Tagging Neighborhoods around Educational Institutions in the battle of neighborhoods

Alfonso Pereda Gálvez

October 2019

That ? (Issue)

Unknowing by investors of the 10 best types of service companies (cafes, restaurants, hotels ...) that are developed around the locations of educational institutions in a city.

For what?

The focus of the project on "The battle of the neighborhoods" is to provide quality and timely information to reduce the risk of investors in service companies (cafes, restaurants, hotels, gyms, etc.) that can be developed or acquired in the neighborhoods close to educational institutions as in this case in New York City.

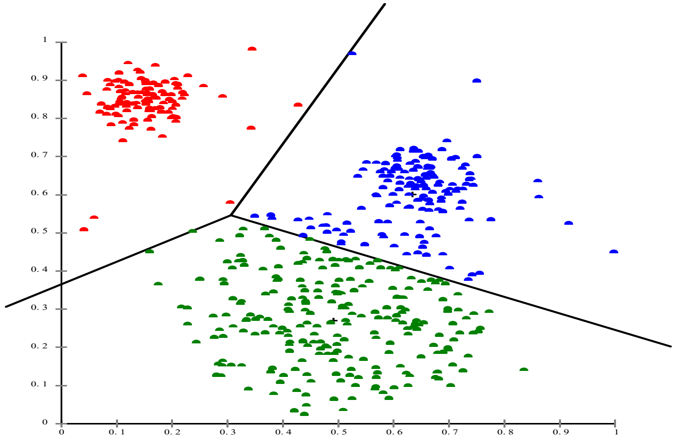
How? (Solve) - Data Perspective

Datasets with key data of Education Institutions (ADDRESS, Population, Number of enrolled annually, It offers dormitories, own casifications of the sector: LEVEL_, TYPE, NAICS, etc.) as the one provided by "Homeland Infrastructure Foundation-Level Data (HIFLD)“.

Data by FOURSQUARE neighborhood of geomarketing type (Company type, distance from a geographical location point, ranking of visits or "check" regarding pivot type point.

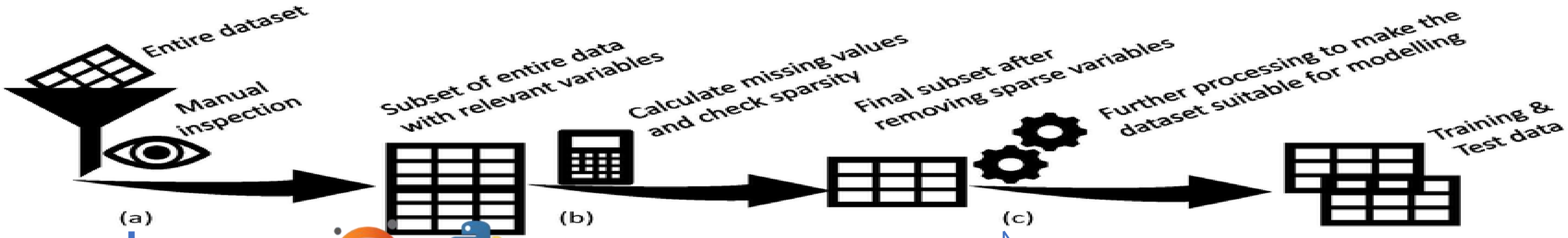
How? - Machine Learning Techniques Perspective

SEGMENTATION AND LABELING WITH NON-SUPERVISED LEARNING TECHNIQUE SUCH AS CLUSTERIZATION WITH K-MEANS



APPLICATION OF METHODOLOGICAL PROCESS OF DATA SCIENCE IN THE DEVELOPMENT OF THE PROJECT WITH THE STAGES: UNDERSTANDING THE PROBLEM, DATA ACQUISITION, DATA WRANGLING, EXPLORATORY ANALYSIS, MODELING AND EVALUATION, RESULTS AND CONCLUSIONS.

Cluster Development –Data Wrangling



DataSet Columns

Columns/ Type (1-15)	Columns/ Type (16-30)	Columns/ Type (31-45)
X, type number	COUNTYFIPS, Int64.Type	HI_OFFER, Int64.Type
Y, type number	COUNTRY, type text	DEG_GRANT, Int64.Type
OBJECTID, Int64.Type	LATITUDE, type number	LOCALE, Int64.Type
IPEDSID, Int64.Type	LONGITUDE, type number	CLOSE_DATE, type text
NAME, type text	NAICS_CODE, Int64.Type	MERGE_ID, Int64.Type
ADDRESS, type text	NAICS_DESC, type text	ALIAS, type text
CITY, type text	SOURCE, type text	SIZE_SET, Int64.Type
STATE, type text	SOURCEDATE, type datetime	INST_SIZE, Int64.Type
ZIP, Int64.Type	VAL_METHOD, type text	PT_ENROLL, Int64.Type
ZIP4, type text	VAL_DATE, type datetime	FT_ENROLL, Int64.Type
TELEPHONE, type text	WEBSITE, type text	TOT_ENROLL, Int64.Type
TYPE, Int64.Type	STFIPS, Int64.Type	HOUSING, Int64.Type
STATUS, type text	COFIPS, Int64.Type	DORM_CAP, Int64.Type
POPULATION, Int64.Type	SECTOR, Int64.Type	TOT_EMP, Int64.Type
COUNTY, type text	LEVEL_, Int64.Type	SHELTER_ID, type text

Data Wrangling

```
#columns
nydata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 86 entries, 0 to 85
Data columns (total 15 columns):
ADDRESS            86 non-null object
NAME               86 non-null object
STATUS             86 non-null object
LATITUDE           86 non-null float64
LONGITUDE          86 non-null float64
NAICS_DESC         86 non-null object
NAICS_CODE         86 non-null int64
TYPE               86 non-null int64
POPULATION         86 non-null int64
LEVEL_             86 non-null int64
INST_SIZE          86 non-null int64
HOUSE_IN           86 non-null int64
TOT_ENROLL         86 non-null int64
TOT_EMP            86 non-null int64
NAICS_DESCshort    86 non-null object
dtypes: float64(2), int64(8), object(5)
memory usage: 10.2+ KB
```

	ADDRESS	NAME	STATUS	LATITUDE	LONGITUDE	NAICS_DESC	NAICS_CODE	TYPE	POPULATION	LEVEL_	INST_SIZE	HOUSE_IN	TOT_
0	145 E BROADWAY	MESIVTHA TIFERETH JERUSALEM OF AMERICA	A	40.713812	-73.991271	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	2	81	1	1	1	
1	290 MADISON AVENUE 5TH FLOOR	CARSTEN INSTITUTE OF COSMETOLOGY	A	40.751893	-73.980278	COSMETOLOGY AND BARBER SCHOOLS	611511	3	108	3	1	0	
2	211 WEST 61ST STREET	AMERICAN MUSICAL AND DRAMATIC ACADEMY	A	40.772309	-73.987638	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	2	1962	1	2	1	
3	3 EAST 43 STREET	BERKELEY COLLEGE-NEW YORK	A	40.753993	-73.979434	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	3	4155	1	2	0	
4	205 EAST 42ND STREET	CUNY SYSTEM OFFICE	A	40.750855	-73.973595	EDUCATIONAL SUPPORT SERVICES	611710	1	1184	1	-2	0	

Cluster Development –Exploratory Analysis of Variables

Descriptive statistics of the New York data set

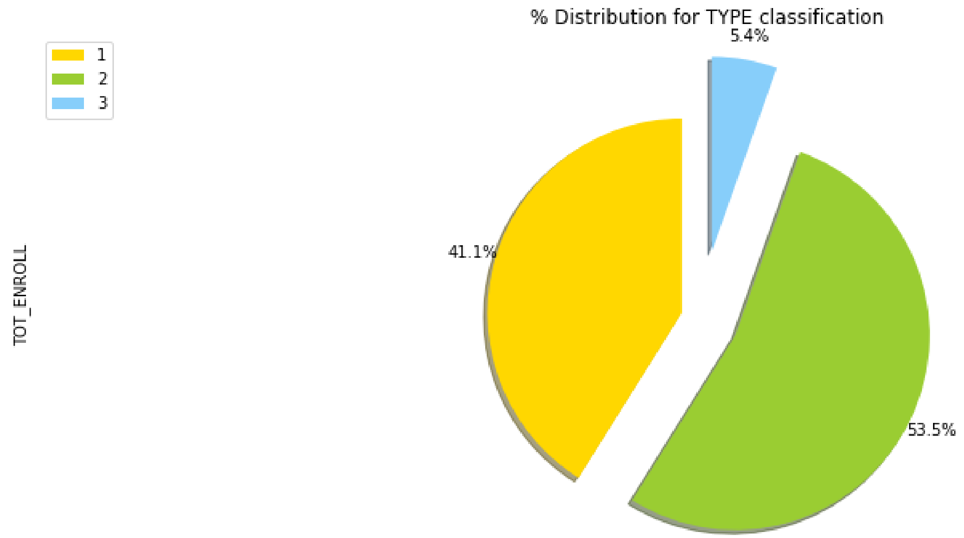
	TYPE	POPULATION	LEVEL_	INST_SIZE	DORM_CAP	TOT_ENROLL	TOT_EMP
count	86.000000	86.000000	86.000000	86.000000	86.000000	86.000000	86.000000
mean	2.313953	4555.104651	1.686047	1.546512	-20.686047	3284.965116	1212.058140
std	0.673212	10997.498020	0.857652	1.289386	2248.987899	7954.533200	3570.209898
min	1.000000	-999.000000	1.000000	-2.000000	-999.000000	-999.000000	-999.000000
25%	2.000000	187.750000	1.000000	1.000000	-999.000000	116.000000	32.500000
50%	2.000000	673.500000	1.000000	1.000000	-999.000000	471.000000	123.500000
75%	3.000000	2089.500000	2.750000	2.000000	218.000000	1438.250000	634.000000
max	3.000000	73997.000000	3.000000	5.000000	13075.000000	51123.000000	22874.000000

jupyter python NumPy

Visual Statistical Analysis

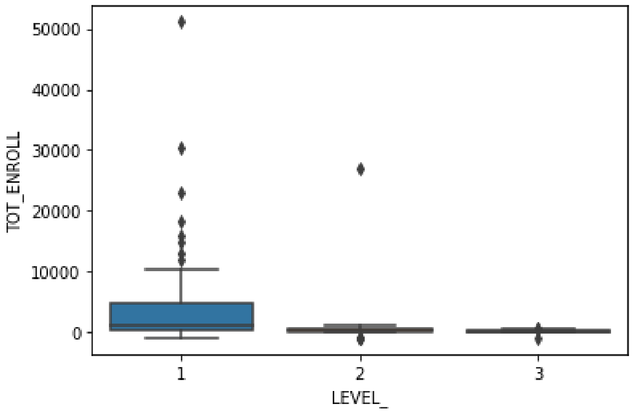
	TOT_ENROLL
NAICS_DESCshort	
COLLE	251242
COMPU	1281
COSME	1460
EDUCA	0
FINE	764
JUNIO	27017
OTHER	1742

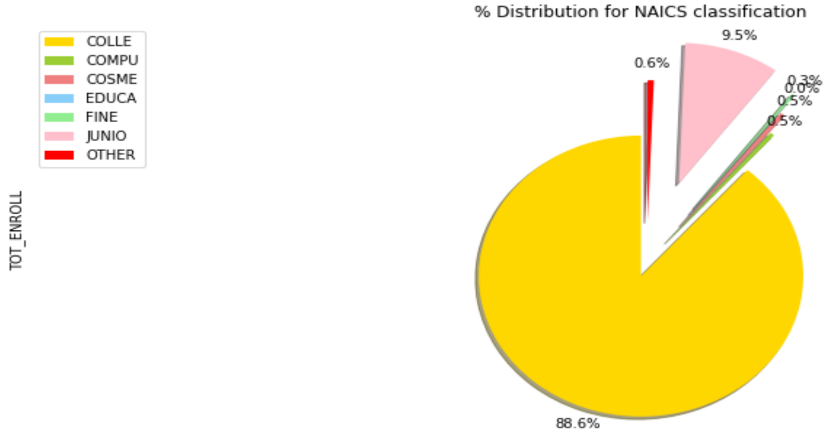
	TOT_ENROLL
TYPE	
1	116061
2	151152
3	15294



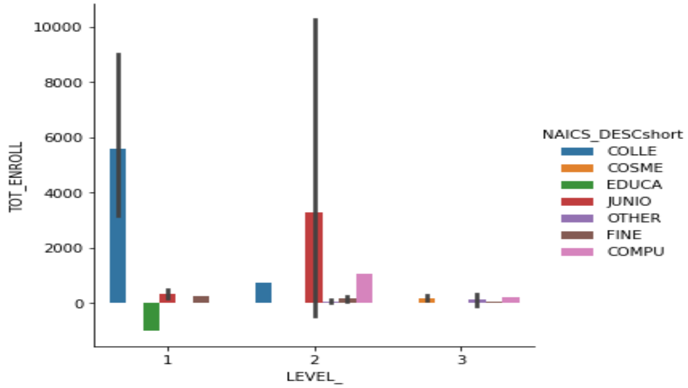
Cluster Development –Exploratory Analysis of Variables

Bi-variable analysis by selecting classification variables of educational institutions with the aggregation data of each observation.

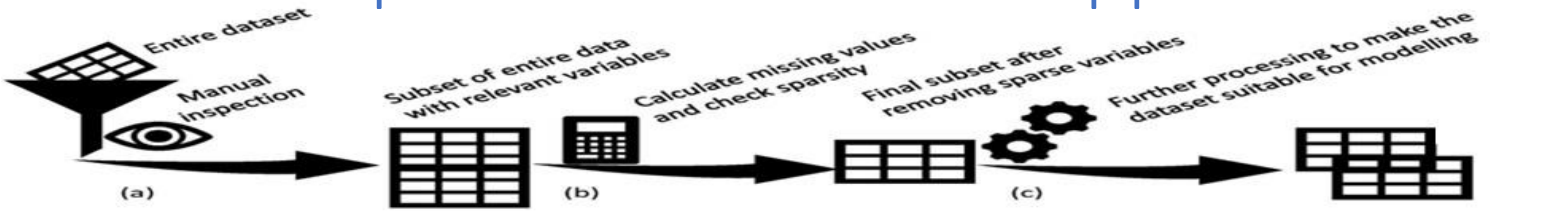




Analysis of three variables by selecting classification variables of educational institutions with the aggregation data of each observation



Cluster Development –DataSet for Applied Model



	ADDRESS	NAME	STATUS	LATITUDE	LONGITUDE	NAICS_DESC	NAICS_CODE	TYPE	POPULATION	LEVEL_	INST_SIZE	HOUSE_IN	TOT_
0	145 E BROADWAY	MESIVTHA TIFERETH JERUSALEM OF AMERICA	A	40.713812	-73.991271	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	2	81	1	1	1	
1	290 MADISON AVENUE 5TH FLOOR	CARSTEN INSTITUTE OF COSMETOLOGY	A	40.751893	-73.980278	COSMETOLOGY AND BARBER SCHOOLS	611511	3	108	3	1	0	
2	211 WEST 61ST STREET	AMERICAN MUSICAL AND DRAMATIC ACADEMY	A	40.772309	-73.987638	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	2	1962	1	2	1	
3	3 EAST 43 STREET	BERKELEY COLLEGE-NEW YORK	A	40.753993	-73.979434	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	3	4155	1	2	0	
4	205 EAST 42ND STREET	CUNY SYSTEM OFFICE	A	40.750855	-73.973595	EDUCATIONAL SUPPORT SERVICES	611710	1	1184	1	-2	0	

	Neighborhood	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	1 PACE PLAZA	Sandwich Place	Coffee Shop	Plaza	Park	Café	Pizza Place	Gym	Bar	Hotel	Restaurant
1	11 PARK PLACE 4TH FLOOR	Hotel	Coffee Shop	Italian Restaurant	Café	Hotel Bar	Bar	Sandwich Place	Gym	Gym / Fitness Center	Park
2	110 WILLIAM ST. 19TH FL	Coffee Shop	Hotel	Italian Restaurant	Deli / Bodega	American Restaurant	Sandwich Place	Juice Bar	Café	Pizza Place	Mediterranean Restaurant
3	111 FRANKLIN ST	Gym / Fitness Center	Italian Restaurant	Cocktail Bar	French Restaurant	Boutique	Theater	Spa	Bakery	Coffee Shop	Diner
4	115 WEST 27TH STREET, 11TH FLOOR	Hotel	Flower Shop	Gym	Bar	Coffee Shop	Japanese Restaurant	Performing Arts Venue	Martial Arts Dojo	Sandwich Place	Gym / Fitness Center
5	12 E 53RD ST	Boutique	Jewelry Store	Hotel	Gym	Italian Restaurant	Steakhouse	Coffee Shop	Spa	Gift Shop	Greek Restaurant

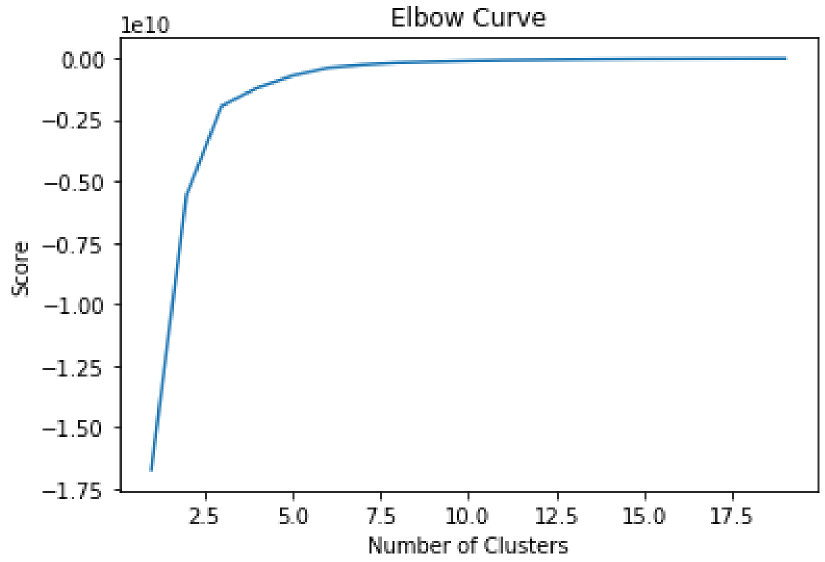
Data Wrangling

ADDRESS	1300 YORK AVE, C-114
NAME	WEILL CORNELL MEDICAL COLLEGE
STATUS	A
LATITUDE	40.7649
LONGITUDE	-73.9547
NAICS_DESC	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS
NAICS_CODE	611310
TYPE	2
POPULATION	7976
LEVEL_	1
INST_SIZE	2
HOUSE_IN	1
TOT_ENROLL	1107
TOT_EMP	6869
NAICS_DESCshort	COLLE
1st Most Common Venue	Coffee Shop
2nd Most Common Venue	Café
3rd Most Common Venue	Japanese Restaurant
4th Most Common Venue	Residential Building (Apartment / Condo)
5th Most Common Venue	Chinese Restaurant
6th Most Common Venue	Sushi Restaurant
7th Most Common Venue	Mexican Restaurant
8th Most Common Venue	Club House
9th Most Common Venue	Cocktail Bar
10th Most Common Venue	Sandwich Place

Model & Evaluation - Number of clusters of the Model (K)

K=4 (Number of Clusters)

Possible range of K values for use in the K-means Model according to the Elwod Curve score technique is in the range of 2 to 4 clusters.



Model K-means applied

```
Model with K=4 : KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=4, n_init=10, n_jobs=None, precompute_distances='auto',
    random_state=None, tol=0.0001, verbose=0)
```

Obtaining the following distribution of the observations in the clusters

	nroclust	quantity
0	0	69
1	1	2
2	2	6
3	3	9

```
{0: Int64Index([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 12, 13, 14, 16, 17, 18, 19,
             21, 22, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39,
             40, 41, 42, 43, 44, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57,
             58, 59, 61, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 78, 80, 81,
             84],
            dtype='int64'),
 1: Int64Index([79, 85], dtype='int64'),
 2: Int64Index([0, 15, 20, 63, 77, 82], dtype='int64'),
 3: Int64Index([11, 23, 31, 46, 60, 62, 64, 76, 83], dtype='int64')}
```

Statistical metrics of the numerical variables of the Dataset applied to the Model

0

	LATITUDE	LONGITUDE	NAICS_CODE	TYPE	POPULATION	LEVEL_	INST_SIZE	HOUSE_IN	TOT_ENROLL	TOT_EMP	Cluster Labels
count	69.000000	69.000000	69.000000	69.000000	69.000000	69.000000	69.000000	69.000000	69.000000	69.000000	69.0
mean	40.752876	-73.985512	611393.391304	2.478261	671.130435	1.840580	1.043478	0.275362	468.507246	130.231884	0.0
std	0.028596	0.017423	131.680788	0.584322	960.697922	0.884893	0.695252	0.449969	817.300661	396.956404	0.0
min	40.705781	-74.015157	611210.000000	1.000000	-999.000000	1.000000	-2.000000	0.000000	-999.000000	-999.000000	0.0
25%	40.739862	-73.995722	611310.000000	2.000000	122.000000	1.000000	1.000000	0.000000	103.000000	25.000000	0.0
50%	40.750855	-73.987638	611310.000000	3.000000	397.000000	2.000000	1.000000	0.000000	268.000000	74.000000	0.0
75%	40.762927	-73.977344	611519.000000	3.000000	1081.000000	3.000000	1.000000	1.000000	634.000000	244.000000	0.0
max	40.833434	-73.940306	611710.000000	3.000000	4155.000000	3.000000	2.000000	1.000000	3635.000000	1552.000000	0.0

1

	LATITUDE	LONGITUDE	NAICS_CODE	TYPE	POPULATION	LEVEL_	INST_SIZE	HOUSE_IN	TOT_ENROLL	TOT_EMP	Cluster Labels
count	2.000000	2.000000	2.0	2.0	2.000000	2.0	2.0	2.0	2.00000	2.000000	2.0
mean	40.768869	-73.979575	611310.0	2.0	62070.500000	1.0	5.0	1.0	40788.50000	21282.000000	1.0
std	0.055744	0.025017	0.0	0.0	16866.618052	0.0	0.0	0.0	14615.19006	2251.427991	0.0
min	40.729452	-73.997264	611310.0	2.0	50144.000000	1.0	5.0	1.0	30454.00000	19690.000000	1.0
25%	40.749160	-73.988419	611310.0	2.0	56107.250000	1.0	5.0	1.0	35621.25000	20486.000000	1.0
50%	40.768869	-73.979575	611310.0	2.0	62070.500000	1.0	5.0	1.0	40788.50000	21282.000000	1.0
75%	40.788577	-73.970730	611310.0	2.0	68033.750000	1.0	5.0	1.0	45955.75000	22078.000000	1.0
max	40.808286	-73.961885	611310.0	2.0	73997.000000	1.0	5.0	1.0	51123.00000	22874.000000	1.0

2

	LATITUDE	LONGITUDE	NAICS_CODE	TYPE	POPULATION	LEVEL_	INST_SIZE	HOUSE_IN	TOT_ENROLL	TOT_EMP	Cluster Labels
count	9.000000	9.000000	9.0	9.000000	9.000000	9.0	9.000000	9.0	9.000000	9.000000	9.0
mean	40.771031	-73.971596	611310.0	1.888889	10135.555556	1.0	2.777778	1.0	6284.000000	3851.555556	3.0
std	0.038702	0.023347	0.0	0.600925	3137.910695	0.0	0.833333	0.0	3801.764985	3491.478774	0.0
min	40.735498	-73.997158	611310.0	1.000000	6144.000000	1.0	2.000000	1.0	1107.000000	1491.000000	3.0
25%	40.747310	-73.989488	611310.0	2.000000	7976.000000	1.0	2.000000	1.0	4393.000000	1751.000000	3.0
50%	40.753362	-73.982240	611310.0	2.000000	9648.000000	1.0	3.000000	1.0	6330.000000	2597.000000	3.0
75%	40.789801	-73.954738	611310.0	2.000000	13216.000000	1.0	3.000000	1.0	8846.000000	3347.000000	3.0
max	40.850800	-73.928541	611310.0	3.000000	14505.000000	1.0	4.000000	1.0	11908.000000	12008.000000	3.0

3

	LATITUDE	LONGITUDE	NAICS_CODE	TYPE	POPULATION	LEVEL_	INST_SIZE	HOUSE_IN	TOT_ENROLL	TOT_EMP	Cluster Labels
count	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.0
mean	40.754924	-73.983978	611293.333333	1.166667	21678.333333	1.166667	4.333333	0.833333	18674.500000	3003.833333	2.0
std	0.040059	0.023313	40.824829	0.408248	5619.261541	0.408248	0.516398	0.408248	5316.845333	602.012431	0.0
min	40.711710	-74.011826	611210.000000	1.000000	16256.000000	1.000000	4.000000	0.000000	12986.000000	2326.000000	2.0
25%	40.724152	-74.000756	611310.000000	1.000000	17556.500000	1.000000	4.000000	1.000000	15125.750000	2596.500000	2.0
50%	40.754453	-73.985910	611310.000000	1.000000	19791.000000	1.000000	4.000000	1.000000	17145.000000	2941.000000	2.0
75%	40.769927	-73.969451	611310.000000	1.000000	25461.250000	1.000000	4.750000	1.000000	21826.000000	3236.750000	2.0
max	40.819794	-73.950550	611310.000000	2.000000	30069.000000	2.000000	5.000000	1.000000	26932.000000	3998.000000	2.0

Applying Labeling Prediction

Jupyter python NumPy

```python
case0 =[3,4155,1,2,0,3635,520]      # idx 3  ; [ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 12, 13, 14, 16, 17, 18, 19,
          #  21, 22, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39,
          #  40, 41, 42, 43, 44, 45, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57,
          # 58, 59, 61, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 78, 80, 81,84]
case1 =[2,50144,1,5,1,30454,19620]        # idx 85  [79,85]
case2 =[2,16256,1,4,1,12986,3270]    # idx  0;  [0, 15, 20, 63, 77, 82]
case3=[2,7976,1,2,1,1107,6869]      # idx  11;  [11, 23, 31, 46, 60, 62, 64, 76, 83]

X_new = np.array([case0])  ## We assign the stock vector
new_labels = kmeans.predict(X_new)  ## Label of the Cluster to which it corresponds
print("cluster to which it belongs : ",new_labels)   ## Print the assigned cluster label
print("with features-> TYPE: {} , POPULATION: {}, LEVEL_: {}, INST_SIZE: {}, HOUSE_IN: {}, TOT_ENROLL: {}, TOT_EMP: {}".format(
    case0[0],case0[1],case0[2],case0[3],case0[4],case0[5],case0[6]))  ## Print the input characteristics

X_new = np.array([case1])  ## We assign the stock vector
new_labels = kmeans.predict(X_new)  ## Label of the Cluster to which it corresponds
print("cluster to which it belongs : ",new_labels)   ## Print the assigned cluster label
print("with features-> TYPE: {} , POPULATION: {}, LEVEL_: {}, INST_SIZE: {}, HOUSE_IN: {}, TOT_ENROLL: {}, TOT_EMP: {}".format(
    case1[0],case1[1],case1[2],case1[3],case1[4],case1[5],case1[6]))  ## Print the input characteristics

X_new = np.array([case2])  ## We assign the stock vector
new_labels = kmeans.predict(X_new)  ## Label of the Cluster to which it corresponds
print("cluster to which it belongs : ",new_labels)   ## Print the assigned cluster label
print("with features-> TYPE: {} , POPULATION: {}, LEVEL_: {}, INST_SIZE: {}, HOUSE_IN: {}, TOT_ENROLL: {}, TOT_EMP: {}".format(
    case2[0],case2[1],case2[2],case2[3],case2[4],case2[5],case2[6]))  ## Print the input characteristics

X_new = np.array([case3])  ## We assign the stock vector
new_labels = kmeans.predict(X_new)  ## Label of the Cluster to which it corresponds
print("cluster to which it belongs : ",new_labels)   ## Print the assigned cluster label
print("with features-> TYPE: {} , POPULATION: {}, LEVEL_: {}, INST_SIZE: {}, HOUSE_IN: {}, TOT_ENROLL: {}, TOT_EMP: {}".format(
    case3[0],case3[1],case3[2],case3[3],case3[4],case3[5],case3[6]))  ## Print the input characteristics
```

Predominant Features (X)

Labelling Prediction (Y)

```
cluster to which it belongs :  [0]
with features-> TYPE: 3 , POPULATION: 4155, LEVEL_: 1, INST_SIZE: 2, HOUSE_IN: 0, TOT_ENROLL: 3635, TOT_EMP: 520
cluster to which it belongs :  [1]
with features-> TYPE: 2 , POPULATION: 50144, LEVEL_: 1, INST_SIZE: 5, HOUSE_IN: 1, TOT_ENROLL: 30454, TOT_EMP: 19620
cluster to which it belongs :  [2]
with features-> TYPE: 2 , POPULATION: 16256, LEVEL_: 1, INST_SIZE: 4, HOUSE_IN: 1, TOT_ENROLL: 12986, TOT_EMP: 3270
cluster to which it belongs :  [3]
with features-> TYPE: 2 , POPULATION: 7976, LEVEL_: 1, INST_SIZE: 2, HOUSE_IN: 1, TOT_ENROLL: 1107, TOT_EMP: 6869
```

Results: Top n of neighboring business premises to each cluster

```
Top 1- Cluster 0:  ['American Restaurant' 'Art Gallery' 'Bookstore' 'Boutique'
 'Clothing Store' 'Cocktail Bar' 'Coffee Shop' 'Deli / Bodega'
 'Donut Shop' 'Gym / Fitness Center' 'Hotel' 'Italian Restaurant'
 'Japanese Restaurant' 'Korean Restaurant' 'Martial Arts Dojo' 'Park'
 'Sandwich Place' 'Shoe Store' 'Tennis Court' 'Theater']
Top 1- Cluster 1:  ['Coffee Shop']
Top 1- Cluster 2:  ['Café' 'Coffee Shop' 'Indian Restaurant' "Men's Store" 'Sandwich Place'
 "Women's Store"]
Top 1- Cluster 3:  ['Coffee Shop' 'Deli / Bodega' 'Hotel' 'Italian Restaurant'
 'Korean Restaurant' 'Park' 'Seafood Restaurant' 'Thrift / Vintage Store']
```

Other results

Characterization and segmentation of the New York neighborhoods identifying commercial premises (sale of products or services) around the study centers within a radius of 300 meters.

Identification of neighborhood patterns associated with the educational institutions of the city of New York allowing to answer this question of the business problem defined in this project.

Discreet labeling by application of the k-means predictor model, by providing as input data values of the predominant variables of the established model.

Uso del conocimiento de los mercados y la capacidad de centrar los esfuerzos en determinados segmentos del mercado objetivo a través de los datos proporcionados por empresas de "geomarketing" como FOURSQUARE.

Conclusions

Clusters are formed by the population density of the variables POPULATION, TOT_ENROLL, TOT_EMP and not by the classifications of educational institutions defined by TYPE, LEVEL_, NAICS. Therefore, cluster 0 with 69 observations has the association of more types of businesses than the other clusters and a diversity of educational institutions by TYPE, LEVEL and NAICS.

The use of K-means to neighborhood problems is useful when combining the geographic data associated with demographic characteristics such as educational institutions with the neighborhood data of business premises provided with the FOURSQUARE API, that is to say enhances the labeling results from at least two perspectives, the first one defined by the clusters obtained and the association with the neighborhood, this provides the characterization of the groups obtained. The second point of view is the prediction when entering new data of the determining variables of the k-means model, we obtain a labeling that corresponds to one of the clusters and consequently we obtain characteristics and neighboring businesses (Top 10 in this project).

Conclusions

K-mean is an algorithm that manages to discover new relationships between features, or it helps us to test or decline hypotheses we have of our business.

Favorable results in the identification of crowded business groups (top 10) developed around educational institutions according to TYPE, POPULATION, LEVEL_, INST_SIZE, HOUSE_IN, TOT_ENROLL, TOT_EMP and geographic location given as a pivot.

The clustering model can be improved by associating more data with each observation, since the purpose is to support the decision making of investors, data such as income, expenses, utility of neighboring businesses would be key variables in the prediction and labeling of clusters that are formed with this new data entry.