IBM Data Science Capstone Project

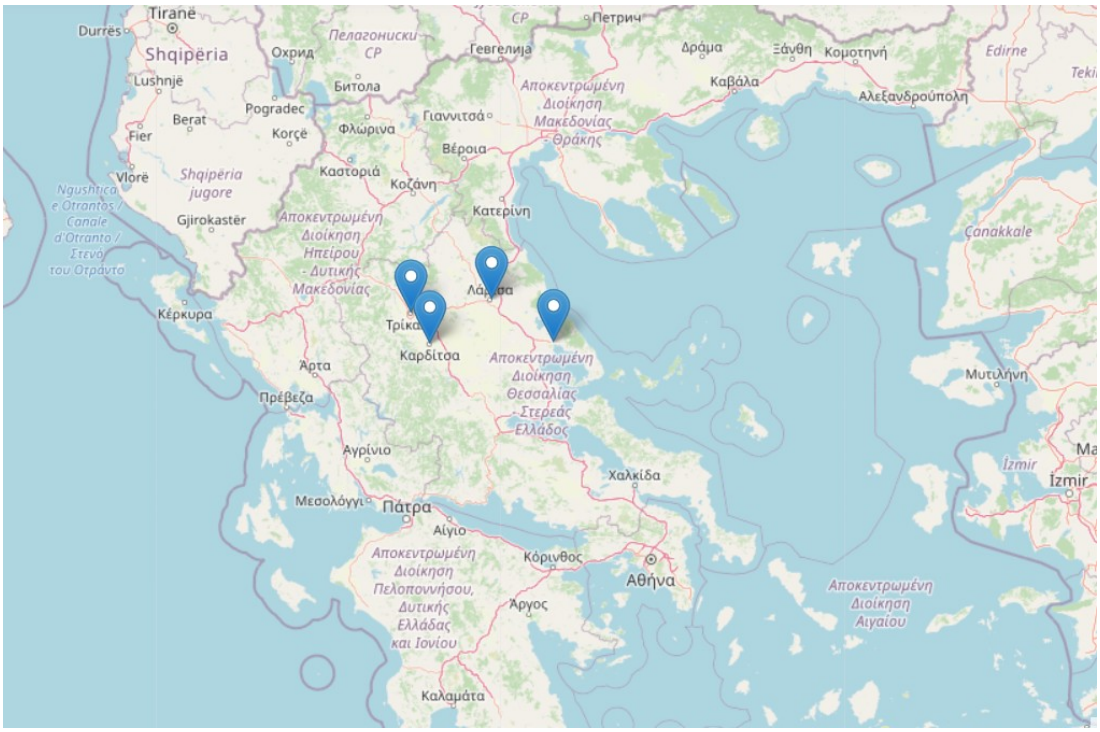
Venues of Thessaly, an overview

Manolis Efthimiou
04/2021

1. Introduction

Thessaly is a geographic and modern administrative region in central Greece, a country known for Tourism. In this project we will make an overview of the venues of the 4 biggest cities in Thessaly. Volos, Larisa, Karditsa and Trikala.

This project will make use of many skills a data scientist must have, like working with an API to gather data, data wrangling, data reduction, and data visualization.



2. Business Problem

The questions that will be asked in this research will be:

1) Which type of venues are found in these cities, and how good are their ratings?

2) How similar or dissimilar are these 4 cities that are geographically close, in regard to the venues they offer?

3. Target Audience

The goal of this research is to give the people of the region an overview of the entertainment options the biggest cities in Thessaly provide. In addition it can be used as a tool for seeing any gaps that can be filled by any businessman willing to make an attempt in the region.

4. Data acquired

The data needed for our research were gathered mostly from the Foursquare API, a platform whose purpose is to help people discover and share information about businesses and attractions around them. It should be mentioned that due to the volume of data we needed (over 50 Premium Calls to the Foursquare API) we had to upgrade from the default Sandbox Account to Personal Non-Commercial verified Account for the platforms developer account.

The data that were collected from Foursquare were about each venue's

- ✓ location, geographical coordinates about latitude and longitude, so as to be able to visualize the venues in maps
- ✓ category of venue, like cafeteria, bar, etc
- ✓ the ratings that Foursquare users gave them

From the Geopy API, we acquired the geographical coordinates of the cities automatically, for ease of receptiveness of the research for other locations.

5. Methodology

The research was done and presented in a Jupiter Notebook. The first step in our project was to find the geographical coordinates of the 4 cities under research. We did not hard code them so as to be easier to repeat the same procedures, with the least changes in the Notebook, for other locations.

After that we created a developer account so as to be able to communicate with the Foursquare API, and wrote some functions for the steps that would be repeated throughout the project. These functions were about creating the URL to hit the Foursquare API, making the HTTP Request to it, and handling the Response as to bring it in a form we could process with Python and it’s libraries.

5.1. Data collection

The functions we created were first to find any trends around our cities of interest. Due to the covid-19 pandemic the country of Greece is under lockdown so no trends were found.

So we proceeded with only with the venues that were visited by the Foursquare users during the previous years. The radius we searched was for 3km around the geographic coordinates of the cities, and for Volos specifically at 6Km because of the fact that it is a city located between the sea and mountain Pilio, so a lot of it’s activities take place in both coastal and the mountain. For every venue we took it’s geographic coordinates, its category, and its rating whenever it had.

The number of venues we acquired from Foursquare was 272.

5.2. Data cleaning

Next we cleaned the data so as to be able to process and interpret them easily. From all the category venues we discarded some as they did not contribute anything to our research, and split the remaining into 7 bigger venue types (*food, cafes, alcohol, leisure, cultural, athletic and sweets).* We then dropped the venues that were of category of the deleted as mentioned before.

The number of venues we had after removing the not needed was 257.

5.3. Data exploration

The next step was to split the total venues per type (of the 7 we created above) for every city

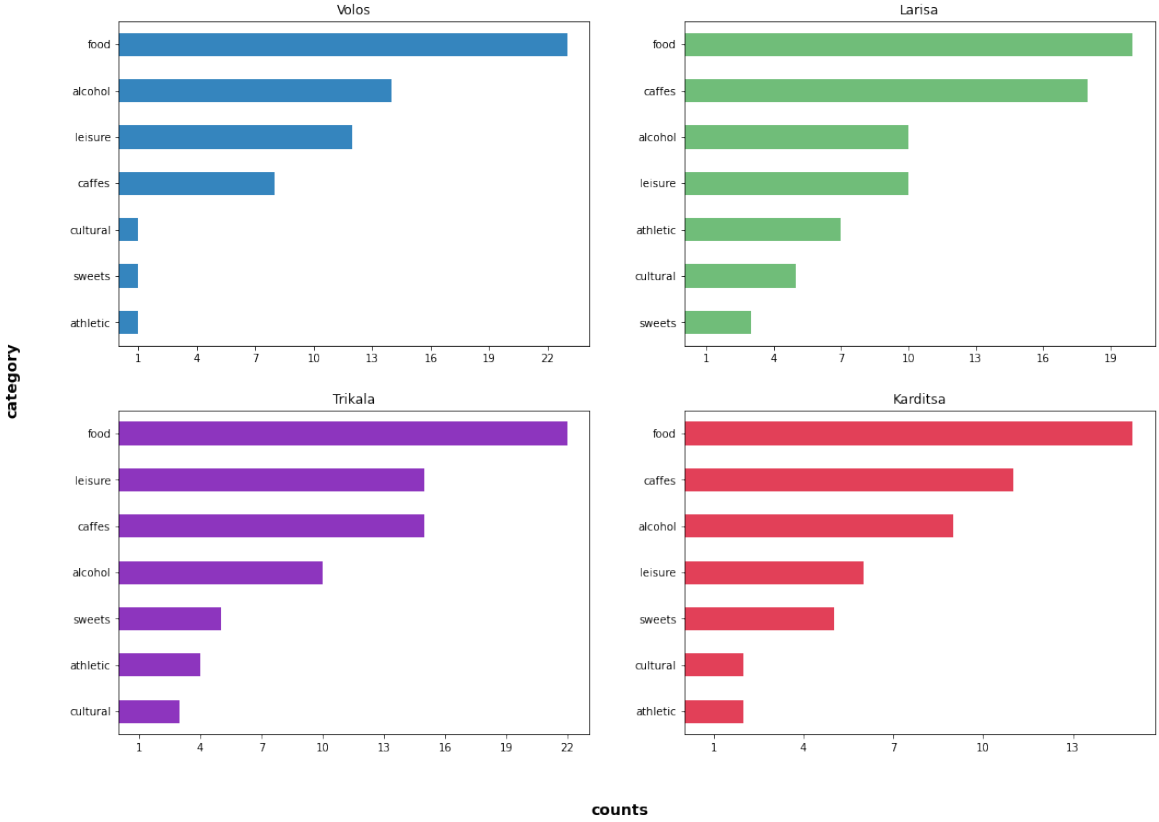


photo 1. venues per type

Next, for every venue type per city we:

✓ calculated the mean rating

✓ found the top venue

	karditsa	karditsa best	karditsa score	larisa	larisa best	larisa score	trikala	trikala best	trikala score	volos	volos best	volos score
food	15	8.1	7.48	20	8.8	7.82	22	8.3	7.67	23	9.2	8.42
caffes	11	8.8	7.81	18	8.9	7.88	15	8.3	7.58	8	9.0	8.36
alcohol	9	8.3	7.49	10	8.8	8.27	10	8.4	7.72	14	8.7	8.30
leisure	6	9.0	7.90	10	8.8	7.89	15	8.8	7.88	12	9.5	7.87
athletic	2	8.4	8.40	7	8.6	7.76	4	8.5	7.75	1	7.4	7.40
cultural	2	8.2	8.05	5	9.2	7.56	3	8.9	8.17	1	6.9	6.90
sweets	5	7.8	7.38	3	8.9	8.33	5	8.0	7.90	1	8.4	8.40

photo 2. total metrics

5.4. Visualization

We created 2 maps for each city. 1 with the top venues per venue type(right), and 1 with the rest (left). For instance we see Volos below

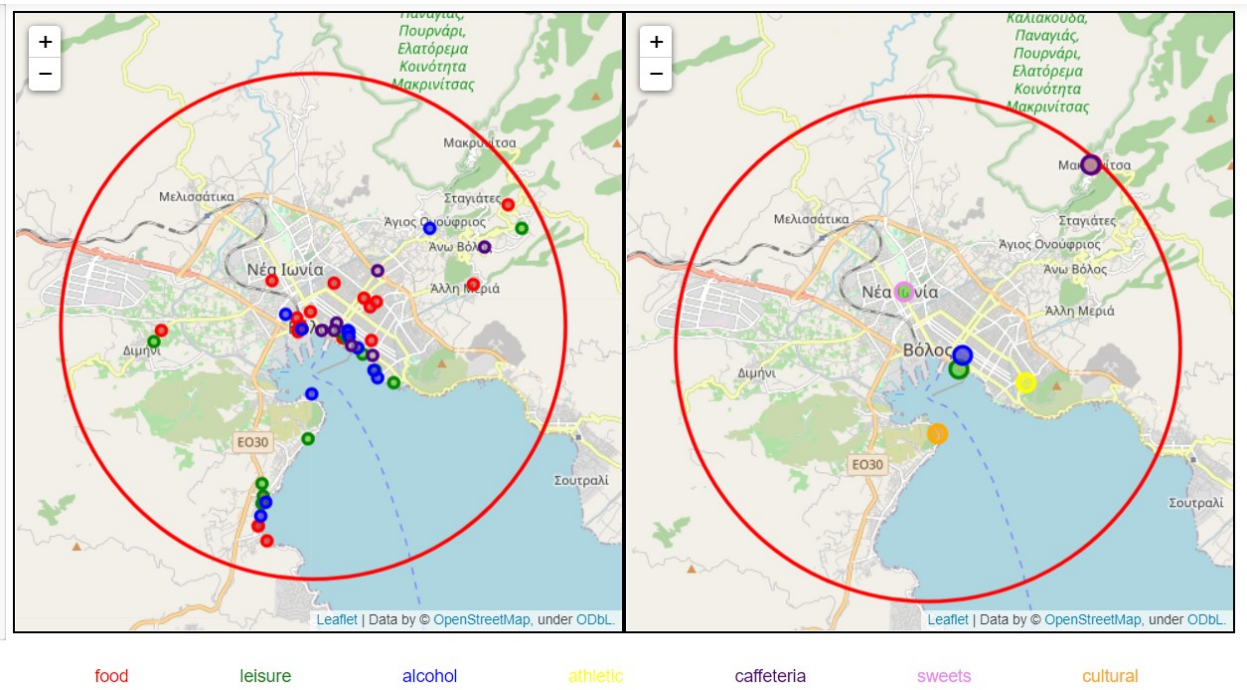


photo 3. map for Volos

The map production was automated by creating a function for it, and we check every venue's name, category and rating by clicking on it on the map.

5.5. Applying Machine Learning - Clustering

We decided to use the Unsupervised Machine Learning technique of clustering to discover natural grouping for the venues of every city.

5.5.1. Feature selection

We split the venues of every city into clusters taking into account their ratings and their type. For the latter since the venue type is a categorical variable we applied One-Hot Encoding as to transform it to form that could be used.

We did not normalize the data because it led to overfitting to the variable ['type'].

5.5.2. Number of clusters selection

We used the k-means clustering algorithm and decided to use 3 clusters because it led to the optimal and more distinct and interpretative clusters in general. Before we concluded to the decision of 3 clusters we checked the inertia *(sum of squared error between each point and its cluster center)* for 1 to 10 clusters tried.

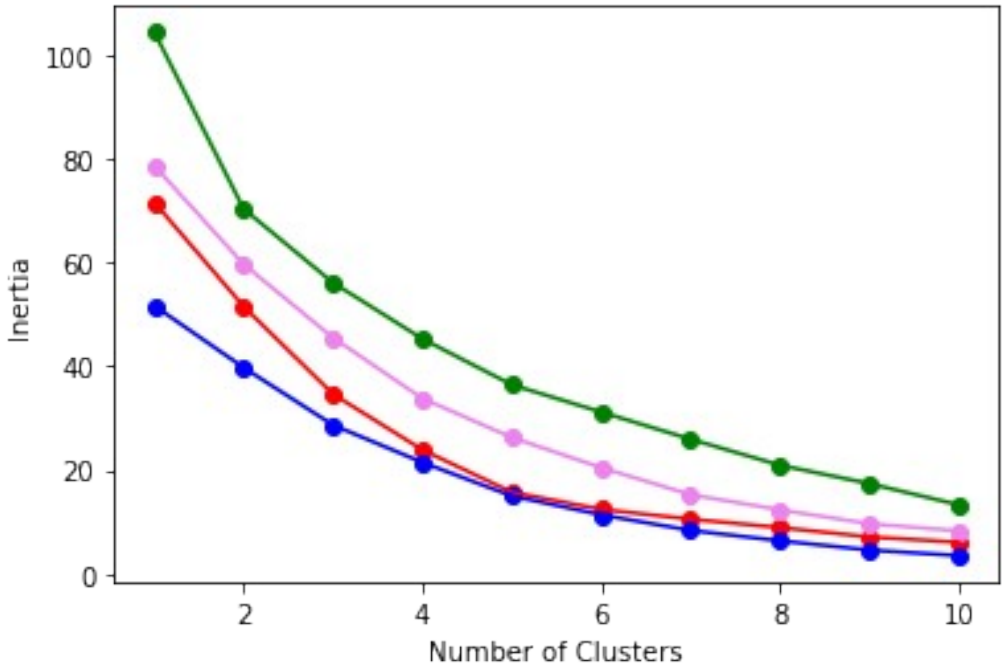


photo 4. inertia per number of Clusters

The elbow in the plot above is not clear for every city

5.5.3. Clustering results

After applying the clustering algorithm we created 1 dataframe with the total venues and the mean rating, found per city and per cluster

ratings

city	cluster	count	mean rating
karditsa	0	11	7.809091
	1	23	7.660870
	2	14	7.478571
larisa	0	16	6.643750
	1	39	8.348718
	2	12	8.158333
trikala	0	33	7.842424
	1	18	7.672222
	2	13	7.576923
volos	0	27	8.470370
	1	8	6.837500
	2	21	8.476190

photo 5. number of venues and mean rating per cluster for all cities

And 1 dataframe to see how many venues per category has each cluster per city

venues per category

city	cluster	type	count
karditsa	0	alcohol	9
		athletic	2
		cultural	2
		leisure	6
		sweets	5
	1	food	15
	2	caffes	11
larisa	0	alcohol	10
		athletic	6
		caffes	13
		cultural	2
		leisure	7
		sweets	3
	1	athletic	1
		caffes	5
		cultural	3
		food	3
		leisure	3
	2	food	17
trikala	0	alcohol	10
		athletic	4
		cultural	3
		leisure	15
		sweets	5
	1	food	22
	2	caffes	15
volos	0	alcohol	13
		caffes	8
		leisure	6
		sweets	1
	1	food	22
	2	alcohol	1
		athletic	1
		cultural	1
		food	1
		leisure	4

photo 6. number of venues per venue category per cluster for all cities

There are to totally 22 unrated venues for the 4 cities in total of 243 That is the reason sometimes we see some small differences in the total counts per city/per cluster in the dataframes above, since in the ratings count dataframe (left) the [count] column takes under account only the venues with numeric rating. The venues without rating will be displayed as we examine every city.

After that we created a map for every city with its clustered venues visualized on it

For example, we see Larisa below

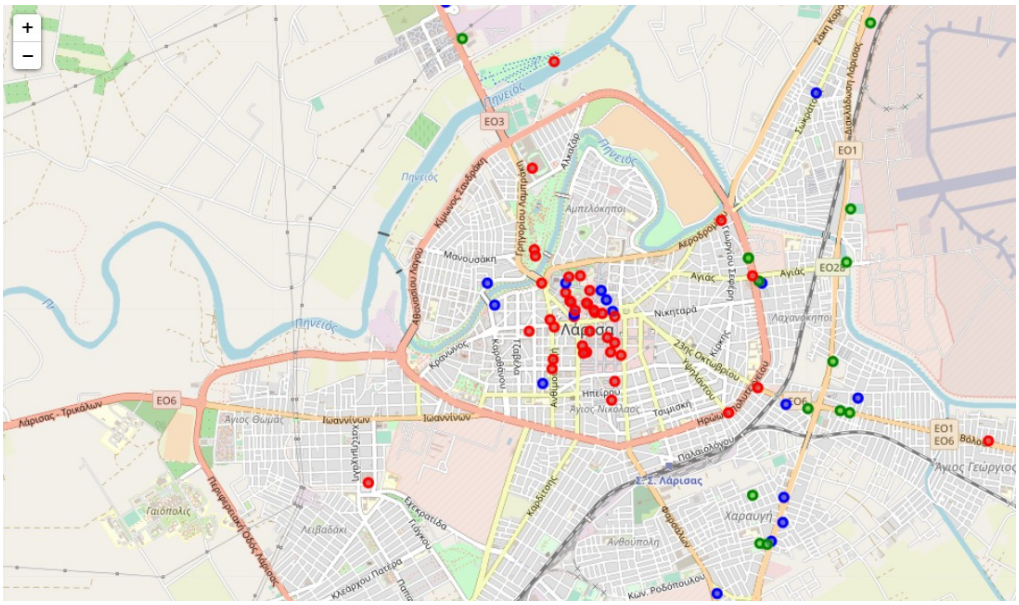


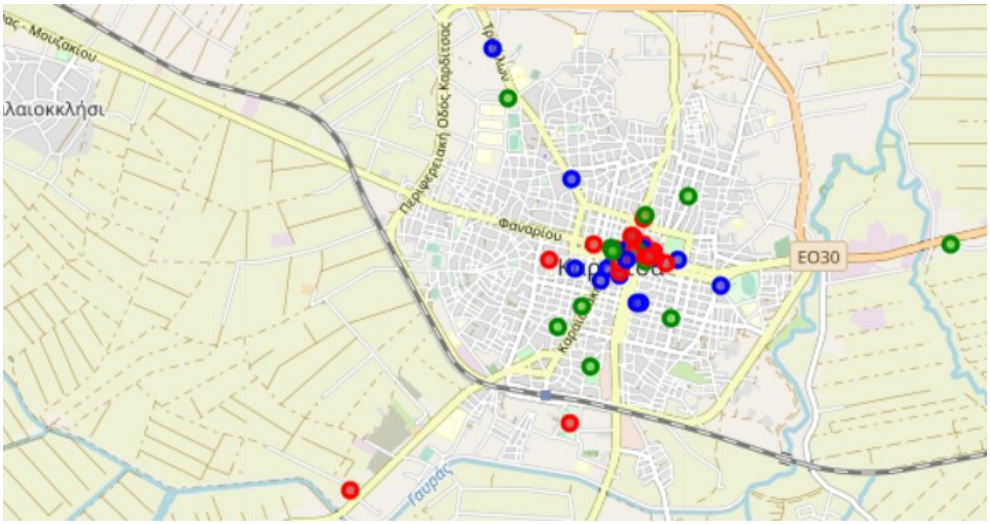
photo 7. map of clustered venues for Larisa

The map production was automated by creating a function for it and we can see every venue's name, category and rating by clicking on it on the map.

6. Results

With the dataframes and maps created at step 5.3.3, we have for every city

6.1. Karditsa



cluster 1 cluster 2 cluster 3

not rated venues

	name	categories	type	rating	latitude	longitude	id	cluster
42	Παλέρμο	Basketball Court	athletic	NaN	39.351821	21.917124	4bbc7eb93de8c9b66dff9aad	0
41	Το Μουράγιο	Greek Restaurant	food	NaN	39.359280	21.916027	52b21635498ea5997ec2e281	1

metrics

Total ratings

cluster	count	mean rating
0	17	7.417647
1	14	7.478571
2	17	8.000000

venues per category

cluster	type	count
0	**alcohol**	9
	athletic	1
	leisure	3
	sweets	5
1	**food**	15
2	**athletic**	1
	caffes	11
	cultural	2
	leisure	3

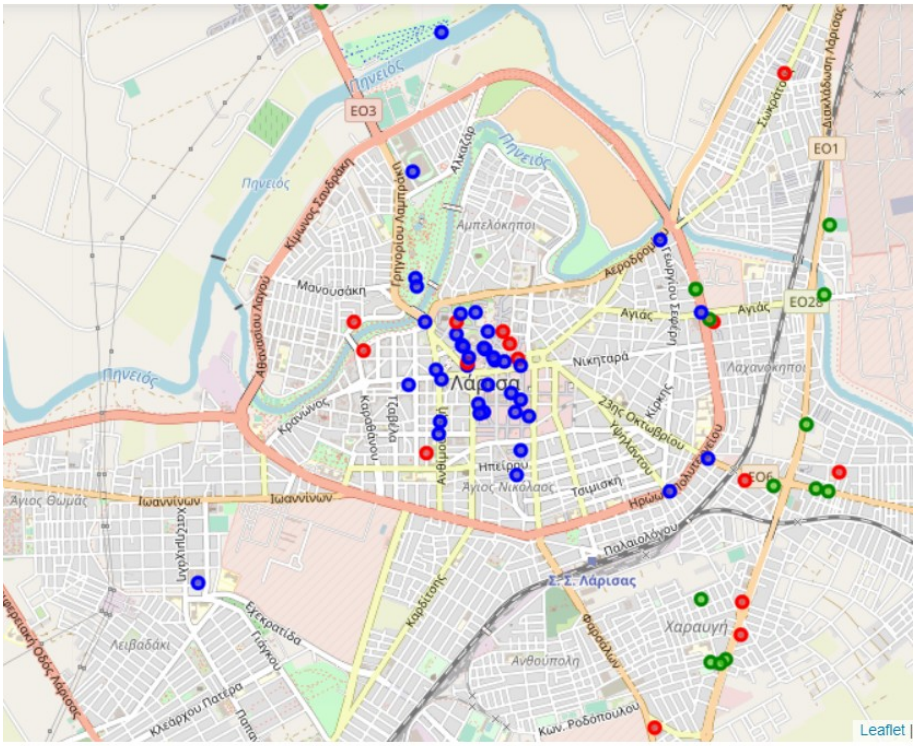
metrics

	karditsa	karditsa best	karditsa score
food	15	8.1	7.48
caffes	11	8.8	7.81
alcohol	9	8.3	7.49
leisure	6	9.0	7.90
athletic	2	8.4	8.40
cultural	2	8.2	8.05
sweets	5	7.8	7.38

- Food venues create a distinct cluster, this is found in all 4 cities
- The cafes also create a distinct cluster, this is also found in Trikala
- In the leisure type we see higher ratings for what was done from the local government in comparison to what is private economic activity

	name	categories	type	rating	latitude	longitude	id	cluster
0	Αθλητικό Πάρκο	Park	leisure	9.0	39.362470	21.932340	4baeff45f964a5200fe63be3	2
1	Παυσίλυπο	Park	leisure	8.8	39.364462	21.928102	4c93857a6b35a14390aa12dc	2
2	Πλατεία Λάππα	Plaza	leisure	7.8	39.361160	21.924297	4cb33b9c562d224b096e2c88	2
3	Kierion Hotel	Hotel	leisure	7.5	39.365671	21.919682	51522038e4b0c5b84881b0bd	0
4	Πλατεία Δικαστηρίων Καρδίτσας	Plaza	leisure	7.2	39.366100	21.923564	4e4b7d4545dd5144016c90eb	0
5	Cosi Summer	Nightclub	leisure	7.1	39.346628	21.895287	4fb7d089e4b0e5da459aae5d	0

6.2. Larisa



cluster 1 cluster 2 cluster 3

not rated venues

	name	categories	type	rating	latitude	longitude	id	cluster
0	Νέα Καστελλα	Greek Restaurant	food	NaN	39.627662	22.432223	4e328447ae60f21828d4c6cf	0
1	Γαλαξίας	Supermarket	food	NaN	39.633894	22.438311	50e81849e4b06746fb33c7f6	0
2	Κικιρίκου	Fried Chicken Joint	food	NaN	39.653231	22.434889	4f707a88e4b07a4bc71bcc70	0
3	Goody's	Fast Food Restaurant	food	NaN	39.658974	22.404417	4ce6e0bb8ef78cfac8658f9b	0
4	Άλσος Λάρισας	Park	leisure	NaN	39.655171	22.413340	4e751314b993a71aa3a36e3d	2
5	Corner Cafe	Café	caffes	NaN	39.631186	22.449050	4cc488bdd43ba14364bf67f8	2

metrics

cluster		
0	12	8.158333
1	16	6.643750
2	39	8.348718

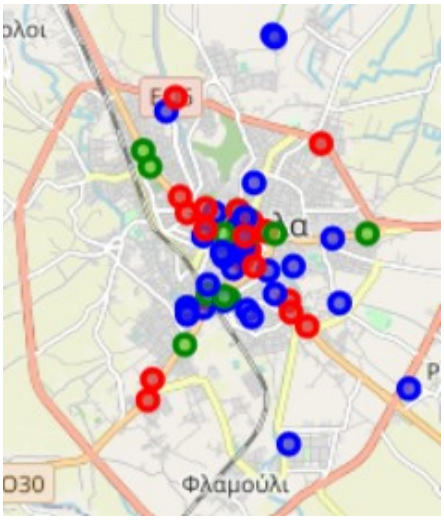
cluster	type	
0	**food**	16
1	**athletic**	1
	caffes	5
	cultural	3
	food	4
	leisure	3
2	**alcohol**	10
	athletic	6
	caffes	13
	cultural	2
	leisure	7
	sweets	3

food	20	8.8	7.82
caffes	18	8.9	7.88
alcohol	10	8.8	8.27
leisure	10	8.8	7.89
athletic	7	8.6	7.76
cultural	5	9.2	7.56
sweets	3	8.9	8.33

- Food venues create a distinct cluster, this is found in all 4 cities
- We see generally high ratings for most of the venues
- Most of the venues are in the city's center and alongside a main road connecting the city with the national highway
- The highest rated venues are spread in the center of the city and are most caffees and food
- The lowest rated cluster is for the furthest of the city's center venues, found mostly in the main road connecting the city with the national highway
- The cultural venues of Larisa have high ratings (with the exception of a music center)

	name	categories	type	rating	latitude	longitude	id	cluster
0	Ancient Theatre of Larissa (Αρχαίο Θέατρο Λάρι...	Historic Site	cultural	9.2	39.640008	22.414713	4c40a11fda3dc928fea3c7b9	2
1	Φρούριο	Historic Site	cultural	8.3	39.641620	22.414491	4f3022b2121d3f490074b848	2
2	Studio 3 (Palace 91,4)	Music Venue	cultural	7.4	39.642809	22.429286	4dc58cb6fa76d685cde3c214	1
3	ΑΒ Βασιλόπουλος	Shopping Mall	cultural	7.4	39.636255	22.436260	57ebe0c0498efbe718341c5f	1
4	Starz Live	Music Venue	cultural	5.5	39.657651	22.439292	4ee3d6f193adf8e1a714474d	1

6.3. Trikala



not rated venues

	name	categories	type	rating	latitude	longitude	id	cluster
57	Billy's "Πλατεία Βουβής"	BBQ Joint	food	NaN	39.555512	21.761352	4e2a8c56091ac5a470e5433e	0
58	Το χωριάτικο ψωμί	Bakery	food	NaN	39.557221	21.758896	51a08de6498e3fa33f233022	0
65	Κουίντα	Ouzeri	food	NaN	39.570562	21.757149	52a3400911d2996c3dc683d1	0
66	Παλιό Μεράκι	Restaurant	food	NaN	39.535752	21.752805	4e0db53de4cd27fc7d27b5f1	0
63	Granello	Cafeteria	caffes	NaN	39.564529	21.752278	57e1031c498e16b14c8c057d	1
64	Τσιμπλής Cafe	Café	caffes	NaN	39.542275	21.758482	4f4d27ffe5e882092459e119	1
59	Πλατεία ΟΣΕ(ΚΔΑΠ)	Park	leisure	NaN	39.547250	21.764129	5022db5be4b0e6fe19fce8f4	2
60	Πλατεία Φιλοσόφων	Park	leisure	NaN	39.546135	21.767827	4ed7ee60f5b915cfe2d92b3a	2
61	Μέλισσα	Dessert Shop	sweets	NaN	39.545355	21.768281	4fb89a22e4b093431c459bdc	2
62	Παγοδρόμιο Μύλου Ξωτικών	Playground	leisure	NaN	39.545635	21.758890	50b91a07e4b0e9ecea75d585	2

metrics

Total ratings

cluster	count	mean rating
0	18	7.672222
1	13	7.576923
2	33	7.842424

venues per category

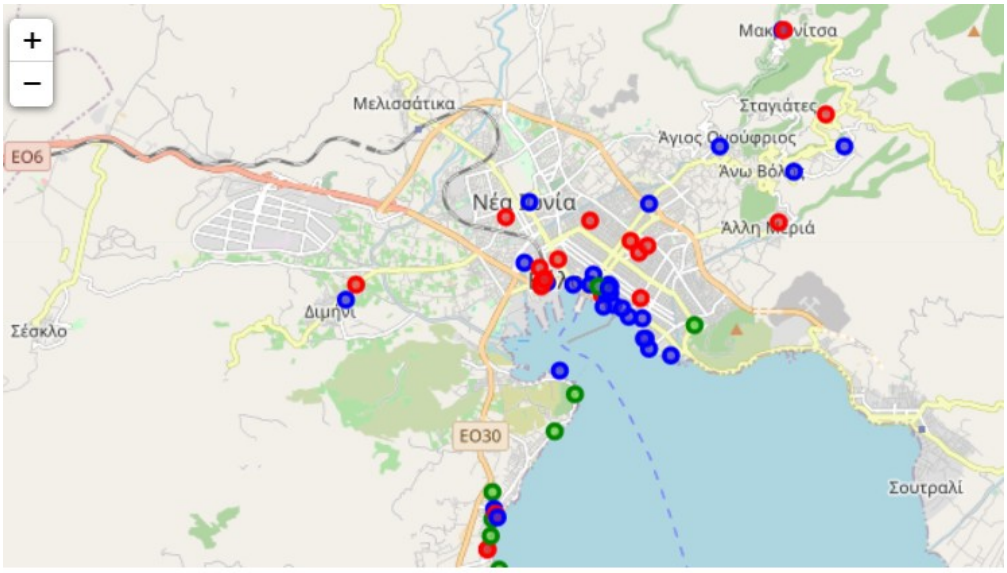
cluster	type	count
0	**food**	22
1	**caffes**	15
2	**alcohol**	10
	athletic	4
	cultural	3
	leisure	15
	sweets	5

metrics

	trikala	trikala best	trikala score
food	22	8.3	7.67
caffes	15	8.3	7.58
alcohol	10	8.4	7.72
leisure	15	8.8	7.88
athletic	4	8.5	7.75
cultural	3	8.9	8.17
sweets	5	8.0	7.90

- Food venues create a distinct cluster, this is found in all 4 cities
- Most of the venues are in the city's center and alongside the city's biggest roads leading to it
- Almost half of the venues are unrated (10/22), in contrast with the rest cities that the unrated are at most 4
- The cafes also create a distinct cluster, this is also found in Karditsa
- The highest ratings are mostly about food, located in the very center of the city and alongside the main roads that lead to it

6.4. Volos



cluster 1 cluster 2 cluster 3

not rated venues

	name	categories	type	rating	latitude	longitude	id	cluster
0	To kalamaki	Restaurant	food	NaN	39.325409	22.926775	575425d2498e02e2045e71ca	0
1	Πλατεία Διμηνίου	Plaza	leisure	NaN	39.359314	22.896027	5611143c498eb70a6ebe058b	2
2	Hotel Filoxenia	Hotel	leisure	NaN	39.326206	22.926289	4debe8231520ed580a544a4e	2
3	Ammos Seaside Lounge	Beach Bar	alcohol	NaN	39.324950	22.926922	5b3fb8b97cd14c004461510f	2

metrics

Total ratings

cluster	count	mean rating
0	21	8.47619
1	8	6.83750
2	27	8.47037

venues per category

cluster	type	count
0	**food**	22
1	**alcohol**	1
	athletic	1
	cultural	1
	food	1
	leisure	4
2	**alcohol**	13
	caffes	8
	leisure	8
	sweets	1

metrics

	volos	volos best	volos score
food	23	9.2	8.42
caffes	8	9.0	8.36
alcohol	14	8.7	8.30
leisure	12	9.5	7.87
athletic	1	7.4	7.40
cultural	1	6.9	6.90
sweets	1	8.4	8.40

- Food venues create a distinct cluster, this is found in all 4 cities
- Most of the venues are alongside the sea, and mostly venues for eating are in the central city
- The highest ratings are inside the city and in mount Pilion
- The lowest ratings are the furthest from the city, but still alongside the sea
- In the leisure type we see higher ratings for what was done from the local government in comparison to what is private economic activity

	name	categories	type	rating	latitude	longitude	id	cluster
0	Παραλία Βόλου	Pedestrian Plaza	leisure	9.5	39.358068	22.948776	4de126ff18380dc4dd41b00d	2
1	Agios Konstantinos Park (Πάρκο Αγίου Κωνσταντί...	Park	leisure	8.9	39.356730	22.953834	4f524211e4b0ac6d0bcef9fd	2
2	Πήλιο	Mountain	leisure	8.7	39.383532	22.997752	514dc843e4b0d91e9293e9ac	2
3	Agios Nikolaos Square (Πλατεία Αγίου Νικολάου)	Plaza	leisure	8.6	39.360179	22.949553	50c3a7fd498ef1fd4464c339	2
4	Domotel Xenia Volos	Hotel	leisure	8.5	39.353272	22.957141	4bd1cd775e0cce724c03a284	2
5	Anavros Beach (Παραλία Αναύρου)	Beach	leisure	8.0	39.350509	22.962294	4e3696c2fa7656ba317c19bd	2
6	Ermou (Ερμού)	Pedestrian Plaza	leisure	7.7	39.361292	22.947714	4fa270dee4b0abdff9a7bcca	1
7	Αλυκές	Beach	leisure	6.9	39.324737	22.925909	4e3a88941838961aff006ded	1
8	Kalloni beach resort & Spa	Resort	leisure	6.1	39.328984	22.926185	5039e5dfe4b0e79b212e7303	1
9	Amaze Club	Nightclub	leisure	5.8	39.338620	22.938719	4dbb583793a08f9274a2b3c3	1
10	Πλατεία Διμηνίου	Plaza	leisure	NaN	39.359314	22.896027	5611143c498eb70a6ebe058b	2
11	Hotel Filoxenia	Hotel	leisure	NaN	39.326206	22.926289	4debe8231520ed580a544a4e	2

7. Discussion

✓ The most frequent venue categories are explained by the nature of Greece's economy which relies a lot on Tourism, local or foreign.

✓ In the venues of type ‘leisure we saw higher ratings for what was done from the local government, like parks and plazas, in comparison to what is private economic activity, like nightclubs and resorts

✓ High rating, so interest too, for venues of historical interest.

✓ The cafes had to be a separate type of venue since the Greek population culture.

✓ The biggest similarity between the 4 cities is that the food venues were always a distinct cluster.

✓ There is a correlation with a venues geographic proximity to the city’s center or at least it’s accessibility with main roads, and its ratin

✓ The highest ratings are for the city of Larisa, the biggest in population and known for its night life.

✓ For Volos, the highest ratings for the venue types of food, cafes, and athletics are inside the radius of 3-6 km, meaning we correctly took a bigger radius for exploring in the case of Volos

✓ The highest rated venues were found in Volos, for the category of food (mean of 8.42). That is explained by the fact that region is well known for its restaurants with tsipouro (a drink with alcohol).