Find the Best City and Neighborhood to set a Bakery Shop business in the U.S.

Applied Data Science Capstone Project (IBM/Coursera¶)

by

Federico Sarrailh

July 27th, 2021

1. INTRODUCTION

1.1 Business Under Study

In this project we advise an investor who is trying to find out where to open a **Bakery Shop in the United States**. He wants to find the **best city in terms of population and income per capita**, where he knows the people spends more; and then compare neighborhoods and their venues to decide where to put his store.

It is reasonable to suppose there are lots of food venues in big cities centers, so we will try to detect **locations that are not already crowded with them**. We are also particularly interested in **neighborhoods with less/no bakeries in it**. We would also prefer locations **as close to city center as possible**.

We will use our data science powers to determine the best cities and the most promising neighborhoods based on these criteria. Advantages of each area will then be expressed so the best possible final location can be chosen by our investor.

2. DATA

2.1 Data Description

Based on the definition of our problem, **factors that will influence** our decision are:

- more crowded and better income per capita cities in the US
- potential competing food venues in each neighborhood (we will consider only the top 10 venues)
- any Bakery shop in the top 10 of the neighborhoods?
- distance of neighborhood from city center (from observation)

2.2 Data Sources

Following data sources will be needed to extract/generate the required information:

- top populated US cities with better per capita income, will be obtained by **web scrapping** using BeautifulSoup

  data link: https://www.statista.com/statistics/205618/per-capita-income-in-the-top-20-most-populated-cities-in-the-us/

- neighborhoods in the chosen cities, will be obtained by **Wikipedia web scrapping** using BeautifulSoup

  data link: https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco

  data link: https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle

- cities and neighborhoods location data, will be obtained using **Nominatim geocoding**
- number of venues and their type and location in every neighborhood will be obtained using **Foursquare API**
- representation maps using Folium
- representation graphs using Plotly and Matplotlib

3. METHODOLOGY

3.1 Procedure steps

For this study, our **First step** is searching for USA most populated cities with higher per capita income. From those candidates cities, we'll choose the top 2 for simplification. We are going to individualize each neighborhood and obtain its coordinates in both. We'll do a map representation of them.

Second step is to explore every neighborhood in terms of venues. We are going to find out the most popular venues.

Now with steps One and Two completed, we're in condition of clean and merge the gathered information.

So, our **Third step** is to analyze and prepare the venues data of the neighborhoods in the chosen cities. We are going to consider the top 10 most common venues. We are particularly interested in food categories, highlighting bakeries if any.

Fourth step is clustering the neighborhoods through a machine learning model, in order to understand and compare them.

Finally, our **Fifth step** is to examine each group and determine the discriminating venue categories that distinguish each one.

Then, with a little help of calculations and simple plots on the previous information, we will be in a good position to recommend the best for the investor.

As explained, we divide this work in steps for a better understanding. The steps and contents are the following:

STEP 1: Scraping info, building and cleaning the dataframes

STEP 2: Exploring Venues

STEP 3: Normalization and grouping by

STEP 4: Clustering through a ML model

STEP 5: Numerical & Examination

3.2 Steps Developing: Obtaining the info and workflow explanation

3.2.1 STEP 1: Scraping info, building and cleaning the dataframes

In order to discover the cities in the USA with the highest income, we decided to do web scraping. For the purpose of this project, we choose to get the info contained in "Per capita income in the most populated cities in the United States in 2019(in U.S. dollars)", link provided in the section **2.2 Data Sources.**

	City	Per capita income 2019
0	San Francisco city, California	75,084
1	Seattle city, Washington	65,205
2	Washington city, District of Columbia	59,808
3	San Jose city, California	51,310
4	Boston city, Massachusetts	48,978
5	Denver city, Colorado	47,802
6	Austin city, Texas	46,217
7	San Diego city, California	43,249
8	New York city, New York	43,046
9	Chicago city, Illinois	40,277
10	Charlotte city, North Carolina	40,071
11	Nashville-Davidson metropolitan government (ba...	38,847
12	Los Angeles city, California	37,779
13	Dallas city, Texas	36,288
14	Houston city, Texas	33,377
15	Columbus city, Ohio	31,843
16	Oklahoma City city, Oklahoma	31,019
17	Jacksonville city, Florida	30,780
18	Phoenix city, Arizona	30,686
19	Fort Worth city, Texas	30,115
20	Philadelphia city, Pennsylvania	29,766
21	Indianapolis city (balance), Indiana	29,008
22	San Antonio city, Texas	26,826
23	El Paso city, Texas	22,583
24	Detroit city, Michigan	21,044

Image 1. PCI in the most populated cities in the US, 2019.

We choose the top 2 most populated and better income cities: **San Francisco and Seattle.**

To find out the neighborhoods in the previous cities we scrape *Wikipedia*, links provided in the section **2.2 Data Sources.** We also use **geopy.geocoders** to obtain Latitude and Longitude data.

In [16]: df_sf

Out[16]:

	Neighborhood	Latitude	Longitude
0	Alamo Square	37.777220	-122.431460
1	Anza Vista	37.780480	-122.443580
2	Ashbury Heights	37.764870	-122.445900
3	Balboa Hollow	37.775890	-122.493600
4	Balboa Terrace	37.731800	-122.467400
5	The Bayview	37.733450	-122.389980
6	Belden Place	37.791275	-122.403785
7	Bernal Heights	37.739040	-122.416250
8	Buena Vista	37.806468	-122.420867
9	Butchertown (Old and New)	37.777120	-122.419640
10	The Castro	37.758490	-122.434770

In [25]: df_se

Out[25]:

	Neighborhood	Latitude	Longitude
0	North Seattle	47.643724	-122.302937
1	Broadview	47.722380	-122.364980
2	Bitter Lake	47.718680	-122.350300
3	North Beach / Blue Ridge	47.700440	-122.384180
4	Crown Hill	47.695200	-122.374100
5	Greenwood	47.690820	-122.355290
6	Northgate	47.713100	-122.319300
7	Haller Lake	47.723200	-122.338700
8	Pinehurst	47.718940	-122.314000
9	North College Park\n(Licton Springs)	47.698550	-122.337630
10	Maple Leaf	47.700130	-122.317650

Image 2. San Francisco's and Seattle's neighborhoods with Lat and Long data, extracts from the dataframes.

Now we can do a map representation of the cities and their neighborhoods. This visual representation also allows us to discover outliers values in the dataframes. We clean the dataframes by getting rid of these values.

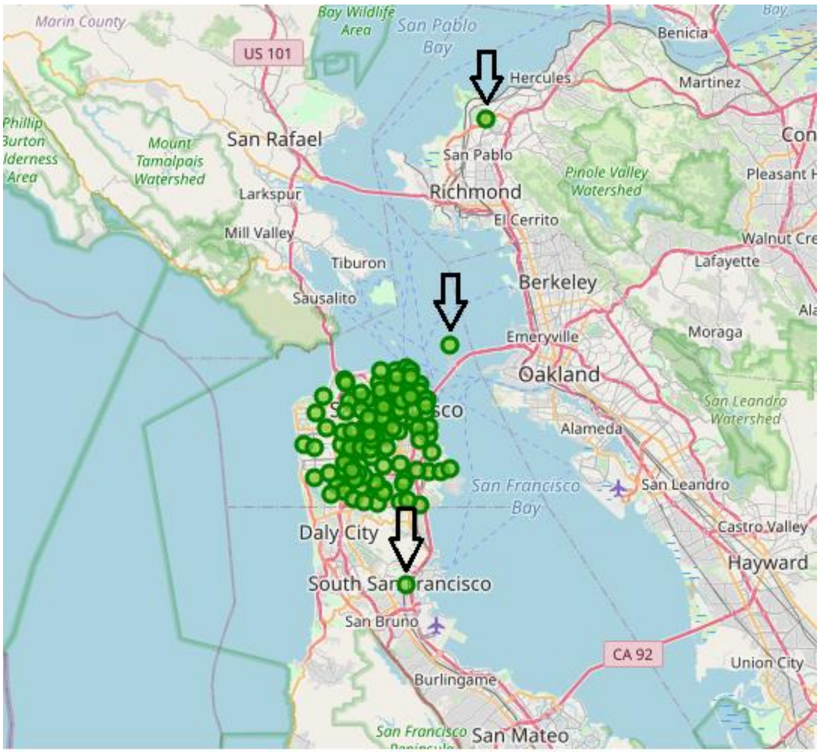


Image 3. Outliers to be removed from San Francisco dataframe.

We can see in the next page a good map representation of the neighborhoods in both cities, which also show us that the dataframes are how we wanted.

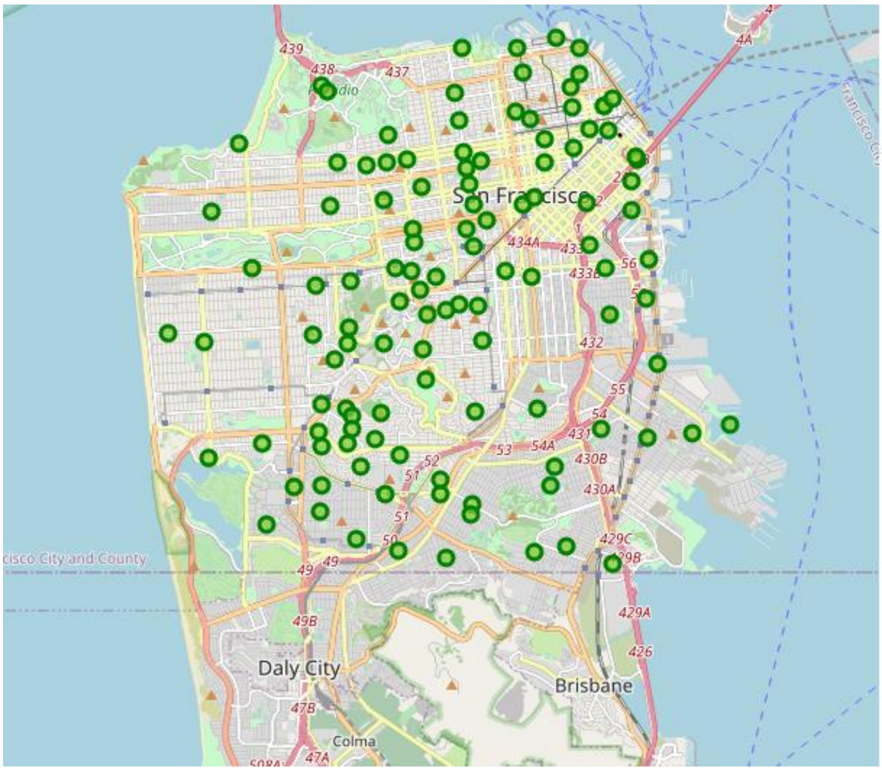


Image 4. SAN FRANCISCO city and its neighborhoods marked.

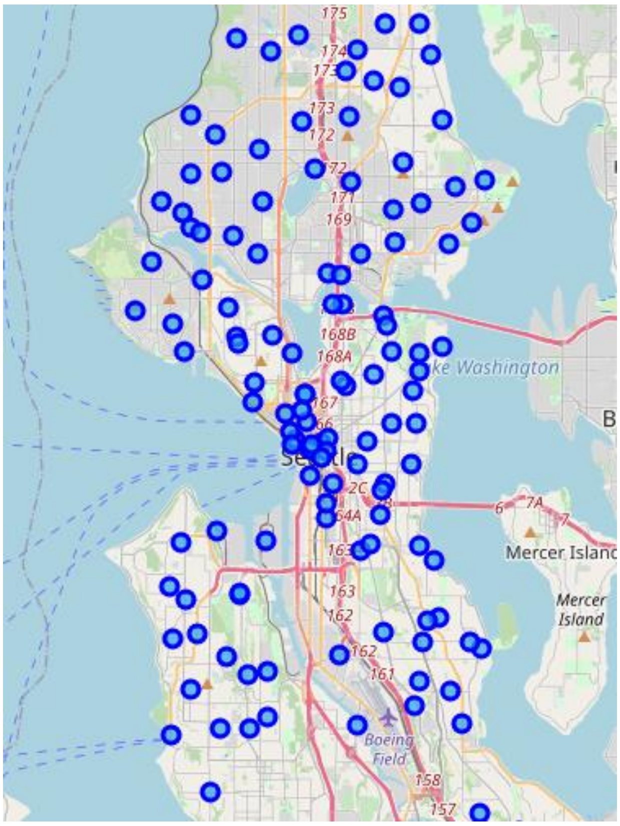


Image 5. SEATTLE city and its neighborhoods marked.

3.2.2 STEP 2: Exploring venues

We will use the **explore** function of the Foursquare API to get the most common **venues** in each neighborhood from SAN FRANCISCO and from SEATTLE.

Let's get the top 50 venues within a radius of 500 meters. Why we chose 500 meters radius: Observing the map we see that some neighborhoods are very close each other, there may be venues overlapping at greater distance, so 500 mts seems to be a reasonable value to consider and prevent it a bit.

Showing up next, extracts of the dataframes built adding venues data to the previous location dataframes in both cities.

Let's check the size of the resulting dataframe

In [55]:
```
print(san_francisco_venues.shape)
san_francisco_venues.head(70)
```

```
(3699, 7)
```

Out[55]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alamo Square	37.77722	-122.43146	Painted Ladies	37.776120	-122.433389	Historic Site
1	Alamo Square	37.77722	-122.43146	Originals Vinyl	37.775835	-122.431227	Record Shop
2	Alamo Square	37.77722	-122.43146	Alamo Square	37.775881	-122.434412	Park
3	Alamo Square	37.77722	-122.43146	Church of 8 Wheels	37.774733	-122.430862	Roller Rink
4	Alamo Square	37.77722	-122.43146	The Center SF	37.774545	-122.430730	Spiritual Center
5	Alamo Square	37.77722	-122.43146	Lady Falcon Coffee Club	37.777255	-122.433998	Food Truck
6	Alamo Square	37.77722	-122.43146	Kebab King	37.779786	-122.431589	Pakistani Restaurant
7	Alamo Square	37.77722	-122.43146	Alamo Square Dog Park	37.775878	-122.435740	Dog Run
8	Alamo Square	37.77722	-122.43146	Suppenküche	37.776324	-122.426382	German Restaurant
9	Alamo Square	37.77722	-122.43146	Salt & Straw	37.776532	-122.426051	Ice Cream Shop

Image 6. Venues in the neighborhoods of San Francisco.

One more line of code shows **There are 342 unique venues**.

Let's check the size of the resulting dataframe

In [59]:
```
print(seattle_venues.shape)
seattle_venues.head(70)
```

```
(2610, 7)
```

Out[59]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	North Seattle	47.643724	-122.302937	Cafe Lago	47.639698	-122.302256	Italian Restaurant
1	North Seattle	47.643724	-122.302937	Montlake Cut	47.647094	-122.304686	Canal
2	North Seattle	47.643724	-122.302937	Seattle Public Library - Montlake	47.640520	-122.302413	Library
3	North Seattle	47.643724	-122.302937	Fuel Coffee - Montlake	47.639688	-122.302009	Coffee Shop
4	North Seattle	47.643724	-122.302937	Montlake Bicycle Shop	47.639380	-122.302340	Bike Shop
5	North Seattle	47.643724	-122.302937	Montlake Blvd Market	47.643480	-122.303915	Grocery Store
6	North Seattle	47.643724	-122.302937	Traveler Montlake	47.639830	-122.302231	American Restaurant
7	North Seattle	47.643724	-122.302937	Metro Bus Stop #25751	47.644848	-122.304488	Bus Stop
8	North Seattle	47.643724	-122.302937	Metro Bus Stop #71344	47.644555	-122.302720	Bus Stop
9	North Seattle	47.643724	-122.302937	King County Metro Bus Route 255	47.642409	-122.303858	Bus Line

Image 7. Venues in the neighborhoods of Seattle.

One more line of code shows **There are 300 unique venues**.

3.2.3 STEP 3: Normalization and grouping by

It is required that we prepare the data in an appropriate way before fitting a machine learning model.

First: we group the venue categories by neighborhood. Second: Since venue categories are text values, they are categorical data. Many machine learning algorithms cannot operate with this type of data directly. They require all input variables and output variables to be numeric. This means that categorical data must be converted to a numerical type.

For this purpose, we create dummy variables using **One-Hot Encoding**, where a binary variable is added for each value.

```
san_francisco_onehot.head()
```

Out[63]:

	Zoo Exhibit	ATM	Acai House	Accessories Store	Acupuncturist	Adult Boutique	African Restaurant	Alternative Healer	American Restaurant	Antique Shop	...	Waterfront	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Win
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

5 rows × 342 columns

And let's examine the new dataframe size.

In [64]:
```
san_francisco_onehot.shape
```

Out[64]: (3699, 342)

Image 8. One-hot encoding in San Francisco venues dataframe.

Next we can do; let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category. These values represent the percentages of appearance of a specific venue category in the total of every neighborhood.

For example, only showing top 5 in SAN FRANCISCO in the two first neighborhoods, we can see the following most common venue categories:

```
----Alamo Square----
                venue  freq
0                Park  0.15
1             Dog Run  0.08
2   German Restaurant  0.04
3      Ice Cream Shop  0.04
4            Bus Line  0.04


----Anza Vista----
             venue  freq
0             Café  0.17
1      Coffee Shop  0.11
2   Cosmetics Shop  0.06
3         Bus Line  0.06
4         Bus Stop  0.06
```

Image 9. Percentages of appearance of categories in the neighborhoods of San Francisco.

This means that in Alamo Square, 15% of the venues are Parks.

Operating in a similar way for SEATTLE, we can see the following most common venue categories:

```
----Adams----
                  venue  freq
0          Burger Joint  0.08
1           Coffee Shop  0.08
2                Bakery  0.05
3  Performing Arts Venue  0.05
4        Thai Restaurant  0.05


----Alki Point----
              venue  freq
0     Scenic Lookout  0.50
1  Convenience Store  0.17
2               Park  0.17
3        Coffee Shop  0.17
4        Yoga Studio  0.00
```

Image 10. Percentages of appearance of categories in the neighborhoods of Seattle.

This means that in Adams, we have 8% for Burger Joints venues and 8% for Coffee Shops.

Now, for further studies and calculations we write a function to sort the most common venues in descending order and compile all the info in a new dataframe. Presented below, extracts from the dataframes displaying **the top 10 venues for every neighborhood**.

```
neighborhoods_venues_sorted.head()
```

Out[69]:

	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
0	Alamo Square	Park	Dog Run	Playground	Coffee Shop	Spiritual Center	Food Truck	French Restaurant	Bus Line	Sushi Restaurant	Café
1	Anza Vista	Café	Coffee Shop	Big Box Store	Tunnel	Sandwich Place	Pet Store	Bus Stop	Health & Beauty Service	Donut Shop	Cosmetics Shop
2	Ashbury Heights	Trail	Breakfast Spot	Wine Bar	Convenience Store	Coffee Shop	Organic Grocery	Toy / Game Store	Restaurant	Bakery	Bar
3	Balboa Hollow	Chinese Restaurant	Café	Japanese Restaurant	Bakery	Bus Station	Sporting Goods Shop	Pizza Place	Flower Shop	Vietnamese Restaurant	Dessert Shop
4	Balboa Terrace	Yoga Studio	Comic Shop	Baseball Field	Gym	Pharmacy	Light Rail Station	Park	Vietnamese Restaurant	Fountain	Playground

Image 11. Venue categories top 10 in the neighborhoods for SF.

```
neighborhoods_venues2_sorted.head()
```

Out[75]:

	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
0	Adams	Coffee Shop	Burger Joint	Performing Arts Venue	Bakery	Thai Restaurant	Sri Lankan Restaurant	Candy Store	Supermarket	Drugstore	Korean Restaurant
1	Alki Point	Scenic Lookout	Coffee Shop	Park	Convenience Store	Field	Event Space	Eye Doctor	Falafel Restaurant	Farmers Market	Fast Food Restaurant
2	Arbor Heights	Spa	Women's Store	Filipino Restaurant	Event Space	Eye Doctor	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Field	Financial or Legal Service
3	Atlantic	Vietnamese Restaurant	Coffee Shop	Intersection	Gym	Burrito Place	Plaza	Skate Park	Seafood Restaurant	Sandwich Place	Bank
4	Ballard	Mexican Restaurant	Cocktail Bar	Coffee Shop	Ice Cream Shop	Sushi Restaurant	Gym	Bar	Thai Restaurant	Dessert Shop	Sandwich Place

Image 12. Venue categories top 10 in the neighborhoods for SE.

3.2.4 STEP 4: Clustering through a ML model

To identify similarities, we need to group the neighborhoods into clusters, based on similarities of venue categories. To be able to do that, we use the **k-means algorithm** to cluster data. It is a form of unsupervised machine learning clustering algorithm. The k-means identifies k number of centroids, then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and most popular unsupervised ML algorithms and is highly suited for this project.

We have a little step to solve previously: determine the optimal number of clusters (k) for the model using **'elbow method'**. This method measures Inertia vs. k / SSE vs. k / Distortion vs. k, which are representations of how may vary the centroids as the number of clusters varies. The optimal k values is when the variable of the ordinate axis changes a small value with respect to the increase of k.

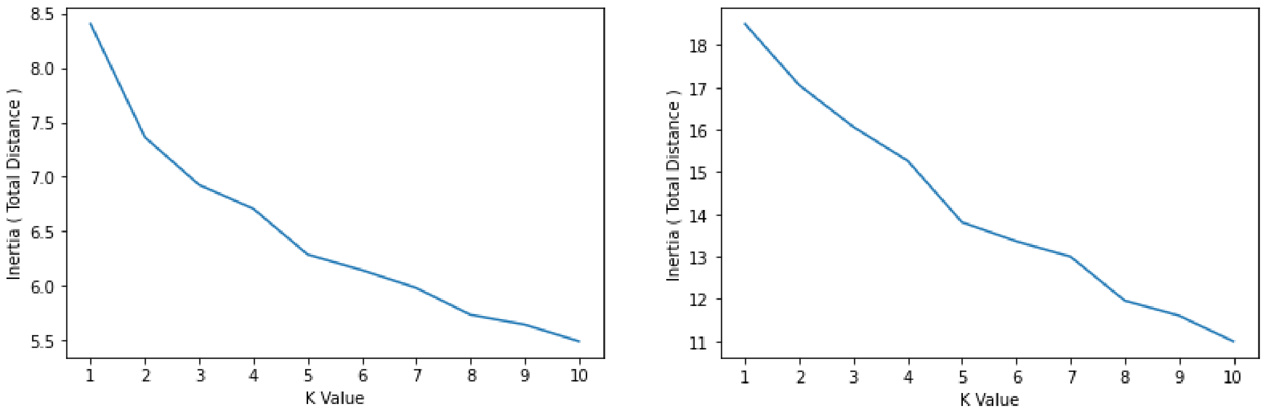


Image 13. Inertia vs. k values in SF dataframe (left) and in SE dataframe (right).

These results show us that reasonable values are:

- k=6 for SAN FRANCISCO
- k =6 for SEATTLE

We run the k-means algorithm with the selected k-values and obtain cluster identification for each neighborhood.

Now we can create new dataframes, including the cluster labels as well as the top 10 venue categories for each neighborhood. We know which cluster each neighborhood belongs to.

```
san_francisco_merged.head() # check the last columns!
```

Out[86]:

Neighborhood	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 Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Alamo Square	37.77722	-122.43146	3	Park	Dog Run	Playground	Coffee Shop	Spiritual Center	Food Truck	French Restaurant	Bus Line	Sushi Restaurant	Café
Anza Vista	37.78048	-122.44358	0	Café	Coffee Shop	Big Box Store	Tunnel	Sandwich Place	Pet Store	Bus Stop	Health & Beauty Service	Donut Shop	Cosmetics Shop
Ashbury Heights	37.76487	-122.44590	0	Trail	Breakfast Spot	Wine Bar	Convenience Store	Coffee Shop	Organic Grocery	Toy / Game Store	Restaurant	Bakery	Bar
Balboa Hollow	37.77589	-122.49360	0	Chinese Restaurant	Café	Japanese Restaurant	Bakery	Bus Station	Sporting Goods Shop	Pizza Place	Flower Shop	Vietnamese Restaurant	Dessert Shop
Balboa Terrace	37.73180	-122.46740	3	Yoga Studio	Comic Shop	Baseball Field	Gym	Pharmacy	Light Rail Station	Park	Vietnamese Restaurant	Fountain	Playground

Image 14. SF dataframe with Cluster Labels included.

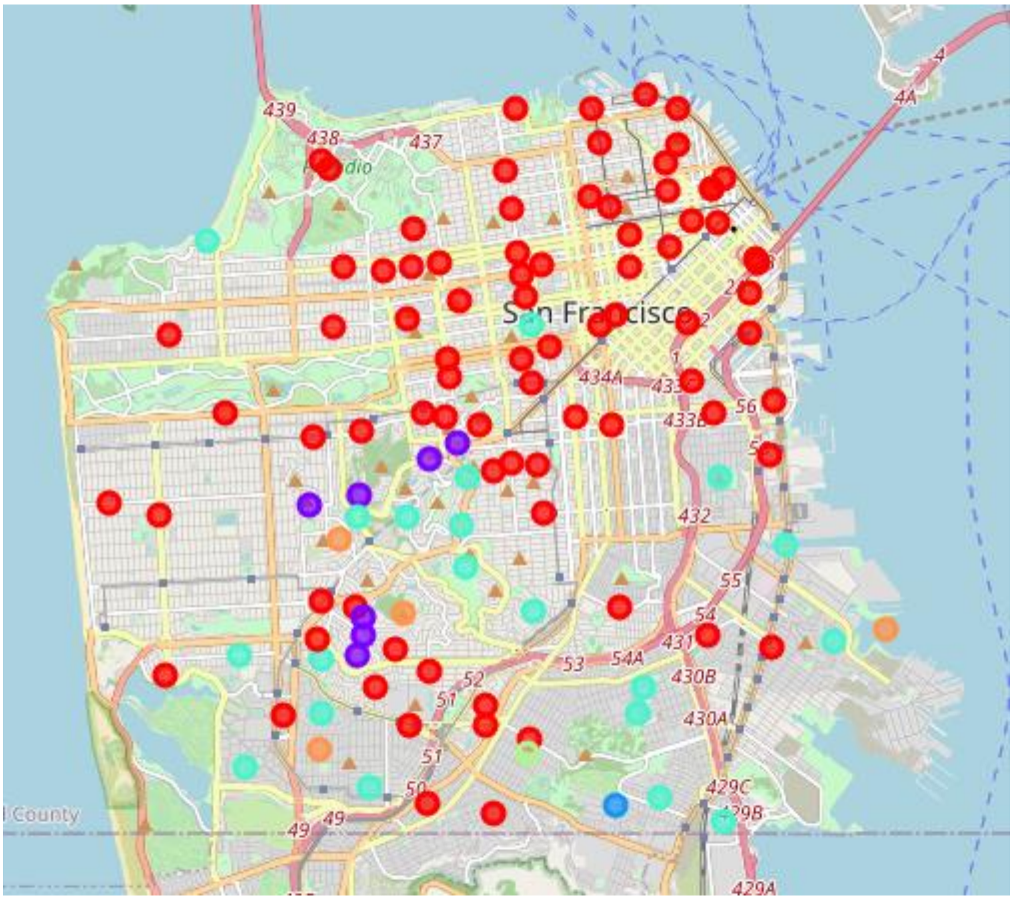
```
seattle_merged.head() # check the last columns!
```

Out[91]:

Neighborhood	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 Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mos Commo Venu
North Seattle	47.643724	-122.302937	0	Bus Stop	Salon / Barbershop	Library	Grocery Store	Canal	Trail	Coffee Shop	Harbor / Marina	Bike Shop	Par
Broadview	47.722380	-122.364980	3	Trail	Concert Hall	Women's Store	Field	Ethiopian Restaurant	Event Space	Eye Doctor	Falafel Restaurant	Farmers Market	Fast Foo Restaurar
Bitter Lake	47.718680	-122.350300	0	Marijuana Dispensary	Automotive Shop	Intersection	Hardware Store	Sandwich Place	Sushi Restaurant	Thai Restaurant	Beer Bar	Donut Shop	Steakhous
North Beach / Blue Ridge	47.700440	-122.384180	1	Garden Center	Photography Studio	Café	Park	Flea Market	Fish Market	Fish & Chips Shop	Financial or Legal Service	Filipino Restaurant	Ethiopia Restaurar
Crown Hill	47.695200	-122.374100	0	Coffee Shop	Pizza Place	Sports Bar	Grocery Store	Pet Store	Rock Club	Sandwich Place	Fast Food Restaurant	Bus Station	Burge Joir

Image 15. SE dataframe with Cluster Labels included.

Finally, let's visualize the resulting clusters:



➔ Cluster 0
➔ Cluster 1
➔ Cluster 2
➔ Cluster 3
➔ Cluster 4
➔ Cluster 5

Image 16. Cluster distribution in San Francisco.

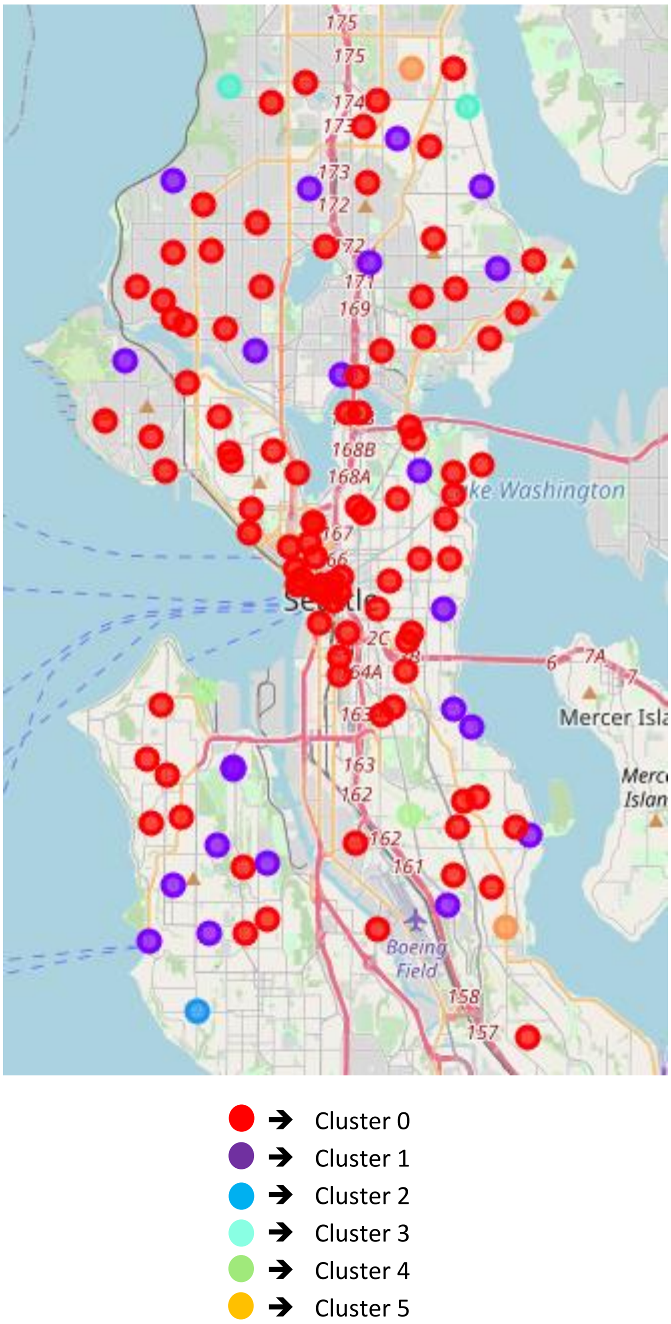


Image 17. Cluster distributions in Seattle.

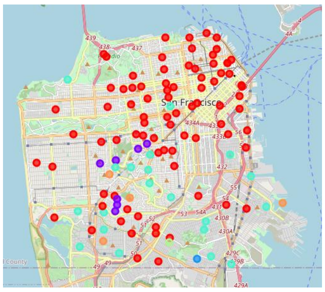
3.2.5 STEP 5: Numerical & Examination

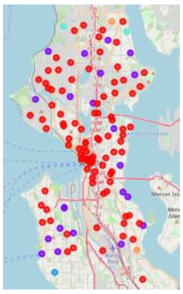
In this last step, counting on all the previous work, we can examine, calculate and plot useful data to make the final evaluations.

We examine each cluster in both cities and determine the discriminating venue categories that distinguish each one. Based on the defining categories, we evaluate according the factors established at the beginning of this report in **2.1 Data Description.**

Cluster 0

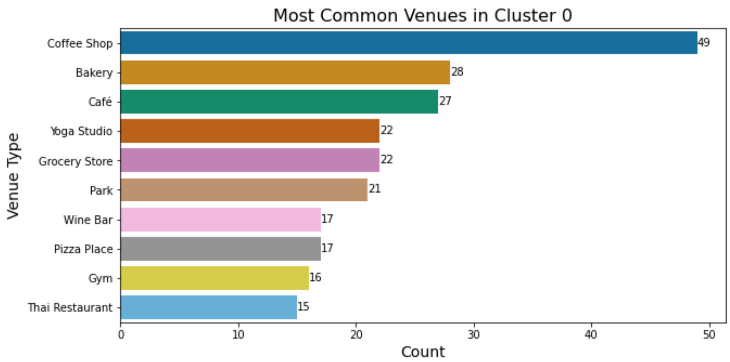
It is the biggest one in both cities. It is identified by the red circle. Predominates **around the entire city** but is **tending to the centre area in both cities**. Below the top 10 venue categories of cluster 0 and the amounts of each.





SAN FRANCISCO

SEATTLE



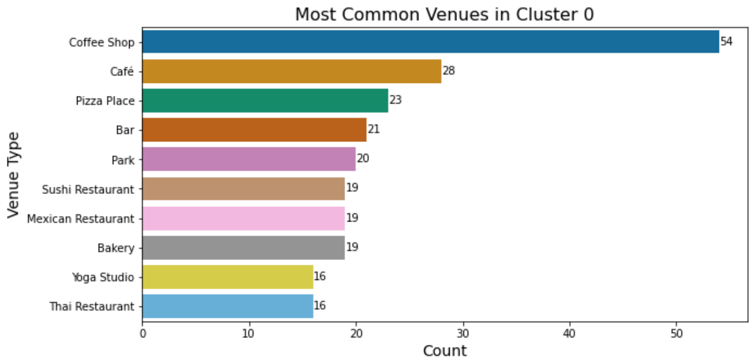


Image 18. Cluster 0 venues distribution.

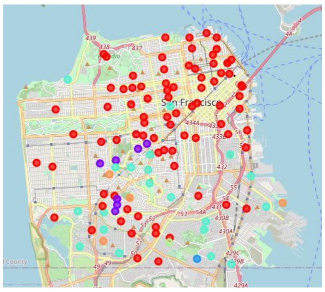
Includes 94 neighborhoods for San Francisco, counting 234 venues in the top 10.

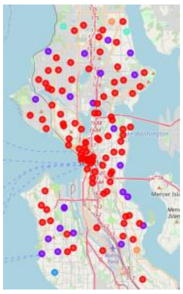
Includes 94 neighborhoods for Seattle, counting 235 venues in the top 10.

How we can see, it is mainly constituted by food venues. **It has 28 bakeries for San Francisco and 19 for Seattle**. They **also have a lot of cafés**, which we consider in this work as probable competitors, since sometimes they sell bakery products too.

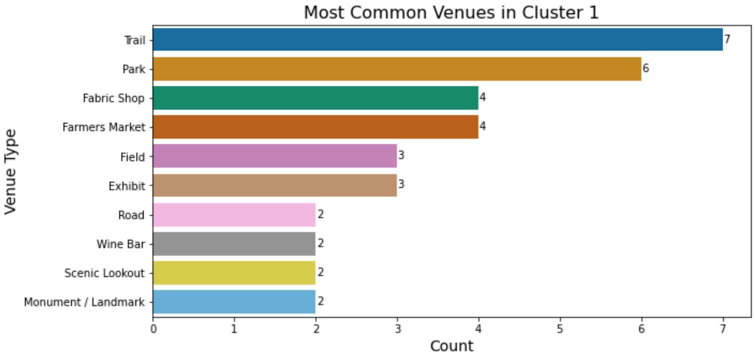
Cluster 1

It is identified by the purple circle. It is located the **middle side in San Francisco** and **around the entire city in Seattle** Below the top 10 venue categories of cluster 1 and the amounts of each.





SAN FRANCISCO



SEATTLE

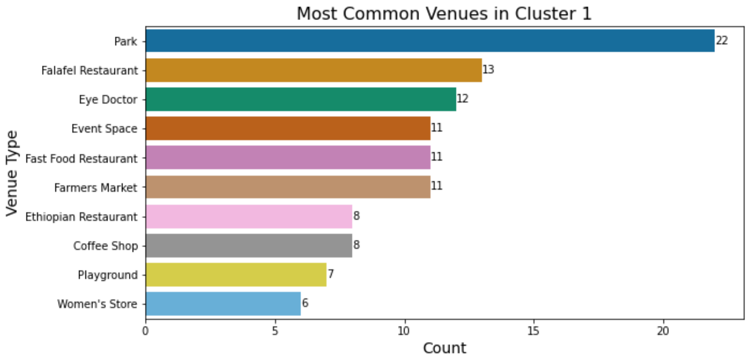


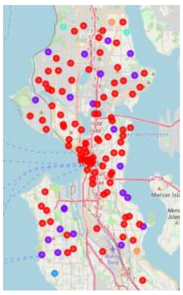
Image 19. Cluster 1 venues distribution.

Includes 7 neighborhoods for San Francisco, counting 35 venues in the top 10. **It has not competitors**.

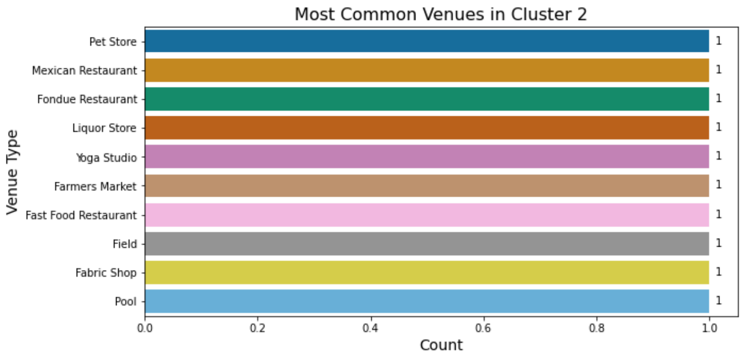
Includes 22 neighborhoods for Seattle, counting 109 venues in the top 10. **It has 8 Coffe Shops**, which as we said, are probable competitors too.

Cluster 2

It is identified by the light blue circle. It is located the **south-east side in San Francisco** and in the **south-west in Seattle.** Below the top 10 venue categories of cluster 2 and the amounts of each.



SAN FRANCISCO



SEATTLE

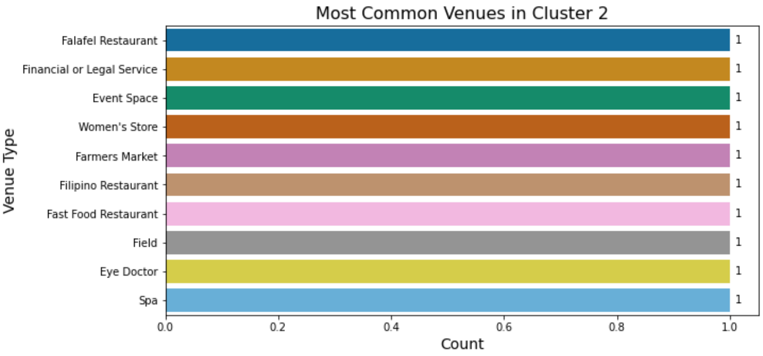


Image 20. Cluster 2 venues distribution.

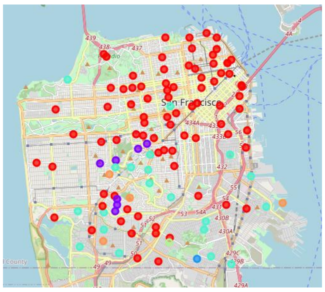
Includes 1 neighborhood for San Francisco, counting 10 venues in the top 10.

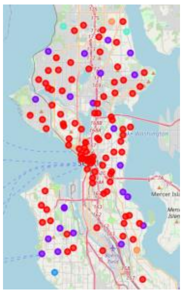
Includes 1 neighborhoods for Seattle, counting 10 venues in the top 10.

This cluster **has not competitors, in both cities**.

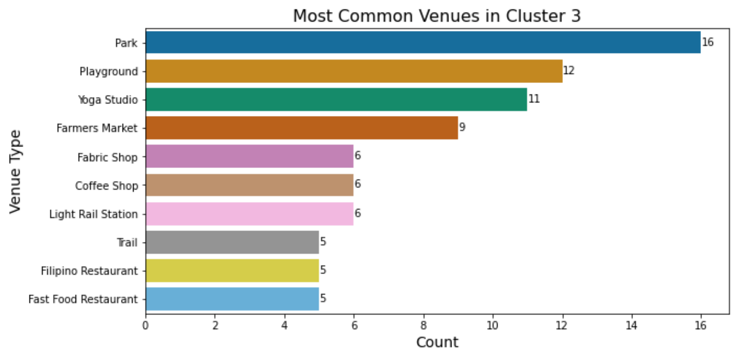
Cluster 3

It is identified by the cyan circle. It is located approximately in the **middle-to-east side in San Francisco** and in the **north side in Seattle.** Below the top 10 venue categories of cluster 3 and the amounts of each.





SAN FRANCISCO



SEATTLE

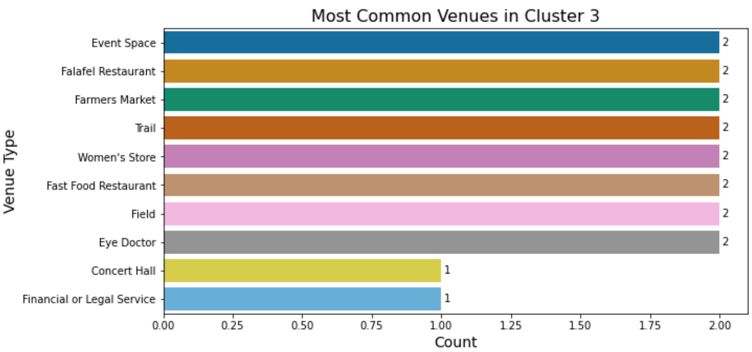


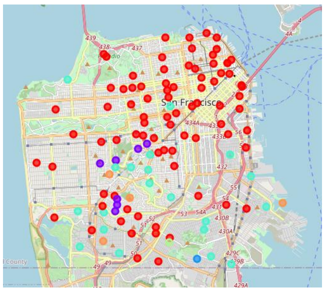
Image 21. Cluster 3 venues distribution.

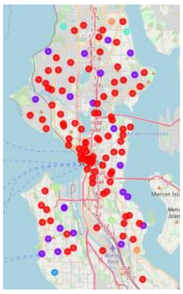
Includes 20 neighborhoods for San Francisco, counting 81 venues in the top 10. **It has 6 Coffee Shop**, which as we said, are probable competitors too.

Includes 2 neighborhoods for Seattle, counting 18 venues in the top 10. **It has not competitors**.

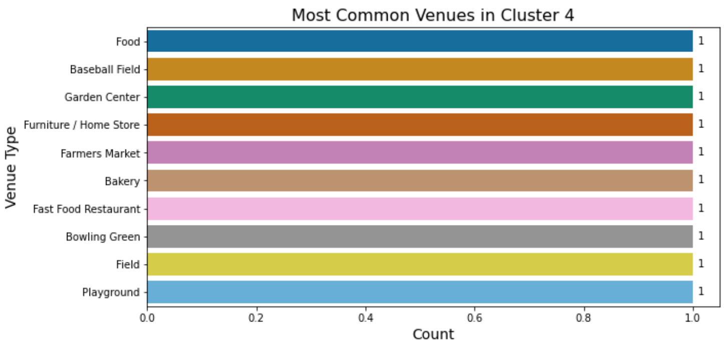
Cluster 4

It is identified by the light green circle. It is in the **south side in San Francisco** and approximately in the **middle side in Seattle**. Below the top 10 venue categories of cluster 4 and the amounts of each.





SAN FRANCISCO



SEATTLE

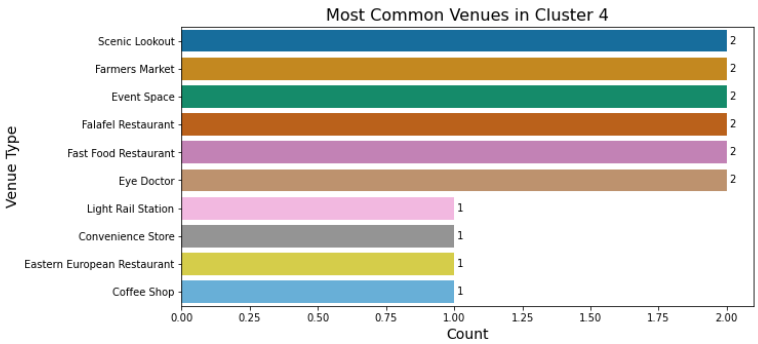


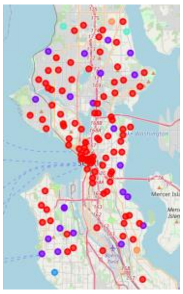
Image 22. Cluster 4 venues distribution.

Includes 1 neighborhood for San Francisco, counting 10 venues in the top 10. **It has 1 Bakery, which is a direct competitor.**

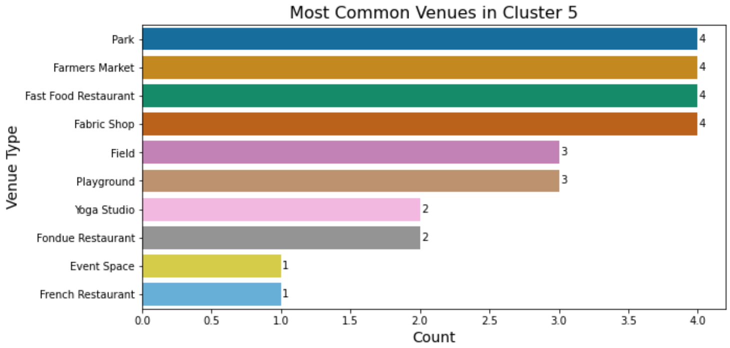
Includes 2 neighborhoods for Seattle, counting 16 venues in the top 10. **It has 1 Coffe Shop**, which as we said, is a probable competitor too.

Cluster 5

It is identified by the orange circle. It is approximately in the **middle-to-south side in San Francisco** and, **1-north-east, 1-south-east sides in Seattle.** Below the top 10 venue categories of cluster 5 and the amounts of each.



SAN FRANCISCO



SEATTLE

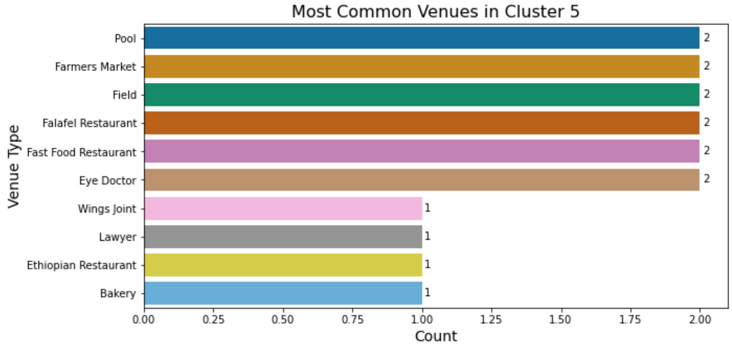


Image 23. Cluster 5 venues distribution.

Includes 4 neighborhoods for San Francisco, counting 28 venues in the top 10. **It has not competitors.**

Includes 2 neighborhoods for Seattle, counting 16 venues in the top 10. **It has 1 Bakery, which is a direct competitor.**

4. RESULTS

In this section we are going to do a summary about what was discovered along the study in the STEPS.

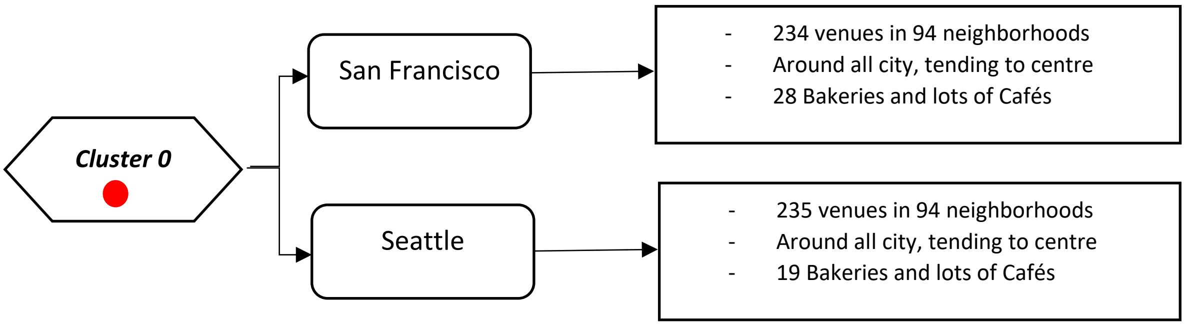
We found the top populated US cities with better per capita income, and then we stay with the 2 best: SAN FRANCISCO and SEATTLE. It can be extended to more cities, but we limited the project by time reasons.

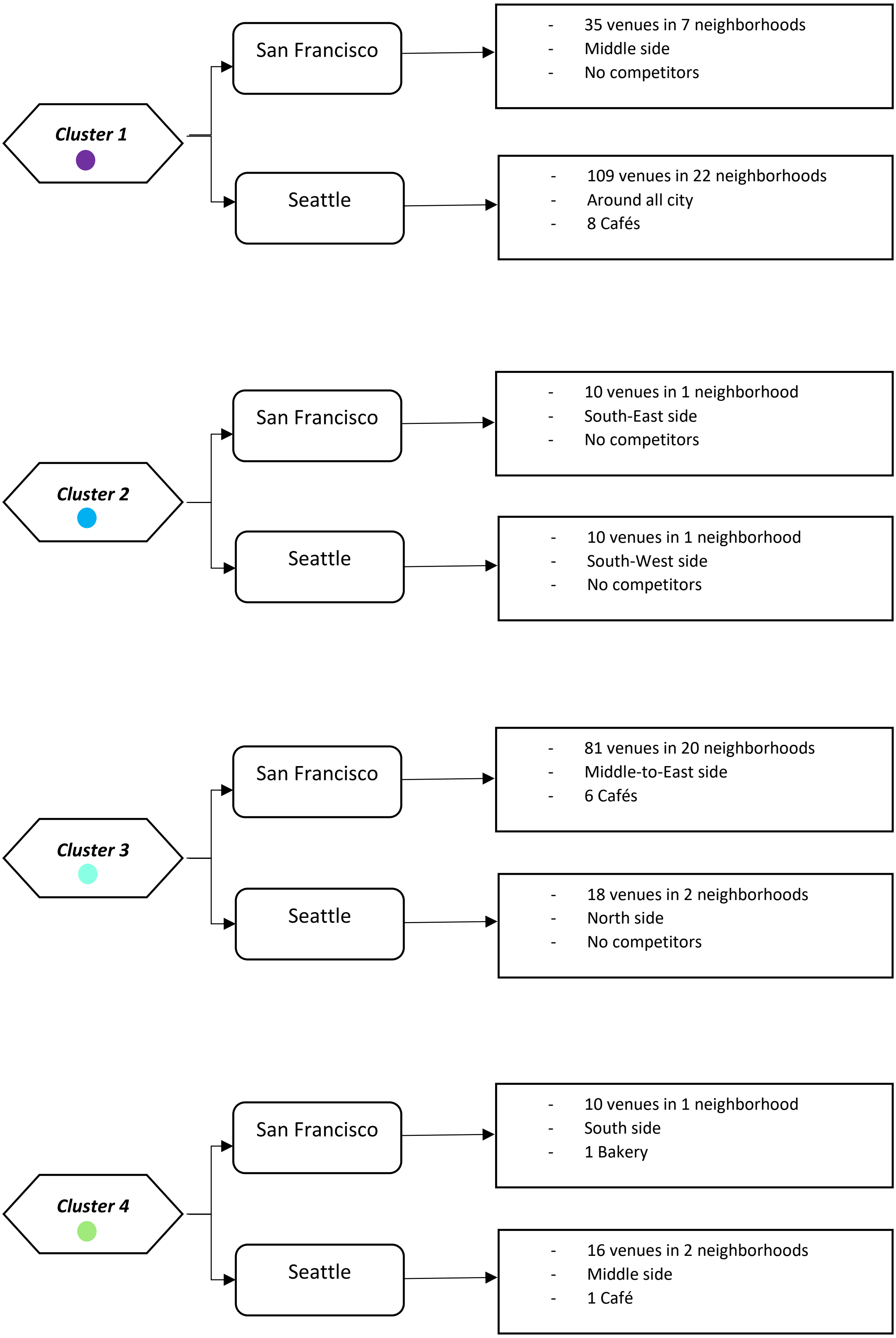
We found the neighborhoods and its coordinates for both cities.

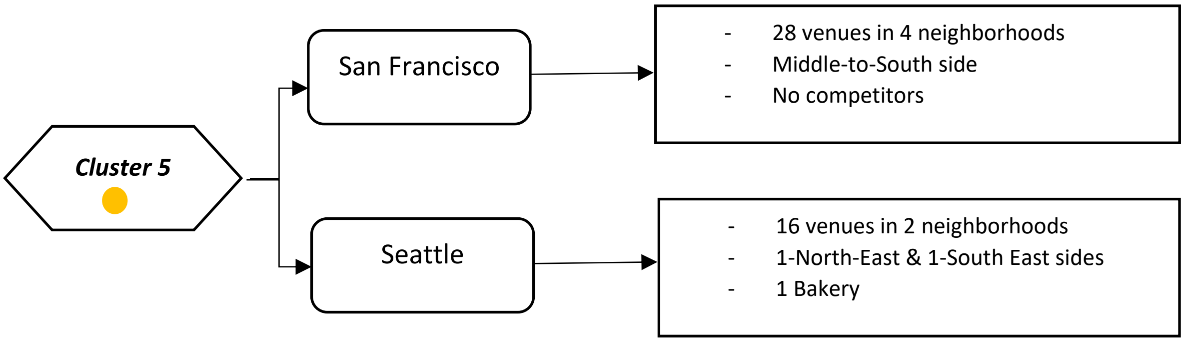
We explored and organized the most common venues to evaluate if there are potential competing food venues; chose the top 10 in every neighborhood. Besides, are in the neighborhoods any bakeries or coffee shops?

We classified the neighborhoods in k groups (clusters) according its venues similarities through a ML model. We chose k=6 as the optimal value for both cities. This means, we have grouped the neighborhoods into 6 clusters.

Based on the defining categories, we can evaluate according the factors established at the beginning of this report in **2.1 Data Description.**




Cluster 1
San Francisco
- 35 venues in 7 neighborhoods
- Middle side
- No competitors
Seattle
- 109 venues in 22 neighborhoods
- Around all city
- 8 Cafés
Cluster 2
San Francisco
- 10 venues in 1 neighborhood
- South-East side
- No competitors
Seattle
- 10 venues in 1 neighborhood
- South-West side
- No competitors
Cluster 3
San Francisco
- 81 venues in 20 neighborhoods
- Middle-to-East side
- 6 Cafés
Seattle
- 18 venues in 2 neighborhoods
- North side
- No competitors
Cluster 4
San Francisco
- 10 venues in 1 neighborhood
- South side
- 1 Bakery
Seattle
- 16 venues in 2 neighborhoods
- Middle side
- 1 Café



5. DISCUSSION

Interpreting the results showed in plots, flowcharts and observing the maps; we can compare them with the decision factors established at the beginning. We are able to discuss about pros and cons and recommend the best for the investor.

SAN FRANCISCO

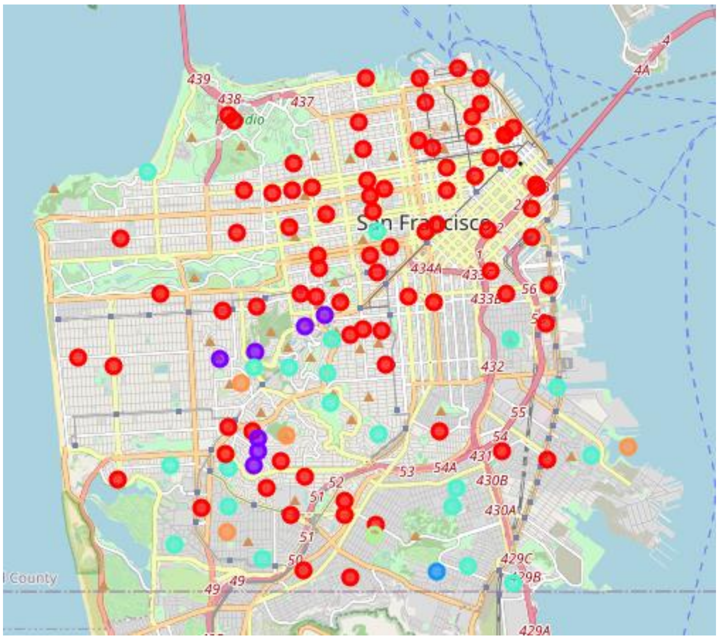


Image 24. Cluster distribution in San Francisco.

According to our criteria, we should **discard Cluster 0** neighborhoods (reds).

Cluster 1 (purples) are **neighborhoods of interest** since they have no competitors and seems to be close enough of city centre.

We discard **Cluster 2 (light blue)** in this approach since it's "far" from city centre.

Cluster 3 (cyan) neighborhoods are interesting too since are very close from city centre and only have 6 cafés, wich we considered indirect competitors. We should analyze in wich neighborhoods are located.

We discard **Cluster 4 (light green)** since it's "far" from city centre and already has 1 bakery.

We discard **Cluster 5 (orange)** in this approach since it's far from city centre.

SEATTLE

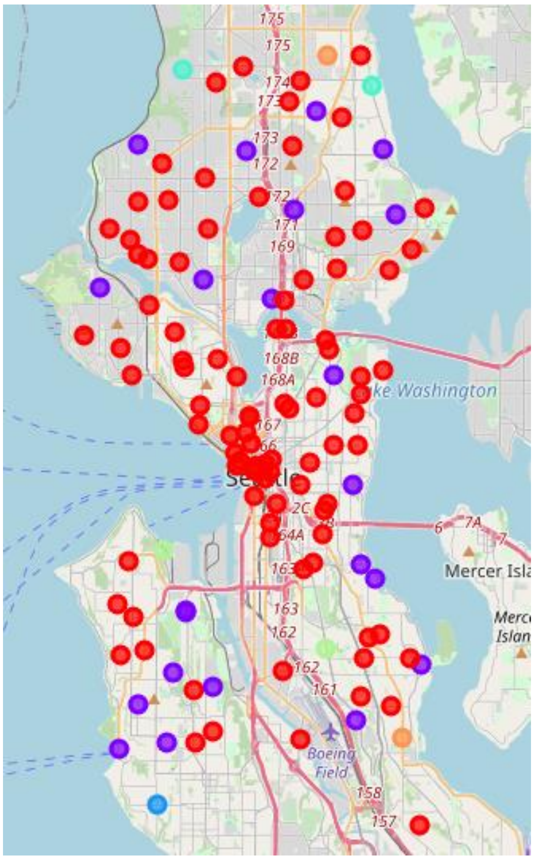


Image 25. Cluster distribution in Seattle.

According to our criteria, we should **discard Cluster 0** neighborhoods (reds).

There are a few neighborhoods of **Cluster 1 (purples) which are interesting** since they are very close from city centre and have only 8 cafés, which we considered indirect competitors. We should analyze in which neighborhoods are located.

We discard **Cluster 2 (light blue)** in this approach since it's far from city centre.

We discard **Cluster 3 (cyan)** in this approach since it's far from city centre.

We could **consider Cluster 4 (light green) of interest** since it's "close" from city centre and has only 1 café.

We discard **Cluster 5 (orange)** in this approach since it's far from city centre.

We are going to make our FINAL CONCLUSION in the next section.

6. CONCLUSSION

We finally decided to stay with **SAN FRANCISCO** city since it has $10 K dollars more in per capita income compared to SEATTLE. This data could be crossed with real estate costs for a better decision.

Then, focusing in SF we stay with the **clusters 1 and 3** according the previous section. And we especially look at the neighborhoods indicated by arrows.

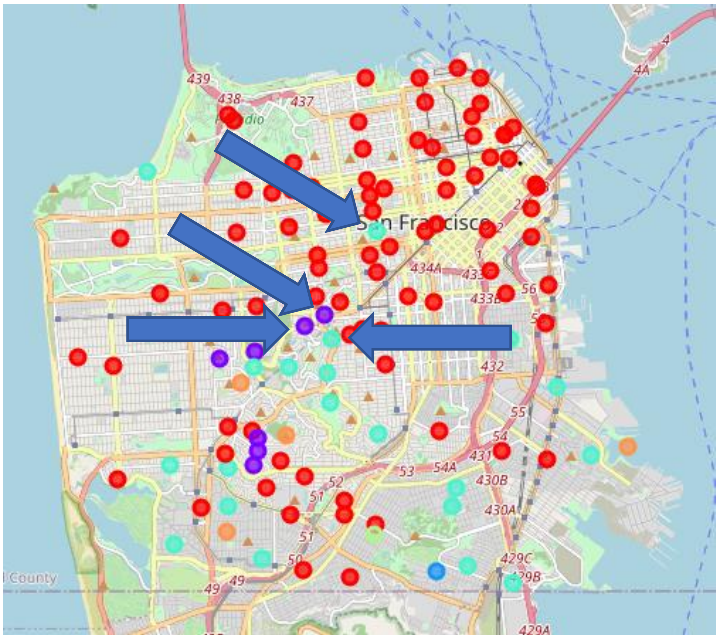


Image 26. Neighborhoods of interest in SF.

These selected neighborhoods are**: 1. Alamo Square, 2. Upper Market, 3. Clarendon Heights and 4. Twin Peaks.**

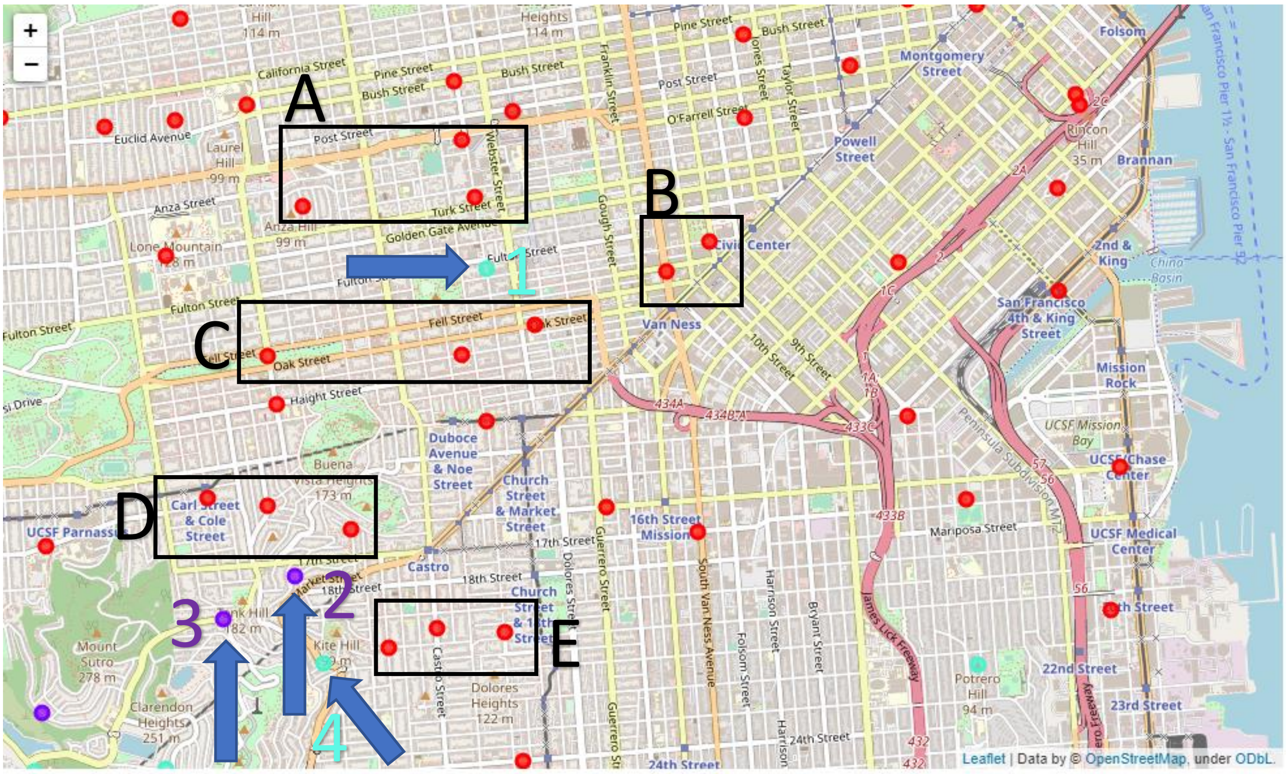


Image 27. Neighborhoods of interest and Cluster 0 neighborhoods in rectangles.

Without further studies we would recommend to set the Bakery in **1. Alamo Square** or in **2. Upper Market** because there are closer to the city centre.

But then, we decided to make a final close look up: Why not to check if any bakeries or coffe shops in the neighborhoods within the rectangles? There are 14 neighborhoods surrounding our points of interest in those rectangles, we can quickly check them.

A) The Fillmore, The Western Addition, Anza Vista

B) Butchertown (Old and New), Civic Center

C) Hayes Valley, The Lower Haight, North of Panhandle

D) Cole Valley, Ashbury Heights, Corona Heights

E) Eureka Valley, The Castro, Dolores Heights

```
aa = san_francisco_venues[san_francisco_venues['Neighborhood']=='The Fillmore']
aa[aa['Venue Category'].str.contains("Bakery|Café|Coffee Shop")]
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1094	The Fillmore	37.78392	-122.43312	Jane the Bakery	37.783797	-122.434283	Bakery

```
ab = san_francisco_venues[san_francisco_venues['Neighborhood']=='The Western Addition']
ab[ab['Venue Category'].str.contains("Bakery|Café|Coffee Shop")]
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
3607	The Western Addition	37.78095	-122.43222	Jane the Bakery	37.783797	-122.434283	Bakery

```
ac = san_francisco_venues[san_francisco_venues['Neighborhood']=='Anza Vista']
ac[ac['Venue Category'].str.contains("Bakery|Café|Coffee Shop")]
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
27	Anza Vista	37.78048	-122.44358	Matching Half Cafe	37.777229	-122.441433	Café
28	Anza Vista	37.78048	-122.44358	Black/Jasmine	37.777526	-122.443268	Coffee Shop
32	Anza Vista	37.78048	-122.44358	Opa Cafe	37.784001	-122.441494	Café
38	Anza Vista	37.78048	-122.44358	Mo'z Cafe	37.782169	-122.447856	Café
40	Anza Vista	37.78048	-122.44358	Starbucks	37.782074	-122.447080	Coffee Shop

Image 28. Rectangle A neighborhoods venues check.

Repeating the same code for neighborhoods in B, C and D as you can see in the main Jupyter notebook. We obtained the following data:

A) 2 Bakeries, 3 Cafés and 2 Coffee Shops
B) 1 Bakery, 2 Cafés and 5 Coffee Shops
C) 2 Bakeries, 4 Cafés and 6 Coffee Shops
D) 2 Bakeries, 2 Cafés and 4 Coffee Shops
E) 1 Bakery, 1 Café and 8 Coffe Shops

FINAL CONCLUSION

We see zones A, B and C are crowded of competitors, so we are going to avoid them at the expense of getting away from the city centre. Discarding neighborhood 1.

We decide finally recommend to the investor the **neighborhood 4 : Twin Peaks**, since is relative at the same distance from the centre than neighborhoods 2 and 3; and besides, the closer rectangle is E wich has only one Bakery and less competitors in general.

Beyond:

We are aware about the limitations of this study, it was made this way in order to simplify certain information and narrow down searches. We consider it is still a good approach for the investor.

Perhaps in a deeper analysis we will have to consider expanding to more cities under study and prefer the neighborhoods where there is a lower real estate cost. Besides, to make focus only in bakeries and calculate their distances to city center and distances between each other. May another analysis could also include which neighborhoods people are most likely to spend in food. For simplicity and time reasons, these points were not considered in this work.

If you got this far, thank you for reading the paper.

Federico Sarrailh

July 27th, 2021

Córdoba, ARGENTINA