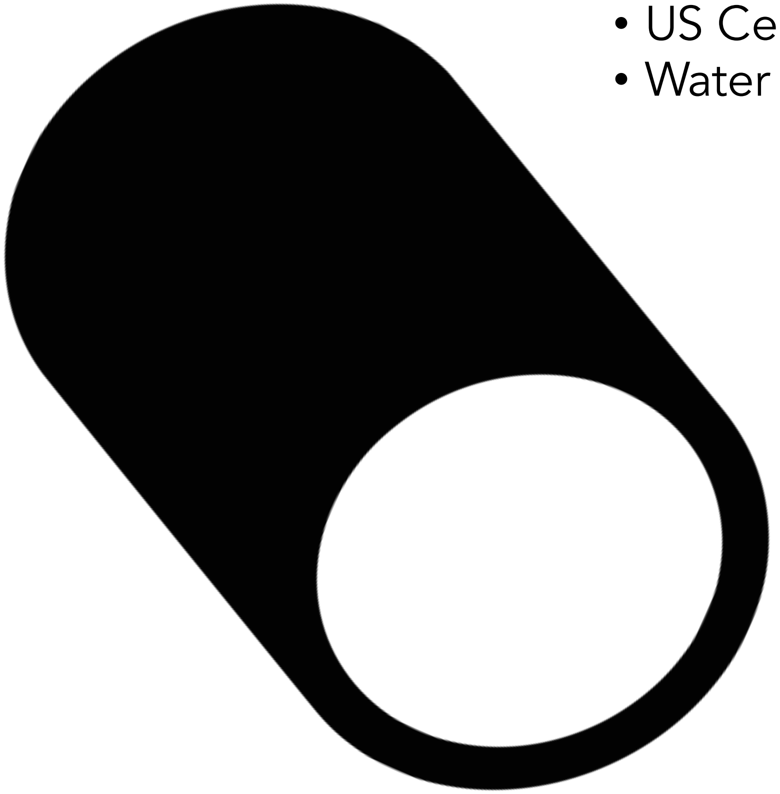
Visualization of Water Usage and Affordability Clusters

Peter Rasmussen
May 13, 2016

Summary

- **Motivation:** Water access, affordability, and quality have re-emerged as serious issues in the US and continue to be major problems in developing countries
- **Objective:** Visualize the extent of the this issue in the US using water price data, demographic information using cluster analyses
- **Analyses:**
  - **Univariate clustering** on water usage, price, and income to identify municipalities and and regions of where water may be less affordable
  - **Multivariate clustering** to see how other factors, including demographic variables, may correlate with income and water usage and price

Methodology: Data pipeline



Sources

- US Census American Consumer Survey (2014)
- Water utilities price data*

Wrangling

- Clean, normalize, and join geospatial and non-geospatial data

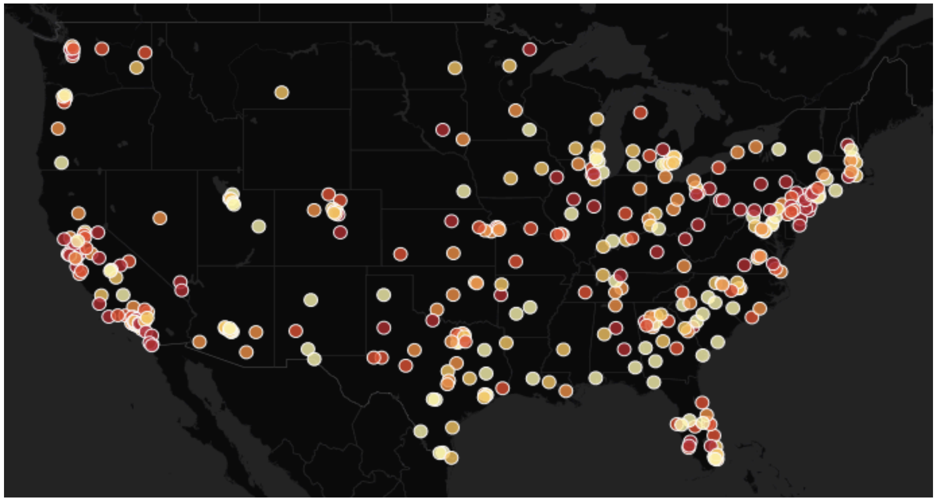
Analysis and visualization

- Univariate modeling and scoring
- Multivariate k means feature & k selection
- Worked with training data only
- Visualize univariate outputs on maps, multivariate outputs on scatterplot matrices

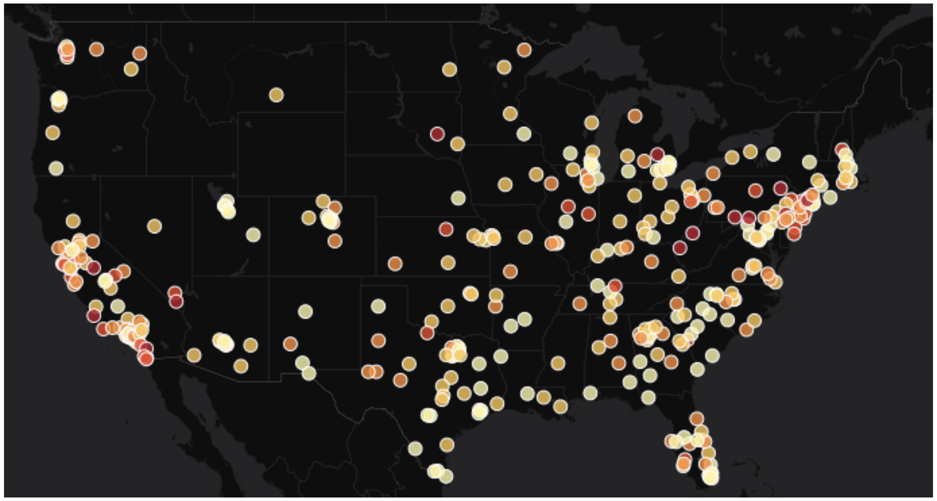
*Pipeline icon created by misirlou from Noun Project

Univariate clustering: Quintiles v. Jenks

Quintiles



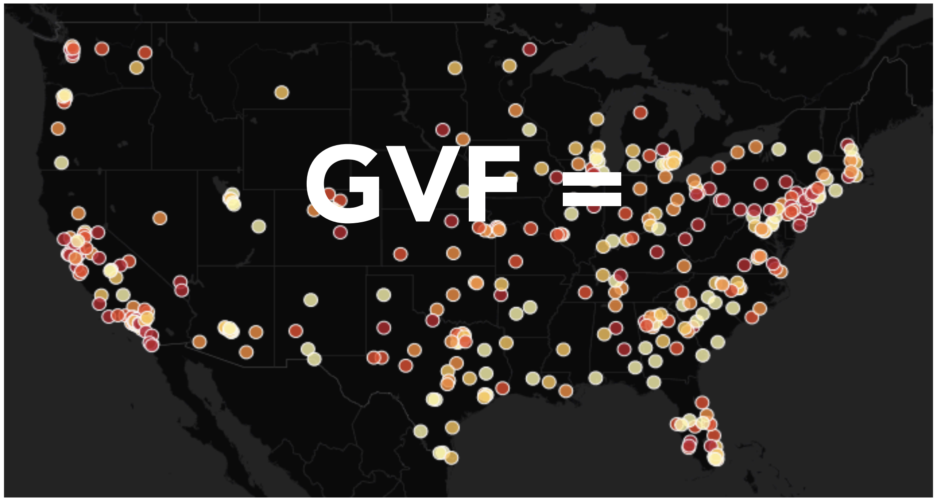
Jenks Breaks



Both look similar above, but goodness of variance fits (GVF) are different

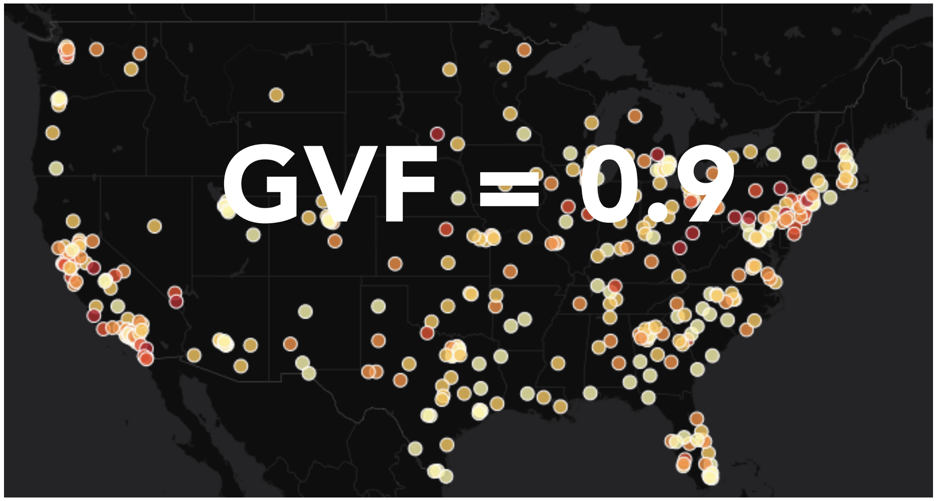
Univariate clustering: Quintiles v. Jenks

Quintiles



GVF =

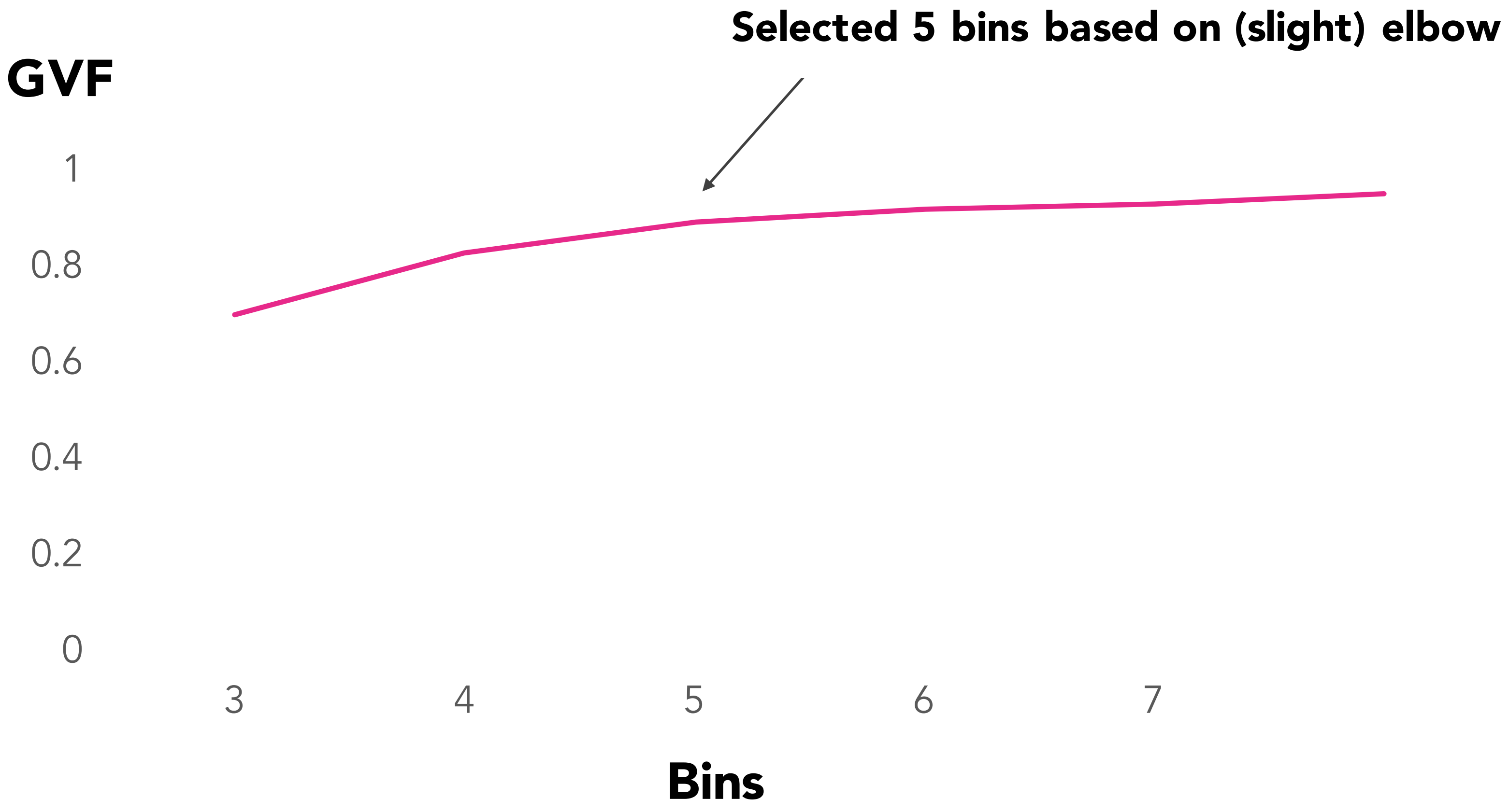
Jenks Breaks



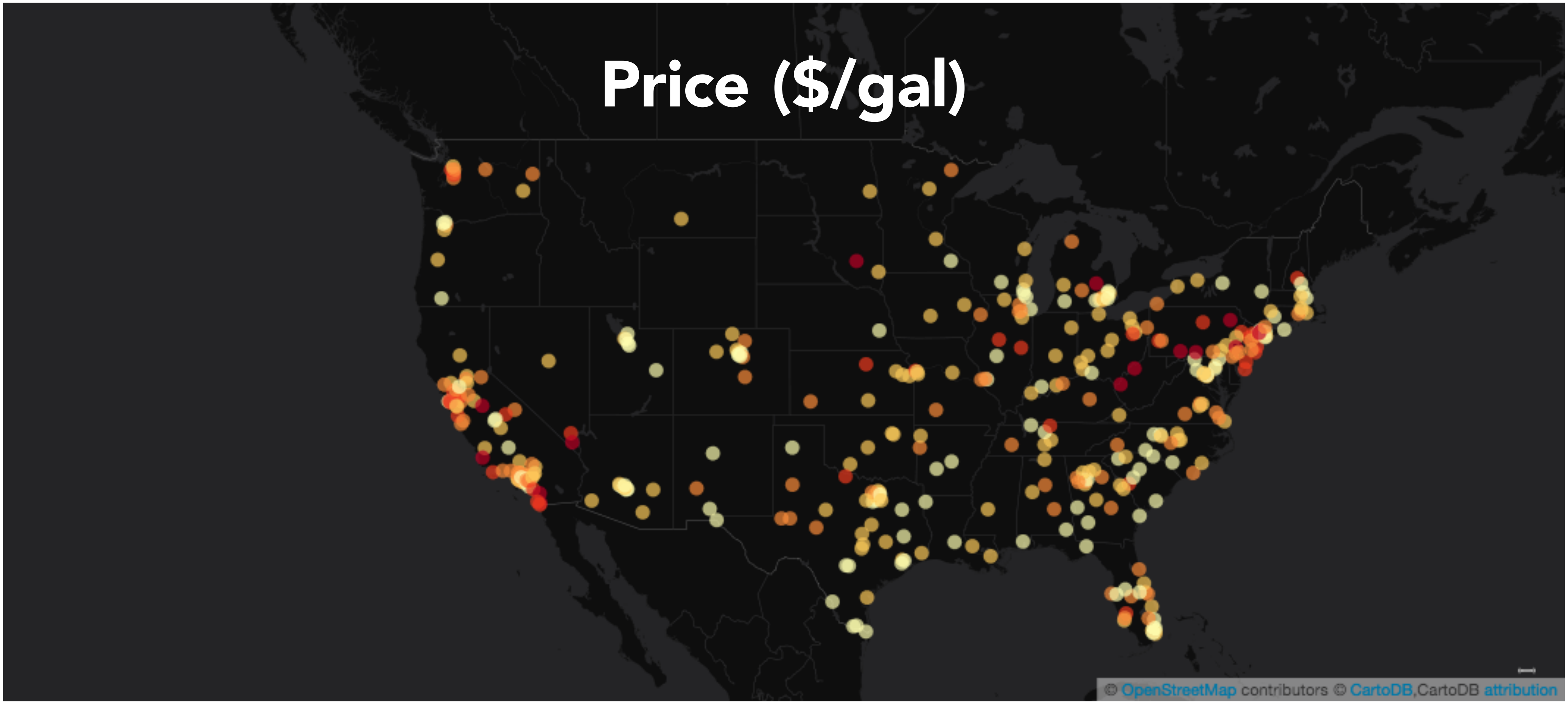
GVF = 0.9

Based on GVF, Jenks wins

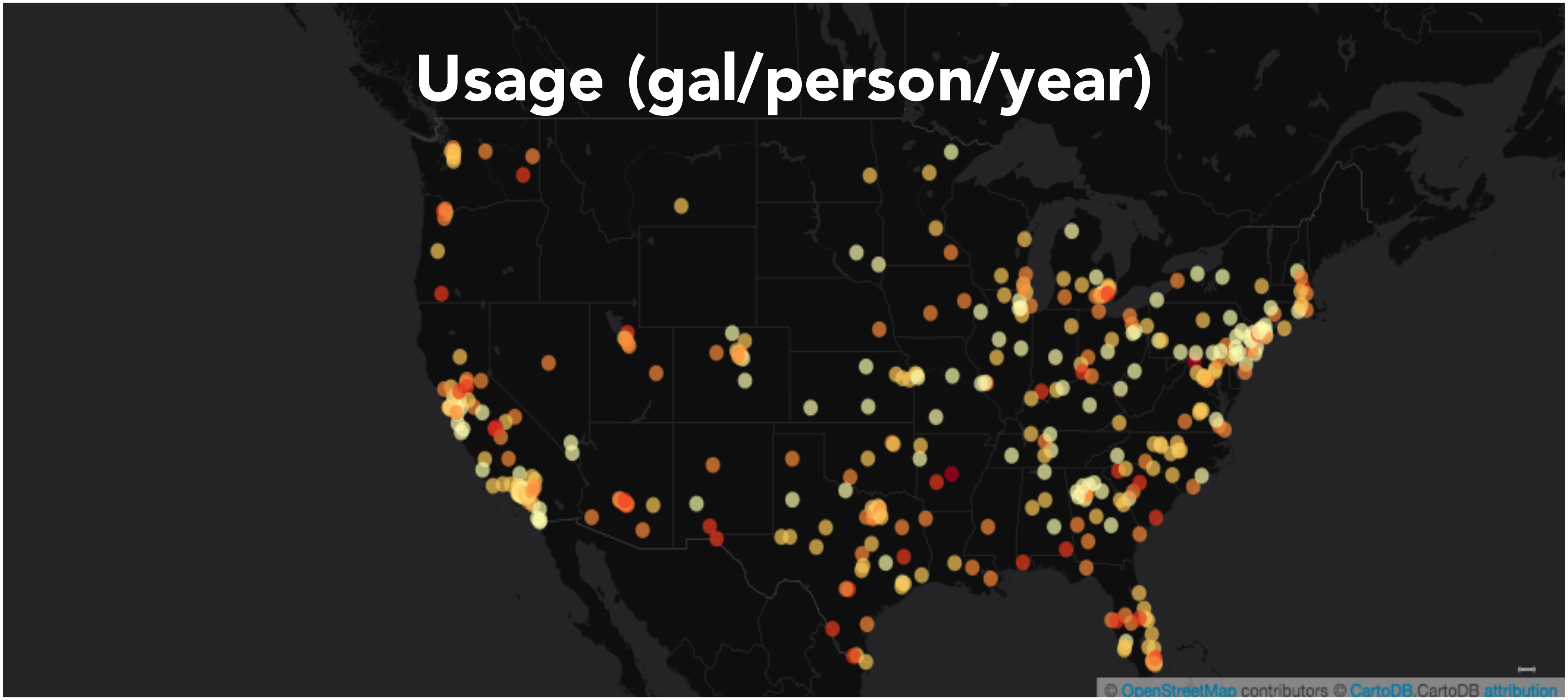
Univariate clustering: Goodness of variance fit (GVF)



Univariate clustering: Results



Univariate clustering: Results



Usage (gal/person/year)

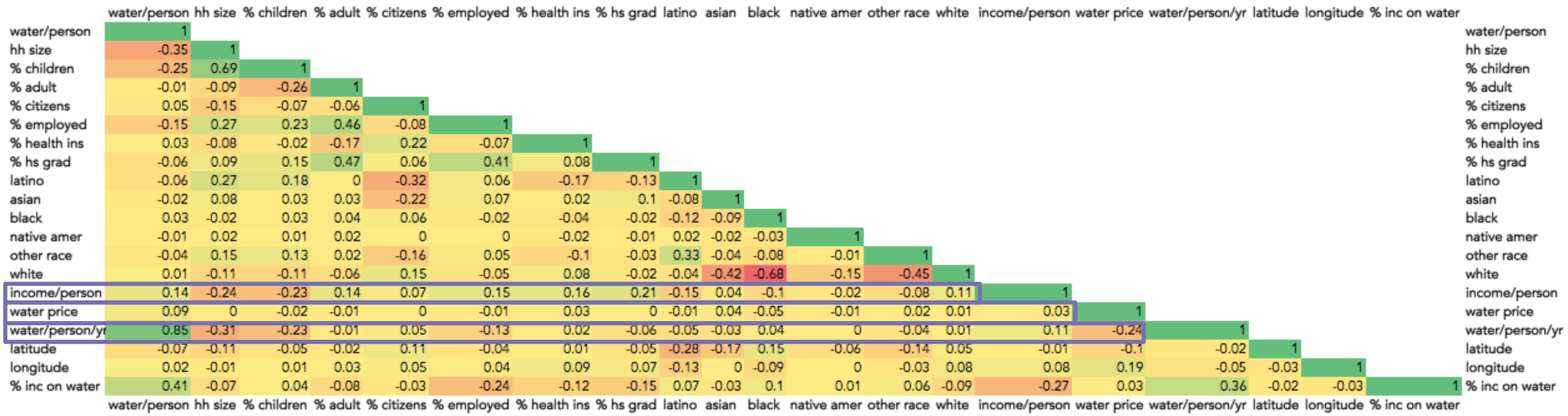
© OpenStreetMap contributors © CartoDB, CartoDB attribution

Multivariate clustering: k means

“k means clustering aims to partition n observations into k clusters in which each observation belongs to cluster with nearest mean” (Wikipedia)

Multivariate clustering: Insights from the correlation matrix

Feature set in a correlation matrix



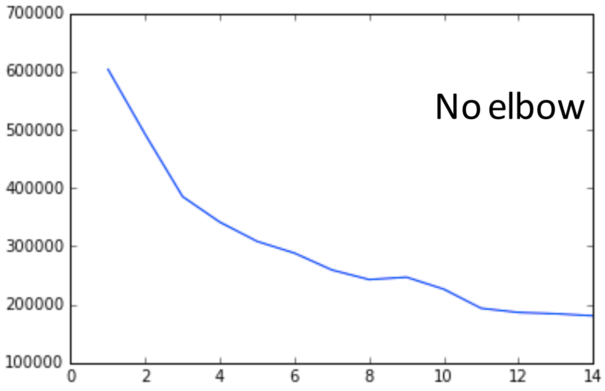
	water/person	hh size	% children	% adult	% citizens	% employed	% health ins	% hs grad	latino	asian	black	native amer	other race	white	income/person	water price	water/person/yr	latitude	longitude	% inc on water	
water/person	1																				water/person
hh size	-0.35	1																			hh size
% children	-0.25	0.69	1																		% children
% adult	-0.01	-0.09	-0.26	1																	% adult
% citizens	0.05	-0.15	-0.07	-0.06	1																% citizens
% employed	-0.15	0.27	0.23	0.46	-0.08	1															% employed
% health ins	0.03	-0.08	-0.02	-0.17	0.22	-0.07	1														% health ins
% hs grad	-0.06	0.09	0.15	0.47	0.06	0.41	0.08	1													% hs grad
latino	-0.06	0.27	0.18	0	-0.32	0.06	-0.17	-0.13	1												latino
asian	-0.02	0.08	0.03	0.03	-0.22	0.07	0.02	0.1	-0.08	1											asian
black	0.03	-0.02	0.03	0.04	0.06	-0.02	-0.04	-0.02	-0.12	-0.09	1										black
native amer	-0.01	0.02	0.01	0.02	0	0	-0.02	-0.01	0.02	-0.02	-0.03	1									native amer
other race	-0.04	0.15	0.13	0.02	-0.16	0.05	-0.1	-0.03	0.33	-0.04	-0.08	-0.01	1								other race
white	0.01	-0.11	-0.11	-0.06	0.15	-0.05	0.08	-0.02	-0.04	-0.42	-0.68	-0.15	-0.45	1							white
income/person	0.14	-0.24	-0.23	0.14	0.07	0.15	0.16	0.21	-0.15	0.04	-0.1	-0.02	-0.08	0.11	1						income/person
water price	0.09	0	-0.02	-0.01	0	-0.01	0.03	0	-0.01	0.04	-0.05	-0.01	0.02	0.01	0.03	1					water price
water/person/yr	0.85	-0.31	-0.23	-0.01	0.05	-0.13	0.02	-0.06	-0.05	-0.03	0.04	0	-0.04	0.01	0.11	-0.24	1				water/person/yr
latitude	-0.07	-0.11	-0.05	-0.02	0.11	-0.04	0.01	-0.05	-0.28	-0.17	0.15	-0.06	-0.14	0.05	-0.01	-0.1	-0.02	1			latitude
longitude	0.02	-0.01	0.01	0.03	0.05	0.04	0.09	0.07	-0.13	0	-0.09	0	-0.03	0.08	0.08	0.19	-0.05	-0.03	1		longitude
% inc on water	0.41	-0.07	0.04	-0.08	-0.03	-0.24	-0.12	-0.15	0.07	-0.03	0.1	0.01	0.06	-0.09	-0.27	0.03	0.36	-0.02	-0.03	1	% inc on water
	water/person	hh size	% children	% adult	% citizens	% employed	% health ins	% hs grad	latino	asian	black	native amer	other race	white	income/person	water price	water/person/yr	latitude	longitude	% inc on water	

Multivariate clustering: iterate with different feature sets across various k's

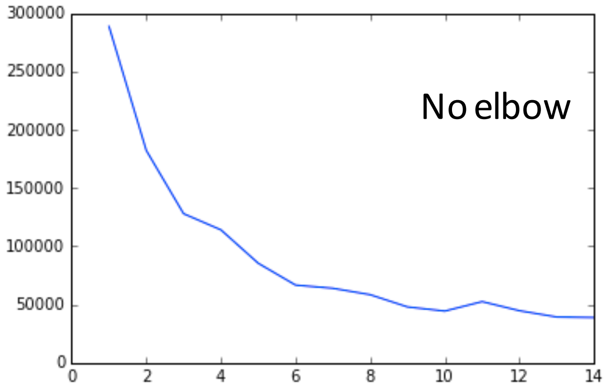
Get optimal* feature set and k clusters

- Remove features that are irrelevant to clusters to avoid curse of dimensionality
- Find the elbow across various feature sets and k's to get optimal feature set and k

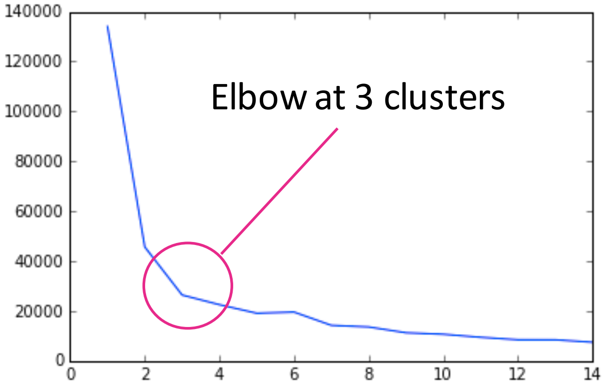
Features = 18



Features = 9

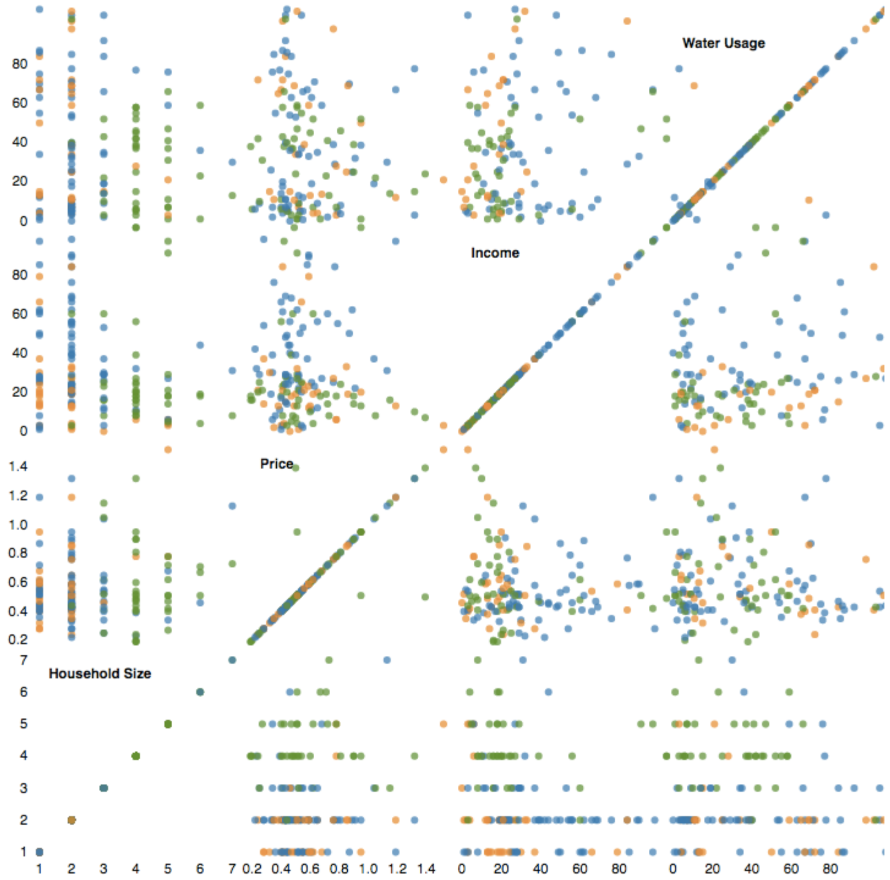


Features = 5



*Optimal meaning to the best of my abilities

Multivariate clustering: visualize outputs



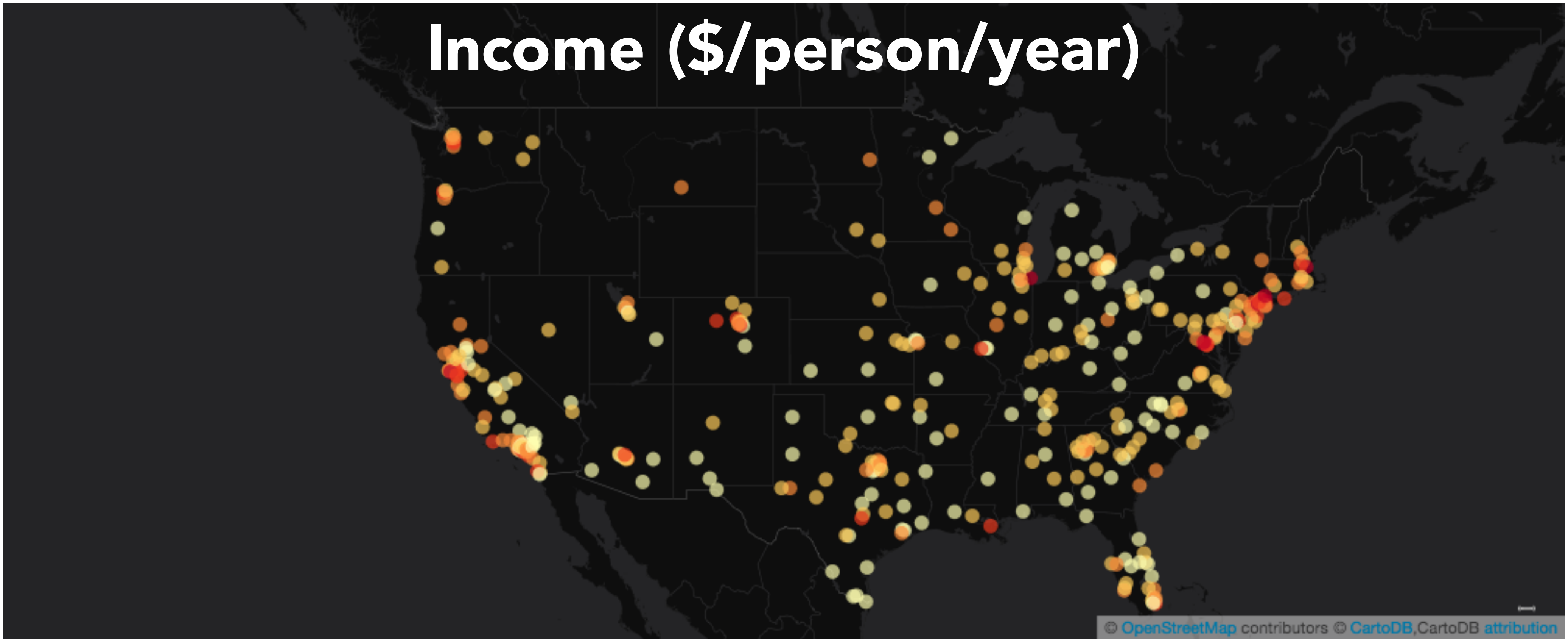
- Visualize facets of higher dimensional space through scatterplot
- Here, the 3 colors each represent a class
- However, classes do not seem coherent based on visual inspection
- Therefore, preliminary multivariate cluster results inconclusive

Conclusions

- **Overall:** Univariate results were more intelligible and conclusive than multivariate ones
- **Univariate clustering:**
  - Jenks often preferable to quintiles → better analyses and visualizations
  - Method good for getting quick, high-level insights
- **Multivariate clustering:**
  - Initial results inconclusive despite good-looking elbow plots
  - May not be the appropriate method for this problem
- **Other approaches v multi-variate clustering:** Treating this as a supervised problem may be more appropriate for this analysis

Appendix

Univariate analysis: Results



Income ($/person/year)

© OpenStreetMap contributors © CartoDB,CartoDB attribution