Assignment 4: Data Wrangling

Molly Bruce

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_A04_DataWrangling.Rmd") prior to submission.

The completed exercise is due on Tuesday, Feb 16 @ 11:59pm.

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. 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
getwd()
```

```
## [1] "C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Assignments"
```

```
# Commented out for purposes of knitting the file
#setwd("C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021")

# Commented out for purposes of knitting the file
#install.packages(tidyverse)
#install.packages(lubridate)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3

## Warning: package 'ggplot2' was built under R version 4.0.3

## Warning: package 'tibble' was built under R version 4.0.3

## Warning: package 'tidyr' was built under R version 4.0.3

## Warning: package 'readr' was built under R version 4.0.3

## Warning: package 'purrr' was built under R version 4.0.3

## Warning: package 'dplyr' was built under R version 4.0.3

## Warning: package 'stringr' was built under R version 4.0.3
```

```
## Warning: package 'forcats' was built under R version 4.0.3
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.3
```

```
#The ideal format for calling these datafiles is a relative path using the format
#EPAair_O3_NC_2019 <- read.csv("./Data/Raw/EPAair_O3_NC2019_raw.csv", stringsAsFactors = TRUE)
#however, this method has thrown errors for me on Assignment 3 and also on this assignment.
#Therefore, I coded the enture path even though this is less resilient and less prefered.
EPAair_O3_NC_2019 <-
  read.csv("C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/Raw/EPAair_O3_NC2019_raw.csv")
EPAair_O3_NC_2018 <-
  read.csv("C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/Raw/EPAair_O3_NC2018_raw.csv")
EPAair_PM25_NC_2019 <-
  read.csv("C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/Raw/EPAair_PM25_NC2019_raw.csv")
EPAair_PM25_NC_2018 <-
  read.csv("C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/Raw/EPAair_PM25_NC2018_raw.csv")

#2
#Exploring the Ozone 2019 dataset
colnames(EPAair_O3_NC_2019)
```

```
##  [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"
```

```
head(EPAair_O3_NC_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
```

```
summary(EPAair_O3_NC_2019)
```

```
##      Date              Source             Site.ID               POC
##  Length:10592       Length:10592       Min.   :370030005   Min.   :1
##  Class :character   Class :character   1st Qu.:370630015   1st Qu.:1
##  Mode  :character   Mode  :character   Median :370870036   Median :1
##                                        Mean   :370960317   Mean   :1
##                                        3rd Qu.:371290002   3rd Qu.:1
##                                        Max.   :371990004   Max.   :1
##
##  Daily.Max.8.hour.Ozone.Concentration    UNITS           DAILY_AQI_VALUE
##  Min.   :0.00000                      Length:10592       Min.   :  0.0
##  1st Qu.:0.03600                      Class :character   1st Qu.: 33.0
##  Median :0.04400                      Mode  :character   Median : 41.0
##  Mean   :0.04331                                         Mean   : 41.2
##  3rd Qu.:0.05000                                         3rd Qu.: 46.0
##  Max.   :0.08100                                         Max.   :136.0
##
##   Site.Name         DAILY_OBS_COUNT PERCENT_COMPLETE AQS_PARAMETER_CODE
##  Length:10592       Min.   :13.00   Min.   : 75.00   Min.   :44201
##  Class :character   1st Qu.:17.00   1st Qu.:100.00   1st Qu.:44201
##  Mode  :character   Median :17.00   Median :100.00   Median :44201
##                     Mean   :18.34   Mean   : 99.69   Mean   :44201
##                     3rd Qu.:17.00   3rd Qu.:100.00   3rd Qu.:44201
##                     Max.   :24.00   Max.   :100.00   Max.   :44201
##
##  AQS_PARAMETER_DESC   CBSA_CODE      CBSA_NAME           STATE_CODE
##  Length:10592       Min.   :11700   Length:10592       Min.   :37
##  Class :character   1st Qu.:16740   Class :character   1st Qu.:37
##  Mode  :character   Median :24660   Mode  :character   Median :37
##                     Mean   :26617                      Mean   :37
##                     3rd Qu.:37080                      3rd Qu.:37
##                     Max.   :49180                      Max.   :37
##                     NA's   :2852
##     STATE            COUNTY_CODE       COUNTY           SITE_LATITUDE
##  Length:10592       Min.   :  3.0   Length:10592       Min.   :34.36
```

```
##  Class :character   1st Qu.: 63.0   Class :character   1st Qu.:35.26
##  Mode  :character   Median : 87.0   Mode  :character   Median :35.59
##                     Mean   : 95.9                      Mean   :35.61
##                     3rd Qu.:129.0                      3rd Qu.:36.03
##                     Max.   :199.0                      Max.   :36.31
##
##  SITE_LONGITUDE
##  Min.   :-83.80
##  1st Qu.:-82.05
##  Median :-80.34
##  Mean   :-80.41
##  3rd Qu.:-78.77
##  Max.   :-76.62
##
```

```
str(EPAair_O3_NC_2019)
```

```
## 'data.frame':    10592 obs. of  20 variables:
##  $ Date                                : chr  "01/01/2019" "01/02/2019" "01/03/2019" "01/04/2019" ..
##  $ Source                              : chr  "AirNow" "AirNow" "AirNow" "AirNow" ...
##  $ 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
##  $ UNITS                               : chr  "ppm" "ppm" "ppm" "ppm" ...
##  $ DAILY_AQI_VALUE                     : int  27 17 15 20 34 34 27 35 35 28 ...
##  $ Site.Name                           : chr  "Taylorsville Liledoun" "Taylorsville Liledoun" "Taylo
##  $ 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
##  $ AQS_PARAMETER_DESC                  : chr  "Ozone" "Ozone" "Ozone" "Ozone" ...
##  $ CBSA_CODE                           : int  25860 25860 25860 25860 25860 25860 25860 25860 25860
##  $ CBSA_NAME                           : chr  "Hickory-Lenoir-Morganton, NC" "Hickory-Lenoir-Morgant
##  $ STATE_CODE                          : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                               : chr  "North Carolina" "North Carolina" "North Carolina" "No
##  $ COUNTY_CODE                         : int  3 3 3 3 3 3 3 3 3 ...
##  $ COUNTY                              : chr  "Alexander" "Alexander" "Alexander" "Alexander" ...
##  $ 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 ...
```

```
dim(EPAair_O3_NC_2019)
```

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

```
#Exploring the Ozone 2018 dataset
colnames(EPAair_O3_NC_2018)
```

```
##  [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"
```

```
head(EPAair_O3_NC_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
```

```
summary(EPAair_O3_NC_2018)
```

```
##      Date              Source             Site.ID               POC
##  Length:9737        Length:9737        Min.   :370030005   Min.   :1
##  Class :character   Class :character   1st Qu.:370650099   1st Qu.:1
##  Mode  :character   Mode  :character   Median :371010002   Median :1
##                                        Mean   :370969118   Mean   :1
##                                        3rd Qu.:371290002   3rd Qu.:1
##                                        Max.   :371990004   Max.   :1
##
##  Daily.Max.8.hour.Ozone.Concentration    UNITS           DAILY_AQI_VALUE
##  Min.   :0.00200                     Length:9737        Min.   :  2.00
##  1st Qu.:0.03400                     Class :character   1st Qu.: 31.00
```

```
##  Median :0.04200                     Mode  :character   Median : 39.00
##  Mean   :0.04194                                        Mean   : 40.22
##  3rd Qu.:0.04900                                        3rd Qu.: 45.00
##  Max.   :0.07700                                        Max.   :122.00
##
##   Site.Name         DAILY_OBS_COUNT PERCENT_COMPLETE AQS_PARAMETER_CODE
##  Length:9737        Min.   :12.00   Min.   : 71.00   Min.   :44201
##  Class :character   1st Qu.:17.00   1st Qu.:100.00   1st Qu.:44201
##  Mode  :character   Median :17.00   Median :100.00   Median :44201
##                     Mean   :16.94   Mean   : 99.65   Mean   :44201
##                     3rd Qu.:17.00   3rd Qu.:100.00   3rd Qu.:44201
##                     Max.   :17.00   Max.   :100.00   Max.   :44201
##
##  AQS_PARAMETER_DESC   CBSA_CODE      CBSA_NAME           STATE_CODE
##  Length:9737        Min.   :11700   Length:9737        Min.   :37
##  Class :character   1st Qu.:16740   Class :character   1st Qu.:37
##  Mode  :character   Median :24660   Mode  :character   Median :37
##                     Mean   :27247                      Mean   :37
##                     3rd Qu.:39580                      3rd Qu.:37
##                     Max.   :49180                      Max.   :37
##                     NA's   :2609
##     STATE            COUNTY_CODE        COUNTY           SITE_LATITUDE
##  Length:9737        Min.   :  3.00   Length:9737        Min.   :34.36
##  Class :character   1st Qu.: 65.00   Class :character   1st Qu.:35.26
##  Mode  :character   Median :101.00   Mode  :character   Median :35.55
##                     Mean   : 96.78                      Mean   :35.62
##                     3rd Qu.:129.00                      3rd Qu.:36.03
##                     Max.   :199.00                      Max.   :36.31
##
##  SITE_LONGITUDE
##  Min.   :-83.80
##  1st Qu.:-82.05
##  Median :-80.34
##  Mean   :-80.42
##  3rd Qu.:-78.90
##  Max.   :-76.62
##
```

```
str(EPAair_O3_NC_2018)
```

```
## 'data.frame':    9737 obs. of  20 variables:
##  $ Date                                : chr  "03/01/2018" "03/02/2018" "03/03/2018" "03/04/2018" ..
##  $ Source                              : chr  "AQS" "AQS" "AQS" "AQS" ...
##  $ 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                               : chr  "ppm" "ppm" "ppm" "ppm" ...
##  $ DAILY_AQI_VALUE                     : int  40 43 44 45 44 28 33 41 45 40 ...
##  $ Site.Name                           : chr  "Taylorsville Liledoun" "Taylorsville Liledoun" "Taylo
##  $ 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                  : chr  "Ozone" "Ozone" "Ozone" "Ozone" ...
##  $ CBSA_CODE                           : int  25860 25860 25860 25860 25860 25860 25860 25860 25860 2
##  $ CBSA_NAME                           : chr  "Hickory-Lenoir-Morganton, NC" "Hickory-Lenoir-Morgant
```

```
##  $ STATE_CODE                          : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                               : chr  "North Carolina" "North Carolina" "North Carolina" "No
##  $ COUNTY_CODE                         : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ COUNTY                              : chr  "Alexander" "Alexander" "Alexander" "Alexander" ...
##  $ 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 ...
```

```
dim(EPAair_O3_NC_2018)
```

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

```
#Exploring the PM25 2019 dataset
colnames(EPAair_PM25_NC_2019)
```

```
##  [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"
```

```
head(EPAair_PM25_NC_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
```

```
summary(EPAair_PM25_NC_2019)
```

```
##      Date              Source             Site.ID               POC
##  Length:8581        Length:8581        Min.   :370110002   Min.   :1.000
##  Class :character   Class :character   1st Qu.:370630015   1st Qu.:3.000
##  Mode  :character   Mode  :character   Median :371190041   Median :3.000
##                                        Mean   :371023743   Mean   :3.032
##                                        3rd Qu.:371290002   3rd Qu.:3.000
##                                        Max.   :371830021   Max.   :5.000
##
##  Daily.Mean.PM2.5.Concentration    UNITS           DAILY_AQI_VALUE
##  Min.   :-3.100                 Length:8581        Min.   : 0.00
##  1st Qu.: 4.900                 Class :character   1st Qu.:20.00
##  Median : 7.400                 Mode  :character   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 AQS_PARAMETER_CODE
##  Length:8581        Min.   :1       Min.   :100      Min.   :88101
##  Class :character   1st Qu.:1       1st Qu.:100      1st Qu.:88101
##  Mode  :character   Median :1       Median :100      Median :88101
##                     Mean   :1       Mean   :100      Mean   :88149
##                     3rd Qu.:1       3rd Qu.:100      3rd Qu.:88101
##                     Max.   :1       Max.   :100      Max.   :88502
##
##  AQS_PARAMETER_DESC   CBSA_CODE      CBSA_NAME           STATE_CODE
##  Length:8581        Min.   :11700   Length:8581        Min.   :37
##  Class :character   1st Qu.:19000   Class :character   1st Qu.:37
##  Mode  :character   Median :25860   Mode  :character   Median :37
##                     Mean   :31099                      Mean   :37
##                     3rd Qu.:40580                      3rd Qu.:37
##                     Max.   :49180                      Max.   :37
##                     NA's   :1058
##     STATE            COUNTY_CODE       COUNTY           SITE_LATITUDE
##  Length:8581        Min.   : 11.0   Length:8581        Min.   :34.36
##  Class :character   1st Qu.: 63.0   Class :character   1st Qu.:35.26
##  Mode  :character   Median :119.0   Mode  :character   Median :35.73
##                     Mean   :102.4                      Mean   :35.63
##                     3rd Qu.:129.0                      3rd Qu.:35.91
##                     Max.   :183.0                      Max.   :36.51
##
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-79.95
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

```
str(EPAair_PM25_NC_2019)
```

```
## 'data.frame':    8581 obs. of  20 variables:
```

```
##  $ Date                          : chr  "01/03/2019" "01/06/2019" "01/09/2019" "01/12/2019" ...
##  $ Source                        : chr  "AQS" "AQS" "AQS" "AQS" ...
##  $ Site.ID                       : int  370110002 370110002 370110002 370110002 370110002 370110002 3
##  $ POC                           : int  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                         : chr  "ug/m3 LC" "ug/m3 LC" "ug/m3 LC" "ug/m3 LC" ...
##  $ DAILY_AQI_VALUE               : int  7 4 5 26 11 5 6 6 15 7 ...
##  $ Site.Name                     : chr  "Linville Falls" "Linville Falls" "Linville Falls" "Linville
##  $ DAILY_OBS_COUNT               : int  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            : chr  "Acceptable PM2.5 AQI & Speciation Mass" "Acceptable PM2.5 A
##  $ CBSA_CODE                     : int  NA NA NA NA NA NA NA NA NA ...
##  $ CBSA_NAME                     : chr  "" "" "" "" ...
##  $ STATE_CODE                    : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                         : chr  "North Carolina" "North Carolina" "North Carolina" "North Ca
##  $ COUNTY_CODE                   : int  11 11 11 11 11 11 11 11 11 11 ...
##  $ COUNTY                        : chr  "Avery" "Avery" "Avery" "Avery" ...
##  $ SITE_LATITUDE                 : num  36 36 36 36 36 ...
##  $ SITE_LONGITUDE                : num  -81.9 -81.9 -81.9 -81.9 -81.9 ...
```

```
dim(EPAair_PM25_NC_2019)
```

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

```
#Exploring the PM25 2018 dataset
colnames(EPAair_PM25_NC_2018)
```

```
##  [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"
```

```
head(EPAair_PM25_NC_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
```

```
summary(EPAair_PM25_NC_2018)
```

```
##      Date              Source             Site.ID               POC
##  Length:8983        Length:8983        Min.   :370110002   Min.   :1.000
##  Class :character   Class :character   1st Qu.:370630015   1st Qu.:3.000
##  Mode  :character   Mode  :character   Median :371010002   Median :3.000
##                                        Mean   :371002405   Mean   :2.812
##                                        3rd Qu.:371230001   3rd Qu.:3.000
##                                        Max.   :371830021   Max.   :5.000
##
##  Daily.Mean.PM2.5.Concentration    UNITS           DAILY_AQI_VALUE
##  Min.   :-2.300                 Length:8983        Min.   : 0.00
##  1st Qu.: 4.900                 Class :character   1st Qu.:20.00
##  Median : 7.000                 Mode  :character   Median :29.00
##  Mean   : 7.491                                    Mean   :30.73
##  3rd Qu.: 9.700                                    3rd Qu.:40.00
##  Max.   :34.200                                    Max.   :97.00
##
##   Site.Name         DAILY_OBS_COUNT PERCENT_COMPLETE AQS_PARAMETER_CODE
##  Length:8983        Min.   :1       Min.   :100      Min.   :88101
##  Class :character   1st Qu.:1       1st Qu.:100      1st Qu.:88101
##  Mode  :character   Median :1       Median :100      Median :88101
##                     Mean   :1       Mean   :100      Mean   :88164
##                     3rd Qu.:1       3rd Qu.:100      3rd Qu.:88101
##                     Max.   :1       Max.   :100      Max.   :88502
##
##  AQS_PARAMETER_DESC   CBSA_CODE      CBSA_NAME           STATE_CODE
##  Length:8983        Min.   :11700   Length:8983        Min.   :37
##  Class :character   1st Qu.:19000   Class :character   1st Qu.:37
##  Mode  :character   Median :25860   Mode  :character   Median :37
##                     Mean   :30946                      Mean   :37
##                     3rd Qu.:40580                      3rd Qu.:37
##                     Max.   :49180                      Max.   :37
##                     NA's   :1263
##     STATE            COUNTY_CODE       COUNTY           SITE_LATITUDE
##  Length:8983        Min.   : 11.0   Length:8983        Min.   :34.36
##  Class :character   1st Qu.: 63.0   Class :character   1st Qu.:35.26
##  Mode  :character   Median :101.0   Mode  :character   Median :35.64
##                     Mean   :100.2                      Mean   :35.61
##                     3rd Qu.:123.0                      3rd Qu.:35.91
##                     Max.   :183.0                      Max.   :36.11
```

```
##
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-79.99
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

```
str(EPAair_PM25_NC_2018)
```

```
## 'data.frame':    8983 obs. of  20 variables:
##  $ Date                          : chr  "01/02/2018" "01/05/2018" "01/08/2018" "01/11/2018" ...
##  $ Source                        : chr  "AQS" "AQS" "AQS" "AQS" ...
##  $ 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                         : chr  "ug/m3 LC" "ug/m3 LC" "ug/m3 LC" "ug/m3 LC" ...
##  $ DAILY_AQI_VALUE               : int  12 15 22 3 10 19 8 10 18 7 ...
##  $ Site.Name                     : chr  "Linville Falls" "Linville Falls" "Linville Falls" "Linville
##  $ 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            : chr  "Acceptable PM2.5 AQI & Speciation Mass" "Acceptable PM2.5 A
##  $ CBSA_CODE                     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ CBSA_NAME                     : chr  "" "" "" "" ...
##  $ STATE_CODE                    : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                         : chr  "North Carolina" "North Carolina" "North Carolina" "North Ca
##  $ COUNTY_CODE                   : int  11 11 11 11 11 11 11 11 11 11 ...
##  $ COUNTY                        : chr  "Avery" "Avery" "Avery" "Avery" ...
##  $ SITE_LATITUDE                 : num  36 36 36 36 36 ...
##  $ SITE_LONGITUDE                : num  -81.9 -81.9 -81.9 -81.9 -81.9 ...
```

```
dim(EPAair_PM25_NC_2018)
```

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

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 I am assuming that by saying "Change date to date,"
#you want us to parse the string turned factor into a traditional date object
EPAair_O3_NC_2019$Date <- as.Date(EPAair_O3_NC_2019$Date, format = "%m/%d/%Y")
EPAair_O3_NC_2018$Date <- as.Date(EPAair_O3_NC_2018$Date, format = "%m/%d/%Y")
EPAair_PM25_NC_2019$Date <- as.Date(EPAair_PM25_NC_2019$Date, format = "%m/%d/%Y")
EPAair_PM25_NC_2018$Date <- as.Date(EPAair_PM25_NC_2018$Date, format = "%m/%d/%Y")

#4
```

```
EPAair_O3_NC_2019_selected <-
  select(EPAair_O3_NC_2019, Date, DAILY_AQI_VALUE, Site.Name,
         AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
EPAair_O3_NC_2018_selected <-
  select(EPAair_O3_NC_2018, Date, DAILY_AQI_VALUE, Site.Name,
         AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
EPAair_PM25_NC_2019_selected <-
  select(EPAair_PM25_NC_2019, Date, DAILY_AQI_VALUE, Site.Name,
         AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
EPAair_PM25_NC_2018_selected <-
  select(EPAair_PM25_NC_2018, Date, DAILY_AQI_VALUE, Site.Name,
         AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

#5
EPAair_PM25_NC_2019_selected$AQS_PARAMETER_DESC <- "PM2.5"
EPAair_PM25_NC_2018_selected$AQS_PARAMETER_DESC <- "PM2.5"

#6
write.csv(EPAair_O3_NC_2019_selected,"C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/Pro
write.csv(EPAair_O3_NC_2018_selected,"C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/Pro
write.csv(EPAair_PM25_NC_2019_selected,"C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/P
write.csv(EPAair_PM25_NC_2018_selected,"C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/P
```

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
# I am combining the four datasets with identical column names
EPAair_NC <- rbind(EPAair_O3_NC_2019_selected,
                   EPAair_O3_NC_2018_selected,
                   EPAair_PM25_NC_2019_selected,
                   EPAair_PM25_NC_2018_selected)

#8
# I am assuming that you want us to use the intersect function to find the sites
# the four data frames have in common, as opposed to filtering based on the list
# you provided, though either method would work.
```

```
EPAair_NC_piped <-
  EPAair_NC %>%
  # Finding the sites the four dataframes have in common using intersect
  filter(Site.Name %in% c(intersect(
    intersect(EPAair_O3_NC_2019_selected$Site.Name,
              EPAair_O3_NC_2018_selected$Site.Name),
    intersect(EPAair_PM25_NC_2019_selected$Site.Name,
              EPAair_PM25_NC_2018_selected$Site.Name))) & Site.Name != "") %>%
  # Split-Apply-Combine
  group_by(Date, Site.Name, AQS_PARAMETER_DESC, COUNTY) %>%
  summarise(mean.AQI = mean(DAILY_AQI_VALUE),
            mean.lat = mean(SITE_LATITUDE),
            mean.long = mean(SITE_LONGITUDE), .groups = "keep") %>%
  # Adding the Month and Year columns
  mutate(Month = month(Date), Year = year(Date))

# Throwing this into a csv because I worked really hard and was worried I'd mess it up
write.csv(EPAair_NC_piped, row.names = FALSE,
          file ="C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/Processed/EPAair_NC_piped

#9
# Pivoting the dataframe
EPAair_NC_piped_spread <-
  EPAair_NC_piped %>%
  pivot_wider(id_cols = c(Date, Month, Year, Site.Name, COUNTY, mean.lat, mean.long),
                names_from = AQS_PARAMETER_DESC, values_from = mean.AQI)

#10
dim(EPAair_NC_piped_spread)
```

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

```
#11
write.csv(EPAair_NC_piped_spread, row.names = FALSE,
          file ="C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/Processed/EPAair_O3_PM25_
```

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.

```
#12a
# Split-Apply-Combine in a pipe again
EPAair_NC_piped_processed <-
  EPAair_NC_piped_spread %>%
  group_by(Site.Name, Month, Year) %>%
  summarise(mean.AQI.PM2.5 = mean(PM2.5),
            mean.AQI.Ozone = mean(Ozone))
```

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

```r
#12b
# Using a separate pipe for this drop_na approach, though I could have done it in the 12a pipe
EPAair_NC_piped_processed_clean <-
  EPAair_NC_piped_processed %>%
  drop_na(c(Month, Year))

#13
dim(EPAair_NC_piped_processed_clean)
```

```
## [1] 308   5
```

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

    Answer: na.omit is general; it removes rows with NA values in ANY column, even if you try to specify a particular column. Meanwhile, drop_na is targeted; it allows users to specify the column where it should search and then drops rows where the NAs are present in that specified column. In our case, running the na.omit function returns a dataframe with dimensions 101 by 5, meaning that, as predicted, it removed a lot more rows because it also tosses rows with NAs in the Mean.AQI.PM2.5 and Mean.AQI.Ozone columns.