# Where to Open a New Peruvian Restaurant

Adolfo G. Ramirez-Aristizabal

## Intro

The data science capstone project aims to gather insights for a common problem for the up and coming businessman. The scenario is framed to help a restaurant owner expand into a new location. The restaurant owner has already established a successful local restaurant for the past 20 years. The restaurant has grown from your standard hole in the wall into a household name and popular name in their local region. The restaurant is essentially successful enough to be able to test out expansion and start a chain of Peruvian restaurants. The opening of the second restaurant will be crucial in the long-term goal of having a successful chain of Peruvian restaurants, so the success of the second restaurant will need some careful planning. We decide to choose Los Angeles County as the region in question. The size and high-density population of this area poses an interesting challenge. We know that Los Angeles County is very culturally diverse, which would make the problem interesting to find a location where an ethnic restaurant can be successful. In this project we explore data science methodology to gather insights of features related to types of venues and frequency of venues related to our Peruvian restaurant.

## Data

The first data frame used in the project had four columns including zip code and city name data along with latitude and longitude for every city. The data was provided by the 'lacounty.gov' page, which provided a total of 370 cities in the Los Angeles County area.

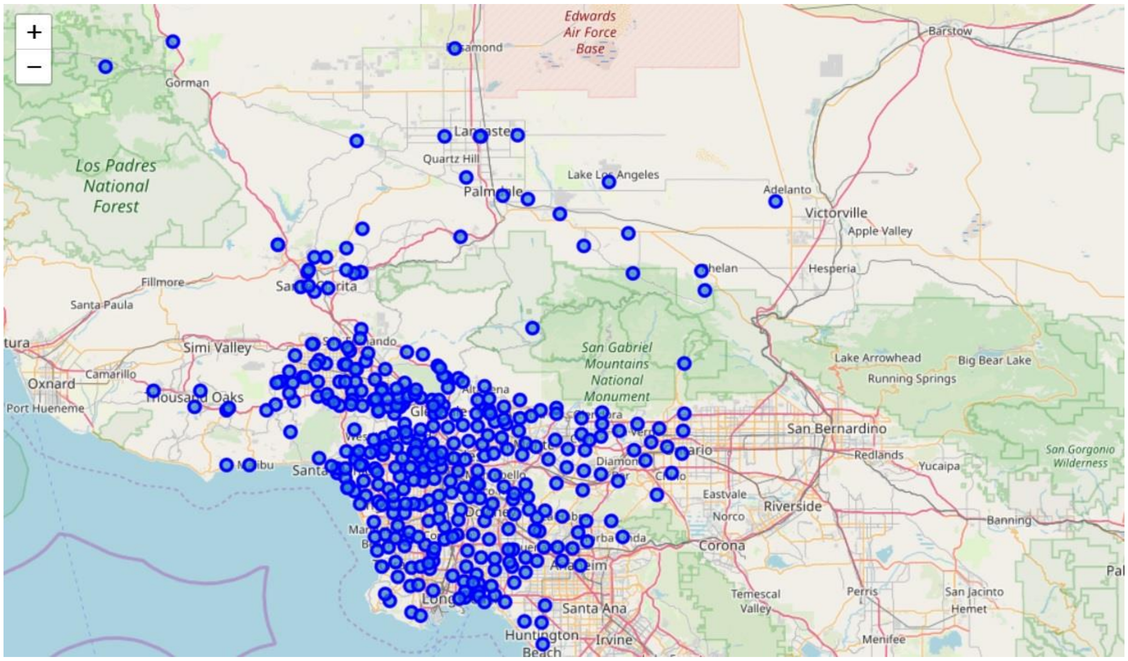

Figure 1: Here is a geo map plot with markers indicating the location of all 370 cities available from the Los Angeles County data set.

Through the use of the Four Square API, we were able to retrieve nearby venue data. An updated data frame now has venue name, venue location, and venue category as new columns of data. The new data frame had a total length of 7296 rows, which are instances of venue locations. The data is simple but we realize we can gather useful insights from semantic categories, location, and item frequency.

## Method

The first part of the project started with finding quality data and putting it into a pandas data frame to use for later. The data set used came from government data, which in this project is considered to be of thorough quality. Like with any data set, the format was not to the specifications needed. The location data came in one column and it had other symbols and unnecessary information with it. The string extract/split methods along with regular expressions are used to clean up the data and define a column for latitude and longitude. This method of data cleaning seems to be common, especially with location data. Next, a 'getNearbyVenues' function is defined to be used with the Four Square API. The function is taken from the previous class project that was used to cluster Toronto venues. The radius parameter is set to 500 so the data retrieved is city centric and avoids low traffic areas. The function is also given a high limit of 1000 for venues extracted. This is meant to capture a range comfortable to big cities.

Next, the project followed the same guidelines as in the Toronto project. A K-Means clustering was used to group venues and order them by most frequent venues per city. To do this, a one-hot encoding was done on the 'Venue Category' column and then it was grouped by mean. To figure out the best K parameter, the elbow method is used.

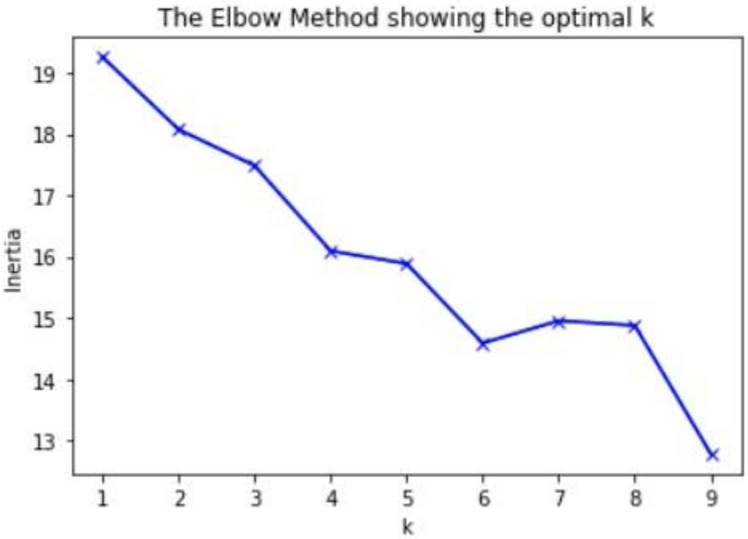

Figure 2: This figure shows a clear break at k=6, which is used to set the k parameter. In this case the inertia attribute of the model is used to measure when performance shows an elbow trend.

The y-axis in figure 2 comes from the model's inertia attribute. This is the same as getting a measure of the sum of squared distances of samples from the cluster center. With a k parameter of 6, the model was fitted and a plot was generated to understand the geographic locations of the clusters. Figure 3 demonstrates the spread of cluster types with the red cluster containing mostly food related venues. The other clusters do not contain many items and it is clear that the cluster of interest is the red one. A problem is encountered through this method because there are many venue types that are irrelevant to our guiding question.

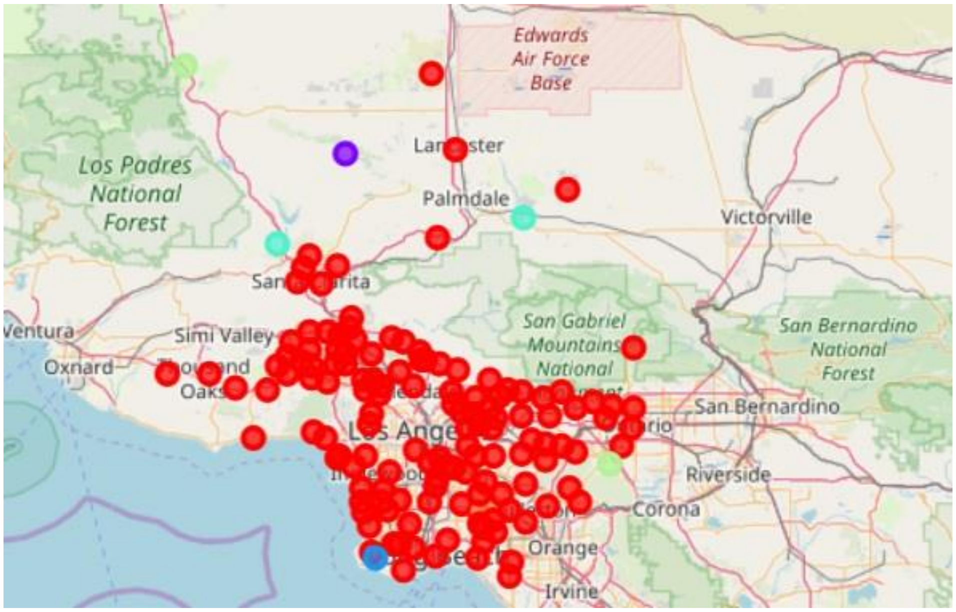

Figure 3: The geographic map of clusters, k=6. The red cluster has many more items in comparison to other clusters and this includes all venue types.

To overcome working with irrelevant data, the text in the venue categories is tokenized before doing K-means. This attempts to use the text data to get some meaningful relationship of the venue category labels. The tokenization used was the *Term Frequency Inverse Document Frequency* (TFID), which treats every venue description as a vector and compares frequency to get a proxy measurement of a semantic score. This is useful because we know that there are types of venue categories that we are more interested in compared to others. This project uses the sklearn feature extraction library. Frequency is usually not very useful for getting a semantic score because many function words such as ‘the’ or ‘and’ will be overrepresented in text data. The inverse document frequency is paired up in this method to downplay the influence of irrelevant function words and focus of semantic rich words.

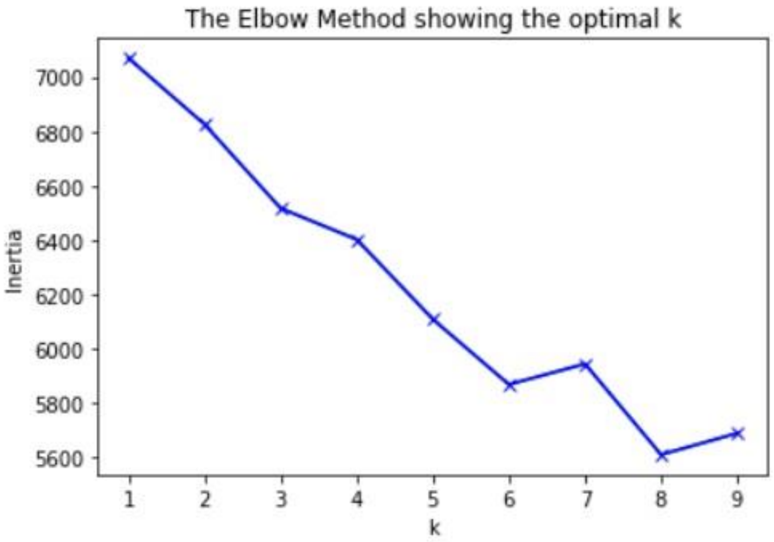

Figure 4: The inertia attribute is used again to graph an elbow trend. A k=6 is chosen because it is the first big elbow trend in the graph.

The TFID vectorization allows us to find a cluster that is the most relevant. Then, K-means is used to then group within that cluster to look at different types of restaurants. Simple exploratory steps are done to understand what type of venues are located where. This includes doing another geographic plot of the new cluster data and sorting data by new clusters to compare most frequent venue types per city.

## Results

The methodology presented above demonstrated that a straightforward k-means clustering technique was not sufficient to understand where a new Peruvian restaurant location can occur. A better way to understand this is by creating a word cloud plot of the venue category column in the data.

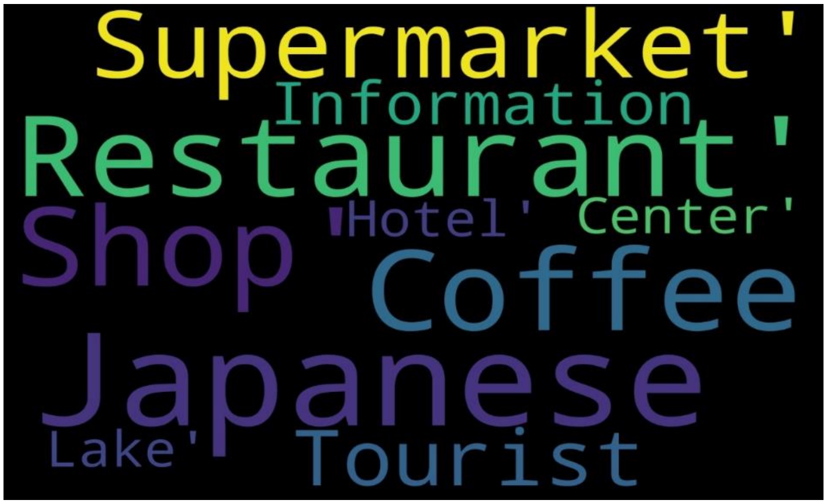

Figure 5: A word cloud of all venue categories before text analysis.

It is clear by looking at Figure 5 that there is a representation of venue categories that are irrelevant to our search without proper clustering. For example, it is useful for us to know where the restaurant venues are mostly located in each city, but it is also useful to know if there are locations in which there are no restaurants. Therefore, TFID vectorization is used to cluster venue categories based on simple semantic features.

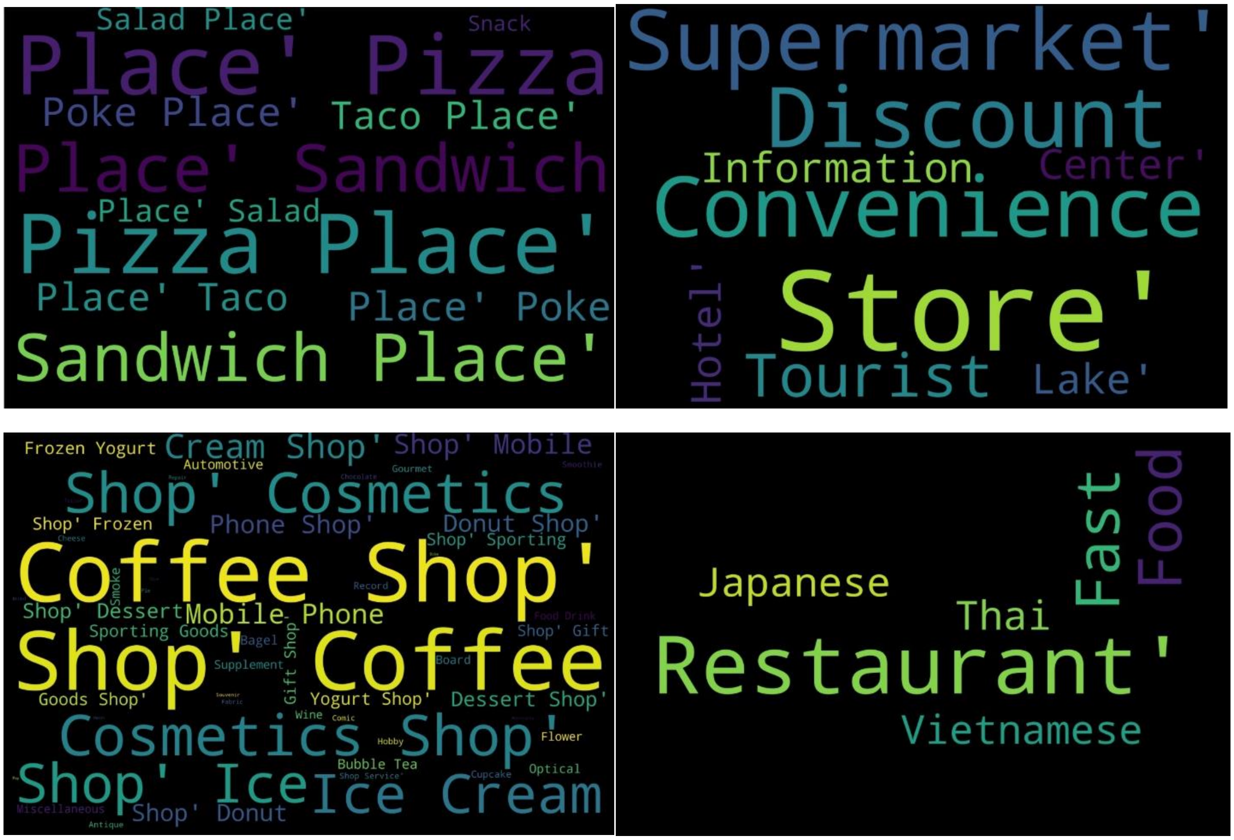

Figure 6: Word clouds from some of the clusters after TFID vectorization. Bottom right shows the cluster of interest.

The success of TFID vectorization is shown from the example clusters in Figure 6. Not only can it tell us where there are food places, but it also allowed for the clustering of different food venues. The location of various food places is important to know, but because we are trying to open a new restaurant, the clustering of restaurant venues is useful. With this we can make informed decision about whether we can locate ourselves in areas that already have restaurants as well as trying to not compete with similar restaurants. The restaurant cluster is then used to refine our search, and a K-means clustering is done to now identify different types of restaurants. The same K-means process is done but now with only venues that fall into the restaurant cluster.

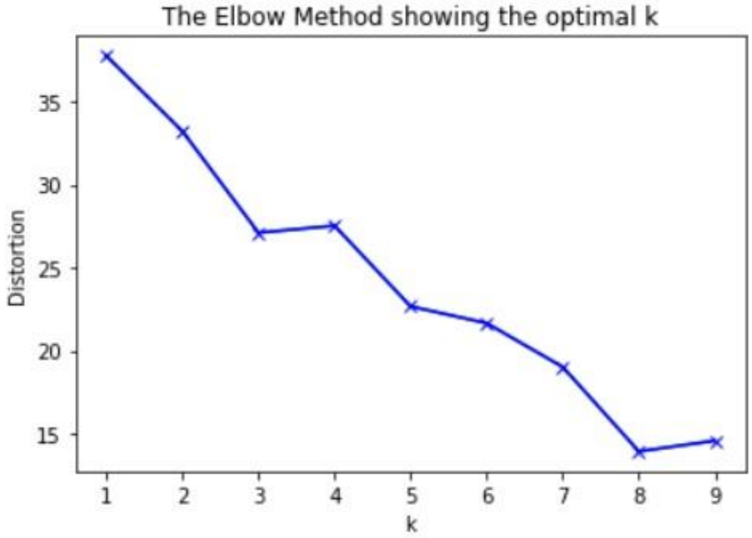

Figure 7: Elbow method plot for the restaurant cluster data.

The elbow method demonstrated that k = 3 could be a choice to explore. With this parameter setting, it was possible to find distinct clusters of restaurant types. New data frames were created and sorted to see the top ten most frequent venues among those clusters. The first cluster was defined as the 'Diverse' cluster, in which there was no pattern to restaurant types. The second cluster was defined as the 'Mexican-Asian' cluster because almost all cities had Mexican restaurants as their top ranked venues and then Asian restaurants following after that. Lastly, the third cluster was defined as 'Fast Food-Asian', with Fast Food being the most prominent and the Asian restaurant category following that frequency. The 'Fast Food-Asian' cluster contains a considerably higher amount of venues compared to the 'Mexican-Asian' cluster with the 'Diverse' cluster containing the least.

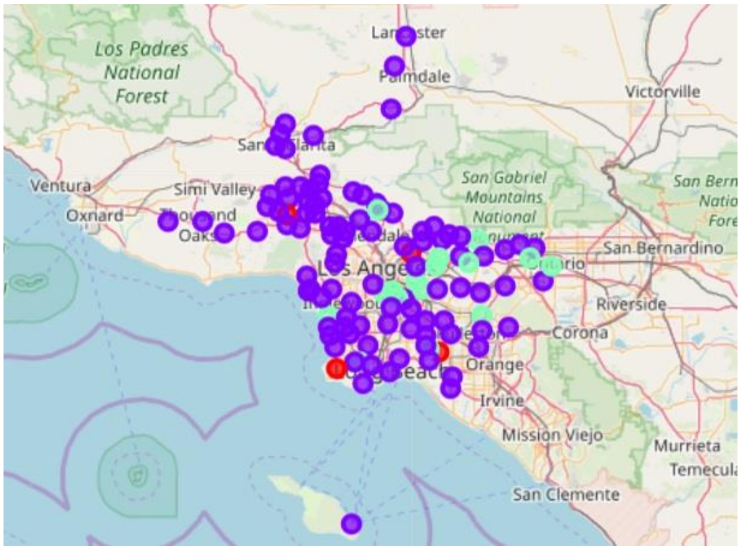

Figure 8: Geographic map of 'Fast Food-Asian' (Purple), 'Mexican-Asian' (Cyan), and 'Diverse' (Red) clusters.

## Conclusion

The presented methodology is used to help in the search for a new Peruvian restaurant location. Some of the criteria the we would need to pay attention to is to look for locations that have similar types of venues but without trying to compete with very similar venues. To do this we separated venues based on simple semantic features. From there we saw that we could add resolution to our search and cluster venues that are restaurants. If we want to pick a location that avoids unnecessary competition, we can avoid culturally similar restaurants such as in the 'Mexican-Asian' cluster. On the other hand, we might also want to avoid restaurant locations that are too different because it might take more time for customer traffic to catch onto the type of service is provided. The 'Diverse' cluster is a safe option because it washes out the effects of cultural friction and contrast in service expectations from consumer traffic. It also allows for flexibility in distribution routes from neighboring restaurants.

The three restaurant clusters give the business owner and investors plenty of initial information to balance out decisions. There are still more analyses that can be done but this initial information should be the first insights to be communicated. Further work can be explored per the investor's expertise in investment risk management. For example, we can do inferential statistics on frequencies to figure out whether there truly are significant differences between locations and cluster types. Once the investors commit to a cluster type of their choice more data can be brought in to predict consumer traffic and a timeline of success.