JULY, 2019

RECOMMENDATION OF LOCATIONS FOR OFFLINE STORES FOR AN ECOMMERCE COMPANY

IBM DATA SCIENCE PROFESSIONAL COURSE -CAPSTONE PROJECT

HOLEY AMIT

Contents

List of Figures

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.

This report is aimed at offline channel stakeholders within ecommerce organizations or Strategy planners within ecommerce organizations.

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	imprec	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

```
{
    "type": "FeatureCollection",
    "features": [
        {
            "type": "Feature",
            "properties": {
                "kind": "ZIP Code Tabulation Area (2012)",
                "external_id": "90001",
                "name": "90001",
                "slug": "90001-zip-code-tabulation-area-2012",
                "set": "/1.0/boundary-set/zip-code-tabulation-areas-2012/",
                "metadata": {"AWATER10": 0, "CLASSFP10": "B5", "ALAND10": 9071359, "INTPTLAT10": "+33.9740268", "FUNCSTAT10": "S", "ZCTA5
                "resource_uri": "/1.0/boundary/90001-zip-code-tabulation-area-2012/"
            },
            "geometry": { "type": "MultiPolygon", "coordinates": [ [ [ [ -118.265151, 33.970249 ], [ -118.265166, 33.974735 ], [ -118.262
        },

        {
            "type": "Feature",
            "properties": {
                "kind": "ZIP Code Tabulation Area (2012)",
                "external_id": "90002",
                "name": "90002",
                "slug": "90002-zip-code-tabulation-area-2012",
                "set": "/1.0/boundary-set/zip-code-tabulation-areas-2012/",
```

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

Methodology

Premise

I have chosen the city of Los Angeles as the seed city for this pilot for offline stores for ecommerce company. The premise for the analysis is that the zip codes with higher probability of footfalls will be a better potential location for the offline store. To understand the probability of footfalls, I have used 2 metrics. One is the number of trending venues and grouping the similar zip codes into clusters and the other metric is the population density of the zip code. Both these metrics lend themselves to arriving at an estimation of probability of footfall.

For the analysis, firstly, we get the relevant data sets:-

Data procurement and inference from data

- Get the set of zip codes for Los Angeles

- For each zip code, get the list of recommended venues from Foursquare
    - Once the zip code and list of venues is gathered, check for the number of venues for each zip code and decide a threshold

In [204]:
```
# Checking the rows in the data frame created
city_venues
# Please note that the terms Area and Zip are used interchangeably
```
Out[204]:

	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
7	90001	33.9740	-118.2495	6295.9	Amapola	33.974076	-118.248034	Fruit & Vegetable Store

Figure 4 : Postal codes/ areas with recommended venues and categories

In [205]:
```
# Observing the data by getting a count of  all venues for the Area
city_venues_category_grouped=city_venues.groupby('Area')[['Venue Category']].count()
city_venues_category_grouped
```
Out[205]:

Area	Venue Category
90001	10
90002	1
90003	4
90004	54
90005	34
90006	13
90007	19
90008	2
90010	23
90011	11
90012	27
90013	88

Figure 5 : Count of venues by zip code

- There are varying number of venues. Number of venues indicates popularity of a zip code and therefore the expected number of footfalls. As a result, a lower number of venues needs to be eliminated as it may distort the cluster. I have set the cutoff to 10. This brings the relevant set of zip codes to **39**

We see that many of the zip codes have less than 10 recommended venues. Lets identify them better

```
In [206]: city_venues_category_grouped.sort_values(by='Venue Category',inplace=True, ascending=False)
```

```
In [207]: # Forming a new data frame containing Zip Codes/Areas with more than 10 venues only
          new_zipcodes=city_venues_category_grouped[city_venues_category_grouped['Venue Category']>=10]
          new_zipcodes.size

Out[207]: 39
```

Lets consider only these 39 zip codes for our analysis. We are assuming that zipcodes with less than 10 recommended venues may not have enough footfalls worthy of setting up offline stores

```
In [208]: new_zipcodes
```

Out[208]:

Area	Venue Category
90071	100
90028	100
90014	100
90048	98
90013	88
90067	74
90079	64
90066	59
90004	54

Figure 6 : Choosing zip codes with more than 10 venues only

- Looking at the zip codes and venues, we see there a multitude of categories for the venues. Let us explore the categories

```
In [211]: # Let's find out how many unique categories can be curated from all the returned venues
          print('There are {} unique categories.'.format(len(new_city_venues['Venue Category'].unique())))

          There are 240 unique categories.
```

Figure 7 : Venue Category exploration

```
In [216]: # Checking the categories ...
          column_headers=city_grouped.columns
          for header in column_headers:
              print (header)

          Area
          ATM
          Accessories Store
          American Restaurant
          Arcade
          Art Gallery
          Art Museum
          Arts & Crafts Store
          Asian Restaurant
          Athletics & Sports
          Automotive Shop
          BBQ Joint
          Bakery
          Bank
          Bar
          Basketball Court
```

Figure 8 : Different venue categories

Algorithm decision point

We see that the zip codes and areas have categories (upto 240 types). These categories range from food joints, entertainment centres, shopping centres, health/fitness areas, etc

As we do not have a definite pattern of labels to form the groups, *we can attempt a unsupervised algorithm* that can help us form clusters and then we can use them to further analyse the set of zipcodes suited for offline stores.

We will use a k-means clustering algorithm. For using a clustering algorithm, we need to first convert the venues into a series of categorical features

Data Preprocessing

- Using *Onehot encoding* from the Pandas library to convert each of the categories into a binary feature for every venue

In [212]:
```
# one hot encoding
city_onehot = pd.get_dummies(new_city_venues[['Venue Category']], prefix="", prefix_sep="")

# add Area column back to dataframe
city_onehot['Area'] = new_city_venues['Area']

# move neighborhood column to the first column
fixed_columns = [city_onehot.columns[-1]] + list(city_onehot.columns[:-1])
city_onehot = city_onehot[fixed_columns]

city_onehot.head()
```

Out[212]:

	Area	ATM	Accessories Store	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Weight Loss Center	Whisky Bar	\
0	90001	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
1	90001	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
2	90001	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
3	90001	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
4	90001	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	

5 rows × 241 columns

Figure 9 : One hot encoding for categories

- I have then aggregated all the categories for each zip code by taking an average of the category across all venues in that zip code/area

In [214]:
```
# Getting a new data frame with venues aggregated by Zipcode
city_grouped = city_onehot.groupby('Area').mean().reset_index()
city_grouped.head(10)
```

Out[214]:

	Area	ATM	Accessories Store	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Weigh Loss Cente
0	90001	0.0	0.0	0.00	0.000000	0.000000	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0
1	90004	0.0	0.0	0.00	0.000000	0.000000	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.018519	0.0
2	90005	0.0	0.0	0.00	0.000000	0.000000	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0
3	90006	0.0	0.0	0.00	0.000000	0.000000	0.0	0.000000	0.000000	0.0	...	0.000000	0.076923	0.0	0.000000	0.0
4	90007	0.0	0.0	0.00	0.000000	0.000000	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0
5	90010	0.0	0.0	0.00	0.000000	0.000000	0.0	0.000000	0.043478	0.0	...	0.000000	0.000000	0.0	0.000000	0.0
6	90011	0.0	0.0	0.00	0.000000	0.000000	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.000000	0.0
7	90012	0.0	0.0	0.00	0.000000	0.037037	0.0	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.074074	0.0
8	90013	0.0	0.0	0.00	0.011364	0.022727	0.0	0.022727	0.000000	0.0	...	0.011364	0.000000	0.0	0.000000	0.0

Figure 10 : Aggregating using mean

- The data set is now pruned and primed for Machine Learning. Next step is to apply the clustering algorithm using K-means

Processing the data using Clustering

- Before I apply the k-means algorithm, we need to determine the optimal K. Using the elbow method, we see that 4 is the optimal number of clusters

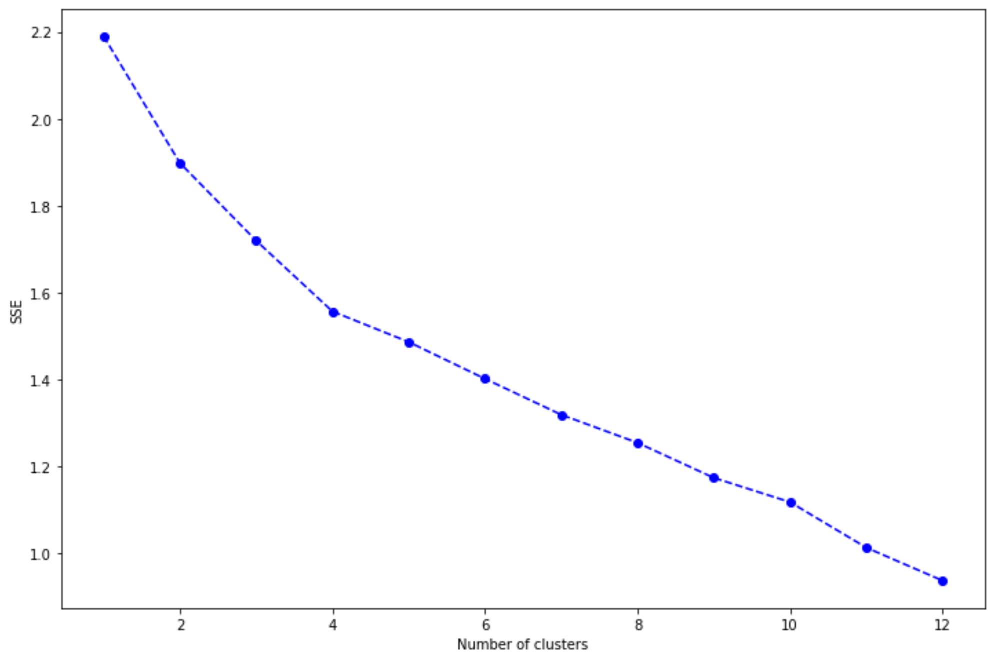

Figure 11 : Elbow for k-means clustering

- Proceeding with k=4, I have rerun the k-means clustering algorithm and get the labels for each cluster. I have also clubbed the top 10 venues for each zipcode along with its cluster label.

```
In [221]: # set number of clusters
          kclusters = 4

          city_grouped_clustering = city_grouped.drop('Area', 1)

          # run k-means clustering
          kmeans = KMeans(init='k-means++',n_clusters=kclusters, random_state=5).fit(city_grouped_clustering)

          # check cluster labels generated for each row in the dataframe
          kmeans.labels_

Out[221]: array([0, 1, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 3, 3, 0, 0, 0, 0,
                 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

```
In [222]: #Let's create a new dataframe that includes the cluster as well as the top 10 venues for each postal code

          # add clustering labels
          city_neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)
          city_neighborhoods_venues_sorted
```

Out[222]:

	Cluster Labels	Area	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	0	90001	Pizza Place	Donut Shop	Mexican Restaurant	Fruit & Vegetable Store	Shoe Store	Fast Food Restaurant	Grocery Store	Sandwich Place	Burger Joint	Pharmacy
1	1	90004	Korean Restaurant	Bakery	Bar	Coffee Shop	Seafood Restaurant	Sandwich Place	Cocktail Bar	Bridal Shop	Japanese Restaurant	Mattress Store
2	1	90005	Korean Restaurant	Café	Japanese Restaurant	Coffee Shop	Yoga Studio	Steakhouse	Gift Shop	Noodle House	Beer Bar	Concert Hall
3	0	90006	Donut Shop	Ice Cream Shop	Video Game Store	Pizza Place	Bus Station	Food Truck	Sandwich Place	Latin American Restaurant	Grocery Store	Spa
4	0	90007	Coffee Shop	Shipping Store	Yoga Studio	College Residence Hall	Caribbean Restaurant	Food Truck	Mediterranean Restaurant	Gastropub	Big Box Store	Farmers Market

Figure 12 : Cluster labels and top 10 venues

- Through a series of operations and merge, we can add the latitude, longitude and population density also to the zip code row

In [233]:

```python
# Merging the dataframes to get a consolidated view
city_merged = pd.merge(df_new,city_neighborhoods_venues_sorted, left_on='zip', right_on='Area')

#city_merged.sort_values("Cluster Labels", inplace=True)

city_merged
```

Out[233]:

	zip	lat	lng	density	Cluster Labels	Area	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 M Comr Ve
0	90001	33.9740	-118.2495	6295.9	0	90001	Pizza Place	Donut Shop	Mexican Restaurant	Fruit & Vegetable Store	Shoe Store	Fast Food Restaurant	Grocery Store	Sandv Pl
1	90004	34.0762	-118.3108	7876.3	1	90004	Korean Restaurant	Bakery	Bar	Coffee Shop	Seafood Restaurant	Sandwich Place	Cocktail Bar	Bridal S
2	90005	34.0591	-118.3064	13421.3	1	90005	Korean Restaurant	Café	Japanese Restaurant	Coffee Shop	Yoga Studio	Steakhouse	Gift Shop	Noc Ho
3	90006	34.0480	-118.2942	11903.1	0	90006	Donut Shop	Ice Cream Shop	Video Game Store	Pizza Place	Bus Station	Food Truck	Sandwich Place	L Ameri Restau
4	90007	34.0281	-118.2849	6403.9	0	90007	Coffee Shop	Shipping Store	Yoga Studio	College Residence	Caribbean Restaurant	Food Truck	Mediterranean Restaurant	Gastro

Figure 13 : Enriched data set

- Now that we have our clusters, we need to consider an additional point which is the population density of the cluster A cluster is formed of zip codes that are similar in nature based on the venue categories. However, the population density of the cluster will determine the number of footfalls in that cluster. *Higher the number of footfalls, better is the suitability of the zip code/area for the offline store*

- Let us first plot the map using the population density and then plot the zip codes on it based on the clusters. Using Folium choropleth and population density for zip codes, below is a map we get
For the folium map, I have downloaded the geo json file for South California and then picked only the zip codes for Los Angeles.

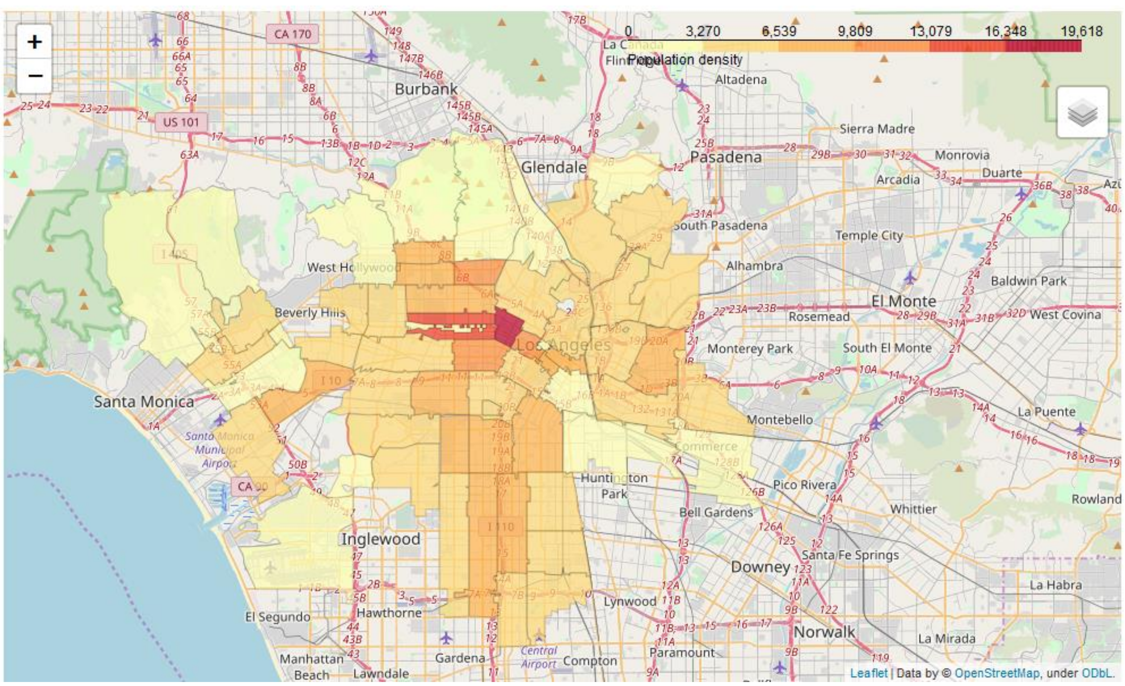

Figure 14 : Choropleth map of Los Angeles based on population density

- On the above map, we now plot the markers for the zip code based on the clusters we identified above. In the below map, the clusters are differentiated by colors
    - Cluster 0 - Red
    - Cluster 1 – Dark Blue
    - Cluster 2 – Light Blue
    - Cluster 3 – Green

    For each zip code popup, we mention the Cluster it belongs to and the population density value

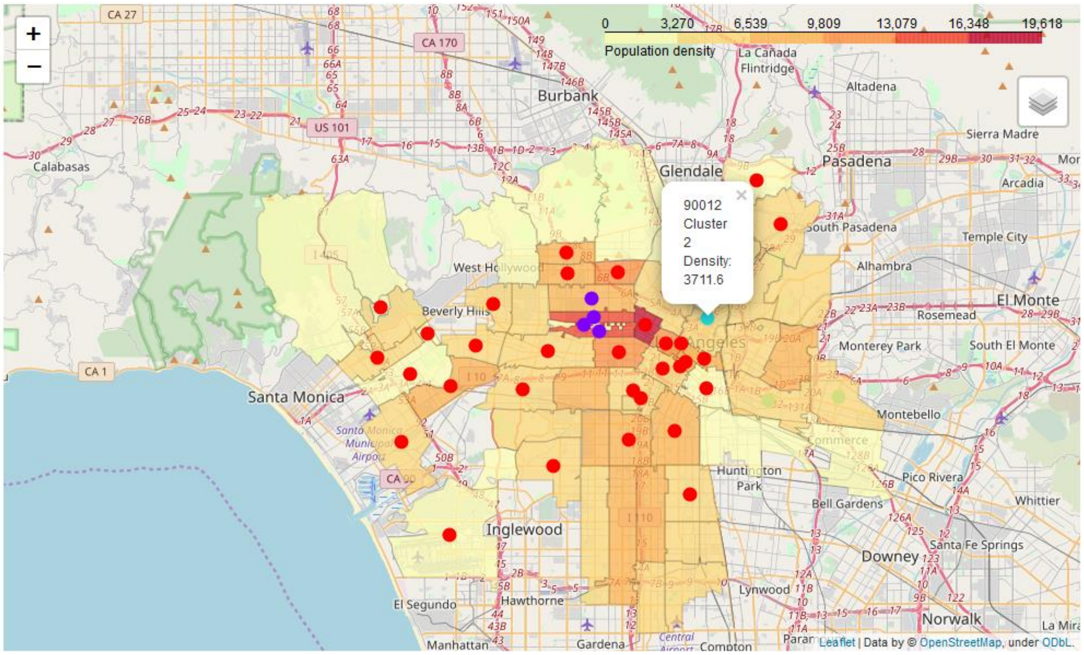

Figure 15 : Plotting the zip code markers

Results

Let us take a closer look at the results of the Clustering Algorithm.

Here is the set of zip codes with their clusters assigned.

Area	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
90001	Pizza Place	Donut Shop	Mexican Restaurant	Fruit & Vegetable Store	Shoe Store	Fast Food Restaurant	Grocery Store	Sandwich Place	Burger Joint	Pharmacy
90004	Korean Restaurant	Bakery	Bar	Coffee Shop	Seafood Restaurant	Sandwich Place	Cocktail Bar	Bridal Shop	Japanese Restaurant	Mattress Store
90005	Korean Restaurant	Café	Japanese Restaurant	Coffee Shop	Yoga Studio	Steakhouse	Gift Shop	Noodle House	Beer Bar	Concert Hall
90006	Donut Shop	Ice Cream Shop	Video Game Store	Pizza Place	Bus Station	Food Truck	Sandwich Place	Latin American Restaurant	Grocery Store	Spa
90007	Coffee Shop	Shipping Store	Yoga Studio	College Residence Hall	Caribbean Restaurant	Food Truck	Mediterranean Restaurant	Gastropub	Big Box Store	Farmers Market
90010	Korean Restaurant	Coffee Shop	Japanese Restaurant	Pizza Place	Construction & Landscaping	Tea Room	Martial Arts Dojo	Fast Food Restaurant	Convenience Store	Asian Restaurant
90011	Fast Food Restaurant	Mexican Restaurant	Ice Cream Shop	Pizza Place	Fried Chicken Joint	Discount Store	Donut Shop	Park	Food	Flea Market

90012	Chinese Restaurant	Bakery	Bar	Vietnamese Restaurant	Monument / Landmark	Café	Brewery	Recreation Center	Bubble Tea Shop	Burger Joint
90013	Japanese Restaurant	Sushi Restaurant	Ice Cream Shop	Ramen Restaurant	Gift Shop	Brewery	Cocktail Bar	Coffee Shop	Bakery	Bubble Tea Shop
90014	Bar	Coffee Shop	Burger Joint	Music Venue	Hotel	Italian Restaurant	Theater	Yoga Studio	Juice Bar	Café
90015	Food Truck	Coffee Shop	Bar	Sports Bar	Mexican Restaurant	Breakfast Spot	Bubble Tea Shop	Snack Place	Office	Smoke Shop
90016	Mexican Restaurant	Latin American Restaurant	Performing Arts Venue	Park	Sandwich Place	Check Cashing Service	Food	Fried Chicken Joint	Deli / Bodega	Wine Bar
90017	Coffee Shop	Clothing Store	Café	Sandwich Place	Gym	Motel	Mexican Restaurant	Steakhouse	Thai Restaurant	Food Truck
90019	Furniture / Home Store	Bank	Sandwich Place	Pizza Place	Chinese Restaurant	Burger Joint	Shopping Mall	Mexican Restaurant	Shipping Store	Mobile Phone Shop
90020	Korean Restaurant	Café	Dessert Shop	Ice Cream Shop	Bakery	Asian Restaurant	Shopping Mall	Coffee Shop	Diner	Burger Joint
90021	Grocery Store	Asian Restaurant	Coffee Shop	Marijuana Dispensary	Restaurant	Farmers Market	Food Stand	Convenience Store	Mexican Restaurant	Burger Joint
90022	Mexican Restaurant	Pizza Place	Donut Shop	Food Truck	Shoe Store	Bank	Mobile Phone Shop	Convenience Store	Discount Store	Flea Market
90023	Mexican Restaurant	Pizza Place	Video Game Store	Food Truck	Seafood Restaurant	Sandwich Place	Discount Store	Grocery Store	Department Store	Dessert Shop
90025	Gym	Intersection	Ramen Restaurant	Shop & Service	Chinese Restaurant	Deli / Bodega	Farmers Market	Cocktail Bar	Coffee Shop	Garden Center
90026	Café	Mexican Restaurant	Food Truck	Bar	Music Venue	Pizza Place	Pet Store	Vegetarian / Vegan Restaurant	Bookstore	Gym
90028	Coffee Shop	Lounge	Mexican Restaurant	Hotel	Bar	Nightclub	American Restaurant	Cocktail Bar	Pizza Place	Burger Joint
90029	Pizza Place	Convenience Store	Bakery	Fast Food Restaurant	Asian Restaurant	Middle Eastern Restaurant	Sandwich Place	Coffee Shop	Restaurant	Donut Shop
90034	Pizza Place	Mexican Restaurant	Optical Shop	Snack Place	Sushi Restaurant	Coffee Shop	Grocery Store	Sandwich Place	Pharmacy	Discount Store
90035	Middle Eastern Restaurant	Kosher Restaurant	Grocery Store	Arcade	French Restaurant	Thai Restaurant	Italian Restaurant	Sandwich Place	Café	Sushi Restaurant
90037	Grocery Store	Fast Food Restaurant	Burger Joint	Chinese Restaurant	Pet Store	Liquor Store	Convenience Store	Park	Smoke Shop	Gas Station
90038	Hotel	Gym	Pharmacy	Bar	Burrito Place	Motel	Fast Food Restaurant	Coffee Shop	Sandwich Place	Music Venue
90041	Italian Restaurant	French Restaurant	Pet Store	Asian Restaurant	Pizza Place	Creperie	Spa	Bubble Tea Shop	Burger Joint	Café
90042	Pizza Place	Italian Restaurant	Fast Food Restaurant	Burger Joint	Dumpling Restaurant	Food Truck	Breakfast Spot	Coffee Shop	Flower Shop	Sandwich Place
90043	Burger Joint	Intersection	Taco Place	American Restaurant	BBQ Joint	Convenience Store	Discount Store	Donut Shop	Dry Cleaner	Fried Chicken Joint
90045	Pizza Place	Burger Joint	Smoothie Shop	Shipping Store	Bar	Supermarket	Sandwich Place	Scenic Lookout	Paper / Office Supplies Store	Park
90048	Clothing Store	Gym / Fitness Center	Mexican Restaurant	Seafood Restaurant	Juice Bar	Italian Restaurant	Breakfast Spot	Café	Bakery	Department Store
90057	Clothing Store	Food Truck	Hotel	Theater	Donut Shop	Mexican Restaurant	Fried Chicken Joint	Seafood Restaurant	Fast Food Restaurant	Coffee Shop
90064	Clothing Store	Sandwich Place	Lingerie Store	ATM	Burger Joint	Food Court	Shopping Mall	Shoe Store	Mediterranean Restaurant	School
90066	Coffee Shop	Pizza Place	Pharmacy	Café	Mexican Restaurant	Japanese Restaurant	American Restaurant	Grocery Store	Pet Store	Bakery
90067	Food Truck	Coffee Shop	Mexican Restaurant	Café	Salad Place	Department Store	Hotel	Chinese Restaurant	Restaurant	Cosmetics Shop
90071	Sandwich Place	Hotel	French Restaurant	Mexican Restaurant	Coffee Shop	Food Truck	Italian Restaurant	Café	Bakery	Irish Pub
90079	Bar	Theater	Hotel	Clothing Store	Coffee Shop	Café	Sushi Restaurant	Burger Joint	Mexican Restaurant	Yoga Studio
90089	Coffee Shop	Mexican Restaurant	Fast Food Restaurant	Sandwich Place	Shipping Store	Burger Joint	Café	Fraternity House	Food Truck	Chinese Restaurant
90095	Coffee Shop	Café	Bus Station	Fountain	Pizza Place	Fast Food Restaurant	Sculpture Garden	Plaza	American Restaurant	Concert Hall

Figure 16 : Zip codes with cluster labels

I have enriched the data with the latitude, longitude, population density as well. Below is a snapshot

	zip	lat	lng	density	Cluster Labels	Area	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	90001	33.9740	-118.2495	6295.9	0	90001	Pizza Place	Donut Shop	Mexican Restaurant	Fruit & Vegetable Store	Shoe Store	Fast Food Restaurant	Grocery Store	Sandwich Place
1	90004	34.0762	-118.3108	7876.3	1	90004	Korean Restaurant	Bakery	Bar	Coffee Shop	Seafood Restaurant	Sandwich Place	Cocktail Bar	Bridal Shop
2	90005	34.0591	-118.3064	13421.3	1	90005	Korean Restaurant	Café	Japanese Restaurant	Coffee Shop	Yoga Studio	Steakhouse	Gift Shop	Noodle House
3	90006	34.0480	-118.2942	11903.1	0	90006	Donut Shop	Ice Cream Shop	Video Game Store	Pizza Place	Bus Station	Food Truck	Sandwich Place	Latin American Restaurant
4	90007	34.0281	-118.2849	6403.9	0	90007	Coffee Shop	Shipping Store	Yoga Studio	College Residence Hall	Caribbean Restaurant	Food Truck	Mediterranean Restaurant	Gastropub
5	90010	34.0621	-118.3162	3208.3	1	90010	Korean Restaurant	Coffee Shop	Japanese Restaurant	Pizza Place	Construction & Landscaping	Tea Room	Martial Arts Dojo	Fast Food Restaurant
6	90011	34.0071	-118.2587	9359.3	0	90011	Fast Food Restaurant	Mexican Restaurant	Ice Cream Shop	Pizza Place	Fried Chicken Joint	Discount Store	Donut Shop	Park
7	90012	34.0660	-118.2382	3711.6	2	90012	Chinese	Bakery	Bar	Vietnamese	Monument /	Café	Brewery	Recreation

Figure 17 : Enriched data

Based on the clustering algorithm, we see that the clusters are larger in size for cluster 0 and 1, while clusters 2 and 3 are single zip code clusters.

A quick snapshot of each cluster is as follows:-

Cluster 0

Cluster 0

```
In [236]: # Looking at cluster 0
          city_merged[city_merged['Cluster Labels']==0]
```

Out[236]:

	zip	lat	lng	density	Cluster Labels	Area	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 Mos Commo Venu
0	90001	33.9740	-118.2495	6295.9	0	90001	Pizza Place	Donut Shop	Mexican Restaurant	Fruit & Vegetable Store	Shoe Store	Fast Food Restaurant	Grocery Store	Sandwic Plac
3	90006	34.0480	-118.2942	11903.1	0	90006	Donut Shop	Ice Cream Shop	Video Game Store	Pizza Place	Bus Station	Food Truck	Sandwich Place	Lati America Restaurar
4	90007	34.0281	-118.2849	6403.9	0	90007	Coffee Shop	Shipping Store	Yoga Studio	College Residence Hall	Caribbean Restaurant	Food Truck	Mediterranean Restaurant	Gastropu
6	90011	34.0071	-118.2587	9359.3	0	90011	Fast Food Restaurant	Mexican Restaurant	Ice Cream Shop	Pizza Place	Fried Chicken Joint	Discount Store	Donut Shop	Par

```
In [237]: df_cl0=city_merged[city_merged['Cluster Labels']==0]

          arealist=df_cl0.Area.tolist()

          df_cl0=city_venues[city_venues.Area.isin(arealist)]

          df_cl0['Venue Category'].value_counts()

Out[237]: Coffee Shop                   62
          Mexican Restaurant            45
          Café                          37
          Pizza Place                   35
          Sandwich Place                33
          Bar                           33
          Clothing Store                31
          Food Truck                    30
          Hotel                         25
          Italian Restaurant            24
          Bakery                        23
          Burger Joint                  22
          Japanese Restaurant           21
          American Restaurant           20
          Grocery Store                 19
          Fast Food Restaurant          19
          Sushi Restaurant              17
          Gym                           16
          Juice Bar                     14
```

Figure 18 : Cluster 0

Cluster 1

Cluster 1

```
In [238]: city_merged[city_merged['Cluster Labels']==1]
```

Out[238]:

	zip	lat	lng	density	Cluster Labels	Area	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 Mc Comm Ven
1	90004	34.0762	-118.3108	7876.3	1	90004	Korean Restaurant	Bakery	Bar	Coffee Shop	Seafood Restaurant	Sandwich Place	Cocktail Bar	Bridal Shop	Japane Restaura
2	90005	34.0591	-118.3064	13421.3	1	90005	Korean Restaurant	Café	Japanese Restaurant	Coffee Shop	Yoga Studio	Steakhouse	Gift Shop	Noodle House	Beer E
5	90010	34.0621	-118.3162	3208.3	1	90010	Korean Restaurant	Coffee Shop	Japanese Restaurant	Pizza Place	Construction & Landscaping	Tea Room	Martial Arts Dojo	Fast Food Restaurant	Convenien Sto
14	90020	34.0664	-118.3099	13310.8	1	90020	Korean Restaurant	Café	Dessert Shop	Ice Cream Shop	Bakery	Asian Restaurant	Shopping Mall	Coffee Shop	Dir

```
In [240]: df_cl1=city_merged[city_merged['Cluster Labels']==1]

          arealist=df_cl1.Area.tolist()

          df_cl1=city_venues[city_venues.Area.isin(arealist)]

          df_cl1['Venue Category'].value_counts()
```

```
Out[240]: Korean Restaurant          39
          Coffee Shop                10
          Japanese Restaurant         7
          Café                        7
          Bakery                      6
          Bar                         4
          Ice Cream Shop              3
          Sandwich Place              3
          Dessert Shop                3
          Convenience Store           3
          Asian Restaurant            3
          Juice Bar                   2
          Bridal Shop                 2
          Sushi Restaurant            2
          Diner                       2
          Seafood Restaurant          2
          Cocktail Bar                2
          Noodle House                2
          BBQ Joint                   2
```

Figure 19 : Cluster 1

Cluster 2

Cluster 2

```
In [241]: city_merged[city_merged['Cluster Labels']==2]
```

Out[241]:

	zip	lat	lng	density	Cluster Labels	Area	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	10 Mo Commo Venu
7	90012	34.066	-118.2382	3711.6	2	90012	Chinese Restaurant	Bakery	Bar	Vietnamese Restaurant	Monument / Landmark	Café	Brewery	Recreation Center	Bubble Tea Shop	Burg Joi

```
In [242]: df_cl2=city_merged[city_merged['Cluster Labels']==2]

          arealist=df_cl2.Area.tolist()

          df_cl2=city_venues[city_venues.Area.isin(arealist)]

          df_cl2['Venue Category'].value_counts()
```

```
Out[242]: Chinese Restaurant        9
          Bakery                    3
          Vietnamese Restaurant     2
          Bar                       2
          Recreation Center         1
          Burger Joint              1
          Art Gallery               1
          Café                      1
          Brewery                   1
          Diner                     1
          Jazz Club                 1
          French Restaurant         1
          Plaza                     1
          Bubble Tea Shop           1
          Monument / Landmark       1
          Name: Venue Category, dtype: int64
```

Figure 20 : Cluster 2

Cluster 3

Cluster 3

In [243]:
```
city_merged[city_merged['Cluster Labels']==3]
```

Out[243]:

	zip	lat	lng	density	Cluster Labels	Area	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	C
16	90022	34.0245	-118.1561	5919.8	3	90022	Mexican Restaurant	Pizza Place	Donut Shop	Food Truck	Shoe Store	Bank	Mobile Phone Shop	Convenience Store	Discount Store	
17	90023	34.0228	-118.1999	4195.7	3	90023	Mexican Restaurant	Pizza Place	Video Game Store	Food Truck	Seafood Restaurant	Sandwich Place	Discount Store	Grocery Store	Department Store	[illegible]

In [244]:
```
df_cl3=city_merged[city_merged['Cluster Labels']==3]

arealist=df_cl3.Area.tolist()

df_cl3=city_venues[city_venues.Area.isin(arealist)]

df_cl3['Venue Category'].value_counts()
```

Out[244]:
```
Mexican Restaurant    5
Pizza Place           2
Food Truck            2
Mobile Phone Shop     1
Department Store      1
Discount Store        1
Convenience Store     1
Donut Shop            1
Bank                  1
Sandwich Place        1
Dessert Shop          1
Video Game Store      1
Shoe Store            1
Grocery Store         1
Seafood Restaurant    1
Name: Venue Category, dtype: int64
```

Figure 21 : Cluster 3

Discussion

We see that the within a cluster, some postal codes are in very high-density area which would suit better the footfall probability for an offline store. As an example, the 90057 in Cluster 0 is in the densest area of Los Angeles and so within Cluster 0, the chances of footfalls for 90057 are much higher than say 90045 which has much lesser density

Let us look through each cluster and note the observations

Cluster 0

Cluster 0 appears to have a good mix of fast food places, cafes, restaurants and shopping areas. Within Cluster 0, the zip codes with high population density are more suited for higher footfalls and therefore the target areas of offline stores

Top 5 Zip codes based on Population Density are 90006, 90017, 90029, 90057, 90014

Cluster 1

Cluster 1 appears to be concentrated around restaurants with few instances of other categories like shops, fitness centers etc. Also, cluster 1 is in relatively high dense areas of the city. With these two conditions, the entire cluster is likely to get high percentage of footfalls throughout the day

Cluster 2

Cluster 2 is only 1 zip code with predominantly Asian Restaurants in it. Also, it has an Art Gallery, Jazz Club and Recreation Center in it. As it is a single zip code cluster and has venues which have infrequent footfalls (possibly higher footfalls on weekends or weekday evenings), it will be have lower chances(as compared to Cluster 0 top 5 and Cluster 1) for offline stores

Cluster 3

Cluster 3 is only zip code with predominantly Mexican Restaurants in it. It has low population density as well.

Based on the above analysis, we can suggest the ecommerce company to setup the offline stores in the below order of priority.

The ecommerce store may choose to launch only a handful offline stores in each of the priority areas or only in the highest priority areas.

Priority 1

Cluster 0 - top 5 zip codes and Cluster 1 - are both concentrated in relatively high population density and with a balanced mix of food joints, shopping centers, fitness areas. These areas will have high footfalls throughout the day/evening on all days.

Priority 2

Cluster 0 remaining zip codes or Cluster 2/3 could be the next locations for offline stores. If the Cluster 0 locations in Priority 1 are handling a good frequency of pickups, it may be a good idea to expand to Cluster 0 next set of zip codes before moving to Cluster 2 and 3 which are anyway a set of single zip codes.

Additional analysis

Assuming there is a customer demographic data from the ecommerce company, we can enrich this analysis using the zip codes of the customers to refine the likelihood of the footfalls and therefore the location of the offline stores.

Conclusion

Utilizing zip codes, population density and foursquare recommended venues, we have deployed a combination of k-means algorithm and map plots to identify zip codes that will have higher likelihood of getting footfalls. This heuristic is expected to determine the locations for the offline stores of the ecommerce company. While this information and analysis is restricted to the information available in the public domain, the ecommerce company can enrich the data with their proprietary data of customers and thus refine the analysis.

It is advisable to start with a handful of offline stores and monitor the utilization of these stores. This utilization data can further optimize the analysis and therefore increase the expected utilization of the next batch of offline stores.