Shivani Chitre

IBM professional Certificate Capstone

Opening a coffee shop/ cafe in Athens, Greece

June 2020

Table of contents	Index

Introduction-
    1.1.   Background
    1.2.   Business Problem
    1.3.   Project Idea
    1.4. Target Audience	

Data	
Methodology	
Results	
Discussion	
Limitations and Suggestions for Future Research	
Conclusion	
References	

1. Introduction

1.1 Background:

Athens is the capital and largest city of Greece, as well as its cultural, industrial, and financial center. Athens dominates the Attica region and is one of the world's oldest cities, with its recorded history spanning over 3,400 years. It is located in the central Athens region of the country, at coordinates: 37.983972, 23.727806 . Athens sprawls across the central plain of Attica that is often referred to as the **Athens Basin** or the **Attica Basin** [1] Furthermore, Athens is divided into several regions as follows:

Regional unit	Population (2011)
Central Athens	1,029,520
North Athens	592,490
South Athens	529,826
West Athens	489,675
Piraeus	448,997
East Attica	502,348
West Attica	160,927

Athens is the center of the Greece economy and industry, accounting for Athens Metropolitan Area, with an area of 2,928.717 km² (1,131 sq mi) and inhabited by 3,753,783 people in 2011. Athen consists of the following neighborhoods:

Athenians the most popular way of dividing the downtown is through its neighborhoods such as Pagkrati, Ambelokipi, Goudi, Exarcheia, Patissia, Ilissia, Petralona, Plaka, Anafiotika, Koukaki and Kypseli, each with its own distinct history and characteristics.:

Neighborhood	Longitude	Latitude
Pagkrati	37.968056	23.743889
Ambelokipi	37.989722	23.764722
Goudi	37.988333	23.775833
Exarcheia	37.986111	23.734722
Patissia	38.033333	23.733333
lissia	37.976944	23.756944
Petralona	37.967222	23.709722
Plaka	37.972222	23.730556
Anafiotika	37.972222	23.727778
Koukaki	37.9625	23.720833
Kypsel	38.0025	23.741389

1.2. Business Problem

There are so many cafe locators such as google, yelp, zomato including FourSquare, that will tell you all about all the cafes in a nearby area when you do a quick search. Cafes are loved by a variety of diverse populations. They make a good hangout spot or even for a quick professional meeting. Especially in a tourist area, cafes are a popular destination to try something local or get fast refreshment from all the traveling. This is where I come in from a business professional's perspective.

Imagine the following scenario:

You would like to invest in a coffee shop in Athens, Greece to start your own business. where would you open it up? There is so much competition already out there.

I would start by researching some geographical and statistical data about the Country and the city. Then I will examine the area of interest and its population, neighborhoods.

1.3 Project Idea

The Question/Problem is to find a solution to where to open up a coffee shop that does not have too much competition but has a good amount of tourists/locals. There are lots of cafes in Athens we will try to detect locations that are not already crowded with coffee stores or cafes

My idea for the Capstone Project is to show how to get venue and location data from FourSquare and analyze that data to come up with potential locations to open up a cafe. A high-level approach is as follows:

1. The business investors pick a city location [for my project I will use athens]
2. The foursquare website is scrapped for the top cafes in the city
3. From this list of top cafes, the list is augmented with additional geographical data
4. Using this additional geographical data the top nearby cafes are selected
5. The number of cafes of all neighborhoods is obtained
6. A map is presented to the user showing all the cafes in the vicinity

1.4 Target Audience

The coffee shop will be for all ages mostly Tourists, working professionals, bloggers, and locals. Age group-wise- teenagers, young adults, or even the older population.

2] Data

In this section, I will describe the data used to solve the problem as described previously

In this project, I will determine which place is good for opening a coffee shop in Athens, Greece.

1. I will convert address data into their equivalent latitude and longitude values.

2. For Athens neighborhood data, I will use wikipedia,

   https://en.wikipedia.org/wiki/Category:Districts of Greece

3. I will use the Foursquare API to explore Athens neighborhoods and to get cafe venues in neighborhoods.

4. I will use the Foursquare API to get venue ratings and likes in neighborhoods.

5. I will use the k-means clustering and Agglomerative algorithms to complete the clustering task

6. I will use the Folium library to visualize the neighborhoods, venues, clusters in Athens

The goals are as follows:

- Download and Explore Neighborhood Dataset

- Load and Analyze Venues of Neighborhoods in Athens

- Cluster Neighborhoods using kmean Clustering

- Discussion

- Conclusion

From Wikipedia pages, I gathered information about the main city of Athens and neighboring areas. Then we will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude

and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the cafe data in those neighborhoods as well as Athens. Foursquare has one of the largest databases 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 cafes 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, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did, and the machine learning technique that was used.

3. Methodology

To begin with, I started by collecting the coordinate values of Athens, Greece as mentioned above by using the geolocator package and found the geographical coordinates of Athens, Greece to be 37.9839412, 23.7283052.

Next, using the BeautifulSoup package I retrieved the total number of regions around Athens. Using wikipedia, I was able to get the data for regions around Athens. I made a CSV file and loaded it into jupyter notebook to create a data frame. There are 7 regions around Athens.

Regional unit	Population (2011)
Central Athens	1,029,520
North Athens	592,490
South Athens	529,826
West Athens	489,675

Piraeus	448,997
East Attica	502,348
West Attica	160,927

Following that, in a similar manner, I collected all the neighborhood data along with their respective coordinates to find out all the locations of the cafes, or coffee shops. These are the most popular neighborhoods from the main central Athens.

Neighborhood	Longitude	Latitude
Pagkrati	37.968056	23.743889
Ambelokipi	37.989722	23.764722
Goudi	37.988333	23.775833
Exarcheia	37.986111	23.734722
Patissia	38.033333	23.733333
lissia	37.976944	23.756944
Petralona	37.967222	23.709722
Plaka	37.972222	23.730556
Anafiotika	37.972222	23.727778
Koukaki	37.9625	23.720833
Kypsel	38.0025	23.741389
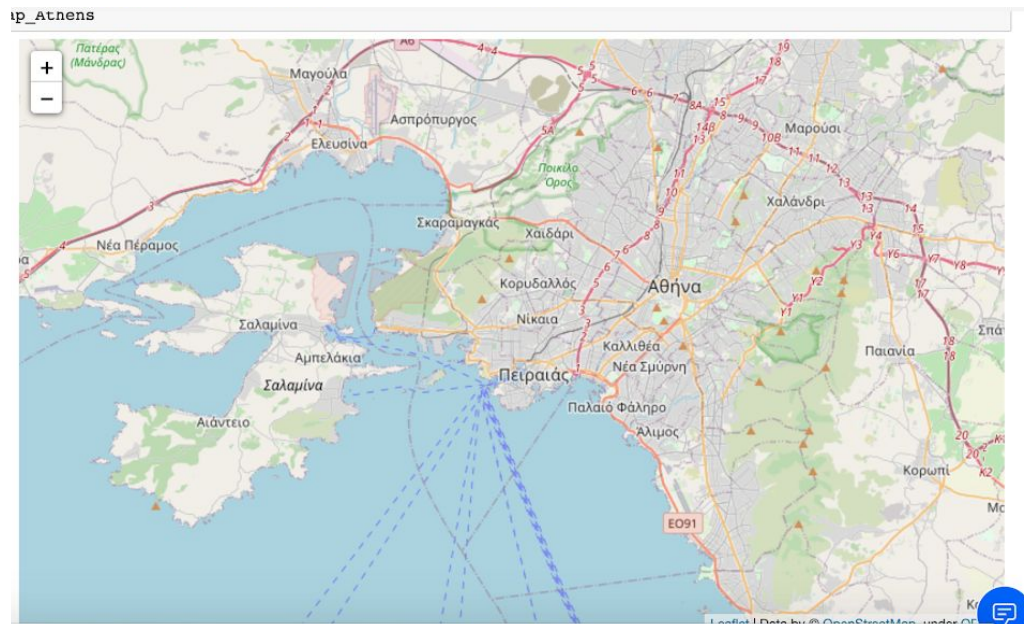
Using folium package, I created a map for all the neighborhoods. (as pictured above).

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 5000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude, and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues.

After collecting data using foursquare API, I had to clean up the data by only keeping the necessary information. I kept only the columns that included the venue name, and anything that is associated with the location. I got the following results.

```
dataframe_cafe.head(100)
```

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	lab
0	Hondos Cafè	Café	Πλ. Ομονοίας	GR	NaN	Ελλάδα	NaN	91	[Πλ. Ομονοίας, Ελλάδα]	[{'label': 37.98470
1	Break Cafe	Café	Πλατεία Ομονοίας 18	GR	Αθήνα	Ελλάδα	NaN	6	[Πλατεία Ομονοίας 18, Αθήνα, Αττική, Ελλάδα]	[{'label': 37.98
2	Café Concerto	Café	Μάγερ 18	GR	Αθήνα	Ελλάδα	NaN	455	[Μάγερ 18, 104 38 Αθήνα, Αττική, Ελλάδα]	[{'label': 37.98744
3	Public Café	Café	Καραγιώργη Σερβίας 1	GR	Αθήνα	Ελλάδα	Πλατεία Συντάγματος 1	986	[Καραγιώργη Σερβίας 1 (Πλατεία Συντάγματος 1),...	[{'label': 37.97619
4	Polis Art Cafe	Café	Πεσματζόγλου 5	GR	Αθήνα	Ελλάδα	στο Αίθριο της Στοάς Βιβλίου	355	[Πεσματζόγλου 5 (στο Αίθριο της Στοάς Βιβλίου)...	[{'label': 37.98155
5	Διπλό Cafe	Café	Θεμιστοκλέους 70	GR	Αθήνα	Ελλάδα	NaN	606	[Θεμιστοκλέους 70, 114 73 Αθήνα, Αττική, Ελλάδα]	[{'label': 37.98629

Furthermore, I created a map showing heatmap/density of restaurants and try to extract some meaningful info from that. Also, I depicted the borders of Greece on our map and a few circles indicating a distance of 2km, 4km, 6km and 10km from Athens center

Red markers are the cafes. (**Heatmap**)

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 2 clusters based on their frequency of occurrence for "cafes". The results will allow us to identify which neighborhoods have a higher concentration of cafes while which neighborhoods have fewer cafes. Based on the occurrence of local cafes in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open a new coffee shop.
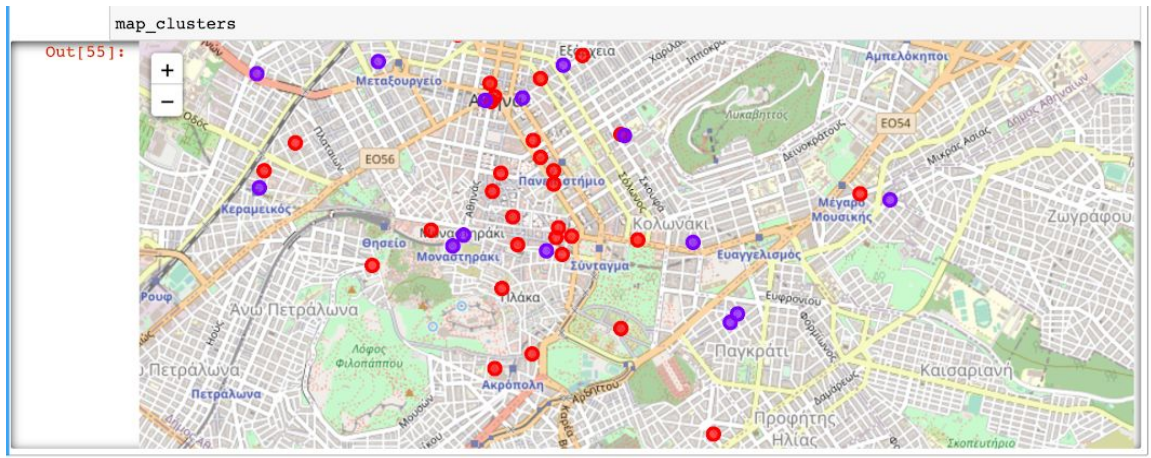
4. Results

After using the foursquare API, I found that there were 50 cafe locations in Athens and nearby neighborhoods. From the data, we can see that the central Athens neighborhood has most cafes. Kypseli and Pangrati do not have a lot of coffee shops but are also popular areas in Athens in the central region as well. There are 9 neighborhoods in Athens. we can choose one of these to open a cafe. I will try to open a cafe in the most populated area which gets a lot of tourists/locals. From the heatmap above, we can see that there are too many cafes in the immediate central Athens vicinity. Enough cafes between the main city towards the museum. However, there are not many cafes near the Pangrati neighborhood. cafes are marked with red circles. Cafes are located pretty much in all directions from the main city including South, North, East, between few miles of the center. There are not enough cafes outside of the man city area. Pangrati is a neighborhood in Central Athens, Greece with an estimated 35,173 residents. Pangrati has attracted artists from all over Greece, who arrive in the city of Athens to educate themselves and to seek inspiration from its vast pool of artistic resources and galleries. Poets, novelists, writers, painters, composers, musicians made Pangrati one of the most important artistic hubs of Athens.

The results from the k-means clustering show that we can categorize the neighborhoods into 2 clusters based on the frequency of occurrence for "cafes" Cluster 0 (red): Neighbourhoods with a lot of cafes and Cluster 1 (purple): Neighbourhoods with not that many of cafes.

5. Discussion :

Observations:

Most of the cafes are concentrated in the central area of Athens city, with the highest number in cluster 0 and moderate number in cluster 1. On the other hand, cluster 1 has a very low number of cafes in the neighborhoods and they are spread out. This represents a great opportunity and high potential areas to open new coffee shops in the areas I suggested before as there is very little to no competition. Meanwhile, cafes in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of other similar locations. From another perspective, this also shows that the crowding of cafes mostly happened in the central area of the city which was the most popular tourist attraction so it makes sense. The neighboring areas a few kilometers away from the central city has fewer stores yet has a high potential. Therefore, this project recommends property developers to capitalize on these findings to open a new coffee shop in neighborhoods in cluster 1 with little to no competition. Kypseli and Pangrati do not have a lot of coffee shops but are also popular areas in Athens in the central region as well. Lastly, property developers are advised to avoid neighborhoods in cluster 0 (Main Athens) which already have a high concentration of cafes, restaurants, and are facing adverse competition.

6. Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of cafes, there are other factors such as population, the income of residents, influx of tourists, and climate that could influence the location decision of a coffee shop. However, to the best knowledge of this researcher, such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a cafe. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid accounts to bypass these limitations and obtain more results.

7. Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 2 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new cafe. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new cafe. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new cafe in Athens, Greece.

8. References

Wiki Athens, Greece. *Wikipedia*. Retrieved from

https://en.wikipedia.org/wiki/Athens

https://en.wikipedia.org/wiki/Pangrati

https://en.wikipedia.org/wiki/Kypseli,_Athens

Foursquare Developers Documentation. *Foursquare*. Retrieved from
https://developer.foursquare.com/docs