8.01 Correlated Data

How do you pick a model?

- When starting new projects, how do you decide which models you want to work with?

Assumptions

- Underlying each modeling tactic we've used, there have been some assumptions.
- Parametric modeling tactics make assumptions about the distributions of our data.
- Nonparametric modeling, while not making assumptions about how our data are distributed, still often assume that our observations are independent of each other.
- The most common assumption we'll make in modeling is that our observations are independent of one another.

Independent Observations

- In many cases, this is perfectly reasonable. If I take a random sample of 300 voters, it's rational for me to assume our data are independent.
- Even in cases where this is slightly violated, we'll believe it to be reasonable. If my random sample of 300 voters included two members of the same household, we'd almost certainly proceed with the assumption that our data are independent.
- Unfortunately, it isn't always reasonable for us to assume that our observations are independent of one another.

Correlated Data Week

- This week, we're going to talk about "correlated" data, which refers to data that is correlated with itself

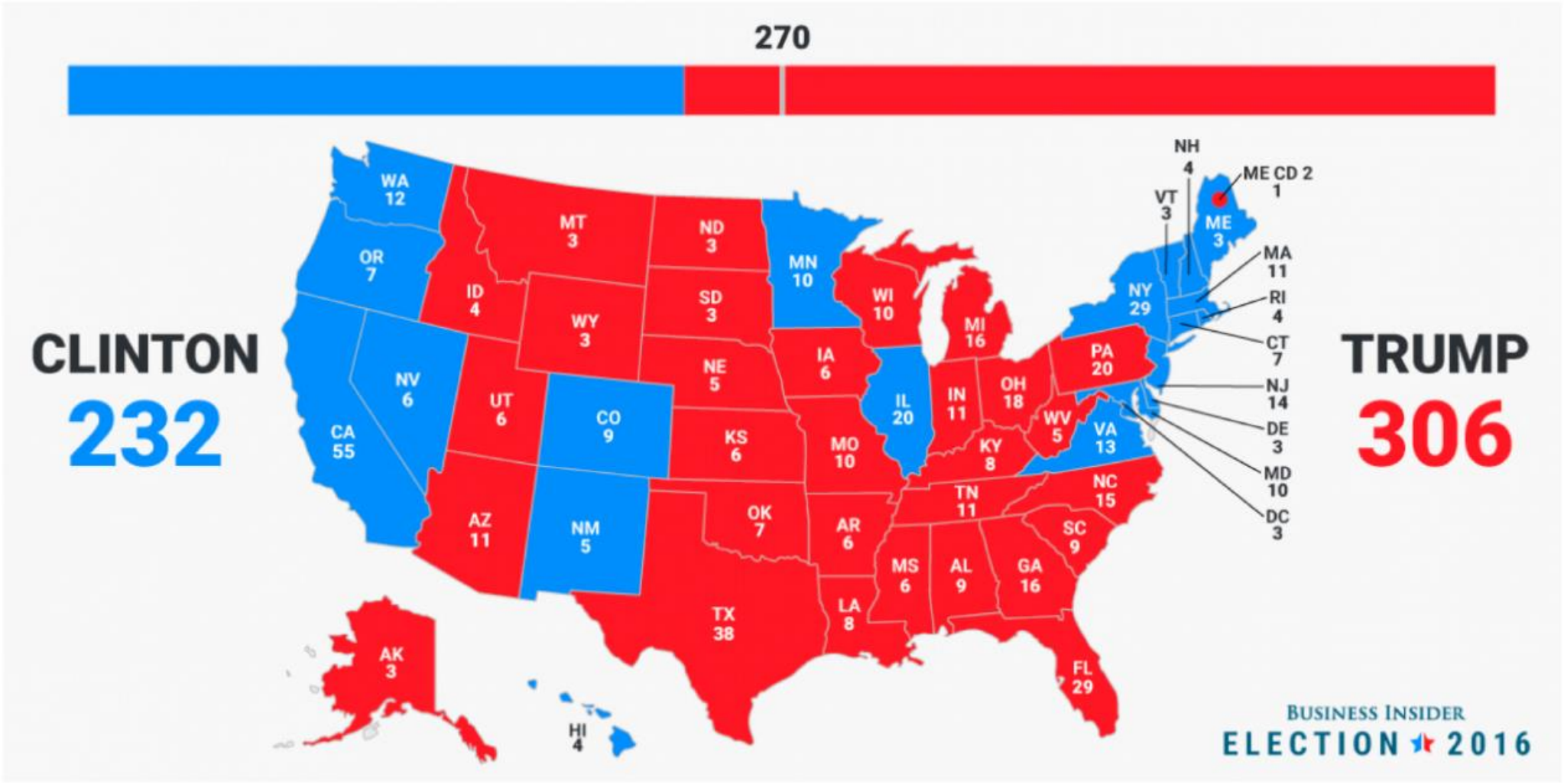
- Examples of correlated data include **time series data**

Correlated Examples

- In thinking about the following examples, let’s consider how the data we observe are not independent of one another.

Correlated Examples – Time Series

Correlated Examples – Spatial



Models with High Error due to Bias

- If your model leaves out important predictors, the bias that occurs is known as omitted variable bias.

- Also known as the bias that exists because we omitted or left out an important predictor

Resolving Omitted Variable Bias

- In a time series model, we could include all old Y-values that we think are relevant in predicting the new Y-values

- In a spatial model, that means we could include all the Y-values that we think are relevant in predicting other values of Y.