David See
IBM – Applied Data Science Capstone

Travel to Toronto, Sydney, London and New York:

The Data Scientist Way!

David See
IBM – Applied Data Science Capstone

Table of Contents

David See
IBM – Applied Data Science Capstone

1. Introduction

Background

The tourism industry has experienced rapid growth and innovation to attract travellers from around the world. With so many places to visit in a given city/country, travellers will need to decide on how best to spend their time and money. Fortunately, this is a question data scientists are able to answer and this report showcases how to travel the "data scientists way"!

The Problem Statement

The *first question* adventurous travellers need to solve before embarking on their holiday is:

"What places should I visit in a given city such that I could maximise my travel experience and minimise visiting places that are similar?"

The answer to the question above depends on the country the traveller intends to visit, which leads to the *second question*:

"Given my travel interests, which countries should I visit and what are the differences/similarities between these countries?"

More specifically, I am a traveller who is keen to explore and review cafés, following which I would like to determine which neighbourhood and country I should travel to in this report.

Purpose of Report and Target Audience

Travellers who wish to visit the following four cities/countries would be interested in this report:

- Toronto, Canada
- Sydney, Australia
- London, United Kingdom
- New York, United States.

This report seeks to answer the questions above, which helps to improve the travel experiences, whereby travellers are able to select the city/country according to their own preferences. Furthermore, by reducing visits to places that are similar, travellers are using their time more wisely and this could translate into tangible money savings. We note the following high level conclusions for each of the four different cities/countries:

- Toronto as a Mixture of London and New York
- Sydney for Travellers Looking for Variety
- London as a Hub for Coffee Shops
- New York – Ideal for Families.

2. Data

Postal Code, Borough, Neighbourhood and Coordinates

The postal code, borough, neighbourhood, latitude and longitude for each city were obtained from Wikipedia, doogal, CostlessQuotes and Distancesto. An example of the data for Sydney is provided in Figure 1 below.

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	2000	Sydney	Dawes Point	-33.860	151.210
1	2000	Sydney	Haymarket	-33.880	151.200
2	2000	Sydney	Millers Point	-33.860	151.200
3	2000	Sydney	Sydney	-33.870	151.210
4	2000	Sydney	The Rocks	-33.860	151.210
5	2002	Sydney	World Square	-33.880	151.210

Figure 1: Postal Code, Borough, Neighbourhood, Latitude and Longitude in Sydney

Explore Venues Using Foursquare

Using the latitude and longitude for each of the neighbourhoods above, we applied the 'venues/explore' endpoint via Foursquare to discover nearby venues that are situated within 500 meters.

	Neighbourhood	American Restaurant	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Australian Restaurant	...
0	Dawes Point	0	0	0	0	0	0	0	0	0	...
1	Dawes Point	0	0	0	0	0	1	0	0	0	...
2	Dawes Point	0	0	0	0	0	0	0	0	0	...
3	Dawes Point	0	0	0	0	0	0	0	0	0	...
4	Dawes Point	0	0	0	0	0	0	0	0	0	...

Figure 2: Dummy Variables for Sydney

Processing Data for Machine Learning

For each of the venues, we obtained the latitude, longitude and category (e.g. café, coffee shop, waterfront etc.), following which we converted the category into dummy variables (Figure 2).

	Neighbourhood	American Restaurant	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Australian Restaurant	...
0	Bondi	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.045455	...
1	Chippendale	0.000000	0.00	0.010204	0.00	0.000000	0.000000	0.00	0.030612	0.010204	...
2	Clovelly	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	...
3	Coogee	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	...
4	Daceyville	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.025000	0.000000	...
5	Darling Point	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	...

Figure 3: Mean Dummy Variables for Sydney

In each of the neighourhoods, we calculated the frequency of the venue categories (Figure 3) and the top 10 most common venues (Figure 4).

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
Dawes Point	-33.86	151.21	0	Café	Hotel	Australian Restaurant	Cocktail Bar	Pub	Hotel Bar	Sandwich Place	Burger Joint	Italian Restaurant
Haymarket	-33.88	151.20	0	Café	Chinese Restaurant	Thai Restaurant	Coffee Shop	Malay Restaurant	Ramen Restaurant	Dumpling Restaurant	Hotel	Asian Restaurant
Millers Point	-33.86	151.20	0	Café	Chinese Restaurant	Seafood Restaurant	Park	Steakhouse	Middle Eastern Restaurant	Brewery	Nature Preserve	Boat or Ferry
Sydney	-33.87	151.21	0	Café	Coffee Shop	Shopping Mall	Cocktail Bar	Hotel	Bar	Clothing Store	Speakeasy	Japanese Restaurant
The Rocks	-33.86	151.21	0	Café	Hotel	Australian Restaurant	Cocktail Bar	Pub	Hotel Bar	Sandwich Place	Burger Joint	Italian Restaurant

Figure 4: Top 10 Most Common Venues in Sydney (excerpt)

We applied machine learning (specifically k-means clustering) to the data above. The idea is that neighbourhoods that are similar will be grouped within the same clusters using machine learning. The venues *within* each cluster are relatively similar, however the venues are relatively different *between* each cluster.

3. Methodology

3.1 Methodology – By Neighbourhood Clusters

Neighbourhood Labelling

Using the information from the previous section, we labelled the neighbourhoods according to their latitude and longitude as shown in Figure 5 below.

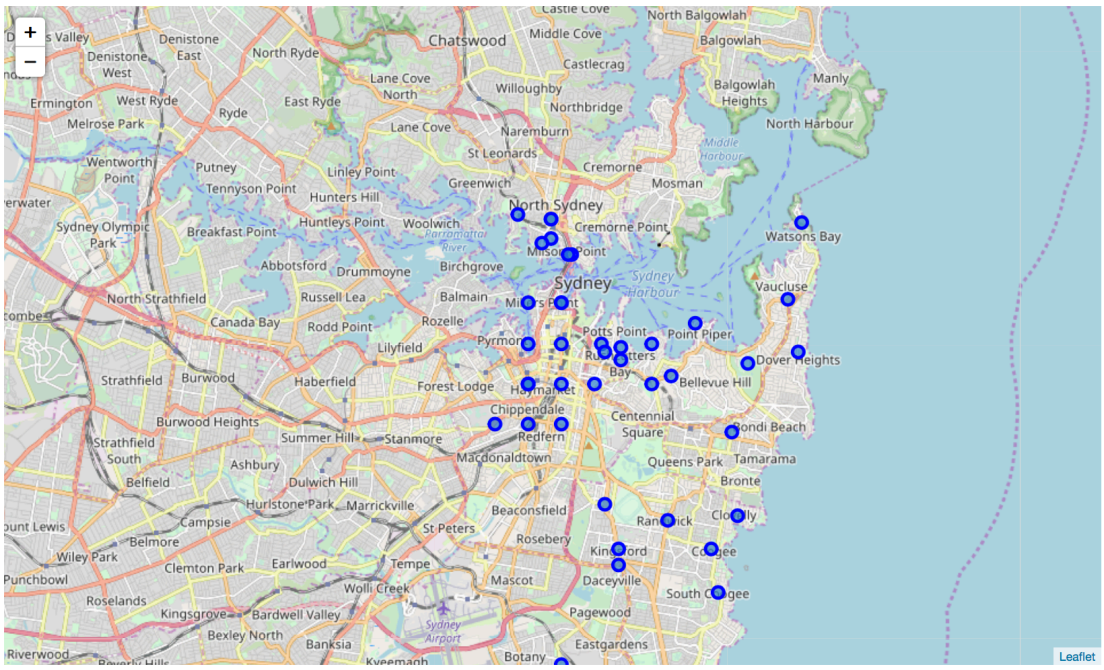

Figure 5: A Map Visualisation of the Neighbourhood in Sydney

Machine Learning Application – k-Means Clustering

We applied machine learning (specifically k-means clustering) to the data in Figure 5. Using machine learning techniques, we first explored the neighbourhood clusters in each city **individually** to address the *first question* that was set out in the introduction.

Traveller Assumption

We assume that travellers will spend 7 days in the city of their choice and as such we have selected 7 clusters using the k-means clustering algorithm. The travellers are expected to spend 1 day in each of the 7 clusters. The places *within* each cluster are relatively similar, however the places are relatively different *between* each cluster.

Data Visualisation

The idea is that neighbourhoods that are similar will be grouped within the same clusters using machine learning. We use a map to visualise the results from the k-means clustering, where similar clusters share the same colours (Figure 6) – this provides the basis to answer the questions posed at the start of the report.

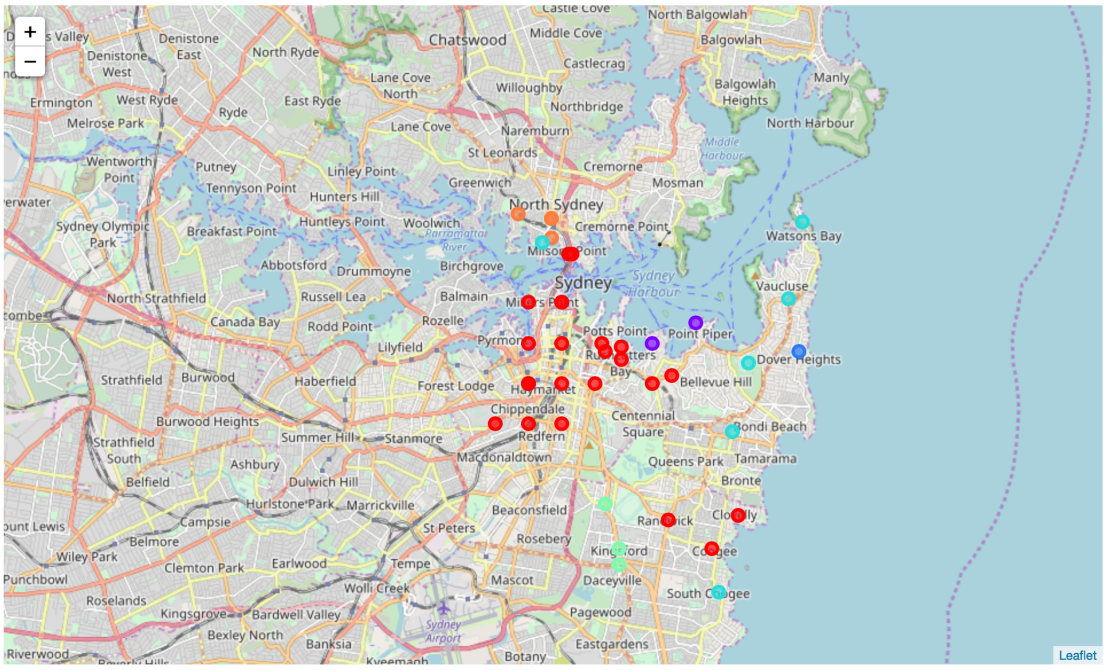

Figure 6: A Map Visualisation of the Neighbourhood Clusters in Sydney

Results Interpretation – Selecting a Neighbourhood to Visit

Using the results above, travellers could seek to improve their travel experiences, whereby they are able to select the neighbourhoods according to their own preferences. Furthermore, by reducing visits to places that are similar (i.e. within the same clusters), travellers are using their time more wisely and this could translate into tangible money savings.

For example, travellers who have an interest in visiting cafés could select one of the neighbourhoods (e.g. The Rocks) listed under Cluster 0 as shown in Figure 7 below. Cluster 0 is visualised in red in Figure 6.

	Neighbourhood	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
0	Dawes Point	0	Café	Hotel	Australian Restaurant	Cocktail Bar	Pub	Hotel Bar
1	Haymarket	0	Café	Chinese Restaurant	Thai Restaurant	Coffee Shop	Malay Restaurant	Ramen Restaurant
2	Millers Point	0	Café	Chinese Restaurant	Seafood Restaurant	Park	Steakhouse	Middle Eastern Restaurant
3	Sydney	0	Café	Coffee Shop	Shopping Mall	Cocktail Bar	Hotel	Bar
4	The Rocks	0	Café	Hotel	Australian Restaurant	Cocktail Bar	Pub	Hotel Bar
5	World Square	0	Café	Thai Restaurant	Japanese Restaurant	Coffee Shop	Burger Joint	Cocktail Bar
7	University Of Sydney	0	Café	Performing Arts Venue	Italian Restaurant	Library	Pub	Park
8	Ultimo	0	Café	Chinese Restaurant	Thai Restaurant	Coffee Shop	Malay Restaurant	Ramen Restaurant
9	Chippendale	0	Café	Chinese Restaurant	Thai Restaurant	Coffee Shop	Malay Restaurant	Ramen Restaurant
10	Darlington	0	Café	Bar	Pub	Pizza Place	Thai Restaurant	Fast Food Restaurant
11	Pyrmont	0	Café	Hotel	Bar	Italian Restaurant	Australian Restaurant	Burger Joint
12	Surry Hills	0	Café	Pub	Lebanese Restaurant	Japanese Restaurant	Coffee Shop	Pizza Place
13	Darlinghurst	0	Café	Italian Restaurant	Bar	Bakery	Bookstore	Burger Joint
14	Woolloomooloo	0	Café	Hotel	Australian Restaurant	Italian Restaurant	Wine Bar	Chinese Restaurant
15	Elizabeth Bay	0	Café	Park	Wine Bar	Italian Restaurant	Hotel	Chinese Restaurant

Figure 7: Neighbourhood Cluster 0 in Sydney

Conversely, Cluster 4 is suitable for travellers looking to explore both Chinese restaurants and Indonesian restaurants as shown in Figure 8 below. Cluster 4 is visualised in green in Figure 6.

	Neighbourhood	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
29	Daceyville	4	Chinese Restaurant	Indonesian Restaurant	Café	Grocery Store	Malay Restaurant	Thai Restaurant
30	Kingsford	4	Indonesian Restaurant	Chinese Restaurant	Café	Italian Restaurant	Grocery Store	Fast Food Restaurant
31	Kensington	4	Chinese Restaurant	Indonesian Restaurant	Indian Restaurant	Convenience Store	Burger Joint	Café

Figure 8: Neighbourhood Cluster 4 in Sydney

3.2 Methodology – By City/Country Clusters

Neighbourhood Labelling

Using the information from the data section, we labelled the neighbourhoods according to their latitude and longitude as shown in Figure 9 below.

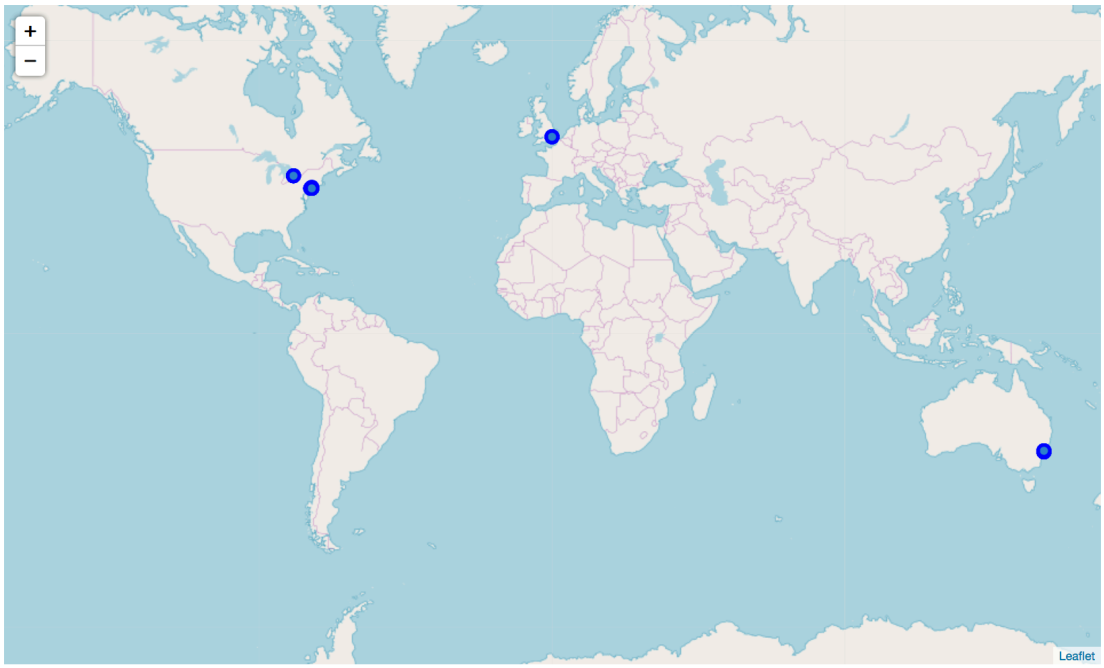

Figure 9: A Map Visualisation of the Four Countries

Machine Learning Application – k-Means Clustering

The *second question* in the introduction of this report is resolved by **combining** the underlying data for each city into one master data, following which we applied the k-means clustering algorithm as described in the previous section.

Data Visualisation

The idea is that neighbourhoods that are similar will be grouped within the same clusters using machine learning. We use a map to visualise the results from the k-means clustering, where similar clusters share the same colours (Figure 10) – this provides the basis to answer the questions posed at the start of the report.

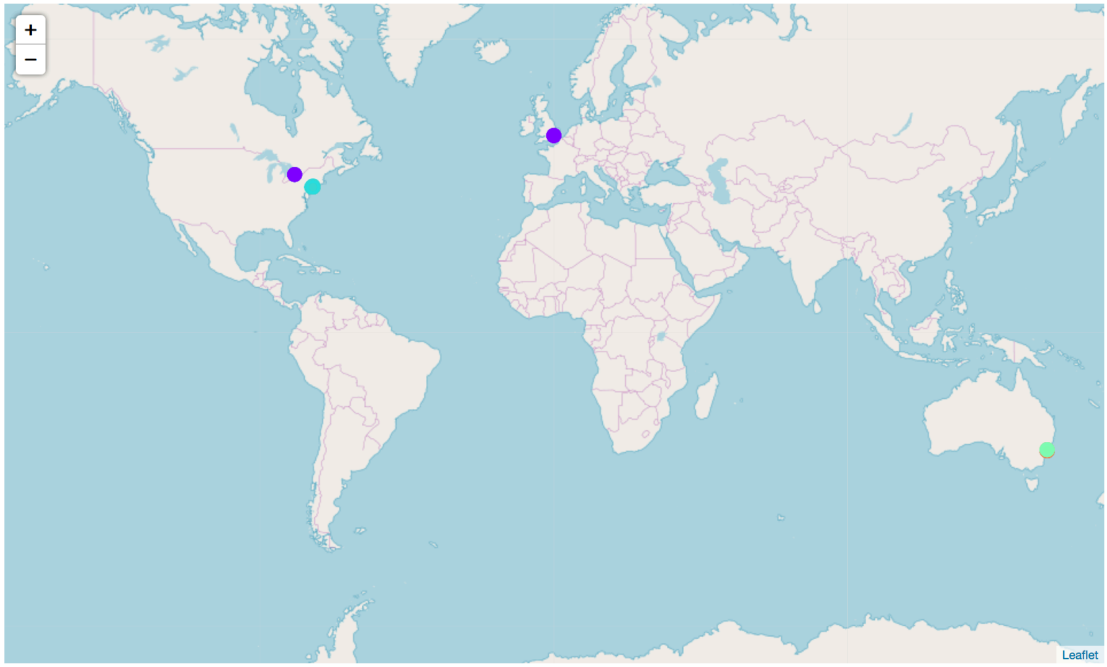

Figure 10: A Map Visualisation of the Neighbourhood Clusters in the Four Countries

Results Interpretation – Selecting a City/Country to Visit

Using the results above, travellers could seek to improve their travel experiences, whereby they are able to select the city/country according to their own preferences.

For example, as Sydney neighbourhoods in Cluster 4 are dominated by cafés, travellers who have an interest in visiting cafés could select one of the neighbourhoods located in Sydney (Figure 11). This cluster is visualised in green in Figure 10.

	Neighbourhood	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
5	Christie	4	Grocery Store	Café	Park	Baby Store	Italian Restaurant	Coffee Shop	Restaurant	Nightclub	Convenience Store	Diner
23	World Square	4	Café	Thai Restaurant	Coffee Shop	Japanese Restaurant	Burger Joint	Cocktail Bar	Breakfast Spot	Sandwich Place	Hotel	Tapas Restaurant
25	University Of Sydney	4	Café	Performing Arts Venue	Pub	Coffee Shop	Farmers Market	Middle Eastern Restaurant	Beer Garden	Beer Bar	Bar	Thai Restaurant
28	Darlington	4	Café	Bar	Pub	Thai Restaurant	Pizza Place	Fast Food Restaurant	Burger Joint	Bakery	Restaurant	Art Gallery
30	Surry Hills	4	Café	Pub	Lebanese Restaurant	Japanese Restaurant	Coffee Shop	Pizza Place	Breakfast Spot	Sandwich Place	Gym	Greek Restaurant
31	Darlinghurst	4	Café	Bar	Italian Restaurant	Pizza Place	Bookstore	Bakery	Pub	Japanese Restaurant	Indian Restaurant	Burger Joint
32	Woolloomooloo	4	Café	Hotel	Australian Restaurant	Italian Restaurant	Chinese Restaurant	Bar	Pub	Wine Bar	Restaurant	Harbor / Marina
33	Elizabeth Bay	4	Café	Park	Wine Bar	Italian Restaurant	Bar	Australian Restaurant	Japanese Restaurant	Chinese Restaurant	Hotel	Thai Restaurant
34	Potts Point	4	Café	Australian Restaurant	Italian Restaurant	Japanese Restaurant	Hotel	Coffee Shop	Bar	Sushi Restaurant	Chinese Restaurant	Wine Bar
35	Rushcutters Bay	4	Café	Italian Restaurant	Bar	Park	Pizza Place	Wine Bar	Hotel	Coffee Shop	Japanese Restaurant	Sushi Restaurant
40	Double Bay	4	Café	Japanese Restaurant	Noodle House	Liquor Store	Cocktail Bar	Hotel	Thai Restaurant	Bar	Malay Restaurant	Beach
45	Clovelly	4	Café	Beach	Coffee Shop	Bus Line	Bakery	Burger Joint	Supermarket	Pizza Place	Hotel	Playground
46	Randwick	4	Café	Thai Restaurant	Supermarket	Gym	Pub	Fast Food Restaurant	Moroccan Restaurant	Middle Eastern Restaurant	Bakery	Gastropub

Figure 11: Sydney Neighbourhoods Dominate Cluster 4

4. Results

As noted in the introduction, I am a traveller who is keen to explore and review cafés, following which I would like to determine which neighbourhood and country I should travel to.

4.1 For a Café Lover – Which City/Country To Visit?

We note that out of the four cities – Toronto, Sydney, London and New York – the latter two cities are fundamentally different according to the results presented in Figures 12-15 below.

Even though Toronto is geographically closer to New York, the city of Toronto houses more neighbourhoods that are more similar to London than New York.

The city of Sydney has the most diversified types of neighbourhood, making it vastly different compared to the other three cities (i.e. Toronto, London and New York).

Being a fan of cafés, the city of **Sydney** offers the best place to explore and review cafés (Figure 11 and 13). Following this, we determine which neighbourhoods to visit within Sydney in the next section.

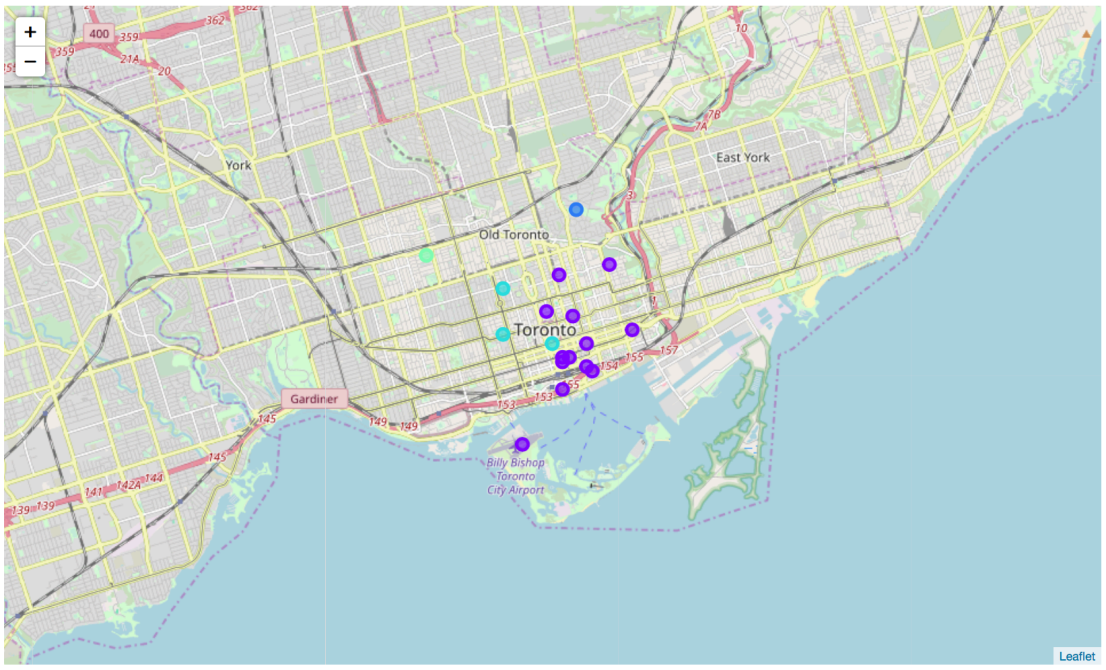

Figure 12: Neighbourhoods in Toronto (By Country Clusters)

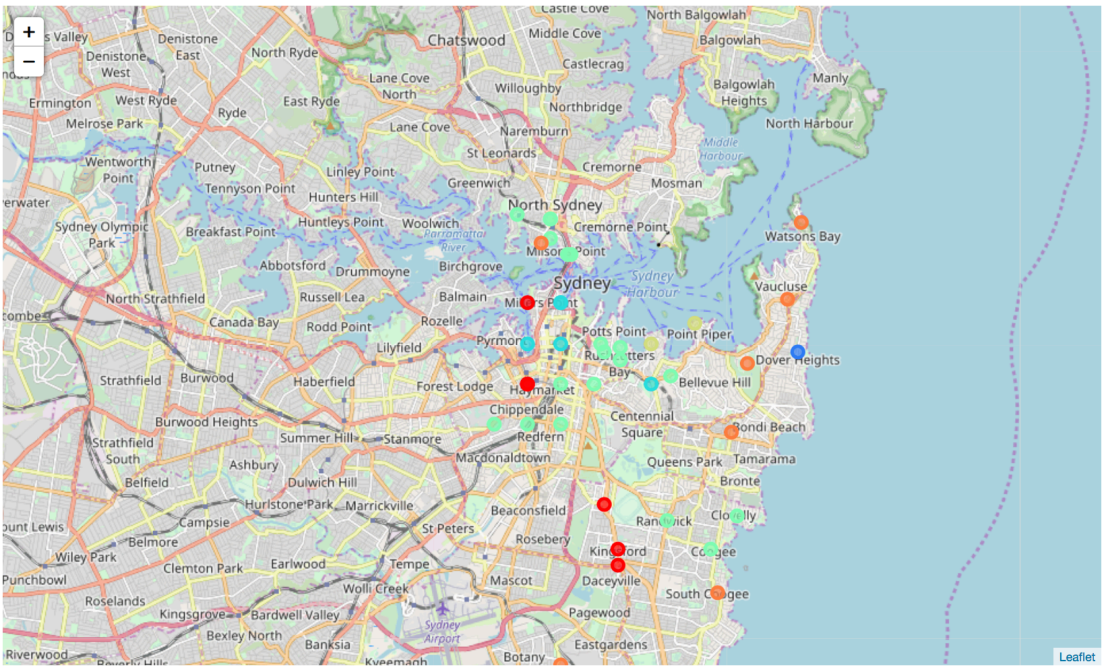

Figure 13: Neighbourhoods in Sydney (By Country Clusters)

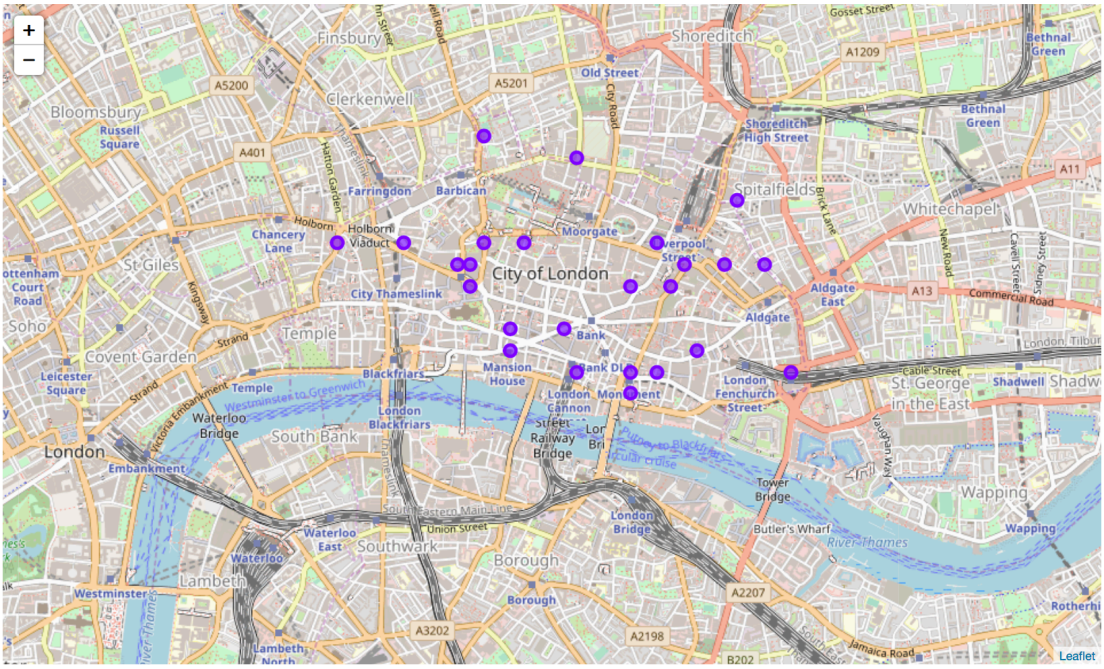

Figure 14: Neighbourhoods in London (By Country Clusters)

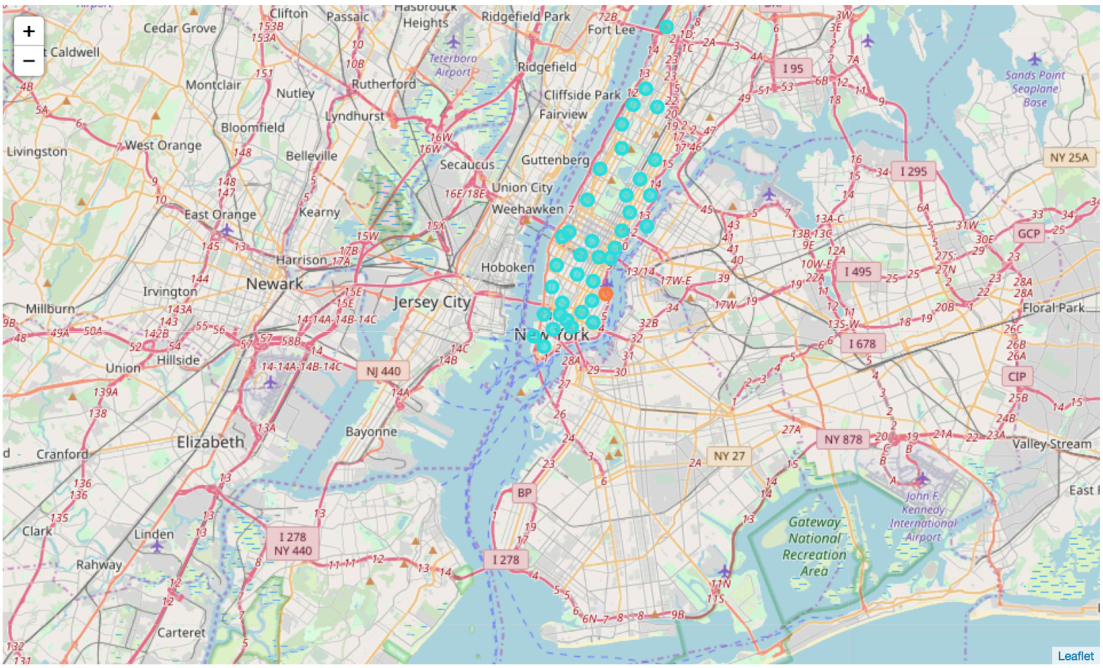

Figure 15: Neighbourhoods in New York (By Country Clusters)

4.2 For a Café Lover – Which Neighbourhood To Visit Within Sydney?

Within Sydney, I would visit the neighbourhood '**The Rocks**' as cafés are the most common venue followed by hotels (i.e. a convenient combination to visit cafés and lodge the necessary accommodation). Note that other options such as '**Dawes Point**' and '**Pyrmont**' are also feasible alternatives (Figures 16-17).

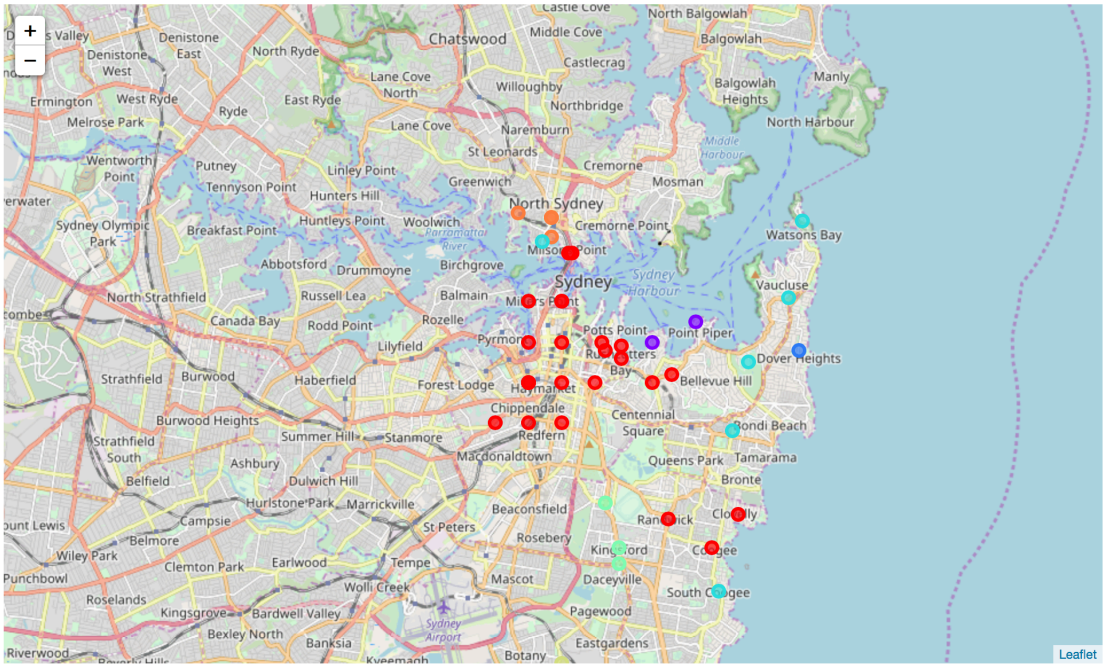

Figure 16: A Map Visualisation of the Neighbourhood Clusters in Sydney

	Neighbourhood	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
0	Dawes Point	0	Café	Hotel	Australian Restaurant	Cocktail Bar	Pub	Hotel Bar
1	Haymarket	0	Café	Chinese Restaurant	Thai Restaurant	Coffee Shop	Malay Restaurant	Ramen Restaurant
2	Millers Point	0	Café	Chinese Restaurant	Seafood Restaurant	Park	Steakhouse	Middle Eastern Restaurant
3	Sydney	0	Café	Coffee Shop	Shopping Mall	Cocktail Bar	Hotel	Bar
4	The Rocks	0	Café	Hotel	Australian Restaurant	Cocktail Bar	Pub	Hotel Bar
5	World Square	0	Café	Thai Restaurant	Japanese Restaurant	Coffee Shop	Burger Joint	Cocktail Bar
7	University Of Sydney	0	Café	Performing Arts Venue	Italian Restaurant	Library	Pub	Park
8	Ultimo	0	Café	Chinese Restaurant	Thai Restaurant	Coffee Shop	Malay Restaurant	Ramen Restaurant
9	Chippendale	0	Café	Chinese Restaurant	Thai Restaurant	Coffee Shop	Malay Restaurant	Ramen Restaurant
10	Darlington	0	Café	Bar	Pub	Pizza Place	Thai Restaurant	Fast Food Restaurant
11	Pyrmont	0	Café	Hotel	Bar	Italian Restaurant	Australian Restaurant	Burger Joint
12	Surry Hills	0	Café	Pub	Lebanese Restaurant	Japanese Restaurant	Coffee Shop	Pizza Place
13	Darlinghurst	0	Café	Italian Restaurant	Bar	Bakery	Bookstore	Burger Joint
14	Woolloomooloo	0	Café	Hotel	Australian Restaurant	Italian Restaurant	Wine Bar	Chinese Restaurant
15	Elizabeth Bay	0	Café	Park	Wine Bar	Italian Restaurant	Hotel	Chinese Restaurant

Figure 17: Neighbourhood Cluster 0 in Sydney

5. Discussion

We noted in the previous section that out of the four cities – Toronto, Sydney, London and New York – the latter two cities are fundamentally different according to the results presented in Figures 12-15.

Toronto as a Mixture of London and New York

Even though Toronto is geographically closer to New York, the city of Toronto houses more neighbourhoods that are more similar to London than New York.

In essence, Toronto has a mixture of neighbourhood styles from both London and New York. Therefore, travellers are able to experience the neighbourhood cultures from both London and New York just by visiting Toronto alone. This proposition is relevant for travellers that have travel budget constraints but would like to make full use of their travel expenses.

Sydney for Travellers Looking for Variety

The city of Sydney has the most diversified types of neighbourhood, making it vastly different compared to the other three cities (i.e. Toronto, London and New York).

Out of the 7 clusters, Sydney houses 6 of the neighbourhood cluster types, making it suitable for travellers looking for a variety of exposures during their travel.

London as a Hub for Coffee Shops

The City of London is dominated by coffee shops, making it a popular destination for coffee lovers. Restaurants (in particular Italian restaurants) are the second most common venue after coffee shops – a point that is highly relevant to pasta and pizza enthusiasts.

New York – Ideal for Families

The popular venues in New York are parks, gyms, clothing stores, American restaurants, Japanese restaurants, theater, spa, exhibitions, supermarket, hotel and ice cream shops, making it an ideal travel destination for families.

6. Conclusion

This report seeks to answer the question on which neighbourhood and city/country a traveller should visit by studying the venues located in Toronto, Sydney, London and New York. Specifically, being a café lover myself, where is the best place to visit? The answer is **The Rocks, Sydney, Australia**.

Through the application of machine learning (specifically the k-means clustering algorithm), we studied the neighbourhood composition in each city. From the analysis, we draw the following conclusions for each of the four different cities/countries:

- **Toronto as a Mixture of London and New York**
  - Travellers are able to experience the neighbourhood cultures from both London and New York by just visiting Toronto alone.
- **Sydney for Travellers Looking for Variety**
  - The city of Sydney has the most diversified types of neighbourhood, making it vastly different compared to the other three cities.
- **London as a Hub for Coffee Shops**
  - The City of London is dominated by coffee shops, making it a popular destination for coffee lovers.
- **New York – Ideal for Families**
  - The popular venues in New York are parks, gyms, clothing stores, American restaurants, Japanese restaurants, theater, spa, exhibitions, supermarket, hotel and ice cream shops.