Assignment 1 - Data Analysis and Statistical Methods

Frans Simanjuntak - S3038971

September 13, 2017

1 Question 1

(A) Below is the yearly average concentration for all pollutants:
**See appendix A1.1 for the R-code*

Year	avg_NH3	avg_CO	avg_O3	avg_NO	avg_NH2
2014	9.481653	213.6479	44.1087	2.827684	16.18873
2015	4.370000	344.7900	32.5800	0.520000	17.95000

(B) Below is the average of NO concentration for Sundays and Thursday:
**See appendix A1.2 for the R-code*

Weekday	avg_NO
1	1.664117
5	3.908968

From the above results we can see that the average value of Nitrogen Oxide (NO) on Thursday is higher than on Sunday. Usually Nitrogen oxides are produced by motor vehicle exhaust, the burning of coal, oil, diesel fuel, and natural gas and by industrial processes such as welding, electroplating, engraving, and dynamite blasting [NO]. However, motor vehicle such as car, trucks, or any other vehicle are most likely to be the combustion source in Netherlands. If we see the average of Nitrogen Oxide released into the air on Thursday, it was almost twice than on Sunday, *meaning that people in Netherlands tend to use the motor vehicle such as car, trucks, etc on Thursday rather than Sunday.*

(C) Below is the average concentration of ammonia (NH3) and Carbon monoxide (CO) in summer (months:6,7,8) and winter (months: 12,1,2) respectively:
**See appendix A1.3 for the R-code*

season	avg_NH3	avg_CO
summer	7.808597	163.7137
winter	3.966393	234.3097

NH3 is commonly known as Ammonia and it is commonly used and needed in fertilizers. From the above result, it can be seen that the average number of ammonia used in summer is twice than in winter. This phenomenon does make sense because people in Netherlands tend to do farming or gardening or livestock raising during summer when the weather is nice. As a result, they used more ammonia on this occasion as a nutrition for their plants or a source of protein in livestock feeds for ruminating animals such as cattle, sheep and goats. Conversely, people used less ammonia in winter because farming or gardening or livestock raising are almost impossible to be done. Beside, some of the plants die or might not grow properly in winter.

CO stands for Carbon Monoxide. It is colorless gas or liquid, practically odorless and burns with a violet flame. The source of CO are usually from the unvented kerosene and gas space

heaters, leaking chimneys and furnaces, back-drafting from furnaces, gas water heaters, wood stoves, fireplaces, gas stoves, generators and other gasoline powered equipment, automobile exhaust from attached garages, and tobacco smoke [CO]. From the above results, we can see that the average number of CO in the winter is 70.596 higher than in summer. The reason behind this trend is most likely because in the winter the average temperature of the weather is very low (normally 2° - 6° C) while the temperature in summer is around 17°-20°C. That said, people tend to stay at home and use heaters or other heating equipment in winter in order to get warm or to cook something which leads to the increasing amount of CO released into the air. Conversely, this is not very likely to happen in summer because people usually spend their time outside enjoying the sun and almost never use heaters or rarely use heating equipment which makes the amount of Carbon Monoxide released into the air is not as much as in winter.

2 Question 2

(A) The histogram of O3 and NO
**See appendix A2.1 for the R-code*

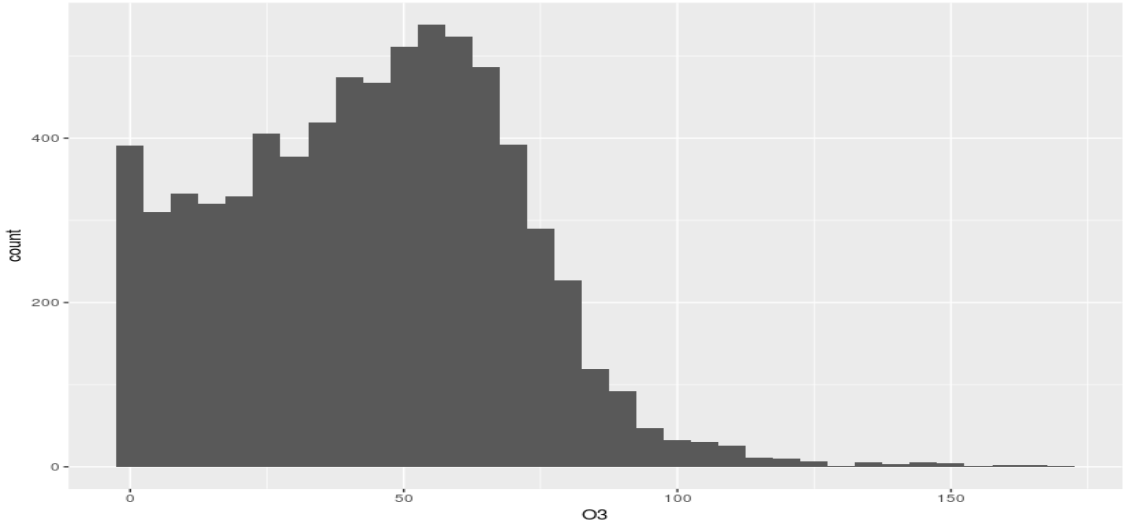

Figure 1: Histogram of O3.

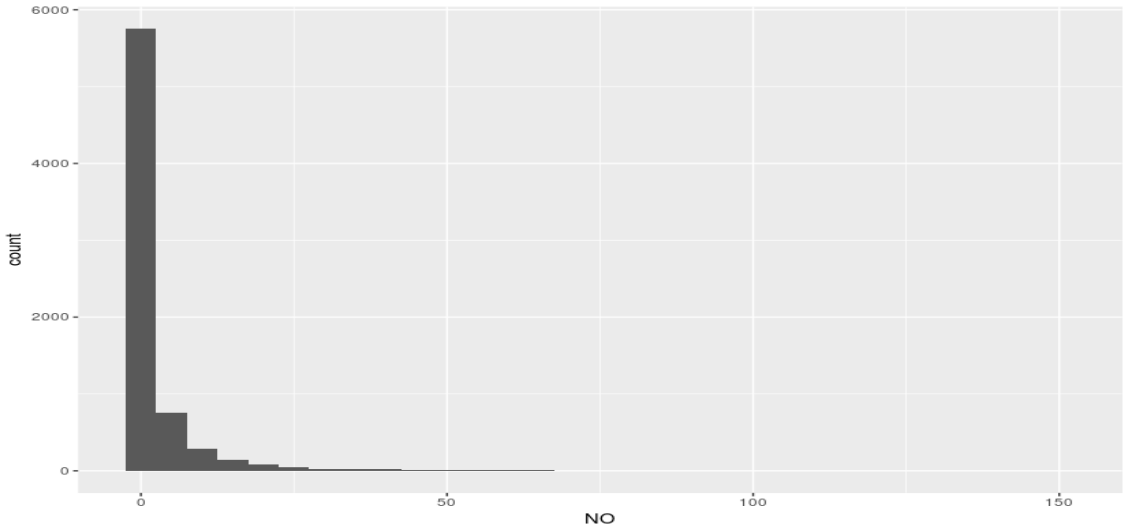

Figure 2: Histogram of NO.

The above histogram depicts that the majority value of Ozone (O3) are in between range of 0 and 85 in which the total count of each value is above 200. The value between 60 - 65 was being the highest one. The minimum value of O3 is 0 and the maximum value is 168 which being the major peak.In comparison to NO, the O3 data forms symmetric, unimodal, or normal distribution. The left and right hand sides of the distribution are roughly equally balanced around the mean.

On the other hand, the NO data forms a right-skewed distribution. This right side of NO histogram contains the larger half of the observations in the data, extends a greater distance than the left side. The majority of NO values are in between 0 and 3 thats almost 6000 data points. The minimum value of NO is 0 and the maximum value is 148.

(B) Below are the mean and the standard deviation of NO.
**See appendix A2.2 for the R-code*

mean_of_NO	standard_deviation_of_NO
2.827363	7.710068

It is true that the standard deviation of NO is much higher than its mean. It happens because the data points of NO form right-skewed distribution rather than normal distribution. This right side of NO histogram contains the larger half of the observations in the data, extends a greater distance than the left side. The tail of the distribution is longer on the left hand side than on the right hand side. Most of data points stay in range 0 to 3 that induces the mean to 2.827363 and cause a slightly higher standard deviation which is almost three times higher than the mean value.

3 Question 3

(A) Below is the box plot of O3 per hour
**See appendix A3.1 for the R-code*

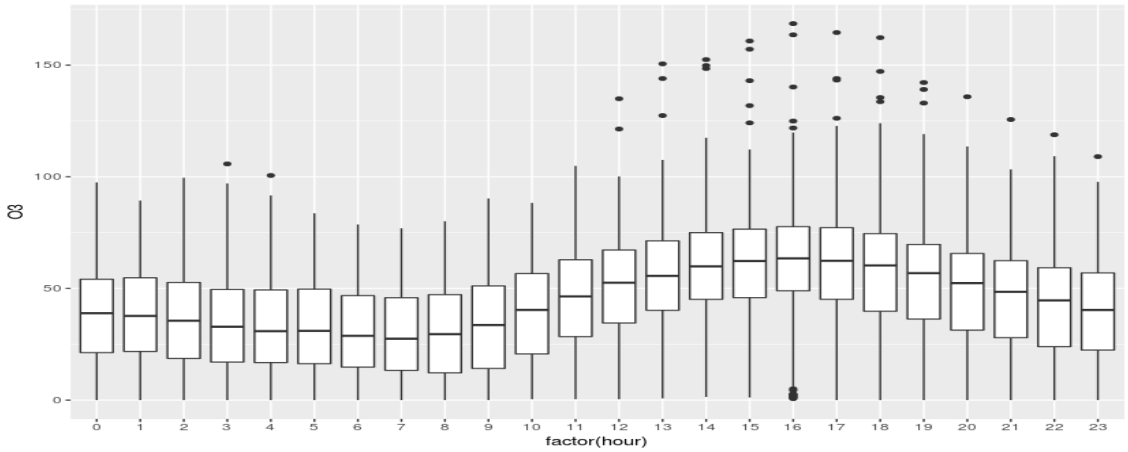

Figure 3: Boxplot of O3.

From the above boxplot, we can obtain some information such as the mean value, 1^{st} quartile and 3^{rd} high quartile of O3 per hour. This plot depicts that the increasing number of ozone happens in the afternoon around 13 to 19. In this occasion, the average value is above 50. The reason why the amount of O3 is increasing in the afternoon is most likely caused by **Photochemical reaction**. This reaction refers to any chemical reaction which occurs as a result of light energy from the sun. As a result of its photochemical origin, O3 displays strong seasonal and diurnal patterns, with higher concentrations in the afternoon [APO].

We can also see in the afternoon some data lie very high above the mean value. However, those data points can be considered as outliers which might be appearing on certain occasions only.

(B) Below are the the individual histograms of O3 for every month which are based on hour and weekday.
**See appendix A3.2 for the R-code*

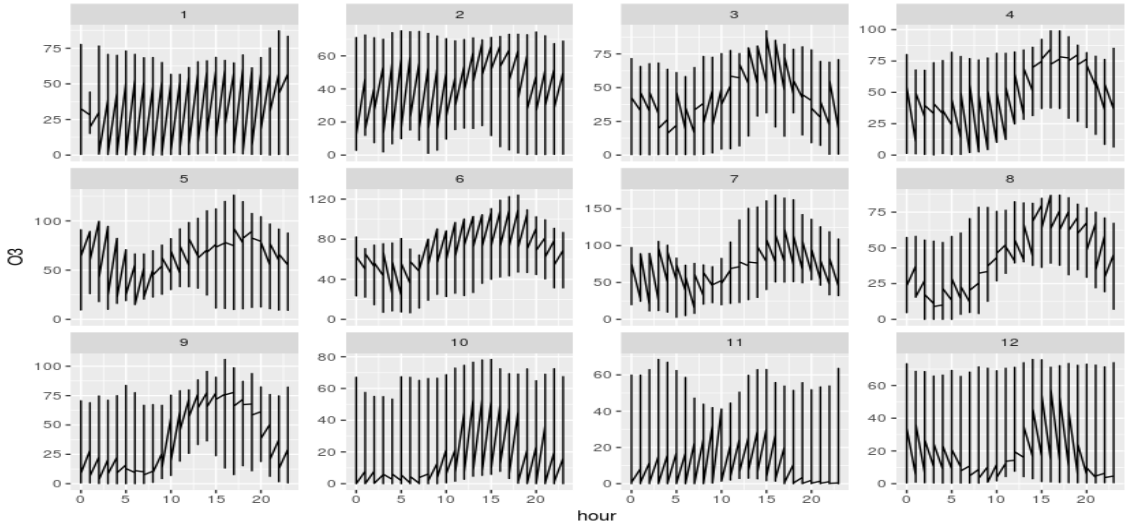

Figure 4: Individual histogram of O3 - hourly based.

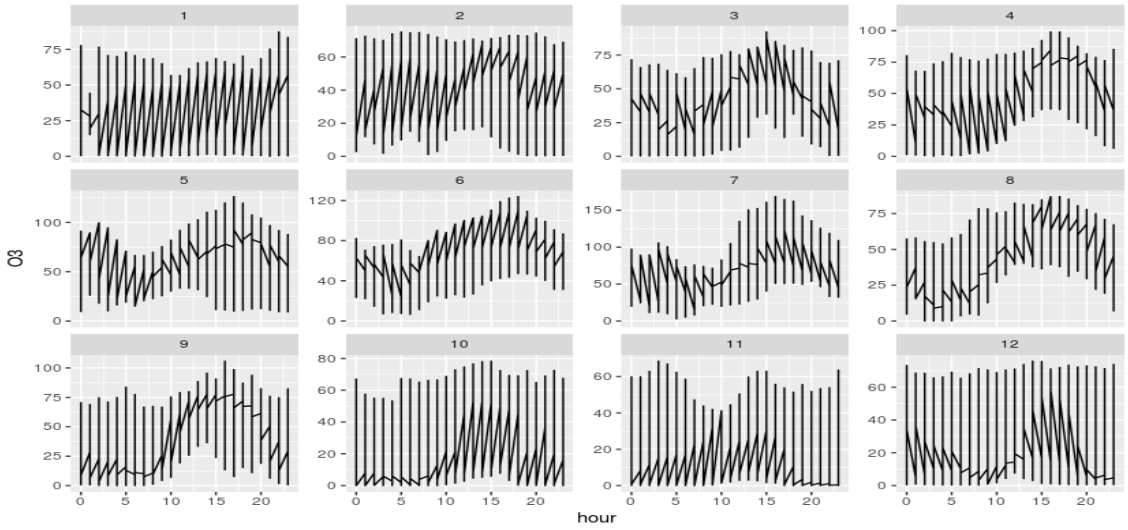

Figure 5: Individual histogram of O3 - weekly based.

From the above figures, we can actually see that the number of O3 is higher during summer (months: 6,7, 8) than other seasons. Conversely, this number decreases in winter(months: 12, 1, 2). The average number of O3 per hour during the summer is around 50 to 155 and in winter, the range value of O3 is in between 0 and 75. Again, the increasing number of O3 in summer is most likely caused by **Photochemical reaction**. As mentioned earlier, this reaction refers to any chemical reaction which occurs as a result of light energy from the sun. As a result of its photochemical origin, O3 displays strong seasonal and diurnal patterns, with higher concentrations in summer.

4 Question 4

(A) The scatter plot of O3 vs NO2 for the month of July colored by hour
**See appendix A4.1 for the R-code*

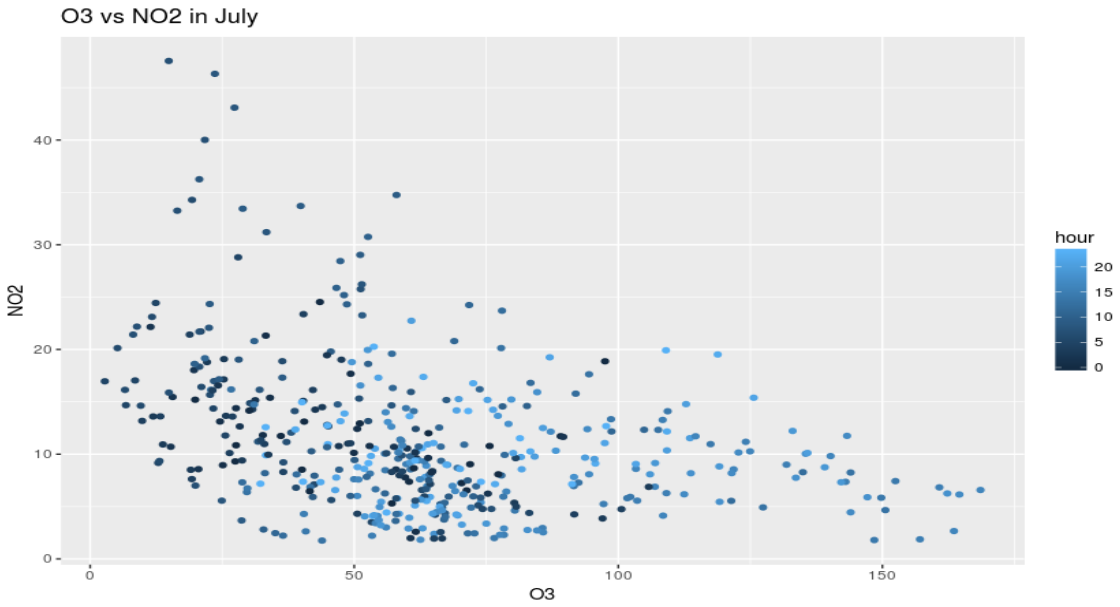

Figure 6: The scatter plot of O3 vs NO2 for the month of July.

(B) The scatter plot of O3 vs NO2 for January and May.
**See appendix A4.2 for the R-code*

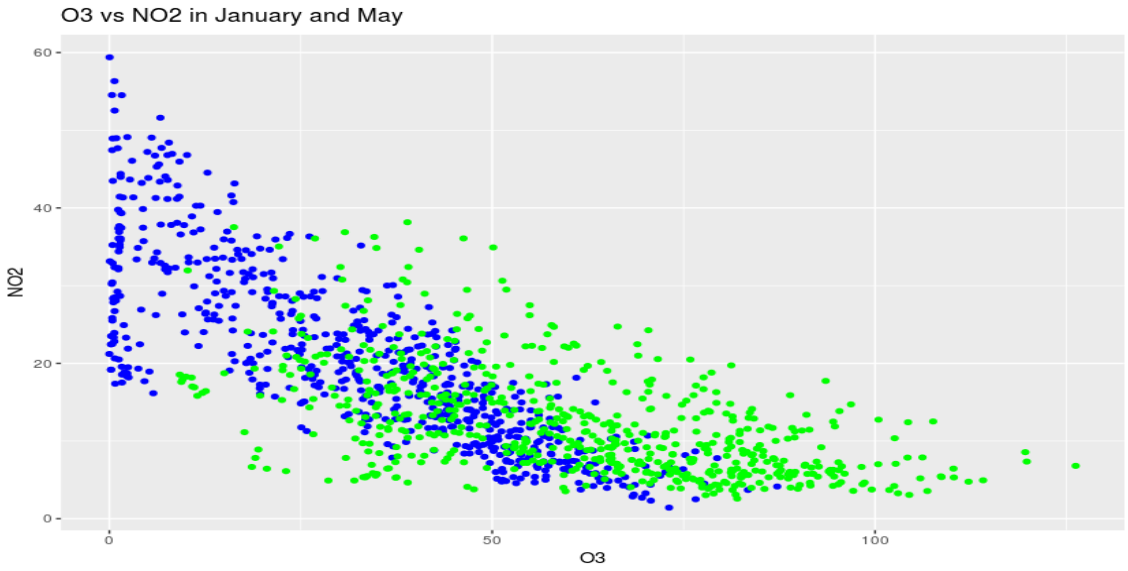

Figure 7: The scatter plot of O3 vs NO2 for January and May.

Appendices

Note:
This preliminary code below must be executed first before executing the codes on each appendix.
Preliminary Code

```
library("dplyr")
library("ggplot2")
setwd("~/Assignments/Assignment1")
df <- read.csv("data1(2).csv", header = TRUE)
df[, "date"] = lapply(df["date"],function(x){strptime(x, "%d/%m/%y")})
df$date <- as.POSIXct(df$date)
```

A1. Question 1

(A) A1.1

```
df$year <- format(df$date,"%Y")

df %>% select(year,NH3,CO,O3,NO,NO2) %>% group_by(year)%>%
summarise(avg_NH3 = mean(NH3), avg_CO = mean(CO), avg_O3 = mean(O3),
avg_NO = mean(NO), avg_NH2 = mean(NO2))
```

(B) A1.2

```
df %>% select(weekday, NO) %>% filter(weekday == 1  | weekday == 5) %>%
group_by(weekday) %>% summarise(avg_NO= mean(NO))
```

(C) A1.3

```
df$season <- ifelse(df$month %in% c(1,2,12),"winter",
ifelse(df$month %in% 3:5, "spring",
ifelse(df$month %in% 6:8, "summer", "autumn")))

df %>% select(season, NH3, CO) %>%
filter(season == "summer"  | season == "winter") %>%
group_by(season) %>%
summarise(avg_NH3= mean(NH3), avg_CO= mean(CO))
```

A2. Question 2

(A) A2.1

```
ggplot(df, aes(x=O3)) + geom_histogram(binwidth=5)
ggplot(df, aes(x=NO)) + geom_histogram(binwidth=5)
```

(B) A2.2

```
df %>% select(NO) %>% summarise(avg_NO = mean(NO), sd_NO = sd(NO))
```

A3. Question 3

(A) A3.1

```
ggplot(df, aes(x=factor(hour), y=O3)) + geom_boxplot()
```

(B) A3.2

```
ggplot(df, aes(x=weekday, y=O3, ymin=0)) + geom_line() +
facet_wrap(~month, scales="free_y")

or

ggplot(df, aes(x=hour, y=O3, ymin=0)) + geom_line() +
facet_wrap(~month, scales="free_y")
```

A4. Question 4

(A) A4.1

```
ggplot(df%>% filter(month == 7), aes(x=O3, y=NO2, colour=hour)) +
geom_point()  + ggtitle("O3 vs NO2 in July")
```

(B) A4.2

```
ggplot() + geom_point(data = df %>%
filter(month == 1), aes(x=O3, y=NO2), colour="blue") +
geom_point(data = df %>%
filter(month == 5), aes(x=O3, y=NO2), colour="green") +
ggtitle("O3 vs NO2 in January and May")
```

References

[APO] Air pollution ozone. https://www.greenfacts.org/en/ozone-o3/l-3/1-presentation.htm. Accessed: 2017-09-11.

[CO] Source of indoor air pollution. http://www.carbonmonoxidekills.com/are-you-at-risk/sources-of-indoor-air-pollution/. Accessed: 2017-09-11.

[NO] Nitrogen oxides. https://toxtown.nlm.nih.gov/text_version/chemicals.php?id=19. Accessed: 2017-09-11.