GENERAL ASSEMBLY

INTRODUCTION TO CORRELATED DATA

Matt Brems, Data Science Immersive

LEARNING OBJECTIVES

- Describe how our previous modeling tactics relate to one another.
- Understand the problem of correlated data.
- Describe methods for identifying the existence of observations correlated across time or space.

HOW DO I PICK A MODELING TACTIC?

FRAMING

- While this hierarchy helps us to understand and organize the different problem-solving methods available, there are other questions we have in mind that help guide our work.

ASSUMPTIONS

- Underlying each modeling tactic we've used, there have been some assumptions.

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.

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.

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.

SPATIOTEMPORAL WEEK

- This week, we're going to talk about "spatiotemporal" data, which just means data that has a space component and a time component.

- In this lecture, we'll introduce both.

- For the first half of the week, we'll spend most of our time on time series data.

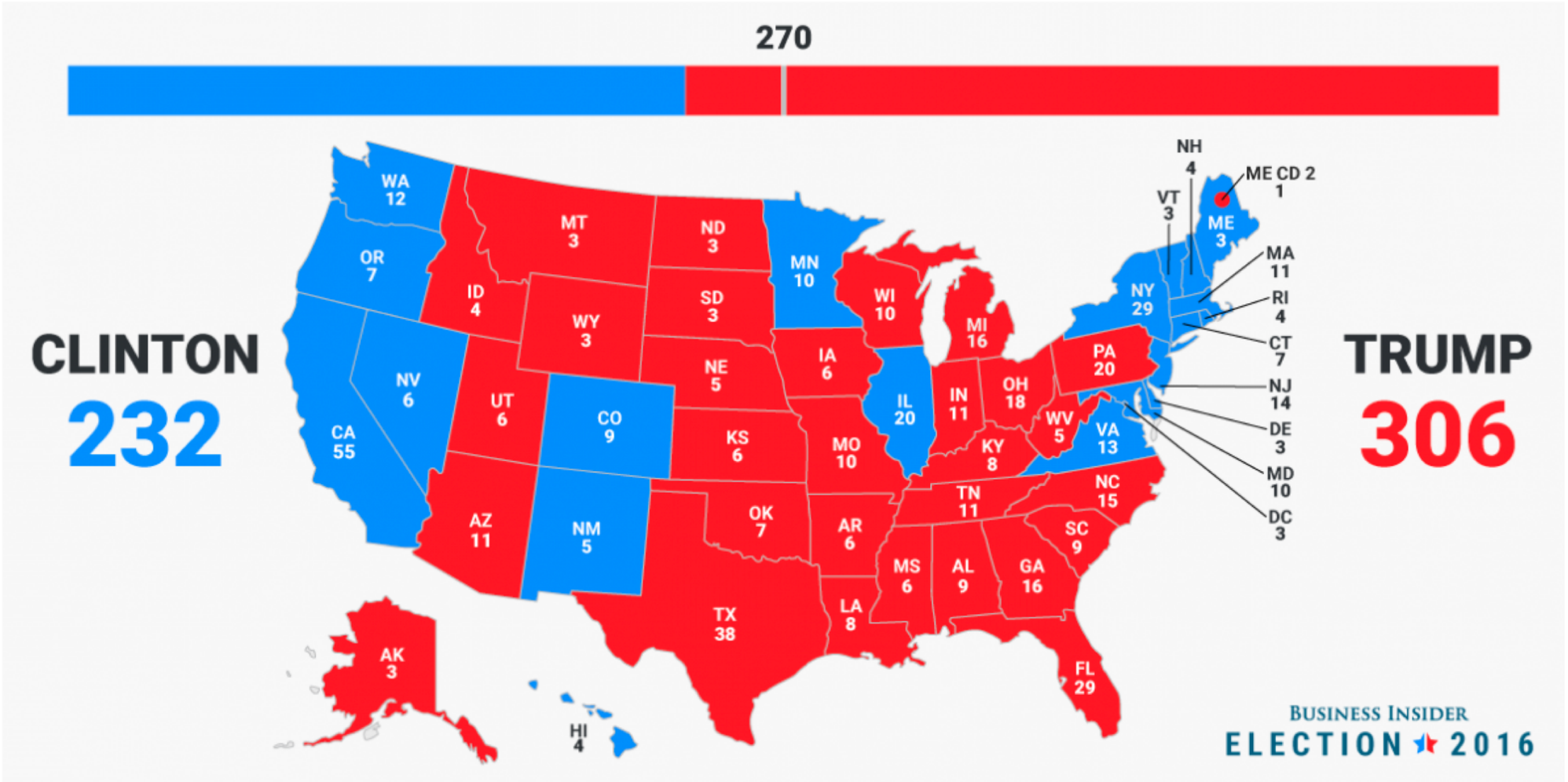
- Near the end of the week, we'll discuss spatial data and how to integrate it with temporal data.

SPATIOTEMPORAL WEEK

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

SPATIOTEMPORAL EXAMPLES

SPATIOTEMPORAL EXAMPLES



SPATIOTEMPORAL EXAMPLES

- It would be possible for us to consider these data as independent of one another. My DataFrame of stock price data might include stock price as the Y with time and other variables as our X.

- **Pandas won't throw an error if we try to fit this model.**

GROUP ACTIVITY #1

- Before building a model with the time series data here…

- 1. How might I detect temporal dependence of my observations?

- 2. When modeling, how could I try to account for this dependence?

GROUP ACTIVITY #1: DISCUSSION

- 1. How might I detect temporal dependence of my observations?

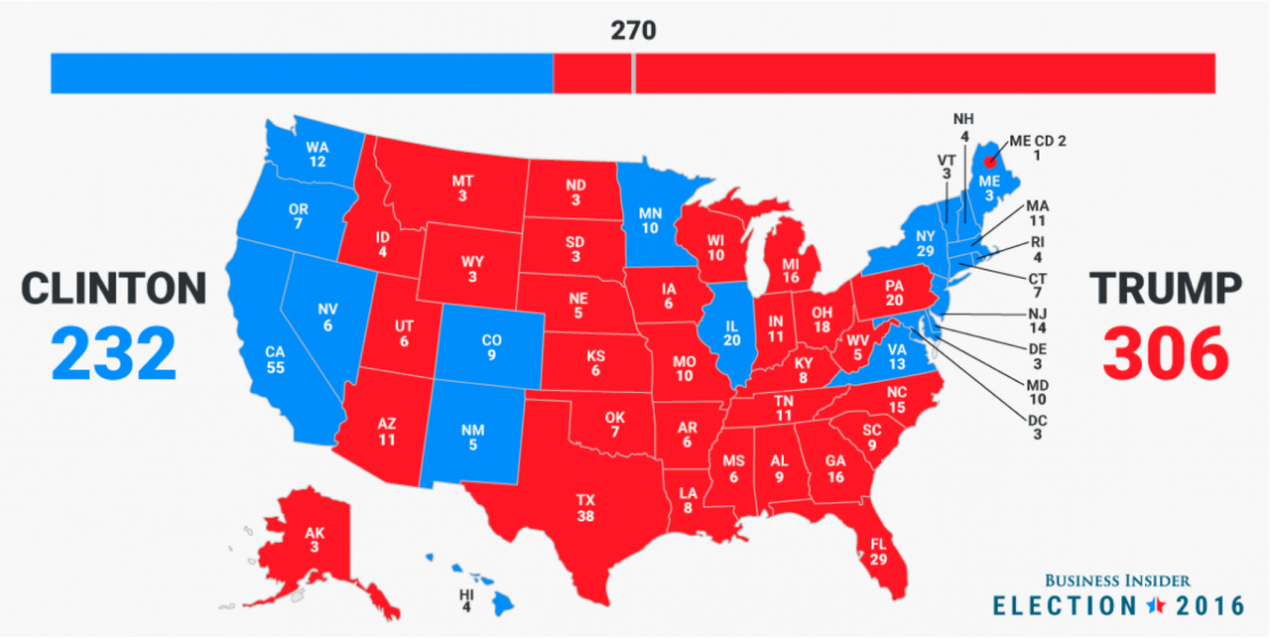
- 2. When modeling, how could I try to account for this dependence?

GROUP ACTIVITY #2

- Before building a model with spatial data here…

- 1. How might I detect spatial dependence of my observations?

- 2. When modeling, how could I try to account for this dependence?



GROUP ACTIVITY #2: DISCUSSION

- 1. How might I detect spatial dependence of my observations?

- 2. When modeling, how could I try to account for this dependence?

TWO STRATEGIES

- Strategy 1: Decorrelate observations.

- Strategy 2: Decompose data into components we care about.

DECORRELATE OBSERVATIONS

- Find out what is causing the correlation, then control for that by including it in our model.

DECOMPOSITION

- We will often use models to decompose spatiotemporal data into different components.

- Decomposing, in a modeling context, means writing out a model that isolates each component.

- $data = f(season) + g(trend) + h(noise)$