PYTHON程序设计

计算机学院 王纯

十 数据预处理

十 数据预处理

- 数据缺失值的处理
- 异常值的处理
- 数据归一化
- 数据连续属性离散化

概述

- 采集的原始数据里存在着各种不利于分析与建模工作的因素，比如数据不完整，格式不正确，数据之间存在矛盾，异常值等。这些因素不仅会影响建模的执行过程，更有甚者在不知不觉间给出错误的建模结果，这就使得数据的预处理显得尤为重要。

通过爬虫取得的二手房数据


```
{"location": ["三环新城8号院 ", "玉泉营"], "house_info": ["2室1厅 | 102.28平米 | 西南 | 精装 | 中楼层(共15层) | 2007年建 | 板塔结合"], "price_info":
{"location": ["刘家窑南里 ", "蒲黄榆"], "house_info": ["1室1厅 | 45.49平米 | 南 | 精装 | 顶层(共6层) | 1991年建 | 板楼"], "price_info": ["258"]}
{"location": ["西府景园 ", "岳各庄"], "house_info": ["3室2厅 | 119.39平米 | 南 北 | 简装 | 中楼层(共23层) | 2004年建 | 板楼"], "price_info": ["685"]
{"location": ["万年花城二期 ", "玉泉营"], "house_info": ["2室1厅 | 104.9平米 | 西南 | 精装 | 低楼层(共15层) | 2007年建 | 板楼"], "price_info": ["753"
{"location": ["玉皇庄小区 ", "丰台其它"], "house_info": ["2室1厅 | 51.77平米 | 南 | 精装 | 顶层(共6层) | 2001年建 | 板楼"], "price_info": ["179"]}
{"location": ["德胜里一区 ", "德胜门"], "house_info": ["4室1厅 | 83.3平米 | 南 北 | 简装 | 中楼层(共5层) | 1980年建 | 板楼"], "price_info": ["1290"]
{"location": ["雅成一里 ", "朝青"], "house_info": ["2室1厅 | 87.49平米 | 西南 | 简装 | 24层 | 2001年建 | 塔楼"], "price_info": ["420"]}
{"location": ["酒仙桥十街坊 ", "酒仙桥"], "house_info": ["2室1厅 | 52.92平米 | 西南 东北 | 简装 | 底层(共6层) | 1986年建 | 板楼"], "price_info": ["31
{"location": ["华瀚国际 ", "欢乐谷"], "house_info": ["3室2厅 | 164.98平米 | 南 北 | 精装 | 中楼层(共20层) | 2008年建 | 板楼"], "price_info": ["1400"
{"location": ["融科橄榄城二期 ", "望京"], "house_info": ["3室2厅 | 139.68平米 | 南 北 | 精装 | 高楼层(共28层) | 2007年建 | 板楼"], "price_info": ["16
{"location": ["车公庄北里 ", "车公庄"], "house_info": ["3室1厅 | 93平米 | 南 北 | 简装 | 中楼层(共6层) | 1995年建 | 板楼"], "price_info": ["1200"]}
{"location": ["晨光家园B区 ", "石佛营"], "house_info": ["3室2厅 | 133.96平米 | 南 北 | 精装 | 21层 | 2008年建 | 板楼"], "price_info": ["880"]}
{"location": ["金隅山墅 ", "玉泉路"], "house_info": ["4室4厅 | 224.81平米 | 南 北 | 毛坯 | 中楼层(共4层) | 2009年建 | 暂无数据"], "price_info": ["160
{"location": ["金茂逸墅 ", "亦庄开发区其它"], "house_info": ["4室2厅 | 160.21平米 | 南 北 | 精装 | 中楼层(共17层) | 2017年建 | 板楼"], "price_info":
```


JSON文件->CSV文件

问题	处理
名字：有空格	去掉空格
价格：字符串，不方便后续计算	转换为数字
描述部分：内容太杂，格式不规整	分成多列，分别是房型、面积、朝向、装修情况等 增加一列单价，并按降序排列

作业一

- 把通过爬虫爬下来的新房数据，进行预处理：
  - 最终的csv文件，应包括以下字段：名称，地理位置（3个字段分别存 储），房型（只保留最小房型），面积（按照最小值），总价（万元， 整数），均价（元，整数）；
  - 对于所有字符串字段，要求去掉所有的前后空格；
  - 如果有缺失数据，不用填充。

新房数据预处理

德贤御府 住宅 在售

朝阳 / 朝阳其它 / 北京市朝阳区诚源三路

2室 / 3室 / 4室

建面 72-144㎡

新房顾问：高志强 沟通

品牌房企 小户型 小型社区

关注

80000 元/㎡(均价)

总价580-1152(万/套)

水岸壹号 别墅 待售

房山 / 良乡 / 良乡大学城西站地铁南侧800米，刺猬河旁

3室 / 4室

建面 185-199㎡

新房顾问：赵海城 沟通

地铁沿线 环线房 成熟商圈 配套齐全

58000 元/㎡(均价)

总价1100-1300(万/套)

名称，地理位置（3个字段分别存储），房型（只保留最小房型），面积（按照最小值，整数），均价（元，整数），总价（万元，整数）。注：放大显示的也可能有总价，爬取该元素的值并判断是总价还是均价，填入相应字段，相应的另一字段由填入的数值与面积计算获得。

雾霾数据预处理

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
No	year	month	day	hour	season	PM_Dongsi	PM_Dongsil	PM_Nongzh	PM_US Po:	DEWP	HUMI	PRES	TEMP	cbwd	Iws	precipitation	Iprec
33927	2013	11	14	14	3	25	22	28	27	-15	11	1017	14	NW	39.79	0	0
33928	2013	11	14	15	3	38	20	26	27	-16	11	1017	14	NW	48.73	0	0
33929	2013	11	14	16	3	71	46	61	69	-16	11	1017	14	NW	55.88	0	0
33930	2013	11	14	17	3	75	62	83	64	-16	11	1017	13	NW	59.9	0	0
33931	2013	11	14	18	3	75	61	56	65	-12	17	1018	12	SE	3.13	0	0
33932	2013	11	14	19	3	95	59	70	65	-11	20	1018	11	SE	6.26	0	0
33933	2013	11	14	20	3	98	54	67	76	-9	30	1018	7	cv	0.89	0	0
33934	2013	11	14	21	3	89	67	75	72	-10	24	1019	9	cv	2.68	0	0
33935	2013	11	14	22	3	85	68	73	78	-10	23	1019	10	SE	1.79	0	0
33936	2013	11	14	23	3	94	74	70	81	-10	24	1019	9	NW	1.79	0	0

<table>
<tr><td>No：记录编号</td><td>Season：季节</td><td rowspan="2">DEWP：露点 (摄氏温度) 指在固定气压之下，空气 中所含的气态水达到饱和而凝结成液态水所需要 降至的温度。</td><td>TEMP：温度（摄氏）</td></tr>
<tr><td>Year：年份</td><td rowspan="2">PM：PM2.5浓度 (ug/m^3)</td><td>cbwd：组合风向</td></tr>
<tr><td>Month：月份</td><td rowspan="2">Precipitation：降水量/时 (mm)</td><td>Iws：累计风速 (m/s)</td></tr>
<tr><td>Day: 日期</td><td>HUMI：湿度（%）</td></tr>
<tr><td>Hour：小时</td><td>PRES：气压 (hPa)c</td><td colspan="2">Iprec：累计降水量 (mm) m</td></tr>
</table>

作业二

- 计算北京空气质量数据
  - 汇总计算PM指数年平均值的变化情况
  - 汇总计算每年中1-12月的PM指数数据变化情况

No：记录编号	Season：季节	DEWP：露点 (摄氏温度) 指在固定气压之下，空气 中所含的气态水达到饱和而凝结成液态水所需要 降至的温度。	TEMP：温度（摄氏）
Year：年份	PM：PM2.5浓度 (ug/m^3)		cbwd：组合风向
Month：月份			Iws：累计风速 (m/s)
Day: 日期	HUMI：湿度（%）	Precipitation：降水量/时 (mm)	
Hour：小时	PRES：气压 (hPa)c	Iprec：累计降水量 (mm) m	

数据缺失值的处理

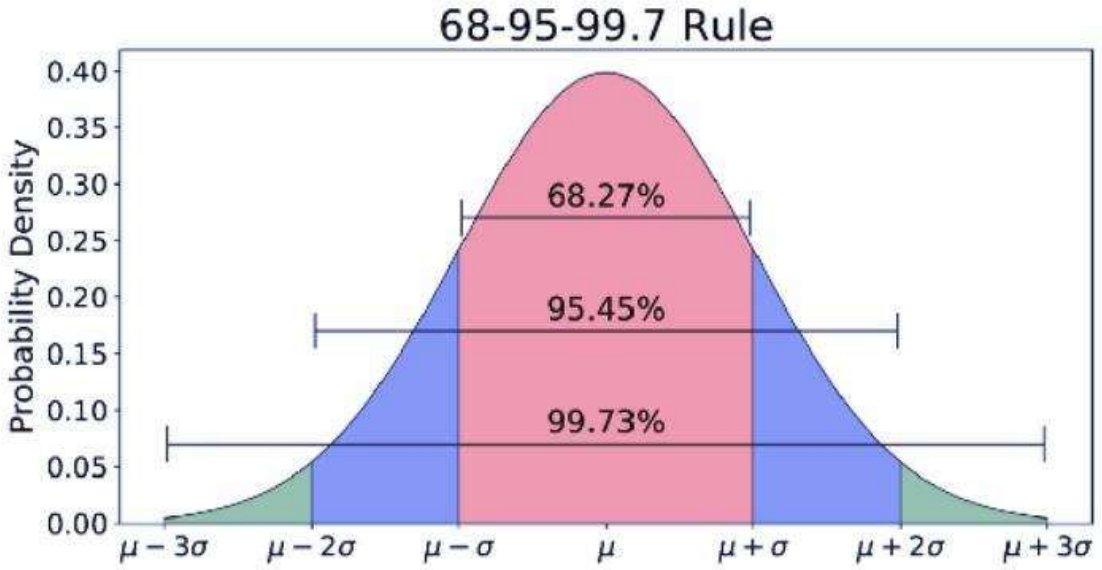
- 忽略，不参与计算
- 删除
- 插值 （下列为针对单元格数据的插值方法）
  - interpolate：线性插值
  - ffill：前向填充
  - bfill：后向填充

■ 沈阳空气质量数据，计算PM指数年平均值的变化情况

No	year	month	day	hour	season	PM_Taiyua	PM_US Po	PM_Xiaohe	DEWP	HUMI	PRES	TEMP	cbwd	Iws	precipitatio	Iprec
1	2010	1	1	0	4	NA	NA	NA	-26	69.79	1024	-22	NE	1.0289	NA	NA
2	2010	1	1	1	4	NA	NA	NA	-26	76.26	1024	-23	NE	2.5722	NA	NA
3	2010	1	1	2	4	NA	NA	NA	-27	69.56	1023	-23	NE	5.1444	NA	NA
4	2010	1	1	3	4	NA	NA	NA	-27	69.56	1023	-23	NE	7.7166	NA	NA
5	2010	1	1	4	4	NA	NA	NA	-27	69.56	1022	-23	NE	9.7744	NA	NA
6	2010	1	1	5	4	NA	NA	NA	-26	76.26	1022	-23	NE	11.8322	NA	NA
7	2010	1	1	6	4	NA	NA	NA	-25	76.46	1021	-22	NE	14.4044	NA	NA
8	2010	1	1	7	4	NA	NA	NA	-24	70.26	1021	-20	NE	16.9766	NA	NA
9	2010	1	1	8	4	NA	NA	NA	-23	70.49	1021	-19	NE	19.0344	NA	NA
10	2010	1	1	9	4	NA	NA	NA	-22	70.71	1021	-18	NE	21.6066	NA	NA
11	2010	1	1	10	4	NA	NA	NA	-20	77.39	1022	-17	NE	24.1788	NA	NA
12	2010	1	1	11	4	NA	NA	NA	-18	77.75	1021	-15	NE	27.2655	NA	NA
13	2010	1	1	12	4	NA	NA	NA	-17	77.92	1020	-14	NE	29.8377	NA	NA
14	2010	1	1	13	4	NA	NA	NA	-16	78.1	1019	-13	NE	32.9244	NA	NA
15	2010	1	1	14	4	NA	NA	NA	-15	84.87	1019	-13	NE	35.4966	NA	NA
16	2010	1	1	15	4	NA	NA	NA	-15	78.27	1019	-12	NE	38.5833	NA	NA
17	2010	1	1	16	4	NA	NA	NA	-15	78.27	1019	-12	NE	41.1555	NA	NA
18	2010	1	1	17	4	NA	NA	NA	-15	78.27	1020	-12	NE	43.2133	NA	NA
19	2010	1	1	18	4	NA	NA	NA	-16	78.1	1020	-13	NE	45.7855	NA	NA
20	2010	1	1	19	4	NA	NA	NA	-17	77.92	1021	-14	NE	48.3577	NA	NA
21	2010	1	1	20	4	NA	NA	NA	-17	84.62	1021	-15	NE	50.4155	NA	NA
22	2010	1	1	21	4	NA	NA	NA	-19	77.57	1022	-16	NE	51.9588	NA	NA
23	2010	1	1	22	4	NA	NA	NA	-20	77.39	1022	-17	NE	53.5021	NA	NA
24	2010	1	1	23	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
25	2010	1	2	0	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

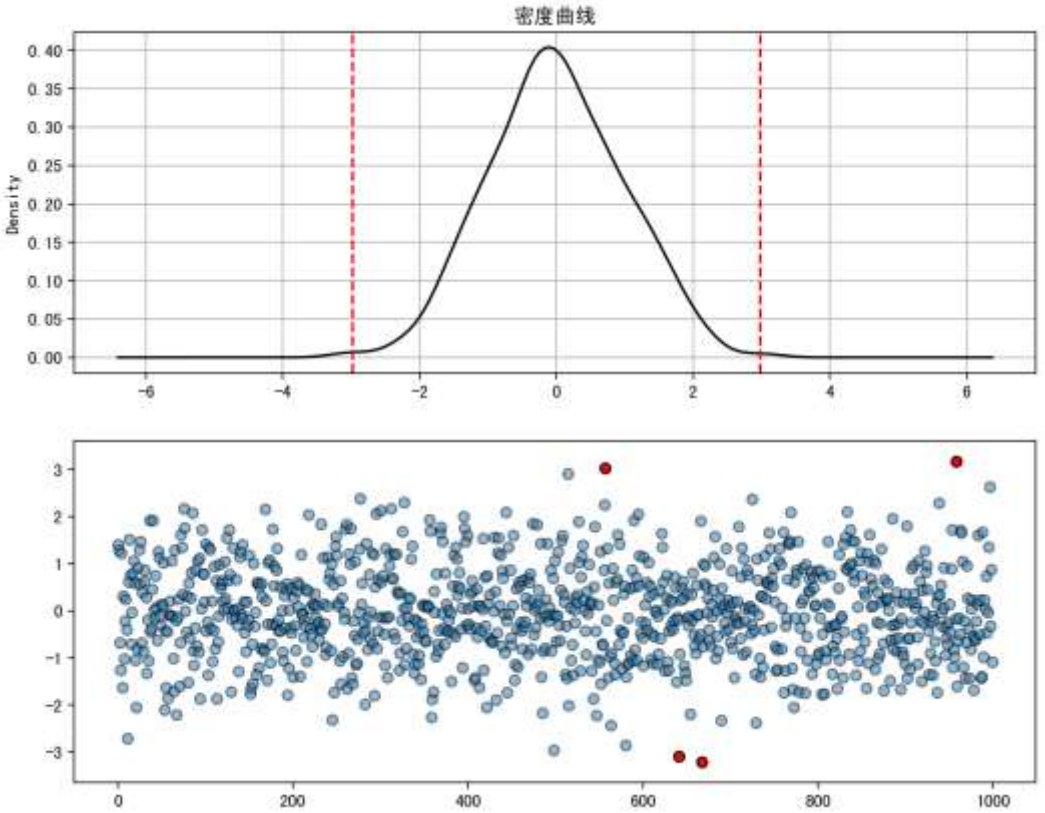
异常值的处理

- 异常值是指一组测定值中与平均值的偏差超过两倍标准差的测定值；
- 与平均值的偏差超过三倍标准差的测定值，称为高度异常的异常值。



一个正态分布的横轴区间（μ−3σ,μ+3σ）内的面积为99.7%。若是不服从正态分布，可以使用原理n倍标准差来描述，如果不合适，可以考虑使用**箱型图**，箱型图的四分位距（IQR）对异常值进行检测，也叫Tukey's Test。

- 发现异常值
  - 观察df的统计信息，使用describe和info函数，查看平均值、最小值和最大值，是否有明显的错误
  - 计算两倍标准差和三倍标准差
  - 使用图形化方式对数据进行展示和分析
- 异常值的处理
  - 直接替换为合理的数据
  - 先置为空，再使用插值的方法进行填充



作业三

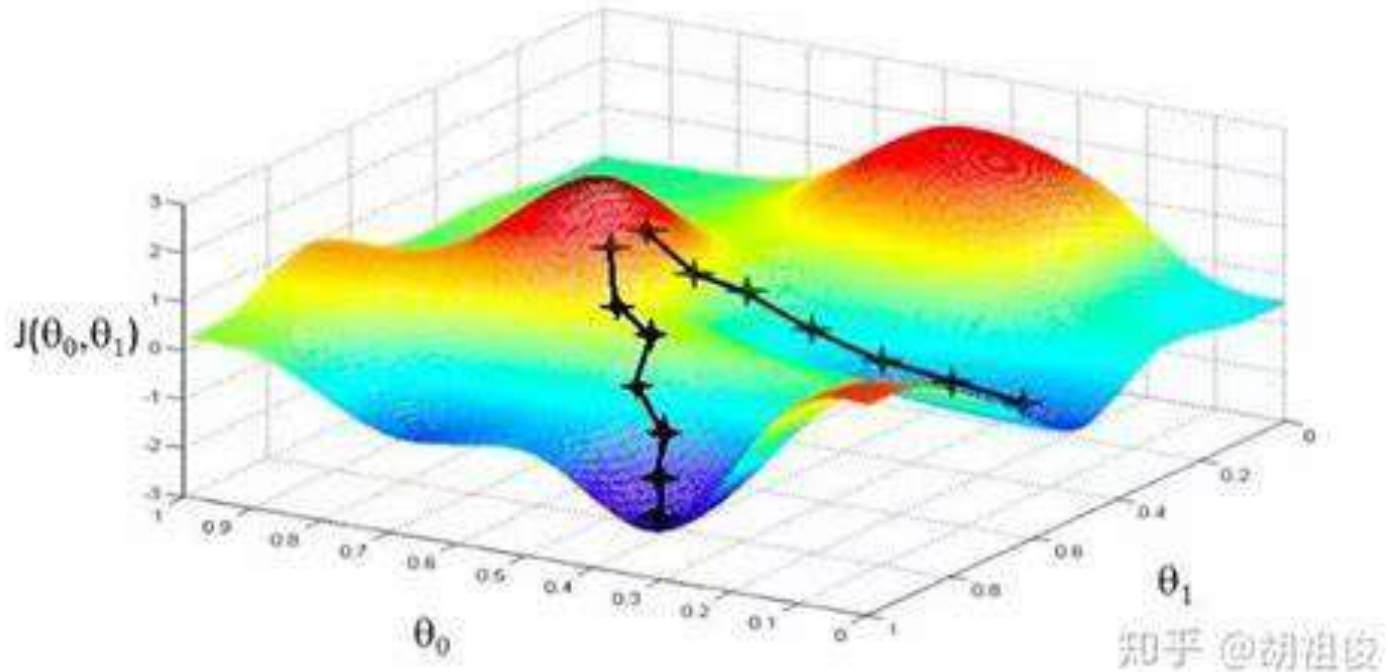
- 处理北京空气质量数据
  - 对HUMI、PRES、TEMP三列，进行线性插值处理。并对其中超过3倍标准差的高度异常数据，修改为3倍标准差的数值。
  - 假设PM指数最高为500，对PM_Dongsi、PM_Dongsihuan、 PM_Nongzhanguan三列中超过500的数据，修改为500PM指数 进行异常值的处理。
  - 修改cbwd列中值为“cv”的单元格，其值用后项数据填充。

数据的归一化

- 一个钢筋的样品，直径是0.95cm，长度是1010cm
- 直径和标准的差距是-0.05，取平方后是0.0025
- 长度和标准的差距是10，取平方后是100
- 直径的残差被忽略，而长度的残差会带来极大的影响
- 需要统一量纲 （类别、单位、量级。。。）

优点

- 归一化后加快了梯度下降求最优解的速度。
- 归一化有可能提高精度（归一化是让不同维度之间的特征 在数值上有一定的比较性）。



Rescaling (Min-Max归一化，最大最小标准化，离差标准化)： 这是一种最简单的归一化，将特征线性映射到[0,1]的范围。

$$x' = \frac{x - minA}{maxA - minA}$$

Standardization (Z-score归一化，标准化)： 在这种归一化中，对特征进行缩放，使其均值为零，方差为1。

$$x' = \frac{x - \mu}{\sigma}$$

两种归一化的比较

Rescaling (Min-Max归一化)

$$x' = \frac{x - minA}{maxA - minA}$$

梯度下降收敛效果好

VS

Standardization (Z-score归一化)

$$x' = \frac{x - \mu}{\sigma}$$

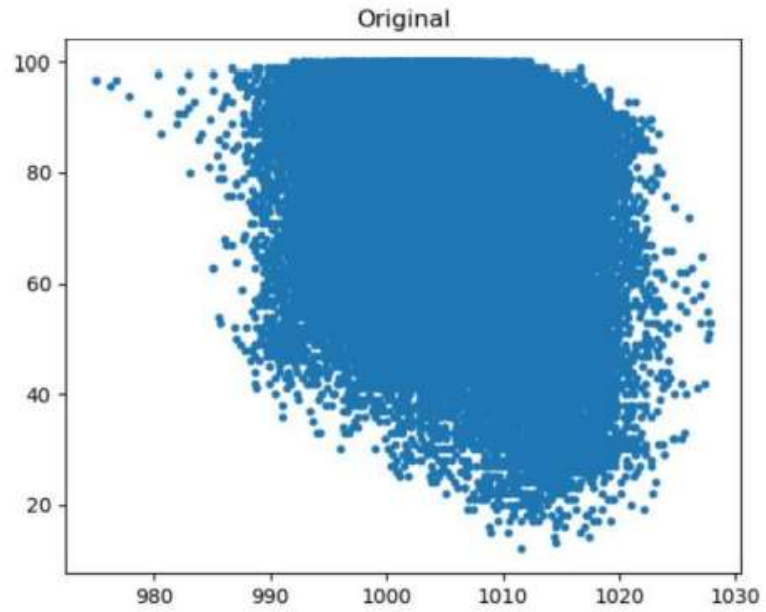
保留了样本原来的分布

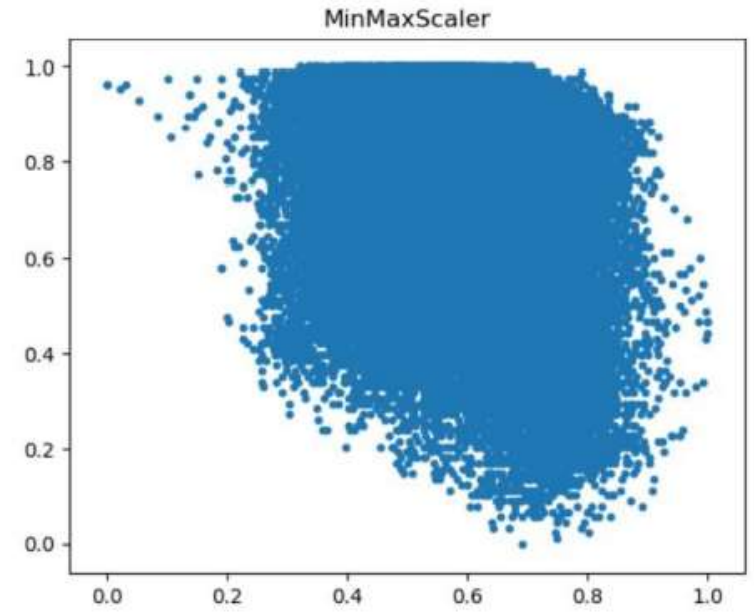
Min-Max归一化

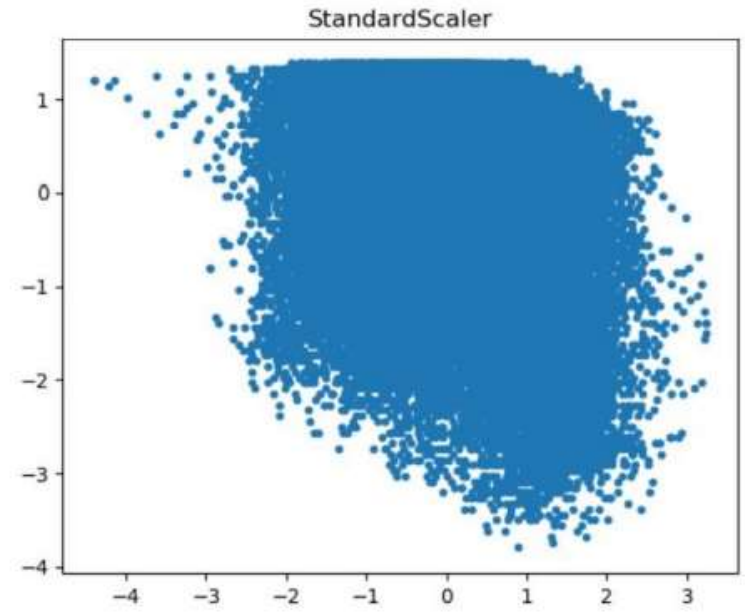
```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler() #x是df中的某一列，即series对象。
x_reshape = x.values.reshape(-1, 1) #变成n行1列的二维矩阵形式
x2 = scaler.fit_transform(x_reshape) #调用MinMaxScaler的fit_transform转换方法，  进行归一化处理
```

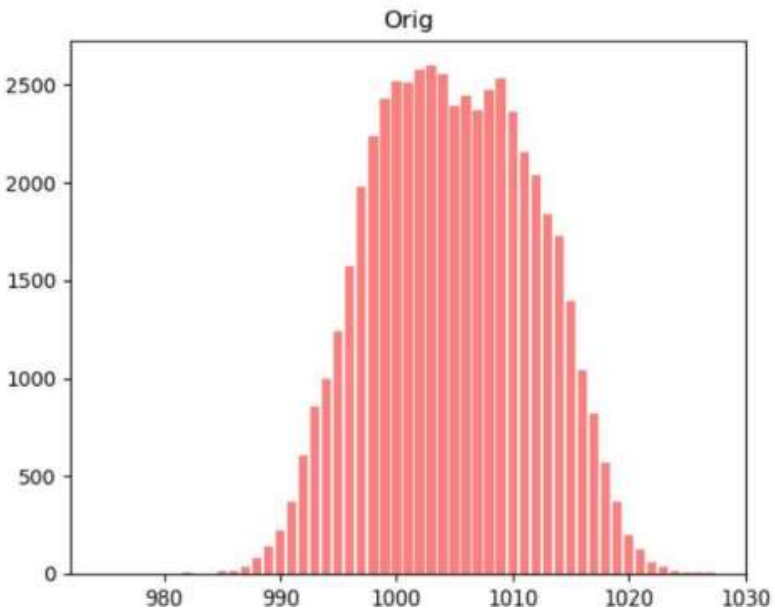
Z-score归一化

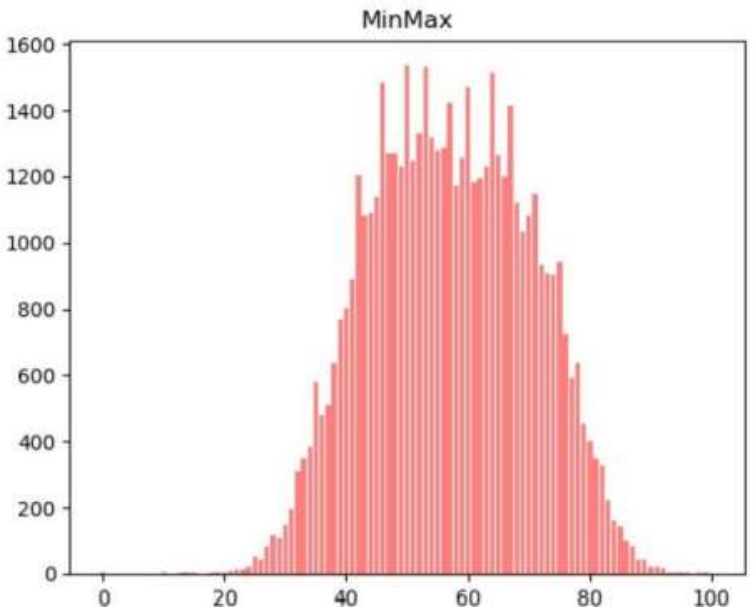
```
from sklearn.preprocessing import StandardScaler
scaler_std = StandardScaler()
x_reshape = x.values.reshape(-1, 1) #变成n行1列的二维矩阵形式
x3 = scaler_std.fit_transform(x_reshape) #调用StandardScaler的fit_transform转换 方法，进行归一化处理
```

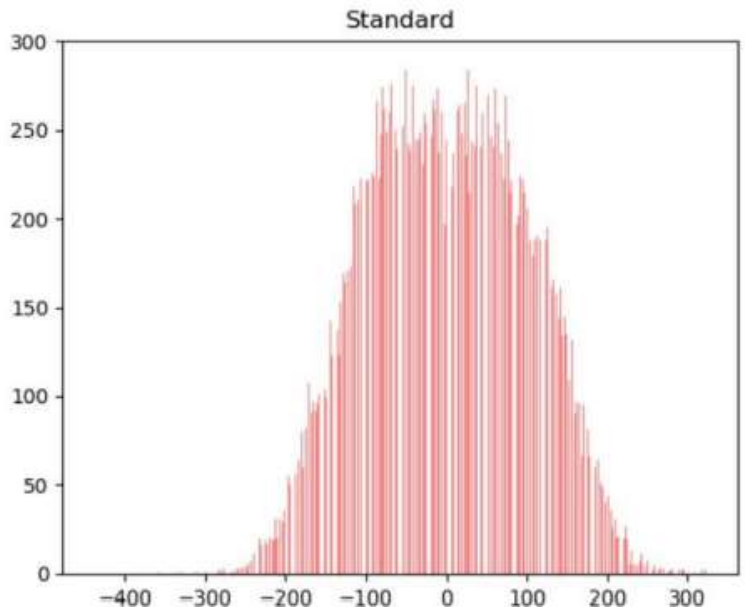

Original
100
80
60
40
20
980
990
1000
1010
1020
1030


MinMaxScaler
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


StandardScaler
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3


Orig
2500
2000
1500
1000
500
0
980
990
1000
1010
1020
1030


MinMax
1600
1400
1200
1000
800
600
400
200
0
0
20
40
60
80
100


Standard
300
250
200
150
100
50
0
−400
−300
−200
−100
0
100
200
300

数据的连续属性离散化

- 数据的特征
  - 数据的属性分为连续和离散两大类。
  - 离散属性比连续属性更接近于知识级的表达。通过对数据连续属性的离散化，数据可以被减少并被简化。对用户而言，离散的数据更易理解、使用和解释。
- 数据的离散化
  - 所谓离散化，就是把无限空间中有限的个体映射到有限的空间中。数据离散化操作大多是针对连续数据进行的，**处理之后的数据值域分布将从连续属性变为离散属性**，这种属性一般包含2个或2个以上的值域。

数据的连续属性离散化

- 数据离散化的好处
  - 节约计算资源，提高计算效率
  - 算法模型的计算需要
  - 增强模型的稳定性和准确度
  - 特定数据处理和分析的必要步骤
  - 模型结果应用和部署的需要
- 如何离散化
  - 时间数据离散化
  - 多值离散数据离散化
  - 连续数据离散化

空气质量数据

A	B	C	D	E	F	G	
No	year	month	day	hour	season	PM_Dongsi	PM
37546	2014	4	14	9	1	299	
37547	2014	4	14	10	1	299	
37548	2014	4	14	11	1	214	
37549	2014	4	14	12	1	280	
37550	2014	4	14	13	1	297	
37551	2014	4	14	14	1	277	
37552	2014	4	14	15	1	234	
37553	2014	4	14	16	1	177	
37554	2014	4	14	17	1	231	
37555	2014	4	14	18	1	245	
37556	2014	4	14	19	1	248	
37557	2014	4	14	20	1	275	
37558	2014	4	14	21	1	193	
37559	2014	4	14	22	1	80	
37560	2014	4	14	23	1	57	
37561	2014	4	15	0	1	50	

表 1　空气质量指数（AQI）分级相关信息

AQI 数值	AQI 级别	AQI 类别及表示颜色		对健康影响情况	建议采取的措施
0～50	一级	优	绿色	空气质量令人满意，基本无空气污染	各类人群可正常活动
51～100	二级	良	黄色	空气质量可接受，但某些污染物可能对极少数异常敏感人群健康有较弱影响	极少数异常敏感人群应减少户外活动
101～150	三级	轻度污染	橙色	易感人群症状有轻度加剧，健康人群出现刺激症状	儿童、老年人及心脏病、呼吸系统疾病患者应减少长时间、高强度的户外锻炼
151～200	四级	中度污染	红色	进一步加剧易感人群症状，可能对健康人群心脏、呼吸系统有影响	儿童、老年人及心脏病、呼吸系统疾病患者避免长时间、高强度的户外锻炼，一般人群适量减少户外运动
201～300	五级	重度污染	紫色	心脏病和肺病患者症状显著加剧，运动耐受力降低，健康人群普遍出现症状	儿童、老年人和心脏病、肺病患者应停留在室内，停止户外运动，一般人群减少户外运动
>300	六级	严重污染	褐红色	健康人运动耐受力降低，有明显强烈症状，提前出现某些疾病	儿童、老年人和病人应当停留在室内，避免体力消耗，一般人群应避免户外活动

cut方法：按值切割，根据数据值的大小范围分成n组，落入这个范围的分别进入到该组。

- 设定区间的个数，每个区间的间距相等
- 也可自定义每个区间的长度

```
pandas.cut(x, bins, right=True, labels=None, retbins=False, precision=3,
include_lowest=False, duplicates='raise')
```

x：数据集，这里一般是**pandas**的**Series**
bins：为一个整数或数组，代表切割成几组或者具体的切割方式
labels：代表切割后的分组名称
right：表示区间右端点的数据是否包含在内，默认为包含

qcut方法：按个数切割，使得每个区间里的元素个数基本相同

```
pandas.qcut(x, q, labels=None, retbins=False, precision=3,duplicates='raise')
```

x：数据集，这里一般是**pandas**的**Series**
q：为一个整数或分位数数组
labels：代表切割后的分组名称

cut方法：按值切割

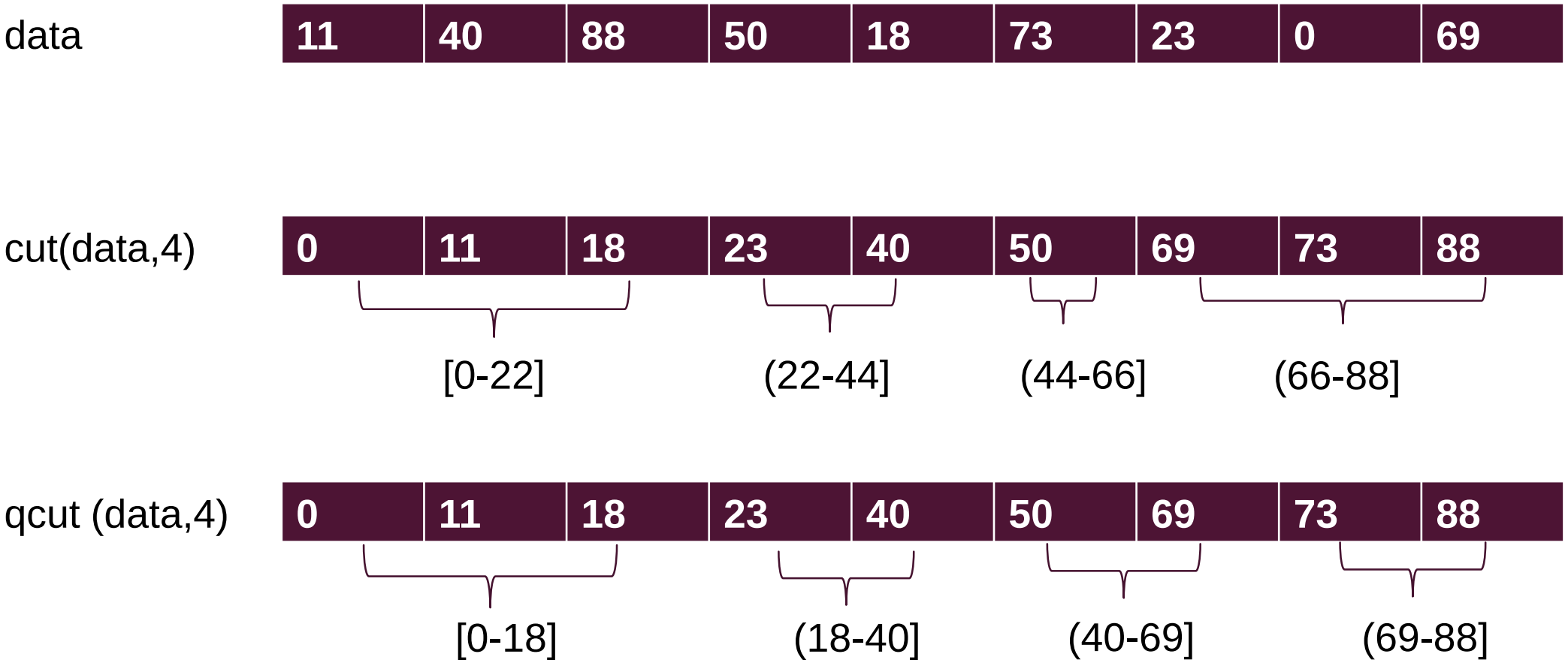
- 设定区间的个数，每个区间的间距相等
- 自定义每个区间的长度

3	1	2	9	5	10	6	4	0	8	7

5等份：　0-2-4-6-8-10 :　(0,1,2),(3,4),(5,6),(7,8),(9,10)

4等份：　0-2.5-5.0-7.5-10 :　(0,1,2),(3,4,5),(6,7),(8,9,10)

指定区间：　0,2,7,10 :　(1,2),(3,4,5,6,7),(8,9,10)

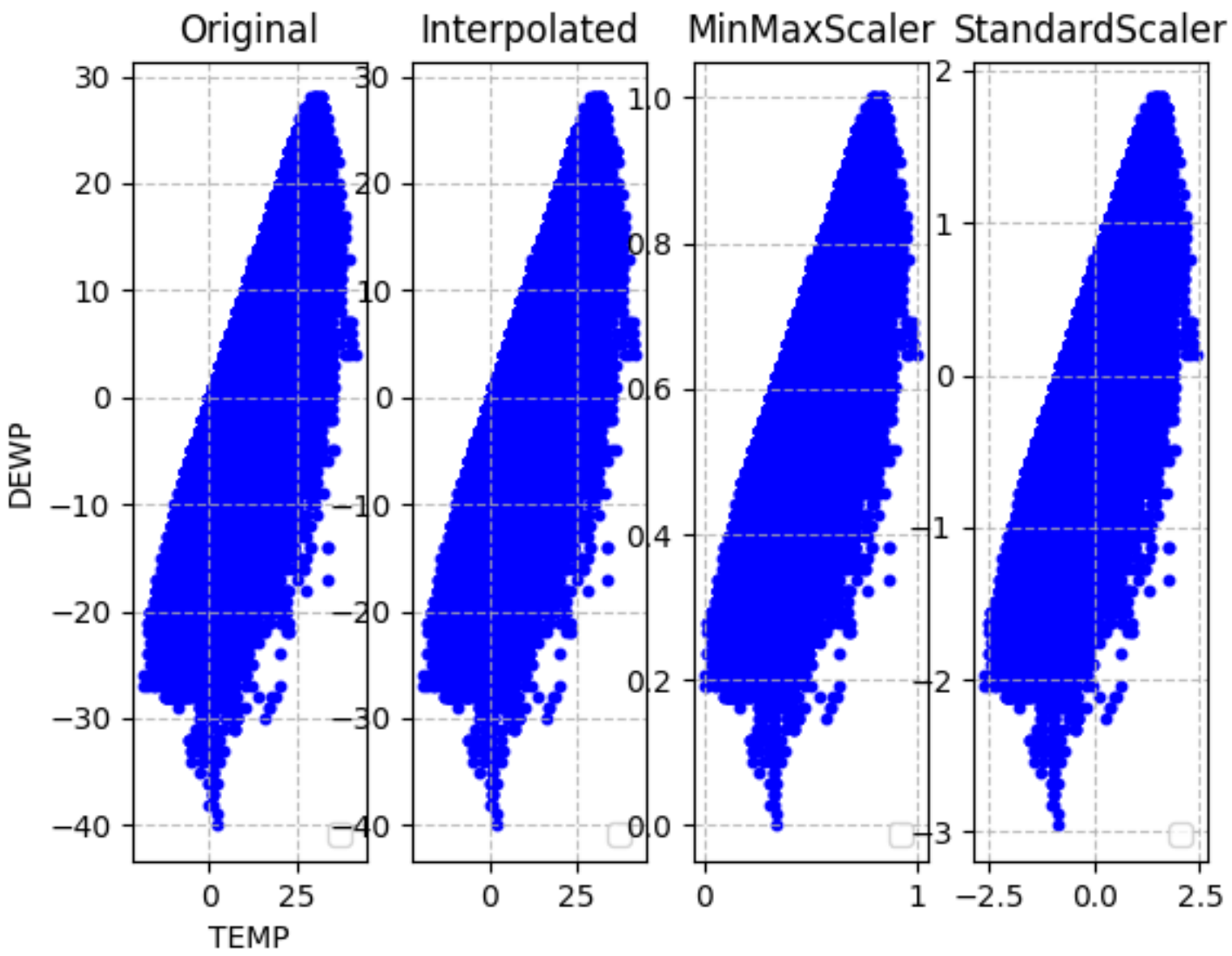

data
11
40
88
50
18
73
23
0
69
cut(data,4)
0
11
18
23
40
50
69
73
88
[0-22]
(22-44]
(44-66]
(66-88]
qcut (data,4)
0
11
18
23
40
50
69
73
88
[0-18]
(18-40]
(40-69]
(69-88]

```
sections = [0,50,100,150,200,300,1200]
#划分为不同长度 的区间
section_names=["green","yellow","orange","red","purple", "Brownish red"]
#设置每个区间的标签
result = pd.cut(df.ave,sections,labels=section_names)
print(pd.value_counts(result))
```

---------- result count----------
green 11587
yellow 8006
orange 3190
red 1518
purple 1121
Brownish red 463
Name: ave, dtype: int64

作业四

- 处理北京空气质量数据
  - 对DEWP和TEMP两列，进行0-1归一化及Z-Score归一化处理。 结果使用散点图的形式表示（参考PPT第19页图形上半部分的表现形式）。
  - 将北京的空气质量数据进行离散化，按照空气质量指数分级标准，计算出每个级别（或颜色值）对应的天数各有多少


Original
Interpolated
MinMaxScaler
StandardScaler
DEWP
TEMP
30
20
10
0
−10
−20
−30
−40
0
25
1.0
0.8
0.6
0.4
0.2
0.0
1
2
−1
−2
−3
−2.5
0.0
2.5

十 数据预处理

- 数据缺失值的处理
- 异常值的处理
- 数据归一化
- 数据连续属性离散化

谢谢