Putting It All Together

Nick Eubank

Data Science is the art of answering questions about the world using quantitative data.

1. Move from problems to questions

1. Move from problems to questions
2. Recognize the type of question you are asking

1. Move from problems to questions
2. Recognize the type of question you are asking
3. Understand how to choose the right tool to answer the question you are asking

Types of Questions

Types of Questions

1. Descriptive Questions
   Identifying patterns in the world

Types of Questions

1. Descriptive Questions
   Identifying patterns in the world
2. Causal Questions
   Understanding the effects of manipulations

Types of Questions

1. Descriptive Questions
   Identifying patterns in the world
2. Causal Questions
   Understanding the effects of manipulations
3. Predictive Questions
   Making out-of-sample predictions

Types of Questions

1. Descriptive Questions
   Identifying Patterns in the World
2. Causal Questions
   Understanding the effects of manipulations
3. Predictive Questions
   Making out-of-sample predictions

Purpose of Descriptive Questions

Purpose of Descriptive Questions

Help identify areas for further investigation / prioritization

Descriptive Questions

Descriptive Questions

When answering a descriptive question, you are always doing dimensionality reduction.

Descriptive Questions

When answering a descriptive question, you are always doing dimensionality reduction.

- Formally: PCA

Descriptive Questions

When answering a descriptive question, you are always doing dimensionality reduction.

- Formally: PCA
- Informally: picking what variables to plot, summary statistics to include, etc.

Descriptive Questions

When answering a descriptive question, you are always doing dimensionality reduction.

Descriptive Questions

When answering a descriptive question, you are always doing dimensionality reduction. And so you will necessarily be discarding information, and so it is your responsibility to:

Descriptive Questions

When answering a descriptive question, you are always doing dimensionality reduction. And so you will necessarily be discarding information, and so it is your responsibility to:

- Ensure what you present faithfully represents the patterns in the underlying data.

Descriptive Questions

When answering a descriptive question, you are always doing dimensionality reduction. And so you will necessarily be discarding information, and so it is your responsibility to:

- Ensure what you present faithfully represents the patterns in the underlying data.
- You make sure to look for ethically-salient patterns (differences by race, ethnicity, gender, etc.)

```
df.head()
```

	example1_x	example1_y	example2_x	example2_y	example3_x	example3
0	32.331110	61.411101	51.203891	83.339777	55.993030	79.277
1	53.421463	26.186880	58.974470	85.499818	50.032254	79.013
2	63.920202	30.832194	51.872073	85.829738	51.288459	82.435
3	70.289506	82.533649	48.179931	85.045117	51.170537	79.165
4	34.118830	45.734551	41.683200	84.017941	44.377915	78.164

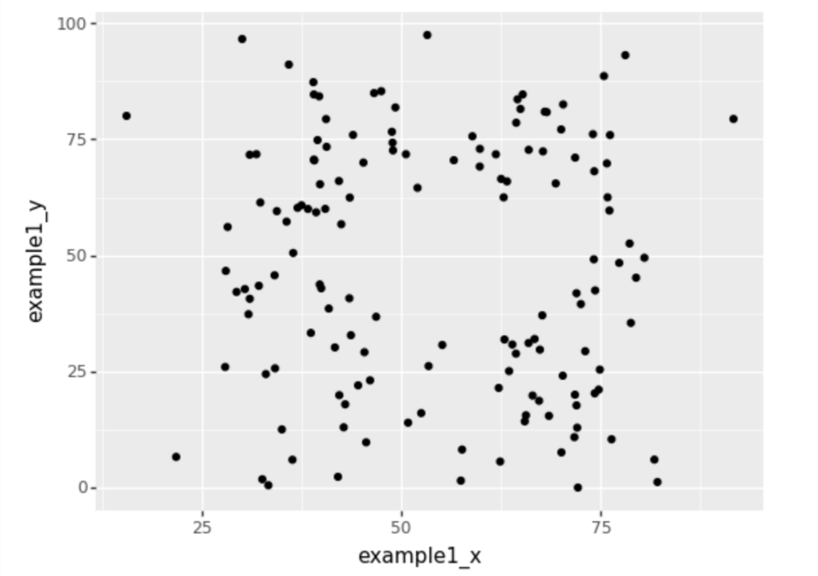
5 rows × 26 columns

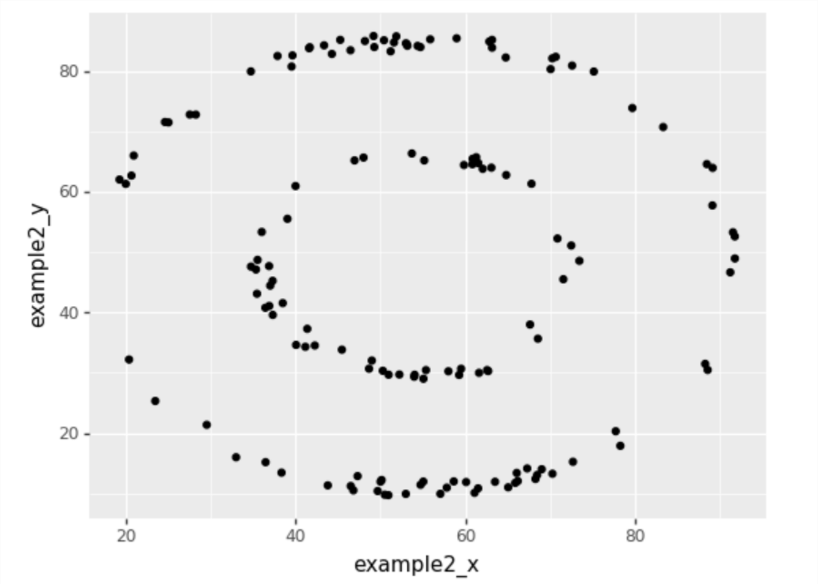
DATA SET 1
Mean x: 54.27
Mean y: 47.83
Std Dev x: 16.77
Std Dev y: 26.94
Correlation: −0.06

DATA SET 2
Mean x: 54.27
Mean y: 47.83
Std Dev x: 16.77
Std Dev y: 26.94
Correlation: −0.07

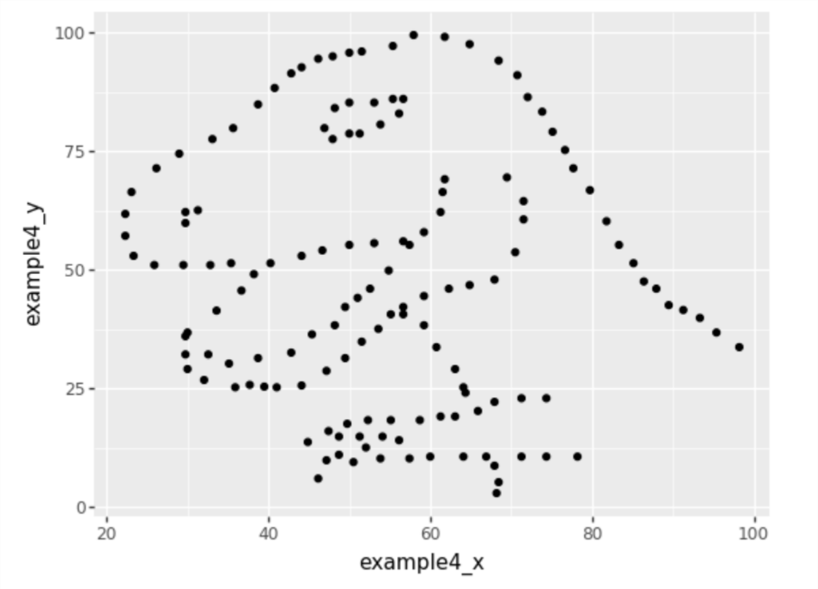
DATA SET 3
Mean x: 54.27
Mean y: 47.84
Std Dev x: 16.76
Std Dev y: 26.93
Correlation: −0.07

DATA SET 4
Mean x: 54.26
Mean y: 47.83
Std Dev x: 16.77
Std Dev y: 26.94
Correlation: −0.06


example1_y
100
75
50
25
0
25
50
75
example1_x



example2_y
80
60
40
20
20
40
60
80
example2_x


100
75
50
25
0
20
40
60
80
100
example4_x
example4_y

These all had the same means, standard deviations, and correlations.

These all had the same means, standard deviations, and correlations.

But if you had only reported those summary statistics, you would not have been faithfully representing the data.

Types of Questions

1. Descriptive Questions
   Identifying Patterns in the World
2. Causal Questions
   Understanding the <u>effects</u> of manipulations
3. Predictive Questions
   Making out-of-sample predictions

Purpose of Causal Questions

Purpose of Causal Questions

Causality is about predicting the consequences of manipulations.

Purpose of Causal Questions

Causality is about predicting the consequences of manipulations.
$\rightarrow$ Generally asked in anticipation of undertaking some action.

Causal Inference

Causal Inference

Causal inference is hard.

Causal inference is hard because of the Fundamental Problem of Causal Inference:

Causal Inference

Causal inference is hard because of the Fundamental Problem of Causal Inference:

We say that “X caused Y” if:

Causal inference is hard because of the Fundamental Problem of Causal Inference:

We say that "X caused Y" if:

1. When X is present, we see Y

Causal inference is hard because of the Fundamental Problem of Causal Inference:

We say that "X caused Y" if:

1. When X is present, we see Y
2. When X is not present, we don't see Y

Causal Inference

Causal inference is hard because of the Fundamental Problem of Causal Inference:
We say that "X caused Y" if:

1. When X is present, we see Y
2. When X is not present, we don't see Y

To know if X causes Y, we would have to see both a world with X, and a world without X, and that's impossible.

Causal Inference

Because we can never see both a world with X, and a world without X, Because we can never see both a world with X, and a world without X, we need to find settings that approximate one of these states of the world.

Counter-factuals: settings with same potential outcomes, but different realizations of treatment.

Ways of Finding Good Counter-Factuals

Ways of Finding Good Counter-Factuals

1. Randomized Control Trials
   Law of large numbers $\rightarrow$ same potential outcomes for C & T

Ways of Finding Good Counter-Factuals

1. Randomized Control Trials
   Law of large numbers $\rightarrow$ same potential outcomes for C & T
2. Regression
   Statistically adjust for baseline differences $\rightarrow$ same potential outcomes after adjustments

Ways of Finding Good Counter-Factuals

1. Randomized Control Trials
   Law of large numbers $\rightarrow$ same potential outcomes for C & T
2. Regression
   Statistically adjust for baseline differences $\rightarrow$ same potential outcomes after adjustments
3. Matching
   Statistically adjust for baseline differences $\rightarrow$ same potential outcomes after adjustments

Ways of Finding Good Counter-Factuals

1. Randomized Control Trials
   Law of large numbers $\rightarrow$ same potential outcomes for C & T
2. Regression
   Statistically adjust for baseline differences $\rightarrow$ same potential outcomes after adjustments
3. Matching
   Statistically adjust for baseline differences $\rightarrow$ same potential outcomes after adjustments
4. Differences-in-Differences
   Adjust for pre-existing baseline differences $\rightarrow$ same potential outcomes in trends

Assumptions

Validity of causal inferences depends on whether assumptions about potential outcomes are met.

Assumptions

Validity of causal inferences depends on whether assumptions about potential outcomes are met.
Fundamentally unverifiable, so evaluation requires critical thinking!

Applies both to your projects, but also anything else you read!

Validity

- **Internal Validity**: have assumptions been met? Did study estimate a causal effect?

Validity

- **Internal Validity**: have assumptions been met? Did study estimate a causal effect?
- **External Validity**: Do I think these causal effects would generalize?

“Correlation does not imply causation”

~~“Correlation does not imply causation”~~

Correlation does not <u>necessarily</u> imply causation, but…

~~“Correlation does not imply causation”~~

Correlation does not necessarily imply causation, but...

- when certain assumptions are met, correlation does imply causation.

~~“Correlation does not imply causation”~~

Correlation does not <u>necessarily</u> imply causation, but...

- when certain assumptions are met, correlation does imply causation.

And now you know those assumptions and how to evaluate them!

Types of Questions

1. Descriptive Questions
   Identifying Patterns in the World
2. Causal Questions
   Understanding the effects of manipulations
3. Predictive Questions
   Making out-of-sample predictions

Purpose of Predictive Questions

Purpose of Predictive Questions

- Anticipate outcomes for subsequent intervention (e.g. high value customers, expensive patients)
  Supervised Machine Learning

Purpose of Predictive Questions

- Anticipate outcomes for subsequent intervention
  (e.g. high value customers, expensive patients)
  Supervised Machine Learning
- Predict result of your actions
  (e.g. advertisements, sales, web design)
  Causal inference tools

Generalizability

Generalizability

1. Parameter values beyond our training data
   Out-of-sample extrapolations

Generalizability

1. Parameter values beyond our training data

   Out-of-sample extrapolations

2. New settings

   Different places, different products

Generalizability

1. Parameter values beyond our training data
   Out-of-sample extrapolations
2. New settings
   Different places, different products
3. Different dynamics
   Adversarial users

Bias

- What constitutes bias is context dependent

Bias

- What constitutes bias is context dependent
- Bias doesn't come from ML models malfunctioning
  If training data biased, ML is designed to replicate!

Bias

- What constitutes bias is context dependent
- Bias doesn't come from ML models malfunctioning
  If training data biased, ML is designed to replicate!
- Interpretable models can help make bias visible
  Often with no performance cost, and benefits to maintainability

Where does that leave you?

If a stake-holder comes to you with a problem…

Where does that leave you?

If a stake-holder comes to you with a problem…

1. articulate a question whose answer will help address their need,

Where does that leave you?

If a stake-holder comes to you with a problem...

1. articulate a question whose answer will help address their need,
2. depending on the type of question, you can reach for the right tool,

Where does that leave you?

If a stake-holder comes to you with a problem...

1. articulate a question whose answer will help address their need,
2. depending on the type of question, you can reach for the right tool,
3. know the types of conceptual problems to bear in mind when answering the question.