IBM Data Science Capstone Project

Christopher Lopez-Simons

November 18, 2020

Problem/Background

An avid European traveler wants to visit one of the big cities in the United States. However, he also loves the outdoors and wants to be able to visit various outdoor venues so he can explore not only the big city on some days but also the natural habitat that the United States offers. Back in Europe, he spends a lot of his free time either going to parks or beaches to just enjoy the sun and trees around him. On other days he loves to go on hiking whether it be on trails or at natural parks. I have been tasked by him to find the top ideal locations for him to travel to in the United States so that he can then pick his final travel destination from those selections. He is not restricted by finances or time, so those variables can be ignored for this task.

Data Required

The first task that needs to be done to complete this task is to scrape the most populous cities in the United States from Wikipedia. From there that data will need to be inserted into a dataframe then cleaned to ensure consistency of each city and state element in the dataframe. A longitude and latitude must be found for each state and inputted into the dataframe. With the longitude and latitude data, an explore api using Foursquare can be done on each city to find the top venues within a large radius of the city to ensure nearby outdoor activities can be found. A pandas dummy dataframe can then be created to assign each city with a mean of each type of venue so that a KMeans algorithm can be run to cluster similar cities. A folium map will then be created to showing which city belongs to which cluster. The clusters will be analyzed for their top venue types to see which includes the most outdoor activities. A new folium map will be created with only the cities belonging to the cluster that exhibits the most outdoor activities. Along with that new folium map, a new data frame will be presented with the top venues for each city in that cluster to the customer so he can decide which city to travel to.

Methodology

Initially the scraped data from Wikipedia needed to be cleaned due to excess data as well as unwanted characters in certain columns. As you can see in figure 1, the following columns needed to be removed: 2010Census, Change, 2016 land area, 2016 land area.1, 2016 population density, and population density.1. Along with removing those columns, the City column had some cities with footnotes which can be seen between two brackets such as [d]. This had to be removed to properly show which city is associated with that row of data.

	2019rank	City	State[c]	2019estimate	2010Census	Change	2016 land area	2016 land area.1	2016 population density	2016 population density.1	Location
0	1	New York City[d]	New York	8336817	8175133	+1.98%	301.5 sq mi	780.9 km2	28,317/sq mi	10,933/km2	40°39′49″N 73°56′19″W / 40.6635°N 73.9387°W
1	2	Los Angeles	California	3979576	3792621	+4.93%	468.7 sq mi	1,213.9 km2	8,484/sq mi	3,276/km2	34°01′10″N 118°24′39″W / 34.0194°N 118.4108°W
2	3	Chicago	Illinois	2693976	2695598	−0.06%	227.3 sq mi	588.7 km2	11,900/sq mi	4,600/km2	41°50′15″N 87°40′54″W / 41.8376°N 87.6818°W
3	4	Houston[3]	Texas	2320268	2100263	+10.48%	637.5 sq mi	1,651.1 km2	3,613/sq mi	1,395/km2	29°47′12″N 95°23′27″W / 29.7866°N 95.3909°W
4	5	Phoenix	Arizona	1680992	1445632	+16.28%	517.6 sq mi	1,340.6 km2	3,120/sq mi	1,200/km2	33°34′20″N 112°05′24″W / 33.5722°N 112.0901°W

Figure 1. Uncleaned data that was scraped from Wikipedia

Once the data was cleaned, the location data needed to be altered to a column of latitude and longitude coordinates so that the mapping process later on will be easier to execute. This can be seen in figure 2 below which has both longitude and latitude columns joined to the dataframe and the location column was removed.

	Rank	City	State	PopEstimate	Latitude	Longitude
0	1	New York City	New York	8336817	40.6635	-73.9387
1	2	Los Angeles	California	3979576	34.0194	-118.4108
2	3	Chicago	Illinois	2693976	41.8376	-87.6818
3	4	Houston	Texas	2320268	29.7866	-95.3909
4	5	Phoenix	Arizona	1680992	33.5722	-112.0901
...	...	...	...	...	...	...
95	96	Hialeah	Florida	233339	25.8699	-80.3029
96	97	Richmond	Virginia	230436	37.5314	-77.4760
97	98	Boise	Idaho	228959	43.6002	-116.2317
98	99	Spokane	Washington	222081	47.6669	-117.4333
99	100	Baton Rouge	Louisiana	220236	30.4422	-91.1309

Figure 2. Cleaned dataframe with all relevant data

Since all the dataframe is now clean, a folium map was created of all top 100 cities as seen below in figure 3.

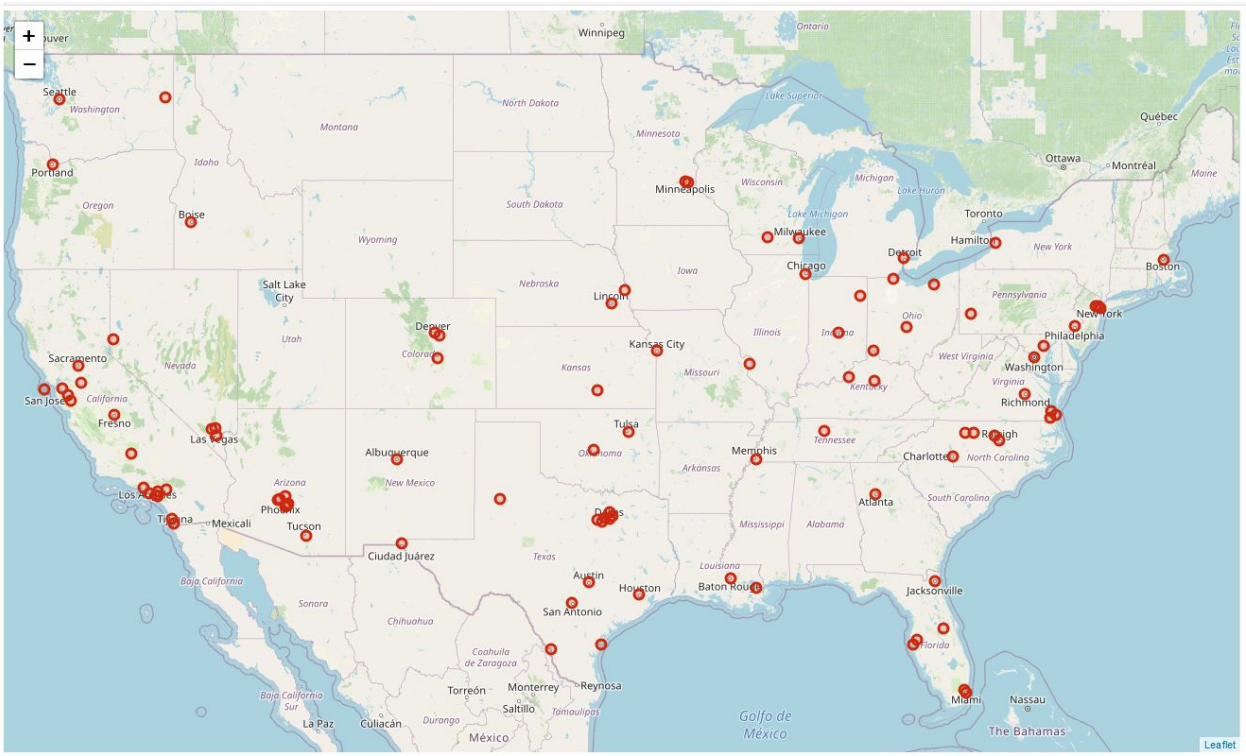

Figure 3. Top 100 cities

With the cleaned dataframe, the next step was to get nearby venues for each city and since the customer is an avid outdoor activity lover, the radius was set to 30km in order to encompass nearby wilderness parks. After the nearby venues were found with Foursqaure, a pandas dummy dataframe was created to be able to numerically analyze the type of venues in each city. All cities were grouped in the dummy dataframe and a mean was found for each venue type so that a kmeans analysis could be performed to cluster the cities. Once all the cities were assigned to one of the ten clusters created, another folium map was created of all the cities along with a popup message with the location, city name, state name, estimated population, and cluster it was associated with. Each cluster had it's own color so the map could be easily read as seen in figure 4 below.

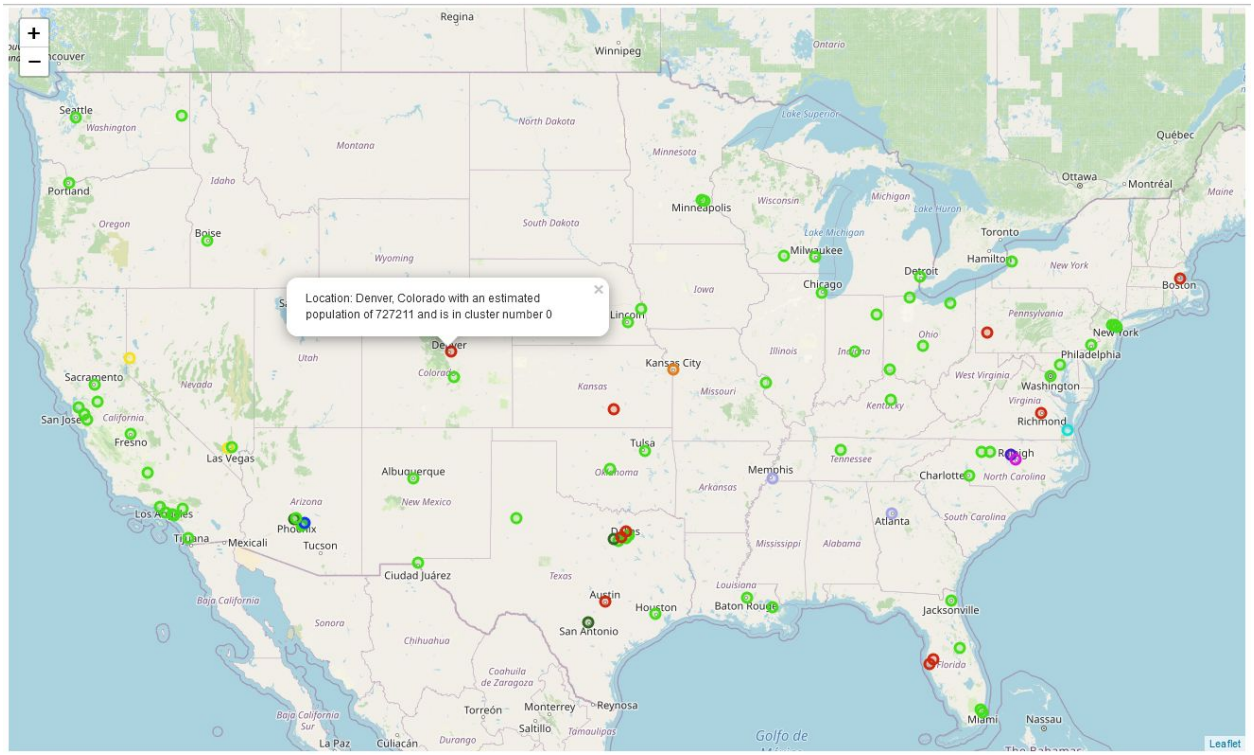

Figure 4. Map of cities along with their associated cluster as seen by the different colors.

The clusters were then grouped together and a sum of their mean of all venue types that an outdoor enthusiast may like. A final score was found for each cluster as seen below in figure 5, which was then used to begin the analysis of each cluster to be presented to the customer.

Cluster	Score
0	0.461082
1	1.000000
2	0.084490
3	0.000000
4	0.000000
5	0.000000
6	0.000000
7	0.333333
8	0.022222
9	0.000000

Figure 5. Outdoor activity score of each cluster.

Results

According to the outdoor activity score of each cluster, the two clusters with the best fit are clusters 0 and 1. The cities within these two clusters were then mapped onto another folium map with a popup message that states the same information as the previous map along with the score of the cluster from the dataframe shown above multiplied by 100 so it's easier to understand, which can be seen below in figure 6.

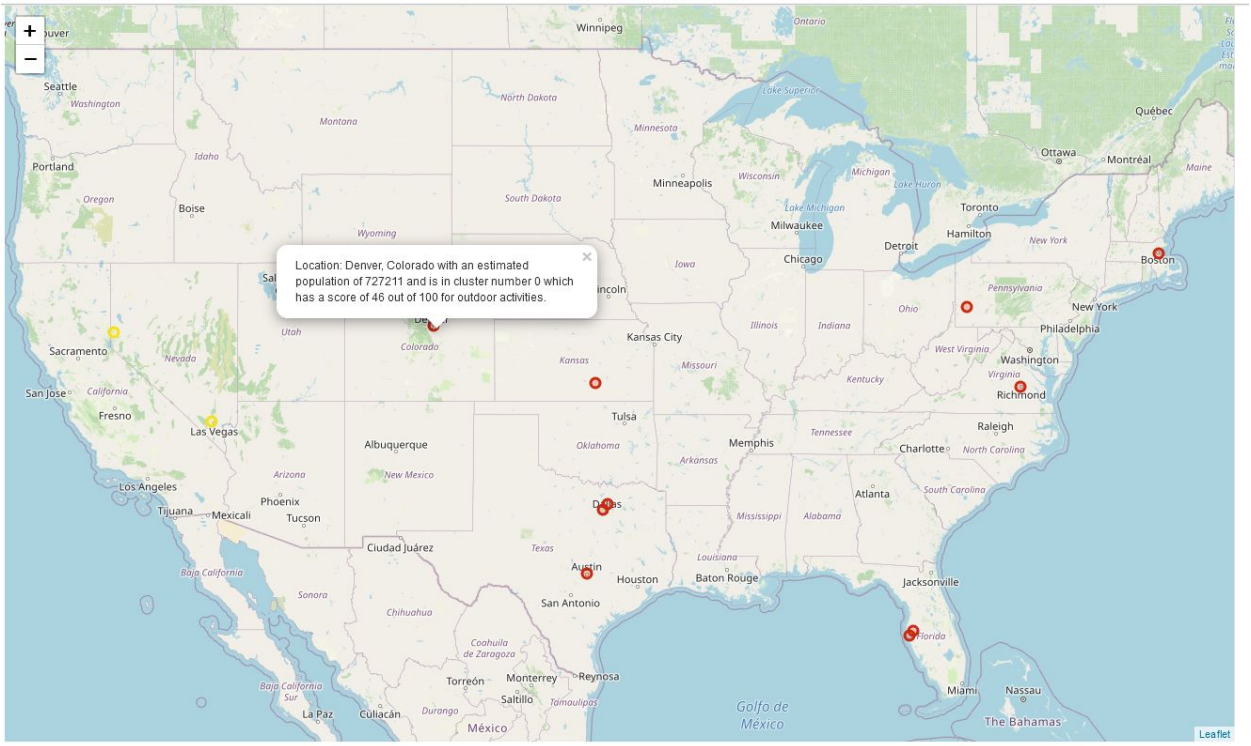

Figure 6. Map of cities in clusters 0 and 1 which have the highest outdoor activity score

In addition to the map in figure 6, a final comprehensive dataframe was created that contained the city, state, score and top venues in that city so that the customer can decide which city he would like to visit based on what the top venues are and the outdoor score. This dataframe can be seen below in figure 7.

City	State	Top Venues	Outdoor Scores (out of 100)
Austin	Texas	Park, Bank, Bridal Shop, Cosmetics Shop, Gym, Ice Cream Shop, Music Store, Playground, Spa, Yoga Studio	46
Boston	Massachusetts	Beach, Park, History Museum, Lighthouse, Playground	46
Denver	Colorado	Park, Bus Stop, Gym / Fitness Center, Pool	46
Irving	Texas	Home Service, Playground	46
Pittsburgh	Pennsylvania	Deli / Bodega, Grocery Store, Neighborhood, Park, Post Office, Sandwich Place	46
Plano	Texas	Garden, Gym, Lounge, Park	46
Richmond	Virginia	Garden, Beach, Garden Center, Lake, Park, Trail	46
St. Petersburg	Florida	Café, Discount Store, Park	46
Tampa	Florida	BBQ Joint, Park	46
Wichita	Kansas	Monument / Landmark, Movie Theater, Park, River, Science Museum, Sculpture Garden	46
Las Vegas	Nevada	Trail	100
Reno	Nevada	Trail	100

Figure 7. Final dataframe with all relevant information succinctly illustrated.

Discussion

Based on figure 7, I would highly recommend Richmond, Virginia because the top venues there are all outdoor activities and it provides a large variety of activities when compared to both cities that received a 100 outdoor score. If the customer really enjoys hikes on trails, then either Las Vegas or Reno would be a great place to visit according to the data retrieved from Foursquare as there looks to be many trails in those two cities. An observation that I noticed was that cluster 2 had quite a large amount of cities associated with it suggesting many of the top cities are quite similar in their types of venues.

Conclusion

In this study I analyzed the top 100 cities in the United States to find which one of them would be ideal for a traveler who enjoys outdoor activities. It was quite a fun project that I learned a lot from when it comes to data analysis and visualizing the conclusions associated with the project. This type of model could be quite helpful for travel package companies that could recommend places to visit based on customer input of what their desirable venue types are.