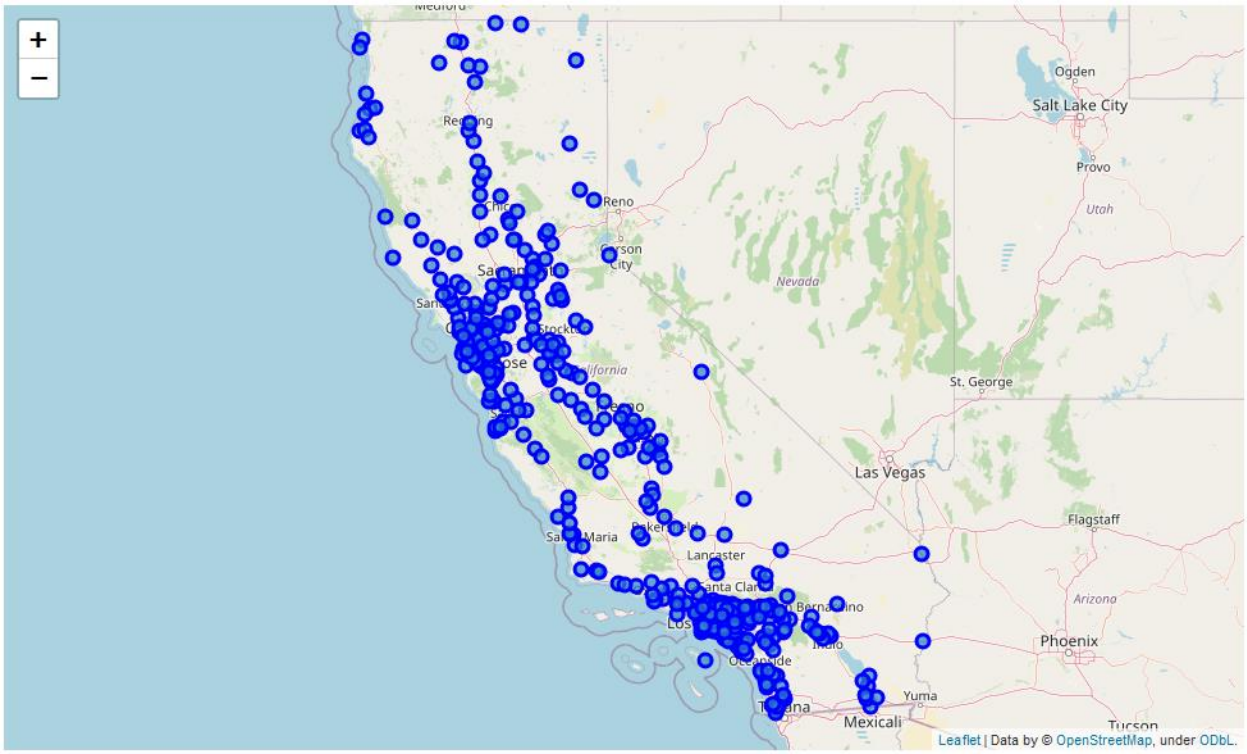
The Battle of Neighborhoods – California Cities

Coursera Capstone Project Report



Introduction/Business Problem

In following suit from prior Ungraded External Tool practice, Peer graded review assignments and supplemental material, an analysis of California, moreover, to assess the feasibility of a new business, considering the represented establishments present in specific clusters.

A Kaggle dataset: https://www.kaggle.com/camnugent/california-housing-feature-engineering, was utilized to represent the latitudinal and longitudinal backbone of the cities, as to support inclusion and integration of the Foursquare API.

The Foursquare API as in prior weeks of this course, is used to analyze the venues and obtain clusters for analysis. In this project, k-means clustering algorithm is utilized to accomplish this analysis. Additionally, we utilize the silhouette score metric in order to arrive at an optimum number of clusters.

Upon arrival of clustering, we will utilize the Folium visualization library to visually highlight specific clusters for our business implementation goals.

The target of this report would be to glean insights from a entrepreneurial aspect, one which maximizes the feasibility of a new business given the prevalence, or lack of presence of businesses within a given cluster.

Data Requirements

The data set utilized from Kaggle contains the latitude and longitudinal coordinates for 459 cities.

For the purposes of comparability and due to computational limitations, a measure, most likely venue count, coupled with a threshold tolerance will needed to be implemented in order to draw any sort of conclusions given that there are only (2) sources of data – Kaggle data set on California cities, and the Foursquare API.

Methodology

Given our 459 unique neighborhoods and the capability via the Foursquare API to obtain information relative to the number of venue types within a given city, we will utilize one hot encoding in order to streamline/refine our return data.

If we were not performing this data sort, we would ultimately find it very challenging to process the data in a timely manner. Especially given the hardware restrictions of my current laptop.

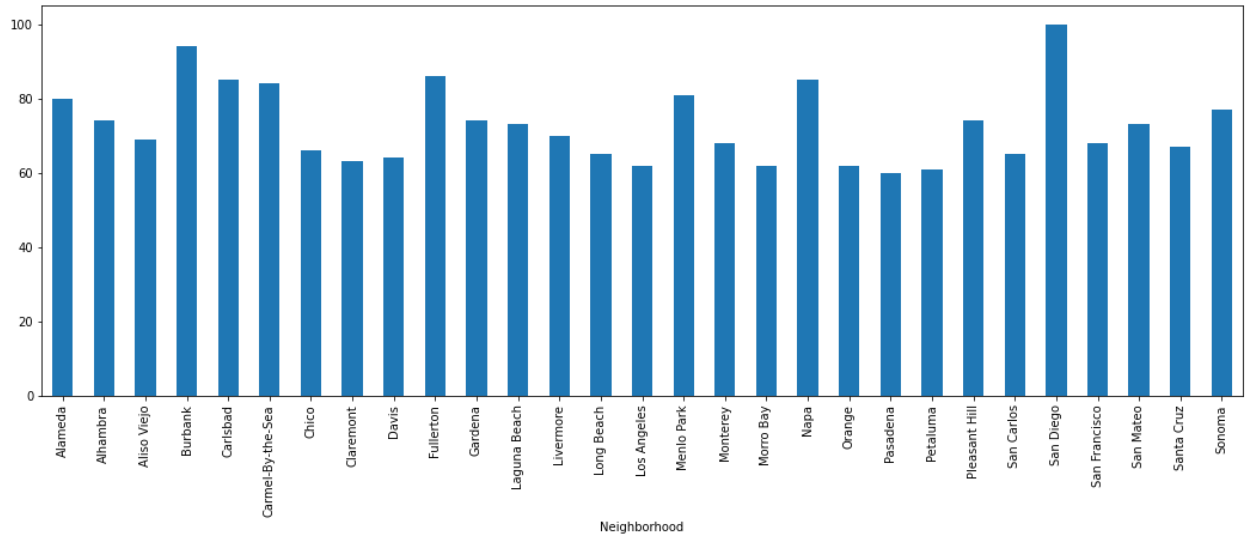
Of the entirety of all the neighborhoods, we ended up utilizing data from 29 California cities. We segregated our search to the neighborhoods who upon combination with Foursquare API, returned greater than 65 venues in each locale. This would support a sizeable populous to provide return on investment for an entrepreneur hoping to start an establishment in a future cluster. We then performed one hot encoding to find the 10 most common venue types as to:

1. Highlight existing businesses in a cluster.
2. Attempt to identify a trend for clusters representative of our California city KNN analysis.

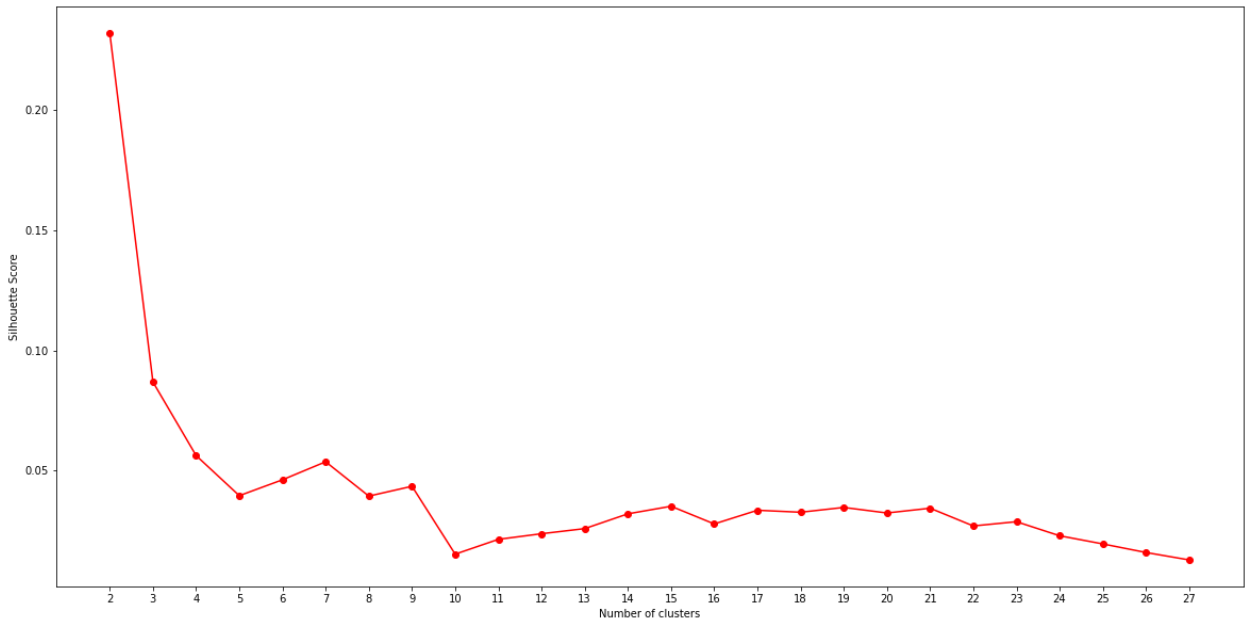
Results

Initial Scope of California Cities by Venue Count (Fig. 1.1)

Refined Scope of California Cities by Venue Count (Fig. 1.2)

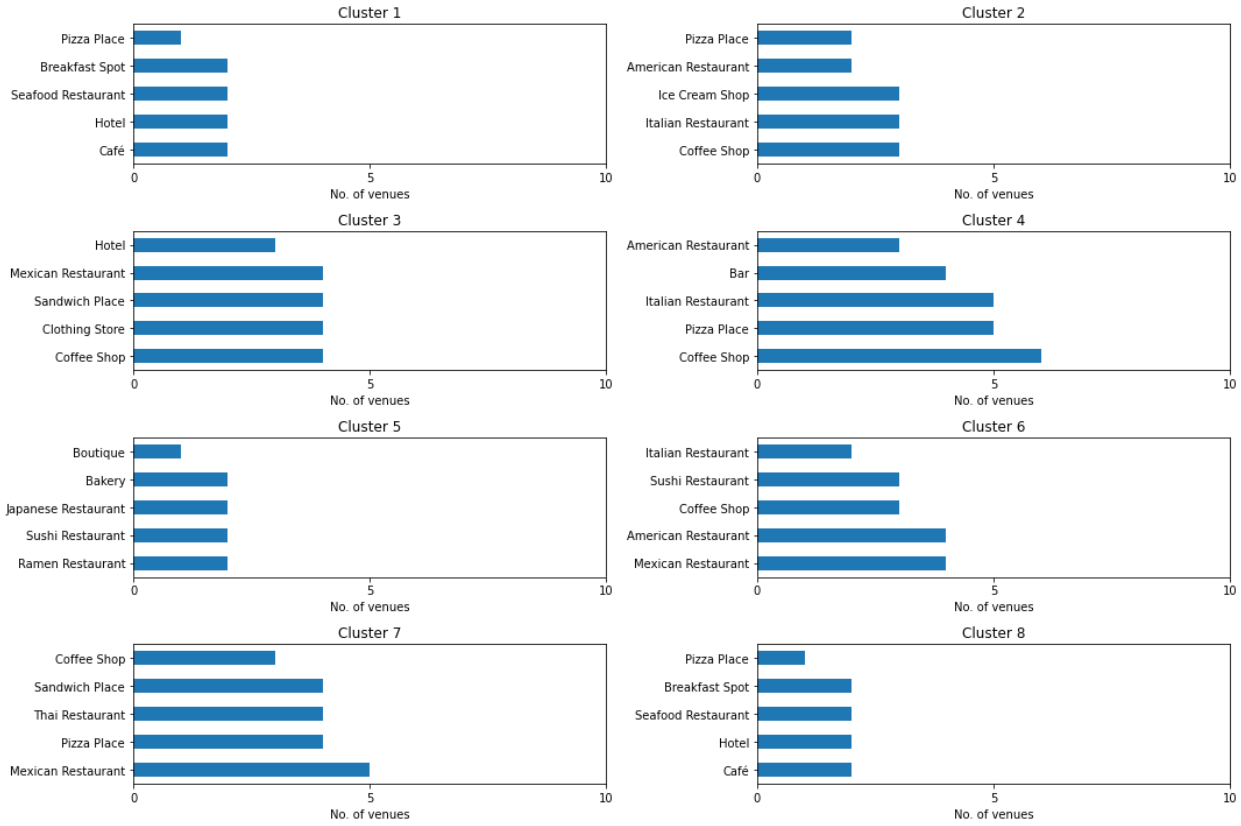


Silhouette Score Metric (Fig. 1-3)



Results Cont.

Venues Count by Cluster as determined by Fig. 1-3 (Fig. 1-4)



Discussion

In following suit from prior Ungraded External Tool practice, Peer graded review assignments and supplemental material, an analysis of California, moreover, to assess the feasibility of a new business, considering the represented establishments present in specific clusters.

A Kaggle dataset: https://www.kaggle.com/camnugent/california-housing-feature-engineering, was utilized to represent the latitudinal and longitudinal backbone of the cities, as to support inclusion and integration of the Foursquare API.

The Foursquare API as in prior weeks of this course, is used to analyze the venues and obtain clusters for analysis. In this project, k-means clustering algorithm is utilized to accomplish this analysis. Additionally, we utilize the silhouette score metric in order to arrive at an optimum number of clusters.

Upon arrival of clustering, we will utilize the Folium visualization library to visually highlight specific clusters for our business implementation goals.

The target of this report would be to glean insights from a entrepreneurial aspect, one which maximizes the feasibility of a new business given the prevalence, or lack of presence of businesses within a given cluster.

Conclusion

Based upon the Fig. 1-4 a gathering place for coffee, is highly desirable. An establishment is within our top 5 venues for 7 of our 8 clusters.

It is only within a very small cluster of cities, namely Gardena and Los Angeles do we see a top 5 venue listing that does not include a café, or coffee shop.

It is worth noting that residing in the U.S., I find it hard to believe that café/coffee shops are underrepresented in LA. However, LA is a densely populated, highly segregated area with various socioeconomic environments occupying it. Perhaps, there are viable areas to perform additional analysis which would yield positive business opportunities.

Additionally, per Fig. 1-3, lower representation is observed and it could also be theorized that as representation, updating of data occurs, this trend may be observed. Or, as population increases, the need for more coffee/cafes will be realized.