Coursera
Applied Data Science Capstone Project

Battle of the Neighborhoods
By: Susan Terrillion, July 2020

Purchasing rental properties in the Dallas-Fort Worth, TX metro area

Introduction

The **Dallas/Fort Worth Metroplex**, commonly referred to as **DFW** or North Texas, encompasses 13 counties and is home to just over 7 million people. The **DFW** area encompasses more than 9,200 square miles of total area, making it the largest inland metropolitan area in the United States.

DFW is home to almost 25 Fortune 500 companies, the third-largest concentration of Fortune 500 in the United States behind New York City (70) and Chicago (34). In 2016, the metropolitan economy surpassed Houston to become the fourth-largest in the U.S. Currently the Metroplex boasts a GDP of just over $620.6 billion in 2020. In 2020, it was recognized as the 36th best metro area for STEM professionals.

The Dallas–Fort Worth metroplex also hosts the highest concentration of colleges and universities in Texas. The UT Southwestern Medical Center is home to six Nobel Laureates and was ranked No. 1 in the world among healthcare institutions in biomedical sciences. DFW is also the second most popular metropolis for megachurches in Texas, ranked the largest Christian metropolitan statistical area in the U.S., and has one of the largest LGBT communities in Texas since 2005.[1]

With all of the cultural, educational, and professional opportunities in Dallas-Fort Worth, the residents there hail from all over the world. The cities offer just about everything that most people are looking for in an urban lifestyle. Both Dallas and Fort Worth are home to the typical city offerings of restaurants, night clubs, the arts, and parks and a few waterways. The population of DFW is increasing by 1200 people a day with newcomers usually preferring to rent upon arrival.

Investors are always looking for a great place to park their money and the Dallas-Fort Worth area is a likely place to invest in rental properties. 57% of the population is currently living in rental units in the DFW area, which makes it ripe for investment opportunities.

Business Problem

The DFW area is spread out over 30 miles which makes it hard for a potential investor to truly get a feel for the different cities where rental investment opportunities may exist. Rental properties which provide the most An ad-hoc analysis via Yelp or a similar tool could provide information on the various amenities available but the information is not quantifiable enough to allow for a trusted analysis.

Our objective here is to cluster and analyze data from both Dallas and Fort Worth. We're looking to answer the question: "Which of two cities, Dallas or Fort Worth, offers the greatest number of amenities for its residents and in which zip codes do those amenities exist?".

This analysis is intended to be used by investors interested in purchasing in the DFW area real estate market. It may also be useful to developers who are interested in converting existing structures into new apartments.

Data sources

The following items were collected for this analysis to be performed.

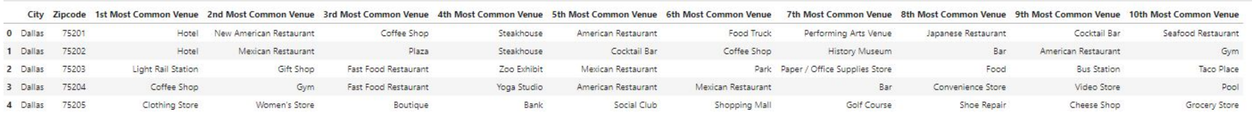
1. For each of the cities, the zip codes and longitude/latitude values were gathered using Opendatasoft's public data and fetched as a .CSV file.

2. We also looked at the venue information which includes the establishments located in each city. The FourSquare API was used to fetch all of the venues available in each defined segment in JSON. This data was summarized by venue category to support the investment opportunity analysis provided here.

These data sources will allow us to segment, cluster and map similar zip codes in the two cities.

Data Cleansing

1. Data was loaded from the Opendatasoft's Dallas .CSV file to a Pandas dataframe for easy manipulation.
2. The data was scrubbed to only include the data required for this analysis. The elements remaining in the dataframe include only ZIP CODE, CITY, LONGITUDE, and LATITUDE.
3. Duplicate zip codes were dropped from the Dallas dataframe where there were multiple rows with the same longitude and latitude values.
4. Data was loaded from the Fort Worth .CSV file to a Pandas dataframe, the same way the Dallas data was loaded.
5. Duplicate zip codes were dropped from the Fort Worth dataframe where there were multiple rows with the same longitude and latitude values.
6. Both dataframes were merged for this analysis to combine the cities. There were a total of 82 zip codes - 52 in Dallas and 32 in Fort Worth.



	City	Zipcode	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	9th Most Common Venue	10th Most Common Venue
0	Dallas	75201	Hotel	New American Restaurant	Coffee Shop	Steakhouse	American Restaurant	Food Truck	Performing Arts Venue	Japanese Restaurant	Cocktail Bar	Seafood Restaurant
1	Dallas	75202	Hotel	Mexican Restaurant	Plaza	Steakhouse	Cocktail Bar	Coffee Shop	History Museum	Bar	American Restaurant	Gym
2	Dallas	75203	Light Rail Station	Gift Shop	Fast Food Restaurant	Zoo Exhibit	Mexican Restaurant	Park	Paper / Office Supplies Store	Food	Bus Station	Taco Place
3	Dallas	75204	Coffee Shop	Gym	Fast Food Restaurant	Yoga Studio	American Restaurant	Mexican Restaurant	Bar	Convenience Store	Video Store	Pool
4	Dallas	75205	Clothing Store	Women's Store	Boutique	Bank	Social Club	Shopping Mall	Golf Course	Shoe Repair	Cheese Shop	Grocery Store

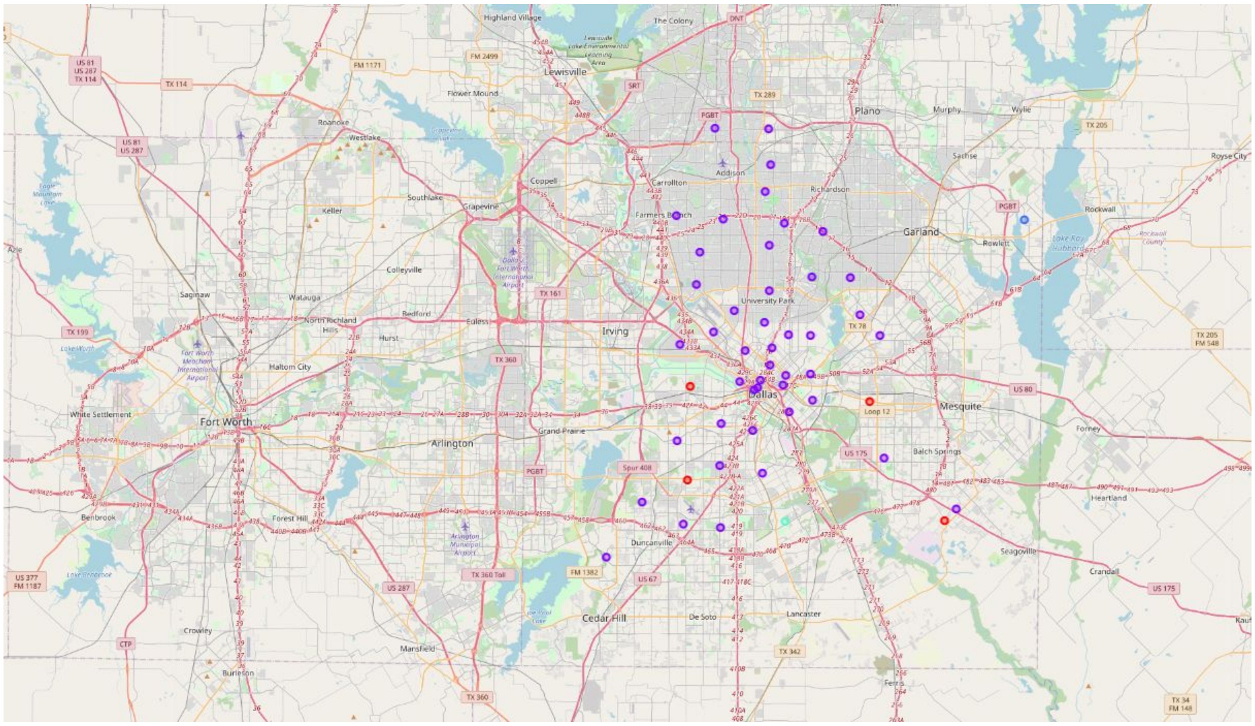
7. Venue data from FourSquare was returned with the venue, the latitude/longitude coordinates, and category elements.

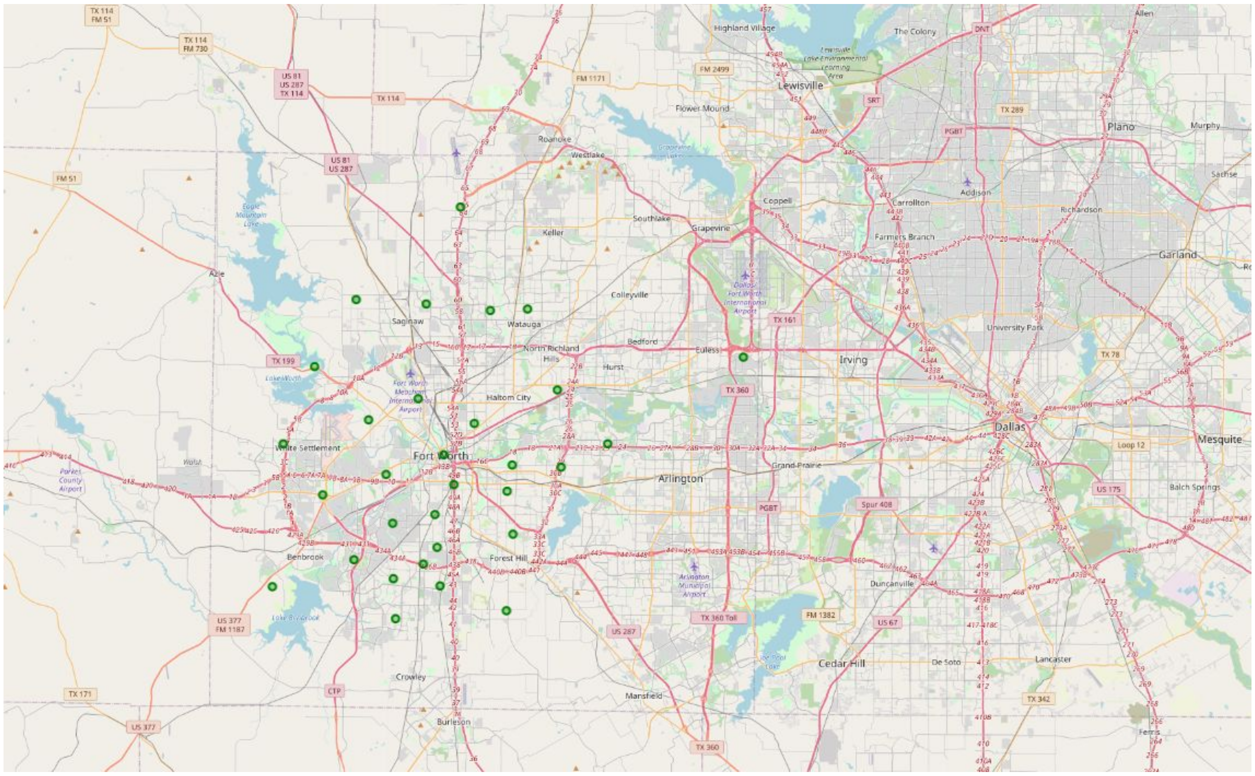
8. The data returned from FourSquare was sorted by category along with the count for each.
9. Onehot encoding was used to convert the category items into numerical values to apply k-means clustering to the sample.
10. Each row was grouped by zip and categories given a frequency.

Methodology

The data cleansing process prepared the Foursquare data for analysis and was used as input to the k-means clustering algorithm. K-means clustering partitions the data into distinct clusters with similar traits via an iterative algorithm which creates unsupervised learning to create the clusters.

After analysis of the source data, it was determined that 8 clusters were appropriate for clustering the zips represented in the data set. Prior to clustering, both zip and city were eliminated from the analysis to focus common venue data to be used in the results. After clustering, the cluster labels were merged into the DFW set with all values present. Geopy was then used to ascertain the coordinates for each city for generating a map of the city and it's clusters. A map for each city was then generated to visualize the clusters of zip codes.


Highland Village
The Colony
Lewisville
Flower Mound
Plano
Murphy
Wylie
Roanoke
Westlake
Coppell
Carrollton
Addison
Richardson
Sachse
Southlake
Grapevine
Keller
Farmers Branch
Garland
Rowlett
Rockwall
Royse City
Colleyville
Dallas Fort Worth International Airport
Saginaw
Watauga
North Richland Hills
Bedford
Euless
Irving
University Park
Hurst
Haltom City
Fort Worth Meacham International Airport
Dallas
Mesquite
Forney
White Settlement
Fort Worth
Arlington
Grand Prairie
Balch Springs
Heartland
Benbrook
Forest Hill
Arlington Municipal Airport
Duncanville
Seagoville
Crandall
Cedar Hill
De Soto
Lancaster
Crowley
Mansfield
Burleson
Ferris
Azle
Lake Worth
Eagle Mountain Lake
Lake Benbrook
Joe Pool Lake
Lake Ray Hubbard
Rockwall County
Lewisville Lake Environmental Learning Area
US 81
US 287
TX 114
TX 199
TX 360
TX 161
TX 289
TX 78
TX 205
TX 342
TX 34
FM 1171
FM 2499
FM 1382
FM 148
FM 548
FM 1187
US 377
US 175
US 80
US 67
Loop 12
Spur 408
PGBT
SRT
DNT
CTP
TX 360 Toll



Highland Village
The Colony
Allen
Lewisville
Flower Mound
Plano
Murphy
Roanoke
Westlake
Addison
Sachse
Coppell
Carrollton
Richardson
Eagle Mountain Lake
Keller
Southlake
Grapevine
Farmers Branch
Garland
Azle
Colleyville
Dallas Fort Worth International Airport
Saginaw
Watauga
University Park
North Richland Hills
Bedford
Euless
Irving
Hurst
Lake Worth
Fort Worth Meacham International Airport
Haltom City
Dallas
White Settlement
Fort Worth
Grand Prairie
Mesquite
Arlington
Hudson Oaks
Parker County Airport
Balch Springs
Benbrook
Forest Hill
Arlington Municipal Airport
Duncanville
Lake Benbrook
Cedar Hill
De Soto
Lancaster
Crowley
Mansfield
Burleson
Ferris
TX 114
FM 51
FM 730
US 81
US 287
FM 1171
FM 2499
DNT
SRT
PGBT
TX 289
TX 161
TX 199
TX 360
TX 78
Loop 12
US 175
Spur 408
FM 1382
US 67
TX 360 Toll
US 377
FM 1187
TX 171
CTP
TX 342

Results

The k-means clustering results showed the largest cluster to include 36 similar zip codes, 19 clusters in Dallas and 17 clusters in Fort Worth. The second largest cluster included 29 zip codes, with Dallas showing 24 of those and Fort Worth with only 5 zips. The next largest cluster included 13 zip codes with 6 zip codes in Dallas and 7 in Fort Worth. The five additional clusters contained only one zip code, making those clusters dissimilar to the other 3 clusters. Of these 5, Dallas was shown to have 3 locations and Fort Worth only 2.

Conclusion

The analysis done here showed that *Dallas* is a more attractive investment opportunity than *Fort Worth* based on the number of venues identified for each city.

As more venues are added to the FourSquare data, we can expect that the analysis will yield an increasingly accurate representation of the venues in the densest parts of the DFW Metroplex. A future analysis may include the introduction of purchase v. rental value for the densest clusters.