Data Science Capstone Project:

A Comparison of Dallas and Fort Worth Neighborhoods

Brian C. Pennington

Data Science Capstone Project:

A Comparison of Dallas and Fort Worth Neighborhoods

Introduction

It is often said that Dallas is where the east ends and Fort Worth is where the west begins. Certainly, it is true that when you think of Fort Worth, you think of pickup trucks and cowboy b oots and Billy Bob's at the Fort Worth Stockyards and when you think of Dallas, you think of B MW's and banking and Highland Park. The question for someone considering moving to the DF W area, though, is: is there a difference in the neighborhoods in Fort Worth and Dallas?

This project will investigate whether there are any systematic differences in Dallas vs. Fort Worth neighborhoods.

Data

To compare neighborhoods in Dallas and Fort Worth we had to gather geo-data to define the neighborhoods and data about the neighborhoods to be the basis of comparison.

Zillow Neighborhoods Dataset

Description

The Zillow U.S. Neighborhoods dataset has geo-data on over 17,300 neighborhoods in the largest cities in the U.S. The Zillow data team used various tactics to collect the data, including calling individual chambers of commerce, tourism and convention boards, speaking with real estate agents and community members in these areas, as well as using available online local sources.

Data Format

In addition to geo-shape data for the neighborhood boundaries, the dataset properties sections include the neighborhood name, region id, two-dimensional geographical coordinates, city, county and state.

Manipulation

Zillow makes this data available for free under a Creative Commons license. All that is required to use the files is to attribute Zillow as the source of the data and make any updates of changes you make to the files available to everyone else via the same or similar license.

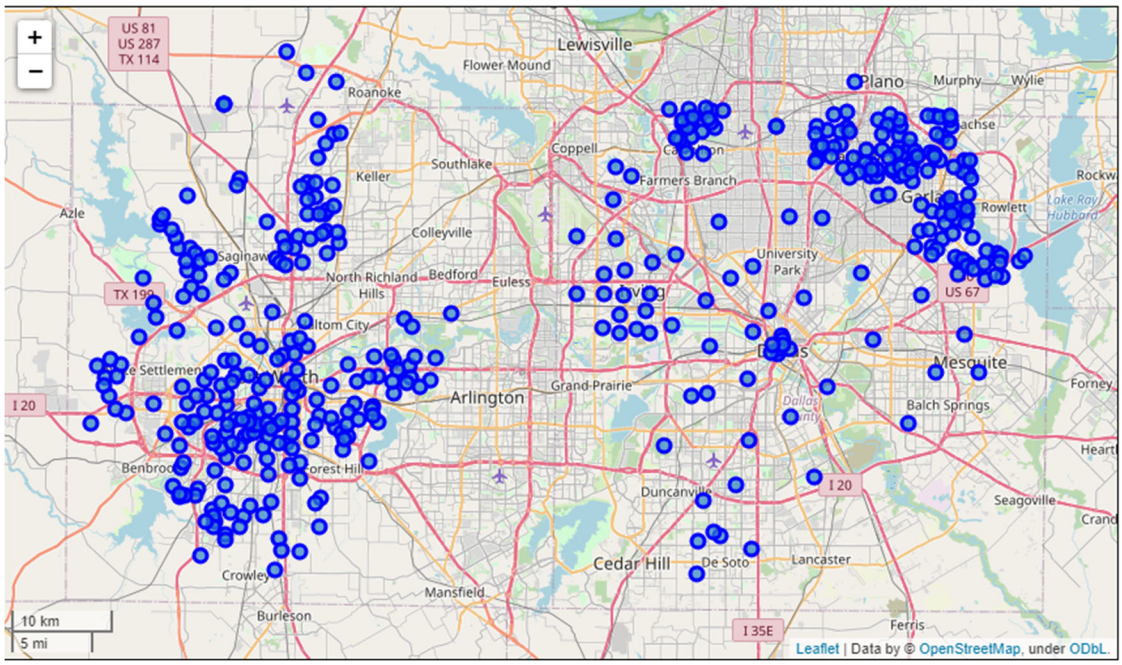
The data can be downloaded from the opendatasoft website: https://data.opendatasoft.com/explore/dataset/zillow-neighborhoods%40public/export/

The data is available in a variety of formats including:

- csv - comma-separated-values;
- JSON - JavaScript Object Notation;
- Excel;
- GeoJSON - an open standard format based on the JSON that was designed for representing simple geographical features along with their non-spatial attributes;
- KML - Keyhole Markup Language.

We used the GeoJSON files.

As you can see in the map below (it may be better to open an interactive html version of the map that can found at DFW Neighborhoods), some neighborhoods are far apart, but many are close together. This created a number of interesting wrinkles in collecting data about the neighborhoods.



Place Data by Foursquare

Information about the neighborhoods was collected using the Foursquare Places database. The Places database has an extensive amount of venue data and user content that can be accessed using geolocation data. The database contains "firmographic" data and "rich community-sourced content" for more than 60 million commercial places around the world.

For each venue, there are over 90 attributes that can be accessed, including the ve nue name, address, gelocation, category as well as ratings, tips and photos.

Our focus was on using the types of venues in each neighborhood to characterize the nature of the neighborhoods.

Setting the radius

As we mentioned above, some neighborhoods are close together, while some are farther apart. To determine the radius, we first used the Shapely Polygon “bounds” property that returns two coordinates which bounds the polygon. Then we used geopy to compute the vincenty distance in meters. For example, if you have two coordinates p1 and p2, you use

1000*geopy.distance.vincenty(p1,p2).km. To be conservative, we would use that as the radius. In addition, for reasons that will be made clear below, the minimum radius we used was 1,600 meters.

Pagination

One thing that we discovered while working on this project was that while there may be hundreds of venues within the radius, the Places API only returns 100 places with each explore endpoint call. So after the initial call, we checked the "totalResults" property and used a while loop where we would increment the "offset=" parameter by 100 and make an additional request until we processed all of the venues.

Neighborhood venues inclusion rule

Boundary check

The first thing we check is whether the venue is "in the neighborhood". To do this we use Shapely's Point within function. The venue coordinates define the Point and the neighborhood boundaries define the Polygon, so it's as easy as Point.within(Polygon)=True.

Minimum radius

We decided that we didn't want to rely solely on the neighborhood boundaries. Even if a venue was not in the polygon, it was included if it was within a minimum radius. We used 800 meters, roughly one-half of a mile, for the minimum radius.

Venue Groups vs. Venue Categories

One thing that we will see in the results section is that the granularity of the venue categories is not conducive to good clustering results. At the same time, we wanted to make a finer distinction than the higher level hierarchy of the Four Square venue classification system. For example, we didn't want to group all restaurants together. We wanted to know if a

neighborhood had a lot of Mexican cuisine or Asian cuisine, but we didn't care whether the former was a Mexican Restaurant, a Tex-Mex Restaurant, a Burrito Place, or a Taco Place or whether the latter was an Asian, Chinese, Indian, Japanese, Korean, Persian, Sushi, Thai or Vietnamese Restaurant or whether it was a Noodle House. Similarly, a fast food place is a fast food place regardless of whether it's serving burgers, chicken or hot dogs. So we made a custom Venue Group in an attempt to group very similar venues while at the same time being able to differentiate distinctive neighborhood characteristics. (See Venue Groups for a full listing.)

Methodology

K-means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. With roots in signal processing, the aim is to partition the observations into a set number of clusters where each observations belongs to the cluster with nearest to the centroid of the cluster.

Pros

1. It is simple to understand.
2. It is fast and efficient.
3. It is easy to implement.
4. You always get a result.

Cons

1. You have to specify the number of clusters.
2. It is sensitive to initialization of the centroids.
3. It is sensitive to outliers.

Determining the number of Factors

Elbow Method

A very common method to determine the number of factors is to plot the distortion scores for a range of clusters and to use the point where there is an "elbow" as the optimal number of clusters. The distortion score is the average of the squared distances from each point to its assigned center. The intuition around this method is that if there are valid clusters, you should see a dramatic improvement in the score, but at some point the improvement "flattens out" creating an "elbow" in the graph of the scores.

The problem is that it is not always easy to tell exactly where this occurs and the choice of the number of clusters can be somewhat arbitrary.

Silhouette Coefficient

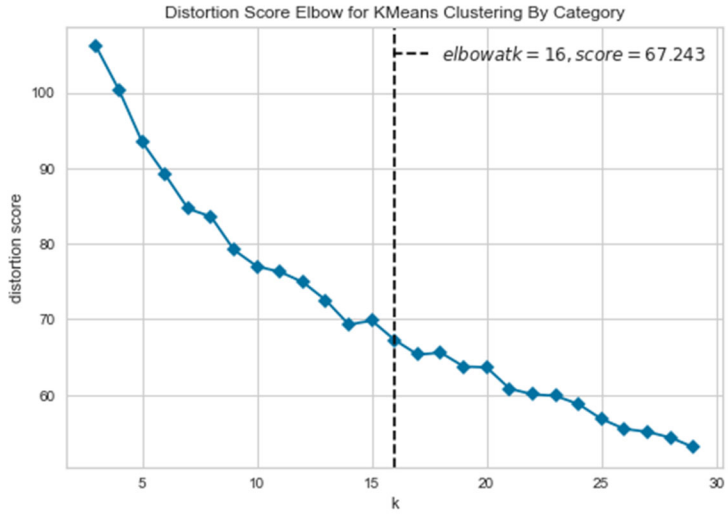
The Silhouette Coefficient is calculated by (x-y)/ max(x, y) where y is the mean intra-cluster distance and x is the mean inter-cluster distance. The coefficient varies between -1 and 1 with a value close to 1 being very good, a value close to 0 indicates a lot of cluster overlap, and negative values indicate very poor results. The reason that it is usually preferred is that it takes into account both how tight the clusters are (y) and how far apart the different clusters are (x).

Results

Clustering By Venue Category

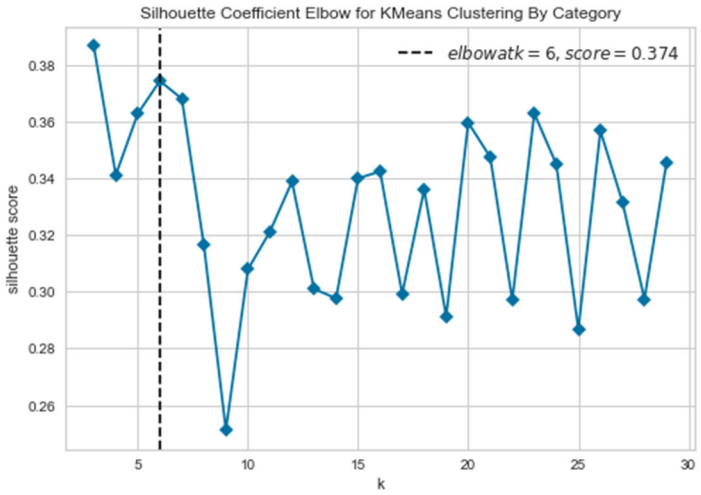
Distortion score elbow

The distortion scores (and the silhouette coefficients below) were plotted using Yellowbrick's KElbowVisualizer. Their visualizer selects 16 as the optimal number of clusters, but you can see how arbitrary that is. Why not 6 clusters? Or 14?

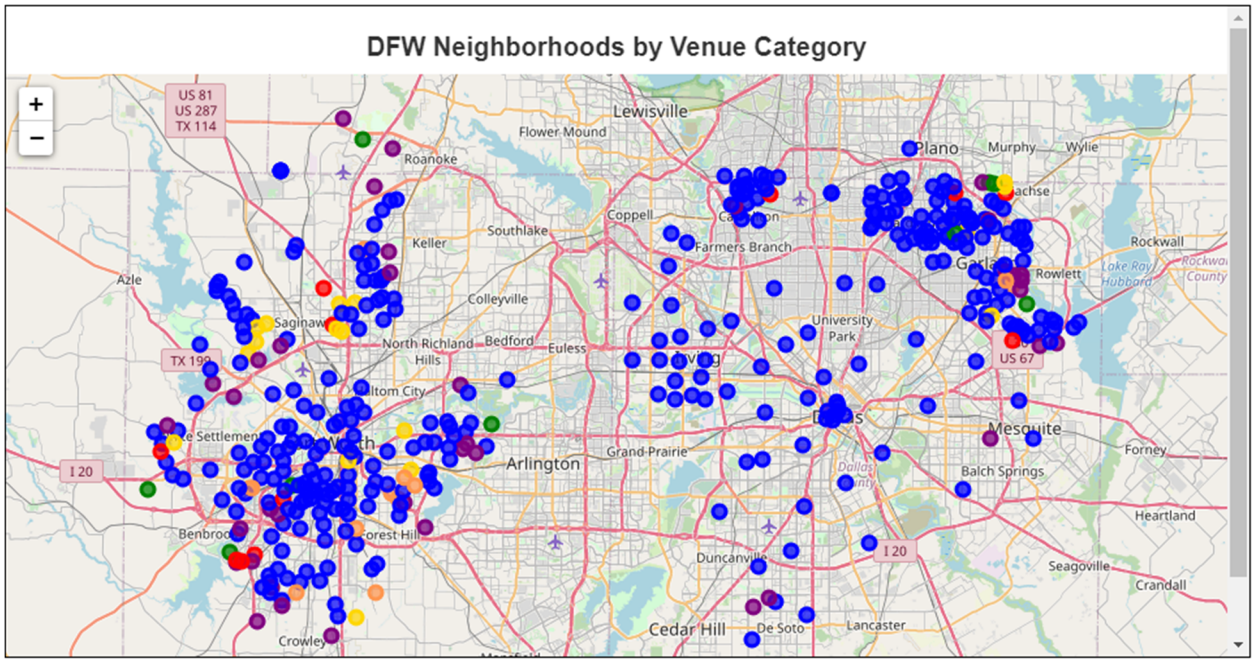


Silhouette Coefficient

The silhouette coefficient seems much more plausible. While, as we will see below, there are some obvious clusters, there is not much to be gained by adding additional clusters. Basically, adding in additional centroids leads to a lower distortion score, but the problem that the silhouette coefficient highlights is that while that reduces the intra-cluster distance, there is not a corresponding improvement in the intra-cluster distance. In addition, if you rerun the cluster analysis additional times, you will observe material differences in the resulting charts. This is indicative that the initialization of the centroids is driving the results which reinforces that clusters large than 6 are random.



Clustering Results, k=6



(Browser version available at DFW by Venue Category)

The details of the clusters can be found in Table 1. We can see that there are some completely residential neighborhoods, and others with parks or golf courses without many other venues. There are a small number of neighborhoods that have mostly convenience stores and another cluster with mostly discount stores. But the vast majority of the neighborhoods—almost 80%--are unable to be differentiated using venue category.

Table 1. K-means clustering results using Venue Category, k=6.

Clusters By Category

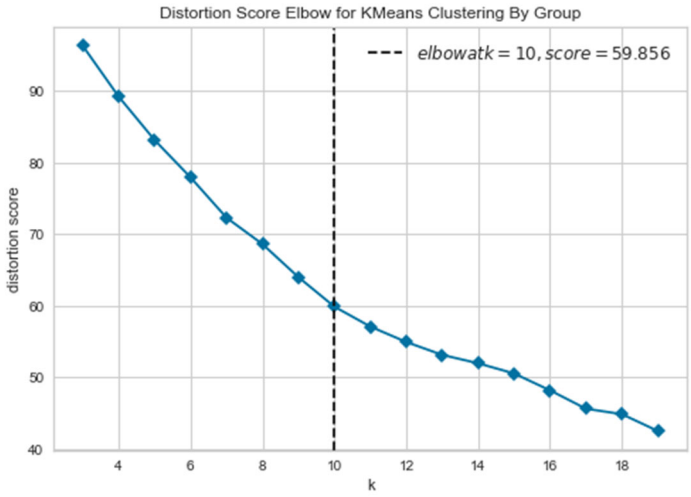
Cluster	Color	Num	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	List
0	yellow	18	Convenience Store-62%	Video Store-5%	Intersection-5%	Mexican Restaurant-3%	Basketball Court-3%	Cluster 0
1	purple	38	Residential-100%					Cluster 1
2	orange	12	Discount Store-67%	Fried Chicken Joint-9%	Grocery Store-4%	Convenience Store-4%	Gym-4%	Cluster 2
3	red	11	Park-77%	Golf Course-6%	American Restaurant-5%	Intersection-3%	Pharmacy-2%	Cluster 3
4	green	9	Golf Course-89%	Other Repair Shop-6%	Hardware Store-6%			Cluster 4
5	blue	337	Fast Food Restaurant-6%	Mexican Restaurant-5%	Pizza Place-4%	Discount Store-3%	Convenience Store-3%	Cluster 5

(The percentages are the average venue percentages of the neighborhoods in the cluster. Return to text)

Clustering by Venue Group

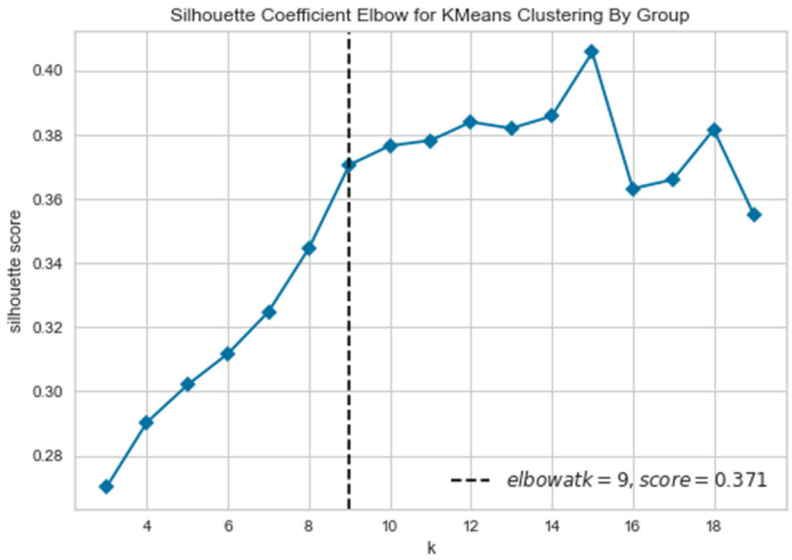
Distortion statistic elbow

The Yellowbrick visualizer detects the elbow at 10 clusters, but, again, that choice looks fairly arbitrary.

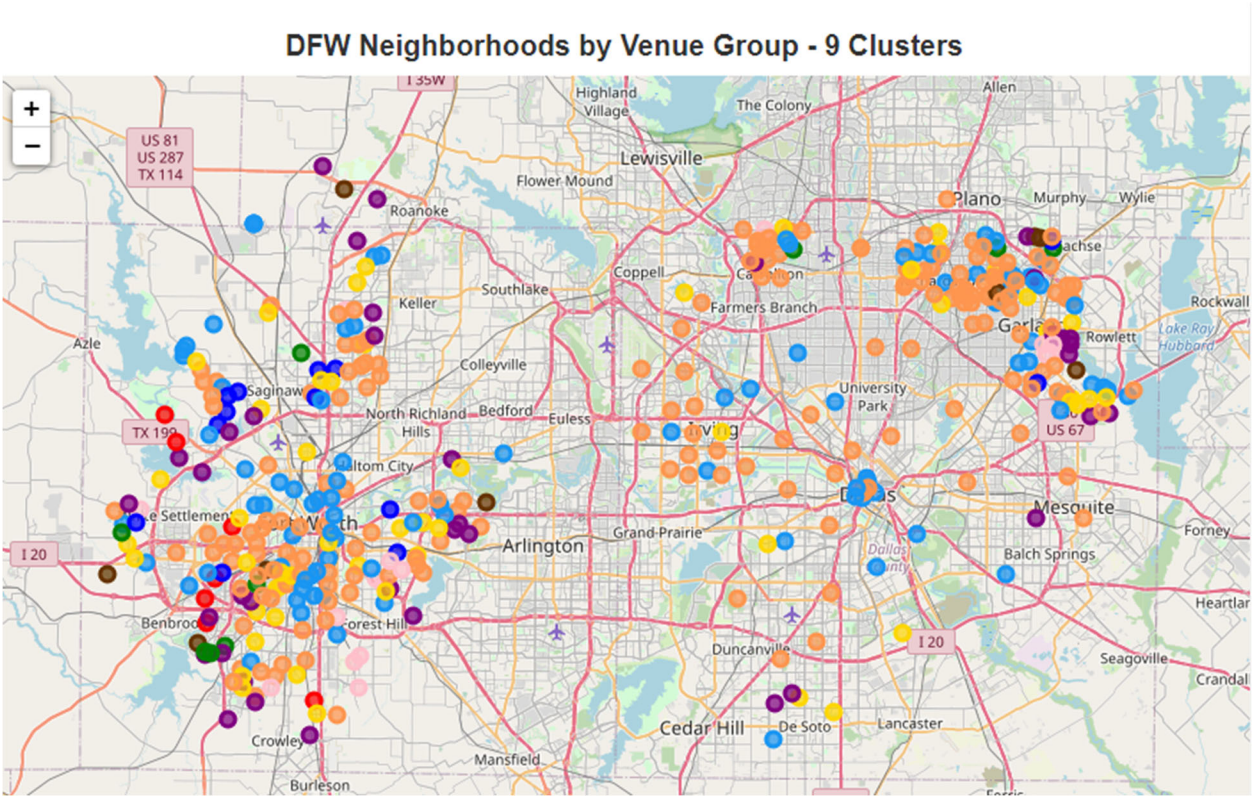


Silhouette Coefficient

The silhouette coefficient looks more reasonable. The visualizer highlights an interesting result. Some practitioners apply the elbow method to the silhouette coefficient—hence the visualizer's choice of 9 clusters—while others argue that since the silhouette coefficient is two-dimensional, i.e. it takes into account the intra-cluster as well as inter-cluster distances as opposed to the one-dimensional distortion score, it is not necessary to use the elbow method and the optimal number of clusters is the global maximum—which in this case would argue for 15 clusters. We decided to do the analysis for both 9 clusters and 15 clusters to see if the results gave us any insight as to which approach might be more reasonable.



Clusters by Venue Group, k=9



(Browser version available at [DFW Neighborhoods by Venue Group - 9 Clusters])

The cluster analysis isolated many of the same clusters that were identified by the category cluster analysis (the details can be found in [Table 2]):

- The 38 residential neighborhoods (with no venues);
- The communities that were characterized by golf courses and those characterized by parks;
- There were a slightly smaller number of communities that had predominantly convenience stores and the communities dominated by shopping venues most likely had a lot of overlap with the communities that were dominated by discount stores in the category clustering.

The venue groups enabled some differentiation in the huge vat of "everything else" in the category clustering:

- There were 53 neighborhoods where fast food venues were by far the dominant feature followed by convenience stores, drinking establishments, gas stations, and Asian cuisine.
- There were 185 neighborhoods where fast-food was the dominant venue followed by a mix of shopping venues, Asian cuisine, Mexican cuisine and grocery stores.
- There were 94 neighborhoods that had mostly Mexican cuisine followed by athletic venues, drinking establishments and home services.

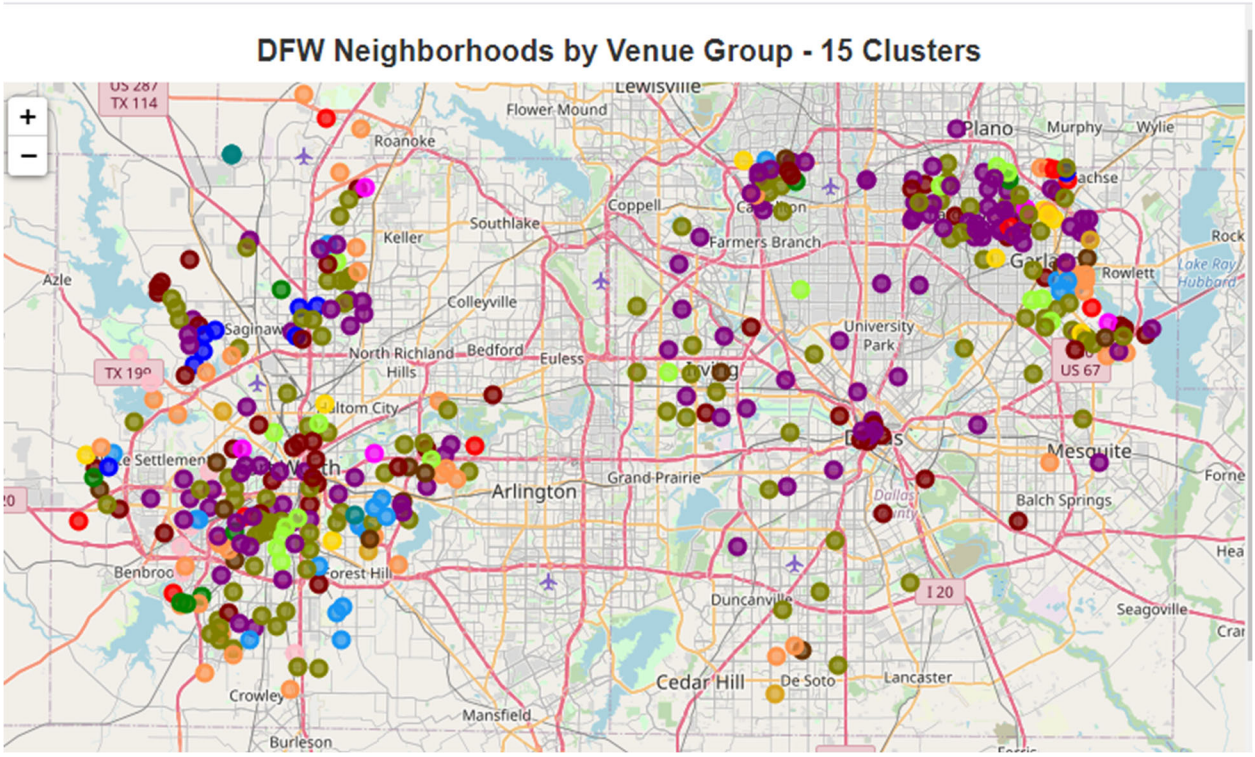
Table 2. K-means clustering results using Venue Group, k=9.

Clusters by Group, k=9

Cluster	Color	Num	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	List
0	yellow	53	Fast-Food Venue-60%	Convenience Store-5%	Drinking Establishment-3%	Gas Station-3%	Asian Cuisine-3%	Cluster 0
1	purple	38	Residential-100%					Cluster 1
2	orange	185	Fast-Food Venue-28%	Shopping Venues-7%	Asian Cuisine-6%	Mexican Cuisine-5%	Grocery Store-5%	Cluster 2
3	red	7	Outdoor Destination-86%	Intersection-7%				Cluster 3
4	green	9	Park-85%	American Cuisine-6%	Golf Course-6%	Intersection-4%		Cluster 4
5	blue	13	Convenience Store-73%	Shopping Venues-8%	Video Store-4%	Intersection-4%	Athletic Venue-4%	Cluster 5
6	sky blue	94	Mexican Cuisine-12%	Athletic Venue-7%	Drinking Establishment-6%	Home Service-5%	American Cuisine-4%	Cluster 6
7	brown	9	Golf Course-89%	Home Shop-6%	Home Service-6%			Cluster 7
8	pink	17	Shopping Venues-62%	Fast-Food Venue-17%	Gym-7%	Grocery Store-4%	Health Services-3%	Cluster 8

(The percentages are the average venue percentages of the neighborhoods in the cluster. Return to text)

Clusters by Venue Group, k=15



(Browser version available at DFW Neighborhoods by Venue Group - 16 Clusters)

We can get some interesting insights into why the silhouette coefficient was higher for 15 clusters (details in Table 3).

First, there were three clusters that were small, but highly distinctive:

- Health services (4 neighborhoods);
- Home services (4 neighborhoods);
- Athletic venues (4 neighborhoods);

In addition, there was the ability to make more differentiation in the types of eating establishments, including some neighborhoods dominated by Asian cuisine and some that had primarily drinking establishments that seemed to be situated around parks and marinas.

Table 3. K-means clustering results using Venue Group, k=15

Cluster	Color	Num	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	List
0	yellow	9	Asian Cuisine-39%	Grocery Store-29%	Fast-Food Venue-9%	Mexican Cuisine-6%	Light Rail Station-6%	Cluster 0
1	purple	120	Fast-Food Venue-22%	Shopping Venues-8%	American Cuisine-6%	Mexican Cuisine-6%	Asian Cuisine-5%	Cluster 1
2	orange	38	Residential-100%					Cluster 2
3	red	10	Golf Course-85%	Home Shop-5%	Park-5%	Home Service-5%		Cluster 3
4	green	8	Park-90%	American Cuisine-6%	Intersection-4%			Cluster 4
5	blue	9	Convenience Store-83%	Video Store-6%	Intersection-6%	Gym-3%	Outdoor Destination-3%	Cluster 5
6	sky blue	19	Shopping Venues-60%	Fast-Food Venue-15%	Convenience Store-6%	Gym-6%	Grocery Store-4%	Cluster 6
7	brown	14	Fast-Food Venue-88%	Gas Station-3%	Shopping Venues-3%	Athletic Venue-2%	Convenience Store-2%	Cluster 7
8	pink	7	Outdoor Destination-86%	Intersection-7%				Cluster 8
9	gold	4	Home Service-88%	Home Shop-6%	Shopping Venues-6%			Cluster 9
10	fuchsia	5	Athletic Venue-80%	Home Service-10%	Convenience Store-10%			Cluster 10
11	lime	18	Mexican Cuisine-45%	Convenience Store-8%	American Cuisine-5%	Athletic Venue-4%	Fast-Food Venue-4%	Cluster 11
12	olive	109	Fast-Food Venue-41%	Convenience Store-5%	Mexican Cuisine-5%	Asian Cuisine-5%	Shopping Venues-4%	Cluster 12
13	teal	4	Health Services-100%					Cluster 13
14	maroon	51	Drinking Establishment-11%	Park-6%	Harbor / Marina-6%	Gas Station-4%	American Cuisine-4%	Cluster 14

Discussion

Neighborhood diversity

Looking at the cluster maps, it is clear that Fort Worth has more diversity in its neighborhoods than Dallas. The northern suburbs of Dallas--Carrollton, Richardson, Garland—have a similar diversity profile to Fort Worth.

Neighborhood size

The other striking feature is how much smaller the neighborhoods are in Fort Worth and the northern suburbs of Dallas than in Dallas proper. It is possible that this is because in 1911, Dallas adopted a city plan developed by George Kessler, who was raised in Dallas and later become a nationally recognized civic engineer. It could be that the larger size of the neighborhoods is masking intra-neighborhood diversity.

Fast Food Venues

The other thing that is notable is how fast food venues dominate the clusters. Maybe it's the ubiquitous nature of fast food venues that is skewing the results. It is possible that we should exclude the fast food venues. That would put more weight on the other types of venues that might be more indicative of the character of a neighborhood.

Is k-means clustering the right technique?

Another assumption that we have made is that k-means clustering is the right tool to be using. Maybe we would get different results if we used, for example, a Gaussian Mixture Model (GMM). K-means clustering works well with spherical clusters, while GMM can work with arbitrarily shaped clusters.

Are venues the right way to characterize neighborhoods?

We have also assumed that venues are appropriate indicators for the nature of neighborhoods. It is possible that there may better ways to characterize neighborhoods, such as using demographic data.

Conclusion

The data indicates that Fort Worth neighborhoods are more diverse. They are also more numerous—which speaks to neighborhood size.

One interesting avenue of inquiry may be to examine whether the venue groups are associated with home values. We used the Zillow U.S. Neighborhoods dataset in this analysis, but Zillow also compute Home Value Indices by neighborhood. As noted in the discussion, it may be fruitful to collect demographic information and investigate the relationship among neighborhood venues, demographic data and home values.

One final thought is that perhaps data science is not well suited to answer this particular question.

Dallas was established as a trading post at the intersection of some Caddo Indian trails on the Trinity floodplain around the same time that Fort Worth was established as an army outpost on a bluff overlooking the Trinity River in 1849.

As a stop on the legendary Chisholm Trail, cattle and ranching were always a big part of Fort Worth's heritage, earning it the nickname of Cowtown. In 1876, the Texas and Pacific Railway was finally completed, stimulating a boom and transforming the Fort Worth Stockyards into a premier center for the cattle wholesale trade. When the railroad came through Dallas, it brought Sanger Brothers department store, E.M. Kahn clothiers and others making Dallas an important regional shopping locale and distribution point.

Fort Worth had “Hell’s Half-Acre”, the biggest collection of saloons, dance halls, and bawdy houses south of Dodge City, while Dallas had “the Park Cities”--University Park and Highland Park--where the tony Neiman-Marcus department store had its roots.

Fort Worth had the Southwestern Exposition and Livestock Show, but Dallas had the Texas State Fair. In 1914, Dallas was named the site of a regional Federal Reserve Bank and within 20 years, Dallas was the financial center of the Southwest.

Maybe that sheds some light on why BMW’s are more common in Dallas, while pickup trucks are more common in Fort Worth.

Maybe it’s not the venues that define the cities at all. Maybe it’s the people. And the history.

References

Mahendru, Khyati (2019). How to Determine the Optimal K for K-Means?, https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

Sarker, Tirhajyoti (2019). Clustering metrics better than the elbow-method, https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6

Scikit-lean. Selecting the number of clusters with silhouette analysis on KMeans clustering, https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Yadav, Jyoti (2019). Selecting optimal number of clusters in KMeans Algorithm (Silhouette Score), https://medium.com/@jyotiyadav99111/selecting-optimal-number-of-clusters-in-kmeans-algorithm-silhouette-score-c0d9ebb11308