

# Big Data Analytics

# 大数据分析

**# 05: In-Memory Analytics with Pandas. Exploratory Data Analysis**  
**#05:用 Pandas 进行内存分析 探索性数据分析**

Instructor: Oleh Tymchuk  
授课教师：奥勒·特姆恰克

## #05: Agenda 课程安排

- Introduction to EDA 探索性数据分析（EDA） 介绍
- Summary statistics 统计概要
- Practical cases 实际案例
- Useful Links 实用链接

# Introduction to EDA

## 探索性数据分析（EDA）介绍

# What is EDA? 什么是EDA?

- Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their key characteristics
- It helps in understanding the structure, distribution, and relationships within the data
- EDA allows us to identify patterns, anomalies, missing values, and outliers before applying machine learning models
- 探索性数据分析 (EDA) 是一种分析数据集以总结其关键特征的方法。
- 它有助于理解数据的结构、分布和关系。
- EDA 使我们能够在应用机器学习模型之前识别模式、异常值、缺失值和离群点。

# Why is EDA important? EDA为什么重要？

- Understand data structure
  - Identify relationships between variables
  - Prepare data for modeling
- 
- 理解数据结构
  - 识别变量之间的关系
  - 准备建模数据

# Types of EDA techniques EDA 技术的类型

- Univariate analysis (examining single variables, e.g., histograms, box plots)
- Bivariate analysis (exploring relationships between two variables, e.g., scatter plots, correlation analysis)
- Multivariate analysis (analyzing more than two variables, e.g., heatmaps, pair plots)
- 单变量分析 ( 考察单个变量 , 例如直方图、箱线图 )
- 双变量分析 ( 探索两个变量之间的关系 , 例如散点图、相关性分析 )
- 多变量分析 ( 分析两个以上的变量 , 例如热力图、配对图 )

# Tools for EDA EDA 工具

- Pandas: Data manipulation and analysis
- NumPy: Numerical computations
- Matplotlib/Seaborn/Plotly: Data visualization

- Pandas : 数据操作与分析
- NumPy : 数值计算
- Matplotlib/Seaborn/Plotly : 数据可视化

# Steps in EDA EDA 步骤

[ x ] Understanding the dataset 理解数据集

[ x ] Handling missing values 处理缺失值

[ x ] Checking data types and conversions 检查数据类型和转换

[ - ] Summary statistics 统计概要

[ - ] Data visualization 数据可视化

[ - ] Identifying outliers and anomalies 识别离群点和异常值

## Scales 尺度

# Nominal Scale

Definition: categorical data **without order**.

Used to classify objects into distinct groups.

Examples:

- Colors (red, blue, green)
- Product types (smartphones, laptops)
- Countries (Ukraine, Germany, Japan)
- Gender (male, female)

Key Features:

- No mathematical meaning in values
- Only **frequency** or **mode** (most frequent category) can be calculated
- Visualization: pie charts, bar plots

翻译见下页

Please see next page for translation

# 名义尺度

定义：**无序**的分类数据。

用于将对象分类为不同的组。

示例:

- 颜色 (红色, 蓝色, 绿色)
- 产品类型 ( 智能手机、笔记本电脑 )
- 国家/地区 (乌克兰、德国、日本 )
- 性别 (男性、女性)

关键特征:

- 数值无数学含义
- 只能计算**频率或众数** (最常见的类别 )
- 可视化：饼图、条形图

# Ordinal Scale

Definition: categorical data **with order**, but intervals between values are not equal or measurable.

Examples:

- Education level (primary < secondary < tertiary)
- Product ratings (1 ★ < 2 ★ < 5 ★)
- Disease stages (mild < moderate < severe)
- Income levels (low, medium, high)

Key Features:

- Order matters, but differences between values are not quantified.
- **Median** and **ranks** are appropriate statistics.
- Visualization: ordered bar plots, Likert scales.

翻译见下页

Please see next page for translation

# 序数尺度

定义：**有序的分类数据**，但值之间的间隔不相等或不可测量。

示例:

- 教育程度 (小学<中学<大学)
- 产品评级 (1 ★ < 2 ★ < 5 ★)
- 疾病阶段 (轻度<中度<重度)
- 收入水平 (低、中、高)

主要特征:

- 顺序很重要，但值之间的差异是无法量化的。
- **中位数和秩次**是合适的统计数据。
- 可视化：有序条形图、李克特量表。

# Quantitative Scale

Definition: numerical data **with mathematical meaning**. Divided into two subtypes:

- Discrete (integers): number of products, children in a family.
- Continuous (decimal numbers): weight, height, temperature.

Examples:

- Age (25 years, 30.5 years)
- Salary (\$50,000)
- Delivery time (2.5 hours)
- Website views (1,000,000)

Key Features:

- All mathematical operations (+, −, ×, ÷) apply.
- Use **mean, standard deviation, variance**.
- Visualization: histograms, box plots, scatter plots.

翻译见下页

Please see next page for translation

# 定量尺度

定义：**具有数学意义的数值数据**。分为两种子类型：

- 离散（整数）：产品数量、家庭中的孩子数量。
- 连续（十进制数）：体重、身高、体温。

示例：

- 年龄（25岁，30.5岁）
- 工资（5万美元）
- 配送时间（2.5小时）
- 网站浏览量（100万）

重要特征：

- 所有数学运算（+、-、 $\times$ 、 $\div$ ）均适用。
- 使用平均值、标准差、方差。
- 可视化：直方图、箱线图、散点图。

# Comparison of Scales

Criterion	Nominal	Ordinal	Quantitative
Order	✗ No	✓ Yes	✓ Yes
Equal Intervals	✗ No	✗ No	✓ Yes
Math Operations	✗ Not meaningful	✗ Limited (on ranks only)	✓ All arithmetic operations allowed
Statistics	Mode, frequency	Median, mode, rank order	Mean, median, mode, variance, standard deviation
Example	Gender, colors, country names	Product ratings, education levels	Weight, height, temperature, age

## Important Notes:

- Common Mistake: Calculating the mean for ordinal data (e.g., "average rating 3.8" is technically incorrect).
- Rule: Statistical methods and visualizations depend on the scale type. Always validate assumptions before analysis.

翻译见下页 Please see next page for translation

# 尺度比较

标准	名义尺度	序数尺度	定量尺度
顺序	✗ 无	✓ 有	✓ 有
等距	✗ 无	✗ 无	✓ 有
数学运算	✗ 无意义	✗ 有限（仅秩次）	✓ 允许所有算术运算
统计	众数、频率	中位数、众数、秩次	均值、中位数、众数、方差、标准差
示例	性别、颜色、国家名称	产品评级、教育水平	体重、身高、温度、年龄

## 重要说明：

- 常见错误：对序数数据计算均值（例如，“平均评分3.8”在技术上是不正确的）。
- 规则：统计方法和可视化取决于尺度类型。在分析前始终验证假设。

# Summary statistics

## 统计概要

# Summary Statistics

## Concept:

- Summary statistics are a subset of descriptive statistics that provide a concise overview of the data
- They summarize key characteristics of the dataset using numerical metrics

## Why is it important?

- Helps us understand the overall structure of the data
- Identifies patterns, trends, and potential issues (e.g., outliers, missing data)
- Provides a foundation for further analysis or modeling

翻译见下页 Please see next page for translation

## 概念:

- 汇总统计数据是描述性统计数据的一个子集，它提供了数据的简明概述
- 他们使用数值指标总结数据集的关键特征

## 为何它如此重要？

- 帮助我们理解数据的整体结构
- 识别模式、趋势和潜在问题（例如离群点、缺失数据）
- 为进一步分析或建模提供基础

# Central Tendency. Mean

- **Calculation:**

ages = [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]

$$\text{Mean} = \frac{\sum x_i}{n} = \frac{23 + 29 + \dots + 65}{10} = \frac{447}{10} = 44.7$$

Interpretation: The average age is 44.7 years.

- **Meaning:** Balances all values equally; "center of gravity"
- **Use when:** Data is symmetric, continuous, no outliers
- **Not for:** Categorical data or ordinal where intervals aren't equal
- **Good use:** Mean income, mean height
- **Bad use:** Mean customer satisfaction (on 1–5 scale) — misleading due to ordinal nature

翻译见下页 Please see next page for translation

# 集中趋势 平均值

- 计算:

年龄 = [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]

$$\text{均值 Mean} = \frac{\sum x_i}{n} = \frac{23 + 29 + \dots + 65}{10} = \frac{447}{10} = 44.7$$

解读：平均年龄为44.7岁。

- 含义: 均值均等地平衡所有数值；“重心”。
- 使用情况：数据对称、连续、无离群点
- 不适用于：间隔不相等的分类数据或序数数据
- 适用于：平均收入、平均身高
- 错误用法：平均客户满意度（按 1-5 级）——序数性质导致误导

# Central Tendency. Mean 集中趋势 平均值

- **Monthly Salaries 月薪:** \$3000, \$3200, \$2800, \$3100, \$2950

Mean 平均值: Yes 是 / No 否?

- **Student Exam Scores 学生考试分数:** 72, 85, 90, 65, 78

Mean 平均值: Yes 是 / No 否?

- **Product Ratings 产品评级:** 3, 4, 2, 5

Mean 平均值: Yes 是 / No 否 ?

- **Zip Codes 邮政编码:** 90210, 10001, 30301

Mean 平均值: Yes 是 / No 否?

# Central Tendency. Mean 集中趋势 平均值

## How to interpret 如何解读:

- High mean → overall tendency toward larger values 高平均值 → 总体趋向于更大的值
- Low mean → most observations are relatively small 低平均值 → 大多数观测值相对较小

## Example insight 示例洞察:

- The average income in the region is \$48,000 — most people earn around this amount  
该地区的平均收入为 4.8 万 美元——大多数人的收入都在这个数额左右

## Warning: Sensitive to outliers! 警告：对离群点敏感！

- One billionaire can completely skew the result  
一位亿万富翁可以完全扭曲结果

# Central Tendency. Median 集中趋势 中位数

- **Calculation 计算:**

ages 年龄 = [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]

Sort data 数据排序: [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]

Median 中位数 =  $(42 + 45) / 2 = 43.5$

Interpretation: 50% of individuals are younger than 43.5 years, and 50% are older. 解读：50% 的人年龄小于 43.5 岁，50% 的人年龄大于 43.5 岁。

- **Meaning:** Middle-ranked value; resistant to outliers

含义：中等排名值；不易受离群值影响

- **Use when 使用情况:** Skewed distributions, ordinal data 偏态分布，序数数据

- **Not for 不适用于:** Nominal data (no order) 名义数据（无顺序）

- **Good use 恰当用法:** Median household income 家庭收入中位数

- **Bad use 不当用法:** Median country name 国家名称中位数

# Central Tendency. Median 集中趋势 中位数

- **Apartment Prices** 公寓价格: \$200k, \$220k, \$180k

Median 中位数: Yes 是 / No 否?

- **Test Scores** 考试分数: 40, 60, 90, 95, 100

- Median 中位数: Yes 是 / No 否?

- **Customer Satisfaction** 客户满意度: 1, 2, 2, 4, 5

- Median 中位数: Yes 是 / No 否?

- **Cities** 城市: Tokyo 东京, Paris 巴黎, London 伦敦, Berlin 柏林

Median 中位数: Yes 是 / No 否?

# Central Tendency. Median 集中趋势 中位数

## How to interpret 如何解读

- Median > Mean → right-skewed distribution (some large outliers)  
中位数 > 平均值 → 右偏分布（一些较大的离群值）
- Median < Mean → left-skewed distribution (some small outliers)  
中位数 < 平均值 → 左偏分布（一些小的离群值）

## Example insight 示例洞察

- The median income is \$35,000, which is lower than the mean — the wealthy pull the average up. Most people earn less than the average.  
中位数收入为3.5万美元，低于平均水平——富人拉高了平均水平。大多数人的收入低于平均水平。

# Central Tendency. Mode 集中趋势 众数

- **Calculation 计算:** Most frequent value(s) 频率最高的值  
ages 年龄 = [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]  
Here, 42 occurs twice 这里, 42 出现了两次  
Interpretation 解读: The most common age is 42 years 最常见的年龄是 42 岁
- **Meaning 意义:** Most frequent observation 最频繁的观察
- **Use when 使用情况:** You care about frequency 关心频率时
- **Applies to 适用于:** all scales 所有的尺度
- **Good use 恰当用法:** Most common customer complaint type  
最常见的客户投诉类型
- **Less useful: Continuous data (e.g., weight with all unique values)**  
不太恰当的用法: 连续数据 (例如, 具有唯一值的重量)

# Central Tendency. Mode 集中趋势 众数

- **Shoe Sizes 鞋码:** 38, 38, 39, 40, 38

Mode 众数: Yes/No 是/否

- **Car Colors 汽车颜色:** Red, Blue, Blue, Black 红、蓝、蓝、黑

Mode 众数: Yes/No 是/否

- **Temperatures 温度(°C):** 22, 23, 22, 21

Mode 众数: Yes/No 是/否

- **Product Codes 产品代码:** A123, B321, A123, A123

Mode 众数: Yes/No 是/否

# Central Tendency. Mode 集中趋势 众数

## How to interpret 如何解读

- Especially useful for categorical data (nominal or ordinal)
- Tells you what's most common, not what's "central"
- 尤其适用于分类数据（名义或序数）
- 告诉你什么是最常见的，而不是什么是“集中”

## Example insight 示例洞察

- The most popular coffee type is “Latte” — we should consider promoting it more  
最受欢迎的咖啡类型是“拿铁”——我们应该考虑更多地推广它

# Central Tendency. Practical cases

## 集中趋势 实际案例

# Measures of Spread. Range 离散程度的度量 极差

- **Calculation 计算:**

ages 年龄 = [23, 29, 35, 42, 42, 45, 50, 56, 61, 65]

Range = Max - Min 极差 = 最大值 - 最小值 = 65 - 23 = 42

Interpretation: The ages span 42 years. 解读：年龄跨度为42年。

- **Meaning 含义:** Difference between extremes (max - min)

极值之间的差异 ( 最大值 - 最小值 )

- **Use when 使用情况:** Quick sense of spread 迅速了解离散程度时

- **Not for 不适用于:** Categorical data, ordinal scales with unclear spacing

分类数据, 间距不明确的序数尺度

- **Good use 恰当用法:** Range of temperatures 温度范围

- **Bad use 不当用法:** Range of product satisfaction ratings (1 to 5) may ignore distribution shape

产品满意度评级范围 ( 1 至 5 ) 可能会忽略分布形状

# Measures of Spread. Range 离散程度的度量 极差

- **Lifespan (years) 寿命（年）** : 70, 85, 90, 95

Range 极差: Yes/No? 是/否？

- **Temperature (°F) 温度（华氏度）** : 32, 45, 60, 55

Range 极差: Yes/No? 是/否？

- **Star Ratings 星级评定**: 1★, 2★, 4★, 5★

Range 极差: Yes/No? 是/否？

- **Country Names 国家名称**: USA 美国, France 法国, Germany 德国

Range 极差: Yes/No? 是/否？

# Measures of Spread. Range 离散程度的度量 极差

## How to interpret 如何解读

- A simple measure of total spread; shows how far apart the smallest and largest values are. 一个简单的总离散程度度量；显示最小值和最大值之间的差距。

## Example insight 示例洞察

- The age range in the group is 22 to 65 — a diverse age group.  
该群体的年龄范围为 22 至 65 岁——是一个多元化的年龄群体。

## Warnings 警告

- Very sensitive to outliers 对离群值非常敏感
- Doesn't reflect variability in the middle of the data 不反映数据中间的变异性

# Measures of Spread. Variance / Standard deviation

## 离散程度的度量 方差/标准差

### Calculation. Variance 计算 方差

情况 Situation	公式 Formula	分母 Denominator	原因 Reason
Population 总数	$\sigma^2 = \frac{1}{N} \sum (x_i - \mu)^2$	$N$	You have all data — no estimation 你有所有数据 — 不需要估计
Sample 样本	$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$	$n - 1$	Corrects bias in small samples 纠正小样本中的偏差

### Calculation. Standard deviation 计算 标准差

$$s = \sqrt{s^2} \quad \text{or} \quad \sigma = \sqrt{\sigma^2}$$

# Measures of Spread. Variance / Standard deviation

## 离散程度的度量 方差/标准差

Let's say we have a sample of **monthly sales** in \$1000:

例如，样本为月度销售额为1000美金

[5, 7, 3, 7, 10]

Step 1: Calculate the mean    第1步：计算均值

$$\bar{x} = \frac{5 + 7 + 3 + 7 + 10}{5} = \frac{32}{5} = 6.4$$

Step 2: Subtract the mean and square the result    第2步：减去平均值并计算结果的平方

$$(5 - 6.4)^2 = 1.96 \quad (7 - 6.4)^2 = 0.36 \quad (3 - 6.4)^2 = 11.56 \quad (7 - 6.4)^2 = 0.36 \quad (10 - 6.4)^2 = 12.96$$

Step 3: Add the squared deviations    第3步：将偏差平方后求和

$$\sum (x_i - \bar{x})^2 = 1.96 + 0.36 + 11.56 + 0.36 + 12.96 = 27.2$$

Step 4: Divide by  $n - 1$  (sample size - 1)    第4步：除以  $n-1$ （样本大小-1）

$$s^2 = \frac{27.2}{4} = 6.8 \quad (\text{Variance}) \quad \text{方差}$$

Step 5: Take the square root    第5步：计算平方根

$$s = \sqrt{6.8} \approx 2.61 \quad (\text{Standard Deviation}) \quad \text{标准差}$$

# Measures of Spread. Variance / Standard deviation

## 离散程度的度量 方差/标准差

- **Calculations** 计算: IQR 四分位距 =  $Q3 - Q1$ ,  
where 其中:  $Q1 = 25\text{th percentile (lower quartile)}$  第25百分位数 (下四分位数) ;  $Q3 = 75\text{th percentile (upper quartile)}$  第75百分位数 (上四分位数)

Data 数据: [3, 7, 8, 5, 12, 14, 21, 13, 18] -> [3, 5, 7, 8, 12, 13, 14, 18, 21]

Median 中位数 = 12

$Q1 = \text{median of lower half}$  下半部分的中位数: 5

$Q3 = \text{median of upper half}$  上半部分中位数: 14

IQR 四分位距 =  $14 - 5 = 9$

- **Meaning** 含义: Spread around the mean (variance is squared, std is same units as data)  
围绕平均值分布 ( 方差为平方 · 标准差与数据单位相同 )
- **Use when** 使用情况: Data is numerical, especially if symmetric 数据是数字, 特别是对称的
- **Not for** 不适用于: Categorical or ordinal without equal intervals 无等距间隔的分类或序数数据
- **Good use** 恰当用法: Std deviation of monthly sales 月销售额的标准差
- **Bad use** 不当用法: Variance of education levels coded as 1–5 编码为 1–5 的教育水平方差

# Measures of Spread. Standard deviation 离散程度的度量 标准差

- **Product Weights 产品重量 (kg):** 2.3, 2.5, 2.1, 2.4

Std Dev: Yes/No? 标准差：是/否？

- **Blood Pressure 血压 (mmHg):** 120, 130, 125

Std Dev: Yes/No? 标准差：是/否？

- **Satisfaction Scores 满意度分数:** 3, 4, 5, 2

Std Dev: Yes/No? 标准差：是/否？

- **ID Numbers 身份证号码:** 102, 105, 110

Std Dev: Yes/No? 标准差：是/否？

## How to interpret 如何解读

- A low standard deviation → data points are close to the mean  
标准差较低→数据点接近平均值
- A high standard deviation → data is more spread out  
标准差较大→数据更加分散

## Example insight 示例洞察

- The standard deviation of monthly sales is \$1500 — sales vary moderately around the average.  
月销售额的标准差为 1500 美元——销售额在平均值附近适度波动。

## Measures of Spread. Interquartile range 离散程度的度量 四分位距

- **Meaning : Middle 50% spread (Q3 – Q1)**
- **Use when : Resistant to outliers; non-normal data**  
使用情况：对离群值具有抵抗力;非正态分布的数据
- **Not for 不适用于: Nominal 名义**  
Good use: IQR of salaries, delivery times  
恰当用法：工资、交货时间的四分位距
- **Bad use : IQR of product names**  
不当用法：产品名称的四分位距

## Measures of Spread. Interquartile range 离散程度的度量 四分位距

- **Daily Steps 每日步数**: 5000, 6000, 7000, 8000, 9000

IQR: Yes/No? 四分位距 : 是/否 ?

- **Exam Scores 考试分数**: 55, 60, 65, 90, 95

IQR: Yes/No? 四分位距 : 是/否 ?

- **Survey Ratings 调查评级**: 2, 3, 3, 4, 5

IQR: Yes/No? 四分位距 : 是/否 ?

- **Phone Numbers 电话号码**: 12345, 23456, 34567

IQR: Yes/No? 四分位距 : 是/否 ?

# Measures of Spread. Interquartile range 离散程度的度量 四分位距

## How to interpret 如何解读

- Describes where the bulk of values lie, ignoring extremes.  
描述大部分值所在的位置，忽略极端值。

## Example insight 示例洞察

- The IQR of exam scores is 20 — most students scored within a 20-point range.  
考试成绩的四分位距为 20——大多数学生的成绩在 20 分范围内。

## Warnings 警告

- Doesn't show the full range of variability 未显示全部变异范围
- May not reflect multimodal distributions 可能无法反映多峰分布

## Measures of Spread (Dispersion). Practical cases

### 离散程度的度量（离散） 实际案例

# Measures of Shape. Skewness / Kurtosis 形状度量 偏度/峰度

- **Meaning:** Shape of distribution — asymmetry and tailedness  
含义：分布的形状——不对称性和尾部性
- **Use when:** You want to assess normality or detect outliers  
使用情况：想要评估正态性或检测离群值
- **Only for:** Quantitative data 仅适用于：定量数据
- **Good use :** Distribution of investment returns  
恰当用法：投资收益分配
- **Bad use :** Shape of nominal variables (e.g., brand names)  
不当用法：名义变量的形状（例如：品牌名称）

# Measures of Shape. Skewness

## How to interpret

- Positive skew: long tail to the right
- Negative skew: long tail to the left
- Skew  $\approx 0$ : fairly symmetric

## Example insight

- Income data shows strong positive skew — a few individuals earn much more than the rest.

## Warnings

- Sensitive to outliers
- Not useful on very small samples
- Skewed data may affect mean-based statistics

翻译见下页 Please see next page for translation

# 形状度量 偏度

## 如何解读

- 正偏态：长尾在右侧
- 负偏态：长尾在左侧
- 偏度 $\approx 0$ ：相对对称

## 示例洞察

- 收入数据呈现出强烈的正偏态——少数人的收入远远高于其他人。

## 警告

- 对离群值敏感
- 对于非常小的样本无用
- 偏态数据可能会影响基于均值的统计

# Measures of Shape. Kurtosi

## How to interpret

- High kurtosis: heavy tails, more outliers
- Low kurtosis: light tails, fewer outliers
- Normal distribution has kurtosis  $\approx 3$  (excess kurtosis = 0)

## Example insight

- Sales data shows high kurtosis — frequent extreme changes month to month

## Warnings

- Often misunderstood as "peakness" (but it's about tails)
- Easily distorted by a few outliers
- Use with other statistics for a full picture

翻译见下页 Please see next page for translation

# 形状度量 峰度

## 如何解读

- 高峰度：尾部较重，离群值较多
- 低峰度：尾部较轻，离群值较少
- 正态分布的峰度  $\approx 3$  (过度峰度 = 0)

## 示例洞察

- 销售数据呈现高峰度——每月频繁出现极端变化

## 警告

- 经常被误解为“峰值”（但它与尾部有关）
- 容易被一些离群值扭曲
- 与其他统计数据一起使用以了解完整情况

# Measures of Shape. Z-scores 形状度量 Z分数

- **Meaning:** How far a point is from the mean in std units  
含义：一个点与平均值的距离（以标准差为单位）
- **Use when:** You need to compare across variables or detect outliers  
使用情况：你需要跨变量进行比较或检测离群值
- **Only for:** Quantitative 仅适用于：定量数据
- **Good use :** Compare student scores across tests with different scales  
恰当用法：用不同的尺度比较不同测试中的学生成绩
- **Bad use :** Z-score of phone brands  
不当用法：手机品牌的Z分数

# Measures of Shape. Z-scores 形状度量 Z分数

## How to interpret 如何解读

- Tells how unusual a value is in the context of the dataset  
说明某个值在数据集中的异常程度

## Example insight 示例洞察

- A z-score of 2.1 for this month's sales means sales were significantly higher than usual  
本月销售额的标准分数为 2.1，这意味着销售额明显高于平常

## Warnings 警告

- Assumes a normal (or roughly symmetric) distribution 假设服从正态（或大致对称）分布
- Not meaningful for categorical or skewed data 对于分类数据或偏态数据无意义
- Outliers will have very large/small z-scores 离群值将具有非常大/非常小的标准分数

# Measures of Shape. Practical cases

## 形状度量 实用案例

# What Can Summary Statistics Tell Us?

## Data Distribution

- Is the data symmetric, skewed, or uniform?
- Are there outliers or extreme values?

## Data Quality

- How much missing data is there?
- Are there unexpected values (e.g., negative values in a positive-only dataset)?

## Insights for Modeling

- Do we need to normalize or scale the data?
- Should we handle outliers or missing values before modeling?

## Business Insights

- What are the typical values for key metrics?
- How much variability exists in the data?

翻译见下页 Please see next page for translation

# 汇总统计数据能告诉我们什么？

## 数据分布

- 数据是对称的、偏态的还是均匀分布的？
- 是否存在离群值或极端值？

## 数据质量

- 有多少缺失数据？
- 是否存在意外值（例如，仅正值的数据集中的负值）？

## 建模洞察

- 我们需要对数据进行标准化或缩放吗？
- 我们是否应该在建模之前处理离群值或缺失值？

## 商业洞见

- 关键指标的典型值是多少？
- 数据中存在多少变化？

## Practical cases 实用案例

# Useful Links 实用链接

[Exploratory Data Analysis with Pandas](#)

使用 Pandas 进行探索性数据分析

[Mastering Exploratory Data Analysis \(EDA\): A Comprehensive Python \(Pandas\) Guide for Data Insights and Storytelling](#)

掌握探索性数据分析 (EDA)：数据洞察和叙事的综合 Python (Pandas) 指南