数据整合和数据清洗

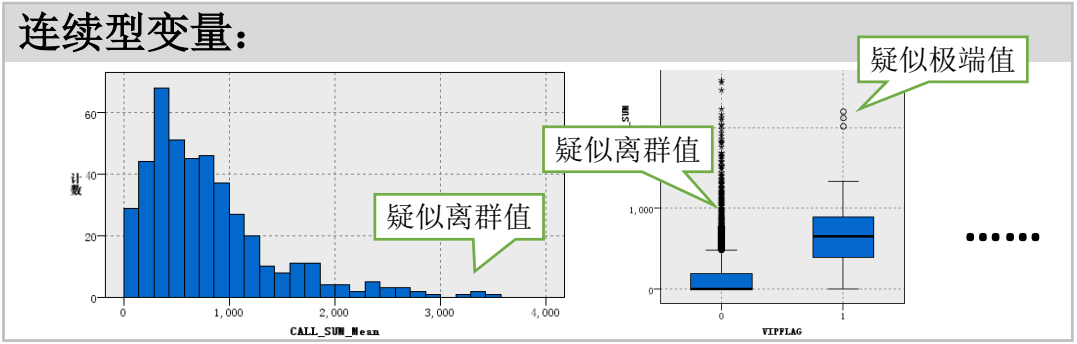
发现数据问题类型

- 脏数据或数据不正确
  - 比如 ‘0’ 代表真实的0，还是代表缺失；Age = -2003
- 数据不一致
  - 比如收入单位是万元，利润单位是元，或者一个单位是美元，一个是人民币
- 数据重复
  - 这个问题在前面已经解决
- 缺失值
- 离群值

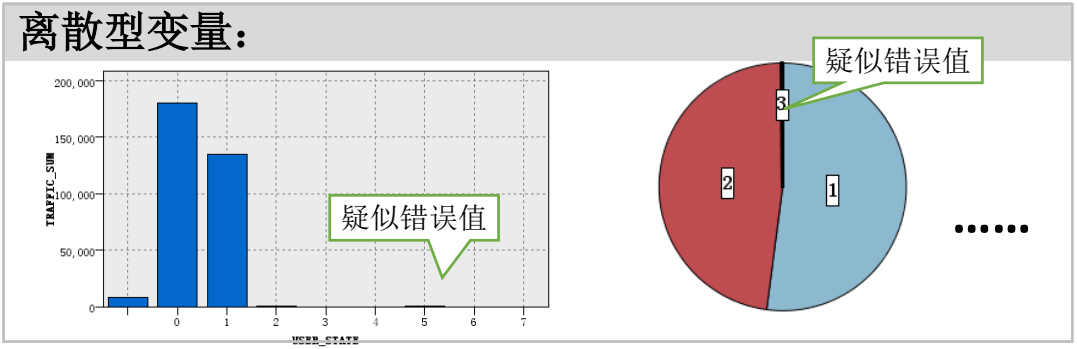
数据探索识别噪声

- 利用图形可以直观快速地对数据进行初步分析：
  - 直方图、饼图、条形图、折线图、散点图等

连续型变量：



离散型变量：



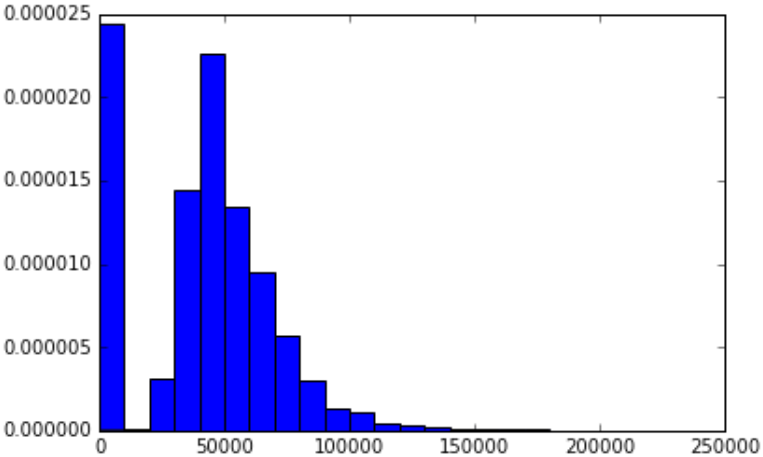
5.6 错误值处理

识别错误方法

通过图形进行探索

发现错误值只能通过描述性统计的方法，逐一核实每个变量是否有问题，比如‘0’代表真实的0，还是代表缺失？

外呼营销数据（teleco_camp_orig）的当地人均收入（AvgIncome），出现了大量0值，我们有理由怀疑是错误值。可以使用缺失值替代，然后再用缺失值填补的方法处理。



处理错误值

- 修正
  - 补充正确信息
  - 对照其他信息源
  - 视为空值
- 删除
  - 删除记录
  - 删除字段

5.6 缺失值处理

发现缺失值

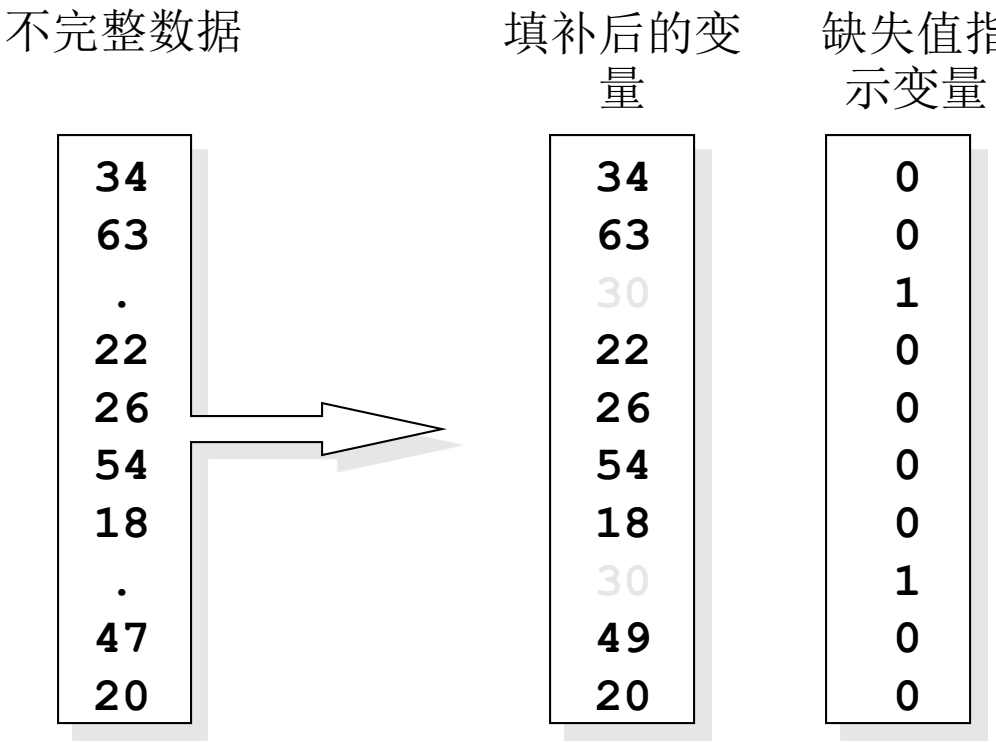
Class	Age	Gender	HomeOwner	AvgARPU	AvgHomeValue	AvgIncome
4	79	M	H	49.894904	33400	39460
3	71	M	H	48.574742	37600	33545
1	79	F	H	49.272646	100400	42091
1	63	F	H	47.334953	39900	39313
1	NaN	F	U	47.827404	47500	NaN
2	81	M	U	48.673449	53000	49487
2	NaN	F	U	48.560389	91000	NaN
3	69	F	H	49.644237	66300	49047

处理原则

首选基于业务的填补方法，其次根据单变量分析进行填补，
多重插补进行所有变量统一填补的方法只有在粗略清洗时才会使用。

- 缺失值少于20%
  - 连续变量使用均值或中位数填补。
  - 分类变量不需要填补，单算一类即可，或者用众数填补
- 缺失值在20%-50%
  - 填补方法同上
  - 另外每个有缺失值的变量生成一个指示哑变量，参与后续的建模
- 缺失值在大于50%
  - 每个有缺失值的变量生成一个指示哑变量，参与后续的建模，原始变量不使用。

填补 + 指示变量



Median = 30

处理缺失值例

ID	gender	年龄	教育程度	所在区域	峰值月时长	营销次数
1	0	22	3	郊区	1232	2
2	1	18	3	市区	522	2
3		45	3	市区	845	1
4	1		2		1321	1
5	0	15	2	市区	611	0
6	1	22	0		967	1
7	1	25	1	郊区	662	0
8				市区	10710	
9	0	27		市区	996	1
10	0	30	3	郊区	422	2
11	0	18	3		776	

对比其它数据源获取

连续变量可填均值/中位数

可填“未知”并增加指示变量

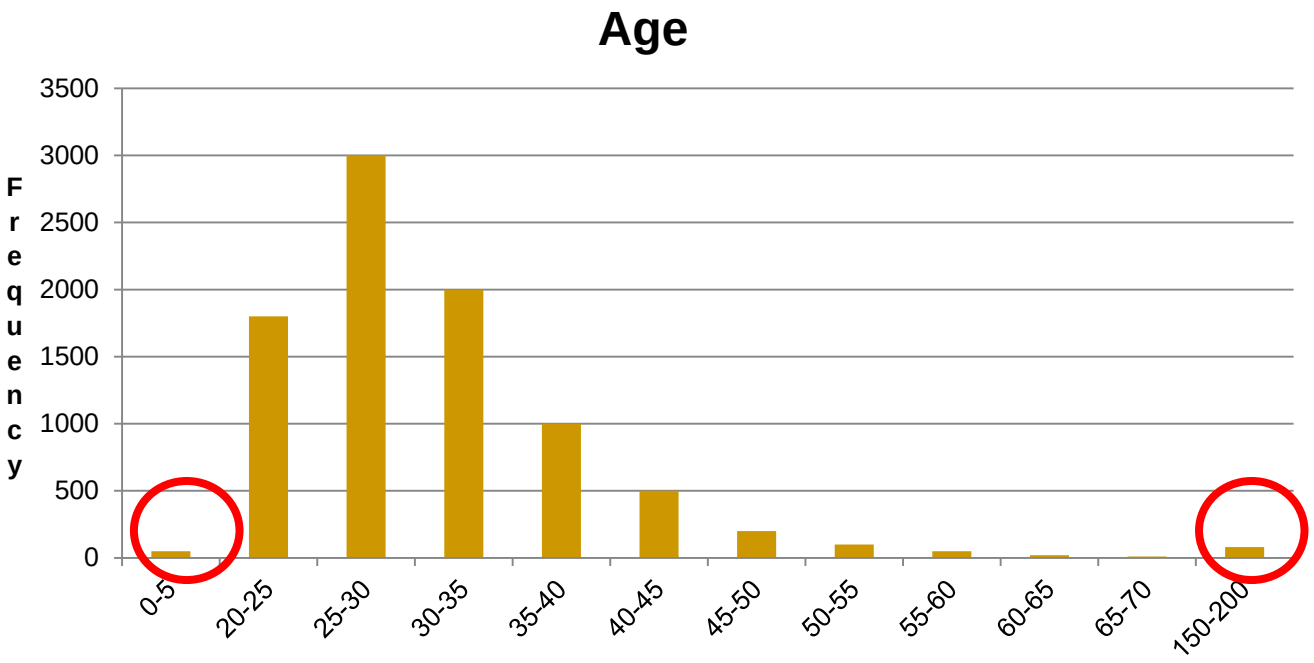
缺失过多直接删除

分类建模，聚类均值

离散变量可填众数

5.7 异常（离群）值处理

单变量离群值发现


Age
Frequency
3500
3000
2500
2000
1500
1000
500
0
0-5
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
150-200

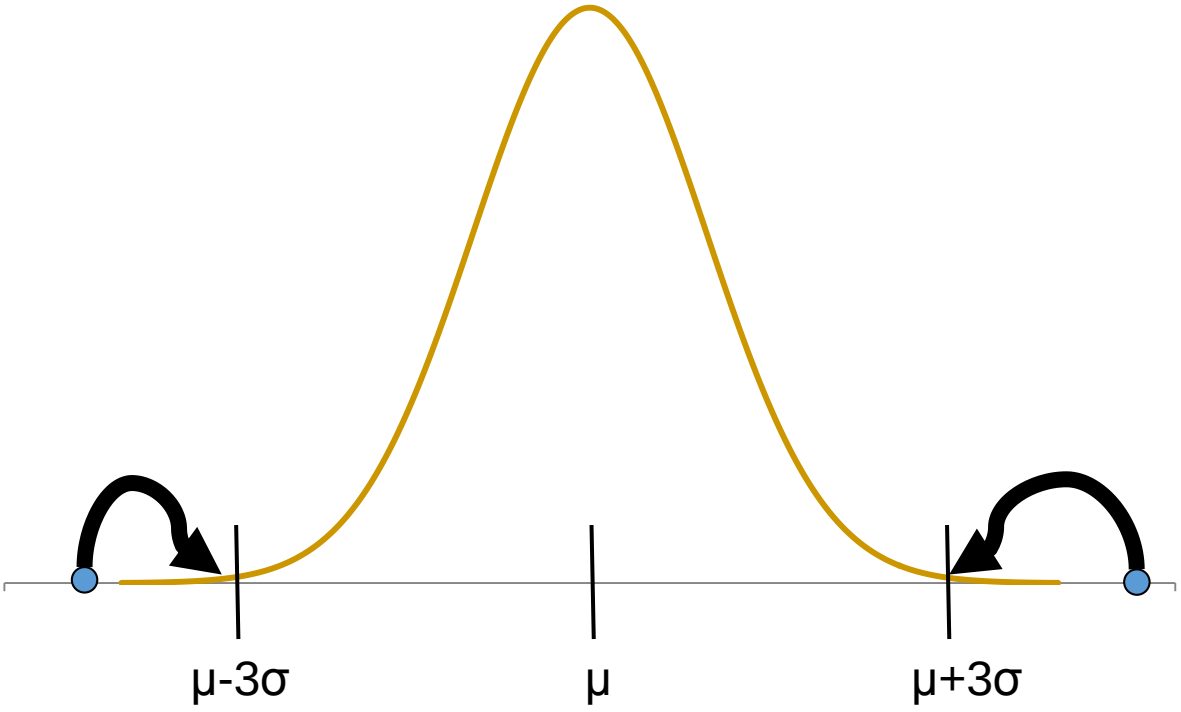
单变量离群值发现

- 极端值
  - 设置标准，如：5倍标准差之外的数据
  - 极值有时意味着错误，应重新理解数据，例如：特殊用户的超大额消费
- 离群值
  - 平均值法：平均值±n倍标准差之外的数据

    建议的临界值：
    - |SR| > 2 ，用于观察值较少的数据集
    - |SR| > 3 ，用于观察值较多的数据集
  - 四分位数法：
    - IQR = Q3 – Q1
    - $Q1 - 1.5 \times IQR \sim Q3 + 1.5 \times IQR$

* 更适用于对称分布的数据

盖帽法处理



分箱法

- 分箱方法通过考察数据的“近邻”来光滑有序数据的值。有序值分布到一些桶或箱中。
- 等深分箱：每个分箱中的样本量一致；
- 等宽分箱：每个分箱中的取值范围一致。

比如价格排序后数据：4, 8, 15, 21, 21, 24, 25, 28, 34

划分为（等深）箱：

- 箱1：4, 8, 15
- 箱2：21, 21, 24
- 箱3：25, 28, 34

划分为（等宽）箱：

- 箱1：4, 8
- 箱2：15, 21, 21, 24
- 箱3：25, 28, 34