# FinalProject_Report

November 27, 2018

## 1 Capstone Project

**Identifying districts in London that are similar to a chosen New York neighborhood** *By Elena Tuchina*

### 1.0.1 Table of Contents



---

### 1.0.2 Business Problem

The relocation company that specializes in relocating finance professionals between New York and London needs a tool to help their customers find suitable neighborhoods/districts of the city they are moving to by narrowing down a list of potential areas based on their location preferences in their home city.

In most cases, customers are not familiar enough with the city they are moving to, making the process of finding the new living location a time consuming and stressful activity.

This process can be greatly facilitated by customers' knowledge of their home city and its areas. By identifying what area(s) in their home city the new location should resemble, the cluster analysis can narrows down the list of the areas to consider to a few that the customer should further research/visit before making final decision.

For example, a customer wants to find areas of London that are similar to the neighborhood of Upper West Side in New York where he/she currently live. Cluster analysis would result in a rather short list of London districts that most closely match Midtown.

The use of the analysis would result in allowing finance professionals to relocate quickly and successfully while staying focused on their professional transition, and with reduced cost for business.

---

### 1.0.3 Data

New York City neighborhood dataset is taken from New York University Spatial Data Repository (https://geo.nyu.edu/catalog/nyu_2451_34572) and contains the boroughs and the neighborhoods that exist in each borough as well as the the latitude and longitude coordinates of each neighborhood.

The London district dataset is created by using information from Wikipedia page (https://en.wikipedia.org/wiki/List_of_areas_of_London) and contains the boroughs and the districts that exist in each borough. Only Inner London boroughs (as per London Government Act 1963) are used in this analysis. The latitude and longitude coordinates for each district were determined using Nominatum geolocator from Geopy library.

Foursquare location data is used for information on existing venues in each area of the city. The venues are aggregated by category for further frequency analysis that is served as basis for area comparison and clustering. Only areas where Foursquare database contains information on more than 100 venues were used for analysis due to smaller informational value of distribution analysis for smaller datasets.

---

### 1.0.4 Methodology

The analysis approach consists of two stages: 1) creating profile for each area, and 2) area segmentation based on their profile similarities.

To create area profiles, venue location data provided by Foursquare is obtained for each area from both New York and London datasets, and then aggregated by venue categories. Only areas with at least 100 venues reported are retained for further analysis. Next, aggregated venue data for the target New York neighborhood is added to the the aggregated venue dataset for London. This combined dataset is then converted into each area profile by calculating the frequency with which various venue categories appear. In other words, the distribution of venue categories in each area serves as that area profile. Using this approach, data in area profile is normalized and allows for direct comparison between areas.

In the second stage of analysis, areas were segmented applying k-means clustering method to area profiles. This relatively simple model is fast and sufficient for the complexity of the task at hand, and will allow for adding further dimensions to area profiles in the future.

---

### 1.0.5 Results

The analysis shows clusters of London districts on the map, with each cluster drawn in a different color.

The result represents a list of London districts, most similar to the chosen neighborhood in New York. For example, there are 9 London districts that most resemble Upper West Side in New York (Upper West Side is solely used to illustrate the workings of the code, other neighborhood choices would result in a different list of London areas).

---

### 1.0.6 Discussion

The city areas returned by the model represents a narrowed down list that most closely match customer's preferences. This list can now be used to perform in-depth research by the customer and relocation agency to take into account additional criteria and housing availability, followed by customer visit to the area before making final decision.

---

### 1.0.7 Conclusion

The application of the analysis by the relocation company would result in allowing finance professionals to relocate quickly and successfully while staying focused on their professional transition, and with reduced cost for business.

Further steps to improve the model could include incorporation of additional information that enhance the description of areas such as property data, educational (schools, universities, etc) and medical facilities, parks and other recreational centers, crime and safety records, ethnic composition etc.

---