Assignment 4: Data Wrangling

Caroline Reents

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics (ENV872L) on data wrangling.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Use the lesson as a guide. It contains code that can be modified to complete the assignment.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file. You will need to have the correct software installed to do this (see Software Installation Guide) Press the `Knit` button in the RStudio scripting panel. This will save the PDF output in your Assignments folder.
6. After Knitting, please submit the completed exercise (PDF file) to the dropbox in Sakai. Please add your last name into the file name (e.g., "Salk_A04_DataWrangling.pdf") prior to submission.

The completed exercise is due on Thursday, 7 February, 2019 before class begins.

Set up your session

1. Check your working directory, load the `tidyverse` package, and upload all four raw data files associated with the EPA Air dataset. See the README file for the EPA air datasets for more information (especially if you have not worked with air quality data previously).
2. Generate a few lines of code to get to know your datasets (basic data summaries, etc.).

```
#1
getwd()
```

```
## [1] "/Users/carolinereents/Desktop/Data Analytics/EnvironmentalDataAnalytics"
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidy
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(stringr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```
EPAair_O3_NC2017_raw <- read_csv("./Data/Raw/EPAair_O3_NC2017_raw.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Date = col_character(),
##   Source = col_character(),
##   `Daily Max 8-hour Ozone Concentration` = col_double(),
##   UNITS = col_character(),
##   `Site Name` = col_character(),
##   AQS_PARAMETER_DESC = col_character(),
##   CBSA_NAME = col_character(),
##   STATE = col_character(),
##   COUNTY = col_character(),
##   SITE_LATITUDE = col_double(),
##   SITE_LONGITUDE = col_double()
## )

## See spec(...) for full column specifications.
```

```
EPAair_O3_NC2018_raw <- read_csv("./Data/Raw/EPAair_O3_NC2018_raw.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Date = col_character(),
##   Source = col_character(),
##   `Daily Max 8-hour Ozone Concentration` = col_double(),
##   UNITS = col_character(),
##   `Site Name` = col_character(),
##   AQS_PARAMETER_DESC = col_character(),
##   CBSA_NAME = col_character(),
##   STATE = col_character(),
##   COUNTY = col_character(),
##   SITE_LATITUDE = col_double(),
##   SITE_LONGITUDE = col_double()
## )
## See spec(...) for full column specifications.
```

```
EPAair_PM25_NC2017_raw <- read_csv("./Data/Raw/EPAair_PM25_NC2017_raw.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Date = col_character(),
##   Source = col_character(),
##   `Daily Mean PM2.5 Concentration` = col_double(),
##   UNITS = col_character(),
##   `Site Name` = col_character(),
##   AQS_PARAMETER_DESC = col_character(),
```

```
##   CBSA_NAME = col_character(),
##   STATE = col_character(),
##   COUNTY = col_character(),
##   SITE_LATITUDE = col_double(),
##   SITE_LONGITUDE = col_double()
## )
## See spec(...) for full column specifications.
```

```
EPAair_PM25_NC2018_raw <- read_csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   Date = col_character(),
##   Source = col_character(),
##   `Daily Mean PM2.5 Concentration` = col_double(),
##   UNITS = col_character(),
##   `Site Name` = col_character(),
##   AQS_PARAMETER_DESC = col_character(),
##   CBSA_NAME = col_character(),
##   STATE = col_character(),
##   COUNTY = col_character(),
##   SITE_LATITUDE = col_double(),
##   SITE_LONGITUDE = col_double()
## )
## See spec(...) for full column specifications.
```

```
#2
head(EPAair_O3_NC2017_raw)
```

```
## # A tibble: 6 x 20
##   Date  Source `Site ID`   POC `Daily Max 8-ho~ UNITS DAILY_AQI_VALUE
##   <chr> <chr>      <int> <int>            <dbl> <chr>           <int>
## 1 3/1/~ AQS    370030005     1            0.041 ppm                38
## 2 3/2/~ AQS    370030005     1            0.046 ppm                43
## 3 3/3/~ AQS    370030005     1            0.046 ppm                43
## 4 3/4/~ AQS    370030005     1            0.046 ppm                43
## 5 3/5/~ AQS    370030005     1            0.046 ppm                43
## 6 3/6/~ AQS    370030005     1            0.048 ppm                44
## # ... with 13 more variables: `Site Name` <chr>, DAILY_OBS_COUNT <int>,
## #   PERCENT_COMPLETE <int>, AQS_PARAMETER_CODE <int>,
## #   AQS_PARAMETER_DESC <chr>, CBSA_CODE <int>, CBSA_NAME <chr>,
## #   STATE_CODE <int>, STATE <chr>, COUNTY_CODE <int>, COUNTY <chr>,
## #   SITE_LATITUDE <dbl>, SITE_LONGITUDE <dbl>
```

```
colnames(EPAair_O3_NC2017_raw)
```

```
##  [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"
```

```
summary(EPAair_O3_NC2017_raw)
```

```
##      Date              Source             Site ID               POC
##  Length:10219       Length:10219       Min.   :370030005   Min.   :1
##  Class :character   Class :character   1st Qu.:370650099   1st Qu.:1
##  Mode  :character   Mode  :character   Median :371010002   Median :1
##                                        Mean   :370962005   Mean   :1
##                                        3rd Qu.:371239991   3rd Qu.:1
##                                        Max.   :371990004   Max.   :1
##
##  Daily Max 8-hour Ozone Concentration    UNITS           DAILY_AQI_VALUE
##  Min.   :0.00500                     Length:10219       Min.   :  5.00
##  1st Qu.:0.03500                     Class :character   1st Qu.: 32.00
##  Median :0.04300                     Mode  :character   Median : 40.00
##  Mean   :0.04211                                        Mean   : 39.87
##  3rd Qu.:0.04900                                        3rd Qu.: 45.00
##  Max.   :0.07500                                        Max.   :115.00
##
##   Site Name         DAILY_OBS_COUNT PERCENT_COMPLETE AQS_PARAMETER_CODE
##  Length:10219       Min.   :13.00   Min.   : 76.00   Min.   :44201
##  Class :character   1st Qu.:17.00   1st Qu.:100.00   1st Qu.:44201
##  Mode  :character   Median :17.00   Median :100.00   Median :44201
##                     Mean   :16.94   Mean   : 99.63   Mean   :44201
##                     3rd Qu.:17.00   3rd Qu.:100.00   3rd Qu.:44201
##                     Max.   :17.00   Max.   :100.00   Max.   :44201
##
##  AQS_PARAMETER_DESC   CBSA_CODE      CBSA_NAME           STATE_CODE
##  Length:10219       Min.   :11700   Length:10219       Min.   :37
##  Class :character   1st Qu.:16740   Class :character   1st Qu.:37
##  Mode  :character   Median :24660   Mode  :character   Median :37
##                     Mean   :27541                      Mean   :37
##                     3rd Qu.:39580                      3rd Qu.:37
##                     Max.   :49180                      Max.   :37
##                     NA's   :2541
##     STATE            COUNTY_CODE        COUNTY          SITE_LATITUDE
##  Length:10219       Min.   :  3.00   Length:10219       Min.   :34.36
##  Class :character   1st Qu.: 65.00   Class :character   1st Qu.:35.26
##  Mode  :character   Median :101.00   Mode  :character   Median :35.55
##                     Mean   : 96.07                      Mean   :35.60
##                     3rd Qu.:123.00                      3rd Qu.:35.99
##                     Max.   :199.00                      Max.   :36.31
##
##  SITE_LONGITUDE
```

```
##  Min.   :-83.80
##  1st Qu.:-82.05
##  Median :-80.23
##  Mean   :-80.32
##  3rd Qu.:-78.77
##  Max.   :-76.62
##
```

```
dim(EPAair_O3_NC2017_raw)
```

```
## [1] 10219    20
```

```
head(EPAair_O3_NC2018_raw)
```

```
## # A tibble: 6 x 20
##   Date  Source `Site ID`   POC `Daily Max 8-ho~ UNITS DAILY_AQI_VALUE
##   <chr> <chr>      <int> <int>            <dbl> <chr>           <int>
## 1 2/16~ AirNow 370030005     1            0.038 ppm                35
## 2 2/17~ AirNow 370030005     1            0.033 ppm                31
## 3 2/18~ AirNow 370030005     1            0.04  ppm                37
## 4 2/19~ AirNow 370030005     1            0.02  ppm                19
## 5 2/20~ AirNow 370030005     1            0.019 ppm                18
## 6 2/21~ AirNow 370030005     1            0.021 ppm                19
## # ... with 13 more variables: `Site Name` <chr>, DAILY_OBS_COUNT <int>,
## #   PERCENT_COMPLETE <int>, AQS_PARAMETER_CODE <int>,
## #   AQS_PARAMETER_DESC <chr>, CBSA_CODE <int>, CBSA_NAME <chr>,
## #   STATE_CODE <int>, STATE <chr>, COUNTY_CODE <int>, COUNTY <chr>,
## #   SITE_LATITUDE <dbl>, SITE_LONGITUDE <dbl>
```

```
colnames(EPAair_O3_NC2018_raw)
```

```
##  [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"
```

```
summary(EPAair_O3_NC2018_raw)
```

```
##      Date              Source             Site ID               POC
##  Length:10781       Length:10781       Min.   :370030005   Min.   :1
##  Class :character   Class :character   1st Qu.:370630015   1st Qu.:1
```

```
##  Mode  :character   Mode  :character   Median :370870036   Median :1
##                                        Mean   :370959550   Mean   :1
##                                        3rd Qu.:371290002   3rd Qu.:1
##                                        Max.   :371990004   Max.   :1
##
##  Daily Max 8-hour Ozone Concentration    UNITS           DAILY_AQI_VALUE
##  Min.   :0.00000                      Length:10781       Min.   :  0.00
##  1st Qu.:0.03400                      Class :character   1st Qu.: 31.00
##  Median :0.04100                      Mode  :character   Median : 38.00
##  Mean   :0.04124                                         Mean   : 39.46
##  3rd Qu.:0.04900                                         3rd Qu.: 45.00
##  Max.   :0.07700                                         Max.   :122.00
##
##   Site Name         DAILY_OBS_COUNT PERCENT_COMPLETE AQS_PARAMETER_CODE
##  Length:10781       Min.   :12.00   Min.   : 71.00   Min.   :44201
##  Class :character   1st Qu.:17.00   1st Qu.:100.00   1st Qu.:44201
##  Mode  :character   Median :17.00   Median :100.00   Median :44201
##                     Mean   :18.69   Mean   : 99.62   Mean   :44201
##                     3rd Qu.:18.00   3rd Qu.:100.00   3rd Qu.:44201
##                     Max.   :24.00   Max.   :100.00   Max.   :44201
##
##  AQS_PARAMETER_DESC   CBSA_CODE      CBSA_NAME           STATE_CODE
##  Length:10781       Min.   :11700   Length:10781       Min.   :37
##  Class :character   1st Qu.:16740   Class :character   1st Qu.:37
##  Mode  :character   Median :24660   Mode  :character   Median :37
##                     Mean   :27015                      Mean   :37
##                     3rd Qu.:39580                      3rd Qu.:37
##                     Max.   :49180                      Max.   :37
##                     NA's   :2802
##     STATE            COUNTY_CODE        COUNTY           SITE_LATITUDE
##  Length:10781       Min.   :  3.00   Length:10781       Min.   :34.36
##  Class :character   1st Qu.: 63.00   Class :character   1st Qu.:35.26
##  Mode  :character   Median : 87.00   Mode  :character   Median :35.59
##                     Mean   : 95.84                      Mean   :35.63
##                     3rd Qu.:129.00                      3rd Qu.:36.03
##                     Max.   :199.00                      Max.   :36.31
##
##  SITE_LONGITUDE
##  Min.   :-83.80
##  1st Qu.:-82.05
##  Median :-80.34
##  Mean   :-80.39
##  3rd Qu.:-78.90
##  Max.   :-76.62
##
```

```
dim(EPAair_O3_NC2018_raw)
```

```
## [1] 10781    20
```

```
head(EPAair_PM25_NC2017_raw)
```

```
## # A tibble: 6 x 20
##   Date  Source `Site ID`   POC `Daily Mean PM2~ UNITS DAILY_AQI_VALUE
##   <chr> <chr>      <int> <int>            <dbl> <chr>           <int>
```

```
## 1 1/1/~ AQS    370110002     1              2.9 ug/m~              12
## 2 1/4/~ AQS    370110002     1              1.2 ug/m~               5
## 3 1/7/~ AQS    370110002     1              3.2 ug/m~              13
## 4 1/10~ AQS    370110002     1              6.4 ug/m~              27
## 5 1/13~ AQS    370110002     1              3.6 ug/m~              15
## 6 1/16~ AQS    370110002     1              5.8 ug/m~              24
## # ... with 13 more variables: `Site Name` <chr>, DAILY_OBS_COUNT <int>,
## #   PERCENT_COMPLETE <int>, AQS_PARAMETER_CODE <int>,
## #   AQS_PARAMETER_DESC <chr>, CBSA_CODE <int>, CBSA_NAME <chr>,
## #   STATE_CODE <int>, STATE <chr>, COUNTY_CODE <int>, COUNTY <chr>,
## #   SITE_LATITUDE <dbl>, SITE_LONGITUDE <dbl>
```

```
colnames(EPAair_PM25_NC2017_raw)
```

```
##  [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"
```

```
summary(EPAair_PM25_NC2017_raw)
```

```
##      Date              Source             Site ID               POC
##  Length:9494        Length:9494        Min.   :370110002   Min.   :1.000
##  Class :character   Class :character   1st Qu.:370630015   1st Qu.:3.000
##  Mode  :character   Mode  :character   Median :371010002   Median :3.000
##                                        Mean   :370980114   Mean   :2.734
##                                        3rd Qu.:371210004   3rd Qu.:3.000
##                                        Max.   :371830021   Max.   :4.000
##
##  Daily Mean PM2.5 Concentration    UNITS           DAILY_AQI_VALUE
##  Min.   :-3.900                 Length:9494        Min.   : 0.00
##  1st Qu.: 5.000                 Class :character   1st Qu.:21.00
##  Median : 7.300                 Mode  :character   Median :30.00
##  Mean   : 7.742                                    Mean   :31.72
##  3rd Qu.:10.000                                    3rd Qu.:42.00
##  Max.   :31.900                                    Max.   :93.00
##
##   Site Name         DAILY_OBS_COUNT PERCENT_COMPLETE AQS_PARAMETER_CODE
##  Length:9494        Min.   :1       Min.   :100      Min.   :88101
##  Class :character   1st Qu.:1       1st Qu.:100      1st Qu.:88101
##  Mode  :character   Median :1       Median :100      Median :88101
##                     Mean   :1       Mean   :100      Mean   :88221
##                     3rd Qu.:1       3rd Qu.:100      3rd Qu.:88502
##                     Max.   :1       Max.   :100      Max.   :88502
##
##  AQS_PARAMETER_DESC   CBSA_CODE      CBSA_NAME           STATE_CODE
##  Length:9494        Min.   :11700   Length:9494        Min.   :37
##  Class :character   1st Qu.:16740   Class :character   1st Qu.:37
##  Mode  :character   Median :25860   Mode  :character   Median :37
```

```
##                    Mean   :30793                      Mean   :37
##                    3rd Qu.:41820                      3rd Qu.:37
##                    Max.   :49180                      Max.   :37
##                    NA's   :1353
##     STATE           COUNTY_CODE     COUNTY          SITE_LATITUDE
##  Length:9494        Min.   : 11   Length:9494        Min.   :34.36
##  Class :character   1st Qu.: 63   Class :character   1st Qu.:35.26
##  Mode  :character   Median :101   Mode  :character   Median :35.64
##                     Mean   : 98                      Mean   :35.60
##                     3rd Qu.:121                      3rd Qu.:35.91
##                     Max.   :183                      Max.   :36.11
##
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-80.23
##  Mean   :-80.03
##  3rd Qu.:-78.82
##  Max.   :-76.21
##
```

```
dim(EPAair_PM25_NC2017_raw)
```

```
## [1] 9494   20
```

```
head(EPAair_PM25_NC2018_raw)
```

```
## # A tibble: 6 x 20
##   Date  Source `Site ID`   POC `Daily Mean PM2~ UNITS DAILY_AQI_VALUE
##   <chr> <chr>      <int> <int>            <dbl> <chr>           <int>
## 1 1/2/~ AQS    370110002     1              2.9 ug/m~              12
## 2 1/5/~ AQS    370110002     1              3.7 ug/m~              15
## 3 1/8/~ AQS    370110002     1              5.3 ug/m~              22
## 4 1/11~ AQS    370110002     1              0.8 ug/m~               3
## 5 1/14~ AQS    370110002     1              2.5 ug/m~              10
## 6 1/17~ AQS    370110002     1              4.5 ug/m~              19
## # ... with 13 more variables: `Site Name` <chr>, DAILY_OBS_COUNT <int>,
## #   PERCENT_COMPLETE <int>, AQS_PARAMETER_CODE <int>,
## #   AQS_PARAMETER_DESC <chr>, CBSA_CODE <int>, CBSA_NAME <chr>,
## #   STATE_CODE <int>, STATE <chr>, COUNTY_CODE <int>, COUNTY <chr>,
## #   SITE_LATITUDE <dbl>, SITE_LONGITUDE <dbl>
```

```
colnames(EPAair_PM25_NC2018_raw)
```

```
##  [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"
```

```
summary(EPAair_PM25_NC2018_raw)
```

```
##      Date               Source              Site ID           POC
##  Length:7611        Length:7611        Min.   :370110002   Min.   :1.000
##  Class :character   Class :character   1st Qu.:370630015   1st Qu.:3.000
##  Mode  :character   Mode  :character   Median :371190041   Median :3.000
##                                        Mean   :371031969   Mean   :3.011
##                                        3rd Qu.:371290002   3rd Qu.:3.000
##                                        Max.   :371830021   Max.   :5.000
##
##  Daily Mean PM2.5 Concentration    UNITS           DAILY_AQI_VALUE
##  Min.   :-2.800                 Length:7611        Min.   : 0.00
##  1st Qu.: 5.000                 Class :character   1st Qu.:21.00
##  Median : 7.200                 Mode  :character   Median :30.00
##  Mean   : 7.554                                    Mean   :31.03
##  3rd Qu.: 9.800                                    3rd Qu.:41.00
##  Max.   :34.200                                    Max.   :97.00
##
##   Site Name         DAILY_OBS_COUNT PERCENT_COMPLETE AQS_PARAMETER_CODE
##  Length:7611        Min.   :1       Min.   :100      Min.   :88101
##  Class :character   1st Qu.:1       1st Qu.:100      1st Qu.:88101
##  Mode  :character   Median :1       Median :100      Median :88101
##                     Mean   :1       Mean   :100      Mean   :88167
##                     3rd Qu.:1       3rd Qu.:100      3rd Qu.:88101
##                     Max.   :1       Max.   :100      Max.   :88502
##
##  AQS_PARAMETER_DESC   CBSA_CODE      CBSA_NAME           STATE_CODE
##  Length:7611        Min.   :11700   Length:7611        Min.   :37
##  Class :character   1st Qu.:19000   Class :character   1st Qu.:37
##  Mode  :character   Median :25860   Mode  :character   Median :37
##                     Mean   :30249                      Mean   :37
##                     3rd Qu.:39580                      3rd Qu.:37
##                     Max.   :49180                      Max.   :37
##                     NA's   :1025
##     STATE            COUNTY_CODE       COUNTY           SITE_LATITUDE
##  Length:7611        Min.   : 11.0   Length:7611        Min.   :34.36
##  Class :character   1st Qu.: 63.0   Class :character   1st Qu.:35.26
##  Mode  :character   Median :119.0   Mode  :character   Median :35.64
##                     Mean   :103.2                      Mean   :35.59
##                     3rd Qu.:129.0                      3rd Qu.:35.87
##                     Max.   :183.0                      Max.   :36.11
##
##  SITE_LONGITUDE
##  Min.   :-83.44
##  1st Qu.:-80.87
##  Median :-79.84
##  Mean   :-79.95
##  3rd Qu.:-78.57
##  Max.   :-76.21
##
```

```
dim(EPAair_PM25_NC2018_raw)
```

```
## [1] 7611   20
```

```
#remove spaces and change to periods
names (EPAair_O3_NC2017_raw) <- str_replace_all(names(EPAair_O3_NC2017_raw), c(" " = "."))
```

```
names (EPAair_O3_NC2018_raw) <- str_replace_all(names(EPAair_O3_NC2018_raw), c(" " = "."))
names (EPAair_PM25_NC2017_raw) <- str_replace_all(names(EPAair_PM25_NC2017_raw), c(" " = "."))
names (EPAair_PM25_NC2018_raw) <- str_replace_all(names(EPAair_PM25_NC2018_raw), c(" " = "."))
```

Wrangle individual datasets to create processed files.

3. Change date to date
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.

```
#3
EPAair_O3_NC2017_raw$Date <- as.Date(EPAair_O3_NC2017_raw$Date, format = "%m/%d/%y")
EPAair_O3_NC2018_raw$Date <- as.Date(EPAair_O3_NC2018_raw$Date, format = "%m/%d/%y")
EPAair_PM25_NC2017_raw$Date <- as.Date(EPAair_PM25_NC2017_raw$Date, format = "%m/%d/%y")
EPAair_PM25_NC2018_raw$Date <- as.Date(EPAair_PM25_NC2018_raw$Date, format = "%m/%d/%y")

class(EPAair_O3_NC2017_raw$Date)
```

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

```
#4
EPAair_O3_NC2017_skinny <- select(EPAair_O3_NC2017_raw, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_

EPAair_O3_NC2018_skinny <- select(EPAair_O3_NC2018_raw, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAMETER_

EPAair_PM25_NC2017_skinny <- select(EPAair_PM25_NC2017_raw, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAM

EPAair_PM25_NC2018_skinny <- select(EPAair_PM25_NC2018_raw, Date, DAILY_AQI_VALUE, Site.Name, AQS_PARAM




#5
EPAair_PM25_NC2017_skinny$AQS_PARAMETER_DESC <- "PM2.5"
EPAair_PM25_NC2018_skinny$AQS_PARAMETER_DESC <- "PM2.5"

#6
write.csv(EPAair_O3_NC2017_skinny, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2017_skinny.
write.csv(EPAair_O3_NC2018_skinny, row.names = FALSE, file = "./Data/Processed/EPAair_O3_NC2018_skinny.
write.csv(EPAair_PM25_NC2017_skinny, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2017_ski
write.csv(EPAair_PM25_NC2018_skinny, row.names = FALSE, file = "./Data/Processed/EPAair_PM25_NC2018_ski
```

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:

- Sites: Blackstone, Bryson City, Triple Oak

- Add columns for "Month" and "Year" by parsing your "Date" column (hint: `separate` function or `lubridate` package)

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_NC1718_Processed.csv"

```
#7
EPA_full_skinny <- rbind(EPAair_O3_NC2017_skinny, EPAair_O3_NC2018_skinny, EPAair_PM25_NC2017_skinny, EI

#8
EPA_Black_Bryson_Oak <- EPA_full_skinny %>%
  filter(Site.Name == "Blackstone" | Site.Name =="Bryson City" | Site.Name == "Triple Oak") %>%
  mutate(month=month(Date), year= year(Date))


#9
EPA_Black_Bryson_Oak_spread <- spread(EPA_Black_Bryson_Oak, AQS_PARAMETER_DESC, DAILY_AQI_VALUE)
View(EPA_Black_Bryson_Oak_spread)
#10
dim(EPA_Black_Bryson_Oak_spread)
```

```
## [1] 1953    9
```

```
#11
write.csv(EPA_Black_Bryson_Oak_spread, row.names=FALSE, file = "./Data/Processed/EPAair_O3_PM25_NC1718_
```

Generate summary tables

12. Use the split-apply-combine strategy to generate two new data frames:
    a. A summary table of mean AQI values for O3 and PM2.5 by month
    b. A summary table of the mean, minimum, and maximum AQI values of O3 and PM2.5 for each site
13. Display the data frames.

```
#12a
AQI_month <- EPA_Black_Bryson_Oak_spread %>%
  group_by(month) %>%
  filter(!is.na(Ozone) &!is.na(PM2.5)) %>%
summarise(meanO3=mean(Ozone), meanPM25= mean(PM2.5))

#12b
AQI_site_summaries <- EPA_Black_Bryson_Oak_spread %>%
  group_by(Site.Name) %>%
  filter(!is.na(Ozone) & !is.na(PM2.5)) %>%
summarise(meanO3=mean(Ozone),
          minimumO3=min(Ozone),
          maximumO3=max(Ozone),
          meanPM25= mean(PM2.5),
          minimumPM25=min(PM2.5),
          maximumPM25=max(PM2.5))

#13
print(AQI_month)
```

```
## # A tibble: 12 x 3
##    month meanO3 meanPM25
##    <dbl>  <dbl>    <dbl>
##  1     1   31.5     34.2
##  2     2   35.4     37.6
##  3     3   42.4     37.4
##  4     4   43.5     31.5
##  5     5   39.5     30.6
##  6     6   39.2     30.9
##  7     7   38.3     31.9
##  8     8   34.4     32.3
##  9     9   32.6     30.7
## 10    10   32.3     30.1
## 11    11   30.1     42.1
## 12    12   29.8     46.6
```

```
print(AQI_site_summaries)
```

```
## # A tibble: 2 x 7
##   Site.Name   meanO3 minimumO3 maximumO3 meanPM25 minimumPM25 maximumPM25
##   <chr>        <dbl>     <dbl>     <dbl>    <dbl>       <dbl>       <dbl>
## 1 Blackstone    38.3         8        97     36.7           0          83
## 2 Bryson City   35.4         5        71     30.3           3          68
```

```
#for some reason Triple Oak is not in the site summary table
```