Searching for Similar Cities

Tom Peters
05/07/2020

Identifying similar cities is important in a variety of circumstances

There are a variety of reasons for which an individual might be interested in assessing what other US cities are most similar to their own . . .

- Job-related relocation
- Disaster-related relocation
- A desire for a change of pace
- Others . . .

In these situations, an individual might be seeking to find 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.

Data Acquisition and Cleaning

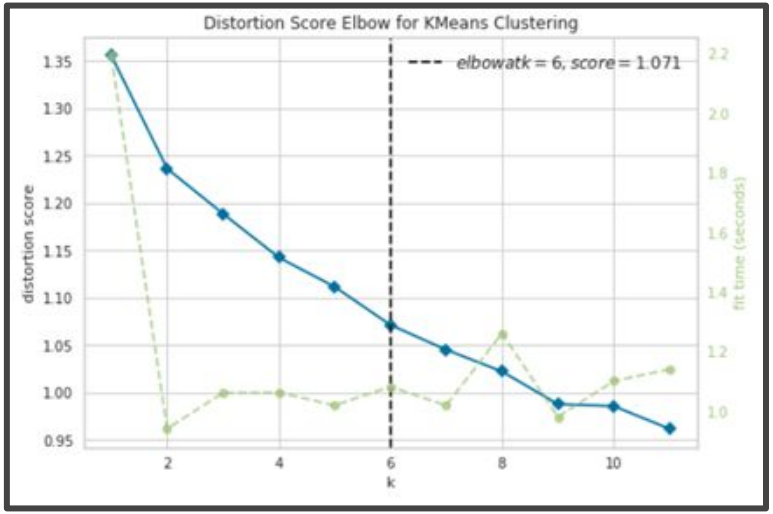
- US city data--name, state, longitude, latitude, population, land area, population density--in CSV format
  - Merged from two sources (1 and 2)
  - Top 100 cities in US by population were used in this project
- Used city location data to request a json file of the top 100 venues in each city using Foursquare API
  - 10,000 rows—100 rows of top venues for each of the 100 cities.
  - 405 unique venue categories.
  - The venue category column was expanded into 405 dummy variable columns of 0 or 1
  - The rows were then grouped and averaged by city to yield the frequency of each venue category among the top 100 venue categories in each city.

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

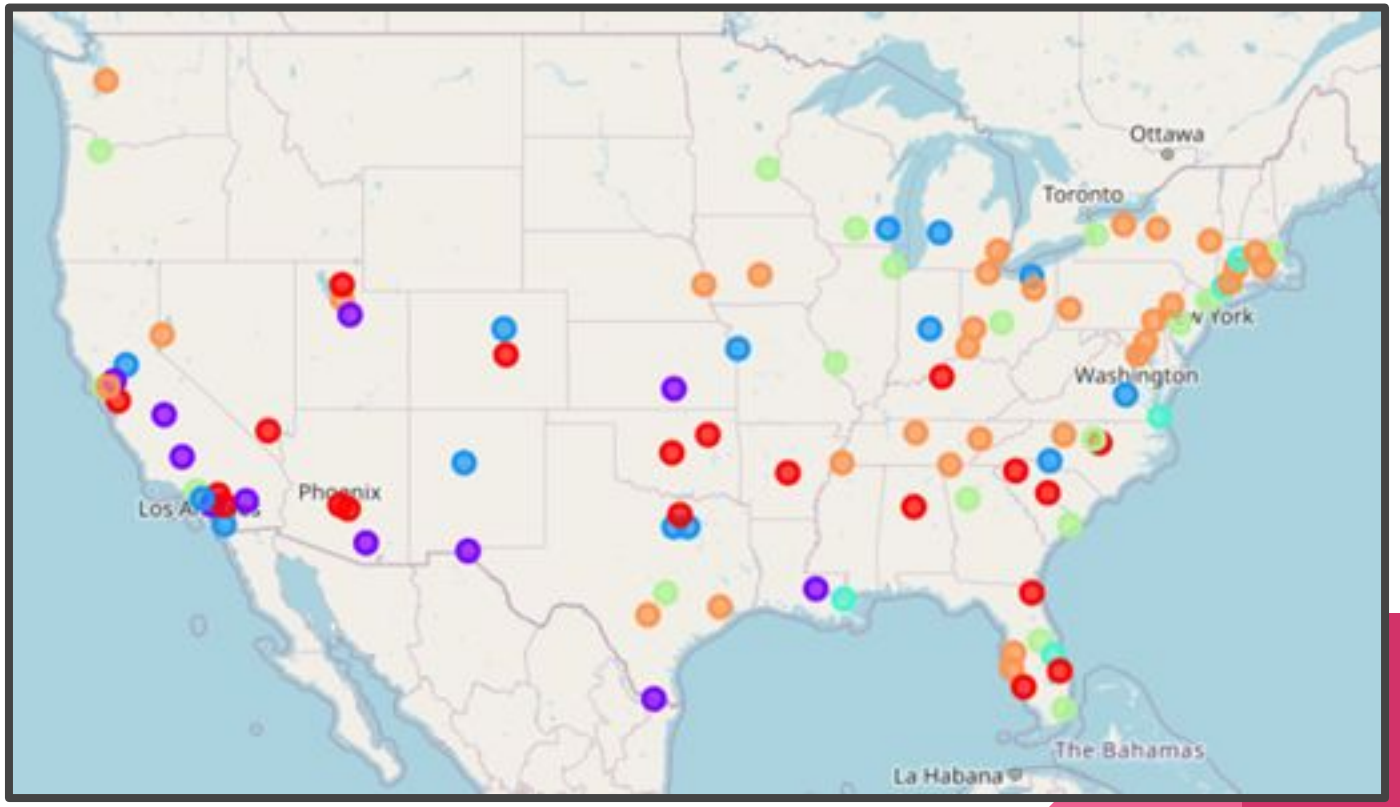
Map of cities prior to clustering/grouping

Appling *k*-means clustering technique

- Because the goal of this project was to group similar cities, *k*-means clustering machine learning methodology was applied.
- In this project, the frequency of each venue category in each city's top 100 venue list served as features or independent variables for determining groups
- The optimal *k* to be 6 was determined using the Elbow method



Results: Map with clusters



Results: List of cities by cluster

	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

Use the list to identify similar cities

- Example:
  - Nashville, Tennessee, is most similar to Houston, Seattle, Detroit, and other cities in Cluster 5.
  - 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.

Conclusions and future directions

- This project does a good job of demonstrating how one could identify similar cities, but its accuracy and applicability is limited by the narrow scope of the feature data.
- A more useful similarity analysis would be possible with . . .
  - More venue information from each city.
  - Additional city demographic data, such as racial and ethnicity distributions, population density data, weather data, economic data, etc...
- To build a better model . . .
  - Define thoroughly what constitutes being *similar* (i.e. what features or characteristics)
  - Acquire appropriate data for representing said features.
  - Apply appropriate clustering model