Is Air Pollution (PM2.5) decreased in the United States?

Mr. Sachin B.

1. Introduction:

- Data analysis looking at changes in fine particulate matter (PM) air pollution in the United States using the Environmental Protection Agencies freely available national monitoring data.
- Particulate matter (less than 2.5 microns in diameter) is a fancy name for dust, and breathing in dust might pose health hazards to the population.
- Our overall hypothesis is that outdoor PM2.5 has decreased on average across the U.S. due to nation-wide regulatory requirements arising from the Clean Air Act.
- To investigate this hypothesis, we obtained PM2.5 data from the U.S. Environmental Protection Agency which is collected from monitors sited across the U.S. We specifically obtained data for the years 1999 and 2012 (the most recent complete year available).

2. Goal: Changes in Fine Particle PM25 Air Pollution in the U.S. from 1999 to 2012

1. on average across the U.S., levels of PM2.5 decreased or not between 1999 and 2012?
2. At one individual monitor, are the levels and that the variability of PM2.5 decreased?
3. Are Most individual states experienced decrease in PM2.5 or not?

2.1 Dataset

```
# download data for 'Air Pollution in 1999'

if(!file.exists("./data")){dir.create("./data")}
download.file("https://raw.githubusercontent.com/jtleek/modules/master/04_ExploratoryAnalysis/CaseStudy/


# download data for 'Air Pollution in 2012'

if(!file.exists("./data")){dir.create("./data")}
download.file("https://raw.githubusercontent.com/jtleek/modules/master/04_ExploratoryAnalysis/CaseStudy/
```

```
# Read in data from 1999

pm0 <- read.table("./data/RD_501_88101_1999-0.txt", comment.char = "#", header = FALSE, sep = "|", na.s

head(pm0)
```

```
##   V1 V2 V3 V4 V5    V6 V7 V8  V9 V10      V11   V12    V13  V14 V15 V16  V17
## 1 RD  I  1 27  1 88101  1  7 105 120 19990103 00:00     NA   AS   3  NA <NA>
## 2 RD  I  1 27  1 88101  1  7 105 120 19990106 00:00     NA   AS   3  NA <NA>
## 3 RD  I  1 27  1 88101  1  7 105 120 19990109 00:00     NA   AS   3  NA <NA>
## 4 RD  I  1 27  1 88101  1  7 105 120 19990112 00:00  8.841 <NA>   3  NA <NA>
## 5 RD  I  1 27  1 88101  1  7 105 120 19990115 00:00 14.920 <NA>   3  NA <NA>
## 6 RD  I  1 27  1 88101  1  7 105 120 19990118 00:00  3.878 <NA>   3  NA <NA>
##   V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28
## 1  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 2  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 3  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 4  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 5  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 6  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

```
dim(pm0)
```

```
## [1] 117421     28
```

```
# Read in data from 2012

pm1 <- read.table("./data/RD_501_88101_2012-0.txt", comment.char = "#", header = FALSE, sep = "|", na.s

head(pm1)
```

```
##   V1 V2 V3 V4 V5    V6 V7 V8  V9 V10      V11   V12 V13  V14 V15 V16  V17  V18
## 1 RD  I  1  3 10 88101  1  7 105 118 20120101 00:00 6.7 <NA>   3  NA <NA> <NA>
## 2 RD  I  1  3 10 88101  1  7 105 118 20120104 00:00 9.0 <NA>   3  NA <NA> <NA>
## 3 RD  I  1  3 10 88101  1  7 105 118 20120107 00:00 6.5 <NA>   3  NA <NA> <NA>
## 4 RD  I  1  3 10 88101  1  7 105 118 20120110 00:00 7.0 <NA>   3  NA <NA> <NA>
## 5 RD  I  1  3 10 88101  1  7 105 118 20120113 00:00 5.8 <NA>   3  NA <NA> <NA>
## 6 RD  I  1  3 10 88101  1  7 105 118 20120116 00:00 8.0 <NA>   3  NA <NA> <NA>
##    V19 V20 V21 V22 V23 V24 V25 V26 V27 V28
## 1 <NA>  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 2 <NA>  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 3 <NA>  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 4 <NA>  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 5 <NA>  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 6 <NA>  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

```
dim(pm1)
```

```
## [1] 1304287      28
```

2.2 Making data ready for Analysis

```
# Reading 1st line in file
cnames <- readLines("./data/RD_501_88101_1999-0.txt",1)
print(cnames)
```

2.2.1 Adding column names to dataset

```
## [1] "# RD|Action Code|State Code|County Code|Site ID|Parameter|POC|Sample Duration|Unit|Method|Date|S
```

```
# Spliting line into vector of string separated by "|"
cnames <- strsplit(cnames,"|", fixed = TRUE)
print(cnames)
```

```
## [[1]]
##  [1] "# RD"                              "Action Code"
##  [3] "State Code"                        "County Code"
##  [5] "Site ID"                           "Parameter"
##  [7] "POC"                               "Sample Duration"
##  [9] "Unit"                              "Method"
## [11] "Date"                              "Start Time"
## [13] "Sample Value"                      "Null Data Code"
## [15] "Sampling Frequency"                "Monitor Protocol (MP) ID"
## [17] "Qualifier - 1"                     "Qualifier - 2"
## [19] "Qualifier - 3"                     "Qualifier - 4"
## [21] "Qualifier - 5"                     "Qualifier - 6"
## [23] "Qualifier - 7"                     "Qualifier - 8"
## [25] "Qualifier - 9"                     "Qualifier - 10"
## [27] "Alternate Method Detectable Limit" "Uncertainty"
```

```
# converting string vector into valid variable names and applying it to both 'pm0' & 'pm1' column name
names(pm0) <- make.names(cnames[[1]])
names(pm1) <- make.names(cnames[[1]])

head(pm0,2)
```

```
##   X..RD Action.Code State.Code County.Code Site.ID Parameter POC
## 1    RD           I          1          27       1     88101   1
## 2    RD           I          1          27       1     88101   1
##   Sample.Duration Unit Method     Date Start.Time Sample.Value Null.Data.Code
## 1               7  105    120 19990103      00:00           NA             AS
## 2               7  105    120 19990106      00:00           NA             AS
##   Sampling.Frequency Monitor.Protocol..MP..ID Qualifier...1 Qualifier...2
## 1                  3                       NA          <NA>            NA
## 2                  3                       NA          <NA>            NA
##   Qualifier...3 Qualifier...4 Qualifier...5 Qualifier...6 Qualifier...7
## 1            NA            NA            NA            NA            NA
## 2            NA            NA            NA            NA            NA
##   Qualifier...8 Qualifier...9 Qualifier...10 Alternate.Method.Detectable.Limit
## 1            NA            NA             NA                                NA
## 2            NA            NA             NA                                NA
##   Uncertainty
## 1          NA
## 2          NA
```

```
head(pm1,2)
```

```
##   X..RD Action.Code State.Code County.Code Site.ID Parameter POC
```

```
## 1    RD           I          1           3      10     88101   1
## 2    RD           I          1           3      10     88101   1
##   Sample.Duration Unit Method     Date Start.Time Sample.Value Null.Data.Code
## 1               7  105    118 20120101      00:00          6.7           <NA>
## 2               7  105    118 20120104      00:00          9.0           <NA>
##   Sampling.Frequency Monitor.Protocol..MP..ID Qualifier...1 Qualifier...2
## 1                  3                       NA          <NA>          <NA>
## 2                  3                       NA          <NA>          <NA>
##   Qualifier...3 Qualifier...4 Qualifier...5 Qualifier...6 Qualifier...7
## 1          <NA>            NA            NA            NA            NA
## 2          <NA>            NA            NA            NA            NA
##   Qualifier...8 Qualifier...9 Qualifier...10 Alternate.Method.Detectable.Limit
## 1            NA            NA             NA                                NA
## 2            NA            NA             NA                                NA
##   Uncertainty
## 1          NA
## 2          NA
```

```
# 1999
dates0 <- pm0$Date

str(dates0)
```

2.2.2 Converting numeric date to Date format

```
##  int [1:117421] 19990103 19990106 19990109 19990112 19990115 19990118 19990121 19990124 19990127 1999
```

```
dates0 <- as.Date(as.character(dates0),"%Y%m%d")

head(dates0)
```

```
## [1] "1999-01-03" "1999-01-06" "1999-01-09" "1999-01-12" "1999-01-15"
## [6] "1999-01-18"
```

```
# 2012
dates1 <- pm1$Date

str(dates1)
```

```
##  int [1:1304287] 20120101 20120104 20120107 20120110 20120113 20120116 20120119 20120122 20120125 201
```

```
dates1 <- as.Date(as.character(dates1),"%Y%m%d")

head(dates1)
```

```
## [1] "2012-01-01" "2012-01-04" "2012-01-07" "2012-01-10" "2012-01-13"
## [6] "2012-01-16"
```

```r
pm0$Date <- dates0
head(pm0$Date)
```

```
## [1] "1999-01-03" "1999-01-06" "1999-01-09" "1999-01-12" "1999-01-15"
## [6] "1999-01-18"
```

```r
pm1$Date <- dates1
head(pm1$Date)
```

```
## [1] "2012-01-01" "2012-01-04" "2012-01-07" "2012-01-10" "2012-01-13"
## [6] "2012-01-16"
```

3. Goal 1: on average across the U.S., levels of PM2.5 decreased or not between 1999 and 2012?

3.1 Seperating pm25 data column i.e. 'Sample.Value'

```r
x0 <- pm0$Sample.Value
x1 <- pm1$Sample.Value
```

3.2 exploring pm25 data for year 1999 & 2012

```r
# 1999
class(x0)
```

```
## [1] "numeric"
```

```r
str(x0)
```

```
##  num [1:117421] NA NA NA 8.84 14.92 ...
```

```r
summary(x0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    7.20   11.50   13.74   17.90  157.10   13217
```

```r
# 2012
class(x1)
```

```
## [1] "numeric"
```

```r
str(x1)
```

```
##  num [1:1304287] 6.7 9 6.5 7 5.8 8 7.9 8 6 9.6 ...
```

```
summary(x1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -10.00    4.00    7.63    9.14   12.00  908.97   73133
```

Explanation:

- Max level of "x1" i.e. 2012 data is very high.
- Min level of "x1" i.e. 2012 data is negative which is practically not possible. It may be a problem with monitor
- Large number of "NA" Values

```
negative <- x1<0

str(negative)
```

3.2.1 Negative Value occurance Investigation

```
##  logi [1:1304287] FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
sum(negative,na.rm = TRUE)
```

```
## [1] 26474
```

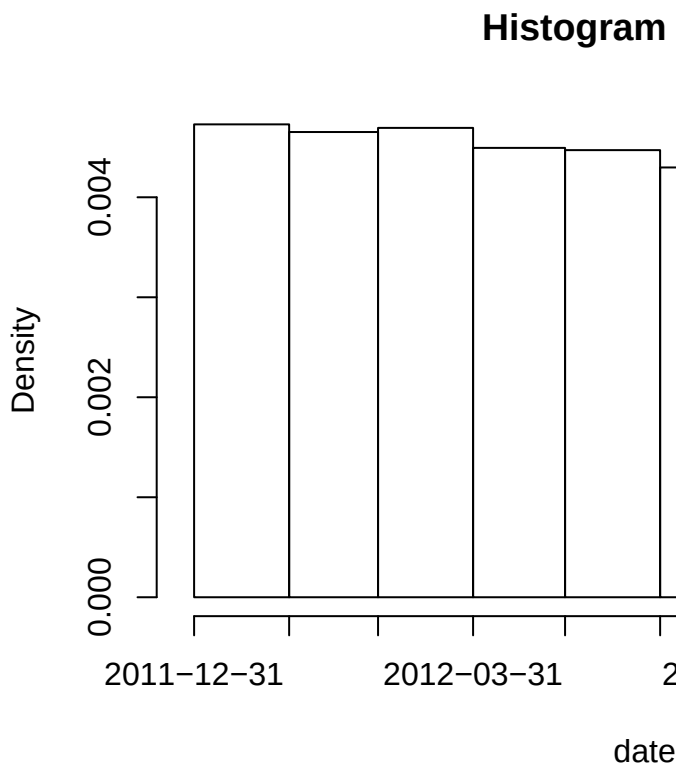
```
mean(negative, na.rm = TRUE)
```

```
## [1] 0.0215034
```

Explanation:

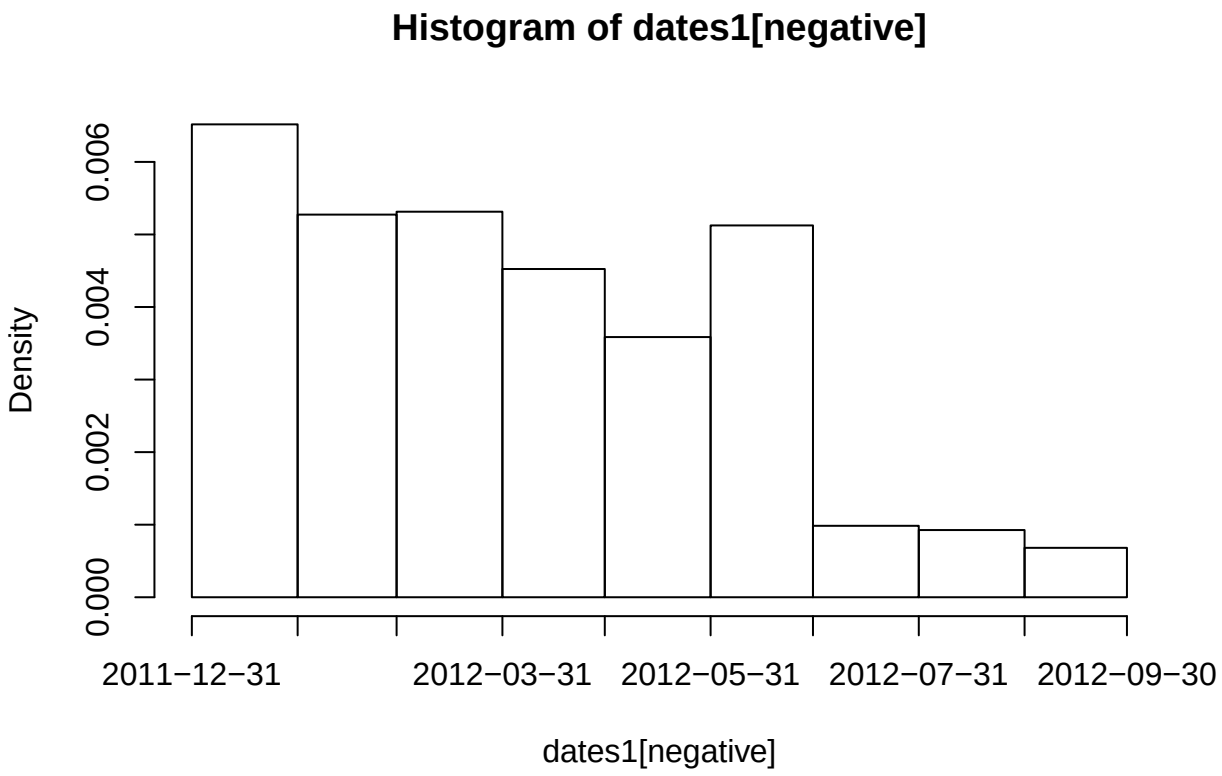
- Percentage of negative values are around 2% which we can ignore.

```
# 2012

# We have already converted Date column in Date format
## dates1 <- as.Date(as.character(pm1$Date),"%Y%m%d")
## head(dates1)

# Normal Reading dates by month
hist(dates1,"month")
```



3.2.2 Negative value occured in particular month/season

```
hist(dates1[negative],"month")
```



Explanation:

- Maximum values for negative occurred from December to March but the reason still not clear so it would be worth investigating. But we have different goal to achieve and 2% negative values can be ignored.

3.3 Are missing data a Problem?

- if missing values are below 5% then we can ignore it but more than that it will going to affect the analysis.

```
# 1999
mean(is.na(x0))
```

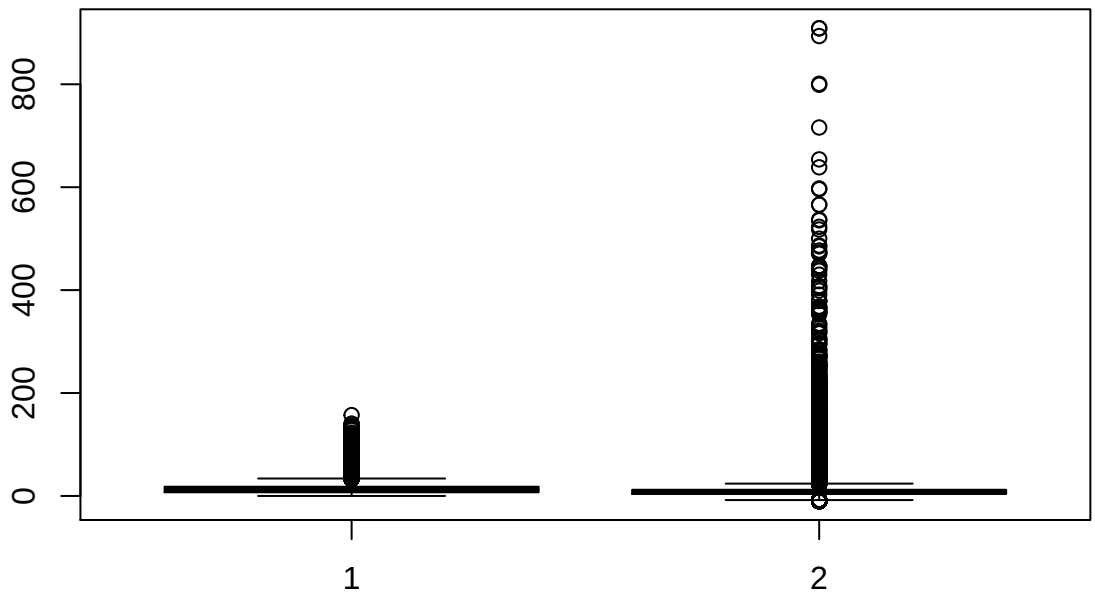
```
## [1] 0.1125608
```

```
# 2012
mean(is.na(x1))
```

```
## [1] 0.05607125
```

3.4 Boxplot

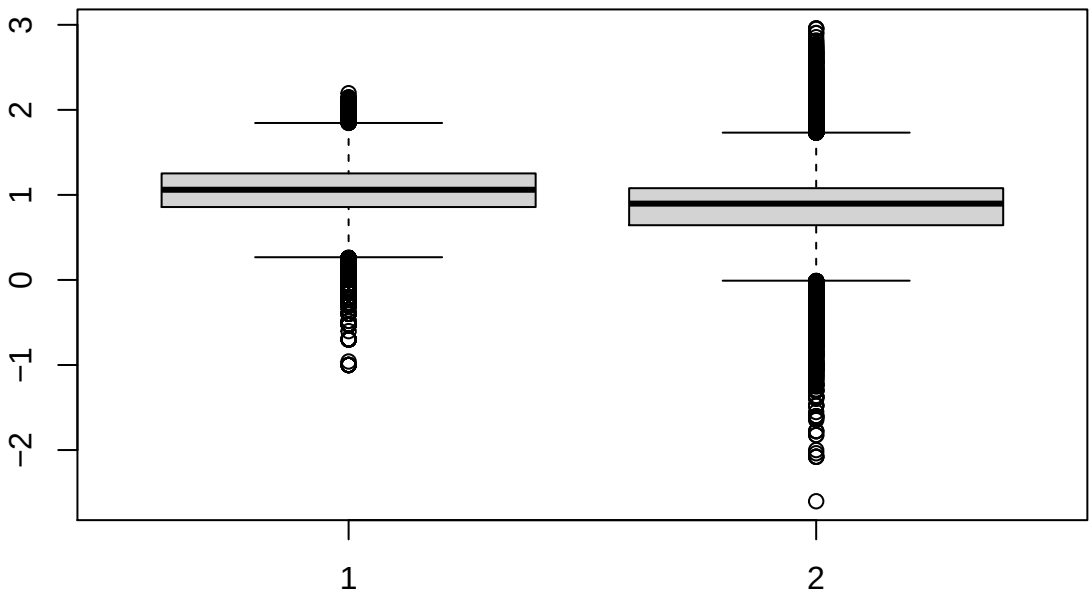
```
boxplot(x0,x1)
```



Explanation:

- Hard to look at the spread of data.
- Data is right skewed.
- Max level of "x1" i.e. 2012 data is very high.

```
boxplot(log10(x0),log10(x1))
```



Explanation:

- Now we can see median of the data is reduced from 1999 to 2012.
- Spread of 2012 is increased with lots of outliers

4. Goal 2: At one individual monitor, are the levels and that the variability of PM2.5 decreased?

4.1 Find a monitor for New York State that exists in both datasets

```
site0 <- unique(subset(pm0,State.Code == 36, c(County.Code, Site.ID)))

site1 <- unique(subset(pm1,State.Code == 36, c(County.Code, Site.ID)))


# Join County.Code and Site.ID column together

site0 <- paste(site0[,1],site0[,2], sep = ".")

site1 <- paste(site1[,1],site1[,2], sep = ".")

str(site0)
```

```
##  chr [1:33] "1.5" "1.12" "5.73" "5.80" "5.83" "5.110" "13.11" "27.1004" ...
```

```
str(site1)
```

```
##  chr [1:18] "1.5" "1.12" "5.80" "5.133" "13.11" "29.5" "31.3" "47.122" ...
```

```
# Select only common county code and site ID available in 1999 and 2012

both <- intersect(site0, site1)

str(both)
```

```
##  chr [1:10] "1.5" "1.12" "5.80" "13.11" "29.5" "31.3" "63.2008" "67.1015" ...
```

4.2 Find how many observations available at each monitor

```
pm0$county.site <- paste(pm0$County.Code,pm0$Site.ID,sep = ".")

pm1$county.site <- paste(pm1$County.Code,pm1$Site.ID,sep = ".")

cnt0 <- subset(pm0, State.Code ==36 & county.site %in% both)

cnt1 <- subset(pm1, State.Code ==36 & county.site %in% both)

table(cnt0$county.site)
```

```
##
##    1.12     1.5   101.3   13.11    29.5    31.3    5.80 63.2008 67.1015   85.55
##      61     122     152      61      61     183      61     122     122       7
```

```
table(cnt1$county.site)
```

```
##
##    1.12     1.5   101.3   13.11    29.5    31.3    5.80 63.2008 67.1015   85.55
##      31      64      31      31      33      15      31      30      31      31
```

4.3 Choose county 63 and side ID 2008

```
pm0sub <- subset(pm0, State.Code == 36 & county.site == 63.2008)
pm1sub <- subset(pm1, State.Code == 36 & county.site == 63.2008)

dim(pm0sub)
```

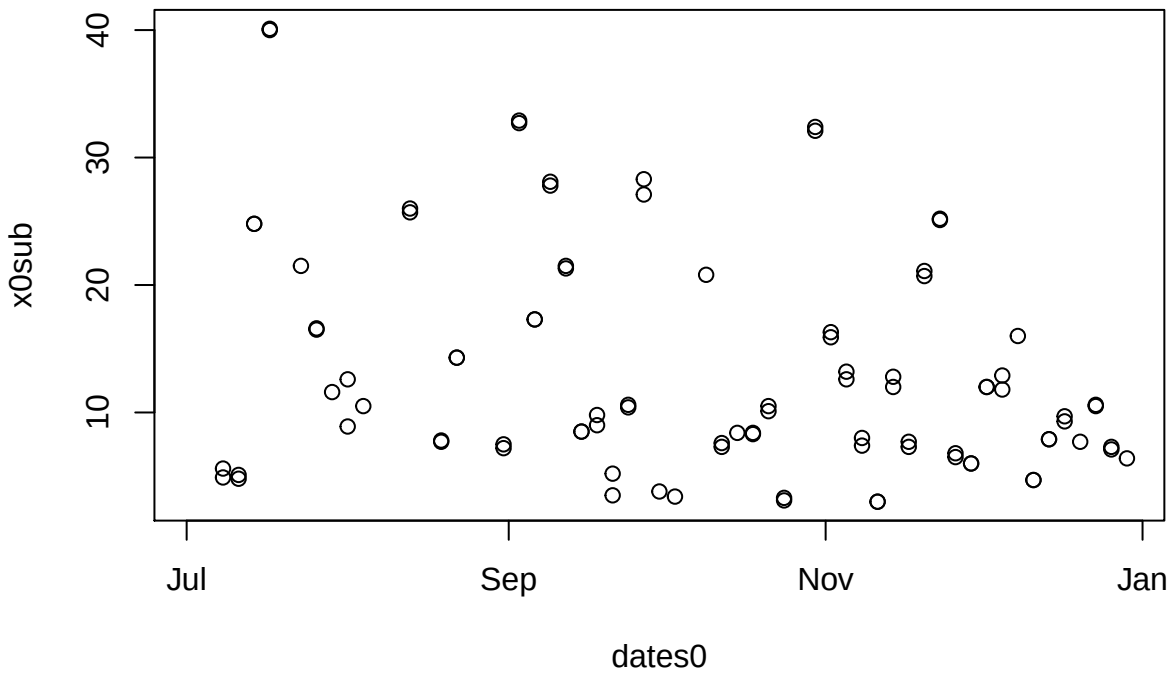
```
## [1] 122  29
```

```r
dim(pm1sub)
```

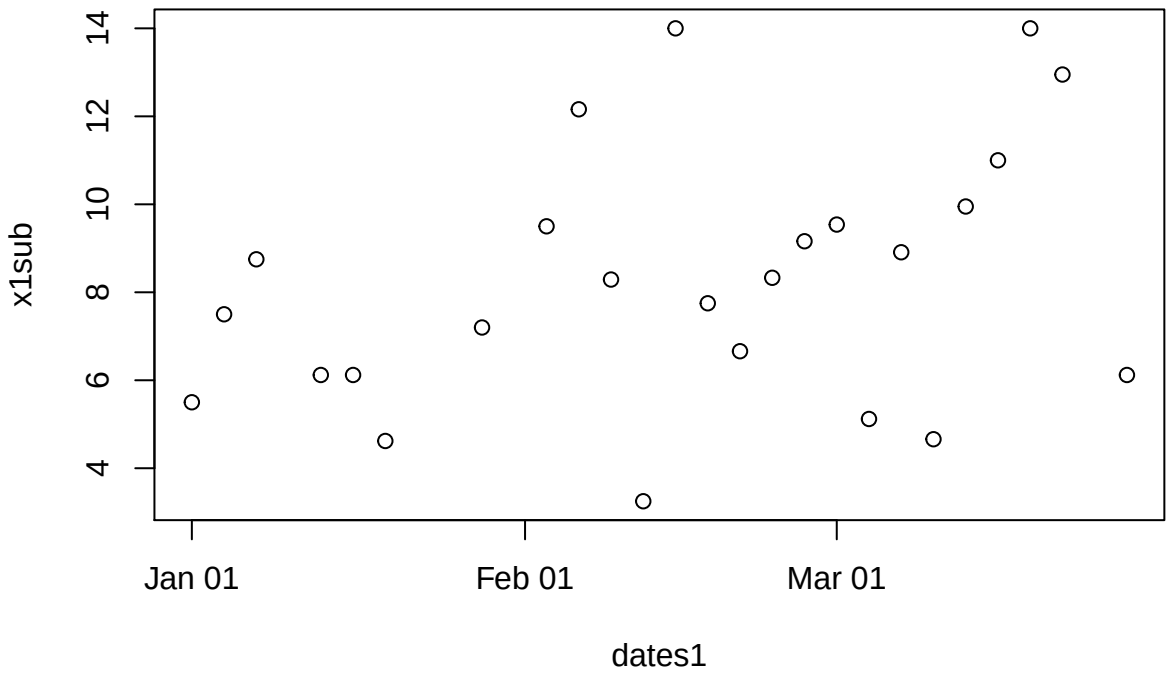
```
## [1] 30 29
```

4.4 Plot data for 1999 and 2012

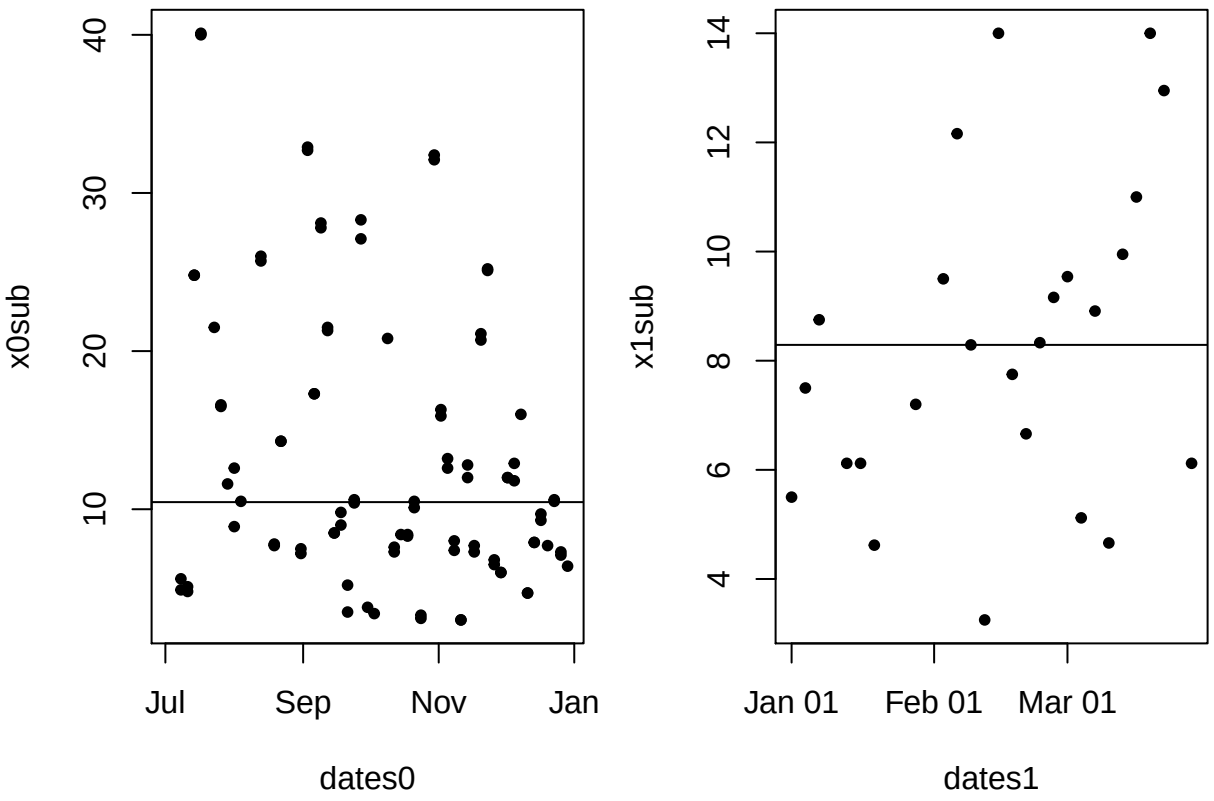
```r
# Plot data for 1999

x0sub <- pm0sub$Sample.Value
dates0 <- pm0sub$Date

plot(dates0,x0sub)
```




```r
# Plot data for 2012

x1sub <- pm1sub$Sample.Value
dates1 <- pm1sub$Date

plot(dates1,x1sub)
```




```
## Plot data for both years in same panel

par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))
plot(dates0, x0sub, pch = 20)
abline(h = median(x0sub, na.rm = T))
plot(dates1, x1sub, pch = 20)
abline(h = median(x1sub, na.rm = T))
```



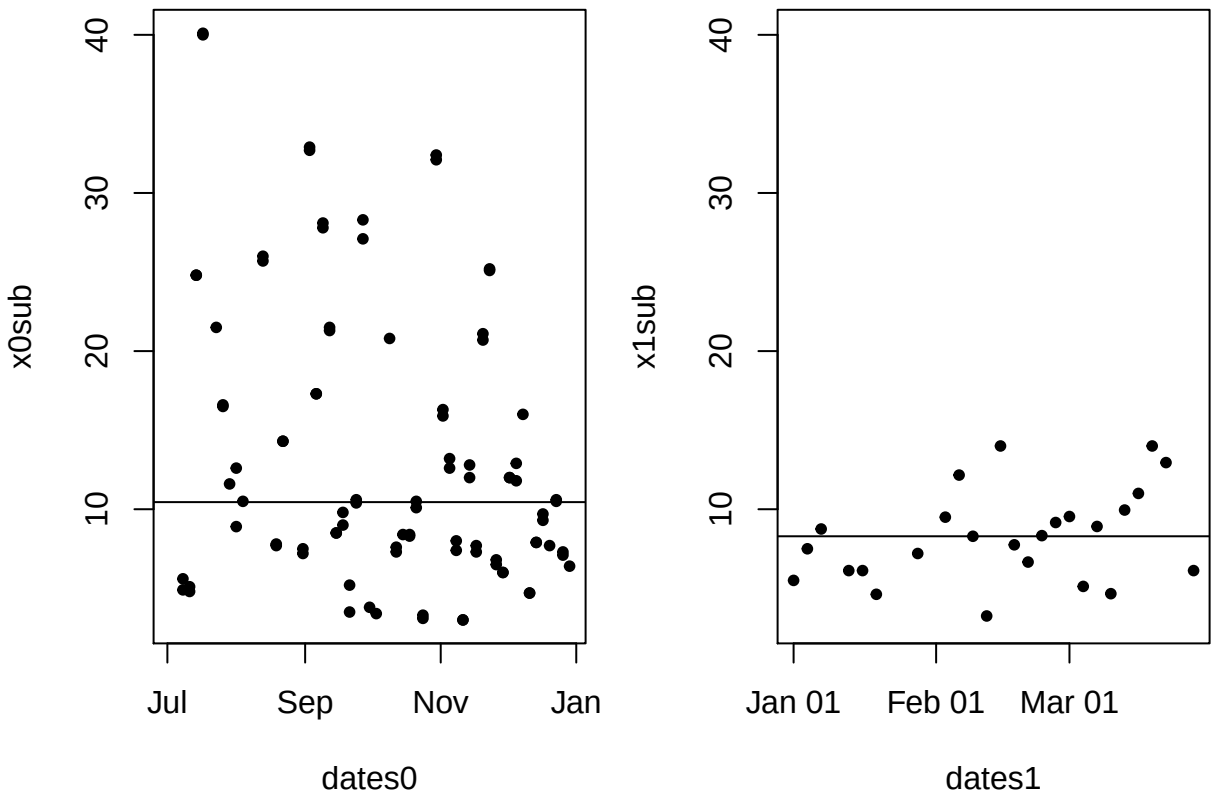
Explanation:

- Though both the plots are correct but their y-lim must be same range so that it would be easy for comparison

```
rng <- range(x0sub,x1sub, na.rm = TRUE)
rng
```

```
## [1]  3.0 40.1
```

```
# copy paste earlier plot with 2 panel but add y-lim

par(mfrow = c(1, 2), mar = c(4, 4, 2, 1))
plot(dates0, x0sub, pch = 20, ylim = rng)
abline(h = median(x0sub, na.rm = T))
plot(dates1, x1sub, pch = 20, ylim = rng)
abline(h = median(x1sub, na.rm = T))
```



5. Goal 3: Are Most individual states experienced decrease in PM2.5 or not?

5.1 state-wide mean calculation

```
head(pm0)
```

```
##   X..RD Action.Code State.Code County.Code Site.ID Parameter POC
## 1    RD           I          1          27       1     88101   1
## 2    RD           I          1          27       1     88101   1
## 3    RD           I          1          27       1     88101   1
## 4    RD           I          1          27       1     88101   1
## 5    RD           I          1          27       1     88101   1
## 6    RD           I          1          27       1     88101   1
##   Sample.Duration Unit Method       Date Start.Time Sample.Value Null.Data.Code
## 1               7  105    120 1999-01-03      00:00           NA             AS
## 2               7  105    120 1999-01-06      00:00           NA             AS
## 3               7  105    120 1999-01-09      00:00           NA             AS
## 4               7  105    120 1999-01-12      00:00        8.841           <NA>
## 5               7  105    120 1999-01-15      00:00       14.920           <NA>
## 6               7  105    120 1999-01-18      00:00        3.878           <NA>
##   Sampling.Frequency Monitor.Protocol..MP..ID Qualifier...1 Qualifier...2
## 1                  3                      NA          <NA>            NA
## 2                  3                      NA          <NA>            NA
```

```
## 3                  3                       NA          <NA>            NA
## 4                  3                       NA          <NA>            NA
## 5                  3                       NA          <NA>            NA
## 6                  3                       NA          <NA>            NA
##   Qualifier...3 Qualifier...4 Qualifier...5 Qualifier...6 Qualifier...7
## 1            NA            NA            NA            NA            NA
## 2            NA            NA            NA            NA            NA
## 3            NA            NA            NA            NA            NA
## 4            NA            NA            NA            NA            NA
## 5            NA            NA            NA            NA            NA
## 6            NA            NA            NA            NA            NA
##   Qualifier...8 Qualifier...9 Qualifier...10 Alternate.Method.Detectable.Limit
## 1            NA            NA             NA                                NA
## 2            NA            NA             NA                                NA
## 3            NA            NA             NA                                NA
## 4            NA            NA             NA                                NA
## 5            NA            NA             NA                                NA
## 6            NA            NA             NA                                NA
##   Uncertainty county.site
## 1          NA        27.1
## 2          NA        27.1
## 3          NA        27.1
## 4          NA        27.1
## 5          NA        27.1
## 6          NA        27.1
```

```
mn0 <- tapply(pm0$Sample.Value,pm0$State.Code,mean,na.rm= TRUE)
str(mn0)
```

```
##  num [1:53(1d)] 19.96 6.67 10.8 15.68 17.66 ...
##  - attr(*, "dimnames")=List of 1
##   ..$ : chr [1:53] "1" "2" "4" "5" ...
```

```
summary(mn0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.862   9.519  12.315  12.406  15.640  19.956
```

```
mn1 <- tapply(pm1$Sample.Value,pm1$State.Code,mean,na.rm= TRUE)
str(mn1)
```

```
##  num [1:52(1d)] 10.13 4.75 8.61 10.56 9.28 ...
##  - attr(*, "dimnames")=List of 1
##   ..$ : chr [1:52] "1" "2" "4" "5" ...
```

```
summary(mn1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.006   7.355   8.729   8.759  10.613  11.992
```

5.2 Makeing separate data frames for states

```
d0 <- data.frame(state = names(mn0), mean = mn0)
d1 <- data.frame(state = names(mn1), mean = mn1)
mrg <- merge(d0, d1, by = "state")
dim(mrg)
```

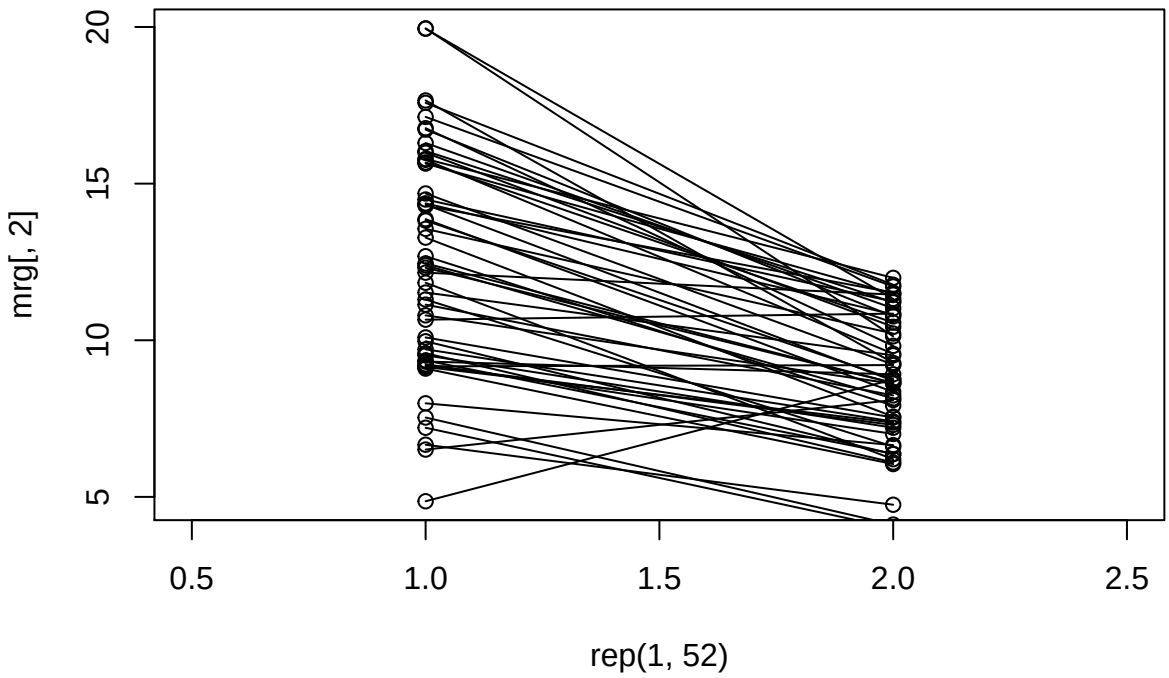
```
## [1] 52  3
```

```
head(mrg)
```

```
##   state    mean.x    mean.y
## 1     1 19.956391 10.126190
## 2    10 14.492895 11.236059
## 3    11 15.786507 11.991697
## 4    12 11.137139  8.239690
## 5    13 19.943240 11.321364
## 6    15  4.861821  8.749336
```

5.3 Plot for states experienced decrease in PM2.5

```
with(mrg, plot(rep(1, 52), mrg[, 2], xlim = c(.5, 2.5)))
with(mrg, points(rep(2, 52), mrg[, 3]))

segments(rep(1, 52), mrg[, 2], rep(2, 52), mrg[, 3])
```


mrg[, 2]
20
15
10
5
0.5
1.0
1.5
2.0
2.5
rep(1, 52)