Lesson 16: Correlation Analysis

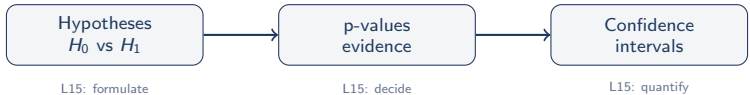
Data Science with Python – BSc Course

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.

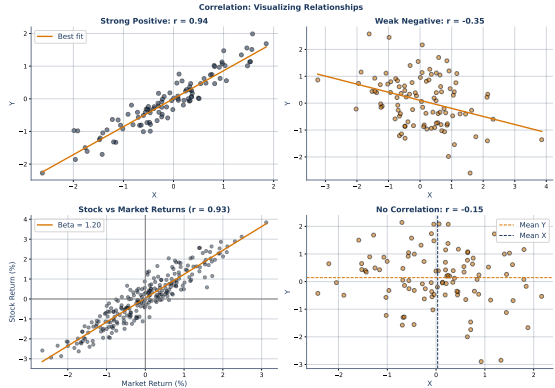
- Portfolios built on historical correlations collapsed
- Diversification benefits vanished overnight
- "Once-in-a-century" losses hit multiple asset classes simultaneously

Two questions for today:

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

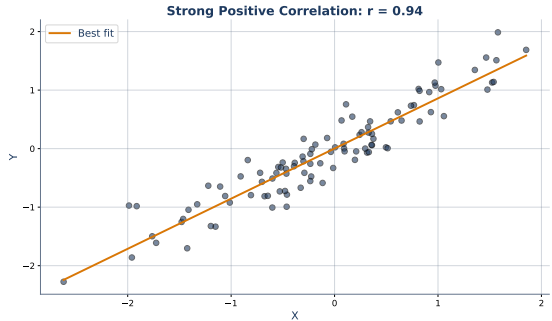
Correlation is the GPS of portfolio construction – but GPS can fail in a storm

Correlation at a Glance



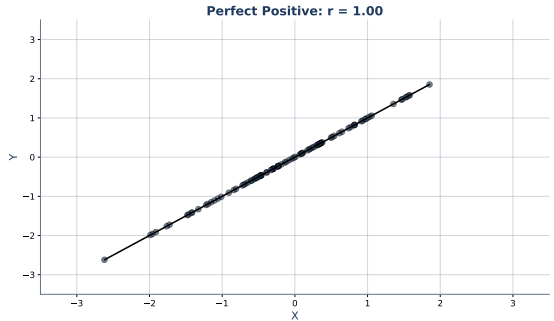
Scatter plots reveal both direction and strength of linear relationships

Strong Positive Correlation



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

The Correlation Scale: -1 to $+1$



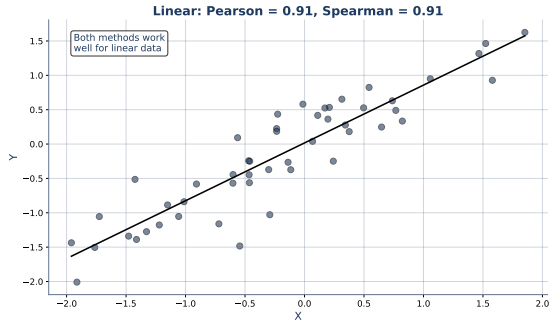
- $r = +1$: perfect positive (points on a line, upward slope)
- $r = 0$: no linear relationship
- $r = -1$: perfect negative (points on a line, downward slope)

Correlation is unitless – compare relationships across any variables

Pearson Correlation: Linear Relationships

Formula:

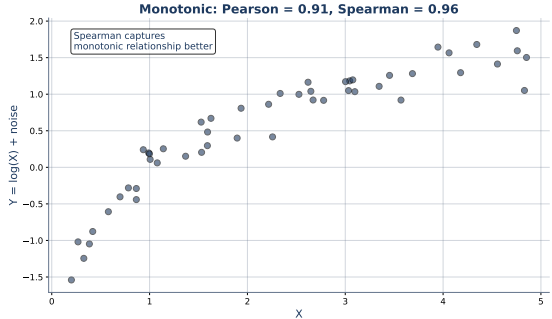
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



Pearson captures how well a straight line fits the data

Spearman Correlation: Monotonic Relationships

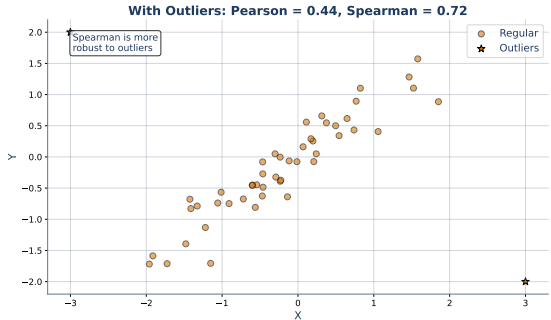
Method: Convert to ranks, then apply Pearson to the ranks.



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

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

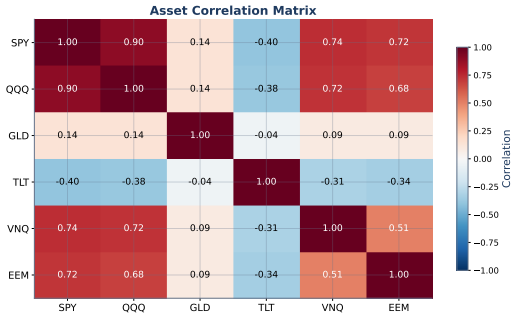
Outlier Effects on Correlation



- 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

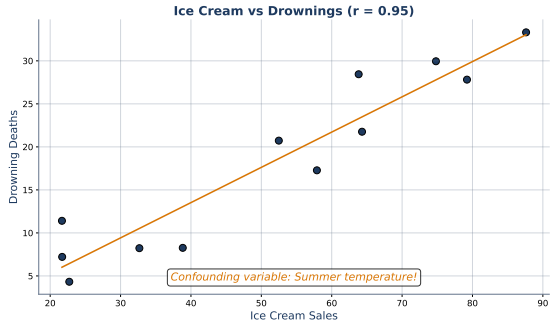
Ice cream sales and drowning deaths have correlation $r = 0.80$ in summer months.

Does ice cream cause drowning?

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

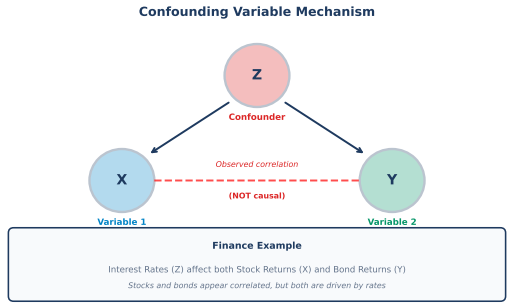
Confounders create correlations where no direct causal link exists

Spurious Correlation: Confounders



A third variable can create a strong correlation between unrelated variables

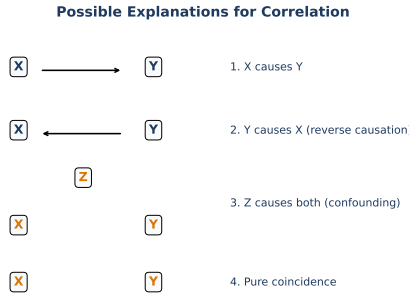
Confounding Variable Diagram



- 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

Finance Examples: Correlation vs Causation

Low VIX vs High Returns

Correlation, not causation!
Both reflect calm markets

Analyst Upgrades vs Price

Does upgrade cause price rise?
Or did price rise prompt upgrade?

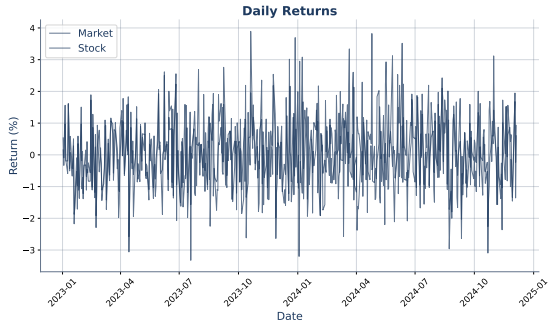
GDP Growth vs Stock Returns

Many confounding factors:
interest rates

Question: What else could explain this relationship?

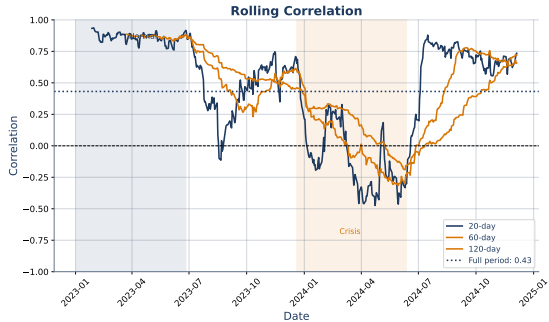
In finance, common causes (market regime, macro shocks) drive many apparent relationships

Stock Returns: Setting Up the Analysis



Two asset return series – do they move together?

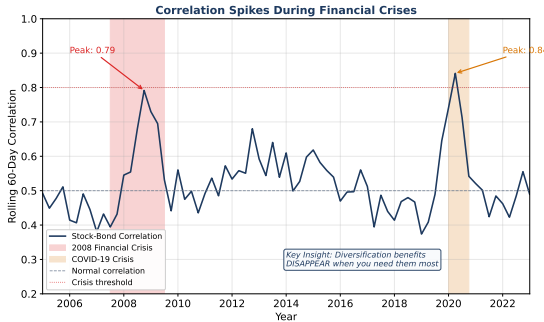
Rolling Correlation: Relationships Change



- 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

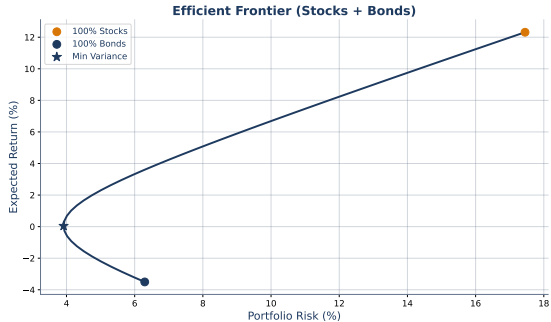
Crisis Correlation: When Diversification Fails



- 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

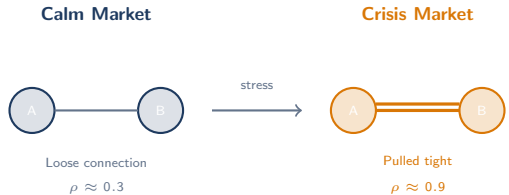
Task: Compute and visualize correlations for a stock portfolio.

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:

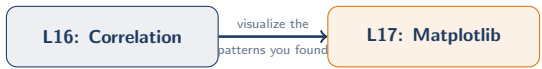
1. **Pearson** measures linear association; **Spearman** measures monotonic
2. Correlation ranges from -1 to $+1$ – magnitude *and* direction matter
3. **Correlation** $\neq$ **causation** – always look for confounders
4. Rolling correlation reveals how relationships *change over time*
5. Lower correlation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

Correlation is the thread that connects every asset in your portfolio

Coming Up: L17 – Matplotlib Basics

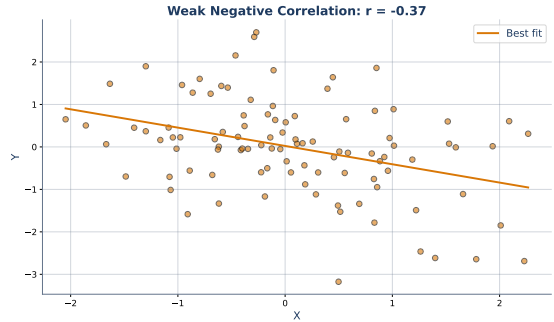
You have been reading charts all module. Now you build your own.

- 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

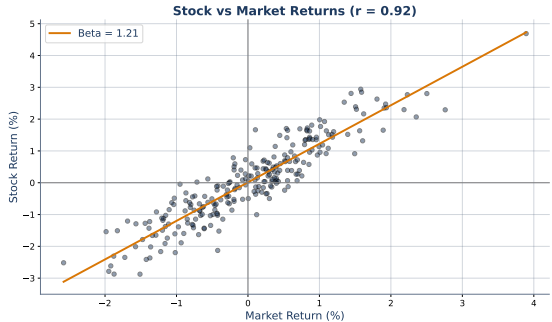
Self-Study: Weak Negative Correlation



- $r \approx -0.3$: variables tend to move in opposite directions, weakly
- Common in hedging pairs that partially offset each other

Weak correlations are noisy – large samples needed to confirm

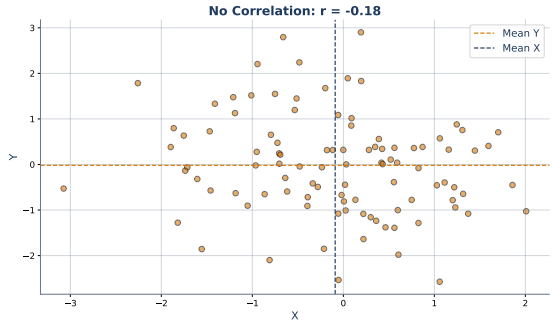
Self-Study: Finance Correlation Example



- Real stock correlations vary by sector and market regime
- Same-sector stocks tend to have higher correlation

Sector membership is one of the strongest drivers of stock correlation

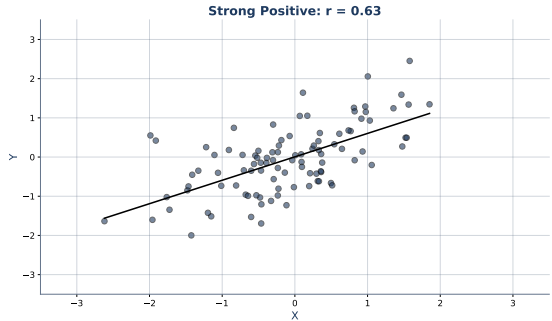
Self-Study: No Correlation



- $r \approx 0$: no linear pattern visible
- Does not mean the variables are independent – could be non-linear

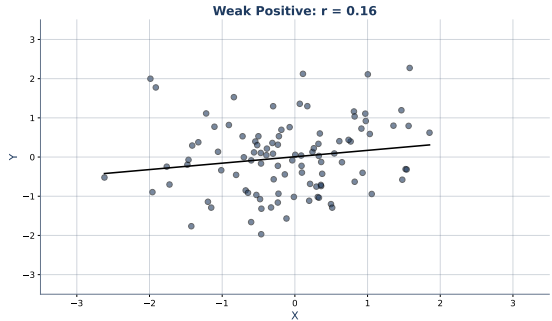
Zero correlation rules out linear links but not all relationships

Self-Study: Correlation Scale – Strong Positive



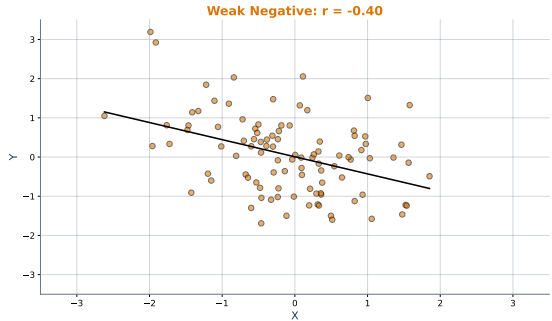
r = +0.8: strong positive – most points cluster tightly around upward trend

Self-Study: Correlation Scale – Weak Positive



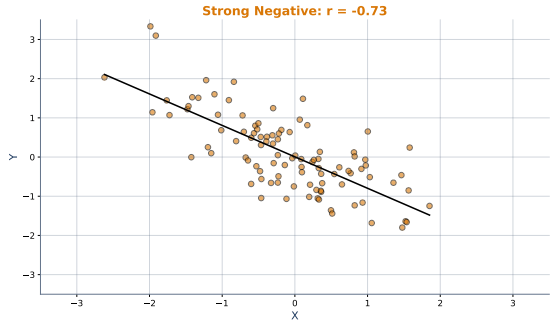
r = +0.3: weak positive – noisy cloud with slight upward trend

Self-Study: Correlation Scale – Weak Negative



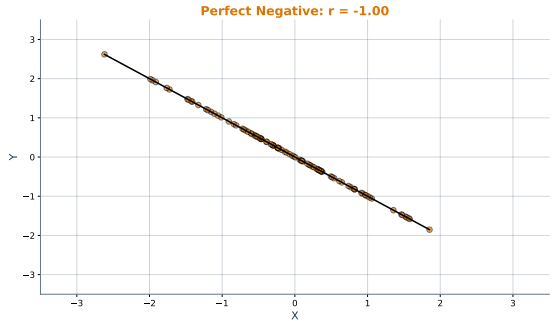
r = -0.3: weak negative – noisy cloud with slight downward trend

Self-Study: Correlation Scale – Strong Negative



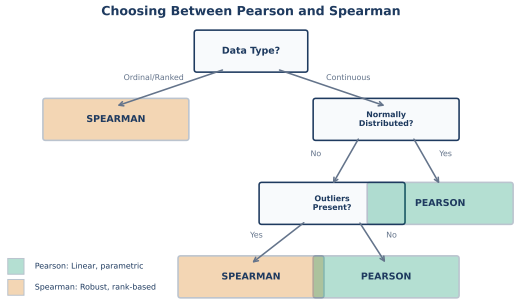
r = -0.8: strong negative – tight cluster around downward trend

Self-Study: Correlation Scale – Perfect Negative



r = -1.0: all points lie exactly on a line with negative slope

Self-Study: Pearson vs Spearman Comparison



- Pearson: linear relationships, assumes normality
- Spearman: monotonic relationships, distribution-free

Use Pearson as default; switch to Spearman for non-normal or ordinal data

Self-Study: Correlation Methods Summary

Pearson vs Spearman: When to Use Each

Pearson:

- Measures LINEAR relationship
- Assumes normal distribution
- Sensitive to outliers
- Use: regression, portfolio optimization

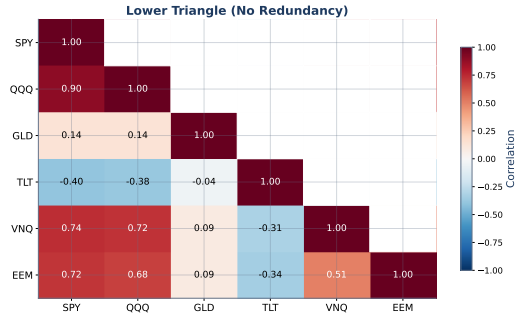
Spearman:

- Measures MONOTONIC relationship
- No distribution assumption
- Robust to outliers
- Use: rankings, ordinal data

`df.corr(method="pearson")` vs `df.corr(method="spearman")`

Quick reference: which method for which data type

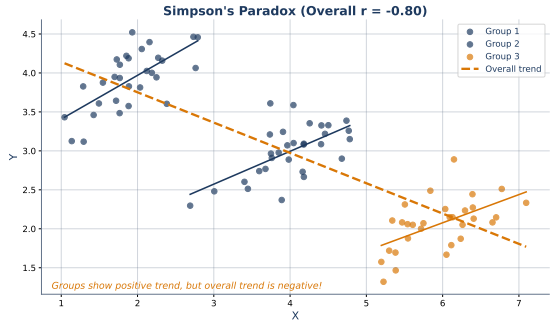
Self-Study: Lower Triangle Heatmap



- Mask the upper triangle to remove redundancy
- `mask = np.triu(np.ones_like(corr, dtype=bool))`

Professional reports use lower-triangle heatmaps for cleaner presentation

Self-Study: Simpson's Paradox

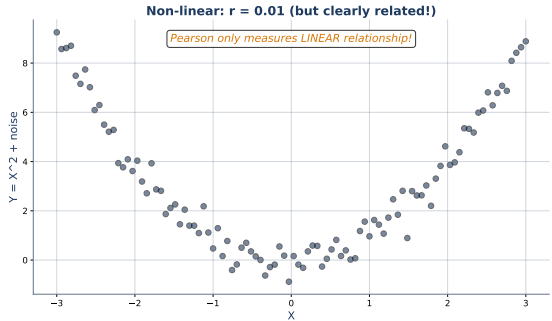


- Aggregate correlation can reverse at subgroup level
- Always check if groups in your data have different patterns

---

Simpson's paradox is a reminder to look beyond the overall number

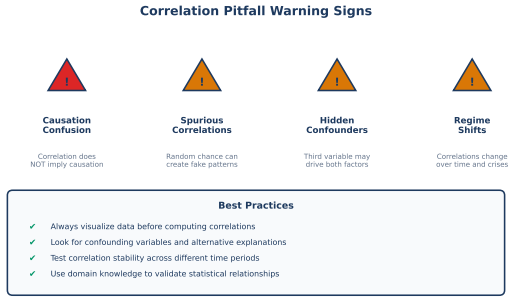
Self-Study: Nonlinear Relationships



- Pearson $r \approx 0$ even though a clear pattern exists
- Always plot your data – correlation misses non-linear structure

Zero correlation does not mean no relationship

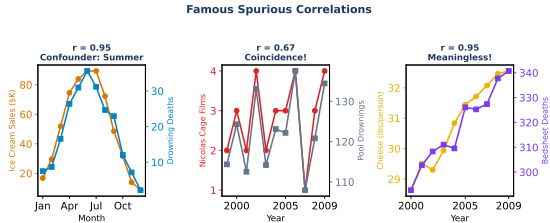
Self-Study: Correlation Pitfalls



- Outliers, non-linearity, subgroups, and restricted range all distort r

Visualize first, compute second – never trust a correlation number alone

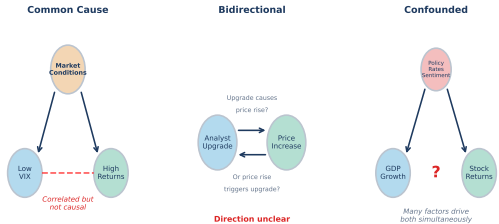
Self-Study: Spurious Correlation Examples



- Nicolas Cage films and pool drownings: $r = 0.87$
- Spurious correlations arise from coincidence, shared trends, or small samples

Thousands of variable pairs will show high correlation by chance alone

Self-Study: Finance Causation Examples

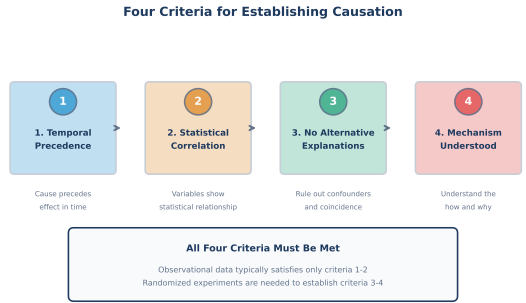


- Low VIX and high returns: both reflect calm markets (common cause)
- Analyst upgrades and price: direction of causation is ambiguous

---

In finance, correlation traps are everywhere – always consider the mechanism

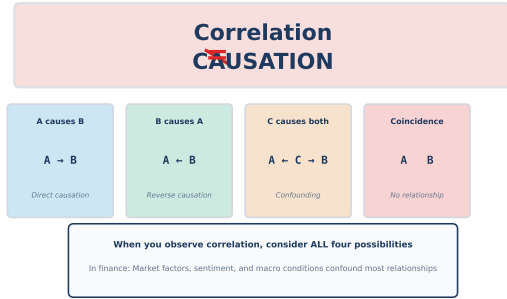
Self-Study: Establishing Causation



- Temporal precedence, no confounding, plausible mechanism
- Randomized experiments are the gold standard (rare in finance)

Establishing causation from observational data is one of the hardest problems in statistics

Self-Study: Causation Key Takeaway

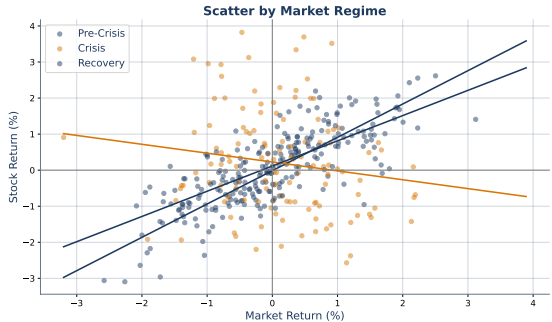


- Always look for confounders and reverse causation
- In finance, be especially skeptical of causal claims from correlations

---

Correlation is a starting point, not an ending point

Self-Study: Scatter Plot by Market Regime



- Calm periods: loose scatter, low correlation
- Stress periods: tight cluster, high correlation

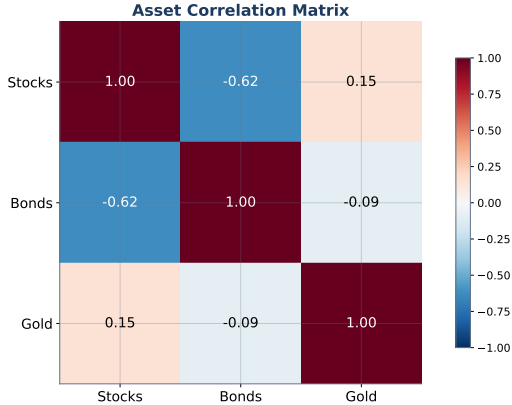
Regime-conditional correlation tells a richer story than a single number

Self-Study: Rolling Correlation Implications

- Monitor correlation stability for risk management
- Consider regime-dependent models (e.g., DCC-GARCH)

Diversification benefits vary over time – static models miss this

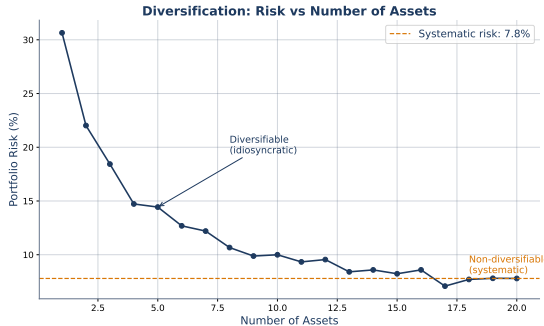
Self-Study: Portfolio Correlation Matrix



- Multi-asset correlation structure drives portfolio risk
- Look for low-correlation pairs to improve diversification

The correlation matrix is the core input to mean-variance optimization

Self-Study: Diversification and Risk Reduction

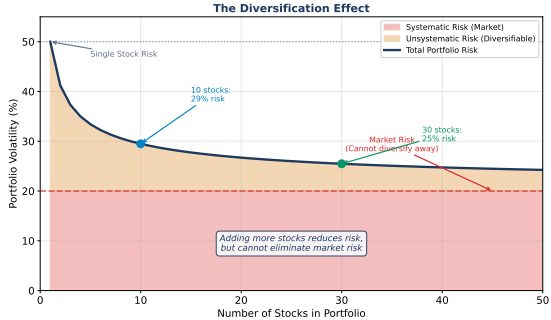


- Lower correlation = greater risk reduction from combining assets
- Perfect negative correlation ($\rho = -1$) can eliminate risk entirely

---

The diversification benefit depends entirely on correlation

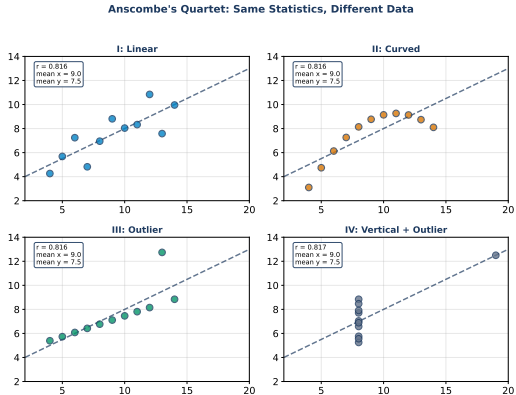
Self-Study: Diversification Curve



- Adding assets reduces portfolio risk – but with diminishing returns
- Most diversification benefit comes from the first 15–20 assets

Beyond 30 stocks, additional diversification is minimal in practice

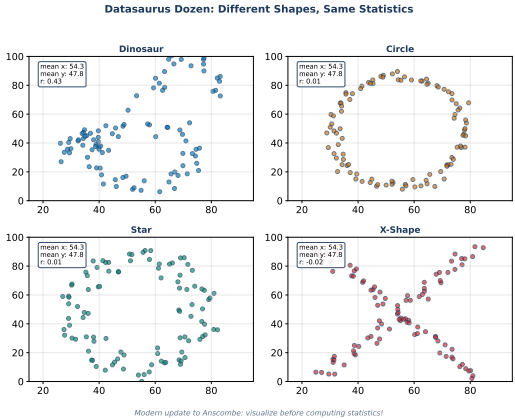
Self-Study: Anscombe's Quartet



- Four datasets with identical r, mean, and variance
- Completely different patterns – only visible when plotted

The original argument for "always visualize your data"

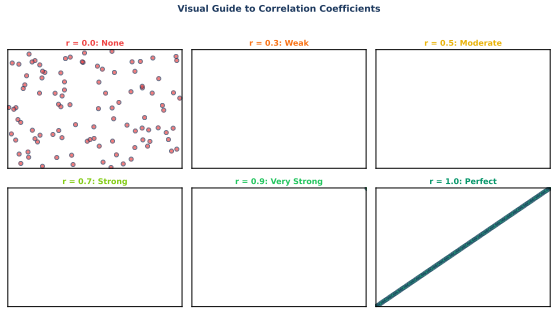
Self-Study: Datasaurus Dozen



- 12 datasets with identical summary statistics but wildly different shapes
- A modern, dramatic extension of Anscombe's point

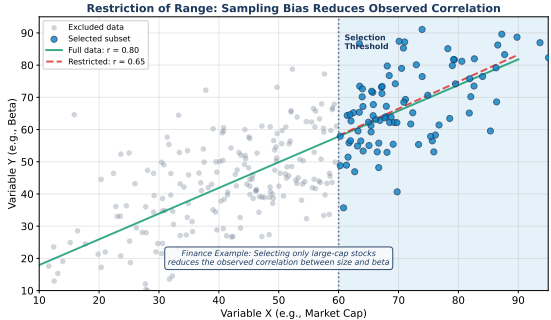
Summary statistics lie – visualization reveals the truth

Self-Study: Visual Correlation Scale



Reference card: what different correlation values look like as scatter plots

Self-Study: Restriction of Range

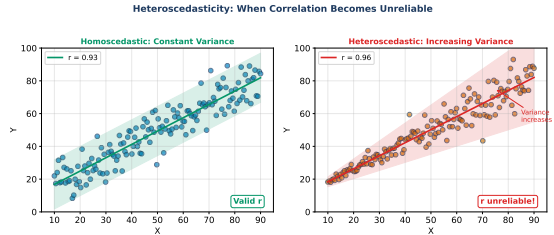


- Truncating data range artificially lowers correlation
- Common trap: analyzing only large-cap stocks, only recent data

---

Always check whether your sample covers the full range of both variables

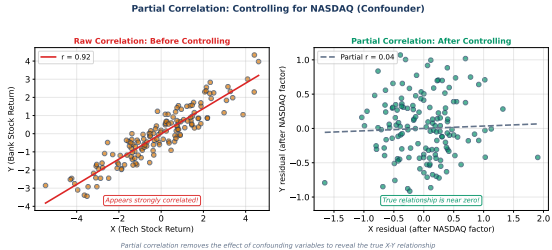
Self-Study: Heteroscedasticity



- Variance changes across the range of X (fan shape)
- Correlation still works but may not tell the full story

Non-constant variance suggests the relationship is more complex than a single number

Self-Study: Partial Correlation



- Removes the effect of confounding variables
- Partial $r_{XY|Z}$: correlation between X and Y after controlling for Z
- Python: `pingouin.partial_corr(data, x, y, covar)`

Partial correlation is the first step toward causal reasoning from observational data