Air Quality PM2.5 Case Study

Anandu R

7/30/2020

Air Quality PM2.5 Case Study

Loading the data

```
fileUrl2012 =
  "https://aqs.epa.gov/aqsweb/airdata/annual_conc_by_monitor_2012.zip"
fileUrl1999 =
  "https://aqs.epa.gov/aqsweb/airdata/annual_conc_by_monitor_1999.zip"
if(!file.exists('./data_1999.csv')&&!file.exists('./data_2012.csv')){
  invisible(download.file(fileUrl1999,"./data_1999.zip"))
  invisible(download.file(fileUrl2012,"./data_2012.zip"))
  unzip("data_2012.zip", exdir = getwd())
  unzip("data_1999.zip", exdir = getwd())
}
unlink('./data_1999.zip')
unlink('./data_2012.zip')
tryCatch(
  invisible(
    file.rename(
      "annual_conc_by_monitor_1999.csv","data_1999.csv"
      )
  ),warning = function(w){
      if(file.exists('./data_1999.csv')){
        print("The file \"data_1999.csv\" already exists!")
      }else{
        print("Some error occured while trying to rename the file
              \"annual_conc_by_monitor_1999.csv\" to \"data_1999.csv\"!")
      }
    }
  )
tryCatch(
  invisible(
    file.rename(
      "annual_conc_by_monitor_2012.csv","data_2012.csv"
      )
  ),warning = function(w){
      if(file.exists('./data_1999.csv')){
        print("The file \"data_2012.csv\" already exists!")
      }else{
        print("Some error occured while trying to rename the file
```

```
              \"annual_conc_by_monitor_2012.csv\" to \"data_2012.csv\"!")
      }
    }
  )
data_1999 = read.csv("data_1999.csv")
data_2012 = read.csv("data_2012.csv")
```

Initial analysis on the data

Identifying the parameter code for PM2.5 data

```
invisible(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
```

```
par_code = grep("PM2.5",data_1999$Parameter.Name)[1]
par_code = data_1999$Parameter.Code[par_code]
```

Filtering dataset with only the required parameter list - PM2.5

```
data_1999 = filter(data_1999, Parameter.Code == par_code)
data_2012 = filter(data_2012, Parameter.Code == par_code)
```

Size of the dataset

```
sprintf("1999 data, Rows = %d, Columns = %d",dim(data_1999)[1],dim(data_1999)[2])
```

```
## [1] "1999 data, Rows = 4801, Columns = 55"
```

```
sprintf("2012 data, Rows = %d, Columns = %d",dim(data_2012)[1],dim(data_2012)[2])
```

```
## [1] "2012 data, Rows = 5547, Columns = 55"
```

Subsetting just the particulate matter measure column from each dataset

```
pm_1999 = data_1999$X1st.Max.Value
pm_2012 = data_2012$X1st.Max.Value
```

Structure of variable

```
summary(pm_1999)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.20   29.00   38.80   41.63   48.00  157.10
```

```
summary(pm_2012)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.80   21.70   27.60   38.59   36.50  801.00
```

0% of the data is missing

```
mean(is.na(pm_1999))
```

```
## [1] 0
```

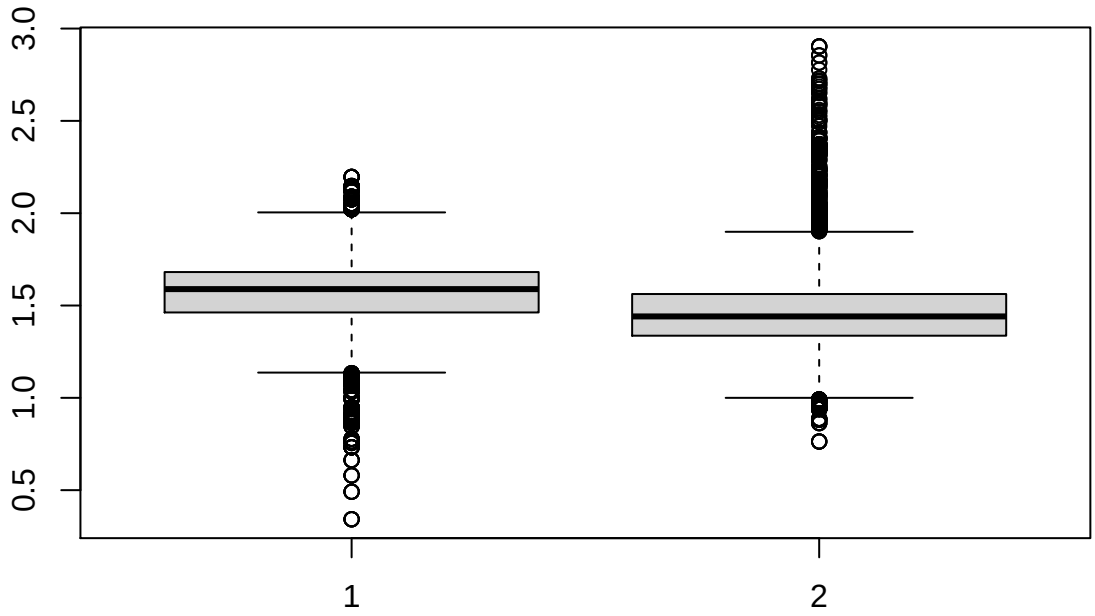
```
mean(is.na(pm_2012))
```

```
## [1] 0
```

Visualization

boxplot

```
boxplot(log10(pm_1999),log10(pm_2012))
```



We observe that eventhough the average values have gone down, the spread of the data has increased to the right extreme, i.e the data has become more right skewed in 2012, from being left skewed in 1999.

Extracting the dates in which the measurements were taken

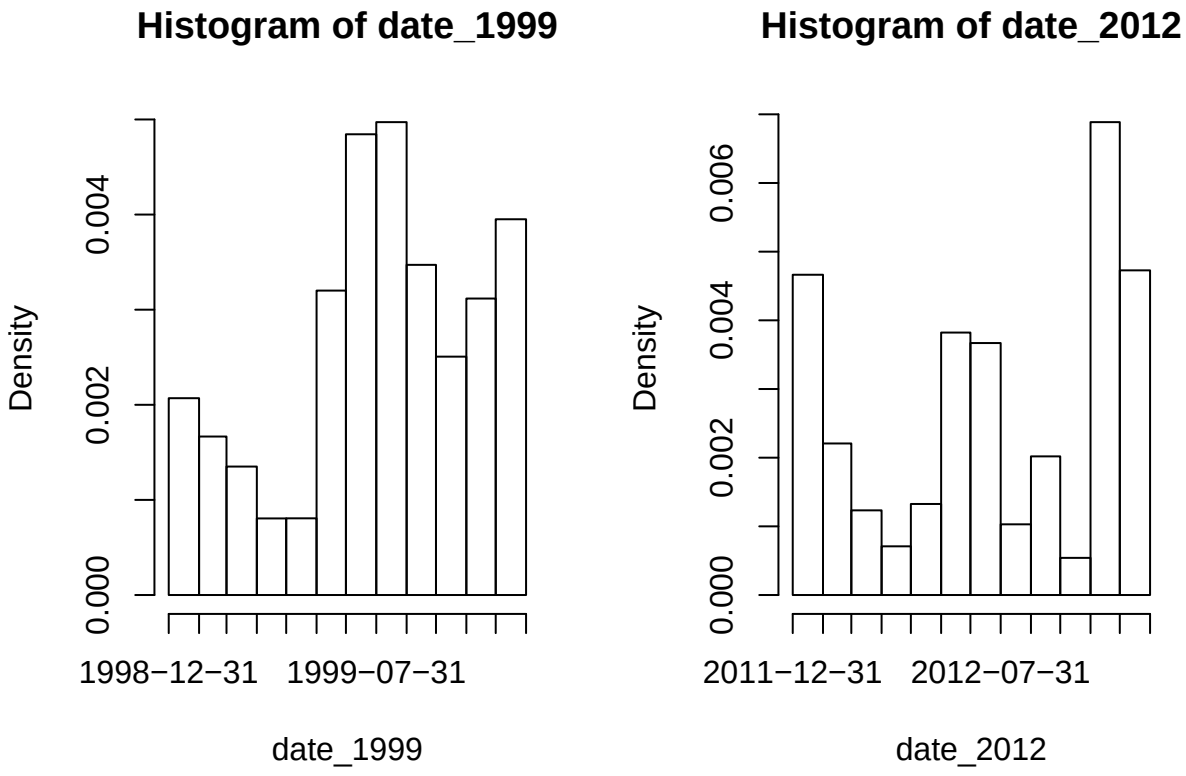
```
date_1999 = as.Date(data_1999$X1st.Max.DateTime)
date_2012 = as.Date(data_2012$X1st.Max.DateTime)
str(date_1999)
```

```
##  Date[1:4801], format: "1999-02-14" "1999-02-14" "1999-02-14" "1999-02-14" "1999-08-28" ...
```

```
str(date_2012)
```

```
##  Date[1:5547], format: "2012-07-29" "2012-07-29" "2012-07-29" "2012-07-29" "2012-06-29" ...
```

```
par(mfrow = c(1,2))
hist(date_1999,"month")
hist(date_2012,"month")
```



To make analysis more simpler we pick a common monitor from 1999 and 2012, and compare the pm2.5 levels between the two instead of the levels of the entire country. This will also allow us to control for possible changes in the monitoring locations between 1999 and 2012. As new monitors are added overtime and this can lead to inaccuracy in the analysis by taking into consideration the entire country.

Now we subset the data to look at the various monitors present in New York City

```
m_NY_1999 = unique(subset(data_1999, State.Code==36, c(County.Code, Site.Num)))
m_NY_2012 = unique(subset(data_2012, State.Code==36, c(County.Code, Site.Num)))

## Creating a new variable that pastes the values of County.Code and Site.Num
## inorder to find the intersecting County-Site combinations in both time periods
## within newyork city
m_NY_1999 = paste(m_NY_1999[,1],m_NY_1999[,2],sep = ".")
m_NY_2012 = paste(m_NY_2012[,1],m_NY_2012[,2],sep = ".")

## Finding the intersecting county-site combinations
common_m = intersect(m_NY_1999,m_NY_2012)
common_m
```

```
##  [1] "1.5"     "1.12"    "5.80"    "5.110"   "13.11"   "29.5"    "31.3"
##  [8] "63.2008" "67.1015" "85.55"   "101.3"
```

Thus we're able to find the common monitor we can use for out analysis Now we need to select the monitor that has the most number of observations, since more data equals better analysis and visualizations

Subsetting the data from NY to only have records of observations that are from the common monitors

```
## Creating a new variable county.site, to subset
m_NY_1999 = data_1999
m_NY_2012 = data_2012
m_NY_1999$County.Site = paste(m_NY_1999$County.Code,m_NY_1999$Site.Num,sep = ".")
m_NY_2012$County.Site = paste(m_NY_2012$County.Code,m_NY_2012$Site.Num,sep = ".")
m_NY_1999 = subset(m_NY_1999, County.Site %in% common_m)
m_NY_2012 = subset(m_NY_2012, County.Site %in% common_m)
```

Looking at the number of observations for each County.Site variable

```
sapply(split(m_NY_1999, m_NY_1999$County.Site),nrow)
```

```
##    1.12     1.5   101.3   13.11    29.5    31.3   5.110    5.80 63.2008 67.1015
##       4       8      12       8       4       4       8       4       8       8
##   85.55
##       4
```

```
sapply(split(m_NY_2012, m_NY_2012$County.Site),nrow)
```

```
##    1.12     1.5   101.3   13.11    29.5    31.3   5.110    5.80 63.2008 67.1015
##       4       8       8       4       4       4       4       4       4       4
##   85.55
##       4
```

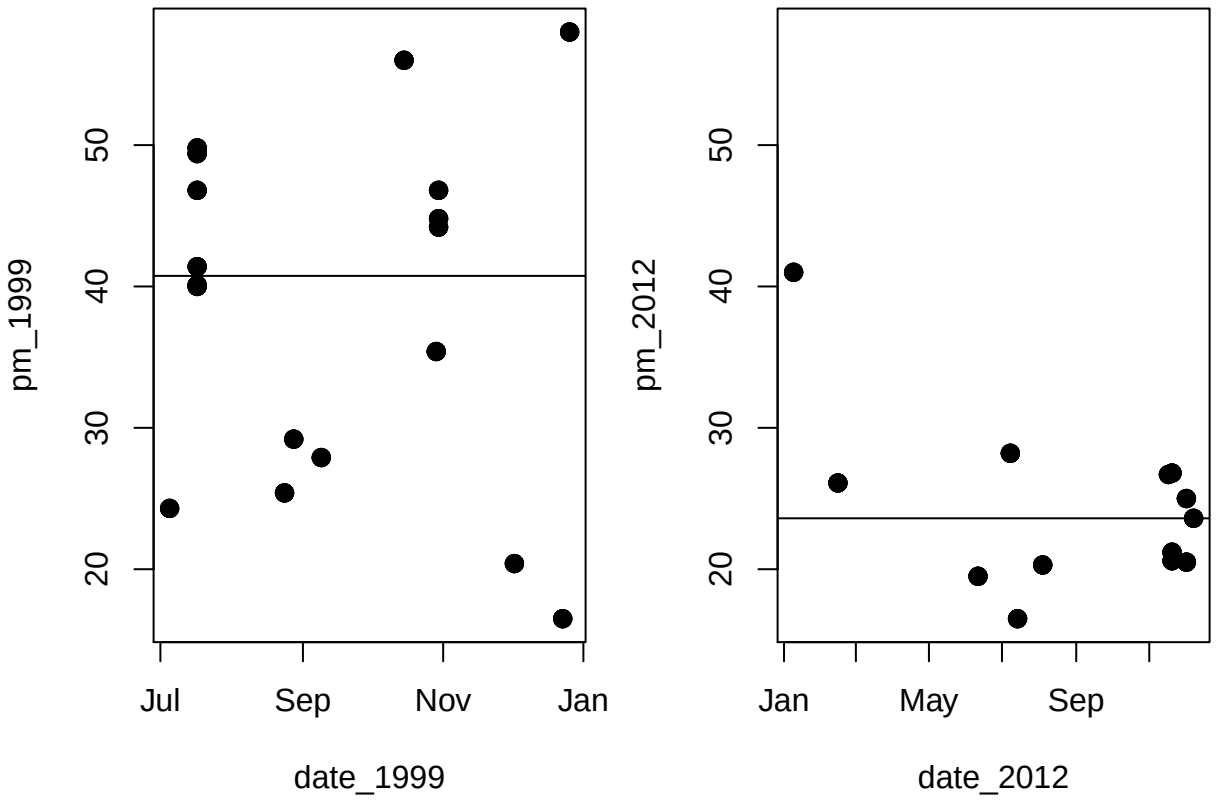
We could pick the county.site combination 1.5 i.e county 1, and site 5 within that county but there is so few observations in each combination of county-site, therefore we shall consider every common County.Site for our analysis.

```
## Originally we were supposed to do the following
## m_1999 = subset(data_1999, State.Code == 36 & County.Code == 1 & Site.Num == 5)
## m_2012 = subset(data_1999, State.Code == 36 & County.Code == 1 & Site.Num == 5)

## But due to lack of observations we proceed with selecting every common
## monitors in New York City
m_1999 = m_NY_1999
m_2012 = m_NY_2012
```

Creating a timeseries analysis for the data

```
pm_1999 = m_1999$X1st.Max.Value
pm_2012 = m_2012$X1st.Max.Value
date_1999 = as.Date(m_1999$X1st.Max.DateTime)
date_2012 = as.Date(m_2012$X1st.Max.DateTime)
par(mfrow = c(1,2), mar = c(5,4,1,1))
rng = range(pm_1999,pm_2012, na.rm = T)
plot(date_1999,pm_1999, pch = 19, cex = 1.2, ylim = rng)
abline(h = median(pm_1999))
plot(date_2012,pm_2012, pch = 19, cex = 1.2, ylim = rng)
abline(h = median(pm_2012))
```



Creating state-wise plot to analyze and compare the PM2.5 levels of the states between the year 1999 and 2012 We calculate the average PM value by each state using the tapply() function

```
mean_1999 = with(data_1999, tapply(X1st.Max.Value,State.Code,mean))
mean_2012 = with(data_2012, tapply(X1st.Max.Value,State.Code,mean))
dc_1999 = data.frame(state = names(mean_1999), mean = mean_1999)
dc_2012 = data.frame(state = names(mean_2012), mean = mean_2012)
```

Viewing the data

1999 means

```
head(dc_1999)
```

```
##    state     mean
## 01    01 48.57778
## 02    02 37.42727
## 04    04 32.29545
## 05    05 32.91364
## 06    06 63.65870
## 08    08 27.52105
```

2012 means

```
head(dc_2012)
```

```
##   state     mean
## 1     1 23.07083
## 2     2 56.21364
## 4     4 55.00199
## 5     5 23.75556
## 6     6 46.56210
## 8     8 28.73895
```

Merging the data of 1999 and 2012

```
mergedData = merge(dc_1999,dc_2012, by = "state")
names(mergedData) = c("State","Mean in 1999","Mean in 2012")
head(mergedData)
```

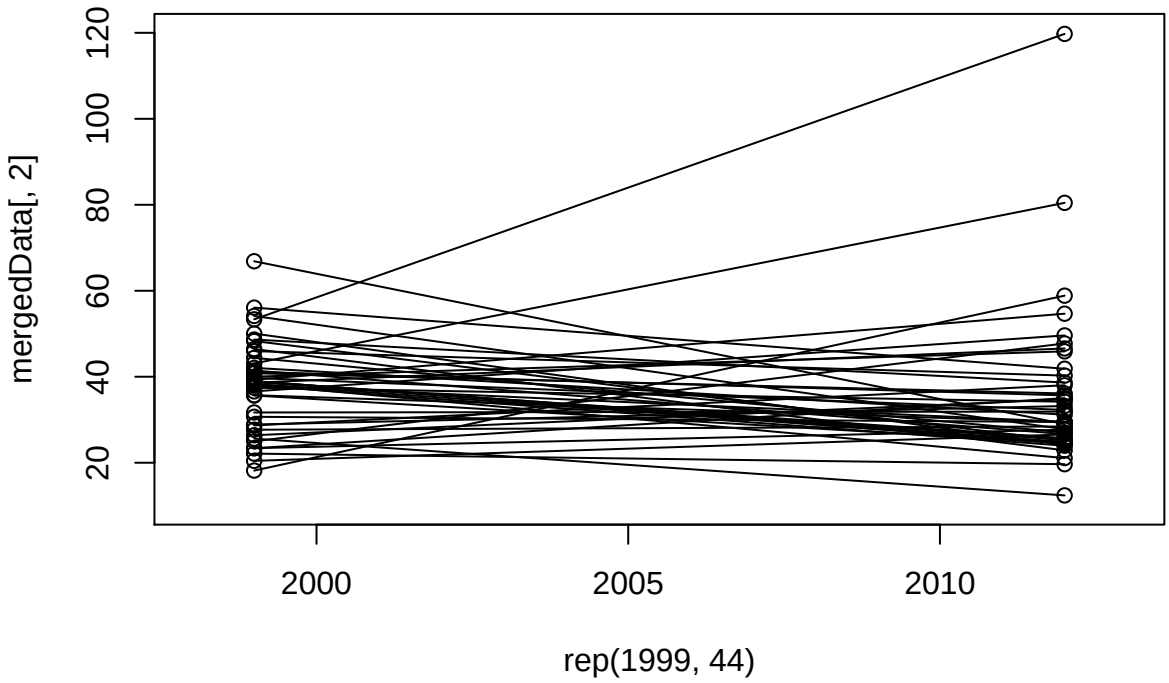
```
##   State Mean in 1999 Mean in 2012
## 1    10     41.43636     26.66364
## 2    11     56.07778     41.86452
## 3    12     38.73929     25.71765
## 4    13     66.86923     28.98909
## 5    15     39.30909     54.66515
## 6    16     53.35385    119.74516
```

```
dim(mergedData)
```

```
## [1] 44  3
```

Visualization of merged data

```
with(mergedData, plot(rep(1999,44),mergedData[,2], xlim = c(1998,2013),ylim = c(10,120)))
with(mergedData, points(rep(2012,44),mergedData[,3]))
segments(rep(1999,44),mergedData[,2],rep(2012,44),mergedData[,3])
```



We observe that most of the states have had their mean PM2.5 value go down over the years when comparing the means between 1999 and 2012.

Deleting files after completion

```
unlink("data_1999.csv")
unlink("data_2012.csv")
```