The Battle of Neighborhoods

Analysis and Clustering of Athens Regions

Coursera Capstone Project

Vasileios Kalyvas

1. Introduction/Business Problem

In this project, we will try to determine the most famous venues for **Athens, Greece**.

Athens, the capital and largest city of Greece, is one of the world's oldest cities with its recorded history spanning over 3,000 years. It was the center for arts, learning, philosophy and the birthplace of democracy, having a great cultural and political impact on the European continent. Athens is home to two UNESCO World Heritage Sites, the Acropolis of Athens and the Daphni Monastery. Because of its ancient monuments, works of art, landmarks and museums, Athens remains nowadays one of the most famous and attractive places for visitors from all over the world. As such, it would be interesting to explore the variety of venues around Athens and gain insights on the city's most popular places. That would be very helpful for both tourists and travel officers to better plan their trips and provide more personalized offers, according to anyone's needs. Furthermore, this analysis would be beneficial for business owners to better understand the different regions and select the appropriate place to open or relocate their business.

2. Data

The analysis will be performed with the help of Wikipedia, some Python libraries and Foursquare, the famous location data platform. Starting with the location, the regions of Athens will be acquired from Wikipedia and passed to *"geopy"* in order to return their coordinates (latitudes and longitudes). *"geopy"* does not provide information about every region in Athens but it is still very sufficient.

After that, we will get the venues for every region with the Foursquare API. A developer account has already been set, in order to acquire venue data for Athens. Then, we will perform the clustering analysis with *"sklearn"* library to find similar regions in terms of their venues and, finally, create a map of the regions and their corresponding clusters.

3. Methodology

As previously stated, *"geopy"* does not have coordinates for all regions of Athens, so we will keep only the regions for which we have data to work with. We address this issue with *"try-except"* and we finally have 52 regions remaining (out of the initial 77). Still, they are sufficient for our analysis.

We then construct a Pandas Dataframe with the Regions, Latitudes and Longitudes and plot them on a map with *“folium”* to get a better understanding of their locations.

Next is the venues’ data acquisition from Foursquare. A developer account had already been created and credentials were provided. With the API requests, we are able to get all venues for every region, along with their coordinates and category type. For simplicity reasons, we decide on a few steps:

1. We will select the 20 most popular types of venues, so as to perform a representative analysis.
2. We are not interested in Gyms (not a tourist attraction).
3. we will combine "cafe" and "coffee shops" into one category.
4. We will combine 'Meze Restaurant" and "Greek Restaurant" into one category (as "meze" refers to traditional Greek food).

As a result, there are 17 venue types remaining.

We then perform one hot encoding in order to transform the venue category of every row into features of 0s and 1s. This is a necessary part when having categorical features.

We have multiple rows per region (equal to the number of venues in this region). So, for clustering, we have to group the data so as to have one row per region and compare their venues to find how similar they are.

Moving on to Clustering Analysis, we first have to scale the data, because clustering is based on distances. That means that very popular venues might "dominate" the less popular ones and the results would not be realistic.

Clustering is an Unsupervised Learning Algorithm and that means it is not based on labeled data and we also do not know the appropriate number of clusters that would give the best performance.

So, we will try to find the best number of clusters based on inertia (a measure of how internally coherent clusters are) and the *"elbow method".*

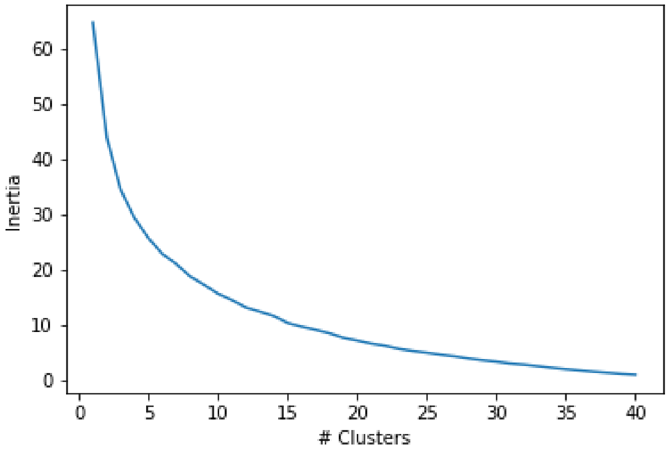

Figure 1: Elbow Method

According to this method, and the diagram constructed, we can set the appropriate number of clusters between 5 and 10 because, as the number of clusters increases, there is no significant decrease of inertia (i.e. increase of performance), so there is no point working with more clusters. However, for simplicity reasons, we set the number of clusters to 4.

The dataset has many features (i.e. every region has multiple venues) and it is not easy to plot the data. For this reason, we introduce PCA (Principal Component Analysis), another Unsupervised Leaning Algorithm, which decreases the number of features in such a way that new features are constructed in order to keep as much explainability as possible. In this way, we can also view the data in two dimensions:

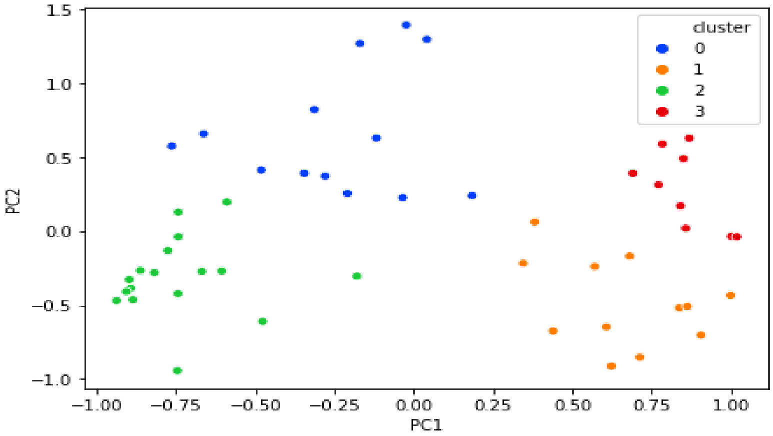

Figure 2: PCA (n=2) plot

We see that 4 clusters are formed (as specified) and they are a bit distinct from each other.

In total, the two new features explain about 61% of the initial variance in the data. It is fair enough, taking into consideration that we have 17 venues for every region.

Another metric for clustering is the Silhouette Score, which is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It ranges from -1 (worst) to 1 (best), while a score of 0 indicates overlapping clusters. In our situation, this score is nearly 27%.

For the final part of our analysis, we have 1) Parallel Plots and 2) Folium Map.

The Parallel Plots show the similarities (in terms of features) between regions of the same cluster, so we should add the information about the cluster of every region to the previous dataset as computed and then scale the data for the parallel plots to work properly (results will be discussed later). We, finally, plot regions in the map, clustered according to their venues with the help of *"folium"*.

4. Results

Our Clustering Analysis results in the following Parallel Plots:

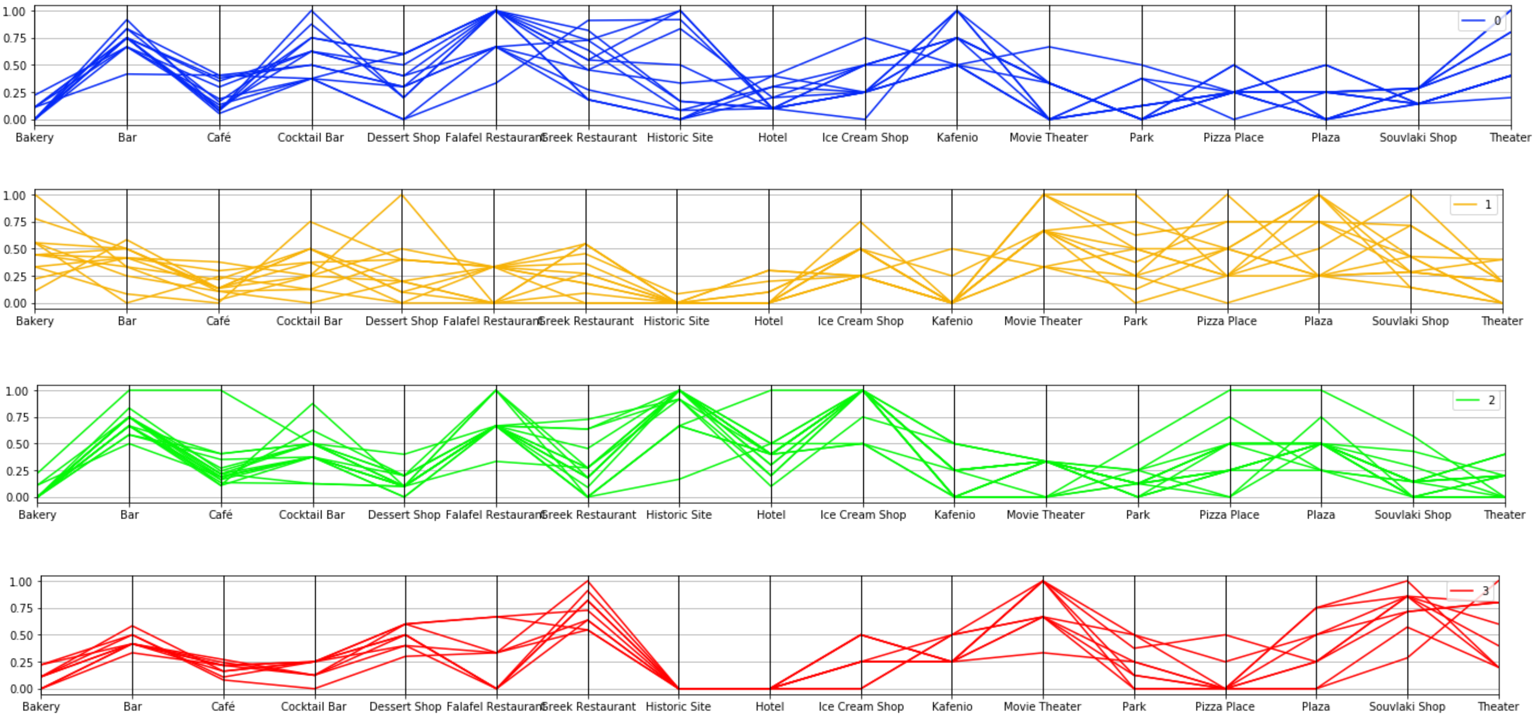

Figure 3: Parallel Plots

5. Discussion

These plots might seem very messy, but they can provide some interesting insights.
Every plot represents a cluster (see label at the top-right) and shows how many venues of all types each cluster has, according to the density of the lines.

- **Cluster 0 (blue)** consists mainly of Bars, Falafel Restaurants, Kafenios (traditional Greek cafeterias), Historic places and some Theaters. This seems like regions in the center of Athens, with their historic places and monuments.
- **Cluster 1 (orange)** seems balanced overall, having a bit of everything but no Historic sites or Hotels. It seems to refer to more suburban areas rather than the city center.
- **Cluster 2 (green)** seems similar to cluster 0 (blue) but with more Historic sites, Ice Cream shops, Plazas, Pizza and Souvlaki shops but, on the contrary, fewer Kafenios, Theaters and Movie Theaters. This also seems like being at the center of Athens.
- **Cluster 3 (red)** mainly consists of Restaurants, Food places and Theaters, probably outside the center. It seems like cluster 1 (orange) but with less options.

6. Conclusion

In this report, we analyzed the regions of Athens and their similarities in terms of their most popular venues, according to Foursquare. We grouped them into clusters, viewed them on a map and gained meaningful insights. These insights could be helpful in personalizing tourist offers or deciding on the correct business at the most appropriate place. As a next step, we could further try different clustering algorithms and test their performance against kmeans.