Correlation and Causation

Professor Widom's Instructional Odyssey

www.professorwidom.org

Stanford University

Google

Instabase

Google Cloud Platform

Association for Computing Machinery

Very Large Data Bases Endowment Inc.

amazon web services

Correlation and Causation

Analyzing data with two (or more) measures X and Y

- Height and shoe size
- Grades and entrance exam scores
- Education level and starting salary
- Temperature and cold drink sales

➢ Correlation (informal): The values of X and Y tend to be interdependent

➢ Causation (informal): X's value tends to influence Y's value

Why Do We Care?

- Discoveries in the medical domain
  - Patients with problem X also tend to have problem Y
  - Taking drug X tends to make symptom Y subside
- Discoveries in the political domain
  - Voters who approve of X tend to also approve of Y
  - Voter turnout is weather-dependent
- Discoveries in the advertising domain
  - Larger fonts tend to result in more click-throughs
  - More purchases are made in the evening

Which are correlation and which are causation?

Categorical versus Numeric Values

- Categorical values: unordered categories
  - color, weather, major
- Numeric values: ordered values
  - height, price, time, age, exam score
  - May be discrete or continuous
- Ordinal values: categories that can be ordered
  - movie rating, letter grade, education level
  - But differences may not be on a meaningful scale

Assume ordered for now

Positive Correlation

Two measures X and Y

When X is higher Y tends to be higher

When X is lower Y tends to be lower

When Y is higher X tends to be higher

When Y is lower X tends to be lower

Examples

X = height, Y = shoe size

X = grades, Y = entrance exam scores

Notation (mine): $X \approx Y$

Positive Correlation by Causation

Two measures X and Y

X being higher causes Y to be higher

X being lower causes Y to be lower

Examples

X = education, Y = starting salary

X = temperature, Y = cold drink sales

Notation (mine): $X \rightarrow Y$

Correlation due to Hidden Causation

Correlation can be the result of causation from a hidden "confounding variable"

A ≈ B because there's a hidden C such that C → A and C → B

Homeless population ≈ crime rate
Confounding variable: unemployment

Forgetfulness ≈ poor eyesight
Confounding variable: age

Negative Correlation by Causation

Two measures X and Y

X being higher causes Y to be lower

X being lower causes Y to be higher

Examples

X = latitude, Y = temperature

X = car weight, Y = gas mileage

X = class absences, Y = final grade

Negative Correlation without Causation

Two measures X and Y

When X is higher Y tends to be lower

When X is lower Y tends to be higher

When Y is higher X tends to be lower

When Y is lower X tends to be higher

Examples

X = cold drink sales, Y = hot tea sales

X = years of schooling, Y = years in jail

Confounding variables?

Is There Such a Thing as Pure Correlation?

Correlation without causation: usually a confounding variable lurking somewhere

X = height, Y = shoe size

X = grades, Y = entrance exam scores

What about the “spurious correlations”?

Bottom Line

1) Want to know when things are correlated
2) But should not assume one causes the other

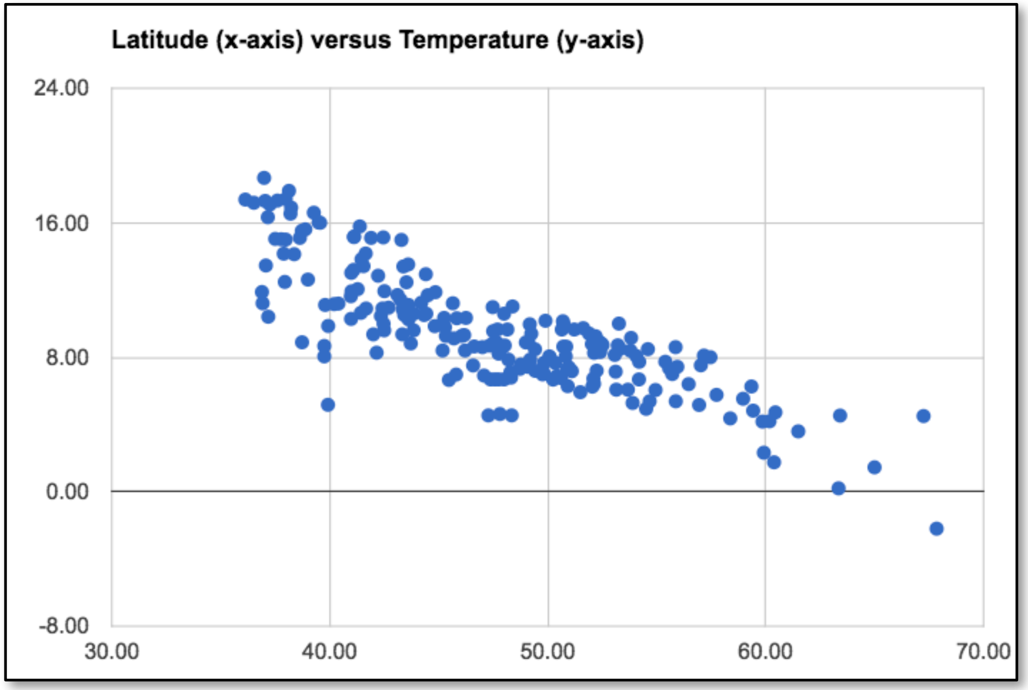
   "Correlation does not imply causation"

Next:

- Determining correlation
- Determining if there's causation

Determining Correlation

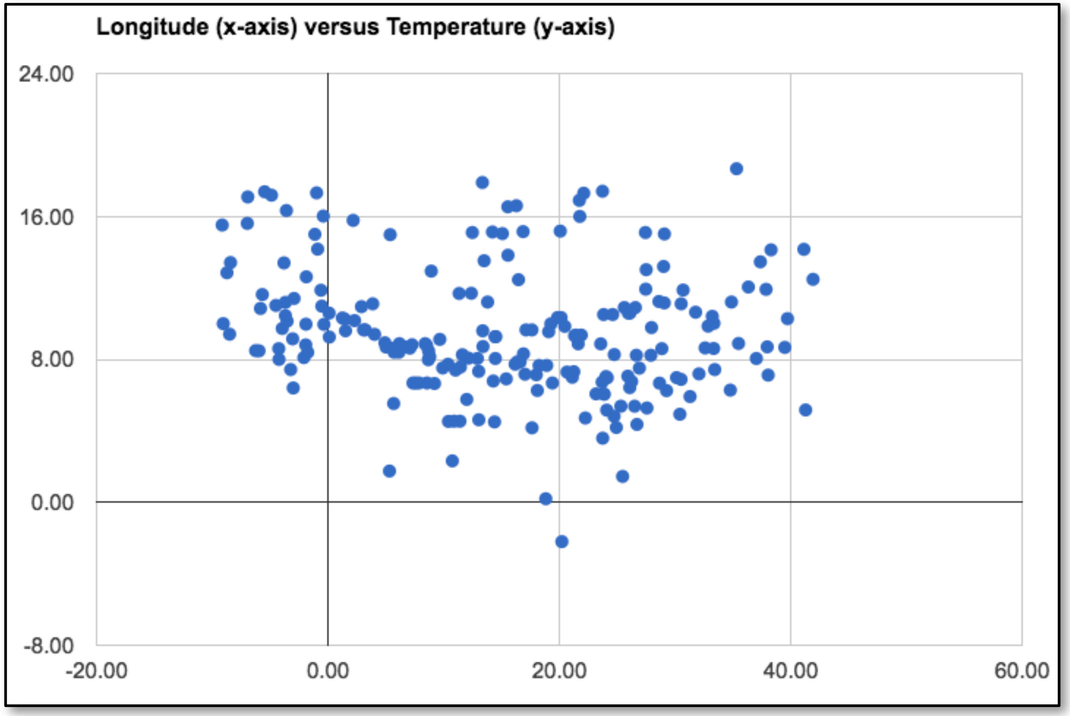
X and Y both ordered: scatterplot, r-values



Significant negative correlation, r = -0.82

Determining Correlation

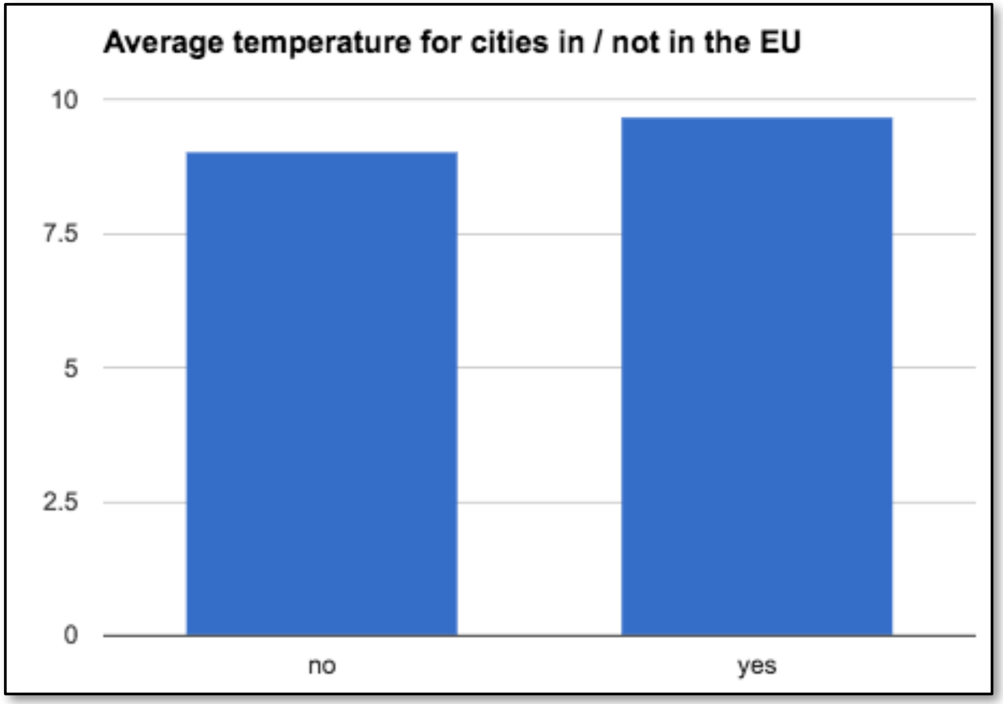
X and Y both ordered: scatterplot, r-values



Little correlation, r = -0.17

Determining Correlation

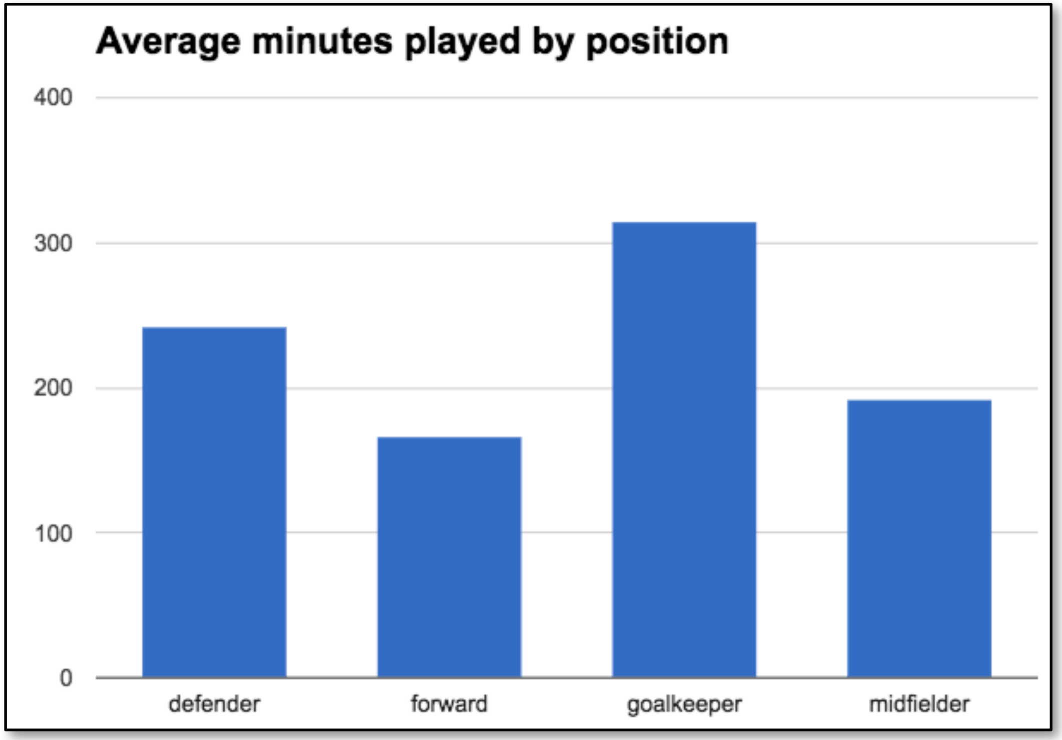
X categorical, Y ordered: bar graph



Similar averages → little correlation

Determining Correlation

X categorical, Y ordered: bar graph



Different averages → correlation

Determining Correlation

X and Y both categorical: table

	EU	Not EU
Cold	12	5
Cool	63	29
Warm	44	18
Hot	31	11

(Ignore that temperature category is ordinal)

Without correlation expect "evenly divided values"

Determining Correlation

X and Y both categorical: table

	EU	% of row	Not EU	% of row
Cold	12	71%	5	29%
Cool	63	68%	29	32%
Warm	31	74%	11	26%
Hot	44	71%	18	29%

Without correlation expect "evenly divided values"

Determining Correlation

X and Y both categorical: table

	EU	% of row	Not EU	% of row
Cold	12	71%	5	29%
Cool	63	68%	29	32%
Warm	31	74%	11	26%
Hot	44	71%	18	29%

If different rows have the same relative percentages, values are uncorrelated

If different rows have **different** relative percentages, values are **correlated**

Determining Correlation

X and Y both categorical: table

	EU	% of column	Not EU	% of column
Cold	12	8%	5	8%
Cool	63	42%	29	46%
Warm	31	21%	11	17%
Hot	44	29%	18	29%

If different *columns* have the same relative percentages, values are uncorrelated

If different *columns* have different relative percentages, values are correlated

Your Turn

Are Position and Passes-per-Minute correlated?
(for purpose of exercise treat ppm as categories)

	Low ppm	Med ppm	High ppm
Defender	94	46	48
Midfielder	79	44	105
Forward	116	15	12

Your Turn

Are Position and Passes-per-Minute correlated?

	Low ppm	% of row	% of col	Med ppm	% of row	% of col	High ppm	% of row	% of col
Defender	94			46			48		
Midfielder	79			44			105		
Forward	116			15			12		

If different rows/columns have the same relative percentages, values are uncorrelated

If different rows/columns have different relative percentages, values are correlated

Determining Causation

Not possible based on data analysis alone

1) “Hill’s Criteria”

2) Run experiments

Hill's Criteria (slightly adapted)

Strength - of correlation between X and Y

Consistency - of correlation across different datasets

Specificity - no other likely explanation for correlation

Temporality - Y occurs after X

Plausibility - there's a reason for causation

Coherence - consistent with related theories

Experimental Validation

To determine if X → Y

- Create "experimental group" with X=value
- Create "control group" with different X values
- *Ensure no other distinctions between two groups*
- See if experimental vs. control ≈ Y

Works for things like drug therapy, advertising font size

Less useful for things like weather, crime rate, forgetfulness

Historical Example

Smoking and Lung Cancer, 1950's

- Strong correlation observed between smoking and lung cancer
- Tobacco proponents implicated confounding variables (e.g., pollution, occupation, genetic predisposition)
- Experimental validation not feasible
- Large-scale data analysis (Big Data!) concluded causation, both using Hill's criteria and eliminating proposed confounding variables

Bottom Line

1) Can measure when things are correlated
2) But should not assume one causes the other

“Correlation does not imply causation”

Correlation and Causation

Professor Widom's Instructional Odyssey

www.professorwidom.org

Stanford University

Google

Instabase

Google Cloud Platform

Association for Computing Machinery

Very Large Data Bases Endowment Inc.

amazon web services™