Capstone Project - The Battle of Neighborhoods

1 A description of the problem and a discussion of the background

We are planning to build a luxury hotel in Barcelona, and we want to know which neighbourhood is the best.

Barcelona was the 20th-most-visited city in the world by international visitors and the fifth most visited city in Europe after London, Paris, Istanbul and Rome, with 5.5 million international visitors in 2011. By 2015, both Prague and Milan had more international visitors. With its Rambles, Barcelona is ranked the most popular city to visit in Spain.

Barcelona as internationally renowned a tourist destination, with numerous recreational areas, one of the best beaches in the world, mild and warm climate, historical monuments, including eight UNESCO World Heritage Sites, and developed tourist infrastructure.

Barcelona is divided into 10 districts. These are administrated by a councillor designated by the main city council, and each of them have some powers relating to issues such as urbanism or infrastructure in their area. The current division of the city into different districts was approved in 1984. In 2009, in Barcelona started using a new division of 73 neighbourhoods (the 10 districts are still in use), a division that was done for a better service from the City Council.

Some of these districts have a previous history as independent municipalities which were integrated into the city of Barcelona during the late 19th century and the first half of the 20th century, such as Sarrià, Les Corts, Sant Andreu de Palomar, Gràcia or Sant Martí de Provençals. However, other municipalities which are contiguous to Barcelona (such as L'Hospitalet de Llobregat, Badalona, Sant Adrià de Besòs or Montcada i Reixac) have remained separate towns to this day and are part of the much larger metropolitan area of Barcelona.

We want to use an index of the family income (RFD) and the quantity of the hotels by neighbourhood to create clusters and see which one fits better with our goal.

2 A description of the data and how it will be used to solve the problem.

For this project we are going to use different datasets:

- Barcelona's last published income index by Neighbourhood (2017)
    - Source: https://opendata-ajuntament.barcelona.cat/data/dataset/79bdf758-dae1-485b-800c-be9f8cfa9360/resource/e7206797-e57b-4ded-8c6c-62e9b4cb54f7/download/2017_distribucio_territorial_renda_familiar.csv
    - Description: RFD (Renta Familiar Disponible per cápita) is the amount of income available to resident families for consumption and savings, once the depreciation or consumption of fixed capital in family economic operations and direct taxes and contributions paid to Social Security have been deducted. Eurostat recommends the use of the RFD as the main regional economic aggregate.

- Barcelona Hotels Foursquare
    - Source: Foursquare API
    - Description: Using the Foursquare API we'll get all the hotels in Barcelona.
- Barcelona Hotels Open Data
    - Source: https://opendata-ajuntament.barcelona.cat/data/dataset/88efe464-2bcd-4794-85b0-8b0bbfd9e4c0/resource/eced0fe8-9892-4926-b035-4fdc7328e31d/download
    - Description: We've looked for another hotel's source because we felt Foursquare API wasn't complete enough. This data set includes an accurate description of every hotel in the city

3 Methodology

For this project we've use Open Data sources to know the family incomes by neighbourhood in Barcelona and for having a complete hotel's list (to compare with Foursquare API) and the Foursquare API to find the hotels in Barcelona city.

We have done three different clusters to approach

3.1 Data Preparation

There is a Open Data Project in Barcelona so we use it to find the list of Barcelona Neighbourhood's with its Name, Code, District, Population and incomes (RFD).

After some manipulation the first 5 rows are like this.

```
In [2]: #Download a Dataset from Barcelona Open Data council
        url = 'https://opendata-ajuntament.barcelona.cat/data/dataset/79bdf758-dae1-485b-800c-be9f8cfa9360/resource/e7206797-e57b-4ded-8c6c-62e9b4cb54f7/download/2017_distribucio_territorial_renda_familiar.csv'
        bcn_rfd = pd.read_csv(url)

        #Some treatment to clean the Data set
        bcn_rfd.loc[bcn_rfd.Nom_Barri=='Vallvidrera, el Tibidabo i les Planes', 'Nom_Barri'] = 'Vallvidrera' #Shortening the neighborhood name
        bcn_rfd = bcn_rfd.drop(['Any'], axis=1)
        bcn_rfd = bcn_rfd.rename(columns={'Índex RFD Barcelona = 100': 'RFD', 'Població': 'Population'})
        bcn_rfd.head()
```

Out[2]:

	Codi_Districte	Nom_Districte	Codi_Barri	Nom_Barri	Population	RFD
0	1	Ciutat Vella	1	el Raval	47986	71.2
1	1	Ciutat Vella	2	el Barri Gòtic	16240	106.1
2	1	Ciutat Vella	3	la Barceloneta	15101	79.8
3	1	Ciutat Vella	4	Sant Pere, Santa Caterina i la Ribera	22923	99.4
4	2	Eixample	5	el Fort Pienc	32048	106.5

This data set doesn't have the coordinates so we used Nominatim to Geocode the Neighbourhoods.

```
In [3]:
# Getting the coordinates of each Neighborhood
locator = Nominatim(user_agent='myGeocoder')
bcn_lon = []
bcn_lat = []

for index, row in bcn_rfd.iterrows():
    try:
        location = locator.geocode(row['Nom_Barri'] + ', Barcelona')
        bcn_lon.append(location.longitude)
        bcn_lat.append(location.latitude)
    except:
        bcn_lon.append('Error')
        bcn_lat.append('Error')
        print("An exception occurred")

        #Adding the coordinates to the data frame
bcn_rfd['Latitude'] = bcn_lat
bcn_rfd['Longitude'] = bcn_lon
bcn_rfd.head()
```

Out[3]:

	Codi_Districte	Nom_Districte	Codi_Barri	Nom_Barri	Population	RFD	Latitude	Longitude
0	1	Ciutat Vella	1	el Raval	47986	71.2	41.379518	2.168368
1	1	Ciutat Vella	2	el Barri Gòtic	16240	106.1	41.383395	2.176912
2	1	Ciutat Vella	3	la Barceloneta	15101	79.6	41.380653	2.189927
3	1	Ciutat Vella	4	Sant Pere, Santa Caterina i la Ribera	22923	99.4	41.389911	2.183068
4	2	Eixample	5	el Fort Pienc	32048	106.5	41.395925	2.182325

For the Hotels List we used two Data sources to compare. Here there is the preparation of the open data list.

```
In [13]:
url2 = 'https://opendata-ajuntament.barcelona.cat/data/dataset/88efe464-2bcd-4794-85b0-8b0bbfd9e4c0/resource/eced0fe8-9892-4926-b035-4fdc7328e31d/download'
bcn_hotels = pd.read_csv(url2)

#Dropping all duplicate values
bcn_hotels.drop_duplicates(subset ="CODI_EQUIPAMENT", keep = 'first', inplace = True)

bcn_hotels
```

Out[13]:

	CODI_EQUIPAMENT	EQUIPAMENT	SECCIO	TIPUS_VIA	NOM_CARRER	NUM_CARRER_1	NUM_CARRER_2	CODI_BARRI	NUM_BARRI	CODI_DISTRICTE	...	HORARI_PERIODE_INICI	HORARI_P
0	12124527	Hotel Amrey Sant Pau - HB-004046	#	C	Sant Antoni Maria Claret	173	173.0	35	el Guinardó	7	...	NaN	
2	75145812	Hotel Sansi Pedralbes - HB-004086	#	Av	Pearson	1	3.0	21	Pedralbes	4	...	NaN	
4	139120246	Hotel Silken Sant Gervasi - HB-004054	#	C	Sant Gervasi de Cassoles	26	26.0	25	Sant Gervasi - la Bonanova	5	...	NaN	
6	182121024	Hotel Attica21 Barcelona Mar - HB-004340	#	C	Provençals	10	10.0	35	el Guinardó	10	...	NaN	
8	273135303	Abba Sants Hotel - HB-004078	#	C	Numància	32	32.0	18	Sants	3	...	NaN	

After some manipulation we'll have a new data frame with the hotels amount by neighbourhood.

```
In [14]: #Grouping the hotels by neighbourhod and cleaning and sorting the result

bcn_hotels['num_hotels'] = bcn_hotels.groupby('CODI_BARRI')['CODI_BARRI'].transform('count')
hotels_barri = bcn_hotels.filter(['CODI_BARRI', 'num_hotels'], axis=1)
hotels_barri.drop_duplicates(subset ="CODI_BARRI", keep = 'first', inplace = True)
hotels_barri = hotels_barri.sort_values('num_hotels', ascending=True)
hotels_barri = hotels_barri.rename(columns={'CODI_BARRI': 'Codi_Barri'})
hotels_barri.head()
```

Out[14]:

	Codi_Barri	num_hotels
810	17	1
334	43	1
374	29	1
194	12	1
412	22	1

3.2 Using Foursquare API

We have to use Foursquare because it's a Project requirement so we prepared the Request with the required data.

```
In [10]: #Setting up the Foursquare API call method
CLIENT_ID = '[redacted]' # your Foursquare ID
CLIENT_SECRET = '[redacted]' # your Foursquare Secret
VERSION = '20180604'
LIMIT = 50
search_query = 'hotel'
radius = 10000

In [11]: #Calling the API
df = pd.DataFrame()
for lat, lon in zip(bcn_rfd['Latitude'], bcn_rfd['Longitude']):
    latitude = lat
    longitude = lon
    url = 'https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll={},{}&v={}&query={}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, latitude, longitude, VERSION, search_query, radius, LIMIT)
    results = requests.get(url).json()
    # assign relevant part of JSON to venues
    hotels = results['response']['venues']

    # tranform venues into a dataframe
    dataframe = json_normalize(hotels)

    #Adding the new dataframe to the existing one
    df = df.append(dataframe)
```

Afther some manipulation, the result is:

```
In [12]: # keep only columns that include venue name, and anything that is associated with location
filtered_columns = ['name', 'categories'] + [col for col in df.columns if col.startswith('location.')] + ['id']
df_filtered = df.loc[:, filtered_columns]

# function that extracts the category of the venue
def get_category_type(row):
    try:
        categories_list = row['categories']
    except:
        categories_list = row['venue.categories']

    if len(categories_list) == 0:
        return None
    else:
        return categories_list[0]['name']

# filter the category for each row
df_filtered['categories'] = df_filtered.apply(get_category_type, axis=1)

# clean column names by keeping only last term
df_filtered.columns = [column.split('.')[-1] for column in df_filtered.columns]

#Dropping all duplicate values
df_filtered.drop_duplicates(subset ="address", keep = 'first', inplace = True)

df_filtered
```

Out[12]:

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lng	neighborhood	p
0	Fairmont Hotel Rey Juan Carlos I (Hotel Rey Ju...	Hotel	Av. Diagonal, 661-671	ES	Barcelona	España	NaN	1021	[Av. Diagonal, 661-671, 08028 Barcelona Catalu...	[{'label': 'display', 'lat': 41.38128935295143...	41.381289	2.109007	Les Corts	
1	Hotel Sansi Pedralbes	Hotel	Av. Pearson, 1-3	ES	Barcelona	España	NaN	369	[Av. Pearson, 1-3, Barcelona Cataluña, España]	[{'label': 'display', 'lat': 41.39343506028381...	41.393435	2.111722	NaN	
2	Piscina Hotel Rey Juan Carlos I	Hotel Pool	Hotel Rey Juan Carlos I	ES	Barcelona	España	NaN	986	[Hotel Rey Juan Carlos I, 08028 Barcelona Cata...	[{'label': 'display', 'lat': 41.38138119681831...	41.381381	2.110405	NaN	
3	abba Garden hotel 4*	Hotel	C/ Santa Rosa, 33	ES	Esplugues de Llobregat	España	NaN	1008	[C/ Santa Rosa, 33, 08950 Esplugues de Llobreg...	[{'label': 'display', 'lat': 41.38532640804012...	41.385326	2.101987	NaN	
4	Hotel Catalonia Rigoletto ****	Hotel	Sabino de Arana 22-24	ES	Barcelona	España	NaN	1215	[Sabino de Arana 22-24, 08028 Barcelona Catalu...	[{'label': 'display', 'lat': 41.38559623122477...	41.385596	2.125446	NaN	
5	Hotel Sofia Barcelona - In the Unbound Collect...	Hotel	Plaça De Pius Xii 4	ES	Barcelona	España	NaN	1035	[Plaça De Pius Xii 4, 08028 Barcelona Cataluña...	[{'label': 'display', 'lat': 41.38598, 'lng':...	41.385980	2.123314	NaN	
6	AC Hotel Victoria Suites by Marriott	Hotel	Beltran I Rózpide, 7-9	ES	Barcelona	España	NaN	767	[Beltran I Rózpide, 7-9, 08034 Barcelona Catal...	[{'label': 'display', 'lat': 41.389324, 'lng':...	41.389324	2.121338	NaN	

3.3 Clustering

We have used three different approaches for clustering. All with kmean function but with different data

1. With RFD Index, Population and Coordinates
2. With RFD Index and Coordinates
3. With RFD and the number of hotels

```
In [31]:
#Clustering neigbourhoods with the number of hotels and RFD
# set number of clusters
kclusters = 5

#We are going to try different clusters with different columns
#Third cluster with RFD Index and number of hotels
bcn_clus3 = bcn.drop(['Codi_Districte', 'Nom_Districte', 'Codi_Barri', 'Nom_Barri', 'Latitude', 'Longitude', 'Population', 'Latitude', 'Longitude', 'Cluster', 'Cluster2'], 1)


# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(bcn_clus3)

# check cluster labels generated for each row in the dataframe
clus3 = kmeans.labels_

bcn['Cluster3'] = clus3
bcn
```

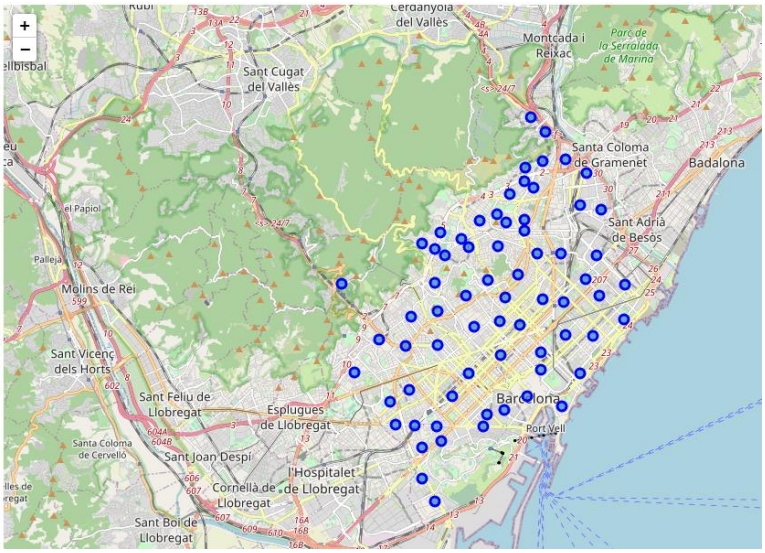
Out[31]:

	Codi_Barri	Codi_Districte	Nom_Districte	Nom_Barri	Population	RFD	Latitude	Longitude	Cluster	Cluster2	num_hotels	Cluster3
0	20	4	Les Corts	Pedralbes	12117	248.8	41.390140	2.112218	0	3	7.0	1
1	23	5	Sarrià-Sant Gervasi	les Tres Torres	16660	215.8	41.397611	2.131184	0	3	3.0	1
2	22	5	Sarrià-Sant Gervasi	Sarrià	25106	193.6	41.399373	2.121513	2	3	1.0	1
3	25	5	Sarrià-Sant Gervasi	Sant Gervasi - Galvany	47753	192.1	41.397807	2.143377	3	3	6.0	1
4	24	5	Sarrià-Sant Gervasi	Sant Gervasi - la Bonanova	25909	184.6	41.405983	2.133405	2	3	3.0	1
5	6	2	Eixample	la Dreta de l'Eixample	44246	175.9	41.395037	2.167207	1	0	3.0	3
6	66	10	Sant Martí	la Vila Olímpica del Poblenou	9404	164.2	41.389868	2.196846	0	0	11.0	3
7	68	10	Sant Martí	Diagonal Mar i el Front Marítim del Poblenou	13710	150.1	41.405228	2.213352	0	0	9.0	3
8	26	5	Sarrià-Sant Gervasi	el Putxet i el Farró	29617	144.6	41.407476	2.143283	2	0	11.0	3
9	21	5	Sarrià-Sant Gervasi	Vallvidrera	4689	144.1	41.415067	2.107482	4	0	2.0	3
10	7	2	Eixample	l'Antiga Esquerra de l'Eixample	42512	137.2	41.390000	2.155000	1	0	82.0	4
11	18	4	Les Corts	les Corts	46104	120.0	41.385244	2.132863	3	2	4.0	2

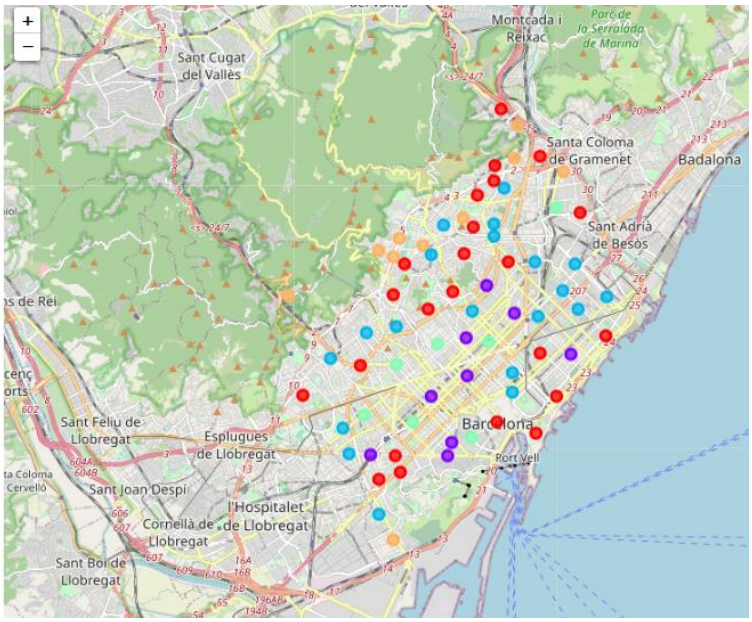
3.4 Data visualization

Barcelona with the 73 neighbourhoods

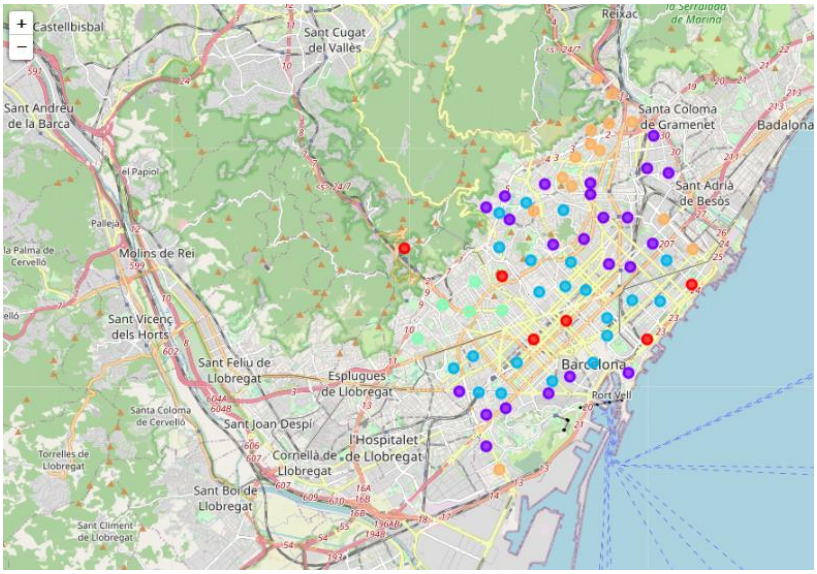
Out[4]:



First Cluster

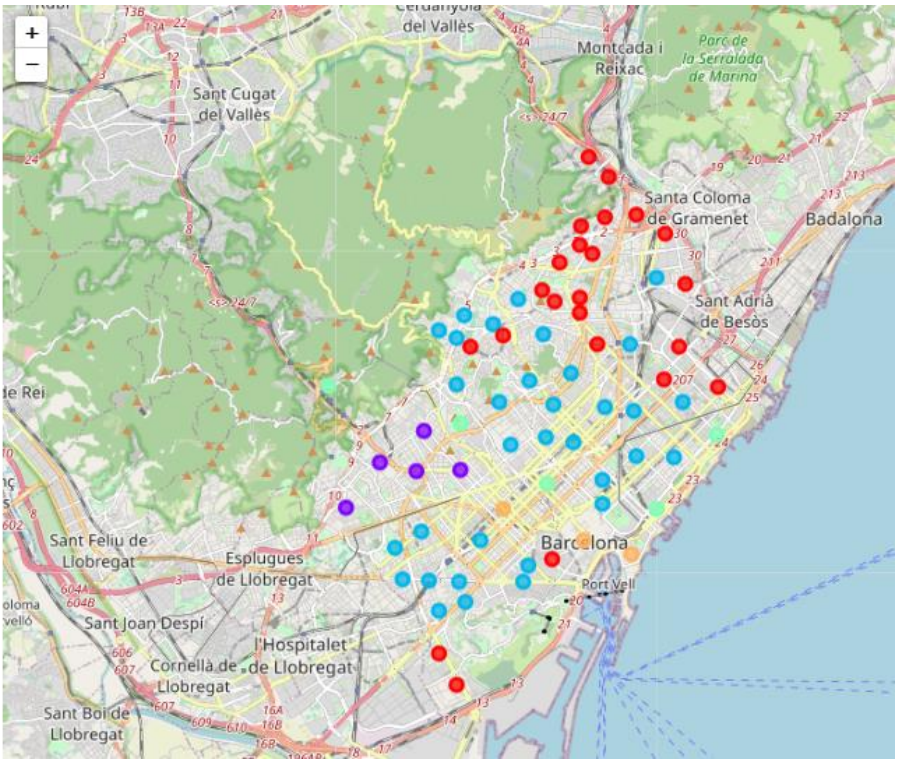


Second Cluster



Third cluster

Out[19]:



4 Results

After some analysis we have found that the third cluster is the best who fits with our proposal, because it offers us a list of neighbourhoods with high income index and a low number of hotels.

Out[18]:

	Codi_Barri	Codi_Districte	Nom_Districte	Nom_Barri	Population	RFD	Latitude	Longitude	Cluster	Cluster2	num_hotels	Cluster3
0	20	4	Les Corts	Pedralbes	12117	248.8	41.390140	2.112218	0	3	7.0	1
1	23	5	Sarrià-Sant Gervasi	les Tres Torres	16660	215.8	41.397611	2.131184	0	3	3.0	1
2	22	5	Sarrià-Sant Gervasi	Sarrià	25106	193.6	41.399373	2.121513	2	3	1.0	1
3	25	5	Sarrià-Sant Gervasi	Sant Gervasi - Galvany	47753	192.1	41.397807	2.143377	3	3	6.0	1
4	24	5	Sarrià-Sant Gervasi	Sant Gervasi - la Bonanova	25909	184.6	41.405983	2.133405	2	3	3.0	1
5	6	2	Eixample	la Dreta de l'Eixample	44246	175.9	41.395037	2.167207	1	0	3.0	3
6	66	10	Sant Martí	la Vila Olímpica del Poblenou	9404	164.2	41.389868	2.196846	0	0	11.0	3
7	68	10	Sant Martí	Diagonal Mar i el Front Marítim del Poblenou	13710	150.1	41.405228	2.213352	0	0	9.0	3
8	26	5	Sarrià-Sant Gervasi	el Putxet i el Farró	29617	144.6	41.407476	2.143283	2	0	11.0	3
9	21	5	Sarrià-Sant Gervasi	Vallvidrera	4689	144.1	41.415067	2.107482	4	0	2.0	3
10	7	2	Eixample	l'Antiga Esquerra de l'Eixample	42512	137.2	41.390000	2.155000	1	0	82.0	4
11	18	4	Les Corts	les Corts	46104	120.0	41.385244	2.132863	3	2	4.0	2
12	19	4	Les Corts	la Maternitat i Sant Ramon	23980	114.2	41.382077	2.125483	2	2	9.0	2
13	27	6	Gràcia	Vallcarca i els Penitents	15615	112.5	41.415526	2.142243	0	2	5.0	2
14	8	2	Eixample	la Nova Esquerra de l'Eixample	58315	110.2	41.383389	2.149000	3	2	25.0	2
15	29	6	Gràcia	la Salut	13207	109.9	41.411866	2.153961	0	2	1.0	2
16	4	2	Eixample	el Fort Pienc	32048	106.5	41.395925	2.182325	2	2	19.0	2
17	1	1	Ciutat Vella	el Barri Gòtic	16240	106.1	41.383395	2.176912	0	2	45.0	4
18	31	6	Gràcia	el Camp d'en Grassot i Gràcia Nova	34431	105.7	41.404589	2.166781	1	2	3.0	2
19	30	6	Gràcia	la Vila de Gràcia	50885	104.4	41.403178	2.157166	3	2	0.0	2
20	9	2	Eixample	Sant Antoni	38412	104.2	41.378412	2.161768	1	2	16.0	2
21	70	10	Sant Martí	Provençals del Poblenou	20649	102.3	41.411948	2.204125	2	2	0.0	2
22	5	2	Eixample	la Sagrada Família	51651	101.8	41.403479	2.174410	3	2	9.0	2
23	65	10	Sant Martí	el Parc i la Llacuna del Poblenou	15204	100.4	41.400733	2.191342	0	2	0.0	2
24	67	10	Sant Martí	el Poblenou	33931	99.9	41.400527	2.201729	1	2	2.0	2
25	3	1	Ciutat Vella	Sant Pere, Santa Caterina i la Ribera	22923	99.4	41.390934	2.182200	2	2	5.0	2
26	14	3	Sants-Montjuïc	Hostafrancs	15949	99.0	41.375088	2.142933	0	2	1.0	2

There is the list of neighbourhoods that are in the Cluster number 3

```
In [20]: goal = bcn[bcn.Cluster3.eq(1)]
         goal
```

Out[20]:

	Codi_Barri	Codi_Districte	Nom_Districte	Nom_Barri	Population	RFD	Latitude	Longitude	Cluster	Cluster2	num_hotels	Cluster3
0	20	4	Les Corts	Pedralbes	12117	248.8	41.390140	2.112218	0	3	7.0	1
1	23	5	Sarrià-Sant Gervasi	les Tres Torres	16660	215.8	41.397611	2.131184	0	3	3.0	1
2	22	5	Sarrià-Sant Gervasi	Sarrià	25106	193.6	41.399373	2.121513	2	3	1.0	1
3	25	5	Sarrià-Sant Gervasi	Sant Gervasi - Galvany	47753	192.1	41.397807	2.143377	3	3	6.0	1
4	24	5	Sarrià-Sant Gervasi	Sant Gervasi - la Bonanova	25909	184.6	41.405983	2.133405	2	3	3.0	1

5 Discussion

There is a list of observations for a future research:

- Postal codes are not very useful in Spain because you can't link it with the neighbourhoods.
- Foursquare API is not very useful in Spain because it hasn't the venues with the required accuracy.

6 Conclusion

This was a Capstone Project so the risk of a bad interpretation is low compared with the required investment to build a luxury hotel in Barcelona.

We would avoid the use of the Foursquare API for a professional job if there is another data source available.