What is the best neighbourhood to live in as a student at Imperial College London?

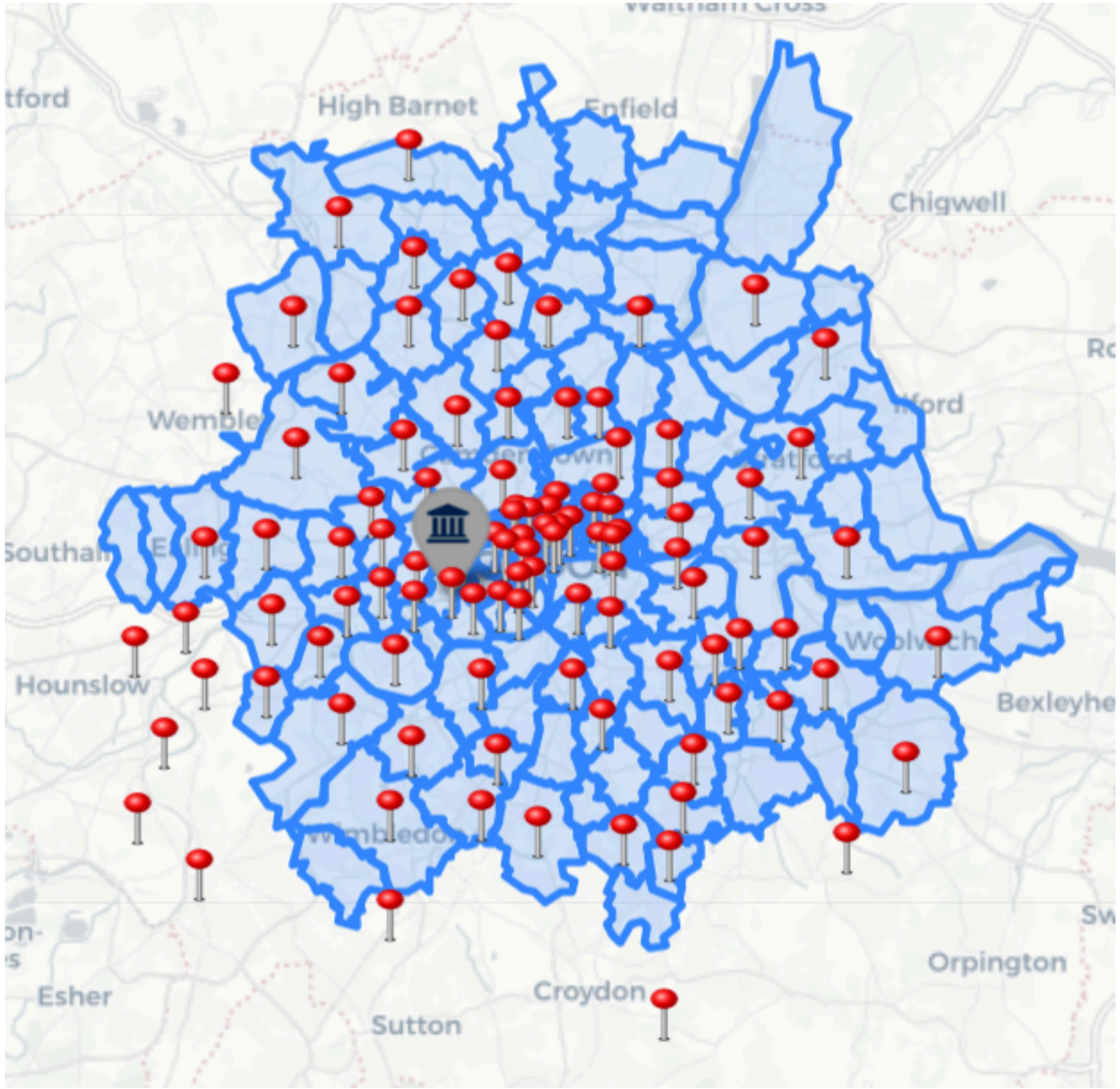
Part I: Initial Attempt

Yousef Nami — 04.10.20

Motivation

A student's perspective

- Imperial College London is in South Kensington, an unsuitable area for students due to high costs
- Students move away to farther (and more crime ridden) areas
- No system that gives students information on the neighbourhoods that they are interested in
- This project is an **initial attempt** at creating such a system, with emphasis on cleaning data and clustering neighbourhoods



Data Attainment and Cleaning

A tough challenge…

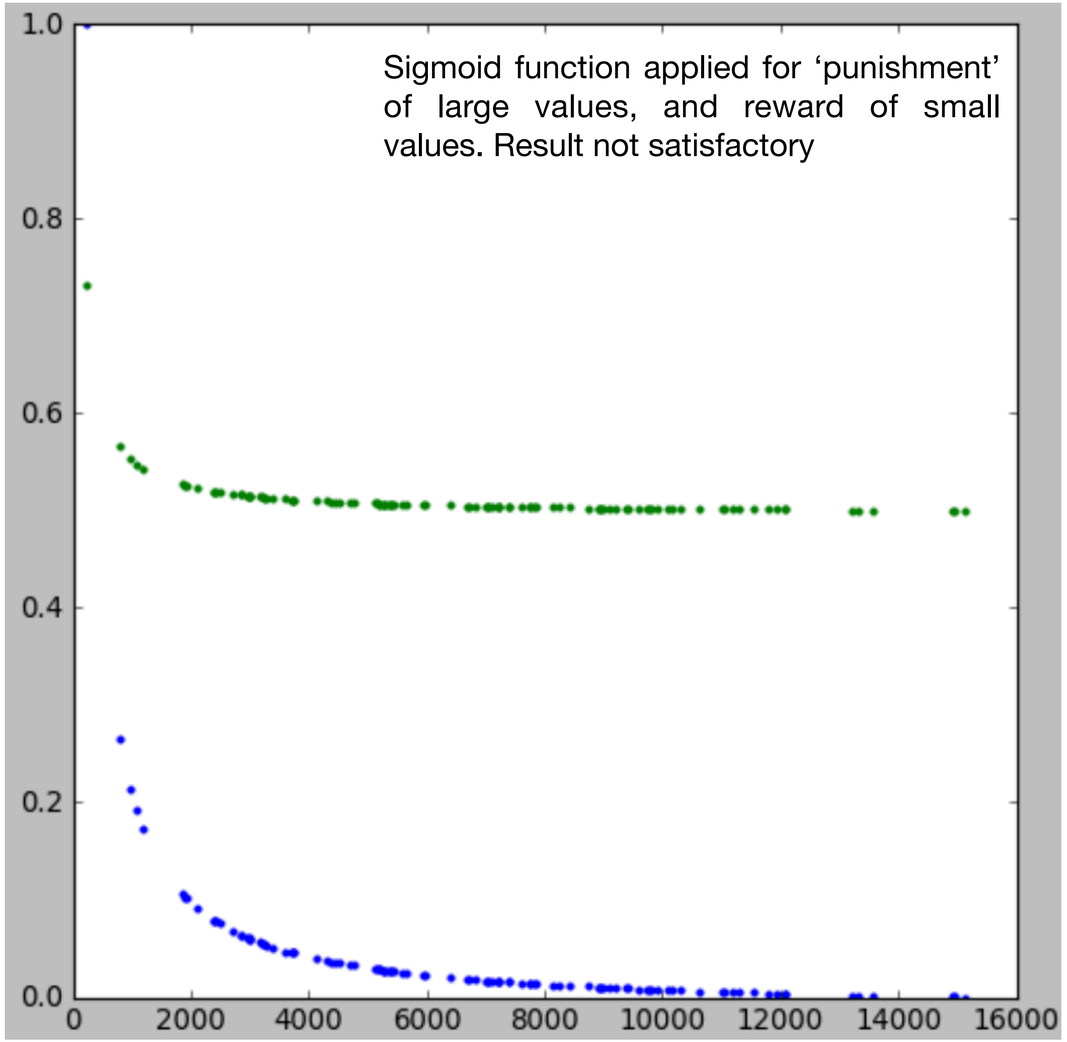
- Average weekly rent per property type (from: https://www.rentbarometer.com/london/all-prices/by-postcode.html#BR)
- Number of venues in (and distance from) the neighbourhood using the **Foursquare API**
- Duration from neighbourhood to university for different travel modes (bus, cycle, walk, tube) using the **Google Directions API**
- Missing weekly rents replaced with median rent per property type
- Distance from university calculated using latitude / longitude values and **Haversine** equation

Scoring

- Raw data converted into ‘scores’ based on suitability for students
- Distance, Duration, Cost scores are reciprocals of actual values (logic being: higher distance / duration —> lower scores)
- Example:

$$S_d = \frac{1}{d}$$

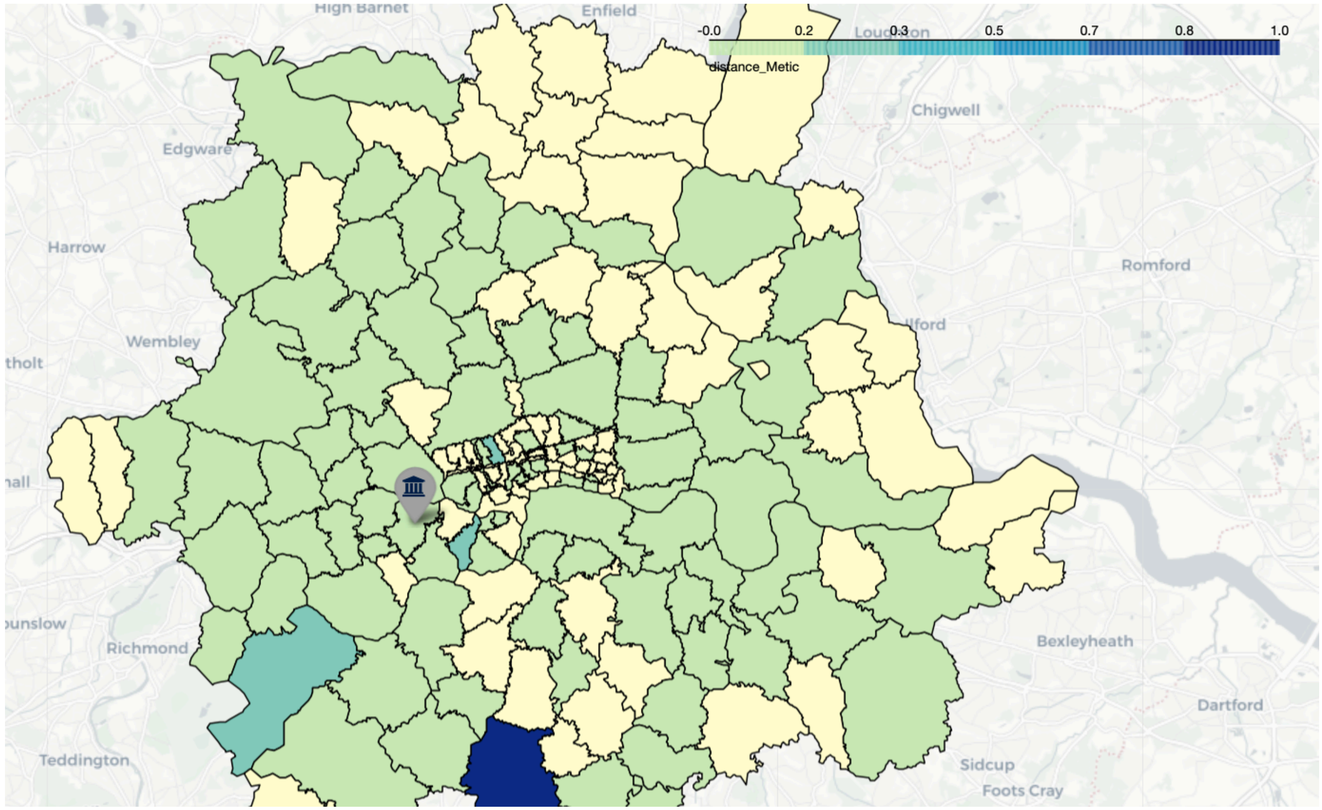
S_d = distance score, d = distance



Data Quality

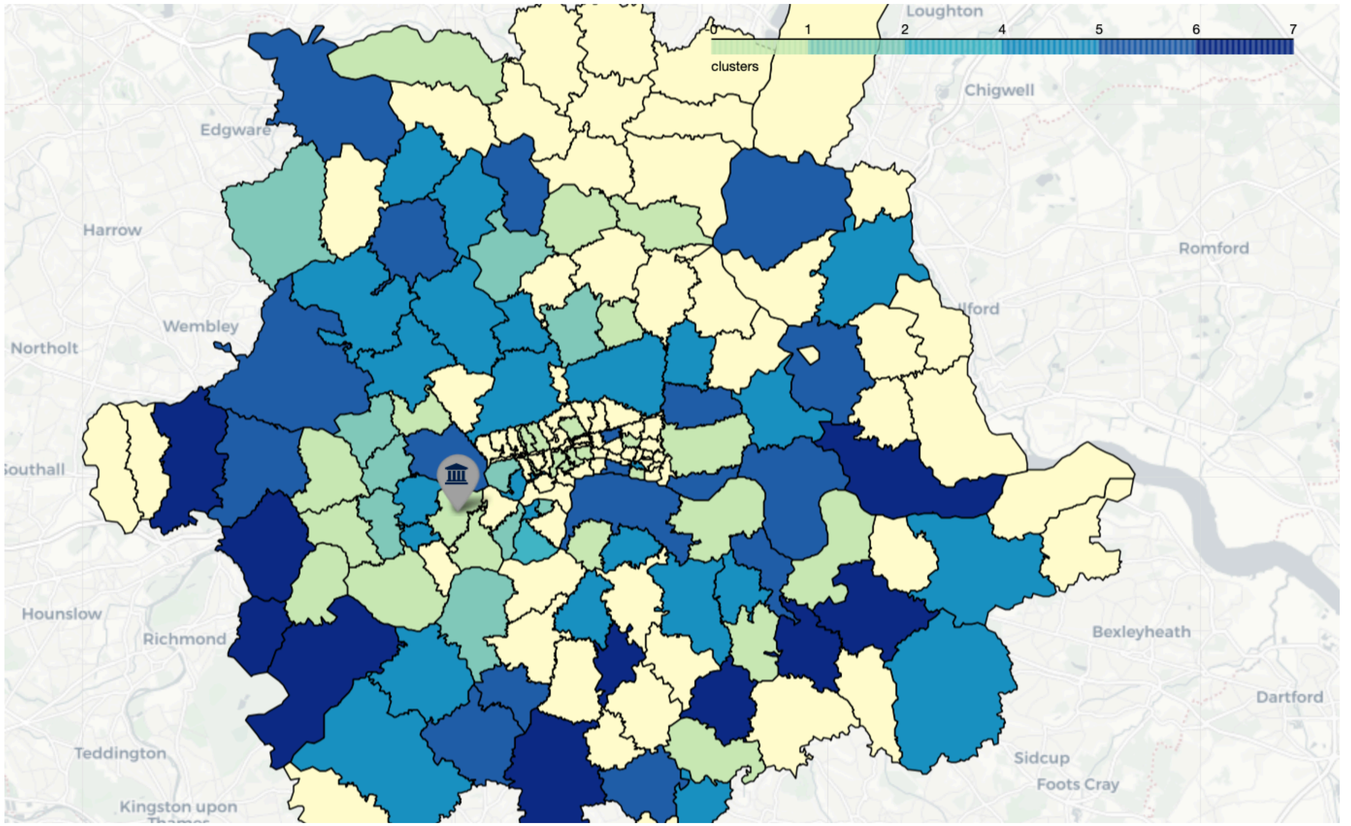
Not great…

- Choropleth map showing distance metric not working correctly
- Most postal districts have similar values, requires special function
- Odd ‘high distance score’ value at London outskirts



Clustering

- Final clustering not entirely informative
- Interesting that outer neighbourhoods have a higher score
- Requires further investigation



Conclusion and Further Work

A long and rewarding journey ahead

- Data highly dimensional — does not lead to very informative outcomes
- That said, potential for improvement is great, and a very good initial attempt
- Scoring functions need refinement
- Better data needed