# 数据处理

## 数据分析与处理步骤

明确思路和目的🡪数据收集🡪数据预处理🡪数据分析🡪数据展现

## 数据选取、分组或排序

从已有的大部分批量数据按照一定的规则抽取部分数据分析、分组讨论或者排序探究的过程

1. 化繁为简
2. 方便处理数据的缺失值、异常值
3. 有利于多角度对数据进行分析，从而客服思维片面化

## 数据缺失值处理

1. 删除：数量较少，对整体数据没什么影响的直接删除该条数据
2. 插补法：方法较多，例如用该属性的均值、众数、中位数插补或者使用回归法，插值拟合法进行插补

使用SPSS软件进行插补：

分析🡪描述统计🡪频率

转换🡪替换缺失值

## 数据异常处理

判别方法：

1. 简单统计分析
2. 3σ原则
3. 箱型图分析

处理方法：

1. 删除
2. 视为缺失值
3. 不处理

简单统计分析：

SPSS🡪描述统计🡪统计🡪最大值，最小值，根据资料辨别

3σ原则:

🡪0.6826

🡪0.9544

🡪0.9974

SPSS🡪分析🡪描述统计🡪频率🡪图标🡪直方图(正态曲线)

箱型图分析：

将数据从小到大排序，将75%分位点对应的值为**上四分位数Q3，**将50%分位点对应的值为**中位数Q2，**将25%分位点对应的值为**下四分位数Q1，上须max=Q3+1.5\*(Q3-Q1)，下须min=Q1-1.5\*(Q3-Q1)，**上须和下须之外就是**异常值**

SPSS🡪分析🡪描述统计🡪探索🡪图🡪因子级别并置，含检验的正态图

## 数据变换

归一化

z-score标准化

数据规约

对减少过多的属性（维度、特征）对预测的模型的影响

降维方法：

1. 合并属性
2. 主成分分析
3. 因子分析

# 数据分析

分类模型：

1. 决策树
2. 朴素贝叶斯分类
3. 支持向量机
4. 最近邻分类

预测类：

1. 朴素估计法
2. 时间序列模型
3. 神经网络预测模型
4. 马尔可夫模型
5. 灰色预测

综合评价：

1. 层次分析
2. 灰色关联度
3. 熵权法
4. TOPSIS

聚类分析：

1. 二阶聚类
2. K-Means聚类
3. 系统聚类