GTECH385 Final Project

Linear Regression Analysis of PM 2.5 in Five Cities of China

- Yu Qiao Chen (Tom)
- Professor Ni-Meister
- 5/21/2019

What is PM 2.5?

Particulate matter (PM2.5) is an air pollutant that is a concern for people's health when levels in air are high. PM2.5 are tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. PM2.5 refers to atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers, which is about 3% the diameter of a human hair. PM 2.5 is measured by micrograms per cubic meter (ug/m^3). It comes from various sources such as: vehicle exhaust, fuel burning, construction, volcano and more. Since PM 2.5 is so tiny and light, it usually stays in air longer and are more likely to be inhaled by people. Due to its size, PM 2.5 can bypass our nose, throat and enter our lungs. PM 2.5 causes heart disease, lung disease, asthma, and respiratory problems

PM 2.5 Measurements and Health Affect

- **0 to 12.0** (Good) Little to no risk
- **12.1 to 35.4** (Moderate) Unusually sensitive individuals may experience respiratory symptoms
- **35.5 to 55.4** (Unhealthy) Increasing likelihood of respiratory symptoms in sensitive individuals, aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly.

- **55.5 to 150.4** (Unhealthy) Increased aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; increased respiratory effects in general population.
- **150.5 to 250.4** (Very Unhealthy) Significant aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; significant increase in respiratory effects in general population.
- **250.5 to 500.4** (Harzardous) Serious aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; serious risk of respiratory effects in general population.
  Source: https://blissair.com/what-is-pm-2-5.htm (https://blissair.com/what-is-pm-2-5.htm)

Air Pollution in China

Air pollution have been a major environmental issue. It has became one of the biggest health threat to citizens of China. The severe air pollution of China in major cities, such as: Beijing, is attributed by an uprise in China's economy. More factors exist to manufacture goods, while residents become more capable of afford vehicles. Population increase has also contributed to worsen air quality in China. Governments of China has made effort in reducing its PM 2.5 by restricting vehicles on the road by their license plate.

中国五城市空气污染状况
之统计学分析
北京
沈阳
PM2.5
Statistical Modeling
PM2.5
成都
综合比较
上海
Air Quality Assessment
Meteorological Confounding
广州
风和降水
冬季供暖
PM2.5
北京大学光华管理学院
北京大学统计科学中心

About the Dataset

Cities:

1. Beijing
2. Chengdu
3. Guangzhou
4. Shanghai
5. Shenyang

Columns:

- No: row number
- year: year of data in this row
- month: month of data in this row
- day: day of data in this row
- hour: hour of data in this row
- season: season of data in this row
- **PM: PM2.5 concentration (ug/m^3)**
- **DEWP: Dew Point (Celsius Degree)**
- **TEMP: Temperature (Celsius Degree)**
- **HUMI: Humidity (%)**
- **PRES: Pressure (hPa)**
- cbwd: Combined wind direction
- **Iws: Cumulated wind speed (m/s)**
- precipitation: hourly precipitation (mm)
- Iprec: Cumulated precipitation (mm)

Part 1. Setting Up

Data Preparation in Excel

1. Concatenated TIMESTAMPS
2. Calculated monthly average of PM2.5, Dew Point, Temperature, Humidity, Pressure, Wind Speed for each city
3. Aggregated all the monthly average spreadsheets into one CSV (FiveCitiesAggregate.csv)

In [1]:

```
# ----------  importing necessary package for the project ----------
import os

import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.impute import SimpleImputer

import warnings
warnings.filterwarnings("ignore")

# ---------- print "success" when successfully imported packages ----------
print('success')
```

```
success
```

In [2]:

```
# ---------- check for the current working directory ----------
os.getcwd()
```

Out[2]:

```
'C:\\Users\\ychen\\Desktop\\Labs\\Final'
```

In [3]:

```
# ---------- creating a column list of the five cities CSV in working director
y ----------
cwd = '.'
fiveCities = []
for file in os.listdir(cwd):
    if 'PM20100101_20151231' in file:
        fiveCities.append(file)

print(fiveCities)

# ---------- creating a column list of the averaged monthly five cities CSV in
working directory ----------
monthlyList = []
for file in os.listdir(cwd):
    if 'Monthly' in file:
        monthlyList.append(file)

print(monthlyList)

# ---------- writing a function to import CSVs into pandas dataframe ---------
-
def import_pd(a):
    city = pd.read_csv(a, na_values = ['NA', '-9999'])
    return city
```

```
['BeijingPM20100101_20151231.csv', 'ChengduPM20100101_20151231.csv', 'Guangzh
ouPM20100101_20151231.csv', 'ShanghaiPM20100101_20151231.csv', 'ShenyangPM201
00101_20151231.csv']
['BeijingPM2010_2015Monthly.csv', 'ChengduPM2010_2015Monthly.csv', 'Guangzhou
PM2010_2015Monthly.csv', 'ShanghaiPM2010_2015Monthly.csv', 'ShenyangPM2010_20
15Monthly.csv']
```

In [4]:

```
# ---------- checking the first couple of rows of the dataframe ----------
for sheet in fiveCities:
    city = import_pd(sheet)
    print(sheet)
    print(city.head())
```

```
BeijingPM20100101_20151231.csv
   No  year  month  day  TIMESTAMP  hour  minuteSec  DATETIME  season  \
0   1  2010      1    1   20100101     0          0  1/1/2010       4   
1   2  2010      1    1   20100101     1          0  1/1/2010       4   
2   3  2010      1    1   20100101     2          0  1/1/2010       4   
3   4  2010      1    1   20100101     3          0  1/1/2010       4   
4   5  2010      1    1   20100101     4          0  1/1/2010       4   

   PM_Dongsi  ...    PM_Nongzhanguan  PM_USPost  DEWP  HUMI    PRES  TEMP  \
0        NaN  ...                NaN        NaN -21.0  43.0  1021.0 -11.0   
1        NaN  ...                NaN        NaN -21.0  47.0  1020.0 -12.0   
2        NaN  ...                NaN        NaN -21.0  43.0  1019.0 -11.0   
3        NaN  ...                NaN        NaN -21.0  55.0  1019.0 -14.0   
4        NaN  ...                NaN        NaN -20.0  51.0  1018.0 -12.0   

   cbwd    Iws  precipitation  Iprec  
0    NW   1.79            0.0    0.0  
1    NW   4.92            0.0    0.0  
2    NW   6.71            0.0    0.0  
3    NW   9.84            0.0    0.0  
4    NW  12.97            0.0    0.0  

[5 rows x 21 columns]
ChengduPM20100101_20151231.csv
   No  year  month  day  TIMESTAMP  hour  minuteSec  DATETIME  season  \
0   1  2010      1    1   20100101     0          0  1/1/2010       4   
1   2  2010      1    1   20100101     1          0  1/1/2010       4   
2   3  2010      1    1   20100101     2          0  1/1/2010       4   
3   4  2010      1    1   20100101     3          0  1/1/2010       4   
4   5  2010      1    1   20100101     4          0  1/1/2010       4   

   PM_Caotangsi  PM_Shahepu  PM_USPost  DEWP   HUMI    PRES  TEMP cbwd  Iws  
\
0           NaN         NaN        NaN   4.0  81.20  1022.0   7.0   cv  1.0   
1           NaN         NaN        NaN   4.0  86.99  1022.0   6.0   cv  1.0   
2           NaN         NaN        NaN   4.0  86.99  1021.0   6.0   cv  1.0   
3           NaN         NaN        NaN   3.0  86.89  1021.0   5.0   cv  1.0   
4           NaN         NaN        NaN   2.0  86.79  1021.0   4.0   cv  1.0   

   precipitation  Iprec  
0            0.0    0.0  
1            0.0    0.0  
2            0.0    0.0  
3            0.0    0.0  
4            0.0    0.0  
GuangzhouPM20100101_20151231.csv
   No  year  month  day  TIMESTAMP  hour  minuteSec  DATETIME  season  \
0   1  2010      1    1   20100101     0          0  1/1/2010     4.0   
1   2  2010      1    1   20100101     1          0  1/1/2010     4.0   
2   3  2010      1    1   20100101     2          0  1/1/2010     4.0   
3   4  2010      1    1   20100101     3          0  1/1/2010     4.0   
4   5  2010      1    1   20100101     4          0  1/1/2010     4.0   

   PM_City Station  PM_5th Middle School  PM_USPost  DEWP  HUMI    PRES  TEMP  
\
0              NaN                   NaN        NaN   9.4  76.0  1015.1  13.5   
1              NaN                   NaN        NaN  10.2  83.0  1015.2  13.0   
```

```
2              NaN                   NaN        NaN  10.4  87.0  1015.0  12.5
3              NaN                   NaN        NaN  10.2  89.0  1014.9  12.0
4              NaN                   NaN        NaN  10.4  91.0  1014.6  11.8

  cbwd  Iws  precipitation  Iprec
0   NW  0.8            0.0    0.0
1   cv  0.5            0.0    0.0
2   NW  0.6            0.3    0.3
3   NW  1.4            0.6    0.9
4   NE  0.6            0.7    1.6
ShanghaiPM20100101_20151231.csv
   No  year  month  TIMESTAMP  day  hour  minuteSec  DATETIME  season  \
0   1  2010      1   20100101    1     0          0  1/1/2010       4   
1   2  2010      1   20100101    1     1          0  1/1/2010       4   
2   3  2010      1   20100101    1     2          0  1/1/2010       4   
3   4  2010      1   20100101    1     3          0  1/1/2010       4   
4   5  2010      1   20100101    1     4          0  1/1/2010       4   

   PM_Jingan  PM_USPost  PM_Xuhui  DEWP   HUMI    PRES  TEMP cbwd  Iws  \
0        NaN        NaN       NaN  -6.0  59.48  1026.1   1.0   cv  1.0   
1        NaN        NaN       NaN  -6.0  59.48  1025.1   1.0   SE  2.0   
2        NaN        NaN       NaN  -7.0  59.21  1025.1   0.0   SE  4.0   
3        NaN        NaN       NaN  -6.0  63.94  1024.0   0.0   SE  5.0   
4        NaN        NaN       NaN  -6.0  63.94  1023.0   0.0   SE  8.0   

   precipitation  Iprec
0            0.0    0.0
1            0.0    0.0
2            0.0    0.0
3            0.0    0.0
4            0.0    0.0
ShenyangPM20100101_20151231.csv
   No  year  month  TIMESTAMP  day  hour  minuteSec  DATETIME  season  \
0   1  2010      1   20100101    1     0          0  1/1/2010       4   
1   2  2010      1   20100101    1     1          0  1/1/2010       4   
2   3  2010      1   20100101    1     2          0  1/1/2010       4   
3   4  2010      1   20100101    1     3          0  1/1/2010       4   
4   5  2010      1   20100101    1     4          0  1/1/2010       4   

   PM_Taiyuanjie  PM_USPost  PM_Xiaoheyan  DEWP   HUMI    PRES  TEMP cbwd  \
0            NaN        NaN           NaN -26.0  69.79  1024.0 -22.0   NE   
1            NaN        NaN           NaN -26.0  76.26  1024.0 -23.0   NE   
2            NaN        NaN           NaN -27.0  69.56  1023.0 -23.0   NE   
3            NaN        NaN           NaN -27.0  69.56  1023.0 -23.0   NE   
4            NaN        NaN           NaN -27.0  69.56  1022.0 -23.0   NE   

      Iws  precipitation  Iprec
0  1.0289            NaN    NaN
1  2.5722            NaN    NaN
2  5.1444            NaN    NaN
3  7.7166            NaN    NaN
4  9.7744            NaN    NaN
```

In [5]:

```
# --------- Describe the monthly averaged PM 2.5 for all cities
for sheet in monthlyList:
    city = import_pd(sheet)
    print(sheet)
    print(city.head())
```

```
BeijingPM2010_2015Monthly.csv
   year  month  TIMESTAMP  PM_USPost   DEWP   HUMI     PRES   TEMP    lws  \
0  2010      1     201001      90.40 -17.01  46.45  1028.01  -6.16  41.23   
1  2010      2     201002      97.24 -13.16  47.64  1023.78  -1.92  13.47   
2  2010      3     201003      94.05  -7.96  49.84  1021.81   3.29  23.28   
3  2010      4     201004      80.07  -3.33  43.12  1017.17  10.81  58.28   
4  2010      5     201005      87.07   7.65  48.20  1007.90  20.83  21.42   

       loc  
0  Beijing  
1  Beijing  
2  Beijing  
3  Beijing  
4  Beijing  
ChengduPM2010_2015Monthly.csv
   year  month  TIMESTAMP  PM_USPost   DEWP   HUMI     PRES   TEMP   lws  \
0  2010      1     201001        NaN   2.52  68.09  1021.58   8.67  3.81   
1  2010      2     201002        NaN   2.89  65.32  1016.37   9.56  4.44   
2  2010      3     201003        NaN   5.22  61.62  1016.94  13.43  5.33   
3  2010      4     201004        NaN  10.49  74.67  1016.01  15.45  4.16   
4  2010      5     201005        NaN  15.77  74.61  1009.08  21.04  4.38   

       loc  
0  Chengdu  
1  Chengdu  
2  Chengdu  
3  Chengdu  
4  Chengdu  
GuangzhouPM2010_2015Monthly.csv
   year  month  TIMESTAMP  PM_USPost   DEWP   HUMI     PRES   TEMP   lws  \
0  2010      1     201001        NaN  10.20  75.50  1016.49  14.84  3.50   
1  2010      2     201002        NaN  13.24  81.70  1012.15  16.52  5.09   
2  2010      3     201003        NaN  13.51  71.48  1012.44  19.28  9.79   
3  2010      4     201004        NaN  16.82  81.25  1010.10  20.33  5.03   
4  2010      5     201005        NaN  21.44  77.15  1003.69  26.08  7.62   

         loc  
0  Guangzhou  
1  Guangzhou  
2  Guangzhou  
3  Guangzhou  
4  Guangzhou  
ShanghaiPM2010_2015Monthly.csv
   year  month  TIMESTAMP  PM_USPost   DEWP   HUMI     PRES   TEMP    lws  \
0  2010      1     201001        NaN   0.04  69.31  1026.16   5.52  38.49   
1  2010      2     201002        NaN   3.10  74.49  1021.02   7.71  37.48   
2  2010      3     201003        NaN   3.77  71.08  1020.74   9.40  54.58   
3  2010      4     201004        NaN   6.85  69.51  1018.23  13.09  51.43   
4  2010      5     201005        NaN  14.28  68.62  1010.29  20.91  35.73   

        loc  
0  Shanghai  
1  Shanghai  
2  Shanghai  
3  Shanghai  
4  Shanghai  
ShenyangPM2010_2015Monthly.csv
```

```
   year  month  TIMESTAMP  PM_USPost   DEWP   HUMI     PRES   TEMP    lws  \
0  2010      1     201001        NaN -17.02  66.79  1025.80 -11.76  10.54   
1  2010      2     201002        NaN -14.29  62.21  1023.94  -7.92   8.63   
2  2010      3     201003        NaN  -8.84  59.09  1021.00  -1.21  14.25   
3  2010      4     201004        NaN  -2.73  58.07  1016.39   6.52  19.05   
4  2010      5     201005        NaN   9.76  66.03  1007.63  16.98  10.18   

        loc  
0  Shenyang  
1  Shenyang  
2  Shenyang  
3  Shenyang  
4  Shenyang  
```

Part 2. Data Analysis

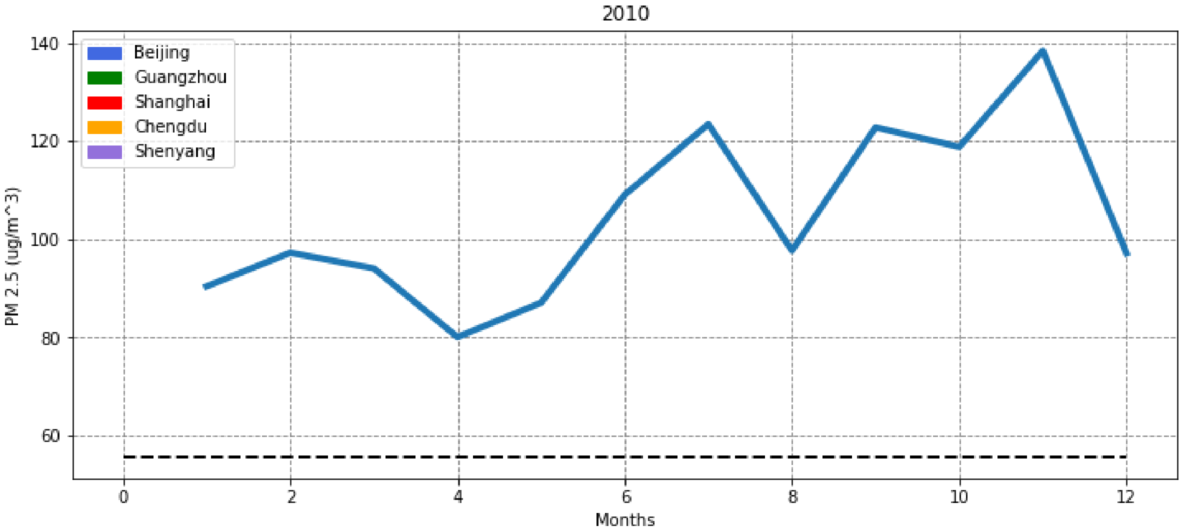
In [6]:

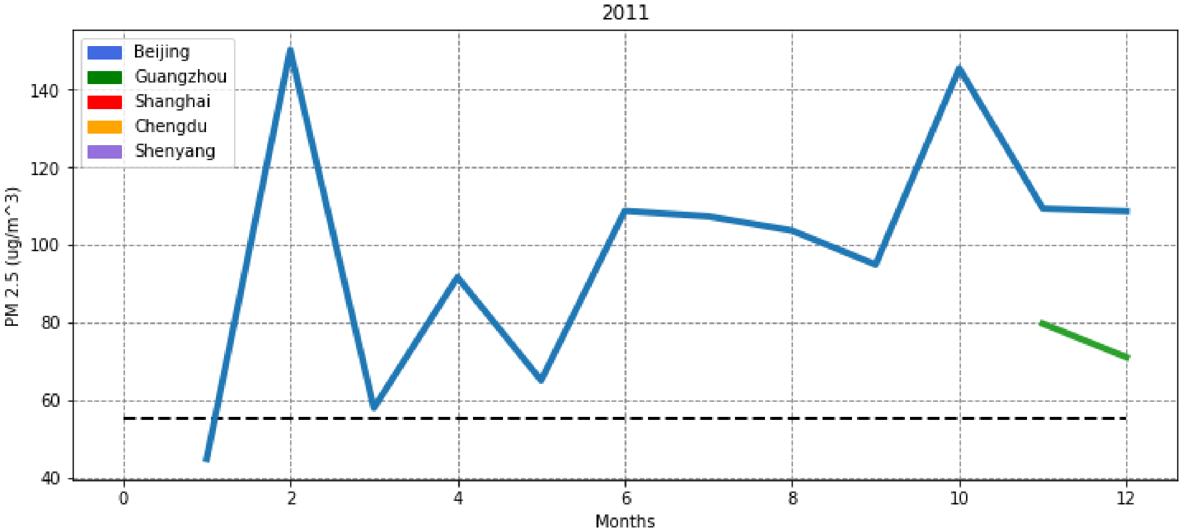
```python
# ---------- writing a function to plot the averaged monthly PM 2.5 by city --
--------
def plotMonthly(e, year):
    plt.rcParams["figure.figsize"] = (12,5)
    fig,(ax1) = plt.subplots()
    ax1.grid(color = 'grey', linestyle = '--')
    ax1.set(xlabel = 'Months', ylabel = 'PM 2.5 (ug/m^3)')
    #plotting a straight line along the x-axis at 55.5 to indicate a a boarder
line between acceptable and dangerous air quality
    marker = 55.5
    points = np.ones(13)
    ax1.plot(marker* points, linestyle='--', color = 'black', linewidth=2)
    # legends
    blue_patch = mpatches.Patch(color = 'royalblue', label = 'Beijing')
    green_patch = mpatches.Patch(color = 'green', label = 'Guangzhou')
    red_patch = mpatches.Patch(color = 'red', label = 'Shanghai')
    orange_patch = mpatches.Patch(color = 'orange', label = 'Chengdu')
    purple_patch = mpatches.Patch(color = 'mediumpurple', label = 'Shenyang')
    plt.legend(handles = [blue_patch, green_patch, red_patch, orange_patch, pu
rple_patch])
    # running a for loop to plot all the cities one by one
    for sheet in e:
        city = import_pd(sheet)
        cityByYear = city[city.year == year]
        ax1.plot(cityByYear.month, cityByYear.PM_USPost, linewidth = 4)
        ax1.set(title = year)

# ---------- print success ----------
print('Success')
```

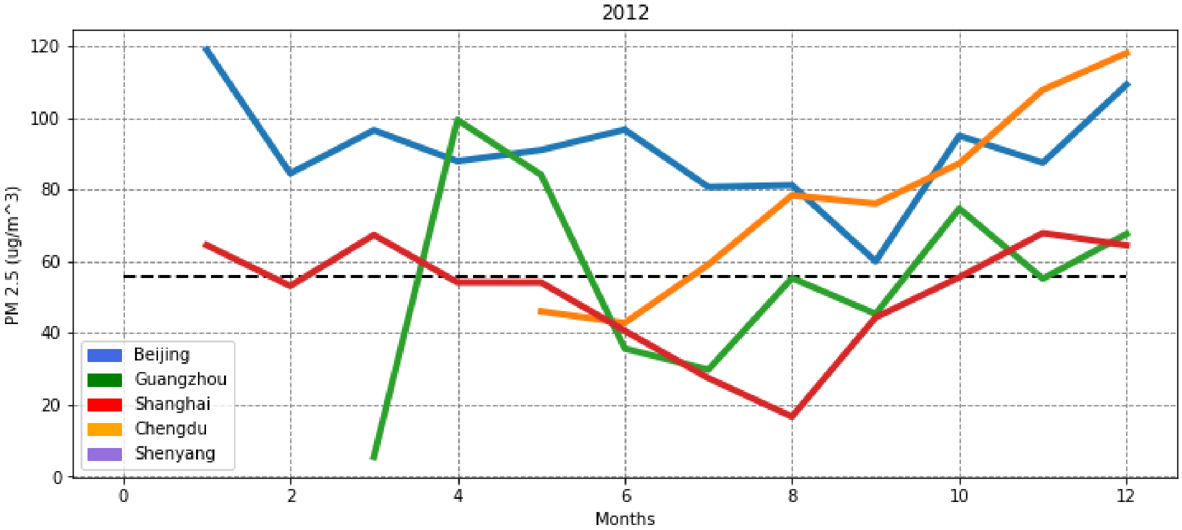
```
Success
```

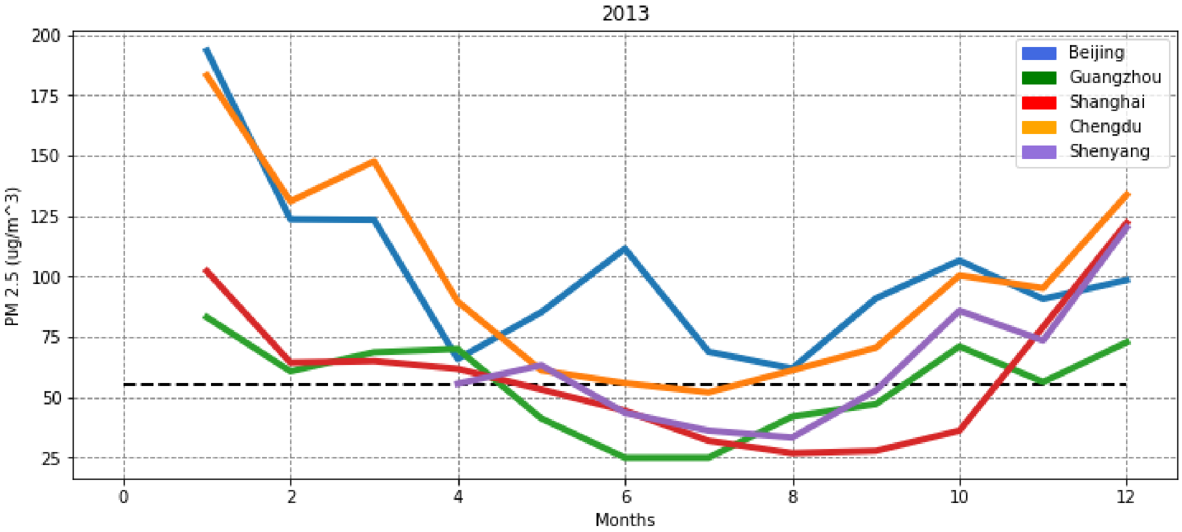
In [7]:

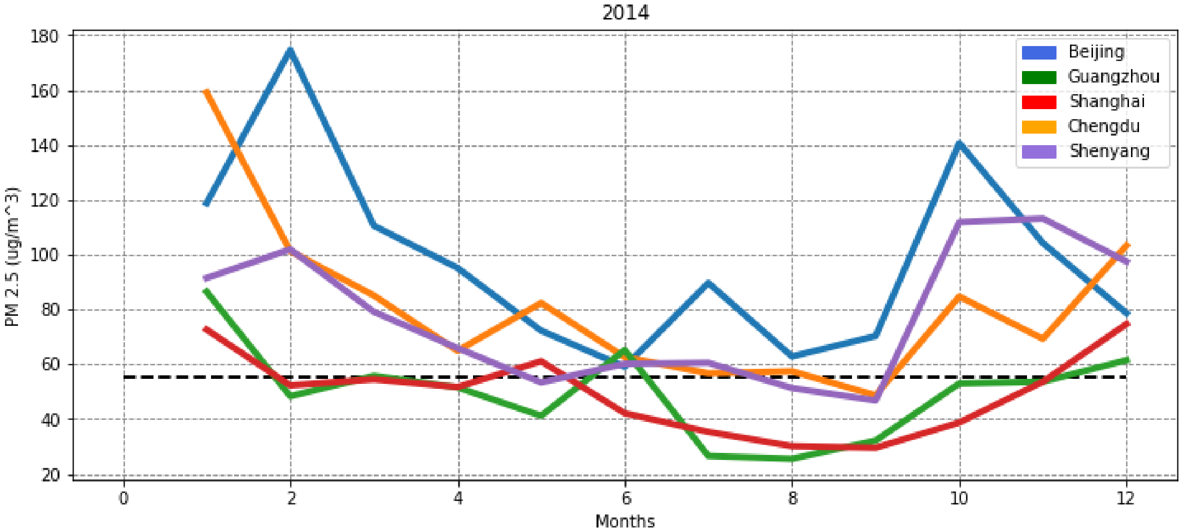
```
# ---------- Run a for loop to plot the data for all years ----------
for year in [2010, 2011, 2012, 2013, 2014, 2015]:
    plotMonthly(monthlyList, year)
```

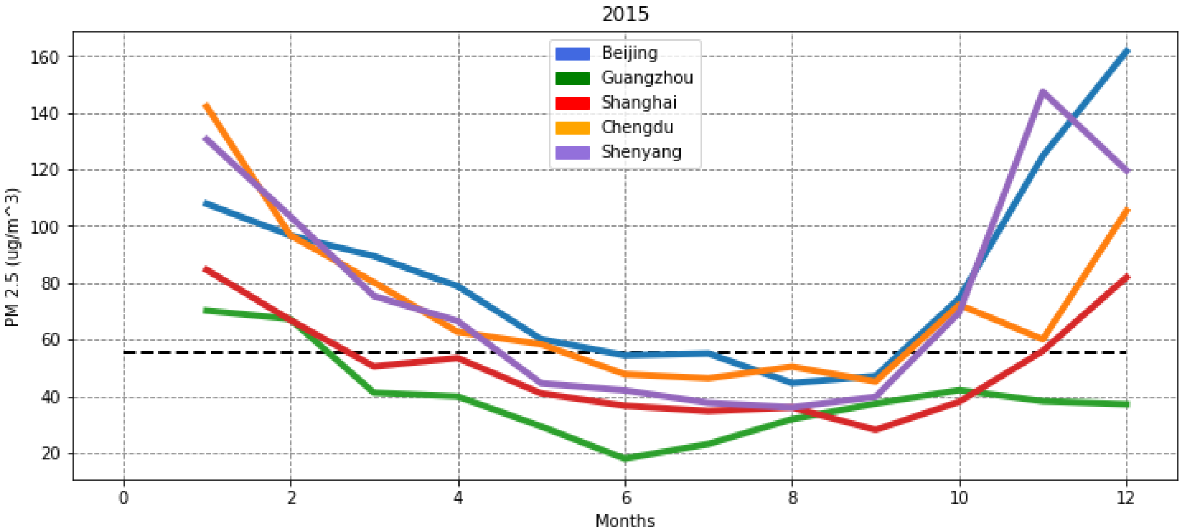

2010
Beijing
Guangzhou
Shanghai
Chengdu
Shenyang
PM 2.5 (ug/m^3)
140
120
100
80
60
0
2
4
6
8
10
12
Months


2011
Beijing
Guangzhou
Shanghai
Chengdu
Shenyang
PM 2.5 (ug/m^3)
140
120
100
80
60
40
0
2
4
6
8
10
12
Months


2012
Beijing
Guangzhou
Shanghai
Chengdu
Shenyang
PM 2.5 (ug/m^3)
120
100
80
60
40
20
0
0
2
4
6
8
10
12
Months


2013
Beijing
Guangzhou
Shanghai
Chengdu
Shenyang
PM 2.5 (ug/m^3)
200
175
150
125
100
75
50
25
0
2
4
6
8
10
12
Months


2014
Beijing
Guangzhou
Shanghai
Chengdu
Shenyang
PM 2.5 (ug/m^3)
180
160
140
120
100
80
60
40
20
0
2
4
6
8
10
12
Months


2015
Beijing
Guangzhou
Shanghai
Chengdu
Shenyang
PM 2.5 (ug/m^3)
160
140
120
100
80
60
40
20
0
2
4
6
8
10
12
Months

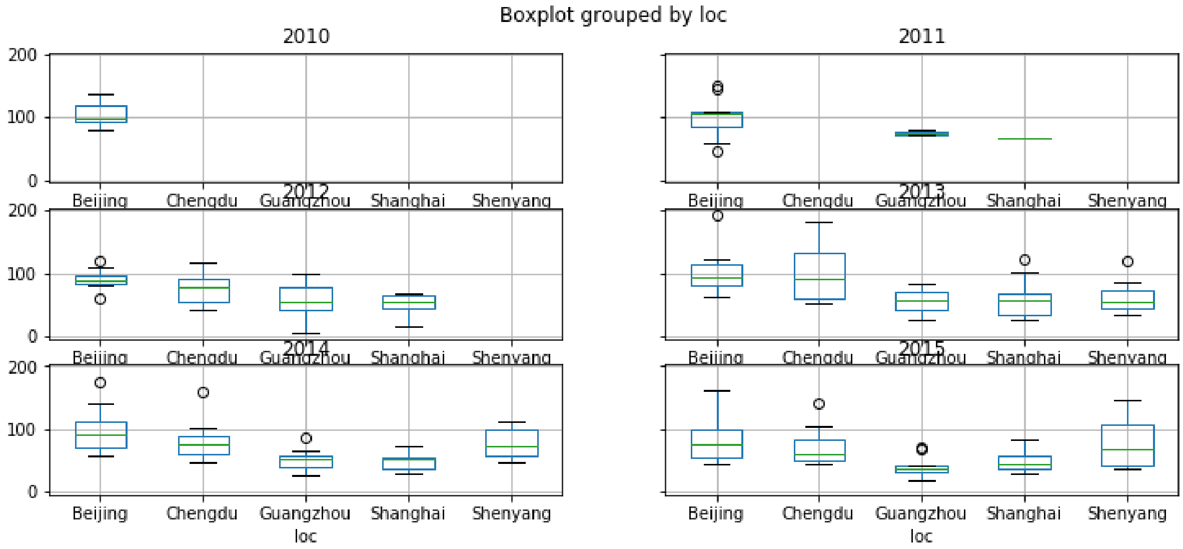
Boxplots

In [8]:

```
# ---------- read the aggregated CSV into pandas dataframe for analysis ------
----
fiveCities_pd = pd.read_csv('FiveCitiesAggregate.csv', na_values = ['NA'])

# ---------- plot the PM 2.5 of all cities in boxplots by year to compare diff
erences between cities ----------
fiveCities_pd.groupby('year').boxplot(column = 'PM_USPost', by = 'loc')
```

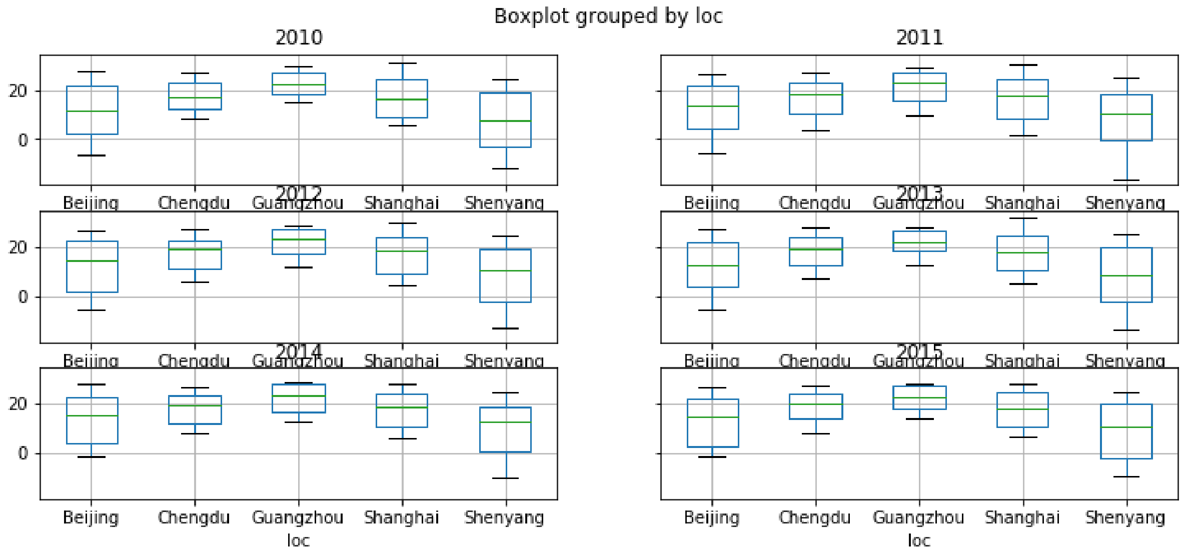
Out[8]:

```
2010         AxesSubplot(0.1,0.679412;0.363636x0.220588)
2011    AxesSubplot(0.536364,0.679412;0.363636x0.220588)
2012         AxesSubplot(0.1,0.414706;0.363636x0.220588)
2013    AxesSubplot(0.536364,0.414706;0.363636x0.220588)
2014            AxesSubplot(0.1,0.15;0.363636x0.220588)
2015       AxesSubplot(0.536364,0.15;0.363636x0.220588)
dtype: object
```



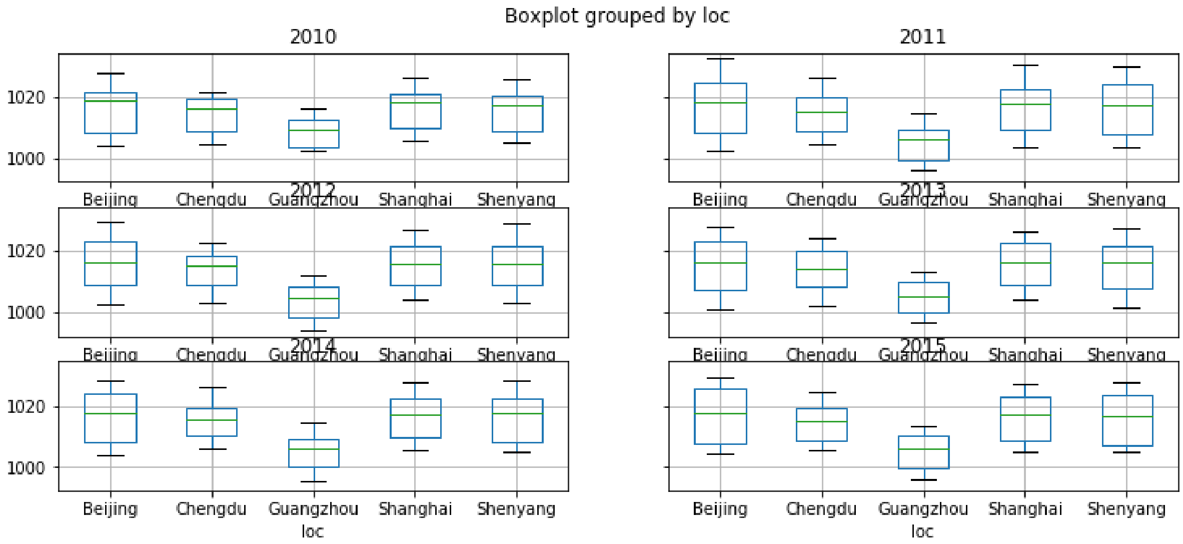
```
In [9]: # ---------- plot the temperature of all cities in boxplots by year to compare
        differences between cities ----------
        fiveCities_pd.groupby('year').boxplot(column = 'TEMP', by = 'loc')
```

```
Out[9]: 2010          AxesSubplot(0.1,0.679412;0.363636x0.220588)
        2011     AxesSubplot(0.536364,0.679412;0.363636x0.220588)
        2012          AxesSubplot(0.1,0.414706;0.363636x0.220588)
        2013     AxesSubplot(0.536364,0.414706;0.363636x0.220588)
        2014              AxesSubplot(0.1,0.15;0.363636x0.220588)
        2015         AxesSubplot(0.536364,0.15;0.363636x0.220588)
        dtype: object
```



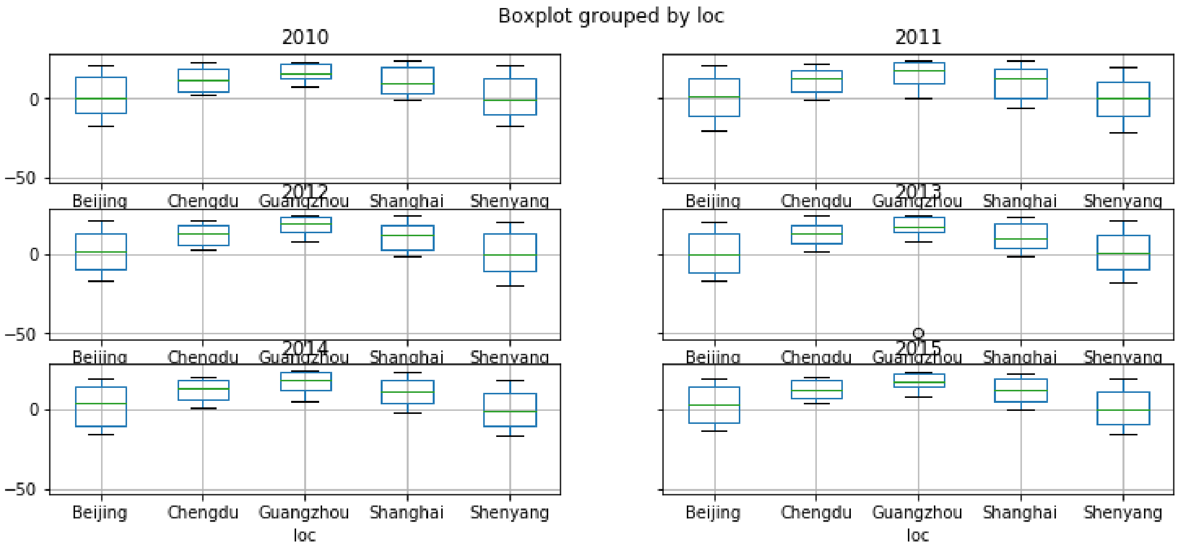
```
In [10]: # ---------- plot the pressure of all cities in boxplots by year to compare di
         fferences between cities ----------
         fiveCities_pd.groupby('year').boxplot(column = 'PRES', by = 'loc')
```

```
Out[10]: 2010          AxesSubplot(0.1,0.679412;0.363636x0.220588)
         2011     AxesSubplot(0.536364,0.679412;0.363636x0.220588)
         2012          AxesSubplot(0.1,0.414706;0.363636x0.220588)
         2013     AxesSubplot(0.536364,0.414706;0.363636x0.220588)
         2014              AxesSubplot(0.1,0.15;0.363636x0.220588)
         2015         AxesSubplot(0.536364,0.15;0.363636x0.220588)
         dtype: object
```



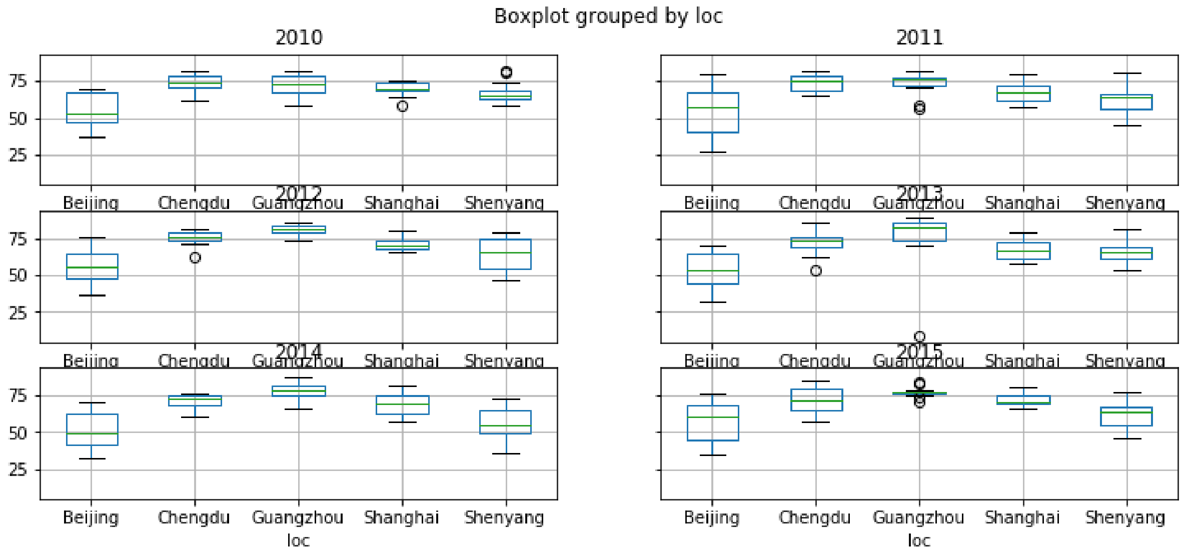
In [11]:
```
# ---------- plot the dew point of all cities in boxplots by year to compare d
ifferences between cities ----------
fiveCities_pd.groupby('year').boxplot(column = 'DEWP', by = 'loc')
```

Out[11]:
```
2010         AxesSubplot(0.1,0.679412;0.363636x0.220588)
2011    AxesSubplot(0.536364,0.679412;0.363636x0.220588)
2012         AxesSubplot(0.1,0.414706;0.363636x0.220588)
2013    AxesSubplot(0.536364,0.414706;0.363636x0.220588)
2014            AxesSubplot(0.1,0.15;0.363636x0.220588)
2015       AxesSubplot(0.536364,0.15;0.363636x0.220588)
dtype: object
```



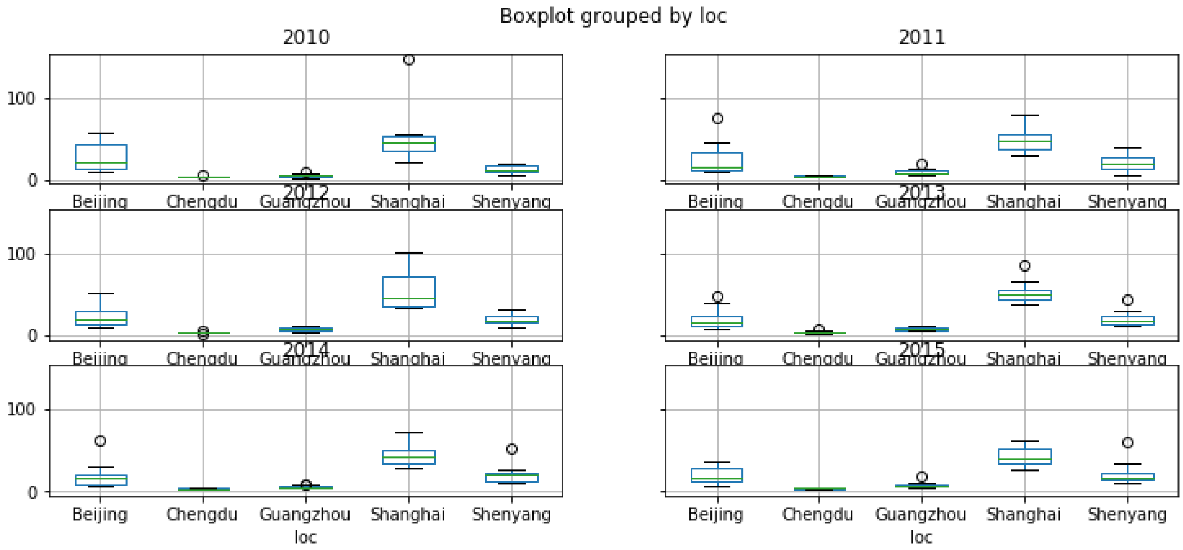
In [12]:
```
# ---------- plot the humidity of all cities in boxplots by year to compare di
fferences between cities ----------
fiveCities_pd.groupby('year').boxplot(column = 'HUMI', by = 'loc')
```

Out[12]:
```
2010         AxesSubplot(0.1,0.679412;0.363636x0.220588)
2011    AxesSubplot(0.536364,0.679412;0.363636x0.220588)
2012         AxesSubplot(0.1,0.414706;0.363636x0.220588)
2013    AxesSubplot(0.536364,0.414706;0.363636x0.220588)
2014             AxesSubplot(0.1,0.15;0.363636x0.220588)
2015        AxesSubplot(0.536364,0.15;0.363636x0.220588)
dtype: object
```



In [13]:
```
# ---------- plot the wind speed of all cities in boxplots by year to compare
 differences between cities ----------
fiveCities_pd.groupby('year').boxplot(column = 'lws', by = 'loc')
```

Out[13]:
```
2010         AxesSubplot(0.1,0.679412;0.363636x0.220588)
2011    AxesSubplot(0.536364,0.679412;0.363636x0.220588)
2012         AxesSubplot(0.1,0.414706;0.363636x0.220588)
2013    AxesSubplot(0.536364,0.414706;0.363636x0.220588)
2014             AxesSubplot(0.1,0.15;0.363636x0.220588)
2015        AxesSubplot(0.536364,0.15;0.363636x0.220588)
dtype: object
```



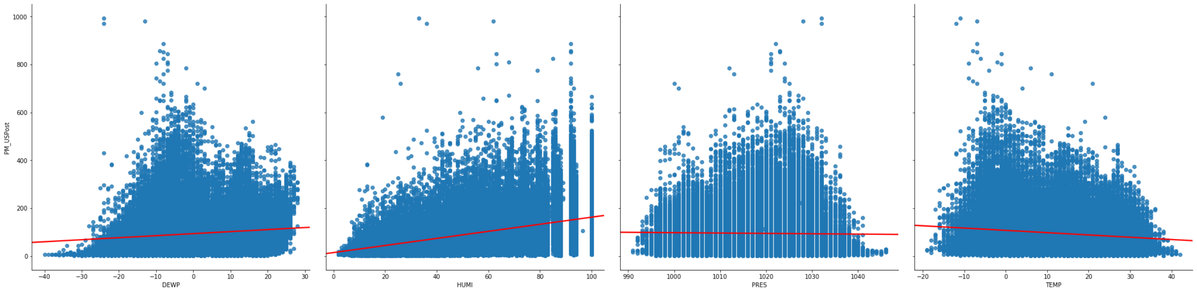
Part 3. Regression

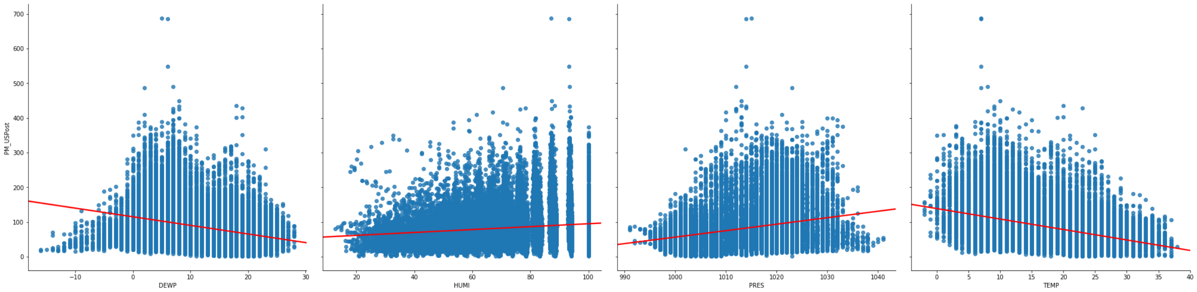
In [18]:

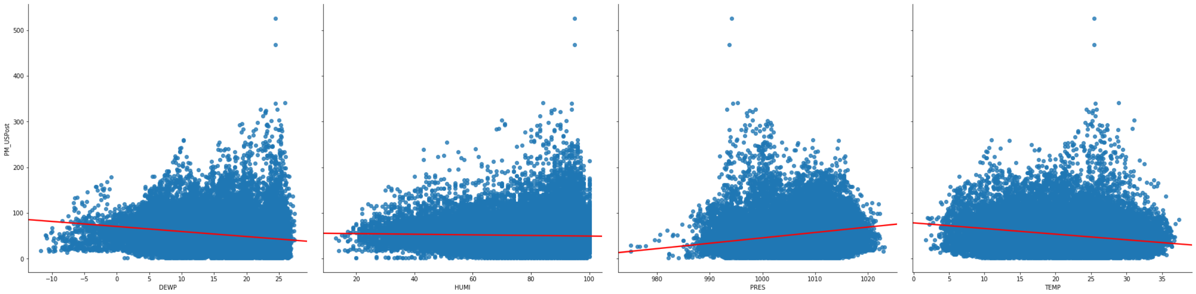
```
# ---------- writing a function that creates pairplots using seaborn ---------
-
def pair_plot(b):
    sns.pairplot(b,
                 # set X and Y variables
                 x_vars = ['DEWP', 'HUMI', 'PRES', 'TEMP', 'lws'],
                 y_vars = 'PM_USPost',
                 height = 7,
                 kind = 'reg',
                 # fit a line through the scatter plot
                 plot_kws={'line_kws':{'color':'red'}}
                )

# ---------- writing a function that runs linear regression and returns the co
efficient, intercept, and R^2 ----------
def linearReg(c, xValue):
    # Independent Variable "X" can be assigned in a list of any number and any
variable
    X = c[xValue]
    y = c['PM_USPost']
    X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 1
)
    # replace NA values with the mean of the data
    X_train.fillna(X_train.mean(), inplace = True)
    X_test.fillna(X_test.mean(), inplace = True)
    y_train.fillna(y_train.mean(), inplace = True)
    y_test.fillna(y_test.mean(), inplace = True)
    linreg = LinearRegression()
    linreg.fit(X_train, y_train)
    scores = linreg.score(X_train, y_train)
    # print out the intercept, coefficient,
    print('Intercept: ', linreg.intercept_, 'Coefficient: ', linreg.coef_, 'R^
2: ', scores)

# ---------- print success when functions are successfully implemented -------
---
print('success')
```

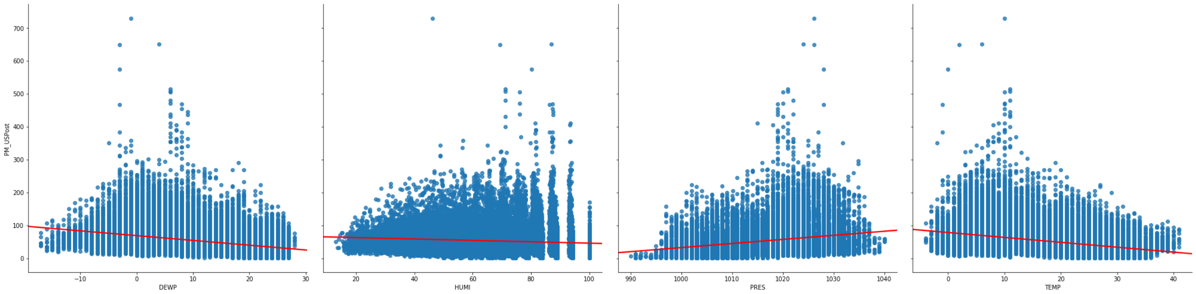
```
success
```

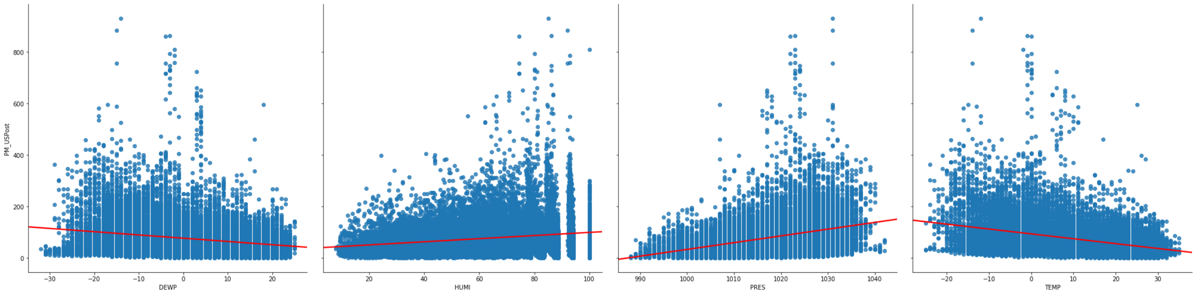
In [17]:
```
# ---------- running function to plot pairplots ----------
for city in fiveCities:
    city_pd = import_pd(city)
    pair_plot(city_pd)
```











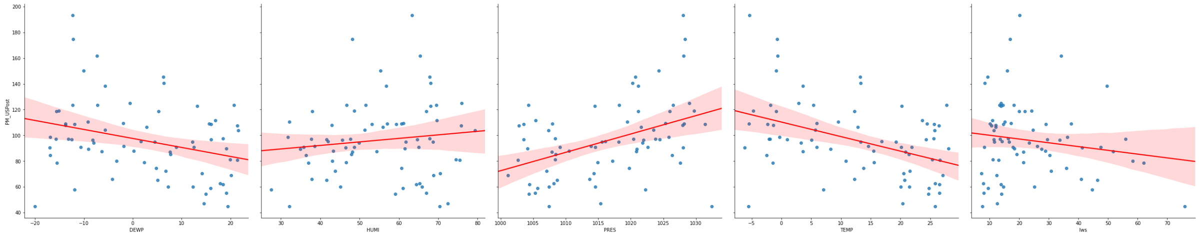
```
In [19]: # ---------- run linear regression on hourly data ----------
         for city in fiveCities:
             print(city)
             cityPd = import_pd(city)
             # input any Y variables in a list for linear regression function
             linearReg(cityPd, ['PRES', 'DEWP', 'TEMP'])
```

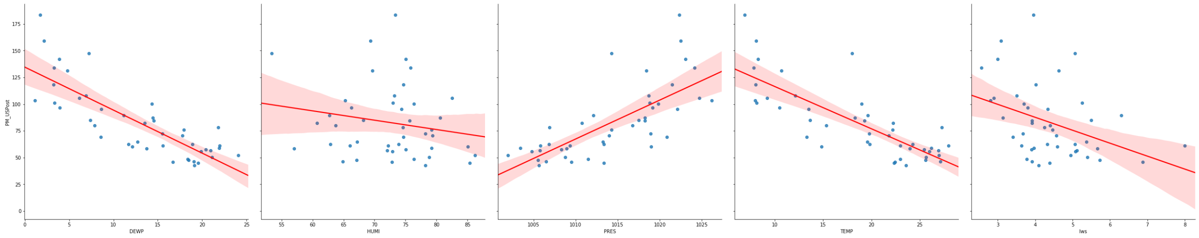
```
BeijingPM20100101_20151231.csv
Intercept:  1700.9116492834921 Coefficient:  [-1.51444797  4.06719769 -5.9079
0083] R^2:  0.18533689181744983
ChengduPM20100101_20151231.csv
Intercept:  644.6207462467463 Coefficient:  [-0.51734229 -0.02847296 -2.02424
703] R^2:  0.09254461843078032
GuangzhouPM20100101_20151231.csv
Intercept:  -340.99441090646616 Coefficient:  [ 0.39825251 -0.12466795 -0.292
2239 ] R^2:  0.033012451022515266
ShanghaiPM20100101_20151231.csv
Intercept:  453.868310492437 Coefficient:  [-0.37654771 -0.75587345 -0.555365
] R^2:  0.07554710064824277
ShenyangPM20100101_20151231.csv
Intercept:  -1123.2723365887662 Coefficient:  [ 1.18910649  1.29693002 -1.221
21695] R^2:  0.08382580052785726
```

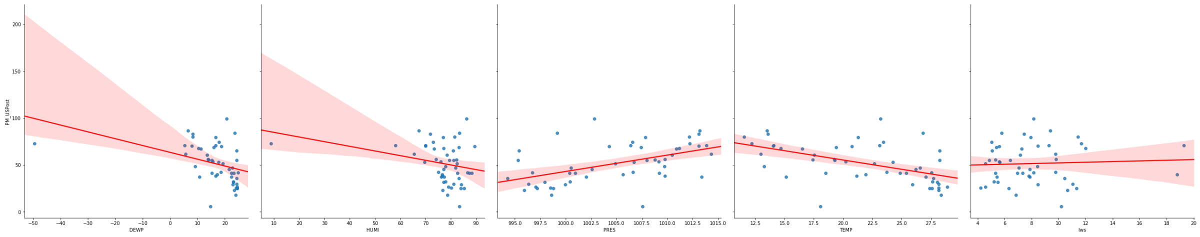
R^2 doesn't look pretty, going to the same process on the monthly averaged data

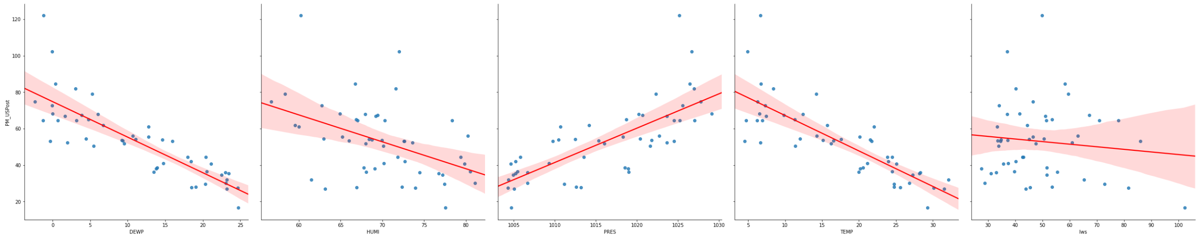
In [20]:

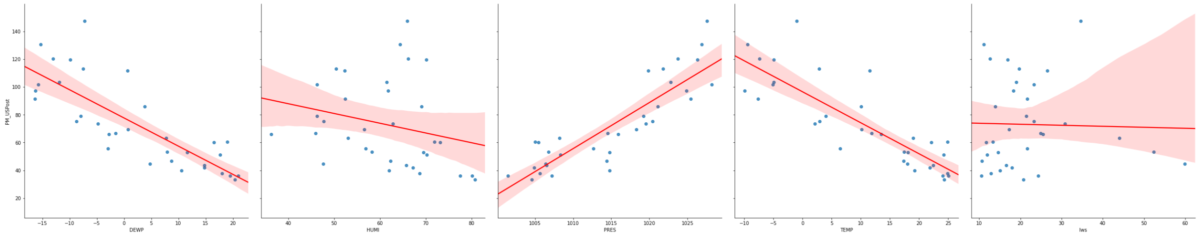
```
# ---------- Creating pairplots for the monthly averaged data ----------
for city in monthlyList:
    city_pd = import_pd(city)
    pair_plot(city_pd)
```











In [21]:

```
# ---------- running linear regression on monthly averaged data ----------
for city in monthlyList:
    print(city)
    cityPd = import_pd(city)
    # input any Y variables in a list for linear regression function
    linearReg(cityPd, ['lws', 'PRES', 'DEWP', 'TEMP'])
```

```
BeijingPM2010_2015Monthly.csv
Intercept:  293.1773771727821 Coefficient:  [-0.83344746 -0.12726569  2.29674
598 -4.1536888 ] R^2:  0.46406703151426515
ChengduPM2010_2015Monthly.csv
Intercept:  83.69522757429075 Coefficient:  [-11.08468789   0.06782768  -1.86
019332   0.05349134] R^2:  0.46026562294869444
GuangzhouPM2010_2015Monthly.csv
Intercept:  -733.0167359812949 Coefficient:  [-0.17160524  0.79468335 -0.1759
8122 -0.4309643 ] R^2:  0.2508643984713521
ShanghaiPM2010_2015Monthly.csv
Intercept:  -24.55303982468788 Coefficient:  [ 0.02113451  0.09122226 -0.8740
0067 -0.27062571] R^2:  0.3940754013221841
ShenyangPM2010_2015Monthly.csv
Intercept:  -2371.5526452662243 Coefficient:  [ 0.23654514  2.40463847  0.638
94104 -0.09723922] R^2:  0.3694245244532483
```

Note to Self, Plot data by year, and also plot the differnce between weekdays and weekend 'scatter_kws': {'alpha': 0.1}}

Conlusion

1. Northern Chinese cities (Beijing and Shenyang) are **more contaminated** with air pollution than Southern Chinese cities (Guangzhou, Shanghai, Chengdu) are. Therefore, it can be inferred that Northern cities/part of China are more industrialized than Southern cities. Transportation could be more heavily used, too.
2. Through plotting the data, Northern cities of China revealed to be dry, windy, and cold, showcasing characteristics of inland climates. Southern cities, on the other hand, are the oppostie. Southern cities are hot, and humid, showcasing tropical climate characteristics.
    - As expected, Northern cities are lower in temperature(C), hence colder, than Southern cities in China.
    - Air pressure(hPa) in Guangzhou is lower than all other cities.
    - Southern cities are also expected to hold a higher dew point(C), with a more humid climate than Northern cities.
    - Northern cities are far more windy than Southern cities with the exception of Chengdu.
3. Seasonal Variation: PM 2.5 concentration is higher in winter than it is in summer. The pattern shows that concentration level pf PM 2.5 changes along with temperature patterns.
4. The combination of **Dew Point (DEWP)**, **Pressure (PRES)**, **wind speed (lws)**, and **temperature (TEMP)** was the most successful meteorological indicator of PM 2.5 in linear regression.

New Skills Learned in Final Project

- Learned and wrote first machine learning codes for linear regression
    - Learned how to replace missing values in dataset when performing linear regression
- Used seaborn to plot pairplots
- Created boxplot using "by" to make separate boxes for each city, and used "groupby" to create sets of plots for each year
- Used for loop to plot data

Improvements for the future

- Use python to perform more in depth data wrangling
- Perform Stepwise regression machine learning
- Learn using .format to format texts

Link to Dataset: https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities (https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities)

The view of Beijing on a clear day (left) versus it on a day of bad air quality (right).

Real Time Air Quality Index: http://aqicn.org/city/beijing/ (http://aqicn.org/city/beijing/)

```
In [ ]:
```