Opening a new Restaurant in Los Angeles

Harsh Diwakar

October 22, 2020

1. Introduction

1.1 Background

Whenever a Businessman is looking for an investment, opening a new Restaurant is a great Investment. Opening a restaurant could be a great investment as the businessman could open the Restaurant in his own hometown and stay near his home while managing his restaurant. Even when opening a new restaurant in own hometown, choosing a neighbourhood to open a restaurant is still an important decision. Restaurant sales, profit and type of restaurant may depend on the location of Neighbourhood.

1.2 Business Problem

The Objective of this Capstone Project is to analyse and select the best locations in the City of Los Angeles, California to open a new Restaurant. Using data science methodology and ML techniques like clustering, this project aims to provide the solution to answer the Business question: If a Businessman or a chain of restaurant is looking to open a new Restaurant in City of Los Angeles, California, Where would you recommend to open it?

1.3 Target Audience

A lot of stakeholders can be helped with this project such as:

i. Businessmen looking to open a new restaurant in city of Los Angeles, California

ii. Chain of Restaurants looking to open a new branch in city of Los Angeles, California.

iii. This analysis can also be helpful for the people of Los Angeles who are looking to dine in some good restaurants near their Neighbourhoods.

2. Data Acquisition

2.1 Required Dataset

To solve the Problem following data will be Required:

- List of Neighbourhoods in the city of Los Angeles.
- Latitude and Longitude of all Neighbourhoods.
- Venue Data, data related to different Venues in that Neighbourhood

2.2 Solving the Problem Using above data

For solving the problem we will begin with the Neighbourhoods data in the City of Los Angeles, then those neighbourhoods will be used to extract the Venus from foursquare location data after that the Venues data will be used be used to cluster the different Neighbourhoods in the city of Neighbourhoods.

2.3 Sources of Data

Below are the sources of data used in this project:

- List of Neighbourhoods in the city of Los Angeles- https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr
- Latitude and Longitude of all Neighbourhoods- https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr
- Venue Data, data related to different Venues in that Neighbourhood- https://foursquare.com

3. Data Cleaning and Feature Selection

3.1 Data Cleaning

Data for list of neighbourhoods with latitude and longitude is downloaded from the above-mentioned website. Data downloaded from website contains a lot of irrelevant information so first Name of Neighbourhoods with latitude and longitude were extracted. The same data was converted in Pandas dataframe so that it could be analysed easily. There are 272 Neighbourhoods in Los Angeles according to the dataset.

Data before cleaning:

```
[8]: neighborhoods_data[1]
```


```
[8]: {'type': 'Feature',
      'properties': {'external_i': 'adams-normandie',
       'name': 'Adams-Normandie',
       'location': 'POINT(34.031461499124156 -118.30020800000011)',
       'latitude': '-118.30020800000011',
       'slug_1': None,
       'sqmi': '0.805350187789',
       'display_na': 'Adams-Normandie L.A. County Neighborhood (Current)',
       'set': 'L.A. County Neighborhoods (Current)',
       'slug': 'adams-normandie',
       'longitude': '34.031461499124156',
       'name_1': None,
       'kind': 'L.A. County Neighborhood (Current)',
       'type': 'segment-of-a-city'},
      'geometry': {'type': 'MultiPolygon',
       'coordinates': [[[[-118.30900800000012, 34.03741099912408],
          [-118.30040800000013, 34.0373119991241],
          [-118.29150800000001, 34.03681199912407],
          [-118.29140800000012, 34.025511999124234],
          [-118.305408, 34.025711999124255],
          [-118.30900800000012, 34.025611999124216],
          [-118.30900800000012, 34.03741099912408]]]]}}
```


Fig.1: Uncleaned Data

Data After it cleaned and Neighbourhoods with latitude and longitude are extracted:

[14]:

	Neighborhood	Latitude	Longitude
0	Acton	34.538990	-118.202617
1	Adams-Normandie	34.037411	-118.309008
2	Agoura Hills	34.168203	-118.761925
3	Agua Dulce	34.558304	-118.254677
4	Alhambra	34.105040	-118.121747

Fig.2: Data after cleaning

3.2 Feature Selection

After the dataset in extracted, it is used to get venues near each neighbourhood using Foursquare API. Then all the venues that contained Restaurant were filtered and neighbourhoods that did not have any restaurants were dropped.

There are total of 78 Numbers of Restaurant Categories.

4. Methodology

4.1 Feature Extraction

For feature extraction One Hot Encoding is used in terms of categories. Each feature is a category that belongs to a venue. Each feature becomes binary that means if that category is found then it will be 1 otherwise 0. Then venues are grouped by neighbourhoods with mean of all the one hot coded category. This gives us venue for each row and each column contains the frequency of occurrence of each category.

Data after One Hot Encoding:

[78]:

	Neighborhood	African Restaurant	American Restaurant	Andhra Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Brazilian Restaurant	Burmese Restaurant	R
3	Adams-Normandie	0	0	0	0	0	0	0	0	
5	Adams-Normandie	0	0	0	0	0	0	0	0	
6	Adams-Normandie	0	0	0	0	0	0	0	0	
8	Adams-Normandie	0	0	0	0	0	0	0	0	
9	Adams-Normandie	0	0	0	0	0	0	0	0	

Fig.3 Data after One hot Encoding

Data after Feature Extraction:

[29]:

	Neighborhood	African Restaurant	American Restaurant	Andhra Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Brazilian Restaurant	Burmese Restaurant
0	Adams-Normandie	0.000000	0.000000	0.000000	0.000000	0.023256	0.000000	0.000000	0.000000
1	Agoura Hills	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Alhambra	0.000000	0.068966	0.000000	0.000000	0.068966	0.000000	0.000000	0.034483
3	Alondra Park	0.000000	0.041667	0.000000	0.000000	0.041667	0.000000	0.000000	0.000000
4	Arleta	0.000000	0.055556	0.000000	0.000000	0.222222	0.000000	0.000000	0.000000
5	Arlington Heights	0.000000	0.000000	0.000000	0.000000	0.020000	0.000000	0.040000	0.000000
6	Artesia	0.000000	0.114286	0.000000	0.000000	0.085714	0.000000	0.000000	0.000000
7	Athens	0.000000	0.090909	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	Atwater Village	0.000000	0.055556	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
9	Avalon	0.000000	0.200000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10	Avocado Heights	0.000000	0.000000	0.000000	0.000000	0.074074	0.000000	0.000000	0.000000
11	Baldwin Hills/Crenshaw	0.000000	0.058824	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
12	Baldwin Park	0.000000	0.090909	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
13	Bell	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
14	Bell Gardens	0.000000	0.083333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
15	Bellflower	0.000000	0.086957	0.000000	0.043478	0.000000	0.000000	0.000000	0.000000
16	Beverly Crest	0.000000	0.000000	0.000000	0.000000	0.142857	0.000000	0.000000	0.000000
17	Beverly Grove	0.000000	0.066667	0.000000	0.000000	0.033333	0.000000	0.033333	0.000000
18	Beverlywood	0.000000	0.085714	0.000000	0.000000	0.057143	0.000000	0.000000	0.000000
19	Boyle Heights	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
20	Bradbury	0.000000	0.250000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Fig.4 Data after feature extraction

4.2 Unsupervised Learning

To find the similarities between neighbourhoods, unsupervised learning is used. Unsupervised learning is used in the datasets with no labels. In this case we need to cluster the neighbourhoods so we will be using K-means Technique.

K-means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with pattern inside the multidimensional features.

K-means uses number of clusters as a parameter, so to find the best value of number of Clusters wen used silhouette score which gave us number of Cluster 3 for the best result.

Plot for Silhouette score with different numbers of Clusters:

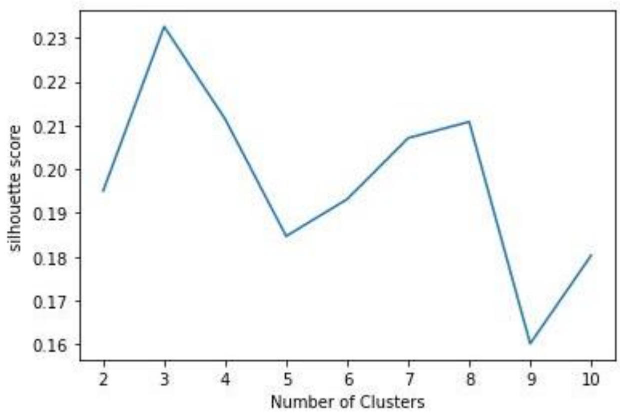

Fig.5: Silhouette Score against Number of Clusters

Data Visualization using Folium:

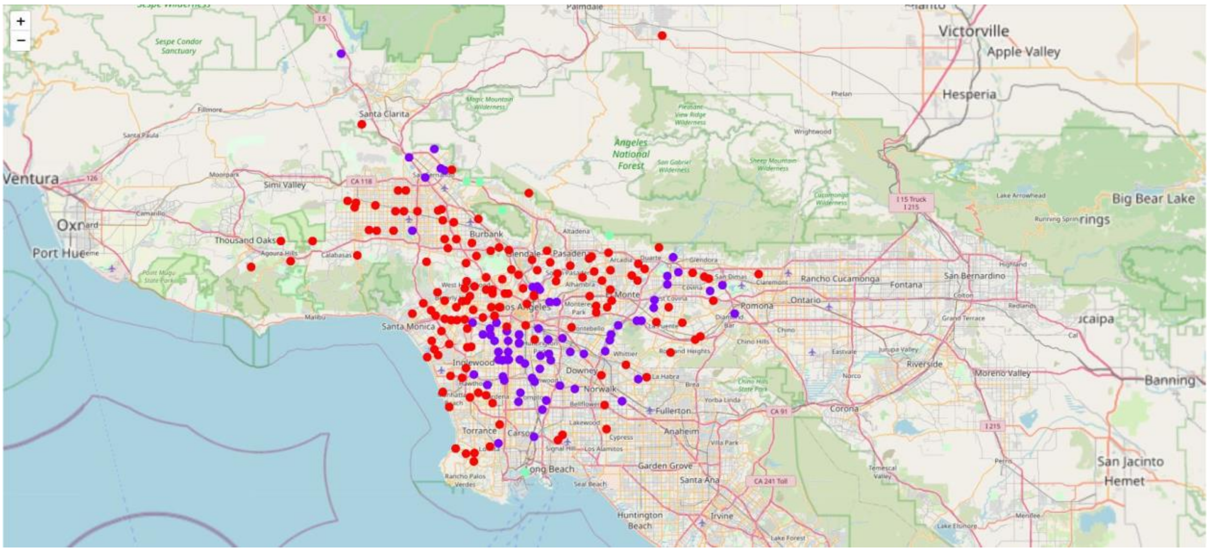

Fig.6: Folium plot of Clustered data

5. Results

After looking at clustered data we can see that neighbourhoods in cluster label 0 have a majority of Asian Restaurant (Red Colour in Folium Map):

[46]:

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Adams-Normandie	34.037411	-118.309008	0	Korean Restaurant	Mexican Restaurant	Fast Food Restaurant	Chinese Restaurant	Restaurant	Sushi Restaurant	Indian Restaurant	Italian Restaurant	Himalayan Restaurant	Vietnamese Restaurant
2	Agoura Hills	34.168203	-118.761925	0	Japanese Restaurant	Fast Food Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Italian Restaurant	Asian Restaurant	American Restaurant	Argentinian Restaurant	Filipino Restaurant	French Restaurant
4	Alhambra	34.105040	-118.121747	0	Chinese Restaurant	Fast Food Restaurant	Dumpling Restaurant	American Restaurant	Vietnamese Restaurant	Thai Restaurant	Asian Restaurant	Mexican Restaurant	Japanese Restaurant	Japanese Curry Restaurant
5	Alondra Park	33.897572	-118.326513	0	Fast Food Restaurant	Vietnamese Restaurant	Mexican Restaurant	Korean Restaurant	Sushi Restaurant	Japanese Restaurant	American Restaurant	Italian Restaurant	Ramen Restaurant	Asian Restaurant
6	Artesia	33.880383	-118.074895	0	Fast Food Restaurant	Chinese Restaurant	American Restaurant	Indian Restaurant	Asian Restaurant	Mexican Restaurant	Filipino Restaurant	Thai Restaurant	Sushi Restaurant	Vietnamese Restaurant
10	Arleta	34.224103	-118.422015	0	Thai Restaurant	Asian Restaurant	Fast Food Restaurant	Japanese Restaurant	American Restaurant	Restaurant	Seafood Restaurant	Sushi Restaurant	French Restaurant	English Restaurant
11	Arlington Heights	34.052611	-118.315909	0	Korean Restaurant	Japanese Restaurant	Mexican Restaurant	Vietnamese Restaurant	Fast Food Restaurant	Sushi Restaurant	Brazilian Restaurant	Seafood Restaurant	Asian Restaurant	Italian Restaurant
13	Atwater Village	34.153007	-118.278325	0	Middle Eastern Restaurant	Fast Food Restaurant	Mexican Restaurant	Chinese Restaurant	American Restaurant	Vietnamese Restaurant	Thai Restaurant	Filipino Restaurant	Sushi Restaurant	German Restaurant
14	Avalon	33.354873	-118.330431	0	Seafood Restaurant	Mexican Restaurant	American Restaurant	Caribbean Restaurant	New American Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant
15	Avocado Heights	34.025974	-117.965869	0	Mexican Restaurant	Fast Food Restaurant	Chinese Restaurant	Vietnamese Restaurant	Asian Restaurant	Sushi Restaurant	South American Restaurant	Seafood Restaurant	Yoshoku Restaurant	English Restaurant

Fig.7: Data in Cluster Label-0

Majority of data in cluster label 1 have a majority of Mexican or fast-food Restaurant (Blue Colour in folium map):

[47]:

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
12	Athens	33.930963	-118.291664	1	Fast Food Restaurant	Mexican Restaurant	American Restaurant	Seafood Restaurant	Gluten-free Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Filipino Restaurant	French Restaurant
17	Vermont-Slauson	33.993314	-118.280607	1	Fast Food Restaurant	Mexican Restaurant	Seafood Restaurant	Chinese Restaurant	American Restaurant	Latin American Restaurant	Ethiopian Restaurant	Falafel Restaurant	Filipino Restaurant	French Restaurant
18	Baldwin Hills/Crenshaw	34.024836	-118.356261	1	Fast Food Restaurant	Mexican Restaurant	Chinese Restaurant	American Restaurant	Seafood Restaurant	French Restaurant	Middle Eastern Restaurant	Mediterranean Restaurant	Southern / Soul Food Restaurant	Ethiopian Restaurant
19	Baldwin Park	34.107064	-117.942953	1	Fast Food Restaurant	Mexican Restaurant	American Restaurant	Thai Restaurant	Seafood Restaurant	Middle Eastern Restaurant	Yoshoku Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant
21	Bellflower	33.908971	-118.138574	1	Fast Food Restaurant	Mexican Restaurant	Japanese Restaurant	American Restaurant	Seafood Restaurant	Chinese Restaurant	Restaurant	Argentinian Restaurant	Thai Restaurant	New American Restaurant
22	Bell Gardens	33.973973	-118.148570	1	Mexican Restaurant	Fast Food Restaurant	Seafood Restaurant	American Restaurant	Spanish Restaurant	Salvadoran Restaurant	Chinese Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant
24	Bell	33.998113	-118.165504	1	Mexican Restaurant	Fast Food Restaurant	Latin American Restaurant	Seafood Restaurant	Chinese Restaurant	Mediterranean Restaurant	Restaurant	Korean Restaurant	Greek Restaurant	Gluten-free Restaurant
31	Boyle Heights	34.061811	-118.192705	1	Mexican Restaurant	Fast Food Restaurant	Seafood Restaurant	Restaurant	Chinese Restaurant	Mediterranean Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant
34	Broadway-Manchester	33.960014	-118.278207	1	Fast Food Restaurant	American Restaurant	Southern / Soul Food Restaurant	Yoshoku Restaurant	German Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Filipino Restaurant	French Restaurant
37	Carson	33.886268	-118.257605	1	Fast Food Restaurant	Mexican Restaurant	Chinese Restaurant	American Restaurant	Restaurant	Asian Restaurant	Italian Restaurant	Hawaiian Restaurant	Seafood Restaurant	Southern / Soul Food Restaurant
44	Cerritos	33.887397	-118.037829	1	Fast Food Restaurant	Asian Restaurant	Korean Restaurant	American Restaurant	Italian Restaurant	Sushi Restaurant	Mediterranean Restaurant	Mexican Restaurant	Yoshoku Restaurant	Gluten-free Restaurant

Fig.8: Data in Cluster label-1

Also it could be seen that very less data is there in cluster-2 and majority of this neighbourhood has American or Yoshuku Restaurant (green colour in folium map):

[48]:

```
#Examine Cluster
LA_merged.loc[LA_merged['Cluster Labels'] == 2]
```

[48]:

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
27	Burbank	34.221654	-118.292761	2	American Restaurant	Yoshoku Restaurant	Gluten-free Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	German Restaurant
98	Hansen Dam	34.273347	-118.368321	2	American Restaurant	Yoshoku Restaurant	Gluten-free Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	German Restaurant
206	San Pedro	33.761717	-118.243705	2	American Restaurant	Yoshoku Restaurant	Gluten-free Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	German Restaurant
209	Shadow Hills	34.272354	-118.340656	2	American Restaurant	Yoshoku Restaurant	Gluten-free Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	German Restaurant
211	Sierra Madre	34.179405	-118.065651	2	American Restaurant	Yoshoku Restaurant	Gluten-free Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	German Restaurant
299	Burbank	34.221654	-118.292761	2	American Restaurant	Yoshoku Restaurant	Gluten-free Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	German Restaurant
370	Hansen Dam	34.273347	-118.368321	2	American Restaurant	Yoshoku Restaurant	Gluten-free Restaurant	English Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	German Restaurant

Fig.9: Data in label-2

6. Discussion

By looking at the different cluster Businessman can now choose which neighbourhood to choose for opening a new restaurant depending on the type of the restaurant he/ she wants to open. Also, the same data can be used by the people of Los Angeles to explore the different restaurants around their neighbourhoods.

7. Conclusion

In this work, we have analysed the data of city of Los Angeles, California to open a new Restaurant by using the data science methodology and ML Techniques. This project successfully makes cluster for different restaurants in Los Angeles and gives an idea about the location where to open a new restaurant.

Additionally, This also provides the information of the new explorable types of restaurants in different Neighbourhoods that can be used by peoples to explore new restaurants in City of Los Angeles.