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)
library(readr)


#1b - check my working directory

here::here()
```

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

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

EPAair_O3_NC2018_raw <- read.csv(
  file=here("Data/Raw/EPAair_PM25_NC2019_raw.csv"),
  stringsAsFactors = TRUE
)

EPAair_O3_NC2019_raw <- read.csv(
  file=here("Data/Raw/EPAair_PM25_NC2019_raw.csv"),
  stringsAsFactors = TRUE
)

EPAair_PM25_NC2018_raw <- read.csv(
  file=here("Data/Raw/EPAair_PM25_NC2019_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.

summary(EPAair_O3_NC2018_raw)
```

```
##          Date         Source        Site.ID               POC
##  02/26/2019:  41   AirNow:1670   Min.   :370110002   Min.   :1.000
##  01/21/2019:  40   AQS   :6911   1st Qu.:370630015   1st Qu.:3.000
##  02/14/2019:  40                 Median :371190041   Median :3.000
##  01/09/2019:  39                 Mean   :371023743   Mean   :3.032
##  01/27/2019:  39                 3rd Qu.:371290002   3rd Qu.:3.000
##  02/02/2019:  39                 Max.   :371830021   Max.   :5.000
##  (Other)   :8343
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-3.100                 ug/m3 LC:8581   Min.   : 0.00
##  1st Qu.: 4.900                                 1st Qu.:20.00
##  Median : 7.400                                 Median :31.00
##  Mean   : 7.684                                 Mean   :31.51
##  3rd Qu.:10.100                                 3rd Qu.:42.00
##  Max.   :31.200                                 Max.   :91.00
##
##                 Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Millbrook School    : 738   Min.   :1       Min.   :100
##  Garinger High School: 629   1st Qu.:1       1st Qu.:100
##  Remount             : 573   Median :1       Median :100
##  Hickory Water Tower : 518   Mean   :1       Mean   :100
##  Hattie Avenue       : 436   3rd Qu.:1       3rd Qu.:100
##  Durham Armory       : 431   Max.   :1       Max.   :100
##  (Other)             :5256
##  AQS_PARAMETER_CODE                              AQS_PARAMETER_DESC
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:1029
##  1st Qu.:88101      PM2.5 - Local Conditions              :7552
```

```
##  Median :88101
##  Mean   :88149
##  3rd Qu.:88101
##  Max.   :88502
##
##    CBSA_CODE                                  CBSA_NAME      STATE_CODE
##  Min.   :11700   Raleigh, NC                      :1441   Min.   :37
##  1st Qu.:19000   Charlotte-Concord-Gastonia, NC-SC:1379   1st Qu.:37
##  Median :25860   Winston-Salem, NC                :1235   Median :37
##  Mean   :31099                                    :1058   Mean   :37
##  3rd Qu.:40580   Hickory-Lenoir-Morganton, NC     : 518   3rd Qu.:37
##  Max.   :49180   Durham-Chapel Hill, NC           : 431   Max.   :37
##  NA's   :1058    (Other)                          :2519
##             STATE       COUNTY_CODE            COUNTY     SITE_LATITUDE
##  North Carolina:8581   Min.   : 11.0   Mecklenburg:1379   Min.   :34.36
##                        1st Qu.: 63.0   Wake       :1083   1st Qu.:35.26
##                        Median :119.0   Forsyth    : 839   Median :35.73
##                        Mean   :102.4   Catawba    : 518   Mean   :35.63
##                        3rd Qu.:129.0   Durham     : 431   3rd Qu.:35.91
##                        Max.   :183.0   Cumberland : 427   Max.   :36.51
##                                        (Other)    :3904
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-79.95
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

```
summary(EPAair_O3_NC2019_raw)
```

```
##          Date          Source        Site.ID                POC
##  02/26/2019:  41   AirNow:1670   Min.   :370110002   Min.   :1.000
##  01/21/2019:  40   AQS   :6911   1st Qu.:370630015   1st Qu.:3.000
##  02/14/2019:  40                 Median :371190041   Median :3.000
##  01/09/2019:  39                 Mean   :371023743   Mean   :3.032
##  01/27/2019:  39                 3rd Qu.:371290002   3rd Qu.:3.000
##  02/02/2019:  39                 Max.   :371830021   Max.   :5.000
##  (Other)   :8343
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-3.100                 ug/m3 LC:8581   Min.   : 0.00
##  1st Qu.: 4.900                                 1st Qu.:20.00
##  Median : 7.400                                 Median :31.00
##  Mean   : 7.684                                 Mean   :31.51
##  3rd Qu.:10.100                                 3rd Qu.:42.00
##  Max.   :31.200                                 Max.   :91.00
##
##                Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Millbrook School    : 738   Min.   :1       Min.   :100
##  Garinger High School: 629   1st Qu.:1       1st Qu.:100
##  Remount             : 573   Median :1       Median :100
##  Hickory Water Tower : 518   Mean   :1       Mean   :100
##  Hattie Avenue       : 436   3rd Qu.:1       3rd Qu.:100
```

```
##  Durham Armory       : 431   Max.   :1       Max.   :100
##  (Other)             :5256
##  AQS_PARAMETER_CODE                                AQS_PARAMETER_DESC
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:1029
##  1st Qu.:88101      PM2.5 - Local Conditions              :7552
##  Median :88101
##  Mean   :88149
##  3rd Qu.:88101
##  Max.   :88502
##
##    CBSA_CODE                                    CBSA_NAME      STATE_CODE
##  Min.   :11700   Raleigh, NC                         :1441   Min.   :37
##  1st Qu.:19000   Charlotte-Concord-Gastonia, NC-SC:1379   1st Qu.:37
##  Median :25860   Winston-Salem, NC                   :1235   Median :37
##  Mean   :31099                                       :1058   Mean   :37
##  3rd Qu.:40580   Hickory-Lenoir-Morganton, NC        : 518   3rd Qu.:37
##  Max.   :49180   Durham-Chapel Hill, NC              : 431   Max.   :37
##  NA's   :1058    (Other)                             :2519
##              STATE       COUNTY_CODE             COUNTY     SITE_LATITUDE
##  North Carolina:8581   Min.   : 11.0   Mecklenburg:1379   Min.   :34.36
##                        1st Qu.: 63.0   Wake       :1083   1st Qu.:35.26
##                        Median :119.0   Forsyth    : 839   Median :35.73
##                        Mean   :102.4   Catawba    : 518   Mean   :35.63
##                        3rd Qu.:129.0   Durham     : 431   3rd Qu.:35.91
##                        Max.   :183.0   Cumberland : 427   Max.   :36.51
##                                        (Other)    :3904
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-79.95
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

```
summary(EPAair_PM25_NC2018_raw)
```

```
##          Date         Source        Site.ID               POC
##  02/26/2019:  41   AirNow:1670   Min.   :370110002   Min.   :1.000
##  01/21/2019:  40   AQS   :6911   1st Qu.:370630015   1st Qu.:3.000
##  02/14/2019:  40                 Median :371190041   Median :3.000
##  01/09/2019:  39                 Mean   :371023743   Mean   :3.032
##  01/27/2019:  39                 3rd Qu.:371290002   3rd Qu.:3.000
##  02/02/2019:  39                 Max.   :371830021   Max.   :5.000
##  (Other)   :8343
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-3.100                ug/m3 LC:8581   Min.   : 0.00
##  1st Qu.: 4.900                                1st Qu.:20.00
##  Median : 7.400                                Median :31.00
##  Mean   : 7.684                                Mean   :31.51
##  3rd Qu.:10.100                                3rd Qu.:42.00
##  Max.   :31.200                                Max.   :91.00
##
##                Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
```

```
##  Millbrook School    : 738   Min.   :1       Min.   :100
##  Garinger High School: 629   1st Qu.:1       1st Qu.:100
##  Remount             : 573   Median :1       Median :100
##  Hickory Water Tower : 518   Mean   :1       Mean   :100
##  Hattie Avenue       : 436   3rd Qu.:1       3rd Qu.:100
##  Durham Armory       : 431   Max.   :1       Max.   :100
##  (Other)             :5256
##  AQS_PARAMETER_CODE                              AQS_PARAMETER_DESC
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:1029
##  1st Qu.:88101      PM2.5 - Local Conditions              :7552
##  Median :88101
##  Mean   :88149
##  3rd Qu.:88101
##  Max.   :88502
##
##    CBSA_CODE                                  CBSA_NAME      STATE_CODE
##  Min.   :11700   Raleigh, NC                       :1441   Min.   :37
##  1st Qu.:19000   Charlotte-Concord-Gastonia, NC-SC:1379   1st Qu.:37
##  Median :25860   Winston-Salem, NC                 :1235   Median :37
##  Mean   :31099                                     :1058   Mean   :37
##  3rd Qu.:40580   Hickory-Lenoir-Morganton, NC      : 518   3rd Qu.:37
##  Max.   :49180   Durham-Chapel Hill, NC            : 431   Max.   :37
##  NA's   :1058    (Other)                           :2519
##             STATE       COUNTY_CODE            COUNTY     SITE_LATITUDE
##  North Carolina:8581   Min.   : 11.0   Mecklenburg:1379   Min.   :34.36
##                        1st Qu.: 63.0   Wake        :1083   1st Qu.:35.26
##                        Median :119.0   Forsyth     : 839   Median :35.73
##                        Mean   :102.4   Catawba     : 518   Mean   :35.63
##                        3rd Qu.:129.0   Durham      : 431   3rd Qu.:35.91
##                        Max.   :183.0   Cumberland  : 427   Max.   :36.51
##                                        (Other)     :3904
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-79.95
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

```
summary(EPAair_PM25_NC2019_raw)
```

```
##          Date         Source        Site.ID               POC
##  02/26/2019:  41   AirNow:1670   Min.   :370110002   Min.   :1.000
##  01/21/2019:  40   AQS   :6911   1st Qu.:370630015   1st Qu.:3.000
##  02/14/2019:  40                 Median :371190041   Median :3.000
##  01/09/2019:  39                 Mean   :371023743   Mean   :3.032
##  01/27/2019:  39                 3rd Qu.:371290002   3rd Qu.:3.000
##  02/02/2019:  39                 Max.   :371830021   Max.   :5.000
##  (Other)   :8343
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-3.100                ug/m3 LC:8581   Min.   : 0.00
##  1st Qu.: 4.900                                1st Qu.:20.00
##  Median : 7.400                                Median :31.00
```

```
##  Mean   : 7.684                                    Mean   :31.51
##  3rd Qu.:10.100                                    3rd Qu.:42.00
##  Max.   :31.200                                    Max.   :91.00
##
##                 Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Millbrook School    : 738   Min.   :1       Min.   :100
##  Garinger High School: 629   1st Qu.:1       1st Qu.:100
##  Remount             : 573   Median :1       Median :100
##  Hickory Water Tower : 518   Mean   :1       Mean   :100
##  Hattie Avenue       : 436   3rd Qu.:1       3rd Qu.:100
##  Durham Armory       : 431   Max.   :1       Max.   :100
##  (Other)             :5256
##  AQS_PARAMETER_CODE                                AQS_PARAMETER_DESC
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:1029
##  1st Qu.:88101      PM2.5 - Local Conditions              :7552
##  Median :88101
##  Mean   :88149
##  3rd Qu.:88101
##  Max.   :88502
##
##    CBSA_CODE                                  CBSA_NAME      STATE_CODE
##  Min.   :11700   Raleigh, NC                      :1441   Min.   :37
##  1st Qu.:19000   Charlotte-Concord-Gastonia, NC-SC:1379   1st Qu.:37
##  Median :25860   Winston-Salem, NC                :1235   Median :37
##  Mean   :31099                                    :1058   Mean   :37
##  3rd Qu.:40580   Hickory-Lenoir-Morganton, NC     : 518   3rd Qu.:37
##  Max.   :49180   Durham-Chapel Hill, NC           : 431   Max.   :37
##  NA's   :1058    (Other)                          :2519
##              STATE       COUNTY_CODE            COUNTY     SITE_LATITUDE
##  North Carolina:8581   Min.   : 11.0   Mecklenburg:1379   Min.   :34.36
##                        1st Qu.: 63.0   Wake       :1083   1st Qu.:35.26
##                        Median :119.0   Forsyth    : 839   Median :35.73
##                        Mean   :102.4   Catawba    : 518   Mean   :35.63
##                        3rd Qu.:129.0   Durham     : 431   3rd Qu.:35.91
##                        Max.   :183.0   Cumberland : 427   Max.   :36.51
##                                        (Other)    :3904
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-79.95
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

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$Date <- as.Date(EPAair_O3_NC2018_raw$Date, format = "%Y-%m-%d")

EPAair_O3_NC2019_raw$Date <- as.Date(EPAair_O3_NC2019_raw$Date, format = "%Y-%m-%d")

EPAair_PM25_NC2018_raw$Date <- as.Date(EPAair_PM25_NC2018_raw$Date, format = "%Y-%m-%d")

EPAair_PM25_NC2019_raw$Date <- as.Date(EPAair_PM25_NC2019_raw$Date, format = "%Y-%m-%d")

#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"

```
#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
Sites_in_Common <- EPAair_O3_PM25_NC1819_Processed$Site.Name %in% c("Linville Falls", "Durham Armory",

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, AQS_PARAMETER_DESC, COUNTY) %>%
  summarise(meanAQI = mean(DAILY_AQI_VALUE), meanLAT = mean(SITE_LATITUDE), meanLON = mean(SITE_LONGITU
  mutate(Month = month(Date), Year = year(Date))
```

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

```
#9


#10


#11
```

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


#13
```

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: