Assignment 4: Data Wrangling

Laura Martinez

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A03_DataExploration.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, Oct7th @ 5:00pm.

Set up your session

1. Check your working directory, load the `tidyverse` and `lubridate` packages, and upload 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. Explore the dimensions, column names, and structure of the datasets.

```
# 1

# setwd('~/Documents/EDA-Fall2022')
getwd()
```

```
## [1] "/Users/laura/Documents/EDA-Fall2022"
```

```
# install.packages(tidyverse)
library(tidyverse)
# install.packages(lubridate)
library(lubridate)

EPA_O3_2018 <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv", stringsAsFactors = TRUE)
EPA_O3_2019 <- read.csv("./Data/Raw/EPAair_O3_NC2019_raw.csv", stringsAsFactors = TRUE)
EPA_PM25_2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv", stringsAsFactors = TRUE)
EPA_PM25_2019 <- read.csv("./Data/Raw/EPAair_PM25_NC2019_raw.csv", stringsAsFactors = TRUE)

# 2 Dimensions & Column names

# Data set 1 number of rows
nrow(EPA_O3_2018)
```

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

```r
# number of columns
ncol(EPA_03_2018)
```

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

```r
# column names
colnames(EPA_03_2018)
```

```
##  [1] "Date"
##  [2] "Source"
##  [3] "Site.ID"
##  [4] "POC"
##  [5] "Daily.Max.8.hour.Ozone.Concentration"
##  [6] "UNITS"
##  [7] "DAILY_AQI_VALUE"
##  [8] "Site.Name"
##  [9] "DAILY_OBS_COUNT"
## [10] "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"
## [12] "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"
## [14] "CBSA_NAME"
## [15] "STATE_CODE"
## [16] "STATE"
## [17] "COUNTY_CODE"
## [18] "COUNTY"
## [19] "SITE_LATITUDE"
## [20] "SITE_LONGITUDE"
```

```r
# number of rows by number of columns
dim(EPA_03_2018)
```

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

```r
# length of objects
length(EPA_03_2018)
```

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

```r
# Data set 2 number of rows
nrow(EPA_03_2019)
```

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

```r
# number of columns
ncol(EPA_03_2019)
```

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

```r
# column names
colnames(EPA_03_2019)
```

```
##  [1] "Date"
##  [2] "Source"
##  [3] "Site.ID"
##  [4] "POC"
##  [5] "Daily.Max.8.hour.Ozone.Concentration"
##  [6] "UNITS"
##  [7] "DAILY_AQI_VALUE"
```

```
##  [8] "Site.Name"
##  [9] "DAILY_OBS_COUNT"
## [10] "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"
## [12] "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"
## [14] "CBSA_NAME"
## [15] "STATE_CODE"
## [16] "STATE"
## [17] "COUNTY_CODE"
## [18] "COUNTY"
## [19] "SITE_LATITUDE"
## [20] "SITE_LONGITUDE"
```

```
# number of rows by number of columns
dim(EPA_O3_2019)
```

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

```
# length of objects
length(EPA_O3_2019)
```

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

```
# Data set 3 number of rows
nrow(EPA_PM25_2018)
```

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

```
# number of columns
ncol(EPA_PM25_2018)
```

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

```
# column names
colnames(EPA_PM25_2018)
```

```
##  [1] "Date"                           "Source"
##  [3] "Site.ID"                        "POC"
##  [5] "Daily.Mean.PM2.5.Concentration" "UNITS"
##  [7] "DAILY_AQI_VALUE"                "Site.Name"
##  [9] "DAILY_OBS_COUNT"                "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"             "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"                      "CBSA_NAME"
## [15] "STATE_CODE"                     "STATE"
## [17] "COUNTY_CODE"                    "COUNTY"
## [19] "SITE_LATITUDE"                  "SITE_LONGITUDE"
```

```
# number of rows by number of columns
dim(EPA_PM25_2018)
```

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

```
# length of objects
length(EPA_PM25_2018)
```

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

```
# Data set 4 number of rows
nrow(EPA_PM25_2019)
```

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

```
# number of columns
ncol(EPA_PM25_2019)
```

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

```
# column names
colnames(EPA_PM25_2018)
```

```
##  [1] "Date"                           "Source"
##  [3] "Site.ID"                        "POC"
##  [5] "Daily.Mean.PM2.5.Concentration" "UNITS"
##  [7] "DAILY_AQI_VALUE"                "Site.Name"
##  [9] "DAILY_OBS_COUNT"                "PERCENT_COMPLETE"
## [11] "AQS_PARAMETER_CODE"             "AQS_PARAMETER_DESC"
## [13] "CBSA_CODE"                      "CBSA_NAME"
## [15] "STATE_CODE"                     "STATE"
## [17] "COUNTY_CODE"                    "COUNTY"
## [19] "SITE_LATITUDE"                  "SITE_LONGITUDE"
```

```
# number of rows by number of columns
dim(EPA_PM25_2019)
```

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

```
# length of objects
length(EPA_PM25_2019)
```

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

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. Use the same file names as the raw files but replace "raw" with "processed".

```
# 3
EPA_O3_2018$Date <- as.Date(EPA_O3_2018$Date, format = "%Y/%m/%d")
EPA_O3_2019$Date <- as.Date(EPA_O3_2019$Date, format = "%m/%d/%Y")
EPA_PM25_2018$Date <- as.Date(EPA_PM25_2018$Date, format = "%m/%d/%Y")
EPA_PM25_2019$Date <- as.Date(EPA_PM25_2019$Date, format = "%m/%d/%Y")

# 4 library(dplyr)
subset_EPA_O3_2018 <- select(EPA_O3_2018, c("Date", "DAILY_AQI_VALUE", "Site.Name",
    "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE"))

subset_EPA_O3_2019 <- select(EPA_O3_2019, c("Date", "DAILY_AQI_VALUE", "Site.Name",
    "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE"))

subset_EPA_PM25_2018 <- select(EPA_PM25_2018, c("Date", "DAILY_AQI_VALUE", "Site.Name",
    "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE"))

subset_EPA_PM25_2019 <- select(EPA_PM25_2019, c("Date", "DAILY_AQI_VALUE", "Site.Name",
```

```
    "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE"))

# 5
subset_EPA_PM25_2018$AQS_PARAMETER_DESC <- "PM2.5"
subset_EPA_PM25_2019$AQS_PARAMETER_DESC <- "PM2.5"

# 6
write.csv(subset_EPA_O3_2018, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2018_processed.csv
write.csv(subset_EPA_O3_2019, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2019_processed.csv
write.csv(subset_EPA_PM25_2018, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2018_processed
write.csv(subset_EPA_PM25_2019, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2019_processed
```

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)
- Some sites have multiple measurements per day. Use the split-apply-combine strategy to generate daily means: group by date, site, 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_NC1718_Processed.csv"

```
# 7
EPAair_combined <- rbind(subset_EPA_O3_2018, subset_EPA_O3_2019, subset_EPA_PM25_2018,
    subset_EPA_PM25_2019)

# 8
EPAair_conditions <- EPAair_combined %>%
    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_Lon = mean(SITE_LO
    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
EPAair_spread <- pivot_wider(EPAair_conditions, names_from = AQS_PARAMETER_DESC,
    values_from = mean_AQI)
```

```
# 10 Dimensions of new data set
dim(EPAair_spread)
```

```
## [1] 9862    9
```

```
# 11
write.csv(EPAair_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 a month and year are not available (use the function `drop_na` in your pipe).

13. Call up the dimensions of the summary dataset.

```
# 12a group_by and summarise first
EPA_summary <- EPAair_spread %>%
    group_by(Site.Name, Month, Year) %>%
    summarise(mean_Ozone = mean(Ozone), mean_PM25 = mean(PM2.5))
```

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

```
# 12b Feed into Pipe
EPA_dropNA <- EPA_summary %>%
    drop_na(mean_Ozone) %>%
    drop_na(mean_PM25)

# 13
dim(EPA_dropNA)
```

```
## [1] 56  5
```

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

    Answer: Drop_na is used to drop rows that contain missing values (NA) in specified columns, whereas na.omit is more generic and will drop any rows that have any missing values in the dataframe. Drop_na allows the user to have more control over which columns will be targeted for NA removal.