CLUSTERING BELO HORIZONTE'S NEIGHBORHOODS BY POPULARITY OF FOOD VENUES' CATEGORIES

Felipe Resende Nogueira | IBM Data Science Professional Certificate's Capstone Project | 05/2020

EXECUTIVE SUMMARY:

This report goes through the process of application of K-Means (a clustering machine learning algorithm), realized to cluster Belo Horizonte's neighborhoods based on the popularity of its food venues' categories. The focus of the study is providing a general data-driven approach for the clustering of neighborhoods in a city, but not necessarily on the accuracy and reliability of this specific implementation's results. Therefore, some assumptions about Belo Horizonte made throughout this particular application lack appropriate founding, nevertheless are still reasonable and acceptable. The final results of the implementation consist of lists of neighborhoods which presented potential for the establishment of a particular category of food venue.

INTRODUCTION:

Opening a food establishment requires great investment of both time and money, especially in Brazil, which is famous for its long bureaucratic processes for registering a new business. Therefore, it is prudent and expected of any entrepreneur to study the market and to search for its best opportunities before making a decision. This report was developed specially for providing entrepreneurs from Belo Horizonte, capital of Brazil's Minas Gerais state, a reliable and data-driven segmentation of the city's neighborhoods in groups formed by similarity of most popular venue categories among its food venues. This analysis is meant to deliver insights on where a particular type of food establishment would have greater chances of acceptance by the public, and also to reveal neighborhoods with untapped potential for particular establishment types.

DATA REQUIREMENTS:

In order to successfully complete the analysis, data from three different sources have been collected:

- A table with the names of all Belo Horizonte's neighborhoods with their area and population values. The table is available on Wikipedia and is based on IBGE's (Brazilian Institute of Geography and Statistics) latest census. Access is available through this link: https://pt.wikipedia.org/wiki/Lista_de_bairros_de_Belo_Horizonte;
- Coordinates for each neighborhood. These will be collected through Python's Geopy library, which provides geolocation services;
- A list with venues' names, categories, latitude and longitude within a determined radius for each neighborhood. Also, a list containing all food venue category names recognized by the Foursquare API. The venues' information will be collected by using the Foursquare API.

METHODOLOGY:

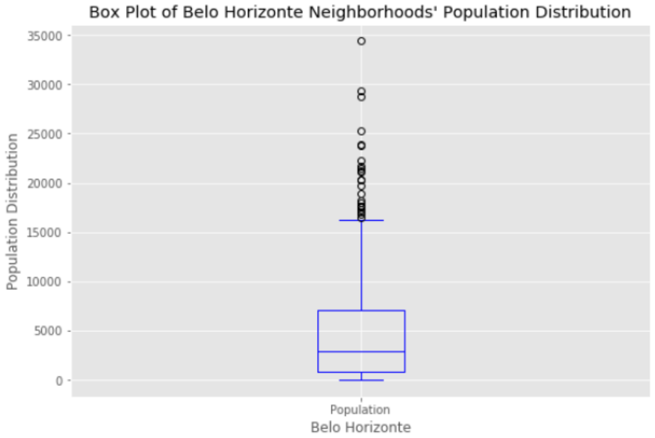
In this section, the main steps taken to reach the study's final results have been addressed in distinct sub-sections, in order to reproduce the real chronological sequence of the project's execution. All the statistical methods and assumptions made through the process are explained below, among some parts that address some of the important programming resources that made the completion of the analysis possible. The entire code can be viewed on my Github account through this link: https://github.com/FelipeNoogueira/Coursera_Capstone/blob/master/Capstone_Project.ipynb

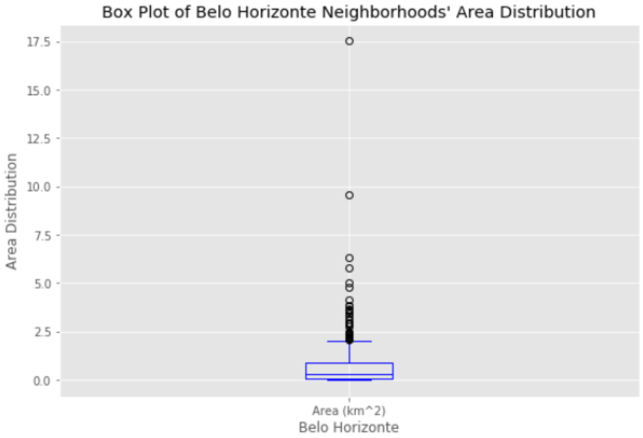
NEIGHBORHOODS' DEMOGRAPHIC AND SPATIAL DATA ACQUISITION AND TREATMENT:

The first part of the study consisted in finding a way of collecting the data from the Wikipedia page's table and converting it to a Pandas dataframe. Pandas dataframes are great for data manipulation and were of great use for filtering only the wanted neighborhoods and using them for making API requests to Foursquare's database and Geopy's location services.

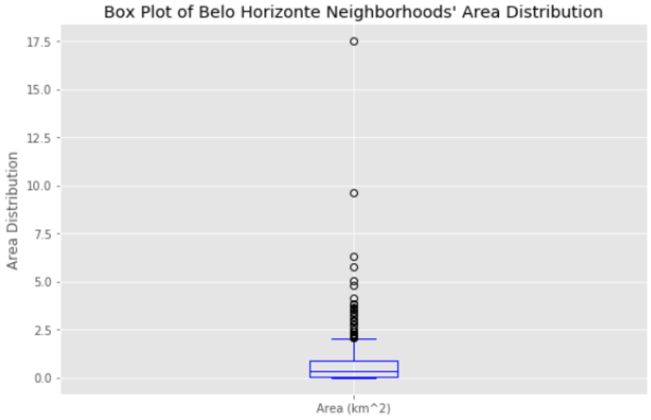
In order to obtain the table's data, a Python's library called BeautifulSoup was used to parse the HTML code and store the wanted data in a Pandas dataframe. It was discovered that some neighborhoods had duplicate rows that divided them in two different regions. As Geopy probably wouldn't recognize the division when requests for coordinates were made, duplicates were aggregated into only one row, by summing area and population values and maintaining the first row's other values, which were all equal to the second's. Also, area and population values needed to be transformed to float values, as they were parsed as strings.

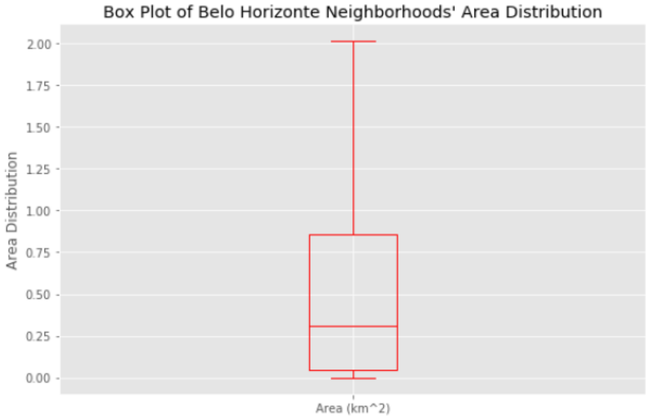
Afterwards, descriptive statistics, namely mean, frequency, standard deviation, minimum, maximum and percentiles were defined and plotted into box plots, in order to further understand the sample.

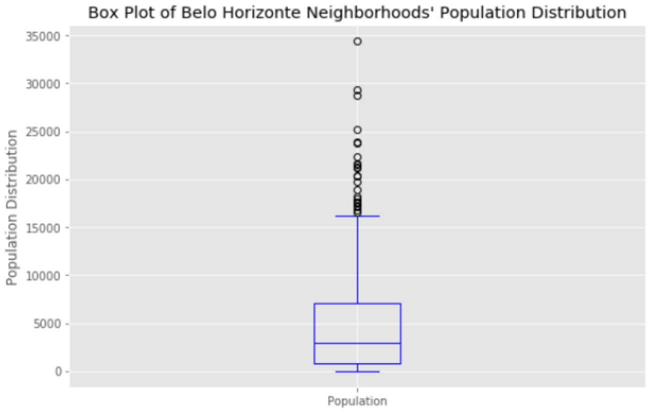


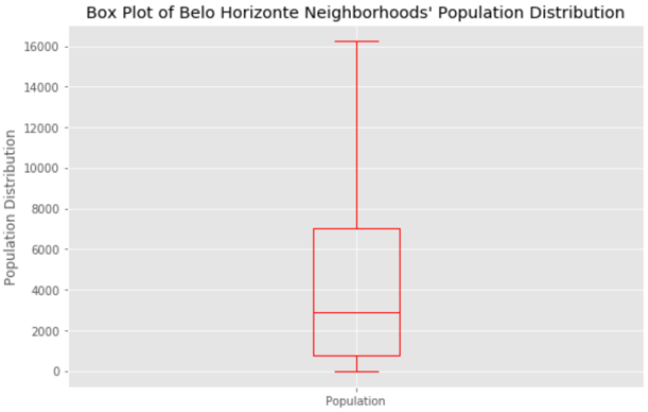


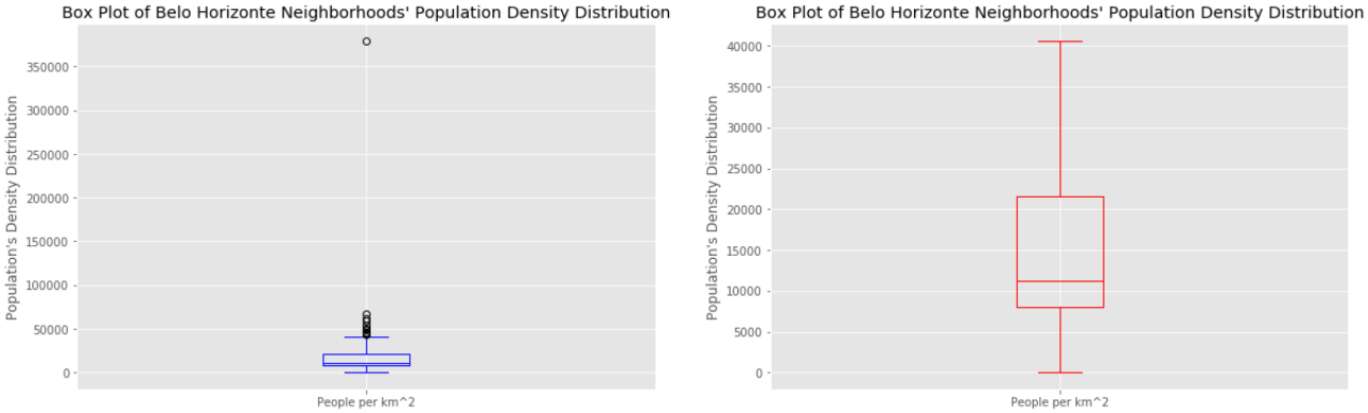
By analyzing it, it was decided that the sample which would be used in the clustering algorithm shouldn't contain neighborhoods with low population, area or population density values. That conclusion was reached due to the fact that these wouldn't be ordinary interesting options of neighborhoods for considering to open a food establishment in, in the case of neighborhoods with low population or population density. As for neighborhoods with low area values, they wouldn't be interesting to consider because, latter, a radius value would have to be set to search for each neighborhood's near venues, and little neighborhoods would mostly consider venues that weren't theirs. After adding a column with population density values to the dataframe and generating new boxplots to analyze the data, the minimum values for each one were defined.











The boxplots on the left side include outliers, while the ones on the right don't.

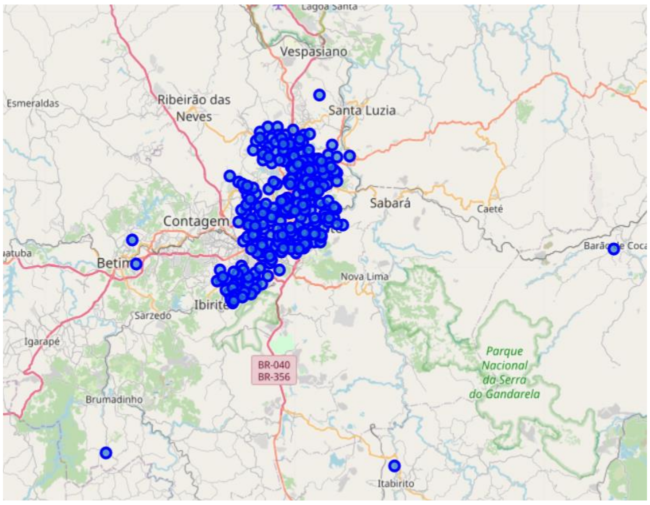
- Population minimum value: 1500 people per neighborhood;
- Area minimum value: 0.02 square kilometers;
- Population density minimum value: 5000 people per square kilometer.

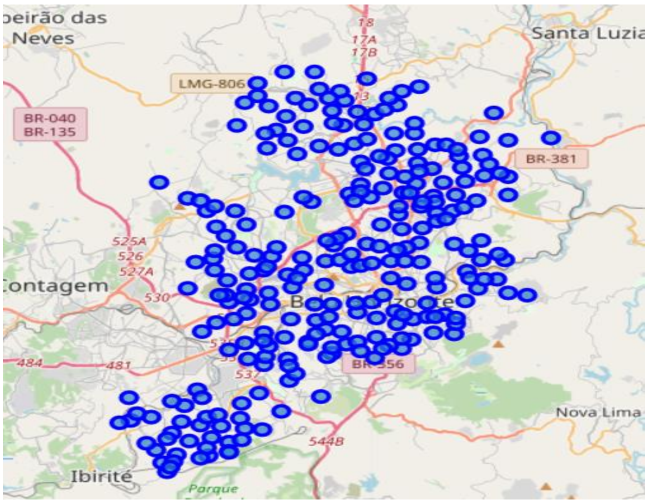
After the dataframe was filtered accordingly to the new requirements, the sample diminished from 487 to 284 neighborhoods.

Neighborhoods' Coordinates Acquisition and Treatment:

The next step was running a loop to acquire for each of the remaining neighborhoods their coordinate values. For that, Geopy's library resources were used and the values were attached to the dataframe.

Eight of the neighborhoods weren't recognized by Geopy, so they were removed from the dataframe. Also, some of the returned coordinates had values that clearly didn't belong to Belo Horizonte, as it is showed on the left Folium map displayed below. After removing each one of them from the sample, some manually and some by filtering the dataframe by removing absurd coordinate values, there were 263 remaining neighborhoods, which can be visualized on the right Folium map displayed below.

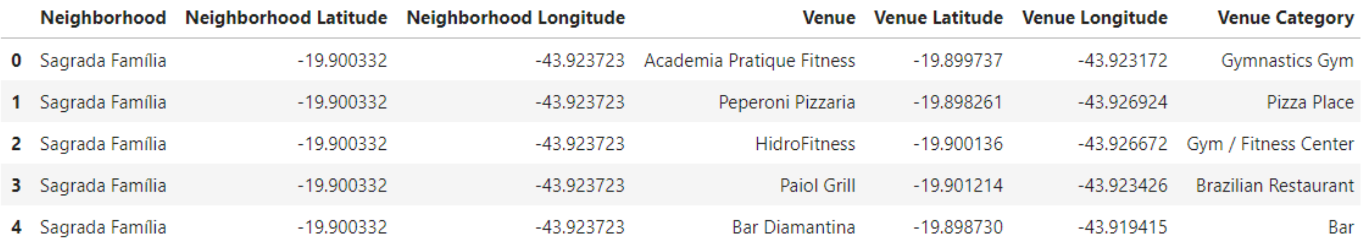




Venues' Data Acquisition and Treatment:

To obtain information about venues within each neighborhood, the Foursquare API was chosen because of its simplicity of usage and because it is free of charge until a certain number of API requests per day. It works by receiving latitude and longitude values and using them for searching for venues around it within a determined radius. The radius chosen for this study was 600 meters, because the area of a circumference with such radius value is slightly bigger than the sample's average value of neighborhood area, which could compensate for elongated neighborhood geometries.

After making the requests, almost 8000 venues were found for all neighborhoods. The values were added to a new dataframe that consisted of one row for each venue in each neighborhood, among with the other column values. A sample of the new dataframe format is displayed below.

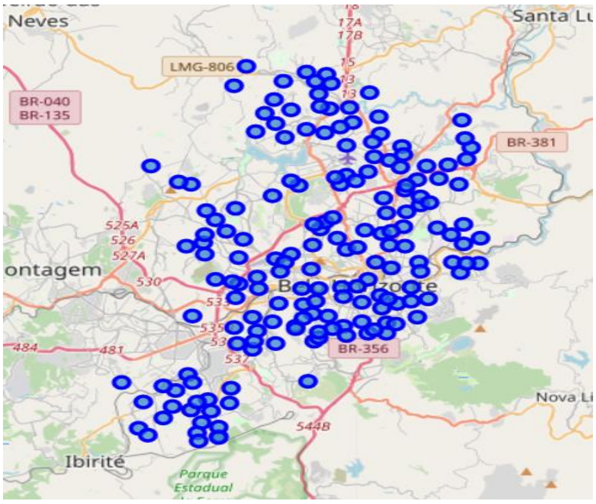


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Sagrada Família	-19.900332	-43.923723	Academia Pratique Fitness	-19.899737	-43.923172	Gymnastics Gym
1	Sagrada Família	-19.900332	-43.923723	Peperoni Pizzaria	-19.898261	-43.926924	Pizza Place
2	Sagrada Família	-19.900332	-43.923723	HidroFitness	-19.900136	-43.926672	Gym / Fitness Center
3	Sagrada Família	-19.900332	-43.923723	Paiol Grill	-19.901214	-43.923426	Brazilian Restaurant
4	Sagrada Família	-19.900332	-43.923723	Bar Diamantina	-19.898730	-43.919415	Bar

As it is evident on the first row, the API request was made to return any kind of venue, but this study requires only food venues' information. Hence, the new dataset needed to be filtered. In order to do that, another API request was made to the Foursquare database to store all names of food venue categories in a list. 349 different possible names were returned for food venues. Then, the dataframe rows that didn't contain a venue category that was among the values in the list were removed from the dataframe.

By counting how many venues were returned for each neighborhood, it was noticed that for some neighborhoods, very few food venues were returned. As our clustering model would work with the percentual values that each venue category's frequency represented among the total number of food venues on that neighborhood, regions with too few establishments would get high percentual values not because of popularity, but due to the size of the sample, which would badly affect the model's results. Therefore, a minimum value of 10 venues per neighborhood was set for a neighborhood to be considered on the model.

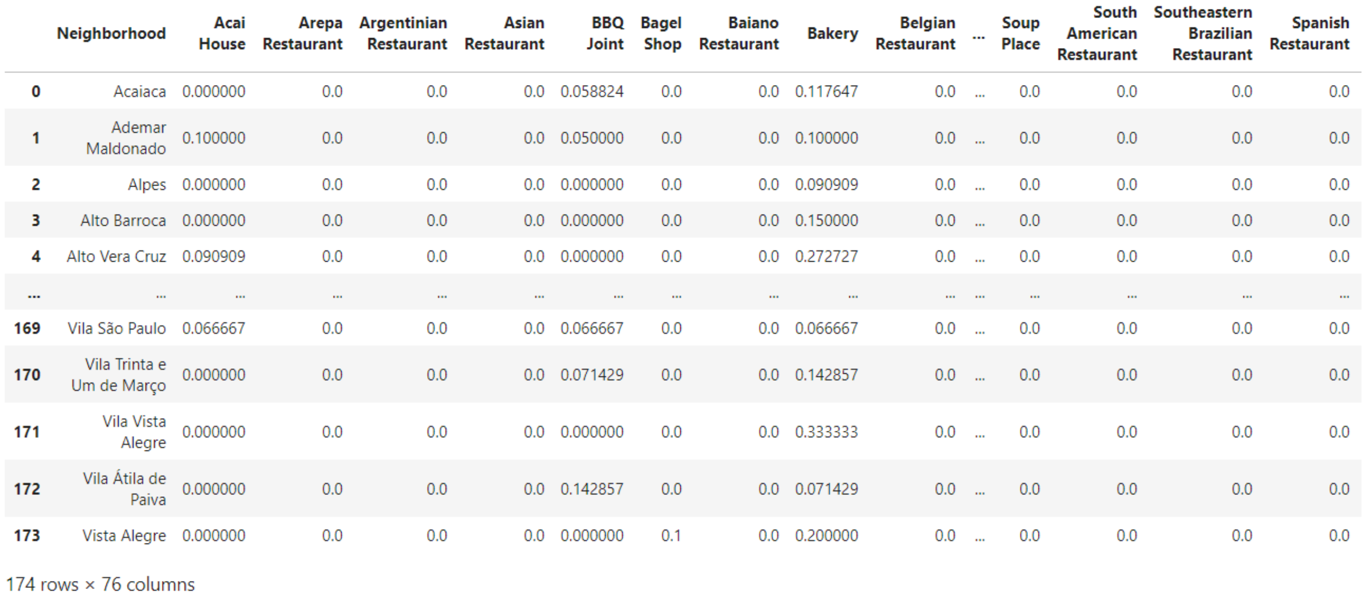
After filtering the dataframe, there were a total of 174 neighborhoods and 3388 venues that would be used to fit the clustering model. The Folium map containing those neighborhoods is displayed below.



Clustering Neighborhoods with K-Means:

At this point, all the data needed for running a clustering model had already been gathered and only needed a few transformations in order to fit a model. The machine learning algorithm chosen was the K-Means method. It works by setting initial values for K centroids within the data, which will move at each iteration trying to minimize the medium distance of a centroid to all its cluster's components. It is important to note that the number of clusters will always be equal to the number of centroids. The main problems involved in the application of this method are the difficulty in choosing a value for K and that depending on the initial values attributed to the centroids, the model may return different results. Although these issues can't be resolved easily, every clustering method has its advantages and disadvantages, so this specific one was chosen for its simplicity of application.

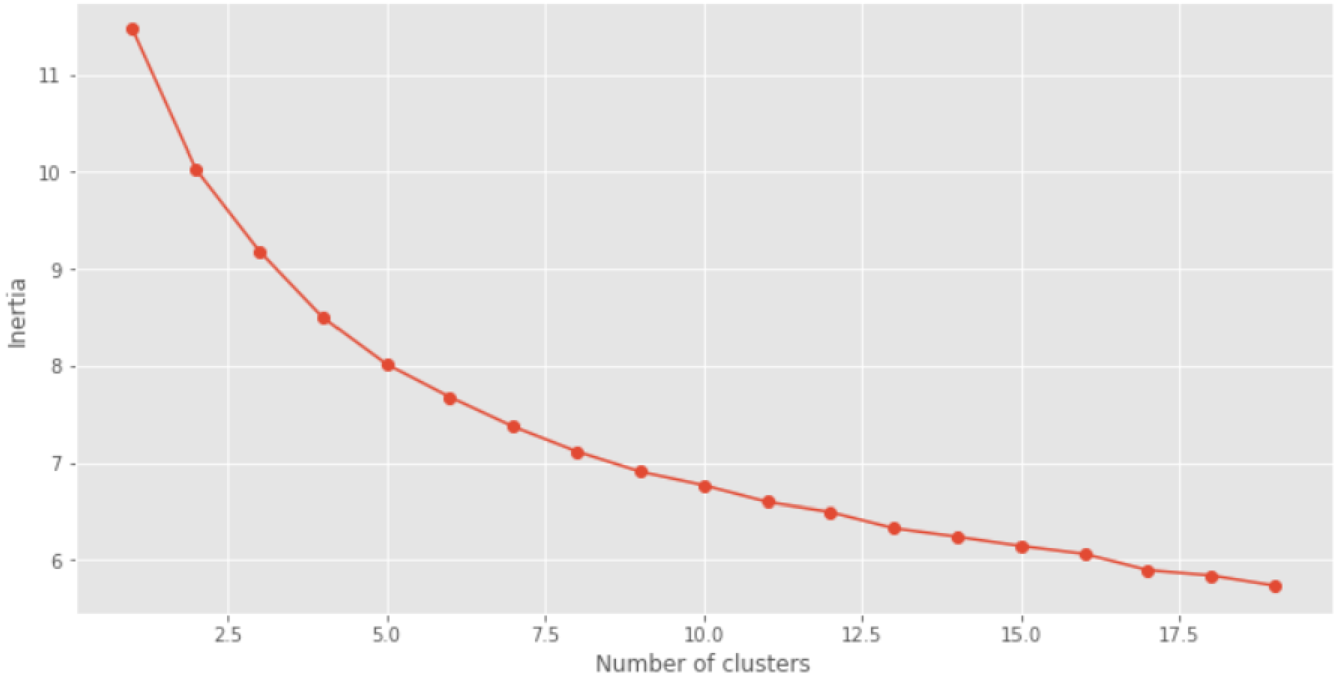
As it has already been stated, the model will be fitted with the percentual frequency of venue categories within each neighborhood. It means that neighborhoods that have similar frequencies for many venue categories will be attributed to the same cluster. A sample of the dataframe used for fitting the model is displayed below.



	Neighborhood	Acai House	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	Baiano Restaurant	Bakery	Belgian Restaurant	...	Soup Place	South American Restaurant	Southeastern Brazilian Restaurant	Spanish Restaurant
0	Acaiaca	0.000000	0.0	0.0	0.0	0.058824	0.0	0.0	0.117647	0.0	...	0.0	0.0	0.0	0.0
1	Ademar Maldonado	0.100000	0.0	0.0	0.0	0.050000	0.0	0.0	0.100000	0.0	...	0.0	0.0	0.0	0.0
2	Alpes	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.090909	0.0	...	0.0	0.0	0.0	0.0
3	Alto Barroca	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.150000	0.0	...	0.0	0.0	0.0	0.0
4	Alto Vera Cruz	0.090909	0.0	0.0	0.0	0.000000	0.0	0.0	0.272727	0.0	...	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
169	Vila São Paulo	0.066667	0.0	0.0	0.0	0.066667	0.0	0.0	0.066667	0.0	...	0.0	0.0	0.0	0.0
170	Vila Trinta e Um de Março	0.000000	0.0	0.0	0.0	0.071429	0.0	0.0	0.142857	0.0	...	0.0	0.0	0.0	0.0
171	Vila Vista Alegre	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.333333	0.0	...	0.0	0.0	0.0	0.0
172	Vila Átila de Paiva	0.000000	0.0	0.0	0.0	0.142857	0.0	0.0	0.071429	0.0	...	0.0	0.0	0.0	0.0
173	Vista Alegre	0.000000	0.0	0.0	0.0	0.000000	0.1	0.0	0.200000	0.0	...	0.0	0.0	0.0	0.0

174 rows × 76 columns

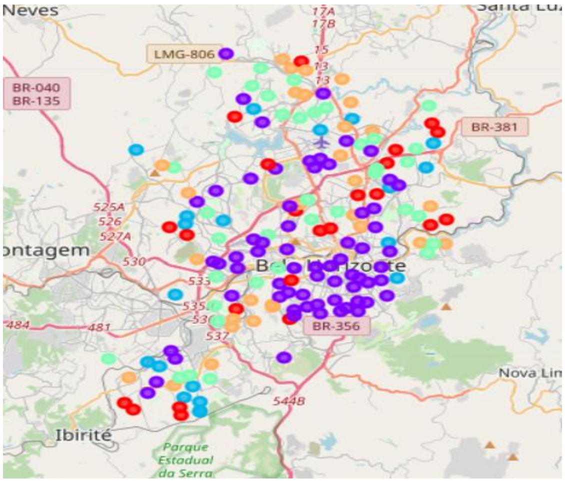
The algorithm ran with 20 different values for K and 40 times for each K value. Each one of the forty times a model was created for each K, an evaluation metric called Inertia was calculated. Inertia captures the aggregate medium distances of cluster components to its centroids for each cluster and is represented in the form of a real value. The bigger the Inertia value, the more distant the instances are from its cluster centroids, which means that big values for Inertia are bad and small values are good. For each K value, the smallest Inertia value that resulted from the forty iterations was stored and displayed in a line plot that can be visualized below.

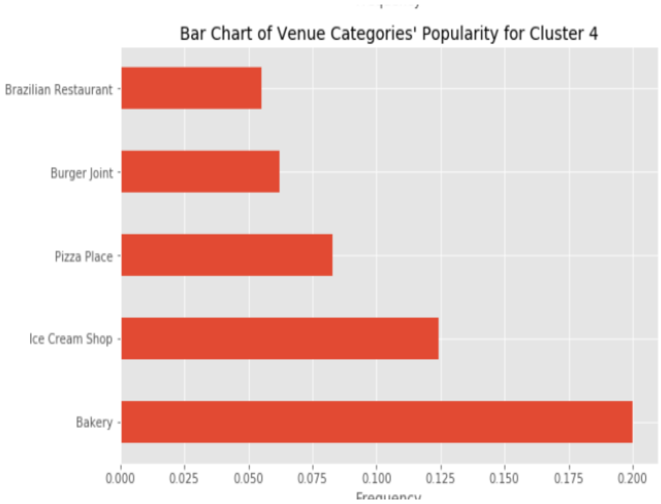


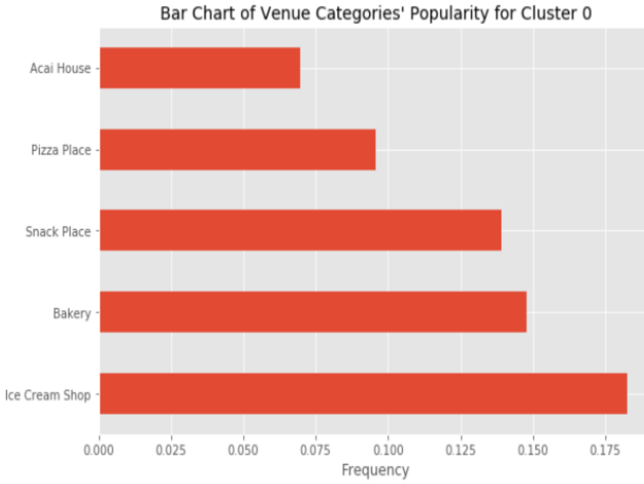
As the slope of the function didn't diminish significantly from one K value to the next, it doesn't help much on choosing the ideal K. Therefore, an assumption that a value of 8 for Inertia would be acceptable for a model was made. It is important to say that smaller Inertia values aren't necessarily better for creating a model, because if it has too many clusters, it gets difficult to take any conclusions from the results, because a lot of the clusters will probably be much alike and the model will probably have overfitted. Thus, it was decided that our cluster model would be made of 5 clusters. The results of the cluster method will be discussed on the next section.

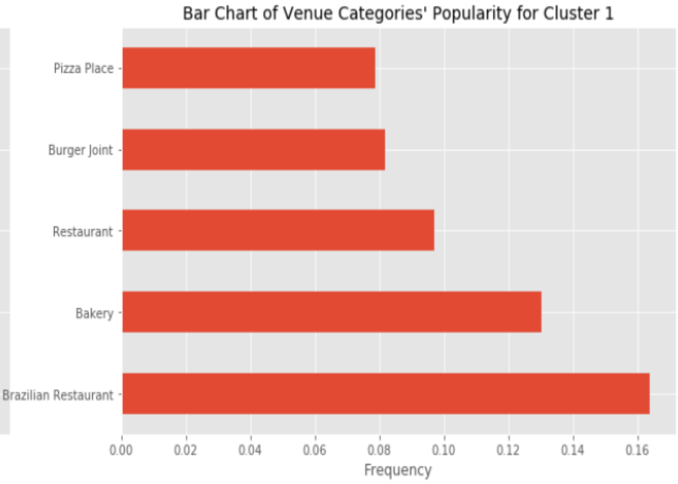
RESULTS:

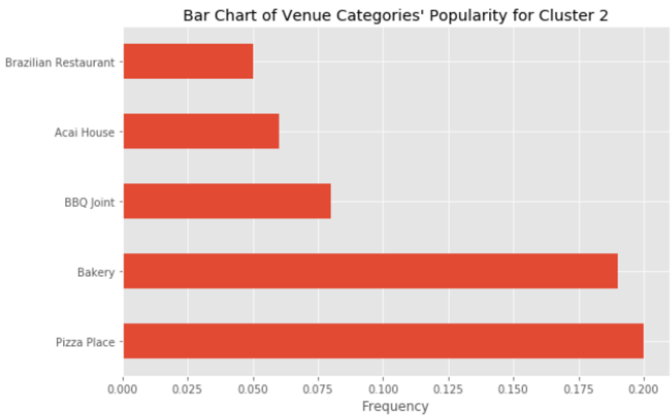
The following Folium map and bar charts were plotted to analyze the results of the clustering model.

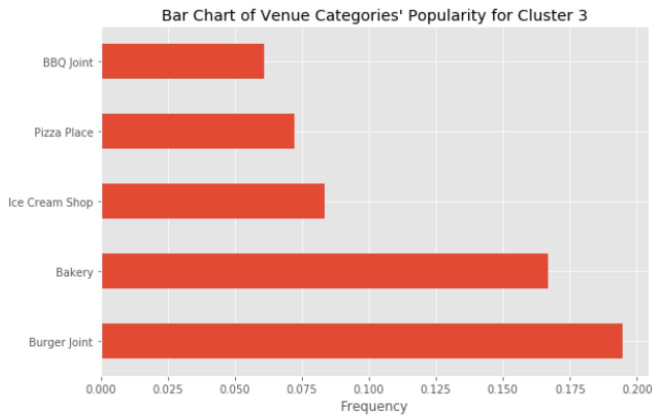



Bar Chart of Venue Categories' Popularity for Cluster 4


Bar Chart of Venue Categories' Popularity for Cluster 0


Bar Chart of Venue Categories' Popularity for Cluster 1


Bar Chart of Venue Categories' Popularity for Cluster 2


Bar Chart of Venue Categories' Popularity for Cluster 3

Each color on the Folium map represents a cluster. As for the bar charts, they show the percentual frequency that each of the clusters' 5 most popular food venue categories represent over all 5 most popular venue categories for each neighborhood in that cluster. Basically, it shows the popularity of a venue category within a cluster among each neighborhood's most popular venue categories.

Having that said, we should search for neighborhoods within each one of the clusters that have few or none establishments of the most popular category. It is reasonable to assume that it is worth it looking further into these neighborhoods, because they were assigned to a determined cluster due to similarity with other neighborhoods. Hence, if they don't possess the most common feature among the group, it must be because of some particularity of the region or for no reason at all, and simply represent an untapped potential for a successful food establishment business.

After searching for neighborhoods that didn't have its cluster's most popular venue category among its 5 most popular venue categories, the following neighborhoods have been classified as with high potential for the opening of a determined establishment.

Neighborhoods with Untapped Potential for Ice Cream Shops	Neighborhoods with Untapped Potential for Brazilian Restaurants	Neighborhoods with Untapped Potential for Burger Joints
• Pedreira Prado Lopes • São Cristóvão	• Alto Barroca • Buritis • Carmo • Conjunto Celso Machado • Fernão Dias • Grajaú • Maria Helena • Mineirão • Sion • São Bernardo • São Pedro • Vila Sumaré	• Rio Branco

DISCUSSION:

This application of a machine learning algorithm to search for neighborhoods that presented possible opportunities for the opening of a particular type of food establishment is far from perfect. Many assumptions were made without much foundation, as were the definition of some of the model's and Foursquare API request's parameters. Many neighborhoods were removed from the sample due to the fact that they weren't recognized either by the Foursquare API or by Geopy, while an alternative would be to find another way of acquiring these neighborhoods' information. Nevertheless, one of the main goals of the study was to present a simple method of clustering a city's neighborhoods based on its venues' categories, and that was certainly accomplished.

As for the clusters' results more specifically, some types of establishments seem to be equally popular among the great majority of neighborhoods, as it happened with bakeries. This kind of venue was among all clusters' 5 most popular venue categories, which raises a question about which venue categories should be considered as model inputs. Maybe, some categories' percentual frequencies are so similar and homogeneous among all neighborhoods that they shouldn't be decisive on the clustering of the regions. This is definitely something that should be considered on future similar applications of K-Means on cities' neighborhoods.

Another questionable feature of this study is the data's reliability. Even though the authors of the Wikipedia page claim to have collected the data from IBGE's census of 2010, its authencity can´t be guaranteed, since Wikipedia is a crowd sourcing website. As the focus here was on the application of the machine learning algorithm and not on the results, it didn't matter much, but real-world applications should consider this factor as of maximum importance, because any results obtained with false data will most certainly lead too wrong conclusions.

CONCLUSION:

Even though the study had many technical flaws, which were appointed on the last section, the main goal of providing entrepreneurs with relevant information about neighborhoods with potential for particular categories of food establishment businesses in Belo Horizonte, by applying a simple machine learning clustering algorithm based on food venue categories' popularity among neighborhoods was achieved. The simplicity of this study was purposeful, so it could serve as basis to more complex and meaningful improved applications. It is expected that the reasoning behind it is used in similar future studies, which should achieve better results with more attention put to sample selection, model parameters' optimization and model inputs' selection.