RECOMMENDING HOTELS BASED ON SIMILAR NEARBY VENUES - CLUSTERING APPROACH

A. INTRODUCTION

The Yellow Woodpecker (fictional company) is a Brazilian travel agency recognized for the quality of its services. In an effort to improve its services and increase customer retention, the company decided to seek together with its Data Science Team, a possible strategy to improve the satisfaction of its customers. During the analysis, it was noticed that the satisfaction rate to travel accommodation of some customers was significantly getting variation, and not always achieve the expected satisfaction. In view of this, the Data Science Team proposed a project to segment all hotels in the destination trip, and then based on the greater satisfaction of accommodation from a previous trip, create a system to suggest Hotels based on the similarity of nearby venues, to increase the likelihood of high satisfaction with the accommodation service provided.

A.2 Data

The data used was obtained from the Foursquare API, which was performed in 3 steps.

First, we will search the entire area for a desired city looking for all the hotels in it, and then saving the following information:

- City
- City latitude
- City longitude
- Hotel name
- Hotel Id
- Hotel latitude
- Hotel longitude

Then, for each hotel found, was make a new request to the API to obtain the evaluation score of each venue (hotel).

- Ratings

Finally, for each hotel, was make a new request to obtain all nearby venues in a 500 meters radius, then save the following information:

- Venue Name
- Venue Latitude
- Venue Longitude

- Venue Category

B. METHODOLOGY

The first step of this Project was use geopy python library to get lat-long coordinates from Paris city - France, which are our reference in the development of this project. The coordinates obtained were:

- Latitude: 48.8566969
- Longitude: 2.3514616

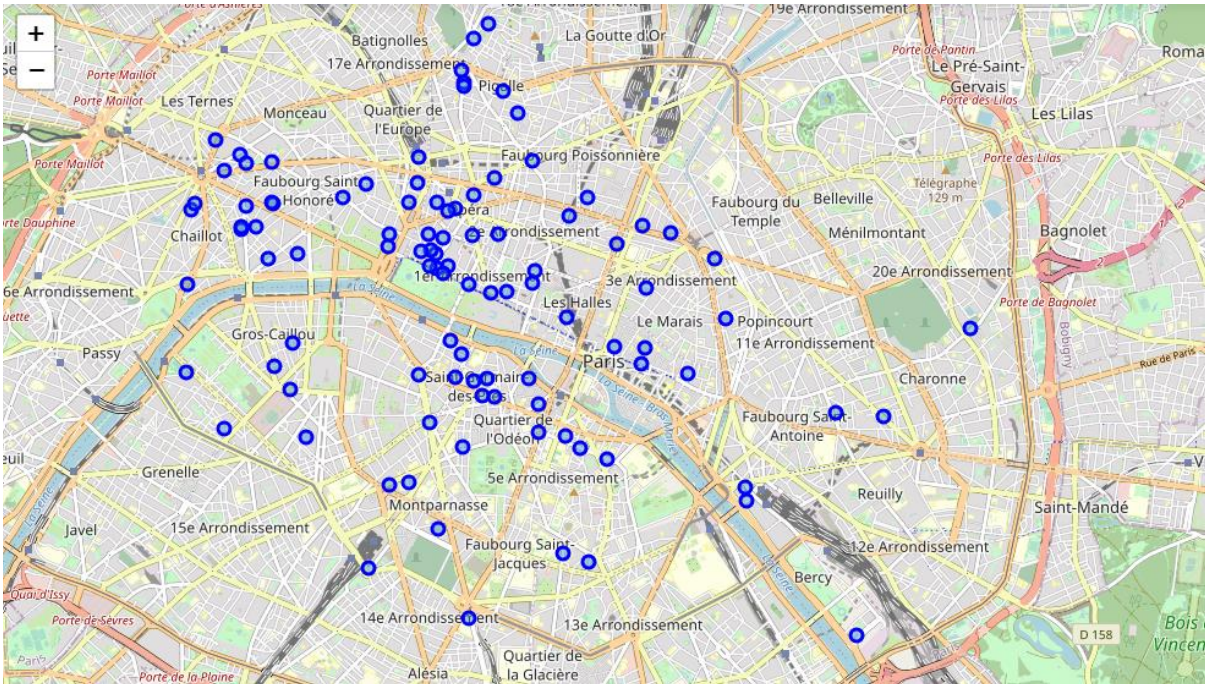
From the Paris coordinates, was used Foursquare API to getting information on 100 Hotels within a radius of 10 Km. The information obtained was, name, Foursquare identification number, latitude, longitude e category. The result header for this query is shown in TABLE 1.

TABLE 1 – PARIS HOTELS

Venue	Venue Id	Venue Latitude	Venue Longitude	Venue Category
Hôtel Le Meurice	4b8b0cbef964a5203b9032e3	48.865333	2.328137	Hotel
Hôtel Bel Ami	4b6ae8dbf964a52066e62be3	48.854918	2.333141	Hotel
InterContinental Paris Le Grand Hôtel	4adcda03f964a520d13121e3	48.870836	2.330725	Hotel
Grand Hôtel du Palais Royal	4b50f4d4f964a5207e3a27e3	48.863183	2.337901	Hotel
W Paris – Opéra	4e8b0ce2f9f464ec8732a56d	48.872098	2.333213	Hotel

With the previously data obtained, the Python Folium library was used to visualize the geographical location of each Hotel, being observed at FIGURE 1.

FIGURE 1 – PARIS HOTELS MAP



After visualizing the hotels, premium requests to Fousquare API were used to obtain the ratings for each hotel. These notes will be used later to classify the Hotels from the best notes. The header of the request result is seen in TABLE 2.

TABLE 2 – HOTELS RATING

Venue	Venue Id	Ratings	Venue Latitude	Venue Longitude	Venue Category
Hôtel Le Meurice	4b8b0cbef964a5203b9032e3	8.7	48.865333	2.328137	Hotel
Hôtel Bel Ami	4b6ae8dbf964a52066e62be3	8.7	48.854918	2.333141	Hotel
InterContinental Paris Le Grand Hôtel	4adcda03f964a520d13121e3	8.6	48.870836	2.330725	Hotel
Grand Hôtel du Palais Royal	4b50f4d4f964a5207e3a27e3	8.3	48.863183	2.337901	Hotel
Hôtel Les Jardins du Marais	4b7051cdf964a5205b122de3	7.8	48.860699	2.368530	Hotel

In sequence, the Foursquare API was used to get up to 100 nearby venues within a radius of 300 meters for each hotel. The result of this request is a list of nearby venues to each analyzed hotel saved in TABLE 3.

TABLE 3 – NEARBY VENUES

	Hotel	Hotel Latitude	Hotel Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Hôtel Le Meurice	48.865333	2.328137	Librairie Galignani	48.864989	2.328570	Bookstore
1	Hôtel Le Meurice	48.865333	2.328137	Hôtel Mandarin Oriental	48.866987	2.327178	Hotel
2	Hôtel Le Meurice	48.865333	2.328137	Balagan	48.865432	2.329680	Israeli Restaurant
3	Hôtel Le Meurice	48.865333	2.328137	Le Dalí	48.865333	2.328137	French Restaurant
4	Hôtel Le Meurice	48.865333	2.328137	Ladurée	48.866121	2.328449	Dessert Shop

Not all machine learning algorithms can work directly with categorical variables, as they need to perform numerical calculations on them. Therefore, it is necessary to look for a way to represent the obtained data numerically. In this Project 'One Hot Encoding' was used, which is a form of numerical representation for categorical data that does not have an ordinal relationship, where a value of 1 is assigned when there is an observation for a given group and 0 for when there is no occurrence.

After that, the data was grouped in 'One Hot Encoding' for each Hotel through the average frequency found for each category. The first rows and columns of the table after such a transformation are shown in TABLE 4.

TABLE 4 – AVERAGE FREQUENCY

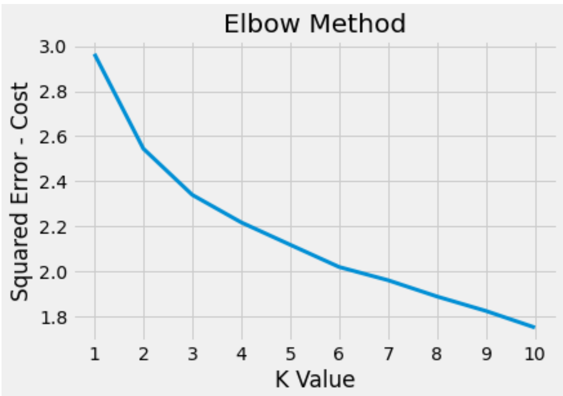
	Hotel	Accessories Store	African Restaurant	Alsatian Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports
0	7 Eiffel Hotel****	0.0	0.0	0.0	0.014085	0.0	0.0	0.000000	0.014085	0.0	0.0	0.014085	0.0
1	Artus Hotel	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.0
2	Champs Elysées Friedland	0.0	0.0	0.0	0.000000	0.0	0.0	0.023810	0.000000	0.0	0.0	0.047619	0.0
3	Citadines Saint-Germain-des-Prés Paris	0.0	0.0	0.0	0.000000	0.0	0.0	0.019608	0.000000	0.0	0.0	0.039216	0.0
4	Cordelia Hotel	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.015152	0.0

With the frequency table properly obtained, the Machine Learning stage began, using the K-Means unsupervised learning technique, which is an algorithm that groups data trying to separate samples into n groups of equal variance, minimizing the known criterion as inertia or through the error of the sum of the squares of a cluster (group). In this algorithm it is necessary to specify previously the k number of clusters (groups) in which the data will be grouped. Therefore, it is needed to find the best k number of clusters that results in the greatest possible similarity between data from the same cluster and at the lowest similarity between the different clusters.

To perform this task, we used several models with k values from 1 to 10, then two metrics were used to evaluate the results, known as, Elbow Method and Silhouette Method.

The first method used was the Elbow Method. This method tests the variance of the data in relation to the number of clusters, in order to present the point k at which the increase in the number of clusters does not represent a significant value of cost reduction when compared to the variance between the previous k. The result of such a method is seen in FIGURE 2.

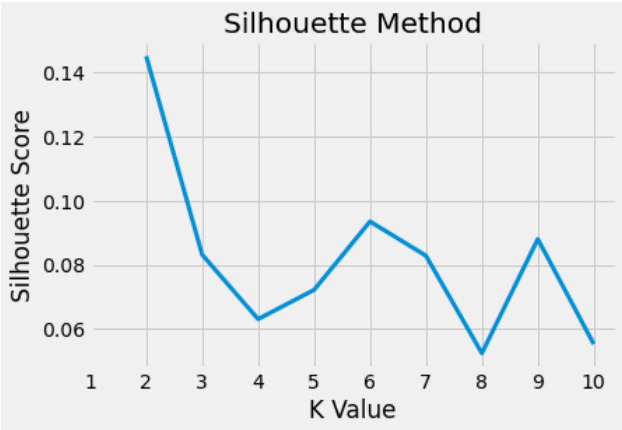
FIGURE 2 – ELBOW METHOD



In some cases (such as the one above), the Elbow method may not clearly represent the best k to be chosen, as this method is a visual inspection of the values, which is not always very clear. The above case, for example, does not make it very clear which k is better, it is only observed that this value is between 2 and 6. Therefore, the ideal is to have more than one form of evaluation.

The second evaluation for finding the best k was performed using the silhouette method, which is used to measure the resulting separation distance between two clusters. The Silhouette coefficient close to 0 represents that the sample is at, or very close to, the border between two neighboring clusters. A value close to 1 indicates that the sample is far from neighboring clusters. The result of this method is seen in FIGURE 3.

FIGURE 3 – SILHOUETTE METHOD



With the graph of FIGURE 3, we observed the highest Silhouette coefficient was achieved for k = 2, thus resulting in 2 clusters, but only 2 clusters for our application would not be very useful, as making a suggestion of hotels based on only 2 groups, limits the variety of choice between different clusters. So, in this project we will use the second-best k (k = 6) according to the Silhouette method, which also coincides with the possible good values observed in the Elbow method. With that we trained our K-Means model with 6 clusters.

C. RESULTS

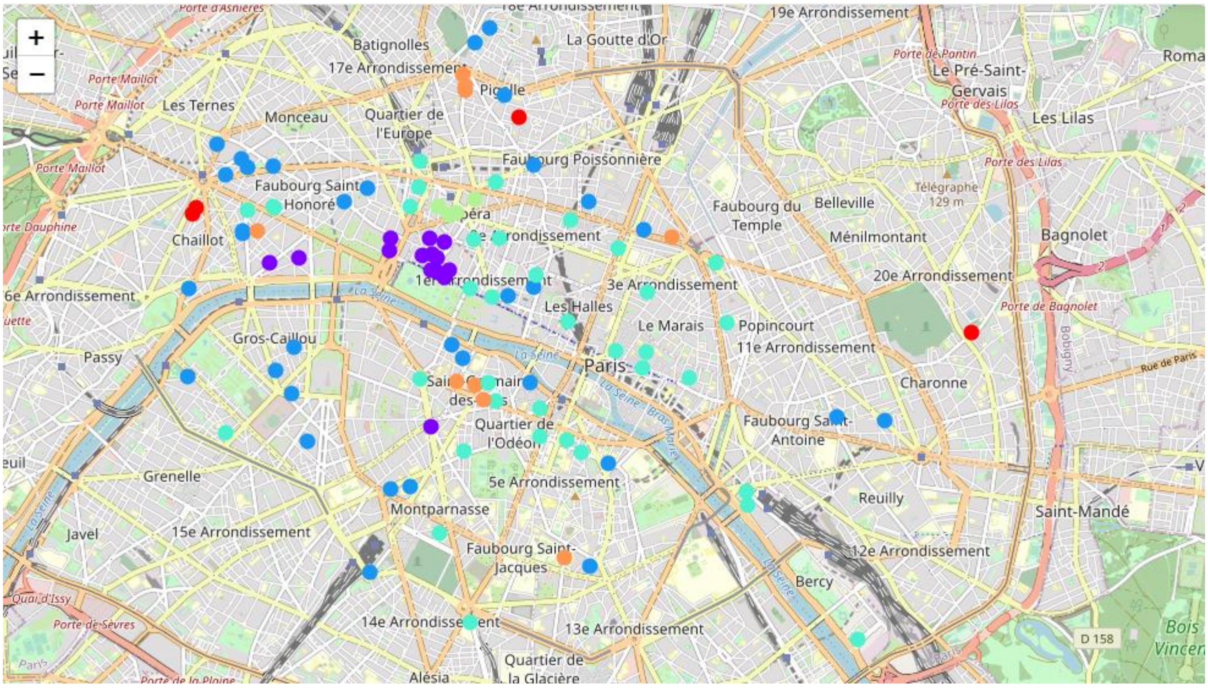
With the result of the K-Means grouping, we were able to obtain the cluster labels belonging to each hotel, and then this result was merged with the table of the 10 most common venues for each hotel, resulting in TABLE 5.

TABLE 5 – 10 MOST COMMON VENUES

Hotel	Hotel Latitude	Hotel Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Hôtel Le Meurice	48.865333	2.328137	1	French Restaurant	Café	Hotel Bar	Chocolate Shop	Japanese Restaurant	Pastry Shop	Beer Garden	Clothing Store	Bookstore	Dessert Shop
Hôtel Bel Ami	48.854918	2.333141	5	Italian Restaurant	French Restaurant	Plaza	Café	Japanese Restaurant	Cosmetics Shop	Clothing Store	Boutique	Sandwich Place	Steakhouse
InterContinental Paris Le Grand Hôtel	48.870836	2.330725	4	Chocolate Shop	Coffee Shop	Clothing Store	Men's Store	Electronics Store	Sandwich Place	Candy Store	French Restaurant	Furniture / Home Store	Roof Deck
Grand Hôtel du Palais Royal	48.863183	2.337901	2	French Restaurant	Plaza	Café	Historic Site	Coffee Shop	Theater	Shoe Store	Spa	Sculpture Garden	Garden
Hôtel Les Jardins du Marais	48.860699	2.368530	3	Bar	Café	French Restaurant	Restaurant	Art Gallery	Clothing Store	Pizza Place	Coffee Shop	Tea Room	Bistro

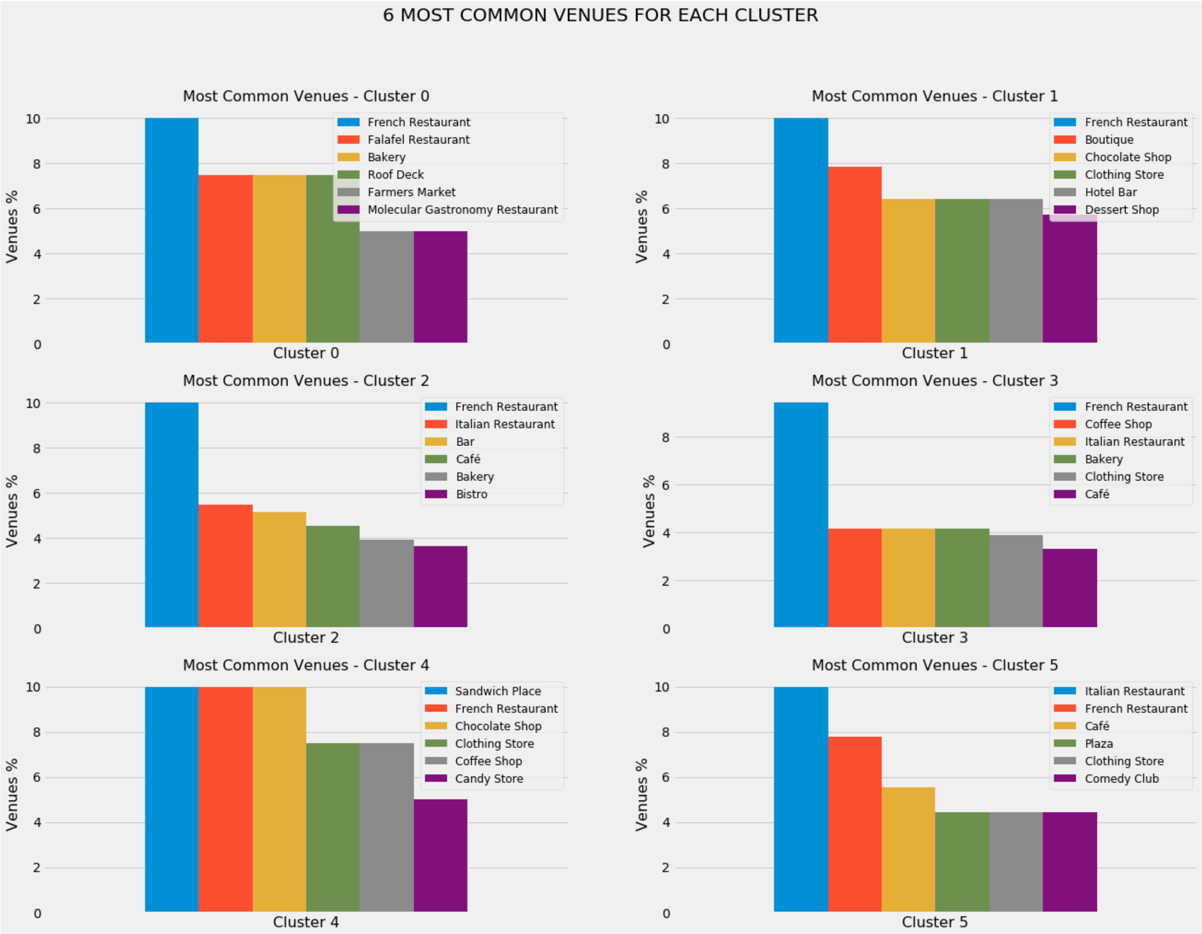
From the Python package Folium, we can see the result of the groupings of the hotels in FIGURE 4, where each color represents one of the 6 clusters.

FIGURE 4 – PARIS HOTELS CLUSTERING MAP



To extract a little more information that represents each cluster, a grouping of locations was performed for each cluster and the graph in FIGURE 5 was plotted, which represents the 6 most common locations for each cluster.

FIGURE 5 – 6 MOST COMMON VENUES FOR EACH CLUSTER



Looking at the graph in FIGURE 5, it was observed that the category 'French Restaurant' appears among the most common in all clusters, being the 1st or 2nd most common venue. This information is interesting and somewhat unexpected, as there is a specific category for French restaurants in a French city, which was not observed similarly in Toronto, as no place classified as Canadian Restaurant was noticed.

As there is such a specific classification for French restaurants, and such a category is suggestively among the most common venues within the city of Paris in France, it may be interesting to remove them from the information used to carry out the grouping, as they may not represent useful information, since it is Expected to have French restaurants around most hotels in France.

D. DISCUSSION

Now that all the hotels are properly grouped, the application proposed in this project can be carried out. So, let's say that a customer has already visited Toronto, and was extremely satisfied with his place of accommodation by assigning a rating of 9.7, we can now use the data from his old trip to suggest a hotel in the city of Paris that has nearby locations similar to that of Toronto.

Toronto hotel location information is shown in TABLE 6.

TABLE 6 – TORONTO HOTEL

	City	City Latitude	City Longitude	Venue	Venue id	Venue Latitude	Venue Longitude	Venue Category
0	Toronto	43.653482	-79.383935	Marriott Downtown at CF Toronto Eaton Centre	4b0563c0f964a5200e5822e3	43.654728	-79.382422	Hotel

From the location of the hotel in Toronto, the same process of preparation and data collection was carried out as applied to hotels in Paris.

With the data of Toronto hotels properly prepared, we used the K-Means model, to classify the hotel in a cluster with the greatest similarity found among the hotels in Paris. The Toronto hotel was classified by the model into cluster 3.

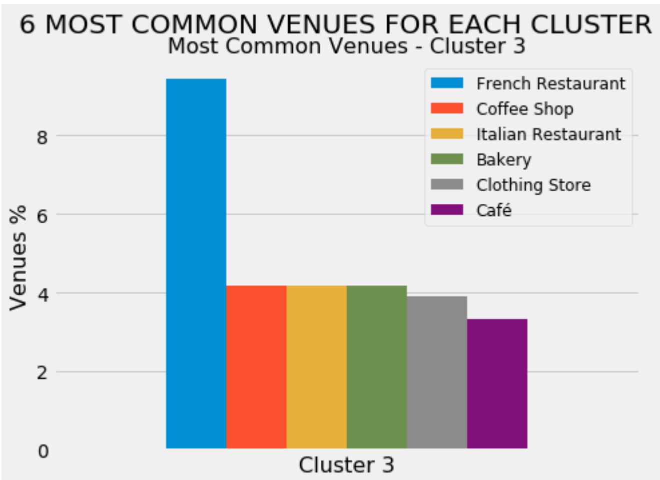
Given this information, we can suggest the top 5 hotels in Paris that have similar nearby venues to Toronto. The suggestion is seen in TABLE 7.

TABLE 7 - RECOMMENDED HOTELS

	Hotel	Ratings	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
27	The Hoxton Paris	9.1	3	Bar	Cocktail Bar	Wine Bar	French Restaurant	Italian Restaurant	Chinese Restaurant	Theater	Salad Place	Gym / Fitness Center	Indian Restaurant
11	Hôtel Barrière Le Fouquet's	8.9	3	French Restaurant	Asian Restaurant	Café	Pastry Shop	Cosmetics Shop	Clothing Store	Electronics Store	Bakery	Hotel Bar	Halal Restaurant
22	Hotel Atmospheres	8.8	3	French Restaurant	Italian Restaurant	Bakery	Coffee Shop	Tapas Restaurant	Portuguese Restaurant	Seafood Restaurant	Ethiopian Restaurant	Flower Shop	Market
23	Hôtel Jules & Jim	8.8	3	French Restaurant	Chinese Restaurant	Café	Art Gallery	Wine Bar	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Japanese Restaurant	Restaurant	Museum
20	Hôtel Caron de Beaumarchais	8.7	3	French Restaurant	Clothing Store	Falafel Restaurant	Pastry Shop	Wine Bar	Plaza	Italian Restaurant	Ice Cream Shop	Bistro	Furniture / Home Store

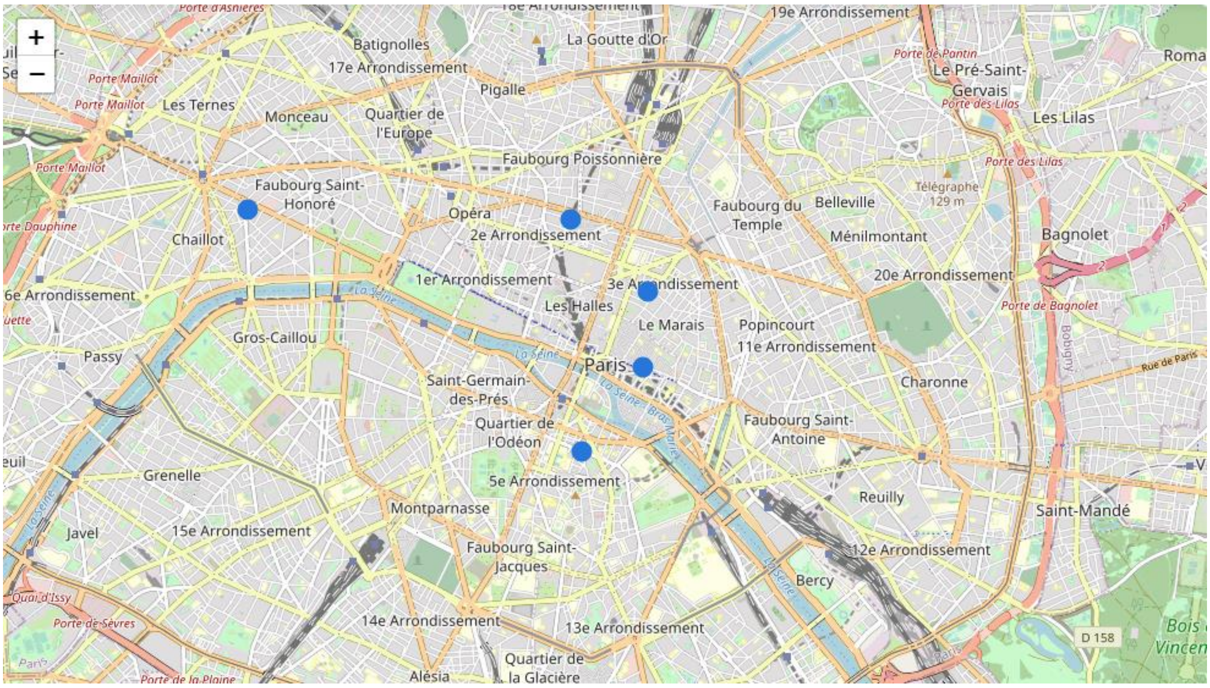
In FIGURE 6, we have the most frequent venues for the predicted cluster.

FIGURE 6 – 6 MOST COMMON VENUES FOR CLUSTER 3



And finally in FIGURE 7, we can see the suggested Hotels.

FIGURE 7 – PARIS RECOMMENDED HOTELS MAP



E. CONCLUSION

With the data obtained via requests to the Foursquare API we were able to collect and group several hotels in the city of Paris, and then with historical data about a given customer, we were able to suggest the best hotels for accommodation based on such data.

Despite the satisfactory result for a first version of this project, there are a lot of things that can be improved in future works, such as using more than one data source, gaining access to a higher number of requests to the Foursquare API, use other metrics to evaluate our model and choose the best number of clusters, perform the segmentation for a set that contains more than one city, in short, there are many possibilities, so now it remains to continue the studies and to improve the projects developed with the knowledge acquired over time.

Thank for your time reading this project :)

F. ACKNOWLEDGEMENTS

My thanks to the Coursera course platform, to IBM and all instructors in the Data Science - Professional Certificate courses, for providing such a courses and materials that enabled the development of this project.

G. REFERENCES

- IBM Data Science Professional Certificate
- One Hot Encode

- An Introduction to Clustering and different methods of clustering
- K-Means - Scikit-learn
- Entenda o algoritmo k-means
- Selecting the number of clusters with silhouette analysis on KMeans clustering
- Forusquare - Documentation for Developers
- Pandas - User Guide
- Folium - Documentation
- A lot of things at StackOverflow :)