IBM DS Capstone Project

Applied Data Science Capstone by IBM/Coursera

by Taras Tsukarev

May 02, 2019

Introduction

Belgium is known all over the world for making unbeatable chocolates. It is paradise for the chocolate lovers. The country has a long and illustrious history of chocolate making. With around 2,000 chocolate companies and shops all over Belgium, the country remains one of the reigning producers and exporters of chocolate in the world. Based on available figures, Belgium exports more than 400,000 tons of chocolate with an annual turnover of over 4 billion euros.

Behind every top chocolate brand, stands a team of top chocolatiers. They use their knowledge, experience and craftsmanship to create the finest and sophisticated pralines, using the best products: high quality Belgian chocolate. They don’t shy away from the latest innovation and technological developments in the chocolate sector. And that makes them award-winning in several international competitions like the Patisserie World Cup.

1.1. Business Problem

A successful Belgian chocolatier is going to expand his business into the United States. Los Angeles is decided to be the starting point to open a new Belgian coffee shop combined with chocolate shop. Since Los Angeles is so big and has lots of different coffee shops and chocolate chops developed by famous brands, my client needs deeper insight from available data in order to decide where to establish his first Belgian coffee shop in the US. Another problem is that LA has very high lease rents for retail property.

To solve this business problem, we are going to cluster LA neighborhoods in order to recommend venues and the current average rent of lease in order business owner could make a decision to start a coffee shop. For this purpose, we will try to find the optimal solution in terms of competitive location, comfortable lease rents, as well as surrounding venues.

1.2. Problem Discussion

Let's discuss the above mentioned problem statements. First of all, we know that our client, famous Belgian chocolatier, wants to lease a retail place for his unique coffee shop combined with chocolate shop. Also he needs to find out the level of competition - how many coffee shops and restaurants are there in different neighborhoods. If there are more than 2-5 coffee shops / café / dessert Shop in a neighborhood, then that would be a great risk to open new coffee shops in that neighborhood. Selecting a place where there is less or no coffee shops / café / dessert shop would be of great choice, considering the lease rent of neighborhood too.

Places like Downtown, Movie theatre, Parks, Malls & Gas stations would help his business running.

1.3. Target Audience

The target audience is broad, it ranges from any company which is going to open new business entity in LA, tourists and those who are passionate about coffee shops with wide range of Belgian chocolate.

2. Data

This project will rely on public data from real estate agencies and Foursquare.

For this project we just need to analyze the current lease rent range. So I collect the lease rent data from open sources like https://www.rentcafe.com/average-rent-market-trends/us/ca/san-francisco/ and https://www.zillow.com/research/data/ according to neighborhoods, so that it's easy for us to check the lease rent data. Prepared data I have uploaded on my github repository.

Los Angeles is really large city (has more than 100 neighborhoods) and due to the limitations in the number of calls for the Foursquare API, we're going to analyze only 50 neighborhoods excluding known in advance the most expensive locations like Santa Monica, North of Montana, Pacific Palisades, etc.

The Foursquare API will be used to obtain the geographical location data for Los Angeles. These will be used to explore the venues in the neighborhoods of LA. The venues will provide the categories needed for the analysis and eventually, these will be used to determine the viability of selected locations for the Belgian coffee shop.

The venues will provide the categories needed for the analysis and eventually, these will be used to determine the viability of selected locations for the Belgian coffee shop.

The data from the lease rent dataset and location, as well as Foursquare will be explored by considering the venues within the neighborhoods of LA. These neighborhoods' coffee shops / restaurants would be checked in terms of the types of coffee shops / café / dessert Shop within a certain mile radius and the size of lease rent. Due to Foursquare restrictions, the number of venues will be limited to 100 venues. The proximity to Downtown, Movie theatre, Parks, Malls & Gas stations and other amenities would be considered.

Explore and Understand Data

After importing the necessary libraries, we download the data from my Github repository.

	State	City	Neighborhood	Average Rent (per SqFoot)
0	CA	Los Angeles	Reseda	2.03
1	CA	Los Angeles	Eagle Rock	2.05
2	CA	Los Angeles	Vermont - Slauson	2.06
3	CA	Los Angeles	Van Nuys	2.10
4	CA	Los Angeles	Tarzana	2.11
5	CA	Los Angeles	Gramercy Park	2.13
6	CA	Los Angeles	Mount Washington	2.14
7	CA	Los Angeles	Baldwin Hills - Crenshaw	2.15
8	CA	Los Angeles	Montecio Heights	2.20
9	CA	Los Angeles	West Hills	2.21

Figure 1. First ten rows from downloaded table.

In obtaining the location data of the locations, the Geocoder package is used with the arcgis_geocoder to obtain the latitude and longitude of the needed locations.

These will help to create a new dataframe that will be used subsequently for LA neighborhoods.

	State	City	Neighborhood	Average Rent (per SqFoot)	Latitude	Longitude
0	CA	Los Angeles	Reseda	2.03	34.193840	-118.547540
1	CA	Los Angeles	Eagle Rock	2.05	34.139270	-118.210870
2	CA	Los Angeles	Vermont - Slauson	2.06	33.989175	-118.237705
3	CA	Los Angeles	Van Nuys	2.10	34.184390	-118.446520
4	CA	Los Angeles	Tarzana	2.11	34.175290	-118.550100
5	CA	Los Angeles	Gramercy Park	2.13	34.033900	-118.312580
6	CA	Los Angeles	Mount Washington	2.14	34.099040	-118.211340
7	CA	Los Angeles	Baldwin Hills - Crenshaw	2.15	34.011570	-118.336460
8	CA	Los Angeles	Montecio Heights	2.20	34.091980	-118.201010
9	CA	Los Angeles	West Hills	2.21	34.200360	-118.629330

Figure 2. LA neighborhoods with the latitude and longitude.

Also we created a function to get the top 100 venues in every neighborhood within a radius of 500 meters.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Reseda	34.19384	-118.54754	YMCA West Valley	34.193438	-118.543146	Gym
1	Eagle Rock	34.13927	-118.21087	Milkfarm	34.138996	-118.212384	Deli / Bodega
2	Eagle Rock	34.13927	-118.21087	The Oinkster	34.139458	-118.210484	American Restaurant
3	Eagle Rock	34.13927	-118.21087	Four Cafe	34.139047	-118.212857	American Restaurant
4	Eagle Rock	34.13927	-118.21087	One Down Dog	34.139031	-118.213691	Yoga Studio
5	Eagle Rock	34.13927	-118.21087	Taco Spot	34.139144	-118.210796	Mexican Restaurant
6	Eagle Rock	34.13927	-118.21087	5 Line Tavern	34.138892	-118.213333	Bar
7	Eagle Rock	34.13927	-118.21087	Room 31	34.138766	-118.213341	Speakeasy
8	Eagle Rock	34.13927	-118.21087	Snow Station	34.139026	-118.212525	Ice Cream Shop
9	Eagle Rock	34.13927	-118.21087	Leanna Lin's Wonderland	34.137762	-118.214294	Gift Shop

Figure 3. Fragment of LA neighborhoods with venues.

Next step is to visualize Los Angeles neighborhoods.

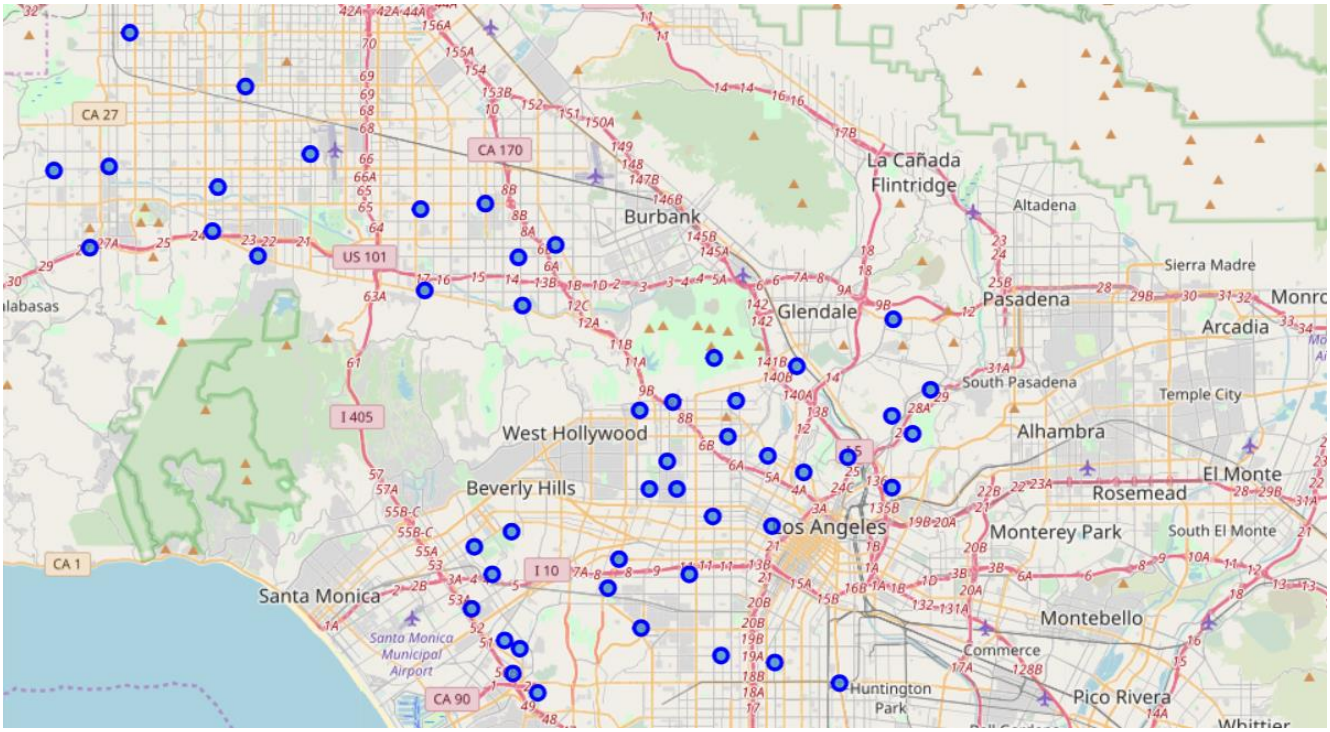

Figure 4. Map of selected LA neighborhoods.

Now let's create the new dataframe and display the top 10 venues for each neighborhood:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Atwater Village	Coffee Shop	Vietnamese Restaurant	Juice Bar	Pub	Italian Restaurant	Mexican Restaurant	Bookstore	Boutique	Mediterranean Restaurant	Sporting Goods Shop
1	Baldwin Hills - Crenshaw	Fast Food Restaurant	Department Store	Shoe Store	Southern / Soul Food Restaurant	Sandwich Place	Lingerie Store	Chinese Restaurant	Mexican Restaurant	Women's Store	Mobile Phone Shop
2	Beverlywood	Museum	Seafood Restaurant	Mobile Phone Shop	Cosmetics Shop	Pharmacy	Grocery Store	Japanese Restaurant	Coffee Shop	Electronics Store	Dumpling Restaurant
3	Canoga Park	Indian Restaurant	Pet Store	Rental Car Location	Burger Joint	Asian Restaurant	Bakery	Theater	Mexican Restaurant	Fried Chicken Joint	Big Box Store
4	Chatsworth	Fast Food Restaurant	Japanese Restaurant	Breakfast Spot	Assisted Living	Pharmacy	Diner	Mexican Restaurant	Food & Drink Shop	Sporting Goods Shop	Sushi Restaurant

Figure 5. Top 10 venues for each neighborhood.

Modeling.

Cluster Neighborhoods.

Run k-means to cluster the neighborhood into 5 clusters:

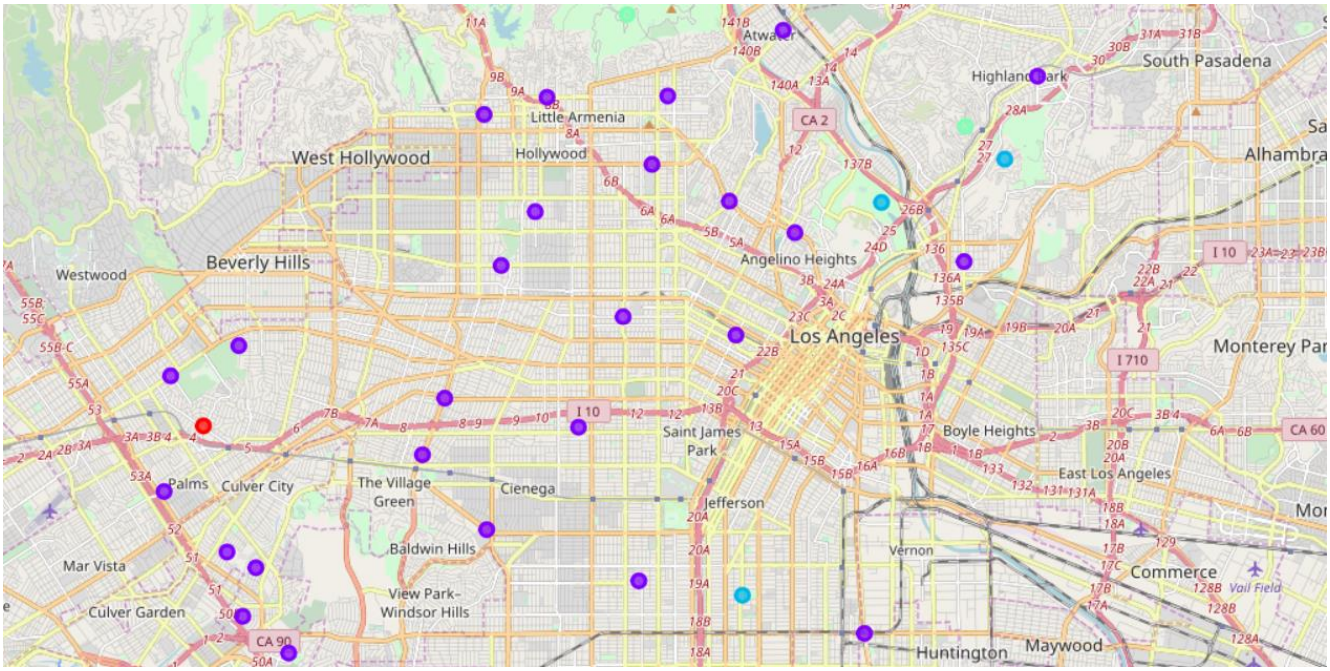

Figure 6. Map of neighborhood clusters.

Then we created a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood. After that, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster.

3. Methodology

In this project we will direct our efforts on detecting areas of Los Angeles that have low coffee shops / café / dessert shop density. We will limit our analysis to area 500 meters around center of neighborhood as well as get the top 100 venues in every neighborhood.

In first step we have collected the required data: location and type of every venue within 500m around center of neighborhood.

Second step in our analysis will be exploration of 'coffee shops density' across different neighborhoods of LA - we will identify a few promising neighborhoods with low number of shops / café / dessert shop in general and focus our attention on those areas.

In third and final step we will focus on most promising neighborhoods and within those create clusters of locations that meet some basic requirements established in discussion with our client: we will take into consideration locations with no more than 2-5 coffee shops / café / dessert shop in radius of 500 meters, and we want appropriate lease rent. Then we will present map of all such locations but also create clusters (using k-means clustering) of those locations.

4. Results and Discussion

Our analysis shows that although there is a great number of coffee shops / café / dessert shop in Los Angeles, in some areas it was found that there are of low coffee shop density. The highest concentration of coffee shops / café / dessert shop as well as different kinds of restaurants was detected in Cluster 2. At the same time not all neighborhoods in Cluster 2 have enough quantity of coffee shops but they have necessary amenities for creating coffee shops (parks, hotels, hostels, etc.). Considering the various amenities in Cluster 2, you must also consider the amount of lease rent. So, the most attractive neighborhoods in Cluster 2 are Vermont-Slauson, West Hills, Vermont Square, Canoga Park. The average lease rents in this areas are acceptable (in range from 2.06 till 2.38).

In Cluster 5 we identified potentially interesting neighborhood, Reseda, which offer a combination of interesting venues - Gym, Flower Shop, Flea Market, Financial or Legal Service, Film Studio. The average lease rent in Reseda is very attractive - 2.03.

Another attractive areas were found in Cluster 3 - Montecio Heights, South Park. In these neighborhoods there are park zones, markets, stores, and almost no any coffee shops / café / dessert shops. The average lease rents are 2.20 and 2.31 respectively.

Cluster 4 also has two acceptable neighborhoods - Mount Washington and Porter Ranch. In these neighborhoods we didn't detected many coffee shops. So, this is a very good result.

Finally, Cheviot Hills from Cluster 1 due to its very high lease rent is not selected.

Result of all this is 9 neighborhoods containing largest number of potential new coffee shops locations based on number of and distance to existing venues - Downtown, Movie theatre, Parks, Malls & Gas stations. This, of course, does not imply that those neighborhoods are actually optimal locations for a new Belgian coffee shop! Purpose of this analysis was to only provide info on areas with acceptable lease rents but not crowded with existing coffee shops /

café / dessert shops - it is entirely possible that there is a very good reason for small number of coffee shops in any of those areas, reasons which would make them unsuitable for a new coffee shop regardless of lack of competition in the neighborhood. Recommended neighborhoods should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

5. Conclusion

Purpose of this project was to identify Los Angeles neighborhoods with low number of coffee shops / café / dessert shop in order to aid client in narrowing down the search for optimal location for a new Belgian coffee shops combined with Belgian chocolate shop. By analyzing coffee shops / café / dessert shop density distribution from Foursquare data we have first identified general neighborhoods that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by client.

Final decision on optimal coffee shop location will be made by the client based on specific characteristics of neighborhoods in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to Downtown, Movie theatre, Parks, Malls & Gas), proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.