CAPSTONE PROJECT:

The battle of the neigbourhoods.

Where shall we open a Mexican food business in Los Angeles, CA?

APRIL 08 / 2020

Authored by:
ROJAS Rafael

1. INTRODUCTION

1.1. Background

Los Angeles is one of the biggest cities at the USA and the main one according to the border with Mexico, laying in a basin in Southern California, adjacent to the Pacific Ocean. The city, which covers about 469 square miles (1,210 km2),[17] is the seat of Los Angeles County, the most populous county in the United States. The Los Angeles metropolitan area (MSA) is home to 13.1 million people, making it the second-largest metropolitan area in the nation after New York. Greater Los Angeles includes metro Los Angeles as well as the Inland Empire and Ventura County. It is the second most populous U.S. combined statistical area, also after New York, with a 2015 estimate of 18.7 million people.

According to the 2010 Census, the racial makeup of Los Angeles included: 1,888,158 Whites (49.8%), 365,118 African Americans (9.6%), 28,215 Native Americans (0.7%), 426,959 Asians (11.3%), 5,577 Pacific Islanders (0.1%), 902,959 from other races (23.8%), and 175,635 (4.6%) from two or more races. Hispanics or Latinos of any race were 1,838,822 persons (48.5%).

1.2. Busines interest

The business to open is a Mexican Restaurant at Los Angeles, CA in the United States of America. Due to the fact that is well known the food preference of the people and the Latin American ethnics established in the city that would benefit this plan. However, what's the best neighbourhood to open this kind of business and what are the factors that would impact our decision?

1.3. Problem

In order to help to make this decision it is imperative to establish the assumptions and parameters that would lead to the best choice, such as:

1. How many schools are located in the neighbourhood?
2. What are the ethnics of the neighbourhood?
3. How many restaurants of the kind are located in the neighbourhood?

Therefore, the following checklist shall be completed:

a) Create a dataframe with Los Angeles schools and Mexican restaurants
b) DBSCAN Clustering the dataframe created and dropping all clusters within the restaurants in it.
c) Creation of a cloropath map from the city showing the neighborhoods by the density of Hispanics population.
d) Add school's markers on the map in order to locate the best neighbourhoods to place the wanted business.

2. DATA SOURCES AND PRE-PROCESSING.

2.1. Data sources.

With the objective of having reliability on this project and results for the consultant, all data has been acquired from official government websites or professional data basis companies. A description of the datasets origin and parameters is listed as follows.

2.1.1 Schools Data Set.

Downloadable files containing general information about California's public schools and Districts. [https://www.cde.ca.gov/ds/si/ds/pubschls.asp]. Table as example of content below.

California Department of Education

Public Schools and Districts Data File

Educational Data
Management Division
March 1, 2020

CDSCode	1.10017E+12	**DOCType**	County Office of Education (COE)
NCESDist	691051	**SOC**	No Data
NCESSchool	No Data	**SOCType**	No Data
StatusType	Active	**EdOpsCode**	No Data
County	Alameda	**EdOpsName**	No Data
District	Alameda County Office of Education	**EILCode**	No Data
School	No Data	**EILName**	No Data
Street	313 West Winton Avenue	**GSoffered**	No Data
StreetAbr	313 West Winton Ave.	**GSserved**	No Data
City	Hayward	**Virtual**	No Data
Zip	94544-1136	**Magnet**	No Data
State	CA	**YearRoundYN**	No Data
MailStreet	313 West Winton Avenue	**FederalDFCDistrictID**	No Data
MailStrAbr	313 West Winton Ave.	**Latitude**	37.658212
MailCity	Hayward	**Longitude**	-122.09713
MailZip	94544-1136	**AdmFName1**	L Karen
MailState	CA	**AdmLName1**	Monroe
Phone	(510) 887-0152	**AdmEmail1**	lkmonroe@acoe.org
Ext	No Data	**AdmFName2**	No Data
WebSite	http://www.acoe.org	**AdmLName2**	No Data
OpenDate	No Data	**AdmEmail2**	No Data
ClosedDate	No Data	**AdmFName3**	No Data
Charter	No Data	**AdmLName3**	No Data
CharterNum	No Data	**AdmEmail3**	No Data
FundingType	No Data	**LastUpDate**	######
DOC	0		

This dataset has many characteristics that does not contributes any to the analyses. Therefore, the cleaning for the data utilisation can be read and evaluated in the section 2.2 Data Pre-processing in which only the columns of CDSCode, Status Type, City, Latitude and Longitude are left (highlighted in yellow in the previous table).

2.1.2 Mexican Restaurants at Los Angeles City.

Only taken into account those restaurants collected at the kaggle data set from the next website [https://www.kaggle.com/datafiniti/fast-food-restaurants]. Table as example of content below.

id	AWrSh_KgsVYjT2BJAzaH
dateAdded	2019-05-19T23:58:05Z
dateUpdated	2019-05-19T23:58:05Z
address	2555 11th Avenue
categories	Fast Food Restaurants,Hamburgers and Hot Dogs,Restaurants
primaryCategories	Accommodation & Food Services
city	Greeley
country	US
keys	us/co/greeley/255511thavenue/554191587
latitude	40.39629
longitude	-104.69699
name	Carl's Jr.
postalCode	80631
province	CO
sourceURLs	https://www.yellowpages.com/greeley-co/mip/carls-jr-7001402
websites	https://www.carlsjr.com/?utm_source=Yextandutm_medium=Visit%20Websiteandutm_campaign=Homepage

This dataset has many characteristics that does not contributes any to the analyses. Therefore, the cleaning for the data utilisation can be read and evaluated in the section 2.2 Data Pre-processing in which only the columns of id, categories, city, latitude and longitude are left (highlighted in yellow in the previous table).

2.1.3 Population ethnics by neighbourhood at Los Angeles City.

Obtained from LOS ANGELES OPEN DATA at https://data.lacity.org/A-Livable-and-Sustainable-City/Census-Data-by-Neighborhood-Council/nwj3-ufba. Table as example of content below.

NC_Name	ARLETA NC
Total Population	34932.84
White_pop	2882.67
Black_pop	409.67
Ameri_es_pop	67.64
Asian_pop	4061.31
Hawn_pi_pop	34.96
Hispanic_pop	27193.96
Other_pop	37.58
Multi_pop	245.05
In_Poverty	34700.56
Owner_occ	5590.27
Renter_occ	2159.94

This dataset has many characteristics that does not contributes any to the analyses. Therefore, the cleaning for the data utilisation can be read and evaluated in the section 2.2 Data Pre-processing in which only the columns of NC_Name, Total Population and Hispanic_pop are left (highlighted in yellow in the previous table).

2.2. Data Pre-processing

2.2.1 Executing step 1 - How many schools are located in the neighbourhood?

Figure 1. Schools final data-frame having the columns used to clean and filters the information of interest.

	CDSCode	StatusType	County	City	Latitude	Longitude
3571	19101990106880	Active	Los Angeles	Los Angeles	34.042460	-118.24912
3581	19101990121897	Active	Los Angeles	Los Angeles	34.063497	-118.20587
3605	19101990127522	Active	Los Angeles	Los Angeles	34.128523	-118.18775
3609	19101990134361	Active	Los Angeles	Los Angeles	33.998537	-118.30871
3611	19101990135582	Active	Los Angeles	Los Angeles	33.995341	-118.30851

Figure 2. Schools data-frame created to merge with the final restaurants location data-frame in order to develop a DBSCAN Clustering Method.

	ID	Latitude	Longitude
3571	19101990106880	34.042460	-118.24912
3581	19101990121897	34.063497	-118.20587
3605	19101990127522	34.128523	-118.18775
3609	19101990134361	33.998537	-118.30871
3611	19101990135582	33.995341	-118.30851
...	...	...	...
7214	19756636120158	34.035653	-118.45535
7274	19770810000000	33.996155	-118.27794
7275	19770810135954	33.996155	-118.27794
7276	19772890000000	34.102430	-118.18311
7277	19772890109942	34.102359	-118.18336

589 rows × 3 columns

2.2.2 Executing step 2 - What are the ethnics of the neighbourhood?

Figure 3. Ethnic's population final data-frame before obtaining the density in percentage.

	NC_Name	Total Population	Hispanic_pop
0	ARLETA NC	34932.84	0.778464
1	ARROYO SECO NC	21711.47	0.581628
2	ATWATER VILLAGE NC	11385.40	0.448178
3	BEL AIR-BEVERLY CREST NC	26789.14	0.056884
4	BOYLE HEIGHTS NC	81900.56	0.941439

Figure 4. Ethnic's density in percentage at the column Hispanic_pop.

	NC_Name	Total Population	Hispanic_pop
0	Arleta	34932.84	0.778464
1	Arroyo Seco	21711.47	0.581628
2	Atwater Village	11385.40	0.448178
3	Bel Air-Beverly Crest	26789.14	0.056884
4	Boyle Heights	81900.56	0.941439
...	...	...	...
92	Wilmington	59140.55	0.876536
93	Wilshire Center - Koreatown	99702.15	0.537840
94	Winnetka	51259.94	0.506450
95	Woodland Hills-Warner Center	68837.33	0.143204
96	Zapata King	50013.53	0.872515

97 rows × 3 columns

- Result of dividing it by the Total Population.
- NC_name has been modified by the function df.str.title() in order to match with the names at the .json file for a choropath mapping visualization.

2.2.3 Executing step 3 - How many restaurants of the kind are located in the neighbourhood?

Figure 5. Restaurants at LA final data-frame before filtering by “Mexican” at the “categories column.

	categories	city	latitude	longitude
213	Mexican,Breakfast,Vegetarian,Fast Food,Restaur...	Los Angeles	34.088300	-118.308800
528	Restaurant,Burger Joint,Mexican,American,Fast ...	Los Angeles	34.063900	-118.287500
1309	Restaurant,Mexican Restaurants,Taco Place,Fast...	Los Angeles	34.092500	-118.280500
1411	Fast Food,Chicken,Restaurants	Los Angeles	34.052900	-118.295200
3288	Restaurant,Fast Food Restaurant,wich Place,Fas...	Los Angeles	34.053538	-118.250897

Figure 6. Restaurants at LA final data-frame ready to merge with the final schools data-frame.

	ID	Latitude	Longitude
213	Mexican,Breakfast,Vegetarian,Fast Food,Restaur...	34.08830	-118.308800
528	Restaurant,Burger Joint,Mexican,American,Fast ...	34.06390	-118.287500
1309	Restaurant,Mexican Restaurants,Taco Place,Fast...	34.09250	-118.280500
5379	Mexican Restaurant,Restaurant,Mexican,Fast Food	34.02748	-118.429292

Result of filtering categories by the string “Mexican” and changing the name of the column “categories” to “ID” as the schools data-frame in order to develop a DBSCAN Clustering Method.

2.2.4 Executing index **a)** from checklist - Create a data-frame with Los Angeles schools and Mexican restaurants.

Figure 7. Merged Data-frame of schools and restaurants.

	ID	Latitude	Longitude
0	19101990106880	34.042460	-118.24912
1	19101990121897	34.063497	-118.20587
2	19101990127522	34.128523	-118.18775
3	19101990134361	33.998537	-118.30871
4	19101990135582	33.995341	-118.30851
...	...	...	...
588	19772890109942	34.102359	-118.18336
589	Mexican,Breakfast,Vegetarian,Fast Food,Restaur...	34.0883	-118.309
590	Restaurant,Burger Joint,Mexican,American,Fast ...	34.0639	-118.287
591	Restaurant,Mexican Restaurants,Taco Place,Fast...	34.0925	-118.281
592	Mexican Restaurant,Restaurant,Mexican,Fast Food	34.0275	-118.429

593 rows × 3 columns

- Result of using df.append() function and df.reset_index() function.
- The columns type for Latitude and Longitude where changed to float64 with the function df.apply(pd.to_numeric) to make the DCSCAN clustering possible.

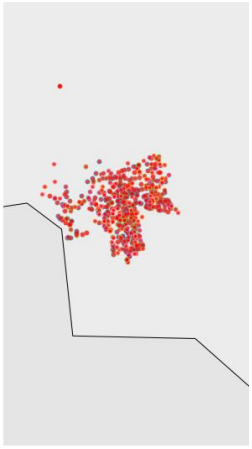
3. Methodology.

3.1. Executing index **b)** from checklist - **DBSCAN Clustering the data-frame created and dropping all clusters within the restaurants in it.**

3.1.1 Visualization of stations on map using basemap package before clustering.

Figure 8. Code and map of the coast of California showing points to be clustered.

```
from mpl_toolkits.basemap import Basemap
import matplotlib.pyplot as plt
from pylab import rcParams
%matplotlib inline
rcParams['figure.figsize'] = (14,10)

llon=-118.6
ulon=-117.8
llat=33.5
ulat=34.5

df_clust = df_clust[(df_clust['Longitude'] > llon) & (df_clust['Longitude'] < ulon) & (df_clust['Latitude'] > llat) &(df_clust['Latitude'] < ulat)]

my_map = Basemap(projection='merc',
            resolution = 'l', area_thresh = 1000.0,
            llcrnrlon=llon, llcrnrlat=llat, #min longitude (llcrnrlon) and latitude (llcrnrlat)
            urcrnrlon=ulon, urcrnrlat=ulat) #max longitude (urcrnrlon) and latitude (urcrnrlat)

my_map.drawcoastlines()
my_map.drawcountries()
my_map.fillcontinents(color = 'white', alpha = 0.3)

# To collect data based on stations

xs,ys = my_map(np.asarray(df_clust.Longitude), np.asarray(df_clust.Latitude))
df_clust['xm']= xs.tolist()
df_clust['ym'] =ys.tolist()

#Visualization1
for index,row in df_clust.iterrows():
#   x,y = my_map(row.Long, row.Lat)
   my_map.plot(row.xm, row.ym,markerfacecolor =([1,0,0]),  marker='o', markersize= 5, alpha = 0.75)
#plt.text(x,y,stn)
```



The previous image shows the correct agglomeration of points for a fine clustering. Be aware that Los Angeles city has one of the most extensive surface as commented at the introduction, and an agglomeration of schools surrounding the business location is a positive characteristic. Therefore the parameters for clustering are very limited to maintain only the schools very near to each other.

Figure 9. Code and data-frame including the cluster number column named “Clus_Db”.

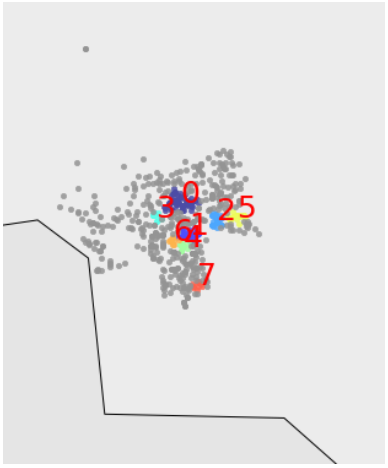
```
from sklearn.datasets.samples_generator import make_blobs
from sklearn.preprocessing import StandardScaler
```

```
import sklearn.utils
from sklearn.preprocessing import StandardScaler
sklearn.utils.check_random_state(1000)
Clus_dataSet = df_clust[['Longitude','Latitude']]
Clus_dataSet = np.nan_to_num(Clus_dataSet)
Clus_dataSet = StandardScaler().fit_transform(Clus_dataSet)

# Compute DBSCAN
db = DBSCAN(eps=0.15, min_samples=10).fit(Clus_dataSet)
core_samples_mask = np.zeros_like(db.labels_, dtype=bool)
core_samples_mask[db.core_sample_indices_] = True
labels = db.labels_
df_clust["Clus_Db"]=labels

realClusterNum=len(set(labels)) - (1 if -1 in labels else 0)
clusterNum = len(set(labels))


# A sample of clusters
df_clust[["ID","Latitude","Longitude","Clus_Db"]]
```

	ID	Latitude	Longitude	Clus_Db
0	19101990106880	34.042460	-118.249120	-1
1	19101990121897	34.063497	-118.205870	-1
2	19101990127522	34.128523	-118.187750	-1
3	19101990134361	33.998537	-118.308710	-1
4	19101990135582	33.995341	-118.308510	-1
...	...	...	...	...
588	19772890109942	34.102359	-118.183360	-1
589	Mexican,Breakfast,Vegetarian,Fast Food,Restaur...	34.088300	-118.308800	-1
590	Restaurant,Burger Joint,Mexican,American,Fast ...	34.063900	-118.287500	0
591	Restaurant,Mexican Restaurants,Taco Place,Fast...	34.092500	-118.280500	-1
592	Mexican Restaurant,Restaurant,Mexican,Fast Food	34.027480	-118.429292	-1

Only coordinates (Longitude and Latitude), were used as variable to define the cluster, as well as an eps of 0.15 used as limiter parameter for the observation regarding the distance between schools. The significance of a cluster number that is equals to -1 means that this point, school or restaurant, cannot be integrated in any cluster.

Figure 10. Visualization of DCSCAN clustering.



- It can be observed that most of the schools were located at the cluster number "-1" leaving a minority with a distance between them to be able to consider as positive for our business location.

Figure 11. Final data-frame after without the cluster numbers of "-1" and "0", at side the quantity of schools in each cluster.

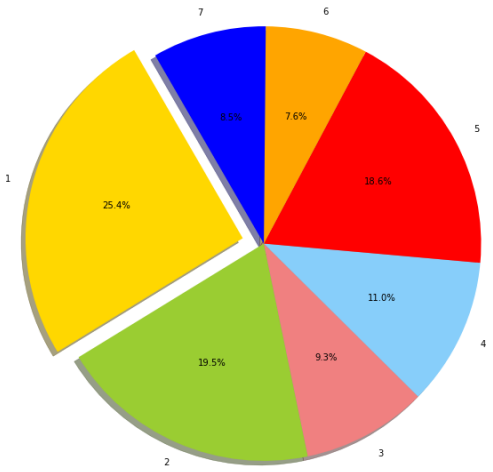
	ID	Latitude	Longitude	xm	ym	Clus_Db
10	19647330100743	34.011574	-118.27364	36289.559172	68419.342196	1
11	19647330100750	34.010434	-118.27393	36257.312658	68266.419595	1
13	19647330100867	34.031652	-118.20961	43409.366972	71112.994611	2
26	19647330102913	33.929468	-118.24623	39337.410676	57410.671338	7
27	19647330106427	34.014575	-118.26081	37716.189409	68821.914054	1
...	...	...	...	...	...	...
563	19647336115794	33.991524	-118.26324	37445.985864	65730.081804	4
568	19647336120471	34.041422	-118.21957	42301.866024	72423.962488	2
583	19753090131383	34.041580	-118.22017	42235.149100	72445.164642	2
585	19770810000000	33.996155	-118.27794	35811.421212	66351.170811	4
586	19770810135954	33.996155	-118.27794	35811.421212	66351.170811	4

118 rows × 6 columns

```
Clus_Db
1    30
2    23
3    11
4    13
5    22
6     9
7    10
dtype: int64
```

As result only 118 schools remains in the scope of the comparison with the choropath map. This amount of schools can be reliable only if there is an agglomeration over a neighbourhood with a considerable density of Hispanic people (>80%).

Figure 12. Pie chart of clusters magnitude in percentage.



- The clusters 1, 2 and 5 hold the biggest quantity of schools with a 25.4%, 19.5% and 18.6% respectively.
- The colours of these clusters at the map in figure 10 are:
  - 1 : Blue
  - 3 : Light Blue
  - 5 : Yellow

3.2. Executing index **c)** from checklist – **Creation of a cloropath map from the city showing the neighborhoods by the density of Hispanics population.**

To be able to use Folium it shall be installed using conda-forge. Once done and before executing the command for a choropath map, a troubleshooting must be effectuated to establish parameters for a correct visualization. For instance, random changes were made in the location, tiles and zoom start until the next map showed up and it was selected as basis to develop the next step.

The LA Times Neighborhood Boundaries information can be obtained at the webpage http://boundaries.latimes.com/sets/ as a .json file.

Figure 13. Code used to deploy the map shown under the cell.

```
# define the LA map
LA_map = folium.Map()

# define the LA map centered around LA with a high zoom level
LA_map = folium.Map(location=[34, -118.2],tiles='Stamen Toner', zoom_start=9.8)

# display LA map
LA_map
```

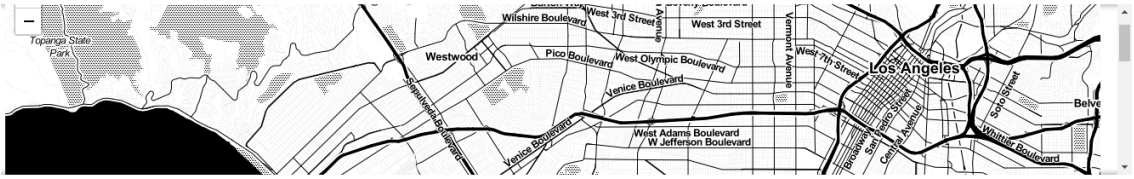
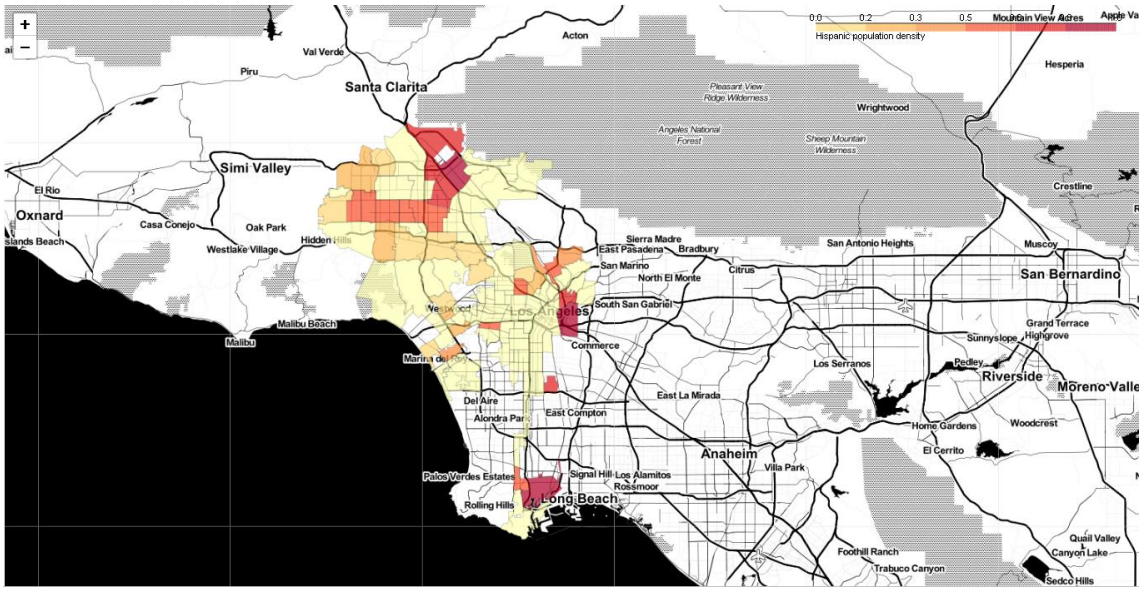


Figure 14. Code used to visualize the choropleth map shown under the cell.

```
LA_map.choropleth(
    geo_data=r'LA_Times_Neighborhood_Boundaries.json',
    data = df_data_2,
    columns=['NC_Name', 'Hispanic_pop'],
    key_on='feature.properties.name',
    fill_color='YlOrRd',
    fill_opacity=0.7,
    line_opacity=0.2,
    legend_name='Hispanic population density'
)

# display map
LA_map
```



For the county of Los Angeles, three neighborhoods are highlighted in dark red due to a density greater than 90%. However, only one of this three is in the city of Los Angeles. This

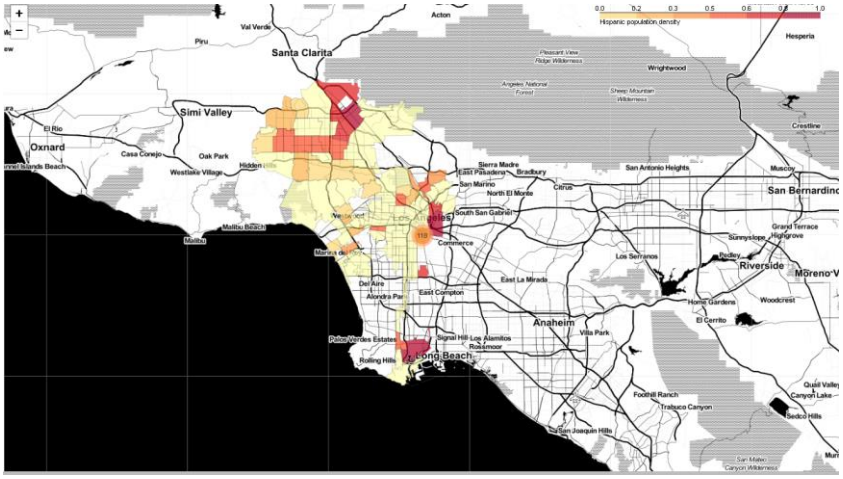
neighbourhood at the centre of the map (which we don't know its name yet) is the one to compare with the schools clusters.

4. RESULTS AND DISCUSSION

4.1. Executing index **d)** from checklist – **Add schools' markers on the map in order to locate the best neighbourhoods to place the Mexican food business.**

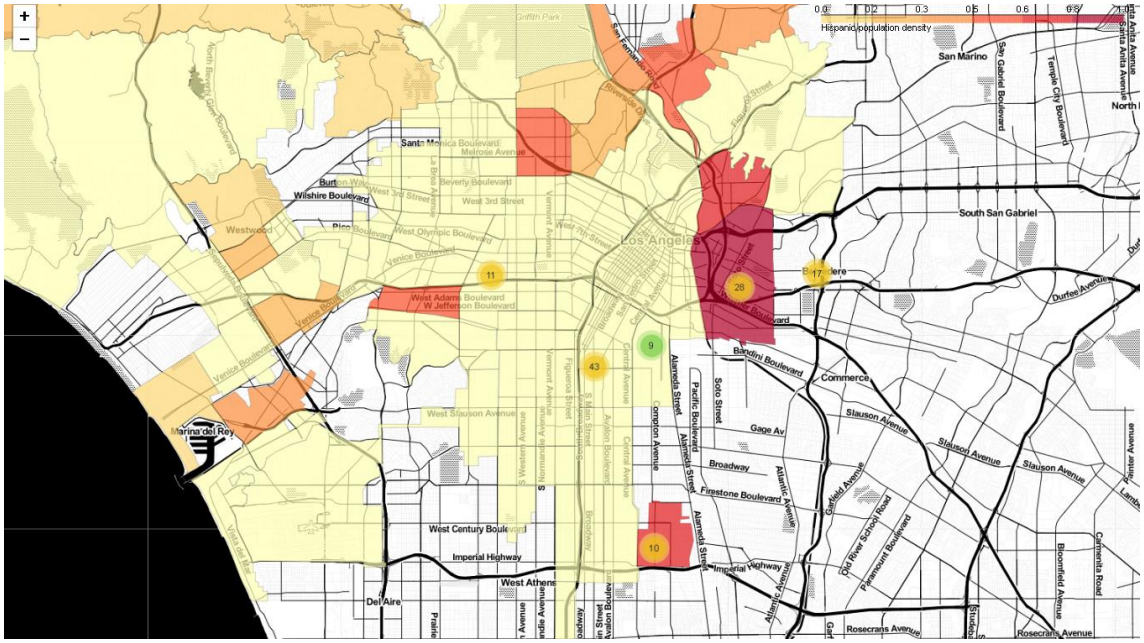
Figure 15. Code used to add the markers and visualize them over the choropleth map.

```
from folium import plugins

# instantiate a mark cluster object for the incidents in the dataframe
sch_mkr = plugins.MarkerCluster().add_to(LA_map)

# loop through the dataframe and add each data point to the mark cluster
for lat, lng, in zip(df_clust.Latitude, df_clust.Longitude):
    folium.Marker(
        location=[lat, lng],
        icon=None
    ).add_to(sch_mkr)

# display map
LA_map
```



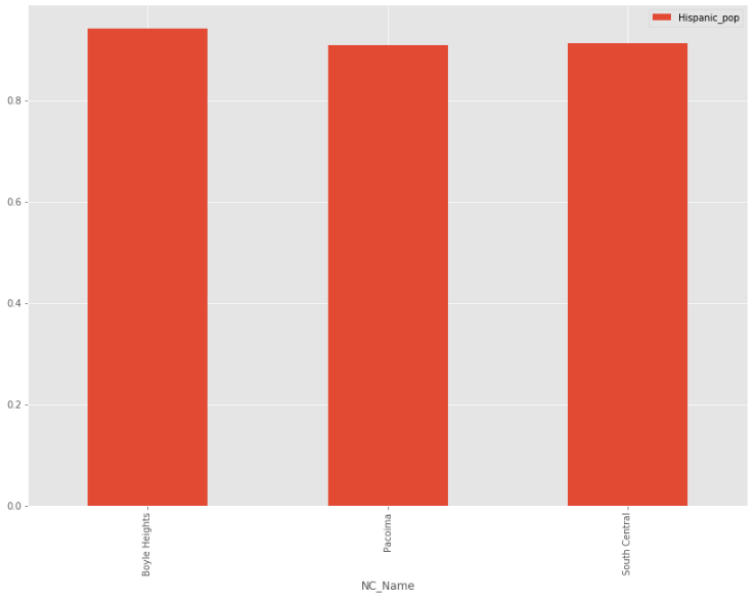
- The markers were added correctly over Los Angeles city but, how many of them are in the wanted neighborhood?

Figure 16. Zoom-in on the Choropleth map with markers distribution.



A zoom-in as been made to observe the schools distribution, with the amazing coincidence of the majority location on the same area than the dark red surface. For a conclusion the name of the neighbourhood shall be known. So, a barchar is consequently made only with the neighborhoods with a density greater than 90% which corresponds to the ones discussed at the figure 14.

Figure 17. Barchar of the county neighborhoods with a density greater than 90.



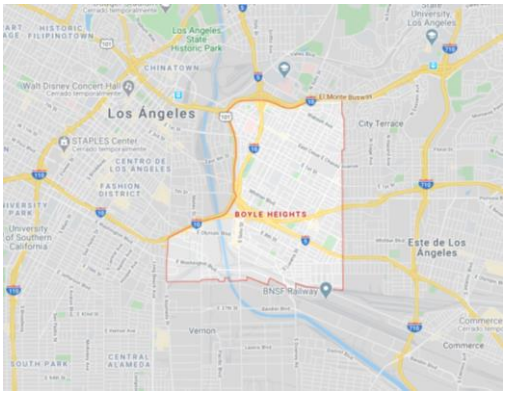
- The resultant neighborhoods are:
  - Boyle Heighths
  - Pacoima
  - South central

To finalize the discussion a verification with Google Maps is necessary not only to locate which one between them is located at the city of Los Angeles, NOT at Los Angeles County. But also as a validation for the work done by verifying the name and area correspondence.

4.2. Verification and validation

Figure 18. Comparison of the chosen neighbourhood name and area in Google maps.

- Boyle Heights.



The neighbourhood “Boule Heigts” is the one located at the city of Los Angeles with similar limits but not with a desired exactitude.

5. CONCLUSSION

After the analysis by DBSCAN clustering, dropping all those with business of the same kind that the wanted, visualizing the top neighborhoods and finally filling it with the schools' markers from the remaining clusters it is imperatively that the best choice would be the "Boyle Heights" neighborhood. However, it is only taking into account the schools at Los Angeles City (metropolitan area), not the County of Los Angeles in which we have similar high Hispanics' density of population.

Also to mention but not less important, the verification of the results vs a Google maps search led us to conclude that the file .json obtained from the Mapping L.A. Boundaries API could not be the correct for this usage. Moreover, the scope and purpose of the project is an acceptable result.

Finally, it could be analysed as well with other variables as security, crimes, neighborhood income, venues, etc. in order to have a more complete and reliable conclusion.