# Applied Data Science Project – Capstone Report

## Contents

# Introduction

Previously in the Applied Data Science Capstone project we compared neighbourhoods within both New York and Toronto. In this project we are going to compare neighbourhoods between 2 different cities - New York, USA and London, UK.

In New York there are 306 different neighbourhoods across 5 different boroughs. There are similarities between the neighbourhoods within each of these boroughs.

I want to be able to compare area from another city to decide which borough that area is most similar to. For example, is Toronto's Scarbourgh Village neighbourhood most similar to neighbourhoods found in the Bronx, Brooklyn, Manhattan, Queens or Staten Island boroughs in New York. As I am from UK, I will be comparing area in London, UK to New York, USA.

## Scenario

Imagine I am a Data Scientist who currently works for Google in New York, work at the Google offices in New York which are based Chelsea, Manhattan and live in the borough of Manhattan.

I have recently been asked to relocate for work to Google's offices near Kings Cross in London, UK. However, I currently know nothing about London or what it is like as a place to live. Before deciding where I would like to move to, I would like to first narrow down my options.

Even though I don't know anything about London I know that I enjoy living in the Manhattan area of New York and would therefore like to know what areas of London are most similar to Manhattan by using available Foursquare API data. In addition, I'd also not like to be too far from my office near Kings Cross.

We first look at all areas of London to establish which New York boroughs they are most similar to. Then from the areas which are most similar to Manhattan, we will then establish how far from the new Google offices near King Cross they are to establish a shortlist of 3 areas I should consider living in when I moved to London.

## Notes

This could be a comparison of any two cities where the data is readily available. London and New York are just a demonstration of it being successfully applied.

Moreover, this general idea could be used in a different context, for example a business that has many successful stores in Manhattan that is looking to expand to London and is unsure of the best area to

# Data

## New York Neighbourhood Data

The data for New York's borough and neighbourhood data will come from, used in the previous practical in the Applied Data Science Capstone

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

This dataset contains every Neighbourhood in New York, the borough they are in as well as their longitude and latitude. The data frame has 5 boroughs and 306 neighbourhoods.

Data scraped from https://en.wikipedia.org/wiki/Boroughs_of_New_York_City reveals:

```
Total population of New York : 8,336,817
Average population of New York Borough: 1,667,363.4
Average population of New York Neighbourhood: 27,244.5
```

|   | Borough | Population Estimate (2019) |
|---|---|---|
| **0** | The Bronx | 1418207 |
| **1** | Brooklyn | 2559903 |
| **2** | Manhattan | 1628706 |
| **3** | Queens | 2253858 |
| **4** | Staten Island | 476143 |

**The population of each of these boroughs is approximately 1,700,000 people per borough**

**The population of each of these boroughs is approximately 27,000 people per neighbourhood**

## London Borough Data

The data for London's borough and data will come from data scraped from Wikipedia

https://en.wikipedia.org/wiki/List_of_London_boroughs

This dataset contains every Borough in London as well as the longitude and latitude coordinate (which will need to be wrangled into the correct form). There are additional features which are not required.

The data frame has 32 boroughs (and we do not include the City of London, as it is not a borough)

Data scraped from https://en.wikipedia.org/wiki/List_of_London_boroughs reveals:

```
Total population of London : 8,408,887
Average population of London Borough: 262,777.71875
```

**The population of each of these boroughs is approximately 260,000 people per borough**

|  | Borough | Population Estimate (2013) |
|---|---|---|
| **0** | Barking and Dagenham [note 1] | 194352 |
| **1** | Barnet | 369088 |
| **2** | Bexley | 236687 |
| **3** | Brent | 317264 |
| **4** | Bromley | 317899 |
| **5** | Camden | 229719 |
| **6** | Croydon | 372752 |
| **7** | Ealing | 342494 |
| **8** | Enfield | 320524 |
| **9** | Greenwich [note 2] | 264008 |
| **10** | Hackney | 257379 |
| **11** | Hammersmith and Fulham [note 4] | 178685 |
| **12** | Haringey | 263386 |
| **13** | Harrow | 243372 |
| **14** | Havering | 242080 |
| **15** | Hillingdon | 286806 |
| **16** | Hounslow | 262407 |
| **17** | Islington | 215667 |
| **18** | Kensington and Chelsea | 155594 |
| **19** | Kingston upon Thames | 166793 |
| **20** | Lambeth | 314242 |
| **21** | Lewisham | 286180 |
| **22** | Merton | 203223 |
| **23** | Newham | 318227 |
| **24** | Redbridge | 288272 |
| **25** | Richmond upon Thames | 191365 |
| **26** | Southwark | 298464 |
| **27** | Sutton | 195914 |
| **28** | Tower Hamlets | 272890 |
| **29** | Waltham Forest | 265797 |
| **30** | Wandsworth | 310516 |
| **31** | Westminster | 226841 |

## New York 'Boroughs' and London 'Boroughs' are not the same

We can see by comparing average population sizes of New York 'boroughs' and London 'boroughs' that they are not the same size as one another. A New York borough has approximately 6 to 7 times the population of a London Borough (~1,700,000 and ~260,000 respectively). As such, the boroughs are not comparable. A New York borough is larger than than a London borough.

However, we can also that we can see by comparing average population sizes of a New York neighbourhoods and London boroughs they are not the same size as one another. A London borough has approximately 10 times the population of a New York Borough (~260,000 and ~27,000 respectively). A London Borough is larger than a New York Borough.

It us worth noting that a London Borough is somewhere in between the size of a New York neighbourhood and borough. Consequently, it will be crucial to normalise features to allow for fair comparison.

## Similarity data - Foursquare API

We will be utilizing the Foursquare API to establish the features of the neighborhoods in New York. We will collect the types of venues within each neighbourhood. We also use the Foursquare API to establish the features of Boroughs in London. We will collect the types of venues within each borough.

To encode a venue list from the Foursquare API into features we will use one-hot encoding over the venues type in both London Boroughs and New York neighbourhoods, this will ensure features between the two data sets are the same (crucial for comparison later).

After one-hot encoding, we then want to group by New York Neighbourhood and London Boroughs by taking the mean of the frequency of occurrence of each category, to normalise features. We can call this data our feature space ($X$). At this point we should clearly label our New York Neighbourhood data for train and cross-validation data and London Borough data as test data ($X_{test}$). To the New York Neighbourhood feature data from Foursquare (X) we can add New York Borough labels which have been given an numerical encoding (y), as well splitting it into train ($X_{train}, y_{train}$) and cross-validation datasets ($X_{cv}, y_{cv}$) (stratified by Borough, 80% - train and 20% Cross Validation).

I then to use the K-Nearest Neighbour Classification model on the New York Neighbourhood data with Borough label to train a Borough classification model. (i.e. given a Neighbourhoods features it should be able to classify what Borough it belongs to). We will use the 20% cross-validation set from New York data to establish the best value for K (the number of nearest neighbours), by finding the best accuracy for a range of possible K values.

I will then use the test data for London Boroughs to classify them according to what New York Boroughs they are most similar to. Then restrict data to London Boroughs most similar to Manhattan.

## Distance from new Office

The current Google office in London can be found on Google Maps
https://www.google.com/maps/place/Google+UK/@51.5332609,-0.1281919,17z/data=!3m1!4b1!4m5!3m4!1s0x48761b3c54efa6e1:0xc7053ab04745950d!8m2!3d51.5332609!4d-0.1260032
Right clicking on the pop-up and clicking What's here reveal the longitude and latitude for the building Latitude = 51.5332609 Longitude = -0.1260032. Using the Longitude and Latitude in the London data set allows us to calculate the Birdseye distance from the Borough to the office. We will then be able to sort our list of London Boroughs like Manhattan by this distance to give a shortlist of 3 London Boroughs which are most like Manhattan and are close to the Google office in London.

# Methodology

## Features established using Foursquare API

Features were established using venue types within that area using the Foursquare API. For New York data, we searched for venues within a 500m radius of the co-ordinates of each neighbourhood, limited to a maximum of 100 venues. For London data, as we observed that their population was around 10 times that of a New York neighbourhood (260,000 average population for a London Borough and for a New York neighbourhood it was 27,000). Consequently, I assumed area would also be approximately 10 times as large, meaning a radius would be around square root of 10 which is approximately 3. As such the radius for London data was 1,500m (3 times 500m), also limited to a maximum of 100 venues.

## Feature Encoding

After importing Foursquare venue data for both New York Neighbourhoods and London Boroughs, it was crucial to encode the venue types using one-hot encoding. Both data sets were concatenated together to ensure that the features set produced by the encoding was the same across both data. The features were then normalised by calculating the proportion of each type of venue was in each area. For example, if an area had 9 Cafes out of total of 90 venues in that area, Cafes were given a value of 0.1. The sum of all normalised venue type features was 1. Once the features had been derived by encoding, the data was separated back into New York and London data. This process produced two separate data sets with the same features.

## Data Separation

The New York data was split into a train (70%) and cross-validation set (30%) using the train_test_split method in sklearn. The reason behind this was that I wanted to the KNN Classifier, and consequently the hyper-parameter k required establishing using a cross-validation dataset. The London data was then used to establish the test dataset. As such we classified the data as follows:

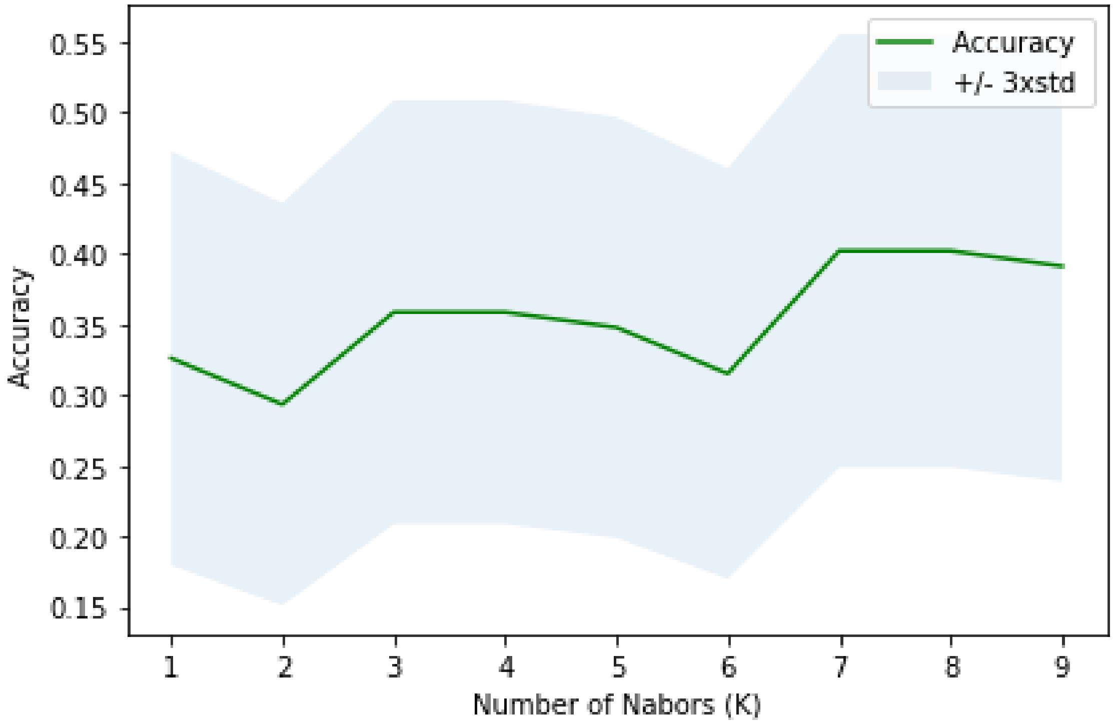
- X_train – feature data only from 70% of New York Neighbourhoods. Used for model fitting
- y_train – the Borough data from 70% of New York Neighbourhoods. Used for model fitting
- X_cv – feature data only from 30% of New York Neighbourhoods. Used for establishing best k
- y_cv – the Borough data from 30% of New York Neighbourhoods. Used for establishing best k
- X_test – feature data of all London Boroughs. Used to predict most similar New York Borough

## Machine Learning Techniques Used

We used a KNN Classifier to build a model which given a list of features (proportion of venue types in that area) predicted which New York Borough a Neighbourhood belonged to. 70% of the New York neighbourhoods were used to train the model and 30% of New York neighbourhoods were used for choosing the best value of k. Deciding which New York Borough a London Borough is most similar to is classification problem and KNN Classifier was used as it is a classifying algorithm technique I was familiar with. A potential draw back of this approach is that it may suffer from overfitting due to dimensionality issues as the number of features is larger than the number of examples. The number of features was 468, but the number of New York neighbourhoods was lower than this 306 in total, 215 in the training set and 91 in the cross-validation set. In future, before training the model it may be advantageous to the reduce the number of features, using a method such as Principal Component Analysis (PCA).

## Selection of k parameter for KNN

The cross-validation New York data set was used to establish the best hyper-parameter k. A KNN Classifier model was constructed for various possible k values then trained using X_train and y_train. The accuracy of each model was then established by using the X_cv dataset to predict which Borough in New York each neighbourhood belong to. This was then compared to the actual Borough it was in using y_cv dataset. After testing several possible values of k, a k hyper-parameter of 7 produced the most accurate model, with a model accuracy of 40.2% on the cross-validation set. It is worth noting that this level of accuracy is relatively poor – randomly guessing Borough would have an expected accuracy of 20%.



```
The best accuracy was with 0.40217391304347827 with k=7
```

## Using KNN Classifier model on London feature data to find most similar NY Borough

The model that had been trained on 70% of New York data with k=7 was then used on the feature data from London to classify each London Borough to the New York Borough it was most like. This is discussed in more detail in the Results. However, all but 2 Boroughs were classified as being most similar to Manhattan, which was not particularly useful for this particular use case. The data set was restricted to only London Boroughs which were most similar to Manhattan.

## Establishing distance from Borough to new office location

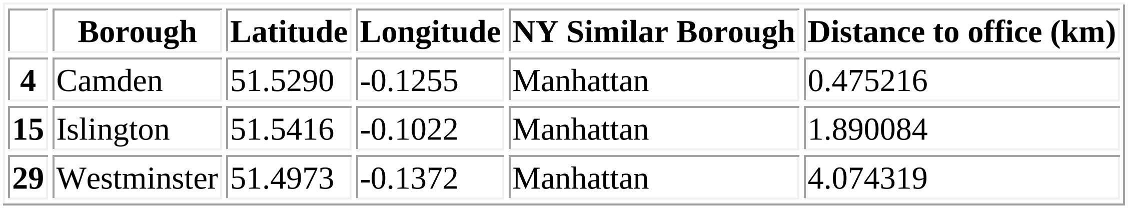
The latitude and longitude of the new office and each of the London Boroughs was used to calculate the distance between the Borough and the new office location. The method used is taken from https://stackoverflow.com/questions/19412462/getting-distance-between-two-points-based-on-latitude-longitude. After calculating the distance to the office the Boroughs most similar to Manhattan are then sorted by their distance from the office, and the closest 3 Boroughs taken.

## Results

The KNN Classifier used to classify each London Borough as the New York Borough it was most similar showed all but two boroughs (Barnet and Hillingdon) were most similar to Manhattan. This didn't help to reduce the Boroughs much for the problem of the best Borough the Data Scientist should consider living in when they relocate to London. It also suggests that most London Boroughs are most similar to Manhattan. Alternatively, this could be due to overfitting caused dimensionality issues as the number of features is larger than the number of examples.

| | Borough | Latitude | Longitude | NY Similar Borough |
|---|---|---|---|---|
| **0** | Barking and Dagenham | 51.5607 | 0.1557 | Manhattan |
| **1** | Barnet | 51.6252 | -0.1517 | Brooklyn |
| **2** | Bexley | 51.4549 | 0.1505 | Manhattan |
| **3** | Brent | 51.5588 | -0.2817 | Manhattan |
| **4** | Bromley | 51.4039 | 0.0198 | Manhattan |
| **5** | Camden | 51.5290 | -0.1255 | Manhattan |
| **6** | Croydon | 51.3714 | -0.0977 | Manhattan |
| **7** | Ealing | 51.5130 | -0.3089 | Manhattan |
| **8** | Enfield | 51.6538 | -0.0799 | Manhattan |
| **9** | Greenwich | 51.4892 | 0.0648 | Manhattan |
| **10** | Hackney | 51.5450 | -0.0553 | Manhattan |
| **11** | Hammersmith and Fulham | 51.4927 | -0.2339 | Manhattan |
| **12** | Haringey | 51.6000 | -0.1119 | Manhattan |
| **13** | Harrow | 51.5898 | -0.3346 | Manhattan |
| **14** | Havering | 51.5812 | 0.1837 | Manhattan |
| **15** | Hillingdon | 51.5441 | -0.4760 | Brooklyn |
| **16** | Hounslow | 51.4746 | -0.3680 | Manhattan |
| **17** | Islington | 51.5416 | -0.1022 | Manhattan |
| **18** | Kensington and Chelsea | 51.5020 | -0.1947 | Manhattan |
| **19** | Kingston upon Thames | 51.4085 | -0.3064 | Manhattan |
| **20** | Lambeth | 51.4607 | -0.1163 | Manhattan |
| **21** | Lewisham | 51.4452 | -0.0209 | Manhattan |
| **22** | Merton | 51.4014 | -0.1958 | Manhattan |
| **23** | Newham | 51.5077 | 0.0469 | Manhattan |
| **24** | Redbridge | 51.5590 | 0.0741 | Manhattan |
| **25** | Richmond upon Thames | 51.4479 | -0.3260 | Manhattan |
| **26** | Southwark | 51.5035 | -0.0804 | Manhattan |
| **27** | Sutton | 51.3618 | -0.1945 | Manhattan |
| **28** | Tower Hamlets | 51.5099 | -0.0059 | Manhattan |
| **29** | Waltham Forest | 51.5908 | -0.0134 | Manhattan |
| **30** | Wandsworth | 51.4567 | -0.1910 | Manhattan |
| **31** | Westminster | 51.4973 | -0.1372 | Manhattan |

After reducing the London data set to Boroughs which are similar to Manhattan, using the longitude and latitude to calculate the distance to the office and then sorting by distance office revealed the top 3 locations to consider living in when relocating to London. Camden, Islington or Westminster.



| | Borough | Latitude | Longitude | NY Similar Borough | Distance to office (km) |
|---|---|---|---|---|---|
| **4** | Camden | 51.5290 | -0.1255 | Manhattan | 0.475216 |
| **15** | Islington | 51.5416 | -0.1022 | Manhattan | 1.890084 |
| **29** | Westminster | 51.4973 | -0.1372 | Manhattan | 4.074319 |

## Discussion

I found that using the KNN Classifier identified all except 2 of the London Boroughs as most similar to Manhattan. This did little reduce the number of Boroughs to consider, so in this particular problem it did little to help reduce the number of Boroughs to consider living in. There are some possible causes for this in the methodology used which could be changed in future. It could be due to the dimensions of the feature set used to train the model. The New York data had 468 features and 306 examples, this increases the likelihood of overfitting. It may be sensible to try to reduce the number of features using a technique such as Principle Component Analysis (PCA).

There were also a few other limitations of the methodology used to collect feature set data. It is worth noting that when collecting venues from the Foursquare API is limited to a maximum limit of 100 venue for each New York Neighbourhood and London Boroughs, there were several which met this limit of 100. This limit is due the limit of 50,500 daily API calls in Foursquares free developer account. This limit could be increased to ensure that all venues in each Neighbourhood/Borough is found to ensure the features have accounts for all venues.

It’s also worth noting that venues from the Foursquare API are fetched based on the distance from the centre of New York neighbourhood/London borough. This is potentially flawed in 2 major ways, the radius used could fail to include all venues in a neighbourhood/borough and there is the potential for overlap between the radii of neighbourhoods/boroughs which could mean that a venue could appear in multiple neighbourhoods/boroughs. Perhaps a more robust approach could be to generate a large list of venues, and use zipcode/postcode data to classify which neighbourhoods/boroughs they belonged to avoid both missing venues and duplicate venues in multiple areas.

## Conclusion

It appears that most London Boroughs (apart from Barnet and Hillingdon) are similar to Manhattan, so in terms of locations to consider relocating to London most are similar to London and should be suitable for our Data Scientist who wants to live somewhere like Manhattan. As such the primary factor in considering where to relocate to is going to be the distance from the office location. Consequently, the best 3 Boroughs to consider living in are Camden, Islington, and Westminster. This should help our Data Scientist to reduce the number of Boroughs to consider when moving to London.