THE TRAVELLING STUDENT

Clustering the neighbourhoods of London and Toronto

INTRODUCTION

- Studying abroad is an essential part of a lot of students University experience. This is because they not only want to study at a great establishment but also want to explore the new city that they are going to be living in for the next few years
- We will be focusing on students that want to study in an English-speaking country as they want to develop their knowledge of the language but are having difficulties deciding which country, they want to explore
- London and Toronto are quite the popular tourist destinations for people all around the world. They are diverse and multicultural and offer a wide variety of experiences that is widely sought after. We try to group the neighbourhoods of London and Toronto respectively and draw insights to what they look like now.

BUSINESS PROBLEM

- The aim is to help international students choose their destinations depending on the experiences that the neighbourhoods have to offer and what they would want to have. This also helps people make decisions if they are thinking about living in London or Toronto
- Our findings will help the international students make informed decisions and address any concerns they have including the different kinds of cuisines; provision stores and what the city has to offer.

DATA DESCRIPTION

- We require geolocation data for both London and Toronto. Postal codes in each city serve as a starting point. Using Postal codes, we use can find out the neighbourhoods, boroughs, venues and their most popular venue categories.
- London – To derive our solution, we scrape our data from: https://en.wikipedia.org/wiki/List_of_areas_of_London to get the neighbourhood, borough and post codes
- Toronto – To derive our solution, we scrape our data from: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M to get the neighbourhood, borough and post codes

DATA COLLECTION - LONDON

- To collect data for London, we scrape the List of areas of London Wikipedia page to take the second table using the following code and producing the following table

```
url_london="https://en.wikipedia.org/wiki/List_of_areas_of_London"
wiki_london_url=requests.get(url_london)
wiki_london_url
wiki_london_data=pd.read_html(wiki_london_url.text)
wiki_london_data
```

	Location	London borough	Post town	Postcode district	Dial code	OS grid ref
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon[8]	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728
...	...	...	...	...	...	...
526	Woolwich	Greenwich	LONDON	SE18	020	TQ435795
527	Worcester Park	Sutton, Kingston upon Thames	WORCESTER PARK	KT4	020	TQ225655
528	Wormwood Scrubs	Hammersmith and Fulham	LONDON	W12	020	TQ225815
529	Yeading	Hillingdon	HAYES	UB4	020	TQ115825
530	Yiewsley	Hillingdon	WEST DRAYTON	UB7	020	TQ063804

531 rows × 6 columns

DATA COLLECTION - TORONTO

- To collect the data from Toronto, we had to scrape the data from the List of postal codes of Canada: M Wikipedia page using the following code and producing the following table

```
source=requests.get("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M").text
soup=BeautifulSoup(source,"lxml")
table=soup.find("table")
table_rows=table.tbody.find_all("tr")
res=[]
for tr in table_rows:
    td=tr.find_all("td")
    row=[tr.text for tr in td]
    # Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.
    if row != [] and row[1] != "Not assigned\n":
        # If a cell has a borough but a "Not assigned" neighborhood, then the neighborhood will be the same as the borough.
        if "Not assigned\n" in row[2]:
            row[2]=row[1]
        res.append(row)
# Dataframe with 3 columns
df=pd.DataFrame(res,columns=["PostalCode","Borough","Neighborhood"])
df.head()
```

	PostalCode	Borough	Neighborhood
0	M3A\n	North York\n	Parkwoods\n
1	M4A\n	North York\n	Victoria Village\n
2	M5A\n	Downtown Toronto\n	Regent Park, Harbourfront\n
3	M6A\n	North York\n	Lawrence Manor, Lawrence Heights\n
4	M7A\n	Downtown Toronto\n	Queen's Park, Ontario Provincial Government\n

DATA PRE-PROCESSING

- In the London data frame, we had to replace the spaces with underscores in the column titles and had to remove the square brackets with numbers in them e.g.: "[7]"

```
wiki_london_data.rename(columns=lambda x: x.strip().replace(" ", "_"), inplace=True)
df1["borough"]=df1["borough"].map(lambda x: x.rstrip("]").rstrip("0123456789").rstrip("["))
```

- For Toronto, we had to remove the "\n" from the table

```
df["Neighborhood"]=df["Neighborhood"].str.replace("\n","")
df["PostalCode"]=df["PostalCode"].str.replace("\n","")
df["Borough"]=df["Borough"].str.replace("\n","")
df.head()
```

FEATURE SELECTION

- For the London data, we only needed the columns "boroughs", "Postal codes" and "Post town" for further steps. Therefore, we can remove the columns' ,"Location", "Dial codes" and "OS grid ref"

```
df1=wiki_london_data.drop([wiki_london_data.columns[0],
wiki_london_data.columns[4],wiki_london_data.columns[5]],axis=1)
```

- We did not have to do anything for the Toronto data as we already had it in the format that we needed.

FEATURE ENGINEERING

- Because we are only focussing on London, we had to then select only the neighbourhoods in London

```
df1=df1[df1["town"].str.contains("LONDON")]
```

- In comparison to the London data, the Toronto data was already how we wanted it with the data just being from the city of Toronto.

FEATURE ENGINEERING (CONT.)

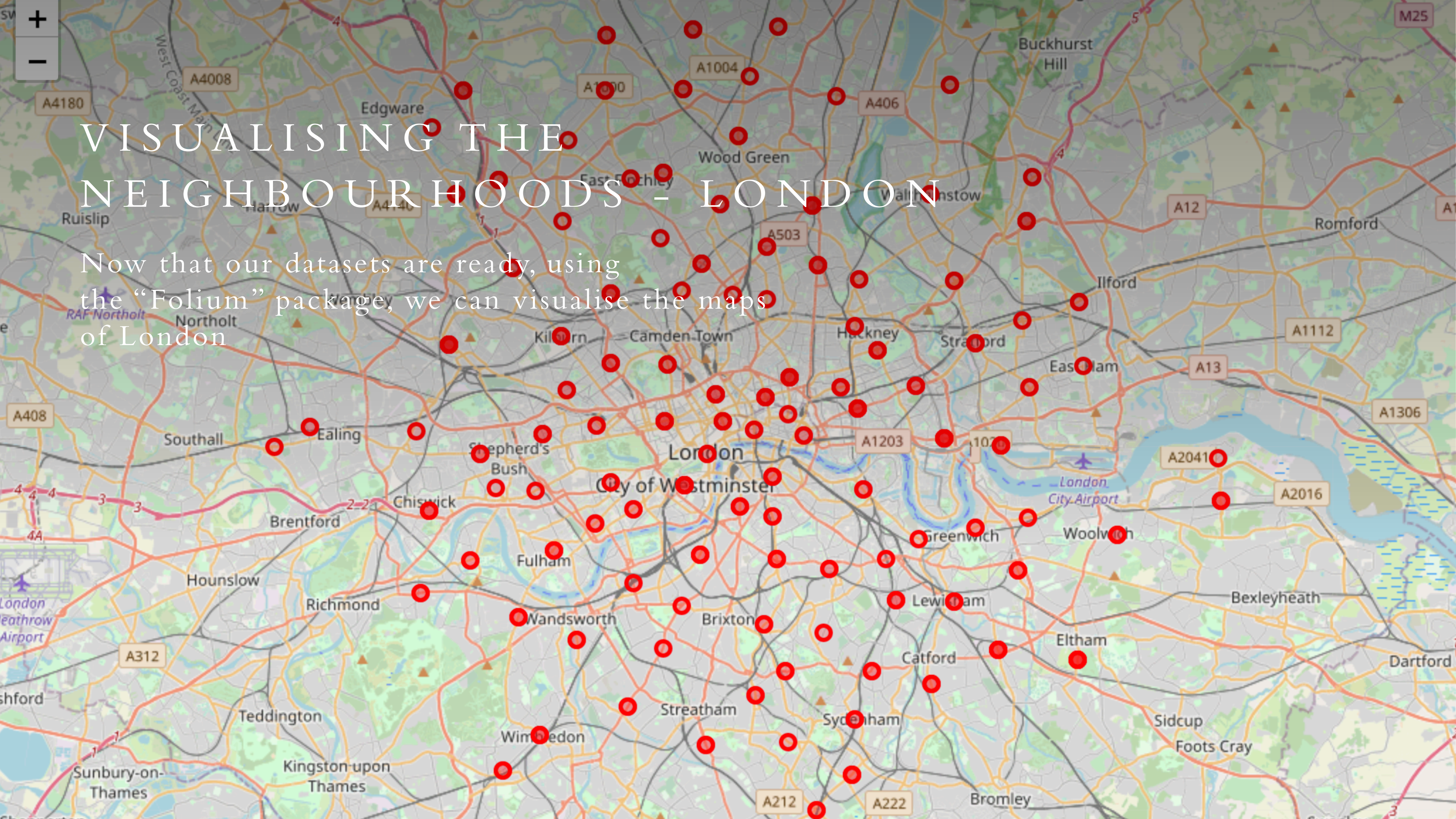
- We then had to get the geolocations of both cities (latitude and longitude) and incorporate them into both tables for London and Toronto

	borough	town	post_code	latitude	longitude
0	Bexley, Greenwich	LONDON	SE2	51.49245	0.12127
1	Ealing, Hammersmith and Fulham	LONDON	W3, W4	51.51324	-0.26746
6	City	LONDON	EC3	51.51200	-0.08058
7	Westminster	LONDON	WC2	51.51651	-0.11968
9	Bromley	LONDON	SE20	51.41009	-0.05683

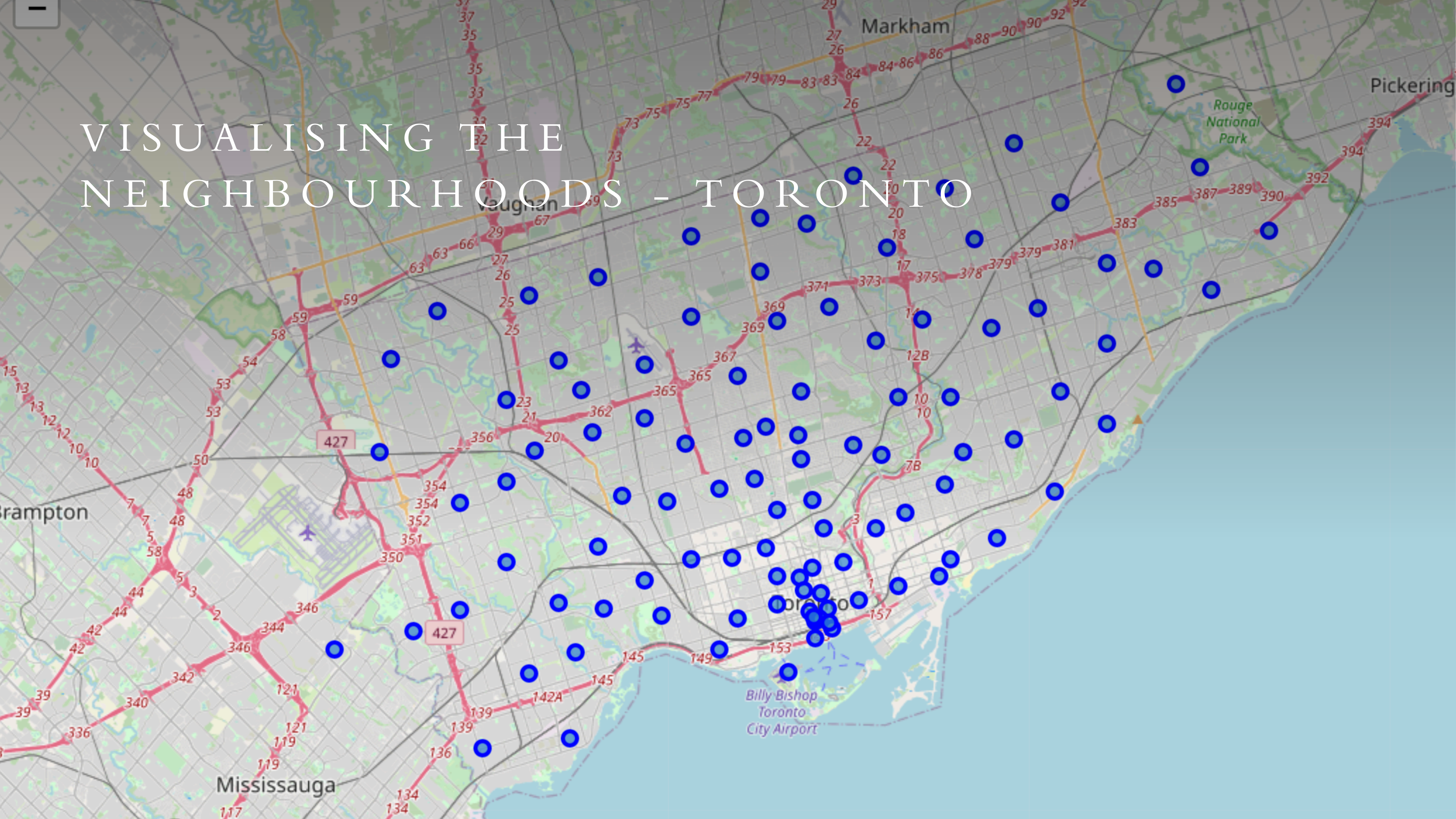
	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

VISUALISING THE NEIGHBOURHOODS - LONDON

Now that our datasets are ready, using the "Folium" package, we can visualise the maps of London



VISUALISING THE NEIGHBOURHOODS - TORONTO



FOURSQUARE

- Now that we have visualised the neighbourhoods, we need to find out what each neighbourhood is like and what are the common venues and the venue categories within a 500m radius.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
0	Bexley, Greenwich	51.49245	0.12127	Lesnes Abbey	Historic Site
1	Bexley, Greenwich	51.49245	0.12127	Sainsbury's	Supermarket
2	Bexley, Greenwich	51.49245	0.12127	Lidl	Supermarket
3	Bexley, Greenwich	51.49245	0.12127	Abbey Wood Railway Station (ABW)	Train Station
4	Bexley, Greenwich	51.49245	0.12127	Bean @ Work	Coffee Shop

Venue Category	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Accessories Store	Lawrence Manor, Lawrence Heights	43.778517	-79.346556	Sunglass Hut	43.777661	-79.344692
Airport	Downsview	43.737473	-79.394420	Toronto Downsview Airport (YZD)	43.738883	-79.396033
Airport Food Court	CN Tower, King and Spadina, Railway Lands, Har...	43.628947	-79.394420	Billy Bishop Café	43.631132	-79.396139
Airport Gate	CN Tower, King and Spadina, Railway Lands, Har...	43.628947	-79.394420	Gate 8	43.631536	-79.394570
Airport Lounge	CN Tower, King and Spadina, Railway Lands, Har...	43.628947	-79.394420	Porter Lounge	43.631360	-79.395756

ONE HOT ENCODING

- Since we are trying to find out what the different kinds of venue categories that are present in each neighbourhood and calculating the top 10 common venues to base our similarity on, we use the "One Hot Encoding" to work with our categorical datatypes of the venue categories. This helps to convert the categorical data into numeric data
- We perform one hot encoding and then calculate the mean of the grouped venue categories for each of the neighbourhoods

	Neighbourhood	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wine Shop	Wings Joint
0	Barnet	0.0	0.0	0.0	0.001795	0.0	0.0	0.0	0.007181	0.0	...	0.007181	0.0	0.0	0.0	0.0	0.0
1	Barnet, Brent, Camden	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
2	Bexley	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
3	Bexley, Greenwich	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0
4	Bexley, Greenwich	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	...	0.000000	0.0	0.0	0.0	0.0	0.0

	Neighborhood	Yoga Studio	Accessories Store	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Train Station	Truck Stop	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant	Warehouse Store	Wine Bar
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.038462	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0

TOP VENUES IN EACH NEIGHBOURHOOD

- we need to rank and label the top venue categories in our neighbourhood. There are many categories, we will consider top 10 categories

	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	Barnet	Coffee Shop	Café	Grocery Store	Pub	Supermarket	Pharmacy	Italian Restaurant	Bus Stop	Sushi Restaurant	Turkish Restaurant
1	Barnet, Brent, Camden	Bus Station	Clothing Store	Gym / Fitness Center	Hardware Store	Supermarket	Fish & Chips Shop	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Filipino Restaurant
2	Bexley	Supermarket	Historic Site	Convenience Store	Coffee Shop	Train Station	Platform	Bus Stop	Golf Course	Park	Fish Market
3	Bexley, Greenwich	Daycare	Construction & Landscaping	Bus Stop	Park	Golf Course	Historic Site	Home Service	Sports Club	Discount Store	Diner
4	Bexley, Greenwich	Supermarket	Train Station	Coffee Shop	Convenience Store	Platform	Historic Site	Film Studio	Exhibit	Falafel Restaurant	Farmers Market

	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	Agincourt	Latin American Restaurant	Lounge	Breakfast Spot	Women's Store	Dumpling Restaurant	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore
1	Alderwood, Long Branch	Pizza Place	Gym	Dance Studio	Pharmacy	Coffee Shop	Athletics & Sports	Pub	Dog Run	Dim Sum Restaurant	Diner
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Fried Chicken Joint	Pizza Place	Intersection	Supermarket	Ice Cream Shop	Sushi Restaurant	Restaurant	Shopping Mall
3	Bayview Village	Café	Bank	Chinese Restaurant	Japanese Restaurant	Women's Store	Doner Restaurant	Discount Store	Distribution Center	Dog Run	Donut Shop
4	Bedford Park, Lawrence Manor East	Sandwich Place	Coffee Shop	Italian Restaurant	Women's Store	Indian Restaurant	Juice Bar	Breakfast Spot	Liquor Store	Locksmith	Restaurant

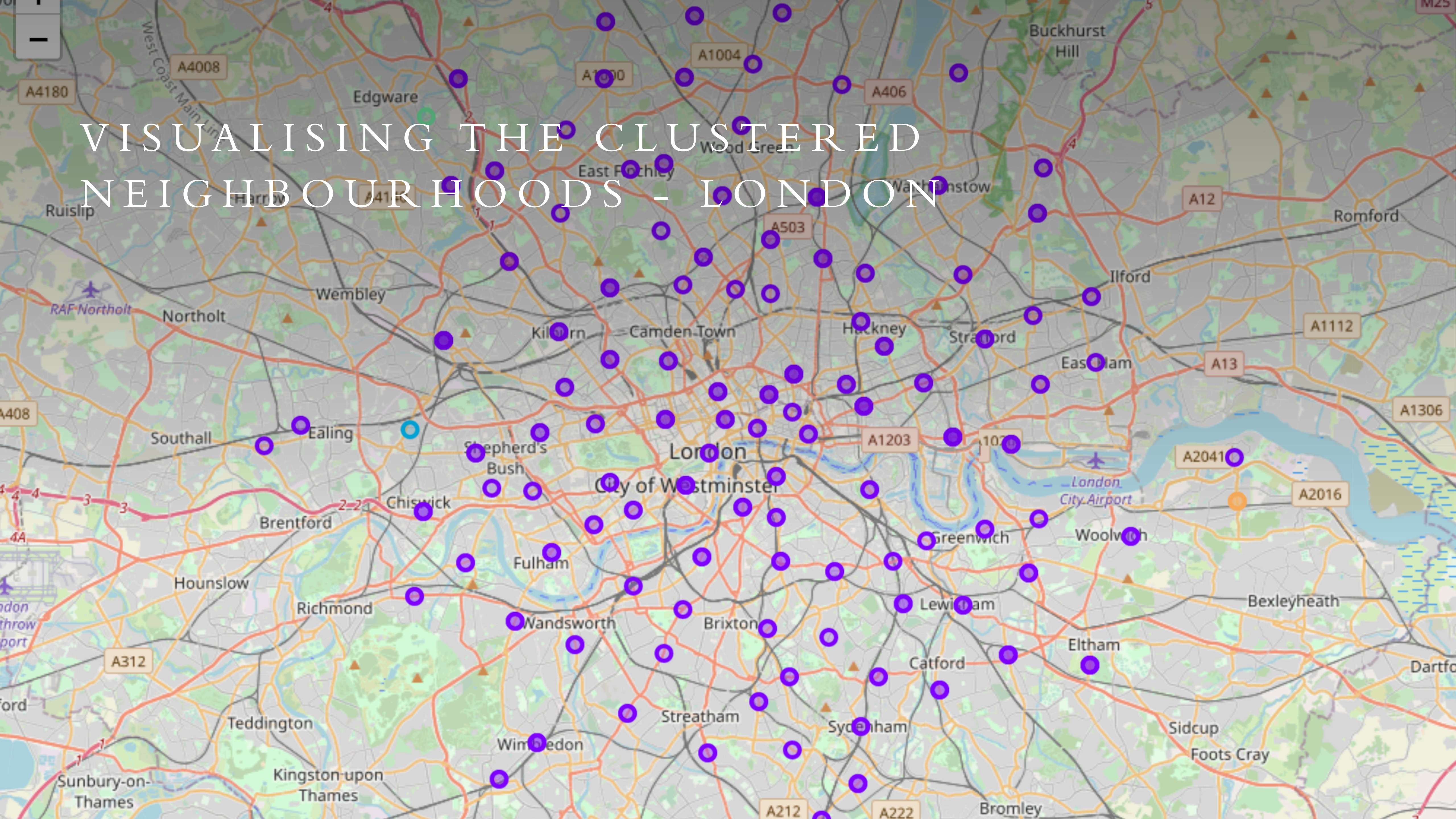
MODEL BUILDING–K MEANS CLUSTERING

- We will now be using K Means Clustering Machine learning algorithm to cluster similar neighbourhoods together. We will be going with the number of clusters as 5.
- We add our labels to the data and then merge with our neighbourhood venues that is sorted to add latitude and longitude for each of the neighbourhoods.

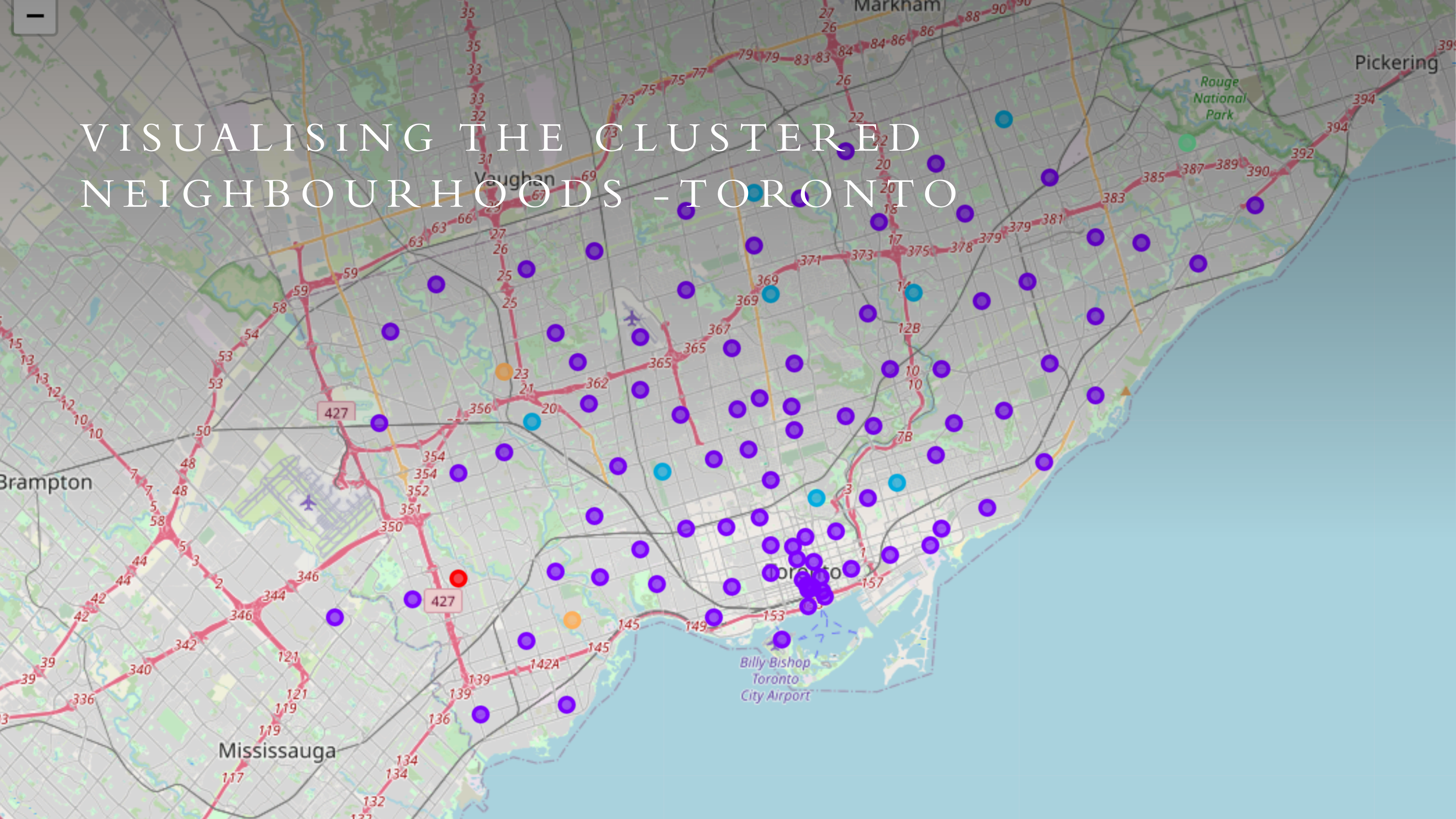
borough	town	post_code	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Bexley, Greenwich	LONDON	SE2	51.49245	0.12127	4	Supermarket	Train Station	Coffee Shop	Convenience Store	Platform	Historic Site	Film Studio	Exhibit	Falafel Restaurant	Farmers Market
Ealing, nmersmith nd Fulham	LONDON	W3, W4	51.51324	-0.26746	2	Grocery Store	Indian Restaurant	Park	Breakfast Spot	Train Station	Zoo Exhibit	Film Studio	Exhibit	Falafel Restaurant	Farmers Market
City	LONDON	EC3	51.51200	-0.08058	1	Hotel	Italian Restaurant	Coffee Shop	Gym / Fitness Center	Pub	Restaurant	Wine Bar	Sandwich Place	Salad Place	Scenic Lookout
estminster	LONDON	WC2	51.51651	-0.11968	1	Hotel	Coffee Shop	Pub	Café	Sandwich Place	Italian Restaurant	Theater	Restaurant	Hotel Bar	Sushi Restaurant
Bromley	LONDON	SE20	51.41009	-0.05683	1	Supermarket	Grocery Store	Convenience Store	Fast Food Restaurant	Hotel	Park	Café	Historic Site	Gym / Fitness Center	Italian Restaurant

Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
North York	Parkwoods	43.753259	-79.329656	2.0	Park	Fireworks Store	Food & Drink Shop	Donut Shop	Diner	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Drugstore
North York	Victoria Village	43.725882	-79.315572	1.0	Hockey Arena	Pizza Place	Coffee Shop	Portuguese Restaurant	Women's Store	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run
Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	1.0	Coffee Shop	Pub	Bakery	Park	Breakfast Spot	Café	Theater	Gym / Fitness Center	Electronics Store	Restaurant
North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	1.0	Clothing Store	Furniture / Home Store	Vietnamese Restaurant	Athletics & Sports	Coffee Shop	Miscellaneous Shop	Event Space	Boutique	Accessories Store	Ethiopian Restaurant
Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	1.0	Coffee Shop	Sushi Restaurant	College Cafeteria	Beer Bar	Bank	Bar	Portuguese Restaurant	Café	Diner	Yoga Studio

VISUALISING THE CLUSTERED NEIGHBOURHOODS - LONDON



VISUALISING THE CLUSTERED NEIGHBOURHOODS -TORONTO



EXAMINING THE CLUSTERS

- Finally, we examined our clusters for each city, London and Toronto respectively

```
london_data_nonan.loc[london_data_nonan["Cluster Labels"]==1,london_data_nonan.columns[[1]+
                                                              list(range(5,london_data_nonan.shape[1]))]]
london_data_nonan.loc[london_data_nonan["Cluster Labels"]==2,london_data_nonan.columns[[1]+
                                                              list(range(5,london_data_nonan.shape[1]))]]
london_data_nonan.loc[london_data_nonan["Cluster Labels"]==3,london_data_nonan.columns[[1]+
                                                              list(range(5,london_data_nonan.shape[1]))]]
london_data_nonan.loc[london_data_nonan["Cluster Labels"]==4,london_data_nonan.columns[[1]
                                                              +list(range(5,london_data_nonan.shape[1]))]]
london_data_nonan.loc[london_data_nonan["Cluster Labels"]==5,london_data_nonan.columns[[1]
                                                              +list(range(5,london_data_nonan.shape[1]))]]
```

```
toronto_merged_nonan.loc[toronto_merged_nonan["Cluster Labels"]==1,toronto_merged_nonan.columns[[1]+
                                                              list(range(5,toronto_merged_nonan.shape[1]))]]
toronto_merged_nonan.loc[toronto_merged_nonan["Cluster Labels"]==2,toronto_merged_nonan.columns[[1]+
                                                              list(range(5,toronto_merged_nonan.shape[1]))]]
toronto_merged_nonan.loc[toronto_merged_nonan["Cluster Labels"]==3,toronto_merged_nonan.columns[[1]
                                                              +list(range(5,toronto_merged_nonan.shape[1]))]]
toronto_merged_nonan.loc[toronto_merged_nonan["Cluster Labels"]==4,toronto_merged_nonan.columns[[1]+
                                                              list(range(5,toronto_merged_nonan.shape[1]))]]
toronto_merged_nonan.loc[toronto_merged_nonan["Cluster Labels"]==5,toronto_merged_nonan.columns[[1]+
                                                              list(range(5,toronto_merged_nonan.shape[1]))]]
```

RESULTS AND DISCUSSION

- The neighbourhoods of London are very multicultural. There are a lot of different cuisines including Indian, Italian, Turkish and Chinese
- It has a lot of shopping options too with that of the Flea markets, flower shops, fish markets, Fishing stores, clothing stores
- For leisure, the neighbourhoods are set up to have lots of parks, golf courses, zoo, gyms and Historic sites
- The main modes of transport seem to be Buses and trains.
- Overall, the city of London offers a multicultural, diverse and certainly an entertaining experience

RESULTS AND DISCUSSION (CONT.)

- Toronto is relatively the same as London. It has a wide variety of cuisines and eateries including Japanese, Vietnamese, Ethiopian, Colombian, Bakeries
- There are a lot of places to relax including cafe's, beer bars and coffee shops
- Toronto has a lot of Diners and people also like to go to places such as farmers markets and visiting the harbour
- Different means of public transport in Toronto which includes the metro station and light rail station
- For leisure, there a lot of dog runs, gyms, spas and people like to also play hockey baseball
- Overall, much like London, Toronto is multicultural, diverse and offers a great experience

CONCLUSION

- The purpose of this project was to explore the cities of London and Toronto and see how attractive it is to potential tourists and international students who are planning on studying in one of these cities
- We could see that each of the neighbourhoods in both the cities have a wide variety of experiences to offer which is unique. The cultural diversity is quite evident which also gives the feeling of a sense of inclusion
- Both London and Toronto seem to offer a vast number of things to do while studying with a lot of places to explore, beautiful landscapes and a wide variety of culture
- Overall, it's up to the international students to decide which city they would most like to study in as both cities offer a lot of the same thing which will make them feel included
- To decide which city to study in they would probably need to do research on the university itself as both cities would be a wonderful place to study in

REFERENCES

- https://en.wikipedia.org/wiki/List_of_areas_of_London
- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Clustering Neighborhoods of London and Paris using Machine Learning
- The Battle of Neighbourhood — My London's Perspective by Dayo John
- https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/DS0701EN/DS0701EN-2-2-1-Foursquare-API-py-v2.0.ipynb?lti=true