Lesson 16: Correlation Analysis

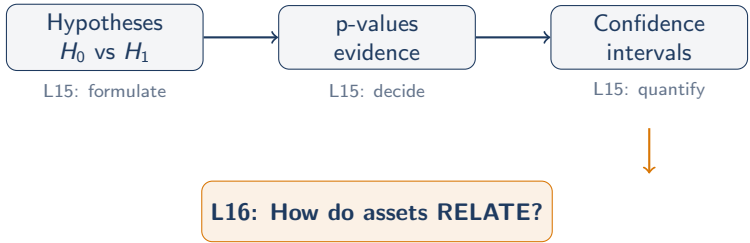
Data Science with Python – BSc Course

Data Science Program

BSc Course

45 Minutes

Previously on Data Science...



L15 asked: "Is this effect real?" **L16 asks:** "How do variables move *together*?"

Correlation measures the strength and direction of relationships between variables

Learning Objectives

After this lesson, you will be able to:

1. Explain the difference between Pearson and Spearman correlation
2. Interpret correlation values on the -1 to $+1$ scale
3. Create and read correlation heatmaps
4. Analyze rolling correlation in financial time series
5. Apply correlation to portfolio diversification

Correlation is the bridge from single-variable statistics to multivariate analysis

The 2008 Problem

In 2008, assets that were "uncorrelated" suddenly moved together.

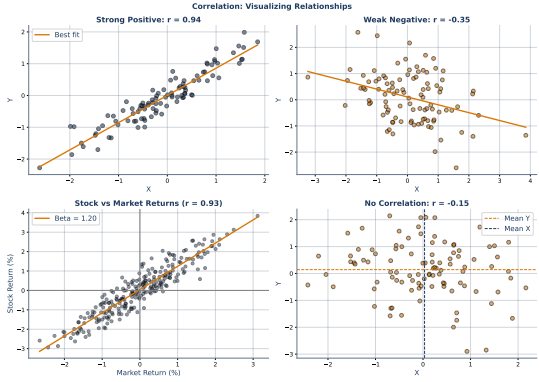
- Portfolios built on historical correlations collapsed
- Diversification benefits vanished overnight
- "Once-in-a-century" losses hit multiple asset classes simultaneously

Two questions for today:

1. How do we *measure* how assets move together?
2. Why do those measurements *break* in a crisis?

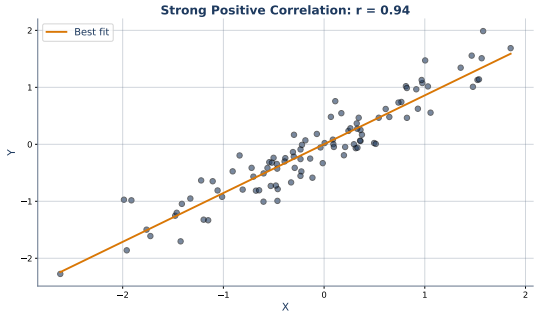
Correlation is the GPS of portfolio construction – but GPS can fail in a storm

Correlation at a Glance



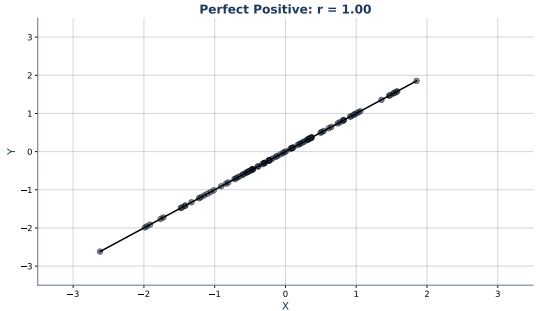
Scatter plots reveal both direction and strength of linear relationships

Strong Positive Correlation



r close to +1: when one variable rises, the other rises too

The Correlation Scale: -1 to $+1$



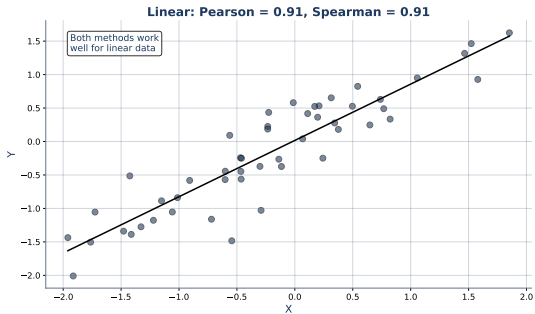
- $r = +1$: perfect positive (points on a line, upward slope)
- $r = 0$: no linear relationship
- $r = -1$: perfect negative (points on a line, downward slope)

Correlation is unitless – compare relationships across any variables

Pearson Correlation: Linear Relationships

Formula:

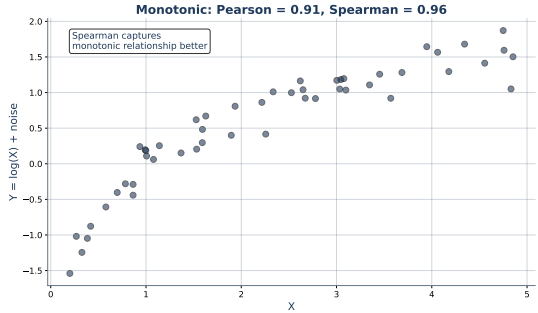
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



Pearson captures how well a straight line fits the data

Spearman Correlation: Monotonic Relationships

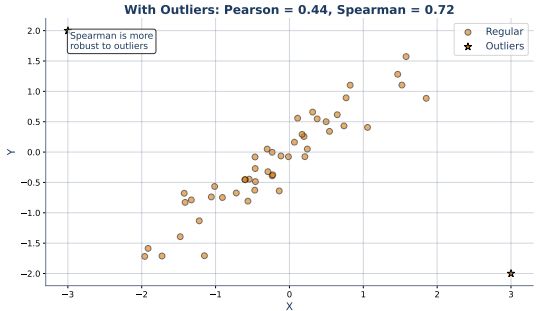
Method: Convert to ranks, then apply Pearson to the ranks.



- Detects monotonic (consistently increasing/decreasing) patterns
- Robust to outliers – ranks dampen extreme values

Use Spearman when data has outliers or non-linear monotonic trends

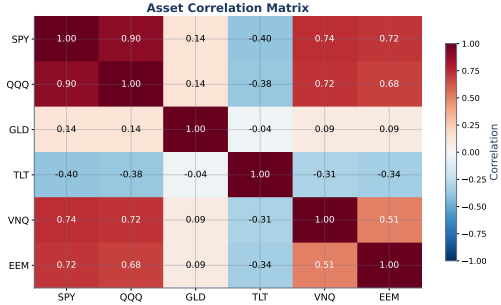
Outlier Effects on Correlation



- A single outlier can drastically change Pearson's r
- Spearman's ρ remains stable – ranks absorb extreme values

Always visualize your data before trusting a correlation number

Correlation Heatmap




```
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
```

Heatmaps reveal correlation structure across many variables at once

Checkpoint: Think About This

Quick Check

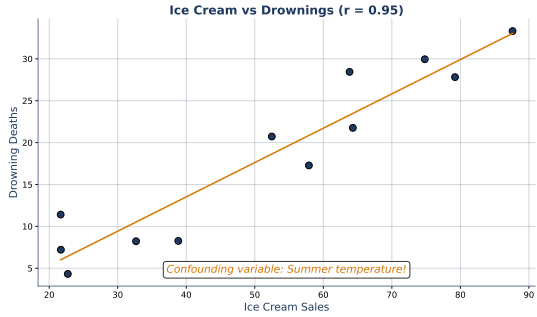
Ice cream sales and drowning deaths have correlation $r = 0.80$ in summer months.

Does ice cream cause drowning?

The hidden variable: hot weather drives *both* ice cream sales and swimming activity.

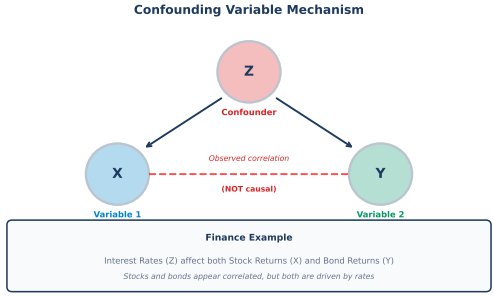
Confounders create correlations where no direct causal link exists

Spurious Correlation: Confounders



A third variable can create a strong correlation between unrelated variables

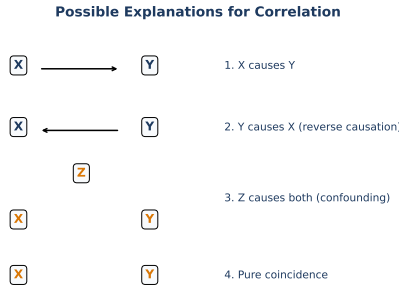
Confounding Variable Diagram



- Variable C drives both X and Y independently
- X and Y appear correlated but share no direct link

Always ask: what else could explain this relationship?

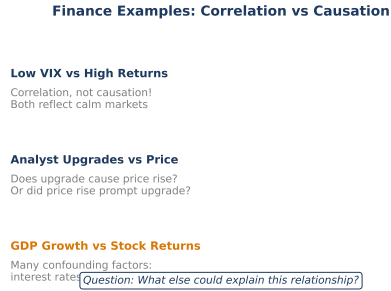
Correlation $\neq$ Causation



- Correlation: two variables move together
- Causation: one variable *makes* the other change
- Establishing causation requires experiments or strong controls

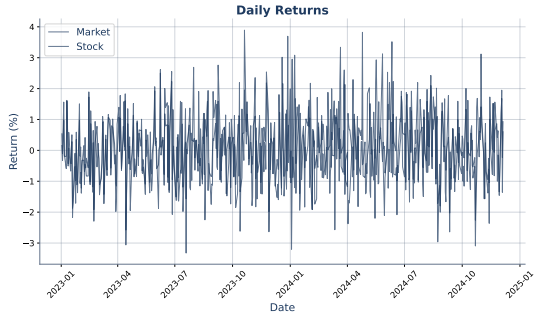
This is the single most important distinction in data analysis

Finance: Correlation Traps



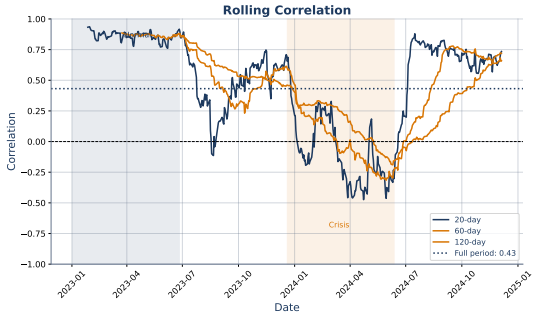
In finance, common causes (market regime, macro shocks) drive many apparent relationships

Stock Returns: Setting Up the Analysis



Two asset return series – do they move together?

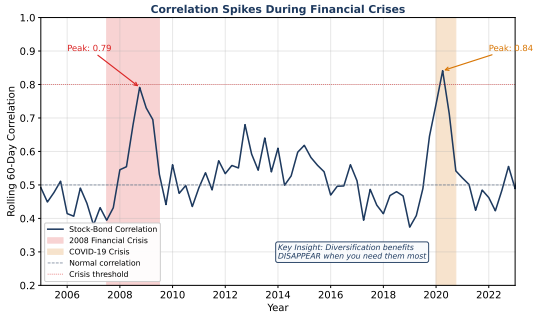
Rolling Correlation: Relationships Change



- A single correlation number hides regime changes
- Rolling windows (e.g., 60 days) reveal how relationships evolve

Static correlation is a snapshot – rolling correlation is the movie

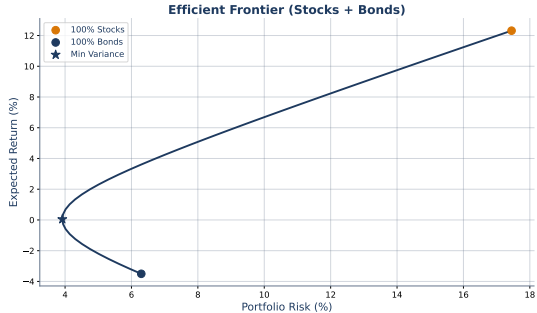
Crisis Correlation: When Diversification Fails



- In calm markets, correlations stay low (diversification works)
- In crises, correlations spike toward 1.0 (everything falls together)

You need diversification most when it works least – the correlation paradox

Efficient Frontier: Correlation in Action



Key insight: When $\rho < 1$, portfolio risk $<$ weighted average risk.
Lower correlation $\rightarrow$ more "bend" in the frontier $\rightarrow$ better diversification.

The efficient frontier is built entirely on correlation structure

Hands-On: Correlation Analysis

Task: Compute and visualize correlations for a stock portfolio.

1. Load daily returns for 4 stocks from different sectors
2. Compute Pearson and Spearman correlation matrices
3. Create a heatmap with `sns.heatmap(..., annot=True)`
4. Calculate 60-day rolling correlation for one pair
5. Identify the period with highest correlation – what happened?

Stretch goal: Compare rolling correlations in 2008 vs 2019.

Hands-on: 10 minutes – use the Colab notebook for starter code

The Elastic Band

"Correlated in calm, tangled in crisis."

The elastic band metaphor: correlation tightens under stress when you need slack the most

Key Takeaways

What you now know:

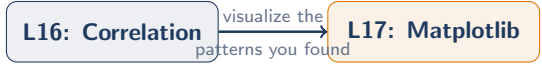
1. **Pearson** measures linear association; **Spearman** measures monotonic
2. Correlation ranges from -1 to $+1$ – magnitude *and* direction matter
3. **Correlation** $\neq$ **causation** – always look for confounders
4. Rolling correlation reveals how relationships *change over time*
5. Lower correlation $\rightarrow$ better diversification (until a crisis hits)

Correlation is the thread that connects every asset in your portfolio

Coming Up: L17 – Matplotlib Basics

You have been reading charts all module. Now you build your own.

- Create line plots, bar charts, scatter plots from scratch
- Control colors, labels, annotations, and layout
- Build publication-quality figures for reports



L17 gives you the tools to create every chart you have seen in L13–L16