Dallas Neighborhoods Analysis—a Combination of Geospatial and Demographical Approaches

March 12, 2020

1.Introduction:

According to numbers released by the U.S. Census Bureau, Three of the 15 fastest-growing large cities — defined by the bureau as having populations of more than 50,000 — are in North Texas, as is the city with the third-highest numeric population growth. So it should be no surprise that U-Haul calculated the Dallas neighbor as one of the top U.S. cities for growth and employment opportunities are keep increasing with the booming of Dallas areas.

My Friend, XY, has been offered a great opportunity to work for a leading firm in Dallas city. He is very excited at this opportunity while a little bit worried about the new environment he is going to move to. Acknowledged that I have learned some geo data analysis skills, he approached me for help. So the key question is: How can my friend find a convenient and enjoyable place to live in Dallas city?

My friend and I sit down and set some basis to select a neighborhood in Dallas city for his new home. We both agree that it is vital to look at some demographic statistics in the neighborhood such as average housing prices, average commute time, populations statistics, etc. Also, my friend is a outdoor person who likes spending his time outside in the fresh air. his favorite is to walk or running in a park nearby after work. So a home with parks around would be an ideal place to live.Thus, the objective is to locate and recommend to the my friend which neighborhood of Dallas city will be best choice to relocate. And my friend expects me to provide the rationale in this challenging, and sometimes confusing task.

With the booming of Dallas economy and its population, this analysis is also applicable for anyone interested in exploring new opportunities in any city. The success criteria of the

project will be a good recommendation of Neighborhood choice to XY based on neighborhood avenues and demographics of the neighborhoods.

.

2. Data acquisition and cleaning

The data acquired for this project is a combination of data from three sources:

2.1. Scraping Dallas Neighborhood Table from website

I first make use of Dallas neighborhoods guides page on 'https://neighborhoods.dmagazine.com/' to scrap the table to create a data-frame. For this, I've used requests and Beautifulsoup4 library to create a data-frame containing name of the 57 neighborhoods of Dallas city with some important demographic variables including: neighborhood , population, area, population density, median age, percent of 18+ age, household income, median home value, owner percent, commute time.

2.2. Getting Coordinates of Neighborhoods by using Geocoder package :

When we have the names of neighborhoods, the next objective is to get the coordinates of these 57 major neighborhoods using geocoder class of arcgis function.

2.3. Using Foursquare Location Data:

Foursquare data is very comprehensive and it powers location data for Apple, Uber etc. For this business problem I have used, as a part of the assignment, the Foursquare API to retrieve information about the popular spots around these 57 neighborhoods. Here I've chosen 100 popular venues for each neighborhood within a radius of 1 km.

The processing of these data will help in answering some key questions : What is the neighborhoods with parks or outside activities easily accessible? What population constititue each neighborhood? How long is the average commute time for the neighborhood? Other interesting findings of the neighborhoods from the data overall?

The insights derived from analysis of neighborhoods will give good understanding of the avenues which help in XY's decisions to target a new home.

Data cleaning:

All the values scraped from the website are object type. I noticed there were a couple of special characters such as $, % populated as part of the string values. In order to conveniently transform them as float numbers using pandas to-numeric function, I have decided to remove those special values from the dataframe, as well as the number separator ‘,’.
Data downloaded or scraped from multiple sources were combined into one table for further analysis.

3. Methodology

Python package BeautifulSoup was used to scrape a list of neighborhood from Dallas magazine website and table was organized to include all the important variables for deciding a home location.

And I also used Geopy to get the geological location of each neighborhood. Some neighborhoods has more than one responses and the longitude and latitude are close to each other. I will only keep the 1st records to avoid any duplications.

The final outputs with latitude and longitude:

	Neighborhood	Latitude	Longitude
0	Bent Tree	32.973411	-96.826306
1	Bluffview	32.976402	-96.908401
2	Casa Linda	34.051474	-117.231948
3	Central Dallas	32.776272	-96.796856
4	Bent Tree	32.973411	-96.826306

I utilized the Foursquare API to explore the neighborhoods and segment them. I designed the limit as 100 venues and the radius 1000 meters for each borough from their given latitude and longitude information. Here is the header of the result, adding venue id, venue name, category, latitude, and longitude information from Foursquare API.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Bent Tree	23	23	23	23	23	23
Bluffview	41	41	41	41	41	41
Casa Linda	11	11	11	11	11	11
Casa View	36	36	36	36	36	36
Central Dallas	100	100	100	100	100	100
Deep Ellum - Expo Park	100	100	100	100	100	100
Design District	100	100	100	100	100	100
Devonshire	76	76	76	76	76	76
Downtown	100	100	100	100	100	100
East Dallas	77	77	77	77	77	77
Eastwood	24	24	24	24	24	24
Far North Dallas	32	32	32	32	32	32
Forest Hills	23	23	23	23	23	23

For each neighborhood, I have calculated the average number of venues.

	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Res
0	Bent Tree	0.0	0.00000	0.043478	0.0	0.130435	0.043478	0.0	0.0	0.0	0.0	0.0	0.0	
1	Bluffview	0.0	0.02439	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	
2	Casa Linda	0.0	0.00000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	
3	Casa View	0.0	0.00000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	
4	Central Dallas	0.0	0.00000	0.000000	0.0	0.000000	0.040000	0.0	0.0	0.0	0.0	0.0	0.0	

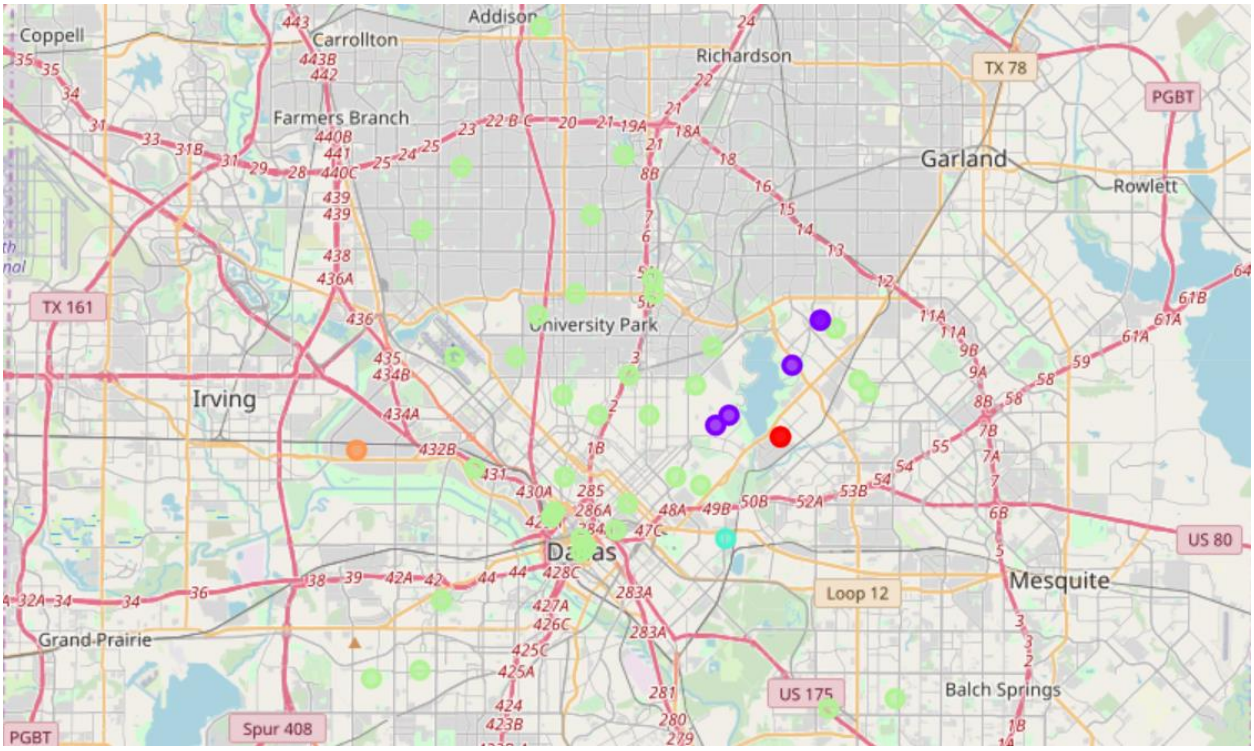
I have explored the data to find out what are the common venues for each neighborhood to get a quick impression first of Dallas neighborhoods. This may give an idea what the final cluster could be.

	Neighborhood	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	
0	Bent Tree	Airport Terminal	Spa	Intersection	History Museum	Italian Restaurant	Shipping Store	Food Truck	Burger Joint	Sandwich Place	
1	Bluffview	Indian Restaurant	Ice Cream Shop	Korean Restaurant	Fried Chicken Joint	Bubble Tea Shop	Burger Joint	Supermarket	Lake	Bookstore	
2	Casa Linda	Park	Sandwich Place	Italian Restaurant	Fast Food Restaurant	Mediterranean Restaurant	Snack Place	Dessert Shop	Grocery Store	Mexican Restaurant	C
3	Casa View	Mexican Restaurant	Chinese Restaurant	Thrift / Vintage Store	Pizza Place	Locksmith	Sandwich Place	Pharmacy	Bank	Dessert Shop	
4	Central Dallas	Hotel	Coffee Shop	Bar	American Restaurant	Cocktail Bar	Mexican Restaurant	Plaza	Park	Italian Restaurant	

Clustering Approach: I have pulled 57 neighborhoods in total in Dallas city. In this project, the 1st part is to clustering of 57 neighborhood using venues in 1000 meters around the longitude and latitude of each neighborhood. The goal of this cluster analysis to find the areas that is more accessible to parks. And second part is re-clustering of the neighborhoods with demographic variables, which is particular important when considering a home purchase.

e	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	9th Most Common Venue	10th Most Common Venue
0	1	Park	Sandwich Place	Italian Restaurant	Fast Food Restaurant	Mediterranean Restaurant	Snack Place	Dessert Shop	Grocery Store	Mexican Restaurant	Coffee Shop
0	1	Park	IT Services	Pharmacy	Food	Video Store	Chinese Restaurant	Fast Food Restaurant	Pizza Place	Cosmetics Shop	Dive Bar
0	1	Park	Fishing Spot	Lake	Boat or Ferry	Trail	Golf Course	Harbor / Marina	Scenic Lookout	Dive Bar	Dog Run
0	1	Locksmith	Golf Course	Park	Zoo	Dog Run	Donut Shop	Dry Cleaner	Electronics Store	Ethiopian Restaurant	Discount Store
0	1	Park	IT Services	Pharmacy	Food	Video Store	Chinese Restaurant	Fast Food Restaurant	Pizza Place	Cosmetics Shop	Dive Bar
0	1	Trail	Convenience Store	Shopping Mall	Gym	Park	Light Rail Station	Fast Food Restaurant	Pizza Place	Fried Chicken Joint	Donut Shop
0	1	Park	Bakery	Harbor / Marina	Gas Station	Used Bookstore	Italian Restaurant	Tex-Mex Restaurant	Beer Bar	Public Art	Wine Bar

Visualization of clusters and dallas map has validated the cluster analysis based on venues. we can see from the map that cluster 1 neighborhood most located around lakes, where are usually ideal sites for beautiful parks establishment.



dcluster	Pop_density	Median_age	age18+	Household_income	Median_home_value	Owner percent	Commute_time
0	5283.071429	32.457143	0.723143	42252.714286	180224.000000	0.446071	29.000000
1	4122.937500	41.595833	0.777208	85448.541667	328356.458333	0.754583	25.791667
2	6218.800000	34.166667	0.900667	66325.833333	281634.333333	0.143833	22.166667
3	4522.600000	37.350000	0.738500	159716.250000	964580.000000	0.787000	19.000000
4	3268.350000	30.700000	0.685000	35234.500000	89608.000000	0.559000	32.500000
5	2337.216667	49.741667	0.842283	101305.833333	646573.000000	0.613850	22.333333

So, to summarize the clusters based on the demographic information: clusters labels are aligned head to head to give my friend a better idea which clusters could be their dream neighborhood. from least populated to most popuated: 5,4,1,3,0,2 from youngest to eldest community: 4,0,2,3,1,5 from highest income to lowest income: 3,5,1,2,0,4 from highest averge houseing price to lowest: 3,5,1,2,0,4 most renter to less renters : 2,0,4,5,1,2
from less commute time to most : 3,2,5,1,0,4 based on differnt standards, the lists above ranked each neighborhood on a sinle measurement. we have to

play the trade off here to select an ideal neighborhood which is not too expansive to buy a home, which is not over populated, which has less commute time and not many people moved in and out frequently. My friend and I agree cluster 1 would be a better choice, if not best.

4. Results

From the demographic clustering, I have pulled following neighborhoods as potential candidates for my friend's new home. These neighborhoods are:

'Bluffview', 'Casa Linda', 'Eastwood', 'Forest Hills', 'Hillside', 'Hollywood Heights - Santa Monica', 'Kiestwood', 'Lake Park Estates', 'Lakewood', 'Lakewood Heights', 'Little Forest Hills', 'Lochwood', 'Love Field', 'Lower Greenville', 'M Streets - Vickery Place', 'Midway Hollow', 'Northaven Park', 'Northwood Hills - Valley View', 'Old Lake Highlands', 'Preston Highlands', 'Prestonwood', 'Ridgewood Park', 'The Peninsula', 'Wilshire Heights'.

From the venue clustering, I have pulled following neighborhoods which are suitable for outdoor people like my friend. These neighborhoods include:

'Casa Linda', 'Lake Highlands', 'Lakewood', 'Lakewood Heights', 'Old Lake Highlands', 'Southwest Dallas', 'The Peninsula'.

The final selection would be the overlap of two clustering methods based on different attributes. As a result, I will recommend following neighborhood for XY's dream house. They are: 'Old Lake Highlands', 'The Peninsula', 'Lakewood', 'Lakewood Heights', 'Casa Linda'.

5. Discussion

As a recommendation to those who plan to move to new cities, location selection is a fundamental problem to think over. The analysis of this report used XY' s selection criteria as variables to be used in clustering.

It can not solve the problem of one particular client who has other standards at selecting their new homes. However, this project certainly gives us some very important preliminary information on possibilities of developing a tool to help new movers in the relocation.

Although in this report, it demonstrates a nice pattern for venues clusting with parks ranked as most often seen in several neighborhood. But to be more reliable, clustering with different number of clusters should also be evaluated to using ‘elbow method’.

With all these analyses done, the report finally becomes constructive for my friend to get familiar with his new environment.

6. Conclusion

As a conclusion of this project, we have got a small glimpse of how real life data-science projects look like. I’ve made use of some frequently used python libraries to scrap web-data, use Foursquare API to explore the neighborhoods in Dallas city and saw the results of clustering using Folium leaflet map. Potential for this kind of analysis in a real life business problem is discussed in great detail. Also, some of the drawbacks and chance for improvements to represent even more realistic pictures are mentioned.