Assignment 4: Data Wrangling

Wynona Curaming

OVERVIEW

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

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.

```
#1a
library(tidyverse)
library(lubridate)
library(here)
library(tidyr)

#1b
getwd()
```

```
## [1] "/home/guest/ENV 872/EDA_Spring2024"
```

```
#1c
EPA.air.data.O3.NC2018 <- read.csv(
  file=here("~/ENV 872/EDA_Spring2024/Data/Raw/EPAair_O3_NC2018_raw.csv"),
  stringsAsFactors = TRUE
)
EPA.air.data.O3.NC2019 <- read.csv(
  file=here("~/ENV 872/EDA_Spring2024/Data/Raw/EPAair_O3_NC2019_raw.csv"),
  stringsAsFactors = TRUE
)
EPA.air.data.PM25.NC2018 <- read.csv(
  file=here("~/ENV 872/EDA_Spring2024/Data/Raw/EPAair_PM25_NC2018_raw.csv"),
  stringsAsFactors = TRUE
)
EPA.air.data.PM25.NC2019 <- read.csv(
  file=here("~/ENV 872/EDA_Spring2024/Data/Raw/EPAair_PM25_NC2019_raw.csv"),
  stringsAsFactors = TRUE
```

```r
)

#2
glimpse(EPA.air.data.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(EPA.air.data.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(EPA.air.data.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(EPA.air.data.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~
```

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
EPA.air.data.O3.NC2018$Date<- as.Date(EPA.air.data.O3.NC2018$Date, format="%m/%d/%Y")
EPA.air.data.O3.NC2019$Date<- as.Date(EPA.air.data.O3.NC2019$Date, format="%m/%d/%Y")
EPA.air.data.PM25.NC2018$Date<- as.Date(EPA.air.data.PM25.NC2018$Date, format ="%m/%d/%Y")
EPA.air.data.PM25.NC2019$Date<- as.Date(EPA.air.data.PM25.NC2019$Date, format ="%m/%d/%Y")

head(EPA.air.data.O3.NC2018$Date)
```

```
## [1] "2018-03-01" "2018-03-02" "2018-03-03" "2018-03-04" "2018-03-05"
## [6] "2018-03-06"
```

```
head(EPA.air.data.O3.NC2019$Date)
```

```
## [1] "2019-01-01" "2019-01-02" "2019-01-03" "2019-01-04" "2019-01-05"
## [6] "2019-01-06"
```

```
head(EPA.air.data.PM25.NC2018$Date)
```

```
## [1] "2018-01-02" "2018-01-05" "2018-01-08" "2018-01-11" "2018-01-14"
## [6] "2018-01-17"
```

```
head(EPA.air.data.PM25.NC2019$Date)
```

```
## [1] "2019-01-03" "2019-01-06" "2019-01-09" "2019-01-12" "2019-01-15"
## [6] "2019-01-18"
```

```
#4

EPA.air.data.O3.NC2018.selection<-select(EPA.air.data.O3.NC2018, Date, DAILY_AQI_VALUE, Site.Name, AQS_
EPA.air.data.O3.NC2019.selection<-select(EPA.air.data.O3.NC2019, Date, DAILY_AQI_VALUE, Site.Name, AQS_
EPA.air.data.PM25.NC2018.selection<-select(EPA.air.data.PM25.NC2018, Date, DAILY_AQI_VALUE, Site.Name, A
EPA.air.data.PM25.NC2019.selection<-select(EPA.air.data.PM25.NC2019, Date, DAILY_AQI_VALUE, Site.Name, A

head(EPA.air.data.O3.NC2018.selection)
```

```
##         Date DAILY_AQI_VALUE             Site.Name AQS_PARAMETER_DESC    COUNTY
## 1 2018-03-01              40 Taylorsville Liledoun              Ozone Alexander
## 2 2018-03-02              43 Taylorsville Liledoun              Ozone Alexander
## 3 2018-03-03              44 Taylorsville Liledoun              Ozone Alexander
## 4 2018-03-04              45 Taylorsville Liledoun              Ozone Alexander
## 5 2018-03-05              44 Taylorsville Liledoun              Ozone Alexander
## 6 2018-03-06              28 Taylorsville Liledoun              Ozone Alexander
##   SITE_LATITUDE SITE_LONGITUDE
## 1       35.9138        -81.191
## 2       35.9138        -81.191
## 3       35.9138        -81.191
## 4       35.9138        -81.191
## 5       35.9138        -81.191
## 6       35.9138        -81.191
```

```
head(EPA.air.data.O3.NC2019.selection)
```

```
##         Date DAILY_AQI_VALUE             Site.Name AQS_PARAMETER_DESC    COUNTY
## 1 2019-01-01              27 Taylorsville Liledoun              Ozone Alexander
## 2 2019-01-02              17 Taylorsville Liledoun              Ozone Alexander
## 3 2019-01-03              15 Taylorsville Liledoun              Ozone Alexander
## 4 2019-01-04              20 Taylorsville Liledoun              Ozone Alexander
## 5 2019-01-05              34 Taylorsville Liledoun              Ozone Alexander
## 6 2019-01-06              34 Taylorsville Liledoun              Ozone Alexander
##   SITE_LATITUDE SITE_LONGITUDE
## 1       35.9138        -81.191
## 2       35.9138        -81.191
## 3       35.9138        -81.191
## 4       35.9138        -81.191
## 5       35.9138        -81.191
## 6       35.9138        -81.191
```

```
head(EPA.air.data.PM25.NC2018.selection)
```

```
##         Date DAILY_AQI_VALUE      Site.Name
## 1 2018-01-02              12 Linville Falls
## 2 2018-01-05              15 Linville Falls
## 3 2018-01-08              22 Linville Falls
## 4 2018-01-11               3 Linville Falls
## 5 2018-01-14              10 Linville Falls
## 6 2018-01-17              19 Linville Falls
##                AQS_PARAMETER_DESC COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 2 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 3 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 4 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 5 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 6 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
```

```
head(EPA.air.data.PM25.NC2019.selection)
```

```
##         Date DAILY_AQI_VALUE      Site.Name
## 1 2019-01-03               7 Linville Falls
## 2 2019-01-06               4 Linville Falls
## 3 2019-01-09               5 Linville Falls
## 4 2019-01-12              26 Linville Falls
## 5 2019-01-15              11 Linville Falls
## 6 2019-01-18               5 Linville Falls
##                AQS_PARAMETER_DESC COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 2 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 3 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 4 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 5 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
## 6 Acceptable PM2.5 AQI & Speciation Mass  Avery      35.97235      -81.93307
```

```
#5
EPA.air.data.PM25.NC2018.selection.mutated <- mutate(EPA.air.data.PM25.NC2018.selection, AQS_PARAMETER_D
EPA.air.data.PM25.NC2019.selection.mutated <- mutate(EPA.air.data.PM25.NC2019.selection, AQS_PARAMETER_D
glimpse(EPA.air.data.PM25.NC2019.selection.mutated)
```

```
## Rows: 8,581
## Columns: 7
## $ Date               <date> 2019-01-03, 2019-01-06, 2019-01-09, 2019-01-12, 20~
## $ DAILY_AQI_VALUE    <int> 7, 4, 5, 26, 11, 5, 6, 6, 15, 7, 14, 20, 8, 10, 8, ~
## $ Site.Name          <fct> Linville Falls, Linville Falls, Linville Falls, Lin~
## $ AQS_PARAMETER_DESC <chr> "PM2.5", "PM2.5", "PM2.5", "PM2.5", "PM2.5", "PM2.5~
## $ COUNTY             <fct> Avery, Avery, Avery, Avery, Avery, Avery, Avery, Av~
## $ SITE_LATITUDE      <dbl> 35.97235, 35.97235, 35.97235, 35.97235, 35.97235, 3~
## $ SITE_LONGITUDE     <dbl> -81.93307, -81.93307, -81.93307, -81.93307, -81.933~
```

```
#6
write.csv(EPA.air.data.O3.NC2018.selection, row.names = FALSE, file = "~/ENV 872/EDA_Spring2024/Data/Pro
write.csv(EPA.air.data.O3.NC2019.selection, row.names = FALSE, file = "~/ENV 872/EDA_Spring2024/Data/Pro
write.csv(EPA.air.data.PM25.NC2018.selection.mutated, row.names = FALSE, file = "~/ENV 872/EDA_Spring202
write.csv(EPA.air.data.PM25.NC2019.selection.mutated, row.names = FALSE, file = "~/ENV 872/EDA_Spring202
```

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
colnames(EPA.air.data.O3.NC2018.selection)
```

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

```
colnames(EPA.air.data.O3.NC2019.selection)
```

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

```
colnames(EPA.air.data.PM25.NC2018.selection.mutated)
```

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

```
colnames(EPA.air.data.PM25.NC2019.selection.mutated)
```

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

```
EPA.air.data.O3andPM25.NC.2018and2019<-rbind(EPA.air.data.O3.NC2018.selection, EPA.air.data.O3.NC2019.s

colnames(EPA.air.data.O3andPM25.NC.2018and2019)
```

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

```
#8
EPA.air.data.O3andPM25.NC.2018and2019.selection<- EPA.air.data.O3andPM25.NC.2018and2019 %>%
 filter(Site.Name %in% c("Linville Falls", "Durham Armory", "Leggett", "Hattie Avenue", "Clemmons Middl
  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.
```

```
head(EPA.air.data.O3andPM25.NC.2018and2019.selection)
```

```
## # A tibble: 6 x 9
## # Groups:   Date, Site.Name, AQS_PARAMETER_DESC [6]
##   Date       Site.Name           AQS_PARAMETER_DESC COUNTY meanAQI meanlatitude
##   <date>     <fct>               <fct>              <fct>    <dbl>        <dbl>
## 1 2018-01-01 Bryson City         PM2.5              Swain       35         35.4
## 2 2018-01-01 Castle Hayne        PM2.5              New H~      13         34.4
## 3 2018-01-01 Clemmons Middle     PM2.5              Forsy~      24         36.0
## 4 2018-01-01 Durham Armory       PM2.5              Durham      31         36.0
## 5 2018-01-01 Garinger High School Ozone             Meckl~      32         35.2
## 6 2018-01-01 Garinger High School PM2.5             Meckl~      20         35.2
## # i 3 more variables: meanlongitude <dbl>, Month <dbl>, Year <dbl>
```

```
dim(EPA.air.data.O3andPM25.NC.2018and2019.selection)
```

```
## [1] 14752     9
```

```
#9
EPA.air.data.O3andPM25.NC.2018and2019.spreadAQI<- spread(EPA.air.data.O3andPM25.NC.2018and2019.selection
head(EPA.air.data.O3andPM25.NC.2018and2019.spreadAQI)
```

```
## # A tibble: 6 x 9
## # Groups:   Date, Site.Name [6]
##   Date       Site.Name COUNTY meanlatitude meanlongitude Month  Year Ozone PM2.5
##   <date>     <fct>     <fct>         <dbl>         <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 2018-01-01 Bryson C~ Swain          35.4         -83.4     1  2018    NA    35
## 2 2018-01-01 Castle H~ New H~         34.4         -77.8     1  2018    NA    13
## 3 2018-01-01 Clemmons~ Forsy~         36.0         -80.3     1  2018    NA    24
## 4 2018-01-01 Durham A~ Durham         36.0         -78.9     1  2018    NA    31
## 5 2018-01-01 Garinger~ Meckl~         35.2         -80.8     1  2018    32    20
## 6 2018-01-01 Hattie A~ Forsy~         36.1         -80.2     1  2018    NA    22
```

```
#10
dim(EPA.air.data.O3andPM25.NC.2018and2019.spreadAQI)
```

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

```
#the dimensions are 11921 by 9.

#11
write.csv(EPA.air.data.O3andPM25.NC.2018and2019.spreadAQI, row.names = FALSE, file = "./Data/Processed/
```

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
EPAair_data_O3_PM25_NC1819 <- read.csv(
  file=here("~/ENV 872/EDA_Spring2024/Data/Processed/EPAair_O3_PM25_NC1819_Processed.csv"),
  stringsAsFactors = TRUE)

colnames(EPAair_data_O3_PM25_NC1819)
```

```
## [1] "Date"          "Site.Name"     "COUNTY"        "meanlatitude"
## [5] "meanlongitude" "Month"         "Year"          "Ozone"
## [9] "PM2.5"
```

```
EPAair_df_O3_PM25_NC_summary<- EPAair_data_O3_PM25_NC1819 %>%
  group_by(Site.Name, Month, Year)%>%
  summarise(meanO3=mean(Ozone),
            meanPM2.5=mean(PM2.5))%>%
  drop_na(meanO3)
```

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

```
head(EPAair_df_O3_PM25_NC_summary)
```

```
## # A tibble: 6 x 5
## # Groups:   Site.Name, Month [4]
##   Site.Name   Month  Year meanO3 meanPM2.5
##   <fct>       <int> <int>  <dbl>     <dbl>
## 1 Bryson City     3  2018   41.6      34.7
## 2 Bryson City     3  2019   42.5      NA
## 3 Bryson City     4  2018   44.5      28.2
## 4 Bryson City     4  2019   45.4      26.7
## 5 Bryson City     5  2019   39.6      NA
## 6 Bryson City     6  2018   37.8      NA
```

```
#13
dim(EPAair_df_O3_PM25_NC_summary)
```

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

```
#The dimensions are 205 by 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.

```
#14

EPAair_df_O3_PM25_NC_summary<- EPAair_data_O3_PM25_NC1819 %>%
  group_by(Site.Name, Month, Year)%>%
  summarise(meanO3=mean(Ozone),
            meanPM2.5=mean(PM2.5))%>%
  na.omit(meanO3)
```

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

```
head(EPAair_df_O3_PM25_NC_summary)
```

```
## # A tibble: 6 x 5
## # Groups:   Site.Name, Month [5]
##   Site.Name   Month  Year meanO3 meanPM2.5
##   <fct>       <int> <int>  <dbl>     <dbl>
## 1 Bryson City     3  2018   41.6      34.7
## 2 Bryson City     4  2018   44.5      28.2
## 3 Bryson City     4  2019   45.4      26.7
## 4 Bryson City     7  2019   30.4      33.6
## 5 Bryson City     9  2018   25.4      25.1
## 6 Bryson City    10  2018   31        31.3
```

```
dim(EPAair_df_O3_PM25_NC_summary)
```

```
## [1] 101   5
```

```
#The dimensions are 45 by 5.
```

> Answer: The row dimensions decrease dramatically from 205 to 45. na.omit() removes rows with NA values. Many observations have NA mean PM2.5 values so many rows were removed.