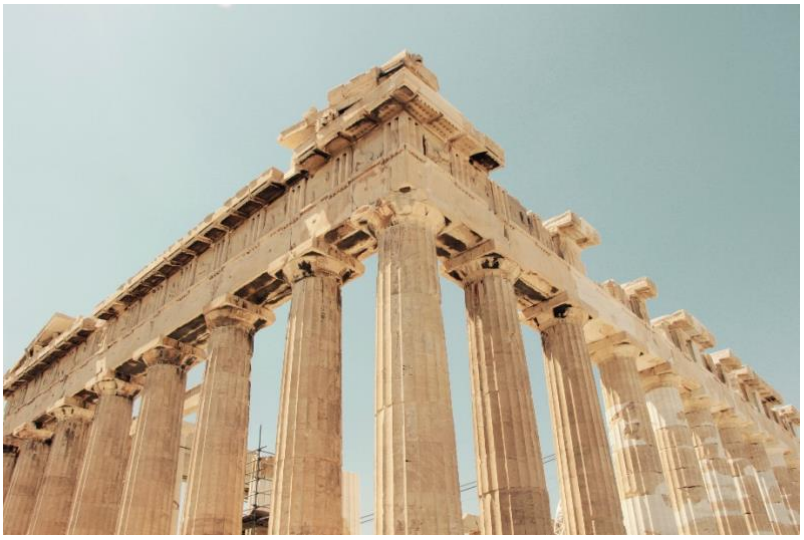
Hotel investment in the neighborhoods of Athens

Thodoris Makridakis

May 28, 2020



1. Introduction

This is the main document for the IBM Data Science Professional certificate capstone project. While scrolling down each code relative to the project will be explained. This project will be accompanied with any possible datasets and other files and there is also a pdf file with the final report and a Power Point presentation. There will be some additional auxiliary files created during this project, such as maps. It will be stated their name and location throughout this project.

After completing successfully the previous eight courses for IBM Data Science Professional Certificate, this capstone allows candidates to demonstrate their abilities and knowledge. The choice for the topic is free and the only restriction is to use Foursquare location data (by using their API) as one of the data sets employed for the project.

2. Business Understanding

Background

Athens is the capital and largest city in Greece. Athens dominates the Attica region and is one of the world's oldest cities, with its recorded history spanning over 3,400 years.

The Municipality of Athens (also City of Athens), which actually constitutes a small administrative unit of the entire city, had a population of 664,046 (in 2011) within its official limits, and a land area of 38.96 km2 (15.04 sq mi). The Athens Urban Area (Greater Athens and Greater Piraeus) extends beyond its administrative municipal city limits, with a population of 3,090,508 (in 2011) over an area of 412 km2 (159 sq mi). According to Eurostat in 2011, the functional urban area (FUA) of Athens was the 9th most populous FUA in the European Union (the 6th most populous capital city of the EU), with a population of 3.8 million people [Wikipedia].

The city is a major center for banking, finance, retailing, trade, media, services and tourism not only in the country but in the Southeast Europe. Greece's primary income source is tourism and so Athens. Greece comes in the 13th place in the world classification of tourist destinations, receiving 34 million tourists in 2019 (according to World Tourism Organization).Tourist's nights in Greece amounted to 236,547 thousand in 2019, compared to 230,727 thousand in 2018, an increase of 2.5%. Specifically Athens which belongs to Attica region is ranked in the 4th place with 34.028 thousands, after Region of South Aegean, Crete and Central Macedonia. [Bank of Greece].

Problem Description

After the successful treatment in Greece concerning COVID-19, Greece is one of the few countries opening the borders for tourism in summer 2020 with great prospects. An international fund plans to expand its operations to Greece and specifically in Athens, which considers not only as a safe choice for tourism in Europe but also a swarming spot for tourists due to the ancient sites and museums. The insights for this analysis will give a better understanding of the spatial analysis of different neighborhoods of the city and will allow to recommend hotel locations to operate a new hotel.

The objective of this capstone project is to analyse and select the best locations in the city of Athens, Greece to open a new hotel. A reliable choice will make the company earn money, with a possible positive effect on the fund's reputation. For the project, data science methodology will be used and machine learning techniques like clustering.

The key indicators employed to analyze Athens' neighborhoods will be all the amenities based on the Foursquare location data.

This project aims to provide solutions to answer the business question: In the city of Athens, if an international fund is looking to open a new hotel, where would you recommend that they open it?

Target Audience

This project is particularly useful to property developers and investors looking to open or invest in new hotels in the capital city of Athens. A recent article in January 2020 from a national famous news site "Ta Nea" wrote that:
"...With a positive seven-year run in terms of abroad travel growth, Athens is constantly welcoming new boutique hotels, as well as four- and five-star hotel units, with industry professionals - domestic and foreign investors - trying to take advantage of its ongoing tourism records' country...."

Success Criteria

The project will be considered successful if a tiered list of Athens' neighborhoods based on the amenities in the neighborhoods can be presented to the client to inform its prospective options of their choices for operating a new hotel in the city.

3. Data Understanding

Geographical Data

In order to process to the spatial analysis we will need the following data:

- A detailed list of neighborhoods in Athens. This list is the base of this project which is related with the capital city of Greece, Athens.
- Then, we will get coordinates (Latitude and Longitude) for every neighborhood. The coordinates are required to plot the map of neighborhoods but also to get the venue data.
- Amenities of the neighborhoods (this category is called "Venue data" in Foursquare), particularly data related to hotels. The amenities will be used in order to perform machine learning and specifically, clustering on the neighborhoods.

For the first bullet, we will get from wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Athens) a list of neighborhoods in Athens, with a total of 63 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and Beautifulsoup packages.

For the second bullet, we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates and combine them with every neighborhood of Athens.

For the last one, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Hotel category in order to help us to solve the business problem put forward. This is a project that will make? use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

In the next section, methodology is explained more thoroughly.

4. Methodology

The first task is to get a detailed list of neighborhoods in the city of Athens. A source is available in the wikipedia page and by web scraping and using request we can extract it. The next step is to combine these neighborhoods with their coordinates in the form of latitude and longitude. The latter step is crucial because the Foursquare will depend on these coordinates. Therefore, using Geocoder will allows us to transform addresses into geographical coordinates. Next procedure is to populate all these data into a pandas Dataframe. Using Folium, a map will be generated visualizing the neighborhoods of Athens.

The second task concerns about enabling Foursquare API in order to get the top 150 venues in radius of 2000 meters. But first you must be registered as Foursquare Developer Account and get the Foursquare ID and Foursquare secret key. Then by making API calls in the neighborhoods of Athens, you can obtain venue data such as venue name, venue category, venue latitude and longitude. After getting the data, we can examine the

number of categories existing in Athens but also the number of each venue category. A quick task is to gather the amenities per category and neighborhood and check their frequency by taking the mean. Last step is to prepare the data about hotels for clustering in the neighborhoods.

The last task is to perform clustering on the data by using k-means clustering. K-means is one of the most common and intuitive clustering algorithms in Machine Learning and identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. In order to choose the optimal K, the elbow method is usually used. In our case, the optimal k was 5. The results from clustering will inform us for higher and lower concentrations of hotels in the neighborhood of Athens. Therefore, the results will provide accurate information in order to deal with the business question of the multinational fund.

5. Data Preparation

After importing all the necessary libraries, we try to find the neighborhoods in Athens city. After parsing the list with BeautifulSoup, we combine the neighborhoods with the coordinates. Next step is to create a map visualizing all the neighborhoods from Athens. Below we can see the neighborhoods from the Folium map.

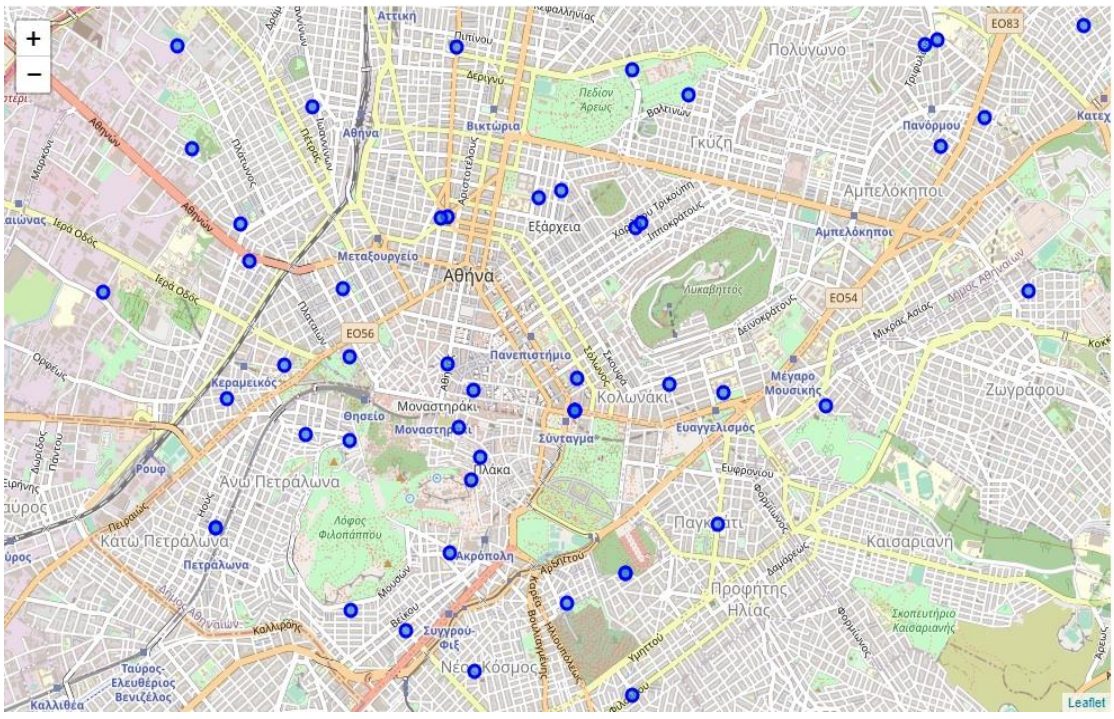


Figure 1 - Neighborhoods in Athens

The critical step is connecting with Foursquare API and setting the radius to 2000 meters and 100 venues max using you Client ID and Client Secret and API version. We can explore more thoroughly the data with the amenities and check the total number of each venue category.

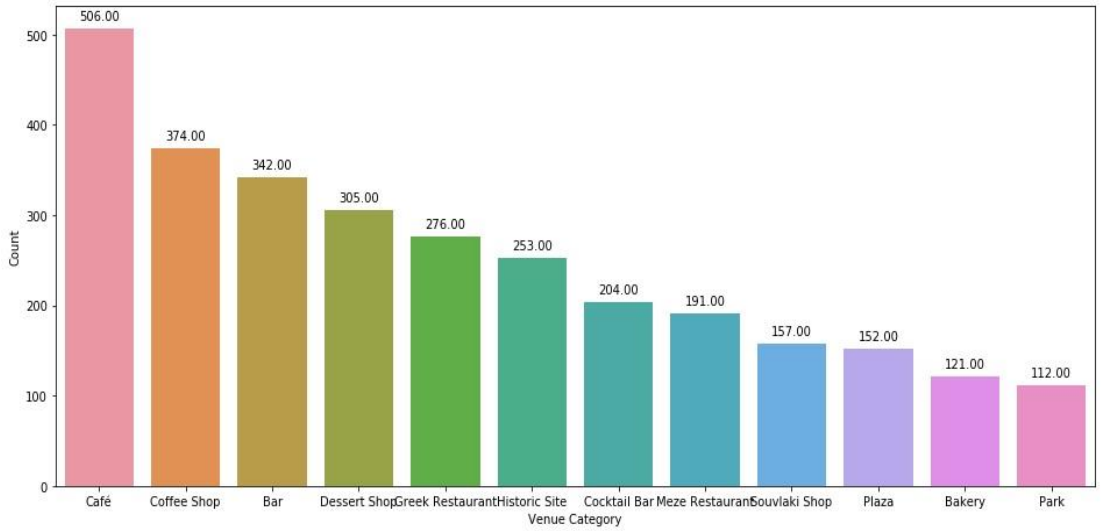


Figure 2 - Total number of each venue category

Another interesting part of the data is exploring the five (5) most common venue categories in each neighborhood.

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
58	Thiseio	Historic Site	Bar	Café	Greek Restaurant	Meze Restaurant
59	Thymarakia	Café	Coffee Shop	Dessert Shop	Greek Restaurant	Souvlaki Shop
60	Treis Gefyres	Café	Plaza	Coffee Shop	Dessert Shop	Bar
61	Vathi, Athens	Bar	Dessert Shop	Coffee Shop	Café	Cocktail Bar
62	Votanikos	Bar	Greek Restaurant	Café	Meze Restaurant	Theater

Figure 3 - Five most common venue categories in each neighborhood

6. Modeling

With all the information we have we will perform a segmentation of the neighborhoods using a K-means clustering algorithm. We will modify our main dataframe with the information from athens_grouped to feed the algorithm.

Applying the elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center. After running Elbow method, it seems like the optimal number of clusters could be 5. The next step is using k-means for 5 clusters and then conclude

Table 1 – Cluster labeling for the first 5 neighborhoods in Athens

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
0	Aerides, Athens	0.02	1	37.97614	23.73640
1	Agios Eleftherios, Athens	0.00	0	38.01997	23.72627
2	Agios Panteleimonas, Athens	0.01	3	37.99672	23.72741
3	Akadimia Platonos	0.00	0	37.99095	23.70718
4	Akadimia, Athens	0.00	0	37.98669	23.71091

7. Evaluation

Examining the clustered data to see how the different neighborhoods are segmented. We can see clearly the five neighborhoods' clusters and next step is to analyze each cluster in order to get more accurate information for the possible location of the new hotel.

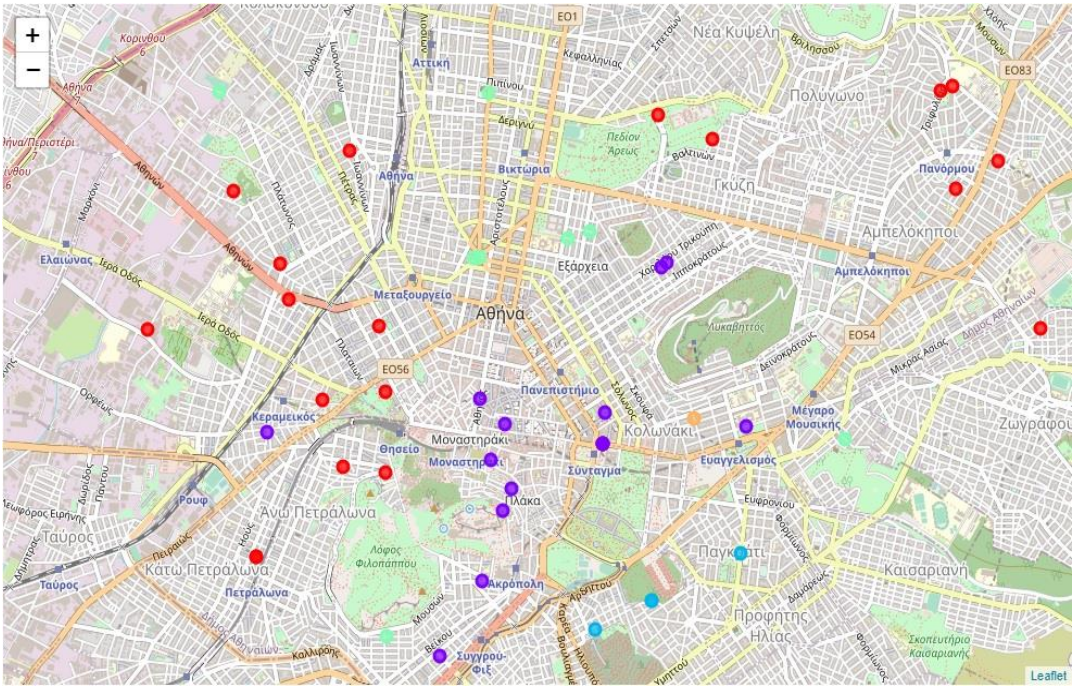


Figure 4 - Clustering map of neighborhoods

Cluster 0 (red color **)**

- The cluster 0 is the biggest cluster among the given clusters.
- It is the densest among the clusters and the cluster becomes sparse as we move inside the city
- The frequency of Hotels in these neighborhoods is very low
- Cluster 0 has some interesting neighborhoods near the center of Athens to operate a new hotel such as Thiseio, Kerameikos

Table 2 - List of neighborhoods for Cluster 0

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
1	Kypriadou	0.0	0	37.993.975	23.745.041
2	Gyzi	0.0	0	38.020.475	23.845.250
3	Kerameikos	0.0	0	37.979.190	23.719.200
4	Kolonos	0.0	0	37.993.340	23.716.370
5	Kynosargous	0.0	0	38.023.996	23.673.818
6	Metaxourgeio	0.0	0	37.983.040	23.718.690
7	Nea Filothei	0.0	0	37.996.820	23.763.110
8	Patisia	0.0	0	38.020.190	23.735.610
9	Pedion tou Areos	0.0	0	37.995.420	23.740.740
10	Petralona	0.0	0	37.969.480	23.708.970
11	Polygono, Athens	0.0	0	38.002.620	23.751.310
12	Profitis Daniil, Athens	0.0	0	37.984.610	23.711.620
13	Profitis Ilias, Athens	0.0	0	38.028.330	23.748.320
14	Rizoupoli	0.0	0	38.029.000	23.738.770
15	Thiseio	0.0	0	37.974.810	23.715.920
16	Thymarakia	0.0	0	37.979.480	23.643.680
17	Gouva, Athens	0.0	0	37.960.030	23.740.770
18	Goudi	0.0	0	37.982.920	23.770.980
19	Votanikos	0.0	0	37.982.830	23.700.430
20	Gazi, Athens	0.0	0	37.978.720	23.714.260
21	Asteroskopeio	0.0	0	37.974.470	23.719.200
22	Ano Petralona	0.0	0	37.969.480	23.708.970
23	Elaionas	0.0	0	37.966.670	23.669.400
24	Ellinoroson	0.0	0	37.997.930	23.775.160
25	Ampelokipoi, Athens	0.0	0	37.991.060	23.764.290
26	Asyrmatos, Athens	0.0	0	37.931.620	23.721.151
27	Erythros Stavros	0.0	0	37.992.681	23.767.639
28	Akadimia, Athens	0.0	0	37.986.690	23.710.910
29	Akadimia Platonos	0.0	0	37.990.950	23.707.180
30	Agios Eleftherios, Athens	0.0	0	38.019.970	23.726.270
31	Girokomeio, Athens	0.0	0	37.997.110	23.764.060

Cluster 1 (purple color **)**

- Cluster 1 is the 3rd cluster with the highest frequency of hotels in neighborhoods
- Cluster 1 is located mainly in the heart of Athens
- Therefore, the competiveness in this cluster is moderate higher

Table 3- List of neighborhoods for Cluster 1

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
1	Rouf, Athens	0.02	1	37.976.830	23.709.890
2	Skouze Hill	0.02	1	37.977.296	23.728.682
3	Anafiotika	0.02	1	37.972.240	23.728.520
4	Plaka	0.02	1	37.973.520	23.729.150
5	Psyri	0.02	1	37.978.760	23.726.670
6	Neos Kosmos, Athens	0.02	1	37.961.420	23.728.710
7	Attiki, Athens	0.02	1	37.986.490	23.741.040
8	Evangelismos, Athens	0.02	1	37.977.150	23.747.730
9	Neapoli, Athens	0.02	1	37.986.780	23.741.480
10	Makrygianni, Athens	0.02	1	37.968.070	23.726.870
11	Koukaki	0.02	1	37.963.720	23.723.530
12	Aerides, Athens	0.02	1	37.976.140	23.736.400
13	Mount Lycabettus	0.02	1	37.977.980	23.736.520
14	Monastiraki	0.02	1	37.975.220	23.727.560

Cluster 2 (cyan color **)**

- Cluster 2 is the no.1 cluster with the highest frequency of hotels
- These neighborhoods are richer compared with the cluster 0, 1 and 3
- The competiveness in this cluster is the highest among all
- All these neighborhoods are next to the other

Table 4 - List of neighborhoods for Cluster 2

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
1	Kallimarmaro	0.04	2	3.796.691	2.374.022
2	Pangrati	0.04	2	3.796.973	2.374.730
3	Mets, Athens	0.04	2	3.796.522	2.373.581

Cluster 3 (light green color **)**

- Cluster 3 has the least richer neighborhoods but also with bad reputation

- It is a moderate clusters and the cluster becomes denser as we move outwards of the city (northwestern of Athens)
- This cluster has the lowest real estate prices for operating a new hotel but with the risk of bad reputation
- Cluster 3 is the 3rd cluster with the most neighborhoods

Table 5 - List of neighborhoods for Cluster 3

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
1	Vathi, Athens	0.01	3	37.987.030	23.726.162
2	Omonoia, Athens	0.01	3	37.987.090	23.726.720
3	Sepolia	0.01	3	38.004.770	23.717.940
4	Kypseli, Athens	0.01	3	38.003.280	23.740.600
5	Philopappou	0.01	3	37.964.825	23.719.321
6	Agios Panteleimonas,	0.01	3	37.996.720	23.727.410
7	Ilisia, Athens	0.01	3	37.976.410	23.755.570
8	Exarcheia	0.01	3	37.988.170	23.733.610
9	Kountouriotika	0.01	3	37.988.563	23.735.319
10	Kolokynthou	0.01	3	37.996.760	23.706.090

Cluster 4 (orange color **)**

- A plausible explanation for the smallest cluster is due to the special characteristics of this neighborhood
- Kolonaki is considered as the most expensive neighborhood among the others in Athens
- It has the 2nd highest frequency of hotels in Athens
- Most hotels in Kolonaki are 4* and 5* stars
- Competiveness in Cluster 4 is very high regarding that consists of only one neighborhood

Table 6 - List of neighborhoods for Cluster 4

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
1	Kolonaki	0.03	4	3.797.762	2.374.359

8. Conclusion

The primary goal of this project was to classify the neighborhoods of Athens based on amenities in order to give information to the international fund about operating a new hotel after the successful treatment of COVID-19 in the country. Greece and particularly Athens is considered as a safe choice for tourism after the pandemic of COVID-19 in Europe.

Information about amenities in the neighborhoods were collected from Foursquare and then the data was converted to a data frame which was normalized and fed to a K Means clustering algorithm that segmented the neighborhoods in the optimal number of clusters using the elbow method, which was five clusters.

Finally, the clusters were examined to search common features for the neighborhoods, as shown in the previous section. Final selection of operating a new hotel will be performed by the international fund.

A suggestion based on this project will be in the Cluster 0 in the neighborhoods very close to the center of Athens. The reasons in order to operate a new hotel there are 3:

- The competiveness is very low in these neighborhoods with great prospects of profit
- In a very close distance, there are ancient monuments and museums but also the center of Athens
- There is a dense public transport network connecting all main destinations not only in Athens but in other districts

Limitations

About 3 neighborhoods in the list of Athens are not located in the correct position due to wrong coordinates from geocoder and 2 are double. I don't think that they play an important role in clustering because they are in the same cluster.

Future Research

Possible suggestions for future research with the inclusion of parameters:

- Real estate prices
- Income
- Crime rates