Assignment 4: Data Wrangling

Curtis Cha

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 Monday, Feb 7 @ 7:00pm.

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. 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.

library(tidyverse)
library(lubridate)

setwd("C:/Users/curtx/Desktop/Environnmental Data Analytic/Environmental_Data_Analytics_2022/Data/Raw/")

O3_18 <- read.csv("EPAair_O3_NC2018_raw.csv", header = T)

O3_19 <- read.csv("EPAair_O3_NC2019_raw.csv", header = T)

PM25_18 <- read.csv("EPAair_PM25_NC2018_raw.csv", header = T)

PM25_19 <- read.csv("EPAair_PM25_NC2019_raw.csv", header = T)

#2.

dim(O3_18)
```

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

```
summary(O3_18)
```

```
##      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  
## 
```

```
dim(O3_19)
```

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

```
summary(O3_19)
```

```
##      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
##
```

```
dim(PM25_18)
```

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

```
summary(PM25_18)
```

```
##      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
##
```

```
dim(PM25_19)
```

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

```
summary(PM25_19)
```

```
##      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  
## 
```

Wrangle individual datasets to create processed files.

3. Change date to a date object
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.

O3_18$Date <- as.Date(O3_18$Date, format = "%m/%d/%y")
O3_19$Date <- as.Date(O3_19$Date, format = "%m/%d/%y")
PM25_18$Date <- as.Date(PM25_18$Date, format = "%m/%d/%y")
PM25_19$Date <- as.Date(PM25_19$Date, format = "%m/%d/%y")

#4.

O3_18_sub <- O3_18 %>% select("Date", "DAILY_AQI_VALUE", "Site.Name", "AQS_PARAMETER_DESC", "COUNTY", "S
O3_19_sub <- O3_19 %>% select("Date", "DAILY_AQI_VALUE", "Site.Name", "AQS_PARAMETER_DESC", "COUNTY", "S

PM25_18_sub <- PM25_18 %>% select("Date", "DAILY_AQI_VALUE", "Site.Name", "AQS_PARAMETER_DESC", "COUNTY
PM25_19_sub <- PM25_19 %>% select("Date", "DAILY_AQI_VALUE", "Site.Name", "AQS_PARAMETER_DESC", "COUNTY

#5.

PM25_18_sub$AQS_PARAMETER_DESC <- "PM2.5"
PM25_19_sub$AQS_PARAMETER_DESC <- "PM2.5"

#6.
setwd("C:/Users/curtx/Desktop/Environnmental Data Analytic/Environmental_Data_Analytics_2022/Data/Proces

write.csv(O3_18_sub, row.names = FALSE, file = "EPAair_O3_NC2018_processed.csv")

write.csv(O3_19_sub, row.names = FALSE, file = "EPAair_O3_NC2019_processed.csv")

write.csv(PM25_18_sub, row.names = FALSE, file = "EPAair_PM25_NC2019_processed.csv")

write.csv(PM25_19_sub, row.names = FALSE, file = "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:

- Filter records to include just the 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 `intersect` function can figure out common factor levels if we didn’t give you this list...)
- 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_NC2122_Processed.csv"

```
#7

colnames(O3_18_sub) == colnames(O3_19_sub)
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
colnames(PM25_18_sub) == colnames(PM25_19_sub)
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
colnames(O3_18_sub) == colnames(PM25_19_sub)
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
O3_PM25_18_19 <- rbind(O3_18_sub, O3_19_sub, PM25_18_sub, PM25_19_sub)
nrow(O3_PM25_18_19) #37893 rows
```

```
## [1] 37893
```

```
#8

sites <- c("Linville Falls", "Durham Armory", "Leggett", "Hattie Avenue", "Clemmons Middle", "Mendenhal

O3_PM25_18_19_sum <- O3_PM25_18_19 %>%
  filter(O3_PM25_18_19$Site.Name %in% sites) %>% #16510 rows
  group_by(Date, Site.Name, AQS_PARAMETER_DESC, COUNTY) %>%
  dplyr::summarise(mean_aqi = mean(DAILY_AQI_VALUE),
            mean_lat = mean(SITE_LATITUDE),
            mean_long = mean(SITE_LONGITUDE)) %>% #7899 rows for some reason
  mutate(month = month(Date), year = year(Date))
```

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

```
dim(O3_PM25_18_19_sum)
```

```
## [1] 7899    9
```

```
#9

O3_PM25_18_19_spread <- pivot_wider(O3_PM25_18_19_sum, names_from = AQS_PARAMETER_DESC, values_from = me

#10

dim(O3_PM25_18_19_spread)
```

```
## [1] 4637    9
```

```
#11
write.csv(O3_PM25_18_19_spread, row.names = FALSE, file = "EPAair_O3_PM25_NC2122_Processed.csv")
```

Generate summary tables

12a. Use the split-apply-combine strategy to generate a summary data frame from your results from Step 9 above. Data should be grouped by site, month, and year. Generate the mean AQI values for ozone and PM2.5 for each group.

12b. BONUS: Add a piped statement to 12a that removes rows where both mean ozone and mean PM2.5 have missing values.

13. Call up the dimensions of the summary dataset.

```
#12(a,b)

O3_PM25_18_19_spread_sum <- O3_PM25_18_19_spread %>%
  group_by(Site.Name, month, year) %>%
  filter(!is.na(Ozone) & !is.na(PM2.5)) %>%
  dplyr::summarise("Mean Ozone" = mean(Ozone), "Mean PM2.5" = mean(PM2.5))
```

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

```
#13

dim(O3_PM25_18_19_spread_sum)
```

```
## [1] 127   5
```

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

    Answer: drop_na() is a function of the tidyr package, while na.omit is a built-in R function. Drop_na can remove rows with na's in one column of a dataframe, but na.omit will remove rows with na's in at least one column. Since we are only concerned with specific rows, na.omit would remove rows with important data.