Assignment 4: Data Wrangling (Spring 2025)

Christopher Starr

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Wrangling

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 does not extend off the page in the PDF.

Set up your session

1a. Load the `tidyverse`, `lubridate`, 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 - Load up packages
library(tidyverse)
library(lubridate)
library(here)


#1b - check my working directory

here::here()
```

```
## [1] "/home/guest/New Git Spring 2025"
```

```
#1c - reading raw data files (all of the EPAair files)

EPAair_O3_NC2018_raw <- read.csv(
```

```r
  file=here("Data/Raw/EPAair_O3_NC2018_raw.csv"),
  stringsAsFactors = TRUE
)

EPAair_O3_NC2019_raw <- read.csv(
  file=here("Data/Raw/EPAair_O3_NC2019_raw.csv"),
  stringsAsFactors = TRUE
)

EPAair_PM25_NC2018_raw <- read.csv(
  file=here("Data/Raw/EPAair_PM25_NC2018_raw.csv"),
  stringsAsFactors = TRUE
)

EPAair_PM25_NC2019_raw <- read.csv(
  file=here("Data/Raw/EPAair_PM25_NC2019_raw.csv"),
  stringsAsFactors = TRUE
)

#2 - checking on the dimensions of the four sets of data.

dim(EPAair_O3_NC2018_raw)
```

```
## [1] 9737   20
```

```r
dim(EPAair_O3_NC2019_raw)
```

```
## [1] 10592    20
```

```r
dim(EPAair_PM25_NC2018_raw)
```

```
## [1] 8983   20
```

```r
dim(EPAair_PM25_NC2019_raw)
```

```
## [1] 8581   20
```

> 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 the Date columns to be date objects.
4. Select 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 cells 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".

```
#3 - change Date columns from Factor to Date

#checking the class of the data before my code
class(EPAair_O3_NC2018_raw$Date)
```

```
## [1] "factor"
```

```
#changing the class to Date with format Y/M/D
EPAair_O3_NC2018_raw$Date <- mdy(EPAair_O3_NC2018_raw$Date)

EPAair_O3_NC2019_raw$Date <- mdy(EPAair_O3_NC2019_raw$Date)

EPAair_PM25_NC2018_raw$Date <- mdy(EPAair_PM25_NC2018_raw$Date)

EPAair_PM25_NC2019_raw$Date <- mdy(EPAair_PM25_NC2019_raw$Date)

#checking the format after the code
class(EPAair_O3_NC2018_raw$Date)
```

```
## [1] "Date"
```

```
#4

#creating four new datasets that only include 7 vectors
EPAair_O3_NC2018_processed <-
  EPAair_O3_NC2018_raw %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY,
         SITE_LATITUDE, SITE_LONGITUDE)

EPAair_O3_NC2019_processed <-
  EPAair_O3_NC2019_raw %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY,
         SITE_LATITUDE, SITE_LONGITUDE)

EPAair_PM25_NC2018_processed <-
EPAair_PM25_NC2018_raw %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY,
         SITE_LATITUDE, SITE_LONGITUDE)

EPAair_PM25_NC2019_processed <-
  EPAair_PM25_NC2019_raw %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY,
         SITE_LATITUDE, SITE_LONGITUDE)

#5
#changing two datasets for which all vectors were measured at PM2.5 so that they all read
#2.5 rather than what they currently have which is a lot of extra text/description
EPAair_PM25_NC2018_processed <-
  EPAair_PM25_NC2018_processed %>%
```

```
  mutate(AQS_PARAMETER_DESC = "PM2.5")

EPAair_PM25_NC2019_processed <-
  EPAair_PM25_NC2019_processed %>%
  mutate(AQS_PARAMETER_DESC = "PM2.5")

#6

#Writing my processed files to be in the processed data folder.
write_csv(EPAair_O3_NC2018_processed,
          file = "Data/Processed/EPAair_O3_NC2018_processed.csv")

write_csv(EPAair_O3_NC2019_processed,
          file = "Data/Processed/EPAair_O3_NC2019_processed.csv")

write_csv(EPAair_PM25_NC2018_processed,
          file = "Data/Processed/EPAair_PM25_NC2018_processed.csv")

write_csv(EPAair_PM25_NC2019_processed,
          file = "Data/Processed/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. 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"

(the function `intersect` can figure out common factor levels - but it will include sites with missing site information, which you don't want…)

- 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 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.

9. 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.
10. Call up the dimensions of your new tidy dataset.
11. Save your processed dataset with the following file name: "EPAair_O3_PM25_NC1819_Processed.csv"

```r
#7
#combine the four datasets into one super-duper dataset
O3.18 <- read.csv("Data/Processed/EPAair_O3_NC2018_processed.csv")
O3.19 <- read.csv("Data/Processed/EPAair_O3_NC2019_processed.csv")
PM25.18 <- read.csv("Data/Processed/EPAair_PM25_NC2018_processed.csv")
PM25.19 <- read.csv("Data/Processed/EPAair_PM25_NC2019_processed.csv")

EPAair_O3_PM25_NC1819_Processed <- rbind(O3.18, O3.19, PM25.18, PM25.19)

#8
#create a list of the sites all four datasets have in common
Sites_in_Common <- 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")

#Take the mean of the AQI value, latitude, and longitude -- by date, site name,
#AQS parameter, and county. then add columns for "Month" and "Year" by parsing your "Date"
EPAair_O3_PM25_NC1819_Processed <-
  EPAair_O3_PM25_NC1819_Processed %>%
  filter(Site.Name %in% Sites_in_Common & !is.na(Site.Name)) %>%
  group_by(Date, Site.Name, COUNTY, AQS_PARAMETER_DESC) %>%
  summarise(meanAQI = mean(DAILY_AQI_VALUE), meanLAT = mean(SITE_LATITUDE),
            meanLON = mean(SITE_LONGITUDE)) %>%
  mutate(Month = month(Date), Year = year(Date))
```

```
## `summarise()` has grouped output by 'Date', 'Site.Name', 'COUNTY'. You can
## override using the `.groups` argument.
```

```r
#9
#spread the AQS_Parameter into two columns, one for Ozone and one for Particle
#Matter larger than 2.5.
EPAair_O3_PM25_NC1819_Processed <- EPAair_O3_PM25_NC1819_Processed %>%
  pivot_wider(
    names_from = AQS_PARAMETER_DESC,
    values_from = meanAQI
  )

#10
#check the new dimensions
dim(EPAair_O3_PM25_NC1819_Processed)
```

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

```r
#11
#save my adjusted dataset
write_csv(EPAair_PM25_NC2019_processed,
          file = "Data/Processed/EPAair_O3_PM25_NC1819_Processed.csv")
```

Generate summary tables

12. Use the split-apply-combine strategy to generate a summary data frame. Data should be grouped by site, month, and year. Generate the mean AQI values for ozone and PM2.5 for each group. Then, add

a pipe to remove instances where 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.

13. Call up the dimensions of the summary dataset.

```
#12
#creating a summary - getting ozone and particular measure means for each month and
#each site
Summary_EPAair_O3_PM25_NC1819_Processed <-
  EPAair_O3_PM25_NC1819_Processed %>%
  group_by(Site.Name, Month, Year) %>%
  summarise(
    meanOzone = mean(Ozone, na.rm = TRUE),
    meanPM2.5 = mean(PM2.5, na.rm = TRUE)
  ) %>%
  drop_na(meanOzone)
```

```
## `summarise()` has grouped output by 'Site.Name', 'Month'. You can override
## using the `.groups` argument.
```

```
# Now with na.omit()
summary_with_na_omit <-
  EPAair_O3_PM25_NC1819_Processed %>%
  group_by(Site.Name, Month, Year) %>%
  summarise(
    meanOzone = mean(Ozone, na.rm = TRUE),
    meanPM2.5 = mean(PM2.5, na.rm = TRUE)
  ) %>%
  na.omit()
```

```
## `summarise()` has grouped output by 'Site.Name', 'Month'. You can override
## using the `.groups` argument.
```

```
#13
#check the new dimensions
dim(Summary_EPAair_O3_PM25_NC1819_Processed)
```

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

```
dim(summary_with_na_omit)
```

```
## [1] 223   5
```

14. Why did we use the function `drop_na` rather than `na.omit`? Hint: replace `drop_na` with `na.omit` in part 12 and observe what happens with the dimensions of the summary date frame.

    Answer: drop_na(meanOzone) only removes rows where the meanOzone column has NA values – na.omit() removes rows where ANY column has NA values, including meanPM2.5 We used drop_na() because we specifically wanted to keep rows that had Ozone measurements but were missing PM2.5 measurements.