A New Way to Compare Major Markets

An Analysis of City Similarity Using Business Type Concentration

Introduction/Business Problem

One interesting question facing a variety of businesses across markets involves the nature of major cities. In the current economy, many companies and businesses serve many major cities throughout the United States and even the world. For this reason, major U.S. businesses need to know about the major markets and cities. Further, as cities are vastly different based on region, location, size, and other demographic factors, it is important to know good information about each city beyond basic facts. A major company may be successful in some cities based on demographics and culture but may not be a good fit in another location even if that location is in close geographic proximity.

For this reason, my goal is to cluster major cities in the United States and Canada based on the types of businesses which exist in that city. Rather than simply a population or geographic approach, this provides another way to assess the similarity of cities. It will attempt to cluster cities which have similar composition of business types and demonstrate to companies which cities are similar based on the types of businesses and industries which flourish in those cities and metro areas.

I was able to gather 128 major cities in the U.S. and Canada from infoplease.com and based on population. Using foursquare data after cleaning up the latitudes and longitudes, I will be able to assess the types of businesses in these 128 metro areas and then come up with a similarity clustering to form another tool which businesses can use to assess the similarity of cities. Then, these businesses can examine which cities they are likely to thrive in, which they may not, and how to improve in the areas which they are not thriving if they'd like to enter new markets which may have different characteristics.

Data

For my project, I will be using the 128 major cities mentioned in the introduction located in the United States and Canada. These were recorded by infoplease and reflect the largest metro areas by population in these two countries. The data included latitudes and longitudes. However, using the infoplease data, the latitudes and longitudes were rounded to the nearest degree leading to large inaccuracies of location when integrating the Foursquare data. For this reason, I took the 128 cities and used Microsoft Excel and the geography tools in the data tab to create accurate latitudes and longitudes for each city. This excel file is used in my included Jupyter Notebook and is also included in my Github repository for easy access of any interested reader.

Once I had the data of the 128 cities with latitude and longitude, I was able to integrate the Foursquare data. Like the previous project, I then located 100 businesses in the radius of the city and broke them down by industry to create scores for the proportion of businesses in each category. This serves as my data which I will be using for clustering as it will identify cities which have similar industry composition. I also ranked the most popular business types of each city for reference.

The data I used is included in the initial steps of my final Juptyer Notebook. I also included a data preparation notebook which is linked in my project submission so that the reader can follow my initial steps in preparation of the clustering analysis. The link is also included below:

https://github.com/AndrewKJacobs2/Coursera_Capstone/blob/master/CourseraFinalProject_DataPrep.ipynb

Methodology

In order to analyze the cities, as set up in the previous data section, I decided to use clustering. This is appropriate as rather than using any target variable, the goal is entirely to cluster similar cities without regard for any specific outcome. Instead, I will be trying simply to group similar cities together allowing companies to understand a new system which goes beyond size and geography which may explain some of the similarities between cities. For this reason, k-means clustering was an appropriate choice for this unsupervised learning problem.

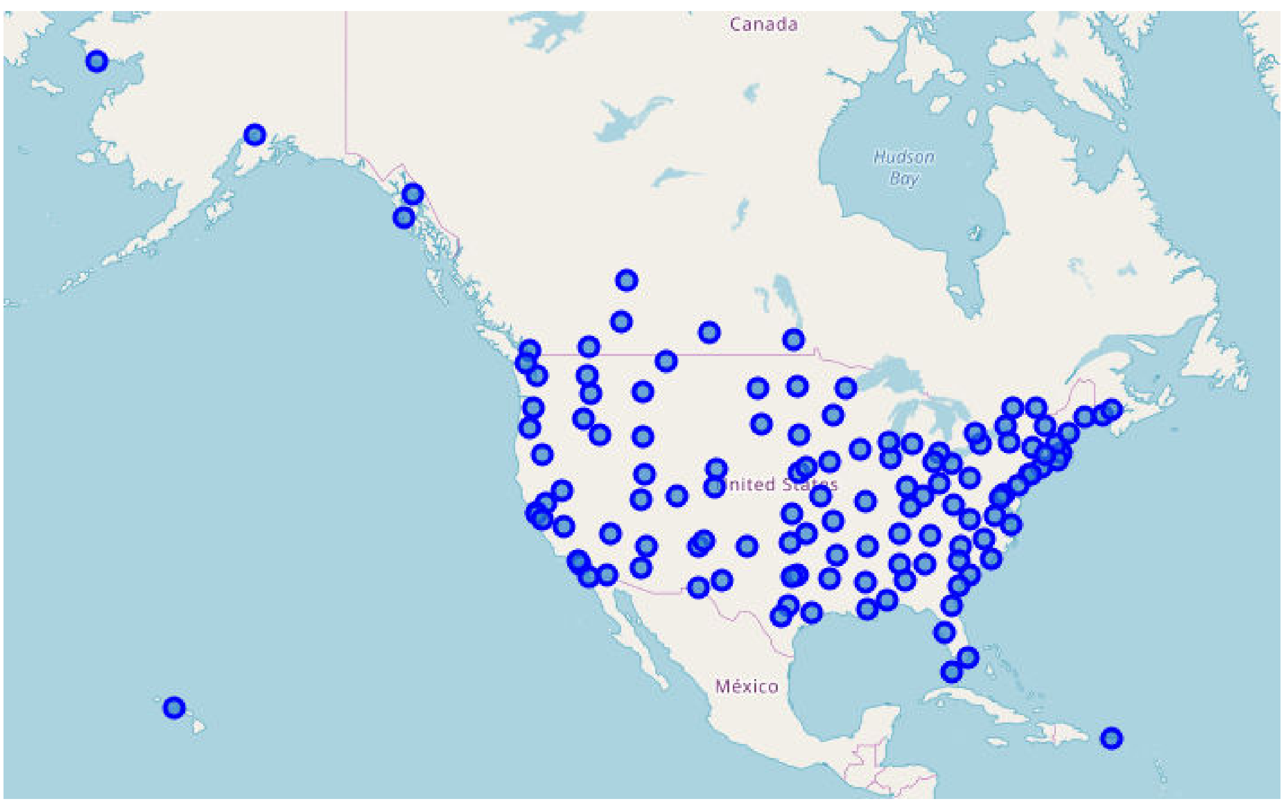
Given that there were over 100 cities included and since I wanted distinct clusters which may group together similar cities and also make key differentiations between cities with different business concentration compositions, I decided to use 10 means for analysis. This means that at the end of the analysis I will have 10 clusters or groups of cities with similar compositions of business contained in each cluster.

This method will iteratively identify the locations of the 10 means and then will classify cities to the cluster corresponding to the mean location which has the minimum distance from that city. In this case, distance is based on the composition score variable discussed in the data section. This means that each city will ultimately be classified to the mean which is closest to that city, or minimizes the error, in terms of the concentration of business types. This will result in the cities with most similar business composition structures being arranged in the same cluster.

Finally, in addition to outputting each cluster and details about each business within for analysis and comparison, I will create a map with colors corresponding to each of the 10 clusters. This allows the reader to quickly locate cities and identify similar cities using this metric. It will also demonstrate how important, if at all, geographic location will end up being associated with business composition clusters. It will also allow for a reader to easily compare cities by locating on the map and examining visually the results. For instance, one can look at major U.S. cities and determine whether the largest populations appeared in the same cluster or whether large population and similar business composition weren't related.

Results

Ultimately, I used 126 major U.S. and Canadian cities to classify using k=10 means clustering. The map of the cities to be clustered is depicted below:



Ultimately, the 10 clusters were obtained and are included below along with the most common business types for each business in the cluster:

Cluster #1

City	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Birmingham, Ala.	33.653330	-86.808890	0	Fast Food Restaurant	Pizza Place	Mexican Restaurant	Gas Station
Carlsbad, N.M.	32.411940	-104.236390	0	Pizza Place	Mexican Restaurant	Hotel	Burger Joint
Cheyenne, Wyo.	41.145560	-104.801940	0	Fast Food Restaurant	Mexican Restaurant	Clothing Store	Sandwich Place
Chicago, Ill.	41.836940	-87.684720	0	Mexican Restaurant	Sandwich Place	Taco Place	Italian Restaurant
El Centro, Calif.	32.800000	-115.567000	0	Fast Food Restaurant	Pizza Place	Mexican Restaurant	Coffee Shop
Fresno, Calif.	36.750000	-119.767000	0	Mexican Restaurant	Grocery Store	Taco Place	Coffee Shop
Grand Junction, Colo.	39.067000	-108.567000	0	Mexican Restaurant	Pizza Place	Coffee Shop	Fast Food Restaurant
Havre, Mont.	48.550000	-109.683000	0	Fast Food Restaurant	Food	Pizza Place	Sandwich Place
Helena, Mont.	46.595806	-112.027031	0	American Restaurant	Fast Food Restaurant	Sandwich Place	Coffee Shop
Hot Springs, Ark.	34.497220	-93.055280	0	Fast Food Restaurant	Hotel	Pizza Place	Mexican Restaurant
Klamath Falls, Ore.	42.225000	-121.781670	0	Pizza Place	Coffee Shop	Café	Mexican Restaurant
Lewiston, Idaho	46.410000	-117.020000	0	Fast Food Restaurant	Pizza Place	Pharmacy	Taco Place
Lincoln, Neb.	40.808890	-96.678890	0	Mexican Restaurant	Convenience Store	Park	Fast Food Restaurant
Montgomery, Ala.	32.361670	-86.279170	0	Fast Food Restaurant	Sandwich Place	Pizza Place	Fried Chicken Joint
Moose Jaw, Sask., Can.	50.393330	-105.551940	0	Fast Food Restaurant	Pizza Place	Pharmacy	Coffee Shop
Nelson, B.C., Can.	49.500000	-117.283330	0	Coffee Shop	Fast Food Restaurant	Restaurant	Grocery Store
Pierre, S.D.	44.368060	-100.336390	0	Fast Food Restaurant	Pizza Place	Bar	Hotel
Santa Fe, N.M.	35.667222	-105.964444	0	Mexican Restaurant	Fast Food Restaurant	Grocery Store	Café

Cluster #2

City	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Austin, Tex.	30.267000	-97.733000	1	Food Truck	Hotel	Coffee Shop	Bar
Cincinnati, Ohio	39.100000	-84.517000	1	Bar	American Restaurant	Hotel	Sandwich Place
Cleveland, Ohio	41.482220	-81.669720	1	Pub	Bar	Lounge	Sushi Restaurant
Columbus, Ohio	39.983000	-82.983000	1	Bar	Pizza Place	Café	American Restaurant
El Paso, Tex.	31.759206	-106.490175	1	Bar	Coffee Shop	Mexican Restaurant	Fast Food Restaurant
Fargo, N.D.	46.877220	-96.789440	1	Coffee Shop	Bar	American Restaurant	Brewery
Flagstaff, Ariz.	35.199170	-111.631110	1	Coffee Shop	Brewery	American Restaurant	Mexican Restaurant
Jacksonville, Fla.	30.336940	-81.661390	1	Sandwich Place	Bar	Coffee Shop	Brewery
Juneau, Alaska	58.300323	-134.417639	1	Seafood Restaurant	Coffee Shop	Bar	Gift Shop
Knoxville, Tenn.	35.961700	-83.923200	1	Bar	American Restaurant	Mexican Restaurant	Hotel
Las Vegas, Nev.	36.175000	-115.136390	1	Bar	Mexican Restaurant	Gastropub	American Restaurant
Nashville, Tenn.	36.166670	-86.783330	1	Bar	Hotel	Restaurant	Music Venue
Oakland, Calif.	37.804440	-122.270830	1	Coffee Shop	Bar	Mexican Restaurant	Beer Garden
Oklahoma City, Okla.	35.482220	-97.535000	1	Bar	Pizza Place	Coffee Shop	Burger Joint
Omaha, Neb.	41.250000	-96.000000	1	Coffee Shop	Bar	Pizza Place	American Restaurant
Phoenix, Ariz.	33.450000	-112.067000	1	Coffee Shop	Hotel	Bar	Art Gallery
Shreveport, La.	32.514720	-93.747220	1	Bar	Hotel	Casino	American Restaurant
Spokane, Wash.	47.658890	-117.425000	1	Bar	Pizza Place	Coffee Shop	American Restaurant

Cluster #3

City	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Albuquerque, N.M.	35.110830	-106.610000	2	Hotel	Mexican Restaurant	Furniture / Home Store	Burger Joint
Atlanta, Ga.	33.755000	-84.390000	2	Hotel	American Restaurant	Aquarium	Coffee Shop
Baker, Ore.	44.775000	-117.834440	2	Coffee Shop	Hotel	Pizza Place	Fast Food Restaurant
Baltimore, Md.	39.283000	-76.617000	2	Hotel	American Restaurant	Aquarium	Seafood Restaurant
Charleston, S.C.	32.783330	-79.933330	2	Hotel	Southern / Soul Food Restaurant	Coffee Shop	Seafood Restaurant
Charlotte, N.C.	35.227220	-80.843060	2	Pizza Place	Hotel	Steakhouse	American Restaurant
Columbia, S.C.	34.000560	-81.034720	2	American Restaurant	Bar	Hotel	Coffee Shop
Dallas, Tex.	32.779170	-96.808890	2	Hotel	Steakhouse	Bar	Coffee Shop
Detroit, Mich.	42.331390	-83.045830	2	American Restaurant	Hotel	Coffee Shop	Steakhouse
Duluth, Minn.	46.786939	-92.098194	2	Hotel	Pizza Place	Brewery	Coffee Shop
Fort Worth, Tex.	32.750000	-97.333000	2	American Restaurant	Hotel	Bar	Mexican Restaurant
Idaho Falls, Idaho	43.500000	-112.033000	2	Hotel	Fast Food Restaurant	Mexican Restaurant	American Restaurant
Indianapolis, Ind.	39.768610	-86.158060	2	American Restaurant	Hotel	Steakhouse	Pizza Place
Jackson, Miss.	32.298890	-90.184720	2	Hotel	Sandwich Place	Bar	American Restaurant
Long Beach, Calif.	33.768330	-118.195560	2	Hotel	American Restaurant	Coffee Shop	Seafood Restaurant
New Orleans, La.	29.950000	-90.080000	2	Hotel	Cajun / Creole Restaurant	Cocktail Bar	Seafood Restaurant
Pittsburgh, Pa.	40.439720	-79.976390	2	Hotel	Bar	Coffee Shop	American Restaurant
Portland, Ore.	45.520000	-122.681940	2	Hotel	Coffee Shop	Bookstore	Sandwich Place
St. Louis, Mo.	38.627220	-90.197780	2	Hotel	Bar	Italian Restaurant	American Restaurant
San Antonio, Tex.	29.417000	-98.500000	2	Hotel	Mexican Restaurant	Theater	Plaza
San Diego, Calif.	32.715000	-117.162500	2	Hotel	Mexican Restaurant	Bar	Italian Restaurant
Seattle, Wash.	47.609720	-122.333060	2	Hotel	Coffee Shop	Seafood Restaurant	Sandwich Place
Sioux Falls, S.D.	43.536390	-96.731670	2	American Restaurant	New American Restaurant	Mexican Restaurant	Hotel
Sitka, Alaska	57.051561	-135.338842	2	Coffee Shop	Hotel	Trail	Zoo
Virginia Beach, Va.	36.850600	-75.977900	2	Beach	Seafood Restaurant	American Restaurant	Hotel
Wichita, Kan.	37.688890	-97.336110	2	American Restaurant	Sandwich Place	Hotel	Bar

Cluster #4

City	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Toledo, Ohio	41.66556	-83.57528	3	Discount Store	Intersection	Fast Food Restaurant	Art Museum

Cluster #5

City	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Albany, N.Y.	42.652500	-73.757220	4	Pub	Café	Coffee Shop	Pizza Place
Anchorage, Alaska	61.217000	-149.900000	4	Seafood Restaurant	Park	Coffee Shop	Clothing Store
Boise, Idaho	43.617000	-116.200000	4	Coffee Shop	Pizza Place	Hotel	Brewery
Boston, Mass.	42.358060	-71.063610	4	Italian Restaurant	Park	Pizza Place	Bakery
Calgary, Alba., Can.	51.050000	-114.067000	4	Coffee Shop	Steakhouse	Hotel	Restaurant
Des Moines, Iowa	41.590830	-93.620830	4	Coffee Shop	Italian Restaurant	Hotel	Bar
Edmonton, Alb., Can.	53.533000	-113.500000	4	Café	Bar	Coffee Shop	Italian Restaurant
Eugene, Ore.	44.051940	-123.086670	4	Brewery	Coffee Shop	Pizza Place	Thai Restaurant
Grand Rapids, Mich.	42.961110	-85.655560	4	Coffee Shop	American Restaurant	Brewery	Café
Honolulu, Hawaii	21.300000	-157.817000	4	Japanese Restaurant	Coffee Shop	Bakery	Dessert Shop
Houston, Tex.	29.762780	-95.383060	4	Park	Trail	Coffee Shop	Pizza Place
Kansas City, Mo.	39.099720	-94.578330	4	Coffee Shop	American Restaurant	Music Venue	Brewery
Kingston, Ont., Can.	44.233000	-76.500000	4	Pub	Coffee Shop	Café	Bar
Los Angeles, Calif.	34.050000	-118.250000	4	Coffee Shop	Bar	French Restaurant	Italian Restaurant
Minneapolis, Minn.	44.983000	-93.267000	4	Coffee Shop	Park	Theater	Music Venue
Newark, N.J.	40.720000	-74.170000	4	Portuguese Restaurant	Brazilian Restaurant	BBQ Joint	Lounge
New Haven, Conn.	41.310000	-72.923610	4	Pizza Place	Coffee Shop	American Restaurant	Italian Restaurant
Ottawa, Ont., Can.	45.424720	-75.695000	4	Coffee Shop	Hotel	Restaurant	Mexican Restaurant
Philadelphia, Pa.	39.952780	-75.163610	4	Coffee Shop	Bar	Italian Restaurant	Wine Bar
Portland, Maine	43.667000	-70.267000	4	Coffee Shop	Brewery	American Restaurant	Bar
Providence, R.I.	41.823610	-71.422220	4	Italian Restaurant	Pizza Place	Bar	American Restaurant
Raleigh, N.C.	35.767000	-78.633000	4	Italian Restaurant	Cocktail Bar	Music Venue	Coffee Shop
Reno, Nev.	39.527220	-119.821940	4	Bar	Pub	Coffee Shop	Breakfast Spot
Sacramento, Calif.	38.555560	-121.468890	4	Coffee Shop	Mexican Restaurant	Vietnamese Restaurant	American Restaurant
St. John, N.B., Can.	45.280560	-66.076110	4	Coffee Shop	Park	Grocery Store	Hotel
Salt Lake City, Utah	40.750000	-111.883000	4	Coffee Shop	Bar	Thai Restaurant	Vegetarian / Vegan Restaurant
San Francisco, Calif.	37.783000	-122.417000	4	Coffee Shop	Sushi Restaurant	Marijuana Dispensary	Gym / Fitness Center
San Jose, Calif.	37.333333	-121.900000	4	Mexican Restaurant	Bar	Coffee Shop	Cocktail Bar
Syracuse, N.Y.	43.046940	-76.144440	4	Coffee Shop	Bakery	Italian Restaurant	Pizza Place
Vancouver, B.C., Can.	49.250000	-123.100000	4	Coffee Shop	Vietnamese Restaurant	Arts & Crafts Store	Indian Restaurant
Victoria, B.C., Can.	48.428610	-123.365560	4	Coffee Shop	Restaurant	Breakfast Spot	Vegetarian / Vegan Restaurant
Washington, D.C.	38.906188	-77.017263	4	Cocktail Bar	Coffee Shop	Bar	Italian Restaurant
Winnipeg, Man., Can.	49.899440	-97.139170	4	Coffee Shop	Asian Restaurant	Café	Hotel

Cluster #6

City	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Eastport, Maine	44.91361	-67.00389	5	Seafood Restaurant	Food	State / Provincial Park	Bakery

Cluster #8

City	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Amarillo, Tex.	35.19917	-101.845280	7	Sandwich Place	Restaurant	American Restaurant	Bank
Bangor, Maine	44.80000	-68.800000	7	Rental Car Location	Pizza Place	Hotel	Airport Terminal
Bismarck, N.D.	46.81333	-100.778890	7	Clothing Store	Pizza Place	Coffee Shop	Sandwich Place
Buffalo, N.Y.	42.90472	-78.849440	7	Discount Store	Bar	Intersection	Gay Bar
Charleston, W. Va.	38.34722	-81.633330	7	Pizza Place	Bar	Discount Store	American Restaurant
Denver, Colo.	39.76185	-104.881105	7	Coffee Shop	Pool	American Restaurant	Sandwich Place
Dubuque, Iowa	42.50000	-90.690000	7	Pizza Place	Bar	Mexican Restaurant	Coffee Shop
Louisville, Ky.	38.22533	-85.741700	7	Bar	Pizza Place	Coffee Shop	Sandwich Place
Manchester, N.H.	42.99083	-71.463610	7	Café	American Restaurant	Pizza Place	Donut Shop
Memphis, Tenn.	35.11750	-89.971110	7	Discount Store	Bar	Café	Convenience Store
Miami, Fla.	25.77528	-80.208890	7	Smoke Shop	Seafood Restaurant	Mexican Restaurant	Bar
Milwaukee, Wis.	43.05000	-87.950000	7	Bar	American Restaurant	Sandwich Place	Grocery Store
Mobile, Ala.	30.69444	-88.043060	7	Intersection	Seafood Restaurant	American Restaurant	Southern / Soul Food Restaurant
Montpelier, Vt.	44.26000	-72.575280	7	Gas Station	Convenience Store	Hotel	Thai Restaurant
Richmond, Va.	37.53300	-77.467000	7	Park	American Restaurant	Pizza Place	Coffee Shop
Roanoke, Va.	37.27083	-79.941670	7	Coffee Shop	American Restaurant	Sandwich Place	Park
Savannah, Ga.	32.01700	-81.117000	7	Department Store	Furniture / Home Store	Hotel	Fast Food Restaurant
Springfield, Mass.	42.10139	-72.590280	7	Donut Shop	Sandwich Place	Gas Station	American Restaurant
Tampa, Fla.	27.96806	-82.476390	7	Cuban Restaurant	Park	Coffee Shop	Spanish Restaurant
Toronto, Ont., Can.	43.74167	-79.373330	7	Coffee Shop	Park	Shopping Mall	Sandwich Place
Tulsa, Okla.	36.13139	-95.937220	7	Sandwich Place	Fast Food Restaurant	Burger Joint	Pizza Place
Wilmington, N.C.	34.22333	-77.912220	7	Fast Food Restaurant	Shoe Store	Department Store	Clothing Store

Cluster #7

City	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Key West, Fla.	24.559720	-81.783610	6	Hotel	Cuban Restaurant	Resort	Bed & Breakfast
Montreal, Que., Can.	45.508890	-73.561670	6	Café	French Restaurant	Hotel	Restaurant
New York, N.Y.	40.661000	-73.944000	6	Caribbean Restaurant	Café	Bakery	Cocktail Bar
San Juan, P.R.	18.451522	-66.069481	6	Caribbean Restaurant	Hotel	Italian Restaurant	Restaurant

Cluster #9

City	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Richfield, Utah	38.76583	-112.0875	8	Pizza Place	Steakhouse	Fast Food Restaurant	Sandwich Place

Cluster #10

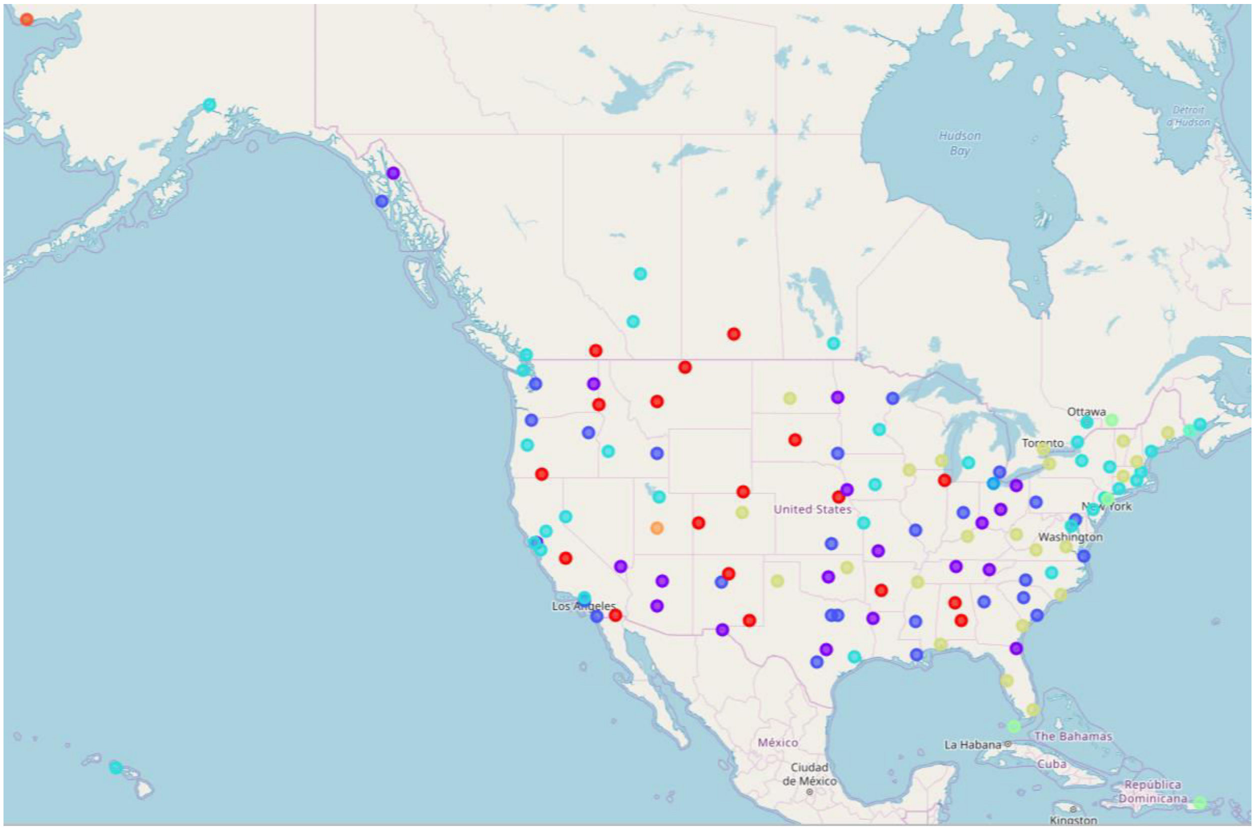
City	Latitude	Longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
Nome, Alaska	64.50389	-165.39944	9	Hotel	Grocery Store	Bakery	Restaurant

The breakdown of the clusters by the size of the cluster is as shown in the table below:

Cluster	
1	18
2	19
3	26
4	1
5	33
6	1
7	4
8	22
9	1
10	1

The largest cluster had 33 cities or approximately 26% of all the major cities analyzed. Four clusters only included one observation. These 4 cities appear to be unusual from the rest of the dataset. However, of the remaining 6 clusters, 5 clusters had at least 18 observations. There were about 5 healthy sized clusters for analysis and comparison. It can be noted that the cities not contained in these clusters are often unusual. For instance, most of the cities not included in a cluster were very small compared to the rest of the data. Richfield, Utah; Nome, Alaska, and Eastport, Maine were 3 out of the 4 smallest cities included of the 126 and made up 3 out of the 4 single observation clusters. Further, the other small cluster (#7) contained cities which are popular tourist destinations and may have unusual characteristics with the majority of major cities which are not as heavily visited. For this reason, it appears that the data reflects several large clusters of similarly business concentrated cites and then differentiates individual or small number of cities which have abnormal populations or traits from commonly included major cities.

A map of the U.S. and Canada with these 126, like included earlier, is provided again below while now including color coded markers to reflect the cluster which each city belongs.



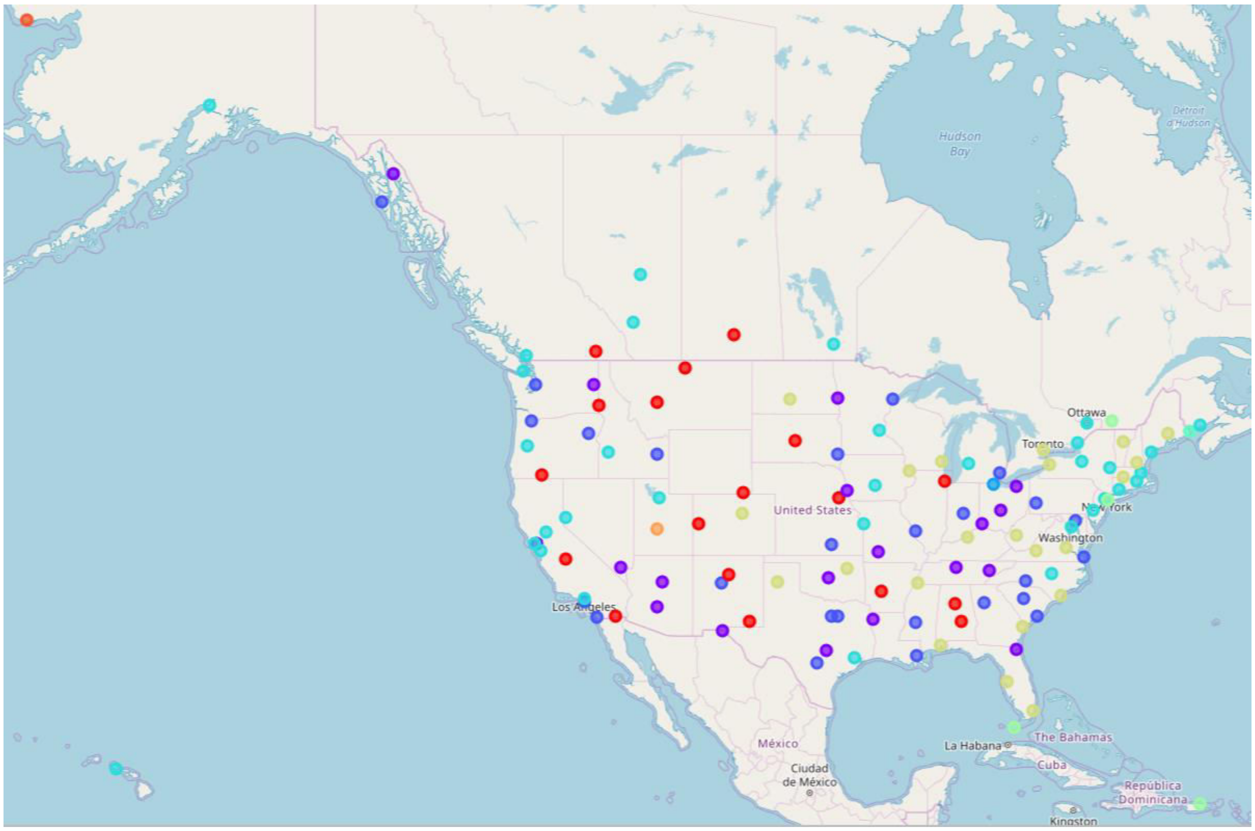
This map will be discussed and analyzed further in later sections.

Discussion

First, I wanted to examine the nature of the clusters. As noted in the results, several clusters were small or only contained one city. However, for the bigger clusters, there are several questions worth examining. First, it was worth examining whether there were any geographic linkages to business type composition of cities. I sorted the latitudes and longitudes of all the cities along with the cluster to which each city belonged. From this I first gathered the geographic center of each cluster in terms of latitude and longitude. This allowed for the examination of where each cluster was centered within the continent. The results are given in the map below:

It can be seen that the majority of clusters were somewhat centered in the United States. There was an unusual point in Alaska as well as the tip of Maine. However, these belonged to the earlier mentioned single observation clusters with very small towns. Instead, of the large clusters containing over 15 cities, the clusters were all somewhat centrally located in the United States. There appears to be no extreme clustering of West Coast or East Coast cities or of Southern vs. Northern cities. The 4 points in cluster 7 were all east coast points and, as mentioned, these cities were unusually popular tourist cities. Aside from this cluster, it appears that the results demonstrate that business type composition is not significantly linked to geography and instead reflects another component of city demographics.

Beyond the central location of the clusters, however, it is worth examining the geographic diversity of the clusters further. An initial observation of the clusters map, provided again below, shows that most clusters have a large amount of geographic diversity:



While the light blue cluster (#5) appears to have a large amount of coastal cities with close proximity to one another on the East Coast and another similar segment on the West Coast, there are also a belt of cities in the center of the country in this cluster including Minneapolis, Des Moines, and Kansas City. Likewise, while the gold region does not include any West Coast cities, it appears that this cluster is spread throughout the East, South, Midwest, and West and does demonstrate geographic diversity. Similarly, the red region has many Western cities and no cities along the East Coast. However, it has cities in the Midwest and the South included in the cluster. For this reason, of the major clusters of size at least 15, it is evident that none of these clusters represent one geographic area. While some geographic trends in business composition similarity may exist, it is not a major factor in dictating what causes these similarities.

To further analyze this, I obtained the standard deviation in latitude and longitude of each cluster as given below:

Cluster	Pop	Lat	Long
1	627969.10	6.18	11.39
2	397486.81	6.85	15.44
3	422432.05	6.55	16.70
4	nan	nan	nan
5	786382.41	6.96	24.86
6	nan	nan	nan
7	3987689.99	12.86	6.42
8	573813.13	5.70	10.04
9	nan	nan	nan
10	nan	nan	nan

Standard Deviation within Each Cluster

While there is no standard deviation to compute for the 4 clusters of a single city, there were large standard deviations amongst the other clusters. Specifically, among these clusters, the minimum latitude standard deviation was 5.7. Since a degree of latitude is always approximately 69 miles due to the parallel nature of the earth and the slightly ellipsoidal shape, this can be very closely computed to miles. Since the minimum cluster standard deviation in latitude was 5.7, this means the standard deviation is very close to 393 miles. This is a substantial North/South distance between a given city and the mean of the cluster. Longitude is slightly different in that it runs through the poles, but a useful rough estimate is that a degree of longitude is 53 miles as this would be the amount at 40 degrees North which equates to the middle of the United States. Then, cluster 7 would have the minimum standard deviation in East/West distance at about 340 miles. However, note that cluster 7 was the lone "small" cluster of popular tourist cities. Among the 5 large clusters of more than 15 cities, the minimum standard deviation in longitude degrees was 10.04 which amounts to around 532 miles. Clearly, this reflects that the business composition similarity clustering goes beyond geography and reflects further understanding of a city and will allow for identifying similar cities in a way which may include some geographic correlation but also goes beyond geography exclusively.

Another important factor which was important to examine in relation to the clusters is population size. Often, cities are compared to similar cities not only on a geographic basis but on a population basis. For this reason, I wanted to gather the results of my clustering in terms of the population sizes. First, the mean population of each cluster is given below:

Cluster	Pop
1	244071.33
2	501913.63
3	523851.77
4	278508.00
5	623328.55
6	1219.00
7	2656107.75
8	382253.91
9	7723.00
10	3797.00

Mean Population within Each Cluster

Note that clusters 4,6,9, and 10 contain only a single city. As noted, three of these were the included cities with abnormally small populations. A more useful analysis would be to examine the large clusters for any substantial population differences. Recall that cluster 7 also only contains 4 cities. One of these cities was New York City which greatly inflated this mean. Instead, the clusters 1, 2, 3, 5, and 8 make up the 5 major clusters of more than 15 cities. These mean populations were all quite similar as the smallest was above 244,000 and the largest was below 624,000. To examine further the city size composition within each cluster, the standard deviations within each cluster were obtained as shown below:

	Pop
Cluster	
1	627969.10
2	397486.81
3	422432.05
4	nan
5	786382.41
6	nan
7	3987689.99
8	573813.13
9	nan
10	nan

Standard Deviation City Population within Each Cluster

Again, clusters with only one city will be blank. Likewise, cluster 7 had a substantially larger standard deviation than any other cluster. As noted, this was a small cluster of only 4 popular tourism cities which also contained New York City and was heavily inflated due to the small amount of datapoints and presence of the outlier, New York City. Instead, the bulk of analysis comes from the remaining 5 major clusters (1, 2, 3, 5, and 8). These cities all had standard deviations of at least 397,000. This leads to a similar analysis as with the discussion of geography and the clustering: it appears that while some similar sized cities may be linked within the same cluster, the majority of business composition clustering reflects city demographics and characteristics which go beyond population size. This demonstrates that this clustering may support common tools to assess similarity such as geographic proximity and population size, but will also provide useful similarity measures of cities based on business type composition which provides insight beyond these two common tools.

Conclusion

The goal of this project was to provide insight into city similarity based on business type composition within the cities and to form a similarity measure which supports businesses when comparing cities and determining which types of cities in which they thrive. The results of this project reflect this goal. First, the clustering of cities fell into 5 major clusters. The vast majority of cities (94%) fell into these clusters and these major clusters ranged only from 14% to 26% of the 126 cities representing a healthy amount in each major cluster, but no individual cluster which would be too large for good analysis and comparison. Then, it can be further analyzed what sources may lead to similar business compositions in these cities. Factors such as income, culture, age, demographics, history, industry and economics, size, and geography all impact cities. A business looking to explore new opportunities in major markets should assess these clusters and examine cities which are similar and how they may thrive in certain markets and can look toward what factors may limit potential in other markets and how they might overcome those limitations.

Two commonly used similarity measures among major cities and metro areas are geography and population size. My analysis examined these two factors within the clustering of business/venue type composition of cities. While there may be linkages between geography and size and which cluster certain cities fall, it appears that the composition of businesses in a city reflects factors beyond these two simple measurements. Instead, a holistic approach may be best. For instance, two cities may be close in size and geography such as Minneapolis and Milwaukee but have unique business type compositions in these cities. Instead, if a company is examining similar markets to Minneapolis, they may choose Des Moines which is close in proximity to Minneapolis but also lies in the same cluster. Further, a model of success or failure could build upon business/venue composition, geography, population, and other key demographic factors for that specific industry to provide a rigorous analysis of markets. Using these clusters as categorical variables could be useful in such advanced analysis including regression or logistic regression models.

Overall, this clustering represents a new way to compare city similarity and group major U.S. and Canadian markets based on similar venue composition. This provides a new tool which offers value beyond simple measures such as geography and population giving it merit in comparing cities. It provides a clear and simple grouping of major clusters and each cluster is representative of multiple geographic regions and varied population sizes. It would be best used as a tool to support thorough analysis. While different industries have different factors in market analysis such as the economics, size, age demographics, income distribution and other factors of a city, the analysis can also examine similarities in the businesses saturating these cities using this clustering tool to provide another layer of insight and analysis to be used for projecting market success or assessing areas of strength and weakness across major North American metro areas.

References

Aklson, A. (n.d.). Applied Data Science Capstone. Retrieved from https://www.coursera.org/learn/applied-data-science-capstone/home/welcome

Latitude and Longitude of U.S. and Canadian Cities. (n.d.). Retrieved from https://www.infoplease.com/world/united-states-geography/latitude-and-longitude-us-and-canadian-cities

Rosenberg, M. (2018, September 28). How to Figure Out the Distance Between Degrees of Latitude and Longitude. Retrieved from https://www.thoughtco.com/degree-of-latitude-and-longitude-distance-4070616