


- 
- **Applied Data Science Capstone**
  - March 2021
  - Pedro Reyes O.

# Exploring Best locations for Bakery store in Santiago, Chile



# Introduction

- Santiago de Chile, is the capital and largest city of Chile as well as one of the largest cities in the Americas. It is the center of Chile's most densely populated region, the Santiago Metropolitan Region, whose total population is 7 million, of which more than 6 million live in the city's continuous urban area.
- Commercial activity is one of the drivers of economic growth and has generated many investors wanting to compete.
- Location, location, location is the key for success





# The Problem

---

Small investors and innovators want to explore specialized bakery chains whose requirements are for convenience and differentiating products

---

The locations must combine an adequate balance between the rental cost per square meter, that there are complementary attractive places and that there is a minimum volume of population residing in the vicinity

# Target Audience

## Interest

- Business personnel who wants to invest or open a Bakery store
- People looking for profitable business alternatives with the possibility of self-employment
- Finding the best location for opening a commercial store

It is interesting to combine geographic analysis with data from commercial places valued by people and other data that could complement the analysis of localities.

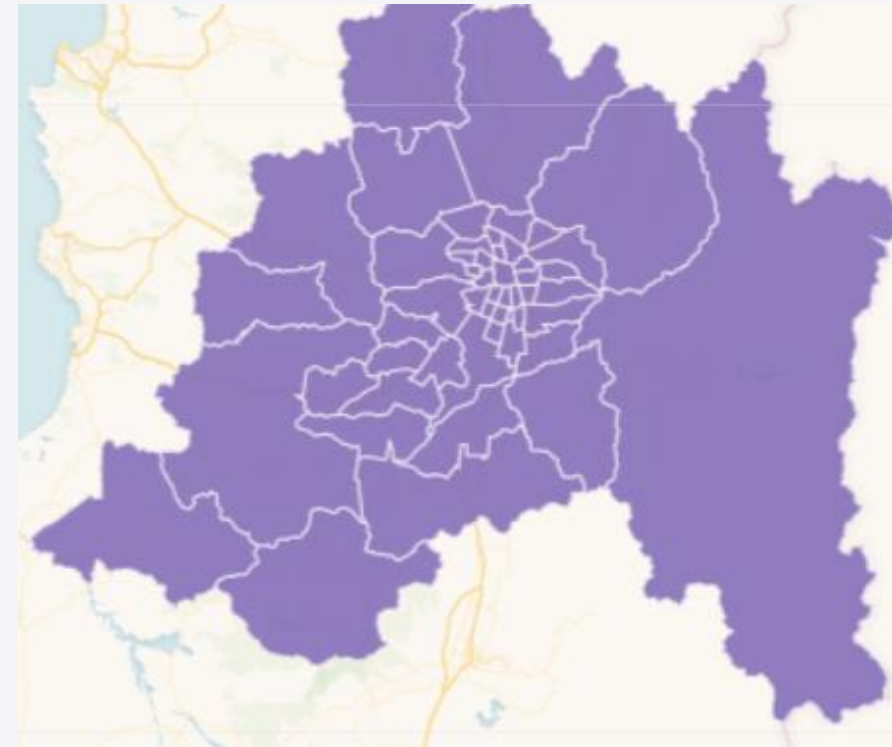
# Data acquisition and cleaning

The following sources of information are identified that will be helpful in addressing the problem:

- Latitude, longitude, Surface (km<sup>2</sup>), Population, 2020 Density (hab./km<sup>2</sup>) of districts in Santiago: [https://es.wikipedia.org/wiki/Anexo:Comunas\\_de\\_Chile](https://es.wikipedia.org/wiki/Anexo:Comunas_de_Chile)
- Lease values per square meter in districts in Santiago: <https://www.buenainversion.cl/blog/valor-metro-cuadrado/>
- Venues in Santiago: [https://api.foursquare.com/v2/venues/explore?&client\\_id={}&client\\_secret={}&v={}&ll={},{}&radius={}&limit={}](https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={})
- GeoJson file for districts in Santiago: [https://cephei.carto.com/tables/comunas\\_santiago/public](https://cephei.carto.com/tables/comunas_santiago/public)

The processes necessary to obtain the data and its preparation include:

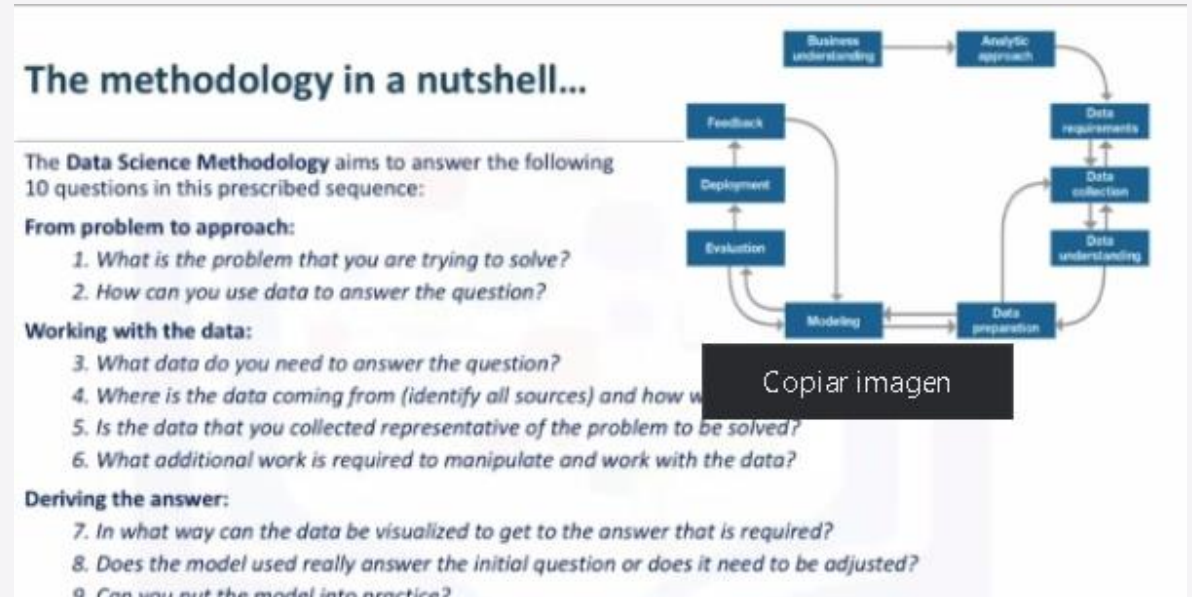
- web scrapping, direct loading, elimination of duplicates, homologate names of districts, join bases and normalize (standardize) the data before starting machine learning processes.





# Methodology





- Business understanding defining the problem and its utility for the stakeholder
- Data and processing.
- Choose analytical approach: in this case the data will be explored through bar graphs, heat maps and choropleths.
- Combining the data and normalizing it, an unsupervised learning technique will be applied to obtain clusters (Kmeans) and visualize them through folium.
- The results will be presented graphically and tabularly with the necessary detail to be able to replicate them,
- Finally, the conclusions and discussions will be analyzed, its scope and possible improvements will be seen.



# First views of data

- Latitude y longitude must be to converted to decimals
- Debugging processes: elimination of columns, null values, correct formats, etc.

<title>Anexo:Comunas de Chile - Wikipedia, la enciclopedia libre</title>

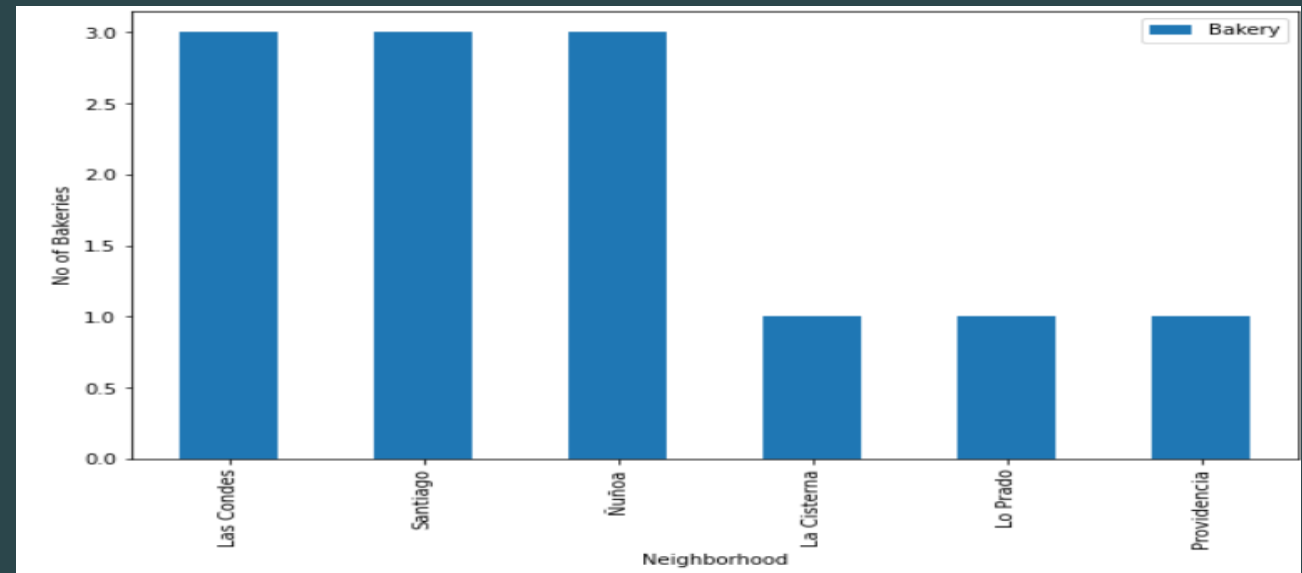
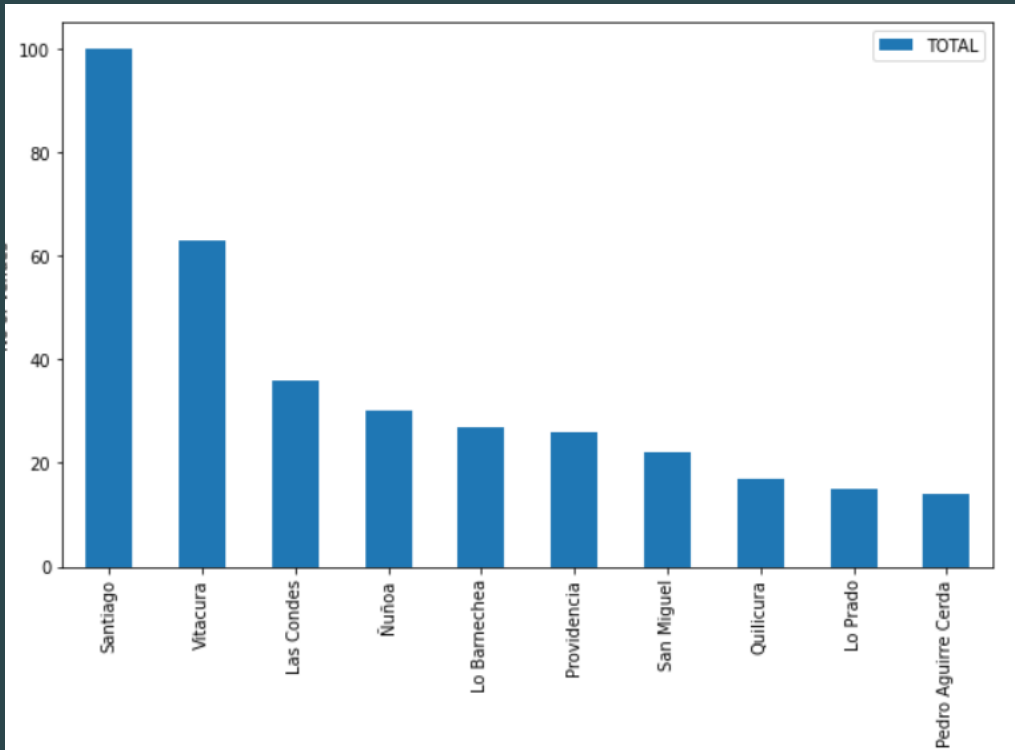
CUT (Código Único Territorial)	Nombre	Provincia	Región	Superficie (km²)	Población 2020	Densidad (hab./km²)	IDH 2005	Latitud	Longitud	
15101	<a href="#">Arica</a>	<a href="#">Arica</a>	 <a href="#">Arica y Parinacota</a>	4.799,4	247.552	51.6	0.866	Alto	-18°27'18"	-70°17'24"
15102	<a href="#">Camarones</a>	<a href="#">Arica</a>	 <a href="#">Arica y Parinacota</a>	3.927	1.233	0.31	0.791	Alto	-19°1'1.2"	-69°52'1.2"
15201	<a href="#">Putre</a>	<a href="#">Parinacota</a>	 <a href="#">Arica y Parinacota</a>	5.902,5	2.515	0.43	0.817	Alto	-18°12'0"	-69°34'58.8"
15202	<a href="#">General Lagos</a>	<a href="#">Parinacota</a>	 <a href="#">Arica y Parinacota</a>	2.244,4	810	0.36	0.773	Medio	-17°39'10.8"	-69°38'6"

Nombre	Unnamed: 2	Provincia	Región	Superficie(km²)	Población2020	Densidad(hab./km²)	IDH 2005	IDH 2005.1	Latitud	Longitud
Santiago	NaN	Santiago	Metropolitana de Santiago	23.2000	503.147	21.8759	0.807	Muy alto	-33.437222	-70.657222
Cerrillos	NaN	Santiago	Metropolitana de Santiago	21.0000	88.956	4.2360	0.743	Alto	-33.500000	-70.716667
Cerro Navia	NaN	Santiago	Metropolitana de Santiago	11.0000	142.465	12.9513	0.683	Medio	-33.422000	-70.735000
Conchalí	NaN	Santiago	Metropolitana de Santiago	10.7000	139.195	12.6540	0.707	Alto	-33.380000	-70.675000

# Looking at data from Foursquare

- 473 venues obtained
- Only 12 venues in Bakery category
- Only 7 districts has more than 20 venues
- → Work with ALL CATEGORIES and not only Bakery

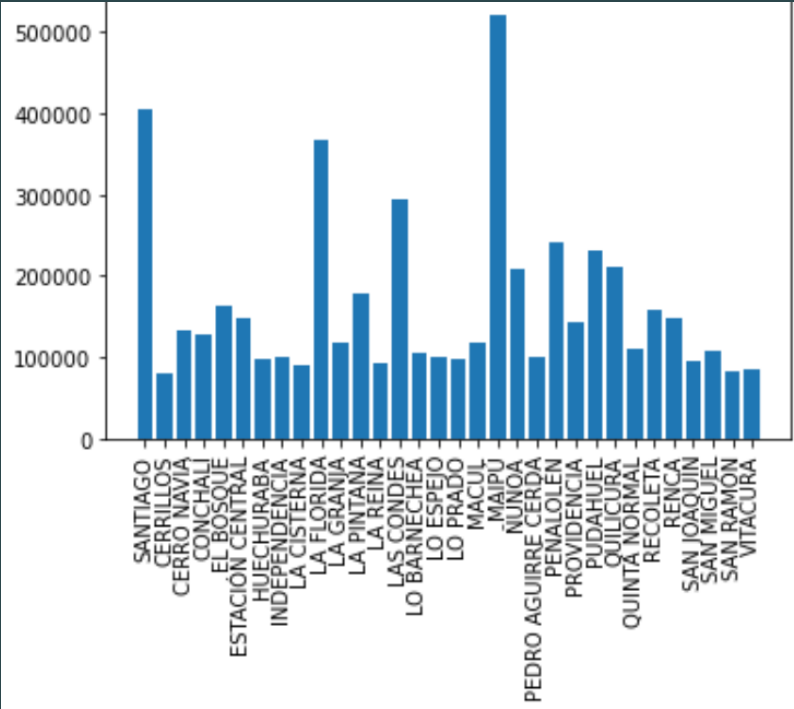
	Neighbourhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Santiago	-33.437222	-70.657222	Plaza de Bolsillo - Santiago Centro	-33.436778	-70.655481	Plaza
1	Santiago	-33.437222	-70.657222	Starbucks	-33.437938	-70.657007	Coffee Shop
2	Santiago	-33.437222	-70.657222	Bambudda	-33.438987	-70.655631	Asian Restaurant
3	Santiago	-33.437222	-70.657222	Amanda's	-33.439206	-70.658247	Arepa Restaurant
4	Santiago	-33.437222	-70.657222	YMCA	-33.439060	-70.656257	Pool
...	...	...	...	...	...	...	...
468	Vitacura	-33.400000	-70.600000	Hotel Director	-33.402812	-70.595808	Hotel
469	Vitacura	-33.400000	-70.600000	Louis Vuitton	-33.401662	-70.595735	Boutique
470	Vitacura	-33.400000	-70.600000	Dap Ducasse	-33.403355	-70.598241	Furniture / Home Store
471	Vitacura	-33.400000	-70.600000	Salcobrand	-33.398748	-70.597841	Pharmacy



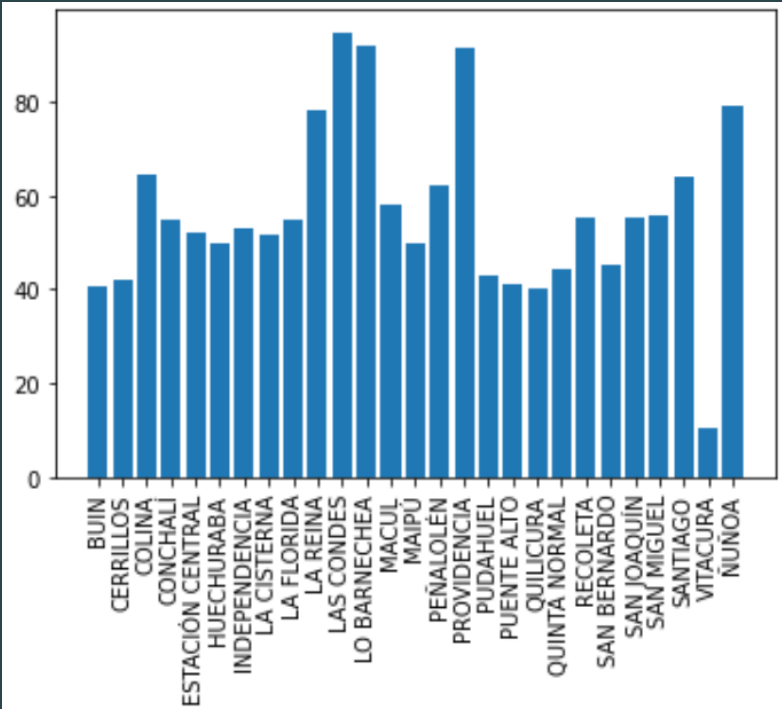


# Exploring Polpulation, density and value of the land...

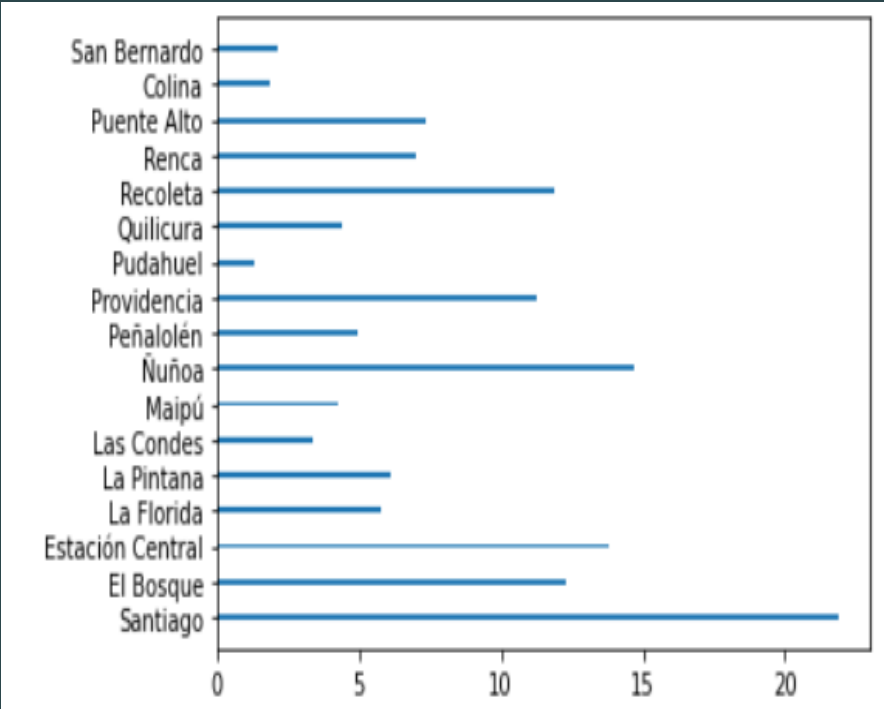
Population



Value UF/ m^2 lease

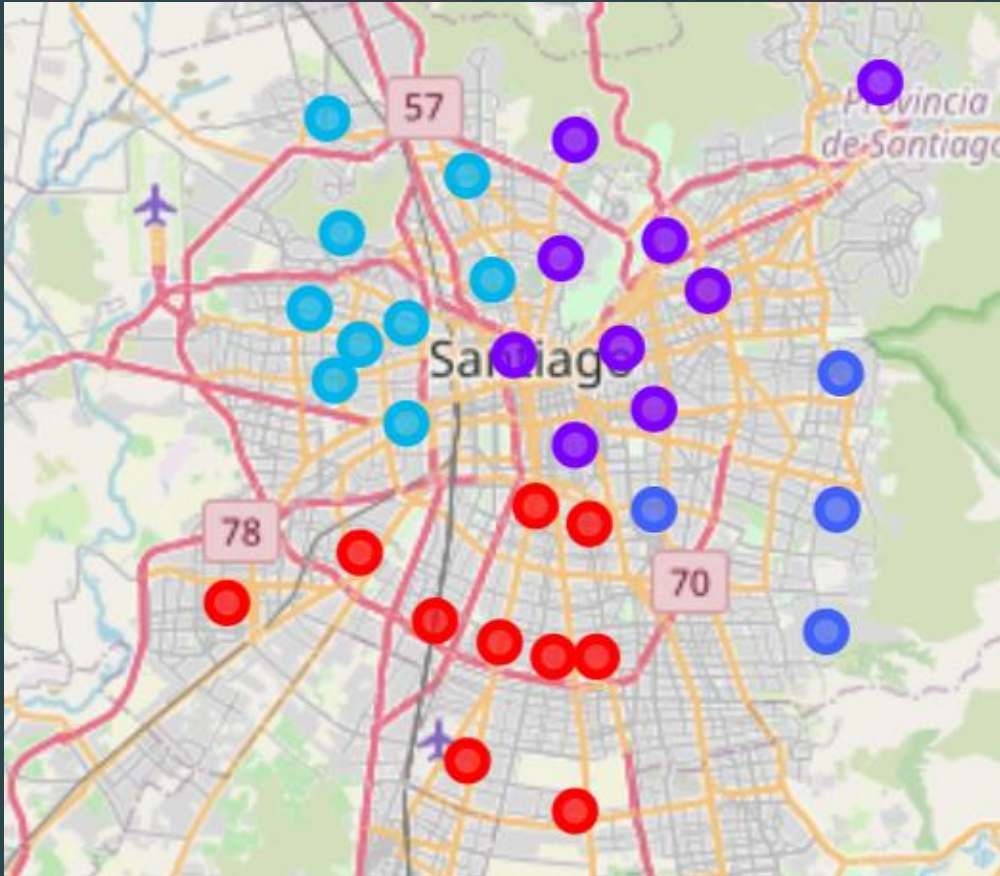


Density :Population / Km^2

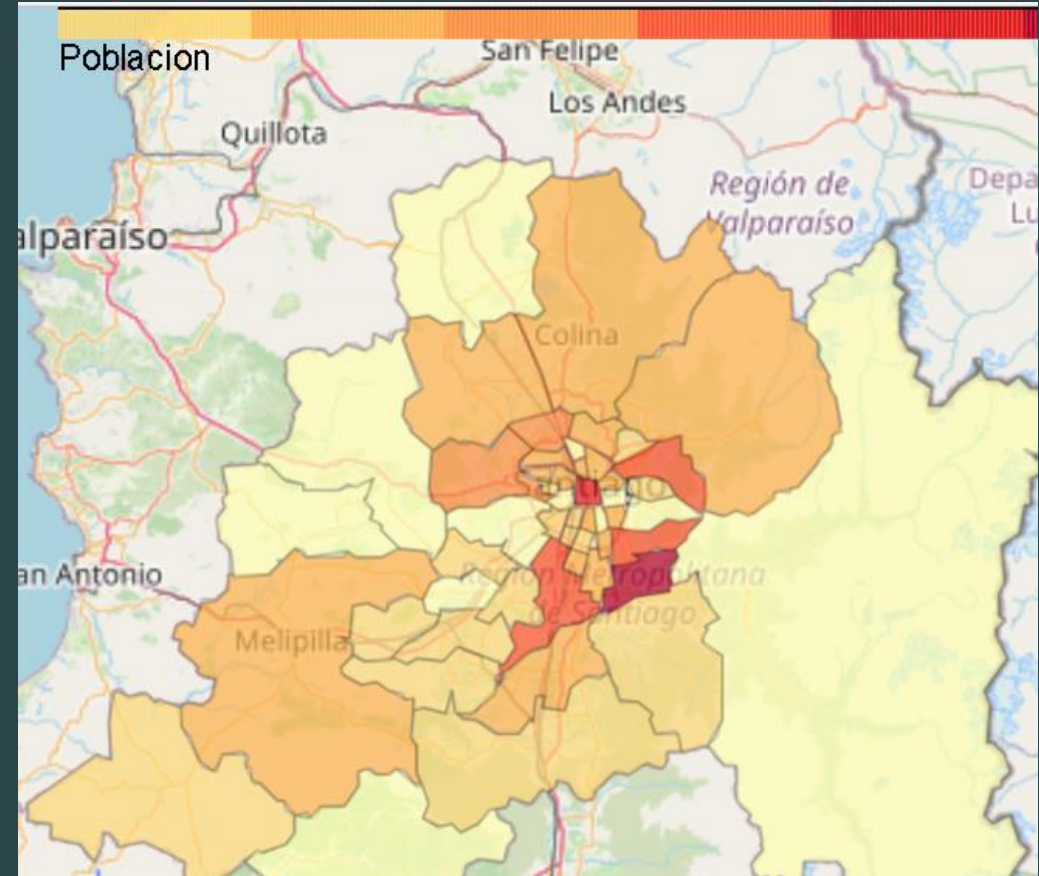


# Putting the data in maps

Geographic clusters, K=4



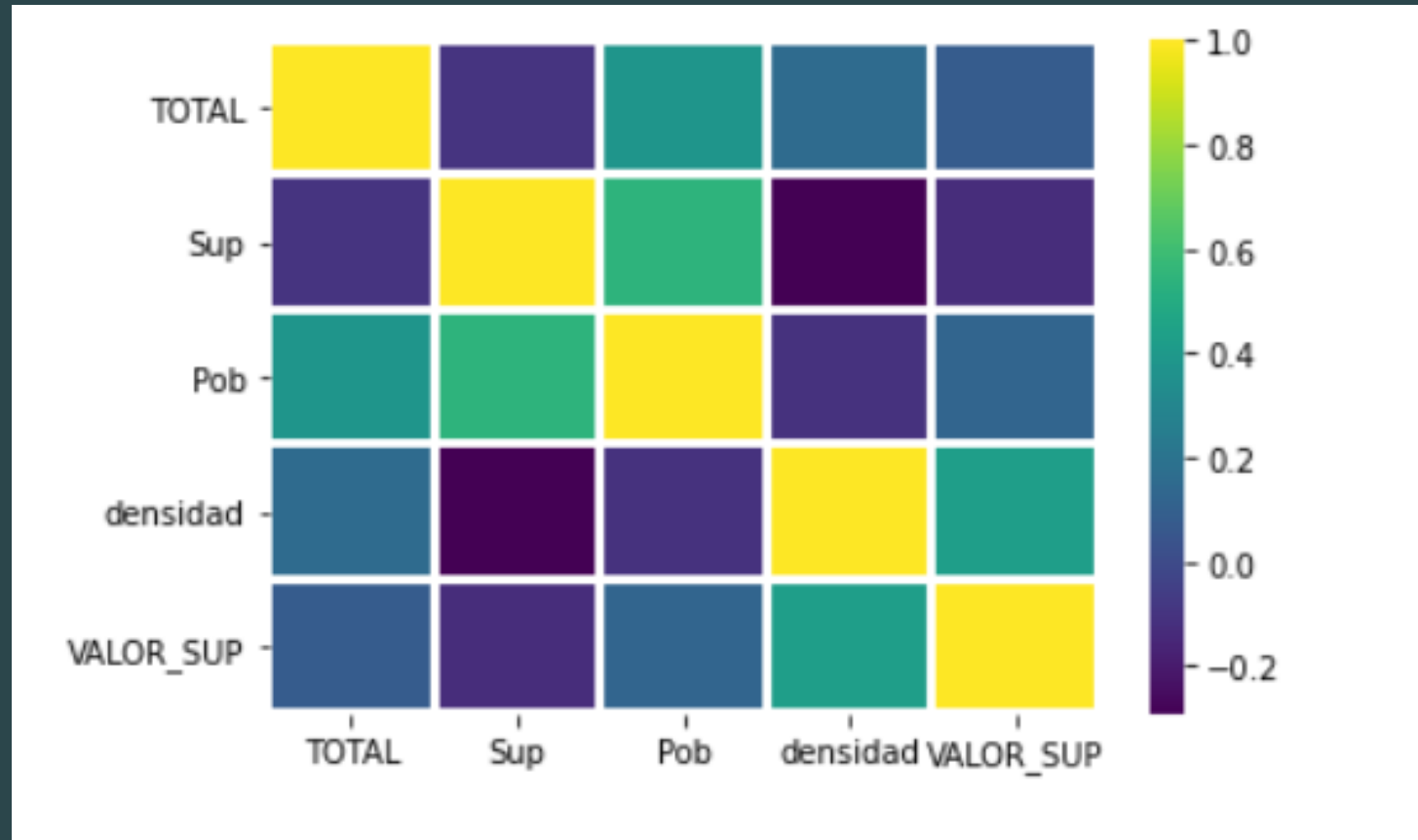
Choropleth, population 2017



- Use of folium library and GeoJson for choropleth maps

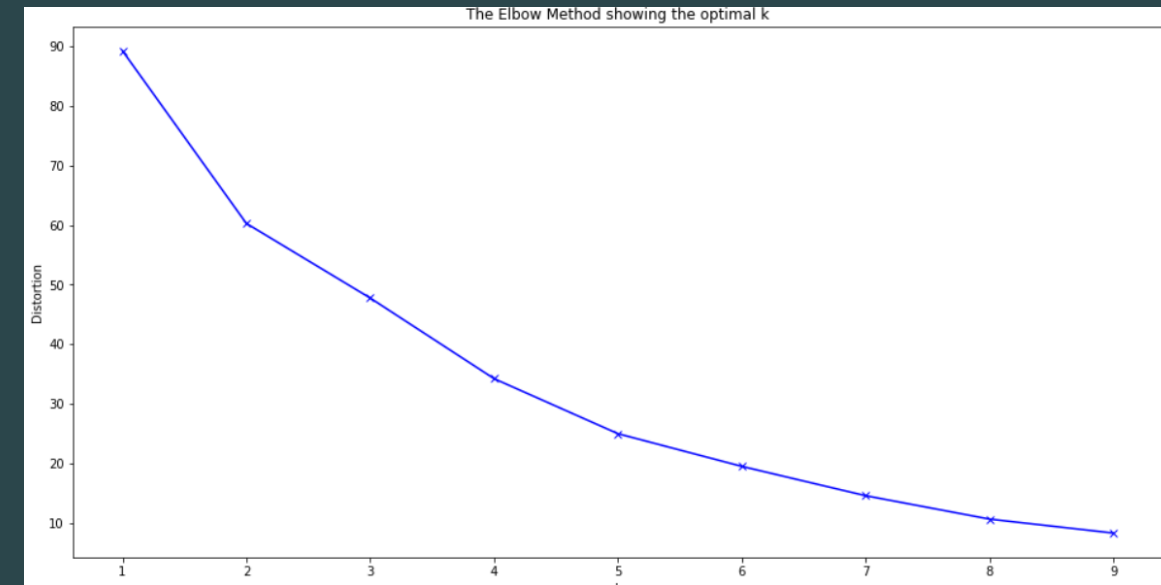
# Correlation analysis

- It is important to try to avoid correlated variables before starting the clustering process
- Through this heat map we visualize the strongest and weakest correlations.
- The one that stands out the most is the correlation between population and total number of venues extracted from Foursquare
- Another interesting positive correlation is that of population density with the value of land use.



# Machine learning technics

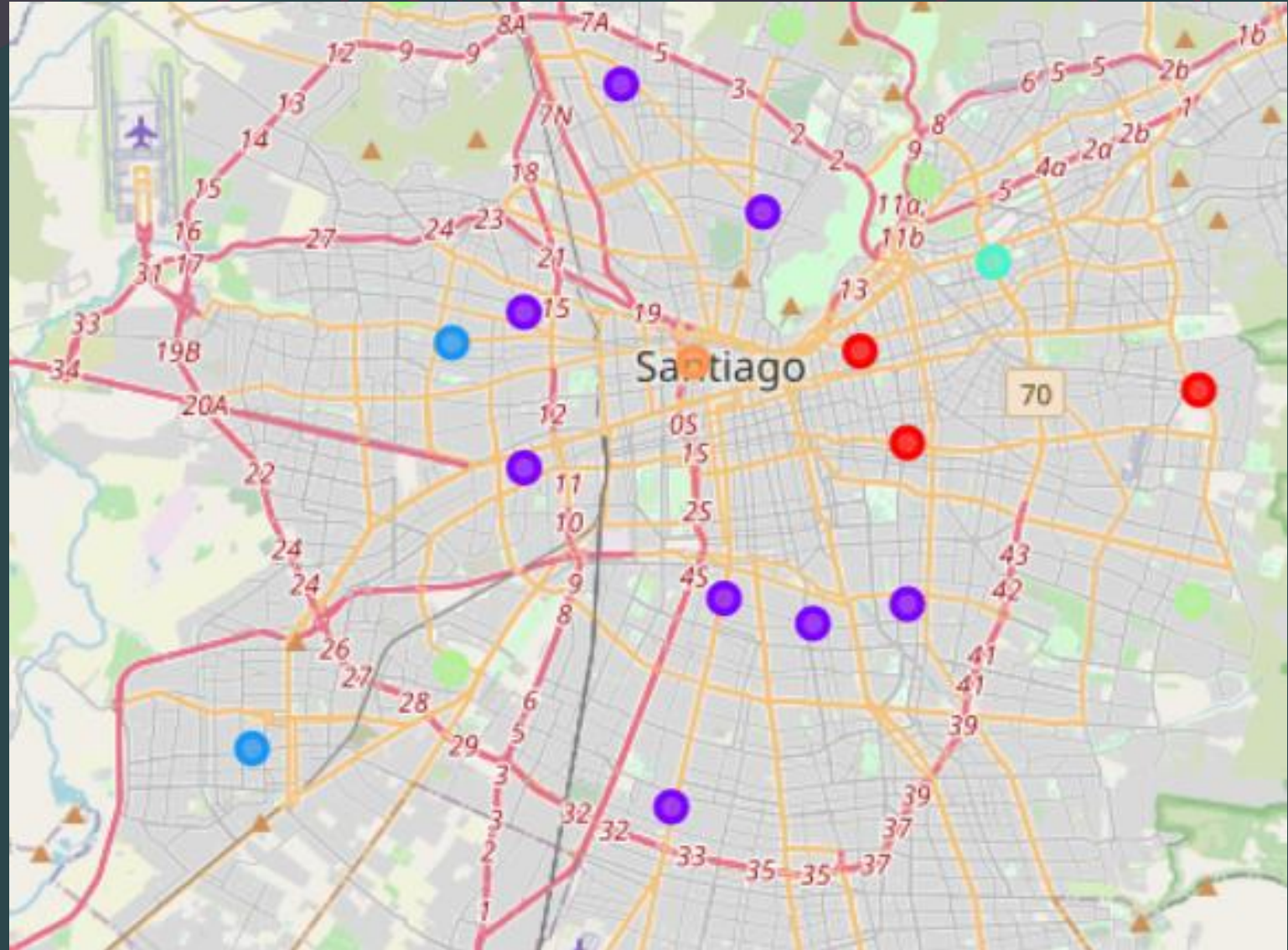
- Clustering techniques vary depending on the bottom-up, top-down (hierarchical) approach, the distance measures used: Euclidean, Hamming (city block), Tchebyshev, Minkowski, Canberra, angular separation and the type of data used, which it can include fuzzy data sets. (Pedrycz, 2005).
- For **customer and market segmentation problems**, the most used techniques are **K-Means**, C-Means and SOM (Self Organization Maps).
- K-Means: one of the problems is its initial randomness at the starting point and the selection of number "k" of partitions. **The elbow method** calculates the total intra-cluster variance as a function of the number of clusters and chooses as optimal the value from which adding more clusters has marginal improvements.
- Before executing the technique, the **data must be standardized**.
- **6 partitions will be used** although the method suggests something else.
- However, the criterion must also be functional to the problem and 6 it is considered that it meets the practical and with the elbow method





# Results: Clustering


- The clusters generated are no longer necessarily geographic.
- The areas with the greatest potential are those in which there is a high concentration of people, high densities and moderate to high values of land costs, in addition to having greater places of interest.
- The best area to start the exploration at the micro level is the cluster in the center of Santiago, where high density, high population, average values of land use and high commercial activity are combined (orange cluster).
- It would be followed by the blue cluster with 8 districts that has as a characteristic a high population density.



# Discussion and Conclusions

- Despite being limited in time and information, it was possible to use the methodological approach.
- The results effectively help people who want to know areas where to invest or start their own business.
- However, at the micro level, there is work to be done. Foursquare is not very popular in Chile, which leads to underestimating the potential of certain areas. A possible improvement would be to supplement the data with more robust trade information.
- Finally, the opportunity to carry out this applied project is appreciated



- 
- **Applied Data Science Capstone**
  - March 2021
  - Pedro Reyes O.

Code available, in this link:

[Segmenting-and-Clustering-Neighborhoods-in-Toronto/PREYES\\_Capstone\\_final.ipynb at master · preyes0/Segmenting-and-Clustering-Neighborhoods-in-Toronto \(github.com\)](https://github.com/preyes0/Segmenting-and-Clustering-Neighborhoods-in-Toronto/blob/master/Segmenting-and-Clustering-Neighborhoods-in-Toronto/PREYES_Capstone_final.ipynb)

# Exploring Best locations for Bakery store in Santiago, Chile