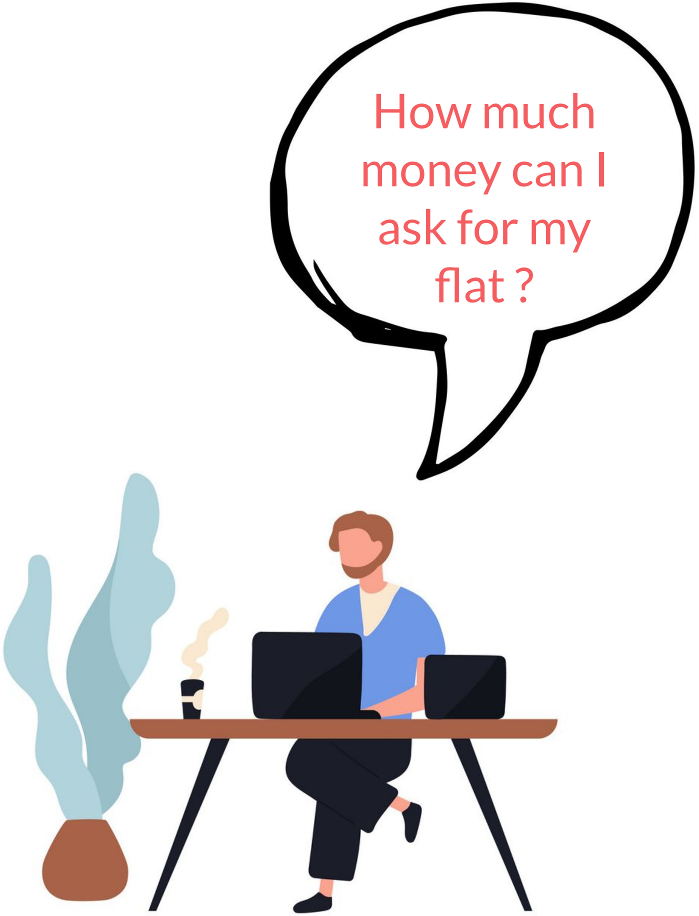
Airbnb price prediction from an host point of view

Team 27 : Lugeon Sylvain, Furter Samuel, Mosser Paul, Hartmann Florian

EPFL

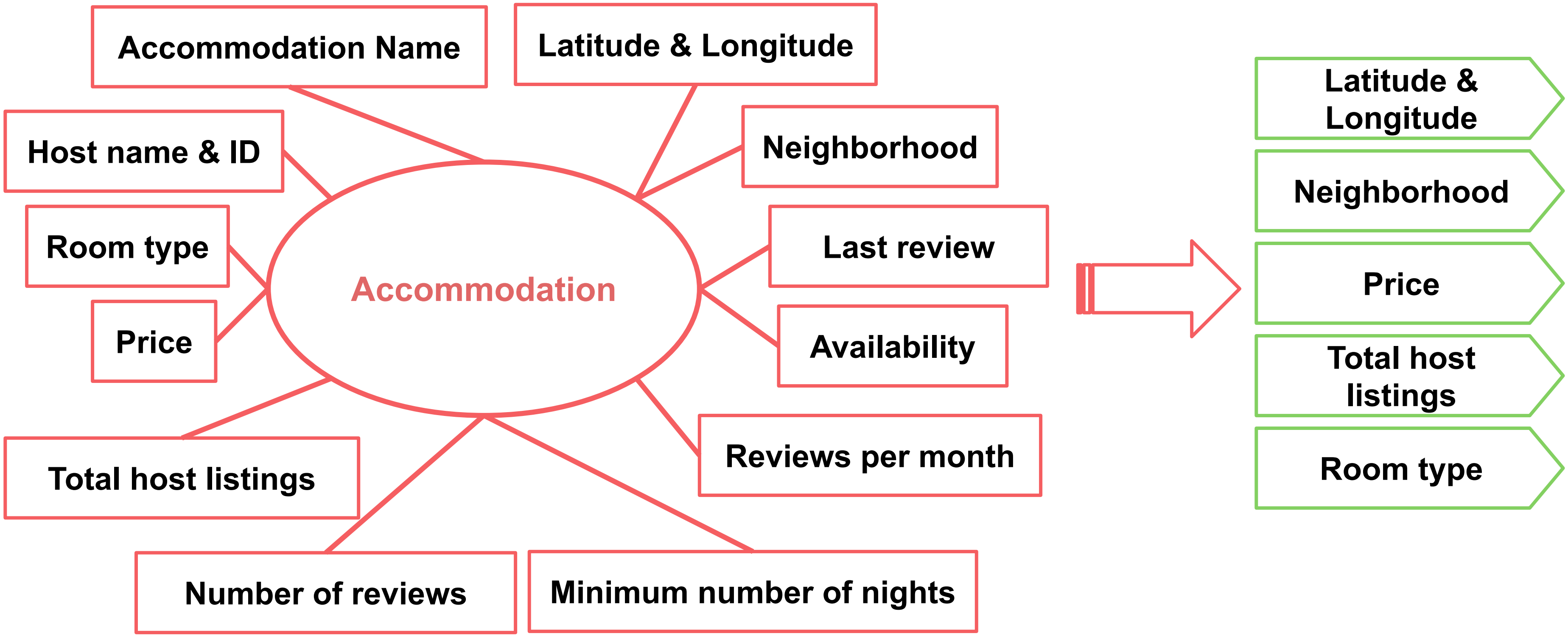
Airbnb - Users

Travellers



Accommodation provider

Features selection



Cleaning and preprocessing

WGS84 coordinates : latitude & longitude → **UTM coordinates**

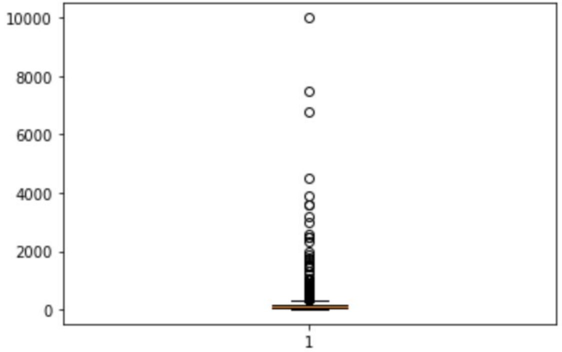
\+ Describe positions as plane coordinates
Allow visualization of geographical structure of accomodations

− Small loss in accuracy

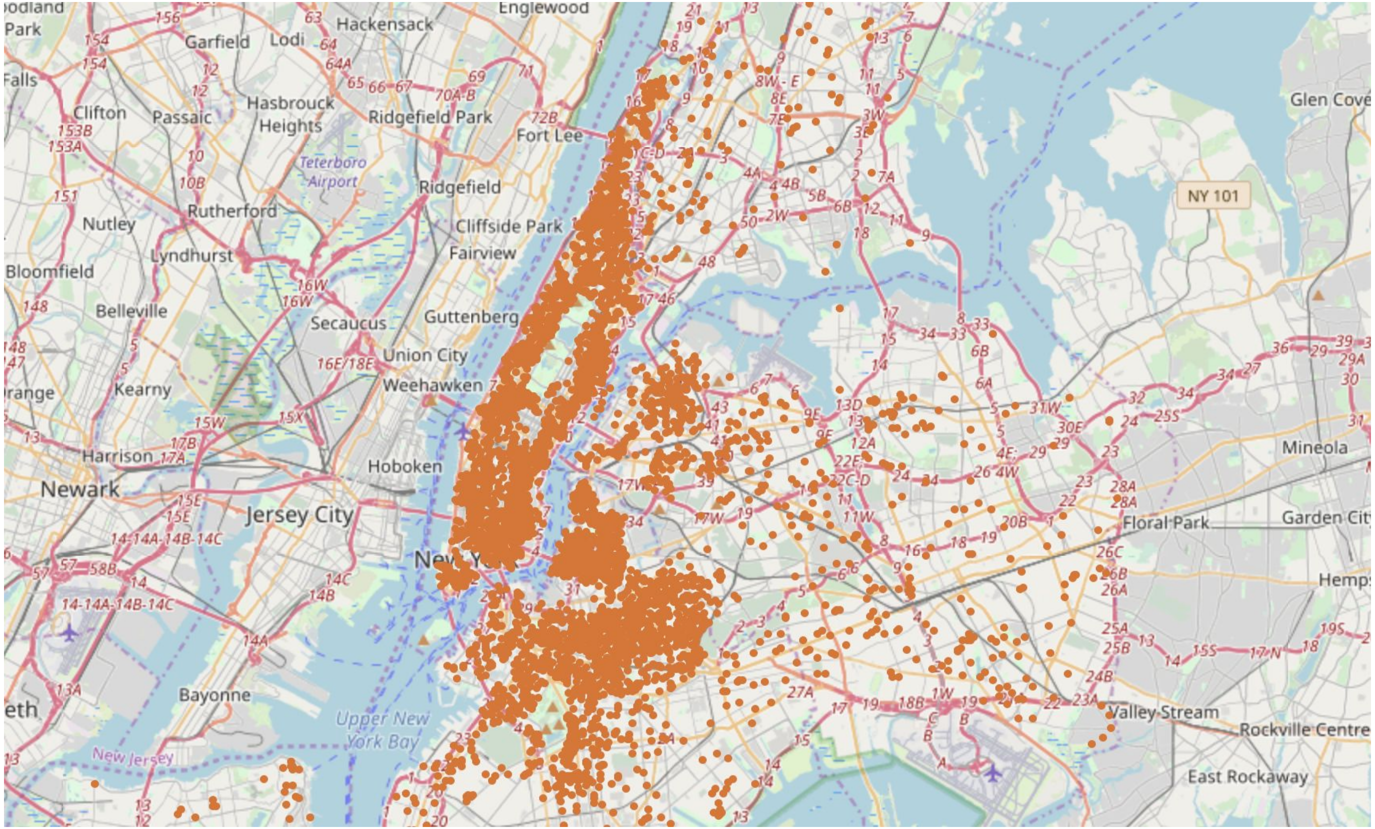
Random sampling : 5000 accomodations out of the 48'895 in the original dataset

Removing outliers : Outliers not appreciated when using ML models

→ Keep the accommodations with a price below the 0.95 quartile of the distribution



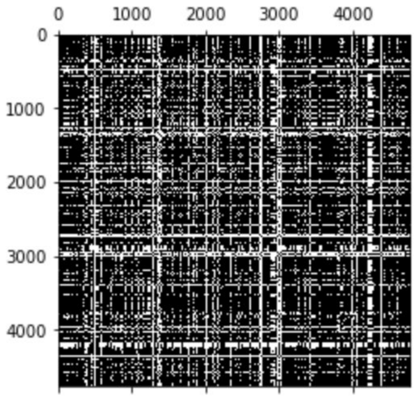
Price distribution

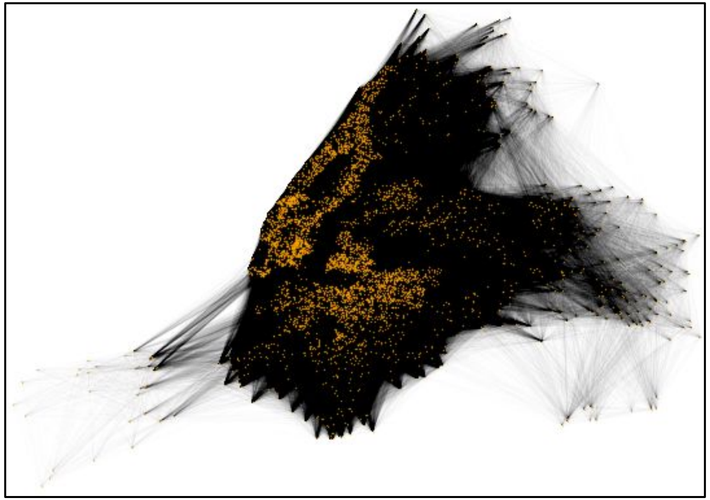


New York Airbnbs after sampling (5000 points)

Exploration - Graph construction

1) RBF Kernel to compute the adjacency matrix
   a) σ = k * mean
   b) ε chosen to sparsify the matrix



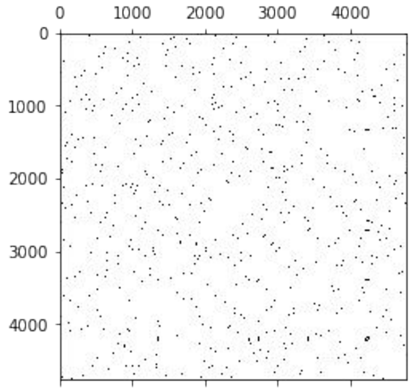


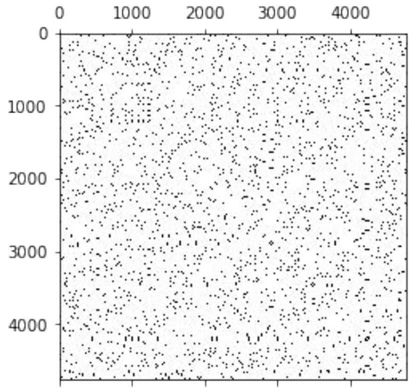
Graph obtained with RBF Kernel

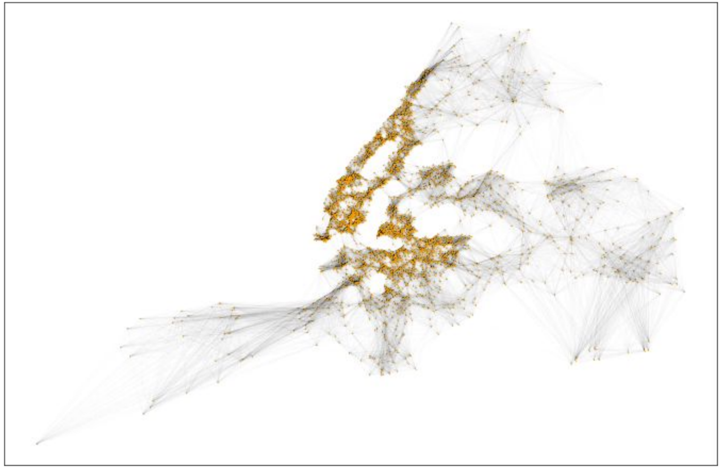
➡ **Goal** : have a connected graph and a sparse matrix

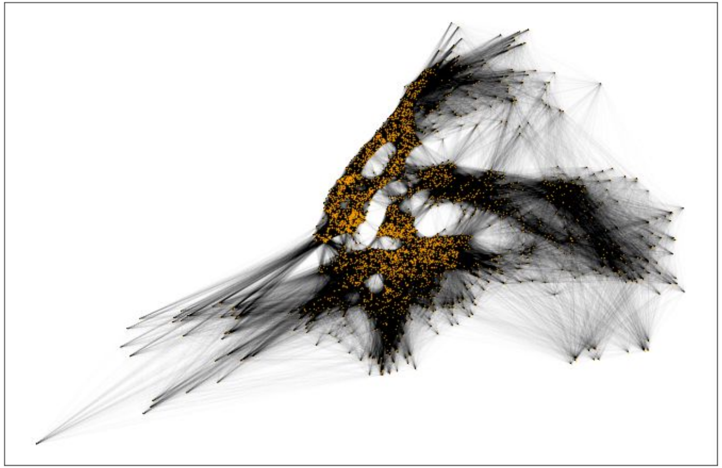
Exploration

2) kNN to compute the adjacency matrix
    a) Compute with k=50, k=200 and k=400



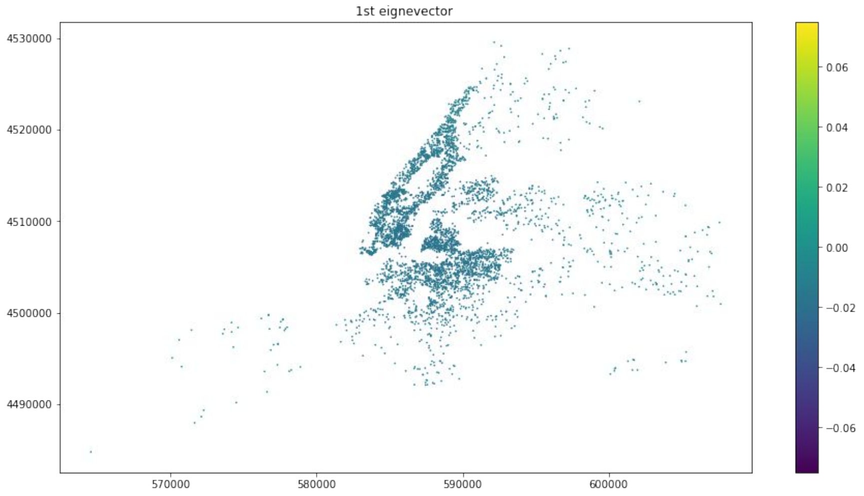


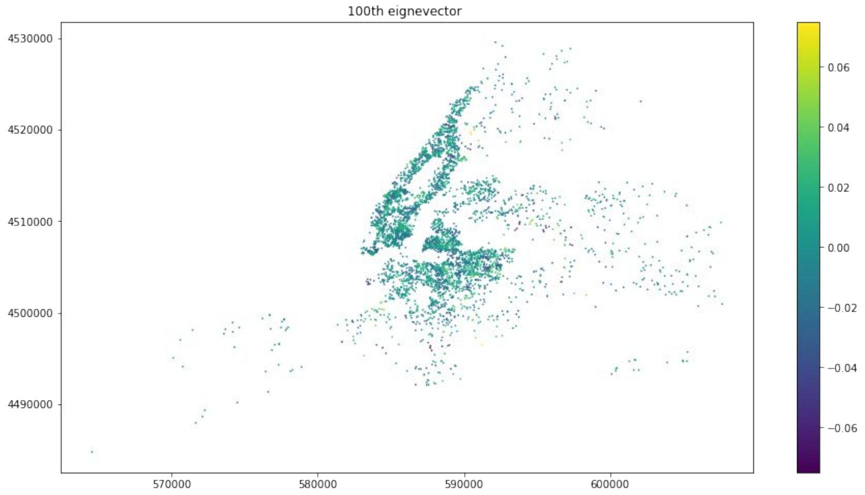




Exploration - Spectral components

Spectral decomposition of the graph's laplacians (normalized laplacians)

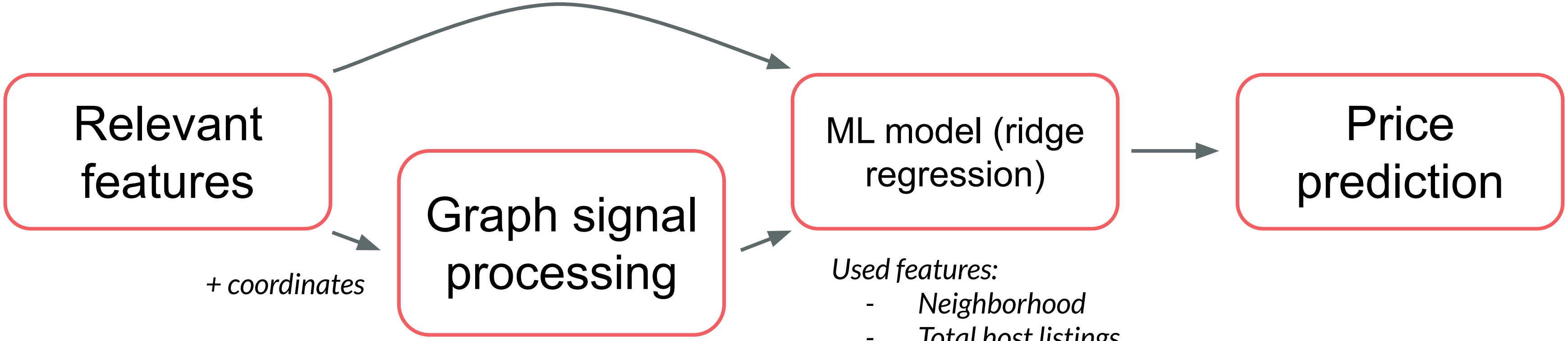




Difference of smoothness : the first eigenvector is smoother than the 100th eigenvector

Exploitation

Process we followed to obtain the **price prediction :**



Used features:

- *Neighborhood*
- *Total host listings*
- *Room type*
- *Price*

Signal filtering

Main idea : Filter the features signals to gives less weight to large eigenvalues

Assumption : Airbnbs that are geographically close should be close in price too

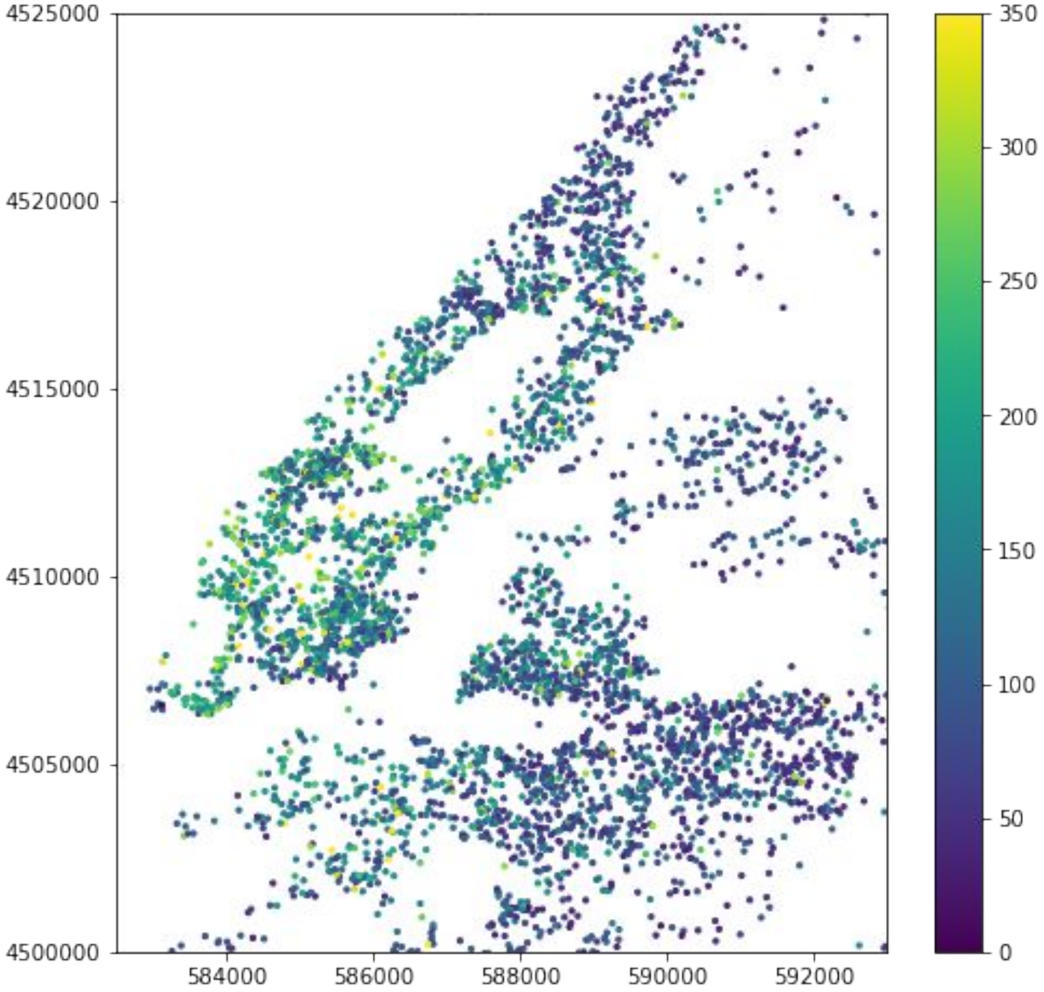
We used the Tikhonov regularization filter seen in course (but slightly different):

tk(e) = 1 / (1 + α * e) with **α = c * 0.99 * e_max** (large c gives less weight to high eigenvalues)

Why use both ridge regression on top of Tikhonov regularization ?

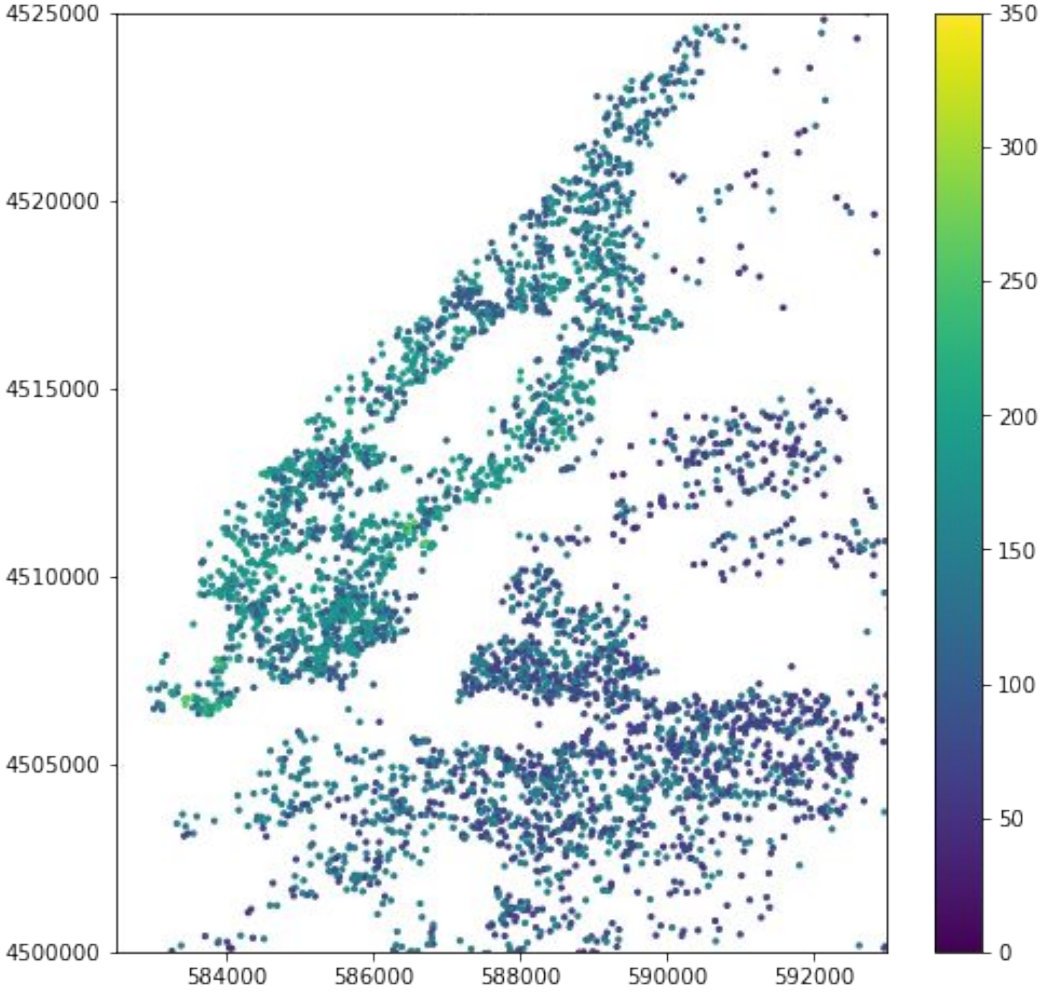
- ➢ Needed to use the same models with and without filtering
- ➢ Tikhonov regularization allows *larger kind* of regularization
- ➢ Tikhonov uses the structure of the graph and the location of the accommodation (where our simple ridge regression model doesn't)

Groundtruth



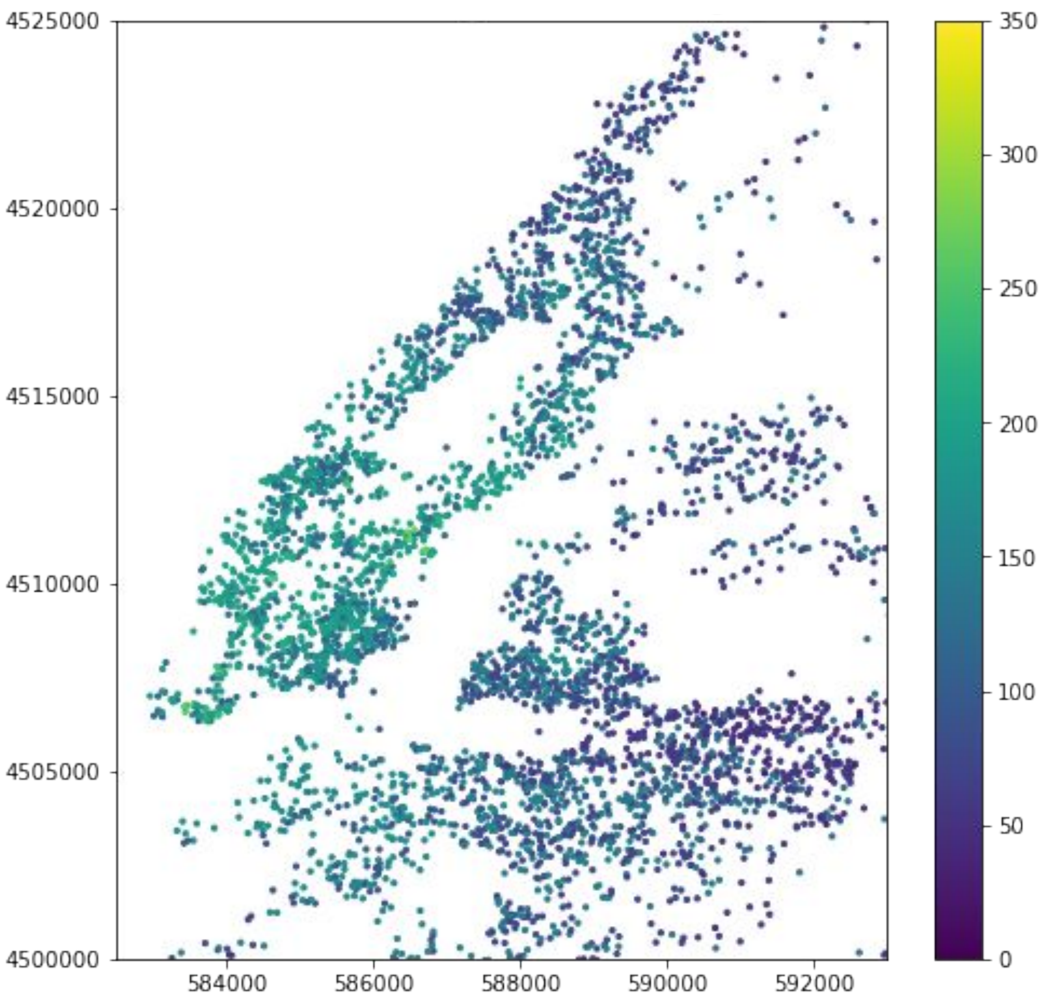
Ridge regression only

MSE : 2950.83



Using graph structure

kNN 200
MSE : 2686.21



Comparison

Model	MSE	MSE using graph structure	Improvement
RBF kernel	2950.83	2864.68	2.92%
KNN-50	2950.83	2801.92	5.05%
KNN-200	2950.83	2686.21	8.97%
KNN-400	2950.83	2705.95	8.30%
KNN-200 (20'000 s.)	2950.90	2806.96	4.88%

Improvements

- Difficult to predict the price with our set of features
- User interface to enter your flat’s data
- Predict price range instead of exact price (gives more freedom to the user)
- Applicable to other cities