Sophie Youk (sy5qm)
Professor Chao Du
Statistics 6130
4 May 2020

Air Pollution in Seoul

1. Background

I obtained data of *Air Pollution in Seoul* on the website Kaggle (Kim, 2020). Seoul Metropolitan Government (SMG) has collected and provided many public data including air pollution information. There are several stations measuring air pollution in South Korea including Seoul. The air pollution has been very serious issue in many countries, and of course, in South Korea too. I have raised and lived in Seoul for so long, so I have heard air pollution issues a lot of times. I wanted to see how much the pollutants are measured in the city and whether there are any relationships between them and other factors such as time or location.

2. Summary of Data

Measurement date: Measurement date and time
Station code: Measuring station code
Address: Address of measuring station
Latitude: Latitude of address
Longitude: Longitude of address
SO2: Sulfur dioxide
NO2: Nitrogen dioxide
O3: Ozone
CO: Carbon monoxide
PM10: Particulate matter
PM2.5: Particulate matter

The 25 mearing stations in Seoul has measured sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), particulate matter (PM_{10}), and particulate matter ($PM_{2.5}$) hourly. The dataset has measurements from 12 AM on 1 January 2017 to 11 PM on 31 December 2019. Every station has its own code (101 to 125), and address, latitude, and longitude indicate where the stations are located. Originally, there were 647,511 observations. However, I used data by 8 hours which were 80,939 observations since I couldn't perform models with the full dataset in my computer. I found additional information about PM_{10} and $PM_{2.5}$ on Australian Government website: PM_{10} is particulate matter 10 micrometers or less in diameter, $PM_{2.5}$ is particulate matter 2.5 micrometers or less in diameter. $PM_{2.5}$ is generally described as fine particles. By way of comparison, a human hair is about 100 micrometers, so roughly 40 fine particles could be placed on its width (Particulate matter (PM10 and PM2.5), 2020).

3. Research Problem

I would like to focus on the data of four air pollutants and the relationship with PMs, and further, how they are related to the location (latitude, longitude, address, or stations code) and time (measurement date). Since the air pollution around the world is neither trivial nor temporary issue, researching air pollution data of Seoul can related to similar issues in other places, especially cities.

4. Model, Method, and Analysis

4-1.

I fitted multivariate regression model with measurement date, station code, latitude, and longitude as responses and SO_2, NO_2, O_3, CO, PM_{10}, and $PM_{2.5}$ as predictors. To see whether the assumption that relationship is linear is met, I fitted a studentized residual vs. fitted value plot. Since there is a linear (U) pattern and negative association (Figure 1), I decided to use $X' = \exp(-X)$ for values of SO_2, NO_2, O_3, and CO. Table 1 shows the coefficients of new fitted model. In table 2, with response of longitude, p-value of SO_2 and O_3 are greater than the significance level of 0.05. I tested hypothesis of H_0: $\beta_1 = \beta_3 = 0$ (coefficients of SO_2 and O_3), and I got a result of table 3. Since all p-values are very small and close to zero, the null hypothesis is rejected and, at least one coefficient of the two is not zero. When I tested H_0: $\beta_1 = 0$ and H_0: $\beta_3 = 0$ separately, I got similar result. All of the p-values are very small and close to zero, so both β_1 and β_3 are not zero. What I found interesting with coefficients is only NO_2 out of other pollutants and PMs was positively related to latitude. NO_2 values were recorded higher in northern counties.

4-2.

I fitted another multivariate regression model with PM_{10} and $PM_{2.5}$ as responses and SO_2, NO_2, O_3, and CO as predictors. As we can see in the figure 2, it is hard to say that there is a pattern, so I decided to use the model for further methods. As we can see from the Table 4 and table 5, all p-values were very small and close to zero. Hence, all of four pollutants were significant at the level of 0.05. Out of four air pollutants, O_3 had p-values relatively greater other three. Interestingly, according to the coefficients, SO_2 was negatively related to PM_{10} and $PM_{2.5}$ values, which meant when SO_2 values were high, values of PM_{10} and $PM_{2.5}$ were recorded low.

4-3.

To do exploratory data analysis, I visualized the data of four air pollutants (SO_2, NO_2, O_3, and CO). Figure 3, 4, 5, and 6 shows values of each pollutants by PM_{10} and $PM_{2.5}$. Generally, the values of PM_{10} and $PM_{2.5}$ are not very different. However, interestingly, 51 datasets have very high PM_{10} values higher than approximately 2,000. As we can see in Table 6, 50 out of 51 stations are 116, 117, or 122. I fitted a linear model with PM_{10} as response and SO_2, NO_2, O_3, CO, code, and $PM_{2.5}$. Except the station codes, based on p-values, four air pollutants and $PM_{2.5}$ value are not significant at the level of 0.05. Since Seoul is a small city, the latitudes and longitudes are not very different, so I didn't include them when fitting the model and focused on station codes. According to the code information, four stations (116, 117, 121, and 122) are located southern or southwestern part of Seoul. Other than that, date and times are very various, so it is hard to find what made the 51 values of PM_{10} extremely high.

4-4.

I used `Mclust` function from `mclust` package to cluster the four air pollutants (SO_2, NO_2, O_3, and CO). As Table 8 tells, there are 9 clusters in the dataset. Then, I performed dimension reduction with $\lambda = 1$ and created a cluster plot (Figure 7). Also, I used `classError` function to see if there were any misclassified variables, but there were not misclassified variables (Table 9). To see average silhouette of dataset under 3, 5, and 7 clusters with K-mean method, I had to created new dataset by a week since R failed running with vector memory exhausted, and then standardized the dataset with `scale` function. According to 3 features in Figure 8, it seemed the optimal number of clusters was 3. There were also no misclassified variables (Table 10). As Figure 8 indicates, data with 3 clusters were mostly centered zero.

5. Conclusion and Discussion

I used several multivariate regression models with geographical data and air pollutant values. When I fitted multivariate regression model with measurement date, station code, latitude, and longitude as responses and SO_2, NO_2, O_3, CO, PM_{10}, and $PM_{2.5}$ as predictors, with response of longitude, p-value of SO_2 and O_3 are greater than the significance level of 0.05. I tested hypothesis of H_0: $\beta_1 = \beta_3 = 0$ (coefficients of SO_2 and O_3), and since all p-values are very small and close to zero, the null hypothesis is rejected. At least one coefficient of the two is not zero. What I found interesting with coefficients is only NO_2 out of other pollutants and PMs was positively related to latitude. NO_2 values were recorded higher in northern counties. Additionally, when I fitted another multivariate regression model with PM_{10} and $PM_{2.5}$ as responses and SO_2, NO_2, O_3, and CO as predictors, all p-values were very small and close to zero (Table 4 and Table 5). Hence, all of four pollutants were significant at the level of 0.05. Out of four air pollutants, O_3 had p-values relatively greater other three. Interestingly, according to the coefficients, SO_2 was negatively related to PM_{10} and $PM_{2.5}$ values, which meant when SO_2 values were high, values of PM_{10} and $PM_{2.5}$ were recorded low.

For exploratory data analysis, I visualized the data of four air pollutants (SO_2, NO_2, O_3, and CO). Figure 3, 4, 5, and 6 shows values of each pollutants by PM_{10} and $PM_{2.5}$. Generally, the values of PM_{10} and $PM_{2.5}$ are not very different, but interestingly, 51 datasets have very high PM_{10} values higher than approximately 2,000. As we can see in Table 6, 50 out of 51 stations are 116, 117, or 122. I fitted a linear model with PM_{10} as response and SO_2, NO_2, O_3, CO, code, and $PM_{2.5}$. Except the station codes, based on p-values, four air pollutants and $PM_{2.5}$ value are not significant at the level of 0.05. According to the code information, four stations (116, 117, 121, and 122) are located southern or southwestern part of Seoul. Other than that, date and times are very various, so it is hard to find what made the 51 values of PM_{10} extremely high. As Table 8 shows, there are 9 clusters in the dataset of the four air pollutants (SO_2, NO_2, O_3, and CO). Also, I used `classError` function to see if there were any misclassified variables, but there were not misclassified variables (Table 9). According to 3 average silhouette of dataset features in Figure 8, it seemed the optimal number of clusters was 3. There were also no misclassified variables (Table 10). As Figure 8 indicates, data with 3 clusters were mostly centered zero.

One county has only one measuring station, but the size, population, number of factories, and etc. are very various and random. Counties with high density of population and factories may have higher air pollutants values than other counties even though they have fewer population and factories in reality. Since there are other factors causing SO_2, NO_2, O_3, and CO and increasing PM_{10} and $PM_{2.5}$, more information of predictors/factors can improve analyzing the air pollution data. Also, rain sometimes decreases the air pollution measurements, so it is better to analyze data collected in the same condition of weather.

Tables and Figures

```
                     date          code      latitude    longitude
(Intercept)   2.017356e+04 16.828360760  7.925557988 12.264539970
I(exp(-SO2)) -7.554426e+03 -1.094484995 -4.780635427  0.689577739
I(exp(-NO2))  1.385402e+04 -9.038084997 23.855775451  3.606041655
I(exp(-O3))  -2.271918e+04  2.773005251 -7.760107688 -0.987494304
I(exp(-CO))   1.489123e+04  4.951097186 -8.761140770 -3.640679907
PM10         -2.923684e+00  0.004277247 -0.003710106 -0.002439335
PM2.5        -6.549215e-01  0.004742441 -0.009087013 -0.005351515
```

Table 1: Matrix of Estimated Coefficients $\hat{\beta}$

```
Response longitude :

Call:
lm(formula = longitude ~ I(exp(-SO2)) + I(exp(-NO2)) + I(exp(-O3)) +
    I(exp(-CO)) + PM10 + PM2.5, data = seoul8)

Residuals:
     Min       1Q   Median       3Q      Max
-18.3316  -6.0596   0.1333   6.2774  17.8520

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2645400  0.2245961  54.607  < 2e-16 ***
I(exp(-SO2))  0.6895777  0.7370332   0.936    0.349
I(exp(-NO2))  3.6060417  0.8332216   4.328 1.51e-05 ***
I(exp(-O3))  -0.9874943  0.5715594  -1.728    0.084 .
I(exp(-CO))  -3.6406799  0.2077119 -17.528  < 2e-16 ***
PM10         -0.0024393  0.0003556  -6.860 6.95e-12 ***
PM2.5        -0.0053515  0.0006519  -8.209 2.26e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2: Summary of Multivariable Regression Model

```
Multivariate Tests:
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            1 0.0149118 306.2663      4  80929 < 2.22e-16 ***
Wilks             1 0.9850882 306.2663      4  80929 < 2.22e-16 ***
Hotelling-Lawley  1 0.0151375 306.2663      4  80929 < 2.22e-16 ***
Roy               1 0.0151375 306.2663      4  80929 < 2.22e-16 ***
```

Table 3: Testing Hypothesis H_0: $\beta_1 = \beta_3 = 0$ (Coefficients of SO_2 and O_3)

```
lm(formula = PM10 ~ SO2 + NO2 + O3 + CO, data = seoul8)

Residuals:
   Min     1Q Median     3Q    Max 
-969.3  -19.3   -8.6    7.1 3497.7 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   27.4516     0.4865  56.424  < 2e-16 ***
SO2         -237.5091    11.8526 -20.039  < 2e-16 ***
NO2          227.9013    10.4945  21.716  < 2e-16 ***
O3            28.1043     7.9481   3.536 0.000407 ***
CO            20.4110     0.6787  30.076  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 4: Summary of Multivariable Regression Model (PM_{10})

```
lm(formula = PM2.5 ~ SO2 + NO2 + O3 + CO, data = seoul8)

Residuals:
    Min      1Q  Median      3Q     Max 
-733.50  -11.78   -5.37    4.49 1111.72 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   13.5717     0.2669  50.851   <2e-16 ***
SO2         -145.7883     6.5020 -22.422   <2e-16 ***
NO2          141.9918     5.7570  24.664   <2e-16 ***
O3             9.2656     4.3601   2.125   0.0336 *  
CO            15.5879     0.3723  41.870   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 5: Summary of Multivariable Regression Model ($PM_{2.5}$)

116	117	118	119	120	121	122
11	12	0	0	0	1	27

Table 6: 51 Stations with Extreme PM_{10} Value

```
Call:
lm(formula = PM10 ~ SO2 + NO2 + O3 + CO + code + PM2.5, data = highPM10)

Residuals:
    Min      1Q  Median      3Q     Max
-386.52  -13.95   -2.04   18.44  302.70

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3445.7764    82.3912  41.822  < 2e-16 ***
SO2           -47.7260  8770.1342  -0.005    0.996
NO2          -907.4200  1817.8556  -0.499    0.620
O3           -605.4637   966.7636  -0.626    0.535
CO           -110.9103   109.3026  -1.015    0.316
code117     -1379.9955    63.6189 -21.692  < 2e-16 ***
code121     -1193.3650   121.6589  -9.809 1.99e-12 ***
code122     -1379.0979    65.3050 -21.118  < 2e-16 ***
PM2.5          -0.0399     0.1122  -0.356    0.724
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 106.1 on 42 degrees of freedom
Multiple R-squared:  0.9689,    Adjusted R-squared:  0.9629
F-statistic: 163.4 on 8 and 42 DF,  p-value: < 2.2e-16
```

Table 7: Summary of Linear Model

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 9 components:

 log-likelihood     n  df     BIC     ICL
       910650.8 80939 134 1819787 1769442

Clustering table:
    1     2     3     4     5     6     7     8     9
 4576 10456  9493   183   536   468 12684 24944 17599
```

Table 8: Summary of Cluster

```
> seoul8_mis <- classError(seoul_clust$classification, Class)$missclassified
> length(seoul8_mis)
[1] 0
```

Table 9: R Codes 1 to Find Misclassified Variables

```
> seoul168_mis3 <- classError(seoul168_clust3$cluster, Class)$missclassified
> length(seoul168_mis3)
[1] 0
```

Table 10: R Codes 2 to Find Misclassified Variables

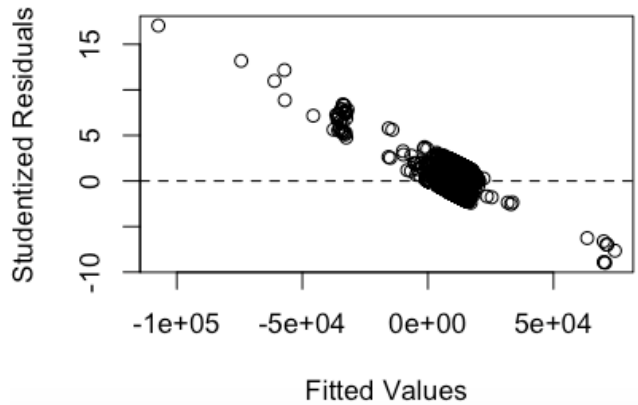

Figure 1: Studentized Residuals vs. Fitted Values 1

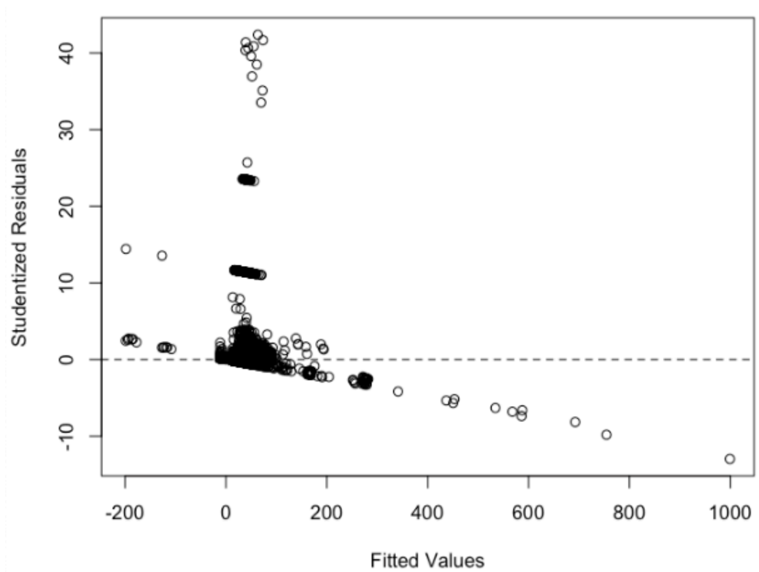

Figure 2: Studentized Residuals vs. Fitted Values 2

Information for Figure 3-6: Red – PM_{10} & Black – $PM_{2.5}$

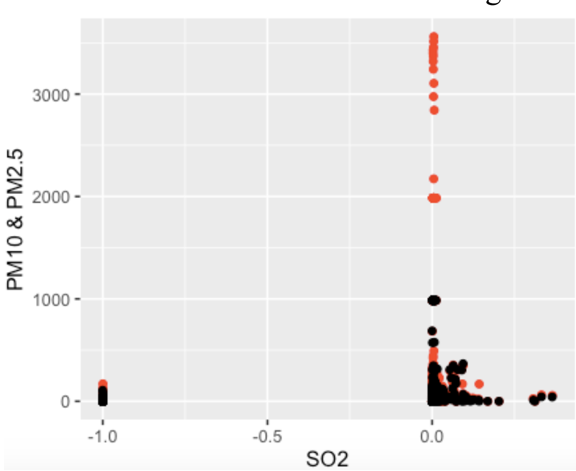

Figure 3: SO_2 vs. PM_{10} & $PM_{2.5}$

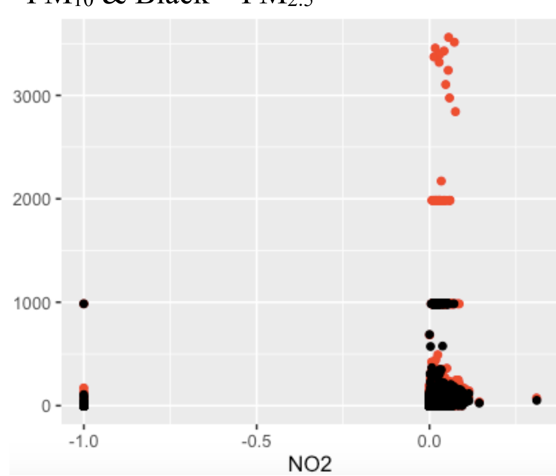

Figure 4: NO_2 vs. PM_{10} & $PM_{2.5}$

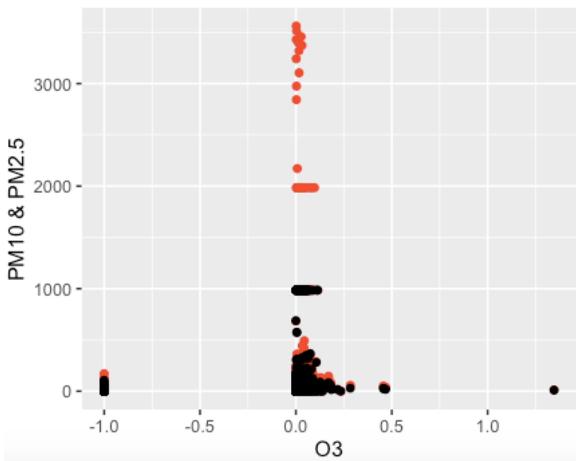

Figure 5: O_3 vs. PM_{10} & $PM_{2.5}$

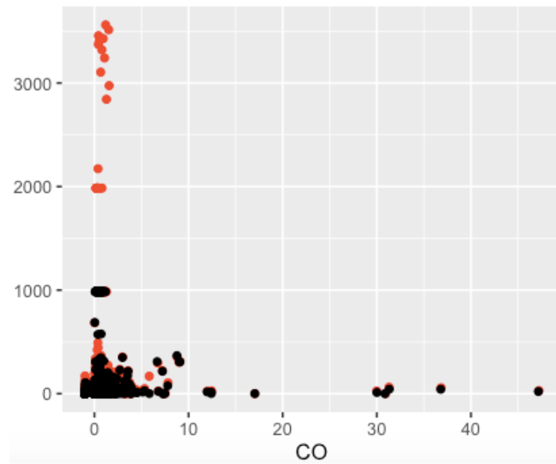

Figure 6: CO vs. PM_{10} & $PM_{2.5}$

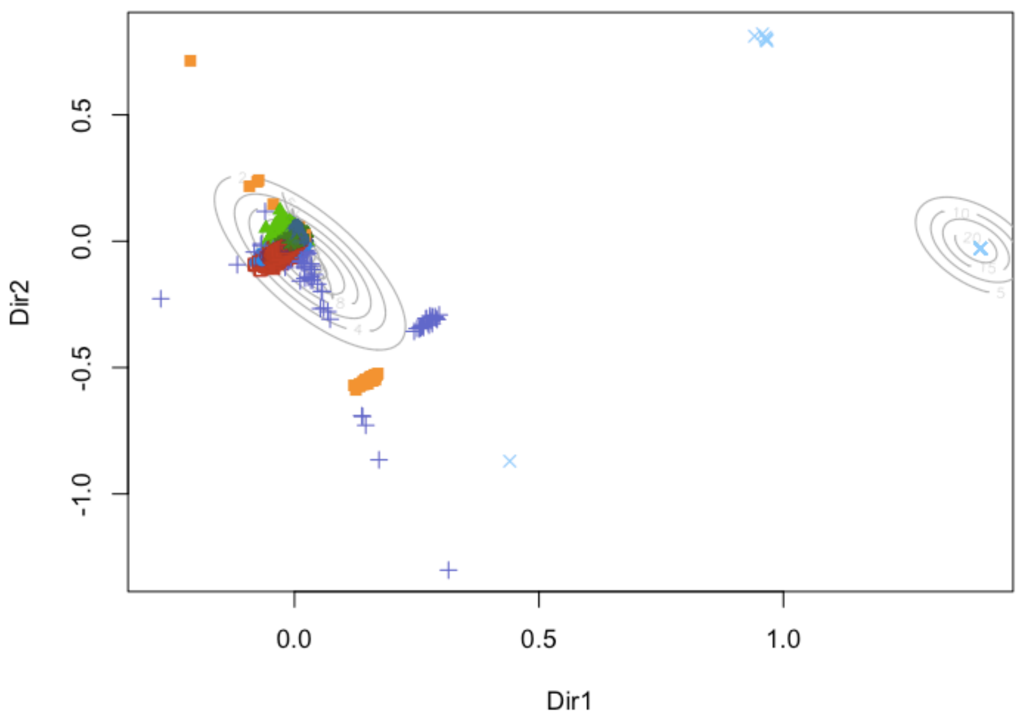

Figure 7: Plot of 9 Clusters

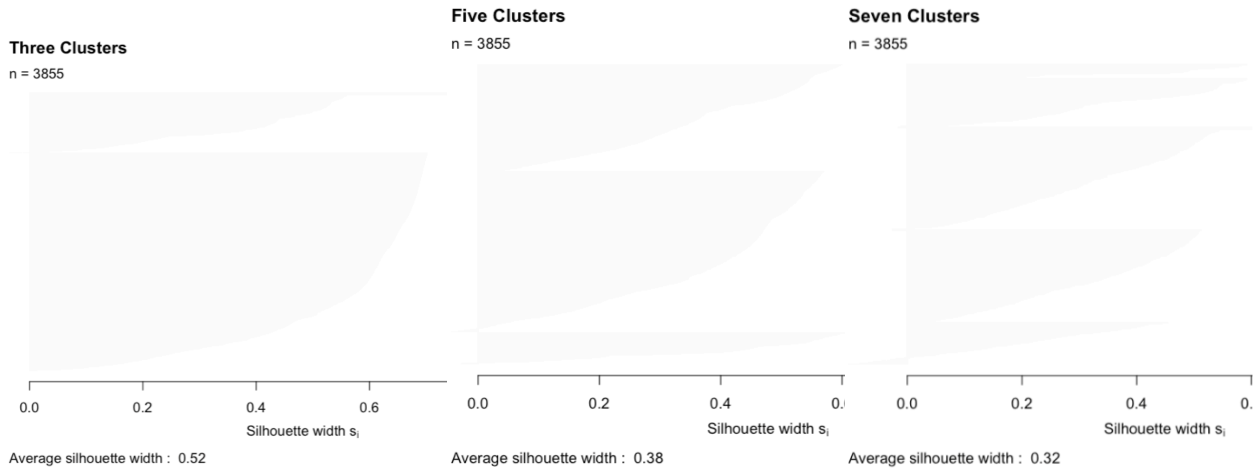

Figure 8: Average Silhouettes under 3, 5, and 7 Clusters

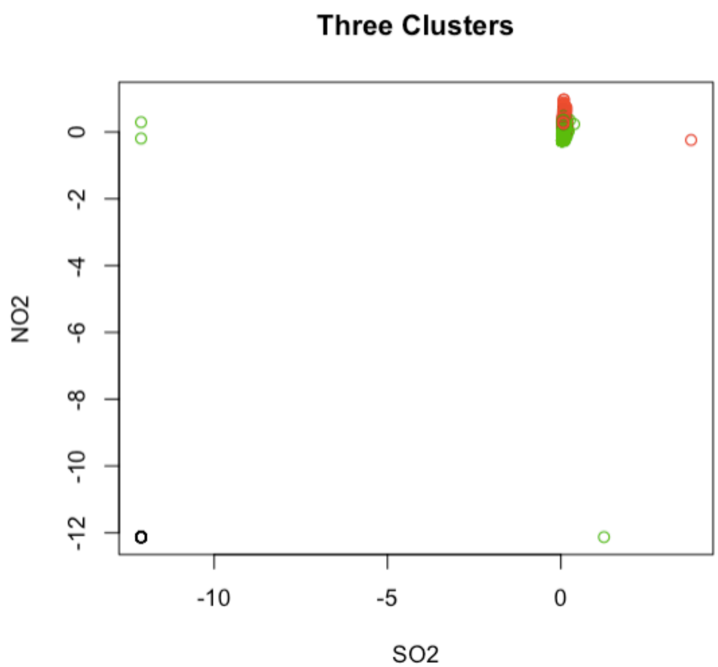

Figure 9: Plot of 3 Clusters

Appendix

```
########## Read and prepare data
seoul <- read.csv("data.csv")
seoul$code <- factor(seoul$code)
seoul$latitude <- factor(seoul$latitude)
seoul$longitude <- factor(seoul$longitude)

seq <- seq(1, 647511, 8)
seoul8 <- seoul[seq,]


########## Multivariate Regression 1

## Multivariate Regression Model
lm1 <- lm(cbind(date,code,latitude,longitude) ~ SO2+NO2+O3+CO+PM10+PM2.5,
data=seoul8)

## Residuals vs. Fitted Values
library(MASS)
fitted <- fitted(lm1)
studres <- studres(lm1)
plot(fitted, studres, xlab='Fitted Values', ylab='Studentized Residuals')
abline(h=0, lty=2)

## Multivariate Regression Model
lm2 <- lm(cbind(date,code,latitude,longitude) ~
            I(exp(-SO2))+I(exp(-NO2))+I(exp(-O3))+I(exp(-CO))+PM10+PM2.5,
data=seoul8)


########## Hypothesis Testing

library(car)

C1 <- matrix(c(0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0), nrow=2, ncol=7)
linearHypothesis(model=lm2, hypothesis.matrix=C1)

C2 <- matrix(c(0, 1, 0, 0, 0, 0, 0), nrow=1, ncol=7)
linearHypothesis(model=lm2, hypothesis.matrix=C2)

C3 <- matrix(c(0, 0, 0, 1, 0, 0, 0), nrow=1, ncol=7)
linearHypothesis(model=lm2, hypothesis.matrix=C3)
```

```
########## Multivariate Regression

## Multivariate Regression Model
lm3 <- lm(cbind(PM10,PM2.5) ~ SO2+NO2+O3+CO, data=seoul8)

## Residuals vs. Fitted Values
fitted3 <- fitted(lm3)
studres3 <- studres(lm3)
plot(fitted3, studres3, xlab='Fitted Values', ylab='Studentized Residuals')
abline(h=0, lty=2)

lm3$coefficients
summary(lm3)


########## Exploratory Data Analysis

library(ggplot2)

## Plot SO2
ggplot(seoul8) +
  geom_jitter(aes(SO2,PM10), colour="red") +
  geom_jitter(aes(SO2,PM2.5), colour="black") +
  labs(x = "SO2", y = "PM10 & PM2.5")

## Plot NO2
ggplot(seoul8) +
  geom_jitter(aes(NO2,PM10), colour="red") +
  geom_jitter(aes(NO2,PM2.5), colour="black") +
  labs(x = "NO2", y = "PM10 & PM2.5")

## Plot O3
ggplot(seoul8) +
  geom_jitter(aes(O3,PM10), colour="red") +
  geom_jitter(aes(O3,PM2.5), colour="black") +
  labs(x = "O3", y = "PM10 & PM2.5")

## Plot CO
ggplot(seoul8) +
  geom_jitter(aes(CO,PM10), colour="red") +
  geom_jitter(aes(CO,PM2.5), colour="black") +
```

```
  labs(x = "CO", y = "PM10 & PM2.5")

## Interestingly high PM10 values
highPM10 <- seoul8[(seoul8$PM10>1500),]
table(highPM10$code)

lm4 <- lm(PM10 ~ SO2+NO2+O3+CO+code+PM2.5, data=highPM10)
summary(lm4)
```

References

Kim, Bappe. *Air Pollution in Seoul*. Kaggle, Mar. 2020, https://www.kaggle.com/bappekim/air-pollution-in-seoul. Accessed 3 Apr. 2020.

Particulate matter (PM10 and PM2.5). Department of Agriculture, Water and the Environment. http://www.npi.gov.au/resource/particulate-matter-pm10-and-pm25. Accessed 3 Apr. 2020.