

THE BATTLE OF THE NEIGHBOURHOODS

THE HEARTS OF EUROPE

Dean Brand
github.com/deanbrand
2021

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.

BUSINESS PROBLEM

- 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.
- 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.

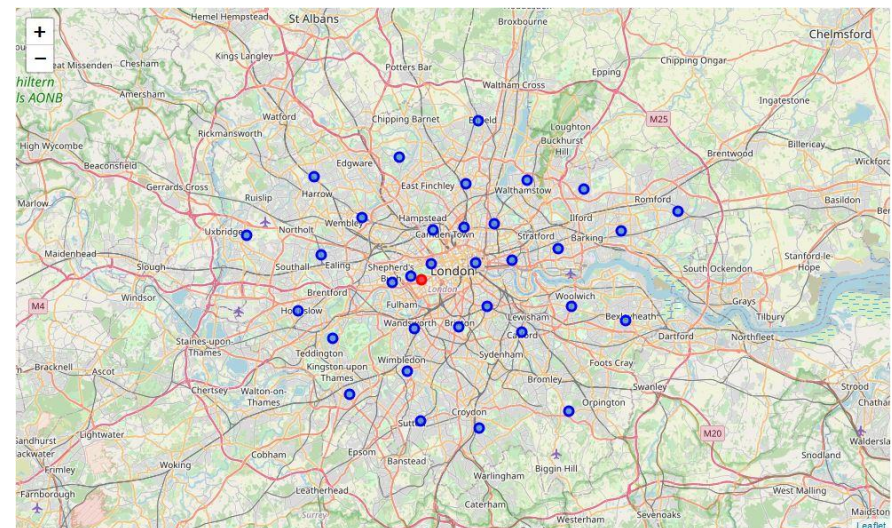
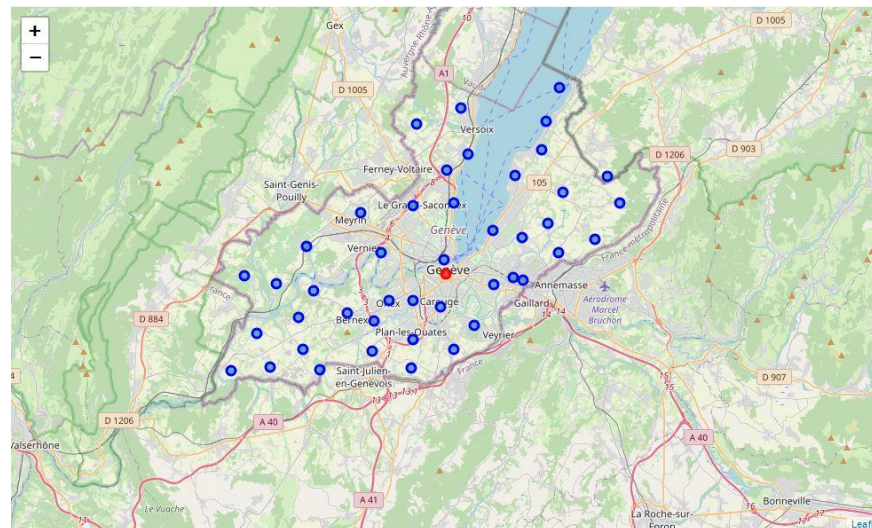
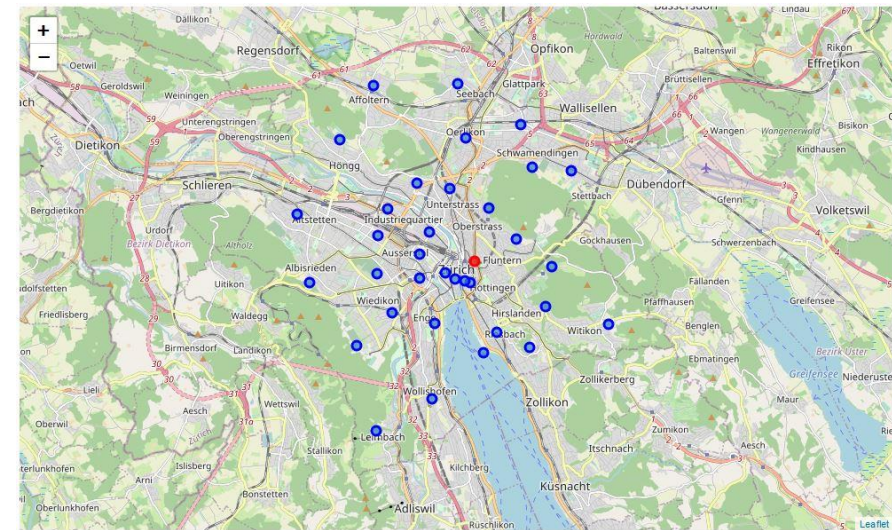
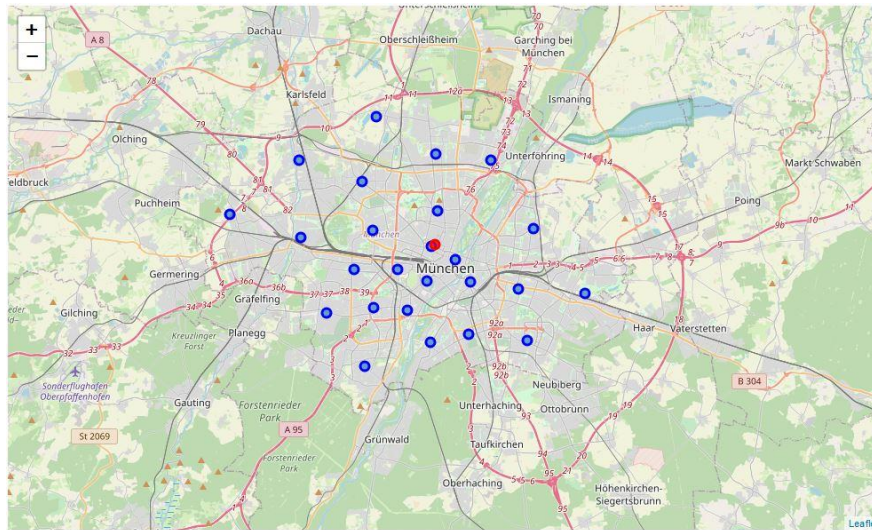
DATA COLLECTION

- 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>).
- The particular file which is used for each city is a geojson file, which provides the neighbourhood name, neighbourhood group, and map shape of each neighbourhood. This location data is then coupled with 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.

METHODOLOGY

- The data could easily be extracted from each geojson file into a geopandas data frame, and from there cleaned into the desired form of having the neighbourhood name, latitude, and longitude.
- Once that data frame had been constructed, the location data of the city in question, and the university, or other institution of interest, was obtained allowing for the mapping to begin.
- The maps created with folium present an overview of each city with the distributed neighbourhoods clearly marked for easy visualisation.

METHODOLOGY



METHODOLOGY

- Making use of the Foursquare API, all of the venue data for each neighbourhood could be easily collected and stitched onto the data frame.
- The most important features are that of the venue name and category, which could then be encoded into numerical values to be used in the machine learning modelling.

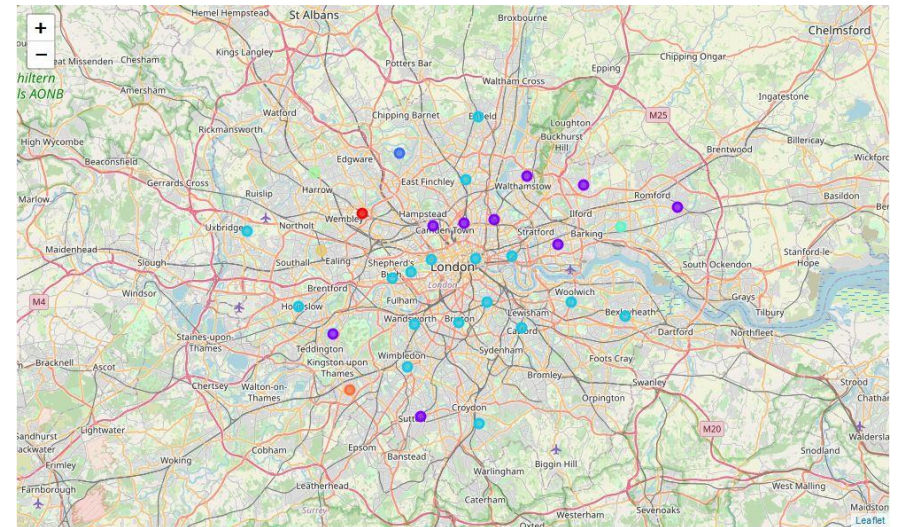
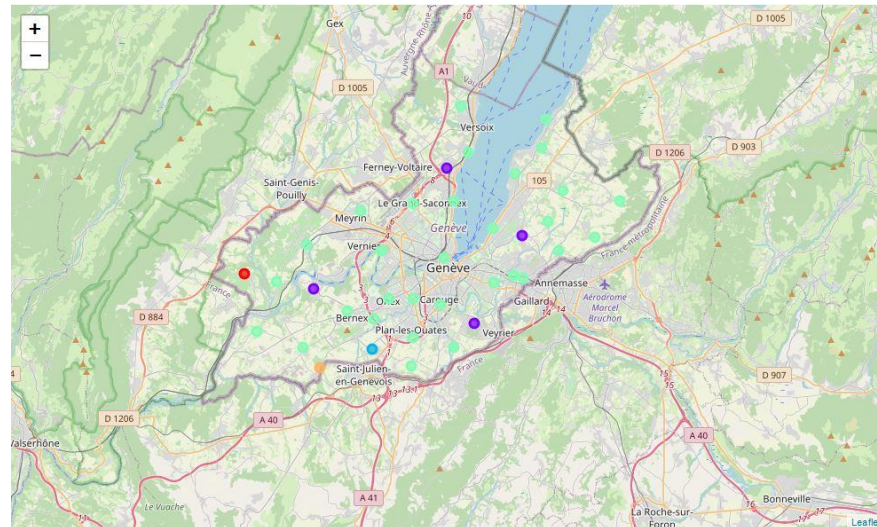
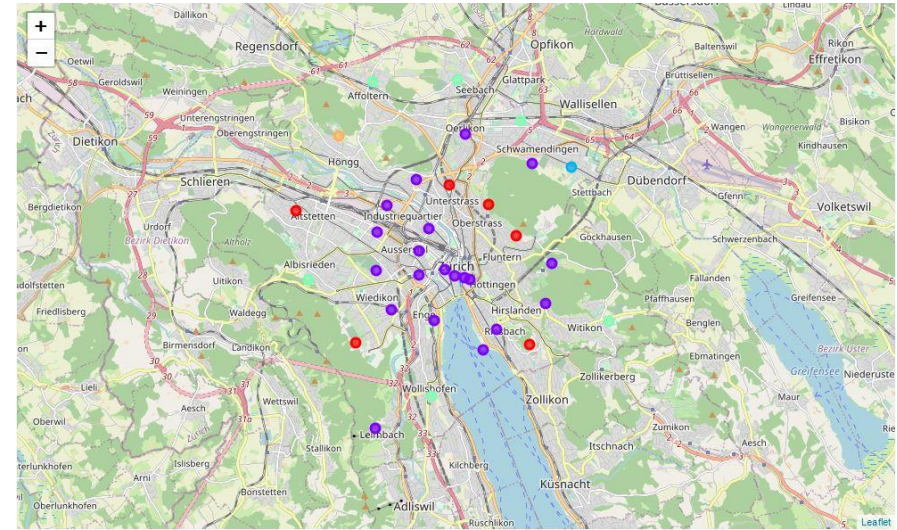
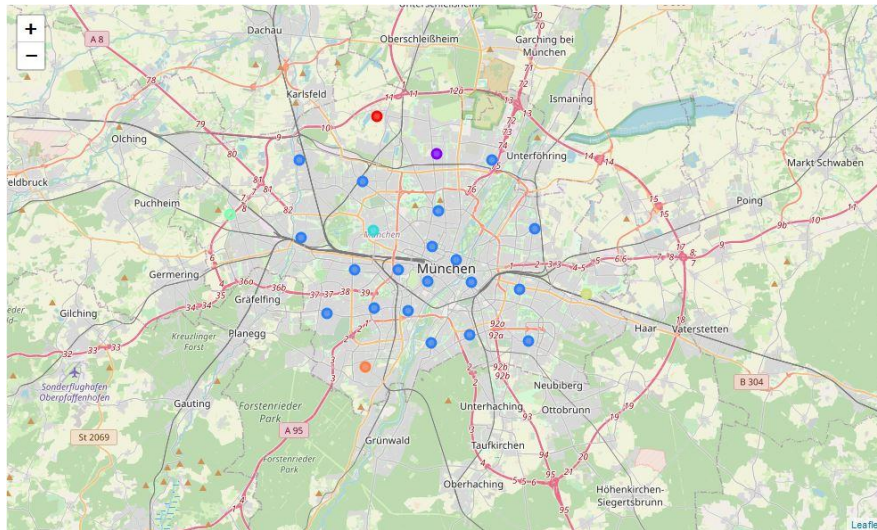
	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
0	Altstadt-Lehel	48.141273	11.583178	SEITZ Trattoria	Trattoria/Osteria
1	Altstadt-Lehel	48.141273	11.583178	Liebighof im Lehel	German Restaurant
2	Altstadt-Lehel	48.141273	11.583178	Kitcho	Japanese Restaurant
3	Altstadt-Lehel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

METHODOLOGY

- Knowing what kind of data we are dealing with, we can begin to train a model to fit the data and give more insight.
- The model chosen for this project is the K-means clustering model.
- Each city had a model fit to it for an optimised number of clusters.

	Neighbourhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Cluster Labels								
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-Hasenberg	48.215176	11.521979	German Restaurant	Greek Restaurant	Motorcycle Shop	Palace	Newsstand

METHODOLOGY



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 fit 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.

CONCLUSION

- 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.

THANK YOU!

Dean Brand
github.com/deanbrand
2021