IBM Coursera Capstone

Battle of Neighborhoods - Week 2

Author: Lucas Gonçalves Temponi Soares

Introduction

The 2020 pandemic, unfortunately, not only took many lives but also brought challenges for the establishments, forcing a great number of those to close doors. Albeit sad, it brings opportunities for those willing to take a risk and who have the means to start a new business.

Using data extracted from Foursquare and data science, a data scientist might be able to shed some light on the best neighborhoods to start a new business.

This work will consider a Pet Store in Buritis, a neighborhood in Belo Horizonte, whose owner wants to open a new store. With this information we will try to find a neighborhood similar to Buritis, in venue's styles and quantities, and which has the biggest popularity in pet stores, with the least number of stores per habitant.

Data

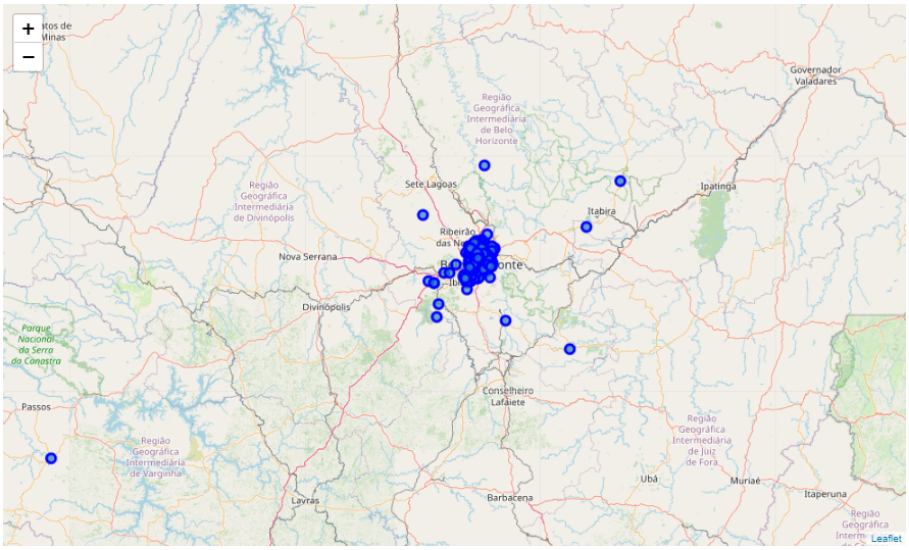
Data sources

Given the problem proposed we first need to list the neighborhoods in Belo Horizonte. For this, we will scrape a Wikipedia page that lists all the neighborhoods and its populations. After scraping the list we can use Nominatim to get the neighborhoods latitude and longitude. This data will later be used to get a list of venues of every neighborhood in the city. With this, we can use k-cluster to find similar neighborhoods, based on the number and style of venues.
After clustering and finding a list of potential candidates we are going to use the Foursquare API to get the likes of every Pet Store in the similar neighborhoods. The likes will be used as a popularity measure and will help us define the best neighborhoods to open the new store.

Data cleaning

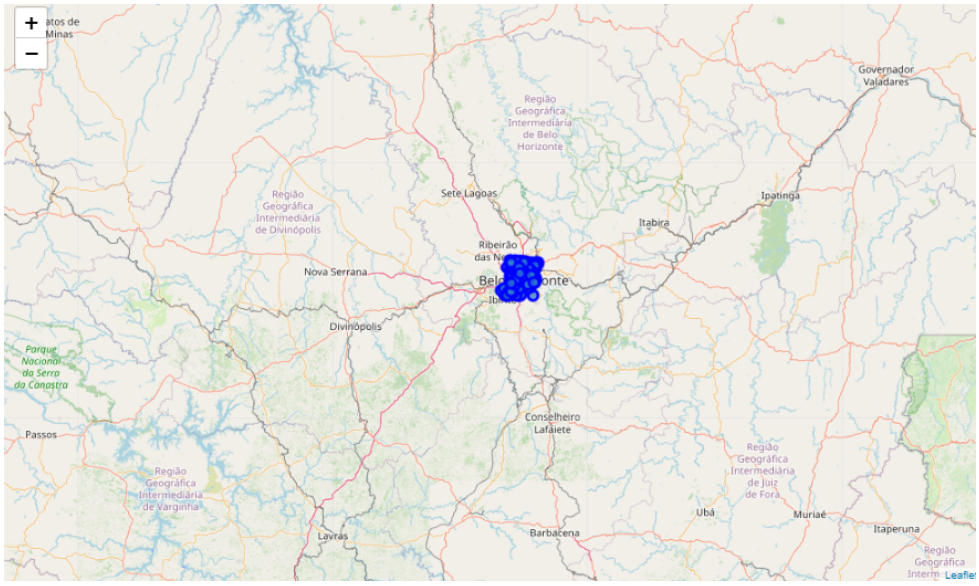
After scraping the list of neighborhoods and getting the latitudes and longitudes, we dropped all the rows with 'NaN' values and plotted a map to visualize the neighborhoods

Unfortunately, some neighborhoods returned a wrong location, even after adding the city name in the Nominatim search. By analyzing the map latitudes and longitudes limits were placed in order to filter the neighborhoods that are not within the city of Belo Horizonte




```
df = df[(df['longitude'].astype(float)<-43.85) &
(df['longitude'].astype(float)>-44.07)\n", " &
(df['latitude'].astype(float)<-19.82) & (df['latitude'].astype(float)>-20)
].reset_index(drop=True,inplace=True)\n
```

Giving the result shown in the map below:



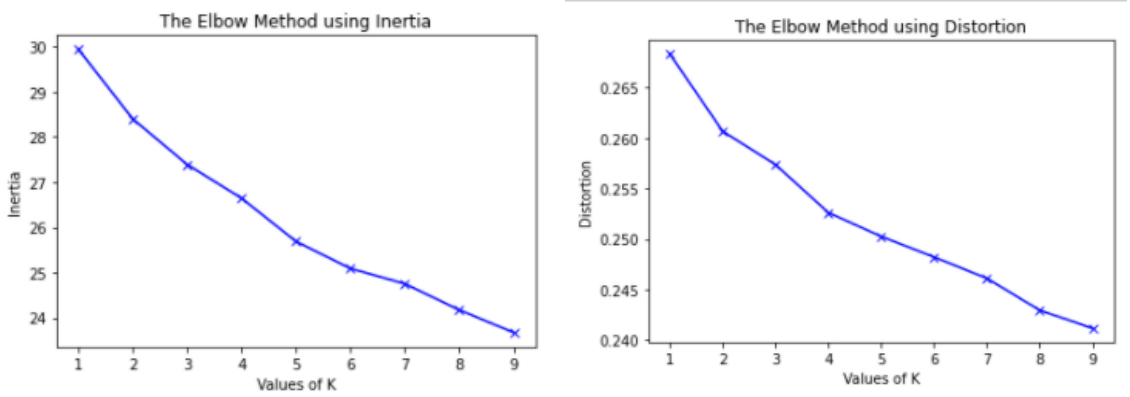
Having the correct list of neighborhoods, we used the Foursquare API to get a list of all the venues,it's name and category. In order to facilitate working in over a few days, a csv file was created with all the relevant data.

Methodology

For this analysis we will start with a k-means cluster in order to group neighborhoods with similar characteristics, like venue types and its frequency within the neighborhood.

K-means clustering is a method that aims to partition observations into *k* clusters. By analysing the top 5 most common venue categories in each neighborhood, we tried to find neighborhoods similar to one another.

As to determine the best k we ran the elbow method using both Distortion and Inertia. Unfortunately, neither gave us a clear best value. A value of k = 5 was used since there seems to be some inclination around this value.



Besides having similar neighborhoods, as a business, we wish to find the place with the highest chance of success. Trying to measure this, we will look for the area where the pet shops have the highest number of likes and with the biggest population per Pet Store. The number of likes will be used as a measure of popularity and it is inferred that the places with the most likes are the regions where pet stores are most popular. The population in the area is directly related to the number of potential customers and, for this reason, we will try to maximize the number of likes and the population of its area.

Results

As stated above, a k = 5 was used returning the clusters shown in the table below. Since, as specified by the given problem, we are looking for neighborhoods similar to Buritis, we identified the cluster label where it was placed and selected all the neighborhoods in that cluster. Buritis was in cluster 1, with another 165 areas.

With this data it's possible to find every Pet Store registered to Foursquare in this given cluster. The number of likes of each venue was extracted from Foursquare over the span of two days, since the API limits this kind of request to 50 and 55 Pet Stores were found in the areas.

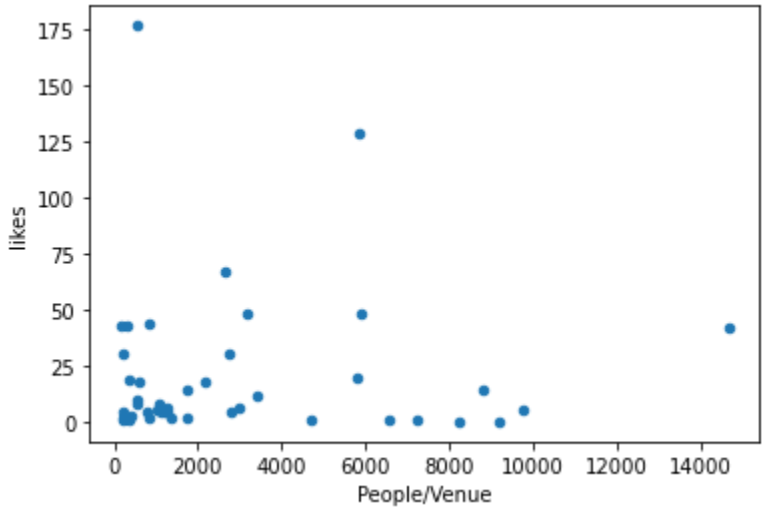
Cluster label	Number of neighborhoods
0	85
1	166
2	42
3	1
4	70

The final dataframe, which contains the data of the venues and its likes are shown below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue id	Venue Latitude	Venue Longitude	Venue Category	likes
0	Buritis	-19.976579	-43.967416	Dog's Shop	4dcc3ac452b19dd12dcd0670	-19.974309	-43.967688	Pet Store	34.0
1	Buritis	-19.976579	-43.967416	Plut Pet Zoo Pet Shop	4d974b0161a3a1cd750ab142	-19.976796	-43.968977	Pet Store	8.0
2	Pindorama	-19.922732	-43.945095	Aquario Show	4f3eaf11e4b0f45920360b15	-19.924817	-43.942433	Pet Store	7.0
3	Pindorama	-19.922732	-43.945095	Aquario Show	4f3eaf11e4b0f45920360b15	-19.924817	-43.942433	Pet Store	7.0
4	Gutierrez	-19.934260	-43.957157	Bom Garotol Pet Shop	4ff5d7c2e4b04619c6c7f9e7	-19.931785	-43.961151	Pet Store	18.0
...	...	...	...	...	...	...	...	...	...
84	Grotinha	-19.850518	-43.892311	Racao Agro Vida	50aa8bbfe4b0bfc91792e02f	-19.848048	-43.896245	Pet Store	2.0
85	Vila São Gabriel	-19.846534	-43.912311	Meu Caotinho - Pet Shop	4dfcef06d4c001cca36e39e5	-19.845508	-43.916225	Pet Store	3.0
86	Flamengo	-19.964009	-44.056256	Clinica Pet Center	4e6a871362e1cd1d54bdfaad	-19.962310	-44.058107	Pet Store	1.0
87	CDI Jatobá	-19.997196	-44.040677	Pet Shop Dog Star	55830351498eb130f5852811	-19.999533	-44.041941	Pet Store	1.0
88	Vila Rica	-19.855422	-43.953409	Pata Aqui Pata Cá	4ed39ad899114b488bc40036	-19.857588	-43.957268	Pet Store	2.0

89 rows × 9 columns

The graph below shows the distribution of “People/venue” and likes of the neighborhoods in cluster 1



Neighborhood	likes	Venue	Population	People/Venue	likes/person
Savassi	48.0	2	11772	5886.000000	0.004077
Gutierrez	129.0	3	17507	5835.666667	0.007368
Coração de Jesus	20.0	1	5812	5812.000000	0.003441
Horto	12.0	1	3429	3429.000000	0.003500
São Pedro	48.0	2	6328	3164.000000	0.007585
São Lucas	6.0	1	2987	2987.000000	0.002009
Campo Alegre	4.0	1	2776	2776.000000	0.001441
Nova Granada (ZS)	30.0	2	5516	2758.000000	0.005439
Vila Barragem Santa Lúcia	67.0	3	7999	2666.333333	0.008376
Palmares (ZL)	18.0	2	4364	2182.000000	0.004125
Minas Brasil	14.0	2	3511	1755.500000	0.003987
Vila Oeste	2.0	1	1339	1339.000000	0.001494
Acaiaca	4.0	2	2565	1282.500000	0.001559
Vila São João Batista	6.0	2	2486	1243.000000	0.002414
Átila de Paiva	4.0	1	1207	1207.000000	0.003314
Belmonte	4.0	2	2236	1118.000000	0.001789
Fernão Dias	8.0	4	4190	1047.500000	0.001909
Vila Califórnia	5.0	3	3100	1033.333333	0.001613
Barroca	44.0	4	3311	827.750000	0.013289
Vila Trinta e Um de Março	4.0	2	1525	762.500000	0.002623
Palmares (ZN)	18.0	2	1157	578.500000	0.015557
Penha	8.0	4	2226	556.500000	0.003594
Dom Joaquim	10.0	6	3279	546.500000	0.003050
Estrela	177.0	3	1591	530.333333	0.111251
Vila São Gabriel	3.0	1	420	420.000000	0.007143
Conjunto Lagoa	19.0	2	662	331.000000	0.028701
Flamengo	1.0	1	328	328.000000	0.003049
Vila Aeroporto Jaraguá	43.0	2	618	309.000000	0.069579
Nova Granada (ZO)	30.0	2	463	231.500000	0.064795
Grotinha	4.0	2	447	223.500000	0.008949
CDI Jatobá	1.0	1	204	204.000000	0.004902
Vila Rica	2.0	1	200	200.000000	0.010000
Vila de Sá	43.0	7	971	138.714286	0.044284

By simply analysing the “People/venue” and the “likes/person” columns in the table above it is possible to conclude that Gutierrez and Savassi would be the 2 best areas, similar to Buritis, to open a new pet store.

Discussion

The number of “likes”, albeit sufficient for a basic analysis, is not a good measure of popularity. For a better result it would be possible to use data from Google Places and find the number of people that actually visit every venue each month. This would also be a better metric for potential customers.

Using machine learning to find the best area to open a new store should be possible if more data about the pet stores are gathered.

Given prior knowledge about the city and some of its neighborhoods, the results observed in this work are realistic and would not surprise a resident of the city. This shows that the methodology applied can be used in cities where no stakeholder has any prior knowledge.

Conclusion

The results obtained were very good, even though there are many possible opportunities for improvement. The basis of the work could be used to develop an algorithm to find the best place to open a franchise, given a list of other places where success was found.