Battle of Neighborhoods in Los Angeles

(Applies Data Science Capstone Project by IBM *via Coursera)*

Aimaiti Rehemanjiang

Marburg(Lahn), February 26, 2020

Contents

1 The Business Problem

Let us imagine that one of the largest coffee shops in Europe MeinCoffee is planning to expand its business to North America. They are planning to open their very first five Coffee Shops in Los Angeles, California, USA. Since there are lots of coffee shops in Los Angeles, the stakeholders must have some reliable information about the best or optimal locations / neighborhoods. Here optimal neighborhood might be neighborhood or neighborhoods with less or no number of existing coffee shops in certain near area e.g. 3 km range. The data science team must be able to deliver a fast report about the necessary location information, then the stakeholder can make safer decision that in which neighborhoods are more suitable to open their new coffee shops.

2 Data Loading and Preparation

The dataset I have used in this small project were collected from several resources such as online source, Foursquare API etc. I have utilized the Python's Pandas library to load and clean the data.

2.1 Data directly from online source

To get the information of Los Angeles neighborhoods and its geological locations like the latitude and the longitude I have searched in Internet and finally I found the needed LA neighborhood data from following link
https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr
Here is a snapshot of part of this data

In [2]:
```
# https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr
neigh_path = r'C:/pythonwork/kaggle/data/us_accidents/la_neighborhoods.csv'
# read the dataset into the pandas DataFrame
neigh_df=pd.read_csv(neigh_path)
# Let's look at the dataset
neigh_df.head()
```

Out[2]:

geom	kind	external_i	name	display_na	sqmi	type	name_1	slug_1	latitude	longitude	location
YGON 20541 897...	L.A. County Neighborhood (Current)	acton	Acton	Acton L.A. County Neighborhood (Current)	39.339109	unincorporated-area	NaN	NaN	-118.169810	34.497355	POINT(34.497355239240846 -118.16981019229348)
YGON 00012 109...	L.A. County Neighborhood (Current)	adams-normandie	Adams-Normandie	Adams-Normandie L.A. County Neighborhood (Curr...	0.805350	segment-of-a-city	NaN	NaN	-118.300208	34.031461	POINT(34.031461499124156 -118.30020800000011)
YGON 00009 029...	L.A. County Neighborhood (Current)	agoura-hills	Agoura Hills	Agoura Hills L.A. County Neighborhood (Current)	8.146760	standalone-city	NaN	NaN	-118.759885	34.146736	POINT(34.146736499122795 -118.75988450000015)
				Agua Dulce							

Figure 1: LA neighborhood dataset from online source

2.2 Data collection by using Foursquare API

To get the information of all types of venues in certain range of area of those neighborhoods and geolocation information of those venues I have use the Foursquare API to query the location data I planned to use in this project. Here is a snapshot of part of this data

In [11]:
```
# type your answer here
latitude =  dwtw_df.longitude.values[0]
longitude = dwtw_df.latitude.values[0]
LIMIT = 100
search_query = 'Italian'
radius = 500
url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&radius={}&limit={}'.format(C
url
```

Out[11]:
```
'https://api.foursquare.com/v2/venues/explore?client_id=4KA2G0XZSTB0RYQRCEKQQB1SFRLNQKIEY4NJJNLKTXMDTOYS&clie
```

Now query the above prepared URL by utilizing the *requests* library.

In [12]:
```
results = requests.get(url).json()
results
```

Out[12]:
```
{'meta': {'code': 200, 'requestId': '5e524db8fb34b5001bc91441'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': '$-$$$', 'key': 'price'}]},
  'headerLocation': 'Fashion District',
  'headerFullLocation': 'Fashion District, Los Angeles',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 19,
```

Figure 2: Location data obtained via Foursquare API

2.3 Prepared dataset

I have checked all the collected datasets and then cleaned the data. At th end I prepared it as a analyse ready Pandas DataFrame. Here is how the prepared data looks like I used the Folium Map library to plot these final location dataset

In [25]:
```
LA_venues.rename(columns={"Unnamed: 0":"ID"},inplace=True)
LA_venues.set_index("ID", inplace=True)
print(LA_venues.shape)
LA_venues.head()
```

```
(3020, 7)
```

Out[25]:

ID	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Acton	34.497355	-118.169810	Epik Engineering	34.498718	-118.168046	Construction & Landscaping
1	Acton	34.497355	-118.169810	Alma Gardening Co.	34.494762	-118.172550	Construction & Landscaping
2	Adams-Normandie	34.031461	-118.300208	Orange Door Sushi	34.032485	-118.299368	Sushi Restaurant
3	Adams-Normandie	34.031461	-118.300208	Shell	34.033095	-118.300025	Gas Station
4	Adams-Normandie	34.031461	-118.300208	Sushi Delight	34.032445	-118.299525	Sushi Restaurant

In [26]:
```
LA_venues.tail()
```

Out[26]:

ID	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
3015	Woodland Hills	34.159409	-118.615217	e-Motion Training	34.158111	-118.615593	Gym / Fitness Center
3016	Woodland Hills	34.159409	-118.615217	Deats Stewart Deats Design	34.159030	-118.612447	Wine Bar
3017	Woodland Hills	34.159409	-118.615217	Woodland Hills Carpet Restoration	34.158725	-118.618393	Carpet Store
3018	Woodland Hills	34.159409	-118.615217	Trunk Peter Productions	34.162271	-118.613903	Concert Hall
3019	Woodland Hills	34.159409	-118.615217	Le Zig Zag Club, Paris	34.162803	-118.615821	Asian Restaurant

Figure 3: Final prepared dataset

superimposed on the map of Los Angeles area, California as shown in Fig.4.

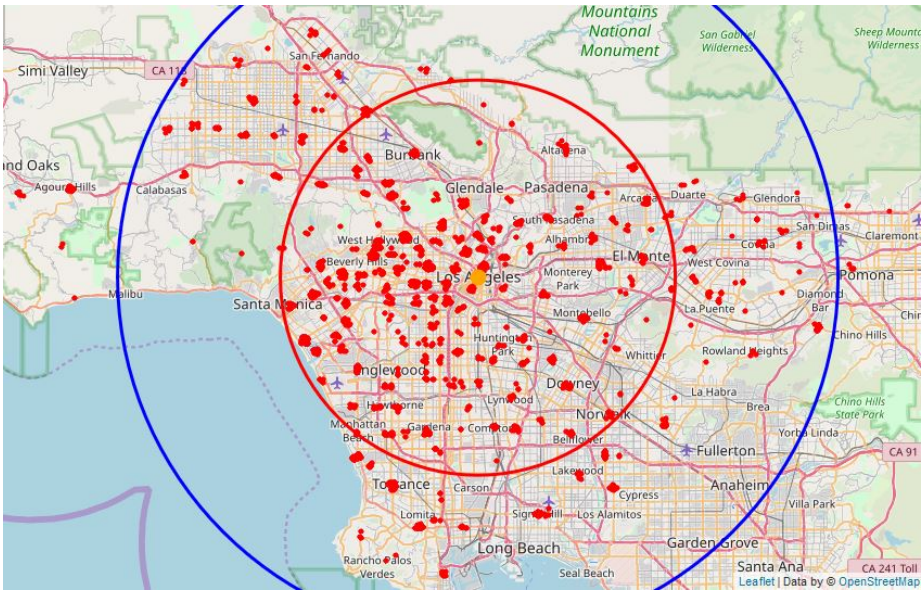

Figure 4: Venues superimposed on the map of Los Angeles. Red circle shows the 23 km range and blue circle shows 42 km range from the LA center.

3 Methodology

The following steps of methods are implement in this project.

- Foucus my analysis on locations within the range of 42 km from the center of Los Angeles
- Find out all venues of coffee shops from the prepared dataset in previous step
- Filter out those venues such that there are at least one coffee shop in 2 km range (by assuming that most residents have their own cars and feel the 2 km is not too far in such a large city like Los Angeles)
- According to previous result, find out those neighborhoods and venues without any coffee shops within 2 km distance.
- Classify these venues into reasonable number of clusters and show corresponding cluster centers
- Show how far these locations, namely these cluster centers

4 Data Analysis

4.1 Venues of Coffee Shops

First of all, I have filtered out all venues of Coffee Shops from the prepared dataframe easily by using Pandas. Here is the code and its output dataframe shows the filtered venues of Coffee Shops as shown in Fig.5. The location of

these Coffee Shops are plotted together with the whole venues of LA on the top of LA map as shown in Fig. 6.

```
In [31]: LA_venues_coffee = LA_venues[LA_venues['Venue Category']=='Coffee Shop']
         # To avoid the 'Quota Ecxceeded' Error of Foursquare API
         LA_venues_coffee.set_index('Neighborhood')
         print(LA_venues_coffee.shape)
         LA_venues_coffee.head(10)

         (89, 7)
```

Out[31]:

ID	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
159	Atwater Village	34.131066	-118.262373	Starbucks	34.129278	-118.258659	Coffee Shop
190	Azusa	34.137470	-117.912469	Starbucks	34.135670	-117.907500	Coffee Shop
336	Beverly Grove	34.076633	-118.376102	Starbucks	34.074911	-118.375322	Coffee Shop
396	Koreatown	34.064510	-118.304958	Bia Coffee	34.063580	-118.308221	Coffee Shop
418	Koreatown	34.064510	-118.304958	Starbucks	34.061339	-118.306407	Coffee Shop
434	Koreatown	34.064510	-118.304958	Starbucks	34.061796	-118.300898	Coffee Shop
523	Century City	34.055326	-118.415083	Starbucks	34.058445	-118.416640	Coffee Shop
532	Century City	34.055326	-118.415083	The Coffee Bean & Tea Leaf	34.058248	-118.413612	Coffee Shop
543	Century City	34.055326	-118.415083	The Coffee Bean & Tea Leaf	34.057721	-118.418984	Coffee Shop
555	Century City	34.055326	-118.415083	The Coffee Bean & Tea Leaf	34.058206	-118.414625	Coffee Shop

Figure 5: Filtering the coffee shops: Code and its output.

4.2 Venues without nearby Coffee Shops

To find out which venues do not have Coffee Shops in certain range of area I calculated the geometrical actual distance from the given geological locations of venues. To do this, I used the *Haversine formula* which simply used the latitudes and longitudes of two locations to find their geometrical distance. The formula is given in following

$$
\begin{aligned}
d &= 2r \arcsin\left(\sqrt{\mathrm{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\mathrm{hav}(\lambda_2 - \lambda_1)}\right) \\
&= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)
\end{aligned}
$$

where hav(x) is the *Haversine function*. And φ_1, λ_1 are latitude, longitude of location 1 respectively and φ_2, λ_2 are latitude, longitude of location 1 correspondingly. Here is the implementation of this Haversine function in Python (see Fig. 7) By using above distance formula I have found out all venues nearby those Coffee Shops in the range of 2 km and then I plotted them on the map together with Coffee Shops as shown in Fig. 8

4.3 Unsupervised Machine Learning - Clustering

Now it is time to use machine learning technique to estimate some optimal area of locations to open new Coffee Shops. Here to be optimal, the estimated area of locations must be such that at least in several km range area does not exist any Coffee Shops. But I do not have given number of such areas. Therefore I am going to use the Clustering technique which is one of the unsupervised machine leaning techniques I learned during the Data Science course by IBM *via Coursera* platform.

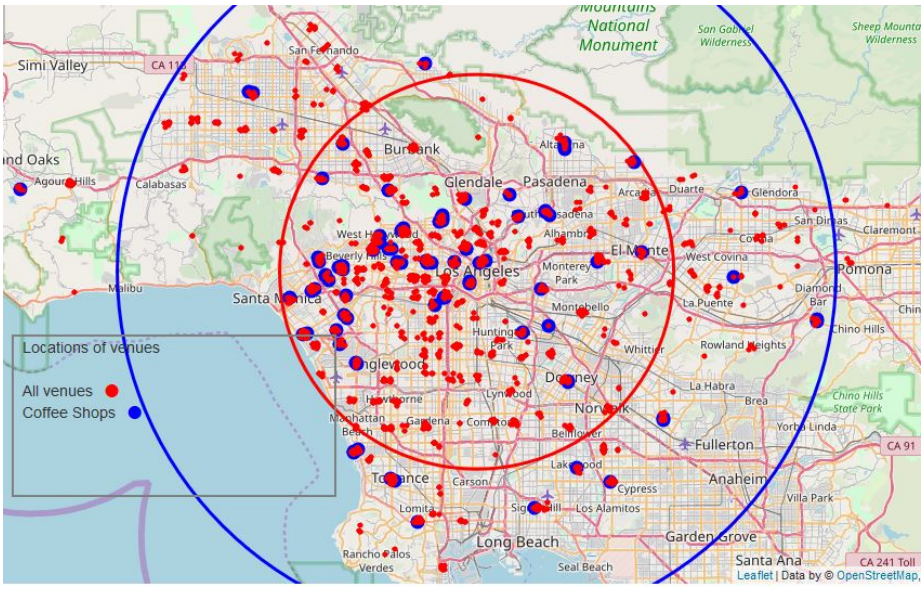

Figure 6: Filtering the coffee shops: Code and its output. Where blue filled circles are showing the locations of the venues of Coffee Shops.

```
In [39]: from math import radians, sin, cos, asin, sqrt
def haversine(lon1, lat1, lon2, lat2):
    lon1, lat1, lon2, lat2 = map(radians, [lon1, lat1, lon2, lat2])
    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat / 2) ** 2 + cos(lat1) * cos(lat2) * sin(dlon / 2) ** 2
    return 2 * 6371 * asin(sqrt(a))
```

Figure 7: Haversine function

First, lets look at the dataset which is going to be clustered. I have obtained this dataset after removing those locations of venues with Coffee Shops in 2 km range of area from the whole dataset. The output dataset is then looks like this as in Fig.9.

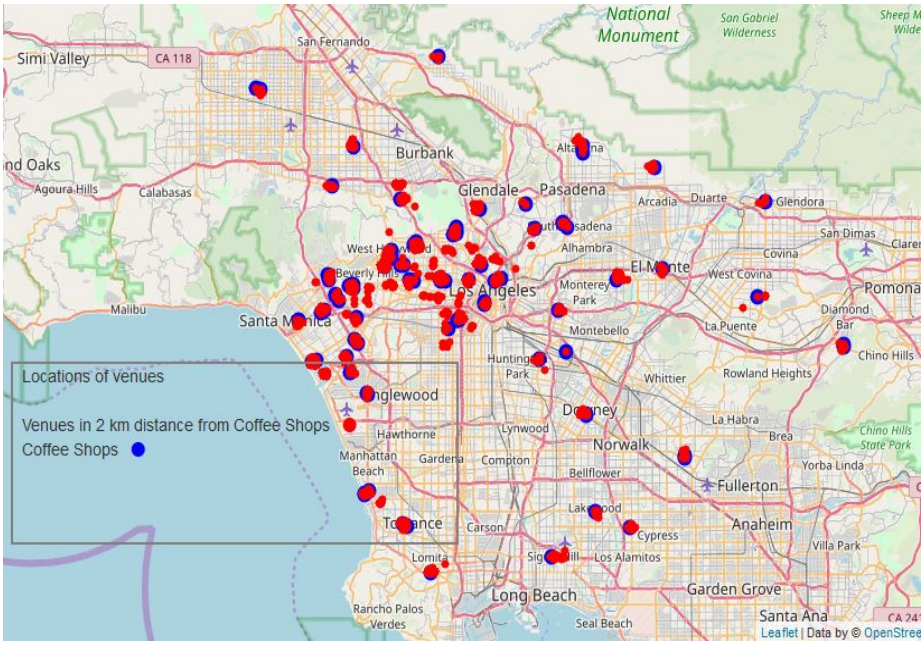

Figure 8: Venues with nearby Coffee Shops: red circles are showing those venues which are nearby Coffee Shops and blue circles are the locations of Coffee Shops.

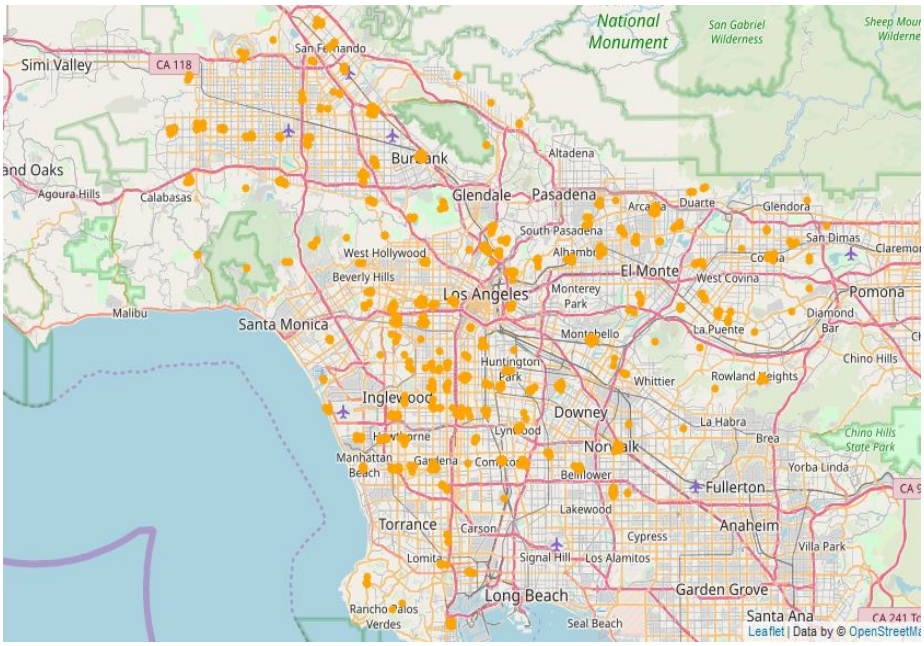

Figure 9: Venues without nearby Coffee Shops

5 Result and Discussion

After many times of trying with difference number of clusters I decided to use 16 clusters, because it gives optically well separated clusters and each cluster also covers only up to few km range of area, in my case it is around 4 km. Here are the clustered venues shown in Fig.10.

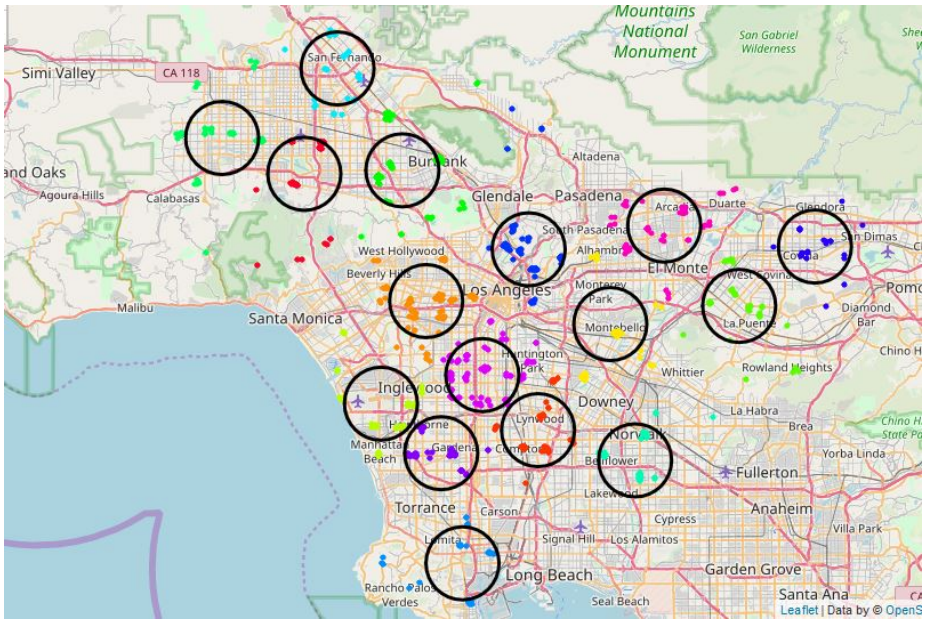

Figure 10: Cluster of venues without nearby Coffee Shops. Black circles are showing 4 km of range.

Since we all know that opening a Coffee Shop in area with less people is not a good idea, to check how the residents are distributed I have found a very informative population density map of Los Angeles area (see Fig.11. This population is denser distributed at those location with denser venues located, as expected.

As the final result of this small project I will select the most suitable three areas of locations as clusters. This clusters must have higher number of venues without nearby Coffee Shops and they should be located at the high population density regions. Here below are these three clusters showed in the map (see Fig.12.)

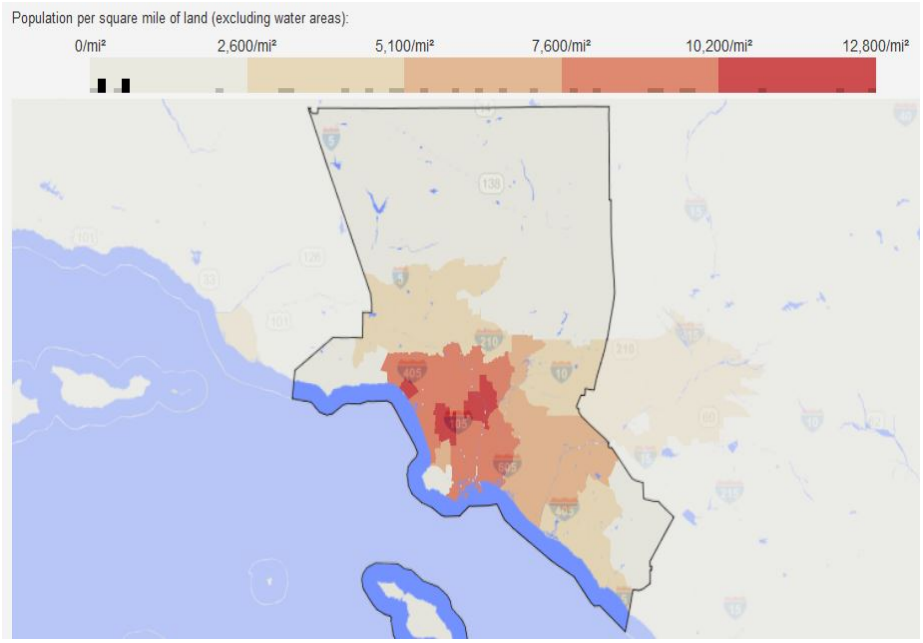

Figure 11: Population density of Los Angeles area.

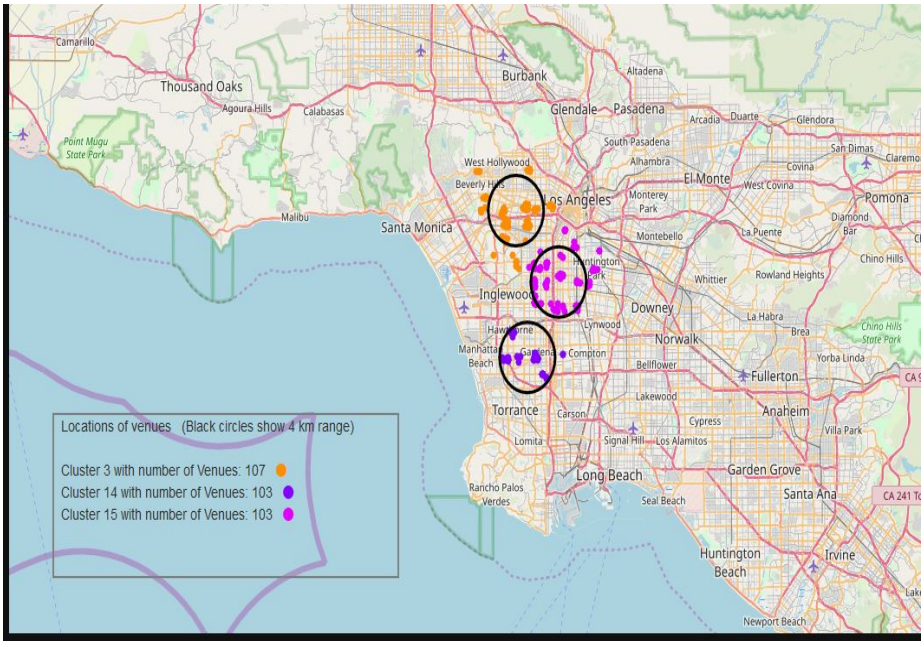

Figure 12: Selected three clusters of locations.

6 Conclusion

Although in this analysis we used only the top 100 venues in 500 meters in each neighbourhood, the result already gave some preliminary hint and idea where one should consider to open new coffee shops. By combining the information with above map of Los Angeles population density, our basic data science analysis is already give quite reasonable direction to stakeholders in the decision making process when it comes to openning new coffee shops in Los Angeles area, California, USA.

At the end of this analysis, I would like to recommend this three locations to open new Coffee Shops:

- Position 1 : Cluster 14 with the center Latitude: 33.890150910087435, Longitude: -118.32150675267954
- Position 2 : Cluster 15 with the center Latitude: 33.96807857657794, Longitude: -118.27197518844822
- Position 3 : Cluster 3 with the center Latitude: 34.041898362314186, Longitude: -118.33957372928336

Above analsis is just very preliminary. But at the same time it is already quite flexible to tweak the analysis, for instance, the distance threshold of filtering venues without coffee shops, how many top venues in which range to be considered in data collection step via Foursquare service, etc. .

References

[1] https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr

[2] https://developer.foursquare.com/docs/api

[3] https://statisticalatlas.com/metro-area/California/Los-Angeles/Population

[4] https://www.coursera.org/professional-certificates/ibm-data-science