Assignment 4: Data Wrangling

Yao Yao

Spring 2023

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.

The completed exercise is due on Friday, Feb 20th @ 5:00pm.

Set up your session

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(here)
```

```
## here() starts at /Users/yaoyao/Desktop/ENV872/EDA-Spring2023
```

1b. Check your working directory.

```
getwd()
```

```
## [1] "/Users/yaoyao/Desktop/ENV872/EDA-Spring2023"
```

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).

```
O3_2018 <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv", stringsAsFactors = TRUE)
O3_2019 <- read.csv("./Data/Raw/EPAair_O3_NC2019_raw.csv", stringsAsFactors = TRUE)
PM2.5_2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv", stringsAsFactors = TRUE)
PM2.5_2019 <- read.csv("./Data/Raw/EPAair_PM25_NC2019_raw.csv", stringsAsFactors = TRUE)
```

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

```
# 1a
glimpse(O3_2018)
```

```
## 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~
```

```
# 1b
glimpse(O3_2019)
```

```
## 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~
```

```
# 1c
glimpse(PM2.5_2018)
```

```
## 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~
```

```
# 2
glimpse(PM2.5_2019)
```

```
## 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 date columns to be date objects.

```
O3_2018$Date <- as.Date(O3_2018$Date, format = "%m/%d/%Y")
O3_2019$Date <- as.Date(O3_2019$Date, format = "%m/%d/%Y")
PM2.5_2018$Date <- as.Date(PM2.5_2018$Date, format = "%m/%d/%Y")
PM2.5_2019$Date <- as.Date(PM2.5_2019$Date, format = "%m/%d/%Y")
```

4. Select the following columns: Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE

```
O3_2018_processed <- select(O3_2018, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
    COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
O3_2019_processed <- select(O3_2019, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
    COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
PM2.5_2018_processed <- select(PM2.5_2018, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
    COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
PM2.5_2019_processed <- select(PM2.5_2019, 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).

```
PM2.5_2018_processed <- mutate(PM2.5_2018_processed, AQS_PARAMETER_DESC = "PM2.5")
PM2.5_2019_processed <- mutate(PM2.5_2019_processed, AQS_PARAMETER_DESC = "PM2.5")
```

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
write.csv(O3_2018_processed, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2018_processed.csv

# 4
write.csv(O3_2019_processed, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2019_processed.csv

# 5
write.csv(PM2.5_2018_processed, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2018_processe

# 6
write.csv(PM2.5_2019_processed, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2019_processe
```

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 all 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...)
- 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
Air <- rbind(O3_2018_processed, O3_2019_processed, PM2.5_2018_processed, PM2.5_2019_processed)
summary(Air)
```

```
##       Date            DAILY_AQI_VALUE                Site.Name
##  Min.   :2018-01-01   Min.   :  0.00   Millbrook School    : 2169
##  1st Qu.:2018-06-27   1st Qu.: 27.00   Garinger High School: 1818
##  Median :2019-01-06   Median : 36.00   Hattie Avenue       : 1432
##  Mean   :2018-12-26   Mean   : 36.27   Durham Armory       : 1405
##  3rd Qu.:2019-06-23   3rd Qu.: 45.00   Pitt Agri. Center   : 1303
##  Max.   :2019-12-31   Max.   :136.00   Clemmons Middle     : 1261
##                                        (Other)             :28505
##  AQS_PARAMETER_DESC         COUNTY      SITE_LATITUDE   SITE_LONGITUDE
```

```
##  Ozone:20329        Mecklenburg: 3903   Min.   :34.36   Min.   :-83.80
##  PM2.5:17564        Forsyth    : 3175   1st Qu.:35.26   1st Qu.:-81.37
##                     Wake       : 2846   Median :35.64   Median :-80.23
##                     Cumberland : 1795   Mean   :35.62   Mean   :-80.21
##                     Haywood    : 1672   3rd Qu.:35.99   3rd Qu.:-78.77
##                     Swain      : 1628   Max.   :36.51   Max.   :-76.21
##                     (Other)    :22874
```

```
# 8
Air_Processed <- 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(meanAQI = mean(DAILY_AQI_VALUE), meanlatitude = mean(SITE_LATITUDE),
        meanlongitude = mean(SITE_LONGITUDE)) %>%
    mutate(Month = month(Date)) %>%
    mutate(Year = year(Date))
```

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

```
# 9
Air_Spread <- pivot_wider(Air_Processed, names_from = AQS_PARAMETER_DESC, values_from = meanAQI)

# 10

dim(Air_Spread)
```

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

```
# 11
write.csv(Air_Spread, 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.

```
# 12
Summary_Table <- Air_Spread %>%
    group_by(Site.Name, Month, Year) %>%
    summarise(meanPM2.5 = mean(PM2.5), meanOzone = mean(Ozone)) %>%
    drop_na(meanOzone)
```

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

```
# 13
dim(Summary_Table)
```

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

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

    Answer: drop_na drops the row with NA in a particular value (e.g. meanOzone); However, na.omit drops any row with NA values in it, no matter what column NA appears at. In this instance, when using drop_na, R only drops the row with NA in meanOzone and rows with NA in PM2.5 are still kept. However, when I tried to use na.omit, R drops all th rows with NA in it, including rows that has value for meanOzone, but has NA in PM2.5.