Clustering US states for fair distribution of US education budget among them

October 12, 2019

Introduction

Let us suppose that the US government wants to classify it's states into three tiers so that it can distribute its education budget among the states in such a manner that the ones belonging to the third tier (having less educational development) get larger share, and the ones belonging to the first tier (having more educational development) get smaller share.

Understanding the problem

Basically, the task here is to divide the states into three clusters.

Tier 1	Tier 2	Tier 3
More educationally developed	Moderately educationally developed	Less educationally developed
Less funds need to be allocated	Average funds need to be allocated	More than average funds need to be allocated

One Important Question:
How can one measure the educational development of a state?

Understanding the data

Data Sources

- assume here that the educational development can be measured by the number of educational institutions

- the Foresquare API is being used to extract the information of the educational institutes for each state

Feature Engineering

Feature Set A

- 'School',
- 'Elementary School'
- 'Private School'
- 'Preschool'
- High School'

Feature Set B

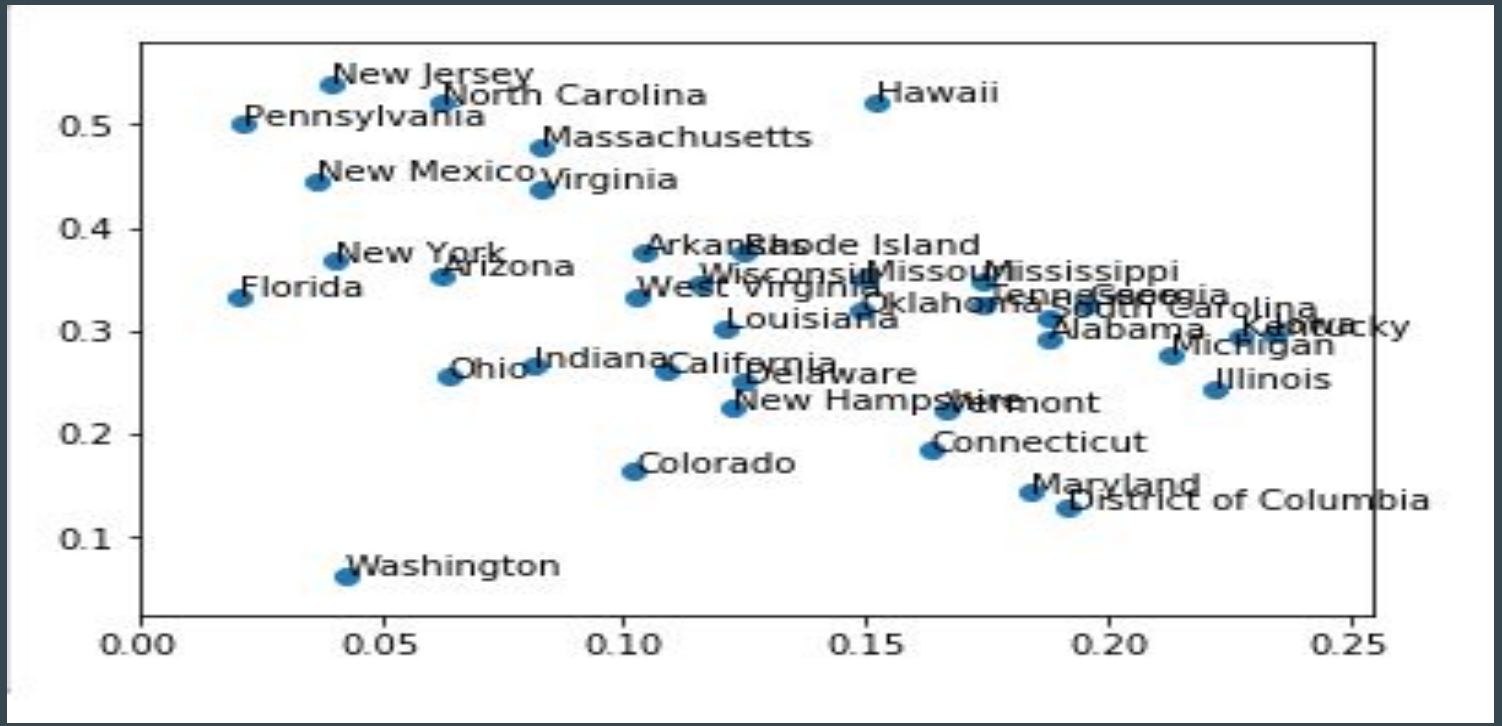
- 'College Academic Building'
- 'College Science Building'
- 'General College & University'
- 'College Engineering Building'
- 'College Classroom'
- 'College Arts Building'
- 'College Library'
- 'Medical School'
- 'Trade School'
- 'Cooking School'
- etc

Feature Set C

- 'History Museum'
- 'Auditorium'
- 'Convention Center'
- 'Library'
- 'Art Gallery'
- 'Bookstore'
- 'Arts & Crafts Store'
- 'Adult Education Center'
- etc

Methodology

Apply K-means clustering algorithm



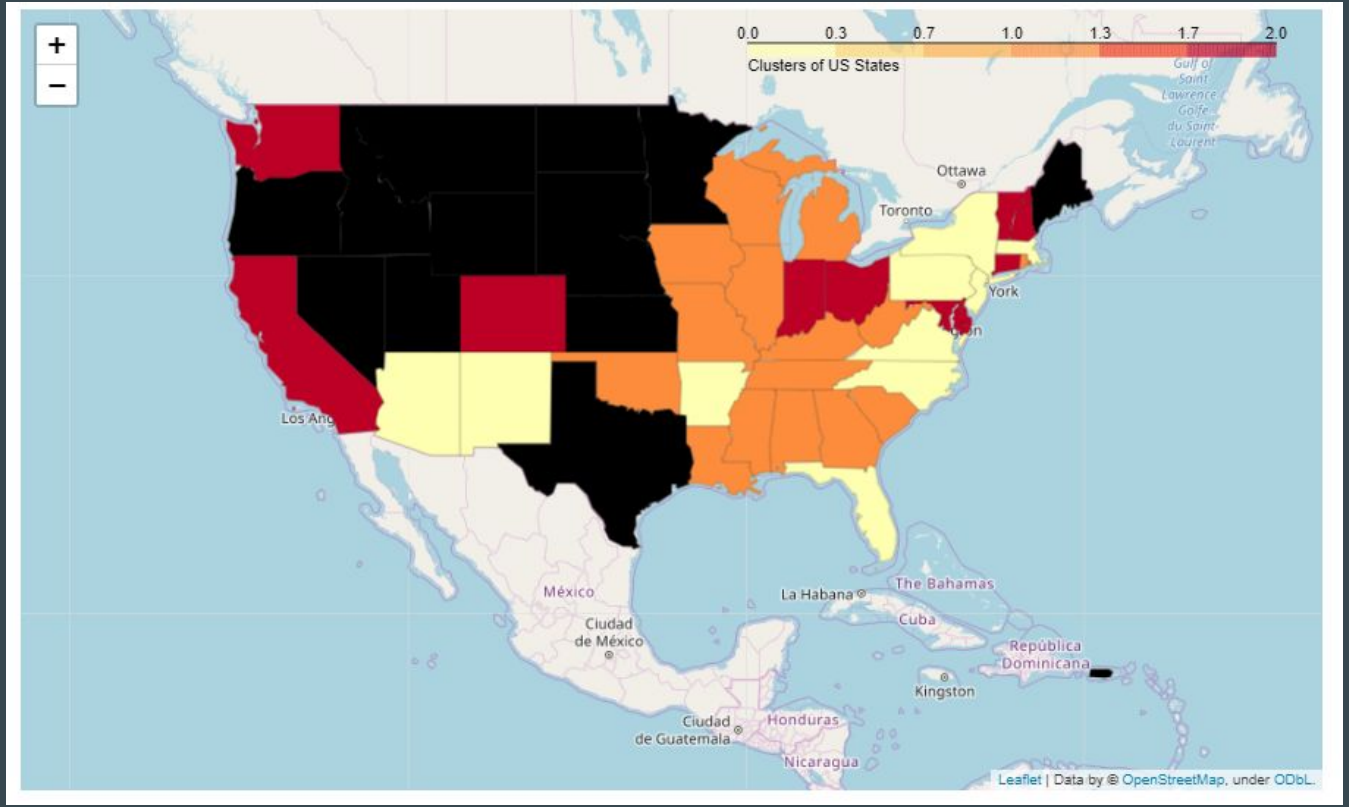
	Neighborhood	feature_set_1	feature_set_2	feature_set_3	Venue
0	Alabama	0.187500	0.291667	0.020833	48
2	Arizona	0.062500	0.354167	0.020833	48
3	Arkansas	0.104167	0.375000	0.083333	48
4	California	0.108696	0.260870	0.065217	46
5	Colorado	0.102041	0.163265	0.040816	49

Features for each state

	region	label
0	Alabama	1
2	Arizona	0
3	Arkansas	0
4	California	2
5	Colorado	2

Labels (Cluster Identifier) for each state

Results


0.0
0.3
0.7
1.0
1.3
1.7
2.0
Clusters of US States
Ottawa
Toronto
York
México
Ciudad de México
La Habana
The Bahamas
Cuba
República Dominicana
Kingston
Ciudad de Guatemala
Honduras
Nicaragua
Leaflet | Data by © OpenStreetMap, under ODbL.

Discussion

- Most of the states in maroon color, e.g Washington, California, Colorado etc have more educational institutions per person with a number of art galleries, libraries, book stores, convention centres, museums, etc and hence belong to tier 1
- Cities in yellow color should come in tier 2 because they have enough educational institutions but lack in libraries, museums, book stores, etc and hence deserve a little more attention than the ones in tier 1.
- Tier 3 cities, e.g. Mississippi, Alabama, etc appear in orange color and they are not well-known for educational institutions.

Conclusion

- The inconsistent data from the Foursquare API can be the major reason for some states being put into unsuitable clusters, e.g. Michigan
- The clustering is successful at a great extent but the results can be further improved if we remove the inconsistency in the data and collect more data.
- Such a clustering can help the government to spend less and achieve better education for every citizen.