Regional Analysis

Marc Los Huertos

10/1/2020

Introduction

Background

Air qualily data varies seasonally, regionally, and with particular events. These create a distribution of values, some of which pose potential health effects.

Goals

We will collect EPA air quality data to compare with the local data we collect to assses if our data fall within the confidence intervals of the EPA database.

Rationale

By collecting and analyzing air quality data, we will be able to estimate the central tendencies of the data. With these estimates, we can determine if our sensor data fall within these estimates.

Learning Outcomes

We have learned how to manipulate data in R (using the Rstudio) to analyze data before. In this case, we will

Grading – low stakes

As an assignment with relatively low stakes, 10 points, you will be assessed using the following criteria:

1. Did you collect 5 years of PM2.5 from a nearby air quality station.
2. Report the mean, standard deviation, and 95% confidence interval.
3. Plot 5-years of data and create a threshold line based on state and federal standards for PM2.5. Please include the folowing:

- Rotated y-axis
- Appropriate axis labels
- Point graph, EPA or state health limits.
- Estimate the mean, S.D. and 95% confidence intervals for each month.

4. Using the peer reviewed literature, discuss the implications of the results.

Assessing EPA Data

We will use the EPA website (https://www.epa.gov/outdoor-air-quality-data/download-daily-data)to collect the data. Select a location and download 5 years of data. From what I can tell, we'll need to download each year separately and then combine them in R.

NOTE: As you can probably appreciate, every project requires some clean up.

Upload to Rstudio Server Folder

First, we'll upload the data to Rstudio. After getting the data from the EPA downloaded, you can upload the data to R.

NOTE: We are not going to be using the github site, so you can start a new project with a new directory. We won't be needing the collaboration work that we used in our previous project.

Read Data

First, we read data into R, using read.csv() and create an dataframe for each year. I prefer to do this in two steps.

1. Create file path that points to the csv file.

```
SCZ2020.csv <- "/home/CAMPUS/mwl04747/github/EJnPi/Air_Quality_Data_Analysis/ad_viz_plotval_data2020.csv
SCZ2019.csv <- "/home/CAMPUS/mwl04747/github/EJnPi/Air_Quality_Data_Analysis/ad_viz_plotval_data2019.csv
SCZ2018.csv <- "/home/CAMPUS/mwl04747/github/EJnPi/Air_Quality_Data_Analysis/ad_viz_plotval_data2018.csv
```

2. Read the csv files into dataframes

```
SCZ2019 = read.csv(SCZ2019.csv); SCZ2020 = read.csv(SCZ2020.csv)
SCZ2018 = read.csv(SCZ2018.csv)
```

3. Always a good idea of checkin gwhat we created:

```
str(SCZ2020)
```

4. Bind the files together. In this case, we "row bind", rbind(), where each dataframe bound together. This works when all the columns are the same – so if you get an error look at the structure of the files to make sure have the same voarabiel names.

```
SCZ=rbind(SCZ2018, SCZ2019, SCZ2020)
```

Fix Dates

As usually, we need to fix the data to make sure R can read them properly.

1. Create character string of dates. When imported they are defined as factors – so first by strip them of this format and then redefine.

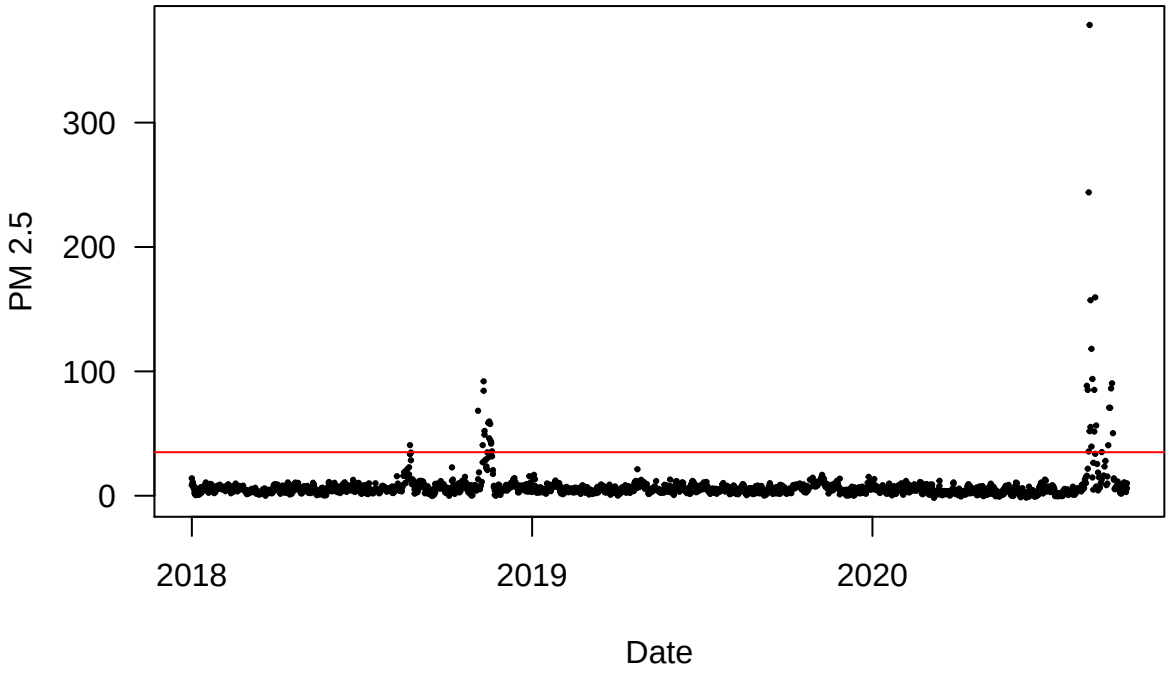
```
Date.char = as.character(SCZ$Date)
# testing to make sure this works...
# as.Date(Date.char, format="%m/%d/%Y")
SCZ$Date = as.Date(Date.char, format="%m/%d/%Y")

str(SCZ)
names(SCZ)
```

Okay, that seems to work.

Graphic Analysis

I have two stations together, might need to decide which one to focus on!



Estimate the Central Tendency for each Month

Next we will determine the central tendency, e.g. average, standard deviation.

So, we will use some packages or libraries to make our job easier. To add packages that are not installed be default, use the `Packages'' tab and click on`Install" – and then search for dpylr and lubridate to install.

the dplyr is described as "a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges."

The lubridate makes it easier manipulate date-times.

We call these libraries in the rcode as below.

```
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
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
##
##     date
str(SCZ)

tmp1 <- mutate(SCZ, Date = ymd(Date),Year = year(Date), Month = month(Date))
tmp2 <- group_by(tmp1, Month, Year)
Monthly <- summarise(tmp2, result = mean(Daily.Mean.PM2.5.Concentration) )
```

We can do the same thing using %>%. This function as has no builtin meaning but the user (or a package) can define operators of the form %whatever% in any way they like. For example, this function will return a string consisting of its left argument followed by a comma and space and then it's right argument.

```
Monthly <- SCZ %>%
  mutate(Date = ymd(Date),Year = year(Date), Month = month(Date)) %>%
  group_by(Month) %>%
  summarise(mean = mean(Daily.Mean.PM2.5.Concentration),
            sd = sd(Daily.Mean.PM2.5.Concentration),
            N = length(Daily.Mean.PM2.5.Concentration) )
```

Estimating Confidence Intervals

Although the standard deviation describe variability, the statistic is not explicilty associated with a probability. We cna convert these to probabilities by mulitplying by a theoretical probability distribution.

```
qnorm(.975)
```

```
## [1] 1.959964
Monthly$UCL95 = Monthly$mean + qnorm(.975)*Monthly$sd/sqrt(Monthly$N)
Monthly$LCL95 = Monthly$mean - qnorm(.975)*Monthly$sd/sqrt(Monthly$N)
```

Table of confidence intervals...

NOTE: At some point, I'll convert the numbers to months – but for now you can translate them for yourself.

```
library(xtable)
Monthly$Months
```

```
## Warning: Unknown or uninitialised column: `Months`.
```

NULL

```
print(xtable(Monthly), include.rownames=FALSE)
```

% latex table generated in R 3.6.0 by xtable 1.8-4 package % Fri Oct 16 08:58:51 2020

Month	mean	sd	N	UCL95	LCL95
1.00	5.94	2.79	182	6.34	5.53
2.00	5.60	2.11	165	5.92	5.28
3.00	3.80	2.22	178	4.13	3.48
4.00	5.14	3.14	175	5.61	4.68
5.00	4.40	2.72	186	4.79	4.01
6.00	5.26	3.15	180	5.72	4.80
7.00	4.86	2.59	179	5.24	4.48
8.00	16.23	39.12	185	21.87	10.59
9.00	8.67	12.82	167	10.62	6.73
10.00	6.53	3.55	111	7.19	5.87
11.00	15.42	17.63	110	18.71	12.12
12.00	5.86	3.13	113	6.44	5.28

Hypothesis Testing

Later Stages – This section is for when we have our pi data (and not done yet :-)

What if want to evalaute a single value and determine if the it the values in inside the confidence intervals – you can use the table above and determine the value, if the parameter is within the expected mean. We'll use the t-test for the test.

If we obtain a mean concentration of 14 using our sensors, we can easily test the value using t.test and all the Septmber values. So, we'll create a vector of all the September readings.

```
September = subset(SCZ, select=Daily.Mean.PM2.5.Concentration, subset = month(SCZ$Date)==9)
```

The null hypothesis is that there is no difference between the mean from the sensor and the values currated by the EPA.

```
t.test(September, mu=14)
```

```
##
##  One Sample t-test
##
## data:  September
## t = -5.3711, df = 166, p-value = 2.609e-07
## alternative hypothesis: true mean is not equal to 14
## 95 percent confidence interval:
##   6.712454 10.630061
## sample estimates:
## mean of x
##  8.671257
```

More likely that we'll have bunch of values from our sensor – not just a mean, then were comparing lots of sensor measurements to the EPA data.

Comparing Sensor Values with EPA Data

First, I'll create simulated data to show how this might be done:

```
simulateddata=rnorm(40, mean=14, sd=12)
```

```r
t.test(September, simulateddata)
```

```
##
##  Welch Two Sample t-test
##
## data:  September and simulateddata
## t = -2.1318, df = 62.579, p-value = 0.03696
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.7837660 -0.2831616
## sample estimates:
## mean of x mean of y
##  8.671257 13.204721
```