The Next Adventure

Data Science Capstone Project

Sriharsha Jayanti

March 28, 2020

Introduction/Business Problem

With nearly 200 countries/territories, each with its own unique draw, how can we help a person choose their next adventure?

A person may choose their next destination based on a multitude of factors. A partial list includes the following:

1. Preference: A person might choose the next excursion for a variety of reasons: nature/beaches, food scene, nightlife, historical significance, cultural diversity
2. Similarities to or differences from previous travels
3. Knowledge of options: Not knowing the options or having to choose between a multitude of options
4. Price and accessibility

The scope for this problem can be very grand and complicated based on the level of the sophistication. For example, we can take all the cities across the world, and analyze each city for its venues, reviews, and costs. To simplify the problem for this project, I will take a subset of places across the world and cluster them based on the top 10 venues. This will provide a simple color-coded map, which we can use to analyze our past choices (have we been inclined to a certain cluster of locations) and the choices available to us (is there something similar, different, or new).

Data Collection and Assumptions

Data will be collected from the following sources:

1. I will use these capital cities as the representative city for each country. It is easy to find a list of national capitals on wikipedia: https://en.wikipedia.org/wiki/List_of_national_capitals.
2. The latitude and longitude data will be obtained using Geopy.
3. Foursquare API venue data (venues, categories of venues) should be available for many of the capital cities.
4. Price and accessibility: While the pricing of venues could be implemented, it does not appear to be free (premium API endpoint, Venue Details), so it will be excluded.

Assumptions

1. The chosen city is representative of the country. This is a big generalization, and each of the countries may be very different/more diverse than the chosen cities.

2. The data available for each of the cities is accurately representing that city. This may not be true depending on how exhaustive the Foursquare API data is.
3. While I will clean up some of the data and unify the categories, it is challenging to unify/clean up all the data and categories. In addition, we lose some of the uniqueness when unifying the categories.

Methodology

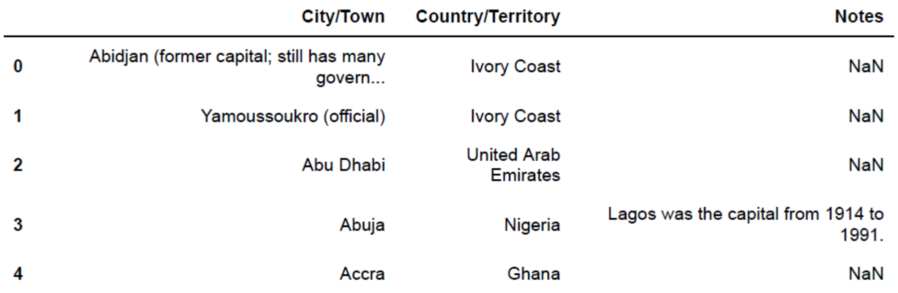
Here's a brief overview of the methodology:

1. First, a table with the cities and countries was compiled and cleaned
2. Second, the location information was added to this table. Any rows with missing critical information were dropped.
3. Venue data for each location was obtained using Foursquare API and then cleaned.
4. A table was compiled containing critical information: city, country, latitude, longitude, and the 5 most common venues.
5. K-means was used to cluster the various cities.
    a. First, the elbow method was used to determine the number of clusters.
    b. Then, K-means was used to cluster the cities and map them
    c. Critical information about each cluster was obtained: cities in the cluster, most common venues in the cluster
    d. The Euclidean distances between the various clusters were obtained
6. A final summary table containing the most critical information was generated. This table includes the cities in each cluster, the most popular venues in the cluster, the Euclidean distances between the clusters, and the farthest cluster (cities that are most different)

Table with Countries and Capitals

The first objective is to compile a table with the countries and capitals. The countries and capital data was obtained by scraping a table from Wikipedia using BeautifulSoup4. Subsequently, Geopy was used to obtain coordinate data. Unfortunately, not all location data was available, so the missing data was dropped from the table. Shown below are the first 5 capitals (City/Town) and countries (Country/Territory). Clearly, some countries have multiple cities listed. For example, Ivory Coast has Abidjan (former capital) and Yamoussoukro (new official). Both the cities were used for the purposes of this work.

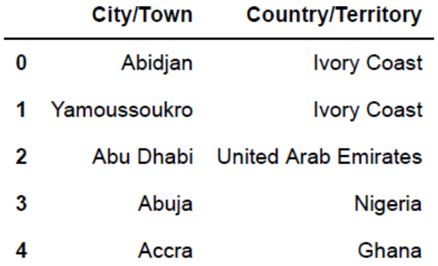
Table 1: Capital and country table scrapped from wikipedia



	City/Town	Country/Territory	Notes
0	Abidjan (former capital; still has many govern...	Ivory Coast	NaN
1	Yamoussoukro (official)	Ivory Coast	NaN
2	Abu Dhabi	United Arab Emirates	NaN
3	Abuja	Nigeria	Lagos was the capital from 1914 to 1991.
4	Accra	Ghana	NaN

There are two main issues: we do not need the Notes column and the additional information in the parentheses in City/Town column. The Notes column was dropped and the information in the parentheses was also dropped to obtain a cleaner table with only the city and country.

Table 2: Capital and country table after clean up

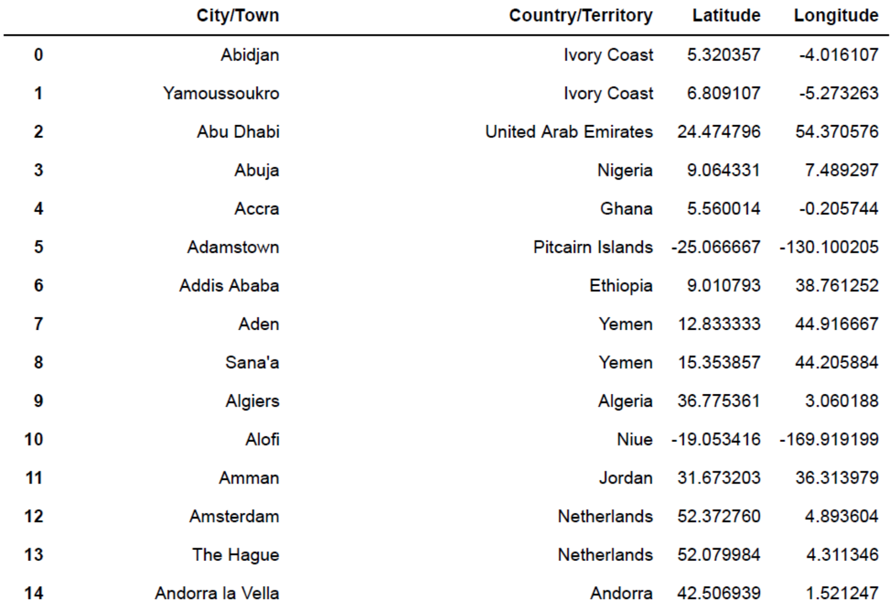


	City/Town	Country/Territory
0	Abidjan	Ivory Coast
1	Yamoussoukro	Ivory Coast
2	Abu Dhabi	United Arab Emirates
3	Abuja	Nigeria
4	Accra	Ghana

Coordinate Information for Each City

The second objective is to add the coordinates for each city. For most of the cities, the latitude and longitude information was obtained. However, after getting multiple timeout errors on the coordinate requests, I had to use the try/except syntax and then drop the locations for which the coordinate information was not obtained. Finally, a table like the one below was obtained, with city and coordinate information. After cleaning up the missing values, there are 262 cities remaining.

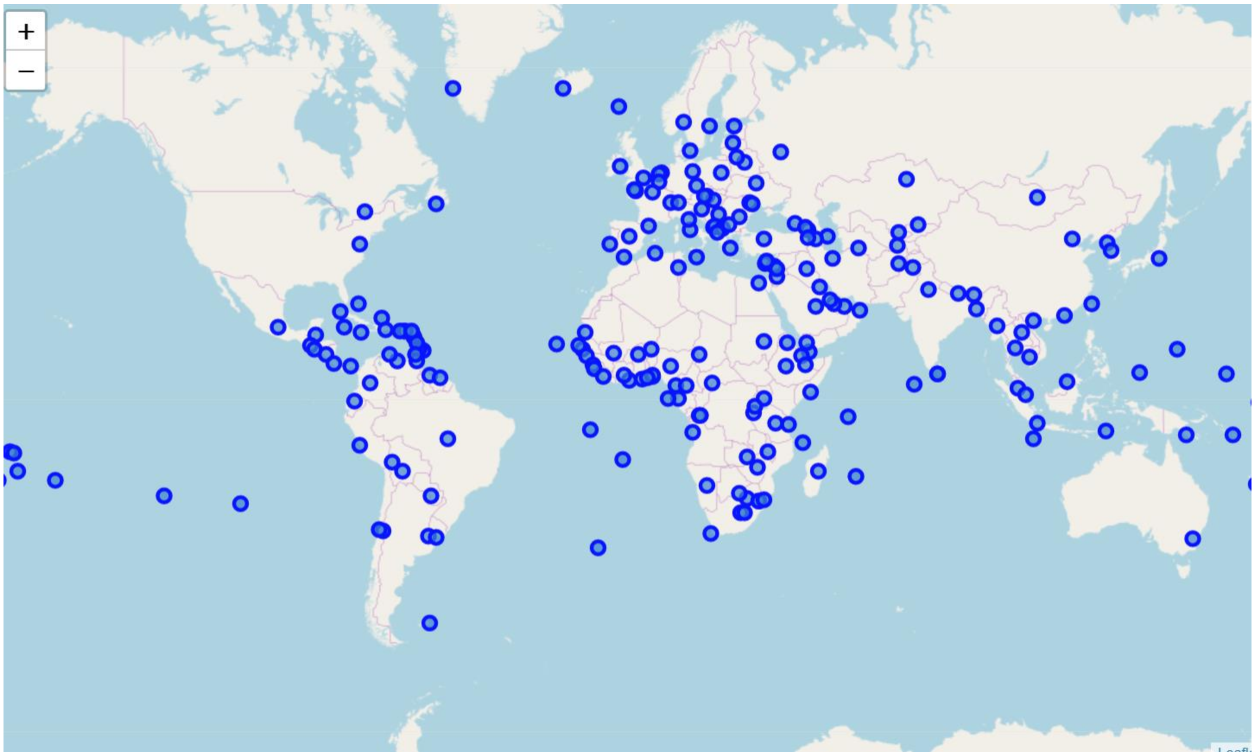
Table 3: Capital, country table, and coordinates



	City/Town	Country/Territory	Latitude	Longitude
0	Abidjan	Ivory Coast	5.320357	-4.016107
1	Yamoussoukro	Ivory Coast	6.809107	-5.273263
2	Abu Dhabi	United Arab Emirates	24.474796	54.370576
3	Abuja	Nigeria	9.064331	7.489297
4	Accra	Ghana	5.560014	-0.205744
5	Adamstown	Pitcairn Islands	-25.066667	-130.100205
6	Addis Ababa	Ethiopia	9.010793	38.761252
7	Aden	Yemen	12.833333	44.916667
8	Sana'a	Yemen	15.353857	44.205884
9	Algiers	Algeria	36.775361	3.060188
10	Alofi	Niue	-19.053416	-169.919199
11	Amman	Jordan	31.673203	36.313979
12	Amsterdam	Netherlands	52.372760	4.893604
13	The Hague	Netherlands	52.079984	4.311346
14	Andorra la Vella	Andorra	42.506939	1.521247

A plot showing all these cities on a world map was displayed using Folium and is reproduced next:

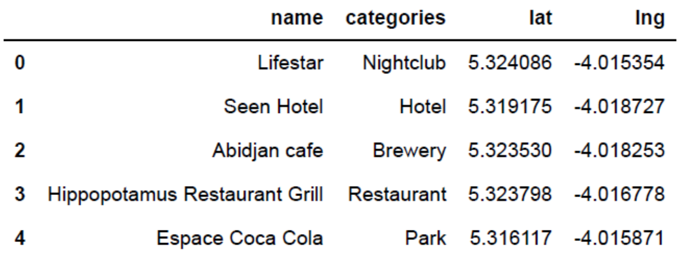
Figure 1: World Map with Markers for All Capital Cities



Generating Venue Data

The venue data was generated using the Foursquare API. By iterating over each city in the table, a list of venues, venue categories, and coordinate data was generated. For example, for the first city, Abidjan, the first five venues are shown below.

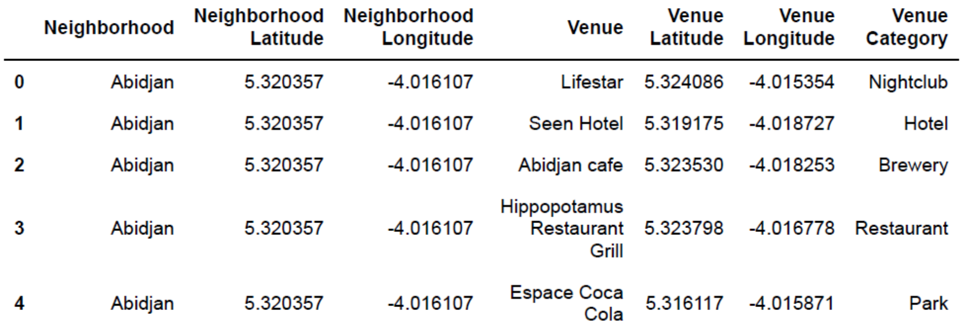
Table 4: Example Venue Information with Venue Name, Categories, and Coordinates



	name	categories	lat	lng
0	Lifestar	Nightclub	5.324086	-4.015354
1	Seen Hotel	Hotel	5.319175	-4.018727
2	Abidjan cafe	Brewery	5.323530	-4.018253
3	Hippopotamus Restaurant Grill	Restaurant	5.323798	-4.016778
4	Espace Coca Cola	Park	5.316117	-4.015871

This was then combined with the city to generate a masterlist of each venue for each city. An example is reproduced below.

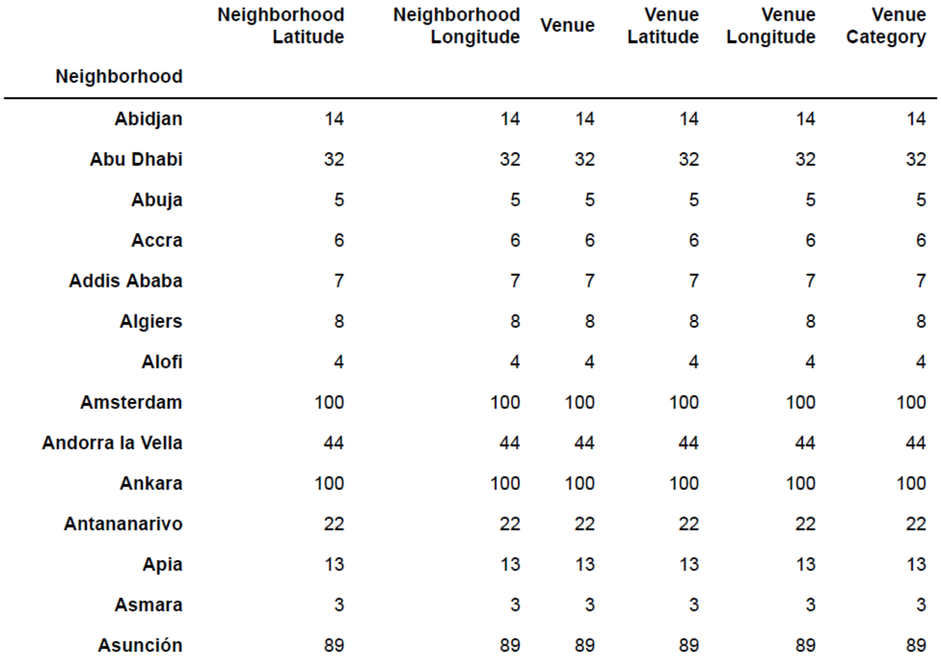
Table 5: Example Venue Information Combined with City



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abidjan	5.320357	-4.016107	Lifestar	5.324086	-4.015354	Nightclub
1	Abidjan	5.320357	-4.016107	Seen Hotel	5.319175	-4.018727	Hotel
2	Abidjan	5.320357	-4.016107	Abidjan cafe	5.323530	-4.018253	Brewery
3	Abidjan	5.320357	-4.016107	Hippopotamus Restaurant Grill	5.323798	-4.016778	Restaurant
4	Abidjan	5.320357	-4.016107	Espace Coca Cola	5.316117	-4.015871	Park

Briefly exploring this data shows that while some cities have a large number of venues, some have a smaller number. For example, Amsterdam has atleast 100 venues of data, while Alofi and Asmara have less than 5 venues each. This might pose a problem later due to the insufficiency of data for clustering.

Table 6: Number of Venues for the Different Cities



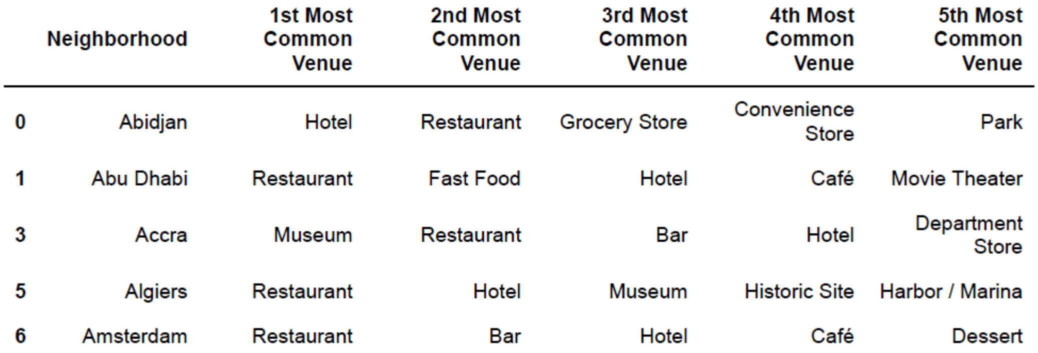
	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Abidjan	14	14	14	14	14	14
Abu Dhabi	32	32	32	32	32	32
Abuja	5	5	5	5	5	5
Accra	6	6	6	6	6	6
Addis Ababa	7	7	7	7	7	7
Algiers	8	8	8	8	8	8
Alofi	4	4	4	4	4	4
Amsterdam	100	100	100	100	100	100
Andorra la Vella	44	44	44	44	44	44
Ankara	100	100	100	100	100	100
Antananarivo	22	22	22	22	22	22
Apia	13	13	13	13	13	13
Asmara	3	3	3	3	3	3
Asunción	89	89	89	89	89	89

To complicate the problem, there are 430 categories, with many redundancies. For example, each different cuisine shows as a different category of restaurant. There were also separate categories for cafes and coffee shops. This was cleaned up by combining the categories to more general categories. The number of categories were reduced to 192. While there are still too many categories, further cutting down the categories seemed arbitrary.

Since ultimately we need to atleast look at 5 unique venues at each city, cities with less than 5 unique venues were dropped from consideration. Cities in India and Australia were dropped. It appears that the data available on Foursquare is far from complete. Perhaps a different or supplemental source could be used to add more data to this set. For now, a significant number of cities were dropped (if the total unique venues were <5), bringing the number of cities down from >200 to 134 cities.

By first generating the frequency of venues at each city, a table comprising the top 5 venues at each city was obtained. It appears that Hotels and Restaurants are the most common in Abidjan, while in Abu Dhabi, fast food and cafes are also popular in addition to hotels and restaurants.

Table 7: Final Table with City and the 5 Most Common Venue Categories



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Abidjan	Hotel	Restaurant	Grocery Store	Convenience Store	Park
1	Abu Dhabi	Restaurant	Fast Food	Hotel	Café	Movie Theater
3	Accra	Museum	Restaurant	Bar	Hotel	Department Store
5	Algiers	Restaurant	Hotel	Museum	Historic Site	Harbor / Marina
6	Amsterdam	Restaurant	Bar	Hotel	Café	Dessert

With the data in the right format, the cities are ready to be grouped into different clusters.

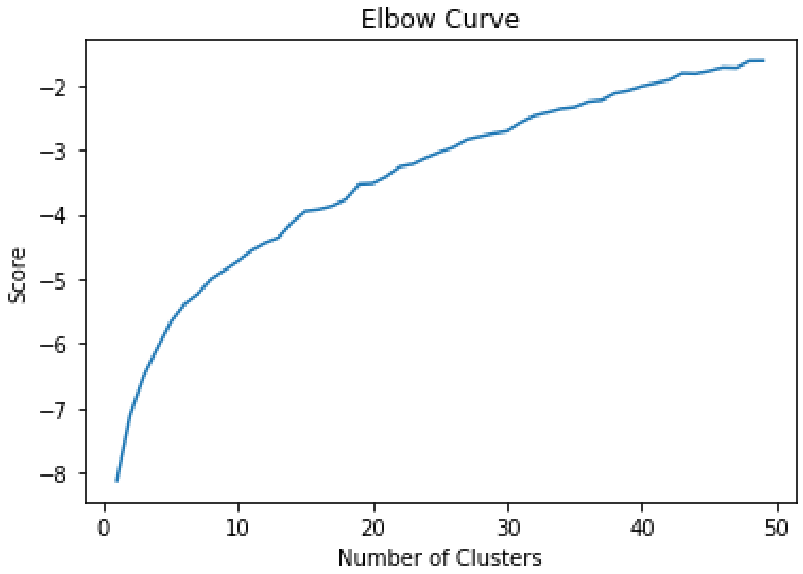
K-means Clustering

Before using k-means, the elbow method was used to determine the number of clusters we should use. The elbow method is a somewhat subjective method, that looks for an inflection in the plot of Accuracy vs. Number of Clusters. The higher the number of clusters, the more accurately we can place each city into a cluster. Yet, we would have too many clusters. On the other side, if there are too few clusters, then we are not differenting the cities well. By using kmeans to fit the City and venue data from the above table, the score was obtained and plotted as shown below. While there is no sign of a clean elbow, the inflection appears to happen around 7 or 8 clusters.

The number of cluster was chosen to be 8. Using k-means fitting, each city was fitted to one of eight clusters. A map was generated using Folium to show the clustering across all the cities.

To find the similarities between the various clusters, Euclidean Distances was imported from sklearn. Using this function, a table was generated showing the distances between all the clusters. In addition, an additional column was added to this table to show the farthest cluster of countries. Using this, one can find the city/cities that are least like the ones in the original cluster. If I were planning to visit a place completely unlike what I have been to in the past, this is a very useful column.

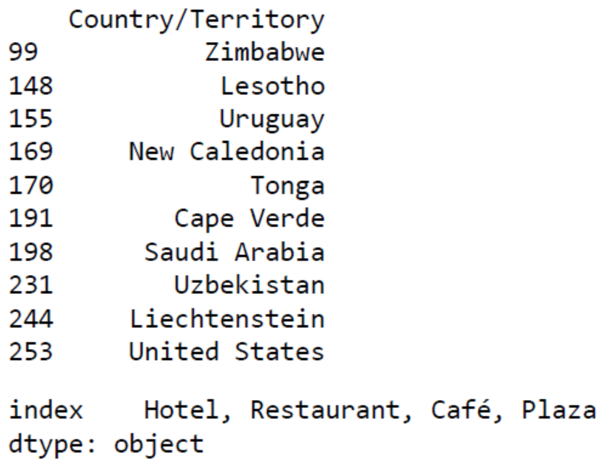
Figure 2: Elbow Curve to Determine accuracy of Kmeans vs. Number of Clusters



Summary Table

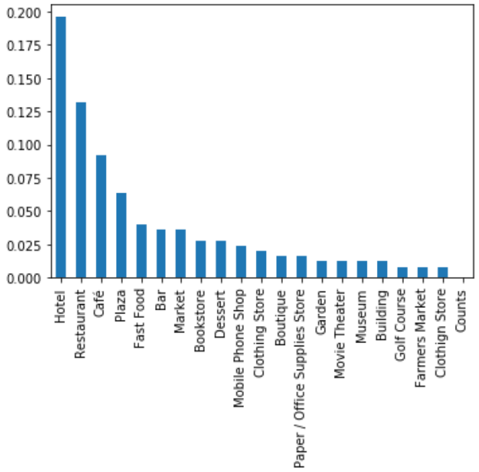
Finally, we had a deeper look at each of the clusters. What are the cities/countries in each of the clusters? What are the common venue categories in these places? A bar plot was also generated for each cluster of cities/countries showing the most common venues. To differentiate between the 1st to the 5th most common venues, a weight was assigned (somewhat arbitrarily) to each of the most common venue categories. The 1st, 2nd, 3rd, 4th, and 5th most common venues had a weight of 1.2, 1.0, 0.8, 0.6, and 0.4, respectively. This is reflected in the bar plots. An example using cluster 0 is reproduced below.

Figure 3: Cluster 0 Example: Countries, Categoris, and Bar Plot of Common Categories



```
    Country/Territory
99           Zimbabwe
148           Lesotho
155           Uruguay
169     New Caledonia
170             Tonga
191        Cape Verde
198      Saudi Arabia
231        Uzbekistan
244     Liechtenstein
253     United States

index    Hotel, Restaurant, Café, Plaza
dtype: object
```



In cluster 0, the countries are Zimbabwe, Lesotho, Uruguay and the United States among others. The most common venues in these cities are hotels, restaurants, cafes and plazas. Similar study was performed on each of the other clusters.

A final summary table comprising the most important learnings was compiled. For each cluster, the countries within the cluster, the representative venue categories, Euclidean distances to the other clusters, and the farthest cluster was displayed.

Results

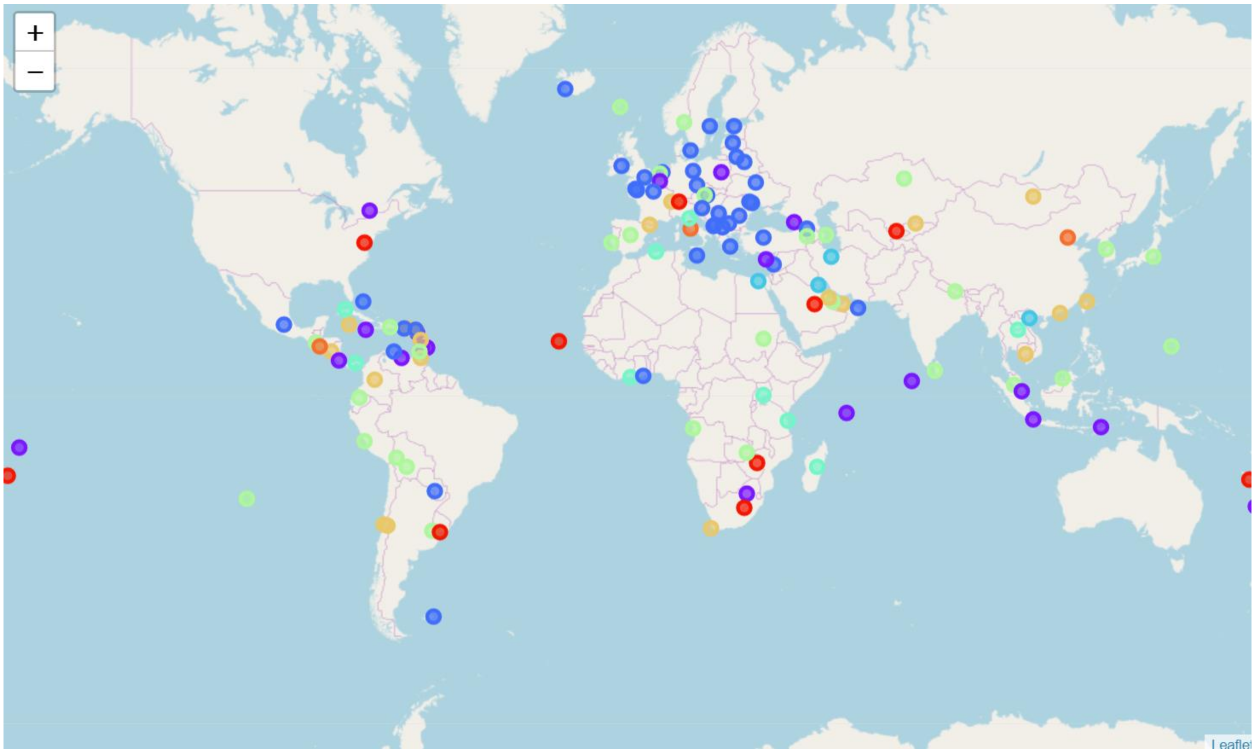
There are two key results that I want to summarize: 1) a map showing how the capitals/countries across the world are clustered, and 2) the summary table with the most critical learnings.

Clustered World map

At a glance, we can see from the map that there are 8 clusters. The biggest cluster appears in blue and is prevalent across a majority of the European cities and countries, and central America. Otherwise the clusters are well mixed.

However, one glaring problem in this map is the lack of a significant number of places. Many of the countries in Africa are missing; so are Australia, and some of the southeast Asian countries. Many of these are missing because there are not many venues that we could obtain in these countries.

Figure 4: World map Showing Capital Cities Clustered into 8 Groups

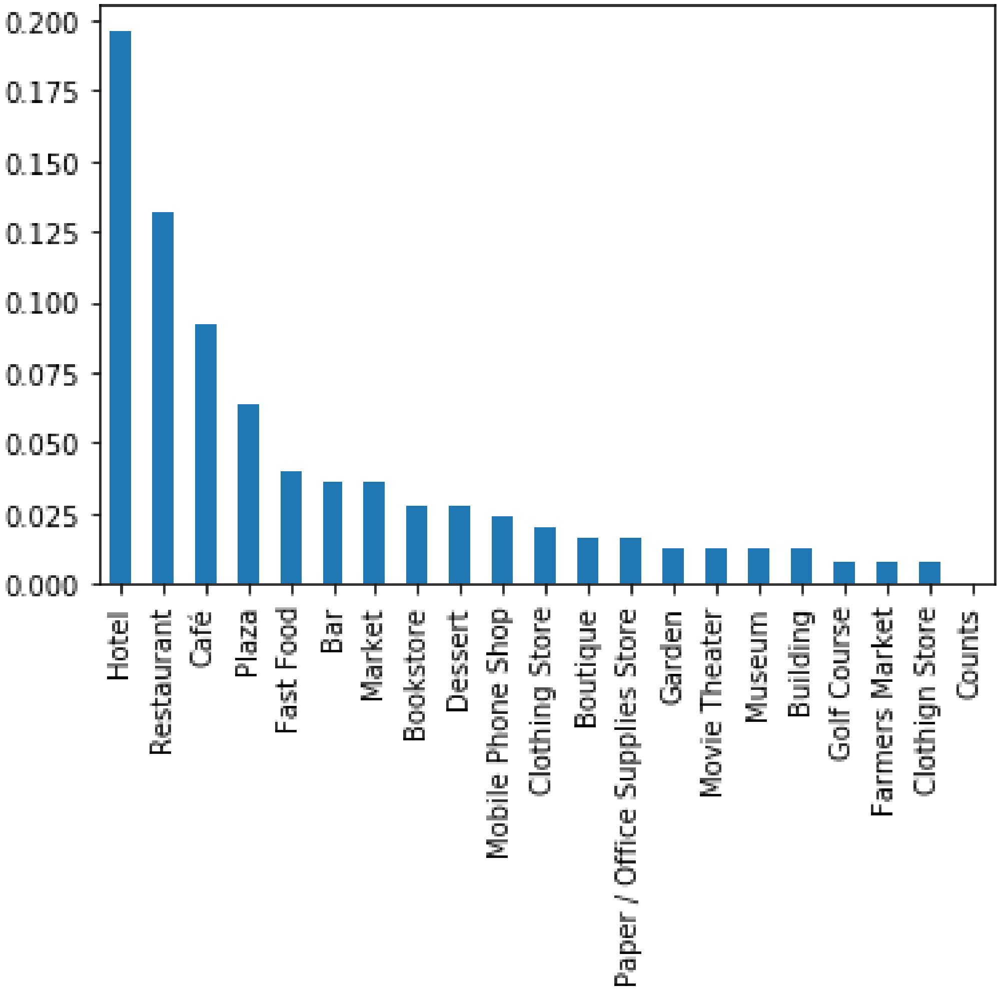


For each of the clusters, the countries and the most common venues are described below.

Cluster 0: Hotels, Restaurants, Café, Plaza

Countries: Zimbabwe, Lesotho, Uruguay, New Caledonia, Tonga, Cape Verde, Saudi Arabia, Uzbekistan, Liechtenstein, and United States

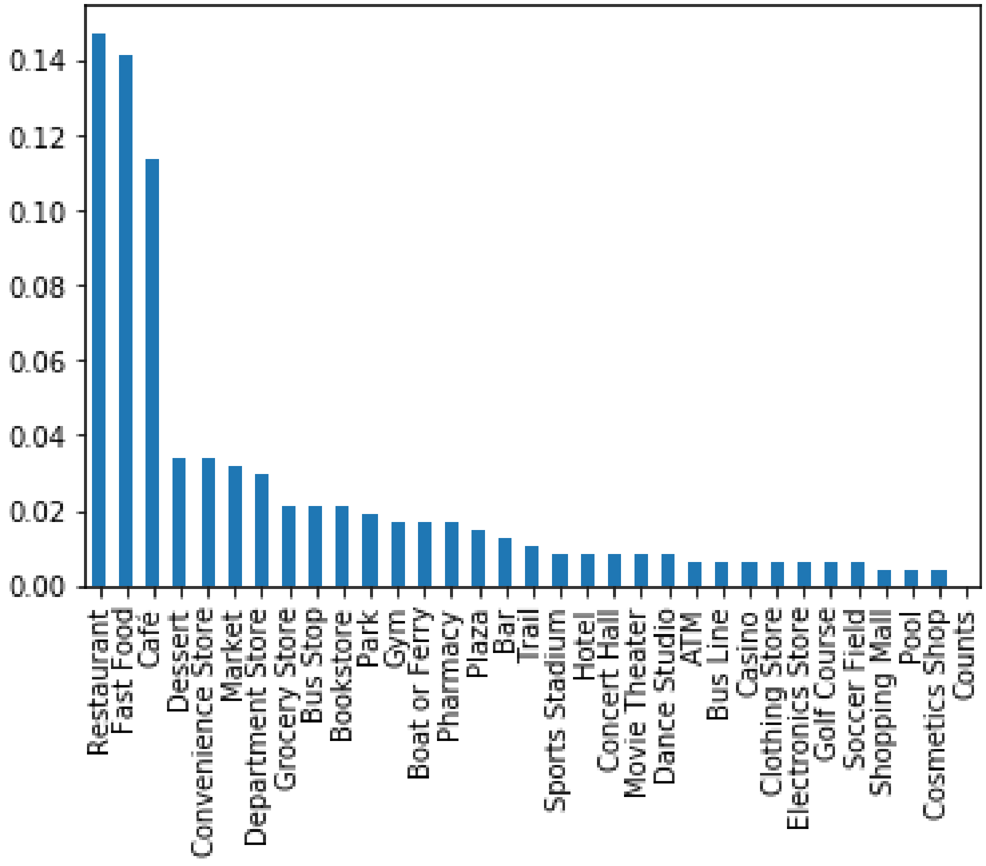
Figure 5: Common Venue Category Bar Plot for Cluster 0



Cluster 1: Restaurants, Fast Food, Café, Dessert (Convenience Store, Department Store)

Countries: Samoa, Saint Kitts and Nevis, South Africa, Barbados, Belgium, Venezuela, East Timor, Indonesia, Jamaica, Norfolk Island, Saint Vincent and the Grenadines, Maldives, Northern Cyprus, Canada, Costa Rica, Singapore, Abkhazia, Seychelles, and Poland

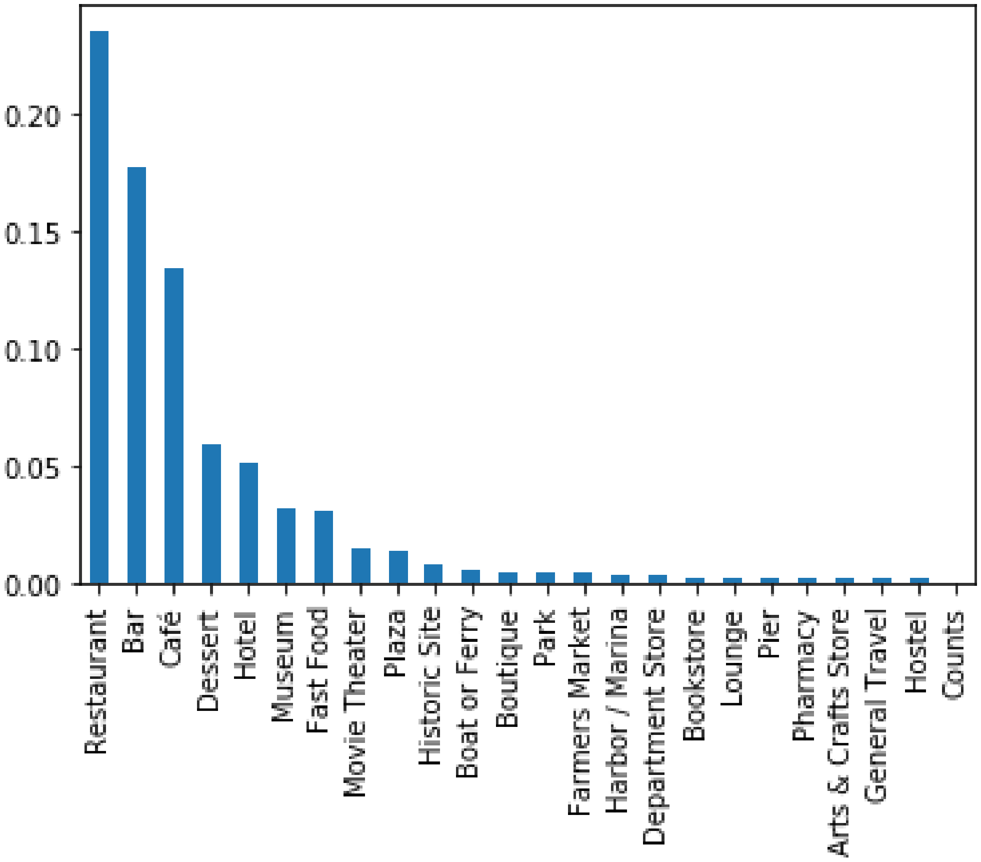
Figure 6: Common Venue Category Bar Plot for Cluster 1



Cluster 2: Restaurants, Bar, Café, Dessert (Museum)

Countries: Ghana, Netherlands, Turkey, Paraguay, Greece, Lebanon, Serbia, Germany, Slovakia, Romania, Montenegro, Moldova, Denmark, Ireland, Ukraine, United Kingdom, Mexico, Belarus, Oman, Bahamas, France, Sint Maarten, Czech Republic, Iceland, Latvia, Puerto Rico, North Macedonia, Bulgaria, Jersey, Guernsey, Falkland Islands, Sweden, Estonia, Georgia, Transnistria, Malta, Lithuania, New Zealand, Curacao, Croatia

Figure 7: Common Venue Category Bar Plot for Cluster 2

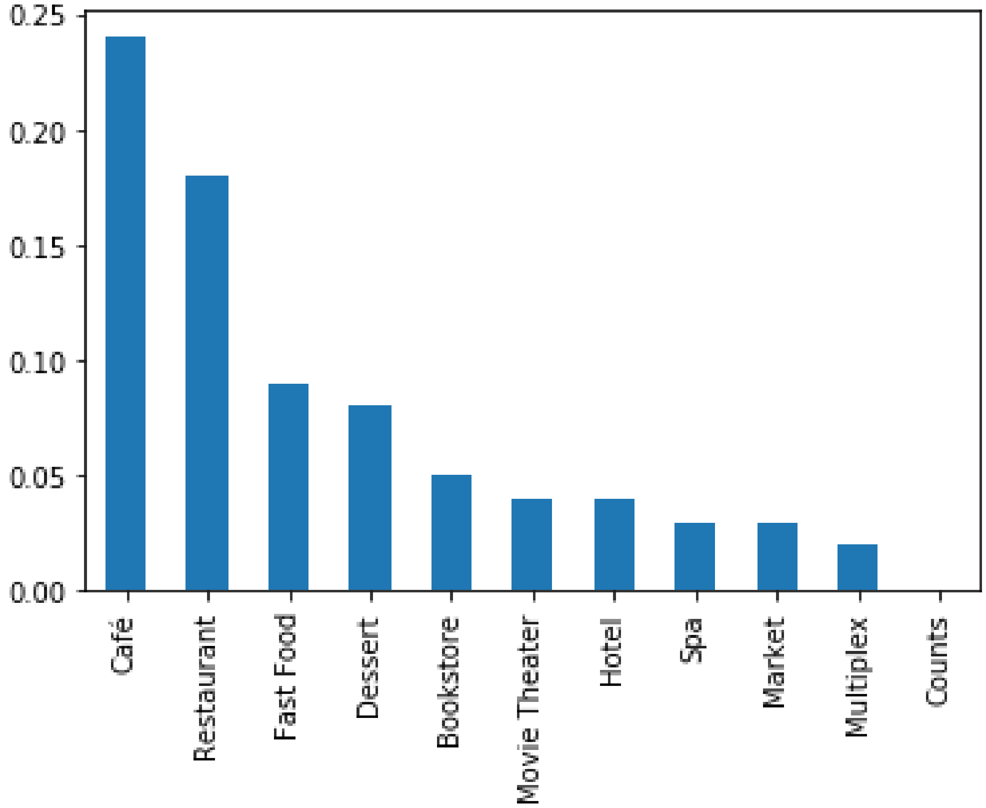


Cluster 3: Café, Restaurant, Fast Food, Dessert (Bookstore, Movie Theater)

Less Common: Museums, Hotels, Sports Stadiums, Jewelry stores

Countries: Egypt, Vietnam, Kuwait, Iran

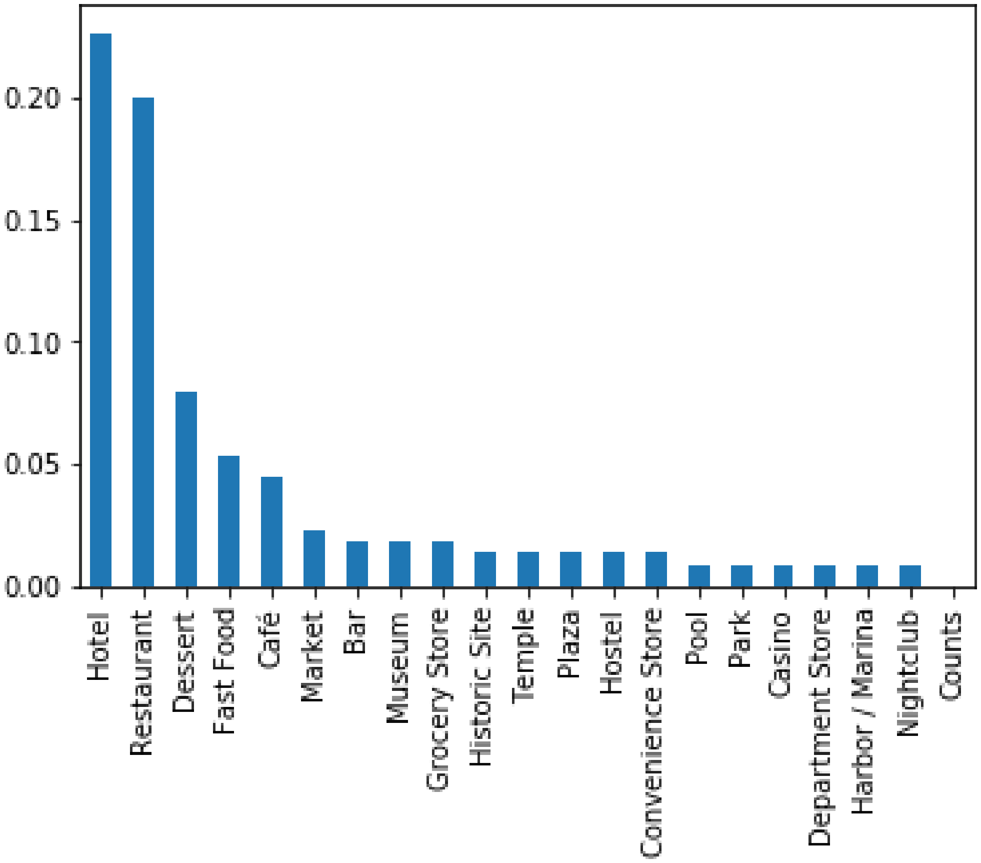
Figure 8: Common Venue Category Bar Plot for Cluster 3



Cluster 4: Hotel, Restaurant, Dessert, Fast Food (Grocery Store, Temple)

Countries: Ivory Coast, Algeria, Madagascar, Tanzania, Cuba, Uganda, Panama, San Marino, Laos

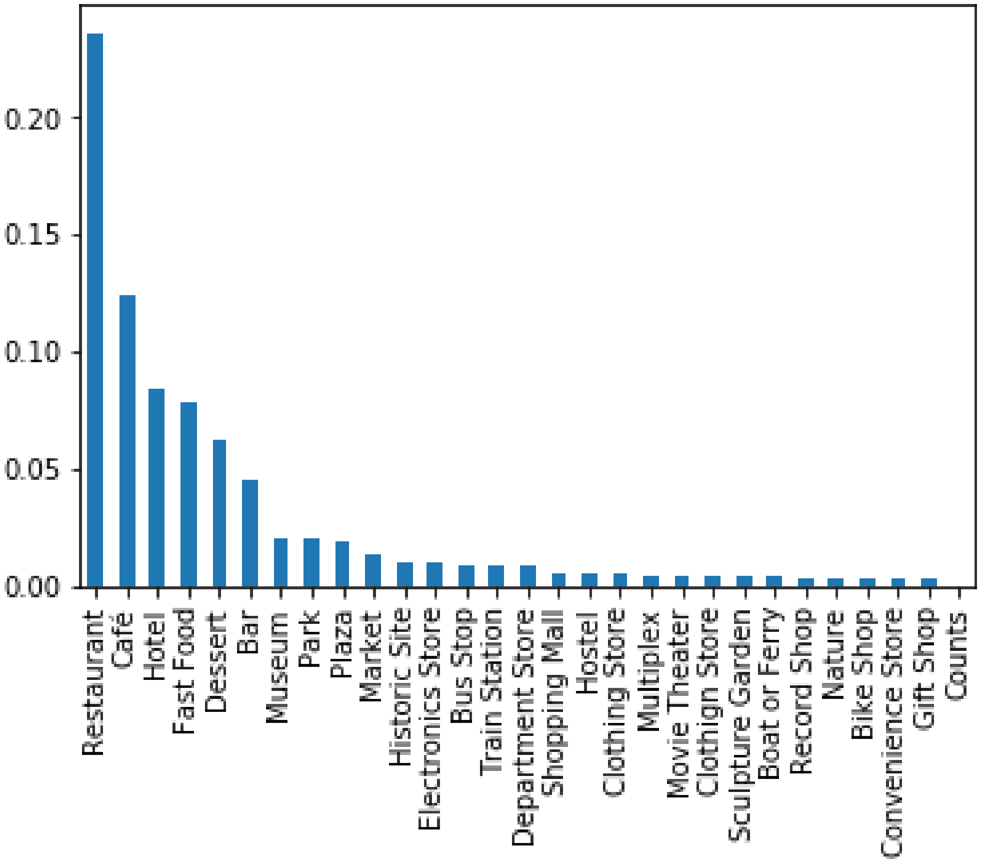
Figure 9: Common Venue Category Bar Plot for Cluster 4



Cluster 5: Restaurant, Café, Hotel, Fast Food (Park, Historic Site, Electronic Store, Public Transport)

Countries: Netherlands, Azerbaijan, Brunei, Argentina, Sri Lanka, Qatar, Guatemala, Guam, Easter Island, Nepal, Sudan, Malaysia, Bolivia, Peru, Portugal, Angola, Zambia, Spain, Kazakhstan, Norway, Ecuador, Dominican Republic, South Korea, Grenada, Japan, Faroe Islands, Austria, Armenia

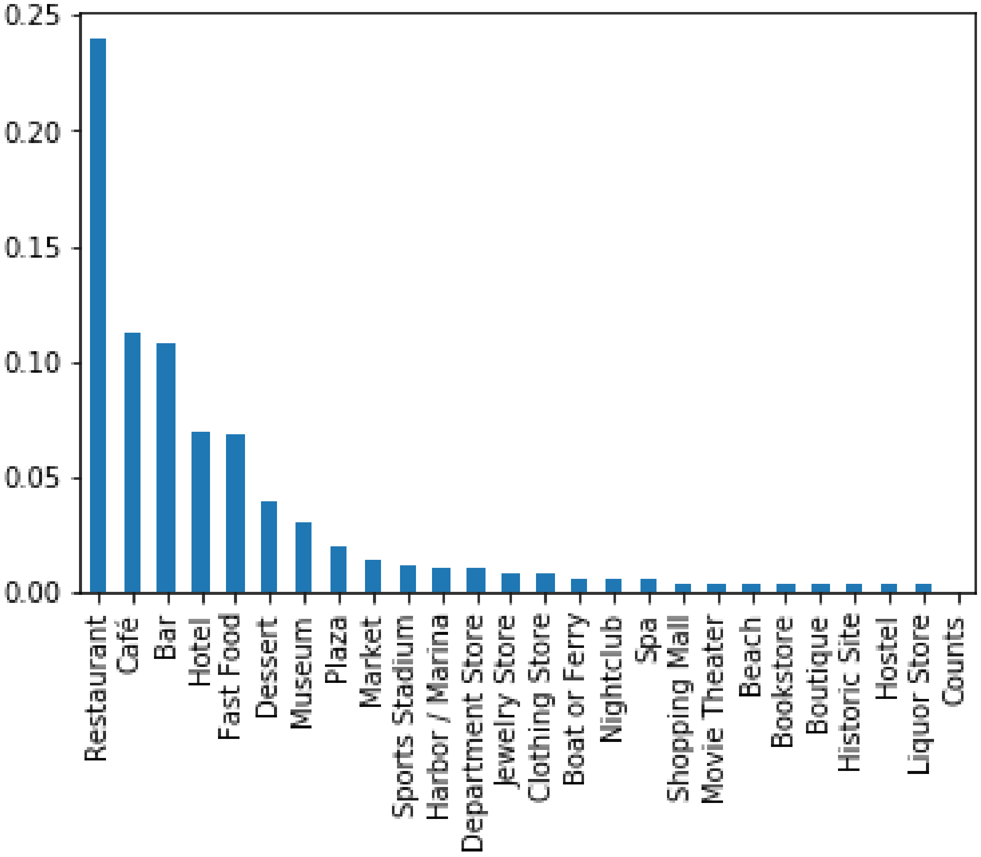
Figure 10: Common Venue Category Bar Plot for Cluster 5



Cluster 6: Restaurant, Café, Bar, Hotel (Sports Stadium, Jewelry Store, Nightclub)

Countries: UAE, Andorra, Switzerland, Kyrgyzstan, South Africa, Colombia, United States Virgin Islands, Cayman Islands, Saint Barthelemy, Nicaragua, Bahrain, Saint Martin, Cambodia, Trinidad and Tobago, Dominica, Chile, Taiwan, Hong Kong, Mongolia

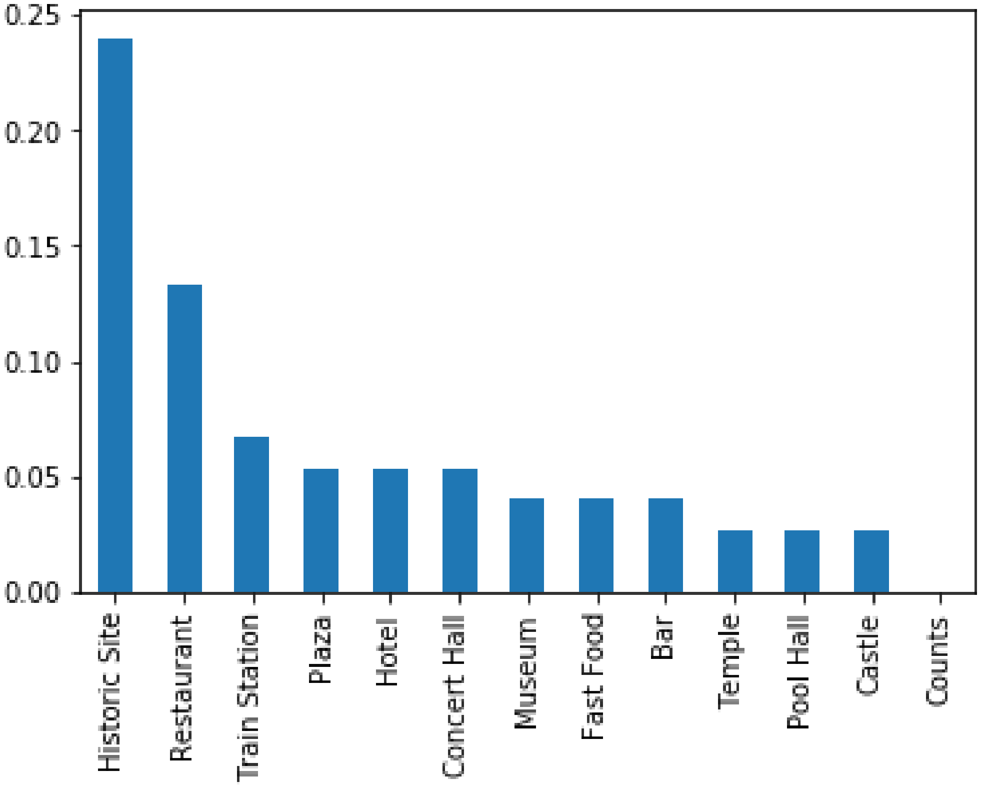
Figure 11: Common Venue Category Bar Plot for Cluster 6



Cluster 7: Historic Site, Restaurant, Train Station, Plaza (Concert Hall)

Countries: China, Italy, El Salvador

Figure 12: Common Venue Category Bar Plot for Cluster 1



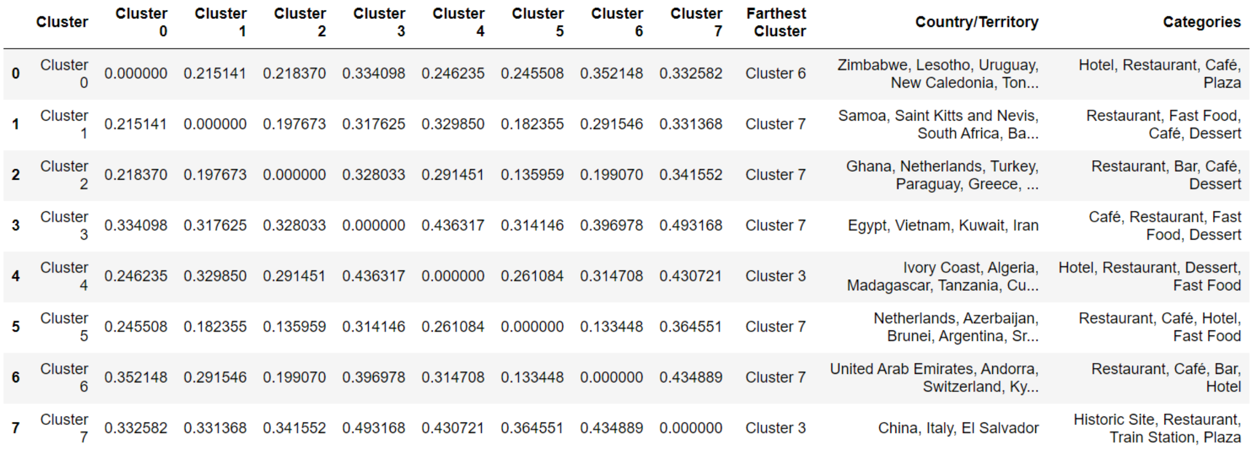
Summary Table

The table below summarizes the key learnings. It includes Euclidean distances for all the clusters, the countries in each cluster, the venue categories in each cluster, and a column which indicates the farthest cluster (cluster of places least like those in the original cluster).

At a glance, cluster 7 seems to be farthest cluster for most clusters. Unlike most of the clusters with a heavy emphasis on restaurants, hotels, fast food, desserts, countries in cluster 7 contain many historic sites and train stations. Cluster 3 appears to contain a lot of Cafés.

At the same time, many clusters have a large overlap. All of the clusters contain a significant number of restaurants, and most contain cafes and hotels. So this analysis doesn't help someone look at those categories. If interested in bars, clusters 2 and 6 appear to be the places to go.

Table 8: Summary Table Showing Cluster Country, Common Categories, Euclidean Distances to Other Clusters, and Most Dissimilar Cluster of Countries



	Cluster	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Farthest Cluster	Country/Territory	Categories
0	Cluster 0	0.000000	0.215141	0.218370	0.334098	0.246235	0.245508	0.352148	0.332582	Cluster 6	Zimbabwe, Lesotho, Uruguay, New Caledonia, Ton...	Hotel, Restaurant, Café, Plaza
1	Cluster 1	0.215141	0.000000	0.197673	0.317625	0.329850	0.182355	0.291546	0.331368	Cluster 7	Samoa, Saint Kitts and Nevis, South Africa, Ba...	Restaurant, Fast Food, Café, Dessert
2	Cluster 2	0.218370	0.197673	0.000000	0.328033	0.291451	0.135959	0.199070	0.341552	Cluster 7	Ghana, Netherlands, Turkey, Paraguay, Greece, ...	Restaurant, Bar, Café, Dessert
3	Cluster 3	0.334098	0.317625	0.328033	0.000000	0.436317	0.314146	0.396978	0.493168	Cluster 7	Egypt, Vietnam, Kuwait, Iran	Café, Restaurant, Fast Food, Dessert
4	Cluster 4	0.246235	0.329850	0.291451	0.436317	0.000000	0.261084	0.314708	0.430721	Cluster 3	Ivory Coast, Algeria, Madagascar, Tanzania, Cu...	Hotel, Restaurant, Dessert, Fast Food
5	Cluster 5	0.245508	0.182355	0.135959	0.314146	0.261084	0.000000	0.133448	0.364551	Cluster 7	Netherlands, Azerbaijan, Brunei, Argentina, Sr...	Restaurant, Café, Hotel, Fast Food
6	Cluster 6	0.352148	0.291546	0.199070	0.396978	0.314708	0.133448	0.000000	0.434889	Cluster 7	United Arab Emirates, Andorra, Switzerland, Ky...	Restaurant, Café, Bar, Hotel
7	Cluster 7	0.332582	0.331368	0.341552	0.493168	0.430721	0.364551	0.434889	0.000000	Cluster 3	China, Italy, El Salvador	Historic Site, Restaurant, Train Station, Plaza

Looking beyond some of the most common categories, some of the less common categories are popular in the following clusters. The list does not mean that these are not available in other clusters, but that they are more prominent in the clusters listed below:

- Cluster 0: Phone Shop, Garden
- Cluster 1: Convenience Store, Department Store, Grocery Store
- Cluster 2:
- Cluster 3: Bookstore, Movie Theater
- Cluster 4: Grocery Store, Temple
- Cluster 5: Park, Historic Site, Electronic Store
- Cluster 6: Sports Stadiums, Jewelry Store, Nightclub
- Cluster 7: Historic Sites, Concert Halls, and Museums

Discussion

Two important results, clustered map and summary table, directly address the business problem that was defined in the first section. The map shows how various capital cities are clustered around the world and enables one to see whether they are inclined towards certain cluster of cities. In addition, at a glance, it provides a visual key to alternate clusters.

From the map, a lot of European cities seem to fall into cluster 2, with many restaurants, bars, cafés, and dessert places. Rest of the places seem to show a more random mix. Personally, another interesting aspect to the map is that it provides locations and context to places I have never considered. For example, there are many different scenes in the dense central America region, southeast Africa, and in the Indian ocean.

Looking more specifically at each of the clusters, there appears to be a big overlap between many of the clusters. In fact, the top 3 or 4 categories in most of the clusters come from a combination of restaurants, hotels, desserts, fast food, bars, and cafés. The emphasis is definitely on the food scene. Yet, there are many other aspects that travelers could be interested in that are lost in this study. This will be addressed further in the conlusions/outlook section.

The summary table addresses the directly the finer points of the business problem. First, it shows the options of categories available within the limited scope of this work. From this, we can more clearly see the categories overlapping between clusters. While this is expected to some extended, it might be masking important reasons to visit some of these places. At a glance, the table enables knowledge of all the options available.

Second, the Euclidean distances in the table show the extent of the overlap between the clusters and how the next adventure is related to the past travel. If the distance is closer to 0, that means two clusters are very similar. If the distance are large, then the clusters are very different. It quantifies the distance to each cluster, providing one with the choice to do something similar, some different but not too different, or something entirely different.

However, there are many limitations to this work.

1. From the clustering, it is clear that there is big overlap between many of the clusters. This is limiting visibility of other important venues in the cities. This is primarily for two reasons:
    a. Lack of information in many of the cities. A lot of cities had barely 5-10 venues.
    b. Only top 5 venue categories were considered at each city. This was a compromise based on the tradeoff between number of cities used for analysis and number of categories to consider. If more venues had been considered, there would be far fewer cities to analyze.
2. The number of venues we were able to obtain at each venue severely limited the analysis. Out of an initial count of >250 cities, after dropping cities where there no coordinates, few venues and few categories of venues, the city count was down to around 130. A majority of the cities were excluded from two continents, Africa and Asia. Perhaps the lack of information is limiting visibility of potentially exciting destinations.
3. The number of categories was cleaned up to some extent, but not fully consolidated. There were still nearly 200 categories, with somewhat arbitrary grouping of categories.

4. The capitals of each country likely doesn't represent all the places there are to see in the country. While this was outside the scope of this study, any serious future study would need to include many more destinations. Personally, when I have travelled internationally, I have traveled to multiple cities that are nearby.

Conclusion/Outlook

A simple map comprising the various clusters around the world was successfully generated. Kmeans was used to cluster over a 100 cities around the world. And a highly useful starting point was generated in the form of the final table. Yet, more sophistication could yield significantly better insights and recommendations of places. Looking forward, here are few items that could improve this study:

1. To further this aspect, multilevel indexing can be applied such that we would start with fewer broad categories (of the order of 10), under which we would further categorize the 400 or so original categories in this study.
2. Instead of choosing the most common venue categories, another way to look at it would be most popular venues by ratings. This will provide more visibility and importance to categories that are fewer in number but attract people. For example, we don't expect to see as many beaches in any location as the number of restaurants and cafés. If the beach is the draw for people in one our cities, this may not even show up in our analysis.
3. Include more destinations to more fully capture what the country has to offer.
4. Combine data from additional sources, not just Foursqaure.