Comparison of Neighborhoods
Berlin – London – New York

Markus Gorges

November 12, 2020

1. Introduction

1.2 General

Three of the biggest and most wanted cities in the world are Berlin, London and New York. They are the first cities that come to everybodys mind when you think about diversity and a place to be. But what do the cities have in common when we talk about possibilities for free time activities. What can you do after work and what would be a ranking of boroughs to live to have everything right at your footstep?

This analysis will be a helper for a decision where to move if you have the opportunity to live in one of these cities.

1.3 Personal Interest

This analysis is for a dear friend who has job offers in these three cities. Hopefully this analisis will help him decide which city to choose. Regarding this the analysis will be done considering his personal interests and make it as personal as possible.

2. Data

2.2 Data Source

For the analysis three different sources will be taken. So different aspects of gathering data will be shown. And also the data was found in different formats on different web pages.

2.2.1 Berlin

The Data for Berlin will come from a normal web page. The page was found through a simple search in the internet.

Adress: http://www.places-in-germany.com/14356-places-within-a-radius-of-15km-around-berlin.html

The Data will be scraped with a tool called BeautifulSoup and directly transformed into a Pandas DataFrame. It needs some cleaning and gathering only the boroughs of the city.

2.2.2 London

The Data for London will be provided through an Excel-File that was found on a different web page. The file can directly be accessed and imported into Pandas.

Adress: https://data.london.gov.uk/download/london-borough-profiles/80647ce7-14f3-4e31-b1cd-d5f7ea3553be/london-borough-profiles.xlsx

2.2.3 New York

The Data for New York is stored in a JSON File also on a web server. The JSON File will be processed with the JSON library.

Adress: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

2.2.4 Geo Data

The Geo Data such as longitude and latitude of the cities, Boroughs and neighborhoods come from the service Geolocator and Nominatim from

Open Street Maps. All Adresses are processed through a simple loop and added to the borough and neighborhood data.

2.2.5 Venue Data

The Data for the venues in the cities will be provided through Foursquare. The received JSON-File through the request will be processed and combined the the previous Data.

2.3 Data Cleaning

Each of the data sources will be stored into a separate Pandas DataFrame. For each city the boroughs and neighborhoods will be combined with the geo data and the venues. Multiple entries will be deleted and also rows with missing data. All can be done through the Pandas package.

In the Data I will concentrate of only the inner circle of each city, so the outer boroughs will not be considered.

Especially for the Berlin data I had some more data cleaning to do. The scrapped table from the web page didn't have the right form. So the zip-code of the boroughs had to be separated from the borough itself. For me the best work around was to generate a little loop to iterate through the rows and split the values into zip-code and borough name. All other rows could be deleted.

For the other cities the data handling was a bit easier because the source was better to handle.

3. Methodology

3.2 Berlin

To fit the given problem most I decided to sit together with my friend and see, what venues he is mostly interested and which venues does he want to have most around his new home. For that I examined the venue categories from Foursquare for the cities and made a list of venues he likes and don’t likes. This was for me the best way to fit the later results most to the given business problem.

With the list of non-interests I was able to erase all the venues he isn’t interested and who have no impact in deciding where to go. The filtering concept that was used was to filter for the string components of the interests and non-interests. So we could separate from Italian restaurants and Asian restaurants for example.

I started with the analysis of Berlin. Here I decided to only go for the boroughs because they aren’t as big as compared to New York. In a first attempt I checked for each borrow what the most occurrence of a venue is.

I started with a plot using Folium to see an overview of the city with all the boroughs.

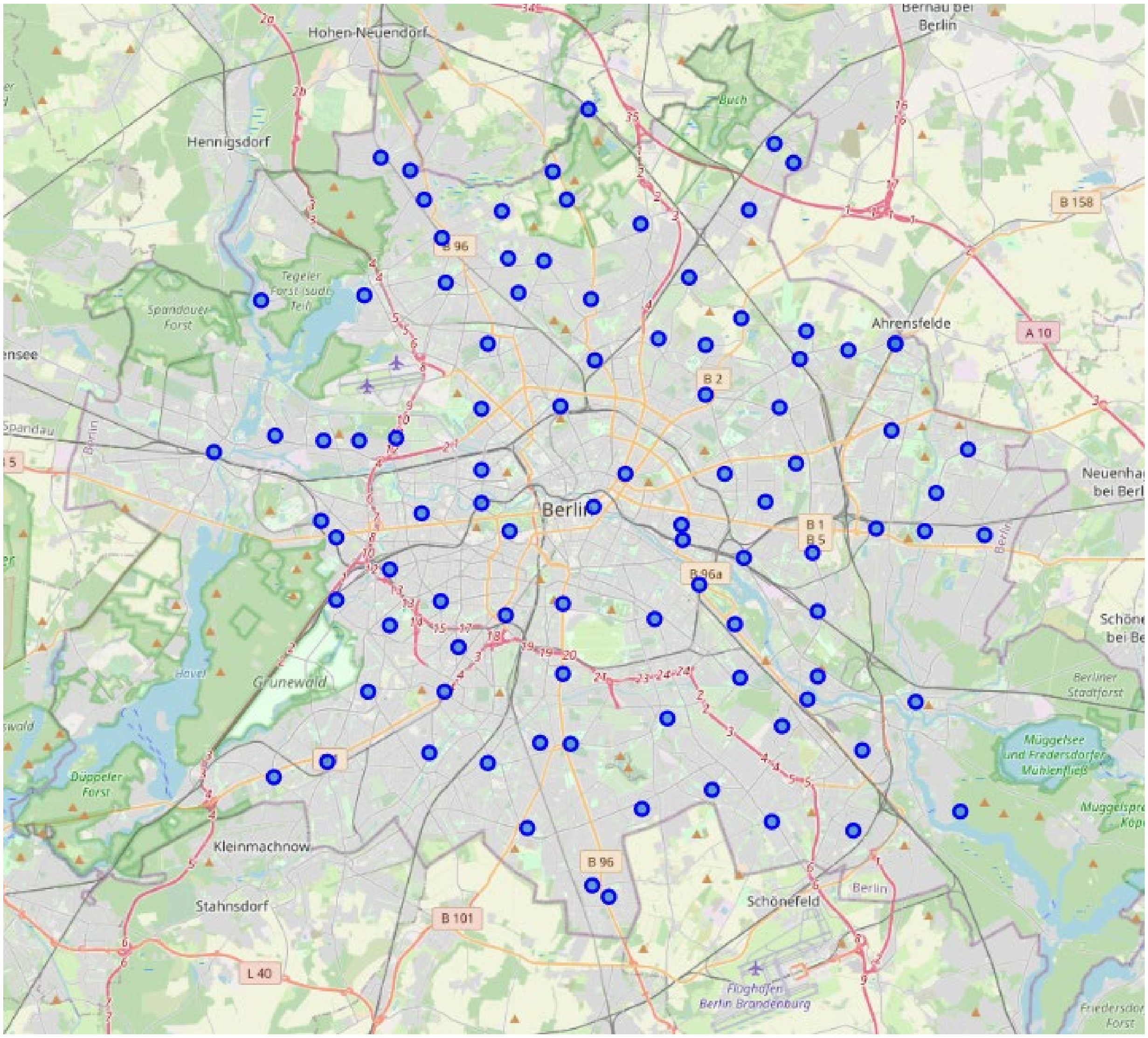
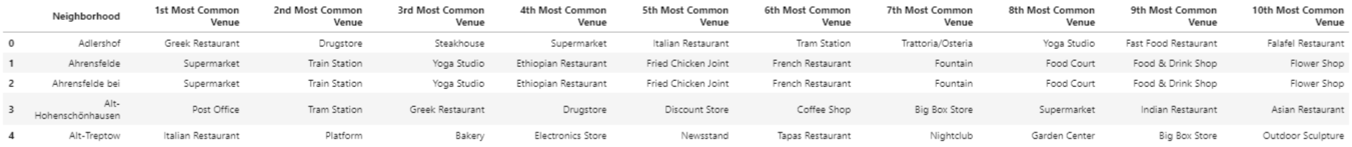


Figure 1 Overview Berlin with Boroughs

With addition of the venue data I decided to have a look at the most occurrent venues of each borough.



	Neighborhood	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	Adlershof	Greek Restaurant	Drugstore	Steakhouse	Supermarket	Italian Restaurant	Tram Station	Trattoria/Osteria	Yoga Studio	Fast Food Restaurant	Falafel Restaurant
1	Ahrensfelde	Supermarket	Train Station	Yoga Studio	Ethiopian Restaurant	Fried Chicken Joint	French Restaurant	Fountain	Food Court	Food & Drink Shop	Flower Shop
2	Ahrensfelde bei	Supermarket	Train Station	Yoga Studio	Ethiopian Restaurant	Fried Chicken Joint	French Restaurant	Fountain	Food Court	Food & Drink Shop	Flower Shop
3	Alt-Hohenschönhausen	Post Office	Tram Station	Greek Restaurant	Drugstore	Discount Store	Coffee Shop	Big Box Store	Supermarket	Indian Restaurant	Asian Restaurant
4	Alt-Treptow	Italian Restaurant	Platform	Bakery	Electronics Store	Newsstand	Tapas Restaurant	Nightclub	Garden Center	Big Box Store	Outdoor Sculpture

Figure 2 top head of DataFrame with most occurent venues

Next step I decided to do a clustering of the city regarding the venues and to see, what differences are between the boroughs. The algorithm that was used for this was KMeans. I started with 5 clusters and it went well. More

ore fewer clusters didn’t satisfy me. Here a plot of the clusters on a Folium Map:

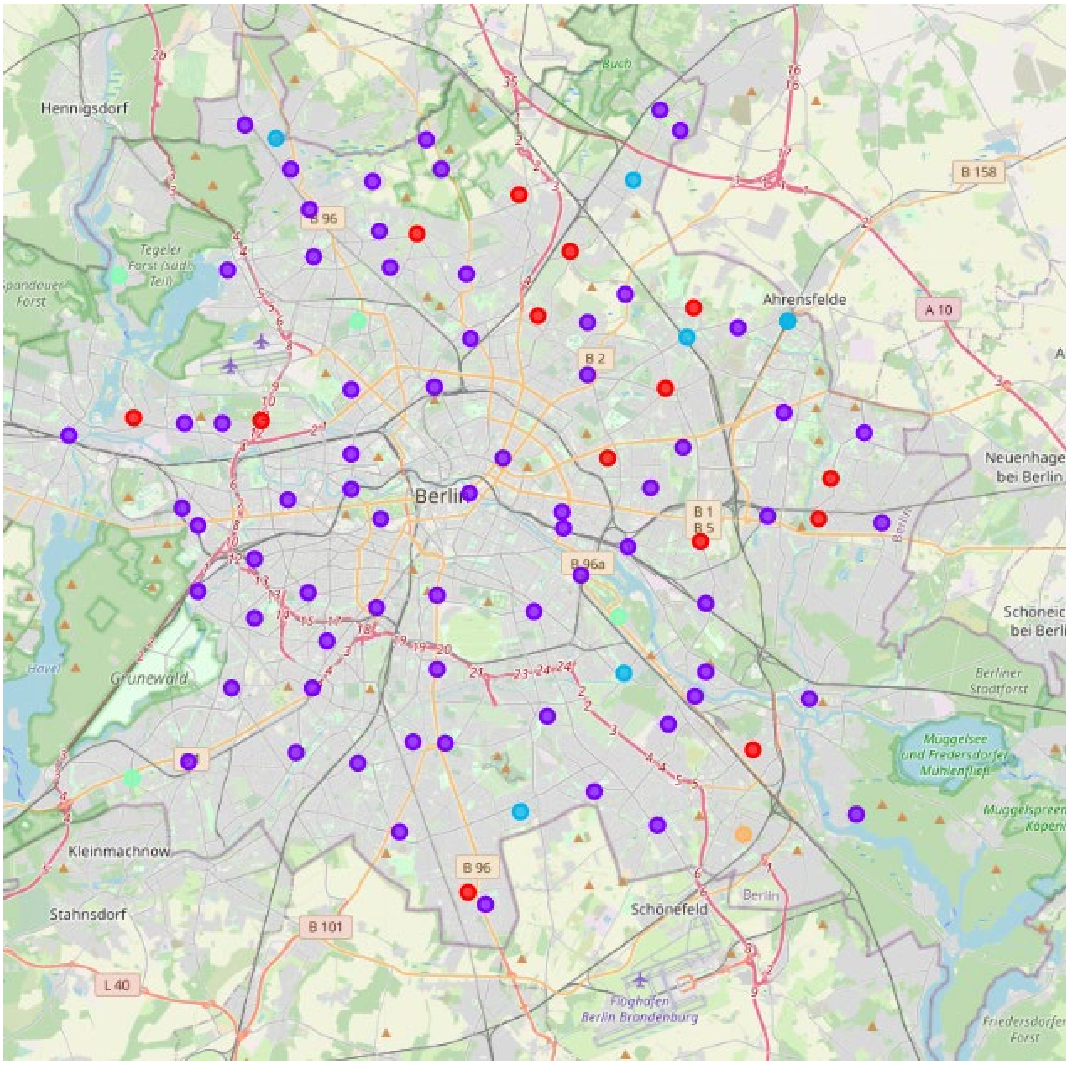
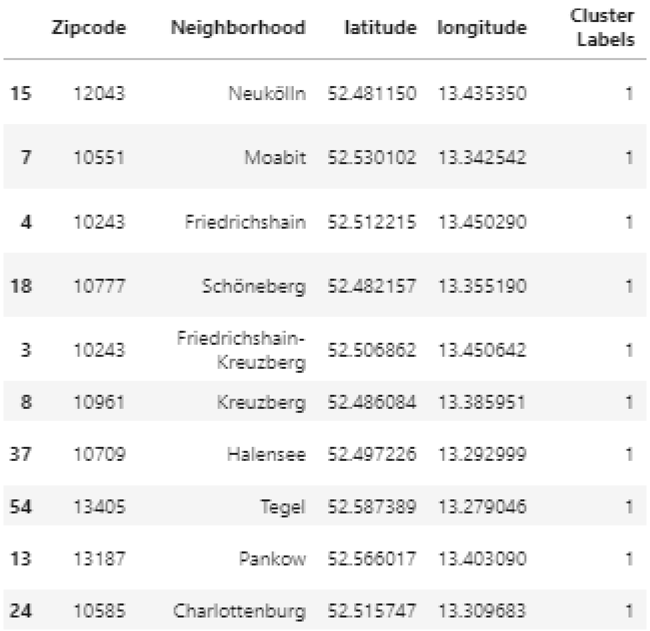


Figure 3 Folium plot with clusters of KMeans

For a first attempt I would recommend to go with the purple cluster.

In a second attempt I decided to bring in the interests of my friend. For this I calculated how much a borough fits his interests and give back the top ten boroughs.



	Zipcode	Neighborhood	latitude	longitude	Cluster Labels
15	12043	Neukölln	52.481150	13.435350	1
7	10551	Moabit	52.530102	13.342542	1
4	10243	Friedrichshain	52.512215	13.450290	1
18	10777	Schöneberg	52.482157	13.355190	1
3	10243	Friedrichshain-Kreuzberg	52.506862	13.450642	1
8	10961	Kreuzberg	52.486084	13.385951	1
37	10709	Halensee	52.497226	13.292999	1
54	13405	Tegel	52.587389	13.279046	1
13	13187	Pankow	52.566017	13.403090	1
24	10585	Charlottenburg	52.515747	13.309683	1

Figure 4 Top 10 regarding interests

All these boroughs are also in the purple cluster which I recommended earlier.

3.3 London

For London I went the same way to come to a point to do a comparison of the both cities.

Also an overview in Folium:

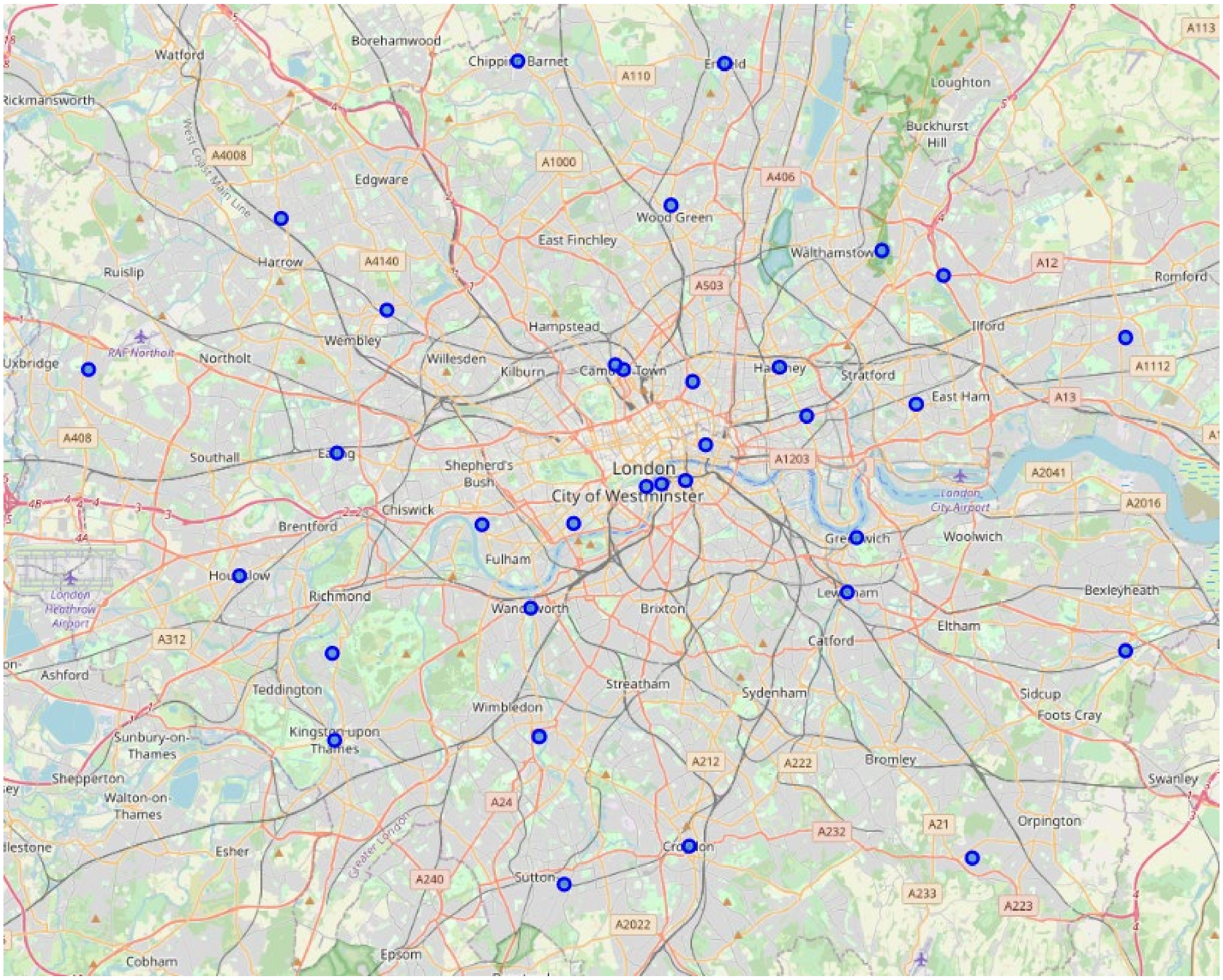


Figure 5 Overview of London

Same clustering algorithm than Berlin and applying the clusters to Folium:

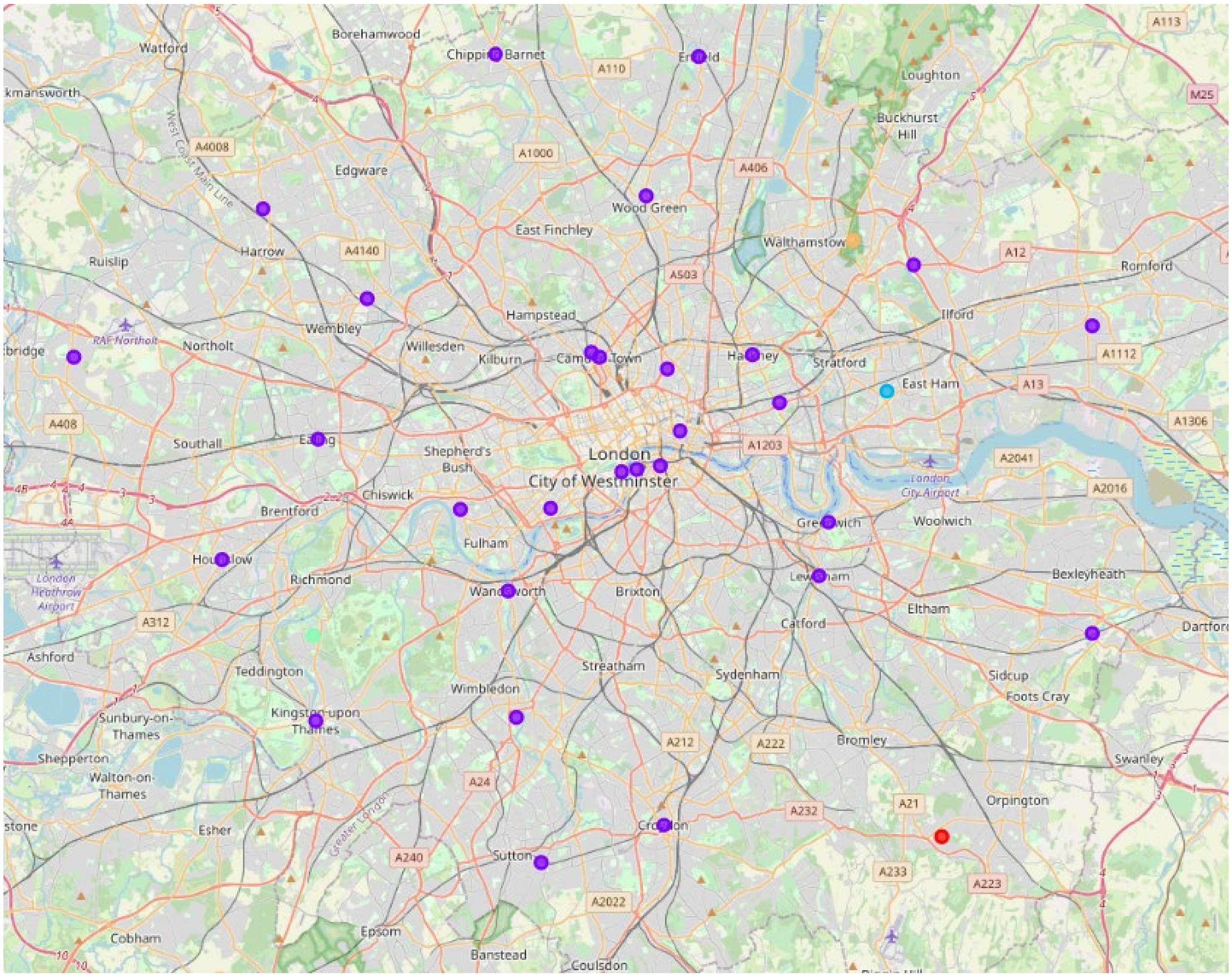
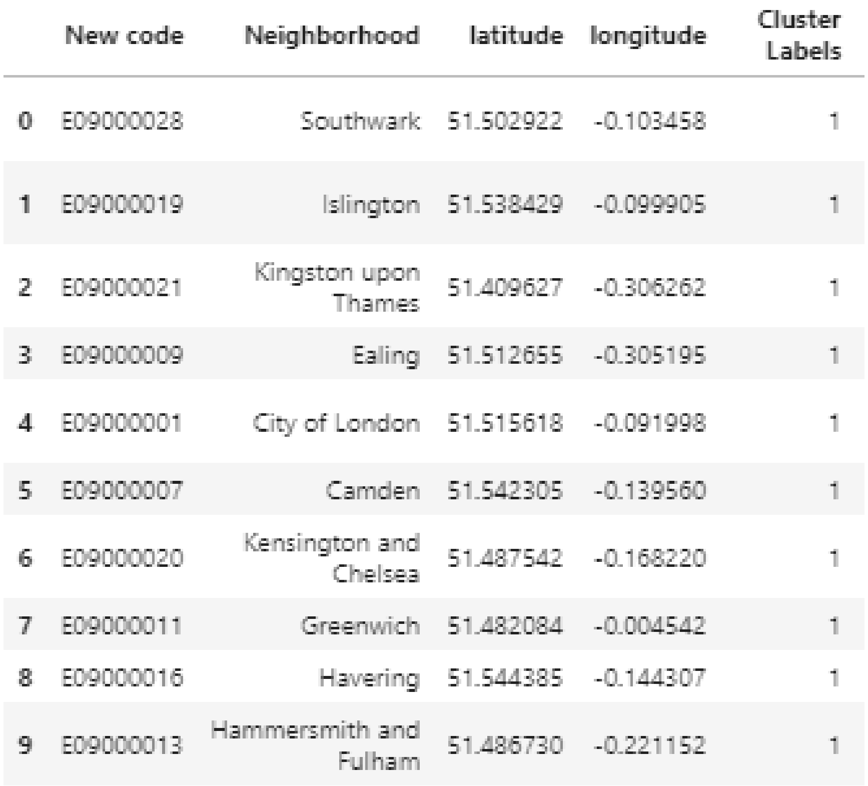


Figure 6 Overview London with clusters

By applying the interests to the venue categories it made more sense to go here for this selection because the clusters were very indifferent.

The top ten list looks like this:



	New code	Neighborhood	latitude	longitude	Cluster Labels
0	E09000028	Southwark	51.502922	-0.103458	1
1	E09000019	Islington	51.538429	-0.099905	1
2	E09000021	Kingston upon Thames	51.409627	-0.306262	1
3	E09000009	Ealing	51.512655	-0.305195	1
4	E09000001	City of London	51.515618	-0.091998	1
5	E09000007	Camden	51.542305	-0.139560	1
6	E09000020	Kensington and Chelsea	51.487542	-0.168220	1
7	E09000011	Greenwich	51.482084	-0.004542	1
8	E09000016	Havering	51.544385	-0.144307	1
9	E09000013	Hammersmith and Fulham	51.486730	-0.221152	1

Figure 7 Top 10 London

3.4 New York

Also for New York I went the same way and first started with an overview in Folium. But for this city I decided to go also for the neighborhoods.

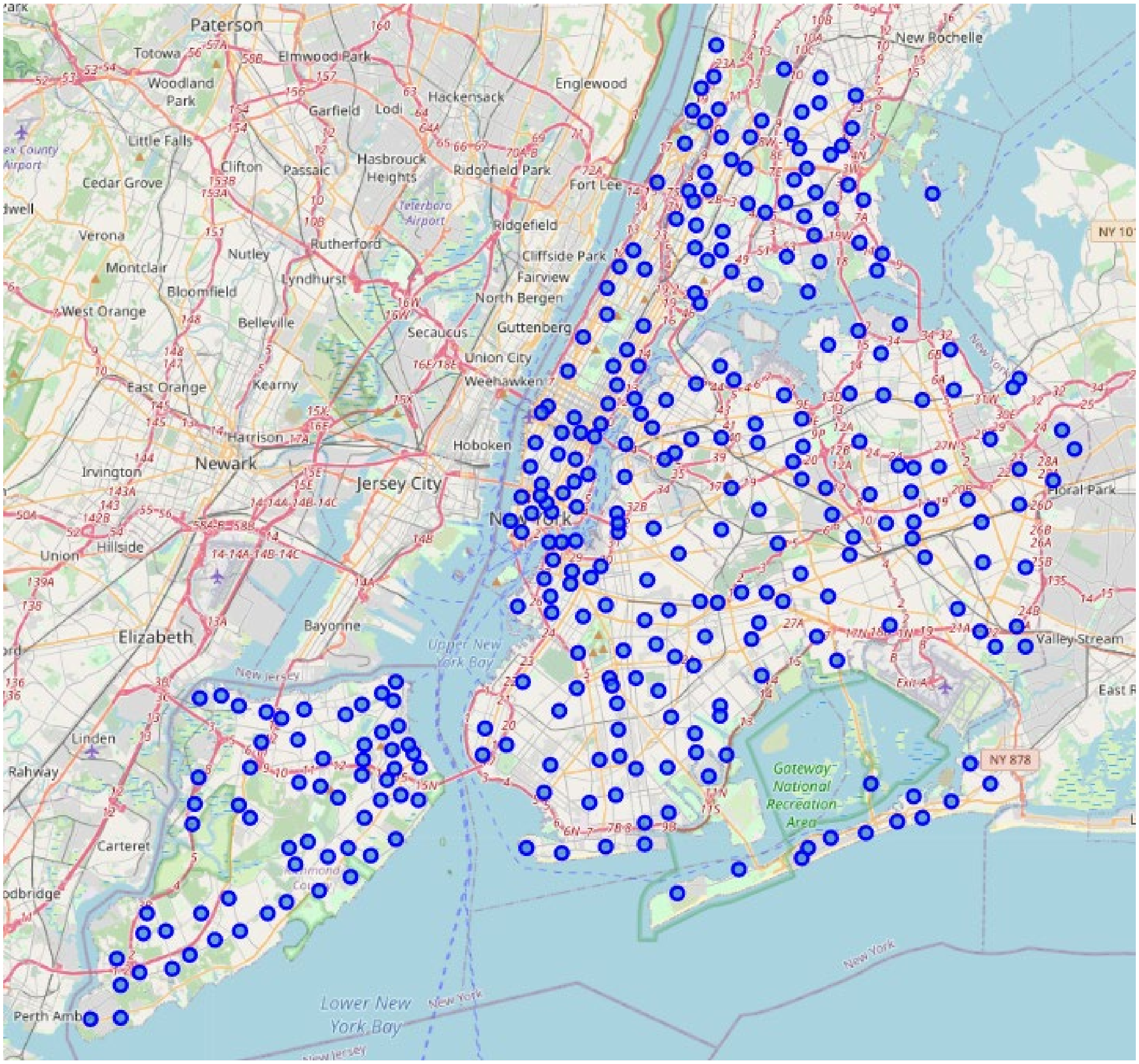


Figure 8 Overview of New York

With the applying of the KMeans algorithm it seemed also a bit indifferent in the beginning but with some tweeks it came out a lot better than London.

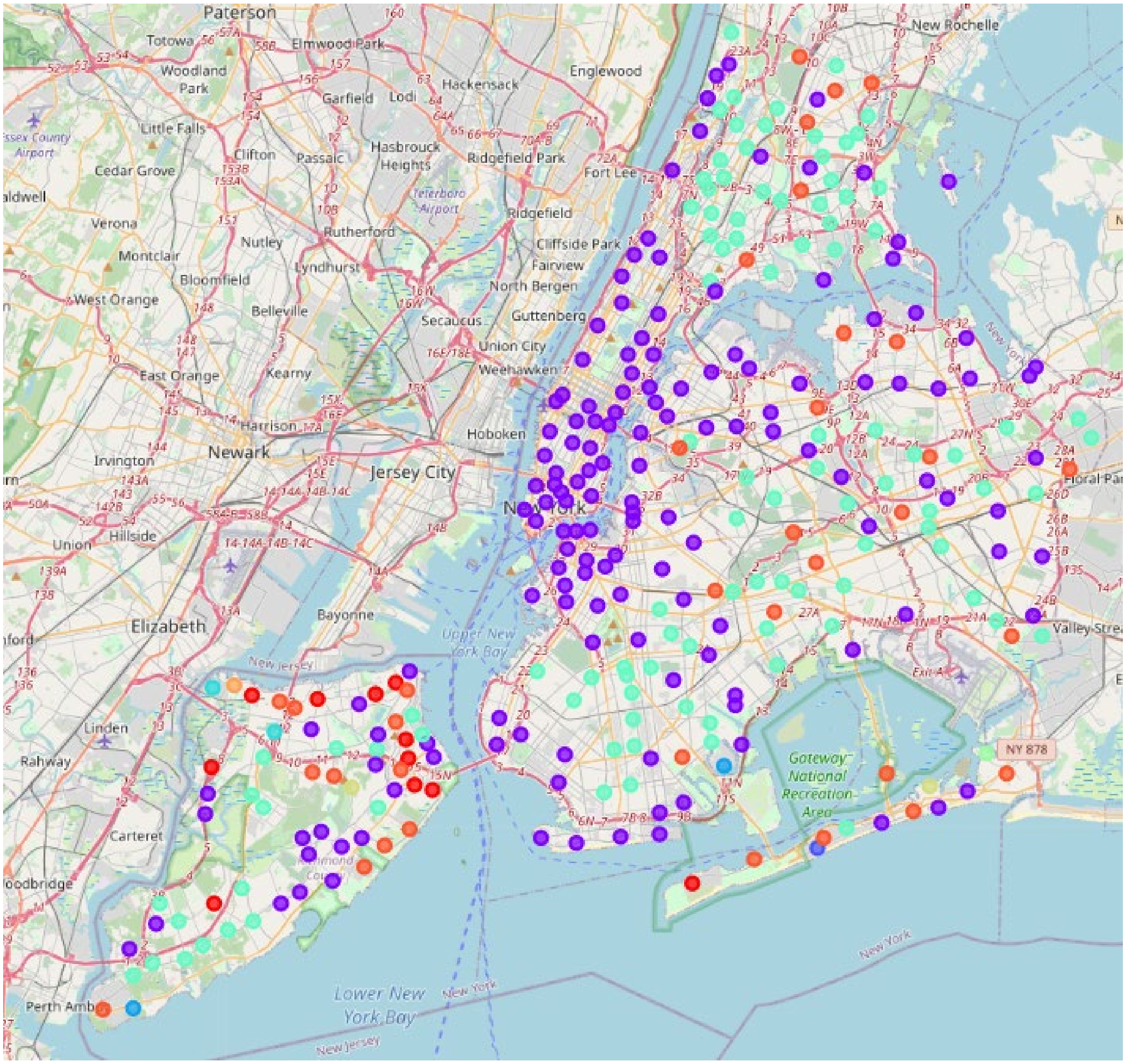
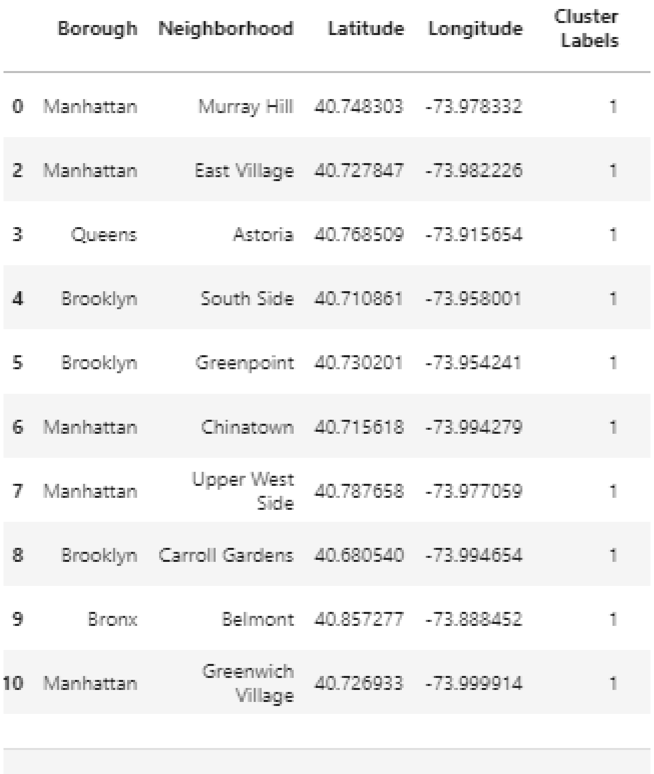


Figure 9 Clustering of New York

Also here I would recommend the purple cluster as the most interesting.

After applying the interests I got following top ten list:



	Borough	Neighborhood	Latitude	Longitude	Cluster Labels
0	Manhattan	Murray Hill	40.748303	-73.978332	1
2	Manhattan	East Village	40.727847	-73.982226	1
3	Queens	Astoria	40.768509	-73.915654	1
4	Brooklyn	South Side	40.710861	-73.958001	1
5	Brooklyn	Greenpoint	40.730201	-73.954241	1
6	Manhattan	Chinatown	40.715618	-73.994279	1
7	Manhattan	Upper West Side	40.787658	-73.977059	1
8	Brooklyn	Carroll Gardens	40.680540	-73.994654	1
9	Bronx	Belmont	40.857277	-73.888452	1
10	Manhattan	Greenwich Village	40.726933	-73.999914	1

Figure 10 Top 10 list of New York

4. Results

After going through all these data sets it seemed that the usage of KMeans and the addition of the interests and sorting of fitting neighborhoods and boroughs was the right attempt to go for a conclusion.

For each city we now know which are the most interesting clusters and so boroughs and neighborhoods. And according to the interests we are able to give an advice to my friend what would be the best fitting neighborhood in each city.

5. Discussion & Conclusion

With the given data and results I would recommend my friend to go to New York. And go for one of the top ten neighborhoods examined through the data analysis. Here he will find the biggest amount of diversity in his interests and has lots of venues to explore. As a second recommendation I would give him the advice to have a look at Berlin. Regarding the data it is more likely to live in Berlin than London.