Final Report | Capstone Project – The Battle of Neighborhoods |Segmenting the USA states Capital into cluster and finding out the most common place in each cluster

1. Introduction:

As we all know that U.S.A is a large country with many states. And many people like to visit these state. But there are many cities to explore in each state. So knowing about the capital of each state will brief us about each state. This project will segment these capitals on the basis of similarity of the venue found around them which will help the explorer/visitor to know about how similar these capitals are with each other and also will be able to find out the most common venue in each cluster.

2. Data Section

Data Link: https://www.britannica.com/topic/list-of-state-capitals-in-the-United-States-2119210

We will scrap the dataset from the above mentioned link using beautiful soup and pandas. The dataset consists of USA state and its Capital city.

Foursquare API Data:

We will need data about different venues in different Capital city of each state. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API. After finding the list of Capital city along with their state, we then connect to the Foursquare API to gather information about venues inside each and every Capital city. For each Capital city, we have chosen the radius to be 500 meter. The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Capital city

2. Capital city Latitude

3. Capital city Longitude

4. Names of the venue e.g. the name of a store or restaurant

5. Venue Latitude

6. Venue Longitude

7. Venue Category

3. Methodology Section

Clustering Approach:

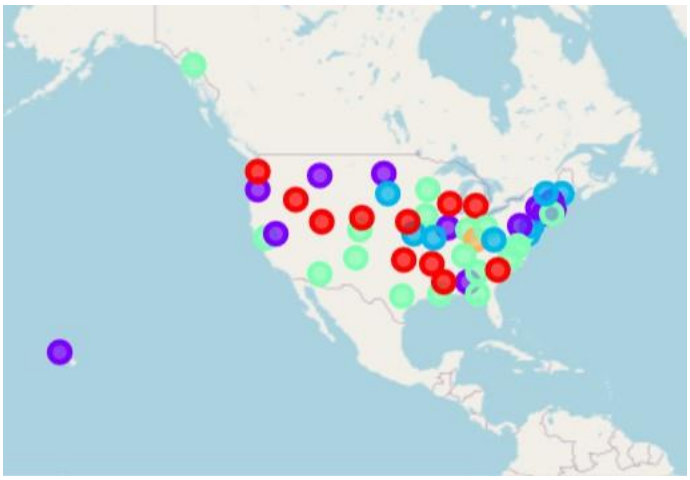
To compare the similarities of two cities, we decided to explore venues, segment them, and group them into clusters to find similar capitals. To be able to do that, we need to cluster data this is a form of unsupervised machine learning: k-means clustering algorithm.

Work Flow:

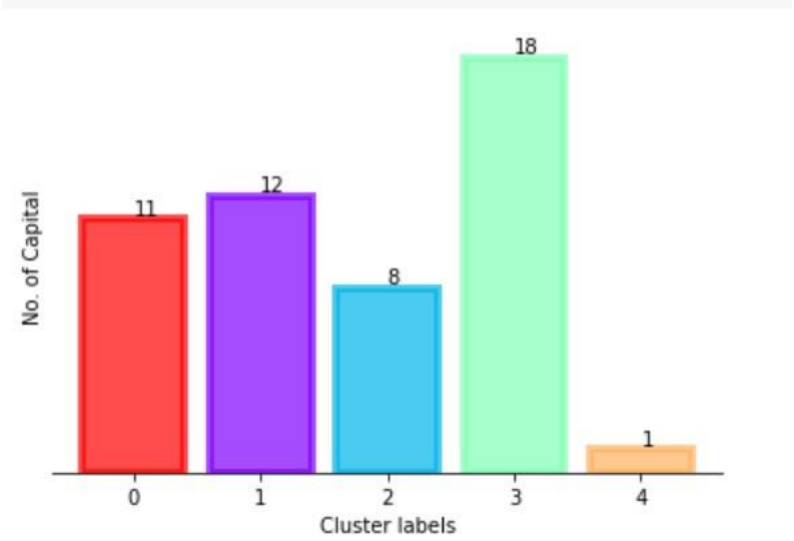
Using credentials of Foursquare API features of near-by places of the capital city would be mined. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

4. Results Section

The USA state capital city was segmented into five clusters and was plotted in map to show the similar cluster group with the same color.



Number of capital city in cluster 0, cluster 1, cluster 2, cluster 3 and cluster 4 was found to be 11, 12, 8, 18, and 1 respectively.



The most common venue found in cluster 0, cluster 1, cluster 2, cluster 3 and cluster 4 was coffee shop, coffee shop, pizza shop, bar, history museum. And cluster 4 only contains 1 capital city so this capital city must be unique.

5. Discussion Section

The major purpose of this project is to suggest visitor in USA about the Capital city of each state and showing them which capital city are similar to other in terms of the venues located nearby.

Problem Which i Tried to Solve:

1) Determine the cluster label for each capital.

2) Visualize the cluster label in the map.

3) Determine most common venue in each cluster.

Conclusion Section

In this project, using k-means cluster algorithm I separated the neighborhood into 5 different clusters. And by locking the most common place in each cluster I came to know that bar, pizza shop, coffee shop are very common and a nice startup business to anyone who wants to start their own business in USA.

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appreciation. This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.