Housekeeping

- Materials posted on BB (waiting for library)
- Will grade first 2 HWs this week

Correlation and Causation

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

Recap

- Pearson's r – linear, continuous data
- Spearman's ρ – rank-based, monotonic
- Kendall's τ – concordant/discordant pairs, robust in small samples
- Point-Biserial – continuous vs. binary
- Partial Correlation – controls for confounders

Correlation in Data Science

- Guides feature selection & redundancy checks
- Identifies candidate relationships for models
  - Remember inference vs exploration
- Helps spot spurious associations
- Sets up the move toward causal reasoning

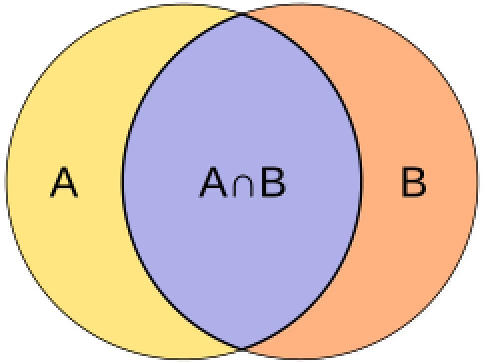
Beyond Classical Correlation

- Why More Measures?
  - Pearson, Spearman, Kendall → continuous / rank data
  - But… data often comes as categories or sets
  - Need similarity metrics for:
    - Recommender systems
    - Text analysis
    - Categorical survey data

Jaccard Similarity Index/Tanimoto

- **Range: 0 = no overlap → 1 = perfect overlap**
  - Examples:
    - Users who liked movie X and movie Y
    - Shared words between two documents

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



Example

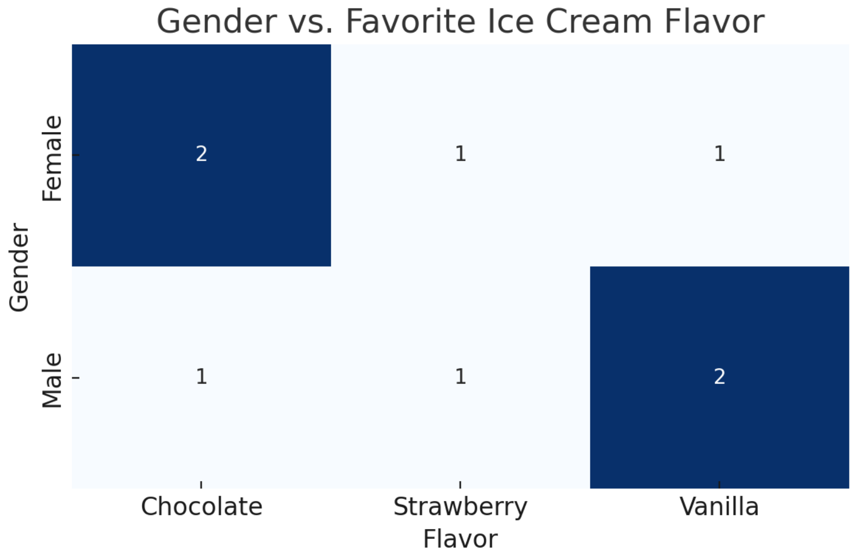
- User A liked {Star Wars, Titanic, Inception, Avatar}
- User B liked {Titanic, Inception, Matrix}
- Intersection = {Titanic, Inception} = 2
- Union = {Star Wars, Titanic, Inception, Avatar, Matrix} = 5
- Tanimoto = 2 / 5 = 0.4

Categorical Correlation: Cramér’s V

- **Definition:** Association between two categorical variables
- k = # of columns, r = # of rows
- Range: **0 = no association, 1 = perfect association**
- Based on Chi-square test of independence
- Example: Gender × Favorite Ice Cream Flavor

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

Cramer’s V Example



$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

- $\chi^2 \approx$ **0.67** and Cramér’s V ≈ **0.289**

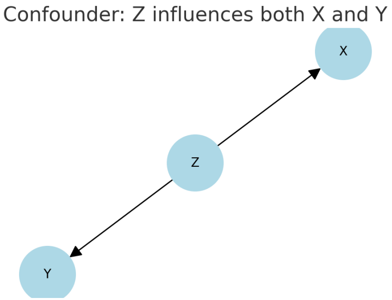
Measures of Association

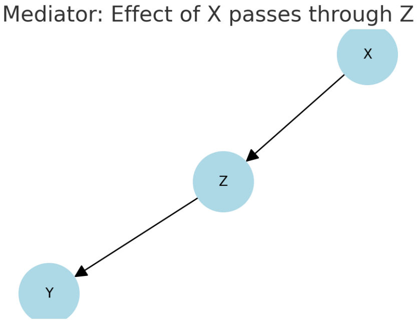
- Jaccard/Tanimoto: set similarity, sparse binary features, recommendations, text mining
- Cramér's V: categorical × categorical relationships, survey/experimental data
- Complements Pearson/Spearman by handling non-numeric data

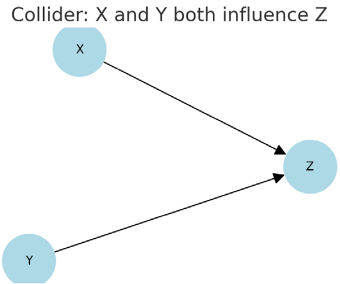
Causation in Data Science

Correlation to Causation

- Correlation: two variables move together.
- Causation: changing one variable changes the other (very hard to measure, this is more conceptual)
- In real data, relationships can be misleading because of:
  - Confounders (hidden common causes)
  - Mediators (indirect causal paths)
  - Colliders (common effects that distort correlations)



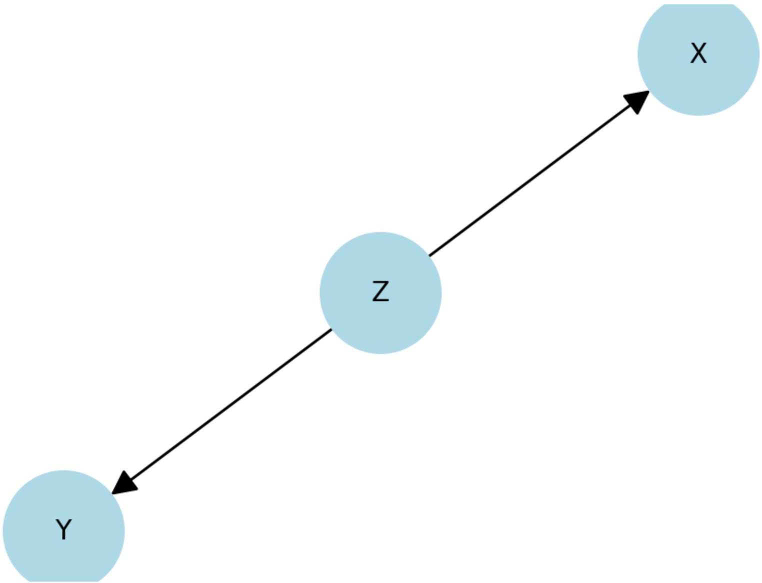




Confounders

- A confounder is a variable that influences both X and Y.
- Creates a spurious association between X and Y.
- Key: If we don’t account for confounders, we may wrongly conclude X causes Y.

Confounder: Z influences both X and Y

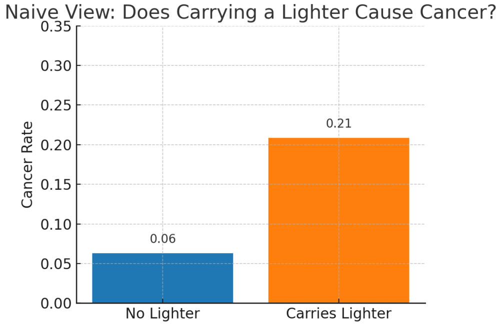


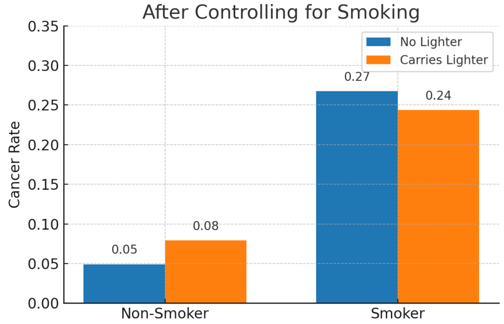
Confounder Generalizable Logic

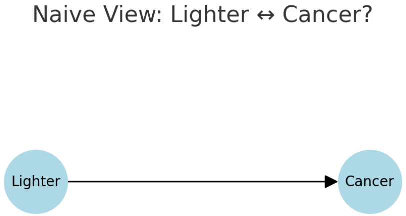
- Confounder must:
  - Be associated with the independent variable (X).
  - Be associated with the dependent variable (Y).
  - Not lie on the causal pathway from X → Y.

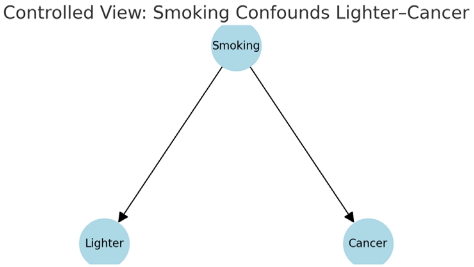
Confounder Example

- Observed: People who carry lighters are more likely to get lung cancer.
- Hidden confounder: Smoking.
- Smoking → people carry lighters.
- Smoking → higher lung cancer risk.
- So the lighter–cancer correlation is spurious.









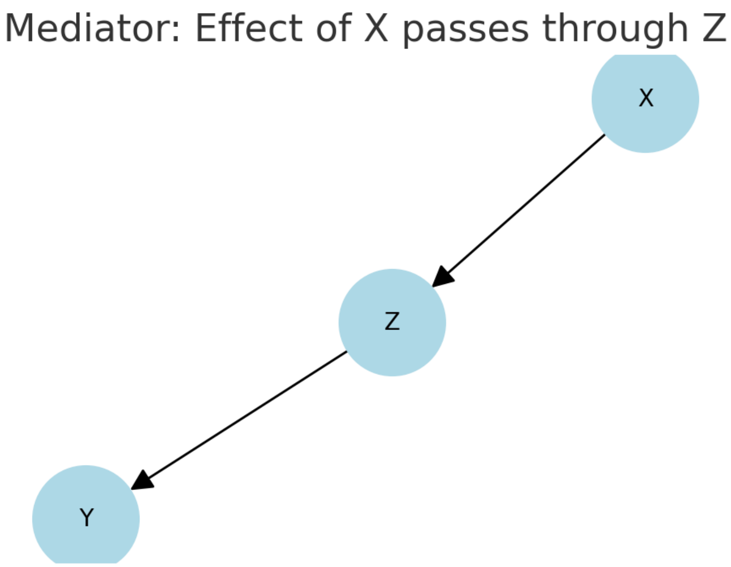
How are we going to do this?

Partial Correlation

- Naive correlation (Lighter ↔ Cancer): 0.22
- Partial correlation (Lighter ↔ Cancer | Smoking): 0.01
- Interpretation: once we control for Smoking, the lighter–cancer link disappears.
- Smoking is the real cause, lighters are just correlated because of the confounder.

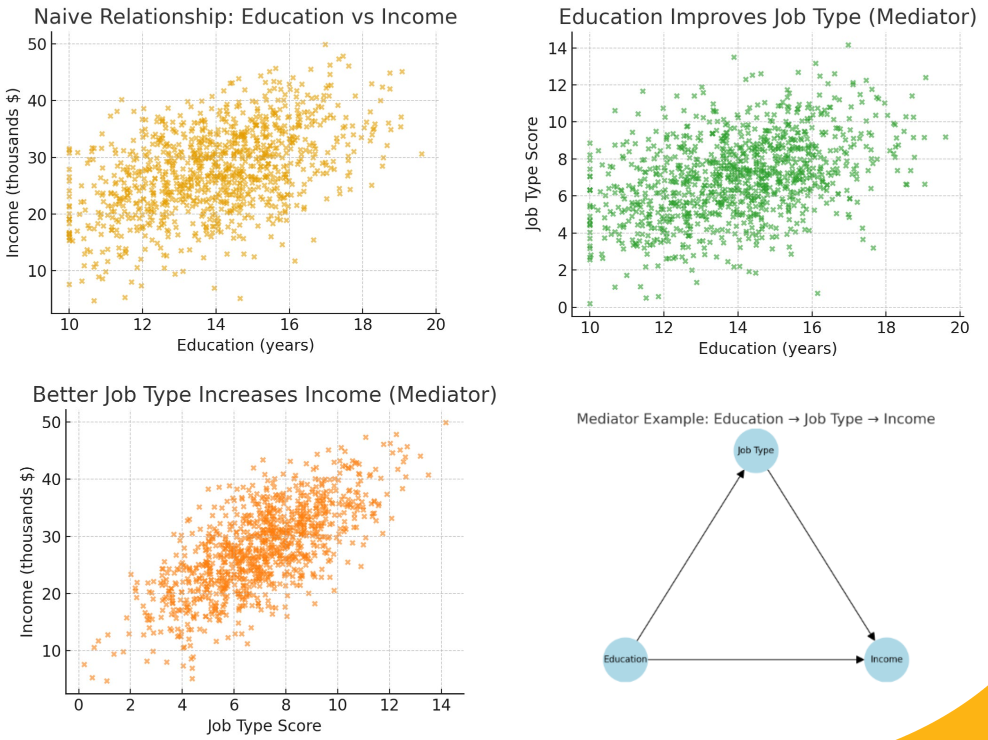
Mediator

- A mediator explains ***how causality flows*** and usually should **not** be adjusted away if you care about the total effect.
  - Be caused by the independent variable (X).
  - Be associated with the dependent variable (Y).
  - Lie on the causal pathway from X → Y.



Mediator Example

- Does higher education lead to higher income — and if so, is the effect direct, or does it work through the kinds of jobs people get?
- Education → better jobs → higher income.
- The effect of education on income is partly explained through job type



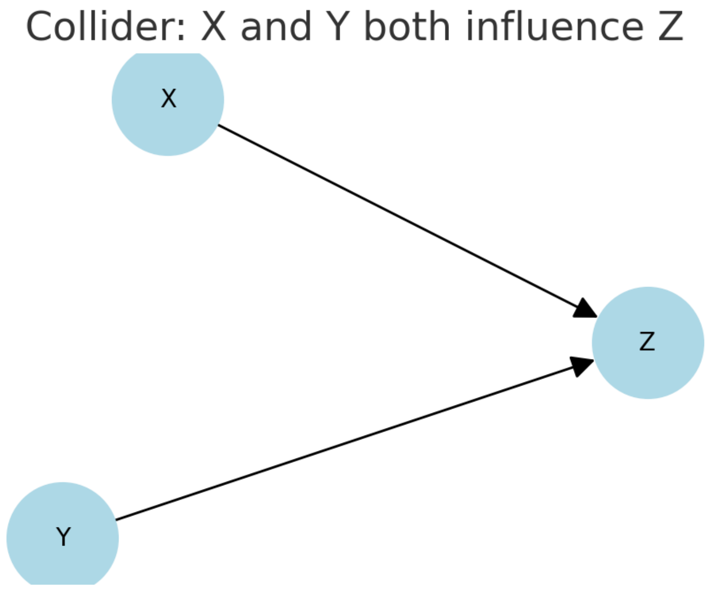
What will the partial correlation show?

Partial Correlation: Mediator Example

- Naive correlation (Education ↔ Income): 0.50
- Partial correlation (Education ↔ Income | Job Type): 0.33
- Interpretation: part of Education’s effect on Income flows through Job Type.
- The correlation weakens **<u>but doesn’t vanish</u>** → evidence of mediation.

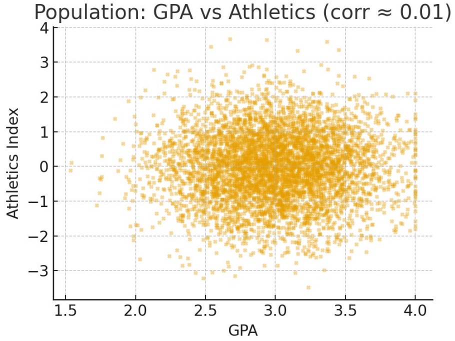
Collider

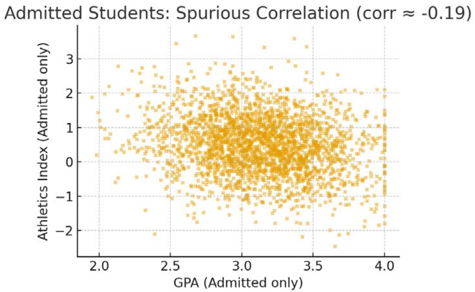
- A collider is a variable that:
  - Is influenced by both X and Y.
  - Conditioning on Z produces an artificial association between X and Y.

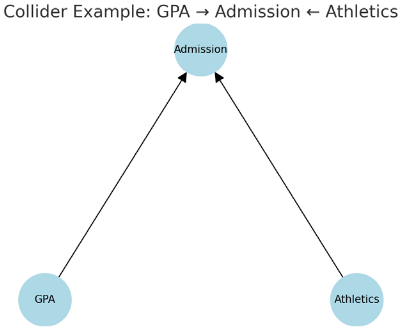


Collider Example

- Among admitted students, why do we see that higher GPA applicants seem less athletic — is there really a trade-off, or is this a statistical artifact of how admissions decisions are made?
- Why? Admissions depends on both GPA and athletics.
  - GPA → Admission
  - Athletics → Admission
- If we condition on “admitted students” (the collider), GPA and athletics appear negatively correlated, even if they aren’t in the full population.
  - Among admitted students:
    - If a student has a low GPA, they probably got in because of strong athletics.
    - If a student has weak athletics, they probably got in because of a high GPA.
    - This creates an artificial negative correlation between GPA and athletics in the admitted group.







What will the partial correlation show?

Collider Partial Correlation

- Scenario: GPA & Athletics, with Admission as a collider
- Naive correlation (GPA ↔ Athletics, whole population): ≈ 0.01
- Partial correlation (GPA ↔ Athletics | Admission): ≈ –0.19
- Interpretation:
  - GPA and Athletics are independent in the population.
  - When we control for Admission (the collider), a false negative correlation is created.
  - Conditioning on a collider can introduce bias instead of removing it.

Summary

- Confounder: must be controlled to avoid false inference.
- Mediator: don’t block it if you want the total effect.
- Collider: never control for it — it creates bias.

Summary

Partial Correlation Outcomes

Scenario	Naive Correlation	Partial Correlation	Interpretation
Confounder (Lighter–Cancer)	~0.22	~0.01	Effect vanishes → confounder explained the spurious link.
Mediator (Education–Income)	~0.50	~0.33	Effect weakens → mediator explains part of the effect.
Collider (GPA–Athletics)	~0.01	~–0.19	Effect appears → conditioning on collider creates a false link.

Moving Forward

- Partial Correlation
  - Measures the relationship between X and Y after removing the influence of Z.
  - Extends correlation to “control for” one or more variables.
  - Useful for conceptual understanding before regression.
- Regression (coming soon)
  - Estimates how much Y changes when X changes, while holding other variables constant.
  - Provides coefficients, significance tests, and predictions.
  - More flexible for multiple confounders and complex models.
- Takeaway:
  - Partial correlation gives a statistical snapshot of adjusted relationships.
  - Regression generalizes this idea and will be our main tool going forward.