Where to buy a house?

A real-state market analysis tool for the city of Santander.

Juan Ramón Santana.

Final Coursera Capstone project for the "IBM Data Science Professional Certificate" course.

Project description

- One of the most important decisions in a lifetime is where to buy a house in a city.
- Sometimes, deciding which area of the city where a house can be bought is difficult, as multiple parameters can have impact (e.g. number of services, location, prices, etc).

Solution proposal

- Provide a choropleth map to easily visualise the similar sections of a city based on the available public services (schools, hospitals, etc).
- To do so, we have chosen the city of Santander to perform such analysis.

Data usage

We have used two different data sources to perform this project:

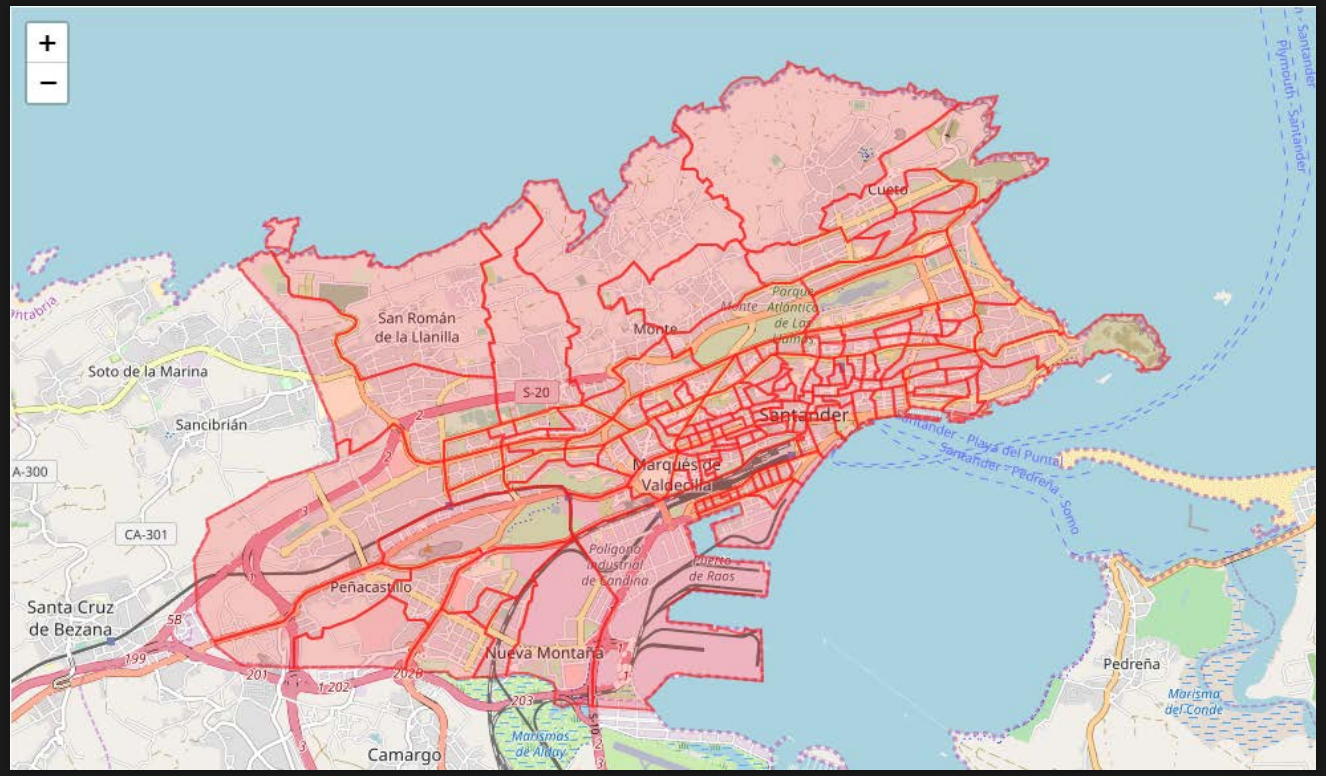
- Santander Open Data

This is a portal from the municipality of Santander with open data from the city. This portal is implemented using a CKAN platform, which guarantees a standardised data access. Thanks to this portal, we got the geographical coordinates of the polygon dividing the city.

- Foursquare API

This API provides access to a huge database including geographical information of venues (among other) from all over the world. Thanks to this API, we have obtained the existing venues for public services in the different city sections.

Santander divided in sections



Methodology followed

- Obtain the geographical data from Santander (polygons in which the city is divided) and create the dataframe. This data is obtained from the Santander Open Data portal using a REST service.
- Format the data into a usable dataframe. The main issue here is to transform the UTM coordinates to Latitude and Longitude used by the Foursquare API.
- Draw the map depending on the inhabitants in each district. To this end, we have used the folium library for python.
- Get information from foursquare about each district: hospitals, schools, etc. Additionally, data obtained is processed to extract the required parameters (e.g. number of venues per section).
- Normalize the information obtained from Foursquare to be used with machine learning techniques.
- Use a non-supervised algorithm to divide the districts depending on their similarity.
- Finally, draw a map using folium to show similarities among the different sections of the city.

Results and future work

- As we can see in the results, there are different areas of the city that are quite similar among them. Even if some of them are in the city centre and in the suburbs. At least regarding to the services provided.

- Some parts of the city center (soft green, between the orange and green) provides similar services to other parts of the city which are further, while the prices are much higher.

- It is clear that the area where the beaches are located, are represented in one single cluster (orange), even if we did not consider beaches as data input.

- This project can be enlarge using more information sources, such as more types of venues (now we are limited by foursquare free plan). Additionally, this report can be enlarge including other cities and more information sources (house prices).