Data Management of PM 2.5 in Five Chinese Cities

STAT 440 Final Project Report

Group #2

Group member: Jinzhe Liu, Yuan Li, Xiruo Li, Chang Ma

● Introduction

a) Problem Statement

Air pollution in China has become a serious problem since 2010. The most important indicator, PM2.5, has been used extensively to measure air qualities. In general, the higher PM2.5 is, the lower quality the air is. In this project, we will be analyzing the PM2.5 data from five major cities in China. The important question we are asking is what factors can cause a high level of PM2.5. Are humidity, wind direction and air pressure some of the deterministic factors in influencing the PM2.5 levels?

Our main purpose is that after analyzing these data, we can gain a better understanding in terms of what decides the air quality and make suggestions on how to alleviate the air pollution issue in China.

b) Background Information

We will use five datasets for five cities respectively:
1) BeijingPM20100101_20151231.csv
2) ChengduPM20100101_20151231.csv
3) GuangzhouPM20100101_20151231.csv
4) ShanghaiPM20100101_20151231.csv
5) ShenyangPM20100101_20151231.csv

These two datasets are chosen from
http://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities#

These datasets were created by *Song Xi Chen* from Guanghua School of Management, Peking University. These datasets include information about different weather conditions, such as temperature, humidity and PM2.5, from 2010 to 2015 in Beijing, Chengdu, Guangzhou, Shanghai and Shenyang. However, since lots of the data have 'NA' values in them, which hinders us from performing the analyses. Thus, we will extract, manage and analyze the data that does not have 'NA' values.

● Methods

a) Datafile Description

The five data files used are all in CSV forms. There are 52584 observations and 15 major variables in each dataset. Some observations are missing and are recorded as “NA” in raw data files.

b) Original Variables Description

	Variable Name	Type	Description
1	No	Numeric	Row number
2	year	Numeric	Year of data in this row
3	month	Numeric	Month of data in this row
4	day	Numeric	Day of data in this row
5	hour	Numeric	Hour of data in this row
6	season	Numeric	Season of data in this row
7	PM	Numeric & Character	PM 2.5 concentration (μg/m^3) Including multiple variables of different sampling places
8	DEWP	Numeric	Dew Point (Degree Celsius)
9	TEMP	Numeric	Temperature (Degree Celsius)
10	HUMI	Numeric	Humidity (%)
11	PRES	Numeric	Pressure (hPa)
12	cbwd	Character	Combined wind direction
13	lws	Numeric	Cumulative wind speed (m/s)
14	precipitation	Numeric	Hourly precipitation (mm)
15	Iprec	Numeric	Cumulated precipitation (mm)

c) Data Input, Check, Cleaning and Validation

1) We concatenate the five datasets into one dataset called originaldata_prep and validate the number of observations in the combined data file matches the sum of five individual ones.

d) Additional Data Preparation

1) Averaged Hourly Data to Daily Data

1.1)Since there are 24 observations per day (one hour per obs), we use “First.Day” “Last.Day” to get the average daily value for the variable that will be analyzed: PM, DEWP, TEMP, HUMI, PRES, lws, cbwd. Specially, for cbwd, since it’s the wind direction and it’s a character variable, we select the cbwd of the last hour as the day’s wind direction. And, we use “output” to get the averaged daily observation

1.2) We use function “mdy” to set the original date format (variable year, month, day) as mm/dd/yyyy.

1.3) We add label and format for the variable we created above.(i.e. Set season 1 2 3 4 as Spring Summer Autumn Winter).

The table below is the first 10 obs for the variables we have created.Another table is their content.

City	Season	Date	PM 2.5	Dew Point (deg)	Temperature (deg)	Humidity (%)	Pressure (hPa)	Cumulated wind speed (m/s)	Combined wind direction
Beijing	Winter	01/01/2010	129.00	-17.00	-5.00	41.00	1020.00	0.89	CV
Beijing	Winter	01/02/2010	144.33	-8.50	-5.13	77.94	1024.75	24.86	SE
Beijing	Winter	01/03/2010	78.38	-10.13	-8.54	87.92	1022.79	70.94	NW
Beijing	Winter	01/04/2010	29.29	-20.88	-11.50	46.21	1029.29	111.16	NW
Beijing	Winter	01/05/2010	43.54	-24.58	-14.46	42.04	1033.63	56.92	NW
Beijing	Winter	01/06/2010	59.38	-23.71	-12.54	39.21	1033.75	18.51	NW
Beijing	Winter	01/07/2010	72.46	-21.25	-12.50	49.00	1034.08	10.17	CV
Beijing	Winter	01/08/2010	174.33	-17.13	-11.71	64.54	1028.00	1.97	NW
Beijing	Winter	01/09/2010	84.75	-16.33	-9.13	57.25	1029.04	13.30	CV
Beijing	Winter	01/10/2010	55.08	-15.96	-8.75	56.50	1032.50	17.42	NW

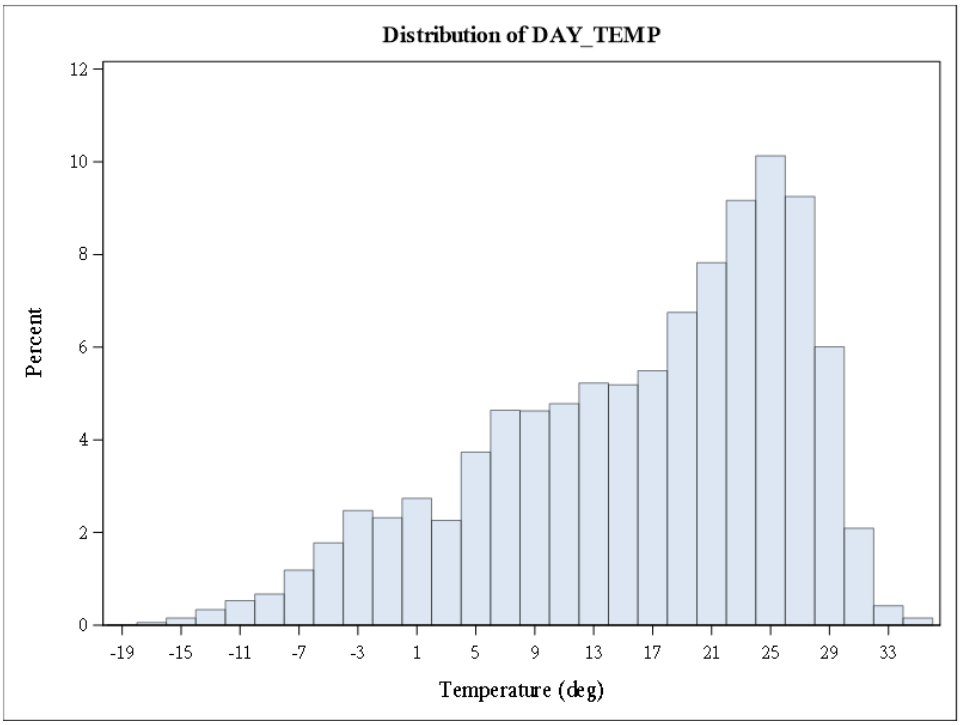
Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
1	City	Char	20		
5	DAY_DEWP	Num	8	10.2	Dew Point (deg)
7	DAY_HUMI	Num	8	10.2	Humidity (%)
4	DAY_PM	Num	8	10.2	PM 2.5
8	DAY_PRES	Num	8	10.2	Pressure (hPa)

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
6	DAY_TEMP	Num	8	10.2	Temperature (deg)
10	DAY_cbwd	Char	4		Combined wind direction
9	DAY_lws	Num	8	10.2	Cumulated wind speed (m/s)
2	Season	Num	8	SEASONFMT.	
3	date_new	Num	8	MMDDYY10.	Date

2) Data Classification

2.1) We use “proc univariate” to get the distribution of variables to be analyzed. For example, for DAY_TEMP variable, we have the quantiles table and histogram:

Quantiles (Definition 5)	
Level	Quantile
100% Max	35.708333
99%	31.260870
95%	28.958333
90%	27.687500
75% Q3	24.625000
50% Median	18.458333
25% Q1	8.895833
10%	0.291667
5%	-3.739130
1%	-10.291667
0% Min	-19.833333



2.2) According to the quantile, we divide each variable into four levels except the wind direction.

3)Tabulate Data

We use "proc tabulate" get the table for DAY_PM with each atmospheric factors (DAY_DEWP, DAY_TEMP, DAY_PRES, DAY_lws, DAY_cbwd) in five different cities. The specific analysis for the table is in the result part.

● Results

1) PM2.5 Level vs. Combined Wind Direction

1.1) Beijing

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Combined wind direction								
CV	95	18.41	78	15.12	109	21.12	234	45.35
NE	56	23.33	37	15.42	44	18.33	103	42.92
NW	193	34.53	82	14.67	113	20.21	171	30.59
SE	125	15.08	134	16.16	226	27.26	344	41.50

From this table we can see in Beijing High PM2.5 level (Level 4) happened more frequently when the wind direction is CV (almost no wind).

1.2) Chengdu

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Combined wind direction								
CV	43	7.48	147	25.57	180	31.30	205	35.65
NE	9	10.23	24	27.27	30	34.09	25	28.41
NW	28	9.86	63	22.18	99	34.86	94	33.10
SE	7	10.77	28	43.08	19	29.23	11	16.92
SW	27	12.56	70	32.56	67	31.16	51	23.72

From this table we can see in Chengdu High PM2.5 level (Level 4) happened more frequently when the wind direction is CV (almost no wind).

1.3) Guangzhou

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Combined wind direction								
CV	16	33.33	9	18.75	15	31.25	8	16.67
NE	156	35.70	120	27.46	114	26.09	47	10.76
NW	166	27.57	198	32.89	170	28.24	68	11.30
SE	113	52.80	52	24.30	35	16.36	14	6.54
SW	55	57.29	26	27.08	9	9.38	6	6.25

From this table we can see in Guangzhou High PM2.5 level (Level 4) happened more frequently when the wind direction is CV (almost no wind).

1.4) Shanghai

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Combined wind direction								
CV	8	17.02	11	23.40	16	34.04	12	25.53
NE	219	39.67	163	29.53	114	20.65	56	10.14
NW	44	21.67	65	32.02	63	31.03	31	15.27
SE	208	38.66	189	35.13	95	17.66	46	8.55
SW	28	24.35	40	34.78	27	23.48	20	17.39

From this table we can see in Shanghai High PM2.5 level (Level 4) happened more frequently when the wind direction is CV (almost no wind).

1.5) Shenyang

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Combined wind direction								
CV	21	27.63	17	22.37	15	19.74	23	30.26
NE	32	22.38	38	26.57	34	23.78	39	27.27
NW	22	22.00	16	16.00	33	33.00	29	29.00
SE	66	18.44	98	27.37	84	23.46	110	30.73
SW	59	23.05	82	32.03	73	28.52	42	16.41

From this table we can see in Shenyang High PM2.5 level (Level 4) happened more frequently when the wind direction is SE (Southwest wind).

2) PM2.5 Level vs. Dew Point Level

2.1) Beijing

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_DEWP								
1	281	27.58	158	15.51	205	20.12	375	36.80
2	81	20.61	67	17.05	96	24.43	149	37.91
3	93	20.22	73	15.87	116	25.22	178	38.70
4	14	5.15	33	12.13	75	27.57	150	55.15

From this table we can see in Beijing High PM2.5 level (Level 4) happened more frequently when the dew point level is 3.

2.2) Chengdu

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_DEWP								
1	.	.	11	18.03	25	40.98	25	40.98
2	18	3.96	95	20.93	101	22.25	240	52.86
3	55	13.65	109	27.05	146	36.23	93	23.08
4	41	13.27	117	37.86	123	39.81	28	9.06

From this table we can see in Chengdu High PM2.5 level (Level 4) happened more frequently when the dew point level is 2.

2.3) Guangzhou

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_DEWP								
1	5	18.52	9	33.33	11	40.74	2	7.41
2	52	20.08	62	23.94	96	37.07	49	18.92
3	79	21.18	112	30.03	131	35.12	51	13.67
4	370	50.14	222	30.08	105	14.23	41	5.56

From this table we can see in Guangzhou High PM2.5 level (Level 4) happened more frequently when the dew point level is 2.

2.4) Shanghai

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_DEWP								
1	38	14.90	73	28.63	85	33.33	59	23.14
2	86	18.94	150	33.04	139	30.62	79	17.40
3	127	38.14	127	38.14	57	17.12	22	6.61
4	256	61.99	118	28.57	34	8.23	5	1.21

From this table we can see in Shanghai High PM2.5 level (Level 4) happened more frequently when the dew point level is 1.

2.5) Shenyang

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_DEWP								
1	56	12.93	86	19.86	110	25.40	181	41.80
2	54	24.22	60	26.91	63	28.25	46	20.63
3	66	30.70	86	40.00	53	24.65	10	4.65
4	24	38.71	19	30.65	13	20.97	6	9.68

From this table we can see in Shenyang High PM2.5 level (Level 4) happened more frequently when the dew point level is 1.

3) PM2.5 Level vs. Temperature Level

3.1) Beijing

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_TEMP								
1	205	23.86	114	13.27	172	20.02	368	42.84
2	102	23.45	80	18.39	91	20.92	162	37.24
3	81	19.33	71	16.95	114	27.21	153	36.52
4	81	18.79	66	15.31	115	26.68	169	39.21

From this table we can see in Beijing High PM2.5 level (Level 4) happened more frequently when the temperature level is 1.

3.2) Chengdu

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_TEMP								
1	1	0.57	24	13.79	36	20.69	113	64.94
2	25	6.44	84	21.65	124	31.96	155	39.95
3	57	15.53	120	32.70	108	29.43	82	22.34
4	31	10.40	104	34.90	127	42.62	36	12.08

From this table we can see in Chengdu High PM2.5 level (Level 4) happened more frequently when the temperature level is 1.

3.3) Guangzhou

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_TEMP								
1	6	35.29	9	52.94	1	5.88	1	5.88
2	83	22.19	87	23.26	135	36.10	69	18.45
3	77	20.64	125	33.51	125	33.51	46	12.33
4	340	53.71	184	29.07	82	12.95	27	4.27

From this table we can see in Guangzhou High PM2.5 level (Level 4) happened more frequently when the temperature level is 2.

3.4) Shanghai

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_TEMP								
1	66	19.24	97	28.28	102	29.74	78	22.74
2	66	17.37	135	35.53	121	31.84	58	15.26
3	176	45.83	134	34.90	52	13.54	22	5.73
4	199	57.18	102	29.31	40	11.49	7	2.01

From this table we can see in Shanghai High PM2.5 level (Level 4) happened more frequently when the temperature level is 1.

3.5) Shenyang

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_TEMP								
1	50	12.63	74	18.69	91	22.98	181	45.71
2	51	23.83	56	26.17	68	31.78	39	18.22
3	77	31.30	94	38.21	56	22.76	19	7.72
4	22	28.57	27	35.06	24	31.17	4	5.19

From this table we can see in Shenyang High PM2.5 level (Level 4) happened more frequently when the temperature level is 1.

4) PM2.5 Level vs. Humidity Level

4.1) Beijing

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_HUMI								
1	372	34.38	203	18.76	252	23.29	255	23.57
2	63	13.13	68	14.17	118	24.58	231	48.13
3	13	4.18	32	10.29	68	21.86	198	63.67
4	21	7.75	28	10.33	54	19.93	168	61.99

From this table we can see in Beijing High PM2.5 level (Level 4) happened more frequently when the humidity level is 3.

4.2) Chengdu

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_HUMI								
1	11	10.00	33	30.00	35	31.82	31	28.18
2	29	8.61	100	29.67	102	30.27	106	31.45
3	32	8.00	100	25.00	127	31.75	141	35.25
4	42	11.05	99	26.05	131	34.47	108	28.42

From this table we can see in Chengdu High PM2.5 level (Level 4) happened more frequently when the humidity level is 3.

4.3) Guangzhou

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_HUMI								
1	5	8.93	15	26.79	27	48.21	9	16.07
2	43	20.98	67	32.68	73	35.61	22	10.73
3	169	35.88	134	28.45	119	25.27	49	10.40
4	289	43.46	189	28.42	124	18.65	63	9.47

From this table we can see in Guangzhou High PM2.5 level (Level 4) happened more frequently when the humidity level is 1.

4.4) Shanghai

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_HUMI								
1	23	11.06	74	35.58	84	40.38	27	12.98
2	156	33.91	149	32.39	82	17.83	73	15.87
3	167	40.14	127	30.53	82	19.71	40	9.62
4	161	43.40	118	31.81	67	18.06	25	6.74

From this table we can see in Shanghai High PM2.5 level (Level 4) happened more frequently when the humidity level is 2.

4.5) Shenyang

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_HUMI								
1	70	20.90	96	28.66	101	30.15	68	20.30
2	53	17.38	81	26.56	81	26.56	90	29.51
3	46	24.08	55	28.80	37	19.37	53	27.75
4	31	30.39	19	18.63	20	19.61	32	31.37

From this table we can see in Shenyang High PM2.5 level (Level 4) happened more frequently when the humidity level is 4.

5) PM2.5 Level vs. Pressure Level

5.1) Beijing

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_PRES								
1	69	17.42	68	17.17	99	25.00	160	40.40
2	88	19.17	69	15.03	119	25.93	183	39.87
3	98	18.01	84	15.44	118	21.69	244	44.85
4	214	28.72	110	14.77	156	20.94	265	35.57

From this table we can see in Beijing High PM2.5 level (Level 4) happened more frequently when the pressure level is 3.

5.2) Chengdu

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_PRES								
1	28	13.21	72	33.96	88	41.51	24	11.32
2	40	12.42	111	34.47	106	32.92	65	20.19
3	30	7.37	91	22.36	113	27.76	173	42.51
4	16	5.59	58	20.28	88	30.77	124	43.36

From this table we can see in Chengdu High PM2.5 level (Level 4) happened more frequently when the pressure level is 4.

5.3) Guangzhou

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_PRES								
1	386	49.11	239	30.41	116	14.76	45	5.73
2	80	18.35	123	28.21	159	36.47	74	16.97
3	40	22.86	43	24.57	68	38.86	24	13.71

From this table we can see in Guangzhou High PM2.5 level (Level 4) happened more frequently when the pressure level is 2.

5.4) Shanghai

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_PRES								
1	138	56.10	74	30.08	28	11.38	6	2.44
2	148	43.02	119	34.59	60	17.44	17	4.94
3	121	30.25	138	34.50	97	24.25	44	11.00
4	100	21.51	137	29.46	130	27.96	98	21.08

From this table we can see in Shanghai High PM2.5 level (Level 4) happened more frequently when the pressure level is 4.

5.5) Shenyang

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_PRES								
1	57	38.00	58	38.67	32	21.33	3	2.00
2	71	29.96	73	30.80	64	27.00	29	12.24
3	42	16.15	72	27.69	76	29.23	70	26.92
4	30	10.49	48	16.78	67	23.43	141	49.30

From this table we can see in Shenyang High PM2.5 level (Level 4) happened more frequently when the pressure level is 4.

6) PM2.5 Level vs. Cumulative wind speed Level

6.1) Beijing

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_lws								
1	16	4.49	28	7.87	61	17.13	251	70.51
2	88	15.91	94	17.00	148	26.76	223	40.33
3	124	18.18	115	16.86	187	27.42	256	37.54
4	241	43.58	94	17.00	96	17.36	122	22.06

From this table we can see in Beijing High PM2.5 level (Level 4) happened more frequently when the cumulative wind speed level is 1.

6.2) Chengdu

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_lws								
1	49	5.78	200	23.58	286	33.73	313	36.91
2	38	12.84	102	34.46	94	31.76	62	20.95
3	21	29.58	26	36.62	15	21.13	9	12.68
4	6	50.00	4	33.33	.	.	2	16.67

From this table we can see in Chengdu High PM2.5 level (Level 4) happened more frequently when the cumulative wind speed level is 1.

6.3) Guangzhou

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_lws								
1	150	31.12	141	29.25	131	27.18	60	12.45
2	191	34.41	171	30.81	135	24.32	58	10.45
3	132	43.42	84	27.63	67	22.04	21	6.91
4	33	58.93	9	16.07	10	17.86	4	7.14

From this table we can see in Guangzhou High PM2.5 level (Level 4) happened more frequently when the cumulative wind speed level is 1.

6.4) Shanghai

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_lws								
1	4	15.38	6	23.08	6	23.08	10	38.46
2	29	17.79	53	32.52	47	28.83	34	20.86
3	84	22.46	141	37.70	100	26.74	49	13.10
4	390	43.72	268	30.04	162	18.16	72	8.07

From this table we can see in Shanghai High PM2.5 level (Level 4) happened more frequently when the cumulative wind speed level is 1.

6.5) Shenyang

	Level_PM							
	1		2		3		4	
	N	RowPctN	N	RowPctN	N	RowPctN	N	RowPctN
Level_lws								
1	15	19.23	21	26.92	14	17.95	28	35.90
2	37	16.74	66	29.86	47	21.27	71	32.13
3	79	22.13	84	23.53	96	26.89	98	27.45
4	69	24.91	80	28.88	82	29.60	46	16.61

From this table we can see in Shenyang High PM2.5 level (Level 4) happened more frequently when the cumulative wind speed level is 1.