江西财经大学
JIANGXI UNIVERSITY OF FINANCE AND ECONOMICS

课程名称：Python语言与数据分析

课 程 报 告

项目名称 2010-2018年18地级市PM2.5浓度数据分析

班　　级 金融202班

学　　号 0204841

姓　　名 杜翻

任课教师 肖　泉

开课学期： 20 至 21 学年 第 二 学期

完成时间： 20 21 年 7 月 03 日

《2010-2018年地级市PM2.5 》数据分析报告

目 录

一、 概述

1. 背景：

随着科技的发展，工业的进步和全球人口急剧增多的因素的影响，人们赖以生存的环境遭到了很大的破坏，很多地区相继出现了酸雨、物种灭绝、土地沙化等环境问题。环境问题已经成为当今世界各国普遍关注的问题之一，也是 21 世纪人类面临的重大挑战。我国是一个人口大国，城市众多，人口密集。但由于工业的发展，我们的很多城市都受到了不同程度的污染，尤其是空气的污染，直接对我们造成伤害，人们疾病的发生率也逐年提高。PM2.5 就是当今对于空气质量情况的一种监测数据。因此根据 PM2.5 数值分析我国当前空气质量极其重要。

2.意义：

（1）.分析各地市不同年份空气质量等级，有助于通过真实数据了解我国近几年来空气质量的变化情况，让人们了解到空气质量防护措施的重要性以及近几年来我国在空气防治方面取得的显著成效。
（2）通过分析监测 PM2.5,能够有效解决 PM2.5 带来的问题,并且调查产生这样颗粒的粉尘的来源，明确空气质量不好的地区有哪些，进而进一步了解造成空气质量问题的原因，制定出有效的解决方案措施.
（3）.通过 PM2.5 的监测,给大家一个现实的数据,让大家知道我们的出行要是选择开车会给环境带来多大的危害，提高人们环境保护的意识。

二、 数据描述

1. 数据来源

数据来源于数据科学竞赛平台“和鲸社区”
网址为：http://fizz.phys.dal.ca/~atmos/martin/?page_id=140

2. 特点

总共 18 条数据，11 字段

3. 字段

City_id　城市编号
City_name 城市名称
City_code　城市编码
Year　　年份（2011-2018）
PM2.5　　空气污染指数
Mean　　各地市 PM2.5 均值
Grade　　空气质量等级

4. 缺失值判断

初步判断数据是否有缺失值，方法如下：
import pandas as pd

data=pd.read_csv('2000-2018 年 286 地级市 PM2.5 浓度数据.csv',encoding='cp936')
print(data.info())

5.数据类型

在获取数据时要注意数据的类型，特别是日期字段的数据，整理数据时可以将其转换成时间格式，以方便后续的数据处理。

6.数据整理

由于一般分析 PM2,5 数据是根据 PM2.5 的平均值将其划分为不同的空气质量等级，所以此数据也需要整理，将其转换为不同的空气质量等级。

三、 数据分析内容

1.将各地级市 2011 年与 2018 年 PM2.5 数值转换为具体的空气的质量等级
2.比较分析 2011 年与 2018 年各地级城市空气质量等级占比变化
3..特点城市 2011-2018 年 PM2.5 变化分析
4.2016-2018PM2.5 排名前六的城市分析

四、 数据分析图表

1. 数据类型查看

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18 entries, 0 to 17
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   city_id     18 non-null     int64
 1   city_name   18 non-null     object
 2   city_code   18 non-null     int64
 3   2011        18 non-null     float64
 4   2012        18 non-null     float64
 5   2013        18 non-null     float64
 6   2014        18 non-null     float64
 7   2015        18 non-null     float64
 8   2016        18 non-null     float64
 9   2017        18 non-null     float64
 10  2018        18 non-null     float64
dtypes: float64(8), int64(2), object(1)
memory usage: 1.7+ KB
None
```

2.数据缺失值的判断

```
各字段是否含有空值情况：
 city_id          False
city_name        False
city_code        False
2011             False
2012             False
2013             False
2014             False
2015             False
2016             False
2017             False
2018             False
dtype: bool
```

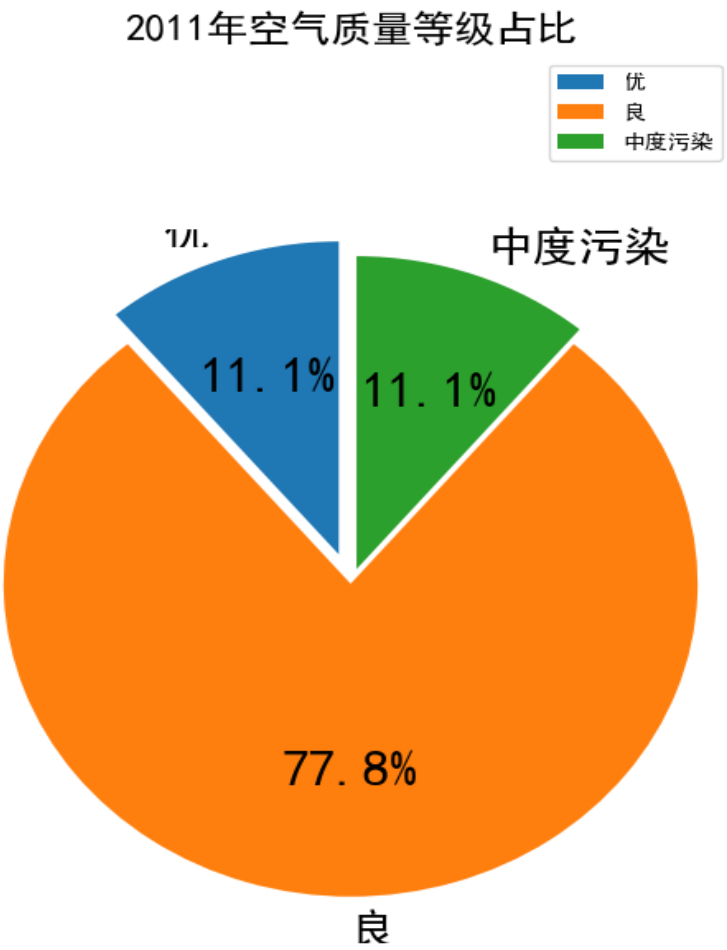
2. 数值数据的统计

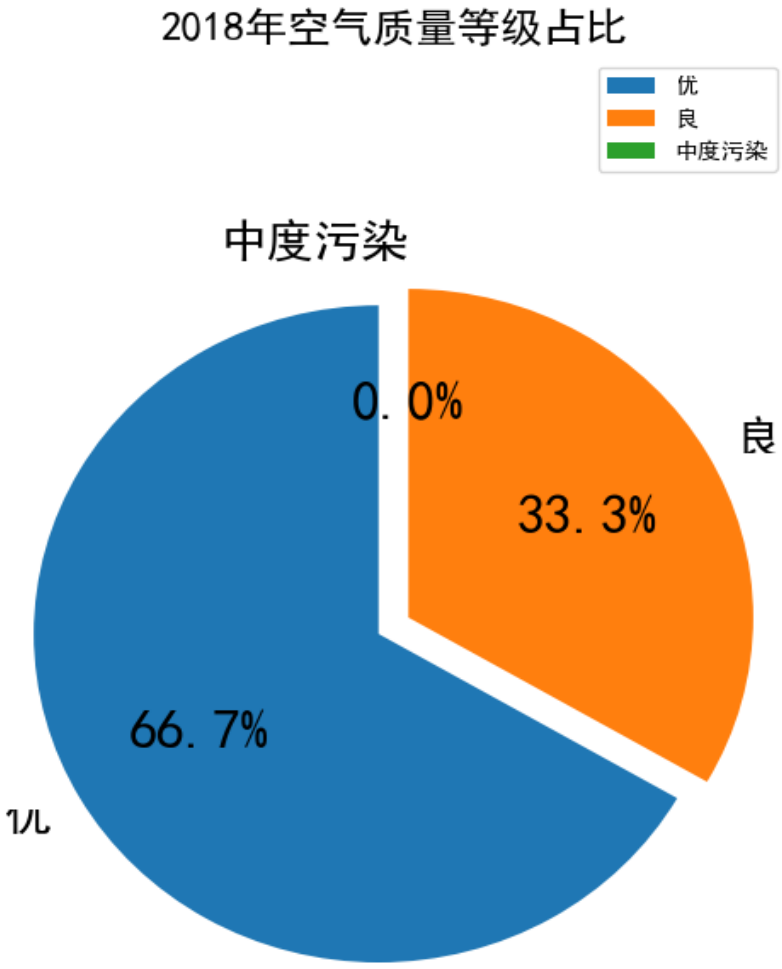
```
2018          False
dtype: bool
          city_id    city_code         2011  ...         2016         2017
2018
count   18.000000    18.000000   18.000000  ...   18.000000   18.000000
18.000000
mean     9.500000  4064.500000   51.020664  ...   38.743736   36.344472
32.509118
std      5.338539  1704.258619   13.470335  ...    8.805414    7.731854
7.041217
min      1.000000  1101.000000   30.014072  ...   22.413896   23.401002
20.094738
25%      5.250000  3301.500000   44.793506  ...   33.492491   32.227551
29.186018
50%      9.500000  3606.500000   47.122432  ...   37.054744   35.588664
30.516883
75%     13.750000  5851.000000   59.216364  ...   43.037579   40.656046
36.231607
max     18.000000  6501.000000   80.086898  ...   62.553569   54.741139
48.279971

[8 rows x 10 columns]

In [21]:
```

3.各地级市 2011 年与 2018 年空气质量等级分析

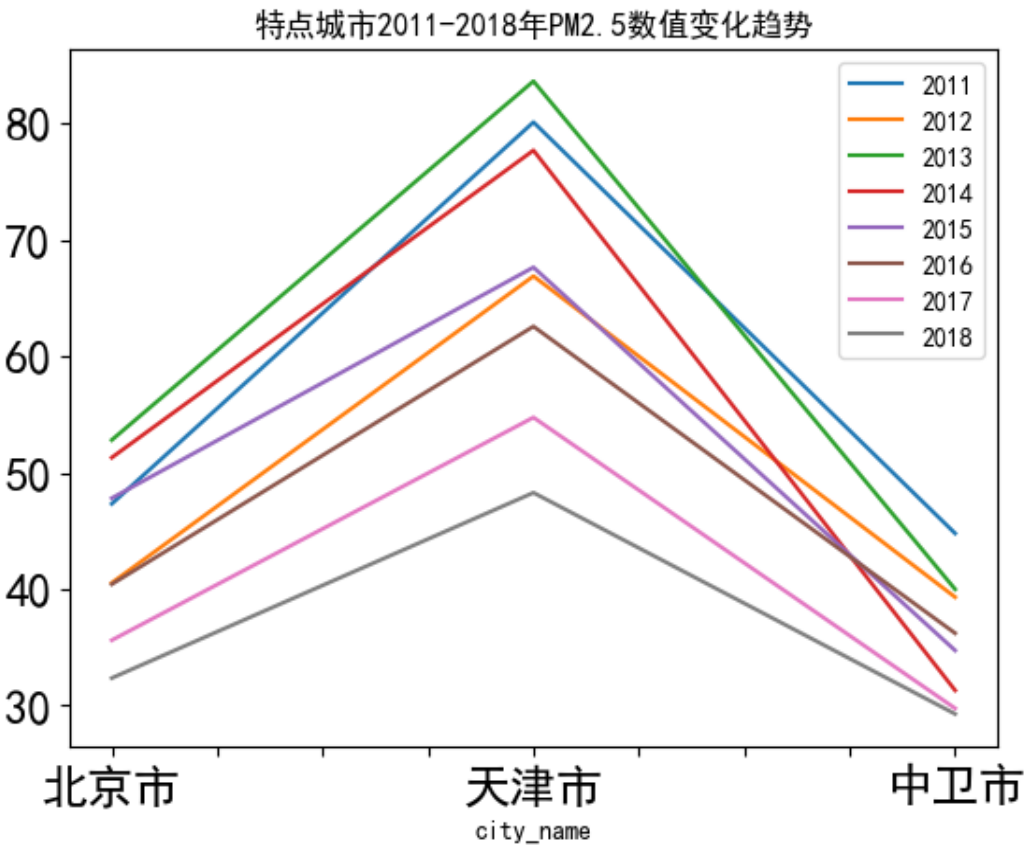
```
In [16]: data
Out[16]:
    city_id city_name  city_code  ...       mean  2011Grade  2018Grade
0         1       北京市       1101  ...  43.522004          良          优
1         2       天津市       1201  ...  67.679887       轻度污染          良
2         3       上海市       3101  ...  44.680991          良          良
3         4       南京市       3201  ...  55.773665          良          良
4         5       杭州市       3301  ...  39.989154          良          优
5         6       温州市       3303  ...  26.468995          优          优
6         7       嘉兴市       3304  ...  51.350027          良          良
7         8       绍兴市       3306  ...  38.787152          良          优
8         9       鹰潭市       3606  ...  38.149065          良          优
9        10       赣州市       3607  ...  30.630475          优          优
10       11       吉安市       3608  ...  38.974117          良          优
11       12       宜春市       3609  ...  42.801061          良          优
12       13       成都市       5101  ...  56.892086       轻度污染          良
13       14       西安市       6101  ...  48.894018          良          良
14       15       银川市       6401  ...  37.903639          良          优
15       16       固原市       6404  ...  34.157683          良          优
16       17       中卫市       6405  ...  35.675000          良          优
17       18      乌鲁木齐市       6501  ...  36.092945          良          优

[18 rows x 14 columns]

In [17]:
```

4.2011 年与 2018 年各地级城市空气质量等级占比变化分析

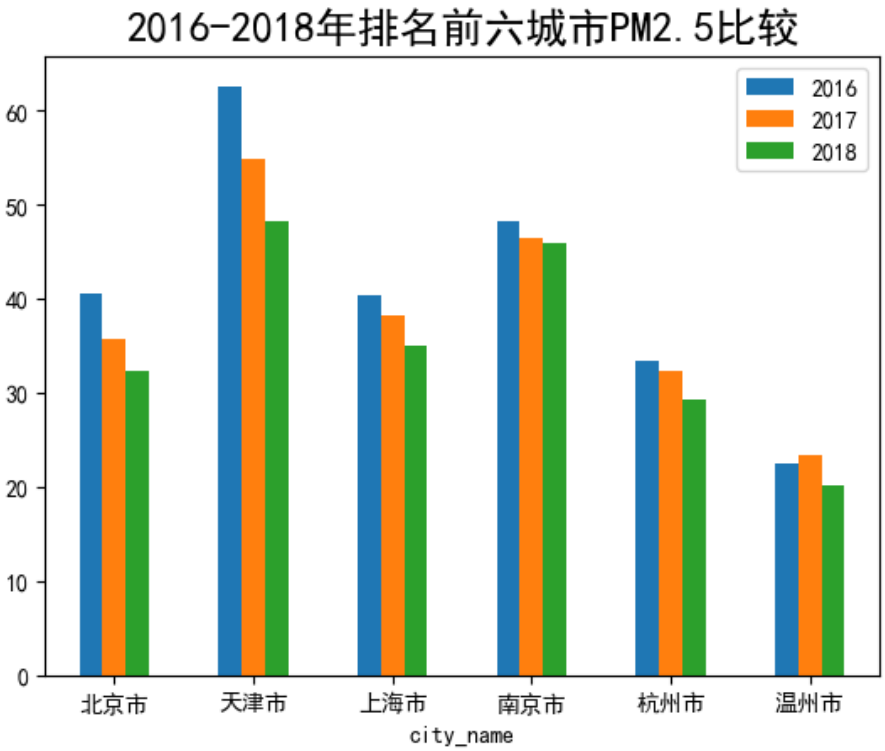

2011年空气质量等级占比
优
良
中度污染
11.1%
11.1%
77.8%
良



5.特点城市 2011-2018 年 PM2.5 变化分析



6.2016-2018PM2.5 排名前六的城市分析



五、数据分析结果

1..各地级市 2011 年与 2018 年空气质量等级分析结果

PM2.5 总体来说是一个比较严格的空气质量数据,监测 PM2.5 使我们能更清楚地认识现在我们空气质量的差距.但其数值不能具体的体现出各地市的空气质量情况。根据我国现在使用的空气质量等级标准 0~35μg/m3 为优；35~75μg/m3 为良；75~115μg/m3 为轻度污染；115~150μg/m3 为中度污染；150~250μg/m3 为重度污染；严重污染大于 250μg/m3 及以上。也就是说随着 PM2.5 数值的增大，空气质量污染越严重。通过数据转化，我们可以清楚的看出 2011 与 2018 的空气质量等级，而且可以看出 2011 年各地级市，空气质量大多数为良，“优”很少，甚至有轻度污染。而到了 2018 年大多数城市空气质量为优，空气质量方面有了很大的提升。

2.2011 年与 2018 年各地级城市空气质量等级占比变化分析结果

（1）.通过分析 2011 与 2018 年空气质量等级占比情况，可以得出以下结论:

2011 年各地级城市空气质量等级偏中等水平，其中 77.8%的城市为良，只有 11%的城市为优，还有 11.1%的城市为中度污染。

（2）.2018 年各地级城市空气质量总体上升偏好。其中有 66.7%的城市空气质量为优，其余的 33.3%为良。

（3）.从图表和数据中我们可以得出，前几年我国的空气质量整体偏低，不是很理想，而究其原因，2011 年我国注重经济效率发展，尤其是一些新兴工业企业的崛起，造成了工业污水大量的排放以及树木土地的破坏，使得当时我国环境质量低下。而自 2012 年之后我国提出了蓝天保卫战的重大措施，提倡绿水青山就是金山银山，我国也开始在环境保护方面有了重大进步，空气质量逐年上升，足见污染防治的重要性。

3..特点城市 2011-2018 年 PM2.5 变化分析结果

（1）.北京处于经济发达地区，工业污染严重，以前经常有大雾霾，空气质量指数极差，但是近些年来由于防治力度加大，空气质量大幅提升；而天津近些年来经济发展主要源于工业，大量工厂的存在必然导致严重的污染；中卫市位于宁夏回族自治区，工业发展较弱，因此空气质量一直以来都比较好，所以以这三个特点城市来分析我国 2011-2018 年以来空气质量的变化情况。

（2）通过图表基本可以看出，我国 PM2.5 逐年下降，因此空气质量逐年提高。但是天津市 PM2.5 数值整体偏高，而北京的 PM2.5 较低，中卫市的最低，空气质量最好。由数据变化可知，可以对于一些工业比较发达的地区加大防治力度，着重有效落实举措。而对于一些经济比较落后的地区，在开发经济的同时保持原

有的良好环境。

6.2016-2018PM2.5 排名前六的城市分析结果

根据图表可以看出，北京、天津、上海、杭州、南京、温州的 PM2.5 排名靠前，空气质量较其他地级城市较差，而且天津最高。并且但是从 2016-2018 年空气质量逐年上升。可以的出经济发达的地区对于环境和空气的破坏较大，需要重视其污染防治，适应新时代脚步，将经济高效率发展转变高质量发展并且贯彻新发展理念。

六．总结

1.从分析的数据可知，我国空气质量偏差的城市主要是一些经济发达的地区，由于注重经济发展，而忽略了空气防治和环境的保护，造成了 PM2.5 过高，影响了人们正常生活的环境。所以，对于一些工业比较集中的地级城市，需要加大污染防治。例如:工业和城区规划要合理，不要排放量过于集中。不要在一个地方排放二次污染，不然会发生严重污染现象；多植树造林。利用树来吸收更多的污染物，减少空气污染的程度；改变燃料构成。如城市工业和民用煤气、液化石油气的发展，低硫燃料和新能源（太阳能、风能、地热等）的采用。要推行采煤，以除去煤中大部分硫（主要是硫铁矿硫）；减少汽车废气排放。主要是改时发动机的燃烧设计和提高油的燃烧质量，加强交通管理等。当今时代是新时代建设社会主义现代化强国阶段，不仅要注重经济高质量发展，而且也要大力保护环境，贯彻新发展理念，走绿色可持续发展道路。

2.此次数据分析的过程中，从数据的查找到筛选都比较艰辛。首先我在可用的网站上查找并下载自己需要分析的空气质量数据，然后运用 Python 工具对于数据进行了初步的整理与筛选。接着我对于自己要分析内容进行具体分类，明确要使用的函数和运用什么样的图形可以加大数据的可视化，最后总结分析结果。

3.而对于此次 Python 分析，我也有许多心得体会。对于一项完整的数据分析报告，要明确分析的内容，运用合适和函数恰当的分析数据，并且过程中还需要用图表突出数据的可视化，使数据的结果更加直观可信。同时 Python 自身强大的优势有着其不可限量的发展前景。Python 被广泛的用在 Web 开发、运维自动化、测试自动化、数据挖掘等多个行业和领域。并且将 Python 作为主要开发语言的开发者数量逐年递增，这表明 Python 正在成为越来越多开发者的开发语言选择。所以我认为学好 python 对于我自身以后的发展和工作有着极大的帮助。

七、附录-数据分析代码

```
import pandas    as pd
import numpy as np
import tushare as ts
import matplotlib.pyplot as plt
pro = ts.pro_api()
data=pd.read_csv('2010-2018 年 18 地级市 PM2.5 浓度数据.csv',encoding='cp936')
```

```
print(data.info())#初步判断数据是否有缺失值
print('各字段是否含有空值情况：\n',data.isna().any())#了解每个字段中的数据是否含有缺失值
print(data.describe())#检查数据是否含有异常值，且仅对数值型数据进行统计
value1=data['2011']
def get_grade(value1):
    if value1 <=35 and value1 >0:
        return '优'
    elif value1 <= 75 and value1 >35:
        return '良'
    elif value1 <= 115 and value1 >75:
        return '轻度污染'
    elif value1 <=150 and value1 >115:
        return '中度污染'
    elif value1 <=250 and value1 >150:
        return '重度污染'
    elif value1 >250:
        return '严重污染'
    elif value1>500:
        return 'Beyond Index'#爆表了 else:return None#输入值无
value2=data['2018']
def get_grade(value2):
    if value2 <=35 and value2 >0:
        return '优'
    elif value2 <= 75 and value2 >35:
        return '良'
    elif value2 <= 115 and value2 >75:
        return '轻度污染'
    elif value2 <=150 and value2 >115:
        return '中度污染'
    elif value2 <=250 and value2 >150:
        return '重度污染'
    elif value2 >250:
        return '严重污染'
    elif value2>500:
        return 'Beyond Index'#爆表了 else:return None#输入值无
data.loc[:,'2011Grade']=data['2011'].apply(get_grade)
data.loc[:,'2018Grade']=data['2018'].apply(get_grade)
plt.rcParams['font.sans-serif']=['SimHei']#指定中文黑体字体
plt.rcParams['axes.unicode_minus']=False#确保-负号显示正常
rate1=[2,14,2]
labels=['优','良','中度污染']
plt.figure(figsize=(6,9)) #设置图形大小
explode =(0.1, 0, 0.05) # explode  设置各部分分割出来的间隙
```

```
patches, ltext, ptext =plt.pie(rate1, explode=explode, labels=labels, autopct='%.1f%%',shadow=False, startangle=90)
# autopct：百分比数字的显示格式，%.1f%%表示保留一位小数
#shadow:是否有阴影
#startangle：起始角度。默认从 0 度逆时针开始为第一块，此处选择从 90 度开始
(一月数据)for inltext:
plt.title('2011 年空气质量等级占比',fontsize=18)
for x in ltext:
    x.set_size(20) #  设置标注文字大小
for x in ptext:
    x.set_size(24) #设置 x/y 轴的单位长度相等
plt.axis('equal') #  设置百分比文字大小
plt.legend()


rate2=[12,6,0]
labels=['优','良','中度污染']
plt.figure(figsize=(6,9)) #设置图形大小
explode =(0.1, 0, 0.05) # explode  设置各部分分割出来的间隙
patches, ltext, ptext =plt.pie(rate2, explode=explode, labels=labels, autopct='%.1f%%', shadow=False, startangle=90)
# autopct：百分比数字的显示格式，%.1f%%表示保留一位小数
#shadow:是否有阴影
#startangle：起始角度。默认从 0 度逆时针开始为第一块，此处选择从 90 度开始
(一月数据)for inltext:
plt.title('2018 年空气质量等级占比',fontsize=18)
for x in ltext:
    x.set_size(20) #  设置标注文字大小
for x in ptext:
    x.set_size(24) #设置 x/y 轴的单位长度相等
plt.axis('equal') #  设置百分比文字大小
plt.legend()
data1=data.set_index('city_name')
data2=data1.loc[['北京市','天津市','中卫市']]
data3=data2.iloc[:,[2,3,4,5,6,7,8,9]]

data3.plot(kind='line',title=' 特 点 城 市  2011-2018  年  PM2.5  数 值 变 化 趋 势 ',fontsize=18)
data4=data1.head(6)
data5=data4[['2016','2017','2018']]
data5.plot(kind='bar',rot='0')
plt.title('2016-2018 年排名前六城市 PM2.5 比较',fontsize=18)
```