AS2: Exploring Data via Visualization

This assignment uses a national monitoring data which contains information on changes in fine particulate matter (PM) air pollution in the United States. This is a public data provided by the Environmental Protection Agencies and is freely available to the public. Using different tools and methods we covered in class, your goal for this assignment is to describe the changes in fine particle (PM2.5) outdoor air pollution in the United States between the years 1999 and 2012.

Your general hypothesis is that *'outdoor PM2.5 has decreased on average across the U.S. due to nationwide regulatory requirements arising from the Clean Air Act.'* To investigate this hypothesis, you will be looking at the PM2.5 data from the U.S. Environmental Protection Agency which is collected from monitors sited across the U.S.

** 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.*

INSTRUCTIONS

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."

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

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().

   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.

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

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'.

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

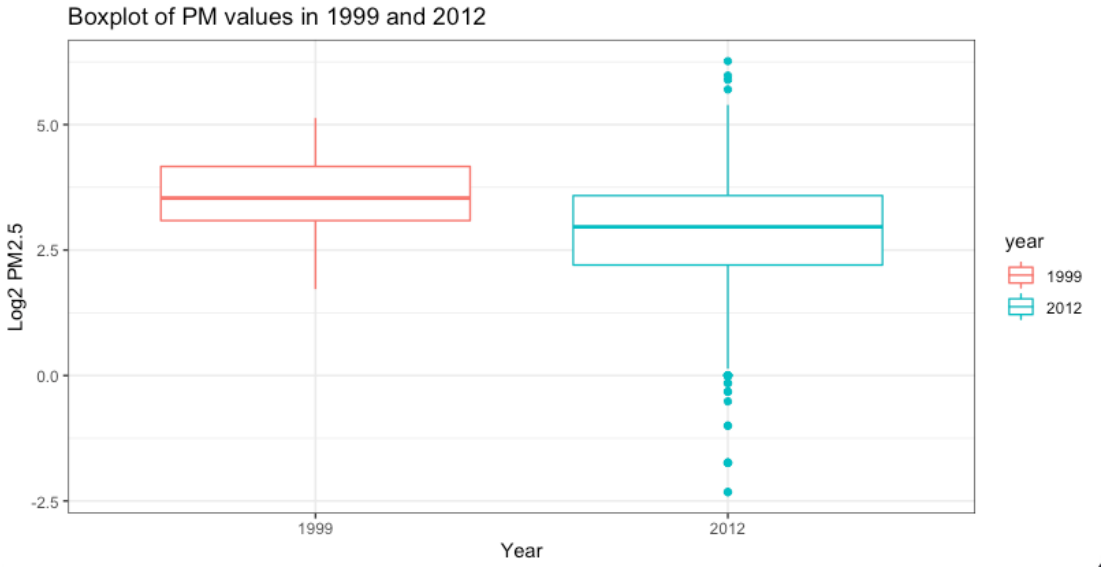
2) Data Exploration with Visualization using *ggplot2*

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.

b) Then, create boxplots of all monitor values in 1999 and 2012 using the randomly sampled data as shown below. (6) Make sure to take the log of the PM values (with base 2; i.e., binary algorithm) to adjust for the skewness in the data, (7) label the title, x-axis & y-axis, and (8) use the base white theme to replicate the graphics.



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

Changes in PM levels at an individual monitor:

So far, we have looked at the change in PM levels on average across the U.S. One potential issue with this approach is that the monitoring network could have changed between 1999 and 2012. That is, if, for some reason, in 2012 there were more monitors concentrated in cleaner areas of the country than there were in 1999, it might appear the PM levels decreased when they did not. We will now focus on a single monitor in New York State to observe/visualize the changes to account for this possibility.

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().

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.

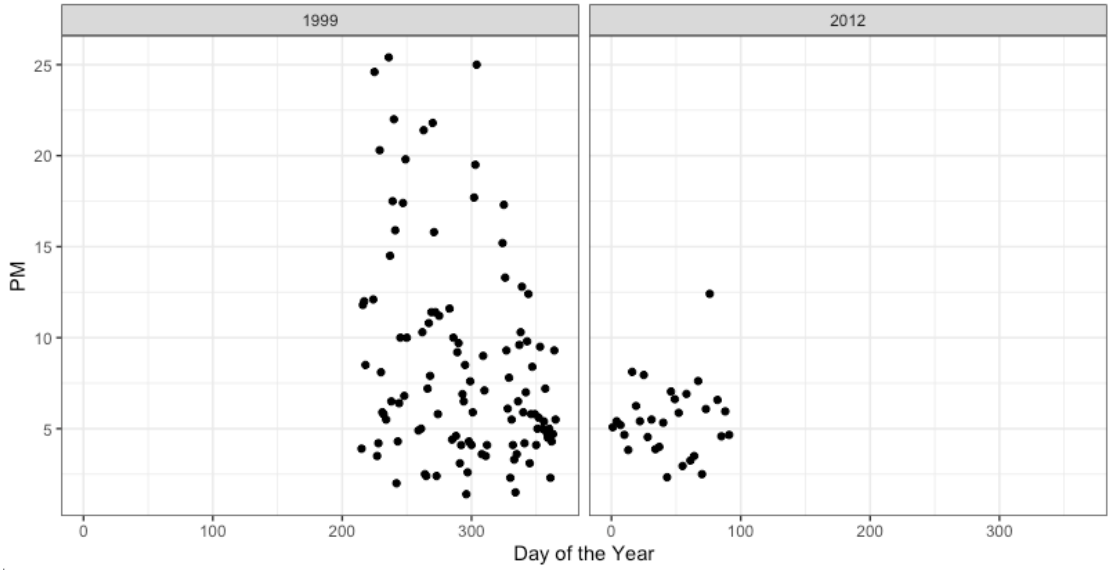
f) (12) Find the intersection of the sites (i.e., monitors) in between 1999 and 2012 which gives us the list of monitors in New York that operated both in 1999 and 2012 using split() and intersect().

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 the original data (i.e., pm) that had the most data using mutate(), filter(), group_by(), summarize(), and arrange().

h) It seems that monitor 101.0003 had collected the most data in the U.S. (i.e., pm) during 1999 and 2012 (n = 527). (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'.

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().

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.



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.