IPEDS Meets Data Science

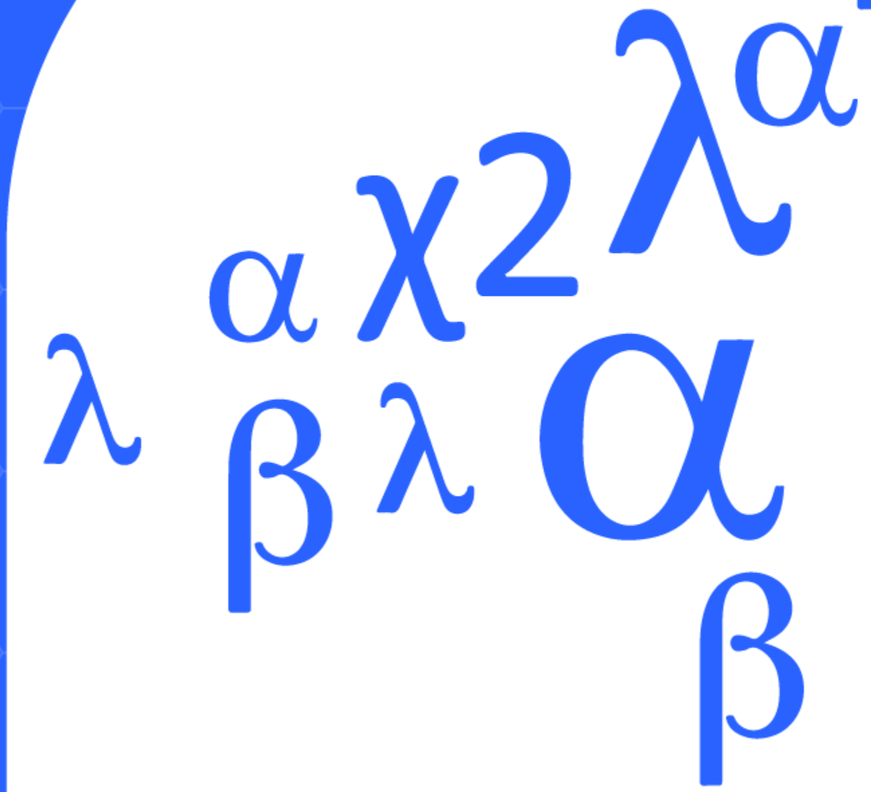
New-ish Methods For Peer Groupings

Adam Ross Nelson JD PhD

August 2021

Click below to find me online!

@adamrossnelson



For Further Reading

Click The Resources Below For Further Reading

A Cookbook: Using Distance To Measure Similarity

Applied Distance Measures; Building Higher Education Comparison Groups

Sourcing Federal Data: Higher Education Data

Why Do We Automate Data Collection?

Adam Ross Nelson

Data Scientist, Consultant

- First job ever ever was as a teacher of English as a foreign language in 1998-99.
- Became a data scientist after finishing a PhD.
- Three First Names!

The Question We Will Answer

Which Institution Is "Most Like" Institution D?

	Inst	Size	Cost	Accept Rt
0	Institution A	19000	22000	0.25
1	Institution B	11500	19000	0.45
2	Institution C	7750	12000	0.76
3	Institution D	23000	10500	0.99

A Related Question

How can we build comparison groups, empirically?

The methods I will discuss today work well with either qualitative or quantitative data.

You Might Ask...

Is This A Solution Looking For A Problem?

No, I don't think so.

The NCES specifies ≈ 240-260 comparison groups for their "Data Feedback Reports."

"The NCES automatic comparison group for degree-granting institutions is based on control type, Carnegie Classification, and enrollment size." (Source).

Agenda

- Distance as a Measure Similarity

- Euclidean Distance
- Jaccardian Distance

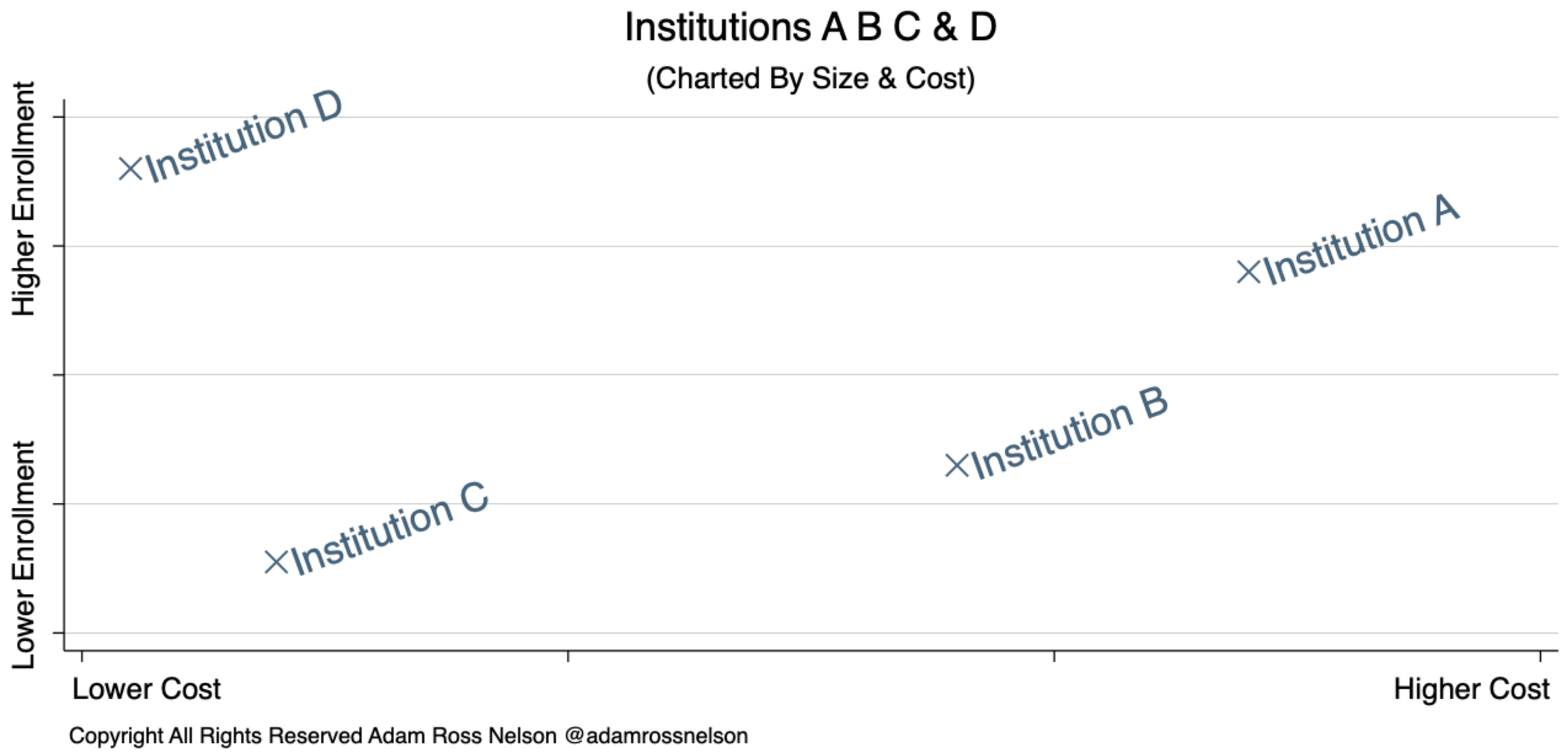
- Live Demonstration

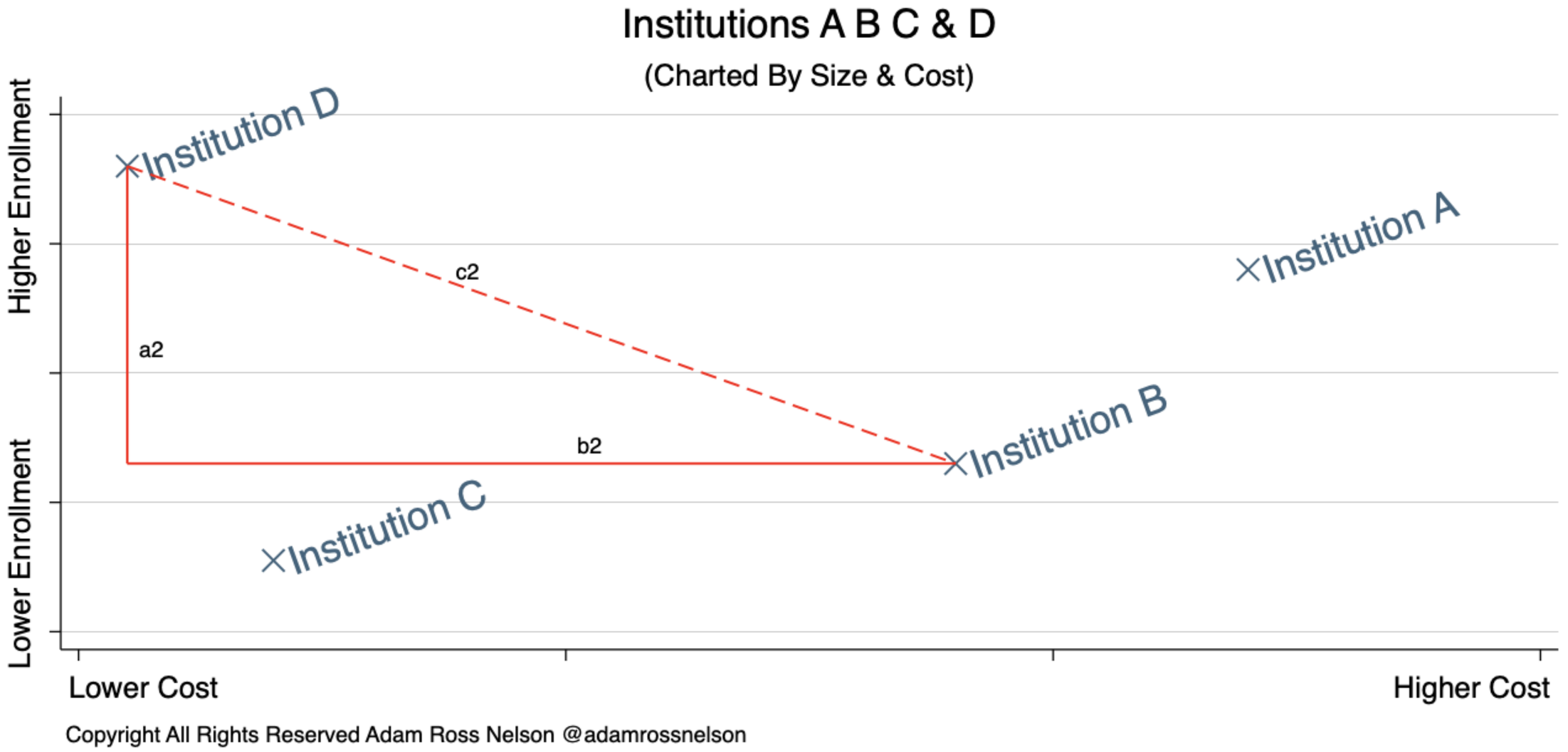
- Discussion + Q&A

The Question We Will Answer

Which Institution Is "Most Like" Institution D?

	Inst	Size	Cost	Accept Rt
0	Institution A	19000	22000	0.25
1	Institution B	11500	19000	0.45
2	Institution C	7750	12000	0.76
3	Institution D	23000	10500	0.99
		- Y	X -	





Agenda

- Distance as a measure similarity
- Euclidean Distance
- Jaccardian Distance
- Live Demonstration
- Discussion + Q&A

The Math (2 Dimensions)

	Inst	Size	Cost
0	Institution A	19000	22000
1	Institution B	11500	19000
2	Institution C	7750	12000
3	Institution D	23000	10500

$$Distance = \sqrt{(23,000 - 11500)^2 + (10,500 - 19,000)^2}$$
$$Distance = \sqrt{(11,500)^2 + (-8,500)^2}$$
$$Distance = \sqrt{132,250,000 + 72,250,000}$$
$$Distance = \sqrt{204,500,000}$$
$$Distance = 14,300.3$$

The Math (2 Dimensions)

	Inst	Size	Cost
0	Institution A	19000	22000
1	Institution B	11500	19000
2	Institution C	7750	12000
3	Institution D	23000	10500

$$d = \sqrt[2]{(rise^2 + run^2)}$$
$$d = \sqrt[2]{(y_1 - y_2)^2 + (x_1 - x_2)^2}$$
$$d = \sqrt[2]{(23k - 7.75k)^2 + (10.5k - 12k)^2}$$
$$d = \sqrt[2]{232562500 - 2250000)}$$
$$d = \sqrt[2]{230312500}$$
$$d = 15176.05$$

The Math (3 Dimensions)

	Inst	Size	Cost	Accept Rt
0	Institution A	19000	22000	0.25
1	Institution B	11500	19000	0.45
2	Institution C	7750	12000	0.76
3	Institution D	23000	10500	0.99

$$d = \sqrt[2]{(rise^2 + run_1^2 + run_2^2)}$$

$$d = \sqrt[2]{(y_1 - y_2)^2 + (x_1 - x_2)^2 + (z_1 - z_2)^2}$$

$$d = \sqrt[2]{(23k - 7.75k)^2 + (10.5k - 12k)^2 + (.99 - .76)^2}$$

...

Hypothetical Data

Which Institution Is "Most Like" Institution D?

	Inst	Size	Cost	Accept Rt	Euclidians*
0	Institution A	19000	22000	0.25	3.611825
1	Institution B	11500	19000	0.45	3.233217
2	Institution C	7750	12000	0.76	2.682701
3	Institution D	23000	10500	0.99	0.000000

Standardization Does Two Things:

1) It converts units to 'standard deviations.'

2) It rescales each variable so that none will overpower the others in the analysis.

* These results standardized by z-scores.

Agenda

- Distance as a measure similarity
- Euclidean Distance
- Jaccardian Distance
- Live Demonstration
- Discussion + Q&A

Jacardian Multi-Demensional

Which Institution Is "Most Like" Institution D?

	Inst	Size	Cost	Accept Rt	Euclidians
0	Institution A	19000	22000	0.25	3.611825
1	Institution B	11500	19000	0.45	3.233217
2	Institution C	7750	12000	0.76	2.682701
3	Institution D	23000	10500	0.99	0.000000

	Inst	isBig	isExpensive	isSelect
0	Institution A	1	1	1
1	Institution B	0	1	1
2	Institution C	0	0	0
3	Institution D	1	0	0

Jacardian Multi-Demensional

Which Institution Is "Most Like" Institution D?

	Inst	isBig	isExpensive	isSelect
0	Institution A	1	1	1
1	Institution B	0	1	1
2	Institution C	0	0	0
3	Institution D	1	0	0

Count the number of matches.	1
Divide matches by the numnber of poss. matches.	1/3
Subtract the previous result from 1.0.	1-(1/3)

Jacardian Multi-Demensional

Which Institution Is "Most Like" Institution D?

	Inst	isBig	isExpensive	isSelect	Jaccardian
0	Institution A	1	1	1	0.6666
1	Institution B	0	1	1	1.0000
2	Institution C	0	0	0	0.3333
3	Institution D	1	0	0	1.0000

Count the number of matches.	1
Divide matches by the numnber of poss. matches.	1/3
Subtract the previous result from 1.0.	1-(1/3)

Agenda

- Distance as a measure similarity
- Euclidean Distance
- Jaccardian Distance
- Live Demonstration
- Discussion + Q&A

Agenda

- Distance as a measure similarity
- Euclidean Distance
- Jaccardian Distance
- Live Demonstration
- Discussion + Q&A

For Further Reading

Click The Resources Below For Further Reading

A Cookbook: Using Distance To Measure Similarity

Applied Distance Measures; Building Higher Education Comparison Groups

Sourcing Federal Data: Higher Education Data

Why Do We Automate Data Collection?

Optimize Recruitment

More ideas at the cusps of institutional research & data science

Use distance measures to find similarities among high schools that send students to your institution; look for more schools that have yet to send students but that are similar to those who are (i.e. untapped recruitment opportunities).

Instructional Support

More ideas at the cusps of institutional research & data science

Use distance measures to find similarities among faculty. This approach will be useful in finding comparison groups among faculty. Use these groups as a method to structure or build instructional colaborations.

More ideas at the cusps of institutional research & data science

Use distance measures to find similarities among students. This approach may be useful in finding comparison groups among students. Use these groups as you look to measure student learning and other outcomes.

More ideas at the cusps of institutional research & data science

Use IPEDS, to ascertain the number of potential transfer students who completed a credential at or stopped out at area institutions. This sizes the transfer student market and helps decide where to focus or re-focus recruitment efforts.

At The Cusp of Data Science

Adam Ross Nelson JD PhD, Representative projects at the cusps of data science and institutional research

- Satisfactory academic progress: Predictive and identified students who may be at risk of not making financial aid's satisfactory academic progress requirements. This predictive model helped deliver academic support to those who were in most need of that support.
- Student learning: Developed measures of and methods suitable for assessing the learning students experience as a result of of extra-curricular activity.
- Research administration: Served as the scientist for association grants (over $2 million) that funded multiple nation-wide educational interventions and randomized controlled trials - evaluation strategies involved IPEDS, National Student Clearinghouse, and related data.