Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Introduction: Business Problem

In this project we will try to find a recommended location for a given type of restaurant. Specifically, this report will be targeted to stakeholders interested in opening a specific type of restaurant in Los Angeles county.

Since there are lots of restaurants in LA we will try to detect locations that are not already crowded with restaurants. We are also particularly interested in areas with no Italian restaurants in vicinity. We would also prefer locations as close to city center as possible, assuming that first two conditions are met.

We will use our data science powers to generate a few most promissing neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Data

Based on definition of our problem, factors that will influence our decission are:

- number of existing restaurants in the neighborhood (any type of restaurant)
- number of and distance to Italian restaurants in the neighborhood, if any
- distance of neighborhood from city center

We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

- number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**
- the region vs city mapping will be done using web scraping from a web page published by LA Times
- coordinate of LA will be obtained using some open data sources made available by the city of Los Angeles

```
#load URLs we want to scrape into an array

BASE_URL = [

'http://maps.latimes.com/neighborhoods/neighborhood/list/'

]
```

In [6]:

```
#convert new array to dataframe
df = pd.DataFrame(postcode_array)
df.columns = ['URL', 'Postal_City', 'Region']
```

URL	Postal_City	Region
http://maps.latimes.com/neighborhoods/neighbor...	Acton	Antelope Valley
http://maps.latimes.com/neighborhoods/neighbor...	Adams-Normandie	South L.A.

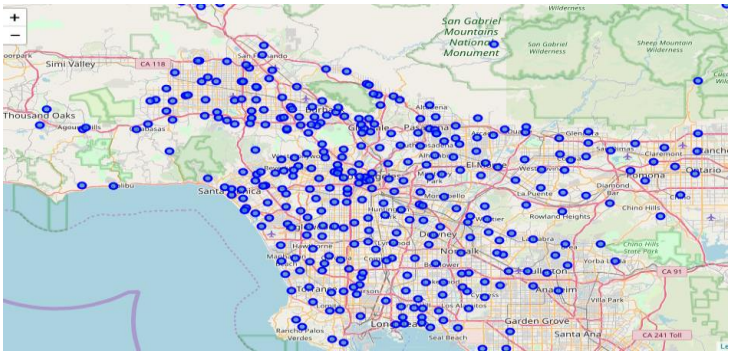
http://maps.latimes.com/neighborhoods/neighbor...	Agoura Hills	Santa Monica Mountains
http://maps.latimes.com/neighborhoods/neighbor...	Agua Dulce	Northwest County
http://maps.latimes.com/neighborhoods/neighbor...	Alhambra	San Gabriel Valley

#	Postal_City	Region
0	Acton	Antelope Valley
1	Adams-Normandie	South L.A.
...		
270	Winnetka	San Fernando Valley
271	Woodland Hills	San Fernando Valley

There are 360 Postal code in LA County. Get their Geo Location Latitude and Longitude

	ZIP Code	Postal City 1	Postal City 2	Postal City 3	Not Acceptable 1	Not Acceptable 2	Not Acceptable 3	Location	Zip	Latitude	Longitude
0	90713	Lakewood	NaN	NaN	NaN	NaN	NaN	90713(33.84871142900005, -118.11357922799999)	90713	33.848711	-118.113579
1	91306	Winnetka	NaN	NaN	NaN	NaN	NaN	91306(34.208404020000046, -118.57593995299999)	91306	34.208404	-118.575940
2	90002	Los Angeles	NaN	NaN	NaN	NaN	NaN	90002(33.94895070600006, -118.24697958699994)	90002	33.948951	-118.246980
3	90506	Torrance	NaN	NaN	NaN	NaN	NaN	90506(33.88535286100006, -118.32659746799999)	90506	33.885353	-118.326598
4	90069	West Hollywood	Los Angeles	NaN	NaN	NaN	NaN	90069(34.08940300900008, -118.37978902499998)	90069	34.089403	-118.379789

Here's a visual representation




```
#Now, let's get the top 100 venues that are in the first city Lakewood
```

```
90713 within a radius of 500 meters using Foursquare API.
#First, let's create the GET request URL. Name your URL url.
# define radius. Ingest JSON file
radius = 500
# create URL
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_s
ecret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    city_latitude,
    city_longitude,
    radius,
    LIMIT)
```

Now we are ready to clean the json and structure it into a pandas dataframe and categoooorize them by Restaurant Category.

	name	categories	postalCode	lat	lng
0	California Sushi & Teriyaki	Japanese Restaurant	90713	33.846888	-118.115684
1	Starbucks	Coffee Shop	90713	33.846731	-118.116127
2	Ralphs	Supermarket	90713	33.848014	-118.115557
3	99 Cents Only Stores	Discount Store	90713	33.847387	-118.117002
4	7-Eleven	Convenience Store	90713	33.846561	-118.116743

Methodology

In this project we will direct our efforts on detecting what type of restaurants are available in each areas of LA County that have low restaurant density. We will limit our analysis to area 500 mtr around each zip code in LA County. In first step we have collected the required data: location and type (category) of every restaurant in LA County. We have to filter the data from foursqare to select only Restaurants / Food joints.

Second step in our analysis will be calculation and exploration of 'restaurant density' across different areas of LA

In third and final step we will focus on most promising areas and within those create clusters of locations that meet some basic requirements established in discussion with stakeholders. We will present map of all such locations but also create clusters (using k-means clustering) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

Explore all Cities in LA County by Zip code. Let's create a function to repeat the same process to all the Cities + Zip Codes in LA

	Neighborhood	Zipcode	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lakewood	90713	33.848711	-118.113579	California Sushi & Teriyaki	33.846888	-118.115684	Japanese Restaurant
1	Lakewood	90713	33.848711	-118.113579	Starbucks	33.846731	-118.116127	Coffee Shop
2	Lakewood	90713	33.848711	-118.113579	Ralphs	33.848014	-118.115557	Supermarket
3	Lakewood	90713	33.848711	-118.113579	99 Cents Only Stores	33.847387	-118.117002	Discount Store
4	Lakewood	90713	33.848711	-118.113579	7-Eleven	33.846561	-118.116743	Convenience Store

```
print(la_venues.shape)
la_venues.head()
```

```
(6425, 8)
```

It appears that all the venue categories retrieved from Foursquare API are not restaurant / food joints.

Find the unique venue categories (it returns 600+) and apply a search only on those which are restaurant / food joints.

```
la_venues["Venue Category"].unique()
```

```
array(['Japanese Restaurant', 'Coffee Shop', 'Supermarket',
       'Discount Store', 'Convenience Store', 'Video Store',
       'Fast Food Restaurant', 'Wings Joint', 'Music Store', 'Bar',
       'Latin American Restaurant', 'Mexican Restaurant',
       'Filipino Restaurant', 'Ice Cream Shop', 'Grocery Store',
       'South American Restaurant', 'Fried Chicken Joint', 'Food Service',
       "Women's Store", 'Park', 'Mediterranean Restaurant', 'Bookstore',
       'Pizza Place', 'Cosmetics Shop', 'Video Game Store', 'Hookah Bar',
       'Sushi Restaurant', 'Concert Hall', 'Asian Restaurant',
       'Sandwich Place', 'Breakfast Spot', 'Donut Shop',
       'Paper / Office Supplies Store', 'Food', 'Italian Restaurant',
       'Hawaiian Restaurant', 'Restaurant', 'New American Restaurant',
       'Boutique', 'Burger Joint', 'Diner', 'Pet Store', 'Plaza',
       'American Restaurant', 'Health Food Store', 'French Restaurant',
```

```
search_values = ['Restaurant','Coffee','Food','Fish','Chips','Pizza','Taco','Café', 'Cafe', 'Bakery','Donut','Pizza','Wings','Ho
t dog','Cake','Steakhouse','Deli','Soup','Dessert','Bistro','Yogurt','Pie','Noodle','Gourmet', 'Bar',]
```

```
#la_venues[la_venues.name.str.contains('|'.join(search_values ))]
la_venues_food = la_venues[la_venues["Venue Category"].str.contains('|'.join(search_values ))]
```

```
print(la_venues_food.shape)
la_venues_food.head()
```

```
(3043, 8)
```

	Neighborhood	Zipcode	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lakewood	90713	33.848711	-118.113579	California Sushi & Teriyaki	33.846888	-118.115684	Japanese Restaurant
1	Lakewood	90713	33.848711	-118.113579	Starbucks	33.846731	-118.116127	Coffee Shop
6	Lakewood	90713	33.848711	-118.113579	McDonald's	33.846284	-118.115973	Fast Food Restaurant
7	Lakewood	90713	33.848711	-118.113579	Taco Bell	33.846726	-118.117500	Fast Food Restaurant
8	Lakewood	90713	33.848711	-118.113579	The Buffalo Spot	33.846601	-118.115791	Wings Joint

Analyze each neighborhood and find the frequency of occurrence of each restaurant category

	Zip_Code	Afghan Restaurant	African Restaurant	American Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Bakery	Bar	Beer Bar	Bistro	Brazilian Restaurant
0	90001	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	90003	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	90004	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	90005	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	90006	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	90007	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	90008	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.100000	0.000000	0.000000	0.000000	0.000000
7	90009	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	90010	0.000000	0.000000	0.000000	0.000000	0.055556	0.000000	0.055556	0.000000	0.027778	0.000000	0.055556
9	90011	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10	90012	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.064516	0.000000	0.000000	0.000000	0.000000
11	90013	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.269231	0.000000	0.000000	0.000000
12	90014	0.000000	0.000000	0.037037	0.000000	0.000000	0.000000	0.000000	0.148148	0.000000	0.000000	0.000000
13	90015	0.000000	0.000000	0.058824	0.000000	0.000000	0.000000	0.029412	0.088235	0.029412	0.000000	0.000000
14	90016	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
15	90017	0.000000	0.000000	0.041667	0.000000	0.000000	0.000000	0.000000	0.041667	0.000000	0.000000	0.000000
16	90018	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000

Analyze each City + Zip Code along with the top 5 most common venues

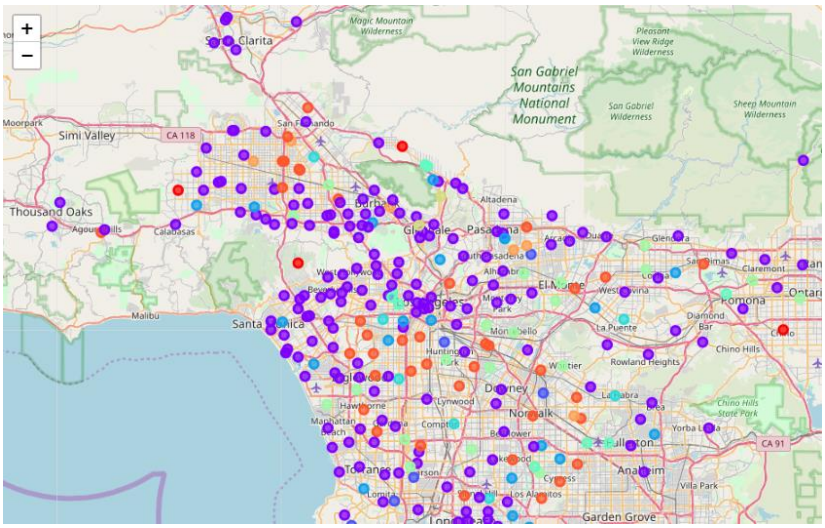
```
----90005----
                venue  freq
0    Korean Restaurant  0.45
1           Restaurant  0.09
2         Karaoke Bar  0.09
3     Sushi Restaurant  0.05
4  Japanese Restaurant  0.05


----90006----
                venue  freq
0    Korean Restaurant  0.67
1          Pizza Place  0.08
2  Fast Food Restaurant  0.08
3          Coffee Shop  0.08
4   Chinese Restaurant  0.08
```

Create the new dataframe and display the top 10 venues for each neighborhood.

	Zip_Code	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	90001	Pizza Place	Donut Shop	Fast Food Restaurant	Mexican Restaurant	Donburi Restaurant	Dongbei Restaurant	Dumpling Restaurant	Eastern European Restaurant	English Restaurant	Ethiopian Restaurant
1	90003	Pizza Place	Fast Food Restaurant	Taco Place	Southern / Soul Food Restaurant	Fish Market	Doner Restaurant	Dongbei Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant
2	90004	Cocktail Bar	Mexican Restaurant	Pizza Place	Sushi Restaurant	Fish Market	Dongbei Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	English Restaurant
3	90005	Korean Restaurant	Restaurant	Karaoke Bar	Café	Sushi Restaurant	Mexican Restaurant	Japanese Restaurant	Steakhouse	Noodle House	Halal Restaurant
4	90006	Korean Restaurant	Pizza Place	Fast Food Restaurant	Coffee Shop	Chinese Restaurant	Food	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	English Restaurant
5	90007	Coffee Shop	Food Truck	Korean Restaurant	Caribbean Restaurant	Italian Restaurant	Mexican Restaurant	Mediterranean Restaurant	Café	Eastern European Restaurant	English Restaurant

Run k-means to cluster the neighborhood into 10 clusters. Set number of clusters (kclusters) to 10 and finally visualize in a map



Results and Discussion

Our analysis shows that although there is a great number of restaurants in Los Angeles County , there are pockets of low restaurant density towwards the north and eastern parts of the county. Highest concentration of restaurants was detected in Los Angeles (downtown) and southern part.

We were able to distintly identfy the clusters where cwertain kinds of cuisinesare popular predominantly (for example for Mexican, Korean, Bakery). This will help us determine some location to open a certain kind of cuisine. Of course, it does not imply that those zones are actually optimal locations for a new restaurant! Purpose of this analysis was to only provide info on areas within LA but not crowded with existing restaurants Recommended zones 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.