Assignment 4: Data Wrangling

Hannah Nelson

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.

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

#1b
here()
```

```
## [1] "/Users/hannahnelson/Desktop/env872/EDA-Fall2023"
```

```
#1c
o3_18 <- read.csv(here("Data/Raw/EPAair_O3_NC2018_raw.csv"))

o3_19 <- read.csv(here("Data/Raw/EPAair_O3_NC2019_raw.csv"))
```

```
pm25_18 <- read.csv(here("Data/Raw/EPAair_PM25_NC2018_raw.csv"))

pm25_19 <- read.csv(here("Data/Raw/EPAair_PM25_NC2019_raw.csv"))

#2
glimpse(o3_18)
```

```
## Rows: 9,737
## Columns: 20
## $ Date                                    <chr> "03/01/2018", "03/02/2018", "03/0~
## $ Source                                  <chr> "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                                   <chr> "ppm", "ppm", "ppm", "ppm", "ppm"~
## $ DAILY_AQI_VALUE                         <int> 40, 43, 44, 45, 44, 28, 33, 41, 4~
## $ Site.Name                               <chr> "Taylorsville Liledoun", "Taylors~
## $ 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                      <chr> "Ozone", "Ozone", "Ozone", "Ozone~
## $ CBSA_CODE                               <int> 25860, 25860, 25860, 25860, 25860~
## $ CBSA_NAME                               <chr> "Hickory-Lenoir-Morganton, NC", "~
## $ STATE_CODE                              <int> 37, 37, 37, 37, 37, 37, 37, 37, 3~
## $ STATE                                   <chr> "North Carolina", "North Carolina~
## $ COUNTY_CODE                             <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ~
## $ COUNTY                                  <chr> "Alexander", "Alexander", "Alexan~
## $ SITE_LATITUDE                           <dbl> 35.9138, 35.9138, 35.9138, 35.913~
## $ SITE_LONGITUDE                          <dbl> -81.191, -81.191, -81.191, -81.19~
```

```
glimpse(o3_19)
```

```
## Rows: 10,592
## Columns: 20
## $ Date                                    <chr> "01/01/2019", "01/02/2019", "01/0~
## $ Source                                  <chr> "AirNow", "AirNow", "AirNow", "Ai~
## $ 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                                   <chr> "ppm", "ppm", "ppm", "ppm", "ppm"~
## $ DAILY_AQI_VALUE                         <int> 27, 17, 15, 20, 34, 34, 27, 35, 3~
## $ Site.Name                               <chr> "Taylorsville Liledoun", "Taylors~
## $ 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                      <chr> "Ozone", "Ozone", "Ozone", "Ozone~
## $ CBSA_CODE                               <int> 25860, 25860, 25860, 25860, 25860~
## $ CBSA_NAME                               <chr> "Hickory-Lenoir-Morganton, NC", "~
## $ STATE_CODE                              <int> 37, 37, 37, 37, 37, 37, 37, 37, 3~
## $ STATE                                   <chr> "North Carolina", "North Carolina~
## $ COUNTY_CODE                             <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ~
## $ COUNTY                                  <chr> "Alexander", "Alexander", "Alexan~
```

```
## $ SITE_LATITUDE                        <dbl> 35.9138, 35.9138, 35.9138, 35.913~
## $ SITE_LONGITUDE                       <dbl> -81.191, -81.191, -81.191, -81.19~
```

```
glimpse(pm25_18)
```

```
## Rows: 8,983
## Columns: 20
## $ Date                           <chr> "01/02/2018", "01/05/2018", "01/08/2018~
## $ Source                         <chr> "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                          <chr> "ug/m3 LC", "ug/m3 LC", "ug/m3 LC", "ug~
## $ DAILY_AQI_VALUE                <int> 12, 15, 22, 3, 10, 19, 8, 10, 18, 7, 24~
## $ Site.Name                      <chr> "Linville Falls", "Linville Falls", "Li~
## $ 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             <chr> "Acceptable PM2.5 AQI & Speciation Mass~
## $ CBSA_CODE                      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ CBSA_NAME                      <chr> "", "", "", "", "", "", "", "", "", "",~
## $ STATE_CODE                     <int> 37, 37, 37, 37, 37, 37, 37, 37, 37, 37,~
## $ STATE                          <chr> "North Carolina", "North Carolina", "No~
## $ COUNTY_CODE                    <int> 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,~
## $ COUNTY                         <chr> "Avery", "Avery", "Avery", "Avery", "Av~
## $ SITE_LATITUDE                  <dbl> 35.97235, 35.97235, 35.97235, 35.97235,~
## $ SITE_LONGITUDE                 <dbl> -81.93307, -81.93307, -81.93307, -81.93~
```

```
glimpse(pm25_19)
```

```
## Rows: 8,581
## Columns: 20
## $ Date                           <chr> "01/03/2019", "01/06/2019", "01/09/2019~
## $ Source                         <chr> "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                          <chr> "ug/m3 LC", "ug/m3 LC", "ug/m3 LC", "ug~
## $ DAILY_AQI_VALUE                <int> 7, 4, 5, 26, 11, 5, 6, 6, 15, 7, 14, 20~
## $ Site.Name                      <chr> "Linville Falls", "Linville Falls", "Li~
## $ 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             <chr> "Acceptable PM2.5 AQI & Speciation Mass~
## $ CBSA_CODE                      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ CBSA_NAME                      <chr> "", "", "", "", "", "", "", "", "", "",~
## $ STATE_CODE                     <int> 37, 37, 37, 37, 37, 37, 37, 37, 37, 37,~
## $ STATE                          <chr> "North Carolina", "North Carolina", "No~
## $ COUNTY_CODE                    <int> 11, 11, 11, 11, 11, 11, 11, 11, 11, 11,~
## $ COUNTY                         <chr> "Avery", "Avery", "Avery", "Avery", "Av~
## $ 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
o3_18$Date <- mdy(o3_18$Date)

head(o3_18)
```

```
##         Date Source   Site.ID POC Daily.Max.8.hour.Ozone.Concentration UNITS
## 1 2018-03-01    AQS 370030005   1                                 0.043   ppm
## 2 2018-03-02    AQS 370030005   1                                 0.046   ppm
## 3 2018-03-03    AQS 370030005   1                                 0.047   ppm
## 4 2018-03-04    AQS 370030005   1                                 0.049   ppm
## 5 2018-03-05    AQS 370030005   1                                 0.047   ppm
## 6 2018-03-06    AQS 370030005   1                                 0.030   ppm
##   DAILY_AQI_VALUE            Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              40 Taylorsville Liledoun              17              100
## 2              43 Taylorsville Liledoun              17              100
## 3              44 Taylorsville Liledoun              17              100
## 4              45 Taylorsville Liledoun              17              100
## 5              44 Taylorsville Liledoun              17              100
## 6              28 Taylorsville Liledoun              17              100
##   AQS_PARAMETER_CODE AQS_PARAMETER_DESC CBSA_CODE                    CBSA_NAME
## 1              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 2              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 3              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 4              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 5              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 6              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
##   STATE_CODE          STATE COUNTY_CODE    COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1         37 North Carolina           3 Alexander       35.9138        -81.191
## 2         37 North Carolina           3 Alexander       35.9138        -81.191
## 3         37 North Carolina           3 Alexander       35.9138        -81.191
## 4         37 North Carolina           3 Alexander       35.9138        -81.191
## 5         37 North Carolina           3 Alexander       35.9138        -81.191
## 6         37 North Carolina           3 Alexander       35.9138        -81.191
```

```
o3_19$Date <- mdy(o3_19$Date)

head(o3_19)
```

```
##         Date Source   Site.ID POC Daily.Max.8.hour.Ozone.Concentration UNITS
## 1 2019-01-01 AirNow 370030005   1                                 0.029   ppm
## 2 2019-01-02 AirNow 370030005   1                                 0.018   ppm
```

```
## 3 2019-01-03 AirNow 370030005   1                                0.016   ppm
## 4 2019-01-04 AirNow 370030005   1                                0.022   ppm
## 5 2019-01-05 AirNow 370030005   1                                0.037   ppm
## 6 2019-01-06 AirNow 370030005   1                                0.037   ppm
##   DAILY_AQI_VALUE            Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              27 Taylorsville Liledoun              24              100
## 2              17 Taylorsville Liledoun              24              100
## 3              15 Taylorsville Liledoun              24              100
## 4              20 Taylorsville Liledoun              24              100
## 5              34 Taylorsville Liledoun              24              100
## 6              34 Taylorsville Liledoun              24              100
##   AQS_PARAMETER_CODE AQS_PARAMETER_DESC CBSA_CODE                   CBSA_NAME
## 1              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 2              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 3              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 4              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 5              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
## 6              44201              Ozone     25860 Hickory-Lenoir-Morganton, NC
##   STATE_CODE          STATE COUNTY_CODE    COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1         37 North Carolina           3 Alexander       35.9138        -81.191
## 2         37 North Carolina           3 Alexander       35.9138        -81.191
## 3         37 North Carolina           3 Alexander       35.9138        -81.191
## 4         37 North Carolina           3 Alexander       35.9138        -81.191
## 5         37 North Carolina           3 Alexander       35.9138        -81.191
## 6         37 North Carolina           3 Alexander       35.9138        -81.191
```

```
pm25_18$Date <- mdy(pm25_18$Date)

head(pm25_18)
```

```
##         Date Source   Site.ID POC Daily.Mean.PM2.5.Concentration    UNITS
## 1 2018-01-02    AQS 370110002   1                            2.9 ug/m3 LC
## 2 2018-01-05    AQS 370110002   1                            3.7 ug/m3 LC
## 3 2018-01-08    AQS 370110002   1                            5.3 ug/m3 LC
## 4 2018-01-11    AQS 370110002   1                            0.8 ug/m3 LC
## 5 2018-01-14    AQS 370110002   1                            2.5 ug/m3 LC
## 6 2018-01-17    AQS 370110002   1                            4.5 ug/m3 LC
##   DAILY_AQI_VALUE      Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              12 Linville Falls               1               100
## 2              15 Linville Falls               1               100
## 3              22 Linville Falls               1               100
## 4               3 Linville Falls               1               100
## 5              10 Linville Falls               1               100
## 6              19 Linville Falls               1               100
##   AQS_PARAMETER_CODE                      AQS_PARAMETER_DESC CBSA_CODE CBSA_NAME
## 1              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 2              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 3              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 4              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 5              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 6              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
##   STATE_CODE          STATE COUNTY_CODE COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1         37 North Carolina          11  Avery      35.97235      -81.93307
## 2         37 North Carolina          11  Avery      35.97235      -81.93307
```

```
## 3         37 North Carolina          11  Avery      35.97235      -81.93307
## 4         37 North Carolina          11  Avery      35.97235      -81.93307
## 5         37 North Carolina          11  Avery      35.97235      -81.93307
## 6         37 North Carolina          11  Avery      35.97235      -81.93307
```

```
pm25_19$Date <- mdy(pm25_19$Date)

head(pm25_19)
```

```
##         Date Source   Site.ID POC Daily.Mean.PM2.5.Concentration    UNITS
## 1 2019-01-03    AQS 370110002   1                             1.6 ug/m3 LC
## 2 2019-01-06    AQS 370110002   1                             1.0 ug/m3 LC
## 3 2019-01-09    AQS 370110002   1                             1.3 ug/m3 LC
## 4 2019-01-12    AQS 370110002   1                             6.3 ug/m3 LC
## 5 2019-01-15    AQS 370110002   1                             2.6 ug/m3 LC
## 6 2019-01-18    AQS 370110002   1                             1.2 ug/m3 LC
##   DAILY_AQI_VALUE      Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1               7 Linville Falls               1              100
## 2               4 Linville Falls               1              100
## 3               5 Linville Falls               1              100
## 4              26 Linville Falls               1              100
## 5              11 Linville Falls               1              100
## 6               5 Linville Falls               1              100
##   AQS_PARAMETER_CODE                     AQS_PARAMETER_DESC CBSA_CODE CBSA_NAME
## 1              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 2              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 3              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 4              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 5              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
## 6              88502 Acceptable PM2.5 AQI & Speciation Mass        NA
##   STATE_CODE          STATE COUNTY_CODE COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1         37 North Carolina          11  Avery      35.97235      -81.93307
## 2         37 North Carolina          11  Avery      35.97235      -81.93307
## 3         37 North Carolina          11  Avery      35.97235      -81.93307
## 4         37 North Carolina          11  Avery      35.97235      -81.93307
## 5         37 North Carolina          11  Avery      35.97235      -81.93307
## 6         37 North Carolina          11  Avery      35.97235      -81.93307
```

```
#4
o3_18 <- o3_18 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

head(o3_18)
```

```
##         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
```

```
o3_19 <- o3_19 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

head(o3_19)
```

```
##         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
```

```
pm25_18 <- pm25_18 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

head(pm25_18)
```

```
##         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
```

```
pm25_19 <- pm25_19 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)

head(pm25_19)
```

```
##         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
pm25_18 <- pm25_18 %>%
  mutate(AQS_PARAMETER_DESC = "PM2.5")

head(pm25_18)
```

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

```
pm25_19 <- pm25_19 %>%
  mutate(AQS_PARAMETER_DESC = "PM2.5")

head(pm25_19)
```

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

```
#6
write.csv(o3_18, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2018_Processed.csv")

write.csv(o3_19, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2019_Processed.csv")

write.csv(pm25_18, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2018_Processed.csv")

write.csv(pm25_19, row.names = FALSE, file = "./Data/ProcessedEPAair_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
#combining into one data frame
o3_pm25_1819 <- rbind(o3_18, o3_19, pm25_18, pm25_19)

head(o3_pm25_1819)
```

```
##         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
```

```
#8
#filtering data to only include sites all four sets have in comon
o3_pm25_1819 <- o3_pm25_1819 %>%
  filter(Site.Name == "Linville Falls" | Site.Name == "Durham Armory" | Site.Name == "Leggett" | Site.Na

head(o3_pm25_1819)
```

```
##         Date DAILY_AQI_VALUE      Site.Name AQS_PARAMETER_DESC COUNTY
## 1 2018-03-01              42 Linville Falls              Ozone  Avery
## 2 2018-03-05              44 Linville Falls              Ozone  Avery
## 3 2018-03-06              38 Linville Falls              Ozone  Avery
## 4 2018-03-07              38 Linville Falls              Ozone  Avery
## 5 2018-03-08              41 Linville Falls              Ozone  Avery
## 6 2018-03-09              45 Linville Falls              Ozone  Avery
##   SITE_LATITUDE SITE_LONGITUDE
## 1      35.97235      -81.93307
## 2      35.97235      -81.93307
## 3      35.97235      -81.93307
## 4      35.97235      -81.93307
## 5      35.97235      -81.93307
## 6      35.97235      -81.93307
```

```
#grouping data & finding the mean for AQI value, latitude, and longitude
o3_pm25_1819 <- o3_pm25_1819 %>%
  group_by(Date, Site.Name, AQS_PARAMETER_DESC, COUNTY) %>%
  filter(!is.na(DAILY_AQI_VALUE) & !is.na(SITE_LATITUDE) & !is.na(SITE_LONGITUDE)) %>%
  summarise(daily_AQI_mean = mean(DAILY_AQI_VALUE),
            latitude_mean = mean(SITE_LATITUDE),
            longitude_mean = mean(SITE_LONGITUDE))
```

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

```
head(o3_pm25_1819)
```

```
## # A tibble: 6 x 7
## # Groups:   Date, Site.Name, AQS_PARAMETER_DESC [6]
##   Date       Site.Name    AQS_PARAMETER_DESC COUNTY daily_AQI_mean latitude_mean
##   <date>     <chr>        <chr>              <chr>           <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 Mi~ PM2.5              Forsy~             24          36.0
## 4 2018-01-01 Durham Armo~ PM2.5              Durham             31          36.0
## 5 2018-01-01 Garinger Hi~ Ozone              Meckl~             32          35.2
## 6 2018-01-01 Garinger Hi~ PM2.5              Meckl~             20          35.2
## # i 1 more variable: longitude_mean <dbl>
```

```r
#parsing date into three columns & renaming columns
o3_pm25_1819 <- o3_pm25_1819 %>%
  separate(Date, c("Y", "m", "d"))

head(o3_pm25_1819)
```

```
## # A tibble: 6 x 9
## # Groups:   Site.Name, AQS_PARAMETER_DESC [6]
##   Y     m     d     Site.Name           AQS_PARAMETER_DESC COUNTY daily_AQI_mean
##   <chr> <chr> <chr> <chr>               <chr>              <chr>           <dbl>
## 1 2018  01    01    Bryson City         PM2.5              Swain              35
## 2 2018  01    01    Castle Hayne        PM2.5              New H~             13
## 3 2018  01    01    Clemmons Middle     PM2.5              Forsy~             24
## 4 2018  01    01    Durham Armory       PM2.5              Durham             31
## 5 2018  01    01    Garinger High Scho~ Ozone              Meckl~             32
## 6 2018  01    01    Garinger High Scho~ PM2.5              Meckl~             20
## # i 2 more variables: latitude_mean <dbl>, longitude_mean <dbl>
```

```r
o3_pm25_1819 <- o3_pm25_1819 %>%
  rename(
    year = Y,
    month = m,
    day = d)


head(o3_pm25_1819)
```

```
## # A tibble: 6 x 9
## # Groups:   Site.Name, AQS_PARAMETER_DESC [6]
##   year  month day   Site.Name           AQS_PARAMETER_DESC COUNTY daily_AQI_mean
##   <chr> <chr> <chr> <chr>               <chr>              <chr>           <dbl>
## 1 2018  01    01    Bryson City         PM2.5              Swain              35
## 2 2018  01    01    Castle Hayne        PM2.5              New H~             13
## 3 2018  01    01    Clemmons Middle     PM2.5              Forsy~             24
## 4 2018  01    01    Durham Armory       PM2.5              Durham             31
## 5 2018  01    01    Garinger High Scho~ Ozone              Meckl~             32
## 6 2018  01    01    Garinger High Scho~ PM2.5              Meckl~             20
## # i 2 more variables: latitude_mean <dbl>, longitude_mean <dbl>
```

```r
#dimension of data set is 14,752 x 9
dim(o3_pm25_1819)
```

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

```r
#9
#separating AQI values for ozone and PM2.5 into two columns & naming columns
o3_pm25_1819 <- o3_pm25_1819 %>%
  pivot_wider(
    names_from = AQS_PARAMETER_DESC,
    values_from = daily_AQI_mean)

head(o3_pm25_1819)
```

```
## # A tibble: 6 x 9
## # Groups:   Site.Name [6]
##   year  month day   Site.Name    COUNTY latitude_mean longitude_mean PM2.5 Ozone
##   <chr> <chr> <chr> <chr>        <chr>          <dbl>          <dbl> <dbl> <dbl>
## 1 2018  01    01    Bryson City  Swain           35.4          -83.4    35    NA
## 2 2018  01    01    Castle Hayne New H~          34.4          -77.8    13    NA
## 3 2018  01    01    Clemmons Mi~ Forsy~          36.0          -80.3    24    NA
## 4 2018  01    01    Durham Armo~ Durham          36.0          -78.9    31    NA
## 5 2018  01    01    Garinger Hi~ Meckl~          35.2          -80.8    20    32
## 6 2018  01    01    Hattie Aven~ Forsy~          36.1          -80.2    22    NA
```

```
o3_pm25_1819 <- o3_pm25_1819 %>%
  rename(
    pm25_daily_AQI = PM2.5,
    o3_daily_AQI = Ozone)

head(o3_pm25_1819)
```

```
## # A tibble: 6 x 9
## # Groups:   Site.Name [6]
##   year  month day   Site.Name COUNTY latitude_mean longitude_mean pm25_daily_AQI
##   <chr> <chr> <chr> <chr>     <chr>          <dbl>          <dbl>          <dbl>
## 1 2018  01    01    Bryson C~ Swain           35.4          -83.4             35
## 2 2018  01    01    Castle H~ New H~          34.4          -77.8             13
## 3 2018  01    01    Clemmons~ Forsy~          36.0          -80.3             24
## 4 2018  01    01    Durham A~ Durham          36.0          -78.9             31
## 5 2018  01    01    Garinger~ Meckl~          35.2          -80.8             20
## 6 2018  01    01    Hattie A~ Forsy~          36.1          -80.2             22
## # i 1 more variable: o3_daily_AQI <dbl>
```

```
#10
#dimension of data set is 8,976 x 9
dim(o3_pm25_1819)
```

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

```
#11
write.csv(o3_pm25_1819, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2018_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
o3_pm25_1819_summary <- o3_pm25_1819 %>%
  group_by(Site.Name, month, year) %>%
```

```
  summarise(mean_o3_daily_AQI = mean(o3_daily_AQI),
            mean_pm25_daily_AQI = mean(pm25_daily_AQI)) %>%
 drop_na(mean_o3_daily_AQI)
```

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

```
head(o3_pm25_1819_summary)
```

```
## # A tibble: 6 x 5
## # Groups:   Site.Name, month [4]
##   Site.Name   month year  mean_o3_daily_AQI mean_pm25_daily_AQI
##   <chr>       <chr> <chr>              <dbl>               <dbl>
## 1 Bryson City 03    2018                41.6                34.7
## 2 Bryson City 03    2019                42.5                NA
## 3 Bryson City 04    2018                44.5                28.2
## 4 Bryson City 04    2019                45.4                26.7
## 5 Bryson City 05    2019                39.6                NA
## 6 Bryson City 06    2018                37.8                NA
```

```
#13
#dimension of data frame is 182 x 5
dim(o3_pm25_1819_summary)
```

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

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

    Answer: The drop_na function was used instead of the na.omit function becuase the drop_na function removed NAs just from the ozone column, while na.omit would remove the PM2.5 values from any rows with an NA value for ozone. This would have removed valuable data from the summary table.