SamanthaSedar_A04_DataWrangling.Rmd

Samantha Sedar

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.

```
#changing wd from Assignments folder
setwd("/home/guest/EDE_Fall2023")


#1a - on-startup code
library(tidyverse)
library(lubridate)
library(here)

#1b - confirm wd
getwd()
```

```
## [1] "/home/guest/EDE_Fall2023"
```

```r
#1c - read/load files
EPA_PM25_19 <- read.csv("./Data/Raw/EPAair_PM25_NC2019_raw.csv",stringsAsFactors = T)
EPA_PM25_18 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv",stringsAsFactors = T)
EPA_O3_19 <- read.csv("./Data/Raw/EPAair_O3_NC2019_raw.csv",stringsAsFactors = T)
EPA_O3_18 <- read.csv("./Data/Raw/EPAair_O3_NC2018_raw.csv",stringsAsFactors = T)

#2 - use glimpse functio for each
glimpse(EPA_PM25_19)
```

```
## 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~
```

```r
glimpse(EPA_PM25_18)
```

```
## 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_O3_19)
```

```
## 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_O3_18)
```

```
## 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~
```

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
#Converting all data to date objects using as date

EPA_PM25_19$Date <- as.Date(EPA_PM25_19$Date, format = "%m/%d/%Y")
EPA_PM25_18$Date <- as.Date(EPA_PM25_18$Date, format = "%m/%d/%Y")
EPA_O3_19$Date <- as.Date(EPA_O3_19$Date, format = "%m/%d/%Y")
EPA_O3_18$Date <- as.Date(EPA_O3_18$Date, format = "%m/%d/%Y")

#4
#Creating new dataframes with the four columns of interest for each
EPA_PM25_19_4 <- select(EPA_PM25_19, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
EPA_PM25_18_4 <- select(EPA_PM25_18, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
EPA_O3_19_4 <- select(EPA_O3_19, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
EPA_O3_18_4 <- select(EPA_O3_18, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC,
COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

#5
#Mutating column to fill all celss with PM2.5
EPA_PM25_19_4 <- mutate(EPA_PM25_19_4, AQS_PARAMETER_DESC = "PM2.5")
EPA_PM25_18_4 <- mutate(EPA_PM25_18_4, AQS_PARAMETER_DESC = "PM2.5")

#6
#Saving to processed
write.csv(EPA_PM25_19_4, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2019_Processed.csv")
write.csv(EPA_PM25_18_4, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2018_Processed.csv")
write.csv(EPA_O3_19_4, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2019_Processed.csv")
write.csv(EPA_O3_18_4, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2018_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 #Confirming identical column names and then combining with rbind

colnames(EPA_PM25_19_4)
```

```
## [1] "Date"                "DAILY_AQI_VALUE"    "Site.Name"
## [4] "AQS_PARAMETER_DESC" "COUNTY"             "SITE_LATITUDE"
## [7] "SITE_LONGITUDE"
```

```
colnames(EPA_PM25_18_4)
```

```
## [1] "Date"                "DAILY_AQI_VALUE"    "Site.Name"
## [4] "AQS_PARAMETER_DESC" "COUNTY"             "SITE_LATITUDE"
## [7] "SITE_LONGITUDE"
```

```
colnames(EPA_O3_19_4)
```

```
## [1] "Date"                "DAILY_AQI_VALUE"    "Site.Name"
## [4] "AQS_PARAMETER_DESC" "COUNTY"             "SITE_LATITUDE"
## [7] "SITE_LONGITUDE"
```

```
colnames(EPA_O3_18_4)
```

```
## [1] "Date"                "DAILY_AQI_VALUE"    "Site.Name"
## [4] "AQS_PARAMETER_DESC" "COUNTY"             "SITE_LATITUDE"
## [7] "SITE_LONGITUDE"
```

```
EPA_Data <- rbind(EPA_O3_18_4,EPA_O3_19_4,EPA_PM25_18_4,EPA_PM25_19_4)

#8
#confirming date class, reformatting, creating a new df using identified specs *note that mutate should

class(EPA_Data)
```

```
## [1] "data.frame"
```

```
EPA_Data_Processed <- EPA_Data  %>% #filter to select site names
 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) %>%
  summarize(mean_aqi = mean(DAILY_AQI_VALUE),
    mean_lat = mean(SITE_LATITUDE),
    mean_long = 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 Spread your datasets such that AQI values for ozone and PM2.5 are in separate columns-- using pivot

EPA_Data_Spread <- pivot_wider(EPA_Data_Processed, names_from ="AQS_PARAMETER_DESC",
values_from ="mean_aqi")


#10
#Calling dimensions
dim(EPA_Data_Spread)
```

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

```
#11
#Saving  processed dataset

write.csv(EPA_Data_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 #In this instance, group_by through drop_na are leveraging spilt-aplly-combine

EPA_Data_Join <- full_join(EPA_Data_Processed, EPA_Data_Spread)
```

```
## Joining with `by = join_by(Date, Site.Name, COUNTY, mean_lat, mean_long, Month,
## Year)`
```

```
EPA_Data_Summary <- EPA_Data_Join %>%
  group_by(Site.Name, Month, Year) %>%
  summarize(mean_aqi_ozone = mean(Ozone),
            mean_aqi_PM2.5 = mean(PM2.5))%>%
  drop_na(mean_aqi_ozone)
```

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

```
#13 Use dim() to call up the dimensions of the summary dataset

dim(EPA_Data_Summary)
```

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

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

    Answer: 'Na.omit' removes all NAs in any column across all, whereas 'drop_na' excludes NAs from a specific column, in this case ozone.