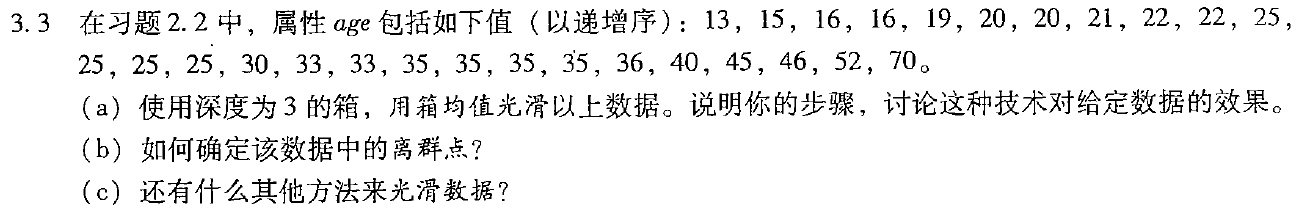
## Python数据分析与挖掘 - 数据预处理

**作业提交要求**：

* 直接用MS Word完成，保存为docx格式提交
* 提交文件名为“Python数据分析第6次作业+班级+姓名.docx” （注：请规范文件名）
* 提交到刘顿同学的邮箱 [810792334@qq.com](mailto:810792334@qq.com" \t "_blank)
* 提交的截止时间为2019年4月9日下午17:00

**（注：以下题目在“数据挖掘 概念与技术（第三版）”P79-80可以找到）**

****

**答：**

**（a）**

**步骤1：将数据划分到大小为3的等频箱中**

**1：13，15，16 2：16，19，20 3：20，21，22 4：22，25，25**

**5：25，25，30 6：33，33，35 7：35，35，35 8：36，40，45**

**9：46，52，70**

**步骤2：计算每个箱中的算数平均值**

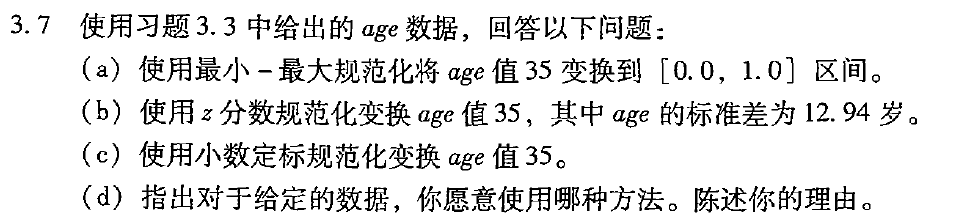
**步骤3：用算出的平均值替换箱中的原本值**

**（b）**

**我们可以使用聚类算法，比如常用的K-Means。因为聚类算法属于无监督学习，可以把相似的点归为一个簇，那么大多数数据应该为一个簇，而数据中的离群点就会在簇外面从而被确定。**

**（c）**

**比如分箱中的箱中位数光滑、箱边界光滑、箱均值光滑；当然我们也可以利用概念分层或者回归（利用函数拟合数据）来光滑数据，当然还有的分类方法有内置的光滑机制。**



答:

(a)

Min=13,max=17,因此利用最小-最大规范化为

(35-13/70-13)\*(1-0)+0=0.386

(b)

对于z分数规范化，给了我们标准差我们还需要去求数据的均值

均值=809/27=29.963

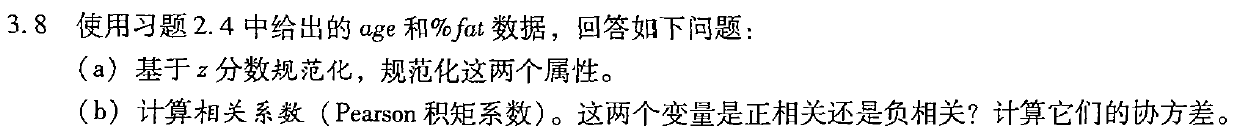
（35-29.963）/12.94=0.38926

（c）

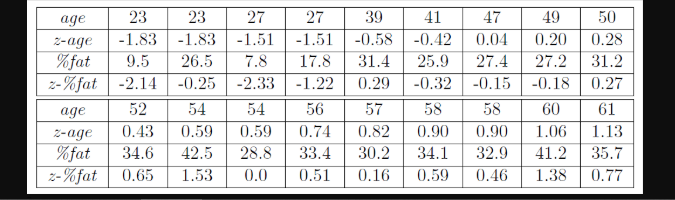
最大值为70，所以除以100，得到0.35

（d）

我愿意用z分数，因为将来添加更多的数据可以使用相同的规范化方法，而最大最小值和小数定标都收极值影响。



（a）

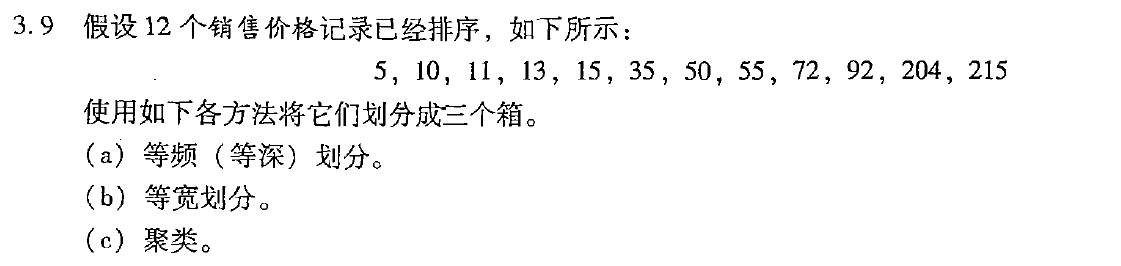


（b）

ra,b=（∑(aibi)-18\*46.44\*28.78）/18\*12.85\*8.99=0.82

0.82>0，所以这两个变量是正相关

Cov(a,b)=E(a,b)-（a的平均值\*b的平均值）=(23\*9.5+23\*26.5+27\*17.8+39\*31.4+41\*25.9+47\*27.4+49\*27.2+50\*31.2+52\*34.6+54\*42.5+54\*28.8+56\*33.4+57\*30.2+58\*34.1+58\*32.9+60\*41.2+61\*0.77)/18-46.44\*28.78=-35.3282‬



（a）

1. 5，10，11，13
2. 15，35，50，55
3. 72，92，204，215

（b）

1. 5,10,11,13,15,35,50,55,72
2. 92
3. 204,215

（c）

1. 5,10,11,13,15
2. 35,50,55,72,92
3. 204,215