Introduction

Summary Statistics about PM2.5 data

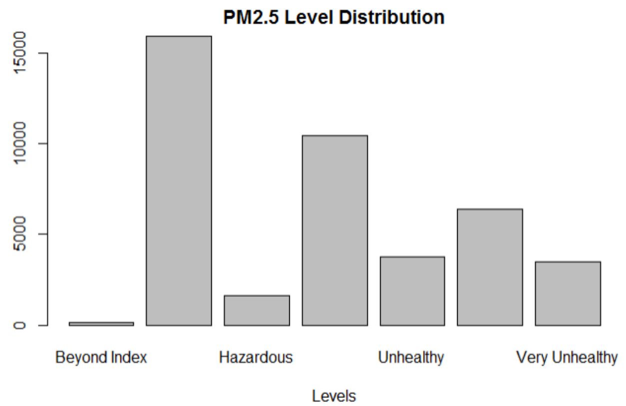
Firstly, we want to provide a basic set of descriptive statistics for the pm2.5 data. There is a total count of 2067 missing values in the data, which we filtered for. We ran a quick summary for the data, shown below:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   29.00   72.00   98.61  137.00  994.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-40.000 -10.000   2.000   1.817  15.000  28.000
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -19.00    2.00   14.00   12.45   23.00   42.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    991    1008    1016    1016    1025    1046
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.45    1.79    5.37   23.89   21.91  585.60

   cv    NE    NW    SE
 9387  4997 14150 15290
```

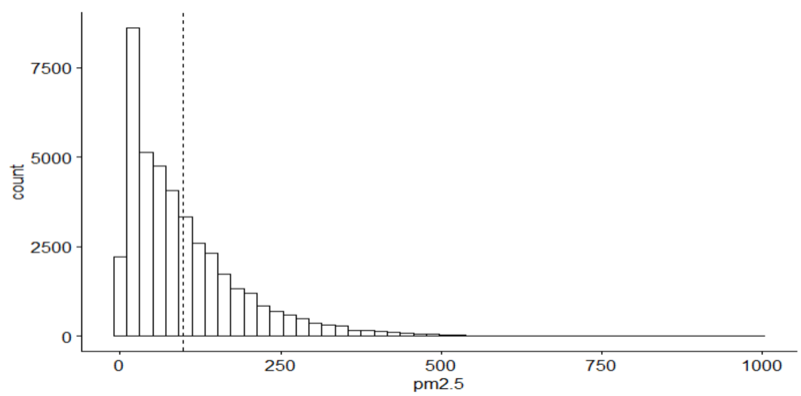
According to the World Health Organization's findings, the highest PM2.5 level for acceptable air quality is 25 micrograms per cubic meter. Our data set shows that the median PM2.5 level in Beijing registers approximately 3 times the recommended level. We used the air quality guide provided by the US government to separate the PM2.5 values into the following levels:

Air Quality Index	PM2.5 Level (microgram per cubic meter)
Good	0 - 50
Moderate	51-100
Unhealthy for Sensitive Groups	101-150
Unhealthy	151-200
Very Unhealthy	201-300
Hazardous	301-500
Beyond Index	>500




```
 Beyond Index                          Good                         Hazardous
          125                         15945                              1634
     Moderate     Unhealthy Unhealthy for Sensitive Groups
        10445          3763                              6376
Very Unhealthy
         3469
```

From the data exploring conducted, we found that, approximately, only 63% of the registered PM2.5 levels in Beijing registered in the "Good" or "Moderate" category. Approximately 12.5% of all levels were registered in the "Very Unhealthy / Hazardous / Beyond Index" categories. We can also see from the graph below that the distribution of PM2.5 levels has a strong positive skew.



We also wanted to examine the cluster months in each of the four years to see how PM2.5 level distribution changed ie. establish a pattern according to the period of the year. We summarized the following about the clustered months:

Month	Proportion of "Very Unhealthy / Hazardous / Beyond Index" Category	Summary
January	17.9%	Higher than total average
February	21.7%	Higher than total average
March	15.0%	Higher than total average
April	5.7%	Lower than total average
May	4.0%	Lower than total average (LOWEST)

June	7.0%	Lower than total average
July	8.3%	Lower than total average
August	4.0%	Lower than total average (LOWEST)
September	7.8%	Lower than total average
October	23.0%	Higher than total average (HIGHEST)
November	17.9%	Higher than total average
December	17.6%	Higher than total average

It is clear that there is a seasonal distribution in PM2.5 levels, as the winter/autumn months tend to have a much higher average PM2.5 level than the summer/spring months. For 6 months continuously, the air quality, on average, falls into the "Very Unhealthy / Hazardous / Beyond Index" category over 15% of the time. We summarized the following about the 24-hour period:

Month	Proportion of "Very Unhealthy / Hazardous / Beyond Index" Category	Summary
Hour 0	16.3%	Higher than average (HIGHEST)
Hour 1	16.3%	Higher than average (HIGHEST)
Hour 2	16.3%	Higher than average (HIGHEST)
Hour 3	15.3%	Higher than average
Hour 4	13.6%	Higher than average
Hour 5	12.3%	Lower than average
Hour 6	11.8%	Lower than average
Hour 7	11.2%	Lower than average
Hour 8	11.1%	Lower than average
Hour 9	10.6%	Lower than average
Hour 10	10.6%	Lower than average
Hour 11	10.5%	Lower than average
Hour 12	10.5%	Lower than average
Hour 13	10.4%	Lower than average
Hour 14	9.7%	Lower than average (LOWEST)
Hour 15	9.7%	Lower than average (LOWEST)
Hour 16	9.8%	Lower than average

Hour 17	10.2%	Lower than average
Hour 18	11.1%	Lower than average
Hour 19	12.6%	Higher than average
Hour 20	14.1%	Higher than average
Hour 21	15.0%	Higher than average
Hour 22	15.6%	Higher than average
Hour 23	16.0%	Higher than average

Similar to the seasonal distribution, it seems that there is a 24-hour distribution of PM2.5 levels too. The levels tend to be higher in the night, and lower in the day. One take note of how there is a smaller variation in results than the seasonal distribution.
Overall, these statistics have given us a better understanding of just how severe the PM2.5 levels are in Beijing.

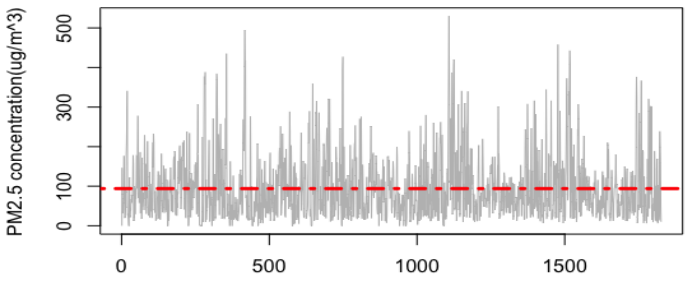
Time series analysis section

This section is mainly about performing time series analysis on average PM 2.5 daily records and monthly records respectively. Our target in this section is to find out the patterns and features of the PM 2.5 data under different frequencies.

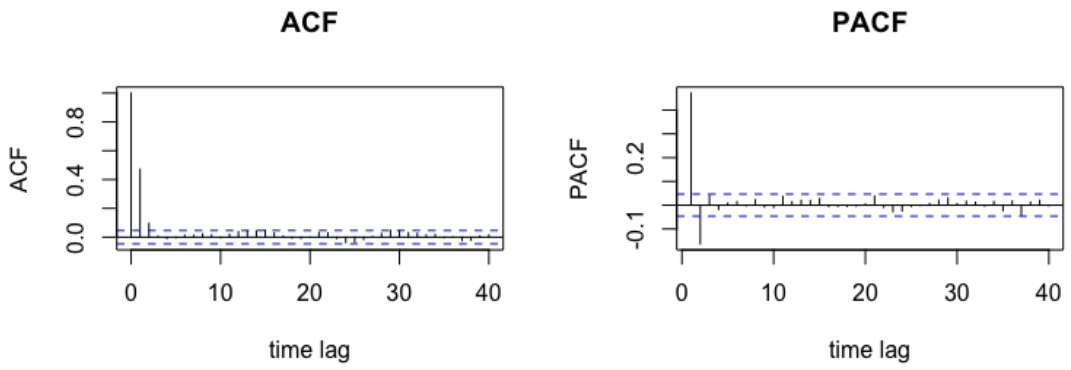
After cleaning the data, we average the PM 2.5 data within 24 hours per day and average the PM 2.5 data in each month respectively. First we analyze the daily PM 2.5 data.

Part1: Daily PM 2.5 data

Before doing any further analysis, we first plot the daily data of PM 2.5 from 01/01/2010 through 31/12/2014：



The time series plot shows that changing variability exists in the daily PM 2.5 data. To remove the changing variability, we performed logarithm on our data and then ploted the ACF and PACF of the data:



Above plots show that both the ACF(autocorrelation function) and PACF(partial autocorrelation function) of daily PM 2.5 records have quick decays so we do not need to do differencing on the data. Also note that in the ACF plot(left), the values of ACF are outside the bounds at lag 1 and lag 2, which suggests that we could fit a MA(2) (moving average 2) model on the records.

Here we used function ***auto.arima( )*** in R package *forecast* to conduct the model fitting:

```
fit_daily = auto.arima(ts(pm_daily));fit_daily$coef
```

```
##       ma1       ma2 intercept
## 0.5553144 0.1135664 4.2399056
```

To evaluate the quality of this fitted model, we need to perform several hypothesis tests on its fitted residuals. If the fitted model is a good fit, we should expect that the residuals behave like white noise.

Test of randomness:

We used Box-Pierce test to test the randomness of the residuals of daily PM 2.5 data.

Box-Pierce test:

1. H_0: the data are independently distributed.
2. Under H_0, the test statistic is $Q = n\sum_{h=1}^{k}\hat{\rho}\,(h) \sim approx.\chi_k^2$.
3. Reject H_0 if $Q > \chi_{k,1-\alpha}^2$ at level α.

```
##  Box-Pierce test
##
## data:  daily_res
## X-squared = 8.4383, df = 20, p-value = 0.9885
```

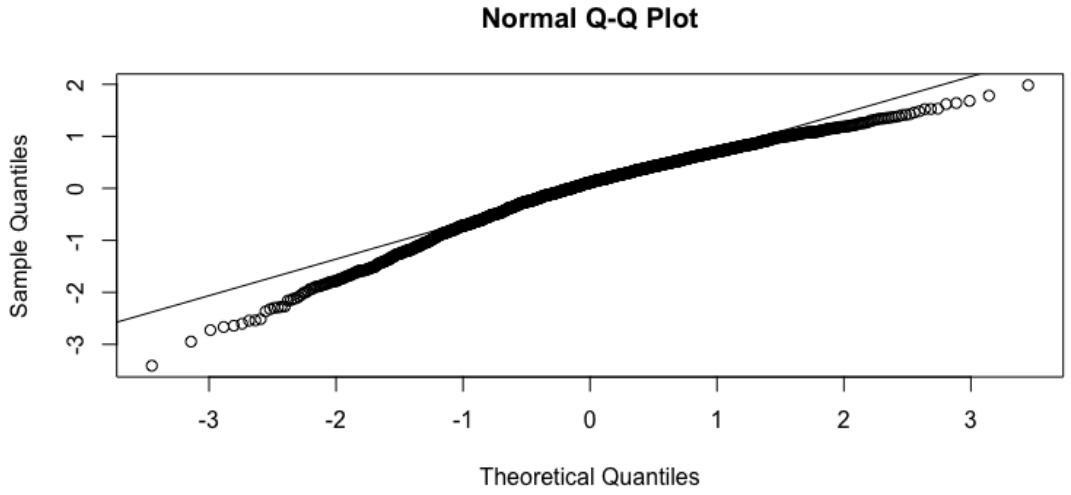
The p-value of Box-Pierce test on the residuals is 0.9885 which is larger than the significance level $\alpha = 0.05$, so we fail to reject H_0, the residuals are independently distributed.

Test of normality:

We first used normal probability plot to evaluate the normality of our data and then used Shapiro-Wilk test to do the hypothesis testing.

The Shapiro-Wilk test is a test of normality in frequentist statistics:

1. H_0: data are normally distributed.
2. Under H_0, the test statistic is $W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$,where $x_{(i)}$ is the i^{th} order statistic and $\overline{x}$ is the sample mean.
3. Reject H_0 if $Q > \chi^2_{k,1-\alpha}$ at level α.




```
##  Shapiro-Wilk normality test
##
## data:  daily_res
## W = 0.96829, p-value < 2.2e-16
```

The Normal Q-Q plot shows that the points are not approximately lie on the $y = x$ line and heavy tails exist on both sides. Also Shapiro-Wilk test gives us a very small p-value. Hence, we should reject the null hypothesis and conclude that residuals are not normally distributed.

Finally, we computed the sample mean and sample variance of the residuals that are -0.00039(almost zero) and 0.55617 respectively. Therefore, the residuals are not Gaussian white noise, but are independently distributed with sample mean 0 and sample variance 0.55617.

Using the same methods but in the next part, we did hypothesis testing on daily PM 2.5 data.

Test of randomness:

```
##  Box-Pierce test
##
## data:  pm_daily
## X-squared = 431.28, df = 20, p-value < 2.2e-16
```

The p-value of Box-Pierce test on the daily PM 2.5 data is 2.2×10^{-16}, so we reject H_0, i.e. the daily PM 2.5 data are dependent.

Test of normality:

```
##  Shapiro-Wilk normality test
##
## data:  pm_daily
## W = 0.9852, p-value = 1.261e-12
```

The p-value of Shapiro-Wilk test on the daily PM 2.5 data is 1.261×10^{-12}, so we reject H_0, i.e. the daily PM 2.5 data are not normally distributed. And this is not surprising because the result of normality test on the fitted residuals shows that the normality is violated.

Test of stationarity:

Here we used Augmented Dickey-Fuller distribution to test whether the daily PM 2.5 data are stationary time series process:

Augmented Dickey-Fuller test:

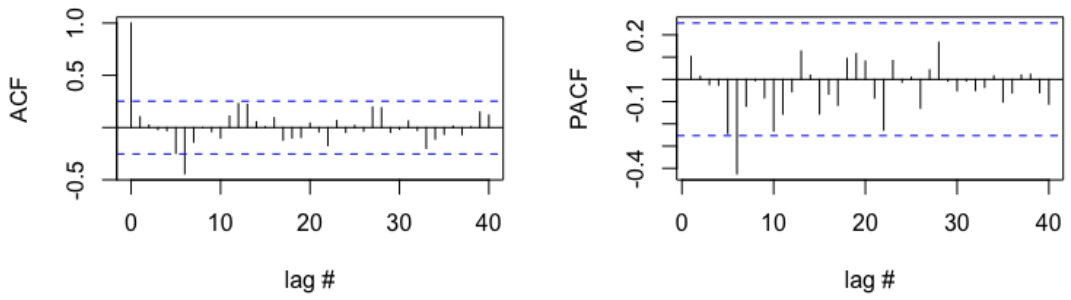
1. The testing procedure is applied to the model:$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \epsilon_t$, where α is a constant, β is the coefficient on a time trend and p is the lag order of the autoregressive process.
2. Hypothesis: H_0: $\gamma = 0$ (the time series data are **not** stationary) $vs.$ H_a: $\gamma < 0$ (the time series data are stationary).
3. Under H_0, the test statistic is $DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$

```
##  Augmented Dickey-Fuller Test
##
## data:  pm_daily
## Dickey-Fuller = -10.407, Lag order = 12, p-value = 0.01
## alternative hypothesis: stationary
```

The p-value of Dickey-Fuller test on the daily PM 2.5 data is 0.01, so we reject H_0, i.e. the daily PM 2.5 data are stationary time series process.

Part 2: Monthly PM 2.5 data

In this section, we did time series analysis on monthly PM 2.5 data by averaging the PM 2.5 values in each month. The analysis approach is very similar to the last section, first we fitted the monthly data a time series model and then performed hypothesis tests on the fitted residuals and the data itself.



The ACF and PACF shown in the figure above are suggestive of an MA(6) or AR(6) model, as the value of ACF drops dramatically after lag 6 and the absolute value of PACF has a sharp decrease after lag 6. Then we used function ***Arima( )*** in R package *forecast* to conduct the model fitting:

```
fit_monthly_ma = Arima(pm_monthly,order=c(0,0,6))
fit_monthly_ar = Arima(pm_monthly,order=c(6,0,0))
```

Here we used **AIC**(Akaike information criterion) to choose which model that we prefer to use for further analysis. The AIC is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. (Wiki link:https://en.wikipedia.org/wiki/Akaike_information_criterion)

Model selection for MA(q) model:

$$AIC = -2log(Maximum\ Gaussian\ likelihood) + 2(q+1)$$

Model selection for AR(p) model:

$$AIC = -2log(Maximum\ Gaussian\ likelihood) + 2(p+1)$$

```
##     MA(6)  AR(6)
## AIC 12.83 18.467
```

We prefer the model with smaller AIC value, so we choose MA(6) model:

```
##          ar1          ar2          ar3          ar4          ar5
## -0.023130956  0.003555347 -0.013987767  0.028318112 -0.208285054
##          ar6    intercept
## -0.432717231  4.508533265
```

Test of randomness:

```
##  Box-Pierce test
##
## data:  monthly_res
## X-squared = 8.7963, df = 20, p-value = 0.9851
```

The p-value of Box-Pierce test on the residuals is 0.9851 which is larger than the significance level $\alpha = 0.05$, so we fail to reject H_0, i.e. the residuals are independently distributed.

Test of normality:

```
##  Shapiro-Wilk normality test
##
## data:  monthly_res
## W = 0.98071, p-value = 0.4591
```

The p-value is 0.4691, we should fail to reject H_0, residuals are normally distributed.

Hence, the residuals are Gaussian white noise, i.e. independently distributed with sample mean -0.00218 and sample variance 0.0025.

Next, we tested the randomness, normality and stationarity of the monthly PM 2.5 records.

```
##  Box-Pierce test
##
## data:  pm_monthly
## X-squared = 27.581, df = 20, p-value = 0.1197

##  Shapiro-Wilk normality test
##
## data:  pm_monthly
## W = 0.9882, p-value = 0.8307

##  Augmented Dickey-Fuller Test
##
## data:  pm_monthly
## Dickey-Fuller = -3.5647, Lag order = 3, p-value = 0.04378
## alternative hypothesis: stationary
```

According to the output, we concluded that monthly PM 2.5 data are stationary time series process, and they are independently and normally distributed.

Summary

Fitting daily PM 2.5 data by MA(2) model, we obtained non-Gaussian but independently distributed noise and the data itself is stationary, dependent but non normal distributed.

Fitting monthly PM 2.5 data by MA(6) model, the fitted residuals are normally and independently distributed. And the monthly PM 2.5 data are stationary and independent Gaussian process.

We want the fitted residuals to behave like Gaussian white noise, thusly we consider that the fitted model MA(6) using monthly PM 2.5 data might be a good fit, while the fitted model MA(2) using daily PM 2.5 data does not work well. To build a better time series model for daily PM 2.5 data, one might need to consider more complicated time series models. Compare the AIC and RMSE (Root Mean Square Error: $RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$) of two fitted models, fitted model MA(6) has much smaller AIC value and lower RMSE.

```
##                       AIC  RMSE
## Daily PM 2.5   4034.693 0.746
## Monthly PM 2.5   12.830 0.221
```

Last but not least, the time series analysis that we have performed in this project is naive due to the lack of knowledge and practical experience. To make more precise time series models on these two records, one could consider a seasonal ARMA model or doing (paramedic and non-parametric) regression with ARIMA errors, which requires more advanced knowledge related to these topics.

Linear Regression Analysis Section

The goal of this part is about selecting an appropriate model consists of variables that influence PM2.5 in Beijing significantly. This process is based on the remaining data that was included in the dataset.

The final model is selected mainly by making transformations of predictor variables and response variable, including interactions of predictor variables and using criteria for model selection.

Model Selection

Analysis of data:

Based on the data, we can classify the type of predictor variables as qualitative and quantitative predictors. The PRES(pressure), DEWP(dew point), TEMP(temperature), Is(cumulated hours of snow), Ir(cumulated hours of rain) and Iws(cumulated hours of wind speed) can be seen as quantitative variables. The cbwd(combined wind direction) can be seen as qualitative variable. They can be represented as follows:

Let w_i mean the response variable, that is PM2.5 concentration. Then let $Y_i = \log w_i$.

The predictor variables are as following:

X_{i1}: pressure(hpa), X_{i2}: dew point, X_{i3}: temperature, X_{i4}: cumulated hours of snow, X_{i5}: cumulated hours of rain, X_{i6}: cumulated hours of wind speed

$$X_{i7} = \begin{cases} 1 & \text{if combined wind direction} = \text{northeast} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{i8} = \begin{cases} 1 & \text{if combined wind direction} = \text{southeast} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{i9} = \begin{cases} 1 & \text{if combined wind direction} = \text{calm and variable} \\ 0 & \text{otherwise} \end{cases}$$

Transformation of response variable:

Firstly, we make the regression model by including every variable without any transformation. Then we can get the model_1 and the QQ-Plot of this model. From the output, we can find that the Adjusted R-squared is 0.26. The value of AIC is 482595. And the QQ-Plot is:

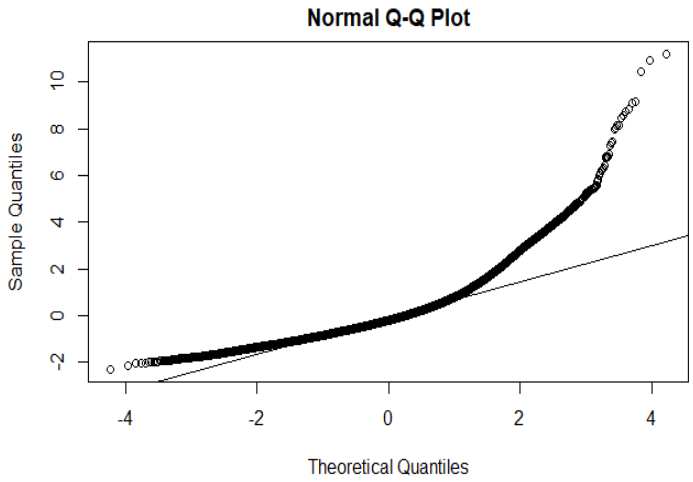


Figure1: QQ-plot of model_1

From this plot, it is clear that the model isn't fitting the data very well and the QQ-plot shows a large deviation from normality. Then we make the transformation of response variable: log(PM2.5). The new model called model_2 can be constructed. For model_2, the Adjusted R-squared is 0.413, which has improved a lot and the value of AIC is 99005, which is lower than the AIC of model_1. The QQ-Plot of model_2 can be seen in figure5. The above result indicates that this model significantly increases the explanatory power of our model. So we can keep the transformation of response variable: log(PM2.5).

Criteria for model selection:

In order to make the final model have a good fitness, we can use some statistical criteria to determine the number of variables in the final model. Firstly, we use the ridge regression and Lasso to select variables. The value of R-squared after making ridge regression is 0.413, which does not improve the fitness of our model. And from the plot of making Lasso:

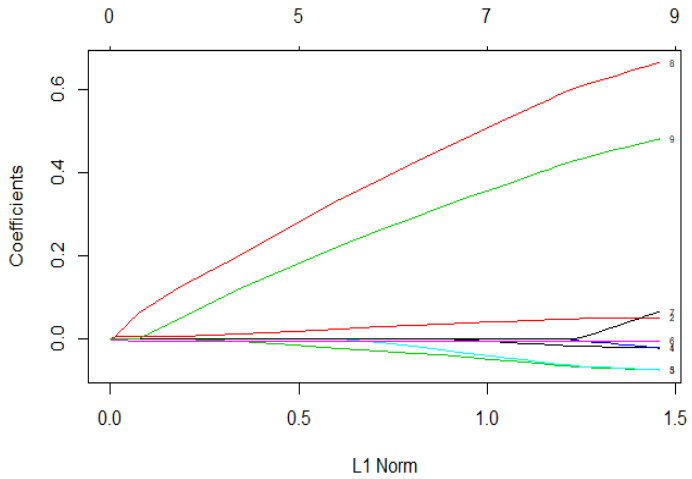


Figure 2: Lasso estimates

From the plot, we should keep the seven variables in the model.

What's more, from the output of selecting best subset by applying the Mallows' Cp criterion, we still find that we should keep the seven variables in our model.

Transformations of predictor variables:

Based on the type of predictor variables (qualitative variables and quantitative variables), we can find that whether the response variable has a statistical interaction between the polynomial of quantitative variables and the interaction between qualitative variables and quantitative variables respectively as well as the interaction between quantitative variables themselves.

Firstly, we want to find whether the quantitative variables(PRES, DEWP, TEMP, Is, Ir, Iws) have interaction between themselves. From R code, we can get their correlation as follows:

	PRES	DEWP	TEMP	Snow	Rain	Wind_speed
PRES	1.0000	-0.7777	-0.8269	0.07054	-0.08053	0.17888
DEWP	-0.7777	1.0000	0.8239	-0.03493	0.12534	-0.29308
TEMP	-0.8269	0.8239	1.0000	-0.09478	0.04955	-0.14976
Snow	0.0705	-0.0349	-0.0948	1.00000	-0.00976	0.02264
Rain	-0.0805	0.1253	0.0496	-0.00976	1.00000	-0.00914
Wind_speed	0.1789	-0.2931	-0.1498	0.02264	-0.00914	1.00000

From this result, we can easily find that there might be correlation between PRES and DEWP, PRES and TEMP, DEWP and TEMP, TEMP and PRES. However, after adding these interactions terms to the model respectively, we get the result that the Adjusted R-Squared hasn't improved much and so is the value of

AIC. That is to say, adding these new terms won't contribute a lot to make our model fits better. Thus, we decide not to add any interaction term between quantitative variables into the model.

Secondly, we want to test whether the response variable has statistical interaction between the polynomial of quantitative variables. After adding the polynomial transformation of each quantitative variable to the model, we find that the values of Adjusted R-Squared and AIC still haven't improved much. In addition, from the plot of response variable and every quantitative variable, we can make some transformations based on the shape of plot. For example, from the following plot about response variable and temperature,

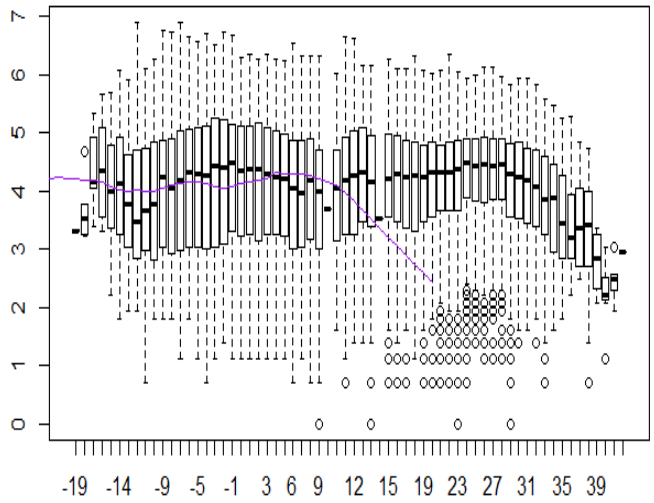


Figure 3: the smooth line between log(PM2.5) and temperature

the shape of this plot is close to the image of $ax^2 + bx + c$, then we can add poly(X_{i3}, 2) to the model. However, after comparing the Adjusted R-Squared 0.415 and the value of AIC 98910 with the current model, these values haven't been improved much. Thus, we decide not to add this polynomial term to our model. Similarly, making this process for other variables and comparing the value of Adjusted R-Squared and AIC. Finally, there is no evident improvement of the model after adding these polynomial terms, so we don't add any polynomial term to our model.

Thirdly, we want to test whether log(PM2.5) depends on a statistical interaction between qualitative variables and quantitative variables, that is the interaction between six quantitative variables and the wind direction. We can conclude the result from another type of plots. If the lines in the plot are parallel, then log(PM2.5) does not depend on a statistical interaction between them. Otherwise, we may consider add the interaction term to the model. For instance, the plot about log(PM2.5) and pressure,

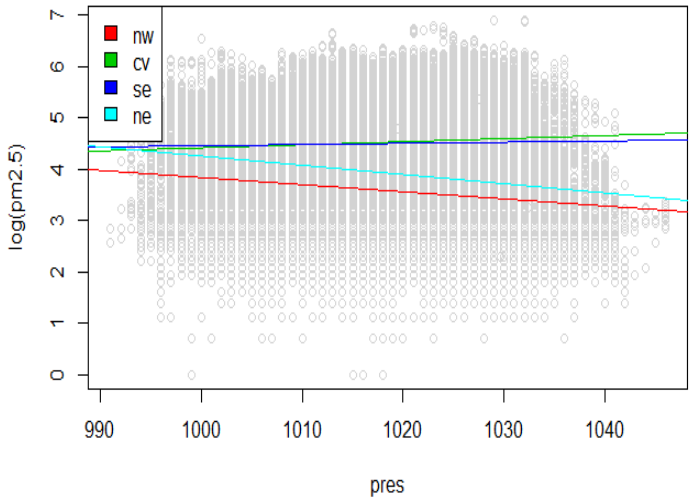


Figure 4: scatter plot about pressure with smoothers for each cbwd level

The lines in this plot are not parallel and this result indicates that log(PM2.5) depends on a statistical interaction between pressure and combined wind direction. However, after adding this interaction term to the model, the Adjusted R-Squared 0.416 and AIC 98790 haven't improved much. Thus we conclude that we should not include this interaction term in the model. Similarly, making the same analysis of other variables, and based on the result that the value of Adjusted R-Squared and AIC haven't improved much, we decide not to add any interaction term to the model.

Diagnostics and Model Validation:

Ideally we want the proportion of the wind direction levels to be the similar for the full data, training data and validation data. Based on the output of R, we can find the output:

MSPR	MSE	MSEearler
0.622	0.628	0.627

Although the value of MSE is not very close to 0, the value of MSPR is very close to MSE. Thus our model can be seen as an appropriate model for this dataset.

Conclusion

Basic exploratory analysis of the final model:

The followings are some basic exploratory analyses of the final model.

Firstly, the final model is:

$Y_i = 26.00996 – 0.02095\ X_{i1} + 0.05221\ X_{i2} - 0.0737\ X_{i3} - 0.02054\ X_{i4} -0.07368\ X_{i5} - 0.00345\ X_{i6} – 0.06649\ X_{i7} + 0.66395\ X_{i8} + 0.4806\ X_{i9} + \varepsilon_i$ (1)

In this model, i means the ith data and i = 1, 2, …… 41755

Assuming that $\varepsilon_i \sim N(0, \sigma^2)$

Secondly, the following plots related to the final model can reflect more information about the fitness of the final model:

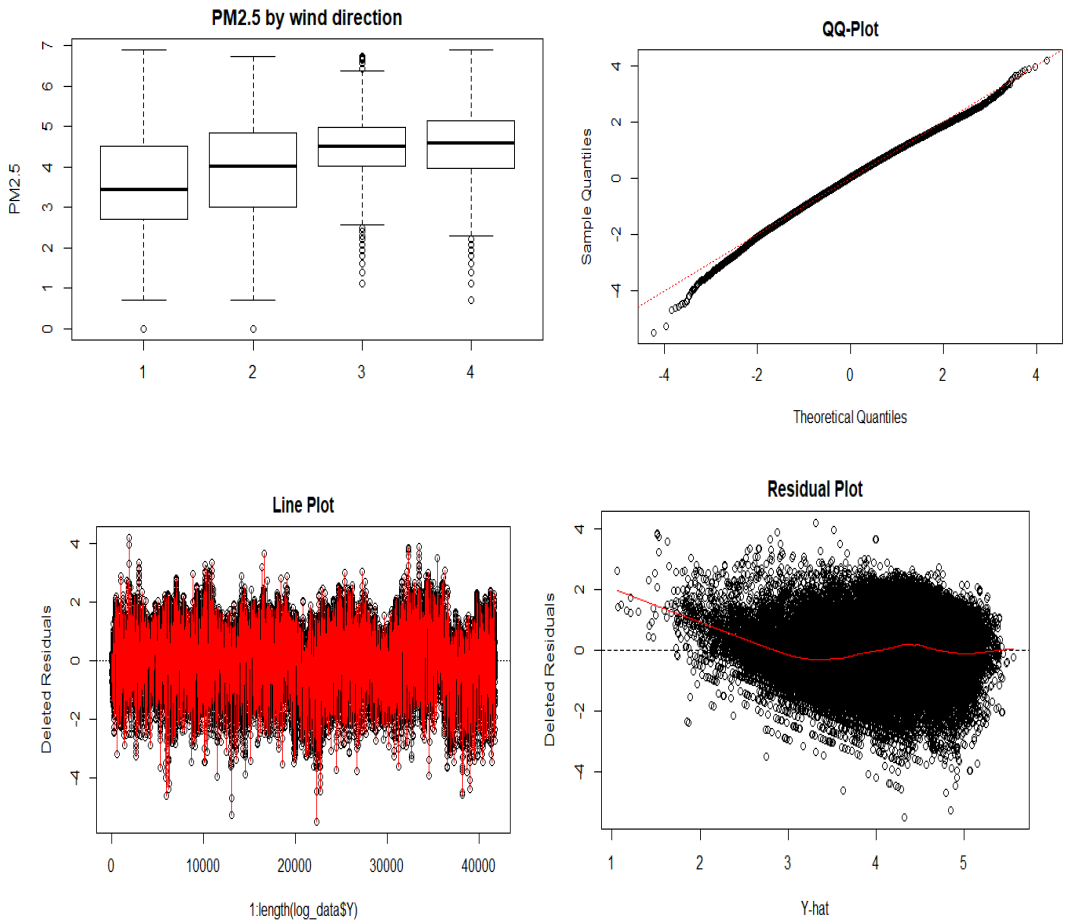


Figure 5: exploratory analysis of the final model

From the boxplot about PM2.5 based on different combined wind direction, number of outliers compared with the huge data can be ignored, so the errors of this model can be seen as normality.

From the QQ-Plot, we can conclude that the errors of the final model are normally distributed.

From the Line Plot, we can conclude that errors of the final model have constant variance and are independent and identically distributed.

From the residual plot, we can find that the response function is linear. Errors have constant variance and are independent and identically normally distributed. To conclude, the final model satisfies major assumptions of regression model.

Summary of the final model:

The following will show some important characteristics of the final model based on the outputs in R. The form of the final model in R is:

formula = log(PM2.5) ~ PRES + DEWP + TEMP + Is + Ir +Iws + cbwd

Table 1: Basic form of the final model

Transformation of response variable	Predictor variables	Transformations of predictor variables	Interactions
log(PM2.5)	PRES, DEWP, TEMP, Is, Ir, Iws, cbwd	no	no

The outputs of some quantities in R can be concluded as:

Table 2: The quantities of the final model

AIC	R^2	R_a^2	$MSPR$
99005	0.413	0.413	0.622

Conclusion and discussion

The whole report mainly discusses the Beijing PM2.5 data from three aspects, the summary statistics, time series analysis and linear regression analysis. Based on the results shown above, we can make some predictions and provide suggestions in order to prevent the worse pollutions caused by PM2.5 in the future.

By exploring data according to different days and different time in a day, we found a seasonal and daily distribution. From the final model, the statistics show that the parametric model may not appropriate for this data set and this suggests that we can use a nonparametric model to fit the Beijing PM2.5 nature. What's more, the model hasn't been improved much by making any transformations of predictor variables or adding any interaction terms. Under this circumstance, we can collect other variables related to the air pollution like the automobile exhaust to see the deep link between PM2.5 and our life.

References:

1. Beijing PM 2.5 Data: https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data

2. 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.

3. The definition of AIC: https://en.wikipedia.org/wiki/Akaike_information_criterion

4. Air quality guide for PM 2.5: https://www.cnn.com/2017/05/04/asia/beijing-sand-storm-pollution-beyond-index/index.html