Assignment 4: Data Wrangling (Spring 2026)

Sky Volz

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
here()
```

```
## [1] "/home/guest/EDE_Spring2026_personal"
```

```
#1c
EPAair.O3.NC2018 <- read.csv(
  file=here("Data/Raw/EPAair_O3_NC2018_raw.csv"),
  stringsAsFactors = TRUE)
```

```
EPAair.03.NC2019 <- read.csv(
  file=here("Data/Raw/EPAair_O3_NC2019_raw.csv"),
  stringsAsFactors = TRUE)

EPAair.PM25.NC2018 <- read.csv(
  file=here("Data/Raw/EPAair_PM25_NC2018_raw.csv"),
  stringsAsFactors = TRUE)

EPAair.PM25.NC2019 <- read.csv(
  file=here("Data/Raw/EPAair_PM25_NC2019_raw.csv"),
  stringsAsFactors = TRUE)

#2
#finding dimensions
dim(EPAair.03.NC2018)
```

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

```
dim(EPAair.03.NC2019)
```

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

```
dim(EPAair.PM25.NC2018)
```

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

```
dim(EPAair.PM25.NC2019)
```

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

```
#each has 20 columns
```

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

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".

```
#3
#using lubridate to make it a date, month day year is the format the date came in
EPAair.03.NC2018$Date <- mdy(EPAair.03.NC2018$Date)
EPAair.03.NC2019$Date <- mdy(EPAair.03.NC2019$Date)
EPAair.PM25.NC2018$Date <- mdy(EPAair.PM25.NC2018$Date)
EPAair.PM25.NC2019$Date <- mdy(EPAair.PM25.NC2019$Date)

#4
#selecting for the columns I want
Air.03.2018.select <- select(EPAair.03.NC2018, Date, DAILY_AQI_VALUE,
                             Site.Name, AQS_PARAMETER_DESC, COUNTY,
                             SITE_LATITUDE, SITE_LONGITUDE)

Air.03.2019.select <- select(EPAair.03.NC2019, Date, DAILY_AQI_VALUE,
                             Site.Name, AQS_PARAMETER_DESC, COUNTY,
                             SITE_LATITUDE, SITE_LONGITUDE)

Air.PM25.2018.select <- select(EPAair.PM25.NC2018, Date, DAILY_AQI_VALUE,
                               Site.Name, AQS_PARAMETER_DESC, COUNTY,
                               SITE_LATITUDE, SITE_LONGITUDE)

Air.PM25.2019.select <- select(EPAair.PM25.NC2019, Date, DAILY_AQI_VALUE,
                               Site.Name, AQS_PARAMETER_DESC, COUNTY,
                               SITE_LATITUDE, SITE_LONGITUDE)

#5
#assigning a value to the entire column to replace the old value
Air.PM25.2018.select$AQS_PARAMETER_DESC <- "PM2.5"

Air.PM25.2019.select$AQS_PARAMETER_DESC <- "PM2.5"

#6
write.csv(Air.03.2018.select,
  file=here("Data/Processed/EPAair_03_NC2018_Processed.csv"))
write.csv(Air.03.2019.select,
  file=here("Data/Processed/EPAair_03_NC2019_Processed.csv"))
write.csv(Air.PM25.2018.select,
  file=here("Data/Processed/EPAair_PM25_NC2018_Processed.csv"))
write.csv(Air.PM25.2019.select,
  file=here("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. 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"

```
#7
EPAair.data.all <- rbind(Air.03.2018.select, Air.03.2019.select,
                         Air.PM25.2018.select, Air.PM25.2019.select)

#8
dim(EPAair.data.all)
```

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

```
#9
#filtering using %in% to show only the sites names included in the vector
EPAair.common.sites <- EPAair.data.all %>%
  filter(Site.Name %in% 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" )) %>%
  group_by(Date, Site.Name, AQS_PARAMETER_DESC, COUNTY) %>% #split=group
  summarise(meanAQI = mean(DAILY_AQI_VALUE), #apply=summarizing
            mean_lat = mean(SITE_LATITUDE),
            mean_long = mean(SITE_LONGITUDE))
```

```
## `summarise()` has regrouped the output.
## i Summaries were computed grouped by Date, Site.Name, AQS_PARAMETER_DESC, and
##   COUNTY.
## i Output is grouped by Date, Site.Name, and AQS_PARAMETER_DESC.
## i Use `summarise(.groups = "drop_last")` to silence this message.
## i Use `summarise(.by = c(Date, Site.Name, AQS_PARAMETER_DESC, COUNTY))` for
##   per-operation grouping (`?dplyr::dplyr_by`) instead.
```

```
#using lubridate to create month and year columns
EPAair.common.sites <- mutate(EPAair.common.sites, month = month(Date), year = year(Date))

#checking dimensions
dim(EPAair.common.sites)
```

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

```
#10
# Spread AQI values into separate columns using pivot_wider
#for every ozone row, value from meanAQI - same for PM2.5
EPAair.common.sites.spread <- EPAair.common.sites %>%
  pivot_wider(names_from = AQS_PARAMETER_DESC, values_from = meanAQI)

#reorganizing columns so the layout makes more sense
EPAair.common.sites.organized <- EPAair.common.sites.spread %>%
  select(Date, year, month, Site.Name:Ozone)


#11
dim(EPAair.common.sites.organized)
```

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

```
#12
write.csv(EPAair.common.sites.organized,
          file=here("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
EPAair.summary <- EPAair.common.sites.organized %>%
  group_by(Site.Name, month, year) %>% #split into groups
  summarise(mean_ozone = mean(Ozone),
            mean_PM2.5 = mean(PM2.5)) %>% #apply summary functions
  drop_na(mean_ozone) #dropping n/a values from the summary ozone column
```

```
## `summarise()` has regrouped the output.
## i Summaries were computed grouped by Site.Name, month, and year.
## i Output is grouped by Site.Name and month.
## i Use `summarise(.groups = "drop_last")` to silence this message.
## i Use `summarise(.by = c(Site.Name, month, year))` for per-operation grouping
##   (`?dplyr::dplyr_by`) instead.
```

```
#14
dim(EPAair.summary)
```

```
## [1] 182   5
```

```
write.csv(EPAair.summary,
          file=here("Data/Processed/EPAair_O3_PM25_NC1819_Summary_Processed.csv"))
```

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: When I used na.omit, it got rid of the n/a in the PM2.5 column, which got rid of 81 rows of data that I had wanted to keep. By using the drop_na and telling it to only drop the n/a values in the ozone column, it still kept rows that had n/a values in the PM2.5 column. The drop_na command is more selective for when I want to omit specific things.

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

    Github repository URL: https://github.com/skyvolz/EDE_Spring2026_personal