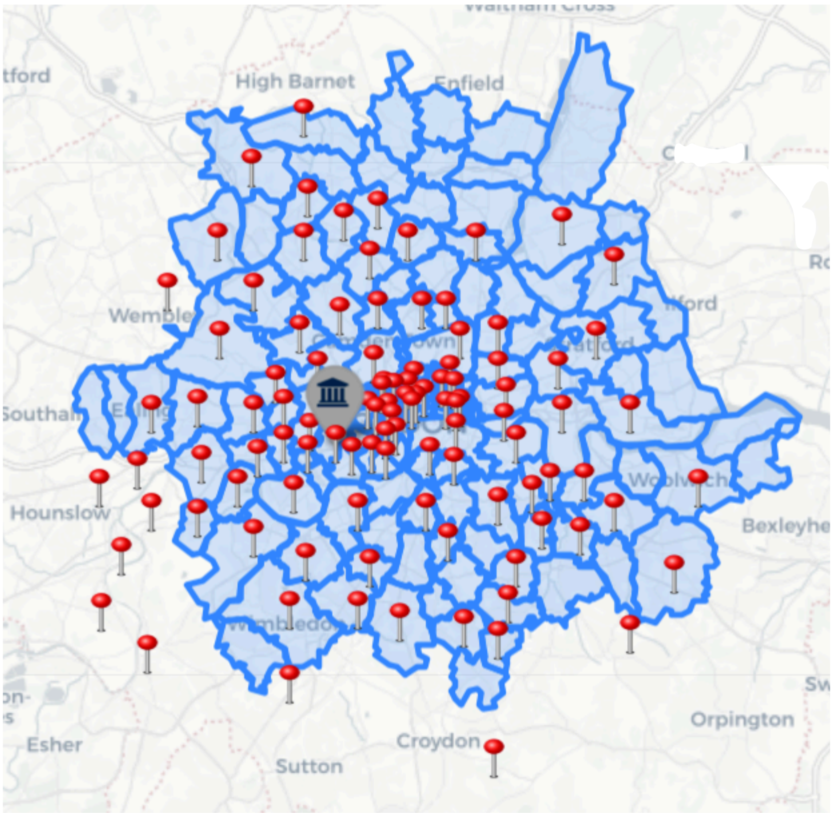
What is the best neighbourhood to live in as a student at Imperial College London?

Part I: Initial Attempt



By: Yousef Nami

Change log:

v.1.0.0.	Initial report prepared for IBM Professional Data Science Certificate
v.1.1.0.	Fixed change in data sources, added images
v.1.2.0.	Added Literature Review
v.1.3.0.	Added Methodology
v.1.4.0.	Completed Methodology
v.1.5.0.	Added results section
v.1.6.0.	Completed discussion, conclusion and further work
v.1.7.0.	Completed Abstract
v.2.0.0.	Proof-read, fixed minor changes

Abstract

Imperial College London, being situated in the heart of South Kensington and lacking in physical space, does not guarantee accommodation for students from their second year onwards. South Kensington lacks student accommodation and has soaring rent prices. This leads to many students moving into flats in areas away from Imperial, leading to longer commute times and lower safety due to increased crime rates.

A student should be able to find an area that suits their needs best: they may enjoy a long commute, but only if they are able to cycle to university. They may prefer living in areas with a high density of restaurants, whereas others may prefer a vibrant nightlife, all while living in a safe and (relatively) crime-free area. The current systems that exist in place for finding accommodation (such as zoopla.co.uk and rightmove.co.uk) do not provide any insight into the areas where the properties are listed.

This report explores the first part of a multistage project to develop a program that can give students an answer for what is the best neighbourhood to live in, based on their preferences. Here, a first attempt to cluster neighbourhoods based on their suitability for students using their distance from university, their average weekly rent, the availability of amenities such as restaurants and grocery stores in the area, and the transport duration to university using different travel modes (bus, tube, walk and cycle) is conducted.

Table of Contents

1. Introduction

1.1. Background

Imperial College London is a STEM university situated in South Kensington: a cultural area full of museums, parks, restaurants, shops and hotels. As such, the university's campus is fairly small, and has minimal accommodation to offer students after their first year at university. This results in the vast majority finding private accommodation, typically a couple of miles away from South Kensington, due to the high flat prices (per meter squared) in the area.

With this, commute to university becomes much longer with many opting to cycle, use the bus or the tube (underground train) instead, while student safety decreases due to increased crime rates further away from university. Of course, other areas have their own positives, as they typically have younger demographics, are more diverse, can be more 'trendy' and offer access to other amenities, such as nightclubs.

At the moment, students are able to find correct properties (in terms of price and location) through trial and error (on website such as rightmove.co.uk and zoopla.co.uk) or by consulting with real-estate agencies, such as Foxtons or Dexters. However, there exists no mechanism by which students can *understand* the type of neighbourhood or area wherein they look for housing. Ideally, students would be able to answer a couple of questions regarding what they consider to be important features in a neighbourhood, and get answers in the form of 'suitable' neighbourhoods for them.

This report aims to explore this problem in more detail and finding a solution.

1.2. Literature Review

Through consultation with many different students at Imperial, a number of parameters which increase student satisfaction have been found. These are listed down below:

- Property price
- Trendiness of area
- Availability of restaurants
- Availability of entertainment venues (excludes 'nightlife')
- Availability of nightlife venues (includes clubs, bars, lounges)
- Time to get to university (walking, cycling, bus or tube)
- Average age in the area
- Crime rate
- Closeness of grocery stores

As such, these parameters will be considered in the analysis.

During research, the geography of London was better understood. In England, areas are split into what is known as 'districts' (also known as Local Authority Districts). Within London specifically, there are 33 districts: 32 of which are known as the 'Boroughs of London', and 1 which is known as 'City of London'. These 33 districts either refer to an area considered part of 'Outer London' or 'Inner London'.

Each borough can be further broken down into neighbourhoods known as 'Postal districts'. There does not seem to be any formal definition for these in the literature, but essentially this refers to sub-divisions of Boroughs that make up the first part of the London Post Codes, for example:

SW7 3BQ

The first portion represents the postal district, and the second is specific to an address within that postal district.

Within ‘Inner London’, the letters within the postal postal codes have a geographical meaning, whereas the number that follows refers to the partition.

E: Eastern	**N:** Northern	**SE:** South Eastern	**W:** Western
EC: Eastern Central	**NW:** North Western	**SW:** South Western	**WC:** Western Central

For example, SW7 3BQ refers to an address within South Western (Inner) London, in the 7th Partition.

There are slight variations to this, as some postal districts[1] will have further sub-divisions, for example:

W1H 7LD

Where the ‘H’ represents a sub-division within the W1 postal district.

The ‘Outer London’ postal district codes have no geographical meaning, and refer simply to the areas they cover.

1.3. Scope

In the long term, the goal of this research project is to create a model that can provide students with the most suitable location to live, based on their preferences. An example could be a deployable model that quantifies each neighbourhood. This model could take an input address from a student, and it would output a score for the address, in terms of different parameters. This is open ended and can change.

For now though, the scope of this project (as part of the Capstone for **IBM Professional Data Science Certificate**), is to explore postal districts (neighbourhoods) to find any that are similar.

[1] These are high density postal districts due to their small area. They are typically located in the centre of London.

2. Data

This project requires more data than that available from Foursquare.

A description of the data required for this project will be given here, as well as links. Note that the methodology for how the data will be used is given under Methodology.

- **Foursquare:** will be used for exploring different types of venues in postal districts
- **Rent Barometer:** includes data on average weekly rent prices (£) in London by postcode districts for different property types (Studios, 1—5 beds)
- **Google Maps:** will be used to explore distances and durations from locations to Imperial College

Note that links to the Rent Barometer dataset is given under Data Sources. Both Google Maps and Foursquare data are accessed through API's, and as such have no 'datasets'.

The next subsections will discuss the datasets in more detail in terms of the parameters that they contain and how they will be used to achieve the objectives of the project.

2.1. Foursquare

Foursquare gives developers a certain number of free daily calls to retrieve information regarding venues[2]. There are premium calls that can be made, such ratings and tips, but for the time being these will be excluded from the analysis.

This project will use the following approaches:

- Postcode districts will be explored for specific venues categories relevant to students[3], these include:

- **Restaurants (**4d4b7105d754a06374d81259**)**
- **Grocery stores (**4bf58dd8d48988d118951735**)**
- **Nightlife (**4d4b7105d754a06376d81259**)**
- **Outdoors (**4d4b7105d754a06377d81259**)**
- **Entertainment (**4d4b7104d754a06370d81259**)**

The data described above will be used to create 'metrics'. This is described in more detail under the methodology.

2.2. Rent Barometer

The rent barometer dataset is a webpage containing tables which present information on weekly rent for different property types in each postcode district.

Webscraping will be used to extract this information. There are some missing values in the datasets, and these will be approximated by suitable models.

The data collected here would also be converted to a metric relevant to students.

2.3. Google Maps

Google gives developers access to Map data.

Google maps will be used to determine average durations from postcode districts to Imperial College London (bus, tube, walk and cycle).

This data will be used to create a 'closeness score' for each postcode district.

[2] I believe this to be ~ 95000 which is more than plenty.

[3] The codes in brackets refer to the relevant Foursquare 'Category ID'. That has more insight into what information is captured.

3. Methodology

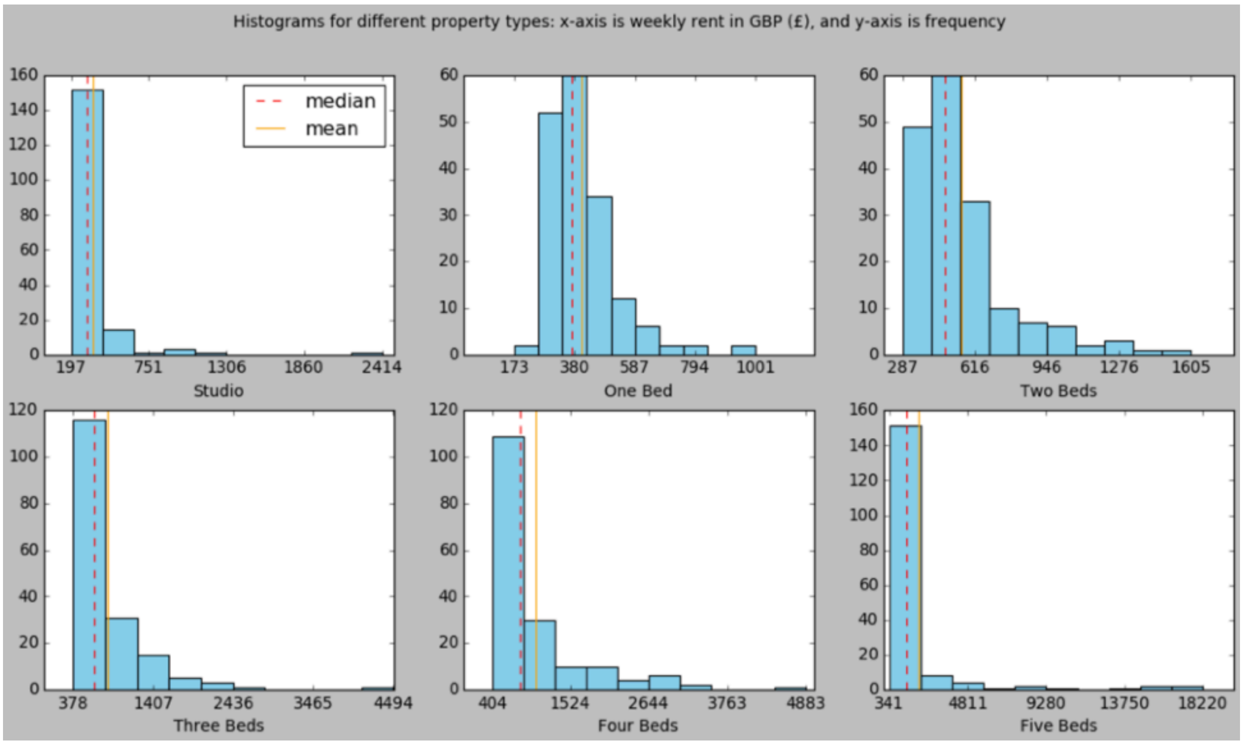
As with most data science projects, the portion of the project that took the most time was actually the preprocessing.

The first couple of notebooks dealt with scraping the data from the different data sources. Though at first crime data and demographics data were considered as part of the analysis, this had to change due to the quality of the data found.

The next subsections will discuss specific exploratory data analysis performed for each of the data sources described in Data.

3.1. House Prices and Distance

The house prices data set from Rent Barometer was quite good, however, it had many missing values for some of the properties. Since these are numerical values, instead of dropping the missing values[4], it was decided that they would be replaced by either the mean or the median.



To decide which one to choose, histograms were created for all the prices (excluding NaN values). As expected, the median was more robust, so it was chosen to replace the NaN values.

Once this had been completed, the distance from each postal district to the university was found. There is an easy way to do this, using the Google Maps API. However, for the sake of learning, it was decided that this would be calculated using the latitude and longitude values.

Using Geopy (with the Google V3 service[5]), the latitude and longitude values for each postal district were found. Then, using the Haversine formula, the distance between them and Imperial College (lat: 51.498356, long: -0.176894) were found.

The distance, d, in kilometres between two coordinates is given by:

[4] In an ideal case, this may be appropriate. However, in this instance it would have decreased the dataset into ~ 25 postal districts instead of the original ~120.

[5] The course suggests using the Nominatim service. This was found to yield terrible results. However, for those trying to reproduce the work it's worth noting that Google has only a limited number of free API calls (specifically 300 GBP; this was found to be more than enough for this project).

$$d = Rc$$

Where:

$$c = 2 \arctan \left(\frac{\sqrt{a}}{\sqrt{1-a}} \right)$$

and:

$$a = \sin^2 \left(\frac{\Delta\phi}{2} \right) + \cos\phi_1 \cos\phi_2 \sin \left(\frac{\Delta\lambda}{2} \right)$$

Note the following definitions for the symbols:

R : radius of the Earth in kilometres (6371)

ϕ : latitude

λ : longitude

Δ : represents change, destination minus origin

Having collected the data, it was time to convert them to scores. Increasing distance should be penalised, so it was decided that the distance score would just be the reciprocal of the distance.

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

The same was adopted for price, since higher price is less desirable, so the price metric (in this case a vector to encompass all the different properties) was give by:

$$\underline{S_P} = \frac{1}{\underline{P}}$$

Note on expression: the above means that you are taking a reciprocal of every value of the price vector.

Having applied this transformation, the results were then normalised using MinMaxScaler.

3.2. Travel time to University

The data collection for this relied on the Google Directions API. For each postal district, the duration to university for four transport modes were found: walking, cycling, subway and tube.

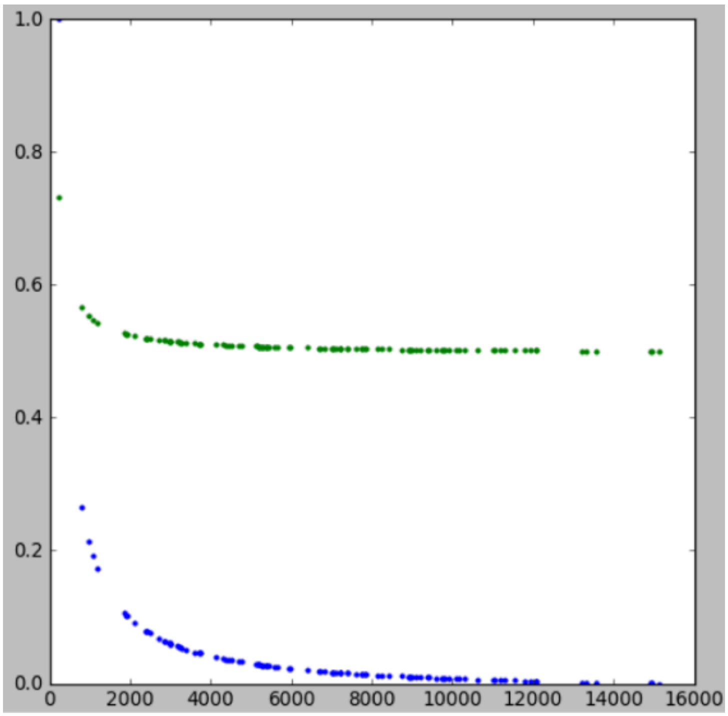
These were all calculated in seconds, and then the same reciprocal transformation was applied (the same logic stands: lower time —> better score).

So:

$$\underline{S_t} = \frac{1}{\underline{t}}$$

In this instance, it was important to consider different distributions for each variables. The reasoning behind this was that they have inherently different meanings, and should have distributions that reflect those meanings. For instance, a postal district with walking time to university, $t_w < 300\,\mathrm{s}$ should get a really high score, whereas others with really high time to university should be penalised for this, such that their scores would be essentially zero. In this case, the distribution that you'd want isn't necessarily a reciprocal, but rather something shaped like a sigmoid.

This was in fact tested, and a sigmoid transformation was applied:



In the figure above, the x axis represents walking time to university; the y axis represents the waking score (after scaling with MinMaxScaler); and each point represents a postal district.

The green line represents the data after the sigmoid function was applied.

As shown, the sigmoid definitely punished the high value times (anything after 6000 seconds essentially has the same score). However, it also decreased the variance in the scores. As such, it was deemed that this transformation was not appropriate.

It was noted that for further iterations of the project, this concept should be explored further, in particular in relation to the other variables such as tube time, cycling time and bus time, since each would require a different distribution (i.e. walking should very be low time —> 1, high time —> 0, whereas for something like bus times, you don't need such a bipolar distribution).

The exploration was then concluded at this point due to time constraints.

3.3. Venues

The Foursquare API was used to collect information on the venues near each postal district. A limit of 50 venues was set (the maximum that Foursquare allows). For each venue, the distance from the venue to the postal district was also calculated, and averaged. The number of venues per postal district (by type of venue) were also calculated. This led to the creation of a venue score based on the 'density' of venues in a given postal district:

$$\underline{S_T} = \frac{\underline{N_T}}{\underline{d_T}}$$

The subscript T in the above equation refers to the type of the venue in question. This score penalises the district for how far the venues are from it's centre, and commends it for the number of venues. In most cases, the number of venues was 50.

In retrospect, setting the limits to 50 was a mistake, as this led to repetitive venues, i.e. a postal district, say SW7, would get venues from SW5. In the future, one should set a radius limit. This way, you could get a measure of 'venue density' for each postal district.

3.4. Clustering Postal Districts — KMeans Clustering

Since this was the first iteration for the project, and KMeans is a fairly intuitive and easy algorithm to apply, it was decided that this would be applied.

Having no knowledge of how many clusters to expect, the default *n_clusters = 8* was chosen. The results are difficult to interpret, but there are some insights to draw from it.

These are discussed in more detail under results.

The conclusion to the report provides recommendations for further iterations of the project.

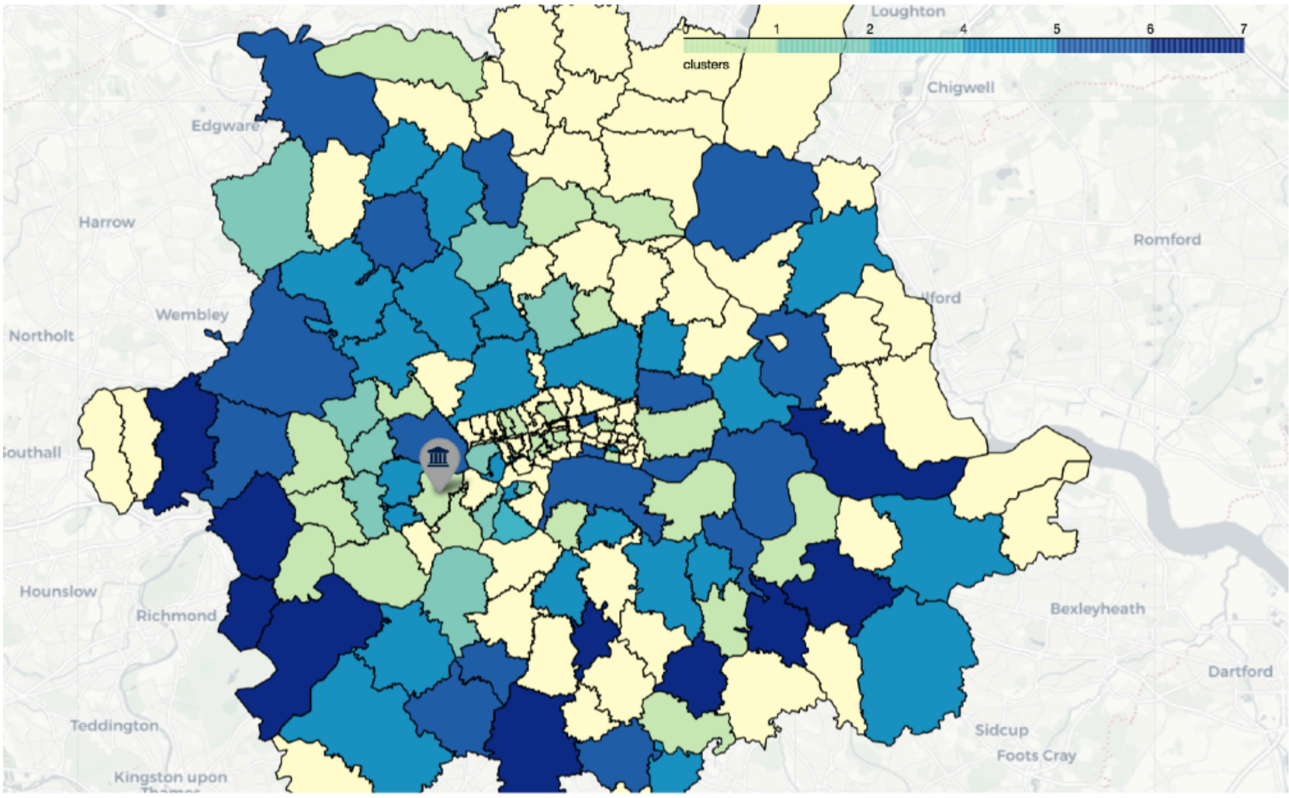
The KMeans method of clustering was chosen because of the high-dimensionality of the data. Since this is an unsupervised learning problem, the pool of algorithms to choose from was fairly small[6]. Compared to the other clustering methods, such as agglomerative, hierarchical and DBSCAN, KMeans was deemed to be the most appropriate for it's ease of use and interpretability.

For future iterations of the project, where there are no restrictions posed by the IBM Certificate, more comprehensive methods will be sought out. These will also be evaluated against different metrics.

[6] The author intended to use algorithms from the IBM Course to ease the marking process of the Peer reviewers.

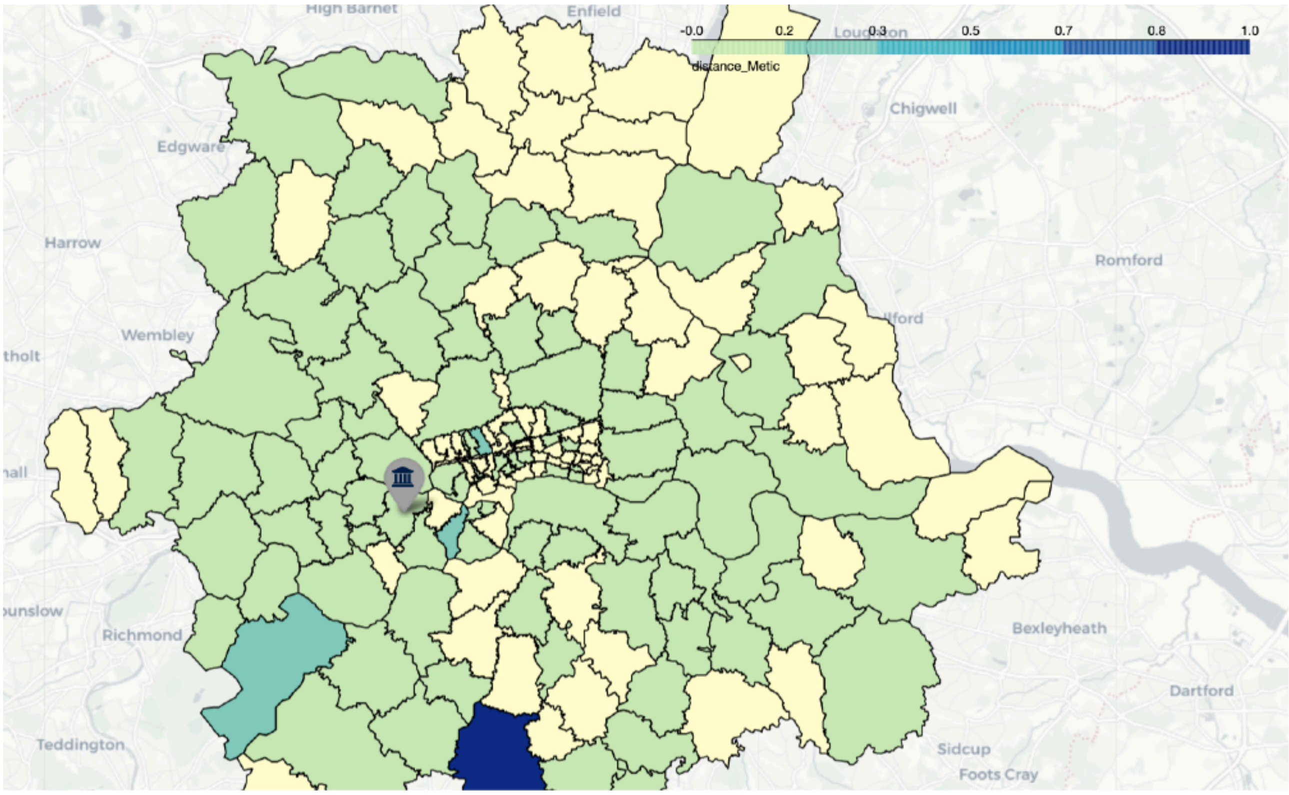
4. Results

The results after applying the clustering algorithm are shown below.



As exhibited by the choropleth map above, the similar postal districts do not seem to show any clear pattern, at least visually.

One interesting thing to note is that a higher score was given to postal districts that are somewhat far away from the university (and in fact, the centre of London). This result is dubious, and does not give confidence that the model has worked very well. In fact, as part of a 'debugging' mode, trying to figure our what went wrong, an individual choropleth for the distance metric was found to show some erroneous results.



This result is quite confusing: how could it be that one of the postal districts that is far away from the university have such a high distance score, while most other postal districts have a very similar (visibly the same — the green colour) distance score?

This suggests, at best, that there has been an error in the data collection phase.

These insights are highly useful as they will be key in defining the next part of this project.

5. Discussion and Further Work

For a first attempt, the project does well in exploring the data that exists. In fact, it uses multiple web scraping tools (from beautifulsoup to Google Direction API to Foursquare API) to achieve this. The project also does well in visualising the data on choropleth maps using the folium package, and showing the different postal districts for which data is available. In fact, it might be worth noting that, much like most data science projects, the biggest weakness of this projects is the data that was found.

5.1. Data

During the course of this research project, it was found that data on demographics and crime is very difficult to attain. In fact, the data sources that were found (see Unused Data Sources) proved to contain much unnecessary information, and were not specific enough to 'Postal Districts', instead referring to Wards[7] and Boroughs.

Other missing data included the events in the vicinity of postal districts, or the 'trendiness' of an area. In theory, it is possible to find such data using the Foursquare API, but limited API calls and time-constrained rendered this as a not a priority for this project.

These are important factors that should be considered in the analysis, and thus should not be dismissed (as was done in this project, due to time pressure). Relevant authorities should be contacted for datasets, should alternative sources not be found.

In terms of the data that was found, the rent data had a substantial number of datapoints missing. This was offset by using the median of the weekly rents, however, this may not be the best method of fixing the problem at hand. An alternative solution would be to design a data scraping tool (using Python or otherwise) that makes calls to websites such as rightmove.co.uk or zoopla.co.uk, finding properties within postal districts, and collecting information on the number of rooms, size of properties, average rent. This would solve two problems: it would fill the gaps found on Rent Barometer, while giving more useful data on the *nature* of properties within a postal district. However, the author of this report acknowledges that such an idea is not feasible in the short term, and that other improvements should be prioritised over this. An easier short term solution might be to cluster postal districts based on similarities in price (for each property type), and hence, having clustered them, use the mean / median of the clusters to fill in missing values. This itself could be a whole project on it's own though id done correctly, perhaps a Part II to this project.

With regards to the data collected from Foursquare, as mentioned previously, there is a need to optimise what the data means, because currently there does not seem to be much variation within the postal districts (due to the lack of specifying a radius). The solution is, of course, to specify a radius (to be decided from student surveys) for the closeness of venues, and then calculating the density of venues on the *assumption* that the venues are uniformly distributed.

5.2. Methodology

The methodology seemed sound, until the point of the analysis. It appears that the models developed for scoring the different parameters did not make sense, either because the models themselves are inadequate because they don't provide sufficient spread (or punishment where appropriate) in the data, or because of mistakes encountered during the data collection / cleaning stage.

The scoring system must be challenged rigorously, as it does not appear to have provided fruitful results. It's very important to consider the data will be affected by any models applied, and whether it should or not. Depending on whether the project is static or dynamic[8], different solutions may be applied. For a static project, the mathematical models applied should consider the following:

[7] Divisions within the City of London. A remnant of a medieval governmental system.

[8] Static here meaning a pre-defined method of scoring that does not take student input into account, dynamic being a data-driven model that modifies the scoring based on the students' preferences.

1) The model must ‘punish’ walking duration, such that a very close postal district gets a very high score, whereas one that is far away gets a score of zero. The model must also ensure that the other duration metrics (bus, cycle, tube) do not dilute the meaning or importance of the waking metric. Given that cycling is the least likely option for students, it must also give it the least importance[9].
2) The model must also accurately depict the postal district distances from Imperial College
3) The model must ensure that the importance of different data degrees of freedom is not diluted. This is often the case in highly dimensional data; perhaps some dimensionality reduction should be applied.

The modelling method used — KMeans — was chosen mostly for pragmatic reasons. Other, more serious algorithms, should be considered. In fact, even the parameters of the KMeans algorithm should be optimised to do more than the default *n_clusters = 8*.

5.3. Next steps

The main steps to follow would be to cleanly document the work done, by adding important functions into .py files to allow reproducibility. This way, the author — or any person intending to continue the project — could do so with ease. The author of the report is for the time being is tied up with other commitments, and hence cannot continue the work straightaway. This project was rushed in the end due to time constraints, and hence the modelling section (and even parts of the preprocessing, data cleaning and visualisation) are not clear / logical. That said, this work will not end here.

Overall, a clear outline of this project is needed to ensure that it gets completed. For the time being, the machine learning part of it is just for exploratory purposes.

The important questions to ask are:

1) Does this project intend to add any functionality over ‘clustering’ postal districts into those that are similar?
2) If it does intend to behave as an engine where a user selects their preferred attributes, is machine learning necessary? Could the project be done algorithmically?
3) Should an initial engine be deployed, would there be a system whereby data from students around the world is collected and added as ‘supervised’ data, from which the machine can continuously learn from (by means of neural networks for example)?

[9] Justifications include: bike thefts very common in London, cycling dangerous in London due to lack of infrastructure, dangerous driving and constant rain (increases chance of slipping).

6. Conclusion

This project made an initial attempt at clustering neighbourhoods in London based on parameters important to students, namely the weekly rent prices, distance from Imperial College, transport duration to Imperial College (in terms fo walking, cycling, bus and tube) as well as information on nearby venues such as restaurants, nightlife venues, entertainment venues, outdoor venues and grocery stores.

The resulting outcome from the clustering process is inconclusive, and needs further work, in particular with regards to the data that has been collected, and the methods used to covert to an importance score.

Overall, this first attempt shows that the work required for the end goal — creating an engine that provides students guidance on the nature of the neighbourhoods that they are considering living in — is far greater than expected, and that this project will require many iterations of work. The single greatest problem facing the project right now is the lack of good data.

Despite this, the work conducted here shows potential. For the purposes of gaining more data, it might be worth leveraging the full power of the Foursquare API by purchasing premium rights, allowing better access to information such as ‘trendiness’ of an area.

7. Resources

7.1. Data Sources

- https://www.rentbarometer.com/london/all-prices/by-postcode.html#BR
- https://developers.google.com/maps/documentation/distance-matrix/overview

7.2. Other

- https://en.wikipedia.org/wiki/Districts_of_England#Metropolitan_boroughs
- https://en.wikipedia.org/wiki/London_boroughs
- https://en.wikipedia.org/wiki/London_postal_district#:~:text=The%20E%2C%20EC%2C%20N%2C,to%20the%20London%20post%20town
- https://en.wikipedia.org/wiki/Wards_of_the_City_of_London

7.3. Unused Data Sources (for Part II)

- https://data.london.gov.uk/dataset/office-national-statistics-ons-population-estimates-borough
- https://data.police.uk/data/fetch/159fe36a-b26d-4bb4-882a-803f490a7b2b/