Comparison of Cities in the Midwestern United States

Chris Butz

Introduction

The Midwestern United States is a geographic area located West of the Atlantic Coastline states, East of the Rocky Mountains, and north of the Mason-Dixon line. A growing trend in young adults is to move to a new city full of opportunity, fun, and often, like-minded individuals. Yet, young adults show hesitance flying too far from the roost. As such, they look for areas that are new but not too far (geographically) from family and friends. For many this may mean a move to Chicago, Indianapolis, or Kansas City, from their rural hometown.

This project aims to make this move easier on younger generations in the Midwestern United States. Data analysis of metropolitan areas around the Midwestern United States will identify similarities to behoove young adults embarking on this venture for independence.

Data Acquisition and Cleaning

Data was pulled from three sources for this analysis. Initially, city name, state, and population data was sourced from the Wikipedia page “List of Midwestern Cities by Size” (https://en.wikipedia.org/wiki/List_of_Midwestern_cities_by_size). Data from this table was mostly clean upon downloading. One change necessary was converting population counts from string values to integer values. The second data sources is OpenDataSoft.com, a French data-sharing software company which hosts a plethora of tables, reports, and maps for the general public. The table from OpenDataSoft consists of all US zip codes, the associated municipality, state, timezone, Daylight-Savings Time indicator, and GPS coordinates. Zip codes, timezone and Daylight-Savings Time indicator, were removed as they were not needed for analysis. A majority of instances were removed from the OpenDataSoft dataset when joining the two datasets, due to their irrelevance of many cities to the task at hand. Finally, the FourSquare API utilized the geographic data provided from the other datasets to identify popular venues in each municipality.

Ultimately, basic features were utilized to maintain simplicity of the project. These features are City, State, Population Change from 2000 to 2013 (%), Population, Longitude, and Latitude. These features were chosen for two reasons. The largest of which was absence of null values. These features, when joined, created a data set without any missing values. The second was the scope of the project. Numerous features could be used to cluster cities by socioeconomic demographics such as political ideology, % of population in poverty, % of male/female/etc., and largest industry, to name a few. It is unknown whether utilizing these features would create an overfit model or if they would enhance the accuracy. Not enough data was available for these features, as well.

Exploratory Data Analysis

Exploration of data was limited due to the low number of features utilized.Using descriptive statistics, it is evident larger metropolitan areas are growing as a whole, though the largest metropolitan areas are not growing as large as their surrounding municipalities. Cities like Lincoln Park, IL and Noblesville, IN, experienced large growth between the years 2000 and 2013. These up-and-coming cities also have a high frequency of venues popular amongst younger generations (coffee shops, asian restaurants, bars, and social venues). Large metropolitan areas in more rural areas are seeing much slower growth or population decline. States further into the midwest with fewer major metropolitan areas have seen a decline in population as a whole, even though larger metropolitan areas in those states experienced growth in population.

My hypothesis is that cities on the outskirts of major metropolitan areas will be clustered together as similar, while major metropolitan areas, and small rural cities will also be clustered together, respectively. Based on the scope of the project, k-means clustering was used to identify similar cities. K-means clustering is a method of vector quantization that partitions observations into K number of clusters, in which each observation belongs to the cluster with the nearest mean distance. For this project, K was set to 10 clusters. Initially, 4 clusters were created. This resulted in overlapping clusters with too large of radii. K was increased to 12 and resulted in overcorrection and a lack of distinct features for each cluster. Observations that should have fallen into one cluster were distributed into multiple. Even with 10 clusters, this effect was present, though at a much lower frequency.

Results & Discussion

Using K-Means, clusters were more or less equally distributed. There are a few instances, as mentioned below, that have few observations. Though on the whole, most clusters contained between 10 and 20 cities. When reviewing the 10 clusters, it is evident the algorithm grouped observations two ways. Some clusters were grouped more geographically, such as with clusters 2, 6 and 9. Most other clusters were driven by population and venues such as with clusters 4 and 8. It is interesting that the outskirt municipalities of the major metropolitan area in each state mostly fell into their own cluster, such as with cluster 4.

Clusters 1 and and 6 showed much similarity, both geographically and venue-wise. Both contained cities northwest of Chicago, extending as far as Minnesota. In both clusters, bars, pizza places, and american restaurants were the most prevalent venues. Entertainment and leisure (parks, performing arts, night clubs) were a strong secondary characteristic for cluster 1. I interpret this as the reason for splitting the cluster.

Cluster 3 was the largest cluster and comprised municipalities surrounding Chicago proper. It is not surprising the pizza place was overwhelmingly the most popular venue in this cluster. This cluster was grouped predominantly by venue and geography, as it has the largest range of populations of any cluster. Overall, this cluster saw the largest population growth. In geographic proximity, cluster 3 was close with clusters 1 and 6. If the model was revised to a lower value for K, it is expected this cluster would include or be included in one of these.

Cluster 9 was the last cluster resulting in a geographically-driven grouping. It consists largely of cities in the eastern Midwest (Indiana, Ohio, Michigan). As a cluster, it saw the largest overall population decline rate, with most cities experiencing small (5-10%) and moderate (15-25%) decline. Sandwich shops and fast food were atop the most popular venues in these cities.

Two clusters contained one or no observations. Interestingly, cluster 10 contained no cities. It is unclear why this occurred. In previous versions of the model, a K set to 9 and 11 both resulted in zero null clusters. Cluster 5 only contained one city, Gary, IN. Given Gary's size and inclusion in the Chicagoland area, it was expected to be included in clusters 1, 3 or 4. I suspect this is due to the differing venues and ratings of such venues in Gary, as the city has a much lower

standard of living than many suburban or outskirt municipalities surrounding Chicago, especially those to the north and west.

Given Gary's standard of living and socioeconomic status, cluster 5 would also fit well within cluster 2. The most geographically dispersed of the clusters, cluster 2, represented the old rust belt. Many of these cities were far from major metropolitan areas, had small decline and populations ranging from 40 to 75,000 people. The most popular venues are fast food and discount stores. Grocery stores fell much further down on this list. Given the lower-than-average socioeconomic status of these aging cities, it is reassuring to see them clustered together.

Cluster 4 represents the growing Midwest. Most all cities are upper-middle class first-ring suburbs of major metropolitan areas such as Kansas City, Indianapolis, Chicago, St. Louis, Milwaukee, and Minneapolis. Combining the proximity to metropolitan centers and higher standard of living, it is suspected many still commute into their respective cities for work. A majority of observations in Cluster 4 range between 50 and 75,000 people with a high mark of 100,000 citizens. Many saw large population growth (>25%), such as with Noblesville, IN experiencing 100%+ growth in the years 2000 to 2013. Cosmetics and clothing stores secured the top 2 highest-rated venues spots with American restaurants representing the most popular food option.

While similar in large population growth, and proximity to larger metropolitan areas, Cluster 7 differed from Cluster 4 in its venues. These cities are second-ring suburbs of major metropolitan areas, meaning they are further from major metropolitan areas than Cluster 4 cities. This is further explained by the most popular venues being grocery stores, delis, and Mexican restaurants, very different from the luxury stores most popular in Cluster 4. A few of the cities in Cluster 7 did not fit the mold were also standalone cities, not near major metropolitan areas. I noticed the standalone cities were all adjacent or straddled state lines. If political boundaries were removed, it would be interesting to find out the total populations of these divided cities.

Cluster 8 represents the proper sections of major metropolitan areas of the Midwest, including Chicago, Indianapolis, Minneapolis, Milwaukee, St. Louis, and Kansas City, among others. Aside from a few outliers on either side, the average population change was stagnant, hovering

right above 0% growth. The most popular venues are driven by tourists and visitors and as a result are hotels, coffee shops, parks, bars and nightclubs.

Conclusions

The clusters formed as a result of the model serve this project well. While they exhibit minor overlap, as a whole, they are diverse, dispersed, and distinct. It seems one downfall was inconsistent methods of selection when clustering. To the human eye, some clusters grouped more with a geographic-orientation, while others grouped more so by venues and population changes. This may be a result of cities in close proximity sharing similar preferences and values. There is not enough data to make a definitive conclusion.

For future models, more features should be included, such as majority political ideology, majority age group, % of population under 40, % population of families, % of homeowners (as opposed to renters), overall area (square miles) of city. While the data utilized provides great insight, additional features would help organize the overlapping clusters.

As a young person living in the Midwest, there are far too many options of cities to inhabit for one to try them all. The model suggests for young people, the best options would be cities in clusters 1, 3, 6, and 8, for the reasons of population growth, proximity to major metropolitan downtowns, and high-ratings for venues popular among younger generations.