Background Introduction

A large multi national firm offers secondment opportunities to different locations around the world for employees; this can enhance both the employee's and the businesses performance by offering a variety of experience and exposure to alternative markets.

The HR team have asked a data scientist to produce an analysis that can make pertinent suggestions of where to base living arrangements for employees who are about to be seconded.

This is important for the firm to provide a high standard of job satisfaction, looking after their employees needs in what can be a difficult transition abroad for many people. Ideally this analysis would recommend regions of cities that are most similar to the employee's home city in terms of nearby amenities.

Problem

- Given the current living location of an employee, can we as the analyst find the most similar postal/zip location in alternative cities given a list of common amenities?
- In our example we will find the most suitable zip code within Boston and New York for an employee who is currently living in Camden, London. We will compare the two cities to London and determine which is the most suitable location for our employee's secondment, considering only local amenities in this case.
- This will be done by performing a similarity analysis in a K Means Classifier on the list of common venues to determine the locations with the most similarity to Camden Town.
- A general visual comparison of the two cities can also be performed from the results of the analysis.

Data Sources and Usage

The following data sources will be used to demonstrate our solution to this location clustering problem:

- City Postal Codes:
    - A list of London Postcodes with latitude and longitude of their centre.
    - A list of Boston, MA zip codes with latitude and longitude of their centre.
    - A list of New York, NY zip codes with latitude and longitude of their centre.
        - To obtain the postcodes, we will perform web-scraping.
        - To obtain the latitude and longitude, we will use a the geopy geocode python module to extract the data from the post/zip address, amending the results to our data frame.
        - This will be a combination of categorical and numeric data, the numeric data will be used in our Foursquare API calls.
- Foursquare venue data for all postcodes/zipcodes to produce a data frame of amenities.
    - We will select the 100 nearest amenities in a 1000m radius to the post code coordinates we have been given.
        - This data will be categorical so to train a K means classifier, will be converted to numeric data via one hot encoding.
        - To perform representative analysis, we will reduce the data frame to contain only common venue types across the three cities. We will then train our classifier on this data.
        - A similarity analysis will allow us to characterise the locations in Boston and New York that are most similar to Camden Town and to perform a general visual comparison of the three cities.