CAPSTONE PROJECT

Recommendation of locations for offline stores for an ecommerce company

IBM Data Science Professional Course - Capstone Project

The Battle of Neighborhoods

Submission by Amit S. Holey, July 2019

Contents

Introduction/Business Problem

Ecommerce as an industry is now well entrenched in the daily lives of all of us. There is an ever- increasing range of products moving to online sales and this has resulted in gargantuan logistics of inventory warehouses, distribution centers, transport facilities and courier delivery mechanisms. One of the impending challenges for these ecommerce companies is meeting the promised timelines for just in time deliveries of their online orders. With rapid urbanization, last mile delivery is fast becoming an obstacle as the infrastructure is not always suitable for accommodating available modes of transports – large containers, pickup trucks, vans or other 4 wheelers.

To combat this challenge, one of the ideas being discussed is setting up of offline-centers of high frequency or fastmoving items on these ecommerce market places. These centers will be stocked with optimal quantity of these fast-moving goods and will also serve as pickup centers for shoppers who cannot commit to a delivery address. Millennials faced with house ownership issues and privacy concerns are increasingly choosing to opt for pickup centers to pick their orders themselves or through delivery agents. These offline centers will thus play a dual role for ecommerce players and will be instrumental in opening a new channel of business and go-to-market for ecommerce landscape

This problem focuses on identifying the right locations suitable for such offline stores. As a pilot, a densely populated urban area of Los Angeles in California in USA is chosen. We will explore the different areas in LA, identify the different factors that impact offline stores and use data science to group similar areas to arrive at likely areas for setting up offline stores.

Data

For getting the different localities in Los Angeles, I have used the list of zip codes associated with Los Angeles. To procure this list I downloaded the data from simplemaps.com. This list consists of zip codes, latitude, longitude and population density also.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	zip	lat	lng	city	state_i	state_r	zcta	parent	population	density	county_fips	county	all_cou	impred	militar	timezo	
29995	90001	33.974	-118.25	Los Angeles	CA	California	TRUE		57110	6295.9	06037	Los Angel	{'06037':1(	FALSE	FALSE	America/Los_Angeles	
29996	90002	33.9491	-118.247	Los Angeles	CA	California	TRUE		51223	6458.8	06037	Los Angel	{'06037':1(	FALSE	FALSE	America/Los_Angeles	
29997	90003	33.9641	-118.274	Los Angeles	CA	California	TRUE		66266	7204.7	06037	Los Angel	{'06037':1(	FALSE	FALSE	America/Los_Angeles	
29998	90004	34.0762	-118.311	Los Angeles	CA	California	TRUE		62180	7876.3	06037	Los Angel	{'06037':1(	FALSE	FALSE	America/Los_Angeles	
29999	90005	34.0591	-118.306	Los Angeles	CA	California	TRUE		37681	13421.3	06037	Los Angel	{'06037':1(	FALSE	FALSE	America/Los_Angeles	
30000	90006	34.048	-118.294	Los Angeles	CA	California	TRUE		59185	11903.1	06037	Los Angel	{'06037':1(	FALSE	FALSE	America/Los_Angeles	
30001	90007	34.0281	-118.285	Los Angeles	CA	California	TRUE		40920	6403.9	06037	Los Angel	{'06037':1(	FALSE	FALSE	America/Los_Angeles	
30002	90008	34.0096	-118.347	Los Angeles	CA	California	TRUE		32327	3398.1	06037	Los Angel	{'06037':1(	FALSE	FALSE	America/Los_Angeles	
30003	90010	34.0621	-118.316	Los Angeles	CA	California	TRUE		3800	3208.3	06037	Los Angel	{'06037':1(	FALSE	FALSE	America/Los_Angeles	
30004	90011	34.0071	-118.259	Los Angeles	CA	California	TRUE		103892	9359.3	06037	Los Angel	{'06037':1(	FALSE	FALSE	America/Los_Angeles	

Figure 1 : Zip codes for Los Angeles from simplemaps.com

For each of these zip codes, I have used an API call to get the recommended venues, categories and related data from Foursquare.com.

	Area	Area Latitude	Area Longitude	Density	Venue	Venue Latitude	Venue Longitude	Venue Category
0	90001	33.9740	-118.2495	6295.9	Superior Grocers	33.973280	-118.247079	Grocery Store
1	90001	33.9740	-118.2495	6295.9	Rite Aid	33.974383	-118.246351	Pharmacy
2	90001	33.9740	-118.2495	6295.9	Jack in the Box	33.975167	-118.250313	Fast Food Restaurant
3	90001	33.9740	-118.2495	6295.9	SUBWAY	33.975311	-118.248038	Sandwich Place
4	90001	33.9740	-118.2495	6295.9	Bill's Drive In	33.974500	-118.244225	Burger Joint
5	90001	33.9740	-118.2495	6295.9	Pizza Hut	33.975158	-118.248129	Pizza Place
6	90001	33.9740	-118.2495	6295.9	WINCHELL'S DONUT HOUSE	33.975075	-118.248211	Donut Shop

Figure 2 : Recommended venues from Foursquare API

With the use of the two data sets, I have attempted to cluster the zip codes using machine learning algorithms. While the algorithm helps build clusters, I have used an additional dimension of population density to refine the analysis. For plotting these points on the map, I have used geo-json data from a git-hub repository

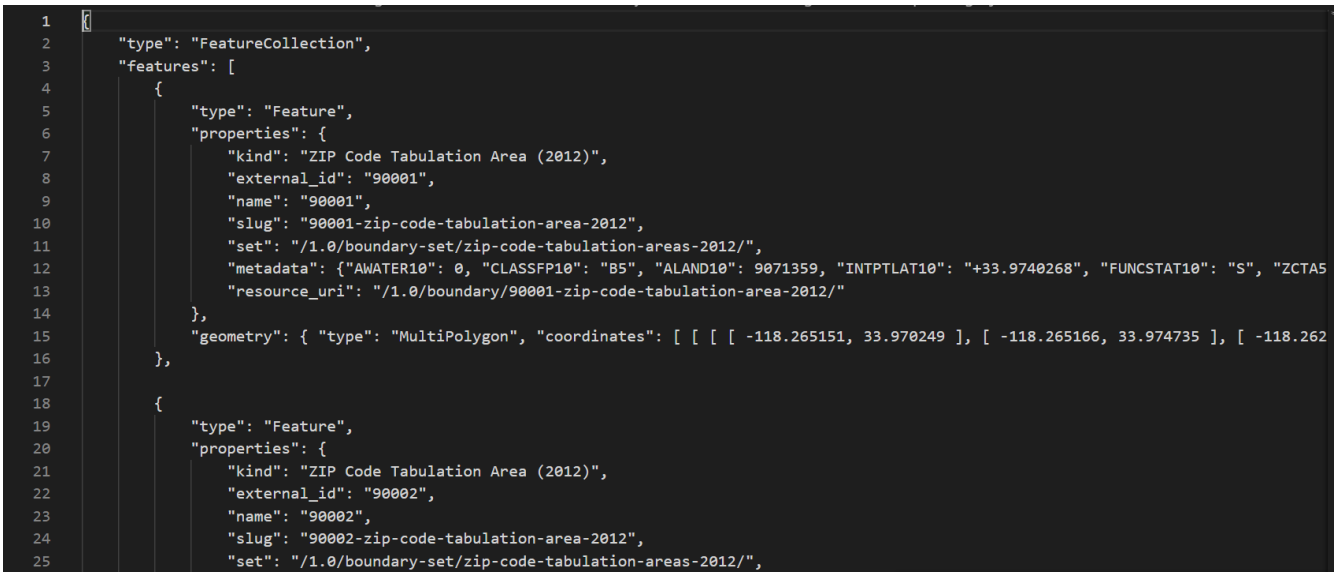

Figure 3 : Geojson data for LA

1. Zip codes with Latitude, Longitude and Population Density downloaded as excel from Simplemaps.com
2. List of recommended venues for each zip codes procured using API calls to Foursquare.com
3. Geo json for all US Postal codes from https://github.com/OpenDataDE/State-zip-code-GeoJSON