Comparing the Top 200 Most Populous Cities in the United States

Daniel Wilcox

Knowing similar cities is valuable

- When people move to a new city, they may want to consider moving to a city that is similar to the one in which they currently live

- Business owners who are looking to open an additional business in a new city may want to look for cities that have similar needs to those of their current city

- The most common types of venues in a city may be good indicators of similarities between cities

Data acquisition and cleaning

- City data pulled from https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

- Location data retrieved using the geopy python module

- 200 cities used, data cleaned of excess information

- Venue information was retrieved using the Foursquare API

- 15,750 different venues used

- Transformed venue categories into categorial variables using one-hot encoding

Cities dataframe before and after cleaning

	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[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

	Rank	City	State	Latitude	Longitude
0	1	New York	New York	40.712728	-74.006015
1	2	Los Angeles	California	34.053691	-118.242767
2	3	Chicago	Illinois	41.875562	-87.624421
3	4	Houston	Texas	29.758938	-95.367697
4	5	Phoenix	Arizona	33.448437	-112.074142

Venue counts by city

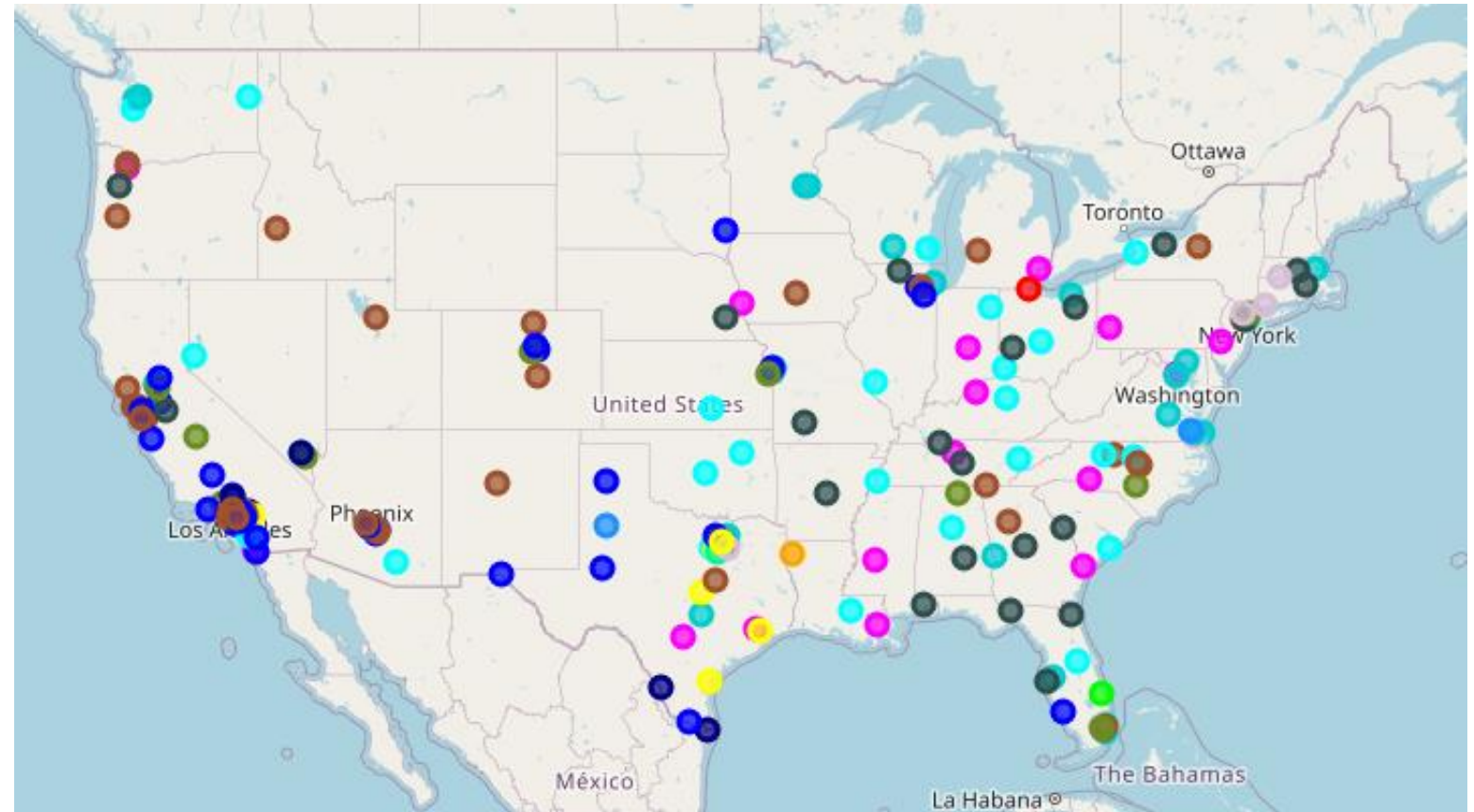| City	State	City Latitude	City Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
Akron	Ohio	91	91	91	91	91	91
Albuquerque	New Mexico	100	100	100	100	100	100
Alexandria	Virginia	100	100	100	100	100	100
Amarillo	Texas	35	35	35	35	35	35
Anaheim	California	64	64	64	64	64	64
...	...	...	...	...	...	...	...
Washington	District of Columbia	100	100	100	100	100	100
Wichita	Kansas	100	100	100	100	100	100
Winston–Salem	North Carolina	92	92	92	92	92	92
Worcester	Massachusetts	100	100	100	100	100	100
Yonkers	New York	56	56	56	56	56	56

Venue sorting

- The venues were grouped by city and then the mean values for each category were calculated to get an idea of the top venue types in each city

- The venue data were sorted to get the top ten most common venue categories for each city

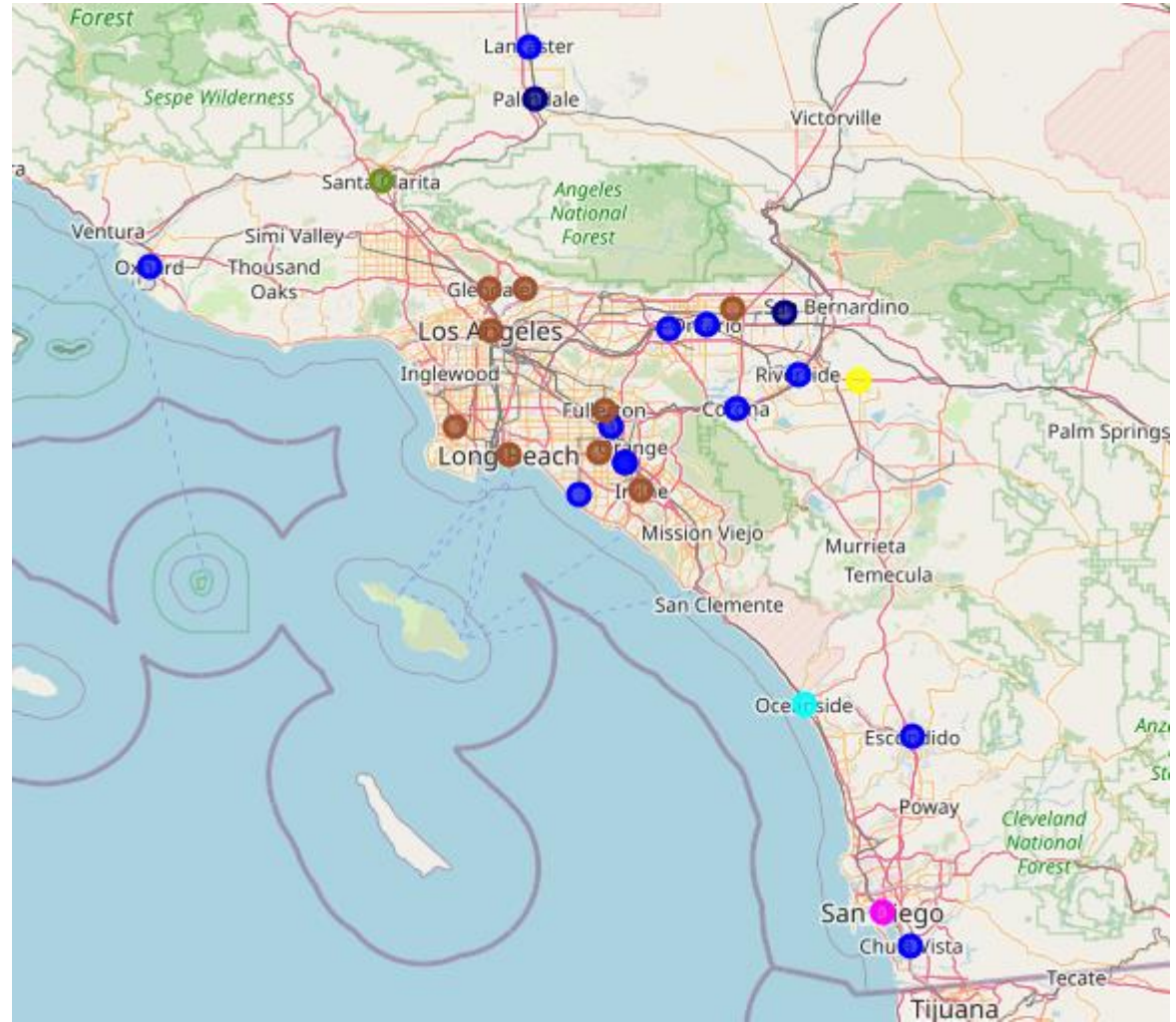# Grouped venue data for each city and top venue types

	City	State	ATM	Accessories Store	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Service	American Restaurant	...	Weight Loss Center	Whisky Bar	Wine Bar
0	Akron	Ohio	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.032967	...	0.0	0.010989	0.010989
1	Albuquerque	New Mexico	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.020000	...	0.0	0.000000	0.000000
2	Alexandria	Virginia	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.040000	...	0.0	0.000000	0.000000
3	Amarillo	Texas	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.057143	...	0.0	0.000000	0.028571
4	Anaheim	California	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.015625	...	0.0	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
194	Washington	District of Columbia	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.030000	...	0.0	0.000000	0.000000
195	Wichita	Kansas	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.050000	...	0.0	0.000000	0.000000

	City	State	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
0	Akron	Ohio	Sandwich Place	Bar	Rental Car Location	Italian Restaurant	Coffee Shop	Music Venue	American Restaurant	Bank	Café	Fast Food Restaurant
1	Albuquerque	New Mexico	Coffee Shop	Pizza Place	Sandwich Place	Brewery	Hotel	Restaurant	Mexican Restaurant	Bar	New American Restaurant	Diner
2	Alexandria	Virginia	Park	Coffee Shop	Pizza Place	American Restaurant	Grocery Store	New American Restaurant	Seafood Restaurant	Spa	Hotel	Gastropub
3	Amarillo	Texas	Mexican Restaurant	Restaurant	Sandwich Place	Café	American Restaurant	Sushi Restaurant	Convenience Store	Performing Arts Venue	Furniture / Home Store	Fast Food Restaurant
4	Anaheim	California	Mexican Restaurant	Convenience Store	Taco Place	Coffee Shop	Indian Restaurant	Burger Joint	Brewery	Ice Cream Shop	Bakery	Steakhouse

Map of the entire country, different colors represent distinct clusters

Zoomed-in map of Southern California

Count of cities in each cluster

Cluster Label	Count
0.0	3
1.0	21
2.0	11
3.0	40
4.0	5
5.0	1
6.0	22
7.0	1
8.0	7
9.0	28
10.0	1
11.0	1
12.0	35
13.0	6
14.0	17

Cluster 12 example

- Even though the clusters were formed based off of most common venue types, some of the clusters share geographic similarities

- Cluster 12 is largely focused in the southwestern portion of the United States

- The top three states represented in cluster 12 are California, Arizona, and Texas

Cities in cluster 12

	City	State			
87	Glendale	Arizona	51	Bakersfield	California
34	Mesa	Arizona	155	Corona	California
146	Peoria	Arizona	159	Hayward	California
80	Chandler	Arizona	189	Thornton	Colorado
162	Lancaster	California	53	Aurora	Colorado
136	Ontario	California	131	Cape Coral	Florida
163	Salinas	California	123	Aurora	Illinois
119	Huntington Beach	California	176	Joliet	Illinois
170	Pomona	California	168	Kansas City	Kansas
109	Oxnard	California	138	Sioux Falls	South Dakota
172	Escondido	California	117	Amarillo	Texas
197	Orange	California	93	Garland	Texas
74	Chula Vista	California	178	Midland	Texas
188	Roseville	California	184	McAllen	Texas
61	Stockton	California	187	Denton	Texas
57	Riverside	California	69	Plano	Texas
56	Santa Ana	California	21	El Paso	Texas
54	Anaheim	California			

Most common venue types in cluster 12

- Mexican restaurants are among the most common types of venues in cluster 12

- This makes sense, as the states in the southwestern United States are close to the border with Mexico and have significant immigrant populations

```
First most common venue types: ['Mexican Restaurant', 'Fast Food Restaurant', 'Burger Joint']
Second most common venue types: ['Convenience Store', 'Mexican Restaurant', 'Coffee Shop']
Third most common venue types: ['Fast Food Restaurant', 'Pizza Place', 'Sandwich Place']
```

Conclusion and future directions

- Meaningful clusters were created that group cities of similar venue types

- The model could be refined by incorporating smaller cities, as this model is highly biased towards large cities

- In this project, Foursquare failed to retrieve data for San Bernardino, California. This may be due to location errors so in the future add in location verification steps to ensure that no cities are missed