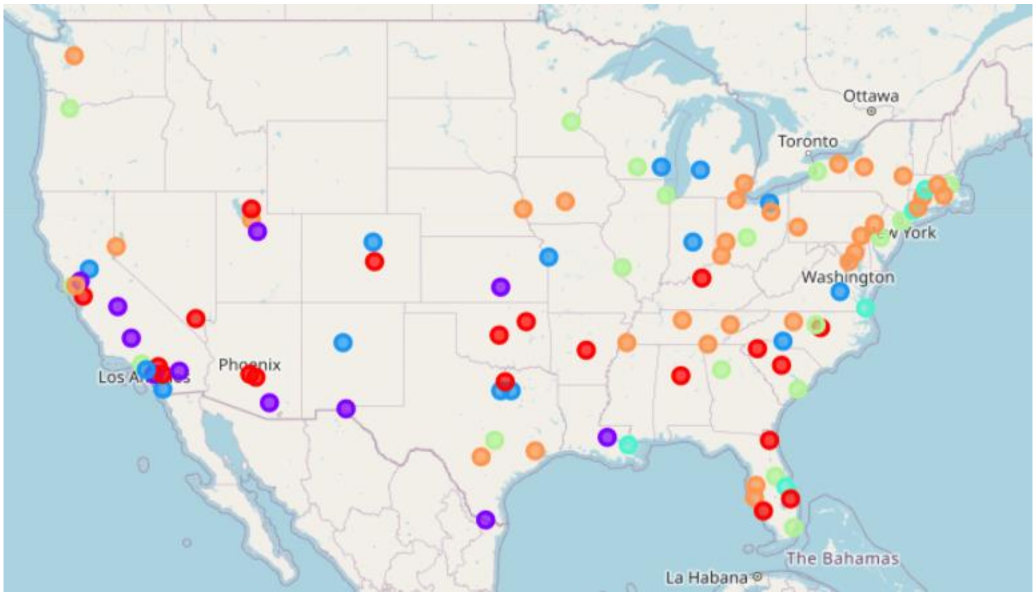

Searching for Similar Cities

Tom Peters

Contact Information Removed

05/07/2020

1. INTRODUCTION

Background

Imagine an individual is working for a large corporation with offices in cities across the United States and is suddenly asked to relocate. It could be for a number of reasons: to help found a new office to re-allocate human resources for efficiency purposes, or another... Their supervisor provides them with a list of cities to choose from to move to according to the individual's preferences.

Or, similarly, imagine another individual has lived and worked in a city for all or the majority of their life but has decided that it is time to make a change to a different city.

In both of these scenarios, what if the individual desired a new city that offers many of the same attractions, sites, foods, etc. that they are familiar with. They seek a new city but at the same time desire a similar scene to their current city in terms of venues and offerings. They want different but not too different. Identifying similar cities can help an individual strike a comfortable balance between new and old.

Problem

How do you identify cities that are characteristically similar?

This project will venture to answer this question via *k*-mean clustering of various cities in the United States by similarity in terms of top venues present in each city.

Interest

As asserted in the background of this project, stakeholders of interest are those who are trying to determine which city to move to. More broadly, stakeholders are those who are interested in evaluating the similarity of different cities across the United States in general.

2. DATA

Acquisition and Cleaning

To complete this project, the following data for every city in the United States was used: city name, state name, longitude, latitude, population, and land area. This data was obtained from Kaggle.com in two separate datasets in csv format: “World City Database”—which contains latitude, longitude, and population data for every city in the world—and “US City Population Densities”—which contains population density, population, and land area data for every city in the United States. Merging these two datasets on US cities yielded the necessary dataset of information.

An example entry in said merged dataset is as follows (Figure 1).

City	State	Area (sq mi)	City Radius (km)	lat	long	population
New York	New York	303	15.80	40.6943	-73.9249	19354922
Chicago	Illinois	228	13.71	41.8373	-87.6862	8675982

Figure 1. Example city data entry

An overview of the merged US city dataset can be found in Figure 2.

	City	lat	Long	State	pop density (people/sq mi)	Land Area (sq mi)	City Radius (km)	pop
count	100	100.00	100.00	100	100.00	100.00	100.00	100.00
unique	100	NaN	NaN	37	NaN	NaN	NaN	NaN
top	Houston	NaN	NaN	California	NaN	NaN	NaN	NaN
freq	1	NaN	NaN	14	NaN	NaN	NaN	NaN
mean	NaN	36.87	-92.91	NaN	4660.24	147.91	9.97	1817325.21
std	NaN	4.85	16.13	NaN	3903.29	144.84	4.78	2616741.97
min	NaN	25.78	-122.65	NaN	1016.00	7.00	2.40	404525.00
25%	NaN	33.60	-106.48	NaN	2388.50	49.50	6.39	572260.00
50%	NaN	37.42	-87.24	NaN	3484.50	104.50	9.28	894459.50
75%	NaN	40.85	-80.36	NaN	5446.50	179.75	12.17	1830098.75
max	NaN	47.62	-71.08	NaN	28211.00	747.00	24.82	19354922.00

Figure 2. Overview of merge US city dataset

Next, venue data for each city was acquired. Using Foursquare API in combination with the latitudes, longitudes, and radius information of every city in the above-described dataset, each city was explored to yield its top 100 venues, their respective locations, and their associated category in json format. This information was cleaned and structured into a *pandas* dataframe for easier processing and manipulation.

Note: Because it is difficult to see a city when all 7000+ cities in the United States are included on a single map, the list of cities to be grouped was narrowed to only the top 100 cities by population size. Another reason for this limit was that having too many cities maxes out the daily requests available to an individual using Foursquare API with a free account. Similarly, the top 100 venues were chosen because it was the max number of venues one can request using a free account.

An example entry in the resulting dataframe of city and venue data is as follows (Figure 3).

City	City Lat	City Long	Venue	Venue Lat	Venue Long	Venue Category
New York	40.6943	-73.9249	Carmenta's	40.70132	-73.92678	Italian Restaurant
New York	40.6943	-73.9249	Henry's Wine and Spirit	40.70105	-73.93025	Wine Shop

Figure 3. Example venue data entry

The resulting dataframe contained 10,000 rows—100 rows of top venues for each of the 100 cities. The venue category column included 405 unique venue categories. The venue category column was expanded into 405 dummy variable columns of 0 or 1 (binary indicators). The rows were then grouped by city, and the mean of the category columns was found to yield the frequency of each venue category among the top 100 venue categories in each city.

Each city was then summarized by its top 10 most common venue categories (see Figure 4).

	City	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	Mexican Restaurant	Park	American Restaurant	Bar	Italian Restaurant	Diner	Brewery	Ice Cream Shop	Burger Joint	Coffee Shop
1	Albany	Coffee Shop	American Restaurant	Café	Italian Restaurant	Ice Cream Shop	Bar	Sandwich Place	New American Restaurant	Pizza Place	Deli / Bodega
2	Albuquerque	Brewery	Coffee Shop	Grocery Store	Mexican Restaurant	Café	Pizza Place	Park	Restaurant	American Restaurant	Science Museum
3	Allentown	Park	Pizza Place	Diner	American Restaurant	Donut Shop	Bar	Italian Restaurant	Pub	Mexican Restaurant	Farmers Market
4	Atlanta	Trail	Park	Mexican Restaurant	Pizza Place	Brewery	Wine Shop	Ice Cream Shop	Music Venue	Mediterranean Restaurant	Italian Restaurant

Figure 4. Top 10 most common venues

3. METHODOLOGY

The goal of this project was to cluster (or group) US cities together to answer the question of which cities have similar characteristics. More specifically, similarity was assessed using frequency of venue category among each city's top 100 venues. That is to say, two cities with top 100 venues of similar category distribution are considered more similar than two cities with top 100 venues of dissimilar category distribution.

Using the Folium map package, the top 100 cities were plotted in order to see the map prior to all the cities being grouped (Figure 5).

Figure 3. Map of cities before clustering

Because the goal of this project was to group similar cities, *k*-means clustering machine learning methodology was applied. *k*-means clustering is a machine learning algorithm that aims to partition a given set of observations into *k* clusters. In this project, the frequency of each venue category in each city served as features or independent variables. The number of clusters (*k*) was determined using the Elbow method. For various values of *k*, the sum of square distances of each data point from its respective cluster center and the train time of each cluster model was evaluated and compared. Using the Elbow method, the optimal number of clusters *k* was found to be 6 (see Figure 6).

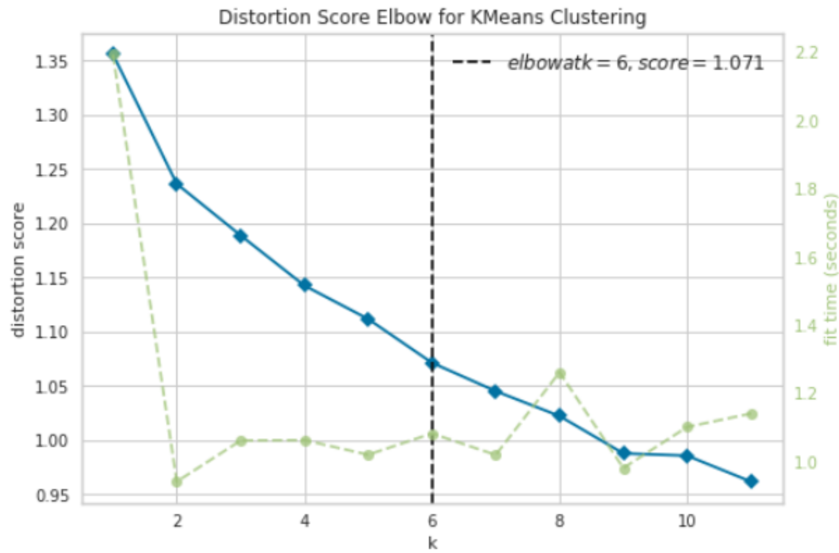

Figure 6. Elbow Method for determine *k*

4. RESULTS

Applying *k*-means clustering on the venue category data, the 100 US cities were grouped into six different clusters. Figure 7 shows distribution of cities among cluster groups across the US.

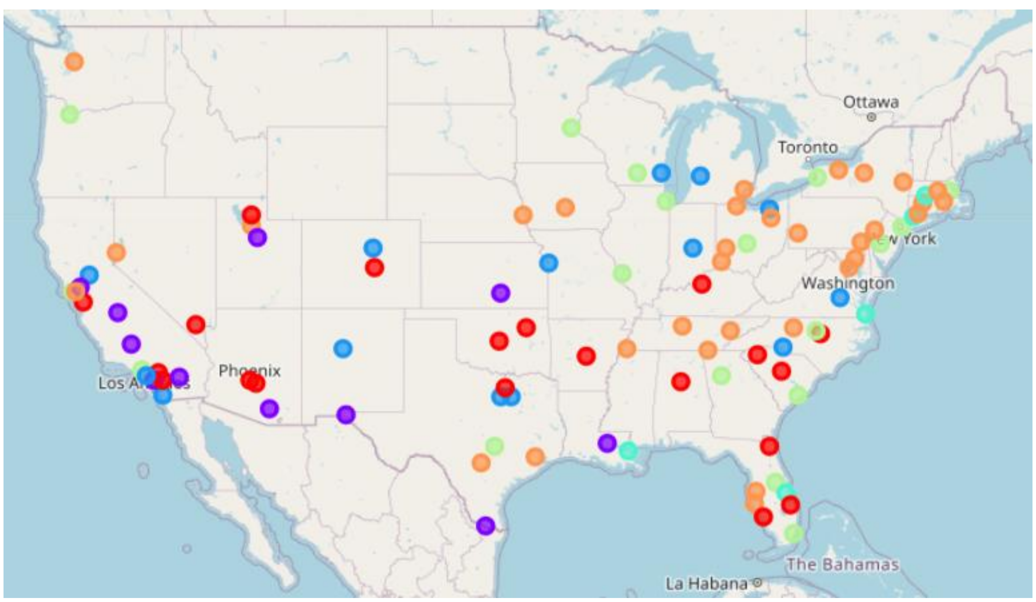

Figure 7. Folium Map of Clusters

Figure 8 provides a more detailed list of the cities assigned to each cluster. Cluster 0 contained 20 of the 100, Cluster 1 contained 11, Cluster 2 contained 14, Cluster 3 contained 5, Cluster 4 contained 18, and Cluster 5 contained 32 cities.

	CLUSTER					
	0	1	2	3	4	5
0	Phoenix, Arizona	Tucson, Arizona	Dallas, Texas	Virginia Beach, Virginia	New York, New York	Houston, Texas
1	Riverside, California	El Paso, Texas	San Diego, California	New Orleans, Louisiana	Los Angeles, California	Washington, District of Columbia
2	Las Vegas, Nevada	McAllen, Texas	Denver, Colorado	Bridgeport, Connecticut	Chicago, Illinois	Seattle, Washington
3	San Jose, California	Fresno, California	Sacramento, California	Springfield, Massachusetts	Miami, Florida	Detroit, Michigan
4	Jacksonville, Florida	Concord, California	Cleveland, Ohio	Palm Bay, Florida	Philadelphia, Pennsylvania	Tampa, Florida
5	Raleigh, North Carolina	Mission Viejo, California	Kansas City, Missouri		Atlanta, Georgia	Baltimore, Maryland
6	Louisville, Kentucky	Baton Rouge, Louisiana	Indianapolis, Indiana		Boston, Massachusetts	San Antonio, Texas
7	Oklahoma City, Oklahoma	Bakersfield, California	Charlotte, North Carolina		San Francisco, California	Pittsburgh, Pennsylvania
8	Birmingham, Alabama	Provo, Utah	Milwaukee, Wisconsin		Minneapolis, Minnesota	Cincinnati, Ohio
9	Tulsa, Oklahoma	Wichita, Kansas	Richmond, Virginia		St. Louis, Missouri	Providence, Rhode Island
10	Cape Coral, Florida	Indio, California	Fort Worth, Texas		Portland, Oregon	Salt Lake City, Utah
11	Colorado Springs, Colorado		Albuquerque, New Mexico		Orlando, Florida	Nashville, Tennessee
12	Ogden, Utah		Grand Rapids, Michigan		Austin, Texas	Memphis, Tennessee
13	Columbia, South Carolina		Long Beach, California		Columbus, Ohio	Hartford, Connecticut
14	Mesa, Arizona				Buffalo, New York	Omaha, Nebraska
15	Murrieta, California				Charleston, South Carolina	Dayton, Ohio
16	Greenville, South Carolina				Madison, Wisconsin	Rochester, New York
17	Little Rock, Arkansas				Durham, North Carolina	Sarasota, Florida
18	Denton, Texas					Allentown, Pennsylvania
19	Port St. Lucie, Florida					Albany, New York
20						Knoxville, Tennessee
21						New Haven, Connecticut
22						Akron, Ohio
23						Worcester, Massachusetts
24						Toledo, Ohio
25						Des Moines, Iowa
26						Reno, Nevada
27						Oakland, California
28						Winston-Salem, North Carolina
29						Syracuse, New York
30						Chattanooga, Tennessee
31						Lancaster, Pennsylvania

Figure 8. List of cities in each cluster

5. DISCUSSION

Using the above table of clusters and their respective cities, an individual can now quickly determine which cities are similar to their own. For example, *k*-means clustering seems to indicate that, in terms of top 100 venue category frequency, a city such as Nashville, Tennessee, is most similar to Houston, Seattle, Detroit, and other cities in Cluster 5. Alternatively, a city such as Virginia Beach, Virginia, is most similar to New Orleans and other cities in Cluster 3. An individual wanting to see what cities are most similar to their own can find their city on the table and see what other cities are included in its cluster.

In order to examine each cluster and determine the discriminating venue categories that distinguish each cluster, the top 10 most common venue categories in each cluster were found and plotted using horizontal bar graphs (See Figures 9-14).

Figure 9: Cluster 0 Venue Breakdown

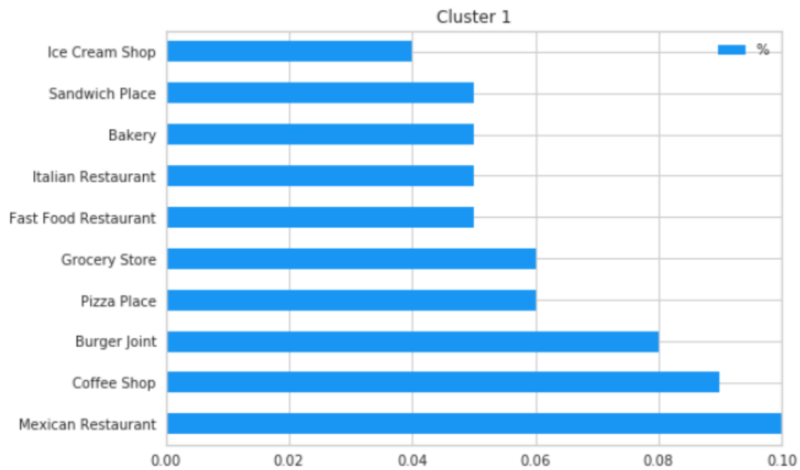

Figure 10: Cluster 1 Venue Breakdown

Figure 11: Cluster 2 Venue Breakdown

Figure 12: Cluster 3 Venue Breakdown

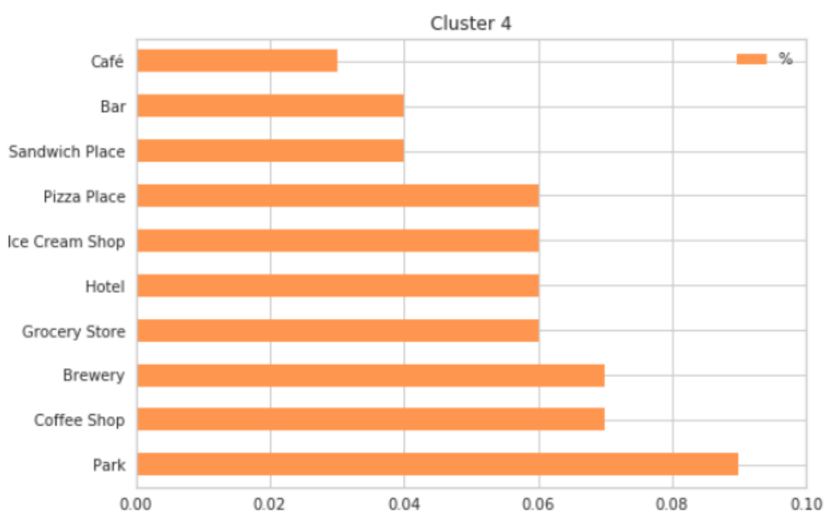

Figure 13: Cluster 4 Venue Breakdown

Figure 14: Cluster 5 Venue Breakdown

Generally speaking, looking at the bar graphs of the top 10 venues in each cluster, all seem to include venue categories relating to coffee, restaurants, and parks. That being said, there are a few categories that seem to be more prominent (proportion of top 10 $\geq$ 0.08) in certain clusters over others.

- In Cluster 0, coffee, grocery stores, and Mexican restaurants are most prominent.
- In Cluster 1, Mexican restaurants, coffee, and burger places are most prominent.
- In Cluster 2, breweries, parks, coffee, and Mexican restaurants are most prominent.
- In Cluster 3, pizza places and pharmacies are most prominent.
- In Cluster 4, parks are most prominent.
- In Cluster 5, coffee shops are most prominent.

Although it might seem as if it is difficult to distinguish one cluster from another using the bar charts of the top 10 venue categories in each cluster, it is important to remember that fewer than 30 of the over 400 venue categories are present in the bar charts above. Consequently, it is likely that many key differences lie beyond the top 10.

6. CONCLUSION

This project serves as a proof of concept for how to compare cities and assess which cities are most similar in the United States using *k*-means clustering and a plethora of features.

The key to finding good clusters is having the data necessary to distinguish one group from another. It is important to note that the application of this project's results to determining city similarity is limited by the fact that only the top 100 venues in each city were included in the analysis. A more useful similarity analysis would be possible with more venue information from each city. Likewise, future projects would benefit from including additional city demographic data, such as racial and ethnicity distributions, population density data, weather data, economic data, etc...

The most important steps to building a better model are to define thoroughly what constitutes being *similar* (i.e. what features or characteristics) and to acquire appropriate data for representing said features. Then, it is a matter of applying an appropriate clustering technique and analyzing the results.

Data Resources

1. https://simplemaps.com/data/us-cities
2. https://www.kaggle.com/max-mind/world-cities-database