THE BATTLE OF THE CUISINES

Derek opel August 5, 2020

1. Introduction

Restaurants serve as a very important attraction to a great number of people who wish to move to a city, town, or community, as well as for open-minded tourists who wish to explore a different variety of cuisines. But to the aspiring entrepreneur whose goal is to open a new restaurant, location is paramount to consider. There are a variety of factors that influence the location for such enterprise, such as anticipated sales volume, accessibility to potential customers, the rent-paying capacity, restrictive ordinances, traffic density, customer parking facilities, proximity to other businesses, history of the site, terms of the lease, and future development.

This project explores the issue of neighboring restaurants in the San Diego, California in addition to their categories--entailing the type of cuisine served--which may be for or against those trying to open restaurants. San Diego has one of the most diversified communities in the United States as well as popular tourist attractions. The city offers a myriad of different cuisines from different parts of the world. Its rich tapestry of cuisines is tremendous and is enticing to almost anyone whose goal is to open their own restaurant whether their objective is to serve cuisine originating from the U.S., Europe, Latin America, Africa, Asia or somewhere else. However, as highlighted above, there are many other factors that influence the decision to locate. This project aims to shed light into the variety of cuisines offered by several San Diego communities and will aid relevant stakeholders in finding the right location to open a new restaurant of their choice.

2. Data

To begin to address the question of where a stakeholder may have a successful feat in opening a restaurant, having a specific or general cuisine in mind, I have to scrape location data. This location data includes zip code and its surrounding neighborhoods, the location coordinates of those zip codes, as well as the top 50 restaurants per designated area with an assigned category. This can be achieved using the Foursquare API. Succinctly, Foursquare is a location data and technology platform, and its API allows developers to interact and harness its power in extracting a plethora of details about venues. Consequently, the Foursquare API is our primary means of procuring pertinent data for the problem at hand.

3. Methods

3.1. Collecting and Preprocessing Data

Before procuring restaurant data using the Foursquare API, I use the BeautifulSoup library to web scrape the zip codes of San Diego county in the city-data page[1]. For every extracted zip code, I then web scrape its list of neighborhoods in the other city-data page[2]. However, the data needs to be cleaned as every even index contains a newline character, so I extract only the meat of the neighborhoods data by filtering out those redundant and unnecessary indices.

The zip codes data is then passed into the google geocoding API to collect the location coordinates of every zip code. First, in order to use the google

[1] https://www.city-data.com/zipmaps/San-Diego-California.html

[2] https://www.city-data.com/zips/{enter_zip_code}.html

geocoding API, one has to create an API key[3] and is well-advised to restrict this key to only the geocoding API as well as designated websites in order to secure it from unauthorized usage.

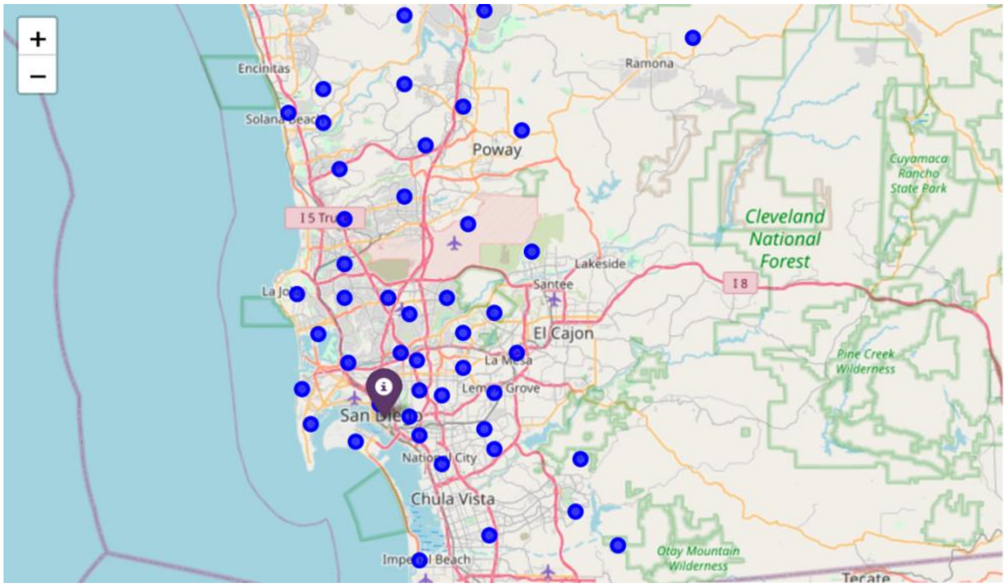
After collecting the location coordinates, those coordinates are coalesced with the zip code and neighborhood data. This consolidation needs minor cleaning in order to polish the data frame into a well-rounded, easy-to-read table.

Once this data table is complete, the next step is to extract the restaurant category data from the Foursquare API[4]. This is achieved by passing the location coordinates, query, and credentials which is obtained in the Foursquare website by creating a free account in the developer portal. For this case, I limit the search to the top 50 most popular restaurants for each zip code. I extract the restaurant name, restaurant category, and restaurant coordinates from this data and combine it with the location data table.

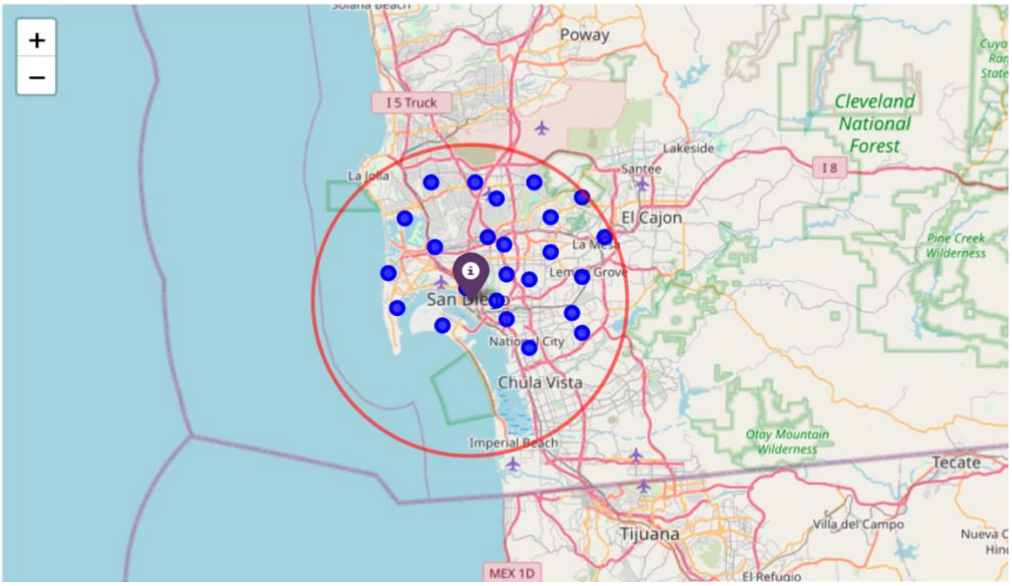
3.2. Exploratory Data Analysis

First of all, I it would be nice to get a feel of the San Diego regions defined in this data, so I use the folium library to visualize the regional spatiality in order to conceptualize the number of regions I will have to explore for restaurant data.

[3] https://developers.google.com/maps/documentation/geocoding/get-api-key
[4] https://developer.foursquare.com/

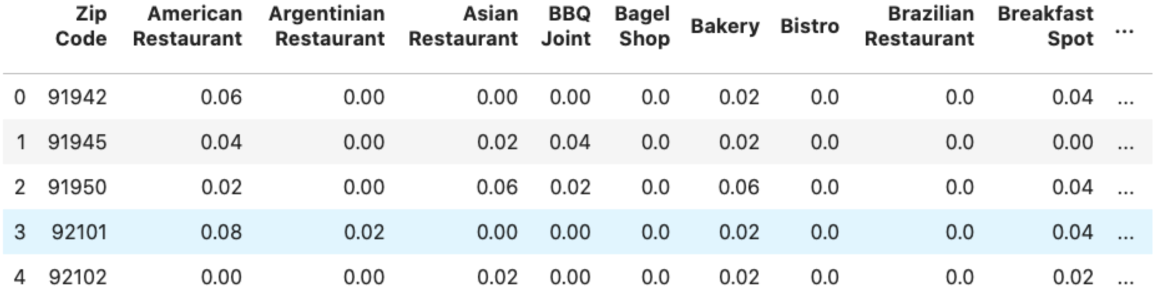


As you can see some of the more rural areas, farther from Downtown San Diego, have somewhat inaccurate location coordinates. This is because the zip code regions tend to larger and irregular in shape as the distance increases away from San Diego. This is okay as my primary focus is geared towards Central San Diego and to simplify the data procurement process. I am going to limit my regions of interest to within roughly 10 miles of Downtown San Diego. I do this by filtering out the location coordinates to specific boundaries that meet this criterion.



Once I define my boundary, the restaurant categories are converted via one hot encoding; that is, they are converted from categorical to integer data so that I may apply my model on them. Once this conversion is complete, I take the mean of the frequency of each restaurant category from each zip code. The result is 25 rows, or zip codes, and 73 columns, or restaurant categories.

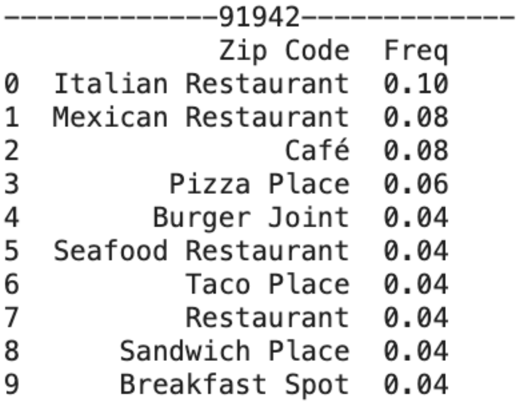
A table showing a subset of the frequency of each restaurant category per zip code:



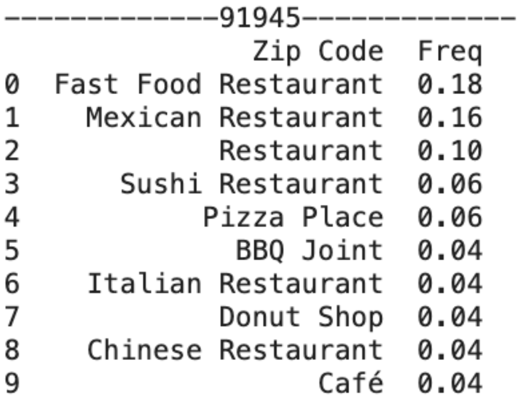
	Zip Code	American Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bistro	Brazilian Restaurant	Breakfast Spot	...
0	91942	0.06	0.00	0.00	0.00	0.0	0.02	0.0	0.0	0.04	...
1	91945	0.04	0.00	0.02	0.04	0.0	0.02	0.0	0.0	0.00	...
2	91950	0.02	0.00	0.06	0.02	0.0	0.06	0.0	0.0	0.04	...
3	92101	0.08	0.02	0.00	0.00	0.0	0.02	0.0	0.0	0.04	...
4	92102	0.00	0.00	0.02	0.00	0.0	0.02	0.0	0.0	0.02	...

To get a better understanding of popular restaurant categories, I take the top 10 restaurant categories for each zip code and print them out in a way that makes it relatively easy to digest.

A printed example of the first two zip codes along with the frequency of restaurant categories:



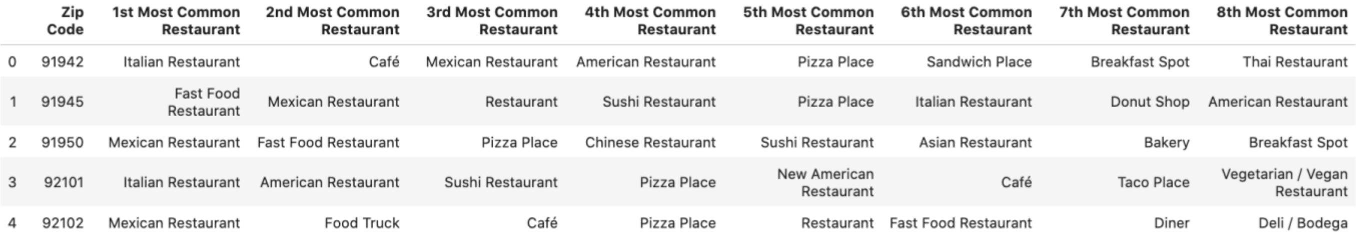
```
-------------91942-------------
             Zip Code  Freq
0  Italian Restaurant  0.10
1  Mexican Restaurant  0.08
2                Café  0.08
3         Pizza Place  0.06
4        Burger Joint  0.04
5  Seafood Restaurant  0.04
6          Taco Place  0.04
7          Restaurant  0.04
8      Sandwich Place  0.04
9      Breakfast Spot  0.04
```



```
-------------91945-------------
                 Zip Code  Freq
0  Fast Food Restaurant  0.18
1    Mexican Restaurant  0.16
2            Restaurant  0.10
3      Sushi Restaurant  0.06
4           Pizza Place  0.06
5             BBQ Joint  0.04
6    Italian Restaurant  0.04
7            Donut Shop  0.04
8    Chinese Restaurant  0.04
9                  Café  0.04
```

The above tables already make a statement in that it highlights San Diego's very diverse dishes and grub. In order to make these tables more readable and simplified, I then convert it to a single ordinal table in order to shed light in to the top 8 restaurant categories per region.

The below ordinal table shows top 8 restaurant categories for the first five zip codes:



	Zip Code	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
0	91942	Italian Restaurant	Café	Mexican Restaurant	American Restaurant	Pizza Place	Sandwich Place	Breakfast Spot	Thai Restaurant
1	91945	Fast Food Restaurant	Mexican Restaurant	Restaurant	Sushi Restaurant	Pizza Place	Italian Restaurant	Donut Shop	American Restaurant
2	91950	Mexican Restaurant	Fast Food Restaurant	Pizza Place	Chinese Restaurant	Sushi Restaurant	Asian Restaurant	Bakery	Breakfast Spot
3	92101	Italian Restaurant	American Restaurant	Sushi Restaurant	Pizza Place	New American Restaurant	Café	Taco Place	Vegetarian / Vegan Restaurant
4	92102	Mexican Restaurant	Food Truck	Café	Pizza Place	Restaurant	Fast Food Restaurant	Diner	Deli / Bodega

So far, it looks like the dominant categories consist of Mexican, American, and Italian restaurants along with Pizza places and Cafés.

3.3. Modeling

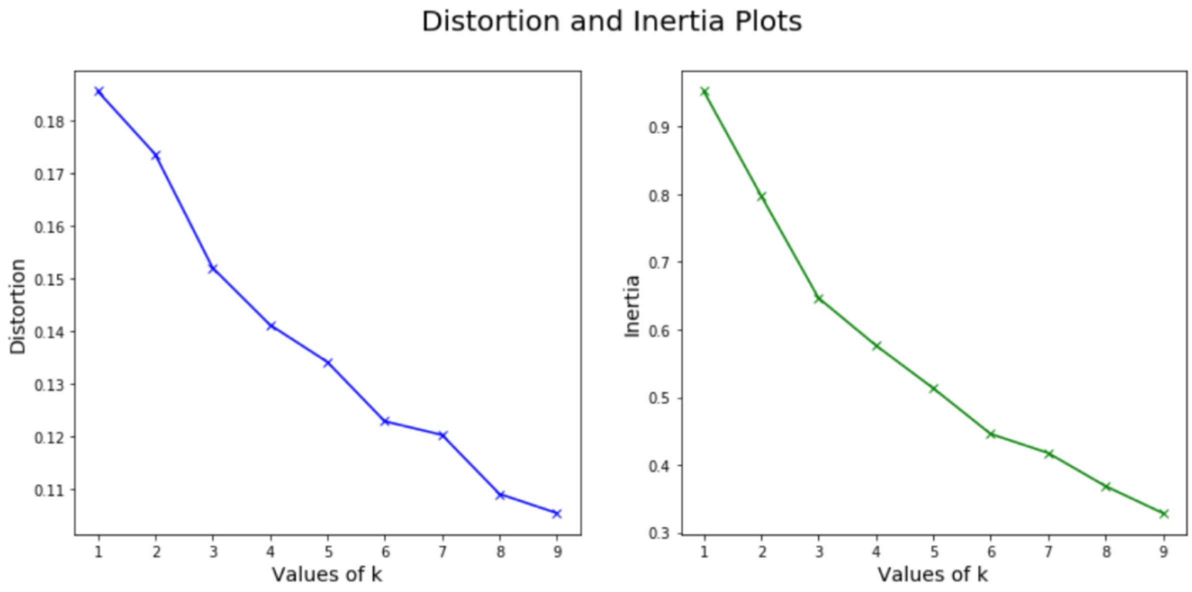
In order to find patterns and cluster the regions using restaurant category, I will choose the K-Means algorithm as my modeling technique. This is a relatively simple, iterative model whereby its goal is to cluster n observations into k distinct classes, or prototypes. These classes are based on the similarity metric—in this case, restaurant category. We can calculate the data points' similarity through different metrics such as the Euclidean distance, correlation-based similarity, cosine distance, Minkowski distance, etc. For this case, we will use the Euclidean distance as our similarity metric. The Euclidean distance is very fundamental and simple to understand. It is simply the straight-line distance between two points.

The Euclidean metric formula:

$$distance = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}$$

However, since the K-Means algorithm is an unsupervised method—that is, there are no class labels to train the model on—I will have to optimize the number of clusters k using other methods. Generally speaking, we want the value k such that we minimize the distance between data points and their respective centroids and maximize the distance between clusters. In order to find the optimal k, we use the elbow method. I use two elbow methods: distortion and inertia. Distortion is the average of the sum of squared distances between the data points and their respective centroids, the mean of the cluster. Inertia is simply the sum of squared distances of data points to their respective centroid; the lower the inertia, the more condensed the clusters are. When using the elbow method, as k increases, we typically find that distortion and inertia decrease at a fast rate and then suddenly decrease at a slower rate, and, hence, forming the elbow point.

I chose nine different values of k ranging from one to nine, and obtained the following results:



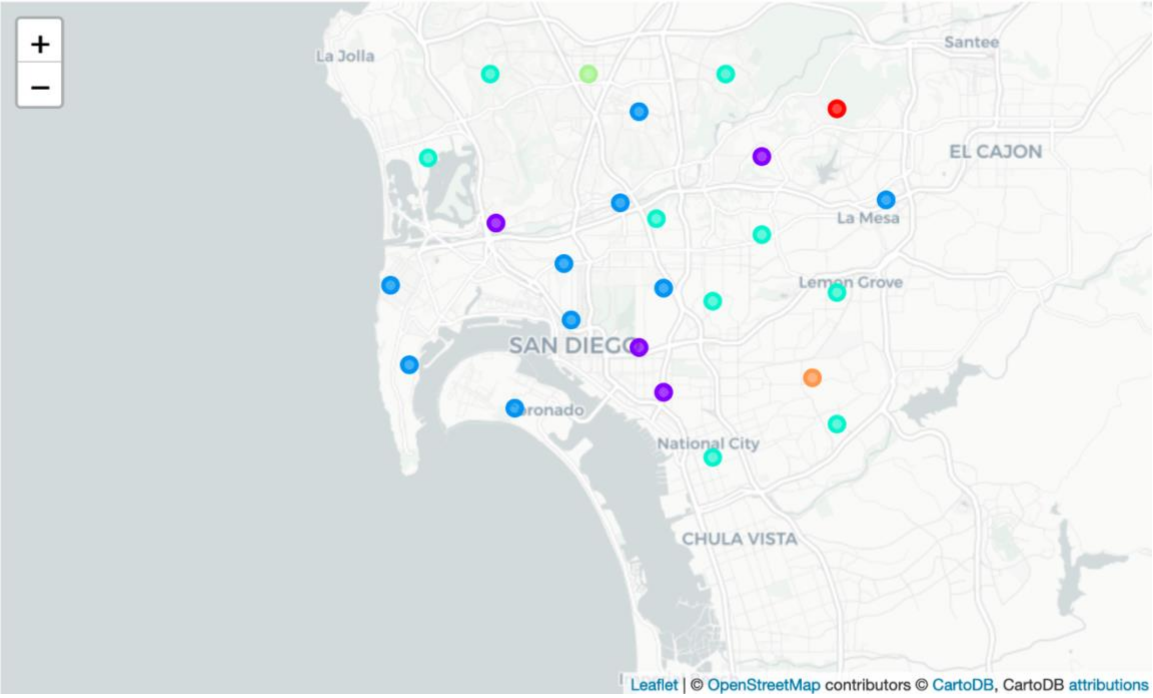
There are two elbow points at k=3 and k=6. Although it is difficult to discern between the two elbow points, distortion and inertia more readily decrease at k=6.

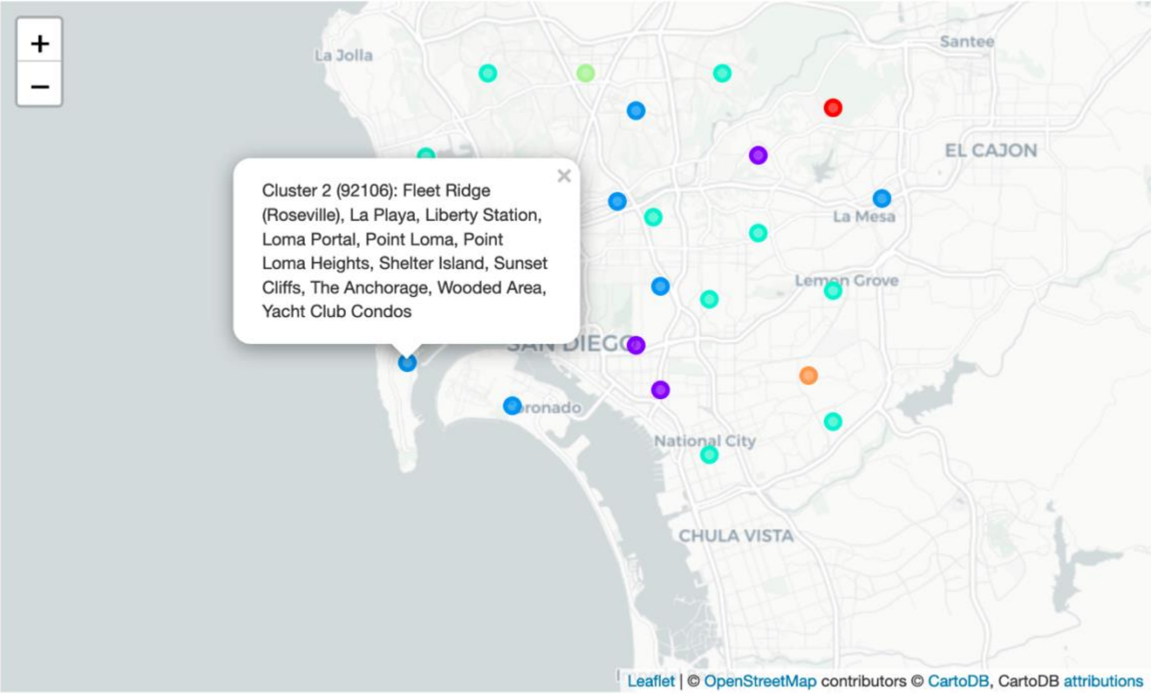
This elbow becomes more apparent when studied by a keen eye. Therefore, I choose six as the optimal value of k for our model.

4. Results

After fitting the K-Means model with our goal of 6 classes, I displayed the resulting clusters on the folium map with each cluster having a different color. The clusters are represented as colored circle markers. Each colored circle marker has an associated set of neighborhoods shown as a popup.

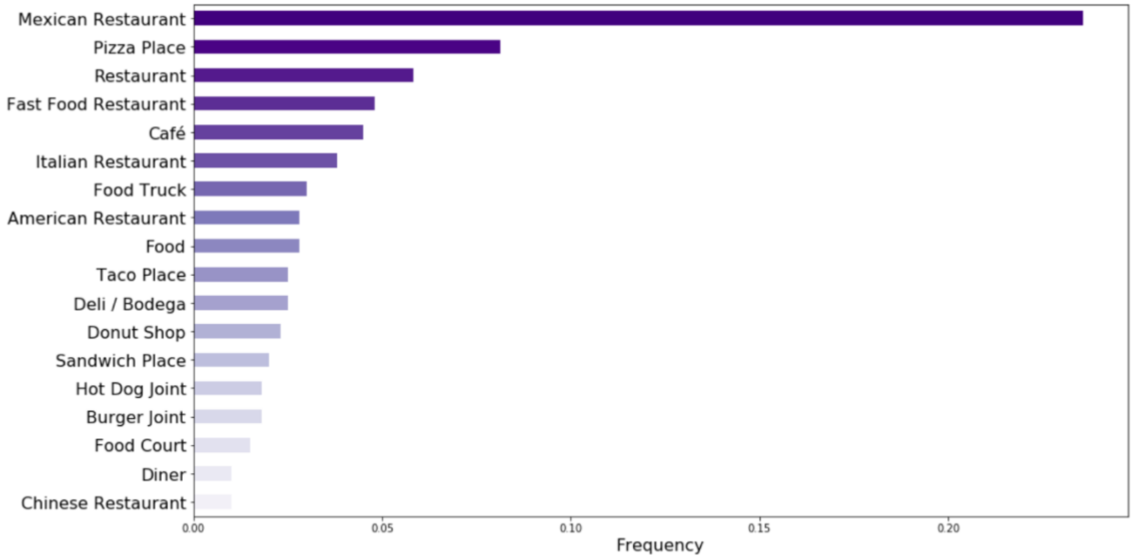
Folium Map showing all 6 clusters representing different classes of restaurant categories:




Cluster 2 (92106): Fleet Ridge (Roseville), La Playa, Liberty Station, Loma Portal, Point Loma, Point Loma Heights, Shelter Island, Sunset Cliffs, The Anchorage, Wooded Area, Yacht Club Condos

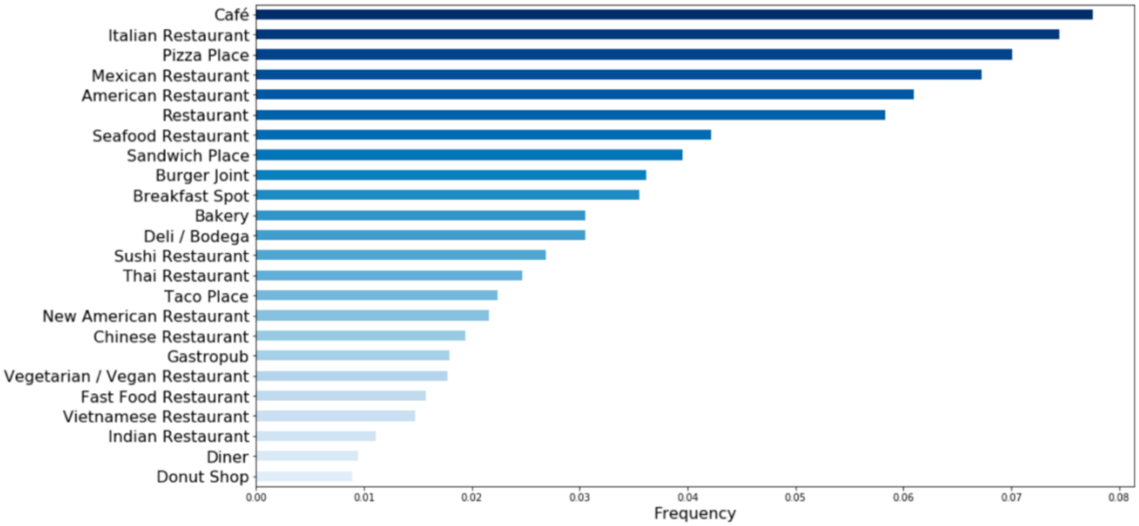
With the folium map in mind, I create a horizontal bar chart displaying the mean frequency distribution of restaurant categories for each cluster in order to manually generate their labels and highlight distinctions between them.

Cluster 1



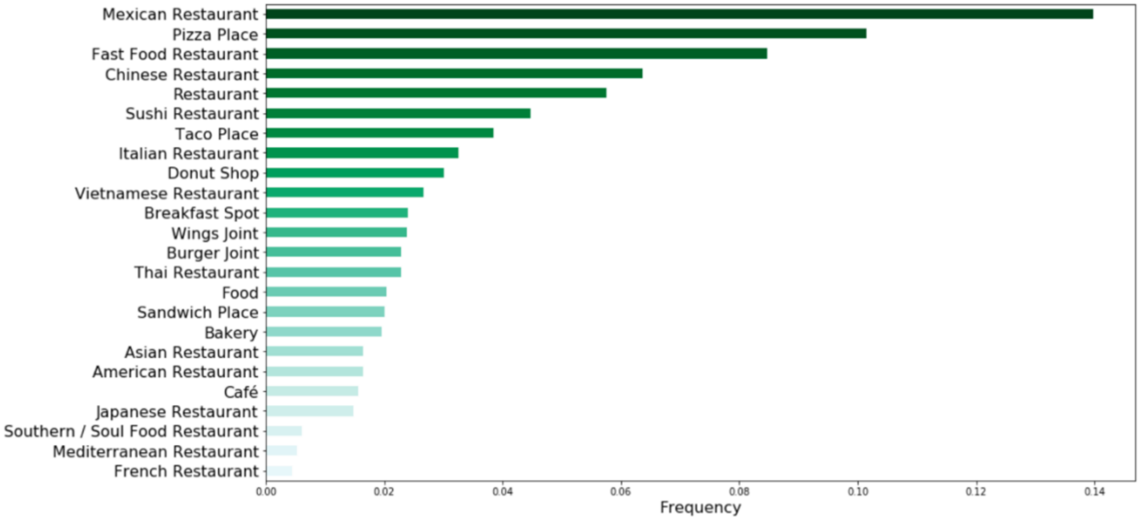
It is very apparent that Cluster 1 most prominently consists of Mexican restaurants.

Cluster 2



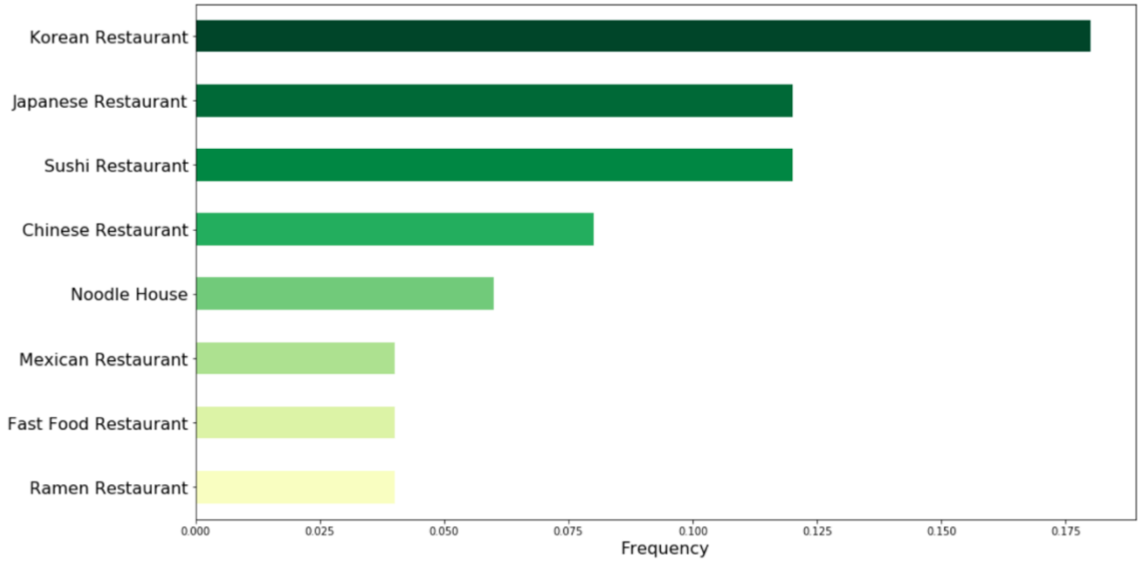
Cluster 2 is more uniform and diversified in restaurant category with more emphasis placed on Cafés as well as Italian, Mexican, and American restaurants.

Cluster 3



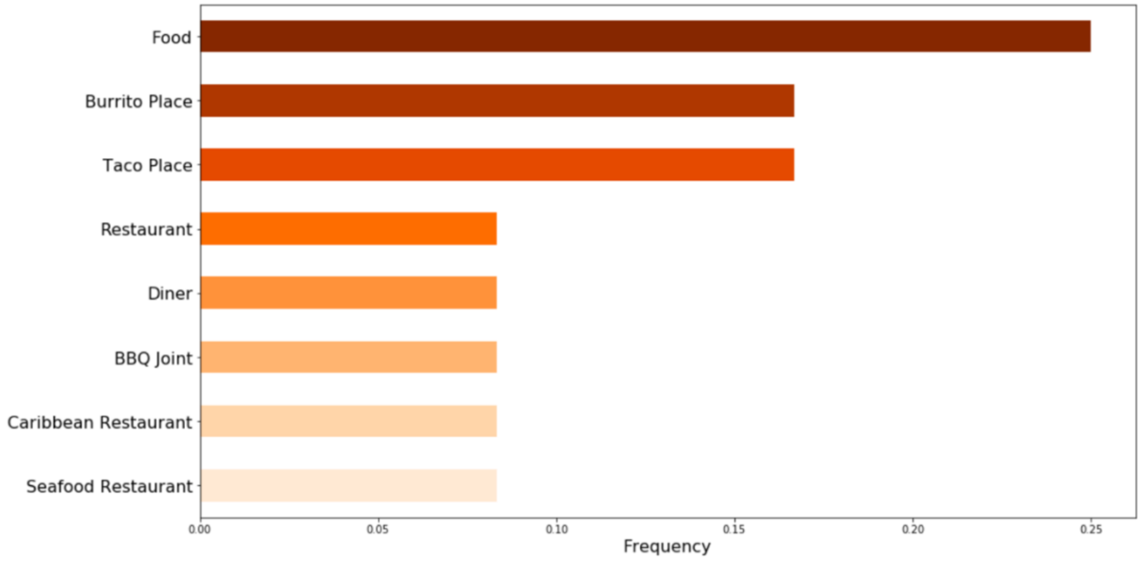
Cluster 3, like Cluster 1, also dominantly contains Mexican restaurants but is more diversified by offering different Asian cuisines. It is also sprinkled with Italian and American cuisines.

Cluster 4



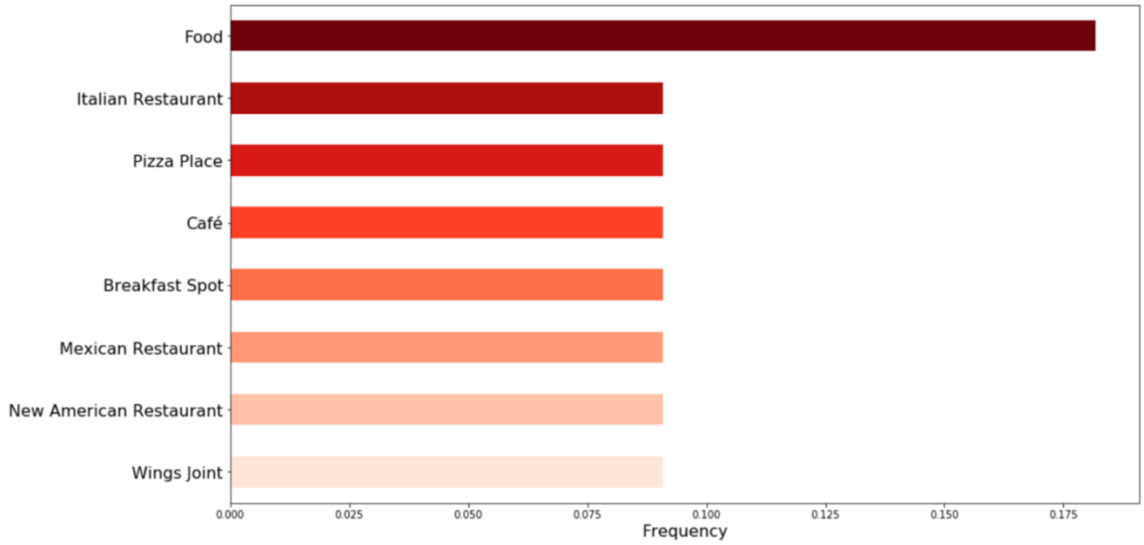
Interestingly, Cluster 4 has a stark difference from the rest of the clusters in that it has a strong emphasis in Asian cuisine.

Cluster 5



Cluster 5 contains more general cuisine with more emphasis placed in fast food Mexican restaurants.

Cluster 6



Like Cluster 5, Cluster 6 contains general cuisine but with more emphasis in Italian and American cuisines as well as cafés.

5. Discussion

Interestingly, based on the results from the map above, there are three clusters that contain only one region and three clusters that contain four or more regions. Again, as aforementioned, this bolsters the prevalent statement that San Diego harbors a rich tapestry of cuisines. It is also interesting to note that the one-region clusters tend to be farther away from Downtown San Diego whereas the larger clusters tend to gather towards its center and along the coastline. This may be a result of cultural autonomy from Downtown San Diego and the beaches.

Another important point to make is that, if you pay careful attention, there is an abundance of Mexican Restaurants in the San Diego area. Intuitively, this makes sense as San Diego is roughly 15 miles North of Mexico. With this in mind, there is ample opportunity for one to open a Mexican restaurant, but the exact location of such opening would have to be further investigated as there are many more variables that come into play as mentioned before.

Lastly, there is also an abundance of Italian restaurants, American restaurants, fast food restaurants, and cafés. Additionally, one area harbors predominantly Asian restaurants. Again, this highlights the fact that San Diego is inclusive with its cuisines. This inclusivity entails ample opportunity for starting a restaurant business of almost any kind of category. And because this project limits the boundaries and scope of the San Diego regions and due to the sole reliance of the Foursquare data, there is a significant possibility that San Diego restaurant categories are more diversified than what our results display.

6. Conclusion

Restaurants and their respective cuisines they offer is a very influential factor when choosing to locate to a specific city, town, or community. It also plays a vital part when tourism is under consideration as many tourists take pleasure in exploring different varieties of cuisines. San Diego is one of the top places in the United States that offers popular tourist attractions. The goal of this project is to aid the aspiring entrepreneur whose goal is to open a new restaurant in this area by highlighting potential and promising locations. It is important to note that other varying factors will influence the decision to locate such as anticipated sales volume, accessibility to potential customers, the rent-paying capacity of the business, restrictive ordinances, traffic density, customer parking facilities, proximity to other businesses, history of the site, terms of the lease, and future development.

This project explores the issues of neighboring restaurants in the San Diego area combined with the need to identify their cuisines, which may serve to positively or negatively impact those trying to open restaurants.

Based on our results, San Diego harbors a large and varied list of cuisines and is subject to location. We found 6 different classes pertaining to the restaurant categories within ten miles of Downtown San Diego. However, our findings should not be limited to the results' categories as the K-Means clustering algorithm deals with only the top eight restaurant categories per region. Each region harbors many more differing cuisines, albeit less common.

7. Future Works

The above analysis shows promising results in helping the entrepreneur start a restaurant business. The analysis can be extended by adding other important variables that influence the decision-making process such as traffic density, future development, anticipated sales volumes, etc. Furthermore, the above analysis may be enhanced by creating an algorithm that consolidates restaurant categories as some categories had overlap. For example, Sushi restaurants and Japanese restaurants were treated as different categories by the Foursquare API. Such algorithm will need meticulous thought, planning, and development.

8. References

[1] Foursquare Developer. (n.d.). Retrieved August 01, 2020, from https://developer.foursquare.com/

[2] (n.d.). Retrieved August 02, 2020, from https://console.cloud.google.com/

[3] Stats about all US cities - real estate, relocation info, crime, house prices, cost of living, races, home value estimator, recent sales, income, photos, schools, maps, weather, neighborhoods, and more. (n.d.). Retrieved August 01, 2020, from https://www.city-data.com/

[4] The Staff of Entrepreneur Media, I. (2009, October 01). How to Start a Restaurant. Retrieved August 05, 2020, from https://www.entrepreneur.com/article/73384