Assignment 4: Data Wrangling (Fall 2024)

Zhaoxin Zhang

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.

```
#1a
library(tidyverse)
library(lubridate)
library(here) #install the packages
#1b
getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
#1c
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
NC_O3_2018<- read.csv(
  file=here("Data/Raw/EPAair_O3_NC2018_raw.csv"),
  stringsAsFactors = TRUE
)

NC_O3_2019<- read.csv(
  file=here("Data/Raw/EPAair_O3_NC2019_raw.csv"),
  stringsAsFactors = TRUE
)

NC_PM25_2018<- read.csv(
  file=here("Data/Raw/EPAair_PM25_NC2018_raw.csv"),
  stringsAsFactors = TRUE
)

NC_PM25_2019<- read.csv(
  file=here("Data/Raw/EPAair_PM25_NC2019_raw.csv"),
  stringsAsFactors = TRUE
)

#2
dim(NC_O3_2018) #9737 rows and 20 columns
```

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

```
dim(NC_O3_2019) #10592 rows and 20 columns
```

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

```
dim(NC_PM25_2018) #8983 rows and 20 columns
```

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

```
dim(NC_PM25_2019) #8581 rows and 20 columns
```

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

All four datasets should have the same number of columns but unique record counts (rows). Do your datasets follow this pattern? Yes. They all have 20 columns, but unique rows.

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" to date format
NC_O3_2018$Date <- mdy(NC_O3_2018$Date)
NC_O3_2019$Date <- mdy(NC_O3_2019$Date)
NC_PM25_2018$Date <- mdy(NC_PM25_2018$Date)
NC_PM25_2019$Date <- mdy(NC_PM25_2019$Date)

#4
#Create new datasets only contain Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
#COUNTY, SITE_LATITUDE, SITE_LONGITUDE
NC_O3_2018_new <-
  NC_O3_2018 %>%
  select( Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
          COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

NC_O3_2019_new <-
  NC_O3_2019 %>%
  select( Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
          COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

NC_PM25_2018_new <-
  NC_PM25_2018 %>%
  select( Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
          COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

NC_PM25_2019_new <-
  NC_PM25_2019 %>%
  select( Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
          COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

#5
library(dplyr) #install package

NC_PM25_2018_new <- NC_PM25_2018_new %>%
  mutate(AQS_PARAMETER_DESC = "PM2.5") #Change to PM2.5

NC_PM25_2019_new <- NC_PM25_2019_new %>%
  mutate(AQS_PARAMETER_DESC = "PM2.5") #Change to PM2.5

#6
write.csv(NC_O3_2018_new, row.names = FALSE,
          file ="./Data/Processed/EPAair_O3_NC2018_processed.csv")
write.csv(NC_O3_2019_new, row.names = FALSE,
          file ="./Data/Processed/EPAair_O3_NC2019_processed.csv")
write.csv(NC_PM25_2018_new, row.names = FALSE,
          file ="./Data/Processed/EPAair_PM25_NC2018_processed.csv")
write.csv(NC_PM25_2019_new, 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
Air_Quality_201819<-rbind(NC_O3_2018_new,NC_O3_2019_new,
                          NC_PM25_2018_new,NC_PM25_2019_new)
#Combine the four datasets into one

#8
Air_Quality_201819_new <- Air_Quality_201819 %>%
   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"))%>% #filter the common factors
  mutate(Month = month(Date),
         Year = year(Date)) %>% #add month and data columns
  group_by(Date, Site.Name, AQS_PARAMETER_DESC, COUNTY, Month, Year) %>%
  summarise(AQI_mean = mean(DAILY_AQI_VALUE),
            Latitude_mean = mean(SITE_LATITUDE),
            Longitude_mean = mean(SITE_LONGITUDE)) #summarise the means
```

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

```
#9
Air_Quality_201819_new_2 <- Air_Quality_201819_new %>%
  pivot_wider(names_from=AQS_PARAMETER_DESC, values_from= AQI_mean)
#Change the columns to means of PM2.5 and O3

#10

dim(Air_Quality_201819_new_2)
```

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

```
#There are 8976 rows and 9 columns

#11
write.csv(Air_Quality_201819_new_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.

```
#12
Summaries_Air_Quality<-
  Air_Quality_201819_new_2 %>%
  group_by(Site.Name, Month, Year) %>%
  summarise(mean_PM2.5= mean(PM2.5),
            mean_O3=mean(Ozone)) %>%
  drop_na(mean_O3)
```

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

```
print(Summaries_Air_Quality)
```

```
## # A tibble: 182 x 5
## # Groups:   Site.Name, Month [109]
##    Site.Name   Month  Year mean_PM2.5 mean_O3
##    <fct>       <dbl> <dbl>      <dbl>   <dbl>
##  1 Bryson City     3  2018       34.7    41.6
##  2 Bryson City     3  2019       NA      42.5
##  3 Bryson City     4  2018       28.2    44.5
##  4 Bryson City     4  2019       26.7    45.4
##  5 Bryson City     5  2019       NA      39.6
##  6 Bryson City     6  2018       NA      37.8
##  7 Bryson City     6  2019       NA      34.0
##  8 Bryson City     7  2018       NA      34.6
```

```
##  9 Bryson City     7  2019       33.6    30.4
## 10 Bryson City     8  2018       NA      30.8
## # i 172 more rows
```

```
#13
dim(Summaries_Air_Quality)
```

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

```
#There are 182 rows and 5 columns
```

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: We use `drop_na` because it only deletes the rows that contain NA in mean_O3 (which we want to remove). If we use `na.omit`, it will delete all the rows of mean_PM2.5 and mean_O3 that contain "NA". The dimiensions will change to have only 101 rows.