

Assignment 4: Data Wrangling

Jake Greif

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/jakegreif/Environmental_Data_Analytics/Assignments"
```

```
library(tidyverse)
```

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

```
## v ggplot2 3.1.0      v purrr  0.2.5  
## v tibble  1.4.2      v dplyr  0.7.8  
## v tidyr   0.8.1      v stringr 1.3.1  
## v readr   1.1.1      v forcats 0.3.0
```

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

```
EPAair_03_17 <- read.csv("../Data/Raw/EPAair_03_NC2017_raw.csv")  
EPAair_03_18 <- read.csv("../Data/Raw/EPAair_03_NC2018_raw.csv")  
EPAair_PM_17 <- read.csv("../Data/Raw/EPAair_PM25_NC2017_raw.csv")  
EPAair_PM_18 <- read.csv("../Data/Raw/EPAair_PM25_NC2018_raw.csv")
```

#2

```
colnames(EPAair_03_17)
```

```
## [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"
```

```
summary(EPAair_03_17)
```

```
##      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
##
```

```
dim(EPAair_03_17)
```

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

```
class(EPAair_03_17$Date)
```

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

```
colnames(EPAair_03_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] "AQ5_PARAMETER_CODE"
## [12] "AQ5_PARAMETER_DESC"
## [13] "CBSA_CODE"
## [14] "CBSA_NAME"
## [15] "STATE_CODE"
## [16] "STATE"
## [17] "COUNTY_CODE"
```

```
## [18] "COUNTY"
## [19] "SITE_LATITUDE"
## [20] "SITE_LONGITUDE"
```

```
summary(EPAair_03_18)
```

```
##      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
##
```

```
dim(EPAair_03_18)
```

```
## [1] 10781 20
```

```
class(EPAair_03_18$Date)
```

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

```
colnames(EPAair_PM_17)
```

```
## [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"
```

```
summary(EPAair_PM_17)
```

```
##      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
## Montclair 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
##
```

```
dim(EPAair_PM_17)
```

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

```
class(EPAair_PM_17$Date)
```

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

```
colnames(EPAair_PM_18)
```

```
## [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"
```

```
summary(EPAair_PM_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
##

```

```
dim(EPAair_PM_18)
```

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

```
class(EPAair_PM_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
```

```
EPAair_03_17$Date <- as.Date(EPAair_03_17$Date, format = "%m/%d/%y")
```

```
EPAair_03_18$Date <- as.Date(EPAair_03_18$Date, format = "%m/%d/%y")
```

```
EPAair_PM_17$Date <- as.Date(EPAair_PM_17$Date, format = "%m/%d/%y")
```

```
EPAair_PM_18$Date <- as.Date(EPAair_PM_18$Date, format = "%m/%d/%y")
```

```
class(EPAair_03_17$Date)
```

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

```
class(EPAair_03_18$Date)
```

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

```
class(EPAair_PM_17$Date)
```

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

```
class(EPAair_PM_18$Date)
```

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

```
#4
```

```
EPAair_03_17.skinny <- select(EPAair_03_17, Date, DAILY_AQI_VALUE, Site.Name,  
                             AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
```

```
EPAair_03_18.skinny <- select(EPAair_03_18, Date, DAILY_AQI_VALUE, Site.Name,  
                             AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
```

```
EPAair_PM_17.skinny <- select(EPAair_PM_17, Date, DAILY_AQI_VALUE, Site.Name,  
                             AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
```

```
EPAair_PM_18.skinny <- select(EPAair_PM_18, Date, DAILY_AQI_VALUE, Site.Name,  
                             AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
```

```
#5
```

```
EPAair_PM_17.skinny <- mutate(EPAair_PM_17.skinny, AQS_PARAMETER_DESC = "PM2.5")
```

```
EPAair_PM_18.skinny <- mutate(EPAair_PM_18.skinny, AQS_PARAMETER_DESC = "PM2.5")
```

```
#6
```

```
write.csv(EPAair_03_17.skinny, row.names = FALSE,  
          file = "../Data/Processed/EPAair_03_NC2017_Processed.csv")
```

```
write.csv(EPAair_03_18.skinny, row.names = FALSE,
```



```

    file = "../Data/Processed/EPAair_O3_NC2018_Processed.csv")
write.csv(EPAair_PM_17.skinny, row.names = FALSE,
    file = "../Data/Processed/EPAair_PM25_NC2017_Processed.csv")
write.csv(EPAair_PM_18.skinny, row.names = FALSE,
    file = "../Data/Processed/EPAair_PM25_NC2018_Processed.csv")

```

Combine datasets

- Combine the four datasets with `rbind`. Make sure your column names are identical prior to running this code.
- 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)
- 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.
- Call up the dimensions of your new tidy dataset.
- Save your processed dataset with the following file name: “EPAair_O3_PM25_NC1718_Processed.csv”

```

#7
EPAair.Combined <- rbind(EPAair_O3_17.skinny, EPAair_O3_18.skinny,
    EPAair_PM_17.skinny, EPAair_PM_18.skinny)

#8
library(lubridate)

##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
##
##     date
EPAair.summary <- EPAair.Combined %>%
    filter(Site.Name == "Blackstone" | Site.Name == "Bryson City" | Site.Name == "Triple Oak") %>%
    mutate(Month = month(Date), Year = year(Date))

dim(EPAair.summary)

## [1] 2986    9

#9
EPAair.summary.spread <- spread(EPAair.summary, AQS_PARAMETER_DESC, DAILY_AQI_VALUE)

#10
dim(EPAair.summary.spread)

## [1] 1953    9

#11
write.csv(EPAair.summary.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

```
Mean.AQI.summary <-  
  EPAair.summary.spread %>%  
  group_by(Month) %>%  
  summarise(meanO3 = mean(Ozone, na.rm = TRUE),  
            meanPM = mean(PM2.5, na.rm = TRUE))
```

#12b

```
AQI.summary <-  
  EPAair.summary.spread %>%  
  group_by(Site.Name) %>%  
  summarise(meanO3 = mean(Ozone, na.rm = TRUE),  
            minO3 = min(Ozone, na.rm = TRUE),  
            maxO3 = max(Ozone, na.rm = TRUE),  
            meanPM = mean(PM2.5, na.rm = TRUE),  
            minPM = min(PM2.5, na.rm = TRUE),  
            maxPM = max(PM2.5, na.rm = TRUE))
```

#13

```
View(Mean.AQI.summary)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):  
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/  
## Resources/modules/R_de.so'' had status 1
```

```
View(AQI.summary)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):  
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/  
## Resources/modules/R_de.so'' had status 1
```