Capstone Project – Where do I move?

Introduction.

The objective of this analysis is to predict which neighborhood from another city will be more suitable to move based on the current one.

To do this, first I need to create a clustering model using the origin city and specific parameters of interest, and later using the data form the destination city see which neighborhoods are in the same cluster.

For this study in particular, the parameters to be consider will be the presence of Hospitals, Schools and Supermarkets, as well as the distance to the city center (because is where most of the job places are).

This analysis can be replicated with other cities and parameters, so it could be used as a feature for some real state app for example.

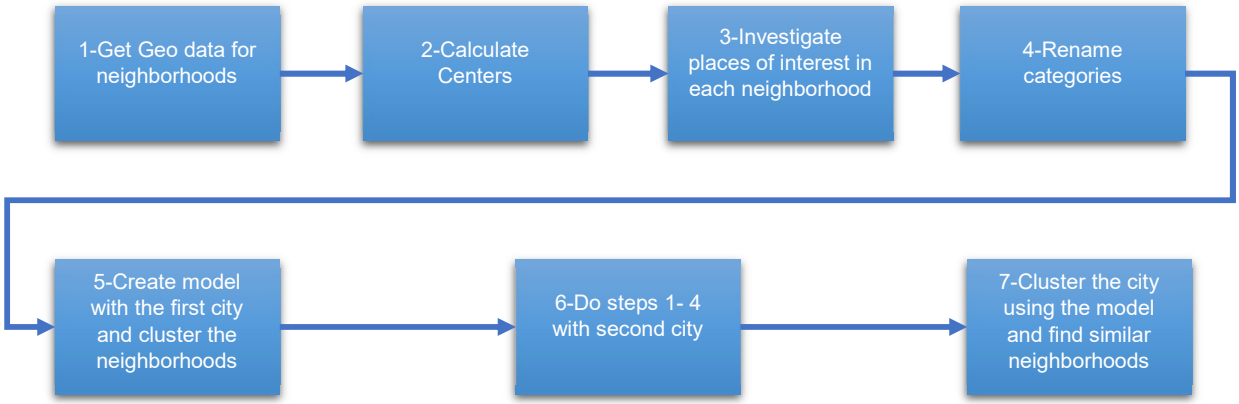
Data

The data I will used is basically the geographical data of both cities, both of which are in geoJson format.

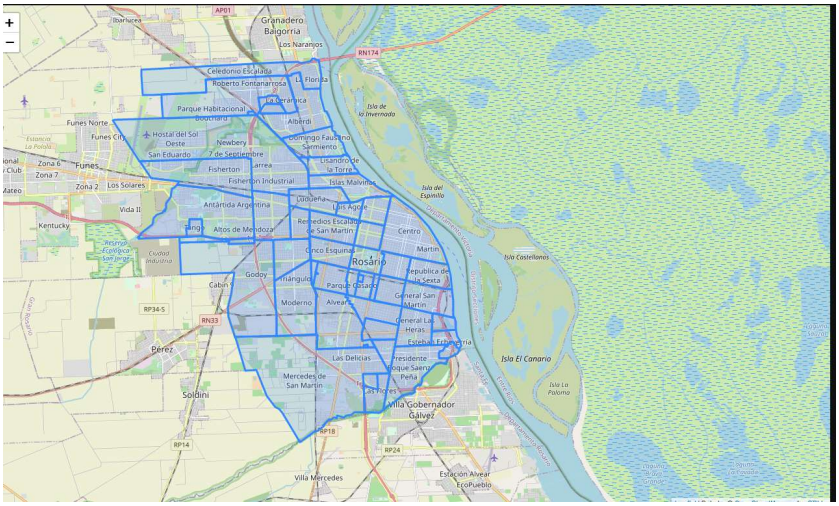
To get information about the places the Foursquare API will be used.

Methodology.

The process can be summarized as:



After completing the first step, I end with all neighborhoods for each city clearly define. For example, for the city of Rosario



Since the GeoJson data has coordinates for each point in the polygon, I can use the maximum and minimum in each coordinate and approximate a center for each neighborhood. This step is important so I can make the foursquare queries around this point and use them to calculate a Euclidian distance to the city center. With step 2 completed, now it's time to use the query function of the Foursquare API to get the "venues "information.

To do this I run a function that make the query for each category (Schools, Hospital and supermarkets) in an 800 mts Radius and store all the results in a Data Frame.

Once all the data is stored, I now have the problem that there are a lot of categories that refer to the same thing with different names. For example, the category clinic and hospital are the different things to foursquare, but to me are both "Hospital". To remedy this, I implement a function that look for all the categories containing some key words a group them in the same one.

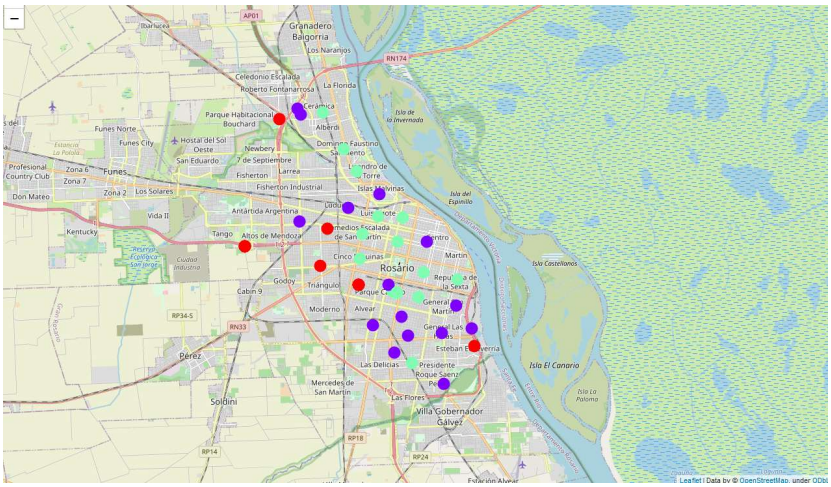
The next two images show this process

	Neighborhood	Neighborhood Latitude	Neighborhood Longit...	Venue	Venue Latitude	Venue Longitude	Venue Category
0							
1	Victoria Walsh	-32.9645365	-60.689285	Escuela Nro 1 Esteb...	-32.9595791	-60.681505400000[illegible]1	Voting Booth
2	14 de Octubre	-32.9969325	-60.66323250000001	Escuela Nro. 1160 "R...	-33.001188338115476	-60.658237257426[illegible]8	School
3	Docente "Hermanas ...	-32.8937765	-60.723695	Escuela Pascual Ech...	-32.8880775	-60.7170[illegible]6	School
4	Latinoamerica	-32.966899	-60.666099	Escuela Secundaria ...	-32.970806	-60.662[illegible]6	High School
5	Bella Vista	-32.9646215	-60.69046	Escuela Nro 1 Esteb...	-32.9595791	-60.681505400000[illegible]1	Voting Booth
6	Bella Vista	-32.9646215	-60.69046	Escuela Nro 1 Esteb...	-32.9595791	-60.681505400000[illegible]1	Voting Booth
7	Parque Casado	-32.970870000000005	-60.6587095	Escuela Secundaria ...	-32.970806	-60.662[illegible]6	High School

	Neighborhood	Neighborhood Latitude	Neighborhood Longit...	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Victoria Walsh	-32.9645365	-60.689285	Escuela Nro 1 Esteb...	-32.9595791	-60.681505400000[illegible]1	ESCUELA
1	14 de Octubre	-32.9969325	-60.66323250000001	Escuela Nro. 1160 "R...	-33.001188338115476	-60.658237257426[illegible]8	ESCUELA
2	Docente "Hermanas ...	-32.8937765	-60.723695	Escuela Pascual Ech...	-32.8880775	-60.7170[illegible]6	ESCUELA
3	Latinoamerica	-32.966899	-60.666099	Escuela Secundaria ...	-32.970806	-60.662[illegible]6	ESCUELA
4	Bella Vista	-32.9646215	-60.69046	Escuela Nro 1 Esteb...	-32.9595791	-60.681505400000[illegible]1	ESCUELA
5	Parque Casado	-32.970870000000005	-60.6587095	Escuela Secundaria ...	-32.970806	-60.662[illegible]6	ESCUELA
6	José Ignacio Rucci	-32.88917	-60.7140695	Escuela Pascual Ech...	-32.8880775	-60.7170[illegible]6	ESCUELA
7	Parque Field	-32.891694	-60.712544	Escuela Pascual Ech...	-32.8880775	-60.7170[illegible]6	ESCUELA
8	Parque Casado	-32.968704	-60.67013	Escuela Secundaria ...	-32.970806	-60.662[illegible]6	ESCUELA

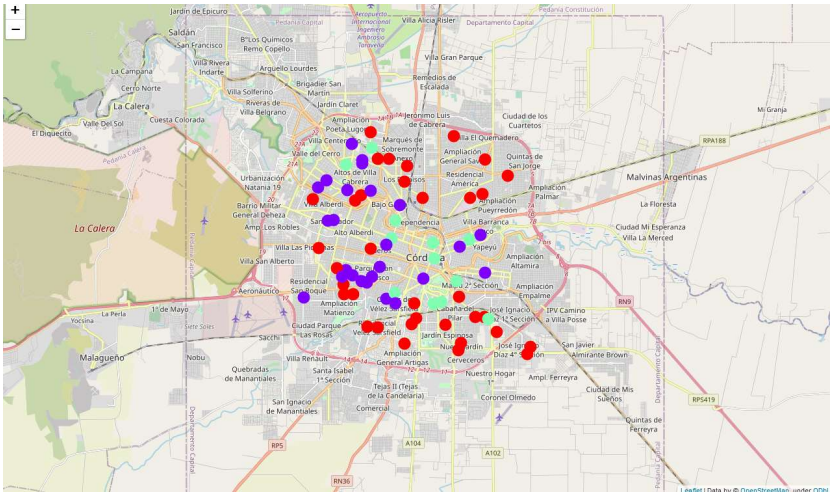
Next, it’s time to train a clustering model, I picked Kean because I can see the number of clusters that give me the least error.

For the city of rosario I got:



As a note, I currently live in a “light green” cluster

So now that I have a clustering model, already trained y repeats all the steps above with the destination city, in this case Cordoba, Argentina to get which neighborhoods there belong to the same cluster that the one I’m currently living.



Results and Discussion

As shown in the images above, I successfully cluster both cities, getting at least a starting point to decide where to move. The clustering clearly shows different areas in a pseudo concentric form, what makes sense since the distance to the city center is a parameter for clustering.

It can be argued that the simplification made by grouping categories could be harmful since the clustering is less specific. I agree with this, but it should be noted that by using the API in two different places, I cannot ensure that I will have the same number of categories to run the model. I learn this the hard way when I tried to run the model without grouping the categories.

Also, if it were available I could have used some pricing information, but unfortunately is not possible to get this kind of information for the cities I choose.

Conclusion

The main challenge of the project was to process all the data I got from various sources, since they don’t follow the same pattern.

This project gave me a pseudo-"real life" experience when working with data, starting with an idea and morph it into something concrete.

Furthermore, I learn a lot about Json, in particular GeoJson and all of its uses, and I think this is a really useful skill for Data Science