Assignment 4: Data Wrangling

Reed Leon-Hinton

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 Setting up the working directory and loading the required packages
getwd()
```

```
## [1] "C:/Users/shado/Documents/Graduate School Stuff/ENVIRON 872 - Environmental Data Analytics/Envir
```

```
# clearing the environment (It's a pet peeve)
remove(list = ls())

# install.packages("tidyverse")
library(tidyverse)
# install.packages("lubridate")
library(lubridate)

# importing the datasets
o3_2018 <- read.csv(file = "./Data/Raw/EPAair_O3_NC2018_raw.csv",
                    stringsAsFactors = TRUE)
o3_2019 <- read.csv(file = "./Data/Raw/EPAair_O3_NC2019_raw.csv",
                    stringsAsFactors = TRUE)
pm25_2018 <- read.csv(file = "./Data/Raw/EPAair_PM25_NC2018_raw.csv",
                      stringsAsFactors = TRUE)
pm25_2019 <- read.csv(file = "./Data/Raw/EPAair_PM25_NC2019_raw.csv",
                      stringsAsFactors = TRUE)

#2 looking at the details of the datasets
```

```r
# looking at the O3 2018 dataset
dim_o3_2018 <- dim(o3_2018)
colnames(o3_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"
```

```r
summary(o3_2018)
```

```
##          Date      Source        Site.ID               POC
##  04/01/2018:  40   AQS:9737   Min.   :370030005   Min.   :1
##  04/12/2018:  40              1st Qu.:370650099   1st Qu.:1
##  04/13/2018:  40              Median :371010002   Median :1
##  04/14/2018:  40              Mean   :370969118   Mean   :1
##  04/15/2018:  40              3rd Qu.:371290002   3rd Qu.:1
##  04/18/2018:  40              Max.   :371990004   Max.   :1
##  (Other)   :9497
##  Daily.Max.8.hour.Ozone.Concentration UNITS      DAILY_AQI_VALUE
##  Min.   :0.00200                      ppm:9737   Min.   :  2.00
##  1st Qu.:0.03400                                 1st Qu.: 31.00
##  Median :0.04200                                 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
##  Coweeta             : 355   Min.   :12.00   Min.   : 71.00
##  Garinger High School: 354   1st Qu.:17.00   1st Qu.:100.00
##  Millbrook School    : 352   Median :17.00   Median :100.00
##  Candor              : 335   Mean   :16.94   Mean   : 99.65
##  Rockwell            : 335   3rd Qu.:17.00   3rd Qu.:100.00
##  Cranberry           : 323   Max.   :17.00   Max.   :100.00
##  (Other)             :7683
##  AQS_PARAMETER_CODE AQS_PARAMETER_DESC   CBSA_CODE
##  Min.   :44201      Ozone:9737          Min.   :11700
##  1st Qu.:44201                          1st Qu.:16740
##  Median :44201                           Median :24660
##  Mean   :44201                           Mean   :27247
##  3rd Qu.:44201                           3rd Qu.:39580
##  Max.   :44201                           Max.   :49180
##                                          NA's   :2609
##                                CBSA_NAME      STATE_CODE             STATE
##                                     :2609   Min.   :37   North Carolina:9737
##  Charlotte-Concord-Gastonia, NC-SC:1338   1st Qu.:37
##  Asheville, NC                    : 927   Median :37
##  Winston-Salem, NC                : 725   Mean   :37
##  Raleigh, NC                      : 585   3rd Qu.:37
##  Hickory-Lenoir-Morganton, NC     : 477   Max.   :37
##  (Other)                          :3076
##   COUNTY_CODE             COUNTY     SITE_LATITUDE   SITE_LONGITUDE
##  Min.   :  3.00   Forsyth    : 725   Min.   :34.36   Min.   :-83.80
##  1st Qu.: 65.00   Haywood    : 683   1st Qu.:35.26   1st Qu.:-82.05
##  Median :101.00   Mecklenburg: 592   Median :35.55   Median :-80.34
##  Mean   : 96.78   Avery      : 558   Mean   :35.62   Mean   :-80.42
##  3rd Qu.:129.00   Swain      : 483   3rd Qu.:36.03   3rd Qu.:-78.90
##  Max.   :199.00   Cumberland : 444   Max.   :36.31   Max.   :-76.62
##                   (Other)    :6252
```

```
lapply(o3_2018, class)
```

```
## $Date
## [1] "factor"
##
## $Source
## [1] "factor"
##
## $Site.ID
## [1] "integer"
##
## $POC
## [1] "integer"
##
## $Daily.Max.8.hour.Ozone.Concentration
## [1] "numeric"
##
## $UNITS
## [1] "factor"
##
## $DAILY_AQI_VALUE
## [1] "integer"
##
## $Site.Name
## [1] "factor"
##
## $DAILY_OBS_COUNT
## [1] "integer"
##
## $PERCENT_COMPLETE
## [1] "numeric"
##
## $AQS_PARAMETER_CODE
## [1] "integer"
##
## $AQS_PARAMETER_DESC
## [1] "factor"
##
## $CBSA_CODE
## [1] "integer"
##
## $CBSA_NAME
## [1] "factor"
##
## $STATE_CODE
## [1] "integer"
##
## $STATE
## [1] "factor"
##
## $COUNTY_CODE
## [1] "integer"
##
## $COUNTY
## [1] "factor"
##
## $SITE_LATITUDE
## [1] "numeric"
##
## $SITE_LONGITUDE
## [1] "numeric"
```

```
# looking at the O3 2019 dataset
dim_o3_2019 <- dim(o3_2019)
colnames(o3_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"
```

```
summary(o3_2019)
```

```
##          Date         Source       Site.ID             POC
##  03/18/2019:   38   AirNow:2126   Min.   :370030005   Min.   :1
##  03/19/2019:   38   AQS   :8466   1st Qu.:370630015   1st Qu.:1
##  03/20/2019:   38                 Median :370870036   Median :1
##  03/23/2019:   38                 Mean   :370960317   Mean   :1
##  03/24/2019:   38                 3rd Qu.:371290002   3rd Qu.:1
##  03/25/2019:   38                 Max.   :371990004   Max.   :1
##  (Other)   :10364
##  Daily.Max.8.hour.Ozone.Concentration UNITS       DAILY_AQI_VALUE
##  Min.   :0.00000                      ppm:10592   Min.   :  0.0
##  1st Qu.:0.03600                                  1st Qu.: 33.0
##  Median :0.04400                                  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
##  Garinger High School: 363   Min.   :13.00   Min.   : 75.00
##  Millbrook School    : 362   1st Qu.:17.00   1st Qu.:100.00
##  Coweeta             : 361   Median :17.00   Median :100.00
##  Rockwell            : 361   Mean   :18.34   Mean   : 99.69
##  Candor              : 358   3rd Qu.:17.00   3rd Qu.:100.00
##  Cranberry           : 351   Max.   :24.00   Max.   :100.00
##  (Other)             :8436
##  AQS_PARAMETER_CODE AQS_PARAMETER_DESC   CBSA_CODE
##  Min.   :44201      Ozone:10592        Min.   :11700
##  1st Qu.:44201                         1st Qu.:16740
##  Median :44201                         Median :24660
##  Mean   :44201                         Mean   :26617
##  3rd Qu.:44201                         3rd Qu.:37080
##  Max.   :44201                         Max.   :49180
##                                        NA's   :2852
##                              CBSA_NAME      STATE_CODE            STATE
##                                   :2852   Min.   :37   North Carolina:10592
##  Charlotte-Concord-Gastonia, NC-SC:1590   1st Qu.:37
##  Asheville, NC                    :1114   Median :37
##  Winston-Salem, NC                : 735   Mean   :37
##  Raleigh, NC                      : 646   3rd Qu.:37
##  Hickory-Lenoir-Morganton, NC     : 567   Max.   :37
##  (Other)                          :3088
##   COUNTY_CODE            COUNTY     SITE_LATITUDE   SITE_LONGITUDE
##  Min.   :  3.0   Haywood    : 864   Min.   :34.36   Min.   :-83.80
##  1st Qu.: 63.0   Forsyth    : 735   1st Qu.:35.26   1st Qu.:-82.05
##  Median : 87.0   Mecklenburg: 657   Median :35.59   Median :-80.34
##  Mean   : 95.9   Avery      : 607   Mean   :35.61   Mean   :-80.41
##  3rd Qu.:129.0   Cumberland : 498   3rd Qu.:36.03   3rd Qu.:-78.77
##  Max.   :199.0   Swain      : 476   Max.   :36.31   Max.   :-76.62
##                  (Other)    :6755
```

```
lapply(o3_2019, class)
```

```
## $Date
## [1] "factor"
##
## $Source
## [1] "factor"
##
## $Site.ID
## [1] "integer"
##
## $POC
## [1] "integer"
##
## $Daily.Max.8.hour.Ozone.Concentration
## [1] "numeric"
##
## $UNITS
## [1] "factor"
##
## $DAILY_AQI_VALUE
## [1] "integer"
##
## $Site.Name
## [1] "factor"
##
## $DAILY_OBS_COUNT
## [1] "integer"
##
## $PERCENT_COMPLETE
## [1] "numeric"
##
## $AQS_PARAMETER_CODE
## [1] "integer"
##
## $AQS_PARAMETER_DESC
## [1] "factor"
##
## $CBSA_CODE
## [1] "integer"
##
## $CBSA_NAME
## [1] "factor"
##
## $STATE_CODE
## [1] "integer"
##
## $STATE
## [1] "factor"
##
## $COUNTY_CODE
## [1] "integer"
##
## $COUNTY
## [1] "factor"
##
## $SITE_LATITUDE
## [1] "numeric"
##
## $SITE_LONGITUDE
## [1] "numeric"
```

```
# looking at the PM 2.5 2018 dataset
dim_pm25_2018 <- dim(pm25_2018)
colnames(pm25_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"
```

```
summary(pm25_2018)
```

```
##          Date      Source        Site.ID               POC
##  01/26/2018:  40   AQS:8983   Min.   :370110002   Min.   :1.000
##  02/01/2018:  40              1st Qu.:370630015   1st Qu.:3.000
##  02/19/2018:  40              Median :371010002   Median :3.000
##  03/21/2018:  40              Mean   :371002405   Mean   :2.812
##  04/02/2018:  40              3rd Qu.:371230001   3rd Qu.:3.000
##  04/08/2018:  40              Max.   :371830021   Max.   :5.000
##  (Other)   :8743
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-2.300                 ug/m3 LC:8983   Min.   : 0.00
##  1st Qu.: 4.900                                 1st Qu.:20.00
##  Median : 7.000                                 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
##  Millbrook School    : 717   Min.   :1       Min.   :100
##  Hattie Avenue       : 510   1st Qu.:1       1st Qu.:100
##  Board Of Ed. Bldg.  : 477   Median :1       Median :100
##  Garinger High School: 472   Mean   :1       Mean   :100
##  Durham Armory       : 466   3rd Qu.:1       3rd Qu.:100
##  Pitt Agri. Center   : 460   Max.   :1       Max.   :100
##  (Other)             :5881
##  AQS_PARAMETER_CODE                              AQS_PARAMETER_DESC
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:1403
##  1st Qu.:88101      PM2.5 - Local Conditions              :7580
##  Median :88101
##  Mean   :88164
##  3rd Qu.:88101
##  Max.   :88502
##
##    CBSA_CODE                                  CBSA_NAME      STATE_CODE
##  Min.   :11700   Raleigh, NC                      :1396   Min.   :37  
##  1st Qu.:19000   Winston-Salem, NC                :1316   1st Qu.:37  
##  Median :25860   Charlotte-Concord-Gastonia, NC-SC:1275   Median :37  
##  Mean   :30946                                    :1263   Mean   :37  
##  3rd Qu.:40580   Asheville, NC                    : 586   3rd Qu.:37  
##  Max.   :49180   Durham-Chapel Hill, NC           : 466   Max.   :37  
##  NA's   :1263    (Other)                          :2681               
##             STATE       COUNTY_CODE            COUNTY     SITE_LATITUDE  
##  North Carolina:8983   Min.   : 11.0   Mecklenburg:1275   Min.   :34.36  
##                        1st Qu.: 63.0   Wake       :1049   1st Qu.:35.26  
##                        Median :101.0   Forsyth    : 876   Median :35.64  
##                        Mean   :100.2   Buncombe   : 477   Mean   :35.61  
##                        3rd Qu.:123.0   Durham     : 466   3rd Qu.:35.91  
##                        Max.   :183.0   Pitt       : 460   Max.   :36.11  
##                                        (Other)    :4380                  
##  SITE_LONGITUDE  
##  Min.   :-83.44  
##  1st Qu.:-80.87  
##  Median :-80.23  
##  Mean   :-79.99  
##  3rd Qu.:-78.57  
##  Max.   :-76.21  
## 
```

```
lapply(pm25_2018, class)
```

```
## $Date
## [1] "factor"
## 
## $Source
## [1] "factor"
## 
## $Site.ID
## [1] "integer"
## 
## $POC
## [1] "integer"
## 
## $Daily.Mean.PM2.5.Concentration
## [1] "numeric"
## 
## $UNITS
## [1] "factor"
## 
## $DAILY_AQI_VALUE
## [1] "integer"
## 
## $Site.Name
## [1] "factor"
## 
## $DAILY_OBS_COUNT
## [1] "integer"
## 
## $PERCENT_COMPLETE
## [1] "numeric"
```

```
##
## $AQS_PARAMETER_CODE
## [1] "integer"
##
## $AQS_PARAMETER_DESC
## [1] "factor"
##
## $CBSA_CODE
## [1] "integer"
##
## $CBSA_NAME
## [1] "factor"
##
## $STATE_CODE
## [1] "integer"
##
## $STATE
## [1] "factor"
##
## $COUNTY_CODE
## [1] "integer"
##
## $COUNTY
## [1] "factor"
##
## $SITE_LATITUDE
## [1] "numeric"
##
## $SITE_LONGITUDE
## [1] "numeric"
```

```
# looking at the PM 2.5 2019 dataset
dim_pm25_2019 <- dim(pm25_2019)
colnames(pm25_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"
```

```
summary(pm25_2019)
```

```
##           Date         Source        Site.ID               POC
##   02/26/2019:  41   AirNow:1670   Min.   :370110002   Min.   :1.000
##   01/21/2019:  40   AQS   :6911   1st Qu.:370630015   1st Qu.:3.000
##   02/14/2019:  40                 Median :371190041   Median :3.000
##   01/09/2019:  39                 Mean   :371023743   Mean   :3.032
##   01/27/2019:  39                 3rd Qu.:371290002   3rd Qu.:3.000
##   02/02/2019:  39                 Max.   :371830021   Max.   :5.000
##  (Other)   :8343
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-3.100                 ug/m3 LC:8581   Min.   : 0.00
##  1st Qu.: 4.900                                 1st Qu.:20.00
##  Median : 7.400                                 Median :31.00
##  Mean   : 7.684                                 Mean   :31.51
##  3rd Qu.:10.100                                 3rd Qu.:42.00
##  Max.   :31.200                                 Max.   :91.00
##
##                 Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Millbrook School    : 738   Min.   :1       Min.   :100
##  Garinger High School: 629   1st Qu.:1       1st Qu.:100
##  Remount             : 573   Median :1       Median :100
##  Hickory Water Tower : 518   Mean   :1       Mean   :100
##  Hattie Avenue       : 436   3rd Qu.:1       3rd Qu.:100
##  Durham Armory       : 431   Max.   :1       Max.   :100
##  (Other)             :5256
##  AQS_PARAMETER_CODE                              AQS_PARAMETER_DESC
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:1029
##  1st Qu.:88101      PM2.5 - Local Conditions              :7552
##  Median :88101
##  Mean   :88149
##  3rd Qu.:88101
##  Max.   :88502
##
##    CBSA_CODE                                 CBSA_NAME      STATE_CODE
##  Min.   :11700   Raleigh, NC                      :1441   Min.   :37
##  1st Qu.:19000   Charlotte-Concord-Gastonia, NC-SC:1379   1st Qu.:37
##  Median :25860   Winston-Salem, NC                :1235   Median :37
##  Mean   :31099                                    :1058   Mean   :37
##  3rd Qu.:40580   Hickory-Lenoir-Morganton, NC     : 518   3rd Qu.:37
##  Max.   :49180   Durham-Chapel Hill, NC           : 431   Max.   :37
##  NA's   :1058    (Other)                          :2519
##             STATE       COUNTY_CODE            COUNTY     SITE_LATITUDE
##  North Carolina:8581   Min.   : 11.0   Mecklenburg:1379   Min.   :34.36
##                        1st Qu.: 63.0   Wake       :1083   1st Qu.:35.26
##                        Median :119.0   Forsyth    : 839   Median :35.73
##                        Mean   :102.4   Catawba    : 518   Mean   :35.63
##                        3rd Qu.:129.0   Durham     : 431   3rd Qu.:35.91
##                        Max.   :183.0   Cumberland : 427   Max.   :36.51
##                                        (Other)    :3904
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-79.95
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

```
lapply(pm25_2019, class)
```

```
## $Date
## [1] "factor"
##
```

```
## $Source
## [1] "factor"
##
## $Site.ID
## [1] "integer"
##
## $POC
## [1] "integer"
##
## $Daily.Mean.PM2.5.Concentration
## [1] "numeric"
##
## $UNITS
## [1] "factor"
##
## $DAILY_AQI_VALUE
## [1] "integer"
##
## $Site.Name
## [1] "factor"
##
## $DAILY_OBS_COUNT
## [1] "integer"
##
## $PERCENT_COMPLETE
## [1] "numeric"
##
## $AQS_PARAMETER_CODE
## [1] "integer"
##
## $AQS_PARAMETER_DESC
## [1] "factor"
##
## $CBSA_CODE
## [1] "integer"
##
## $CBSA_NAME
## [1] "factor"
##
## $STATE_CODE
## [1] "integer"
##
## $STATE
## [1] "factor"
##
## $COUNTY_CODE
## [1] "integer"
##
## $COUNTY
## [1] "factor"
##
## $SITE_LATITUDE
## [1] "numeric"
##
```

```
## $SITE_LONGITUDE
## [1] "numeric"
```

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 changing the format of the date column in each dataset
o3_2018$Date <- mdy(o3_2018$Date)
o3_2019$Date <- mdy(o3_2019$Date)
pm25_2018$Date <- mdy(pm25_2018$Date)
pm25_2019$Date <- mdy(pm25_2019$Date)

#4 creating processed datasets with the selected columns only
o3_2018_p <- select(.data = o3_2018,
                    Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
                    COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
o3_2019_p <- select(.data = o3_2019,
                    Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
                    COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
pm25_2018_p <- select(.data = pm25_2018,
                      Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
                      COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
pm25_2019_p <- select(.data = pm25_2019,
                      Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
                      COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

#5 setting the value of the AQS_PARAMETER_DESC column to PM2.5
pm25_2018_p$AQS_PARAMETER_DESC <- as.factor("PM2.5")
pm25_2019_p$AQS_PARAMETER_DESC <- as.factor("PM2.5")

#6 creating the processed datasets in the processed folder
write.csv(o3_2018_p, row.names = FALSE,
          file = "./Data/Processed/EPAair_O3_NC2018_processed.csv")
write.csv(o3_2019_p, row.names = FALSE,
          file = "./Data/Processed/EPAair_O3_NC2019_processed.csv")
write.csv(pm25_2018_p, row.names = FALSE,
          file = "./Data/Processed/EPAair_PM25_NC2018_processed.csv")
write.csv(pm25_2019_p, row.names = FALSE,
          file = "./Data/Processed/EPAair_PM25_NC2019_processed.csv")
```

Combine datasets

7. Combine the four datasets with `rbind`. Make sure your column names are identical prior to running this code.
8. Wrangle your new dataset with a pipe function (%>%) so that it fills the following conditions:

- Include 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 combining the datasets with the rbind function
o3_pm25_p <- rbind(o3_2018_p, o3_2019_p, pm25_2018_p, pm25_2019_p)

#8 completing the pipe function to meet the conditions specified
o3_pm25_p <- o3_pm25_p %>%
  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_at(vars(DAILY_AQI_VALUE, SITE_LATITUDE, SITE_LONGITUDE), list(name = mean)) %>%
  mutate(Month = month(Date, label = TRUE, abbr = TRUE)) %>%
  mutate(Year = year(Date)) %>%
  ungroup() %>%
  droplevels()
dim(o3_pm25_p) #just checking to see if I got the 14,752 x 9.
```

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

```
#9 spreading the dataset
o3_pm25_p <- pivot_wider(o3_pm25_p, names_from = AQS_PARAMETER_DESC,
                         values_from = DAILY_AQI_VALUE_name)

colnames(o3_pm25_p)[4] <- "SITE_LATITUDE" # fixing the unusual naming resulting from the
                                          # process where I meaned the values

colnames(o3_pm25_p)[5] <- "SITE_LONGITUDE" # doing the same as above

#10 looking at the dimensions of the dataset
final_dimensions <- dim(o3_pm25_p)
print(paste("The final dataset has", final_dimensions[1], "recorded values for",
            final_dimensions[2], "variables."))
```

```
## [1] "The final dataset has 8976 recorded values for 9 variables."
```

```
#11
write.csv(o3_pm25_p, row.names = FALSE,
          file = "./Data/Processed/EPAair_O3_PM25_NC1718_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 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
o3_pm25_p_summaries <-
  o3_pm25_p %>%
  group_by(Site.Name, Month, Year) %>%
  summarise_at(vars(PM2.5, Ozone), list(name = mean)) %>%
  drop_na(Month, Year) %>%
  ungroup()

# fixing the naming convention like I did previously in the assignment
colnames(o3_pm25_p)[4] <- "PM2.5"
colnames(o3_pm25_p)[5] <- "Ozone"

#13
summary_dimensions <- dim(o3_pm25_p_summaries)
print(paste("The summary dataset has", summary_dimensions[1], "recorded values for",
            summary_dimensions[2], "variables."))
```

```
## [1] "The summary dataset has 308 recorded values for 5 variables."
```

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

    Answer: na.omit drops all values which have NA associated with every row, whereas drop_na only drops rows with NA values from the columns specified. We get many 207 more observations when we use drop_na rather than na.omit. drop_na works better in this situation because we want to keep the NA values for Ozone and PM2.5, but not for the month and year.