Assignment 4: Data Wrangling

Sara Sayed

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] "/Users/sarasayed/Documents/Data Analytics/Environmental_Data_Analytics_2021"
```

```
library(tidyverse)
library(lubridate)


O3_2018 <- read_csv("./Data/Raw/EPAair_O3_NC2018_raw.csv")
O3_2019 <- read_csv("./Data/Raw/EPAair_O3_NC2019_raw.csv")
PM2.5_2018 <- read_csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")
PM2.5_2019 <-read_csv("./Data/Raw/EPAair_PM25_NC2019_raw.csv")
#2
dim(O3_2018)
```

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

```
dim(O3_2019)
```

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

```
dim(PM2.5_2018)
```

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

```
dim(PM2.5_2019)
```

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

```
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"
```

```
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"
```

```
colnames(PM2.5_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"
```

```
colnames(PM2.5_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"
```

```
str(O3_2018, width=80, strict.width="cut")
```

```
## tibble [9,737 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Date                                : chr [1:9737] "03/01/2018" "03/02/201"..
##  $ Source                              : chr [1:9737] "AQS" "AQS" "AQS" "AQS" ..
##  $ Site ID                             : num [1:9737] 3.7e+08 3.7e+08 3.7e+08 ..
##  $ POC                                 : num [1:9737] 1 1 1 1 1 1 1 1 1 ...
##  $ Daily Max 8-hour Ozone Concentration: num [1:9737] 0.043 0.046 0.047 0.049 ..
##  $ UNITS                               : chr [1:9737] "ppm" "ppm" "ppm" "ppm" ..
##  $ DAILY_AQI_VALUE                     : num [1:9737] 40 43 44 45 44 28 33 41 ..
##  $ Site Name                           : chr [1:9737] "Taylorsville Liledoun""..
##  $ DAILY_OBS_COUNT                     : num [1:9737] 17 17 17 17 17 17 17 17 ..
##  $ PERCENT_COMPLETE                    : num [1:9737] 100 100 100 100 100 100 ..
##  $ AQS_PARAMETER_CODE                  : num [1:9737] 44201 44201 44201 44201 ..
##  $ AQS_PARAMETER_DESC                  : chr [1:9737] "Ozone" "Ozone" "Ozone""..
##  $ CBSA_CODE                           : num [1:9737] 25860 25860 25860 25860 ..
##  $ CBSA_NAME                           : chr [1:9737] "Hickory-Lenoir-Morgant"..
##  $ STATE_CODE                          : num [1:9737] 37 37 37 37 37 37 37 37 ..
##  $ STATE                               : chr [1:9737] "North Carolina" "North"..
##  $ COUNTY_CODE                         : chr [1:9737] "003" "003" "003" "003" ..
##  $ COUNTY                              : chr [1:9737] "Alexander" "Alexander""..
##  $ SITE_LATITUDE                       : num [1:9737] 35.9 35.9 35.9 35.9 35.9..
##  $ SITE_LONGITUDE                      : num [1:9737] -81.2 -81.2 -81.2 -81.2 ..
##  - attr(*, "spec")=
##   .. cols(
##   ..   Date = col_character(),
##   ..   Source = col_character(),
##   ..   `Site ID` = col_double(),
##   ..   POC = col_double(),
##   ..   `Daily Max 8-hour Ozone Concentration` = col_double(),
##   ..   UNITS = col_character(),
##   ..   DAILY_AQI_VALUE = col_double(),
##   ..   `Site Name` = col_character(),
##   ..   DAILY_OBS_COUNT = col_double(),
##   ..   PERCENT_COMPLETE = col_double(),
```

```
##   ..   AQS_PARAMETER_CODE = col_double(),
##   ..   AQS_PARAMETER_DESC = col_character(),
##   ..   CBSA_CODE = col_double(),
##   ..   CBSA_NAME = col_character(),
##   ..   STATE_CODE = col_double(),
##   ..   STATE = col_character(),
##   ..   COUNTY_CODE = col_character(),
##   ..   COUNTY = col_character(),
##   ..   SITE_LATITUDE = col_double(),
##   ..   SITE_LONGITUDE = col_double()
##   .. )
```

```
str(O3_2019, width=80, strict.width="cut" )
```

```
## tibble [10,592 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Date                                : chr [1:10592] "01/01/2019" "01/02/20"..
##  $ Source                              : chr [1:10592] "AirNow" "AirNow" "Air"..
##  $ Site ID                             : num [1:10592] 3.7e+08 3.7e+08 3.7e+08..
##  $ POC                                 : num [1:10592] 1 1 1 1 1 1 1 1 1 ...
##  $ Daily Max 8-hour Ozone Concentration: num [1:10592] 0.029 0.018 0.016 0.022..
##  $ UNITS                               : chr [1:10592] "ppm" "ppm" "ppm" "ppm"..
##  $ DAILY_AQI_VALUE                     : num [1:10592] 27 17 15 20 34 34 27 35..
##  $ Site Name                           : chr [1:10592] "Taylorsville Liledoun"..
##  $ DAILY_OBS_COUNT                     : num [1:10592] 24 24 24 24 24 24 24 24..
##  $ PERCENT_COMPLETE                    : num [1:10592] 100 100 100 100 100 100..
##  $ AQS_PARAMETER_CODE                  : num [1:10592] 44201 44201 44201 44201..
##  $ AQS_PARAMETER_DESC                  : chr [1:10592] "Ozone" "Ozone" "Ozone"..
##  $ CBSA_CODE                           : num [1:10592] 25860 25860 25860 25860..
##  $ CBSA_NAME                           : chr [1:10592] "Hickory-Lenoir-Morgan"..
##  $ STATE_CODE                          : num [1:10592] 37 37 37 37 37 37 37 37..
##  $ STATE                               : chr [1:10592] "North Carolina" "Nort"..
##  $ COUNTY_CODE                         : chr [1:10592] "003" "003" "003" "003"..
##  $ COUNTY                              : chr [1:10592] "Alexander" "Alexander"..
##  $ SITE_LATITUDE                       : num [1:10592] 35.9 35.9 35.9 35.9 35...
##  $ SITE_LONGITUDE                      : num [1:10592] -81.2 -81.2 -81.2 -81.2..
##  - attr(*, "spec")=
##   .. cols(
##   ..   Date = col_character(),
##   ..   Source = col_character(),
##   ..   `Site ID` = col_double(),
##   ..   POC = col_double(),
##   ..   `Daily Max 8-hour Ozone Concentration` = col_double(),
##   ..   UNITS = col_character(),
##   ..   DAILY_AQI_VALUE = col_double(),
##   ..   `Site Name` = col_character(),
##   ..   DAILY_OBS_COUNT = col_double(),
##   ..   PERCENT_COMPLETE = col_double(),
##   ..   AQS_PARAMETER_CODE = col_double(),
##   ..   AQS_PARAMETER_DESC = col_character(),
##   ..   CBSA_CODE = col_double(),
##   ..   CBSA_NAME = col_character(),
##   ..   STATE_CODE = col_double(),
##   ..   STATE = col_character(),
##   ..   COUNTY_CODE = col_character(),
##   ..   COUNTY = col_character(),
```

```
##   ..   SITE_LATITUDE = col_double(),
##   ..   SITE_LONGITUDE = col_double()
##   .. )
```

```
str(PM2.5_2018, width=80, strict.width="cut")
```

```
## tibble [8,983 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Date                           : chr [1:8983] "01/02/2018" "01/05/2018" "01"..
##  $ Source                         : chr [1:8983] "AQS" "AQS" "AQS" "AQS" ...
##  $ Site ID                        : num [1:8983] 3.7e+08 3.7e+08 3.7e+08 3.7e+0..
##  $ POC                            : num [1:8983] 1 1 1 1 1 1 1 1 1 ...
##  $ Daily Mean PM2.5 Concentration: num [1:8983] 2.9 3.7 5.3 0.8 2.5 4.5 1.8 2...
##  $ UNITS                          : chr [1:8983] "ug/m3 LC" "ug/m3 LC" "ug/m3 "..
##  $ DAILY_AQI_VALUE                : num [1:8983] 12 15 22 3 10 19 8 10 18 7 ...
##  $ Site Name                      : chr [1:8983] "Linville Falls" "Linville Fa"..
##  $ DAILY_OBS_COUNT                : num [1:8983] 1 1 1 1 1 1 1 1 1 1 ...
##  $ PERCENT_COMPLETE               : num [1:8983] 100 100 100 100 100 100 100 10..
##  $ AQS_PARAMETER_CODE             : num [1:8983] 88502 88502 88502 88502 88502 ..
##  $ AQS_PARAMETER_DESC             : chr [1:8983] "Acceptable PM2.5 AQI & Speci"..
##  $ CBSA_CODE                      : num [1:8983] NA NA NA NA NA NA NA NA NA NA ..
##  $ CBSA_NAME                      : chr [1:8983] NA NA NA NA ...
##  $ STATE_CODE                     : num [1:8983] 37 37 37 37 37 37 37 37 37 37 ..
##  $ STATE                          : chr [1:8983] "North Carolina" "North Carol"..
##  $ COUNTY_CODE                    : chr [1:8983] "011" "011" "011" "011" ...
##  $ COUNTY                         : chr [1:8983] "Avery" "Avery" "Avery" "Aver"..
##  $ SITE_LATITUDE                  : num [1:8983] 36 36 36 36 36 ...
##  $ SITE_LONGITUDE                 : num [1:8983] -81.9 -81.9 -81.9 -81.9 -81.9 ..
##  - attr(*, "spec")=
##   .. cols(
##   ..   Date = col_character(),
##   ..   Source = col_character(),
##   ..   `Site ID` = col_double(),
##   ..   POC = col_double(),
##   ..   `Daily Mean PM2.5 Concentration` = col_double(),
##   ..   UNITS = col_character(),
##   ..   DAILY_AQI_VALUE = col_double(),
##   ..   `Site Name` = col_character(),
##   ..   DAILY_OBS_COUNT = col_double(),
##   ..   PERCENT_COMPLETE = col_double(),
##   ..   AQS_PARAMETER_CODE = col_double(),
##   ..   AQS_PARAMETER_DESC = col_character(),
##   ..   CBSA_CODE = col_double(),
##   ..   CBSA_NAME = col_character(),
##   ..   STATE_CODE = col_double(),
##   ..   STATE = col_character(),
##   ..   COUNTY_CODE = col_character(),
##   ..   COUNTY = col_character(),
##   ..   SITE_LATITUDE = col_double(),
##   ..   SITE_LONGITUDE = col_double()
##   .. )
```

```
str(PM2.5_2019, width=80, strict.width="cut")
```

```
## tibble [8,581 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Date                           : chr [1:8581] "01/03/2019" "01/06/2019" "01"..
```

```
##  $ Source                        : chr [1:8581] "AQS" "AQS" "AQS" "AQS" ...
##  $ Site ID                       : num [1:8581] 3.7e+08 3.7e+08 3.7e+08 3.7e+0..
##  $ POC                           : num [1:8581] 1 1 1 1 1 1 1 1 1 ...
##  $ Daily Mean PM2.5 Concentration: num [1:8581] 1.6 1 1.3 6.3 2.6 1.2 1.5 1.5 ..
##  $ UNITS                         : chr [1:8581] "ug/m3 LC" "ug/m3 LC" "ug/m3 "..
##  $ DAILY_AQI_VALUE               : num [1:8581] 7 4 5 26 11 5 6 6 15 7 ...
##  $ Site Name                     : chr [1:8581] "Linville Falls" "Linville Fa"..
##  $ DAILY_OBS_COUNT               : num [1:8581] 1 1 1 1 1 1 1 1 1 ...
##  $ PERCENT_COMPLETE              : num [1:8581] 100 100 100 100 100 100 100 10..
##  $ AQS_PARAMETER_CODE            : num [1:8581] 88502 88502 88502 88502 88502 ..
##  $ AQS_PARAMETER_DESC            : chr [1:8581] "Acceptable PM2.5 AQI & Speci"..
##  $ CBSA_CODE                     : num [1:8581] NA NA NA NA NA NA NA NA NA NA ..
##  $ CBSA_NAME                     : chr [1:8581] NA NA NA NA ...
##  $ STATE_CODE                    : num [1:8581] 37 37 37 37 37 37 37 37 37 37 ..
##  $ STATE                         : chr [1:8581] "North Carolina" "North Carol"..
##  $ COUNTY_CODE                   : chr [1:8581] "011" "011" "011" "011" ...
##  $ COUNTY                        : chr [1:8581] "Avery" "Avery" "Avery" "Aver"..
##  $ SITE_LATITUDE                 : num [1:8581] 36 36 36 36 36 ...
##  $ SITE_LONGITUDE                : num [1:8581] -81.9 -81.9 -81.9 -81.9 -81.9 ..
##  - attr(*, "spec")=
##   .. cols(
##   ..   Date = col_character(),
##   ..   Source = col_character(),
##   ..   `Site ID` = col_double(),
##   ..   POC = col_double(),
##   ..   `Daily Mean PM2.5 Concentration` = col_double(),
##   ..   UNITS = col_character(),
##   ..   DAILY_AQI_VALUE = col_double(),
##   ..   `Site Name` = col_character(),
##   ..   DAILY_OBS_COUNT = col_double(),
##   ..   PERCENT_COMPLETE = col_double(),
##   ..   AQS_PARAMETER_CODE = col_double(),
##   ..   AQS_PARAMETER_DESC = col_character(),
##   ..   CBSA_CODE = col_double(),
##   ..   CBSA_NAME = col_character(),
##   ..   STATE_CODE = col_double(),
##   ..   STATE = col_character(),
##   ..   COUNTY_CODE = col_character(),
##   ..   COUNTY = col_character(),
##   ..   SITE_LATITUDE = col_double(),
##   ..   SITE_LONGITUDE = col_double()
##   .. )
```

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
O3_2018$Date <- as.Date(O3_2018$Date, format = "%m/%d/%Y")
O3_2019$Date <- as.Date(O3_2019$Date, format = "%m/%d/%Y")
PM2.5_2018$Date <- as.Date(PM2.5_2018$Date, format = "%m/%d/%Y")
PM2.5_2019$Date <- as.Date(PM2.5_2019$Date, format = "%m/%d/%Y")

#4
O3_2018.aqi.value <- select(O3_2018,Date, DAILY_AQI_VALUE,`Site Name`,
                            AQS_PARAMETER_DESC, COUNTY,SITE_LATITUDE,
                            SITE_LONGITUDE)
O3_2019.aqi.value <- select(O3_2019,Date, DAILY_AQI_VALUE,`Site Name`,
                            AQS_PARAMETER_DESC, COUNTY,
                            SITE_LATITUDE, SITE_LONGITUDE)
PM2.5_2018.aqi.value <- select(PM2.5_2018,Date, DAILY_AQI_VALUE,`Site Name`,
                               AQS_PARAMETER_DESC, COUNTY,SITE_LATITUDE,
                               SITE_LONGITUDE)
PM2.5_2019.aqi.value <- select(PM2.5_2019,Date, DAILY_AQI_VALUE,`Site Name`,
                               AQS_PARAMETER_DESC, COUNTY,SITE_LATITUDE,
                               SITE_LONGITUDE)
#5

PM2.5_2018.aqi.value.mutate <- mutate(PM2.5_2018.aqi.value, AQS_PARAMETER_DESC = "PM2.5")

PM2.5_2019.aqi.value.mutate <- mutate(PM2.5_2019.aqi.value, AQS_PARAMETER_DESC = "PM2.5")

#6
write.csv(O3_2018, row.names = FALSE,
          file = "./Data/Processed/EPAair_O3_NC2018_Processed.csv")
write.csv(O3_2019, row.names = FALSE,
          file = "./Data/Processed/EPAair_O3_NC2019_Processed.csv")
write.csv(PM2.5_2018, row.names = FALSE,
          file = "./Data/Processed/EPAair_PM25_NC2018_Processed.csv")
write.csv(PM2.5_2019, 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

AQI_Values <- rbind(O3_2018.aqi.value, O3_2019.aqi.value, PM2.5_2018.aqi.value.mutate
                    , PM2.5_2019.aqi.value.mutate)
dim(AQI_Values)
```

```
## [1] 37893     7
```

```
#8

AQI_Values.processed <-
  AQI_Values %>%
  filter(`Site Name` == "Linville Falls"|`Site Name` == "Durham Armory"
         |`Site Name` == "Leggett"|`Site Name` == "Hattie Avenue"
         |`Site Name` =="Clemmons Middle" |`Site Name` == "Mendenhall School"
         |`Site Name` == "Frying Pan Mountain"|`Site Name` == "West Johnston Co."
         |`Site Name` == "Garinger High School" |`Site Name` =="Castle Hayne"
         |`Site Name` == "Pitt Agri. Center" |`Site Name` =="Bryson City"
         |`Site Name` == "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(year = year(Date))%>%
  mutate(month = month(Date))

dim(AQI_Values.processed)
```

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

```
#9

AQI_Values.processed.spread <- pivot_wider(AQI_Values.processed,
                                           names_from = AQS_PARAMETER_DESC,
                                           values_from = mean_AQI)


#10

dim(AQI_Values.processed.spread)
```

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

```
#11

write.csv(AQI_Values.processed.spread, 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.

```r
#12a


AQI_Values.processed.summary <-
  AQI_Values.processed.spread %>%
  group_by(`Site Name`, month, year) %>%
  summarise(mean_Ozone = mean(Ozone),
            mean_PM2.5 = mean(PM2.5))


#12b

AQI_Values.processed.summary2 <-
  AQI_Values.processed.summary %>%
  drop_na(month, year)

#13

dim(AQI_Values.processed.summary)
```

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

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

    Answer: `drop_na` removes NA values within the selection you select while `na.omit` removes all the NA's within the data frame. We selected `drop_na` because we looking to remove NA's from specific rows rather than the whole dataset.