Capstone Project - The Battle of Neighborhoods: Athens

Konstantinos Chronas

May 14, 2021

Introduction

The city of Athens is chosen for the analysis of this project, because the author has domain expertise of the venues located in Athens since he grew up there. Athens is the capital of Greece and had a population of 664.046 (in 2011). The city has great history and receives a great number of tourists every year. For example, in 2019 Athens received 6 million tourists in 2019 according to the data statistics. Additionally, the neighbourhood of the city hosts a lot of different venues. Therefore, finding the places that correlate with people interests and needs could be a challenge.

The goal of this project is to provide a list with all the similar neighbourhoods of Athens. This would help the tourists to get an initial overview of the city and be able to find easier the venues they would like visit. Additionally, it could help the people who would like to open a business to find the beachhead market that will bring regarding the area of where they should open.

Data Description

The data that will be used for the project would be a list of the names of the neighbourhoods and cities of Athens and Attica respectively, with their latitudes and longitudes parameters. Those will be used to gather all the relevant venues in the city by using the foursquare API to find the Venues in Athens and cities of the Attica region as was requested by the assignment. Foursquare is a US tech company from New York focusing on location data. Their technology and data help global companies and developers to build better user experiences powered by the location data they have gathered [1] .

To gather the data for the analysis of the Attica region and Athens for this project, the list of areas was used from their own Wikipedia pages [2] [3]. Therefore, first, it was needed to collect the text data from the webpage using the BeautifulSoup package [4] and after to find the Latitude and Longitude coordinates the geopy

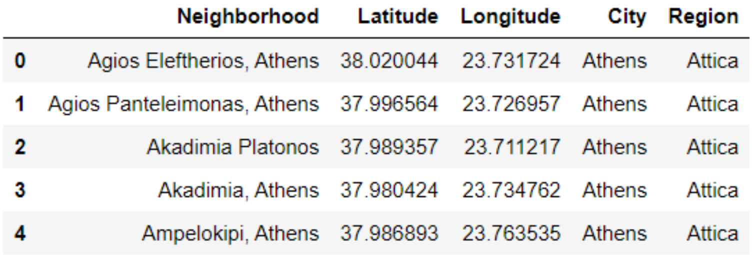
[1] https://foursquare.com/about/

[2] https://en.wikipedia.org/wiki/List_of_settlements_in_Attica

[3] https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Athens

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

package [5], which is a client for several geocoding web services like OpenStreetMap Nominatim [6] to find the Latitude and Longitude coordinates. The following figures showcase the two initial data sets required for the analysis:



	Neighborhood	Latitude	Longitude	City	Region
0	Agios Eleftherios, Athens	38.020044	23.731724	Athens	Attica
1	Agios Panteleimonas, Athens	37.996564	23.726957	Athens	Attica
2	Akadimia Platonos	37.989357	23.711217	Athens	Attica
3	Akadimia, Athens	37.980424	23.734762	Athens	Attica
4	Ampelokipi, Athens	37.986893	23.763535	Athens	Attica

Figure 1: Athens Neighbourhoods Dataset

The Athens neighbourhood dataset is incomplete; thus, the municipality of Athens provides a dataset with a complete list of the neighbourhoods [7]. However, the names are in the Greek language and the target group of the project might not be accustomed to the Greek alphabet therefore manual translation of the names would be a solution. This would require a lot of time from the author, and it is outside of the scope of this project.

The next step is to acquire a dataset with the venues that each neighbourhood and city hosts. As was mentioned above, the foursquare API is used [8]. The API requires the user to define the radius of the search, defining a high value for the radius will result in overlapping venues for each neighbourhood or city and thus create a bias model. Thus, the value of the radius has to be plotted on a map to carefully define its value so that similar neighbourhoods would not have many same venues.

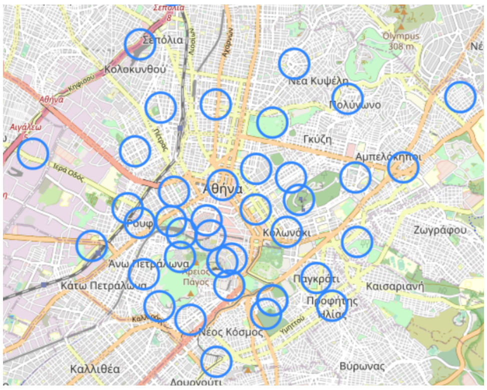
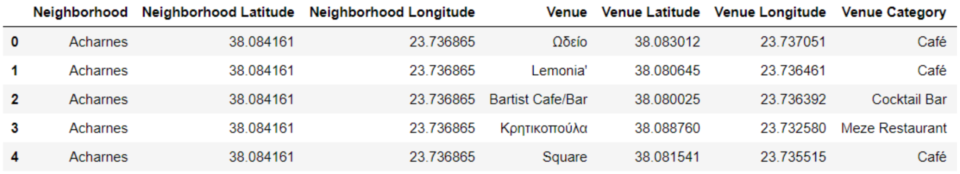


Figure 2: This figure showcases the map radius for the neighbourhoods in Athens

[5] https://pypi.org/project/geopy/
[6] https://nominatim.org/
[7] https://geodata.gov.gr/dataset/op1a-euvo1k1wv
[8] https://developer.foursquare.com/docs/places-api/

Additionally, to narrow down the categories of the resulted venues to align with the interest of tourists such as finding bars and restaurants, venues categories such as stores and playgrounds were filtered out.



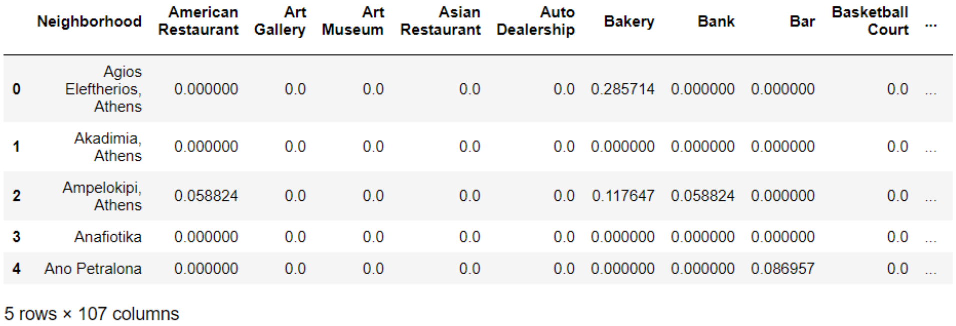
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Acharnes	38.084161	23.736865	Ωδείο	38.083012	23.737051	Café
1	Acharnes	38.084161	23.736865	Lemonia'	38.080645	23.736461	Café
2	Acharnes	38.084161	23.736865	Bartist Cafe/Bar	38.080025	23.736392	Cocktail Bar
3	Acharnes	38.084161	23.736865	Κρητικοπούλα	38.088760	23.732580	Meze Restaurant
4	Acharnes	38.084161	23.736865	Square	38.081541	23.735515	Café

Figure 3: Athens Venues Dataset

Methodology

To find and group similar neighbourhoods and combine them with the associated labels the K-Means clustering algorithm will be used[1]. The K-means algorithm creates clusters of n observation into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centroid). It tries to create clusters of equal variance by minimizing within-cluster variances (squared Euclidean distances).

To use the algorithm, we need to create a dataset that is relevant for the analysis. From the venues datasets that were gathered the only feature that provides discreet information is the venue category feature. Therefore, one hot encoding was used to create new features for each category in the Venues Category column, and then the dataset was grouped by the Neighbourhoods column showcasing the mean value of each category for the neighbourhoods or cities.



	Neighborhood	American Restaurant	Art Gallery	Art Museum	Asian Restaurant	Auto Dealership	Bakery	Bank	Bar	Basketball Court	...
0	Agios Eleftherios, Athens	0.000000	0.0	0.0	0.0	0.0	0.285714	0.000000	0.000000	0.0	...
1	Akadimia, Athens	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	...
2	Ampelokipi, Athens	0.058824	0.0	0.0	0.0	0.0	0.117647	0.058824	0.000000	0.0	...
3	Anafiotika	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	...
4	Ano Petralona	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.086957	0.0	...

5 rows × 107 columns

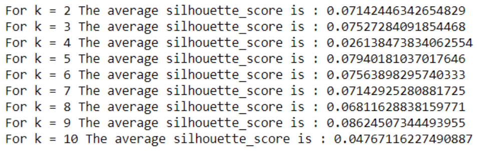
Figure 4: Athens onehot encoding of Venues Category

The next step was to define the number of k clusters to be used. The question now is how to determine the optimal number of clusters for our dataset. In this project, the following two methods were used: the elbow and the silhouette coefficient score methods[2].

- **Silhouette Coefficient Score:** shows the model with better-defined clusters and is calculated by b -a/max (a,b) where a is the mean distance between a sample point and all others in the same class and b mean distance between a sample and all other points in the next nearest class.

- **Elbow method:** for the elbow method we have to calculate and visualize the values o inertia which is the sum of squared distances of samples to the closest cluster center and select the value of k at the “elbow” ie the point after which the inertia start decreasing in a linear fashion.

Fig 5 and fig 4 showcase the results of the search for the value of k from the range of 10 k for Athens Neighbourhoods.



```
For k = 2 The average silhouette_score is : 0.07142446342654829
For k = 3 The average silhouette_score is : 0.07527284091854468
For k = 4 The average silhouette_score is : 0.026138473834062554
For k = 5 The average silhouette_score is : 0.07940181037017646
For k = 6 The average silhouette_score is : 0.07563898295740333
For k = 7 The average silhouette_score is : 0.07142925280881725
For k = 8 The average silhouette_score is : 0.06811628838159771
For k = 9 The average silhouette_score is : 0.08624507344493955
For k = 10 The average silhouette_score is : 0.04767116227490887
```

Figure 5: Silhouette Coefficient Score

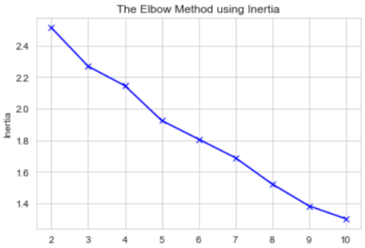


Figure 6: Elbow Method

The Elbow Method is more of a decision rule, while the Silhouette is a metric used for validation while clustering. Thus, it can be used in combination with the Elbow Method. Therefore, the Elbow Method and the Silhouette Method are not alternatives to each other for finding the optimal K. Thus they are tools to be used together for a more confident decision. For both the Athens neighborhoods and the Attica cities, 5 clusters were chosen.

Results

The Clustering analysis produced the following results:

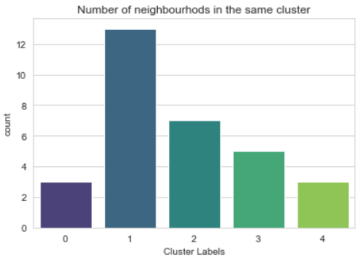


Figure 7: Neighbourhoods in the same cluster

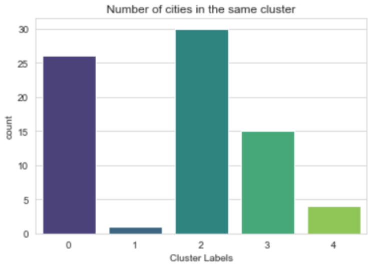


Figure 8: Cities in the same cluster

In the above figures, we can see the number of neighborhoods and cities that belong to the same clusters. Bellow, the clustering labels are showcased on the maps.

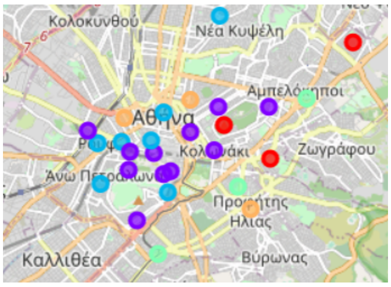


Figure 9: Clustered Athens

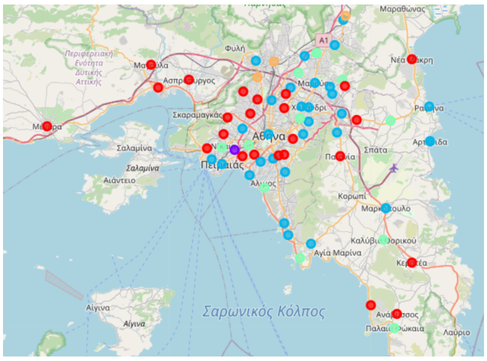


Figure 10: Clustered Attica

People that live in Athens or in Attica would be able to agree that these clusters are pretty reasonable regarding the similarity of the neighborhoods and cities. Detailed results of the analysis are hosted on Github [9].

Discussions

The data gathered to make a concrete analysis are not enough, because premium access to foursquare API, is required to gather features such as ratings, hours, tips, description, and many more [10] for each venue. These features would enrich the dataset used for clustering and assign sufficient labels. Additionally, graph theory analysis could be used by assigning each neighborhood as a node in a graph containing all the venues as features and the edges would define the location of the nodes.

Conclusion

To conclude this project of the battle of the neighbors, tourists can now get an initial perspective of the similarity between the neighbourhoods of Athens as well as the cities of Attica. Therefore, they would be able to see before their trip which places they want to visit.

References

[1] John A Hartigan and Manchek A Wong. “AK-means clustering algorithm”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28.1 (1979), pp. 100–108.

[2] Sukavanan Nanjundan et al. “Identifying the number of clusters for K-Means: A hypersphere density based approach”. In: *arXiv preprint arXiv:1912.00643* (2019).

[9] https://github.com/kon91/Applied-Data-Science-Capstone-/tree/main/Attica%20Clustering.

[10] https://developer.foursquare.com/docs/places-database/details/