

Note that some of this is review and covered in our intro class, but we know that some people haven't taken that.

Tidyverse: Collection of packages that make working with data easier.

- **dplyr**::mutate, group_by, filter, select, summarize, case_when, recode
- **tidyr**::gather, spread
- piping

[11-11:15 am]
Start with RStudio & general coding practices

- Quick description of various panes
- Start a project (from GitHub)
- General organization of projects that use code
    - raw_data
    - Cache_data - intermediate forms of the data that you don't need, but are nice to save
    - Release_data - publication/release-ready data
    - R - scripts
    - Figures - plots/maps/etc that are created from the code
- General coding things
    - Talk about file paths in projects - relative paths
    - Don't use setwd or install.packages in scripts
    - Always include library() at top of scripts
- When to and when not to save workspaces
    - Change global option so it doesn't ask every time (Tools → Global → General)

[11:15-11:30 am]
Reading in data as correct class

- data.table package for big(ger) data
- fread to read in data
- No need to do stringsAsFactors = FALSE
- Sites should be character, dates should be dates
- Read data back in with colClasses defined for sites (date classes can't be set in fread)

[11:30-11:45 am]

Fixing dates & introducing mutate/pipes

- R format - YYYY-mm-dd (vs others)
- Just use as.Date for already formatted
- Explain mutate + %>%
- Show what happens when you open in Excel
- Walk through handling Excel dates

[11:45 am - 12:30 pm]

Cleaning up data (adding/combining columns, inserting missing values)

- Add column for aquifer zone
  - Use **filter**() when figuring out cutoff value for shallow vs deep wells
  - Use **mutate**() + **case_when**() to add new column based on values of another column
- Aggregate landuse columns into categories of interest (developed, semi-developed, low-usage, and agriculture) by year
  - Reshape data using **gather**(), then separate to parse out year from category
  - Reshape again (**spread**) with years still as a column, and categories back as columns
  - Add new columns (**mutate**) that are combos of previous columns, then remove old columns (**select**)
- Insert missing landuse values when sample year is past the landuse year
  - Reshape to make the category and value columns using **gather**()
  - For each observation, insert NA if the sample year is past the landuse year (**mutate** + ifelse)
  - Combine the year and the landuse category (**unite**), then reshape (**spread**) to get back to how original data looks
- Show all in one big chain, then save intermediate

[12:30 - 12:45 pm]

Summarize data

- group_by and summarize
- recode

[12:45 - 1:00 pm]

Questions/discussion