Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

Downtown vs Suburbs comparison for the city of Santander

by Angel San Emeterio Herrera

Table of contents

Introduction: Business Problem

In this project our goal will be to identify which streets in the furthest areas of the city of Santander, a medium size town in the north coast of Spain, are similar to the ones belonging to innermost downtown, in terms of venues and services.

In other words, we will make a comparison between two zones of the city of Santander: most internal downtown (identify as the first postal code of the city, 39001) and most external suburbs (identify as the last postal code, 39012), in order to pinpoint which areas of the latter would be similar to the former, according to the closeness of settings and places of interest.

Given that the urban development in Spain, and more precisely in Santander, focus services and venues on the downtown, while leaving certain zones in the suburbs lacking places of interest, what we will try to do is finding out to what extent this is true for the case of Santander.

That is to say, what we actually want to know is which streets in the most exterior area of the city of Santander are similar to the ones in the very center of the town.

Data

Based on the above definition of our problem, we will need:

First of all, data on postal codes, streets and latitude and longitude about the city of Santander. To get this we will use **Santander city council's API**, located in the city open data repository.

http://datos.santander.es/documentacion-api/

http://datos.santander.es/api/rest/datasets/callejero_numpostales.json

Secondly, data on venues and their location for the city of Santander, which will be provided by using **Foursquare API**.

Methodology

Data cleaning

Data from Santander city council's open repository have been obtained from the following url:

http://datos.santander.es/api/datos/callejero_numpostales.json

We first transformed the json we obtained as response into a dataframe such as this:

	callej:sigla	callej:distrito	rdf:type	dct:spatial	gn:postalCode	callej:portal-bis	dc:modified	callej:num-portal	callej:seccion	dc:identifier	callej:nomb cl
0	CL	08	callej:Numero-postal	POINT (430592.56000000 4810310.05000000)	39011		2020-04-28T22:06:28.35Z	7	020	4728	FAUST CAVAI
1	CL	08	callej:Numero-postal	POINT (429633.68000000 4813410.42000000)	39012	B	2020-04-28T22:06:28.35Z	6	007	3411	CO QUEBR/
2	CL	08	callej:Numero-postal	POINT (434920.66000000 4814743.53000000)	39012		2020-04-28T22:06:28.35Z	162	012	1399	INE NO
3	CL	08	callej:Numero-postal	POINT (434890.16000000 4814688.66000000)	39012		2020-04-28T22:06:28.35Z	174	012	1399	INE NO
4	CL	07	callej:Numero-postal	POINT (435178.29000000 4813424.31000000)	39006		2020-04-28T22:06:28.35Z	62	012	172	FERNAN DE LOS R

This initial dataframe has been handled until we have reached a new one, with just the columns useful for our analysis:

	StreetType	District	PostalCode	Section	StreetName	Lat_UTM	Long_UTM
0	CL	08	39011	020	FAUSTINO CAVADAS	430592.56000000	4810310.05000000
1	CL	08	39012	007	COSTA QUEBRADA	429633.68000000	4813410.42000000
2	CL	08	39012	012	INES D. NOVAL	434920.66000000	4814743.53000000
3	CL	08	39012	012	INES D. NOVAL	434890.16000000	4814688.66000000
4	CL	07	39006	012	FERNANDO DE LOS RIOS	435178.29000000	4813424.31000000

However, the geographical coordinates used by Spanish government bodies are UTM (Universal Tranverse Mercator)[1], so we needed to convert them into GPS coordinates.

We did it by using the Bidirectional UTM-WGS84 converter for python, available at https://github.com/Turbo87/utm, obtaining a new dataframe with the coordinates in the desired GPS format:

[1] https://en.wikipedia.org/wiki/Universal_Transverse_Mercator_coordinate_system

	StreetType	District	PostalCode	Section	StreetName	Latitude	Longitude
0	CL	08	39011	020	FAUSTINO CAVADAS	43.442475	-3.857716
1	CL	08	39012	007	COSTA QUEBRADA	43.470299	-3.869965
2	CL	08	39012	012	INES D. NOVAL	43.482780	-3.804766
3	CL	08	39012	012	INES D. NOVAL	43.482284	-3.805136
4	CL	07	39006	012	FERNANDO DE LOS RIOS	43.470925	-3.801423

At this point, we have the full set for the city, so we filtered to obtain a new dataframe with only the first and last postal code of the town, because the city center corresponds to the first one (39001, with 34 records) and the furthest area is the last one (39012 with 322 rows).

PostalCode	StreetType	District	Section	StreetName	Latitude	Longitude
39001	34	34	34	34	34	34
39012	322	322	322	322	322	322

Provided we have duplicate streets in our dataset, because several streets appears several times, we decided to reduce each street to one ocurrence only, to make easier our analysis.

As we can see in this example, there are duplicates:

StreetName	StreetType	District	PostalCode	Section	Latitude	Longitude
ALBERICIA	3	3	3	3	3	3
ALFONSINA STORNI	1	1	1	1	1	1
ALSEDO BUSTAMANTE	1	1	1	1	1	1
ARCILLERO	1	1	1	1	1	1
ARRIBA	2	2	2	2	2	2
ARSENIO ODRIOZOLA	2	2	2	2	2	2
ASILO	2	2	2	2	2	2
ATALAYA	2	2	2	2	2	2
AURELIO RUIZ CRESPO	1	1	1	1	1	1
AUTONOMIA	5	5	5	5	5	5

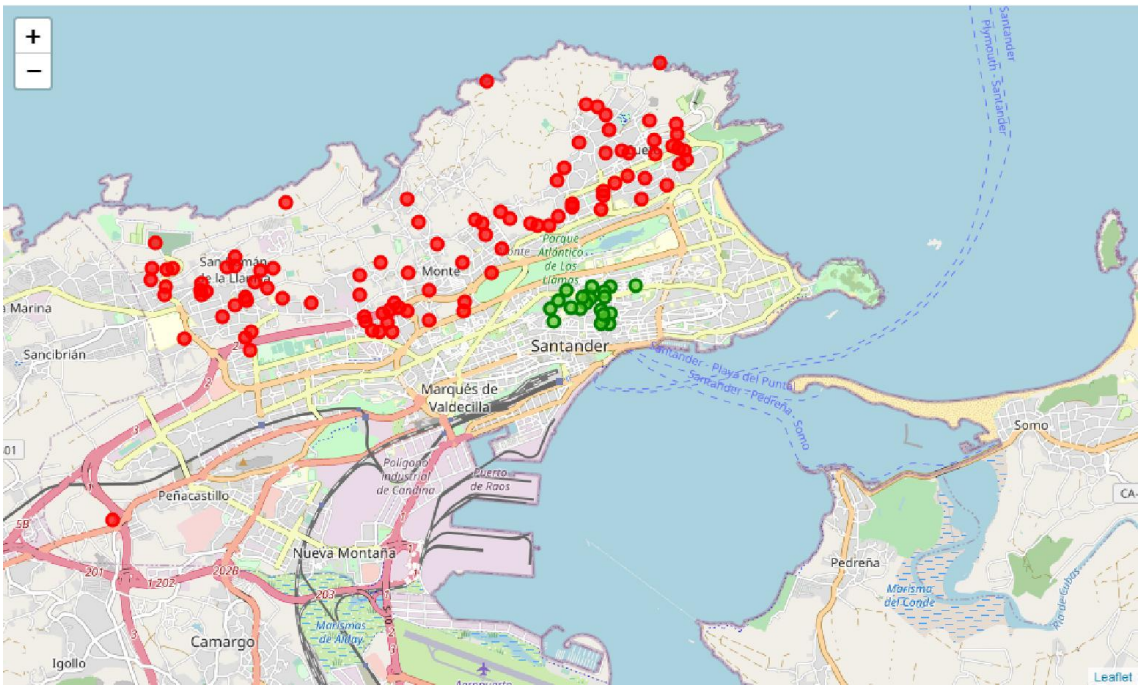
So afted this reduction, we obtain a new dataset smaller, with just one occurrence by street:

PostalCode	StreetType	District	Section	StreetName	Latitude	Longitude
39001	21	21	21	21	21	21
39012	96	96	96	96	96	96

And with this format:

	StreetType	District	PostalCode	Section	StreetName	Latitude	Longitude
0	CL	08	39012	007	COSTA QUEBRADA	43.470299	-3.869965
1	CL	08	39012	012	INES D. NOVAL	43.482780	-3.804766
2	CL	08	39012	007	CORBAN	43.467523	-3.870217
3	CL	07	39012	020	REPUENTE	43.465691	-3.834289
4	CL	07	39012	001	PRONILLO	43.464794	-3.830979
5	CL	08	39012	008	AVICHE	43.474001	-3.822505
6	CL	08	39012	024	SOMONTE	43.470577	-3.859825
7	CL	08	39012	009	CORBANERA	43.477869	-3.834161
8	CL	08	39012	024	ELENA QUIROGA	43.467943	-3.864890
9	CL	08	39012	024	MAZO DE ABAJO	43.470427	-3.861043
10	CL	08	39012	008	LA TORRE	43.475281	-3.815871

If the plot this dataset in a map, we see a clear differentiation, as expected:



In this case, green dots corresponds to postal code 39001, downtown, and red dots belongs to 39012 postal code, outer most suburbs of the town.

We then request data from Foursquare, to construct a dataframe with all the near venues for each dot in the map.

Street	Street Latitude	Street Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
ALBERICIA	5	5	5	5	5	5
ALFONSINA STORNI	6	6	6	6	6	6
ALSEDO BUSTAMANTE	66	66	66	66	66	66
ARCILLERO	57	57	57	57	57	57
ARRIBA	2	2	2	2	2	2
ARSENIO ODRIOZOLA	8	8	8	8	8	8
ASILO	35	35	35	35	35	35
ATALAYA	29	29	29	29	29	29
AURELIO RUIZ CRESPO	31	31	31	31	31	31
AUTONOMIA	8	8	8	8	8	8

Finally, such dataframe will be reduced to just the 10 top common venues per location:

	Street	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	ALBERICIA	Pizza Place	Supermarket	Restaurant	Big Box Store	Food & Drink Shop	Fast Food Restaurant	Electronics Store	Diner	Dessert Shop	Department Store
1	ALFONSINA STORNI	Park	Seafood Restaurant	Supermarket	Stadium	Bakery	Hotel	Castle	Church	Clothing Store	Cocktail Bar
2	ALSEDO BUSTAMANTE	Spanish Restaurant	Bar	Café	Tapas Restaurant	Nightclub	Plaza	Cocktail Bar	Bakery	Frozen Yogurt Shop	Restaurant
3	ARCILLERO	Café	Bar	Tapas Restaurant	Restaurant	Spanish Restaurant	Plaza	Ice Cream Shop	Wine Bar	Coffee Shop	Convention Center
4	ARRIBA	Spanish Restaurant	Italian Restaurant	Wine Bar	Clothing Store	Cocktail Bar	Coffee Shop	Concert Hall	Convenience Store	Convention Center	Department Store

Analysis

Once we have the desired dataset, we proceed to begin the analysis.

We'll make a three-clustering comparison, to analyze and find the optimal option, which'll be the one with the least streets belonging to the postal code 39012 in the same cluster as the streets of the 39001 postal code.

So we will implement three clustering options:

a) 2 clusters (provided we are comparing 2 sets of streets),

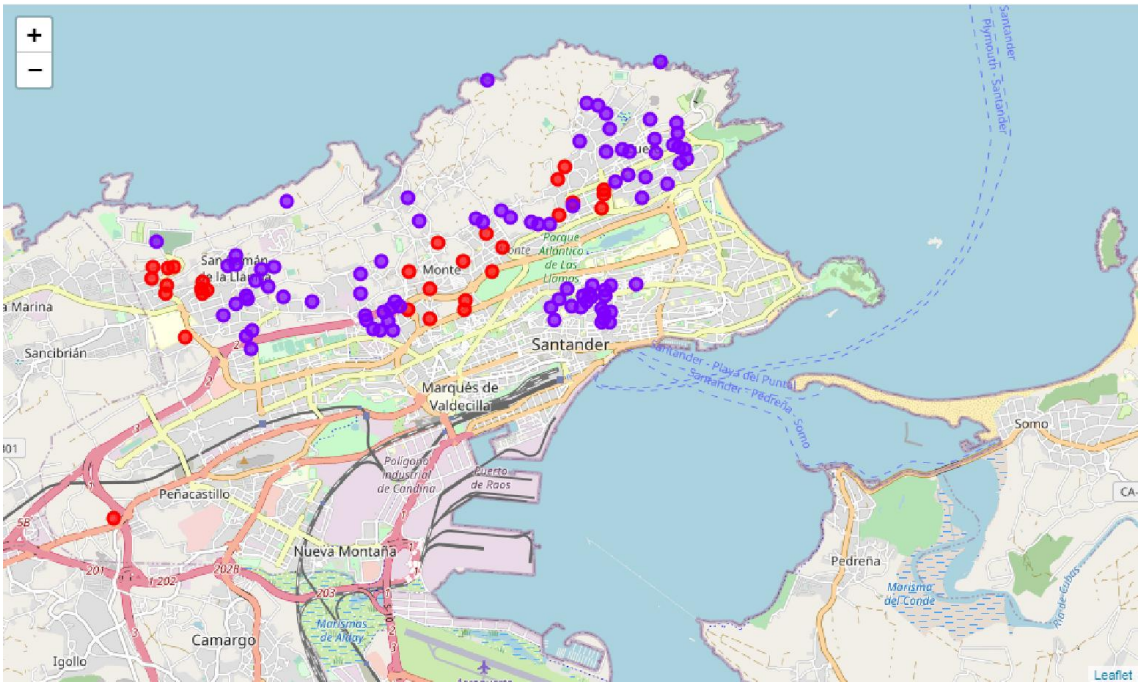
b) 5 clusters, as an intermediate solution and

c) 20 clusters, as a way of obtaining the least possible number of streets from postal code 39012 clustered with 39001 ones

Run k-means to cluster the neighborhood into 2 clusters.

Given we are comparing two areas, if we cluster our dataset into just 2 clusters, this would be the minimum possible cluster option.

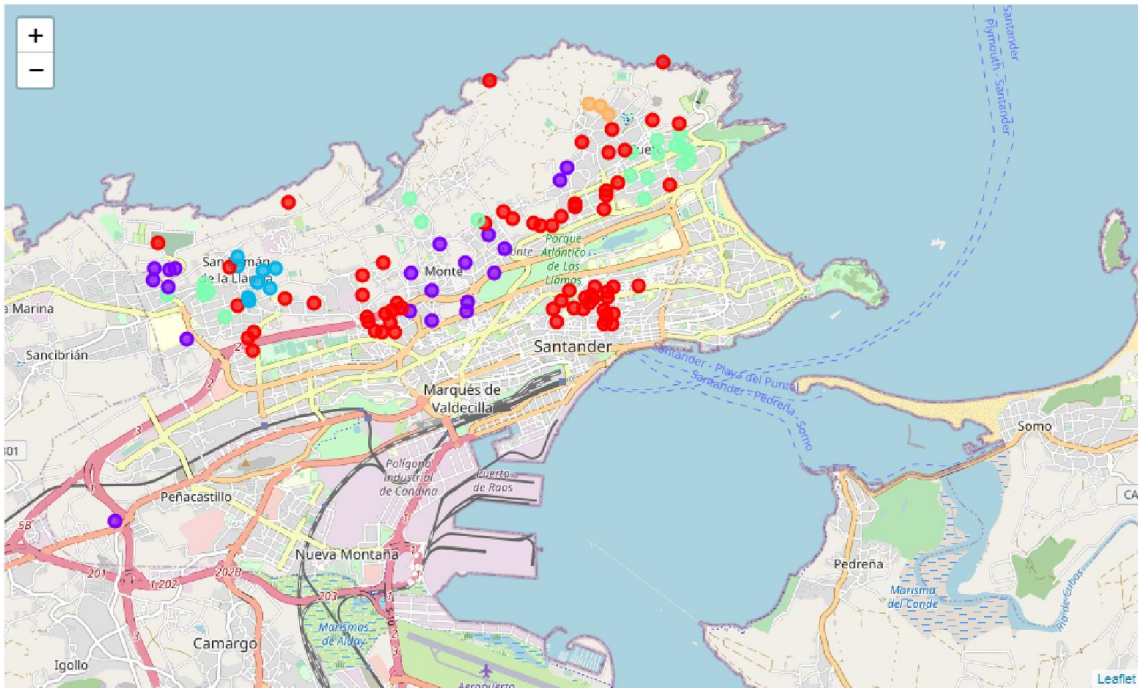
In this case, we obtain a clear separation: all the streets from postal code 39001 (21) are in one clusters and all street from postal code 39012 (66) are in the other.



PostalCode	
39001	21
39012	66

Run k-means to cluster the neighborhood into 5 clusters.

As an intermediate option, we more than double the number of clusters, and decide to explore a 5-clustered agrupation:



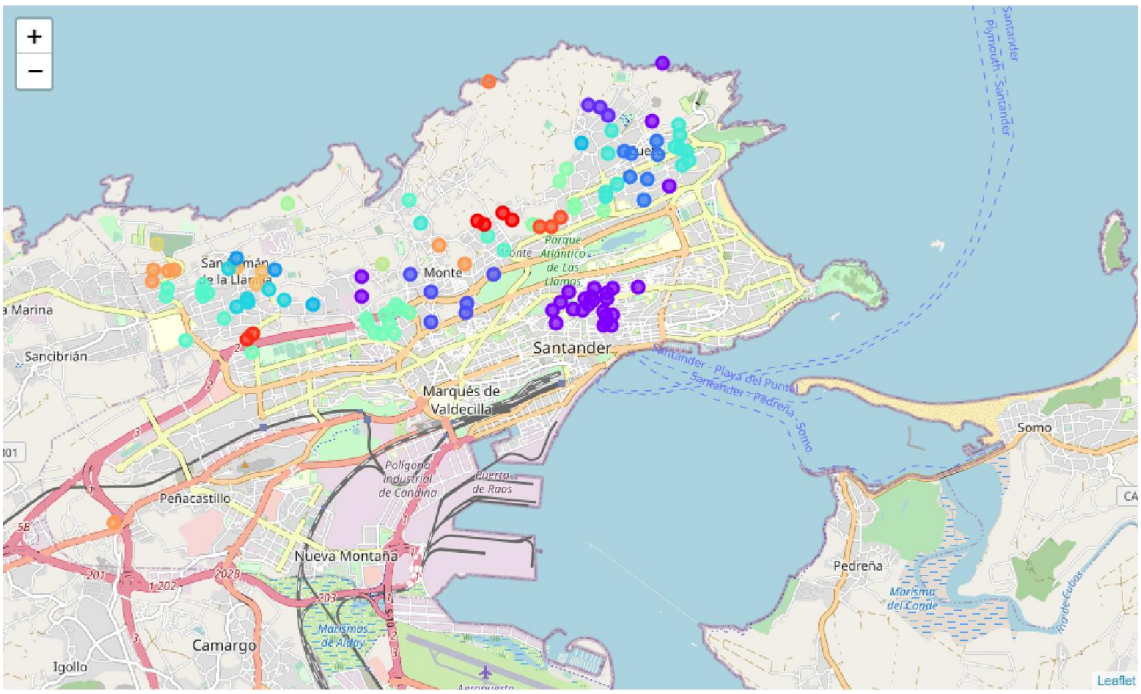
In this case, we get the 21 locations of postal code 39001 in the same agrupation that 44 streets of the postal code 39012.

PostalCode	
39001	21
39012	44

A better approximation than the starting one, but still insufficient.

Run k-means to cluster the neighborhood into 20 clusters

So we run a potential maximum approach, by choosing a 20 clustered gruoping.



This is clearly a better option, as it produces one of the clusters with the 21 streets of the 39001 postal code and just 5 locations of the 39012 postal code.

PostalCode	
39001	21
39012	5

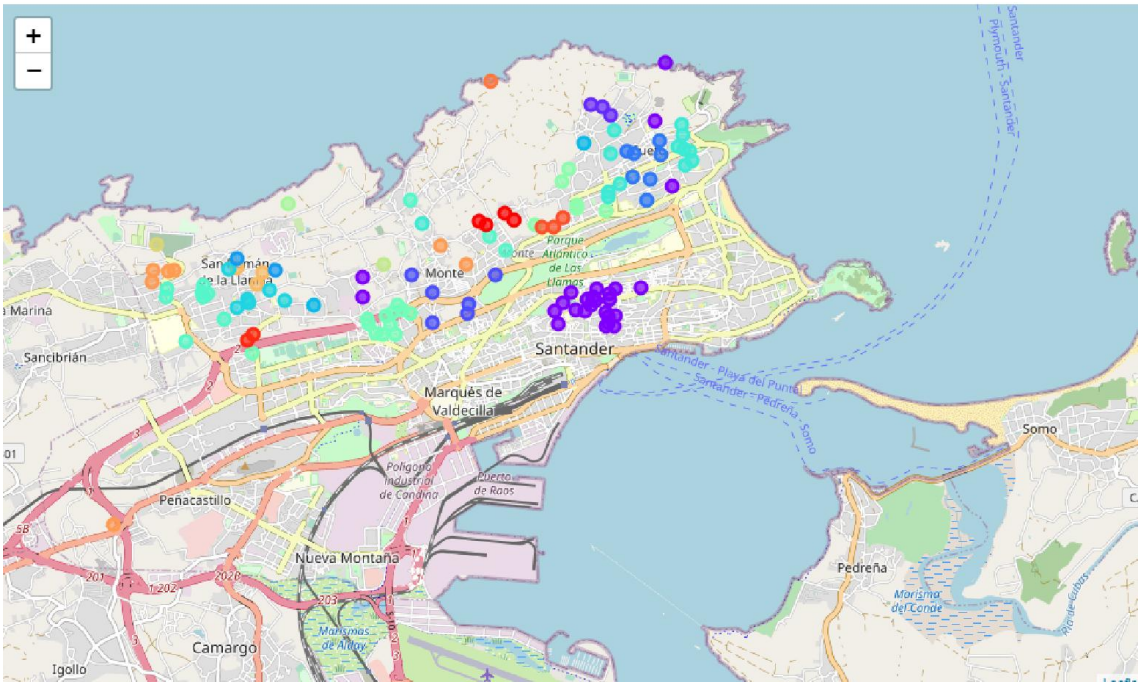
Results and Discussion

The purpose of the projects was, by using data from public sources and from Foursquare, and by applying k-means to such data, find the streets in the outer most area of the city of Santander which were similar in terms of venues and places to go to those in the downtown of the city.

After applying a triple clustering approach, we have found an aggrupation with all the streets in the city center (postal code 39001) and the least locations from the postal code 39012 (the suburbs).

In other words, as we increase the number of clusters, we see how the number of similar streets between both areas decreases.

So, by focusing in the maximum-numbered clustering, with 20 labels, we get a desired aggrupation with only 5 postal code 39012 locations:



	StreetType	District	PostalCode	Section	StreetName	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Cor
0	CL	08	39012	011	CAMUS	43.486326	-3.798123	1	Campground	BBQ Joint	Bar	Lighthouse	Snack Place	Depar
1	CL	08	39012	011	RICARDO LORENZO	43.492528	-3.796664	1	Bar	Lighthouse	Wine Bar	Department Store	Clothing Store	Cc
2	AV	04	39012	014	CANTABRIA	43.479272	-3.795595	1	Park	Bakery	Mexican Restaurant	Hotel	Gym / Fitness Center	Gi
3	CL	08	39012	005	LA GLORIA	43.467476	-3.841331	1	Bike Rental / Bike Share	Football Stadium	Coffee Shop	Boutique	Wine Bar	Cc
4	CL	08	39012	005	LOS FORAMONTANOS	43.469578	-3.841350	1	Bike Rental / Bike Share	Football Stadium	Coffee Shop	Wine Bar	Clothing Store	Cc

Conclusion

Our goal when starting this project was, by means of joining data from two main sources, the city of Santander's open repository and Foursquare, make a comparison between two areas of the city, the very center and the outer most suburb, in order to find out which places in the latter would be similar to the ones in the former in terms of venues and places to go. And, as we have already seen, our analysis have showed that there are 5 possible candidates to be considered. In other words, 5 streets in the suburbs similar to the downtown.