Comparing Cities Worldwide by Popular Venues

Nick Freedman

August 2021

1. Introduction

In today's world, most major cities offer very similar amenities and venues. While each city is unique, one could travel to nearly any destination and find, for example, any world cuisine they desired. However, the tastes and habits of any local population (and indeed tourists) would be expected to differ between cities. Therefore, the question to answer is: How similar are major cities based on their most popular destinations? An answer to this particular question could be useful in many ways. For one, someone who wishes to travel may find this useful in deciding their next destination if they enjoy the types of venues which are popular there. Someone attempting to open or expand a business internationally may find the data useful in determining which cities would be best suited to a new location. There are numerous other opportunities for this data to become useful.

2. Data

I used 3 data sources for this analysis. First, this Wikipedia article contains a list of the top 81 world cities by population. I extracted this information (city name, country name and area) to determine which cities to compare. One restriction will be that out of the list, I only analyzed one city per country so as to limit bias in the analysis.

Second, using the acquired city names, I used the geopy package in Python to obtain each city's coordinates.

Third, I extracted the most popular venues of each city using Foursquare. The radius from which to pull the data for each city is proportional to the area of the city. Additionally, I used the categories endpoint to get the Foursquare venue categories tree.

3. Methodology

First, I scraped the data from the aforementioned wikipedia table. I took the City name, Country name, and Area of the city proper. Again. I then filtered out duplicates for the countries, allowing only 1 city per country. This resulted in 36 cities.

Next, I used geopy to obtain the coordinates for each city and attached that to the cities data frame. The "Area" column was a string, so this needed to be converted to an integer. After doing that, I filled the cities which did not have areas with the average area of the other cities.

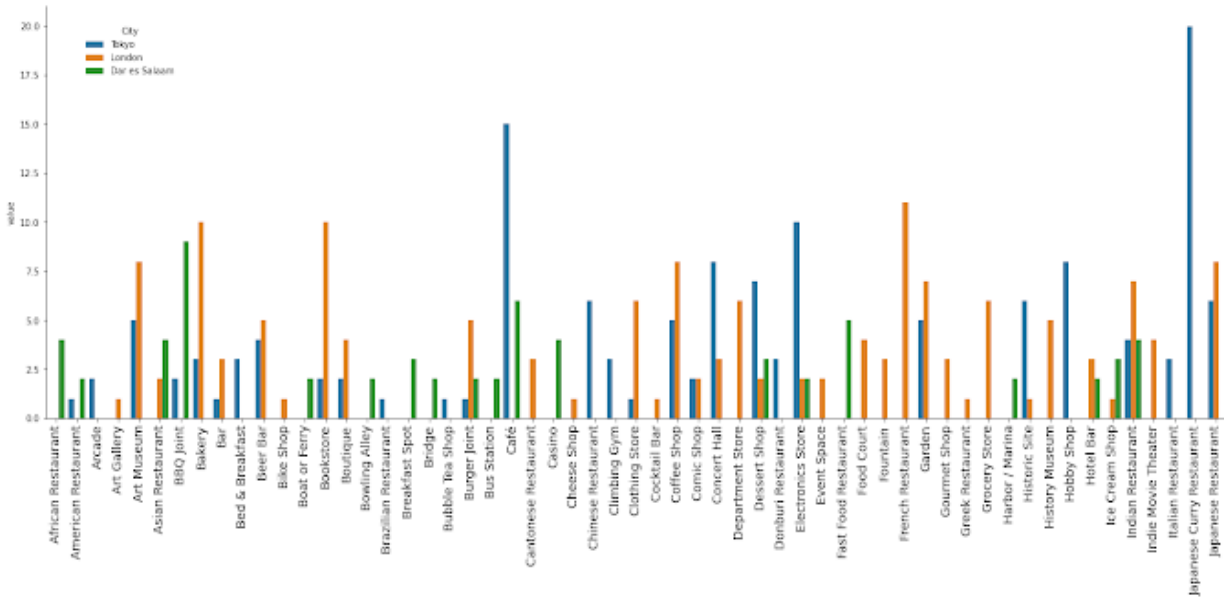
Finally, I calculated the radius based on the area and attached that to the dataframe as well.

The next step was to obtain the Foursquare category hierarchy. This contains the entire structure of hierarchical categories. For example, "Food" is a parent of "Asian Food" which is a parent of "Chinese Food." Using the categories endpoint, I was able to obtain this information. The reason for this was that I determined that doing analysis on two sets of data would be prudent. One would be the "generalized" dataset while the other would be the "non-generalized" dataset. For the generalized set, with the exception of venues below the top level categories (i.e "Food" or "Art and Entertainment"), I would promote each venue category obtained to a higher level in the hierarchy. For example, when the category "Japanese Curry Restaurant" appeared, it would be promoted to "Japanese Restaurant." The non-generalized set would retain the categories as it was retrieved from Foursquare.

I obtained 3 pages worth (roughly 150 results) of venues per city. I removed the categories "Hostel" and "Hotel" because these did not seem relevant to the goal of this analysis. This resulted in 7870 samples. At this point, I created two datasets: the generalized and non-generalized set. Then I used one hot encoding for each set and summed them to get the total number of each venue category for each city.

In the next section, I performed some exploratory data analysis. I first explored the non-generalized dataset. I used bar charts to compare venues in three cities on different continents, Tokyo, London and Dar es Salaam. A sample is shown in Figure 1.

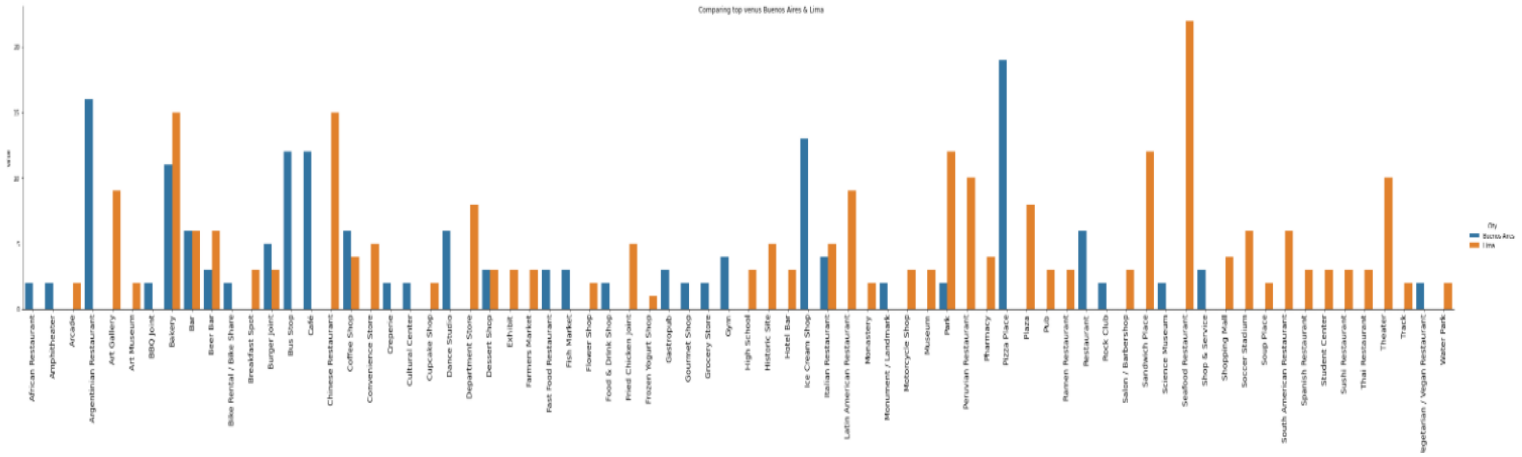
Figure 1



Based on this comparison, it can be seen that there is a wide distribution of venue types in the data. There is certainly some overlap, for example, dessert shops and electronics stores are similar for all three cities, as are Indian restaurants. Certain categories do dominate, however, such as “Japanese Curry Restaurant.”

I also compared two cities from the same continent. Lima, Peru and Buenos Aires, Argentina. I limited cities to one per country, however my expectation was that cities from the same continent or region might be similar in regards to their popular venues.

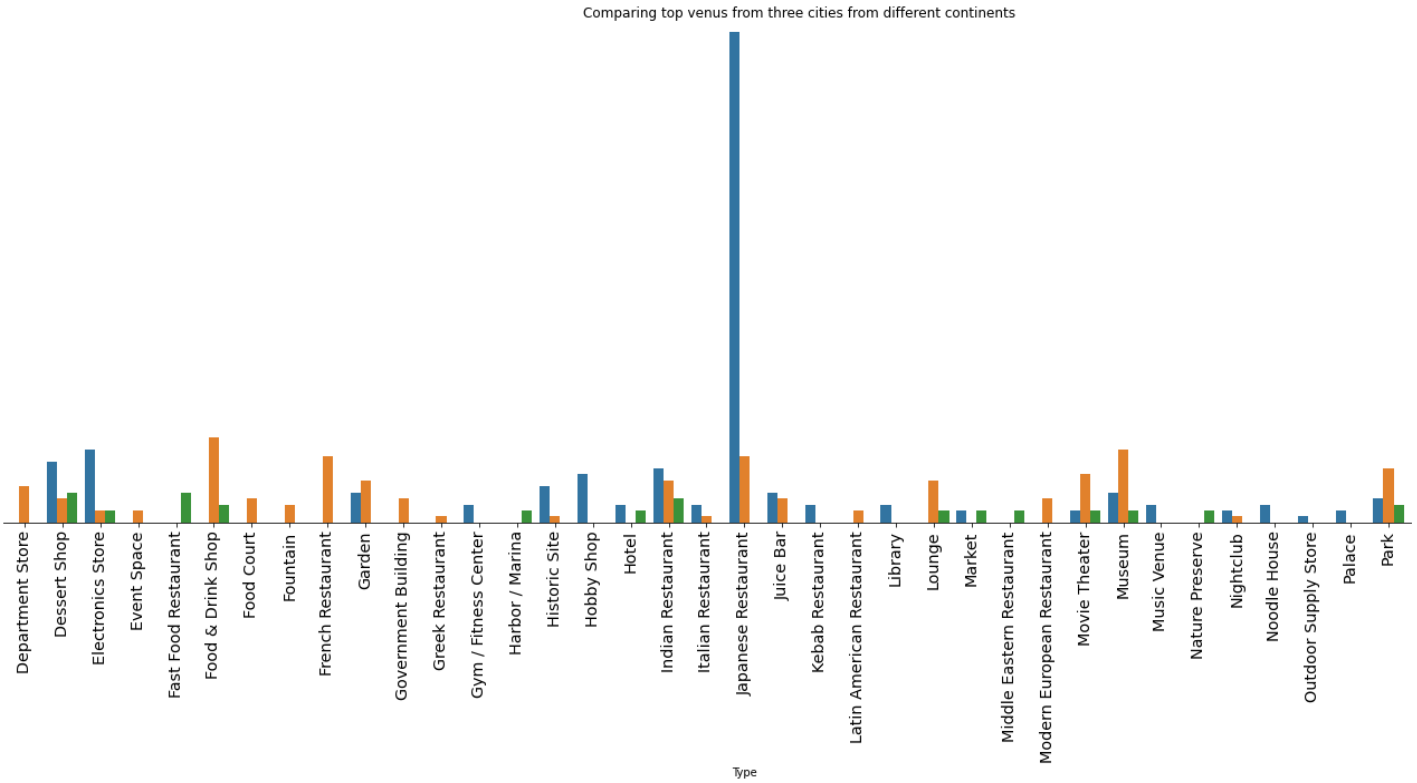
Figure 2



Lima (Orange) and Buenos Aires (Blue) have significant differences between them. While there is some overlap, it is clear that many categories are only occupied by one city.

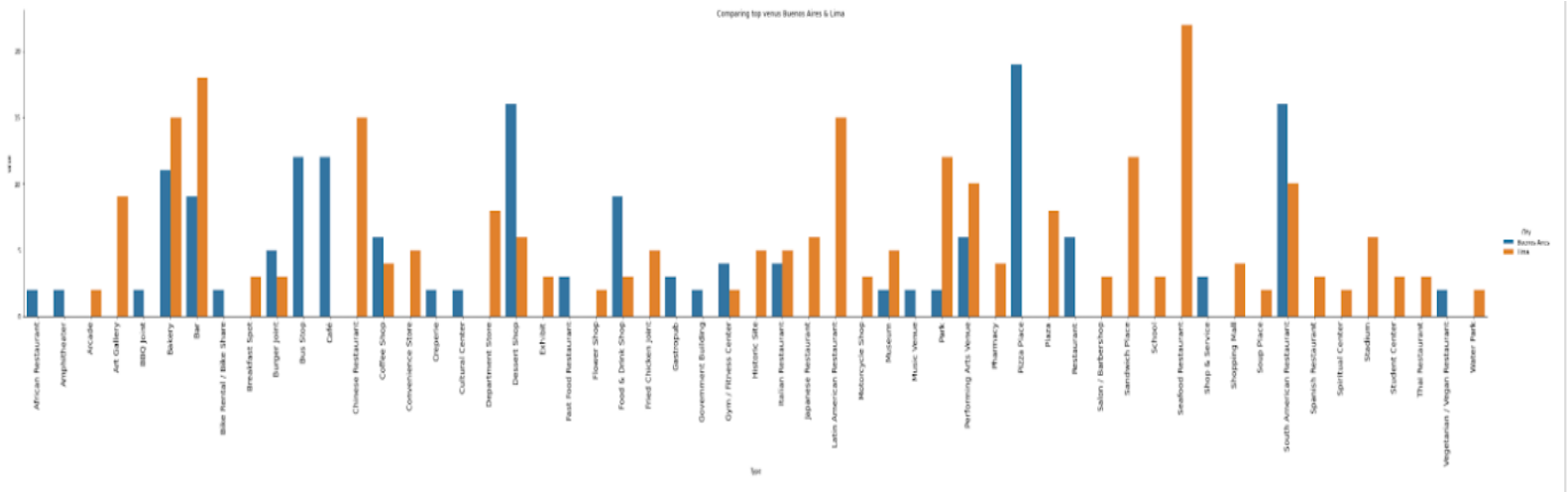
Next, I analyzed the generalized set, using the same example cities as before. A sample of the first analysis (cities from 3 different continents) is shown in figure 3.

Figure 3



For the most part, the distribution of the venues seems to be more even with the generalized set. However, for Tokyo (Blue) “Japanese Restaurant” greatly dominates the data. I also conducted the same analysis for Lima and Buenos Aires, shown in figure 4.

Figure 4



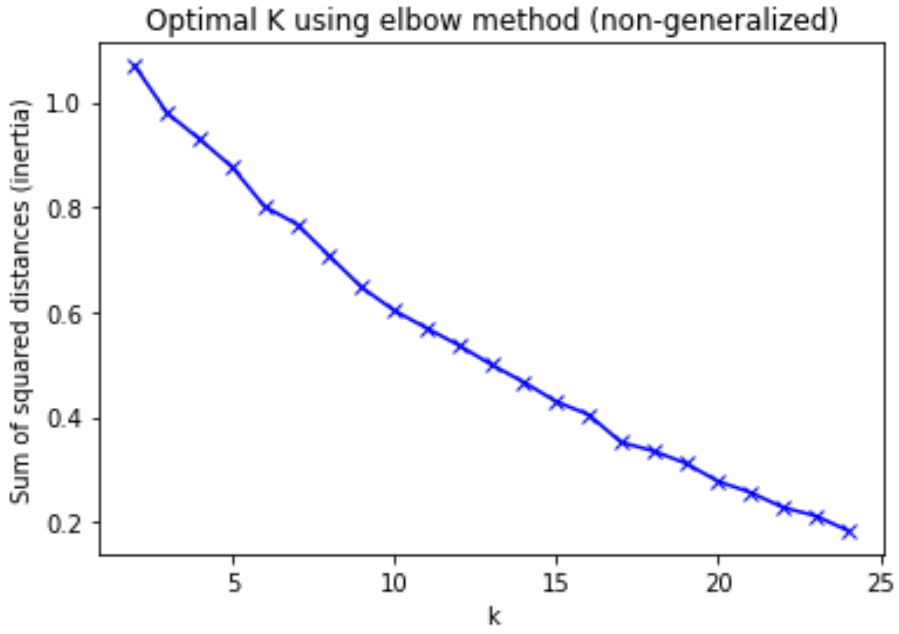
As can be seen in the above figure, certain categories such as "South American Restaurant" became more prominent after being generalized and thus the cities can be seen as more similar. However, other categories remained prominent and unique, such as "Seafood Restaurant" for Lima.

Overall, the analysis resulted in similar findings to the non-generalized set with several categories merged into single categories.

The next step was to cluster the data. For clustering, I used the KMeans algorithm as this is unlabeled data. First, I normalized the data by converting each venue category to its percentage of the total sum of categories for each city. For example, in Cairo "African Restaurant" makes up about 3.2% of the total venues. I did this for both sets.

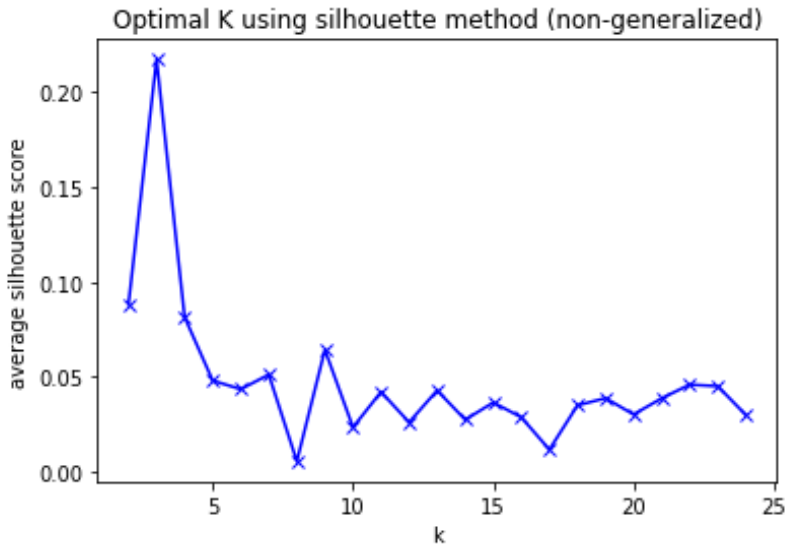
Next, I ran a K-Means algorithm from 2 to 25 clusters to determine the optimal K. Using the inertia attribute, I plotted the Sum of Squared distances for an elbow test. Figure 5 shows the results of the test.

Figure 5



Because the graph did not indicate an optimal K (an elbow is present), I decided to run a silhouette score analysis to try to determine the optimal K. I obtained the average silhouette scores for the data when clustered with K groups from 2 to 25, as with the elbow analysis. Figure 6 shows the average silhouette score for each K.

Figure 6



This analysis seems to indicate the optimal K is 3. I ran these algorithms for the generalized set as well (Figure 7 & 8).

Figure 7

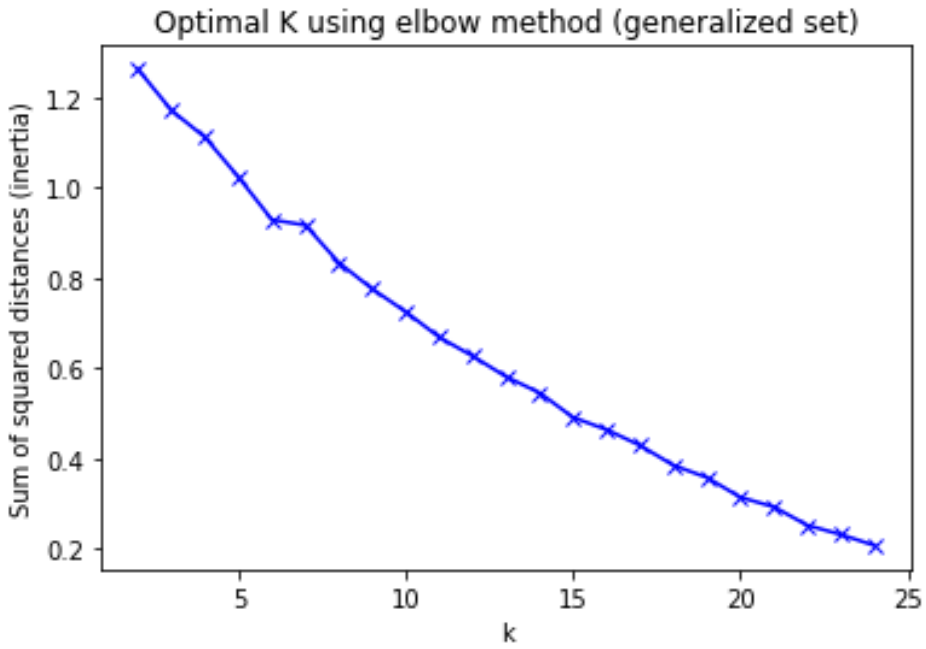
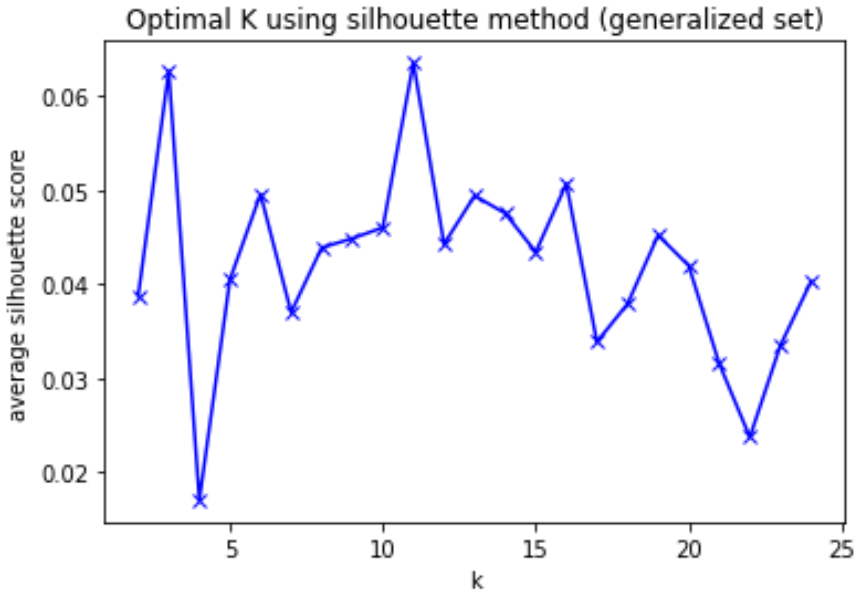


Figure 8

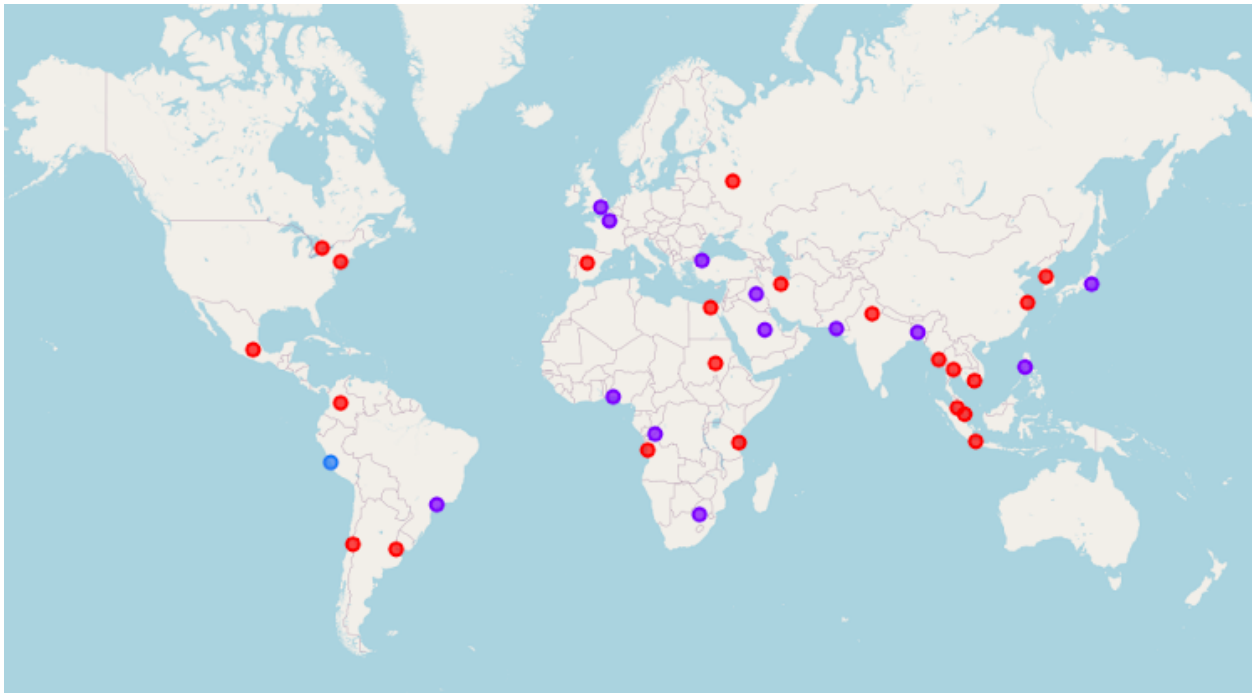


Once again, there is no obvious elbow shown in the elbow test (sum of squared distances). The silhouette test seems to show that K = 11 clusters is the optimal number of clusters for this set.

4. Results

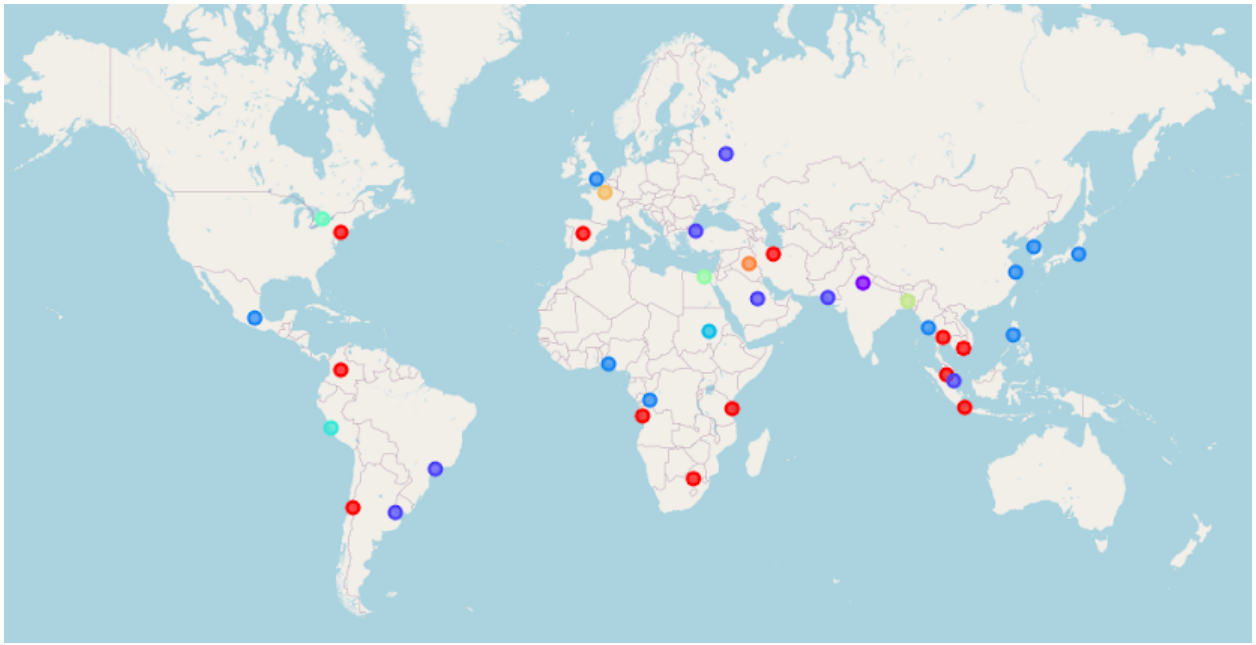
First, I mapped the cities with their clusters color-coded based on the results of the non-generalized K-Means analysis as shown in Figure 9.

Figure 9



Next, I did the same for the generalized set (K = 11 clusters) in figure 10.

Figure 10



I then created a final data frame which contained each city's cluster and top venues for both the generalized and non-generalized sets as shown in figure 11.

	City	Country	top1_NG	top2_NG	top3_NG	Cluster_NG	top1_G	top2_G	top3_G	Cluster_G
0	Tokyo	Japan	Japanese Curry Restaurant	Café	Ramen Restaurant	1	Japanese Restaurant	Café	Bar	3
1	Delhi	India	Indian Restaurant	Café	Snack Place	0	Indian Restaurant	Café	Snack Place	1
2	Seoul	South Korea	Korean Restaurant	Coffee Shop	Chinese Restaurant	0	Korean Restaurant	Coffee Shop	Japanese Restaurant	3
3	Shanghai	China	Coffee Shop	Shopping Mall	Cocktail Bar	0	Bar	Coffee Shop	Shopping Mall	3
4	São Paulo	Brazil	Bakery	Bar	Coffee Shop	1	Food & Drink Shop	Bar	Bakery	2

5. Discussion

The results of this experiment were not ideal. Based on the elbow method, an optimal number of clusters was not readily apparent. This was the case for both sets. The silhouette method gave somewhat more clear results, although not perfect. A silhouette score of 1 would indicate that the clusters are far apart and very clear. The average score for the non-generalized set given K = 3 was around 0.24, which is closer to 0, meaning there is less significance to the clusters. The generalized set actually had an even lower silhouette score.

Given the results of this analysis however, we can see some interesting trends. In the non-generalized set, the algorithm clustered Lima, Peru on its own. The top 3 venues there are "Seafood Restaurant," "Bakery," and "Peruvian Restaurant." "Seafood Restaurant" made up 8% of Lima's popular venues, which is far higher than any other city for that category. Similarly, in the generalized set, Paris was clustered on its own. "French Restaurant" made up 17% of Paris' popular venues.

Tokyo, Seoul, Shanghai, and Manilla were all grouped together in the generalized set and share some venues in their top 3. These include "Cafe", "Coffee Shop", "Bar" and most notably," Japanese Restaurant." It is interesting that they are both regionally close and ostensibly close in their popular venues.

Using K-Means seemed the best approach as it is fairly general purpose. I limited the cities to one per country so that the results wouldn't be too skewed. It may be insightful to limit the analysis to two cities per country in order to obtain more data. Alternatively, getting a longer list of cities with more entries could be useful.

6. Conclusion

Every city is unique in its own way and the most popular venues differ greatly between them. That said, there are certainly similarities that can be drawn between cities which may be useful for tourism, business, and commerce.

With the cities clearly visualized on the map, we can see which are potentially similar. This info, in combination with the most popular venues could give business owners an idea about where to expand or what type of business to start.