Data Analysis

Kushal Kharel

1/11/2021

Use the readr package to read the daily_SPEC_2014.csv.bz2 data file in to R. This file contains daily levels of fine particulate matter (PM2.5) chemical constituents across the United States. The data are measured at a network of federal, state, and local monitors and assembled by the EPA. In this dataset, the "Sample.Value" column provides the level of the indicated chemical constituent and the "Parameter.Name" column provides the name of the chemical constituent. The combination of a "State.Code", a "County.Code", and a "Site.Num", uniquely identifies a monitoring site (the location of which is provided by the "Latitude" and "Longitude" columns).

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   `Parameter Code` = col_double(),
##   POC = col_double(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   `Pollutant Standard` = col_logical(),
##   `Date Local` = col_date(format = ""),
##   `Observation Count` = col_double(),
##   `Observation Percent` = col_double(),
##   `Arithmetic Mean` = col_double(),
##   `1st Max Value` = col_double(),
##   `1st Max Hour` = col_double(),
```

```
##   AQI = col_logical(),
##   `Method Code` = col_double(),
##   `Date of Last Change` = col_date(format = "")
## )
## i Use `spec()` for the full column specifications.
```

What is average Arithmetic.Mean for "Bromine PM2.5 LC" in the state of Wisconsin in this dataset?

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 1 x 2
##   Parameter.Name      mean
##   <chr>              <dbl>
## 1 Bromine PM2.5 LC 0.00396
```

Calculate the average of each chemical constituent across all states, monitoring sites and all time points.

```
## `summarise()` regrouping output by 'State.Name', 'Site.Num', 'Date.Local' (override with `.groups` a
```

```
## # A tibble: 1,690,291 x 5
## # Groups:   State.Name, Site.Num, Date.Local [111,096]
##    State.Name           Site.Num Date.Local Parameter.Name                   mean
##    <chr>                <chr>    <date>     <chr>                           <dbl>
##  1 District Of Columbia 0043     2014-01-12 OC CSN Unadjusted PM2.5 LC TOT 5453
##  2 District Of Columbia 0043     2014-01-13 OC CSN Unadjusted PM2.5 LC TOT 5453
##  3 District Of Columbia 0043     2014-01-14 OC CSN Unadjusted PM2.5 LC TOT 5453
##  4 District Of Columbia 0043     2014-01-11 OC CSN Unadjusted PM2.5 LC TOT 5370.
##  5 District Of Columbia 0043     2014-03-02 OC CSN Unadjusted PM2.5 LC TOT 4184.
##  6 District Of Columbia 0043     2014-02-09 OC CSN Unadjusted PM2.5 LC TOT 3728.
##  7 District Of Columbia 0043     2014-01-10 OC CSN Unadjusted PM2.5 LC TOT 3663.
##  8 District Of Columbia 0043     2014-03-11 OC CSN Unadjusted PM2.5 LC TOT 3591.
##  9 District Of Columbia 0043     2014-01-09 OC CSN Unadjusted PM2.5 LC TOT 3187.
## 10 District Of Columbia 0043     2014-03-12 OC CSN Unadjusted PM2.5 LC TOT 3168.
## # ... with 1,690,281 more rows
```

Which monitoring site has the highest average level of "Sulfate PM2.5 LC" across all time?

```
## `summarise()` regrouping output by 'State.Code', 'County.Code' (override with `.groups` argument)
```

```
## # A tibble: 358 x 4
## # Groups:   State.Code, County.Code [313]
##    State.Code County.Code Site.Num  mean
##    <chr>      <chr>       <chr>    <dbl>
##  1 39         081         0017      3.18
##  2 42         003         0064      3.06
##  3 54         039         1005      2.94
##  4 18         019         0006      2.74
##  5 39         153         0023      2.71
##  6 39         035         0060      2.64
##  7 39         087         0012      2.64
##  8 54         051         1002      2.62
##  9 21         111         0067      2.55
## 10 18         037         2001      2.52
## # ... with 348 more rows
```

What is the absolute difference in the average levels of "EC PM2.5 LC TOR" between the states California and Arizona, across all time and all monitoring sites?

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 1 x 3
##   Arizona California    diff
##     <dbl>      <dbl>   <dbl>
## 1   0.179      0.198 -0.0186
```

What is the median level of "OC PM2.5 LC TOR" in the western United States, across all time? Define western as any monitoring location that has a Longitude LESS THAN -100?

```
## `summarise()` regrouping output by 'Parameter.Name' (override with `.groups` argument)
```

```
## # A tibble: 1 x 3
## # Groups:   Parameter.Name [1]
##   Parameter.Name   east  west
##   <chr>           <dbl> <dbl>
## 1 OC PM2.5 LC TOR  0.88  0.43
```

How many monitoring sites are labelled as both RESIDENTIAL for "Land Use" and SUBURBAN for "Location Setting"?

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting numeric in A20237 / R20237C1: got 'CC'
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting numeric in A20238 / R20238C1: got 'CC'
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting numeric in A20239 / R20239C1: got 'CC'
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting numeric in A20240 / R20240C1: got 'CC'
```

```
## # A tibble: 1 x 1
##       N
##   <int>
## 1  3527
```

```
##                       Location.Setting
## Land.Use               RURAL SUBURBAN UNKNOWN URBAN AND CENTER CITY
##   AGRICULTURAL          2233       62       5                    10
##   BLIGHTED AREAS           5        0       0                     3
##   COMMERCIAL             353     1610      26                  3208
##   DESERT                 140        2       1                     1
##   FOREST                 620       15       1                     1
##   INDUSTRIAL            1330     1207       3                  1008
##   MILITARY RESERVATION     7        6       0                     1
##   MOBILE                  20      110       0                   130
##   RESIDENTIAL            753     3527      12                  1625
##   UNKNOWN                145        0     896                     0
```

What is the median level of "EC PM2.5 LC TOR" amongst monitoring sites that are labelled as both "RESIDENTIAL" and "SUBURBAN" in the eastern U.S., where eastern is defined as Longitude greater than or equal to -100?

```
## tibble [20,239 x 6] (S3: tbl_df/tbl/data.frame)
##  $ State.Code      : num [1:20239] 1 1 1 1 1 1 1 1 1 1 ...
##  $ County.Code     : num [1:20239] 1 1 1 3 3 3 3 5 5 7 ...
##  $ Site.Num        : num [1:20239] 1 2 3 1 2 3 10 1 2 1 ...
##  $ Longitude       : num [1:20239] -86.5 -86.4 -86.8 0 -87.7 ...
##  $ Land.Use        : chr [1:20239] "RESIDENTIAL" "AGRICULTURAL" "FOREST" "UNKNOWN" ...
##  $ Location.Setting: chr [1:20239] "SUBURBAN" "RURAL" "RURAL" "RURAL" ...
```

```
## tibble [2,108,467 x 6] (S3: tbl_df/tbl/data.frame)
##  $ State.Code     : num [1:2108467] 1 1 1 1 1 1 1 1 1 1 ...
##  $ County.Code    : num [1:2108467] 3 3 3 3 3 3 3 3 3 3 ...
##  $ Site.Num       : num [1:2108467] 10 10 10 10 10 10 10 10 10 10 ...
##  $ Parameter.Name : chr [1:2108467] "Ambient Temperature" "Ambient Temperature" "Ambient Temperature
##  $ Arithmetic.Mean: num [1:2108467] 10.9 14.1 0.7 17.9 12.7 11 9.1 6.3 8.6 -4.1 ...
##  $ Date.Local     : Date[1:2108467], format: "2014-01-02" "2014-01-05" ...
```

```
## tibble [2,108,467 x 9] (S3: tbl_df/tbl/data.frame)
##  $ State.Code      : num [1:2108467] 1 1 1 1 1 1 1 1 1 1 ...
##  $ County.Code     : num [1:2108467] 3 3 3 3 3 3 3 3 3 3 ...
##  $ Site.Num        : num [1:2108467] 10 10 10 10 10 10 10 10 10 10 ...
##  $ Parameter.Name  : chr [1:2108467] "Ambient Temperature" "Ambient Temperature" "Ambient Temperatur
##  $ Arithmetic.Mean : num [1:2108467] 10.9 14.1 0.7 17.9 12.7 11 9.1 6.3 8.6 -4.1 ...
##  $ Date.Local      : Date[1:2108467], format: "2014-01-02" "2014-01-05" ...
##  $ Longitude       : num [1:2108467] -87.9 -87.9 -87.9 -87.9 -87.9 ...
##  $ Land.Use        : chr [1:2108467] "COMMERCIAL" "COMMERCIAL" "COMMERCIAL" "COMMERCIAL" ...
##  $ Location.Setting: chr [1:2108467] "SUBURBAN" "SUBURBAN" "SUBURBAN" "SUBURBAN" ...
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 1 x 2
##   Parameter.Name  median
##   <chr>            <dbl>
## 1 EC PM2.5 LC TOR   0.61
```

Amongst monitoring sites that are labeled as COMMERCIAL for "Land Use", which month of the year has the highest average levels of "Sulfate PM2.5 LC"?

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 12 x 2
##    month  mean
##    <ord> <dbl>
##  1 Feb    2.02
##  2 Mar    1.81
##  3 Jul    1.78
##  4 Aug    1.76
##  5 Jun    1.75
##  6 Sep    1.65
```

```
##  7 Apr    1.57
##  8 May    1.56
##  9 Dec    1.54
## 10 Jan    1.32
## 11 Oct    1.31
## 12 Nov    1.30
```

Which monitoring site in the dataset has the highest correlation between "Sulfate PM2.5 LC" and "Total Nitrate PM2.5 LC" across all dates? Identify the monitoring site by it's State, County, and Site Number code

```
## `summarise()` regrouping output by 'State.Code', 'County.Code', 'Site.Num', 'Parameter.Name' (overri
```

```
## `summarise()` regrouping output by 'State.Code', 'County.Code' (override with `.groups` argument)
```

```
## # A tibble: 358 x 4
## # Groups:   State.Code, County.Code [313]
##    State.Code County.Code Site.Num correlation
##    <chr>      <chr>       <chr>          <dbl>
##  1 02         090         0035           0.898
##  2 08         001         0006           0.896
##  3 34         001         0006           0.881
##  4 42         045         0002           0.874
##  5 02         090         0010           0.864
##  6 53         033         0030           0.847
##  7 02         090         0034           0.841
##  8 41         033         0010           0.792
##  9 16         037         0002           0.791
## 10 38         017         1004           0.790
## # ... with 348 more rows
```