Data Analysis

Reading the Data

First we read in the data.

```
library(readr)
dat <- read_csv("SPEC_2014.csv.gz")
dat
```

```
## # A tibble: 1,519,790 x 26
##    State.Code County.Code Site.Num Parameter.Code   POC Latitude Longitude
##    <chr>      <chr>       <chr>             <dbl> <dbl>    <dbl>     <dbl>
##  1 01         073         0023              88102     5     33.6     -86.8
##  2 01         073         0023              88102     5     33.6     -86.8
##  3 01         073         0023              88102     5     33.6     -86.8
##  4 01         073         0023              88102     5     33.6     -86.8
##  5 01         073         0023              88102     5     33.6     -86.8
##  6 01         073         0023              88102     5     33.6     -86.8
##  7 01         073         0023              88102     5     33.6     -86.8
##  8 01         073         0023              88102     5     33.6     -86.8
##  9 01         073         0023              88102     5     33.6     -86.8
## 10 01         073         0023              88102     5     33.6     -86.8
## # ... with 1.52e+06 more rows, and 19 more variables: Datum <chr>,
## #   Parameter.Name <chr>, Sample.Duration <chr>, Pollutant.Standard <lgl>,
## #   Date.Local <date>, Units.of.Measure <chr>, Event.Type <chr>,
## #   Observation.Count <dbl>, Observation.Percent <dbl>,
## #   Sample.Value <dbl>, X1st.Max.Value <dbl>, X1st.Max.Hour <dbl>,
## #   AQI <lgl>, Method.Code <dbl>, State.Name <chr>, County.Name <chr>,
## #   City.Name <chr>, CBSA.Name <chr>, Date.of.Last.Change <date>
```

This dataset has 1519790 rows and 26 columns.

The dataset has the following variables.

```
names(dat)
```

```
##  [1] "State.Code"          "County.Code"         "Site.Num"
##  [4] "Parameter.Code"      "POC"                 "Latitude"
##  [7] "Longitude"           "Datum"               "Parameter.Name"
## [10] "Sample.Duration"     "Pollutant.Standard"  "Date.Local"
## [13] "Units.of.Measure"    "Event.Type"          "Observation.Count"
## [16] "Observation.Percent" "Sample.Value"        "X1st.Max.Value"
## [19] "X1st.Max.Hour"       "AQI"                 "Method.Code"
## [22] "State.Name"          "County.Name"         "City.Name"
## [25] "CBSA.Name"           "Date.of.Last.Change"
```

How many pollutants are there in the dataset?

```
unique(dat$Parameter.Name)
```

```
##  [1] "Antimony PM2.5 LC"
##  [2] "Arsenic PM2.5 LC"
##  [3] "Aluminum PM2.5 LC"
##  [4] "Barium PM2.5 LC"
##  [5] "Bromine PM2.5 LC"
```

```
##  [6] "Cadmium PM2.5 LC"
##  [7] "Calcium PM2.5 LC"
##  [8] "Chromium PM2.5 LC"
##  [9] "Cobalt PM2.5 LC"
## [10] "Copper PM2.5 LC"
## [11] "Chlorine PM2.5 LC"
## [12] "Cerium PM2.5 LC"
## [13] "Cesium PM2.5 LC"
## [14] "Iron PM2.5 LC"
## [15] "Lead PM2.5 LC"
## [16] "Indium PM2.5 LC"
## [17] "Manganese PM2.5 LC"
## [18] "Nickel PM2.5 LC"
## [19] "Magnesium PM2.5 LC"
## [20] "Phosphorus PM2.5 LC"
## [21] "Selenium PM2.5 LC"
## [22] "Tin PM2.5 LC"
## [23] "Titanium PM2.5 LC"
## [24] "Vanadium PM2.5 LC"
## [25] "Silicon PM2.5 LC"
## [26] "Silver PM2.5 LC"
## [27] "Zinc PM2.5 LC"
## [28] "Strontium PM2.5 LC"
## [29] "Sulfur PM2.5 LC"
## [30] "Rubidium PM2.5 LC"
## [31] "Potassium PM2.5 LC"
## [32] "Sodium PM2.5 LC"
## [33] "Zirconium PM2.5 LC"
## [34] "Chloride PM2.5 LC"
## [35] "Ammonium Ion PM2.5 LC"
## [36] "Sodium Ion Pm2.5 LC"
## [37] "Potassium Ion PM2.5 LC"
## [38] "Total Nitrate PM2.5 LC"
## [39] "OC PM2.5 LC TOR"
## [40] "EC PM2.5 LC TOR"
## [41] "OC1 PM2.5 LC"
## [42] "OC2 PM2.5 LC"
## [43] "OC3 PM2.5 LC"
## [44] "OC4 PM2.5 LC"
## [45] "OP PM2.5 LC TOR"
## [46] "EC1 PM2.5 LC"
## [47] "EC2 PM2.5 LC"
## [48] "EC3 PM2.5 LC"
## [49] "OC CSN_Rev Unadjusted PM2.5 LC TOT"
## [50] "EC CSN_Rev Unadjusted PM2.5 LC TOT"
## [51] "OC CSN_Rev Unadjusted PM2.5 LC TOR"
## [52] "OC1 CSN_Rev Unadjusted PM2.5 LC"
## [53] "OC2 CSN_Rev Unadjusted PM2.5 LC"
## [54] "OC3 CSN_Rev Unadjusted PM2.5 LC"
## [55] "OC4 CSN_Rev Unadjusted PM2.5 LC"
## [56] "OP CSN_Rev Unadjusted PM2.5 LC TOR"
## [57] "EC CSN_Rev Unadjusted PM2.5 LC TOR"
## [58] "EC1 CSN_Rev Unadjusted PM2.5 LC"
## [59] "EC2 CSN_Rev Unadjusted PM2.5 LC"
```

```
## [60] "EC3 CSN_Rev Unadjusted PM2.5 LC"
## [61] "OP CSN_Rev Unadjusted PM2.5 LC TOT"
## [62] "Sulfate PM2.5 LC"
## [63] "Europium PM2.5 LC"
## [64] "Gallium PM2.5 LC"
## [65] "Hafnium PM2.5 LC"
## [66] "Iridium PM2.5 LC"
## [67] "Molybdenum PM2.5 LC"
## [68] "Mercury PM2.5 LC"
## [69] "Gold PM2.5 LC"
## [70] "Lanthanum PM2.5 LC"
## [71] "Niobium PM2.5 LC"
## [72] "Samarium PM2.5 LC"
## [73] "Scandium PM2.5 LC"
## [74] "Tantalum PM2.5 LC"
## [75] "Terbium PM2.5 LC"
## [76] "Uranium PM2.5 LC"
## [77] "Yttrium PM2.5 LC"
## [78] "Tungsten PM2.5 LC"
## [79] "Total Carbon PM2.5 LC TOT"
## [80] "OP PM2.5 LC TOT"
## [81] "OC CSN Unadjusted PM2.5 LC TOT"
## [82] "EC CSN PM2.5 LC TOT"
## [83] "Optical EC PM2.5 LC TOT"
## [84] "Black Carbon PM2.5 at 880 nm"
## [85] "UV Carbon PM2.5 at 370 nm"
## [86] "Non-volatile Nitrate PM2.5 LC"
```

The data were collected over the following range of dates.

```
range(dat$Date.Local)
```

```
## [1] "2014-01-01" "2014-12-31"
```

Summary statistics

Here are the data for Baltimore County.

```
library(dplyr)
```

```
##
## 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
```

```
balt <- filter(dat, State.Code == "24" &
                        County.Code == "005" &
                        Parameter.Name == "Sulfate PM2.5 LC") %>%
        select(Date.Local, Sample.Value)
balt
```

```
## # A tibble: 111 x 2
##    Date.Local Sample.Value
##    <date>            <dbl>
##  1 2014-01-05        2.15
##  2 2014-01-08        1.16
##  3 2014-01-11        1.04
##  4 2014-01-14        0.851
##  5 2014-01-17        5.45
##  6 2014-01-23        2.53
##  7 2014-01-26        1.23
##  8 2014-01-29        1.78
##  9 2014-02-04        2.54
## 10 2014-02-10        6.42
## # ... with 101 more rows
```

Here's a plot of the Sulfate Baltimore County data.

```
library(ggplot2)
qplot(Date.Local, Sample.Value, data = balt)
```