WANG Jiarun
SONG Yuxuan

Juin 2020

M8 Projet

find the best solution to solve the problem of $PM_{2.5}$ in Beijing

Advisor Stéphane CANU

Summary

1. Introduction

We have learned the principal of information processing in our course M8.
In this project, we will do some statistical study on a database by using the methods in class.

Our project is concerned about the air pollution in Beijing, China, which is caused by fine particulate matters, and $PM_{2.5}$ in particular. $PM_{2.5}$ includes all particulate matter that has an aerodynamic diameter of 2.5 microns or smaller. (2.5 micrometers is approximately 30% of the size of a human hair.) It could lodge deeply into the tissue of the lung and is not easily dispelled. It causes both respiratory and cardiac illness, and has been linked to premature mortality even in healthy individuals.

So we aim to find out the important factors that lead to produce this pollutant and come up with a best idea to decrease the level of $PM_{2.5}$ pollution.

The dataset consists of the meteorological records from 2010-2014, the severity of $PM_{2.5}$ pollution is quantified with a set of statistical measures hourly, which includes the measure of dew point, temperature, pressure, combined wind direction, cumulated wind speed and cumulated hours of rain and snow. Since $PM_{2.5}$ can be formed in industrial processes and the atmosphere when gases such as sulfur dioxide, nitrogen oxides, and volatile organic compounds (all of which are also products of fuel combustion) are transformed in the air by chemical reactions, we should also combined the result of our study with some large factories's location in Beijing (which is also associated with the impact of combined wind direction).

2. Description of data

• Presentation of data

We have 12 variables and only one of them is quality (combined wind direction). The concentration of $PM_{2.5}$ is recorded hourly by the US Embassy in Beijing and the hourly meteorological measurements at Beijing Capital International Airport.(The two places experience the same weather.)

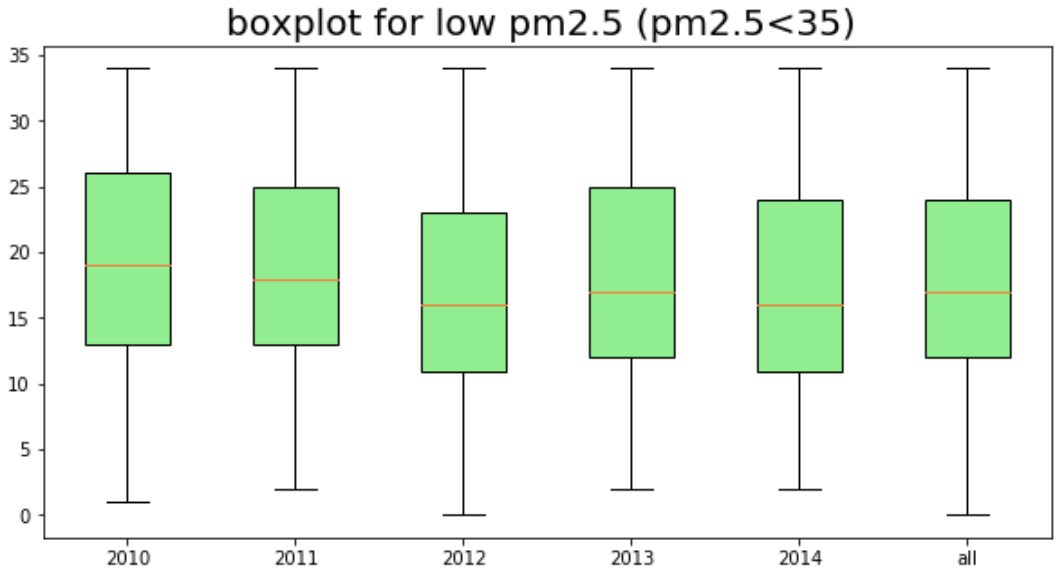
No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	Is	Ir
97	2010	1	5	0	30	-26	-17	1035	NW	201.58	0	0
98	2010	1	5	1	34	-26	-18	1035	NW	205.6	0	0
99	2010	1	5	2	27	-26	-19	1035	NW	208.73	0	0
100	2010	1	5	3	25	-27	-18	1035	NW	213.65	0	0
101	2010	1	5	4	28	-27	-19	1035	NW	218.57	0	0
102	2010	1	5	5	28	-27	-16	1034	NE	4.92	0	0
103	2010	1	5	6	27	-26	-16	1035	NE	8.05	0	0
104	2010	1	5	7	27	-27	-16	1034	NE	13.86	0	0
105	2010	1	5	8	27	-26	-16	1035	NE	18.78	0	0
106	2010	1	5	9	29	-26	-15	1035	NE	24.59	0	0
107	2010	1	5	10	36	-25	-14	1035	NE	29.51	0	0
108	2010	1	5	11	30	-25	-13	1035	NE	34.43	0	0
109	2010	1	5	12	27	-25	-12	1034	NE	39.35	0	0
110	2010	1	5	13	39	-24	-11	1032	NE	41.14	0	0
111	2010	1	5	14	41	-22	-11	1032	cv	0.89	0	0
112	2010	1	5	15	33	-23	-11	1031	NW	1.79	0	0
113	2010	1	5	16	50	-24	-11	1031	NW	3.58	0	0
114	2010	1	5	17	56	-23	-11	1031	NW	5.37	0	0
115	2010	1	5	18	59	-23	-11	1032	NW	7.16	0	0
116	2010	1	5	19	60	-22	-13	1033	NW	10.29	0	0
117	2010	1	5	20	84	-22	-12	1033	NW	13.42	0	0
118	2010	1	5	21	106	-24	-18	1033	NW	16.55	0	0
119	2010	1	5	22	66	-22	-13	1034	NW	20.57	0	0
120	2010	1	5	23	50	-22	-16	1033	NW	23.7	0	0
121	2010	1	6	0	56	-25	-17	1033	NW	26.83	0	0
122	2010	1	6	1	77	-25	-14	1033	NE	4.02	0	0
123	2010	1	6	2	50	-26	-14	1034	NE	8.04	0	0

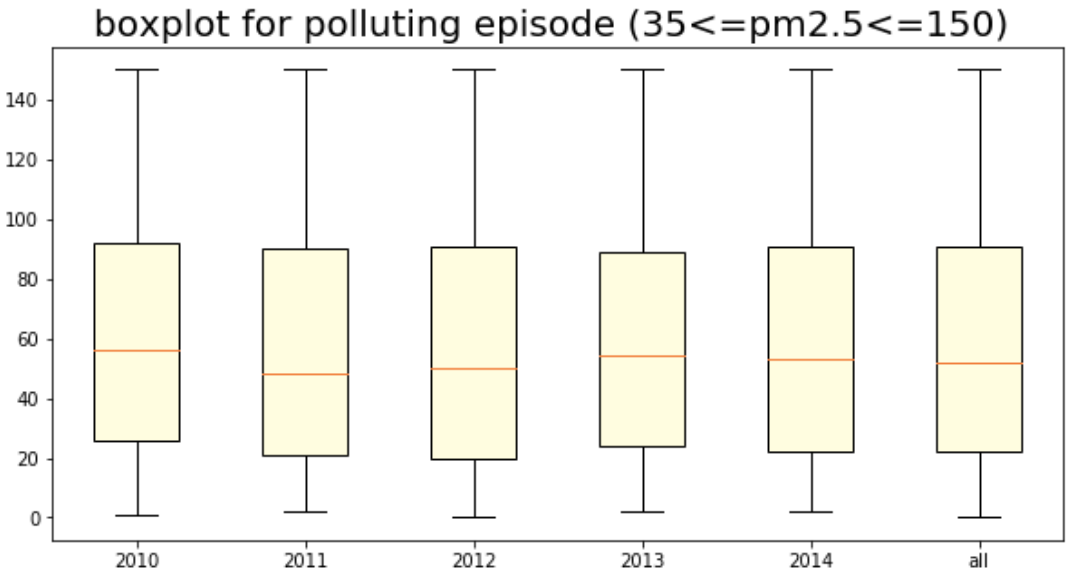
- **Box plot**

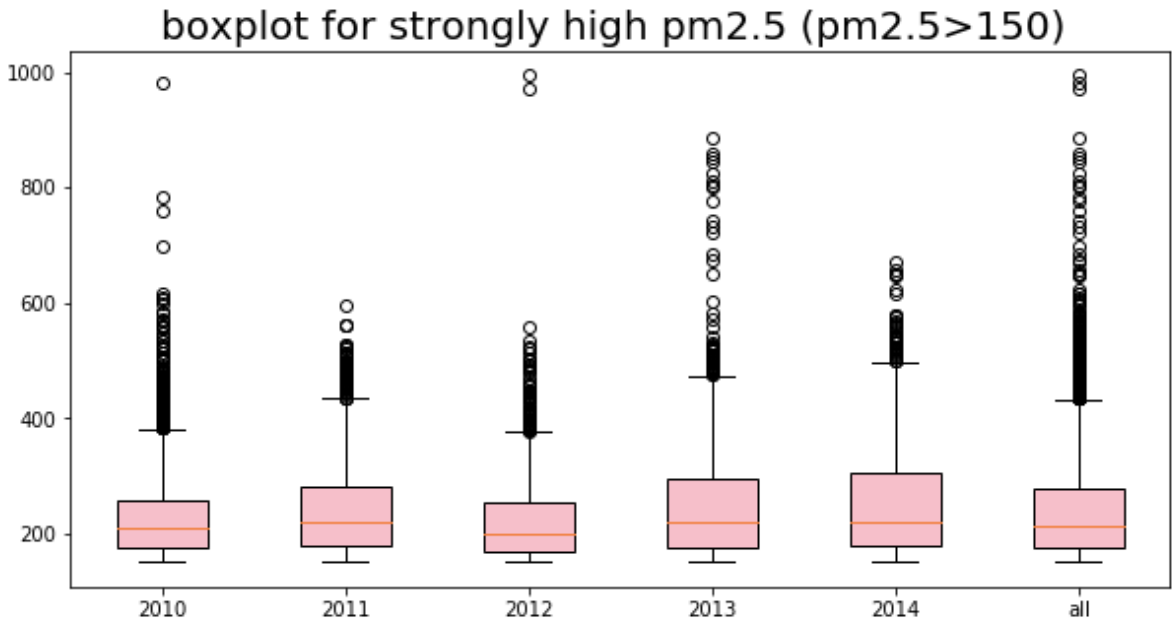
We firstly divide the data into three categories to easily understand the severity of the $PM_{2.5}$ pollution: (unit $\mu g/m^3$)

$PM_{2.5}$ state	Range
Low	< 35
polluting episode	35≤ pm2.5 ≤ 150
Strongly harmful	>150

By using the box plot, we have:

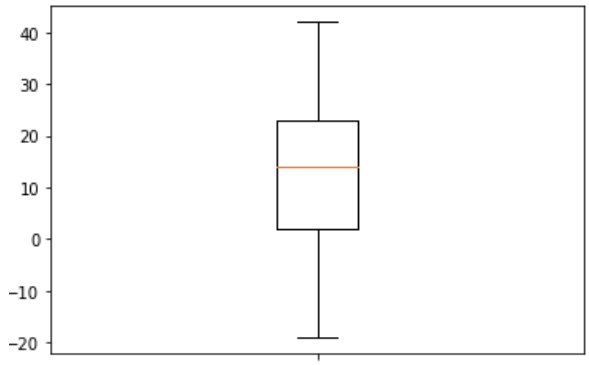




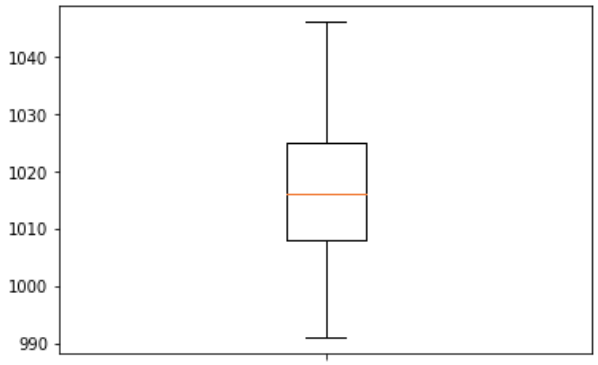


China's State Council set up a pollution reduction target in 2012 , but from the result we could find the fact that the concentration of the $PM_{2.5}$ in the years 2013 and 2014 increases, rather than decreases, compared with those in year 2012, unfortunately. In the picture 'box plot for strongly high pm2.5', there are some data out of the range, but we could not say abnormality since the concentration has a large variety, we divide it into three categories to simplify the data and we could also have the fourth categories 'extremely harmful episode' for those higher than 150 .

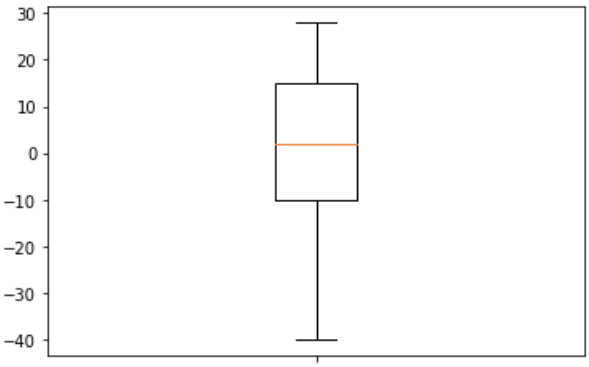
Then, we make the box plot for other variables.



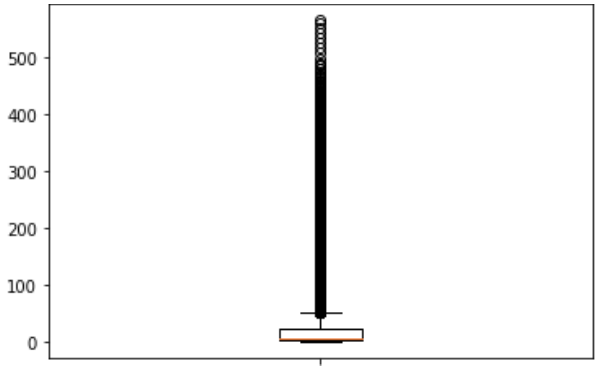
Temperature



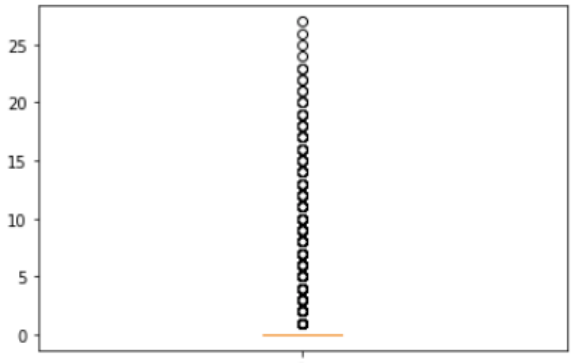
Pressure



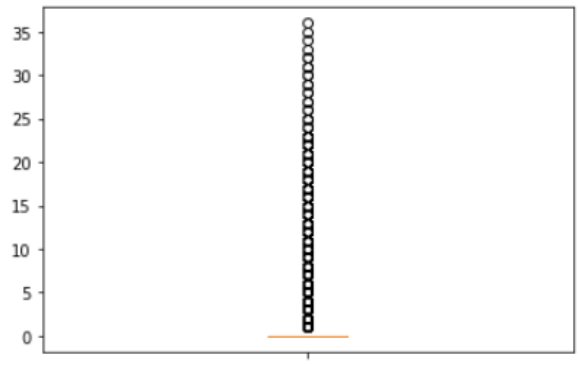
Dew point



Cumulated wind speed



Cumulated hours of snow



Cumulated hours of rain

We can then observe the fact that the distribution of measurements is homogeneous for the first three figures but quite heterogeneous for the last three. It is not surprised to have this result since the strong wind is believed to decrease the air pollution and the lack of wind is often blamed for the high concentration of $PM_{2.5}$ in Beijing. As for the cumulated hours of snow and rain, which depends much on the seasons (winter and summer) is reasonable to be accumulated in one season but very closed to 0 for the whole five years.

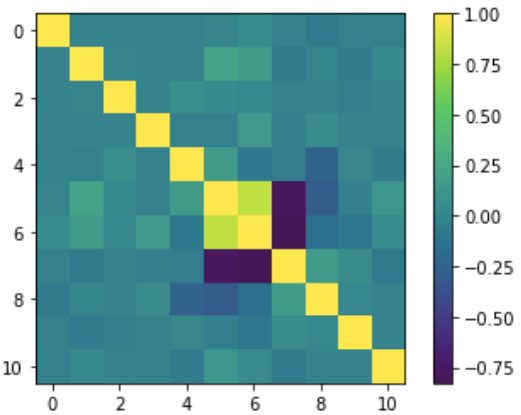
• PCA (Principal Component Analysis)

We use the method learned in class:

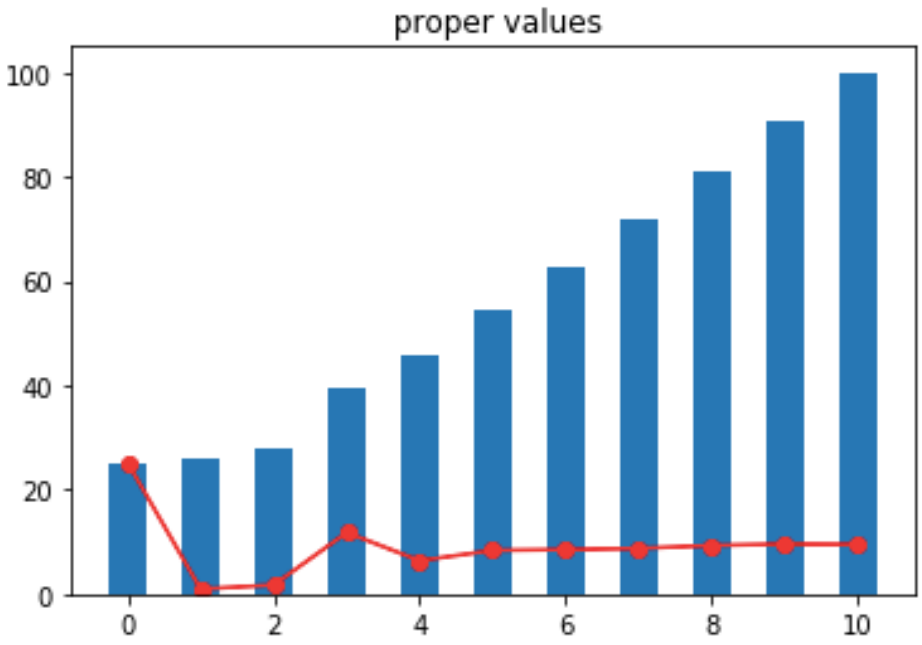
$$\text{Fonction } V_n, U, \lambda \leftarrow \texttt{ACP}(X, k)$$

$$X = (X - un * \texttt{mean}(X))./(un * \texttt{std}(X))$$
$$(V, \lambda) = \texttt{eig(X'*X,k)} \qquad ou \quad (U, V, \mu) = \texttt{svd(X,k)}$$
$$U = X * V$$
$$V_n = V * \sqrt{\lambda}/\sqrt{n} \qquad ou \quad V_n = V * \mu/\sqrt{n}$$

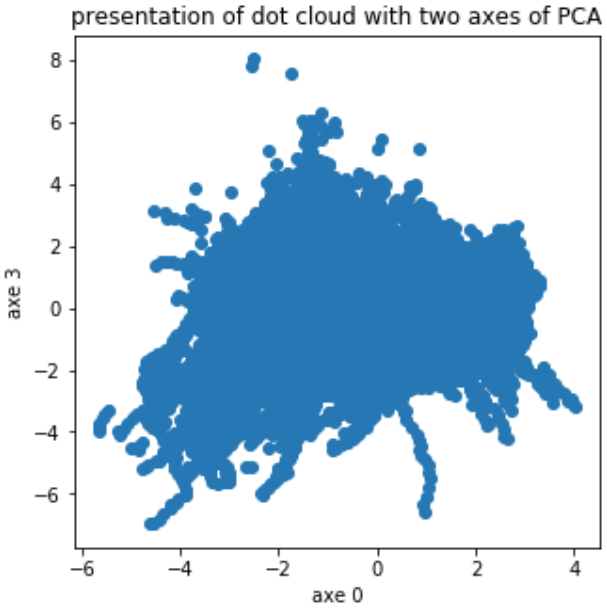
Then we could get the correlation matrix shown below. Obviously, the values in diagonals are 1 because they are the variances of the variables themselves. Also, the figure reveals the fifth variable (dew point) has a strong linear correlation with the sixth variable (temperature), whereas the seventh variable (pressure) has no relation with these two variables.



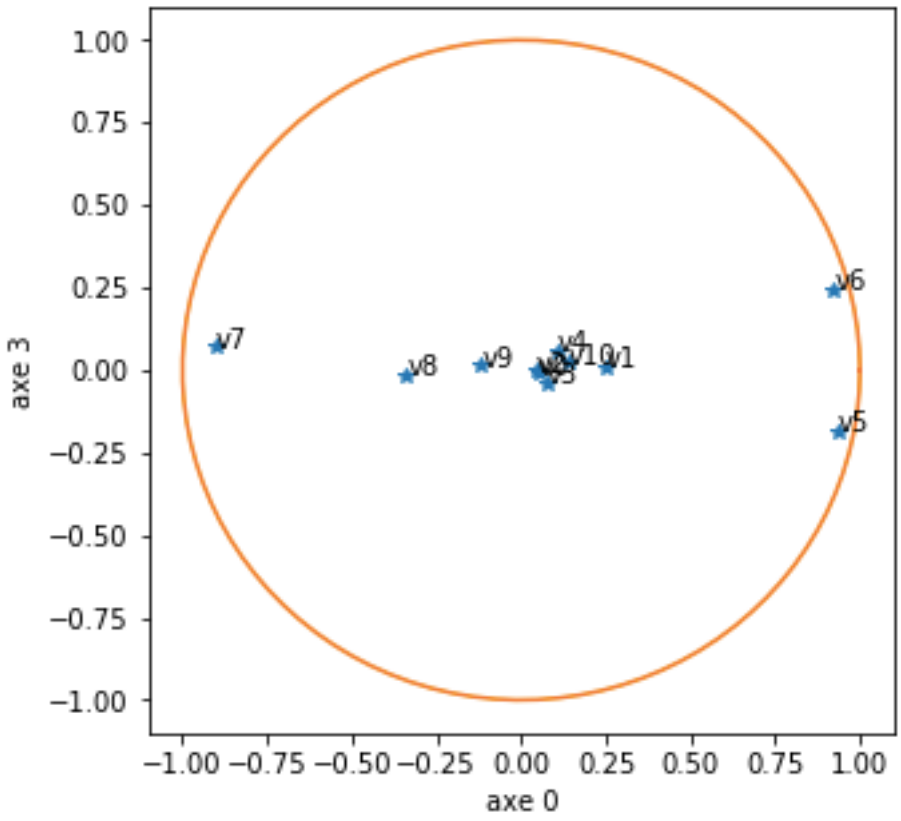
After the normalization of the data, we can visualize the proper values and their importance as a percentage of information.



Avoiding to the figure, the first and the third proper values can better describe our data and avoid the loss of information as much as possible. So after the calculation of the main components, we choose axe0 and axe3 to be the two principal axes and visualize the dot cloud as following :



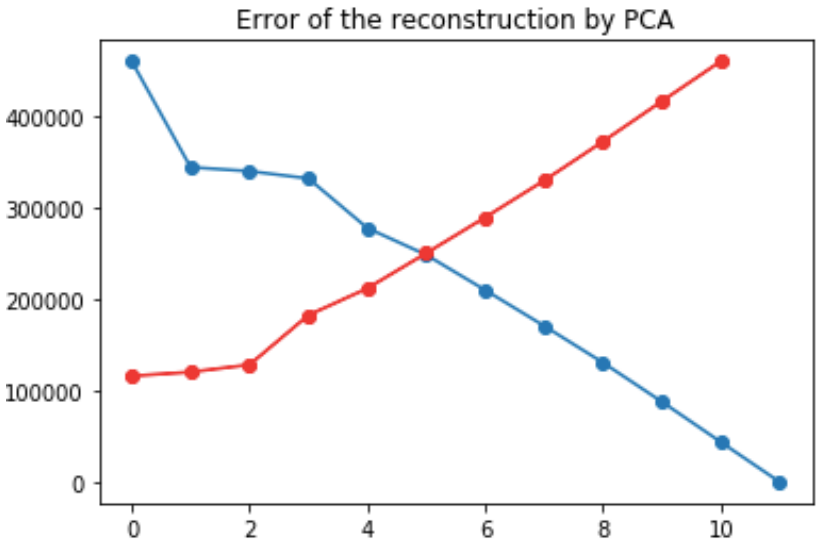
Owing to the large amount of our data, the individuals are not clearly shown but we could say that the mean part is in the middle of the figure and we could then find the relation between the variables by considering the circle of correlation.



We can firstly converge to the same conclusion as the correlation matrix above, which reveals the '' none relation'' between the seventh variable and the fifth, sixth variable.

Then the variables in the middle of the circle have strong relationship with one another, including year, month, day, hour, pm2.5 and cumulated wind speed, cumulated hours of snow and cumulated hours of snow as well but not so strong as them. We could also think these variables are better represented by the two axes.

Afterwards, we verify the error of the reconstruction by PCA is equal to the sum of rest of the proper values.



So, according to the calculation all above, we can interpret individuals' projections of the first two main components by separating the individuals to 4 parts:

- Year, month, day, hour and cumulated wind speed
- cumulated hours of snow and cumulated hours of snow
- Dew point and temperature
- Pressure

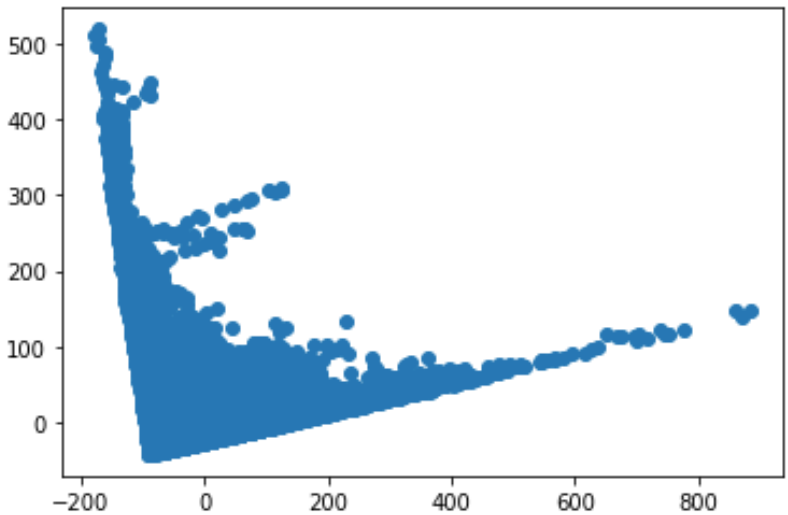
At last, we want to use another method to prove the contribution of the PCA and the percentage of the confidence we could have to represent and simplify the data by PCA.

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
pca.fit(X)
print(pca.explained_variance_ratio_)
```

```
[0.7529861 0.1976071]
```

From the figure above, we have the result that shows if we choose two components for PCA, we have approximately 75% of the confidence to describe the data well by using PCA. If we transforme the whole data in the first two components, it is shown as following :

```
X_new = pca.transform(X)
plt.scatter(X_new[:, 0], X_new[:, 1],marker='o')
plt.show()
```



3. Regression

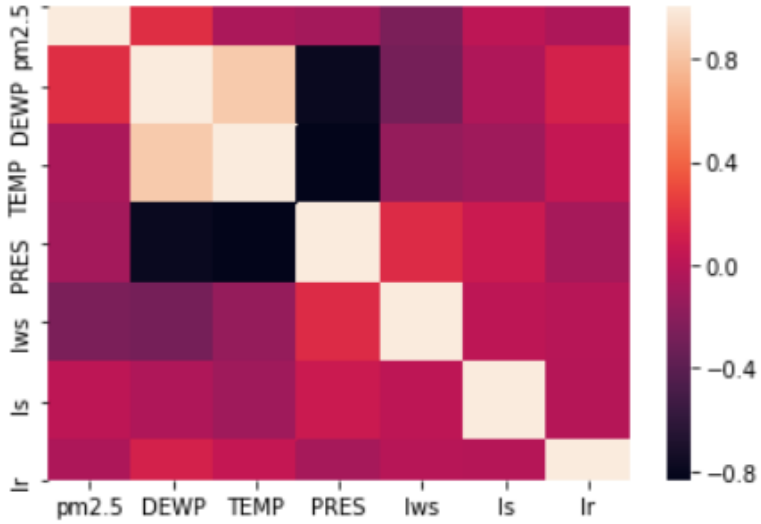
As we have seen that the situation of those 5 years are almost the same in the first part of box plot, we decide to do linear regression analysis only on the data.of year 2011. Besides, as we have observed before in the PCA, dew point and temperature are well related, so in the following analysis of regression, we just choose temperature rather than considering the 2 variables.

Moreover, unfortunately, we have so much data that makes the situation more difficult to do linear regression very well. There exists some kinds of complicated functional relationship between these variables, but we can tell certain relationship applying what we have learnt this semester.

Firstly, we calculate the correlation as following :

We can tell from the result above that there is a positive correlation between temperature and pm2.5. While for Ir, Is, PRES and Iws ,they have a negative correlation with pm2.5.

	pm2.5
pm2.5	1.000000
TEMP	0.018467
Is	-0.003448
Ir	-0.060243
PRES	-0.130904
Iws	-0.278571



Next, we calculate the linear regression between pm2.5 and those variables:

In order to simplify the code, we define a function to do the linear regression in the form of $y_p = a^*x+b$. We calculate R2 coefficient of determination with the following formula:

$$R^2 = \frac{SC_{Expliqué}}{SC_{Total}} = \frac{\sum_{i=1}^{n} (zi - \bar{y})^2}{\sum_{i=1}^{n} (yi - \bar{y})^2}$$

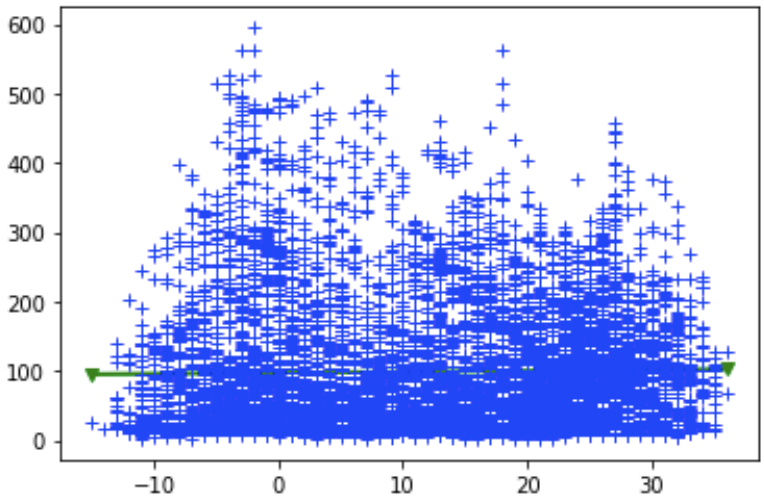
```
y= data3['pm2.5'] #all data of pm2.5

def regression (variable):
    x= data3[variable]
    sx2 = np.sum((x-np.mean(x))**2) #calculate the variance
    sy2 = np.sum((y-np.mean(y))**2)
    sxy=((x-np.mean(x)).T @ (y-np.mean(y)))
    a = sxy/sx2                          #calculate a
    b = np.mean(y) - a*np.mean(x)  #calculate b
    n = len(x)
    print('a=',a,'b=',b)
    xm = np.min(x);
    xM = np.max(x);
    ym = a*xm+b;
    yM = a*xM+b;
    plt.plot([xm,xM],[ym,yM],'gv-', linewidth=2) #print the figure
    plt.plot(x,y,'b+')
    plt.show()
    yp = a*x+b
    e = y - yp
    SCT = np.sum((y-np.mean(y))**2)  # calculate total square sum
    SCR = np.sum(e**2)  # square sum residue
    SCM = np.sum(((y-e)-np.mean(y))**2)    # sum of the squares of the part explained
    SCT, SCM, SCR
    R2 = 1- SCR/SCT
    print('R2=',R2)
```

Thanks to the function above, we get the information below:

Regression for temperature:

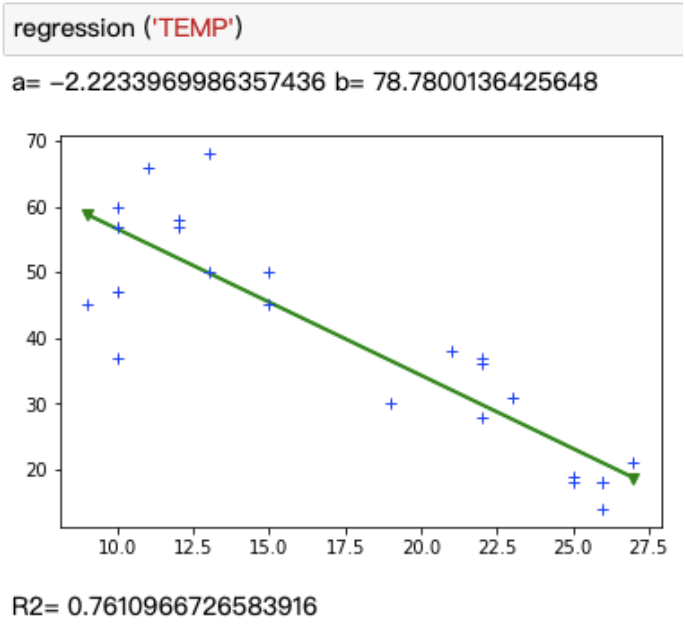
```
: regression ('TEMP')
```

a= 0.1419318273868936 b= 97.31026818757746



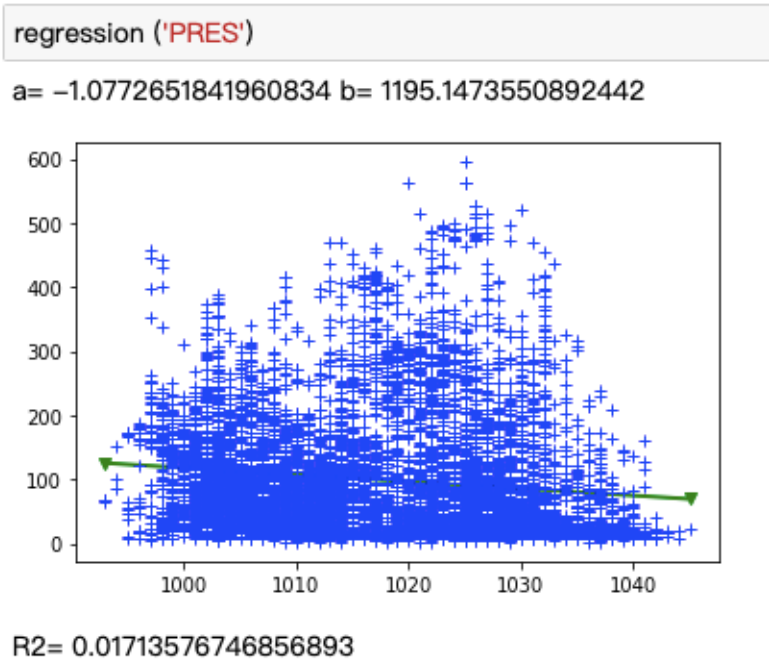
R2= 0.0003410130856985605

We sadly see that the R2 is so small that we can not conclude any linear relation. However, we can find that the higher the temperature is, the less concentration of $PM_{2.5}$ is. However, if we just look at the data for one day but not one year by removing the influence of other variables, we could have the figure as following:



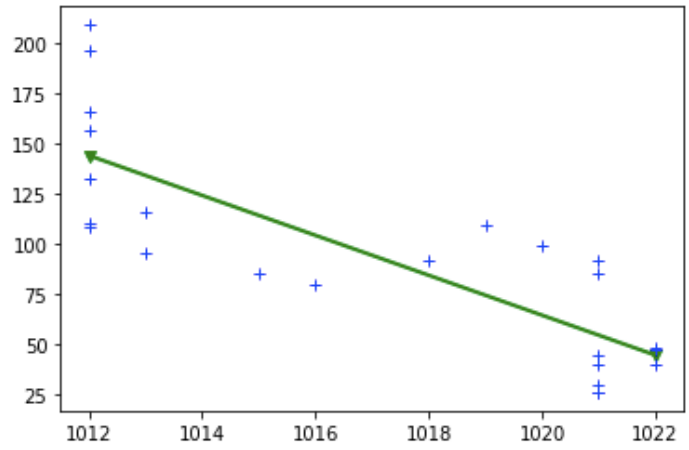
Although the value of R^2 is still not so ideal but compared with the last figure, it is much better. We could have the equation as : $y = -2.223\ x + 78.78$.
we will verify this equation in next part : student's t-test.

Next, we do the regression for variable pressure:



The value of R2 is also too small to conclude any linear relation. However, we can find something interesting in the data for one day:

```
regression ('PRES')
a= -9.945774007876402 b= 10209.053721094619
```



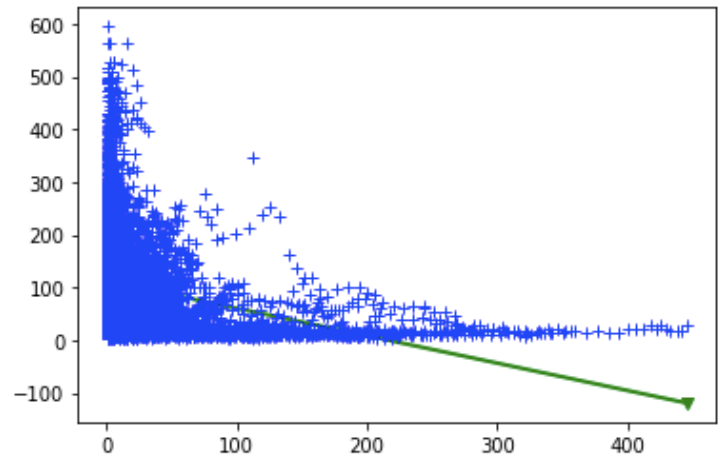

```
R2= 0.6622895211989539
```

We get the equation as: $y = -9.946\ x + 10209.054$

R2 is 0.66, which is not nearly 1, but in contrast to the result before, we can admit the linear regression and we find that the higher the pressure is, the less concentration of $PM_{2.5}$ there is. We will testify the equation in student's t-test.

Then the regression for variable lws (Cumulated wind speed)

```
regression ('lws')
a= -0.519825890199937 b= 112.32269567198342
```

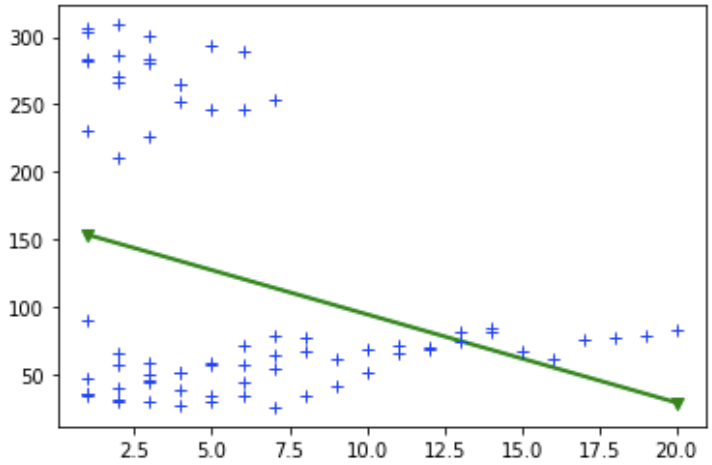



```
R2= 0.0776016212594578
```

The regression for the variable 'ls' (cumulated hours of rain):

```
regression ('ls')
```

```
a= -6.532631107472991 b= 160.0443498876387
```

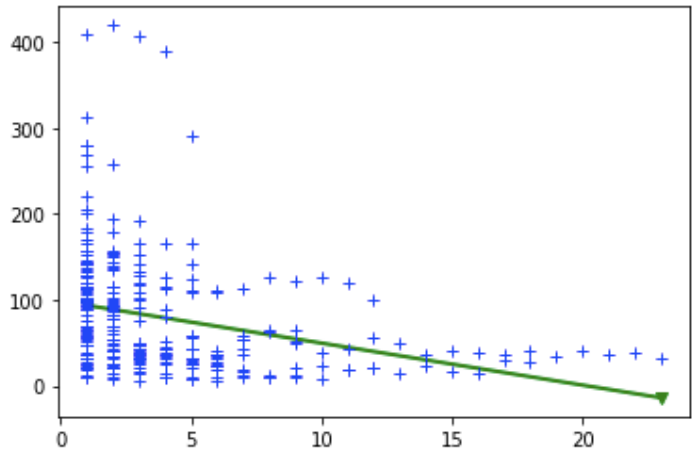



```
R2= 0.10070868565076929
```

The regression for the variable 'ls' (cumulated hours of snow) :

```
regression('lr')
```

```
a= -4.890129719544894 b= 98.5435683745199
```




```
R2= 0.08778586265421162
```

We see that they all have weak linear relation with $PM_{2.5}$ which means the linear regression does not fit our situation, the relation between them is much more complicated. But we have the same conclusion as the correlation table at the beginning of this part. That is to say we at least are able to conclude that the increment of the rain, snow and wind can help decrease the concentration of $PM_{2.5}$ and solve the air pollution.

4. Test

• Chi 2-Test

Until now, we still have a variable not tested yet: combined wind direction. By using the test du chi 2, we try to find whether different wind directions have influence on the concentration of $PM_{2.5}$ or not. At first, we have five different wind directions.

SE	cv	SW	NE	NW
92	76	248	16	53
106	225	37	78	17
96	70	114	298	21
17	82	66	29	22
149	154	211	38	120
41	23	185	13	13
38	144		77	24
98	250		43	25
41	158		55	11
249	288		64	11

.....

166	63		54	8
416	63		122	115
70	48		298	451
21	15		17	103
65	43		38	8
70	198		79	87
5	43		242	20
54	75		113	37

According to the table above, we give their respective probabilities as

$P_{cv} = 0.2$, $P_{SE} = 0.4$, $P_{NE} = 0.1$, $P_{NW} = 0.3$, $P_{SW} = 0.0$

We then make the hypothesis : H0: it is not raisonnable

H1: it is raisonnable

Then we get the distance of chi 2 between the observation and the theoretical values: 1.64

Finally, we have the p_value: 0.65

That means we could look at H0, and think the probabilities are raisonnable which is to say we can ignore the variable of the wind direction southwest. Also, after the calculation of the mean concentration of $PM_{2.5}$ in different wind directions, the result of that in direction SW (southwest) is shown as ''nan'', which stands for the negligence of this direction. So we just study four directions in the following test. The repartition table now is shown as :

Combined wind direction	Mean concentration of PM2.5
cv (calm and variable)	164.26781326781327
SE(southeast)	117.62799401197604
NE(northeast)	95.09134615384616
NW(northwest)	59.35689655172414

- **Student's *t*-test**

From the repartition table, we could see that the concentration of $PM_{2.5}$ is quite high in direction of southeast while the lowest is in the opposite direction northwest. So we use this method to examine whether the wind direction of southeast increases the concentration of $PM_{2.5}$ compared with its opposite direction northwest. We already have the mean value, so we calculate the standard deviation next:

$$\sigma(r) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - r)^2}$$

We could get the observation below :

	SE(southeast)	NW(northwest)
Mean value	117.62799401197604	59.35689655172414
Amount	148	84
standard deviation	1115.1782399046158	674.2991070098534

We first make the hypothesis:

H0: wind direction southeast does not increase the $PM_{2.5}$ concentration.

H1: wind direction southeast increases the $PM_{2.5}$ concentration

$$\begin{cases} \mathcal{H}_0 : \mu_t - \mu_p = 0 \\ \mathcal{H}_1 : \mu_t - \mu_p > 0 \end{cases}$$

We then use:

$$t = \frac{\overline{x}_t - \overline{x}_p}{\sqrt{\widehat{\sigma}^2\left(\frac{1}{n_t} + \frac{1}{n_p}\right)}} = 4.4849998$$

Since H1 is increment, p-value is changed as:

$$pval = \mathbb{P}(T \geqslant t) = 1.151426^{-5} < 0.05$$

Assuming a first-species risk of 0.05. The p-value is too small so we do not think about the first hypothesis and conclude that the wind direction southeast increases the concentration of $PM_{2.5}$.

By using the same method, we can conclude that if the wind direction is calm and variable (cv), the concentration of $PM_{2.5}$ is even higher while the wind direction northwest can decrease the pollution and the wind direction northeast has little influence on it.

- **Student's *t*-test on the slope of regression**

From the regression part, we finally get two equations for variables 'temperature','pm2.5' and variables 'pressure','pm2.5'. Now we want to testify if these two results are just accidents. We start with the first equation:

$$y = -2.223\,x + 78.78$$

Then the hypothesis :

$$\begin{cases} H_0 : Indépendance & \boldsymbol{a = 0} \\ H_1 : Dépendance & \boldsymbol{a \neq 0} \end{cases}$$

We have the estimated variance with the value : $\begin{cases} \hat{a} = -2.223 \\ \hat{b} = 78.78 \end{cases}$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{a}x_i + \hat{b}))^2 = 0.00128$$

The variance:

$$Sx^2 = 977.333$$

$$t = \frac{\hat{a}}{\sqrt{\frac{\hat{\sigma}^2}{S_x^2}}} = -0.36992$$

At last, we get the

$$p\text{-value} = 0.7157 > 0.05$$

So we have to conclude that the equation is got by accident and the variable temperature is not linear with the concentration of $PM_{2.5}$, we should still look at the first picture in the regression part for temperature which converges to the same conclusion as the student's *t*-test.

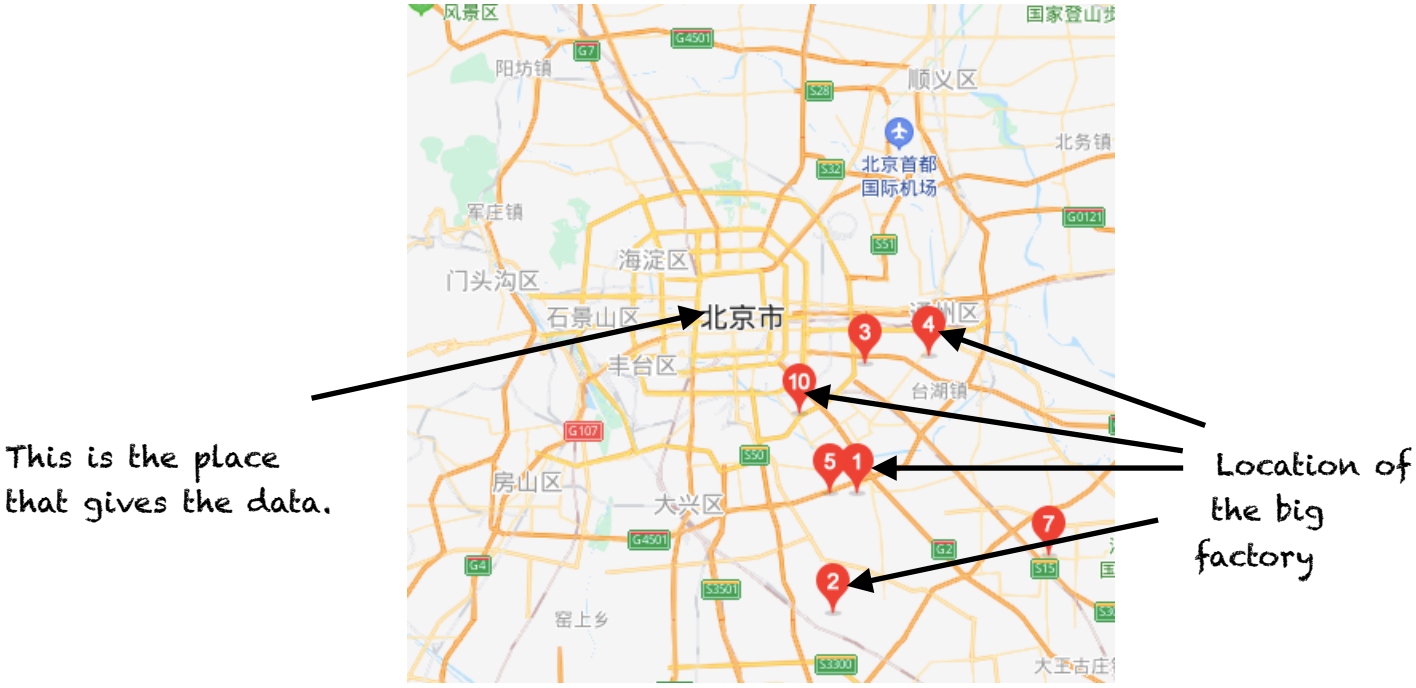
We then do the same calculation for the variable 'pressure','pm2.5' to do the student's test and unfortunately have the same conclusion as above, that is to say the pressure is not linear with the concentration of $PM_{2.5}$, either. The relation between them is much more complicated to be precise.

5. Conclusion

Until now, we have used nearly all the methods learned in class to analyze our data, the first part by box plot, PCA, the second part by the regression and the third part by chi 2 test, student's *t*-test and the test on the slope of regression.

Thanks to these methods, we have a cognition on our data. We could conclude that temperature and pressure has no obvious linear relation with the concentration of $PM_{2.5}$. More rain and snow can help to decrease the air pollution but Beijing is lack of rain and snow (unlucky).

Next, the wind direction is an important factor among these variables, and we find out that the high level of the air pollution is associated with the increment of the southeast wind and the calm and variable wind. But the northwest wind brings less pollution. However, from the data, we can see that the most of the time in Beijing is lack of wind which is also 'unlucky'. As for the raison of the extremely different influence of the two wind direction (NW, SE), we look at the map of Beijing. We first search the big big factories around Beijing.



As we can see, most of the factories are located in the southeast of Beijing, which means that when the wind blows from the southeast, the emissions from the factories will blow towards the city and increase the level of the air pollution.

To conclude, except the natural condition that we are not able to change, what we should take action as soon as possible is to deal with the big factories in the southwest of the city. If it is difficult to do the removal of these factories, at least……plant more trees!

Last but not least, much thanks for our teacher Stéphane CANU!

ANNEX

We define some functions to help avoid the repetition of typing since we originally have the data for 5 years. We need to extract the data to not only simplify the calculation but also remove the interference.

```
def extractlow(year):
    global timelow
    time = data[(data.year==year)]
    timelow = time[(time['pm2.5'])<35]
    return timelow
```

```
def extractmid(year):
    global timemid
    time= data[(data.year==year)]
    timemid=time[(time['pm2.5'])<=150]
    return timemid
```

```
def extracthigh(year):
    global timehigh
    time= data[(data.year==year)]
    timehigh=time[(time['pm2.5'])>150]
    return timehigh
```

```
datase= data3[(data3.cbwd=='SE')&(data3.hour==21)]
datacv= data3[(data3.cbwd=='cv')&(data3.hour==21)]
datasw= data3[(data3.cbwd=='SW')&(data3.hour==21)]
datane= data3[(data3.cbwd=='NE')&(data3.hour==21)]
datanw= data3[(data3.cbwd=='NW')&(data3.hour==21)]
#print(len(datase),(len(datacv)),(len(datane)),len(datanw))
datase.to_csv('data11.csv')
datacv.to_csv('data12.csv')
datasw.to_csv('data13.csv')
datane.to_csv('data14.csv')
datanw.to_csv('data15.csv')
```

```
x= datase['pm2.5']
y= datanw['pm2.5']
nx=len(x)
ny=len(y)
mx= mse
my= mnw
Sx = np.sum(x**2) - np.sum(x)**2/nx
Sy = np.sum(y**2) - np.sum(y)**2/ny
shat = (Sx + Sy)/(nx+ny-2)
t = (mx-my)/np.sqrt(shat*(1/nx+1/ny))
pval =2*(1 - stats.t.cdf(np.abs(t),nx+ny-2))
print(pval,t)
```

Link for location of the big factories in Beijing :

https://map.baidu.com/search/%E5%8C%97%E4%BA%AC%E8%BE%B9%E5%B7%A5%E5%8E%82%E5%9C%B0%E5%9B%BE/@12953768.621763784,4852694.540000006,10.89z?querytype=s&da_src=shareurl&wd=%E5%8C%97%E4%BA%AC%E8%BE%B9%E5%B7%A5%E5%8E%82%E5%9C%B0%E5%9B%BE&c=131&src=0&pn=0&sug=0&l=10&b=(12852905.42395985,4734631.557836366;13107769.558776183,4875538.315151741)&from=webmap&biz_forward=%7B%22scaler%22:1,%22styles%22:%22pl%22%7D&device_ratio=1