The **Cuban Meal Crisis**

Capstone Project for
IBM Applied Data Science Course

June 2020

By Krishna Prasad S Rao

Introduction

Los Angeles (LA) is one of the most populous and multi-cultural city in the west coast of the USA. Just a few hundred miles south of LA is the country of Mexico, with a populace that majorly speaks Spanish. Cuba, on the other hand, is a small island country located not far away from the USA, in Atlantic Ocean. Cuba is just under 100 miles from Miami, another metropolitan city within the USA.

Although both Cuba and Mexico are Spanish speaking nations, culturally they are significantly different, which also reflects in their unique colorful and spicy foods. Thus, although Mexican restaurants are aplenty in LA, authentic Cuban food is hard to come by. This creates an opportunity, especially in a metropolitan city like LA.

This project is inspired from the 2014 movie, "Chef". The title of the report is a word play on the geopolitical event, famously known as the 'Cuban Missile Crisis' that occurred in the cold-war era between the US and the Soviet Republic.

Business Problem

The goal of this project is to methodically explore the best possible locations of selling authentic Cuban food, in a food-truck, within the populous LA. Food trucks enjoy the unique advantage of mobility, allowing varied locations to be targeted for selling food. Using data science methods and predictive statistical techniques such as clustering, the project aims at providing answers to the following questions:

a.   *What are the ideal locations for selling Cuban food, in the city of Los Angeles, in a food truck?*
b.   *What are the ideal times (of the day) for selling Cuban food at the locations obtained from (a)?*

Target Audience

The report is of key interest to the proprietor of the Cuban Food truck. Besides, LA being a densely populated city with good amount of locals and tourists, a food truck can create traffic congestion as well. Thus, the findings of the report can be key to the city's traffic and police authorities as well.

Data & Sources

To answer the questions put forth in the 'business problem' section, the following data sets are essential:

a. *List of businesses to identify commercial / residential locations which can be useful for targeting the right crowd at varied times of the day. The dataset was found on kaggle.com, at this particular URL, consists of fields like business name, co-ordinates, zip-codes and business category.*

b. *The above dataset does not contain 'neighborhood names'. To obtain those, this URL is scraped and then merged with the above data set, with Zip Code as the key.*

c. *While the combination of (a) and (b) yields information about businesses, Foursquare API data is used to additionally provide locations of competing Cuban restaurants, so that other locations can be targeted for selling Cuban food. To achieve this goal, the venue and search APIs of foursquare.com is utilized.*

Sources mentioned in (a) and (b) above are available as '.csv' file and html respectively, while the data source in (c) is available in json format, which are all directly readable by various APIs of Python's Pandas package.

Methodology

Selling Cuban food in a food truck around the city of Los Angeles, California is the goal. Food trucks generally have limit supply, limited options and an affordable price, with the USP of offering authentic and delicious food on the go. This is especially attractive for small-medium businesses where time is of the essence for the employees in such businesses, while being able to afford. Since such small offices also have a tendency of 'going out for lunch' as a group, targeting sale of Cuban food around such areas warrants a high chance of high sale volume. As a longer term goal, with the obtained profit, opening of more branches around the city can be targeted. Thus, the brand of this fictional food truck business, say 'El Jefe' can now be opened at multiple locations within the city and have a brand-loyal customer base. With the goal of the business established, the following section describes the methods by which the goal was achieved.

Initially, data is downloaded from two sources mainly, as described in (a) and (b) of the Data & Sources section. Both the datasets contain co-ordinates and were obtained from the internet. Foursquare was not used for this purpose since a list of businesses in general, were needed, to identify the location clusters to sell food. The datasets from (a) and (b) are merged such that a collection of businesses along with their location, address, coordinates, zip code and neighborhood names are thus finally under one master dataframe. Unsurprisingly, the dataset obtained from the internet includes a lot of null type data, and is thus cleaned, to remove such non-workable data.

The next important part is identifying competitors. El Jefe, being a new brand in a city has to conquer customer loyalty, and as such cannot be let fall prey to established brands in the area. Hence, from the master data frame generated above, all businesses within a custom radius (in this case, 3 km) from the Cuban restaurants are filtered out. This would eventually generate a subset of locations which are not in proximity to any other Cuban restaurants and as such, presents an opportunity for seeking success in the sale of tasty Cuban food.

Having filtered the uninteresting locations out, the rest of the data frame is used for performing analysis to understand the demographics of the area. Although lot of assumptions need to be made here (owning to the lack of economic data of the neighborhoods), with the concentration of the kinds of business in a certain locality, paints a reasonable picture to work with. Thus, the dataset is 'one-hot encoded' and grouped by neighborhoods to understand the density of concentration of business per locality. Eventually, the dataset is used to understand the top 10 common businesses in a neighborhood.

Finally, the unsupervised learning method called 'Clustering' is used. Clustering, being a popular machine learning technique picks a center, computes distances from all points and forms 'clusters'. In this case, clustering is used to identify clusters of neighborhoods with similarities that are non-obvious. A total of four clusters get generated, thus highlighting that these locations are ideal as business locations for selling El Jefe branded Cuban food.

Results

The results of clustering reveal four clusters, with three clearly dominant clusters, ideal for business. The other cluster is so tiny that it can effectively be discarded. In addition, a list of neighborhoods that would be of interest to the proprietor of the El Jefe food truck is generated, as seen in the appendix. Since the dataset is primarily a list of small-medium size businesses, the lack of them in one cluster can be assumed to be the affluent part of the city.
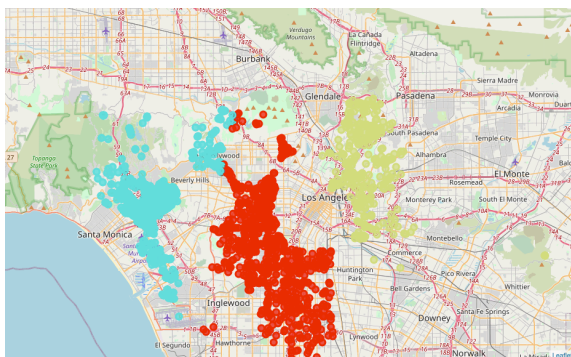
Figure showing the three dominant clusters in the city for El Jefe to operate

Discussion

From the cluster map, it is clear that there are three significant geographical locations where operating a food truck during the day, when the businesses operate, and conveniently managing to avoid competition. The other perspective would be to understand that the existing Cuban food locations are mostly spread in the either side of the city center to serve all the interested customers in the enormous city. El Jefe has a unique opportunity here, since the business enjoys mobility. With these options, the proprietor can now establish a time table so that one cluster can be targeted two days of the week. This spreads the name of the business throughout the city, and potentially helps identify locations where possible future 'branches' of the trucks can be stationed.

Future Work

This analysis has categorized information based on the available small/medium size businesses, which operate during the day. For food trucks that operate in the night, datasets on businesses that operate in the night such as bars, nightclubs, movies and theaters can be gathered (perhaps using foursquare) and a suitable analysis can then be performed. Additionally, more data indicating economic strata of the population, weekday and weekend hotspots can all serve to narrow down the possibilities and help increase revenue.

Conclusion

Thus, an attempt to answer the questions raised in the 'Business problem' section of the report has been made. Clusters of conducting the Cuban food business has been identified with the neighborhoods clearly listed out. Scope for future work has also been specified to serve the best interests of the proprietors.

Appendix

1. Neighborhoods in Cluster 1 in order of priority:

Los Angeles (South Los Angeles)
Los Angeles (Jefferson Park, Leimert Park)
Los Angeles (Hancock Park, Wilshire Center, Windsor Square)
Los Angeles (South Los Angeles, Southeast Los Angeles)
Los Angeles (Baldwin Hills, Crenshaw, Leimert Park)
Los Angeles (Hyde Park, View Park, Windsor Hills)
Athens, Los Angeles (South Los Angeles)
Los Angeles (Arlington Heights, Country Club Park, Mid-City)
Los Angeles (Fairfax, Melrose, Miracle Mile, Park La Brea, Wilshire-La Brea)
Los Angeles (West Adams)
Los Angeles (Southeast Los Angeles)
Los Angeles (Griffith Park, Hollywood, Los Feliz)
Los Angeles (Southeast Los Angeles, Watts), Willowbrook
Los Angeles (Hancock Park, Western Wilton, Wilshire Center, Windsor Square)
Los Angeles (Hancock Park, Koreatown, Wilshire Center, Wilshire Park, Windsor Square)
Los Angeles (Byzantine-Latino Quarter, Harvard Heights, Koreatown, Pico Heights)
Los Angeles (Southeast Los Angeles, Watts)
Los Angeles (South Los Angeles), Florence-Graham
Los Angeles (East Hollywood)
Los Angeles (Hollywood)

2. Neighborhoods in Cluster 2 in order of priority:

Los Angeles (Sawtelle, West Los Angeles)
Los Angeles (Westwood)
Los Angeles (Cheviot Hills, Rancho Park)
Los Angeles (Bel Air Estates, Brentwood)
Los Angeles (Century City)

Culver City, Los Angeles (Mar Vista)
Los Angeles (Hollywood), West Hollywood
Los Angeles (Hollywood, Melrose), West Hollywood
Los Angeles (Los Angeles International Airport, Westchester)
Los Angeles (Bel Air Estates, Beverly Glen)
Los Angeles (Fairfax, Melrose, Miracle Mile, Park La Brea, Wilshire-La Brea)

3. Neighborhoods in Cluster 3 in order of priority:

Los Angeles (Highland Park)
Los Angeles (El Sereno, Monterey Hills, University Hills)
Los Angeles (Eagle Rock)
Los Angeles (Lincoln Heights, Montecito Heights)
Los Angeles (Boyle Heights)
Los Angeles (Cypress Park, Glassell Park, Mt Washington)
Commerce, East Los Angeles
Los Angeles (Downtown Civic Center, Chinatown, Arts District, Bunker Hill, Historic Core, Little Tokyo)
City Terrace, Los Angeles (Boyle Heights)
East Los Angeles