# SPEC REU R Resources: Data Management I with tidyverse – Homework

Ben Graham, Alix Ziff, Jasmine Chu, and Claudia Salas Gimenez

Summer 2024

Welcome to the homework assignment for the Data Management I module. In this assignment, you'll put into practice the concepts we've covered so far, including data manipulation, subsetting, creating new variables, and summarizing data using R's tidyverse and dplyr packages.

As we continue to work with the `IDC_training_2021.rds` dataset, the tasks are designed to give you hands-on experience in data management. By the end of the assignment, you'll have a stronger grasp of how to load, manipulate, and save datasets in R.

Save your responses in your personal subfolder in the `412_413 shared AY24-25` Google Drive folder. The R script should be titled `HW_DM1_[YOUR INITIALS]`. You can also save a copy of your R script to your own computer for future reference.

And remember: annotating your script is key! Clear comments will help you understand what your code does when you revisit it in the future and are especially helpful in collaborative settings.

## Initial Setup

Start by setting up your R workspace. Create a header for your R script, save it in your personal folder, set your working directory to the folder containing the training data, and load the required libraries (`tidyverse` and `readr`) and the data.

For this assignment, you will continue working with the `IDC_training_2021.rds` dataset, a country-year dataset collected by the SPEC Lab. Included for your reference are the "Safeguarding Democracy: Power-sharing and Democratic Survival" research paper, along with the codebook. These documents offer detailed insights into the dataset and will help you complete the following exercises.

```r
# Set working directory
#setwd("YourFolderPath")

# Load required libraries
library(tidyverse)
library(readr)

# Load the data
dt <- readRDS("IDC_training_2021.rds")
```

Keep in mind that the code in the answer key is just one possible solution. There are many correct ways to approach the task, so it's perfectly fine if your code differs from the answer key.

## Exercise 1: Data Subsetting

Create a subset of the data, focusing on the U.S., China, Russia, and France from 2015 to 2018, and including variables that refer to country names, country codes, years, and those related to subnational policy authorities. Use piping to accomplish this task.

```
# Subset the data to include only specific countries, variables, and years
dt_subset <-dt %>%
  filter(country %in% c("United States of America", "China", "France",
                        "Russia (Soviet Union)"))%>%
  select(country, gwno, year, subed_IDC, subtax_IDC, subpolice_IDC)%>%
  filter(year %in% (2015:2018))
```

### Bonus Exercise: Calculating Averages

Create a new subset that includes the same countries, years, and variables as in Exercise 1. Additionally, to take it a step forward, incorporate the global averages for subnational policy authorities for each country from 2015 to 2018.

```
# Subset the data and calculate averages
dt_bonus_1 <- dt %>%
  filter(country %in% c("United States of America", "China", "France",
                        "Russia (Soviet Union)"),
         year %in% (2015:2018)) %>%
  select(country, gwno, year, subed_IDC, subtax_IDC, subpolice_IDC, auton_IDC,
         stconst_IDC) %>%
  group_by(country) %>%
  summarise(Avg_subed_IDC = mean(subed_IDC, na.rm = TRUE),
            Avg_subtax_IDC = mean(subtax_IDC, na.rm = TRUE),
            Avg_subpolice_IDC = mean(subpolice_IDC, na.rm = TRUE))
```

## Exercise 2: Save the Subset

Save the dataset you created in Exercise 1 as `Minipowersharing_[YOUR NAME].rds` to your personal subfolder in the `412_413 shared AY24-25` Google Drive folder.

```
# Save dataset as .rds file
saveRDS(dt_subset, file = "Minipowersharing_JASMINECHU.rds")
```

## Exercise 3: Create New Variables

Using the complete dataset, create a new variable named `subpower_additive` that represents the sum of the subnational policy authorities. This index should be assigned a value of `NA` if any of the three components are missing.

```
# Create new variables
dt <- dt %>%
  mutate(subpower_additive = subed_IDC + subtax_IDC + subpolice_IDC)
```

### Bonus Exercise: Handle Missing Values

Create a new version of the `subpower_additive` variable from Exercise 3, treating missing values as 0 to ensure there are no `NA` values. Name this new variable `subpower_additive_nm`.

```r
# Create new variables replacing NA values with 0
## Approach 1: Create a new variable 'subpower_additive_nm' and then replace NA
## values in 'subpower_additive_nm' with 0
dt_bonus_2 <- dt %>%
  mutate(subpower_additive_nm = subed_IDC + subtax_IDC + subpolice_IDC)

dt_bonus_2$subpower_additive_nm[is.na(dt_bonus_2$subpower_additive_nm)] <- 0

## Approach 2: Create a new variable 'subpower_additive_nm' and automatically
## replace NA values with 0 using 'replace_na()'
dt_bonus_2 <-  dt %>%
  mutate(subpower_additive_nm = subed_IDC + subtax_IDC + subpolice_IDC) %>%
  mutate(subpower_additive_nm = replace_na(subpower_additive_nm, 0))
```

# Exercise 4: Summarizing Data

## Exercise 4.1: Mean of `subpower_additive` (2010-2019)

Can you calculate the average value of `subpower_additive` for all countries over the years 2010 to 2019?

```r
# Calculate the mean of 'subpower_additive' for all countries between 2010 and 2019
dt_1 <- dt %>%
  summarise(AverageSubpower = mean(subpower_additive, na.rm = T))

# Display the result
dt_1
```

```
##   AverageSubpower
## 1        1.154657
```

```r
## For the entire sample, the mean value is 1.154657
```

## Exercise 4.2: Mean in 2019

Find the mean of `subpower_additive` for the year 2019.

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

# Display the result
dt_2
```

```
##   AverageSubpower
## 1        1.416185
```

```r
## The mean value of the first subpower index in the year 2019 is 1.416185
```

## Exercise 4.3: Mean of `subpower_additive_nm`

Calculate the mean of the modified `subpower_additive_nm` index across the entire sample.

```r
# Summarize the dataset
dt_3 <- dt_bonus_2 %>%
  summarise(AverageSubpower = mean(subpower_additive_nm, na.rm = T))
```

```r
# Display the result
dt_3
```

```
##   AverageSubpower
## 1        1.048962
```

```r
## For the entire sample, the mean value of the _nm version is 1.048962
```

### Bonus Exercise: Compare Versions

Determine the number of countries in 2019 that have a value for `subpower_additive_nm` but lack one for the original `subpower_additive`.

```r
# Filter the dataset for the year 2019
dt_2019 <- dt %>%
  filter(year == 2019)

# Count how many NA values there are in the original subpower_additive for 2019
na_count_original <- sum(is.na(dt_2019$subpower_additive))

## There are 2 NA values for original version

# Now filter the same year (2019) for the dataset with subpower_additive_nm
dt_bonus_2019 <- dt_bonus_2 %>%
  filter(year == 2019)

# Get the total number of rows (countries) in the _nm version for 2019
total_count_nm <- nrow(dt_bonus_2019)

# Calculate how many countries have a value in subpower_additive_nm but not in the
# original subpower_additive
countries_with_nm_not_original <- total_count_nm - na_count_original

## There are 1023 values in the _nm version

## 1023 - 2 = 1021
## There are 1021 countries in 2019 that have a value for the _nm version but not
## for the original version
```

## Conclusion

This final assignment in Module 2: Data Management I solidifies your understanding of key data management techniques, including subsetting data, creating new variables, and summarizing datasets. Completing these exercises has equipped you with the foundational skills needed to efficiently manage and analyze real-world data, preparing you for more advanced tasks in the next module.