APPLIED DATA SCIENCE CAPSTONE PROJECT

The Battle of Neighborhoods

Data

This document describes the data to be used for the applied data science capstone project.

Boštjan Keber
bostjan.keber@gmail.com
2019/02/01

Data

Districts and Neighborhoods

As we aim to compare Manhattan, New York City with Barcelona, Spain, first we have to collect data of districts and neighborhoods of the two cities.

For the city of New York, we shall use the data provided by the Applied data science course "newyork_data.json" which can be obtained from https://cocl.us/new_york_dataset. An example of a record in the dataset is:

```
{'type': 'Feature',
   'id': 'nyu_2451_34572.105',
   'geometry': {'type': 'Point',
    'coordinates': [-73.9573853935188, 40.8169344294978]},
   'geometry_name': 'geom',
   'properties': {'name': 'Manhattanville',
    'stacked': 2,
    'annoline1': 'Manhattanville',
    'annoline2': None,
    'annoline3': None,
    'annoangle': 0.0,
    'borough': 'Manhattan',
    'bbox': [-73.9573853935188,
     40.8169344294978,
     -73.9573853935188,
     40.8169344294978]}}
```

The dataset provides boroughs, neighborhoods, and their coordinates. We are only interested in the borough of Manhattan. We shall rename the borough column to "District" and add a column "District_ID" to the dataset. As there is only one Borough/District (Manhattan), all rows get District_Id = 0. In Barcelona, however, there will be several districts, each district will get its unique ID.

	District_ID	District	Neighborhood	Latitude	Longitude
0	0	Manhattan	Marble Hill	40.876551	-73.910660
1	0	Manhattan	Chinatown	40.715618	-73.994279
2	0	Manhattan	Washington Heights	40.851903	-73.936900
3	0	Manhattan	Inwood	40.867684	-73.921210
4	0	Manhattan	Hamilton Heights	40.823604	-73.949688
5	0	Manhattan	Manhattanville	40.816934	-73.957385

FIGURE 1: MANHATTAN, NYC DISTRICTS (BOROUGHS) AND NEIGHBORHOODS

The source for Barcelona districts and neighborhoods will be Wikipedia[i]. Barcelona has 10 districts and each district has several neighborhoods. By scrapping the Wikipedia page we will get the data in the format suitable for data analysis.

Number	District	Size km²	Population	Density hab/km	Neighbourhoods	Councilman	Party
1	Ciutat Vella	4.49	111,290	24,786	La Barceloneta, El Gòtic, El Raval, Sant Pere, Santa Caterina i la Ribera	Gala Pin Ferrando	Barcelona en Comú
2	Eixample	7.46	262,485	35,586	L'Antiga Esquerra de l'Eixample, La Nova Esquerra de l'Eixample, Dreta de l'Eixample, Fort Pienc, Sagrada Família, Sant Antoni	Gerardo Pisarello Prados	Barcelona en Comú
3	Sants-Montjuïc	21.35	177,636	8,321	La Bordeta, la Font de la Guatlla, Hostafrancs, la Marina de Port, la Marina del Prat Vermell, El Poble-sec, Sants, Sants-Badal, Montjuïc*, Zona Franca - Port*	Laura Pérez Castaño	Barcelona en Comú
4	Les Corts	6.08	82,588	13,584	les Corts, la Maternitat i Sant Ramon, Pedralbes	Agustí Colom Cabau	Barcelona en Comú
5	Sarrià-Sant Gervasi	20.09	140,461	6,992	El Putget i Farró, Sarrià, Sant Gervasi - la Bonanova, Sant Gervasi - Galvany, les Tres Torres, Vallvidrera, Tibidabo i les Planes	Jaume Asens Llodrà	Barcelona en Comú
6	Gràcia	4.19	120,087	28,660	Vila de Gràcia, el Camp d'en Grassot i Gràcia Nova, la Salut, el Coll, Vallcarca i els Penitents.	Eloi Badia Casas	Barcelona en Comú
7	Horta-Guinardó	11.96	169,920	14,217	El Baix Guinardó, El Guinardó, Can Baró, El Carmel, la Font d'en Fargues, Horta, la Clota, Montbau, Sant Genís dels Agudells, la Teixonera, La Vall d'Hebron.	Mercedes Vidal Lago	Barcelona en Comú
8	Nou Barris	8.04	164,981	20,520	Can Peguera, Canyelles, Ciutat Meridiana, La Guineueta, Porta, La Prosperitat, les Roquetes, Torre Baró, la Trinitat Nova, El Turó de la Peira, Vallbona, Verdum, Vilapicina i la Torre Llobeta	Janet Sanz Cid	Barcelona en Comú
9	Sant Andreu	6.56	142,598	21,737	Baró de Viver, Bon Pastor, El Congrés i els Indians, Navas, Sant Andreu de Palomar, La Sagrera i Trinitat Vella	Laia Ortiz Castellví	Barcelona en Comú
10	Sant Martí	10.80	221,029	20,466	El Besòs i el Maresme, el Clot, El Camp de l'Arpa del Clot, Diagonal Mar i el Front Marítim del Poblenou, el Parc i la Llacuna del Poblenou, Poblenou, Provençals del Poblenou, Sant Martí de Provençals, La Verneda i la Pau, la Vila Olímpica del Poblenou	Josep Maria Montaner Martorell	Barcelona en Comú

FIGURE 2: WIKI PAGE WITH BARCELONA DISTRICTS

Wiki does not provide location data (coordinates) for Barcelona neighborhoods. We have to obtain that data using GeoPy geocoder.

Finally, Barcelona scrapped and location enriched data has to be merged with the NYC data. For easier manipulation we shall persist the result of processing into a .csv file (bcn_nyc.csv).

	District_ID	District	Neighborhood	Latitude	Longitude
...	...	...	...	...	...
26	0	Manhattan	Morningside Heights	40.808000	-73.963896
27	0	Manhattan	Gramercy	40.737210	-73.981376
28	0	Manhattan	Battery Park City	40.711932	-74.016869
29	0	Manhattan	Financial District	40.707107	-74.010665
...	...	...	...	...	...
80	7	Horta-Guinardó	la Font d'en Fargues	41.425480	2.166410
81	7	Horta-Guinardó	Horta	41.438200	2.157950
82	7	Horta-Guinardó	la Clota	41.429120	2.152680
83	7	Horta-Guinardó	Montbau	41.432510	2.142300
84	7	Horta-Guinardó	Sant Genís dels Agudells	41.428020	2.133480
85	7	Horta-Guinardó	la Teixonera	41.423070	2.144820

FIGURE 3: MERGED NYC AND BARCELONA DISTRICTS AND NEIGHBORHOODS WITH COORDINATES

Foursquare Data

The next step is to identify most popular venue categories (types) for each neighborhood of NYC and Barcelona. We shall use Foursquare data for this purpose. For coordinates of each neighborhood we shall invoke Foursquare's venues/explore API which returns venues in the round of the given radius around given coordinates.

A sample Foursquare request URL:

```
url =
'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&l
l={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)
```

Below is a part of a sample response from the Foursquare API:

```
{'id': '4fa862b3e4b0ebff2f749f06',
 'name': "Harry's Italian Pizza Bar",
 'contact': {'phone': '2126081007', 'formattedPhone': '(212) 608-1007'},
 'location': {'address': '225 Murray St',
  'lat': 40.71521779064671,
  'lng': -74.01473940209351,
  'labeledLatLngs': [{'label': 'display',
    'lat': 40.71521779064671,
    'lng': -74.01473940209351}],
  'postalCode': '10282',
  'cc': 'US',
  'city': 'New York',
  'state': 'NY',
  'country': 'United States',
  'formattedAddress': ['225 Murray St',
   'New York, NY 10282',
   'United States']},
 'canonicalUrl': 'https://foursquare.com/v/harrys-italian-pizza-
bar/4fa862b3e4b0ebff2f749f06',
 'categories': [{'id': '4bf58dd8d48988d1ca941735',
   'name': 'Pizza Place',
   'pluralName': 'Pizza Places',
   'shortName': 'Pizza',
   'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/pizza_',
    'suffix': '.png'},
   'primary': True},
  {'id': '4bf58dd8d48988d110941735',
   'name': 'Italian Restaurant',
   'pluralName': 'Italian Restaurants',
   'shortName': 'Italian',
   'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/italian_',
    'suffix': '.png'}}],
 …
```

We are interested in the categories section as we will cluster neighborhoods based on frequencies of category venues of neighborhoods.

Barcelona Property Prices

As we have to propose suitable locations for 2 different offerings (premium and budget), we are interested in property prices (rent or purchase) in Barcelona. Premium coffee shop generates more revenues and can justify a more expensive location, whereas budget coffee shop has to be placed in a cheaper location. A good source of real estate price information is statista.com[ii] web page. The figure below shows an example of average real estate prices for Q2/2018 for districts of Barcelona.

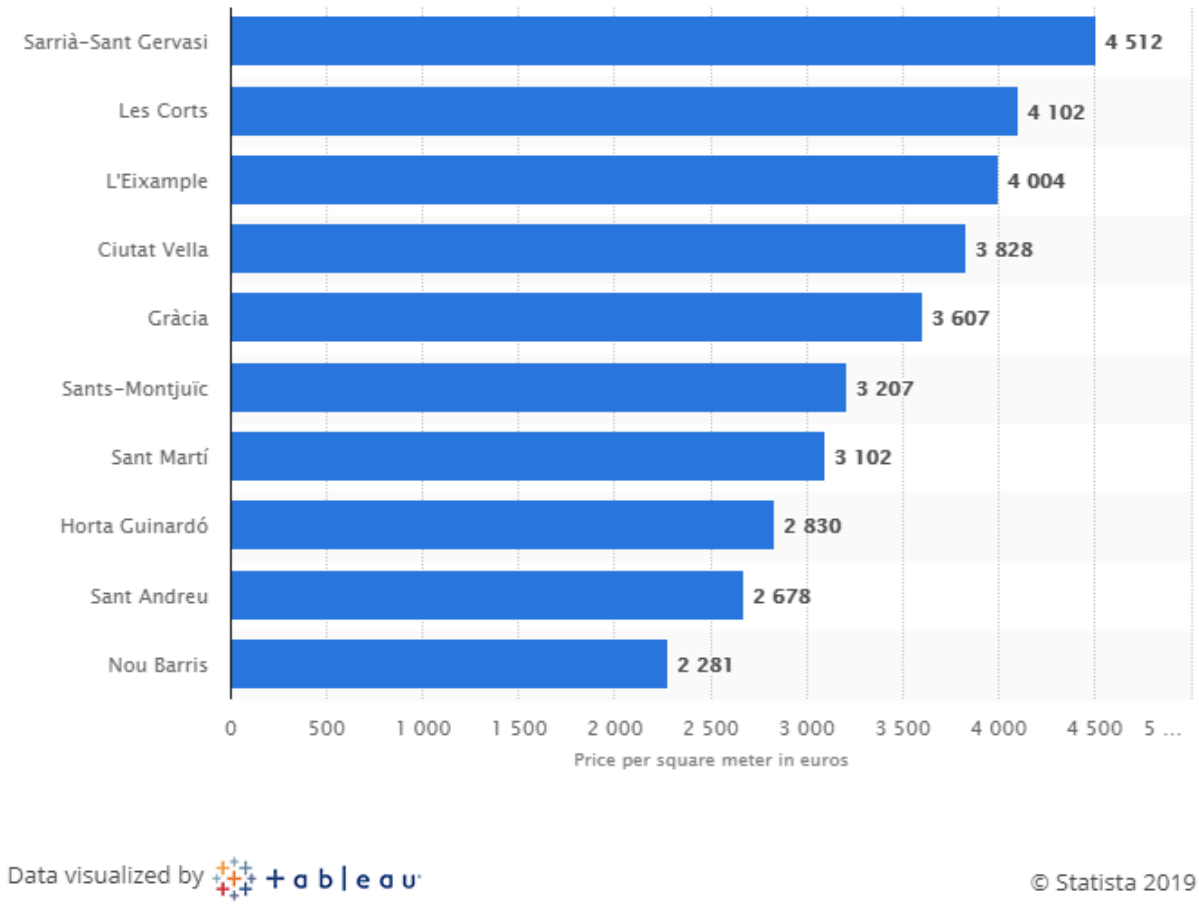

FIGURE 4: REAL ESTATE PRICES BY DISTRICTS OF BARCELONA, Q2/2018

Barcelona GeoJSON

For enhanced visualizations we will need a GeoJSON file which determines districts of Barcelona. GeoJSON is provided by Ajuntament de Barcelona / CartoBCN and can be obtained from https://cdn.rawgit.com/martgnz/bcn-geodata/master/barris/barris_geo.json.

Data Processing Pipeline

Data shall be collected, cleaned, merged and enhanced using Foursquare. Before further processing, data shall be visually inspected and verified.

Later processing stages include one-hot encoding of venue categories, grouping and extracting most common venue types by district. This data will be the input for K-means clustering which will determine a cluster label for each neighborhood of Barcelona and Manhattan. The figure below shows a data processing pipeline.

This way, we will be able to find neighborhoods of Manhattan and Barcelona with similar characteristics (i.e. same cluster label) and propose good locations for coffee shop in Barcelona based on NYC data.

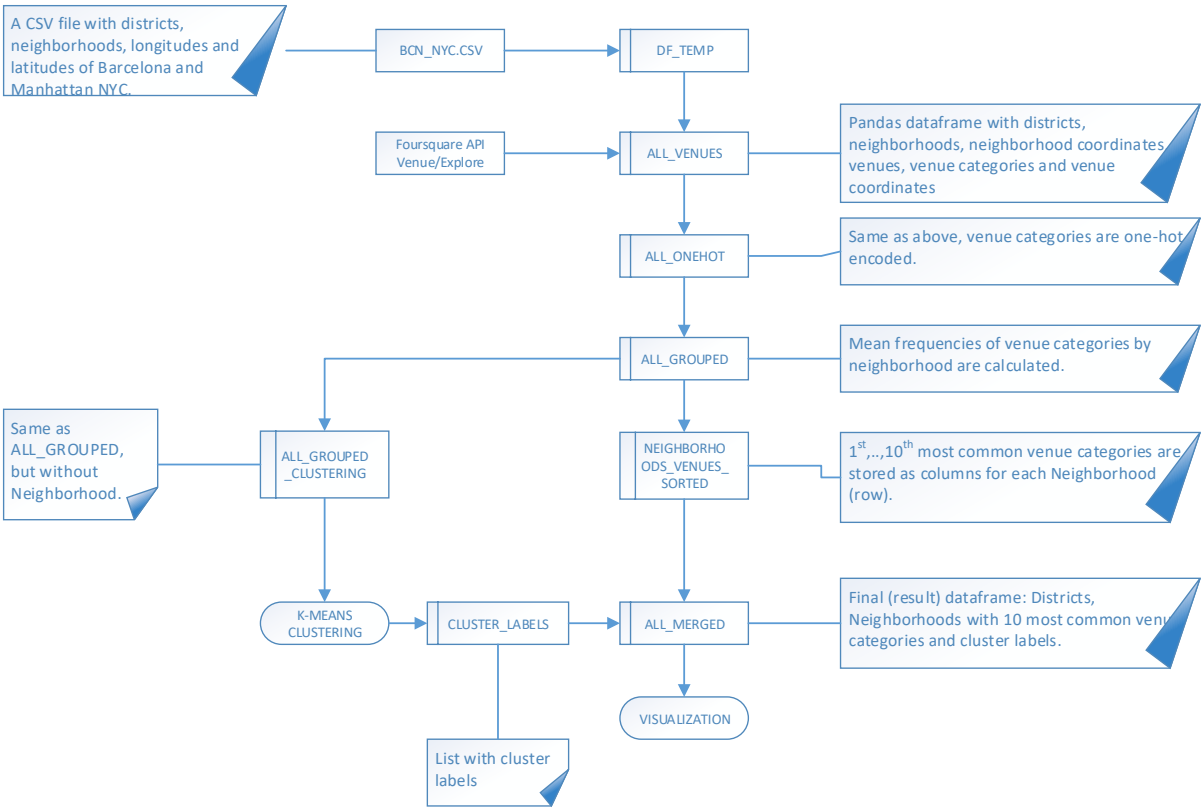

FIGURE 5: DATA PROCESSING PIPELINE

---

[i] https://en.wikipedia.org/wiki/Districts_of_Barcelona

[ii] https://www.statista.com/statistics/765380/average-price-per-square-meter-of-houses-in-barcelona-by-district/