Clustering US States for fair distribution of education budget among them

Shivam Kumar

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

- Tier 1: More Educationally Developed
- Tier 2: Moderately Educationally Developed
- Tier 3: Less Educationally Developed

One very important question arises here.

How can one measure the educational development of a state?

2. Data acquisition and cleaning

Data Sources and Cleaning:
Let's talk about the 'search' endpoint of Foresquare API. One cool thing about it is that if you enter a word in the query when you use this end point, the Foresquare API, behind the scene, searches not for the query word you mentioned but also for similar words.

Let us assume here that the educational development can be measured by the number of educational institutions. The more the educational institutions a state has, the more it is educatationally developed.

And the best thing about the Foursquare API is that if you search for school, it will return results of schools, colleges, universities, etc.

So, now we can us the Foresquare API to extract the information of the educational institutes for each state. Since we have to apply clustering on the states, we will try to keep the educational institutions into various categories, e.g. number of primary schools, number of secondary schools, universities, etc. We can then apply clustering on the states to divide them into three clusters. And, finally we will try to see whether the literacy data (literacy data for each of the states) justifies the results of clustering. Through the same data, we will deduce which clusters should be labelled as tier 1, tier2, and tier 3.

Once we have geojson data (data used for chropleth maps) of the US states and the data of neaby places, we first go on and remove all the states for which sufficient data is not avialable in any of the above two datasets.

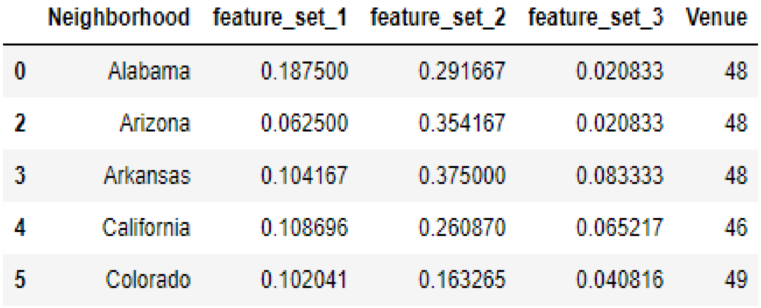
Feature Engineering:
Once we have the data of all nearby places, we will find the category of each place and then create a one-hot encoding. We then mean over the rows concerning to the same state. Now, we will use three features:

1. Schools
2. Colleges
3. Other Learning Environments

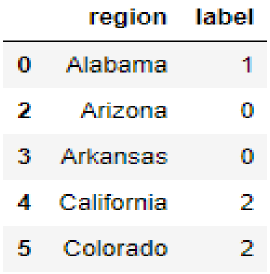
After the above dataframe is changed to accomodate these new features, we drop the rest of the columns.

3. Methodology

Since the problem is to divide the states into three tiers on the basis of the features, we can use K-Means clustering technique to divide the data points (states) ino three clusters (K = 3). We can then manually inspect each of the clusters to decide which of the three should be called tier 1 or tier 2 or tier 3. The machine learning pipeline takes the generated feature set and label each of the state as 0, 1, 2 (identifiers for the three clusters).



	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



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

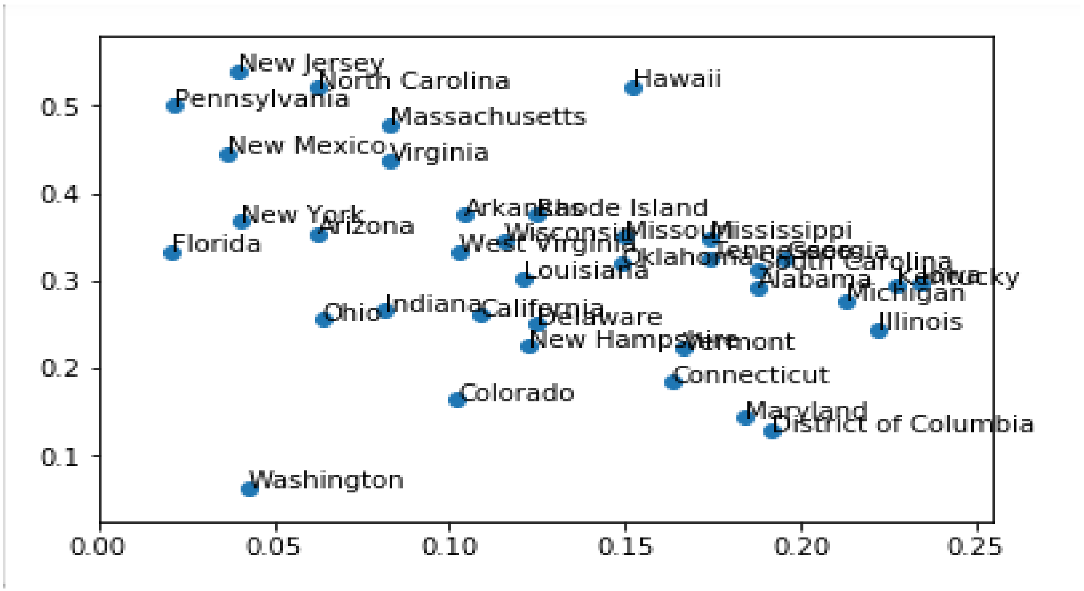
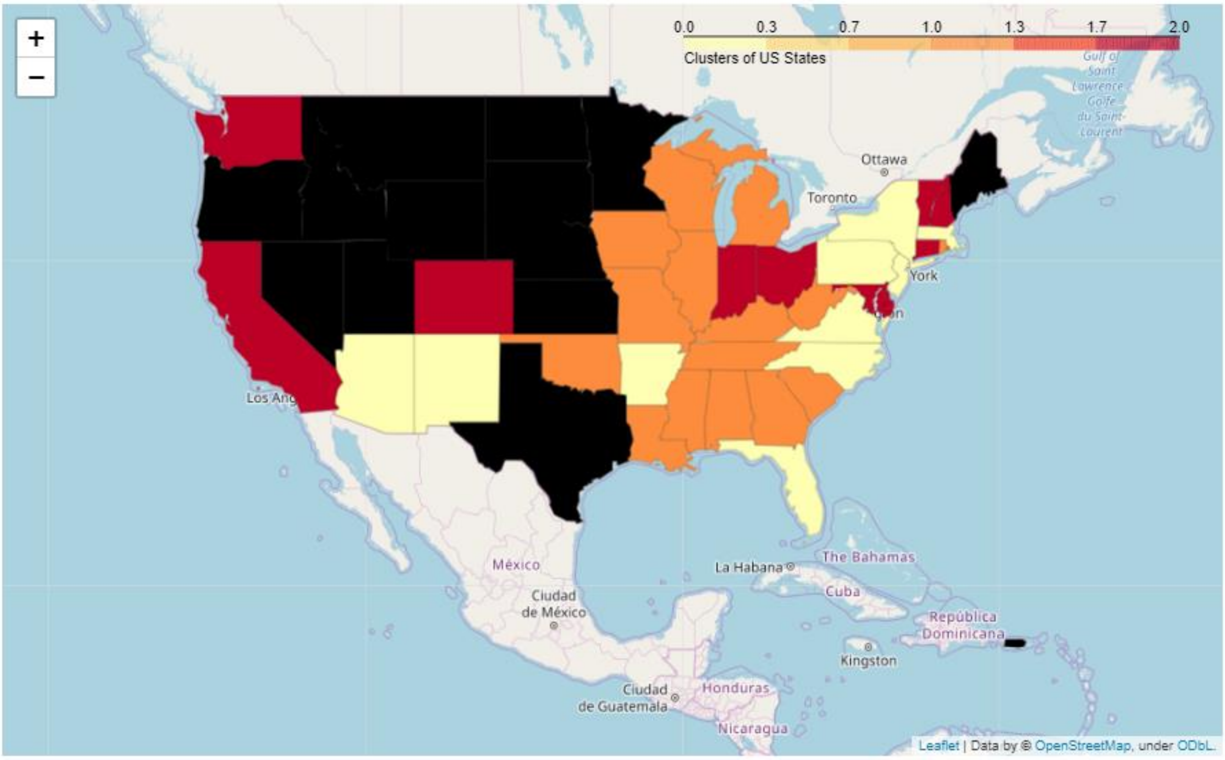


Fig: ML Pipeline

Fig: Plotting data points using the first two features

4. Results

Once the labels are generated, we can visualise them in a choropleth diagram, as shown.



Leaflet | Data by © OpenStreetMap, under ODbL.

5. Discussion

Since there is not enough data, a few states could not be processed, marked by the black color. However, the clusters look reasonable. Most of the states in maroon color, e.g Washington, California, Colorado etc have more educational institutions per person and they also have large number of art galleries, libraries, book stores, convention centres, museums, etc and hence can do good with less funds too.

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.

5. Conclusion

However, many states have been placed in tiers which fit them well but there are some states, e.g. Michigan which fit in a different tier than the predicted one. However, the inconsistent data from the Foursquare API can be the major reason for this. Even if the search query is 'Education', it is returning the data of restaurents and clubs in case of some states. Also, the maximum radius allowed is 100km which don't cover properly many of the states.

We have clustered the states into three tiers succesfully at a great extent but there is always room for improvement. The results can be further improved if we remove the incosistency in the data and collect more data.

Such a clustering can help the government to spend less and acheive better education for every citizen.