Assignment 4: Data Wrangling

Hanna Karnei

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 Upload packages
library(tidyverse)
library(lubridate)
library(here)
library(dplyr)

#1b Set and check working directory
setwd("~/Desktop/EDA/EDE_Fall2023")
getwd()
```

```
## [1] "/Users/hannakarnei/Desktop/EDA/EDE_Fall2023"
```

```r
#1c Read files
epa_air1=read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv", stringsAsFactors = TRUE)
epa_air2=read.csv("./Data/Raw/EPAair_O3_NC2019_raw.csv", stringsAsFactors = TRUE)
epa_air3=read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv", stringsAsFactors = TRUE)
epa_air4=read.csv("./Data/Raw/EPAair_PM25_NC2019_raw.csv", stringsAsFactors = TRUE)

#2 Get information on files
glimpse(epa_air1)
```

```
## 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~
```

```r
glimpse(epa_air2)
```

```
## 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(epa_air3)
```

```
## 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(epa_air4)
```

```
## 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~
```

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 Change the Date columns to be date objects

epa_air1$Date<-mdy(epa_air1$Date)
epa_air2$Date<-mdy(epa_air2$Date)
epa_air3$Date<-mdy(epa_air3$Date)
epa_air4$Date<-mdy(epa_air4$Date)

#4 Select columns

epa_air1<-select(epa_air1, Date, DAILY_AQI_VALUE, Site.Name,
                 AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
epa_air2<-select(epa_air2, Date, DAILY_AQI_VALUE, Site.Name,
                 AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
epa_air3<-select(epa_air3, Date, DAILY_AQI_VALUE, Site.Name,
                 AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
epa_air4<-select(epa_air4, Date, DAILY_AQI_VALUE, Site.Name,
                 AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

#5 Fill a column with "PM2.5"

epa_air3$AQS_PARAMETER_DESC<-'PM2.5'
epa_air4$AQS_PARAMETER_DESC<-'PM2.5'

#6 Save files in Processed folder

write.csv(epa_air1, file = "./Data/Raw/EPAair_O3_NC2018_processed.csv", row.names=FALSE)
write.csv(epa_air2, file = "./Data/Raw/EPAair_O3_NC2019_processed.csv", row.names=FALSE)
write.csv(epa_air3, file = "./Data/Raw/EPAair_PM25_NC2018_processed.csv", row.names=FALSE)
write.csv(epa_air4, file = "./Data/Raw/EPAair_PM25_NC2018_processed.csv", row.names=FALSE)

head(epa_air1, 3)
```

```
##         Date DAILY_AQI_VALUE             Site.Name AQS_PARAMETER_DESC    COUNTY
## 1 2018-03-01              40 Taylorsville Liledoun              Ozone Alexander
## 2 2018-03-02              43 Taylorsville Liledoun              Ozone Alexander
## 3 2018-03-03              44 Taylorsville Liledoun              Ozone Alexander
##   SITE_LATITUDE SITE_LONGITUDE
## 1       35.9138        -81.191
## 2       35.9138        -81.191
## 3       35.9138        -81.191
```

```
head(epa_air2, 3)
```

```
##         Date DAILY_AQI_VALUE             Site.Name AQS_PARAMETER_DESC    COUNTY
## 1 2019-01-01              27 Taylorsville Liledoun              Ozone Alexander
## 2 2019-01-02              17 Taylorsville Liledoun              Ozone Alexander
## 3 2019-01-03              15 Taylorsville Liledoun              Ozone Alexander
##   SITE_LATITUDE SITE_LONGITUDE
## 1       35.9138        -81.191
## 2       35.9138        -81.191
## 3       35.9138        -81.191
```

```
head(epa_air3, 3)
```

```
##         Date DAILY_AQI_VALUE      Site.Name AQS_PARAMETER_DESC COUNTY
## 1 2018-01-02              12 Linville Falls              PM2.5  Avery
## 2 2018-01-05              15 Linville Falls              PM2.5  Avery
## 3 2018-01-08              22 Linville Falls              PM2.5  Avery
##   SITE_LATITUDE SITE_LONGITUDE
## 1      35.97235      -81.93307
## 2      35.97235      -81.93307
## 3      35.97235      -81.93307
```

```
head(epa_air4, 3)
```

```
##         Date DAILY_AQI_VALUE      Site.Name AQS_PARAMETER_DESC COUNTY
## 1 2019-01-03               7 Linville Falls              PM2.5  Avery
## 2 2019-01-06               4 Linville Falls              PM2.5  Avery
## 3 2019-01-09               5 Linville Falls              PM2.5  Avery
##   SITE_LATITUDE SITE_LONGITUDE
## 1      35.97235      -81.93307
## 2      35.97235      -81.93307
## 3      35.97235      -81.93307
```

Combine datasets

7. Combine the four datasets with `rbind`. Make sure your column names are identical prior to running this code.

```
#7 Combine datasets
epa_air<-rbind(epa_air1, epa_air2, epa_air3, epa_air4)
tail(epa_air)
```

```
##             Date DAILY_AQI_VALUE  Site.Name AQS_PARAMETER_DESC COUNTY
## 37888 2019-12-26              38 Triple Oak              PM2.5   Wake
## 37889 2019-12-27              48 Triple Oak              PM2.5   Wake
## 37890 2019-12-28              41 Triple Oak              PM2.5   Wake
## 37891 2019-12-29              27 Triple Oak              PM2.5   Wake
## 37892 2019-12-30              15 Triple Oak              PM2.5   Wake
## 37893 2019-12-31              18 Triple Oak              PM2.5   Wake
##       SITE_LATITUDE SITE_LONGITUDE
```

```
## 37888       35.8652       -78.8197
## 37889       35.8652       -78.8197
## 37890       35.8652       -78.8197
## 37891       35.8652       -78.8197
## 37892       35.8652       -78.8197
## 37893       35.8652       -78.8197
```

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"

```
#8 Wrangle data

epa_air_modified <- epa_air %>%
  filter(Site.Name=='Linville Falls' | Site.Name=='Durham Armory' |
           Site.Name=="Leggett" | Site.Name=="Hattie Avenue" |
           Site.Name=="Clemmons Middle" | Site.Name=="Mendenhall School" |
           Site.Name=="Frying Pan Mountain" | Site.Name=="West Johnston Co." |
           Site.Name=="Garinger High School" | Site.Name=="Castle Hayne" |
           Site.Name=="Pitt Agri. Center" | Site.Name=="Bryson City" |
           Site.Name=="Millbrook School") %>%
  group_by(Date, Site.Name, AQS_PARAMETER_DESC, COUNTY) %>%
  summarise(mean_AQI=mean(DAILY_AQI_VALUE),
            mean_lat=mean(SITE_LATITUDE),
            mean_long=mean(SITE_LONGITUDE),
            .groups = 'drop')  %>%
  mutate(Month = month(Date),
         Year = year(Date))

head(epa_air_modified)
```

```
## # A tibble: 6 x 9
##   Date       Site.Name        AQS_P~1 COUNTY mean_~2 mean_~3 mean_~4 Month  Year
##   <date>     <fct>            <fct>   <fct>    <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 2018-01-01 Bryson City      PM2.5   Swain       35    35.4   -83.4     1  2018
```

```
## 2 2018-01-01 Castle Hayne     PM2.5   New H~      13    34.4   -77.8     1  2018
## 3 2018-01-01 Clemmons Middle  PM2.5   Forsy~      24    36.0   -80.3     1  2018
## 4 2018-01-01 Durham Armory    PM2.5   Durham      31    36.0   -78.9     1  2018
## 5 2018-01-01 Garinger High S~ Ozone   Meckl~      32    35.2   -80.8     1  2018
## 6 2018-01-01 Garinger High S~ PM2.5   Meckl~      20    35.2   -80.8     1  2018
## # ... with abbreviated variable names 1: AQS_PARAMETER_DESC, 2: mean_AQI,
## #   3: mean_lat, 4: mean_long
```

```
dim(epa_air_modified)
```

```
## [1] 14752     9
```

```
#9 Spread datasets with pivot
epa_air_pivot <- epa_air_modified %>%
  filter(AQS_PARAMETER_DESC %in% c("Ozone", "PM2.5")) %>%
  pivot_wider(names_from = AQS_PARAMETER_DESC, values_from = mean_AQI)

head(epa_air_pivot)
```

```
## # A tibble: 6 x 9
##   Date       Site.Name            COUNTY mean_~1 mean_~2 Month  Year PM2.5 Ozone
##   <date>     <fct>                <fct>    <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 2018-01-01 Bryson City          Swain     35.4   -83.4     1  2018    35    NA
## 2 2018-01-01 Castle Hayne         New H~    34.4   -77.8     1  2018    13    NA
## 3 2018-01-01 Clemmons Middle      Forsy~    36.0   -80.3     1  2018    24    NA
## 4 2018-01-01 Durham Armory        Durham    36.0   -78.9     1  2018    31    NA
## 5 2018-01-01 Garinger High School Meckl~    35.2   -80.8     1  2018    20    32
## 6 2018-01-01 Hattie Avenue        Forsy~    36.1   -80.2     1  2018    22    NA
## # ... with abbreviated variable names 1: mean_lat, 2: mean_long
```

```
#10 Check dimensions
dim(epa_air_pivot)
```

```
## [1] 8976    9
```

```
#11 Save dataset
write.csv(epa_air_pivot, file = "./Data/Raw/EPAair_O3_PM25_NC1819_Processed.csv",
          row.names=FALSE)
```

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 Generate summary tables
epa_summary_table <- epa_air_pivot %>%
  group_by(Site.Name, Month, Year) %>%
  summarise(mean_AQI_ozone=mean(Ozone),
            mean_AQI_PM25=mean(PM2.5),
            .groups = 'drop') %>%
  drop_na(mean_AQI_ozone)

head(epa_summary_table)
```

```
## # A tibble: 6 x 5
##   Site.Name   Month  Year mean_AQI_ozone mean_AQI_PM25
##   <fct>       <dbl> <dbl>          <dbl>         <dbl>
## 1 Bryson City     3  2018           41.6          34.7
## 2 Bryson City     3  2019           42.5          NA
## 3 Bryson City     4  2018           44.5          28.2
## 4 Bryson City     4  2019           45.4          26.7
## 5 Bryson City     5  2019           39.6          NA
## 6 Bryson City     6  2018           37.8          NA
```

```r
#13 Check dimensions

dim(epa_summary_table)
```

```
## [1] 182   5
```

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

    Answer: We did not use na.omit because this function removes a row if it contains an "NA" in any column. In contrast, "drop_na" allows us to specify which column to check for "NA".