Assignment 4: Data Wrangling

Shubhangi Gupta

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A04_DataWrangling.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. Ensure that code in code chunks does not extend off the page in the PDF.

Set up your session

1a. Load the `tidyverse`, `lubridate`, and `here` packages into your session.

1b. Check your working directory.

1c. Read in all four raw data files associated with the EPA Air dataset, being sure to set string columns to be read in a factors. See the README file for the EPA air datasets for more information (especially if you have not worked with air quality data previously).

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

```
#1b
setwd("~/RStudio Project Folder/EDA_Spring2024")
getwd()
```

```
## [1] "/home/guest/RStudio Project Folder/EDA_Spring2024"
```

```
#1c
EPAair_O3_NC2018_raw <- read.csv("Data/Raw/EPAair_O3_NC2018_raw.csv", stringsAsFactors = TRUE)
EPAair_O3_NC2019_raw <- read.csv("Data/Raw/EPAair_O3_NC2019_raw.csv", stringsAsFactors = TRUE)
EPAair_PM25_NC2018_raw <- read.csv("Data/Raw/EPAair_PM25_NC2018_raw.csv", stringsAsFactors = TRUE)
EPAair_PM25_NC2019_raw <- read.csv("Data/Raw/EPAair_PM25_NC2019_raw.csv", stringsAsFactors = TRUE)
```

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

```
#2
glimpse(EPAair_O3_NC2018_raw)
```

```
## 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(EPAair_O3_NC2019_raw)
```

```
## 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(EPAair_PM25_NC2018_raw)
```

```
## 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(EPAair_PM25_NC2019_raw)
```

```
## 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.

```
EPAair_O3_NC2018_raw$Date <- as.Date(EPAair_O3_NC2018_raw$Date)
class(EPAair_O3_NC2018_raw$Date)
```

```
## [1] "Date"
```

```
EPAair_O3_NC2019_raw$Date <- as.Date(EPAair_O3_NC2019_raw$Date)
class(EPAair_O3_NC2019_raw$Date)
```

```
## [1] "Date"
```

```
EPAair_PM25_NC2018_raw$Date <- as.Date(EPAair_PM25_NC2018_raw$Date)
class(EPAair_PM25_NC2018_raw$Date)
```

```
## [1] "Date"
```

```
EPAair_PM25_NC2019_raw$Date <- as.Date(EPAair_PM25_NC2019_raw$Date)
class(EPAair_PM25_NC2019_raw$Date)
```

```
## [1] "Date"
```

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

```
EPAair_O3_NC2018_processed <- EPAair_O3_NC2018_raw[,c("Date", "DAILY_AQI_VALUE",
"Site.Name", "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE")]
head(EPAair_O3_NC2018_processed)
```

```
##      Date DAILY_AQI_VALUE             Site.Name AQS_PARAMETER_DESC    COUNTY
## 1 3-01-20              40 Taylorsville Liledoun              Ozone Alexander
## 2 3-02-20              43 Taylorsville Liledoun              Ozone Alexander
## 3 3-03-20              44 Taylorsville Liledoun              Ozone Alexander
## 4 3-04-20              45 Taylorsville Liledoun              Ozone Alexander
## 5 3-05-20              44 Taylorsville Liledoun              Ozone Alexander
## 6 3-06-20              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
```

```
EPAair_O3_NC2019_processed <- EPAair_O3_NC2019_raw[,c("Date", "DAILY_AQI_VALUE",
"Site.Name", "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE")]
head(EPAair_O3_NC2019_processed)
```

```
##      Date DAILY_AQI_VALUE             Site.Name AQS_PARAMETER_DESC    COUNTY
## 1 1-01-20              27 Taylorsville Liledoun              Ozone Alexander
## 2 1-02-20              17 Taylorsville Liledoun              Ozone Alexander
```

```
## 3 1-03-20              15 Taylorsville Liledoun              Ozone Alexander
## 4 1-04-20              20 Taylorsville Liledoun              Ozone Alexander
## 5 1-05-20              34 Taylorsville Liledoun              Ozone Alexander
## 6 1-06-20              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
```

```
EPAair_PM25_NC2018_processed <- EPAair_PM25_NC2018_raw[,c("Date", "DAILY_AQI_VALUE",
"Site.Name", "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE")]
head(EPAair_PM25_NC2018_processed)
```

```
##      Date DAILY_AQI_VALUE      Site.Name                      AQS_PARAMETER_DESC
## 1 1-02-20              12 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
## 2 1-05-20              15 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
## 3 1-08-20              22 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
## 4 1-11-20               3 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
## 5    <NA>              10 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
## 6    <NA>              19 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
##   COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1  Avery      35.97235      -81.93307
## 2  Avery      35.97235      -81.93307
## 3  Avery      35.97235      -81.93307
## 4  Avery      35.97235      -81.93307
## 5  Avery      35.97235      -81.93307
## 6  Avery      35.97235      -81.93307
```

```
EPAair_PM25_NC2019_processed <- EPAair_PM25_NC2019_raw[,c("Date", "DAILY_AQI_VALUE",
"Site.Name", "AQS_PARAMETER_DESC", "COUNTY", "SITE_LATITUDE", "SITE_LONGITUDE")]
head(EPAair_PM25_NC2019_processed)
```

```
##      Date DAILY_AQI_VALUE      Site.Name                      AQS_PARAMETER_DESC
## 1 1-03-20               7 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
## 2 1-06-20               4 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
## 3 1-09-20               5 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
## 4 1-12-20              26 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
## 5    <NA>              11 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
## 6    <NA>               5 Linville Falls Acceptable PM2.5 AQI & Speciation Mass
##   COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1  Avery      35.97235      -81.93307
## 2  Avery      35.97235      -81.93307
## 3  Avery      35.97235      -81.93307
## 4  Avery      35.97235      -81.93307
## 5  Avery      35.97235      -81.93307
## 6  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).

```
EPAair_PM25_NC2019_processed$AQS_PARAMETER_DESC <- "PM2.5"
count(EPAair_PM25_NC2019_processed, EPAair_PM25_NC2019_processed$AQS_PARAMETER_DESC)
```

```
##   EPAair_PM25_NC2019_processed$AQS_PARAMETER_DESC    n
## 1                                          PM2.5 8581
```

```
EPAair_PM25_NC2018_processed$AQS_PARAMETER_DESC <- "PM2.5"
count(EPAair_PM25_NC2018_processed, EPAair_PM25_NC2018_processed$AQS_PARAMETER_DESC)
```

```
##   EPAair_PM25_NC2018_processed$AQS_PARAMETER_DESC    n
## 1                                          PM2.5 8983
```

6. Save all four processed datasets in the Processed folder. Use the same file names as the raw files but replace "raw" with "processed".

```
write.csv(EPAair_O3_NC2018_processed, row.names = FALSE, file =
            "./Data/Processed/EPAair_O3_NC2018_processed.csv")
write.csv(EPAair_O3_NC2019_processed, row.names = FALSE, file =
            "./Data/Processed/EPAair_O3_NC2019_processed.csv")
write.csv(EPAair_PM25_NC2018_processed, row.names = FALSE, file =
            "./Data/Processed/EPAair_PM25_NC2018_processed.csv")
write.csv(EPAair_PM25_NC2019_processed, row.names = FALSE, file =
            "./Data/Processed/EPAair_PM25_NC2019_processed.csv")
```

Combine datasets

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

```
#Checking final processed datasets
glimpse(EPAair_O3_NC2018_processed)
```

```
## Rows: 9,737
## Columns: 7
## $ Date               <date> 3-01-20, 3-02-20, 3-03-20, 3-04-20, 3-05-20, 3-06-~
## $ DAILY_AQI_VALUE    <int> 40, 43, 44, 45, 44, 28, 33, 41, 45, 40, 31, 43, 42,~
## $ Site.Name          <fct> Taylorsville Liledoun, Taylorsville Liledoun, Taylo~
## $ AQS_PARAMETER_DESC <fct> Ozone, Ozone, Ozone, Ozone, Ozone, Ozone, Ozone, Oz~
## $ COUNTY             <fct> Alexander, Alexander, Alexander, Alexander, Alexand~
## $ SITE_LATITUDE      <dbl> 35.9138, 35.9138, 35.9138, 35.9138, 35.9138, 35.913~
## $ SITE_LONGITUDE     <dbl> -81.191, -81.191, -81.191, -81.191, -81.191, -81.19~
```

```
glimpse(EPAair_O3_NC2019_processed)
```

```
## Rows: 10,592
## Columns: 7
## $ Date               <date> 1-01-20, 1-02-20, 1-03-20, 1-04-20, 1-05-20, 1-06-~
## $ DAILY_AQI_VALUE    <int> 27, 17, 15, 20, 34, 34, 27, 35, 35, 28, 27, 25, 31,~
## $ Site.Name          <fct> Taylorsville Liledoun, Taylorsville Liledoun, Taylo~
```

```
## $ AQS_PARAMETER_DESC <fct> Ozone, Ozone, Ozone, Ozone, Ozone, Ozone, Ozone, Oz~
## $ COUNTY             <fct> Alexander, Alexander, Alexander, Alexander, Alexand~
## $ SITE_LATITUDE      <dbl> 35.9138, 35.9138, 35.9138, 35.9138, 35.9138, 35.913~
## $ SITE_LONGITUDE     <dbl> -81.191, -81.191, -81.191, -81.191, -81.191, -81.19~
```

```
glimpse(EPAair_PM25_NC2018_processed)
```

```
## Rows: 8,983
## Columns: 7
## $ Date               <date> 1-02-20, 1-05-20, 1-08-20, 1-11-20, NA, NA, NA, NA~
## $ DAILY_AQI_VALUE    <int> 12, 15, 22, 3, 10, 19, 8, 10, 18, 7, 24, 5, 9, 14, ~
## $ 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~
```

```
glimpse(EPAair_PM25_NC2019_processed)
```

```
## Rows: 8,581
## Columns: 7
## $ Date               <date> 1-03-20, 1-06-20, 1-09-20, 1-12-20, NA, NA, NA, NA~
## $ 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~
```

```
#Combining four datasets
EPAAir_AQ_MasterDataSet <- rbind(EPAair_O3_NC2018_processed, EPAair_O3_NC2019_processed,
                                 EPAair_PM25_NC2018_processed, EPAair_PM25_NC2019_processed)
glimpse(EPAAir_AQ_MasterDataSet)
```

```
## Rows: 37,893
## Columns: 7
## $ Date               <date> 3-01-20, 3-02-20, 3-03-20, 3-04-20, 3-05-20, 3-06-~
## $ DAILY_AQI_VALUE    <int> 40, 43, 44, 45, 44, 28, 33, 41, 45, 40, 31, 43, 42,~
## $ Site.Name          <fct> Taylorsville Liledoun, Taylorsville Liledoun, Taylo~
## $ AQS_PARAMETER_DESC <fct> Ozone, Ozone, Ozone, Ozone, Ozone, Ozone, Ozone, Oz~
## $ COUNTY             <fct> Alexander, Alexander, Alexander, Alexander, Alexand~
## $ SITE_LATITUDE      <dbl> 35.9138, 35.9138, 35.9138, 35.9138, 35.9138, 35.913~
## $ SITE_LONGITUDE     <dbl> -81.191, -81.191, -81.191, -81.191, -81.191, -81.19~
```

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.

```
#Filtering combined dataset to only include the rows with the given sites.
#Saving in new object.
EPAAir_AQ_MasterDataSet_wrangled <- EPAAir_AQ_MasterDataSet %>%
  filter(Site.Name %in% c("Linville Falls", "Durham Armory", "Leggett", "Hattie Avenue",
                          "Clemmons Middle", "Mendenhall School", "Frying Pan Mountain",
                          "West Johnston Co.", "Garinger High School", "Castle Hayne",
                          "Pitt Agri. Center", "Bryson City", "Millbrook School"))

#checking if the filtering worked
count(EPAAir_AQ_MasterDataSet_wrangled, EPAAir_AQ_MasterDataSet_wrangled$Site.Name)
```

```
##    EPAAir_AQ_MasterDataSet_wrangled$Site.Name    n
## 1                                Bryson City 1171
## 2                              Castle Hayne 1108
## 3                            Clemmons Middle 1261
## 4                              Durham Armory 1405
## 5                          Frying Pan Mountain  638
## 6                         Garinger High School 1818
## 7                              Hattie Avenue 1432
## 8                                    Leggett 1184
## 9                               Linville Falls  627
## 10                           Mendenhall School 1172
## 11                            Millbrook School 2169
## 12                           Pitt Agri. Center 1303
## 13                           West Johnston Co. 1222
```

```
glimpse(EPAAir_AQ_MasterDataSet_wrangled)
```

```
## Rows: 16,510
## Columns: 7
## $ Date               <date> 3-01-20, 3-05-20, 3-06-20, 3-07-20, 3-08-20, 3-09-~
## $ DAILY_AQI_VALUE    <int> 42, 44, 38, 38, 41, 45, 47, 41, 43, 43, 43, 61, 50,~
## $ Site.Name          <fct> Linville Falls, Linville Falls, Linville Falls, Lin~
## $ AQS_PARAMETER_DESC <fct> Ozone, Ozone, Ozone, Ozone, Ozone, Ozone, Ozone, Oz~
## $ 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~
```

```
#Applying the split-apply-combine strategy
EPAAir_AQ_MasterDataSet_grouped <-
  EPAAir_AQ_MasterDataSet_wrangled %>%
  group_by(Date, Site.Name, AQS_PARAMETER_DESC, COUNTY)%>%
  summarise(meanAQI = mean(DAILY_AQI_VALUE),
            meanlatitude = mean(SITE_LATITUDE),
            meanlongitude = mean(SITE_LONGITUDE),
```

```
            .groups = "keep")

glimpse(EPAAir_AQ_MasterDataSet_grouped)
```

```
## Rows: 3,128
## Columns: 7
## Groups: Date, Site.Name, AQS_PARAMETER_DESC, COUNTY [3,128]
## $ Date               <date> 1-01-20, 1-01-20, 1-01-20, 1-01-20, 1-01-20, 1-01-~
## $ Site.Name          <fct> Bryson City, Castle Hayne, Clemmons Middle, Durham ~
## $ AQS_PARAMETER_DESC <fct> PM2.5, PM2.5, PM2.5, PM2.5, Ozone, Ozone, PM2.5, PM~
## $ COUNTY             <fct> Swain, New Hanover, Forsyth, Durham, Haywood, Meckl~
## $ meanAQI            <dbl> 29.50000, 13.50000, 24.00000, 33.00000, 47.00000, 2~
## $ meanlatitude       <dbl> 35.43477, 34.36417, 36.02600, 36.03296, 35.37917, 3~
## $ meanlongitude      <dbl> -83.44213, -77.83861, -80.34200, -78.90404, -82.792~
```

```
#Making and parsing the date columns
EPAAir_AQ_MasterDataSet_grouped <- mutate(EPAAir_AQ_MasterDataSet_grouped, Month = month(Date))
EPAAir_AQ_MasterDataSet_grouped <- mutate(EPAAir_AQ_MasterDataSet_grouped, Year = year(Date))
glimpse(EPAAir_AQ_MasterDataSet_grouped)
```

```
## Rows: 3,128
## Columns: 9
## Groups: Date, Site.Name, AQS_PARAMETER_DESC, COUNTY [3,128]
## $ Date               <date> 1-01-20, 1-01-20, 1-01-20, 1-01-20, 1-01-20, 1-01-~
## $ Site.Name          <fct> Bryson City, Castle Hayne, Clemmons Middle, Durham ~
## $ AQS_PARAMETER_DESC <fct> PM2.5, PM2.5, PM2.5, PM2.5, Ozone, Ozone, PM2.5, PM~
## $ COUNTY             <fct> Swain, New Hanover, Forsyth, Durham, Haywood, Meckl~
## $ meanAQI            <dbl> 29.50000, 13.50000, 24.00000, 33.00000, 47.00000, 2~
## $ meanlatitude       <dbl> 35.43477, 34.36417, 36.02600, 36.03296, 35.37917, 3~
## $ meanlongitude      <dbl> -83.44213, -77.83861, -80.34200, -78.90404, -82.792~
## $ Month              <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, ~
## $ Year               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
```

```
# Reordering columns to put month with the rest of the date variables
EPAAir_AQ_MasterDataSet_grouped <- select(EPAAir_AQ_MasterDataSet_grouped, Date,
                                  Month, Year, Site.Name:meanlongitude)
glimpse(EPAAir_AQ_MasterDataSet_grouped)
```

```
## Rows: 3,128
## Columns: 9
## Groups: Date, Site.Name, AQS_PARAMETER_DESC, COUNTY [3,128]
## $ Date               <date> 1-01-20, 1-01-20, 1-01-20, 1-01-20, 1-01-20, 1-01-~
## $ Month              <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, ~
## $ Year               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ Site.Name          <fct> Bryson City, Castle Hayne, Clemmons Middle, Durham ~
## $ AQS_PARAMETER_DESC <fct> PM2.5, PM2.5, PM2.5, PM2.5, Ozone, Ozone, PM2.5, PM~
## $ COUNTY             <fct> Swain, New Hanover, Forsyth, Durham, Haywood, Meckl~
## $ meanAQI            <dbl> 29.50000, 13.50000, 24.00000, 33.00000, 47.00000, 2~
## $ meanlatitude       <dbl> 35.43477, 34.36417, 36.02600, 36.03296, 35.37917, 3~
## $ meanlongitude      <dbl> -83.44213, -77.83861, -80.34200, -78.90404, -82.792~
```

```
#TWO ISSUES: number of rows is wrong and year is coming as 1
```

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.

```
EPAAir_AQ_MasterDataSet_grouped_spread <- pivot_wider(EPAAir_AQ_MasterDataSet_grouped,
                        names_from = AQS_PARAMETER_DESC, values_from = meanAQI)
glimpse(EPAAir_AQ_MasterDataSet_grouped_spread)
```

```
## Rows: 1,845
## Columns: 9
## Groups: Date, Site.Name, COUNTY [1,845]
## $ Date          <date> 1-01-20, 1-01-20, 1-01-20, 1-01-20, 1-01-20, 1-01-20, 1~
## $ Month         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2,~
## $ Year          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ Site.Name     <fct> Bryson City, Castle Hayne, Clemmons Middle, Durham Armor~
## $ COUNTY        <fct> Swain, New Hanover, Forsyth, Durham, Haywood, Mecklenbur~
## $ meanlatitude  <dbl> 35.43477, 34.36417, 36.02600, 36.03296, 35.37917, 35.240~
## $ meanlongitude <dbl> -83.44213, -77.83861, -80.34200, -78.90404, -82.79250, -~
## $ PM2.5         <dbl> 29.50000, 13.50000, 24.00000, 33.00000, NA, 21.33333, 22~
## $ Ozone         <dbl> NA, NA, NA, NA, 47.0, 28.0, NA, NA, NA, 32.5, NA, 26.0, ~
```

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

```
dim(EPAAir_AQ_MasterDataSet_grouped_spread)
```

```
## [1] 1845    9
```

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

```
write.csv(EPAAir_AQ_MasterDataSet_grouped_spread, row.names = FALSE, file="./Data/Processed/EPAair_O3_PM
```

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_AQ_MasterDataSet_Qn12 <-
  EPAAir_AQ_MasterDataSet_grouped_spread %>%
  group_by(Site.Name, Month, Year)%>%
  summarise(meanAQIOzone = mean(Ozone), meanAQIPM2.5 = mean(PM2.5), .groups = "keep")%>%
  drop_na(meanAQIOzone)

#13

dim(EPAAir_AQ_MasterDataSet_Qn12)
```

```
## [1] 1431    5
```

```
glimpse(EPAAir_AQ_MasterDataSet_Qn12)
```

```
## Rows: 1,431
## Columns: 5
## Groups: Site.Name, Month, Year [1,431]
## $ Site.Name    <fct> Bryson City, Bryson City, Bryson City, Bryson City, Bryso~
## $ Month        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, ~
## $ Year         <dbl> 3, 4, 5, 6, 7, 8, 9, 10, 3, 4, 5, 6, 7, 8, 9, 10, 3, 4, 5~
## $ meanAQIOzone <dbl> 30.5, 45.5, 54.5, 47.0, 34.5, 32.5, 33.5, 29.0, 37.5, 44.~
## $ meanAQIPM2.5 <dbl> 35.0, 22.5, 35.0, 31.5, 45.0, 30.0, 29.0, 38.0, 19.5, 36.~
```

14. Why did we use the function `drop_na` rather than `na.omit`? Hint: replace `drop_na` with `na.omit` in part 12 and observe what happens with the dimensions of the summary date frame.

    Answer: The dimensions with na.omit are [1270,5] whereas with drop_na(), it's [1431,5]. This is because na.omit removes all NAs from the dataframe whereas drop_na only drops the NAs in the specified column. #CHECK!!!