The Battle of Neighborhoods - Search for Similar Cities for Graduating University Students

Dexin Wang for Capstone Project Assignment
Feb 28, 2020

Introduction

A survey shows some graduating students from a New York university are searching for information to support their decision of coming relocation. Assuming that the US job market is prospering, the students want to find US cities or city neighborhoods much like New York where they enjoy the nearby venues or amenities. Though there may be many different criteria, numbers and variety of the venues as well as population density are among the most important factors to support the relocation decision. Now the question is, which cities or city neighborhoods are similar to New York the students could choose from? The students prefer to have multiple choices to leverage with job opportunities and other considerations.

Project Objective

To help the students make decision of relocation, the project objective is to collect and analyze city or neighborhood venue data, from which draw insights about which cities are more similar to the current city they live in. In addition to the lists of similar cities, the most popular venues are also presented for further information along with grouped city visualization.

Analytic Approach

Based on the needs of the customers– in this case the students, it is appropriate to find city neighborhood similarity by clustering the similar cities as groups.

Data Description

Apparently, we need data of as many cities as possible across the country. So we started with the 200 largest cities by population in 2020 at http://worldpopulationreview.com/us-cities/#cities. We can read a table from the website, which has city names, states, population, population density, location coordinates, etc. (Figure 1). For the neighborhood venues, we use the data from the Foursquare https://foursquare.com by exploring nearby venues defined by a radius of the given city neighborhoods (Figure 2). Let us assume that the city neighborhood centered at the location coordinates is the neighborhood the students might be interested to relocate to. In fact, there are multiple city neighborhoods included in the city list for some metropolitan areas.

Here is the table read from the http://worldpopulationreview.com/us-cities:

	Rank	Name	State	2020 Population	2010 Census	Change	2020 Density	Latitude/Longitude	Area (km²)
0	1	New York	New York	8,622,357	8,175,133	0.25%	11,084/km²	40.66/-73.94	778
1	2	Los Angeles	California	4,085,014	3,792,621	0.67%	3,365/km²	34.02/-118.41	1,214
2	3	Chicago	Illinois	2,670,406	2,695,598	-0.32%	4,535/km²	41.84/-87.68	589
3	4	Houston	Texas	2,378,146	2,099,451	0.79%	1,443/km²	29.79/-95.39	1,649
4	5	Phoenix	Arizona	1,743,469	1,445,632	1.88%	1,300/km²	33.57/-112.09	1,341

Figure 1 - Illustration of City Information

In summary, our data is composed of city population densities and nearby venues around the locations for 200 largest US cities. The data has the information as the most important factors mentioned earlier to support the students' relocation decision.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	New York	40.789624	-73.959894	North Meadow	40.792027	-73.959853	Park
1	New York	40.789624	-73.959894	Central Park Tennis Center	40.789313	-73.961862	Tennis Court
2	New York	40.789624	-73.959894	East Meadow	40.790160	-73.955498	Field
3	New York	40.789624	-73.959894	Central Park - Woodman's Gate	40.787786	-73.955924	Park
4	New York	40.789624	-73.959894	The Jewish Museum	40.785276	-73.957411	Museum

Figure 2 - Illustration of City Venues

Method - Feature Extraction

There are total 12981 venues found from the Four Square. Venues from each city neighborhood are categorized and normalized as features for sequent city clustering. On average, there are about 68 venues for each city neighborhood. There are 434 unique categories (Figure 3). The city population density is also normalized as an additional feature.

	City	ATM	Accessories Store	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Service	Airport Terminal	American Restaurant	Antique Shop
0	Akron	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.031250	0.0
1	Albuquerque	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.041667	0.0
2	Alexandria	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
3	Amarillo	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.035714	0.0
4	Anaheim	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.034483	0.0

Figure 3 - Illustration of City Venue Categories

Method - Clustering

K-means clustering is used for grouping the cities based on the selected features. The number of clusters is selected based on Silhouette score and Dendrogram (Figure 4). Eight clusters are selected to have fine granularity though two or three clusters have higher Silhouette scores. Agglomerative clustering is used to cross-check the K-means clustering results. There are 121 city neighborhoods in the same clusters as New York from K-means clustering. Cluster match percentage between K-means & agglomerative Clustering is about 98%.

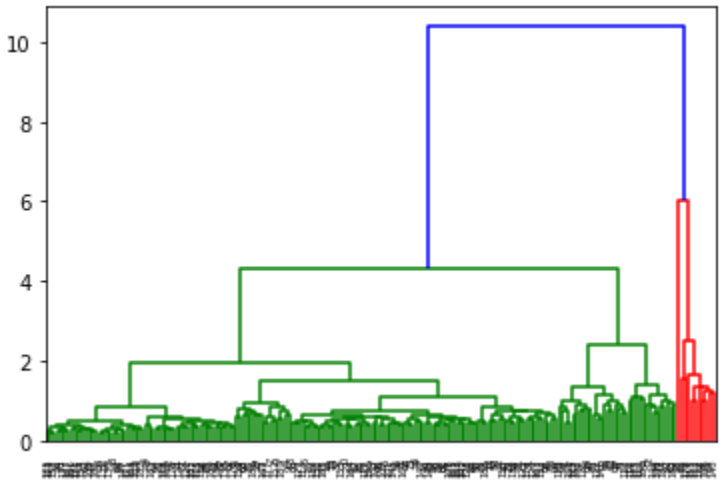

Figure 4 - Dendrogram of City Clustering

Results and Discussion

It is a surprising fact that most of the cities in the same cluster as New York (shown as green dots in Figure 5). The cosine similarity is also calculated and the highest similarity is Chicago. Comparing the two cities in this cluster in www.areavibes.com, it shows both are rated with A+ for amenities. Since there are 435 features or venue categories, it is not intuitive to explain the similarity of the cities in the same cluster by the most common venues (Figure 6).

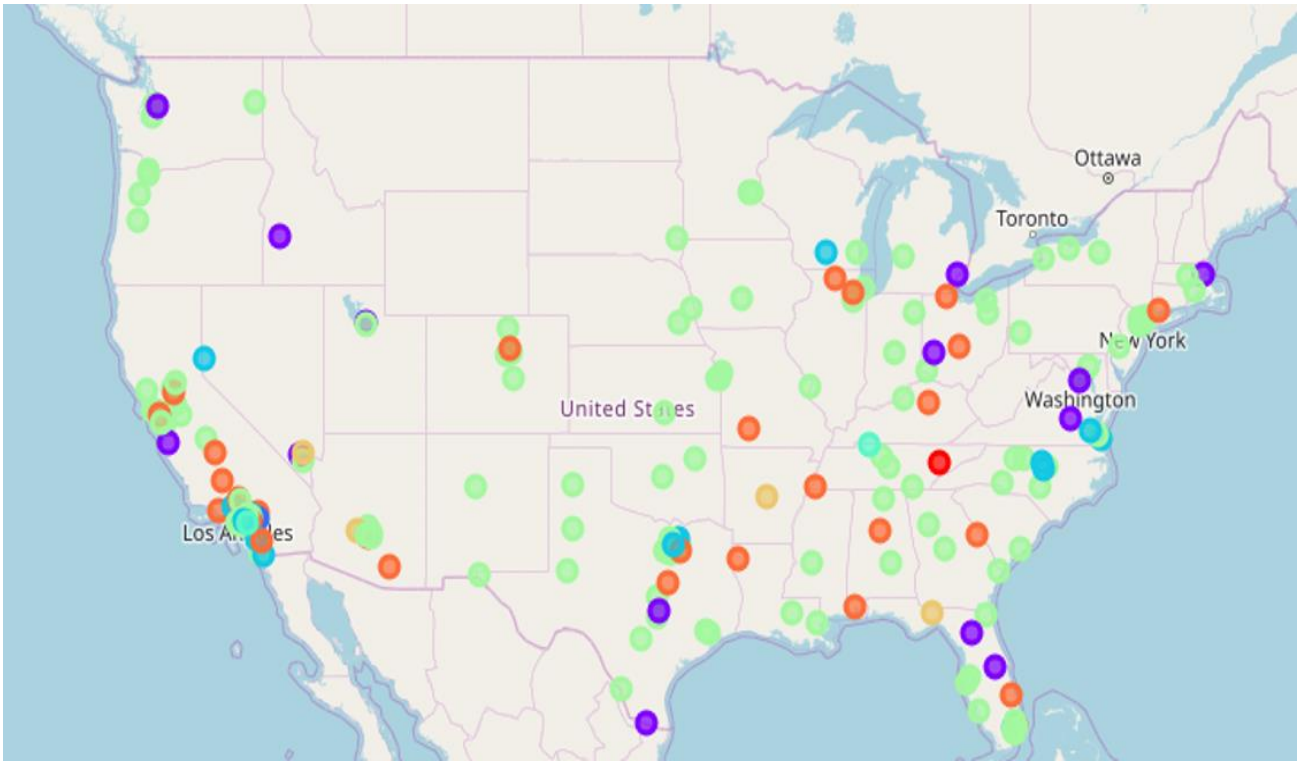

Figure 5 - Cities Clustered as Eight Groups with Different Colors

City	2020 Density	Latitude	Longitude	Agg Cluster Labels	Cluster Labels	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
New York	11084	40.789624	-73.959894	0.0	5.0	Park	Grocery Store	Café	Pizza Place	Pharmacy	Deli / Bodega	Playground	Bakery
Los Angeles	3365	34.053691	-118.242767	0.0	5.0	Coffee Shop	Sushi Restaurant	Mexican Restaurant	Bakery	Indian Restaurant	Thai Restaurant	Japanese Restaurant	Ice Cream Shop
Chicago	4535	41.875562	-87.624421	0.0	5.0	Italian Restaurant	Coffee Shop	Hotel	Grocery Store	Pizza Place	Garden	Theater	Gym / Fitness Center
Houston	1443	29.758938	-95.367697	0.0	5.0	Hotel	Mexican Restaurant	Coffee Shop	Bar	Park	Burger Joint	American Restaurant	Southern / Soul Food Restaurant
Phoenix	1300	33.448437	-112.074142	0.0	5.0	Coffee Shop	Pizza Place	Hotel	Mexican Restaurant	American Restaurant	Fast Food Restaurant	Bar	Music Venue

Figure 6 - Top 8 Most Common Venues of Selected Cities

Conclusion

It is found that there are over 120 big city neighborhoods are similar to New York City in terms of venues. Chicago has the relatively higher similarity score to the city. The features used in this study are unfortunately very limited due to number of venues could be searched in each neighborhood. One could further refine the cluster similar to New York if more venue data is available. Also, amenities are just one aspect of likability about cities. More data means more insights could be drawn. However it is good news from this study there are many similar cities to choose from.