第15章

科学计算与分析

本章的主要内容

第1节 数据特征分析

第2节 数学建模

数据特征分析

数据特征分析是指用适当的统计分析方法对收集来的大量数据进行分析，为了提取其中有用信息并形成结论，而对数据进行详细研究和总结的过程，是为了寻求问题的答案而实施的有计划、有步骤的行为。

数据特征分析的分类：

- 描述性统计分析
- 验证性统计分析
- 探索性统计分析

一、统计量分析

- 集中（中心）趋势度量
  - 均值
  - 中位数：将所有数据排序，位于正中的那个数据
  - 众数：是变量中出现频率最大的值，通常用于对定性数据确定众数
- 离中趋势度量
  - 极差=最大值-最小值
  - 标准差：数据偏离均值的程度

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

  - 四分位数

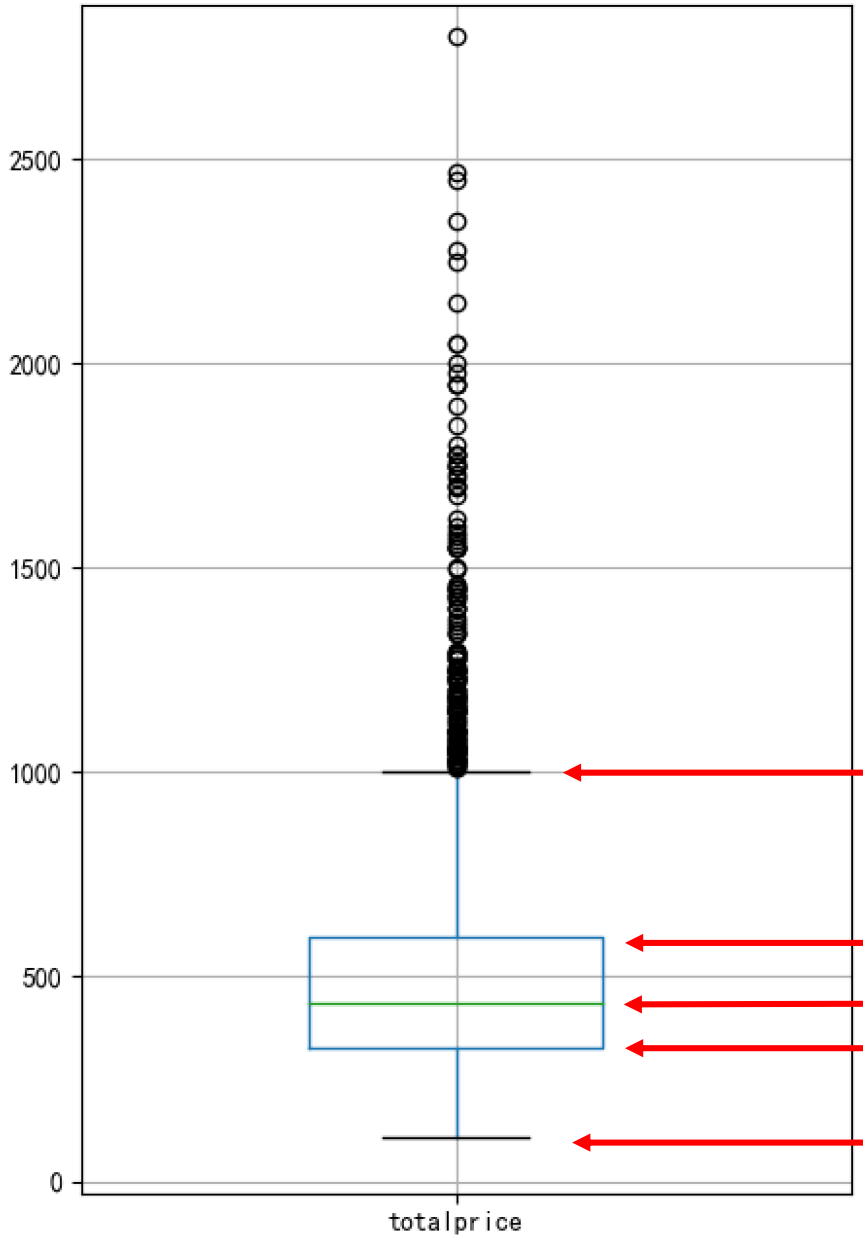
一、统计量分析

	Unnamed: 0	area	year	totalprice	unitprice
count	3000.000000	3000.000000	2988.000000	3000.000000	3000.000000
mean	1499.500000	84.210810	2000.196787	506.618500	6.137601
std	866.169729	32.958665	9.339562	282.059772	2.430205
min	0.000000	19.140000	1950.000000	105.000000	1.583400
25%	749.750000	60.637500	1994.000000	328.000000	4.316000
50%	1499.500000	78.410000	2002.000000	435.000000	5.574750
75%	2249.250000	96.805000	2007.000000	598.000000	7.379250
max	2999.000000	363.350000	2016.000000	2800.000000	16.857600

数据特征分析

四分位数：

箱型图



异常值

上限=(Q3+IQR*1.5)和最大值中的较小者

内限

Q3:上四分位数

Q2: 中位数

Q1:下四分位数

四分位距IQR=Q3-Q1

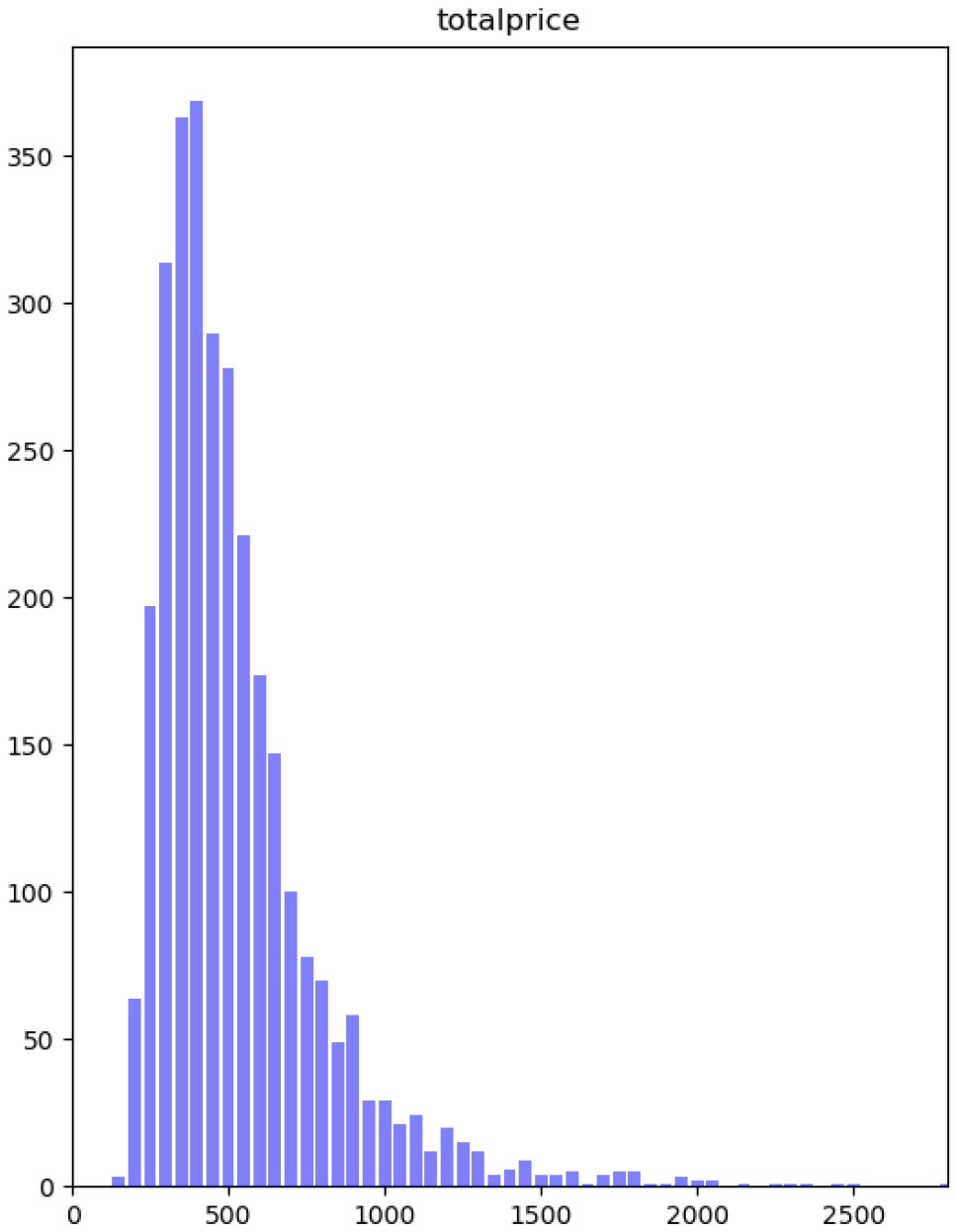
下限=(Q1-IQR*1.5)和最小值中的较大者

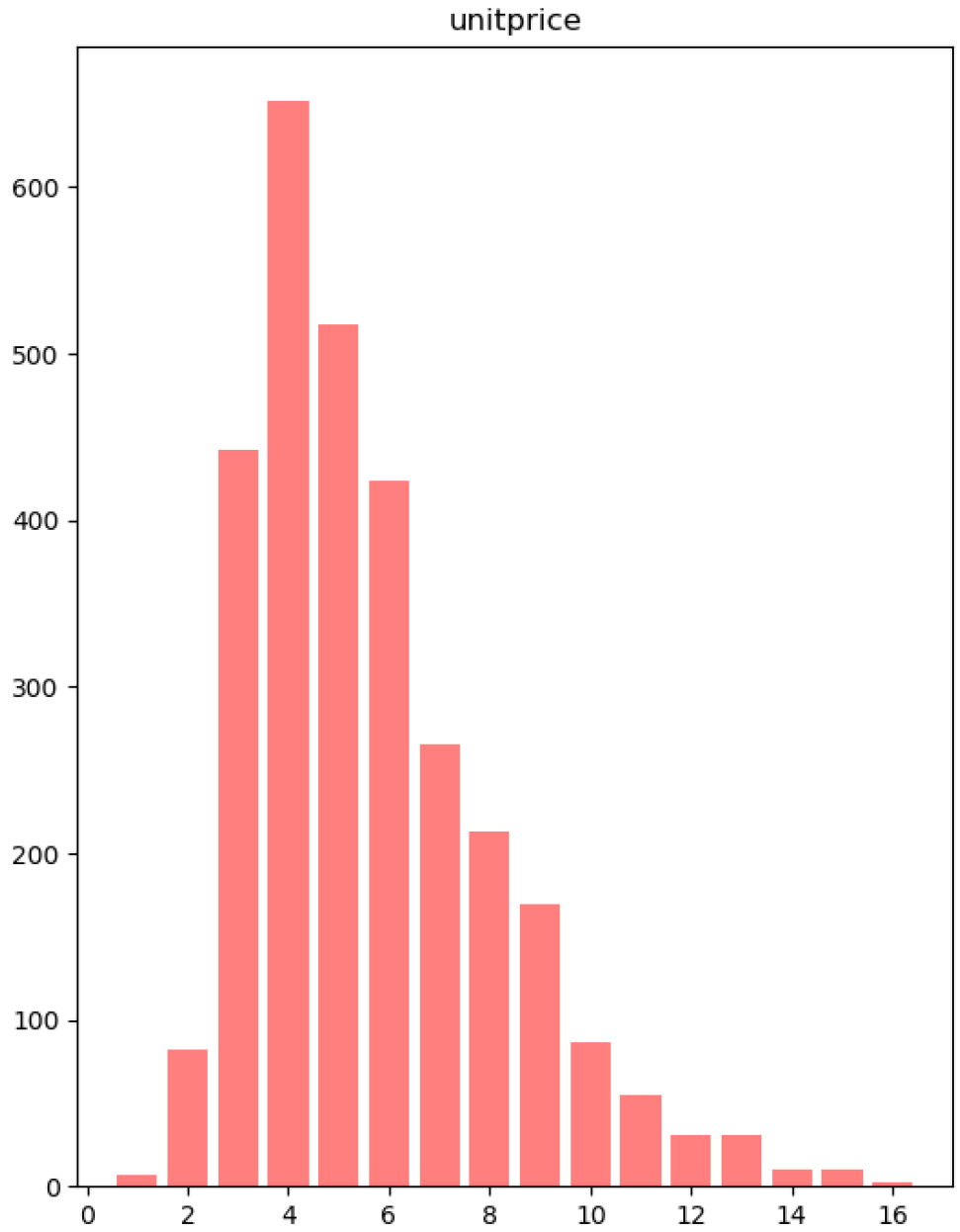
二、分布分析

- 是研究数据的分布特征和分布类型;
- 对于定量数据，可以做出频率分布表、绘制频率分布直方图显示分布特征;
- 对于定性数据，可用饼图和条形图等显示分布情况。

数据特征分析

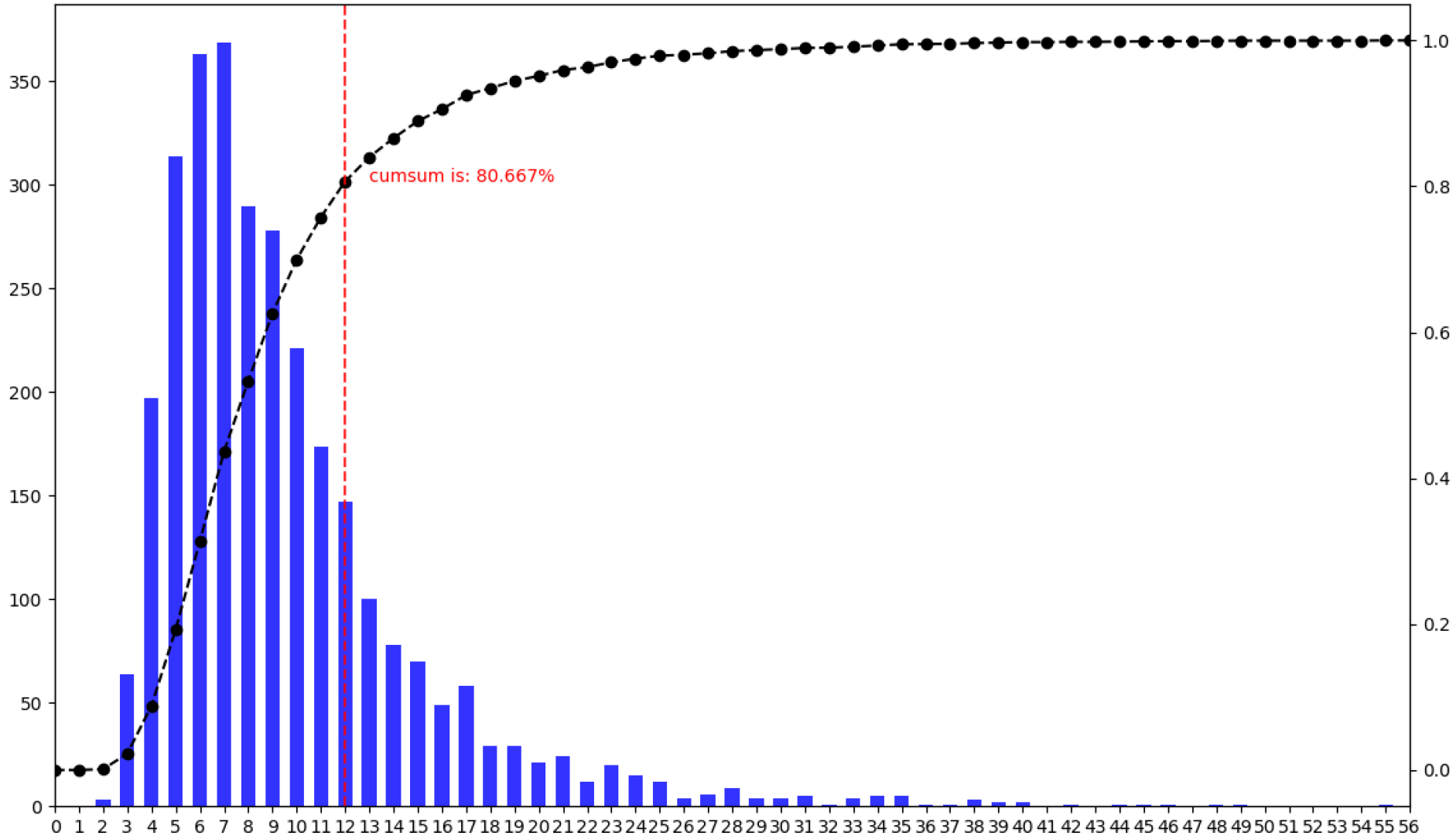
二、分布分析





数据特征分析

三、帕累托分析（二八定律）



数据预处理

作业1:

1.在链家官网上，找一个你感兴趣的城市，将其二手房数据爬取下来（取前100页3000个数据，如果不够3000个则取全部），并进行清洗整理（去掉不必要的空格，算出单价，房屋建成年份用整数表示等）；

2.查看总价和均价的统计量分析;

3.完成总价和单价的分布分析：要求画出直方图;

4.进行帕累托分析：看看是否符合二八定律。

四、对比分析

是指把两个相互联系的指标进行比较，从数量上展示和说明研究对象规模的大小、水平的高低、速度的快慢，以及各种关系是否协调等。特别适用于指标间的横纵向比较、时间序列的比较分析。

- 绝对数比较
- 相对数比较

数据特征分析

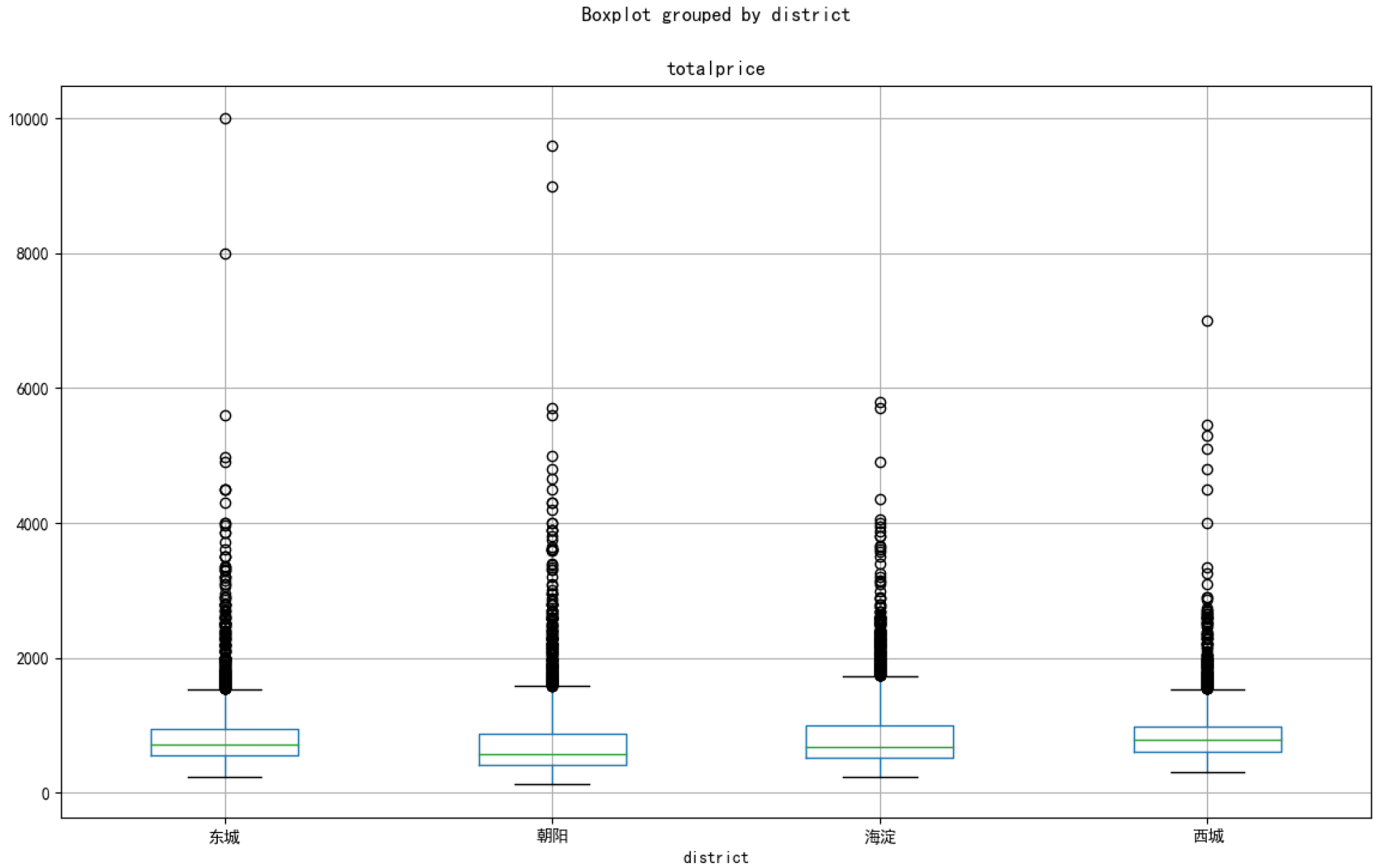

Bronx
Manhattan
Queens
New York City
Brooklyn
Staten
Island
Lower
Bay
米宅海外

四、对比分析

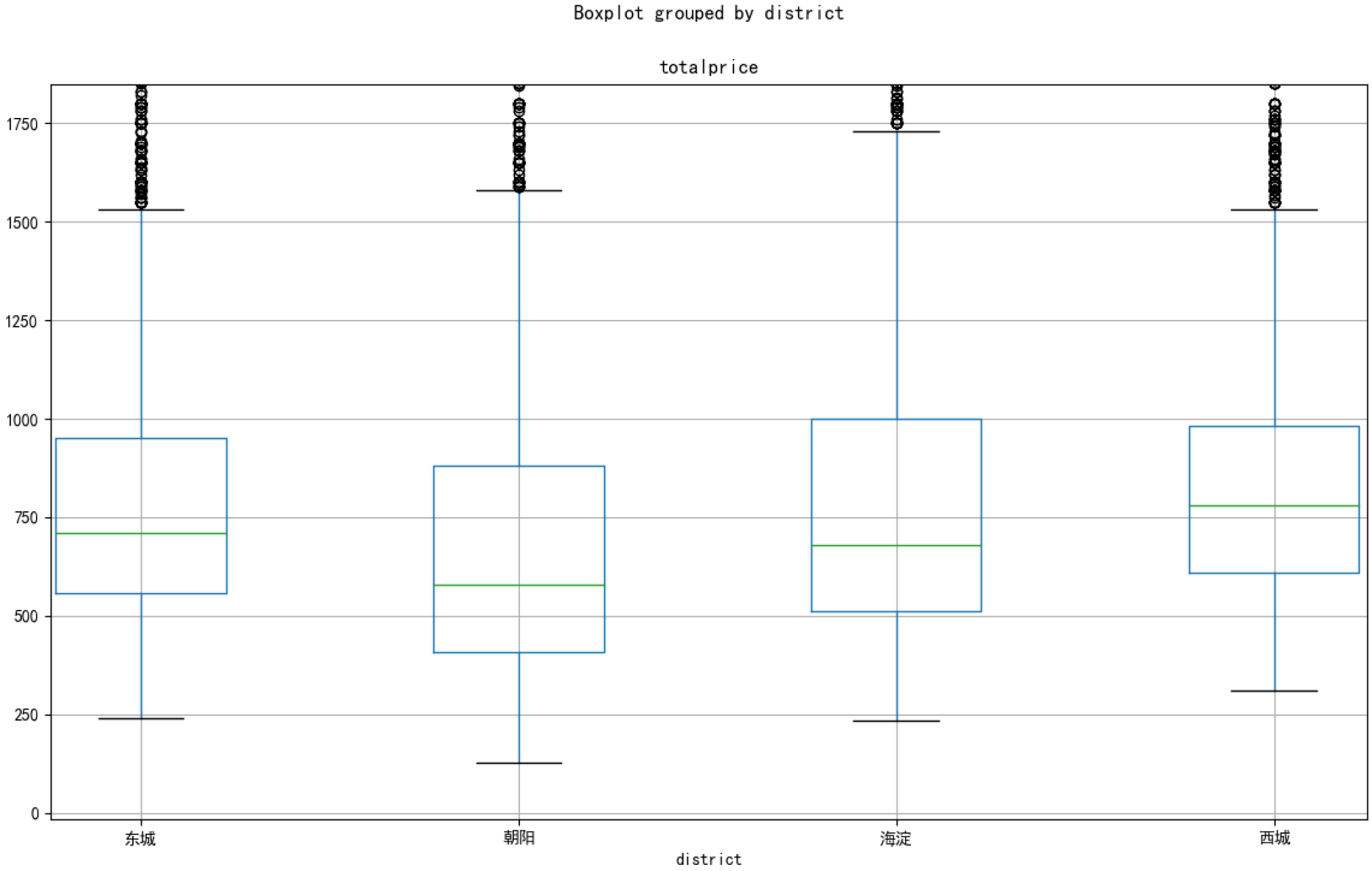
- 相对数比较
  - 结构相对数
  - 比例相对数
  - 比较相对数
  - 强度相对数
  - 计划完成程序相对数
  - 动态相对数

数据特征分析

四、对比分析（箱型图）



数据特征分析

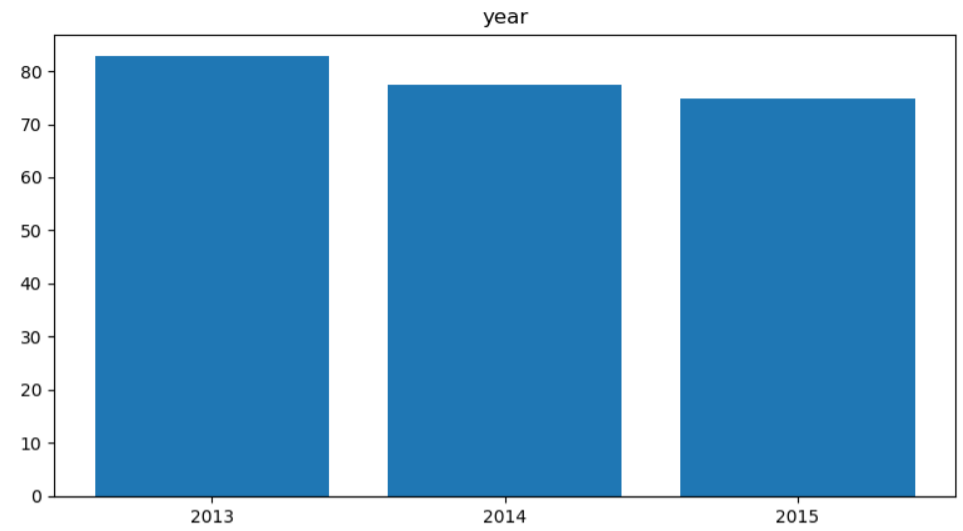

Boxplot grouped by district
totalprice
1750
1500
1250
1000
750
500
250
0
东城
朝阳
海淀
西城
district

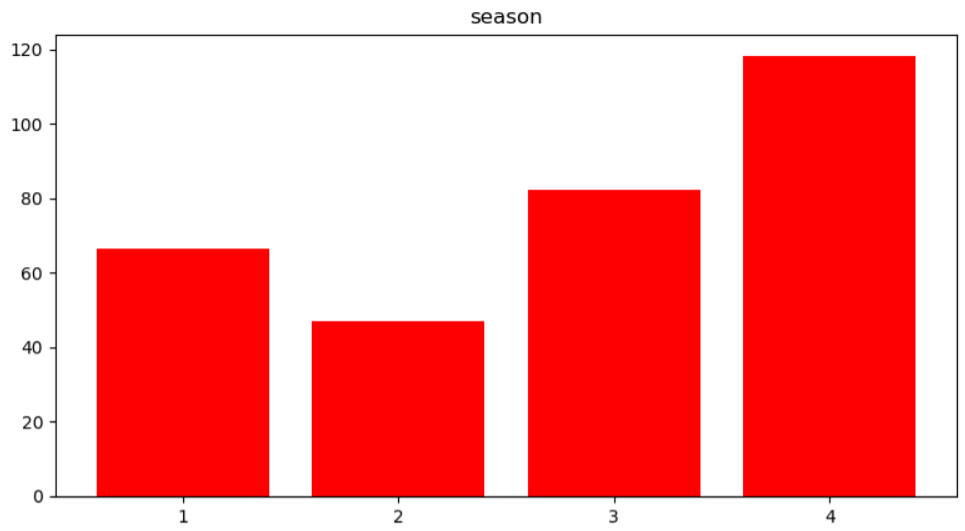
五、周期分析

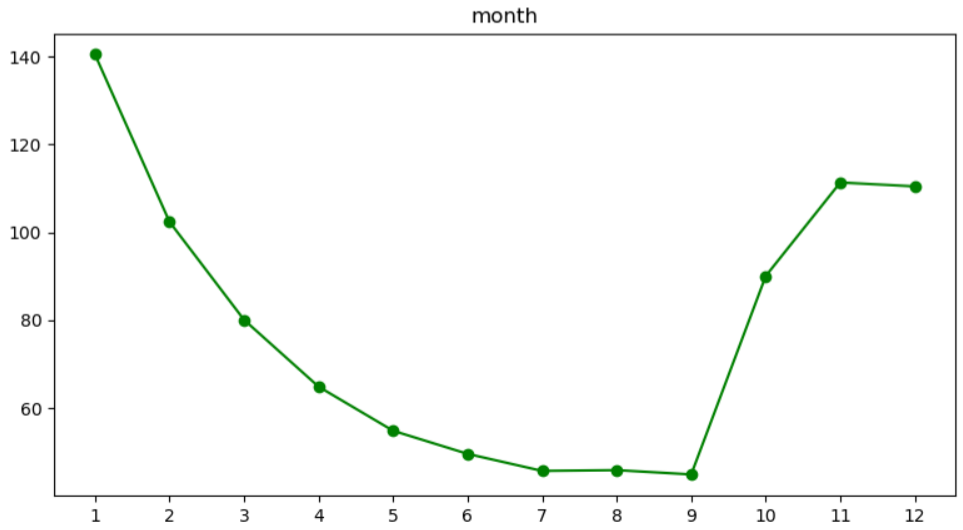
时间序列分析

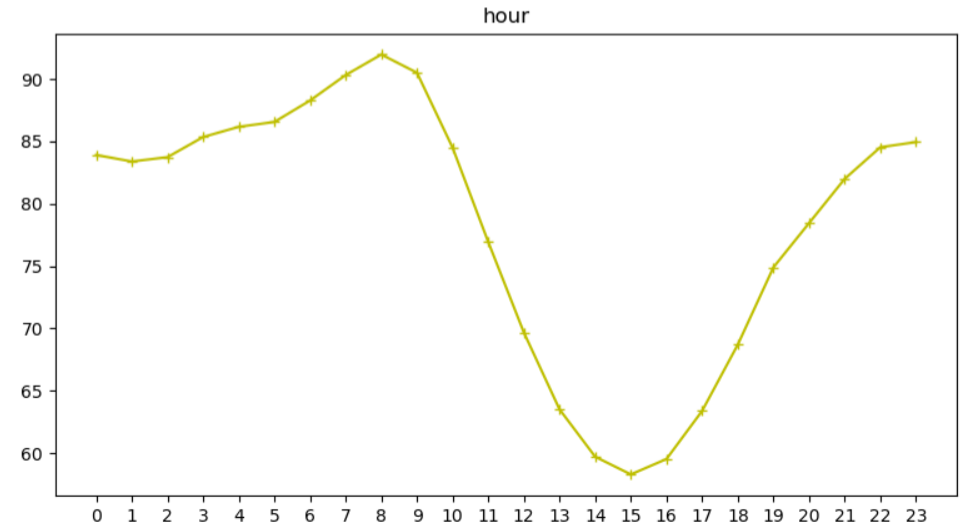
- 每日、每周、每月、每年
- 注意节假日

数据特征分析

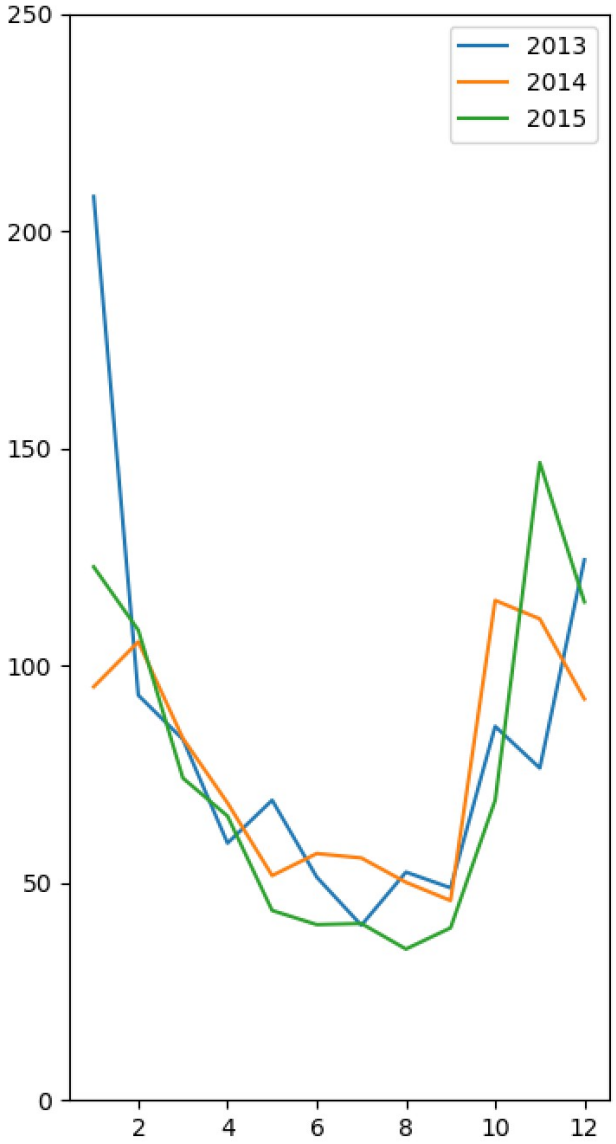

year
80
70
60
50
40
30
20
10
0
2013
2014
2015

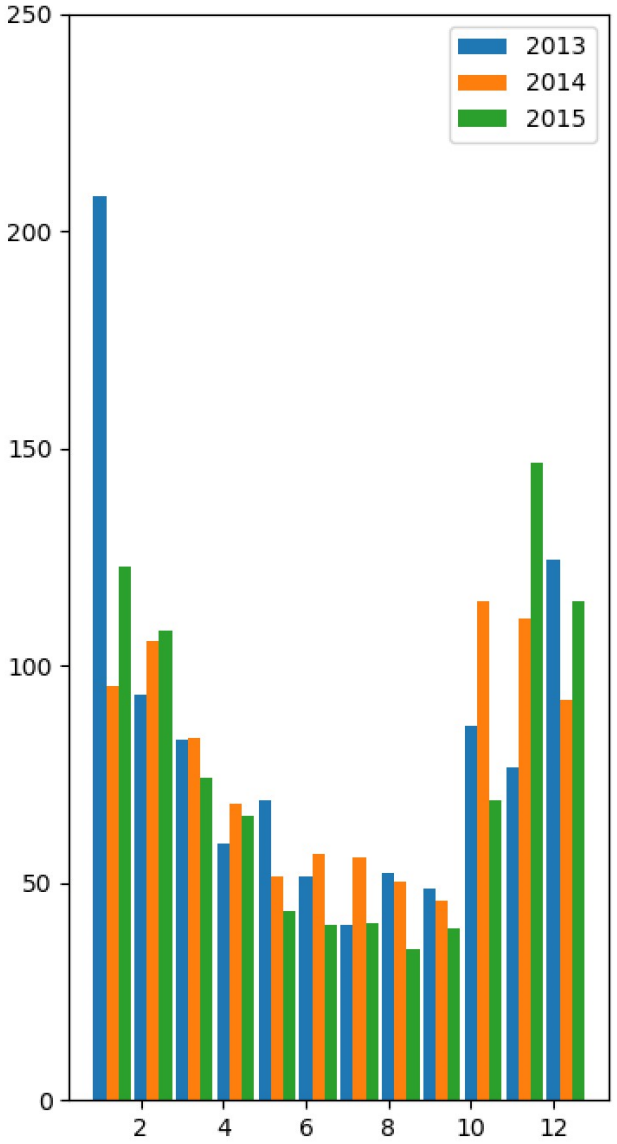

season
120
100
80
60
40
20
0
1
2
3
4

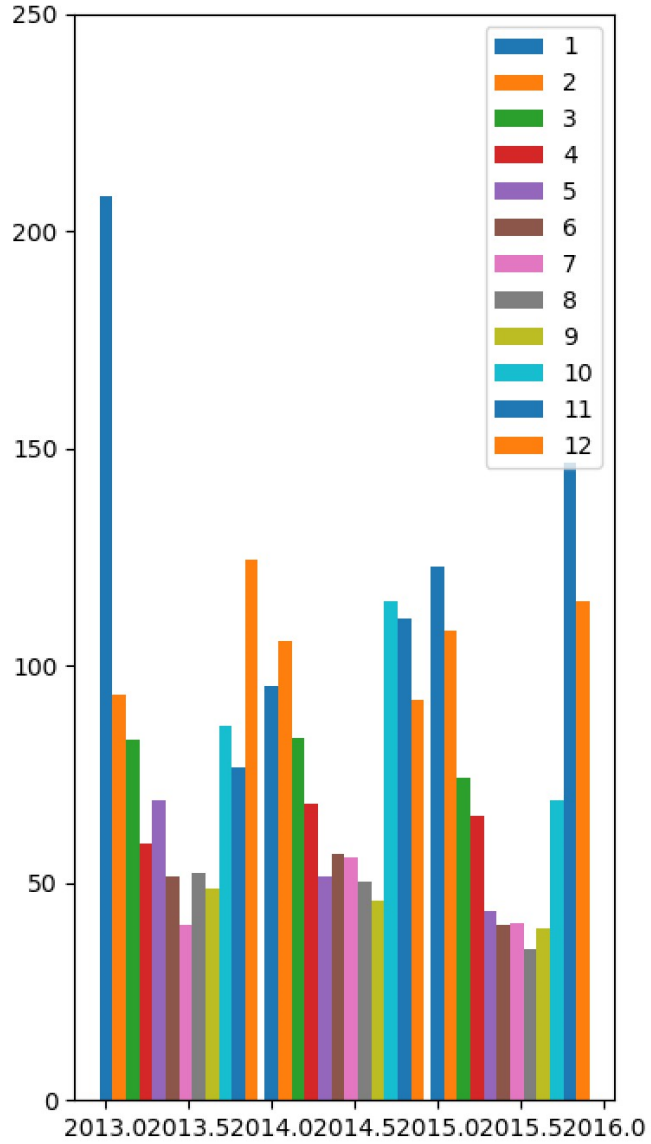

month
140
120
100
80
60
1
2
3
4
5
6
7
8
9
10
11
12


hour
90
85
80
75
70
65
60
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

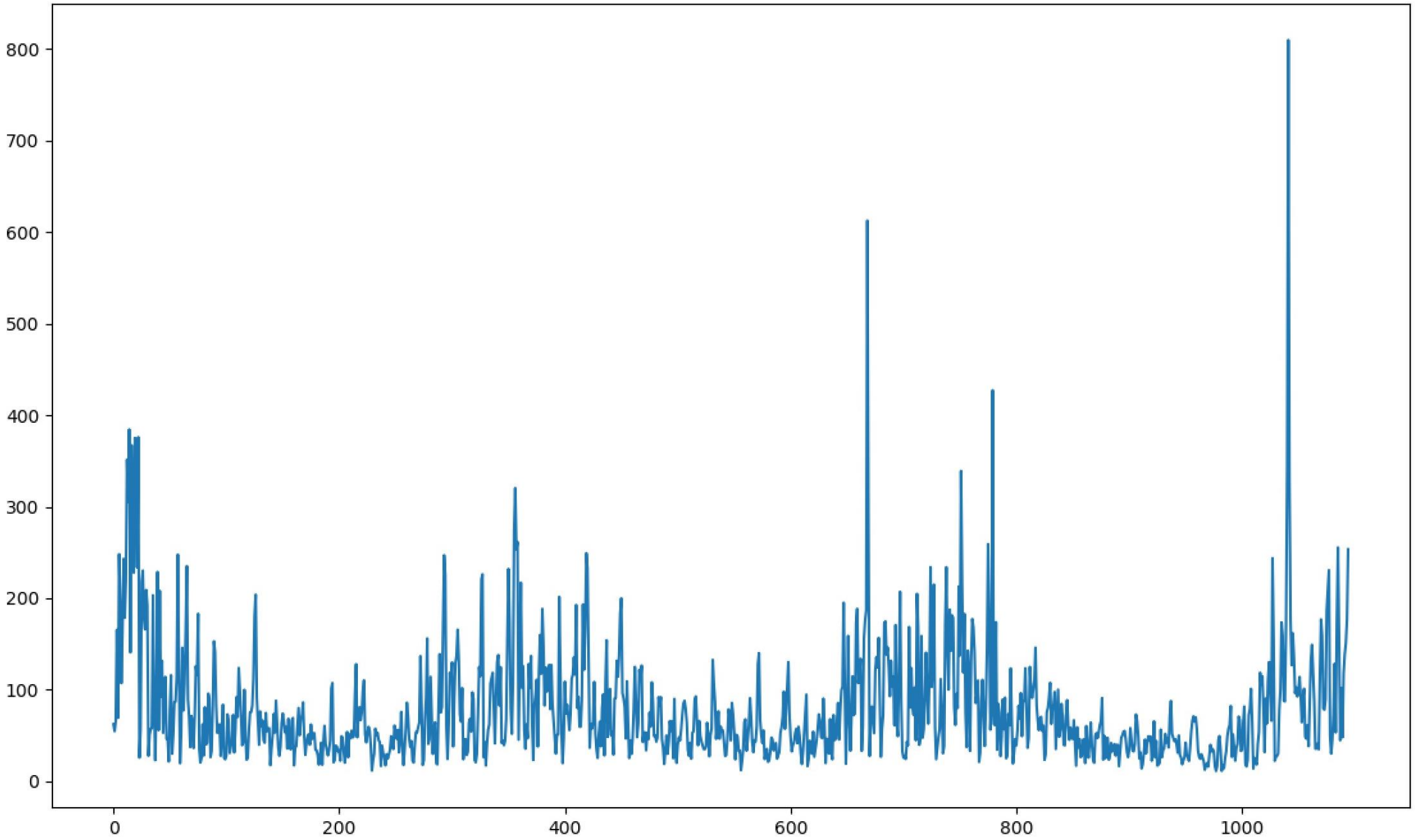
数据特征分析


2013
2014
2015
250
200
150
100
50
0
2
4
6
8
10
12

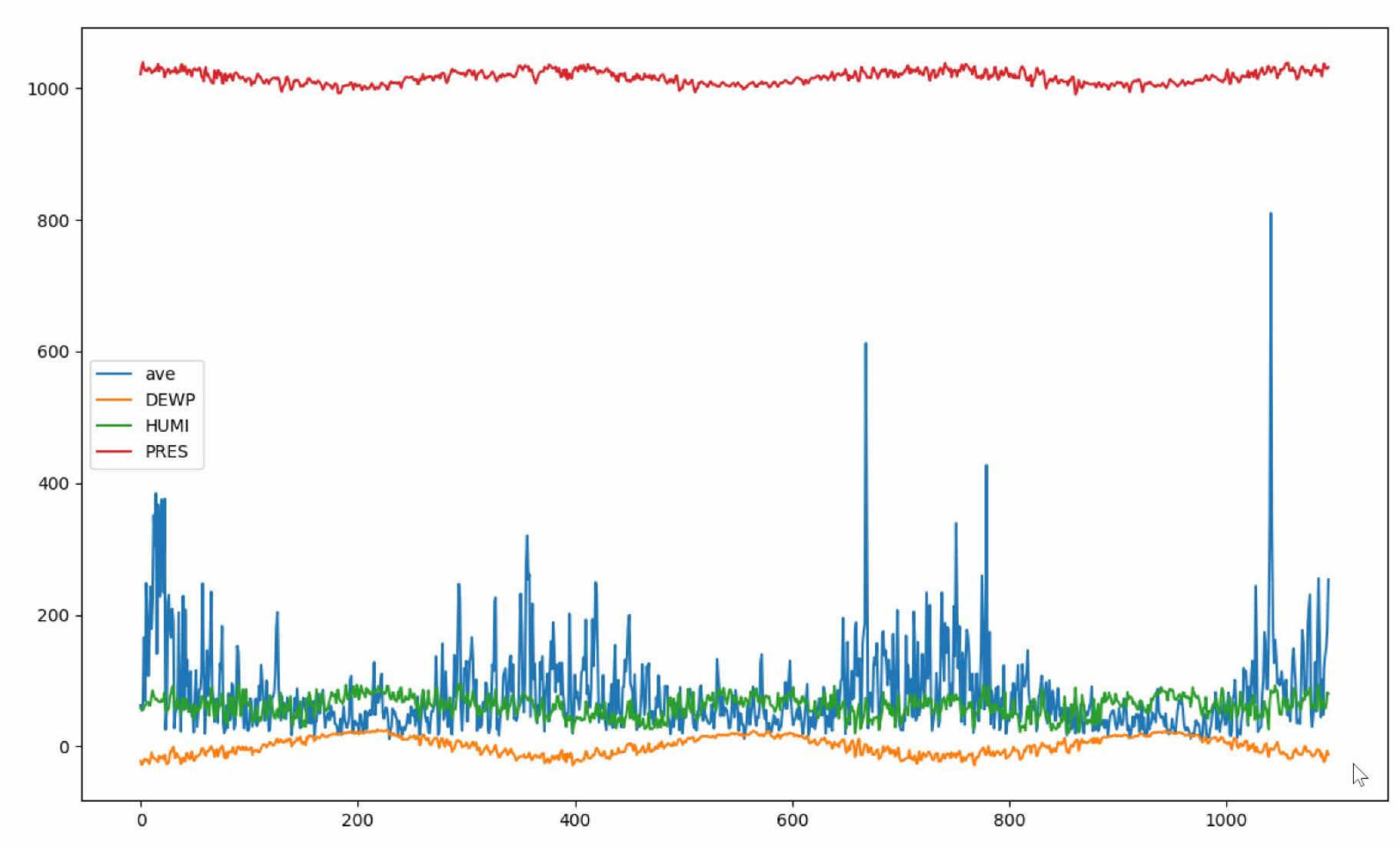

2013
2014
2015
250
200
150
100
50
0
2
4
6
8
10
12


1
2
3
4
5
6
7
8
9
10
11
12
250
200
150
100
50
0
2013.0
2013.5
2014.0
2014.5
2015.0
2015.5
2016.0

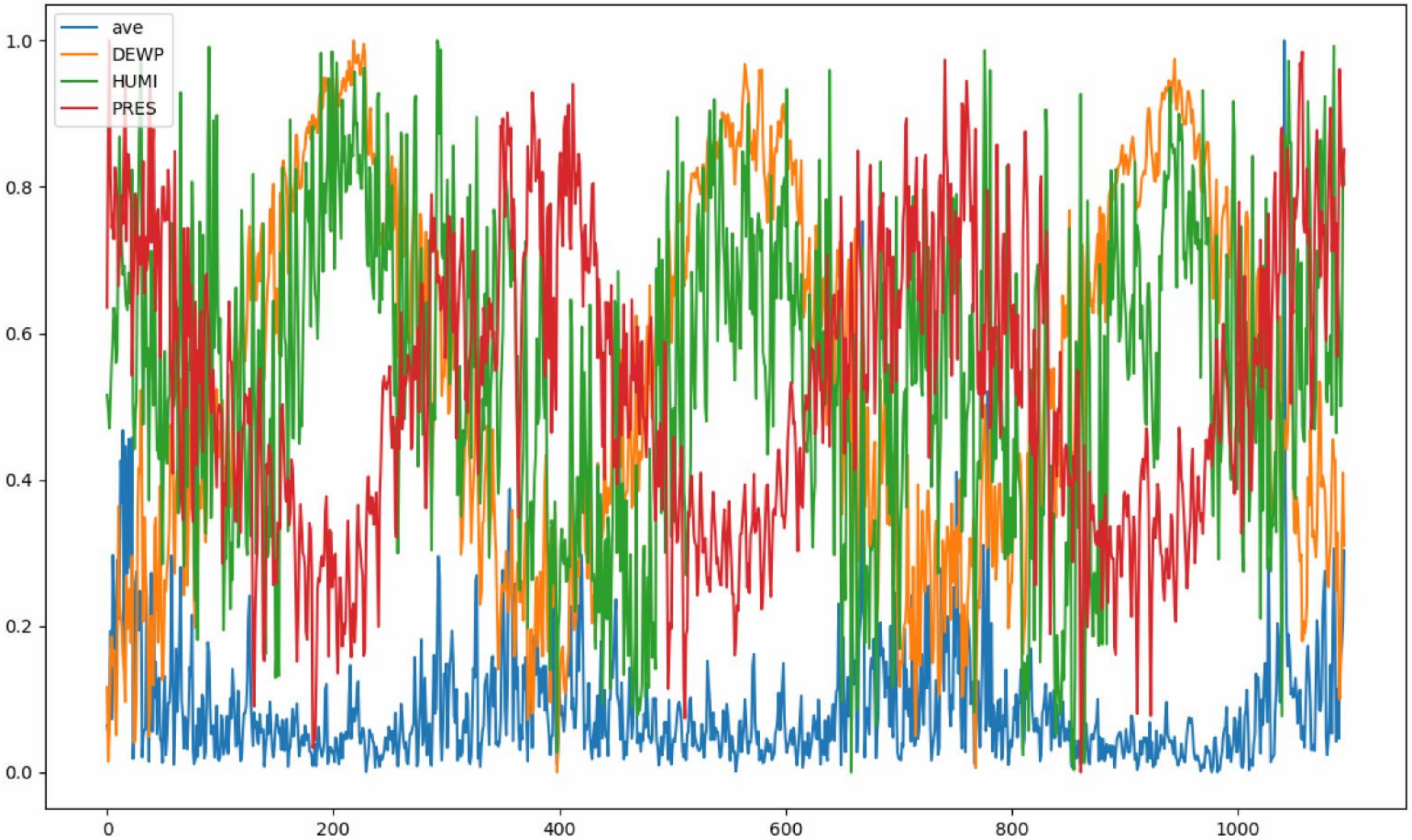
数据建模



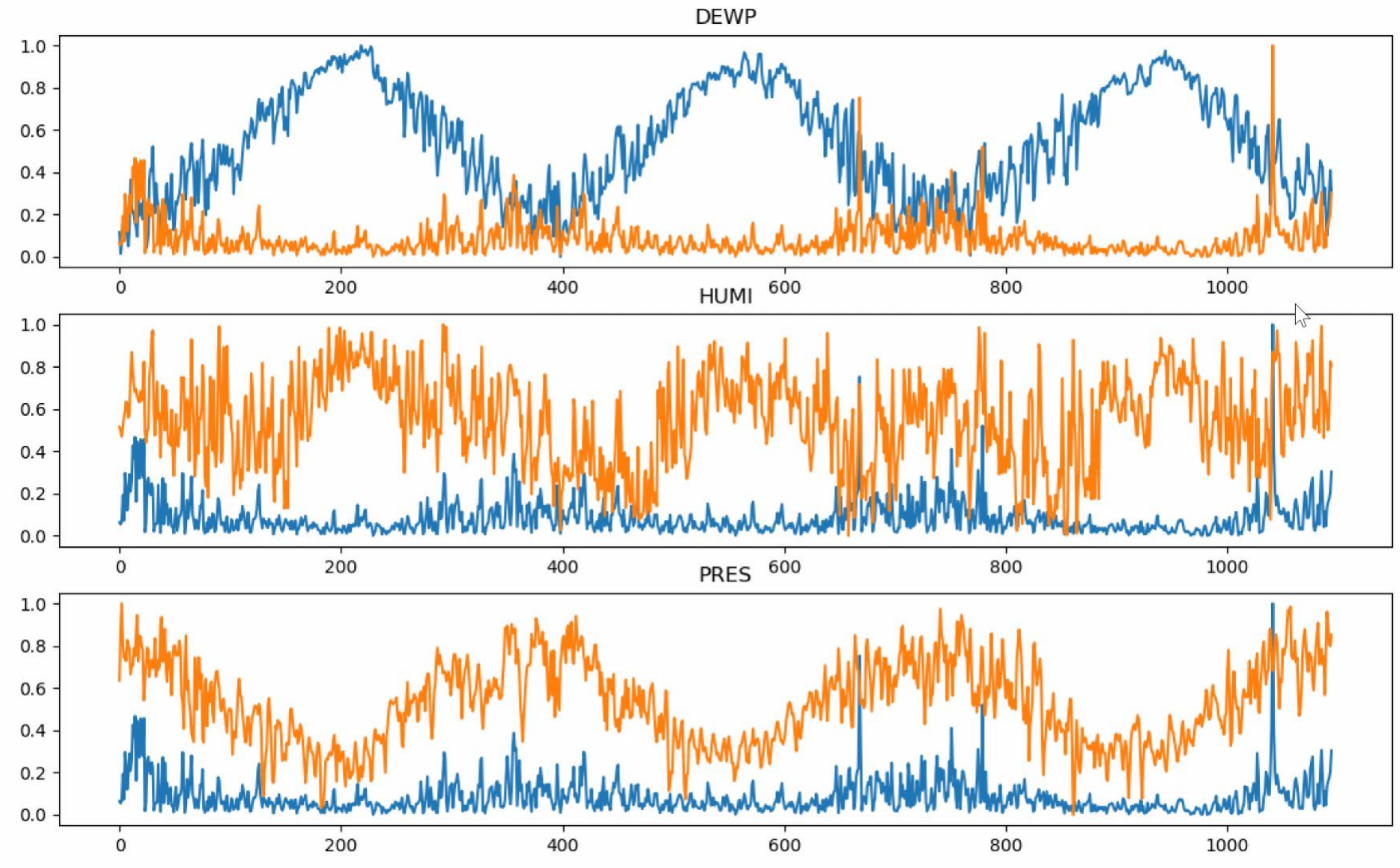
数据特征分析



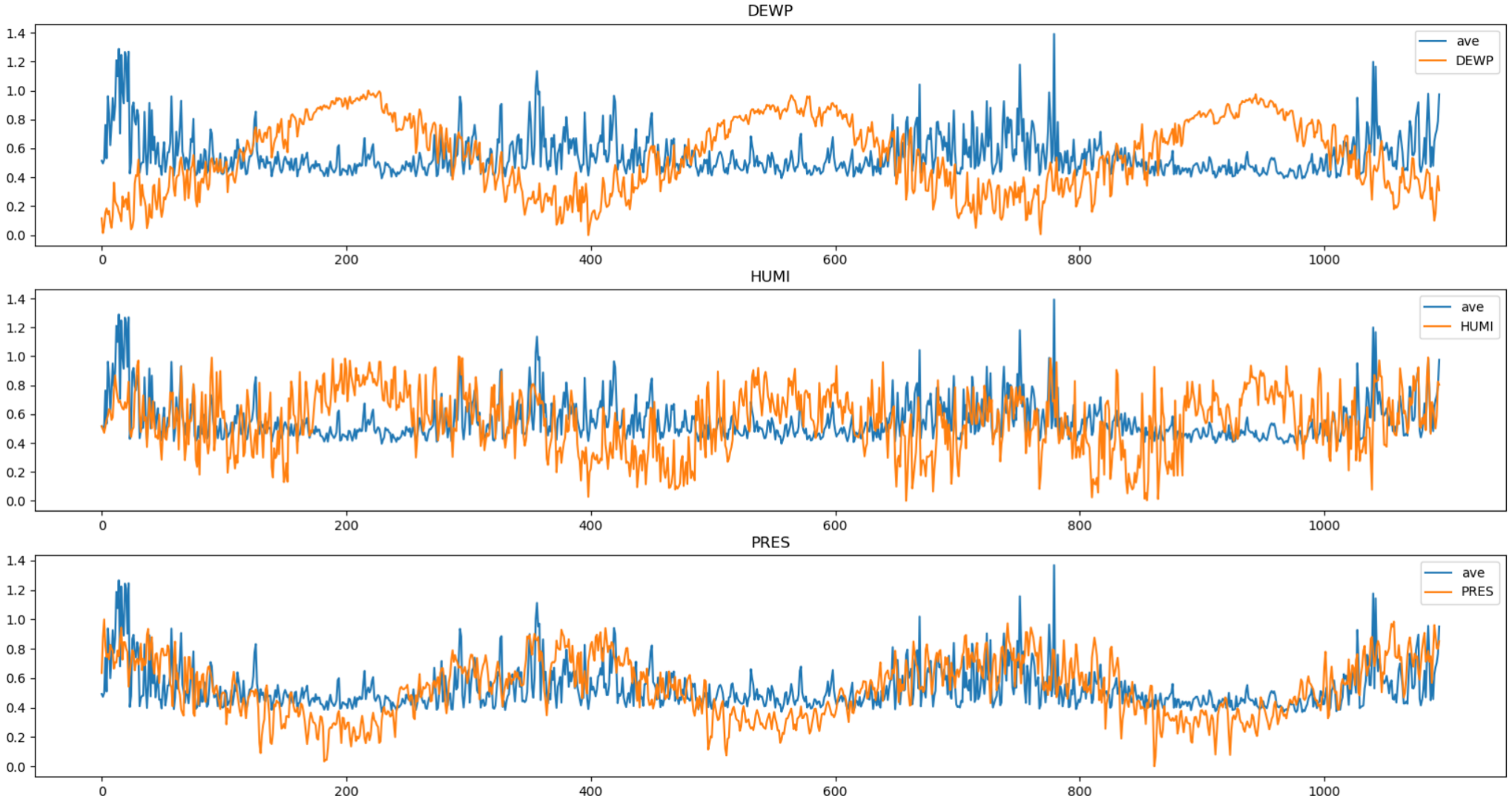
数据特征分析



数据特征分析


DEWP
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
HUMI
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
PRES
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000

数据特征分析


DEWP
ave
DEWP
HUMI
ave
HUMI
PRES
ave
PRES
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
200
400
600
800
1000

本章的主要内容

第1节 数据特征分析

第2节 数学建模

1. 基于时间序列的分析和预测

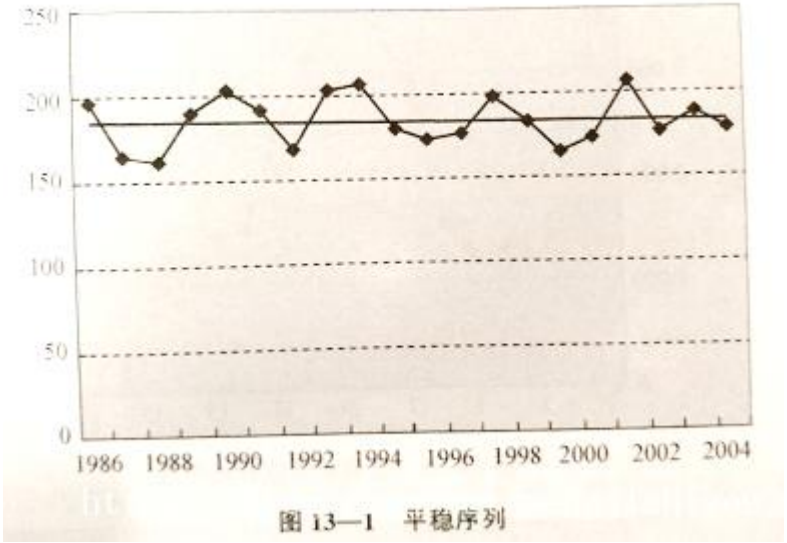


图 13—1　平稳序列

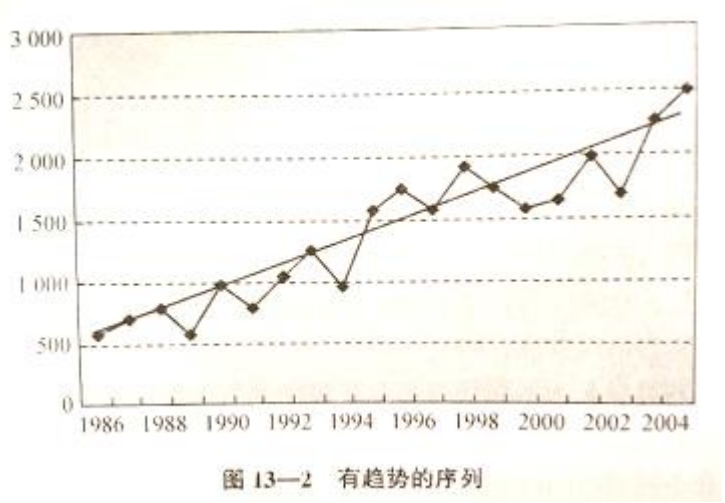


图 13—2　有趋势的序列

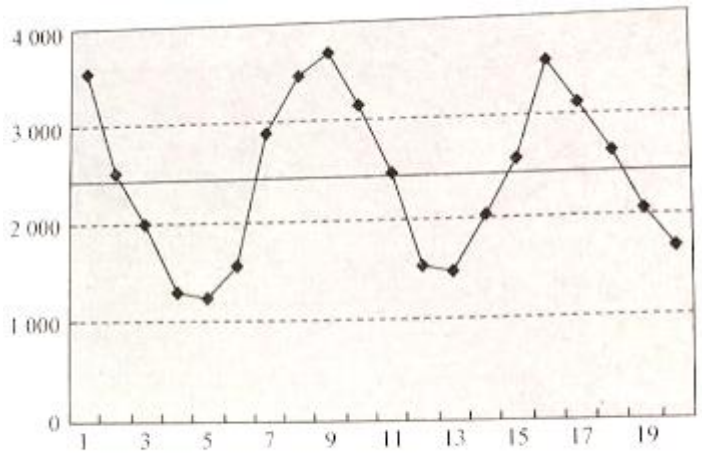


图 13—3　含有季节成分的序列

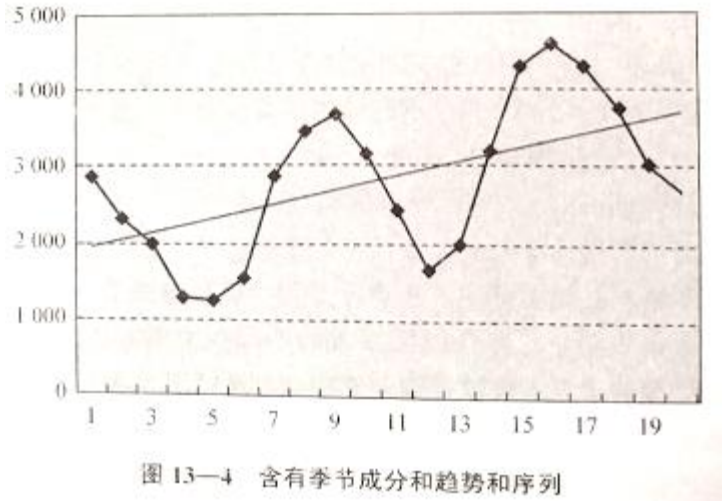
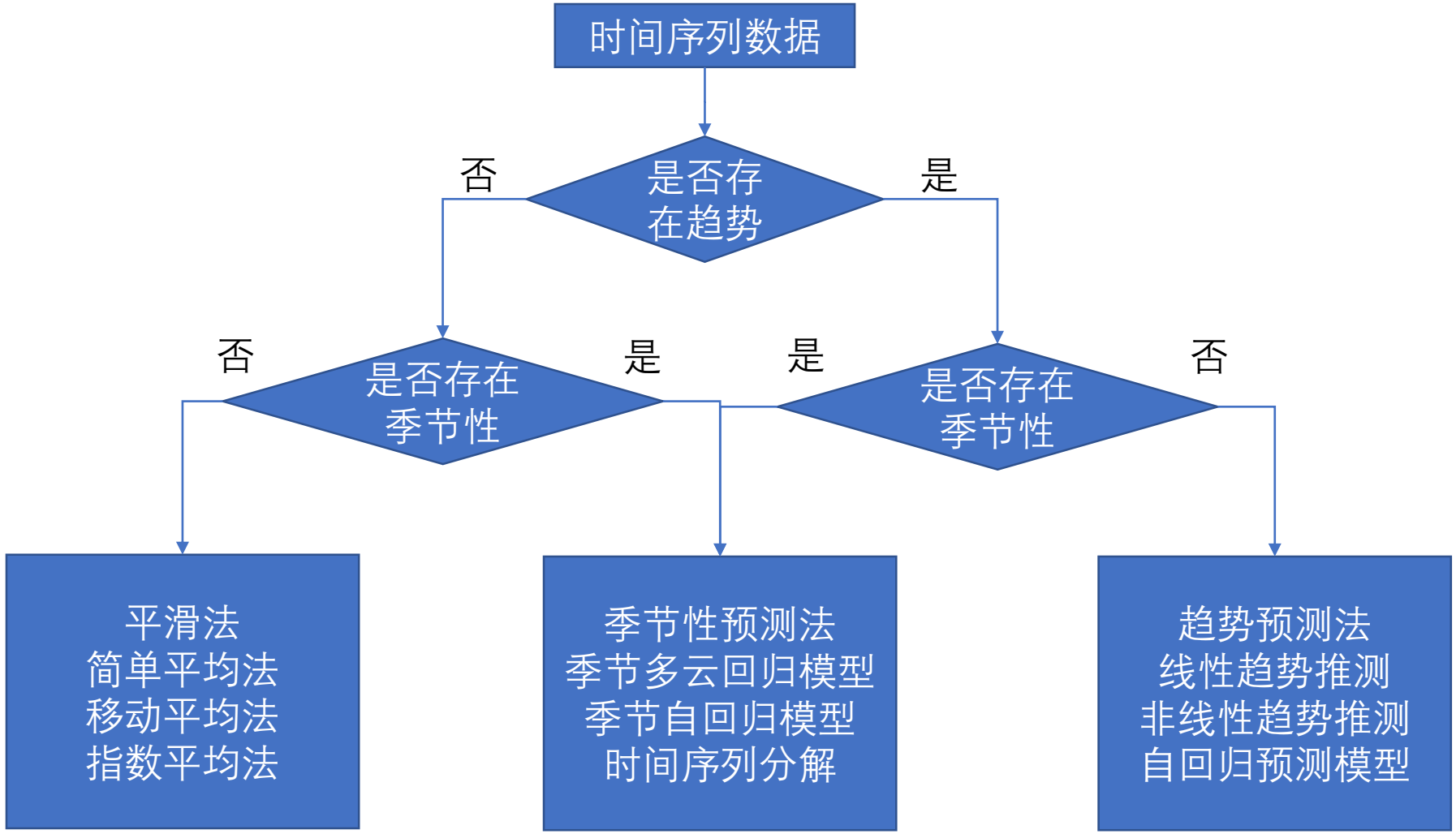


图 13—4　含有季节成分和趋势和序列

数学建模

1. 基于时间序列的分析和预测

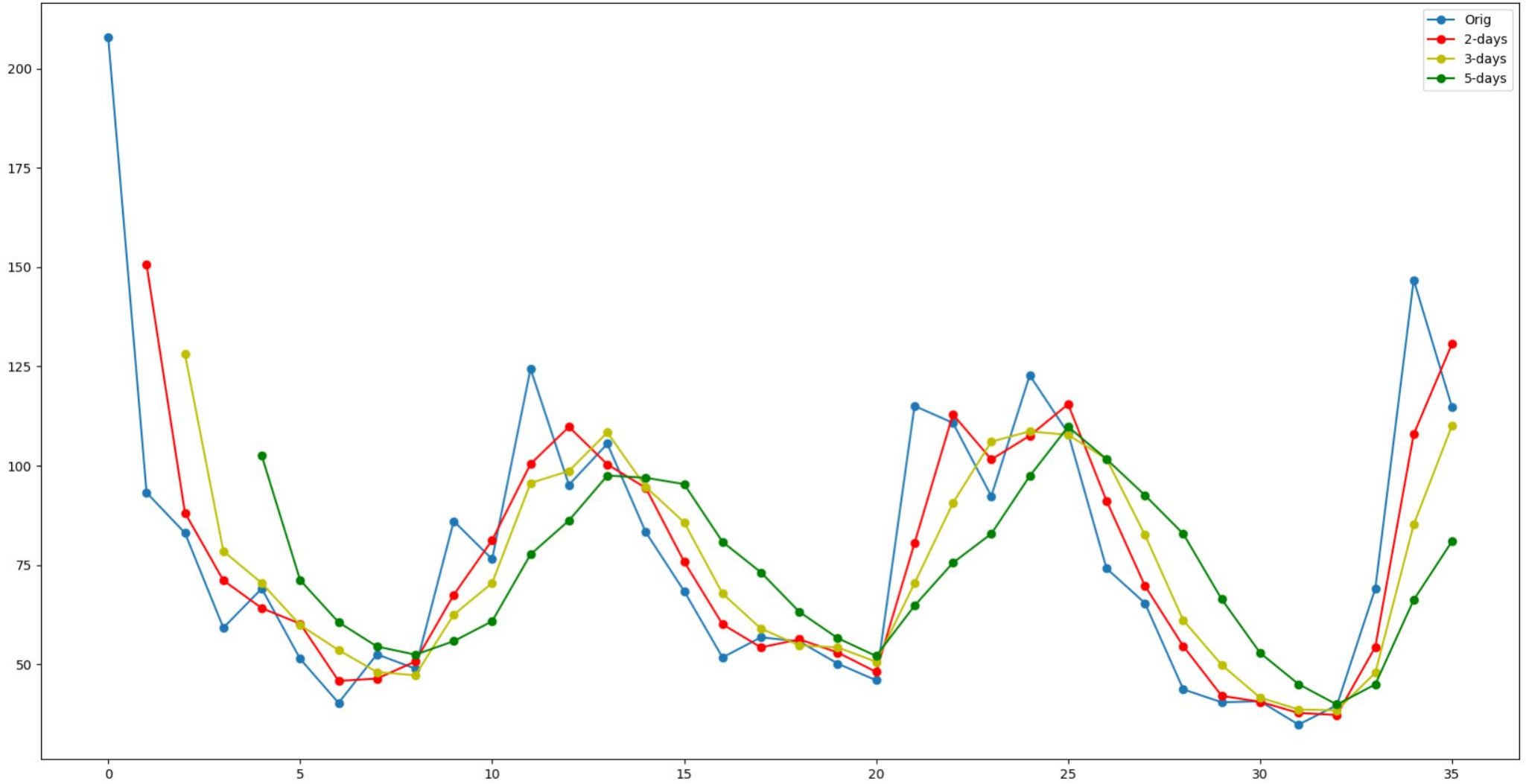


1. 基于时间序列的分析和预测

评价方法

- **平均误差**（mean error：由于预测误差的数值可能有正有负，求和的结果就会相互抵消，这种情况下，平均误差可能会低估误差。
- **平均绝对误差**mean absolute deviation：平均绝对误差可避免误差相互抵消的问题，因而可以准确反映实际预测误差的大小。
- **均方误差**（mean square error）：通过平方消去误差的正负号后计算的平均误差。
- **均方标准差**（mean square error）：通过平方消去误差的正负号后计算的平均误差。

1. 基于时间序列的滑动窗口预测



作业

根据北京（或者其它城市）的数据，按照天的频度，展示雾霾的变化情况；

尝试使用不同的滑动窗口，看看哪个参数得到的误差最小。

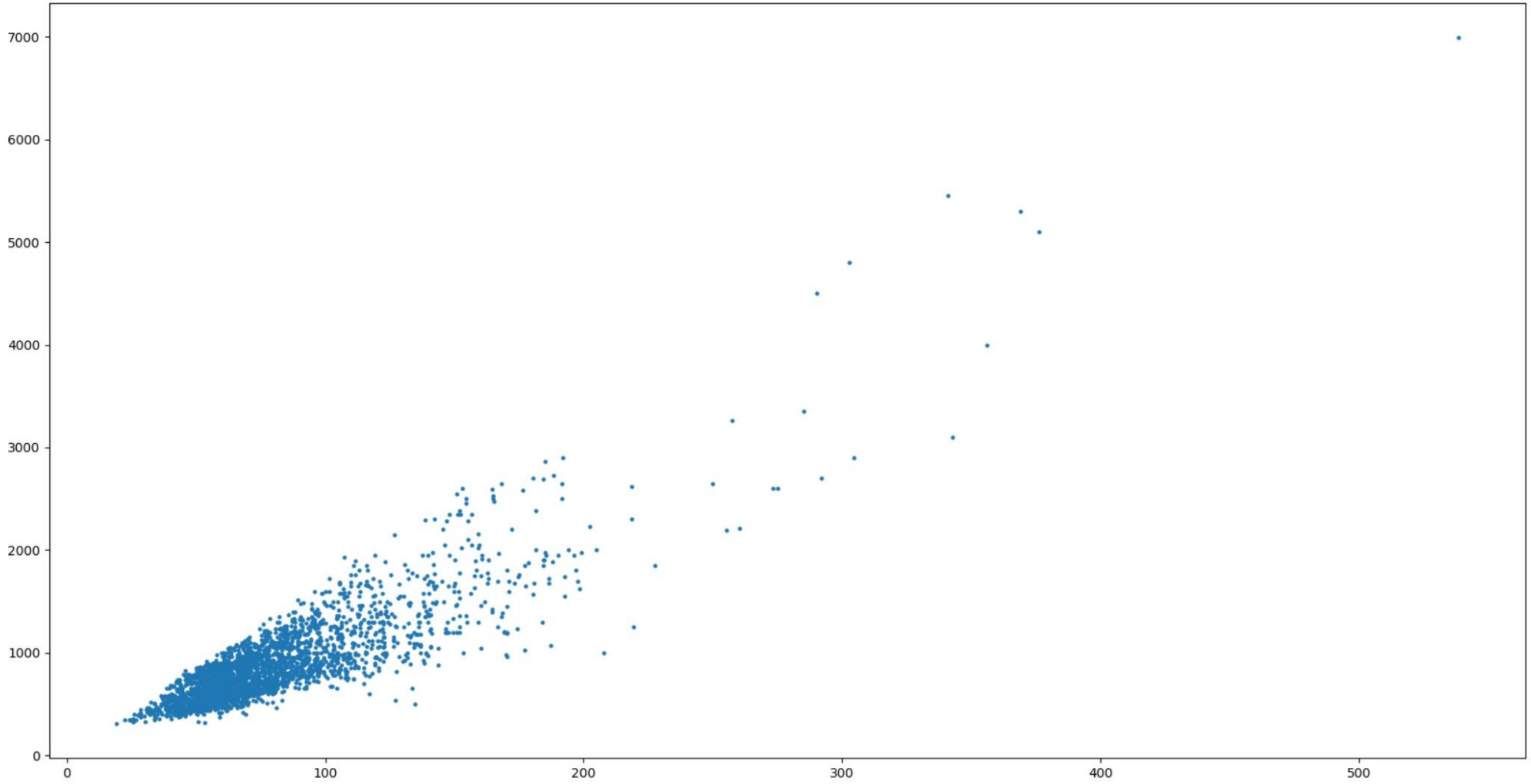
2. 线性回归

线性回归是利用称为线性回归方程的最小二乘函数，对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种函数是一个或多个称为回归系数的模型参数的线性组合。

一元线性回归：$Y=aX+b$

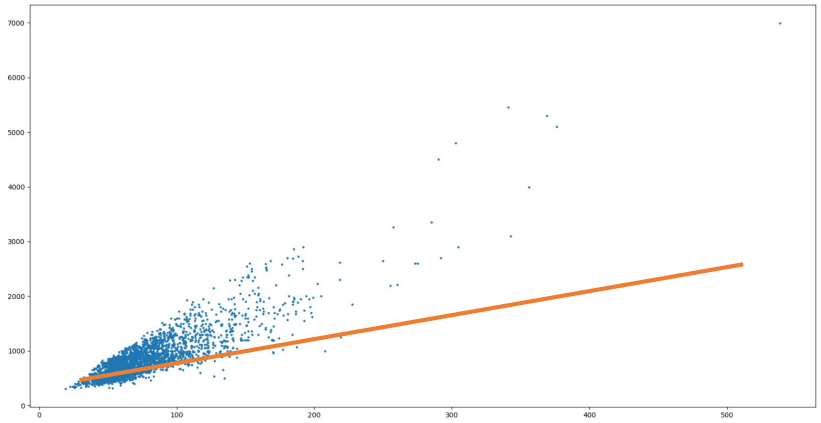
多元线性回归：$Y=a_1X1+ a_2X2+\cdots+ a_nXn+b$

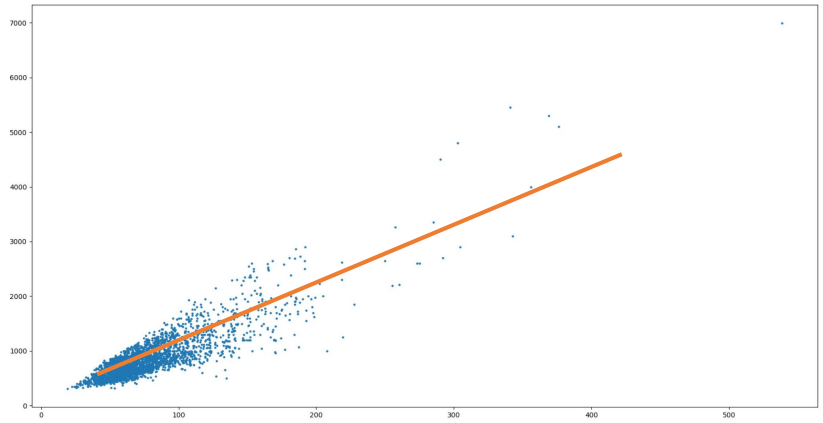
2. 线性回归

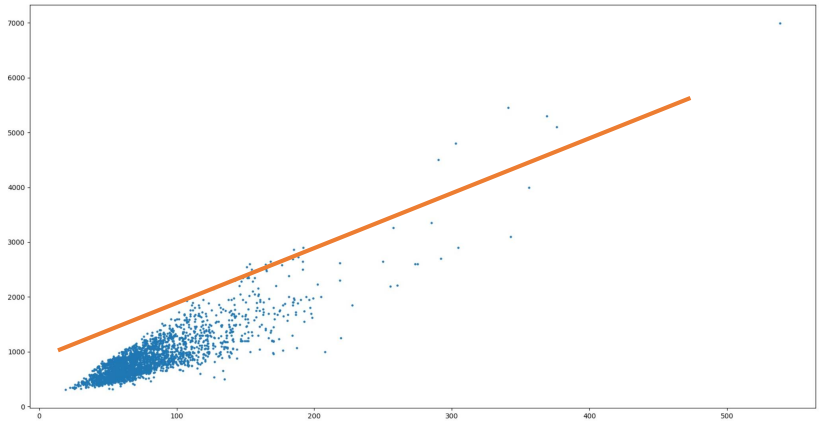


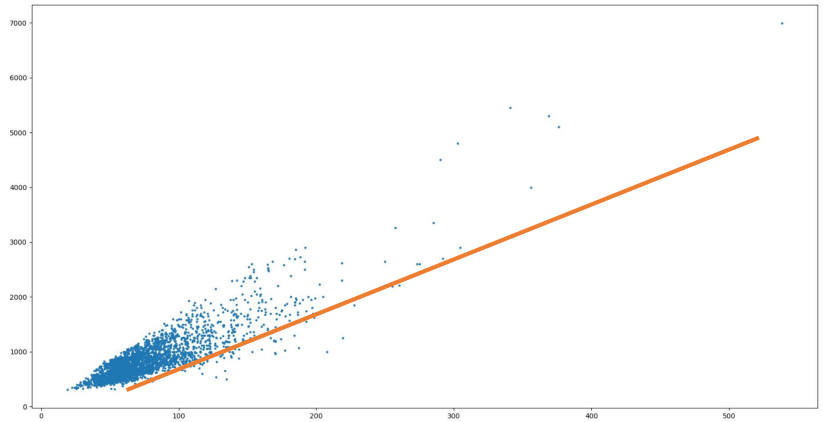
数学建模

2. 线性回归

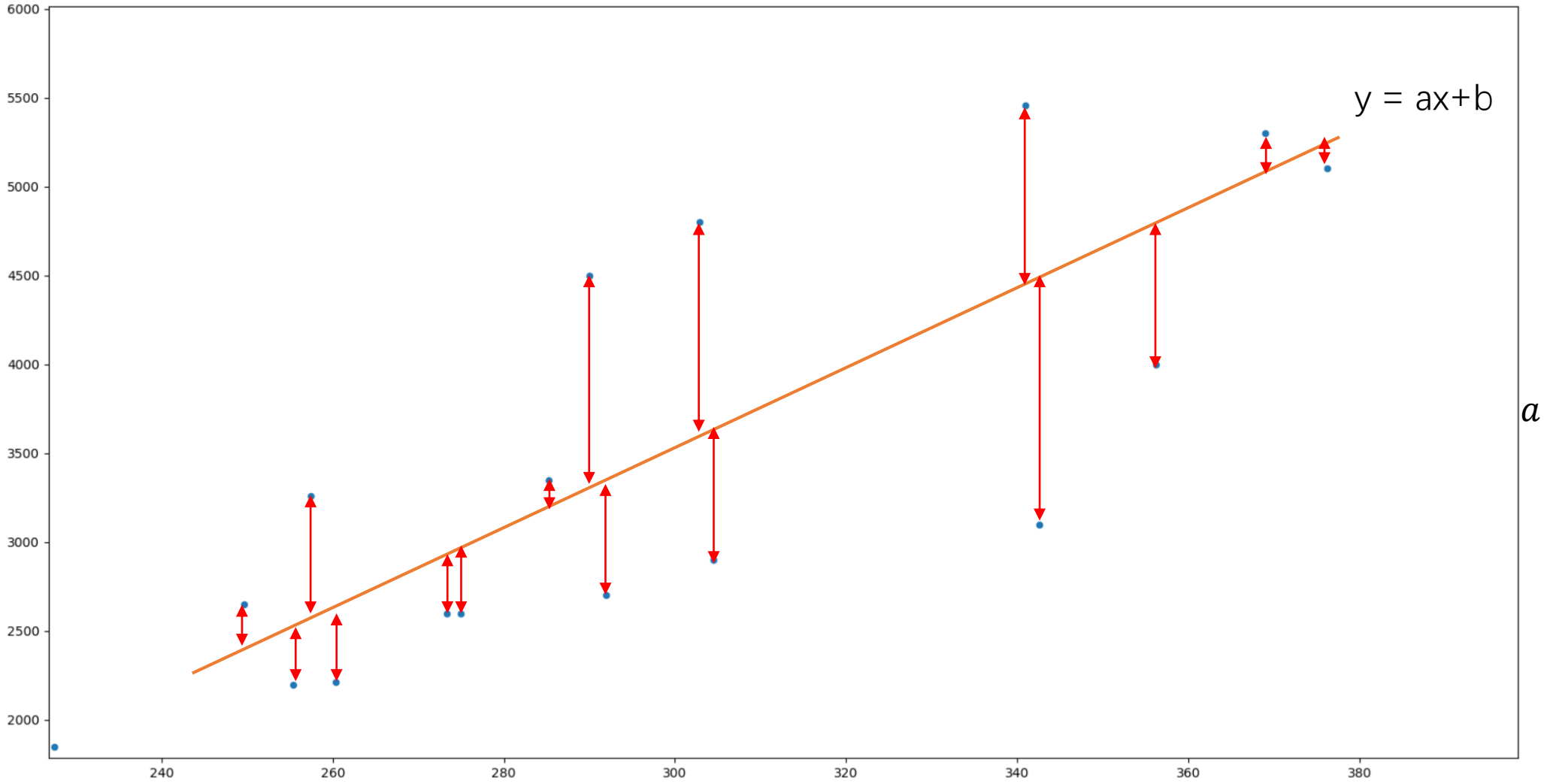

7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500


7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500


7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500


7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500

2. 线性回归

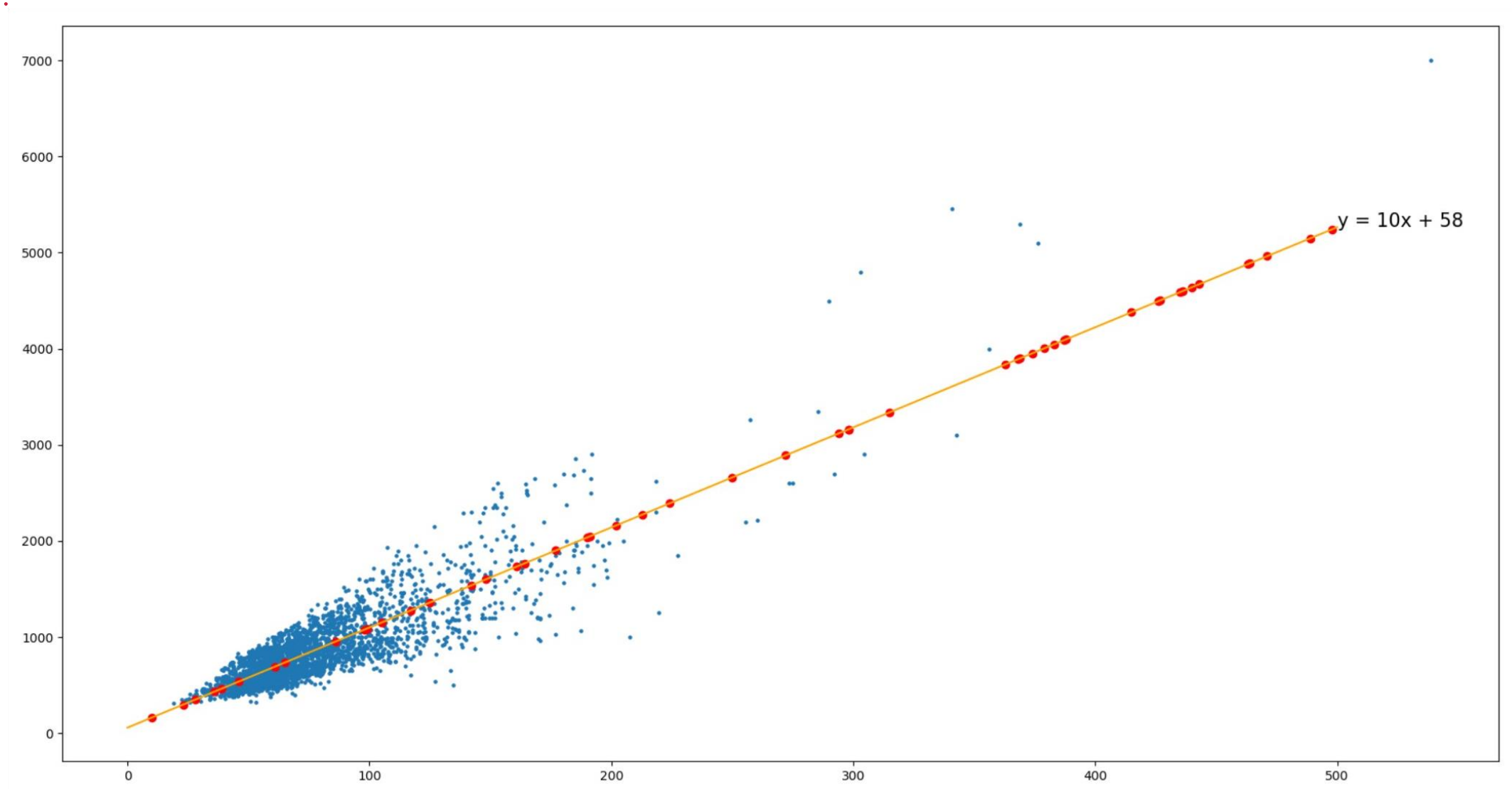


$$min \sum (y_i - \hat{y})^2$$

$$a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\mathrm{b} = \bar{y} - a\bar{x}$$

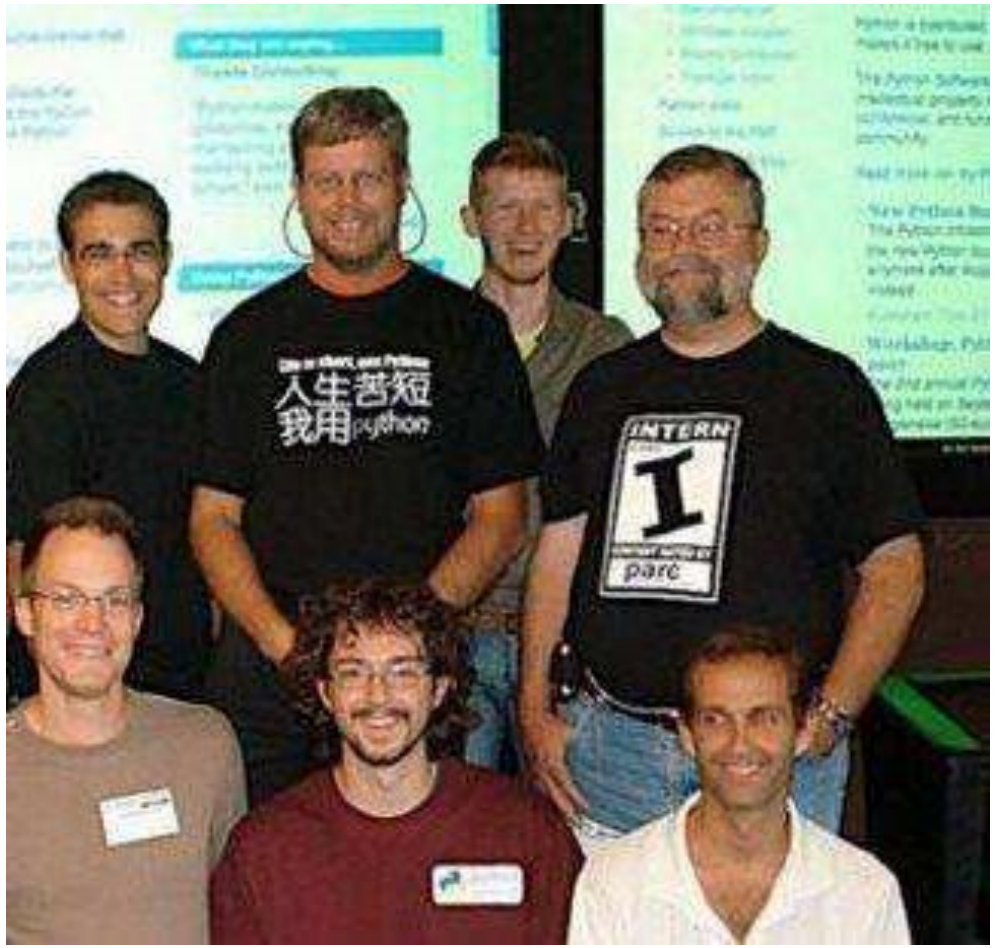
2. 线性回归


y = 10x + 58
7000
6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500

作业

从链家官网的二手房数据中，通过线性回归方法，找出一个城市或者一个区，房屋面积和总价的对应关系，给出公式，并画出图形。

人生苦短，我用python



人生苦短，我用Python！