SPEC Lab R Resources: Data Management I - Homework

Ben Graham, Alix Ziff, and Jasmine Chu

Summer 2022

Transforming Data using the Tidyverse

Complete the following assignment. Save your R script and the mini dataset from step 2 into your personal subfolder in the Homework Submission google drive folder. The R script should be titled HW_DM1_[YOUR INTIALS]. Please also make sure you save your R script to your own computer for future reference.

Don't forget to annotate your script thoroughly! And save a copy to your personal R script library. These scripts are a resource for future you!

Exercise 1:

A Load in the .rds file of the IDC 2021 powersharing data: "Training Data Summer 2022/IDC_training_2021.rds"

```
rm(list=ls()) #clear everything
library(dplyr) #load the correct library

dt <- readRDS("IDC_training_2021.rds")
```

B Complete step B in a single command, piped together. Create a subset of the data that includes only the following countries, years, and variables:

1. Countries: U.S., China, Russia, France
2. Variables: country, gwno, year, subed_IDC, subtax_IDC, subpolice_IDC, auton_IDC, stconst_IDC
3. Years: 2015-2018

```
# Don't forget that the code provided here is only ONE way to do this.
# There are multiple correct ways to do things, so it is totally OK
# if your code doesn't match the answer key.

dt <-dt %>%
  filter(country %in% c("United States of America", "China", "France",
                        "Russia (Soviet Union)"))%>%
  select(country, gwno, year, subed_IDC, subtax_IDC, subpolice_IDC,
         auton_IDC, stconst_IDC)%>%
  filter(year %in% (2015:2018))
```

Exercise 2: Save this smaller dataset as "Minipowersharing_YOURNAME.rds". Ideally, specify a filepath in your save command so that it saves straight to your homework submission folder in google drive.

```
saveRDS(dt, file = "Minipowersharing_JASMINECHU.rds") #save as rds file
```

Exercise 3: Using the full dataset again, create a new variable, "subpower_additive" that is the sum of subed_IDC, subtax_IDC, and subpolice_IDC. This index should take a value of N/A if any of the three component indicators is missing.

```
dt <- readRDS("IDC_training_2021.rds")%>%
  mutate(subpower_additive=subed_IDC+subtax_IDC+subpolice_IDC)
```

Bonus: Create a new version of the index, "subpower_additive_nm", that assumes subed_IDC, subtax_IDC, and subpolice_IDC take a value of 0 if they are missing. This version of the index should have no missing values.

```
dt_bonus <- readRDS("IDC_training_2021.rds")%>%
  mutate(subpower_additive_nm=subed_IDC+subtax_IDC+subpolice_IDC)
#replace NA values with 0
dt_bonus$subpower_additive_nm[is.na(dt_bonus$subpower_additive_nm)] <- 0


# OR
dt_bonus <- readRDS("IDC_training_2021.rds") %>%
  mutate(subpower_additive_nm = subed_IDC+subtax_IDC+subpolice_IDC) %>%
  mutate(subpower_additive_nm = replace_na(subpower_additive_nm, 0))
```

Exercise 4: Use summarise() or summarise_at() to answer the following:

1.What is the mean value of your first subpower index (`subpower_additive`) in the entire sample of countries, across the years 2010 through 2019? *Hint:* This should be one value.

```
dt_1 <- dt %>%
  summarise(AverageSubpower = mean(subpower_additive, na.rm=T))
# For the entire sample, the mean value is 1.410197
```

2.What about in the year 2019 only?

```
dt_2 <- dt %>%
  filter(year==2019)%>%
  summarise(AverageSubpower=mean(subpower_additive, na.rm=T))

#In the year 2019, the mean value of the first subpower index in the year 2019
# is 1.44186
```

3. DOUBLE BONUS: How about the mean value of your second version of the index (`subpower_additive_nm`) for the entire sample?

```
dt_bonus1 <- dt_bonus %>%
  summarise(AverageSubpower=mean(subpower_additive_nm, na.rm=T))
# For the entire sample, the mean value of the _nm version is 1.391652
```

4. TRIPLE BONUS: How many countries in 2019 have a value for the _nm version but not for the original version?

```r
dt<-dt%>%
  filter(year==2019) #filter for when the year is 2019 for original version
sum(is.na(dt$subpower_additive)) #there are 3 NA values for original version
```

```
## [1] 2
```

```r
dt_bonus<-dt_bonus%>%
  filter(year==2019) #filter for _nm version
ncol(dt_bonus) #49 values
```

```
## [1] 1023
```

```r
#49 - 3 = 46
# There are 46 countries in 2019 that have a value for the _nm version
# but not for the original version
```