STAT 443 Professional Statistics

Project Report: PM2.5 in Guangyuan

Client: Darren Glosemeyer

By Group 3

Wanxing Dai Yipin Luo

Darren Lim Bowei Ye

Introduction

With the development of industry and more concerns about healthcare, people are paying more and more attention to the changing air quality and how their lives may be at risk from it. Guangyuan is a city in Sichuan Province, China with an area of 16313.78 square kilometers (~276 Champaign) and a population of 2,484,123 in 2010. Guangyuan's economy is based on a diverse array of heavy industry, as well as mining and agriculture. This has been shown to cause many severely negative effects to the air quality in the area as harmful chemicals and particles are released into the air. PM2.5 is one of these harmful byproducts, as it is a particulate matter that is 2.5mm in diameter, which is about 3% the diameter of a human hair, which can be harmful to those who breathe it in high amounts. Studies have found a close link between exposure to fine particles and premature death from heart and lung disease. Fine particles like PM2.5 are also known to trigger or worsen chronic disease such as asthma, heart attack, bronchitis and other respiratory problems which can be an issue especially to those who have these conditions and live in these areas where they continually breathe these in as they live their daily lives.

We hope to raise public awareness about PM2.5 and its possible health concern, conduct public education about PM2.5 and set up an alert system on the next day's air quality by finding the season, month, day and time when PM 2.5 levels are typically higher and predicting what it may be like for the time that the user is searching for. We aim to predict the PM2.5 level of the next day by the hour in order to help citizens of Guangyuan decide whether they should be staying inside or not based on an alert system with 4 levels of classification of PM2.5 based on WHO air quality guidelines 2015: 0-35 for "Good air quality", 35-75 for "Moderate air quality",

75-150 for “Unhealthy air quality for sensitive groups”, and 150+ for “Unhealthy air quality for all individuals”.

We received this data from the Beijing Multi-Site Air-Quality Data Set in the UCI Machine Learning Database. We are convinced that this source is reliable as the meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017. Missing data are denoted as NA. There are 35064 samples in the dataset. We have 16 numeric variables, which are *No, year, month, day, hour, PM2.5, PM10, S02, N02, CO, O3, TEMP, PRES, DEWP, RAIN, WSPM*, and 2 categorical variables, which are *station, and wd*. The description of variables is listed below.

Variables	*functions*
No	Row number
year	Year of data in the row
month	month of data in the row
day	day of data in the row
hour	hour of data in the row
PM2.5	PM2.5 concentration (ug/m^3)
PM10	PM10 concentration (ug/m^3)
SO2	SO2 concentration (ug/m^3)
NO2	NO2 concentration (ug/m^3)
CO	CO concentration (ug/m^3)
O3	O3 concentration (ug/m^3)
TEMP	Temperature (degree Celsius)
PRES	Pressure (hPa)

DEWP	Dew point temperature (degree Celsius)
RAIN	Precipitation (mm)
wd	Wind direction
WSPM	Wind speed (m/s)
Station	Name of the air-quality monitoring site

Data Preprocessing

Nearly 8% of the data (2801 pieces) contains missing (NA) values between the variables. For the data visualizations, we removed the missing data and combined *Year, Month, Day, and Hour* variables into a single date variable in the format of year-month-day-hour. For the prediction, we filled the missing data with neighboring values. We removed the *wd* variable in KNN but contained *wd* in Lasso to increase prediction accuracy. Because we follow the threshold (35,75,150), we have four regions in total (0-35,35-75,75-150,150+). We created a new column named *Level* for indicating PM2.5's level, and we compute the prediction accuracy based on this *Level* parameter. For the Wind Direction *wd*, we transferred it from a character variable to a dummy variable (new columns) using 1 or 0 to indicate the wind direction for some very specific time point.

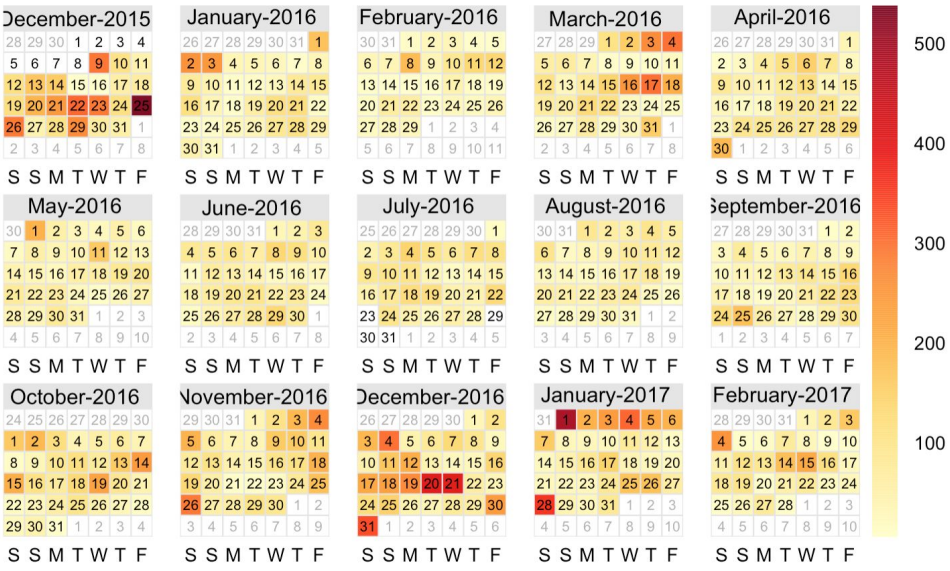
Method

We started with data visualization in order to understand how variables, especially PM2.5 change over time by plotting the PM2.5 and correlated values by year, season, month, day and hour. We tracked the values over time and how they fluctuated so that in this way we could see

trends in PM2.5 values and see what recommendations would make sense for a person who is living in this environment at those times.

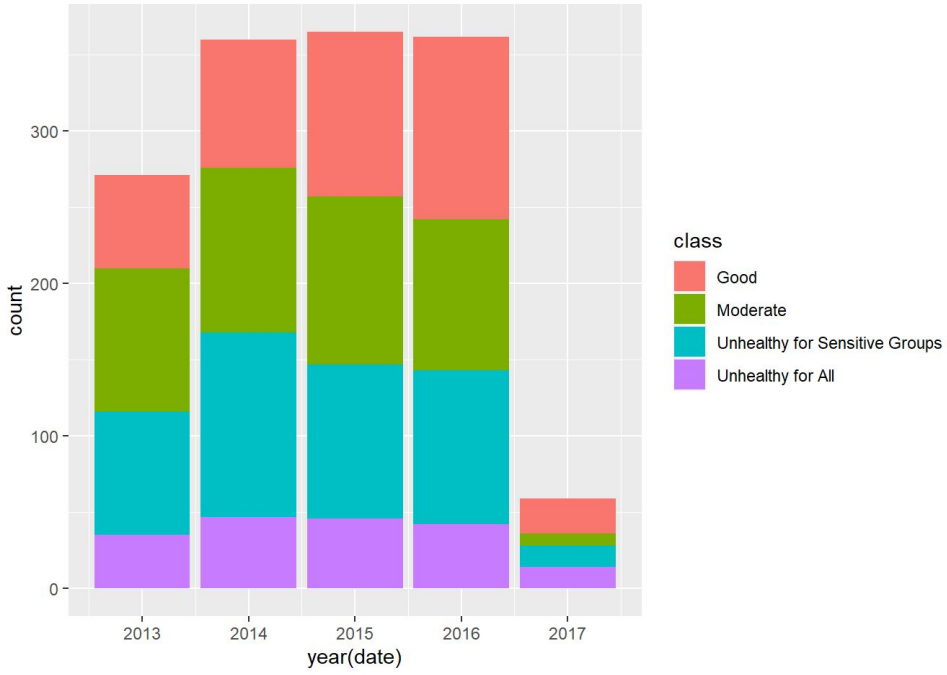
For prediction, we use Lasso (The Least Absolute Shrinkage and Selection operator), which could reduce some parameters to 0 for parameter selection and KNN (K-Nearest Neighbors algorithm). Because there is strong correlation among air quality data based on date, KNN could be a good option for extracting useful information and performing prediction. For Lasso, we are able to perform only short-term (one hour) predictions, but for KNN, we perform both short and long term(one day) predictions.

Results

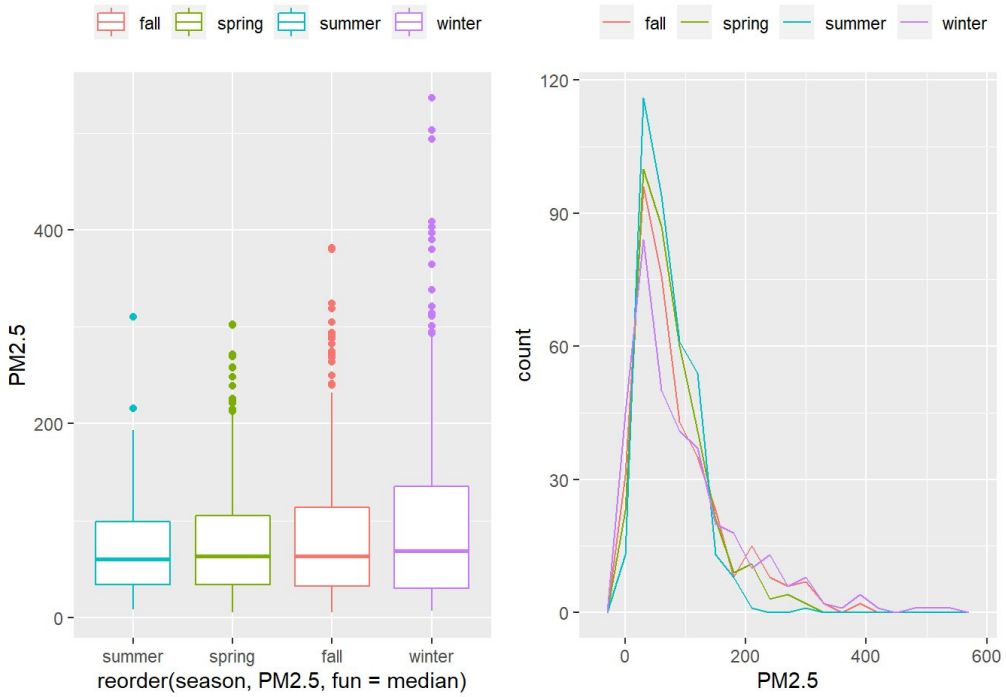


This is a calendar plot that highlights certain days of high average PM2.5 values in redder colors. From the highlighted portions that we can see for the year of 2016, the plot indicates that the

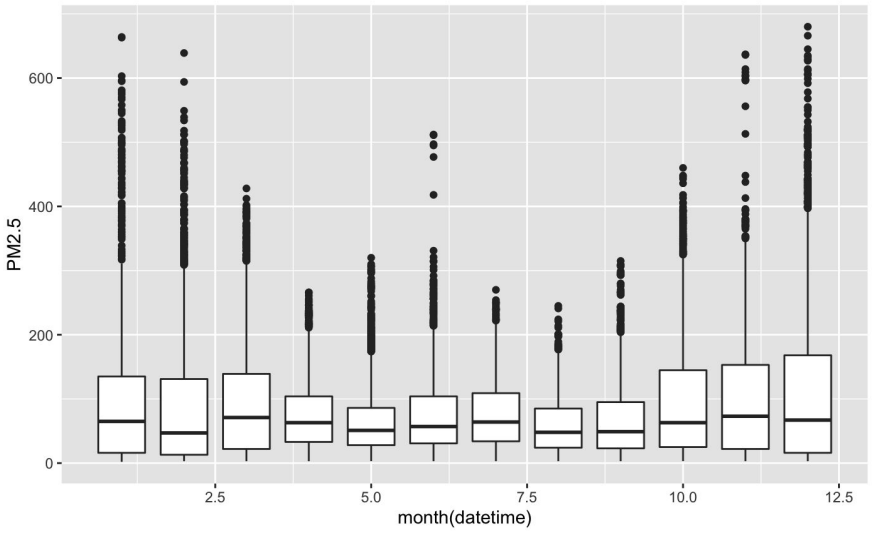
PM2.5 level is more severe during winter time, and even in March so we believe that winter time is usually the most dangerous time to go out.



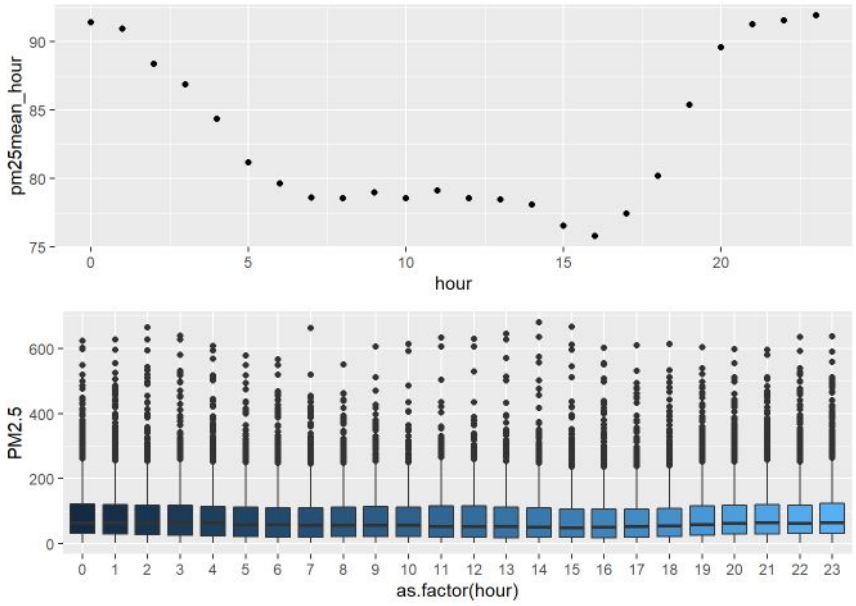
This graph shows a stacked barplot with frequencies of days within each level out of each year. We can see that out of the full years (the bars with almost 365 days each), the percentage of days with “Good” air quality has been increasing from 2014-2016 which shows that air quality has been slowly getting better over time.



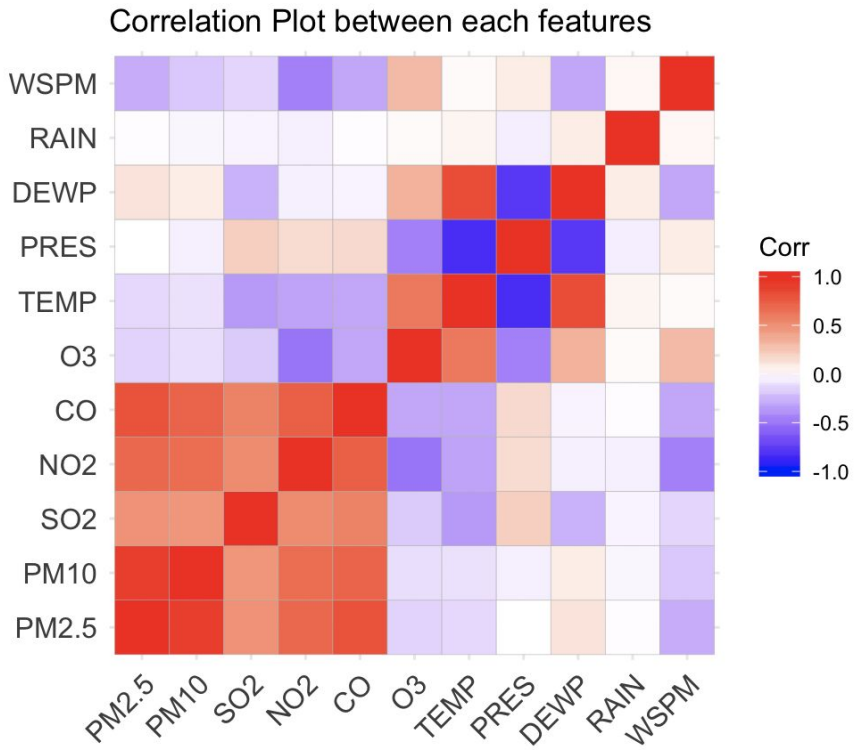
For this visualization we included a boxplot and a frequency line plot of PM2.5 values separated by season. The average PM2.5 value seems to be higher and has more outliers during the winter season which is corroborated by our earlier finding that the end of the year has higher PM2.5 values. There are also lesser frequencies of lower PM2.5 values and higher frequencies of dangerous PM2.5 values during winter as shown by the line plot, which is the exact opposite in the summer.



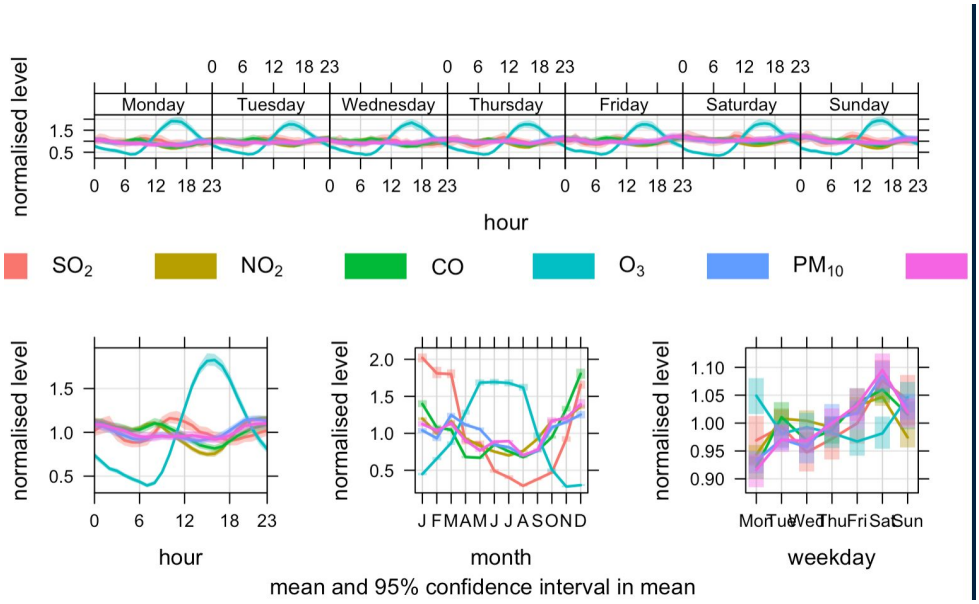
The boxplot contains PM2.5 values separated by month again but this time contains all the years unlike the calendar plot from before. This plot easily shows the highs and lows over the year easier than the calendar plot does and yet again shows us there are higher outliers near the beginning and ends of the year with a small spike in the middle of the year which is interesting as we saw summer months having less high values from the previous graph.



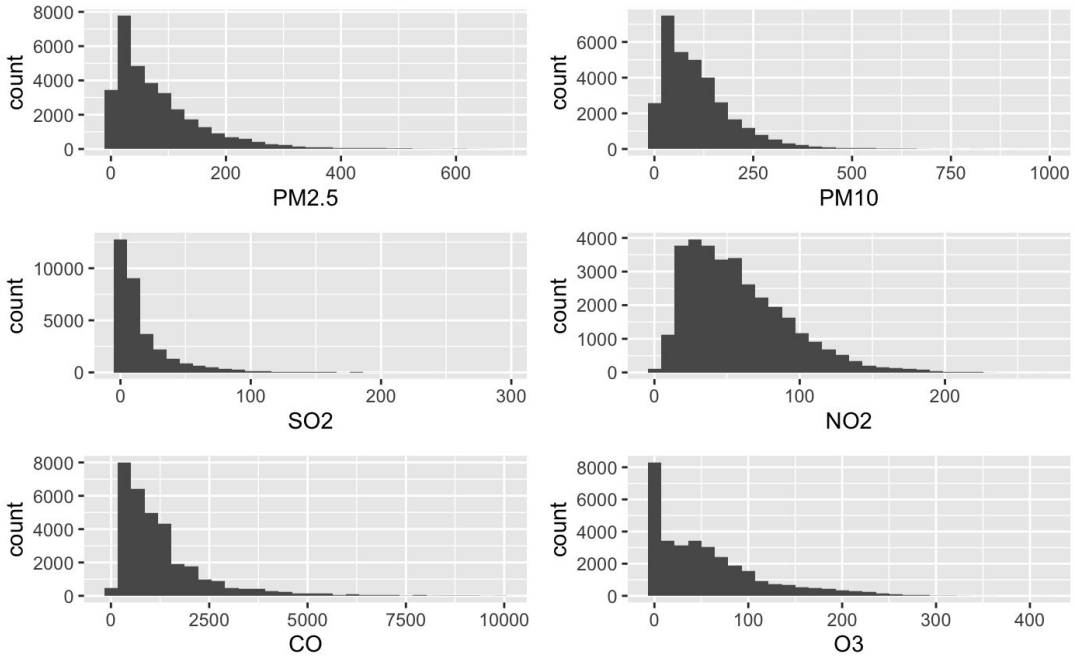
This box and scatter plot contains the PM2.5 values but this time separated by hour in the day. This one was really important as our prediction later on has a lot to do with predicting on an hourly short-term basis and we were able to see how PM2.5 could fluctuate so much even just over one day.



The correlation plot indicates that PM2.5 are closely related to PM10, SO2, NO2, and CO. This could show us which variables are more likely to correlate with PM2.5 levels for prediction purposes.



This graph shows trends of each air component based on time in terms of month, day, and even hour. As seen, we found that most of the components follow similar trends, except for O3 and SO2 for the month and O3 for the hour. This could help, especially seeing as we can see increases in all the components around January and December time as we saw was when PM 2.5 levels are highest during the year.

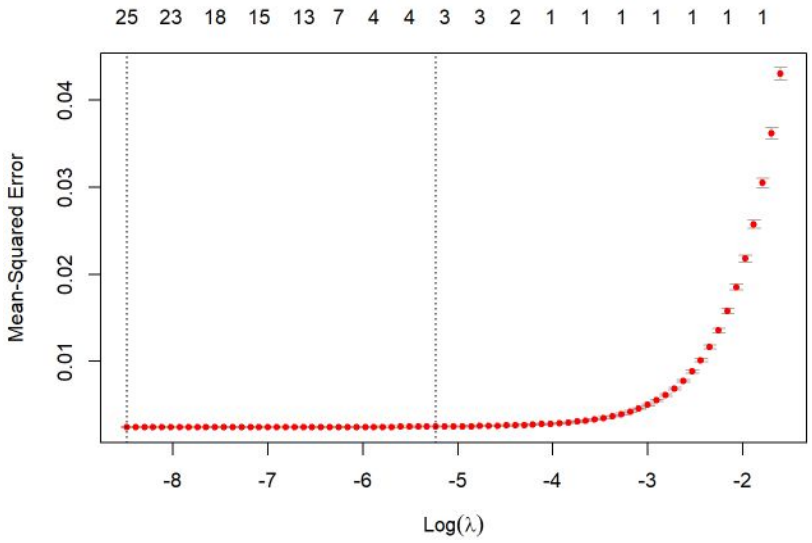


A simple frequency histogram that shows whether the components are more often in lower or higher concentrations. The distribution of levels of particles are all tailed and PM 2.5 concentrated around 50 which shows that they are tending towards lower numbers on average.

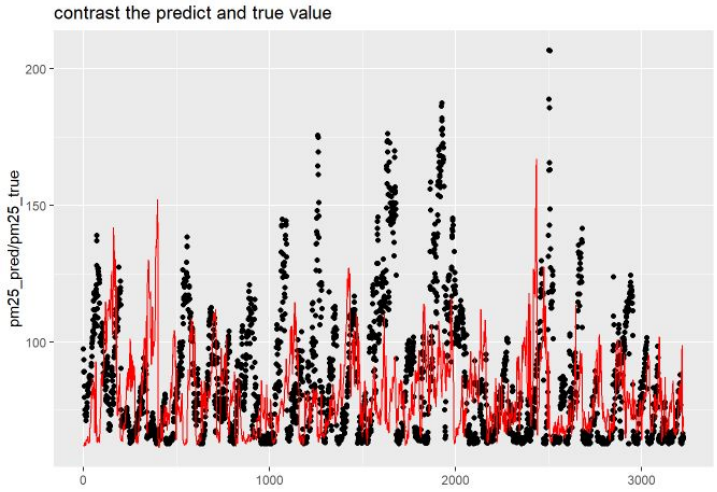
Lasso (Short term)

(Intercept)	-2.42078
PM2.5	0.90834
PM10	0.03149
SO2	0.01257
NO2	0.05238
CO	0.00017
O3	0.00703
TEMP	-0.11680
PRES	0.00047
DEWP	0.07897
RAIN	-0.63756
WSPM	-1.06867
wd_E	0.44815
wd_ENE	-0.60432
wd_ESE	0.48043
wd_N	-2.74825
wd_NE	-1.46707
wd_NNE	-2.38898
wd_NNW	-2.49342
wd_NW	-2.70033
wd_S	0.74164
wd_SE	0.05365
wd_SSE	0.41119
wd_SSW	0.00000
wd_SW	1.43011
wd_W	0.00000
wd_WNW	-0.65157
wd_WSW	1.87714

We checked the coefficients for the terms that were used in the Lasso Regression. Wind direction variables which we at first thought to be useless from our earlier analysis after parameter selection, we actually found that the wind direction does have impact on PM2.5 in Lasso regression so we ended up creating dummy variables to use in the prediction instead.



Based on Mean-Squared error, the minimum value of lambda is 0.0002060101, showing that the model fits well as a prediction.

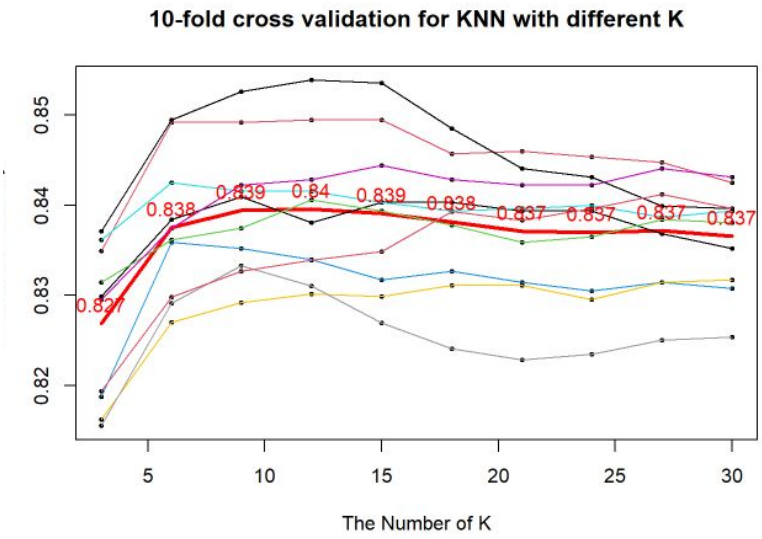


Above is our plotted Lasso Regression model. The red line is the predicted values, while the black scatterplot behind it are the true values for the time period that we are predicting on. We computed accuracy by comparing the predicted value's level and real value, which we got to be around 73% for Lasso in one-hour PM2.5 prediction.

KNN

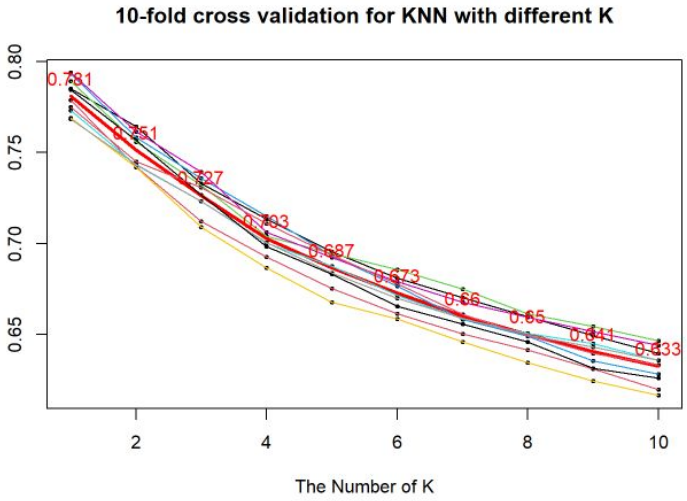
We used ten fold cross-validation to demonstrate the accuracy of the prediction model, and we also selected ten different K for deciding the best K numbers. Because the accuracy of a model with wind direction is worse than the model without it, we exclude the 16 generated features of wind direction in KNN. For the prediction based on KNN, we are able to predict the response of the PM2.5 value in the next one hour (short-term) or 24 hours (long-term).

Short term prediction



The red line is the average prediction accuracy among ten-folds cross-validation. The short-term accuracy based on KNN is better than Lasso for about ten percent as it averages around 83.7% for all values of K.

Long term prediction



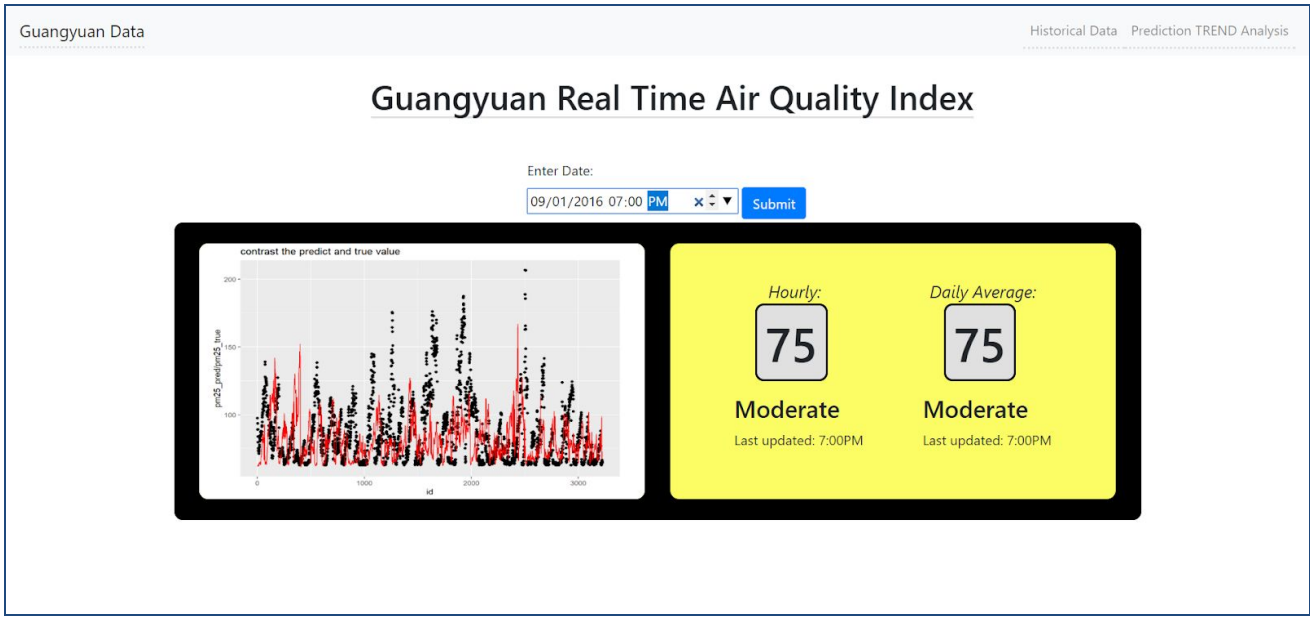
Again, the red line with the values is the average prediction accuracy among ten fold cross-validation. The long-term accuracy based on KNN has a decreasing trend, and this may be because of the relation of PM2.5 among different time points. Namely, the closer, the better for prediction.

Overall, based on the short term prediction, the KNN does better than Lasso but with more computation. For the long term prediction, if we want to perform KNN for one week or longer, it's better to not include too many neighbors.

Alert System

In order to warn the general public of incoming high PM2.5 values for the coming days, we decided to create a website that would be able to track the PM2.5 values using the above

prediction methods and output the predicted value for the time that the user puts in. We decided to create a website over a shiny app as we believed that it would be much more easily accessible to the public and with less barrier of entry in terms of usage. With the website, there is just a simple time and date input that the user puts in that will determine the output of what the hourly value of the PM2.5 will be at that exact time and what the daily average would be for that day that was input. The user will easily be able to inform themselves this way whether or not to go out and how to plan their day accordingly.



Conclusion and Discussion

Based on our findings, outdoor activity during the winter is not recommended since PM2.5 tends to be extremely high in November, December, January, and February based on our seasonal analysis. On average, the PM2.5 level tends to be the lowest between 3pm to 4pm, and the peak is around the midnight, so these would be times that could be considered most safe to

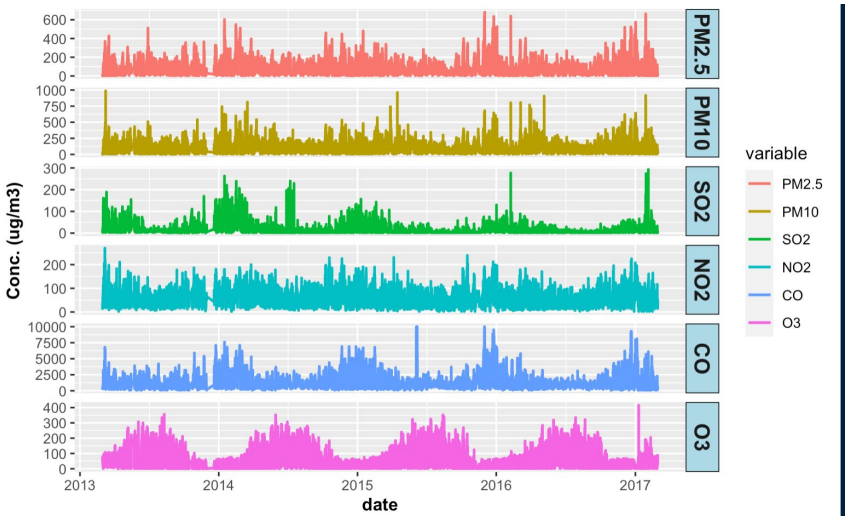
go outside. When PM2.5 level is high, we recommend people to stay indoor and close all windows and openings that allow polluted air to enter if possible, and to turn on an air purifiers that are equipped with a HEPA (High Efficiency Particulate Air) filter. For people who work on the road and must drive for long periods of time in all weather conditions, get an air purifier for your car that comes with at least HEPA and activated carbon filters. There is also the option of boosting your body's resistance against PM2.5 by increasing your intake of nutrients like fish oil, Vitamin C and Vitamin E which can all help strengthen the body against inhalation of PM2.5. If you must go outdoors, make it short and quick, and wear a N95 or higher face mask. Based on full year data from 2014 to 2016, we could find the overall air quality especially PM2.5 is becoming better and better with each year, which is a good trend, but it is still important to stay safe when there is a large volume of particulate matter in the air.

For the prediction part, short term (hourly) prediction is more reliable than long term prediction (daily). If we want to predict PM2.5 based on a very long time interval, the prediction model should be tailored with many pretraining assumptions.

Although the alert system is still in a pretty basic state, we believe that it gets the job done for citizens in the area who need to know the information about PM2.5 fast and accessibly. They do not need to read any charts or graphs and the number is there for them to see with colors to warn them. In the future it could possibly be improved by creating weekly forecasts based on highly trained prediction models and even include the public awareness information we were hoping to spread on the homepage or about pages for the website.

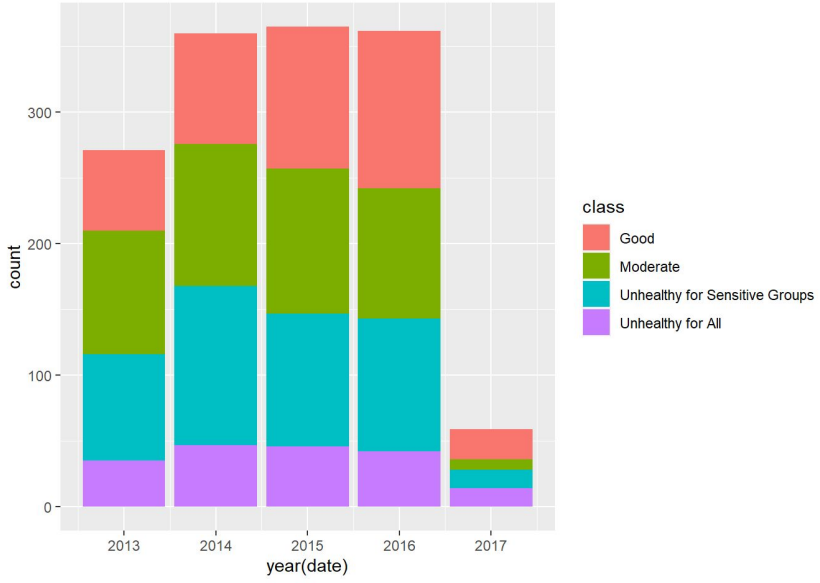
Appendices

1. **PM2.5, PM10, SO2, NO2, CO, O3 over time**



The above three plots are time series plots that could help us understand the trend of the main components of the air pollution based on timeline. We have a similar plot as part of the appendix below, but we think this one has brighter visuals and is easier to analyze at a glance.

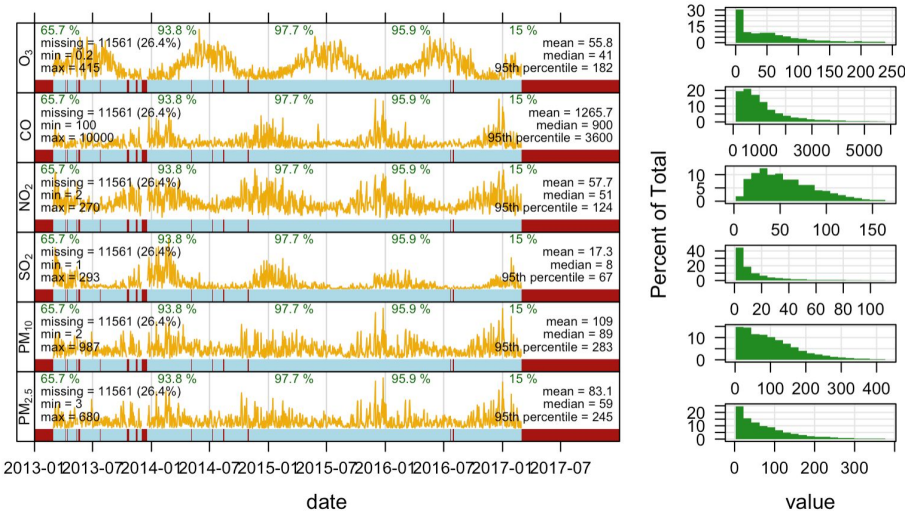
2. **Percentage of days with "Good" air quality 2014-2017**



Since 2013 and 2017 have incomplete data, we only look at the data from 2014 to 2017.

Percentage of days with “Good” air quality has been increasing from 2014-2017

3. Distribution of PM2.5, PM10, SO2, NO2, CO, O3



These graphs depict the concentration levels of each variable: O3, CO, NO2, SO2, PM10 and PM2.5 over time. It also has a graph on the right which counts the frequencies of values of each variable as well to see if there’s any strong reason to believe these variables are highly correlated with PM2.5.

Reference

Beijing Multi-Site Air-Quality Data Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

The geographical location of Guangyuan in China. Retrieved from https://www.researchgate.net/figure/The-geographical-location-of-Guangyuan-in-China_fig1_328612842

What is PM2.5 and Why You Should Care. Retrieved from https://blissair.com/what-is-pm-2-5.htm

Air Quality Index Scale and Color Legend. Retrieved from https://aqicn.org/scale/

Code & Data

See attached Rmd and html files for code and alert system.