A Review of Accessible Education Resources in Taipei City

Prepared by To-Hong WU

3. Data Acquisition and Cleaning

3.1 GeoJSON of Taipei Neighborhoods and Their Geometric Centers

3.1.1 GeoJSON of Taipei Neighborhoods

This study needs a map containing all neighborhoods' coordinates to work on. Unfortunately such ready-for-use information couldn't be found during the study. However, there are a few GeoJSON and Shapely file of Taipei City. After searching and comparing different accuracy rate of GeoJSON files, this study uses the GeoJSON file of Taipei contributed by littlebtc, as the infrastructure which this study will be constructed upon. Using GeoJSON instead of the coordinates also brings the benefits of better visualization effect on the map.

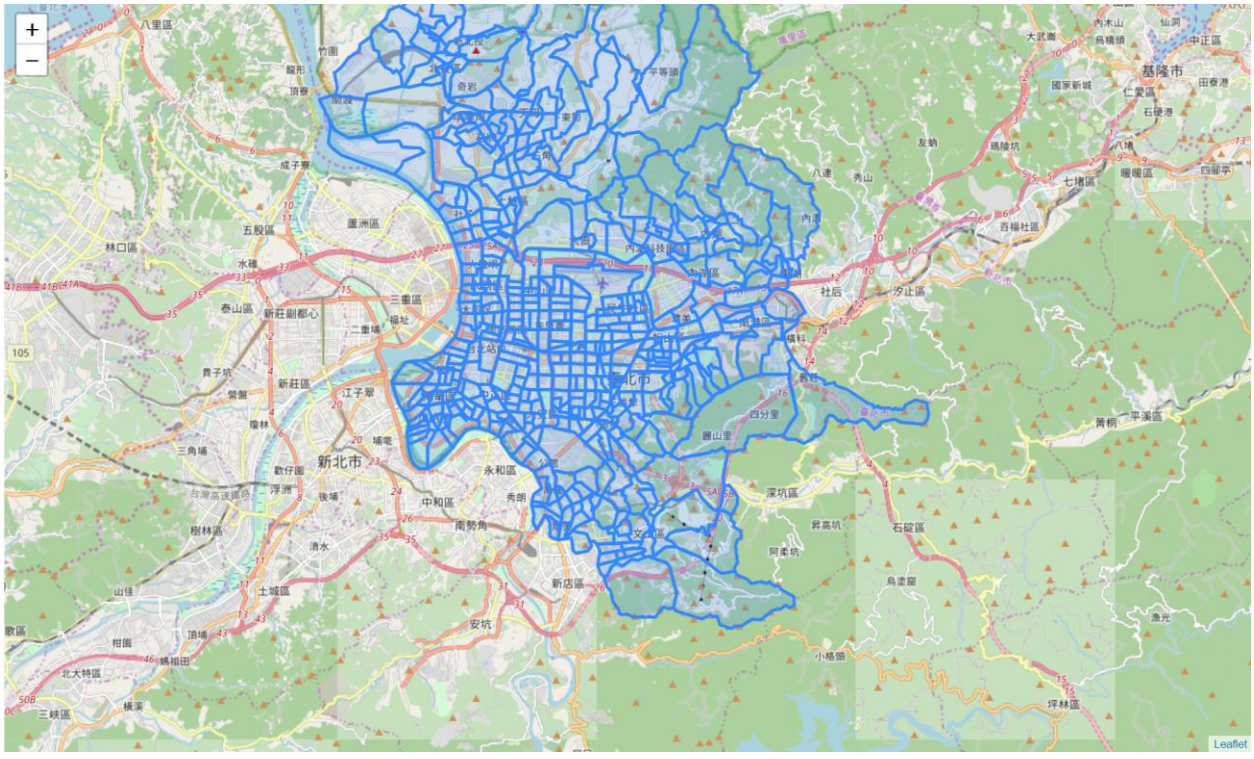

Figure 2. GeoJSON map view of Taipei City

3.1.2 Calculate the geometric centers of neighborhoods

The next task is to calculate the geometric center of each neighborhood polygon within the GeoJSON file, which I will use as the reference point to explore venues on Foursquare.

Though there are a few libraries working on different calculation of polygons, it seems that they are not created for GeoJSON. Thanks to brandonxiang who created a super useful project geojson-python-utils that solves the problem. With the centroid function, I can check these neighborhoods on the map.

It appears that the Chinese characters cannot *popup* properly on the folium map, and not too many discussion covers the encoding issues like this. Thanks to Yafei's Blog who solves this issue by bypassing it with html popups.

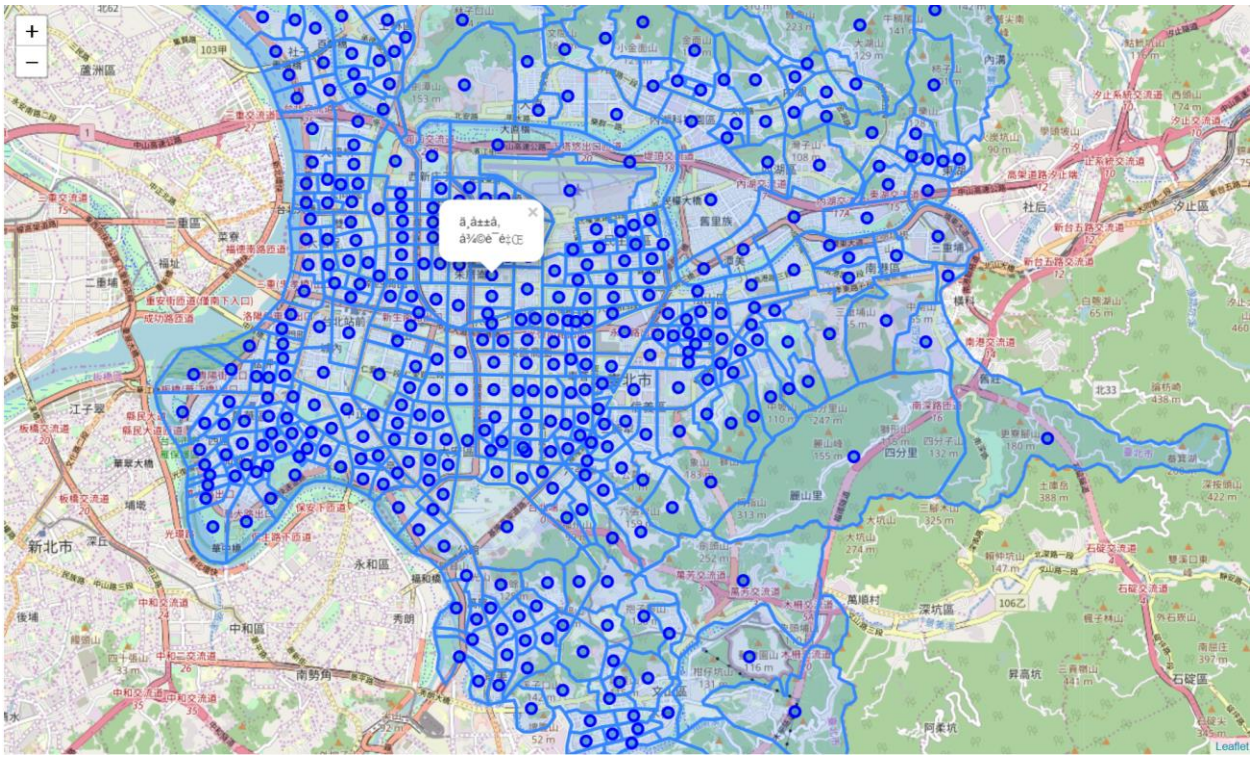

Figure 3. GeoJSON map with the geometric centers of neighborhoods with Chinese character issues

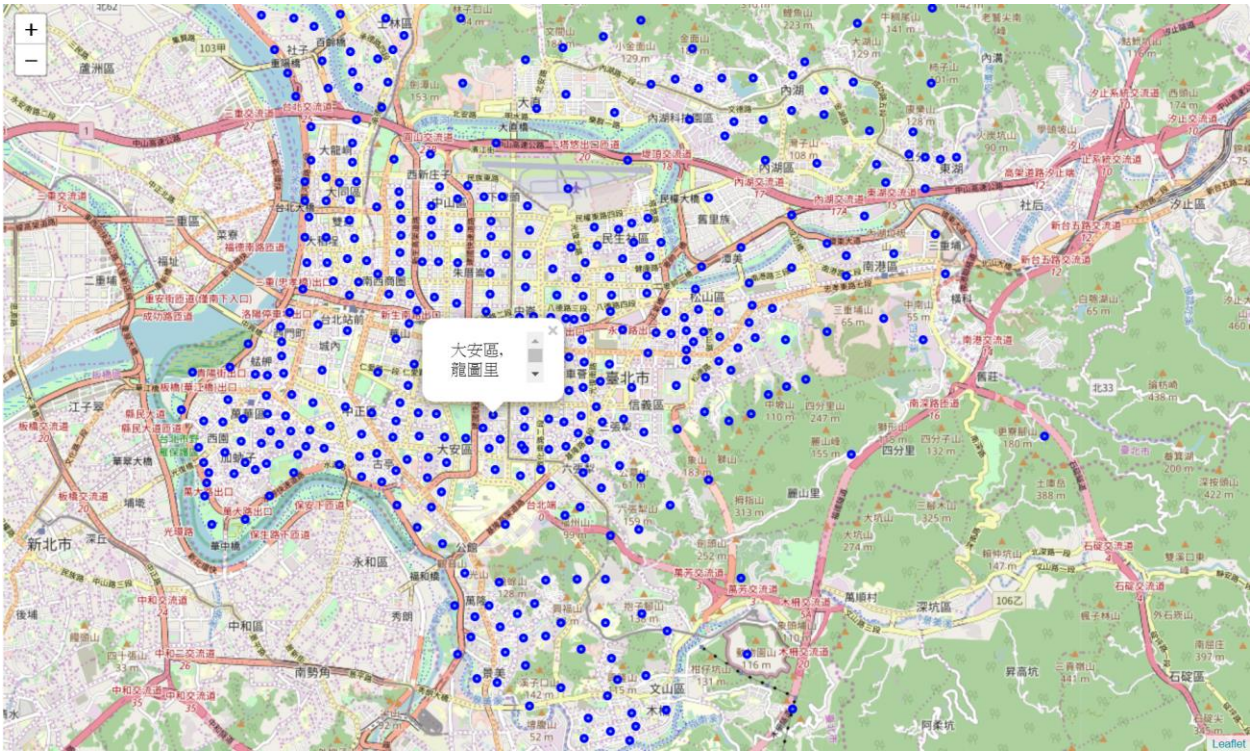

Figure 4. GeoJSON map with the geometric centers of neighborhoods (Chinese character issues fixed)

3.2 Taipei Household Income and Population under 15

3.2.1 Data cleaning

The statistics data is collected from the government website and stored on [github] for use.

The household income data is read from an excel file. Some of the titles of columns are renamed for future use, where "dist" and "neigh" represent district and neighborhood respectively. The population data was separated by ages and sex. A new column "under_15" is created to store the sum of the population. The two dataframes are then merged with the GeoJSON dataframe which contains the coordinates of all neighborhoods.

Table 1 shows the required information for later study, where dist_neigh, under_15 and income_median represent the neighborhood, population under 15, and the median value of household income.

When comparing with the original GeoJSON data where there are 456 neighborhoods, the new dataframe contains only 451. The five missing neighborhoods could be because of the lack of the socio-economic data, change of administration neighborhoods or any other reasons, which is beyond the scope of this study.

Table 1. List of neighborhoods, coordinates and its socio-economic data

	dist_neigh	Latitude	Longitude	under_15	income_median
0	松山區莊敬里	25.070698	121.563780	753	688
1	松山區東榮里	25.060350	121.558050	1476	919
2	松山區三民里	25.060024	121.562244	1045	800
3	松山區新益里	25.061795	121.567122	621	714
4	松山區富錦里	25.060958	121.564352	806	749
...	...	...	...	...	...
446	北投區關渡里	25.117710	121.467772	1679	641
447	北投區泉源里	25.153236	121.520774	247	584
448	北投區湖山里	25.149328	121.539538	188	614
449	北投區大屯里	25.160564	121.505817	166	570
450	北投區湖田里	25.179281	121.547548	101	589

451 rows × 5 columns

3.2.2 A quick look at the socio-economic data

Quickly checking the distribution of the population and household income of the neighborhoods, both figures show right-skewed distribution, meaning that they have longer tails to the right. In such case median instead of mean value could usually be a better choice to represent the group.

Below figures show how the neighborhoods distributed geometrically. When visualizing the data on map, it's hard not to notice that how similar these two figures are. Is it possible that household income and population under 15 are correlated for an unknown reason? Does wealthier family tend to have more kids than others? These topics are certainly deserved a more careful look in the future.

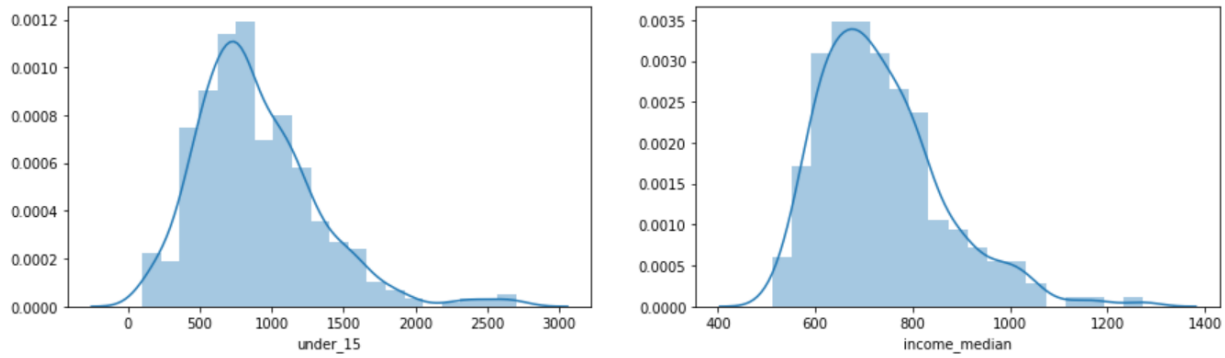

Figure 5. Distribution of neighborhoods for population under 15 and household income median

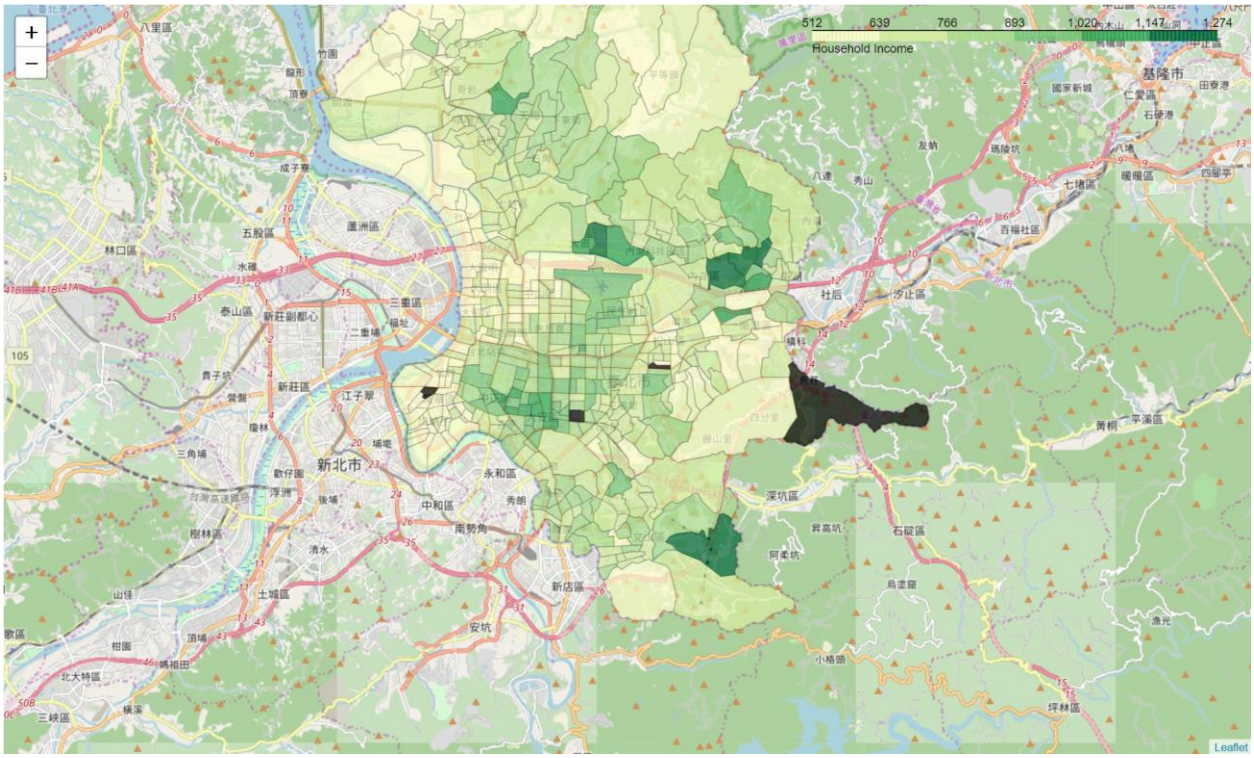

Figure 6. Neighborhoods distribution based on household income

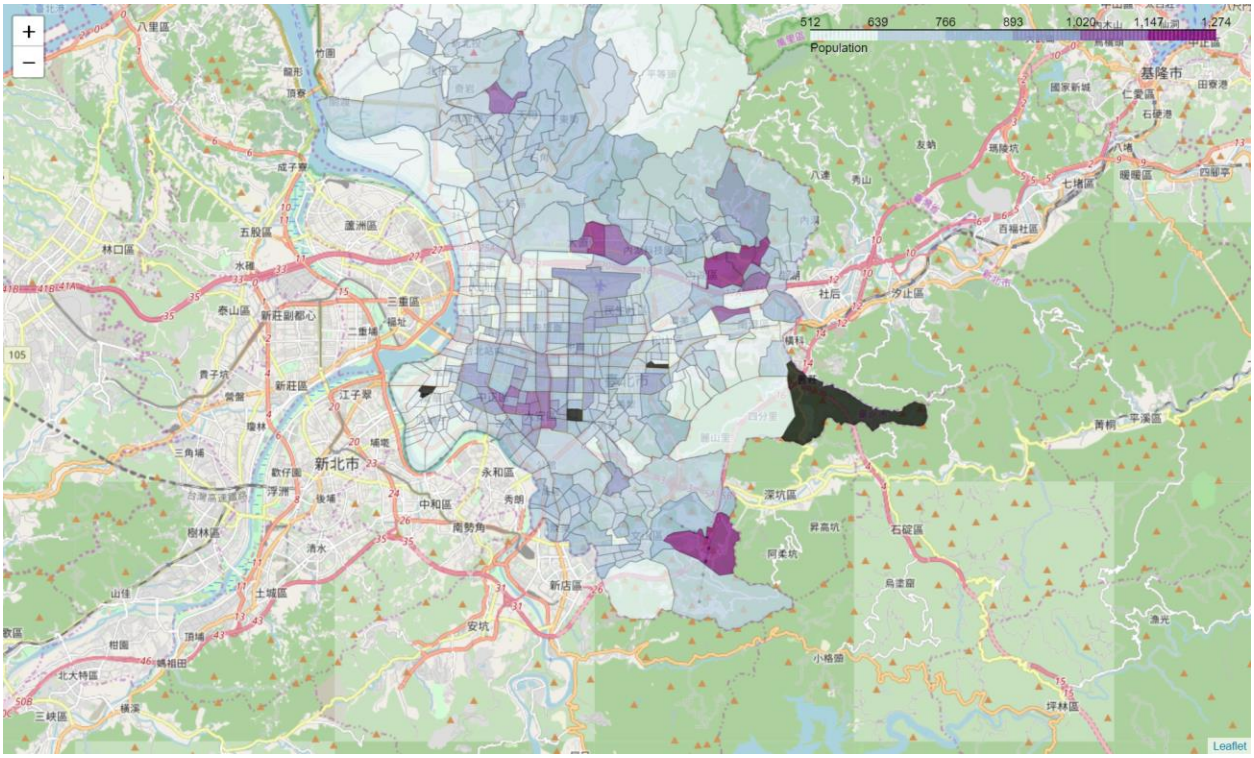

Figure 7. Neighborhoods distribution based on under-15 population

3.3 Education Resources by Foursquare

3.3.1 Data cleaning

To get to know what categories of venues we can find on Foursquares, first this study pulls out the entire categories list provided by Foursquare. However, it's not very easy to understand even if we convert the list to a dataframe because of its hierarchical structure.

I turned out to directly scrap it on Foursqaure webpage of Venues Categories, and find "Museum" subcategory under "Arts & Entertainment", and "School" and "Library" under "Professional & Other Places". These category IDs are then used to get the list of accessible education resources of each neighborhood, where more than 9,000 venues are successfully retrieved.

However the results still show a wide range of categories where many of them should not be identified as education resources (Figure 8). This study narrows down the category "School" to only include School, Elementary School, Middle School and Preschool, and "Museum" to all related kinds of museums. The list is significantly cut down by 2000 to 7225 venues.

Table 2 List of Education Resources on Foursquare

	categoryName	categoryId
0	School	4bf58dd8d48988d13b941735
1	Library	4bf58dd8d48988d12f941735
2	Museum	4bf58dd8d48988d181941735

```
array(['Community College', 'Library', 'School', 'Elementary School',
       'Middle School', 'Museum', 'Memorial Site', 'High School',
       'Comedy Club', 'Nursery School', 'Art Museum', 'Public Art',
       'Café', 'Music School', 'Factory', 'Preschool', 'History Museum',
       'College Academic Building', 'Indoor Play Area',
       'Performing Arts Venue', 'Trade School', 'Medical School',
       'Monument / Landmark', 'Language School', 'College Arts Building',
       'Music Venue', 'Temple', 'Park', 'Government Building',
       'Private School', 'University', 'Church', 'Swim School',
       'Historic Site', 'Office', 'Music Store', 'Science Museum',
       'Coffee Shop', 'Driving School', 'Planetarium'], dtype=object)
```

Figure 8. List of categories of venues retrieved from Foursquare

Table 3. Refined list of venues

	Neighbourhood	venue_id	venue_name	venue_latitude	venue_longitude	venue_category_name
2	松山區莊敬里	4eb253b0f5b94483884b11fa	Happy Marian	25.058909	121.557636	School
4	松山區莊敬里	4cd258364aa9f04dcdde56cb	臺北市立三民國民小學 Taipei Municipal SanMin Elementary ...	25.064342	121.565906	Elementary School
5	松山區莊敬里	4c81bbbe74d7b60cbe1d7ad8	台北市立民權國民小學 Taipei Municipal MinQuan Elementary...	25.062611	121.563132	Elementary School
6	松山區莊敬里	4c46fd491ddec928e9f19b32	臺北市立濱江國民中學 Taipei Municipal BinJiang Junior Hi...	25.079964	121.559820	Middle School
7	松山區莊敬里	4cdd91967e2e236af2b3791b	臺北市立民生國民中學 Taipei Municipal MinSheng Junior Hi...	25.059093	121.566732	Middle School

3.3.2 A quick look at the accessible education resources

Checking on how the neighborhoods access the education resources, one can easily find that the top 5 neighborhoods which have the most accessible resources are all in the old town, the southwestern part of Taipei City, while the last five are located in outer area.

However, there are 5 neighborhoods missing due to that they have zero access to education. These neighborhoods have been brought back by setting the numbers to zero, for later analysis.

Table 4. Numbers of accessible education resources of each neighborhood

v_cat	Neighbourhood	Library	Museum	School	edu_Total
45	中正區南門里	7	18	19	44
52	中正區愛國里	7	18	19	44
71	中正區龍福里	8	15	19	42
69	中正區黎明里	7	17	17	41
57	中正區東門里	8	16	16	40
...	...	...	...	...	...
148	內湖區金瑞里	0	0	1	1
179	北投區湖田里	0	0	1	1
115	內湖區內溝里	0	1	0	1
212	士林區公館里	0	0	1	1
156	北投區八仙里	0	0	1	1

446 rows × 5 columns

3.3.3 Check the unique venues and where they locate

There are more than 7,000 venues in the list, where most of them are duplicated simply because of the area covered by the radius of 1,500 meters. The search results are refined by removing duplicates and find 333 unique venues, of which 310 are located within only 204 neighborhoods in Taipei City.

It's not surprising that more than half neighborhoods have no education resources given that the land area of neighborhoods could be very small inside the old town.

Table 5. List of where unique venues locate

v_cat	venue_neigh	Library	Museum	School	Total
29	中正區黎明里	2	3	5	10
19	中正區建國里	2	3	1	6
17	中正區南門里	0	3	2	5
113	士林區臨溪里	1	3	1	5
106	士林區福佳里	2	2	0	4
...	...	...	...	...	...
89	南港區舊莊里	1	0	0	1
90	南港區萬福里	0	0	1	1
91	南港區西新里	0	0	1	1
11	中山區永安里	0	0	1	1
102	士林區永倫里	0	0	1	1

204 rows × 5 columns

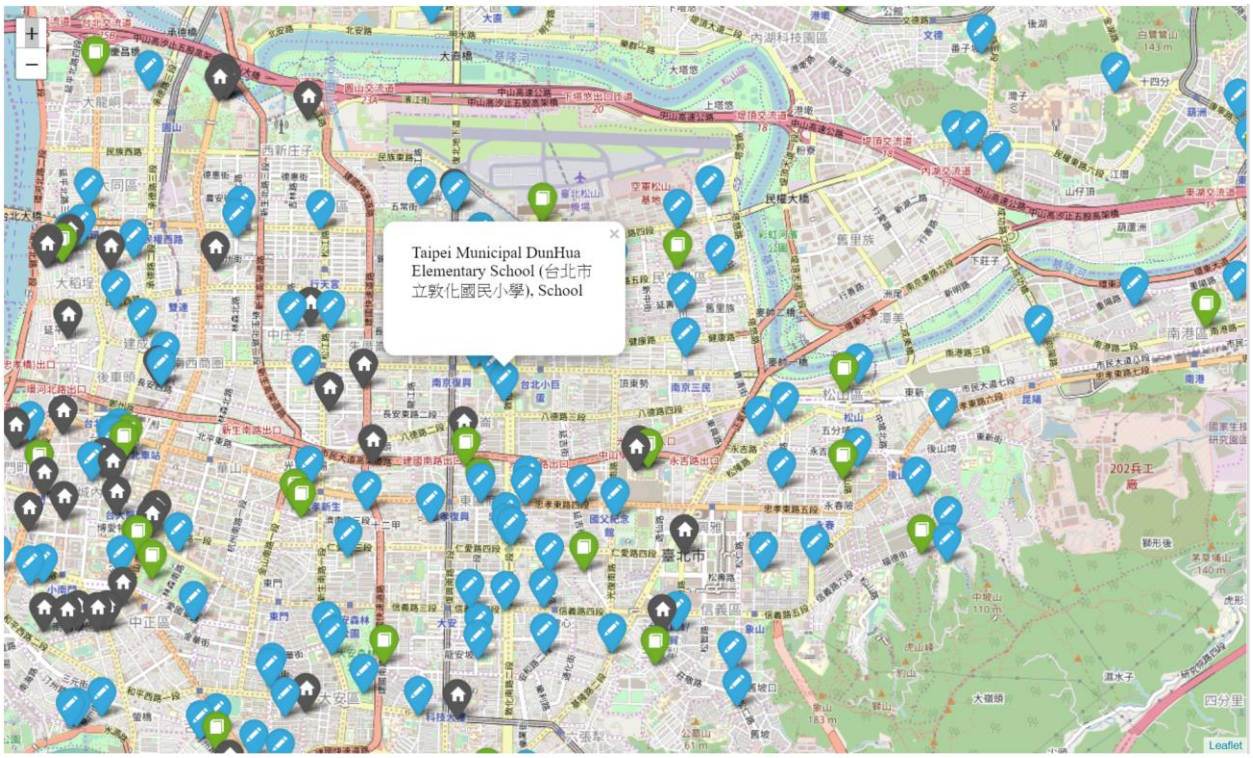

Figure 8. Map where the unique venues are located