Tagging Neighborhoods around Educational Institutions in the battle of neighborhoods

Alfonso Pereda Gálvez

October 6, 2019

1. Introduction to the Business Problem

In general, investors seek to reduce the risk of their investment, so obtaining clear and accurate information (processed data) allows you to assess the risk in the evaluation of your investment and make your decision. Considering the service businesses such as coffee shops, restaurants, hotels, bars, it is located in centers with high population density and that are concentrated not only by housing areas but also by the economic activities that take place in these neighborhoods. The main idea is that we consider educational institutions (Colleges, Universities, technical institutes) as points of concentration of people, we consider the location information of these educational institutions and some characteristics (Population, staff, qualifications determined by their activity in the education) as foci where you can invest for the development of businesses around them, inferring that it is important to indicate the analysis of location and the determination of the variables and characteristics that can change the selection of an investment in a commercial or business type business. services around a focus like the one we indicated is favorable to know in the evaluation of the investment.

Then, the support and provision of this characterization and location patterns of a business can limit the investment risk and is currently decisive, especially if the objective is of small or medium investors who need to project specific income and expenses of the selected alternative. This project attempts to provide as a result and achieve the following objectives:

- Characterization and segmentation of New York neighborhoods located in commercial premises (sale of products or services) around the study centers.
- Recognize neighborhood patterns associated with the approach of educational institutions. Is it important to invest in a cafeteria or gym near a university or college, associated with the number of people and businesses associated with a study center?
- Discreet labeling of business groups around educational institutions. This classification is characteristic of the variables and the location of the observations of the data set used.
- For this, we will apply a K-mean Cluster model to the dataset of New York educational institutions as a case that can be extended to other cities.

- Apply knowledge of markets and the ability to focus efforts on certain segments of the target market through the data provided from "Geomarketing" companies such as FOURSQUARE.

1.1. Business Problem

This project approach in "The battle of the neighborhoods" is to provide quality and timely information to reduce the risk of investors in service businesses (Cafeterias, Restaurants, Hotels, gymnasiums, etc.) that can be developed or acquired in neighboring neighborhoods to educational institutions. The questions we can answer are:

a) What are the main n companies (n: 5,10) that develop around a university, considering its population, number of staff, has dormitories?
b) What types of businesses are developed around schools for children and that characterize them according to their population, enrollment per year?
c) In what type of clusters are a bar in general installed within the first 2 businesses?

These types of problems try to solve the result of the project.

1.2. Interested audience

The interest groups are varied, which can be people or institutions, here are some:

- Investors of small and medium enterprises require information to determine what types of companies are developed in neighborhoods close to educational institutions.
- Evaluation Professionals oriented to the economic development of the neighborhoods of a city, using the characterization of groups of these.
- Companies looking for (local) locations to install franchise-based businesses (Starbucks, Subway, Pizza Hut ...)
- Financial companies to assess and limit the risks associated with loans to investors who wish to invest in neighborhoods associated with this project.
- Real estate companies for the sustained development of real estate construction or renovation projects in the neighborhoods associated with the results of the grouping.

These are some of the interest groups to which the solution of this project is oriented.

2. Description of the data and use in solving the problem

For this project, we will use a K Means Grouping model, to group types of businesses located around the United States Study Centers, specifically New York. The data source is the website of "Homeland Infrastructure Foundation-Level Data (HIFLD)" which considers as Centers of Studies to Universities, Colleges, Institutes, for our case we consider the data to this reference of the city of New York.

The Source is https://hifld-geoplatform.opendata.arcgis.com/datasets/colleges-and-universities, we will use a data frame with 7150 observations in the following 45 variables (Last update 2 months ago).

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

There are variables of the Dataset that does not contribute to the analysis of business characterization that one wishes to obtain from the data, such as columns X and Y, which are LATITUDE and LONGITUDE. Additionally, we will add data from FOURSQUARE, about neighboring locations to selected study centers in New York City.

Description of columns to use

Column Name	Type	Description
ADDRESS	Text	The Dirección de una institución educativa de EEUU, which is unique.
NAME	Text	Name of the education institution.
CITY	Text	City where the institution is located.
STATE	Text	State where the institution is located.
TYPE	Int	educational level classification.
POPULATION	Int	Población del Centro de estudios.
LATITUDE	Float	Geospatial Coordinate.
LONGITUDE	Float	Geospatial Coordinate.
NAICS_DESC	Text	Description of the NAICS Classification of the Studies Center.
LEVEL_	Int	Group codes that indicate, 1: Colleges, Universities, 2: Kindergarten, Children schools, 3: Specialties such as Computing, Cosmetology and others.
TOT_ENROLL	Int	Number of people enrolled.
TOT_EMP	Int	Number of employees of the institution.

3. Methodology

The data science methodology that we will apply is based on a process that starts from the understanding of the Business Problem and ends in the Evaluation of Results by applying the K-means Cluster model that will be used to obtain prediction results based on clusters and labeling that will support the investment decisions in services businesses developed around Education

Institutions.
To provide information on the data that allows reducing the risks of investors in neighborhoods close to study institutions such as; Universities, colleges, nursery schools and other institutions in the sector. After the Understanding of the Business Problem we will pass to the stage of acquiring data on educational institutions, this data set contains observations from the entire country of the United States. Then we apply to select the observations of the city of New York, this is called Subseting and we will reduce variables (columns) that are redundant for the application of the grouping model (Clustering); such as the geographic coordinates X, Y with latitude and longitude, which are also geographical coordinates; The NAICS code with the NAICS description represents the same content, this is summarized in the descriptions of each variable.

We will continue with the realization of an exploratory analysis of this pre-processed data set that contains the relevant variables and that shows group characterizations (classification) determined by the institution that manages and organizes the data of these study centers, some examples are classification of levels (variable LEVEL_), this classification of an educational institution can be 1, 2, 3. Another representation of the exploratory data analysis will be applied with the variable NAICS_DESCshort (Brief description of the NAICS classification), this variable indicates the classification of an institution education in one of these classes: COLLE, COMPU, COSME, EDUCA, FINE, JUNE, OTHERS, which will be added (sum) the population of education institutions by this classification and is represented graphically, this classification will allow us to understand depending on the aggregation the density of colleges and universities in COLLE; Infant schools in JUNE, and the other categories of NAICS.

A relevant milestone prior to the application of the Model, is to attach to the "Homeland Infrastructure Foundation-Level Data (HIFLD)" dataset the FOURSQUARE dataset, which aggregates the businesses that are developed as part of the neighborhood of educational institutions; To achieve this, we will pass a REST invocation with the FOURSQUARE API with the latitude and longitude of each educational institution in NEW YORK to the quadrangular invocation function "explore". FOURSQUARE data will be the 10 companies with the highest frequency of visits (check) and that are neighbors of an educational institution. This way of adding the information to the data of the filtered and reduced NAICS data set will be the input to apply the K-means grouping model for this we chose as initial 4 clusters (obtained by the analysis of "Elbow Curve") , to avoid the disaggregation of the groups associated to the businesses.

Finally, we will go to the Evaluation and Results stages where we will select and describe some of the groups obtained with the Kmeans model applied as Predictor of cluster labeling. In the presentation of the conclusions these results of the project will be summarized from the perspective of compliance with the questions posed as a business problem and its continuity as a problem in continuous improvement.

As mentioned earlier, we will use the * k-means * model, which is widely used to group in many data science applications, especially useful if you need to quickly discover information from unlabeled data, such as the case presented, the project documents that detail each phase is organized in the following structure with comments in the case of the code in the Notebook Jupiter cells.

3.1. Data Acquisition Source & Data Wrangling

For the development of the case, the phases of data acquisition and Data Wrangling (Cleaning, Basic statistical analysis and transformation) were carried out in two stages. In stage 1, only HIFLD data is acquired, data with variables from educational institutions in New York City that indicate: the name of the institution (NAME), address (ADDRESS), geospatial coordinates (LATITUDE, LENGTH), classifications typical of the education sector such as LEVEL, TYPE, NAICS, Population (POPULATION), Total student enrolled per year, Number of workers, size of residence facilities or dormitories.

In the following table (table 1) we show a set of data from the dataframe resulting from the Data Wranling process necessary for exploratory analysis and for modeling the kmeans cluster.

Table 1: DataFrame nydata after Data Wrangling (nydata.head()), Partial view

	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	

The definitions of the dataframe columns are as follows (table 2), his view of the variables indicates the types of data of the variables and that there is no null data in them, the number of observations and variables were determined in 86 observations (rows) and 15 variables (columns) for the defined "nydata" dataframe in the jupiter notebook "02CapstoneProjectBattleNeighborhoods.ipynb" of this project.

Table 2: nydata.Info() nydata DataFrame

```
#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
```

3.3. Exploratory Analysis

We will apply Data Analysis and Visual Analysis, to achieve the best understanding of the data, in this process we will define data sets adapted to understand the variables selected in the analysis in this project context.

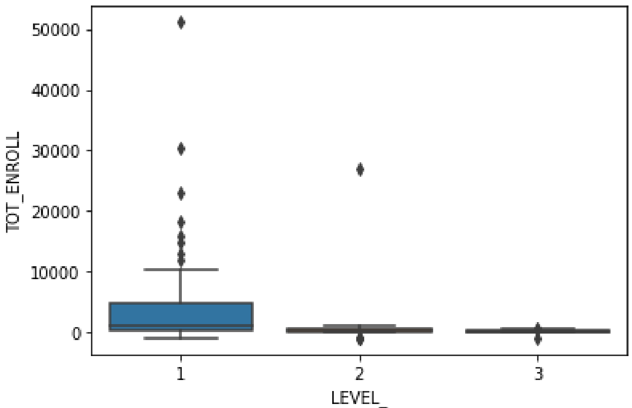
Descriptive statistics of the New York data set, this view shows the statistical values of the data set variables that are numerical, it should be noted that the variables TYPE LEVEL_, INST_SIZE, represent classification categories of educational institutions with respect to the type At the level of the institutions, INST_SIZE is a classification of the type of size of the facilities. On the other hand, the variables POPULATION, DORM_CAP, TOT_ENROLL, TOT_EMP, are numerical values that indicate the number of people associated with the variable, as an example TOT_EMP is the total employees of the institution. Table 4 shows the descriptive statistics view of these variables.

Table 4: Descriptive statistics of the NY dataset

	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

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

- LEVEL_ vs TOT_ENROLL:
  Fig. 1: Boxplot LEVEL_ vs TOT_ENROLL



Visual representation with Boxplot of LEVEL_ vs TOT_ENROLL, we observe that:

  - Level 1 has the largest number of students enrolled in institutions such as Colleges and Universities, recruitments are dispersed and contain some data that are outside the average range of institutions in this category or boxplot analysis.
  - Level 2 corresponds to schools for children and pre-K is less dispersed in terms of total enrollment per year than level 1, so we observe a value outside the boxplot range.
  - Level 3 corresponds to educational institutions of a technical type or other groups such as cosmetology, computing, the total number of enrollments of these institutions are very similar, so that their dispersion is less than the previous level categories.
- NAICS vs TOT_ENROLL:

Table 5: Enrolled by NAICS Class

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

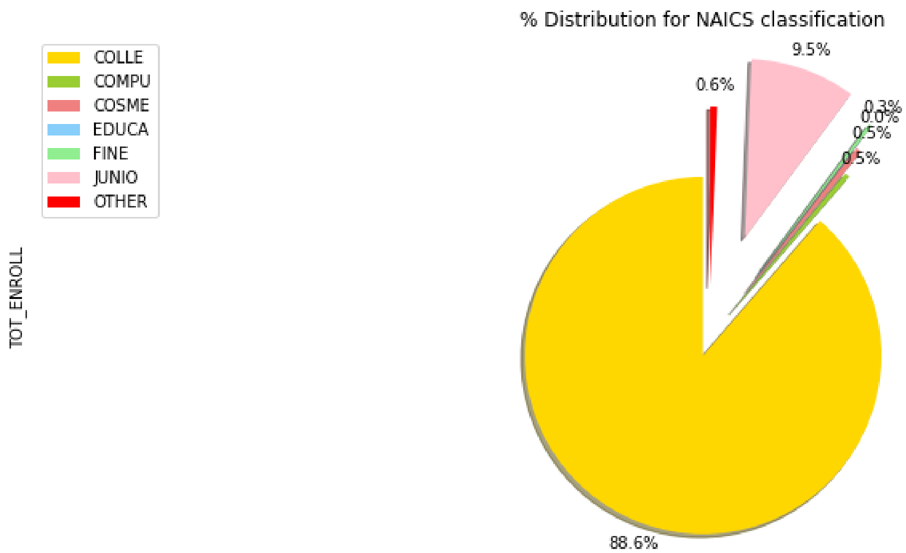
Tabular representation of values of the variables NAICS_DESCshort vs TOT_ENROLL, where the NAICS classification groups educational institutions by regular training that is COLLE

(Colleges, Universities), JUNE (children's schools) and those of functional specialty such as COMPU, FINE, COSME (Computing, Gym, Cosmotology) and others. We observed that:

- The COLL category has the largest number of enrolls per year with a total of 251242.
- The JUNE category has the second position of enrolling with 27017.
- The EDUC category has 0 enrollments.

The graphical representation of these variables (NAICS vs TOT_ENROLL) are shown below in a pie chart.

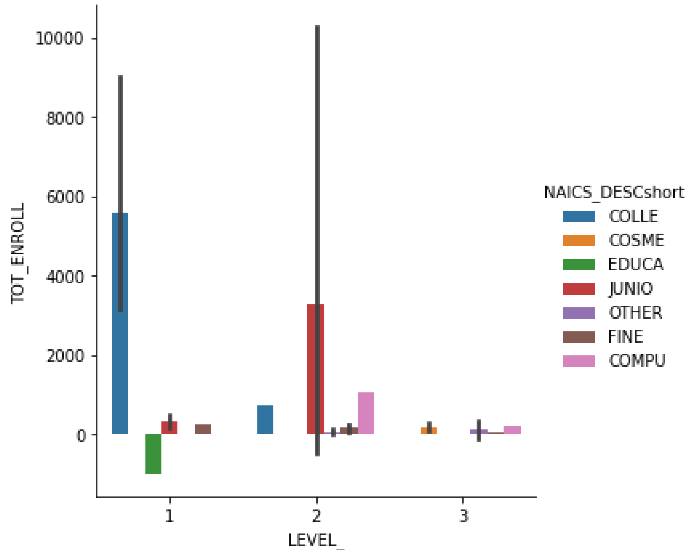
Fig. 2: Distribution for NAICS



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

- LEVEL_ , NAICS vs TOT_ENROLL

Fig. 3: BoxPlot LEVEL_ , NAICS vs TOT_ENROLL



Visual representation with Boxplot of LEVEL_ & NAICS vs TOT_ENROLL, where we observe that:

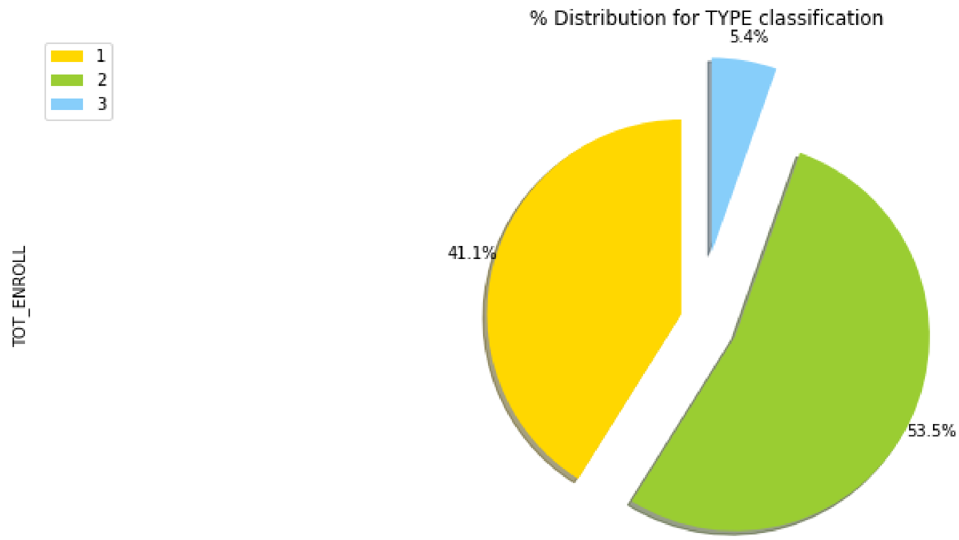
- The classification of levels includes several categories of NAICS.
- There is still a tendency for level 1 with the NACIS categories to be the largest number of enrolls.
- In level 1 the highest concentration is given in the COLICS category of NAICS.
- A special interpretation that the EDUCA category, which corresponds to pedagogical institutions is considered part of level 1 and in this case the negative enrollment value is that no recent information is available

Another classification variable is TYPE, and the relationship with enrollment is indicated in the following table and with a percentage distribution of the categories in the pie chart.

Table 6: Enrollment by TYPE class

TYPE	TOT_ENROLL
1	116061
2	151152
3	15294

Fig. 4: Pie Chart of Distribution for TYPE class



The variables POBLATION, TOT_EMP, are similar distributions, with the uniqueness associated with the physical size of the facilities.

The HOSING variable is transformed to 1 if it has student residence facilities or O if it does not. In the next stage we will add variables associated with the businesses in the neighborhood to which each observation of the educational institutions of the city of New York belongs.

Visual representation of educational institutions in the top 10 by population of the nydata dataframe

Finally, in this exploratory analysis stage, we use the Follium API to geographically visualize the geographic distribution of the 10 institutions with the largest population of the nydata dataframe. This map shows and the dataframe indicates the data of these educational institutions.

Fig. 5: Map of Top 10 Educational Institutions by POPULATION

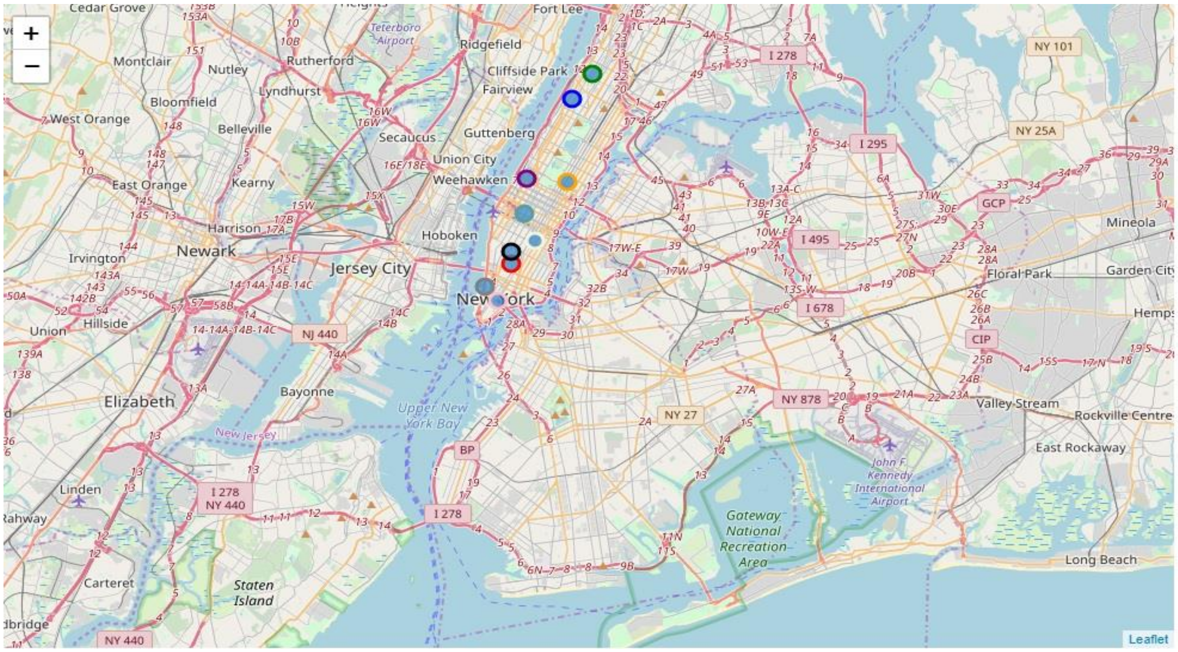


Table 7: List of Educational Institutions in the Top 10 by Population

	ADDRESS	NAME	STATUS	LATITUDE	LONGITUDE	NAICS_DESC	NAICS_CODE	TYPE	POPULATION	LEVEL_	INST_SIZE	DORM_CAP	TO
40	70 WASHINGTON SQ SOUTH	NEW YORK UNIVERSITY	A	40.729452	-73.997264	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	2	73997	1	5	13075	
51	WEST 116 ST AND BROADWAY	COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK	A	40.808286	-73.961885	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	2	50144	1	5	12953	
71	199 CHAMBERS ST	CUNY BOROUGH OF MANHATTAN COMMUNITY COLLEGE	A	40.718790	-74.011826	JUNIOR COLLEGES	611210	1	30069	2	5	-999	
36	695 PARK AVE	CUNY HUNTER COLLEGE	A	40.768669	-73.964795	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	1	27003	1	5	650	
21	ONE BERNARD BARUCH WAY (55 LEXINGTON AVE AT 24...	CUNY BERNARD M BARUCH COLLEGE	A	40.740238	-73.983417	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	1	20836	1	4	414	
72	160 CONVENT AVE	CUNY CITY COLLEGE	A	40.819794	-73.950550	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	1	18746	1	4	590	
46	524 W 59TH ST	CUNY JOHN JAY COLLEGE OF CRIMINAL JUSTICE	A	40.770346	-73.988403	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	1	17160	1	4	176	
25	1 PACE PLAZA	PACE UNIVERSITY-NEW YORK	A	40.711710	-74.004874	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	2	16256	1	4	3726	
10	500 7TH AVENUE	TOURO COLLEGE	A	40.753362	-73.989488	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	2	14505	1	4	388	
9	66 WEST 12TH STREET	THE NEW SCHOOL	A	40.735498	-73.997158	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS	611310	2	13736	1	4	1960	

4. Model & Evaluation

Before applying the K-means model of the python "sklearn.cluster" library, we will add the business expiration columns to the model's Data Framework with the characteristics associated to the location of each ADDRESS using latitude and longitude coordinates of the dataset "nydata"; for this we will use the API of "FOURSQUARE", passing as input every observation of "nydata".

Table 8: Top 10 of businesses neighboring (dataframe: neighborhoods_venues_sorted)

	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

The data of the resulting dataframe contains in one of the columns the address of each educational Institution and the other columns have the top 10 businesses neighboring the address that is the pivot from which they were obtained from "FOURSQUARE" according to the following scope definitions (300 meter radius), a sample of the dataframe obtained is shown in table 8 of this document. We call this dataframe "neighborhoods_venues_sorted" which we will link to the "nydata" dataframe by the ADDRESS column to obtain all the elements that characterize and allow us to perform the labeling to solve what is proposed in "Business Problem". To obtain the resulting dataframe we use the following python code line:

ny_merged = nydatam.join (neighborhoods_venues_sorted.set_index ('ADDRESS'), on = 'ADDRESS')

This dataframe will allow to have information associated with a labeling of new data or to consult about the best 10 businesses that are located around a prediction by providing the values that are required in the Kmeans Model Prediction. An example is Table 9, with information from an observation of the consolidated dataframe called “ny_merged”.

Table 9: Sample of data ny_merget DataFrame

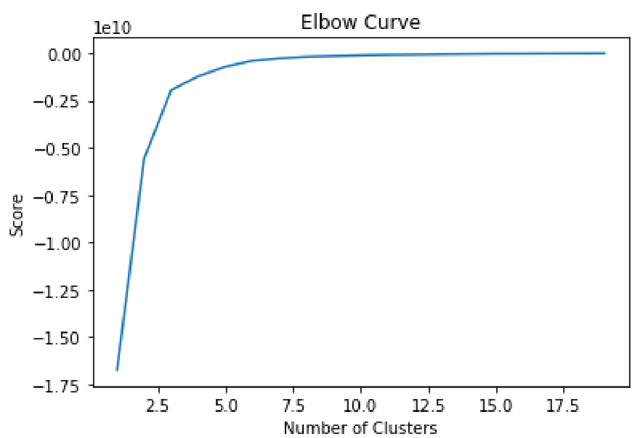
ADDRESS	1300 YORK AVE, C-114
NAME	WEILL CORNELL MEDICAL COLLEGE
STATUS	A
LATITUDE	40.7649
LONGITUDE	-73.9547
NAICS_DESC	COLLEGES, UNIVERSITIES, AND PROFESSIONAL SCHOOLS
NAICS_CODE	611310
TYPE	2
POPULATION	7976
LEVEL_	1
INST_SIZE	2
HOUSE_IN	1
TOT_ENROLL	1107
TOT_EMP	6869
NAICS_DESCshort	COLLE
1st Most 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

This observation already relates to an address of an educational institution the 10 best businesses that are located in the vicinity of the institution named "WEILL CORNELL MEDICAL COLLEGE", this will be explained in the part of results when assigning a label as a prediction of the model.

4.1. Define the number of clusters of the Model (K) and Evaluation

We will use the "Elbow Curve" technique and this indicates that we can choose a K between 2 to 4, for a particular case raised in this project we will use 4 to have more clusters to label the predictions as a function of the variables considered as characteristics, Although the score is higher in K = 2, as shown in Figure 5 in the curve of the "Elbow Curve" technique.

Fig. 5: Elbow Curve



4.2. K-means model

The configuration of the K-means model of the library "" is the one shown in Figure 6, it indicates parameters such as the number of clusters, the algorithm "k-means ++", maximum number of iterations 300 as shown.

Fig. 6; 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)
```

4.3. Adding the cluster tag to each observation

With the configured model, assigned the independent variables or determining characteristics of the vector X (['TYPE','POPULATION','LEVEL_','INST_SIZE','HOUSE_IN','TOT_ENROLL','TOT_EMP']), and executing the python statement "kmeans = KMeans (n_clusters = K) .fit (X)", we can obtain the labels of the clusters of each observation (row) with the python expression "kmeans.labels_". In table 10, we set some rows of the "ny_merged" dataframe with the label of the assigned cluster number.

Table 10: DataFrame with assigned Cluster labeling

	ADDRESS	Cluster Labels	NAICS_CODE	TYPE	POPULATION	LEVEL_	TOT_ENROLL	1st Most Common Venue
0	1 PACE PLAZA	2	611310	2	16256	1	12986	Sandwich Place
1	11 PARK PLACE 4TH FLOOR	0	611210	3	-999	2	-999	Hotel
2	110 WILLIAM ST. 19TH FL	0	611310	3	732	1	578	Coffee Shop
3	111 FRANKLIN ST	0	611310	2	224	1	119	Gym / Fitness Center
4	115 WEST 27TH STREET, 11TH FLOOR	0	611519	3	83	3	70	Hotel

5. Results

5.1. Number of Tags per Cluster

Table 11 summarizes the number of observations for each cluster.

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

5.2. Indexes of observations by Cluster of DataFrame ny_merged

The following list shows the indices of each observation associated with the cluster that belongs.

```
{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')}
```

5.5. Metrics of the numerical variables of each cluster

This result defines the statistical values of each variable for the determination of the characteristic as a membership metric, we must consider that it is the combination of the set of variables and the values that delimit for each cluster.

The following tables are the statistical values of each cluster.

Table 11: Statistical metrics of numerical variables of cluster 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

Table 12: Statistical metrics of numerical variables of cluster 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

Table 13: Statistical metrics of numerical variables of cluster 2

	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

Table 14: Statistical metrics of numerical variables of cluster 3

	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

Tables 11 through 14 simplify the characterization of the determining variables that were defined in the k-means model. Tables 11 through 14 simplify the characterization of the determining variables that were defined in the k-means model.

```
cluster to which it belongs: [3]
with features-> TYPE: 2, POPULATION: 7976, LEVEL_: 1, INST_SIZE: 2, HO
USE_IN: 1, TOT_ENROLL: 1107, TOT_EMP: 6869
```

5.6. Applying Labeling Prediction

We make 4 predictions with the data from the ny_merged dataset, we will choose an observation for each Cluster, so we ensure that the labeling and the chosen characteristics are correct if it is the same with the result of the Model prediction, to execute this we will use the python statement:

We make 4 predictions with the data from the ny_merged dataset, we will choose an observation for each Cluster, so we ensure that the labeling and the chosen characteristics are correct if it is the same with the result of the Model prediction, To execute this we will use the following lines of python code and the output of each prediction is observed at the end of the code.:

```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
```

```
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
```

The labeling results correspond to the cluster.

5.7. Top n of neighboring business premises to each cluster:

In this result we will show the top 1 of the neighboring businesses of each cluster, if in the case you want the top 3 or top 4 you must add the list in order.

```
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']
```

With this result we can answer the questions raised in "Business problem" (a) and (c). For question (b) we can use the Prediction results that label the cluster to which the new data obtained belongs.

A relevant result is that the density variables populate the educational institutions (POPULATION, TOT_ENROLL, TOT_EMP, HOUSE_IN), are the key variables of the characterization of the groups and not the variables of grouping of educational entities (TYPE, LEVE_, NAICS

6. 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).
- 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 K-means algorithm allows us to create clusters when we have unlabeled data groups.
- 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.