Common Manipulations Tutorial

Election Data Science

Peter Licari

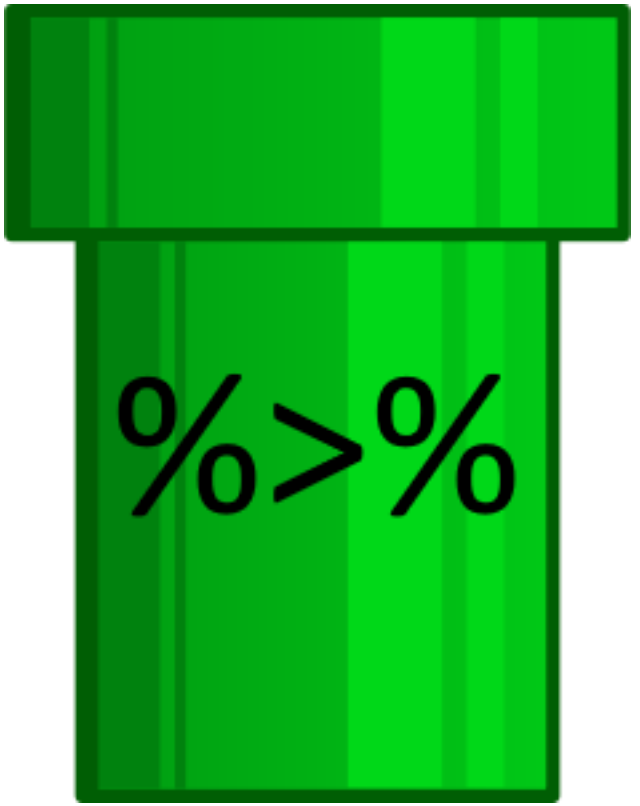
2020-09-15

It's good to learn about a bunch of wrangling techniques and practices---and you will, given time---but I wanted to walk you through the the most common ones (at least in my experience).

Most common `tidyverse` (`dplyr`) wrangling functions

- `filter`
- `mutate` & `case_when`
- `group_by` & `summarise`
- `select`
- `pivot`
- `across`
- `join`

But first, let's talk about pipes.



Let's-a go!

When you write out code in your R script (or R markdown chunk) you'll often do a number of steps:

```
data
data1 <- select(data, var1, var2, var3)
data1 <- filter(data1, var1 > 7)
data1 <- mutate(data1, var4 = var2 / var3)
```

There are redundancies in each step that makes errors more likely.

Use %>% instead!

```
data1 <- data %>%
  select(var1, var2, var3) %>%
  filter(data1, var1 > 7) %>%
  mutate(data1, var4 = var2 / var3)
```

When reading through the code, replace `%>%` with the phrase "*and then, with the same dataframe.*"

Bonus! `%>%` doesn't just work with `tidyverse` functions. As long as your custom-functions include an input for dataframes/matrices/tibbles, you can chain them in as well!

If you want to follow along:

```
#Set your working directory

setwd("your/directory/here")

# Voter registration records Tyrell County NC
url1 <- "https://s3.amazonaws.com/dl.ncsbe.gov/data/ncvoter89.zip"
download.file(url1, destfile = paste0(getwd(),"/voterfile.zip"))
unzip("voterfile.zip")
file.rename(from = "ncvoter89.txt", to = "voterregistration.txt")

# Voter history files Tyrrell County NC
url2 <- "https://s3.amazonaws.com/dl.ncsbe.gov/data/ncvhis89.zip"
download.file(url2, destfile = paste0(getwd(),"/voterhistory.zip"))
unzip("voterhistory.zip")
file.rename(from = "ncvhis89.txt", to = "voterhistory.txt")

# Load in data
TY_Records <- readr::read_delim("voterregistration.txt", delim = "\t")

TY_History <- readr::read_delim("voterhistory.txt", delim = "\t")
```

Filter

The Code | How it Works | Output

```
TY_Records_Filt <- TY_Records %>%
  filter(status_cd != "A", race_code == "B", birth_age < 40, !is.na(full
```

Mutate

The Code | How it Works | Output

```
TY_Records_Mut <- TY_Records %>%
  mutate(registr_dt = as.Date(registr_dt, format = "%m/%d/%Y"),
         registr_dur = Sys.Date() - registr_dt,
         yrs_since_el = birth_age - 18,
         hello = "world")
```

Mutate & Case_When

The Code | How it Works | Output

```
TY_Records_MutCase <- TY_Records %>%
   mutate(generation = case_when(
   birth_age <= 24 ~ "Gen Z",
   birth_age > 24 & birth_age <= 40 ~ "Millennial",
   birth_age > 40 & birth_age <= 55 ~ "Gen X",
   birth_age > 55 & birth_age <= 74 ~ "Boomer",
   birth_age > 74 & birth_age <= 92 ~ "Silent",
   TRUE ~ "Greatest"
  ))
```

group_by & summarise

The Code | How it Works | Output

```
TY_Records_Group <- TY_Records_MutCase %>%
 group_by(generation, party_cd) %>%
   summarise(Number = n())
```

Select

The Code | How it Works | Output

```
TY_Records_Select <- TY_Records%>%
 select(voter_reg_num, party_cd, gender_code, race_code)
```

Pivot

The Code | How it Works | Output

```
TY_History_Pivot <- TY_History %>%
 mutate(n = 1) %>%
 pivot_wider(names_from = election_desc, id_cols = voter_reg_num, value
             values_fill = 0)
```

join

The Code | How it Works | Output

```
TY_Reg_Hist <-
    TY_Records_Select %>% full_join(TY_History_Pivot, by = "voter_reg_nu
```