Swiss UMEF
UNIVERSITY OF APPLIED SCIENCES INSTITUTE

Big Data Analytics

06: In-Memory Analytics with Pandas. Chart Visualization

Instructor: Oleh Tymchuk

#06: Agenda

- Introduction
- Legend: Dataset Overview
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Practical cases
- Useful Links

Introduction

Key Principles of Effective Visualization

Clarity: Ensure the message is clear

Accuracy: Represent data truthfully

Efficiency: Convey insights with minimal clutter

Aesthetics: Make it visually appealing

Python Libraries for Data Visualization

Matplotlib:

Low-level library for creating static, animated, and interactive plots.

Seaborn:

High-level library built on Matplotlib for statistical visualizations.

Plotly:

Interactive and web-based visualizations.

Pandas Plotting:

Built-in plotting tools for quick visualizations.

Types of Visualizations

Univariate Analysis:

- Histograms
- Boxplots
- KDE plots

Bivariate Analysis:

- Scatterplots
- Line plots
- Bar plots

Multivariate Analysis:

- Heatmaps
- Pairplots
- 3D plots

Specialized Visualizations:

- Geospatial maps
- Network graphs
- Word clouds

Legend: Dataset Overview

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`

Dataset Example

Year	Category	Revenue	Customers	Rating	Region
2019	Clothing	120000	5000	4.5	North America
2020	Clothing	95000	7000	4.2	Europe
2021	Clothing	130000	6000	4.3	Asia
2022	Clothing	140000	5200	4.6	North America
2023	Clothing	110000	7500	4.4	Europe
2019	Home & Kitchen	150000	6000	4.7	Asia
2020	Home & Kitchen	140000	6500	4.5	North America
2021	Home & Kitchen	160000	6200	4.8	Europe
2022	Home & Kitchen	155000	6300	4.6	Asia
2023	Home & Kitchen	145000	6700	4.7	North America

Univariate Analysis

Histograms

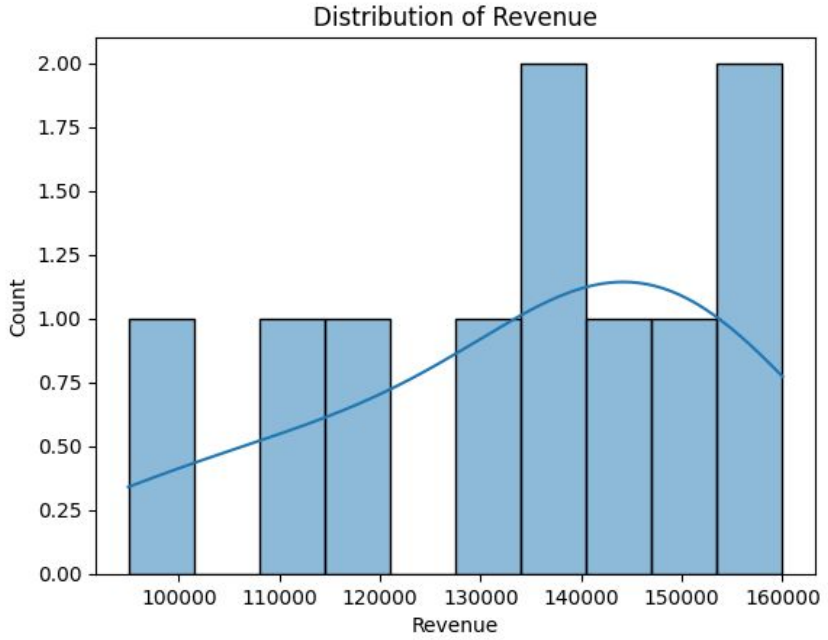
A histogram shows the frequency distribution of numerical data. It groups data into bins and counts occurrences in each bin.

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



Boxplots

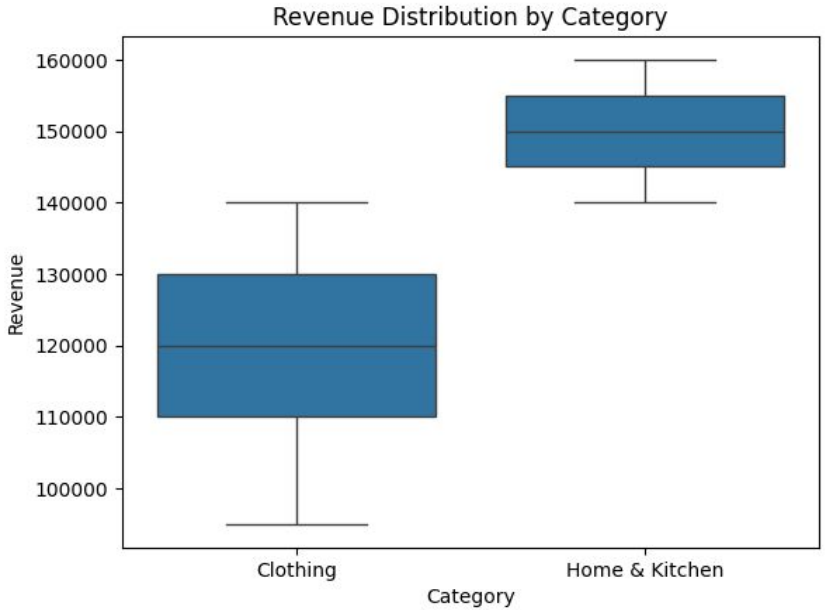
A boxplot shows the distribution of data, including quartiles and potential outliers.

Use Cases:

- Comparing distributions across different categories
- Identifying outliers

Insights:

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



KDE plots

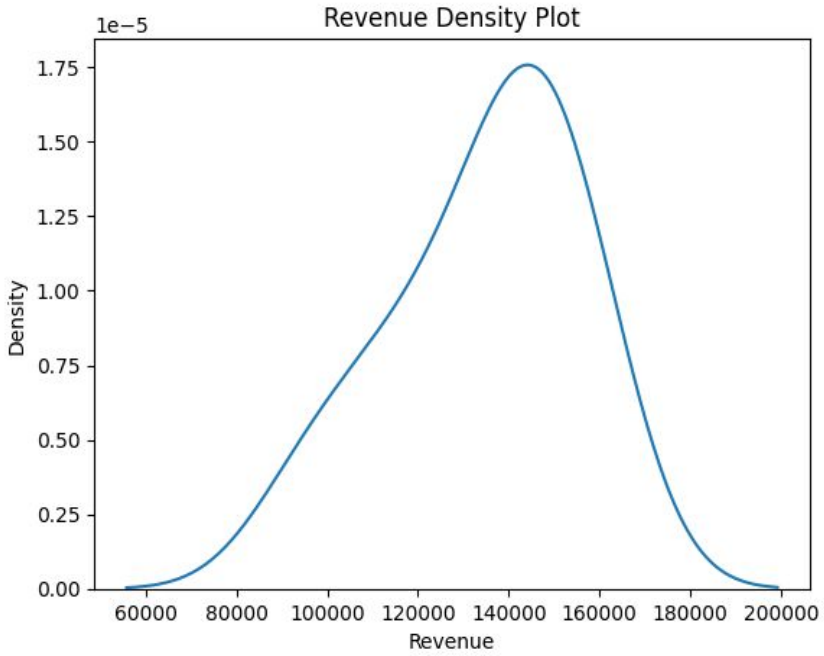
A KDE (Kernel Density Estimate) plot is a smoothed version of a histogram that shows the probability density function.

Use Cases:

- Understanding distribution trends
- Finding peaks and troughs in data

Insights:

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



Bivariate Analysis

Scatterplots

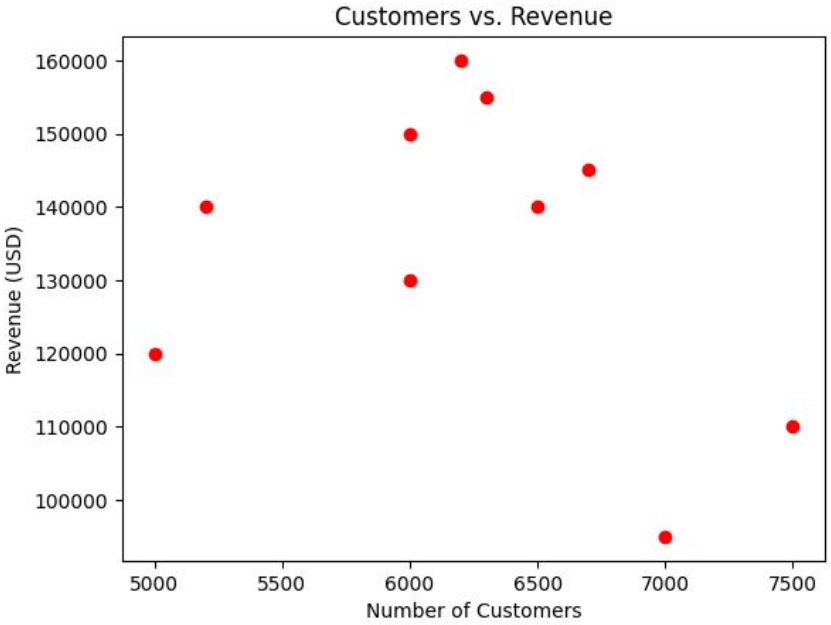
A scatter plot shows how two variables are related. Each point represents a data pair.

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



Line plots

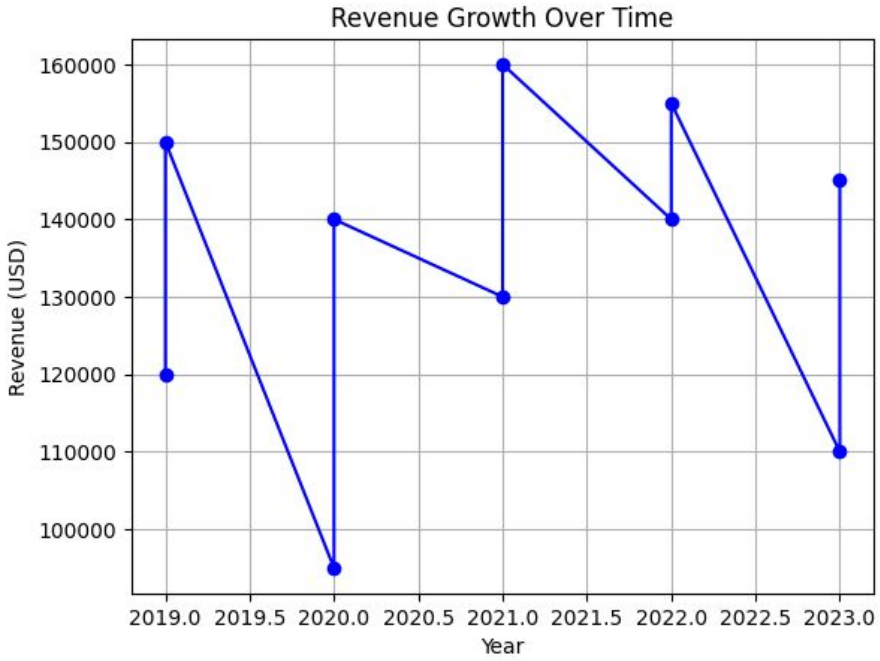
A line plot shows trends over time. It connects data points sequentially.

Use Cases:

- Analyzing revenue growth over years
- Identifying seasonal patterns

Insights:

- Shows revenue trends over years
- Detects dips or spikes in revenue



Bar plots

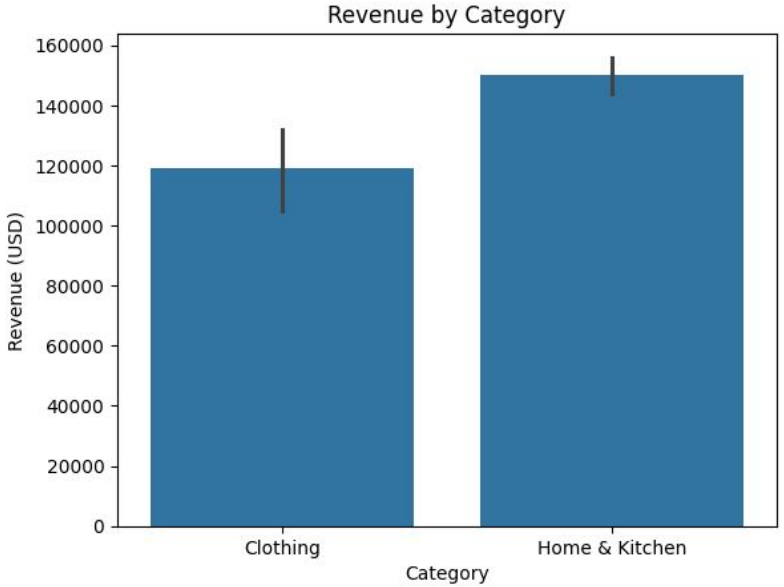
A bar plot compares values across different categories.

Use Cases:

- Comparing revenue between product categories
- Identifying top-performing segments

Insights:

- Shows which category generates higher revenue
- Highlights performance differences



Multivariate Analysis

Heatmaps

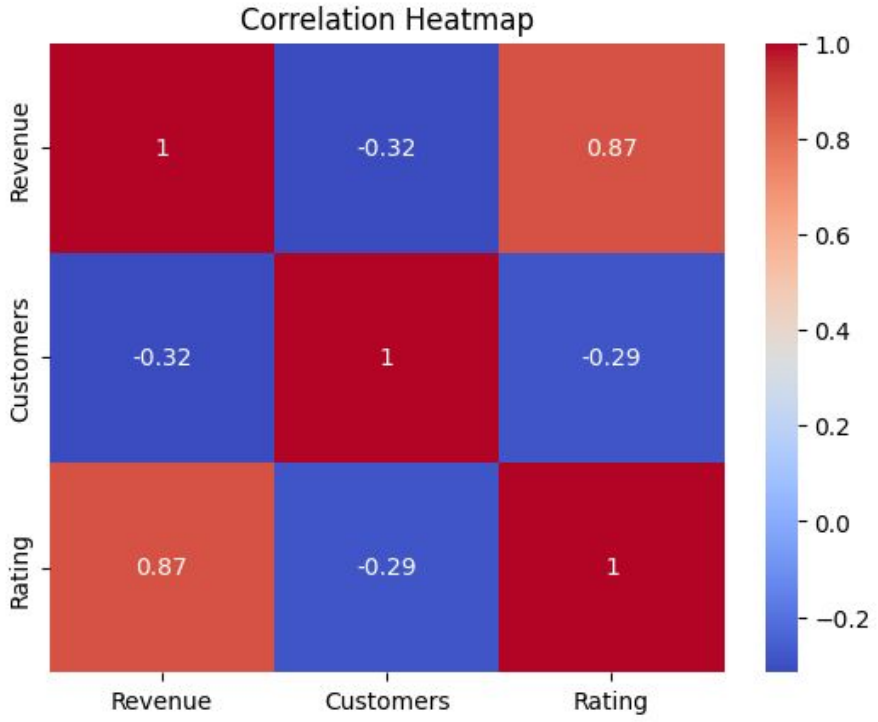
A heatmap visualizes correlations between multiple variables.

Use Cases:

- Understanding how different variables relate to each other.
- Identifying strong or weak correlations.

Insights:

- Shows if revenue correlates with customer count or ratings.
- Identifies strong positive or negative correlations.



Pairplots

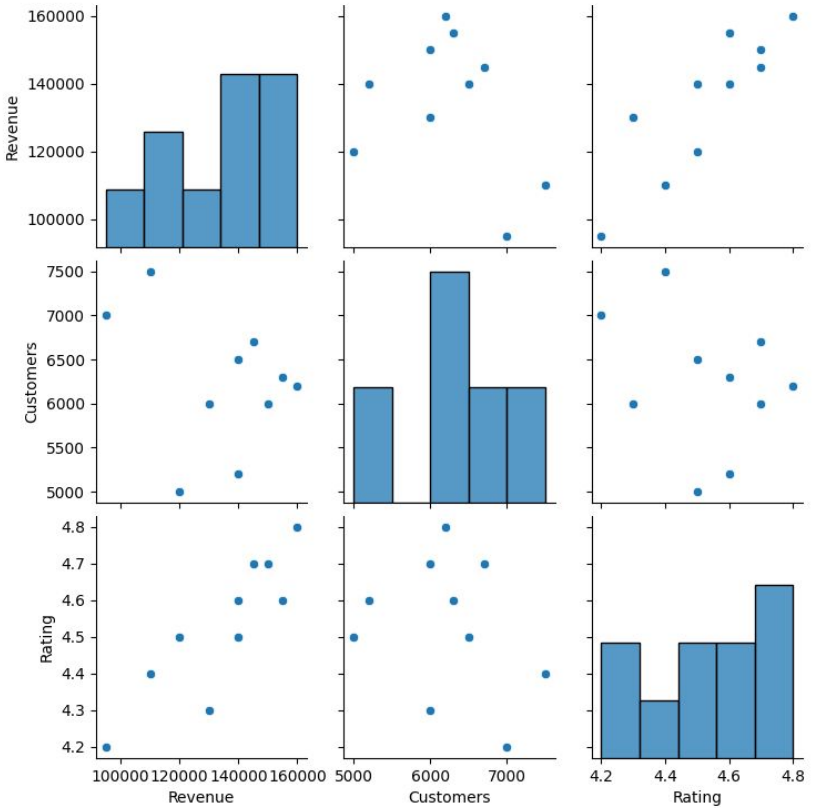
A pairplot creates scatter plots for all numeric variable combinations.

Use Cases:

- Exploring variable relationships.
- Detecting clustering patterns.

Insights:

- Visualizes how variables interact.
- Highlights potential outliers.



3D plots

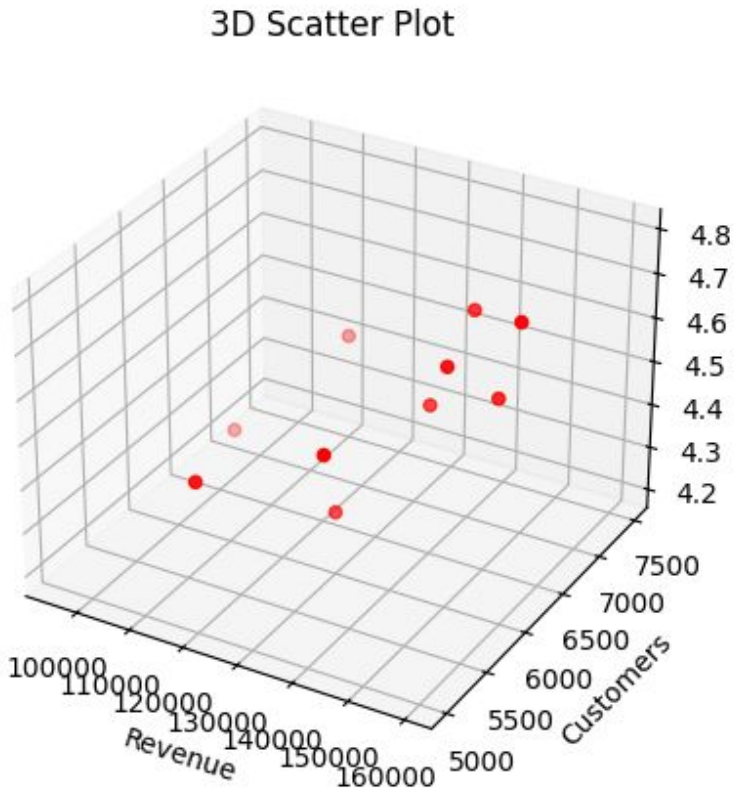
A 3D scatter plot adds a third variable to a standard scatter plot.

Use Cases:

- Showing interactions between three numerical variables.

Insights:

- Displays trends in three-dimensional space.
- Shows how customer count, revenue, and rating interact.



Practical cases

Useful Links

Matplotlib 3.10.3 documentation

Seaborn. User guide and tutorial

Pandas. Chart visualization

Plotly Open Source Graphing Library for Python

Q&A