Visualization of
Beijing Multi-Site Air Quality

REPORT BY,

Aihan Liu

DATS 6401 Visualization of Complex Data
Professor Reza Jafari
May 4, 2022

Contents

Abstract

The human body inhales oxygen necessary for life from the air to maintain normal physiological activities. Therefore, the standard chemical composition of the air is an essential condition to ensure the physiological function and health of the human body. This project is based on the hourly air pollutants data from 12 nationally controlled air-quality monitoring sites in Beijing, China. Use different graphs to show the changes in various air indicators from 2013 to 2017.

INTRODUCTION

Breathing clean air can lessen the possibility of disease from stroke, heart disease, lung cancer, and chronic and acute respiratory illnesses such as asthma. Lower levels of air pollution are better for heart and respiratory health, both long- and short-term.

Air quality reflects the degree of air pollution, which is judged according to the concentration of pollutants in the air. Air pollution is a complex phenomenon, and the concentration of air pollutants at a given time and place is affected by many factors. Artificial pollutant emissions from stationary and mobile pollution sources are among the most critical factors affecting air quality, including exhaust from vehicles, ships, and airplanes, industrial pollution, residential living and heating, and waste incineration. Urban development density, topography, and weather also affect air quality.

This project based on the hourly air pollutants data from 12 nationally controlled air-quality monitoring sites in Beijing, China. Use different graphs to show the changes in various air indicators from 2013 to 2017.

The student used different plots and graphs to show the data, and also created an interactive dash for data visualization.

DATA DESCRIPTION

The dataset related to this project is the Beijing Multi-Site Air-Quality Data Data Set which is provided in UCI Machine Learning Repository [1],

The city of Beijing established an air pollution monitoring network in January 2013 as part of the national monitoring network [2]. There are 26 air-quality monitoring sites, 15 meteorological sites, and 12 national controlled sites in Beijing, and they are marked in the below picture.

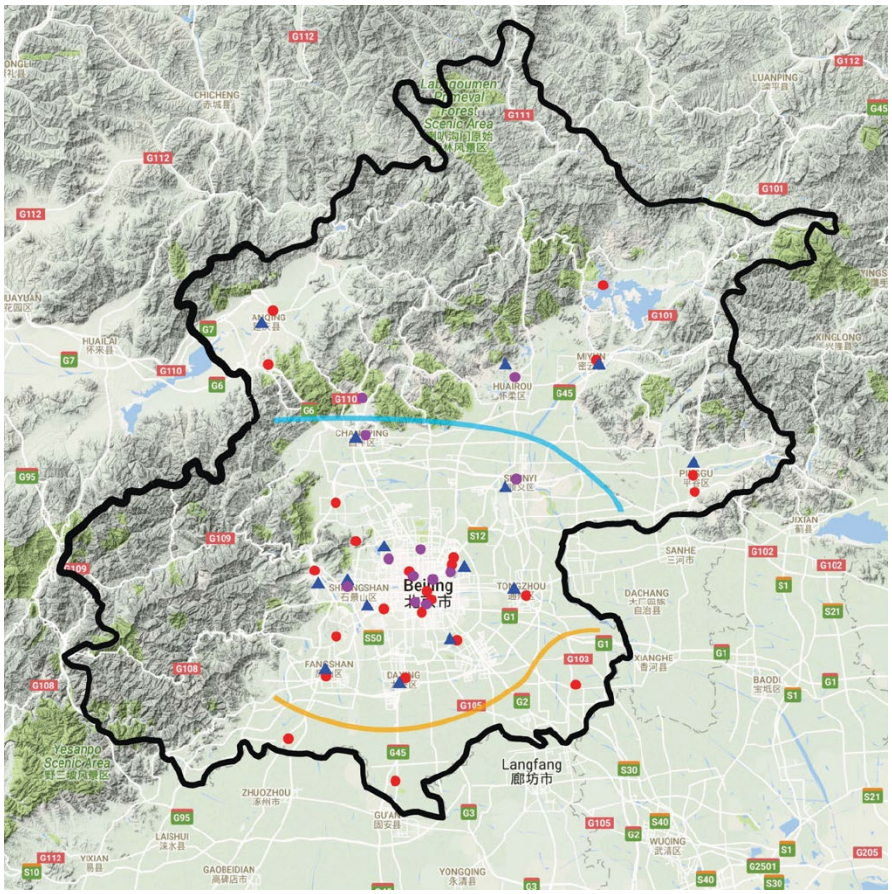

Figure 1 Locations of the 36 air-quality monitoring sites (red or purple dots) and 15 meteorological sites (blue triangles) in Beijing. The purple dots mark the 12 Nationally controlled sites, and the red dots show the other 24 sites.

The student's data set in this project includes hourly air pollutants data from 12 nationally controlled air-quality monitoring sites from March 1, 2013, to February 28, 2017. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in

each air-quality site are matched with the nearest weather station from the China Meteorological Administration.

The dataset contains 18 variables which show below:

no	name	description
1	No	row number
2	year	Year of date
3	month	Month of date
4	day	Day of date
5	hour	Hour of date
6	PM2.5	PM2.5 concentration (ug/m^3)
7	PM10	PM10 concentration (ug/m^3)
8	SO2	PM10 concentration (ug/m^3)
9	NO2	PM10 concentration (ug/m^3)
10	CO	PM10 concentration (ug/m^3)
11	O3	PM10 concentration (ug/m^3)
12	TEMP	temperature (degree Celsius)
13	PRES	pressure (hPa)
14	DEWP	dew point temperature (degree Celsius)
15	RAIN	precipitation (mm)
16	wd	wind direction
17	WSPM	wind speed (m/s)
18	station	name of the air-quality monitoring site

Table 1 Attribute information

The dependent variable in this data is PM2.5. PM2.5 refers to fine particulate matter with an aerodynamic diameter of less than 2.5μm. Both PM2.5 and PM10 can penetrate deep into the lungs, but PM2.5 can even enter the bloodstream, mainly affecting the cardiovascular and respiratory systems and other organs. PM is primarily produced by fuel combustion in different sectors, including transport, energy, households, industry, and agriculture. In 2013, the World Health Organization's International Agency for Research on Cancer (IARC) classified outdoor air pollution and particulate matter as carcinogens.
The independent variables are the left.

PRE-PROCESSING

Merge data: The original data is saved in different files based on the different stations. The student first merges these files into a single file named 'All.csv'. The original merged file contains 10 features and 420,768 observations.

Remove missing value: As the description of the dataset shown, missing data are denoted as NA. The student removes the missing value by using Pandas build-in function dropna(). I remove the rows that contain the missing value. After removing the missing value, there are 382,168 observations left.

Date format: The date is saved into four columns: 'year', 'month', 'day', and 'hour'. The student creates another column that contains all the date columns together.

OUTLIER DETECTION

An outlier is an observation that lies at an abnormal distance from other values in a random sample of a population. It could be calculated by the lower and upper quartiles (defined as the 25^{th} and 75^{th} percentiles). If the lower quartile is defined as Q1 and the upper quartile as Q3, the difference (Q3-Q1) is called the interquartile range or IQ. The lower inner fence is Q1-1.5IQ, and the upper inner fence is Q3+1.5IQ. The lower outer fence is Q1-3IQ, and the upper outer fence is Q1+3IQ. Any value beyond the inner fence on either side is considered a mild outlier. Any value beyond the outer fence is conditioned as an extreme outlier.

It is another way to determine outliers. The further away an observation's Z-score is from zero, the more unusual it is. A standard cut-off value for finding outliers is Z-scores of +/- 3 or further from zero. Z-score can be calculated as $Z = \frac{X-\mu}{\sigma}$. After removing the missing value, there are 348,705 observations left.

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is the process of computing the principal components that are most commonly used to reduce the dimension of the dataset. It computes a matrix that represents data variance and ranks them by relevance. In Python, PCA could be implemented in the sklearn.preprocessing package easily.

Singular value decomposition (SVD) is a matrix factorization method that generalizes the eigendecomposition of a square matrix to any matrix. SVD is more general than PCA. The singular values in SVD assume the correlation between one or more features. It could be calculated by calculating the H martix and then using the NumPy package by np.linalg.svd(H). Any singular values close to zero mean one or more features are correlated.

Normalize Feature Space: The student first sliced the dataframe into a new one containing only the numerical variables and then applied the StandardScaler function in the sklearn.preprocessing package to the sliced dataframe. Related code shows below:

```
numeric_col = df[df._get_numeric_data().columns.to_list()[6:]]   numeric_col: PM10   SO2   NO2
X = StandardScaler().fit_transform(numeric_col)   X: [[-1.23799113 -0.59534536 -1.33176805 ...
```

Figure 2 code for normalized data

Explained Variance Ratio: After applying PCA with Minka's MLE to guess the dimension in the original feature space, the transformed dataframe contains four components with explained variance ratio shown in figure 4.4.4.

```
Original Dim (346589, 11)
Transformed Dim (346589, 10)
explained variance ratio [0.32062115 0.23509271 0.1125351  0.09051912 0.08708915 0.04692703
 0.03364179 0.02699112 0.02294824 0.01692678]
```

Figure 3-1 Console output of explained variance ratio for original data

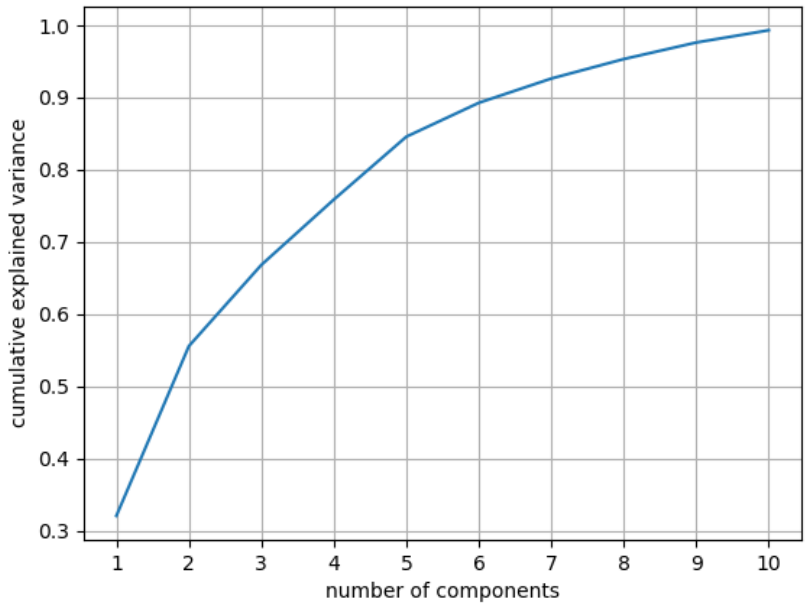

Figure 3-2 Explained variance ratio plot

The explained variance ratio and the plot show that the first seven components contain more than 90% of the variance (0.32062115+0.23509271+0.1125351+0.09051912+0.08708915+0.04692703+0.03364179= 0.92642605), so the first seven components should be maintained, the last one should be removed from the PCA analysis.

Singular Values and Condition Number

```
Original Data: singular Values [1222361.40308876  896286.01674235  429037.7146412   345102.25114317
  332025.57190611  178908.3019987   128258.62102764  102903.07013421
   87489.66666368   64532.99834865   25573.38430552]
Original Data: condition number 6.913623284152144

Reduced Dimensions Data: singular Values [1222361.40308875  896286.01674235  429037.7146412   345102.25114317
  332025.57190611  178908.3019987   128258.62102764  102903.07013421
   87489.66666368   64532.99834865]
Reduced Dimensions Data: condition number 4.352200476941864
```

Figure 4 Singular values and conditional number for original data and reduced dimensions data

The minimal singular value is 102903.070. the larger the singular value, the better the data.

The condition number for the original data is 6.91365, and for the reduced dimensions data is 4.35220. They both indicated a weak degree of co-linearity.

NORMALITY TEST

In scientific research, it is often necessary to perform a difference test on the data, and the commonly used parametric test requires the data to obey a normal distribution. Therefore, it is required to test the data for normality before deciding whether to use a normality test.

Normality test is a nonparametric test, and there are many test methods, and the basic principles and applicable conditions of each test method are different. Among all normality testing methods, its

The basic principle can be mainly explained from the following aspects:

1. Directly compares the data with a standard normal distribution curve by plotting a histogram of the data. Quantile-quantile (Q-Q) plot is the most common graph in this test.
2. Through probability statistics and hypothesis testing, to determine whether the data obeys the normal distribution, you can calculate the hypothesis test with the null hypothesis of "data obeys the normal distribution" and construct relevant statistics to calculate the test result.

   Based on empirical distribution functions, such as the Kolmogorov-Smirnov test (K-S test). It is one of the most used normality test methods that mainly calculates the distance between the empirical and theoretical distribution and uses the most significant distance as the test statistic.

   The D'Agostino's K-squared test (Skewness-Kurtosis test). It mainly quantifies the difference and asymmetry between the data distribution curve and the standard normal distribution curve by calculating the skewness and kurtosis and then calculates the degree

of difference between these values and the expected value of the normal distribution.

The Shapiro–Wilk test is one of the most effective methods for normality testing, testing normality in frequentist statistics. It works when the sample size is smaller than 5000.

From the above analysis, the student applied the Q-Q plot and K-S test to show the normality for the continuous data.

```
for idx in colindex:
    data = df.iloc[:, idx]
    names = colnames[idx]
    norm_test = stats.kstest(data, 'norm')
    print(f'{names} K-S test: statistics= {norm_test[0]:.3f} p-value = {norm_test[1]:.2f}.')
    if norm_test[1] > alpha:
        print(f'{names} is normal in the {alpha} level.')
    else:
        print(f'{names} is not normal in the {alpha} level.')
```

Figure 5 Sample code for k-s test

	statistics	p-value	result
PM2.5	0.999	0.00	Not normal
PM10	0.998	0.00	Not normal
SO2	0.969	0.00	Not normal
NO2	0.988	0.00	Not normal
CO	1.000	0.00	Not normal
O3	0.960	0.00	Not normal
TEMP	0.780	0.00	Not normal
PRES	1.000	0.00	Not normal
DEWP	0.518	0.00	Not normal
RAIN	0.500	0.00	Not normal
WSPM	0.632	0.00	Not normal

Table 2 Results of K-S test

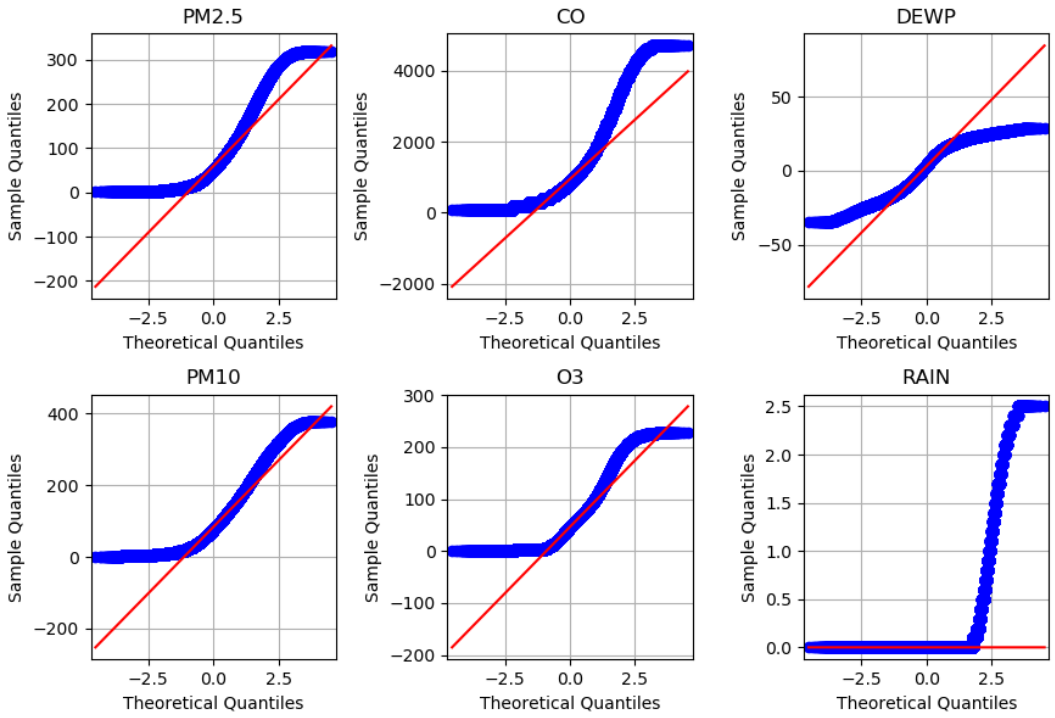

Figure 6-1 Q-Q plot

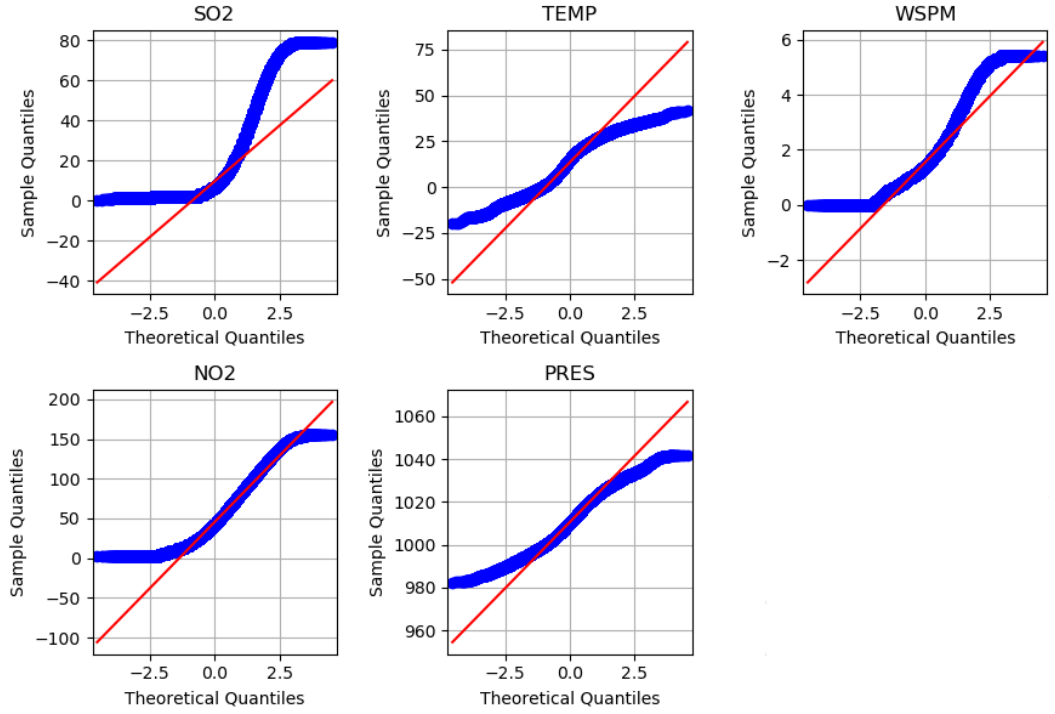

Figure 6-2 Q-Q plot

Both K-S test and the Q-Q plot shows that all the continuous data are not normally distributed. Since it is a time series problem, there is no need to transform the data to normal distribution for visualization.

HEATMAP AND PEARSON CORRELATION

The correlation for the continuous variables shows below:

	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	WSPM
PM2.5	1.00	0.88	0.45	0.60	0.76	-0.12	-0.04	-0.06	0.20	-0.03	-0.28
PM10	0.88	1.00	0.45	0.62	0.67	-0.09	-0.00	-0.09	0.16	-0.05	-0.22
SO2	0.45	0.45	1.00	0.48	0.55	-0.16	-0.33	0.24	-0.30	-0.08	-0.08
NO2	0.60	0.62	0.48	1.00	0.67	-0.50	-0.22	0.14	0.01	-0.05	-0.40
CO	0.76	0.67	0.55	0.67	1.00	-0.34	-0.28	0.15	-0.00	-0.01	-0.33
O3	-0.12	-0.09	-0.16	-0.50	-0.34	1.00	0.58	-0.43	0.29	-0.01	0.32
TEMP	-0.04	-0.00	-0.33	-0.22	-0.28	0.58	1.00	-0.81	0.82	0.02	0.02
PRES	-0.06	-0.09	0.24	0.14	0.15	-0.43	-0.81	1.00	-0.75	-0.05	0.10
DEWP	0.20	0.16	-0.30	0.01	-0.00	0.29	0.82	-0.75	1.00	0.11	-0.30
RAIN	-0.03	-0.05	-0.08	-0.05	-0.01	-0.01	0.02	-0.05	0.11	1.00	-0.02
WSPM	-0.28	-0.22	-0.08	-0.40	-0.33	0.32	0.02	0.10	-0.30	-0.02	1.00

Figure 7 Pearson correlation

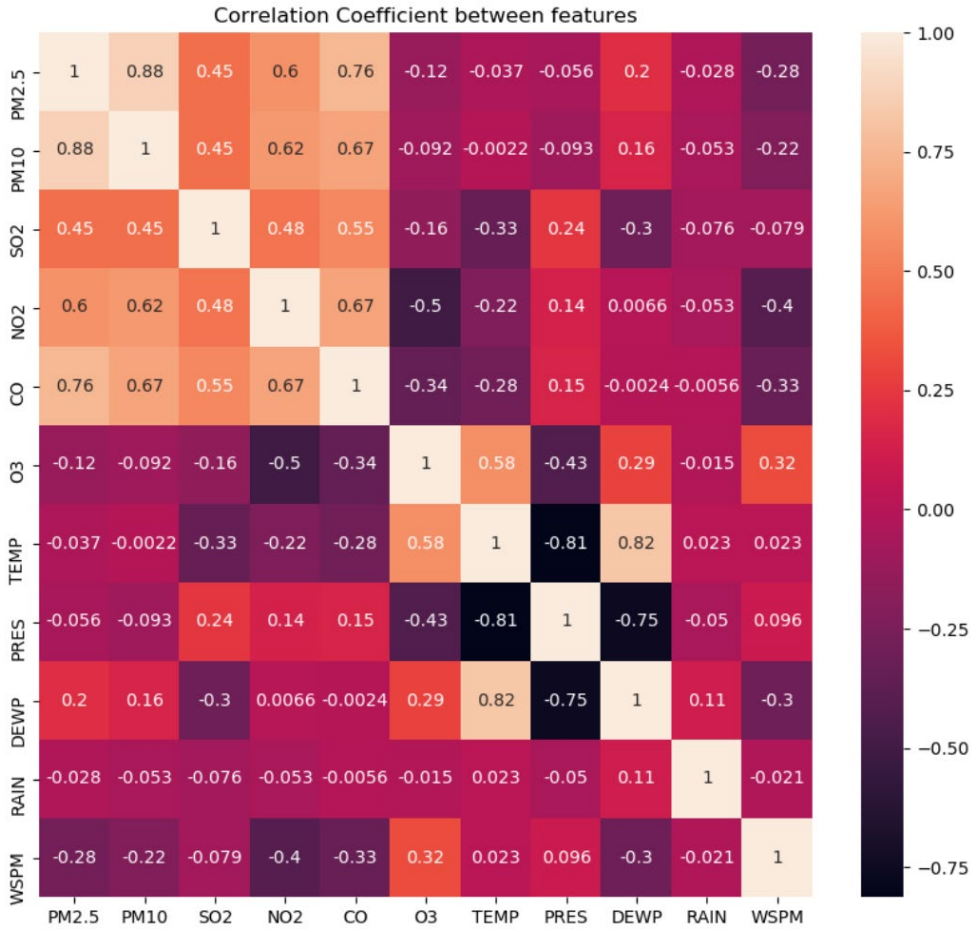

Figure 8 Heatmap for correlation coefficient

The pearson correlation and heatmap for correlation coefficient both gives the same result of correlation between continuous variables. It shows that TEMP and DEWP are highly negative correlated with PRES. And TEMP and DEWP have positive correlation (they are both temperature measurements). For the chemical content, apart from O3, others are positive correlated with each other. The target variable PM2.5 has highly positive relationship with PM10, SO2, NO2, CO, but not very correlated with other meteorological variables.

DATA VISUALIZATION

To visualize the data correctly, the student sliced the data in February 2017 and selected four stations, including Aotizhongxin, Dongsi, Gucheng, and Wanliu, which indicated the four districts in the urban Beijing.

Line plot

The line plot shows the PM2.5 changes for four different stations in a month. There is no clear trend for time series data, and these four-station shows similar changes.

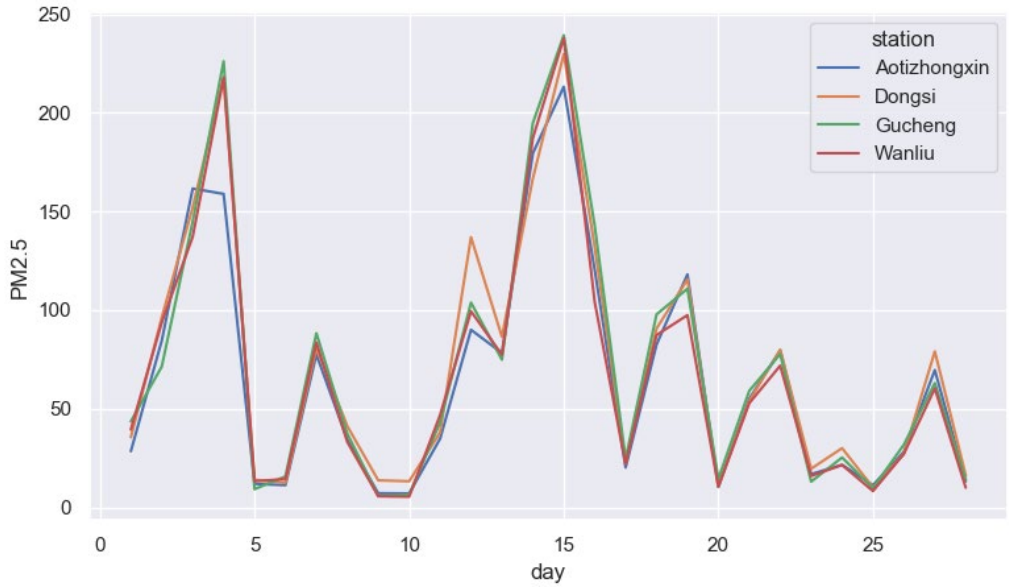

Figure 9 Line plot for four stations of PM2.5

Bar plot

The grouped and stack bar plot shows the frequency of PM2.5 in a month. It shows that most of the time, PM2.5 is lower than 3.5. However, some hours also show that the concentration of PM2.5 is more extensive than 250, which indicates that it is severe pollution.

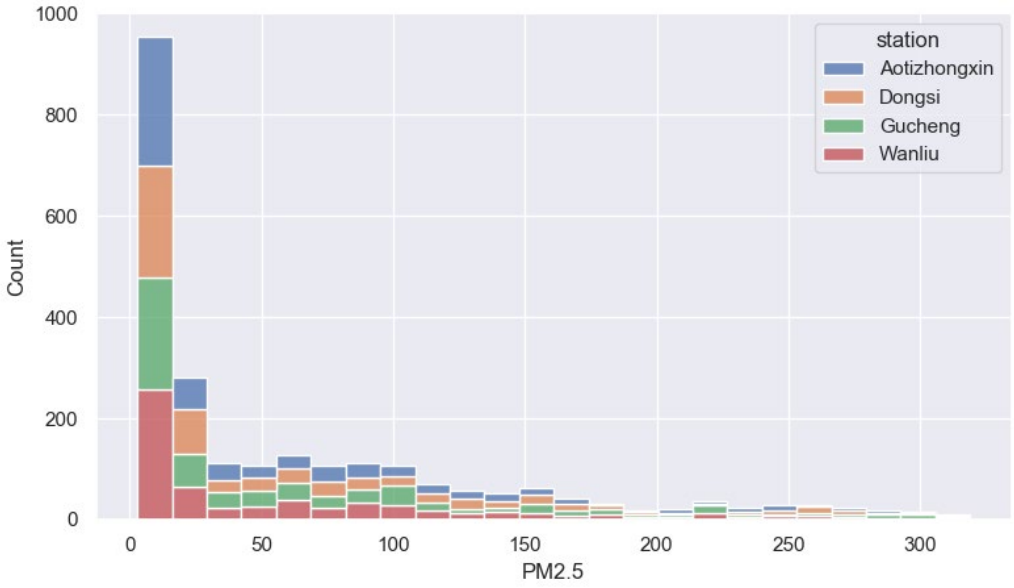

Figure 10 Stack bar plot

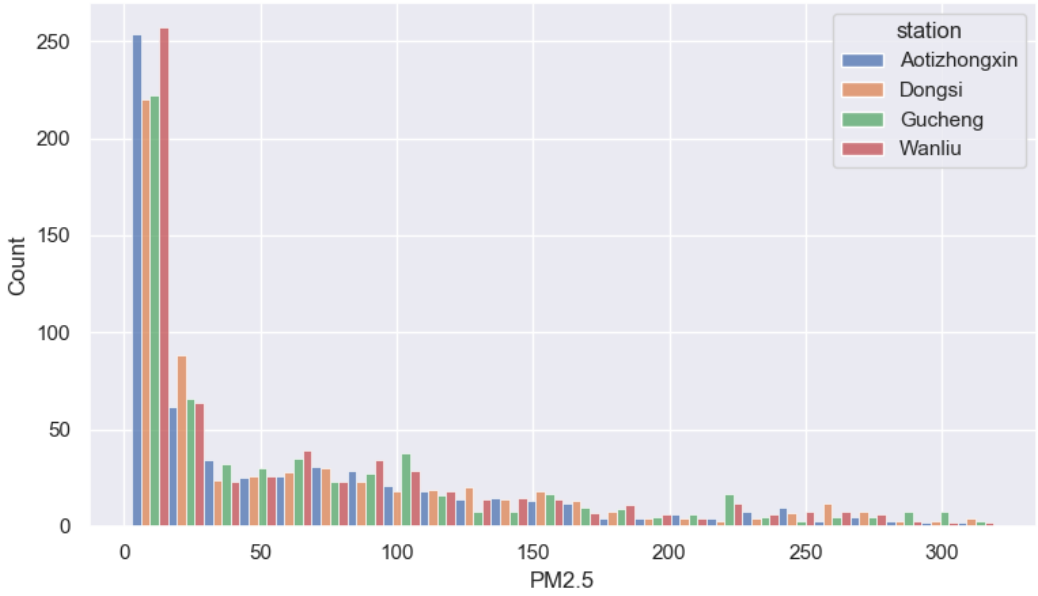

Figure 11 Group bar plot

Count plot

The count plot shows the count of wind directions in February 2017. The most significant number of northwesterly winds. The most wind came from

north directions (NNE, N, NE, ENE, NW, NNW, WNW).

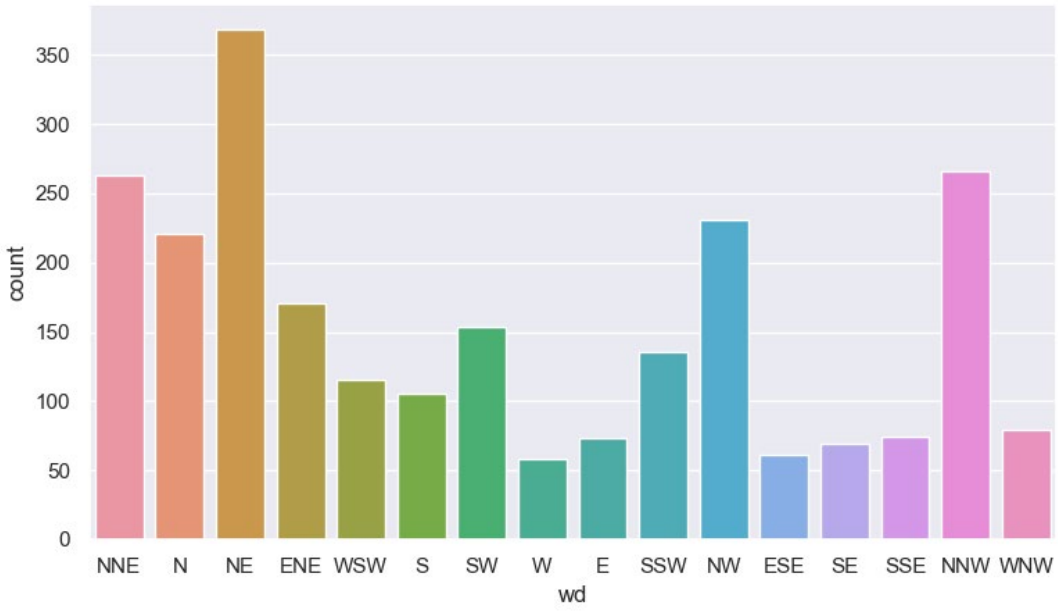

Figure 12 Count plot for winds

Cat-plot

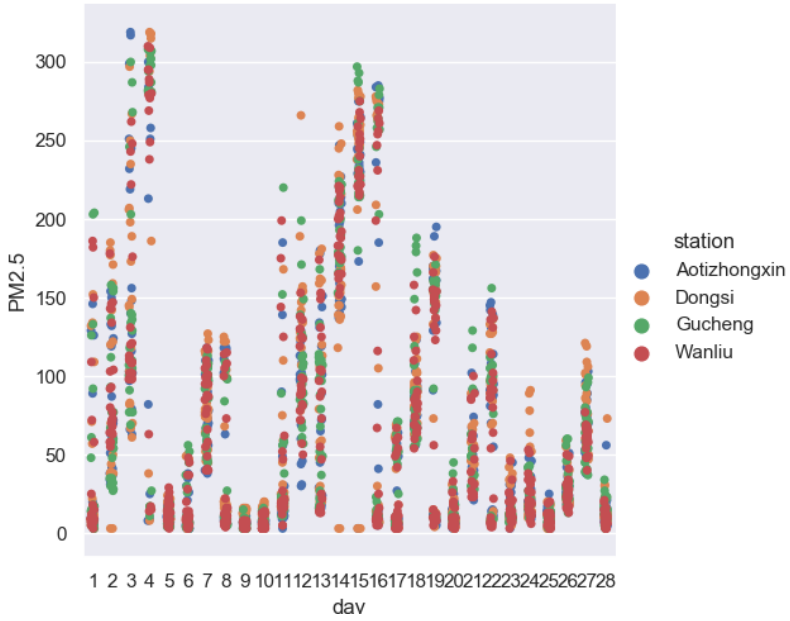

Figure 13 Cat-plot of PM2.5 for four stations

The cat-plot gives the PM2.5 distribution of four stations in a month. The high PM2.5 happened at the beginning and middle of this month.

Pie chart

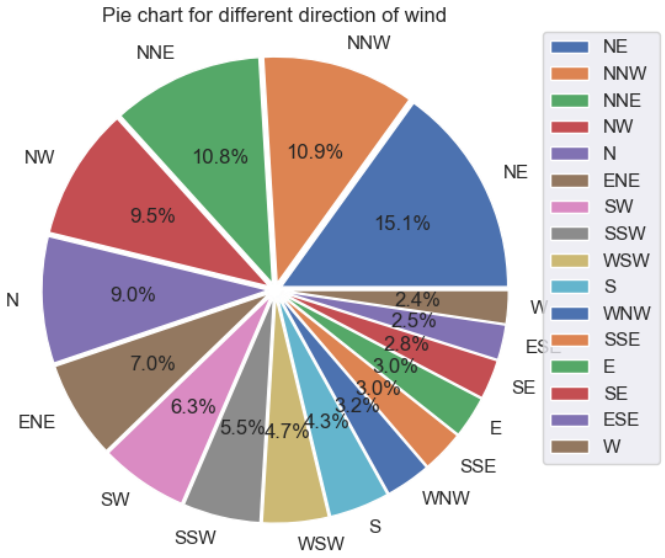

Figure 14 Pie chart for wind direction

The pie chart shows the number of wind directions. It gives the same result as the count plot. Northeast wind took the highest proportion, followed by North-Northwest and North-Northeast.

Displot

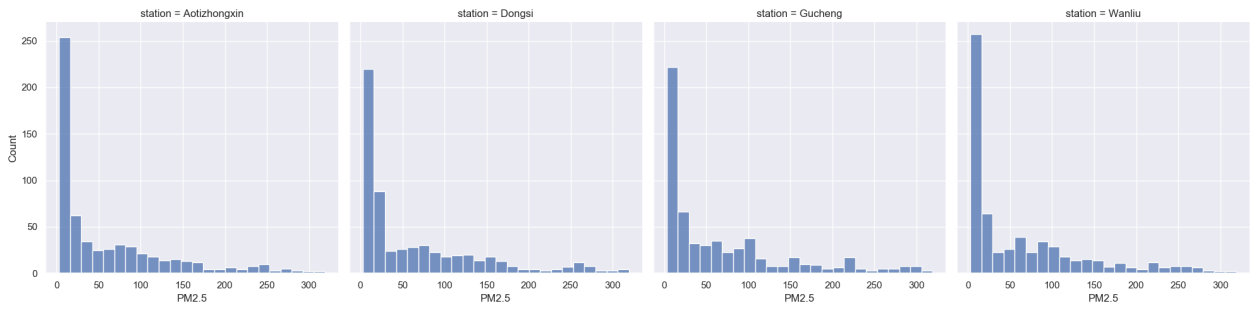

Figure 15 Displot for PM2.5 in different stations

The displots show the PM2.5 counts for four stations. It indicates that Gucheng and Wanliu had a larger count of high PM2.5. These two stations are located in the west of Beijing, which means that the west of Beijing had worth air quality than the eastern Beijing this month.

Pair plot

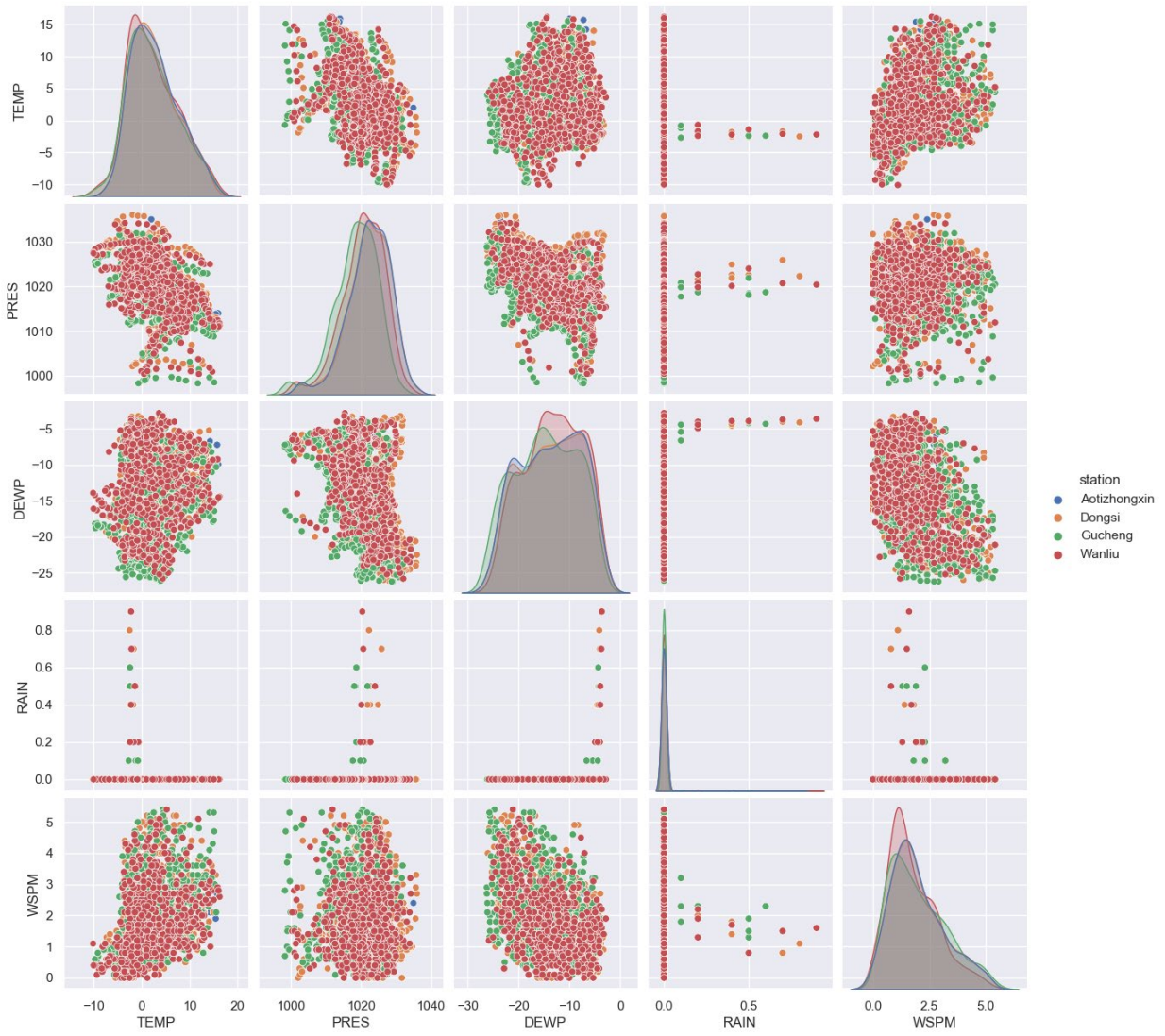

Figure 16 Pair plot for meteorological data

The pair plot shows the relationship between meteorological data, including temperature, pressure, dew point temperature, rainfall capacity, and wind speed. There is no clear relationship between these variables.

Kernel density Plot

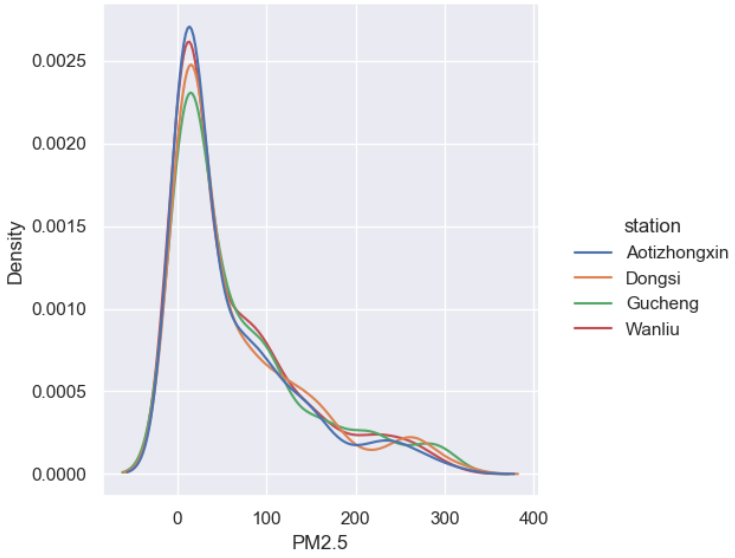

Figure 17 Kernel density plot for PM2.5

The kernel density plot for PM2.5 shows that the distribution is not normally distributed for monitoring stations. PM2.5 concentrations were less than 100 most of the time.

Scatter plot and regression line

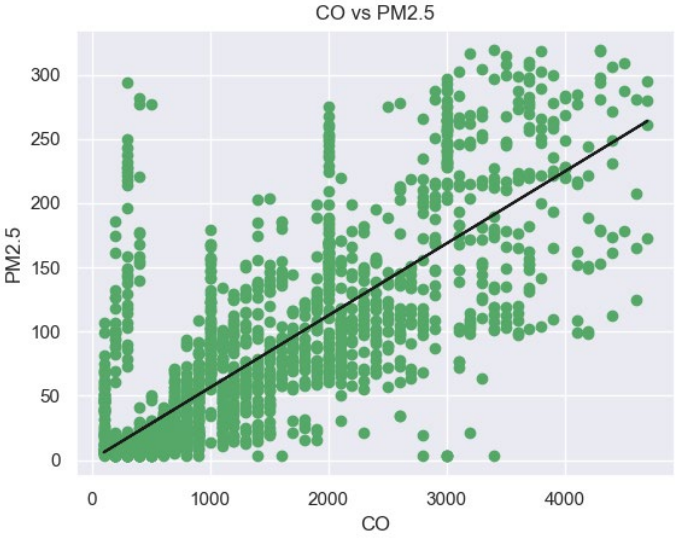

Figure 18 Scatter plot and regression line for CO versus PM2.5

The scatter plot has similar results as the correlation matrix that CO and PM2.5 have a positive relationship.

Boxplot

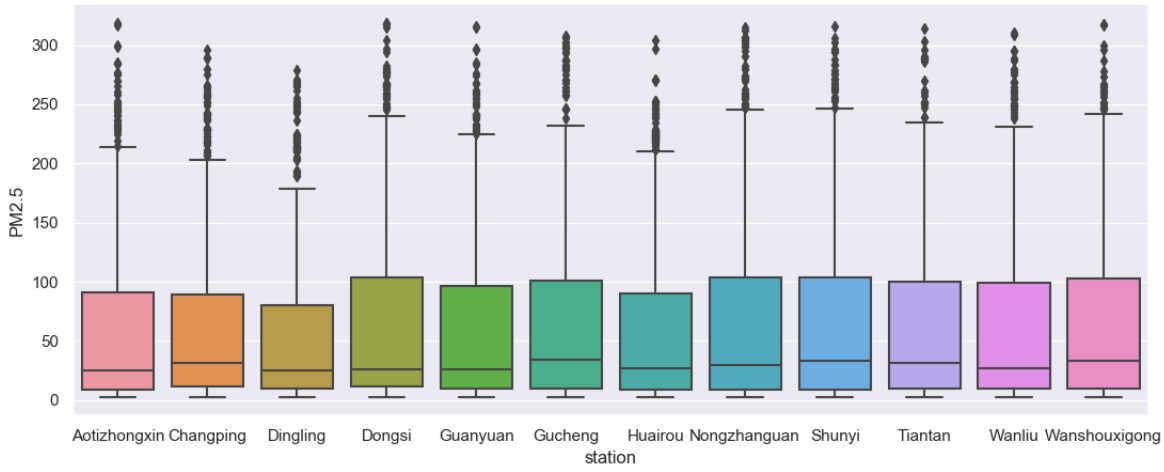

Figure 19 Boxplot for PM2.5

Violin plot

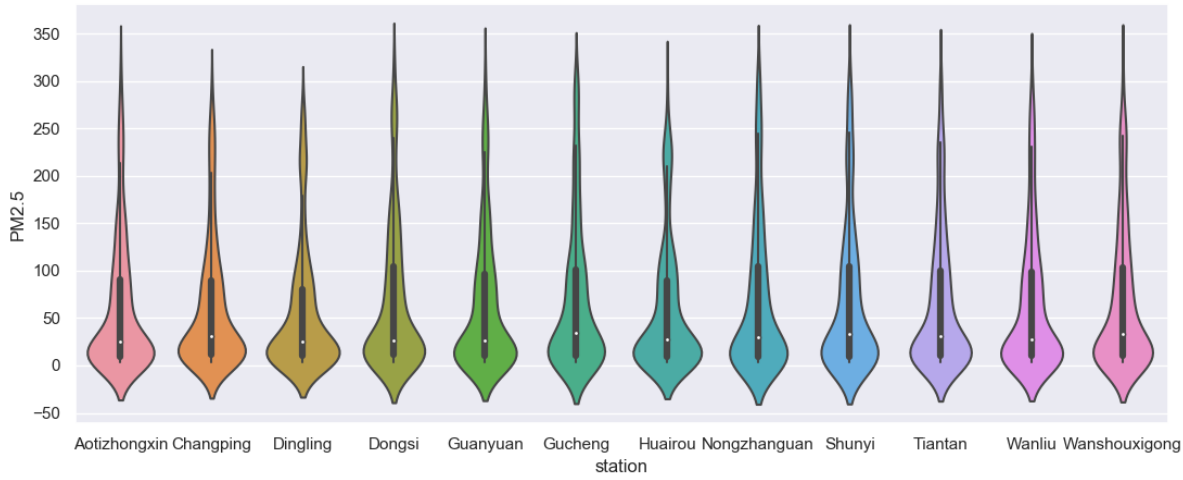

Figure 20 Violin plot for PM2.5

The boxplot and violin plot give the same result for the distribution of PM2.5 in all 12 stations. Dongsi, Tiantan, Wanliu, and Wanshouxigong had much higher median and third quantities than other stations. Aotizhongxin (the Olympic Sports Center) had a lower PM2.5 value. It might be because the density of building and roads near Aotizhongxin is lower than other places, and there is a nearby forest park. Changping and Dinling also have lower PM2.5 than other places since these two stations are located in the suburbs.

DASH

Dash is the original low-code framework for rapidly building data apps in python, R, Julia, and F#. It provides a simple way to make a user interface. Dash by Plotly is an excellent way for a Python developer to create interactive web apps without learning Javascript and Front End Web Development.

The student created a dashboard to visualize the data, allowing the user to make variable selections.

Below two figure shows the shortcut of the dash.

There are three tabs in this dash. The first one is the location-based tab, in which users can choose the station location in the dropdown menu and the date range. The right cards will show the mean value for all the measurements. When turning the switch on in each card, a time series line would show that the user could compare the measurements between multiple variables in the below graph region.

Figure 21 Shortcut of dash-1

The measurement-based tabs allow the user to pick multi-stations. And compare the chosen measurement between different stations. At the same time, users can perform hierarchical selection in this tab, such as time range, time granularity, etc. The final image is updated based on the information selected by the user. Three graphs are shown here, including the data change displayed by the line graph over time and the histogram and boxplot showing the data distribution.

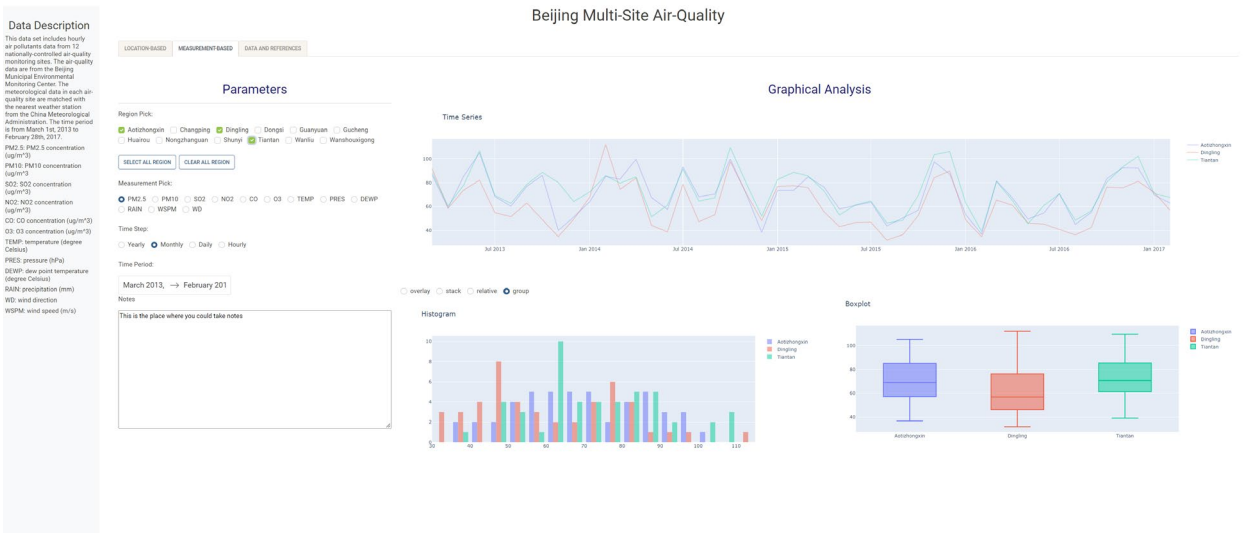

Figure 21 Shortcut of dash-2

The third tab allows users to upload an image. And it shows the reference of data and the paper.

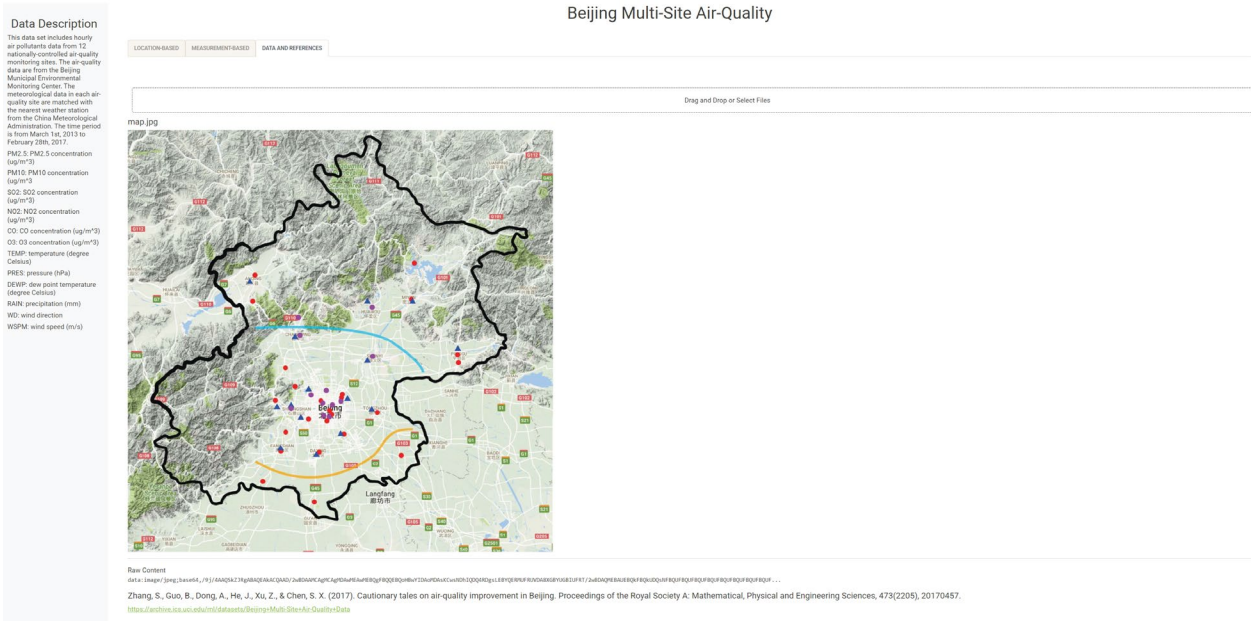

Figure 21 Shortcut of dash-3

Dash provides users with interactive visualization options, which play a more influential role in displaying massive data sets.

More details on how to use dash have been shown in the presentation.

RECOMMENDATIONS

The student applied various graphs from this project, including lineplot, barplot, countplot, catplot, pie chart, displot, pair plot, heatmap, histogram, QQ plot, kernel density estimate, scatter plot and regression line, multivariate box plot and violin plot. Each graph serves a different purpose for processing different types of data.

Line plot is used to visualize time series data; bar plot and count plot are suitable for presenting the number of continuous or discrete data; histogram is generally used to show the distribution of continuous variables, whereas pie chart is only helpful for discrete data. Scatter plots, heatmap, boxplots, and violin plots offer the relationship and comparison between variables. Q-Q plot is used to compare the shapes of distributions and is an image discrimination method of normality test.

Using different data visualization methods and choosing other visualization methods makes it easier to understand the structure and content of the data. In addition to visualizing data locally, the student also created a dash for data visualization.

Dash is the original low-code framework for rapidly building data apps in python, R, Julia, and F#. It provides a simple way to build a user interface. The student created a dashboard to visualize the data, allowing the user to make variable selections.

At the same time, the graph in the dash can be zoomed in, which solves the situation where the graphs are superimposed and cannot be seen clearly due to the overlay of variables.

Since the dash cannot be put into GCP, the current app cannot be opened to the public. The specific operations can be seen in the presentation. The dash provides many options for visualizing data, including time selection, location selection, and variable selection. Users can choose according to their needs. Due to a large amount of raw data, some graphs are generated slowly in the dash, but the overall performance is efficient.

It is recommended that using the time series specific model to fit the data of PM2.5, such as AR model. It is also recommended to use the original data in modeling since the PCA result does not really reduce the dimension of data much. With original data, model is easier to interpret.

Through this project, the student learned to use different image visualization methods and various components in the dash to create interactive visualization results.

Unlike previous work, when creating the dash, the students used the overlay selection method, so the user has more choices in the visualization process. While the dash is currently unavailable to the public, this project will be a great way to showcase the work when the GCP issue is resolved.

APPENDIX

Please see the .py file.

Vis-Project.py: the preprocessing of data and the visualization.

Vis-Project-dash.py: the code for dash

REFERENCES

[1] data: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data

[2] Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., & Chen, S. X. (2017). Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *473*(2205), 20170457.

[3] (2008) Kolmogorov–Smirnov Test. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_214

[4] Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3–4): 591–611. doi:10.1093/biomet/52.3-4.591. JSTOR 2333709. MR 0205384. p. 593

[5] D'Agostino, R. B. (1971), "An omnibus test of normality for moderate and large sample size", Biometrika, 58, 341-348