Assignment 4: Data Wrangling

Jiaqi Li

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics (ENV872L) on data wrangling.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Use the lesson as a guide. It contains code that can be modified to complete the assignment.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file. You will need to have the correct software installed to do this (see Software Installation Guide) Press the `Knit` button in the RStudio scripting panel. This will save the PDF output in your Assignments folder.
6. After Knitting, please submit the completed exercise (PDF file) to the dropbox in Sakai. Please add your last name into the file name (e.g., "Salk_A04_DataWrangling.pdf") prior to submission.

The completed exercise is due on Thursday, 7 February, 2019 before class begins.

Set up your session

1. Check your working directory, load the `tidyverse` package, 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. Generate a few lines of code to get to know your datasets (basic data summaries, etc.).

```
#1 Preparation
getwd()
```

```
## [1] "/Users/ljq/Desktop/Blue Devils/Data Analysis/ENV872_02/Assignments"
```

```
setwd("/Users/ljq/Desktop/Blue Devils/Data Analysis/ENV872_02")
library(tidyverse)
pm2017 <- read.csv("./Data/Raw/EPAair_PM25_NC2017_raw.csv")
pm2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")
ozone2017 <- read.csv("./Data/Raw/EPAair_O3_NC2017_raw.csv")
ozone2018 <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv")
#2 Data summary
dim(pm2017)
```

```
## [1] 9494   20
```

```
head(pm2017)
```

```
##      Date Source   Site.ID POC Daily.Mean.PM2.5.Concentration    UNITS
## 1  1/1/17    AQS 370110002   1                             2.9 ug/m3 LC
## 2  1/4/17    AQS 370110002   1                             1.2 ug/m3 LC
## 3  1/7/17    AQS 370110002   1                             3.2 ug/m3 LC
```

```
## 4 1/10/17    AQS 370110002   1                           6.4 ug/m3 LC
## 5 1/13/17    AQS 370110002   1                           3.6 ug/m3 LC
## 6 1/16/17    AQS 370110002   1                           5.8 ug/m3 LC
##   DAILY_AQI_VALUE      Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              12 Linville Falls               1              100
## 2               5 Linville Falls               1              100
## 3              13 Linville Falls               1              100
## 4              27 Linville Falls               1              100
## 5              15 Linville Falls               1              100
## 6              24 Linville Falls               1              100
##   AQS_PARAMETER_CODE                     AQS_PARAMETER_DESC CBSA_CODE
## 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
##   CBSA_NAME STATE_CODE          STATE COUNTY_CODE COUNTY SITE_LATITUDE
## 1                  37 North Carolina          11  Avery      35.97235
## 2                  37 North Carolina          11  Avery      35.97235
## 3                  37 North Carolina          11  Avery      35.97235
## 4                  37 North Carolina          11  Avery      35.97235
## 5                  37 North Carolina          11  Avery      35.97235
## 6                  37 North Carolina          11  Avery      35.97235
##   SITE_LONGITUDE
## 1      -81.93307
## 2      -81.93307
## 3      -81.93307
## 4      -81.93307
## 5      -81.93307
## 6      -81.93307
```

```
colnames(pm2017)
```

```
##  [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"
```

```
dim(pm2018)
```

```
## [1] 7611   20
```

```
head(pm2018)
```

```
##      Date Source   Site.ID POC Daily.Mean.PM2.5.Concentration    UNITS
## 1  1/2/18    AQS 370110002   1                           2.9 ug/m3 LC
## 2  1/5/18    AQS 370110002   1                           3.7 ug/m3 LC
## 3  1/8/18    AQS 370110002   1                           5.3 ug/m3 LC
## 4 1/11/18    AQS 370110002   1                           0.8 ug/m3 LC
```

```
## 5 1/14/18    AQS 370110002   1                          2.5 ug/m3 LC
## 6 1/17/18    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
## 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
##   CBSA_NAME STATE_CODE          STATE COUNTY_CODE COUNTY SITE_LATITUDE
## 1                  37 North Carolina          11  Avery      35.97235
## 2                  37 North Carolina          11  Avery      35.97235
## 3                  37 North Carolina          11  Avery      35.97235
## 4                  37 North Carolina          11  Avery      35.97235
## 5                  37 North Carolina          11  Avery      35.97235
## 6                  37 North Carolina          11  Avery      35.97235
##   SITE_LONGITUDE
## 1      -81.93307
## 2      -81.93307
## 3      -81.93307
## 4      -81.93307
## 5      -81.93307
## 6      -81.93307
```

```
colnames(pm2018)
```

```
##  [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"
```

```
dim(ozone2017)
```

```
## [1] 10219    20
```

```
head(ozone2017)
```

```
##     Date Source   Site.ID POC Daily.Max.8.hour.Ozone.Concentration UNITS
## 1 3/1/17    AQS 370030005   1                                 0.041   ppm
## 2 3/2/17    AQS 370030005   1                                 0.046   ppm
## 3 3/3/17    AQS 370030005   1                                 0.046   ppm
## 4 3/4/17    AQS 370030005   1                                 0.046   ppm
## 5 3/5/17    AQS 370030005   1                                 0.046   ppm
```

```
## 6 3/6/17    AQS 370030005   1                                0.048   ppm
##   DAILY_AQI_VALUE             Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              38 Taylorsville Liledoun              17              100
## 2              43 Taylorsville Liledoun              17              100
## 3              43 Taylorsville Liledoun              17              100
## 4              43 Taylorsville Liledoun              17              100
## 5              43 Taylorsville Liledoun              17              100
## 6              44 Taylorsville Liledoun              17              100
##   AQS_PARAMETER_CODE AQS_PARAMETER_DESC CBSA_CODE
## 1              44201              Ozone     25860
## 2              44201              Ozone     25860
## 3              44201              Ozone     25860
## 4              44201              Ozone     25860
## 5              44201              Ozone     25860
## 6              44201              Ozone     25860
##                      CBSA_NAME STATE_CODE          STATE COUNTY_CODE
## 1 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
## 2 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
## 3 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
## 4 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
## 5 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
## 6 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
##      COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1 Alexander       35.9138        -81.191
## 2 Alexander       35.9138        -81.191
## 3 Alexander       35.9138        -81.191
## 4 Alexander       35.9138        -81.191
## 5 Alexander       35.9138        -81.191
## 6 Alexander       35.9138        -81.191
```

```
colnames(ozone2017)
```

```
##  [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"
```

```
dim(ozone2018)
```

```
## [1] 10781    20
```

```
head(ozone2018)
```

```
##      Date Source   Site.ID POC Daily.Max.8.hour.Ozone.Concentration UNITS
## 1 2/16/18 AirNow 370030005   1                                 0.038   ppm
## 2 2/17/18 AirNow 370030005   1                                 0.033   ppm
## 3 2/18/18 AirNow 370030005   1                                 0.040   ppm
## 4 2/19/18 AirNow 370030005   1                                 0.020   ppm
## 5 2/20/18 AirNow 370030005   1                                 0.019   ppm
## 6 2/21/18 AirNow 370030005   1                                 0.021   ppm
##   DAILY_AQI_VALUE            Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              35 Taylorsville Liledoun              24              100
## 2              31 Taylorsville Liledoun              24              100
## 3              37 Taylorsville Liledoun              24              100
## 4              19 Taylorsville Liledoun              24              100
## 5              18 Taylorsville Liledoun              24              100
## 6              19 Taylorsville Liledoun              24              100
##   AQS_PARAMETER_CODE AQS_PARAMETER_DESC CBSA_CODE
## 1              44201              Ozone     25860
## 2              44201              Ozone     25860
## 3              44201              Ozone     25860
## 4              44201              Ozone     25860
## 5              44201              Ozone     25860
## 6              44201              Ozone     25860
##                      CBSA_NAME STATE_CODE          STATE COUNTY_CODE
## 1 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
## 2 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
## 3 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
## 4 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
## 5 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
## 6 Hickory-Lenoir-Morganton, NC         37 North Carolina           3
##      COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1 Alexander       35.9138        -81.191
## 2 Alexander       35.9138        -81.191
## 3 Alexander       35.9138        -81.191
## 4 Alexander       35.9138        -81.191
## 5 Alexander       35.9138        -81.191
## 6 Alexander       35.9138        -81.191
```

```
colnames(ozone2018)
```

```
##  [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"
```

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.

```
#3 Change date type
pm2017$Date <- as.Date(pm2017$Date, format="%m/%d/%y")
pm2018$Date <- as.Date(pm2018$Date, format="%m/%d/%y")
ozone2017$Date <- as.Date(ozone2017$Date, format="%m/%d/%y")
ozone2018$Date <- as.Date(ozone2018$Date, format="%m/%d/%y")
#4 Select columns
pm2017_new <- select(pm2017, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
                     COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
pm2018_new <- select(pm2018, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
                     COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
ozone2017_new <- select(ozone2017, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
                        COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
ozone2018_new <- select(ozone2018, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
                        COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
#5 Fill AQS_PARAMETER_DESC cells
pm2017_new$AQS_PARAMETER_DESC <- "PM2.5"
pm2018_new$AQS_PARAMETER_DESC <- "PM2.5"
#6 Save files
write.csv(pm2017_new, row.names = FALSE,
          file = paste("/Users/ljq/Desktop/Blue Devils/Data Analysis/ENV872_02/Data",
                       "/Processed/EPAair_PM25_NC2017_Processed.csv",sep =""))
write.csv(pm2018_new, row.names = FALSE,
          file = paste("/Users/ljq/Desktop/Blue Devils/Data Analysis/ENV872_02/Data",
                       "/Processed/EPAair_PM25_NC2018_Processed.csv",sep =""))
write.csv(ozone2017_new, row.names = FALSE,
           file = paste("/Users/ljq/Desktop/Blue Devils/Data Analysis/ENV872_02/Data",
                       "/Processed/EPAair_O3_NC2017_Processed.csv",sep =""))
write.csv(ozone2018_new, row.names = FALSE,
          file = paste("/Users/ljq/Desktop/Blue Devils/Data Analysis/ENV872_02/Data",
                       "/Processed/EPAair_O3_NC2018_Processed.csv",sep =""))
```

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:

- Sites: Blackstone, Bryson City, Triple Oak

- Add columns for "Month" and "Year" by parsing your "Date" column (hint: `separate` function or `lubridate` package)

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 Combine the four datasets
air_combine <- rbind(pm2017_new, pm2018_new, ozone2017_new, ozone2018_new)
#8 Pipe function
library(lubridate)
air_combine.processed <-
  air_combine %>%
  filter(air_combine$Site.Name %in% c("Blackstone", "Bryson City", "Triple Oak")) %>%
  mutate(Month = month(Date), Year = year(Date))
#9 Spread the dataset
air_combine.processed <- spread(air_combine.processed, AQS_PARAMETER_DESC, DAILY_AQI_VALUE)
#10 Dimension of the dataset
dim(air_combine.processed)
```

```
## [1] 1953    9
```

```
#11 Save the dataset
write.csv(air_combine.processed, row.names = FALSE,
          file = paste("/Users/ljq/Desktop/Blue Devils/Data Analysis/ENV872_02/Data",
                       "/Processed/EPAair_O3_PM25_NC1718_Processed.csv",sep =""))
```

Generate summary tables

12. Use the split-apply-combine strategy to generate two new data frames:
    a. A summary table of mean AQI values for O3 and PM2.5 by month
    b. A summary table of the mean, minimum, and maximum AQI values of O3 and PM2.5 for each site
13. Display the data frames.

```
#12a Mean AQI values for O3 and PM2.5 by month
air.month.summaries <-
  air_combine.processed %>%
  group_by(Month) %>%
  filter(!is.na(Ozone) & !is.na(PM2.5)) %>%
  summarise(mean_ozone = mean(Ozone), mean_pm = mean(PM2.5))
#12b AQI values summary for O3 and PM2.5 by site
air.site.summaries <-
  air_combine.processed %>%
  group_by(Site.Name) %>%
  filter(!is.na(Ozone) & !is.na(PM2.5)) %>%
  summarise(mean_ozone = mean(Ozone), min_ozone = min(Ozone), max_ozone = max(Ozone),
            mean_pm = mean(PM2.5), min_pm = min(PM2.5), max_pm = max(PM2.5))
#13 Display the data frames
knitr::kable(air.month.summaries, caption = 'Mean AQI values for O3 and PM2.5 by month')
```

Table 1: Mean AQI values for O3 and PM2.5 by month

Month	mean_ozone	mean_pm
1	31.48276	34.24138
2	35.41176	37.57353
3	42.40164	37.40984
4	43.48598	31.52336
5	39.49057	30.63208
6	39.16981	30.92453
7	38.32787	31.92623
8	34.40449	32.33708
9	32.64000	30.65333
10	32.29412	30.12941
11	30.06897	42.13793
12	29.78378	46.62162

```
knitr::kable(air.site.summaries, caption = 'AQI summary for O3 and PM2.5 by site')
```

Table 2: AQI summary for O3 and PM2.5 by site

Site.Name	mean_ozone	min_ozone	max_ozone	mean_pm	min_pm	max_pm
Blackstone	38.30237	8	97	36.66485	0	83
Bryson City	35.42769	5	71	30.32231	3	68