Capstone Project - The Battle of Neighborhoods

I) Introduction

São Paulo is a city in the Southern Region of Brazil. The city is the capital of the state (São Paulo analogue with the city), the wealthiest city in the country and of South Hemisphere, one of the 20th wealthiest cities in the world, representing alone 10.7% of all Brazil GDP and the 13th most populous. Home of the São Paulo Stock Exchange. It exerts strong influence in commerce, finance, arts and entertainment in the country. Brazil is a country of many contrasts and São Paulo is not different. With a Human Development Index like a European country in some regions and low as a score of 0.700 in other regions.
Problems with high demographical density and with High Crime Rates. Given its diversity, the analysis will give good insight to cluster city´s districts with the same pattern and, in some level, help to understand where a new business should be placed.
The analysis will help to find the correct location for a business in the city, decreasing potential risks. Showing the profile of each cluster to open branches in districts with the same behavior.

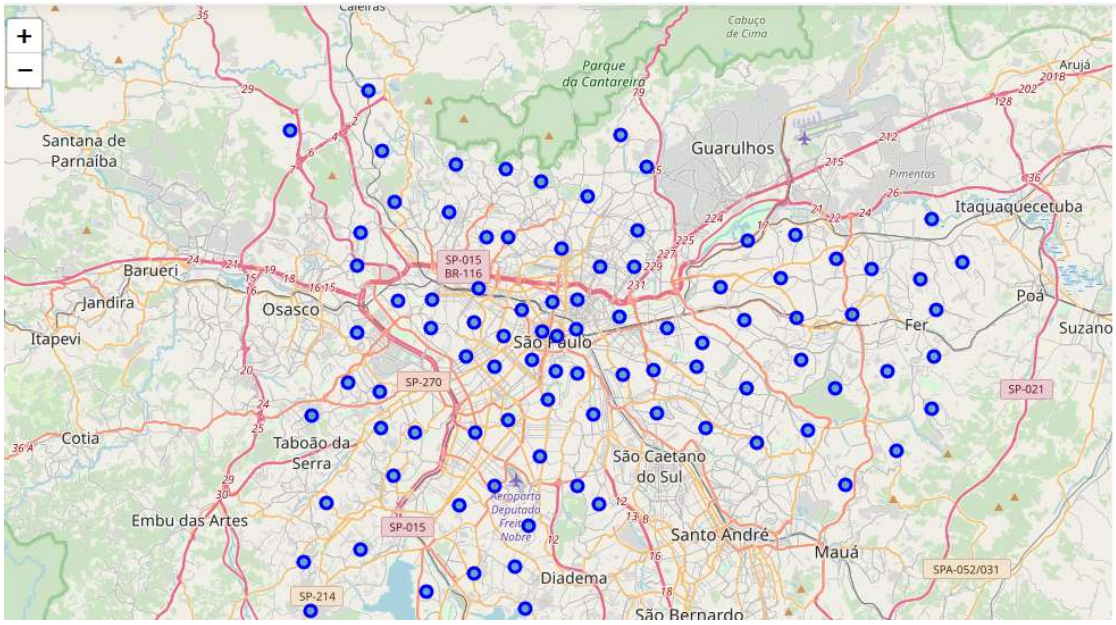
II) Data

The necessary information needed in the analysis will come from different sources like Wikipedia, city hall homepage, foursquare and secretary of public security of the state of São Paulo. The description of each data and the cleansing needed for it will be explained bellow. And all the granularity of the information is by City´s Districts. The District is an administrative division that aggregates neighborhoods. The aggregation of Districts becomes a sub-prefecture. And finally the sub-prefecture becomes a city.

- **Geographic Boundaries:** Come from the City Hall Homepage, with each district in the city and each respective geographic boundary. Using the coordinates, I calculated the mean to discover the center. And the file is in a shape format and it will be converted a json format. Below follows an example.

Borough	Latitude	Longitude
MANDAQUI	-23.462176	-46.640467
MARSILAC	-23.949529	-46.724111
MOEMA	-23.594287	-46.659987
PARQUE DO CARMO	-23.576289	-46.462777
PERDIZES	-23.540115	-46.680767

With the conversion from shp to json done, the next step is to view the map with the coordinates read from the json file. The centroids can be seen in the map below.



- **Population and Area:** Come from the City Hall Homepage, with the population of each district as consolidated in the last demographic census of 2010. And derived from these two pieces of information, the demographic density.
- **Aging Index:** Come from the City Hall Homepage, with the aging index for each district. The aging index is the population equal or over 60 years old divided by the population under 14 years old multiplied by 100.
- **HDI:** Come from the Wikipedia. The data is from 2000 but is the most recent and trustworthy data that can be found in the internet.
- **Crime Rate:** Come from the secretary of public security of the state of São Paulo. The Rate is the aggregation of all types of crimes divided by the population and multiplied by 1000. The data is from 2019. Not all the districts have information about crimes and to complete the required information, I used the mean given by the districts that the missing line is inserted.

The four bases above was collected manually from each respective site and merged in excel resulting in the file data.csv. Below follows an example.

subprefecture	districts	area_km2	population_2010	demographical_density	aging_index	HDI	crimes	crimes_per_1000pop
Mooca	AGUA RASA	6,9	84963	12313,48	146,11	0,886	2883	33,932
Pinheiros	ALTO DE PINHEIROS	7,7	43117	5599,61	229,4	0,955	16118	373,82
Perus	ANHANGUERA	33,3	65859	1977,75	38,94	0,774	7077	107,457
Aricanduva	ARICANDUVA	6,6	89622	13579,09	108,6	0,83	4523	50,468
Penha	ARTUR ALVIM	6,6	105269	15949,85	100,11	0,833	8106	77,003

- **Foursquare API:** The API will be used to generate the most Commons venues of each district. Due to the great diversity of venues given by the API, it will be necessary to cluster types of business such as restaurants in only one type.

As an initial setup, I need to set the Client_ID, Client_Secret and the Version to access the API. Next, create a function to connect with the API and a function to read the json file. Then I used the function with the required parameters.

```
saopaulo_venues = getNearbyVenues(names=sp_districts['Borough'],
                                  latitudes=sp_districts['Latitude'],
                                  longitudes=sp_districts['Longitude']
                                 )

saopaulo_venues.shape
```

```
(2333, 7)
```

As it can be seen above, the dataframe has 2332 rows and 7 columns.

III) Methodology

In this analysis, the goal is to find the districts cluster that have similar parameters and could be potential places to install a new business. To do the analysis I did the steps as follow:

- In the first step, I prepared a file to cluster the same kind of business in the Foursquare but come with different names. Merged the file with the base of Foursquare venues and dropped the unnecessary columns.
- In the second step, transposed the base and aggregated the values by districts.
- In the third step, due to the different ranges of each variable and this fact could influence the analysis to the variable with the largest range, I normalized of them. And replace any NAN value with a 0.
- In the fourth step, I explored the data, using the describe and correlation functions to identify possible relations between the variables.
- And in the final step, I used the unsupervised learning algorithm, K-Means, because, in this analysis, the base didn´t have a target variable (Y). And as a result, the algorithm generated the clusters with districts with similar pattern. To finally understand and describe with one of them.

IV) Analysis

First, open the csv file with the new names for the Venues Categories to aggregate then. And result in the "Venue Category" as shown below.

```
saopaulo_venues['Venue Category'].unique()
```

```
array(['Gym', 'Market', 'Regional Restaurant', 'International Restaurant',
       'Coffee Shop', 'Park', 'Plaza', 'Stadium', 'Store',
       'fast food restaurant', 'Pharmacy', 'Service', 'Hospital',
       'Food Shop', 'Bar', 'Sports Field', 'General Entertainment',
       'Bakery', 'Theater', 'Shopping Plaza', 'Museum', 'Cultural Center',
       'Bank', 'Bus Station', 'Hotel', 'Club', 'Event Space', 'School',
       'Movie Theater'], dtype=object)
```

As a next step, to transform and transpose the base. With the result set group by "districts". Then merge the dataframes and cast the variables from string to float.
As the data have different ranges, it is necessary to normalize them so that the values with greater variation do not influence the result of the algorithm that will be applied. And replace any "NAN" value with 0.

Then drop any unnecessary column of the dataframe. With the dataframe ready, the first analysis is to get the mean, standard deviation, min, max and the quartiles with the describe function.

```
df2.describe()
```

	demographical_density	aging_index	HDI	crimes_per_1000pop	Bakery	Bank	Bar	Bus Station	Club	Coffee Shop	...
count	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000	...
mean	0.412344	0.234549	0.580208	0.100229	0.074632	0.000795	0.054698	0.004804	0.005764	0.026679	...
std	0.196399	0.154391	0.249326	0.162388	0.088116	0.005415	0.124247	0.021470	0.023293	0.053824	...
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	0.287501	0.121013	0.383654	0.025750	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
50%	0.408798	0.205258	0.586538	0.047621	0.053713	0.000000	0.000000	0.000000	0.000000	0.000000	...
75%	0.529667	0.322840	0.771154	0.085586	0.102778	0.000000	0.063322	0.000000	0.000000	0.036045	...
max	1.000000	1.000000	1.000000	1.000000	0.500000	0.050000	1.000000	0.142857	0.173913	0.333333	...

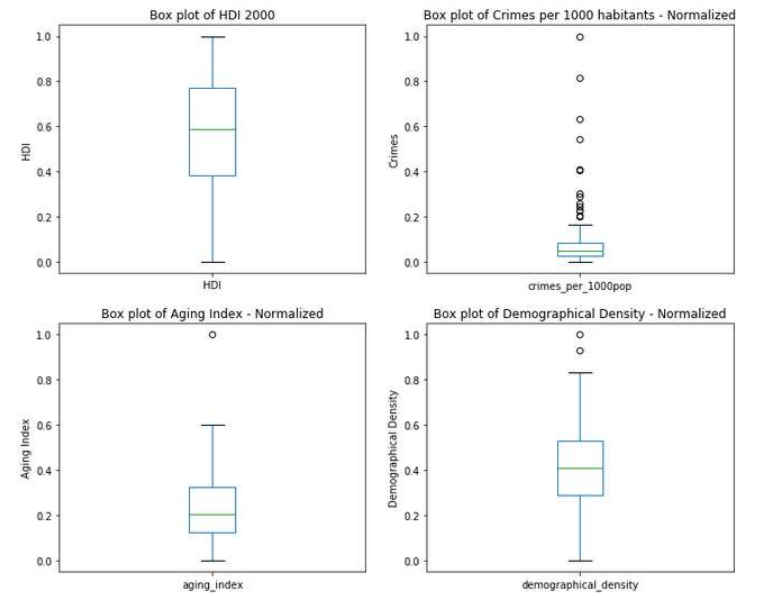
8 rows × 33 columns

Nothing of note here, but it´s good to keep the result in mind for the comparative with the resulting clusters. Next calculate the correlation matrix for the dataframe.

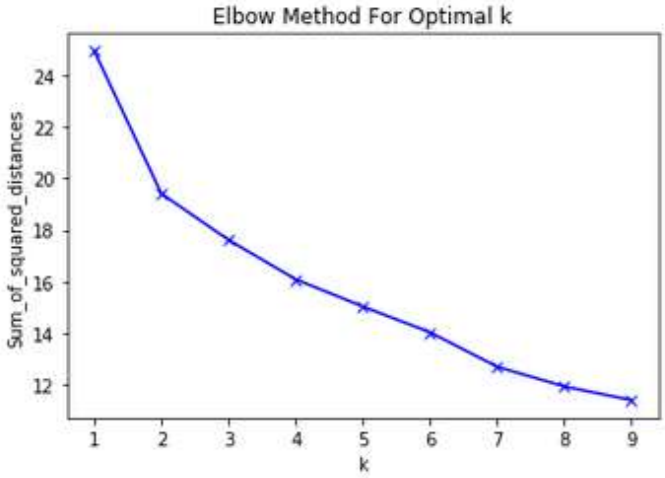
```
df2.corr()
```

	demographical_density	aging_index	HDI	crimes_per_1000pop	Bakery	Bank	Bar	Bus Station	Club	Coffee Shop	.
demographical_density	1.000000	0.061098	0.033454	-0.155616	0.051841	0.129444	-0.050027	-0.061759	0.002751	0.033506	.
aging_index	0.061098	1.000000	0.607240	0.053983	-0.128034	-0.010798	0.131386	-0.067683	0.028469	0.374106	.
HDI	0.033454	0.607240	1.000000	0.197694	-0.068316	0.007357	0.201226	-0.099096	0.131516	0.190212	.
crimes_per_1000pop	-0.155616	0.053983	0.197694	1.000000	-0.137934	0.116114	-0.005116	-0.029649	0.226024	0.097751	.
Bakery	0.051841	-0.128034	-0.068316	-0.137934	1.000000	-0.043762	-0.194290	0.160922	-0.060415	-0.274682	.
Bank	0.129444	-0.010798	0.007357	0.116114	-0.043762	1.000000	-0.004526	-0.033192	-0.036713	-0.029765	.
Bar	-0.050027	0.131386	0.201226	-0.005116	-0.194290	-0.004526	1.000000	-0.048200	-0.012309	-0.041995	.
Bus Station	-0.061759	-0.067683	-0.099096	-0.029649	0.160922	-0.033192	-0.048200	1.000000	-0.055952	-0.100451	.
Club	0.002751	0.028469	0.131516	0.226024	-0.060415	-0.036713	-0.012309	-0.055952	1.000000	0.027549	.
Coffee Shop	0.033506	0.374106	0.190212	0.097751	-0.274682	-0.029765	-0.041995	-0.100451	0.027549	1.000000	.
Cultural Center	0.299618	-0.017541	0.108255	0.447486	-0.065632	0.129197	0.046702	-0.038517	0.083301	0.260746	.
Event Space	0.066013	0.037183	0.021408	-0.076105	-0.012001	-0.026008	-0.006039	-0.039637	-0.043841	-0.056771	.
Food Shop	0.230180	-0.114519	-0.131685	-0.195402	0.137446	-0.094867	-0.220696	-0.038949	-0.132445	-0.122262	.
General Entertainment	-0.187048	-0.050640	-0.061706	-0.015432	-0.139437	-0.031013	-0.092143	0.136713	-0.018117	-0.053918	.
Gym	-0.107264	-0.118408	0.008179	-0.147411	-0.077853	0.021478	-0.159932	0.018676	-0.067740	0.053974	.
Hospital	-0.034986	0.159924	0.173651	-0.014656	-0.087354	-0.015141	-0.021688	-0.023076	-0.025523	-0.051122	.
Hotel	0.067156	-0.015260	0.043005	0.091798	-0.140687	-0.036371	-0.013202	-0.055431	0.313024	-0.033696	.
International Restaurant	0.009179	0.106200	0.408863	0.197415	-0.228362	-0.064934	-0.045590	-0.187432	0.099969	0.202059	.
Market	0.186099	0.075677	-0.099475	-0.165219	0.018354	-0.030634	-0.109235	0.127470	-0.117563	0.081138	.
Movie Theater	0.016318	-0.083566	-0.117720	-0.050951	-0.091277	-0.017152	-0.044393	-0.026140	-0.018571	-0.041774	.
Museum	0.086508	0.005834	0.133275	0.275970	-0.053230	0.059777	0.063045	-0.065863	0.014208	0.110650	.
Park	-0.173952	-0.080662	-0.032176	-0.039011	0.325631	-0.044387	-0.095304	-0.013041	-0.019951	-0.125603	.
Pharmacy	0.104976	-0.003124	-0.000657	-0.110652	0.094394	0.136917	-0.121402	0.024783	-0.072399	-0.089860	.
Plaza	-0.156354	0.226125	-0.009628	0.015623	-0.170338	-0.053583	-0.083734	-0.042519	-0.027453	0.125252	.
Regional Restaurant	0.121620	0.089147	0.266032	0.212177	0.054788	0.091979	-0.060521	-0.117200	-0.012685	0.107792	.
School	-0.010054	0.047323	0.024097	-0.045061	-0.042336	-0.015141	0.050376	-0.023076	-0.025523	0.022577	.
Service	0.020727	0.016849	0.145210	-0.029553	0.013956	-0.032989	-0.037159	-0.003523	0.118872	-0.112104	.
Shopping Plaza	0.117642	-0.069514	-0.014652	0.174275	-0.150182	-0.030995	-0.055877	-0.047237	-0.022763	0.042420	.
Sports Field	-0.106862	0.036833	0.084575	-0.061747	-0.009455	-0.059652	0.021664	-0.014024	0.169149	-0.100693	.
Stadium	-0.037890	-0.066086	-0.039874	-0.070430	-0.076045	-0.021610	-0.059136	-0.046075	-0.005370	0.044933	.
Store	0.159663	-0.099370	0.072050	0.136810	0.031706	0.230399	-0.128536	-0.029605	0.005299	-0.115020	.
Theater	0.412431	0.200991	0.323890	0.276984	-0.083097	0.020416	0.065554	-0.035901	0.058822	0.118552	.
fast food restaurant	0.149796	-0.091455	-0.054220	-0.161867	0.006976	0.023700	0.102292	0.116063	0.038690	-0.126369	.

All the correlations are very week to average. For example, the "Theater" has a average correlation with the demographical density. That result is expected. Another expected result is the average correlation between "HDI" and "aging index", because in the calculation of HDI is used the life expectancy. Two interesting results is the correlation between HDI and Theater and HDI and International Restaurant. In other words, the higher the HDI, the higher the quantity of the Theater and International Restaurants. The next step of the analysis is to verify the boxplot of the variables.



An interesting result of the boxplot is to see the quartiles and candidates to be outliers. The outliers are all the points in the plot that are below the inferior limit and above the superior limit. In some cases, these points can be discarded, but because the dataframe have so few lines for this analysis I'll keep all the rows.
To the next step, I need to evaluate the optimum value for the number of clusters with the elbow method. The graphic can be seen below.

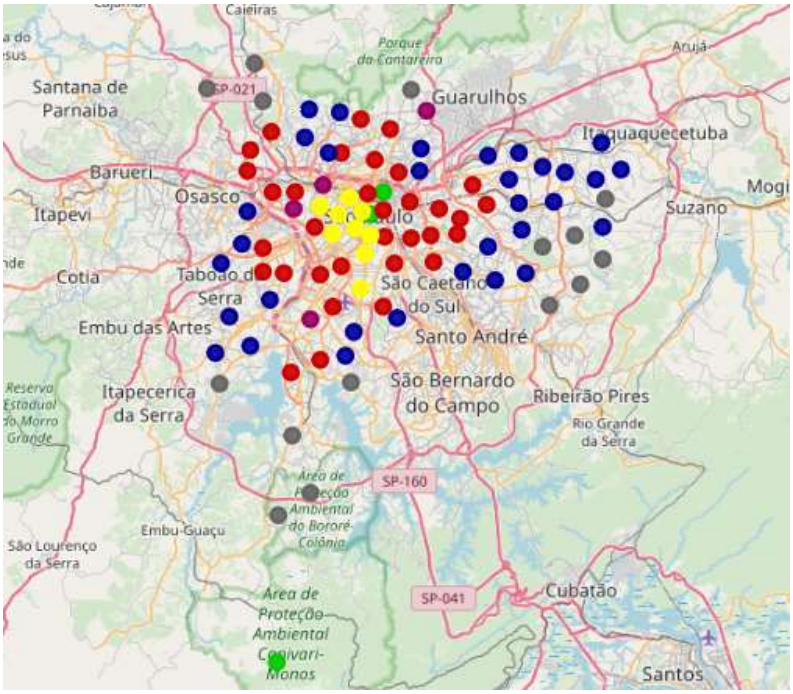


We have an elbow around the k value equal 2, but to the analysis the value is too low. Then, we tried different values of k and analyze and the value that did the best clustering was 6. Next, I created a function to get top nth most common venues and using the

function above, the num_top_venues was set to 10 and a new dataframe was created it the result with the labels.

	Cluster Labels	districts	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	AGUA RASA	Bar	Food Shop	Market	Plaza	International Restaurant	fast food restaurant	Store	Gym	Regional Restaurant	Pharmacy
1	5	ALTO DE PINHEIROS	Plaza	fast food restaurant	Regional Restaurant	Food Shop	General Entertainment	Gym	Market	Bus Station	Coffee Shop	International Restaurant
2	0	ANHANGUERA	Food Shop	General Entertainment	Store	Bus Station	Service	Park	Gym	fast food restaurant	Hotel	Bar
3	4	ARICANDUVA	Store	Food Shop	fast food restaurant	Stadium	Regional Restaurant	Pharmacy	Gym	Hotel	Bar	Bus Station
4	4	ARTUR ALVIM	Store	Regional Restaurant	fast food restaurant	Gym	Pharmacy	Bar	Food Shop	Market	Bank	Plaza

Then merged the resulting dataframe with the dataframe with geographical coordinates and the clusters generated by the k-means algorithm can be seen in the map below.



Then merged the dataframe with the columns with the HDI, crimes, demographical density and aging index. And the districts and their respective values are shown below.

Cluster Labels	demographical_density	aging_index	HDI	crimes_per_1000pop
0	0.245525	0.099269	0.288718	0.039738
1	0.330136	0.296171	0.757343	0.083851
2	0.705601	0.368039	0.920513	0.175710
3	0.214252	0.135558	0.408974	0.816405
4	0.533994	0.163991	0.427644	0.033064
5	0.231680	0.571810	0.795192	0.292543

The clusters formed by the k-means algorithm present some interesting insights that will be explained more in the results and discussion section.

V) Results and Discussion

The result obtained with the k-means algorithm brought some interesting insights. Analyzing the first cluster, whose label is equal to 0, we can see that it is composed of districts with HDI, on average, low with value around 0.289, has a low degree of criminality, a low aging index and a low demographic density. Among the most common places in these districts are "Gym", "Restaurants" and "Bar". From the map we can see that the districts are concentrated on the outskirts of the city.
For the cluster with a label of 1, the HDI has a medium-high value of 0.757. Low crime, with a low population density and an aging index slightly higher than that of the previous cluster. Among the most common locations in these districts are "Restaurants" and "Store". From the map we can see that the districts are the ones around the center, with some points on the periphery.
For the cluster with a label equal to 2, HDMI has a high value of 0.920. Low crime, with a high population density and a relatively high aging index when compared to the average of the other clusters. Among the most common places are "Restaurants" and "Store". From the map we can see that the districts are concentrated in central points of the city in tourist points of the city and some considered to be of a high standard.
For the cluster with a label equal to 3, HDMI has a medium-low value of 0.409. High crime rate, low population density and a low aging index. Among the most common places are "Restaurants" and "Bar". From the map we can see that the districts correspond to the center of the city and one in the extreme south on the periphery.
For the cluster with a label equal to 4, HDMI has a medium-low value of 0.428. Low crime rate, average population density and low aging index. Among the most common places in the neighborhood are the "Store", "Market" and "Restaurant". From the map we can see that the districts are concentrated on the outskirts of the city in the south, in the north and east of the city.
For the cluster with a label equal to 5, HDMI has a high value of 0.795, criminality among the top 2, low demographic density and with a higher aging index among all clusters. Among the most common places in the neighborhood are the "Market", "Plaza" and "Restaurant". From the map we can see that the districts are spread across the city.
It is important to note that it is visible that due to the k-means algorithm trying to fit all the data in a cluster, we have data in some groups that differ from the group average. For example, as we can see in the cluster label equal to 5. Where the district of Jacana is clearly distorting the average of the cluster.
Another interesting point that is visible is that in most clusters, the Venues do not have a very well-defined pattern, becoming partly random as we could anticipate with the correlation matrix. To improve the analysis, in a next step, some information such as transportation (metro and train station) should be placed, because the accessibility of the low influence in a certain way on the profile of the neighborhood. In addition to using more filters in types of locations used in the analysis such as school, hospital, among others. And collect more API-specific information focused on the business the stakeholders want to open.

VI) Conclusion

The purpose of the study was to identify patterns between the districts of the city of São Paulo to identify possible clusters with the potential to open a business. For this, in addition to the location of locations given by the Foursquare API, demographic and social indicators were used. Thus, the study resulted in some interesting results, forming clusters with a certain consistency in the average and distribution of its indicators. It is important to note here that due to the unsupervised K-means method trying to fit all data in one of the clusters, some districts were allocated to groups whose value was very different from the rest. Even so, the information is useful if you want to obtain possible candidates to receive a business affiliate.

As a next step, it would be interesting to add more metrics and filter the information from the Foursquare API for the business area that the stakeholders really want to establish in order to assess competition and possible risks in implementing the unit.