Assignment 4: Data Wrangling

Jake Whisler

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Wrangling

Directions

1. Rename this file `<FirstLast>_A04_DataWrangling.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. Ensure that code in code chunks does not extend off the page in the PDF.

The completed exercise is due on Thursday, Sept 28th @ 5:00pm.

Set up your session

1a. Load the `tidyverse`, `lubridate`, and `here` packages into your session.

1b. Check your working directory.

1c. Read in all four raw data files associated with the EPA Air dataset, being sure to set string columns to be read in a factors. See the README file for the EPA air datasets for more information (especially if you have not worked with air quality data previously).

2. Apply the `glimpse()` function to reveal the dimensions, column names, and structure of each dataset.

```
#1a Loading the tidyverse, lubridate, and here packages
library(tidyverse)
library(lubridate)
library(here)

#1b Checking working directory, it shows that it's the EDE_Fall2023
setwd("/home/guest/EDE_Fall2023/")
getwd()
```

```
## [1] "/home/guest/EDE_Fall2023"
```

```r
#1c  Reading in the raw data files associated with EPA Air
pm25_nc2019 <- read.csv("./Data/Raw/EPAair_PM25_NC2019_raw.csv", stringsAsFactors = TRUE)
pm25_nc2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv", stringsAsFactors = TRUE)
o3_nc2019 <- read.csv("./Data/Raw/EPAair_O3_NC2019_raw.csv", stringsAsFactors = TRUE)
o3_nc2018 <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv", stringsAsFactors = TRUE)

#2 Using glimpse() to reveal the dimensions of each data set
glimpse(pm25_nc2019)
```

```
## Rows: 8,581
## Columns: 20
## $ Date                           <fct> 01/03/2019, 01/06/2019, 01/09/2019, 01/~
## $ Source                         <fct> AQS, AQS, AQS, AQS, AQS, AQS, AQS, AQS,~
## $ Site.ID                        <int> 370110002, 370110002, 370110002, 370110~
## $ POC                            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ Daily.Mean.PM2.5.Concentration <dbl> 1.6, 1.0, 1.3, 6.3, 2.6, 1.2, 1.5, 1.5,~
## $ UNITS                          <fct> ug/m3 LC, ug/m3 LC, ug/m3 LC, ug/m3 LC,~
## $ DAILY_AQI_VALUE                <int> 7, 4, 5, 26, 11, 5, 6, 6, 15, 7, 14, 20~
## $ Site.Name                      <fct> Linville Falls, Linville Falls, Linvill~
## $ DAILY_OBS_COUNT                <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ PERCENT_COMPLETE               <dbl> 100, 100, 100, 100, 100, 100, 100, 100,~
## $ AQS_PARAMETER_CODE             <int> 88502, 88502, 88502, 88502, 88502, 8850~
## $ AQS_PARAMETER_DESC             <fct> Acceptable PM2.5 AQI & Speciation Mass,~
## $ CBSA_CODE                      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ CBSA_NAME                      <fct> "", "", "", "", "", "", "", "", "", "",~
## $ STATE_CODE                     <int> 37, 37, 37, 37, 37, 37, 37, 37, 37, 37,~
## $ STATE                          <fct> North Carolina, North Carolina, North C~
## $ COUNTY_CODE                    <int> 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,~
## $ COUNTY                         <fct> Avery, Avery, Avery, Avery, Avery, Aver~
## $ SITE_LATITUDE                  <dbl> 35.97235, 35.97235, 35.97235, 35.97235,~
## $ SITE_LONGITUDE                 <dbl> -81.93307, -81.93307, -81.93307, -81.93~
```

```r
glimpse(pm25_nc2018)
```

```
## Rows: 8,983
## Columns: 20
## $ Date                           <fct> 01/02/2018, 01/05/2018, 01/08/2018, 01/~
## $ Source                         <fct> AQS, AQS, AQS, AQS, AQS, AQS, AQS, AQS,~
## $ Site.ID                        <int> 370110002, 370110002, 370110002, 370110~
## $ POC                            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ Daily.Mean.PM2.5.Concentration <dbl> 2.9, 3.7, 5.3, 0.8, 2.5, 4.5, 1.8, 2.5,~
## $ UNITS                          <fct> ug/m3 LC, ug/m3 LC, ug/m3 LC, ug/m3 LC,~
## $ DAILY_AQI_VALUE                <int> 12, 15, 22, 3, 10, 19, 8, 10, 18, 7, 24~
## $ Site.Name                      <fct> Linville Falls, Linville Falls, Linvill~
## $ DAILY_OBS_COUNT                <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ PERCENT_COMPLETE               <dbl> 100, 100, 100, 100, 100, 100, 100, 100,~
## $ AQS_PARAMETER_CODE             <int> 88502, 88502, 88502, 88502, 88502, 8850~
## $ AQS_PARAMETER_DESC             <fct> Acceptable PM2.5 AQI & Speciation Mass,~
## $ CBSA_CODE                      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ CBSA_NAME                      <fct> "", "", "", "", "", "", "", "", "", "",~
## $ STATE_CODE                     <int> 37, 37, 37, 37, 37, 37, 37, 37, 37, 37,~
## $ STATE                          <fct> North Carolina, North Carolina, North C~
```

```
## $ COUNTY_CODE                     <int> 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,~
## $ COUNTY                          <fct> Avery, Avery, Avery, Avery, Avery, Aver~
## $ SITE_LATITUDE                   <dbl> 35.97235, 35.97235, 35.97235, 35.97235,~
## $ SITE_LONGITUDE                  <dbl> -81.93307, -81.93307, -81.93307, -81.93~
```

```
glimpse(o3_nc2019)
```

```
## Rows: 10,592
## Columns: 20
## $ Date                                <fct> 01/01/2019, 01/02/2019, 01/03/201~
## $ Source                              <fct> AirNow, AirNow, AirNow, AirNow, A~
## $ Site.ID                             <int> 370030005, 370030005, 370030005, ~
## $ POC                                 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ Daily.Max.8.hour.Ozone.Concentration <dbl> 0.029, 0.018, 0.016, 0.022, 0.037~
## $ UNITS                               <fct> ppm, ppm, ppm, ppm, ppm, ppm, ppm~
## $ DAILY_AQI_VALUE                     <int> 27, 17, 15, 20, 34, 34, 27, 35, 3~
## $ Site.Name                           <fct> Taylorsville Liledoun, Taylorsvil~
## $ DAILY_OBS_COUNT                     <int> 24, 24, 24, 24, 24, 24, 24, 24, 2~
## $ PERCENT_COMPLETE                    <dbl> 100, 100, 100, 100, 100, 100, 100~
## $ AQS_PARAMETER_CODE                  <int> 44201, 44201, 44201, 44201, 44201~
## $ AQS_PARAMETER_DESC                  <fct> Ozone, Ozone, Ozone, Ozone, Ozone~
## $ CBSA_CODE                           <int> 25860, 25860, 25860, 25860, 25860~
## $ CBSA_NAME                           <fct> "Hickory-Lenoir-Morganton, NC", "~
## $ STATE_CODE                          <int> 37, 37, 37, 37, 37, 37, 37, 37, 3~
## $ STATE                               <fct> North Carolina, North Carolina, N~
## $ COUNTY_CODE                         <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ~
## $ COUNTY                              <fct> Alexander, Alexander, Alexander, ~
## $ SITE_LATITUDE                       <dbl> 35.9138, 35.9138, 35.9138, 35.913~
## $ SITE_LONGITUDE                      <dbl> -81.191, -81.191, -81.191, -81.19~
```

```
glimpse(o3_nc2018)
```

```
## Rows: 9,737
## Columns: 20
## $ Date                                <fct> 03/01/2018, 03/02/2018, 03/03/201~
## $ Source                              <fct> AQS, AQS, AQS, AQS, AQS, AQS, AQS~
## $ Site.ID                             <int> 370030005, 370030005, 370030005, ~
## $ POC                                 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ Daily.Max.8.hour.Ozone.Concentration <dbl> 0.043, 0.046, 0.047, 0.049, 0.047~
## $ UNITS                               <fct> ppm, ppm, ppm, ppm, ppm, ppm, ppm~
## $ DAILY_AQI_VALUE                     <int> 40, 43, 44, 45, 44, 28, 33, 41, 4~
## $ Site.Name                           <fct> Taylorsville Liledoun, Taylorsvil~
## $ DAILY_OBS_COUNT                     <int> 17, 17, 17, 17, 17, 17, 17, 17, 1~
## $ PERCENT_COMPLETE                    <dbl> 100, 100, 100, 100, 100, 100, 100~
## $ AQS_PARAMETER_CODE                  <int> 44201, 44201, 44201, 44201, 44201~
## $ AQS_PARAMETER_DESC                  <fct> Ozone, Ozone, Ozone, Ozone, Ozone~
## $ CBSA_CODE                           <int> 25860, 25860, 25860, 25860, 25860~
## $ CBSA_NAME                           <fct> "Hickory-Lenoir-Morganton, NC", "~
## $ STATE_CODE                          <int> 37, 37, 37, 37, 37, 37, 37, 37, 3~
## $ STATE                               <fct> North Carolina, North Carolina, N~
## $ COUNTY_CODE                         <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ~
## $ COUNTY                              <fct> Alexander, Alexander, Alexander, ~
## $ SITE_LATITUDE                       <dbl> 35.9138, 35.9138, 35.9138, 35.913~
## $ SITE_LONGITUDE                      <dbl> -81.191, -81.191, -81.191, -81.19~
```

Wrangle individual datasets to create processed files.

3. Change the Date columns to be date objects.
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. Use the same file names as the raw files but replace "raw" with "processed".

```
#3 Changing the "Date" columns in each dataset to be date objects using as.Date()
pm25_nc2019$Date <- as.Date(pm25_nc2019$Date, format = "%m/%d/%Y")
pm25_nc2018$Date <- as.Date(pm25_nc2018$Date, format = "%m/%d/%Y")
o3_nc2019$Date <- as.Date(o3_nc2019$Date, format = "%m/%d/%Y")
o3_nc2018$Date <- as.Date(o3_nc2018$Date, format = "%m/%d/%Y")

#4 Selecting only the desired columns
pm25_nc2019 <- pm25_nc2019[c(1,7,8,12,18,19,20)]
pm25_nc2018 <- pm25_nc2018[c(1,7,8,12,18,19,20)]
o3_nc2019 <- o3_nc2019[c(1,7,8,12,18,19,20)]
o3_nc2018 <- o3_nc2018[c(1,7,8,12,18,19,20)]

#5 Filling AQS_PARAMETER_DESC cells with "PM2.5"
pm25_nc2019$AQS_PARAMETER_DESC <- "PM2.5"
pm25_nc2018$AQS_PARAMETER_DESC <- "PM2.5"

#6 Saving the datasets to the Processed folder
write.csv(pm25_nc2019, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2019_processed.csv")
write.csv(pm25_nc2018, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2018_processed.csv")
write.csv(o3_nc2019, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2019_processed.csv")
write.csv(o3_nc2018, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2018_processed.csv")
```

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:

- Include only sites that the four data frames have in common: "Linville Falls", "Durham Armory", "Leggett", "Hattie Avenue", "Clemmons Middle", "Mendenhall School", "Frying Pan Mountain", "West Johnston Co.", "Garinger High School", "Castle Hayne", "Pitt Agri. Center", "Bryson City", "Millbrook School" (the function `intersect` can figure out common factor levels - but it will include sites with missing site information, which you don't want...)
- Some sites have multiple measurements per day. Use the split-apply-combine strategy to generate daily means: group by date, site name, AQS parameter, and county. Take the mean of the AQI value, latitude, and longitude.
- Add columns for "Month" and "Year" by parsing your "Date" column (hint: `lubridate` package)
- Hint: the dimensions of this dataset should be 14,752 x 9.

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_NC1819_Processed.csv"

```
#7 Combining the four datasets using rbind()
EPA_air <- rbind(pm25_nc2019, pm25_nc2018, o3_nc2019, o3_nc2018)

#8 Wrangling dataset to meet conditions
EPA_air_common_sites <-
  EPA_air %>%
  filter(Site.Name %in% c("Linville Falls","Durham Armory",
                                                        "Leggett","Hattie Avenue",
                                                        "Clemmons Middle","Mendenhall School",
                                                        "Frying Pan Mountain","West Johnston Co.",
                                                        "Garinger High School","Castle Hayne",
                                                        "Pitt Agri. Center","Bryson City",
                                                        "Millbrook School")) %>%
  group_by(Date,Site.Name,AQS_PARAMETER_DESC,COUNTY) %>%
  summarise(meanAQI = mean(DAILY_AQI_VALUE),
            meanLat = mean(SITE_LATITUDE),
            meanLong = mean(SITE_LONGITUDE))
```

```
## `summarise()` has grouped output by 'Date', 'Site.Name', 'AQS_PARAMETER_DESC'.
## You can override using the `.groups` argument.
```

```
EPA_air_common_sites <- mutate(EPA_air_common_sites, month = month(Date))

EPA_air_processed <- mutate(EPA_air_common_sites, year = year(Date))

#9 Creating two separate columns for Ozone and PM2.5
EPA_air_processed_2 <-
  EPA_air_processed %>%
  pivot_wider(names_from=AQS_PARAMETER_DESC, values_from = meanAQI)

#10 Finding dimensions of tidy dataset, it returns 9 columns and 8,976 rows
glimpse(EPA_air_processed_2)
```

```
## Rows: 8,976
## Columns: 9
## Groups: Date, Site.Name [8,976]
## $ Date      <date> 2018-01-01, 2018-01-01, 2018-01-01, 2018-01-01, 2018-01-01,~
## $ Site.Name <fct> Bryson City, Castle Hayne, Clemmons Middle, Durham Armory, G~
## $ COUNTY    <fct> Swain, New Hanover, Forsyth, Durham, Mecklenburg, Forsyth, E~
## $ meanLat   <dbl> 35.43477, 34.36417, 36.02600, 36.03296, 35.24010, 36.11069, ~
## $ meanLong  <dbl> -83.44213, -77.83861, -80.34200, -78.90404, -80.78568, -80.2~
## $ month     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ year      <dbl> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, ~
## $ PM2.5     <dbl> 35.0, 13.0, 24.0, 31.0, 20.0, 22.0, 14.0, 28.0, 15.0, 24.0, ~
## $ Ozone     <dbl> NA, NA, NA, NA, 32, NA, NA, 34, NA, NA, NA, NA, NA, NA, NA, ~
```

```r
#11 Writing a .csv file from our processed EPA air dataset that we cleaned
write.csv(EPA_air_processed_2, row.names = FALSE,
          file = "./Data/Processed/EPAair_O3_PM25_NC1819_Processed.csv")
```

Generate summary tables

12. Use the split-apply-combine strategy to generate a summary data frame. Data should be grouped by site, month, and year. Generate the mean AQI values for ozone and PM2.5 for each group. Then, add a pipe to remove instances where mean **ozone** values are not available (use the function `drop_na` in your pipe). It's ok to have missing mean PM2.5 values in this result.

13. Call up the dimensions of the summary dataset.

```r
#12 Creating a summary of O3 and PM2.5 means by location, month, and year
EPA_air_summary <-
  EPA_air_processed_2 %>%
  group_by(Site.Name,month,year) %>%
  na.omit(Ozone) %>%
  summarise(meanAQI_PM25 = mean(PM2.5),
            meanAQI_O3 = mean(Ozone))
```

```
## `summarise()` has grouped output by 'Site.Name', 'month'. You can override
## using the `.groups` argument.
```

```r
#13 Grabbing the dimensions of the summary dataset, it returns 5 columns and 223 rows
glimpse(EPA_air_summary)
```

```
## Rows: 223
## Columns: 5
## Groups: Site.Name, month [127]
## $ Site.Name    <fct> Bryson City, Bryson City, Bryson City, Bryson City, Bryso~
## $ month        <dbl> 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 2, 3~
## $ year         <dbl> 2019, 2018, 2019, 2018, 2019, 2018, 2019, 2018, 2019, 201~
## $ meanAQI_PM25 <dbl> 26.66667, 34.74194, 32.46154, 28.16667, 26.73333, 33.4814~
## $ meanAQI_O3   <dbl> 32.40000, 41.58065, 41.03846, 44.53333, 45.40000, 37.8888~
```

14. Why did we use the function `drop_na` rather than `na.omit`?

    Answer: drop_na() allows us to drop rows with NA values from a single column, in this case "Ozone", within our pipe just by specifying the column name. From what I understand, na.omit CAN do the same thing, it just may involve more steps.