# IBM Data Science Capstone

*Opening a new fast food joint in CABA, Argentina*

Joaquin Minvielle

2020



Contents

[IBM Data Science Capstone 1](#_Toc39829530)

[1. Introduction 2](#_Toc39829531)

[2. Data 2](#_Toc39829532)

## 1. Introduction

There is a new trend in Ciudad de Buenos Aires (CABA), Argentina, and that is to open your own fast-food joint to compete against the well-known franchises. The problem is that there are so many fast-food joints that the market seems to be saturated. This means that if you want to open a new joint and survive in this very competitive market, you must have a competitive advantage, like a good location.

For my final Data Science Capstone, I will analyze CABA and choose the best neighborhood to open a new fast-food joint using a clustering algorithm. The results from the analyses could be of help for anyone trying to open a new joint.

This analysis will be useful for everyone who wants to start their own fast food joint in the near future and doesn’t know where to locate it.

## 2. Data

To do my analysis I will use data found in Wikipedia with all the neighborhoods from CABA. The information will be added to a Pandas DataFrame. The information can be found on the following link:

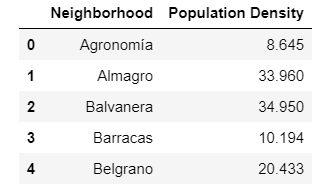
<https://es.wikipedia.org/wiki/Anexo:Barrios_de_la_ciudad_de_Buenos_Aires>



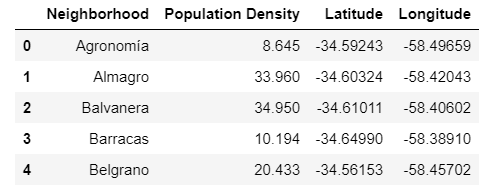
This table gives the names of the neighborhoods and the populational density. The names will be used to locate all the venues in CABA using Foursquare. The populational density will be used to factor in the clusters, because the higher the density the higher the chance of having more customers.

## 3. Methodology

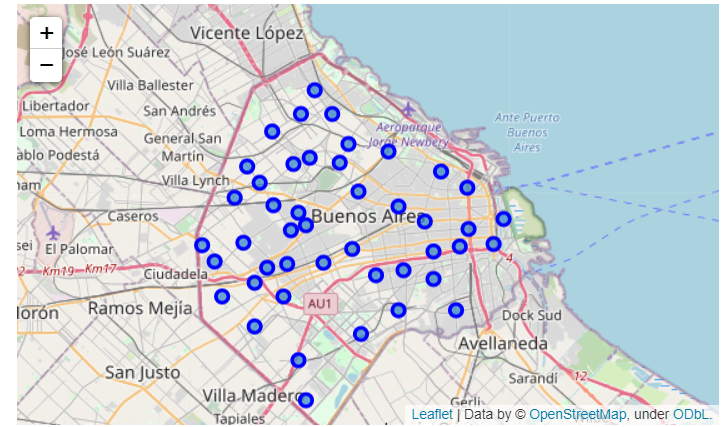
After creating the data frame from the Wikipedia table, I cleaned it by taking out the unnecessary columns and changing the names of the remaining columns.



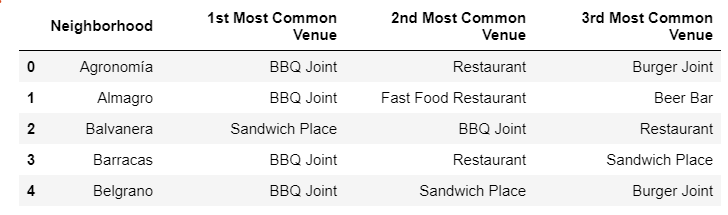
Then, I used the Geocoder library to get the latitudes and longitudes for each of the neighborhoods:



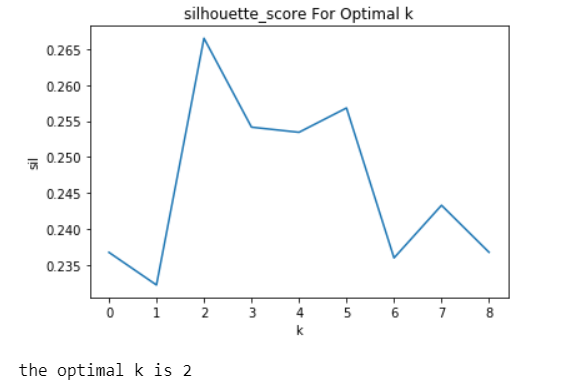
With the coordinates of the neighborhoods I was able to plot them on a geographic using the Folium library.



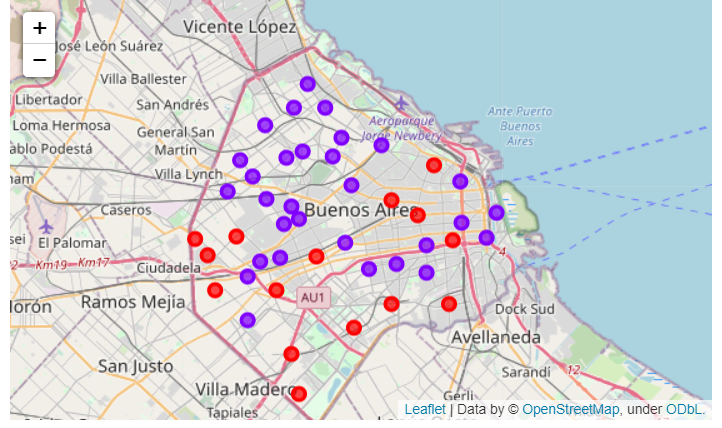
Nest, using the Foursquare API I created a data frame with the venues of each neighborhood. I limited to 100 venues per neighborhood and a radius of 1500 meters. The values were converted to their discrete values, and I used the means to determine the frequency in which eac venue repted itself on a neighborhood. Then I just kept the 8 most relevant venues for this type of analysis.



With this data I clustered the data using KMeans. Using the Silhouette Score I determine K=2 as the best value.



Finally I ploted the 2 clusters:



## 4. Results

There are just two clusters:

Cluster 1 has a more varied mix of joints.

Cluster 2 is also varied but with more BBQ joints over any other type of joint.

## 5. Discussion

The results aren’t the best, because we don’t have many clusters, a possible reason for this is the covid-19 pandemic which may have caused some venues to shut down (the case study was made on may 2020)

If a decision had to be made right now, I would go with cluster 1, but because of the pandemic I would wait until the quarantine is over and do a new analysis.

6. Conclusion

I was able to cluster CABA into just 2 clusters and tried to identify which one is the best. The reality of the situation is that this isn’t the best time to do an analysis or to have a new joint. The analysis should be done again in a few months after the quarantine and once the economy re activates.