Battle of Neighborhoods. Los Angeles California.

Coursera Capstone Project

Victoria Tokmakova

05.30.2020

Part 1: Problem Description

Every time people want to find better place to live, they explore the place and try to get as much information as possible about it.

It can be the neighborhood, locality, market, price of the place, schools nearby and many more factors including neighborhood analysis.

In my research, I suggest create a search algorithm which usually returns the requested features such as population rate, median house price, school ratings, crime rates, weather conditions, recreational facilities etc.

It would be useful to have an application which could make easy by considering a comparative analysis between the neighborhood with provided factors.

This project helps the end user or the stakeholder to achieve the results which will not only recommend but also saves a lot of time in manual search. It can be used by the user at the time of rental apartment or buy house in a locality based on the distribution of various facilities available around the neighborhood. As an example, this project would compare 2 randomly picked neighborhoods and analyses some common venues in each of those two neighborhoods.

Also, this project uses K-mean clustering unsupervised machine learning algorithm to cluster the venues based on the place category such as restaurants, park, coffee shop, gym, clubs etc. This would give a better understanding of the similarities and dissimilarities between the two chosen neighborhoods to retrieve more insights and to conclude with ease which neighborhood wins over other.

Part 2: Data Sets

For this project we need the following data:

Los Angeles data that contains list Boroughs, Neighborhoods along with their latitude and longitude. Also, we get information about population, school rating, housepricing etc.

Data source: https://docs.gaslamp.media/wp-content/uploads/2013/08/zip_codes_states.csv

We will need geo-locational information about that specific borough and the neighborhoods in that borough and finding categories of areas.

We a going to use:

• Foursquare API:

This API has a database of more than 105 million places. This project would use Four-square API as its prime data gathering source. Many organizations are using to geo-tag their photos with detailed info about a destination, while also serving up contextually relevant locations for those who are searching for a place to eat, drink or explore. This API provides the ability to perform location search, location sharing and details about a business. Foursquare users can also use photos, tips and reviews in many productive ways to add value to the results.

• Work Flow:

HTTP requests would be made to this Foursquare API server using zip codes of the Los Angeles city neighborhoods to pull the location information (Latitude and Longitude). Foursquare API search feature would be enabled to collect the nearby places of the neighborhoods. 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 700.

• Folium - Python visualization library would be used to visualize the neighborhoods cluster distribution of Los Angeles city over an interactive leaflet map. Extensive comparative analysis of two randomly picked neighborhoods world be carried out to derive the desirable insights from the outcomes using python’s scientific libraries Pandas, NumPy and Scikit-learn.

• Unsupervised machine learning algorithm K-mean clustering would be applied to form the clusters of different categories of places residing in and around the neighborhoods.

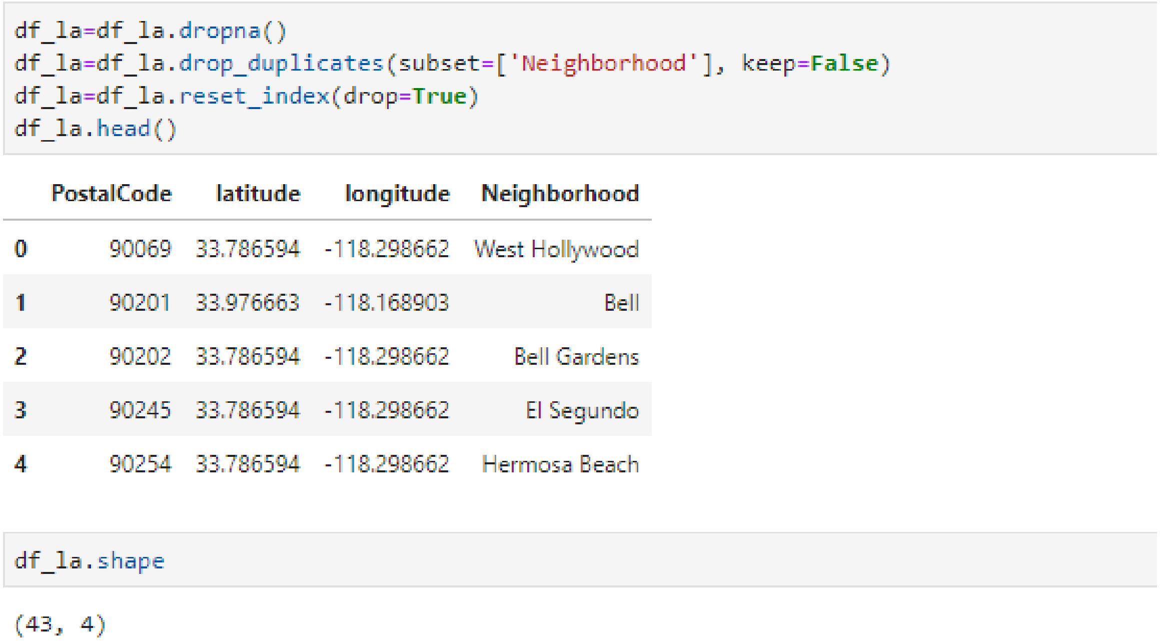
These clusters from each of those two chosen neighborhoods would be analyzed individually collectively and comparatively to derive the conclusions.

Python packages and Dependencies:

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- JSON – Library to handle JSON files
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests
- Matplotlib – Python Plotting Module
- Sklearn – Python machine learning Library
- Folium – Map rendering Library

Implementation

After getting and preparing data we get following dataframe:



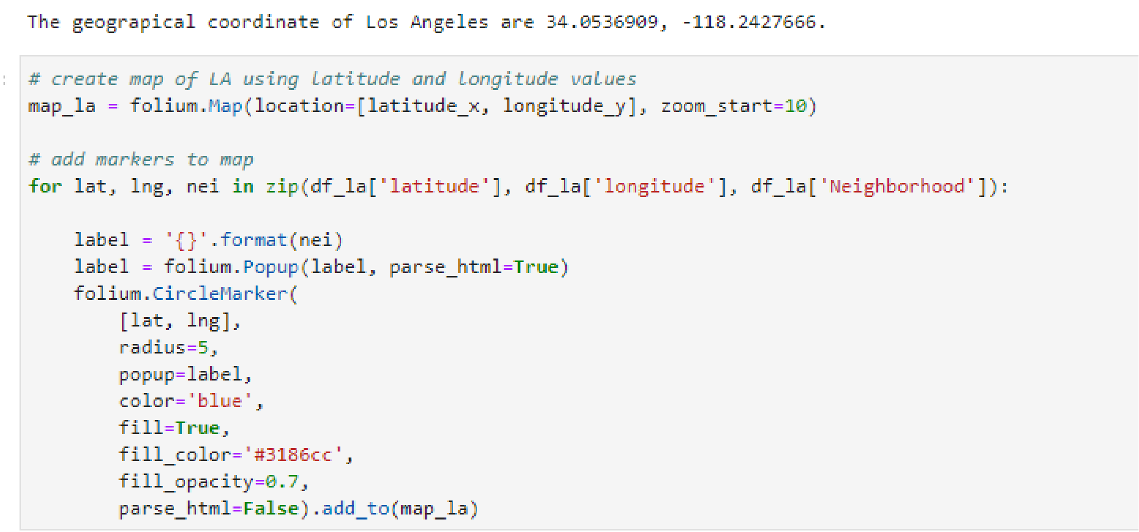
```
df_la=df_la.dropna()
df_la=df_la.drop_duplicates(subset=['Neighborhood'], keep=False)
df_la=df_la.reset_index(drop=True)
df_la.head()
```

	PostalCode	latitude	longitude	Neighborhood
0	90069	33.786594	-118.298662	West Hollywood
1	90201	33.976663	-118.168903	Bell
2	90202	33.786594	-118.298662	Bell Gardens
3	90245	33.786594	-118.298662	El Segundo
4	90254	33.786594	-118.298662	Hermosa Beach

```
df_la.shape
```

```
(43, 4)
```

Folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map.

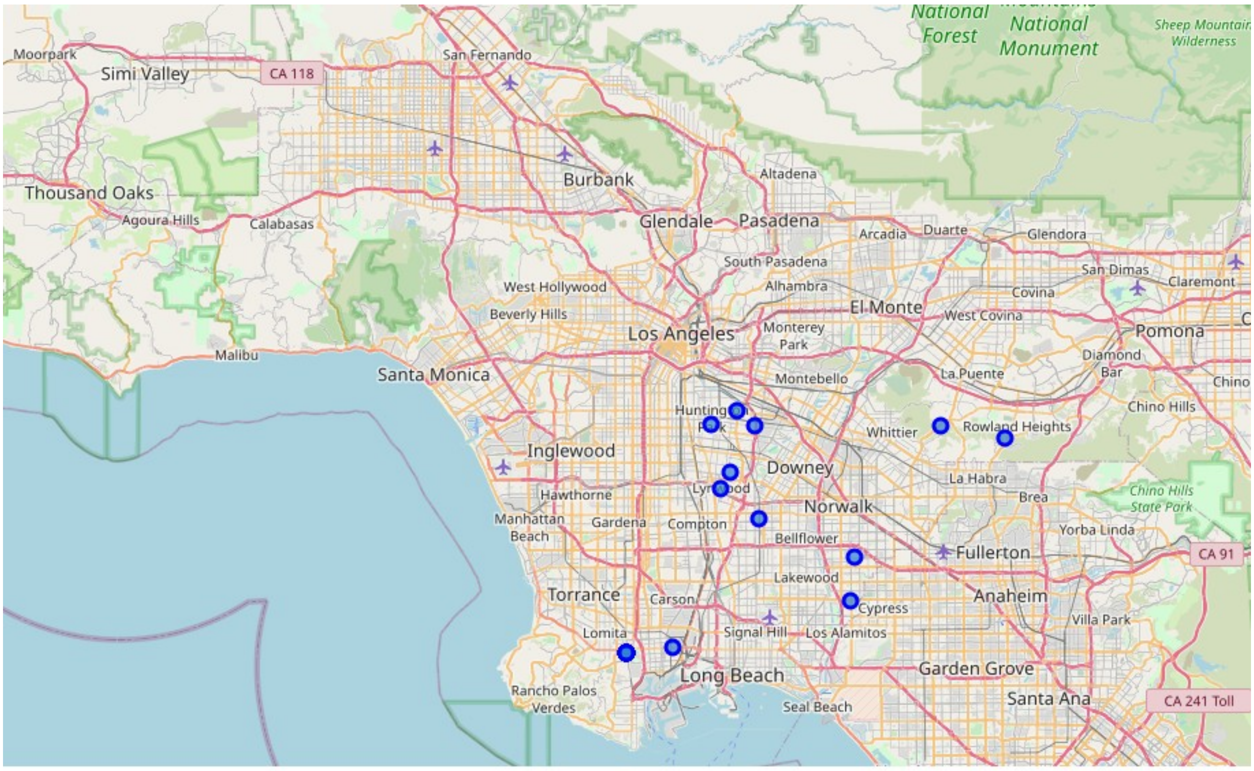


```
The geograpical coordinate of Los Angeles are 34.0536909, -118.2427666.
```

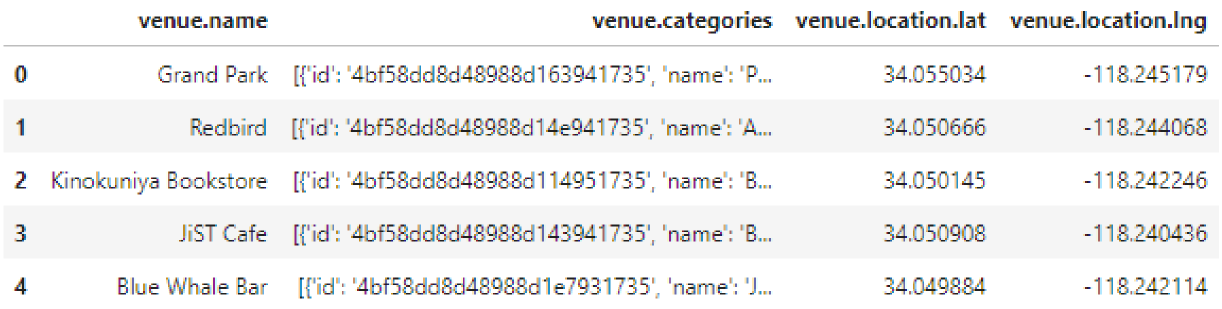
```
# create map of LA using latitude and longitude values
map_la = folium.Map(location=[latitude_x, longitude_y], zoom_start=10)

# add markers to map
for lat, lng, nei in zip(df_la['latitude'], df_la['longitude'], df_la['Neighborhood']):

    label = '{}'.format(nei)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_la)
```



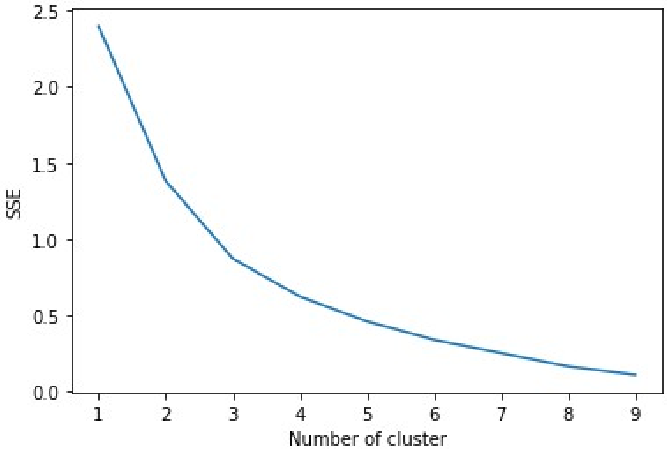
FourSquare API helped us search and collect all the popular Venues in LA Neighborhoods

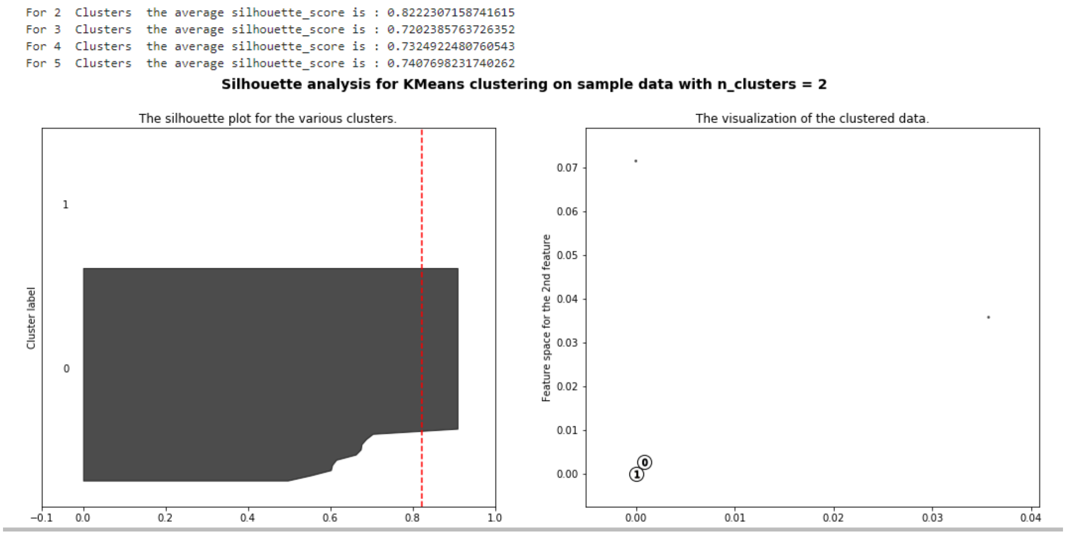


	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	Grand Park	[{'id': '4bf58dd8d48988d163941735', 'name': 'P...	34.055034	-118.245179
1	Redbird	[{'id': '4bf58dd8d48988d14e941735', 'name': 'A...	34.050666	-118.244068
2	Kinokuniya Bookstore	[{'id': '4bf58dd8d48988d114951735', 'name': 'B...	34.050145	-118.242246
3	JiST Cafe	[{'id': '4bf58dd8d48988d143941735', 'name': 'B...	34.050908	-118.240436
4	Blue Whale Bar	[{'id': '4bf58dd8d48988d1e7931735', 'name': 'J...	34.049884	-118.242114

Part 3. Methodology section.

Find optimum number of Clusters and Visualize them using K-Mean.



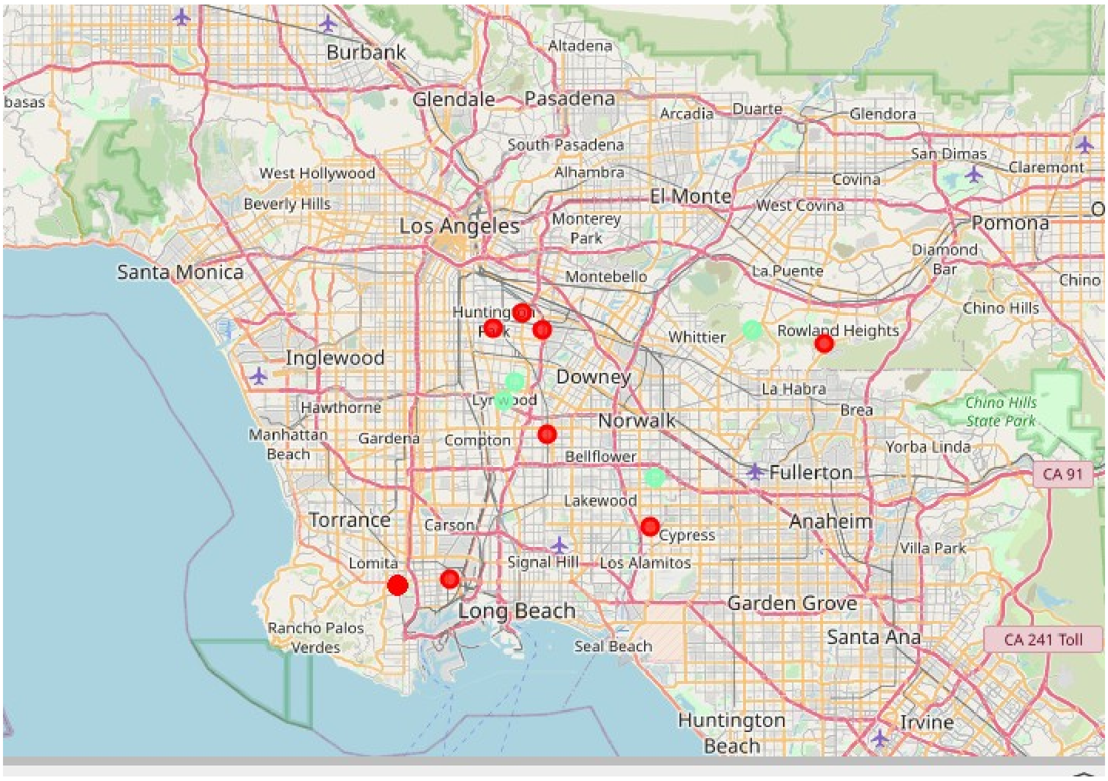


```
For 2  Clusters  the average silhouette_score is : 0.8222307158741615
For 3  Clusters  the average silhouette_score is : 0.7202385763726352
For 4  Clusters  the average silhouette_score is : 0.7324922480760543
For 5  Clusters  the average silhouette_score is : 0.7407698231740262
```

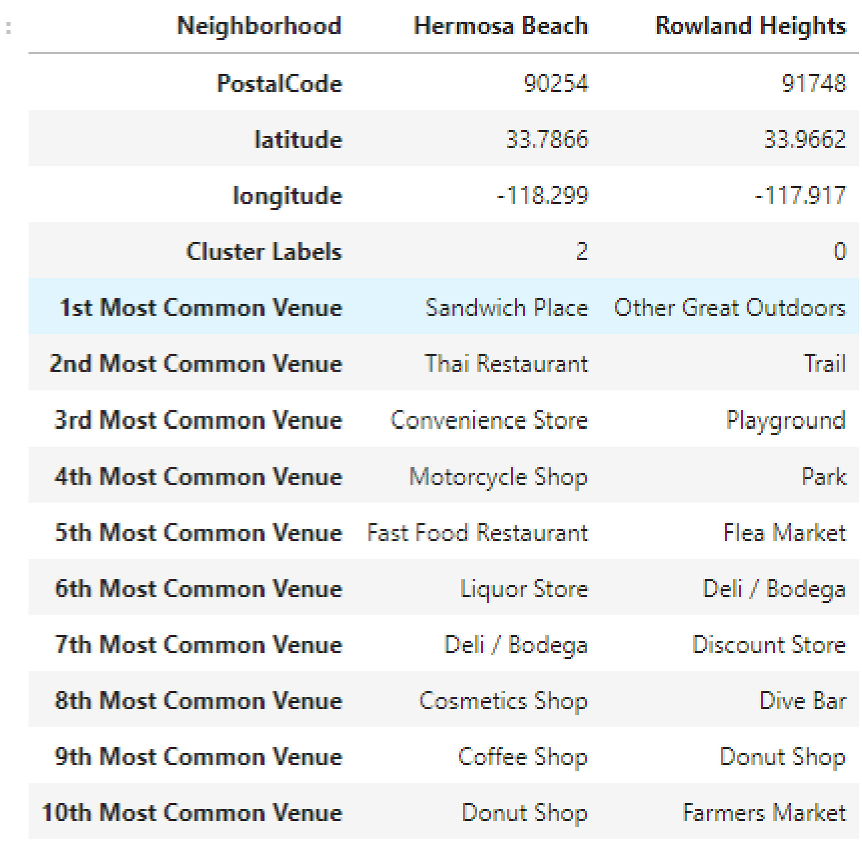
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2

The silhouette plot for the various clusters.

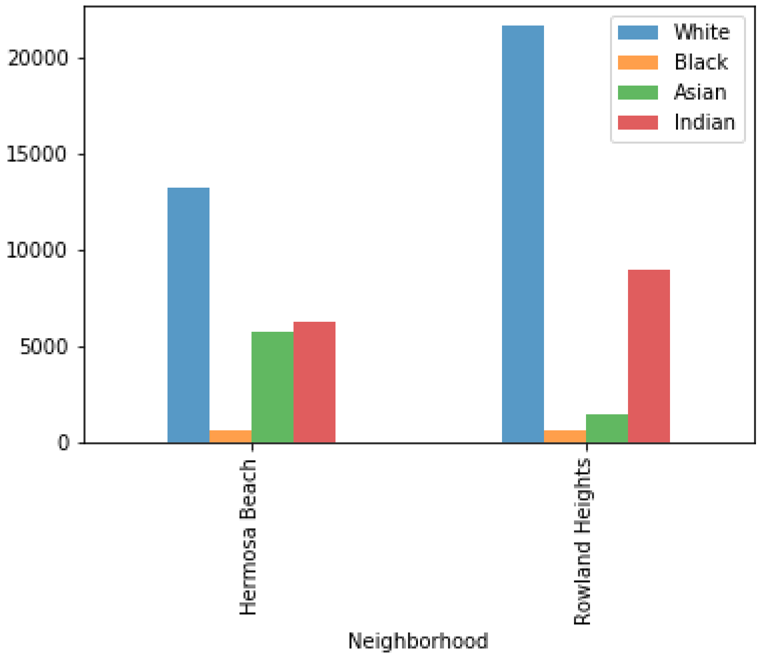
The visualization of the clustered data.

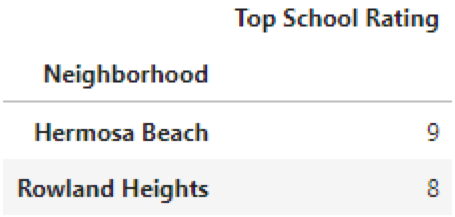


After getting Population, school rating and house pricing data analyze 2 neighborhoods.

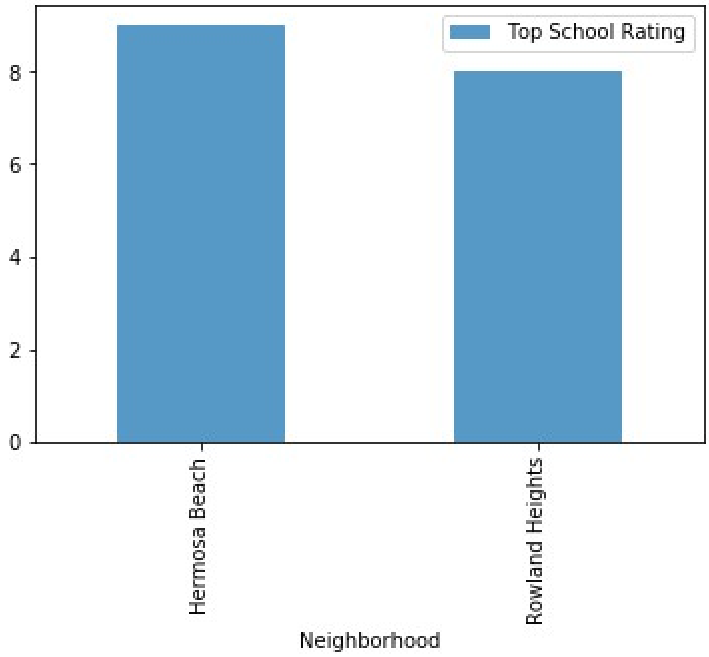


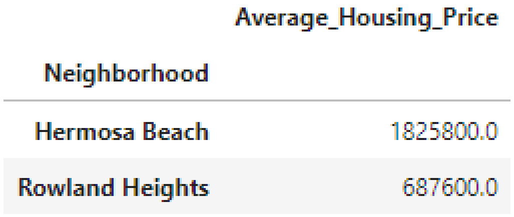
Neighborhood	Hermosa Beach	Rowland Heights
PostalCode	90254	91748
latitude	33.7866	33.9662
longitude	-118.299	-117.917
Cluster Labels	2	0
1st Most Common Venue	Sandwich Place	Other Great Outdoors
2nd Most Common Venue	Thai Restaurant	Trail
3rd Most Common Venue	Convenience Store	Playground
4th Most Common Venue	Motorcycle Shop	Park
5th Most Common Venue	Fast Food Restaurant	Flea Market
6th Most Common Venue	Liquor Store	Deli / Bodega
7th Most Common Venue	Deli / Bodega	Discount Store
8th Most Common Venue	Cosmetics Shop	Dive Bar
9th Most Common Venue	Coffee Shop	Donut Shop
10th Most Common Venue	Donut Shop	Farmers Market



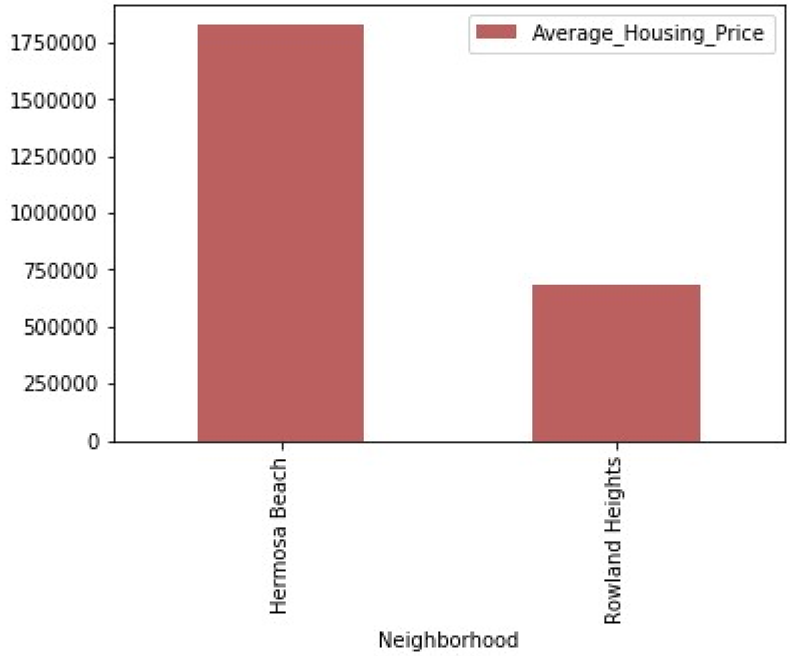


	Top School Rating
Neighborhood	
Hermosa Beach	9
Rowland Heights	8





	Average_Housing_Price
Neighborhood	
Hermosa Beach	1825800.0
Rowland Heights	687600.0



Part 4. Conclusions and discussions

This Analysis concludes that the two places of Los Angeles Hermosa Beach and Rowland Heights.

Both has great amenities and locality, but quite different: Hermosa is seashore, Rowland Heights in and below the Puente Hills in the San Gabriel Valley, near the National Parks.

Of course the have different housepricing: out of these two Rowland Heights has better prospects for buying houses or choose for rental houses. Rowland Heights has the higher number of Indian population, but school rating is good in both areas 8+. Top 10 common venues shows Hermosa Beach has got a good neighborhood with Restaurants, Convenience Store, Cosmetic Shop, Donot Shop and many more. But housing price there is very high.

If you have money - Hermosa is perfect place! But if you need more comfortable prices and Parks and Trails nearby Rowland Heights is for you.

So, we can see that our research helps user to compare two neighborhood and recommend options with facts.

For further Development I can suggest:

1. Get more data with different criteria for search such as:
    - criminal level
    - prices in common places
    - distance to other cities, national parks
    - unemployment level etc.
2. Use more machine learning for testing division to clusters.

Thank you for attention, to my project!