Assignment 4: Data Wrangling

Yosia Theo Napitupulu

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A03_DataExploration.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.

The completed exercise is due on Friday, Oct7th @ 5:00pm.

Set up your session

1. Check your working directory, load the `tidyverse` and `lubridate` packages, and upload 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. Explore the dimensions, column names, and structure of the datasets.

```
# 1 loading tidyverse and lubridate packages
getwd()
```

```
## [1] "E:/ENV872/EDA-Fall2022"
```

```
library(tidyverse)
library(lubridate)

# Uploading the raw data
EPAair_O3_NC2018 <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv", stringsAsFactors = TRUE)
EPAair_O3_NC2019 <- read.csv("./Data/Raw/EPAair_O3_NC2019_raw.csv", stringsAsFactors = TRUE)
EPAair_PM25_NC2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv", stringsAsFactors = TRUE)
EPAair_PM25_NC2019 <- read.csv("./Data/Raw/EPAair_PM25_NC2019_raw.csv", stringsAsFactors = TRUE)

# 2 Explore the dimensions, column names, and structure of the datasets
dim(EPAair_O3_NC2018)
```

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

```
colnames(EPAair_O3_NC2018)
```

```
##  [1] "Date"
##  [2] "Source"
##  [3] "Site.ID"
##  [4] "POC"
##  [5] "Daily.Max.8.hour.Ozone.Concentration"
##  [6] "UNITS"
##  [7] "DAILY_AQI_VALUE"
##  [8] "Site.Name"
##  [9] "DAILY_OBS_COUNT"
## [10] "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"
## [12] "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"
## [14] "CBSA_NAME"
## [15] "STATE_CODE"
## [16] "STATE"
## [17] "COUNTY_CODE"
## [18] "COUNTY"
## [19] "SITE_LATITUDE"
## [20] "SITE_LONGITUDE"
```

```
str(EPAair_O3_NC2018)
```

```
## 'data.frame':    9737 obs. of  20 variables:
##  $ Date                                : Factor w/ 364 levels "01/01/2018","01/02/2018",..: 60 61 62
##  $ Source                              : Factor w/ 1 level "AQS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Site.ID                             : int  370030005 370030005 370030005 370030005 370030005 3700
##  $ POC                                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Daily.Max.8.hour.Ozone.Concentration: num  0.043 0.046 0.047 0.049 0.047 0.03 0.036 0.044 0.049 0
##  $ UNITS                               : Factor w/ 1 level "ppm": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DAILY_AQI_VALUE                     : int  40 43 44 45 44 28 33 41 45 40 ...
##  $ Site.Name                           : Factor w/ 40 levels "","Beaufort",..: 35 35 35 35 35 35 35 3
##  $ DAILY_OBS_COUNT                     : int  17 17 17 17 17 17 17 17 17 17 ...
##  $ PERCENT_COMPLETE                    : num  100 100 100 100 100 100 100 100 100 100 ...
##  $ AQS_PARAMETER_CODE                  : int  44201 44201 44201 44201 44201 44201 44201 44201 44201 4
##  $ AQS_PARAMETER_DESC                  : Factor w/ 1 level "Ozone": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CBSA_CODE                           : int  25860 25860 25860 25860 25860 25860 25860 25860 25860 1
##  $ CBSA_NAME                           : Factor w/ 17 levels "","Asheville, NC",..: 9 9 9 9 9 9 9 9 9
##  $ STATE_CODE                          : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                               : Factor w/ 1 level "North Carolina": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COUNTY_CODE                         : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ COUNTY                              : Factor w/ 32 levels "Alexander","Avery",..: 1 1 1 1 1 1 1 1
##  $ SITE_LATITUDE                       : num  35.9 35.9 35.9 35.9 35.9 ...
##  $ SITE_LONGITUDE                      : num  -81.2 -81.2 -81.2 -81.2 -81.2 ...
```

```
head(EPAair_O3_NC2018)
```

```
##         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
```

```
dim(EPAair_O3_NC2019)
```

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

```
colnames(EPAair_O3_NC2019)
```

```
##  [1] "Date"
##  [2] "Source"
##  [3] "Site.ID"
##  [4] "POC"
##  [5] "Daily.Max.8.hour.Ozone.Concentration"
##  [6] "UNITS"
##  [7] "DAILY_AQI_VALUE"
##  [8] "Site.Name"
##  [9] "DAILY_OBS_COUNT"
## [10] "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"
## [12] "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"
## [14] "CBSA_NAME"
## [15] "STATE_CODE"
## [16] "STATE"
## [17] "COUNTY_CODE"
## [18] "COUNTY"
## [19] "SITE_LATITUDE"
## [20] "SITE_LONGITUDE"
```

```
str(EPAair_O3_NC2019)
```

```
## 'data.frame':    10592 obs. of  20 variables:
##  $ Date                                : Factor w/ 365 levels "01/01/2019","01/02/2019",..: 1 2 3 4 5
##  $ Source                              : Factor w/ 2 levels "AirNow","AQS": 1 1 1 1 1 1 1 1 1 ...
##  $ Site.ID                             : int  370030005 370030005 370030005 370030005 370030005 3700
##  $ POC                                 : int  1 1 1 1 1 1 1 1 1 ...
##  $ Daily.Max.8.hour.Ozone.Concentration: num  0.029 0.018 0.016 0.022 0.037 0.037 0.029 0.038 0.038 0
##  $ UNITS                               : Factor w/ 1 level "ppm": 1 1 1 1 1 1 1 1 1 ...
##  $ DAILY_AQI_VALUE                     : int  27 17 15 20 34 34 27 35 35 28 ...
##  $ Site.Name                           : Factor w/ 38 levels "","Beaufort",..: 33 33 33 33 33 33 33 3
##  $ DAILY_OBS_COUNT                     : int  24 24 24 24 24 24 24 24 24 24 ...
##  $ PERCENT_COMPLETE                    : num  100 100 100 100 100 100 100 100 100 100 ...
##  $ AQS_PARAMETER_CODE                  : int  44201 44201 44201 44201 44201 44201 44201 44201 44201 4
##  $ AQS_PARAMETER_DESC                  : Factor w/ 1 level "Ozone": 1 1 1 1 1 1 1 1 1 ...
##  $ CBSA_CODE                           : int  25860 25860 25860 25860 25860 25860 25860 25860 25860 1
##  $ CBSA_NAME                           : Factor w/ 15 levels "","Asheville, NC",..: 8 8 8 8 8 8 8 8
##  $ STATE_CODE                          : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                               : Factor w/ 1 level "North Carolina": 1 1 1 1 1 1 1 1 1 ...
##  $ COUNTY_CODE                         : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ COUNTY                              : Factor w/ 30 levels "Alexander","Avery",..: 1 1 1 1 1 1 1 1
##  $ SITE_LATITUDE                       : num  35.9 35.9 35.9 35.9 35.9 ...
##  $ SITE_LONGITUDE                      : num  -81.2 -81.2 -81.2 -81.2 -81.2 ...
```

```
head(EPAair_O3_NC2019)
```

```
##         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
```

```
dim(EPAair_PM25_NC2018)
```

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

```
colnames(EPAair_PM25_NC2018)
```

```
##  [1] "Date"                           "Source"
##  [3] "Site.ID"                        "POC"
##  [5] "Daily.Mean.PM2.5.Concentration" "UNITS"
##  [7] "DAILY_AQI_VALUE"                "Site.Name"
##  [9] "DAILY_OBS_COUNT"                "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"             "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"                      "CBSA_NAME"
## [15] "STATE_CODE"                     "STATE"
## [17] "COUNTY_CODE"                    "COUNTY"
## [19] "SITE_LATITUDE"                  "SITE_LONGITUDE"
```

```
str(EPAair_PM25_NC2018)
```

```
## 'data.frame':    8983 obs. of  20 variables:
##  $ Date                           : Factor w/ 365 levels "01/01/2018","01/02/2018",..: 2 5 8 11 14 17
##  $ Source                         : Factor w/ 1 level "AQS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Site.ID                        : int  370110002 370110002 370110002 370110002 370110002 370110002 3
##  $ POC                            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Daily.Mean.PM2.5.Concentration: num  2.9 3.7 5.3 0.8 2.5 4.5 1.8 2.5 4.2 1.7 ...
##  $ UNITS                          : Factor w/ 1 level "ug/m3 LC": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DAILY_AQI_VALUE                : int  12 15 22 3 10 19 8 10 18 7 ...
##  $ Site.Name                      : Factor w/ 25 levels "","Blackstone",..: 15 15 15 15 15 15 15 15 15
##  $ DAILY_OBS_COUNT                : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PERCENT_COMPLETE               : num  100 100 100 100 100 100 100 100 100 100 ...
##  $ AQS_PARAMETER_CODE             : int  88502 88502 88502 88502 88502 88502 88502 88502 88502 88502
##  $ AQS_PARAMETER_DESC             : Factor w/ 2 levels "Acceptable PM2.5 AQI & Speciation Mass",..: 1
##  $ CBSA_CODE                      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ CBSA_NAME                      : Factor w/ 14 levels "","Asheville, NC",..: 1 1 1 1 1 1 1 1 1 1 ..
##  $ STATE_CODE                     : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                          : Factor w/ 1 level "North Carolina": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COUNTY_CODE                    : int  11 11 11 11 11 11 11 11 11 11 ...
##  $ COUNTY                         : Factor w/ 21 levels "Avery","Buncombe",..: 1 1 1 1 1 1 1 1 1 1 ..
##  $ SITE_LATITUDE                  : num  36 36 36 36 36 ...
##  $ SITE_LONGITUDE                 : num  -81.9 -81.9 -81.9 -81.9 -81.9 ...
```

```
head(EPAair_PM25_NC2018)
```

```
##         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
```

```
dim(EPAair_PM25_NC2019)
```

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

```
colnames(EPAair_PM25_NC2019)
```

```
##  [1] "Date"                           "Source"
##  [3] "Site.ID"                        "POC"
##  [5] "Daily.Mean.PM2.5.Concentration" "UNITS"
##  [7] "DAILY_AQI_VALUE"                "Site.Name"
##  [9] "DAILY_OBS_COUNT"                "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"             "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"                      "CBSA_NAME"
## [15] "STATE_CODE"                     "STATE"
## [17] "COUNTY_CODE"                    "COUNTY"
## [19] "SITE_LATITUDE"                  "SITE_LONGITUDE"
```

```
str(EPAair_PM25_NC2019)
```

```
## 'data.frame':    8581 obs. of  20 variables:
##  $ Date                           : Factor w/ 365 levels "01/01/2019","01/02/2019",..: 3 6 9 12 15 18
##  $ Source                         : Factor w/ 2 levels "AirNow","AQS": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Site.ID                        : int  370110002 370110002 370110002 370110002 370110002 370110002 3
##  $ POC                            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Daily.Mean.PM2.5.Concentration: num  1.6 1 1.3 6.3 2.6 1.2 1.5 1.5 3.7 1.6 ...
##  $ UNITS                          : Factor w/ 1 level "ug/m3 LC": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DAILY_AQI_VALUE                : int  7 4 5 26 11 5 6 6 15 7 ...
##  $ Site.Name                      : Factor w/ 25 levels "","Board Of Ed. Bldg.",..: 14 14 14 14 14 14
##  $ DAILY_OBS_COUNT                : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ PERCENT_COMPLETE             : num  100 100 100 100 100 100 100 100 100 100 ...
##  $ AQS_PARAMETER_CODE           : int  88502 88502 88502 88502 88502 88502 88502 88502 88502 88502
##  $ AQS_PARAMETER_DESC           : Factor w/ 2 levels "Acceptable PM2.5 AQI & Speciation Mass",..: 1
##  $ CBSA_CODE                    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ CBSA_NAME                    : Factor w/ 14 levels "","Asheville, NC",..: 1 1 1 1 1 1 1 1 1 ..
##  $ STATE_CODE                   : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                        : Factor w/ 1 level "North Carolina": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COUNTY_CODE                  : int  11 11 11 11 11 11 11 11 11 11 ...
##  $ COUNTY                       : Factor w/ 21 levels "Avery","Buncombe",..: 1 1 1 1 1 1 1 1 1 ..
##  $ SITE_LATITUDE                : num  36 36 36 36 36 ...
##  $ SITE_LONGITUDE               : num  -81.9 -81.9 -81.9 -81.9 -81.9 ...
```

```
head(EPAair_PM25_NC2019)
```

```
##         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
```

Wrangle individual datasets to create processed files.

3. Change date to date
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 to date
EPAair_O3_NC2018$Date <- as.Date(EPAair_O3_NC2018$Date, format = "%m/%d/%Y")
EPAair_O3_NC2019$Date <- as.Date(EPAair_O3_NC2019$Date, format = "%m/%d/%Y")
EPAair_PM25_NC2018$Date <- as.Date(EPAair_PM25_NC2018$Date, format = "%m/%d/%Y")
EPAair_PM25_NC2019$Date <- as.Date(EPAair_PM25_NC2019$Date, format = "%m/%d/%Y")

# 4 Select and create processed dataset
EPAair_O3_NC2018 <- select(EPAair_O3_NC2018, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
    COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

EPAair_O3_NC2019 <- select(EPAair_O3_NC2019, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
    COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

EPAair_PM25_NC2018 <- select(EPAair_PM25_NC2018, Date, DAILY_AQI_VALUE, Site.Name,
    AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

EPAair_PM25_NC2019 <- select(EPAair_PM25_NC2019, Date, DAILY_AQI_VALUE, Site.Name,
    AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

# 5 Fill AQS_PARAMETER_DESC with PM2.5
EPAair_PM25_NC2018$AQS_PARAMETER_DESC <- "PM2.5"

EPAair_PM25_NC2019$AQS_PARAMETER_DESC <- "PM2.5"

# 6 Saving processed datasets
write.csv(EPAair_O3_NC2018, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2018.processed.csv"

write.csv(EPAair_O3_NC2019, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2019.processed.csv"

write.csv(EPAair_PM25_NC2018, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2018.processed.

write.csv(EPAair_PM25_NC2019, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2019.processed.
```

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 all 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)
- Some sites have multiple measurements per day. Use the split-apply-combine strategy to generate daily means: group by date, site, 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_NC1718_Processed.csv"

```
# 7 Combine the four datasets with `rbind`

EPAair_O3_PM25_NC1819.combine <- rbind(EPAair_O3_NC2018, EPAair_O3_NC2019, EPAair_PM25_NC2018,
    EPAair_PM25_NC2019)

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

EPAair_O3_PM25_NC1819.filter <- filter(EPAair_O3_PM25_NC1819.combine, 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_AQI = mean(DAILY_AQI_VALUE), mean_LATITUDE = mean(SITE_LATITUDE),
        mean_LONGITUDE = 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.
```

```
dim(EPAair_O3_PM25_NC1819.filter)
```

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

```
# 9 Spread your datasets such that AQI values
EPAair_O3_PM25_NC1819.filter_spread <- pivot_wider(EPAair_O3_PM25_NC1819.filter,
    names_from = AQS_PARAMETER_DESC, values_from = mean_AQI)

# 10 Call out the dimension of dataset

dim(EPAair_O3_PM25_NC1819.filter_spread)
```

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

```
# 11 Save the dataset
write.csv(EPAair_O3_PM25_NC1819.filter_spread, row.names = FALSE, file = "./Data/Processed/EPAair_O3_PM
```

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 a month and year are not available (use the function `drop_na` in your pipe).

13. Call up the dimensions of the summary dataset.

```
# 12 generating AQI value

EPAair_O3_PM25_NC1819.filter_spread_summary <- EPAair_O3_PM25_NC1819.filter_spread %>%
    group_by(Site.Name, Month, Year) %>%
    summarise(mean_AQI_Ozone = mean(Ozone), mean_AQI_PM2.5 = mean(PM2.5)) %>%
    drop_na(mean_AQI_Ozone, mean_AQI_PM2.5)
```

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

```
# 13 Call up the dimensions of the summary dataset.
dim(EPAair_O3_PM25_NC1819.filter_spread_summary)
```

```
## [1] 101   5
```

14. Why did we use the function `drop_na` rather than `na.omit`?

    Answer: We use drop_na() because it is more simple and more comprehensively processing the n/a, while the na.omit() function can't inspect a subset of all columns. Drop_na() drops rows where any column contains a missing value.It also keeps the "complete" rows (where no rows contain missing values).