ANDERSON BALRAJ IBM DATA SCIENCE CAPSTONE: BATTLE OF THE NEIGHBORHOODS WEEK 2

NEIGHBORHOOD RESTAURANT CUISINE ANALYSIS OF LOS ANGELES

1. Introduction

Los Angeles (LA) is the largest city in California. LA has a population of approximately four (4) million people and ranks third in having the largest population in North America. There are many characteristics that define LA including the city having ethnic diversity, a large metropolis and a wide-ranging economy. As such it has one of the largest Gross Domestic Product (GDP) in the world that has reached 1 trillion USD.

Development of any business within LA will require an in-depth analysis of several factors. This report however looks specifically into the restaurant business and how data analysis can point to several key indicators in determining specifics of the business which can suggest a possible location and a recommended type of cuisine to maximize profit. The following are the queries that the report will answer to determine these:

1) Determining neighborhood clusters according to the most popular type of restaurants cuisine e.g. Italian, Chinese.

2) Determining current rating of the different types of restaurant cuisine

3) Analyze the tips for a restaurant cuisine to determine popularity

4) Determining current trending types of restaurants cuisine

These queries will have the effect of showing the relationship between a location and a preferred cuisine as LA is ethnically diverse as mentioned before. In addition, even changes to colloquial preferences as per current trends in demand for certain foods can be shown by the queries.

Business personnel making the decisions can get an idea how the accumulated data points to via a simplified map showing the clusters. The use of graphs will also give insight into which location and cuisine can be most profitable based on indicators of restaurant ratings, trending and tips analysis for restaurants. These will then be compiled into a conclusion that can surmise the best location and cuisine a business owner starting a restaurant in LA can choose.

A businessperson's interest in knowing this data will be due to him/her making an informed decision such that they gain the most amount of profit from their initial investment. By knowing the analysis of the existing data, they reduce the risk of a loss since location and cuisine of a restaurant are key decision factors that can determine how popular a restaurant is and the amount of revenue that can be drawn.

2. Data Description

2.1 Description of dataframe with Restaurant Name, Co-ordinates and Cuisines

The data will be obtained from Foursquare Places API for Los Angeles, California. In order to determine the most suitable neighborhoods to analyze within Los Angeles, data was obtained from "SOCR Data - Los Angeles City Neighborhoods Data" in terms of all the neighborhoods and the median income per neighborhood. A snippet of this data is shown in Table 2.1.1 and used to generate a dataframe in the notebook.

Table 2.1.1: This table shows part of the raw data to be used to select the requisite neighborhoods to be analyzed

LA Neighborhood	Median Income	Schools	Diversity	Population	Area	Longitude	Latitude
Bel-Air	208861	924	0.2	7928	6.6	-118.46356	34.096148
Beverly Crest	168104	0	0.1	10610	7.9	-118.42471	34.112107
Pacific Palisades	168008	879	0.1	23940	22.7	-118.54461	34.07586
Porter Ranch	121428	886	0.7	24923	5.7	-118.56048	34.27848
Brentwood	112927	882	0.2	31344	15.2	-118.473	34.052
Cheviot Hills	109980	0	0.3	5776	1.3	-118.40851	34.04061
Hollywood Hills West	108199	961	0.1	14860	4.9	-118.36791	34.10771

From this data the top ranked neighborhoods in terms of median income were selected and a dataframe created, a snippet of which is shown in Figure 2.1.1. This gives the co-ordinate data for each neighborhood in Los Angeles.

	Neighborhood	Latitude	Longitude
0	Bel-Air	34.096148	-118.416000
1	Beverly_Crest	34.112107	-118.448009
2	Pacific_Palisades	34.075860	-118.605726
3	Porter_Ranch	34.278482	-118.272689
4	Brentwood	34.052000	-118.334603

Figure 2.1.1: This shows a snippet of a dataframe that will be used in the Foursquare API queries

The co-ordinates for the latitude and longitude were used to generate a JSON file containing the information of venues inclusive of restaurant data for each neighborhood in the dataframe utilizing a Foursquare API URL. A snippet of the JSON file is shown in Figure 2.1.2.

```
{'meta': {'code': 200, 'requestId': '5ed4624abe61c9001b274a3a'},
 'response': {'venues': [{'id': '4c59b2584ed9ef3b5bded96d',
   'name': 'Wok Inn Restaurant',
   'location': {'address': '201 N Los Angeles St Ste 102',
    'crossStreet': 'btw Los Angeles & Main',
    'lat': 34.0543171973882,
    'lng': -118.24175235053629,
    'labeledLatLngs': [{'label': 'display',
      'lat': 34.0543171973882,
      'lng': -118.24175235053629}],
    'distance': 116,
    'postalCode': '90012',
    'cc': 'US',
    'city': 'Los Angeles',
    'state': 'CA',
    'country': 'United States',
    'formattedAddress': ['201 N Los Angeles St Ste 102 (btw Los Angeles & Main)',
     'Los Angeles, CA 90012',
     'United States']},
   'categories': [{'id': '4bf58dd8d48988d142941735',
     'name': 'Asian Restaurant',
     'pluralName': 'Asian Restaurants',
     'shortName': 'Asian',
     'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/asian_',
      'suffix': '.png'},
     'primary': True}],
   'referralId': 'v-1590977096',
   'hasPerk': False},
```


Figure 2.1.2: The figure shows a snippet of the JSON generated from the venues query

This will then be converted into a dataframe which is shown in Figure 2.1.3. As seen in this dataframe the categories are described via ID characters.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bel-Air	34.096148	-118.416000	Franklin Canyon Park	34.096533	-118.412046	Park
1	Bel-Air	34.096148	-118.416000	Coldwater Canyon Park	34.091264	-118.411829	Park
2	Bel-Air	34.096148	-118.416000	Secure Live Scan	34.098159	-118.410645	Health Food Store
3	Bel-Air	34.096148	-118.416000	Coldwater Canyon Park Soft Track	34.089688	-118.412130	Trail
4	Beverly_Crest	34.112107	-118.448009	The Real Towing Services Company Since 1985	34.111911	-118.444805	Business Service

Figure 2.1.3: The figure shows a snippet of the JSON data converted into a dataframe

The important components of this dataframe from Figure 2.1.3 is the generated venue names, co-ordinates i.e. latitude and longitude and the venue category that can be used after. Filtering will then be used in order to get the categories i.e. the cuisines required as mentioned above for the data analysis. This is shown in Figure 2.1.4 where each restaurant's cuisine is described in the dataframe under the 'categories' column. The data that shows cuisines can then be grouped to show the actual numbers of each for the respective neighborhoods this can give an idea of popularity of the cuisines and will be illustrated via graphs.

Figure 2.1.4 shows some of the restaurant categories that will be used in the analysis. This data will be used for clustering in the respective neighborhoods based on cuisine.

	Neighborhood	American Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant	...	Latin American Restaurant	Mediterranean Restaurant	Mexican Restaurant
0	Bel-Air	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
1	Beverly_Crest	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
2	Beverlywood	0.075000	0.000000	0.025	0.025000	0.000000	0.00	0.050000	0.00	0.00	...	0.025000	0.000000	0.025000
3	Brentwood	0.016393	0.000000	0.000	0.016393	0.032787	0.00	0.000000	0.00	0.00	...	0.032787	0.000000	0.049180
4	Century_City	0.011494	0.011494	0.000	0.000000	0.022989	0.00	0.011494	0.00	0.00	...	0.000000	0.000000	0.057471
5	Chatsworth	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
6	Cheviot_Hills	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
7	Encino	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
8	Granada_Hills	0.000000	0.086957	0.000	0.000000	0.000000	0.00	0.086957	0.00	0.00	...	0.000000	0.000000	0.000000
9	Hancock_Park	0.040816	0.000000	0.000	0.000000	0.000000	0.00	0.020408	0.00	0.00	...	0.000000	0.000000	0.000000
10	Hollywood_Hills_West	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
11	Pacific_Palisades	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
12	Playa_del_Rey	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000

Figure 2.1.4: The figure shows the filtered dataframe showing the cuisines for each restaurant

2.2 Description of JSON data to be used for ratings, trending and tips analysis

The next JSON was generated to get specific venue information including the ratings, trending and tips for the venue, this is shown in Figure 2.2.1. This data will be used to determine how popular a restaurant is in each location based on cuisine. This data will be used to generate graphs for analysis.

This data will be placed into a dataframe and cleaned. The issues with this data are some venues are missing for example rating's data and thus for analyses purposes will be removed from the dataset, this holds true for trending and tips analysis.


```
dict_keys(['id', 'name', 'contact', 'location', 'canonicalUrl', 'categories', 'verified', 'stats', 'url', 'price', 'hasMen
u', 'likes', 'dislike', 'ok', 'rating', 'ratingColor', 'ratingSignals', 'menu', 'allowMenuUrlEdit', 'beenHere', 'specials',
'photos', 'venuePage', 'reasons', 'description', 'page', 'hereNow', 'createdAt', 'tips', 'shortUrl', 'timeZone', 'listed',
'hours', 'popular', 'seasonalHours', 'defaultHours', 'pageUpdates', 'inbox', 'attributes', 'bestPhoto', 'colors'])
```

```
{'id': '49ebc74af964a5202b671fe3',
 'name': 'Yang Chow Restaurant',
 'contact': {'phone': '2136250811',
  'formattedPhone': '(213) 625-0811',
  'facebook': '109603625743324',
  'facebookUsername': 'YangChowRestaurant',
  'facebookName': 'Yang Chow Restaurant'},
 'location': {'address': '819 N Broadway',
  'crossStreet': 'btwn College & Alpine',
  'lat': 34.06292584487055,
  'lng': -118.2380586952736,
  'labeledLatLngs': [{'label': 'display',
    'lat': 34.06292584487055,
    'lng': -118.2380586952736},
   {'label': 'entrance', 'lat': 34.062969, 'lng': -118.238165}],
  'postalCode': '90012',
  'cc': 'US',
```


Figure 2.2.1: This figure shows a snippet of the JSON that will be used to get specifics about the venue including ratings, trending and tips

3. Methodology

3.1 Obtaining, Validating and Cleaning the Data

As stated previously a pandas dataframe was created to store the data obtained from the "SOCR Data - Los Angeles City Neighborhoods Data" data source. In order to validate if the dataframe had good data the co-ordinates of Los Angeles were obtained utilizing the 'Nominatim' function. A map of Los Angeles was then created, and the neighborhoods co-ordinates superimposed on the map via blue circle markers using 'folium'. This is shown in Figure 3.1.1, all blue markers were in the Los Angeles area in the map indicating that the data was accurate.

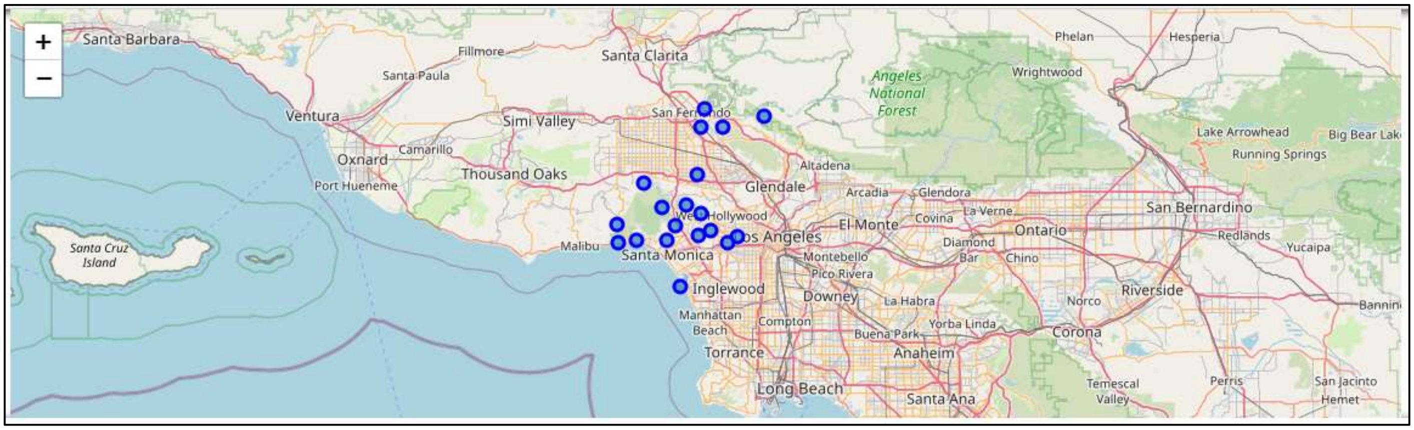

Figure 3.1.1: The figure shows the neighborhood data location superimposed on the Los Angeles map indicating the validity of the data

A specific venue function was then created to make the "GET" request that generates the JSON file. This function has the ability to read the created dataframe mentioned previously to generate the URL and then obtain the JSON. The function then extracts certain data from the JSON inclusive of Venue, Venue Co-ordinates and Venue Category and builds a new dataframe containing this information. Figure 3.1.2 is a snippet of the summary of the output counts for each category pulled.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bel-Air	5	5	5	5	5	5
Beverly_Crest	3	3	3	3	3	3
Beverlywood	43	43	43	43	43	43
Brentwood	61	61	61	61	61	61
Century_City	84	84	84	84	84	84
Chatsworth	6	6	6	6	6	6
Cheviot_Hills	6	6	6	6	6	6
Encino	2	2	2	2	2	2
Granada_Hills	23	23	23	23	23	23
Hancock_Park	53	53	53	53	53	53

Figure 3.1.2: The figure shows a summary of the count data generated via the venue function

The one hot encoding technique was applied to then applied in the ‘Venue Category’ column in order to create a new data frame where the previous ‘Venue Category’ strings are now the column names which were then grouped. Figure 3.1.3 shows the resulting dataframe, however the dataframe that is required is one that contains only Restaurant i.e. cuisine data. Also note the cells contains the frequency of each Venue Category.

	Neighborhood	Women's Store	Accessories Store	Airport	Airport Terminal	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	...	Temple	Thai Restaurant	Theater	Tourist Information Center	Toy / Game Store
0	Bel-Air	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
1	Beverly_Crest	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
2	Beverlywood	0.000000	0.00000	0.000000	0.000000	0.075000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.025000	0.000000	0.000000
3	Brentwood	0.000000	0.00000	0.000000	0.016393	0.016393	0.016393	0.032787	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
4	Century_City	0.011494	0.00000	0.000000	0.000000	0.011494	0.000000	0.011494	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.011494
5	Chatsworth	0.000000	0.00000	0.142857	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.142857	0.000000	0.000000	0.000000
6	Cheviot_Hills	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000

Figure 3.1.2: The figure shows the results of the one hot encoding technique (compare to Figure 2.1.3)

Utilizing ‘regex’ a filter was then applied to get only the ‘Restaurant’ data required for the analysis. The final ‘cleaned’ dataframe that will be used for the analysis consist of the ‘Neighborhood’ column which we can utilize to query in the ‘Neighborhood co-ordinates’ dataset to obtain the respective locations and then the ‘Restaurant Categories’ columns which indicate the frequency of visits to the respective Neighborhoods. This dataframe is now ready for further data analysis and a snippet of the dataframe is shown in Figure 3.1.3.

	Neighborhood	American Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant	...	Latin American Restaurant	Mediterranean Restaurant	Mexican Restaurant
0	Bel-Air	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
1	Beverly_Crest	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
2	Beverlywood	0.075000	0.000000	0.025	0.025000	0.000000	0.00	0.050000	0.00	0.00	...	0.025000	0.000000	0.025000
3	Brentwood	0.016393	0.000000	0.000	0.016393	0.032787	0.00	0.000000	0.00	0.00	...	0.032787	0.000000	0.049180
4	Century_City	0.011494	0.011494	0.000	0.000000	0.022989	0.00	0.011494	0.00	0.00	...	0.000000	0.000000	0.057471
5	Chatsworth	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
6	Cheviot_Hills	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
7	Encino	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
8	Granada_Hills	0.000000	0.086957	0.000	0.000000	0.000000	0.00	0.086957	0.00	0.00	...	0.000000	0.000000	0.000000
9	Hancock_Park	0.040816	0.000000	0.000	0.000000	0.000000	0.00	0.020408	0.00	0.00	...	0.000000	0.000000	0.000000
10	Hollywood_Hills_West	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
11	Pacific_Palisades	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
12	Playa_del_Rey	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
13	Rancho_Park	0.026316	0.000000	0.000	0.000000	0.000000	0.00	0.026316	0.00	0.00	...	0.000000	0.026316	0.000000
14	Shadow_Hills	0.000000	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.000000
15	West_Hills	0.073171	0.000000	0.000	0.000000	0.000000	0.00	0.000000	0.00	0.00	...	0.000000	0.000000	0.048780
16	West_Los_Angeles	0.060000	0.000000	0.000	0.000000	0.010000	0.01	0.000000	0.01	0.01	...	0.000000	0.020000	0.010000

Figure 3.1.2: The figure shows the dataframe that will be mainly used for the data analysis

3.2 Data Analysis of Cleaned Dataframe

The cleaned dataframe is now ready to be used using data analysis methodologies. The first exploratory analysis is looking at the top 5 restaurants per neighborhood according to the frequencies in the cleaned dataframe. This is to get an initial idea of how the data will look like before further processing. Each neighborhood has the respective restaurants sorted from the largest frequencies using the 'sort_values' function and then printing the top 5 for each neighborhood. Figure 3.2.1 shows a snippet of the results obtained for Century City, Los Angeles.

```
----Century_City----
                               venue  freq
0                 Mexican Restaurant  0.06
1                 Italian Restaurant  0.05
2                 Chinese Restaurant  0.02
3                American Restaurant  0.01
4  Southern / Soul Food Restaurant  0.01
```

Figure 3.2.1: The figure shows the results from the exploratory data analysis on the top 5 Restaurants/cuisines for Century City, Los Angeles.

The data is looking valid thus far therefore the next step is to obtain a more substantial dataframe utilizing similar methods in the exploratory analysis. A function 'return_most_common_venues' was created to determine the most common restaurants and basically sort the venues for each neighborhood via frequency. The data range created expands to the top 10 restaurants/cuisines in contrast to the initial analysis and utilizing the sorting function a new dataframe is created. The dataframe consists of the 'Neighborhood' column and then columns illustrating a ranking of 1 to 10 of the most common restaurants as per the respective neighborhood. Figure 3.2.2 shows a snippet of the new dataframe.

	Neighborhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
0	Bel-Air	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
1	Beverly_Crest	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
2	Beverlywood	American Restaurant	Fast Food Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Restaurant	Mexican Restaurant	Latin American Restaurant	Italian Restaurant	Asian Restaurant	Chinese Restaurant
3	Brentwood	Mexican Restaurant	Chinese Restaurant	Latin American Restaurant	Vegetarian / Vegan Restaurant	Caribbean Restaurant	Kosher Restaurant	Korean Restaurant	American Restaurant	New American Restaurant	Mediterranean Restaurant
4	Century_City	Mexican Restaurant	Italian Restaurant	Chinese Restaurant	Korean Restaurant	Asian Restaurant	Fast Food Restaurant	Japanese Restaurant	American Restaurant	Restaurant	Southern / Soul Food Restaurant

Figure 3.2.2: The figure shows a snippet of the dataframe showing the top ten restaurants per neighborhood

3.3 K-means Clustering Analysis

Another part of data analysis undertaken is the K-means clustering utilizing the dataframe from Figure 3.2.2. The reason this analysis is undertaken is to determine which neighborhoods are similar type in terms of cuisine preference based on each top ten restaurants/cuisines. A standard cluster of 5 is chosen to give appropriate variety.

The 'KMeans' function is utilized in order to generate 'cluster labels' for each row of the dataframe based on the clustering analysis the function performed. Once this is completed three major data aspects are merged into one dataframe which include the Neighborhoods and co-ordinates, the cluster labels and finally the top ten most common restaurants/cuisines per neighborhood. Figure 3.3.1 illustrates the final dataframe.

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
0	Bel-Air	34.096148	-118.416000	0	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
1	Beverly_Crest	34.112107	-118.448009	0	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
2	Pacific_Palisades	34.075860	-118.605726	0	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
4	Brentwood	34.052000	-118.334603	0	Mexican Restaurant	Chinese Restaurant	Latin American Restaurant	Vegetarian / Vegan Restaurant	Caribbean Restaurant	Kosher Restaurant	Korean Restaurant	American Restaurant	New American Restaurant	Mediterranean Restaurant
5	Cheviot_Hills	34.040610	-118.602000	0	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant

Figure 3.3.1: The figure shows a snippet of the final dataframe showing the cluster label assigned to each neighborhood.

The clusters were then placed on the map of Los Angeles to illustrate which neighborhoods are similar in terms of preference in cuisine. The function was setup in such a way each cluster will generate a different color on the map using 'folium'. The results of the clustering for each of the five clusters was broken out from the dataframe to show the neighborhoods that was grouped together in a cluster utilizing the 'cluster label' as a grouping index.

3.4 Foursquare API Ratings and Tips Analysis

In order to query the Foursquare API for Ratings and Tips analysis (people who have marked a venue as a 'to do') the venue ID for each Restaurant needs to be obtained. To do this a function is created 'getNearbyVenuesID' to obtain from a JSON file the required data. A dataframe is then created that stores the Venue Name, ID and Category. Once this dataframe is obtained it is filtered according to the cuisines listed in Table 4.2.1. The most popular restaurants/cuisines were chosen for the sample based on the clustering information.

The results of this filtering are the Venue ID example shown in Figure 3.4.1. A second URL is created to query Foursquare API based on the Venue ID to obtain both the rating and tips for each restaurant since they are located in different parts of the JSON.

A sample of the Venue IDs for each cuisine is taken i.e. between 2-5 samples in order to conduct exploratory analysis of each cuisine ratings and the number of tips. An average of the ratings and tip count was calculated to compare the cuisines based on the samples.

The results of the ratings and tip count is shown on two horizontal bar graphs respectively utilizing the 'matplotlib' library to compare the results. In addition to this an area graph was generated to show the popularity of the cuisines in a respective Neighborhood in terms of highest frequencies.

	Venue Name	Venue ID	Venue Category
19	El Compita	4aeb7b84f964a52088c221e3	Mexican Restaurant
27	Chipotle Mexican Grill	524b20ba11d298598ce1c917	Mexican Restaurant
72	Zapoteca	4c3a43421e06d13a410e7b3e	Mexican Restaurant
101	Alta California Fonda	5327ba6311d2f453c9f4f223	Mexican Restaurant
138	Frida Taqueria	4b0082e3f964a5200d3f22e3	Mexican Restaurant
143	Kayndaves	4adb5719f964a520272621e3	Mexican Restaurant
168	Javier's	59f125e69e3b655e73f48d07	Mexican Restaurant
171	Tocaya Organica - Century City	5a2ecd0c345cbe4a6ac3ab88	Mexican Restaurant
174	Chipotle Mexican Grill	51cdef4d498ed4304a79b15c	Mexican Restaurant
235	Habanero Grill	4ace5e7cf964a52023d020e3	Mexican Restaurant
249	Taco Limón	4de81ccbe4cd157353eec669	Mexican Restaurant
258	Sharky's Woodfired Mexican Grill	4b46c026f964a520c42726e3	Mexican Restaurant
307	Frida Mexican Cuisine	4a42d464f964a52048a61fe3	Mexican Restaurant

Figure 3.4.1: The figure shows the Venue IDs obtained for Mexican Restaurants

3.5 Determining Trending Venues

In order to pull the trending data, a request was made Foursquare API for the current JSON file and then the venue information can be extracted. A trending values dataframe is created from the normalized JSON containing the venue and trending data. The trending data was then placed on the map of Los Angeles using a blue circle marker. There was also provision placed in the event of there being no trending data that this is fed back to the user.

4. Results

4.1 Determining neighborhood clusters according to the most popular type of restaurants cuisine

The neighborhood clusters were illustrated in the map of Los Angeles shown in Figure 4.1.1. As seen from the map clustering occurs in similar locations for clusters that have more than one neighborhood.

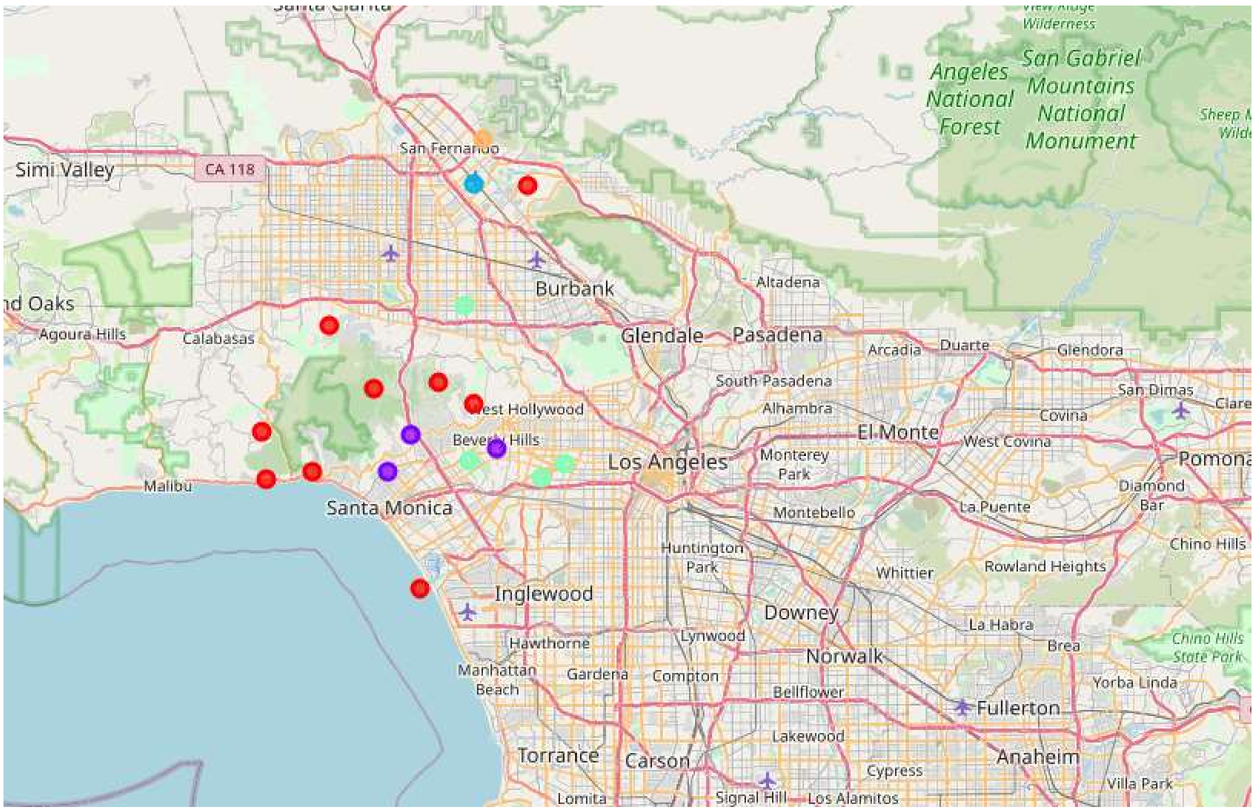

Figure 4.1.1: The figure shows the allocation of neighborhoods to the respective clusters according to colour

4.1.2 Cluster 1

Figure 4.1.2 shows Cluster 1 and the neighborhoods that have a similar preference in restaurants and cuisines. As seen from these results these neighborhoods have Vegetarian and Japanese as their most popular cusines.

	Neighborhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
0	Bel-Air	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
1	Beverly_Crest	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
2	Pacific_Palisades	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
4	Brentwood	Mexican Restaurant	Chinese Restaurant	Latin American Restaurant	Vegetarian / Vegan Restaurant	Caribbean Restaurant	Kosher Restaurant	Korean Restaurant	American Restaurant	New American Restaurant	Mediterranean Restaurant
5	Cheviot_Hills	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
6	Hollywood_Hills_West	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
10	Playa_del_Rey	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
16	Shadow_Hills	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
17	Encino	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant	Greek Restaurant
18	Rancho_Park	Vegetarian / Vegan Restaurant	New American Restaurant	Fast Food Restaurant	Japanese Restaurant	Mediterranean Restaurant	American Restaurant	Seafood Restaurant	Restaurant	French Restaurant	Asian Restaurant

Figure 4.1.2: The figure shows the neighborhoods belonging to Cluster 1

4.1.3 Cluster 2

Figure 4.1.3 shows Cluster 2 and the neighborhoods that have a similar preference in restaurants and cuisines. As seen from these results there is only one neighborhood in this cluster.

	Neighborhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
11	Woodland_Hills	Asian Restaurant	Chinese Restaurant	Mexican Restaurant	Thai Restaurant	Vegetarian / Vegan Restaurant	Italian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Falafel Restaurant	Fast Food Restaurant

Figure 4.1.3: The figure shows the neighborhood belonging to Cluster 2

4.1.4 Cluster 3

Figure 4.1.4 shows Cluster 3 and the neighborhoods that have a similar preference in restaurants and cuisines. As seen from these results there is only one neighborhood in this cluster.

	Neighborhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
14	Chatsworth	Thai Restaurant	Vegetarian / Vegan Restaurant	Japanese Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant	French Restaurant

Figure 4.1.4: The figure shows the neighborhood belonging to Cluster 3

4.1.5 Cluster 4

Figure 4.1.5 shows Cluster 4 and the neighborhoods that have a similar preference in restaurants and cuisines. As seen from these results there is only one neighborhood in this cluster.

	Neighborhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
15	Granada_Hills	Asian Restaurant	Fast Food Restaurant	Japanese Restaurant	Restaurant	Vegetarian / Vegan Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	French Restaurant

Figure 4.1.5: The figure shows the neighborhood belonging to Cluster 4

4.1.6 Cluster 5

Figure 4.1.6 shows Cluster 5 and the neighborhoods that have a similar preference in restaurants and cuisines. As seen from these results this cluster prefers American and Italian type cuisines.

	Neighborhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
7	Beverlywood	American Restaurant	Fast Food Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Restaurant	Mexican Restaurant	Latin American Restaurant	Italian Restaurant	Asian Restaurant	Chinese Restaurant
8	West_Hills	American Restaurant	Italian Restaurant	Mexican Restaurant	Sushi Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant	Fast Food Restaurant
9	Century_City	Mexican Restaurant	Italian Restaurant	Chinese Restaurant	Korean Restaurant	Asian Restaurant	Fast Food Restaurant	Japanese Restaurant	American Restaurant	Restaurant	Southern / Soul Food Restaurant
12	West_Los_Angeles	American Restaurant	Sushi Restaurant	Restaurant	Mediterranean Restaurant	Italian Restaurant	New American Restaurant	Mexican Restaurant	Chinese Restaurant	Falafel Restaurant	Japanese Restaurant
13	Hancock_Park	Italian Restaurant	American Restaurant	Restaurant	New American Restaurant	Fast Food Restaurant	Asian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant	Falafel Restaurant

Figure 4.1.6: The figure shows the neighborhood belonging to Cluster 5

4.1.7 Most frequented Restaurant Cuisine

Figure 4.1.7 shows the most frequented restaurant in the area graph. Note this graph is not meant to show individual data but to give an idea of popularity of overall cuisine. As seen American Restaurant cuisine is most popular.

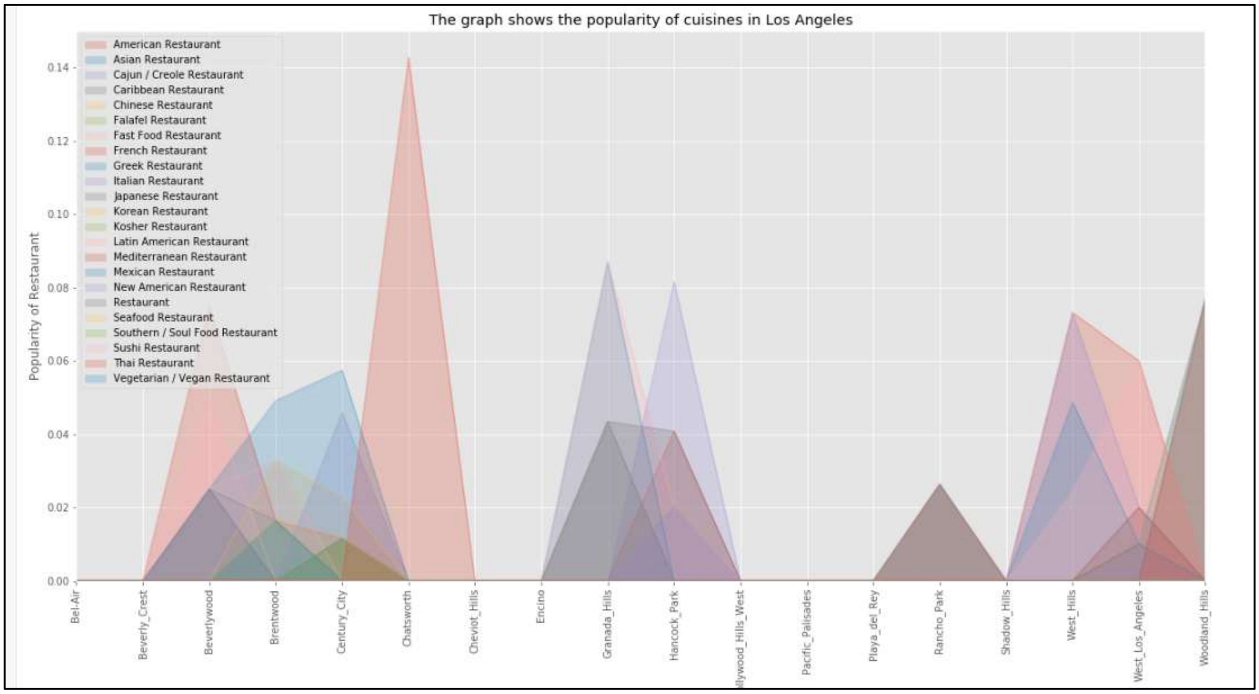

Figure 4.1.7: The figure shows the most popular cuisine in Los Angeles

4.2 Determining restaurant/cuisine ratings and tip counts

A summary of results is shown in Table 4.2.1 for the average sample ratings and average tip counts.

Table 4.2.1: The table shows the cuisine and average rating and tip count

Cuisine	Rating	Count
Vegan	8.50	7.5
American	8.14	42.0
Japanese	7.95	37.3
Mexican	7.55	5.3
Asian	7.43	53.7
Thai	6.65	5.0
Italian	7.35	12.0

These were then transferred to bar graphs for ease of view Figure 4.2.1 shows the Ratings graphs for the top restaurant cuisines. As seen the Vegetarian and American cuisines get the highest ratings whilst Asian, Thai and Italian cuisines get the lowest.

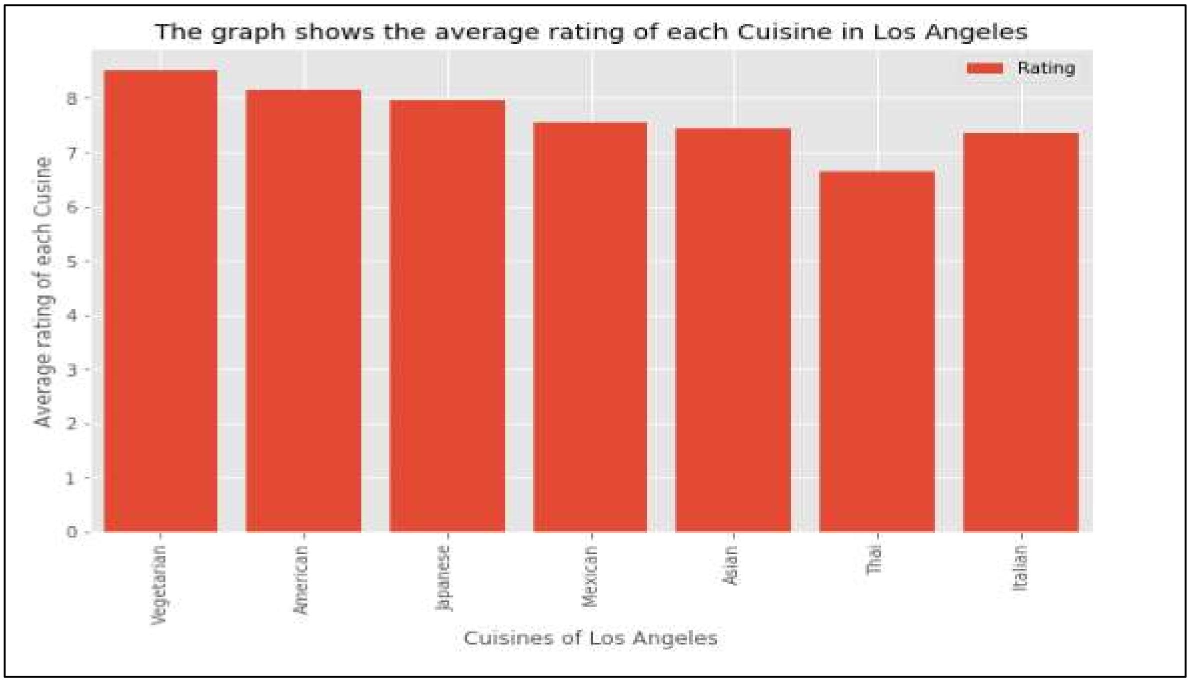

Figure 4.2.1: The table shows the average rating per a cuisine in Los Angeles

Figure 4.2.1 shows the Tips Count graphs for the top restaurant cuisines. As seen the Asian and American tip counts are the highest whilst the Mexican and Thai are the lowest.

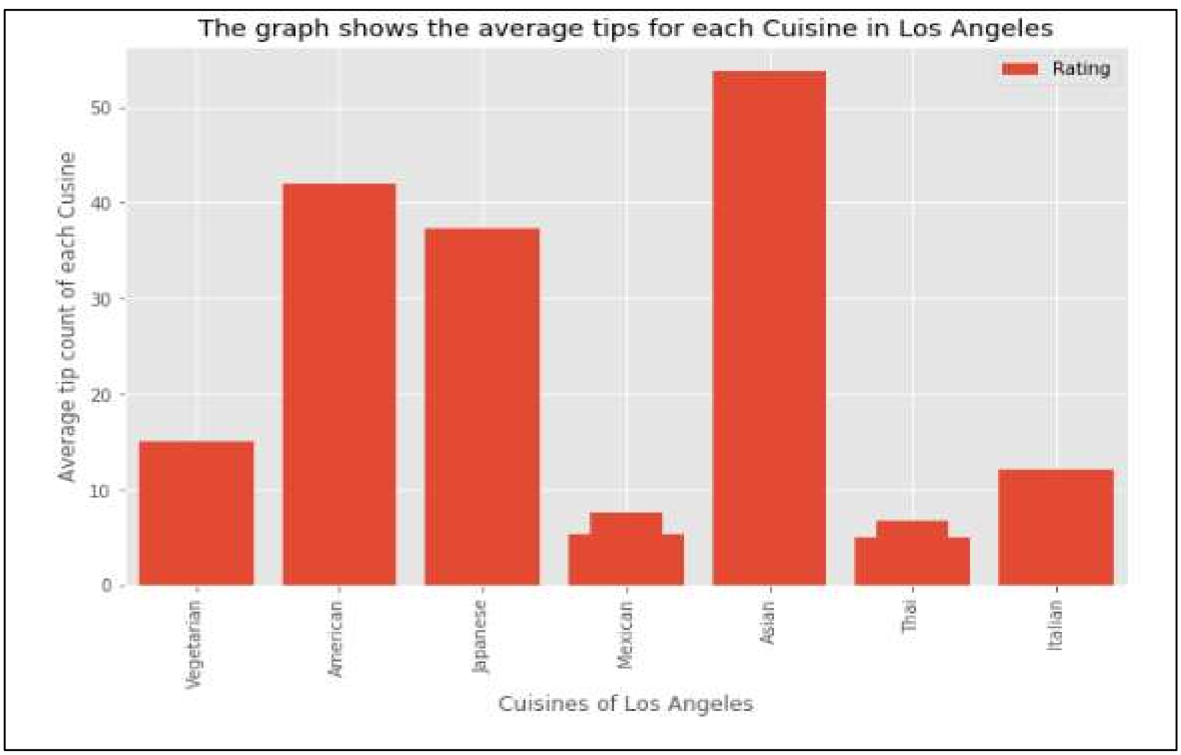

Figure 4.2.2: The table shows the average tip count per a cuisine in Los Angeles

4.3 Determining the trending venues

Figure 4.3.1 shows the only trending venue in Los Angeles called 'Redbird' an American Restaurant. Figure 4.3.2 shows the location of 'Redbird'.

	name	categories	location.distance	location.city	location.postalCode	location.state	location.country	location.lat	location.lng
0	Redbird	American Restaurant	357	Los Angeles	90012	CA	United States	34.050666	-118.244068

Figure 4.3.1: The figure shows the trending venue in Los Angeles

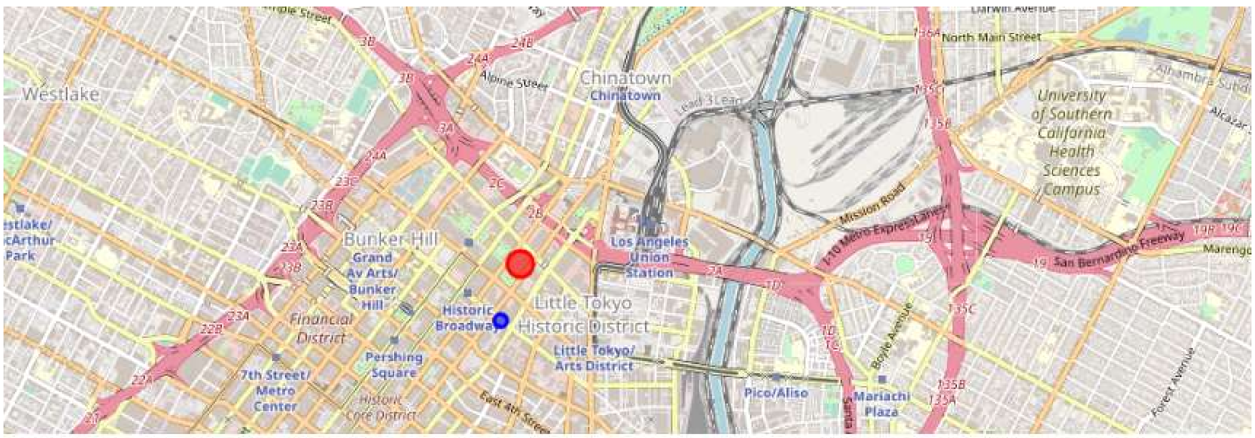

Figure 4.3.2: The figure shows the location of the trending venue in Los Angeles

5. Discussion

In terms of the observation of the results one of the biggest keys in the analysis is the clustering of the neighborhoods according to restaurant cuisine popularity. Cluster 1 shows that Vegetarian and Japanese are the most popular cuisines however the availability of these cuisines are limited as there weren't as many hits from Foursquare API. So, although these cuisines are popular there are few places in which customers can access them which makes it ideal to open more restaurants of this type i.e. Vegetarian and Japanese. If you also look at the map for Cluster 1 these neighborhoods are also close together, with the exception of neighborhoods Playa del Rey and Shadow Hills, so that there is a larger consumer base for a new restaurant that is opened in any neighborhood in Cluster 1.

Clusters 2, 3 and 4 have one neighborhood per cluster and thus any new restaurant located in these neighborhoods will have a smaller consumer base however they wouldn't be competing with a lot of other venues.

Cluster 5 neighborhoods are popular with American and Italian type cuisines however these places are in abundance and are some of the most frequented places according to the data in terms of popularity. Any new restaurant located in a neighborhood in Cluster 5 will face a lot of competition no matter the cuisine.

Similarly, as per the cluster analysis the ratings analysis on the sample set shows that Vegetarian, American and Japanese cuisines have the highest ratings. This infers that customers have better preference for these cuisines. The data is slightly skewed for the Vegetarian cuisine however due to the limited number of venues.

In terms of tip count analysis Asian, American and Japanese have the highest counts. This is expected for American and Japanese but not for the Asian cuisine since it has limited hits in the clustering. This suggests that the tip count is generated for the Asian cuisine in one neighborhood in Cluster 4 and new Asian restaurant located in a Cluster 4 neighborhood wouldn't have as large of a consumer base. The tip count data is slightly skewed for the Vegetarian cuisine however due to the limited number of venues.

Finally, the trending data did not show much presumably due to current worldwide issues where people movement is limited and doesn't give a clear indication on what is the current trending cuisine.

5. Conclusion

In conclusion the data points to opening a Vegetarian or Japanese type restaurant in any of the following neighborhoods:

- Bel Air
- Beverley Crest
- Pacific Palisades
- Brentwood
- Cheviot Hills
- Hollywood Hills West
- Encino
- Rancho Park

The reasons being for this include:

- There are limited venues for both Vegetarian and Japanese cuisines meaning less competition amongst venues
- These places are near each other resulting in a larger consumer base
- Vegetarian and Japanese venues currently have good rating and good tip counts
- Cluster 5 neighborhoods are not recommended due to the immense popularity and abundance of American cuisine
- Cluster 2, 3 and 4 is not recommended due to the lower consumer base as these have single neighborhoods in each