# 第三章 数据探索

2023年7月19日 17:07

1. 数据质量分析
    a. 脏数据：缺失值、异常值、不一致的值、重复、有特殊符号的数据
    b. 缺失值处理方式：
        i. 统计分析：缺的个数、缺失率
        ii. 删除、插补、不处理
    c. 异常值（离散点）分析
        i. 描述统计量分析:max,min
        ii. 3σ原则
        iii. 箱线图
2. 数据特征分析
    a. 分布分析
        i. 定量：频率分布直方图、茎叶图、频率分布表
        ii. 定性：饼图、条形图
3. 统计量分析
    a. 集中趋势：（加权）均值/截断均值（去除极端值）；中位数；众数（多用于定性变量）
    b. 变异/离散：标准差/方差；四分位间距；极差；变异系数
4. 周期性分析
    a. 时间较长：年度/季节性 的周期性趋势
    b. 时间较短：月度/周度/天/小时 的周期性趋势
5. 贡献度分析
    a. 帕累托分析（20/80定律-同样的投入放在不同的地方会产生不同的效益）-重点改善盈利最高的产品。--绘制图形的时候可以添加箭头，指向85%的盈利点。
6. 相关性分析
    a. 直观：散点图/散点图矩阵
    b. 计算相关系数：
        i. 二元变量：Pearson相关系数（两个连续性变量~正态分布）-相关系数矩阵、spearman秩相关系数（等级相关系数）
            1) 这两种相关系数都需要进行“假设检验”，使用t检验检验其显著性水平以确定其相关程度
        ii. 判定系数：是相关系数的平法r2，衡量回归方程对y的解释程度。
7. python主要数据探索函数（数据分析：pandas；数据可视化：matplotlib）

a. 基本统计特征函数
    i. 反映数据的整体分布：均值、方差、标准差、相关系数
b. 统计绘图函数
    i. 直观反映数据以及统计量的性质和内在规律