Coursera Capstone - The Battle of the Neighbourhoods

By Gulliver Handley

Introduction

Barcelona is a city with a lot of contrasts, from a gothic centre, to modern, highrise beach living, and everything in between.
One of the things that has really marked out the city recently is large changes in the property prices, with some neighbourhoods nearby one another having very different property prices.

Business problem

The objective of the capstone project is to identify through data if there are types of venues that correlate with higher or lower property prices.
The more visibility we have into the property market, the better we will be able to make choices related to purchasing properties.

Target audience of this project

This project is designed to be useful to property investors in Barcelona, or Spain in general. It will also potentially be of interest to highlight the types of venues that correlate with higher and lower price areas to help better understand social trends.

Data
To solve this problem, we will need to access the following data sources:

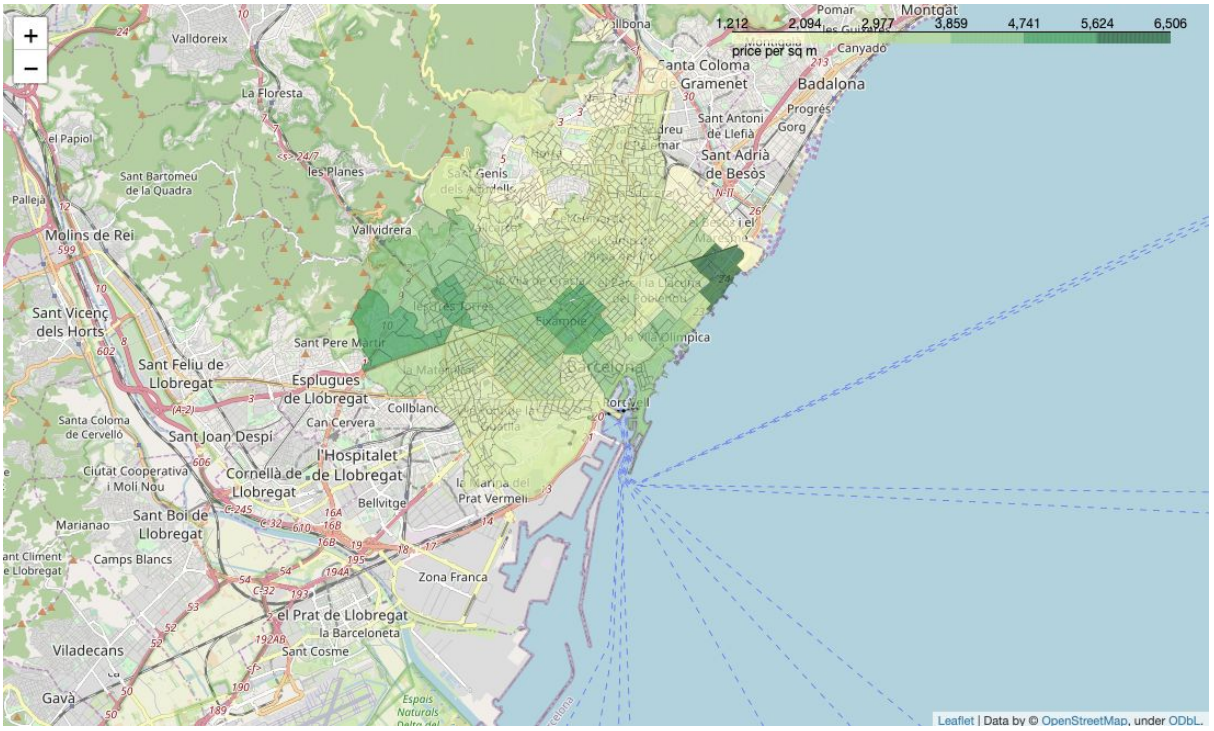
- A list of all the neighbourhoods in Barcelona
- Geospatial data, including boundary and lat/lng data to be able to perform the geospatial analysis
- A list of average property prices per square meter for all of the neighbourhoods - this is to allow us to tell higher value from lower value neighbourhoods
- Venue data across the city - to allow us to see the type of venues and their location

Sources of data:

Barcelona city council has an open data policy, and we are able to access this for the first two data sources above with a list of (sub)neighbourhoods, and a list of avg property prices in 2015.
By accessing Foursquare's API, we can get venue data in a programmatic way for the whole of the city.

Data manipulation
Barcelona is divided into districts, neighbourhoods and sub neighbourhoods, to begin, we need to identify them - Barcelona City Council provide this data at the level of sub-neighbourhood here
The council also offers average house price data from sales in 2015 at the level of neighbourhood here
Geospatial data of all of the neighbourhoods is available from here

By using geopandas, we can merge all this data together, we can also plot avg property price on an interactive map.



Map showing avg 2015 property prices per m^2 by Barcelona neighbourhood.

We want to specifically identify venues within the actual boundaries of the neighbourhoods, to do this we need to identify:

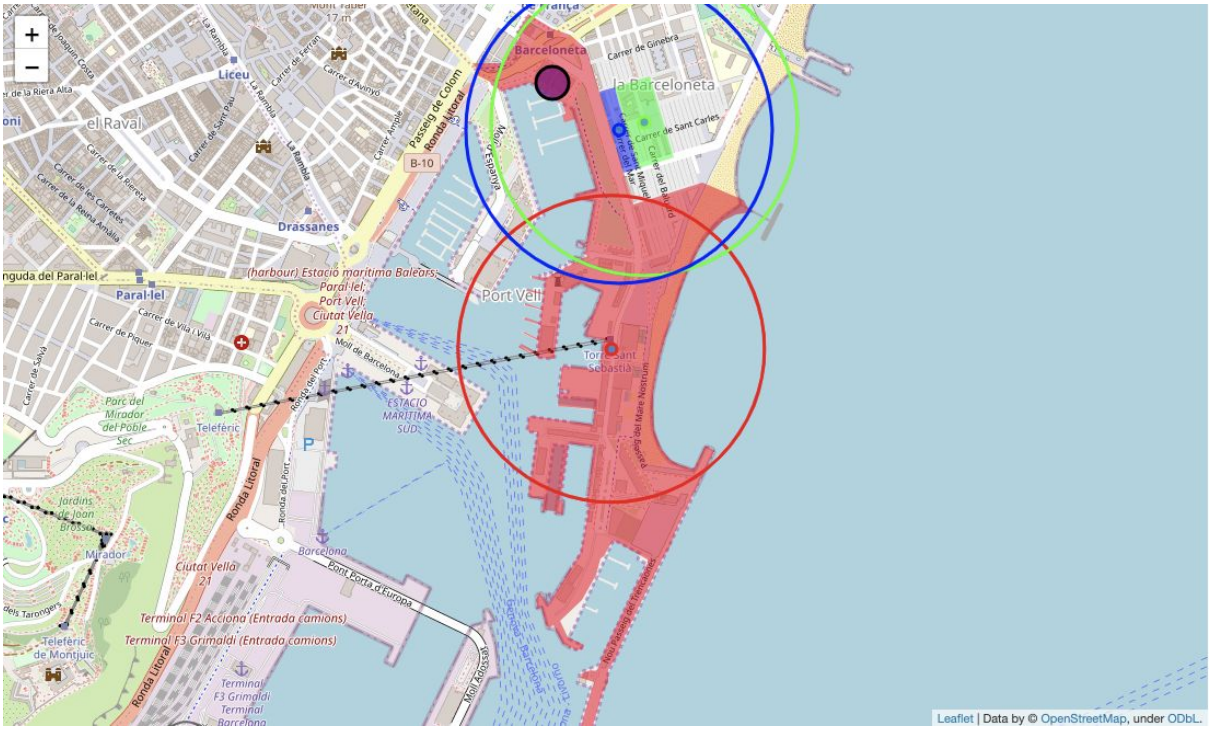
- The lat/lng of the centre of each neighbourhood
- The encompassing radius of each neighbourhood.

By using GeoPandas, we can identify the centers like this

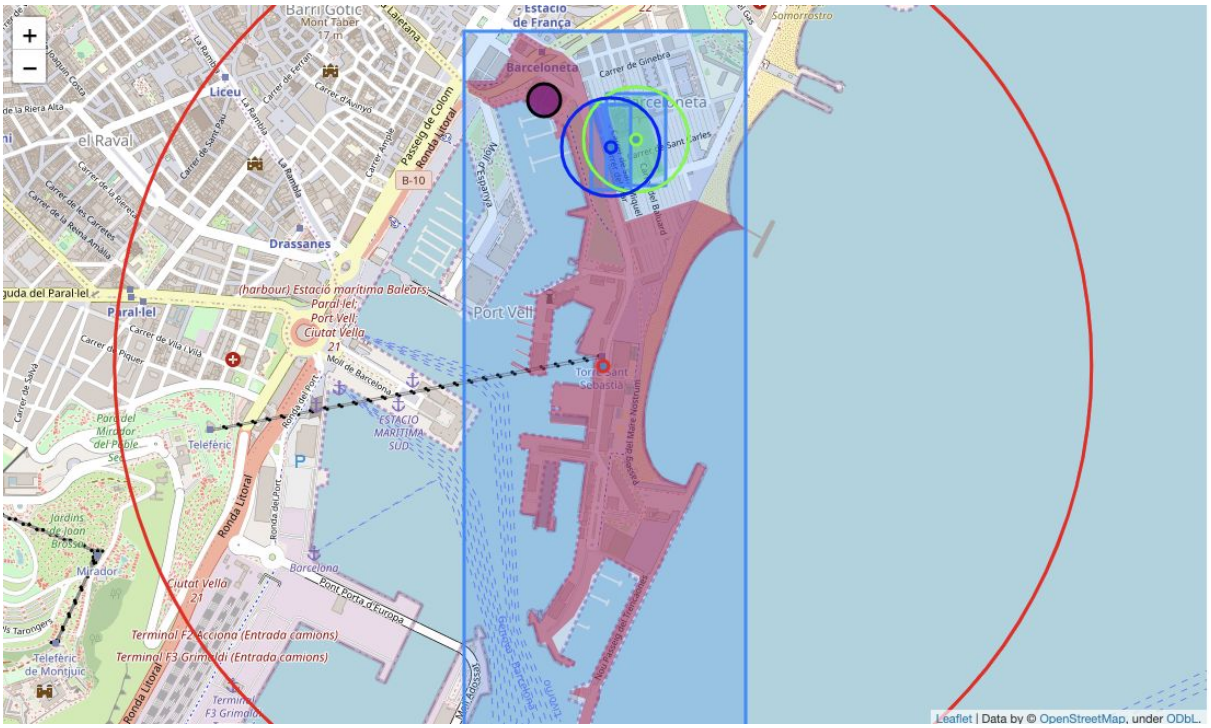
We can use the built in GeoPandas centroid functionality to calculate the centre point for each sub neighbourhood like this

```
]: df_hoods['centre_lat']=df_hoods['geometry'].centroid.y
   df_hoods['centre_lng']=df_hoods['geometry'].centroid.x
```

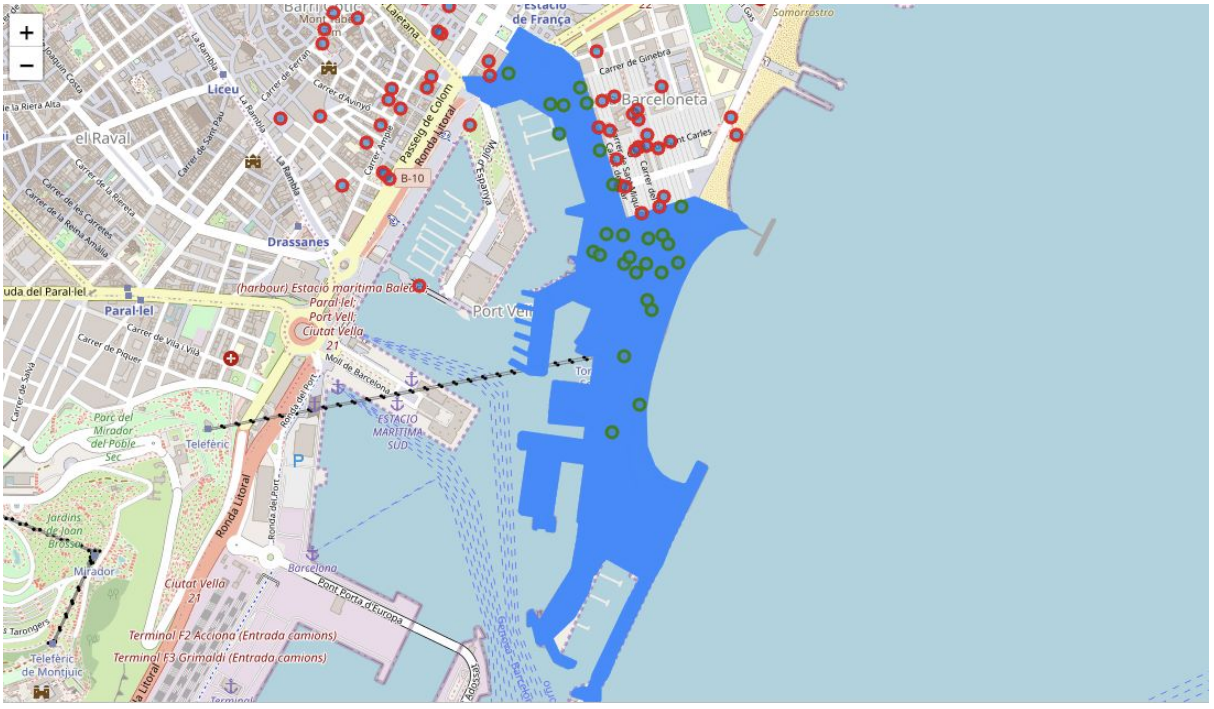
If we use the standard 500m radius around each of the centeroids, we can see that unusual shaped regions will not be captured properly - for example in the figure below, we show 3 sub neighborhoods, in red, green and blue, with their 500m radiuses around.
We can see that that the Catalan History Museum showed in the purple circle would be included in both the green and blue neighbourhood's capture areas, but not the red neighbourhood where' it's actually located.



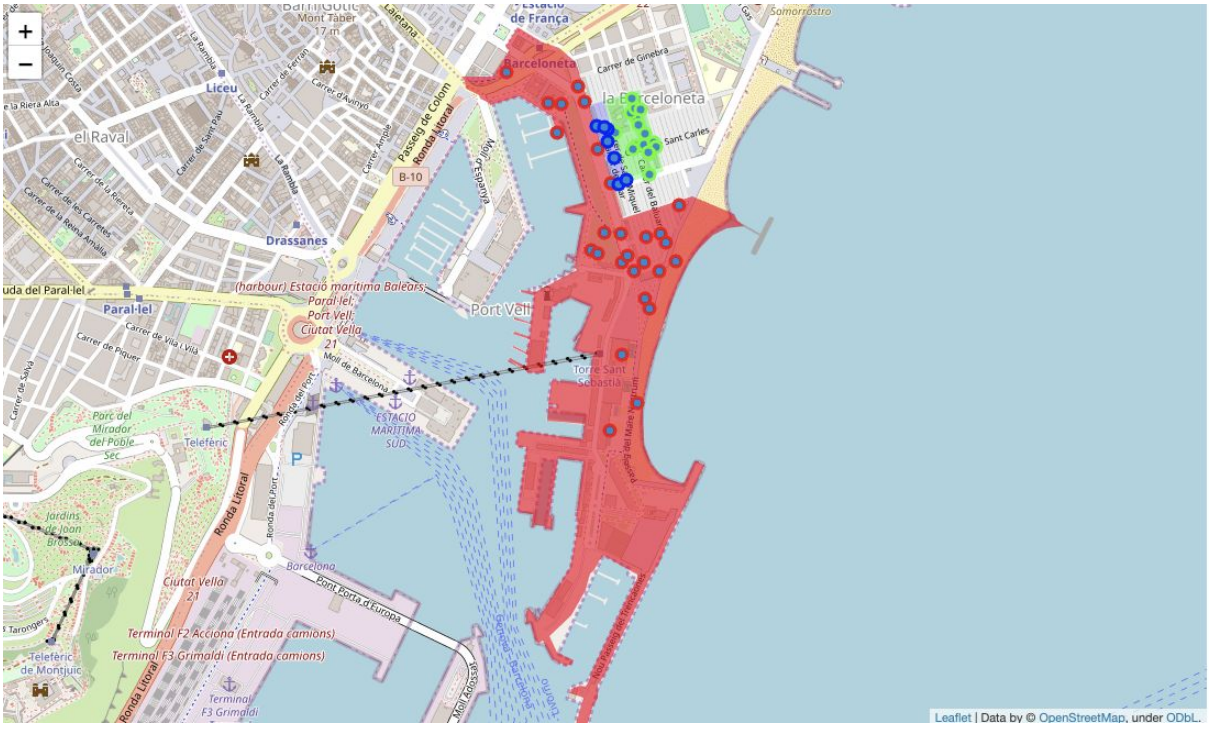
By using GeoPanda's envelope function, we can identify a rectangular bounding box around each region, and then we can use Haversine's formula to convert the coords into distances and use the longest side to calculate a radius that would enclose the rectangular - for example - in the following charts the coloured circles show the radiuses for each region.



Now we're set to start using our centre point and custom radiuses to pull in FourSquare data.
As an example, we've pulled all venues within the radius for the are highlighted in blue. Because we're using a radius, we also get a number of venues in surrounding areas, however, by using GeoPandas, we can identify these (highlighted in red)



When we put all of this together, we can ensure we can correctly identify venues directly within each region like this



Once we've run this for the whole of Barcelona, we have venue names, their location, but also the venue types - below we see an example of this data

centre_lat	centre_lng	radius	received_venues	received_venues_cnt	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	point	within
41.377801	2.180396	459.0	False	0	41.37937	2.188125	Circo Raluy	41.379887	2.182670	Circus	POINT (2.182670452780761 41.37988718883992)	True
41.377801	2.180396	459.0	False	0	41.37937	2.188125	Plaça de l'Ictineo	41.379149	2.183294	Plaza	POINT (2.183293995822424 41.37914910292008)	True
41.377801	2.180396	459.0	False	0	41.37937	2.188125	Circ Raluy Barcelona	41.380307	2.182953	Circus	POINT (2.182953301646579 41.38030659068007)	True
41.377801	2.180396	459.0	False	0	41.37937	2.188125	Terraza Sweet	41.380457	2.182927	Cocktail Bar	POINT (2.182926630324814 41.38045681860642)	True
41.386344	2.187169	649.0	False	0	41.37937	2.188125	Perikete	41.381363	2.183432	Gastropub	POINT (2.183432132005691 41.38136308403704)	True