Coursera Capstone

IBM Applied Data Science Capstone

Predicting the Success of a Café in Los Angeles, California

By: George Freund

May 2020

1. Introduction

1.1 Background

Los Angeles is famous for its beautiful beaches, luxurious lifestyles, and a film and television industry that has shaped popular culture around the world for the past 100 years. It takes a lot of caffeine to fuel those cinematic ideas, and one might be correct to assume that Hollywood is saturated with cafés to provide such liquid inspiration. But the county of Los Angeles extends well beyond its movie star neighborhood. As the most populous county in the United States of America — with over 10 million residents as of 2018 — there are plenty of thirsty locals that would gladly welcome a café in their neck of the woods.

1.2 Business Problem

Predicting which cities in Los Angeles County could benefit from a new java joint can be the difference between an A-list business venture, or a D-list dud. From beach communities to mountain towns, sprawling suburbs to a vibrant downtown city, Los Angeles County is home to a diverse number of demographics including population, culture, income and education — all of which can factor into where a prospective owner will start a business. For the scope of this capstone project, however, we will be looking only at the competition aspect of opening a café in each neighborhood, using venue data provided by Foursquare.

1.3 Target Audience of this Project

This project is particularly useful to investors looking to open or invest in cafés in neighborhoods not typically associated with such businesses. Coffee shops have typically been a trend in neighborhoods associated with wealth and a certain demographic of people. But the industry is growing and adapting to different neighborhoods recently, as discussed in a January 2020 article on kcrw.com, that highlights coffee shops in gentrifying neighborhoods.

2. Data acquisition and cleaning

2.1 Data sources

A list of cities within Los Angeles County can be found here on Wikipedia. This dataset does not provide geographical coordinates however, so that will be attained using Python's geocoder attribute.

The venue data used to locate existing cafés and visualize these competitors will be attained from Foursquare's database.

2.2 Data cleaning

The list of cities in Los Angeles County will be obtained from its Wikipedia site using the beautifulsoup package for Python. Once scraped, the data will then be merged with location information provided by Python's geocoded package.

Once we have the cities' geographical locations, we can then call on Foursquare's API to obtain café venues in the listed cities. From Foursquare's API, I will retrieve the following for each venue:

- **Name:** The name of the venue
- **Category:** The category type defined by the API
- **Latitude:** The latitude value of the venue
- **Longitude:** The longitude value of the venue

Once merged, we can then visualize and cluster the data to help prospective stakeholders decide which locations would be areas of interest for opening a café.

3. Exploratory Data Analysis

3.1 Web scraping

Scraping the data provided a database comprised of “City”, “Date Incorporated” and “Population” columns. I dropped the “Date Incorporated” column, but kept the “Population” column as that might provide some useful information for later. Below is the database before dropping the “Date Incorporated” column.

	City	Date incorporated	Population as of(2010 Census)
0	Agoura Hills	December 8, 1982	20330
1	Alhambra	July 11, 1903	83653
2	Arcadia	August 5, 1903	56364
3	Artesia	May 29, 1959	16522
4	Avalon	June 26, 1913	3728
5	Azusa	December 29, 1898	46361
6	Baldwin Park	January 25, 1956	75390
7	Bell	November 7, 1927	35477
8	Bell Gardens	August 1, 1961	42072
9	Bellflower	September 3, 1957	76616

3.2 Retrieve coordinates

The next step was to grab the coordinates of the cities returned in the database. I did this using Python’s “geocoder” attribute. Once the coordinates were retrieved, I merged them with the list of cities and created the new database head seen below.

	City	Population as of(2010 Census)	Latitude	Longitude
0	Agoura Hills	20330	34.14611	-118.77812
1	Alhambra	83653	34.09370	-118.12727
2	Arcadia	56364	34.13614	-118.03887
3	Artesia	16522	33.86114	-118.07968
4	Avalon	3728	33.34411	-118.32139
5	Azusa	46361	34.13361	-117.90589

3.3 Map the data

I mapped the location of the cities in Los Angeles County using Python's folium attribute.

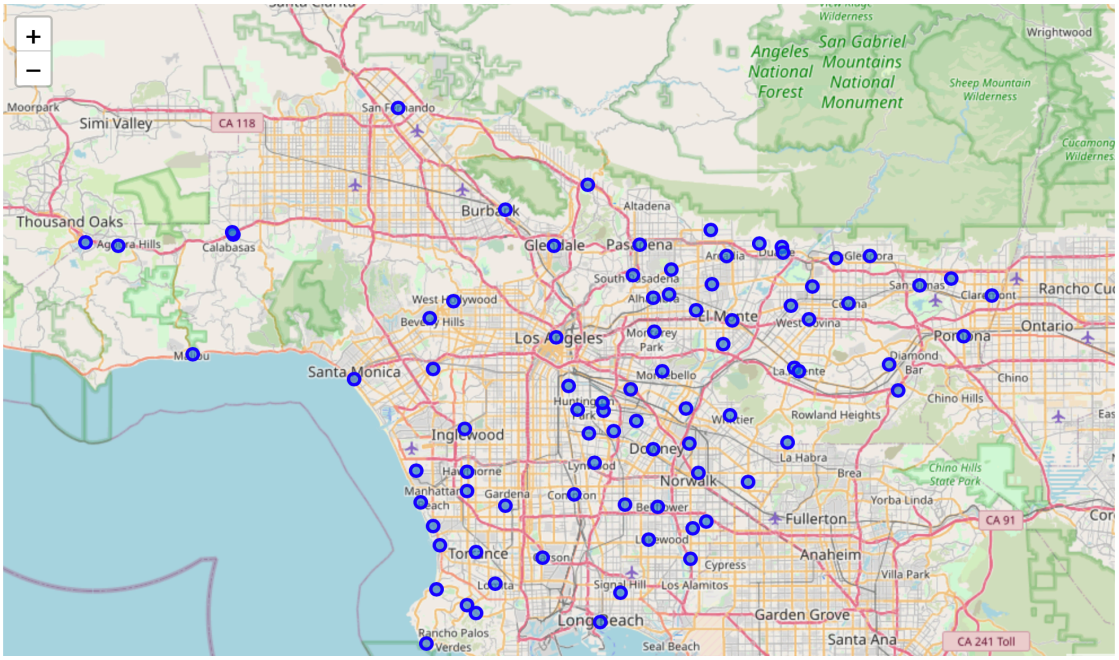

Figure 1: list of cities in Los Angeles County

3.4 Retrieve Venue Data

I then used Foursquare's API to retrieve a list of venues matching these coordinates. I merged the data and created a new dataframe that included the columns "City", "Latitude", "Longitude", "Venue Name", Venue Latitude", "Venue Longitude", and "Venue Category".

	City	Latitude	Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Agoura Hills	34.14611	-118.77812	Future Track Running Center	34.145819	-118.779251	Sporting Goods Shop
1	Agoura Hills	34.14611	-118.77812	Twisted Oak Tavern	34.145308	-118.778679	Gastropub
2	Agoura Hills	34.14611	-118.77812	Cafe Bizou	34.148410	-118.782587	French Restaurant
3	Agoura Hills	34.14611	-118.77812	Grissini Ristorante	34.145815	-118.778534	Italian Restaurant
4	Agoura Hills	34.14611	-118.77812	Pizza Nosh	34.148311	-118.782181	Pizza Place
5	Agoura Hills	34.14611	-118.77812	Forest Cove Park	34.152290	-118.774749	Park

3.5 Clean Venue Data

I then analyzed each neighborhood and got a list of 351 unique categories of venues. I grouped the dataframe by the “City” column, and by taking the mean of the frequency of occurrence in each category.

I then created a new dataframe that grouped the info for Café data only…

	City	Café
0	Agoura Hills	0.010000
1	Alhambra	0.040000
2	Arcadia	0.000000
3	Artesia	0.040000
4	Avalon	0.014286
5	Azusa	0.011111
6	Baldwin Park	0.011628
7	Bell	0.000000
8	Bell Gardens	0.000000
9	Bellflower	0.000000

3.6 K-means

Now that I had my Café dataframe, I clustered the data using K-means. After several iterations, I determined that a cluster set of 4 was appropriate. I then merged the cluster data with the top 10 venues for each city and created a final dataframe.

	City	Café	Cluster Labels	Population as of(2010 Census)	Latitude	Longitude
0	Agoura Hills	0.010000	0	20330	34.14611	-118.77812
29	Glendora	0.014286	0	50073	34.13602	-117.86452
32	Hermosa Beach	0.010000	0	19506	33.86404	-118.39535
33	Hidden Hills	0.014706	0	1856	34.15918	-118.64025
34	Huntington Park	0.010000	0	58114	33.98143	-118.21914
36	Inglewood	0.010000	0	109673	33.96178	-118.35674

3.7 Visualize and Cluster Data

The final step was to visualize the clustered data in a map, using Python's Folium attribute, then examine each of the clusters.

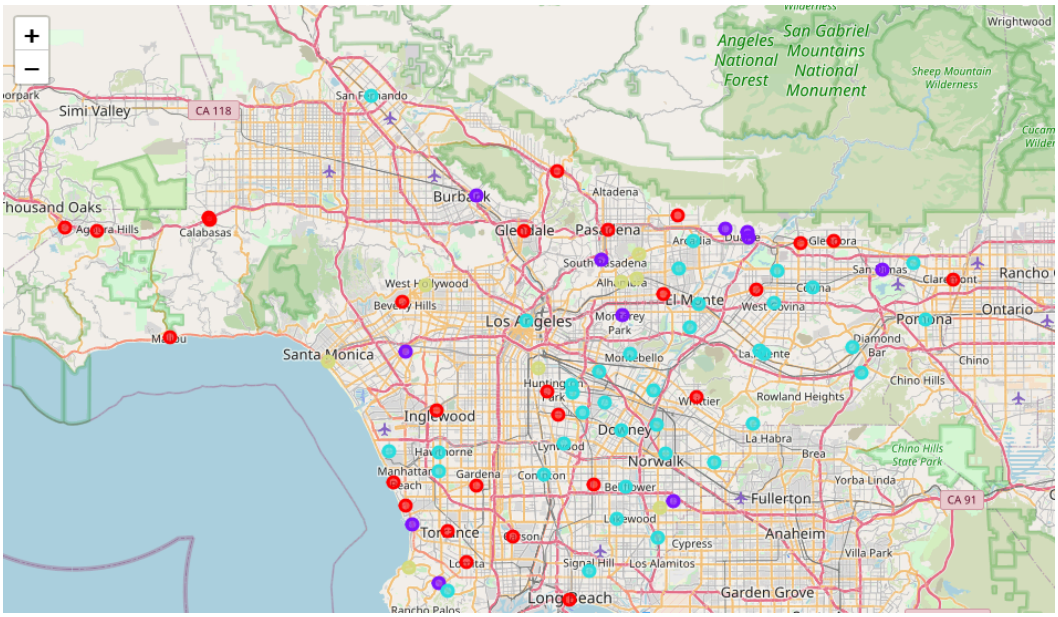

Figure 2: Venue clusters in Los Angeles County

4. Results and Discussion

As previously mentioned, Los Angeles County is a sprawling area with over 10 million residents, which means that both opportunity and competition exist hand in hand. Several factors outside the scope of this project can come into play in deciding where to open a café; including population, income and demographic of each city, though that is changing due to gentrification and social trends.

In the future, the data set can be expanded upon with the factors mentioned above, which can then bring more analysis and prediction modeling, including linear regression and other classification models, to make a more informed decision on where to open a café in Los Angeles.

6. Conclusion

After clustering data from the Foursquare API, we can see that cluster 2 is compromised of cities without cafés. This represents a great opportunity and high potential areas to open new cafés as there is no competition present. Meanwhile, cafés in cluster 3 are likely suffering from intense competition due to high concentration. Therefore, this project recommends property developers to open new cafés in cities in cluster 2 to avoid competition. Property developers with unique selling propositions to stand out from the competition can also open new cafés in cities in cluster 0 with minimal competition. Lastly, property developers are advised to avoid cities in cluster 1 and cluster 3, which already have a moderate to high number of cafés.