Search for a bakery location in Barcelona (Spain)

Battle of Neighborhoods – Capstone Project

1. INTRODUCTION

Barcelona (BCN) is the capital from Catalonia (region and province from Spain). With a population of 1,7 million people is the second city in Spain and one of the most populated in the EU. As a global leading tourist, business and cultural centre it offers to residents and visitants all kind of products and services and has an extensive transportation network by sea, air, train and motorways to Spain and the rest of Europe. Small businesses and start-ups are continuously establishing due to all this facts plus a nice climate and lively/friendly character from citizens.

Business Problem

In this project an investor is interested in placing a "gourmet" bakery in Barcelona city and since "Location" is one of the most important success Bakery's KPIs (others are: product quality, price and range, processes, plan and persons) has asked us to help him to decide which neighborhood (one or three) could be the most convenient having into account:

- Barcelona city has 73 neighborhoods (included in 10 districts)
- **Competition**: more than 400 hundred bakeries spread all over the city. Number of bakeries in selected neighborhood/s should be low. Avoid neighborhoods having more than one recognized as "best bakeries"
- **Population and density** should be high rated
  (high people flow could be a plus -tourists and citizens coming from other neighborhoods- but residents population should be high however
- Prices will be higher than average due to quality. Avoid neighborhoods having comparatively low **Family income**.
- Store **Rental prices** (€/m2 month) should not be excessive in any case.

2. DATA REQUIRED

1/ Foursquare API: For getting data about **bakeries** and **neighborhood venues** in Barcelona

2/ BCN **Figures by neighborhoods**-Urban statistics (**Population, Area and Density**) (data 2018)

http://www.bcn.cat/estadistica/angles/dades/barris/timm/tterr/sup418.htm

3/ BCN **Figures by neighborhoods**-Economic statistics-Territorial distribution of household income (**Family income**) (data 2017)

http://www.bcn.cat/estadistica/angles/dades/economia/renda/rdfamiliar/evo/rfbarris.htm

4/ BCN **Figures by neighborhoods**-Urban statistics > The real state market in Barcelona > Prices registered contracts and rental housing (**Rental€m2**) (data 2018, 4Q summed up)

http://www.bcn.cat/estadistica/angles/dades/barris/timm/ipreus/habllo/ls2018.htm

5/ **Latitude** and **longitude** by neighborhood :

Wikipedia+ Geohack (example for ‘Raval’ neighborhood: https://es.wikipedia.org/wiki/Raval, following link “Ubicación” to get decimal coordinates

https://tools.wmflabs.org/geohack/geohack.php?language=es&pagename=El_Raval&params=41.38_N_2.16861_E_type:city)

Latitude and **longitude** are not neighborhood geographical center. It’s supposed to be neighborhood residential center. I have detected 4 neighborhoods having wrong coordinates (placed in other neighborhood): *Sants*, *les Roquetes*, *Clot* and *Canyelles*. Their coordinates have been replaced by other manually estimaded.

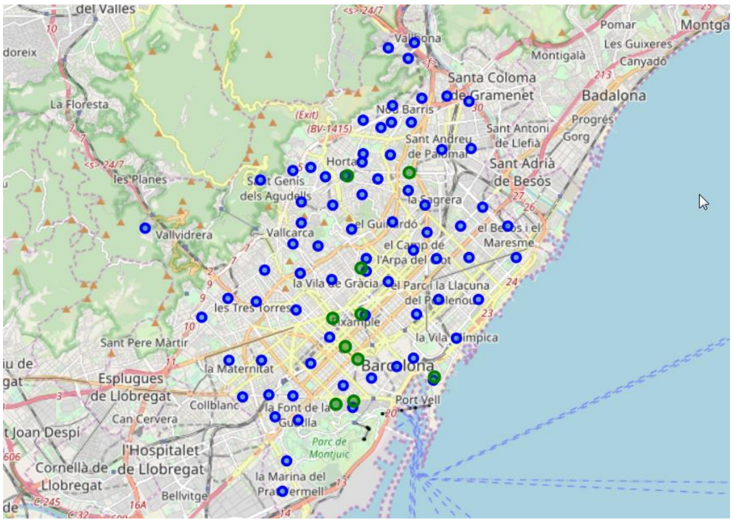
6/ Newspaper article “**The 10 best bakeries in Barcelona**” (January 11,2019)

https://www.elperiodico.com/es/onbarcelona/visitar/20190111/mejores-panaderias-barcelona-7233367

Coordinates obtained through geocode and further individual checking through Google maps.

3. METHODOLOGY

After loading data from files (neighborhoods, best bakeries) and from Foursquare(FS) (bakeries) we use geocode and folium library to display a map of Barcelona with best bakeries and neighborhood centers dotted with different colors. We observe than centers are not geographically “centered” but we will use them because usually represent the center of venues, commerce, cultural and people movement what is most relevant for our project.



We will follow somehow a “funnel” process to propose just one (or five) neighborhoods from 73 candidates. Three stages , ***Bakeries analysis, Neighborhood and Clustering analysis and Neighborhood similarity*** will end us to our Result (final top ranking).

Bakeries analysis

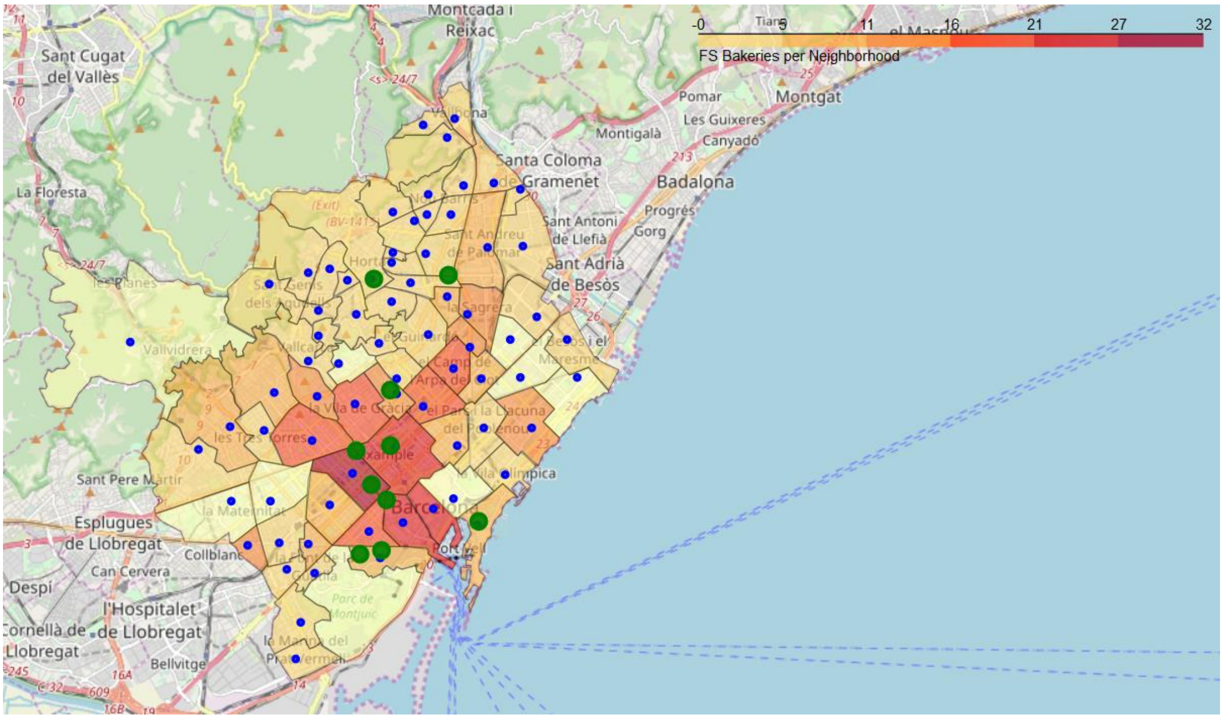
First stage, ***Bakeries analysis*** is subdivided in *Displaying and Analysing Bakeries, Neighborhood variable analysis, Getting FS Venues* and *First Discarding neighborhoods.* This stage will conclude discarding a set of neighborhoods resulting from the union of different criteria discarding subsets.

Displaying and Analysing Bakeries

After cleaning bakeries data obtained from Foursquare(FS) –dropping categories diferent from "bakery", duplicates and ensuring "best bakeries" are included without missing values in resulting dataframe-, we group bakeries by neighborhood to display mean (7,35), min(1) and max(32)..

Next thing is determine the amount of bakeries by neighborhood and show a sample with ten most bakery crowded neighborhoods. An histogram of the distribution of *number of bakeries vs number of neighborhoods* will help us to propose our first discard set including 5 most bakery crowded neighborhoods.

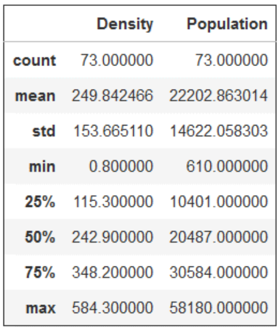
After that we show maps of Barcelona containing bakeries, best bakeries and neighborhood centers: first using simple pointing (brown, green and blue colors), second with marker cluster for bakeries and the last one using choropleth map on neighborhoods based on number of bakeries. This maps will allow us to observe the great bakeries density in Barcelona, downtown high bakeries concentration and differences among neighborhoods based on color scale(choropleth). Also zoom option will allow further detailed analysis by neighborhood or even street.



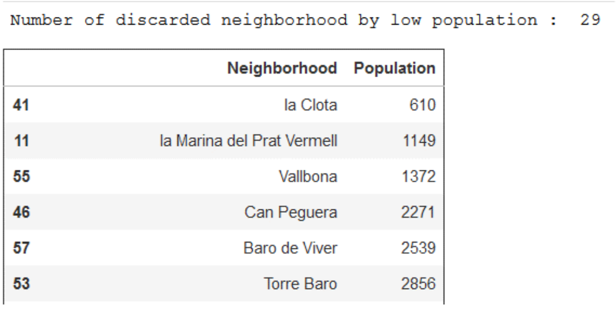
Neighborhood Variable Analysis

First we test to find neighborhood variable correlation. Family income vs Residential Rental €m2-month have 0.72 but we don't consider enough to discard anyone of them. We will be interested in having simultaneously high Family income and low rental€m2

Following that, we observe neighborhood **population and population density**. According to initial business given criteria we create two discard sets: *low_density* (including only neighborhoods in quantile 15%) and *low_pop* (including neighborhoods in quantile 40%). Being more restrictive in population is intended because of giving not so many relevance to movements inside a neighborhood.



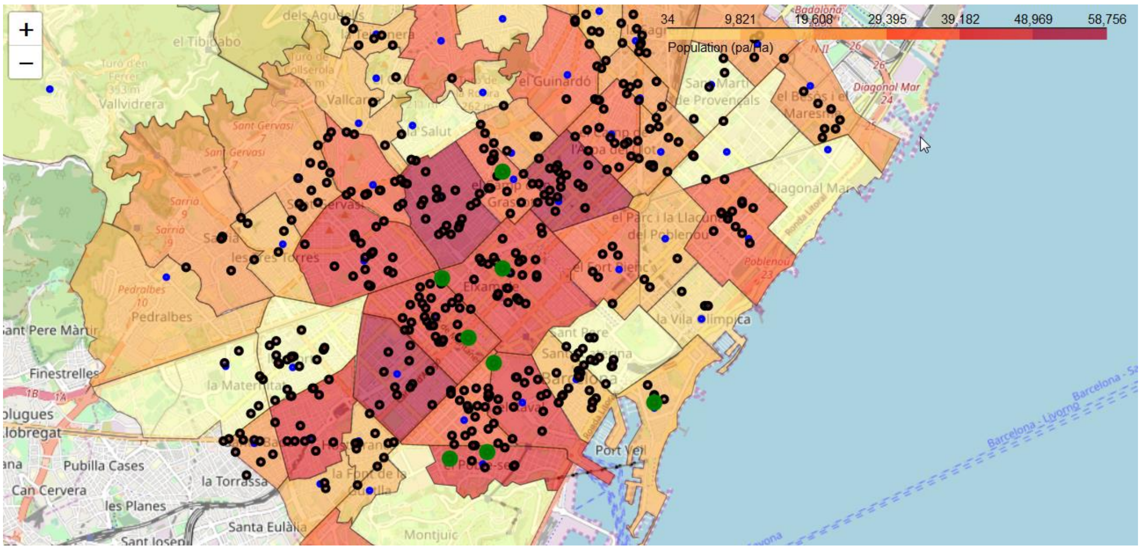
	Density	Population
count	73.000000	73.000000
mean	249.842466	22202.863014
std	153.665110	14622.058303
min	0.800000	610.000000
25%	115.300000	10401.000000
50%	242.900000	20487.000000
75%	348.200000	30584.000000
max	584.300000	58180.000000



```
Number of discarded neighborhood by low population :  29
```

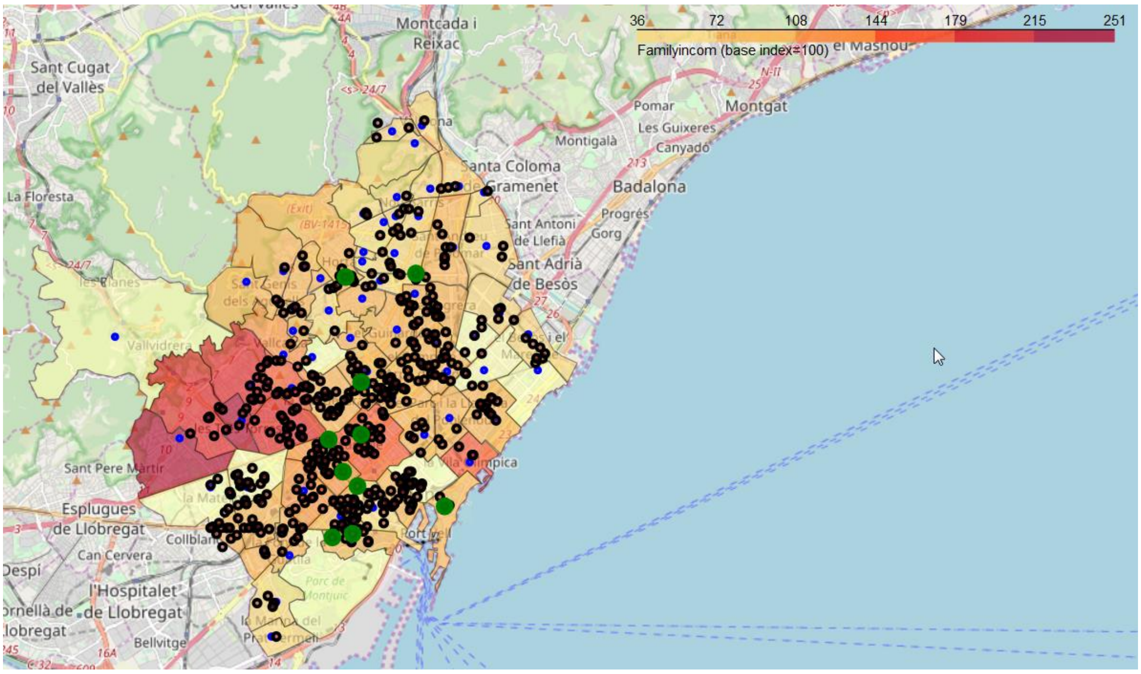
	Neighborhood	Population
41	la Clota	610
11	la Marina del Prat Vermell	1149
55	Vallbona	1372
46	Can Peguera	2271
57	Baro de Viver	2539
53	Torre Baro	2856

A choropleth map of Barcelona on neighborhood population, combined with dotted bakeries(black), best bakeries(green) and neighborhood centers(blue) will confirm higher bakeries concentration on most populated neighborhoods, and the fact that some few "best bakeries" are not in this.



Family income, using an index developed by BCN public administration –base index all BCN city is 100- is our next variable to analyse. We observe min,mean, max, percentiles and create a new discard set *low_income* including neighborhoods in quantile 30% (supermarkets, convenience stores, low price bakeries will be probably more competitive than our "gourmet" bakery in this area).

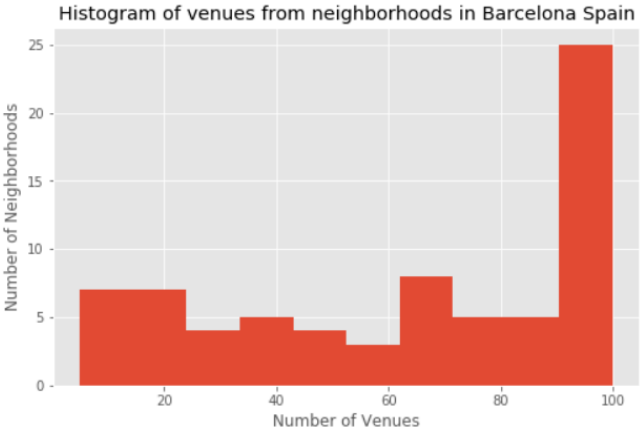
Another choropleth map, this time based on Family income will show us than bakerys concentration is lower on lighter colored (lower) Family income neighborhoods. Excepcionally, two very high Family income neighborhoods (Pedralbes and Tres Torres) almost have not bakeries.



With Store **Rental€m2**-month we observe again min,mean, max, percentiles. Investor does not want to have high rental expenses with this first shop and we create a new discard set *toohigh_rental* including neighborhoods with values higher than quantile 75% but also limited to a maximum of ten neighborhoods

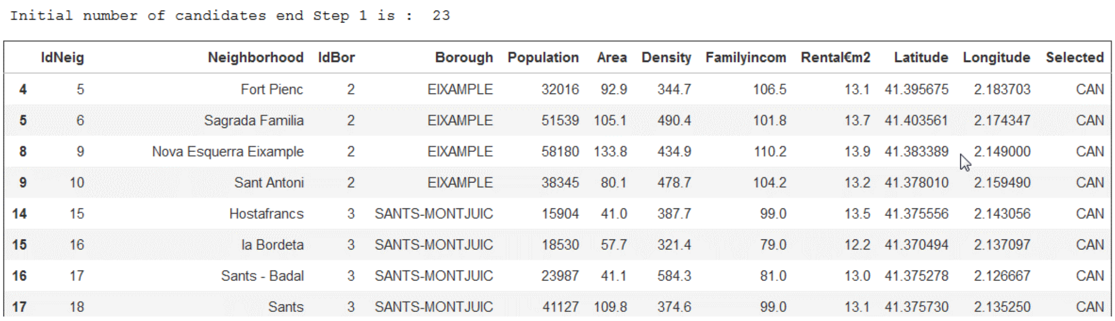
Getting FS Neighborhood Venues

First we get venues for all neighborhoods (radius=700 m) obtaining about 4700 venues grouped in 292 different categories. We can see *Tapas, Spanish and Restaurant* in general as the most frequent ones. An histogram of the distribution of *number of venues vs number of neighborhoods* will give us about 25 neighborhoods having 100 venues and about 13 having less than 20 venues. We assume a very low touristic people movement in the last ones and create another discard set, named *low_movement*.



First Discarding neighborhoods

This time we create a discard set *(bbdiscard)* formed by neighborhoods with more than one "best bakery" and a merged list including them and all discard sets created previously (*high_nobkyrs, low_density ,low_pop,low_income,toohigh_rental,low_movement*) by using a "multiple" outer join. We show total number of candidates and sample list of discarded and candidates using a new filter column "Selected" with values "CAN" or "NO".



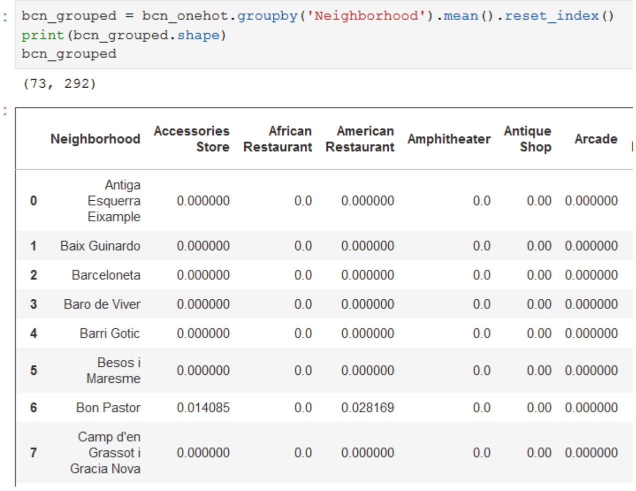
```
Initial number of candidates end Step 1 is :  23
```

	IdNeig	Neighborhood	IdBor	Borough	Population	Area	Density	Familyincom	Rental€m2	Latitude	Longitude	Selected
4	5	Fort Pienc	2	EIXAMPLE	32016	92.9	344.7	106.5	13.1	41.395675	2.183703	CAN
5	6	Sagrada Familia	2	EIXAMPLE	51539	105.1	490.4	101.8	13.7	41.403561	2.174347	CAN
8	9	Nova Esquerra Eixample	2	EIXAMPLE	58180	133.8	434.9	110.2	13.9	41.383389	2.149000	CAN
9	10	Sant Antoni	2	EIXAMPLE	38345	80.1	478.7	104.2	13.2	41.378010	2.159490	CAN
14	15	Hostafrancs	3	SANTS-MONTJUIC	15904	41.0	387.7	99.0	13.5	41.375556	2.143056	CAN
15	16	la Bordeta	3	SANTS-MONTJUIC	18530	57.7	321.4	79.0	12.2	41.370494	2.137097	CAN
16	17	Sants - Badal	3	SANTS-MONTJUIC	23987	41.1	584.3	81.0	13.0	41.375278	2.126667	CAN
17	18	Sants	3	SANTS-MONTJUIC	41127	109.8	374.6	99.0	13.1	41.375730	2.135250	CAN

This step ends Stage 1, ***Bakeries Analysis.***

Neighborhood and Clustering analysis

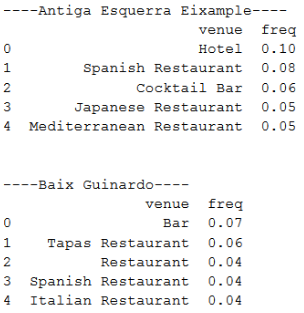
First at this second stage, we construct a neighborhood grouped dataframe with dummy variables from column "Venue category" in FS returned venues . Values shown are the mean of the frequency of occurrence for each category.



```
bcn_grouped = bcn_onehot.groupby('Neighborhood').mean().reset_index()
print(bcn_grouped.shape)
bcn_grouped
```

```
(73, 292)
```

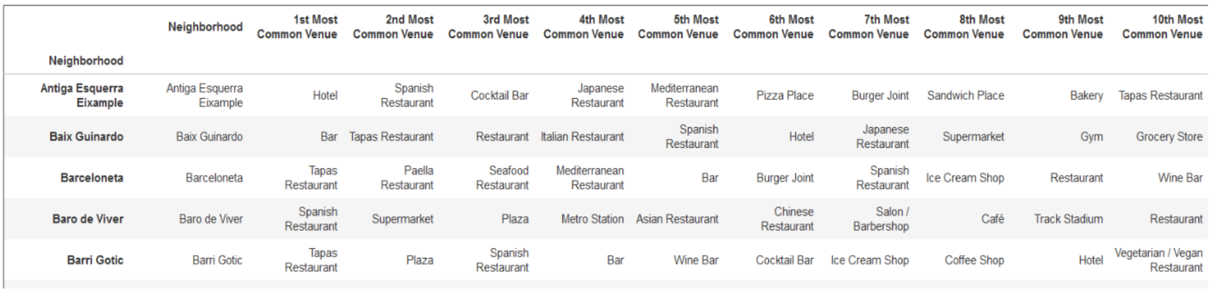
	Neighborhood	Accessories Store	African Restaurant	American Restaurant	Amphitheater	Antique Shop	Arcade
0	Antiga Esquerra Eixample	0.000000	0.0	0.000000	0.0	0.00	0.000000
1	Baix Guinardo	0.000000	0.0	0.000000	0.0	0.00	0.000000
2	Barceloneta	0.000000	0.0	0.000000	0.0	0.00	0.000000
3	Baro de Viver	0.000000	0.0	0.000000	0.0	0.00	0.000000
4	Barri Gotic	0.000000	0.0	0.000000	0.0	0.00	0.000000
5	Besos i Maresme	0.000000	0.0	0.000000	0.0	0.00	0.000000
6	Bon Pastor	0.014085	0.0	0.028169	0.0	0.00	0.000000
7	Camp d'en Grassot i Gracia Nova	0.000000	0.0	0.000000	0.0	0.00	0.000000



```
----Antiga Esquerra Eixample----
                      venue  freq
0                     Hotel  0.10
1        Spanish Restaurant  0.08
2               Cocktail Bar  0.06
3       Japanese Restaurant  0.05
4  Mediterranean Restaurant  0.05


----Baix Guinardo----
                venue  freq
0                 Bar  0.07
1    Tapas Restaurant  0.06
2          Restaurant  0.04
3  Spanish Restaurant  0.04
4  Italian Restaurant  0.04
```

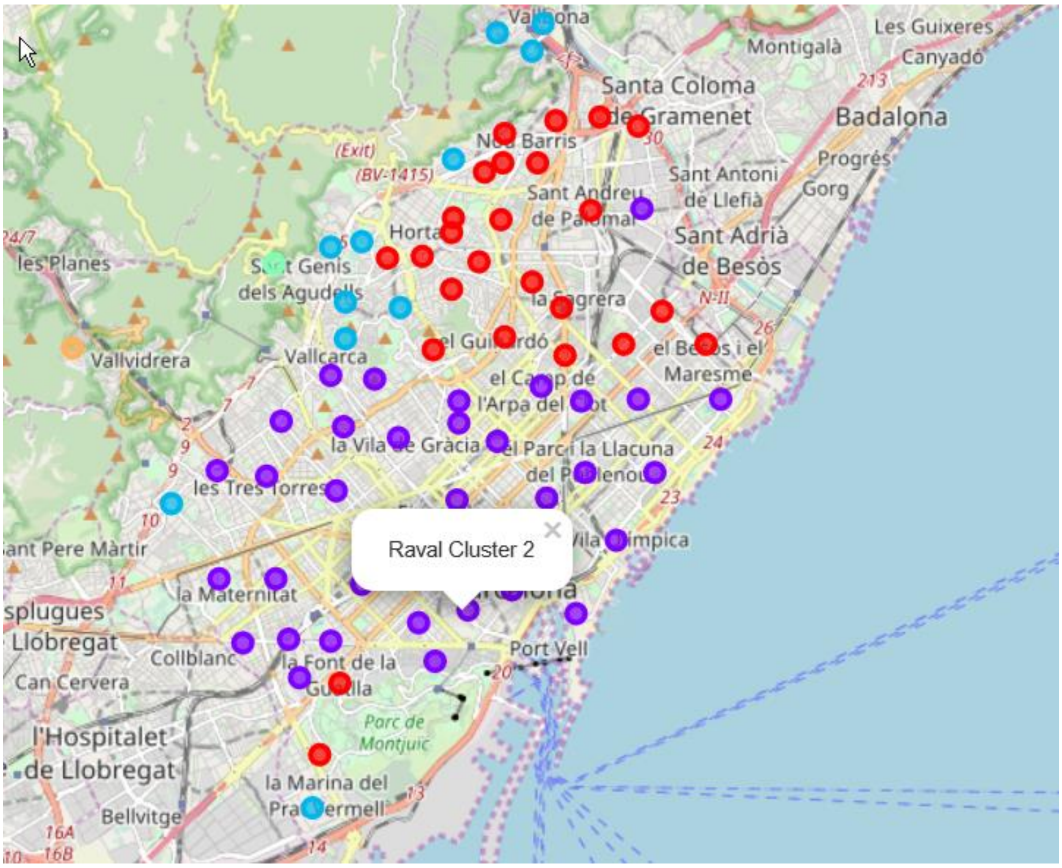
After that we print each neighborhood along with their five most common venue categories plus frequency, and create a neighborhood indexed dataframe displaying top 10 venues categories in descending order.



	Neighborhood	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
Neighborhood											
Antiga Esquerra Eixample	Antiga Esquerra Eixample	Hotel	Spanish Restaurant	Cocktail Bar	Japanese Restaurant	Mediterranean Restaurant	Pizza Place	Burger Joint	Sandwich Place	Bakery	Tapas Restaurant
Baix Guinardo	Baix Guinardo	Bar	Tapas Restaurant	Restaurant	Italian Restaurant	Spanish Restaurant	Hotel	Japanese Restaurant	Supermarket	Gym	Grocery Store
Barceloneta	Barceloneta	Tapas Restaurant	Paella Restaurant	Seafood Restaurant	Mediterranean Restaurant	Bar	Burger Joint	Spanish Restaurant	Ice Cream Shop	Restaurant	Wine Bar
Baro de Viver	Baro de Viver	Spanish Restaurant	Supermarket	Plaza	Metro Station	Asian Restaurant	Chinese Restaurant	Salon / Barbershop	Café	Track Stadium	Restaurant
Barri Gotic	Barri Gotic	Tapas Restaurant	Plaza	Spanish Restaurant	Bar	Wine Bar	Cocktail Bar	Ice Cream Shop	Coffee Shop	Hotel	Vegetarian / Vegan Restaurant

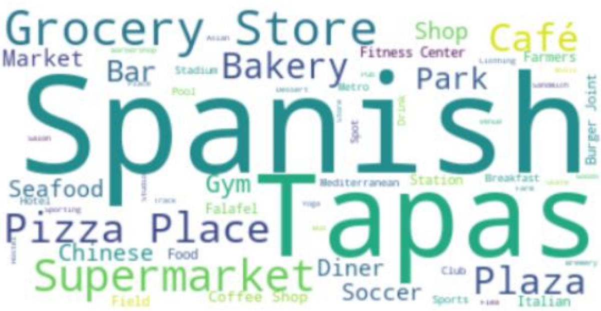
K-means algorithm from machine learning Scikit-learn library for Python is used to cluster the neighborhood into 5 clusters, an easy modifiable number (we also tested with 4 and 6 values but not getting a significant improvement in grouping semantics).

A Barcelona colored-cluster dotted map displays different neighborhood clusters



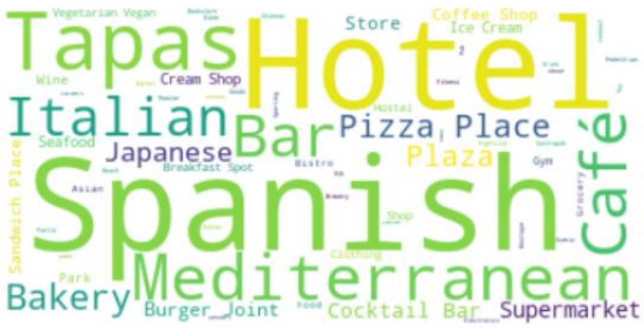
We examine clusters using word clouds (adding "Restaurant" as stopword to give more relevance to other words) based on neighborhood table with 10 most frequent venues:

Cluster 1 (red points), is comprised by 25 neighborhoods. Almost all of them -except two- are in the north of the city well grouped. Word cloud has a reasonable density (that means venue density) and number of words. *Spanish and tapas restaurant, Supermarket, Grocery Store, Plaza, Pizza Place, Bakery and Café* are most remarkable words.



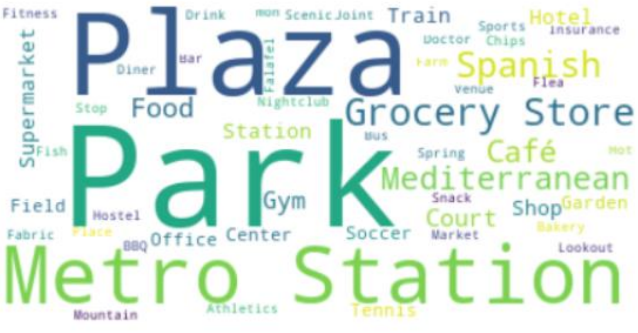
Neighborhoods seem to be just well populated residential and non-touristic areas, as we can confirm -by our native knowledge- checking the list of neighborhoods included.

Cluster 2 (purple points), is comprised by 35 neighborhoods. All of them except one (Bon Pastor) are well grouped and spread all over downtown, coastal and west areas of the city. Word cloud has a higher density (more and more different venues) and *Spanish restaurant* and *Hote*l are the top words, followed by *Tapas, Café, Italian restaurant, Bakery* and diferent kinds of specialized restaurants (Italian, Japanese, Burger, Pizza, Vegetarian…).



This cluster represents more people diversity, both residential and touristic and includes most centric neighborhoods in the city. "Bon Pastor" outlier neighborhood is probably included because of having one of the biggest shopping centers in the city *(La Maquinista)*

Cluster 3 (blue points), is comprised by 11 neighborhoods close to mountain line going from southwest to north and far from the sea and downtown. Word cloud density and diferent number of venues is lower than in clusters 1 and 2 and most significative are words *Park, Plaza, and Metro Station*.



This cluster represent residential neighborhoods, not as much "diversity venue" crowded as clusters 1 and 2, well communicated and provided of urban facilities *(Park, Plaza, Scenic joint, Garden, Gym, Lookout, Tennis, Mountain, Metro, Train, Bus).* "Pedralbes" outlier neighborhood is included here because clustering is based only in neighborhood venues and not in demographic variables such as population, family income or rental prices considered in Stage 1.

Cluster 4 (green point), is comprised by only 1 neighborhood (S.Genis dels Agudells) noted mainly by *Trails*, *Scenic Lookouts and Tennis courts*. Apart from cluster 5 his center coordinates are the closest to the mountain line (includes words such as *Farm* and *Farmers*)

Lookout
Tennis
Bodega
Court
Tapas
Trail
Deli
Farm
Flower
Market
Sports
Farmers
Scenic
Shop
Bar

And finally ***Cluster 5*** (brown point) includes one neighborhood (Vallvidrera, Tibidabo y Les Planes). It is formed by three isolated surrounding residential developments inside the green mountain area out from the city. So our clustering algorithm (and data) have worked well considering it an special cluster.

Not surprinsingly, remarkable words are now *Lake, Tourist, Information* and *Train, Hostel, Station.*

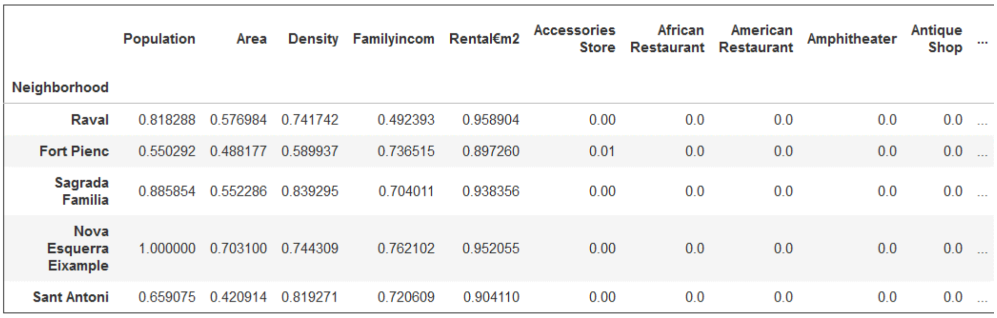
Information
Field
Station
Center
Food
Soccer
Hostel
Fast
Lake
Train
Fabric
Tourist
Shop
Exhibit

From the analysis of cluster characteristics we can set our preferred one. Number two is our choice because of better geographical centrality (closer to downtown) and diversity/type of venues suggesting higher people flow. Making the intersection with neighborhoods resulting from Stage 1 (Selected='CAN') we get a group of 18 neighborhood as candidates at the end of this Stage 2.

Neighborhood similarity and Final top ranking

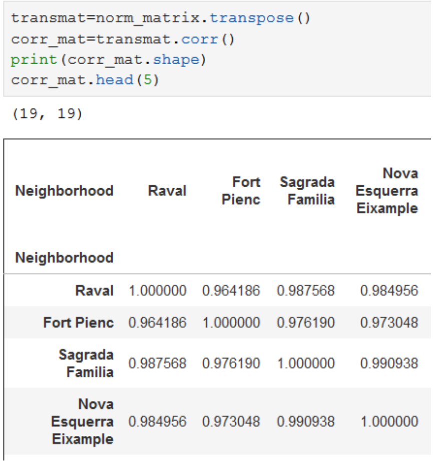
At this Stage 3, inspired by content-based recommender systems, we determine which is/are the most similar neighborhoods to the profile of a “location model” bakery selected from our list of ten “best bakeries” -location model because that is the kpi we are analyzing-.

Similarity is calculated using *Pearson correlation* on a neighborhood matrix composed by “simple-scaling normalized” demographic values (population, density, area, family income, rental€/m2) and the mean of the frequency of occurrence of each venue category.



	Population	Area	Density	Familyincom	Rental€m2	Accessories Store	African Restaurant	American Restaurant	Amphitheater	Antique Shop	...
Neighborhood											
Raval	0.818288	0.576984	0.741742	0.492393	0.958904	0.00	0.0	0.0	0.0	0.0	...
Fort Pienc	0.550292	0.488177	0.589937	0.736515	0.897260	0.01	0.0	0.0	0.0	0.0	...
Sagrada Familia	0.885854	0.552286	0.839295	0.704011	0.938356	0.00	0.0	0.0	0.0	0.0	...
Nova Esquerra Eixample	1.000000	0.703100	0.744309	0.762102	0.952055	0.00	0.0	0.0	0.0	0.0	...
Sant Antoni	0.659075	0.420914	0.819271	0.720609	0.904110	0.00	0.0	0.0	0.0	0.0	...

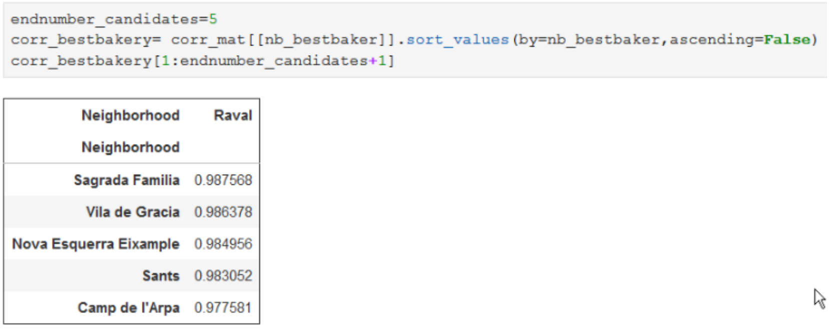
After matrix transposition and correlation we get a table we can filter by our selected best bakery neighborhood and top endnumber of candidates to obtain sorted results.



```
transmat=norm_matrix.transpose()
corr_mat=transmat.corr()
print(corr_mat.shape)
corr_mat.head(5)
```

```
(19, 19)
```

Neighborhood	Raval	Fort Pienc	Sagrada Familia	Nova Esquerra Eixample
Neighborhood				
Raval	1.000000	0.964186	0.987568	0.984956
Fort Pienc	0.964186	1.000000	0.976190	0.973048
Sagrada Familia	0.987568	0.976190	1.000000	0.990938
Nova Esquerra Eixample	0.984956	0.973048	0.990938	1.000000



```
endnumber_candidates=5
corr_bestbakery= corr_mat[[nb_bestbaker]].sort_values(by=nb_bestbaker,ascending=False)
corr_bestbakery[1:endnumber_candidates+1]
```

Neighborhood	Raval
Neighborhood	
Sagrada Familia	0.987568
Vila de Gracia	0.986378
Nova Esquerra Eixample	0.984956
Sants	0.983052
Camp de l'Arpa	0.977581

4. RESULTS AND DISCUSSION

We have found an answer to our investor question, ¿what neighborhood (one or top reduced number ones) would be most convenient to establish a bakery in Barcelona taking into account competition density, population and population density, family income and rental prices per month. That would be the starting point for final 'street level' exploration where "crossing" external information about selected neighborhood/s locals available for rental and zooming in some of our generated maps including bakeries distribution could help in a great way.

Further research and improvements could be made: checking actual bakeries data vs FS returned data (limited by a 500m radius parameter), using actual store rental prices by neighborhood (creating specific dataset from most relevant real estate agencies) or removing FS limit of 100 returning events by call (paid).

Also trying results using "candidate" sets intersection instead of joining "discarding" sets (stage 1), including stage 1 variables into clustering, and even replacing "best-bakery model" neighborhood (stage 3) by a more complex "n-best bakeries" model for calculating an "average ranking" of rankings by tuple "best bakery-neighborhood" would provide different perspectives and chances for a better generalization. A higher level analysis on same-owner bakery shop groups could even be made if realiable bakery guild data are available.

5. CONCLUSION

More than answering our title report, *Searching for a bakery location in Barcelona*, a "funnel" three stage methodology extensible to other finding-business-location cases has been proposed. This methodology applies to public and Foursquare data sources and combines some discarding rules assumptions based on initial business problem definition and a mix of tools and machine learning algorithms (such as interactive geodata maps, choropleth, wordclouds, histogram, descriptive statistics, k-means clustering and Pearson correlation similarity).

Nowadays, decision-support systems based on data analysis and AI machine learning algorithms are a wider spreaded reality than in last years due to higher computational power, easier tools, education and experience. Automatic and semi-autonomous systems will even have more impact in our future lives.