Capstone Project: The Battle of the Neighbourhoods

- The Hearts of Europe -

Dean Brand

1 Introduction

Moving to a new city can be a very daunting adventure, especially because of the lack of familiarity with the area. It can be things as simple as knowing the closest supermarket to your daily route or finding your favourite restaurant that can make life exponentially better when they are incorporated into one's life. This is perhaps even more of a significant effect on one's life if they are a student, living on a constrained budget and free time allowance between studies. So, it seems like a natural benefit to know as much as possible in a snapshot analysis of the prospective city. I am facing this situation currently, where I have an option of cities to move to in continuation of postgraduate studies, and knowing how these cities compare and how they are each laid out will be incredibly useful.

I know that I am not alone in this situation of being entirely unfamiliar with a city but needing to know how best to adapt to the prospective lifestyle, without needing to spend a long time exploring the town to become familiar with it the long way. With the growing globalisation of education, this simple snapshot analysis will become an increasingly useful tool for students when browsing their options, as well as for the institutions which want to advertise the "student life" associated with them. I have seen firsthand how this specific topic is heavily advertised, and it makes sense why, people want to live comfortably, and being in a strange place is often too daunting for it to be worth relocating just because of the university. This will be a mutually beneficial tool to mediate the first impressions of each of the cities and how well they will cater to the individual. The project will rely heavily on location data, in the form of mappable area coordinates as well as the matching Foursquare data, which will then be able to describe each neighbourhood comprehensively in terms of what kinds of venues each has to offer, which creates a snapshot of the essential details of the possible new home of the prospective traveller.

2 Data

The data of this problem is obviously the central factor in the project. As such, it is worth a quick glance to see how the overall aim of the project will be formed from the handful of sources. One of the most valuable data sources in this project is that which provides the location data of the neighbourhoods in the cities under analysis, so that it can be cleanly and comprehensively visualised. To accomplish this goal, with the added benefit of consistency among the cities, the data is sourced from the Inside Airbnb database (`http://insideairbnb.com/get-the-data.html`). This database provides an incredible wealth of data, in many different forms. The particular file which is used for each city is a `geojson` file, which provides the neighbourhood name (`neighbourhood`), borough name (`neighbourhood_group`), and the map shape of each neighbourhood (`geometry`), from which we can extract the coordinates and nicely present the maps of the cities. An example of this data frame is shown in Figure 1 below.

	neighbourhood	neighbourhood_group	geometry
0	Altstadt-Lehel	None	MULTIPOLYGON (((11.59520 48.14170, 11.59500 48...
1	Ludwigsvorstadt-Isarvorstadt	None	MULTIPOLYGON (((11.55600 48.14080, 11.55930 48...
2	Maxvorstadt	None	MULTIPOLYGON (((11.58430 48.14420, 11.58310 48...
3	Schwabing-West	None	MULTIPOLYGON (((11.58170 48.17630, 11.58320 48...
4	Au-Haidhausen	None	MULTIPOLYGON (((11.59560 48.14050, 11.59590 48...

Figure 1: Example of entries in the data frame.

This location data can then be seamlessly coupled with the Foursquare databases, through the name of each neighbourhood. This then allows for the access to all of the venue data of the neighbourhoods, and as such only these two databases are required for the project, minimising the room for error and maximising the consistency in the data as if all of the information was being sourced from one enormous database.

3 Methodology

The process of wrangling and moulding the data was a simple enough process which will be walked through in this section. The first step of the methodology was to read all of the data files, in `geojson` format, into Pandas data frames, as in Figure 1. Now, the data did not come in perfect form for the task, as the `geometry` column gives the coordinates of the boarders of the neighbourhoods, not their central coordinates. Fortunately, the Geopandas (based on Pandas) package allows for better manipulation of these kinds of datatypes. The package has a method for extracting the centroids of the `multipolygon` data type, which simply calculates the average point of the boarders to give a central location point, which could then be split into latitude and longitude through some string manipulation. An example of a resulting cleaned data frame is displayed in Figure 2 below.

	Neighbourhood	Latitude	Longitude
0	Altstadt-Lehel	48.141273	11.583178
1	Ludwigsvorstadt-Isarvorstadt	48.130152	11.561218
2	Maxvorstadt	48.148069	11.564298
3	Schwabing-West	48.166539	11.569307
4	Au-Haidhausen	48.129727	11.594758

Figure 2: Example of entries in the cleaned and reduced data frame.

The `geopy` package could then be used to find the location data of each of the cities through a simple search, and the coordinates of each university was obtained through a manual search, as they are only for reference in my particular use of the program. All of these location data could then be used in the form of a Folium map, to give points on each city map to easily identify the neighbourhoods at a glance. These maps are displayed in Figures 3 to 6 below.

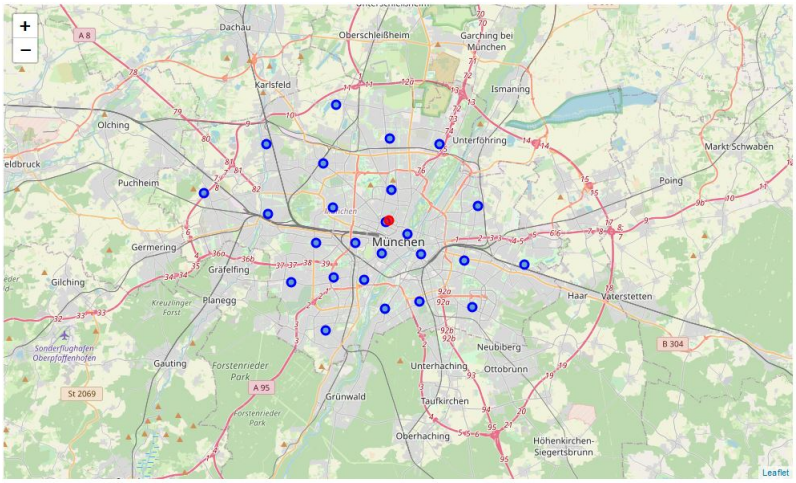

Figure 3: Map of Munich, with neighbourhoods in blue and the university in red.

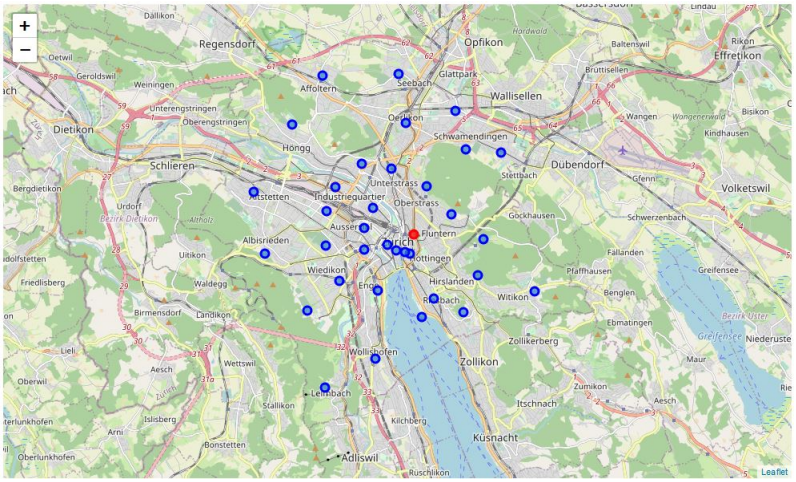

Figure 4: Map of Zurich, with neighbourhoods in blue and the university in red.

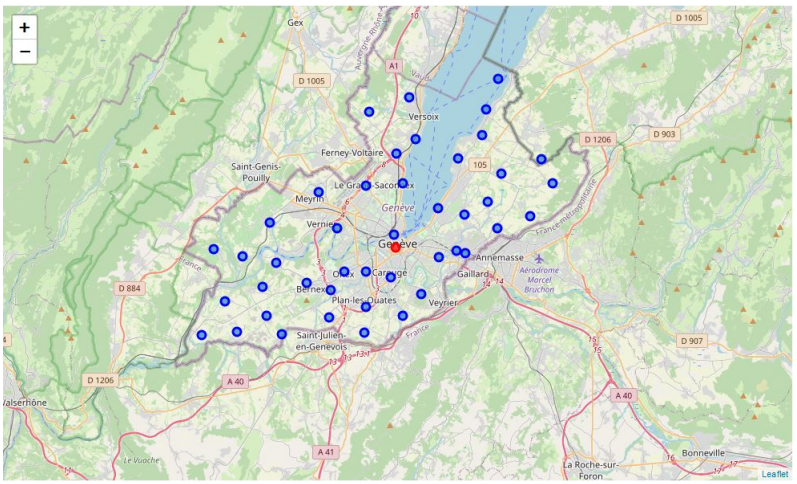

Figure 5: Map of Geneva, with neighbourhoods in blue and the university in red.

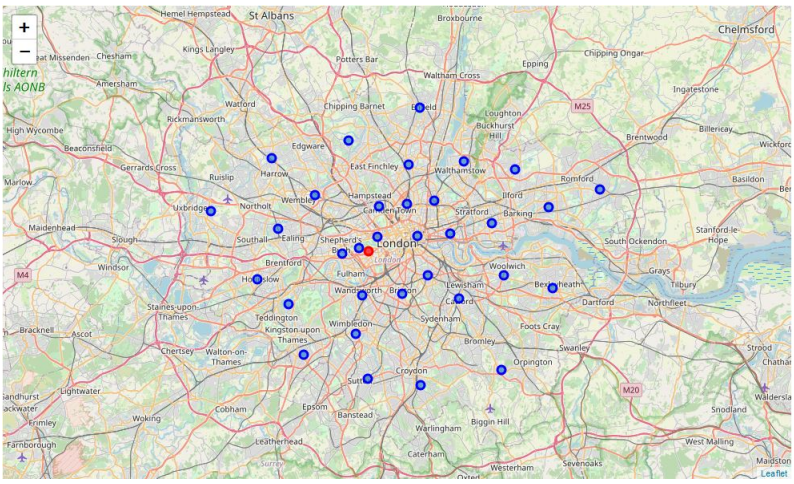

Figure 6: Map of London, with neighbourhoods in blue and the university in red.

This gives a nice glance at each of the cities and how distributed their areas are, so that we can see how many different options there are and their relative distance to the university and city centres. To improve this analysis of the neighbourhoods, we can now incorporate the Foursquare data. The API is accessed in the usual way, with the client ID, secret, and access token. This is then used to pull data from the database, specifically through a function designed to obtain the venue and venue category, as well as the associated neighbourhood location data, and compile each entry into a new data frame. This is done for each city, but an example is displayed in Figure 7 below.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
0	Altstadt-Lehel	48.141273	11.583178	SEITZ Trattoria	Trattoria/Osteria
1	Altstadt-Lehel	48.141273	11.583178	Liebighof im Lehel	German Restaurant
2	Altstadt-Lehel	48.141273	11.583178	Kitcho	Japanese Restaurant
3	Altstadt-Lehel	48.141273	11.583178	Hotel Vier Jahreszeiten Kempinski	Hotel
4	Altstadt-Lehel	48.141273	11.583178	Hofgarten	Garden

Figure 7: Venue data for each associated neighbourhood in Munich.

This provides us with another opportunity for deeper analysis, as we can see the types and counts of each venue and how they are distributed among the neighbourhoods. A good metric for a glance analysis is to get a visualisation of the density of each of the neighbourhoods for each city. This is done in Figures 8 to 11 below as bar charts.

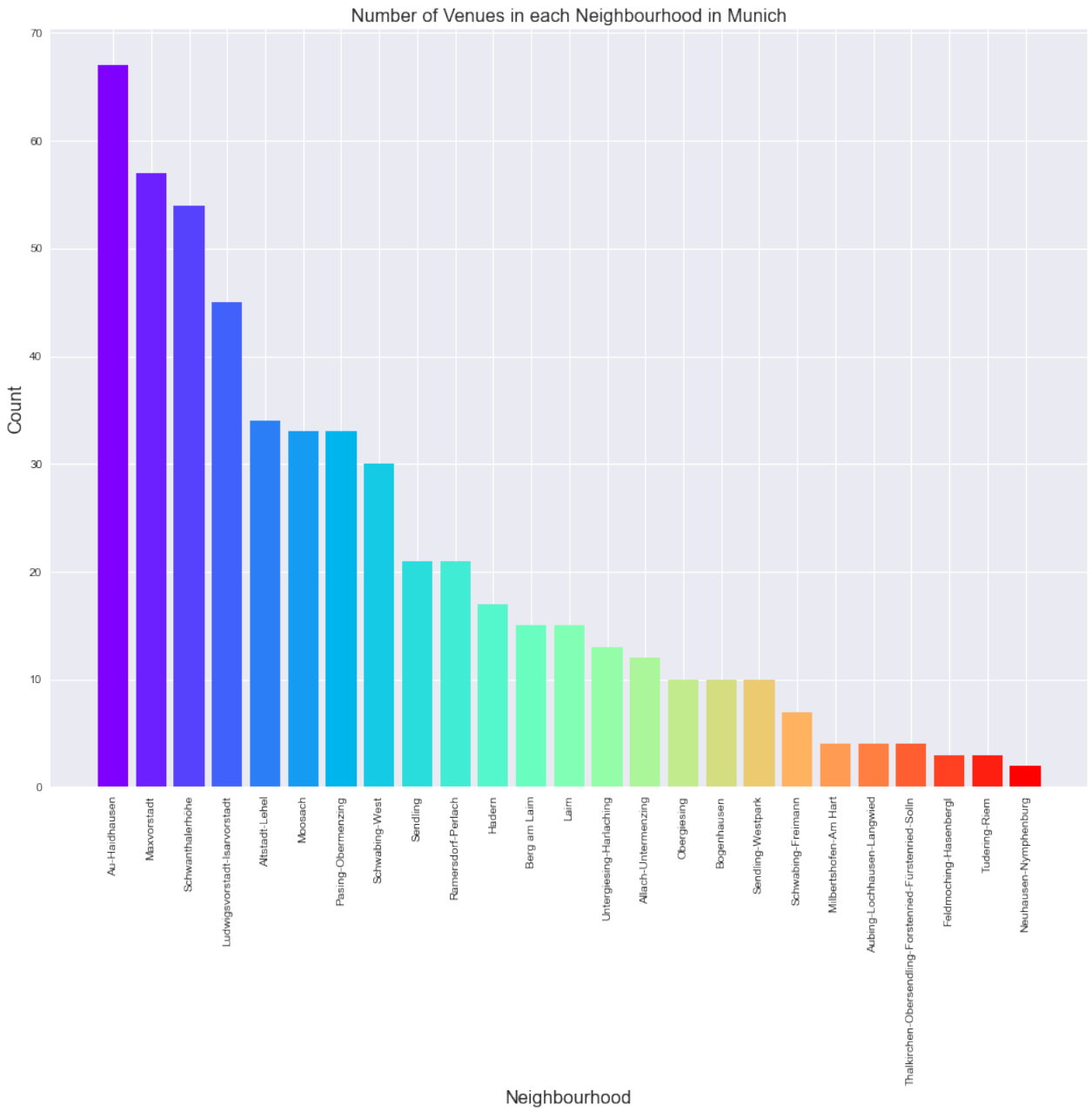

Figure 8: Bar chart of the venue count for each neighbourhood in Munich.

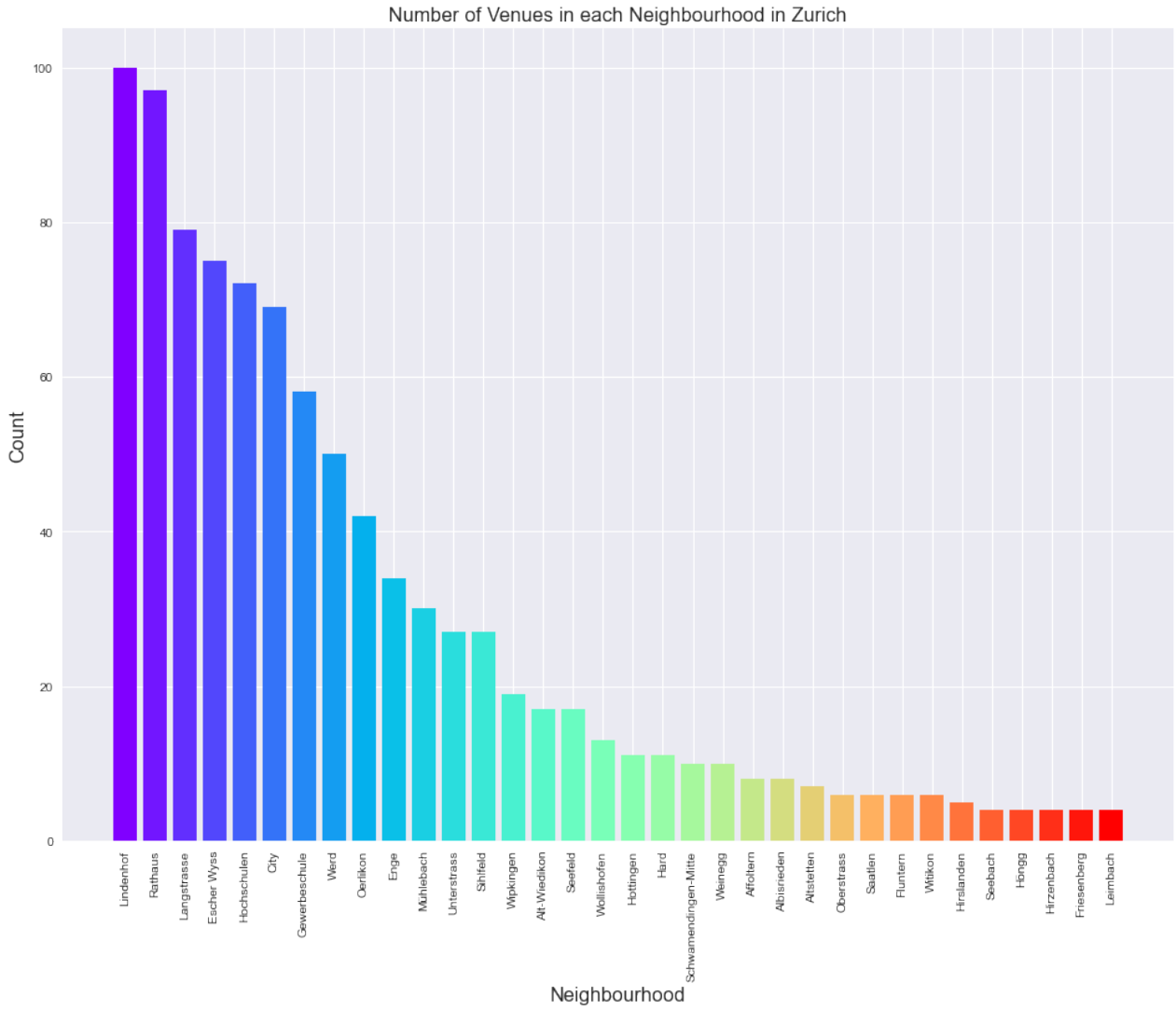

Figure 9: Bar chart of the venue count for each neighbourhood in Zurich.

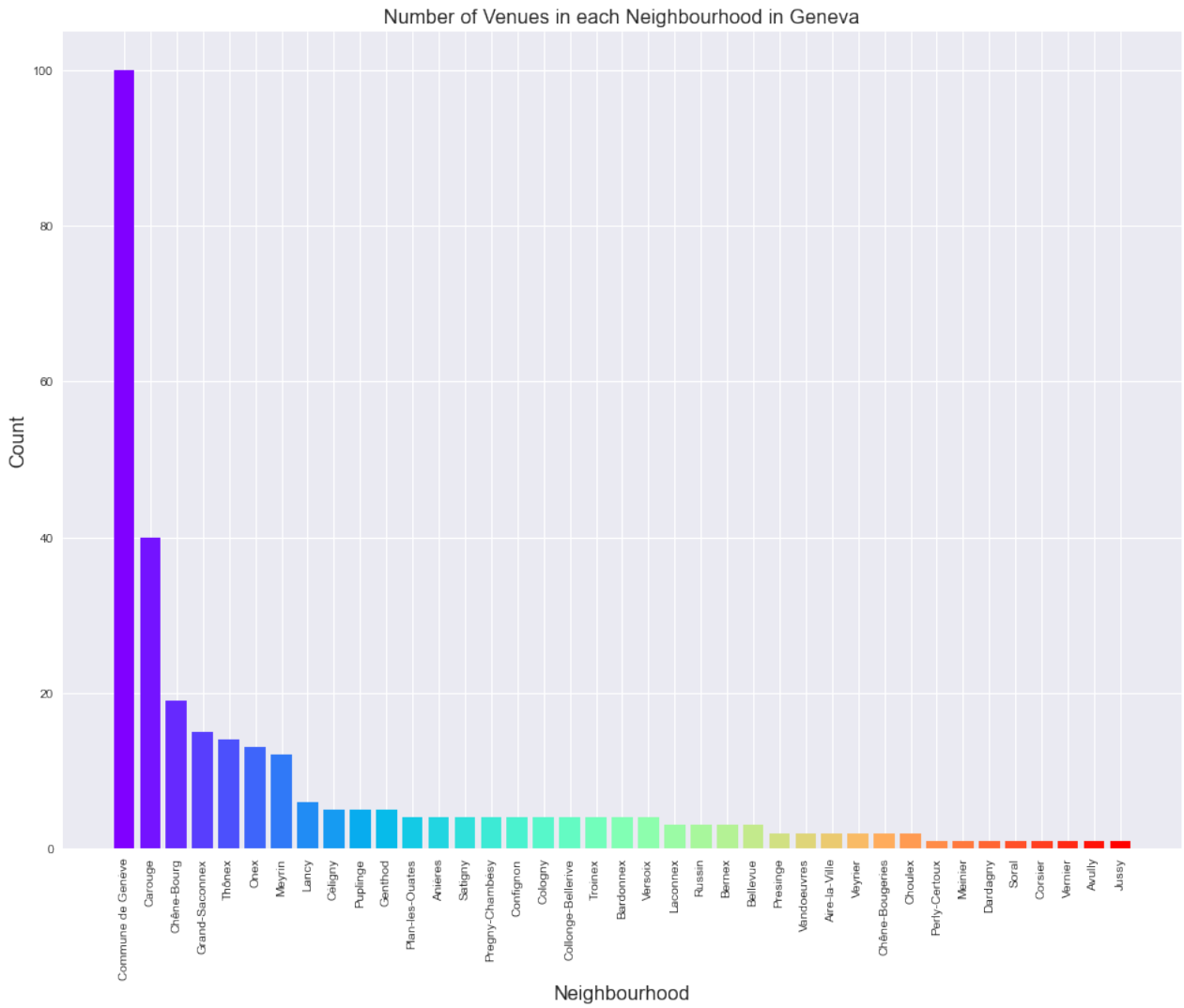

Figure 10: Bar chart of the venue count for each neighbourhood in Geneva.

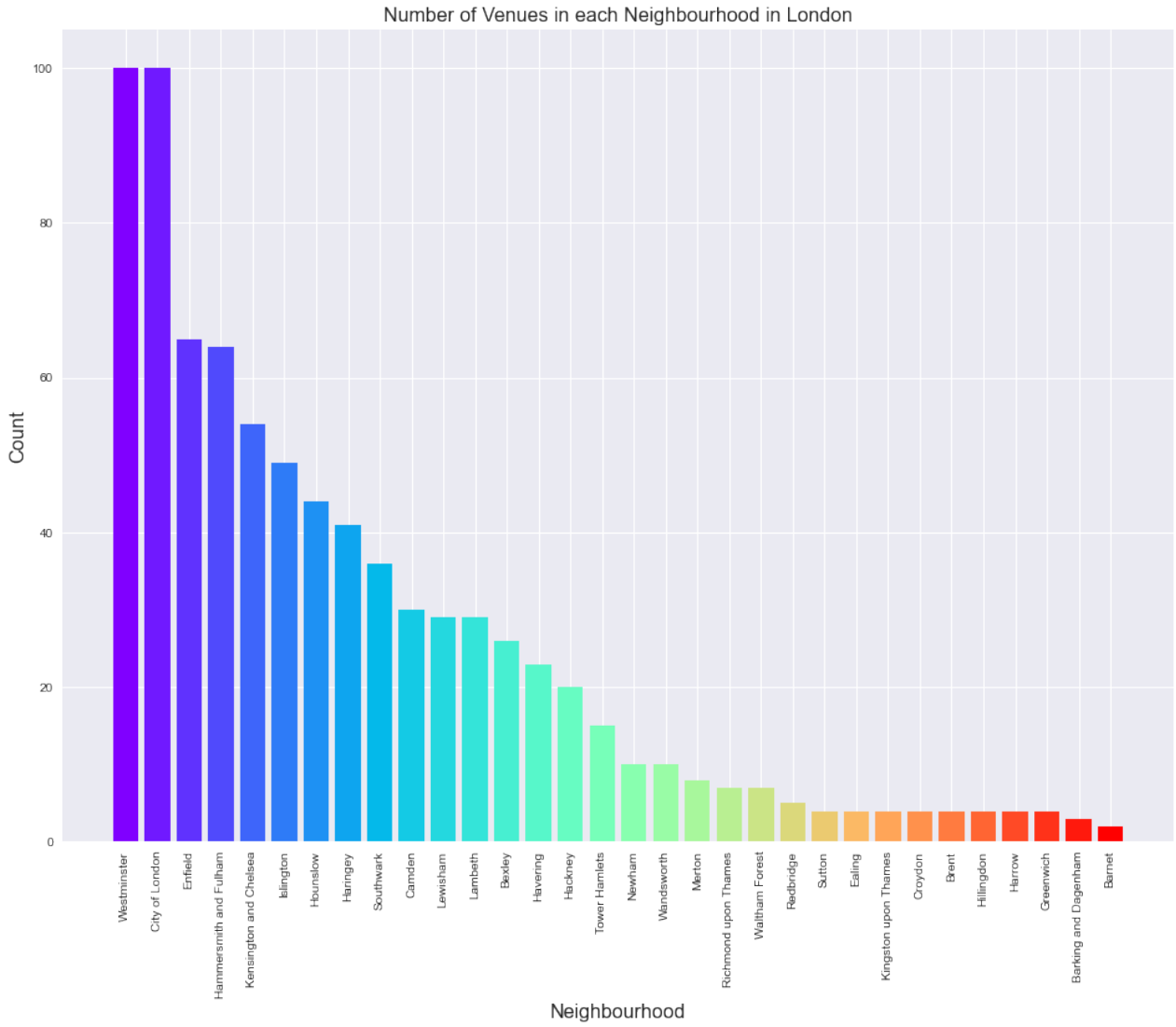

Figure 11: Bar chart of the venue count for each neighbourhood in London.

The details of each of the plots can be seen in greater detail in the Github notebook. These charts give great insight into how dense each of the neighbourhoods are, and as such give a feeling for what kind of experience can be expected in each of them, ranging from a quiet small town with a handful of essential venues, to bustling city centres with all kinds of venues all over the place. It would, however, be a lot better to see what kind of venues these neighbourhoods host as well. So, the next step in the methodology is to compile the venues in an analysable data frame. This is done through encoding each of the venues into a numerical value which can then be measured and used in calculation. The first of these calculations is that of finding popularity of the venues for each neighbourhood, to not only see what kind of venues are present but also to see which kinds are the most visited and most defining of the neighbourhood. However, the names of each venue might not be that familiar or insightful, so the grouping and ranking is done by the venue category instead. An example of this for Munich is displayed in Figure 12 below.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allach-Untermenzing	Supermarket	Drugstore	Park	Trattoria/Osteria	Bus Stop	German Restaurant	Bakery	Sporting Goods Shop	Italian Restaurant	Light Rail Station
1	Altstadt-Lehel	Plaza	Cocktail Bar	Hotel	Café	Opera House	Japanese Restaurant	Convenience Store	Historic Site	Nightclub	Palace
2	Au-Haidhausen	Café	Italian Restaurant	Plaza	Bakery	French Restaurant	German Restaurant	Bar	Ice Cream Shop	Cocktail Bar	Turkish Restaurant
3	Aubing-Lochhausen-Langwied	Design Studio	Sporting Goods Shop	Pharmacy	Supermarket	Palace	Newsstand	Nightclub	Opera House	Organic Grocery	Outdoor Gym
4	Berg am Laim	Supermarket	Bakery	Dog Run	Tram Station	Hotel	Bavarian Restaurant	Gastropub	Discount Store	Italian Restaurant	Asian Restaurant

Figure 12: Venue category popularity ranking for each associated neighbourhood in Munich.

Now we have all of the venues and neighbourhoods related in a very useful and insightful way, we can begin building models for the data to find more hidden patterns in the data. This is done through a clustering method, to find similar neighbourhoods and cluster them into groups, which will narrow down the options into new neighbourhood categories. In this project the clustering was done through the K-means model, and as such it required a predetermined number of clusters. While it would be nice to have a constant set number of clusters, say 5, to make for neat analysis, the data did not work nicely with this bound as in most cases one large cluster was made with single-entry clusters being the outliers. This problem is echoed when taking a deeper look at an inertia plot which ideally allows for a clear outstanding value at which the optimal cluster number is found, but as in Figure 13 we can see that there was no definite 'elbow' in the curve, meaning that the clustering gets consistently better until each entry is its own cluster, which is the trivial solution.

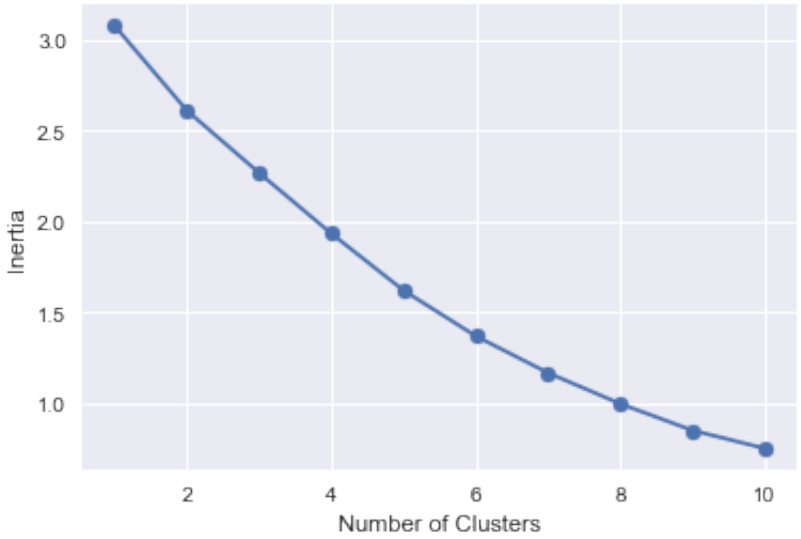

Figure 13: Inertia plot of the ideal cluster number. The example here is for the Munich analysis, but the other plots are nearly identical and thus redundant.

The solution for this was to select a few candidate cluster numbers for each city and show the results on the map to see if there was any insightful result. Most were either just a very large city centre with some outliers, or too many clusters of only one or two neighbourhoods. The final results decided upon for each city do, fortunately, provide some nice insight as there are a few large clusters with some outliers, which gives a good middle ground between the extremes. With the clusters constructed, they could then be used to group the most representative neigh-

bourhoods for each, and their most commonly visited venues, to give a feel as to what defines each cluster. These cluster rankings are displayed in Figures 14 to 17 below.

Cluster Labels	Neighbourhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Milbertshofen-Am Hart	48.195994	11.567868	Museum	Bus Stop	Pastry Shop	Metro Station	Afghan Restaurant
2	Untergiesing-Harlaching	48.192805	11.643526	Zoo Exhibit	Tram Station	Vietnamese Restaurant	Trattoria/Osteria	Supermarket
3	Neuhausen-Nymphenburg	48.156452	11.519305	Italian Restaurant	Canal	Afghan Restaurant	Paper / Office Supplies Store	Nightclub
4	Aubing-Lochhausen-Langwied	48.164726	11.408340	Design Studio	Sporting Goods Shop	Pharmacy	Supermarket	Palace
5	Tudering-Riem	48.123764	11.683573	Home Service	Outdoor Gym	Park	Paper / Office Supplies Store	Newsstand
6	Thalkirchen-Obersendling-Forstenried-Fürstenri...	48.086525	11.513238	Bus Stop	Arts & Crafts Store	Trail	Paper / Office Supplies Store	Newsstand
7	Feldmoching-Hasenbergl	48.215176	11.521979	German Restaurant	Greek Restaurant	Motorcycle Shop	Palace	Newsstand

Figure 14: Venue category popularity ranking for each cluster and representative neighbourhood in Munich.

Cluster Labels	Neighbourhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Wipkingen	47.408739	8.577778	Swiss Restaurant	Swiss Restaurant	Swiss Restaurant	Vietnamese Restaurant	Swiss Restaurant
2	Hirzenbach	47.400156	8.585452	Tram Station	Steakhouse	Accessories Store	Other Great Outdoors	Moroccan Restaurant
3	Wollishofen	47.422893	8.599741	Supermarket	Tram Station	Supermarket	Tram Station	Trail
4	Höngg	47.408293	8.495744	Bus Station	Steakhouse	Soccer Field	Accessories Store	Optical Shop
5	Weinegg	47.395503	8.569414	Tram Station	Wine Shop	Tram Station	Tram Station	Tram Station

Figure 15: Venue category popularity ranking for each cluster and representative neighbourhood in Zurich.

Cluster Labels	Neighbourhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Veyrier	46.254665	6.203664	Restaurant	Spa	Trail	Sandwich Place	Sandwich Place
2	Perly-Certoux	46.158084	6.087883	Soccer Field	Airport	Shoe Store	Sandwich Place	Salad Place
3	Versoix	46.347291	6.279121	Winery	Tram Station	Vineyard	Tram Station	Smoke Shop
4	Soral	46.148108	6.047346	Brewery	Airport	Shoe Store	Sandwich Place	Salad Place
5	Dardagny	46.198490	5.989070	Winery	Paper / Office Supplies Store	Sandwich Place	Salad Place	Restaurant

Figure 16: Venue category popularity ranking for each cluster and representative neighbourhood in Geneva.

Cluster Labels	Neighbourhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Waltham Forest	51.594033	0.221108	Supermarket	Yoga Studio	Pub	Yoga Studio	Trail
2	Barnet	51.616027	-0.210017	Athletics & Sports	Stables	African Restaurant	Pizza Place	Optical Shop
3	Westminster	51.650995	0.140355	Women's Store	Supermarket	Sandwich Place	Wine Bar	Supermarket
4	Barking and Dagenham	51.545277	0.133528	Construction & Landscaping	Lake	Home Service	Plaza	Pakistani Restaurant
5	Harrow	51.597723	-0.341267	Indian Restaurant	Kitchen Supply Store	Clothing Store	Coffee Shop	African Restaurant
6	Ealing	51.522475	-0.331026	Rugby Pitch	Sports Club	Park	Train Station	African Restaurant
7	Kingston upon Thames	51.387906	-0.286900	Garden	Theater	Coffee Shop	Colombian Restaurant	Optical Shop
8	Brent	51.558556	-0.267821	Music Store	Scenic Lookout	Warehouse Store	Food Court	African Restaurant

Figure 17: Venue category popularity ranking for each cluster and representative neighbourhood in London.

With a nice selection of clusters and knowing what they each represent, we can visualise these to get a final look at what the narrowed down options are. These clusters were all added to the same maps as before, to see how they are distributed relative to each other as well as the city centres. These new maps are displayed in below.

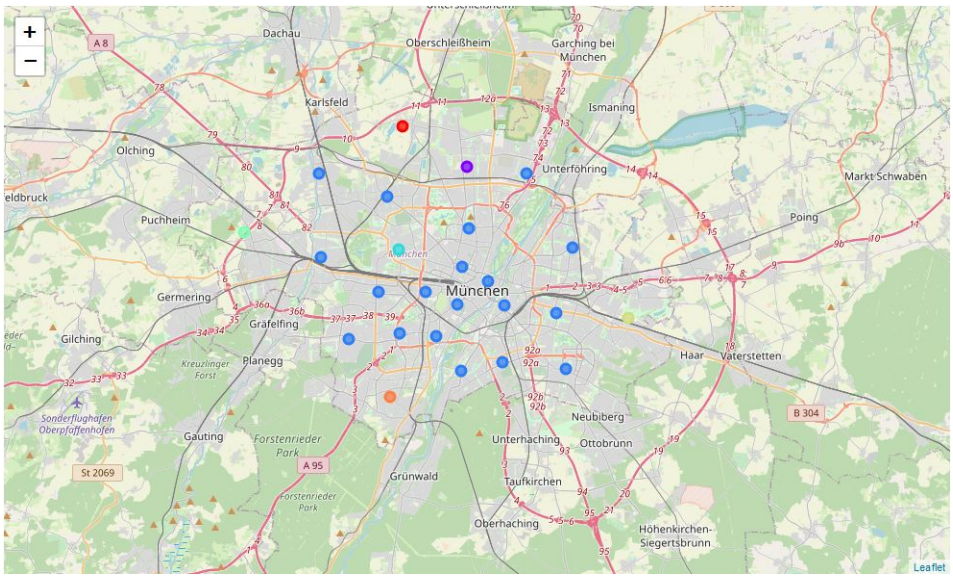

Figure 18: Map of Munich, with the clusters of neighbourhoods grouped by colour.

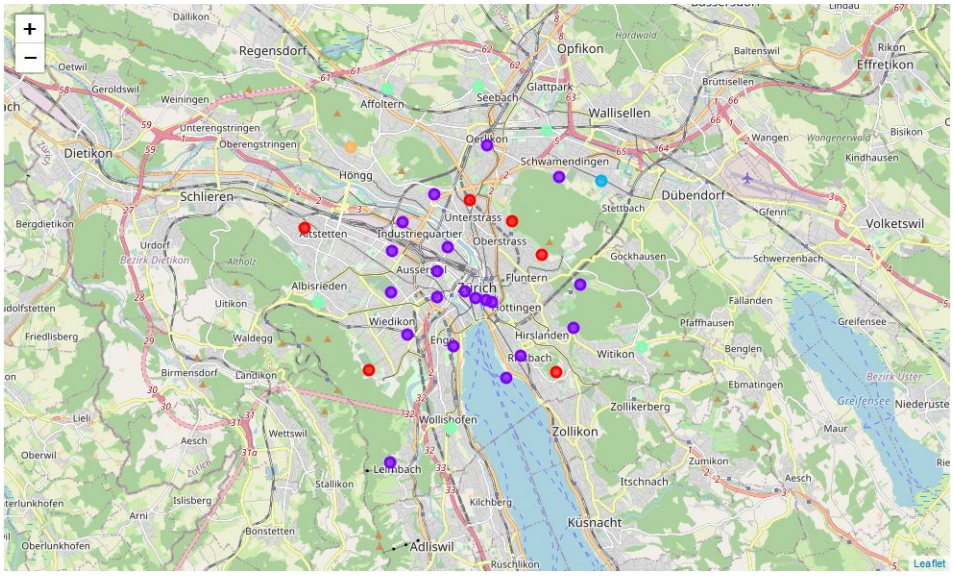

Figure 19: Map of Zurich, with the clusters of neighbourhoods grouped by colour.

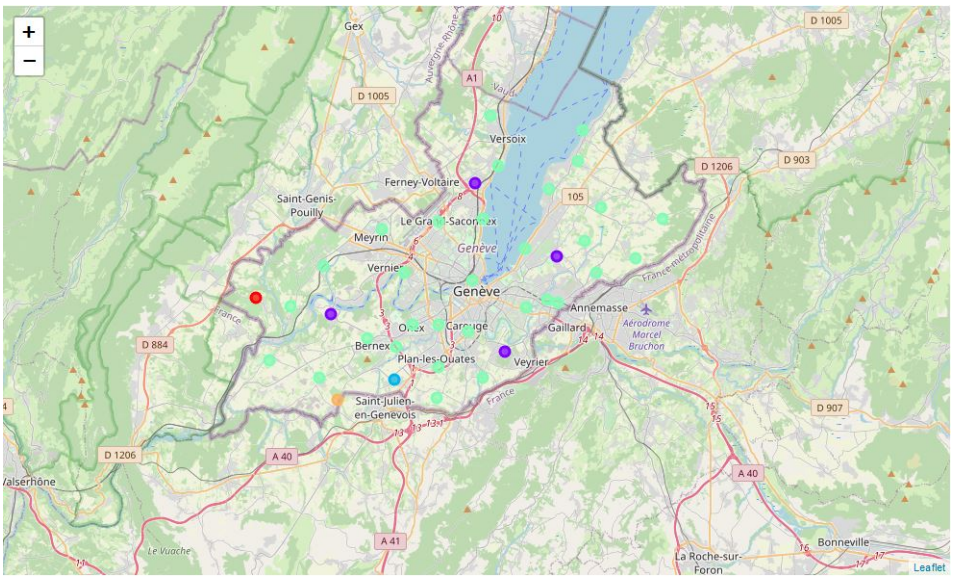

Figure 20: Map of Geneva, with the clusters of neighbourhoods grouped by colour.

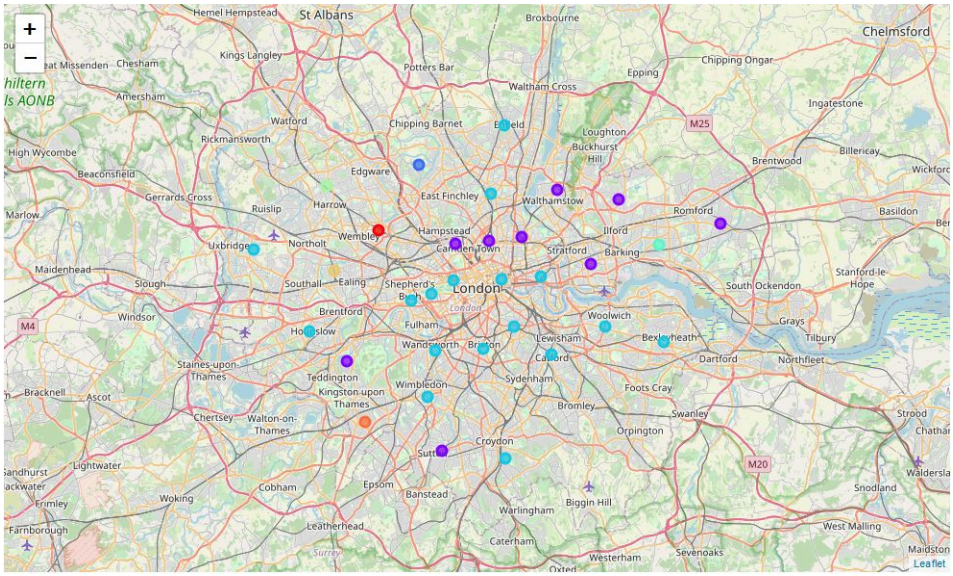

Figure 21: Map of London, with the clusters of neighbourhoods grouped by colour.

4 Discussion

As may be expected, each city has one dominant cluster around the city centre, with the rest being distributed around the outskirts. Each map does, however, provide some nice insight into the decision to be made about which neighbourhood is ideal, although the results may seem trivial. For example, in the case of Munich in Figure 18, we have the case of one dominant cluster with the rest of the clusters being single-entry neighbourhoods. While this may seem a useless outcome, it rather shows that there is not much variation in what each neighbourhood has to offer, and in general they can all be considered to have the same range of venue categories, so there is not much benefit in that sense of being closer or further from the city centre. So, one could choose to stay closer to the outskirts for whatever reason, probably living costs, and not suffer from lack of venues. A similar sentiment is echoed through the results of the other

cities, although there is a bit more of a gradient between the clusters. For example, in Zurich in Figure 19, the second most dominant cluster does have a sizeable population in terms of contained neighbourhoods, and as such fits well for a step away from the city centre with still a few options within the same cluster which could fit well to the browser's needs. This also works well as a perspective as the clusters are not grouped solely with proximity, as can be seen in Geneva in Figure 20, which has a few locations in the purple cluster, with some being inland and some being lakeside, which offers nice variation in options for someone who is more suited to the venue range in that cluster.

5 Conclusion

For this project, four cities were investigated for their prospective 'student life' and the associated metrics for that, to provide an insightful snapshot at a glance for what each city has to offer and what kinds of areas would be ideal for a prospective newcomer who may not be that well acquainted with what each city has to offer. The baseline distribution of the cities was made through a large dataset of neighbourhood locations, coupled with a python mapping package, which gave a good first glance at each city. The dataset was then enriched with the Foursquare API which provided more details about venues and venue categories in each city and neighbourhood. Together, all of these data were used in tandem to give a quick one-line glance at which neighbourhoods provide what the user may be looking for, in the form of a 'top 10' list of most popular venues. Finally, a clustering model was fir to each city to find clusters of similar neighbourhoods, to narrow down the options even further to make it easily digestible for the user to make simple decisions about which area and city seems most appealing to them before endeavouring with deeper research.

As a program designed to give a first look at each city with a consistent and comprehensive analysis process, the project achieved all that it set out to, and has proved to be personally very useful to myself, and I hope that means a similar result will follow for others. The project is, however, only meant to be a first glance, and as such does lack some more detailed metrics which may be useful in further research, such as the living costs of each neighbourhood and the relative distance to the university in question. These useful metrics can make for even better analysis and feeling for each city, but would be delegated to other programs built upon this foundation for each of those purposes. In further spirit of generality with this project, beyond the first mapping of the neighbourhoods, the data has not been directed at relating specifically to universities, so it could be specialised to work for anyone immigrating to a new country and looking for the best area to move to.