属性类型总结 (Attribute Types)

Nominal (名义型)

- 用于表示类别、标签或代码。
- 值没有意义上的顺序，仅是枚举。
- 示例：头发颜色 (hair color) = {red, black, grey, brown}，婚姻状况 (marital status) = {single, married, divorced}。
- 不支持计算操作，因为不是定量的。
- 常用统计：众数 (Mode)，即最常出现的值。

Binary (二元型)

- 特殊的名义型属性，仅有两个状态。
- 分类：
  - 对称型 (Symmetric Binary)：两个状态同等重要，例如性别 (gender)。
  - 非对称型 (Asymmetric Binary)：两个状态重要性不同，例如医疗测试结果 (medical test results)。
- 通常用 {1, 0} 表示存在/不存在或正/负。

Ordinal (序数型)

- 值具有有意义的顺序或排名，但相邻值之间的差距未知。
- 示例：饮料大小 (drink size) = {small, medium, large}，教育水平 (education level) = {high school, bachelor, master}。
- 应用：用于主观评估，例如问卷调查中的满意度评分。

Numeric (数值型)

- 提供对象的定量测量。
- 分类：
  - Interval-Scaled (区间型)：值有顺序且差值有意义，但没有绝对零点。例如温度 (temperature)。
  - Ratio-Scaled (比率型)：值有顺序且差值和比率均有意义，且有绝对零点。例如重量 (weight)、金钱 (money)。
- 示例：
  - 区间型：年份 (year) = {2010, 2020}，差值有意义但比率无意义。
  - 比率型：收入 (income) = {$10,000$,20,000}，差值和比率均有意义。

Discrete vs. Continuous (离散型 vs. 连续型)

Discrete (离散型)

- 属性值是有限的或可数的无限集合。
- 示例：
  - 名义型 (Nominal)：婚姻状况 (marital status) = {single, married, divorced}。
  - 二元型 (Binary)：医疗测试 (medical test) = {positive, negative}。
  - 序数型 (Ordinal)：职业等级 (profession levels) = {assistant, associate, full}。
  - 数值型 (Numeric)：鞋码 (shoe size)、分数 (scores)、售出商品数量 (# items sold)。

- 通常表示为整数变量，例如年龄范围 [0, 100]。

Continuous (连续型)

- 属性值可以是范围内的实数，包括小数或分数。
- 示例：
  - 温度 (temperature)、身高 (height)、体重 (weight)、时间 (time)、金钱 (money)。
- 值是不可数的无限集合，例如在 50kg 和 51kg 之间可以有 50.1, 50.2 等。
- 通常存储为浮点数变量 (float, double) 以提高精度。

Basic Statistical Descriptions of Data (数据的基本统计描述)

1. Central Tendency (集中趋势)

- 描述数据的中心或中间值。
- 常用指标：
  - **Mean (均值)**: 数据的平均值，易受极端值影响。
  - **Median (中位数)**: 排序后数据的中间值，适合偏态 (skewed) 分布。
  - **Mode (众数)**: 数据中出现频率最高的值。

2. Data Dispersion (数据分散性)

- 描述数据的分布范围和离散程度。
- 常用指标：
  - **Range (极差)**: 最大值与最小值的差。
  - **Interquartile Range (IQR, 四分位距)**: 中间 50% 数据的范围 (Q3 - Q1)。
  - **Variance (方差)**: 数据偏离均值的平方平均值。
  - **Standard Deviation (标准差)**: 方差的平方根，表示数据的离散程度。

3. Graphic Displays (图形化展示)

- 用于直观地观察数据分布和特性。
- 常见图表：
  - **Bar Charts (柱状图)**: 适合分类数据。
  - **Pie Charts (饼图)**: 显示比例。
  - **Histograms (直方图)**: 展示频率分布。
  - **Boxplots (箱线图)**: 显示数据的五数概括 (最小值, Q1, 中位数, Q3, 最大值) 和异常值。
  - **Line Graphs (折线图)**: 适合时间序列数据。

Histogram vs. Bar Chart (直方图 vs. 柱状图)

特性	Histogram (直方图)	Bar Chart (柱状图)
目的	展示数值数据的分布	比较不同类别或离散组的数据
数据类型	数值型数据，通常是连续型	分类数据或离散型数据
X轴	分组的数值范围（区间）	离散的类别

特性	Histogram (直方图)	Bar Chart (柱状图)
Y轴	每个区间的频率或计数	每个类别的频率、计数或其他统计值
条形间隔	条形相连，表示数据的连续性	条形分隔，表示类别的独立性
应用场景	分析数据分布形状、集中趋势和离散性	比较类别之间的大小差异，展示趋势
示例	温度分布、收入分布	不同产品的销售量、不同地区的人口数量

利用 Scatter Plots 展示 Data Correlation

- **Scatter Plots (散点图)**: 用于图形化展示两个数值型属性之间的关系，每个点代表数据集中的一个观测值。
- **目的**: 直观地观察两个变量之间的关系、模式或趋势。

使用步骤:

1. **选择两个数值型属性**:
   - 选择需要分析的两个属性，例如收入 (income) 和信用额度 (credit limit)，汽车年龄 (age of car) 和转售价格 (resale price)。
2. **绘制数据点**:
   - 在散点图中，每个点的横坐标 (x-axis) 表示一个属性，纵坐标 (y-axis) 表示另一个属性。
3. **观察模式**:
   - **正相关 (Positive Correlation)**: 数据点呈上升趋势，例如收入增加时信用额度增加。
   - **负相关 (Negative Correlation)**: 数据点呈下降趋势，例如汽车年龄增加时转售价格降低。
   - **无相关 (No Correlation)**: 数据点分布随机，没有明显的趋势。
4. **识别异常值 (Outliers)**:
   - 观察偏离主要模式的点，这些点可能是异常值。

应用场景:

- Scatter Plots 广泛用于**探索性数据分析 (Exploratory Data Analysis, EDA)**，例如：
  - 检测变量之间的关系。
  - 识别数据中的聚类 (clusters) 或分组。
  - 突出潜在的异常值以供进一步分析。

Statistical Measures (统计指标)

1. Mean (均值)

- $\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}$ 其中 x_i 是数据值，n 是数据点的数量。
- **统计意义**: 数据的中心值，表示所有数据的平均水平。
- **应用场景**: 用于衡量总体趋势，例如收入、温度等。易受极端值影响。

2. Median (中位数)

- **公式**: 如果数据点数量为奇数，中位数是排序后中间的值。
  - 如果数据点数量为偶数，中位数是排序后中间两个值的平均值。

- **统计意义**: 数据的中间值，适合偏态 (skewed) 分布的数据。
- **应用场景**: 用于描述不对称分布的数据，例如房价、收入分布。

3. Midrange (极值平均)

- **公式**: $\text{Midrange} = \frac{\text{Max}+\text{Min}}{2}$ 其中 Max 和 Min 分别是数据的最大值和最小值。
- **统计意义**: 数据范围的中心点，易受极端值影响。
- **应用场景**: 用于快速估计数据中心，但不常用于正式分析。

4. Variance (方差)

- $\text{Variance} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$ 其中 μ 是均值。
- **统计意义**: 衡量数据的离散程度，表示数据点偏离均值的平方平均值。
- **应用场景**: 用于风险评估（如金融领域）或数据分布分析。

5. Standard Deviation (标准差)

- $\text{Standard Deviation} = \sqrt{\text{Variance}}$
- **统计意义**: 数据离散程度的直观度量，单位与原始数据一致。
- **应用场景**: 用于评估数据波动性，例如股票价格波动、实验数据的精确度。

6. Quantile (分位数)

- **定义**: 将数据分为若干等份的点，表示数据分布的百分比位置。常见分位数包括：
  - 四分位数 (Quartiles): 将数据分为四等份，分别为 Q1 (25%)、Q2 (50%，即中位数)、Q3 (75%)。
  - 百分位数 (Percentiles): 将数据分为 100 等份，例如 P90 表示前 90% 的数据。
- **统计意义**: 帮助理解数据的分布特性，识别数据的集中趋势和分散程度。
- **应用场景**: 用于描述数据分布，例如考试成绩排名、收入分布等。

7. Interquartile Range (IQR, 四分位距)

- **定义**: IQR = Q3 - Q1，即数据中间 50% 的范围。
- **统计意义**: 衡量数据的分散程度，减少极端值的影响，比极差更稳健。
- **应用场景**: 用于检测异常值 (outliers)，例如箱线图 (Boxplot) 中的应用。

8. Symmetric vs. Positively-Skewed vs. Negatively-Skewed (对称分布 vs. 正偏态 vs. 负偏态)

特性	**对称分布 (Symmetric)**	**正偏态分布 (Positively-Skewed)**	**负偏态分布 (Negatively-Skewed)**
分布形状	左右对称	长尾在右侧	长尾在左侧
均值、中位数、众数关系	均值 = 中位数 = 众数	均值 > 中位数 > 众数	均值 < 中位数 < 众数
应用场景	身高、考试成绩	收入分布、房价分布	老龄化人口寿命分布