DATA MANIPULATION (IFN509)

PROJECT

A ASSESSMENT SUBMITTED TO

THE SCIENCE AND ENGINEERING FACULTY

OF QUEENSLAND UNIVERSITY OF TECHNOLOGY

Group Member (ID): Yuchen Jiang (N9573950) (Submitter)
Chaoran Li (N10298827)
Jinning Guo (N9858598)
Tutor: Philip Eichinski

Due Date: Sunday, 2/06/2019, 11.59pm

1. Explain the choice of technology

In this report, our group only use the R studio as the compile tool to analyse the factors which can influence the air quality. In order to better analyse the correlations between different indicators, firstly, we used the R studio to pre-processing the data. In pre-processing stage, we used some codes to identify the outliers, smooth the noisy data and fill the missing values. And then, the regression plot, decision tree diagram and clustering diagram can help us to analyse the relationship between different indicators. Comparing the R with other tools for analysing data, R language has powerful package ecosystem and chart advantage. The "dplyr" and "ggplot2" software packages are respectively used for data processing and plotting, and the charts are intuitive and easy to analyse. In addition, the R language is extensible and has a wealth of functional options to help developers to build their own tools and methods for data analysis. However, correct using of package is the most important part of using R language. It has to fully understand the manual and confirm the validity of the function before using the package. Therefore, it will waste much time if we cannot understand or use the wrong package. In addition, the ability to process large amounts of data is also the limitation of R studio. But in this report, the data size after data cleaning and integration is not large, so that we choose the R studio to analyse the data. In future data processing tasks, it is necessary to select the most suitable analysis software and language based on the data size and the limitations of different analysis tools.

2. Data Summary

In this section, we will use some tables to illustrate the steps that we use to analyse the data quality, data cleaning, data preparation and data integration. In the tables, it will also provide the justification and the results of each step.

2.1. Data quality analysis

Step	Code	Justifications
1	Brisbane <- read.csv(file = "southbrisbane-aq-2018.csv", header = TRUE)	Import the original data to analyse the quality
2	BNEWeather<- read.csv(file = "BNEWeather.csv", header = TRUE, sep = ",")	Choose the requires data manually (Location: Brisbane, Date: 1/01/2018 – 31/12/2018

3	summary(Brisbane)
	summary(BNEWeather)
		Use the summary function to check the outliers, miss value (NA) and noisy data

4	install.packages("ggplot2")
	plot(Brisbane$PM10..ug.m.3.)
	boxplot(Brisbane$PM10..ug.m.3.)
		Use the plot and boxplot function to find the outliers of each column data, so as to analyse the quality of data.

2.2. Data Cleaning

Step	Code	Justifications

1	for (i in 1 : ncol(BNEWeather)) {
	  BNEWeather[is.na(BNEWeather[,i]), i] <- mean(BNEWeather[,i], na.rm = TRUE)
	}
	
	for (i in 1 : ncol(Brisbane)) {
		Through the step 3 in the last table, we found there are many miss values in these two data documents. In addition, we found the majority missing value are numerical type, such as rainfall, wind speed, temperature etc. Replacing these missing values with averages is not only

	Brisbane[is.na(Brisbane[,i]), i] <- mean(Brisbane[,i], na.rm = TRUE) }	improves data integrity, but also can ensure the stability of the data. So that we fill the missing values of the numeric types with the average of the corresponding column values.

2	summary(Brisbane)
	summary(BNEWeather)
		Double check to ensure that there are no missing values of the numeric types in the corresponding column

3	Brisbane[Brisbane < 0] <- 0 summary(Brisbane)	In the data set named Brisbane, we found outliers (negative numbers) in the pm2.5 and pm10 columns. However, the normal index of pm2.5 and pm10 cannot be negative. We decided that the index in the present time period should be the healthiest index (0). So we replace the negative numbers in pm2.5 and pm10 to 0
4	write.csv(BNEWeather, file = "BNEWeather.csv")	Export the cleaning data for data integration

2.3. Data Integration

Step	Code	Justification
1	tmp <- aggregate(Brisbane[,3:14], by = list(Brisbane$Date), FUN = mean)	Since the data is recorded hourly in this data list, so that we calculated the average of the data every 24 hours for better integration
2	summary(tmp)	Check the data to ensure there are 365 averages in that data list
3	write.csv(tmp, file = "BrisbaneMean.csv")	Export the list of transformed data to integrate with another data list (BNEWeather). In order to create a unique column name, we change the label (Group.1 to Date) manually

4	bm <- read.csv(file = "BrisbaneMean.csv", header = TRUE)
	MERGE <- merge(BNEWeather, bm, by = "Date")
		Import the transformed data and merge the two data lists by Date

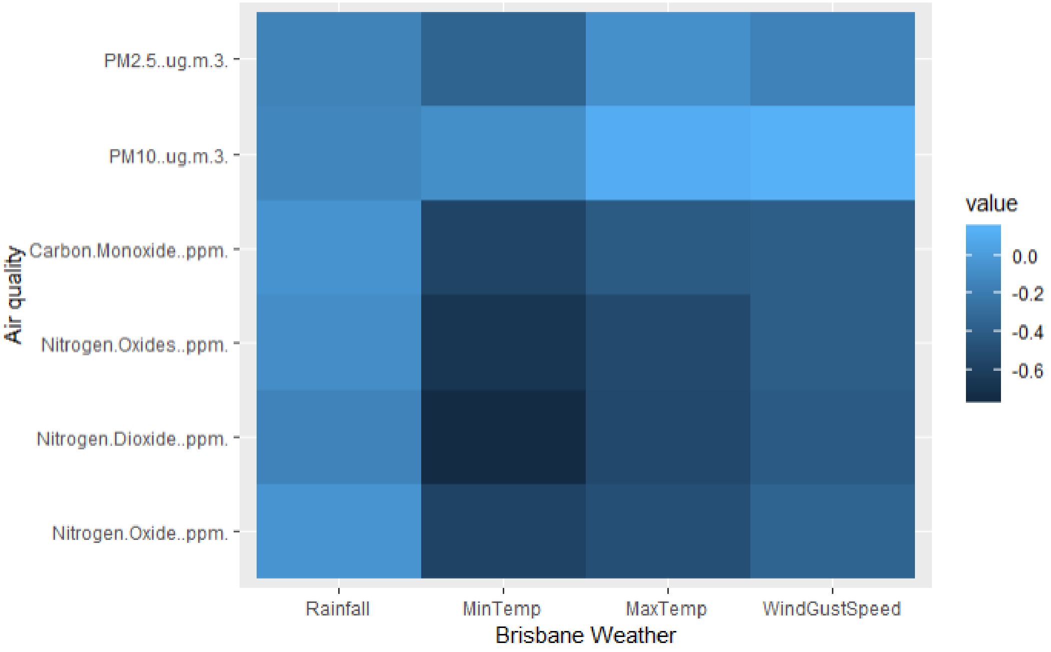
5	summary(MERGE)	Check the data to ensure there are 365 objectives in the merge data list.
6	write.csv(MERGE, file = "Merge.csv")	Export the merge data to draw the correlations diagram, decision tree diagrams and clusters

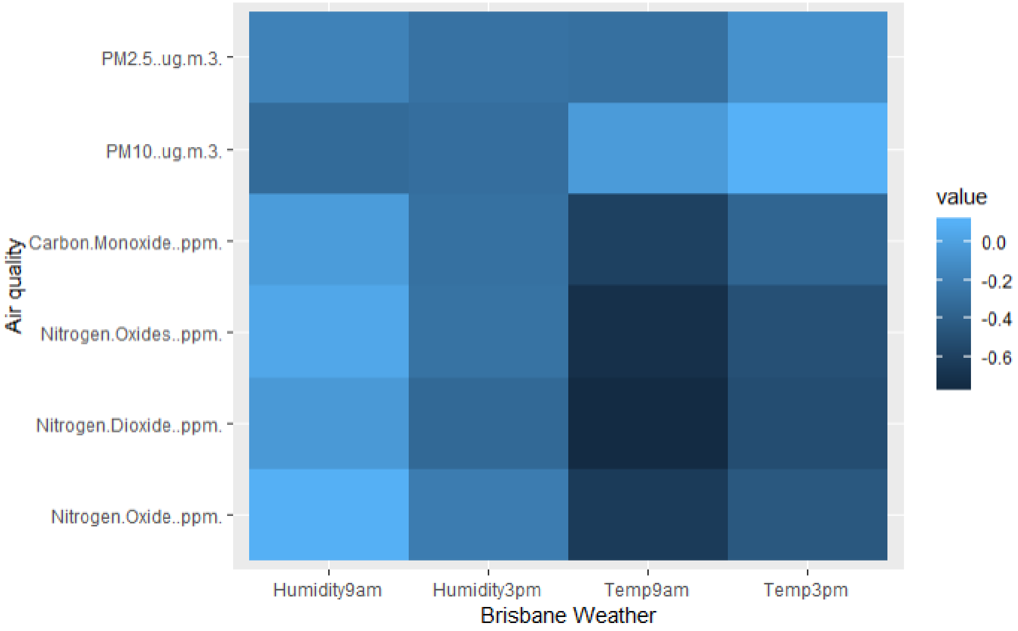
3. Correlations and Visualisation

In order to better analyse the relationship between weather and air quality, we investigated kinds of literature. After investigation, we found that the change of meteorological factors such as temperature, humidity and air pressure will influence the air quality. In addition, the temperature, wind, precipitation and other meteorological conditions will directly affect air quality without significant changes in pollutant discharge (Rosas, Mccartney & Payne, 1998). Meanwhile, we also found that the liveable index, nitrogen and oxygen compound content, pm10 and pm2.5 index all directly reflect the quality of air (McGregor & Bamzelis, 1995).

Therefore, we install the “ggplot2” and “reshape2” to visual the correlations between weather factors ("Rainfall", "MinTemp", "MaxTemp", "WindGustSpeed", "Humidity9am", "Humidity3pm", "Temp9am", "Temp3pm") and air quality factors ("Nitrogen.Oxide..ppm.", "Nitrogen.Dioxide..ppm.", "Nitrogen.Oxides..ppm.", "Carbon.Monoxide..ppm.", "PM10..ug.m.3.", "PM2.5..ug.m.3.").

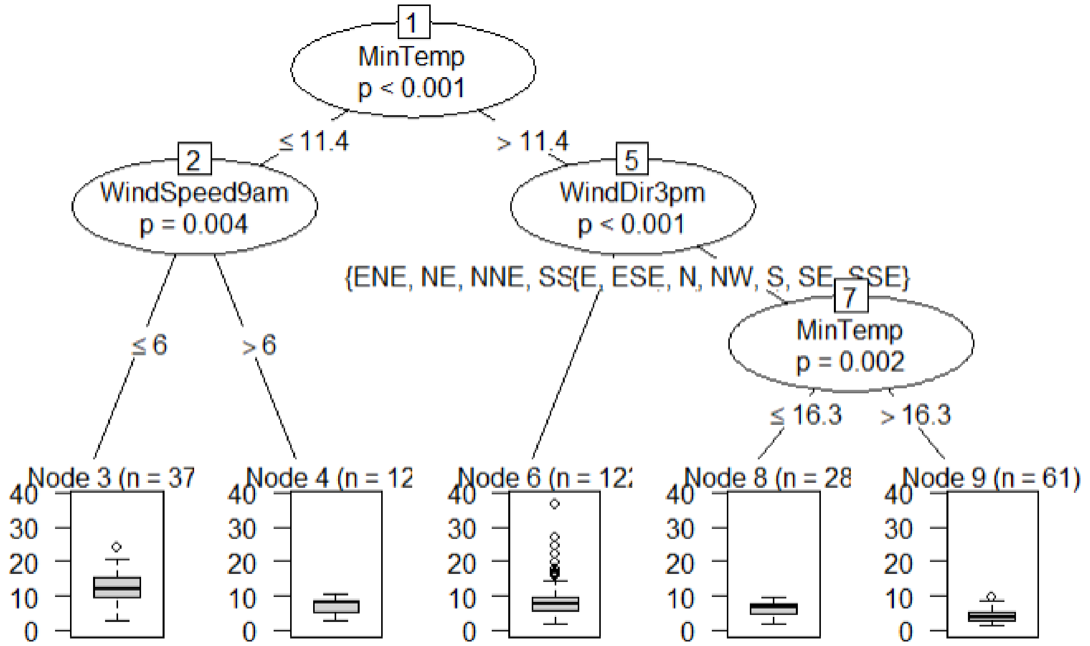
```
                         Rainfall     MinTemp     MaxTemp WindGustSpeed Humidity9am
Humidity3pm
Nitrogen.Oxide..ppm.   -0.03863601 -0.57245993 -0.48946843    -0.3459796  0.09552821
-0.2125021
Nitrogen.Dioxide..ppm. -0.14402766 -0.75243320 -0.53199781    -0.4050962 -0.04041587
-0.3343177
Nitrogen.Oxides..ppm.  -0.08369563 -0.67152525 -0.52811304    -0.3854384  0.04407313
-0.2718489
Carbon.Monoxide..ppm.  -0.04542993 -0.56574043 -0.40975882    -0.3872390 -0.02156607
-0.2810116
PM10..ug.m.3.          -0.12234577 -0.07078799  0.11425396     0.1384955 -0.31865819
-0.2976467
PM2.5..ug.m.3.         -0.14512183 -0.34570701 -0.06940662    -0.1492320 -0.17695888
-0.2729766
                          Temp9am    Temp3pm
Nitrogen.Oxide..ppm.   -0.63763633 -0.4437612
Nitrogen.Dioxide..ppm. -0.75232521 -0.5120654
Nitrogen.Oxides..ppm.  -0.71256140 -0.4910117
Carbon.Monoxide..ppm.  -0.59397287 -0.3584670
PM10..ug.m.3.          -0.02845511  0.1031010
PM2.5..ug.m.3.         -0.28655795 -0.0859505
```





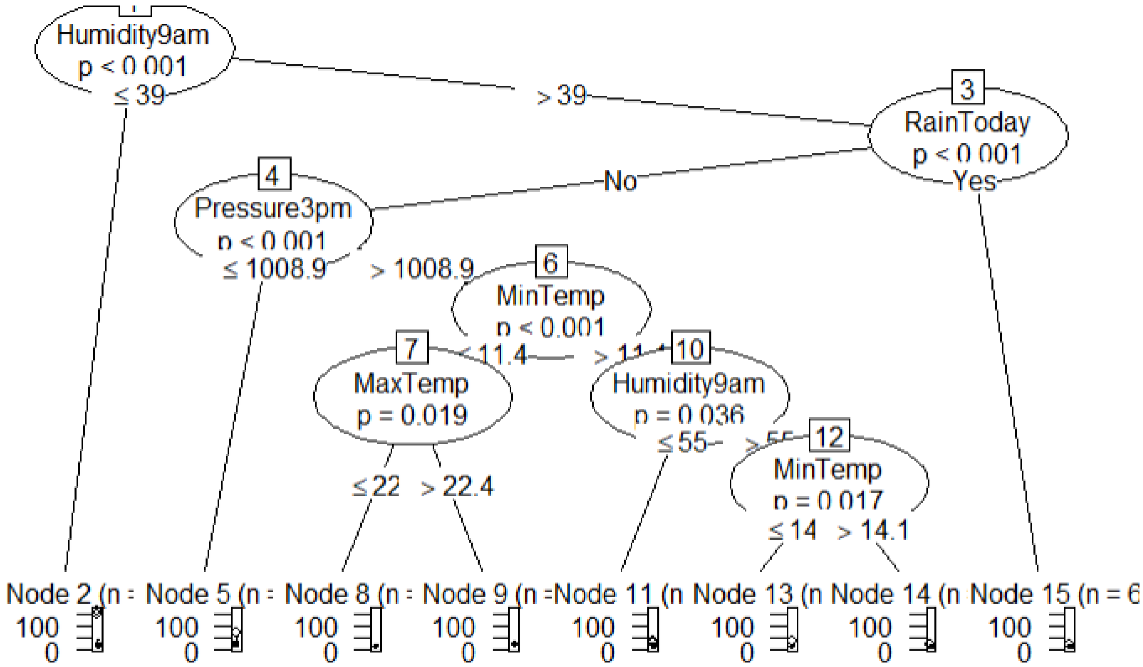
The correlation diagram shows the most weather indicators are negatively correlated with the air quality indicators, which means the higher value in weather indicators has a lower value in air quality indicators. In addition, in the visual diagram, the deeper colour represents the stronger negative correlations. Because all the air quality indicators are harmful, a lower index means the better air quality. For example, the correlation between MinTemp and air quality indicators are -0.57, -0.75, -0.67, -0.56, -0.07 and -0.34, which means when the MinTemp in Brisbane comes to higher, Brisbane will have the better air quality. Although there is a positive correlation coefficient in these diagrams, it can be ignored based on the value of the coefficient (too small) and the approach used in data cleaning.

4. Decision Tree



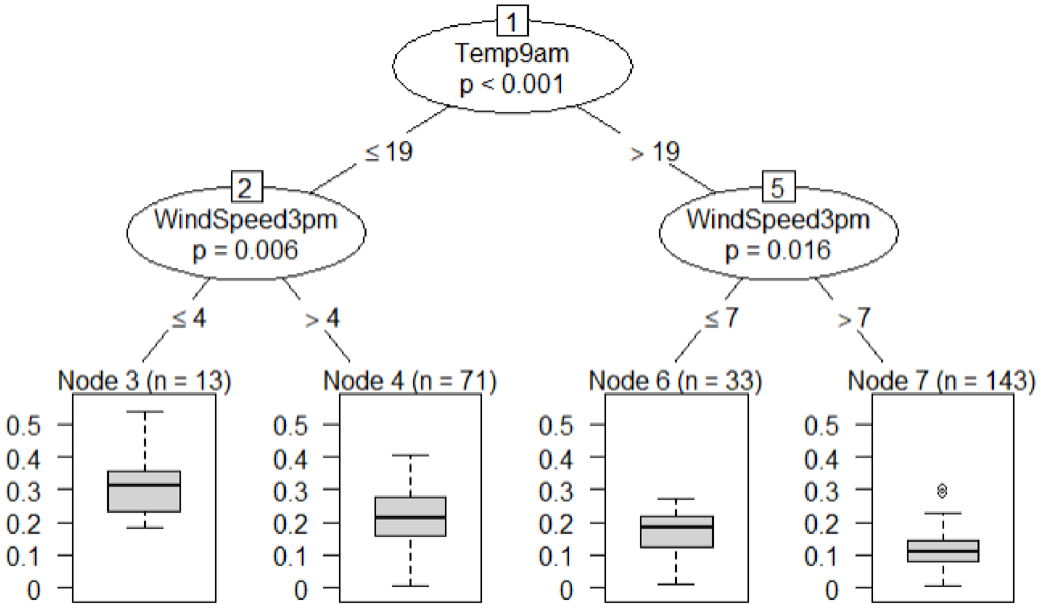
Pm2.5

Firstly, we made the indicator of Pm2.5 as the target data and found the wind speed will influence the index of Pm2.5 in two sides. The diagram shows the index of Pm2.5 is much lower when the wind speed above 6m/s. In addition, the index of Pm2.5 will become higher when the wind is blowing from West and the index of Pm2.5 will become lower when the wind is blowing from East. According to the Google Map, we found the west of Brisbane is inland and the west is ocean. Therefore, the wind from ocean are cleaner than wind from inland.



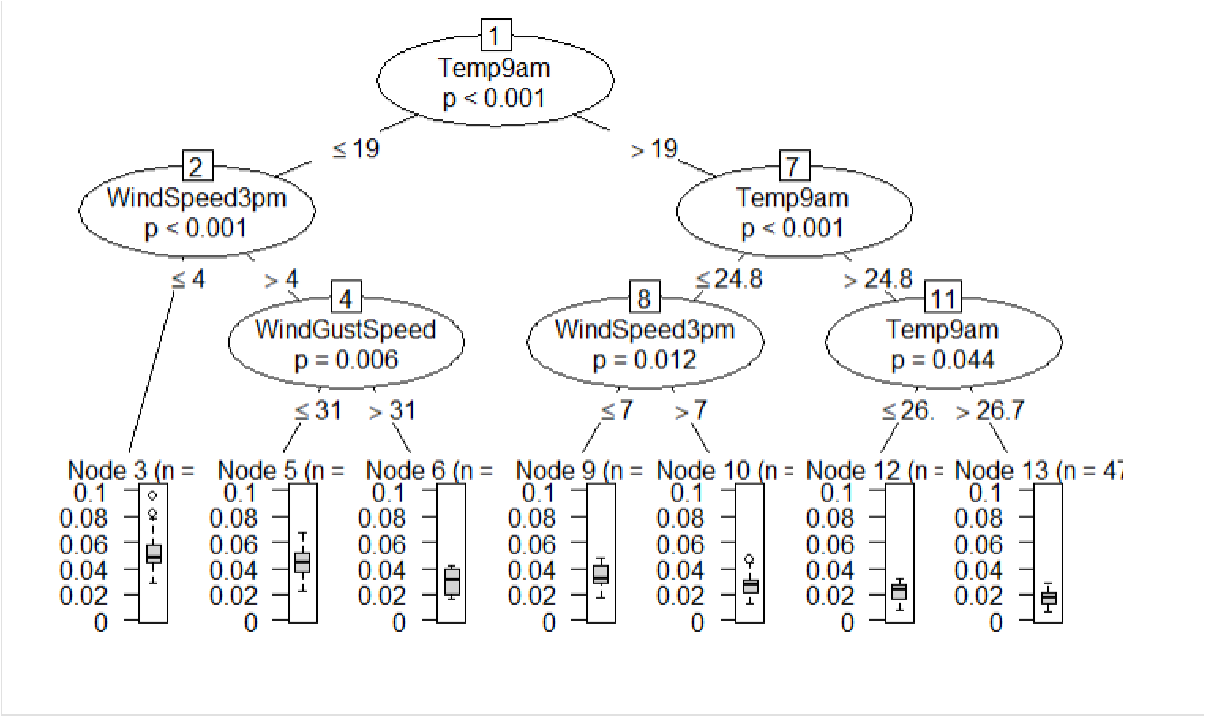
Pm10

In that diagram, we made the indicator of Pm10 as the target data. The diagram shows us the humidity will influence the index of Pm10. When the index of humidity is lower than 39, the index of Pm10 is much higher than that when the humidity is higher than 55. Meanwhile, the index is significantly lower when the weather is rainy. In addition, when the air pressure lower than 1008.9, the index of Pm10 become higher. And the index is higher when the temp becomes higher.



Carbo Monoxide

Through the analysis of the above diagram, we found that the temperature and wind speed are the mainly influencing factors when we made the carbo monoxide as the target data in decision tree. The trend of the four nodes show the index of Carbo Monoxide is gradual decreasing from left to the right, which reflects that with the increasing of wind speed and decreasing of temperature, the index of the Carbo Monoxide will be decreased.

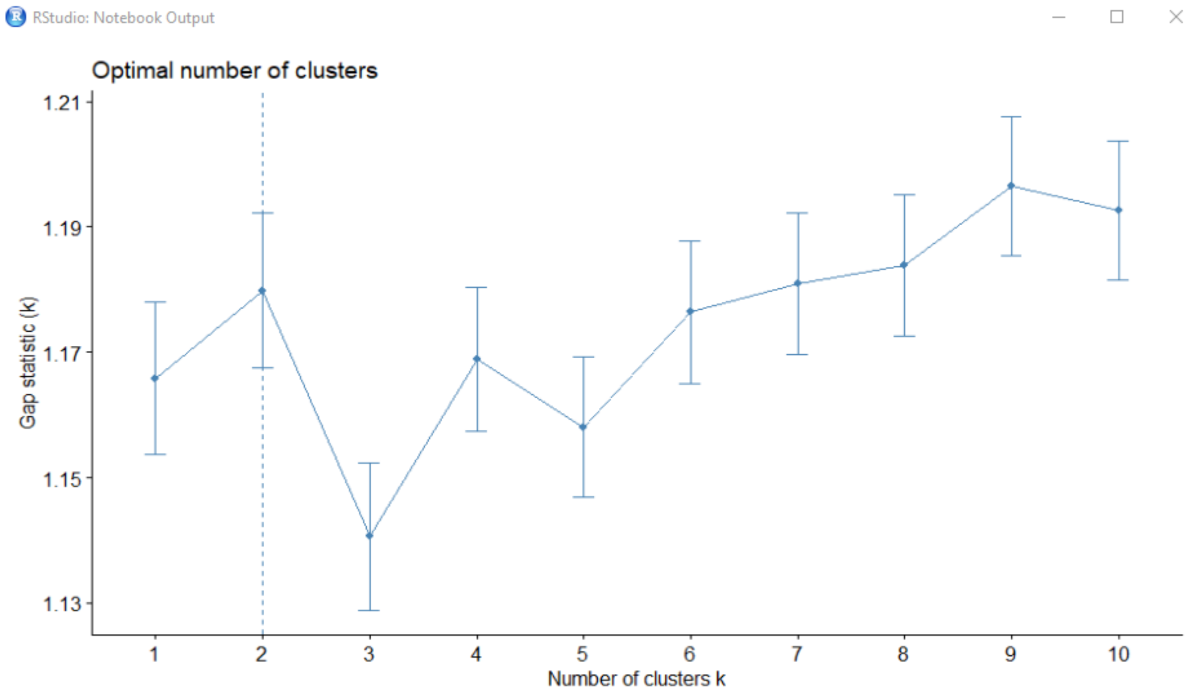


Nitrogen Oxides

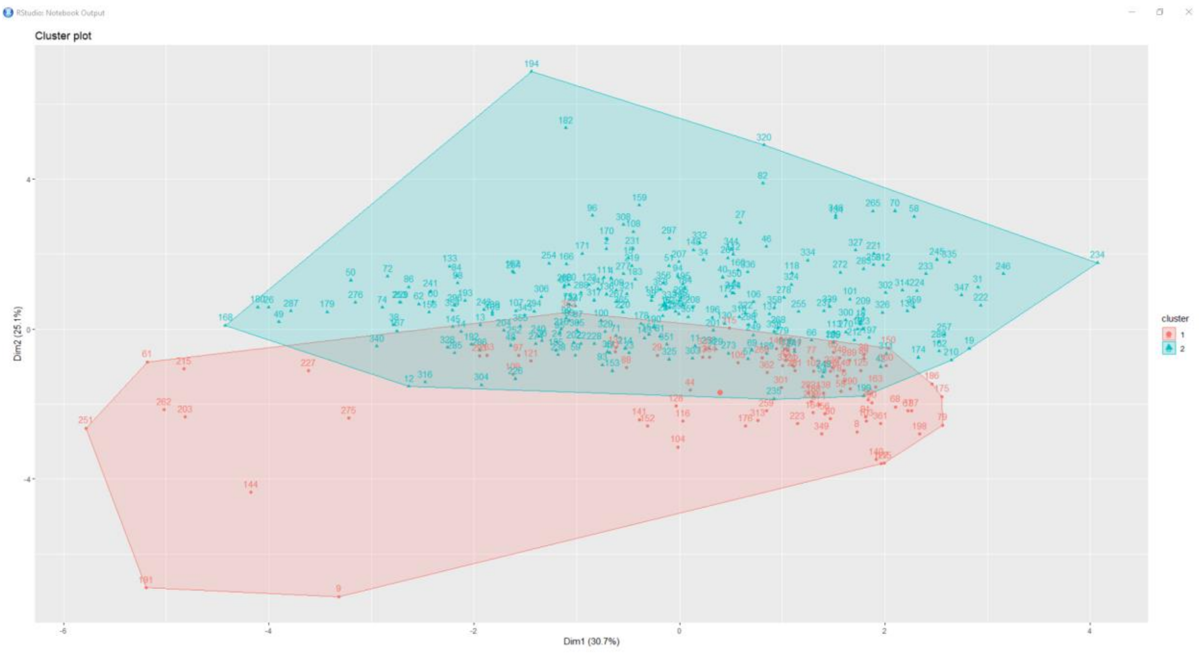
In the decision of Nitrogen Oxides, the trend of the temperature becomes higher from left to the right. But the trend of the index of Nitrogen Oxides becomes lower from left to the right. According to the McGregor and Bamzelis (1995), we know that the structural of Nitrogen Oxides is unstable, they will become NO, NO2 and other are pollutants when it exposed to the light and heat.

To sum up, through the analysis of the above decision trees, we found that the humidity and wind (speed & direction) are the main factors which are affecting the air quality. The air quality is good when the index of humidity and wind speed is high (Wind Direction: East). Therefore, we can know that the air quality is better in rainy days and the air quality is worse in sunny days.

5. Clustering



Through the analysis, we find the 2 is the local optimal, so that we draw the following diagram with 2 clusters.



6. Reference

McGregor, G. R., & Bamzelis, D. (1995). Synoptic typing and its application to the investigation of weather air pollution relationships, Birmingham, United Kingdom. *Theoretical and Applied Climatology*, *51*(4), 223-236.

Rosas, I., Mccartney, H., Payne, R., Calderón, C., Lacey, J., Chapela, R., & Ruiz-Velazco, S. (1998). Analysis of the relationships between environmental factors (aeroallergens, air pollution, and weather) and asthma emergency admissions to a hospital in Mexico City. Allergy, 53(4), 394–401. https://doi.org/10.1111/j.1398-9995.1998.tb03911.x