Assignment 4: Data Wrangling (Fall 2024)

Lauren Shohan

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
library(tidyverse)
library(lubridate)
library(here)

#1b
#getwd()

here::here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
#1c
Airdata_03_2018 <- read.csv(here('./Data/Raw/EPAair_O3_NC2018_raw.csv'),
                            stringsAsFactors = TRUE)
head(Airdata_03_2018)
```

```
##         Date Source   Site.ID POC Daily.Max.8.hour.Ozone.Concentration UNITS
## 1 03/01/2018    AQS 370030005   1                                 0.043   ppm
## 2 03/02/2018    AQS 370030005   1                                 0.046   ppm
## 3 03/03/2018    AQS 370030005   1                                 0.047   ppm
## 4 03/04/2018    AQS 370030005   1                                 0.049   ppm
## 5 03/05/2018    AQS 370030005   1                                 0.047   ppm
## 6 03/06/2018    AQS 370030005   1                                 0.030   ppm
##   DAILY_AQI_VALUE            Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              40 Taylorsville Liledoun              17              100
## 2              43 Taylorsville Liledoun              17              100
## 3              44 Taylorsville Liledoun              17              100
## 4              45 Taylorsville Liledoun              17              100
## 5              44 Taylorsville Liledoun              17              100
## 6              28 Taylorsville Liledoun              17              100
##   AQS_PARAMETER_CODE AQS_PARAMETER_DESC CBSA_CODE                     CBSA_NAME
## 1              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 2              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 3              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 4              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 5              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 6              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
##   STATE_CODE          STATE COUNTY_CODE    COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1         37 North Carolina           3 Alexander       35.9138        -81.191
## 2         37 North Carolina           3 Alexander       35.9138        -81.191
## 3         37 North Carolina           3 Alexander       35.9138        -81.191
## 4         37 North Carolina           3 Alexander       35.9138        -81.191
## 5         37 North Carolina           3 Alexander       35.9138        -81.191
## 6         37 North Carolina           3 Alexander       35.9138        -81.191
```

```
Airdata_03_2019 <- read.csv(file = here('./Data/Raw/EPAair_O3_NC2019_raw.csv'),
                            stringsAsFactors = TRUE)
head(Airdata_03_2019)
```

```
##         Date Source   Site.ID POC Daily.Max.8.hour.Ozone.Concentration UNITS
## 1 01/01/2019 AirNow 370030005   1                                 0.029   ppm
## 2 01/02/2019 AirNow 370030005   1                                 0.018   ppm
## 3 01/03/2019 AirNow 370030005   1                                 0.016   ppm
## 4 01/04/2019 AirNow 370030005   1                                 0.022   ppm
## 5 01/05/2019 AirNow 370030005   1                                 0.037   ppm
## 6 01/06/2019 AirNow 370030005   1                                 0.037   ppm
##   DAILY_AQI_VALUE            Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              27 Taylorsville Liledoun              24              100
## 2              17 Taylorsville Liledoun              24              100
## 3              15 Taylorsville Liledoun              24              100
## 4              20 Taylorsville Liledoun              24              100
## 5              34 Taylorsville Liledoun              24              100
## 6              34 Taylorsville Liledoun              24              100
##   AQS_PARAMETER_CODE AQS_PARAMETER_DESC CBSA_CODE                     CBSA_NAME
## 1              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 2              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 3              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 4              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 5              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 6              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
```

```
##   STATE_CODE          STATE COUNTY_CODE    COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1         37 North Carolina           3 Alexander       35.9138        -81.191
## 2         37 North Carolina           3 Alexander       35.9138        -81.191
## 3         37 North Carolina           3 Alexander       35.9138        -81.191
## 4         37 North Carolina           3 Alexander       35.9138        -81.191
## 5         37 North Carolina           3 Alexander       35.9138        -81.191
## 6         37 North Carolina           3 Alexander       35.9138        -81.191
```

```
Airdata_pm25_2018 <- read.csv(file = here('./Data/Raw/EPAair_PM25_NC2018_raw.csv'),
                              stringsAsFactors = TRUE)
head(Airdata_pm25_2018)
```

```
##         Date Source   Site.ID POC Daily.Mean.PM2.5.Concentration    UNITS
## 1 01/02/2018    AQS 370110002   1                             2.9 ug/m3 LC
## 2 01/05/2018    AQS 370110002   1                             3.7 ug/m3 LC
## 3 01/08/2018    AQS 370110002   1                             5.3 ug/m3 LC
## 4 01/11/2018    AQS 370110002   1                             0.8 ug/m3 LC
## 5 01/14/2018    AQS 370110002   1                             2.5 ug/m3 LC
## 6 01/17/2018    AQS 370110002   1                             4.5 ug/m3 LC
##   DAILY_AQI_VALUE      Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              12 Linville Falls               1              100
## 2              15 Linville Falls               1              100
## 3              22 Linville Falls               1              100
## 4               3 Linville Falls               1              100
## 5              10 Linville Falls               1              100
## 6              19 Linville Falls               1              100
##   AQS_PARAMETER_CODE                     AQS_PARAMETER_DESC CBSA_CODE CBSA_NAME
## 1              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 2              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 3              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 4              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 5              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 6              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
##   STATE_CODE          STATE COUNTY_CODE COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1         37 North Carolina          11  Avery      35.97235      -81.93307
## 2         37 North Carolina          11  Avery      35.97235      -81.93307
## 3         37 North Carolina          11  Avery      35.97235      -81.93307
## 4         37 North Carolina          11  Avery      35.97235      -81.93307
## 5         37 North Carolina          11  Avery      35.97235      -81.93307
## 6         37 North Carolina          11  Avery      35.97235      -81.93307
```

```
Airdata_pm25_2019 <- read.csv(file = here('./Data/Raw/EPAair_PM25_NC2019_raw.csv'),
                              stringsAsFactors = TRUE)
head(Airdata_pm25_2019)
```

```
##         Date Source   Site.ID POC Daily.Mean.PM2.5.Concentration    UNITS
## 1 01/03/2019    AQS 370110002   1                             1.6 ug/m3 LC
## 2 01/06/2019    AQS 370110002   1                             1.0 ug/m3 LC
## 3 01/09/2019    AQS 370110002   1                             1.3 ug/m3 LC
## 4 01/12/2019    AQS 370110002   1                             6.3 ug/m3 LC
## 5 01/15/2019    AQS 370110002   1                             2.6 ug/m3 LC
## 6 01/18/2019    AQS 370110002   1                             1.2 ug/m3 LC
##   DAILY_AQI_VALUE      Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
```

```
## 1               7 Linville Falls               1              100
## 2               4 Linville Falls               1              100
## 3               5 Linville Falls               1              100
## 4              26 Linville Falls               1              100
## 5              11 Linville Falls               1              100
## 6               5 Linville Falls               1              100
##   AQS_PARAMETER_CODE                     AQS_PARAMETER_DESC CBSA_CODE CBSA_NAME
## 1              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 2              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 3              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 4              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 5              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 6              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
##   STATE_CODE          STATE COUNTY_CODE COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1         37 North Carolina          11  Avery      35.97235      -81.93307
## 2         37 North Carolina          11  Avery      35.97235      -81.93307
## 3         37 North Carolina          11  Avery      35.97235      -81.93307
## 4         37 North Carolina          11  Avery      35.97235      -81.93307
## 5         37 North Carolina          11  Avery      35.97235      -81.93307
## 6         37 North Carolina          11  Avery      35.97235      -81.93307
```

```
#2
dim(Airdata_03_2018)
```

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

```
dim(Airdata_03_2019)
```

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

```
dim(Airdata_pm25_2018)
```

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

```
dim(Airdata_pm25_2019)
```

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

All four datasets should have the same number of columns but unique record counts (rows). Do your datasets follow this pattern? Yes, they all have 20 columns and unique record counts for the rows.

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
class(Airdata_03_2018$Date) #they were factors before
```

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

```
Airdata_03_2018$Date <- as.Date(Airdata_03_2018$Date, format = "%m/%d/%Y")
Airdata_03_2019$Date <- as.Date(Airdata_03_2019$Date, format = "%m/%d/%Y")
Airdata_pm25_2018$Date <- as.Date(Airdata_pm25_2018$Date, format = "%m/%d/%Y")
Airdata_pm25_2019$Date <- as.Date(Airdata_pm25_2019$Date, format = "%m/%d/%Y")

class(Airdata_03_2018$Date)
```

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

```
class(Airdata_03_2019$Date)
```

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

```
class(Airdata_pm25_2018$Date)
```

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

```
class(Airdata_pm25_2019$Date)
```

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

```
#4

Airdata_03_2018_subset <- Airdata_03_2018 %>%
  select(Date,DAILY_AQI_VALUE,Site.Name,
         AQS_PARAMETER_DESC,COUNTY,SITE_LATITUDE,SITE_LONGITUDE)
head(Airdata_03_2018_subset)
```

```
##         Date DAILY_AQI_VALUE             Site.Name AQS_PARAMETER_DESC    COUNTY
## 1 2018-03-01              40 Taylorsville Liledoun              Ozone Alexander
## 2 2018-03-02              43 Taylorsville Liledoun              Ozone Alexander
## 3 2018-03-03              44 Taylorsville Liledoun              Ozone Alexander
## 4 2018-03-04              45 Taylorsville Liledoun              Ozone Alexander
## 5 2018-03-05              44 Taylorsville Liledoun              Ozone Alexander
## 6 2018-03-06              28 Taylorsville Liledoun              Ozone Alexander
##   SITE_LATITUDE SITE_LONGITUDE
## 1       35.9138        -81.191
## 2       35.9138        -81.191
## 3       35.9138        -81.191
## 4       35.9138        -81.191
## 5       35.9138        -81.191
## 6       35.9138        -81.191
```

```
Airdata_03_2019_subset <- Airdata_03_2019 %>%
  select(Date,DAILY_AQI_VALUE,Site.Name,
         AQS_PARAMETER_DESC,COUNTY,SITE_LATITUDE,SITE_LONGITUDE)
head(Airdata_03_2019_subset)
```

```
##         Date DAILY_AQI_VALUE             Site.Name AQS_PARAMETER_DESC    COUNTY
## 1 2019-01-01              27 Taylorsville Liledoun              Ozone Alexander
## 2 2019-01-02              17 Taylorsville Liledoun              Ozone Alexander
## 3 2019-01-03              15 Taylorsville Liledoun              Ozone Alexander
## 4 2019-01-04              20 Taylorsville Liledoun              Ozone Alexander
## 5 2019-01-05              34 Taylorsville Liledoun              Ozone Alexander
## 6 2019-01-06              34 Taylorsville Liledoun              Ozone Alexander
##   SITE_LATITUDE SITE_LONGITUDE
## 1       35.9138        -81.191
## 2       35.9138        -81.191
## 3       35.9138        -81.191
## 4       35.9138        -81.191
## 5       35.9138        -81.191
## 6       35.9138        -81.191
```

```
Airdata_pm25_2018_subset <-  Airdata_pm25_2018 %>%
  select(Date,DAILY_AQI_VALUE,Site.Name,
         AQS_PARAMETER_DESC,COUNTY,SITE_LATITUDE,SITE_LONGITUDE)
head(Airdata_pm25_2018_subset)
```

```
##         Date DAILY_AQI_VALUE      Site.Name
## 1 2018-01-02              12 Linville Falls
## 2 2018-01-05              15 Linville Falls
## 3 2018-01-08              22 Linville Falls
## 4 2018-01-11               3 Linville Falls
## 5 2018-01-14              10 Linville Falls
## 6 2018-01-17              19 Linville Falls
##                       AQS_PARAMETER_DESC COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 2 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 3 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 4 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 5 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 6 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
```

```
Airdata_pm25_2019_subset <- Airdata_pm25_2019 %>%
  select(Date,DAILY_AQI_VALUE,Site.Name,
         AQS_PARAMETER_DESC,COUNTY,SITE_LATITUDE,SITE_LONGITUDE)
head(Airdata_pm25_2019_subset)
```

```
##         Date DAILY_AQI_VALUE      Site.Name
## 1 2019-01-03               7 Linville Falls
## 2 2019-01-06               4 Linville Falls
## 3 2019-01-09               5 Linville Falls
## 4 2019-01-12              26 Linville Falls
## 5 2019-01-15              11 Linville Falls
## 6 2019-01-18               5 Linville Falls
```

```
##                       AQS_PARAMETER_DESC COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 2 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 3 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 4 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 5 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 6 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
```

```
#5
#mutating that current column instead of creating new one
Airdata_pm25_2018_subset <- mutate(Airdata_pm25_2018_subset, AQS_PARAMETER_DESC = 'PM2.5')
head(Airdata_pm25_2018_subset)
```

```
##         Date DAILY_AQI_VALUE      Site.Name AQS_PARAMETER_DESC COUNTY
## 1 2018-01-02              12 Linville Falls              PM2.5  Avery
## 2 2018-01-05              15 Linville Falls              PM2.5  Avery
## 3 2018-01-08              22 Linville Falls              PM2.5  Avery
## 4 2018-01-11               3 Linville Falls              PM2.5  Avery
## 5 2018-01-14              10 Linville Falls              PM2.5  Avery
## 6 2018-01-17              19 Linville Falls              PM2.5  Avery
##   SITE_LATITUDE SITE_LONGITUDE
## 1      35.97235      -81.93307
## 2      35.97235      -81.93307
## 3      35.97235      -81.93307
## 4      35.97235      -81.93307
## 5      35.97235      -81.93307
## 6      35.97235      -81.93307
```

```
Airdata_pm25_2019_subset <- mutate(Airdata_pm25_2019_subset, AQS_PARAMETER_DESC = 'PM2.5')
head(Airdata_pm25_2019_subset)
```

```
##         Date DAILY_AQI_VALUE      Site.Name AQS_PARAMETER_DESC COUNTY
## 1 2019-01-03               7 Linville Falls              PM2.5  Avery
## 2 2019-01-06               4 Linville Falls              PM2.5  Avery
## 3 2019-01-09               5 Linville Falls              PM2.5  Avery
## 4 2019-01-12              26 Linville Falls              PM2.5  Avery
## 5 2019-01-15              11 Linville Falls              PM2.5  Avery
## 6 2019-01-18               5 Linville Falls              PM2.5  Avery
##   SITE_LATITUDE SITE_LONGITUDE
## 1      35.97235      -81.93307
## 2      35.97235      -81.93307
## 3      35.97235      -81.93307
## 4      35.97235      -81.93307
## 5      35.97235      -81.93307
## 6      35.97235      -81.93307
```

```
#6
write.csv(Airdata_pm25_2019_subset, row.names = FALSE,
          file = './Data/Processed/EPAair_PM25_NC2019_processed.csv')

write.csv(Airdata_pm25_2018_subset, row.names = FALSE,
          file = './Data/Processed/EPAair_PM25_NC2018_processed.csv')
```

```
write.csv(Airdata_03_2019_subset, row.names = FALSE,
          file = './Data/Processed/EPAair_O3_NC2019_processed.csv')

write.csv(Airdata_03_2018_subset, row.names = FALSE,
          file = './Data/Processed/EPAair_O3_NC2018_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 with `rbind`. Make sure your column names are identical prior to running
Airdata_allfour <- rbind(Airdata_03_2018_subset, Airdata_03_2019_subset, Airdata_pm25_2018_subset, Airda
head(Airdata_allfour)
```

```
##         Date DAILY_AQI_VALUE             Site.Name AQS_PARAMETER_DESC    COUNTY
## 1 2018-03-01              40 Taylorsville Liledoun              Ozone Alexander
## 2 2018-03-02              43 Taylorsville Liledoun              Ozone Alexander
## 3 2018-03-03              44 Taylorsville Liledoun              Ozone Alexander
## 4 2018-03-04              45 Taylorsville Liledoun              Ozone Alexander
## 5 2018-03-05              44 Taylorsville Liledoun              Ozone Alexander
## 6 2018-03-06              28 Taylorsville Liledoun              Ozone Alexander
##   SITE_LATITUDE SITE_LONGITUDE
## 1       35.9138        -81.191
## 2       35.9138        -81.191
## 3       35.9138        -81.191
```

```
## 4       35.9138        -81.191
## 5       35.9138        -81.191
## 6       35.9138        -81.191
```

```
#8 -Wrangle your new dataset with a pipe function (%>%)

Airdata_allfour_piped <-
  Airdata_allfour %>%
  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) %>%
  summarise(mean_daily_AQI = mean(DAILY_AQI_VALUE),
            mean_latitude = mean(SITE_LATITUDE),
            mean_long = mean(SITE_LONGITUDE)) %>%
  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.
```

```
head(Airdata_allfour_piped)
```

```
## # A tibble: 6 x 9
## # Groups:   Date, Site.Name, AQS_PARAMETER_DESC [6]
##   Date       Site.Name    AQS_PARAMETER_DESC COUNTY mean_daily_AQI mean_latitude
##   <date>     <fct>        <fct>              <fct>           <dbl>         <dbl>
## 1 2018-01-01 Bryson City  PM2.5              Swain              35          35.4
## 2 2018-01-01 Castle Hayne PM2.5              New H~             13          34.4
## 3 2018-01-01 Clemmons Mi~ PM2.5              Forsy~             24          36.0
## 4 2018-01-01 Durham Armo~ PM2.5              Durham             31          36.0
## 5 2018-01-01 Garinger Hi~ Ozone              Meckl~             32          35.2
## 6 2018-01-01 Garinger Hi~ PM2.5              Meckl~             20          35.2
## # i 3 more variables: mean_long <dbl>, Month <dbl>, Year <dbl>
```

```
#9 Spread your datasets such that AQI values for ozone and PM2.5 are in separate columns

Airdata_allfour_piped_spread <- pivot_wider(Airdata_allfour_piped, names_from = AQS_PARAMETER_DESC, valu
head(Airdata_allfour_piped_spread)
```

```
## # A tibble: 6 x 9
## # Groups:   Date, Site.Name [6]
##   Date       Site.Name    COUNTY mean_latitude mean_long Month  Year PM2.5 Ozone
```

```
##   <date>     <fct>        <fct>          <dbl>     <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 2018-01-01 Bryson City  Swain           35.4     -83.4     1  2018    35    NA
## 2 2018-01-01 Castle Hayne New H~          34.4     -77.8     1  2018    13    NA
## 3 2018-01-01 Clemmons Mi~ Forsy~          36.0     -80.3     1  2018    24    NA
## 4 2018-01-01 Durham Armo~ Durham          36.0     -78.9     1  2018    31    NA
## 5 2018-01-01 Garinger Hi~ Meckl~          35.2     -80.8     1  2018    20    32
## 6 2018-01-01 Hattie Aven~ Forsy~          36.1     -80.2     1  2018    22    NA
```

```
#10 - Call up the dimensions of your new tidy dataset.
dim(Airdata_allfour_piped_spread)
```

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

```
#11 Save your processed dataset with the following file name: "EPAair_O3_PM25_NC1819_Processed.csv"
write.csv(Airdata_allfour_piped_spread, row.names = FALSE,
          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
Airdata_piped_spread_summary <-
  Airdata_allfour_piped_spread %>%
  group_by(Site.Name, Month, Year) %>%
  summarise(mean_AQI_ozone = mean(Ozone),
            mean_AQI_pm25 = mean(PM2.5)) %>%
  drop_na(mean_AQI_ozone)
```

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

```
head(Airdata_piped_spread_summary)
```

```
## # A tibble: 6 x 5
## # Groups:   Site.Name, Month [4]
##   Site.Name   Month  Year mean_AQI_ozone mean_AQI_pm25
##   <fct>       <dbl> <dbl>          <dbl>         <dbl>
## 1 Bryson City     3  2018           41.6          34.7
## 2 Bryson City     3  2019           42.5          NA
## 3 Bryson City     4  2018           44.5          28.2
## 4 Bryson City     4  2019           45.4          26.7
## 5 Bryson City     5  2019           39.6          NA
## 6 Bryson City     6  2018           37.8          NA
```

```r
#13
dim(Airdata_piped_spread_summary)
```

```
## [1] 182   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: If I use na.omit, the dimensions went from 182 x 5 to 101 x 5. The na.omit deleted entire rows that have NA while with drop_na I can specify which column I want to filter NAs out of and in this case I specified the Ozone column to filter NAs out of so the PM2.5 column kept the NAs.