Assignment 4: Data Wrangling

Ying Wei Jong

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
getwd()
```

```
## [1] "/Users/YwJong/Documents/NSOE/Spring 2019/ENV 872 Environment Data Analytics/Labs"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

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

```
library(lubridate) #For question 8
```

```
##
## Attaching package: 'lubridate'
```

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

```
O3_2017 <- read.csv("./Data/Raw/EPAair_O3_NC2017_raw.csv")
O3_2018 <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv")
PM25_2017 <- read.csv("./Data/Raw/EPAair_PM25_NC2017_raw.csv")
PM25_2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")

#2 I will not repeat the same line over all four datasets, it occupies too much space
head(O3_2017)
```

```
##     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
```

```
summary(O3_2018)
```

```
##       Date          Source        Site.ID               POC
##  3/10/18:   39   AirNow:2718   Min.   :370030005   Min.   :1
##  3/11/18:   39   AQS   :8063   1st Qu.:370630015   1st Qu.:1
##  3/13/18:   39                 Median :370870036   Median :1
##  3/14/18:   39                 Mean   :370959550   Mean   :1
```

```
##  3/15/18:   39                  3rd Qu.:371290002   3rd Qu.:1
##  3/16/18:   39                  Max.   :371990004   Max.   :1
##  (Other):10547
##  Daily.Max.8.hour.Ozone.Concentration UNITS       DAILY_AQI_VALUE
##  Min.   :0.00000                      ppm:10781   Min.   :  0.00
##  1st Qu.:0.03400                                  1st Qu.: 31.00
##  Median :0.04100                                  Median : 38.00
##  Mean   :0.04124                                  Mean   : 39.46
##  3rd Qu.:0.04900                                  3rd Qu.: 45.00
##  Max.   :0.07700                                  Max.   :122.00
##
##                 Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Coweeta             : 340   Min.   :12.00   Min.   : 71.00
##  Millbrook School    : 338   1st Qu.:17.00   1st Qu.:100.00
##  Candor              : 337   Median :17.00   Median :100.00
##  Garinger High School: 333   Mean   :18.69   Mean   : 99.62
##  Bethany sch.        : 332   3rd Qu.:18.00   3rd Qu.:100.00
##  Cranberry           : 319   Max.   :24.00   Max.   :100.00
##  (Other)             :8782
##  AQS_PARAMETER_CODE AQS_PARAMETER_DESC   CBSA_CODE
##  Min.   :44201      Ozone:10781        Min.   :11700
##  1st Qu.:44201                         1st Qu.:16740
##  Median :44201                         Median :24660
##  Mean   :44201                         Mean   :27015
##  3rd Qu.:44201                         3rd Qu.:39580
##  Max.   :44201                         Max.   :49180
##                                        NA's   :2802
##                              CBSA_NAME      STATE_CODE
##                                   :2802   Min.   :37
##  Charlotte-Concord-Gastonia, NC-SC:1469   1st Qu.:37
##  Asheville, NC                    :1159   Median :37
##  Winston-Salem, NC                : 754   Mean   :37
##  Raleigh, NC                      : 636   3rd Qu.:37
##  Greensboro-High Point, NC        : 595   Max.   :37
##  (Other)                          :3366
##            STATE        COUNTY_CODE            COUNTY
##  North Carolina:10781   Min.   :  3.00   Haywood    : 879
##                         1st Qu.: 63.00   Forsyth    : 754
##                         Median : 87.00   Mecklenburg: 632
##                         Mean   : 95.84   Avery      : 613
##                         3rd Qu.:129.00   Cumberland : 467
##                         Max.   :199.00   Swain      : 447
##                                          (Other)    :6989
##  SITE_LATITUDE   SITE_LONGITUDE
##  Min.   :34.36   Min.   :-83.80
##  1st Qu.:35.26   1st Qu.:-82.05
##  Median :35.59   Median :-80.34
##  Mean   :35.63   Mean   :-80.39
##  3rd Qu.:36.03   3rd Qu.:-78.90
##  Max.   :36.31   Max.   :-76.62
##
```

```
str(PM25_2017)
```

```
## 'data.frame':    9494 obs. of  20 variables:
```

```
##  $ Date                          : Factor w/ 365 levels "1/1/17","1/10/17",..: 1 26 29 2 5 8 11 15 18
##  $ Source                        : Factor w/ 1 level "AQS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Site.ID                       : int  370110002 370110002 370110002 370110002 370110002 370110002 3
##  $ POC                           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Daily.Mean.PM2.5.Concentration: num  2.9 1.2 3.2 6.4 3.6 5.8 3.6 1.5 1.4 1.4 ...
##  $ UNITS                         : Factor w/ 1 level "ug/m3 LC": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DAILY_AQI_VALUE               : int  12 5 13 27 15 24 15 6 6 6 ...
##  $ Site.Name                     : Factor w/ 25 levels "","Blackstone",..: 15 15 15 15 15 15 15 15 15
##  $ DAILY_OBS_COUNT               : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ PERCENT_COMPLETE              : int  100 100 100 100 100 100 100 100 100 100 ...
##  $ AQS_PARAMETER_CODE            : int  88502 88502 88502 88502 88502 88502 88502 88502 88502 88502
##  $ AQS_PARAMETER_DESC            : Factor w/ 2 levels "Acceptable PM2.5 AQI & Speciation Mass",..: 1
##  $ CBSA_CODE                     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ CBSA_NAME                     : Factor w/ 14 levels "","Asheville, NC",..: 1 1 1 1 1 1 1 1 1 1 ..
##  $ STATE_CODE                    : int  37 37 37 37 37 37 37 37 37 37 ...
##  $ STATE                         : Factor w/ 1 level "North Carolina": 1 1 1 1 1 1 1 1 1 1 ...
##  $ COUNTY_CODE                   : int  11 11 11 11 11 11 11 11 11 11 ...
##  $ COUNTY                        : Factor w/ 21 levels "Avery","Buncombe",..: 1 1 1 1 1 1 1 1 1 1 ..
##  $ SITE_LATITUDE                 : num  36 36 36 36 36 ...
##  $ SITE_LONGITUDE                : num  -81.9 -81.9 -81.9 -81.9 -81.9 ...
```

```
dim(PM25_2018)
```

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

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
O3_2017$Date <- as.Date(O3_2017$Date, format = "%m/%d/%y")
O3_2018$Date <- as.Date(O3_2018$Date, format = "%m/%d/%y")
PM25_2017$Date <- as.Date(PM25_2017$Date, format = "%m/%d/%y")
PM25_2018$Date <- as.Date(PM25_2018$Date, format = "%m/%d/%y")

#4
O3_2017_Pro <- select(O3_2017, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITU
O3_2018_Pro <- select(O3_2018, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITU
PM25_2017_Pro <- select(PM25_2017, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LA
PM25_2018_Pro <- select(PM25_2018, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LA

#5
PM25_2017_Pro$AQS_PARAMETER_DESC <- "PM2.5"
PM25_2018_Pro$AQS_PARAMETER_DESC <- "PM2.5"

#6
write.csv(O3_2017_Pro, row.names=FALSE, file = "./Data/Processed/EPAair_O3_NC2017_processed.csv")
write.csv(O3_2018_Pro, row.names=FALSE, file = "./Data/Processed/EPAair_O3_NC2018_processed.csv")
write.csv(PM25_2017_Pro, row.names = FALSE, file =
```

```
"./Data/Processed/EPAair_PM25_NC2017_processed.csv")
write.csv(PM25_2018_Pro, row.names = FALSE, file =
"./Data/Processed/EPAair_PM25_NC2018_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:

- 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
O3_com <- full_join(O3_2017_Pro,O3_2018_Pro)
```

```
## Joining, by = c("Date", "DAILY_AQI_VALUE", "Site.Name", "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITU
```

```
## Warning: Column `Site.Name` joining factors with different levels, coercing
## to character vector
```

```
## Warning: Column `COUNTY` joining factors with different levels, coercing to
## character vector
```

```
PM25_com <- full_join(PM25_2017_Pro,PM25_2018_Pro)
```

```
## Joining, by = c("Date", "DAILY_AQI_VALUE", "Site.Name", "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITU
```

```
## Warning: Column `Site.Name` joining factors with different levels, coercing
## to character vector
```

```
Combined_data <- full_join(O3_com, PM25_com)
```

```
## Joining, by = c("Date", "DAILY_AQI_VALUE", "Site.Name", "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITU
```

```
## Warning: Column `AQS_PARAMETER_DESC` joining factor and character vector,
## coercing into character vector
```

```
## Warning: Column `COUNTY` joining character vector and factor, coercing into
## character vector
```

```
str(Combined_data)
```

```
## 'data.frame':    38105 obs. of  7 variables:
##  $ Date              : Date, format: "2017-03-01" "2017-03-02" ...
##  $ DAILY_AQI_VALUE   : int  38 43 43 43 43 44 44 49 54 44 ...
##  $ Site.Name         : chr  "Taylorsville Liledoun" "Taylorsville Liledoun" "Taylorsville Liledoun" "
##  $ AQS_PARAMETER_DESC: chr  "Ozone" "Ozone" "Ozone" "Ozone" ...
##  $ 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 ...
```

```r
#8
Combined_data <-
  Combined_data %>%
  filter(Site.Name == "Blackstone" | Site.Name == "Bryson City" | Site.Name == "Triple Oak") %>%
  mutate(Date, month= month(Date)) %>%
  mutate(Date, year=year(Date))

#separate(Combined_data, Date, c("Year", "Month", "d")) <- cannot specify df name otherwise command won

#Apparently I cannot specify the name of dataframe again when doing the lubridate/separate in piping op

#9
Combined_data <- spread(Combined_data, AQS_PARAMETER_DESC, DAILY_AQI_VALUE)

#10
dim(Combined_data)
```

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

```r
#11
write.csv(Combined_data, row.names = FALSE, file = "./Data/Processed/EPAair_O3_PM25_NC1718_Processed.cs
```

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 of O3 and PM2.5 for each site
13. Display the data frames.

```r
#12a
MonthlyAQI <-
  Combined_data %>%
  group_by(month) %>%
  filter(!is.na(Ozone) & !is.na(PM2.5)) %>%
  summarise(meanO3 = mean(Ozone),
            meanPM25 =mean(PM2.5))

#Have to remove NAs otherwise the resulting summary tables are full of NAs
#12b
SiteAQI <-
  Combined_data %>%
  group_by(Site.Name) %>%
  filter(!is.na(Ozone) & !is.na(PM2.5)) %>%
  summarise(meanO3 = mean(Ozone),
            maxO3 = max(Ozone),
            minO3 = min(Ozone),
            meanPM25 = mean(PM2.5),
            maxPM25 = max(PM2.5),
            minPM25 = min(PM2.5))
#Apparently Triple oak does not have complete (non NA) entries for both O3 and PM2.5

#13
MonthlyAQI
```

```
## # A tibble: 12 x 3
##    month meanO3 meanPM25
##    <dbl>  <dbl>    <dbl>
##  1     1   31.5     34.2
##  2     2   35.4     37.6
##  3     3   42.4     37.4
##  4     4   43.5     31.5
##  5     5   39.5     30.6
##  6     6   39.2     30.9
##  7     7   38.3     31.9
##  8     8   34.4     32.3
##  9     9   32.6     30.7
## 10    10   32.3     30.1
## 11    11   30.1     42.1
## 12    12   29.8     46.6
```

```
SiteAQI
```

```
## # A tibble: 2 x 7
##   Site.Name   meanO3 maxO3 minO3 meanPM25 maxPM25 minPM25
##   <chr>        <dbl> <dbl> <dbl>    <dbl>   <dbl>   <dbl>
## 1 Blackstone    38.3    97     8     36.7      83       0
## 2 Bryson City   35.4    71     5     30.3      68       3
```