Capstone Project - Dry Bars/ Beauty Express

Viana, Edlaine

April 25,2021

1. **Introduction**

In order to study possible locations for opening a business that has not yet been explored in Brazil, this project brings the DryBar concept, and aims to identify more suitable locations for the location of this type of establishment. The result may allow extrapolating the profile of the most successful locations to other cities with similar characteristics.

**1.1 The concept**

DryBar can be defined as spaces of beauty, at more convenient prices, when compared to traditional services. It appeared in California in 2010 and became a common presence in cities such as New York and Los Angeles, due to its agility, facilities and inviting prices.

For the development of the project, the option was to map a region of the city of New York, seeking to understand elements of geolocation and its relevance, for the installation of DryBar. It is considered that this type of service is highly dependent on the flow of female people, with little time available and a profile that values ​​good presentation in the work environment. These are factors considered to be decisive for the success of this business model.

**1.2 Project's goal**

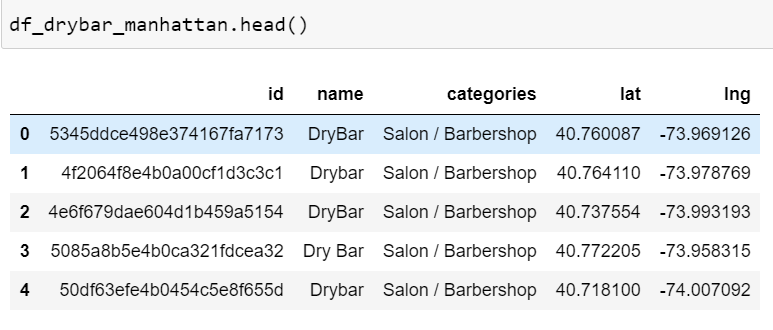
Finding DryBar in the observation region, analyzing its surroundings from the study of clusters, complemented with customer evaluations, considering that, in addition to the quality of the service, ease of access, an element related to location, is implicit.

A map of the Midtown region on Manhattan Island was made, one of the main hubs of large companies in New York City. Due to these characteristics, the region concentratespublic with a profile for this type of business; women with lots of activities and little time for traditional beauty salons.

Based on these criteria, Foursquare (https://developer.foursquare.com/docs) was used as the only source of data. The ‘k-means clustering’ unsupervised learning algorithm was the means to identify the advantages of each area, so that the profile of the best location can be defined, for future referrals to those interested in implementing this business in large Brazilian capitals.

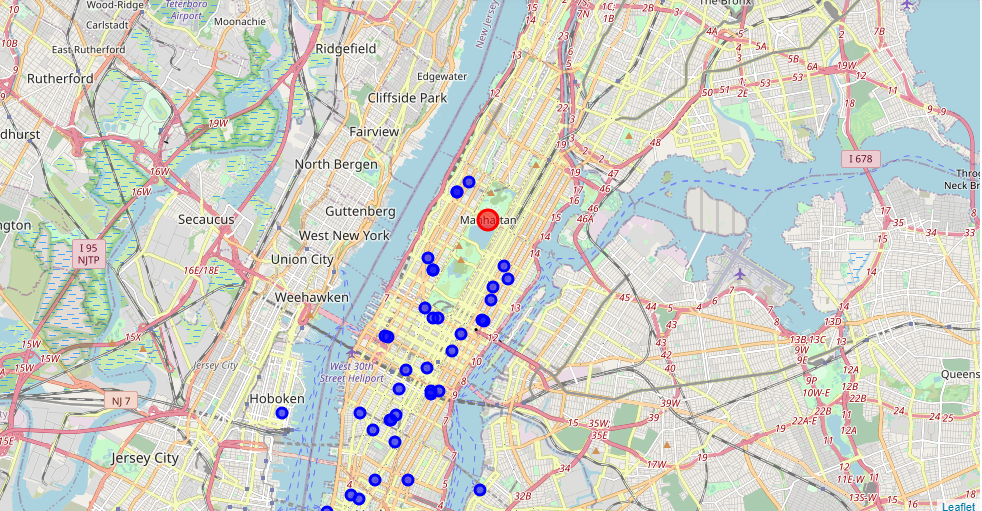
1. **Development**

In the Foursquare API a search was made with a radius of 9.5km from the center of the Manhattan district, to find the drybar establishments in the region and afterwards, manipulate and organize the data in a data frame



Parte data frame com a localização dos Drybars

41 Drybars returned, in the graph below it is possible to see how they are distributed on the island of Manhattan.



Localização gráfica dos Drybars

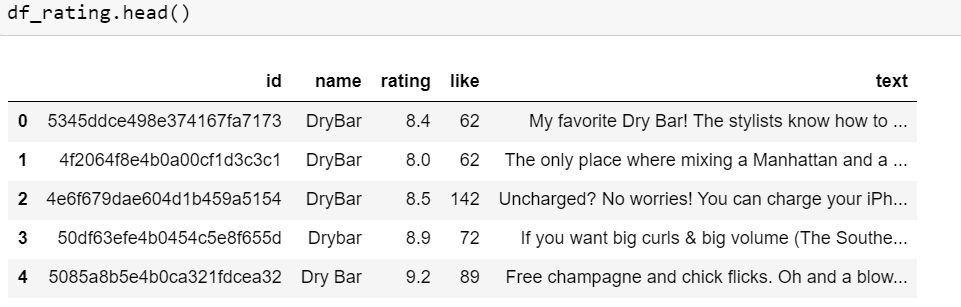
It is possible to observe that the Drybars are concentrated in a part of the island of Manhattan and not in its entire extension. The spots are in the Midtown, Upper and Lower regions, three of the five that divide Manhattan Island. It is in these regions that large companies, various government agencies and different businesses are concentrated, in fact the proximity to other establishments seems to have an influence on the choice of the location for the opening of this business model, since the other side of the island are areas considered residential.

<https://www.rodei.com.br/amp/para-entender-melhor-os-bairros-de-manhattan-parte-ii/>

* 1. **Drybars Notes**

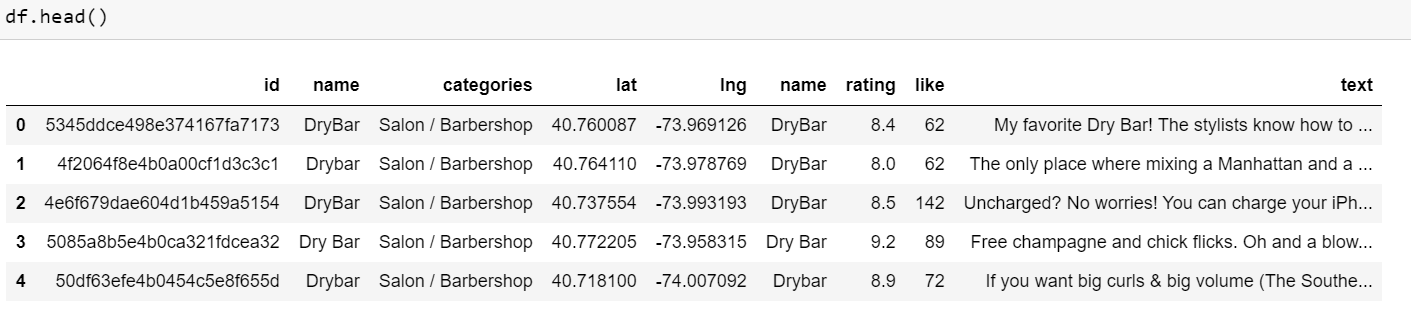
The scope of the project is to verify if there is a relationship between the success of the business and its location, in this case, to know what types of categories, services that permeate these establishments and to measure the success of the business that is associated with the quality of service will be used. customer notes, which is implicitly located, is based on the hypothesis that establishments with good scores tend to be at strategic points.

Again, a search is made on Foursquare with the ID of each Drybar in order to obtain the evaluations given by the customers (details of the search in the notebook), after manipulating a dataframe, df\_rating, as shown below.



Part of the dataframe with customer notes

The step now is to concatenate the first data frame that contains the categories and the location of each drybar, with the second data frame, which contains notes.



Dataframe part after concatenation

With the Fousquare search data previously organized, the next step will be to use the Machine learning algorithm to extract information that will later be transformed into knowledge.

**2.2 Machine Learning**

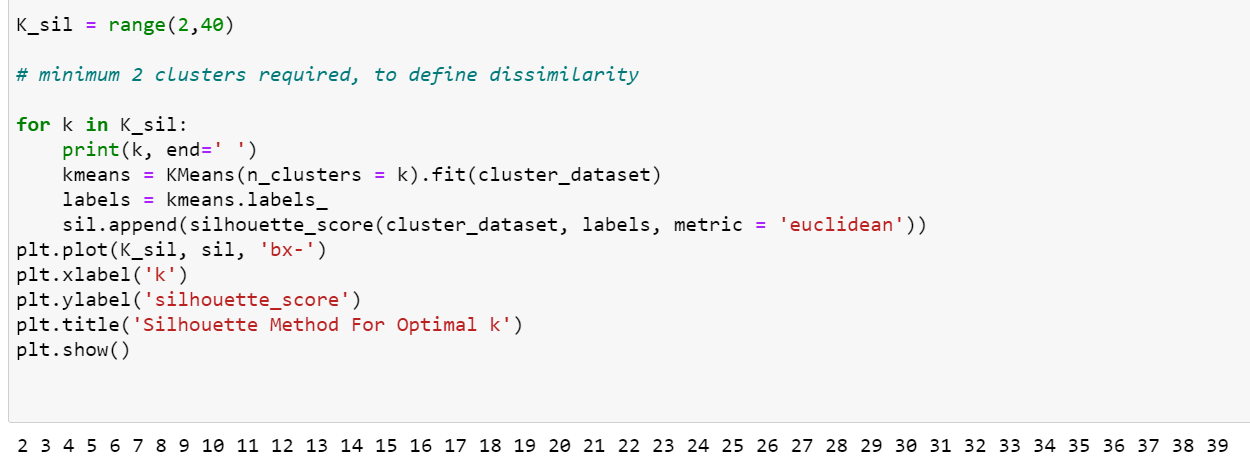
The unsupervised Machine Learning algorithm “K-means” groups data by similarities around each other around a central point, the centroid. For processing, input variables are given, in this case, latitudes and longitudes, and it returns groups, called clusters, as output.

Once you have the clusters, the next step is to calculate the average of the grades evaluated by the customers of each group. The two clusters that have the highest average score, will make a new search on Foursquare from the centroid of these two clusters to discover the categories of the neighboring establishments and draw conclusions from that.

**2.3 K-means**

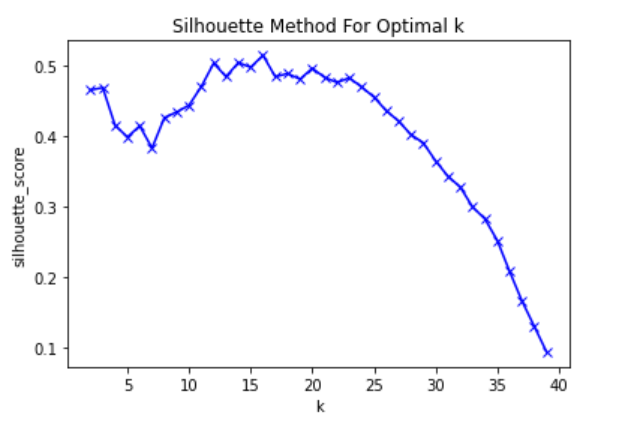
To start the algorithm it is necessary to define the best value of k, as it is not known what is the best value to choose, so we use the silhouette method, which calculates the average between how cohesive the points are within the cluster and how distant are from the other clusters so the closer to 1 the better the k value.

To carry out this process, it is necessary to provide values ​​for the algorithm so that it tests and finds out what is the best value for k, there is no rule for choosing these numbers, it is up to each one, in the case of the project the choice is numbers ranging from 2 to 39 represented by range (2.40) as shown in the image below.



Execution to find out which best value for k

The code above returned the graph to assess which k will be chosen.

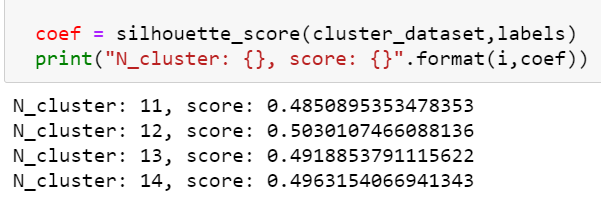


Silhouette graphic

It is possible to see in the graph two peaks, the first approximately between 11 and 14 and the second between 15 and 17, for the data scientist the decision between choosing which one has the best for k is personal, but in the case of the project with a small sample of data is not very interesting to have many clusters.

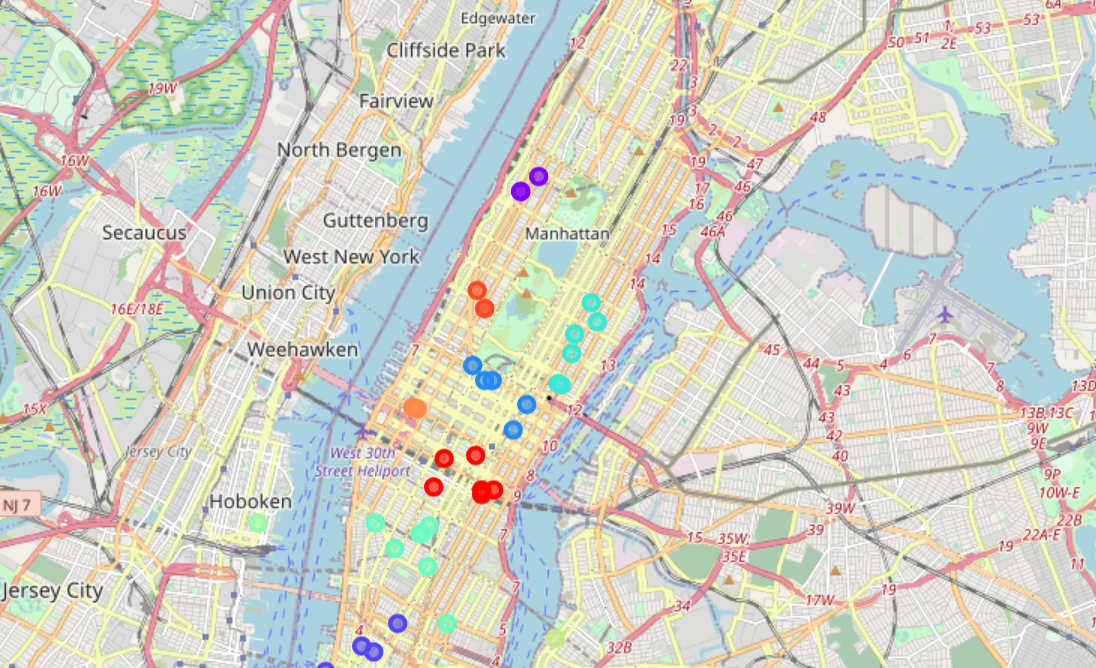
When choosing the first peak of the graph it is not yet known which is the best value for k, it could be 11.12, 13 or 14, which is the best among them?

It is not always possible to make this decision just by looking at the graphs, so it is important to check through the values of the silhouette coefficients, as shown below.



Values of the coefficients of clusters 11,12,13 and 14

This processed generated the information that the best value for k is 12, it means that the unsupervised k-means algorithm based on the input variables: latitude and longitude, will return 12 similar clusters, as can be seen in the graph below.

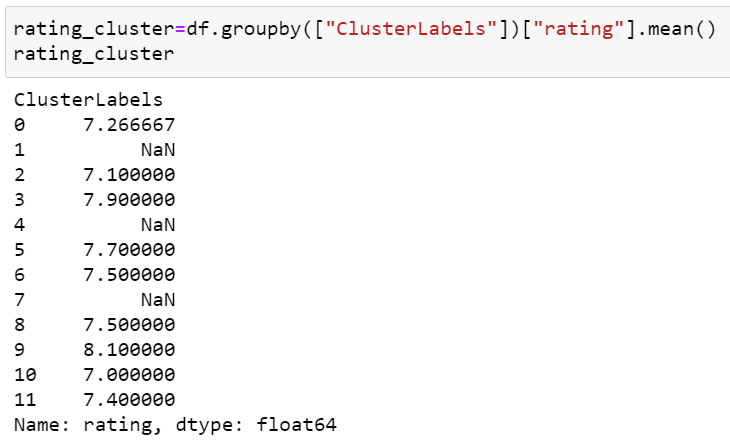


The clusters

**2.4 Getting to know the clusters**

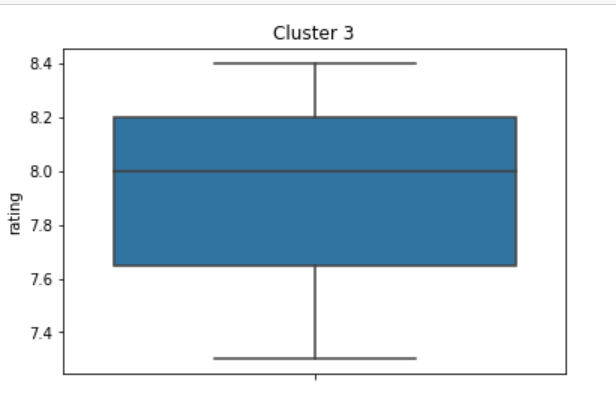
As part of the project, it is not exploring all the clusters, but the establishments considered as success cases based on customer evaluations. One way to obtain this information is by averaging the scores for each cluster.

As the idea is to know the influence of neighboring establishments, the comparison between two clusters of higher average seems to be considerable.To know the averages of each cluster, it is necessary to use the groupby function, as shown below.



Average customer ratings for each cluster

Clusters 3 and 9 have the highest average, only these two groups will serve as a parameter for the completion of the project, for which all establishments related to each of these clusters are selected.

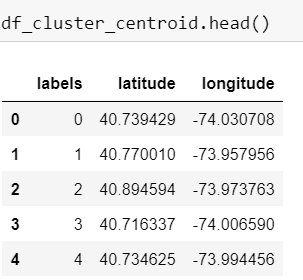


Cluster 5 boxplot

Cluster 3 has a good concentration of grades considered good above 7.5 (as shown in the graph below). Cluster 9, on the other hand, has only 1 element, that is, 1 establishment, for this reason it is not necessary to plot the graph, but as its score had a significant value, it is worth knowing the categories of its neighboring establishments. (details on notebook)

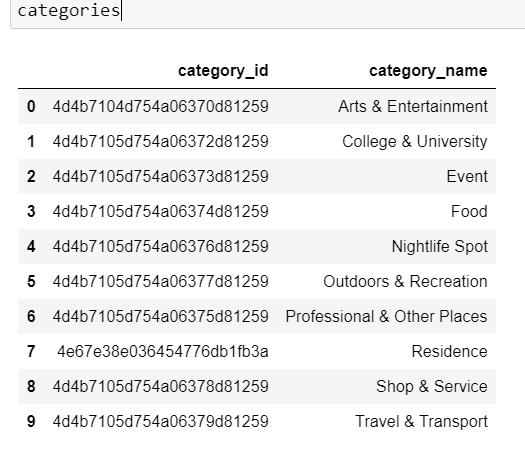
**2.5 In search of the neighboring categories of clusters 3 and 9**

As stated earlier, the k-means algorithm returns the clusters and also the central point of each one with the “kmeans.cluster\_centers” function, these centroides are coordinated. Thus, after being handled, they are organized in a data frame. (details on notebook)



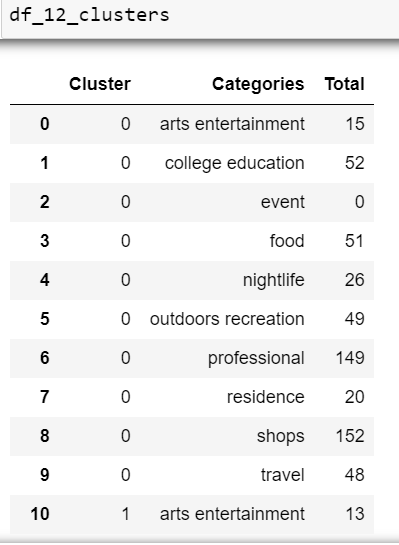
Location of the 12 centroides

To search for categories from a location point first, you need to know how many categories the Fourquare API makes available, so with the code and registration keys previously made it is possible to have this information on a dataframe like the one below.



The 10 categories that the Foursquare API provides

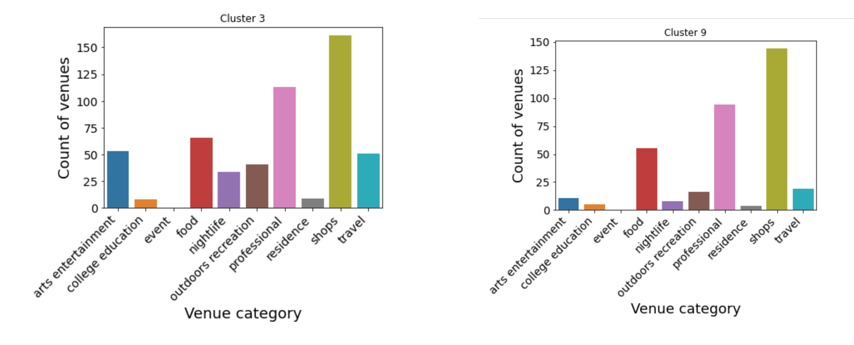
With the information of the categories and knowing the location of the centroides, it is already possible to search how many categories there are for each cluster. After manipulated and organized in a dataframe, this is the result. (Details of this process on the notebook)



Part of the iframe with total categories within a radius of 300 meters from the clusters' centroides

1. **Results**

A way to visualize which categories are within a radius of 300 meters from the centroides of clusters 3 and 9 and through the graph.



Categories within a radius of 300 meters from the centroid

There is a high concentration of categories such as shop and professional in both cases, these categories have establishments such as: Business Center, Distribution Center, Town Hall, Industrial Estate, Wedding Hall etc. To know more details, it is necessary to specify the search even more, which will certainly be for a little while, with this information it is possible to draw some conclusions.

1. **Conclusion**

According to the Brazilian Franchising Association - ABF, the definition of the best location is a little more complex than it appears, because it involves antagonistic variables, such as the flow of people and costs. The best location is not necessarily the one that will provide the highest revenue, but the one that will bring the best result, being one of the factors to be considered in the evaluations given by customers.

Several factors are taken into account when starting a business, such as the flow of people, the level of demand and operating costs. With this, it is possible to conjecture, with the information extracted, which would be the best places to open the business, based on the K-means algorithm, serving as one of the reference factors.

When clustering, k-means returned groups with shorter distances from each other. Associated with the verification of the marks given by the customers, it was observed that the clusters with the highest averages had around them business categories such as shop and professional, which indicates a significant presence of trades and companies and which helps to justify the hypothesis that, well-evaluated and well-frequented places are strategically positioned, in relation to an environment composed of companies and extensive commerce.

This is the contribution that this project offers, and the future interested in the promising business of express salons should also consider other elements, such as the flow of public with interest in this type of service, demand capacity, among others, for a greater possibility of success in choosing the localization.