Assignment 4: Data Wrangling

Siying Chen

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
# Check working directory
getwd()
```

```
## [1] "/Users/Sylvia/Downloads/ENV872/ENV872"
```

```
# Load package
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidy
```

```
## 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_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Import EPA air dataset
EPA_O3_17 <- read.csv("./Data/Raw/EPAair_O3_NC2017_raw.csv")
EPA_O3_18 <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv")
```

```
EPA_PM25_17 <- read.csv("./Data/Raw/EPAair_PM25_NC2017_raw.csv")
EPA_PM25_18 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")

#2
head(EPA_O3_17)
```

```
##     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(EPA_O3_18)
```

```
##  [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(EPA_PM25_17)
```

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

```
summary(EPA_PM25_18)
```

```
##       Date          Source        Site.ID               POC
##  1/26/18:  39   AirNow: 783   Min.   :370110002   Min.   :1.000
##  2/1/18 :  39   AQS   :6828   1st Qu.:370630015   1st Qu.:3.000
##  2/19/18:  39                 Median :371190041   Median :3.000
##  1/14/18:  38                 Mean   :371031969   Mean   :3.011
##  1/8/18 :  38                 3rd Qu.:371290002   3rd Qu.:3.000
##  2/7/18 :  38                 Max.   :371830021   Max.   :5.000
##  (Other):7380
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-2.800                 ug/m3 LC:7611   Min.   : 0.00
##  1st Qu.: 5.000                                 1st Qu.:21.00
##  Median : 7.200                                 Median :30.00
##  Mean   : 7.554                                 Mean   :31.03
##  3rd Qu.: 9.800                                 3rd Qu.:41.00
##  Max.   :34.200                                 Max.   :97.00
##
##                  Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Millbrook School     : 621   Min.   :1       Min.   :100
##  Board Of Ed. Bldg.   : 428   1st Qu.:1       1st Qu.:100
##  Garinger High School : 421   Median :1       Median :100
##  Durham Armory        : 415   Mean   :1       Mean   :100
##  Lexington water tower: 411   3rd Qu.:1       3rd Qu.:100
##  Pitt Agri. Center    : 409   Max.   :1       Max.   :100
##  (Other)              :4906
##  AQS_PARAMETER_CODE                              AQS_PARAMETER_DESC
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:1246
##  1st Qu.:88101      PM2.5 - Local Conditions              :6365
##  Median :88101
##  Mean   :88167
##  3rd Qu.:88101
##  Max.   :88502
##
##    CBSA_CODE                                  CBSA_NAME      STATE_CODE
##  Min.   :11700   Raleigh, NC                      :1274   Min.   :37
##  1st Qu.:19000   Charlotte-Concord-Gastonia, NC-SC:1171   1st Qu.:37
##  Median :25860                                    :1025   Median :37
##  Mean   :30249   Winston-Salem, NC                : 803   Mean   :37
##  3rd Qu.:39580   Asheville, NC                    : 447   3rd Qu.:37
```

```
##  Max.   :49180   Durham-Chapel Hill, NC           : 415   Max.   :37
##  NA's   :1025    (Other)                          :2476
##             STATE       COUNTY_CODE            COUNTY     SITE_LATITUDE
##  North Carolina:7611   Min.   : 11.0   Mecklenburg:1171   Min.   :34.36
##                        1st Qu.: 63.0   Wake       : 947   1st Qu.:35.26
##                        Median :119.0   Buncombe   : 428   Median :35.64
##                        Mean   :103.2   Durham     : 415   Mean   :35.59
##                        3rd Qu.:129.0   Davidson   : 411   3rd Qu.:35.87
##                        Max.   :183.0   Pitt       : 409   Max.   :36.11
##                                        (Other)    :3830
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-79.84
##  Mean   :-79.95
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

```
class(EPA_O3_18$Date)
```

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

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
EPA_O3_17$Date <- as.Date(EPA_O3_17$Date, format = "%m/%d/%y")
EPA_O3_18$Date <- as.Date(EPA_O3_18$Date, format = "%m/%d/%y")
EPA_PM25_17$Date <- as.Date(EPA_PM25_17$Date, format = "%m/%d/%y")
EPA_PM25_18$Date <- as.Date(EPA_PM25_18$Date, format = "%m/%d/%y")

#4
EPA_O3_17.select <- select(EPA_O3_17, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE
EPA_O3_18.select <- select(EPA_O3_18, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE
EPA_PM25_17.select <- select(EPA_PM25_17, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY,
EPA_PM25_18.select <- select(EPA_PM25_18, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY,

#5
EPA_PM25_17.select$AQS_PARAMETER_DESC <- "PM2.5"
EPA_PM25_18.select$AQS_PARAMETER_DESC <- "PM2.5"

#6
write.csv(EPA_O3_17.select, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2017_Processed.csv")
write.csv(EPA_O3_18.select, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2018_Processed.csv")
write.csv(EPA_PM25_17.select, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2017_Processed.
write.csv(EPA_PM25_18.select, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2018_Processed.
```

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
EPAair_combined <- rbind(EPA_O3_17.select, EPA_O3_18.select, EPA_PM25_17.select, EPA_PM25_18.select)

#8
EPAair_combined.filter <-
  EPAair_combined %>%
  filter(Site.Name == "Blackstone" | Site.Name == "Bryson City" | Site.Name == "Triple Oak") %>%
  separate(Date, c("Y", "m"), remove = FALSE)
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 2986 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
#9
EPAair_combined.tidy <- spread(EPAair_combined.filter, AQS_PARAMETER_DESC, DAILY_AQI_VALUE)

#10
dim(EPAair_combined.tidy)
```

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

```
#11
write.csv(EPAair_combined.tidy, row.names = FALSE, file = "./Data/Processed/EPAair_O3_PM25_NC1718_Proces
```

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
EPAair_combined.tidy.summaryA <-
  EPAair_combined.tidy %>%
  group_by(m) %>%
  filter(!is.na(Ozone) & !is.na(PM2.5)) %>%
  summarise(meanO3 = mean(Ozone),
            meanPM25 = mean(PM2.5))

#12b
EPAair_combined.tidy.summaryB <-
```

```
  EPAair_combined.tidy %>%
  group_by(Site.Name) %>%
  filter(!is.na(Ozone) & !is.na(PM2.5)) %>%
  summarise(meanO3 = mean(Ozone),
            minO3 = min(Ozone),
            maxO3 = max(Ozone),
            meanPM25 = mean(PM2.5),
            minPM25 = min(PM2.5),
            maxPM25 = max(PM2.5))

#13
print(EPAair_combined.tidy.summaryA)
```

```
## # A tibble: 12 x 3
##    m     meanO3 meanPM25
##    <chr>  <dbl>    <dbl>
##  1 01      31.5     34.2
##  2 02      35.4     37.6
##  3 03      42.4     37.4
##  4 04      43.5     31.5
##  5 05      39.5     30.6
##  6 06      39.2     30.9
##  7 07      38.3     31.9
##  8 08      34.4     32.3
##  9 09      32.6     30.7
## 10 10      32.3     30.1
## 11 11      30.1     42.1
## 12 12      29.8     46.6
```

```
print(EPAair_combined.tidy.summaryB)
```

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