Correlation

- Correlation is a **statistical measure of association** that describes how strongly and in what direction two variables are related.
- Correlation ≠ Causation

Warm Up

- Over the summer of 2025 we measured ice cream sales and drowning events in Key West
- Key Lime flavored ice cream and drowning events both went up over the summer compared to the winter
- Are Key Lime ice cream and drowning related?

Obviously Not

- You will see lots of examples where people over interpret correlation
- If one thing goes up (or down) and the other goes up (or down) then they must be related.
- Not true! Always think it through

Core concepts

- Measures association between two variables
- Numeric range: –1 (perfect negative) to +1 (perfect positive)
  - True of most (if not, all) measures of correlation (there are a lot of different types)

Correlation and Data Science

- Correlation measures the **degree of association** between two (or more) variables.
- It tells us how much information one variable carries about another.
- Useful for:
  - Detecting redundancy in features (highly correlated predictors).
  - Identifying candidate relationships for feature selection and feature engineering.
  - Exploring relationships between independent ↔ dependent variables.
- At its heart: correlation is about shared information content and whether variables move together in a systematic way.

UMBC

Feature Engineering & Selection

- Redundant features
  - If two independent variables are highly correlated, they contain overlapping information.
  - Example: “height in inches” and “height in cm” → drop one.
- Multicollinearity in modeling
  - Strongly correlated predictors can distort regression coefficients.
  - Example: “age” and “years since college” in a salary model.
- Feature reduction
  - Correlation heatmaps can guide which variables to keep or combine.
  - Example: many correlated survey items → reduce with PCA.
- Creating new features
  - Weakly correlated features may be combined to capture interaction.
  - Example: “hours studied” and “sleep” may individually correlate weakly with GPA, but together have stronger predictive power.
- Correlation with target variable
  - Helps prioritize variables for exploration.
  - Example: checking which predictors are most associated with churn (dependent variable).

Types of Correlation

- Comes in different *flavors* depending on:
  - **Shape of relationship** (linear vs. nonlinear).
  - **Data type** (continuous, ordinal, binary).
  - **Assumptions** (parametric vs. non-parametric).

Types of Correlation 2

- Pearson's r (parametric)
  - Measures linear association between two continuous variables.
  - Sensitive to outliers.
- Spearman's ρ (rank-based)
  - Measures monotonic association using ranked data.
  - Works with ordinal data, robust to outliers.
- Kendall's τ (pairwise concordance)
  - Based on agreement/disagreement of pairs.
  - More interpretable in small samples or with ties.
- Point-Biserial
  - One variable continuous, one binary (0/1).
  - Example: gender (binary) vs. test score.
- Partial Correlation
  - Correlation between two variables controlling for a third (or more).
  - Useful for handling confounding variables.

- Why Different Correlation Types?
  - Different data, different tools → continuous, ordinal, binary, or confounded variables each need their own measure.
  - Shape matters → Pearson only sees linear; Spearman/Kendall catch monotonic curves.
  - Avoid misinterpretation → the "right" method can reveal strong links that look weak otherwise.
- Bottom line: The right correlation type helps you find and interpret real relationships.

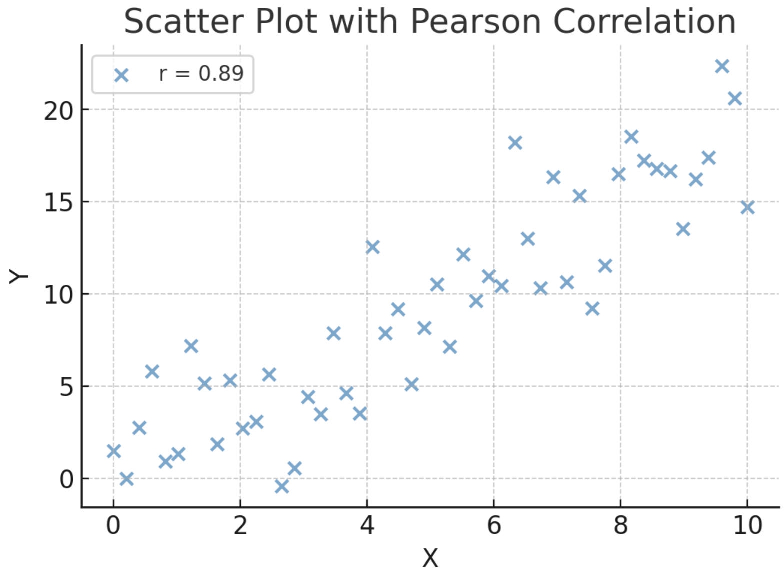
Pearson Correlation

- The most common type, *r*, ranges from -1 to 1
- For the correlation coefficient itself (<u>descriptive use</u>):
  - Linearity: The relationship between X and Y should be approximately linear.
  - Continuous variables: Both should be measured on an interval or ratio scale.
  - No significant outliers: Outliers can drastically inflate or deflate r
- For significance testing (<u>inference</u>):
  - Bivariate normality: The pair (X,Y) should follow a joint normal distribution.
  - Homoscedasticity: The spread of Y values is similar across the range of X (equal variance).
  - Independence of observations: Each pair is independent of others
- Which do we need in data science? Description or inference?

Pearson's Correlation Coefficient (r)

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

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$



Spearman's ρ

- **Definition:** Non-parametric measure of **monotonic association** between two variables
- Based on **ranks**, not raw values → robust to outliers & skewed data
- Captures increasing or decreasing trends (not just linear)

Spearman’s cont’d.

Spearman's Rank Correlation Calculation (ρ = 0.50)

X	Rank X	Y	Rank Y	d = RankX - RankY	d^2
10.0	1.0	15.0	1.0	0.0	0.0
20.0	2.0	40.0	4.0	-2.0	4.0
30.0	3.0	25.0	2.0	1.0	1.0
40.0	4.0	50.0	5.0	-1.0	1.0
50.0	5.0	35.0	3.0	2.0	4.0

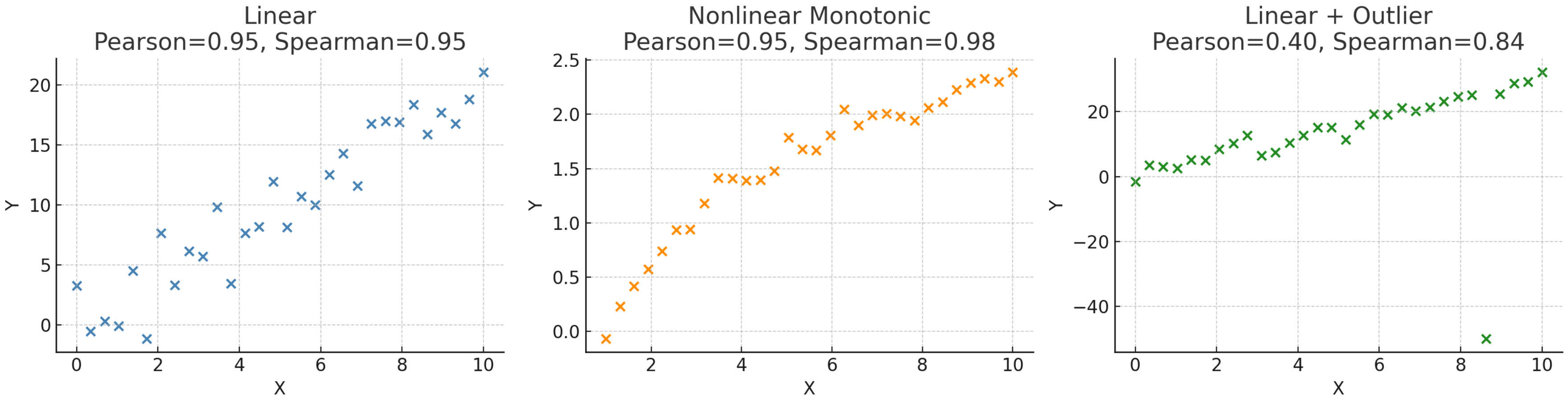
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$d_i = \mathrm{rank}(x_i) - \mathrm{rank}(y_i)$

$n = \text{number of pairs}$

$$\rho = \mathrm{Pearson}(\,\mathrm{rank}(X),\ \mathrm{rank}(Y)\,)$$

Spearman vs Pearson



Kendall’s Tau (τ) – Definition

- Non-parametric correlation measure
- Based on concordant vs. discordant pairs
- Concordant: for any two observations, if ranks of X and Y move in the same direction.
- Discordant: if ranks of X and Y move in opposite directions.

Kendall's Formula

$$\tau = \frac{C - D}{\binom{n}{2}}$$

- C = number of concordant pairs
- D = number of discordant pairs
- (n choose 2) = total number of pairs

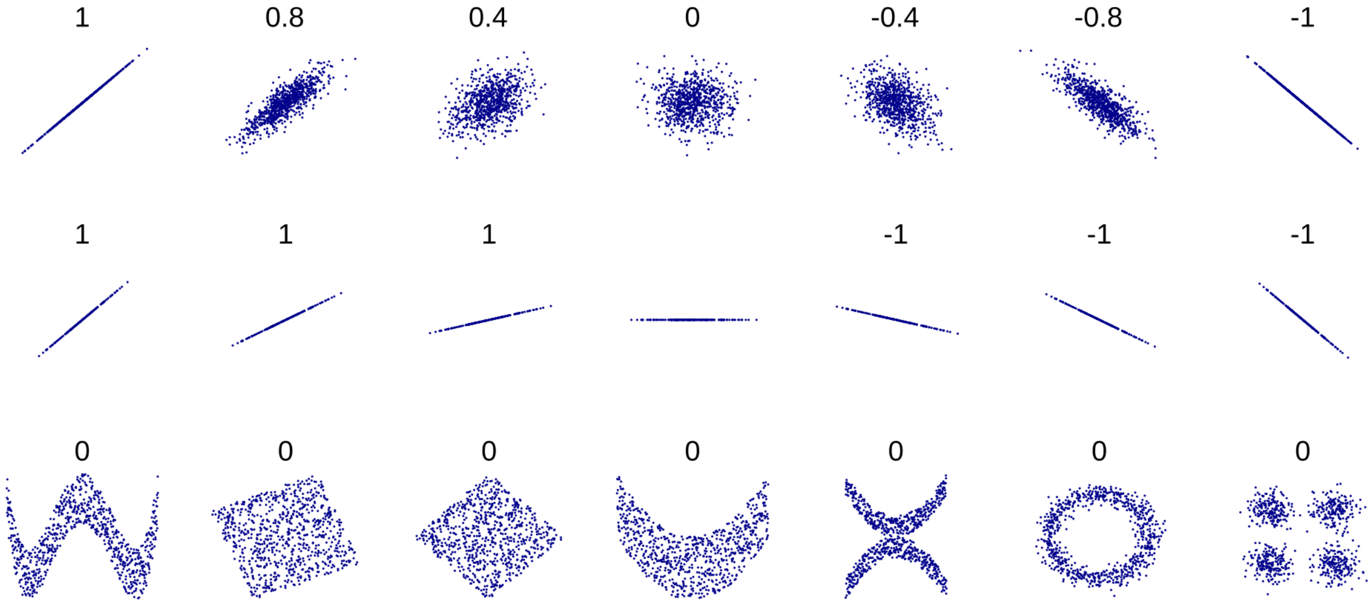
Kendall’s Derivation

Obs	X	Y
A	1	12
B	2	15
C	3	14
D	4	10

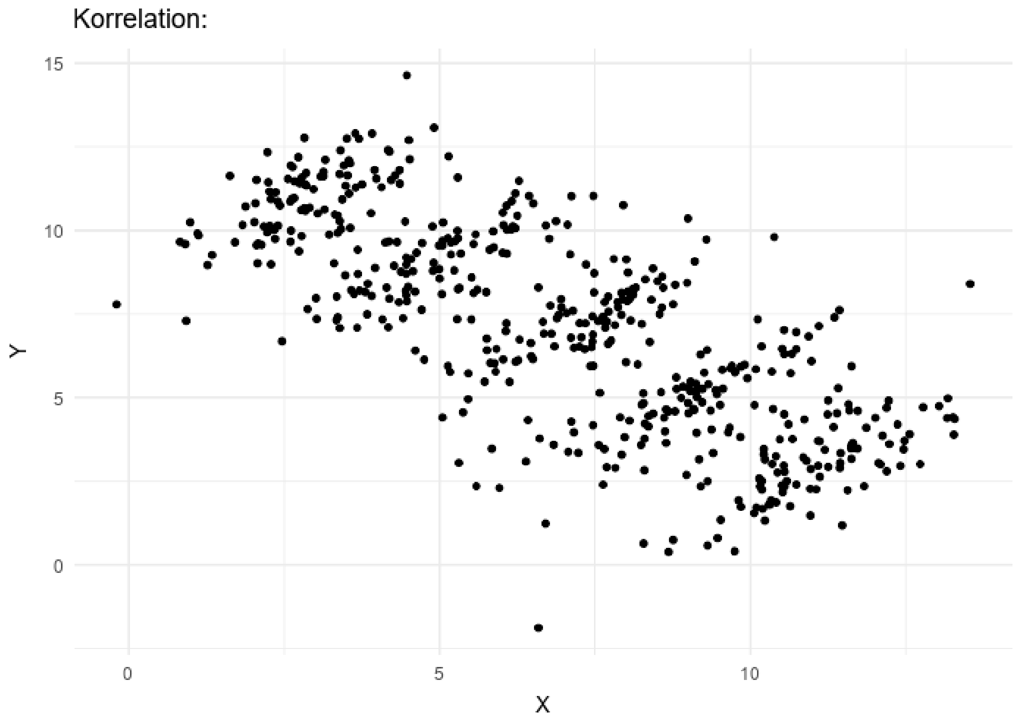
Pair	Compare X	Compare Y	Result
(A, B)	1 < 2	12 < 15	Concordant
(A, C)	1 < 3	12 < 14	Concordant
(A, D)	1 < 4	12 > 10	Discordant
(B, C)	2 < 3	15 > 14	Discordant
(B, D)	2 < 4	15 > 10	Concordant
(C, D)	3 < 4	14 > 10	Concordant

$$\tau = \frac{C - D}{\binom{n}{2}} = \frac{4 - 2}{6} = \frac{2}{6} = 0.333$$

Correlation Pitfalls



Sampson’s Paradox



Summary

Method	Data Type / Assumptions	Captures	Formula / Idea	Pros	Cons
Pearson's r	Continuous, linear, approx. normal	Linear association (strength & direction)	Covariance standardized by SDs	Simple, widely used, intuitive	Sensitive to outliers; misses nonlinear or monotonic-only trends
Spearman's ρ	Ordinal or continuous (nonlinear OK)	Monotonic association via ranks	Pearson's r on ranks, or 1−6∑d2n(n2−1)1 - \frac{6\sum d^2}{n(n^2-1)} (no ties)	Handles skew/outliers, good for large n, similar to Pearson	Less robust than Kendall in small nn; still influenced by big rank changes
Kendall's τ	Ordinal or continuous (robust to ties)	Pairwise agreement probability	τ=C−D(n2)\tau = \frac{C - D}{\binom{n}{2}}	Interpretable as probability, robust in small n, good with ties	Usually smaller values than Spearman; more conservative, lower power

Point Biserial Correlation

- Special case of Pearson's correlation.
- Used when:
  - One variable is continuous (e.g., exam score, height).
  - The other is binary/dichotomous (e.g., male/female, treatment/control, yes/no).
- Tells us whether the two groups (0 vs. 1) differ systematically on the continuous variable.
- Values range from –1 to +1, just like Pearson.

Point Biserial Formula

- $M1$ = mean of group coded "1"
- $M0$ = mean of group coded "0"
- s = standard deviation of all scores
- $n1$, $n0$ = group sample sizes
- n = total sample size

$$r_{pb} = \frac{M_1 - M_0}{s} \cdot \sqrt{\frac{n_1 n_0}{n^2}}$$

Point Biserial Example

Student	Group (0 = No, 1 = Yes)	Exam Score
A	0	72
B	0	68
C	0	75
D	1	85
E	1	90
F	1	88

$$r_{pb} = \frac{87.7 - 71.7}{8.35} \cdot \sqrt{\frac{3\times 3}{6^2}} \approx 0.96$$

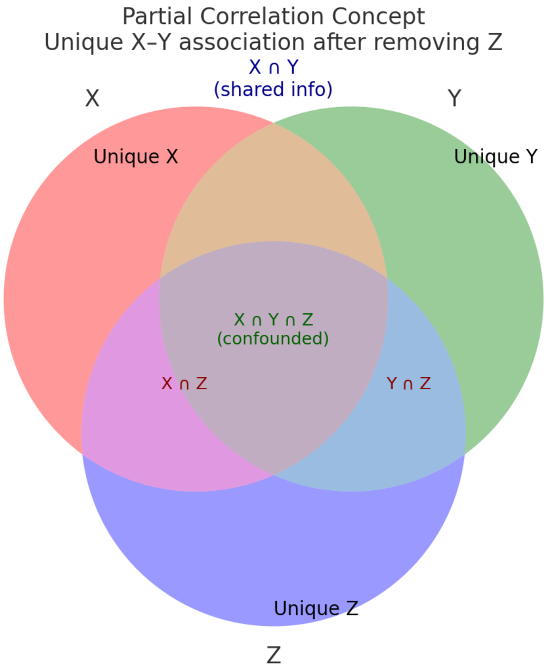
Partial Correlation

- Definition: Correlation between X and Y after removing the effect of a third variable Z.
- Controls for confounders → isolates the "direct" relationship.
- Bridges correlation ↔ causation ideas.

Partial Correlation and Data Science

- Multicollinearity: avoid redundant predictors.
- Causal thinking: helps distinguish spurious correlations.
- Model diagnostics: closer to regression coefficients.
- Example:
  - Shoe size ↔ Reading ability (correlated).
  - Both related to Age.
  - Partial correlation controlling for Age → near zero.

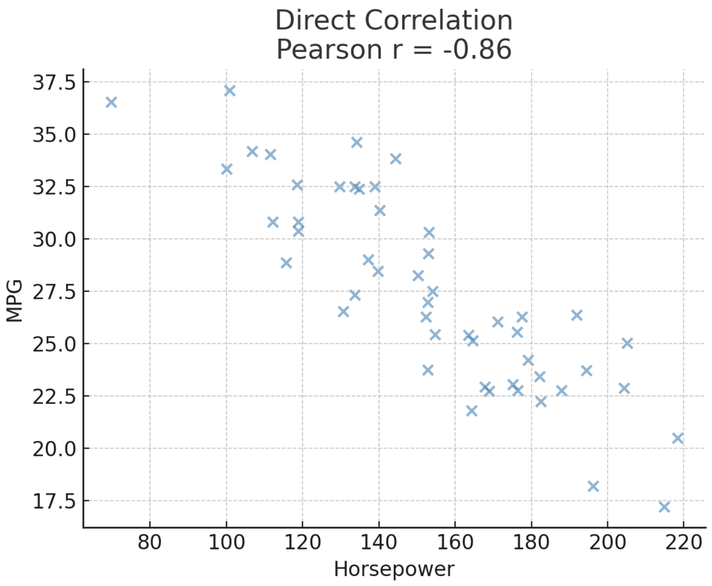
Concept and Formula

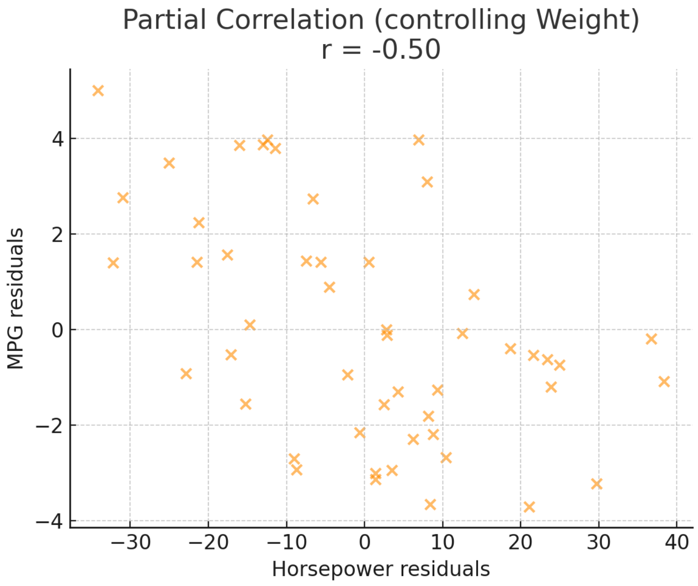


$$r_{XY\cdot Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

UMBC

In Practice





Correlation in Data Science Takeaways

- Correlation = information overlap between variables
  - Measures strength and direction of association
  - Helps detect redundancy, spurious patterns, and potential predictive power
- Different flavors for different data
  - Pearson: linear, continuous
  - Spearman & Kendall: monotonic, rank-based, robust
  - Point-biserial: binary variables
  - Partial: isolates association by controlling confounders
- Why it matters in Data Science
  - Guides feature selection & engineering
  - Detects multicollinearity in models
  - Forms the foundation for causal reasoning

Up Next (Next Lecture)

- More correlation measures & when to use them
- Hands-on: heatmaps and correlation matrices for multi-variable EDA
- Visual workflows: pair plots & feature redundancy checks
- Transition: how correlation ≠ causation → confounding