Assignment 4: Data Wrangling (Spring 2026)

HE GAO

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Wrangling.
Do not use any AI tools in completing this assignment.

Directions

1. Rename this file `<FirstLast>_A04_DataWrangling.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. **Ensure that code in code chunks is tidy and does not extend off the page in the PDF**.
7. **Push your completed RMD to your GitHub account**

Set up your session

1a. Load the `tidyverse` and `here` packages into your session.

1b. Check your working directory.

1c. Read in all four raw data files associated with the EPA Air dataset, being sure to set string columns to be read in a factors. See the README file for the EPA air datasets for more information (especially if you have not worked with air quality data previously).

2. Add the appropriate code to reveal the dimensions of the four datasets.

```
# 1a
library(tidyverse)
library(here)

# 1b
getwd()
```

```
## [1] "C:/Users/ /EDE_Spring2026"
```

```
# 1c: read the FOUR EPA Air raw files from Data/Raw
raw_dir <- here("Data", "Raw")

raw_files <- file.path(
```

```
  raw_dir,
  c(
    "EPAair_O3_NC2018_raw.csv",
    "EPAair_O3_NC2019_raw.csv",
    "EPAair_PM25_NC2018_raw.csv",
    "EPAair_PM25_NC2019_raw.csv"
  )
)

stopifnot(all(file.exists(raw_files)))
raw_files
```

```
## [1] "C:/Users/ /EDE_Spring2026/Data/Raw/EPAair_O3_NC2018_raw.csv"
## [2] "C:/Users/ /EDE_Spring2026/Data/Raw/EPAair_O3_NC2019_raw.csv"
## [3] "C:/Users/ /EDE_Spring2026/Data/Raw/EPAair_PM25_NC2018_raw.csv"
## [4] "C:/Users/ /EDE_Spring2026/Data/Raw/EPAair_PM25_NC2019_raw.csv"
```

```
EPA_raw_list <- map(raw_files, ~ read.csv(.x, stringsAsFactors = TRUE))
names(EPA_raw_list) <- basename(raw_files)

# 2
map(EPA_raw_list, dim)
```

```
## $EPAair_O3_NC2018_raw.csv
## [1] 9737   20
##
## $EPAair_O3_NC2019_raw.csv
## [1] 10592    20
##
## $EPAair_PM25_NC2018_raw.csv
## [1] 8983   20
##
## $EPAair_PM25_NC2019_raw.csv
## [1] 8581   20
```

TIP: All four datasets should have the same number of columns but unique record counts (rows). Do your datasets follow this pattern?

Wrangle individual datasets to create processed files.

3. Change any date columns to be date objects.
4. Create new dataframes with just the following columns:
   'Date', 'DAILY_AQI_VALUE', 'Site.Name', 'AQS_PARAMETER_DESC', 'COUNTY', 'SITE_LATITUDE', 'SITE_LONGITUDE'
5. For the PM2.5 datasets, fill all cells in AQS_PARAMETER_DESC with "PM2.5" (all cell values in this column should be identical).
6. Save all four processed datasets in the Processed folder. Use the same file names as the raw files but replace "raw" with "processed".

```r
library(lubridate)

parse_epa_date <- function(x) {
  as.Date(parse_date_time(x, orders = c("ymd", "mdy", "dmy")))
}

# 3-5: wrangle each dataset
EPA_processed_list <- map2(
  EPA_raw_list,
  names(EPA_raw_list),
  function(df, nm) {
#3
    df <- df %>%
      mutate(Date = parse_epa_date(Date))

#4
    df <- df %>%
      select(
        Date,
        DAILY_AQI_VALUE,
        Site.Name,
        AQS_PARAMETER_DESC,
        COUNTY,
        SITE_LATITUDE,
        SITE_LONGITUDE
      )
#5
    if (str_detect(tolower(nm), "pm25")) {
      df <- df %>% mutate(AQS_PARAMETER_DESC = "PM2.5")
    }

    df
  }
)

#6
processed_dir <- here("Data", "Processed")
dir.create(processed_dir, recursive = TRUE, showWarnings = FALSE)

processed_files <- names(EPA_raw_list) %>%
  str_replace("_raw\\.csv$", "_processed.csv")

walk2(
  EPA_processed_list,
  file.path(processed_dir, processed_files),
  ~ write.csv(.x, .y, row.names = FALSE)
)

processed_files
```

```
## [1] "EPAair_O3_NC2018_processed.csv"   "EPAair_O3_NC2019_processed.csv"
## [3] "EPAair_PM25_NC2018_processed.csv" "EPAair_PM25_NC2019_processed.csv"
```

Combine datasets

7. Combine the four datasets with `rbind`. Make sure your column names are identical prior to running this code.
8. Use code to display the dimensions of the combined dataset.
9. Wrangle your new dataset with a pipe function (%>%) so that it fills the following conditions:

- Include only sites that the four data frames have in common:
  "Linville Falls", "Durham Armory", "Leggett", "Hattie Avenue",
  "Clemmons Middle", "Mendenhall School", "Frying Pan Mountain", "West Johnston Co.", "Garinger High School", "Castle Hayne", "Pitt Agri. Center", "Bryson City", "Millbrook School"
- Some sites have multiple measurements per day. Use the split-apply-combine strategy to generate daily means: group by date, site name, AQS parameter, and county. Take the mean of the AQI value, latitude, and longitude.
- Add new columns for "Month" and "Year" by parsing your "Date" column (hint: `lubridate` package)
- Hint: the dimensions of this dataset should be 14,752 x 9.

10. Spread your datasets such that AQI values for ozone and PM2.5 are in separate columns. Each location on a specific date should now occupy only one row.
11. Call up the dimensions of your new tidy dataset.
12. Save your processed dataset with the following file name: "EPAair_O3_PM25_NC1819_Processed.csv"

```
library(lubridate)

# 7
EPA_combined <- do.call(rbind, EPA_processed_list)

# 8
dim(EPA_combined)
```

```
## [1] 37893     7
```

```
# 9
common_sites <- c(
  "Linville Falls", "Durham Armory", "Leggett", "Hattie Avenue",
  "Clemmons Middle", "Mendenhall School", "Frying Pan Mountain",
  "West Johnston Co.", "Garinger High School", "Castle Hayne",
  "Pitt Agri. Center", "Bryson City", "Millbrook School"
)

EPA_daily <- EPA_combined %>%
  filter(Site.Name %in% common_sites) %>%
  group_by(Date, Site.Name, AQS_PARAMETER_DESC, COUNTY) %>%
  summarise(
    DAILY_AQI_VALUE = mean(DAILY_AQI_VALUE, na.rm = TRUE),
    SITE_LATITUDE   = mean(SITE_LATITUDE,   na.rm = TRUE),
    SITE_LONGITUDE  = mean(SITE_LONGITUDE,  na.rm = TRUE),
    .groups = "drop"
```

```
  ) %>%
  mutate(
    Month = month(Date),
    Year  = year(Date)
  )

# Hint check (should be 14752 x 9)
dim(EPA_daily)
```

```
## [1] 14752     9
```

```
# 10
EPA_tidy <- EPA_daily %>%
  mutate(
    Param = case_when(
      str_detect(tolower(AQS_PARAMETER_DESC), "o3|ozone") ~ "Ozone",
      str_detect(tolower(AQS_PARAMETER_DESC), "pm") ~ "PM2.5",
      TRUE ~ NA_character_
    )
  ) %>%
  drop_na(Param) %>%   # only keep Ozone + PM2.5 rows
  select(Date, Site.Name, COUNTY, SITE_LATITUDE, SITE_LONGITUDE, Month, Year, Param, DAILY_AQI_VALUE) %
  pivot_wider(
    names_from = Param,
    values_from = DAILY_AQI_VALUE
  )

# 11
dim(EPA_tidy)
```

```
## [1] 8976    9
```

```
# 12
final_file <- here("Data", "Processed", "EPAair_O3_PM25_NC1819_Processed.csv")
write.csv(EPA_tidy, final_file, row.names = FALSE)
final_file
```

```
## [1] "C:/Users/ /EDE_Spring2026/Data/Processed/EPAair_O3_PM25_NC1819_Processed.csv"
```

Generate summary tables

13. Use the split-apply-combine strategy to generate a summary data frame. Data should be split into groups by site, month, and year. Then compute the mean AQI values for ozone and PM2.5 for each group. Finally, add a pipe to remove instances where the mean **ozone** values are not available (use the function `drop_na` in your pipe). It's ok to have missing mean PM2.5 values in this result.

14. Call up the dimensions of the summary dataset.

```
#13
EPA_summary <- EPA_tidy %>%
  group_by(Site.Name, Month, Year) %>%
```

```
  summarise(
    mean_ozone = mean(Ozone, na.rm = TRUE),
    mean_pm25  = mean(`PM2.5`, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  drop_na(mean_ozone)    # drop only when mean ozone is missing

#14
dim(EPA_summary)
```

```
## [1] 239   5
```

15. Why did we use the function `drop_na` rather than `na.omit`? Hint: replace `drop_na` with `na.omit` in part 13 and observe what happens with the dimensions of the summary date frame.

    Answer: We used drop_na(mean_ozone) because it lets us remove rows only when the mean ozone value is missing. In this assignment, it's okay for the mean PM2.5 value to be missing, so we don't want to delete those rows. If we used na.omit(), it would drop any row that has an NA in any column (including PM2.5), which would remove additional rows and make the summary dataset smaller than intended.

16. Stage, commit, and push your Assignment to your GitHub account. Provide a link to your repository below.

    Github repository URL: https://github.com/hegao188/EDE_Spring2026