Applied Data Science Capstone Project

Adam Rubins

Business Problem

- My project is devoted to exploring and segmenting elementary schools in Tel Aviv (a major city located in the centre of Israel), based on their proximity to other venues that are of particular interest to the client. The project aims to provide clients, who wish to live in close proximity to a school, with data concerning other venues in the area, in order to assist them in understanding and narrowing down their choices and eventually deciding where to rent an apartment.

Target Audience - the clients

- My project was designed for a specific client - my friend, who is a 32 year old male, married with 2 children and a pet dog. This is a middle class family that intends to move to the city of Tel Aviv. However, the target audience for this project is far wider, and includes young families that wish to live in Tel Aviv, in close proximity to an elementary school and in a child and family-friendly environment.

The Client's Particular Interest

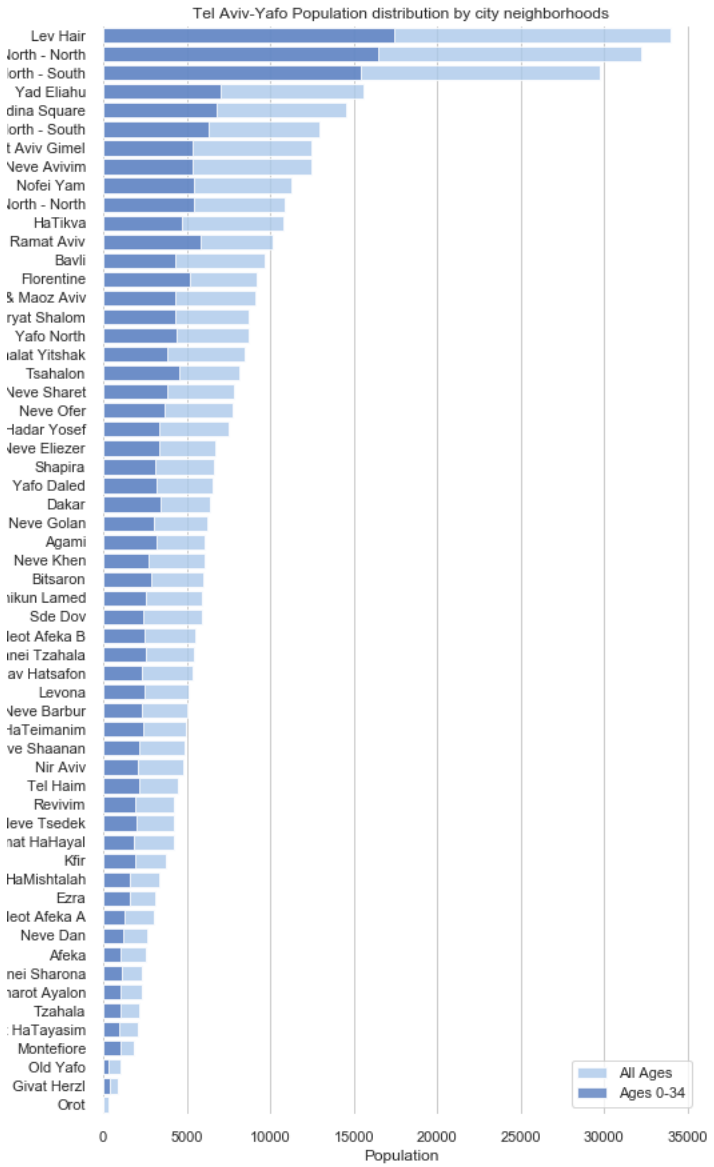
- The client's main particular interest is the proximity to elementary schools. These schools must abide by several specific parameters:
- State (public); ownership: city.
- Stream of education: Jewish
- Type of education: regular (secular)
- There are additional parameters that are of particular interest to the client, as listed bellow. He wishes to live in a walking distance from these venues, which he considers to be 500 meters. The client strongly prefers there to be a large variety of these venues of interest in close proximity (for example, there is great advantage in a wide variety of kindergartens and daycares, as many may have long waiting lists):
  - kindergartens (see specific parameters listed above for elementary schools)
  - Playgrounds with at least one facility for toddlers (mostly interested in the number of playground facilities available nearby)
  - Green public areas (mostly interested in the total green area in meters)
  - Dog gardens
  - Pizza places
  - Ice cream parlors
  - Pubs/bars (specific types of venues with no interest to the client were filtered out, as detailed in the full report in foursquare data collection).
  - Client does not wish to live in close proximity to strip clubs.
  - Population distribution by age – the client prefers to live in a relatively young neighbourhood and considers 34 to be a young age.

Project data sources and API’s

- **TLV OpenData -** A free publicly available website provided by the Tel Aviv-Yafo Municipality
  - GisLayers - A REST API for GIS (Geographic Information System) Layers
- **Foursquare Places API -** Which offers real-time access to Foursquare’s global database of venue data and user content.
- **Google Maps -** I use Google Maps for two purposes. First, to retrieve the Tel Aviv-Yafo coordinates, which I use as the center for the map visualization in the project. Second, to obtain the properly formatted (English) addresses of the schools from the school geocode (latitude, longitude) data. I will obtain this data through the Tel-Aviv Municipality API (GisLayers REST API for Geographic Information System Layers).
- **Google Cloud Translate** - most of the data that is returned by the Tel-Aviv Municipality API is in Hebrew. I will translate (to English) the fields’ names and any fields’ values that will be used for filtering or understanding the data analysis. I will not translate the names of the venues (such as schools, day cares, kindergartens, etc.) and will be using system Id's in the report.

DATA

- **Tel-Aviv Municipality API - Population distribution by city neighbourhoods:** Unfortunately, the most recent data available is from 2017. After consulting with the client, I will not use this data in the clustering process. However, I will use it to give the client an indication of the age distribution of the population, as well as to filter out schools located in neighbourhoods where there are no people aged under 34.



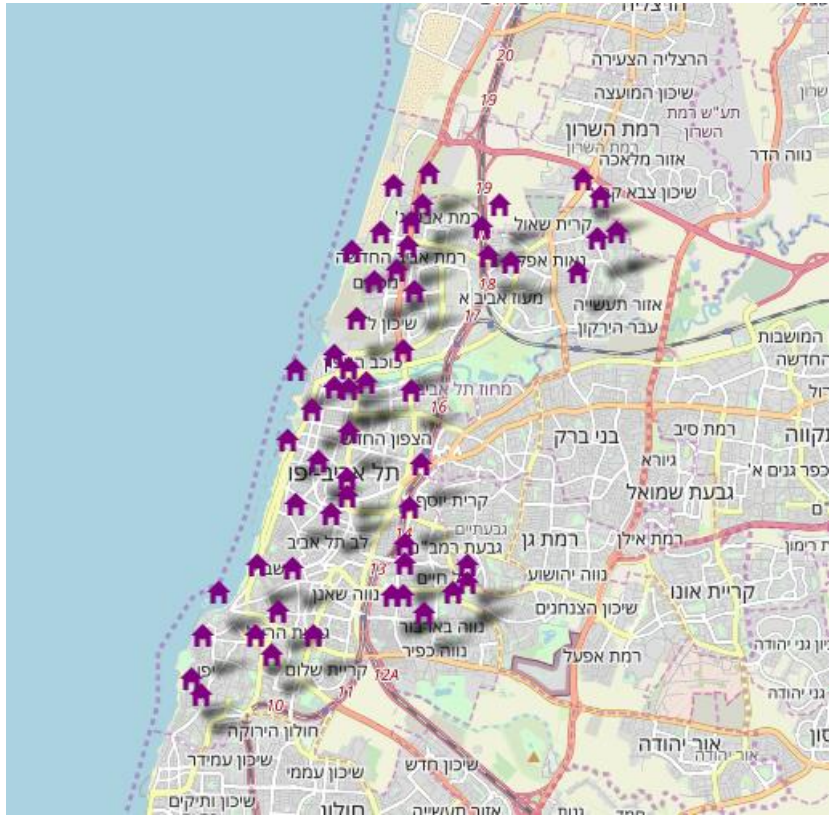
DATA

- **The Schools Data:** 56 schools.

Filter only:

- State (public); ownership: city.
- Stream of education: Jewish
- Type of education: regular (secular)

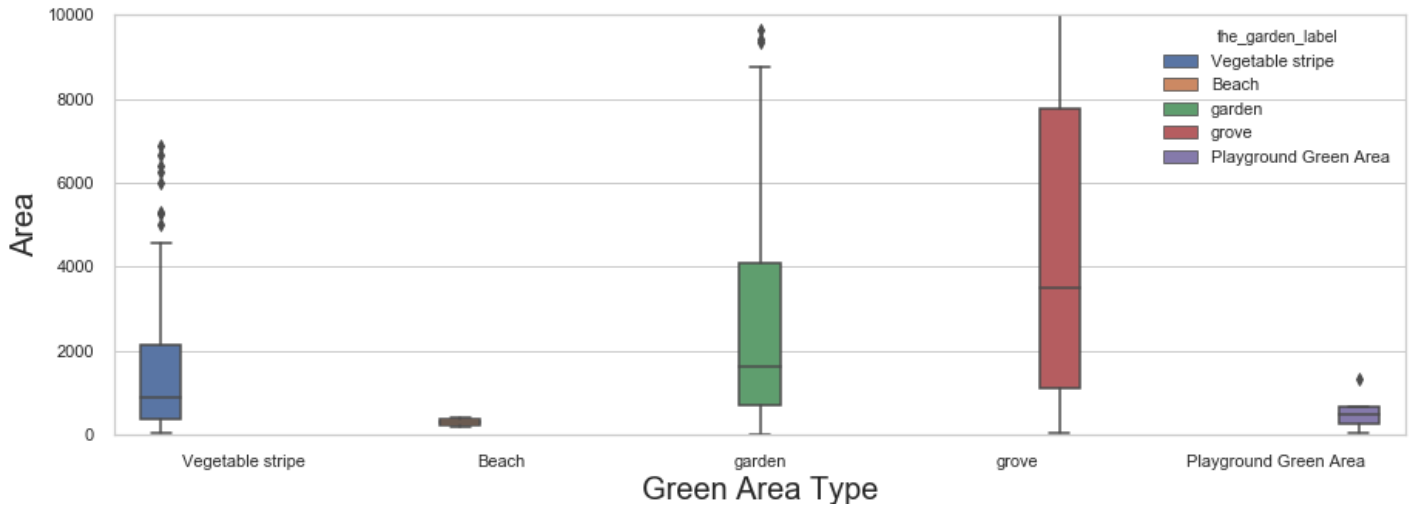
	system_id	institution_name	longitude	latitude	points	school_address
0	599-300070	ארן	34.790260	32.108128	POINT (34.790259756694 32.10812820184408)	Yehuda Burla St 25, Tel Aviv-Yafo, Israel
1	599-300150	ארנון	34.784813	32.088542	POINT (34.7848129994323 32.08854244352671)	David Yellin St 11, Tel Aviv-Yafo, Israel
2	599-300230	צוקי אביב	34.794272	32.125635	POINT (34.79427209646924 32.12563548292588)	Yair Rozenblum St 11, Tel Aviv-Yafo, Israel
3	599-300310	כוכב הצפון	34.786380	32.101471	POINT (34.78637968904724 32.10147101961964)	Abba Kovner St 16, Tel Aviv-Yafo, Israel
4	599-300490	יהודה מכבי	34.784702	32.092574	POINT (34.78470188636309 32.0925735999618)	Antigonus St 6, Tel Aviv-Yafo, Israel

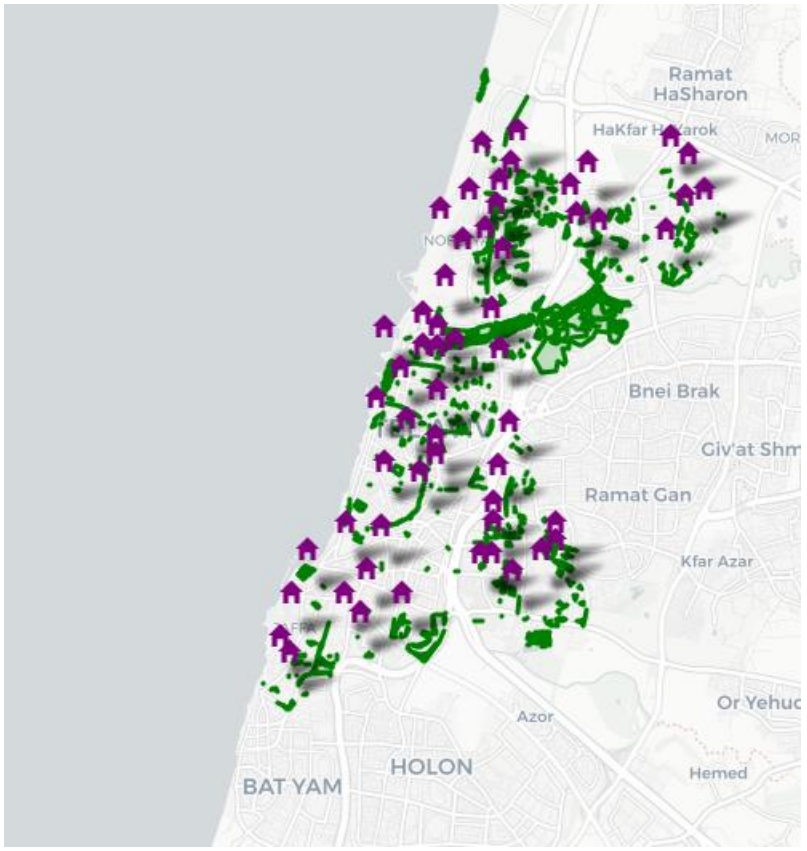


DATA

- **The Green Areas Data**

Note that the green types: "Traffic island" and "Temporary grove" were filtered out

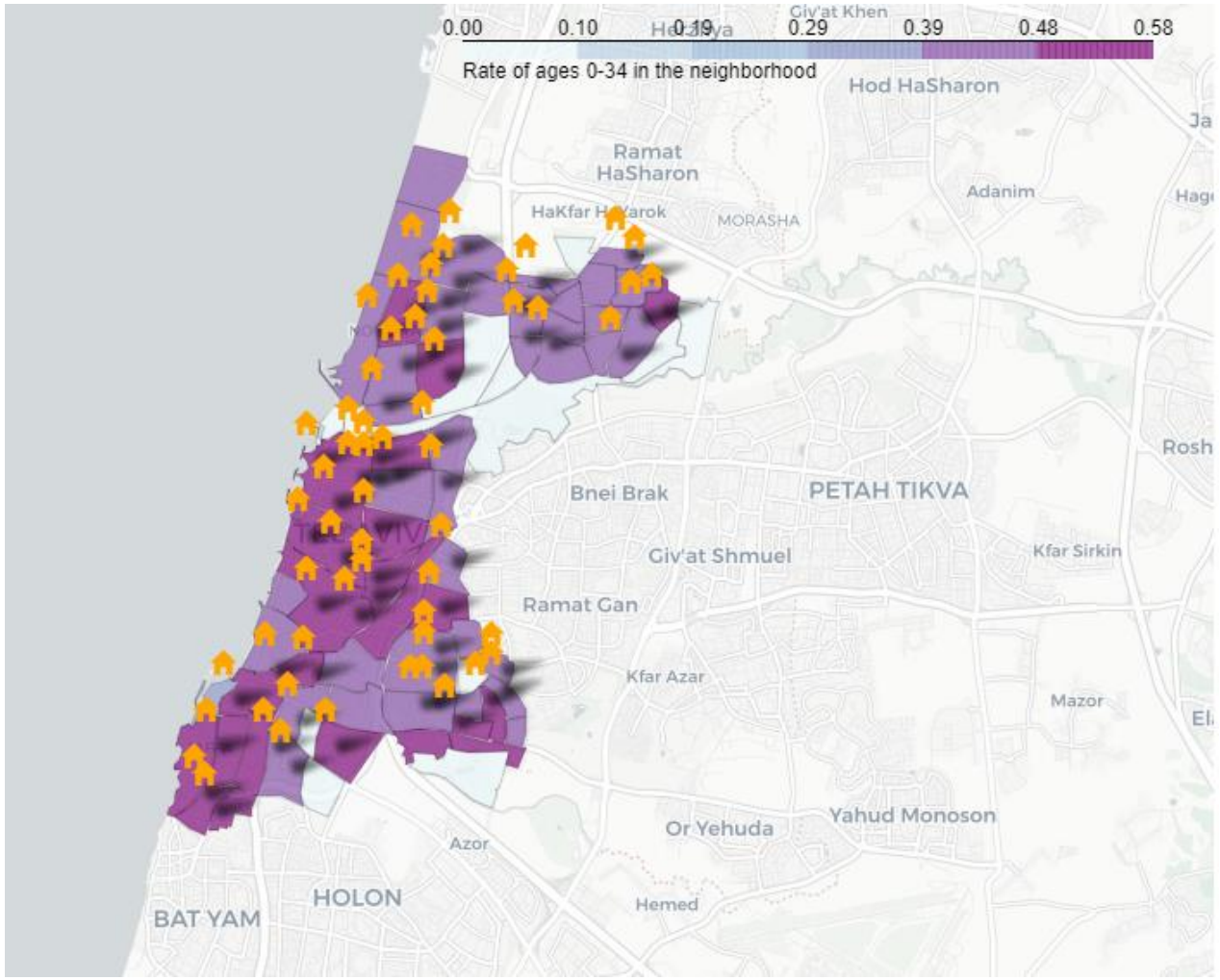




DATA

- **The Neighborhoods Data –**

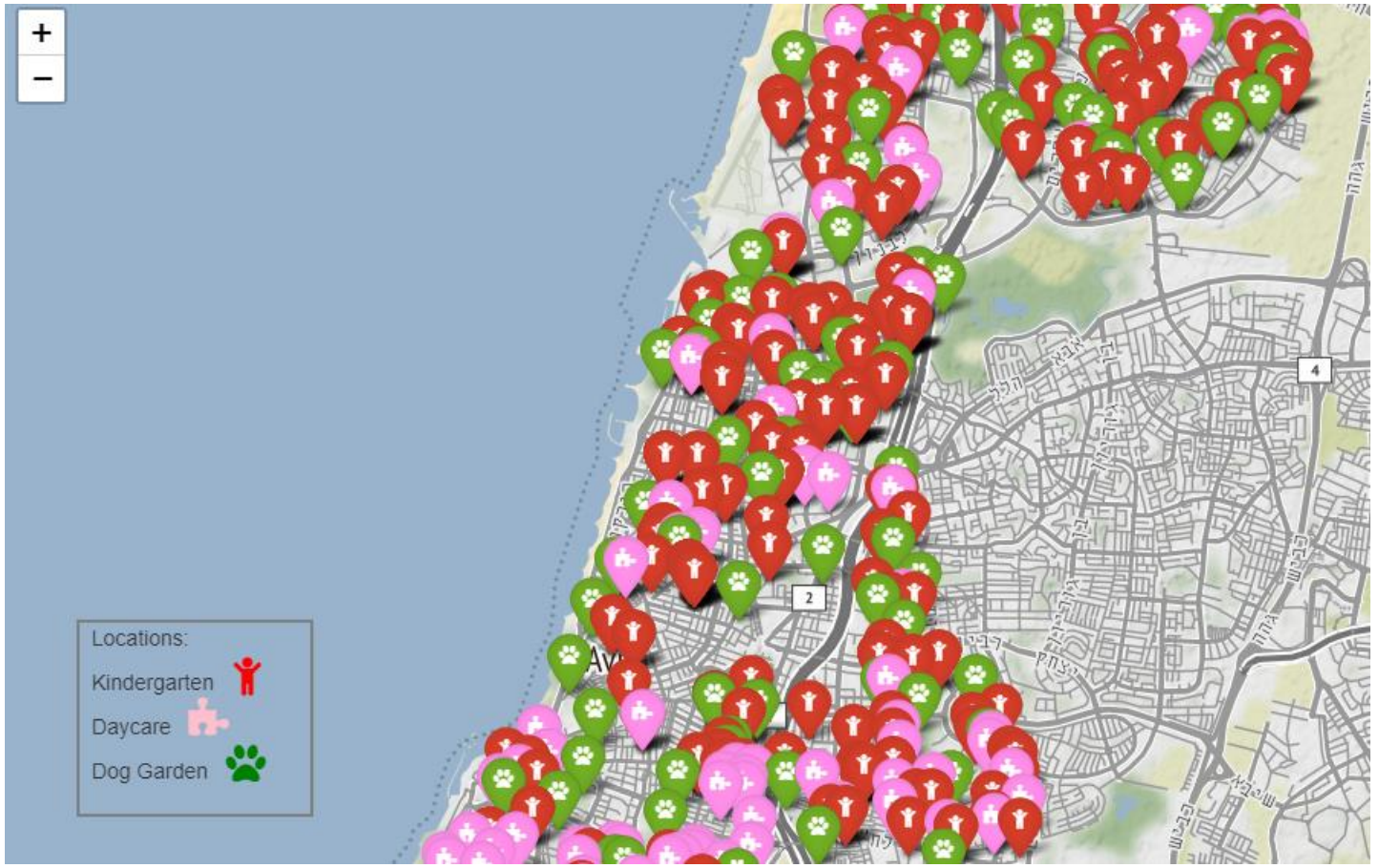
I will use the neighbourhoods data to match schools to neighbourhoods (each school is in only one neighbourhood). And incorporate a measure of age distribution (that is available for neighbourhoods) to the schools data.



DATA

- **The *Dog Gardens***
- ***Daycares* Data**
- ***kindergartens***

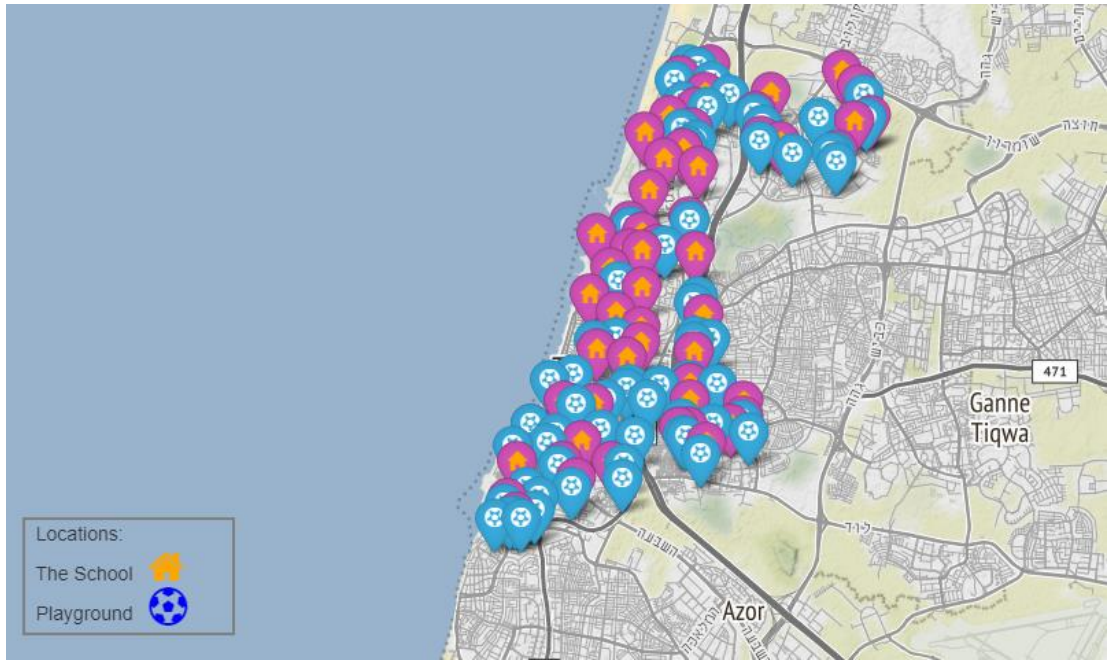
(same filters to *kindergarten as to school)*

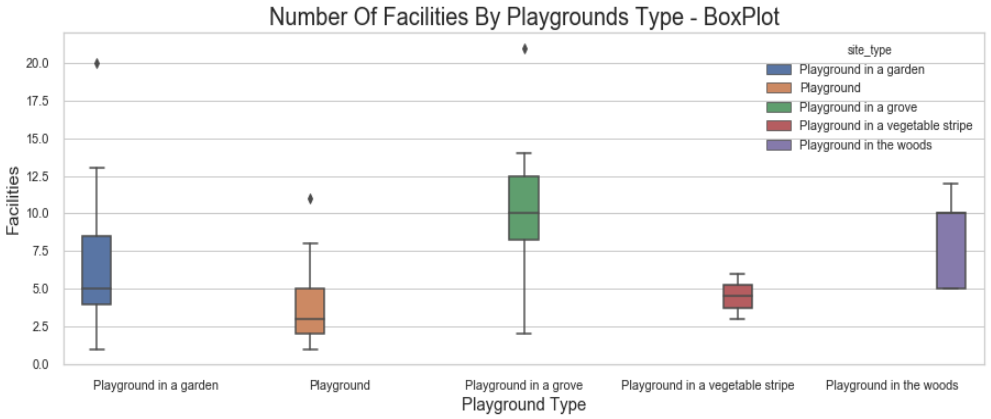


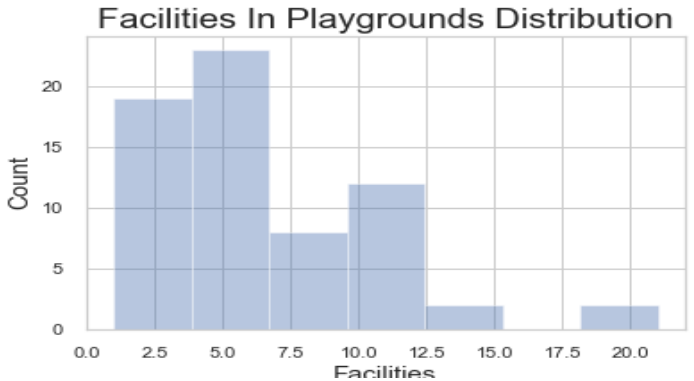
DATA

- **The Playground Data**

Note that the playground with no facilities for toddlers was filtered out



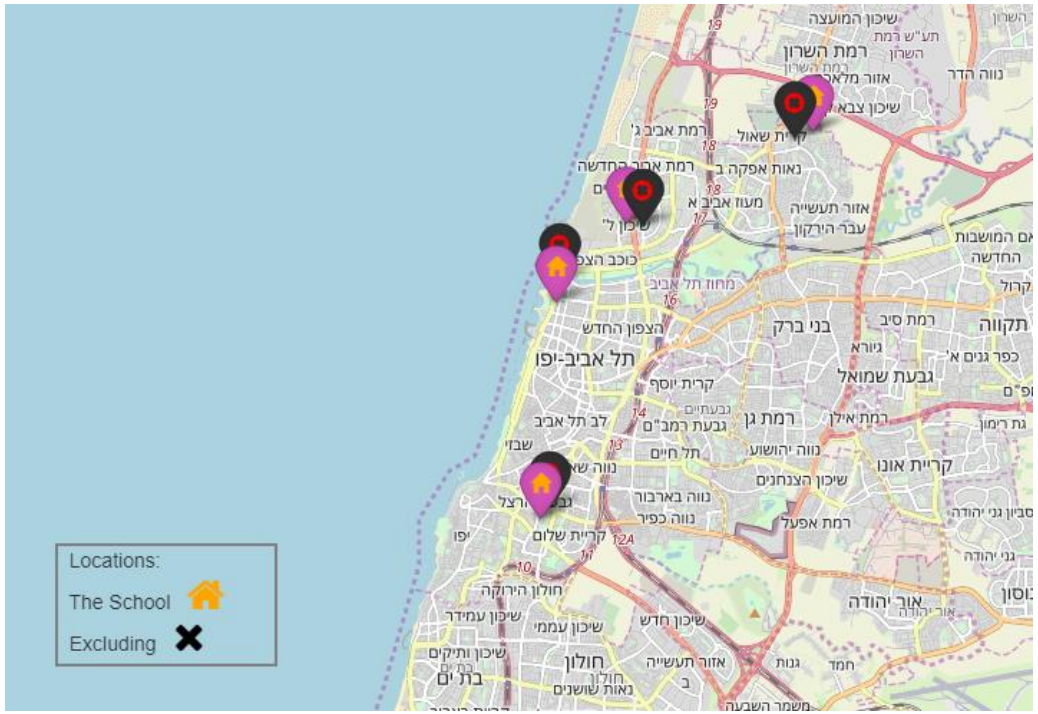




DATA

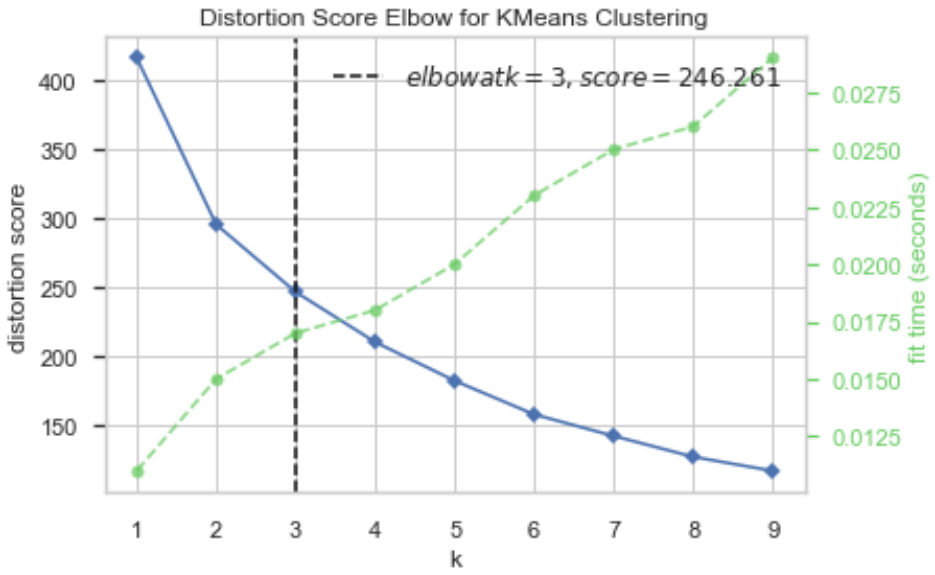
- **More Venues**
  - **Ice Cream Shop - 86**
  - **Pizza place – 121**
  - **Parents fun (bars/pubs) – 314**
  - **School excluding strip clubs - 3**

- **Filtering schools**
  - **near a strip clubs**
  - **With no population (up to age 34)**



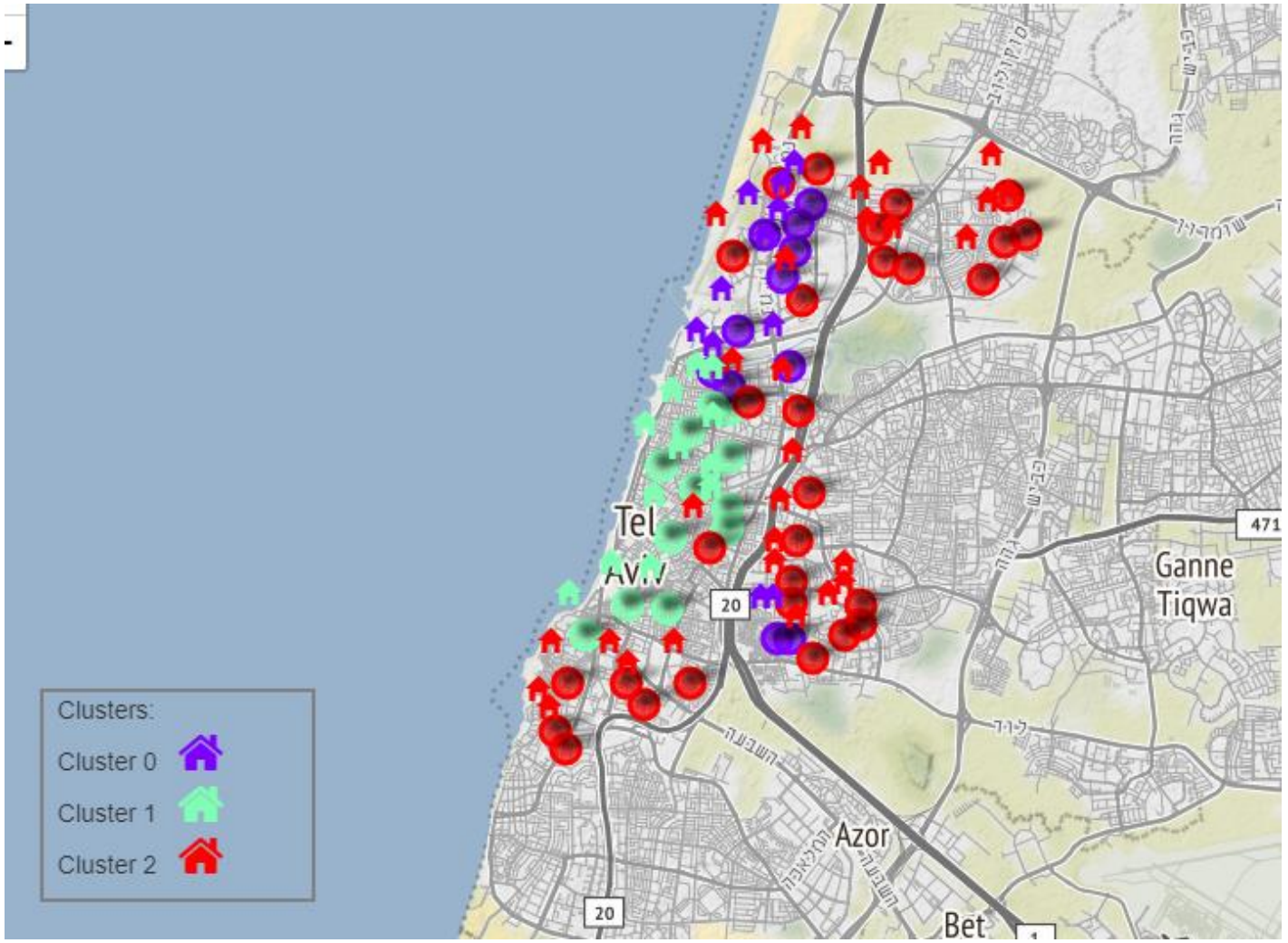
Methodology

- Pre-processing
  - calculate the distance for all locations from all the schools
  - Sum the Green areas & playground facilities near each school
  - Merging, transforming and filtering the data
  - Normalize the data over the standard deviation
- Modelling
  - Finding The Optimal K (number of clusters) Using The Elbow Method
  - Apply k means clustering

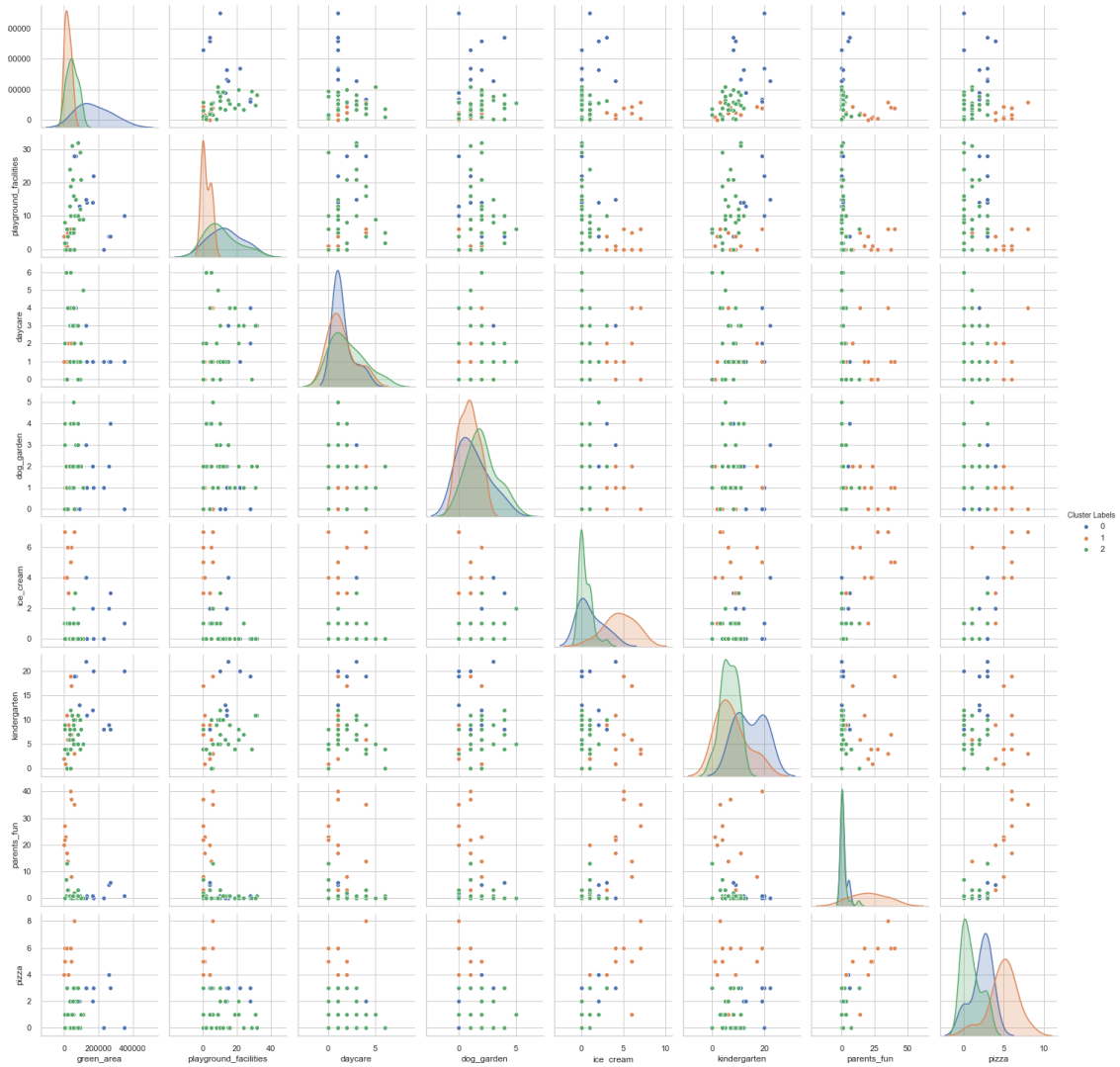


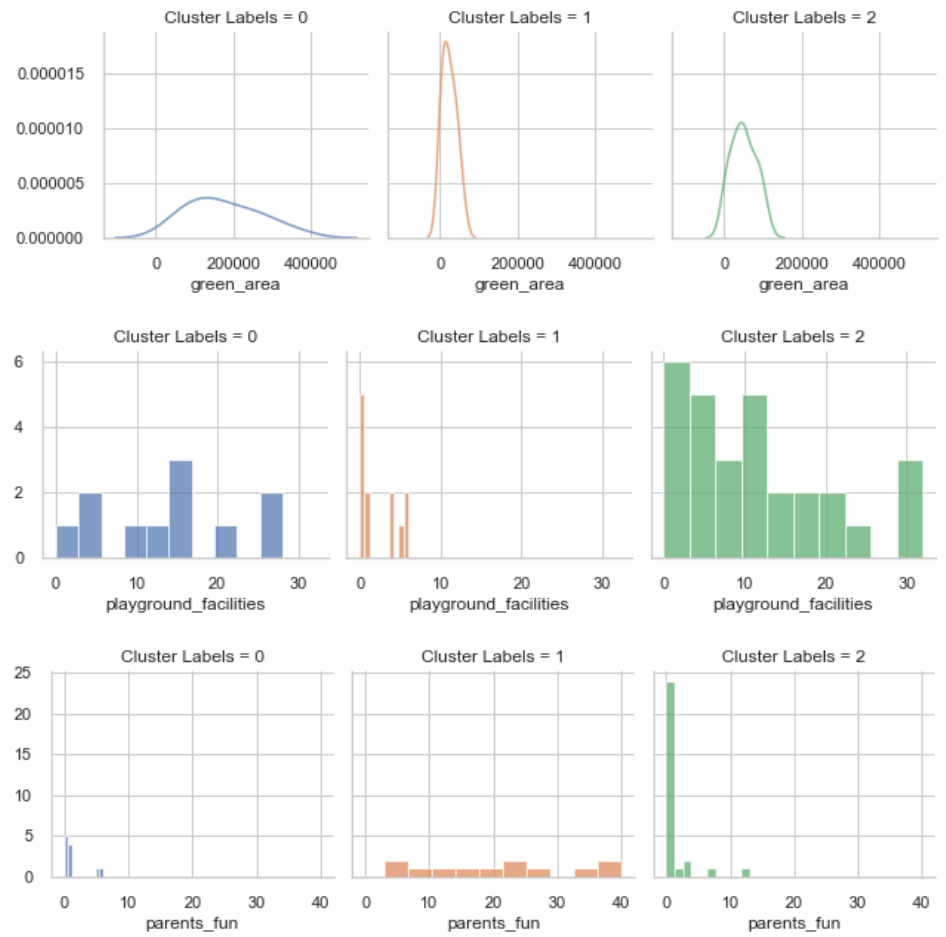
Results

Cluster Labels	School Count
0	11
1	12
2	29



Characterization of the clusters



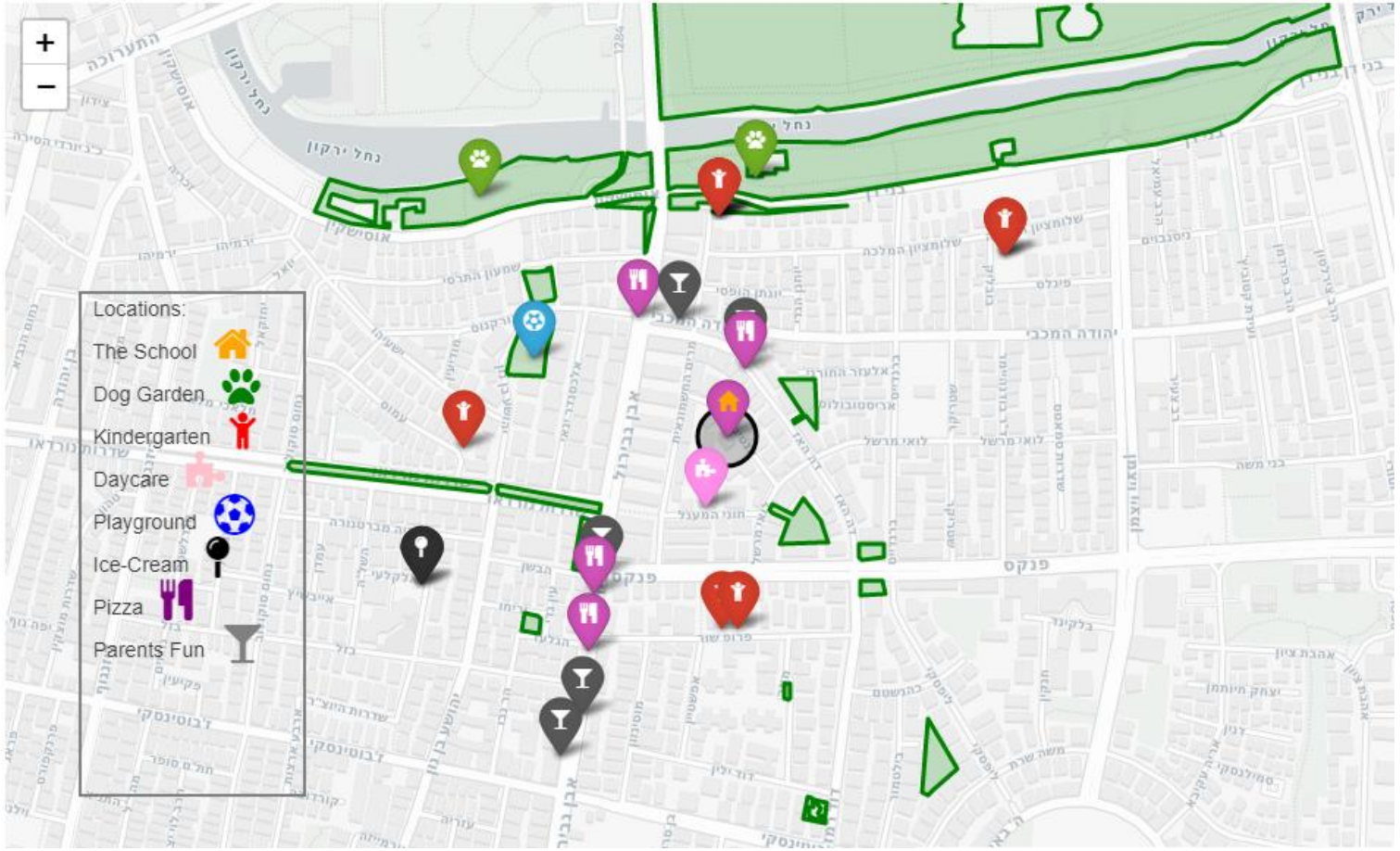


Discussion

- According to my conclusions, **cluster 0** seems most suitable for the client's needs. This cluster averages highest on green areas and children's playgrounds, both of which are important to the client. Cluster 0 is also very diverse in terms of types of venues. Only in this cluster does each school have more than one venue of each location type on average. Geographically speaking, we see that while it is not entirely homogenous, cluster 0 is mainly located in the northern neighbourhoods of Tel Aviv. This finding seems consistent with the strong social- economic profile of these neighborhoods.
- As the visualization shows, **cluster 1** is located in the center of the city. This cluster is less suitable for my client, as it contains the smallest amount of green areas and playgrounds of the three clusters. However, cluster 1 does meet some of the client's needs, as it has a wide variety of pubs/bars, pizza parlours and ice cream parlours.
- **Cluster 2** is somewhat in the middle. It contains more green areas and playgrounds than cluster 1 but less than cluster 0. It has more daycares and dog gardens than cluster 0, but less pubs/bars, pizza and ice cream parlours. It is far less diverse in terms of location types. This cluster seems least suited for the client's needs, mainly because geographically, it is located mostly in the southern, eastern and northern margins of the city (the west is occupied by the Mediterranean Sea). These locations, most of which are far from the city center, make it difficult to manage without a private vehicle, as my client plans to do. Additionally, even assuming the client is willing to sacrifice the proximity to the city center (including such parameters as bars/pubs or family weekly tradition), in favor of green areas and playgrounds, it would make more sense to choose cluster 0 and not cluster 2.
- As we can see in the illustrations, there is a variance in the distribution of location types within the clusters. Therefore, I will examine the other clusters as well and will highlight one particularly diverse and interesting school in each cluster.

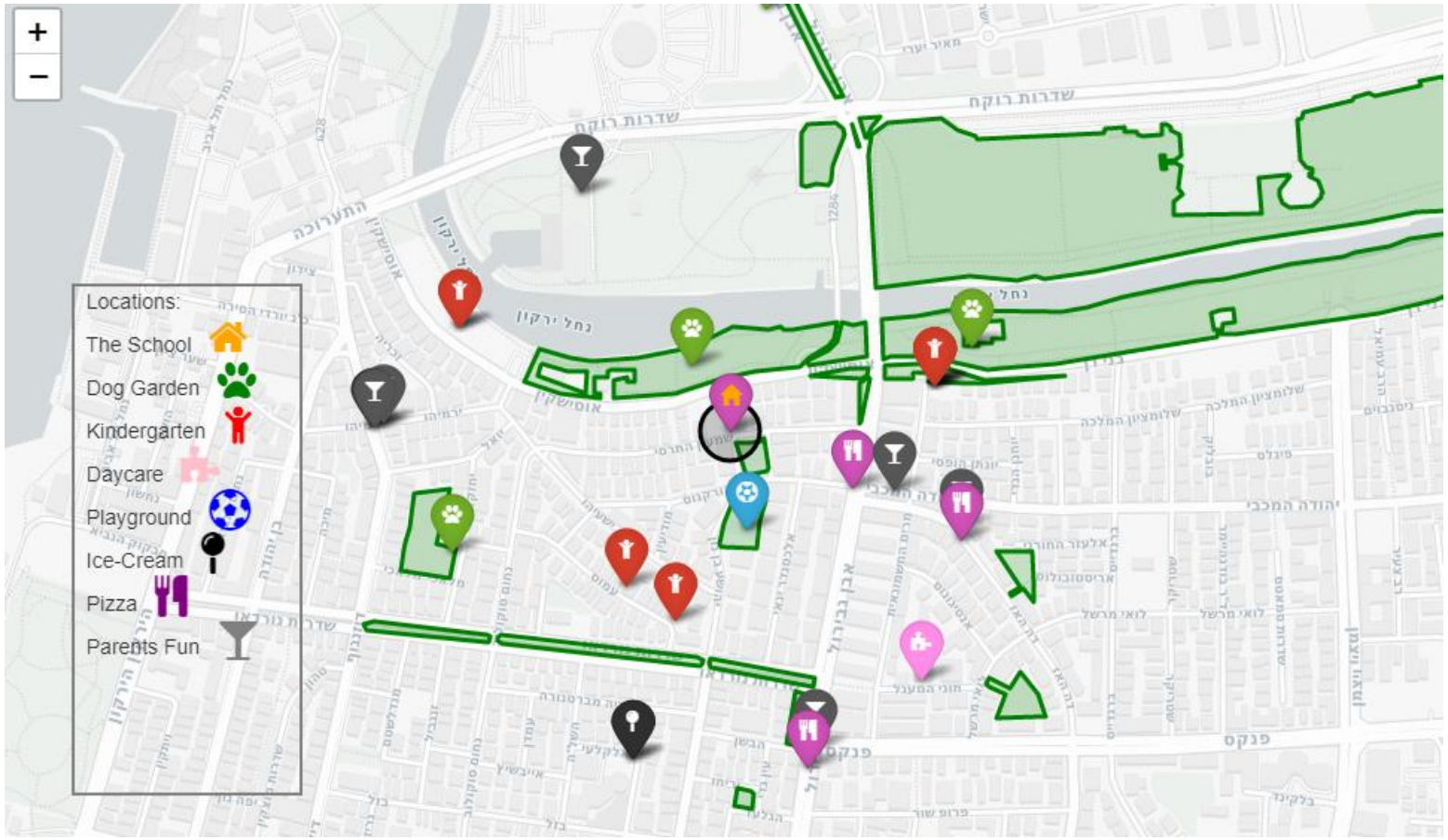
Highlighting candidate Schools in each cluster

Cluster 0: School - 599-300490



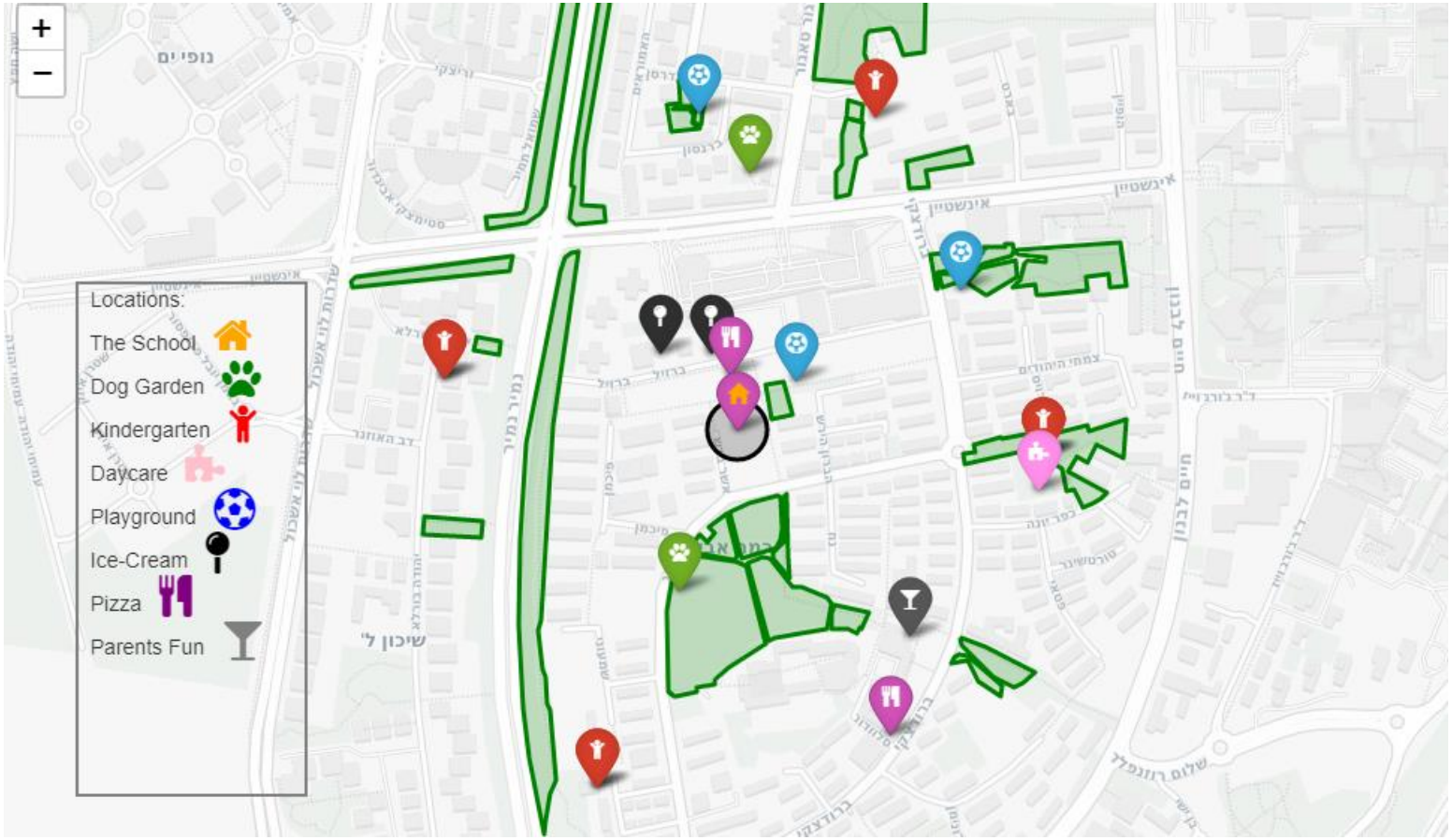
Highlighting candidate Schools in each cluster

Cluster 0: School - 599-304030



Highlighting candidate Schools in each cluster

Cluster 0: School - 599-302470



Highlighting candidate Schools in each cluster

Cluster 1: School - 599-309650

Highlighting candidate Schools in each cluster

Cluster 2: School - 599-302130

Conclusion

This project was devoted to exploring and segmenting elementary schools in Tel Aviv, based on their proximity to other venues that were of particular interest to the client - a friend of mine who has decided to move to the city with his family. The project provides clients, who wish to live near an elementary school, with data concerning other venues in the area, to inform their decision regarding where to rent an apartment. To do so, I collected the relevant data from a number of sources. I did some basic data analysis and some exploratory data visualization. I then segmented the schools based on the nearby location of interest, using K-means clustering algorithm, and found three clusters. This resulted in an interesting geographical differentiation between the city center (cluster 1), the Southern and Eastern neighbourhoods (cluster 2) and the Northern neighbourhoods (cluster 0). The latter cluster was found to be most suitable for the clients needs, as it is most diverse in terms of location types and contains the largest green areas and most playgrounds. This is consistent with the strong social-economic background of the Northern neighbourhoods of Tel Aviv, in which most of cluster 0 resides. In addition to the general characterization of the segments, I also highlighted three highly diverse schools in cluster 0 and one school in each of the other clusters.

In conclusion, it is noteworthy that at the request of the client, some data that could have also been relevant is currently not factored into the analysis, such as quality of school; cost of rent; other recreational facilities (restaurants, movie theaters, etc.).

Thus, there is room for expanding and improving the analysis by adding parameters that are currently outside the scope of the project. I believe that adding these dimensions to the clustering algorithm could have yielded very interesting results. Additionally, it is possible to combine weight for each location type in accordance with different preferences, so that the segmentation could give more weight to certain places. It is also possible to think of variations for the features. However, we must remember that possible variations are endless and that our main objective is to create a model that is valuable for solving a concrete problem, not a perfect one (no model is perfect!).