STAT 306
Group Project Report: Beijing PM2.5 data

Presented to:
Professor Bruce Dunham

Date Submitted: December 7, 2022

Group B1:

Zejun Su - 72603681
Pontus Sjostrand - 90888793
Yulu Duan - 23366081
Yifan Wu - 99677320

1. Introduction

Air quality has become a significant factor influencing human health as urban development and modernization accelerate. According to the National Weather Service (n.d.), an estimation of more than 100,000 premature deaths would arise from poor air quality and the costs from air pollution-related illness are expected to be at $150 billion in the United States per year. Air pollutants are considered to be made up of gaseous pollutants and particulate matters (Xing et al., 2016), and meteorological factors would significantly impact the amount of pollution in the atmosphere (Shenfeld, 1970).

In this report, we will analyze the relationship between different meteorological parameters and the air quality in Beijing, China. The air quality is in this report defined as fine particulate matter (PM 2.5). The data set we use in the analysis contains PM2.5 data from the US embassy in Beijing and meteorological data from Beijing International airport. The data time period is from Jan 1st, 2010 to Dec 31st, 2014. The data consist of around 40 000 data points, measured hourly during the time period. Source of data is from Songxi, Chen, csx@gsm.pku.edu.cn, Guanghua School of Management, Center for Statistical Science, Peking University.

In the dataset, we use the PM 2.5 concentration measured in micrograms per cubic meter ($\mu g/m^3$) as the response variable. As explanatory variables, we use air temperature measured in °C, air pressure measured in hPa, dew point also measured in °C, cumulated hours of rain, cumulated hour of snow and the season (winter, spring, summer or fall) and wind direction (NE, NW, SE and “no wind”).

The reason why we choose to study this topic is because it is of interest to predict air quality based on variables which are easy to measure. If we find a relationship between measured air quality (in this case PM 2.5), temperature, air pressure and season, we can predict air quality based on metrics that are easy and cheap to measure. The need for an expensive air quality monitor would become less important and recommendations to people regarding air quality could be made even without a specific air quality measurement device if a clear relationship between the explanatory variables and the response variable can be observed.

In the proposal for this project, we stated that we would include wind speed as an explanatory variable. However, after looking closer into the raw data, we found out that the cumulative wind speed was rather hard to interpret and had to be looked at together with the current wind speed in order to understand it correctly. We therefore decided to use “season” and “wind direction” as explanatory variables instead since we think those could also impact the air quality.

2. Analysis

Model Explanation:

To build our model, we will use the linear regression algorithm to model the relationship between air quality (pm2.5) in Beijing and seasons, dew point temperature, temperature, pressure, and precipitation. We begin by cleaning the dataset by removing 24 hours of data for one day if one hour is missing for that specific day. To further explore the seasonal characteristics of the pm2.5 index, we divide the data into 4 seasons:

- **Winter** for Dec, Jan, Feb
- **Spring** for Mar, Apr, May
- **Summer** for Jun, Jul, Aug
- **Fall** for Sep, Oct, Nov

We also transform the response variable by taking the logarithm of it before making any plots or fitting any models. The reason why we do this is that the PM2.5 concentration can (of course) never be negative. By taking the log of the concentration before fitting a model, predicting a response and taking *e* to the power of that response, we make sure that a predicted response is never negative. A problem we face when making this transformation is that it is not possible to take the logarithm of 0, and there are two observations in the dataset where the PM2.5 concentration is 0. These observations probably appeared because of some measurement error (it's not reasonable to have zero concentration), and the problem is easily solved by removing those two observations.

Scatter Plots and Boxplots Analysis

Before building our model, we will first analyze a few characteristics of how the meteorological factors influence air quality (measured by the concentration of PM2.5) based on the visualization we have made.

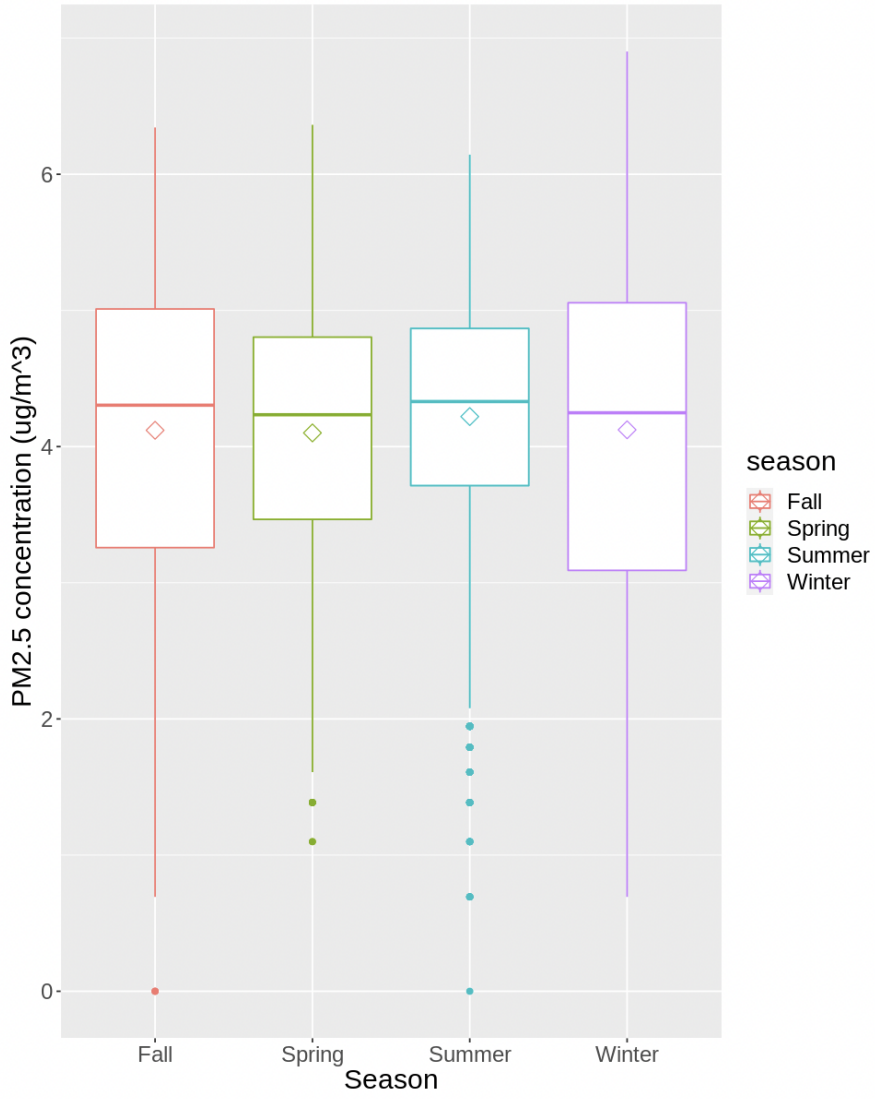

Figure 1

In the boxplot showing the distribution of PM2.5 concentration by the type of season, it could be seen that the mean of the PM2.5 concentration for each season are very close, with the variance larger in fall and winter and smaller in spring and summer. We can also see that there are more outliers in the winter.

We are then going to analyze the relationship between PM2.5 concentration and meteorological factors by looking into the scatterplot we have made.

PM2.5 concentration vs. dewpoint (by season)

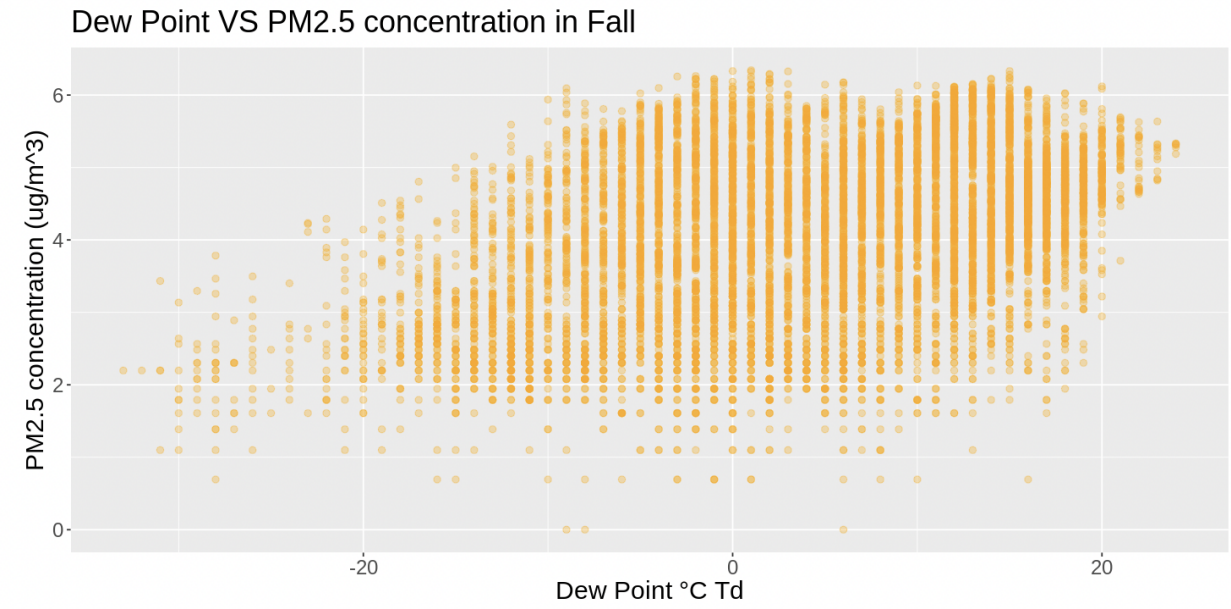

Figure 2

Fall: there is a very weak positive relationship between these two variables in fall, as the concentration of PM2.5 tends to increase when the dew point increases.

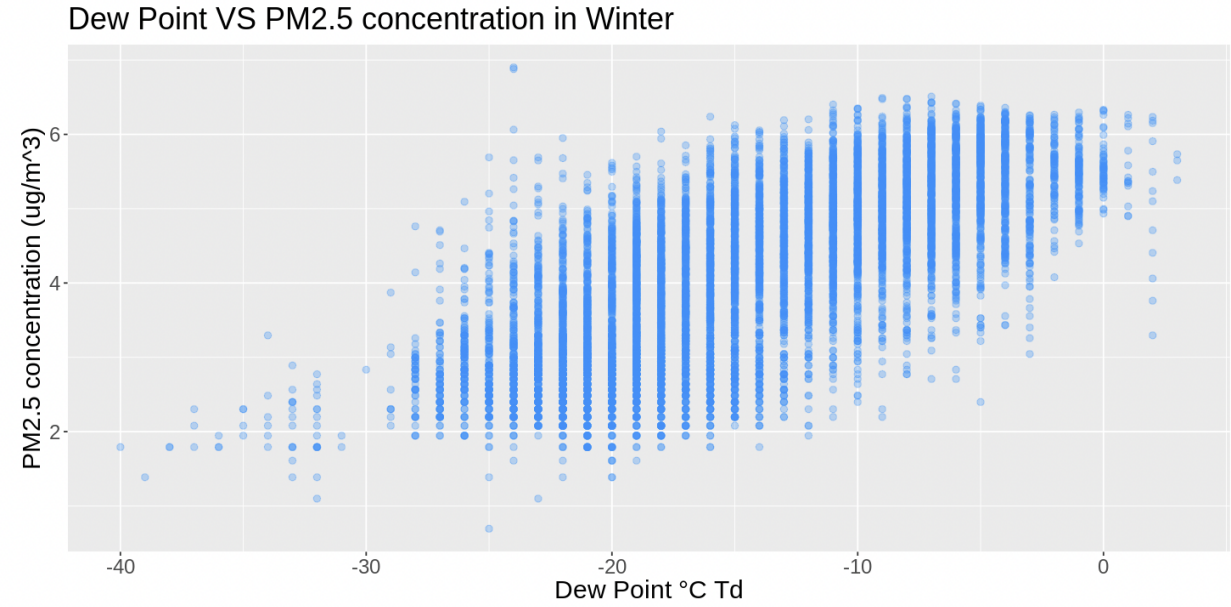

Figure 3

Winter: there is a weak positive relationship between the two variables in winter as the concentration of PM2.5 increases with the dew point increasing.

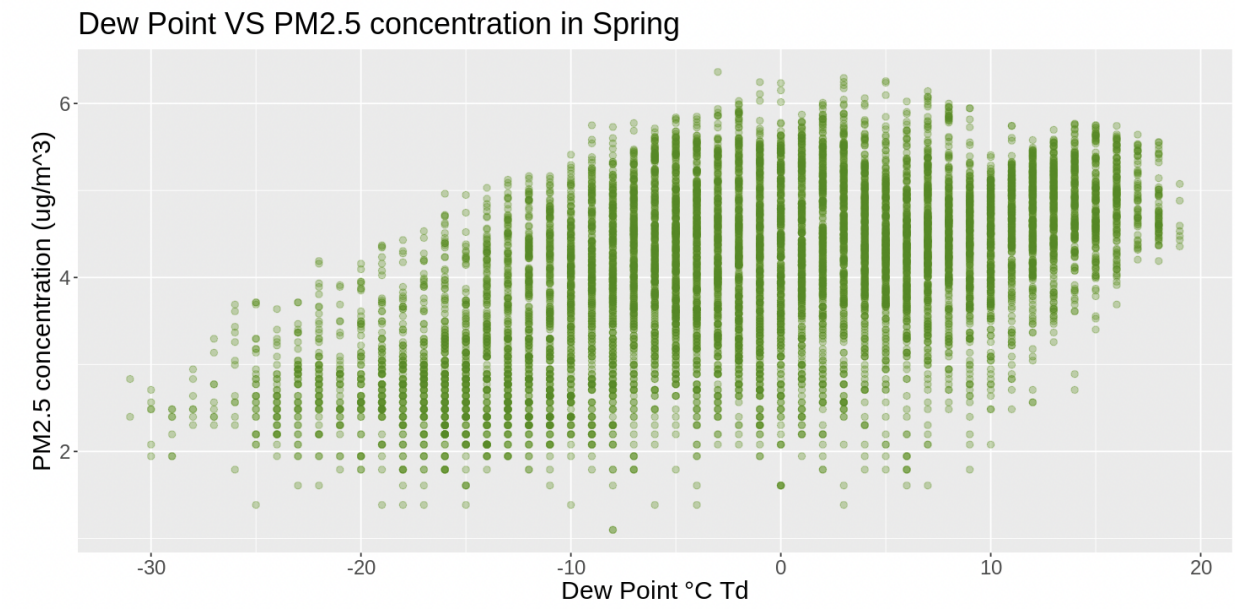

Figure 4

Spring: there is a very weak positive relationship between the two variables in spring as the concentration of PM2.5 tends to increase when the dew point increases. The scatterplot is still distracted with a few outliers.

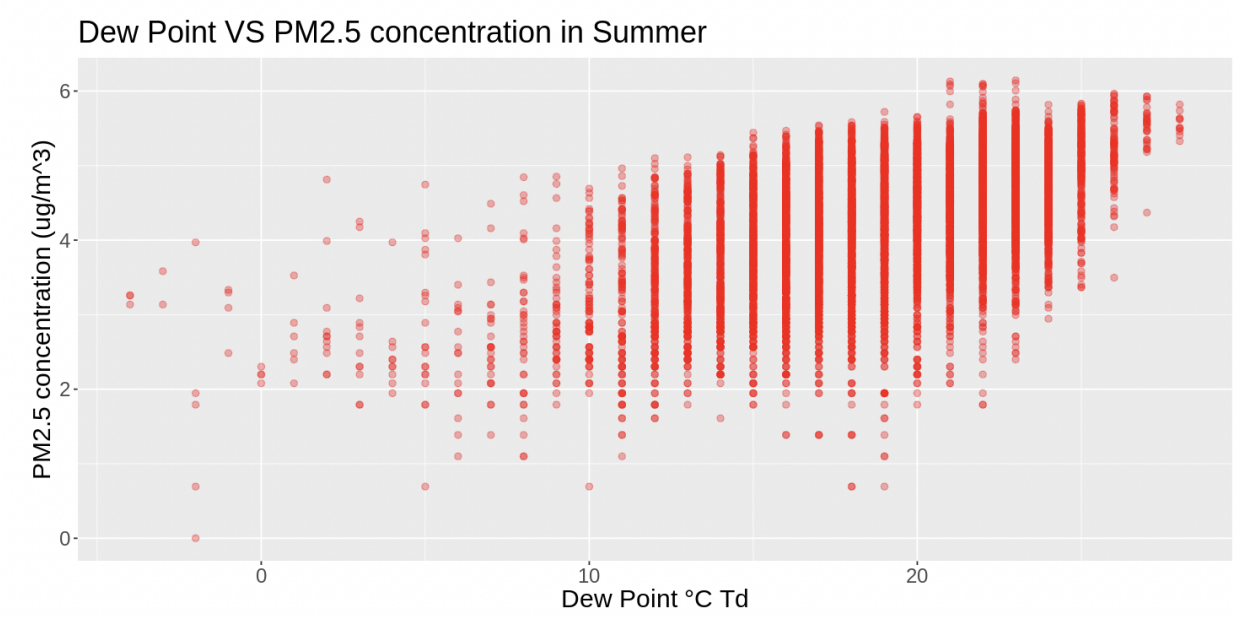

Figure 6

Summer: there is a weak positive relationship between the two variables in summer as the concentration of PM2.5 tends to increase when the dew point increases. The plot is more distracted when the dew point is around -5 to 5°C.

PM2.5 concentration vs. dewpoint (by combined wind direction)

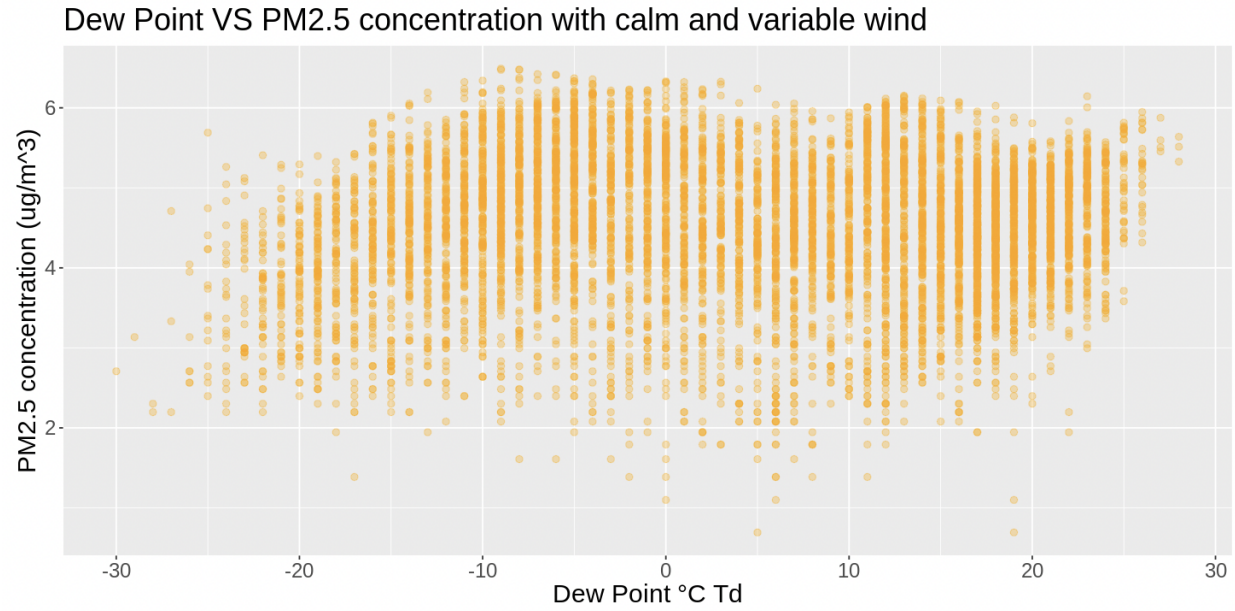

Figure 7

Calm and variable wind: the data points are completely spread out in the scatterplot above, which means there is no correlation between the concentration of PM2.5 and dew point with the calm and variable wind.

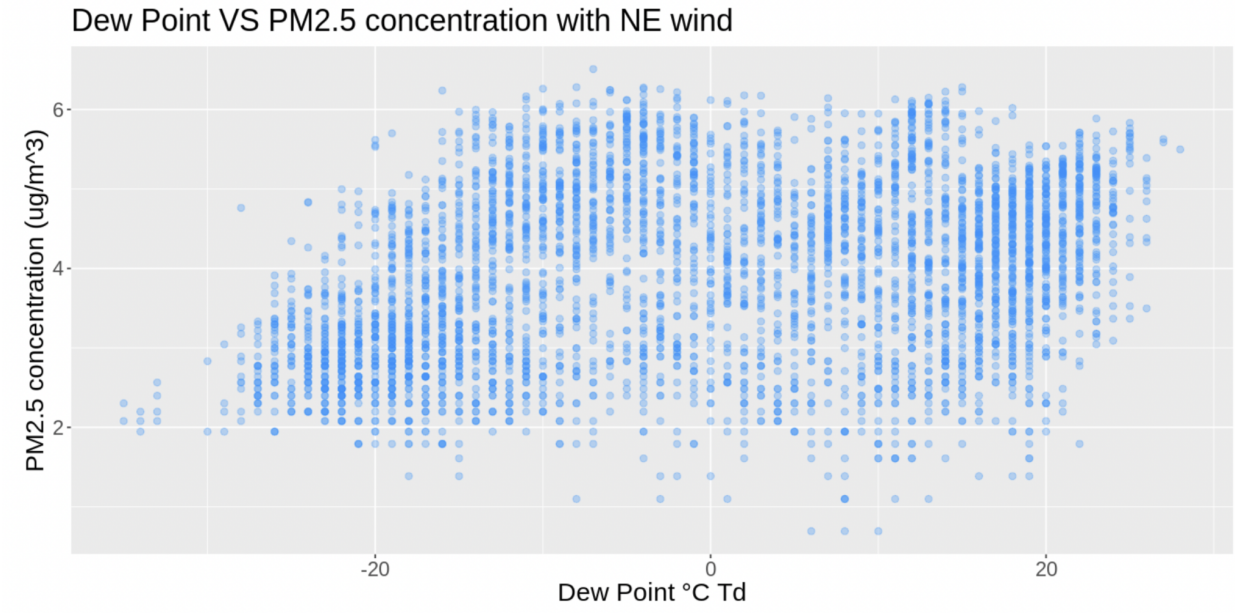

Figure 8

NE wind: there is a weak positive relationship between the concentration of PM2.5 and dew point with the northeast wind as the concentration of PM2.5 tends to increase when the dew point increases.

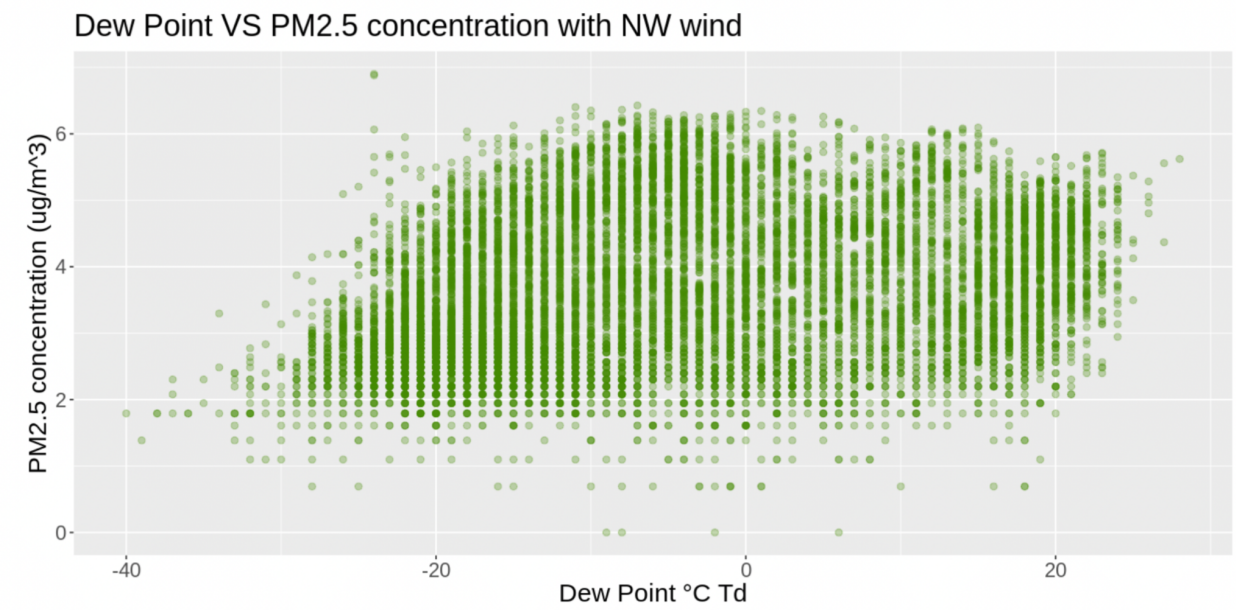

Figure 9

NW wind: the data points show no trend for the two variables, which means there is no correlation between the concentration of PM2.5 and dew point with the northwest wind.

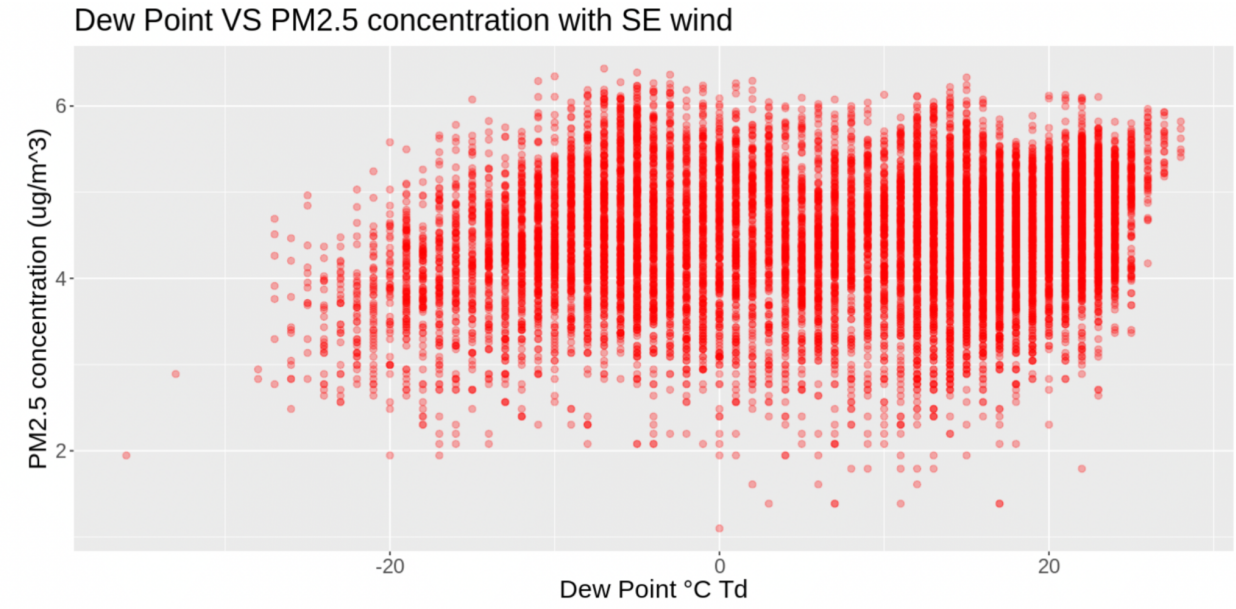

Figure 10

SE wind: the data points are distracted in the scatterplot, which means there is no correlation between the concentration of PM2.5 and dew point with the southeast wind.

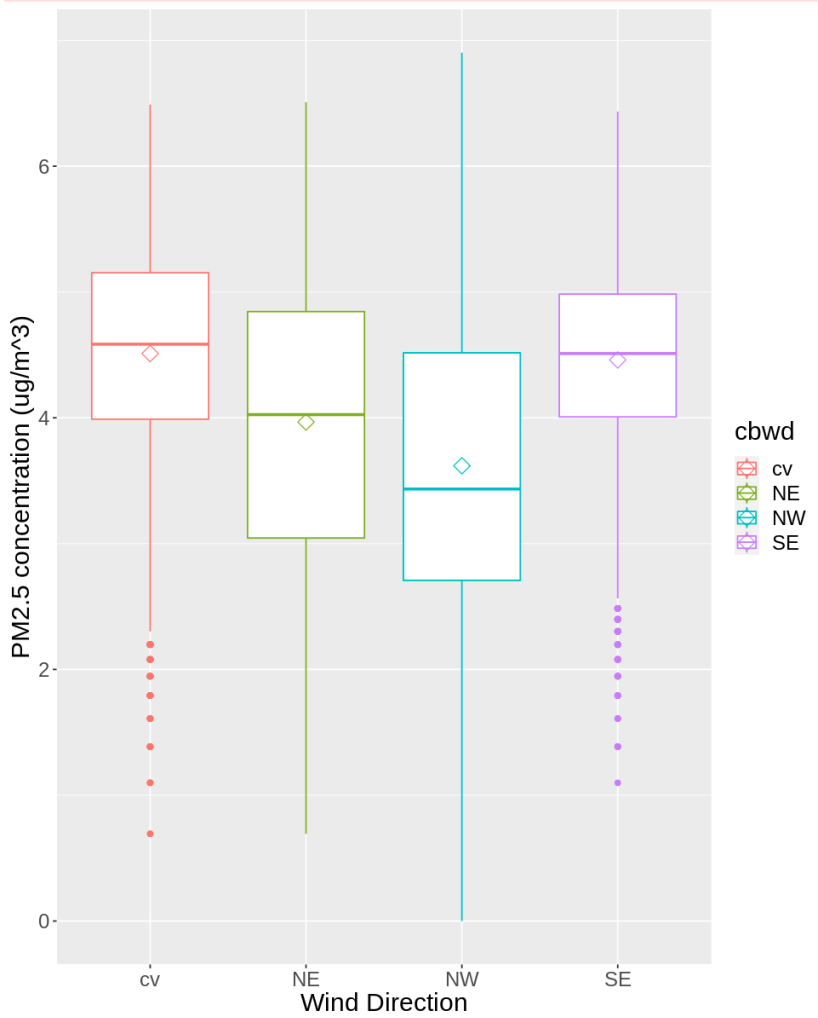

Figure 11

Another Boxplot indicates that wind direction can also affect the PM2.5 concentration. Due to the mountains in the northwest of Beijing, the south wind cannot blow through. Therefore, only the north wind will bring down the PM2.5 concentration.

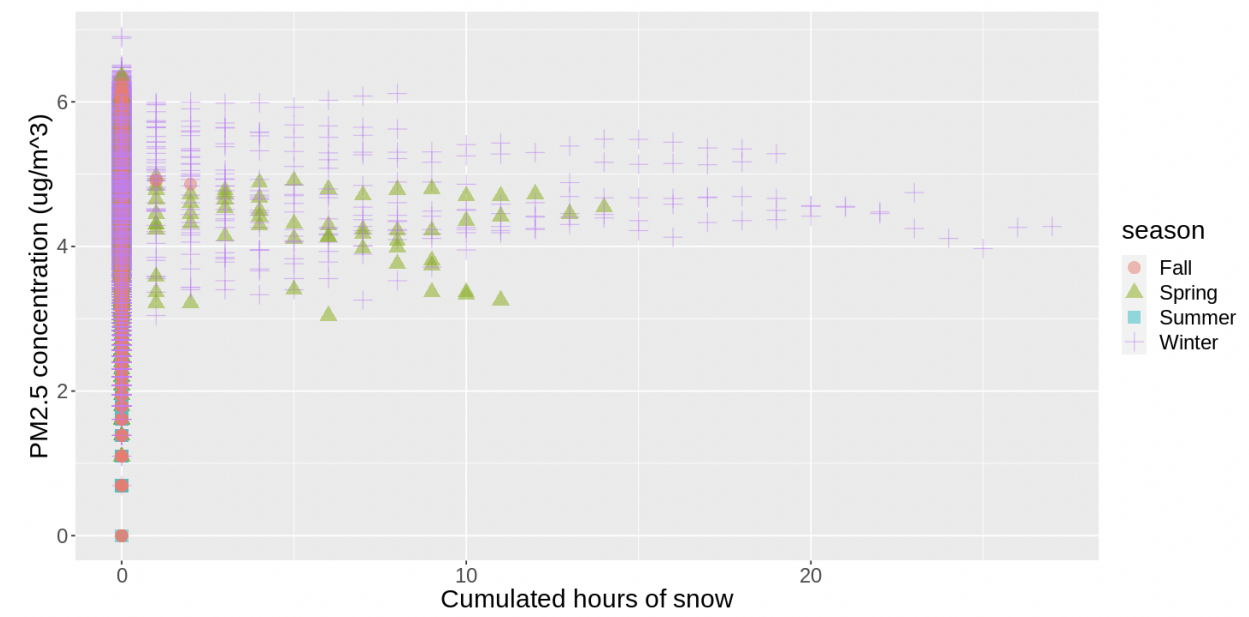

Figure 12

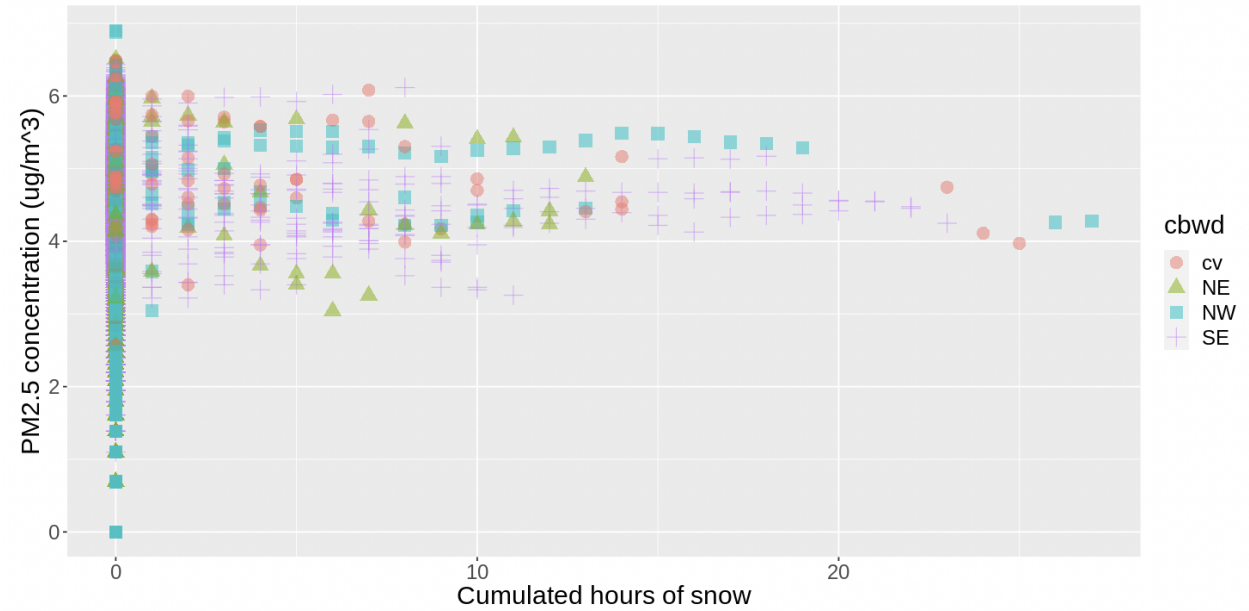

Figure 13

PM2.5 concentration vs. cumulated hours of snow

The data points are largely gathered around 0 which cannot show any pattern between the concentration of PM2.5 and cumulated hours of snow regardless the data points are sorted by combined wind direction or season. Therefore, the variable “cumulated hours of snow” is not going to be used to build our model.

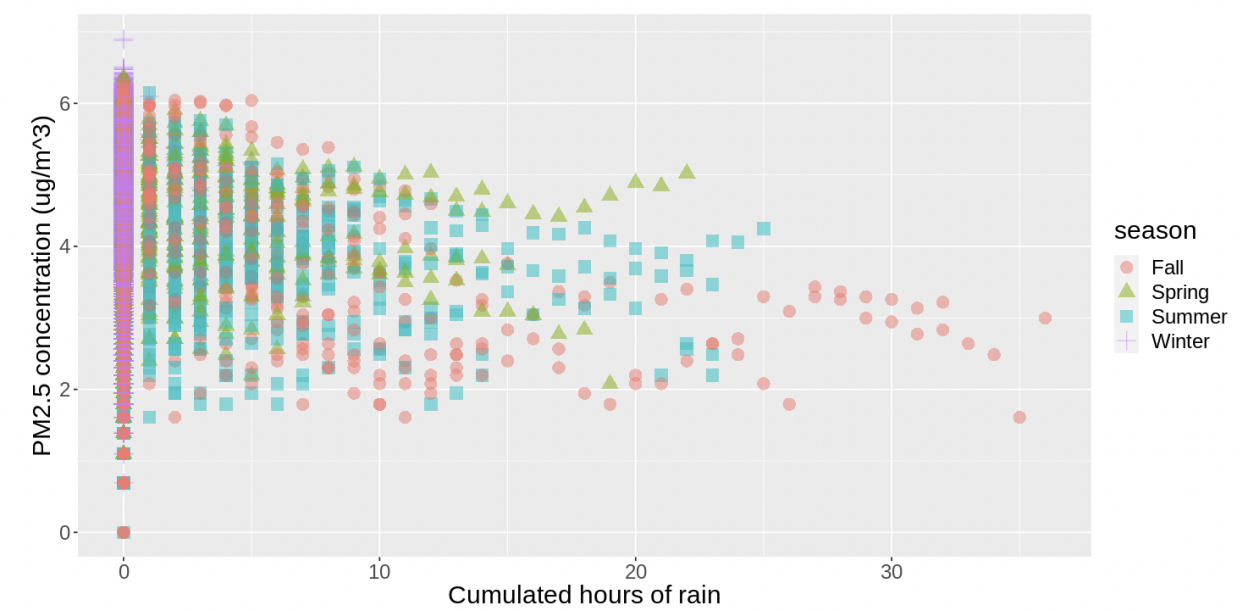

Figure 14

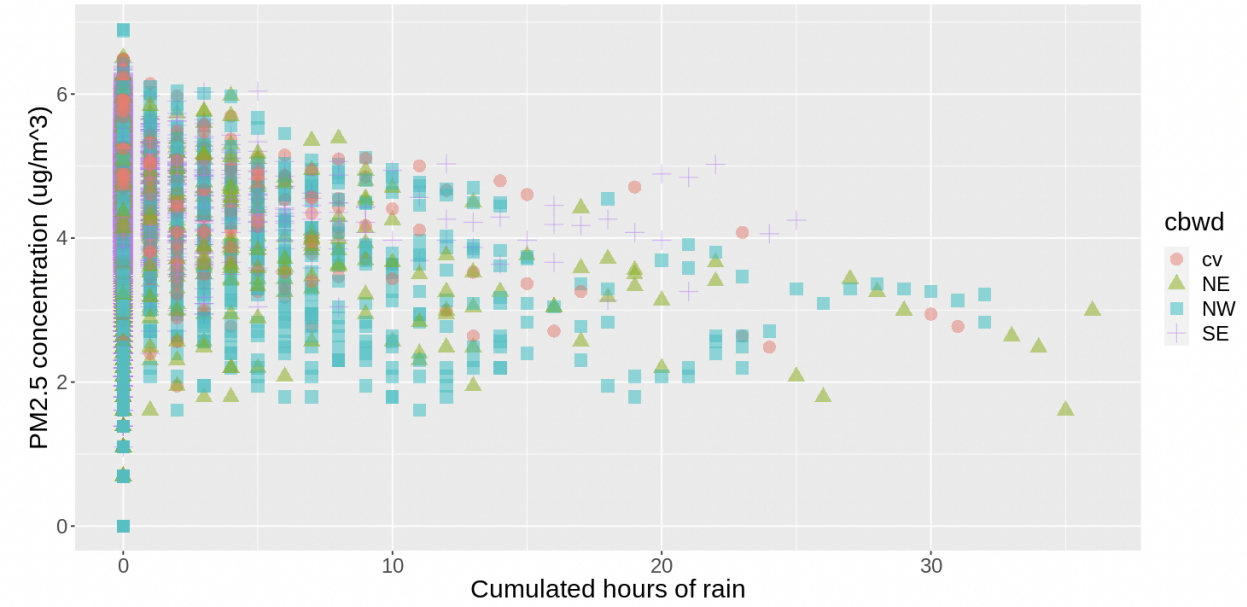

Figure 15

PM2.5 concentration vs. cumulated hours of rain

Similar to the variable “cumulated hours of rain”, the data points are mostly gathered around 0 which cannot show any pattern between the concentration of PM2.5 and cumulated hours of rain no matter the data points are sorted by combined wind direction or season. Therefore, we are not going to use the variable “cumulated hours of rain” to build our model.

After plotting the relationships between each measurable explanatory variable and the air quality (pm2.5 index), we find that the dew point temperature, temperature, pressure, precipitation, and seasons are the main factors that seem to affect the air quality in Beijing.

Model fitting

We have a very large sample which contains about 40k observations in our study of air quality model fitting. The forward model selection method will not work under such a large sample size. The goal of the forward model selection method is to address the overfitting problem by considering the penalization term of the number of parameters because a model's residual sum of squares always decreases as additional variables are included. Therefore, the model with all variables would always be chosen if our goal is to choose the model that produces the least residual sum of squares. However, in our case, we have 40k observations and only 5 predictors. Following Mallows' *Cp* statistics, the prediction error will always dominate the formula, and the penalization of the number of variables that we selected will not contribute much to the process of model selection.

$$C_p = \frac{SSE_p}{S^2} - N + 2(P+1)$$

Therefore, we choose all the variables that are easy to measure and show a relatively strong correlation with our response variable, the index of PM2.5, by analyzing the scatter plots. We have two categorical data: season and wind direction which show a correlation with PM2.5, and we try to explore the relationship between air quality, dew point temperature, temperature, pressure, wind direction, and season.

From this reasoning and the reasoning from the plots above, we will fit three different models. In all models, we include dew point, temperature and pressure which seemed to be the most important explanatory variables looking at the plot.

Model	**Explanatory variables**	**P-value**	**Adjusted R^2**
Model 1	Dew point Temperature Pressure Season	<2e-16 <2e-16 <2e-16 <2e-16	0.3573
Model 2	Dew point Temperature Pressure Wind direction	<2e-16 <2e-16 <2e-16 <2e-16	0.3818

Model 3	Dew point Temperature Pressure Wind Direction Season	<2e-16 <2e-16 <2e-16 <2e-16 <2e-16	0.4265

In the first model, we selected the dew point, temperature, pressure, and season as our explanatory variables according to the analysis of the scatter plots. From the table above we can see that the p-value for the dummy variable seasons spring, summer, and winter is very significant, suggesting that there is statistical evidence of a difference in pm2.5 between the seasons. Moreover, the Adjusted R-squared is 0.3573 which means that about 36% of the variation in the air quality can be explained by the temperature, dew point temperature, pressure, and seasons.

Furthermore, we try to explore the relationship between wind direction and PM2.5, and how the wind direction contributes to the prediction. In the second model, we change the season to wind direction and refit the model. From the table above, the p-value for wind direction is also very significant, suggesting that there is evidence that the change in wind direction will affect PM2.5. Compared with the first model, the second model shows a higher adjusted r squared (0.3818), suggesting that about 38% of the variation in the air quality can be explained by the temperature, dew point temperature, pressure, and wind direction. Also, the wind direction provides more information for predicting the level of the PM2.5.

In our third model, we use both season and wind direction as categorical variables. We use three dummy variables to show the different seasons, and three dummy variables to show the different wind directions that will affect the pm2.5 concentration in different ways.

The P-values for all the dummy variables “spring”, “summer”, “winter”, “NE”, “NW” and “SE” are all very significant, suggesting that there is statistical evidence of a difference in pm2.5 between the seasons and wind direction. Moreover, the Adjusted R-squared is 0.4265 which means that about 43% of the variation in the air quality can be explained by the temperature, dew point temperature, pressure, seasons and wind direction.

When comparing this model to the two previous ones, we conclude that the adjusted R-squared is approximately 0.07 and 0.05 higher respectively. A higher adjusted R-squared means that we can better predict the air quality, but we have to pay the price of adding an extra explanatory variable compared to the other two models. However, such an increase in the R-squared value should be worth the trade-off and hence we want to investigate this model further by looking at the residuals.

Residual Plots Evaluation

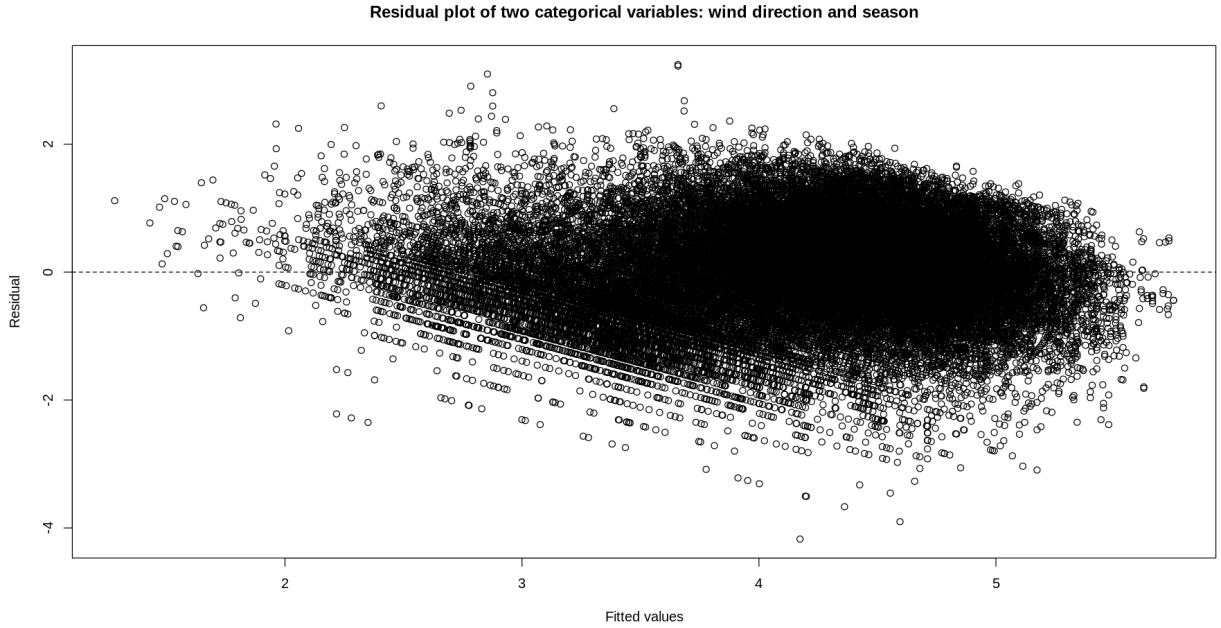

Figure 16

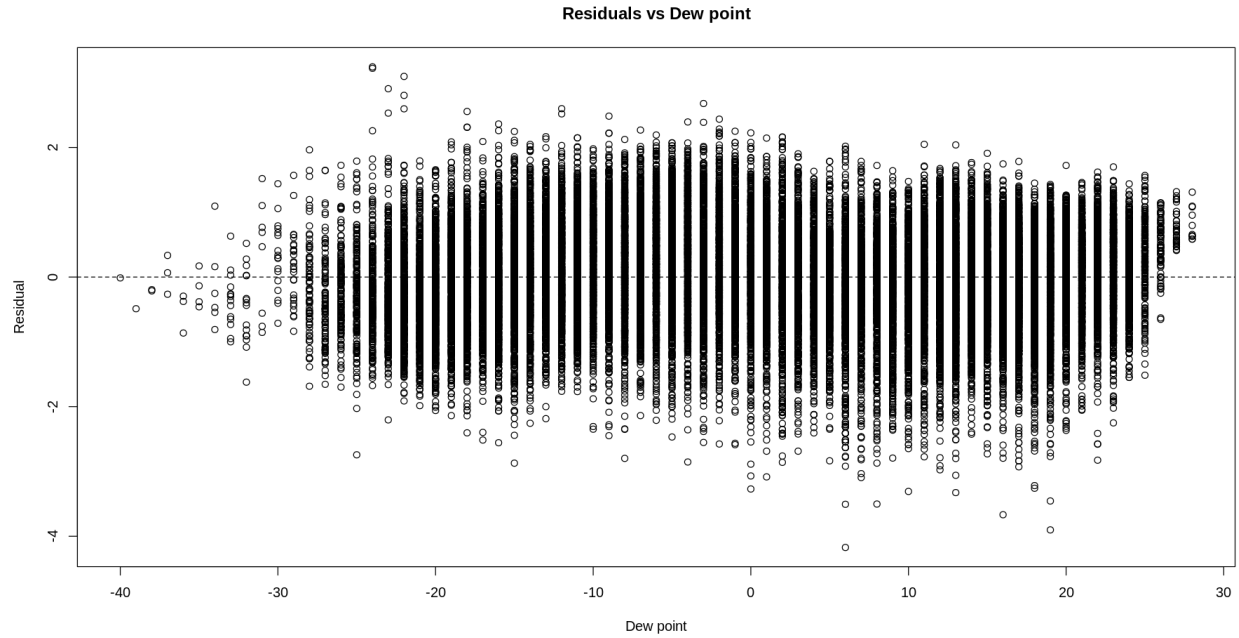

Figure 17

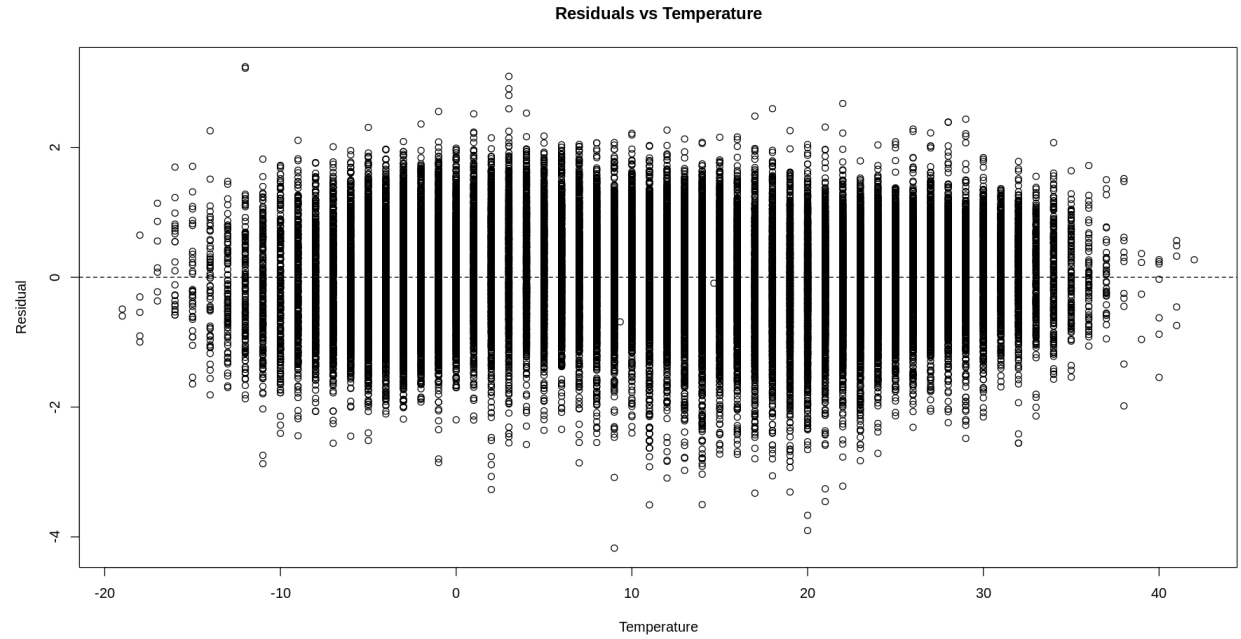

Figure 18

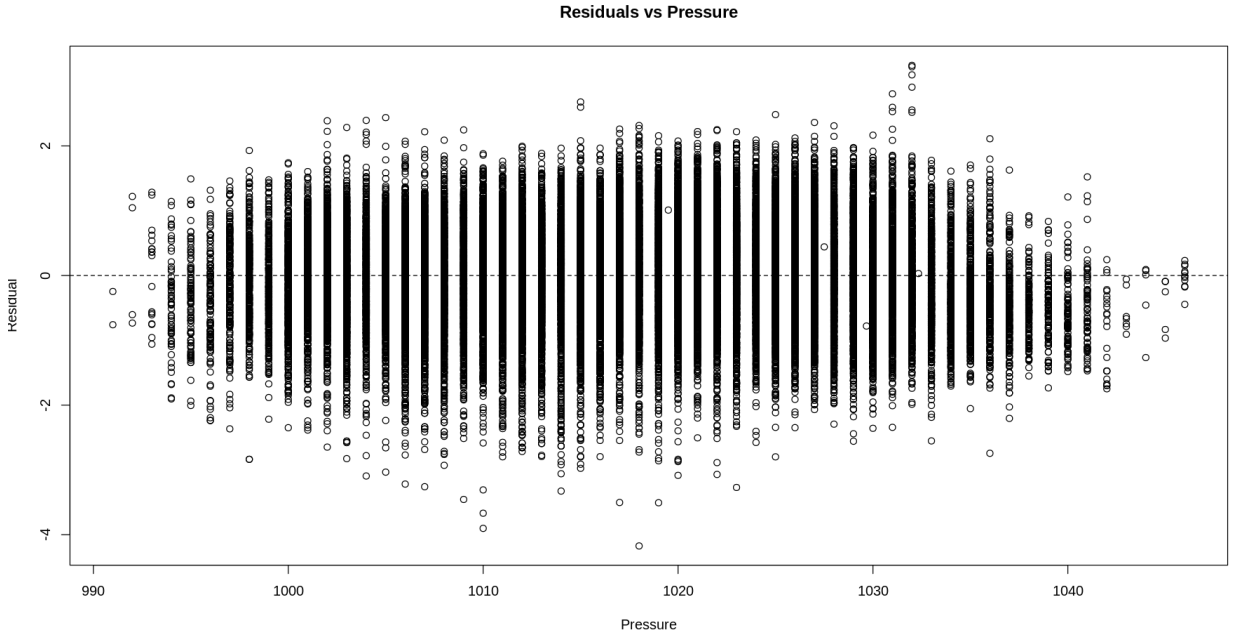

Figure 19

From the residual plots for this model, we can see that the observations are scattered randomly around the residual = 0 line, and no obvious pattern appears, so we can conclude that the linear model is appropriate for modeling the air quality.

To be more precise about model diagnostics, we drew the residual plots for the covariates of the third fitted model. For all three continuous variables pressure, temperature, and dew point, the

observations are all randomly distributed, and there are no obvious patterns found in their corresponding residual plots. Thus, we can conclude that the third linear model is appropriate.

The final model is then going to be:

$$PM2.5 = 27.373 + -0.4010\gamma_1 - 0.5756\gamma_2 + 0.0881\gamma_3 + 0.2339\gamma_4 - 0.5526\gamma_5 + 0.5832\gamma_6 + 0.0704d - 0.0630t - 0.0221p$$

where

$PM2.5$ is the PM2.5 concentration in $\mu g/m^3$

d is the dew point temperature in °C

t is the temperature in °C

p is the pressure in hPa

$$\gamma_1 = \begin{cases} 1 \; if \; wind \; direction \; is \; NE \\ 0 \; else \end{cases}$$

$$\gamma_2 = \begin{cases} 1 \; if \; wind \; direction \; is \; NW \\ 0 \; else \end{cases}$$

$$\gamma_3 = \begin{cases} 1 \; if \; wind \; direction \; is \; SE \\ 0 \; else \end{cases}$$

$$\gamma_4 = \begin{cases} 1 \; if \; season \; is \; Spring \\ 0 \; else \end{cases}$$

$$\gamma_5 = \begin{cases} 1 \; if \; season \; is \; Summer \\ 0 \; else \end{cases}$$

$$\gamma_6 = \begin{cases} 1 \; if \; season \; is \; Winter \\ 0 \; else \end{cases}$$

From the model above we can see that compared with the no-wind situation, PM2.5 concentration is 0.0881 μg/m^3 higher if the wind direction is in the southeast which means the air quality is worse with SE wind. Compared with the no-wind situation, PM2.5 concentration is 0.4010 μg/m^3 lower if the wind direction is in the northeast which means the air quality is worse with no wind compared with NE wind. Also, PM2.5 concentration is 0.5756 μg/m^3 lower if the wind direction is northeast, compared with no wind.

We also see that the air quality seems to be the worst during winter, since the categorical variable corresponding to winter is the highest. We have the next highest PM2.5 concentration during spring followed by fall and summer.

It is worth noting that the differences in intercepts that these two categorical variables contribute with is relatively small (all six coefficients are less than 1, compared to the baseline which is 27.373).

When looking at the continuous variables, we see that the PM2.5 concentration increases with an increased dew point temperature and decreases with higher air temperature. It also decreases with an increased air pressure, but the air pressure does not affect the PM2.5 concentration as much as the other two variables.

To sum it up, the air quality is predicted to be the best when there is wind from **northwest** during **summer**, and when the dew point temperature is **low,** and the air temperature and air pressure are **high**.

3. Conclusion

After fitting and choosing our final model, we can conclude that the explanatory variables in our dataset can predict the PM2.5 concentration to a limited degree. The R-squared for the final model is around 0.43, which means that 43% of the variation in the air quality can be explained by the explanatory variables in the final model.

A model for the PM2.5 concentration got continuously better as more explanatory variables were added (better as in higher adjusted R-squared). We also see that the P-value for all explanatory variables is very small, suggesting that it is very unlikely that any of them does not affect the PM2.5 concentration. This is a lot because of the fact that the dataset is huge (over 40 000 observations) and with so many observations, the penalty of adding an extra variable can be neglected. When choosing our final model, we therefore could not solely look at the adjusted R-squared nor Mallow's C_p at different models. Since the purpose of the analysis was prediction (rather than explanation) we did not mind a little complexity to the model if we increased the chances of predicting the air quality accurately.

A thing to note for this dataset is that many of the explanatory variables have some sort of correlation. This resulted in the fact that the model only got slightly better every time a new variable was added, since the explanatory variables often explained the same thing. If we would choose a new dataset to analyze, we would probably choose a dataset with explanatory variables that are not so correlated, since this could open up for more interesting analysis'.

There might be other variables that explain the PM2.5 concentration and that we have not included in this analysis. Since the R-squared in our final model is only 0.43, it suggests that that is indeed the case. Such variables may include other meteorological data. The air quality could also be affected by other variables, like wildfires in the area, if the observation is made on a weekend or not, if the measurement is on a holiday or not, the status of nearby industries and the amount of traffic that day.

In conclusion, our final model can make a prediction of the air quality in the Beijing area, but only to a limited degree.

References:

National Weather Service. (n.d.). Why Air Quality is Important. https://www.weather.gov/safety/airquality

Xing, Y., Xu, Y., Shi, M., & Lian, Y. (2016). The impact of PM2.5 on the human respiratory system. *Journal of Thoracic Disease, 8(1)*: 69-74. doi: 10.3978/j.issn.2072-1439.2016.01.19

Shenfeld, L. (1970). Meteorological aspects of air pollution control. *Atmosphere, 8:1*, 3-13. doi: 10.1080/00046973.1970.9676578