IBM CAPSTONE PROJECT FOR DATA SCIENCE

Analysis of grouping of clothing stores in the city of Recife through its adjacent locations

Nicólly Lira de Albuquerque

INTRODUCTION

This report refers to the explanations and results of the final project of the Coursera Aplied Data Science Capstone.

The purpose of this analysis is to verify whether it is possible to categorize clothing stores in groups in the city of Recife - PE, through the similarity of nearby establishments within a 200 meter radius. In order to find a pattern of common establishments for each group, helping as one of the possible indicators to support the best choice of location for the opening of other stores, which have similarity with stores of any of these groups that were formed.

In order to construct the problem, and properly answer it, in this project we will use the Coursera specifications, the Foursquare API and the k-means clustering algorithm as mandatory.

Clustering is the separation of data into groups of similarities, performed by unsupervised machine learning algorithms. Where, this grouping is based on the distance of the Cartesian distribution of points relative to the categories that were chosen as parameters for this similarity analysis.

K-means is an algorithm that groups data by means of centroid, where it is necessary to define a k number of clusters, for each cluster a centroid is assigned and each point of the distribution that belongs to a cluster is closer to its own centroid that of the others.

Foursquare is a city guide that allows users to register establishments, review them and assign grades. This information is available for consultation by any user through the Foursquare website and application and through the API. The site is mostly used by ordinary users to do various searches on locations, as it has a more friendly and interactive interface to access this data. But that same data can be searched by programmers through its API, which was the way used to build this project.

METHODOLOGY

1. Of data collection

In carrying out this study, two main sources of data were used for the collection: the Foursquare API and the website of the city of Recife.

Data from 50 clothing stores were extracted from the Foursquare API, within a radius of 6,000 meters from the coordinates of the city center of Recife. For this first call in the API, the endpoint 'venues' was used, passing in addition to the authentication and locality parameters, the id parameter of the clothing stores category, provided by Foursquare itself on its presentation and category explanation page.

From that call, the saved data was saved store ID, store name, latitude and longitude.

Before leaving for the second call on Foursquare, we will access the Nominatim API and use the latitude and longitude to retrieve information from the neighborhoods to which these stores belong, as this endpoint does not return this information to us in your json file.

In the second Foursquare API call, we will use the 'explore' endpoint, to search, for each clothing store found above, a maximum of 200 establishments and facilities within a radius of 200 meters away. This time, we store the name of these establishments and what type / category of Foursquare classification they belong to.

At this point, we found our first obstacle to our goal of grouping clothing stores based on the category of establishments in their surroundings: Foursquare has a hierarchical classification system, where there are classes, within subclasses, within sub-subclasses and so on. Given this, the API returns us the smallest possible class of categorization of that establishment, thus creating an absurd amount of different classes in the vicinity of the stores, making it impossible to identify a pattern.

To solve this problem, we will again resort to another call in the Foursquare API, this time with its 'categories' endpoint, to access an available ontology that determines the larger categories of understanding to which each subcategory belongs. We thus created an ontological dictionary of Foursquare categories that made it possible to simply replace the smaller categories with their respective larger categories.

Now using the Recife City Hall website, we will perform a page-by-page webcrapping, from a list already created with the unique neighborhood names of our clothing stores, and save for each neighborhood its "Average Monthly Nominal Income Value of Households ".

This socioeconomic metric from IBGE refers to the average income calculated with the sum of any money received at home, not just the amounts formally documented. On the city hall website there is no information on whether this nominal income is per capita or not, that is, the sum of the total income divided by the number of residents of the residence, but in a brief survey of metrics, we were able to confirm that it is indeed a value per capto.

2. Pre-processing of data

As we already know, we will perform the grouping based on the similarity of categories of establishments in the surroundings of our sample of clothing stores in the city of Recife. These categories are words that designate to which class of establishment these locations belong, of the ten general classes provided by Foursquare, seven were found in this case.

To count how many times each of these seven categories appears for a respective clothing store, we will isolate the column of categories and represent it in a binary form (one hot encoding), transforming the unique values of the column into categories.

So that for each existing category for an entry, your column will receive a value of 1 and for each nonexistent category, your column will receive a value of 0.

Once this is done, the ID information of the clothing stores will be repositioned for their respective lines with binary columns. , which represent each establishment close to that ID.

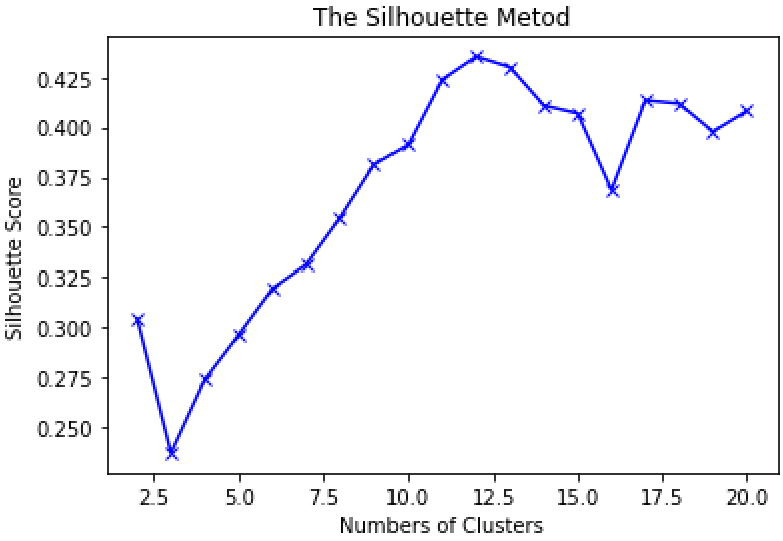
Finally, we group the IDs together by adding the values of the binary category columns, thus finding the number of categories next to each of our clothing stores.

The last step of our pre-processing phase is scaling, that is, we will place our data on the same scale of values, in order to improve the metrics of the algorithm. For this case, we will use Normalization scaling, placing our values on a scale between -1 and 1.

3. Clustering

k-means is the clustering algorithm that we will use and as mentioned above, it performs the grouping by means of centroid distance. However, indicating as a parameter the wrong number of clusters, can make our separation unsupervised by similarity of variance (clustering) little precise, either because it has too many groups that generalize little, or groups of less that cannot give visibility to other nuances that may be important in the analysis.

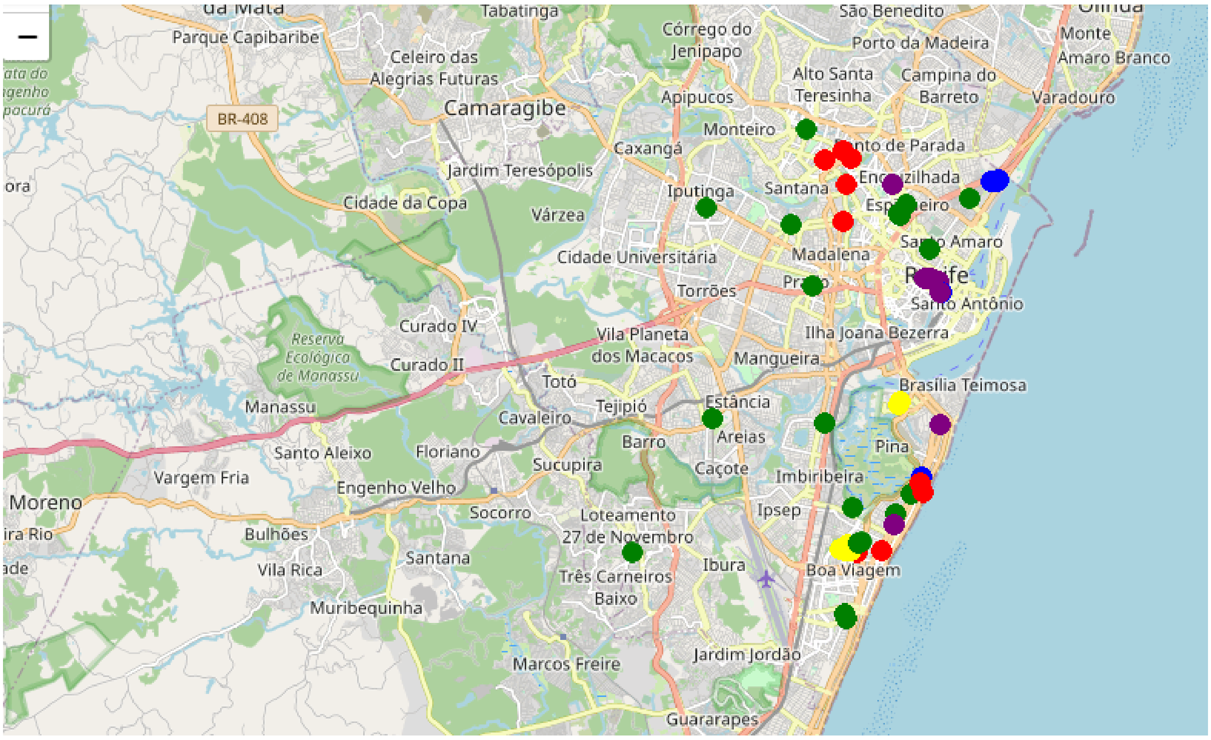
To try to solve this problem, finding an optimal amount of clusters, two Heuristic methods were used: the elbow method, which gave very inconclusive results, and the silhouette method, which gave us slightly less cloudy results.



Although this graph points to more or less ten or eleven clusters, this is a very large number of groups in relation to the amount of our data.

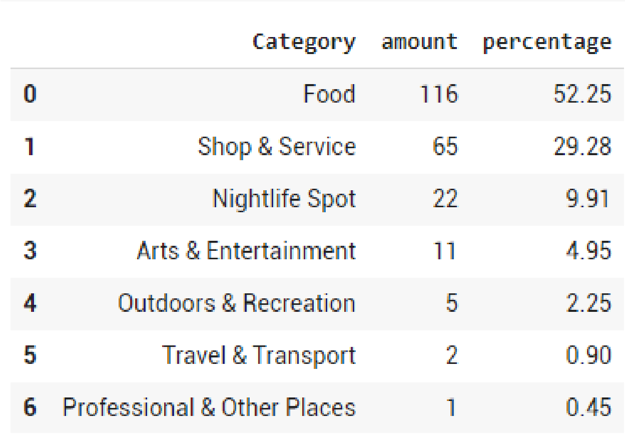
Thus, to avoid an overfit, a number of 5 clusters was assumed, because it is the best silhouette index before the score goes up indefinitely and it is the number of clusters that more evenly distributes the number of stores among the clusters.

CLUSTERS UNDERSTANDING



Cluster 4 (Purple): It has eight stores, the vast majority of which are located in the central region of the city of Recife, in the Boa Vista neighborhood. However, it still has two stores located further south of the city and only one in the north in the Aflitos neighborhood. The average monthly nominal income per household, in the neighborhoods in which the stores in this cluster are located, varies considerably, starting at one thousand and reaching up to seven thousand reais.

It is also the cluster that has the highest percentage of 'Nightlife Spot' in relation to the other groups.

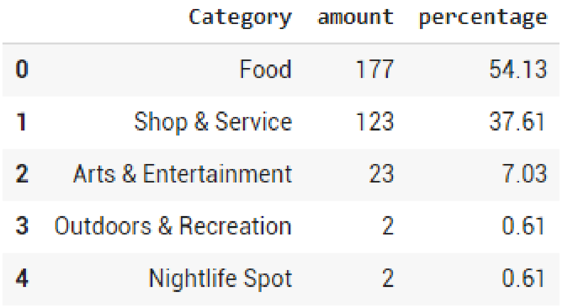


	Category	amount	percentage
0	Food	116	52.25
1	Shop & Service	65	29.28
2	Nightlife Spot	22	9.91
3	Arts & Entertainment	11	4.95
4	Outdoors & Recreation	5	2.25
5	Travel & Transport	2	0.90
6	Professional & Other Places	1	0.45

Cluster 3 (Yellow): It has six clothing stores, all within shopping centers, located in the south zone of Recife, in the neighborhoods of Boa Viagem and Pina, which have a value of the average monthly nominal income of households of seven and two thousand reais respectively. Another similarity between the region is the fact that both are coastal regions, with beaches that are also neighboring.

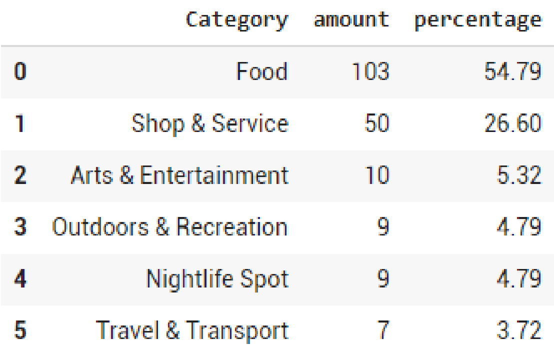
These malls also have a pattern among themselves: they are the largest malls in the city.

With this group, we can assume that the layout of clothing stores in shopping malls in this area of the city, have a pattern of establishments in the surroundings.



	Category	amount	percentage
0	Food	177	54.13
1	Shop & Service	123	37.61
2	Arts & Entertainment	23	7.03
3	Outdoors & Recreation	2	0.61
4	Nightlife Spot	2	0.61

Cluster 2 (Green): There are nineteen stores classified as part of this cluster. With a very varied geographical spread, without the presence of an identifiable pattern in this respect. However, the value of the average monthly nominal income of the households has a smaller variance between them, with the exception of the Boa Vista neighborhood, this socioeconomic metric in this cluster varies between one thousand and two thousand reais. With this, it is also the cluster with the lowest average monthly nominal income, indicating that this is the pattern of clothing store surroundings for neighborhoods in Recife where residents have, on average, a lower purchasing power.

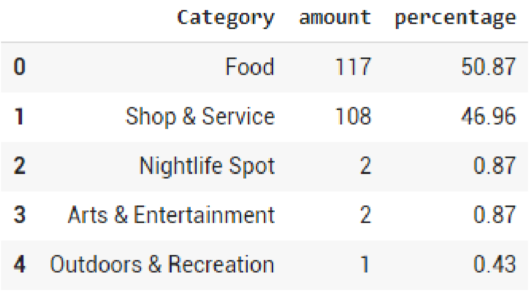


	Category	amount	percentage
0	Food	103	54.79
1	Shop & Service	50	26.60
2	Arts & Entertainment	10	5.32
3	Outdoors & Recreation	9	4.79
4	Nightlife Spot	9	4.79
5	Travel & Transport	7	3.72

Cluster 1 (Blue): This group has six stores, with its largest concentration in the center of Recife, in the Santo Amaro neighborhood, within the Tacaruna shopping mall, with only two stores located in the southern region of Recife in the Pina and Boa neighborhoods. Travel.

The value of the average monthly nominal income of households in these last two neighborhoods is two and seven thousand, respectively. The neighborhood of Santo Amaro is around a thousand, but there are relativizations that can be made due to the fact that all these stores are inside a mall.

It is also important to note that, despite being stores in a shopping mall, they were not grouped in cluster 3. The most likely hypothesis is the indication of different categories of establishments available in shopping malls according to the region, since cluster 3 has a percentage much larger in the 'Arts & Entertainment' category and slightly higher in the 'Outdoors & Recreation' category. But another hypothesis is just a different arrangement of clothing stores between these malls.

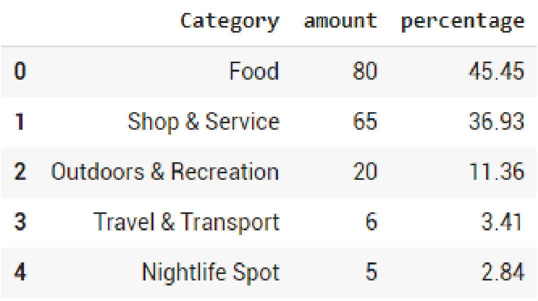


	Category	amount	percentage
0	Food	117	50.87
1	Shop & Service	108	46.96
2	Nightlife Spot	2	0.87
3	Arts & Entertainment	2	0.87
4	Outdoors & Recreation	1	0.43

Cluster 0 (Red): It has nine stores, all of which are completely concentrated in the south and north of the city, with no store in the center and in other regions.

In this cluster, in addition to the first two categories present in a greater percentage in all clusters, "Travel & Transport" establishments stand out, followed by a also good index of "Outdoors & Recreation".

All the neighborhoods in which the stores in this cluster are located also have a high value of the average monthly nominal income of households that vary between 4 and 11 thousand reais, thus being a cluster of stores in a locality of middle and upper class residents.



	Category	amount	percentage
0	Food	80	45.45
1	Shop & Service	65	36.93
2	Outdoors & Recreation	20	11.36
3	Travel & Transport	6	3.41
4	Nightlife Spot	5	2.84

DISCUSSION AND FINAL CONSIDERATIONS

Before we focus on discussions about the results of this study, I would like to highlight two pieces of information relevant to the understanding of this result as a whole.

The first point is the difficulty of finding data referring to neighborhoods in the city of Recife, the open data sources found had more general data related to the city as a whole, with entries without divisions or identification variables by neighborhood. In view of this, I carried out a webscrapping of the Recife City Hall website, but it is still a long way from a robust database.

The second and last is that unfortunately, Foursquare has gradually lost popularity in Brazil, and today there are few records of establishments still registered and users feeding the platform. Thus, an alternative that can answer the question proposed in this work, is the collection of more data, both from clothing stores, as well as other data sources for purposes of comparison and cross-checking with the main data of the stores.

Having said that, let's look at the results. With a brief observation of the percentages of categories per cluster, it is possible to determine that all of them have mostly establishments in the category "Food" and "Shop & Service". Indicating that all clothing stores, according to data and techniques used in this study, regardless of the region of the city of Recife, will have a significant number of establishments of this type around, with more than 50% of the total of nearby establishments for all customers. clusters.

It is possible to glimpse slight indications of possible patterns in some of the clusters grouped here, for example, clusters that only appear in more valued regions of the city or only inside shopping malls, as was the case with some. But, based on the quantity and quality of data available here, for the application of this specific methodology, the arguments ofsimilarities that arise are still very fragile. It can be concluded that, with the k-means clustering algorithm and the data used, it is not possible to identify a pattern of clothing store surroundings for each cluster.

Of the possible ideas for optimizing the response to this problem, one of them is to use the average ticket information of establishments in the vicinity of clothing stores, instead of the data from Recife city hall by neighborhood. Because, the average is not a good value to generalize income in this problem, although at first glance it seems specific enough, most of the neighborhoods do not have a homogeneous income distribution. There are blocks with greater economic vulnerability in the same neighborhood that has other blocks with good economic development rates.

REFERENCES

- Recife City Hall website: http://www2.recife.pe.gov.br/
    - Method Elbow and Silhouette: https://ichi.pro/pt/metodo-da-silhueta-melhor-do-que-o-elbow-method-to-find-ideal-clusters-61080390822033
    - New York City neighborhood flavors: https://towardsdatascience.com/exploring-the-taste-of-nyc-neighborhoods-1a51394049a4
    - São Paulo metro cluster: https://medium.com/@felipe.testaa/unsupervised-machine-learning-kmeans-clustering-of-s%C3%A3o-paulo-subway-stations-using-foursquare-c5101727dd85