**Coursera IBM Data Science Capstone project**

**Report Michael Weinberger**

**January 2021**

**1. Introduction/Business Problem**

In today's world people frequently need to relocate, be it due to work requirements or for personal reasons. Often, this involves moving to a city that you have not lived in before and may never have visited.

So how do you decide which part of a city that you do not yet know you would prefer to live in?

One way would be to base this decision on past experience and start from a neighbourhood that you know and like. You would then choose a new neighbourhood based on the fact that is very similar to the one you like. This project allows you to do just that: It uses data on the presence of various kinds of local venues (<https://foursquare.com/>) to determine neighbourhoods that best match a location freely chosen by the user.

The project therefore is aimed at a broad audience, namely anyone who would like a personalised recommendation on which new neighbourhood best to move to. It therefore addresses a common need. It is distinct from existing property marketing resources in that it helps to choose a neighbourhood, rather than a specific property.

To reach its audience, the project can be implemented as a web-based app. The app would furthermore be able to give additional information, such as the kinds of venues that define a chosen neighbourhood.

## 2. Data

The project uses neighbourhood data (names and locations) from four cities: San Francisco, New York, Toronto and London.

The San Francisco data have been used in the Data Visualization with Python course (<https://cocl.us/sanfran_geojson>). Names were isolated from the geojson file "features" sections and "properties" subsections. The mean values of the "geometry" section and "coordinates" subsection values for each neighbourhood were used as neighbourhood locations.

The New York neighbourhood name and location data are taken from <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json> (as used in week 3 of the Data Science Capstone Project course). Names were isolated from the json file "properties" sections and "name" subsections. Locations were isolated from the "geometry" sections and "coordinates" subsections.

The Toronto neighbourhood name data are taken from <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>, more specifically from the "Toronto - 103 FSAs" table on the website. The neighbourhood location data was acquired from <http://cocl.us/Geospatial_data> (as used in week 3 of the Data Science Capstone Project course) and merged with the name data using the postal code information.

The London neighbourhood name and location data are taken from <https://en.wikipedia.org/wiki/List_of_London_boroughs>. The "List of boroughs and local authorities" table on the website was used to get names and coordinates of the London neighbourhoods.

Data on venues was acquired using the Foursquare API (<https://foursquare.com/>). Venue data was collected for each neighbourhood location using a search radius of 500 meters and an upper limit of 100 found venues. Venue data was collected using personal Foursquare developers account credentials.