



# 数据预处理：数据清洗等



覃雄派



# 提纲

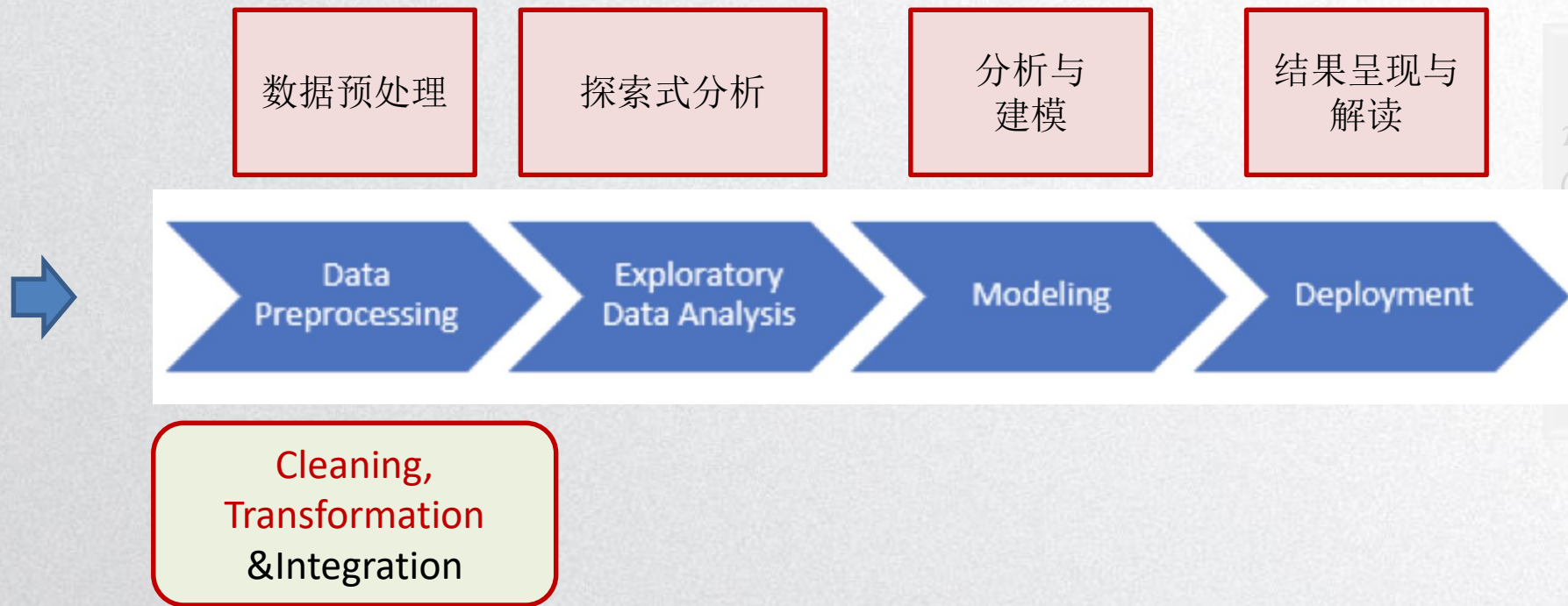


数据预处理：数据清洗  
等

- 为什么要进行数据清洗
- 如何进行数据清洗
  - 重复值
  - 缺失值
  - 异常值
- 编码和转换
- 规范化和缩放

# 数据预处理：数据清洗等

- 数据预处理：数据清洗等



# 数据预处理：数据清洗等





# 数据预处理：数据清洗等

- 为什么要做数据清洗与集成？
  - 现实世界中，数据通常是脏的 → Garbage In, Garbage Out
    - 数据存在错误和不一致



↓	C1	C2
	Total Cholesterol_1	Total Cholesterol_2
682	214.4	214.4
683	184.4	184.4
684	183.5	183.5
685	240.7	240.7
686	215.1	215.1
687	198.6	198.6
688	2800.0	280.0
689	210.8	210.8
690	182.5	182.5
691	192.6	192.6

# 数据预处理：数据清洗等

- 为什么要做数据清洗与集成？
  - 现实世界中，数据通常是脏的 → Garbage In, Garbage Out
    - 数据存在错误和不一致
    - 数据存在缺失 (Missing)

**Exhibit 2: Examples of variables that are set to unknown values**

**Administrative dates:** set to 0101YY, 010199, 999999

**Date of Birth** 0101YY, 1506YY, 3006YY, 0107YY, 1507YY, 0101YEAR

**Names:** set to spaces, NK, UNKNOWN, or ZZZZ  
BABY, MALE, FEMALE, TWIN, TRIPLET, INFANT

**Other variables:** set to 9, 99, 9999, -1  
NK (Not Known)  
NA (Not applicable)  
NC (Not coded)  
U (Unknown)

[Gill et al; Univ of Oxford 2003]



# 数据预处理：数据清洗等

- 为什么要做数据清洗与集成？
  - 现实世界中，数据通常是脏的 → **Garbage In, Garbage Out**
    - 数据存在错误和不一致
    - 数据存在缺失 (Missing)
    - 名称/属性的**二义性**

Michael  
Jordan



人大

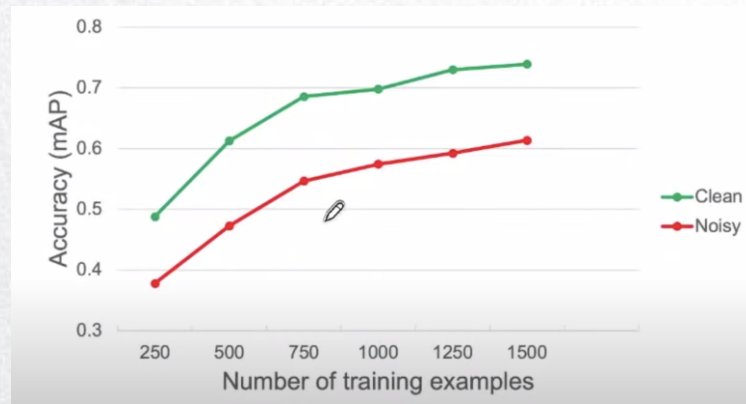
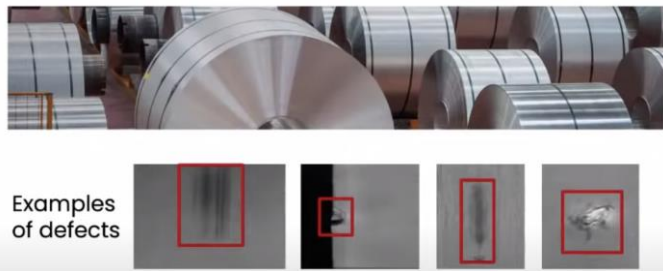
有个大叔想去人大告状  
结果出租车司机给拉到了人民大学  
从早上到现在，大叔一直呆在学校操场最高的探照  
灯上不肯下来  
论正确表达以及进门看牌子的重要性 😊😊





# From Big Data to Good Data

- 当数据科学方法
  - 利用计算机视觉算法检测钢板缺陷
  - 存在“脏数据”：12%被错误标注



	钢板缺陷检测	Solar Panel	Surface Inspection
基线方法	76.2%	75.68%	85.05%
Improving the <b>Code</b>	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Improving the <b>Data</b>	<b>+16.9%</b> (93.1%)	<b>+3.06%</b> (78.74%)	<b>+0.4%</b> (85.45%)

- 大数据 vs. 好数据
  - 实验表明：以下情况效果相同
  - 1250个“脏”样本
  - 500个“干净”样本

Andrew's talk:

<https://www.youtube.com/watch?v=06-AZXmwHjo>



# 数据预处理：数据清洗等





# 数据预处理：数据清洗等

- 数据清洗与集成的主要任务

- 将文本**拆分**成不同的属性 (Fields) → 解决分隔符问题

- 例：教师列表 Ju Fan: Associate Prof., Computer Science | 35

- 补充**缺失**的数据

- 例：如果Ju Fan的年龄信息缺失，应该如何填充呢？

- 平均值填充、用最近似教师年龄、贝叶斯估计

- 格式**转换**问题

- 日期的表示：20190329, 03/29/2019, 29/03/2019

- 异常值**检测**

- 例：Salary = -10; Age = 222

- 同一实体不同表示的**识别**

- 例：iPhone 2 vs iPhone 2<sup>nd</sup> generation



请写**Python**代码帮助  
范老师进行文本拆分，  
输出一个字符串数组

下一个PPT讲数据集成

# 数据预处理：数据清洗等

- 数据清洗与集成的主要任务

- 将文本**拆分**成不同的属性 (Fields) → 解决分隔符问题

- 例：教师列表 Ju Fan: Associate Prof., Computer Science | 35

- 补充**缺失**的数据

- 例：如果Ju Fan的年龄信息缺失，应该如何填充呢？

- 平均值填充、用最近似教师年龄、贝叶斯估计

- 格式**转换**问题

- 日期的表示：20190329, 03/29/2019, 29/03/2019

- 异常值**检测**

- 例：Salary = -10; Age = 22

- 同一实体不同表示的**识别**

- 例：iPhone 2 vs iPhone 2<sup>nd</sup> generation



请写**Python**代码帮助范老师进行文本拆分，输出一个字符串数组

```
info = 'Ju Fan: Associate Prof., Computer Science | 35'
fields = info.split(':')
print(fields)
fields = re.split('[:|]', info)
print(fields)
```

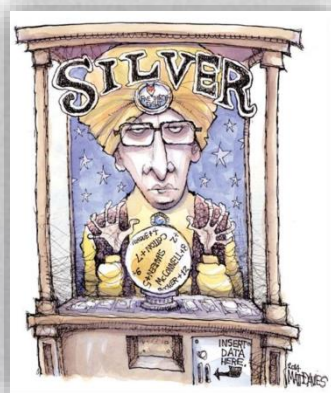
```
['Ju Fan', ' Associate Prof., Computer Science | 35']
['Ju Fan', ' Associate Prof.', ' Computer Science ', ' 35']
```

下一个PPT讲数据集成



# 数据预处理：数据清洗等

- 如何做数据清洗与集成
  - 查看数据描述
    - 检查数据量和数据属性
    - 属性的类型、**值域区间**、**关联**，数据的可获得性
    - 从**业务角度**理解数据属性和属性值的含义
    - 计算每个属性的**统计信息**（如最大值、最小值、均值、方差等）
  - 检测数据质量问题并修复
    - 检查数据值是否有**错误**
    - 有无**缺失值**
    - 有无**重复属性**
    - 检查数据值是否有**异常值**
    - 值和属性本身的含义是否符合



*New York Times*: 一个数据科项目通常**80%**的时间花在清洗和集成数据上，**20%**的时间花在实质数据分析上



# 数据预处理：数据清洗等





# 数据预处理：数据清洗等

- 数据清洗

近似的数据行的去重deduplicate, 参考下一个PPT数据集成

- 重复值duplicates: removing the duplicate
- 缺失值missing value : 删除整行记录(pandas dropna)/用均值、中位数填充numerical column/用众数填充category column(pandas fillna)
  - For categorical column, you can replace the missing values with mode values i.e the frequent ones
  - Predict Missing values with an ML Algorithm
- 异常值outliers
  - Using boxplot
  - it is advisable to check that the variable shouldn't have extreme values .i.e. outliers
  - Drop the outlier value/ Replace the outlier value using the IQR
  - Using Z-score : if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers

# 数据预处理：数据清洗等


- 数据清洗

- 缺失值及其处理详解

- 有数据的地方，通常也就有缺失的数据
      - 我们没有上帝视角
    - 数据科学，必须面对数据缺失的问题
      - 仪器故障
    - 因为与其他记录冲突被删除
    - 人为因素没有输入
    - 当初录入数据时数据项没有被重视
    - 数据变化没有被记录

- 缺失数据对分析建模有很大影响

忽略还是弥补，这是个问题

A large blue speech bubble pointing towards the left, containing the text '数据缺失的原因'.

数据缺失的原因



# 数据预处理：数据清洗等

- 数据清洗
  - 缺失值及其处理详解
    - 把整行记录删除，保留那些完整的记录
    - 把整个属性忽略掉，保留那些完整的属性
    - 人工填充某些值
      - 用一个全局值代替，例如-1， unknown
      - 使用对应属性的**平均值/中位数**
        - » 使用相同类标签样本对应属性上
        - » 的**平均值/中位数**
      - 用最大可能的值推理
        - » 如找**最相似的点**推理缺失值
        - » 使用贝叶斯或决策树**推理**
    - 缺失值填充：猜的越多，离真实数据越远



	missing	

	missing	

数据缺失的处理办法



# 数据预处理：数据清洗等

- 数据清洗
  - 处理缺失值

```
test_df.isnull().any()
```

PassengerId	False
Pclass	False
Name	False
Sex	False
Age	True
SibSp	False
Parch	False
Ticket	False
Fare	True
Cabin	True
Embarked	False

dtype: bool



请问你会如何处理  
**Age**缺失的记录？

- A. 将这些记录删除
- B. 使用平均年龄填充
- C. 使用年龄的中位数填充

# 数据预处理：数据清洗等

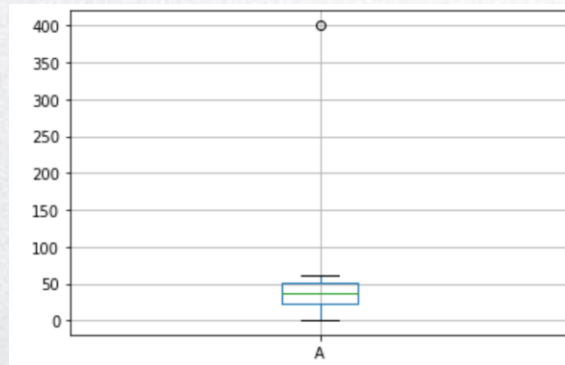
- 缺失值的处理

- 均值和中位数的差别
- 假设有如下数据列表

1 20 21 21 22 26 33 35 36 37 39 42 45 47 54 57 61 62 400



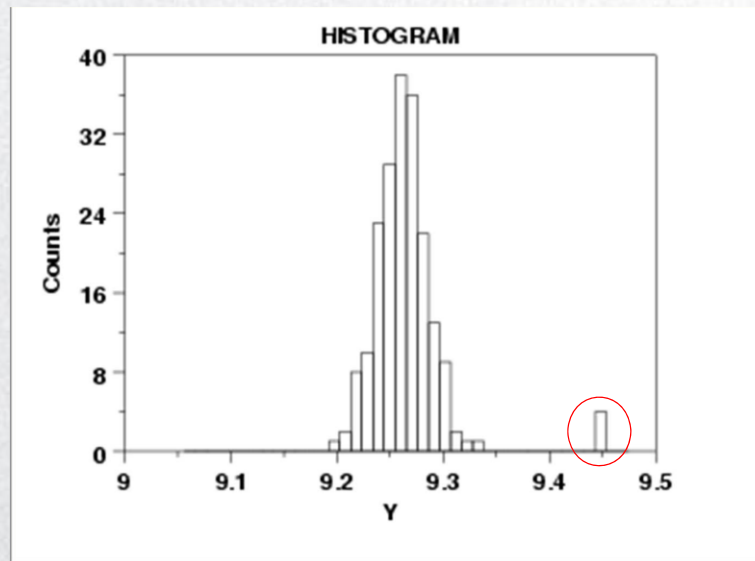
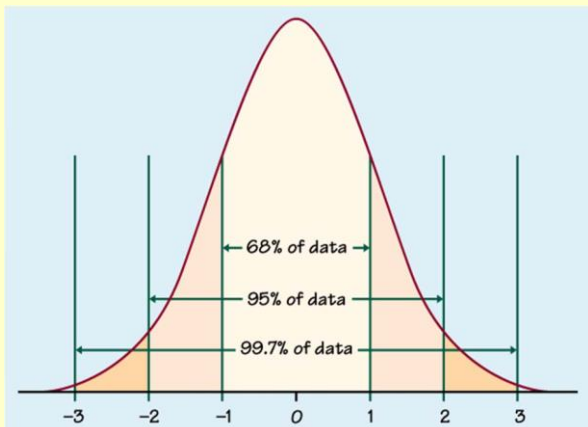
- 其均值为55.74
  - 该均值受到400的影响很大，400是个异常值
- 在这种情况下可以使用中位数
  - 中位数为37



# 数据预处理：数据清洗等

- 发现和处理异常值 (outlier)
  - 单变量
    - 统计方法  $\mu \pm 2\sigma$  或者  $\bar{x} \pm ks$

$$(\bar{x} - ks, \bar{x} + ks)$$



# 数据预处理：数据清洗等

- 发现和处理异常值 (outlier)

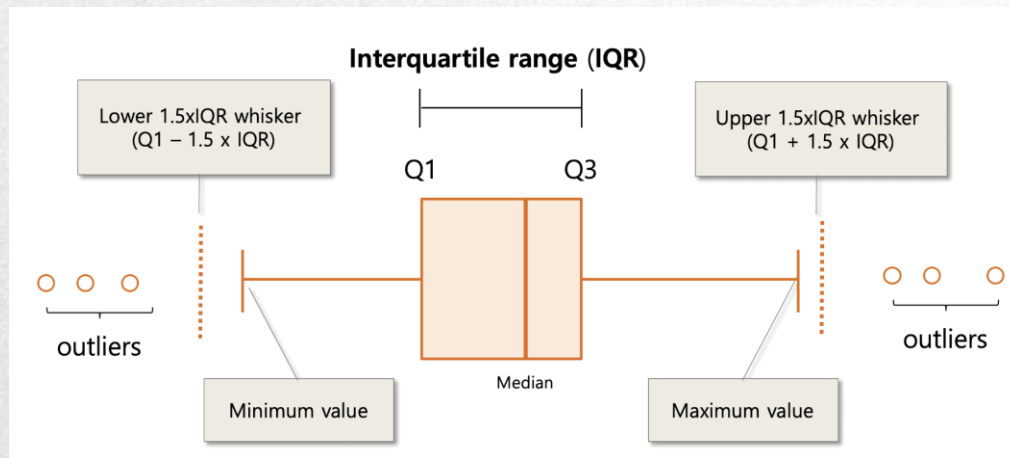
- 单变量

- 普通异常值

$(Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR)$

- 超级异常值

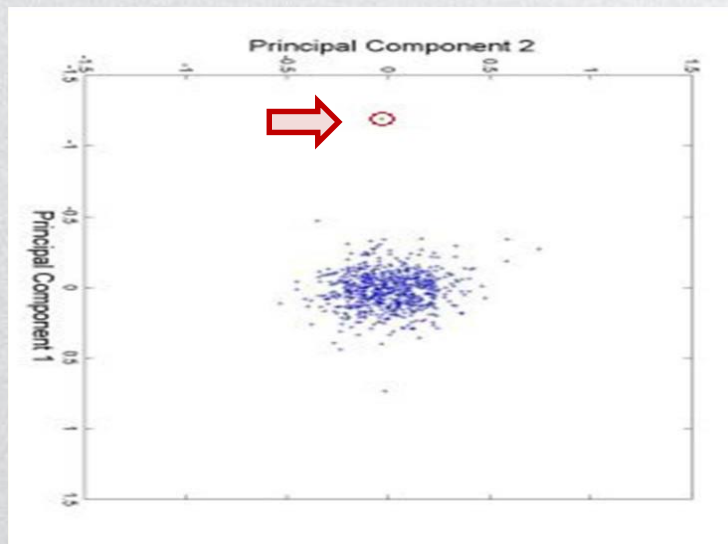
$(Q1 - 3 \times IQR, Q3 + 3 \times IQR)$ , where  $IQR = Q3 - Q1$



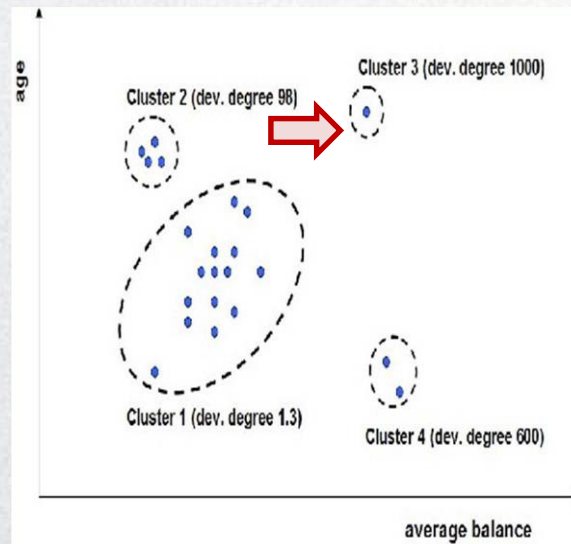


# 数据预处理：数据清洗等

- 发现和处理异常值 (Outlier)
  - 多变量
    - 数据降维与可视化

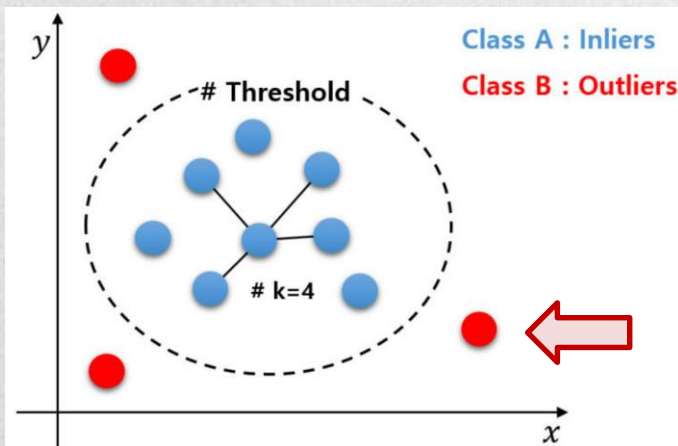


## 数据聚类与可视化

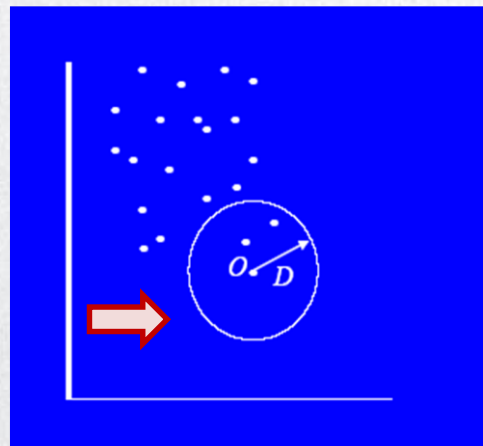


# 数据预处理：数据清洗等

- 发现和处理异常值 (Outlier)
  - 多变量
    - 基于距离如何发现异常值?
      - 不合群的点



低密度区域



# 数据预处理：数据清洗等

- 数据的编码和转换
  - 编码和转换
    - 对数据(一般是Categorical variables)进行适当编码
      - 比如把性别的男/女编码为1/0, 年龄编码为老中青 (2/1/0) 等
    - 对数据(一般是Numerical variables)进行分桶, 离散化处理
      - 把0-8000的工资, 按照1000元一档, 转换为0、1、....7
  - 数据的缩放、规范化、标准化(一般是Numerical Variables)
    - min max scaler缩放到[0,1]
    - standard scaler transform the data
      - using the formula  $(x - \text{mean}) / \text{standard deviation}$



# 数据预处理：数据清洗等

- 数据的编码和转换
  - 编码和转换的实例
    - ID属性比如用户ID
      - 一般保持不变
    - 多值属性处理
      - 颜色、国家、省/州等有限数值，可以用数值编号字典化处理数据
    - 属性值分组
      - 比如全国省份，按区域转换为东部、西部、北部、南部、中部等分组
    - 排序型转换为数值型
      - ABCDE，按照该规则转换A→4.0, B→3.3, C→2.9, D→2.1, E→1.7

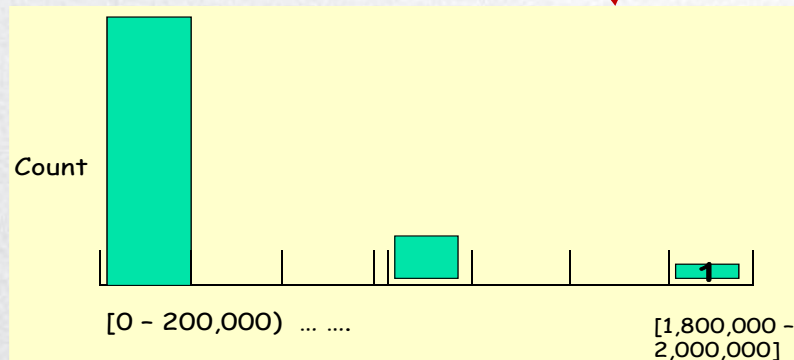
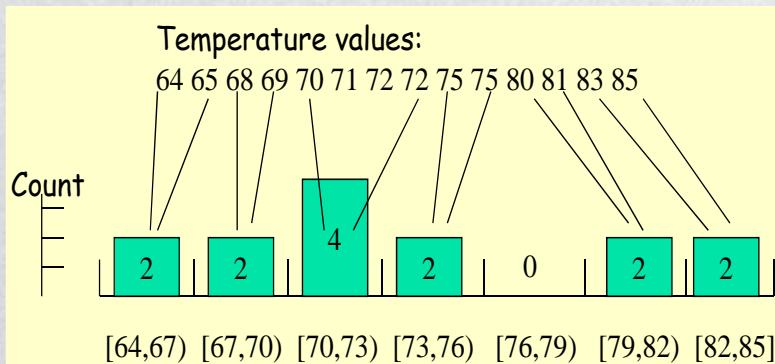


# 数据预处理：数据清洗等

## • 数据的编码和转换

### – 编码和转换的实例

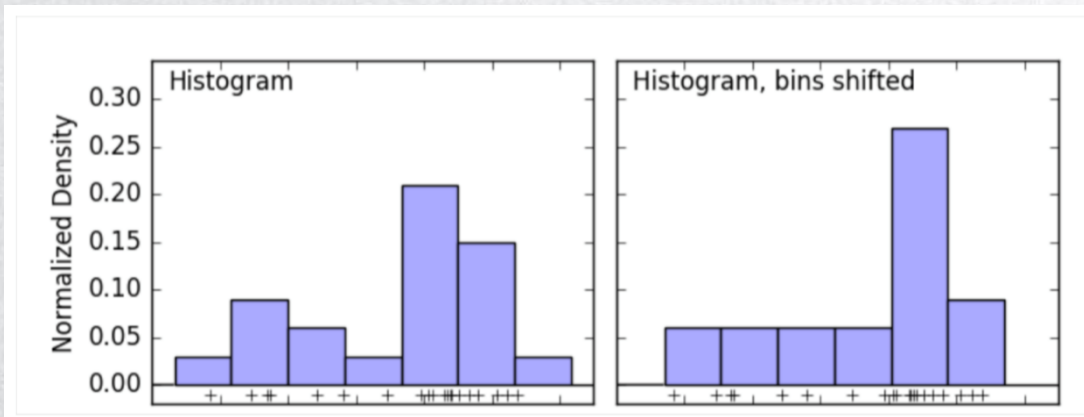
- 数值型变量的分桶 **Binning**
  - 可以减少数据量、压缩数据
- 等宽度划分离散化 **Equal-Width Binning**
  - 优点：简单、易懂
  - 缺点：分成多少个桶合适？受噪音影响大



# 数据预处理：数据清洗等

- 数据的编码和转换

- 直方图的局限性
- 如何分桶 (binning) 会带来很大的视觉差异



分桶的边界不同，  
可视化效果不同

Data is like people – interrogate it hard enough and it will tell you whatever you want to hear.

interrogate

英[ɪn'terəgeɪt] 美[ɪn'terəgeɪt]

v. 讯问; 审问; 盘问; (在计算机或其他机器上)查询, 询问;

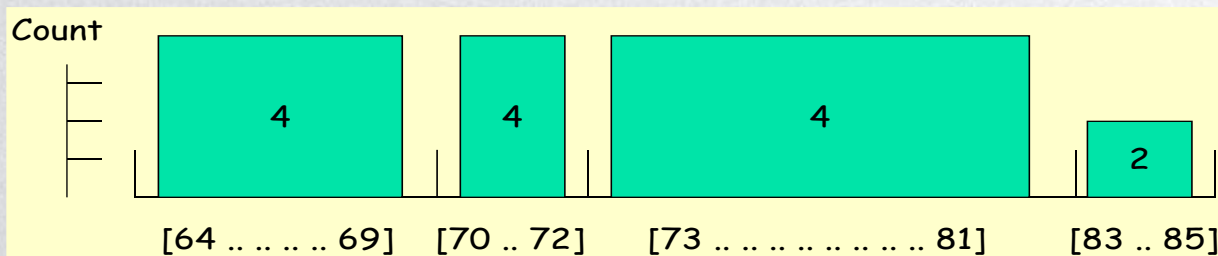


# 数据预处理：数据清洗等

- 数据的编码和转换

- 编码和转换的实例

- 数值型变量的分桶Binning
      - 可以减少数据量、压缩数据
    - 等宽度划分离散化Equal-Width Binning
      - 优点：简单、易懂
      - 缺点：分成多少个桶合适？受噪音影响大
    - 等高划分离散化 Equal-Depth Binning
      - 对比一下与EW的优缺点？





# 数据预处理：数据清洗等

- 数据的编码和转换

- 编码和转换的实例

- 数值型变量的分桶Binning
      - 可以减少数据量、压缩数据
    - 等宽度划分离散化Equal-Width Binning
      - 优点：简单、易懂
      - 缺点：分成多少个桶合适？受噪音影响大
    - 等高划分离散化 Equal-Depth Binning
      - 对比一下与EW的优缺点？

请将如下数据用 Equal-Width Binning和Equal-Depth Binning 离散化处理  
13,15,16,16,19,20,21,22,22,  
25,30,33,35,35,36,40,45

# 数据预处理：数据清洗等

## • 数据的编码和转换

### – 编码和转换的实例

- 数值型变量的分桶 **Binning**
  - 可以减少数据量、压缩数据
- 等宽度划分离散化 **Equal-Width Binning**
  - 优点：简单、易懂
  - 缺点：分成多少个桶合适？受噪音影响大
- 等高划分离散化 **Equal-Depth Binning**
  - 对比一下与EW的优缺点？

请将如下数据用 **Equal-Width Binning** 和 **Equal-Depth Binning** 离散化处理  
13,15,16,16,19,20,21,22,22,25,30,33,35,35,36,40,45



[13,21)	[21,29)	[29,37)	[37,45)	[45, 53)
13,15,16,16,19,20	21,22,22,25	30,33,35,35,36	40	45
6个	4个	5个	1个	1个

[13,20)	[20,30)	[30,40)	[40,...)
13,15,16,16,19	20,21,22,22,25	30,33,35,35,36	40,45
5个	5个	5个	2个

# 数据预处理：数据清洗等

- 数据的规范化和缩放

- Min max scaler

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Z-Score/Standard scaler

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

$\mu$  = Mean

$\sigma$  = Standard Deviation

- 十位数归一化

- j取值刚好让最大的数小于等于1

$$v' = \frac{v}{10^j}$$



# 数据预处理：数据清洗等

- 数据的规范化和缩放
  - 规范化的实例

Age	min-max (0-1)	z-score	dec. scaling
44	0.421	0.450	0.44
35	0.184	-0.450	0.35
34	0.158	-0.550	0.34
34	0.158	-0.550	0.34
39	0.289	-0.050	0.39
41	0.342	0.150	0.41
42	0.368	0.250	0.42
31	0.079	-0.849	0.31
28	0.000	-1.149	0.28
30	0.053	-0.949	0.3
38	0.263	-0.150	0.38
36	0.211	-0.350	0.36
42	0.368	0.250	0.42
35	0.184	-0.450	0.35
33	0.132	-0.649	0.33
45	0.447	0.550	0.45
34	0.158	-0.550	0.34
65	0.974	2.548	0.65
66	1.000	2.648	0.66
38	0.263	-0.150	0.38
28	minimum		
66	maximum		
39.50	avgerage		
10.01	standard deviation		

# 数据预处理：数据清洗等

