Predicting the best location for a new restaurant in Los Angeles

IBM Data Science Capstone Project

Mohammed Essam Madbouli

September 12, 2020

I. Introduction

The IBM Data Science Professional certificate course on Coursera concludes with a Capstone Project. This project is about using data science toolset on a real-life problem and demonstrating the creation of value by applying the learned skills. This report presents this capstone project. The analysis was performed in Python.

II. Problem Definition

a. The Problem

For this project, I chose a hypothetical business problem. The question that we are trying to answer is the following.

A successful owner of multiple mid to high-end restaurants decided to open a new restaurant in Los Angeles, California. Having visited the city many times in recent years, he couldn't disregard the big boom in gastronomy. He is keen on opening a new unit, which will focus on the American and Mexican fusion kitchen.

Considering the price level at which the restaurant will operate, the intent is to find an optimal location in an area, where gastronomy is booming, and which is easily accessible for tourists and for wealthier local citizens as well.

b. Assumptions and business logic

The assumption behind the analysis is that we can use unsupervised machine learning to create clusters of neighborhoods that will provide us with a list of areas for consideration for the restaurant. The intent is that the restaurant to be situated close to one of the gastronomical centers and touristic hotspots.

c. Audience

While here we are assuming a concrete business owner to whom we are addressing this report, but this restaurant owner can be treated as a persona and thus this analysis could be useful for a group of market players (restaurant owners).

III. Data

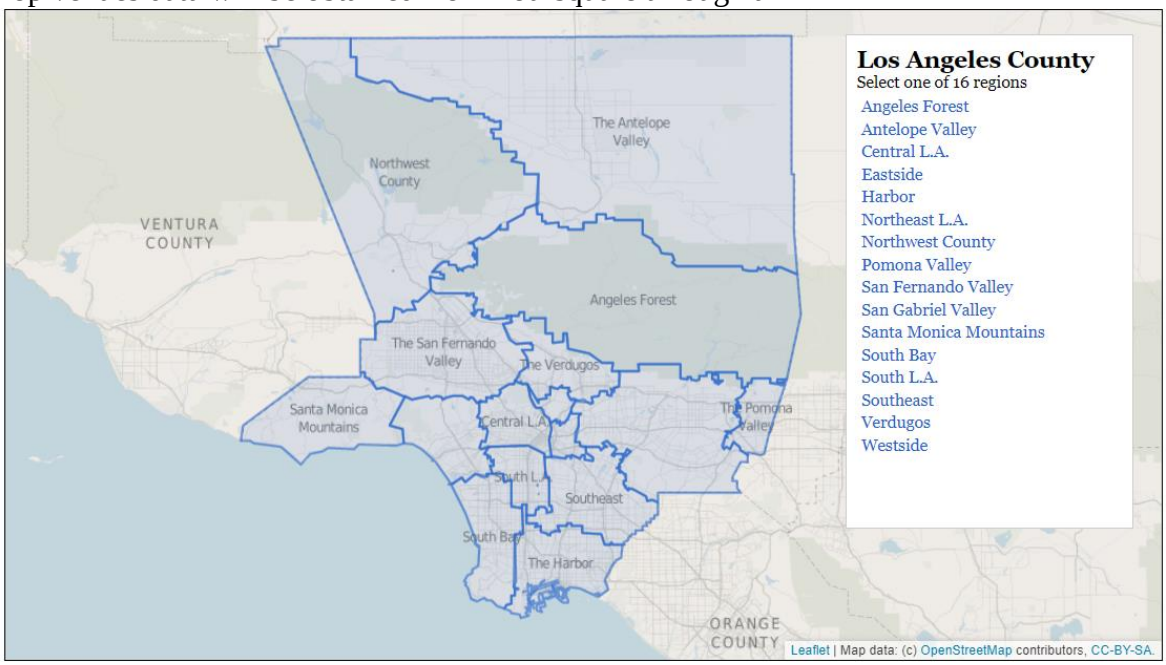
To perform this analysis, we will need the following data:

1. Dataset of the neighborhoods of LA with their Geo-coordinates
2. Top venues of the neighborhoods

List of neighborhoods will be obtained from Los Angeles Times (http://maps.latimes.com/neighborhoods/)

Geo-coordinates of neighborhoods will be obtained with the help of the ARCGIS geocoder tool.

Top venues data will be obtained from Foursquare through an API



IV. Methodology

a. Use of data and a high-level roadmap

After tidying up and exploring the data, we will apply the K-means machine learning technique for creating clusters of neighborhoods. We will use the silhouette score for choosing the optimal number of clusters.

b. Analysis

- Data Preparation and exploration

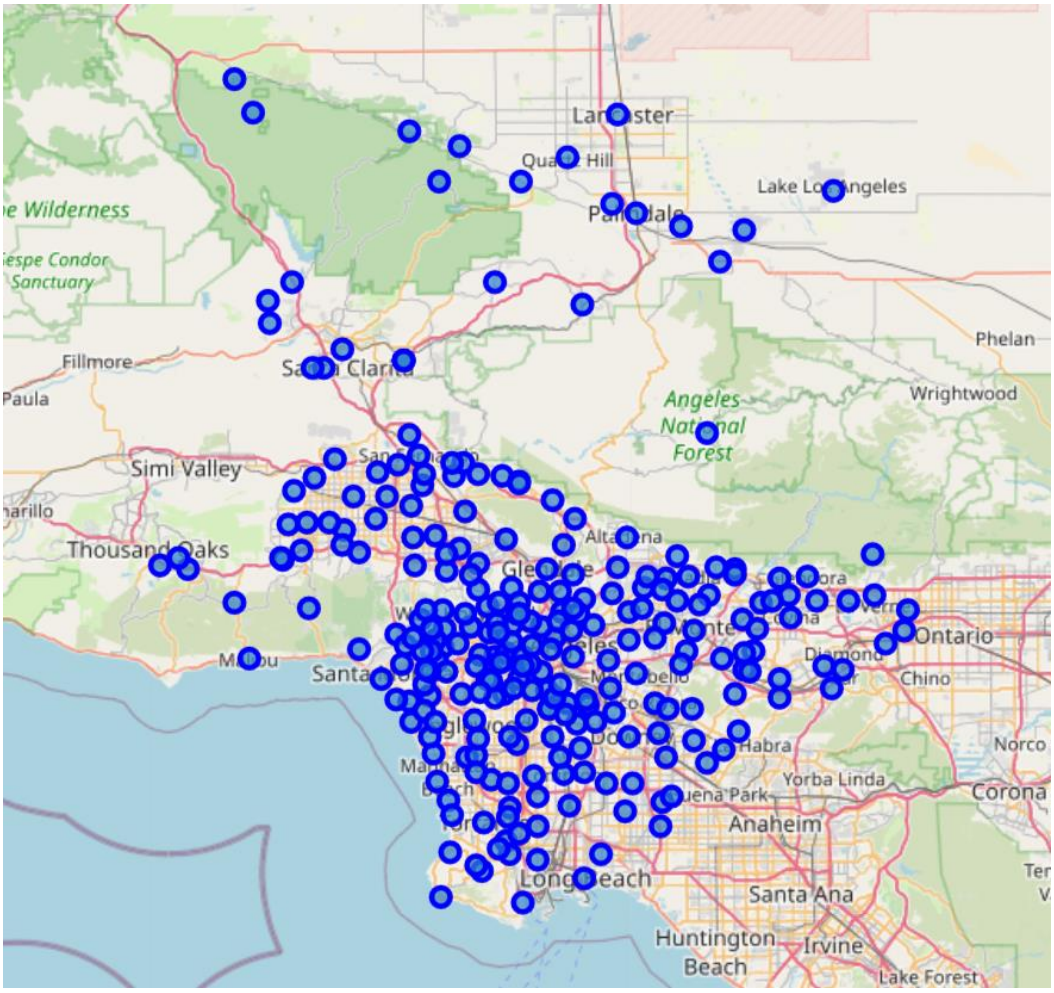
As part of preparing the data, we start by creating a list of neighborhoods in LA and add the geo-coordinates of each district to this table. That is done by first importing a list of neighborhoods and then using this list and geocode python library, we add the latitude and longitude coordinates to each district. After performing this task, we get the following table that we use in pandas dataframe format.

	NEIGHBORHOOD	DISTRICT	LATITUDE	LONGITUDE
0	Acton	Antelope Valley	34.46815	-118.19513
1	Adams-Normandie	South L.A.	34.07809	-118.30120
2	Agoura Hills	Santa Monica Mountains	34.14611	-118.77812
3	Agua Dulce	Northwest County	34.49570	-118.32621
4	Alhambra	San Gabriel Valley	34.09370	-118.12727

LA has 272 neighborhoods, so this is a real dataset.

In the next step, we create a visual representation of how the neighborhoods are situated in LA.

For this, the folium library was used.



In the next step of the analysis, the neighborhoods were explored in greater detail. It means venues were collected for each district via Foursquare API. The data from Foursquare is received in json format. After arranging the data, we have up to 100 venues for each district. Venues are collected within a radius of 1000 meters from the point of district coordinates. The collected and arranged data looks like this. The following table shows some venues from the first district.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Acton	34.46815	-118.19513	Acton Market & Country Store	34.468595	-118.197626	Grocery Store
1	Acton	34.46815	-118.19513	Fox Hay Feed and Grain	34.469565	-118.195481	Pet Store
2	Acton	34.46815	-118.19513	Acton Market	34.467628	-118.195892	Grocery Store
3	Acton	34.46815	-118.19513	TSW Social Media Marketing	34.470898	-118.192307	Market
4	Acton	34.46815	-118.19513	specialty truss	34.470898	-118.192307	Construction & Landscaping

We can check how many venues have been collected for each district. The following table gives that summary.

Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Acton	8	8	8	8	8	8
Adams-Normandie	83	83	83	83	83	83
Agoura Hills	17	17	17	17	17	17
Agua Dulce	11	11	11	11	11	11
Alhambra	73	73	73	73	73	73
Alondra Park	40	40	40	40	40	40
Altadena	29	29	29	29	29	29
Arcadia	52	52	52	52	52	52
Arleta	13	13	13	13	13	13
Arlington Heights	33	33	33	33	33	33
Artesia	100	100	100	100	100	100
Athens	19	19	19	19	19	19
Atwater Village	100	100	100	100	100	100
Avalon	71	71	71	71	71	71
Avocado Heights	19	19	19	19	19	19
Azusa	36	36	36	36	36	36
Baldwin Hills/Crenshaw	81	81	81	81	81	81
Baldwin Park	46	46	46	46	46	46
Bel-Air	8	8	8	8	8	8
Bell	40	40	40	40	40	40

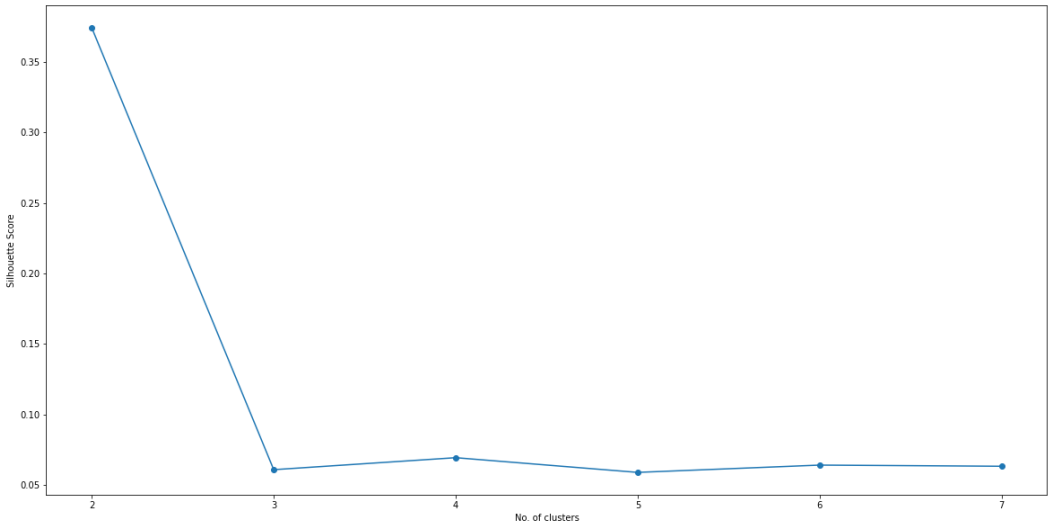
For analyzing the neighborhoods, we focus on venue categories. For that purpose, we use the one-hot encoding. This creates dummy variables for categories so the data set could be used for machine learning.

After performing manipulations with the dataset, we get the following table, which shows the top ten most common venues for each district (first five shown in the table).

	NEIGHBORHOOD	DISTRICT	LATITUDE	LONGITUDE	Cluster Labels	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
0	Acton	Antelope Valley	34.46815	-118.19513	1.0	Park	Grocery Store	Nature Preserve	Construction & Landscaping	Market	Pet Store	Deli / Bodega	Department Store
1	Adams-Normandie	South L.A.	34.07809	-118.30120	0.0	Thai Restaurant	Coffee Shop	Korean Restaurant	Bar	Seafood Restaurant	Pizza Place	Sandwich Place	Latin American Restaurant
2	Agoura Hills	Santa Monica Mountains	34.14611	-118.77812	0.0	Park	Ramen Restaurant	Fast Food Restaurant	Casino	Car Wash	Health & Beauty Service	General College & University	Laundry Service
3	Agua Dulce	Northwest County	34.49570	-118.32621	0.0	Bakery	Park	Grocery Store	Home Service	Mexican Restaurant	Gift Shop	Café	Pizza Place
4	Alhambra	San Gabriel Valley	34.09370	-118.12727	0.0	Ice Cream Shop	Bank	Bakery	Bubble Tea Shop	Gym	Park	Seafood Restaurant	Fast Food Restaurant

- Clustering

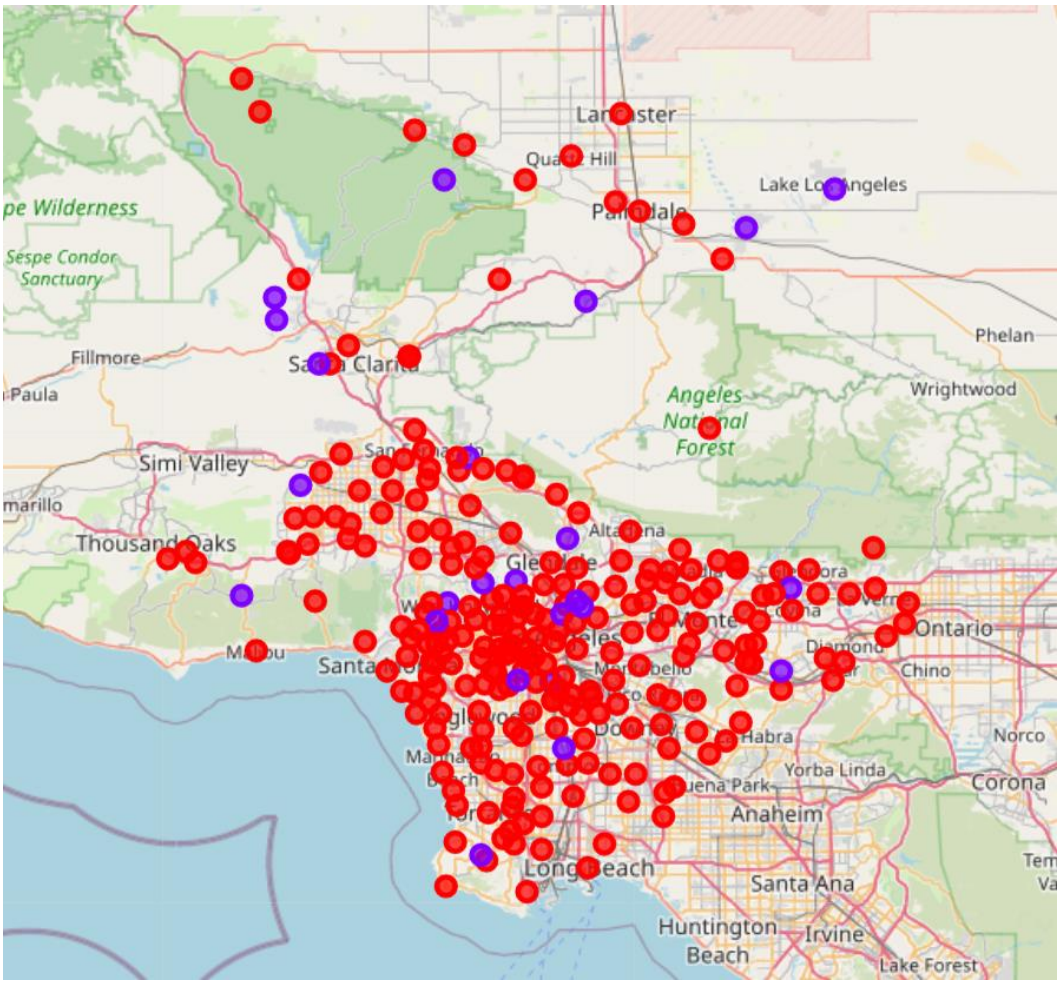
Now that we have the dataset ready, we perform clustering. For this, unsupervised machine learning technique will be used based on K-means. For K-means clustering, we need to decide the number of clusters that we want to use. To avoid the trial and error approach, the silhouette score was used. The following graph shows the silhouette scores for a range of clusters variations.



From the graph, we can read that the optimal number of clusters to use is 2 (where the score is the highest). In the next step, we run the K-means clustering algorithm with the parameter of 2 as the number of clusters. When done, we add the cluster labels to the dataset. We get the following table.

	NEIGHBORHOOD	DISTRICT	LATITUDE	LONGITUDE	Cluster Labels	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
0	Acton	Antelope Valley	34.46815	-118.19513	1.0	Park	Grocery Store	Nature Preserve	Construction & Landscaping	Market	Pet Store	Deli / Bodega	Department Store
1	Adams-Normandie	South L.A.	34.07809	-118.30120	0.0	Thai Restaurant	Coffee Shop	Korean Restaurant	Bar	Seafood Restaurant	Pizza Place	Sandwich Place	Latin American Restaurant
2	Agoura Hills	Santa Monica Mountains	34.14611	-118.77812	0.0	Park	Ramen Restaurant	Fast Food Restaurant	Casino	Car Wash	Health & Beauty Service	General College & University	Laundry Service
3	Agua Dulce	Northwest County	34.49570	-118.32621	0.0	Bakery	Park	Grocery Store	Home Service	Mexican Restaurant	Gift Shop	Café	Pizza Place
4	Alhambra	San Gabriel Valley	34.09370	-118.12727	0.0	Ice Cream Shop	Bank	Bakery	Bubble Tea Shop	Gym	Park	Seafood Restaurant	Fast Food Restaurant

Also, we can visualize the clusters on the map that we created earlier.



c. Limitations

The analysis has some limitations that should be taken into account:

1. The analysis is performed on a neighborhood level.
2. When collecting venues, a 1000-meter radius is used around the center coordinates of the neighborhoods.
3. The number of collected venues is limited to 100 per neighborhood.

V. Results

Understanding the Clusters

By looking at the cluster data, we can see that cluster 1 is the one that we are the most interested in.

i. Cluster 1

The first cluster (Cluster label 0) is the biggest cluster, but this is where we see lots of gastronomy related venues (coffee shop, pizza place, American Restaurant, bar, Mexican Restaurant, etc..).

	DISTRICT	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
1	South L.A.	Thai Restaurant	Coffee Shop	Korean Restaurant	Bar	Seafood Restaurant	Pizza Place	Sandwich Place	Latin American Restaurant	Bakery	Grocery Store
2	Santa Monica Mountains	Park	Ramen Restaurant	Fast Food Restaurant	Casino	Car Wash	Health & Beauty Service	General College & University	Laundry Service	Pizza Place	Gastropub
3	Northwest County	Bakery	Park	Grocery Store	Home Service	Mexican Restaurant	Gift Shop	Café	Pizza Place	Business Service	Restaurant
4	San Gabriel Valley	Ice Cream Shop	Bank	Bakery	Bubble Tea Shop	Gym	Park	Seafood Restaurant	Fast Food Restaurant	Shoe Store	Café
5	South Bay	Fast Food Restaurant	Mexican Restaurant	Pizza Place	Coffee Shop	Restaurant	Italian Restaurant	Baseball Field	Mediterranean Restaurant	Park	Chinese Restaurant
6	Verdugos	Grocery Store	Coffee Shop	Hardware Store	Gym / Fitness Center	Automotive Shop	Scenic Lookout	Pet Store	Pharmacy	Bakery	Bank
8	San Gabriel Valley	Racetrack	American Restaurant	Food Truck	Mexican Restaurant	Coffee Shop	Sandwich Place	Breakfast Spot	Bar	Convenience Store	Park

ii. Cluster 2

Cluster 2 (Cluster label 1) is neighborhoods where public travel rated at top, but behind that parks, playgrounds are also present. These are mainly areas with family houses where people live, but not really the vibrant, lively part of the city.

	DISTRICT	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	Antelope Valley	Park	Grocery Store	Nature Preserve	Construction & Landscaping	Market	Pet Store	Deli / Bodega	Department Store	Exhibit	Eye Doctor
23	Westside	Park	Clothing Store	Other Great Outdoors	Zoo	Film Studio	Event Space	Exhibit	Eye Doctor	Fabric Shop	Falafel Restaurant
35	Central L.A.	Park	Food Truck	Skate Park	Baseball Field	Trail	Food & Drink Shop	Food	Food Stand	Event Space	Exhibit
38	South L.A.	Food	Park	Taco Place	Burger Joint	Bus Line	Construction & Landscaping	Food Truck	Flea Market	Mexican Restaurant	Light Rail Station
43	San Fernando Valley	Trail	Park	Gourmet Shop	Business Service	Film Studio	Event Space	Exhibit	Eye Doctor	Fabric Shop	Falafel Restaurant
47	San Gabriel Valley	Park	Bakery	Home Service	Donut Shop	Gas Station	Coffee Shop	Mexican Restaurant	Asian Restaurant	Convenience Store	Food
74	Central L.A.	Park	Scenic Lookout	Food Truck	Mexican Restaurant	Trail	Playground	Breakfast Spot	Fast Food Restaurant	Hardware Store	Garden
88	Northwest County	Food	Playground	Home Service	Zoo	Film Studio	Event Space	Exhibit	Eye Doctor	Fabric Shop	Falafel Restaurant

VI. Discussion and Recommendations

Based on what we learned about the clusters, we can advise the restaurant owner to consider the neighborhoods from cluster 1 as a potential location for the new restaurant. These are the neighborhoods where gastronomy is well represented and also hotels are frequent. These satisfy the two original criteria that the location should be in a gastronomical center and in a location that is easily accessible for tourists.

VII. Conclusion

This paper discussed the process of coming up with an answer for a hypothetical though real-life like business problem. The analysis was performed based on the toolset of data science and relied heavily on the use of Python and Python libraries such as Pandas, Scikit, Folium to name a few. Data was collected from a different type of sources and in different formats. For analysis, machine learning technique was used. The output of the analysis provided a thorough base for the recommendation for the business problem in question.

VIII. References

The Jupyter notebook of the analysis can be found on GitHub.

https://github.com/MohammedMadbouliWorkspace/Coursera_Capstone/blob/master/notebooks/capstone_project_notebook.ipynb