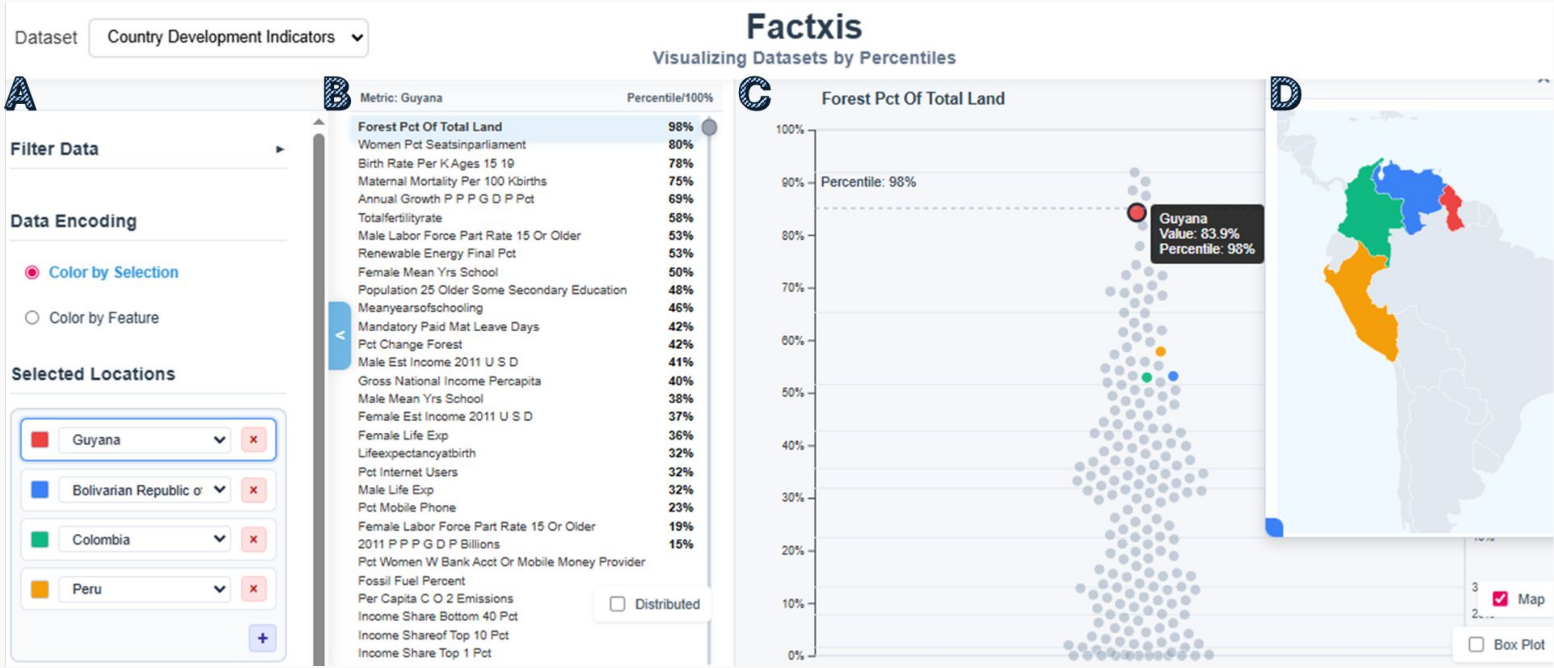
Factxis: A Percentile-Ranking System for Multivariate Tabular Data



Atlanta Regional Commission Lunch and Learn

Monday Sept. 21, 2026

Colin Vu and Clio Andris

friendlycities.gatech.edu

INTRODUCTIONS

COLIN VU
MS Student, School of Computer Science

CLIO ANDRIS
Associate Prof., School of City & Regional Planning & Interactive Computing

PREVIOUS WORK TOGETHER

Clio Andris • You
Associate Professor at Georgia Institute of Technology
1d •

We want to share a free data resource with you all: https://lnkd.in/eY4XJaAE

"County Buddy: A Companion Dataset for Socioeconomic Data Analysis and Exploration of U.S. Datasets"

CountyBuddy is a free data spreadsheet with locations of four different types of populations: Incarcerated People, Colleges, Military, and Native Americans at county and tract levels. The data has the institution's population and name (e.g., Fort Drum, Cornell University).

In total, it accounts for 8,846,310 U.S. residents across 5,067 institutions in 924 counties and 2,797 census tracts (years: mostly the 2020 decennial census). We worked on it for over a year.

This spreadsheet can be used to add variables to regressions and statistical models, describe disparities, explain geographic and colocation trends, and to provide tooltips and labels to interactive maps. It was developed by Colin Vu and Leila Baniassad.

County Buddy: A Companion Dataset for Socioeconomic Data Analysis and Exploration of U.S. Datasets

Colin Vu[a,b], Leila Baniassad[a], Clio Andris[a,b]
[a]College of Computing, Georgia Institute of Technology, Atlanta, Georgia, USA
[b]Corresponding authors: colinhvu@gmail.com, clio@gatech.edu

County Buddy: A Companion Dataset for Socioeconomic Data Analysis and Exploration of U.S. Datasets

Colin Vu[a,b], Leila Baniassad[a], Clio Andris[a,b]
[a]College of Computing, Georgia Institute of Technology, Atlanta, Georgia, USA
[b]Corresponding authors: colinhvu@gmail.com, clio@gatech.edu

Keywords: Demographics, Geospatial Data, Incarceration, Prisons, Universities, Colleges, Students, Military, Native Americans, Geographic Information Systems (GIS), Data Visualization, Geospatial Mapping.

Introduction

County Buddy is a new dataset that represents information on the presence, number of, and institution names of special populations at the U.S. county and census tract levels. These *distinguished populations* include the **incarcerated population**, **college student population**, **military population**, and **Native American population**.

The objective of this dataset is to provide geographic and demographic context to help explain variations in socio-economic indicators (such as life expectancy, household income, educational attainment, etc.). The data on distinguished populations can help provide a potential 'explainer' for why these values (e.g. high marriage rates) appear or provide a landscape to put the values into context.

By Anna Traykova • 9/18/2026

Georgia Tech School of City & Regional Planning
3,175 followers
2d •

We're excited to share this new open data resource from the research team of Prof. Clio Andris - County Buddy: A Companion Dataset for Socioeconomic Data Analysis and Exploration of U.S. Datasets. ...more

Clio Andris • You
Associate Professor at Georgia Institute of Technology
2w •
We want to share a free data resource with you all: https://lnkd.in/eY4XJaAE
"County Buddy: A Companion Dataset for Socioeconomic Data Analysis ...more

Vu, Colin; Andris, Clio; Baniassad, Leila, 2025, "County Buddy: A Companion Dataset for Socioeconomic Data Analysis and Exploration of U.S. Datasets", https://doi.org/10.7910/DVN/V7LNJK, Harvard Dataverse, V1.

AGENDA

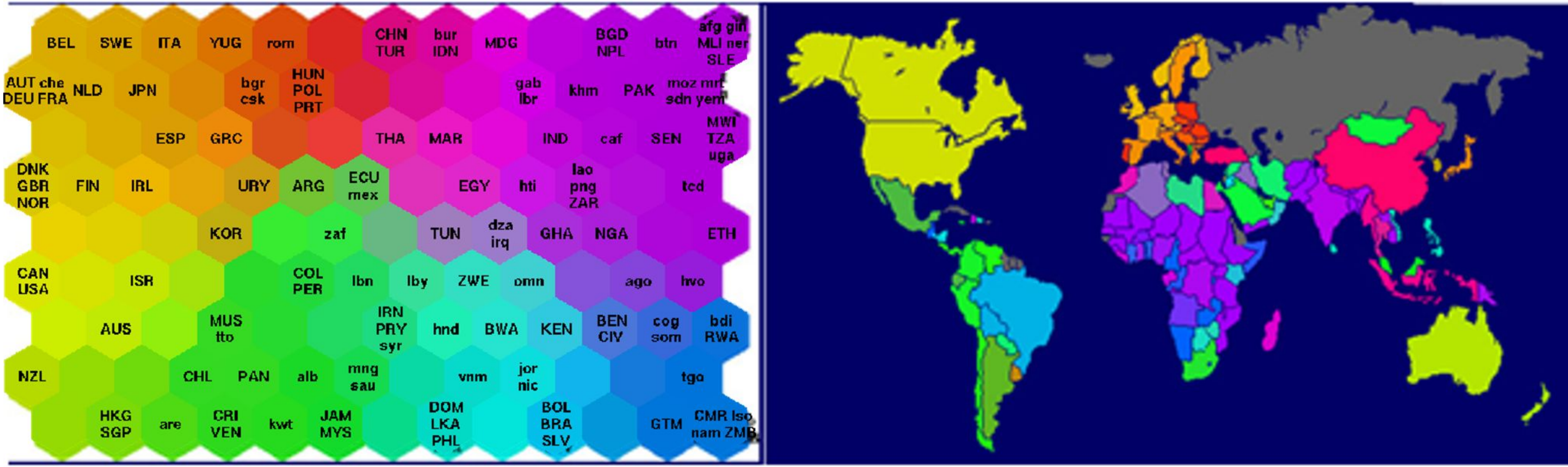
1. Objective
2. Development Process & Demo
3. User Study
4. Open Exploration

What problem are we trying to solve?

NAME_1	NPU	X_LON	Y_LAT	TOTALHH_E2	TOTALHH_M2	HHINCLT10K	HHINCLT1_1	PHHINCLT10	PHHINCLT_1	HHINC10_15	HHINC10_16	PHHINC10_1	PHHINC10_2	HHINC15_25	HHINC15_26	PHHINC15_2	PHHINC15_3	HHINC25_35	HHINC25_36	PHHINC25_3	PHHINC25_4	HHINC35_50	HHINC35_51	PHHINC35_5	PHHINC3
Paces	A	-84.444838	33.847267	1660	217	19	22	1.2	1.3	9	16	0.5	1	20	25	1.2	1.5	20	29	1.2	1.7	113	99	6.8	
Mt. Paran/Northside	A	-84.422304	33.862292	2019	270	32	36	1.6	1.8	24	33	1.2	1.6	134	137	6.6	6.7	66	50	3.3	2.4	252	179	12.5	
Chastain Park	A	-84.397172	33.864333	1426	211	12	25	0.9	1.7	29	47	2.1	3.3	14	26	1	1.8	43	49	3	3.4	55	53	3.8	
Peachtree Heights West	B	-84.389733	33.834482	2589	334	223	137	8.6	5.2	129	80	5	3	205	155	7.9	5.9	70	61	2.7	2.3	269	157	10.4	
South Tuxedo Park	B	-84.382244	33.846104	2608	314	144	116	5.5	4.4	31	31	1.2	1.2	214	176	8.2	6.7	122	78	4.7	3	124	80	4.8	
East Chastain Park	B	-84.384025	33.867959	942	166	10	23	1	2.5	29	46	3	4.8	5	15	0.5	1.6	26	36	2.7	3.8	41	48	4.4	
North Buckhead	B	-84.369662	33.863442	6142	596	265	168	4.3	2.7	113	82	1.8	1.3	304	184	4.9	3	270	207	4.4	3.3	271	164	4.4	
Brookhaven	B	-84.353097	33.868133	801	181	5	19	0.7	2.4	0	14	0	1.7	0	19	0	2.4	17	51	2.1	6.4	30	45	3.8	
Lenox	B	-84.357432	33.849699	2689	247	116	83	4.3	3.1	35	39	1.3	1.4	204	117	7.6	4.3	189	119	7	4.4	133	89	5	
Peachtree Park	B	-84.367723	33.838667	2421	320	209	143	8.6	5.8	72	50	3	2	55	52	2.3	2.1	119	110	4.9	4.5	113	133	4.7	
Pine Hills	B	-84.353907	33.835986	5142	550	417	204	8.1	3.9	125	107	2.4	2.1	112	101	2.2	2	121	96	2.3	1.8	420	210	8.2	
Garden Hills	B	-84.375696	33.830582	2015	242	111	94	5.5	4.6	91	81	4.5	4	107	121	5.3	6	147	97	7.3	4.7	165	103	8.2	
Peachtree Hills	B	-84.382062	33.820337	2392	372	13	32	0.5	1.3	75	69	3.1	2.8	40	49	1.7	2	194	120	8.1	4.9	312	268	13	
Lindbergh/Morosgo	B	-84.365855	33.823614	3976	608	242	130	6.1	3.1	135	111	3.4	2.7	147	119	3.7	2.9	707	534	17.8	13.2	367	181	9.2	
Wesley Battle	C	-84.440004	33.823944	1336	226	4	16	0.3	1.2	26	35	1.9	2.6	12	26	0.9	1.9	71	82	5.3	6.1	65	49	4.8	
Cross Creek	C	-84.425035	33.817234	1626	304	23	36	1.4	2.2	67	80	4.1	4.9	21	35	1.3	2.1	85	76	5.2	4.5	101	75	6.2	
Brandon	C	-84.417452	33.833259	1834	283	39	44	2.1	2.4	37	41	2	2.2	170	167	9.3	9	10	29	0.5	1.6	98	110	5.4	
Peachtree Battle Allian...	C	-84.40062	33.830278	1071	158	29	31	2.7	2.9	18	32	1.7	3	54	61	5	5.7	16	27	1.5	2.5	34	37	3.2	
Collier Hills	C	-84.401344	33.812089	2225	239	109	93	4.9	4.1	58	63	2.6	2.8	39	62	1.7	2.8	201	135	9	6	267	162	12	
Wildwood (NPU-C)	C	-84.414298	33.814672	1839	199	49	78	2.7	4.2	0	18	0	1	6	26	0.3	1.4	139	120	7.6	6.5	216	163	11.8	
Bolton	D	-84.464374	33.814836	2006	240	82	81	4.1	4	68	60	3.4	3	76	74	3.8	3.7	86	97	4.3	4.8	77	59	3.9	
Underwood Hills	D	-84.426059	33.807272	1584	269	22	56	1.4	3.5	10	27	0.7	1.7	29	62	1.8	3.9	67	73	4.2	4.6	112	104	7.1	
Hills Park	D	-84.430923	33.798031	2673	419	147	125	5.5	4.6	81	125	3	4.7	101	111	3.8	4.1	165	155	6.2	5.7	188	199	7	
Ansley Park	E	-84.380069	33.796895	1910	208	68	43	3.6	2.2	18	32	0.9	1.7	146	72	7.6	3.7	49	42	2.5	2.2	117	87	6.2	
Georgia Tech	E	-84.402119	33.77805	1699	286	275	175	16.2	10	32	49	1.9	2.9	98	69	5.7	3.9	65	37	3.8	2.1	353	209	20.8	1
Brookwood	E	-84.396734	33.804242	2076	263	224	171	10.8	8.1	50	45	2.4	2.1	28	37	1.4	1.8	71	91	3.4	4.3	198	166	9.5	
Loring Heights	E	-84.401403	33.794943	2955	409	299	153	10.1	5	152	157	5.1	5.3	90	104	3.1	3.5	276	201	9.3	6.7	303	219	10.2	
Brookwood Hills	E	-84.386078	33.808775	1802	244	71	70	3.9	3.9	97	62	5.4	3.3	19	40	1.1	2.2	171	128	9.5	7	144	106	8	
Home Park	E	-84.39882	33.785456	2152	307	290	162	13.5	7.3	88	70	4.1	3.2	128	117	5.9	5.4	101	83	4.7	3.8	273	166	12.7	
Midtown	E	-84.380749	33.782563	15919	1189	1094	487	6.9	3	359	197	2.3	1.2	638	266	4	1.6	855	559	5.4	3.5	899	287	5.6	
Piedmont Heights	F	-84.371323	33.807427	1813	265	75	74	4.2	4	0	15	0	0.8	41	47	2.2	2.6	43	43	2.4	2.4	202	138	11.1	
Lindridge/Martin Manor	F	-84.355384	33.816975	2179	286	86	79	3.9	3.6	134	160	6.2	7.3	242	136	11.1	6	212	138	9.7	6.2	202	148	9.3	
Virginia Highland	F	-84.357763	33.781075	4649	532	206	124	4.4	2.6	123	115	2.6	2.5	367	256	7.9	5.4	118	101	2.5	2.2	242	105	5.2	
Morningside/Lenox Park	F	-84.356123	33.799058	4748	515	62	64	1.3	1.3	216	217	4.5	4.5	286	316	6	6.6	178	137	3.7	2.9	102	86	2.1	
Emory	F	-84.322863	33.798357	1205	214	264	97	21.9	7	52	41	4.3	3.3	27	28	2.3	2.2	96	81	7.9	6.6	186	120	15.4	
Atlanta Industrial Park	G	-84.487474	33.795894	998	174	77	95	7.7	9.4	176	104	17.6	10	119	89	11.9	8.7	180	106	18	10.2	174	167	17.4	1

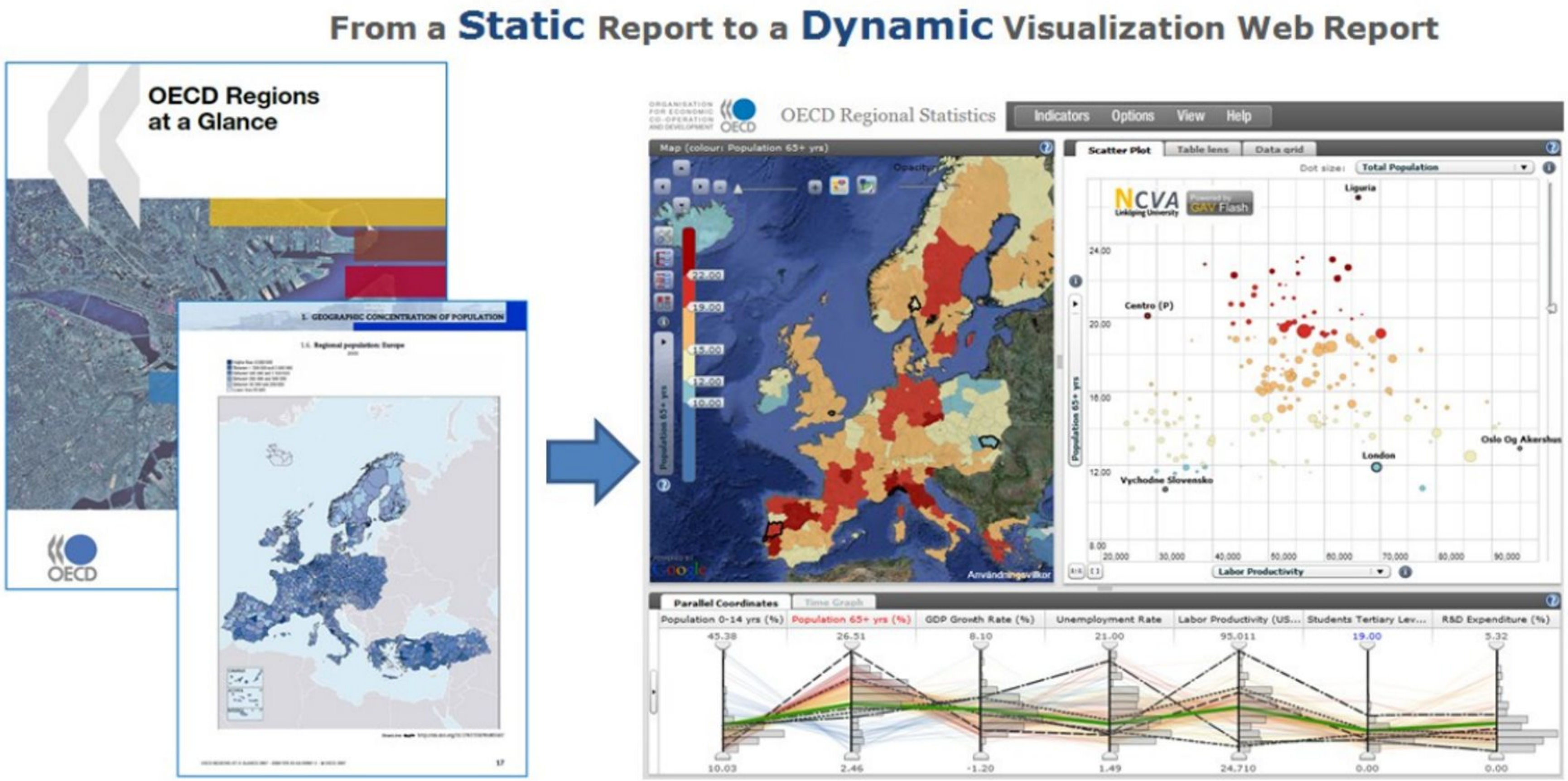
Feature reduction methods are great!

These include PCA, Unsupervised Classification, Self Organizing Maps, Indexes, etc.

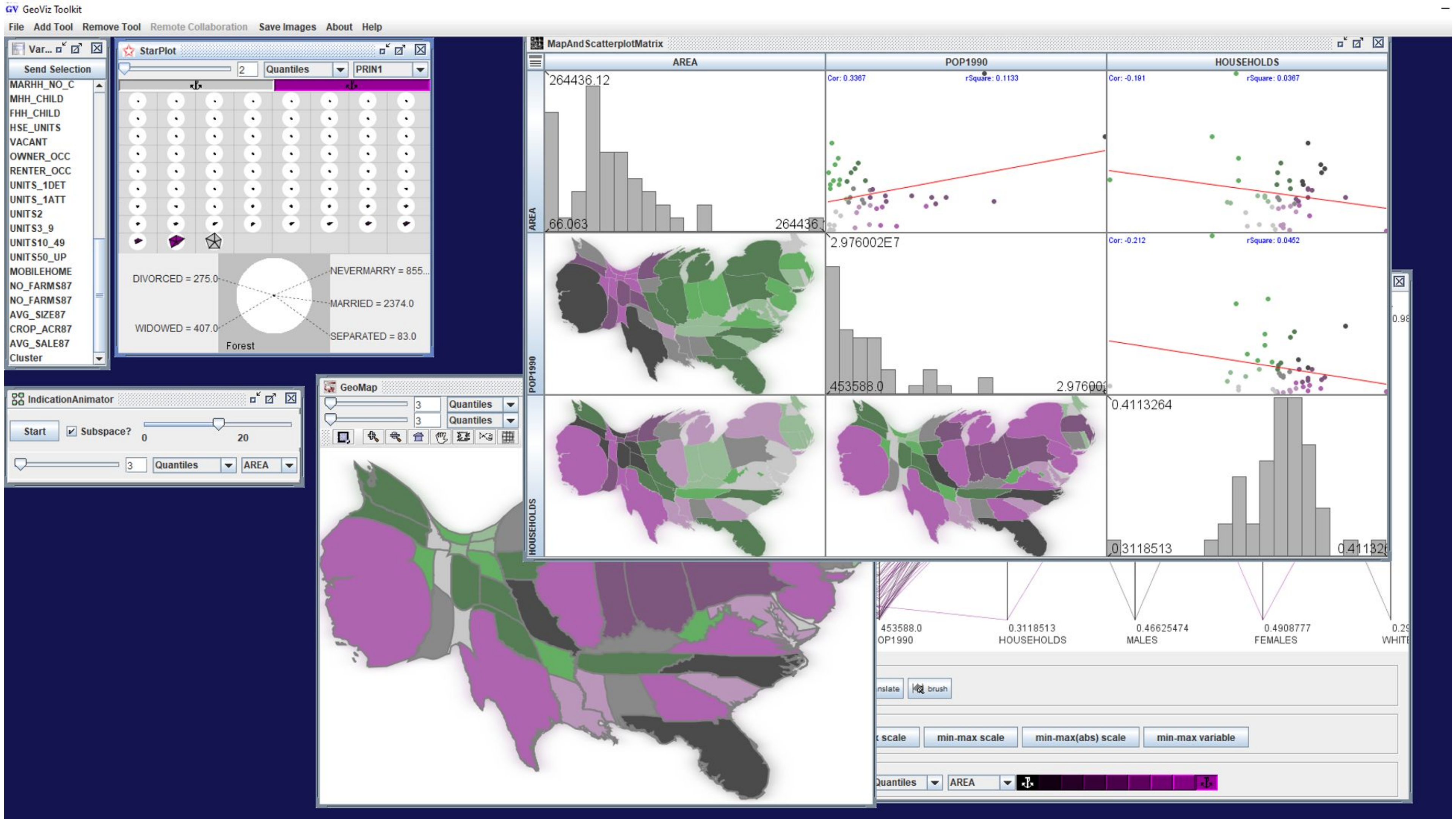


https://ldocao.wordpress.com/2016/07/04/self-organising-maps/

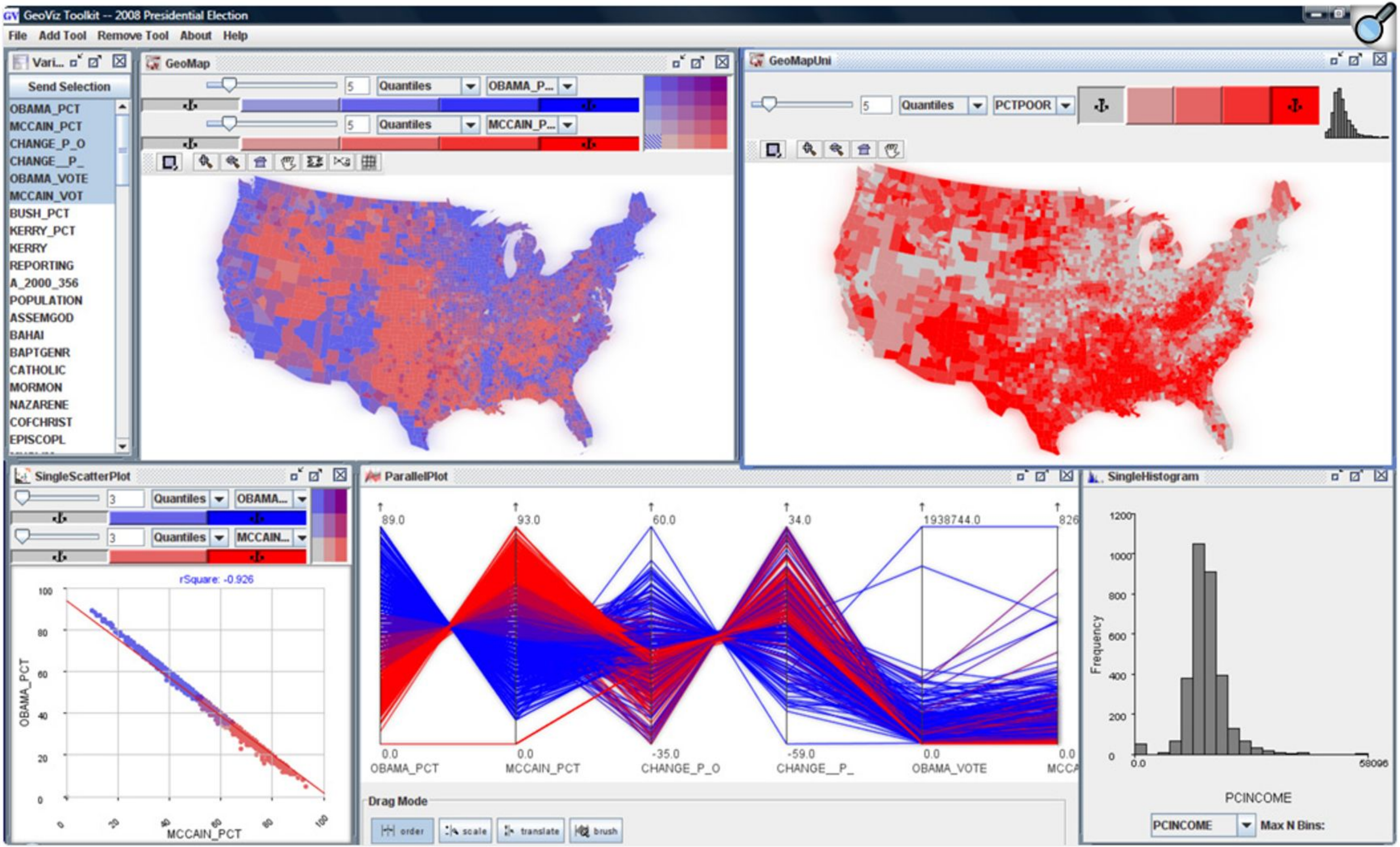
Image 1: A dynamic OECD Visualisation Web Report



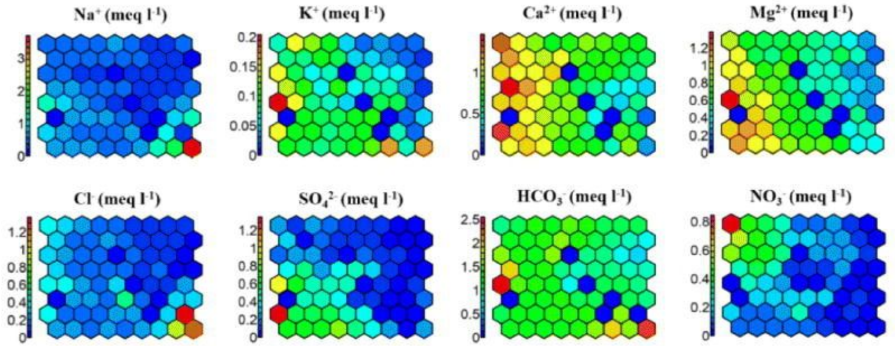
Organization for the Economic Cooperation and Development, Paris, France. Comparing a region's competitiveness in the global arena requires sound statistics and data, but such information is often limited and complex to be visualized. [PAPER]



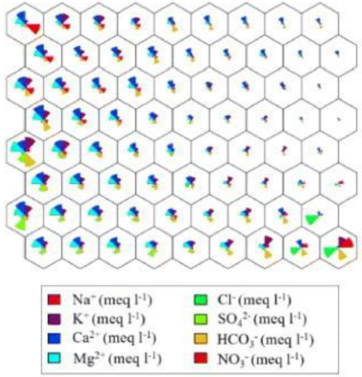
GeoViz Toolkit
File
Add Tool
Remove Tool
Remote Collaboration
Save Images
About
Help
Send Selection
MARHH_NO_C
MHH_CHILD
FHH_CHILD
HSE_UNITS
VACANT
OWNER_OCC
RENTER_OCC
UNITS_1DET
UNITS_1ATT
UNITS2
UNITS3_9
UNITS10_49
UNITS50_UP
MOBILEHOME
NO_FARMS87
NO_FARMS87
AVG_SIZE87
CROP_ACR87
AVG_SALE87
Cluster
StarPlot
Quantiles
PRIN1
DIVORCED = 275.0
NEVERMARRY = 855...
MARRIED = 2374.0
WIDOWED = 407.0
SEPARATED = 83.0
Forest
IndicationAnimator
Start
Subspace?
0
20
Quantiles
AREA
GeoMap
Quantiles
Quantiles
MapAndScatterplotMatrix
AREA
POP1990
HOUSEHOLDS
264436.12
66.063
Cor: 0.3367
rSquare: 0.1133
Cor: -0.191
rSquare: 0.0367
2.976002E7
453588.0
Cor: -0.212
rSquare: 0.0452
0.4113264
0.3118513
453588.0
0.3118513
HOUSEHOLDS
0.46625474
MALES
0.4908777
FEMALES
brush
min-max scale
min-max(abs) scale
min-max variable
AREA


GeoViz Toolkit -- 2008 Presidential Election
File
Add Tool
Remove Tool
About
Help
Vari...
Send Selection
OBAMA_PCT
MCCAIN_PCT
CHANGE_P_O
CHANGE__P_
OBAMA_VOTE
MCCAIN_VOT
BUSH_PCT
KERRY_PCT
KERRY
REPORTING
A_2000_356
POPULATION
ASSEMGOD
BAHAI
BAPTGENR
CATHOLIC
MORMON
NAZARENE
COFCHRIST
EPISCOPL
GeoMap
Quantiles
OBAMA_P...
MCCAIN_P...
GeoMapUni
PCTPOOR
SingleScatterPlot
OBAMA...
MCCAIN...
rSquare: -0.926
OBAMA_PCT
MCCAIN_PCT
ParallelPlot
89.0
93.0
60.0
34.0
1938744.0
0.0
-35.0
-59.0
CHANGE_P_O
CHANGE__P_
OBAMA_VOTE
Drag Mode
order
scale
translate
brush
SingleHistogram
Frequency
PCINCOME
Max N Bins:

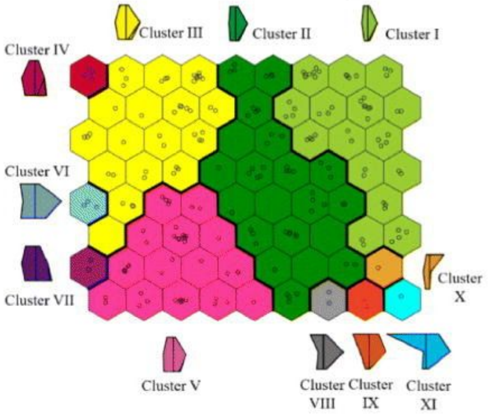
SOM component planes

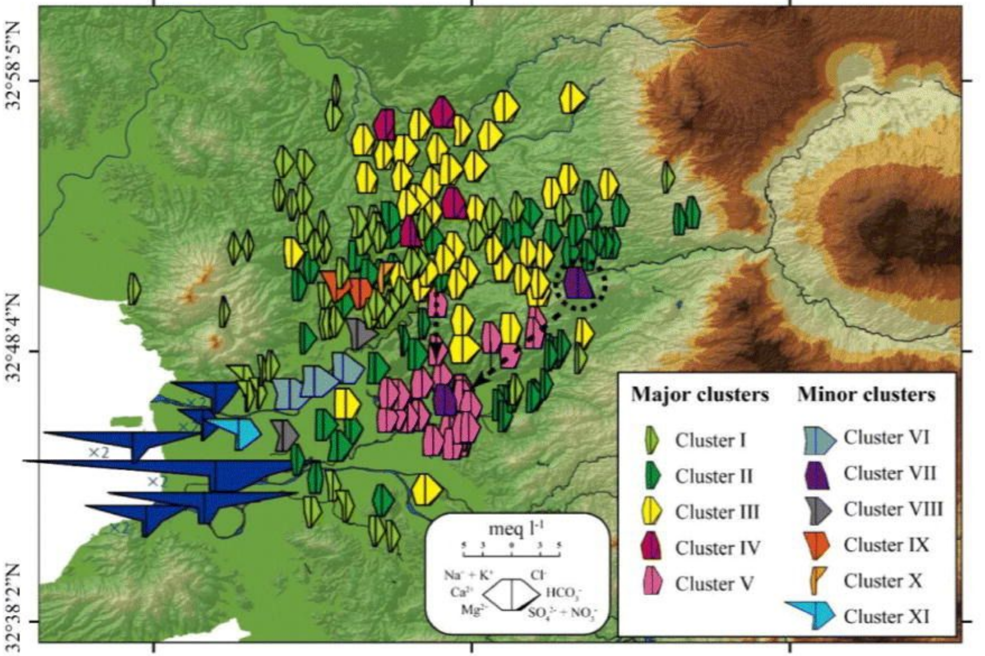


SOM node



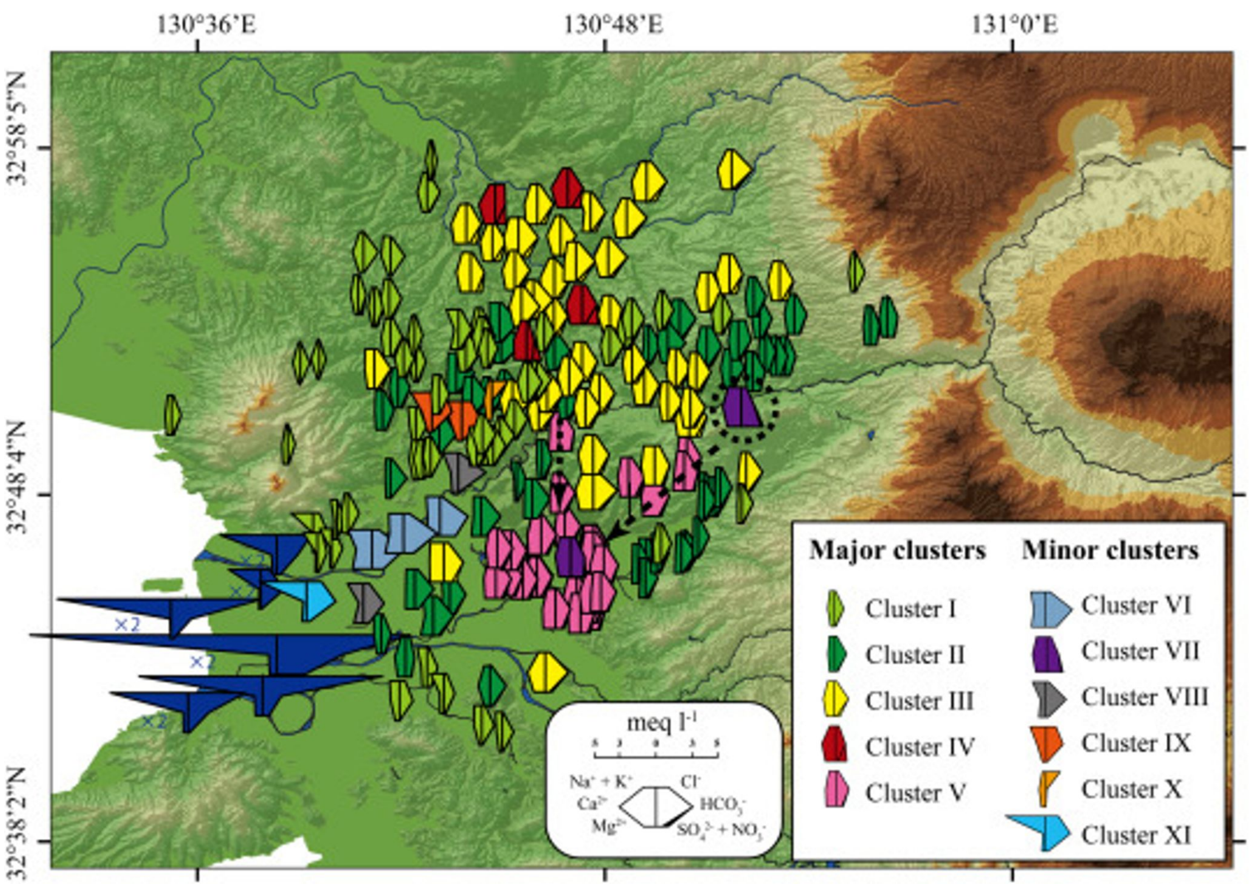
Clustering





SOM categorized groundwater chemistry type into 11 clusters based on major ion concentrations that reasonably explain the mixing of waters from different sources and comprehensive hydrochemical processes of aquifer systems in Kumamoto on reginal scale.

Rahman, A. S., Kono, Y., & Hosono, T. (2022). Self-organizing map improves understanding on the hydrochemical processes in aquifer systems. *Science of the Total Environment, 846*, 157281. https://www.sciencedirect.com/science/article/pii/S0048969722043790



Spatial distribution of clusters categorized by the SOM analysis presented by Stiff diagrams for the second aquifer systems in Kumamoto region. 6 samples in the coastal area shown by blue Stiff diagrams were not subjected to SOM analysis (see text for its reason) but categorized by visual inspection that are mostly

Again, not best for the average person.

Again, not best for the average user.

We want to help users:

-See all variables simultaneously, explore variables they didn't think to explore.

-Determine what is "**special**" about a place.

-Find and rank places by "**specialness**".

-Discover **how a place compares** to a distribution.

-Get interested in **new variables.**

Formal Design Goals

DG1: Capturing Entity 'Specialness'. For any individual entity in the dataset, the system should show how it ranks in terms of its percentile values to communicate the ways in which it is special or unique among the set of entities. For example, in a dataset of colleges, users should be able to discover which variables are most noteworthy for a specific focal school (e.g., "What's special about Michigan State University?"). This can support the inquiry needs of those at the college (the provost or development office), or those who are interested in that particular college (an alumnus or a neighbor).

DG2: Discovering and Exploring Variables. Users should be able to discover variables that they may not have considered exploring. Interesting variables may 'pop out' when an entity of interest is in a top of bottom percentile (e.g., 90th or 10th percentile). They can then pivot between entities and variables, as in a directed walk exercise (i.e., select an entity E1, then select a variable of interest V 1, then select an entity of interest E2 based on V 1's distribution, then select a new variable of interest V 2, etc.). Perhaps an alumnus noticed that their university had a high percentage of students receiving Pell Grants, and becomes newly interested in learning more about this variable.

DG3: Comparing Entities. The system should allow users to simultaneously contrast two or more entities by selecting multiple entities across dropdown menus, or "pinning" them by using the shift key. For example, a user may want to compare University of Michigan with University of Texas to see what their different strengths and weaknesses are in terms of how they rank across variables such as admissions rates, tuition costs, number of undergraduates, and graduation rates.

Formal Tasks

· **Identify Extrema:** for particular variables or variables where specific entities act as outliers. (DG1, DG2)
· **Sort**: Order entities by their value for a particular variable, or order variables by their percentiles for a particular entity. (DG1)
· **Compare:** Contrast selected entities by ”pinning” them to track their relative positions across variables. (DG3)
· **Rank**: List entities in the beeswarm plot from high (top) to low (bottom) based the percentile that they fall in for a given variable. In addition, list variables from high (top) to low (bottom) in a list, ranked by a focal entity’s specific percentile for each variable. (DG1, DG3)
· **Characterize Distribution:** View a distribution of values (e.g., normal, uniform, long tailed) to learn about value frequency. (DG1, DG2)
· **Contextualize:** Locate an entity within a distribution (e.g., finding high/lower neighbors). (DG1, DG2, DG3)
· **Find Similar:** Identify entities and variables with generally aligned patterns. (DG3)

FORMAL DESIGN GOALS & TASKS

DG1: Capturing Entity 'Specialness'.

- Identify Extrema
- Sort, Rank
- Contextualize

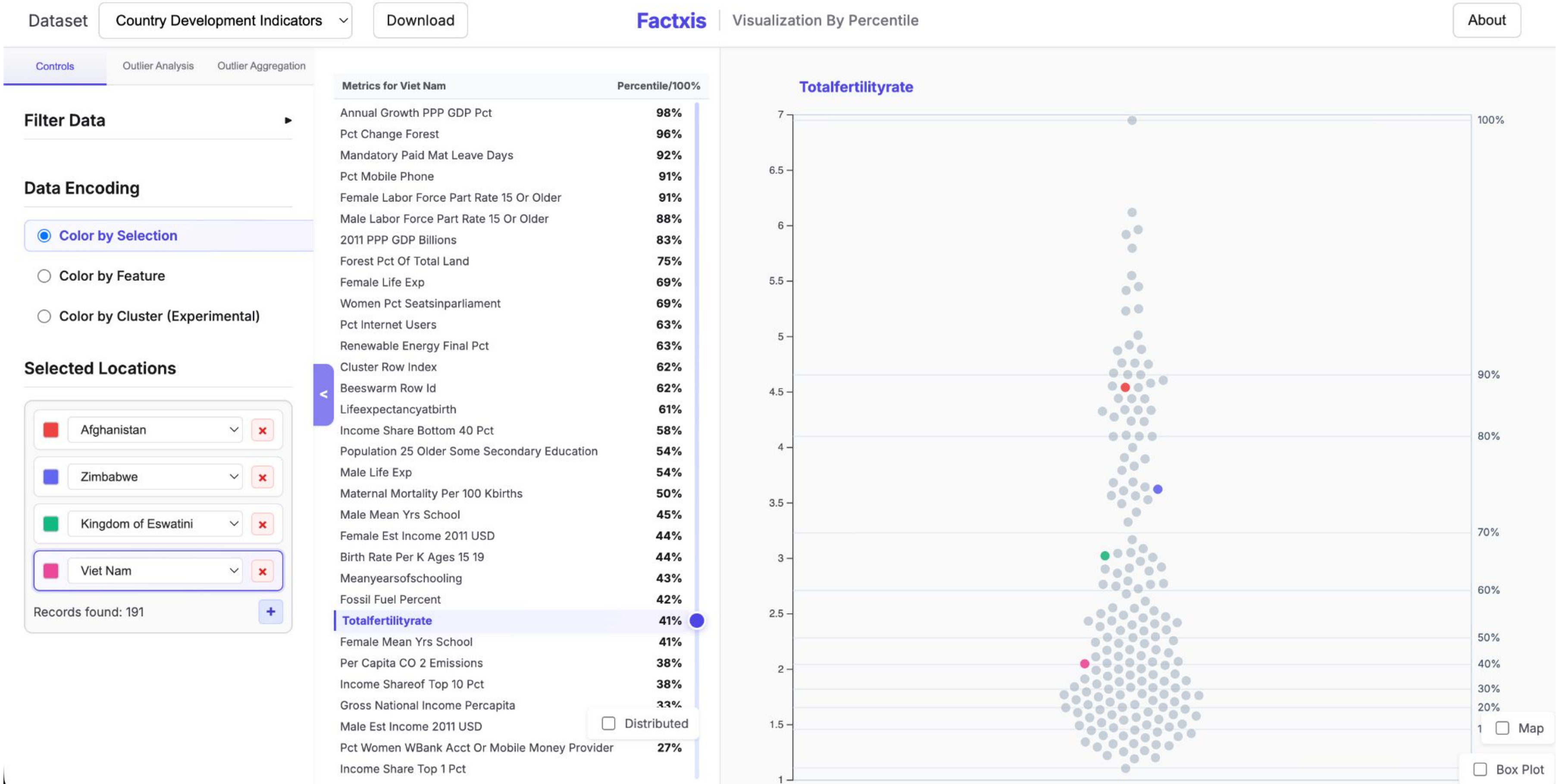
DG2: Discovering and Exploring Variables.

- Identify Extrema
- Characterize Distribution
- Contextualize

DG3: Comparing Entities.

- Rank, Compare
- Find Similar
- Contextualize

Our development so far...



Country Development Indicators - Percentile Scroll

Upload Excel File *No file selected* Japan

Japan

90%

Lifeexpectancyatbirth

Value: 84.69

Percentile: 90%

Pct Internet Users

Value: 91.8%

Percentile: 90%

Pct Mobile Phone

Value: 139.2%

Percentile: 90%

90%

90th percentile

Country Development Indicators - Percentile Scroll

Upload Excel File

No file selected Japan

Can we add:

- Map integration
- Direct comparison between regions

Japan

90%

Lifeexpectancyatbirth
Value: 84.69
Percentile: 90%

Pct Internet Users
Value: 91.8%
Percentile: 90%

Pct Mobile Phone
Value: 139.2%
Percentile: 90%

90%
90th percentile

Specifications

- Data is only shown on ONE axis
- We want to see **distributions** and **clusters** of percentiles for a given region, while still **contextualizing** how that region fares against peer regions
- Consider two objectives:
  - **Comparing** individual, selected regions to one another
  - See an overview of the data and **delve** into specific data points

Metrics for United States	Percentile/100%
Meanyearsofschooling	99%
2011 PPP GDP Billions	99%
Female Mean Yrs School	99%
Male Mean Yrs School	99%
Gross National Income Percapita	95%
Female Est Income 2011 USD	95%
Male Est Income 2011 USD	94%
Per Capita CO 2 Emissions	94%
Pct Internet Users	86%
Population 25 Older Some Secondary Education	85%
Income Share Top 1 Pct	85%
Cluster Row Index	85%
Beeswarm Row Id	85%
Pct Women WBank Acct Or Mobile Money Provider	84%
Lifeexpectancyatbirth	81%
Female Life Exp	80%
Male Life Exp	78%
Pct Mobile Phone	69%
Pct Change Forest	65%
Fossil Fuel Percent	64%
Women Pct Seatsinparliament	62%
Income Shareof Top 10 Pct	60%
Annual Growth PPP GDP Pct	58%
Female Labor Force Part Rate 15 Or Older	58%
Forest Pct Of Total Land	56%
Birth Rate Per K Ages 15 19	33%
Male Labor Force Part Rate 15 Or Older	32%
Totalfertilityrate	29%
Renewable Energy Final Pct	27%
Maternal Mortality Per 100 Kbirths	
Income Share Bottom 40 Pct	23%
Mandatory Paid Mat Leave Days	

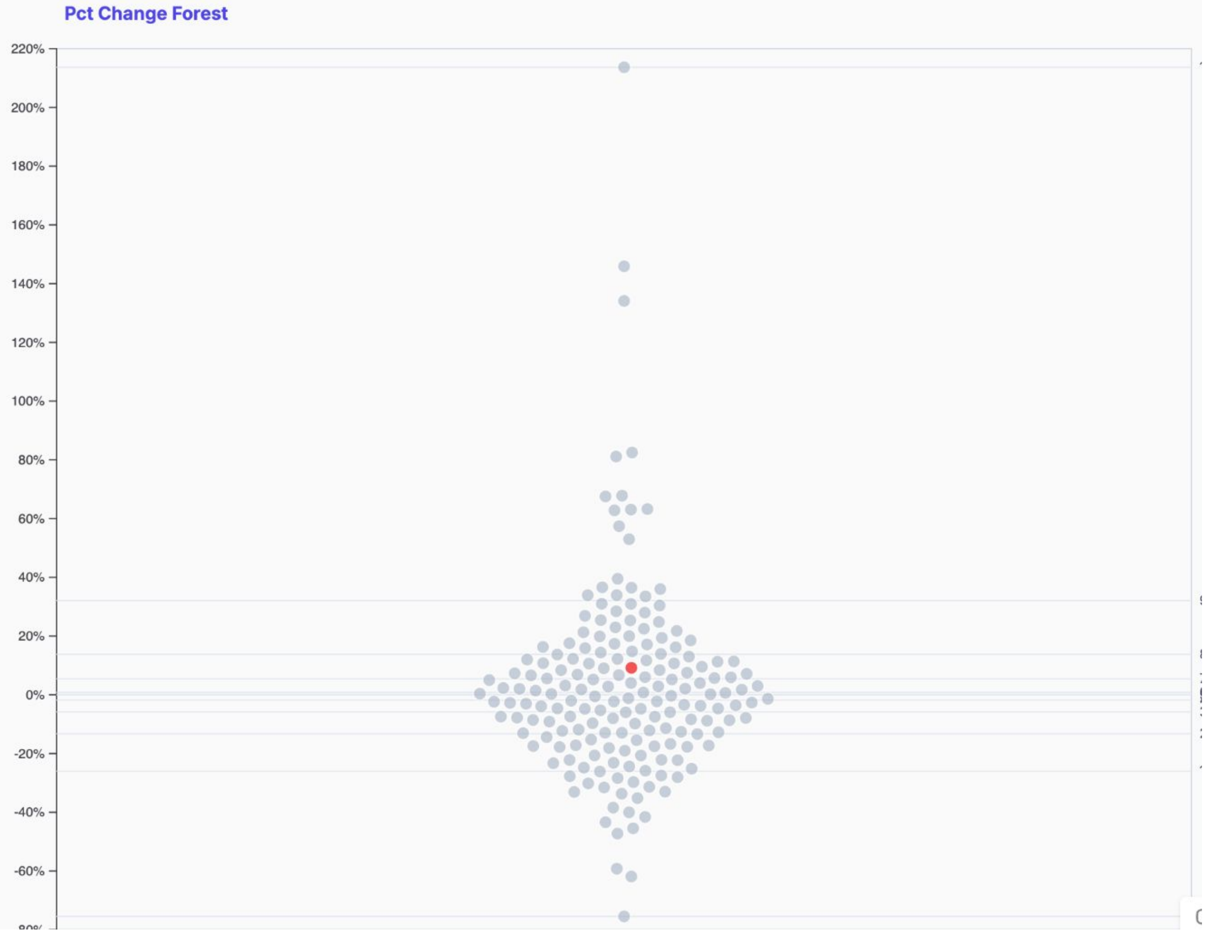
Distributed

Metrics for United States	Percentile/100%
Meanyearsofschooling	**99%**
2011 PPP GDP Billions	**99%**
Female Mean Yrs School	**99%**
Male Mean Yrs School	**99%**
Gross National Income Percapita	**95%**
Female Est Income 2011 USD	**95%**
Male Est Income 2011 USD	**94%**
Per Capita CO 2 Emissions	**94%**
Pct Internet Users	**86%**
Population 25 Older Some Secondary Education	**85%**
Income Share Top 1 Pct	**85%**
Cluster Row Index	**85%**
Beeswarm Row Id	**85%**
Pct Women WBank Acct Or Mobile Money Provider	**84%**
Lifeexpectancyatbirth	**81%**
Female Life Exp	**80%**
Male Life Exp	**78%**
Pct Mobile Phone	**69%**
Pct Change Forest	**65%**
Fossil Fuel Percent	**64%**
Women Pct Seatsinparliament	**62%**
Income Shareof Top 10 Pct	**60%**
Annual Growth PPP GDP Pct	**58%**
Female Labor Force Part Rate 15 Or Older	**58%**
Forest Pct Of Total Land	**56%**
Birth Rate Per K Ages 15 19	**33%**
Male Labor Force Part Rate 15 Or Older	**32%**
Totalfertilityrate	**29%**
Renewable Energy Final Pct	**27%**
Maternal Mortality Per 100 Kbirths	
Income Share Bottom 40 Pct	**23%**
Mandatory Paid Mat Leave Days	

Distributed

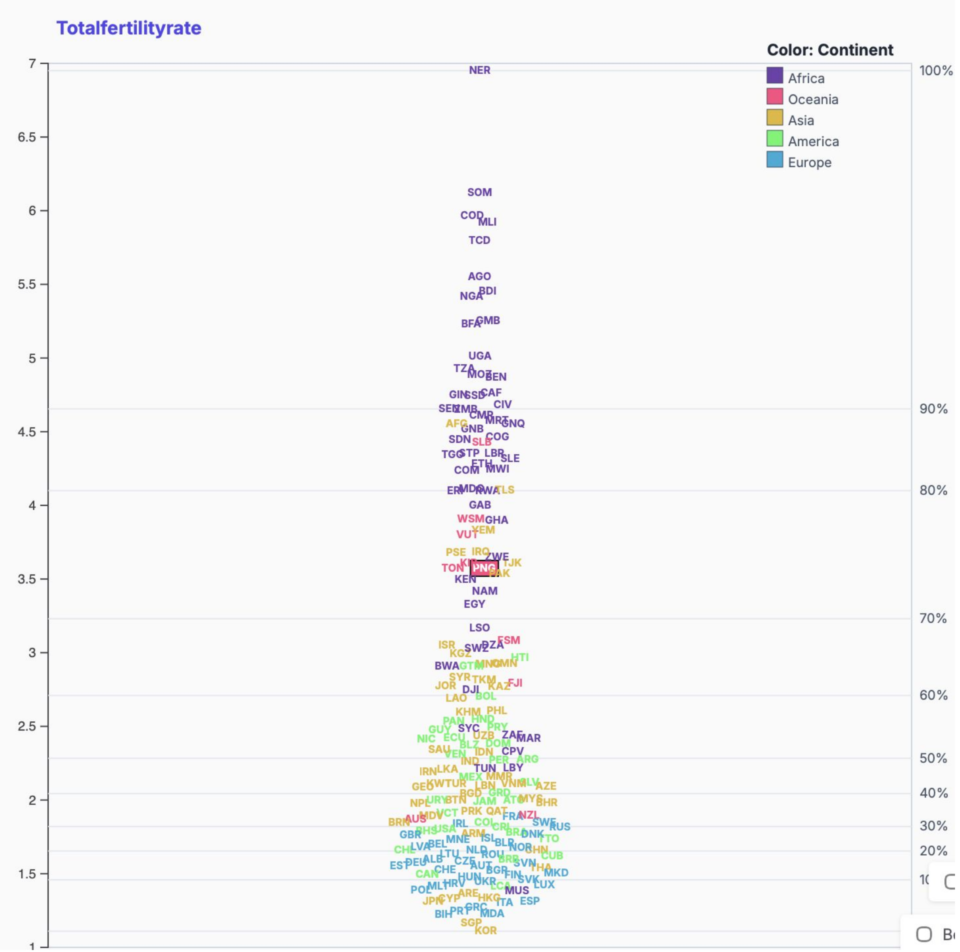
Metrics for United States	Percentile/100%
Meanyearsofschooling	99%
2011 PPP GDP Billions	99%
Female Mean Yrs School	99%
Male Mean Yrs School	99%
Gross National Income Percapita	95%
Female Est Income 2011 USD	95%
Male Est Income 2011 USD	94%
Per Capita CO 2 Emissions	94%
Pct Internet Users	86%
Population 25 Older Some Secondary Education	85%
Income Share Top 1 Pct	85%
Cluster Row Index	85%
Beeswarm Row Id	85%
Pct Women WBank Acct Or Mobile Money Provider	84%
Lifeexpectancyatbirth	81%
Female Life Exp	80%
Male Life Exp	78%
Pct Mobile Phone	69%
Pct Change Forest	**65%**
Fossil Fuel Percent	64%
Women Pct Seatsinparliament	62%
Income Shareof Top 10 Pct	60%
Annual Growth PPP GDP Pct	58%
Female Labor Force Part Rate 15 Or Older	58%
Forest Pct Of Total Land	56%
Birth Rate Per K Ages 15 19	33%
Male Labor Force Part Rate 15 Or Older	32%
Totalfertilityrate	29%
Renewable Energy Final Pct	27%
Maternal Mortality Per 100 Kbirths	
Income Share Bottom 40 Pct	23%
Mandatory Paid Mat Leave Days	

Distributed



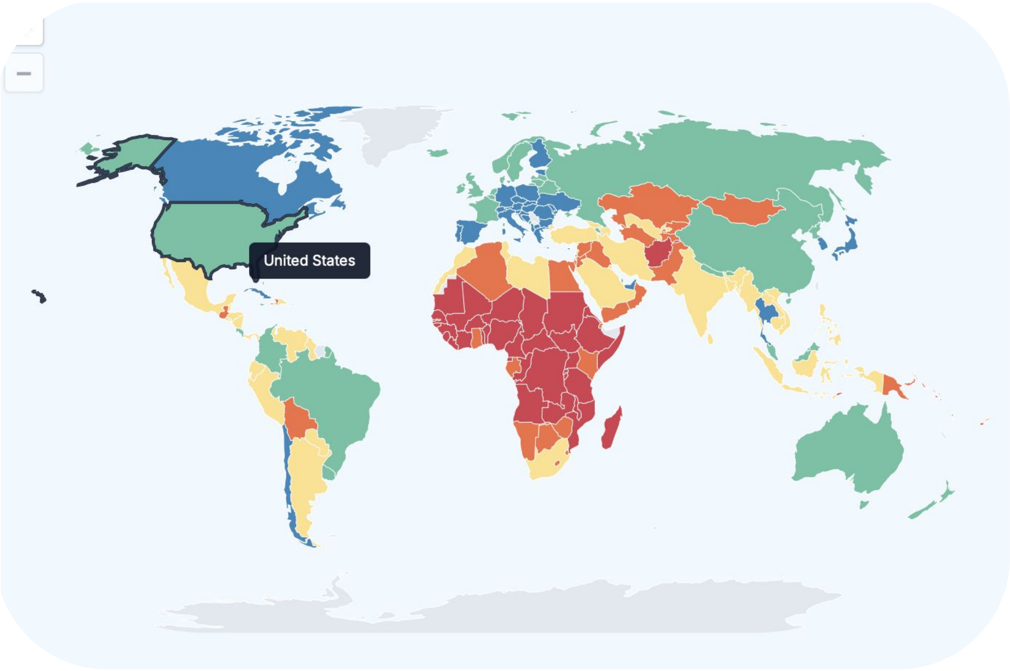
Demo

- Map View
- Data Encoding
  - Multi-Select vs Feature Encoding
- Filtering
- Distributed View
- Outlier Analysis



Feature Timeline

1. Visualization by Percentile
2. Metrics List & Beeswarm
3. Distributed View
4. Multi-Select
5. Feature Encoding
6. Map Integration
7. Text Encoding
8. Filtering
9. Identifying Similar Regions
10. Outlier Analysis
11. **User Study**



Our User Test (Part 1): April 2026

In person: demo and paper/pencil.

On campus / weekends.

34 Students / Faculty / Staff.
(Avg. Age = 23.4)
(11 F, 22 M, 2 NB)
12 Undergrads.

Dataset about U.S. colleges.

Brief interview questions.

Users got $15 cash.
Participants took 30-60 mins.
IRB approved.

ID Number: ________ Date: ________________ Staff Initials: ________

FACTXIS USER STUDY QUESTIONS

Directions: Please circle any questions you find to be difficult to answer.

PART 1

(1) LEARNING ABOUT A COLLEGE

1. Choose the following: **[Georgia Tech]** in the dropdown menu.

a. For which of the different variables does Georgia Tech have the highest percentile? ________________ What is the *percentile*? ________________

2. Display the following variable: **[Average Cost]**

a. At what *percentile* is **[Georgia Tech]** at for this variable? ________________

b. What is the **[Average Cost]** of Georgia Tech? ________________

c. What are some colleges with an average cost within $1,000 of GT's cost?

________________ ________________ ________________

(2) FINDING OUTLIERS

A college is considered an outlier if it is in the top or bottom percentile (above 95th or below 5th) for 3 or more variables.

3. What are some outlier colleges you see in the data?

________________ ________________ ________________

(3) RANKINGS / DISTRIBUTIONS

If you can't find it, that's ok! Leave it blank. Ask if you want a definition of any of these terms.

4. Can you find a variable that has a:

a. Uniform distribution? (evenly spread values) ________________

b. Normal distribution? (a bell curve) ________________

c. Bimodal distribution? (two groups) ________________

d. A top *or* bottom heavy distribution? ________________

Est. cumulative time 10-15 minutes.

(4) MULTI-CRITERIA QUERIES

5. Among colleges whose average ACT is 30 or above, which has the HIGHEST admissions rate (it's the least selective)? ________________ *(hint: you can use the filter.)*

(5) COMPARISON

6. Compare **[Georgia Tech]** with **[Berklee College of Music]**. ***This is a small music college in Boston where John Mayer went to school.*** What variables are similar and what are different?

a. Name a variable where the schools have a similar value: ________________

b. Name a variable where the schools differ: ________________

(6) SCENARIOS

7. Your younger sister is applying to college, and she wants a school that is "diverse" and low cost. She thinks high selectivity (low acceptance rate) is also important. What are a few schools you would recommend for her?

________________ ________________

________________ ________________

8. You are the president at **[Case Western Reserve]**. What are some ways you are doing "better" than other schools and which do you think your school could improve on?

Ways that you are doing "better" than other schools:

____

____

____

Things you can improve on:

____

____

____

Est cumulative time 25 minutes.

(7) CORRELATION & SIMILARITY

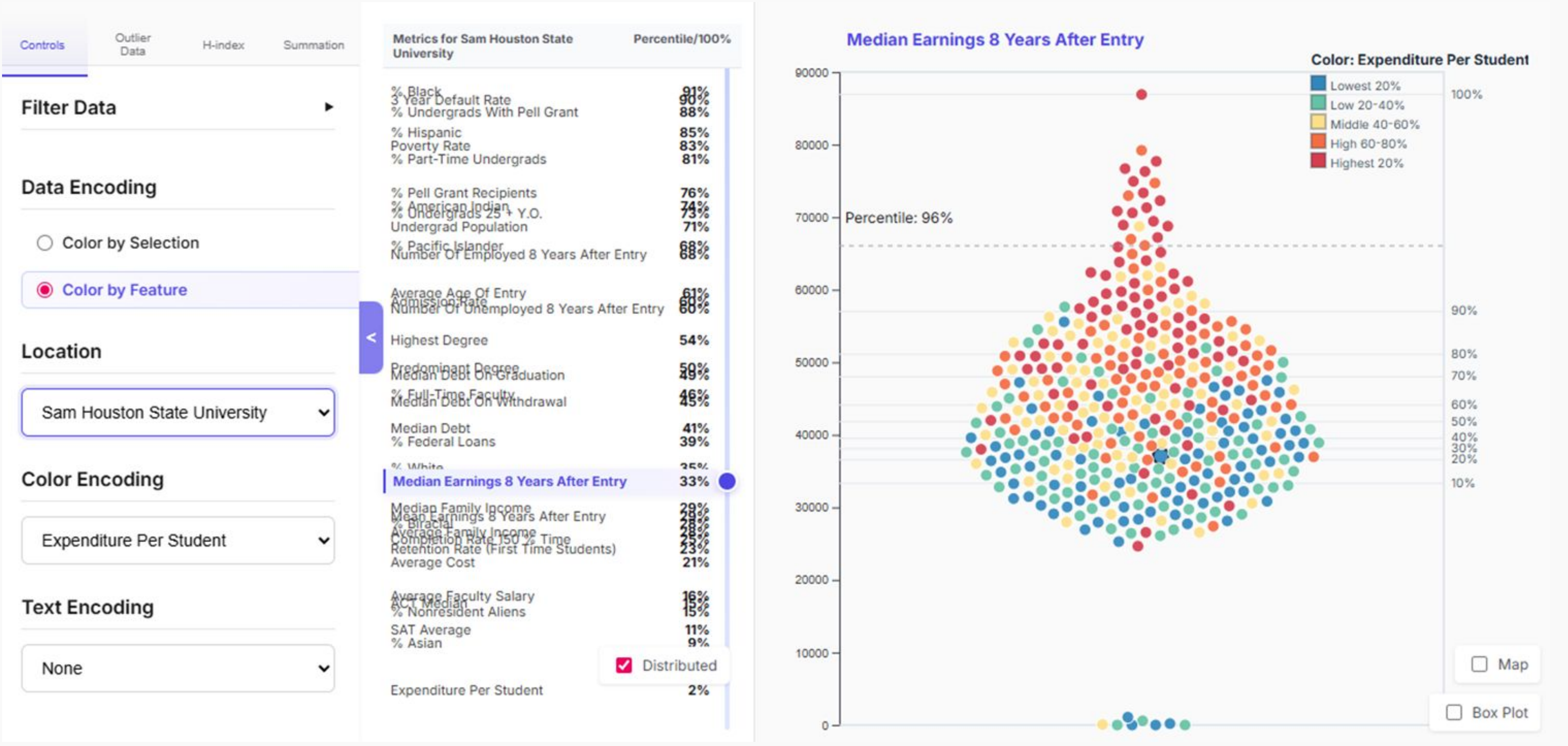
Max time: 6 minutes total (3 mins per question).

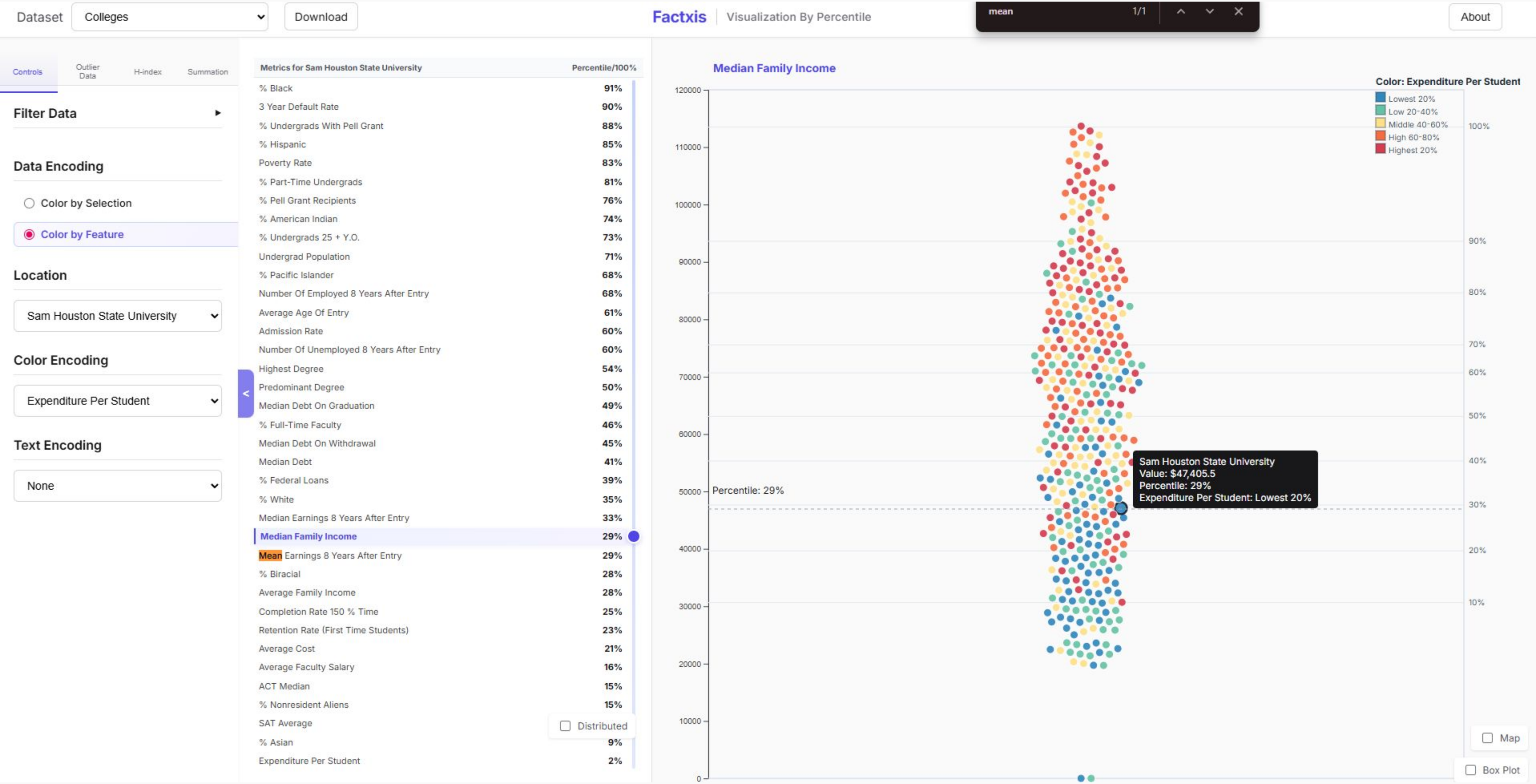
9. What **variable** correlates the most strongly in the dataset with the variable **[Expenditure Per Student]** (not based on common sense, but based on your visualization of the dataset.) *Hint: You can color each dot according to a variable of interest.*

____

10. Please find a college that is the most like (hast the most similar variables) as **[Grinnell College]**? Do not take longer than 3 minutes to find your answer. ________________

Colleges Data




Dataset
Colleges
Download
Factxis
Visualization By Percentile
mean
1/1
About
Controls
Outlier Data
H-index
Summation
Filter Data
Data Encoding
Color by Selection
Color by Feature
Location
Sam Houston State University
Color Encoding
Expenditure Per Student
Text Encoding
None
Metrics for Sam Houston State University
Percentile/100%
% Black 91%
3 Year Default Rate 90%
% Undergrads With Pell Grant 88%
% Hispanic 85%
Poverty Rate 83%
% Part-Time Undergrads 81%
% Pell Grant Recipients 76%
% American Indian 74%
% Undergrads 25 + Y.O. 73%
Undergrad Population 71%
% Pacific Islander 68%
Number Of Employed 8 Years After Entry 68%
Average Age Of Entry 61%
Admission Rate 60%
Number Of Unemployed 8 Years After Entry 60%
Highest Degree 54%
Predominant Degree 50%
Median Debt On Graduation 49%
% Full-Time Faculty 46%
Median Debt On Withdrawal 45%
Median Debt 41%
% Federal Loans 39%
% White 35%
Median Earnings 8 Years After Entry 33%
Median Family Income 29%
Mean Earnings 8 Years After Entry 29%
% Biracial 28%
Average Family Income 28%
Completion Rate 150 % Time 25%
Retention Rate (First Time Students) 23%
Average Cost 21%
Average Faculty Salary 16%
ACT Median 15%
% Nonresident Aliens 15%
SAT Average
% Asian 9%
Expenditure Per Student 2%
Distributed
Median Family Income
Color: Expenditure Per Student
Lowest 20%
Low 20-40%
Middle 40-60%
High 60-80%
Highest 20%
120000
110000
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
Percentile: 29%
Sam Houston State University
Value: $47,405.5
Percentile: 29%
Expenditure Per Student: Lowest 20%
Map
Box Plot

RESULTS

Percentage correct answers by task.

Scenarios were tough!

Finding exact values was easy

Task Type	Pct. Correct	Details
Find Extremum	100%	
Retrieve Value	97.2*%	
Retrieve Value	97.2*%	
Retrieve Value	100%	
Query with Range (Name 3)	90.2%	
Detect Outliers (Name 3)	87.3%	(G:3 VG:17 E:62)
Describe Distribution (Uniform)	85.3%	
Describe Distribution (Normal)	70.6%	
Describe Distribution (Bimodal)	94.1%	
Describe Distribution (Skewed)	85.3%	
Query with Filter	88.2%	
Compare	94.1%	
Contrast	91.2%	
Scenario (Success Areas)	80.7%	
Scenario (Improvement Areas)	72.5%	
Correlate	76.4%	(G:2 VG:6 E:18)
Find Similar (Advanced)	73.5%	(G:6 VG:3 E:16)

We scored correct results as G = good, VG = very good and E = excellent (see supplementary materials). *One respondent could not select dots briefly; their responses were 1 and 3 percentile values different from the correct answers and are marked incorrect. Color is used to group sets of questions.

RESULTS

Likert scale results from the System Usability Scale (SUS). The mean user response is provided, where **1 is low** and **5 is high**, and cells are colored using a diverging color ramp to help display variation.

May not use it often

Easy to use

SUS Question Keywords	Mean Value	St. Dev.
Support Needed for Use	1.38	0.73
Inconsistent	1.42	0.60
Cumbersome	1.52	0.86
Need for Pre-Training	1.72	0.91
Too Difficult	1.91	0.96
Would use Frequently	3.51	0.94
Confident with Use	3.68	0.87
Can Learn Quickly	3.75	0.93
Easy to Use	3.88	0.68
Integrated Components	3.93	0.71

What would they have used instead?

1 Microsoft Excel (n=15)
Google Sheets (n=2)

2 R and RStudio (n=9) Python (n=2)

3 Tableau (n=4)

4 LLM (e.g., ChatGPT) (n=2)

5 ArcGIS, CodeApp, Matlab, and Radiant were mentioned by one respondent each.

WHAT WOULD YOU USE THIS FOR?

Housing and infrastructure, including apartments and homes, buildings, and parks (n=6)

Cities, neighborhoods, countries, regions, weather, traffic, disasters, transportation, cost of living (n=18)

Companies, job types majors, and product testing (n=6)

Consumer and popular culture including sports, restaurants, movies, cars, hotels (n=8)

Future development areas

LLM Integration

- Give context: answer the “why?”

Use Cases

- Dots as sensors

Related Entity Network

- Find what makes entities “similar”

New User Study (Groups)

NEXT STEPS FOR TODAY

1. Q& A

2. Sandbox Session

3. Open Conversation about Needs, Use Cases, and Features

4. Closing