Statistical solutions on finding optimal places to start a local business in Barcelona

Applied Data Science Capstone

Guillem Cáceres Clerias

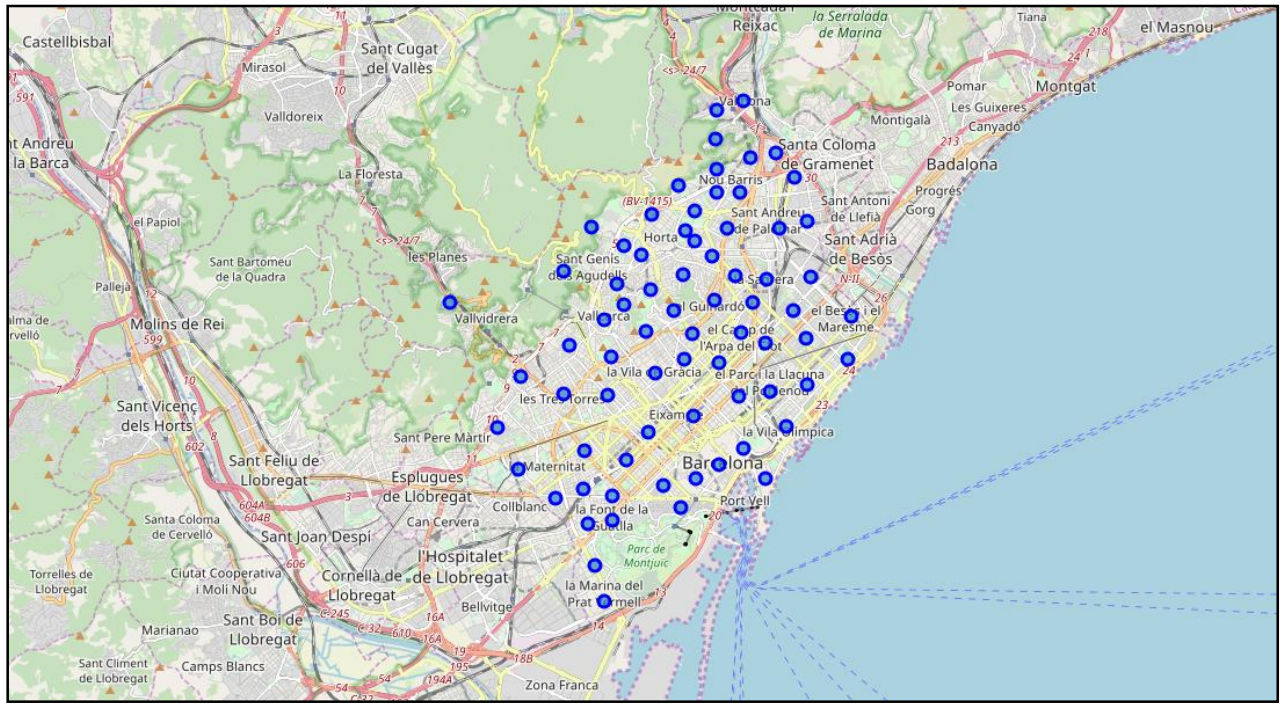
28/09/2019

1. Find optimal neighborhood in Barcelona

Barcelona has an increasing amount of new opportunities to start a new business, but a lack of new constructions space threat.

It is more important than ever choose the correct spot to open your new local.

The objective is to help to select the optimal neighborhood to start a new project both in private and in public sector.



2. Data acquisition and cleaning

The datasets used during this study was extracted by the following source:

- Barcelona statistical department webpage:
  - ❑ Population and net_density by neighborhood
  - ❑ Population age by neighborhood grouped by large age groups.
- Foursquare API:
  - ❑ Total of venues and his category of every Barcelona Neighboarhood.
- Manually created dataset with Google Maps support:
  - ❑ Latitude and longitude of every Neighborhood.

	index	Venue Category	Count
0	248	Spanish Restaurant	59
1	224	Restaurant	58
2	51	Café	56
3	261	Tapas Restaurant	54
4	171	Mediterranean Restaurant	49
5	213	Plaza	49
6	201	Park	47
7	210	Pizza Place	45
8	126	Grocery Store	45
9	20	Bakery	44
10	256	Supermarket	43
11	144	Hotel	42
12	45	Burger Joint	40
13	65	Coffee Shop	40
14	21	Bar	39

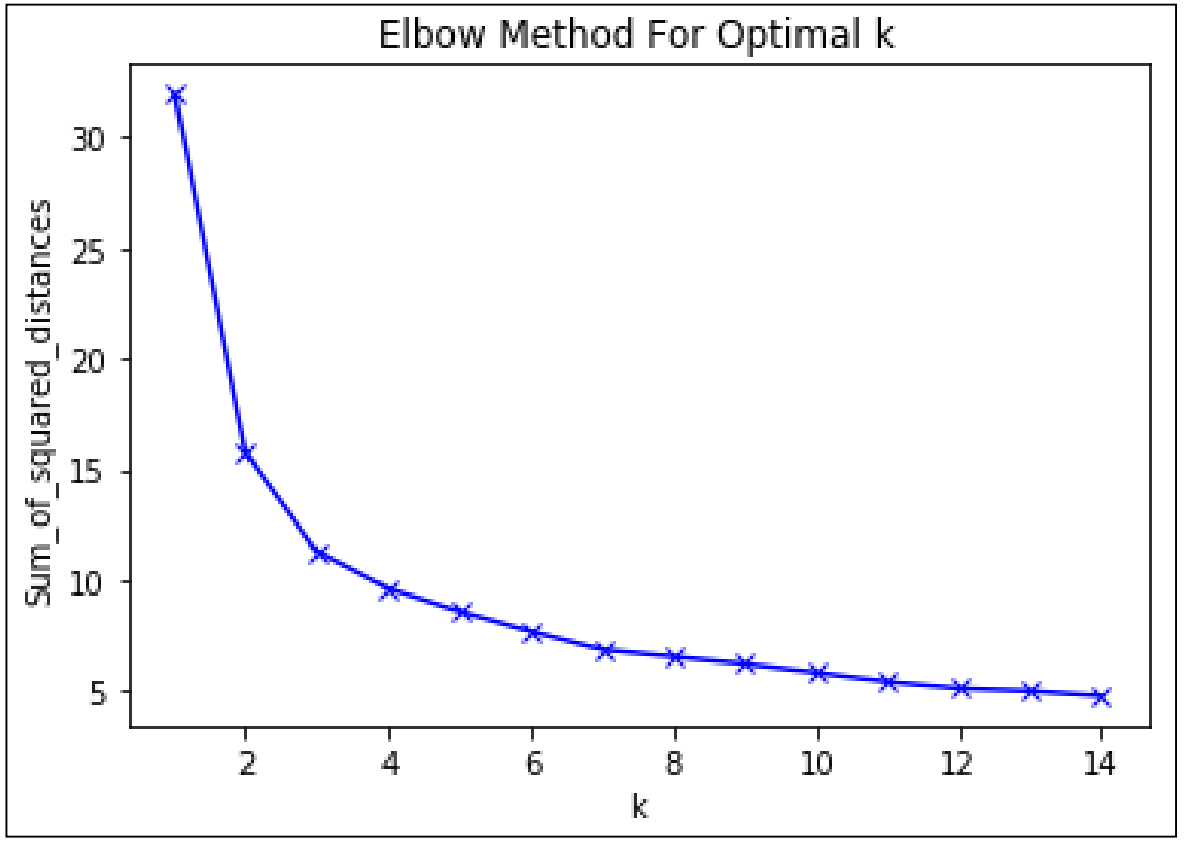
This datasets have been upload to a Github repository as CSV files, processed as dataframes and formatted, filtered and merged during the cleaning procedure.

3. Using K-means to create model

In order to cluster the neighborhoods of the city, I used K-means clustering algorithm for the unlabeled data, like number of same venue category or population density.

- First of all, I normalized the data to compare easily the values of the different fields.

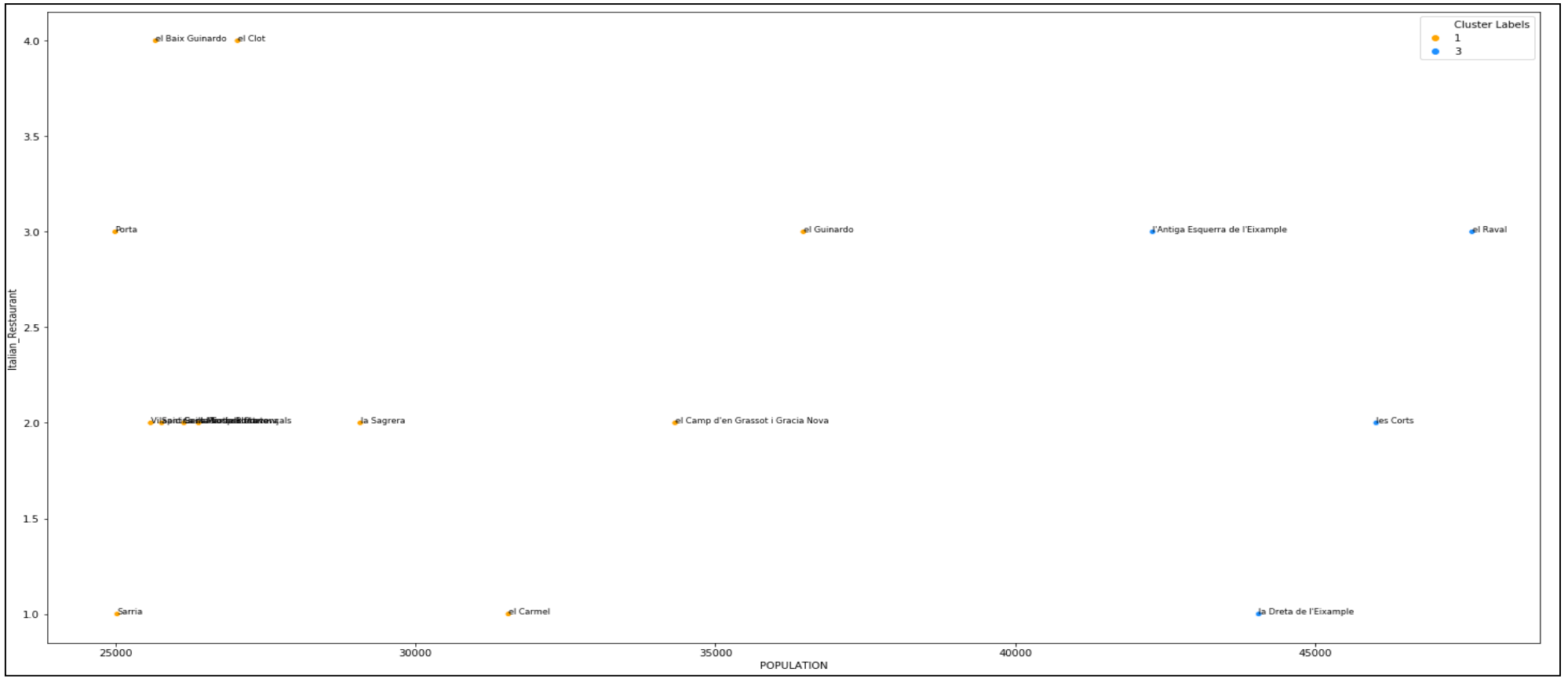
- Before executing the k-mean algorithm, the elbow method is used to find the optimal number of k cluster, in this case, 4.



4. More population and less competence

Which are the most interesting neighborhoods to start an “Italian restaurant”?

Properly Filtering the data and finding the zones with less restaurants of this kind and a population as potential clients.



5. Solving public requirments

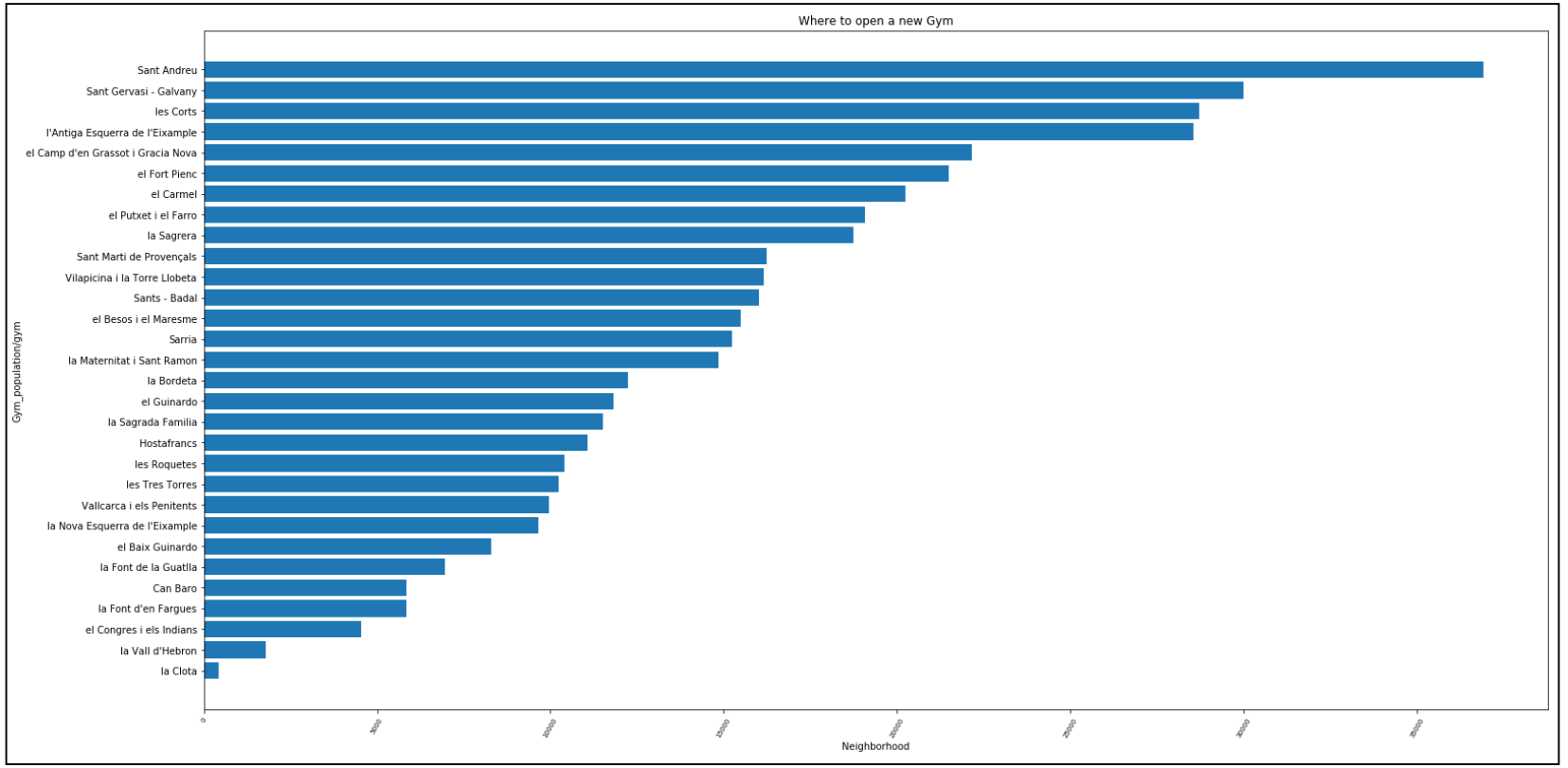
Which neighborhoods are in greater need of a green area (Park)?

In this case, the data was processed to get the neighborhoods but a higher population density per green area, and show how many parks have and which clustering is assigned.

	Neighborhood_main_field	NET_DENSITY	Park	Net_density/Park	Cluster Labels
0	la Salut	680.1	5	136.0	0
1	el Coll	597.6	4	149.0	2
2	Vallcarca i els Penitents	335.1	4	83.0	0
3	Sant Gervasi - la Bonanova	314.0	3	104.0	1
4	la Vila Olimpica del Poblenou	387.3	3	129.0	2
5	la Vall d'Hebron	724.0	3	241.0	2
6	Sant Andreu	746.2	2	373.0	3
7	el Poble Sec	1043.1	2	521.0	1
8	el Baix Guinardo	1081.9	2	540.0	1
9	Verdun	866.7	2	433.0	0
10	Sant Pere, Santa Caterina i la Ribera	700.1	2	350.0	0

6. Choose population section

Which are the zones with a higher potential clients to open a new Gym?



In this last situation I expected a relative importance in the age of the potential clients for a Gym. Having this in mind, processed the data show the neighborhoods with a higher population around 16 -65 years old and that already have at least one Gym near.

7. Conclusions and future implementations

- New investors will need more information to start a new business in Barcelona.
- This study expects to be a support element during the decision making of where to open a new local.
- The value of the results is connected to the amount of data extracted from the venues, which in some cases is not well categorized.
- In order to enable the use of this information to the higher number of users, it has to be done some upgrades in the automation of the visualization and obtaining results.