Swiss UMEF
UNIVERSITY OF APPLIED SCIENCES INSTITUTE

Big Data Analytics
大数据分析

# 07: In-Memory Analytics with Pandas. Chart Visualization
# 07:使用 Pandas 进行内存分析 图表可视化

Instructor: Oleh Tymchuk
授课教师：奥勒·特姆恰克

#07: Agenda 课程安排

- Introduction 介绍
- Legend: Dataset Overview 图例：数据集概览
- Univariate Analysis 单变量分析
- Bivariate Analysis 双变量分析
- Multivariate Analysis 多变量分析
- Practical cases 实际案例
- Useful Links 实用链接

Introduction 介绍

Key Principles of Effective Visualization 有效可视化的关键原则

Clarity: Ensure the message is clear

Accuracy: Represent data truthfully

Efficiency: Convey insights with minimal clutter

Aesthetics: Make it visually appealing

清晰度：确保信息清晰明了

准确性：真实呈现数据

效率：以最少的混乱传达洞见

美观性：使其具有视觉吸引力

Python Libraries for Data Visualization 用于数据可视化的 Python 库

Matplotlib:

Low-level library for creating static, animated, and interactive plots.
用于创建静态、动画和交互式图表的低级库。

Seaborn:

High-level library built on Matplotlib for statistical visualizations.
基于 Matplotlib 构建的用于统计可视化的高级库。

Plotly:

Interactive and web-based visualizations. 交互式和基于网络的可视化。

Pandas Plotting:

Built-in plotting tools for quick visualizations. 内置绘图工具，可快速实现可视化。

Types of Visualizations 可视化类型

Univariate Analysis 单变量分析:

- Histograms 直方图
- Boxplots 箱线图
- KDE plots 核密度估计图

Bivariate Analysis 双变量分析:

- Scatterplots 散点图
- Line plots 线图
- Bar plots 条形图

Multivariate Analysis 多变量分析:

- Heatmaps 热力图
- Pairplots 配对图
- 3D plots 三维图

Specialized Visualizations 专业可视化:

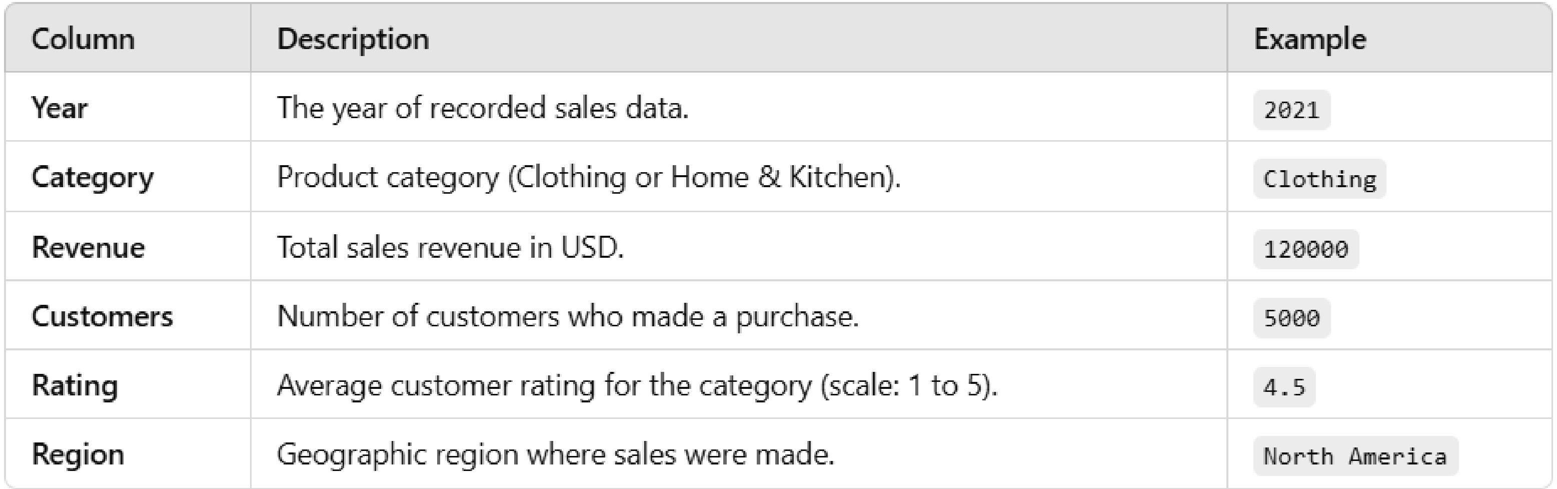
- Geospatial maps 地理空间地图
- Network graphs 网络图
- Word clouds 词云

Legend: Dataset Overview
图例：数据集概览

Dataset Overview

Before diving into visualizations, let's define the dataset we'll use for examples.

Context: This dataset represents sales data from an online store, covering two product categories—Clothing and Home & Kitchen—over a five-year period (2019-2023)



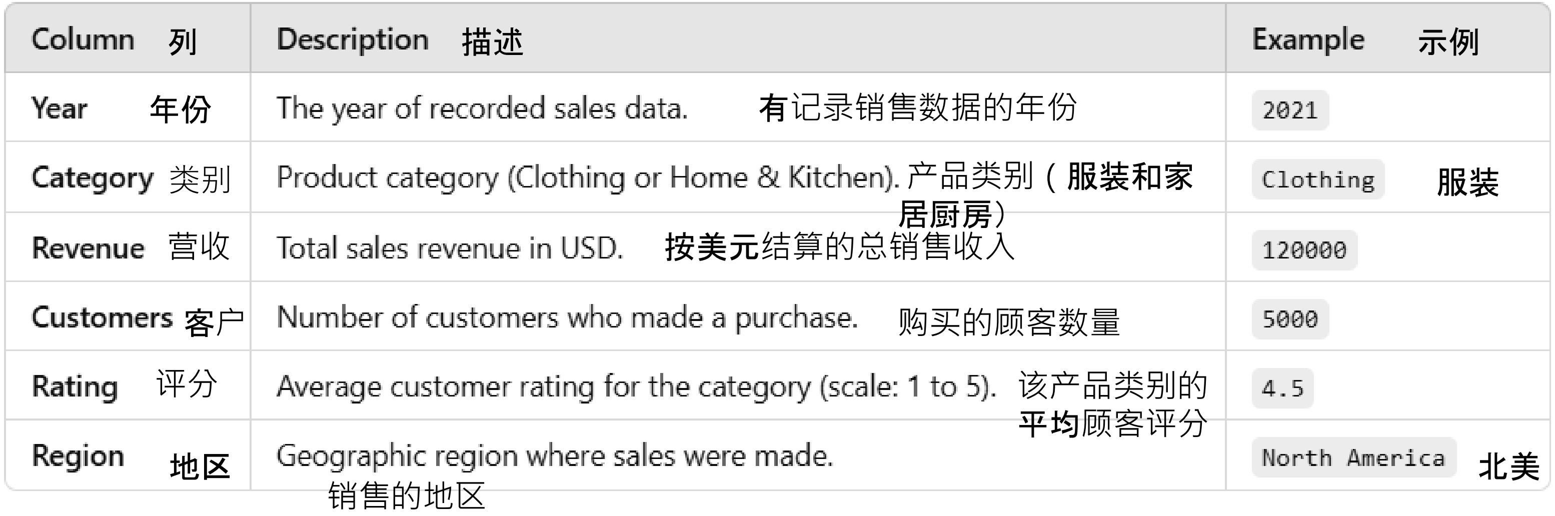
Column	Description	Example
Year	The year of recorded sales data.	`2021`
Category	Product category (Clothing or Home & Kitchen).	`Clothing`
Revenue	Total sales revenue in USD.	`120000`
Customers	Number of customers who made a purchase.	`5000`
Rating	Average customer rating for the category (scale: 1 to 5).	`4.5`
Region	Geographic region where sales were made.	`North America`

翻译见下页 Please see next page for translation

数据集概览

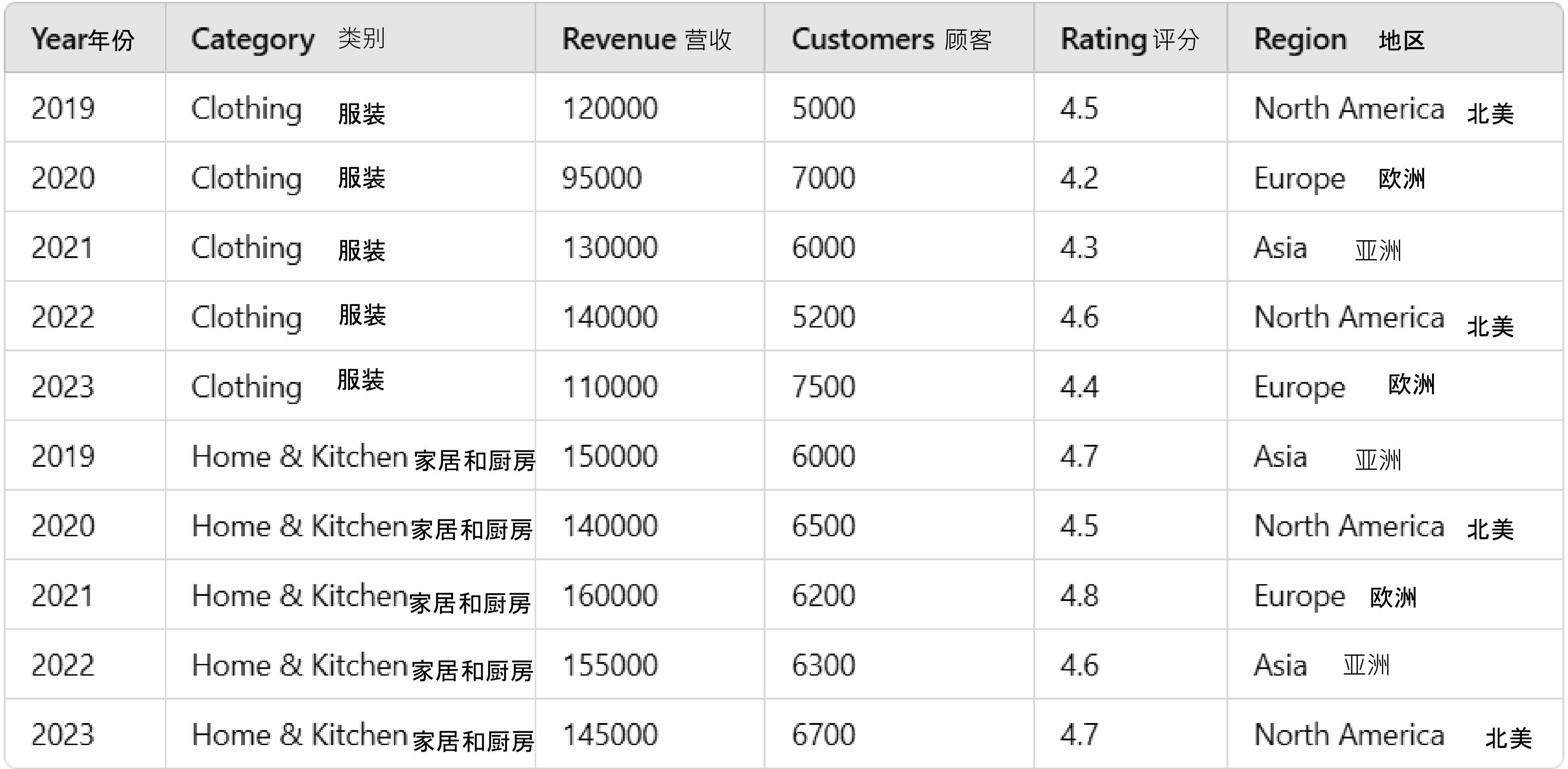
在深入探讨可视化之前，我们先来定义一下示例数据集。

背景：该数据集代表一家网店的销售数据，涵盖五年期间（2019-2023 年）的服装和家居厨房用品两大产品类别。



Column 列	Description 描述	Example 示例
Year 年份	The year of recorded sales data. 有记录销售数据的年份	`2021`
Category 类别	Product category (Clothing or Home & Kitchen). 产品类别（服装和家居厨房）	`Clothing` 服装
Revenue 营收	Total sales revenue in USD. 按美元结算的总销售收入	`120000`
Customers 客户	Number of customers who made a purchase. 购买的顾客数量	`5000`
Rating 评分	Average customer rating for the category (scale: 1 to 5). 该产品类别的平均顾客评分	`4.5`
Region 地区	Geographic region where sales were made. 销售的地区	`North America` 北美

Dataset Example 数据集示例



Year年份	Category 类别	Revenue 营收	Customers 顾客	Rating 评分	Region 地区
2019	Clothing 服装	120000	5000	4.5	North America 北美
2020	Clothing 服装	95000	7000	4.2	Europe 欧洲
2021	Clothing 服装	130000	6000	4.3	Asia 亚洲
2022	Clothing 服装	140000	5200	4.6	North America 北美
2023	Clothing 服装	110000	7500	4.4	Europe 欧洲
2019	Home & Kitchen 家居和厨房	150000	6000	4.7	Asia 亚洲
2020	Home & Kitchen 家居和厨房	140000	6500	4.5	North America 北美
2021	Home & Kitchen 家居和厨房	160000	6200	4.8	Europe 欧洲
2022	Home & Kitchen 家居和厨房	155000	6300	4.6	Asia 亚洲
2023	Home & Kitchen 家居和厨房	145000	6700	4.7	North America 北美

Univariate Analysis
单变量分析

Histograms

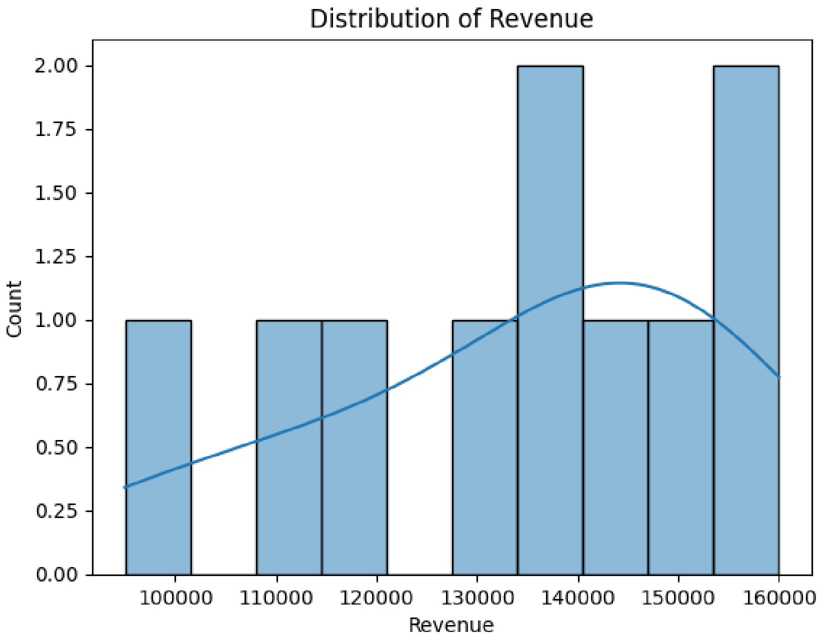
A histogram shows the frequency distribution of numerical data. It groups data into bins and counts occurrences in each bin.

Use Cases:

- understanding the shape of the data (normal, skewed, etc.)
- Identifying outliers
- Checking for multimodal distributions

Insights:

- Reveals if revenue follows a normal distribution
- Identifies peaks and gaps in revenue values



翻译见下页 Please see next page for translation

直方图

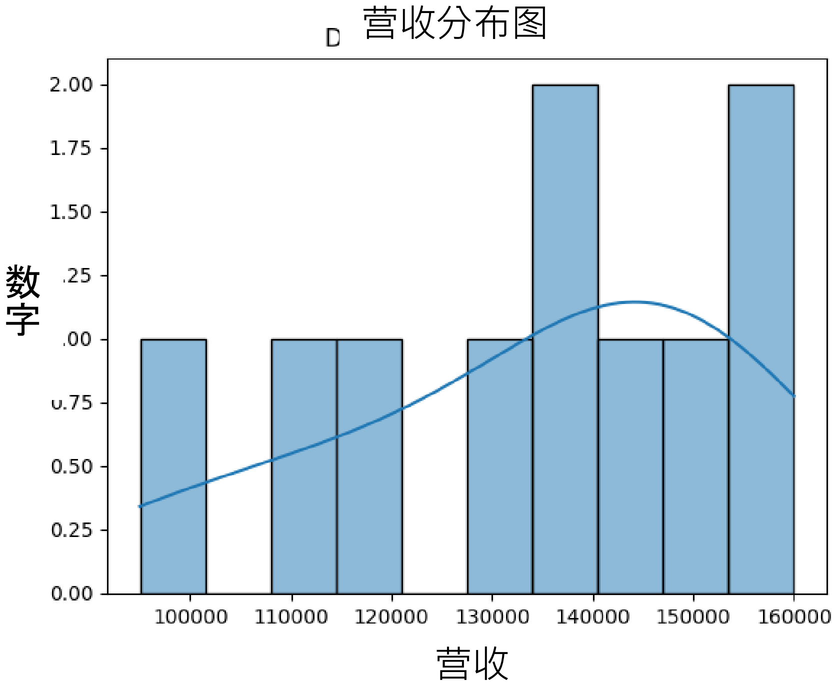
直方图显示数值数据的频率分布。它将数据分组到各个区间内，并计算每个区间内出现的次**数**。

用例:

- **理解数据形状**（**正**态分布、偏态分布等）
- 识别**离群**值
- 检查多峰分布

洞见:

- **揭示收入是否遵循正**态分布
- 识别收入值的峰值和缺口



Boxplots

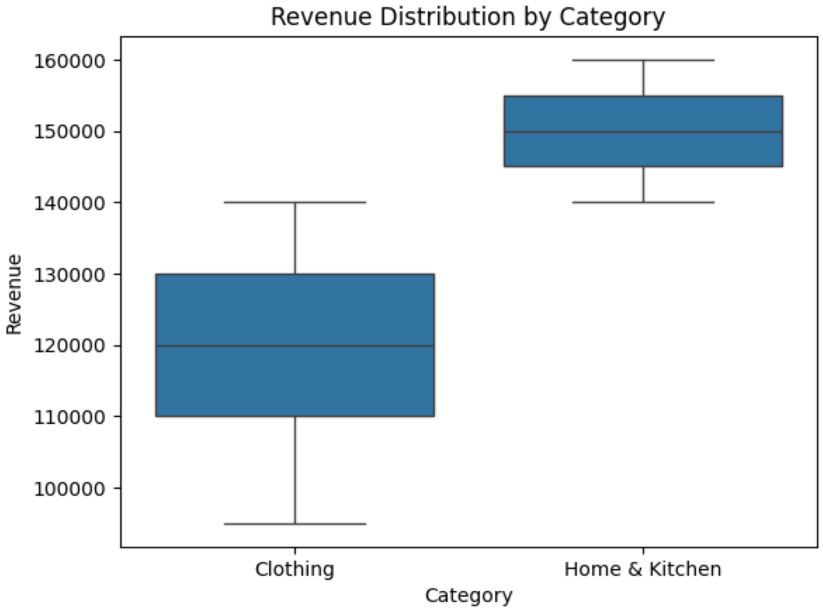
A boxplot shows the distribution of data, including quartiles and potential outliers.

Use Cases:

- Comparing distributions across different categories
- Identifying outliers

Insights:

- Displays median revenue for each category
- Shows variability and presence of outliers



翻译见下页 Please see next page for translation

箱线图

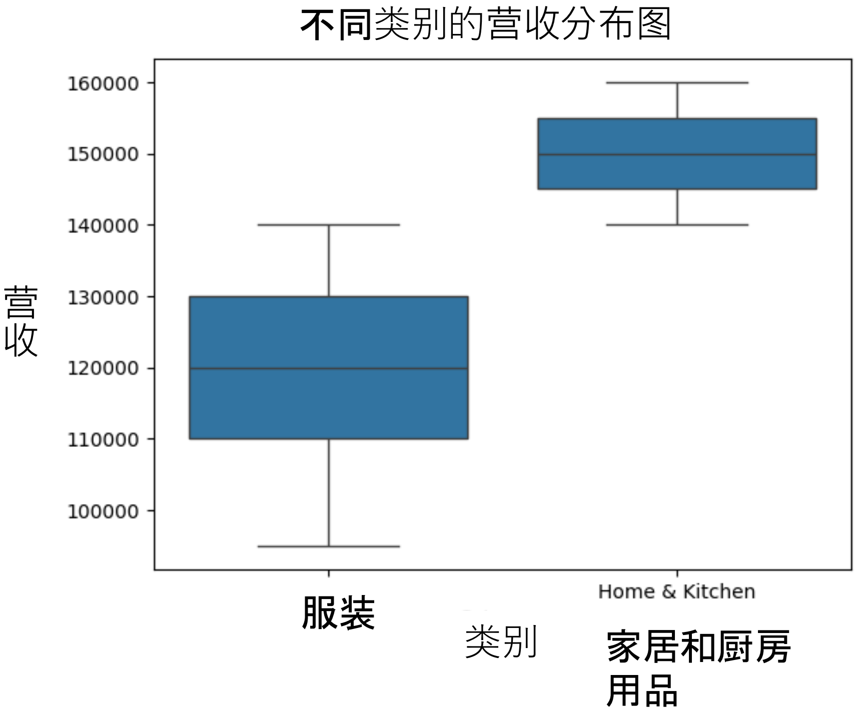
箱线图显示数据分布，包括四分位数和潜在**离群值**。

用例:

- **比**较不同类别的分布
- 识别**离群值**

洞见:

- 显示每个类别的平均收入
- 显示变异性及**离群值**的存在



KDE plots

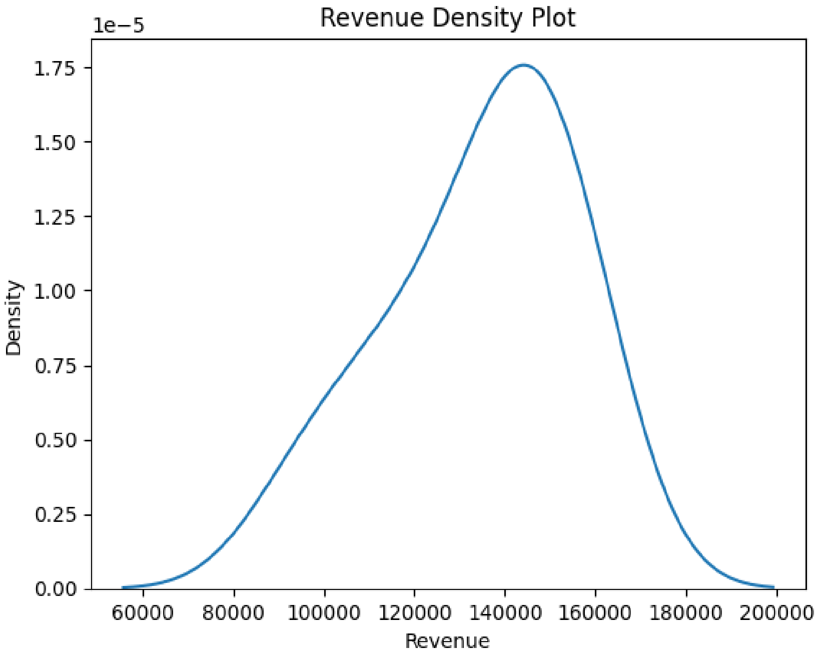
A KDE (Kernel Density Estimate) plot is a smoothed version of a histogram that shows the probability density function.

Use Cases:

- Understanding distribution trends
- Finding peaks and troughs in data

Insights:

- Shows revenue concentration around specific values
- Helps detect multiple peaks (bimodal distribution)



翻译见下页 Please see next page for translation

核密度估计图

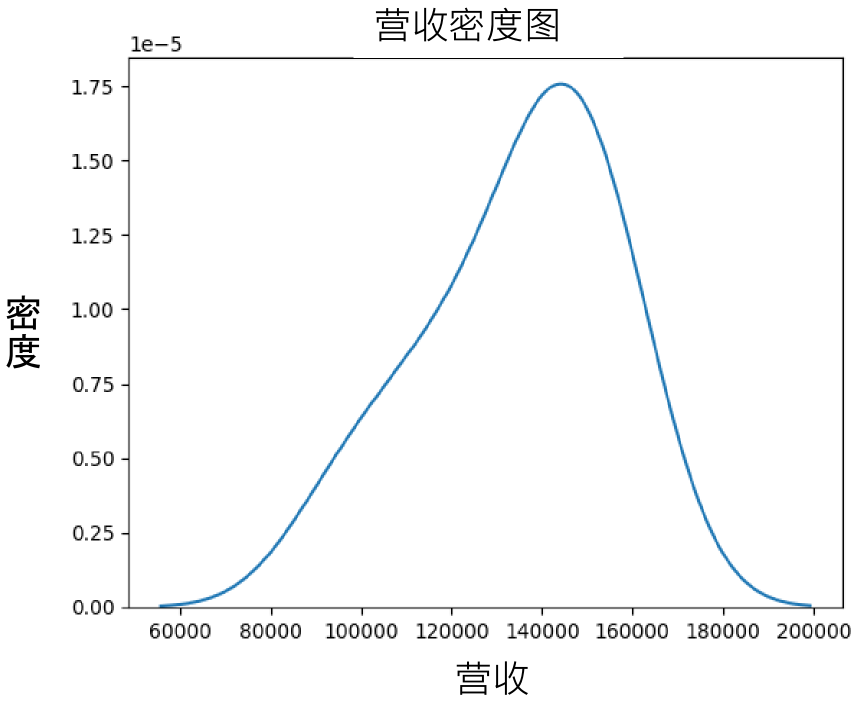
KDE（**核密度估计**）图是直方图的平滑版本，显示概率密度函数。

用例:

- **了解分布**趋势
- 寻找数据中的峰值和谷值

洞见:

- 显示收入集中于特定价值
- **帮助**检测多个峰值（双峰分布）



Bivariate Analysis
双变量分析

Scatterplots

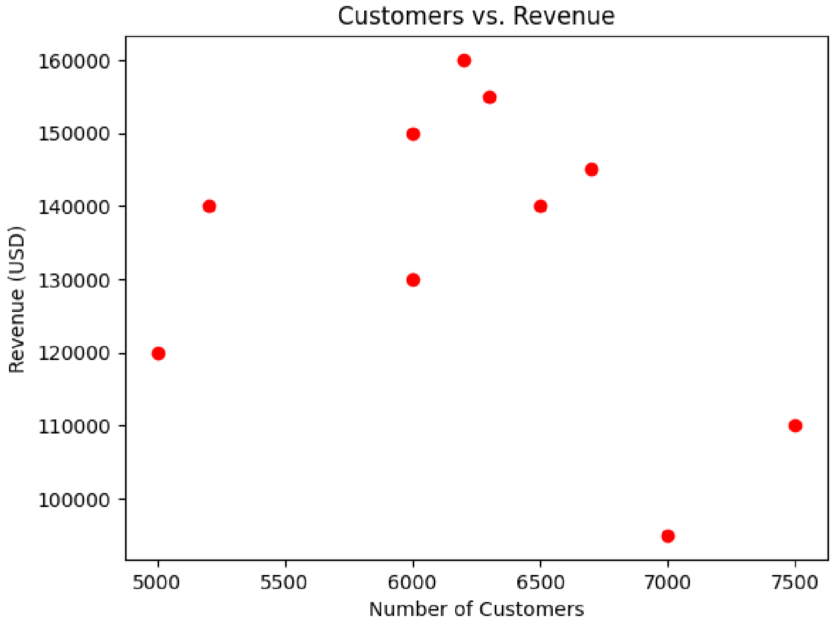
A scatter plot shows how two variables are related. Each point represents a data pair.

Use Cases:

- Finding correlations between numerical variables
- Detecting patterns and anomalies

Insights:

- Identifies if more customers lead to higher revenue
- Detects unusual customer-revenue relationships



翻译见下页 Please see next page for translation

Scatterplots 散点图

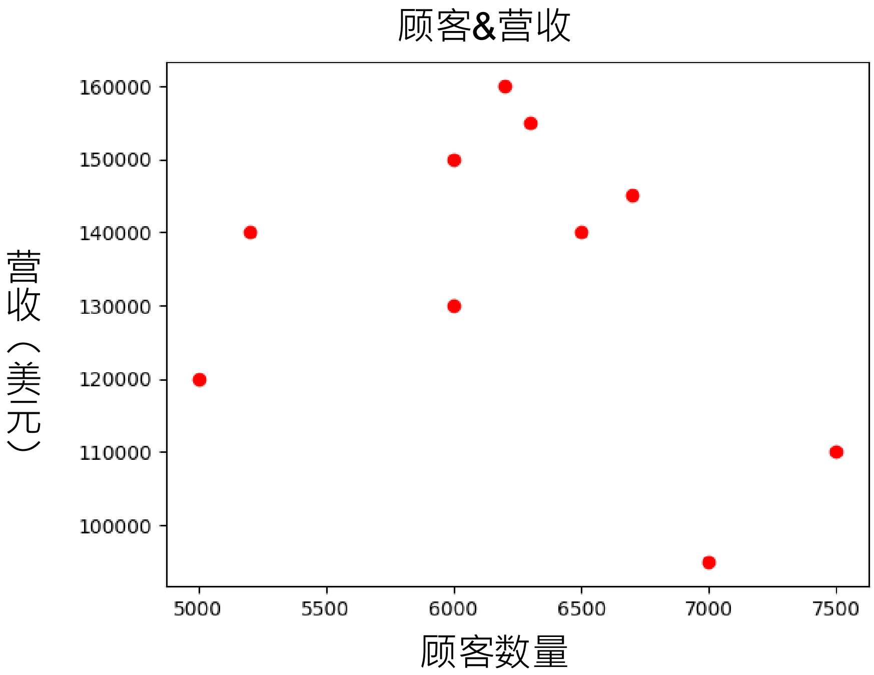
散点图显示两个变量之间的关系。每个点**代表一**对数据。

用例:

- 寻找数值变量之间的相关性
- 检测模式和异常值

洞见:

- 识别更多客户是否能带来更高收入
- 检测异常的客户-**收入关系**



Line plots 线图

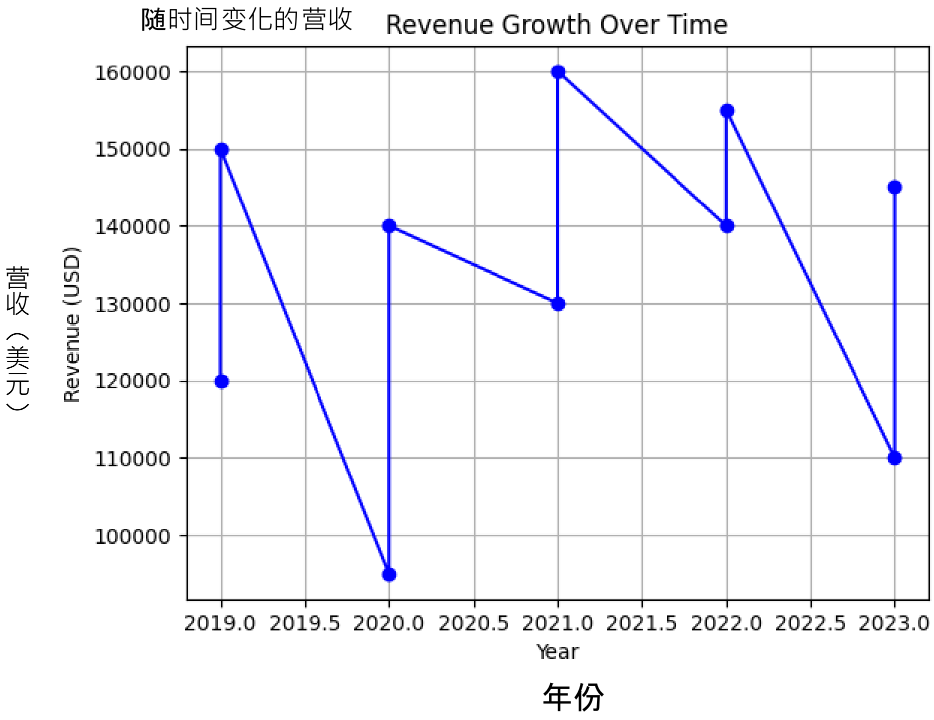
A line plot shows trends over time. It connects data points sequentially.
线图显示了随时间变化的趋势。它按顺序连接数据点。

Use Cases 用例:

- Analyzing revenue growth over years
  分析多年来的收入增长
- Identifying seasonal patterns
  识别季节性模式

Insights 洞见:

- Shows revenue trends over years
  显示多年来的收入趋势
- Detects dips or spikes in revenue
  检测收入的下降或飙升



Bar plots 条形图

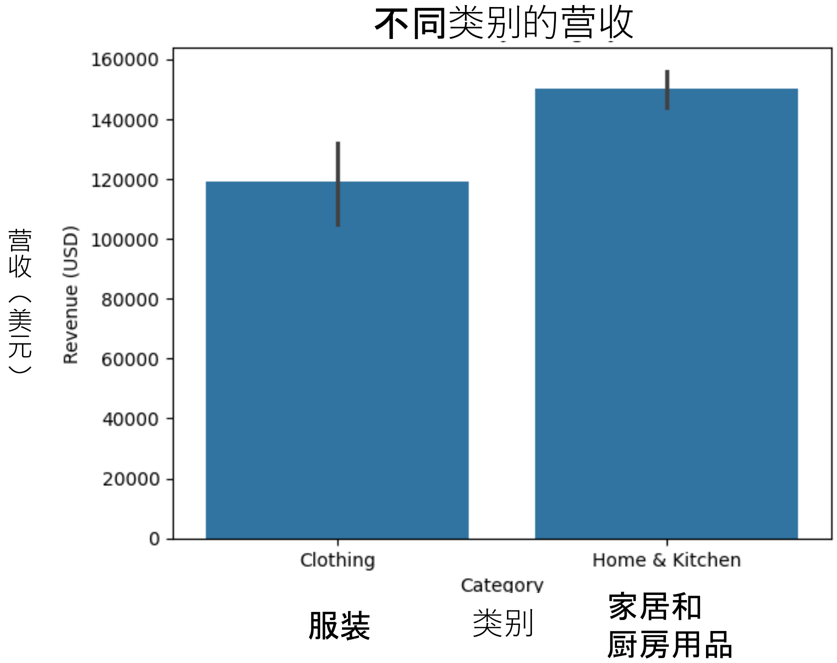
A bar plot compares values across different categories.
条形图比较不同类别的值。

Use Cases 用例:

- Comparing revenue between product categories
- 比较不同产品类别的收入
- Identifying top-performing segments
  识别表现最佳的细分市场

Insights 洞见:

- Shows which category generates higher revenue
  显示哪些类别的收入更高
- Highlights performance differences
  突出业绩差异



Multivariate Analysis 多变量分析

Heatmaps 热力图

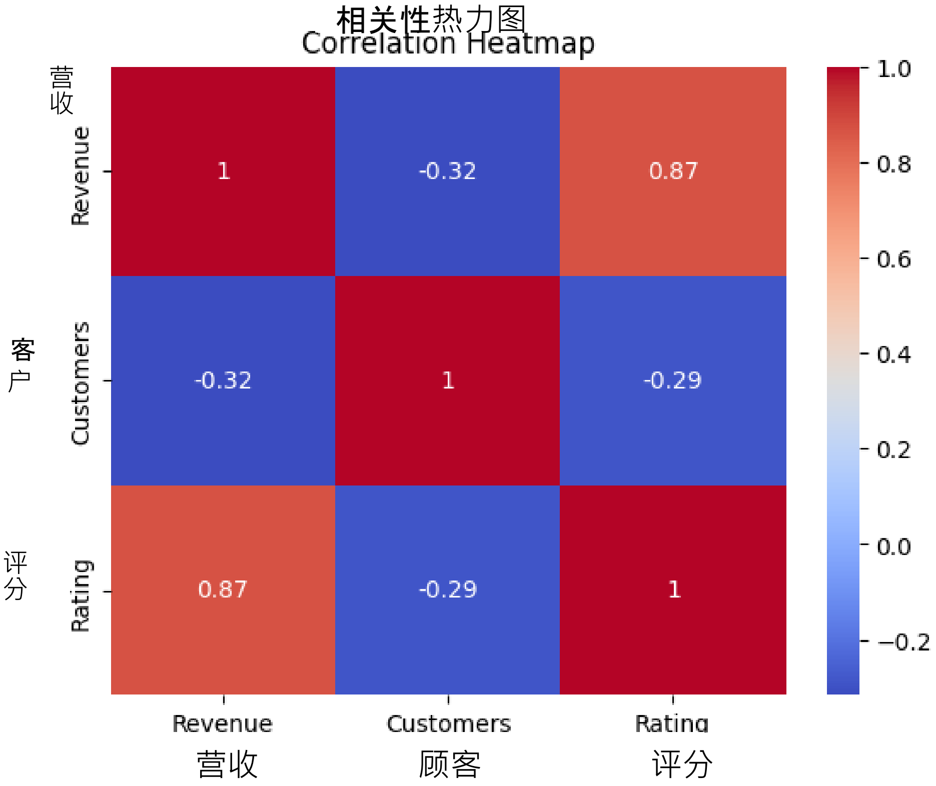
A heatmap visualizes correlations between multiple variables.
热图直观地显示了多个变量之间的相关性。

Use Cases 用例:

- Understanding how different variables relate to each other.
  了解不同变量之间的关系。
- Identifying strong or weak correlations.
  识别强相关性或弱相关性。

Insights 洞见:

- Shows if revenue correlates with customer count or ratings.
  显示收入是否与客户数量或评分相关。
- Identifies strong positive or negative correlations.
  识别强正相关性或负相关性。



Pairplots 配对图

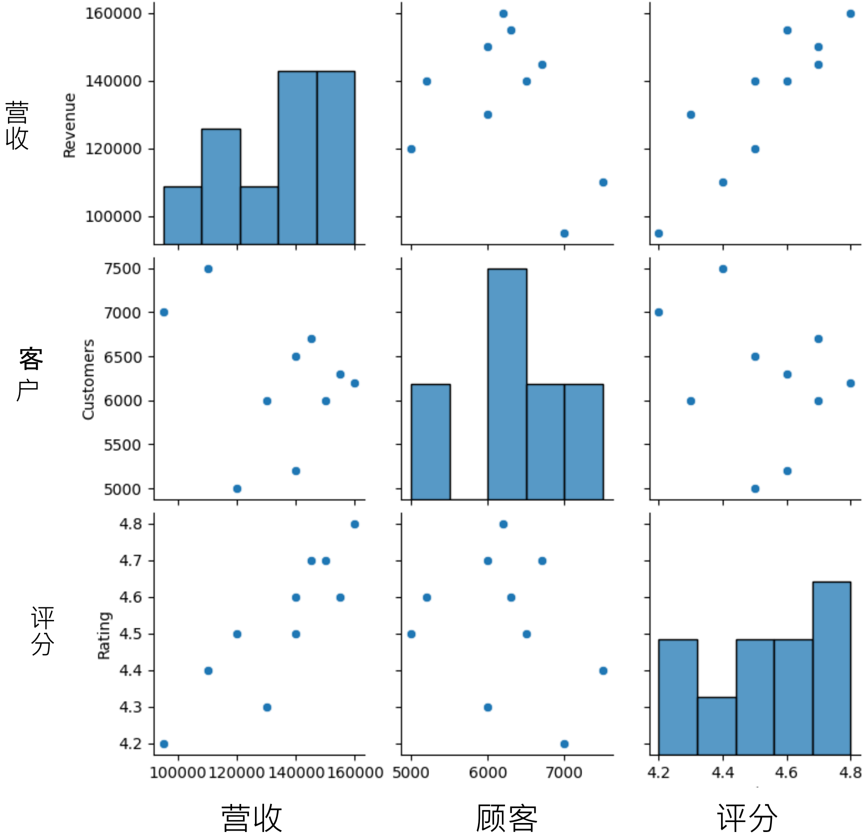
A pairplot creates scatter plots for all numeric variable combinations.
配对图为所有数值变量组合创建散点图。

Use Cases 用例:

- Exploring variable relationships.
  探索变量关系。
- Detecting clustering patterns.
  检测聚类模式。

Insights 洞见:

- Visualizes how variables interact.
  可视化变量间的相互作用。
- Highlights potential outliers.
  突出显示潜在的离群值。



3D plots 三维图

A 3D scatter plot adds a third variable to a standard scatter plot.
3D 散点图在标准散点图中添加了第三个变量。

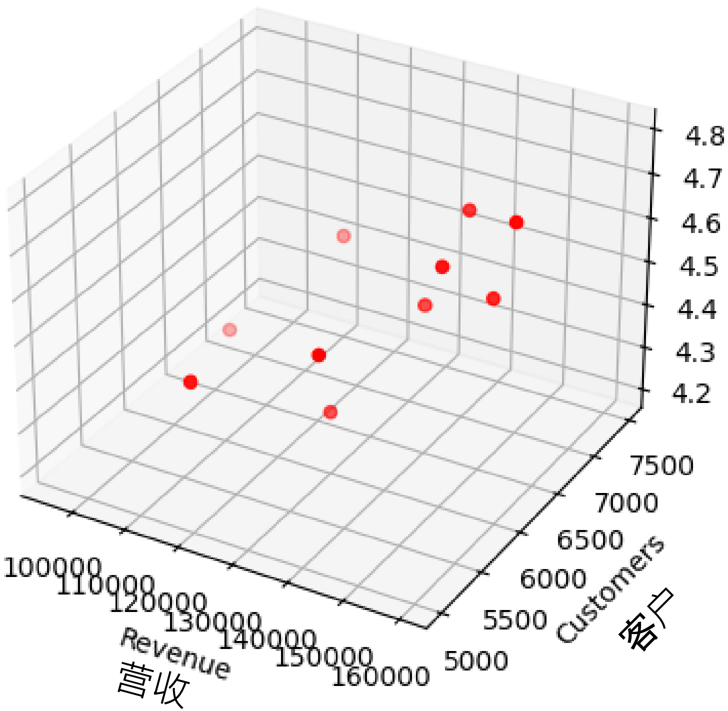
Use Cases 用例:

- Showing interactions between three numerical variables.
  显示三个数值变量之间的相互作用。

Insights 洞见:

- Displays trends in three-dimensional space.
  显示三维空间中的趋势。
- Shows how customer count, revenue, and rating interact.
  显示客户数量、营收和评分如何相互作用。

3D Scatter Plot 3D 散点图



Practical cases 实际案例

Useful Links 实用链接

Matplotlib 3.10.3 documentation

Seaborn. User guide and tutorial

Pandas. Chart visualization

Plotly Open Source Graphing Library for Python

Matplotlib 3.10.3 文档

Seaborn. 用户指南和教程

Pandas. 图表可视化

Plotly 开源 Python 图形库

Q&A 问答