Assignment 4: Data Wrangling

Xin Wang

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] "Y:/19spring/872/Environmental_Data_Analytics/Assignments"
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

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

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     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 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

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

```
o3_17 <- read.csv("../Data/Raw/EPAair_O3_NC2017_raw.csv")
o3_18 <- read.csv("../Data/Raw/EPAair_O3_NC2018_raw.csv")
pm25_17 <- read.csv("../Data/Raw/EPAair_PM25_NC2017_raw.csv")
pm25_18 <- read.csv("../Data/Raw/EPAair_PM25_NC2018_raw.csv")
#2
head(o3_18)
```

```
##      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
```

```
head(pm25_18)
```

```
##      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
```

```
summary(o3_17$DAILY_AQI_VALUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   32.00   40.00   39.87   45.00  115.00
```

```
summary(pm25_18$Daily.Mean.PM2.5.Concentration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -2.800   5.000   7.200   7.554   9.800  34.200
```

```
summary(o3_18$Site.Name)
```

```
##                                  Beaufort             Bent Creek
##                       155               223                    280
##            Bethany sch.        Blackstone           Bryson City
##                       332               215                    292
##              Bushy Fork            Butner                Candor
##                       275               287                    337
##            Castle Hayne      Cherry Grove       Clemmons Middle
##                       241               255                    254
##                 Coweeta         Cranberry                Crouse
##                       340               319                    265
##           Durham Armory  Frying Pan Mountain  Garinger High School
```

```
##                   291                   311                   333
##          Hattie Avenue       Honeycutt School      Jamesville School
##                   251                   232                   239
##            Joanna Bald               Leggett          Lenoir (city)
##                   309                   253                   287
## Lenoir Co. Comm. Coll.         Linville Falls      Mendenhall School
##                   255                   294                   263
##       Millbrook School          Monroe School           Mt. Mitchell
##                   338                   254                   262
##      Pitt Agri. Center          Purchase Knob              Rockwell
##                   287                   311                   318
##  Taylorsville Liledoun            Union Cross     University Meadows
##                   285                   249                   299
##                  Wade     Waynesville School      West Johnston Co.
##                   235                   257                   298
```

```
class(pm25_17$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
o3_17$Date <- as.Date(o3_17$Date,format="%m/%d/%y")
o3_18$Date <- as.Date(o3_18$Date,format="%m/%d/%y")
pm25_17$Date <- as.Date(pm25_17$Date,format="%m/%d/%y")
pm25_18$Date <- as.Date(pm25_18$Date,format="%m/%d/%y")
#4
o317skinny <- select(o3_17,Date,DAILY_AQI_VALUE,Site.Name,AQS_PARAMETER_DESC,COUNTY,SITE_LATITUDE,SITE_
o318skinny <- select(o3_18,Date,DAILY_AQI_VALUE,Site.Name,AQS_PARAMETER_DESC,COUNTY,SITE_LATITUDE,SITE_
pm2517skinny <- select(pm25_17,Date,DAILY_AQI_VALUE,Site.Name,AQS_PARAMETER_DESC,COUNTY,SITE_LATITUDE,S
pm2518skinny <- select(pm25_18,Date,DAILY_AQI_VALUE,Site.Name,AQS_PARAMETER_DESC,COUNTY,SITE_LATITUDE,S
#5
pm2517skinny$AQS_PARAMETER_DESC <- "PM2.5"
pm2518skinny$AQS_PARAMETER_DESC <- "PM2.5"
#6
write.csv(o317skinny,file = "../Data/Processed/EPAair_O3_NC2017_processed.csv",row.names = F)
write.csv(o318skinny,file = "../Data/Processed/EPAair_O3_NC2018_processed.csv",row.names = F)
write.csv(pm2517skinny,file = "../Data/Processed/EPAair_PM25_NC2017_processed.csv",row.names = F)
write.csv(pm2518skinny,file = "../Data/Processed/EPAair_PM25_NC2018_processed.csv",row.names = F)
```

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"

```
#8
airdat <- rbind(o317skinny,o318skinny,pm2517skinny,pm2518skinny)
#9
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.5.2
```

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

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

```
airdat.processed <-
  airdat %>%
  filter(Site.Name %in% c("Blackstone","Bryson City","Triple Oak")) %>%
  mutate(Month = month(Date)) %>%
  mutate(Year = year(Date))
```

```
## Warning: package 'bindrcpp' was built under R version 3.5.2
```

```
#10
airdat.processed2 <- spread(airdat.processed,AQS_PARAMETER_DESC,DAILY_AQI_VALUE)
#11
write.csv(airdat.processed2,file = "../Data/Processed/EPAair_O3_PM25_NC1718_Processed.csv",row.names = F
```

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
airdat.sum1 <-
  airdat.processed2 %>%
  group_by(Month) %>%
  summarise(meanO3AQI = mean(Ozone,na.rm = T), meanPM25AQI = mean(PM2.5,na.rm = T))
#12b
airdat.sum2 <-
  airdat.processed2 %>%
  group_by(Site.Name) %>%
  summarise(meanO3AQI=mean(Ozone,na.rm = T),minO3AQI=min(Ozone,na.rm = T),maxO3AQI=max(Ozone,na.rm = T)
            meanPM25AQI=mean(PM2.5,na.rm = T),minPM25AQI=min(PM2.5,na.rm = T),maxPM25AQI=max(PM2.5,na.r
```

```
## Warning in min(Ozone, na.rm = T): no non-missing arguments to min;
## returning Inf
```

```
## Warning in max(Ozone, na.rm = T): no non-missing arguments to max;
```

```
## returning -Inf
```

```
airdat.sum2[airdat.sum2 == Inf | airdat.sum2 == -Inf] <- NA #No O3 data in Triple Oak. Inf generated fo
#13
print(airdat.sum1)
```

```
## # A tibble: 12 x 3
##    Month meanO3AQI meanPM25AQI
##    <dbl>     <dbl>       <dbl>
##  1     1      31.5        34.6
##  2     2      35.5        36.7
##  3     3      42.4        35.1
##  4     4      44.3        32.5
##  5     5      38.9        31.7
##  6     6      38.7        33.3
##  7     7      38.2        33.1
##  8     8      34.0        33.7
##  9     9      32.6        31.9
## 10    10      32.1        29.3
## 11    11      30.1        36.8
## 12    12      29.8        41.1
```

```
print(airdat.sum2)
```

```
## # A tibble: 3 x 7
##   Site.Name   meanO3AQI minO3AQI maxO3AQI meanPM25AQI minPM25AQI maxPM25AQI
##   <fct>           <dbl>    <dbl>    <dbl>       <dbl>      <int>      <int>
## 1 Blackstone       38.5        8       97        36.7          0         83
## 2 Bryson City      35.2        5       71        32.3          3         78
## 3 Triple Oak      NaN         NA       NA        33.5          0         74
```