APPLIED DATA SCIENCE CAPSTONE:

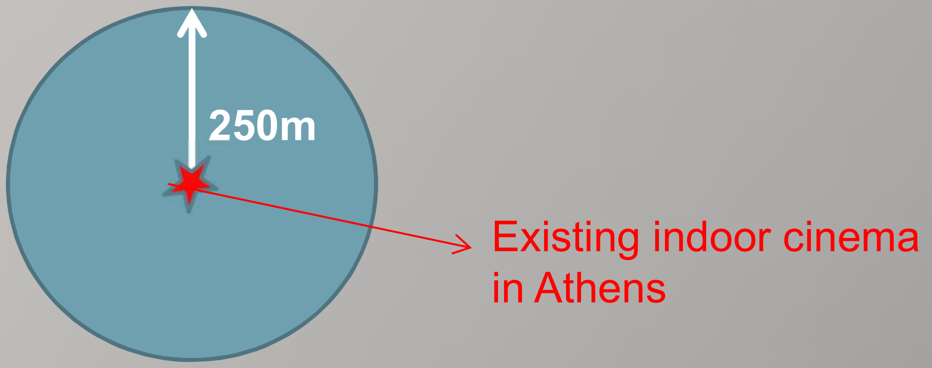
THE BATTLE OF NEIGHBORHOODS

By **DIMITRA DIONYSIOU**

STARTING A NEW CINEMA IN ATHENS, GREECE

Problem Description (I)

- Consider the hypothetical scenario of clients interested in starting an indoor cinema in Athens, Greece
- Candidate location: a cyclic area, with a radius of 250m, centered on an already existing indoor cinema



Candidate location

Problem Description (II)

❖ Requirements:

Scale and number of competitor cinemas: large multiplexes with more than 4 screens should be excluded as candidate locations. Candidate locations for which more than 5 other cinemas exist in a 500m distance should be also excluded.

Leisure facilities near the planned cinema: calculate the number of available leisure facilities in a candidate location in terms of restaurants, nightlife spots and shopping malls.

Transportation options: calculate the number of available transportation options in a candidate location in terms of metro stations, bus stops and private parking areas.

Question to answer:

➡ "In the vicinity (250m-radius cyclic area) of which existing cinema should the new cinema be located based on the specific requirements outlined above?"

Data sources

- List of indoor cinemas in Athens (names and addresses): scraped from a well-known Athens city guide: https://www.athinorama.gr/cinema/guide.aspx?show=1&seltab=1&sec=2
- Geographical coordinate data for each cinema: retrieved from OpenStreetMap: https://www.openstreetmap.org
- Foursquare location data:
    - Number of cinemas for each candidate area
    - Number of leisure facilities (restaurants, nightlife spots, shopping malls) for each candidate area
    - Number of transportation options (metro stations, bus stops, parking areas) for each candidate area

Methodology (I)

- Web page scraping: Beautiful Soup package

- Geographical coordinate retrieval: Nominatim search engine for OpenStreetMap, geopy library of Python

- Foursquare location data:

  - API calls for each candidate location: search venues with category id ‘movie theater’, radius 500m ⇨ processing to get the total number of cinemas in the area

  - API calls for each candidate location: search venues with category ids: ‘food’, ‘nightlife spot’, ’shopping mall’, ‘metro station’, ‘bus stop’, ‘parking’ . Radius 250m

    ⇨ processing to get the corresponding total numbers of venues

    ⇨ processing to create a new feature named “transport” by adding the retrieved numbers of metro stations, bus stops, parking places

Methodology (II)

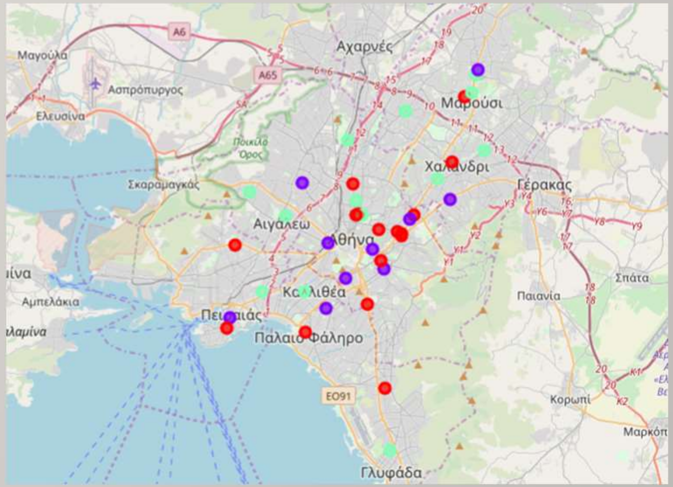
- K-means Clustering: partition the candidate locations (centered on each existing cinema) into 3 different clusters
- Select the best cluster: based on the mean values of the studied features, which should be larger in the best cluster

⇨ PROVIDE RECOMMENDATION AT THE CLUSTER LEVEL

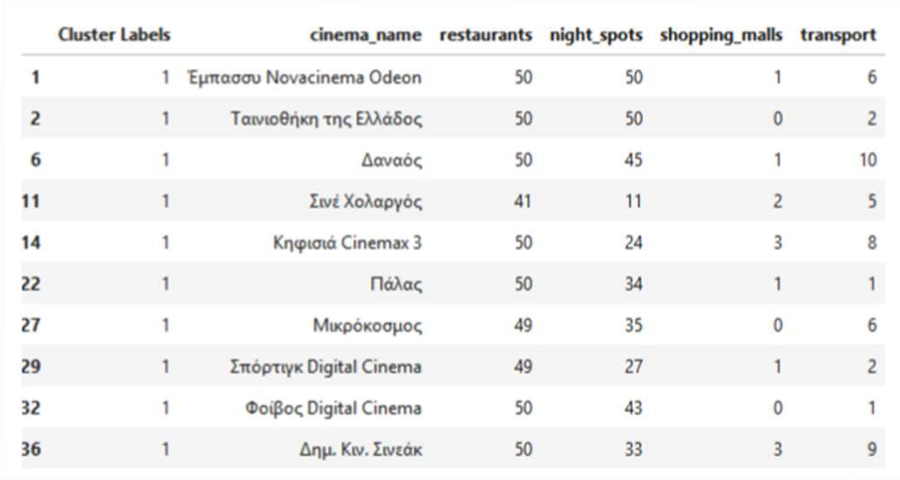
- Basic statistical processing on the best cluster:
  - Normalize the values of the features in the best cluster
  - Compute scores for each cinema of the best cluster:
    - Score neutral: Restaurants, nightlife spots and transportation features are considered equally important (by assigning appropriate weights)
    - Score Leisure: multiple options for leisure activities is considered of greater importance than the existence of transportation options (by assigning appropriate weights)
    - Score Transportation: transportation options are considered of greater importance than leisure activity options (by assigning appropriate weights)

⇨ PROVIDE RECOMMENDATION AT THE SINGLE CINEMA LEVEL

Results (I)

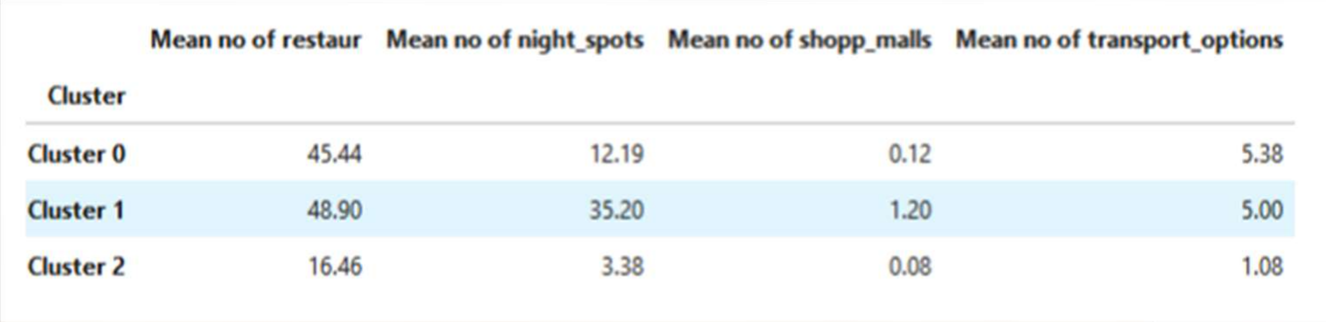


The three cinema clusters identified through k-means clustering. Red: Cluster 0, Purple: Cluster 1, Light Green: Cluster 2



	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

The best cluster: Cluster 1

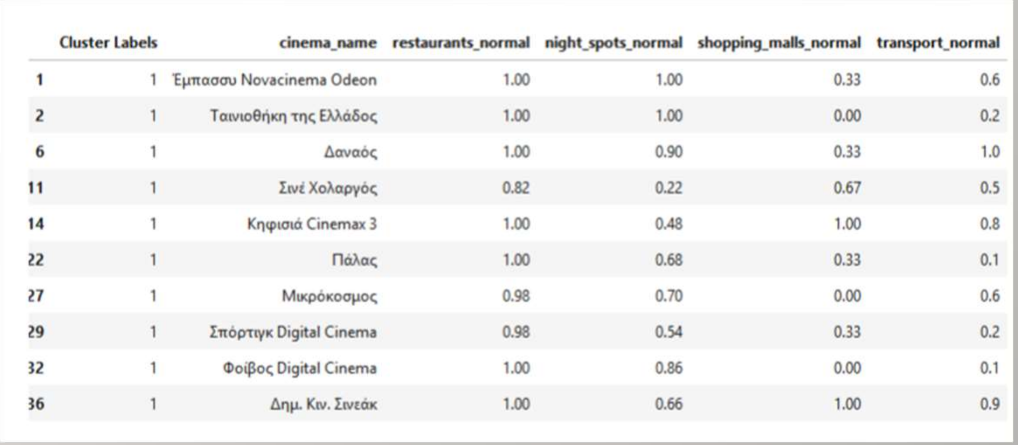


	Mean no of restaur	Mean no of night_spots	Mean no of shopp_malls	Mean no of transport_options
Cluster				
Cluster 0	45.44	12.19	0.12	5.38
Cluster 1	48.90	35.20	1.20	5.00
Cluster 2	16.46	3.38	0.08	1.08

Mean values of the studied features in each cluster

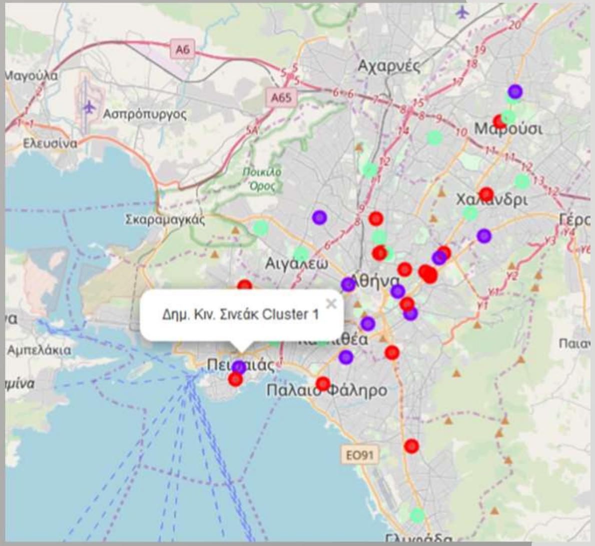
RECOMMENDATION AT THE CLUSTER LEVEL: Start a new cinema in any of the candidate locations of Cluster 1

Results (II)

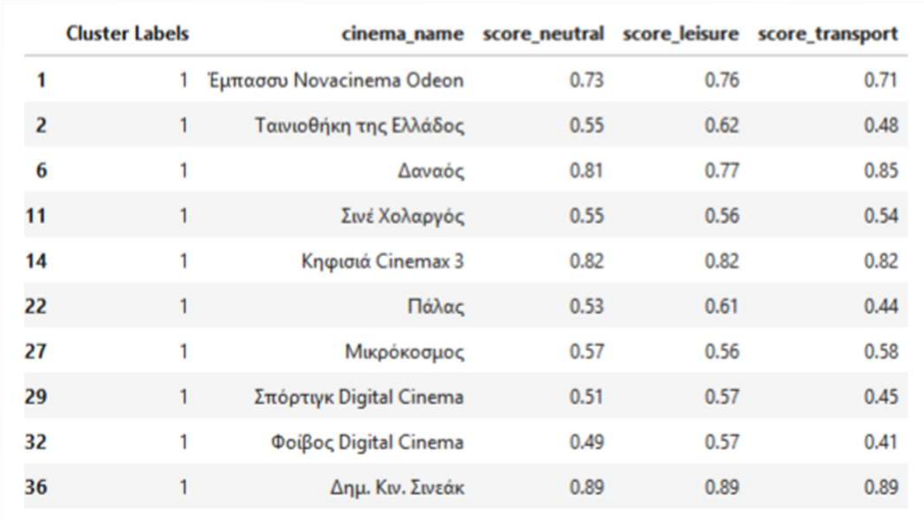


	Cluster Labels	cinema_name	restaurants_normal	night_spots_normal	shopping_malls_normal	transport_normal
1	1	Έμπασσυ Novacinema Odeon	1.00	1.00	0.33	0.6
2	1	Ταινιοθήκη της Ελλάδος	1.00	1.00	0.00	0.2
6	1	Δαναός	1.00	0.90	0.33	1.0
11	1	Σινέ Χολαργός	0.82	0.22	0.67	0.5
14	1	Κηφισιά Cinemax 3	1.00	0.48	1.00	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

Normalized feature values of Cluster 1



Location of the selected cinema



	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

Computed Neutral, Leisure and Transport scores for the cinemas (candidate locations) of Cluster 1

RECOMMENDATION AT THE SINGLE CINEMA LEVEL: Start a new cinema within the 250m-radius cyclic area centered on the cinema with index 36

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
  - the scale and number of competitor cinemas in an area,
  - the existence of nearby leisure facilities and
  - the availability of transportation options.
- 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.