DATA SCIENCE PROJECT ON FINDING RESTAURANTS IN LOS ANGELES

1. Introduction:

- ❖ Business Problem

With an estimated population of nearly four million people, Los Angeles is the second-most populous city in the United States. It has a diverse economy and hosts businesses in a broad range of professional and cultural fields. It is also arguably the most amazing place to eat in America, owing to an incredible variety of international cuisines and some of the most talented chefs in the world. LA's great seasonal produce and access to ingredients makes it an ideal place for restaurants to thrive — but how do you know which ones to go to? How do you know where to set up your restaurant? As daunting this may sound, it is possible to know what the best places to get something to eat are with Foursquare.

- ❖ Target Audience

i. Entrepreneurs seeking to open a restaurant in Los Angeles and would like to map the competition in order to choose the best location.

ii. People seeking to find the best restaurant to go to based on Foursquare likes, restaurant category and geographic location data for restaurants in Los Angeles.

2. Data Description

I will be use the Foursquare API to pull the following location data on restaurants in Los Angeles, California:

- Venue Name
- Venue ID
- Venue Location
- Venue Category
- Count of Likes

To acquire the aforementioned data, I will need to do the following:

- Get the latitude and longitude coordinates for Los Angeles from the Geocoder library
- Use Foursquare API to get a list of all venues in Los Angeles

I will then take the gathered data and create a k-means clustering algorithm that groups restaurants into 4-5 clusters so that people looking to start a restaurant or eat in Los Angeles can easily see which restaurants are the best to eat at and what cuisine is available.

3. Methodology

I utilized the Foursquare API to explore the venues. I designed the limit as **100 venue** and the radius **500 meter** from 34.0536909N, 118.2427666W (Los Angeles). Here is a head of the list Venues name, category, latitude and longitude informations from Forsquare API:

Out[7]:

	venue.name	venue.id	venue.categories	venue.location.lat	venue.location.lng
0	Grand Park	4fecf601067d351381ea64fa	Park	34.055034	-118.245179
1	Badmaash	518471e6498e1c0b5f1401f9	Indian Restaurant	34.051342	-118.244571
2	Redbird	54938133498ed65f02e8c4ba	American Restaurant	34.050666	-118.244068
3	Kinokuniya Bookstore	4a8e024bf964a520ba1120e3	Bookstore	34.050145	-118.242246
4	JiST Cafe	51dccd46498e4f9ac4865270	Breakfast Spot	34.050908	-118.240436

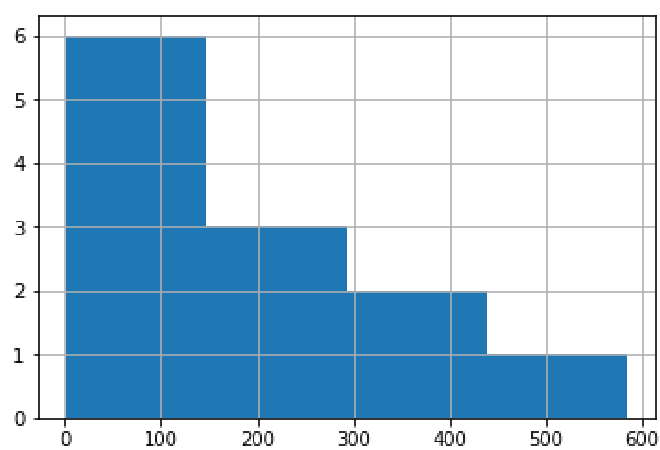
I used the unique () function to get a list of unique categories from the API in order to see what may or may not fit for restaurants. I removed the venues that are not restaurants and obtained the dataframe of restaurants only. Here is a head of the list Venues name, category, latitude and longitude informations from Forsquare API:

Out[11]:

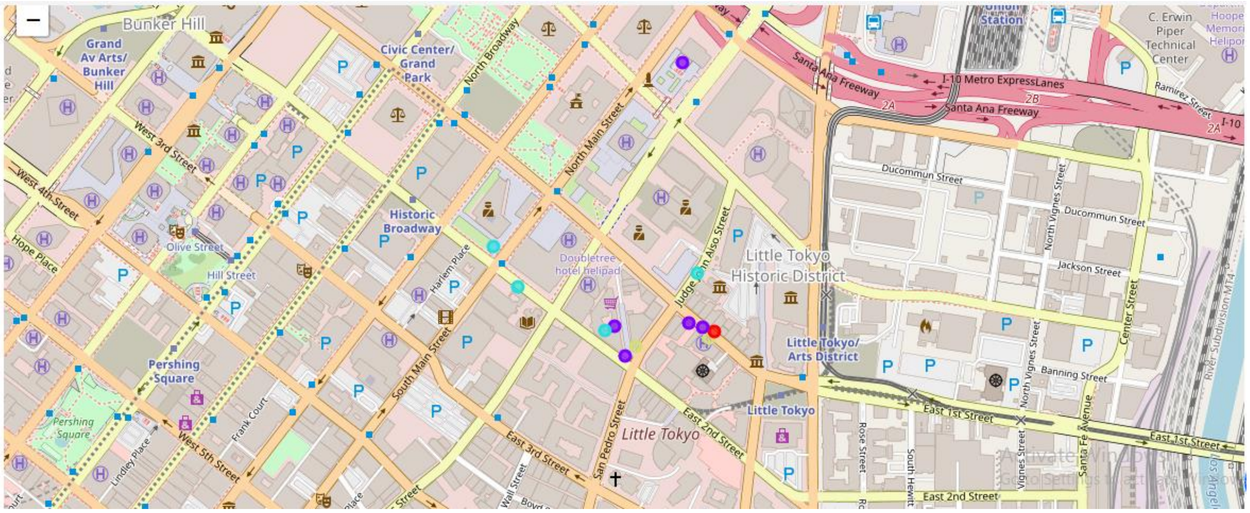
	name	id	categories	lat	lng
1	Badmaash	518471e6498e1c0b5f1401f9	Indian Restaurant	34.051342	-118.244571
2	Redbird	54938133498ed65f02e8c4ba	American Restaurant	34.050666	-118.244068
4	JiST Cafe	51dccd46498e4f9ac4865270	Breakfast Spot	34.050908	-118.240436
9	Cafe Demitasse	4e2071738130e92fc6a3f821	Coffee Shop	34.049668	-118.241696
11	Marugame Monzo	5143f2d7e4b039102cf9793f	Udon Restaurant	34.049807	-118.240202

From the list of venue ids, I pulled the likes and added them to the dataframe. I bin the total likes and visualize the data with a histogram as shown below:

In [19]:
```
# let's visualize our total likes based on a histogram
%matplotlib inline
import matplotlib.pyplot as plt
losangeles_venues['total likes'].hist(bins=4)
plt.show()
```



I categorized the data based on likes and used one hot encoding to represent the categorical data more expressively. I used unsupervised learning **K-means algorithm** to cluster the restaurants. I used python **folium** library to visualize the clusters as shown below:



I went ahead to cluster the data 4 ways, based on the total likes of each restaurant and their similarities.

4. Results

I represented the observations in the following clusters:

CLUSTER 1

characteristics

```
Poor quality food
```

In [50]:
```
losangeles_venues.loc[losangeles_venues['label']== 0]
```

Out[50]:

	name	id	categories	lat	lng	total likes	categories_new	label
13	Daikokuya	4127e200f964a520540c1fe3	Ramen Restaurant	34.049914	-118.240095	585	euro asia indian food	0

CLUSTER 2

characteristics

```
below average quality food
Mostly Europe / Asia inspired food
```

In [48]:
```
losangeles_venues.loc[losangeles_venues['label']==1]
```

Out[48]:

	name	id	categories	lat	lng	total likes	categories_new	label
15	Mitsuru Sushi and Grill	4b5b6561f964a520b2fa28e3	Sushi Restaurant	34.050066	-118.240620	12	euro asia indian food	1
16	Midori Matcha	5869aa300037eb49446d5351	Food & Drink Shop	34.050011	-118.242124	28	other	1
17	Starbucks	57fd4578498e20e69bc98c2a	Coffee Shop	34.049518	-118.241908	9	other	1
22	My Ramen Bar	54aae895498e545686bde596	Noodle House	34.049993	-118.240341	36	euro asia indian food	1
23	Quiznos	4c50911b5ee81b8d33cacefe	Sandwich Place	34.054424	-118.240744	1	other	1

CLUSTER 3

characteristics

```
High quality food
Mostly Mexican and South American food
```

```
In [49]: losangeles_venues.loc[losangeles_venues['label']==2]
```

Out[49]:

	name	id	categories	lat	lng	total likes	categories_new	label
1	Badmaash	518471e6498e1c0b5f1401f9	Indian Restaurant	34.051342	-118.244571	213	euro asia indian food	2
2	Redbird	54938133498ed65f02e8c4ba	American Restaurant	34.050666	-118.244068	218	american food	2
4	JiST Cafe	51dccd46498e4f9ac4865270	Breakfast Spot	34.050908	-118.240436	123	other	2
19	Orochon Ramen	46ddce98f964a520934a1fe3	Noodle House	34.049939	-118.242319	162	euro asia indian food	2

CLUSTER 4

characteristics

```
Above average quality food
```

```
In [52]: losangeles_venues.loc[losangeles_venues['label']==3]
```

Out[52]:

	name	id	categories	lat	lng	total likes	categories_new	label
9	Cafe Demitasse	4e2071738130e92fc6a3f821	Coffee Shop	34.049668	-118.241696	340	other	3
11	Marugame Monzo	5143f2d7e4b039102cf9793f	Udon Restaurant	34.049807	-118.240202	356	euro asia indian food	3

5. Discussion

The thought process behind this is that likes are a proxy for quality. The more likes there are, the better the restaurant is. This might be incorrect but API call issues (how many I can use for free) holds me back from getting price / rating data.

I ended the study by visualizing the data and clustering the information.

We have divided the restaurants in Los Angeles into the 4 Clusters below:

Cluster 1: Poor quality food

Cluster 2: Below average quality food

Cluster 3: High quality food

Cluster 4: Above average quality food

6. Conclusion

In conclusion, there are different types of restaurants in Los Angeles and data analysis can provide plenty of useful information to meet one’s needs.