Assignment 4: Data Wrangling

Ina Liao

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.

```
#install.packages("tidyverse")
#install.packages("lubridate")
#install.packages("here")
#install.packages("dplyr")
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(here)
```

```
## here() starts at /Users/inaliao/Desktop/EDE_Fall2023
```

```
library(dplyr)
```

1b. Check your working directory.

```
getwd()
```

```
## [1] "/Users/inaliao/Desktop/EDE_Fall2023"
```

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).

```
df_EPA_O3_2018<-read.csv("/Users/inaliao/Desktop/EDE_Fall2023/Data/Raw/EPA/EPAair_O3_NC2018_raw.csv",st
df_EPA_O3_2019<-read.csv("/Users/inaliao/Desktop/EDE_Fall2023/Data/Raw/EPA/EPAair_O3_NC2019_raw.csv",st
df_EPA_PM_2018<-read.csv("/Users/inaliao/Desktop/EDE_Fall2023/Data/Raw/EPA/EPAair_PM25_NC2018_raw.csv",
df_EPA_PM_2019<-read.csv("/Users/inaliao/Desktop/EDE_Fall2023/Data/Raw/EPA/EPAair_PM25_NC2019_raw.csv",
```

2. Apply the `glimpse()` function to reveal the dimensions, column names, and structure of each dataset.

```
#1a structure of each data set
glimpse(df_EPA_O3_2018)
```

```
## 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~
```

```
glimpse(df_EPA_O3_2019)
```

```
## 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(df_EPA_PM_2018)
```

```
## 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(df_EPA_PM_2019)
```

```
## 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~
```

```
#1b dimentions of each data set
dim(df_EPA_O3_2018)
```

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

```
dim(df_EPA_O3_2019)
```

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

```
dim(df_EPA_PM_2018)
```

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

```
dim(df_EPA_PM_2019)
```

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

```
#1c column names of each dataset
colnames(df_EPA_O3_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"
```

```
colnames(df_EPA_O3_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"
```

```
colnames(df_EPA_PM_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"
```

```
colnames(df_EPA_PM_2019)
```

```
##  [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"
```

```
#2 distribution of values in each column of the dataset.
summary(df_EPA_O3_2018)
```

```
##          Date      Source        Site.ID               POC
##  04/01/2018:  40   AQS:9737   Min.   :370030005   Min.   :1
##  04/12/2018:  40              1st Qu.:370650099   1st Qu.:1
##  04/13/2018:  40              Median :371010002   Median :1
##  04/14/2018:  40              Mean   :370969118   Mean   :1
##  04/15/2018:  40              3rd Qu.:371290002   3rd Qu.:1
##  04/18/2018:  40              Max.   :371990004   Max.   :1
##  (Other)   :9497
##  Daily.Max.8.hour.Ozone.Concentration UNITS      DAILY_AQI_VALUE
##  Min.   :0.00200                      ppm:9737   Min.   :  2.00
##  1st Qu.:0.03400                                 1st Qu.: 31.00
##  Median :0.04200                                 Median : 39.00
##  Mean   :0.04194                                 Mean   : 40.22
##  3rd Qu.:0.04900                                 3rd Qu.: 45.00
##  Max.   :0.07700                                 Max.   :122.00
##
##                Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
```

```
##  Coweeta             : 355   Min.   :12.00   Min.   : 71.00
##  Garinger High School: 354   1st Qu.:17.00   1st Qu.:100.00
##  Millbrook School    : 352   Median :17.00   Median :100.00
##  Candor              : 335   Mean   :16.94   Mean   : 99.65
##  Rockwell            : 335   3rd Qu.:17.00   3rd Qu.:100.00
##  Cranberry           : 323   Max.   :17.00   Max.   :100.00
##  (Other)             :7683
##  AQS_PARAMETER_CODE AQS_PARAMETER_DESC   CBSA_CODE
##  Min.   :44201      Ozone:9737         Min.   :11700
##  1st Qu.:44201                         1st Qu.:16740
##  Median :44201                         Median :24660
##  Mean   :44201                         Mean   :27247
##  3rd Qu.:44201                         3rd Qu.:39580
##  Max.   :44201                         Max.   :49180
##                                        NA's   :2609
##                              CBSA_NAME      STATE_CODE            STATE
##                                   :2609   Min.   :37   North Carolina:9737
##  Charlotte-Concord-Gastonia, NC-SC:1338   1st Qu.:37
##  Asheville, NC                    : 927   Median :37
##  Winston-Salem, NC                : 725   Mean   :37
##  Raleigh, NC                      : 585   3rd Qu.:37
##  Hickory-Lenoir-Morganton, NC     : 477   Max.   :37
##  (Other)                          :3076
##   COUNTY_CODE             COUNTY     SITE_LATITUDE   SITE_LONGITUDE
##  Min.   :  3.00   Forsyth    : 725   Min.   :34.36   Min.   :-83.80
##  1st Qu.: 65.00   Haywood    : 683   1st Qu.:35.26   1st Qu.:-82.05
##  Median :101.00   Mecklenburg: 592   Median :35.55   Median :-80.34
##  Mean   : 96.78   Avery      : 558   Mean   :35.62   Mean   :-80.42
##  3rd Qu.:129.00   Swain      : 483   3rd Qu.:36.03   3rd Qu.:-78.90
##  Max.   :199.00   Cumberland : 444   Max.   :36.31   Max.   :-76.62
##                   (Other)    :6252
```

```
summary(df_EPA_O3_2019)
```

```
##          Date          Source        Site.ID               POC
##  03/18/2019:   38   AirNow:2126   Min.   :370030005   Min.   :1
##  03/19/2019:   38   AQS   :8466   1st Qu.:370630015   1st Qu.:1
##  03/20/2019:   38                 Median :370870036   Median :1
##  03/23/2019:   38                 Mean   :370960317   Mean   :1
##  03/24/2019:   38                 3rd Qu.:371290002   3rd Qu.:1
##  03/25/2019:   38                 Max.   :371990004   Max.   :1
##  (Other)   :10364
##  Daily.Max.8.hour.Ozone.Concentration UNITS       DAILY_AQI_VALUE
##  Min.   :0.00000                      ppm:10592   Min.   :  0.0
##  1st Qu.:0.03600                                  1st Qu.: 33.0
##  Median :0.04400                                  Median : 41.0
##  Mean   :0.04331                                  Mean   : 41.2
##  3rd Qu.:0.05000                                  3rd Qu.: 46.0
##  Max.   :0.08100                                  Max.   :136.0
##
##                Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Garinger High School: 363   Min.   :13.00   Min.   : 75.00
##  Millbrook School    : 362   1st Qu.:17.00   1st Qu.:100.00
##  Coweeta             : 361   Median :17.00   Median :100.00
##  Rockwell            : 361   Mean   :18.34   Mean   : 99.69
```

```
##  Candor              : 358   3rd Qu.:17.00   3rd Qu.:100.00
##  Cranberry           : 351   Max.   :24.00   Max.   :100.00
##  (Other)             :8436
##  AQS_PARAMETER_CODE AQS_PARAMETER_DESC   CBSA_CODE
##  Min.   :44201      Ozone:10592        Min.   :11700
##  1st Qu.:44201                         1st Qu.:16740
##  Median :44201                         Median :24660
##  Mean   :44201                         Mean   :26617
##  3rd Qu.:44201                         3rd Qu.:37080
##  Max.   :44201                         Max.   :49180
##                                        NA's   :2852
##                               CBSA_NAME      STATE_CODE            STATE
##                                   :2852   Min.   :37   North Carolina:10592
##  Charlotte-Concord-Gastonia, NC-SC:1590   1st Qu.:37
##  Asheville, NC                    :1114   Median :37
##  Winston-Salem, NC                : 735   Mean   :37
##  Raleigh, NC                      : 646   3rd Qu.:37
##  Hickory-Lenoir-Morganton, NC     : 567   Max.   :37
##  (Other)                          :3088
##   COUNTY_CODE            COUNTY     SITE_LATITUDE   SITE_LONGITUDE
##  Min.   :  3.0   Haywood    : 864   Min.   :34.36   Min.   :-83.80
##  1st Qu.: 63.0   Forsyth    : 735   1st Qu.:35.26   1st Qu.:-82.05
##  Median : 87.0   Mecklenburg: 657   Median :35.59   Median :-80.34
##  Mean   : 95.9   Avery      : 607   Mean   :35.61   Mean   :-80.41
##  3rd Qu.:129.0   Cumberland : 498   3rd Qu.:36.03   3rd Qu.:-78.77
##  Max.   :199.0   Swain      : 476   Max.   :36.31   Max.   :-76.62
##                  (Other)    :6755
```

```
summary(df_EPA_PM_2018)
```

```
##          Date      Source        Site.ID               POC
##  01/26/2018:  40   AQS:8983   Min.   :370110002   Min.   :1.000
##  02/01/2018:  40              1st Qu.:370630015   1st Qu.:3.000
##  02/19/2018:  40              Median :371010002   Median :3.000
##  03/21/2018:  40              Mean   :371002405   Mean   :2.812
##  04/02/2018:  40              3rd Qu.:371230001   3rd Qu.:3.000
##  04/08/2018:  40              Max.   :371830021   Max.   :5.000
##  (Other)   :8743
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-2.300                ug/m3 LC:8983   Min.   : 0.00
##  1st Qu.: 4.900                                1st Qu.:20.00
##  Median : 7.000                                Median :29.00
##  Mean   : 7.491                                Mean   :30.73
##  3rd Qu.: 9.700                                3rd Qu.:40.00
##  Max.   :34.200                                Max.   :97.00
##
##                Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Millbrook School    : 717   Min.   :1       Min.   :100
##  Hattie Avenue       : 510   1st Qu.:1       1st Qu.:100
##  Board Of Ed. Bldg.  : 477   Median :1       Median :100
##  Garinger High School: 472   Mean   :1       Mean   :100
##  Durham Armory       : 466   3rd Qu.:1       3rd Qu.:100
##  Pitt Agri. Center   : 460   Max.   :1       Max.   :100
##  (Other)             :5881
##  AQS_PARAMETER_CODE                          AQS_PARAMETER_DESC
```

```
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:1403
##  1st Qu.:88101      PM2.5 - Local Conditions              :7580
##  Median :88101
##  Mean   :88164
##  3rd Qu.:88101
##  Max.   :88502
##
##    CBSA_CODE                                  CBSA_NAME      STATE_CODE
##  Min.   :11700   Raleigh, NC                       :1396   Min.   :37
##  1st Qu.:19000   Winston-Salem, NC                 :1316   1st Qu.:37
##  Median :25860   Charlotte-Concord-Gastonia, NC-SC:1275   Median :37
##  Mean   :30946                                     :1263   Mean   :37
##  3rd Qu.:40580   Asheville, NC                     : 586   3rd Qu.:37
##  Max.   :49180   Durham-Chapel Hill, NC            : 466   Max.   :37
##  NA's   :1263    (Other)                           :2681
##             STATE       COUNTY_CODE            COUNTY     SITE_LATITUDE
##  North Carolina:8983   Min.   : 11.0   Mecklenburg:1275   Min.   :34.36
##                        1st Qu.: 63.0   Wake       :1049   1st Qu.:35.26
##                        Median :101.0   Forsyth    : 876   Median :35.64
##                        Mean   :100.2   Buncombe   : 477   Mean   :35.61
##                        3rd Qu.:123.0   Durham     : 466   3rd Qu.:35.91
##                        Max.   :183.0   Pitt       : 460   Max.   :36.11
##                                        (Other)    :4380
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-79.99
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

```
summary(df_EPA_PM_2019)
```

```
##          Date         Source        Site.ID               POC
##  02/26/2019:  41   AirNow:1670   Min.   :370110002   Min.   :1.000
##  01/21/2019:  40   AQS   :6911   1st Qu.:370630015   1st Qu.:3.000
##  02/14/2019:  40                 Median :371190041   Median :3.000
##  01/09/2019:  39                 Mean   :371023743   Mean   :3.032
##  01/27/2019:  39                 3rd Qu.:371290002   3rd Qu.:3.000
##  02/02/2019:  39                 Max.   :371830021   Max.   :5.000
##  (Other)   :8343
##  Daily.Mean.PM2.5.Concentration      UNITS      DAILY_AQI_VALUE
##  Min.   :-3.100                ug/m3 LC:8581   Min.   : 0.00
##  1st Qu.: 4.900                                1st Qu.:20.00
##  Median : 7.400                                Median :31.00
##  Mean   : 7.684                                Mean   :31.51
##  3rd Qu.:10.100                                3rd Qu.:42.00
##  Max.   :31.200                                Max.   :91.00
##
##                Site.Name    DAILY_OBS_COUNT PERCENT_COMPLETE
##  Millbrook School    : 738   Min.   :1       Min.   :100
##  Garinger High School: 629   1st Qu.:1       1st Qu.:100
##  Remount             : 573   Median :1       Median :100
##  Hickory Water Tower : 518   Mean   :1       Mean   :100
```

```
##  Hattie Avenue       : 436   3rd Qu.:1       3rd Qu.:100
##  Durham Armory       : 431   Max.   :1       Max.   :100
##  (Other)             :5256
##  AQS_PARAMETER_CODE                              AQS_PARAMETER_DESC
##  Min.   :88101      Acceptable PM2.5 AQI & Speciation Mass:1029
##  1st Qu.:88101      PM2.5 - Local Conditions              :7552
##  Median :88101
##  Mean   :88149
##  3rd Qu.:88101
##  Max.   :88502
##
##    CBSA_CODE                                    CBSA_NAME      STATE_CODE
##  Min.   :11700   Raleigh, NC                         :1441   Min.   :37
##  1st Qu.:19000   Charlotte-Concord-Gastonia, NC-SC:1379   1st Qu.:37
##  Median :25860   Winston-Salem, NC                   :1235   Median :37
##  Mean   :31099                                       :1058   Mean   :37
##  3rd Qu.:40580   Hickory-Lenoir-Morganton, NC        : 518   3rd Qu.:37
##  Max.   :49180   Durham-Chapel Hill, NC              : 431   Max.   :37
##  NA's   :1058    (Other)                             :2519
##             STATE       COUNTY_CODE             COUNTY     SITE_LATITUDE
##  North Carolina:8581   Min.   : 11.0   Mecklenburg:1379   Min.   :34.36
##                        1st Qu.: 63.0   Wake        :1083   1st Qu.:35.26
##                        Median :119.0   Forsyth     : 839   Median :35.73
##                        Mean   :102.4   Catawba     : 518   Mean   :35.63
##                        3rd Qu.:129.0   Durham      : 431   3rd Qu.:35.91
##                        Max.   :183.0   Cumberland  : 427   Max.   :36.51
##                                        (Other)     :3904
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-79.95
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

Wrangle individual datasets to create processed files.

3. Change the Date columns to be date objects.

```
df_EPA_O3_2018$Date<-as.Date(df_EPA_O3_2018$Date,format="%m/%d/%Y")
df_EPA_O3_2019$Date<-as.Date(df_EPA_O3_2019$Date,format="%m/%d/%Y")
df_EPA_PM_2018$Date<-as.Date(df_EPA_PM_2018$Date,format="%m/%d/%Y")
df_EPA_PM_2019$Date<-as.Date(df_EPA_PM_2019$Date,format="%m/%d/%Y")
```

4. Select the following columns: Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE

```
#select specific columns and create another subset
df_EPA_O3_2018_subset<-df_EPA_O3_2018 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
head(df_EPA_O3_2018_subset)
```

```
##         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
```

```
df_EPA_O3_2019_subset<-df_EPA_O3_2019 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
head(df_EPA_O3_2019_subset)
```

```
##         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
```

```
df_EPA_PM_2018_subset<-df_EPA_PM_2018 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
head(df_EPA_PM_2018_subset)
```

```
##         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
```

```
df_EPA_PM_2019_subset<-df_EPA_PM_2019 %>%
  select(Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_DESC, COUNTY, SITE_LATITUDE, SITE_LONGITUDE)
head(df_EPA_PM_2019_subset)
```

```
##         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. For the PM2.5 datasets, fill all cells in AQS_PARAMETER_DESC with "PM2.5" (all cells in this column should be identical).

```
PM_replace_AQS<-function(x){
  ifelse(x=="Acceptable PM2.5 AQI & Speciation Mass","PM2.5","PM2.5")
}
df_EPA_PM_2018_subset$AQS_PARAMETER_DESC<-PM_replace_AQS(df_EPA_PM_2018_subset$AQS_PARAMETER_DESC)
df_EPA_PM_2019_subset$AQS_PARAMETER_DESC<-PM_replace_AQS(df_EPA_PM_2019_subset$AQS_PARAMETER_DESC)
head(df_EPA_PM_2018_subset)
```

```
##         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
```

```
head(df_EPA_PM_2019_subset)
```

```
##         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. Save all four processed datasets in the Processed folder. Use the same file names as the raw files but

replace "raw" with "processed".

```
#3
write.csv(df_EPA_O3_2018_subset, file="/Users/inaliao/Desktop/EDE_Fall2023/Data/Processed/EPAair_O3_NC2

#4
write.csv(df_EPA_O3_2019_subset, file="/Users/inaliao/Desktop/EDE_Fall2023/Data/Processed/EPAair_O3_NC2

#5
write.csv(df_EPA_PM_2018_subset, file="/Users/inaliao/Desktop/EDE_Fall2023/Data/Processed/EPAair_PM_NC2

#6
write.csv(df_EPA_PM_2019_subset, file="/Users/inaliao/Desktop/EDE_Fall2023/Data/Processed/EPAair_PM_NC2
```

Combine datasets

7. Combine the four datasets with `rbind`. Make sure your column names are identical prior to running this code.

```
#combine data frame based on AIQ parameters
#combine the original data sets rather than subsets
df_EPA_O3<-rbind(df_EPA_O3_2018,df_EPA_O3_2019)
df_EPA_PM <- rbind(df_EPA_PM_2018,df_EPA_PM_2019)

#check if the column names are identical
colnames(df_EPA_O3)
```

```
##  [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"
```

```
colnames(df_EPA_PM)
```

```
##  [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"
```

```
head(df_EPA_PM);tail(df_EPA_PM)
```

```
##         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
```

```
##             Date Source   Site.ID POC Daily.Mean.PM2.5.Concentration    UNITS
## 17559 2019-12-26 AirNow 371830021   3                             9.2 ug/m3 LC
## 17560 2019-12-27 AirNow 371830021   3                            11.5 ug/m3 LC
## 17561 2019-12-28 AirNow 371830021   3                             9.9 ug/m3 LC
## 17562 2019-12-29 AirNow 371830021   3                             6.5 ug/m3 LC
## 17563 2019-12-30 AirNow 371830021   3                             3.6 ug/m3 LC
## 17564 2019-12-31 AirNow 371830021   3                             4.3 ug/m3 LC
##       DAILY_AQI_VALUE  Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 17559              38 Triple Oak               1              100
## 17560              48 Triple Oak               1              100
## 17561              41 Triple Oak               1              100
## 17562              27 Triple Oak               1              100
## 17563              15 Triple Oak               1              100
## 17564              18 Triple Oak               1              100
##       AQS_PARAMETER_CODE       AQS_PARAMETER_DESC CBSA_CODE   CBSA_NAME
## 17559              88101 PM2.5 - Local Conditions     39580 Raleigh, NC
## 17560              88101 PM2.5 - Local Conditions     39580 Raleigh, NC
## 17561              88101 PM2.5 - Local Conditions     39580 Raleigh, NC
## 17562              88101 PM2.5 - Local Conditions     39580 Raleigh, NC
```

```
## 17563              88101 PM2.5 - Local Conditions     39580 Raleigh, NC
## 17564              88101 PM2.5 - Local Conditions     39580 Raleigh, NC
##       STATE_CODE          STATE COUNTY_CODE COUNTY SITE_LATITUDE SITE_LONGITUDE
## 17559         37 North Carolina         183   Wake       35.8652       -78.8197
## 17560         37 North Carolina         183   Wake       35.8652       -78.8197
## 17561         37 North Carolina         183   Wake       35.8652       -78.8197
## 17562         37 North Carolina         183   Wake       35.8652       -78.8197
## 17563         37 North Carolina         183   Wake       35.8652       -78.8197
## 17564         37 North Carolina         183   Wake       35.8652       -78.8197
```

```
#rename columns names so PM2.5 and O3 data frames could be combined
colnames(df_EPA_O3)[colnames(df_EPA_O3)=="Daily.Max.8.hour.Ozone.Concentration"]<-"Daily.Max.8.hour Conc
colnames(df_EPA_PM)[colnames(df_EPA_PM)=="Daily.Mean.PM2.5.Concentration"]<-"Daily.Max.8.hour Concentra
colnames(df_EPA_O3)
```

```
##  [1] "Date"                           "Source"
##  [3] "Site.ID"                        "POC"
##  [5] "Daily.Max.8.hour 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"
```

```
colnames(df_EPA_PM)
```

```
##  [1] "Date"                           "Source"
##  [3] "Site.ID"                        "POC"
##  [5] "Daily.Max.8.hour 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"
```

```
#replace cells values: fill all cells in AQS_PARAMETER_DESC with "PM2.5"
#use previous defined function, PM_replace_AQS
df_EPA_PM$AQS_PARAMETER_DESC<-PM_replace_AQS(df_EPA_PM$AQS_PARAMETER_DESC)
head(df_EPA_PM)
```

```
##         Date Source   Site.ID POC Daily.Max.8.hour 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 STATE_CODE
## 1              88502              PM2.5        NA                  37
## 2              88502              PM2.5        NA                  37
## 3              88502              PM2.5        NA                  37
## 4              88502              PM2.5        NA                  37
## 5              88502              PM2.5        NA                  37
## 6              88502              PM2.5        NA                  37
##            STATE COUNTY_CODE COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1 North Carolina          11  Avery      35.97235      -81.93307
## 2 North Carolina          11  Avery      35.97235      -81.93307
## 3 North Carolina          11  Avery      35.97235      -81.93307
## 4 North Carolina          11  Avery      35.97235      -81.93307
## 5 North Carolina          11  Avery      35.97235      -81.93307
## 6 North Carolina          11  Avery      35.97235      -81.93307
```

```
df_EPA<-rbind(df_EPA_O3,df_EPA_PM)
head(df_EPA);tail(df_EPA)
```

```
##         Date Source   Site.ID POC Daily.Max.8.hour 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
```

```
##             Date Source   Site.ID POC Daily.Max.8.hour Concentration    UNITS
## 37888 2019-12-26 AirNow 371830021   3                         9.2 ug/m3 LC
## 37889 2019-12-27 AirNow 371830021   3                        11.5 ug/m3 LC
## 37890 2019-12-28 AirNow 371830021   3                         9.9 ug/m3 LC
## 37891 2019-12-29 AirNow 371830021   3                         6.5 ug/m3 LC
## 37892 2019-12-30 AirNow 371830021   3                         3.6 ug/m3 LC
```

```
## 37893 2019-12-31 AirNow 371830021   3                          4.3 ug/m3 LC
##       DAILY_AQI_VALUE  Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 37888              38 Triple Oak               1              100
## 37889              48 Triple Oak               1              100
## 37890              41 Triple Oak               1              100
## 37891              27 Triple Oak               1              100
## 37892              15 Triple Oak               1              100
## 37893              18 Triple Oak               1              100
##       AQS_PARAMETER_CODE AQS_PARAMETER_DESC CBSA_CODE   CBSA_NAME STATE_CODE
## 37888              88101              PM2.5     39580 Raleigh, NC         37
## 37889              88101              PM2.5     39580 Raleigh, NC         37
## 37890              88101              PM2.5     39580 Raleigh, NC         37
## 37891              88101              PM2.5     39580 Raleigh, NC         37
## 37892              88101              PM2.5     39580 Raleigh, NC         37
## 37893              88101              PM2.5     39580 Raleigh, NC         37
##                STATE COUNTY_CODE COUNTY SITE_LATITUDE SITE_LONGITUDE
## 37888 North Carolina         183   Wake       35.8652       -78.8197
## 37889 North Carolina         183   Wake       35.8652       -78.8197
## 37890 North Carolina         183   Wake       35.8652       -78.8197
## 37891 North Carolina         183   Wake       35.8652       -78.8197
## 37892 North Carolina         183   Wake       35.8652       -78.8197
## 37893 North Carolina         183   Wake       35.8652       -78.8197
```

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.

```
df_EPA_subset<-df_EPA %>%
  #filter sites names
  filter(Site.Name=="Linville Falls" | Site.Name=="Durham Armory"| Site.Name=="Leggett"| Site.Name=="Ha
  #calculate mean AIQ, mean latitude, and mean longitude
  group_by(Date,Site.Name,AQS_PARAMETER_DESC,COUNTY) %>%
  summarize(Mean_AQI = mean(DAILY_AQI_VALUE),
            Mean_Latitude = mean(SITE_LATITUDE),
            Mean_Longitude = mean(SITE_LONGITUDE))%>%
  mutate(Month=month(Date),Year=year(Date))
```

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

```
head(df_EPA_subset)
```

```
## # A tibble: 6 x 9
## # Groups:   Date, Site.Name, AQS_PARAMETER_DESC [6]
##   Date       Site.Name          AQS_PARAMETER_DESC COUNTY Mean_AQI Mean_Latitude
##   <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 Sch~ Ozone              Meckl~       32          35.2
## 6 2018-01-01 Garinger High Sch~ PM2.5              Meckl~       20          35.2
## # i 3 more variables: Mean_Longitude <dbl>, Month <dbl>, Year <dbl>
```

```
dim(df_EPA_subset)
```

```
## [1] 14752     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.

```
#seperate mean AIQ values based on AIQ parameters
df_EPA_subset_spread<- df_EPA_subset %>%
  pivot_wider(names_from="AQS_PARAMETER_DESC",values_from="Mean_AQI")

#rename columns' names
colnames(df_EPA_subset_spread)[colnames(df_EPA_subset_spread)=="PM2.5"]<-"Mean_AQI_PM2.5"
colnames(df_EPA_subset_spread)[colnames(df_EPA_subset_spread)=="Ozone"]<-"Mean_AQI_O3"
```

```
df_EPA_subset_spread<- df_EPA_subset_spread %>%
  group_by(Date,COUNTY)
head(df_EPA_subset_spread);tail(df_EPA_subset_spread)
```

```
## # A tibble: 6 x 9
## # Groups:   Date, COUNTY [5]
##   Date       Site.Name           COUNTY Mean_Latitude Mean_Longitude Month  Year
##   <date>     <fct>               <fct>          <dbl>          <dbl> <dbl> <dbl>
## 1 2018-01-01 Bryson City         Swain           35.4          -83.4     1  2018
## 2 2018-01-01 Castle Hayne        New H~          34.4          -77.8     1  2018
## 3 2018-01-01 Clemmons Middle     Forsy~          36.0          -80.3     1  2018
## 4 2018-01-01 Durham Armory       Durham          36.0          -78.9     1  2018
## 5 2018-01-01 Garinger High Scho~ Meckl~          35.2          -80.8     1  2018
## 6 2018-01-01 Hattie Avenue       Forsy~          36.1          -80.2     1  2018
## # i 2 more variables: Mean_AQI_PM2.5 <dbl>, Mean_AQI_O3 <dbl>
```

```
## # A tibble: 6 x 9
## # Groups:   Date, COUNTY [6]
##   Date       Site.Name         COUNTY   Mean_Latitude Mean_Longitude Month  Year
##   <date>     <fct>             <fct>            <dbl>          <dbl> <dbl> <dbl>
## 1 2019-12-31 Hattie Avenue     Forsyth           36.1          -80.2    12  2019
## 2 2019-12-31 Leggett           Edgecom~          36.0          -77.6    12  2019
## 3 2019-12-31 Mendenhall School Guilford          36.1          -79.8    12  2019
## 4 2019-12-31 Millbrook School  Wake              35.9          -78.6    12  2019
## 5 2019-12-31 Pitt Agri. Center Pitt              35.6          -77.4    12  2019
## 6 2019-12-31 West Johnston Co. Johnston          35.6          -78.5    12  2019
## # i 2 more variables: Mean_AQI_PM2.5 <dbl>, Mean_AQI_O3 <dbl>
```

10. Call up the dimensions of your new tidy dataset.

```
dim(df_EPA_subset_spread)
```

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

11. Save your processed dataset with the following file name: "EPAair_O3_PM25_NC1819_Processed.csv"

```
#7 O3, 2018 & 2019
write.csv(df_EPA_O3, file="/Users/inaliao/Desktop/EDE_Fall2023/Data/Processed/EPAair_O3_NC1819_Processe

#8 PM, 2018 & 2019
write.csv(df_EPA_PM, file="/Users/inaliao/Desktop/EDE_Fall2023/Data/Processed/EPAair_PM_NC1819_Processe

#9 O3 & PM, 2018 & 2019
write.csv(df_EPA, file="/Users/inaliao/Desktop/EDE_Fall2023/Data/Processed/EPAair_O3_PM_NC1819_Processe

#10 group by dates and location, generate daily mean AIQ values
write.csv(df_EPA_subset, file="/Users/inaliao/Desktop/EDE_Fall2023/Data/Processed/EPAair_O3_PM_NC1819_P

#11 Spread AIQ values for ozone and PM2.5
write.csv(df_EPA_subset_spread, file="/Users/inaliao/Desktop/EDE_Fall2023/Data/Processed/EPAair_O3_PM25
```

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
df_EPA_subset_summary <- df_EPA_subset_spread %>%
  group_by(Site.Name, Month,Year) %>%
  summarize(Mean_AQI_PM2.5 = mean(Mean_AQI_PM2.5),
            Mean_AQI_O3 =mean(Mean_AQI_O3),
            Mean_Latitude = mean(Mean_Latitude),
            Mean_Longitude = mean(Mean_Longitude)) %>%
  drop_na(Mean_AQI_O3)
```

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

```
head(df_EPA_subset_summary)
```

```
## # A tibble: 6 x 7
## # Groups:   Site.Name, Month [4]
##   Site.Name  Month  Year Mean_AQI_PM2.5 Mean_AQI_O3 Mean_Latitude Mean_Longitude
##   <fct>      <dbl> <dbl>          <dbl>       <dbl>         <dbl>          <dbl>
## 1 Bryson Ci~     3  2018           34.7        41.6          35.4          -83.4
## 2 Bryson Ci~     3  2019           NA          42.5          35.4          -83.4
## 3 Bryson Ci~     4  2018           28.2        44.5          35.4          -83.4
## 4 Bryson Ci~     4  2019           26.7        45.4          35.4          -83.4
## 5 Bryson Ci~     5  2019           NA          39.6          35.4          -83.4
## 6 Bryson Ci~     6  2018           NA          37.8          35.4          -83.4
```

```
#13
dim(df_EPA_subset_summary)
```

```
## [1] 182   7
```

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

    Answer: na.omit will remove rows with any missing values, and thus, we can not keep the rows

with missing mean PM2.5 values.