Assignment 4: Data Wrangling

Njeri Kara

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
#Setting the working directory
setwd("C:/Users/jerik/OneDrive/Documents/Spring 2019 Semenster/Environmental Data Analytics/EDA_R_Work/
#Confirming that it is the correct working directory
getwd()
```

```
## [1] "C:/Users/jerik/OneDrive/Documents/Spring 2019 Semenster/Environmental Data Analytics/EDA_R_Work
```

```
#Loading necessary packages
library(tidyverse)
```

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

## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.2     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()
```

```
library(lubridate)
```

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

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

```
library(knitr)

#Uploading the four raw datafiles associated with EPA Air dataset.
NC.03.2017.raw.data <- read.csv("./Data/Raw/EPAair_O3_NC2017_raw.csv")
NC.03.2018.raw.data <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv")
NC.PM25.2017.raw.data <- read.csv("./Data/Raw/EPAair_PM25_NC2017_raw.csv")
NC.PM25.2018.raw.data <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")

#2
#Getting to know NC.03.2017 data
dim(NC.03.2017.raw.data) #shows number of rows and columns in the dataset
```

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

```
str(NC.03.2017.raw.data) #shows the names and class of each variable and a sample of its values
```

```
## '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 37003
##  $ 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 0
##  $ 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 3
##  $ 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 4
##  $ 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 2
##  $ CBSA_NAME                           : Factor w/ 17 levels "","Asheville, NC",..: 9 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 ...
```

```
head(NC.03.2017.raw.data) #shows the first six observations in the dataset
```

```
##     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(NC.03.2017.raw.data$Daily.Max.8.hour.Ozone.Concentration) #summary stats of 03 concentration
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.00500 0.03500 0.04300 0.04211 0.04900 0.07500
```

```
#Getting to know NC.03.2018 data
dim(NC.03.2018.raw.data) #shows number of rows and columns in the dataset
```

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

```
str(NC.03.2018.raw.data) #shows the names and class of each variable and a sample of its values
```

```
## 'data.frame':    10781 obs. of  20 variables:
##  $ Date                                : Factor w/ 343 levels "1/1/18","1/10/18",..: 109 110 111 112
##  $ Source                              : Factor w/ 2 levels "AirNow","AQS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Site.ID                             : int  370030005 370030005 370030005 370030005 370030005 37003
##  $ POC                                 : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Daily.Max.8.hour.Ozone.Concentration: num  0.038 0.033 0.04 0.02 0.019 0.021 0.031 0.022 0.038 0.0
##  $ UNITS                               : Factor w/ 1 level "ppm": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DAILY_AQI_VALUE                     : int  35 31 37 19 18 19 29 20 35 29 ...
##  $ Site.Name                           : Factor w/ 39 levels "","Beaufort",..: 34 34 34 34 34 34 34 3
##  $ DAILY_OBS_COUNT                     : int  24 24 24 24 24 24 24 24 24 24 ...
##  $ 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 4
##  $ 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 2
##  $ CBSA_NAME                           : Factor w/ 16 levels "","Asheville, NC",..: 8 8 8 8 8 8 8 8 8
##  $ 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/ 31 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 ...
```

```
head(NC.03.2018.raw.data) #shows the first six observations in the dataset
```

```
##      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
```

```
#summary stats of daily 03 concentration
summary(NC.03.2018.raw.data$Daily.Max.8.hour.Ozone.Concentration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.03400 0.04100 0.04124 0.04900 0.07700
```

```
#Getting to know NC.PM25.2017 data
dim(NC.PM25.2017.raw.data) #shows number of rows and columns in the dataset
```

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

```
str(NC.PM25.2017.raw.data) #shows the names and class of each variable and a sample of its values
```

```
## '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 ...
```

```
head(NC.PM25.2017.raw.data) #shows the first six observations in the dataset
```

```
##      Date Source   Site.ID POC Daily.Mean.PM2.5.Concentration    UNITS
## 1  1/1/17    AQS 370110002   1                            2.9 ug/m3 LC
## 2  1/4/17    AQS 370110002   1                            1.2 ug/m3 LC
## 3  1/7/17    AQS 370110002   1                            3.2 ug/m3 LC
## 4 1/10/17    AQS 370110002   1                            6.4 ug/m3 LC
## 5 1/13/17    AQS 370110002   1                            3.6 ug/m3 LC
## 6 1/16/17    AQS 370110002   1                            5.8 ug/m3 LC
##   DAILY_AQI_VALUE      Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              12 Linville Falls               1              100
## 2               5 Linville Falls               1              100
## 3              13 Linville Falls               1              100
## 4              27 Linville Falls               1              100
## 5              15 Linville Falls               1              100
## 6              24 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 stats of daily PM25 concentation
summary(NC.PM25.2017.raw.data$Daily.Mean.PM2.5.Concentration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -3.900   5.000   7.300   7.742  10.000  31.900
```

```
#Getting to know NC.PM25.2018 data
dim(NC.PM25.2018.raw.data) #shows number of rows and columns in the dataset
```

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

```
str(NC.PM25.2018.raw.data) #shows the names and class of each variable and a sample of its values
```

```
## 'data.frame':    7611 obs. of  20 variables:
##  $ Date                          : Factor w/ 343 levels "1/1/18","1/10/18",..: 12 27 30 3 6 9 13 16 1
##  $ Source                        : Factor w/ 2 levels "AirNow","AQS": 2 2 2 2 2 2 2 2 2 2 ...
##  $ 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 3.7 5.3 0.8 2.5 4.5 1.8 2.5 4.2 1.7 ...
##  $ UNITS                         : Factor w/ 1 level "ug/m3 LC": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DAILY_AQI_VALUE               : int  12 15 22 3 10 19 8 10 18 7 ...
##  $ Site.Name                     : Factor w/ 24 levels "","Blackstone",..: 14 14 14 14 14 14 14 14 14
##  $ 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 ...
```

```
head(NC.PM25.2018.raw.data) #shows the first six observations in the dataset
```

```
##      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
```

```r
#summary stats of daily PM25 concentation
summary(NC.PM25.2018.raw.data$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
```

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.

```r
#3
#Changing date variable of NC.03.2017 data to date format
NC.03.2017.raw.data$Date <- as.Date(NC.03.2017.raw.data$Date, format = "%m/%d/%y")
class(NC.03.2017.raw.data$Date) #confirming date change
```

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

```r
#Changing date variable of NC.03.2018 data to date format
NC.03.2018.raw.data$Date <- as.Date(NC.03.2018.raw.data$Date, format = "%m/%d/%y")
class(NC.03.2018.raw.data$Date) #confirming date change
```

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

```r
#Changing date variable of NC.PM25.2017 data to date format
NC.PM25.2017.raw.data$Date <- as.Date(NC.PM25.2017.raw.data$Date, format = "%m/%d/%y")
class(NC.PM25.2017.raw.data$Date) #confirming date change
```

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

```
#Changing date variable of NC.PM25.2018 data to date format
NC.PM25.2018.raw.data$Date <- as.Date(NC.PM25.2018.raw.data$Date, format = "%m/%d/%y")
class(NC.PM25.2018.raw.data$Date) #confirming date change
```

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

```
#4
#selecting specific columns in the NC.03.2017 data
NC.03.2017.proccessed.v1 <- select(NC.03.2017.raw.data, "Date", "DAILY_AQI_VALUE",
                                   "Site.Name", "AQS_PARAMETER_DESC",
                                   "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE")

#selecting specific columns in the NC.03.2018 data
NC.03.2018.proccessed.v1 <- select(NC.03.2018.raw.data, "Date", "DAILY_AQI_VALUE",
                                   "Site.Name", "AQS_PARAMETER_DESC",
                                   "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE")

#selecting specific columns in the NC.PM25.2017 data
NC.PM25.2017.proccessed.v1 <- select(NC.PM25.2017.raw.data, "Date", "DAILY_AQI_VALUE",
                                     "Site.Name", "AQS_PARAMETER_DESC",
                                     "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE")

#selecting specific columns in the NC.PM25.2018 data
NC.PM25.2018.proccessed.v1 <- select(NC.PM25.2018.raw.data, "Date", "DAILY_AQI_VALUE",
                                     "Site.Name", "AQS_PARAMETER_DESC", "COUNTY",
                                     "SITE_LATITUDE", "SITE_LONGITUDE")

#5
#filling all cells in dataset NC.PM25.2017.proccessed.v1, variable AQS_PARAMETER_DESC with "PM2.5"
NC.PM25.2017.proccessed.v2 <- mutate(NC.PM25.2017.proccessed.v1,AQS_PARAMETER_DESC = "PM2.5")

#filling all cells in dataset NC.PM25.2018.proccessed.v1, variable AQS_PARAMETER_DESC with "PM2.5"
NC.PM25.2018.proccessed.v2 <- mutate(NC.PM25.2018.proccessed.v1,AQS_PARAMETER_DESC = "PM2.5")

#6
#Saving NC.03.2017.proccessed.v1 in processed data folder
write.csv(NC.03.2017.proccessed.v1, row.names = FALSE, file = "./Data/Processed/NC.03.2017.proccessed.v

#Saving NC.03.2018.proccessed.v1 in processed data folder
write.csv(NC.03.2018.proccessed.v1, row.names = FALSE, file = "./Data/Processed/NC.03.2018.proccessed.v

#Saving NC.PM25.2017.proccessed.v2 in processed data folder
write.csv(NC.PM25.2017.proccessed.v2, row.names = FALSE, file = "./Data/Processed/NC.PM25.2017.proccess

#Saving NC.PM25.2018.proccessed.v2 in processed data folder
write.csv(NC.PM25.2018.proccessed.v2, row.names = FALSE, file = "./Data/Processed/NC.PM25.2018.proccess
```

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
#Ensuring all column names are identical
colnames(NC.03.2017.proccessed.v1)
```

```
## [1] "Date"              "DAILY_AQI_VALUE"    "Site.Name"
## [4] "AQS_PARAMETER_DESC" "COUNTY"             "SITE_LATITUDE"
## [7] "SITE_LONGITUDE"
```

```
colnames(NC.03.2018.proccessed.v1)
```

```
## [1] "Date"              "DAILY_AQI_VALUE"    "Site.Name"
## [4] "AQS_PARAMETER_DESC" "COUNTY"             "SITE_LATITUDE"
## [7] "SITE_LONGITUDE"
```

```
colnames(NC.PM25.2017.proccessed.v2)
```

```
## [1] "Date"              "DAILY_AQI_VALUE"    "Site.Name"
## [4] "AQS_PARAMETER_DESC" "COUNTY"             "SITE_LATITUDE"
## [7] "SITE_LONGITUDE"
```

```
colnames(NC.PM25.2018.proccessed.v2)
```

```
## [1] "Date"              "DAILY_AQI_VALUE"    "Site.Name"
## [4] "AQS_PARAMETER_DESC" "COUNTY"             "SITE_LATITUDE"
## [7] "SITE_LONGITUDE"
```

```
#Combining all datasets using rbind
NC.03.PM25.2017.2018.data <- rbind(NC.03.2017.proccessed.v1,NC.03.2018.proccessed.v1,NC.PM25.2017.procc

#8 #Wrangling dataset
#displaying the different factor levels of Site.name
levels(NC.03.PM25.2017.2018.data$Site.Name)
```

```
##  [1] ""
##  [2] "Beaufort"
##  [3] "Bent Creek"
##  [4] "Bethany sch."
##  [5] "Blackstone"
##  [6] "Bryson City"
##  [7] "Bushy Fork"
##  [8] "Butner"
##  [9] "Candor"
## [10] "Castle Hayne"
## [11] "Cherry Grove"
## [12] "Clemmons Middle"
## [13] "Coweeta"
## [14] "Cranberry"
## [15] "Crouse"
## [16] "Durham Armory"
## [17] "Frying Pan Mountain"
## [18] "Garinger High School"
## [19] "Hattie Avenue"
## [20] "Honeycutt School"
## [21] "Jamesville School"
## [22] "Joanna Bald"
## [23] "Leggett"
## [24] "Lenoir (city)"
## [25] "Lenoir Co. Comm. Coll."
## [26] "Linville Falls"
## [27] "Mendenhall School"
## [28] "Millbrook School"
## [29] "Monroe School"
## [30] "Mt. Mitchell"
## [31] "OZONE MONITOR ON SW SIDE OF TOWER/MET EQUIPMENT 10FT ABOVE TOWER"
## [32] "Pitt Agri. Center"
## [33] "Purchase Knob"
## [34] "Rockwell"
## [35] "Taylorsville Liledoun"
## [36] "Union Cross"
## [37] "University Meadows"
## [38] "Wade"
## [39] "Waynesville School"
## [40] "West Johnston Co."
## [41] "Board Of Ed. Bldg."
## [42] "Candor: EPA CASTNet Site"
## [43] "Hickory Water Tower"
## [44] "Lexington water tower"
## [45] "Montclaire Elementary School"
## [46] "PM2.5 COLOCATED MONITORS LOCATED ON TOP OF BUILDING"
## [47] "Remount"
## [48] "Spruce Pine Hospital"
## [49] "Triple Oak"
## [50] "William Owen School"
```

```
NC.03.PM25.2017.2018.data.v1 <- NC.03.PM25.2017.2018.data %>%
  #filtering out data from sites Blackstone, Bryson City, Triple Oak
  filter(Site.Name=="Blackstone"|Site.Name=="Bryson City"|Site.Name=="Triple Oak") %>%
  mutate(Month = month(Date)) %>% #including a month column
  mutate(Year = year(Date)) #including a year column



#9
#spreading dataset to include 2 columns for DAILY_AQI_VALUEs,broken down by AQS_PARAMETER_DESC factors
NC.03.PM25.2017.2018.data.v2 <- NC.03.PM25.2017.2018.data.v1 %>%
  spread(AQS_PARAMETER_DESC,DAILY_AQI_VALUE) %>%
  rename(Ozone_Daily_AQI=Ozone,PM2.5_Daily_AQI=PM2.5) #renaming columns to more descriptive data label

#10
#Dimensions of the new dataset
dim(NC.03.PM25.2017.2018.data.v2)
```

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

```
#11
#saving the dataset in the processed folder
write.csv(NC.O3.PM25.2017.2018.data.v2, row.names = FALSE, file = "./Data/Processed/EPAair_O3_PM25_NC17
```

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 values for O3 and PM2.5 by month
NC.O3.PM25.2017.2018.data.month.summ <-
  NC.O3.PM25.2017.2018.data.v2 %>%
  group_by(Month) %>%
  summarise(mean.AQI.O3 = mean(Ozone_Daily_AQI,na.rm=TRUE),
            mean.AQI.PM2.5 = mean(PM2.5_Daily_AQI,na.rm=TRUE))
            #na.rm=TRUE excludes NA values in the mean computation

#12b
#summary table of mean, minimum, and maximum AQI values of O3 and PM2.5 for each site
NC.O3.PM25.2017.2018.data.site.summ <-
  NC.O3.PM25.2017.2018.data.v2 %>%
  group_by(Site.Name) %>%
  summarise(mean.AQI.O3 = mean(Ozone_Daily_AQI,na.rm=TRUE),
            mean.AQI.PM2.5 = mean(PM2.5_Daily_AQI,na.rm=TRUE),
            min.AQI.O3 = min(Ozone_Daily_AQI,na.rm=TRUE),
            min.AQI.PM2.5 = min(PM2.5_Daily_AQI,na.rm=TRUE),
             max.AQI.O3 = max(Ozone_Daily_AQI,na.rm=TRUE),
            max.AQI.PM2.5 = max(PM2.5_Daily_AQI,na.rm=TRUE))
           #na.rm=TRUE excludes NA values in the mean computation

#13
#Displaying the summary table of mean AQI values for O3 and PM2.5 by month
kable(NC.O3.PM25.2017.2018.data.month.summ, caption = "Summary table of mean AQI values by month")
```

Table 1: Summary table of mean AQI values by month

Month	mean.AQI.O3	mean.AQI.PM2.5
1	31.48276	34.58192
2	35.52174	36.70659
3	42.40164	35.13978
4	44.30000	32.52147
5	38.90826	31.68333
6	38.71429	33.28743
7	38.16129	33.07609
8	33.95960	33.68667
9	32.59036	31.88889
10	32.12644	29.32639
11	30.06897	36.83333
12	29.78378	41.12150

Month	mean.AQI.O3	mean.AQI.PM2.5

```
#Displaying the summary table of mean, minimum, and maximum AQI values of O3 and PM2.5 for each site
kable(NC.03.PM25.2017.2018.data.site.summ, caption = "Summary table of mean,min and max AQI values by si
```

Table 2: Summary table of mean,min and max AQI values by site

Site.Name	mean.AQI.O3	mean.AQI.PM2.5	min.AQI.O3	min.AQI.PM2.5	max.AQI.O3	max.AQI.PM2.5
Blackstone	38.48246	36.72613	8	0	97	83
Bryson City	35.18252	32.29955	5	3	71	78
Triple Oak	NaN	33.48000	Inf	0	-Inf	74