2021 年第八届中国可视化与可视分析大会

数据可视分析挑战赛

（ChinaVis Data Challenge 2021）

作品说明文档

参赛队名称： muybien

作品名称： Air Pollutants Analysis in China

作品主题关键词： 大气污染源分析、大气污染时空态势分析

团队成员：

尹潇翯，北师港浸大，n830026144@mail.uic.edu.cn

伍恩妍，北师港浸大，n830026118@mail.uic.edu.cn

谭子洛，北师港浸大，n830026105@mail.uic.edu.cn

李鉴龙，北师港浸大，n830026061@mail.uic.edu.cn

团队成员是否与报名表一致（是或否）：是

是否学生队（是或否）： 是

使用的分析工具或开发工具（如果使用了自己研发的软件或工具请具体说明）： chart.js, bootstrap.js, echart.js

共计耗费时间（人天）： 7 人天

本次比赛结束后，我们是否可以在网络上公布该文档与相关视频（是或否）：是

一、作品简介：请围绕作品主题、要解决的问题\场景、目标用户\读者、应用价值等方面简要介绍作品（建议参赛者描述本部分内容不多于 500 字，图表不多于 1 个）

At present, China's air pollution prevention and control has achieved remarkable progress, which benefits the improvement of China's air quality monitoring network. In recent years, air quality monitoring stations have collected a large number high-dimensional and time-series air quality data. Thus, it is of great significance to translate these data to be more user-friendly. Using big data analysis technology and various visualization methods, we mainly analyzed the causes and problems of air pollution, monitored the development trend of air pollution, and analyzed the regional correlation of air pollution.

For the public, policy makers, scholars in various fields and other different types of users to explore the air pollution situation and raise the awareness of protecting the environment. For instance, it can also help the readers to clearly figure out the air quantity in their own cities as well as assist the government to formulate prevention and control strategies towards potential air pollutions.

二、数据介绍：请围绕数据来源、数据格式、数据严谨性、数据清洗等方面简要介绍（建议参赛者描述本部分内容不多于 500 字, 图表不多于 3 个）

We used the open data set of China's high-resolution air pollution reanalysis from 2013 to 2018 provided by the *ChinaVis* competition, and using visual analysis technology and visualization method, the hidden patterns and laws behind the big data of air quality are explored and found. The data is the national air quality reanalysis data based on geospatial grid and corresponding meteorological data, including 13 attributes including six conventional pollutants, wind speed, temperature, air pressure, relative humidity, as well as longitude and latitude. The data are in the *csv* file and each row represents one place’s air pollution in detail while each column represents the pollutant types. The data is relatively strict since there are few null values and irrational data. There are few locations lies on the boundary of China that our conversion failed to transfer these points into geographical city locations. For these points, we manually classified them and do the transformation into according cities. Most of the data are accurately given by several decimal places.

For data preprocessing, we firstly calculated each location’s specific AQI according to the given various pollutant’s value. Then we convert the given locations (based on its latitude and longitude) into corresponding cities or provinces for better classifications. Lastly, by averaging the AQI values, we assigned each province an AQI value to evaluate the region’s general air quality.

Below is the table of the attributes and keys of the data.

Field	Type	Unit	Example
Longitude	Double	degree	109.25
Latitude	Double	degree	18.34

PM2.5	Double	μg/ m^3	26.98
PM10	Double	μg/ m^3	30.66
AQI	Double	/	80.5
SO_2	Double	μg/ m^3	5.44
NO_2	Double	μg/ m^3	3.68
O_3	Double	μg/ m^3	56.70
CO	Double	mg/ m^3	0.36
TEMP (temperature)	Double	K	280.66
RH (Relative Humidity)	Double	%	76.99
PSFC	Double	Pa	100371.2
Province	String	/	浙江省
City	String	/	杭州市

Table 1 Data description

三、分析任务与可视分析总体流程（建议参赛者描述本部分内容不多于 500 字，图表不多于 3 个）

Task 1: Analyze the overall air pollution situations in different provinces.

Task 2: Identify the main air pollution sources and analyze the key pollution causes.

Task 3: Analyzes the temporal and spatial distribution pattern of air pollution and monitors the temporal and spatial evolution trend of air pollution.

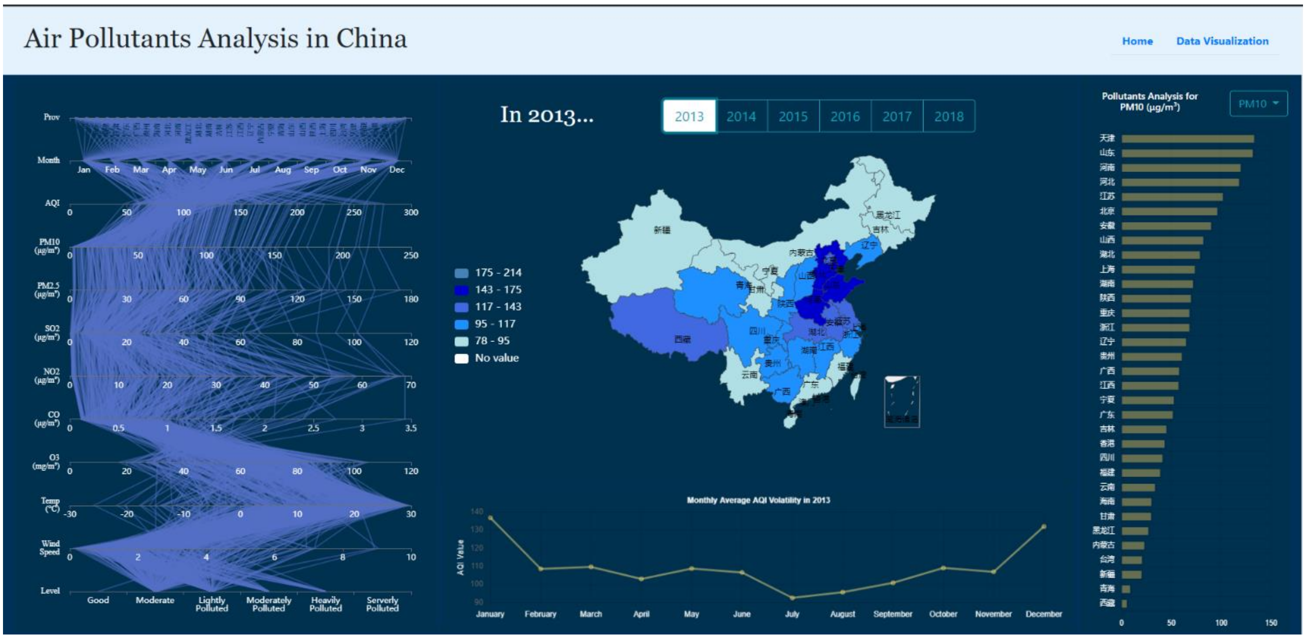


Figure 1: General interface

Our interface and visualization were designed and implemented based on the above data analysis tasks and combined with the air pollution dataset.

In the interface we use bar chart and line chart are used to show the basic data of air pollution situations, and the results are good. From the map of China, we can see a whole picture of air pollution in different provinces. The overall AQI map in years enable us to have a general insight that China's air pollution is still serious and differs in different regions.

In the bar chart, users can select a specific air pollutant and see the rank of all provinces in China. Besides, in below line graph, users can have a general information about the AQI value along with time by selecting a province in the map.

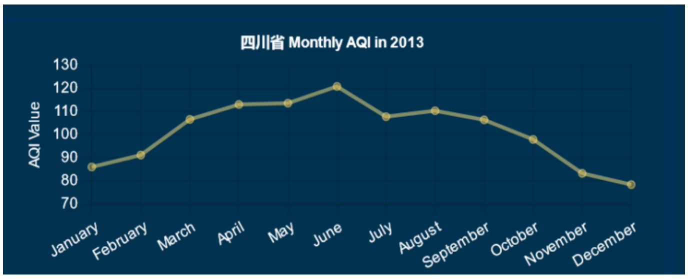


Figure 2: line chart for monthly AQI in a selected province

四、数据处理与算法模型（建议参赛者描述本部分内容不多于 1000 字，图表不多于 5 个）

4.1 Data Aggregation

After downloading all provided data from the website, we assign the province names to all data records according to their latitude and longitude and get rid of the record that lie outside China. Then, we group the data by province, calculate the average value of all attributes except the latitude and longitude attribute. After that, we also do the aggregation of the data in terms of the months. Finally, we can obtain all the average values of the attributes in every province and every month.

4.1 AQI Calculation

Based on the average amount of pollutants, like PM 2.5 and PM 10, we can calculate the AQI of each province and each month. Air Quality Index (AQI) is a measure of air condition, which is a dimensionless index that quantitatively describes the state of air quality. In order to access to the AQI value, we first need to calculate the Individual Air Quality Index (IAQI) of each pollutant by following formula.

$$IAQI(p_{ij}) = \begin{cases} \frac{50-0}{35-0}(p_{ij}-0)+0, & if\ \ 0 \le value < 36 \\ \frac{100-50}{75-35}(p_{ij}-35)+50, & if\ \ 36 \le value < 75 \\ \frac{150-100}{115-75}(p_{ij}-75)+100, & if\ \ 75 \le value < 115 \\ \frac{200-150}{150-115}(p_{ij}-115)+150, & if\ \ 115 \le value < 150 \\ \frac{300-200}{250-150}(p_{ij}-150)+200, & if\ \ 150 \le value < 250 \\ \frac{400-300}{350-250}(p_{ij}-250)+300, & if\ \ 250 \le value < 350 \\ \frac{500-400}{500-350}(p_{ij}-350)+500, & if\ \ 350 \le value < 500 \end{cases}$$

AQI is the maximum value of all IAQI value of each pollutant.

$$IAQI(x_i) = \max[IAQI(p_{ij})]$$

According to the formulas above, we can obtain the air quality by measuring the value of AQI. After data preprocessing, we have totally 12 quantitative attributes, and 2 categorical attributes as showing in **Table 1**.

Table 1

Table attributes after preprocessing

Attribute	Type	Explanation	Example
Province	String	Indicate which province the record is	上海市
Month	String	Indicate which month the record is	April
AQI	Float values	Indicate Air quality	115.791146
PM2.5 (μg/m³)	Float values	Indicate the amount of PM2.5 in the air	42.811917
PM10 (μg/m³)	Float values	Indicate the amount of PM10 in the air	70.154181
SO_2 (μg/m³)	Float values	Indicate the amount of SO_2 in the air	11.923931
NO_2 (μg/m³)	Float values	Indicate the amount of NO_2 in the air	42.133042
CO (mg/m³)	Float values	Indicate the amount of CO in the air	0.700083
O_3 (μg/m³)	Float values	Indicate the amount of O_3 in the air	87.632917
U (m/s)	Float values	Indicate the zonal wind speed	-0.863847

V (m/s)	Float values	Indicate the meridional wind speed	0.757819
TEMP (K)	Float values	Indicate the temperature	288.677347
RH (%)	Float values	Indicate relative humidity	70.001056
PSFC (Pa)	Float values	Indicate the ground pressure	101491.613514

4.3 Data Classification

After obtaining the AQI values of each province, we have found that The AQI of every province in China is roughly in the range of 80-140, which means it is hard to classify the provinces according to the normal measurement. Therefore, here we make use of Geometric Progression to classify 4 classes so that the users are available to get idea of which air pollution level the province is. In Geometric Progression, we first need to obtain the common multiplier of the interval by the following formula.

$$X_{min} \cdot r^n = X_{max},$$

$$where\ X_{min}\ and\ X_{max}\ are\ the\ minimum\ and\ maximum\ value, n\ is\ the\ number\ of\ the\ record$$

Then, we can get 4 class limits which are obtain by the following formula

$$l^{(i)} = X_{min} \cdot r^{n-1}$$

4.4 Clustering

In order to have a better understanding of the pollution distribution in China, we conduct the clustering in terms of both provinces and months using the aggregate AQI values. Since the AQI values is a one-dimensional data, therefore, we choose Jenks Natural Breaks method to do the clustering. In general, the principle of the Jenks Natural Breaks is to put the sample which are close to each other together and consider them as a clustering. Statistically, it can be measured by variances. By calculating the variances of each clustering, and then calculating the sum of these variances, the size of the sum of variances can be used to compare the quality of the clustering. In our project, we decide to generally divide the AQI values into 3 clustering for both provinces and months. Figure 3 shows the 3 clusters which are generated by Jenks Natural Breaks for provinces and months respectively.

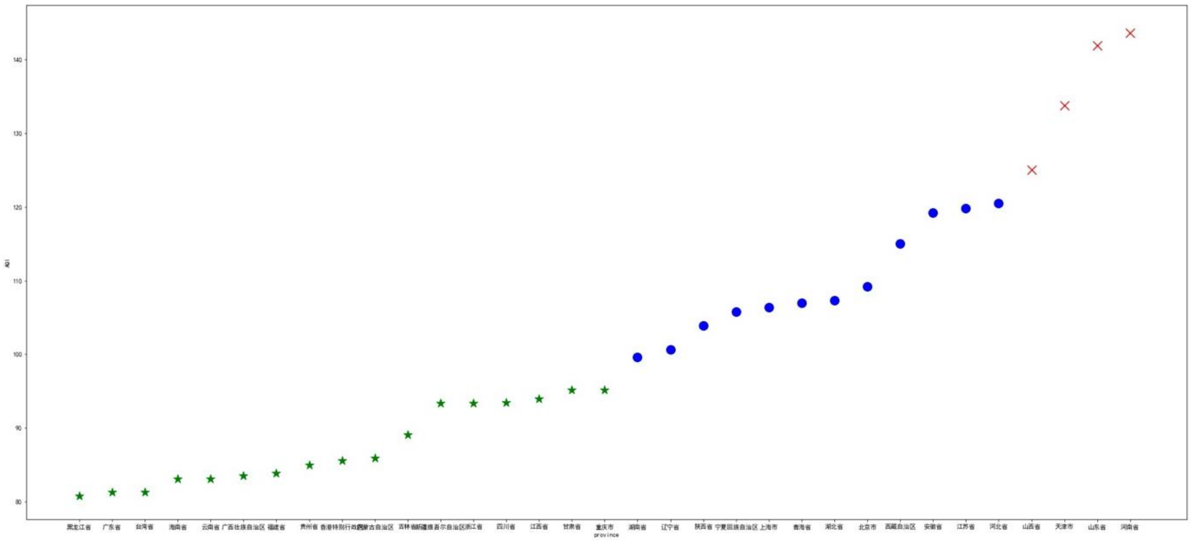


Figure 3: Clusters of provinces

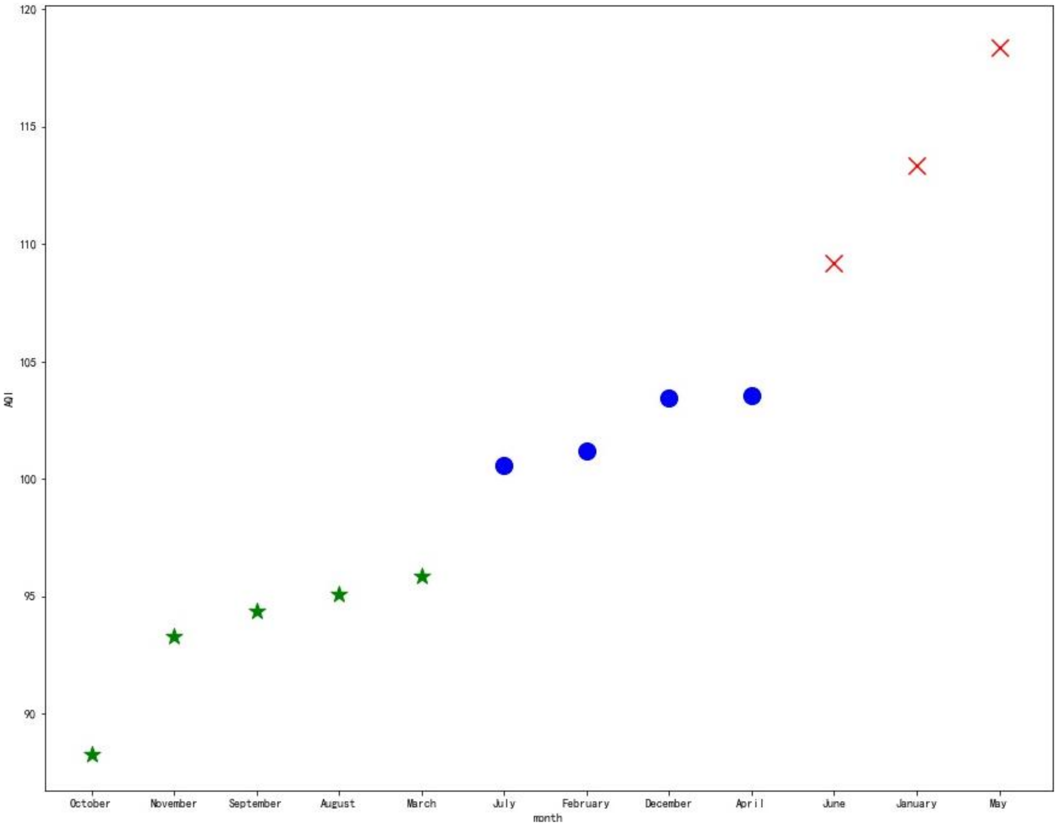


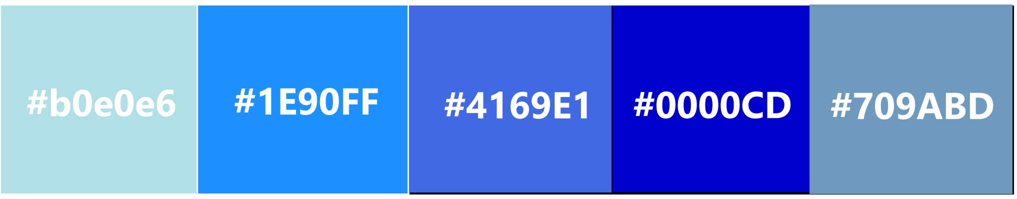
Figure 4: Clusters of months

五、可视化与交互设计（建议参赛者描述本部分内容不多于 1500 字，图表不多于 5 个）

In this project, we utilize Javascript Language to complete the whole visualization. For the whole CSS layout in the visualization, we apply bootstrap.js and bootstrap.css to generate the four modules for the graph. As for China map and parallel coordinate plot, we use echart.js library. Meanwhile, we apply chart.js library to create line chart of time series and bar chart.

More specifically, we mainly have four modules in the whole design, and design various interactive operations within and between views facilitating the user to understand internal details like specific provinces or time.

The first module is map graph. By echarts, we apply an overview Chinese AQI indices map. Here we use the geometric progression to provide intuitive data classification for yearly AQI value in different provinces. To give consideration to both clarity and art, the gradient color design (from low to high AQI value) is as followed, the more polluted, the darker and opaquer the color appears:



The user can click on different legends to see the classification of provinces in different AQI levels. (Figure 5)

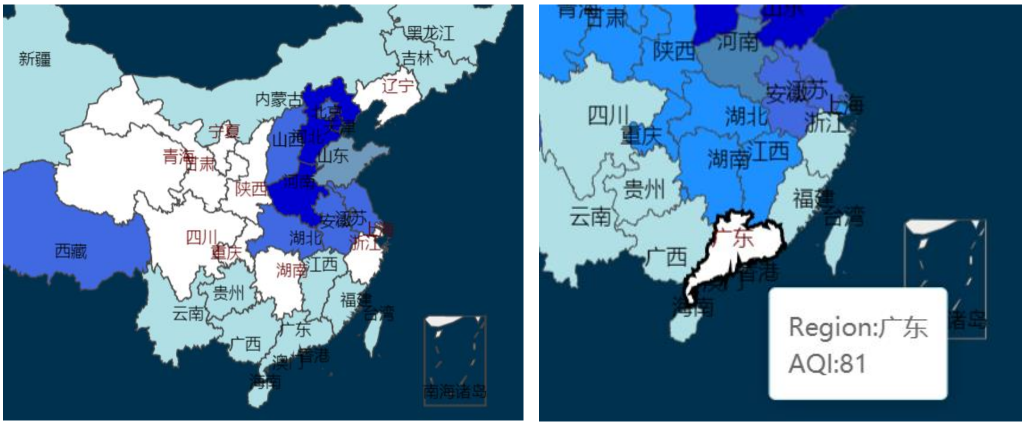


Figure 5: Interaction on map

Then, in the second module, referring to the section in the left side of visualization, we draw a parallel coordinate plot in a vertical way. The characteristic and superiority of parallel coordinates plot is that multiple attributes can be rearranged in parallel, connecting each data record with polylines. Therefore, it is of great help to observe the trend and distribution of data in high-dimension. In this project parallel coordinates plot is suitable for visualize the high-dimensional data of more than 30 provinces. Generally, the user can have an overall size and trend perception of all the average values such as air component, temperature and wind speed and so on, in modules of each province in each year.

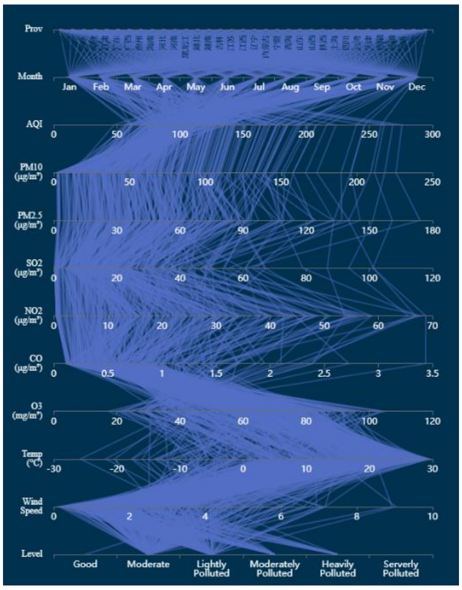


Figure 6: PCP

The interaction design is that, once the user clicks one of the cities on map, the name of province will be highlighted in parallel coordinate plot. Then we can select that area to check a specific year. Also, we plot the last two models by using a line chart to present the national average AQI changes in years, and a bar chart to present province average AQI in decreasing order in the default setting.

Since generalization is not enough to make users have a better understanding of the development of air pollutants, firstly, we set a group of choosing buttons like years choosing for users, which they can select each button to check air pollutants analysis in a specific year. When one of years is selected, the modules are also changed after clicking.

Then we can check the indicator of each province in 12 months. For example, clicking provinces on the graph can show all indicators in a specific province. For the line charts, the user can watch the monthly AQI volatility. Similarly, if the user clicks the region, the line chart will be changed into the monthly AQI volatility in that province. For the last modules, bar chart, once the year is selected, we find out which provinces have worse air pollution on the map graph intuitively. Then to explore the issues of what kinds of pollutants cause those provinces have higher AQI value. In the bar chart, we do a ranking to each specific year. In this way, we can find which pollutants affect the provinces most.

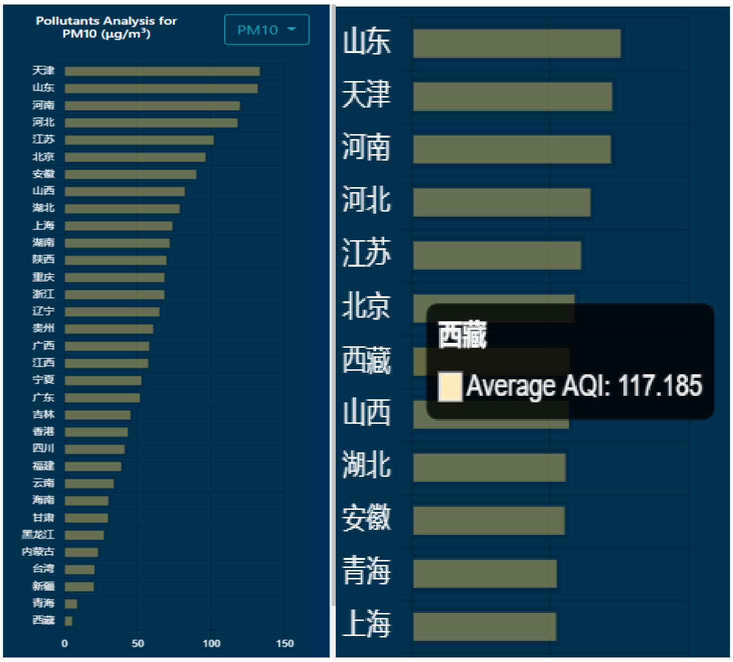


Figure 7 bar chart

六、实验\案例\场景分析（建议参赛者描述本部分内容不多于 2000 字，图表不多于 10 个）

In this part, via Chinese AQI indices map, overview parallel coordinates plot, volatility line chart as well as air component bar chart, we initially identify the **major component** in air corresponding to atmospheric pollution in the whole country, and detail to each province. Thus, indicate the **spatial and temporal distribution** of pollution in China. Additionally, we further explore the correlation between air pollution and **other factors** such as temperature and intensity of pressure. Last but not least, we analyze the **spread of pollutants** among provinces through time series line chart.

6.1 Spatial distribution of pollution

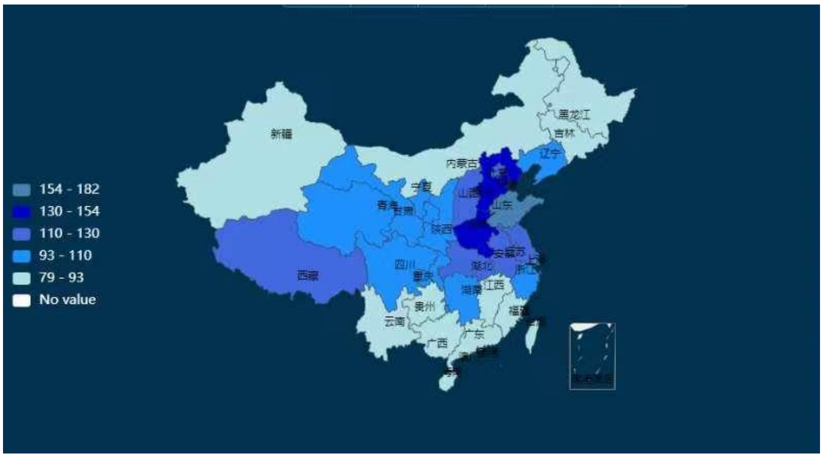


Figure 8: Chinese AQI indices map

The overall AQI numerical map in years enable us to have a general sense of the pollution that China's pollution is still serious and differs in different regions. Most of the heavily polluted cities are concentrated in the central and inland region. China's air pollution belongs to soot type pollution, and the pollution in small and medium-sized cities is higher than that in big cities.

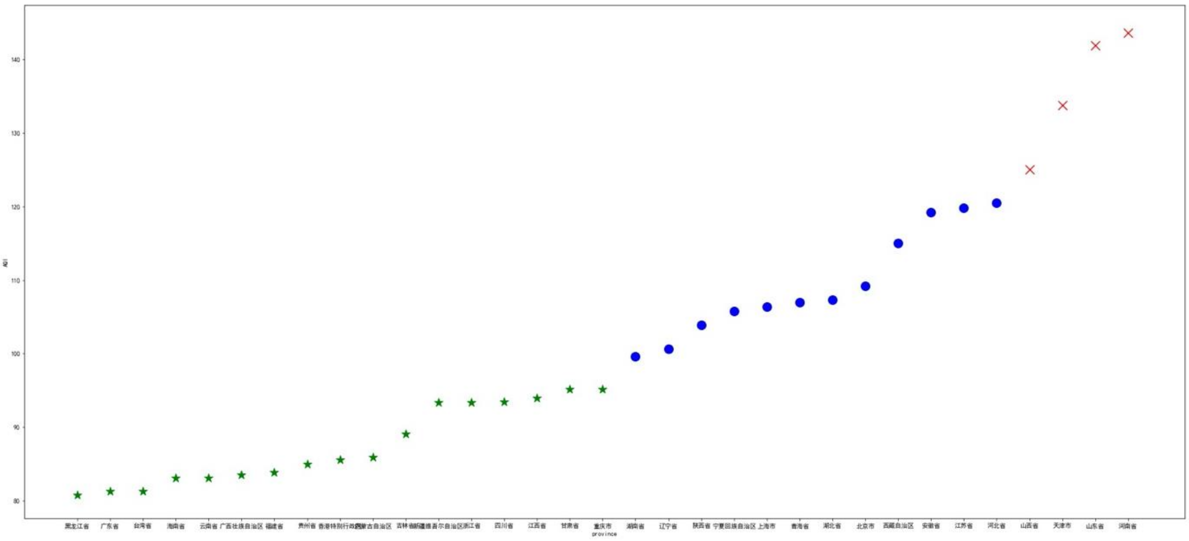


Figure 9: Clustering of provinces

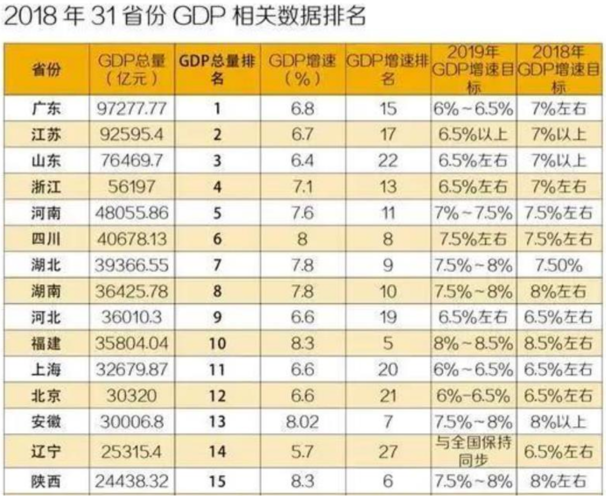
In addition to intuitively obtaining the pollution levels from map, we also use cluster analysis to present each class of provinces in form of points. After clustering, we detect there are mainly three classes according to the AQI values, so the provinces can be roughly divided in three levels of pollution. From the scatter plot (Figure. 9) of the clustering of the provinces, we can see that most of the northern provinces in the third and heavily pollution level.

According to the degree of pollution, we can conclude that, China's trend of pollution from south to north basically presents light – medium – heavy – light. The most polluted provinces are Shandong, Henan and Tianjin, as for the least polluted provinces, Heilongjiang and Fujian are major representatives. Firstly, the slightly polluted provinces are mainly concentrated in south region.

Represented by the Pearl River Delta, most provinces and cities in the south are dominated by textile and light industry, besides, in terms of electricity and fuel, natural gas and new sources are widely used among those provinces. While the AQI indices of inland provinces in central and western China, is generally high over years. We speculate the result is highly related to central and western provinces' focus on manufacturing. In terms of daily source, these provinces and cities mainly use coal for heating and power generation. The coal consumption of China's iron and steel industry is generally heavy in north while light in south, which is also well reflected in numerical AQI of our visualization.

It is worth mentioning that our results finds that both Heilongjiang and Tibet were not consistent with people's common sense that, Heilongjiang is heavily industrial which might be the reason of heavily polluted, while Tibet is sparsely populated so it is much less polluted. However, it is interesting to find out that Tibet with a relative high pollution may cause by Special highland position where is exposed to a number of ozone. Because the amount of ozone in Tibet is the highest among all the provinces in China during the 6 years period. By looking up the research, we acquire that the ozone is a very special component in the air, whose strong oxidation, the acute impact on the human body is still relatively significant. Therefore, it might cause that the large amount of ozone will result in the high-level pollution in the specific area, although other components are not considerably large.

From figure 10, in recent years, due to the vigorous development of economic construction, Henan and Shandong provinces have become the heavily polluted areas, and the pollution level in special economic zones such as Beijing and Shanghai are also relatively serious, which indicates that China has not achieved enough balance among economic construction, environmental protection and pollution control.



2018 年 31 省份 GDP 相关数据排名

省份	GDP总量（亿元）	GDP总量排名	GDP增速（%）	GDP增速排名	2019年GDP增速目标	2018年GDP增速目标
广东	97277.77	1	6.8	15	6%~6.5%	7%左右
江苏	92595.4	2	6.7	17	6.5%以上	7%以上
山东	76469.7	3	6.4	22	6.5%左右	7%以上
浙江	56197	4	7.1	13	6.5%左右	7%左右
河南	48055.86	5	7.6	11	7%~7.5%	7.5%左右
四川	40678.13	6	8	8	7.5%左右	7.5%左右
湖北	39366.55	7	7.8	9	7.5%~8%	7.50%
湖南	36425.78	8	7.8	10	7.5%~8%	8%左右
河北	36010.3	9	6.6	19	6.5%左右	6.5%左右
福建	35804.04	10	8.3	5	8%~8.5%	8.5%左右
上海	32679.87	11	6.6	20	6%~6.5%	6.5%左右
北京	30320	12	6.6	21	6%-6.5%	6.5%左右
安徽	30006.8	13	8.02	7	7.5%~8%	8%以上
辽宁	25315.4	14	5.7	27	与全国保持同步	6.5%左右
陕西	24438.32	15	8.3	6	7.5%~8%	8%左右

Figure 10: Chinese GDP condition

6.2 Temporal distribution of pollution

Similar to the spatial clustering, we divided 12 months into 3 classes to get some idea of which months will be the polluted most.

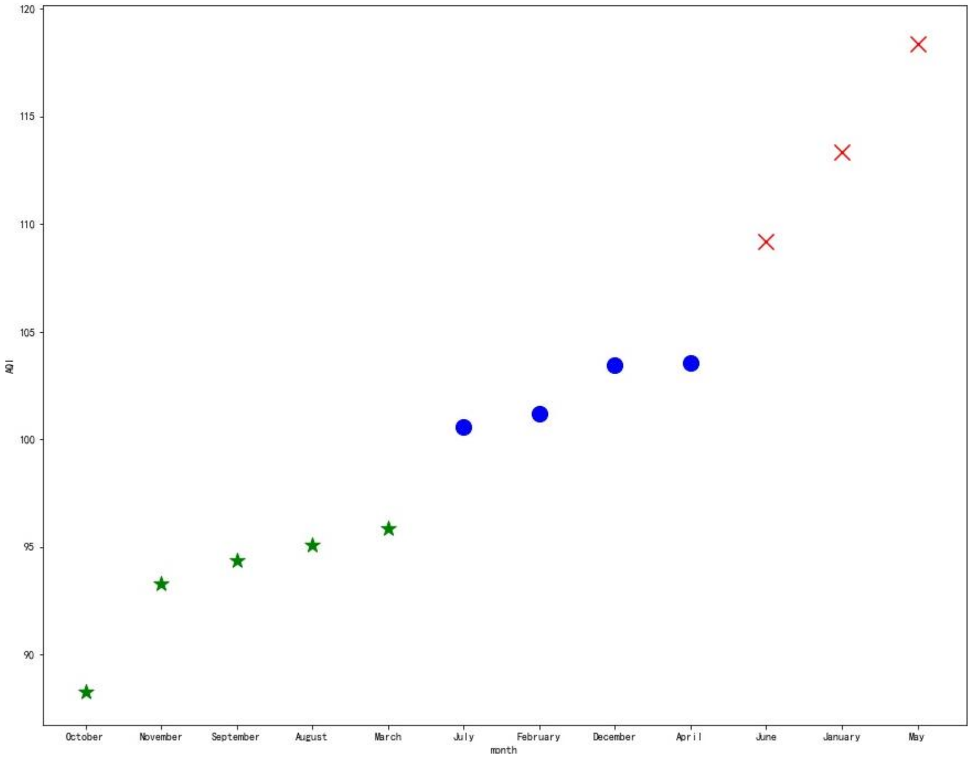


Figure 11: Clustering of months

From Figure 7 we are able to know, generally, colder season, like the October and November, will have lower pollution, while hotter weather, like the June and May, will have higher pollution, from which we can deduce that there is a direct proportional relation between temperature and pollution level.

6.3 Other factors affecting air pollution

In our visualization, we can conclude that firstly in temperature aspect, with increasing in temperature, the stability of the atmosphere is affected and pollution is more serious. While more serious pollution like CO2 will increase AQI value, resulting a vicious circle. Second in intensity of pressure aspect, the decrease of air pressure difference makes it difficult for cold air to move south, thus increasing the value of PM2.5. In addition, humidity will also have an impact on air pollution. With the decrease of precipitation across the country, the washing effect on subsidence is reduced, which also increases AQI value. What’s more, the wind speed also determines the degree air pollutants. The higher the wind speed, the lower the pollutant concentration, and the lower the wind speed, the higher the pollutant concentration, since the pollutants are hard to spread.

6.4 Spread of pollutants among provinces

Take the most heavily polluted province Shandong as example, its monthly AQI value reaches maximum on December. Through comparing its neighboring provinces like Hebei, we find that Hebei reaches maximum on about October, which is a little earlier than Shandong. Therefore, from the time series line chart we can analyze the spread of pollutants from Hebei to Shandong.

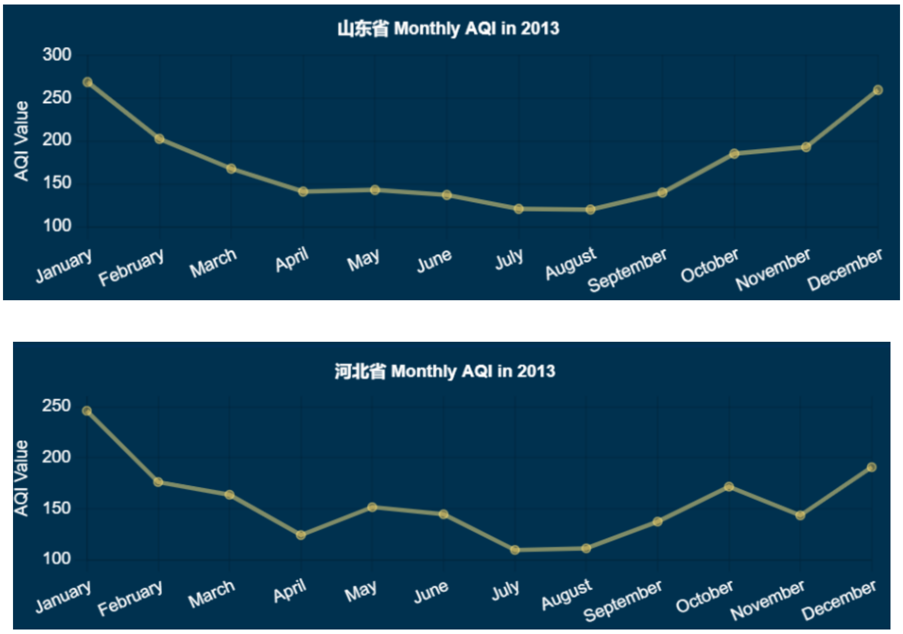


Figure 12: line chart

七、讨论与总结（建议参赛者描述本部分内容不多于 500 字）

From above visualization and analysis, we have a general insight of the China air pollution situation. By analyzing the source and the cause of the pollutant, as well as the temporal and spatial distribution pattern of the pollutant, we can observe the pollution situation in detail. After that, we figure out many regular patterns. For example, how does the industry affect the air pollution geographically, and what is the key pollution sources that contribute to a high AQI value. We find out many facts such as the slightly polluted provinces are mainly located in the south region. For the users, they can view the visualization in detail location for a specific time and corresponding various pollutants to get an overview of how and why the area are being polluted.

Summary

Nowadays, more and more people are concerning our air conditions that are close related to our daily life. Also, the meteorology department has abundant raw data in this field. In our project, we generate a visualization interactive interface, by using these detailed data and try to give users and scholars a brief insight and understanding of the status quo. We fully utilized many types of graphs, tables and charts such as parallel coordinates graph, line graph, bar chart, etc., to present various phenomenon and current problems. Based on these, we do the analysis on the source and the cause of the pollutant, as well as the temporal and spatial distribution pattern of the pollutant to observe the pollution situation in detail.