Capstone Project
The Battle of the Neighborhoods

Hector A. Barriga-Acosta
(Applied Data Science Capstone by IBM/Coursera)

1. Introduction

1.1 Background

New York City, Toronto and Paris belong to the top ten multicultural cities in the world. From the "center" of each city we will consider 50 coffee shops within a radius of 1km. We will analyze similarities among clusters of coffee shops in each area.

For example, is it true that in all clusters of coffee shops there are an executive offices, or supermarkets, or restaurants, or parks nearby?

1.2 Problem

In this project we will compare the (dis)similarities that multicultural cities have from the perspective of clusters of coffee shops.

1.3 Discussion

This problem fits to the audience that wish to open a coffee shop branch in a potentially multicultural city. This analysis would suggest that the audience will success by opening a new coffee shop regardless there is already competition around.

Let's say I want to open a new coffee shop in Mexico City which is known to have a wide variety of culture (you can thing of any other city with this feature, not necessarily a big city). I might fear I won't have success because there are already coffee shops in the location that I like, I'm afraid of the competition. Now let's assume that clusters of coffee shops in multicultural cities share the fact that there are parks and museums close by. The following report will tell me whether is reliable to take the risk by opening a new branch only by looking at the type of venues close to the location.

2. Data

2.1 Description of the Data

The data that we are going to work with are 3 data frames (one for each city) that include information about each coffee shop such as name, category, address, latitude, longitude, distance, postal code, city, state, country, neighborhood and id. The information is obtained from the Foursquare API.

2.2 Obtaining the Data

For each city we set a point of reference. This could be a popular venue in the middle of the city such as a cathedral, a super center or a museum. From this point we take a sample of 50 coffee shops within a radius of 1 kilometer (Figure 1).

2.3 Data Cleaning

After obtaining robust data I cleaned it up to extract only essential information. Our concern is only about the name and location of coffee shops as well as name and location of venues around. In other words, after cleaning the data, there were 50 samples per city with 3 features: name, latitude and longitude (Figure 1). For venues around these coffee shops we obtained the features: name, latitude, longitude and category.

	name	lat	lng
0	Coffee Crêpes	48.858841	2.340802
1	Coffee Parisien	48.852607	2.334536
2	Starbucks Coffee - kiosque	48.875763	2.358857
3	Super Wild Coffee	48.875763	2.358857
4	Thalone's Coffee	48.859264	2.348910

Figure 1: Essential data for coffee shops (Paris)

3. Exploratory Data Analysis

3.1 Target Visualization

Since out main purpose is to find similarities among coffee shops it is important to have a picture in mind. I present the maps with the 50 coffee shops per city. The red point represents the center city whereas blue points are the location of coffee shops (Figure 2, 3, 4).

From each center city, 50 coffee shops were captured within a radius of 2 km. It is important to note that the density of coffee places is higher around the center city. This might be a relevant factor if stake holders wish to open a new branch coffee shop.

Also, it is noticeable that Manhattan has a special geographical shape whereas Paris and Toronto have similar density among coffee shops nearby center city.

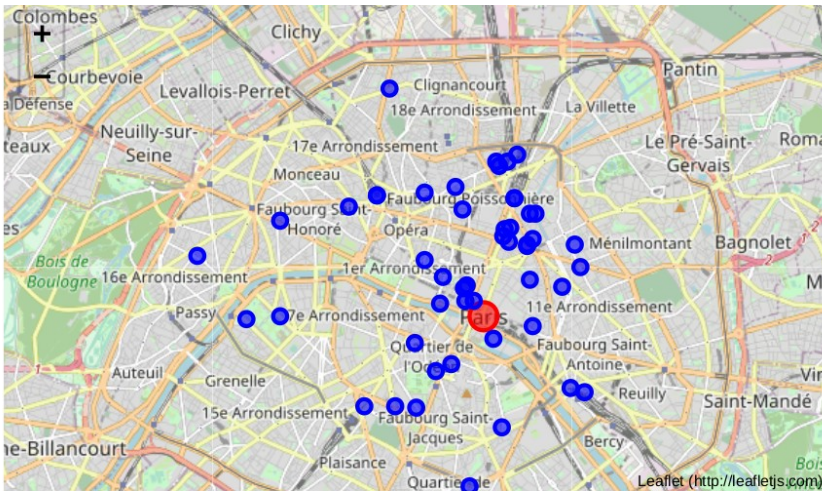


Figure 2: View of Paris

Figure 3: View of Manhattan

Figure 4: View of Toronto

3.2 Exploring More Venues

After target's visualization I proceed to explore venues around each coffee shop. The additional features captured were: venue, latitude, longitude and category (Figure 5). Once these venues are grouped by category and normalized, it gets easier to see what are the most common venues (Figure 6). This analysis will be important too to perform clusters of coffee shops with similarities.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Coffee Crêpes	48.858841	2.340802	Place du Louvre	48.859841	2.340822	Plaza
1	Coffee Crêpes	48.858841	2.340802	Église Saint-Germain-l'Auxerrois (Église Saint...	48.859520	2.341306	Church
2	Coffee Crêpes	48.858841	2.340802	Coffee Crêpes	48.858841	2.340802	Coffee Shop
3	Coffee Crêpes	48.858841	2.340802	Cour Carrée du Louvre	48.860360	2.338543	Pedestrian Plaza
4	Coffee Crêpes	48.858841	2.340802	Pont des Arts	48.858565	2.337635	Bridge

Figure 5: Venues around coffee shops (Paris)

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Araku Coffee	70	70	70	70	70	70
Arômes Coffee Shop	100	100	100	100	100	100
Blackburn Coffee	45	45	45	45	45	45
Bombay Coffee	74	74	74	74	74	74
Caoua Coffee Stop	63	63	63	63	63	63

Figure 6: Frequency of venues

4. Classification Model

4.1 Filtering and transforming the data

The application of classification models is appropriate since we search (dis)similarities between the numerical variables of frequency among venues around coffee shops. I used the k-means classification model and found the best fit value for k for each city.

This analysis was made by grouping the most common venues around coffee shops and considering only the 10 most popular venues 500 meters around (Figure 7).

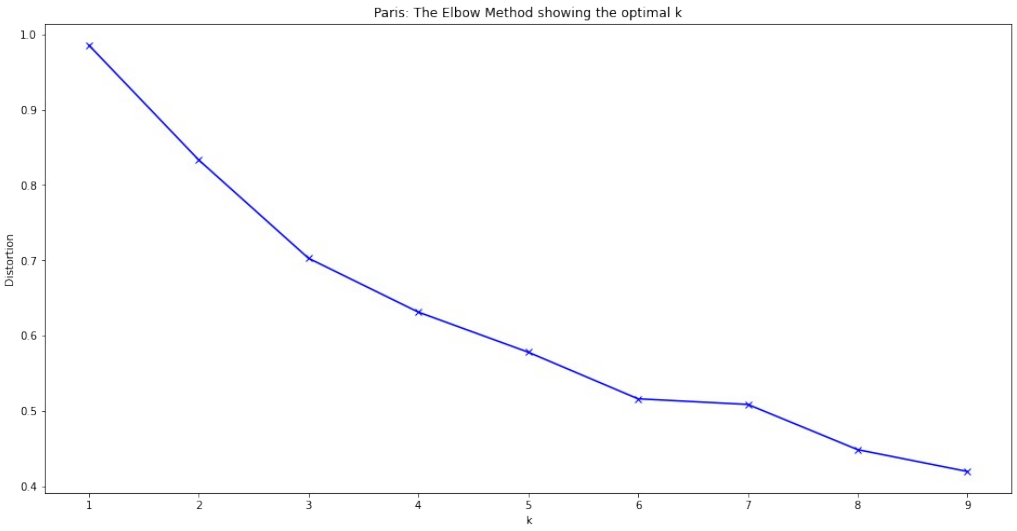
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Mos Commo Venu
0	Araku Coffee	Hotel	Art Gallery	French Restaurant	Bistro	Wine Bar	Vietnamese Restaurant	Japanes Restaurar
1	Arômes Coffee Shop	French Restaurant	Bar	Bistro	Restaurant	Bakery	Japanese Restaurant	Coffe Sho
2	Blackburn Coffee	Cocktail Bar	French Restaurant	Bakery	Breakfast Spot	Hotel	Pizza Place	Seafoo Restaurar
3	Bombay Coffee	French Restaurant	Bakery	Café	Italian Restaurant	Hotel	Science Museum	Morocca Restaurar
4	Caoua Coffee Stop	French Restaurant	Coffee Shop	Bar	Asian Restaurant	Bakery	Garden	Wine Ba

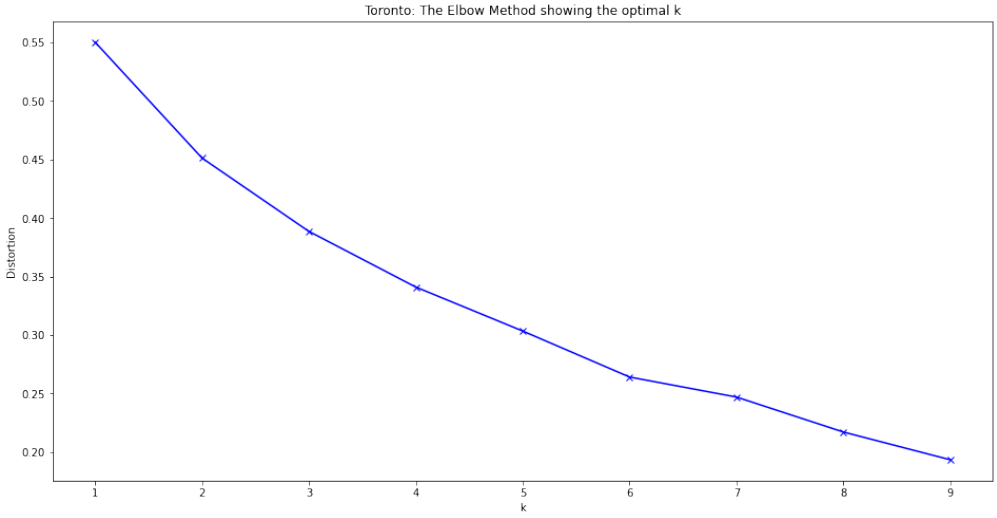
Figure 7: Most common venues around coffee shops

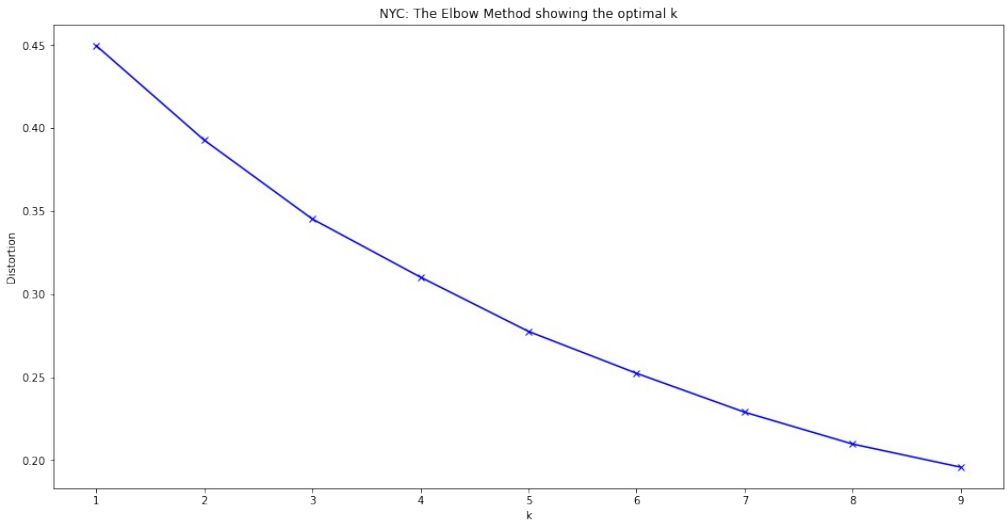
4.2 Optimizing the Model

With help of the Elbow Method I obtained that the best fit for k are as follows.

- Paris, k=8
- Manhattan, k = 5
- Toronto, k = 7


Paris: The Elbow Method showing the optimal k
Distortion
k


Toronto: The Elbow Method showing the optimal k
Distortion
k


NYC: The Elbow Method showing the optimal k
Distortion
k

4.3 Clustering Visualization

Once found the best fit for the classification model I was able to visualize the clusters of coffee shops. Each cluster will have coffee shops with similar venues in a neighborhood of 500 meters.

One might think that two coffee shops that are physically distant will be distinct, but this analysis shows that this is not the case.

4.3.1 Clusters in Paris

The best fit for Paris is k = 8. However there is a huge cluster formed by the blue points. Coffee shops in this cluster have the following venues around: Hotels, Plazas, Bars and Restaurants (Figure 8). Other clusters have a similar amount of elements such as green and orange, and red and purple cluster.

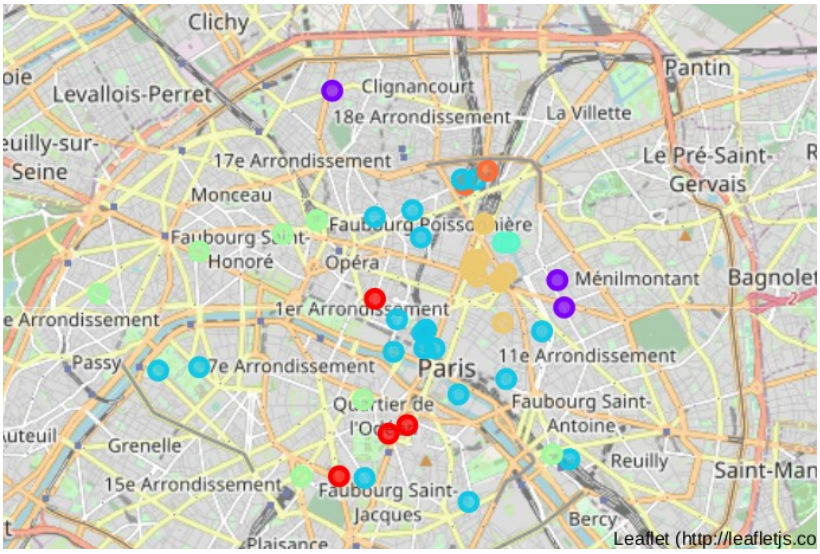


Figure 8: Clusters in Paris

4.3.2 Clusters in Manhattan

The best fit for Manhattan is k = 5. In this case we have two main clusters: blue and green clusters (Figure 9). Observe that the density of the green cluster is extremely high whereas the blue cluster is dispersed. Red and purple clusters are alike.

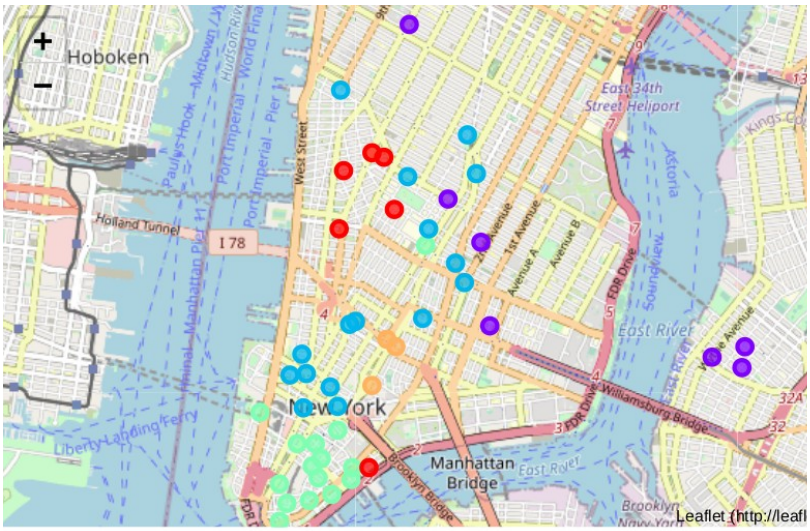


Figure 9: Clusters in Manhattan

4.3.1 Clusters in Toronto

The best fit for Toronto is k = 7. Like Paris, Toronto has a huge cluster formed by the red points. Coffee shops in this cluster have the following venues around: Hotels, Restaurants, Gyms and Bars (Figure 10). Other clusters have a similar amount of elements such as blue, light blue and yellow, and orange and purple clusters.

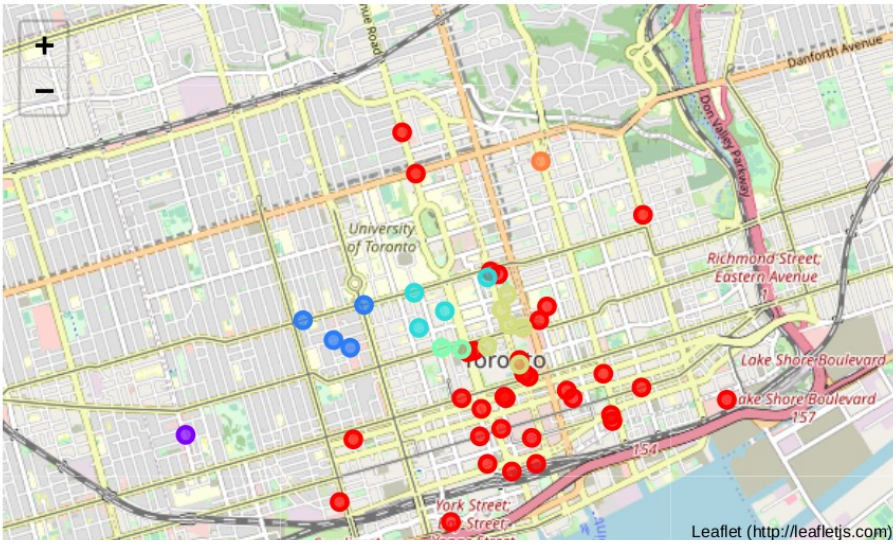


Figure 10: Clusters in Toronto

5. Correlation

5.1 Data for Correlation

Finally, with the purpose of finding a correlation I merged the data set including the 10 most common venues for each of the 50 coffee places in the three cities.

The data set was filtered after merging by considering only venues with more than 60 counts in total (3 cities). I believe that venues with less than 60 counts are not so relevant since it is unlikely that they are correlated to coffee shops. A description of basic statistics are shown below (Figure 11).

	Salad Place	Wine Bar	Vegetarian / Vegan Restaurant	Bookstore	Burger Joint	Park	Cocktail Bar	Sandwicl Plac
count	64.000000	66.000000	68.000000	69.000000	72.000000	75.000000	80.000000	80.00000
mean	0.014621	0.019269	0.016316	0.015117	0.016767	0.016419	0.021451	0.02529
std	0.006156	0.011550	0.010846	0.007160	0.008972	0.008765	0.012861	0.01973
min	0.005000	0.003333	0.003333	0.005000	0.002500	0.004149	0.006002	0.00500
25%	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.012487	0.01324
50%	0.012274	0.014939	0.011696	0.011765	0.014036	0.014406	0.020000	0.02000
75%	0.020000	0.028750	0.020000	0.020000	0.020102	0.020000	0.025539	0.03000
max	0.030000	0.050000	0.053333	0.040816	0.040000	0.060000	0.066667	0.14285

Figure 11: Statistics after merging

5.2 Heat Map

I created a heat map. It's very easy now to check what venue is most correlated to the presence of coffee shops (Figure 12).

It was surprising to see that there is no correlation at all between venues and coffee places perhaps only with the exception of Sandwich Places which have 0.52. This is not a great correlation but stakeholders may want to consider locations with sandwich places around within a radius of 500 meters.

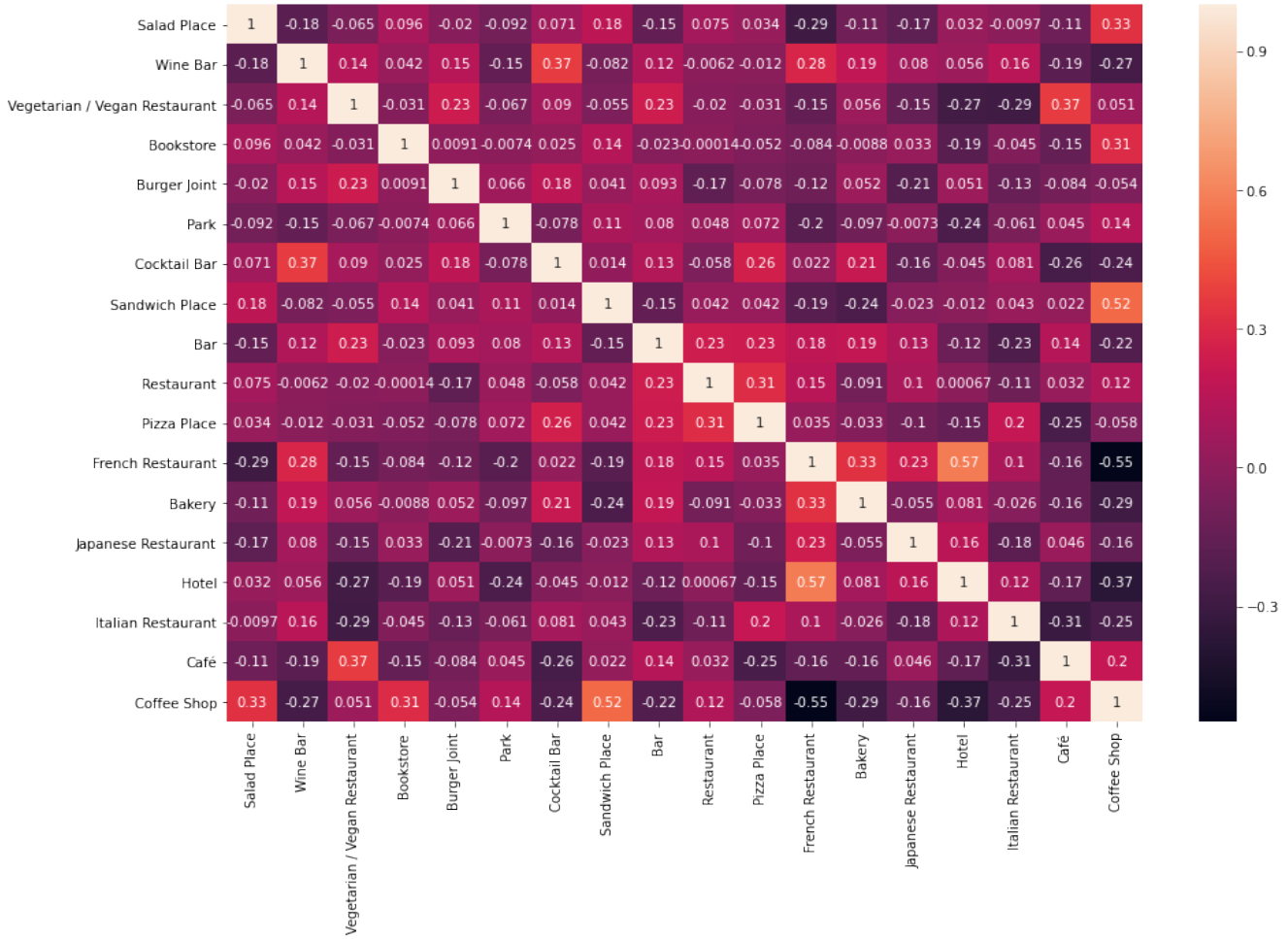


Figure 12: Heat map for correlation

6. Conclusion

In this study, I analyzed the (dis)similarities between coffee shops in the cities of Paris, NYC and Toronto which are know as multicultural cities. Further, it was studied whether coffee places are correlated to any other venue. I identified that Paris and Toronto are more alike than the area of Manhattan, and there only possible correlation to coffee places is sandwich places.

I built a classification model for each city with the best fit. These models can be very useful in helping visualize such (dis)similarities among coffee shops. It was found that physical distance is not the only reason why two coffee shops can be similar, and moreover, not because two coffee places are near implies they are similar.

Stakeholders interested in opening a new branch coffee place should focus on locations with no more than 3 competition establishments and near sandwich places, hotels, gyms, restaurants and bars.

7. Scope of the Analysis

A similar analysis can be use to study other potentially multicultural cities and explore the market such as opening a new restaurant, a new gym, etc.

Moreover, it can be useful to categorize the biggest cities around the world and find insights of why they're such big cities. Since cities are constantly growing, stakeholders might wish to mimic the market applied to smaller cities by placing new coffee shops in similar locations (given by this report) where successful coffee shop are located in those big cities.