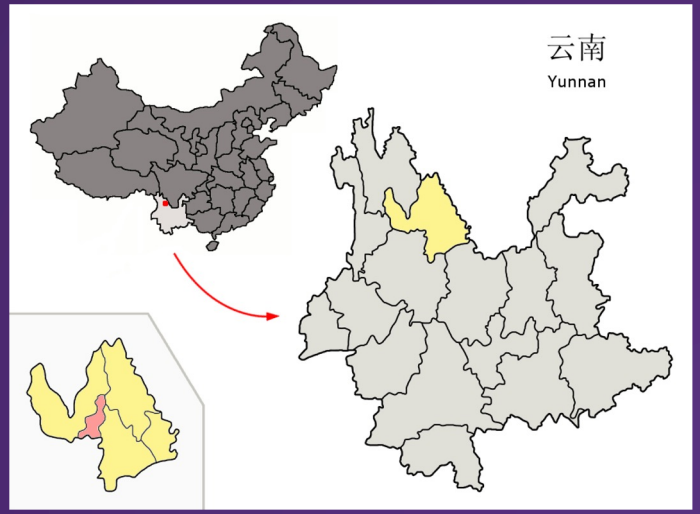
GUCHENG DATA: CONSULTING PROJECT

Group 4: Mengjia Zeng, Junseok Yang, Lavanya Upadhyaya, Wasay Siddiqui



Introduction

- A lot of people in Beijing, China face respiratory issues such as asthma, which can lead to hospitalizations, as well as long-term heart and lung conditions.

- Therefore, we investigated the impact of $PM_{2.5}$, or particulate matter, on air pollution and human health

- 6 main air pollutants and 6 relevant meteorological variables from 12 nationally-controlled air-quality monitoring sites in Beijing, China were observed.

Objective and Methodology

- **Visualizations:** Used to identify specific aspects that may influence $PM_{2.5}$ levels. These could be time of day, time of year, seasonal changes

- **Machine learning techniques:** Used to develop an "early warning system," or take information from past days/weeks and determine whether this data can be used to predict future pollutant level. This information will tell the public of when to take extra precautions.

    - **Linear regression models:** Used to examine the current data; for examples, relationships between $PM_{2.5}$ and other variables are explored to predict future trends.

Data description

- Categorical variable with the maximum daily concentration of $PM_{2.5}$.
- Continuous variable description (predictors): concentrations of PM_{10}, SO_2, NO_2, CO, and O_3; particular values of TEMP, (temperature), PRES (, DEWP, Rain, WSPM
    - Indicates variables that were not missing values, meaning that they contribute to the dataset. There were some variables such as No, year, month, day, hour, and station that were not as significant
- Categorical (response) variable: $PM_{2.5}$ level.

Statistical terminology

- **R^2:** A measure of how close the data fits the regression line/model to determine goodness of fit.
- **Correlation:** A statistical relationship between two or more variables.
- **Regularization:** A technique used to tune the function by adding additional penalty terms in the error function.

Data: Concentration levels

- Daily $PM_{2.5}$ level types:

  Low: Level ≤ 35

  Medium: 35 < Level ≤ 75

  High: 75 < Level ≤ 105

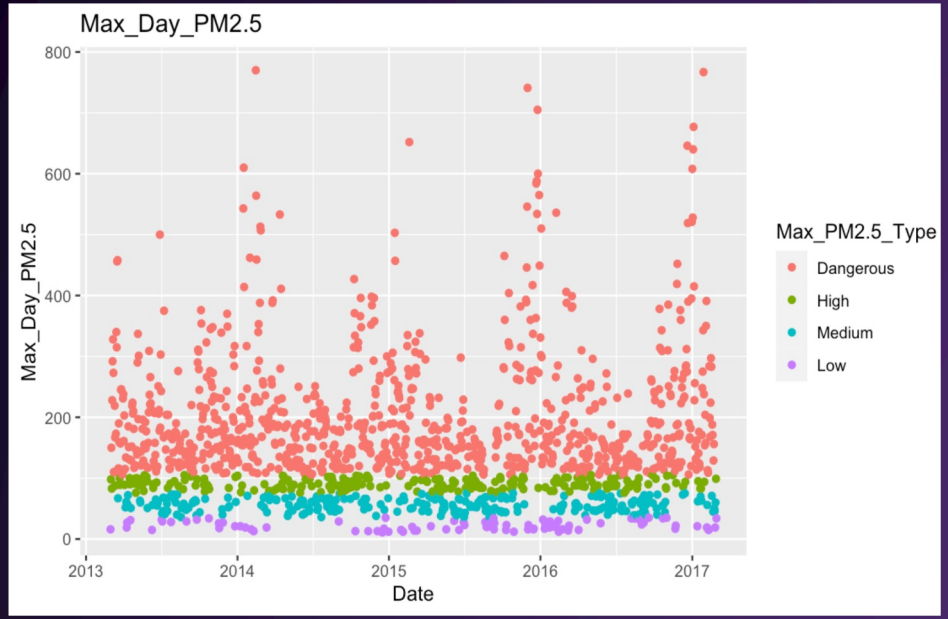
  Dangerous: Level > 105

	Dangerous	High	Medium	Low
2013	200	41	49	15
2014	232	58	62	13
2015	209	59	66	31
2016	211	54	70	31
2017	39	5	9	6

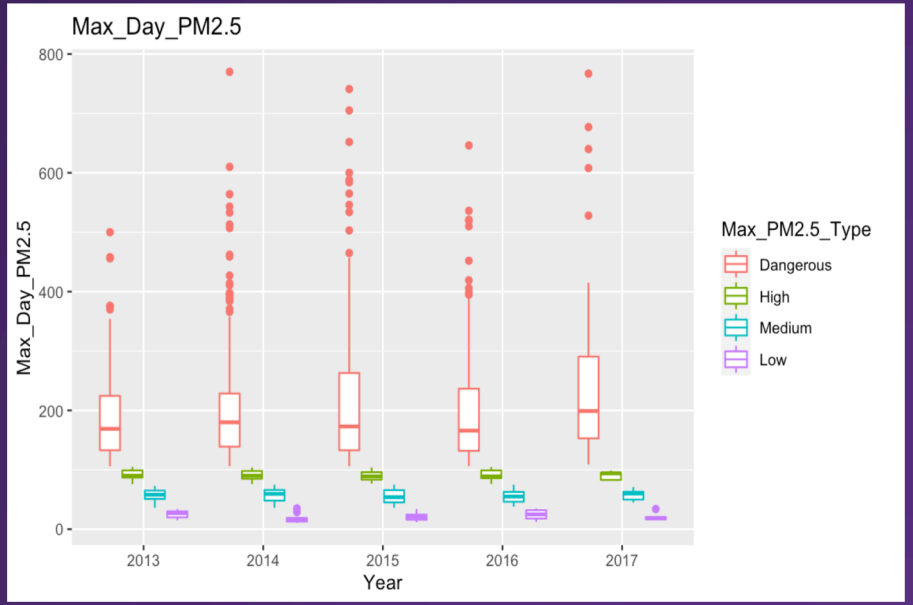
- Table shows concentration levels from the years of 2013-2017
  - From 2013-2016, there were “dangerous” amounts of $PM_{2.5}$
  - Lowest concentrations in 2017

$PM_{2.5}$ levels by YYYY-MM-DD and year (only)

YYYY-MM-DD



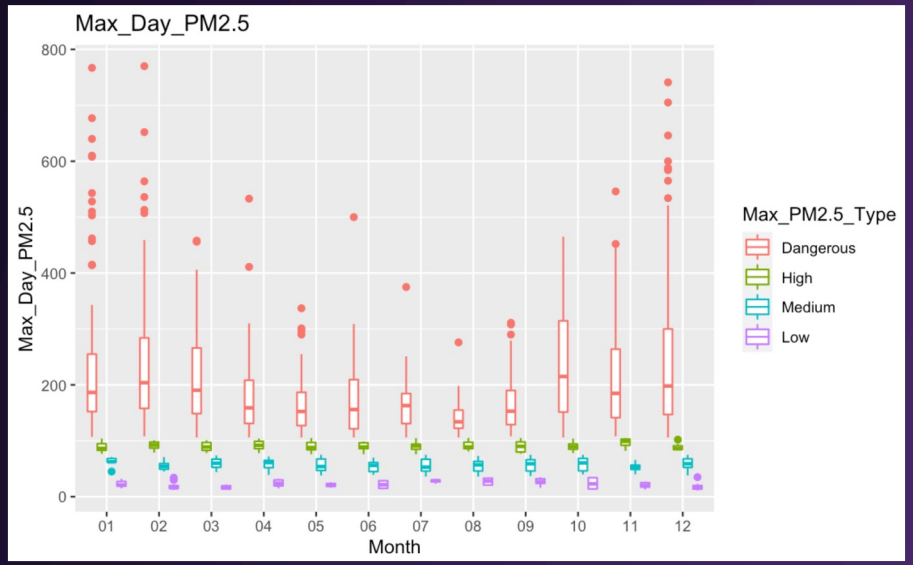
Year



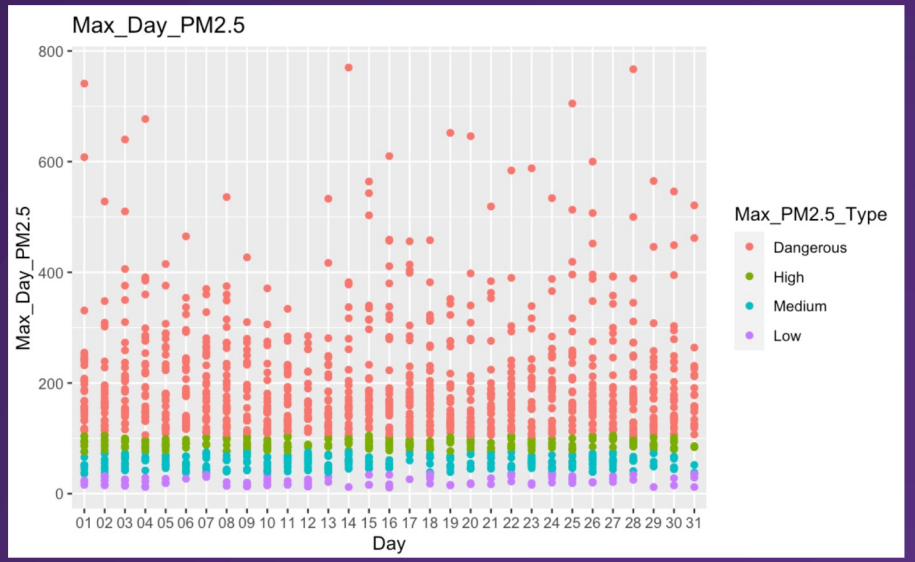
- Overall, dangerous levels increased as time went on; more points in the "dangerous" category in 2017

$PM_{2.5}$ levels by day of the month vs. weekday

Month



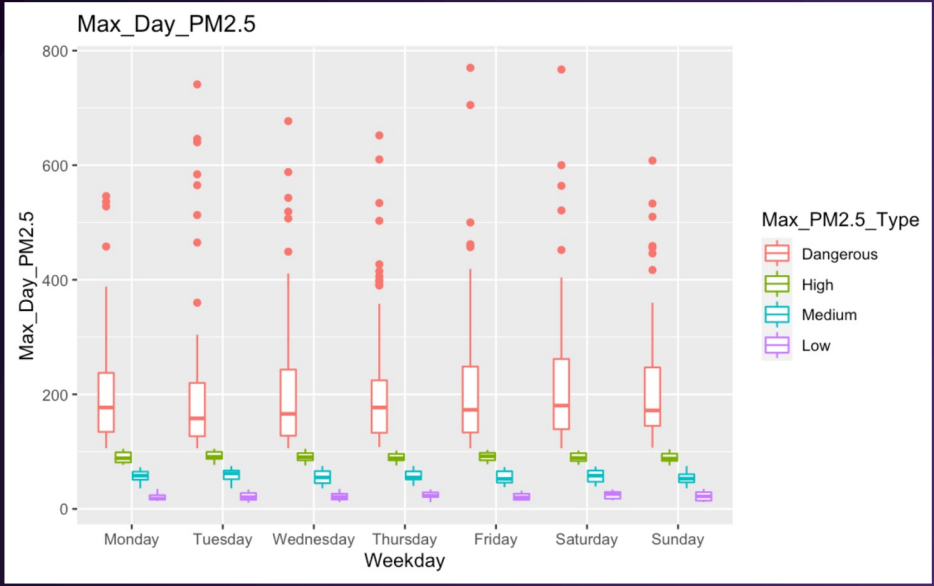
Day (in month)



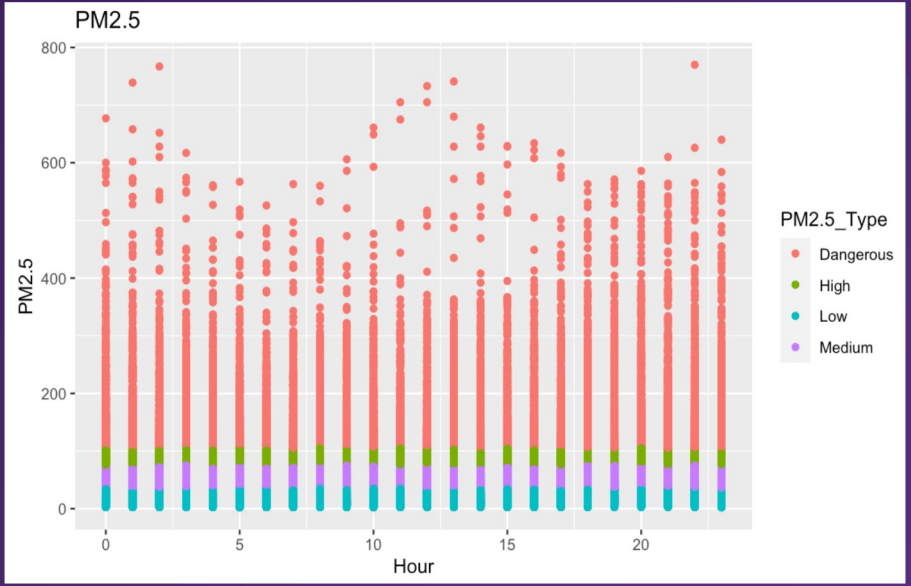
- Did not change to Box plot for "Day" since it was too narrow of a range

$PM_{2.5}$ levels by weekdays vs. hours in a day

Weekday

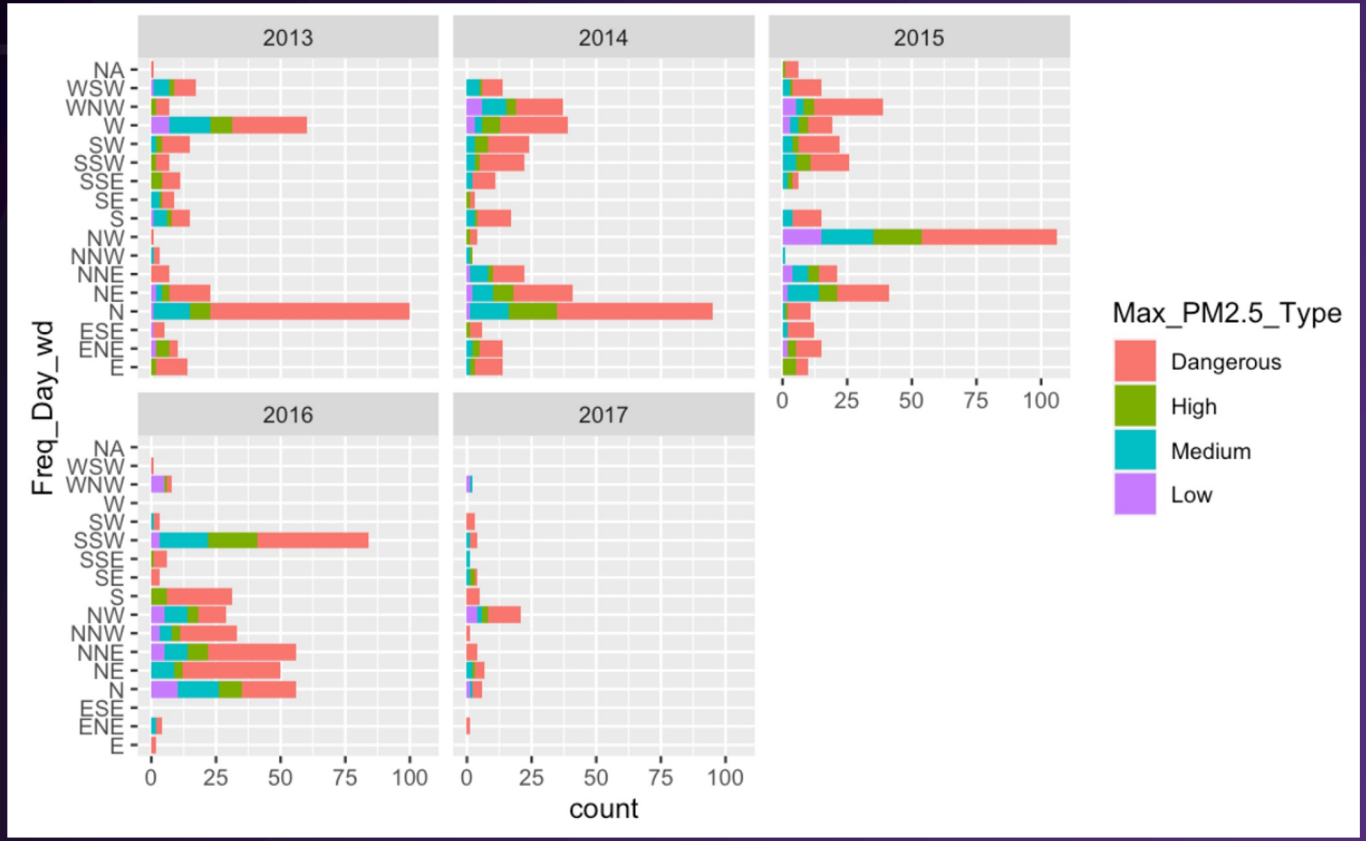


Hours in a day



When considering the maximum values, the highest $PM_{2.5}$ concentration seems to fall on Friday, and concentrations are higher during the middle of the day.

Frequency of wind direction by year



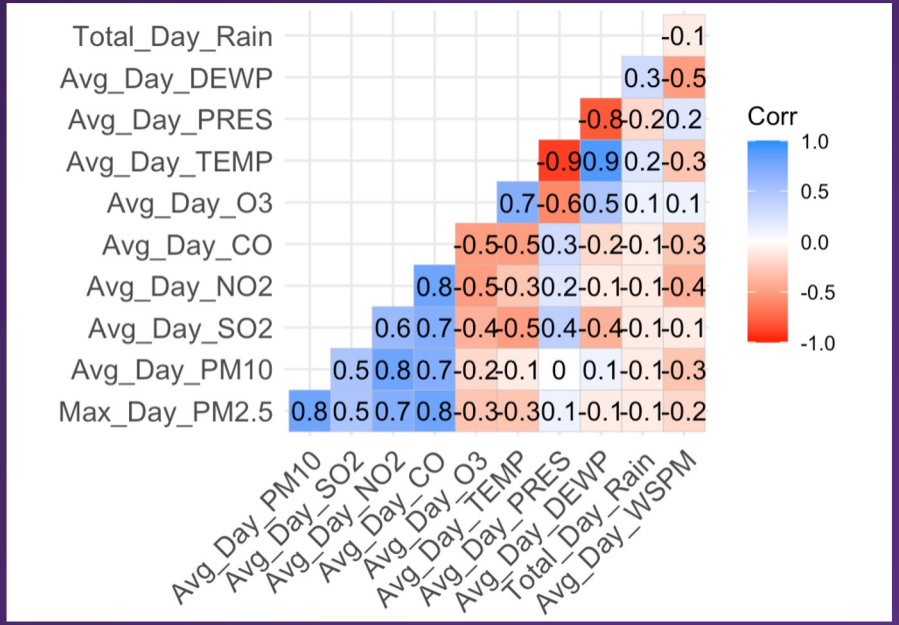
- Extra visualization: wind direction by year
- We can see that dangerous levels are particularly high for the “ESE” wind-speed (southeast)

Linear Regression Model

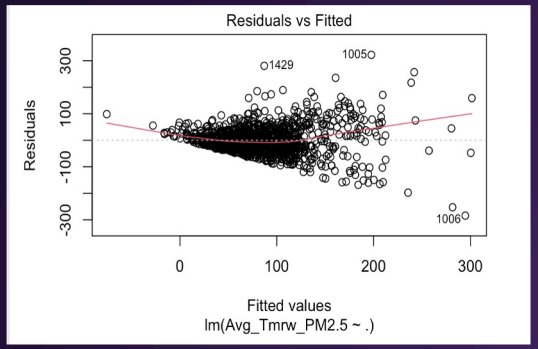
- Predict tomorrow's maximum $PM_{2.5}$ based on its yesterday's maximum $PM_{2.5}$ and various features values
- Split the dataset into train & test datasets - 80%:20%
- Fit models with diverse methods with the training dataset
    - Numeric Response: Variable Selection & Transformation / Regularization
    - Categorical Response: Multinomial and KNN
- Check their performances (R^2 / Adjusted R^2 / Level Prediction Accuracy) using test data and choose the best model
    - Higher the performances scores, the better that the model explains the data

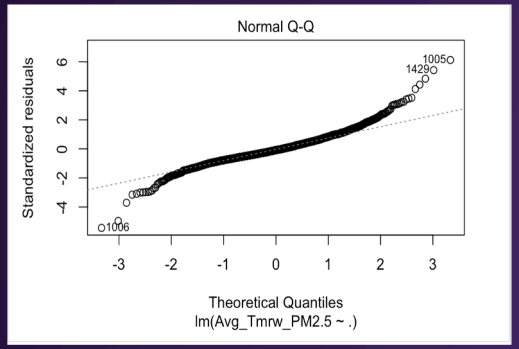
Correlations

- Correlation matrix heatmap
- Highly correlated with each other (Blue - Positive, Red - Negative)
  - Max_Day_$PM_{2.5}$- Avg_Day_CO (0.8)
  - Avg_Day_TEMP - Avg_Day_PRES (-0.9)
- Indication of Redundancy when fitting a model
  - May not add any meaningful predictive power to the model



Why Variable Selection & Transformation / Regularization?





- Possible threat of 'Overfitting'
  - Too many explanatory variables with highly correlated to each other - Redundant
  - Assumptions (ex. Constant Variance / Normality)
  - Outliers
- Yield a model with poor performance

➔ Find the best model by transformation & regularization

Variable Selection & Transformation / Regularization

1) **Selection Method (Stepwise):** Examined statistical significance of each independent variable
    a) Removed Variables
        i) Avg_Day_PM10
        ii) Avg_Day_DEWP
        iii) Freq_Day_wd
2) **Transformation Method (Box-Cox):** Transformed data to resemble a normal distribution
3) **Regularization Methods (LASSO / Ridge):** Shrunk the values towards a central point (mean)
    a) Removed Variables (both methods)
        i) Freq_Day_wd

Predicting Levels (Multinomial and KNN)

- Multinomial - Try different predictors

- KNN - Predicting with k nearest neighbors
  - Try model with k from 1 to 30

Performance Comparison & Reflection

Model	R^2	Adjusted R^2	Level Prediction Accuracy	Number of Underpredicted
Transformed	0.285	0.269	165/284 = 0.604	17
LASSO	0.365	0.347	178/284 = 0.627	19
Ridge	0.362	0.339	177/284 = 0.623	19
Multinomial	X	X	183/284 = 0.644	27
KNN	X	X	177/284 = 0.623	27

- Multinomial and KNN do not have R^2 and Adj. R^2 because they are predicting labels
- Similar performances by LASSO & Ridge
- Better performance by Multinomial
- What affects performance?
  - Imputing Missing values
  - Different measurements for explanatory variables
    - Median...
  - Explanatory variables selection
  - Train/Test split proportion

Possible Options when Choosing Models

Transformed Model

	Actual_Obs			
Predicted	Dangerous	High	Medium	Low
Dangerous	140	19	16	10
High	19	16	20	6
Medium	8	9	15	5
Low	0	0	0	0

Multinomial Model

	Actual_Obs			
Predicted	Dangerous	High	Medium	Low
Dangerous	156	29	24	13
High	0	0	0	0
Medium	10	13	26	7
Low	1	3	1	1

- The transformed model has less values that are under fitted than the multinomial model
  - Under fitted values → the model will illustrate medium or low levels when the actual concentrations are dangerous or high
  - It also performs better in predicting high and dangerous concentrations.

- However, the multinomial model has a higher accuracy.

Model Interpretation

- $PM_{2.5}$, SO_2, NO_2, O_3, TEMP, Rain, WSPM are significant in predicting PM2.5

- $PM_{2.5}$, NO_2, O_3 are positively correlated with the future $PM_{2.5}$
- SO_2, TEMP, Rain are negatively correlated with the future $PM_{2.5}$.

	Coefficients	P - Value
Max_Day_PM2.5	0.0141	<2e-16
Avg_Day_SO2	-0.0398	1.30e-05
Avg_Day_NO2	0.0650	<2e-16
Avg_Day_O3	0.0420	<2e-16
Avg_Day_TEMP	-0.203	<2e-16
Total_Day_Rain	-0.0930	2.27e-07
Avg_Day_WSPM	-1.436	4.54e-15

Summary of results

- Some of the models gave less promising results than others due to violations of various assumptions.
- **For the whole accuracy, the best model is the multinomial model.**
- **For the least underpredicted errors, the best model is the transformed model.**
- Higher concentrations of $PM_{2.5}$, NO_2, O_3 in the present will result in higher concentrations of $PM_{2.5}$ in the future
- Lower SO_2, temperature, rain and wind speed will also result in higher future $PM_{2.5}$ concentrations