DS 3003: Research Fallacies

Spring 2022

Youmi Suk

School of Data Science, University of Virginia

1. 18. 2022

Overview

- Data driven vs. Data informed
- Simpson's Paradox
- Ecological Fallacy
- Atomistic Fallacy
- Correlation is Not Causation
- Other Fallacies

What does data tell us?

- We live in a society obsessed with data
- Data helps us make decision.
    - data-driven decision
    - data-informed decision

The difference between driven and informed

- There is a fundamental difference between being data-driven and being data-informed.
- **Data-driven**: You let the data guide your decision-making process
- **Data-informed**: You let data act as a check on your intuition
- When data guides your decision-making, it's recommended that you collect more data to have more accurate models and find trends.
- Being data-informed means using both intuition and data to develop your hypotheses/ideas/thoughts.

When does data guide our decision making?

AlphaGo versus Lee Sedol

A five-game Go match between 18-time world champion Lee Sedol and AlphaGo, a computer Go program developed by Google DeepMind, played in Seoul, South Korea between 9 and 15 March 2016

When does data guide our decision making?

AlphaGo versus Lee Sedol

- AlphaGo initially learned from thousands of games played by professional human players.
- It decides on the next move based on a probability to win associated with each of the possible moves (using Monte Carlo simulations).
- AlphaGo improves by playing against itself (using reinforcement learning).
- As computing power grew exponentially, computers have become better than human at computation by many orders of magnitude.
- At the end of the game, *AlphaGo won by 4 against 1*.
- But in many situations, we want to leverage data to check on our intuition and make data-informed decisions.

Being data-informed is all well and good?

- Of course, it's good because data is never going to tell you the full story.
- But using it at face value to drive decision making can be rather dangerous.
- This isn't a case of "garbage in = garbage out."

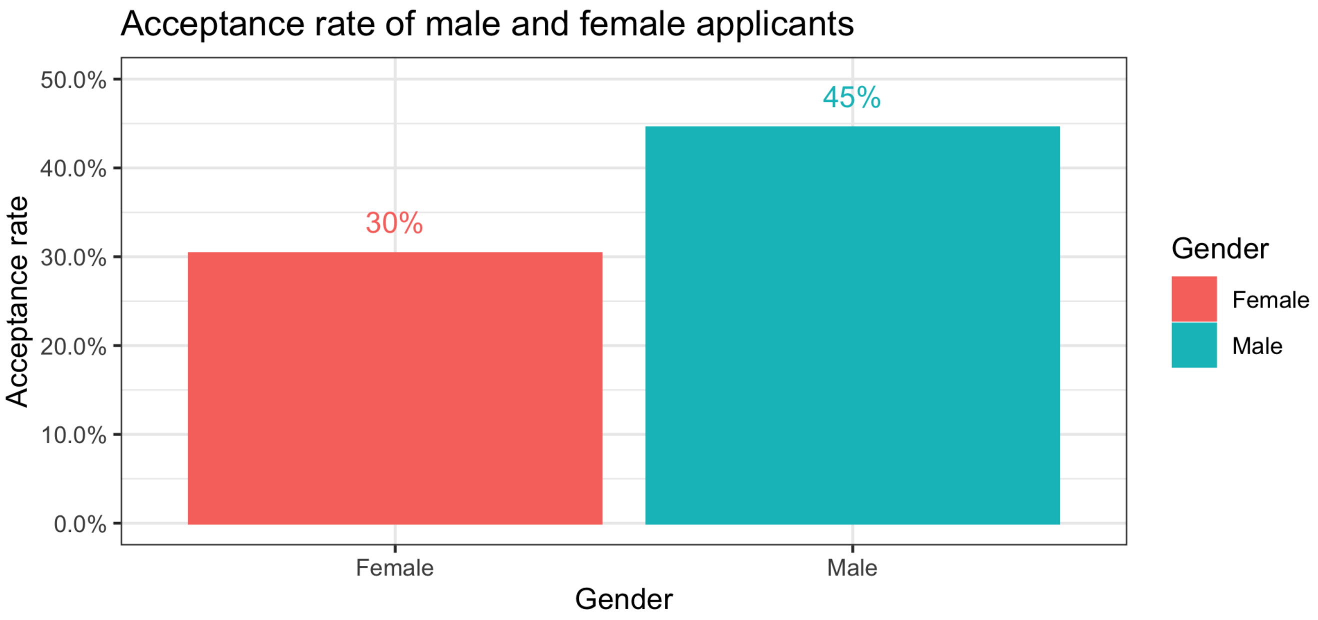
How UC Berkeley almost got sued for sex discrimination

- In 1973, UC Berkeley was sued for sex-discrimination. It turned out of all the female students who applied — only 30% of them were admitted. While out all the male students who applied 45% of them were admitted.

Gender	Admitted Rate (%)
Female	30
Male	45

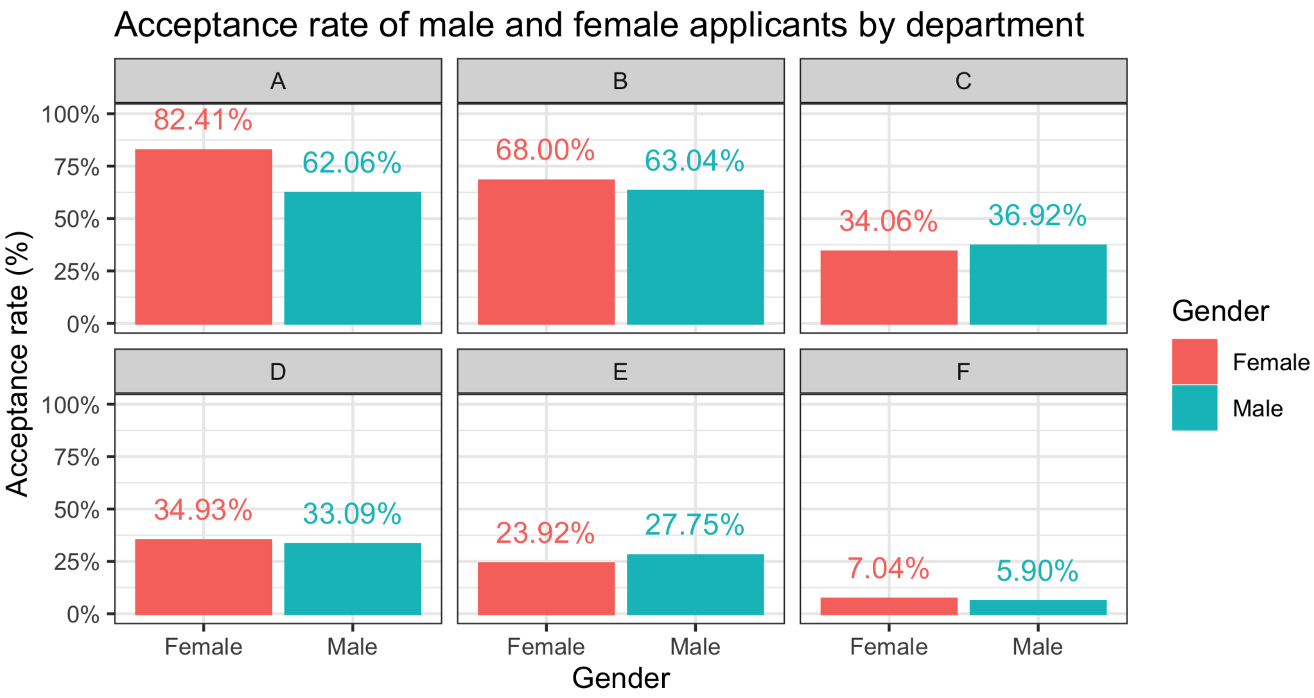
How UC Berkeley almost got sued for sex discrimination

- In 1973, UC Berkeley was sued for sex-discrimination. It turned out of all the female students who applied — only 30% of them were admitted. While out all the male students who applied 45% of them were admitted.



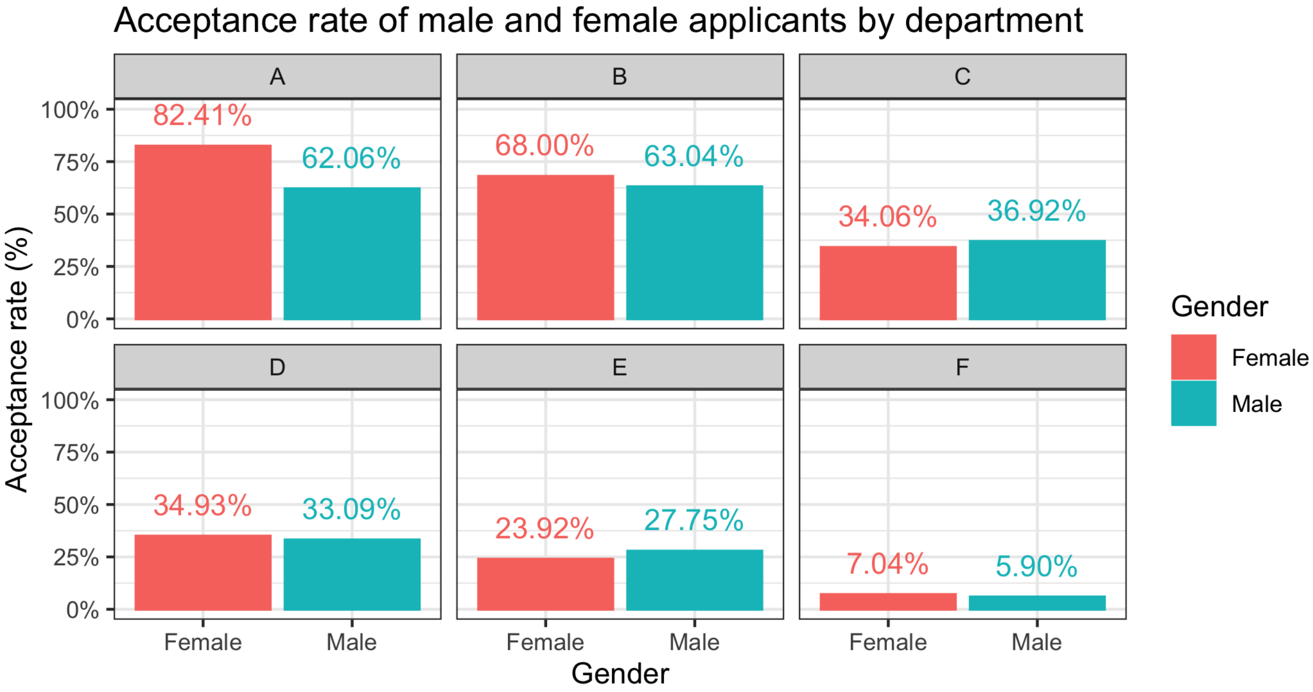
How UC Berkeley almost got sued for sex discrimination

- UC Berkeley set out to find the main culprits of this gender discrimination.
- They broke open the data to see which departments were mainly responsible for this gender bias.



How UC Berkeley almost got sued for sex discrimination

- UC Berkeley set out to find the main culprits of this gender discrimination.
- Out of the 6 departments, 4 of the departments accepted women more than men.



How UC Berkeley almost got sued for sex discrimination

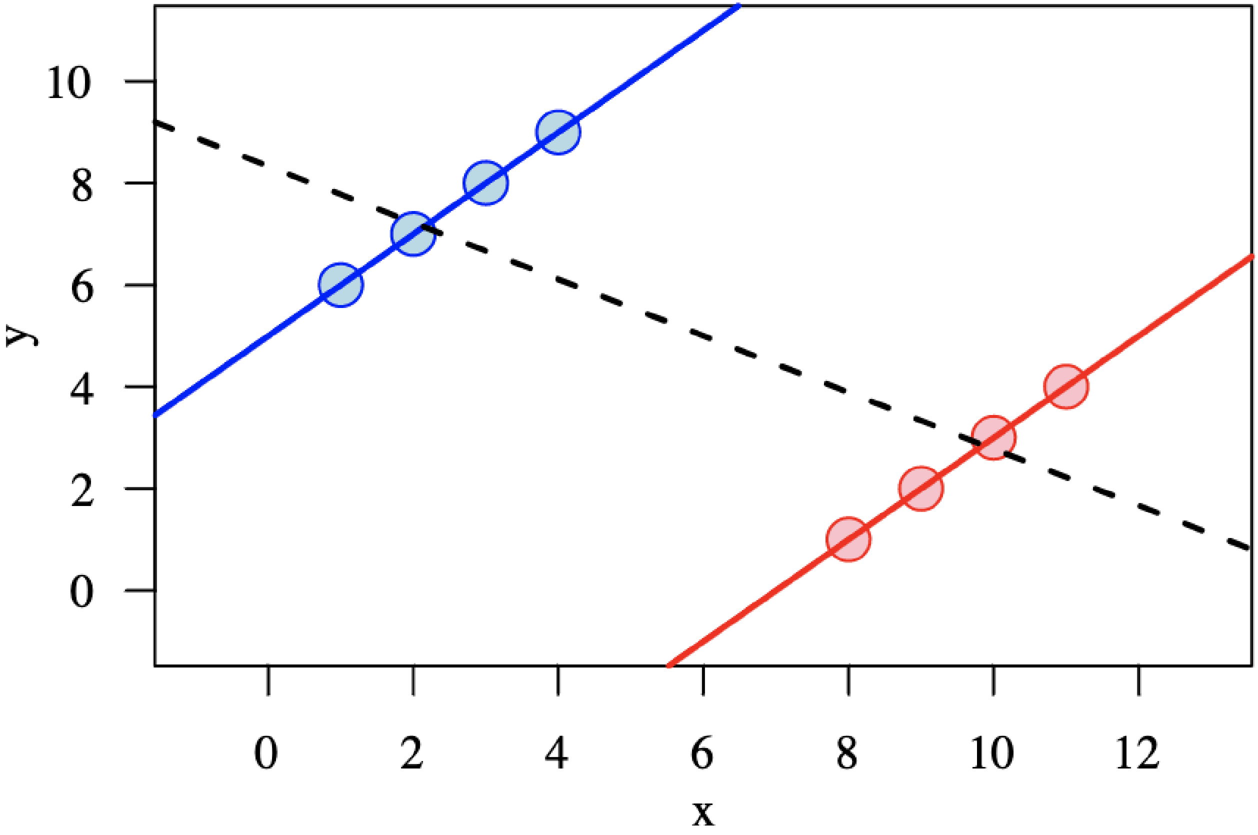
- UC Berkeley set out to find the main culprits of this gender discrimination.
- Out of the 6 departments, 4 of the departments accepted women more than men.

Department	# of Men	# of Women	Men Accepted	Women Accepted
A	825	108	62%	82%
B	560	25	63%	68%
C	325	593	37%	34%
D	417	375	33%	35%
E	191	393	28%	24%
F	373	341	6%	7%
Total	8442	4321	44%	35%

Simpson's Paradox

- A trend or result that is present when data is put into groups that reverses or disappears when the data is ungrouped/combined.
    - Grouped data tells the opposite story of the ungrouped data.
- More formally, Simpson's Paradox occurs when the marginal association between two variables is qualitatively different from the partial association between the same two variables after controlling for one or more other variables.
- This happens because of a confounding factor that is hidden from sight within the data.

Simpson's Paradox

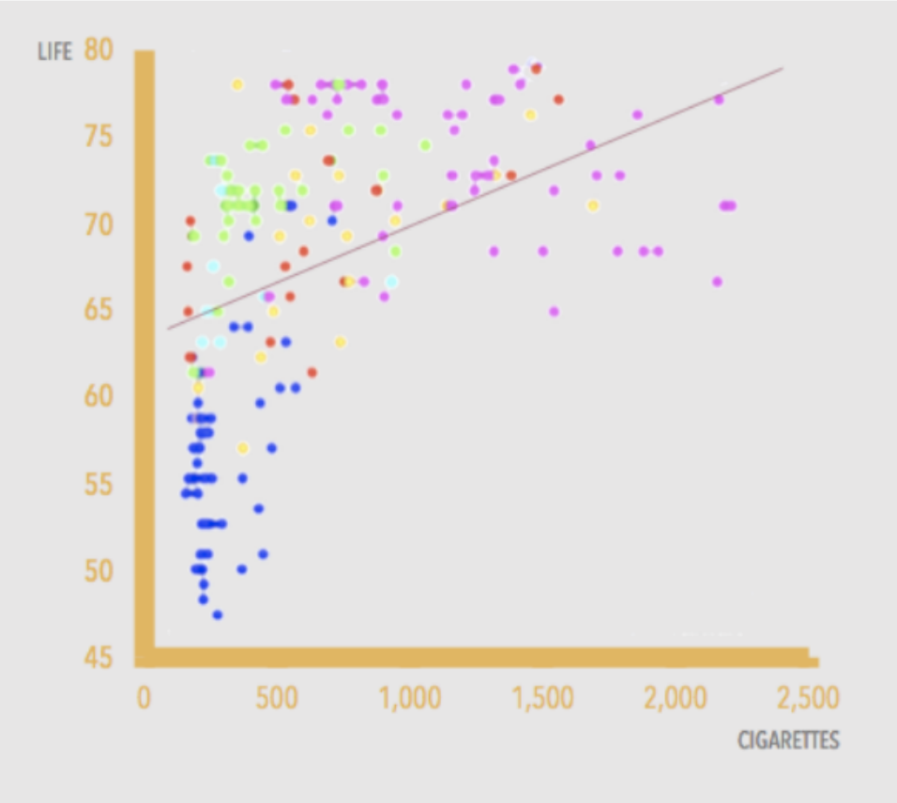


Ecological Fallacy

- Drawing conclusions about individuals based on aggregate data on a larger group.
    - Group → Individual
- A striking ecological fallacy is Simpson's paradox.
- Hierarchical levels - e.g., students, schools, districts, countries, etc.

Ecological Fallacy: Example

- Question: Is smoking cigarettes good for your health?
- See a relationship between average life expectancy and average cigarette consumption of different countries.



Ecological Fallacy: Example

- Question: Is smoking cigarettes good for your health?
- The problem is not with the data, the analysis, or the visualization. **The problem is with the title.**
- If we look at individual-level data, we can see the following trend.

Atomistic fallacy

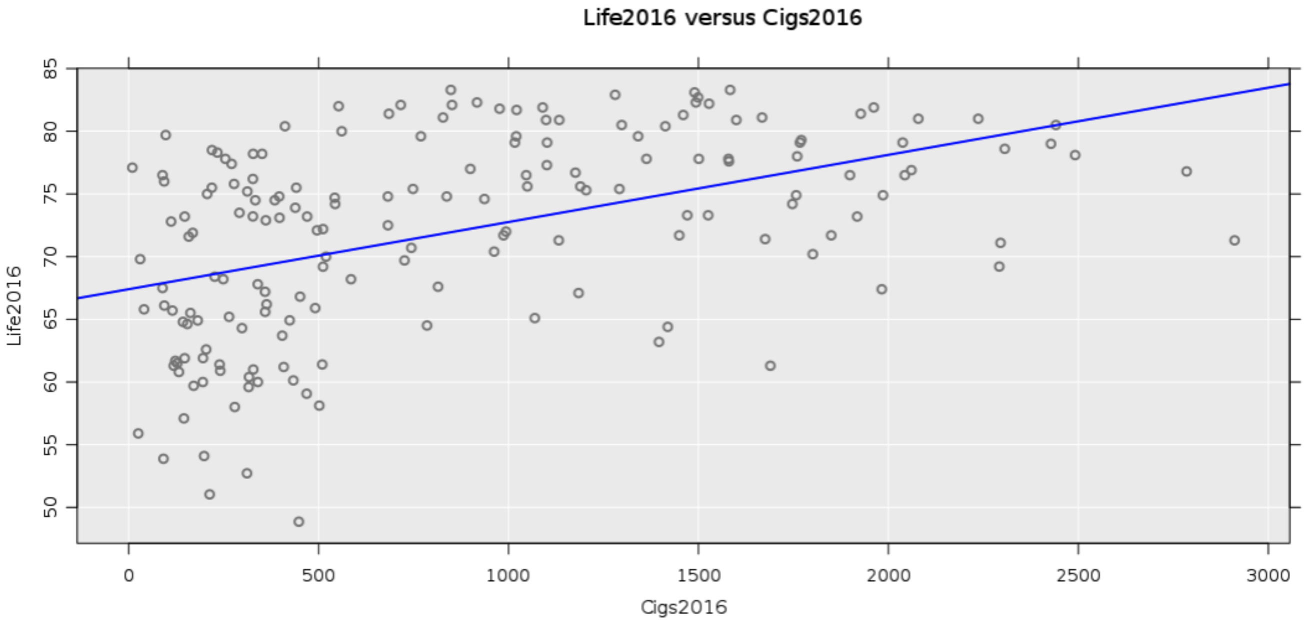
- Drawing conclusions about groups from analyses performed at a lower level.
    - Individual → Group
- The reverse of the ecological fallacy

Do you want to tell a compelling story based on data?

Make sure that you use data correctly and avoid traps like ecological or atomistic fallacy.

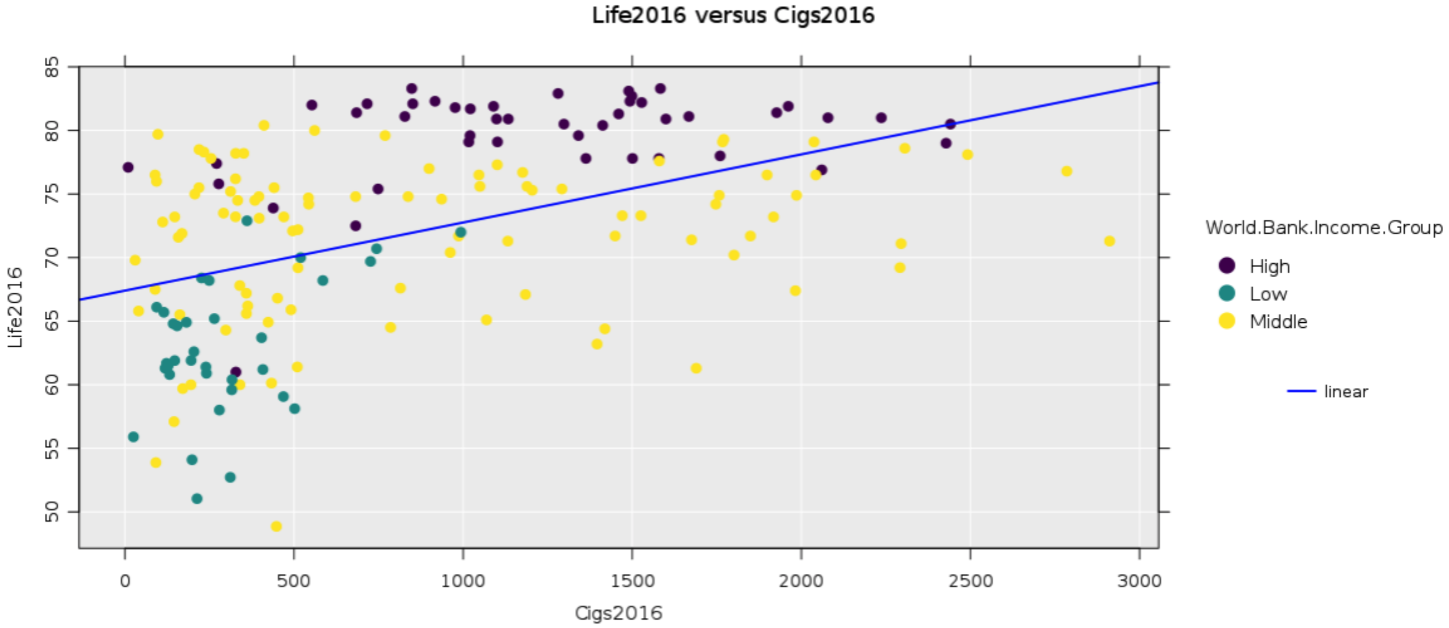
Correlation is Not Causation

- Although correlation can indicate the existence of a causal relationship, it is not a sufficient condition to definitively establish such a relationship.
- Let's revisit the previous example of average life expectancy and average cigarette consumption of different countries.



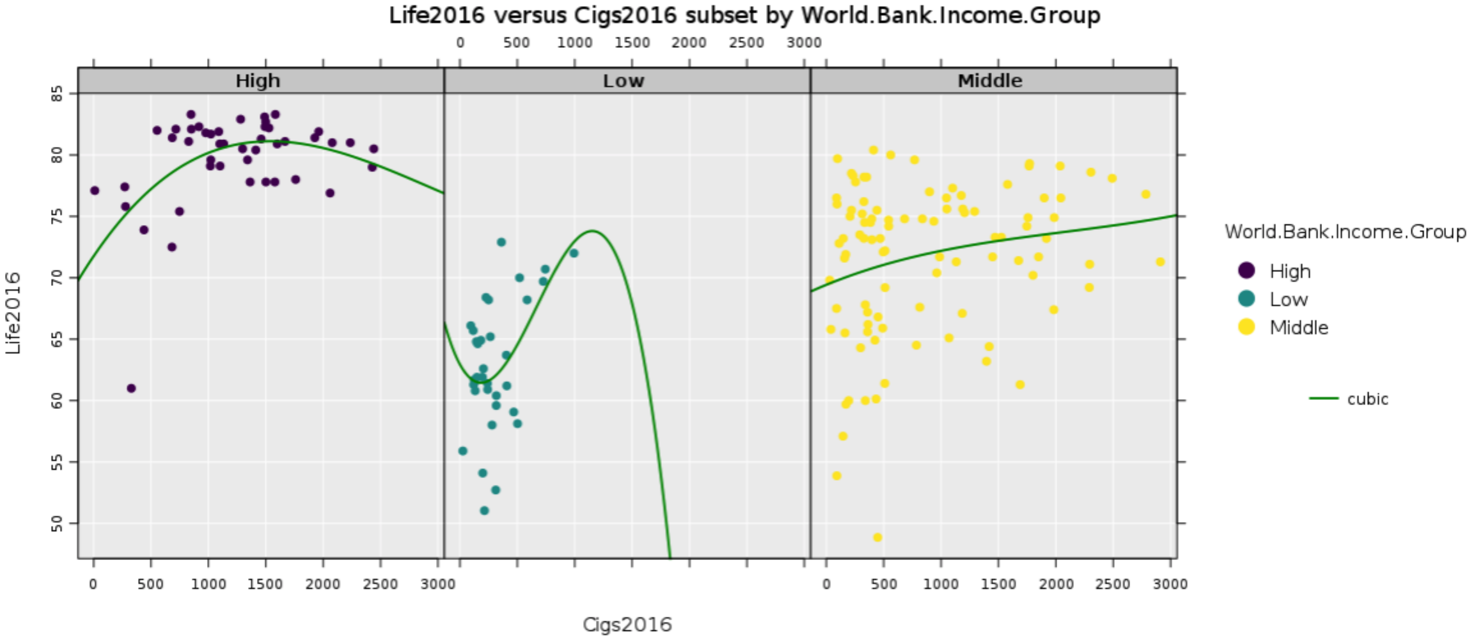
Correlation is Not Causation: Example

- We need to remove confounding variables in observational data.



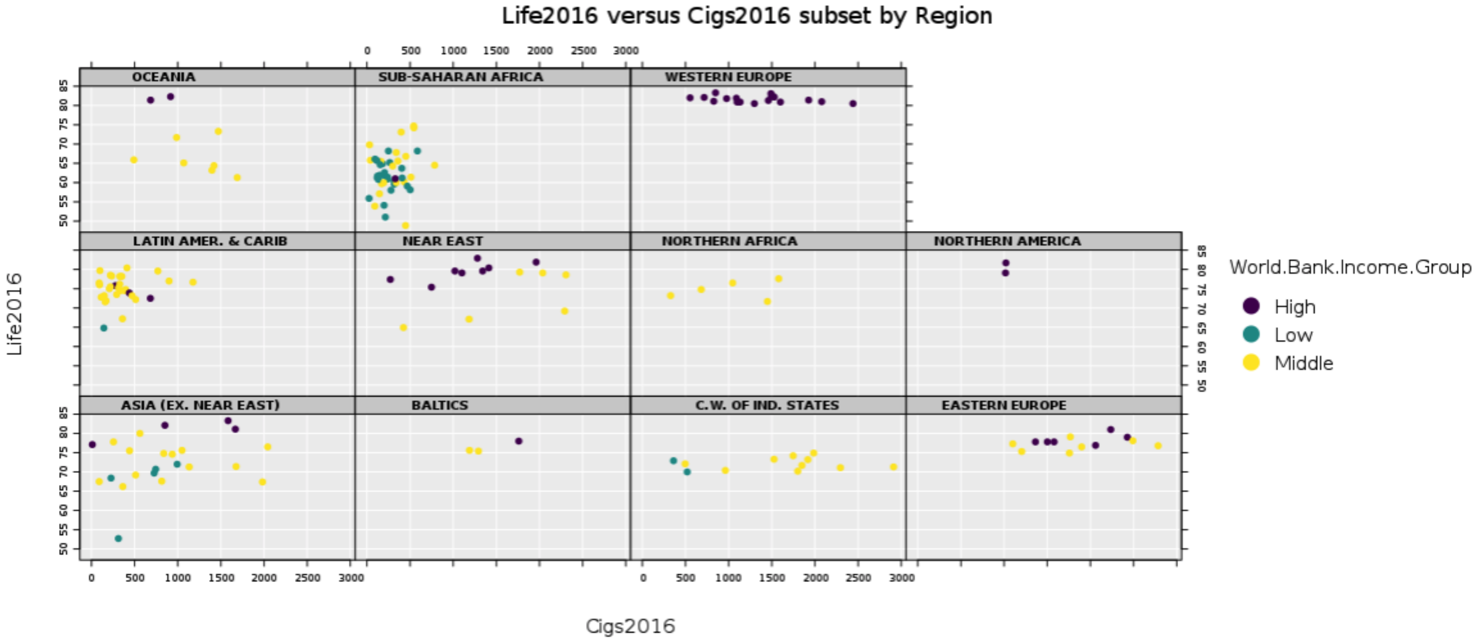
Correlation is Not Causation: Example

- Remove the confounding bias by the income group.



Correlation is Not Causation: Example

- Remove the confounding bias by regions.



Causal interpretation vs. Descriptive interpretation

- **Causal interpretation**:

If we were to *intervene* in the real world and first set $X = x$ and then $X = x + 1$ (e.g., $X = 0$ for the control group and $X = 1$ for the treatment group), the difference in the average outcomes would be $E[Y|do(x+1)] – E[Y|do(x)]$. *Intervening* means fixing the values of X in the real world. We change the real-world data-generating process, and the values of other variables (Y) change as a result of the intervention.

- **Descriptive interpretation**:

If we condition on $X = x$ and $X = x + 1$, the difference in conditional means is $E[Y|x+1] – E[Y|x]$. In conditioning on a variable we do not change the real world, we only narrow our focus to a subset of cases (i.e., our perception of the world changes but not the world itself).

Other Fallacies

- False generalizations
  - A sample is not representative of the target population (**biased-sample fallacy**).
  - An inadequate sample is used to justify the conclusion drawn (**insufficient sample fallacy**).
- Outliers/exceptions (**exception fallacy**)
- Inadequate evidence
- Misleading surveys or statistics