AS2: Exploring Data via Visualization

* The following instructions contain a set of questions (highlighted in purple) you will need to solve using the R statistical programming language.

Please make sure to submit 1) a document (e.g., MS Word) containing your answers to each question, and 2) the script file used for the assignment.

(1) Import and Preprocess Data

(a) First, import the datasets using the following links 1) "https://bit.ly/3c4AHbL" for 1999 data, and 2) "https://bit.ly/3nZicL2" for 2012 data using the data.table package.p.s., set colClasses of the first 5 variables to "character" and the rest of it to "numeric."

```r
data_1999 <- fread('https://bit.ly/3c4AHbL',colClasses=list(character=1:5, numeric=6:12))
data_2012 <- fread('https://bit.ly/3nZicL2',colClasses=list(character=1:5, numeric=6:12))
#to know each classes of the column
lapply(data_1999, class)
lapply(data_2012, class)
```

```
> #to know each class
> lapply(data_1999, c
$X..RD
[1] "character"

$Action.Code
[1] "character"

$State.Code
[1] "character"

$County.Code
[1] "character"

$Site.ID
[1] "character"

$Parameter
[1] "numeric"

$POC
[1] "numeric"

$Sample.Duration
[1] "numeric"

$Unit
[1] "numeric"

$Method
[1] "numeric"

$Date
[1] "numeric"

$Sample.Value
[1] "numeric"
```

```
> lapply(data_2012, class)
$X..RD
[1] "character"

$Action.Code
[1] "character"

$State.Code
[1] "character"

$County.Code
[1] "character"

$Site.ID
[1] "character"

$Parameter
[1] "numeric"

$POC
[1] "numeric"

$Sample.Duration
[1] "numeric"

$Unit
[1] "numeric"

$Method
[1] "numeric"

$Date
[1] "numeric"

$Sample.Value
[1] "numeric"

>
```

(b) Take a look at the 1999 data by (1) printing out the dimensions and (2) the first 3 rows.

```
> dim(data_1999)#[1] 117421      12
[1] 117421      12
> head(data_1999,3)
   X..RD Action.Code State.Code County.Code Site.ID Parameter POC Sample.Duration Unit Method     Date Sample.Value
1:    RD           I         01         027    0001     88101   1               7  105    120 19990103           NA
2:    RD           I         01         027    0001     88101   1               7  105    120 19990106           NA
3:    RD           I         01         027    0001     88101   1               7  105    120 19990109           NA
> 
```

(c) The variable of our interest is Sample.Value which contains the PM2.5 measurements. (3) Using the 1999 data, print the summary statistics of the variable with summary().

```
> summary(data_1999$Sample.Value)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00    7.20   11.50   13.74   17.90  157.10   13217 
> 
```

i) We observe some missing values in the observations of the PM2.5 measurements (n = 13,217). Compute the number of NAs using table() and is.na(), then divide the numbers by the total number of observations in the data to calculate the proportions.

```
> #[3-i]
> #sum(is.na(data_1999))
> temp <- table(is.na(data_1999$Sample.Value));temp

 FALSE   TRUE 
104204  13217 
> theNAData <-temp[2]/sum(temp)#0.1125608
> theNAData
     TRUE 
0.1125608 
> 
```

ii) (4) What is the percentage of the PM2.5 observations that are missing (round up to 3 decimal places)?

```
> # [3-ii]
> temp <- as.character(round(theNAData*100,3))
> sprintf("NA portion :%s percent",temp)
[1] "NA portion :11.256 percent"
```

(d) Bind the 1999 data and 2012 data and assign the aggregated data to an object called ‘pm’. Then, subset the years from the Date variable and convert it into a factor variable called ‘year’.

```
> pm <- rbind(data_1999,data_2012)
> pm <- mutate(pm,Year = as.factor(year(ymd(pm$Date))))
> class(pm$Year)
[1] "factor"
> 
```

(e) Next, rename the Sample.Value variable to PM which better expresses the values stored in the variable.

```
> #[e] Next, rename the Sample.Value variable to PM which better expresses the values stored in the variable.
> str(pm)
Classes ‘data.table’ and 'data.frame':  1421708 obs. of  13 variables:
 $ X..RD          : chr  "RD" "RD" "RD" "RD" ...
 $ Action.Code    : chr  "I" "I" "I" "I" ...
 $ State.Code     : chr  "01" "01" "01" "01" ...
 $ County.Code    : chr  "027" "027" "027" "027" ...
 $ Site.ID        : chr  "0001" "0001" "0001" "0001" ...
 $ Parameter      : num  88101 88101 88101 88101 88101 ...
 $ POC            : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Sample.Duration: num  7 7 7 7 7 7 7 7 7 7 ...
 $ Unit           : num  105 105 105 105 105 105 105 105 105 105 ...
 $ Method         : num  120 120 120 120 120 120 120 120 120 120 ...
 $ Date           : num  2e+07 2e+07 2e+07 2e+07 2e+07 ...
 $ Sample.Value   : num  NA NA NA 8.84 14.92 ...
 $ Year           : Factor w/ 2 levels "1999","2012": 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, ".internal.selfref")=<externalptr>
> pm <- pm %>% rename(PM = Sample.Value)
> colnames(pm)
 [1] "X..RD"           "Action.Code"     "State.Code"      "County.Code"     "Site.ID"         "Parameter"
 [7] "POC"             "Sample.Duration" "Unit"            "Method"          "Date"            "PM"
[13] "Year"
> 
```

(2) Aggregate data analysis:

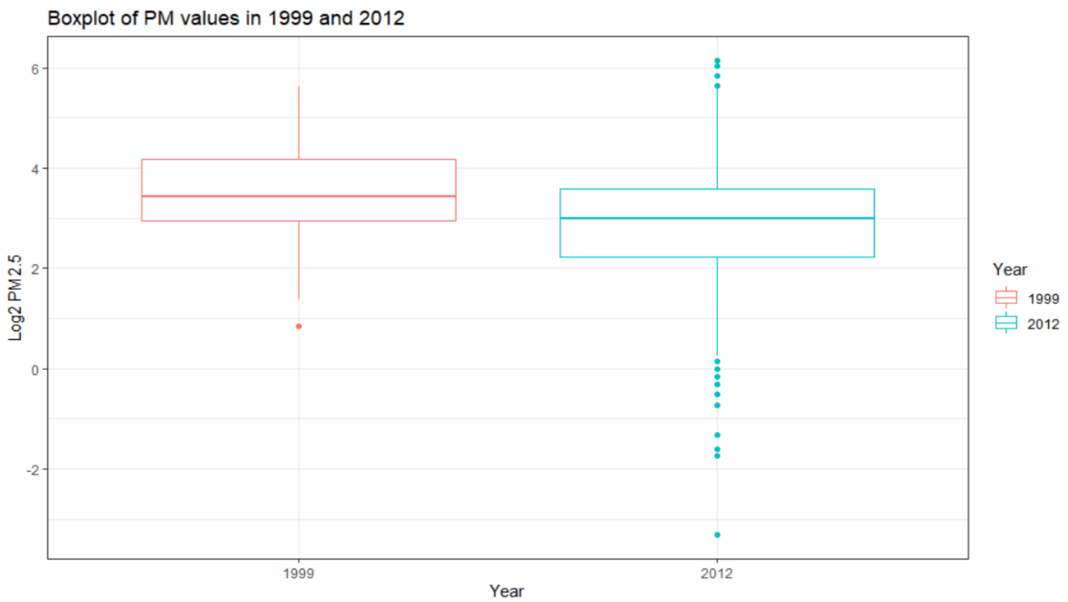
We want to visualize the aggregate changes in PM across the entire monitoring network.

(a) First, for better visibility and reproducibility**, (5) set the seed at 2021 and draw 1,000 randomly** selected samples from the data (i.e., pm) using the sampling function **in dplyr package.**

```
# [a] First, for better visibility and reproducibility,
#(5) set the seed at 2021 and draw 1,000 randomly
# selected samples from the data (i.e., pm) using the sampling function in dplyr package.

set.seed(2021)
sub_sample <- sample_n(pm,1000) #i use sample_n function  in dplyr package
```

(b) create boxplots of all monitor values in 1999 and 2012 using the randomly sampled data as shown below. (6) log of the PM values (7) label the title, x-axis & y-axis, and (8) use the base white theme

```
ggplot(sub_sample, aes(x = Year, y = log2(PM), color = Year))+
    geom_boxplot()+
    labs(title = "Boxplot of PM values in 1999 and 2012", x = 'Year',y= 'Log2 PM2.5')+
    theme(legend.position = "white")+
    theme_bw()
```



(c) (9) Describe what you observe in terms of the means and variances of the observations in 1999 and 2012?

we can use the method below to reach the value of the mean and variance for 1999 and 2012.

```
#mean and variance
mean(data_1999$Sample.Value, na.rm = TRUE)#[1] 13.7381

mean(data_2012$Sample.Value, na.rm = TRUE)#[1] 9.139924

var(data_1999$Sample.Value, na.rm = TRUE)#[1] 88.54687

var(data_2012$Sample.Value, na.rm = TRUE)#[1] 73.21078
```

The mean of PM value in 1999 is higher than the 2012's. However, the variance of the PM value of 1999 is higher than the 2012's too.

(d) Our first task is to identify a monitor in New York State that has data in 1999 and 2012 (not all monitors operated during both time periods).

(10) Subset the data to include only the observations from New York (i.e., State.Code == 36) and only include the County.Code and the Site.ID (i.e. monitor number) variables using filter(), select(), and unique().

```
ny_data <- filter(pm, pm$State.Code == 36); ny_data
subset_ny_data  <- select(ny_data, County.Code:Site.ID);subset_ny_data

colnames(subset_ny_data)
count(unique(subset_ny_data))
```

```
> colnames(subset_ny_data)
[1] "County.Code" "Site.ID"
```

focus on a single monitor in **NY** to observe/visualize the changes to account for this possibility.

e) (11) Create a new variable called Site.Code that combines the county code and the site ID into a single string by using paste() with "." as the separator**.**

```
#Changes in PM levels at an individual monitor:
#[e](11) Create a new variable called Site.Code by using paste() with "."

colnames(ny_data) #沒有Site.Code欄位
ny_data <- mutate(ny_data,Site.Code = paste(ny_data$County.Code,ny_data$Site.ID, sep = '.') )
head(ny_data)
```

144:40 (Untitled)

Console Terminal Jobs

R 4.1.1 · ~/

```
3264: 20120319  9.58 2012  103.0002
3265: 20120322 10.33 2012  103.0002
3266: 20120325  5.41 2012  103.0002
3267: 20120328  9.62 2012  103.0002
3268: 20120331  6.25 2012  103.0002
> ny_data <- mutate(ny_data,Site.Code = paste(ny_data$County.Code,ny_data$Site.ID, sep = '.') )
> head(ny_data)
   X..RD Action.Code State.Code County.Code Site.ID Parameter POC Sample.Duration Unit Method
1:    RD           I         36         001    0005     88101   1               7  105    118
2:    RD           I         36         001    0005     88101   1               7  105    118
3:    RD           I         36         001    0005     88101   1               7  105    118
4:    RD           I         36         001    0005     88101   1               7  105    118
5:    RD           I         36         001    0005     88101   1               7  105    118
6:    RD           I         36         001    0005     88101   1               7  105    118
       Date   PM year Site.Code
1: 19990702   NA 1999  001.0005
2: 19990705   NA 1999  001.0005
3: 19990708   NA 1999  001.0005
4: 19990711   NA 1999  001.0005
5: 19990714 11.8 1999  001.0005
6: 19990717 49.4 1999  001.0005
```

(f) (12) Find the intersection of the sites (i.e., monitors) in between 1999 and 2012 which gives us the list of monitors in New York that operated both in 1999 and 2012 using split() and intersect().

```
# [f] (12) Find the intersection of the sites  between 1999 and 2012

#這邊使用的是單純只有紐約的探測器
head(ny_data)#使用ny_data，看看我們資料樣子

ny_1999.siteId <- filter(ny_data, ny_data$year == 1999)
ny_2012.siteId <- filter(ny_data, ny_data$year == 2012)
monitors <- unique(intersect(ny_1999.siteId$Site.ID, ny_2012.siteId$Site.ID))
monitors
```

146:1 (Untitled)

Console Terminal Jobs

R 4.1.1 · ~/

```
6: 19990717 49.4 1999  001.0005
> ny_1999.siteId <- filter(ny_data, ny_data$year == 1999)
> ny_2012.siteId <- filter(ny_data, ny_data$year == 2012)
> monitors <- unique(intersect(ny_1999.siteId$Site.ID, ny_2012.siteId$Site.ID))
> monitors
 [1] "0005" "0012" "0080" "0011" "0002" "1007" "0003" "2008" "1015" "0055"
>
```

(g) We observe that the list contains 10 monitors. Rather than choosing a monitor at random, it would make more sense to choose one that had the most observations. (13) Write a block of code to identify the monitor in New York State that had the most data using mutate(), filter(), group_by(), summarize(), and arrange().

```
# A tibble: 19 x 2
   Site.Code data_count
   <chr>          <int>
 1 031.0003         198
 2 001.0005         186
 3 101.0003         183
 4 067.1015         153
 5 063.2008         152
 6 029.0005          94
 7 001.0012          92
 8 005.0080          92
 9 013.0011          92
10 047.0011          69
11 029.0002          61
12 059.0005          61
13 059.0011          61
14 093.0003          61
15 085.0055          38
16 055.1007          31
17 071.0002          31
```

(h) It seems that monitor ~~101.0003~~ 031.0003 had collected the most data in New York State during 1999 and 2012 ~~(n = 527)~~. (n=198)

(14) Subset the data (i.e., pm) that contains observations from the monitor we just identified (State.Code = 36 & County.Code = 101 & Site.ID = 0003) and assign the subset data to an obj. called 'pmsub'.

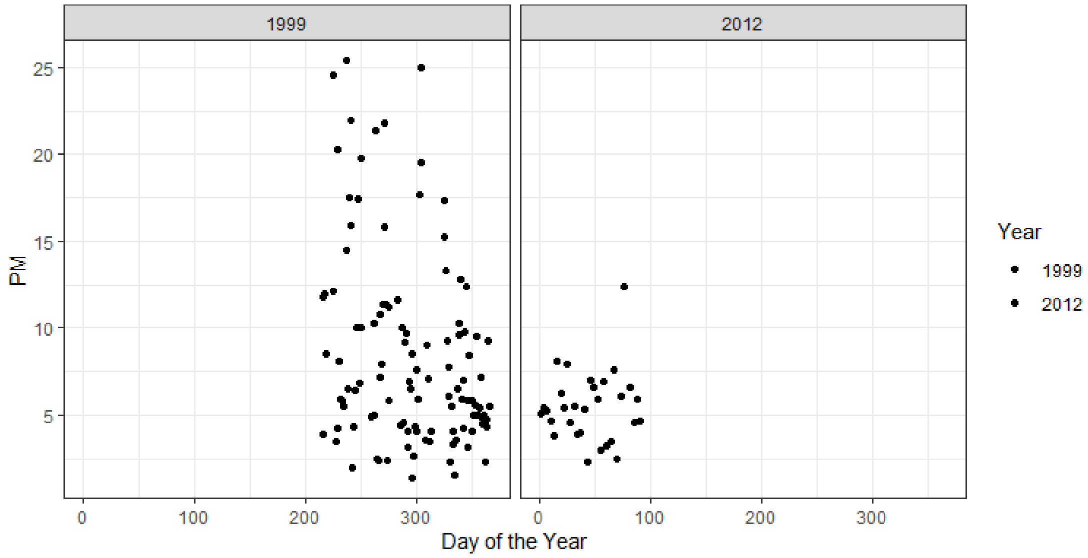
```
## h)
### (14) Subset the data that contains observations from the monitor
pmsub<- subset(pm, State.Code == '36'
                   & County.Code == '101'
                   & Site.ID == '0003')
pmsub

```

(i) Next, using the lubridate package, (15) convert the Date variable into a date obj. and then create a variable called 'yday' containing info. on day of the year using yday().

```
# (i) Next, using the lubridate package,
# (15) convert the Date variable into a date obj. and then create a variable
#called 'yday' containing info. on day of the year using yday().
pmsub$yday <- yday(ymd(pmsub$Date))
head(pmsub)
str(pmsub)
```

(j) Draw a scatter plot by mapping the year-day variable on the x-axis, PM2.5 level on the y-axis separately for 1999 and 2012. (16) Make sure to label the x-axis, (17) separate the plots using the facet function and (18) use the base white theme to replicate the graphics shown below.

```r
ggplot(pmsub, aes(yday, PM)) +
    geom_point() +
    facet_wrap(. ~ year) +
    labs(x = "Day of the Year",  y = 'PM') +
    theme_bw()
```



(k) Interesting pattern observed is that the variation (spread) in the PM values in 2012 is much smaller (vs. larger in aggregate) than it was in 1999. The plot shows that not only are the average levels of PM lower in 2012, but that there are fewer large spikes from day to day in 2012.