Peer-graded Assignment: Capstone Project - The Battle of Neighbourhoods

Locating a suitable borough and then neighbourhood for personal trainer relocation: Toronto vs New York.

1.0 INTRODUCTION

1.1 Project Discussion

The project involves the analysis of boroughs and then neighbourhoods within New York and Toronto. A personal trainer is interested in relocating to increase their client base. The trainer is interested in which area may have the greatest potential customers. It is likely the area with the most gyms or related venues represents an area with significant potential customer volume. A comparison of mean occurrence of gym venues, relative to other venues, within the two cities will indicate which city, and specifically which neighbourhood, maximum customers are likely to occur.

1.2 Audience

The investigation is beneficial for personal trainers who are flexible regarding the location in which they operate. It is also useful to anyone interested in the spatial distribution of venues relating to gyms. This could be gym users, city planners, local councils, or current and future fitness centre owners.

1.3 Project Data

Foursquare location data will be utilised to highlight the venues, specifically gyms and related venues, within each borough, for New York and Toronto. The borough with the highest mean occurrence of gym related venues for both cities will then be compared. K means clustering will be utilised to indicate the distribution of gym venues within neighbourhoods. Within the boroughs that hosts the highest mean occurrence of gym venues, the neighbourhood(s) within them with the greatest sum of mean occurrence of gym related venues will be selected as the proposed area for relocation This will include the following data sources:

	Data	Source
1)	Longitude and latitudes of New York and Toronto Boroughs and Neighbourhoods.	Python GeoCoder package

2)	Venue information; specifically no. of gyms	foursquare API
3)	Boroughs and Neighbourhoods in New York	https://cocl.us/new_york_dataset
4)	Boroughs and Neighbourhoods in Toronto	https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,

2.0 METHODOLOGY

- Obtain relevant boroughs and neighborhoods for Toronto and New York; including longitude and latitudes. Toronto data scraped from Wikipedia using the Beautiful Soup library. Whilst New York data obtained from JSON file. Where necessary the Geopy Python library has been utilized to provide addition longitude and latitude information.
- Visualise the neighborhoods and then boroughs within these cities utilising the folium library.
- Use the foursquare API to pull 100 venues for each borough. Focus on the mean occurrence of venues of interest to a personal trainer; relative to others.
- Compare the sum of mean occurrence of venues of interest for boroughs within the two cities. This included categories such as: ['Gym'], ['Athletics & Sports'], ['Dance Studio'], ['Gym / Fitness Center']. Display the sum of mean occurrence of venues of interest using matplotlib library.
- Repeat steps 3 and 4 for the neighborhoods which display the greatest no. of venues of interest.
- Partition these neighborhoods into non-overlapping subsets by k-means clustering using randomly generated centroids. The aim being to identify and then visualise via folium where neighborhoods' of interest are situated.
- Compare the sum of mean occurrence of venues of interest to understand which neighborhood has the highest occurrence and therefore the greatest potential for relocation.
- Use foursquare API to pull gym related venues for final selected neighborhoods of interest. Visualise the location of these venues with the folium library.

3.0 RESULTS

Results indicated one borough in New York (Manhattan) and three boroughs in Toronto (Central, East and West Toronto) which had the highest sum of the mean occurrence of venues of interest (Fig. 1). These are the boroughs that were then investigated in more detail, looking at the neighbourhoods within them.

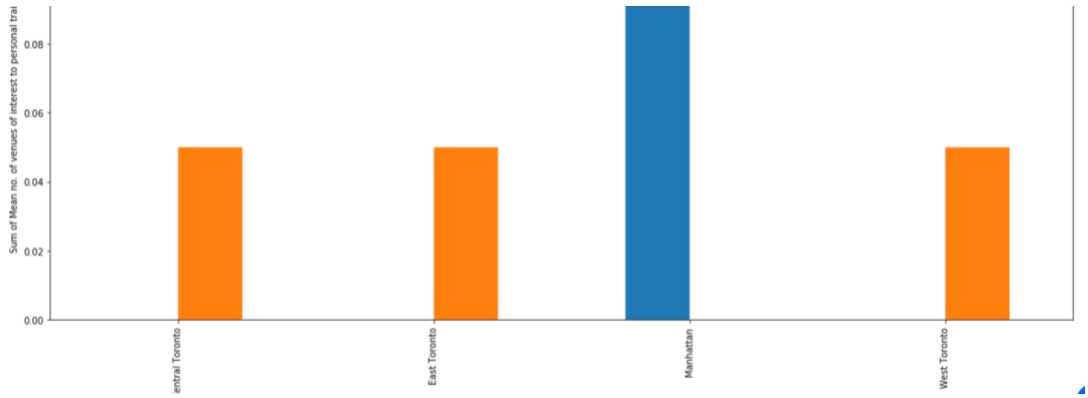

Figure 1 – Boroughs of interest that were investigated further.

Further analysis identified the neighbourhoods within these boroughs which had the greatest sum of mean occurrence of venues of interest. On the final run, these were Central Harlem, Manhattanville and Hamilton Heights for Manhattan, New York (Fig. 2a) For the 3 boroughs identified for Toronto, The Beaches proved the highest (Fig. 2b).

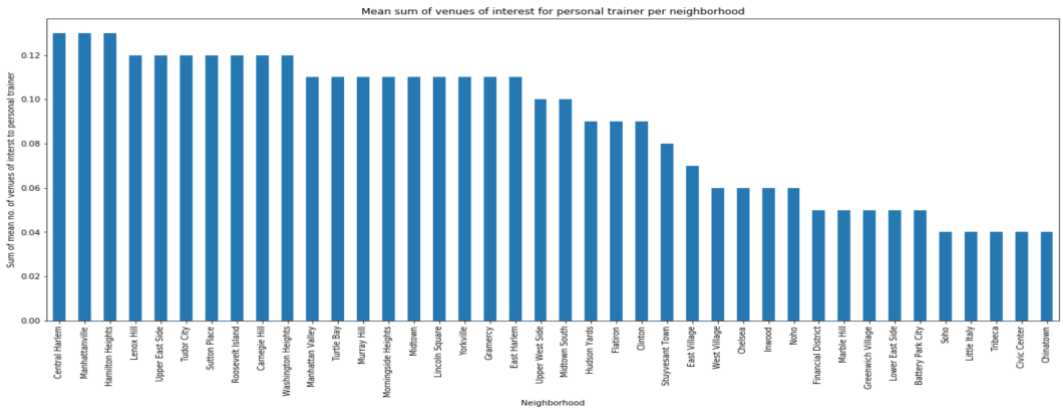

Figure 2a – Sum of mean occurrence of venues of interest in New York

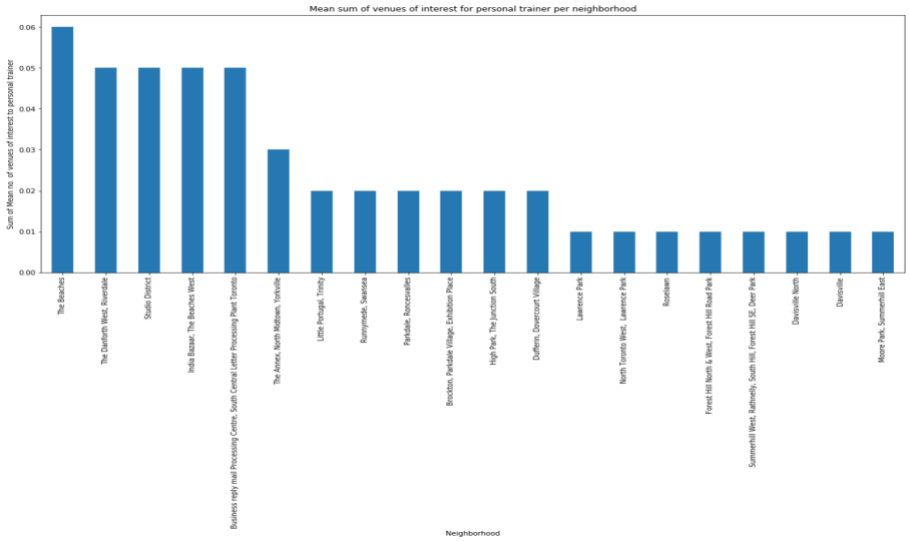

Figure 2b – Sum of mean occurrence of venues of interest in Toronto.

Clusters were produced using the random k-means clustering (Fig. 3) and were dependent on mean occurrence of identified venues of interest. Clusters were produced for both New York and Toronto (Table 1). This provides a visual representation of areas where a personal trainer may target.

Table 1 – Clusters of gym related venues

Cluster	**Note**
1 (purple)	Few gym related venues of interest.
2 (red)	Average gym related venues of interest.
3 (green)	Significant gym related venues of interest.

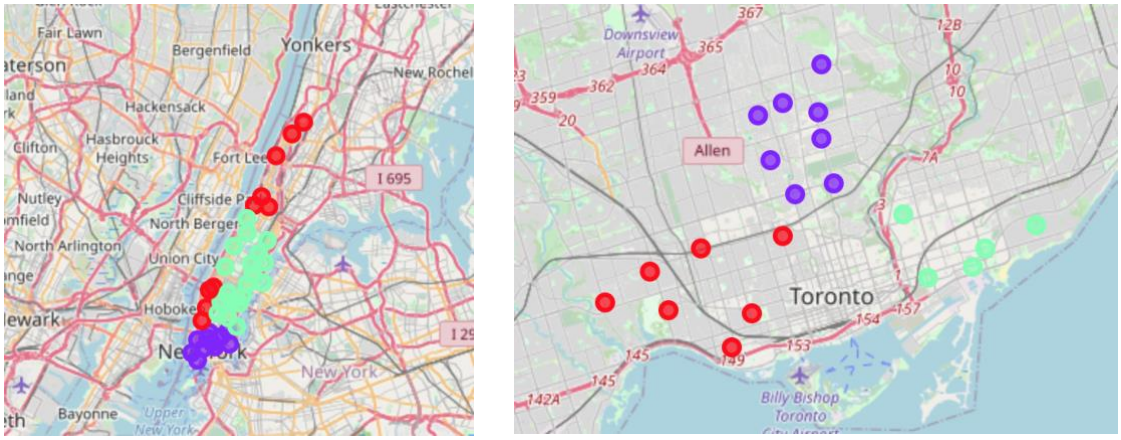

Figure 3 – K-means clusters a) New York b) Toronto.

4.0 DISCUSSION

The results indicate a series of neighbourhoods that should be investigated in more detail, these can be identified primarily by the green and secondarily by the red k-means clusters.

There are a few factors to note which influence the results of this study. In this case, foursquare returned 100 venue per location e.g. borough or neighbourhood. Therefore, the mean occurrence of venues of interest is only approximate depending on the 100 venues and may not therefore be truly representative of the wider borough or neighbourhood. The boroughs amplify this uncertainty as cover a much wider geographical area than neighbourhoods, meaning 100 venues will be less representative. The result of this is multiple runs of the API may result in different neighbourhoods appearing to have the greatest number of gym venues. Moreover, New York data pulled by foursquare API had a more comprehensive range of venue categories; this played a crucial role in creating consistently higher venue occurrence. Finally, due to the heuristic nature of the k-means algorithm, there may be value to repeat the clustering process multiple times to identify which clusters appear the most often. However, a visual inspection of the data highlights a good match between cluster segregation and the data.

5.0 CONCLUSION

On the basis of occurrence of venues of interest, **Central Harlem, Manhattanville and Hamilton Heights** has the most to offer with a sum of 0.13 of the mean occurrence of venues of interest. This is assuming a personal trainer has complete freedom to relocate.

For **Toronto, The Beaches** is the highest neighbourhood with a sum of **0.05.** Although, less venue categories of interest appeared within the foursquare API request.

Clearly other factors would have to be considered in the viability of any neighbourhood for potential relocation, such as:

- More venues per area to increase representativeness of study

- House price within each neighbourhood
- Average age of residents and their disposable income
- Presence of other personal trainers in the area
- Available accommodation for future house relocation
- Likelihood of gyms wanting to partner with personal trainer.