The Battle of Neighborhoods In SAO PAULO-BR

Etienne Vanhaecke

06 June 2020

I) Introduction

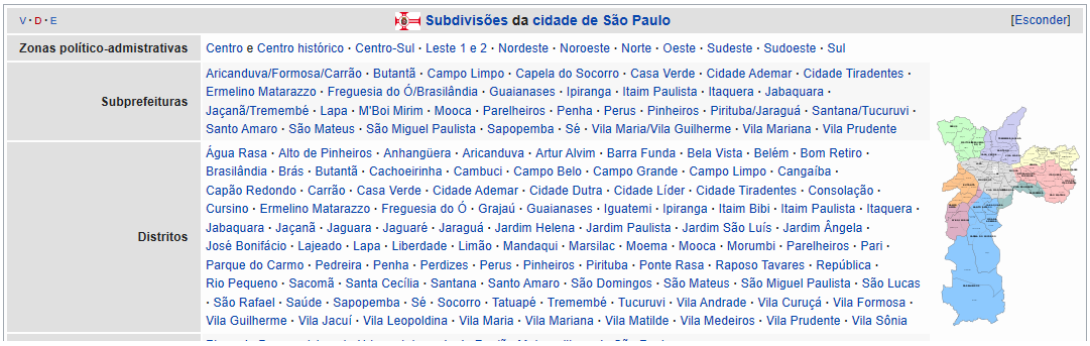
1) Background

The city of Sao Paulo, Sao Paulo state in Brazil, is one of the bigger world city with more of twenty millions of habitants.

The city has multiple territory division, some of which are not compatible. This incompatibility causes doubts about spatial references, for both the population and the state representatives.

For this work, it will have use of three territory divisions:

- First level: The 32 sub prefectures.]
- Second level: The 96 districts (each district belongs to one sub prefecture).]
- Third level: The hundreds of neighborhoods. Unlike the sub prefectures and the districts, whose list is fixe, being administrative structures, there is no single list of neighborhood in Sao Paulo. For this work, we will consider 525 neighborhoods (the richest in information) on more than 1700 listed.



V · D · E	Subdivisões da cidade de São Paulo	[Esconder]
Zonas político-admistrativas	Centro e Centro histórico · Centro-Sul · Leste 1 e 2 · Nordeste · Noroeste · Norte · Oeste · Sudeste · Sudoeste · Sul	
Subprefeituras	Aricanduva/Formosa/Carrão · Butantã · Campo Limpo · Capela do Socorro · Casa Verde · Cidade Ademar · Cidade Tiradentes · Ermelino Matarazzo · Freguesia do Ó/Brasilândia · Guaianases · Ipiranga · Itaim Paulista · Itaquera · Jabaquara · Jaçanã/Tremembé · Lapa · M'Boi Mirim · Mooca · Parelheiros · Penha · Perus · Pinheiros · Pirituba/Jaraguá · Santana/Tucuruvi · Santo Amaro · São Mateus · São Miguel Paulista · Sapopemba · Sé · Vila Maria/Vila Guilherme · Vila Mariana · Vila Prudente	
Distritos	Água Rasa · Alto de Pinheiros · Anhangüera · Aricanduva · Artur Alvim · Barra Funda · Bela Vista · Belém · Bom Retiro · Brasilândia · Brás · Butantã · Cachoeirinha · Cambuci · Campo Belo · Campo Grande · Campo Limpo · Cangaíba · Capão Redondo · Carrão · Casa Verde · Cidade Ademar · Cidade Dutra · Cidade Líder · Cidade Tiradentes · Consolação · Cursino · Ermelino Matarazzo · Freguesia do Ó · Grajaú · Guaianases · Iguatemi · Ipiranga · Itaim Bibi · Itaim Paulista · Itaquera · Jabaquara · Jaçanã · Jaguara · Jaguaré · Jaraguá · Jardim Helena · Jardim Paulista · Jardim São Luís · Jardim Ângela · José Bonifácio · Lajeado · Lapa · Liberdade · Limão · Mandaqui · Marsilac · Moema · Mooca · Morumbi · Parelheiros · Pari · Parque do Carmo · Pedreira · Penha · Perdizes · Perus · Pinheiros · Pirituba · Ponte Rasa · Raposo Tavares · República · Rio Pequeno · Sacomã · Santa Cecília · Santana · Santo Amaro · São Domingos · São Mateus · São Miguel Paulista · São Lucas · São Rafael · Saúde · Sapopemba · Sé · Socorro · Tatuapé · Tremembé · Tucuruvi · Vila Andrade · Vila Curuçá · Vila Formosa · Vila Guilherme · Vila Jacuí · Vila Leopoldina · Vila Maria · Vila Mariana · Vila Matilde · Vila Medeiros · Vila Prudente · Vila Sônia	

The clustering of these neighborhoods, in function of the more represented venues categories and in function of socio-economic indicators, should give interesting insights to help deciding where open a new business.

2) Problem

The data might help to visualize the neighborhoods with similarities about socio-economic indicators and about categories venues.

3) Interest

The executives should have interest to visualize neighborhood distributions of the São Paulo city, economic capital of Brazil and of the South America continent, to help them to define the best locals to open new business.

The general public, interested in geography, should also be interested in this study.

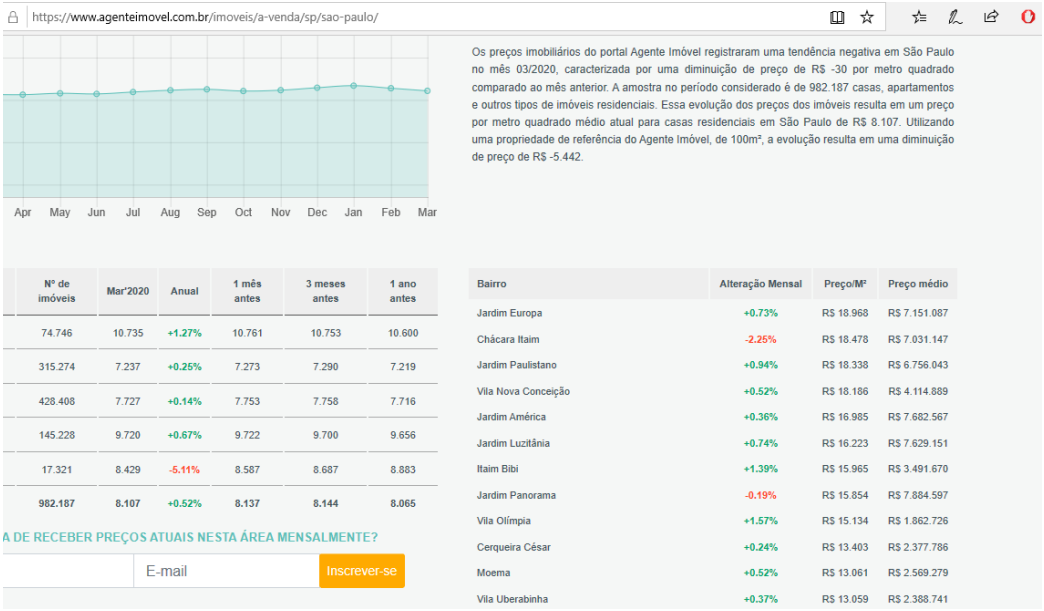
II) Data acquisition and cleaning

1) Data Source

The data have been recovered from different sources:

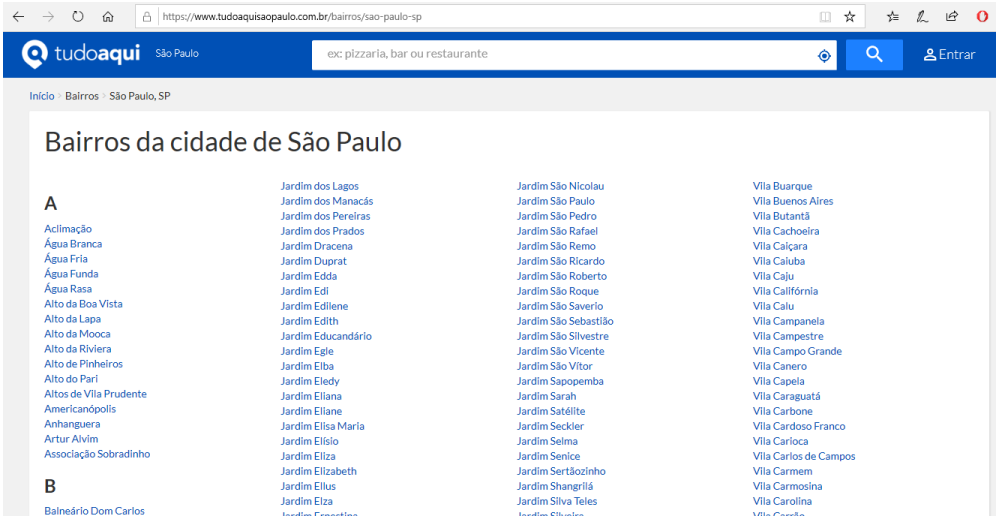
- The list of the neighborhood by the scrap of one table in the Web Site page:
  - **https://www.agenteimovel.com.br/imoveis/a-venda/sp/sao-paulo/**

    At most of the name of the neighborhood, it has been recovered information about the real state at march 2020, like the price at m2, but limited at 525 neighborhood.



  - https://www.tudoaquisaopaulo.com.br/bairros/sao-paulo-sp

    The name of more of 1700 neighborhood referenced in Sao Paulo city.



- The latitude and longitude coordinates of the center of each neighborhood calling the API of the provider ARCGIS with the Python library Geocoder.

```
for ind in dfNbh.iloc[0:].index:
  nbh=dfNbh.loc[ind, "Neighborhood"]
  #br=dfNbh.loc[ind, "Borough"]
  # initialization
  lat_lng_coords = None
  # loop until you get the coordinates
  while(lat_lng_coords is None):
      #g = geocoder.google('{}, Toronto, Ontario'.format(postalCode))
      #g = geocoder.arcgis('{}, {}, Sao Paulo, Sao Paulo'.format(nbh, br))
      g = geocoder.arcgis('{}, Sao Paulo, Sao Paulo'.format(nbh))
      lat_lng_coords = g.latlng
```

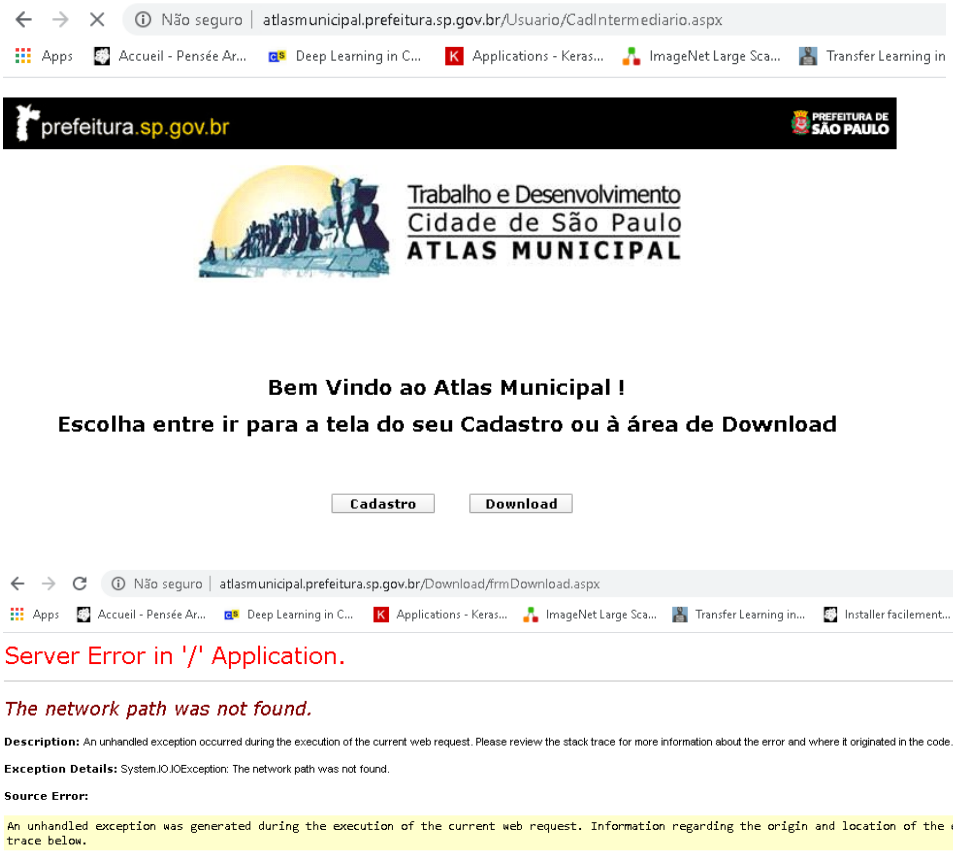
- The limits of the Sao Paulo city sub prefectures (32) and districts (96), downloading shape files from the official Sao Paulo web site http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx

- The IDH (Human Development Index) by district of Sao Paulo City by scrap of the Brazilian Wiki page https://pt.wikipedia.org/wiki/Lista_dos_distritos_de_S%C3%A3o_Paulo_por_%C3%8Dndice_de_Desenvolvimento_Humano

Posição	Distrito	IDH Dados de 2000
1	Moema	0,961
2	Pinheiros	0,960
3	Perdizes	0,957
4	Jardim Paulista	0,957
5	Alto de Pinheiros	0,955

Posição	Distrito	IDH Dados de 2000
16	Campo Belo	0,935
17	Santa Cecília	0,930
18	Butantã	0,928
19	Santana	0,925
20	Campo Grande	0,921

The origin of this data IDH is the *"Atlas do Trabalho de Desenvolvimento do Município de São Paulo 2007"*, available, after creating a login, with the official Web Site: http://atlasmunicipal.prefeitura.sp.gov.br/Login/Login.aspx
This source should be excellent to enrich the data to use for this study, but unfortunately, the download of this archive does not currently work:



- The venues of each neighborhood has been recovered calling a API of the Foursquare site:

```
[ ] import requests # library to handle requests
    from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe
    def getNearbyVenues(names, latitudes, longitudes, radius=1000):
        LIMIT = 200 # limit of number of venues returned by Foursquare API
        venues_list=[]
        for name, lat, lng in zip(names, latitudes, longitudes):
            print(name)

            # create the API request URL
            url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
                CLIENT_ID,
                CLIENT_SECRET,
                VERSION,
                lat,
                lng,
                radius,
                LIMIT)

            # make the GET request
            try:
              jsonResult = requests.get(url).json()
              results = jsonResult["response"]['groups'][0]['items']
```

2) Data Transformation/Cleaning
   - List of the neighborhoods
     Like it has not be possible to find information about the complete list of the neighborhood (more of 1700), it has been used the limited list of 525 neighborhood, scraped with some information about the real state:

```
Index(['Neighborhood', 'RateMonth', 'PriceByM2', 'MedianPrice'], dtype='object')
```

	Neighborhood	RateMonth	PriceByM2	MedianPrice
0	Jardim Europa	+0.73%	R$ 18.968	R$ 7.151.087
1	Chácara Itaim	-2.25%	R$ 18.478	R$ 7.031.147
2	Jardim Paulistano	+0.94%	R$ 18.338	R$ 6.756.043
3	Vila Nova Conceição	+0.52%	R$ 18.186	R$ 4.114.889
4	Jardim América	+0.36%	R$ 16.985	R$ 7.682.567
...	...	...	...	...
520	Parque Savoy City	-1.13%	R$ 3.540	R$ 417.503
521	Parque Casa de Pedra	+0.06%	R$ 3.531	R$ 454.416
522	Conjunto Residencial José Bonifácio	+0.18%	R$ 3.287	R$ 192.838
523	Cidade São Mateus	-1.04%	R$ 3.180	R$ 360.973
524	Parque São Rafael	-0.29%	R$ 3.123	R$ 334.544

525 rows × 4 columns

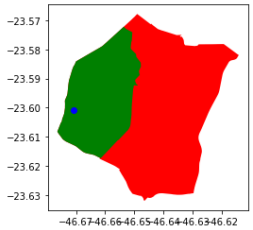
- Coordinates of the 32 sub prefectures and 96 districts:
  Like the library folium, used to create map of the Sao Paulo city with his division, runs with the coordinates defined in spherical system (latitude/longitude degrees **EPSG:4326**), it has been necessary to convert, to this unit, the coordinates of the administrative limits (in form of polygon) of the sub prefecture and district, present in the shape file with the unit UTM Cartesian (**EPSG:29193)**:

```
/usr/local/lib/python3.6/dist-packages/pyproj/crs/crs.py:53: FutureWarning: '+init=<a
  return _prepare_from_string(" ".join(pjargs))
There is 96 districts.
  ds_codigo          ds_nome                                           geometry
0        51         MANDAQUI  POLYGON ((-46.65469 -23.43064, -46.65468 -23.4...
1        52         MARSILAC  POLYGON ((-46.60987 -23.90551, -46.60989 -23.9...
2        32            MOEMA  POLYGON ((-46.65360 -23.57220, -46.65358 -23.5...
3        57  PARQUE DO CARMO  POLYGON ((-46.44460 -23.59273, -46.44470 -23.5...
4        60         PERDIZES  POLYGON ((-46.66355 -23.53692, -46.66361 -23.5...
/usr/local/lib/python3.6/dist-packages/pyproj/crs/crs.py:53: FutureWarning: '+init=<a
  return _prepare_from_string(" ".join(pjargs))
There is 32 subprefectures.
```

	sp_codigo	sp_nome	geometry
0	02	PIRITUBA-JARAGUA	POLYGON ((-46.77040 -23.47841, -46.77040 -23.4...
1	03	FREGUESIA-BRASILANDIA	POLYGON ((-46.68975 -23.50904, -46.68995 -23.5...
2	04	CASA VERDE-CACHOEIRINHA	POLYGON ((-46.67362 -23.48004, -46.67359 -23.4...
3	05	SANTANA-TUCURUVI	POLYGON ((-46.62430 -23.51942, -46.62509 -23.5...
4	06	JACANA-TREMEMBE	POLYGON ((-46.56112 -23.48727, -46.56148 -23.4...

- Sub prefecture and district of each neighborhood
  Like it hasn't been found a web site with the indications, easy to recover, of the sub prefecture and district of each Sao Paulo city neighborhood, it has been used a geometric approach to define these: for each neighborhood loop on all the sub prefectures and all districts to find these whose the polygon contains the neighborhood central point.

```
<class 'geopandas.geoseries.GeoSeries'>
<class 'geopandas.geoseries.GeoSeries'>
   sp_codigo      sp_nome                                           geometry
19        12  VILA MARIANA  POLYGON ((-46.66175 -23.61726, -46.66187 -23.6...
  ds_codigo ds_nome                                           geometry
2        32   MOEMA  POLYGON ((-46.65360 -23.57220, -46.65358 -23.5...
       Neighborhood RateMonth  PriceByM2  MedianPrice  Latitude  Longitude
11  Vila Uberabinha    +0.37%  R$ 13.059  R$ 2.388.741 -23.60078  -46.67094
Vila Uberabinha is in the district of MOEMA? 2    True
dtype: bool
Vila Uberabinha is in the subprefecture of VILA MARIANA? 19    True
dtype: bool
```



	Neighborhood	RateMonth	PriceByM2	MedianPrice	Latitude	Longitude	Subprefecture	District
0	Jardim Europa	+0.73%	R$ 18.968	R$ 7.151.087	-23.57621	-46.68416	PINHEIROS	PINHEIROS
1	Chácara Itaim	-2.25%	R$ 18.478	R$ 7.031.147	-23.59182	-46.67881	PINHEIROS	ITAIM BIBI
2	Jardim Paulistano	+0.94%	R$ 18.338	R$ 6.756.043	-23.57191	-46.68685	PINHEIROS	PINHEIROS
3	Vila Nova Conceição	+0.52%	R$ 18.186	R$ 4.114.889	-23.59183	-46.67246	VILA MARIANA	MOEMA
4	Jardim América	+0.36%	R$ 16.985	R$ 7.682.567	-23.56921	-46.67180	PINHEIROS	JARDIM PAULISTA
...	...	...	...	...	...	...	...	...
520	Parque Savoy City	-1.13%	R$ 3.540	R$ 417.503	-23.56462	-46.48631	ITAQUERA	CIDADE LIDER
521	Parque Casa de Pedra	+0.06%	R$ 3.531	R$ 454.416	-23.45212	-46.60123	JACANA-TREMEMBE	TREMEMBE
522	Conjunto Residencial José Bonifácio	+0.18%	R$ 3.287	R$ 192.838	-23.54756	-46.43445	ITAQUERA	JOSE BONIFACIO
523	Cidade São Mateus	-1.04%	R$ 3.180	R$ 360.973	-23.60128	-46.47860	SAO MATEUS	SAO MATEUS
524	Parque São Rafael	-0.29%	R$ 3.123	R$ 334.544	-23.62790	-46.47014	SAO MATEUS	SAO RAFAEL

525 rows × 8 columns

3) Feature Selections
   - The sub prefecture and district limits are used to show these administrative limits in a Sao Paulo map.
   - The center neighborhood coordinates are showed in the same map to have an idea of the selected neighborhoods distribution between the different sub prefectures and districts of Sao Paulo city.
   - The first cluster of these neighborhoods has been mounted with the top ten venues categories of each neighborhood.
   - The second cluster of these neighborhoods has been mounted with the normalized data about the IDH and price/m2 of each neighborhood.

III) Methodology

1) Presentation on a folium map of the neighborhood studied with their administrative divisions (sub prefecture and district):
   - Creation of a folium map centered on Sao Paulo city:

```
[ ] from geopy.geocoders import Nominatim # convert an address into latitude and longitude values
    import folium # map rendering library

    address = 'Sao Paulo, Sao Paulo'
    geolocator = Nominatim(user_agent="tr_explorer")
    location = geolocator.geocode(address)
    latitude = location.latitude
    longitude = location.longitude
    print('The geograpical coordinate of Sao Paulo City are {}, {}.'.format(latitude, longitude))
    # create map of New York using latitude and longitude values
    # create map of New York using latitude and longitude values
    mapSaoPaulo = folium.Map(location=[latitude, longitude], zoom_start=12.3)
```

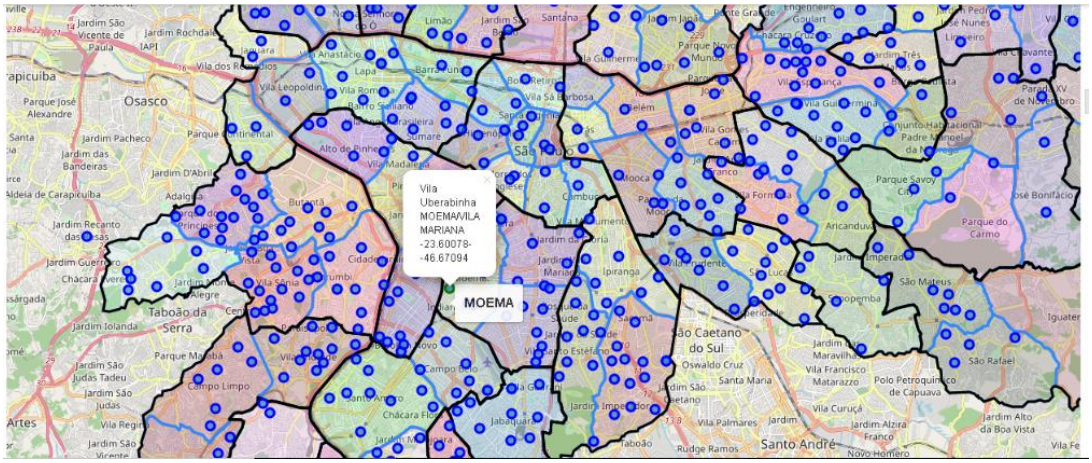
   - Addition on this map, with GeoJson, of the limit of each subprefecture (using the black color to show these limits and of each district (using the blue color for the limits and aleatory color to fill each area district):

```
import random
def random_html_color():
    r = random.randint(0,256)
    g = random.randint(0,256)
    b = random.randint(0,256)
    return '#%02x%02x%02x' % (r, g, b)
def style_fcn2(x):
    return { 'fillColor': random_html_color() }
def style_fcn(x):
    return { 'color':'black', 'fill':False, 'fill_color':'white', 'fill_opacity':0.0 }

style2 = "font-size: 15px; font-weight: bold"
folium.GeoJson(jsDistrit, style_function=style_fcn2,
               tooltip=folium.features.GeoJsonTooltip(['ds_nome'], style=style2, labels=False)
               ).add_to(mapSaoPaulo)
folium.GeoJson(jsSubPref, style_function=style_fcn,
               #tooltip=folium.features.GeoJsonTooltip(['sp_nome'], style=style2, labels=False)
               ).add_to(mapSaoPaulo)
```

- Addition on this map, with folium CircleMarker, of the center of each neighborhood, using the blue color, except for the marker of the neighborhood “Vila Uberabinha”, used for unitary tests, colored in green:

```
[ ] # add markers to map
    #for lat, lng, borough, neighborhood in zip(dfNbh['Latitude'], dfNbh['Longitude'], dfNbh['Borough'], dfNbh['Neighborhood']):
    for lat, lng, neighborhood, district, subprefecture in zip(dfNbh['Latitude'], dfNbh['Longitude'], dfNbh['Neighborhood'], dfNbh['District'], dfNbh['Subprefe
    #    label = '{} of {}'.format(neighborhood, borough)
      label = '{} {}/{} {}-{}'.format(neighborhood, district, subprefecture, lat, lng)
      label = folium.Popup(label, parse_html=True)
      if neighborhood=='Vila Uberabinha':
        coloracao = 'green'
      else:
        coloracao = 'blue'
      folium.CircleMarker(
            [lat, lng],
            radius=5,
            popup=label,
            color=coloracao,
            fill=True,
            fill_color='#3186cc',
            fill_opacity=0.7,
            parse_html=False).add_to(mapSaoPaulo)
```

- Visualization of the map:



2) Recuperation of the TOP 10 venue categories of each studied neighborhood

- Use of the Foursquare API to recover until 100 venues present in a circle of 1km around the center of each neighborhood, with indication of the venue category:

- Exemple of venues of the neighborhood Vila Uberabinha

```
[ ] venuesSaoPaulo[venuesSaoPaulo.Neighborhood=='Vila Uberabinha']
```

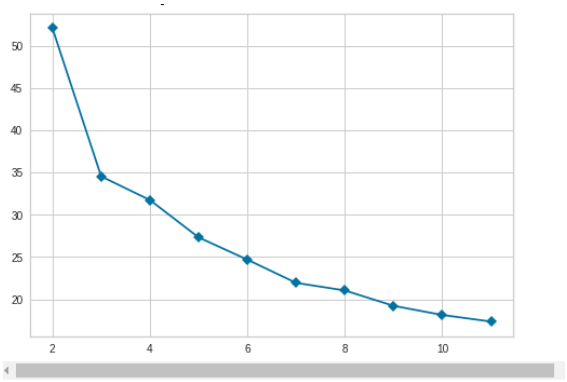
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1100	Vila Uberabinha	-23.60078	-46.67094	Atrium Saúde	-23.601070	-46.671485	Health & Beauty Service
1101	Vila Uberabinha	-23.60078	-46.67094	Bráz Pizzaria	-23.599948	-46.672866	Pizza Place
1102	Vila Uberabinha	-23.60078	-46.67094	Museo Veronica	-23.600247	-46.669106	Spanish Restaurant
1103	Vila Uberabinha	-23.60078	-46.67094	Câmara Fria	-23.600048	-46.672806	Speakeasy
1104	Vila Uberabinha	-23.60078	-46.67094	Kopenhagen	-23.600870	-46.668023	Chocolate Shop
...	...	...	...	...	...	...	...
1195	Vila Uberabinha	-23.60078	-46.67094	Sorvetes Rochinha	-23.606795	-46.667994	Ice Cream Shop
1196	Vila Uberabinha	-23.60078	-46.67094	Ruella	-23.593423	-46.674789	French Restaurant
1197	Vila Uberabinha	-23.60078	-46.67094	Drogaria São Paulo	-23.604824	-46.668842	Pharmacy
1198	Vila Uberabinha	-23.60078	-46.67094	Hi Pin Shan	-23.594034	-46.673859	Chinese Restaurant
1199	Vila Uberabinha	-23.60078	-46.67094	Au Vin Wine Shop and Tasting Bar	-23.596126	-46.667001	Liquor Store

- Transformation of the neighborhood venues data frame in a TOP 10 venue categories data frame by neighborhood:

```
# create a new dataframe
venuesSortedNbh = pd.DataFrame(columns=columns)
venuesSortedNbh['Neighborhood'] = groupedSaoPaulo['Neighborhood']

for ind in np.arange(groupedSaoPaulo.shape[0]):
    venuesSortedNbh.iloc[ind, 1:] = return_most_common_venues(groupedSaoPaulo.iloc[ind, :], num_top_venues)

venuesSortedNbh.head()
```

	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	Aclimação	Gym / Fitness Center	Pizza Place	Bakery	Pet Store	Coffee Shop	Hostel	Dance Studio	Korean Restaurant	BBQ Joint	Bar
1	Alto da Boa Vista	Brazilian Restaurant	Park	Bakery	Japanese Restaurant	Steakhouse	Butcher	Gym	Pizza Place	Music Venue	Department Store
2	Alto da Lapa	Bakery	Gym / Fitness Center	Dessert Shop	Plaza	Restaurant	Ice Cream Shop	Pizza Place	Coffee Shop	Burger Joint	Donut Shop
3	Alto da Mooca	Bakery	Bar	Pizza Place	Burger Joint	Italian Restaurant	Gym / Fitness Center	Dessert Shop	Deli / Bodega	Dance Studio	Middle Eastern Restaurant
4	Alto de Pinheiros	Plaza	Restaurant	Athletics & Sports	Bar	Convenience Store	Park	Bike Rental / Bike Share	Trail	Dog Run	Gym / Fitness Center

3) Clustering of the neighborhood by the TOP 10 venue categories, using the K-Means ML algorithm
   - Verification of the optimum k hyper parameter by the elbow method of the library yellowbrick.cluster:



We will use of 3 for the parameter k of the K-Mean algoritm

   - Execution of the K-Means, using the sklearn.cluster library, with K to 3 and the default hyper parameters:

- Use of the K-Means algoritm to mount 3 clusters of neighborhood acording to their TOP 10 venues categories

```
# set number of clusters
kClust1 = 3

# run k-means clustering
kmeans = KMeans(n_clusters=kClust1, random_state=1).fit(groupedClusteringSaoPaulo)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([0, 1, 0, 0, 0, 2, 2, 0, 1, 0], dtype=int32)
```

- Conservation of the label cluster by Top venues categories of each neighborhood

```
# add clustering labels
#venuesSortedNbh.drop(columns=["Cluster Labels"], inplace=True)
venuesSortedNbh.insert(0, 'Cluster By TOP Venues Categ.', kmeans.labels_)
venuesSortedNbh
```

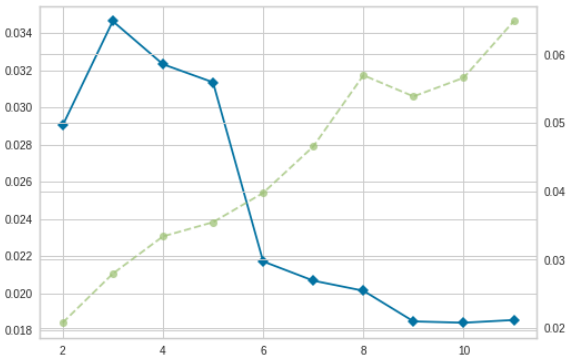
4) Clustering of the neighborhood by the socio-economic indicators (Human Development Indicator and house price by m2), using the K-Means ML algorithm
   - Cleaning and normalization of these data:

```
from sklearn import preprocessing

#New data frame of the Sao Paulo neighborhood considering the IDH and the price/m2
dfCluster = dfNbh.loc[0:, ['PriceByM2','Idh']]

#Transformation of the price in number
dfCluster.PriceByM2=dfCluster.PriceByM2.str.replace(pat='R$ ', repl='', regex=False).str.replace(pat='.', repl='', regex=False).astype(np.float)
#Force 0 as Idh for the samples without Idh to can normalize afterward all the data frame
dfCluster.loc[dfCluster[dfCluster.Idh.isnull()].index, 'Idh']=0
#Normalization of the data frame to use for the clustering
dfCluster=preprocessing.normalize(X=dfCluster, norm='max', axis=0, )
print(dfCluster[0:10])
```

```
[[1.         0.99895942]
 [0.97416702 0.99167534]
 [0.96678617 0.99895942]
 [0.95877267 1.        ]
 [0.8954555  0.99583767]
 [0.85528258 1.        ]
 [0.84168073 0.99167534]
 [0.83582876 0.9760666 ]
 [0.7978701  0.99167534]
 [0.70661113 0.99583767]]
```

   - Verification of the optimum k hyper parameter by the elbow method of the library yellowbrick.cluster:

```
0.018409329767544228,
0.018556247474571824]
```



We will use of 6 for the parameter k of the K-Mean algoritm

- Execution of the K-Means, using the sklearn.cluster library, with K to 6 and the default hyper parameters:

- Use of the K-Means algoritmo to mount 6 clusters of neighborhood acording to their IDH and house price by m2

```
[ ] # set number of clusters
    kClust2 = 6

    # run k-means clustering
    kmeans = KMeans(n_clusters=kClust2, random_state=1).fit(dfCluster)

    # check cluster labels generated for each row in the dataframe
    kmeans.labels_[0:10]
```

```
array([3, 3, 3, 3, 3, 3, 3, 3, 3, 4], dtype=int32)
```

- Conservation of the label cluster by socio-economic indicators of each neighborhood

```
[ ] #Creation of a new data frame, copy of the data frame with the coordinates of each neighborhood
    mergedSaoPaulo = dfNbh.copy()
    #Add in this new data frame the second clustering label of each neighborhood
    mergedSaoPaulo.insert(0, 'Cluster By IDH-M2Price', kmeans.labels_)
    mergedSaoPaulo
```

	Cluster By IDH-M2Price	Neighborhood	RateMonth	PriceByM2	MedianPrice	La
0	3	Jardim Europa	+0.73%	R$ 18.968	R$ 7.151.087	-2

5) Representation on a Sao Paulo city map of the two clusters of each studied neighborhood

- Merge in a same data frame of the label of the two clusters for each neighborhood

```
[ ] # merge SaoPaulo_grouped with SaoPaulo_data to add latitude/longitude for each neighborhood
    mergedSaoPaulo = mergedSaoPaulo.join(venuesSortedNbh.set_index('Neighborhood'), on='Neighborhood')

    mergedSaoPaulo.head() # check the first rows
```

	Cluster By IDH-M2Price	Neighborhood	RateMonth	PriceByM2	MedianPrice	Latitude	Longitude	Subprefecture	District	Idh	Cluster By TOP Venues Categ.	1st Most Common Venue	2n
0	3	Jardim Europa	+0.73%	R$ 18.968	R$ 7.151.087	-23.57621	-46.68416	PINHEIROS	PINHEIROS	0.960	0.0	Italian Restaurant	Res
1	3	Chácara Itaim	-2.25%	R$ 18.478	R$ 7.031.147	-23.59182	-46.67881	PINHEIROS	ITAIM BIBI	0.953	0.0	Italian Restaurant	Res
2	3	Jardim Paulistano	+0.94%	R$ 18.338	R$ 6.756.043	-23.57191	-46.68685	PINHEIROS	PINHEIROS	0.960	0.0	Middle Eastern Restaurant	

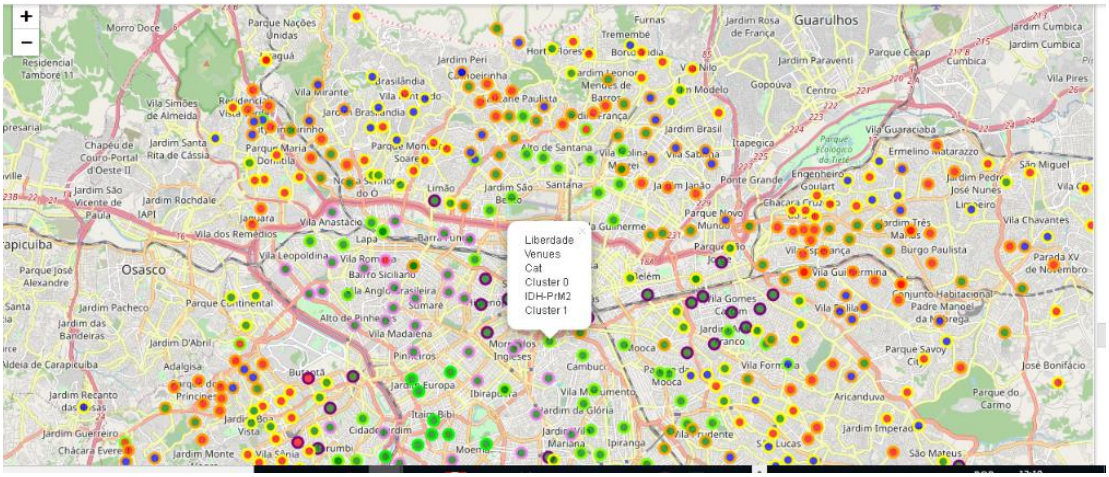
- Setting to the label 3 (not use for the clustering by TOP 10 venue categories) of the neighborhoods without venue recovered by the Foursquare API

- We force these neighborhoods to a new label 3 for this clustering by TOP venues caterories

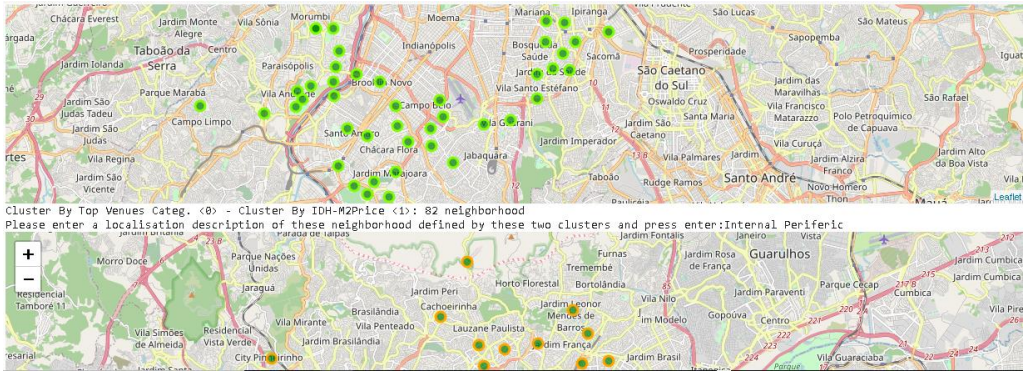
```
[ ] mergedSaoPaulo.loc[mergedSaoPaulo["Cluster By TOP Venues Categ."].isnull(), "Cluster By TOP Venues Categ."] = kClust1
    mergedSaoPaulo[(mergedSaoPaulo["Cluster By TOP Venues Categ."]==kClust1)]
```

	Cluster By IDH-M2Price	Neighborhood	RateMonth	PriceByM2	MedianPrice	Latitude	Longitude	Subprefecture	District	Idh	Cluster By TOP Venues Categ.	1st Most Common Venue	Co
162	0	Campininha	-0.59%	R$ 6.444	R$ 523.945	-23.23433	-47.87848	NaN	NaN	NaN	3.0	NaN	
178	0	Maranhão	+0.12%	R$ 6.207	R$ 467.925	-1.63333	-44.96667	NaN	NaN	NaN	3.0	NaN	

- Representation, in a Sao Paulo city folium map, of the two neighborhood clusters, using a circle color different by label of the IDH/priceM2 cluster and a filling color different by label of the TOP 10 venue categories cluster:



6) Verification, visually, if some neighborhood clusters group (cartesian product of the labels of the two clusters) have a localization concentration
Loop on each neighborhood cluster group (by TOP 10 venue categories and by IDH/priceM2 labels) to consult on a map the localization of their neighborhood and eventually input manually a description of this localization:



- Neighborhood cluster groups with e description localisation inputed

```
[ ] groupCluster = mergedSaoPaulo.loc[(mergedSaoPaulo.DescLocal.str.len()>0), ['Cluster By TOP Venues Categ.', 'Cluster By IDH-M2Price', 'DescLocal']]
    groupCluster=groupCluster.groupby(by=['Cluster By TOP Venues Categ.', 'Cluster By IDH-M2Price', 'DescLocal'])
    groupCluster.size()
    groupCluster.agg({'DescLocal':['count']})
```

			DescLocal
			count
Cluster By TOP Venues Categ.	Cluster By IDH-M2Price	DescLocal	
0	0	Periferic	45
	1	Internal Periferic	82
	2	External Periferic	69
	3	South East	9
	4	South-North East	44
	5	Center, South, East	35

7) Study of the socio-economic indicators of each neighborhood cluster group:
- Grouping of the neighborhoods by their two label clusters

- Grouping of the neighborhoods by theirs two label clusters (TOP venue categories and IDH-price by m2

```
[ ] groupCluster=groupCluster.groupby(by=['Cluster By TOP Venues Categ.', 'Cluster By IDH-M2Price'])
```

- Statistical description of the IDH and house price/m2 of each neighborhood cluster group

		Idh								PriceByM2							
		count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
Cluster By TOP Venues Categ.	Cluster By IDH-M2Price																
0	0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	45.0	5392.266667	758.322593	3180.0	4729.00	5574.0	6008.00	6400.0
	1	82.0	0.907317	0.033136	0.806	0.88500	0.9090	0.93800	0.957	82.0	7004.219512	877.574309	5288.0	6347.50	6892.0	7641.00	8754.0
	2	69.0	0.859449	0.035571	0.795	0.83400	0.8580	0.88400	0.957	69.0	5108.739130	582.288554	4008.0	4648.00	5225.0	5521.00	6274.0
	3	9.0	0.955111	0.007305	0.938	0.95300	0.9570	0.96000	0.961	9.0	17125.666667	1395.648147	15134.0	15965.00	16985.0	18338.00	18968.0
	4	44.0	0.947614	0.012301	0.907	0.94100	0.9530	0.95700	0.961	44.0	10351.363636	1405.101521	8602.0	9218.75	9823.5	11410.00	13403.0
	5	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	35.0	7924.657143	1145.501148	6703.0	7055.50	7618.0	8424.00	11253.0
1	0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	71.0	4754.985915	678.347582	3287.0	4295.00	4685.0	5217.00	6273.0
	1	10.0	0.867500	0.053284	0.777	0.82725	0.8735	0.91375	0.935	10.0	6669.000000	947.669656	5520.0	6007.75	6404.5	7048.50	8394.0
	2	72.0	0.841986	0.034379	0.766	0.82275	0.8410	0.86400	0.935	72.0	4897.444444	627.144386	3685.0	4502.50	4895.5	5298.25	6390.0
	5	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	5.0	7875.800000	1582.875295	6520.0	6528.00	7268.0	9024.00	10039.0
2	0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	30.0	4764.333333	764.484245	3123.0	4188.50	4757.0	5220.25	6310.0

- Loop on each neighborhood cluster group to consult the statistics of the IDH and price by m2 series and eventually input a description about these statistics

	Idh	PriceByM2
count	0.0	45.000000
mean	NaN	5392.266667
std	NaN	758.322593
min	NaN	3180.000000
25%	NaN	4729.000000
50%	NaN	5574.000000
75%	NaN	6008.000000
max	NaN	6400.000000

```
Cluster By Top Venues Categ. <0> - Cluster By IDH-M2Price <0>: 45 neighborhood
Please enter a social-economic description of these neighborhood defined by these two clusters and press enter:Without IDH and very low price by m2
```

	Idh	PriceByM2
count	82.000000	82.000000
mean	0.907317	7004.219512

- Neighborhood cluster groups with a socio-economic description inputed

```
groupCluster = mergedSaoPaulo.loc[(mergedSaoPaulo.DescSocEco.str.len()>0), ['Cluster By TOP Venues Categ.', 'Cluster By IDH-M2Price', 'DescSocEco']]
groupCluster=groupCluster.groupby(by=['Cluster By TOP Venues Categ.', 'Cluster By IDH-M2Price', 'DescSocEco'])
groupCluster.size()
```

```
Cluster By TOP Venues Categ.  Cluster By IDH-M2Price  DescSocEco
0                             0                      Without IDH and very low price by m2    45
                              3                      High IDH and Very High m2 price          9
                              4                      High IDH and high m2 price              44
dtype: int64
```

8) Study of the venue categories of each neighborhood cluster group

- Grouping of the neighborhoods by their two label clusters

- Grouping of the neighborhoods by theirs two label clusters (TOP venue categories and IDH-price by m2

```
[ ] groupCluster=groupCluster.groupby(by=['Cluster By TOP Venues Categ.', 'Cluster By IDH-M2Price'])
```

- Loop on each neighborhood cluster group to consult the statistics of their venue categories and eventually input a description about these statistics

	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
count	82	82	82	82	82	82	82	82	82	82
unique	18	22	26	31	32	39	39	47	51	45
top	Bar	Pizza Place	Bakery	Pharmacy	Dessert Shop	Gym / Fitness Center	Gym	Burger Joint	Bakery	Pizza Place
freq	15	13	14	12	7	9	8	6	6	6

```
Cluster By Top Venues Categ. <0> - Cluster By IDH-M2Price <1>: 82 neighborhood
Please enter a venues categories description of these neighborhood defined by these two clusters and press enter:Bar and Pizza
```

	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
173	Bar	Bakery	Gym / Fitness Center	Pharmacy	Brazilian Restaurant	Burger Joint	Pet Store	Pizza Place	Restaurant	Ice Cream Shop

- Neighborhood cluster groups with a TOP venue categories description inputed

```
[ ] groupCluster = mergedSaoPaulo.loc[(mergedSaoPaulo.DescVenCat.str.len()>0), ['Cluster By TOP Venues Categ.', 'Cluster By IDH-M2Price', 'DescVenCat']]
    groupCluster=groupCluster.groupby(by=['Cluster By TOP Venues Categ.', 'Cluster By IDH-M2Price', 'DescVenCat'])
    groupCluster.size()
```

```
Cluster By TOP Venues Categ.  Cluster By IDH-M2Price  DescVenCat
0                             0                      Restaurant                45
                              1                      Bar and Pizza             82
                              2                      Bakery and Pizza          69
                              3                      Restaurant (Italian...)    9
dtype: int64
```

9) Descriptions inputted (about localization, IDH-price/M2 indicators and TOP 10 venue categories) for each neighborhood cluster groups
   - The descriptions are saved in a Json file:

IV.4) Saving in a Json file of the result data frame of the neighborhoods with their cluster label and descriptions inputed

```
[ ] #Saving of the Foursquare API results in a Json file
    mergedSaoPaulo.to_json(path_or_buf="/content/drive/My Drive/Colab Notebooks/mergedSaoPaulo.json")
```

   - Cluster groups with descriptions inputted:

Cluster By TOP Venues Categ.	Cluster By IDH-M2Price	DescLocal	DescSocEco	DescVenCat
0	0	Internal Perifery and Center	Without IDH and low price/m2	Restaurant (principally Brazilian)
	1	Internal Periphery	Medium IDH and price/m2	Restaurant (principally Brazilian) an
	2	External Periphery	Low IDH and price/m2	Bakery and Pizzeria
	3	Center West	Very High IDH and price/m2	Restaurant (principally Italian)
	4	West North-Center-South	Very High IDH and high price/m2	Restaurant (principally Brazilian) an
	5	Center, East and South	Without IDH and medium price/m2	Restaurant and Bar
1	0	External Periphery	Without IDH and very low price/m2	Bakery
	1		Low IDH and medium price/m2	Pizzeria and Bakery
	2	External Periphery	Low IDH and price/M2	Pizzeria
	5		Without IDH and medium price/m2	Pizzeria
2	0		Without IDH and very low price/m2	Bakery
	2	External Periphery	Very low IDH and low price/m2	Bakery

IV) Results and discussion:

As was to be expected, given the size and population of the city, Sao Paulo presents a wide variety of neighborhood.

At most of the administrative divisions of these neighborhoods, it has been shown:

- Three clusters of these neighborhoods considering their principal venue categories. The principals venue categories are restaurant, pizzeria and bakery.
- Six clusters of these neighborhoods considering there IDH and house price by m2. The neighborhood cluster group with the better socio-economics indicators is formed of 9 neighborhoods and is located in the south east of the extended center of Sao Paulo (label 0 for the TOP 10 venue categories cluster and 3 for the IDH-price/m2 cluster).

The study could have been more precise, if we has been able to access data from all of the neighborhood of Sao Paulo (more than 1700).

It would have been interesting to use more socio-economics indicators of these neighborhoods, if the Internet site with the las referencing had been on-line.

V) Conclusion

This study allows having a better vision of the disparities between the thousands of neighborhood of the megalopolis Sao Paulo.

It should give insights for the businessman to identify the best local to open a new business.