Assignment 4: Data Wrangling

Jack Eynon

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., "Salk_A04_DataWrangling.Rmd") prior to submission.

The completed exercise is due on Tuesday, February 4 at 1:00 pm.

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/jackeynon/Courses/EnvDataAnalytics/Environmental_Data_Analytics_2020/Assignments"
```

```
library(tidyverse)
library(lubridate)
EPAair_O3_NC2018 <- read_csv("~/Courses/EnvDataAnalytics/Environmental_Data_Analytics_2020/Data/Raw/EPAa
EPAair_O3_NC2019 <- read_csv("~/Courses/EnvDataAnalytics/Environmental_Data_Analytics_2020/Data/Raw/EPAa
EPAair_PM25_NC2018 <- read_csv("~/Courses/EnvDataAnalytics/Environmental_Data_Analytics_2020/Data/Raw/E
EPAair_PM25_NC2019 <- read_csv("~/Courses/EnvDataAnalytics/Environmental_Data_Analytics_2020/Data/Raw/E
#2
## Dimensions
dim(EPAair_O3_NC2018) # 9737 observations with 20 columns
```

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

```
dim(EPAair_O3_NC2019) # 10592 observations with 20 columns
```

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

```
dim(EPAair_PM25_NC2018) # 8983 observations with 20 columns
```

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

```r
dim(EPAair_PM25_NC2019) # 8581 observations with 20 columns
```

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

```r
## Exploring column names
colnames(EPAair_03_NC2018)
```

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

```r
colnames(EPAair_O3_NC2019)
```

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

```r
colnames(EPAair_PM25_NC2018)
```

```
##  [1] "Date"                           "Source"
##  [3] "Site ID"                        "POC"
##  [5] "Daily Mean PM2.5 Concentration" "UNITS"
```

```
##  [7] "DAILY_AQI_VALUE"                "Site Name"
##  [9] "DAILY_OBS_COUNT"                "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"             "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"                      "CBSA_NAME"
## [15] "STATE_CODE"                     "STATE"
## [17] "COUNTY_CODE"                    "COUNTY"
## [19] "SITE_LATITUDE"                  "SITE_LONGITUDE"
```

```
colnames(EPAair_PM25_NC2019)
```

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

```
## Exploring structure
str(EPAair_03_NC2018)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and '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                             : num  3.7e+08 3.7e+08 3.7e+08 3.7e+08 3.7e+08 ...
##  $ POC                                 : num  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                     : num  40 43 44 45 44 28 33 41 45 40 ...
##  $ Site Name                           : chr  "Taylorsville Liledoun" "Taylorsville Liledoun" "Taylo
##  $ DAILY_OBS_COUNT                     : num  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                  : num  44201 44201 44201 44201 44201 ...
##  $ AQS_PARAMETER_DESC                  : chr  "Ozone" "Ozone" "Ozone" "Ozone" ...
##  $ CBSA_CODE                           : num  25860 25860 25860 25860 25860 ...
##  $ CBSA_NAME                           : chr  "Hickory-Lenoir-Morganton, NC" "Hickory-Lenoir-Morgant
##  $ STATE_CODE                          : num  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                               : chr  "North Carolina" "North Carolina" "North Carolina" "No
##  $ COUNTY_CODE                         : chr  "003" "003" "003" "003" ...
##  $ 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 ...
##  - 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(EPAair_O3_NC2019)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and '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                             : num  3.7e+08 3.7e+08 3.7e+08 3.7e+08 3.7e+08 ...
##  $ POC                                 : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Daily Max 8-hour Ozone Concentration: num  0.029 0.018 0.016 0.022 0.037 0.037 0.029 0.038 0.038 0
##  $ UNITS                               : chr  "ppm" "ppm" "ppm" "ppm" ...
##  $ DAILY_AQI_VALUE                     : num  27 17 15 20 34 34 27 35 35 28 ...
##  $ Site Name                           : chr  "Taylorsville Liledoun" "Taylorsville Liledoun" "Taylo
##  $ DAILY_OBS_COUNT                     : num  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                  : num  44201 44201 44201 44201 44201 ...
##  $ AQS_PARAMETER_DESC                  : chr  "Ozone" "Ozone" "Ozone" "Ozone" ...
##  $ CBSA_CODE                           : num  25860 25860 25860 25860 25860 ...
##  $ CBSA_NAME                           : chr  "Hickory-Lenoir-Morganton, NC" "Hickory-Lenoir-Morganto
##  $ STATE_CODE                          : num  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                               : chr  "North Carolina" "North Carolina" "North Carolina" "No
##  $ COUNTY_CODE                         : chr  "003" "003" "003" "003" ...
##  $ 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 ...
##  - 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(EPAair_PM25_NC2018)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and '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                        : num  3.7e+08 3.7e+08 3.7e+08 3.7e+08 3.7e+08 ...
##  $ POC                            : num  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                : num  12 15 22 3 10 19 8 10 18 7 ...
##  $ Site Name                      : chr  "Linville Falls" "Linville Falls" "Linville Falls" "Linville
##  $ DAILY_OBS_COUNT                : num  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             : num  88502 88502 88502 88502 88502 ...
##  $ AQS_PARAMETER_DESC             : chr  "Acceptable PM2.5 AQI & Speciation Mass" "Acceptable PM2.5 A
##  $ CBSA_CODE                      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ CBSA_NAME                      : chr  NA NA NA NA ...
##  $ STATE_CODE                     : num  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                          : chr  "North Carolina" "North Carolina" "North Carolina" "North Ca
##  $ COUNTY_CODE                    : chr  "011" "011" "011" "011" ...
##  $ 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 ...
##  - 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(EPAair_PM25_NC2019)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and '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                       : num  3.7e+08 3.7e+08 3.7e+08 3.7e+08 3.7e+08 ...
##  $ POC                           : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Daily Mean PM2.5 Concentration: num  1.6 1 1.3 6.3 2.6 1.2 1.5 1.5 3.7 1.6 ...
##  $ UNITS                         : chr  "ug/m3 LC" "ug/m3 LC" "ug/m3 LC" "ug/m3 LC" ...
##  $ DAILY_AQI_VALUE               : num  7 4 5 26 11 5 6 6 15 7 ...
##  $ Site Name                     : chr  "Linville Falls" "Linville Falls" "Linville Falls" "Linville
##  $ DAILY_OBS_COUNT               : num  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            : num  88502 88502 88502 88502 88502 ...
##  $ AQS_PARAMETER_DESC            : chr  "Acceptable PM2.5 AQI & Speciation Mass" "Acceptable PM2.5 A
##  $ CBSA_CODE                     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ CBSA_NAME                     : chr  NA NA NA NA ...
##  $ STATE_CODE                    : num  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                         : chr  "North Carolina" "North Carolina" "North Carolina" "North Ca
##  $ COUNTY_CODE                   : chr  "011" "011" "011" "011" ...
##  $ 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 ...
##  - 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
EPAair_O3_NC2018$Date <- as.Date(EPAair_O3_NC2018$Date, format = "%m/%d/%Y")
EPAair_O3_NC2019$Date <- as.Date(EPAair_O3_NC2019$Date, format = "%m/%d/%Y")
EPAair_PM25_NC2018$Date <- as.Date(EPAair_PM25_NC2018$Date, format = "%m/%d/%Y")
EPAair_PM25_NC2019$Date <- as.Date(EPAair_PM25_NC2019$Date, format = "%m/%d/%Y")
#4
## Selecting relevant columns
EPAair_O3_NC2018 <- EPAair_O3_NC2018 %>% select("Date", "DAILY_AQI_VALUE", "Site Name", "AQS_PARAMETER_
EPAair_O3_NC2019 <- EPAair_O3_NC2019 %>% select("Date", "DAILY_AQI_VALUE", "Site Name", "AQS_PARAMETER_
EPAair_PM25_NC2018 <- EPAair_PM25_NC2018 %>% select("Date", "DAILY_AQI_VALUE", "Site Name", "AQS_PARAME
EPAair_PM25_NC2019 <- EPAair_PM25_NC2019 %>% select("Date", "DAILY_AQI_VALUE", "Site Name", "AQS_PARAME
#5
EPAair_PM25_NC2018 <- EPAair_PM25_NC2018 %>% mutate(AQS_PARAMETER_DESC = "PM2.5")
EPAair_PM25_NC2019 <- EPAair_PM25_NC2019 %>% mutate(AQS_PARAMETER_DESC = "PM2.5")
#6
write.csv(EPAair_O3_NC2018, row.names = TRUE, file = "~/Courses/EnvDataAnalytics/Environmental_Data_Anal
write.csv(EPAair_O3_NC2019, row.names = TRUE, file = "~/Courses/EnvDataAnalytics/Environmental_Data_Anal
write.csv(EPAair_PM25_NC2018, row.names = TRUE, file = "~/Courses/EnvDataAnalytics/Environmental_Data_An
write.csv(EPAair_PM25_NC2019, row.names = TRUE, file = "~/Courses/EnvDataAnalytics/Environmental_Data_An
```

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
EPAair.combined <- rbind(EPAair_O3_NC2018, EPAair_O3_NC2019, EPAair_PM25_NC2018, EPAair_PM25_NC2019)
#8
EPAair.summary <- EPAair.combined %>% filter(`Site Name` %in% intersect(intersect(EPAair_O3_NC2018$`Site

#9
EPAair_O3_PM25 <- spread(EPAair.summary, AQS_PARAMETER_DESC, mean.AQI)
#10
dim(EPAair_O3_PM25)
```

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

```
#11
write.csv(EPAair_O3_PM25, file = "~/Courses/EnvDataAnalytics/Environmental_Data_Analytics_2020/Data/Proc
```

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
summary.df <- EPAair_03_PM25 %>% group_by(`Site Name`, Month, Year) %>% summarise(mean.AQI.Ozone = mean
#12b
summary.df <- summary.df %>% drop_na(Month, Year)
#13
dim(summary.df)
```

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

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

    Answer: The function na.omit would drop any row containing any NAs. We are still interested in rows that contain NAs (for example, if there is an AQI value for PM2.5 but not ozone), so dropping all rows containing NAs would not be appropriate.