neue fische: EDA project

Project: King County housing data

Author: Colin von Negenborn

Client: Amy Williams

- Mafiosi
- Needs average outskirt houses to hide from FBI
- Sells several central houses (top10%) over time

The data

Dataset on King County house prices

Source: neue fische

Range: 21k data points, years 2014 - 2015

Focus: For each house...

- Geolocation
- Size of lot and home
- Size of neighbours' lot and home

Additional data: King County ZIP geography

Goals

We are looking for a place to hide from the FBI...

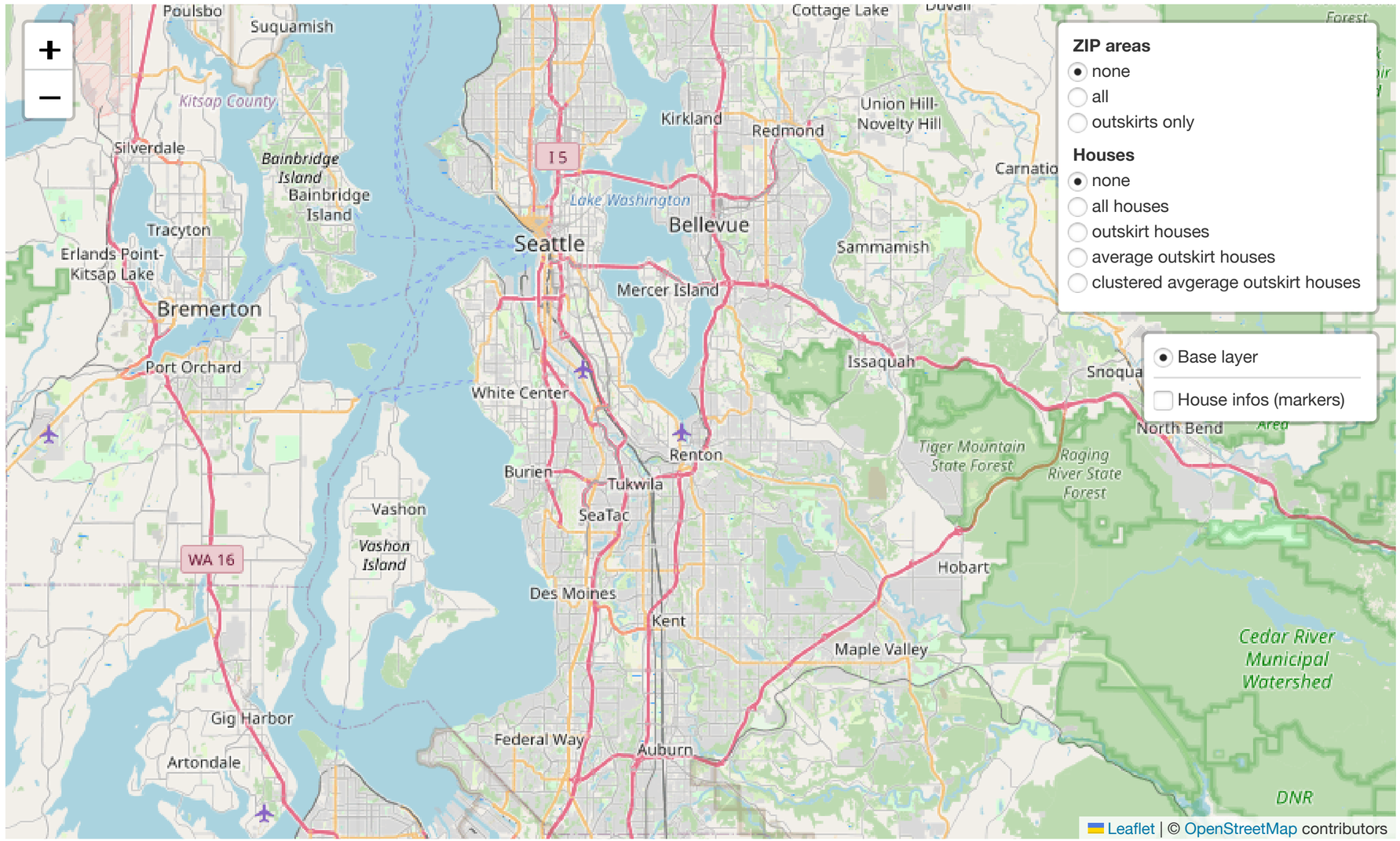
1. in the outskirts
2. in an average house
3. with other average houses nearby

Ideas/Hypotheses

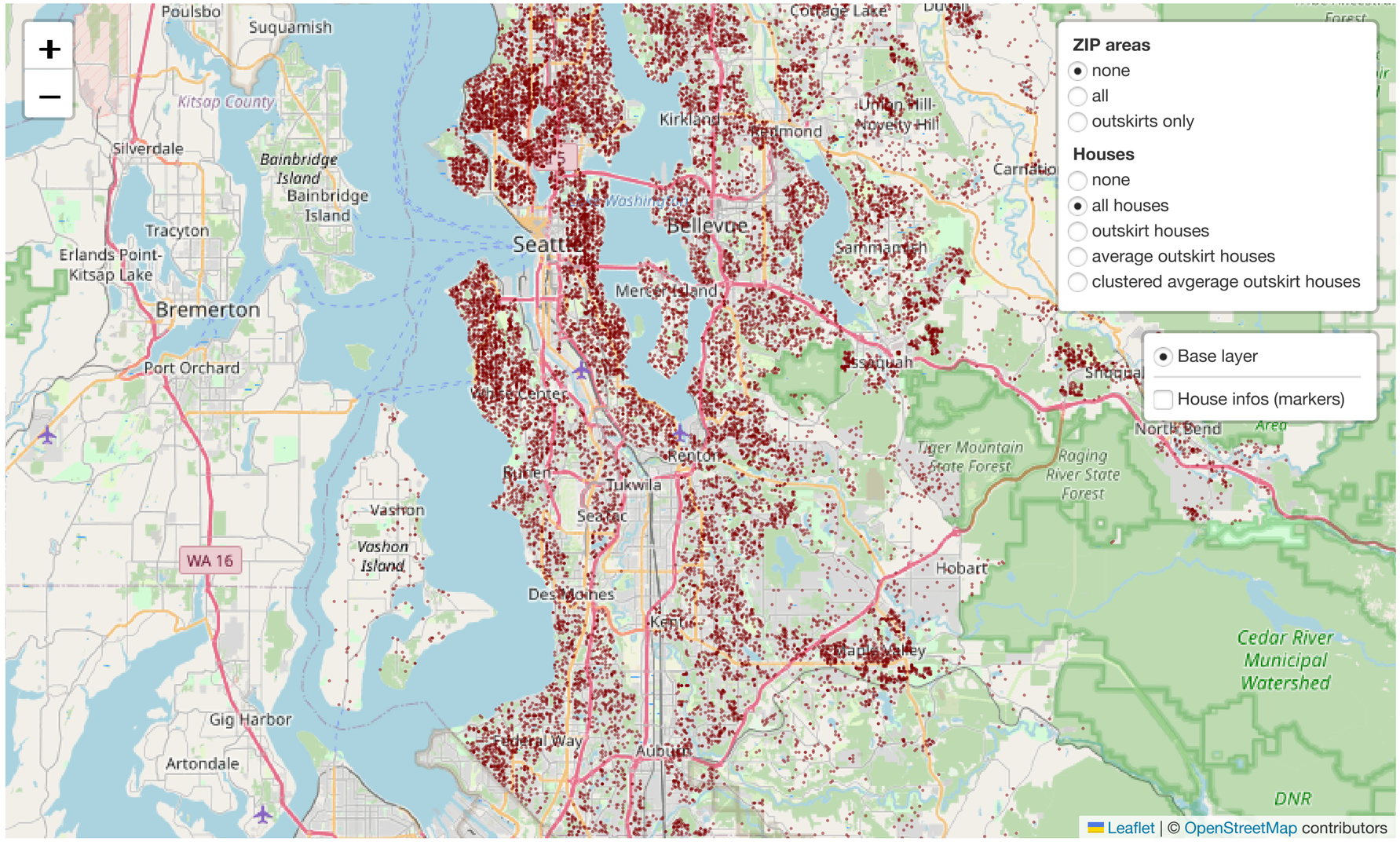
We can identify such houses via...

1. ZIP code
2. similarity to neighboring houses
3. clusters of similar houses

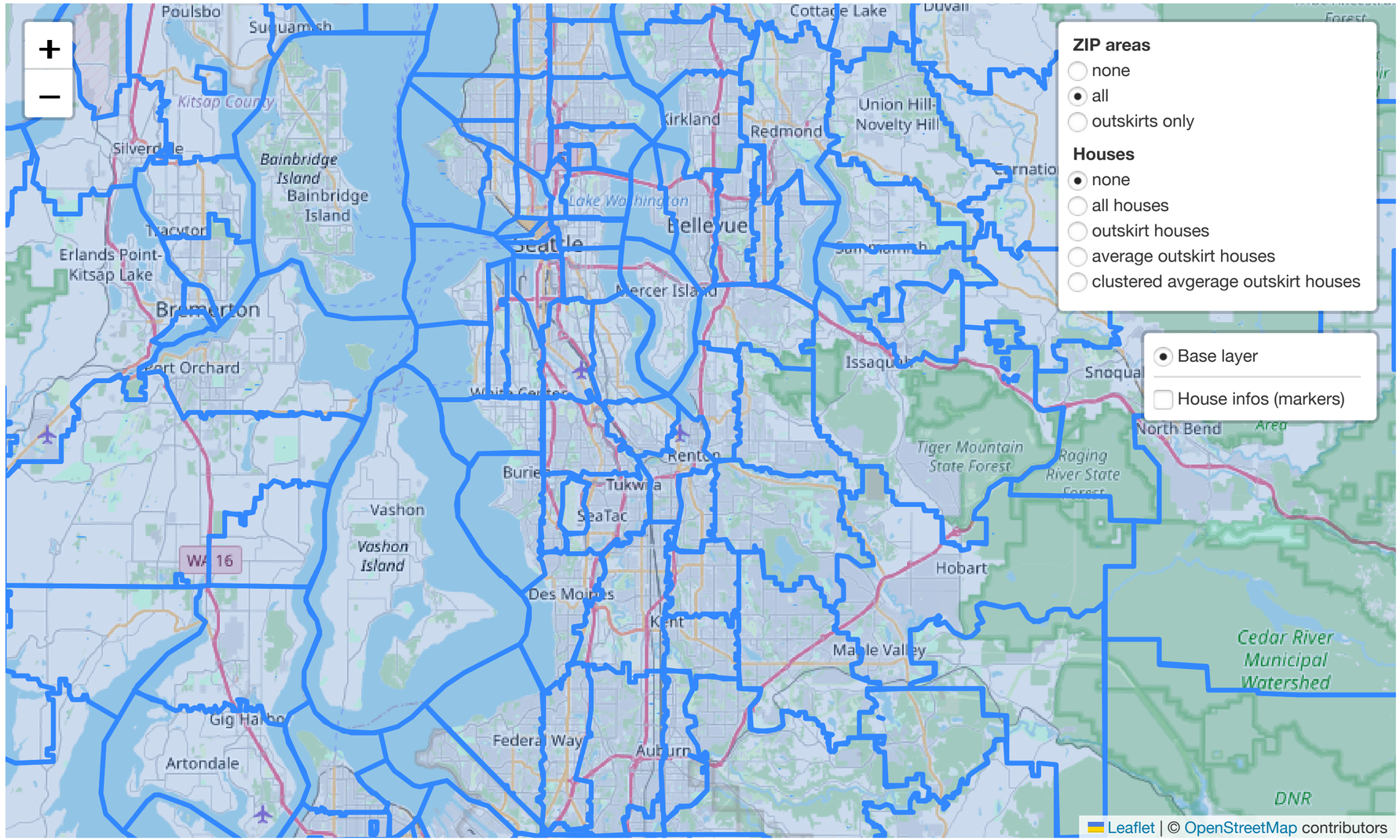
Technical details → Appendix


ZIP areas
none
all
outskirts only
Houses
none
all houses
outskirt houses
average outskirt houses
clustered avgerage outskirt houses
Base layer
House infos (markers)
Poulsbo
Suquamish
Kitsap County
Silverdale
Bainbridge Island
Bainbridge Island
Tracyton
Erlands Point-Kitsap Lake
Bremerton
Port Orchard
Cottage Lake
Kirkland
Redmond
Union Hill-Novelty Hill
I 5
Lake Washington
Seattle
Bellevue
Sammamish
Mercer Island
Issaquah
White Center
Renton
Burien
Tukwila
SeaTac
Vashon
Vashon Island
WA 16
Des Moines
Kent
Hobart
Maple Valley
Gig Harbor
Federal Way
Auburn
Artondale
Tiger Mountain State Forest
Raging River State Forest
North Bend
Cedar River Municipal Watershed
DNR
Leaflet | © OpenStreetMap contributors

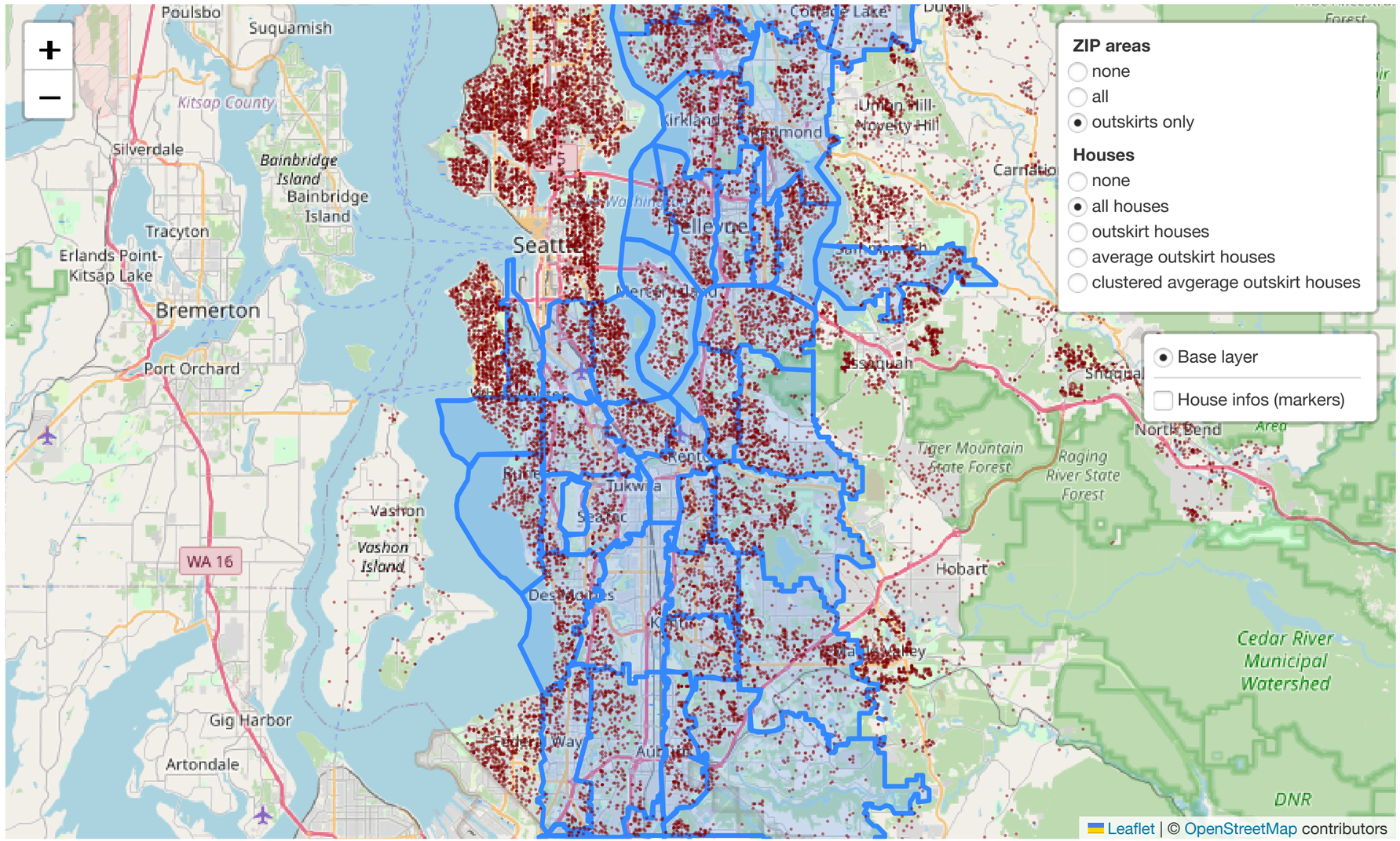
0. overview


ZIP areas
none
all
outskirts only
Houses
none
all houses
outskirt houses
average outskirt houses
clustered avgerage outskirt houses
Base layer
House infos (markers)
Poulsbo
Suquamish
Kitsap County
Silverdale
Bainbridge Island
Bainbridge Island
Tracyton
Erlands Point-Kitsap Lake
Bremerton
Port Orchard
Seattle
Kirkland
Redmond
Bellevue
Cottage Lake
Issaquah
North Bend
Tiger Mountain State Forest
Raging River State Forest
Renton
Tukwila
Seatac
Vashon
Vashon Island
WA 16
Des Moines
Kent
Hobart
Cedar River Municipal Watershed
Gig Harbor
Artondale
Federal Way
Auburn
DNR
Leaflet | © OpenStreetMap contributors

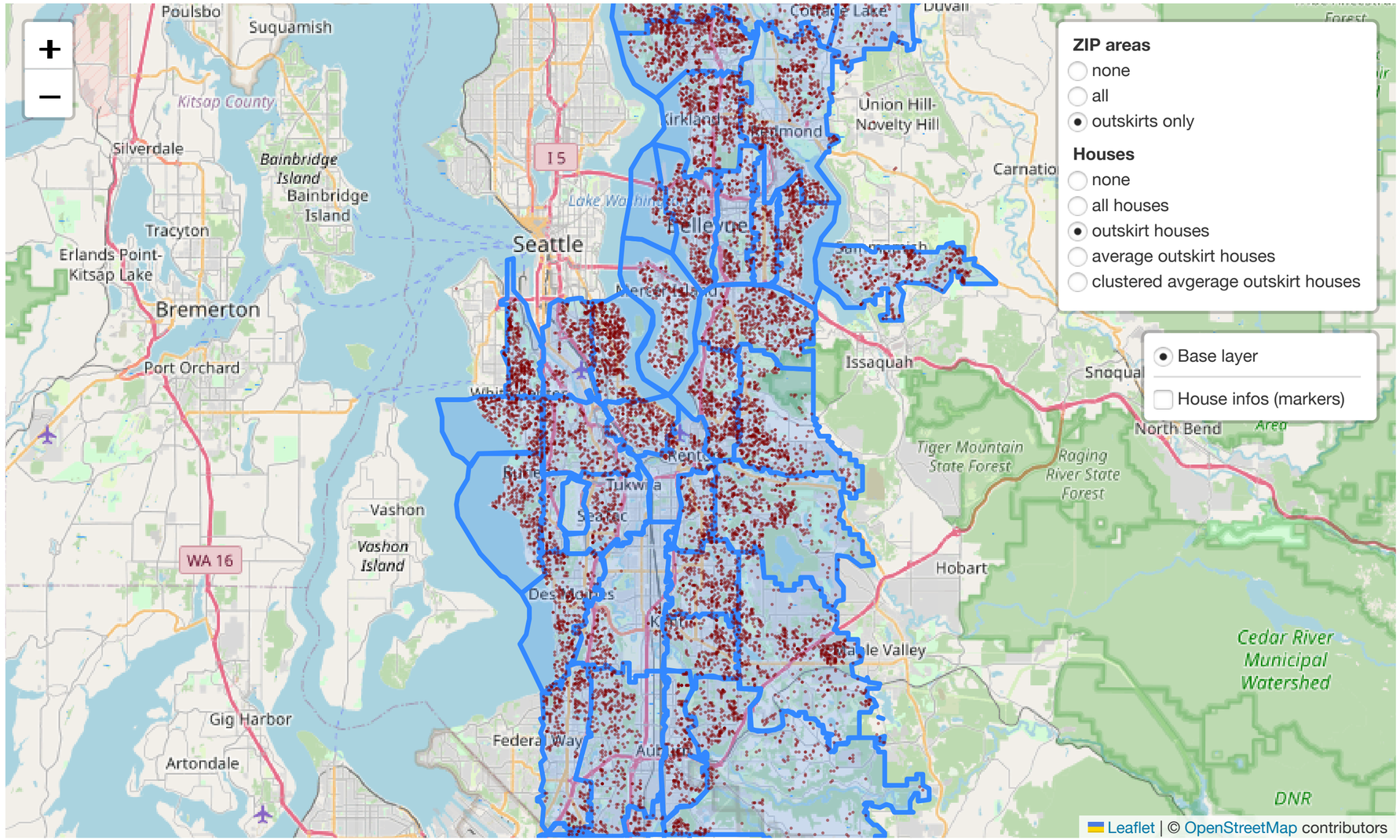
0. overview



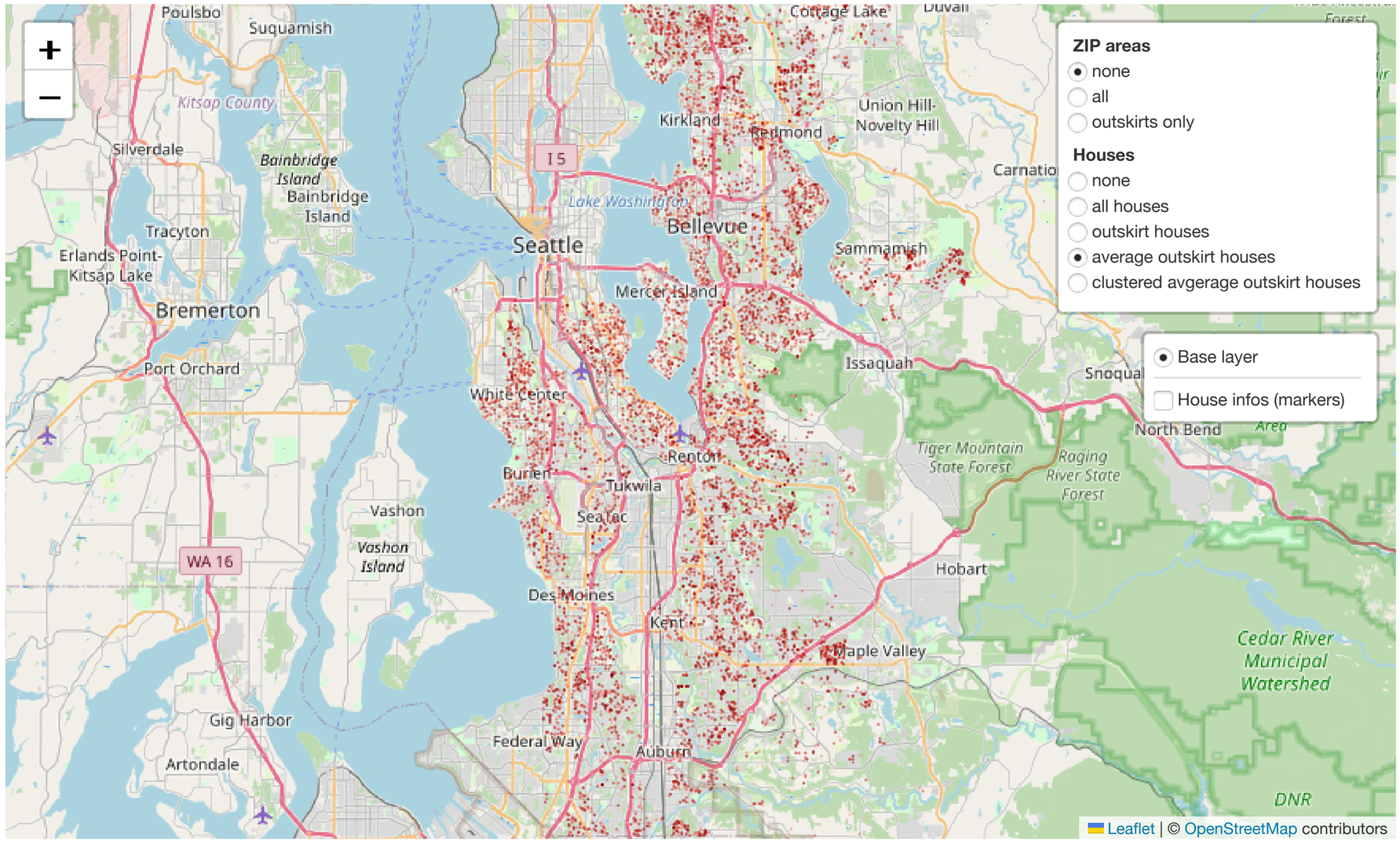
1. identify **outskirt** houses via **ZIP code**



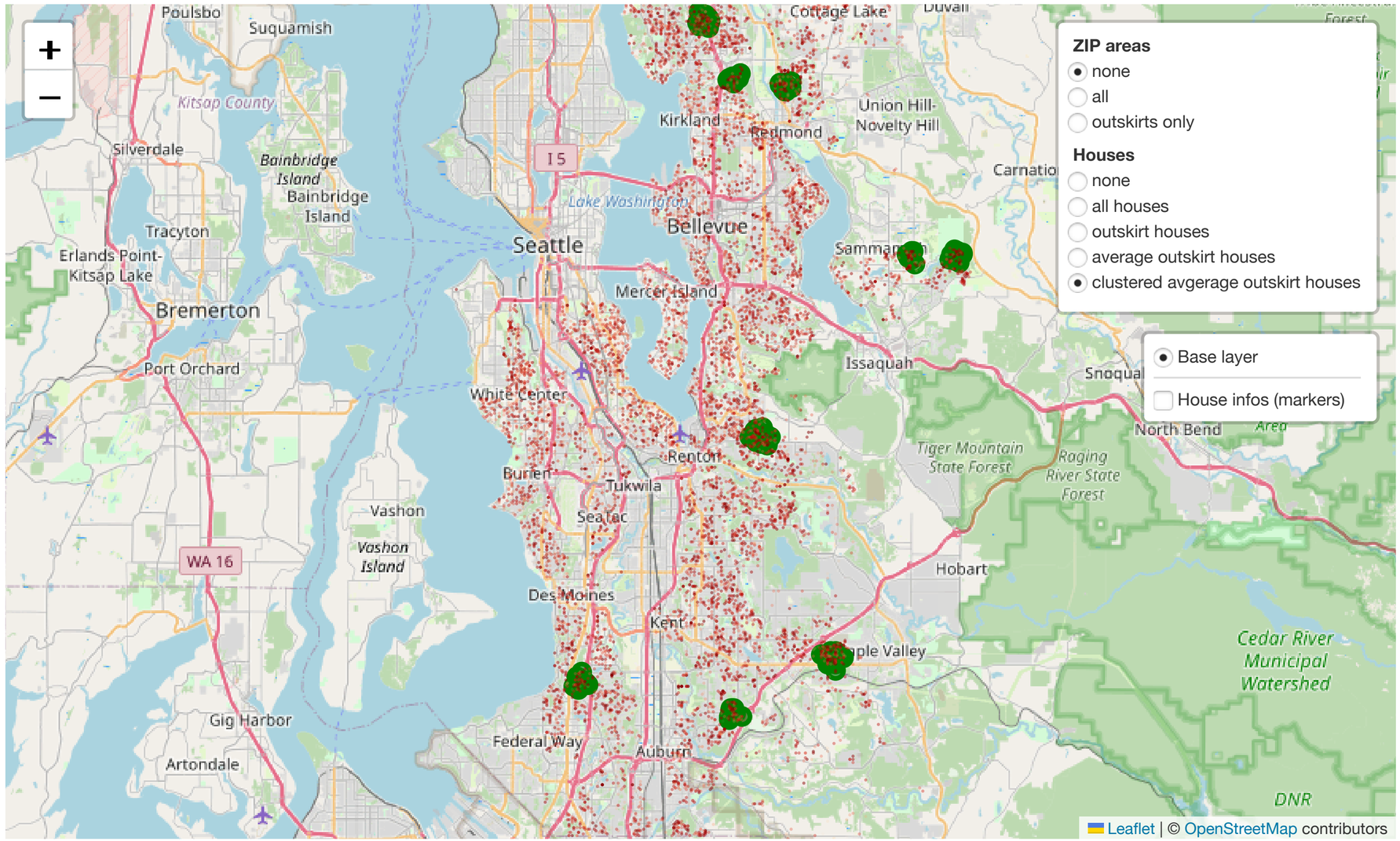
1. identify **outskirt** houses via **ZIP code**



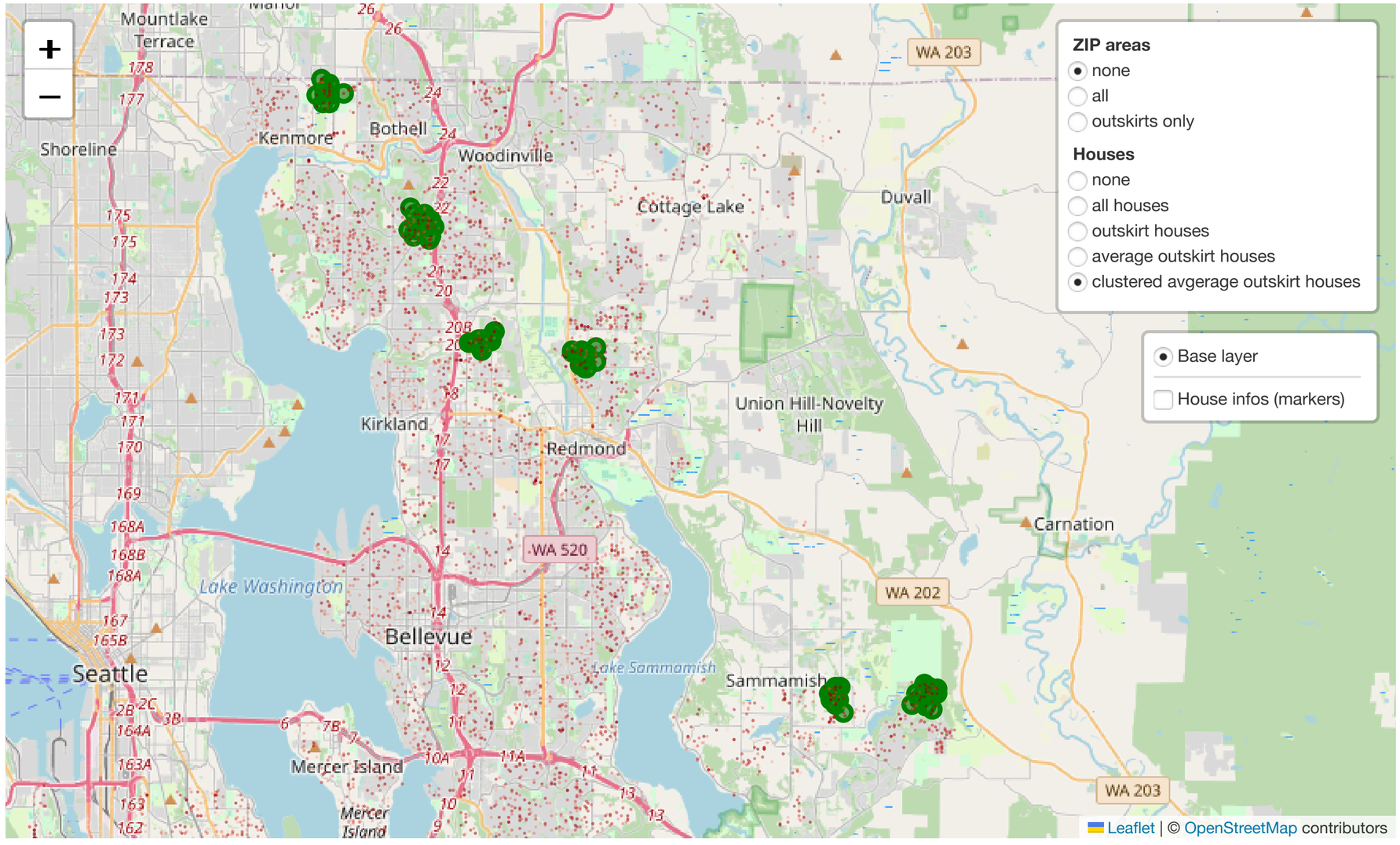
1. identify **outskirt** houses via **ZIP code**



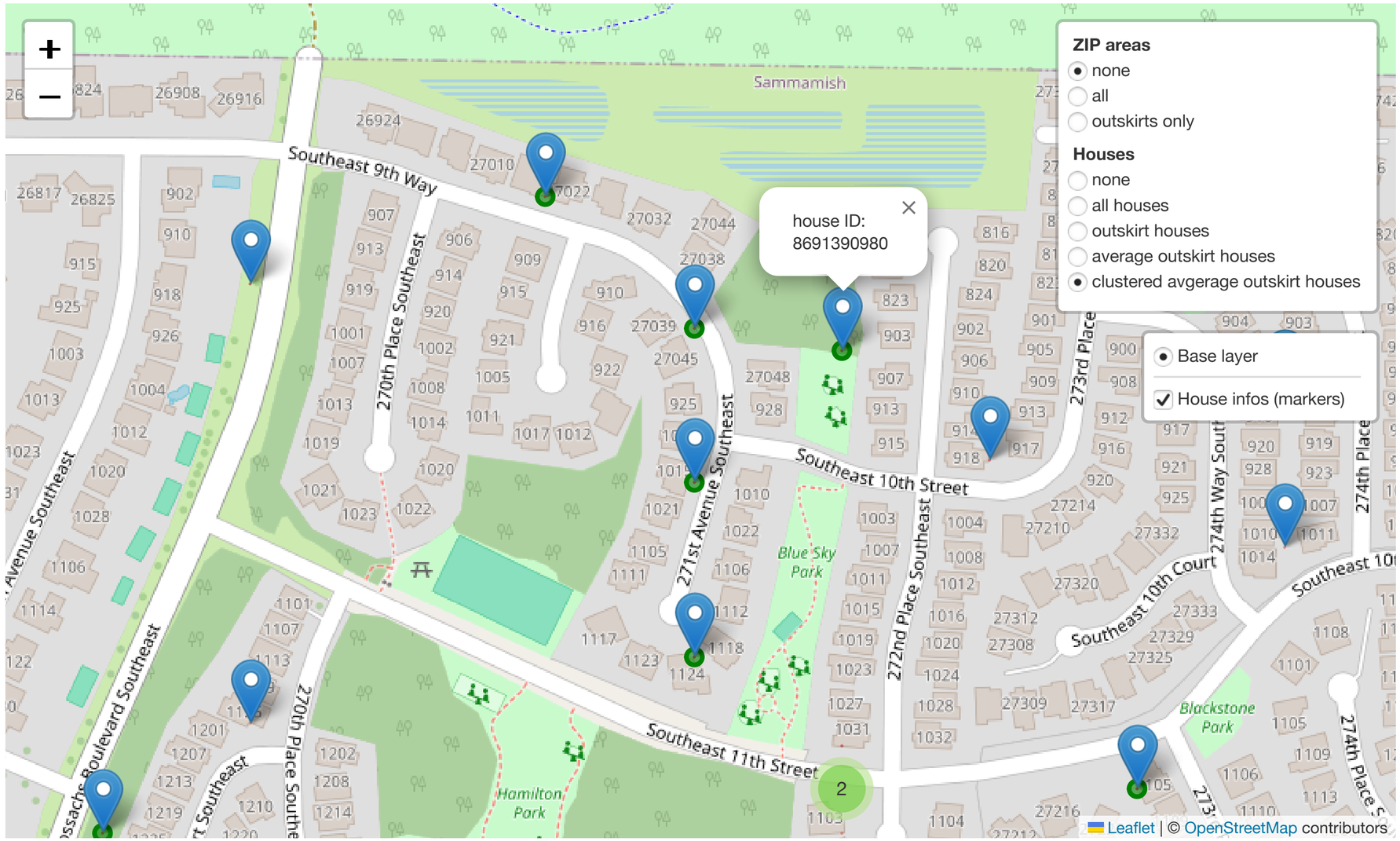
2. identify **average** houses via **similarity** to neighbours



3. identify **proximity** to other average houses via **clustering**



→ clusters of average outskirt houses (green) as optimal hideouts



→ clusters of average outskirt houses (green) as optimal hideouts

Insights

We can **identify average outskirt houses** as hiding places

Criteria: data on...

- ZIP code
- similarity to neighboring houses
- clusters of similar houses

Recommendations

1. Choose a hideout house in one of the clusters
2. Make an offer the owner cannot refuse

Thank you & nice try, FBI

Appendix (go back)

1. Deviation: With $s_{lot} =$ `sqft_lotsquare`, $s_{liv} =$ `sqft_livingsquare`, $s_{lot15} =$ `sqft_lot15`, and $s_{liv15} =$ `sqft_living15`, we define a house's **deviation** via its root mean square deviation:

$$d_i := \sqrt{\frac{1}{2}\left[\left(\frac{s_{lot} - s_{lot15}}{s_{lot}}\right)^2 + \left(\frac{s_{liv} - s_{liv15}}{s_{liv}}\right)^2\right]}$$

2. Outliers: We remove houses that are particularly different from their neighbours. We define such **outliers** as follows:

$$d_i > \mathrm{Q3} + 1.5 * \mathrm{IQR} = \mathrm{Q3} + 1.5 * (\mathrm{Q3} - \mathrm{Q1})$$

3. Similarity index: From the remaining houses, we create similarity index. Defining the highest deviation $\bar{d} = \max_i d_i$, the **similarity index** is defined via:

$$s_i = \frac{\bar{d} - d_i}{\bar{d}} \in [0, 1]$$

4. Clusters: We conduct spatial clustering via the `DBSCAN` algorithm (density-based spatial clustering of applications with noise) from the `scikit-learn` package. We set `eps` (maximum distance between cluster points to $(1/6371)/2$ in haversine metric, corresponding to about 1/2 km, and we set the minimum cluster sample size to 12.

Future work

Add data on crimes and offenses from King County Sheriff's Office (Link), in particular the spatial distribution of offenses. This way, we can identify areas of high crime/offense rates, which allow our client to have a shorter commute to work.

Furthermore, we could look at optimal means to fund her business. In particular, she sells "central (top 10%) houses over time". A first look at the data did not indicate that there were months or weekdays were selling is particularly profitable for this subset of houses, though.

Finally, we may want to exclude clusters that are close to police stations.