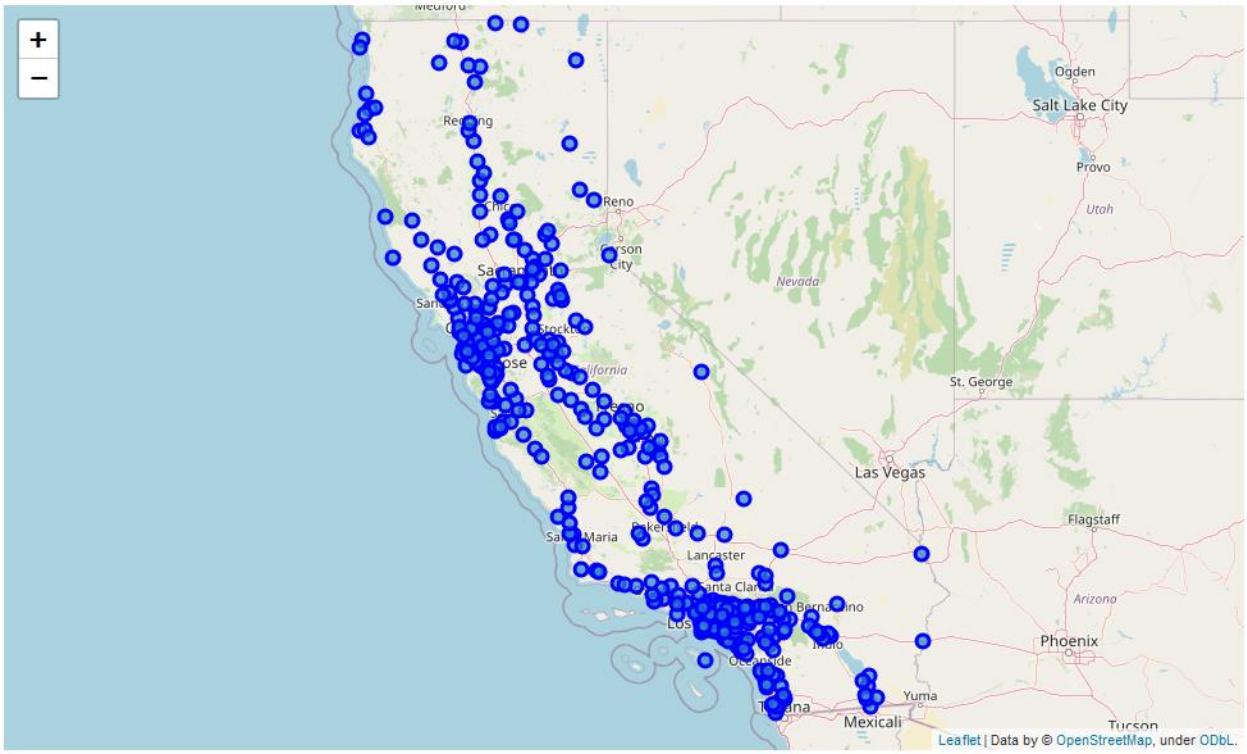
The Battle of Neighborhoods – California Cities

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.

The simple and direct methodology will be to utilize the coordinates of the California cities, joining the dataset with the venue information from the Foursquare API, in order to numerically assess the count of venue types within given California cities of a to be determined cluster size.

Upon designation of an optimal cluster size for analysis, further comparison will be conducted and optimal areas for specific businesses will be presented in my final report.