Backwards Design in Data Science Causal Inference Edition

Nick Eubank

Backwards Design

Approach to planning data science projects

Backwards Design

Approach to planning data science projects

- (Though backwards design isn't unique to DS)

Goals:

- Minimize wasted effort

Backwards Design

Approach to planning data science projects

- (Though backwards design isn't unique to DS)

Goals:

- Minimize wasted effort
- Make sure you develop explicit goals
    - Not get lost in your tools and data

Backwards Design

Start with where you want to end up, then work backwards

Backwards Design

Backwards Design

1. Determine Problem / Topic Area

Backwards Design

1. Determine Problem / Topic Area
2. What *causal question* are you seeking to answer?

Backwards Design

1. Determine Problem / Topic Area
2. What *causal question* are you seeking to answer?
3. What would your ideal experiment be?

Backwards Design

1. Determine Problem / Topic Area
2. What *causal question* are you seeking to answer?
3. What would your ideal experiment be?
4. Where can you (a) measure your outcome variable, and (b) find variation in your treatment variable?

Backwards Design

1. Determine Problem / Topic Area
2. What *causal question* are you seeking to answer?
3. What would your ideal experiment be?
4. Where can you (a) measure your outcome variable, and (b) find variation in your treatment variable?
5. What would a feasible design be?

Backwards Design

1. Determine Problem / Topic Area
2. What *causal question* are you seeking to answer?
3. What would your ideal experiment be?
4. Where can you (a) measure your outcome variable, and (b) find variation in your treatment variable?
5. What would a feasible design be?
6. What does an answer look like?

Backwards Design

1. Determine Problem / Topic Area
2. What *causal question* are you seeking to answer?
3. What would your ideal experiment be?
4. Where can you (a) measure your outcome variable, and (b) find variation in your treatment variable?
5. What would a feasible design be?
6. What does an answer look like?
7. What variables do you need to generate that answer?

Backwards Design

1. Determine Problem / Topic Area
2. What *causal question* are you seeking to answer?
3. What would your ideal experiment be?
4. Where can you (a) measure your outcome variable, and (b) find variation in your treatment variable?
5. What would a feasible design be?
6. What does an answer look like?
7. What variables do you need to generate that answer?
8. What data contains those variables?

Step 0: Define the Problem / Topic

Why are you doing this project?

Step 0: Define the Problem / Topic

Why are you doing this project?

What motivates your investigation?

Step 0: Define the Problem / Topic

Why are you doing this project?
What motivates your investigation?

Examples:

- Polling places change locations all the time. Are there consequences for voters? Could this be manipulated for political gain?
- My company (Zarbucks) recently did major store renovations. Did they improve sales?

Step 1: What question are you seeking to answer?

The tools of data science are fundamentally designed to answer questions,

Step 1: What question are you seeking to answer?

The tools of data science are fundamentally designed to answer questions, so to before you pick your tools, you have to decide what question you wish to answer.

Step 1: What question are you seeking to answer?

The tools of data science are fundamentally designed to answer questions, so to before you pick your tools, you have to decide what question you wish to answer.

$\Rightarrow$ The MOST important part of your project

Step 1: What question are you seeking to answer?

Most important because:

Step 1: What question are you seeking to answer?

Most important because:

- if you can't define the question you are seeking to answer, you'll find yourself lost in your data, or worse

Step 1: What question are you seeking to answer?

Most important because:

- if you can't define the question you are seeking to answer, you'll find yourself lost in your data, or worse
- after finishing your project, you'll realizing the question you answered doesn't help solve the problem that motivated you.

Step 1: What question are you seeking to answer?

Most important because:

- if you can't define the question you are seeking to answer, you'll find yourself lost in your data, or worse
- after finishing your project, you'll realizing the question you answered doesn't help solve the problem that motivated you.

$\Rightarrow$ Invest in this stage of your project *before* you dive into the data!

Step 1: What question are you seeking to answer?

A critical feature of a good question is that it is *tractable* and *answerable* in a data science project.

- If your question does not directly imply a course of action in your data science project, it's too vague.

Step 1: What question are you seeking to answer?

Not answerable:

- What happens when you move a polling place?
- Are Zarbucks renovations worth it?

Step 1: What question are you seeking to answer?

Not answerable:

- What happens when you move a polling place?
- Are Zarbucks renovations worth it?

Answerable:

- What is the effect of moving a polling place on voter turnout?
- Do Zarbucks renovations result in increased sales?

Step 1: What question are you seeking to answer?

How do I know if my answer is answerable / tractable?

Step 1: What question are you seeking to answer?

How do I know if my answer is answerable / tractable?

1. Can you hypothesize an answer to your question?
   i.e. Can you state what you think might be the answer to your question?

Step 1: What question are you seeking to answer?

How do I know if my answer is answerable / tractable?

1. Can you hypothesize an answer to your question?
   i.e. Can you state what you think might be the answer to your question?
2. Can you imagine what the answer to your question looks like?

Step 2: What would the ideal experiment look like?

Ignore feasibility – if you were a god, what experiment would you want to run?

Step 2: What would the ideal experiment look like?

Ignore feasibility – if you were a god, what experiment would you want to run? Why is this helpful?

- Helps you think through what your treatment is,
- What your outcome is, and
- What *variation* in your treatment looks like.

Separately from worrying about *feasibility*.

Step 2: What would the ideal experiment look like?

- Randomly assign voters to have new polling places or not
- Randomly renovate stores and observe consequences

Step 3: Pick a study context

Answering causal questions requires variation in treatment assignment.

Step 3: Pick a study context

Answering causal questions requires variation in treatment assignment. Need a place where you can:

- Observe variation in treatment assignment, and
- Measure your target outcome

Step 3: Pick a study context

- North Carolina: Polling places move, and North Carolina has public data on voter registration, polling place locations, and turnout

Step 3: Pick a study context

- North Carolina: Polling places move, and North Carolina has public data on voter registration, polling place locations, and turnout
- Zarbucks: If we worked for Zarbucks: was there variation in when people had renovations? If not, can we get data from before and after renovations?

Step 4: What design might be feasible?

In your ideal experiment, you identified your "treatment."

Now you need to find variation in your treatment somewhere in the real world.

Step 4: What design might be feasible?

Once you find variation, decide what comparisons you want to make.

Common strategies:

- Experiment (e.g. AB testing)
- Pre-Post (variation over time within a unit)
- Cross-sectional (variation across units)
- Differences-in-differences (variation across time and units)

Step 4: What design might be feasible?

North Carolina

- If data from one election, then cross-section (maybe matching?)
  Could use census data for community demographics
- If data from multiple elections, we can do a difference-in-difference

Step 4: What design might be feasible?

North Carolina

- If data from one election, then cross-section (maybe matching?)
  Could use census data for community demographics
- If data from multiple elections, we can do a difference-in-difference

Zarbucks:

- If all renovations happened at once, pre-post
- If variation in timing, difference-in-difference

Step 4: What design might be feasible?

North Carolina

- If data from one election, then cross-section (maybe matching?)
  Could use census data for community demographics
- If data from multiple elections, we can do a difference-in-difference

Zarbucks:

- If all renovations happened at once, pre-post
- If variation in timing, difference-in-difference
- If you have POWER, randomize rollouts

Step 5: What does an answer to your question look like?

Write down what the answer to your question will look like!

Step 5: What does an answer to your question look like?

Write down what the answer to your question will look like!

- A figure
- A table or regression
- A dataset with predicted values

Step 5: What does an answer to your question look like?

Write down what the answer to your question will look like!

- A figure
- A table or regression
- A dataset with predicted values

$\Rightarrow$ Ask yourself: if I gave that to my stakeholder / put it in a paper, would people be pleased?

Step 5: What does an answer to your question look like?

Write down what the answer to your question will look like!

- A figure
- A table or regression
- A dataset with predicted values

$\Rightarrow$ Ask yourself: if I gave that to my stakeholder / put it in a paper, would people be pleased?
(OK, they might want robustness, and extensions, but at its core, is this an answer?)

Step 4: What does an answer to your question look like?

But it's not enough to imagine *one* answer. You should be able to imagine what an answer to your question looks like if your hypothesis is true and the if your hypothesis is false.

Step 4: What does an answer to your question look like?

But it’s not enough to imagine *one* answer. You should be able to imagine what an answer to your question looks like if your hypothesis is true and the if your hypothesis is false. Otherwise your question isn’t falsifiable!

Write down what your answer looks like if your hypothesis is true, *and* if it’s false!

Step 5: What do you need to generate that answer?

Congratulations! You've just specified the goal of your analysis!

Step 5: What do you need to generate that answer?

Congratulations! You've just specified the goal of your analysis! In my view, that is actually the hardest part of being a good data scientist.

Step 5: What do you need to generate that answer?

Congratulations! You've just specified the goal of your analysis! In my view, that is actually the hardest part of being a good data scientist.

...Though probably not the part that will take up the majority of your time.

Step 5: What do you need to generate that answer?

So you now have in mind a table you want to generate. What data and variables do you need to create that result?

Step 5: What do you need to generate that answer?

So you now have in mind a table you want to generate. What data and variables do you need to create that result?

Step 5: What do you need to generate that answer?

So you now have in mind a table you want to generate. What data and variables do you need to create that result? For each variable, specify:

1. What do you need the variable to measure?
2. For what population do you need the variable defined?

Step 5: Where can you get those variables?

1. Where can you get those variables?, and
2. How will you relate your different datasets?

In Class

You’ve been hired by an real estate agent industry group that wants to know if it should campaign against laws that require mandatory disclosure of problems with houses.

They aren’t sure if mandatory disclosure will increase sales (since buyers will be less worried about hidden problems), or decrease sales (since more problems may be made evident to customers).