WHERE DO WE BUILD?

Analysis of Dallas/Fort Worth Neighborhoods to Target Location of Catering Business

Coursera Data Science Specialization - Capstone Project

Battle of the Neighborhoods

Prepared by DavidL

September/October 2019

Table of Contents

Introduction (A description of the problem and a discussion of the background) 1

Data 1

Methodology (A description of the data and how it will be used to solve the problem.) 2

Results 2

Discussion 3

Conclusion 3

Acknowledgements 3

Introduction
(A description of the problem and a discussion of the background)

Our hypothetical client is a Catering firm (FoodForAll, or FFA) looking to establish a location in the Dallas/Fort Worth (DFW) metroplex region in the northern part of Texas in the United States. The DFW region is a sprawling metropolitan area comprised of 2 major cities (Dallas and Fort Worth) and many smaller cities/communities. We will limit our analysis to the cities of the 2 major counties in the area, Tarrant County and Dallas County.

FFA is a startup catering company being created with the intent to serve primarily business functions and, to a lesser degree, residential clients. The client requests an analysis of the neighborhoods in the DFW area to help determine a location for their company location.

The primary factors for the analysis should include:

- Proximity to multiple businesses within an approximately 5 mile radius of the selected location
- Lower number of nearby restaurants/other food service locations
- A secondary factor is the number of parks and community use areas nearby where residential customer parties could be catered

Data

The Acknowledgement section below lists the details most of the data points used in this analysis. As specified above, the study focused on the counties of Dallas and Tarrant, representing the bulk of the population in the Dallas/Fort Worth metroplex.

Our first source of data was to the individual county sites (Sources #1 and #2 in Acknowledgments) to determine which cities were included in each country. These were listed for Dallas county on the “About Us” page, and for Tarrant county on the “Incorporated Areas” page.

The next, most relevant crucial source was the extremely helpful North Central Texas Council of Governments website (Source #3 in Acknowledgements). This site hosts a variety of location/shape GIS files. For the purposes of this study, the “**City Limits (2010 Census)**” data proved to be most useful. This dataset lists all of the cities in the North Texas area, including other counties besides the two that I focused on. Among the many features listed in this dataset, I was able to find the Latitude and Longitude for each of the city centers that I was researching.

A key source of information for this assignment was the Explore functionality of the FourSquare API (Source #4 in Acknowledgements). The explore API takes a location (specified by Latitude/Longitude) and a radius (in meters) and returns a list of features (restaurants, parks, retail, etc) within that radius of the given location.

Finally, because we’re creating a Recommendation engine, I created a dataset to represent the preferences for my hypothetical client. This is a one row record containing the rankings (from 0=Low to 10=High) of each of the features represented in the search radius of the cities being researched.

Methodology
(A description of the data and how it will be used to solve the problem.)

Our methodology that we use in our analysis of the neighborhood data will include several tools, techniques and sources learned in the Data Science specialization coursework.

The ultimate output of this analysis will be a list of 5 cities which are the best fits for our client's preferences for locating their new Catering business. The study we did on Recommendation Engines seems like the most relevant model to explore for this. The primary steps involved are described below:

1. Gather city/location data
    a. Here, I gathered city data from the websites into a dataframe and removed any duplicates because there were some cities that were on boundaries of both counties and listed in both cases.
    b. Then I created a new dataset to store all north Texas cities along with their location data.
    c. Then I created a new dataframe by joining the researched cities with the appropriate location data
2. Query Location data for Features and Transform
    a. I used the FourSquare API to pull in the first 100 items within an 8000m (approx. 5miles) radius of each of the city centers
    b. I aggregated/summarized and transformed the data into a table with a row representing each city – each feature/column represented the counts of each feature available within the searched radius (in this case 8000m) of the city center.
3. Determine Client preferences
    a. Created a survey for the client to update their preferences for each feature type represented in the data. The key factors were that the hypothetical client did not want to be too close to too many restaurant/food service outlets to minimize the competition, so these features were rated with a 0. There were some features that were highly desirable because of the anticipated customer demand (Medical Center, Harbor/Marina, Hotel, etc), so these were ranked with a 10. Several other features were ranked between these two extremes on the spectrum and were ranked accordingly to try to maximize possible catering sales.
4. Compare Preferences to weighted city feature list to develop an index used to rate cities (high to low)

Results

Neighborhood	Rating
Newark	3.71
Fort Worth	3.27
Azle	2.98
Pelican Bay	2.98
Dallas	2.85

The above are the top 5 (out of 70 north Texas cities) when the features are ranked in accordance to the customers preferences.

Discussion

There was some surprise at the composition of the top cities in the result set. Some cities that were expected to be higher, such as Addison and Irving were not in the top 5. The expectation was that because of the city populations, they might be good candidates. However, upon deeper review of the data, a huge disqualification was the fact that these cities already host many restaurants which were rated as 0 in the client survey.

Although there is data to support the results achieved, there are several areas to follow up with this analysis:

1. Incorporate population of each of the cities to assist with segmentation. Without that data in this analysis, we don't know if the suggested areas could support another business – this would require some additional work.
2. Possibly increase the weighting of any feature categories to ensure proper consideration in scoring. In some cases, it looks like some cities were ranked higher due to a higher quantity of mid-tier possible customers (per the client's preference survey). We would have to look at ways to allow more desired options to influence the scoring accordingly.
3. Highly dependent on the Client's feature ratings. With the quantity of features in the data set, better ways to group the different types of features could help improve the recommendation score.
4. If the client had more of an idea of their target group/demographic, the analysis could include some segmenting based on population and income levels of each of the various cities.
5. This analysis currently only focuses on areas within approx. 5 miles of the city centers. To be more accurate, we should look at businesses using a wider radius, or look at neighborhoods at a more granular basis.

Conclusion

The DFW metroplex is a large, dispersed extended community. Based on the data used, we're confident that the top 5 locations suggested would be good initial considerations for where to launch the new catering business, however, I do feel that further analysis would be warranted to see if affluence, population and/or other factors could lend more insights.

The other big consideration is related to the issue described in Discussion point #5 above – the limitations of looking only 5 miles within a city center. If the client truly wants to look across the entire metroplex, they may want to look at sub-neighborhoods within each city; this may be a good second stage view once the client narrows their choices for the locations.

Acknowledgements

Below is a list of the data sources used in this analysis:

1. Dallas County website: Used to get the list of cities within Dallas County https://www.dallascounty.org/about-us/cities/

2. Tarrant County website: Used to get the list of cities within Tarrant County
   http://access.tarrantcounty.com/en/county/about-tarrant/incorporated-areas.html
3. North Central Texas Council of Governments – this site has GIS shape and location data for cities in the North Central Texas area (http://data-nctcoggis.opendata.arcgis.com/)
   Data used is the 'City Limits (2010 Census)' data set (http://data-nctcoggis.opendata.arcgis.com/datasets/city-limits-2010-census)
4. FourSquare API Location Data: used to find features within a certain radius of each city center
   https://api.foursquare.com/v2/venues/explore