INTRODUCTION

I currently live in Manchester (UK) city centre, but I am looking to move to either Toronto or New York. How do I decide which city and, more specifically, which neighbourhood I want to live in?

Manchester is an international university city with lots of students from different backgrounds who want to move to the world's financial centres for post-graduation employment. Everyone is different, but we all agree that the Mancunian social scene is great.
Thus, my selection criteria will be geared towards moving to a neighbourhood in Toronto or New York which has similar amenities.

DATA

Data I used and the respective sources:

- Used mapawi to get Manchester postcodes and longitude/latitude data (http://zip-code.en.mapawi.com/united-kingdom/2/greater-manchester/2/60/manchester/m1/25483/)
- I will use Foursquare API to get the most common venues of given neighbourhoods in Manchester, Toronto and New York (https://foursquare.com/developers/)
- Toronto neighbourhood data taken from Wikipedia (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Toronto)
- New York neighbourhood data taken from work completed on the IBM Data Science Professional Certificate

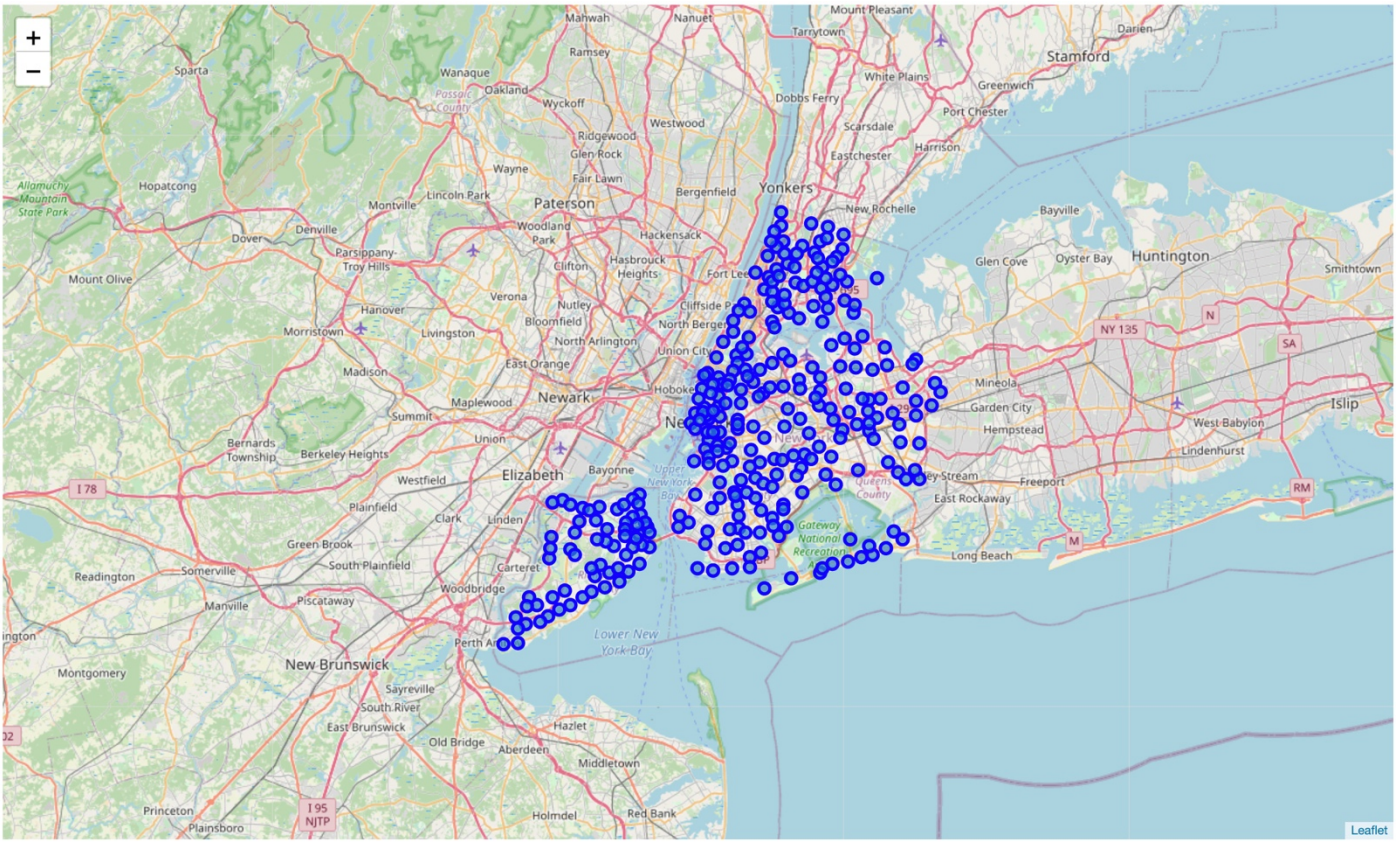
METHODOLOGY

Part A - Data Mining
First, I will explain how I collected the data.
For the New York data I simply imported the same location dataset that was used in week 3 of the Applied Data Science Capstone. As seen in my notebook, this data is imported via json and its original form is an incredibly long dictionary with all the data imbedded in it (some of it useful and some not). In order to extract the relevant neighbourhood, borough, longitude and latitude information I simply created a new dictionary containing only the 'features' key of the original dataset.
OK, now I have my raw New York dataset but I need to insert it into a Pandas dataframe in order to properly use it. I first created an empty dataframe with the following columns: Borough, Neighbourhood, Latitude and Longitude. To fill this table with data I had to use a loop that sequentially inserted each relevant row of data one at a time.
Here's a sample of the finished table:

```
ny_neighbourhoods.head()
```

	Borough	Neighbourhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

This dataframe contained 5 boroughs and a total of 306 neighbourhoods.
The data can be visualised as such:



If you go to the notebook you can click each blue dot to see the respective neighbourhood and borough.

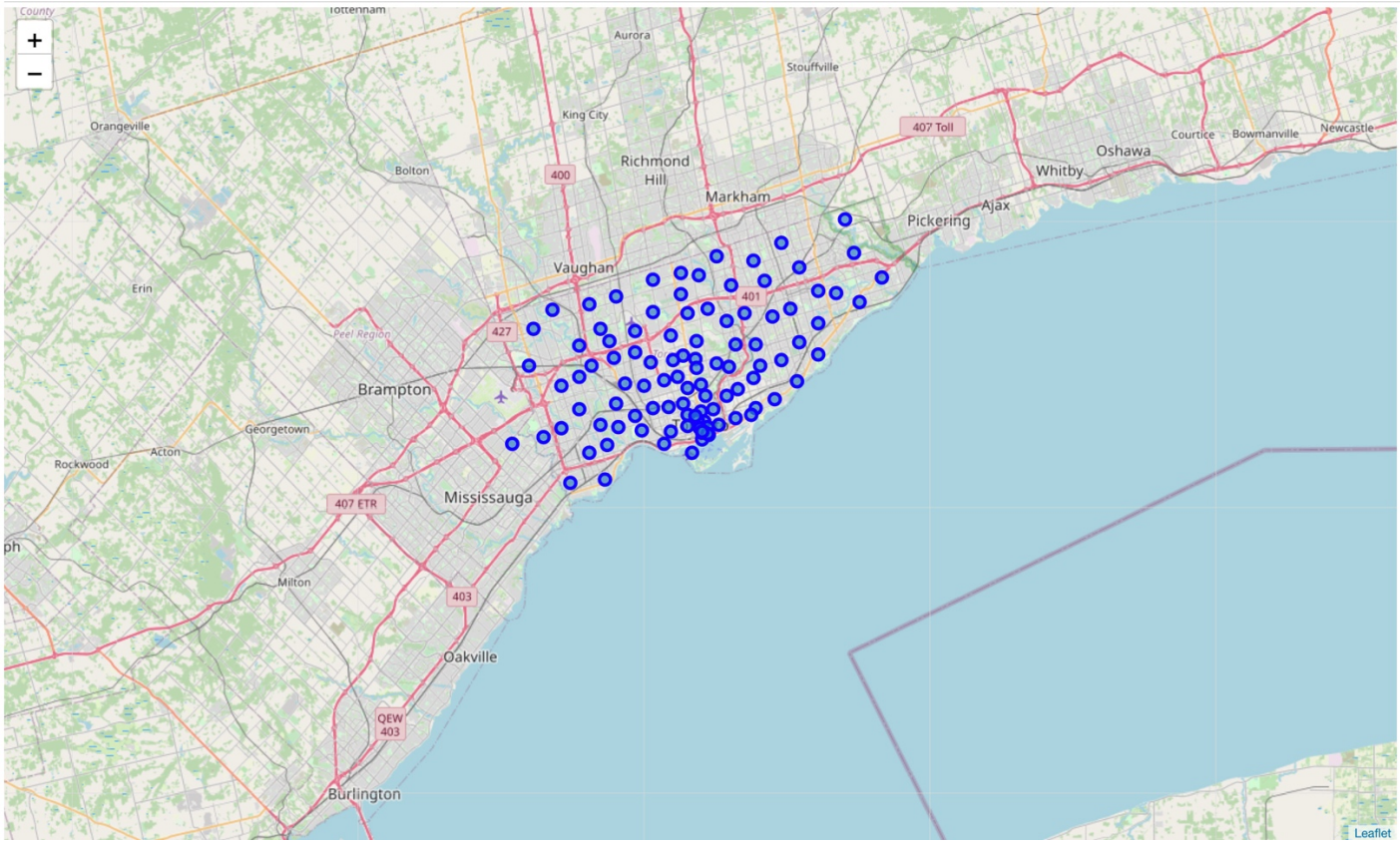
For my Toronto data I had to first go to the Toronto postal codes Wikipedia page (link can be found in the data section of the report). I copied the table in this webpage and loaded it into excel where I cleaned it and added longitude and latitude data that was also provided during week 3 of the module. After this I imported the table back into my Notebook using the read_csv Pandas function.
Here's a quick view of the resulting dataframe:

```
toronto_neighbourhoods.head()
```

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

In total there are 10 boroughs and 103 neighbourhoods.

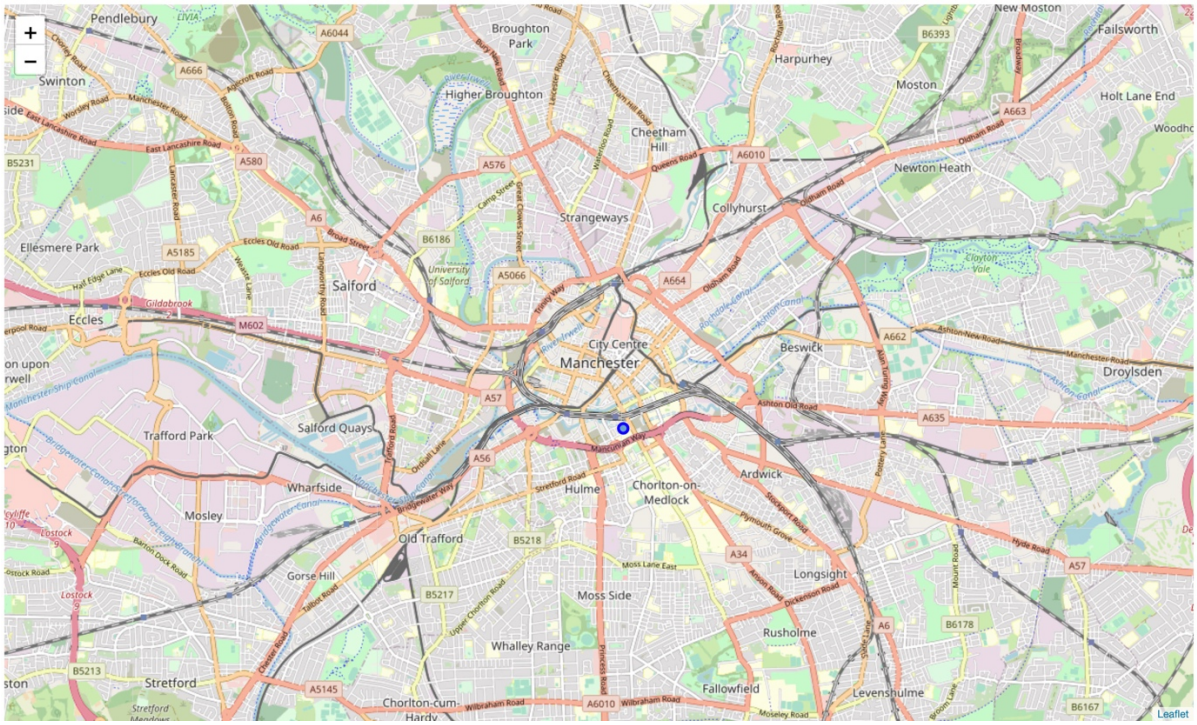
Similar to New York, the boroughs can be visualised as such:



Last but not least, I need location data for my neighbourhood in Manchester, UK. To find this I went to a website called Mapawi (link in Data section) and exported their .csv file which contains every Manchester Post Code's longitude and latitude. I cleared all other non-relevant rows of data then imported the resulting spreadsheet so that I was left with the following 1-row dataframe:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M1 5QD	Deansgate	Oxford Road	53.472531	-2.240887

Visualisation of my neighbourhood:



Now I have all my location data for New York, Toronto and Manchester. The final step is to amalgamate the data into one table for the analysis, I did this in excel then loaded the data back into the notebook. Here's the top 5 rows of the final dataset:

	Borough	Neighbourhood	Latitude	Longitude
0	Deansgate	Oxford Road	53.472531	-2.240887
1	Scarborough	Malvern, Rouge	43.806686	-79.194353
2	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
3	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
4	Scarborough	Woburn	43.770992	-79.216917

Part B – Finding Venue Data

To find the relevant venue data in each neighbourhood I used the Foursquare API. I defined a new function to retrieve the data (this can be seen in my notebook) and set my radius variable at 300 units. This function was then run for the neighbourhood data which produced the following table:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Oxford Road	53.472531	-2.240887	Costa Coffee	53.472734	-2.239394	Coffee Shop
1	Oxford Road	53.472531	-2.240887	Hatch	53.471972	-2.238277	Pop-Up Shop
2	Oxford Road	53.472531	-2.240887	Takk Espresso Bar	53.471834	-2.238618	Café
3	Oxford Road	53.472531	-2.240887	Zouk Tea Bar & Grill	53.472321	-2.240544	Indian Restaurant
4	Oxford Road	53.472531	-2.240887	The Salisbury Ale House	53.473969	-2.241055	Pub

This is useful information but I still needed to do a fair bit of work to get it into a usable format for k-Nearest-Neighbour analysis.

The first step was to one hot encode each venue category respective to each neighbourhood then group the data by neighbourhood. This gives us a table with the ratio of each venue category in every neighbourhood. From here I wanted to look at the top 5 venues in each neighbourhood in descending order so I wrote a function to do so. Finally, I created a new dataframe that displayed the top 10 venues for each neighbourhood, a sample of the final product can be seen 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	Agincourt	Breakfast Spot	Latin American Restaurant	Yoga Studio	Farmers Market	Entertainment Service	Escape Room	Ethiopian Restaurant	Event Service	Event Space	Eye Doctor
1	Alderwood, Long Branch	Pub	Coffee Shop	Pizza Place	Gym	Pharmacy	Yoga Studio	Factory	Empanada Restaurant	English Restaurant	Entertainment Service
2	Allerton	Pizza Place	Spa	Discount Store	Chinese Restaurant	Martial Arts School	Electronics Store	Bike Trail	Gas Station	Donut Shop	Fast Food Restaurant
3	Annadale	American Restaurant	Train Station	Yoga Studio	Farmers Market	English Restaurant	Entertainment Service	Escape Room	Ethiopian Restaurant	Event Service	Event Space
4	Arden Heights	Deli / Bodega	Coffee Shop	Bus Stop	Pharmacy	Yoga Studio	Farm	Entertainment Service	Escape Room	Ethiopian Restaurant	Event Service

Awesome! I also quickly sense-checked the most common venues in my neighbourhood:

	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
248	Oxford Road	Pub	Bar	Coffee Shop	Burrito Place	Pop-Up Shop	Bakery	Indian Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Pool

Pubs and bars are the most common, no surprises here!

Part C – k Nearest Neighbour Analysis

I am not looking to segment these neighbourhoods as one would typically, but rather find 10-15 neighbourhoods in Toronto and/or New York that are similar to where I currently live. Furthermore, I will not use the 'elbow' method typically utilised for finding the best k, but instead arbitrarily set my k at 40 to begin with. After analysing the results, I will adjust my k accordingly until I have found the results I am looking for.

In total I did 6 rounds of k clustering, starting with k-clusters = 40. I found this number of clusters to be insufficient as I was left with over 100 neighbourhoods that were considered similar to where I live but this is not specific enough for my liking.

For round 2 I increased the number of clusters to 80 and found the same problem as in round 1.
In round 3 the number of clusters was 120 but still the same issue arose.
From here I increased the number of clusters all the way up to 350 but found my neighbourhood to be in a cluster by itself which is also no good.
Round 5 had 250 clusters and I was left with 14 similar neighbourhoods, but I felt I could refine the results a bit more.
My final round, round 6, had 265 clusters. I will outline the results in the following section.

RESULTS AND DISCUSSION

Here's the final table of results:

	Borough	Neighbourhood	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
0	Deansgate	Oxford Road	53.472531	-2.240887	108.0	Pub	Bar	Coffee Shop	Burrito Place	Pop-Up Shop	Bakery	Indian Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Pool
78	West Toronto	Little Portugal, Trinity	43.647927	-79.419750	108.0	Bar	Asian Restaurant	Art Gallery	Theater	Bakery	Beer Store	Brewery	Cocktail Bar	Record Shop	Yoga Studio
153	Brooklyn	Greenpoint	40.730201	-73.954241	108.0	Bar	Coffee Shop	Cocktail Bar	Grocery Store	Mexican Restaurant	Supermarket	Spa	Café	Sushi Restaurant	Pizza Place
163	Brooklyn	Prospect Heights	40.676822	-73.964859	108.0	Bar	Cocktail Bar	Mexican Restaurant	Brewery	Thai Restaurant	Grocery Store	Greek Restaurant	Sushi Restaurant	Garden Center	Garden
165	Brooklyn	Williamsburg	40.707144	-73.958115	108.0	Taco Place	Pizza Place	Bar	Latin American Restaurant	Breakfast Spot	Lounge	Liquor Store	Grocery Store	Gym	Coffee Shop
166	Brooklyn	Bushwick	40.698116	-73.925258	108.0	Bar	Mexican Restaurant	Deli / Bodega	Coffee Shop	Thrift / Vintage Store	Sandwich Place	Liquor Store	Chinese Restaurant	Latin American Restaurant	Nightclub
191	Brooklyn	Boerum Hill	40.685683	-73.983748	108.0	Furniture / Home Store	Bar	Italian Restaurant	Thrift / Vintage Store	Sandwich Place	Men's Store	Kids Store	Concert Hall	Jewelry Store	Middle Eastern Restaurant
201	Brooklyn	South Side	40.710861	-73.958001	108.0	Pizza Place	Bar	Coffee Shop	Latin American Restaurant	Yoga Studio	American Restaurant	Dive Bar	Burger Joint	South American Restaurant	Deli / Bodega
207	Manhattan	Hamilton Heights	40.823604	-73.949688	108.0	Café	Mexican Restaurant	Yoga Studio	Bar	Donut Shop	Coffee Shop	Cocktail Bar	Caribbean Restaurant	Pizza Place	Deli / Bodega
209	Manhattan	Central Harlem	40.815976	-73.943211	108.0	Caribbean Restaurant	Deli / Bodega	African Restaurant	Fried Chicken Joint	French Restaurant	Gym / Fitness Center	Beer Bar	Library	Lounge	Breakfast Spot
222	Manhattan	East Village	40.727847	-73.982226	108.0	Bar	Korean Restaurant	Mexican Restaurant	Cocktail Bar	Vegetarian / Vegan Restaurant	Pizza Place	Ice Cream Shop	Wine Bar	Italian Restaurant	Ramen Restaurant
243	Queens	Long Island City	40.750217	-73.939202	108.0	Hotel	Bar	Coffee Shop	Café	Deli / Bodega	Bubble Tea Shop	Mexican Restaurant	Pizza Place	Mediterranean Restaurant	Donut Shop
317	Staten Island	Great Kills	40.549480	-74.149324	108.0	Bar	Italian Restaurant	Pizza Place	Japanese Restaurant	Sushi Restaurant	Train Station	Liquor Store	Cafeteria	Food & Drink Shop	Pharmacy
381	Queens	Sunnyside Gardens	40.745652	-73.918193	108.0	Bar	Pizza Place	Coffee Shop	Spa	Grocery Store	Pub	Donut Shop	Thai Restaurant	Bank	Dog Run

The common theme is a high prevalence of bars which is exactly what I was looking for. I am also incredibly happy with the awesome variety of places to eat (e.g. coffee shops, Italian restaurants, Mexican restaurants, etc.).
I think any university student from Manchester would be comfortable moving to these neighbourhoods.

Thank you for reading my report!