IBM Data Science Coursera Capstone Project

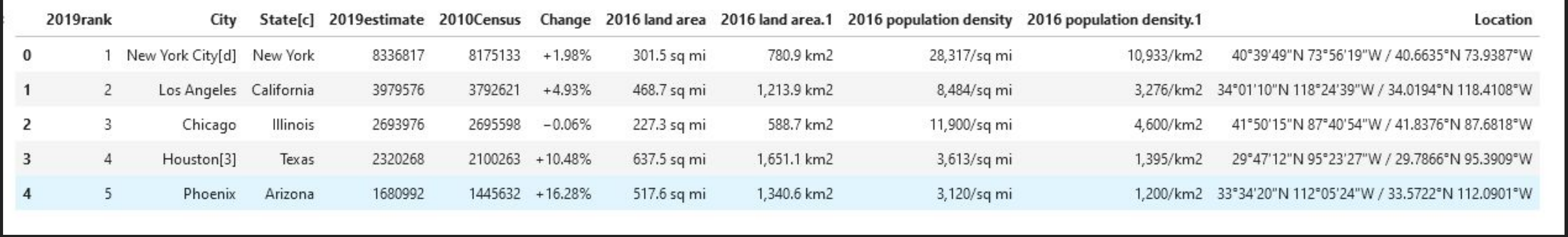
Christopher Lopez-Simons
November 19, 2020

Customer Request

- Avid traveler from Europe wants to travel to a major city in the United States
- He is loves outdoor activities ranging from hanging out on a beach to hiking
- The goal of this project is to find cities in the United States that meet both his requirements (major city, numerous outdoor activities) so that he can decide which city he would like to visit

https://media.istockphoto.com/photos/venetberg-austria-august-02-2017-single-hiker-in-the-early-morning-at-picture-id1054012932?k=6&m=1054012932&s=612x612&w=0&h=ZD-tIT-bRDyrRl5TXaxjNJ2r3eqWr1rYWdkFpieJjI8=

Web Scraping

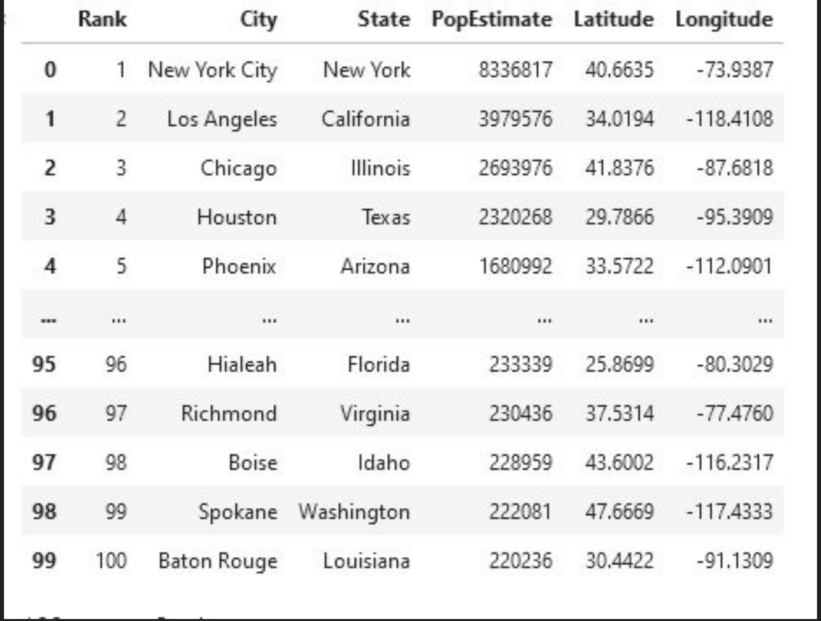


	2019rank	City	State[c]	2019estimate	2010Census	Change	2016 land area	2016 land area.1	2016 population density	2016 population density.1	Location
0	1	New York City[d]	New York	8336817	8175133	+1.98%	301.5 sq mi	780.9 km2	28,317/sq mi	10,933/km2	40°39'49"N 73°56'19"W / 40.6635°N 73.9387°W
1	2	Los Angeles	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

- Table of most populous cities in the United States was scraped from Wikipedia
- Dataframe is inconsistent, ill-formatted and contains excess information

Data Cleaning

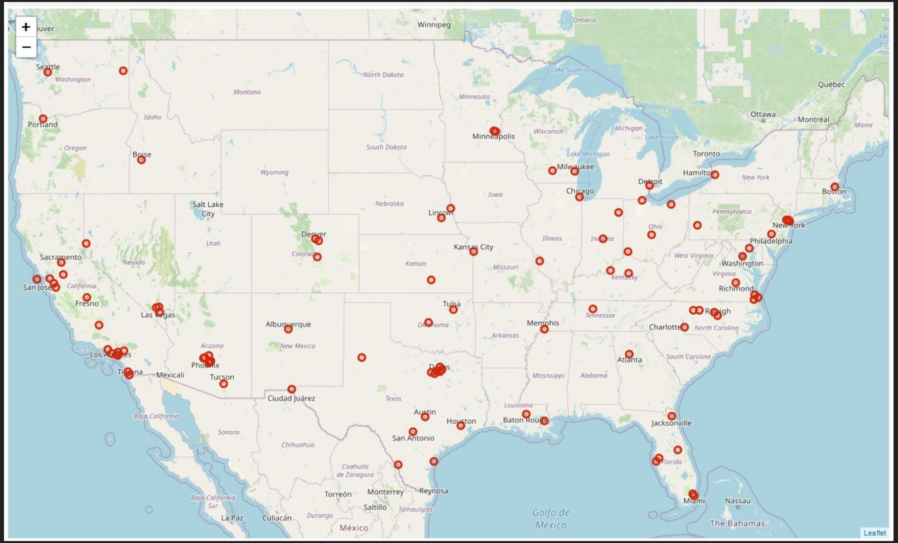
- Initial dataframe collected from Wikipedia was cleaned
  - Top 100 cities were sliced from original dataframe
  - Unnecessary information was removed
  - Characters in the City column that represented footnotes were removed
  - Location column was split up into latitude and longitude then removed



	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

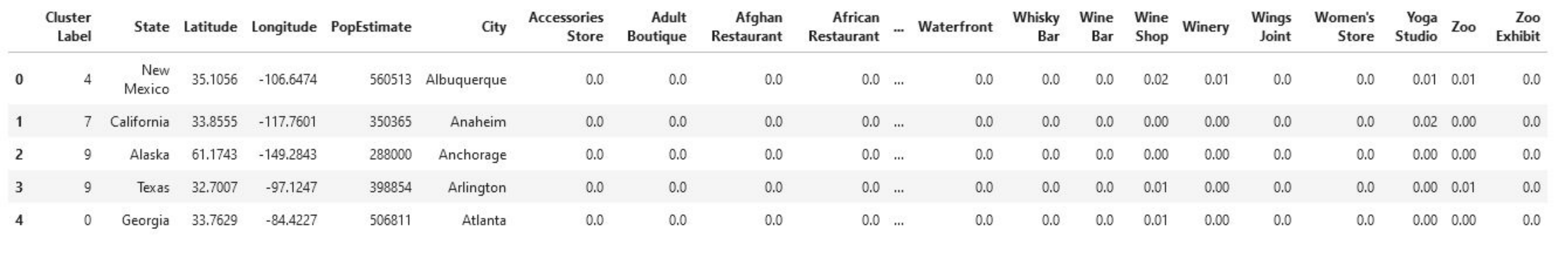
Map of Top 100 Cities

- A map was made with the folium package with circles marking each city in the top 100 dataframe



KMeans Clustering

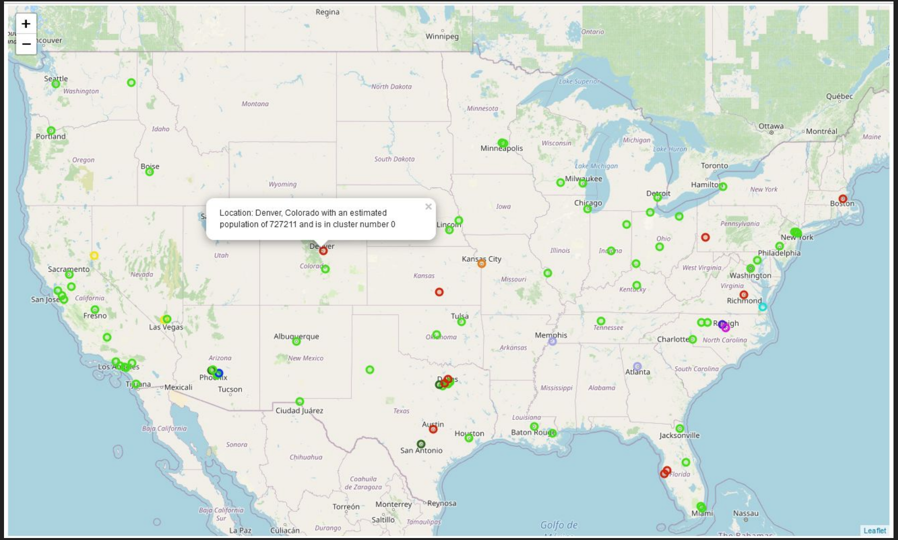
- Foursquare data of top venues and their types was acquired for each location
- KMeans clustering was applied to the venues
- Each city was grouped so they could be mapped with their respective cluster



	Cluster Label	State	Latitude	Longitude	PopEstimate	City	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	...	Waterfront	Whisky Bar	Wine Bar	Wine Shop	Winery	Wings Joint	Women's Store	Yoga Studio	Zoo	Zoo Exhibit
0	4	New Mexico	35.1056	-106.6474	560513	Albuquerque	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.02	0.01	0.0	0.0	0.01	0.01	0.0
1	7	California	33.8555	-117.7601	350365	Anaheim	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.00	0.0	0.0	0.02	0.00	0.0
2	9	Alaska	61.1743	-149.2843	288000	Anchorage	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.00	0.0	0.0	0.00	0.00	0.0
3	9	Texas	32.7007	-97.1247	398854	Arlington	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.01	0.00	0.0	0.0	0.00	0.01	0.0
4	0	Georgia	33.7629	-84.4227	506811	Atlanta	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.01	0.00	0.0	0.0	0.00	0.00	0.0

Map of Top Cities in Clusters

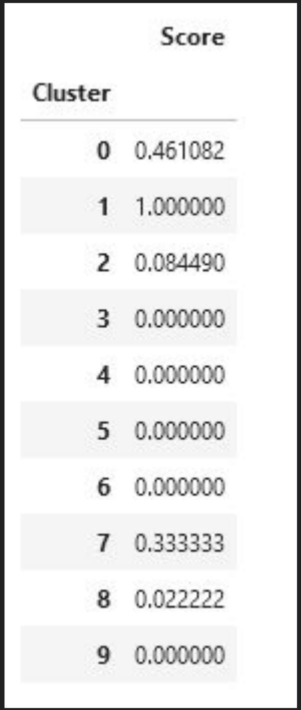
- A folium map was created with top cities
- Color of city correlates to which cluster that city was associated with
- A popup label with location, estimated population and cluster number was added to map for each city



Location: Denver, Colorado with an estimated population of 727211 and is in cluster number 0

Cluster Outdoor Score

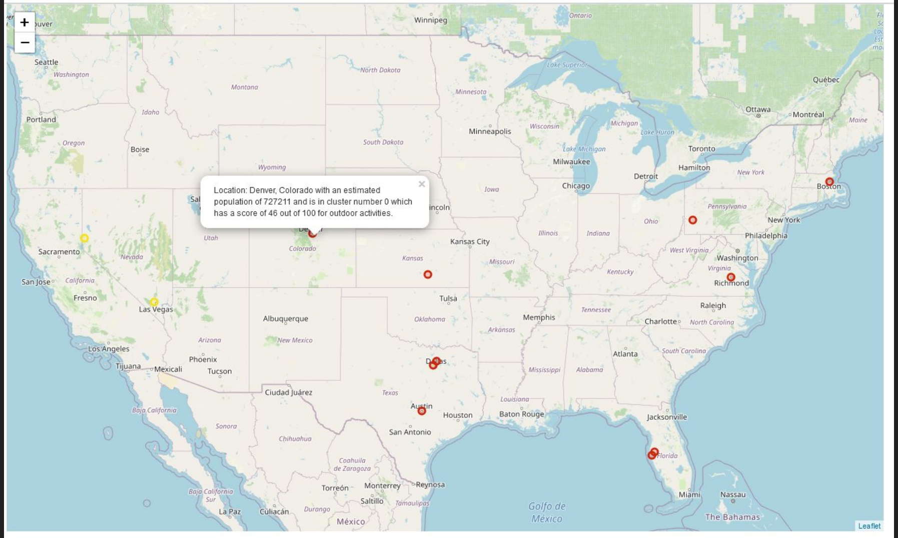
- A set of venue types were created that were deemed outdoor centered
- A score was found for each cluster based on that set
- As seen on the right, clusters 0 and 1 scored the highest according to outdoor activities



	Score
Cluster	
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

Map of Clusters 0 and 1

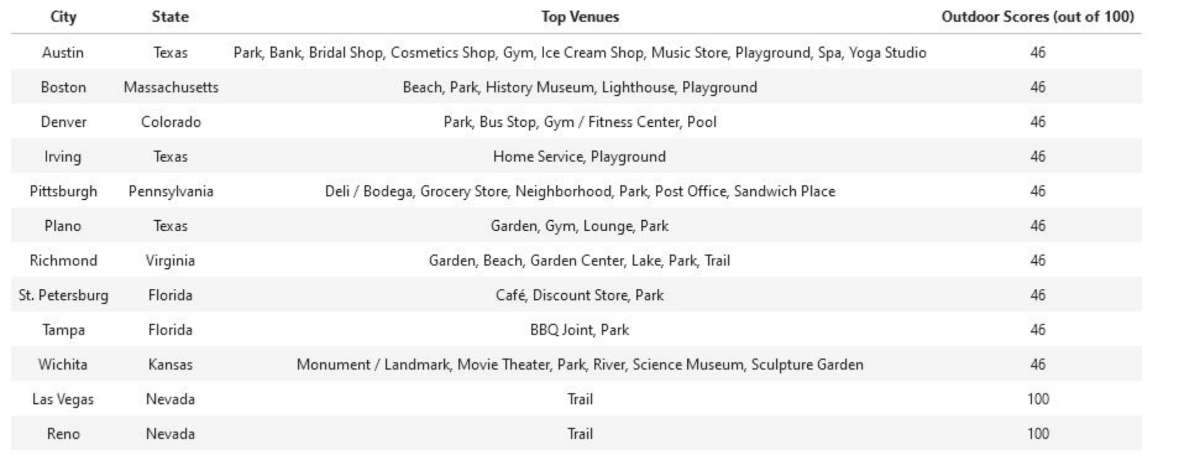
- A map of cities in clusters 0 and 1 was created
- A popup label with the same information as previous popup along with the score associated with the cluster that city was grouped in



Location: Denver, Colorado with an estimated population of 727211 and is in cluster number 0 which has a score of 46 out of 100 for outdoor activities.

Summary Dataframe

- A final dataframe was created to encompass all relevant data for the customer
- Each city and state from clusters 0 and 1 were listed along with their individual top venues and their outdoor scores of their respective cluster



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

Conclusion

- The final dataframe with all relevant data along with the final map of cities will be presented to the customer so he can decide which city he wants to travel to
- Ideal recommendation would be to travel to Richmond Virginia because all of their top venues are outdoor related and there is a wide variety of activities to choose from