Capstone Project Presentation

THE BATTLE OF THE NEIGBOORHODS

Autor:
Rafael Rojas Cárdenas

Objective

- To define the best neighborhood to open a Mexican food business at the city of Los Angeles taking into accout the number of schools, current Mexican restaurants in the city and Hispanic density of population as variables.

Background

- 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%).

Business 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.

Problem aproach

- 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?

Methodology Check-list

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

Data sources.

- 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].

- Mexican Restaurants at Los Angeles City. Only taken into account those restaurants collected at the kaggle dataset from the next website [https://www.kaggle.com/datafiniti/fast-food-restaurants].

- 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

Data Pre-processing

How many schools are located in the neighbourhood?

Schools' final data-frame

	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

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

Data Pre-processing

What are the ethnics of the neighbourhood?

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.

Data Pre-processing

How many restaurants of the kind are located in the neighbourhood?

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.

Data Pre-processing

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

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.

Methodology

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

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
```

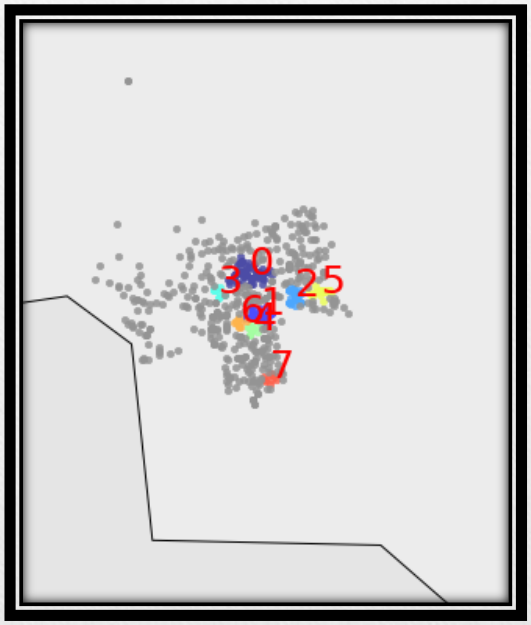
```
import sklearn.utils
from sklearn.preprocessing import StandardScaler
sklearn.utils.check_random_state(1000)
Clus_dataSet = df_clust[['Longitude','Latitude']]
Clus_dataSet = np.nan_to_num(Clus_dataSet)
Clus_dataSet = StandardScaler().fit_transform(Clus_dataSet)

# 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

Methodology

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

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

Methodology

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

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

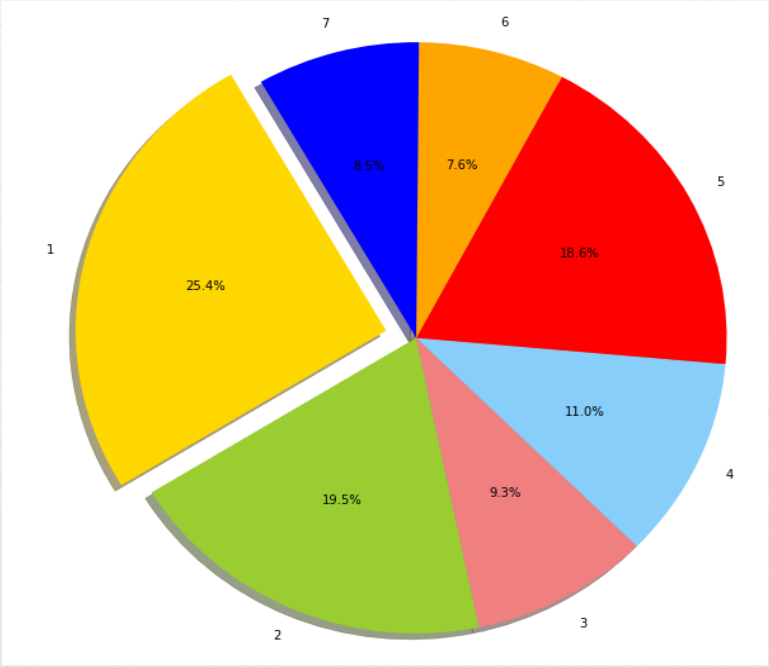
118 rows × 6 columns

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

Methodology

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

Pie chart of clusters magnitude in percentage.

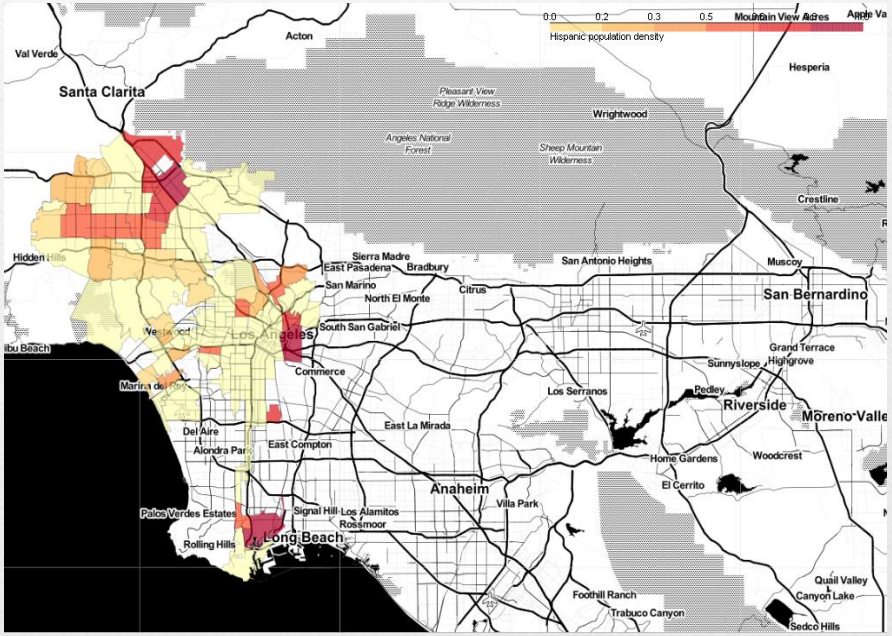


- The clusters 1, 2 and 5 hold the biggest quantity of schools with a 25.4%, 19.5% and 18.6% respectively.
- The colours of these clusters at the map in figure 10 are:

  1 : Blue
  3 : Light Blue
  5 : Yellow

Methodology

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

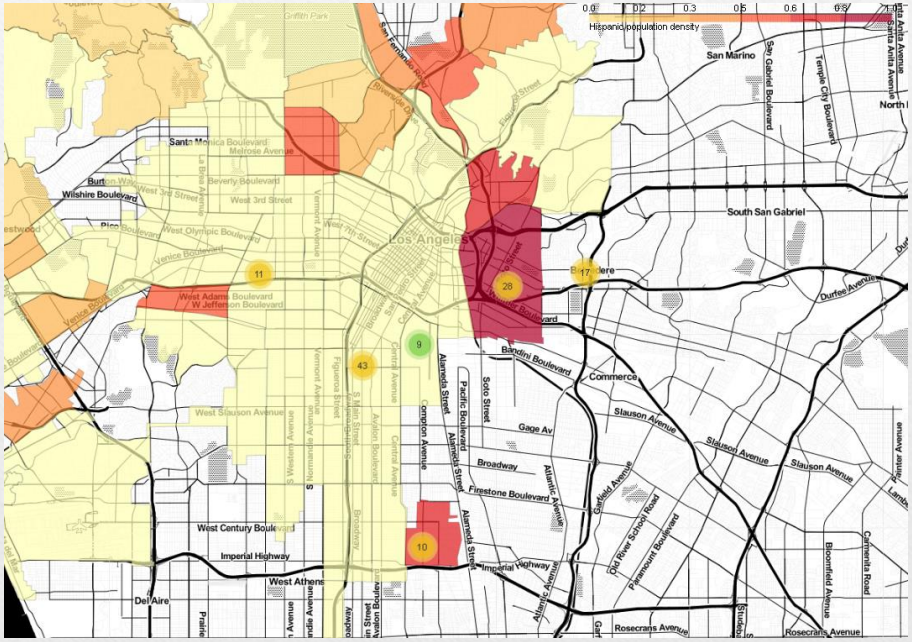


- For the county of Los Angeles, three neighborhoods are highlighted in dark red due to a density greater than 90%. However, only one of this three is in the city of Los Angeles.

RESULTS AND DISCUSSION

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

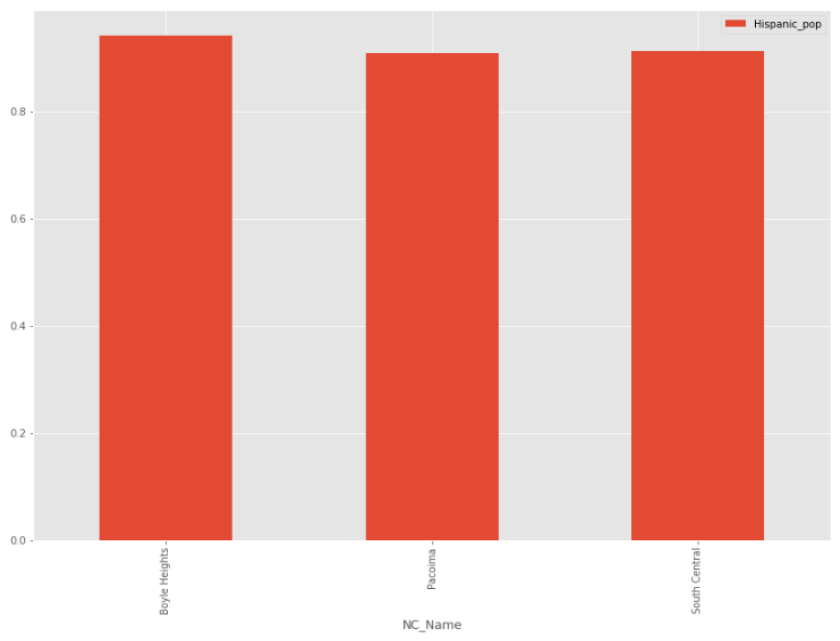
Zoom-in on the Choropleth map with markers distribution



- 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%

RESULTS AND DISCUSSION

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

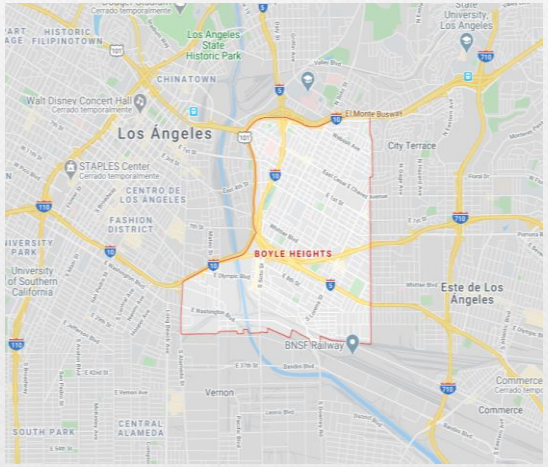


- The resultant neighborhoods are:
  - Boyle Heighths
  - Pacoima
  - South central

Verification and validation

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

Boyle Heights comparison.



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

CONCLUSSIONS

- 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.