Assignment 4: Data Wrangling

Kat Horan

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
# Set working directory
# setwd("/Users/kathleenhoran/Desktop/Duke/Spring 2019/Env. Data Analytics/Env_Data_Analytics")

# Load package
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------- tidyverse 1.2
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------ tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
## 
##     date
```

```
library(pander)

# Upload data
EPAair_O3_NC17_raw <- read.csv("./Data/Raw/EPAair_O3_NC2017_raw.csv")
EPAair_O3_NC18_raw <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv")
EPAair_PM25_NC17_raw <- read.csv("./Data/Raw/EPAair_PM25_NC2017_raw.csv")
EPAair_PM25_NC18_raw <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")

#2 Getting to know the data
head(EPAair_O3_NC17_raw)
```

```
##     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
```

```
head(EPAair_PM25_NC17_raw)
```

```
##      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(EPAair_O3_NC17_raw)
```

```
##  [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(EPAair_PM25_NC17_raw)
```

```
##  [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(EPAair_O3_NC17_raw)
```

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

```
dim(EPAair_PM25_NC17_raw)
```

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

```
summary(EPAair_O3_NC17_raw)
```

```
##       Date      Source         Site.ID               POC
##  4/13/17:  40   AQS:10219   Min.   :370030005   Min.   :1
##  4/15/17:  40               1st Qu.:370650099   1st Qu.:1
##  4/18/17:  40               Median :371010002   Median :1
##  4/3/17 :  40               Mean   :370962005   Mean   :1
##  4/5/17 :  40               3rd Qu.:371239991   3rd Qu.:1
##  4/8/17 :  40               Max.   :371990004   Max.   :1
##  (Other):9979
##  Daily.Max.8.hour.Ozone.Concentration UNITS       DAILY_AQI_VALUE
##  Min.   :0.00500                      ppm:10219   Min.   :  5.00
##  1st Qu.:0.03500                                  1st Qu.: 32.00
##  Median :0.04300                                  Median : 40.00
##  Mean   :0.04211                                  Mean   : 39.87
##  3rd Qu.:0.04900                                  3rd Qu.: 45.00
##  Max.   :0.07500                                  Max.   :115.00
##
##                Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Garinger High School: 358   Min.   :13.00   Min.   : 76.00
##  Blackstone          : 355   1st Qu.:17.00   1st Qu.:100.00
##  Rockwell            : 354   Median :17.00   Median :100.00
##  Coweeta             : 344   Mean   :16.94   Mean   : 99.63
##  Millbrook School    : 339   3rd Qu.:17.00   3rd Qu.:100.00
##  Beaufort            : 338   Max.   :17.00   Max.   :100.00
##  (Other)             :8131
##  AQS_PARAMETER_CODE AQS_PARAMETER_DESC   CBSA_CODE
##  Min.   :44201      Ozone:10219        Min.   :11700
##  1st Qu.:44201                         1st Qu.:16740
##  Median :44201                         Median :24660
##  Mean   :44201                         Mean   :27541
##  3rd Qu.:44201                         3rd Qu.:39580
##  Max.   :44201                         Max.   :49180
##                                        NA's   :2541
##                               CBSA_NAME      STATE_CODE
##                                    :2541   Min.   :37  
##  Charlotte-Concord-Gastonia, NC-SC:1428   1st Qu.:37  
##  Asheville, NC                    : 940   Median :37  
##  Winston-Salem, NC                : 725   Mean   :37  
##  Raleigh, NC                      : 584   3rd Qu.:37  
##  Durham-Chapel Hill, NC           : 486   Max.   :37  
##  (Other)                          :3515               
##             STATE        COUNTY_CODE             COUNTY    
##  North Carolina:10219   Min.   :  3.00   Forsyth    : 725  
##                         1st Qu.: 65.00   Haywood    : 700  
##                         Median :101.00   Mecklenburg: 601  
##                         Mean   : 96.07   Avery      : 541  
##                         3rd Qu.:123.00   Cumberland : 464  
##                         Max.   :199.00   Swain      : 429  
##                                          (Other)    :6759  
##  SITE_LATITUDE   SITE_LONGITUDE  
##  Min.   :34.36   Min.   :-83.80  
##  1st Qu.:35.26   1st Qu.:-82.05  
##  Median :35.55   Median :-80.23  
##  Mean   :35.60   Mean   :-80.32  
##  3rd Qu.:35.99   3rd Qu.:-78.77  
##  Max.   :36.31   Max.   :-76.62  
## 
```

```
summary(EPAair_PM25_NC17_raw)
```

```
##       Date      Source        Site.ID               POC       
##  1/31/17:  45   AQS:9494   Min.   :370110002   Min.   :1.000  
##  1/19/17:  44              1st Qu.:370630015   1st Qu.:3.000  
##  11/3/17:  44              Median :371010002   Median :3.000  
##  2/12/17:  44              Mean   :370980114   Mean   :2.734  
##  4/1/17 :  44              3rd Qu.:371210004   3rd Qu.:3.000  
##  5/31/17:  44              Max.   :371830021   Max.   :4.000  
##  (Other):9229                                                 
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-3.900                 ug/m3 LC:9494   Min.   : 0.00  
##  1st Qu.: 5.000                                 1st Qu.:21.00  
##  Median : 7.300                                 Median :30.00  
##  Mean   : 7.742                                 Mean   :31.72  
##  3rd Qu.:10.000                                 3rd Qu.:42.00  
##  Max.   :31.900                                 Max.   :93.00  
##                                                                
##                         Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Board Of Ed. Bldg.          : 542   Min.   :1       Min.   :100     
##  Hattie Avenue               : 505   1st Qu.:1       1st Qu.:100     
##  Lexington water tower       : 501   Median :1       Median :100     
##  Montclaire Elementary School: 489   Mean   :1       Mean   :100     
##  Pitt Agri. Center           : 483   3rd Qu.:1       3rd Qu.:100     
##  West Johnston Co.           : 478   Max.   :1       Max.   :100     
##  (Other)                     :6496                                   
##  AQS_PARAMETER_CODE                              AQS_PARAMETER_DESC
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:2842    
##  1st Qu.:88101      PM2.5 - Local Conditions              :6652    
##  Median :88101                                                     
##  Mean   :88221
##  3rd Qu.:88502
##  Max.   :88502
##
##    CBSA_CODE                                 CBSA_NAME      STATE_CODE
##  Min.   :11700   Charlotte-Concord-Gastonia, NC-SC:1411   Min.   :37
##  1st Qu.:16740   Winston-Salem, NC                :1366   1st Qu.:37
##  Median :25860                                    :1353   Median :37
##  Mean   :30793   Raleigh, NC                      :1285   Mean   :37
##  3rd Qu.:41820   Asheville, NC                    : 657   3rd Qu.:37
##  Max.   :49180   Greenville, NC                   : 483   Max.   :37
##  NA's   :1353    (Other)                          :2939
##             STATE       COUNTY_CODE          COUNTY     SITE_LATITUDE
##  North Carolina:9494   Min.   : 11   Mecklenburg:1411   Min.   :34.36
##                        1st Qu.: 63   Forsyth    : 865   1st Qu.:35.26
##                        Median :101   Wake       : 807   Median :35.64
##                        Mean   : 98   Buncombe   : 542   Mean   :35.60
##                        3rd Qu.:121   Davidson   : 501   3rd Qu.:35.91
##                        Max.   :183   Pitt       : 483   Max.   :36.11
##                                      (Other)    :4885
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-80.03
##  3rd Qu.:-78.82
##  Max.   :-76.21
##
```

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
# Checking to see what class it is
class(EPAair_O3_NC17_raw$Date)
```

```
## [1] "factor"
```

```
# Changing the class
EPAair_O3_NC17_raw$Date <- as.Date(EPAair_O3_NC17_raw$Date, format = "%m/%d/%y")

EPAair_O3_NC18_raw$Date <- as.Date(EPAair_O3_NC18_raw$Date, format = "%m/%d/%y")

EPAair_PM25_NC17_raw$Date <- as.Date(EPAair_PM25_NC17_raw$Date, format = "%m/%d/%y")

EPAair_PM25_NC18_raw$Date <- as.Date(EPAair_PM25_NC18_raw$Date, format = "%m/%d/%y")

# Confirming that class was changed to Date
```

```
class(EPAair_O3_NC17_raw$Date)
```

```
## [1] "Date"
```

```
#4 Selecting only the desired columns
EPAair_O3_NC17_skinny <- select(EPAair_O3_NC17_raw, Date, DAILY_AQI_VALUE, Site.Name,
      AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

EPAair_O3_NC18_skinny <- select(EPAair_O3_NC18_raw, Date, DAILY_AQI_VALUE, Site.Name,
      AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

EPAair_PM25_NC17_skinny <- select(EPAair_PM25_NC17_raw, Date, DAILY_AQI_VALUE, Site.Name,
      AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

EPAair_PM25_NC18_skinny <- select(EPAair_PM25_NC18_raw, Date, DAILY_AQI_VALUE, Site.Name,
      AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

#5 Changing the values in AQS_PARAMETER_DESC
EPAair_PM25_NC17_skinny$AQS_PARAMETER_DESC <- "PM2.5"
EPAair_PM25_NC18_skinny$AQS_PARAMETER_DESC <- "PM2.5"

#6 Saving to Processed folder


write.csv(EPAair_O3_NC17_skinny, row.names = FALSE,
    file = "./Data/Processed/EPAair_O3_NC17_skinny.csv")

write.csv(EPAair_O3_NC18_skinny, row.names = FALSE,
    file = "./Data/Processed/EPAair_O3_NC18_skinny.csv")

write.csv(EPAair_PM25_NC17_skinny, row.names = FALSE,
    file = "./Data/Processed/EPAair_PM25_NC17_skinny.csv")

write.csv(EPAair_PM25_NC18_skinny, row.names = FALSE,
    file = "./Data/Processed/EPAair_PM25_NC18_skinny.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:

- 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 Combining the datasets

EPAair_combined <- rbind(EPAair_O3_NC17_skinny, EPAair_O3_NC18_skinny,
                  EPAair_PM25_NC17_skinny, EPAair_PM25_NC18_skinny)
```

```
#8 Pipe
EPAair_combined_processed <-
  EPAair_combined %>%
  filter(Site.Name == "Blackstone" | Site.Name == "Bryson City" |
         Site.Name == "Triple Oak") %>%
  mutate(Month = month(Date)) %>%
  mutate(Year = year(Date))

#9 Spread
EPAair_combined_processed_spread <- spread(EPAair_combined_processed,
                                    AQS_PARAMETER_DESC, DAILY_AQI_VALUE)

#10 Finding dimensions
dim(EPAair_combined_processed_spread)
```

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

```
#11 Saving processed dataset

write.csv(EPAair_combined_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 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 A summary table of mean AQI values for O3 and PM2.5 by month
EPAair_O3_PM25_NC1718_MeanAQISbyMonth_summary <-
  EPAair_combined_processed_spread %>%
  group_by(Month) %>%
  filter(!is.na(Ozone) & !is.na(PM2.5)) %>%
  summarise(Monthly.Ozone.Avg = mean(Ozone),
            Monthly.PM25.Avg = mean(PM2.5))

#12b A summary table of the mean, minimum, and maximum AQI values of O3 and PM2.5 for each site
EPAair_O3_PM25_NC1718_Sites_summary <-
  EPAair_combined_processed_spread %<>%
  group_by(Site.Name) %>%
  filter(!is.na(Ozone) & !is.na(PM2.5)) %>%
  summarise(Ozone.Avg.by.Site = mean(Ozone),
            Ozone.Min.by.Site = min(Ozone),
            Ozone.Max.by.Site = max(Ozone),
            PM25.Avg.by.Site = mean(PM2.5),
            PM25.Min.by.Site = min(PM2.5),
            PM25.Max.by.Site = max(PM2.5)
  )

#13 Display dataframe
```

```
pander(EPAair_O3_PM25_NC1718_MeanAQISbyMonth_summary, type = 'grid')
```

Month	Monthly.Ozone.Avg	Monthly.PM25.Avg
1	31.48	34.24
2	35.41	37.57
3	42.4	37.41
4	43.49	31.52
5	39.49	30.63
6	39.17	30.92
7	38.33	31.93
8	34.4	32.34
9	32.64	30.65
10	32.29	30.13
11	30.07	42.14
12	29.78	46.62

```
pander(EPAair_O3_PM25_NC1718_Sites_summary, type = 'grid')
```

Table 2: Table continues below

Site.Name	Ozone.Avg.by.Site	Ozone.Min.by.Site	Ozone.Max.by.Site
Blackstone	38.3	8	97
Bryson City	35.43	5	71

PM25.Avg.by.Site	PM25.Min.by.Site	PM25.Max.by.Site
36.66	0	83
30.32	3	68