APPLIED DATA SCIENCE CAPSTONE

https://www.coursera.org/learn/applied-data-science-capstone

ASSIGNMENT: THE BATTLE OF NEIGHBORHOODS

"STARTING A NEW CINEMA IN ATHENS, GREECE

Dimitra Dionysiou

1. Introduction

In this project the **hypothetical scenario of clients who are interested in starting an indoor cinema in Athens, Greece**, is considered. The presented analysis according to the assignment instructions aims mainly at demonstrating the use of Foursquare location data to solve business problems. Obviously, a formal analysis of such an issue would be very complex and multi-parametric and is out of the scope of the presented work. Assuming, therefore, that financial and demographic aspects are not an issue, the presented analysis identifies candidate locations ("catchment areas") for starting a new cinema based on the **requirements** outlined below:

- **Scale and number of competitor cinemas**: The cinema should be located in an area where other cinema theatres exist as well, so that availability of audience can be generally assumed. More specifically, **a candidate location is defined as a cyclic area, with a radius of 250m, centered on an already existing cinema**. Whereas at least one already existing cinema in the area is a prerequisite, too many cinemas in the vicinity imply that it could be much more difficult for the new business to be competitive and draw audience. Therefore, large multiplexes (with more than 4 screens) with established presence in Athens are excluded from the analysis. Candidate locations which include more than 5 other cinemas in a 500m distance are also excluded.

- **Leisure facilities near the planned cinema**: Since visits to cinemas are usually accompanied by other leisure activities (eating, drinking, shopping etc.), a candidate location should offer many such opportunities in its vicinity. The existence of large shopping malls will be considered as an extra plus.

- **Transport options**: A candidate location should be well served by public transport. Availability of nearby parking areas will be also considered. This aspect of the candidate locations will be evaluated based on the total number of available metro stations, bus stops and parking places in the area, since all three options will be assumed equally important.

In summary, the question which the following analysis seeks to answer can be formulated as follows: **"In the vicinity (250m cyclic area) of which existing cinema should the new cinema be located based on the specific requirements outlined above?"**

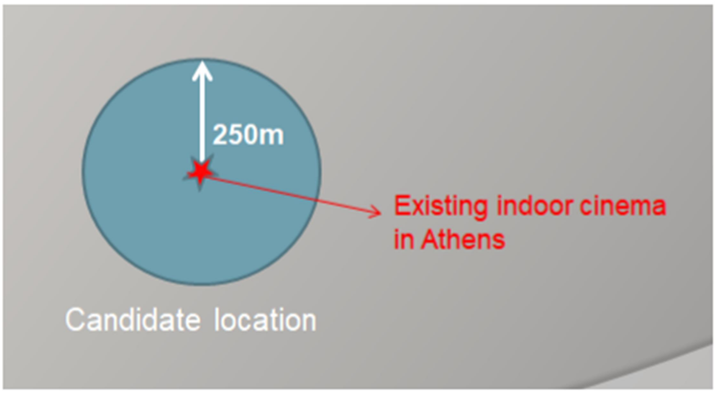

Figure 1. Definition of candidate locations for starting a new cinema

2. Data

The data sources that have been used in this analysis are the following:

- A list of the existing indoor cinemas in Athens, along with their addresses is retrieved from a well-known Athens city guide:
  https://www.athinorama.gr/cinema/guide.aspx?show=1&seltab=1&sec=2
  Each cinema of the retrieved list will serve as the center of candidate location for starting a new cinema.

- Geographical coordinate data for each cinema is retrieved from OpenStreetMap: https://www.openstreetmap.org using each cinema's address (Nominatim search engine, geopy Python package).

- Foursquare location data are used to decide which cinemas from the initially derived cinema list will be considered in the subsequent analysis, by excluding cases where more than 5 other cinemas exist in a 500m distance, according to the requirements defined in the Introduction section. The corresponding venue category id ("movie theater") is used for these API calls.

- Foursquare location data are also used to define the characteristics of each candidate location (centered on an already existing cinema) in terms of the features of interest: restaurants, nightlife spots, shopping malls and transport options (metro stations, bus stops, private parking places), according to the requirements defined in the Introduction section. The corresponding venue category ids are used to perform the necessary Foursquare API calls, by defining a radius of interest equal to 250m.

3. Methodology

3.1. Web page scraping and data preprocessing

The above specified web page was scraped with the use of the Beautiful Soup package. The extracted information included: name and address of each existing indoor cinema in Athens. 58 cinemas were retrieved.

Large multiplexes with more than 4 screens have been excluded from the subsequent analysis. As a result 44 candidate locations have been specified at this stage.

3.2. Geographic coordinate retrieval

The geographic coordinate of each cinema is retrieved with the help of the Nominatim search engine for OpenStreetMap data and the geopy library of Python. A little preprocessing was required at this stage since some of the extracted cinema addresses (in greek characters) had not been initially recognized by Nominatim and they had to be reformatted. The following figure displays the first 10 rows of the resulting dataframe.

	cinema_name	cinema_address	latitude	longitude
0	Ααβόρα	Ααβόρα Ιπποκράτους 180 Νεάπολη	37.988000	23.746283
1	Άστορ	Άστορ Σταδίου 28	37.979521	23.732192
2	Άστυ	Asti Korai 4	37.979778	23.732302
3	Έλλη	Έλλη Ακαδημίας 64	37.982776	23.733563
4	Έμπασσυ Novacinema Odeon	Πατριάρχου Ιωακείμ 5 κολωνάκι	37.977705	23.742040
5	Ιντεάλ	Ideal, 46, Ελευθερίου Βενιζέλου, Exarcheia	37.982464	23.731459
6	Odeon Όπερα	akadimias 57	37.982289	23.733577
7	Ταινιοθήκη της Ελλάδος	48, iera odos, gkazi	37.980923	23.712603
8	Αθήναιον	αθήναιον, 124, Βασιλίσσης Σοφίας	37.985962	23.761325
9	Ανδόρα	Σεβαστουπόλεως 117	37.995862	23.769469

Figure 2. The first 10 rows of the dataframe containing the name and address of each cinema and its geographical coordinate

3.3. Retrieval of Foursquare location data

Based on the previously identified longitude and latitude data, the necessary calls to the Foursquare API have been made.

Firstly, candidate locations centered on existing cinemas having more than 5 other cinemas in a cyclic area of 500m were excluded. For this purpose, Foursquare API calls have been made using the venue category id "movie theater" (id = 4bf58dd8d48988d17f941735) and the results were processed to get for each candidate location the total number of cinemas. After this step of the analysis, the list of candidate locations included 39 cinemas.

Subsequently, for each of the 39 cinemas new Foursquare calls were performed in the form of searches for specific types of venues around each cinema, in cyclic areas with a radius of 250m, which can be considered as an acceptable walking distance:

- Food: Foursquare id = '4d4b7105d754a06374d81259'
- Night spot: Foursquare id = '4d4b7105d754a06376d81259'
- Shopping mall: Foursquare id = '4bf58dd8d48988d1fd941735'
- Metro station: Foursquare id = '4bf58dd8d48988d1fd931735'
- Bus stop: Foursquare id = '52f2ab2ebcbc57f1066b8b4f'
- Parking: Foursquare id = '4c38df4de52ce0d596b336e1'

The retrieved data were processed to get the number of available venues of each category around each cinema (i.e. total number of restaurants, total number of night spots etc.). The data about metro stations, bus stops and parkings for each cinema area were added together to constitute a new total number representing the total transportation options in the vicinity of a cinema. It should also be noted that the "food" venue category of Foursquare returns all types of restaurants as well as similar venues (such as bakeries, cafeterias etc.).

The first 5 rows of our dataframe after the retrieval of location data are the following:

	cinema_name	cinema_address	latitude	longitude	restaurants	night_spots	shopping_malls	metro_stations	bus_stops	parkings
0	Αaβόρα	Αaβόρα Ιπποκράτους 180 Νεάπολη	37.988000	23.746283	39	9	0	0	4	0
1	Έμπασσυ Novacinema Odeon	Πατριάρχου Ιωακείμ 5 κολωνάκι	37.977705	23.742040	50	50	1	0	2	4
2	Ταινιοθήκη της Ελλάδος	48, iera odos, gkazi	37.980923	23.712603	50	50	0	1	0	1
3	Αθήναιον	αθήναιον, 124, Βασιλίσσης Σοφίας	37.985962	23.761325	50	12	0	0	3	9
4	Ανδόρα	Σεβαστουπόλεως 117	37.995862	23.769469	49	6	1	0	2	3

Figure 3. The dataframe with the cinema info and the retrieved data from Foursquare

And after adding the numbers of metro stations, bus stops and parkings into a variable named “transport”:

	cinema_name	cinema_address	latitude	longitude	restaurants	night_spots	shopping_malls	transport
0	Αàβόρα	Αàβόρα Ιπποκράτους 180 Νεάπολη	37.988000	23.746283	39	9	0	4
1	Έμπασσυ Novacinema Odeon	Πατριάρχου Ιωακείμ 5 κολωνάκι	37.977705	23.742040	50	50	1	6
2	Ταινιοθήκη της Ελλάδος	48, iera odos, gkazi	37.980923	23.712603	50	50	0	2
3	Αθήναιον	αθήναιον, 124, Βασιλίσσης Σοφίας	37.985962	23.761325	50	12	0	12
4	Ανδόρα	Σεβαστουπόλεως 117	37.995862	23.769469	49	6	1	5

Figure 4. A modified version of the dataframe in which metro station, bus stop and parking data for each cinema have been combined into a new variable named “transport” according to the initially defined requirements for the considered problem.

3.4. K-means Clustering

Having gathered all the necessary data, k-means clustering has been used in order to partition cinemas in Athens in 3 different clusters. The number of clusters has been chosen by trial and error, as well as taking into account that the total number of cinemas (39) was relatively small. Each cluster would contain cinemas similar in terms of our studied features: restaurants, night spots, shopping malls and transportation options, and the selection of the best cluster (i.e the list of cinemas that are located in areas with the desired features) would be made possible. The mean values of the studied features are expected to be larger in the best cluster.

The data have been normalized using sklearn’s StandardScaler before proceeding to clustering. At this stage of the analysis **a recommendation about the best possible cinema locations would be made at the cluster level** (i.e. the analysis would recommend starting a new cinema in the vicinity of any of the cinemas belonging to the best cluster).

3.5. Further basic statistical processing for the cinemas of the best cluster

Having gathered the results of k-means clustering and thereby specified the cinemas belonging to the best cluster, the window of possible candidate locations can be narrowed down by some further basic statistical processing. The aim at this stage is to provide **a recommendation at the single cinema level (single area level),** by formulating in a more intuitive way the characteristics of the candidate cinemas (candidate areas).

This will involve the definition of several “scores”, wherein each score will take values from 0 to 1 (1 is the best possible score). Each score will emphasize a different aspect of the “performance” of a candidate location. For this purpose, firstly each value for restaurants, night-spots, shopping malls and transportation options is divided by the maximum observed value of the respective category in the cluster, which transforms the data in the 0-1 range. For example, in the case of a cinema with 29 night spots in its vicinity we will get:

Norm_night_spots = Initial_night_spots/Max_night_spots = 29/49 = 0.59

Subsequently, several scores can be defined as follows:

- **Score Neutral**:

  Restaurants, nightlife spots and transportation features are considered equally important.

$$Score\_Neutral_{cin\ i} = \frac{Norm\_restaur_{cin\ i} + Norm\_night\_sp_{cin\ i} + Norm\_malls_{cin\ i} + Norm\_transp_{cin\ i}}{4},$$

  for each cinema i in the best cluster,

  where $Norm\_restaur_{cin\ I}$, $Norm\_night\_sp_{cin\ I}$, $Norm\_malls_{cin\ I}$, $Norm\_transp_{cin\ I}$ are the normalized values for the restaurants, night spots, shopping malls and transportation options, respectively.

- **Score Leisure Activities**:

  In this case, the existence of multiple options for leisure activities is considered of greater importance than the existence of transportation options (the magnitude of weights can be adjusted , in the following equation they are set to 0.3, 0.3, 0.3, 0.1 for restaurants, nightlife spots, shopping malls and transportation options, respectively)

$$Score\_Leisure_{cin\ i} = 0.3 * Norm\_restaur_{cin\ i} + 0.3 * Norm\_night\_sp_{cin\ i} + 0.3 * Norm_{malls_{cin\ i}} + 0.1 * Norm\_transp_{cin\ i}$$

  for each cinema i in the best cluster.

  where $Norm\_restaur_{cin\ I}$, $Norm\_night\_sp_{cin\ I}$, $Norm\_malls_{cin\ I}$, $Norm\_transp_{cin\ I}$ are the normalized values for the restaurants, night spots, shopping malls and transportation options, respectively.

- **Score Transportation**:

  In this case, the existence of transportation options (which include metro stations, bust stops and parking areas) is considered of greater importance than the existence of leisure activity options (the magnitude of weights can be adjusted, in the following equation they are set to 0.2, 0.2, 0.2, 0.4 for restaurants, night spots, shopping malls and transportation options, respectively)

$$Score\_Transport_{cin\ i} =$$

$$0.2 * Norm\_restaur_{cin\ i} + 0.2 * Norm\_night\_sp_{cin\ i} + 0.2 * Norm_{malls\ cin\ i} + 0.4 * Norm\_transp_{cin\ i}$$

for each cinema i in the best cluster.

where $Norm\_restaur_{cin\ I}$, $Norm\_night\_sp_{cin\ I}$, $Norm\_malls_{cin\ I}$, $Norm\_transp_{cin\ I}$ are the normalized values for the restaurants, night spots, shopping malls and transportation options, respectively.

With the help of the previously defined scores different aspects of the "performance" of each cinema of the best cluster can be quantified and in close contact with the client(s) a **recommendation at the level of a single cinema be reached.**

4. Results

4.1. K-means clustering results

The three resulting cinema clusters identified by k-means clustering can be seen superimposed on the Athens map in the following figure:

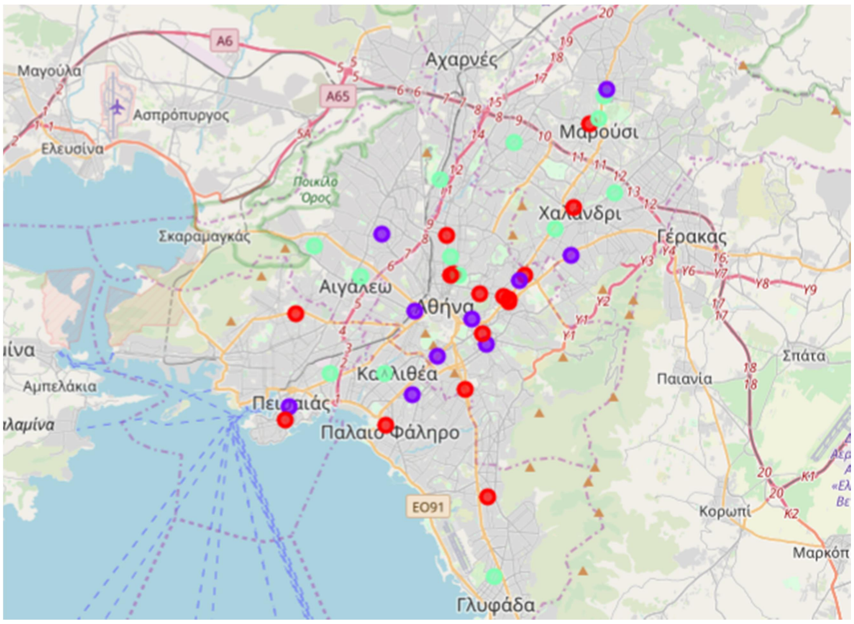

Figure 5. The three cinema clusters identified through the use of k-means clustering. Red: Cluster 0, Purple: Cluster 1, Light Green: Cluster 2

Figures 6-8 help us inspect the resulting clusters and Figure 9 depicts the mean values of the studied features for each cluster.

	Cluster Labels	cinema_name	restaurants	night_spots	shopping_malls	transport
0	0	Ααβόρα	39	9	0	4
3	0	Αθήναιον	50	12	0	12
4	0	Ανδόρα	49	6	1	5
5	0	Γαλαξίας	50	8	0	13
7	0	Νιρβάνα 1 & 2 Cinemax	50	21	0	9
9	0	Athena	43	13	0	5
12	0	Διάνα	50	21	0	5
17	0	Αλεξάνδρα Europa Cinemas Digital	38	11	1	4
19	0	Όσκαρ Digital	37	5	0	4
21	0	Τριανόν	40	10	0	5
23	0	Πτι-Παλαι	47	26	0	4
24	0	Ατλαντίς Classic Cinemas	50	12	0	6
25	0	Σοφία HD DIGITAL	50	22	0	1
28	0	Cinerama Digital cinema	46	4	0	5
37	0	Ζέα Digital Cinema	48	7	0	2
38	0	Cine Παράδεισος 2+1 (Δημ. Κιν/φος)	40	8	0	2

Figure 6. Cluster 0: the first cinema cluster retrieved by k-means clustering

	Cluster Labels	cinema_name	restaurants	night_spots	shopping_malls	transport
1	1	Έμπασσυ Novacinema Odeon	50	50	1	6
2	1	Ταινιοθήκη της Ελλάδος	50	50	0	2
6	1	Δαναός	50	45	1	10
11	1	Σινέ Χολαργός	41	11	2	5
14	1	Κηφισιά Cinemax 3	50	24	3	8
22	1	Πάλας	50	34	1	1
27	1	Μικρόκοσμος	49	35	0	6
29	1	Σπόρτιγκ Digital Cinema	49	27	1	2
32	1	Φοίβος Digital Cinema	50	43	0	1
36	1	Δημ. Κιν. Σινεάκ	50	33	3	9

Figure 7. Cluster 1: the second cinema cluster retrieved by k-means clustering

	Cluster Labels	cinema_name	restaurants	night_spots	shopping_malls	transport
8	2	Αβάνα	16	3	0	1
10	2	Αίγλη 3D Digital	15	5	0	1
13	2	Κηφισιά Cinemax	17	5	1	2
15	2	Novacinema Odeon Μαρούσι	5	4	0	2
16	2	Τρία Αστέρια 3D Digital	29	4	0	3
18	2	Ίλιον Cinema & Stage	25	8	0	0
20	2	Studio new star art cinema	21	4	0	4
26	2	Αλεξάνδρα Digital Cinema	23	5	0	0
30	2	Άνοιξη Digital Cinema (δημ. Κιν/φος) 2+1	13	2	0	0
31	2	Λάμπρος Κωνσταντάρας - Ρένα Βλαχοπούλου	14	1	0	1
33	2	Μαρία Έλενα-Όναρ Digital Cinema (Δημ. Κιν/φος)	5	1	0	0
34	2	Novacinema Odeon Γλυφάδα	15	1	0	0
35	2	Δημ. Κιν. Όνειρο Ρέντη	16	1	0	0

Figure 8. Cluster 2: the third cinema cluster retrieved by k-means clustering

Cluster	Mean no of restaur	Mean no of night_spots	Mean no of shopp_malls	Mean no of transport_options
Cluster 0	45.44	12.19	0.12	5.38
Cluster 1	48.90	35.20	1.20	5.00
Cluster 2	16.46	3.38	0.08	1.08

Figure 9. Mean values of the studied features in each cluster

4.2. Cinema scores for the best cluster

The following is a modified version of Cluster 1 data (best cluster: see Discussion section) with the values of interest transformed in the range 0-1, as described in Section 3.5 of Methodology.

	Cluster Labels	cinema_name	restaurants_normal	night_spots_normal	shopping_malls_normal	transport_normal
1	1	Έμπασσυ Novacinema Odeon	1.00	1.00	0.33	0.6
2	1	Ταινιοθήκη της Ελλάδος	1.00	1.00	0.00	0.2
6	1	Δαναός	1.00	0.90	0.33	1.0
11	1	Σινέ Χολαργός	0.82	0.22	0.67	0.5
14	1	Κηφισιά Cinemax 3	1.00	0.48	1.00	0.8
22	1	Πάλας	1.00	0.68	0.33	0.1
27	1	Μικρόκοσμος	0.98	0.70	0.00	0.6
29	1	Σπόρτιγκ Digital Cinema	0.98	0.54	0.33	0.2
32	1	Φοίβος Digital Cinema	1.00	0.86	0.00	0.1
36	1	Δημ. Κιν. Σινεάκ	1.00	0.66	1.00	0.9

Figure 10. Cluster 1: normalized data

In the following figure the three scores defined in section 3.5 of Methodology have been computed as well.

	Cluster Labels	cinema_name	score_neutral	score_leisure	score_transport
1	1	Έμπασσυ Novacinema Odeon	0.73	0.76	0.71
2	1	Ταινιοθήκη της Ελλάδος	0.55	0.62	0.48
6	1	Δαναός	0.81	0.77	0.85
11	1	Σινέ Χολαργός	0.55	0.56	0.54
14	1	Κηφισιά Cinemax 3	0.82	0.82	0.82
22	1	Πάλας	0.53	0.61	0.44
27	1	Μικρόκοσμος	0.57	0.56	0.58
29	1	Σπόρτιγκ Digital Cinema	0.51	0.57	0.45
32	1	Φοίβος Digital Cinema	0.49	0.57	0.41
36	1	Δημ. Κιν. Σινεάκ	0.89	0.89	0.89

Figure 11. Cluster 1: computed scores for each cinema

5. Discussion

Based on the inspection of the clusters it can be derived that cinemas of clusters 1 are generally preferable, because they are characterized by comparatively larger values of restaurants, night spots and shopping malls, while transportation options are also good. This can be easily seen in Figure 9, where the mean values of the studied features for each cluster are depicted. **So based on the k-means cluster analysis we would recommend starting a new cinema in the vicinity (250m-radius cyclic areas) of any of the cinemas included in cluster 1 (recommendation at the cluster level).**

If our client requests a **recommendation at the cinema (location) level,** then we can use the results of the basic statistical processing on cluster 1. It can be easily seen in Figure 11, that the best candidate location will be the area centered on **the cinema with index 36 in our dataframe** (greek name: 'Δημοτικός Κινηματογράφος Σινεάκ', latitude: 37.941869, longitude: 23.647198) which has the maximum value for all three computed scores defined in section 3.5 of the methodology. The location of this specific cinema can be seen on the following map:

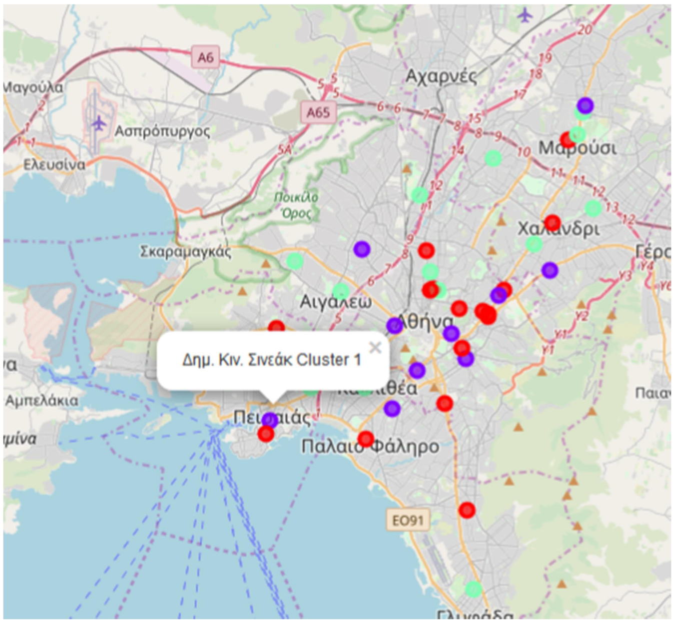

Figure 12. Folium Map displaying the location the cinema in the vicinity of which starting a new cinema would be recommended, if a recommendation at the single area level was requested.

6. Summary - Conclusions

The presented analysis dealt with the problem of selecting the best candidate locations for starting a new cinema in Athens, Greece, based on a number of specific predefined requirements which were related to a) the scale and number of competitor cinemas in an area, b) the existence of nearby leisure facilities and c) the availability of transportation options. The exact requirements have been outlined in the Introduction section. The problem to be solved was formulated as the following question: "In the vicinity (250m cyclic area) of which existing cinema should the new cinema be located based on the specific requirements?".

Foursquare location data were the main data source for solving the problem, after having retrieved a list of the existing cinemas in Athens through web page scraping. After all necessary data prepropressing, k-means clustering was used in order to partition cinemas in Athens into three different clusters and select the best cluster based on the predefined features of interest. A recommendation at the cluster level was made at this stage of the analysis. Additionally, basic statistical processing on the data of the best cluster permitted a recommendation at the single cinema (single candidate location) level.