# Report: Comparison of G20 Largest Cities

## Introduction/Business Problem

A business consultant is working on an urban planning project to try to develop a city and help it in becoming better. The consultant decides that the standard that they want to achieve would be something similar to the largest cities of the top world economies. The top 20 economies of the world are those of the G20 which consists of 19 countries plus the European Union.

The purpose of this study is to compare the largest cities of these 19 countries and see how similar or different they are to each other. The consultant hopes to be able to cluster these cities into four distinct clusters so they can later on decide which city type they would like to imitate through the development of their cities.

To solve this problem, the consultant turns to a data scientist to pull out the data, analyze it and get the results.

## Data

The data scientist will first rely on web scraping to get the information they need regarding the countries of the G20 and their largest cities.

Once the names of the cities are found, the data scientist will use GeoPy's geocoders to get the latitudes and longitudes of the cities from the OpenStreetMap data.

The data scientist will then pull from Foursquare, using the Foursquare Places API, the venues within a 1000m radius from the city latitudes and longitudes.

This data will then be analyzed to cluster the cities into 4 groups.

## Methodology

The first step is to get data on the G20 countries from the web. Wikipedia has a page for the G20 in which it lists the G20 countries in a table. The list of countries is extracted from the Wikipedia table using BeautifulSoup and stored in a data frame. The G20 list on Wikipedia lists the European Union as a distinct country, so it will be removed from the data frame.

The analysis is based on cities not countries, so the next step would be to get the names of the largest cities of G20 countries. The largest cities for each of the G20 countries can be obtained through web scraping the individual country pages on Wikipedia. Fortunately, Wikipedia has a standard format for its pages' URL that uses the country name in the URL. The standard format is "https://en.wikipedia.org/wiki/*[country_name]*" where [country_name] is the name of the country. In case the country's name is made of two or more words, the words are separated with an underscore. Each country's Wikipedia page is then scraped for the name of the largest city. For ease of future reference, the URLs and the largest cities are stored in the same data frame as the countries. Using GeoPy's geocoder Nominatim, the latitude and longitude for each city can be determined. The coordinates are then stored in the same data frame.

Now that the data frame is ready with the names and coordinates of the cities, it is possible to start the analysis. Using Foursquare's API and the cities' coordinates, the aim is to search for venues within a 1000m-radius of the coordinates. It would be better to expand the radius as much as possible, but the free version of the API is limited in the number of results it would produce, so extending it too much won't make a lot of difference. Since this will be a repetitive task for each city, a function is defined to perform this action and return the results as a new data frame.

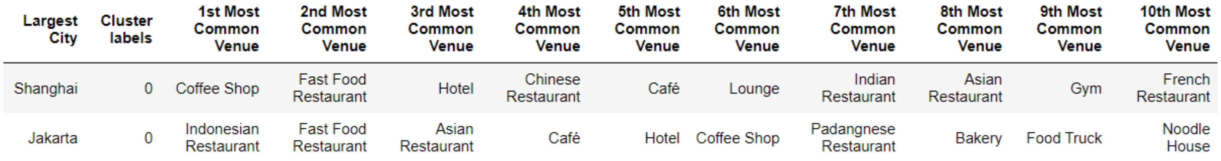
To get a better understanding of the data, one-hot encoding is used to create a separate column for each venue category and link it to the city it pertains to. After that a summary is made by grouping the data by city. Since the numbers vary wildly from one city to the other, the mean function is used during grouping to make the city data more comparable to each other.

To have a more meaningful set of results for the project and to avoid clutter, the top ten venues for each city are found and returned in a separate data frame. Based on the new data frame, the cities are clustered into four distinct clusters using K-means clustering.

## Results

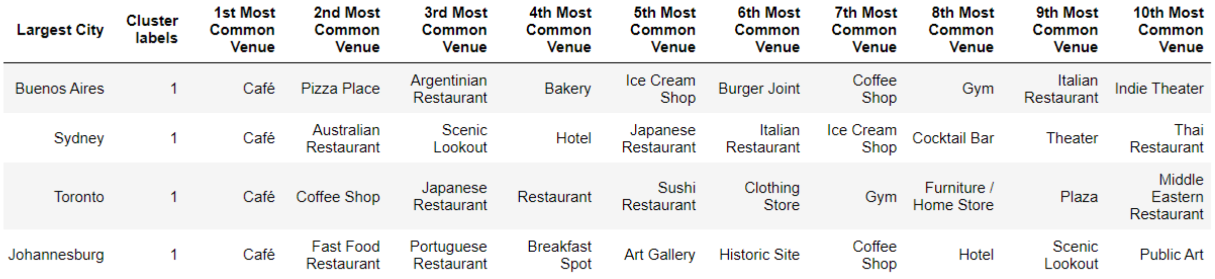
As of the writing of this report, the search obtains 1712 venues in 292 unique categories. The K-clustering method clusters the 19 cities into 4 clusters based on similarities.

The first cluster (Cluster 0) is comprised of Shanghai and Jakarta.



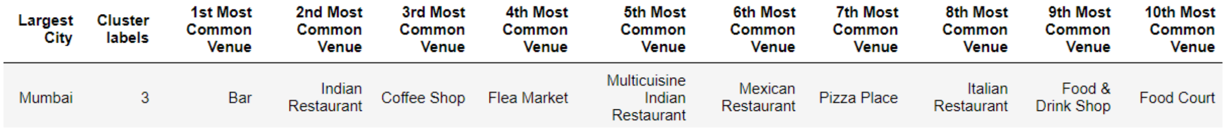
| Largest City | Cluster labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Shanghai | 0 | Coffee Shop | Fast Food Restaurant | Hotel | Chinese Restaurant | Café | Lounge | Indian Restaurant | Asian Restaurant | Gym | French Restaurant |
| Jakarta | 0 | Indonesian Restaurant | Fast Food Restaurant | Asian Restaurant | Café | Hotel | Coffee Shop | Padangnese Restaurant | Bakery | Food Truck | Noodle House |

The second cluster (Cluster 1) contains Buenos Aires, Sydney, Toronto and Johannesburg.



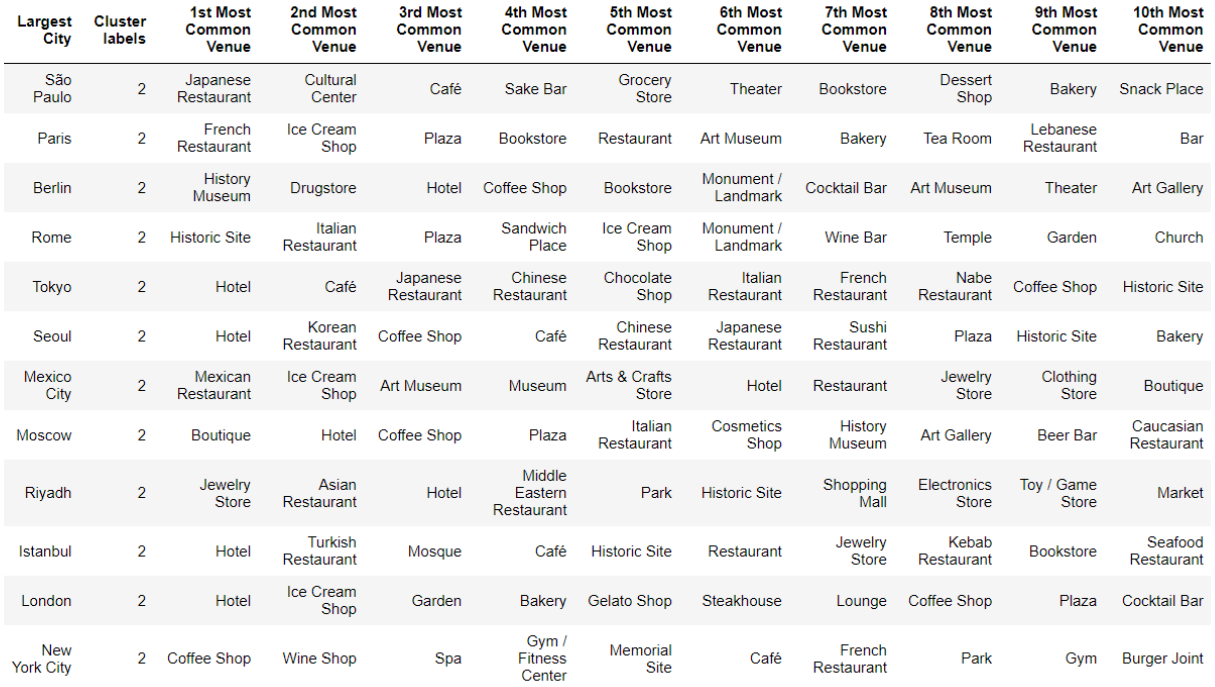
| Largest City | Cluster labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Buenos Aires | 1 | Café | Pizza Place | Argentinian Restaurant | Bakery | Ice Cream Shop | Burger Joint | Coffee Shop | Gym | Italian Restaurant | Indie Theater |
| Sydney | 1 | Café | Australian Restaurant | Scenic Lookout | Hotel | Japanese Restaurant | Italian Restaurant | Ice Cream Shop | Cocktail Bar | Theater | Thai Restaurant |
| Toronto | 1 | Café | Coffee Shop | Japanese Restaurant | Restaurant | Sushi Restaurant | Clothing Store | Gym | Furniture / Home Store | Plaza | Middle Eastern Restaurant |
| Johannesburg | 1 | Café | Fast Food Restaurant | Portuguese Restaurant | Breakfast Spot | Art Gallery | Historic Site | Coffee Shop | Hotel | Scenic Lookout | Public Art |

The third cluster (Cluster 3) has only one city, Mumbai.



| Largest City | Cluster labels | 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 Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mumbai | 3 | Bar | Indian Restaurant | Coffee Shop | Flea Market | Multicuisine Indian Restaurant | Mexican Restaurant | Pizza Place | Italian Restaurant | Food & Drink Shop | Food Court |

The fourth and final cluster (Cluster 2) has the remaining 12 cities.



| Largest City | Cluster labels | 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 Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| São Paulo | 2 | Japanese Restaurant | Cultural Center | Café | Sake Bar | Grocery Store | Theater | Bookstore | Dessert Shop | Bakery | Snack Place |
| Paris | 2 | French Restaurant | Ice Cream Shop | Plaza | Bookstore | Restaurant | Art Museum | Bakery | Tea Room | Lebanese Restaurant | Bar |
| Berlin | 2 | History Museum | Drugstore | Hotel | Coffee Shop | Bookstore | Monument / Landmark | Cocktail Bar | Art Museum | Theater | Art Gallery |
| Rome | 2 | Historic Site | Italian Restaurant | Plaza | Sandwich Place | Ice Cream Shop | Monument / Landmark | Wine Bar | Temple | Garden | Church |
| Tokyo | 2 | Hotel | Café | Japanese Restaurant | Chinese Restaurant | Chocolate Shop | Italian Restaurant | French Restaurant | Nabe Restaurant | Coffee Shop | Historic Site |
| Seoul | 2 | Hotel | Korean Restaurant | Coffee Shop | Café | Chinese Restaurant | Japanese Restaurant | Sushi Restaurant | Plaza | Historic Site | Bakery |
| Mexico City | 2 | Mexican Restaurant | Ice Cream Shop | Art Museum | Museum | Arts & Crafts Store | Hotel | Restaurant | Jewelry Store | Clothing Store | Boutique |
| Moscow | 2 | Boutique | Hotel | Coffee Shop | Plaza | Italian Restaurant | Cosmetics Shop | History Museum | Art Gallery | Beer Bar | Caucasian Restaurant |
| Riyadh | 2 | Jewelry Store | Asian Restaurant | Hotel | Middle Eastern Restaurant | Park | Historic Site | Shopping Mall | Electronics Store | Toy / Game Store | Market |
| Istanbul | 2 | Hotel | Turkish Restaurant | Mosque | Café | Historic Site | Restaurant | Jewelry Store | Kebab Restaurant | Bookstore | Seafood Restaurant |
| London | 2 | Hotel | Ice Cream Shop | Garden | Bakery | Gelato Shop | Steakhouse | Lounge | Coffee Shop | Plaza | Cocktail Bar |
| New York City | 2 | Coffee Shop | Wine Shop | Spa | Gym / Fitness Center | Memorial Site | Café | French Restaurant | Park | Gym | Burger Joint |

## Discussion

Cluster 2 has 12 cities in it and so represents 63% of the G20 countries. The biggest concentration of venues in Cluster 2 is in restaurants, coffee shop / café, hotels, and cultural venues like historical sites, museums, and galleries.

Cluster 1 is the next largest cluster with 4 cities in it. It represents 21% of the G20 countries. The most common venue in all 4 cities is café, and the biggest concentration of venues in the cluster is in restaurants, gyms, hotels, and scenic lookouts.

Cluster 0 is the third cluster with 2 cities in it representing 11% of the G20 countries. The biggest concentration of venues in Cluster 0 is in restaurants, coffee shop / café, and hotels.

The last cluster is Cluster 3 with only 1 city in it representing 5% of the G20 countries.

The results show that the largest cities of G20 countries are very similar in terms of which venues are the most common in those cities. In all 19 cities, it was clear that restaurants, coffee shops, and hotels are in the top ten most common venues.

## Conclusion

Since Cluster 2 has the biggest representation of G20 countries, following the lead of the countries in that cluster in developing a city would be the safest bet for the customer to achieve their goals.