DSC 498 FINAL REPORT

Tian Xia

12/8/2019

1. Abstract

This report is a temporary report for the China mainland air quality index. My task in this report is to find the potential factors by using cluster analysis. First, I use Elbow method to find the optimal number of clusters. Then, I use the results above to create 4 k-means clusters. Through following the results of k-means cluster analysis, I find 10 new factors to doing the scatterplot matrix for the highest AQI providences Hebei. And finally, I found the relationship between the variables and the air quality index.

1. Background

The air pollution index is an index for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you. In recent years, air pollution has gradually become a serious environmental issue in China. Today, most people who live in high air pollution cities would check the AQI(air pollution index) every day. If the AQI is high, they have to avoid going out or wearing the mask. For example, Beijing is the capital of China, and it is also the biggest city in China. Beijing has twenty-one million people, and it has No.7 GDP compare to other providences(Beijing is a city). However, Beijing's air pollution is one of the worst provinces in China. The high air pollution affects millions of people in these provinces. I come from China, and I hope I can doing something to make this situation better. For the government, this project may help them to decide how to decrease air pollution. For citizens, I hope they can know how dangerous the API is, how it affects human health, and how to avoid it. And also, I hope my project can advocate for most people in China to do something to decrease the API.

2. Introduction and Data

The data source is from Ming Cheng's project.[1], which is a public project created by author Ming Cheng. The project includes every China mainland cities' AQI data for 2000-2015. The author scrapes the data from the Ministry of Ecology and Environment of the People's Republic of China.[2] The AQI data set include five kinds of pollutions in the air: Tropospheric Ozone(O3), Particulates, Sulfur Dioxide(SO2), Carbon Monoxide(CO), Nitrogen Dioxide(NO2). However, in this project, we don't consider each element of air quality separately. The Ministry of Ecology and Environment of the People's Republic of China (MEG) already detected and collected all of them to calculate the AQI in every city of China mainland. The government calculates AQI data hourly, then SO2, NO2, CO concentrations are measured as average per 24h, O3 concentration is measured as an average per hour, and the moving average per 8h, PM2.5, and PM10 concentrations are measured as an average per hour and per 24h. This more details for how AQI calculated are on the airnowtech site.[3]

Air Quality Index (AQI) Values	Levels of Health Concern	Colors
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

age

The AQI divide by 6 levels from low to high air quality. We can see the difference in the following photo. Unfortunately, In China, most cities' AQI is in yellow and orange. Therefore, the AQI problems are urgent to solve for China government.

3. Method

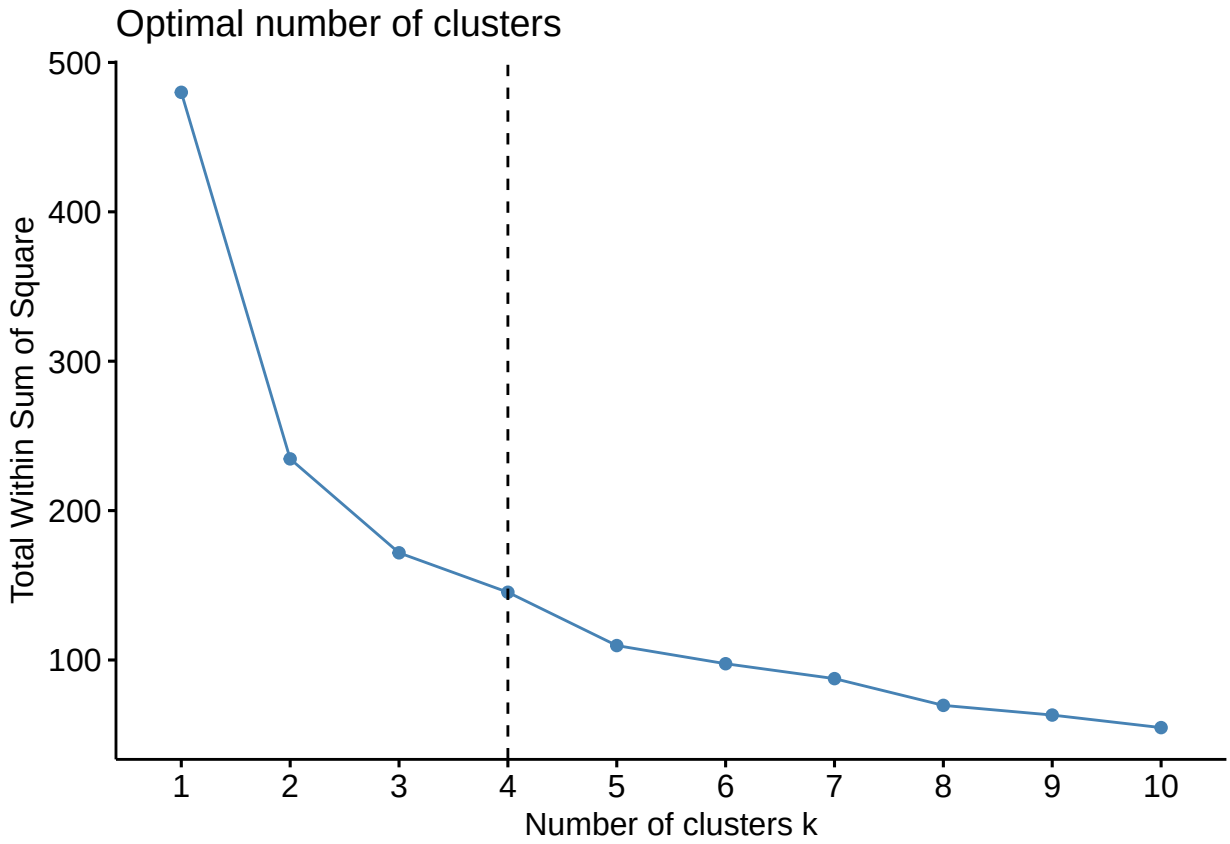
3.1 k-means cluster

K_means clustering is the most common clustering algorithm. The basic idea of this algorithm can be summarized as follows: 1. Choose the number of k clusters 2. Randomly select k center points 3. Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid 4. For each of the k clusters, update the cluster centroid by calculating the new mean values of all the data points in the cluster. The centroid of a Kth cluster is a vector of length p containing the means of all variables for the observations in the kth cluster; p is the number of variables. 5. Iteratively minimize the total within the sum of the square (Eq. 7). That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached. By default, the R software uses 10 as the default value for the maximum number of iterations. We can implement the k-means algorithm in r with the k-means function by following code: code part In the code, centers means the number of clusters, nstart option meanns that attempts multiple initial configurations and reports on the best one. [4] #3.2 Find the optimal number of clusters To do the cluster analysis for the dataet. The Elbow method will help us to find the optimal number of cluster. As I mentioned above that the basic idea behind cluster methods such as k-means, is to define clusters such that the total intra-cluster variation(known as total within-cluster variation or total within-cluster sum of square) We can use the following algorithm to define the optimal clusters: 1. Compute the clustering algorithm for diifferent clusters of k 2. For each key, calculate the total within-cluster sum of square(wss) 3. plot the wss 4. the location of a bend in the plot is generally considered as the optimal number of clusters We can do this algorithm in R by using the fviz_nbclust function.[5] #3.3 Scatterplot matrix The scatterplot matrix is an excellent way to roughly determine if you have a linear correlation between multiple variables. I would use it to show the relationship between each of the variables in the most represented cities Hebei.

```
row.names(data2) <- data[,1]
head(data2)
```

```
##           X2000 X2001 X2002 X2003 X2004 X2005 X2006 X2007 X2008 X2009
## Beijing     101   113   112    98   104    99   110   100    88    85
## Tianjin     110   115    98    94    84    83    85    75    74    78
## Hebei       144   137   124   115    89    92    98    89    84    77
## Shanxi      147   129   120   115   112    97    99    90    77    82
## Neimenggu   100   111    93    83    70    72    75    73    60    63
## Liaoning     88   114   116    96    95    85    84    85    85    81
##           X2010 X2011 X2012 X2013 X2014 X2015
## Beijing      86    81    79   126   134   152
## Tianjin      76    73    78   155   122   136
## Hebei        75    76    74   384   171   180
## Shanxi       73    70    68   113   110   116
## Neimenggu    60    65    71   110    93   100
## Liaoning     77    75    73   153   112   157
```

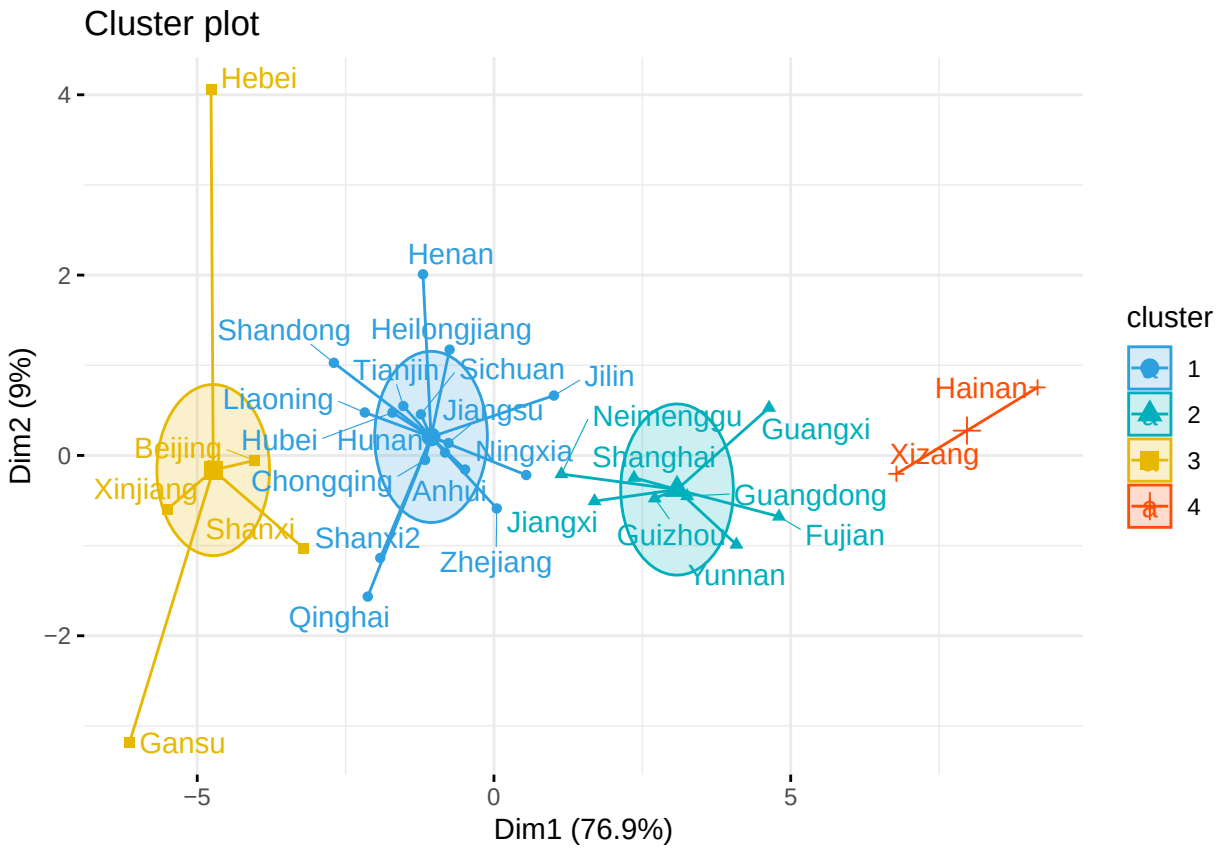
```
df <- scale(data2)
fviz_nbclust(df, kmeans, method = "wss") + geom_vline(xintercept = 4, linetype = 2)
```



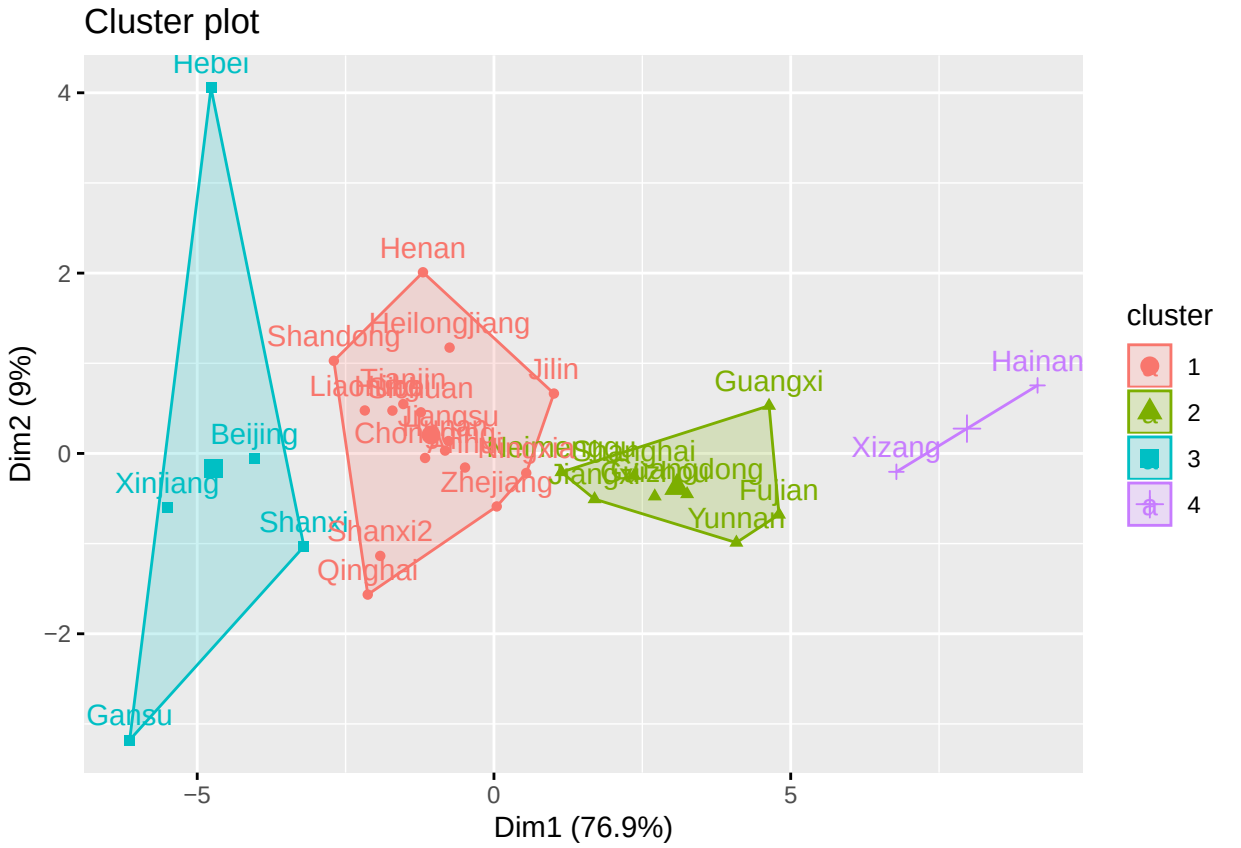
We can see that the 4 cluster is the optimal clusters for my dataset, so I will use the 4 cluster on the following cluster analysis. We can implement k-means cluster analysis by using the kmeans function.

```
fviz_cluster(km_result, data = df,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
             ellipse.type = "euclid",
             star.plot = TRUE,
             repel = TRUE,
             ggtheme = theme_minimal()
)
```

```
## Too few points to calculate an ellipse
```

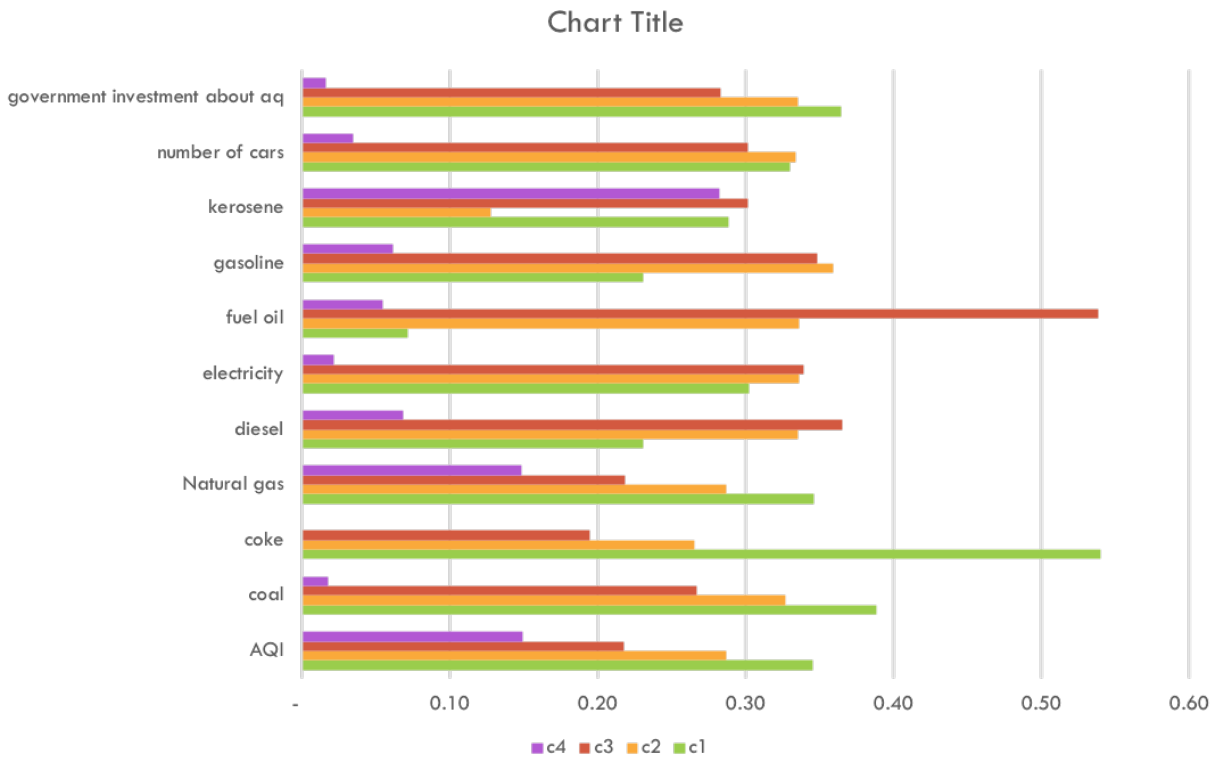



```
fviz_cluster(km_result,data=df)
```



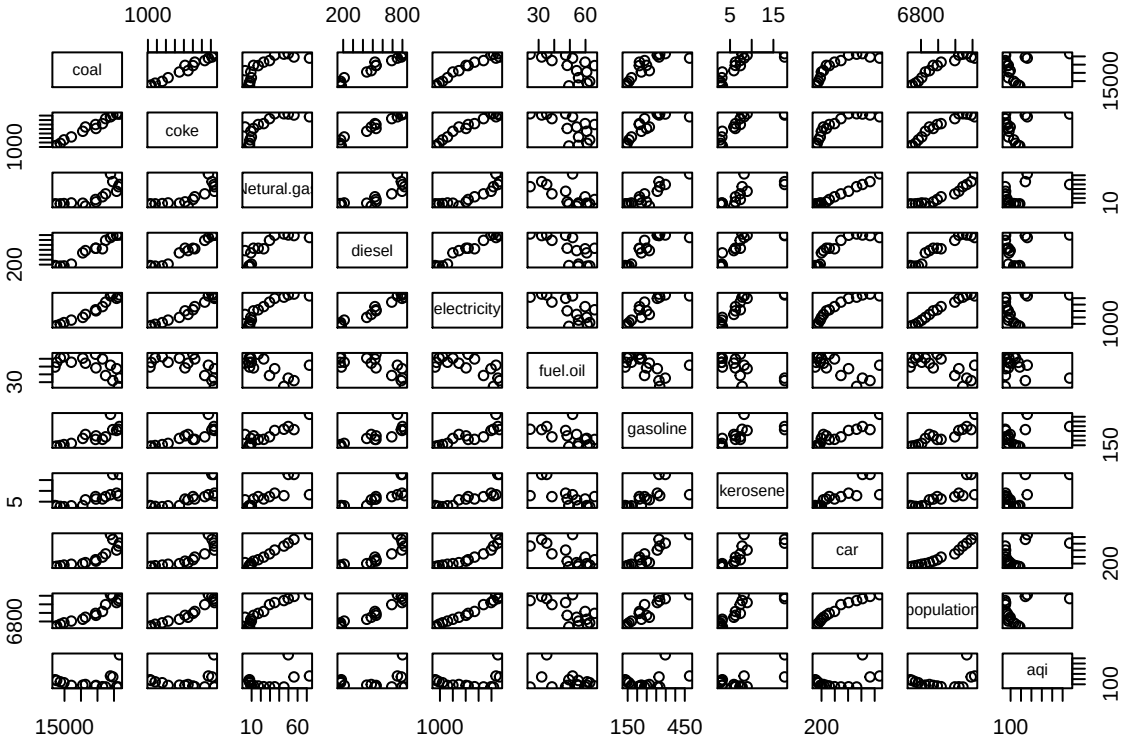
4. Results

There are two different visualizations for the k-means cluster analysis results. We can see that there has 5 cities/province in the first cluster(Hebei,Beijing,Xinjiang,Shanxi,Gansu), 16 cities/province in the second cluster(Henan,Heilongjiang,Tianjin,Sichuan,Jilin,Jiangsu,Ningxia,Anhui,Zhejiang,Shanxi2,Qinghai,Chongqing,Hunan,Hubei,Liaoni cities/province in the third cluster(Guangxi,Shanghai,Guangdong,Fujian,Yunnan,Guizhou,Neimenggu,Jiangxi), two cities/province in the fourth cluster(Xizang,Hainan). After we finished the cluster analysis, we can start to find the potential factors by the classification. My first hypothesis from the plot is about the geographical position. The most north providences is distributing in the first two clusters, which means most of them's AQI is between medium-high and high AQI range. There is two social impression about north of china may relate to the AQI data, cold and industry. Cold-area people need to use some natural resource to make them warm, and most of them would discharge the pollution to the gas like Carbon monoxide and. Industry is the same as the natural resources, and the north of China is the most significant and most crucial industry area in China. After that, I have noticed the lowest AQI cluster, including Hainan and Xizang. These two provinces are famous tourism provinces in China. They both have the lowest population in China, so I also consider the population factors in my list. Expect the population and industry factors, I thought the government might also play an important role in AQI data. For example, Beijing gets better air quality near the 2008 year. May it is because of 2008 Beijing Olympic Games, government increase the budget about environment protecting.<<>plot> Finally, for this semester, I found 10 factors in my list, including the government budget, number of cars, kerosene consumption, gasoline consumption, fuel oil consumption, electricity consumption, diesel consumption, natural gas consumption, coke consumption, coal consumption. Here is the plot about the average comparison between 4 clusters.



We can see that the first cluster's coal and coke data more significant than the other clusters. Then probably relate to the air pollution. To see more details of the relationship between these factors, I choose to use the scatterplot matrix for them. I choose the Hebei as my objective because Hebei has worst AQI data in total 31 providences. We can implement the scatterplot matrix by using the pairs function.

```
Hebei<-read.csv("Hebei.csv",header=TRUE)
pairs(Hebei[c("coal","coke","Netural.gas","diesel","electricity","fuel.oil","gasoline","kerosene","car"
```



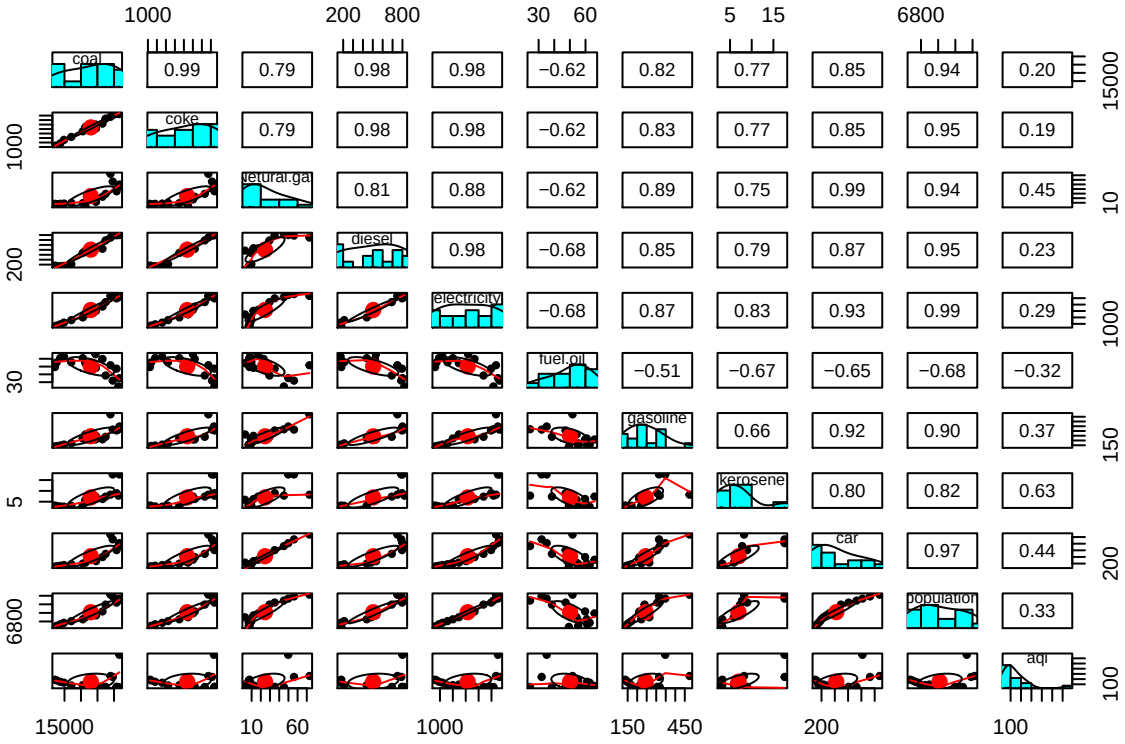
We can see there has some apparent relationship between each factor. For example, coal-coke plot has a straight line. The straight line in the plot points that when the consumption of coke increased, the consumption of coal also increased. However, if we can add more information to the plot, it would be clear to see the relationship between each factor. We can use parirs.panels function from psych library to add more information to the scatterplot.

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
pairs.panels(Hebei[c("coal","coke","Netural.gas","diesel","electricity","fuel.oil","gasoline","kerosene
```



pairs.panels function add two more information to the scatterplot matrix. The ellipse in each scatterplot is a correlation ellipse; it shows the more close relationship between each variable when the ellipse is stretched. For example, we can see in the coal-coke plot; the ellipse is even twisted into a straight line, which shows that there was a strong relationship between coal and coke consumption. Moreover, we can see that all variables have a certain extent affect the AQI data because all ellipse have some deformation on the plot. The second information is the redline; it is the loess smooth; It shows the variation relationship between the variables. The all redline in the AQI scatterplot matrix means that there is no noticeable trend of change. I guess that there may have a threshold in the relationship between variables and AQI. When the variables are over a threshold, they can have a more obvious influence on the AQI data. In most of the AQI scatterplot, the redline has leveled off first and finally go down or up.

5. Conclusion & Future works

This semester, I am trying to find the potential data factors about the air pollution data for the future works. For this semester, I found the 10 factors in the natural resource, industry, government area. I would found 10 more in the next semester. These new variables would focus on government, weather, and economy area. In my plan, finally, I would use the 20+ factors on my models. The scatterplot matrix shows that there exists a relationship between variables an nd AQI. Therefore, in the next semester, I would find more details about the relationship. I am planning to use the regression and principal component analysis for this dataset. The regression analysis is a common method to find the relationships between a dependent variable and one or more independent variables. The principal component analysis is also useful methods for this dataset. Because in my datasets, some of these factors are not independent. For example, the natural source consumption data is no independent; because the energy requirement is fixed; if people use more on coke, they will use less on other resources. The principal component analysis would help us to find several "most important" factors, which are helpful for this situation.

6. Timeline

Jan-Feb Finding 10 more variables for my models, these variables would focus on govnerment, weather and economy area.

Feb-Mar Starting do the regression analysis and principla component analysis for the dataset. Remove some variables, which can't affect to the AQI data.

Mar-Apr Finishing the regression and PCA analysis, start to do some the data vasulization.

Apr-May Finish the data vasulization, writing the final report.

References

[1].Ming Cheng's project: https://github.com/mingcheng/AQI

[2].Ministry of Ecology and Environment of the People's Republic of China: http://datacenter.mep.gov.cn/

[3].airnowtech site:https://forum.airnowtech.org/t/the-aqi-equation/169.

[4].UC Business Analytics R Programming Guide:https://uc-r.github.io/kmeans_clustering

[5].Scatter Plot Matrices- R Base Graphs: http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs