Answers to AS2 – PBA

Daniele Melotti (ID: 110077432)

1) The dimensions are found using the function dim(), and for the 1999 data it shows a total of 117421 rows and 12 variables.

2) The first 3 rows of 1999 data can be printed out by using head(d1, 3), where d1 is the name of the data table containing the 1999 data.

3) Using the function summary() on the 1999 data, we obtain the following summary statistics:

Min	1st Quartile	Median	Mean	3rd Quartile	Max	NAs
0.00	7.20	11.50	13.74	17.90	157.10	13217

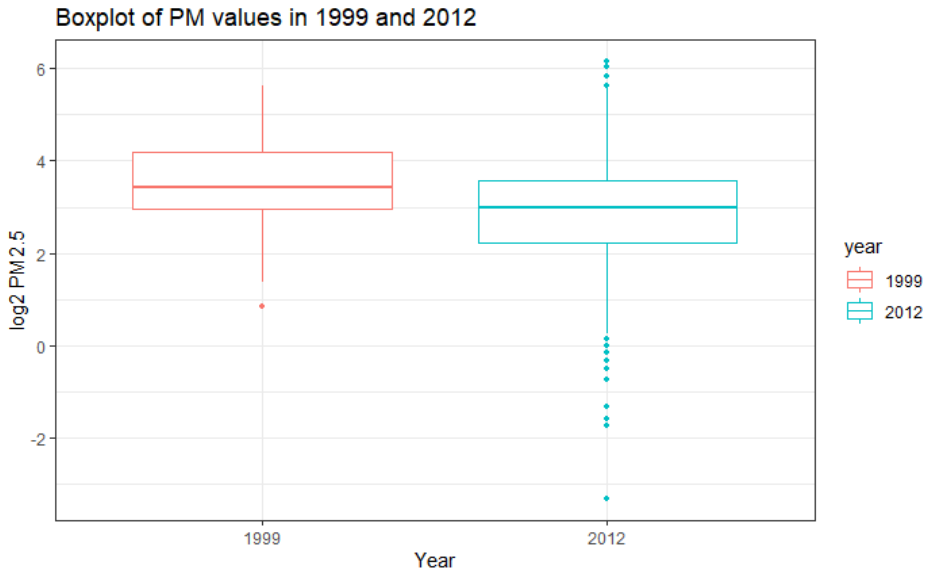
4) The percentage of the PM2.5 observations that are missing is equal to 11.256%.

5) I set seed using the function set.seed(2021), and extracted 1,000 randomly selected samples using sample_n(pm, 1000) from dplyr. Then I assigned the sampled data to a new data table called “sample”.

6) I calculated the log of the PM values using the function log(sample$PM, 2) and assigned it to the data table “sample”.

7) I added labels by using labs in the boxplot instructions in the following way:
labs(title = “Boxplot of PM values in 1999 and 2012”,
x = “Year”,
y = “log2 PM2.5”)

8) I used the base white theme by adding theme_bw() at the end of the instructions.
Here’s the final plot:



9) If the distribution was normal, we would be able to say that the median shown on the boxplots is equal to the mean (and the mode as well). However, the PM data is probably not normally distributed, so we may only discuss about the median. If we look at the median value of the boxplots, we can see that PM air pollution in the US was a little higher in 1999 than in 2012, on average. The difference is more visible when we compare the two upper quartiles, as we see that the one for the 1999 data is well above the one for 2012 data. The same is true in the case of the lower quartile.
However, we can also see that there are way more outliers for the 2012 data; moreover, the whiskers are a little longer when compared to those of the 1999 data. These features can be taken as a sign that the variance is higher in the case of 2012 data.

10) I used the following code to obtain only the observations from New York and keep only the County.Code and Site.ID:

```
ny <- pm %>%
        filter(State.Code == 36) %>%
        select(County.Code, Site.ID, year) %>%
        unique()
```

Actually, I also included the year variable simply because it will be useful later in Question 12.

11) I created a new variable Site.Code in the ny regional data with the following line of code:
ny$Site.Code <- paste(ny$County.Code, ny$Site.ID, sep = ".")

12) In order to get the intersection of the sites between 1999 and 2012, I used the following code:

```
monitor <- split(ny$Site.Code, ny$year)
inter <- intersect(monitor$"1999", monitor$"2012")
inter
```

I could split the Site.Code variable thanks to the inclusion of the year variable in the select() within Question 10.
The code returns a vector containing 10 site codes which are present both in 1999 and 2012. I put them in the following table:

Site.Code
001.0005
001.0012
005.0080
013.0011
029.0005
031.0003
063.2008
067.1015
085.0055
101.0003

13) In order to identify the monitor in the NY state that had the most data using the required functions, I wrote the following code:

```
pm %>%
  mutate(Site.Code = paste(County.Code, Site.ID, sep = "." )) %>%
  filter(Site.Code %in% inter) %>%
  group_by(Site.Code) %>%
  summarize(n = n()) %>%
  arrange(desc(n))arrange(desc(n))
```

The code above outputs the following table:

Site.Code	n
101.0003	527
013.0011	213
031.0003	198
001.0005	186
067.1015	153
063.2008	152
029.0005	94
001.0012	92
005.0080	92
085.0055	38

We can see that the monitor with the most observations is 101.0003, with 527 observations.

14) I subset the data according to the given instructions and assigned it to the new object pmsub as follows:
pmsub <- subset(pm, State.Code == "36" & County.Code == "101" & Site.ID == "0003")

15) I converted the Date variable into a date object using lubridate in the following way:
pmsub$Date <- as.character(pmsub$Date)
pmsub$Date <- as_date(pmsub$Date, format = "%Y%m%d")

Then, I created a new variable yday containing info on day of the year in the following way:
pmsub$yday <- yday(pmsub$Date)

16) & 17) & 18) I drew the required scatterplot by writing the following block of code:

```
sct <- ggplot(pmsub, aes(x = yday, y = PM))

sct +
 geom_point() +
 labs(x = "Day of the Year", y = "PM") +
 facet_wrap(~ year, ncol = 2) +
 theme_bw()
```

The result is the same as on the assignment instructions, as can be seen below: