Assignment 4: Data Wrangling (Fall 2024)

Kamil Burak Karayel

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.

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. Add the appropriate code to reveal the dimensions of the four datasets.

```r
#1a
library(tidyverse)
library(lubridate)
library(here)
#1b
getwd()
```

```
## [1] "/home/guest/EDA_Spring2025_kbk"
```

```r
here()
```

```
## [1] "/home/guest/EDA_Spring2025_kbk"
```

```r
#1c
EPAair.O3.NC2018 <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv",stringsAsFactors = TRUE)
EPAair.O3.NC2019 <- read.csv("./Data/Raw/EPAair_O3_NC2019_raw.csv",stringsAsFactors = TRUE)
EPAair.PM25.NC2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv",stringsAsFactors = TRUE)
EPAair.PM25.NC2019 <- read.csv("./Data/Raw/EPAair_PM25_NC2019_raw.csv",stringsAsFactors = TRUE)

#2
dim(EPAair.O3.NC2018) #9737 rows and 20 columns
```

```
## [1] 9737   20
```

```r
dim(EPAair.O3.NC2019) #10592 rows and 20 columns
```

```
## [1] 10592    20
```

```r
dim(EPAair.PM25.NC2018) #8983 rows and 20 columns
```

```
## [1] 8983   20
```

```r
dim(EPAair.PM25.NC2019) #8581 rows and 20 columns
```

```
## [1] 8581   20
```

```r
#alternative way
glimpse(EPAair.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~
```

```r
glimpse(EPAair.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~
```

```r
glimpse(EPAair.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~
```

```r
glimpse(EPAair.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~
```

All four datasets should have the same number of columns but unique record counts (rows). Do your datasets follow this pattern?

#Answer: Yes, all four datasets have 20 columns but different observations.

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".

```r
#3
EPAair.O3.NC2018$Date <- as.Date(EPAair.O3.NC2018$Date, format = "%m/%d/%Y")
#Change vector type of column "Date" from Factor to Date
EPAair.O3.NC2019$Date <- as.Date(EPAair.O3.NC2019$Date, format = "%m/%d/%Y")
#Change vector type of column "Date" from Factor to Date
EPAair.PM25.NC2018$Date <- as.Date(EPAair.PM25.NC2018$Date, format = "%m/%d/%Y")
#Change vector type of column "Date" from Factor to Date
EPAair.PM25.NC2019$Date <- as.Date(EPAair.PM25.NC2019$Date, format = "%m/%d/%Y")
#Change vector type of column "Date" from Factor to Date

#4
Q4.EPAair.O3.NC2018 <- EPAair.O3.NC2018 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY,
         SITE_LATITUDE, SITE_LONGITUDE)
#Seven columns selected from EPAair.O3.NC2018 and a ne dataframe created.

Q4.EPAair.O3.NC2019 <- EPAair.O3.NC2019 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY,
         SITE_LATITUDE, SITE_LONGITUDE)
#Seven columns selected from EPAair.O3.NC2019 and a ne dataframe created.

Q4.EPAair.PM25.NC2018 <- EPAair.PM25.NC2018 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY,
         SITE_LATITUDE, SITE_LONGITUDE)
#Seven columns selected from EPAair.PM25.NC2018 and a ne dataframe created.

Q4.EPAair.PM25.NC2019 <- EPAair.PM25.NC2019 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY,
         SITE_LATITUDE, SITE_LONGITUDE)
#Seven columns selected from EPAair.PM25.NC2019 and a ne dataframe created.

#5
Q5.EPAair.PM25.NC2018 <- Q4.EPAair.PM25.NC2018 %>% mutate(AQS_PARAMETER_DESC = "PM2.5")
#Because it is a PM2.5 dataset, all AQS_PARAMETER_DESC cells are changed with "PM2.5"

Q5.EPAair.PM25.NC2019 <- Q4.EPAair.PM25.NC2019 %>% mutate(AQS_PARAMETER_DESC = "PM2.5")
#Because it is a PM2.5 dataset, all AQS_PARAMETER_DESC cells are changed with "PM2.5"

#6
write.csv(Q4.EPAair.O3.NC2018, row.names = FALSE,
          file = "./Data/Processed/EPAair_O3_NC2018_Processed.csv")
write.csv(Q4.EPAair.O3.NC2019, row.names = FALSE,
          file = "./Data/Processed/EPAair_O3_NC2019_Processed.csv")
write.csv(Q5.EPAair.PM25.NC2018, row.names = FALSE,
          file = "./Data/Processed/EPAair_PM25_NC2018_Processed.csv")
write.csv(Q5.EPAair.PM25.NC2019, row.names = FALSE,
          file = "./Data/Processed/EPAair_PM25_NC2019_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
Q7 <- rbind(Q4.EPAair.O3.NC2018, Q4.EPAair.O3.NC2019, Q5.EPAair.PM25.NC2018, Q5.EPAair.PM25.NC2019)
#Four datasets from Q4 and Q5 are merged by rbind. Rows reached to 37893, column number is same as 7.

#8
common_sites_O3 <- intersect(Q4.EPAair.O3.NC2018$Site.Name, Q4.EPAair.O3.NC2019$Site.Name)
#common sites from O3 datasets.
common_sites_PM25 <- intersect(Q5.EPAair.PM25.NC2018$Site.Name, Q5.EPAair.PM25.NC2019$Site.Name)
#common sites from PM2.5 datasets.
common_sites <- intersect(common_sites_O3, common_sites_PM25)
#common sites from all four datasets.

Q7$Site.Name[Q7$Site.Name == ""] <- NA #Change "" values in Site.Name column to NA.
Q8_v1 <- Q7 %>% filter(Site.Name %in% c(common_sites)) %>% drop_na(Site.Name)
#Take only sites that are included in common_sites and drop missing values.

Q8_v2 <- Q8_v1 %>%
   group_by(Date, Site.Name, AQS_PARAMETER_DESC, COUNTY) %>%
   summarize(
         Mean_AQI = mean(DAILY_AQI_VALUE, na.rm = TRUE), #Take mean values for different observations.
         Mean_Latitude = mean(SITE_LATITUDE, na.rm = TRUE),
         Mean_Longitude = mean(SITE_LONGITUDE, na.rm = TRUE)
   )
```

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

```
Q8_v3 <- Q8_v2 %>% mutate(Month = month(Date), Year = year(Date))
#Add year and month columns from "Date" column by using Lubridate functions month and year.

#9
Q9 <- pivot_wider(Q8_v3, names_from = AQS_PARAMETER_DESC, values_from = Mean_AQI)
#Create new columns as PM2.5 and Ozone (from AQS_PARAMETER_DESC) and write AQI values in the cells.

#10
dim(Q9)
```

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

```
#8976 rows and 9 columns. Two columns disappeared and two new columns added in Question 9.

#11
write.csv(Q9, 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
Q12 <- Q9 %>% group_by(Site.Name, Month, Year) %>%
  #Group the rows by these column values and take average value for AQI values.
  summarize(mean_Ozone_AQI = mean(Ozone, na.rm = TRUE),
            mean_PM25_AQI = mean(PM2.5, na.rm = TRUE)) %>%
  drop_na(mean_Ozone_AQI)
```

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

```
#13
dim(Q12) #239 rows and 5 columns.
```

```
## [1] 239   5
```

14. Why did we use the function `drop_na` rather than `na.omit`? Hint: replace `drop_na` with `na.omit` in part 12 and observe what happens with the dimensions of the summary date frame.

    Answer: drop_na() drops all NA values in the desired column as I used mean_Ozone_AQI here. But na.omit drops every NA values from other columns, so we have a 223x5 dataframe. This decrease in rows is because of 16 NA values in PM2.5 column.