Data science capstone project – The battle of the Neighborhoods:

Introduction/Problem definition:

A businessman has lived in New York, United states and now lives in Toronto, Canada. He has been asked to work for Fictive incorporation in Europe. They have offices in Paris and Berlin. He has a beautiful wife and two children. In New York and Toronto they have lived in nice neighborhoods. He does want to work in Europe, but in a similar Neighbourhood. In this project we will similarities between neighborhoods in the different cities.

Data

Data that will be used:
Data containing boroughs of new York:
https://cocl.us/new_york_dataset

Data containing boroughs of Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Data containing boroughs of Berlin:
https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin#External_links

Data containing boroughs of Paris:
https://en.wikipedia.org/wiki/Arrondissements_of_Paris

Additional Data:
Additional data will be gathered by using the Foursquare API.

Use of data:

In this project we will gather venue information about the different Boroughs and Neighbourhoods in the four different cities. This will be done by using the foursquare API. Some of the borough data has already been collected, other data will be collected by use of Beautifulsoup and/or pandas.

Methodology

Collecting and preparing of data

New York
The dataset of New York could be downloaded from the link and already contained the following values:
Borough, Neighborhood, Latitude, Longitude. The dataset of New York could be downloaded as JSON and had to be turned into a pandas dataframe.

Toronto
The dataset of Toronto could be scraped into a Pandas dataframe and contained the following values:
Postcode, Borough, Neighborhood. The dataset of Toronto contained a two abnormalities:
- Missing boroughs ("not assigned")
- Missing Neighborhoods ("not assigned")

The missing boroughs are excluded in this project. The missing neighborhoods are named after the borough and are included in this report.

Berlin
The dataset of Berlin was gathered from Wikipedia pages and contained the following values:**Postcode, Borough, Neighborhood.**

Paris

The dataset of Paris was gathered from Wikipedia pages and contained the following values:**Postcode, Borough, Neighborhood.**

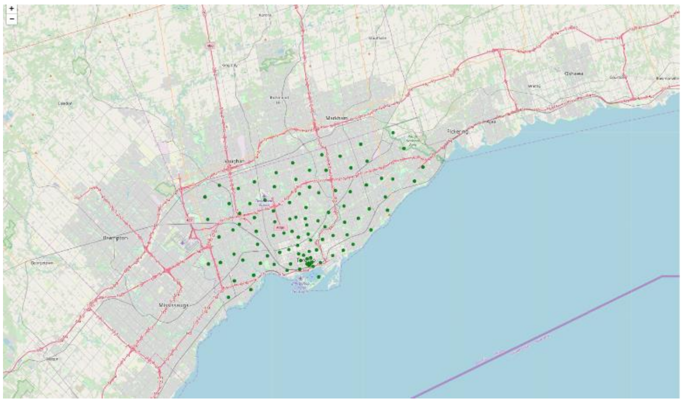
To work with this data, all datasets should contain at least Neighborhood, borough and Postcode. The new York dataset does not contain Postcode, but already has coordinates. The datasets of Paris, berlin and Toronto are missing coordinates. I added coordinates by using the geopy library and using the ArcGis geocoder. By defining a function that loops through the dataframe "latitude" and "longitude" were added to the datasets.

	postcode	Borough	Neighbourhood	Address	Latitude	Longitude
0	0401	Charlottenburg-Wilmersdorf	Charlottenburg	Charlottenburg-Wilmersdorf, Charlottenburg, Be...	52.51915	13.30639
1	0402	Charlottenburg-Wilmersdorf	Wilmersdorf	Charlottenburg-Wilmersdorf, Wilmersdorf, Berli...	52.48976	13.31519
2	0403	Charlottenburg-Wilmersdorf	Schmargendorf	Charlottenburg-Wilmersdorf, Schmargendorf, Ber...	52.51915	13.30639
3	0404	Charlottenburg-Wilmersdorf	Grunewald	Charlottenburg-Wilmersdorf, Grunewald, Berlin,...	52.49963	13.32297
4	0405	Charlottenburg-Wilmersdorf	Westend	Charlottenburg-Wilmersdorf, Westend, Berlin, G...	52.49963	13.32297

(figure 1) figure 1: example of added latitude and longitude

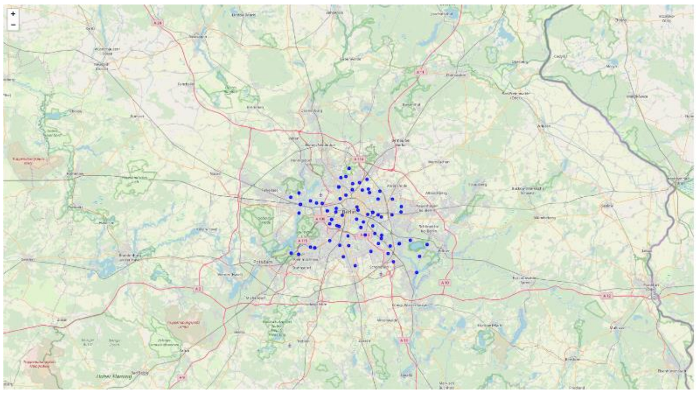
After adding the coordinates to the datasets it became possible to plot all neighborhoods in the different cities by using the folium library.

Toronto (green dots)

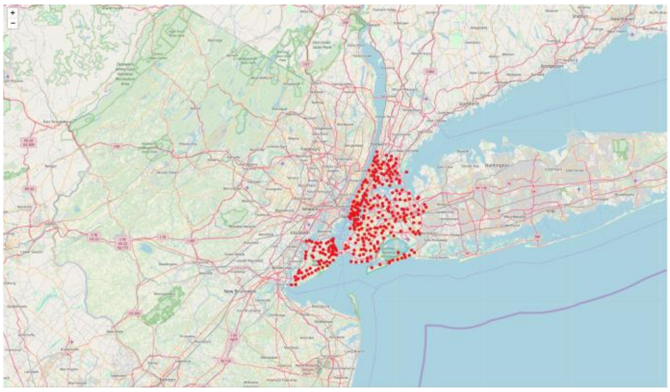


Paris (yellow dots)

Berlin (blue dots)



New York (red dots)



Foursquare data collection:

After setting up the Foursquare API, a GET request was made to find venues in a radius of 500 m with a limit of 50 venues per location. The GET request delivers a JSON structure, which had to be transformed into a pandas dataframe.

Figure 3: example of found venues

	Unnamed: 0	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	0	Parkwoods	43.75242	-79.329242	Brookbanks Park	43.751976	-79.332140	Park
1	1	Parkwoods	43.75242	-79.329242	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	2	Parkwoods	43.75242	-79.329242	Corrosion Service Company Limited	43.752432	-79.334661	Construction & Landscaping
3	3	Parkwoods	43.75242	-79.329242	TTC stop - 44 Valley Woods	43.755402	-79.333741	Bus Stop
4	4	Victoria Village	43.73060	-79.313265	Wigmore Park	43.731023	-79.310771	Park

Onehot encoding and grouping

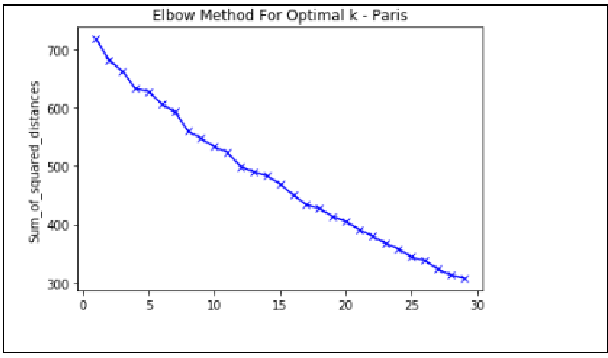
To use the venues later as input for the K-means machine learning algorithm, all categorical values in the dataframe were converted by using onehot encoding. After the encoding I made a dataframe which contained the 10 most common venues that were found by Foursquare.

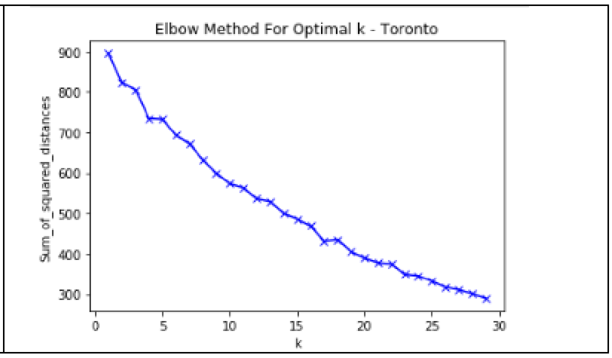
Figure 4: example of most common venues.

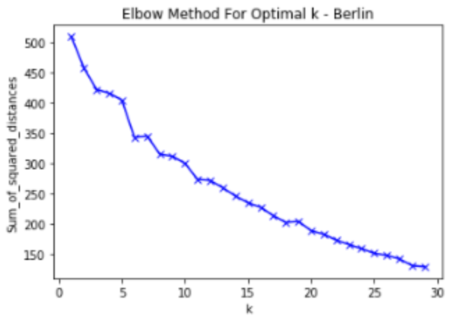
NeighborhoodName	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Allerton	Pizza Place	Chinese Restaurant	Cosmetics Shop	Supermarket	Deli / Bodega	Playground	Discount Store	Grocery Store	Gas Station	Breakfast Spot
Annadale	Pizza Place	Restaurant	Train Station	Dance Studio	American Restaurant	Deli / Bodega	Sports Bar	Pharmacy	Sushi Restaurant	Diner
Arden Heights	Deli / Bodega	Pharmacy	Bus Stop	Coffee Shop	Home Service	Pizza Place	Yoga Studio	Filipino Restaurant	Event Space	Exhibit
Arlington	Bus Stop	Intersection	American Restaurant	Deli / Bodega	Yoga Studio	Fish Market	Exhibit	Factory	Falafel Restaurant	Farm
Arrochar	Bus Stop	Italian Restaurant	Bagel Shop	Deli / Bodega	Liquor Store	Pizza Place	Hotel	Taco Place	Middle Eastern Restaurant	Athletics & Sports

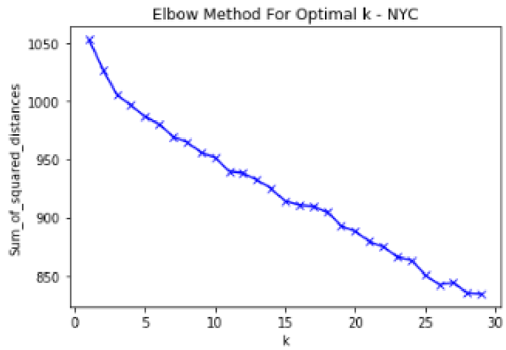
Elbow, silhouette and hierarchical dendograms

After processing al the data into a usable format, I used the elbow method, the silhouette method and a dendogram to find the best K for the k-means algorithm.

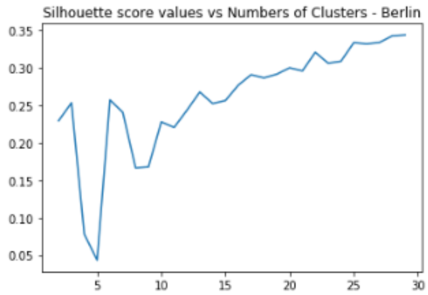




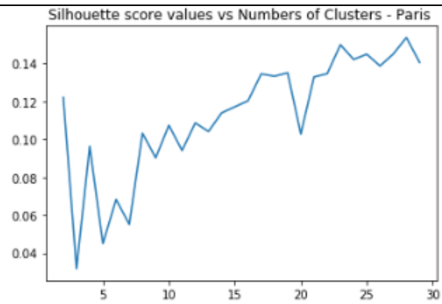




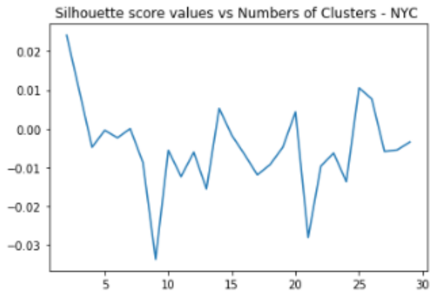
The Elbow method was, in my opinion not very conclusive, so I decided to use the silhouette method to find the best K.



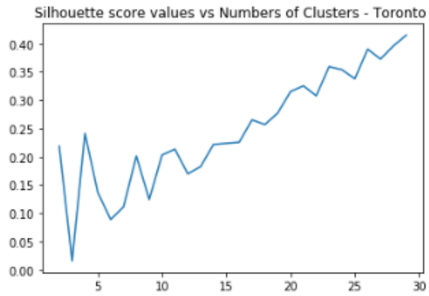

```
Optimal number of components is:
29
```




```
Optimal number of components is:
28
```

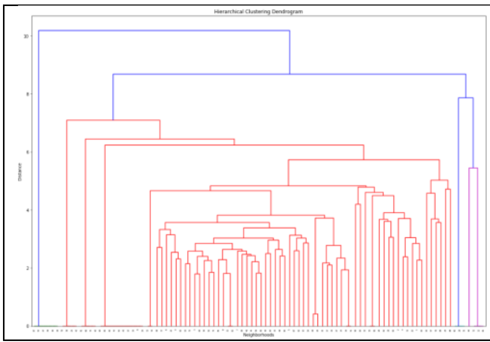


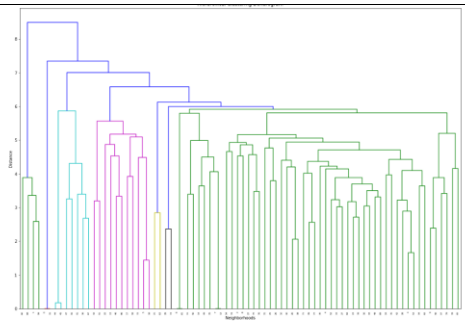

```
Optimal number of components is:
2
```

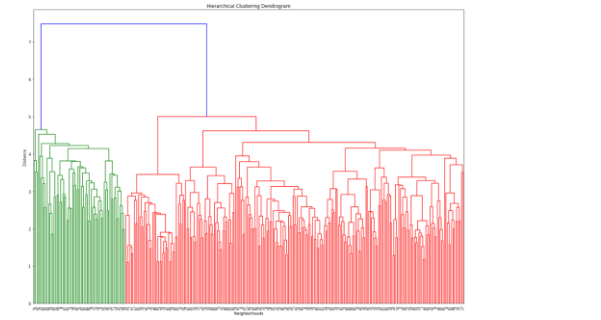


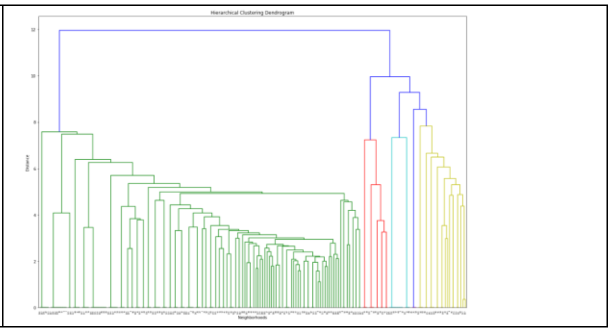

```
Optimal number of components is:
29
```

Because I wanted to see how it looked in a more visual format, I tried to make dendrograms:





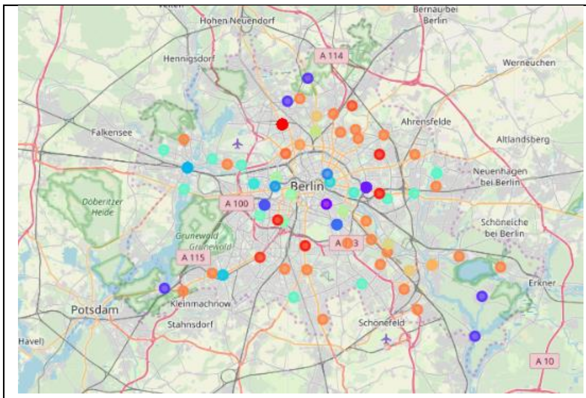




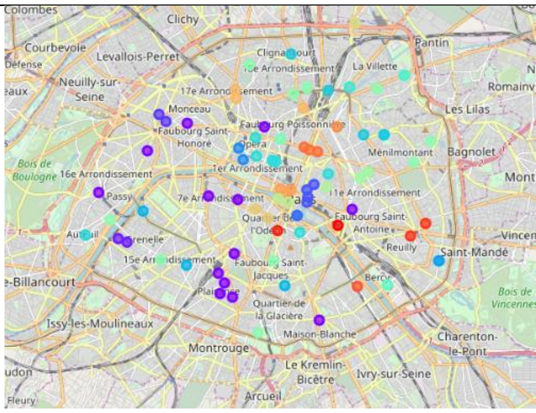
Results

results per city

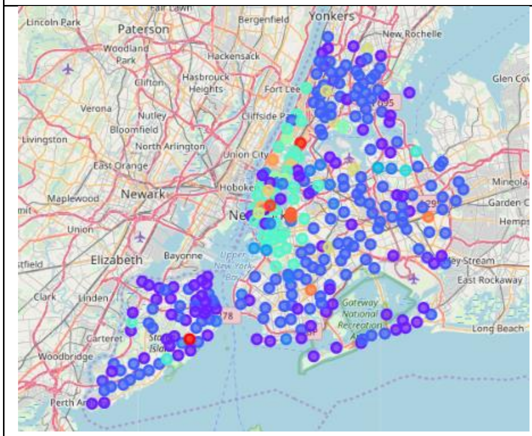
I first wanted to see the results per city to get an impression of the K-means clustering. As we can see in the image, different clusters are plotted on the map. I used the silhouette plots as input for the number of K's.



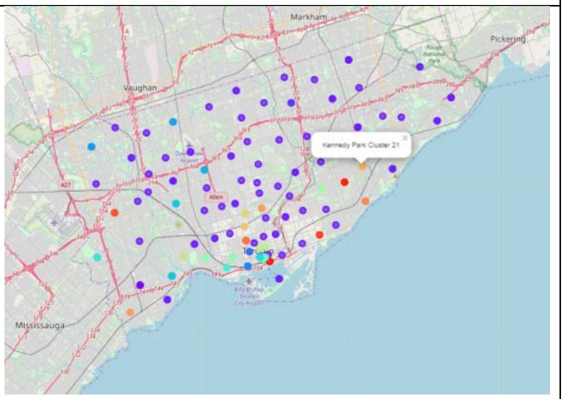
Berlin



Paris



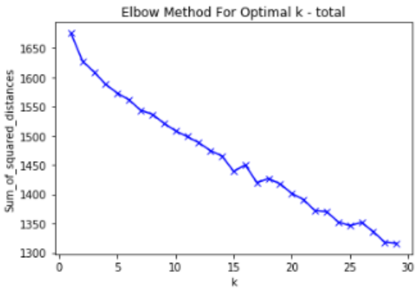
NYC



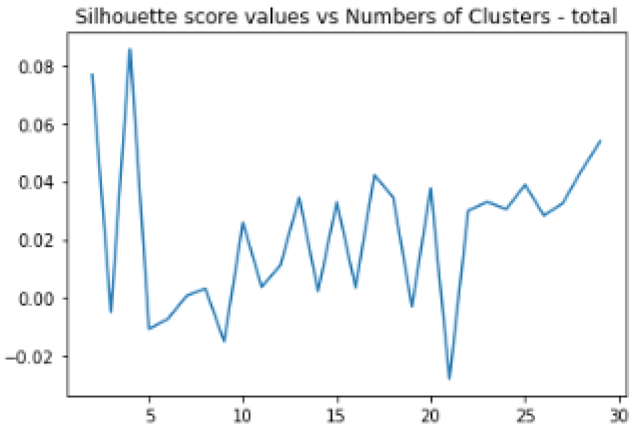
Toronto

Results total

We want to see if there are (dis)similarities between cities. So the total results are more important. I repeated the same steps to create a total dataframe, which contains all values of all four cities. Again I used the elbow, silhouette and dendogram method to determine the optimal number of K's.



Again, the elbow method is a bit inconclusive .




```
Optimal number of components is:
4
```

The silhouette method determined that the optimal number of clusters would be 4. In the end, I picked 16 as the number of clusters, because I expected that 4 clusters would be to superficial.

```
Total number of neighborhoods in cluster 0 is 79
Total number of neighborhoods in cluster 1 is 5
Total number of neighborhoods in cluster 2 is 254
Total number of neighborhoods in cluster 3 is 6
Total number of neighborhoods in cluster 4 is 5
Total number of neighborhoods in cluster 5 is 5
Total number of neighborhoods in cluster 6 is 84
Total number of neighborhoods in cluster 7 is 128
Total number of neighborhoods in cluster 8 is 14
Total number of neighborhoods in cluster 9 is 6
Total number of neighborhoods in cluster 10 is 19
Total number of neighborhoods in cluster 11 is 2
Total number of neighborhoods in cluster 12 is 15
Total number of neighborhoods in cluster 13 is 26
Total number of neighborhoods in cluster 14 is 6
Total number of neighborhoods in cluster 15 is 10
```

In the figure above the distribution can be seen throughout the clusters. I will only look at Cluster 0, 2, 6,7 and 13. They contain the most values.

Cluster 0: French dining area.

	NeighborhoodName	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
16	Arsenal	French Restaurant	Cocktail Bar	Thai Restaurant	Canal Lock	Gastropub	Camera Store	Garden	Boat or Ferry	Vegetarian / Vegan Restaurant	Supermarket
17	Arts-et-Métiers	French Restaurant	Hotel	Seafood Restaurant	Vegetarian / Vegan Restaurant	Beer Store	Thai Restaurant	Bar	Japanese Restaurant	Vietnamese Restaurant	Coffee Shop
25	Batignolles	French Restaurant	Bar	Wine Bar	Hotel	Pub	Bakery	Bookstore	Gym / Fitness Center	Beer Bar	Noodle House
42	Belleville	Bar	French Restaurant	Café	Chinese Restaurant	Restaurant	Pizza Place	Dim Sum Restaurant	Coffee Shop	Cocktail Bar	Hostel
45	Bercy	Hotel	Italian Restaurant	French Restaurant	Japanese Restaurant	Bar	Museum	Wine Bar	Cosmetics Shop	Burger Joint	Multiplex
...	...	...	...	...	...	...	...	...	...	...	...
618	Vivienne	French Restaurant	Bistro	Clothing Store	Italian Restaurant	Wine Bar	Japanese Restaurant	General College & University	Historic Site	Garden	Lebanese Restaurant
632	Westend	Italian Restaurant	Café	Dessert Shop	Hotel	French Restaurant	Vietnamese Restaurant	Thai Restaurant	Gourmet Shop	Greek Restaurant	Seafood Restaurant
647	Wilmersdorf	Hotel	French Restaurant	Plaza	Café	Grocery Store	Doner Restaurant	Drugstore	Market	Bavarian Restaurant	Park
662	École-Militaire	Bar	French Restaurant	Japanese Restaurant	Bakery	Café	Park	Italian Restaurant	Music Venue	Theater	Bistro
663	Épinettes	French Restaurant	Hotel	Creperie	Japanese Restaurant	Comedy Club	Auvergne Restaurant	Theater	Seafood Restaurant	Belgian Restaurant	Bakery

79 rows × 11 columns

This cluster contains mostly French restaurants in different cities.

Cluster 2: Common area

[886]:

	NeighborhoodName	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
1	Adlershof	Supermarket	Tanning Salon	Greek Restaurant	Steakhouse	Italian Restaurant	Bank	Plaza	Trattoria/Osteria	Drugstore	Tram Station
5	Alderwood	Pub	Convenience Store	Gym	Athletics & Sports	Yoga Studio	Falafel Restaurant	Fast Food Restaurant	Farmers Market	Farm	Exhibit
8	Alt-Treptow	Italian Restaurant	Park	Coffee Shop	Café	Dessert Shop	Farm	Bistro	Hookah Bar	Hostel	Bakery
9	Altglienicke	Supermarket	German Restaurant	Indian Restaurant	Bowling Alley	Discount Store	Event Service	Event Space	Ethiopian Restaurant	Exhibit	Film Studio
13	Arden Heights	Pharmacy	Pizza Place	Coffee Shop	Home Service	Deli / Bodega	Bus Stop	Yoga Studio	Falafel Restaurant	Farmers Market	Farm
...	...	...	...	...	...	...	...	...	...	...	...
651	Woburn	Korean Restaurant	Park	Indian Restaurant	Business Service	Coffee Shop	Falafel Restaurant	Fast Food Restaurant	Farmers Market	Farm	Yoga Studio
653	Woodbine Heights	Pharmacy	Bus Line	Café	Grocery Store	Doctor's Office	Arts & Crafts Store	Coffee Shop	Middle Eastern Restaurant	Gas Station	Pizza Place
655	Woodlawn	Pub	Deli / Bodega	Pizza Place	Playground	Food Truck	Italian Restaurant	Train Station	Donut Shop	Park	Supermarket
658	York Mills West	Bank	Park	Convenience Store	Speakeasy	Factory	Fast Food Restaurant	Farmers Market	Farm	Falafel Restaurant	Exhibit
660	Yorkville	Italian Restaurant	Coffee Shop	Sandwich Place	Gym	Café	Deli / Bodega	Park	Liquor Store	Mexican Restaurant	Bar

254 rows × 11 columns

This cluster contains a mix of restaurants, bars, supermarkets and banks.

Cluster 6: Food court

```
totalvenues_sorted.loc[totalvenues_sorted['Cluster_Labels'] == 6, totalvenues_sorted.columns[[1] + list(range(2, totalvenues_sorted.shape[1]))]]
```

	NeighborhoodName	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	Adelaide	Coffee Shop	Café	Asian Restaurant	Steakhouse	Seafood Restaurant	Hotel	Gastropub	American Restaurant	Gym	Salad Place
10	Amérique	Bar	French Restaurant	Burger Joint	Japanese Restaurant	Sandwich Place	Theater	Grocery Store	Metro Station	Bike Rental / Bike Share	Gastropub
19	Astoria	Bar	Greek Restaurant	Mediterranean Restaurant	Seafood Restaurant	Middle Eastern Restaurant	Gourmet Shop	Pub	Hookah Bar	Gym	Indian Restaurant
28	Bay Ridge	Spa	Italian Restaurant	Greek Restaurant	Hookah Bar	Grocery Store	Pizza Place	American Restaurant	Bagel Shop	Ice Cream Shop	Chinese Restaurant
36	Bedford Stuyvesant	Bar	Coffee Shop	Pizza Place	Café	Fruit & Vegetable Store	Discount Store	Thrift / Vintage Store	Gourmet Shop	Cocktail Bar	Bagel Shop
...	...	...	...	...	...	...	...	...	...	...	...
617	Vinegar Hill	Food Truck	Café	Wine Shop	Art Gallery	Coffee Shop	Antique Shop	Scenic Lookout	Bakery	Men's Store	Ice Cream Shop
623	Washington Heights	Café	New American Restaurant	Deli / Bodega	Park	Coffee Shop	Tapas Restaurant	Mobile Phone Shop	Bakery	Wine Shop	Breakfast Spot
630	West Village	Italian Restaurant	Cocktail Bar	Cosmetics Shop	American Restaurant	Speakeasy	New American Restaurant	Bakery	Chinese Restaurant	Gastropub	Coffee Shop
642	Williamsburg	Coffee Shop	Bar	Bagel Shop	Yoga Studio	Breakfast Spot	Latin American Restaurant	Tapas Restaurant	Pet Store	Liquor Store	Lounge
648	Windsor Terrace	Deli / Bodega	Grocery Store	Café	Plaza	Diner	Park	Chinese Restaurant	Bar	Coffee Shop	Italian Restaurant

84 rows × 11 columns

This cluster contains mostly restaurants from different nationalities.

Cluster 7: Traffic zone

```
totalvenues_sorted.loc[totalvenues_sorted['Cluster_Labels'] == 7, totalvenues_sorted.columns[[1] + list(range(2, totalvenues_sorted.shape[1]))]]
```

	NeighborhoodName	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
6	Allerton	Pizza Place	Chinese Restaurant	Supermarket	Deli / Bodega	Cosmetics Shop	Spanish Restaurant	Breakfast Spot	Spa	Fried Chicken Joint	Playground
11	Annadale	Pizza Place	Dance Studio	Diner	Sports Bar	Restaurant	Pharmacy	American Restaurant	Sushi Restaurant	Deli / Bodega	Train Station
15	Arrochar	Bus Stop	Italian Restaurant	Liquor Store	Bagel Shop	Deli / Bodega	Pizza Place	Food Truck	Polish Restaurant	Middle Eastern Restaurant	Outdoors & Recreation
23	Bath Beach	Chinese Restaurant	Pizza Place	Pharmacy	Sushi Restaurant	Italian Restaurant	Asian Restaurant	Bubble Tea Shop	Fast Food Restaurant	Donut Shop	Surf Spot
30	Baychester	Discount Store	Donut Shop	Pizza Place	Men's Store	Electronics Store	Mexican Restaurant	Baseball Field	Bank	Pet Store	Fast Food Restaurant
...	...	...	...	...	...	...	...	...	...	...	...
628	West Farms	Bus Station	Bus Stop	Park	Metro Station	Diner	Hotel	Coffee Shop	Bank	Sandwich Place	Pizza Place
631	Westchester Square	Fast Food Restaurant	Pharmacy	Donut Shop	Sandwich Place	Pizza Place	Latin American Restaurant	Asian Restaurant	Gym	Metro Station	Mexican Restaurant
635	Weston	Pharmacy	Fried Chicken Joint	Diner	Convenience Store	Park	Pizza Place	Factory	Farmers Market	Farm	Falafel Restaurant
656	Woodrow	Pharmacy	Pizza Place	Coffee Shop	Home Service	Chinese Restaurant	Donut Shop	Liquor Store	Sushi Restaurant	Bagel Shop	Bakery
657	Woodside	Thai Restaurant	Bakery	Latin American Restaurant	Grocery Store	Pub	American Restaurant	Filipino Restaurant	Bar	Gourmet Shop	Donut Shop

128 rows × 11 columns

Lots of bus stops in this area.

Cluster 13: Sports and events

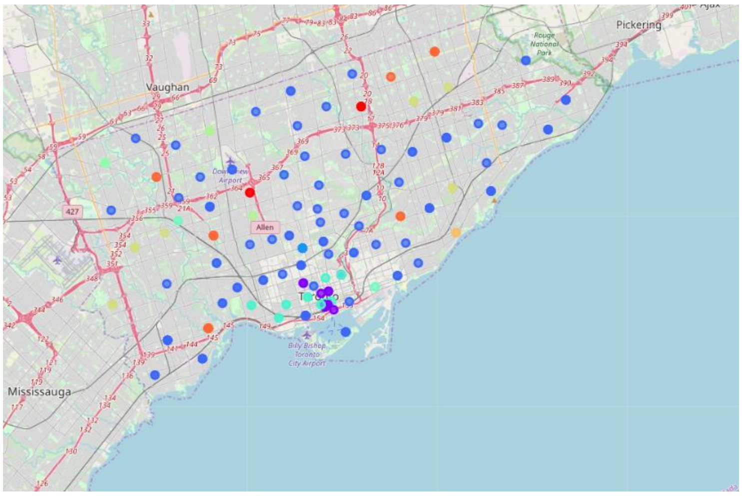
```
totalvenues_sorted.loc[totalvenues_sorted['Cluster_Labels'] == 13, totalvenues_sorted.columns[[1] + list(range(2, totalvenues_sorted.shape[1]))]]
```

	NeighborhoodName	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
3	Agincourt North	Pharmacy	Sushi Restaurant	Film Studio	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Service	Event Space	Exhibit	Factory
63	Britz	Fast Food Restaurant	Automotive Shop	Hardware Store	Harbor / Marina	Intersection	Hookah Bar	Pizza Place	Falafel Restaurant	Factory	Exhibit
134	Del Ray	Construction & Landscaping	Coffee Shop	Fast Food Restaurant	Yoga Studio	Filipino Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Service	Event Space
170	Emery	Coffee Shop	Nightclub	Park	Yoga Studio	Filipino Restaurant	English Restaurant	Ethiopian Restaurant	Event Service	Event Space	Exhibit
268	Humber Bay	Italian Restaurant	Fast Food Restaurant	Sushi Restaurant	Coffee Shop	Yoga Studio	Factory	Farmers Market	Farm	Falafel Restaurant	Event Space
272	Humberlea	Coffee Shop	Nightclub	Park	Yoga Studio	Filipino Restaurant	English Restaurant	Ethiopian Restaurant	Event Service	Event Space	Exhibit
292	Keelesdale	Construction & Landscaping	Coffee Shop	Fast Food Restaurant	Yoga Studio	Filipino Restaurant	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Service	Event Space
300	King's Mill Park	Italian Restaurant	Fast Food Restaurant	Sushi Restaurant	Coffee Shop	Yoga Studio	Factory	Farmers Market	Farm	Falafel Restaurant	Event Space
304	Kingsway Park South East	Italian Restaurant	Fast Food Restaurant	Sushi Restaurant	Coffee Shop	Yoga Studio	Factory	Farmers Market	Farm	Falafel Restaurant	Event Space
310	L'Amoreaux East	Pharmacy	Sushi Restaurant	Film Studio	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Service	Event Space	Exhibit	Factory
311	L'Amoreaux West	Fast Food Restaurant	Chinese Restaurant	Camera Store	Pharmacy	Gym Pool	Grocery Store	Coffee Shop	Pizza Place	Sandwich Place	Thrift / Vintage Store
358	Marienfelde	Chinese Restaurant	Farmers Market	Restaurant	Tennis Court	Park	Fast Food Restaurant	Factory	Farm	Falafel Restaurant	Yoga Studio
374	Milliken	Pharmacy	Sushi Restaurant	Film Studio	Empanada Restaurant	English Restaurant	Ethiopian Restaurant	Event Service	Event Space	Exhibit	Factory

A lot of event places and gyms are found in this cluster.

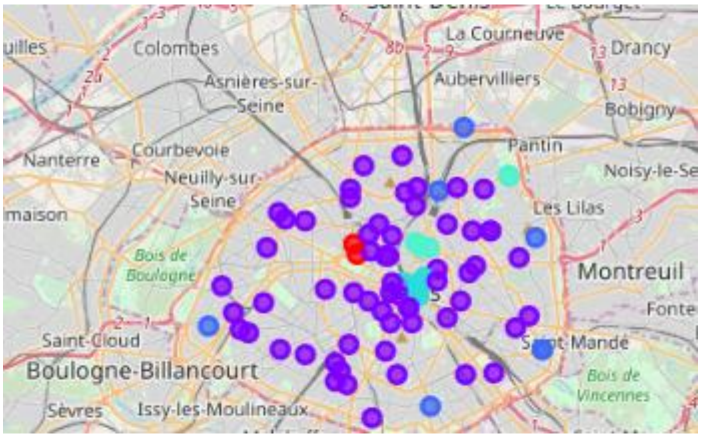
MAPS

Toronto

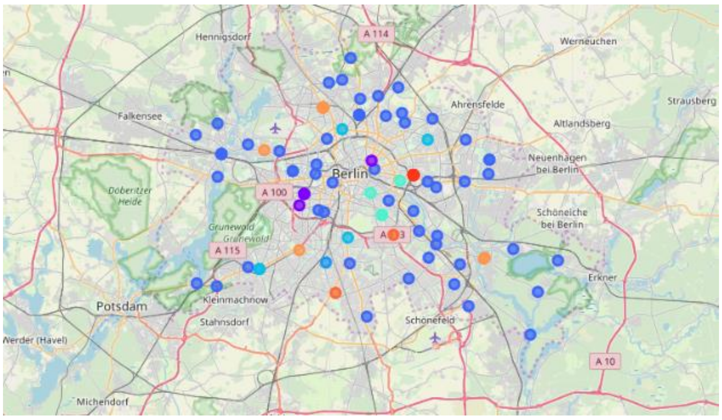


NYC

Paris



Berlin



Discussion:

In my opinion, the results are not very conclusive. To get more specific results, improvements would have to be made to the model and maybe other variables can be used to enrich the data :

- Income per capita
- Crime rates in the boroughs
- Cost of travel
- Average restaurant prices
- Average housing costs.

Conclusion:

In general there are differences to be seen in the clusters. However, to really spot differences between neighborhoods in the different cities, the model has to be refined. More variables can be added. The foursquare data also has to be examined more closely. Example: sushi restaurant, ramen restaurant, udon restaurant can all be classified as Japanese restaurants. This will refine the model, so better differences can be spotted.

I hope that you did enjoy reading this capstone. I would recommend reading the Jupyter Notebook that belongs to this project.

Thank you for reading.