Data Analysis for starting a new restaurant in California.

Capstone Project - The Battle of Neighborhoods

IBM Applied Data Science Capstone

Varun Kumar Anand

30-Dec-2020

Contents

Introduction

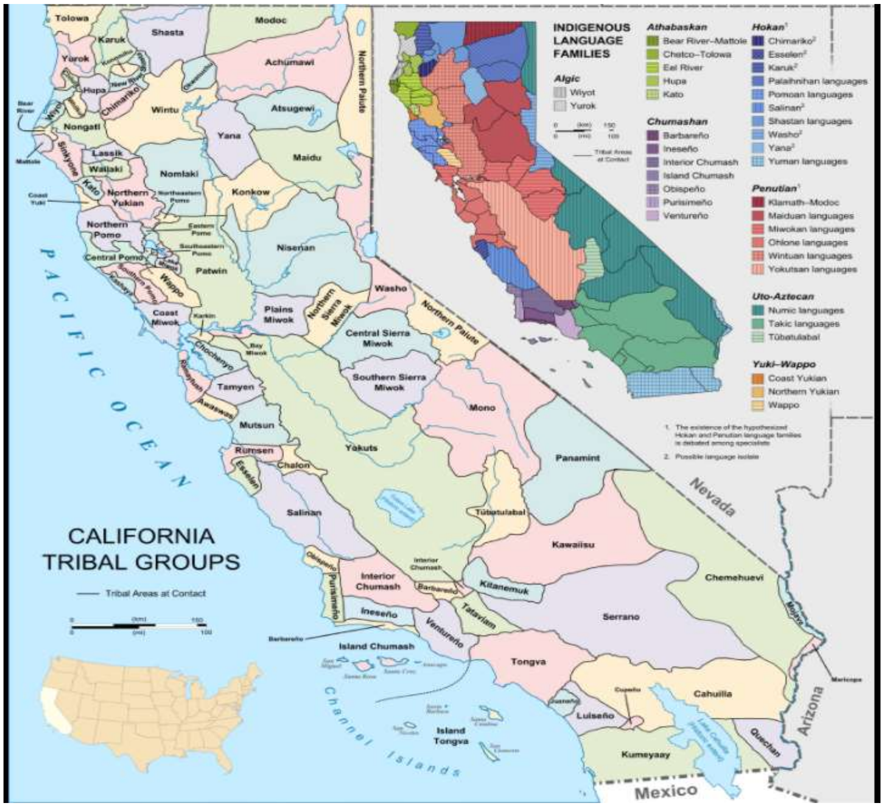
Opening a new restaurant is a challenge for anyone considering the major hurdle of selecting a location for that restaurant which can increase its probability of getting a good Business.

Location is not the only factor involved but it can give you an upper edge in business.

This project concentrates on finding locations in California where opening a restaurant can give good business.

Below are the factors which have considered for selecting California for this study.

- California is the most populous U.S. state and the third-largest by area, and is also the world's thirty-fourth most populous subnational entity.
- California's economy, with a gross state product of $3.0 trillion, is the largest sub-national economy in the world.
- California integrates foods, languages, and traditions from other areas across the country and around the globe.
- California's agriculture industry has the highest output of any U.S. state.



Problem

To Provide locations in California for staring a new Restaurant which are more likely to give a good business.

Data Acquisition and Cleaning

Data Sources

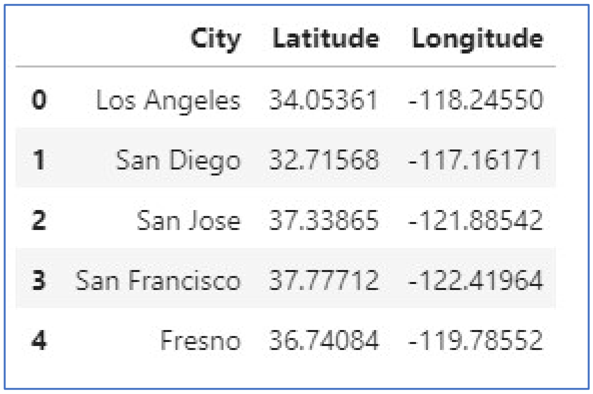
Data for this Study has been taken from “California demographics by Cubit” where cities for California are provided based on their population in Decreasing Order. For this Study we have used only first 100 entries.

https://www.california-demographics.com/cities_by_population

Data Cleaning

Data was extracted from “California demographics by Cubit” Webpage using web scraping with BeautifulSoup. Irrelevant data was removed and only name of cities were kept.

Later Latitude and Longitude values for these cities were calculated using GeoPy and two new columns were added in our Dataframe as shown below.



	City	Latitude	Longitude
0	Los Angeles	34.05361	-118.24550
1	San Diego	32.71568	-117.16171
2	San Jose	37.33865	-121.88542
3	San Francisco	37.77712	-122.41964
4	Fresno	36.74084	-119.78552

Venue Data for these cities was calculated using Foursquare API. This data was used to study the venues in various cities in California. This data provided us information regarding various restaurants available in the area and helped in drawing main conclusion for this project.

Libraries Used

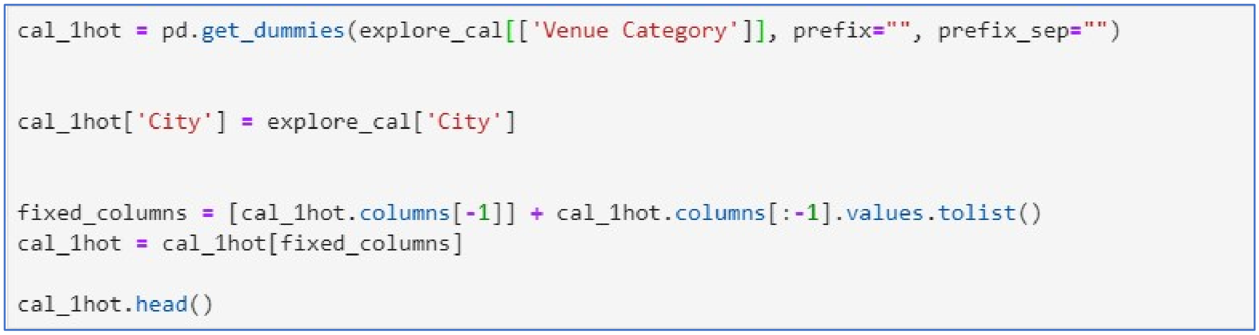
- Pandas – For Data Analysis
- Json – For handling json files
- Geopy – For converting address into Latitude and Longitude
- sklearn.cluster - for Kmeans Clustering
- folium – For Maps rendering
- BeautifulSoup - For parsing HTML and XML files
- Numpy – For Numerical analysis
- Matplotlib – Python plotting module

Methodology

Feature Extraction

Feature extraction was carried out through one hot encoding. In this method, each feature is a category that belongs to a venue which is then converted into binary, this means that 1 means this category is found and 0 means the opposite.

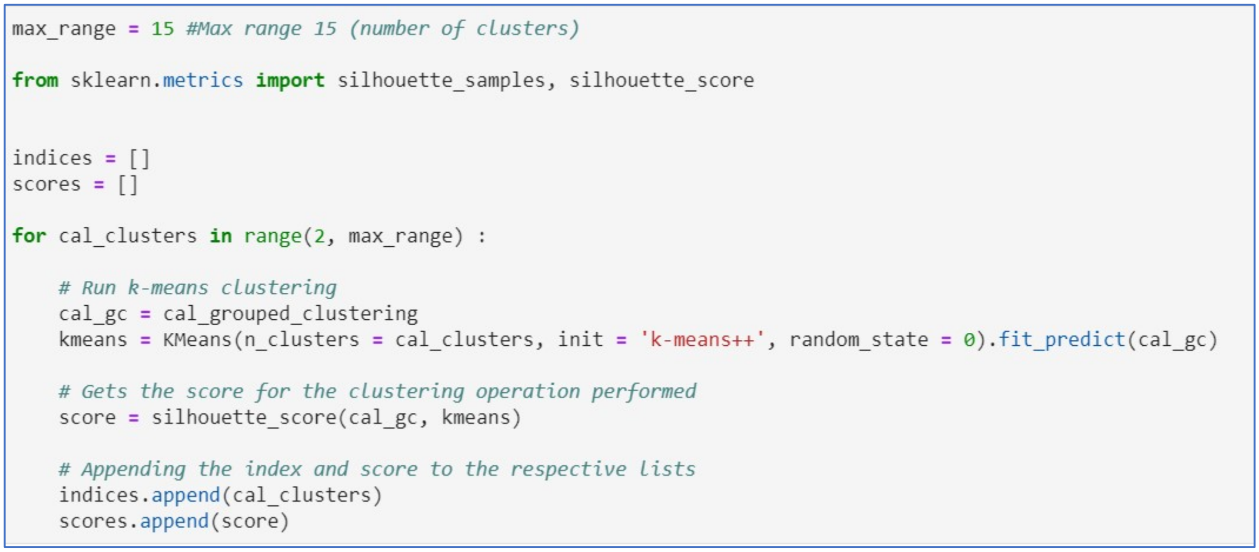
Then all the venues are grouped by the cities, computing mean at the same time. This will give us a venue for each row and each column will contain the frequency of occurrence of that particular category.



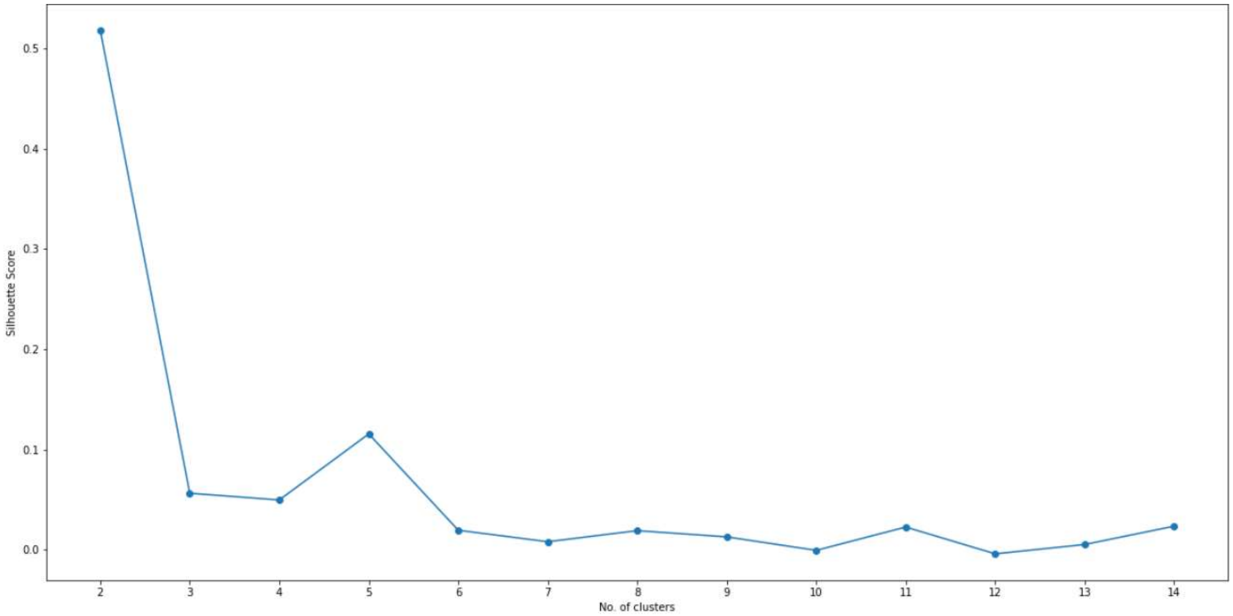
```
cal_1hot = pd.get_dummies(explore_cal[['Venue Category']], prefix="", prefix_sep="")


cal_1hot['City'] = explore_cal['City']


fixed_columns = [cal_1hot.columns[-1]] + cal_1hot.columns[:-1].values.tolist()
cal_1hot = cal_1hot[fixed_columns]

cal_1hot.head()
```

Unsupervised Learning- K-Means

K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective is to minimize the data dispersion for each cluster. This, each group found represents a set of data with a pattern inside the multi-dimensional features. It is necessary for this algorithm to have a prior idea about the number of clusters since it is considered an input of this algorithm. For this reason, the elbow method is implemented. A chart that compares error vs number of cluster is done and the elbow is selected. Then, further analysis of each cluster is done.



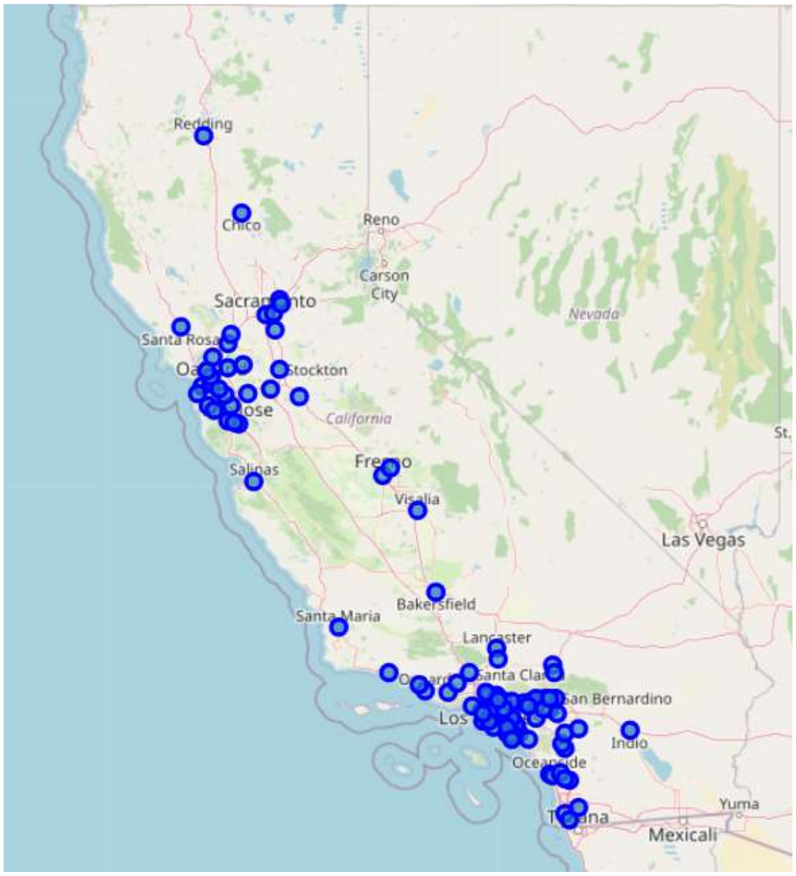
```python
max_range = 15 #Max range 15 (number of clusters)

from sklearn.metrics import silhouette_samples, silhouette_score


indices = []
scores = []

for cal_clusters in range(2, max_range) :

    # Run k-means clustering
    cal_gc = cal_grouped_clustering
    kmeans = KMeans(n_clusters = cal_clusters, init = 'k-means++', random_state = 0).fit_predict(cal_gc)

    # Gets the score for the clustering operation performed
    score = silhouette_score(cal_gc, kmeans)

    # Appending the index and score to the respective lists
    indices.append(cal_clusters)
    scores.append(score)
```



Plotting

Visualizing data often gives a clear understanding of the data as it is easier to spot patterns in visualized data as compared to quantitative data.

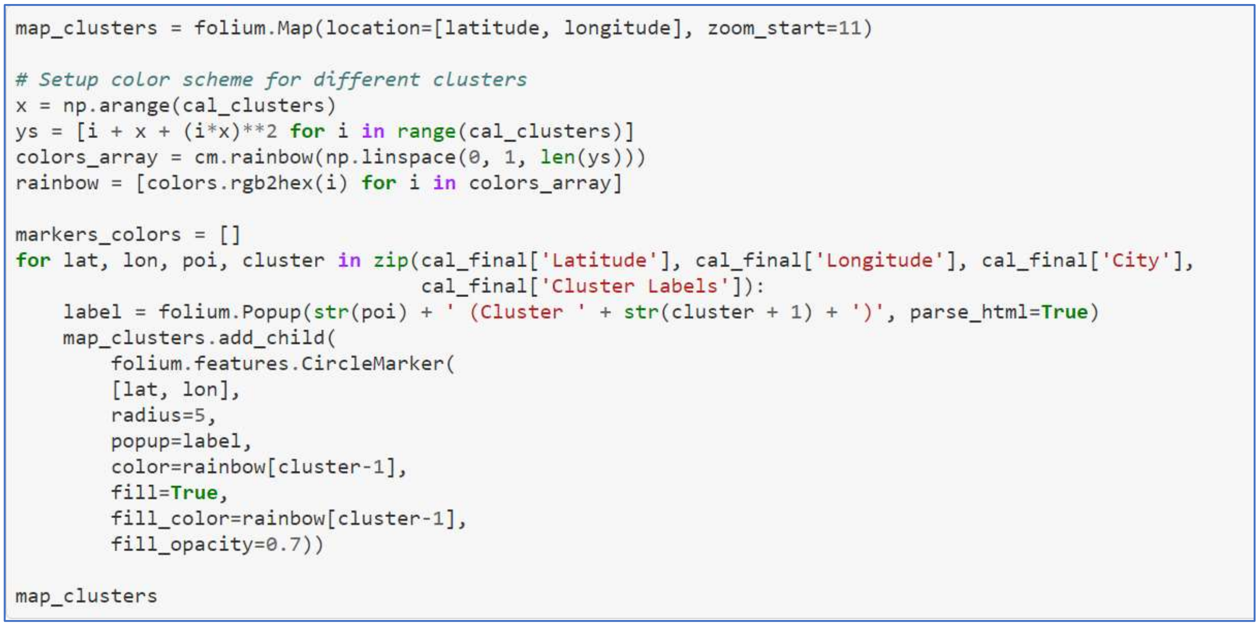
Folium Library was used to plot maps of California as well as its cities. Folium was also used to visualize the cluster.



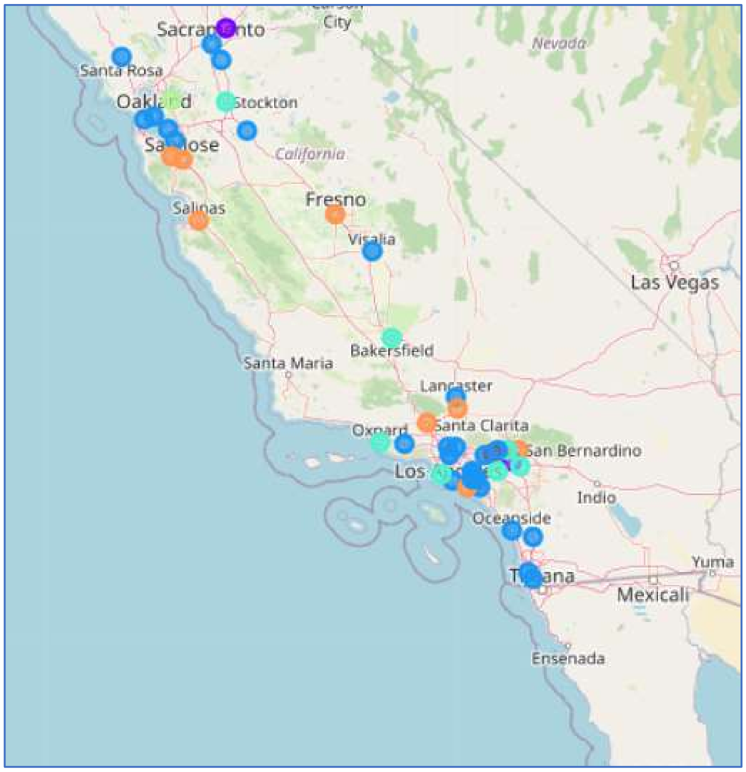
Results

K Means method was applied to the dataframe. The value of optimal K was calculated as 6 using Elbow method.

The code and plotting of clusters is given below



```
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=11)

# Setup color scheme for different clusters
x = np.arange(cal_clusters)
ys = [i + x + (i*x)**2 for i in range(cal_clusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

markers_colors = []
for lat, lon, poi, cluster in zip(cal_final['Latitude'], cal_final['Longitude'], cal_final['City'],
                                  cal_final['Cluster Labels']):
    label = folium.Popup(str(poi) + ' (Cluster ' + str(cluster + 1) + ')', parse_html=True)
    map_clusters.add_child(
        folium.features.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7))

map_clusters
```



Discussion

After Visualizing the cluster, Individual clusters were studies and below are the observations

- Cluster 3 contains most suitable cities for opening new restaurants and Cluster 4 has a greater concentration of Mexican Restaurants.
- Based on the Cuisine of the new Restaurant proper location can be selected from Cluster 3 and 4.
- Prices and other variables are excluded from this study.
- Improvements can be made with paid/premium account of foursquare, where Ratings and Reviews can also be considered.

Conclusion and Future Directions

- Machine Learning is very helpful in solving various Business problems.
- Python’s inbuilt libraries are very helpful for applying machine learning technique and Data analysis.
- This project can be re-used for other states also provided data is available in Foursquare API