Project for Multivariate Statistical Analysis

Haoyu_Zhao, 2016012390

June 3, 2017

Overview

The air quality is always a hot topic in China in these days, and the word 'PM2.5' is familiar to almost every people in China now. As a student in Tsinghua University, I also concern about the weather, the air quality and the index of the PM2.5 very much, because we have to finish 27 3-kilomiter running in every semester. In this report, I want to use the data to find some characteristics or features of the index of pm2.5 and our air quality, and try to apply some of the multivariate statistics tools to find some information from the data.

Data Set Description

In this project, I found the data set from the UCI machine learning repository. The data set is called 'Beijing PM2.5 Data Data Set', the website is https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data#. This data set is originally used in a air quality analysis, and the reference of that article is in the References part.

The data contains the data from Jan 1,2010 to Dec 31,2014, with 1 record an hour. There are 13 columns in the data set, the first denote the number of the record, the second to fifth denote the year, month, day, hour of each record. The sixth column denote the pm2.5 index. It is integer. The 7th column denotes the dew point(Dew point is the temperature to which air must be cooled to become saturated with water vapor). The 8th column is the temperature of the time. The 9th column is the air pressure of the time. The 10th is the combined wind direction at that time. The 11th column, is the cumulated wind speed of the time. If the wind change direction, the cumualated wind speed will be set to 0. The 12th and 13th columns are the cumulated hours of snow and rain.

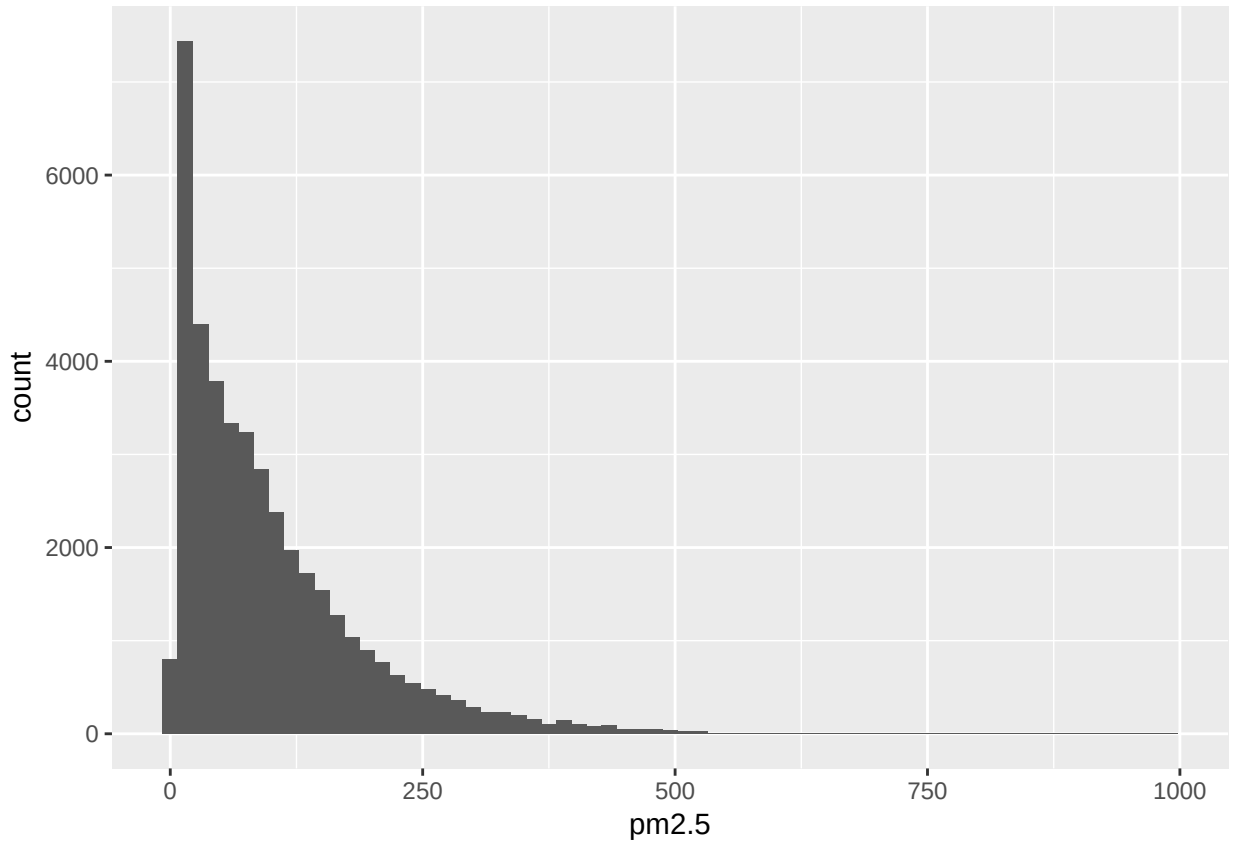
Some Descriptive Statistics

In this section, I will present the analysis, the results and plots of the data, and the codes that generate these results togther.

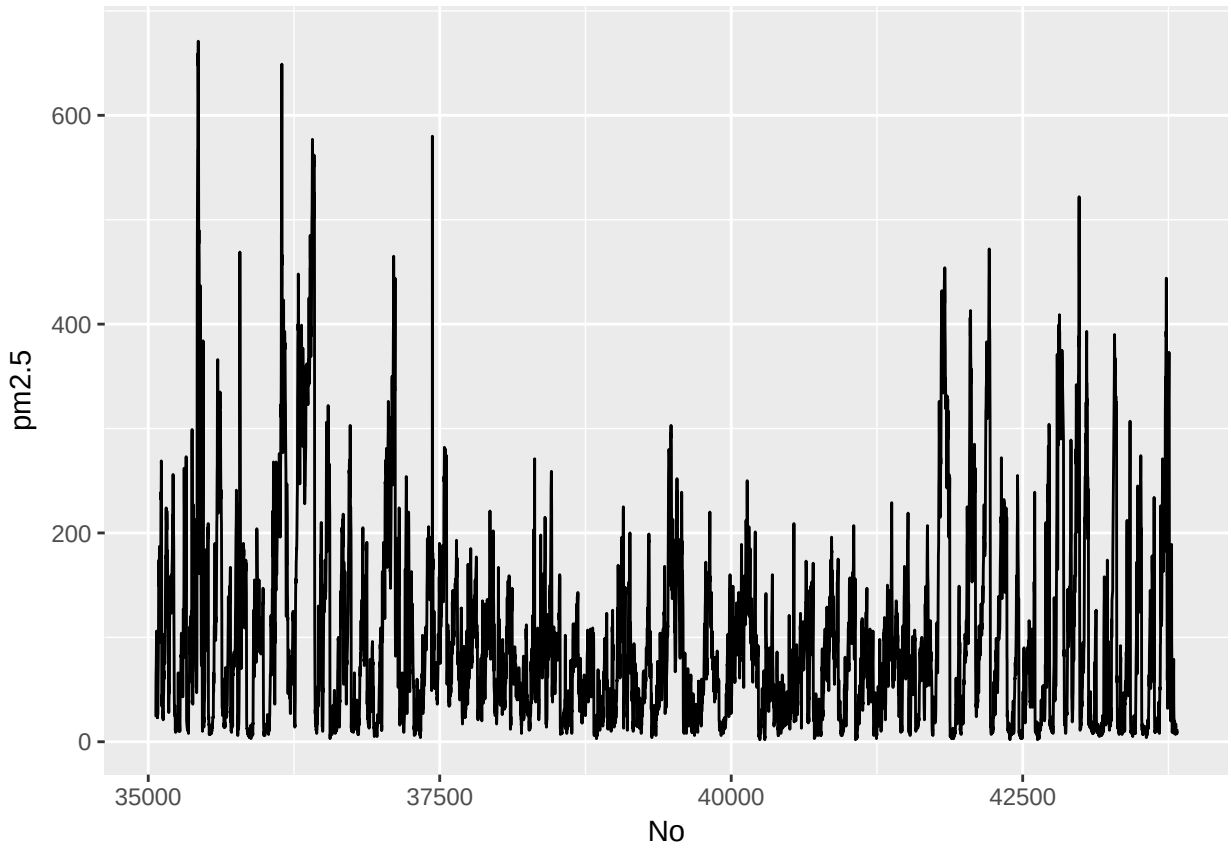
```
##############################################################################
#----------------------begin of the project code-------------------------
#----------------------load the packages---------------------------------
library(ggplot2)
library(corrplot)
```

```
#----------------------read the data from the data set------------------
raw_data <- read.csv('PRSA_data_2010.1.1-2014.12.31.csv', header = TRUE)
#there are some NA values in the data set, so I omit them
data_basic <- subset(raw_data, is.na(pm2.5)==FALSE)
```

```
#----------------------some descriptice statistics----------------------
#----------------------barplot of the observations of pm2.5 data--------
ggplot(data_basic) + geom_histogram(aes(x=pm2.5), binwidth = 15)
```




```
#---------------print the change of the pm2.5 in 2014--------------------
ggplot(subset(data_basic, year == 2014)) + geom_line(aes(x = No, y = pm2.5))
```



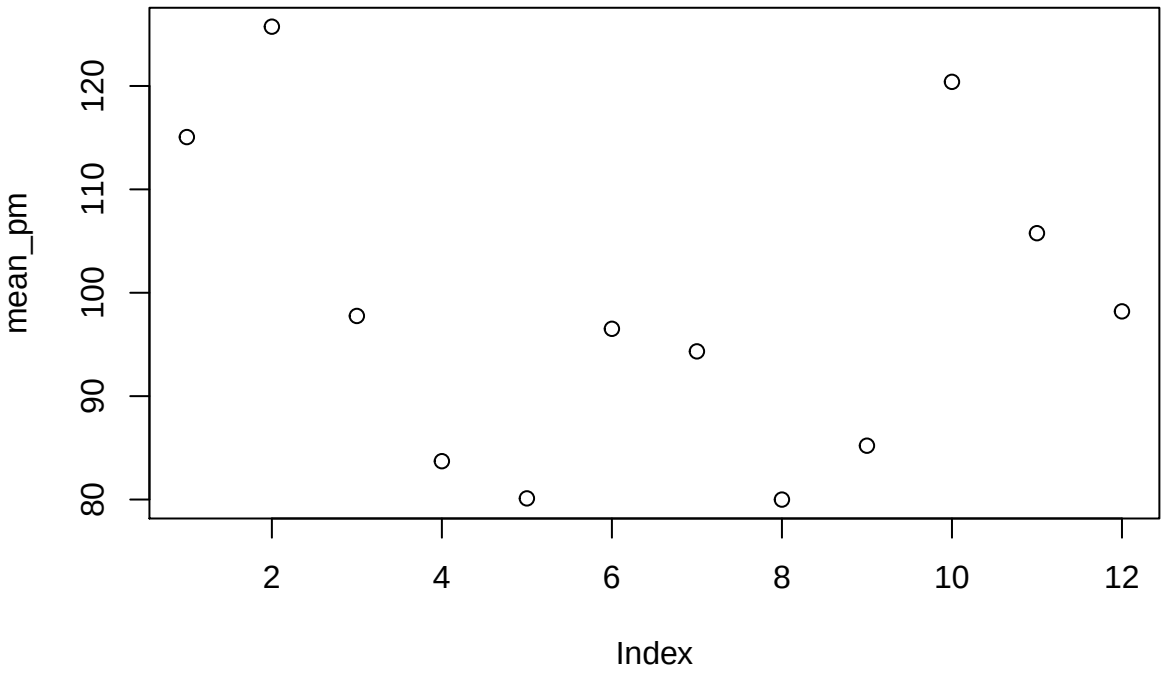

```
#print some
summary(data_basic$pm2.5)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   29.00   72.00   98.61  137.00  994.00
```

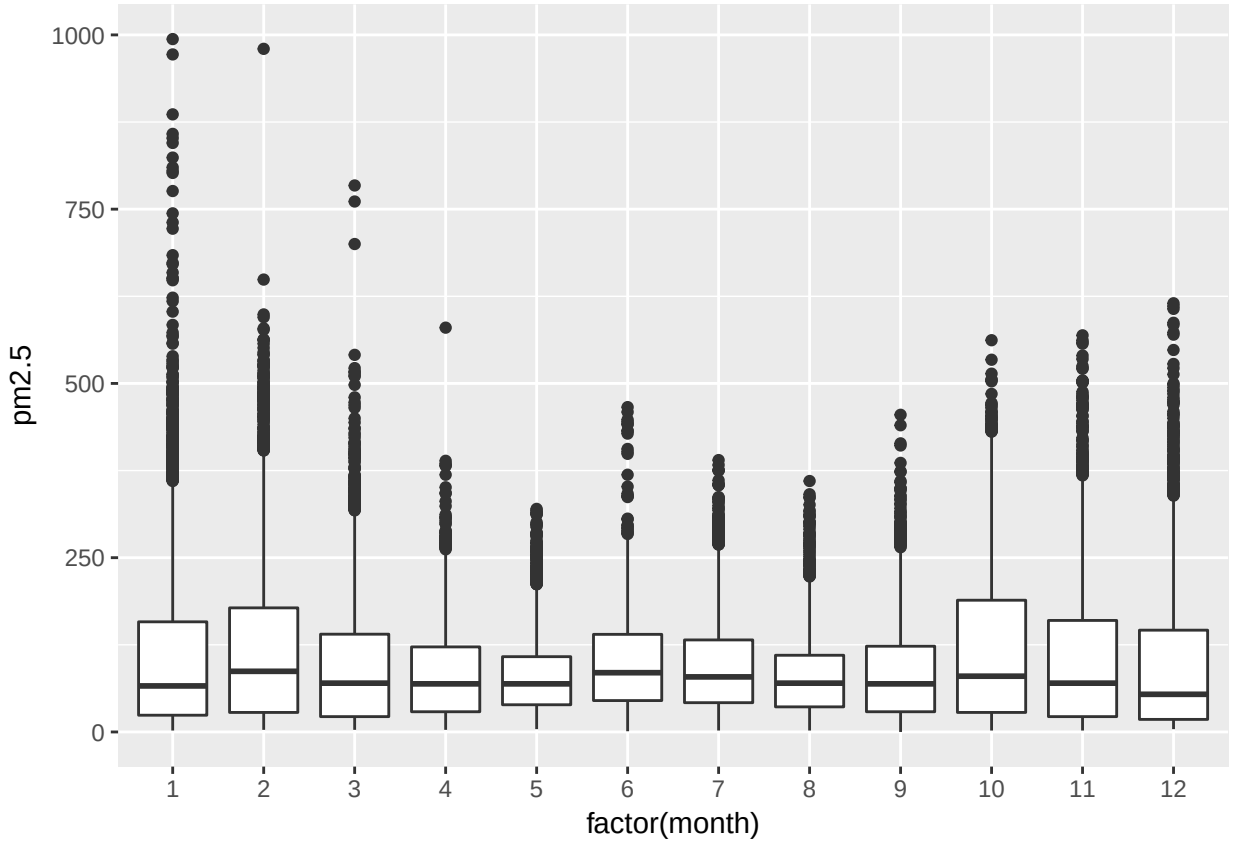
From the histagram and the summary, we can find that actually, most of the time the pm2.5 concentration is not so high, but there are times that the air is polluted seriously. From the plot, we can find that there are days that the index is higher than 500, and many of them are higher than 250.

From the second plot, we can find that most of the high pm2.5 index days are in winter.

```
#--------------------the mean of the pm2.5 of each month----------------
mean_pm <- rep(0,12)
for(i in 1:12) {
  mean_pm[i] = mean(subset(data_basic, month == i)$pm2.5)
}
plot(mean_pm)
```

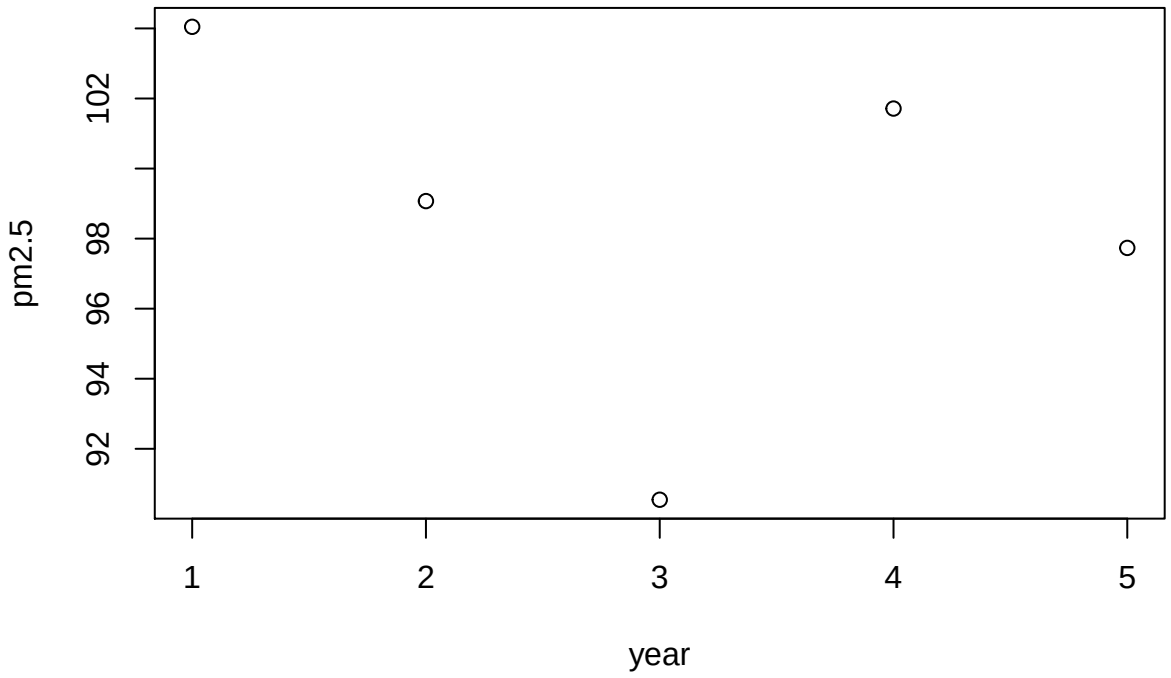



```
ggplot(data_basic, aes(x=factor(month), y=pm2.5)) + geom_boxplot()
```



From the mean and the boxplot, we can find that the mean of the pm2.5 concentration in summer is less than that in winter. The first and third quater of the observations are largely the same, but in winter, there are many days in which the air is badly polluted.

```
#--------------------the mean pm2.5 index of each year-----------------
t <- rep(0,5)
for (i in 2010:2014) {
  temp <- subset(data_basic, year==i)
  t[i-2009] = mean(temp$pm2.5)
}

plot(t, xlab = 'year', ylab = 'pm2.5')
```



From the plot, we can see that the index has a decreasing trend.

```
#---------------then convert to the statistics of each hour--------------
#----------first observe the distinction between day and night---------
data_basic.night <- subset(data_basic, hour < 8 | hour >= 20)
data_basic.day <- subset(data_basic, hour >= 8 & hour < 20)

mean(data_basic.day$pm2.5)
```

```
## [1] 90.61315
```

```
mean(data_basic.night$pm2.5)
```

```
## [1] 106.6037
```

```
var(data_basic.day$pm2.5)
```

```
## [1] 7328.108
```

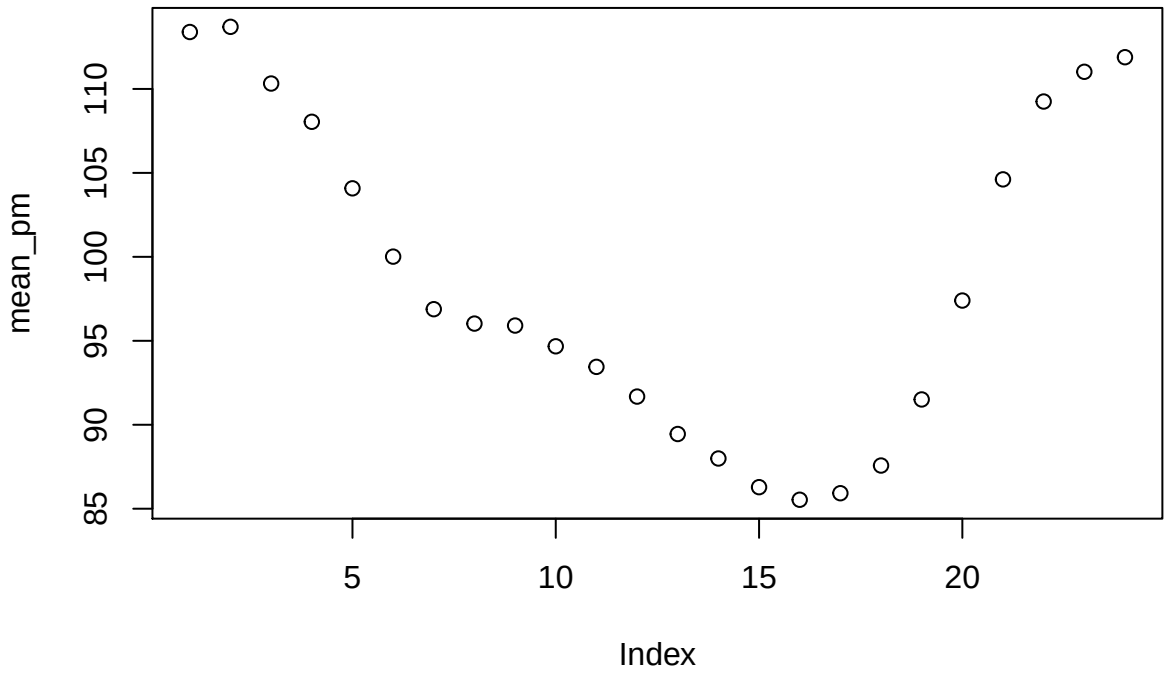
```
var(data_basic.night$pm2.5)
```

```
## [1] 9489.696
```

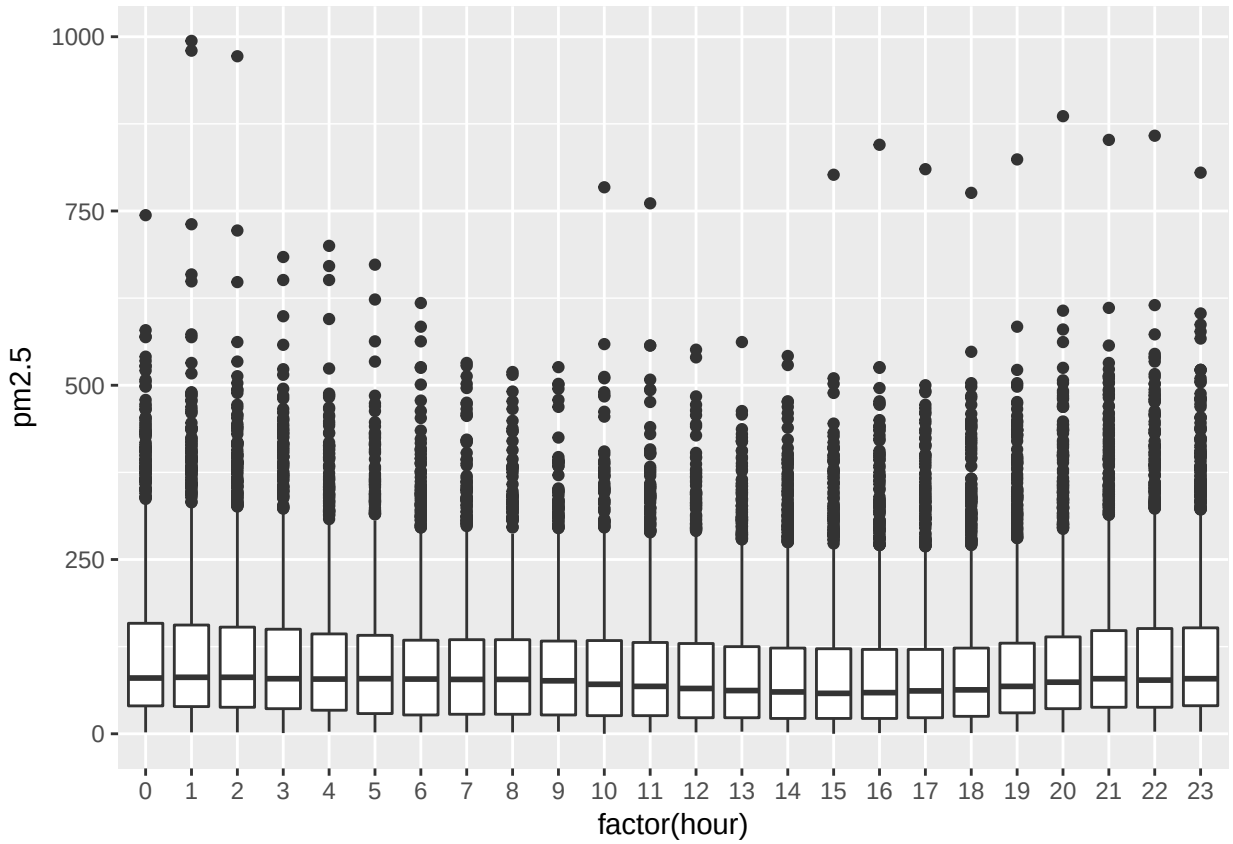
From the statistics, we find that the average air quality of the day is better of that of the night.

```
#--------------the mean of the pm2.5 of each hour----------------------
mean_pm <- rep(0,24)
```

```r
for(i in 0:23) {
  mean_pm[i+1] = mean(subset(data_basic, hour == i)$pm2.5)
}
plot(mean_pm)
```

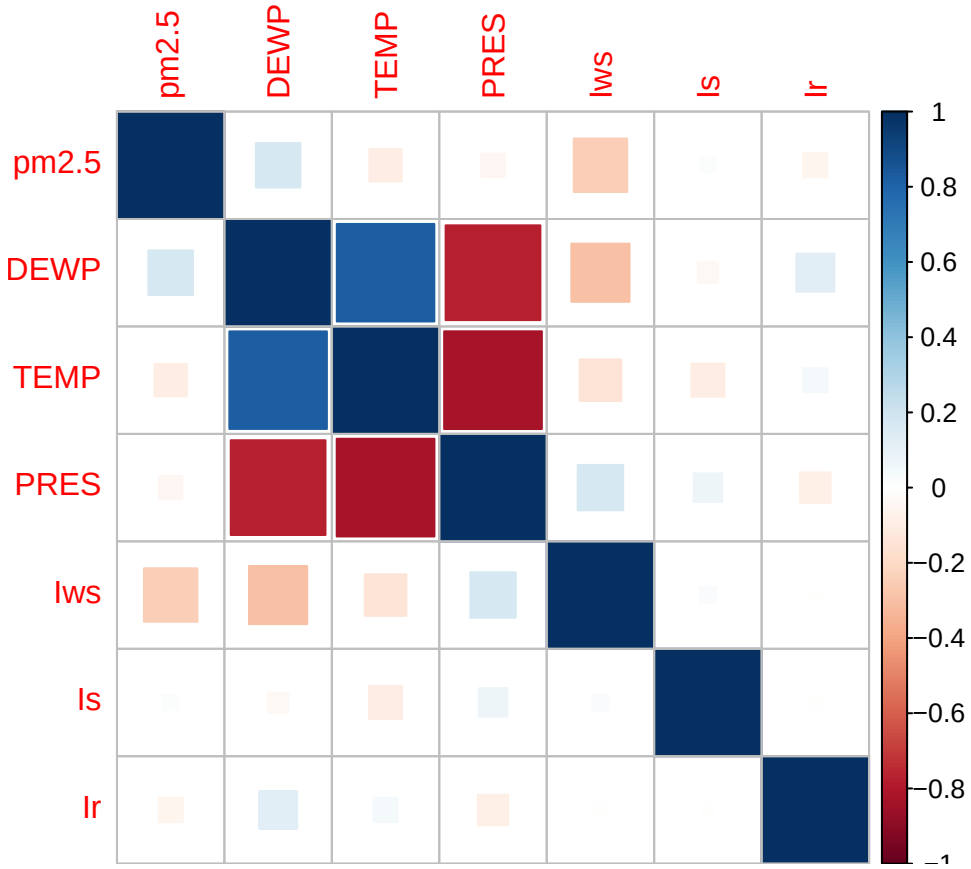



```r
ggplot(data_basic, aes(x=factor(hour), y=pm2.5)) + geom_boxplot()
```



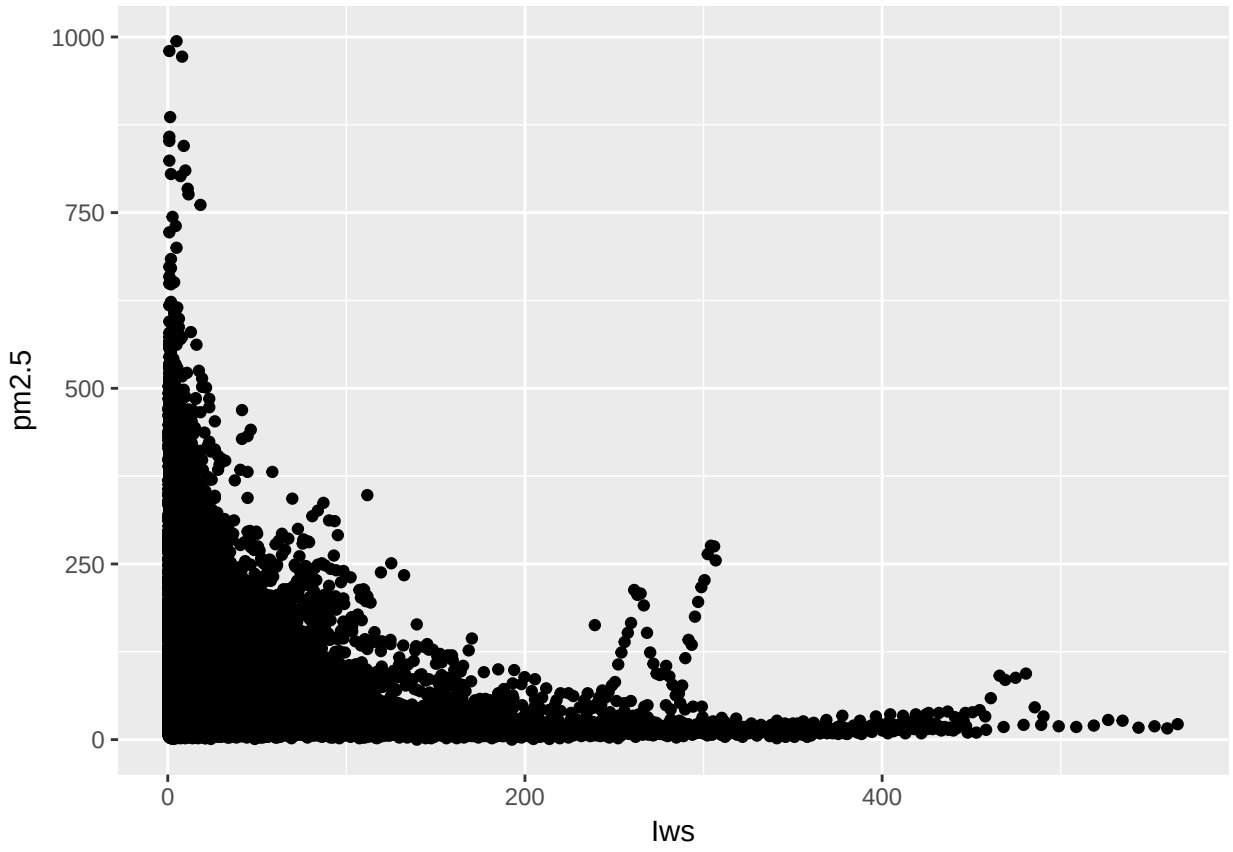
From the results, we find that the mean pm2.5 concentration is at its lowest at about 4:00 pm, and is at its peak at about 1:00 am. So it is better to go outside and do sports at afternoon, and do not get outside and do sport too late. We always omit the pm2.5 at night because we cannot figure out the air quality by just looking outside ,and we go out and do sports without thinking about the air pollution. But in fact, from the graph, we can find that the pm2.5 index rises fast after 6:00 pm, maybe this is due to the evening rush hour.

```
#-------then do some descriptive statistics to more than 1 variable---------
#-----------------the correlation of different variables--------------------
#------------delete the time and the wind direction variables---------------
reg_var <- data_basic[,c(6,7,8,9,11,12,13)]
corrplot(cor(reg_var), method='square')
```

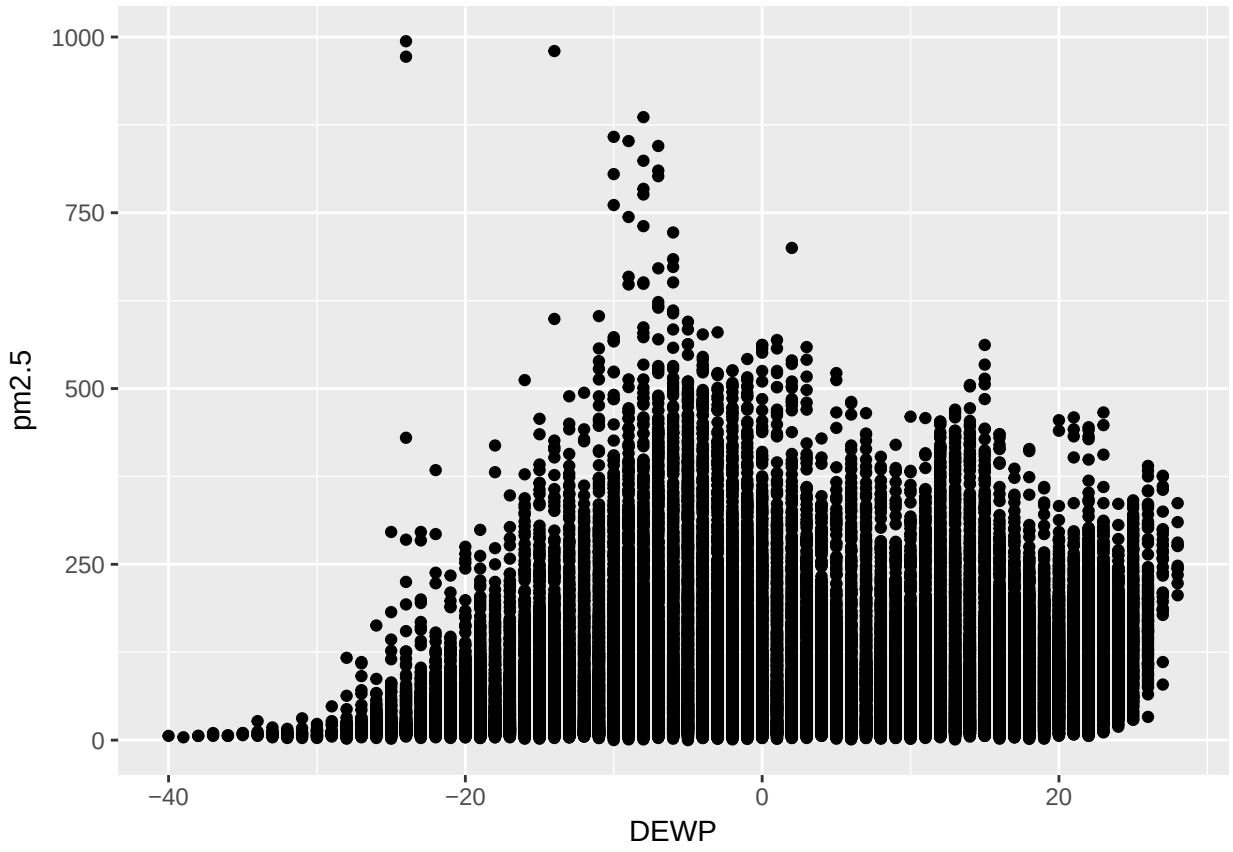


From the visualized correlation plot, we find that the variables 'DEWP'(Dew Point), 'TEMP'(Temperature), and 'PRES'(Pressure) has large correlation. It is normal because these three variables has the physics relation. We also find that 'DEWP', 'TEMP' and 'Iws' has some relation with the index of the pm2.5. We then plot the scatter plot of each these 3 variables with the pm2.5 concentration variable.

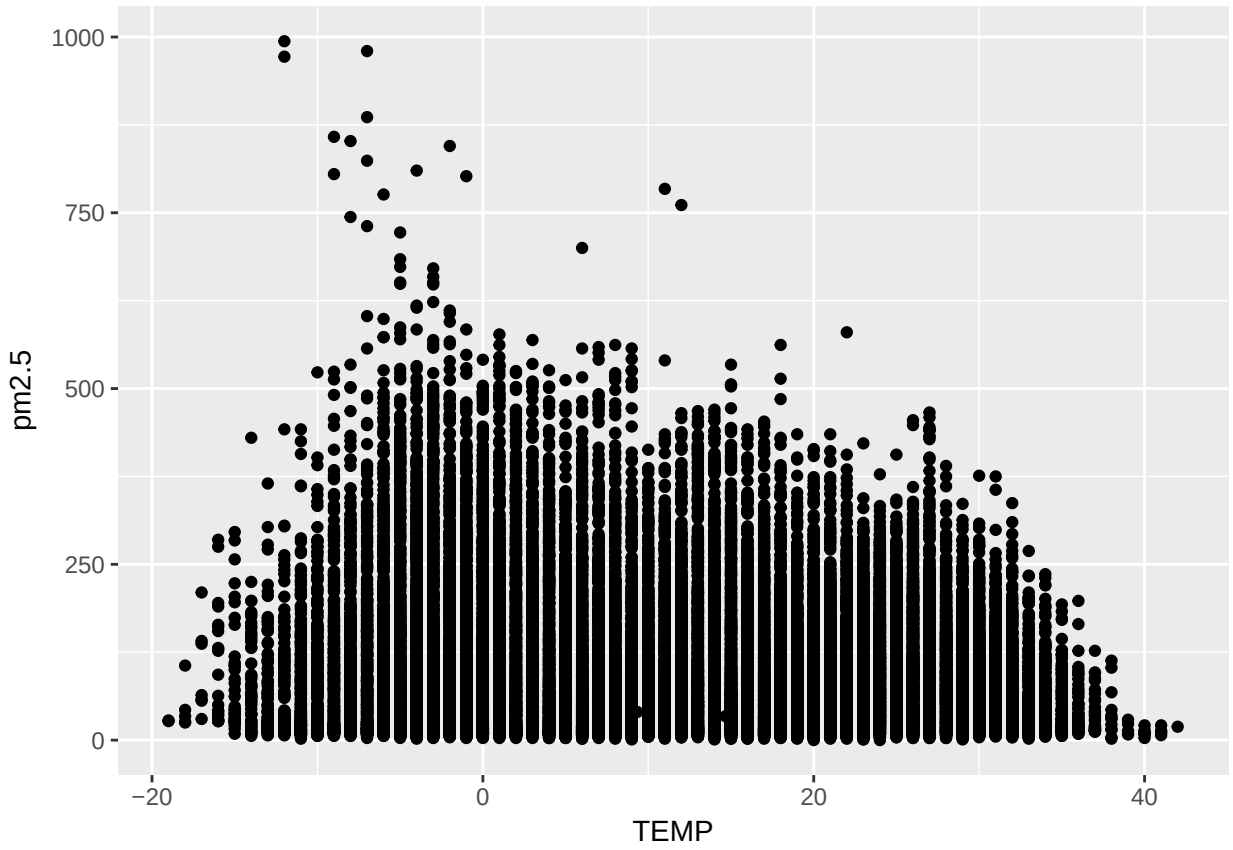
```
#------------------pair scatter plot----------------
qplot(Iws, pm2.5, data=data_basic)
```




```
qplot(DEWP, pm2.5, data=data_basic)
```

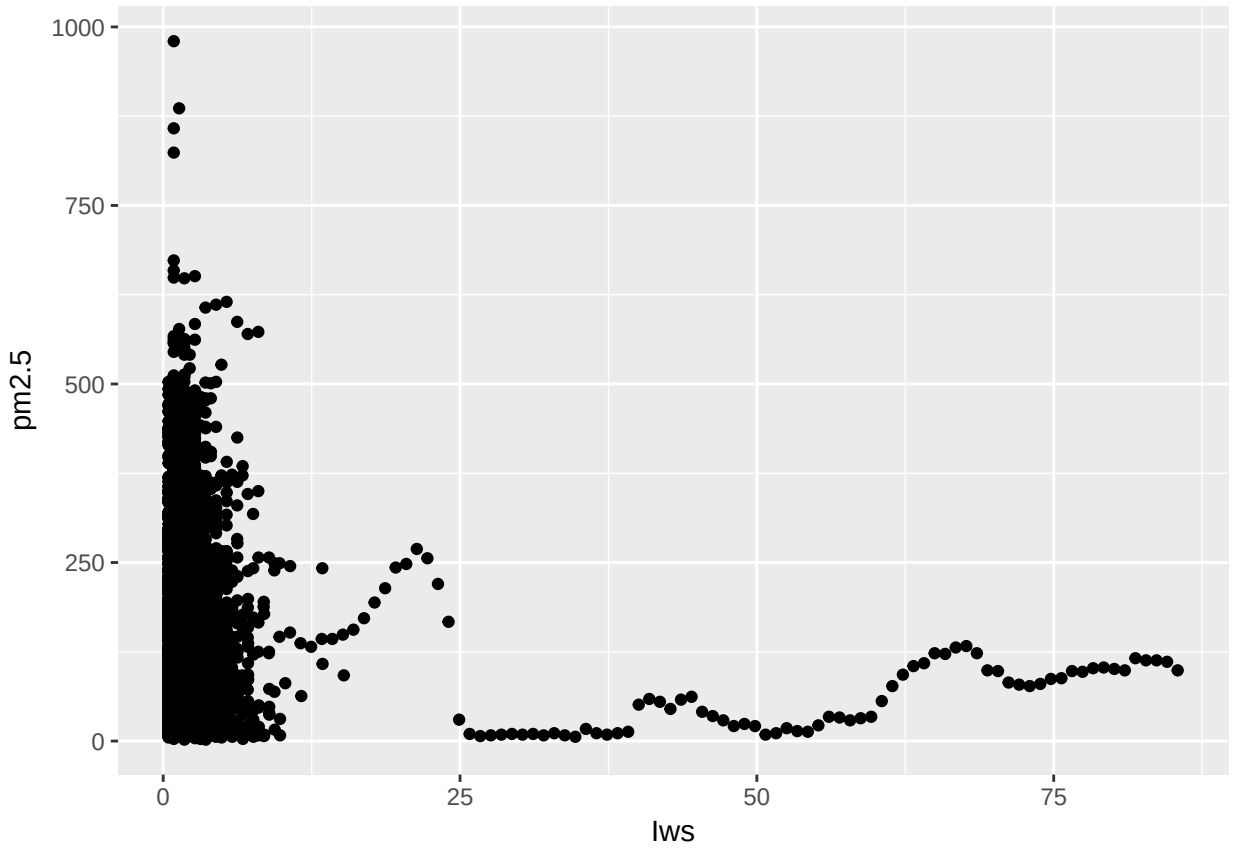



```
qplot(TEMP, pm2.5, data=data_basic)
```

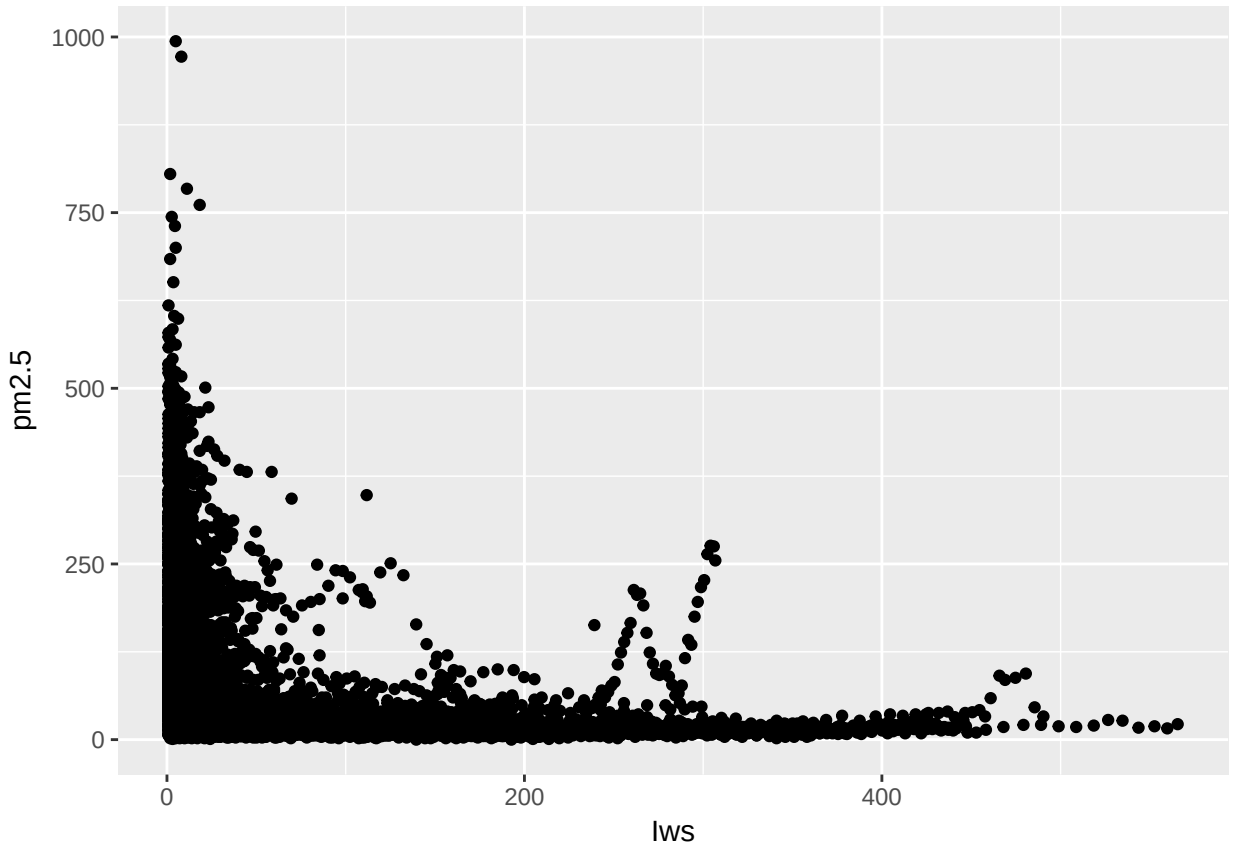


From the plots, we can see that the pm2.5 has soem correlation with the cumulated wind speed. When the cumulated wind speed is larger, the index of the pm2.5 tends to decrease. But the other 2 plots do not show very good trend of the x-axis variable and the index of the pm2.5. In the plot of the cumualted wind speed, we omit the direction of the wind, then we plot the graph considering the wind direction.

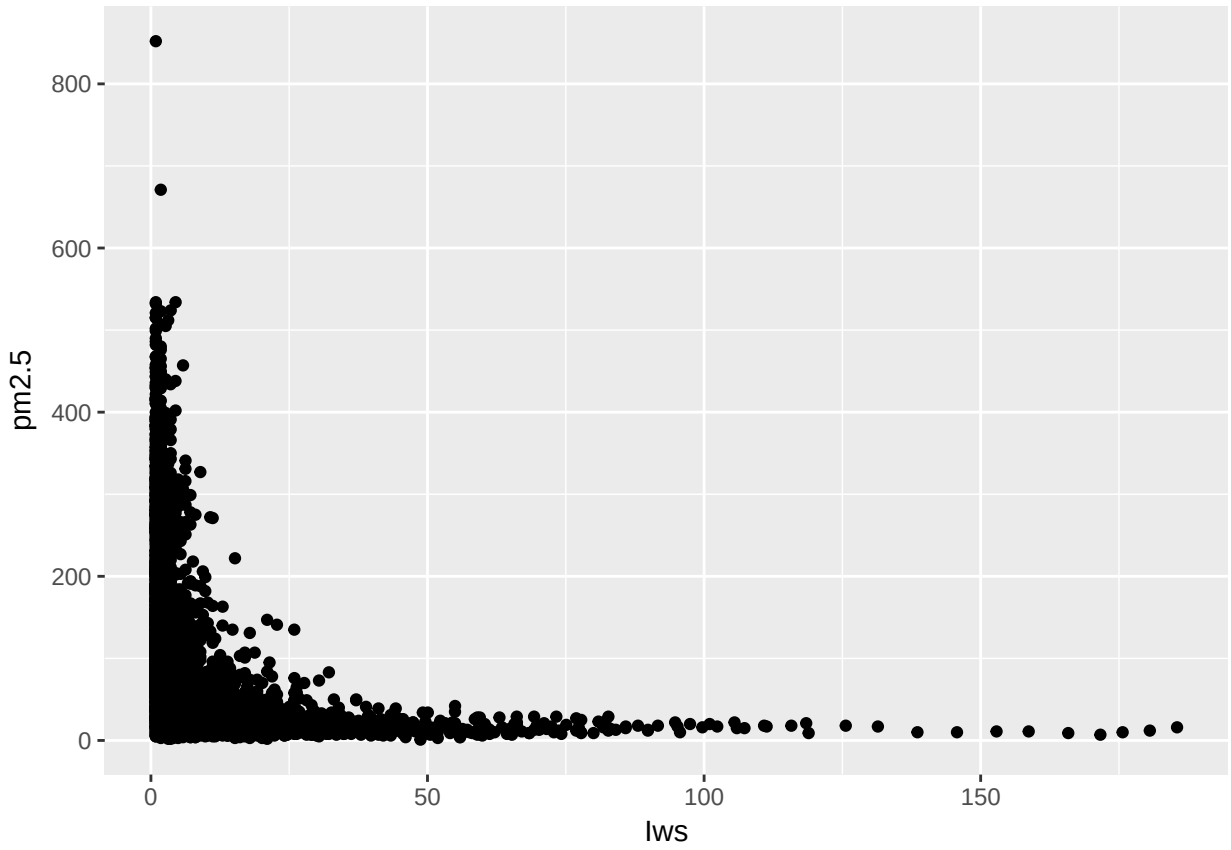
```
#--------plot the cumulated wind speed and the index of pm2.5-----------
#there are only 4 value in this feature:
#cv stands for no wind (the speed is less than 0.5) or no constant direction
#NE, NW, SE is the same as what we say in our daily life
qplot(Iws, pm2.5, data=subset(data_basic, cbwd=='cv'))
```



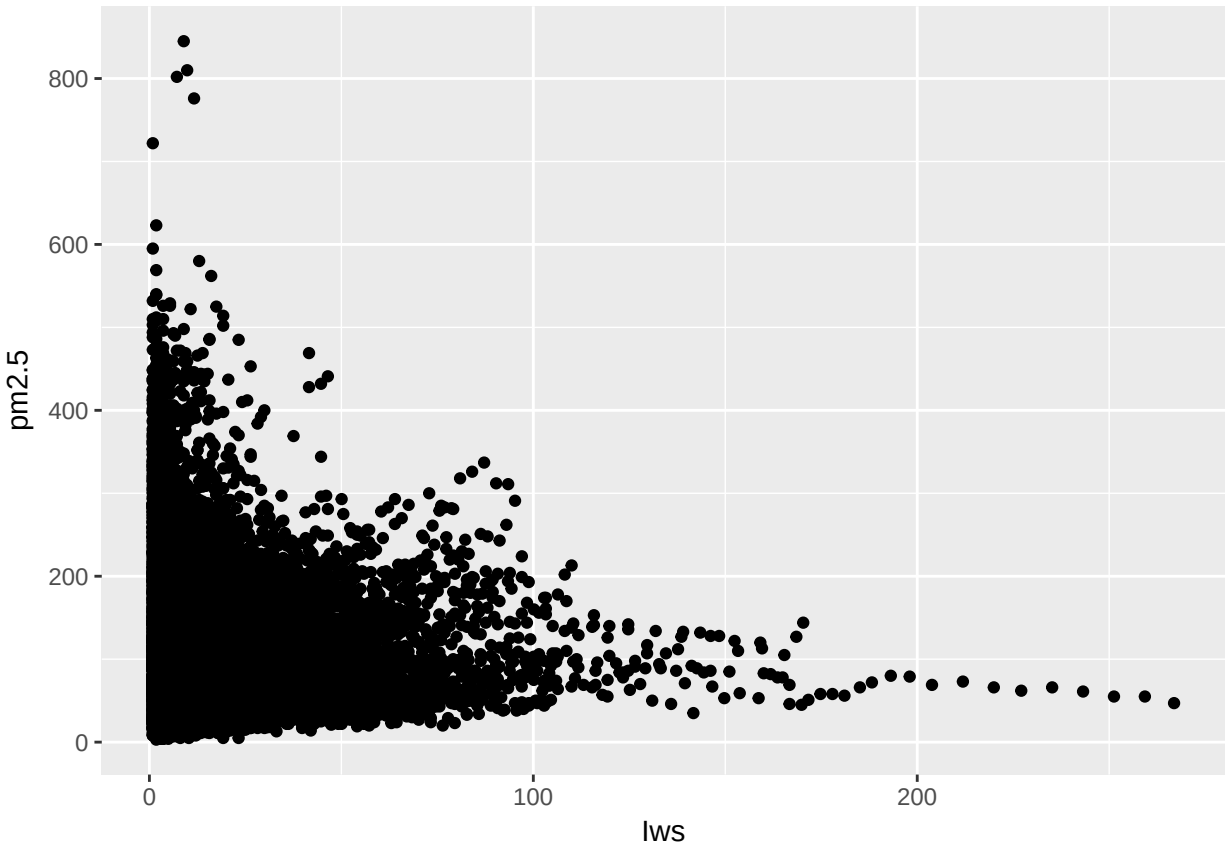

```
qplot(Iws, pm2.5, data=subset(data_basic, cbwd=='NW'))
```




```
qplot(Iws, pm2.5, data=subset(data_basic, cbwd=='NE'))
```




```
qplot(Iws, pm2.5, data=subset(data_basic, cbwd=='SE'))
```



From the plots we can see that the north wind can obviously decrease the index of the pm2.5, but the south wind cannot decrease the concentration of the pm2.5 so obviously. I think that it is due to the industrial distribution of China. There are many industry city lie in the south of Beijing, but there is only little lies in the North of Beijing. The city that has more factory tends to have higher pm2.5 index and the wind will blow the polluted air from one place to another. Although the wind can blow the pm2.5 away, but the south wind may also take some from the industrial city.

Application of the Multivariate Statistical Tools

In this section, I try to apply some of the multivariate statistical tools to the air quality data.

First, I want to find the linear component of the different features that has the largest correlation with the pm2.5 feature. This is possible by applying the CCA. The set that has smaller number of the feature is the pm2.5 feature, the other set is the whole data set except the pm2.5 feature.

```
#---------------Apply the CCA to the data----------------
#---------------Just use the numerical variables---------
cormat <- cor(reg_var)
R12 <- cormat[1,2:7]
R21 <- cormat[2:7,1]
R22 <- cormat[2:7,2:7]

#compute the coefficient of the correlation pair
#in the second set of the variables
R2 <- solve(R22) %*% R21 %*% R12
eigen(R2)
```

```
## $values
## [1]  2.361366e-01 -5.551115e-17  7.459788e-18  3.934158e-18  3.336109e-19
## [6] -2.348877e-19
##
## $vectors
##             [,1]       [,2]       [,3]        [,4]        [,5]        [,6]
## [1,]  0.62426726 -0.6433127 -0.2950879  0.25666352 -0.05530802 -0.02015216
## [2,] -0.74620097 -0.7312931  0.6774688 -0.06860506 -0.01096810 -0.05573112
## [3,] -0.15908968 -0.1559113  0.2209841 -0.75719530 -0.45309690  0.30087412
## [4,] -0.13111964 -0.1285001 -0.5531418  0.42151807 -0.11399367 -0.25325664
## [5,] -0.01783308 -0.0174768  0.0610605 -0.12162033  0.17065486 -0.24075673
## [6,] -0.10322112 -0.1011589  0.3089190 -0.40447926  0.86568256  0.88535867
```

From the eigenvector of the matrix, we find the best factor in for the set of variables except the pm2.5 index that has max correlation with the pm2.5 concentration of contrast the first variable in the second set, the Dew Point, with the other variables, because the Dew Point has the positive correlation with the pm2.5 index, but the remaining variables have negative correlation with the pm2.5 index.

```
#-------------compare the CCA result and the correlation matrix-----
#print the correlation matrix
cormat
```

```
##             pm2.5        DEWP        TEMP        PRES          Iws
## pm2.5  1.00000000  0.17142327 -0.09053400 -0.04728231 -0.247784449
## DEWP   0.17142327  1.00000000  0.82382123 -0.77772212 -0.293105921
## TEMP  -0.09053400  0.82382123  1.00000000 -0.82690281 -0.149612519
## PRES  -0.04728231 -0.77772212 -0.82690281  1.00000000  0.178871492
## Iws   -0.24778445 -0.29310592 -0.14961252  0.17887149  1.000000000
## Is     0.01926558 -0.03492523 -0.09478480  0.07053712  0.022630317
## Ir    -0.05136871  0.12534076  0.04954445 -0.08053221 -0.009156939
##                Is           Ir
## pm2.5  0.019265576 -0.051368706
## DEWP  -0.034925232  0.125340756
## TEMP  -0.094784798  0.049544454
## PRES   0.070537123 -0.080532209
## Iws    0.022630317 -0.009156939
## Is     1.000000000 -0.009763862
## Ir    -0.009763862  1.000000000
```

```
#print the max correlation computed by CCA
R2 <- solve(R22) %*% R21 %*% R12
t <- eigen(R2)
cor(reg_var[,1], as.matrix(reg_var[,-1]) %*% as.matrix(t$vectors[,1]))
```

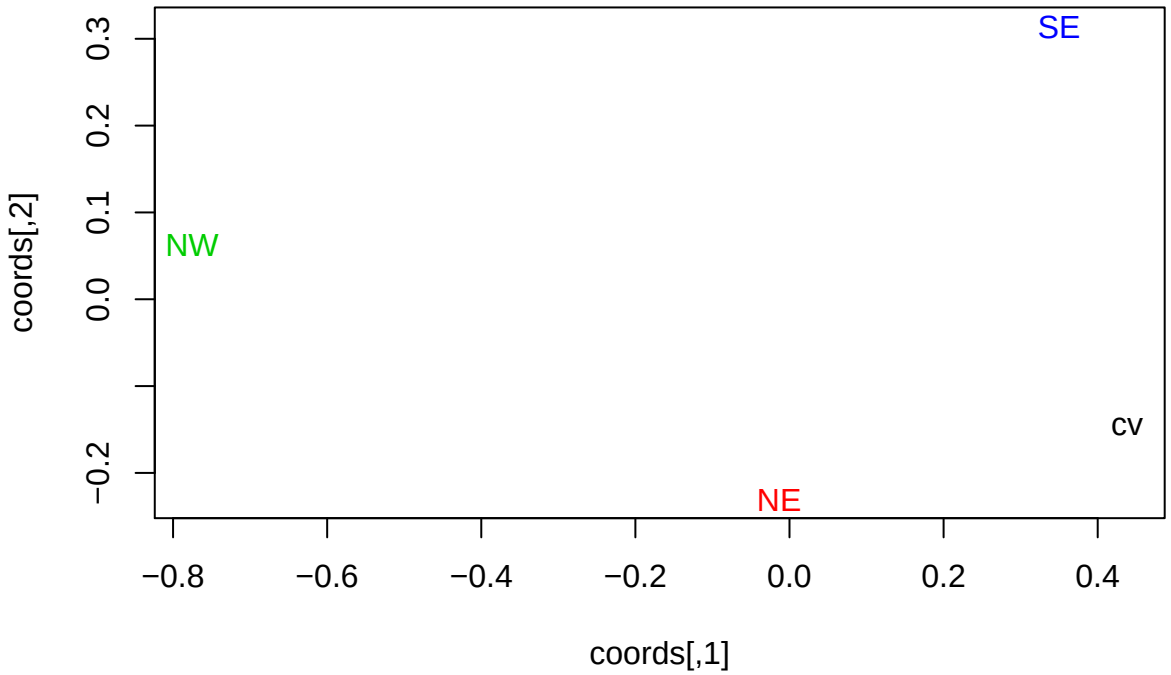
```
##           [,1]
## [1,] 0.4226416
```

From the result, we can know that the CCA can really impove the correlation a lot, but 0.423 is also a relatively small correlation. I think this is due to the fact that pm2.5 is actually caused by many factors, and the factor listed just interpret some of the reason, and the factors listed above can just affect the pm2.5 concentration, but not decide it. Besides, we just omit the direction of the wind, but we can find that the wind direction actually has some influence on the pm2.5 index.

Next we consider the wind direction. We do a MDS to the data, and see whether the wind direction has some influence to the data of the air.

```
#--------apply the MDS to the different wind direction-------
data_new <- data_basic[,6:13]
```

```
data.cv <- subset(data_new, cbwd == 'cv')[,-5]
data.NE <- subset(data_new, cbwd == 'NE')[,-5]
data.NW <- subset(data_new, cbwd == 'NW')[,-5]
data.SE <- subset(data_new, cbwd == 'SE')[,-5]

centers <- matrix(0,4,7)

centers[1,] = apply(data.cv, 2, mean)
centers[2,] = apply(data.NE, 2, mean)
centers[3,] = apply(data.NW, 2, mean)
centers[4,] = apply(data.SE, 2, mean)

S <- cov(reg_var)

mahal <- matrix(0,4,4)
for (i in 1:4) {
  mahal[i,] = mahalanobis(centers, centers[i,], S)
}

coords <- cmdscale(mahal)
plot(coords, type = 'n')
text(coords, labels = c('cv','NE','NW','SE'),
     col = c(1,2,3,4))
```



From the MDS result, we can find that the wind direction really affects the average air data in Beijing, and every 2 wind direction has huge difference for the average air data.

Conclusion

From the descriptive statistics, we can find that the pm2.5 index changed regularly between months and different time in each day. We find that the air quality in summer is better than that in winter, and the

average air quality is best in the afternoon. We also know that the Dew Point, the Temperature, and the cumualated wind speed affect the pm2.5 concentration, and the wind direction will also affect the pm2.5 index.

In the second part, by applying the canonical component analysis to the data, we find the linear combination of the variables that have the largest correlation with the pm2.5 index. The component is largely a contrast of the Dew Point with the Temperature , the Pressure, and the cumulated wind speed. From the MDS result, we can learn that the dirction of the wind affects the air data a lot, and it is easy to distinguish different data centers of different wind direction.

References

[1] Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257.

[2] https://en.wikipedia.org/wiki/Dew_point

[3] The lecture notes and the R-Demos of the course, Multivariate Statistical Analysis