Assignment 4: Data Wrangling

Laurie Muzzy

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/laurie/Desktop/Envtl_Data_Analytics/MuzzyGitFile"
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.4.2
```

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

```
## v ggplot2 3.1.0     v purrr   0.3.0
## 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
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
## Warning: package 'tibble' was built under R version 3.4.4
```

```
## Warning: package 'tidyr' was built under R version 3.4.4
```

```
## Warning: package 'readr' was built under R version 3.4.4
```

```
## Warning: package 'purrr' was built under R version 3.4.4
```

```
## Warning: package 'dplyr' was built under R version 3.4.4

## Warning: package 'stringr' was built under R version 3.4.4

## Warning: package 'forcats' was built under R version 3.4.3

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

```
EPA.NCAir.O3.2017 <- read.csv("./Data/Raw/EPAair_O3_NC2017_raw.csv")
EPA.NCAir.O3.2018 <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv")
EPA.NCAir.PM25.2017 <- read.csv("./Data/Raw/EPAair_PM25_NC2017_raw.csv")
EPA.NCAir.PM25.2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")

#2
dim(EPA.NCAir.O3.2017)
```

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

```
summary(EPA.NCAir.O3.2017)
```

```
##       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
##
```

```
head(EPA.NCAir.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
```

```
str(EPA.NCAir.O3.2017)
```

```
## 'data.frame':    10219 obs. of  20 variables:
##  $ Date                                : Factor w/ 364 levels "1/1/17","1/10/17",..: 151 162 173 176
##  $ Source                              : Factor w/ 1 level "AQS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Site.ID                             : int  370030005 370030005 370030005 370030005 370030005 3700
##  $ POC                                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Daily.Max.8.hour.Ozone.Concentration: num  0.041 0.046 0.046 0.046 0.046 0.048 0.047 0.053 0.056
##  $ UNITS                               : Factor w/ 1 level "ppm": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DAILY_AQI_VALUE                     : int  38 43 43 43 43 44 44 49 54 44 ...
##  $ Site.Name                           : Factor w/ 40 levels "","Beaufort",..: 35 35 35 35 35 35 35
##  $ DAILY_OBS_COUNT                     : int  17 17 17 17 17 17 17 17 17 17 ...
##  $ PERCENT_COMPLETE                    : int  100 100 100 100 100 100 100 100 100 100 ...
##  $ AQS_PARAMETER_CODE                  : int  44201 44201 44201 44201 44201 44201 44201 44201 44201
##  $ AQS_PARAMETER_DESC                  : Factor w/ 1 level "Ozone": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CBSA_CODE                           : int  25860 25860 25860 25860 25860 25860 25860 25860 25860
##  $ CBSA_NAME                           : Factor w/ 17 levels "","Asheville, NC",..: 9 9 9 9 9 9 9 9
##  $ 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  3 3 3 3 3 3 3 3 3 3 ...
##  $ COUNTY                              : Factor w/ 32 levels "Alexander","Avery",..: 1 1 1 1 1 1 1 1
##  $ 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 ...
```

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 changing date from factor to date (select function doesn't allow factors)
class(EPA.NCAir.O3.2017$Date)
```

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

```
 #change factor to Date

EPA.NCAir.O3.2017$Date <- as.Date(EPA.NCAir.O3.2017$Date, format = "%m/%d/%y")
```

```
## Warning in strptime(x, format, tz = "GMT"): unknown timezone 'default/
## America/New_York'
```

```
EPA.NCAir.O3.2018$Date <- as.Date(EPA.NCAir.O3.2018$Date, format = "%m/%d/%y")
EPA.NCAir.PM25.2017$Date <- as.Date(EPA.NCAir.PM25.2017$Date, format = "%m/%d/%y")
EPA.NCAir.PM25.2018$Date <- as.Date(EPA.NCAir.PM25.2018$Date, format = "%m/%d/%y")

#4
EPA.NCAir.O3.2017.sitespecific <- select(EPA.NCAir.O3.2017, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAM
```

```
EPA.NCAir.O3.2018.sitespecific <- select(EPA.NCAir.O3.2018, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAM

EPA.NCAir.PM25.2017.sitespecific <- select(EPA.NCAir.PM25.2017, Date, DAILY_AQI_VALUE, Site.Name, AQS_P

EPA.NCAir.PM25.2018.sitespecific <- select(EPA.NCAir.PM25.2018, Date, DAILY_AQI_VALUE, Site.Name, AQS_P

#5

EPA.NCAir.PM25.2017.sitespecific$AQS_PARAMETER_DESC <- "PM2.5"
EPA.NCAir.PM25.2018.sitespecific$AQS_PARAMETER_DESC <- "PM2.5"

#6


write.csv(EPA.NCAir.O3.2017.sitespecific, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2017_

write.csv(EPA.NCAir.O3.2018.sitespecific, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2018_

write.csv(EPA.NCAir.PM25.2017.sitespecific, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2

write.csv(EPA.NCAir.PM25.2018.sitespecific, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2
```

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"

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.4.4
```

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

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

```
#7
EPA.NCAir.O3.PM25.2017to18 <- rbind(EPA.NCAir.O3.2017.sitespecific, EPA.NCAir.O3.2018.sitespecific, EPA
dim(EPA.NCAir.O3.PM25.2017to18) #38105 rows, 7 columns
```

```
## [1] 38105     7
```

```
#8
```

```
EPA.NCAir.O3.PM25.2017to18.B.BC.TO <- EPA.NCAir.O3.PM25.2017to18 %>%
filter(Site.Name %in% c("Blackstone", "Bryson City", "Triple Oak")) %>%
mutate(month = month(Date), year = year(Date))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
 #added month and year, while keeping Date; this is why lubridate is cool

dim(EPA.NCAir.O3.PM25.2017to18.B.BC.TO) #[1] 2986    9
```

```
## [1] 2986    9
```

```
#9
EPA.NCAir.201718.tidy <- spread(EPA.NCAir.O3.PM25.2017to18.B.BC.TO, AQS_PARAMETER_DESC, DAILY_AQI_VALUE

#10

dim(EPA.NCAir.201718.tidy) #[1] 1953    8
```

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

```
#11

write.csv(EPA.NCAir.201718.tidy, row.names = FALSE, file = "./Data/Processed/EPA_NCAir_2017_18_Processe
```

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 summary table of mean AQI for O3 and PM2.5 #group by month and find mean; summarise mean AQI

EPA.NCAir.201718.AQI <-
  EPA.NCAir.201718.tidy %>%
  group_by(month) %>%
  summarise(meanPM2.5 = mean(PM2.5, na.rm = TRUE),
            meanO3 = mean(Ozone, na.rm = TRUE))

#12b
EPA.NCAir.201718.Sites <-
  EPA.NCAir.201718.tidy %>%
  group_by(Site.Name) %>%
  summarise(minPM2.5 = min(PM2.5, na.rm = TRUE),
            meanPM2.5 = mean(PM2.5, na.rm = TRUE),
            maxPM2.5 = max(PM2.5, na.rm = TRUE),
            minO3 = min(Ozone, na.rm = TRUE),
            meanO3 = mean(Ozone, na.rm = TRUE),
            maxO3 = max(Ozone, na.rm = TRUE))

#13 data frames

print(EPA.NCAir.201718.Sites)
```

```
## # A tibble: 3 x 7
##   Site.Name   minPM2.5 meanPM2.5 maxPM2.5 minO3 meanO3 maxO3
##   <fct>          <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl>
## 1 Blackstone         0      36.7       83     8   38.5    97
## 2 Bryson City        3      32.3       78     5   35.2    71
## 3 Triple Oak         0      33.5       74   Inf  NaN    -Inf
```

```
print(EPA.NCAir.201718.AQI)
```

```
## # A tibble: 12 x 3
##    month meanPM2.5 meanO3
##    <dbl>     <dbl>  <dbl>
##  1     1      34.6   31.5
##  2     2      36.7   35.5
##  3     3      35.1   42.4
##  4     4      32.5   44.3
##  5     5      31.7   38.9
##  6     6      33.3   38.7
##  7     7      33.1   38.2
##  8     8      33.7   34.0
##  9     9      31.9   32.6
## 10    10      29.3   32.1
## 11    11      36.8   30.1
## 12    12      41.1   29.8
```