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. Upon completion of the analysis, we aim to draw a number of high level conclusions such as those outlined below 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.