# Investigating the relationship between climate-related measurements and air pollution

***Abstract:*** ***Beijing is one of the cities affected by air pollution. Given the severity of air pollution, it is important to investigate the various predictors of air pollution in hopes of tackling the issue in Beijing. The relationships between the concentration of pm2.5 and year, month, wind direction, temperature, air pressure and wind speed are investigated using various statistical techniques. We conclude that the concentration of pm2.5 varies across most years (except for 2011 and 2014), months, and wind direction, with summer months having a higher concentration of log(pm2.5) than winter months. Cumulated wind speed is the most significant predictor of the concentration of pm2.5 in winter months.***

# Table of Contents

# 1 Introduction

Air pollution is a major issue worldwide. One of the cities that is affected by air pollution is Beijing. With poor air quality, there is increased mortality from diseases such as stroke and lung cancer. In contrast, the lower levels of air pollution will see better respiratory and cardiovascular health in the long-run and short-run. (WHO, 2021) Furthermore, air pollution also causes changes in the climate. (EPA, 2021) In view of the severe consequences of air pollution, this report aims to study the various factors affecting air pollution in Beijing. This is done using the hourly data set of the PM2.5 data of the US Embassy in Beijing and Beijing Capital International Airport. The data set from January 1, 2010, to December 31, 2014, contains information relating to PM2.5 and other variables relating to the weather in China.

We aim to use the dataset to answer the following questions:

1. Does the PM2.5 concentration vary across the years?
2. Does the PM2.5 concentration vary across the months?
3. Does the PM2.5 concentration vary with combined wind direction?
4. Is temperature a significant predictor of PM2.5 concentration?
5. Is pressure a significant predictor of the PM2.5 concentration?
6. Is cumulated wind speed a significant predictor of the PM2.5 concentration?
7. Is there a single continuous variable that is more important than the others in predicting the PM2.5 concentration in the winter months?

In this report, the dataset was examined in relation to the research objectives. This was done through statistical analysis using the R language. Subsequently, conclusions were formed using the appropriate methods.

## 2 Data Description

The dataset, titled “beijingpm”, is obtained from the UCI Machine Learning Repository. The original data consists of 1 csv data frame, titled “PRSA_data_2010.1.1-2014.12.31.csv”. It was originally used in the research paper assessing Beijing’s PM2.5 pollution from the perspectives of the severity, impact on weather and winter heating. (Liang et al., 2015)

Before proceeding to data analysis, we first performed a preliminary data cleaning to ensure that:

- Irrelevant columns are eliminated (i.e. “Is: Cumulated hours of snow”, “Ir: Cumulated hours of rain”, “DEWP: dew point”, “day: day of observation” and “hour: hour of observation”); and

- Entries with “NA” recorded in the PM2.5 columns were excluded.

After all the preparation, 41757 observations (players) with 8 variables are retained for analysis:

1. *No*: row number (1, 2, …, 41757)

2. *month*: Month of data (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)

3. *pm2.5*: PM2.5 concentration ($\mu g^{-3}$)

4. *year*: Year of data (2010, 2011, 2012, 2013, 2014)

5. *TEMP*: Temperature (°C)

6. *PRES*: Pressure (hPa)

7. *cbwd*: Combined wind direction (Northwest (NW), Northeast (NE), Southeast (SE) and calm and variable (CV))

8. Iws: Cumulated wind speed (m/s)

# 3 Description and Cleaning of Dataset

## 3.1 Summary Statistics for Main Variable of Interest, *pm2.5*

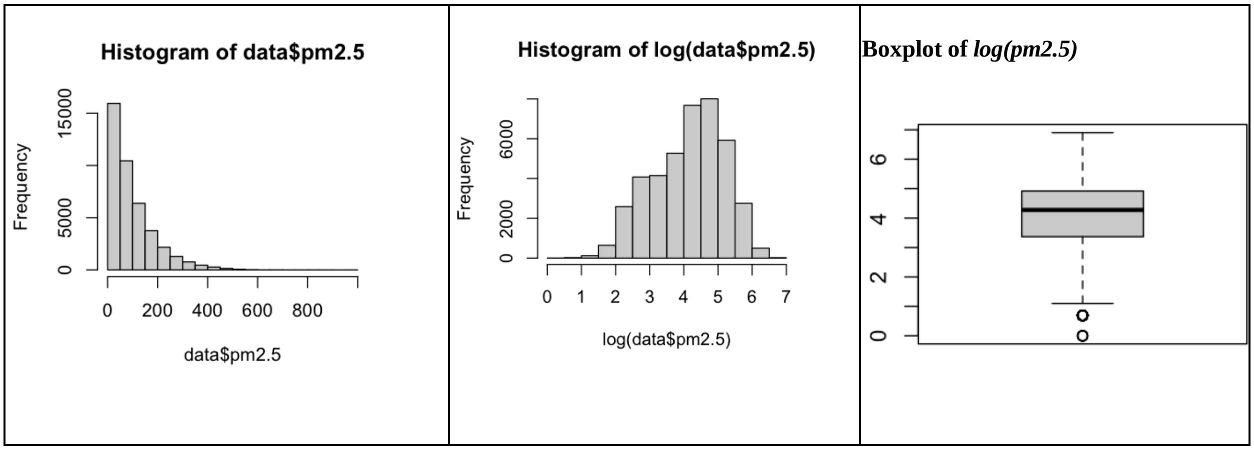
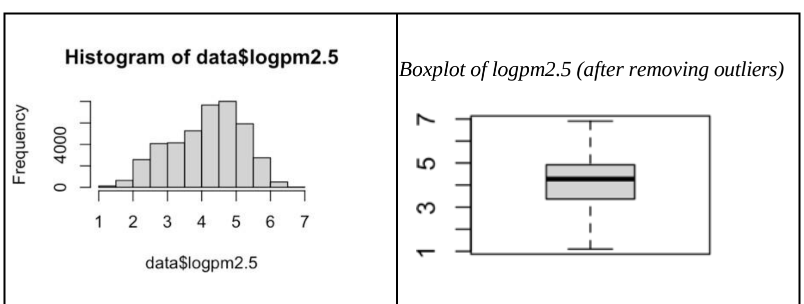



*Figure 3.1: Histogram of pm2.5, Histogram of log(pm2.5) and boxplot of log(pm2.5) showing the overall distribution of pm2.5*

*pm2.5* is a continuous variable. From the histogram of *pm2.5*, we can infer that the histogram is skewed to the right. (Figure 3.1) Hence, we apply a log-transformation (base *e*) to *pm2.5*. The log-transformed data has a distribution that is closer to the normal distribution and therefore, more suitable for analysis. However, the corresponding boxplot shows that there are some outliers. Upon further investigation, we notice that some of the observations recorded a value of zero of *pm2.5*. As it is highly unlikely for *pm2.5* values to be zero, we regard them to be outliers. In response, we remove the outliers, which comprises less than 1% of the data. (i.e.: only 31 observations out of 41757 were removed to get a final dataset of 41726 observations)

The histogram and boxplot of the log-transformed variable, with the outliers removed are shown along with summary statistics in Figure 3.11. The dataset is now more symmetric and does not have any outliers.

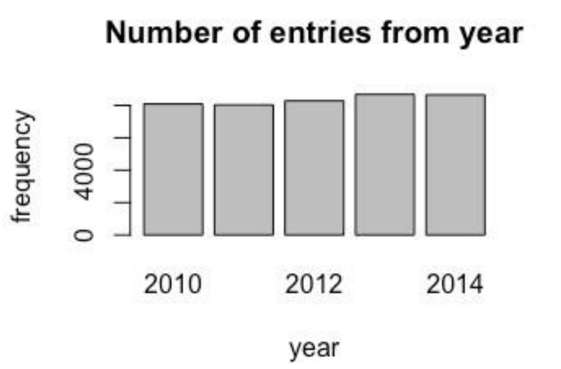



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.099   3.367   4.277   4.138   4.920   6.902
```

*Figure 3.11: Histogram of log(pm2.5), boxplot of log(pm2.5) and summary statistics of log(pm2.5) after removing outliers*

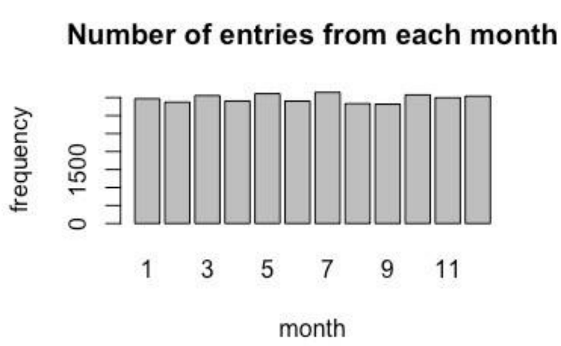
## 3.2 Summary statistics for other variables

### 3.2.1 Year of data, *year*



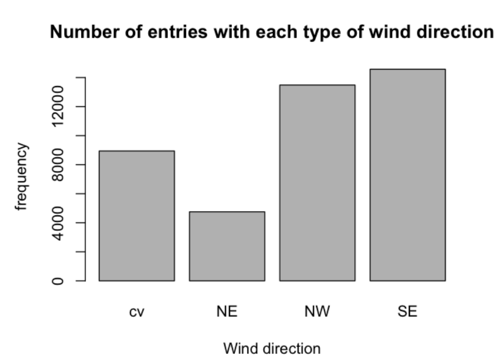

*years* is regarded as a categorical variable which can take 5 values. y*ear=2010* denotes 2010 while *year*=2014 denotes 2014. The dataset contains observations spanning from 2010 to 2014. There are no outliers to be removed. There are also no huge differences in number of observations for each month.

### 3.2.2 Month of data, *month*



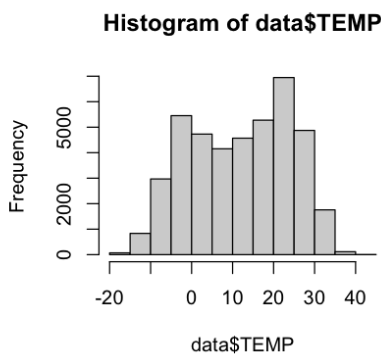
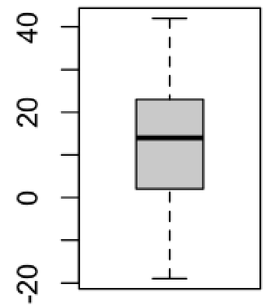

*month* is regarded as a categorical variable which can take 12 values. *month*=1 represents January while *month*=12 represents December. There are no outliers to be removed. There are also no huge differences in number of observations for each month.

### 3.2.3 Combined wind direction, *cbwd*

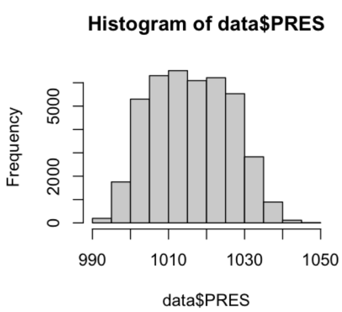
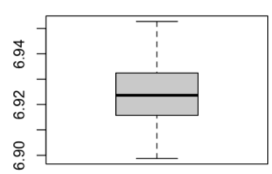



*cbwd* is regarded as a categorical variable which can take 4 factors (cv, NE, NW and SE). These factors are northeast (NE), northwest (NW), southeast (SE) and calm and variable (cv). There are no outliers to be removed. However, there appears to be more observations with wind directions being NW and SE as compared to cv and NE. Without loss of generality, we consider "cv" to be Southwest ("SW").

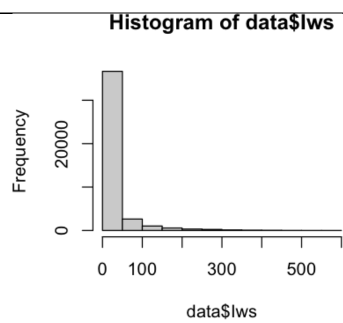
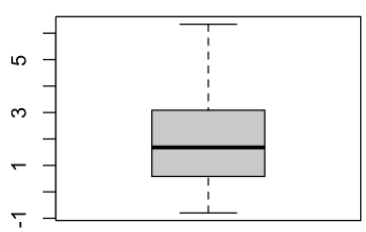
### 3.2.4 Temperature, *TEMP*

| | | |
|---|---|---|
|  |  | Histogram appears to have dual modes. However, we are not able to apply log transformation (base *e*) as there are negative values for *TEMP*. Hence, we consider *TEMP* to not be normally distributed. There does not appear to be any outliers from the boxplot. |

### 3.2.5 Pressure, *PRES*

| | | |
|---|---|---|
|  | Boxplot of *log(PRES)*<br> | Histogram appears to be right-skewed. Thus, we apply log transformation (base *e*). There does not appear to be any outliers from the boxplot. |

### 3.2.6 Cumulated wind speed, *Iws*

| | | |
|---|---|---|
|  | Boxplot of *log(Iws)*<br> | Histogram appears to be right skewed. Thus, we apply log transformation (base *e*). There does not appear to be any outliers from the boxplot. |

## 3.3 Final Dataset for Analysis

Based on the above analysis, the dataset is further reduced to 41726 observations with log-transformation (base *e*) being applied to *Iws*, *PRES and pm2.5*.

# 4. Statistical Analysis

## 4.1 Correlation between *log(pm2.5)* and other Continuous Variables

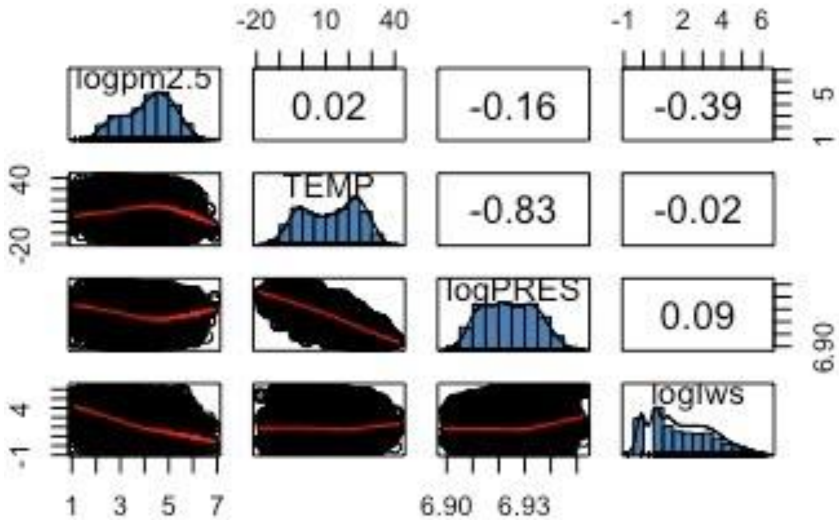



*Figure 4.1: Correlation matrix for log(pm2.5) and other continuous variables (i.e.:TEMP, log(PRES), log(Iws))*

The scatter plots and correlation coefficients in Figure 4.1 are used to study the possible linear relationship between the level of pm2.5 in the air and other climate indicators. The plots suggest that *log(pm2.5)* is more highly correlated to *log(Iws)* (r = -0.39) than to *TEMP* (r = 0.02) and *log(PRES)* (r = -0.16). In addition, *log(pm2.5)* appears to be negatively correlated with *log(PRES)* and *log(IWS).*

Furthermore, there are other notable observations which can be deduced from the plots. These include:

- *TEMP* and *log(PRES)* are quite highly negatively correlated (r = -0.83)
- *log(PRES)* and *log(Iws)* are positively correlated. However, the correlation is weak. (r = 0.09)

Further statistical tests will be performed in the subsequent sections to confirm these observations.

## 4.2 Statistical Tests

### 4.2.1 Relation between *pm2.5* and *year*

In this section, we determine if the concentration of PM2.5 varies across the years. As *year* is a categorical variable, the ANOVA test will be used to assess whether *log(pm2.5)* varies with *year*.

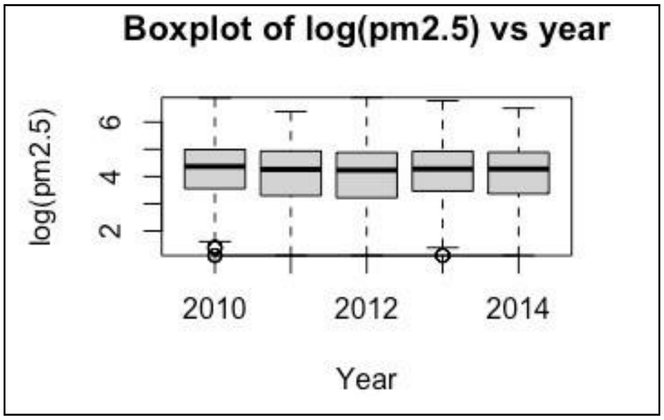



*Figure 4.2.11: Boxplot of log(pm2.5) vs year illustrating the distribution of log(pm2.5) across the different years*

The boxplot (Figure 4.2.11) suggests that the spread of *log(pm2.5)* is similar for all 5 years (factor levels). Hence, the ANOVA test can be used to test for the equality of mean *log(pm2.5)* values for the different years.

The ANOVA test is conducted in the following manner:

$H_0$: $\mu_{2010} = \mu_{2011} = \mu_{2012} = \mu_{2013} = \mu_{2014}$ vs $H_1$: not all $\mu_i$'s are equal

where $\mu_i$: mean *log(pm2.5)* of year *i* (*i* = 2010, 2011, 2012, 2013, 2014)

```
                     Df Sum Sq Mean Sq F value Pr(>F)    
factor(data$year)     4    152   37.94    35.9 <2e-16 ***
Residuals         41721  44087    1.06                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 4.2.412 : Output of ANOVA test*

It is evident that the ANOVA test returns a value of p-value of less than $2e^{-16}$. Hence, we reject $H_0$ and conclude that not all the mean *log(pm2.5)* values across the different years are equal at the 0.001 significance level. Thus, we conclude that the PM2.5 varies across the years.

As the ANOVA test does not show us the detailed comparisons between the mean *log(pm2.5)* values of the different years, we perform pairwise comparisons between the years to derive greater insight.

```
        Pairwise comparisons using t tests with pooled SD

data:  data$logpm2.5 and data$year

     2010    2011    2012    2013
2011 2.7e-11 -       -       -
2012 < 2e-16 3.3e-06 -       -
2013 2.1e-06 0.04132 1.0e-11 -
2014 < 2e-16 0.14873 0.00101 0.00038

P value adjustment method: none
```

*Figure 4.2.13 : Pairwise comparisons between mean log(pm2.5) values of the different years*

We use the p-values derived from the pairwise comparisons to make our conclusions at the 0.05 significance level. It was found that only the mean *log(pm2.5)* values of 2011 and 2014 were the same. Hence, we conclude that the mean concentration of PM2.5 varied for most years but were the same for years 2011 and 2014.

### 4.2.2 Relation between *pm2.5* and *month*

In this section, we determine if the concentration of PM2.5 varies with month.

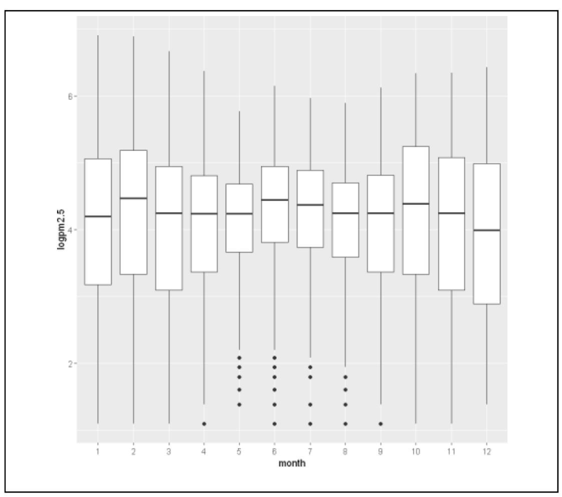



*Figure 4.2.21: boxplot of log(pm2.5) vs month*

Figure 4.2.21 shows the boxplot of log(pm2.5) vs month. From the boxplot, the median value of *log(pm2.5)* appears to be higher in the summer months (*month* = 6,7,8) than that in the winter months (*month* = 12,1,2). Moreover, considering that there are many outliers in *month = 5,6,7,8*, we assume that the spread of the *log(pm2.5)* values are the same across the months. Thus, to further investigate if *log(pm2.5)* varies across the months, we conduct the Analysis of Variance (ANOVA) test as *month* is a categorical variable.

The test is conducted in the following manner:

$H_0$: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_{12}$ vs $H_1$: not all $\mu_i$'s are equal

where $\mu_i$: mean value of *log(pm2.5)* in month *i (i = 1,2,3,4,…,12)*

```
                     Df Sum Sq Mean Sq F value Pr(>F)
factor(data$month)    11    394   35.82   34.08 <2e-16 ***
Residuals          41714  43844    1.05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Figure 4.2.22: Output of ANOVA test*

The ANOVA test returns a value of p-value of less than $2e^{-16}$. Hence, we reject $H_0$ and conclude that not all the mean *log(pm2.5)* values are the same at the 0.001 significance level. Additionally, we also do some further analysis to investigate if the *log(pm2.5)* values are higher in Summer than in Winter. This is done in the following manner:

$H_0$: $\mu_{Summer} = \mu_{Winter}$ vs $H_1$: $\mu_{Summer} > \mu_{Winter}$

First, the F test is done to check if the variances of *log(pm2.5)* are the same for Summer and Winter. As the variances are different, the two-sample t-test is conducted.

```
        F test to compare two variances

data:  summerdata$logpm2.5 and winterdata$logpm2.5
F = 0.53065, num df = 10381, denom df = 10383, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5106255 0.5514677
sample estimates:
ratio of variances
         0.5306538


        Welch Two Sample t-test

data:  summerdata$logpm2.5 and winterdata$logpm2.5
t = 5.3565, df = 18981, p-value = 4.291e-08
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.05222926        Inf
sample estimates:
mean of x mean of y
 4.210480  4.135103
```

*Figure 4.2.22: Output of T test*

As the p-value = $4.291e^{-08}$ is less than 5% significance level, we reject $H_0$ and conclude that mean values of *log(pm2.5)* are greater in Summer than those in Winter. Overall, we conclude that the PM2.5 concentration varies across the months.

### 4.2.3 Relation between *pm2.5* and *cbwd*

In this section, we determine if the concentration of PM2.5 varies with wind direction.

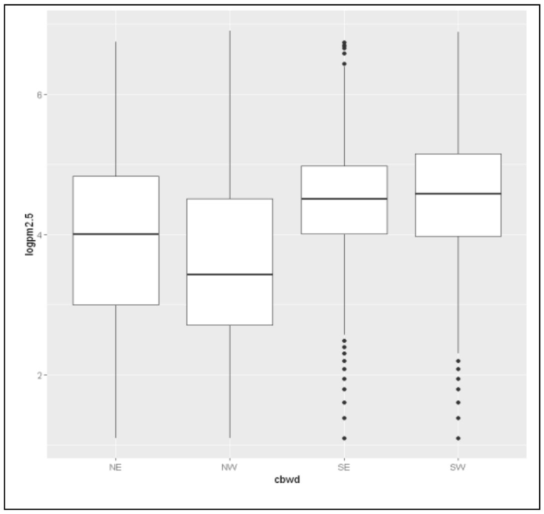



*Figure 4.2.31: Boxplot of pm2.5 across different wind directions*

Figure 4.2.31 reveals that the median concentration of PM2.5 is higher for Southeast and Southwest wind directions. However, the median concentration of PM2.5 is lower for Northeast and Northwest wind directions. As there appears to be many outliers for Southeast and Southwest direction, we assume that the spread of *log(pm2.5)* for all wind directions are the same. Given that *cbwd* is a categorical variable, we proceed to perform the ANOVA test to assess if the mean *log(pm2.5)* concentrations are the same across all wind directions.

The ANOVA test is conducted in the following manner:

$H_0$: $\mu_{NE} = \mu_{\mathrm{NW}} = \mu_{\mathrm{SE}} = \mu_{\mathrm{SW}}$ vs $H_1$: not all $\mu_i$'s are equal.

where $\mu_i$ denotes mean *log(pm2.5)* concentrations of the wind direction *i*. (*i = NE, NW, SE, SW)*

```
                      Df Sum Sq Mean Sq F value Pr(>F)
factor(data$month)    11    394   35.82   34.08 <2e-16 ***
Residuals          41714  43844    1.05
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Figure 4.2.32: Output of ANOVA test*

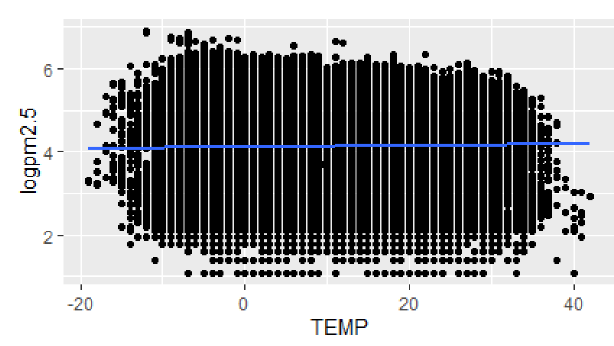
Since the p-value is less than $2e^{-16}$, we can reject $H_0$ at a 0.001 significance level. Thus, we observe that the mean values of *log(pm2.5)* varies across *cbwd*. Hence, we conclude that the concentration of PM2.5 varies across the different wind directions.

### 4.2.4 Relation between *pm2.5* and *TEMP*

In this section we want to find out if the concentration of PM2.5 depends on the temperature. We use a simple linear regression to determine the relationship between *log(pm2.5)* and *TEMP*.

The p-value of *TEMP* is $4.6e^{-05}$. This indicates that the relationship between *log(pm2.5)* and *TEMP* is statistically significant at 0.001 level of significance. However, the R-squared of this test is 0.0003979. This is close to zero which concurs with the findings in Section 4.1 that shows a weak linear correlation (r = 0.02) between these two variables.

Despite its statistical significance, we observe that *TEMP* can only explain about 0.04% of variation in *log(pm2.5).* Thus, temperature appears to not be a practically significant predictor of PM2.5 concentration.




```
Call:
lm(formula = data$logpm2.5 ~ data$TEMP)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0862 -0.7684  0.1429  0.7795  2.8046 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.1173459  0.0071927 572.432  < 2e-16 ***
data$TEMP   0.0016869  0.0004139   4.076  4.6e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.029 on 41724 degrees of freedom
Multiple R-squared:  0.0003979,	Adjusted R-squared:  0.00037
4 
F-statistic: 16.61 on 1 and 41724 DF,  p-value: 4.6e-05
```

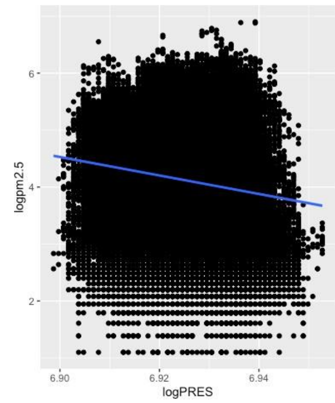
*Figure 4.2.4: Scatter plot of log(pm2.5) against TEMP (left) and results of linear regression (right)*

### 4.2.5 Relation between PM2.5 and *PRES*

In this section, we determine if the concentration of PM2.5 depends on pressure. We perform a simple linear regression between *log(PRES)* and *log(pm2.5).*

The regression model provides a p-value of $< 2e^{-16}$, which is lesser than 0.001. This implies a statistically significant relationship between *log(pm2.5)* and *log(PRES)* at a significance level of 0.001. The R-squared value for this model is 0.0255 which implies that *log(PRES)* only explains about 2.55% of the variation in *log(pm2.5).* This concurs with what was observed in Section 4.1 where *log(pm2.5)* and *log(PRES)* were weakly linearly correlated. (r = -0.16).

Hence, despite its statistical significance, we conclude that *log(PRES)* is not practically significant. Thus, we conclude that the pressure is not a significant predictor of PM2.5 concentration.



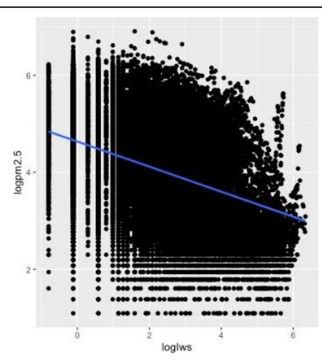

```
Call:
lm(formula = data$logpm2.5 ~ data$logPRES)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3685 -0.7528  0.1162  0.7420  3.0108

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  116.5072     3.4007   34.26   <2e-16 ***
data$logPRES -16.2289     0.4911  -33.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.016 on 41724 degrees of freedom
Multiple R-squared:  0.0255,    Adjusted R-squared:  0.02548
F-statistic:  1092 on 1 and 41724 DF,  p-value: < 2.2e-16
```

*Figure 4.2.5: Scatter plot of log(pm2.5) against log(PRES) (left) and results of linear regression (right)*

### 4.2.6 Relation between *pm2.5* and *Iws*

Next, we determine if the concentration of PM2.5 depends on cumulated wind speed. We perform a simple linear regression between *log(Iws)* and *log(pm2.5).*

The regression model provides a p-value of less than $2e^{-16}$ at a significance level of 0.001. Hence, *log(Iws)* is statistically significant. The R-squared for this model is 0.1524, which indicates that *log(Iws)* explains approximately 15.24% of the variation in *log(pm2.5).* This concurs with the finding in Section 4.1 where *log(pm2.5)* and *log(Iws)* were fairly linearly correlated. (r = -0.39). Thus, *log(Iws)* is not only statistically significant but it is also practically significant.




```
Call:
lm(formula = data$logpm2.5 ~ data$logIws)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5634 -0.6687  0.0741  0.7071  2.7852

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.631953   0.007349  630.25   <2e-16 ***
data$logIws -0.257864   0.002977  -86.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.948 on 41724 degrees of freedom
Multiple R-squared:  0.1524,    Adjusted R-squared:  0.1524
F-statistic:  7505 on 1 and 41724 DF,  p-value: < 2.2e-16
```

*Figure 4.2.6: Scatter plot of log(pm2.5) against log(Iws) (left) and results of linear regression (right)*

### 4.2.7 The single most important continuous predictor of PM2.5 in winter

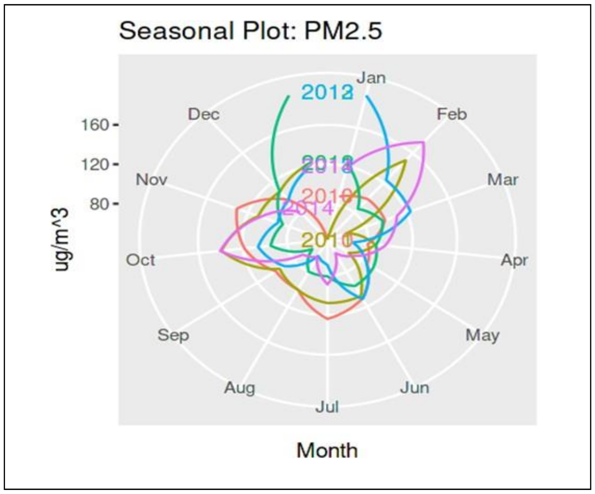


*Figure 4.2.7: seasonal plot of PM2.5 with respect to month*

Figure 4.2.7 shows the concentration of PM2.5 is extremely high in some winter months. Sections 4.2.4 and 4.2.5 shows that *TEMP* and *log(PRES)* may not be practically significant despite being statistically significant. However, Section 4.2.6 shows that *log(Iws)* might be both practically and statistically significant. These results were obtained without consideration of the different months. As there might be doubts about the significance of the predictors in the different months, this section seeks to investigate which factor affects PM2.5 most in winter months. We now use simple linear regression to determine the most important predictor.

$$\boldsymbol{log(pm2.5) = \beta_0 + \beta_1 * X + \varepsilon}$$

where *X* can be any one of *TEMP*, *log(Iws)* or *log(Pres)*

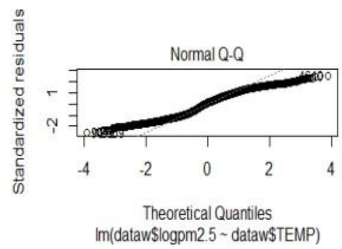
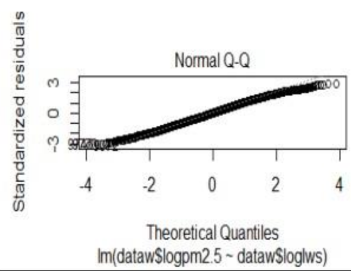
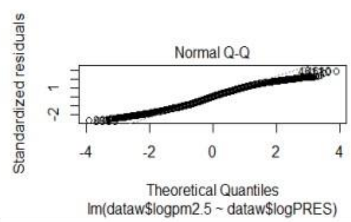
| Variable | Fitted model | P-value | R-squared | QQ-plot of residuals |
|---|---|---|---|---|
| TEMP | Y=4.167+0.012X | 2.6e-07 | 0.002552 |  |
| log(lws) | Y=4.934-0.388X | <2e-16 | 0.3167 |  |
| log(Pres ) | Y=469.224-67.069X | <2e-16 | 0.1407 |  |

*Table 4.2.7: Results of linear regression*

By comparing the R-squared and the residual plot (Table 4.2.7), *log(lws)* can be concluded to be the most important predictor to model *log(pm2.5)* when a simple linear model is used.

# 5 Conclusion and Discussion

Air pollution is one of the most notorious causes of respiratory and cardiovascular diseases, and changes in climate. (WHO, 2021; EPA, 2021) Seeing that PM2.5 is one of the main air pollutants, it is key for relevant authorities to keep track of the climate-related variables that can predict PM2.5 concentration. The main variables are namely temperature, pressure, cumulated wind speed, months, years, and wind direction. In this report, we attempt to answer some of the important questions regarding the relationship between PM2.5 concentration and these aforementioned factors.

We conclude that:

1. The mean PM2.5 concentration varied across most years except for years 2011 and 2014.
2. PM2.5 concentration varies across the months, with mean *log(pm2.5)* concentration being higher in the summer months than in the winter months.
3. The mean PM2.5 concentration varies with combined wind direction.
4. Temperature is not a significant predictor of PM2.5 concentration.
5. Pressure is not a significant predictor of PM2.5 concentration.
6. Cumulated wind speed is a significant predictor of PM2.5 concentration.
7. Cumulated wind speed is the most important predictor that affects the PM2.5 concentration in the winter months.

Although the results of this report might be interesting, it is important to note that this report is only based on data from the years 2010 to 2014, and only from the city of Beijing. Thus, the findings might not be applicable to recent years or to different cities. More in-depth analysis using weather data from recent years and different cities may be needed to better ascertain the relationship between climate-related variables and PM2.5 concentration.

**Appendix: R Code**

```
data_raw <- read.csv("/Users/Hello/Downloads/beijingpm.csv")
summary(data_raw)
str(data_raw)
library(dplyr)
library(ggplot2)
#43824 variables of 13 variables
dataint <- select(data_raw, c(month, cbwd, pm2.5, year, TEMP, PRES, Iws, No))
#select 8 variables + 1 variable ('no' column included for reference),
#others for data analysis
data <- na.omit(dataint) #removing all "NA"
str(data)
summary(data)
##
hist(data$pm2.5)
boxplot(data$pm2.5)
hist(log(data$pm2.5))
boxplot(log(data$pm2.5))
summary(log(data$pm2.5))
data <- mutate(data, logpm2.5 = log(pm2.5))
Q1logpm2.5 <- quantile(data$logpm2.5, .25)
Q3logpm2.5 <- quantile(data$logpm2.5, .75)
IQR <- IQR(data$logpm2.5)
data <- subset(data, data$logpm2.5 > (Q1logpm2.5-1.5*IQR) & data$logpm2.5 <
(Q3logpm2.5+1.5*IQR))
hist(log(data$pm2.5))
boxplot(log(data$pm2.5))
summary(log(data$pm2.5))
```

```
## month

plot(as.factor(data$month), main="Number of entries from each month", xlab="month",
ylab="frequency")

### year

plot(as.factor(data$year), main="Number of entries from year", xlab="year", ylab="frequency")

### cbwd

plot(as.factor(data$cbwd), main="Number of entries with each type of wind direction",
xlab="Wind direction", ylab="frequency")

## temp

hist(data$TEMP)

boxplot(data$TEMP)

## pres

hist(data$PRES)

boxplot(data$PRES)

hist(log(data$PRES))

boxplot(log(data$PRES))

data <- mutate(data, logPRES = log(PRES))

## Iws

hist(data$Iws)

boxplot(data$Iws)

hist(log(data$Iws))

boxplot(log(data$Iws))

data <- mutate(data, logIws = log(Iws))

##corr plot

data_continuous <- data[,c(9,5,10,11)]

install.packages("psych")

library(psych)

pairs.panels(data_continuous, method = "pearson", # correlation method

        hist.col = "steelblue",

        pch = 21,
```

```
            density = TRUE,
            ellipses = FALSE)

#corr without importing new library
variable_list <- c("month", "DEWP", "TEMP", "PRES","Iws")
for(i in colnames(data)[3:ncol(data)-2]) {
   #print(length(data[,variable_list[i]]))
   COR=cor.test(data[,"pm2.5"],data[,i],method="pearson")}

##
##relationship between pm2.5 and year
boxplot(data$logpm2.5~data$year, ylab="log(pm2.5)", xlab="Year", main="Boxplot of
log(pm2.5) vs year")
summary(aov(data$logpm2.5~factor(data$year)))
pairwise.t.test(data$logpm2.5, data$year, p.adjust.method = "none")
##relationship between pm2.5 and month
#boxplot
print(ggplot(data, aes(factor(data$month), logpm2.5)) +
      geom_boxplot()+ xlab("month")
     )
summary(aov(data$logpm2.5~factor(data$month)))
#t-test between winter and summer
summerdata = data[data$month %in% c(6,7,8),]
winterdata = data[data$month %in% c(12, 1, 2),]
var.test(summerdata$logpm2.5,winterdata$logpm2.5)
t.test(summerdata$logpm2.5,winterdata$logpm2.5,alt='greater',var.equal = FALSE)

##relationship between pm2.5 and cbwd
#pre-process data
# change "cv"" in cbwd to "SW"
```

```
levels(data$cbwd)[1] <- "SW"
# sort it to NE, NW, SE, SW
data$cbwd <- factor(data$cbwd, levels = c("NE", "NW", "SE", "SW"))
summary(data$cbwd)

#visualization boxplot
print(ggplot(data, aes(data$cbwd, logpm2.5)) +
    geom_boxplot() +
    xlab("cbwd"))
summary(aov(data$logpm2.5~factor(data$cbwd)))

#relationship between pm2.5 and TEMP
modelT<-lm(data$logpm2.5~data$TEMP)
summary(modelT)
ggplot(data,aes(x=TEMP,y=logpm2.5))+geom_point()+geom_smooth(method="lm")

##
#relationship between pm2.5 and PRES
modelp<-lm(data$logpm2.5~ data$logPRES)
summary(modelp)
plot(modelp)
ggplot(data,aes(x=logIws,y=logPRES))+geom_point()+geom_smooth(method="lm")



##
#relationship between pm2.5 and Iws
modelws<-lm(data$logpm2.5~ data$logIws)
summary(modelws)
plot(modelws)
ggplot(data,aes(x=logIws,y=logpm2.5))+geom_point()+geom_smooth(method="lm")
```

```
##
#which factor affects pm2.5 most in winter
#seasonal plot
library(xts)
library(seasonal)
#pre-process data
data <- mutate(data,dates=paste(year,month,day,sep="-"))

#covert to xct type
dataxts <- as.xts(data$pm2.5, order.by = as.POSIXct(data$dates))
#calculate the mean of pm2,5 for every month
datamonth <- apply.monthly(dataxts, FUN = "mean")

#time series
inds <- seq(as.Date("2010-1-1"), as.Date("2014-12-31"), by = "day")
datamonthts <- ts(as.numeric(datamonth),
        frequency = 12,
        start=c(2010, as.numeric(format(inds[1], "%j"))))
p2 <- ggseasonplot(datamonthts,
            polar = T,
            year.labels = T,
            year.labels.left = T) +
  ggtitle("Seasonal Plot: PM2.5") +
  ylab('ug/m^3')

#find the most important factor
dataw<-data%>%filter(month==1|month==2|month==12)
modelT1<-lm(dataw$logpm2.5~dataw$TEMP)
summary(modelT1)
```

```
plot(modelT1)
modelws1<-lm(dataw$logpm2.5~ dataw$logIws)
summary(modelws1)
plot(modelws1)
modelp1<-lm(dataw$logpm2.5~ dataw$logPRES)
summary(modelp1)
plot(modelp1)
```

## References


Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., & Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *471*(2182), 20150257. https://doi.org/10.1098/rspa.2015.0257

United States Environment Protection Agency (US EPA). (2021, July 29). *Air Quality and Climate Change Research*. US EPA. Retrieved 25 March 2021, from https://www.epa.gov/air-research/air-quality-and-climate-change-research#:%7E:text=Emissions%20of%20pollutants%20into%20the,cooling%20effects%20on%20the%20climate

World Health Organization. (2021, September 22). *Ambient (outdoor) air pollution*. WHO | World Health Organization. Retrieved 25 March 2021, from https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health