**Predicting the Similarity between Areas in Different Cities**

**Research Report**

**Nikita Ravi**

**Words:**

Contents

[**Introduction** 3](#_Toc16114321)

[*Background* 3](#_Toc16114322)

[*Problem* 3](#_Toc16114323)

[*Interest* 3](#_Toc16114324)

[**Data Acquisition and Cleaning** 3](#_Toc16114325)

[*Data Sources* 3](#_Toc16114326)

[*Data Cleaning and Feature Selection* 3](#_Toc16114327)

[**Methodology** 5](#_Toc16114328)

[**Results** 6](#_Toc16114329)

[**Conclusions** 6](#_Toc16114330)

[**Limitations and Future Directions** 6](#_Toc16114331)

[**References** 7](#_Toc16114332)

# **Introduction**

## *Background*

The world is becoming more globalized day by day. With an increase in the rate at which information and ideas are exchanged, people are becoming more aware of the infinite number of cultures that exist in our society, and etc. Keeping this in mind, many multinational companies take this globalization as an opportunity and aim to open their branches in different parts of the world for better profit and this accounts for increased Gross Domestic Product (GDP) and better stabilizes the host country’s economy. As of 2006, there were 63,000 recorded multinational corporations with over 700,000 branches worldwide according to the United Nations Conference on Trade and Development (Shoo, 2017).

## *Problem*

In order to start a chain in another city of another country, often times the company would like to choose a location that is similar to that of the original location of the first chain store because it is the best way to prevent an economic loss and gain profit. This project aims to predict the similarity between places in different cities, by recording the different types of venues in the area to assess the location’s entertainment preferences. This way, if a company wants to start a branch in another area in a different part of the world, they can choose an area similar to that of their original location so that they can potentially target the same audience and gain more profit / minimize economic loss.

## *Interest*

This report is targeted to multinational corporations who plan on opening a new branch in a different part of the world and help them assess which location would be the best at minimizing the company’s economic loss.

# **Data Acquisition and Cleaning**

## *Data Sources*

The postal codes and area names for Tokyo, Japan were obtained from [UPS](https://www.ups.com/assets/resources/media/en_CA/LmtServArea_JP%20IDG%20Export_Accpt%20Postal%20Codes.xlsx) website. For London, the data was obtained from the [Doogal](https://www.doogal.co.uk/AdministrativeAreas.php?district=E09000001) website, Toronto information was retrieved from [Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), and the Los Angeles information was obtained from a [travel](http://www.heartandcoeur.com/heart_travel/area/california_323.php) website.

## *Data Cleaning and Feature Selection*

A screenshot of a cell phone

Description automatically generatedThe data on Japan downloaded from the UPS site had to be cleaned by getting rid of unwanted information such as an introduction to the data on the csv file. This was crucial to do so because otherwise unwanted information might appear in the Pandas DataFrame when calling on the *pd.read\_csv()* function. However, there were no missing values in the data UPS provided. Some columns were removed while reading the data set as they were insignificant to this program. Finally, two more columns were added to include information on the latitude and longitude of each location.

Figure 1 - Tokyo DataFrame

A screenshot of a cell phone

Description automatically generatedThe information on London from the Doogal website needed no cleaning as all the information was there with no introductions and etc. However, some columns were removed while reading the data as they were unnecessary to the program.

Figure 2 - London DataFrame

A screenshot of a cell phone

Description automatically generatedThe information obtained about Toronto from Wikipedia had a lot of missing information. For instance, if the borough name was not assigned, then that entire row was dropped. In addition to this, if the neighborhood column has a value of not assigned, then it was replaced by its Borough name. Also, the last extra character ‘\n’ was removed from each neighborhood name. Furthermore, the same boroughs with different neighborhoods are combined and put into one row. Two extra columns were later included to indicate the longitude and latitude of each borough and the neighborhood columns were dropped as it was redundant relative to the borough names.

Figure 3 - Toronto DataFrame

A screenshot of a cell phone

Description automatically generatedThe Los Angeles dataset was cleaned up to only include the city name while reading the table. The latitude and longitude columns were later included.

Figure 4 - LA DataFrame

# **Methodology**

A screenshot of a cell phone

Description automatically generatedAfter collecting the data necessary for each city that is being investigated, the Foursquare API to get the venue data for each area of each city. The program collects information on the top 100 venues in a radius of 500m. By making use of API calls, Foursquare will return the venue data in a JSON file with information on Venue Name, Venue Latitude, Venue Longitude, and Venue Category. All this information is compiled into a DataFrame. An example is shown below of Tokyo’s venues:

Figure - Tokyo's Venues in each Area

A screenshot of a cell phone

Description automatically generatedAfter getting tables like the one in figure 5 for each city, merge all of them to form one huge giant table which is grouped by Area name and illustrates the mean frequency of each venue category at each location in each city, as shown below:

Figure - Final Combined DataFrame

Using the data from figure 6, we now cluster the data by using the K-means clustering method. The K-means clustering algorithm finds all the k-number of centroids, and then assigns every data point to the nearest cluster. The feature variables are all the venue categories and by using these columns, each Area Name is clustered with one of k-number of clusters.

A close up of a map

Description automatically generatedTo choose the optimum number of clusters, the elbow method is used. The elbow method involves graphing out the Sum of Squared Distances (SSD) of each cluster point to its respective centroids, against each k-value. The optimum k-value is later determined to be where the SSD doesn’t change as much on a drastic level.

Figure - Elbow Method

# **Results**

The results from this type of unsupervised machine learning algorithm shows there are nine cluster groups. Each different cluster is given a different color with cluster 0 being red, cluster 1 being purple, cluster 2 being blue, cluster 3 being light blue, cluster 4 being turquoise, cluster 5 being teal, cluster 6 being green, cluster 7 being light orange, and cluster 8 being dark orange. The different clusters is a visualization of how each area in each city are similar or dissimilar to each other as shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| Tokyo | London | Toronto | Los Angeles |
| A picture containing text, map  Description automatically generated | A close up of a map  Description automatically generated | A picture containing text, map  Description automatically generated | A picture containing text, map  Description automatically generated |

Figure 8 - Visual Clustering of Data

# **Conclusions**

From the results shown above, majority of the areas in London and Los Angeles are in cluster 0, and some of the areas in Toronto are in cluster 0 and this indicates that these locations are similar in terms of the types of venues they share. Only one location in Los Angeles is part of the same cluster, cluster 1, as majority of the locations in Tokyo. Hence, if a company, with its main branch being in London, wants to open a different branch in a different part of the world, it is more statistically appropriate for the branch to be opened in Toronto and Los Angeles as potentially the company may be targeting the same type of audience, thus maximizing profit.

# **Limitations and Future Directions**

In this project, only the venue categories were taken into account to assess the similarities between areas of different cities. However, this may not be the only factor that needs to be taken into account when considering the similarity between locations. Future research for this project may have to include a methodology to investigate similarity or dissimilarity between the cultural background, history, venue categories of each particular location and see if there is an improvement in predicting the optimum location of opening a new branch for a multinational corporation.

# **References**

Shoo, D. (2017, September 26). *Economic Effects of Multinational Corporations*. Retrieved from bizfluent: https://bizfluent.com/info-8444236-economic-effects-multinational-corporations.html