Analyzing neighborhoods in Dallas, TX for starting a restaurant (The Battle of the Neighborhoods)

Table of contents

1. Introduction: Business Problem

Dallas is the business and financial services center for the state of Texas and has evolved into a major high-tech hub. Dallas has become a popular migrant destination, attracting residents from abroad as well as from other states. Located in North Texas, the city of Dallas is the main core of the largest metropolitan area in the Southern United States and the largest inland metropolitan area in the U.S. that lacks any navigable link to the sea. It is the most populous city in the Dallas–Fort Worth metroplex, the fourth-largest metropolitan area in the country at 7.5 million people. Given the rapid rise of this multicultural and financially booming city, it is home to numerous cuisines and flavors from all over the world. This multicultural hotpot drives people to explore new cuisines with friends from the world over and so it is no surprise that the restaurant industry is flourishing here. Having lived in the Dallas-Fort Worth multiplex (DFW) for two years during Master's, I have grown fond of the city and its people, and through this project I hope to provide potential restaurant business owners an exploratory analysis on which areas of Dallas to target to open a restaurant.

Given my love for Indian and Chinese cuisine, this project will be targeted for stakeholders interested in opening an Indian, Asian, or Chinese restaurant in Dallas.

2. Data Collection

The data required for this project has been collected from multiple sources. A summary of the data required for this project is given below.

1. Neighborhood's data

The names of the neighborhoods and areas in Dallas is scraped from https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Dallas. The data is read into a pandas data frame using the read_html() method. The neighborhoods will be referred to as a combination of Area and Neighborhood i.e., Area_Neighborhood henceforth, to avoid any chance of neighborhoods containing the same name.

2. House property value data

The latest (July 2021) house property prices of mid-tier properties are obtained from: https://www.zillow.com/research/data/

The ZHVI is defined below:
Zillow Home Value Index (ZHVI): A smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type. It reflects the typical value for homes in the 35th to 65th percentile range.

3. Geographical coordinates

The geographical coordinates for Dallas data have been obtained from the GeoPy library. This data is relevant for plotting the map of Dallas using the Folium library. The geocoder library has been used to obtain latitude, longitude, zip code data for various neighborhoods in Dallas. The geographical coordinates are then further used for plotting using the Folium library in python.

4. Shape of Dallas boundary

The outline of the Dallas boundary on the world map is obtained from a shapely file obtained from: https://www2.census.gov/geo/tiger/TIGER2019/ZCTA5/tl_2019_us_zcta510.zip
The accompanying files need to be in the same folder when reading the .shp file. Beyond that, the .shp file is converted to geojson to ensure that it can be used to plot the polygon, multipolygon and other shapes forming the Dallas map boundary.

This will be useful when augmented with HZVI data in determining which regions (by zip code) have cheaper property prices to open a restaurant in Dallas.

5. Venue data

The venue data has been extracted using the Foursquare API. This data contains venue recommendations for all neighborhoods in Dallas and is used to study the popular venues of different neighborhoods.

3. Methodology

	Area	Neighborhood	Area_type	Area_Neighborhood	Latitude	Longitude	Zip_Code
0	Downtown Dallas	Baylor District	Mixed	Baylor District, Downtown Dallas	32.778220	-96.795120	75201
1	Downtown Dallas	The Cedars	Mixed	The Cedars, Downtown Dallas	32.688525	-96.569610	75353
2	Downtown Dallas	Civic Center District	Mixed	Civic Center District, Downtown Dallas	32.778220	-96.795120	75201
3	Downtown Dallas	Dallas Arts District	Mixed	Dallas Arts District, Downtown Dallas	32.789440	-96.797170	75201
4	Downtown Dallas	Dallas Farmers Market	Mixed	Dallas Farmers Market, Downtown Dallas	32.777510	-96.788240	75201
...	...	...	...	...	...	...	...
202	West Dallas	Los Altos	Residential	Los Altos, West Dallas	32.956317	-96.806857	75248
203	West Dallas	Muncie	Residential	Muncie, West Dallas	32.774990	-96.849958	75212
204	West Dallas	Western Heights	Residential	Western Heights, West Dallas	32.690290	-96.733950	75241
205	West Dallas	Westmoreland Heights	Residential	Westmoreland Heights, West Dallas	32.647609	-96.874271	75237
206	West Dallas	Trinity Grove	Residential	Trinity Grove, West Dallas	32.655200	-96.812117	75232

207 rows × 7 columns

Dallas has 206 neighborhoods (some with similar/same names, hence the need for a column called 'Area_Neighborhood' to differentiate them) and the Latitude, Longitude, Zip Code have been found using the library GeoPy.

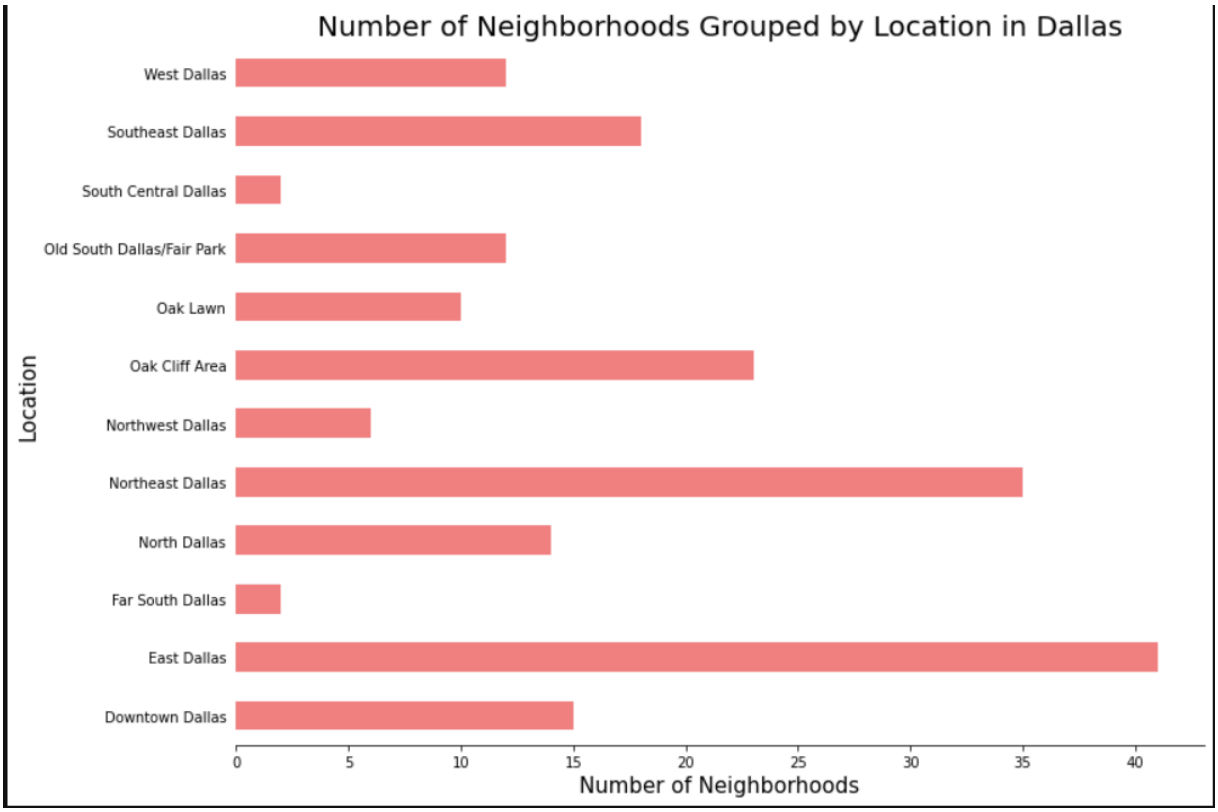
While searching for the zip codes for each neighborhood, the GeoPy library returns zip codes from places named Dallas but not in TX, despite specifying 'TX' or 'Texas' in the search queries. To avoid this error, only the neighborhoods in the Zip code list obtained from: https://www.maxleaman.com/mortgage-resources/texas-zip-code-maps/dfw-zip-code-map/ were filtered out.

This is also because the GeoPy library returns zip codes:

1. near Dallas which fall into the Dall-Fort Worth multiplex but not in Dallas

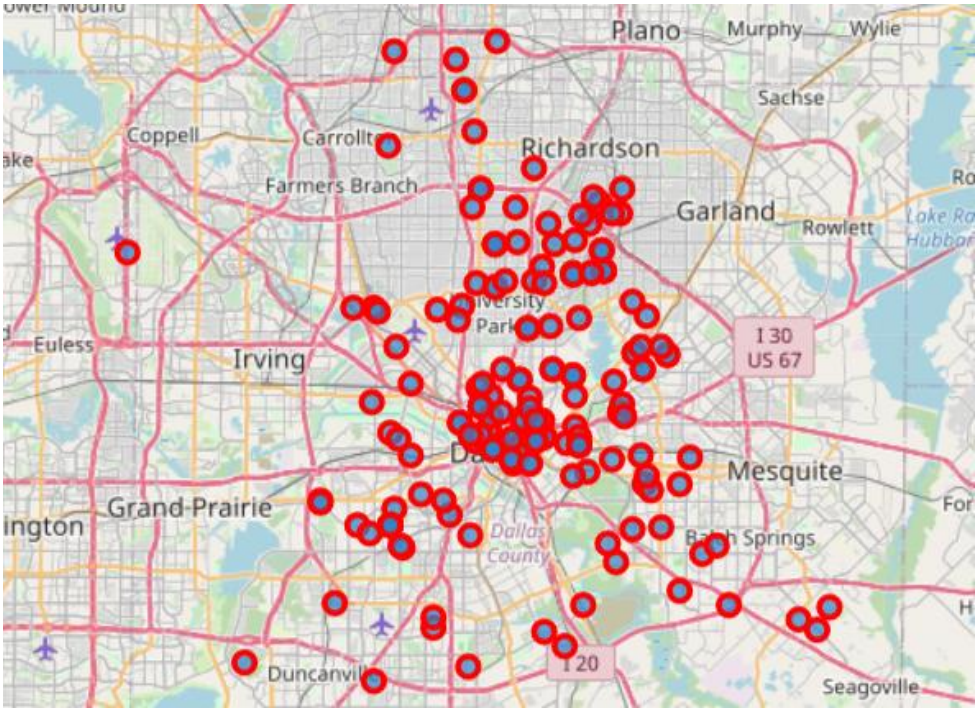
2. which are not from Dallas, TX but from other states as there are [14 places](https://geotargit.com/citiespercountry.php?qcountry_code=US&qcity=Dallas) named Dallas in USA and on occasion, GeoPy gets it wrong despite specifying 'Dallas, Texas' in the query.

This is the neighborhoods grouped by locations in Dallas:

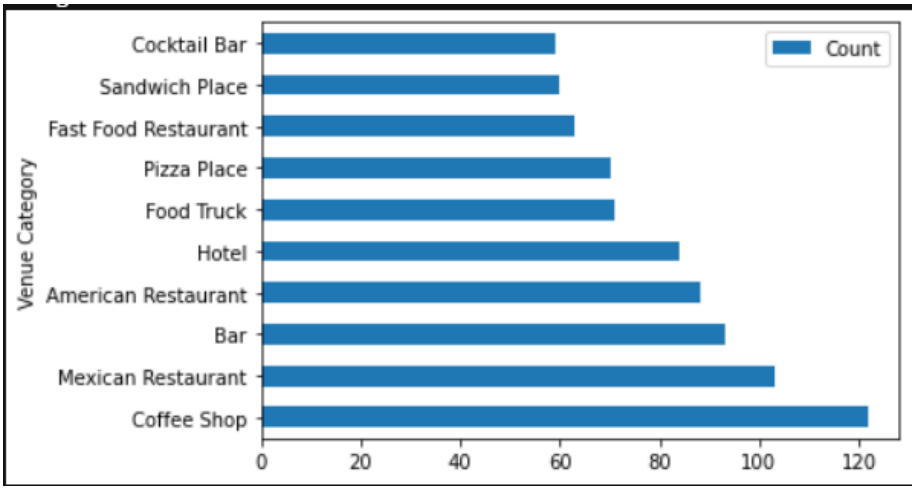


We note that East Dallas and Northeast Dallas contain the greatest number of neighborhoods which might give us insights later how the breakdown of restaurants are there in these neighborhoods.

These are the distribution of the neighborhoods in Dallas:

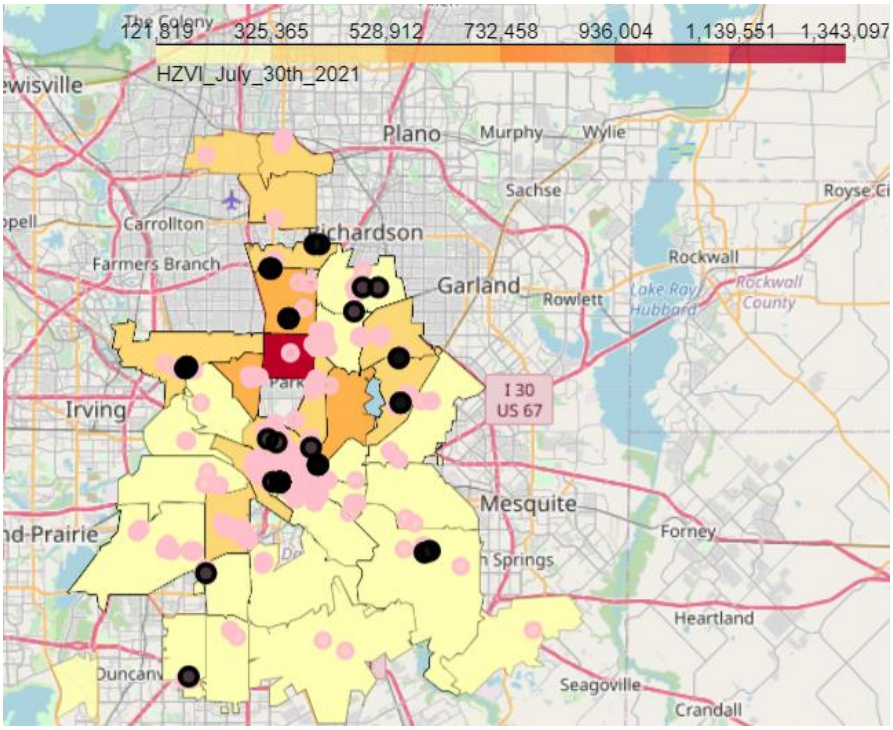


After limiting the Foursquare API's maximum venues per neighborhood to 100 and in a radius of 1 kilometer, we obtain the following Top 10 most common venue categories in Dallas:



Among all the venue categories in Dallas, Mexican Restaurant, Coffee Shop, Bar, Hotel and American Restaurant are the most frequently occurring venue categories. However, we do not treat a Bar as a type of restaurant when filtering all the restaurants despite their ability to serve limited options of food. So, we filter out that, among other venues that are not strictly restaurants to aid us in further analysis.

Next, we augment the current data frame with property prices for the latest (at the time) available prices and further add the shape file of the city of Dallas to this data frame. A choropleth map is then rendered to understand the distribution of Indian, Asian, or Chinese restaurants against all other restaurants in the differing property price landscape of Dallas as shown below:



It seems from the map that the center of Dallas is heavily populated with restaurants which may prove to a stumbling block when opening a new Indian, Asian, or Chinese restaurant. Note that this is just a preliminary visual conclusion, whereas further analytical judgements will be made in due course.

We would like to hit the sweet center of locating a neighborhood where there are:

1. candidate neighborhood is close to any of the booming restaurant neighborhoods

2. candidate neighborhood does not have any restaurants in a 300m radius (this value has been chosen since Area of Dallas is 999 sq.km)

3. candidate neighborhood has no Indian, Asian, or Chinese restaurants in a 500m radius (this value has been chosen since Area of Dallas is 999 sq.km)

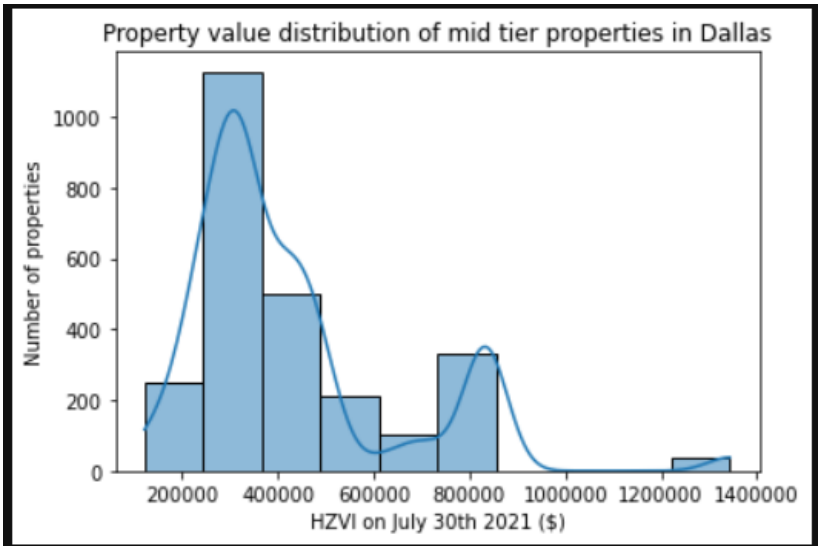
4. candidate neighborhood is in an area with relatively cheaper property price

We will present map of all such locations but also create clusters (using k-means clustering, DBSCAN clustering, HDBSCAN clustering) of those locations to identify general zones / neighborhoods which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

The above factors have been obtained from articles like this.

4. Analysis

Before we get into the nitty gritty ML clustering algorithms, looking at the average housing price in Dallas in the mid-tier range (35 to 65 percentile) yields the following distribution:



We can see that the average house price in Dallas is about $300k in the mid-tier range (35 to 65 percentile)

This can be confirmed here:

https://www.zillow.com/dallas-fort-worth-arlington-metro-tx_r394514/home-values/

"This value is seasonally adjusted and only includes the middle price tier of homes. Dallas-Fort Worth-Arlington Metro home values have gone up 16.9% over the past year and Zillow predicts they will rise 15.6% in the next year."

Next, from the review here, we find out the neighborhoods in Dallas which are the best in terms of the food scene, which is influenced by a number of other factors like: an unemployment rate of 3.4%, easy public transit with DART, and a thriving entertainment scene.

Finally, candidate neighborhoods are filtered out from the main data frame containing the neighborhoods based on the two conditions mentioned above:

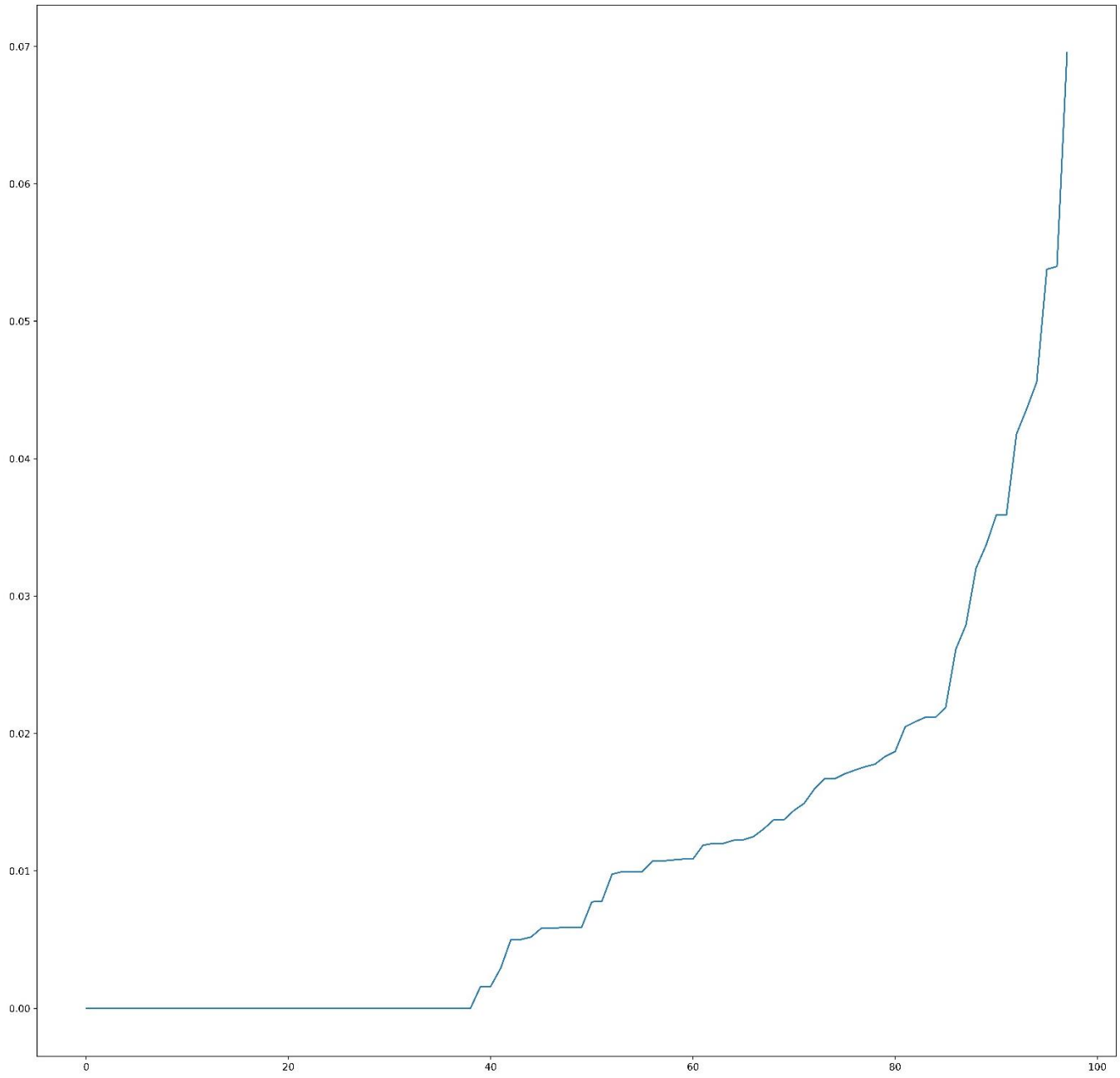
- candidate neighborhood does not have any restaurants in a 300m radius (this value has been chosen since Area of Dallas is 999 sq.km)
- candidate neighborhood has no Indian, Asian, or Chinese restaurants in a 500m radius (this value has been chosen since Area of Dallas is 999 sq.km)
- candidate neighborhood is close to any of the booming restaurant neighborhoods

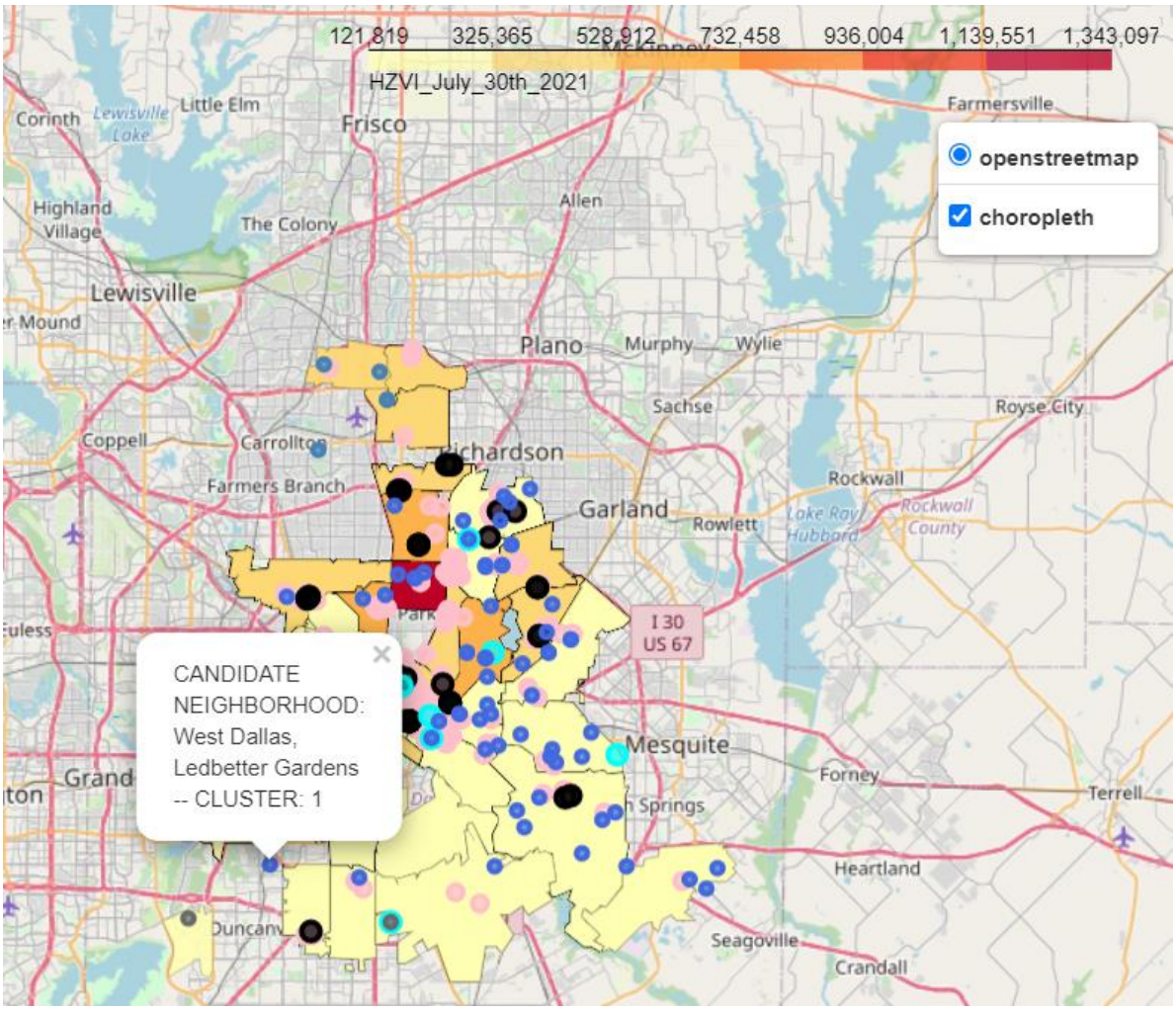
Clustering algorithms are now used to ease the identification process of candidate neighborhoods in the less expensive property zip codes overlaid in the choropleth map.

1. DBSCAN Clustering

Upon optimizing the two DBSCAN parameters (1. minPts = 4 for 2-D data) and displaying the map for clustering the candidate neighborhoods we get the map.

2. The ideal value for ε will be equal to the distance value at the “crook of the elbow”/”maximum curvature” on the graph titled ‘Points sorted by distance to the 4th nearest neighbor’. This point represents the optimization point where diminishing returns are no longer worth the additional cost. This concept of diminishing returns applies here because while increasing the number of clusters will always improve the fit of the model, it also increases the risk that overfitting will occur.





The clustering is good. However, it can be better. Far too many candidate neighborhoods belong to cluster 1 that are not in the same proximity as each other. To improve this, a variant of DBSCAN, HDBSCAN will be used.

Based on above criteria, cheaper property prices and proximity to booming food neighborhoods, potential attractive neighborhoods could be:

1. Downtown Dallas, Uptown

2. West Dallas, Trinity Grove

3. East Dallas, Lower Greenville

4. South Central Dallas, Skyline Heights

5. East Dallas, Baylor/Meadows

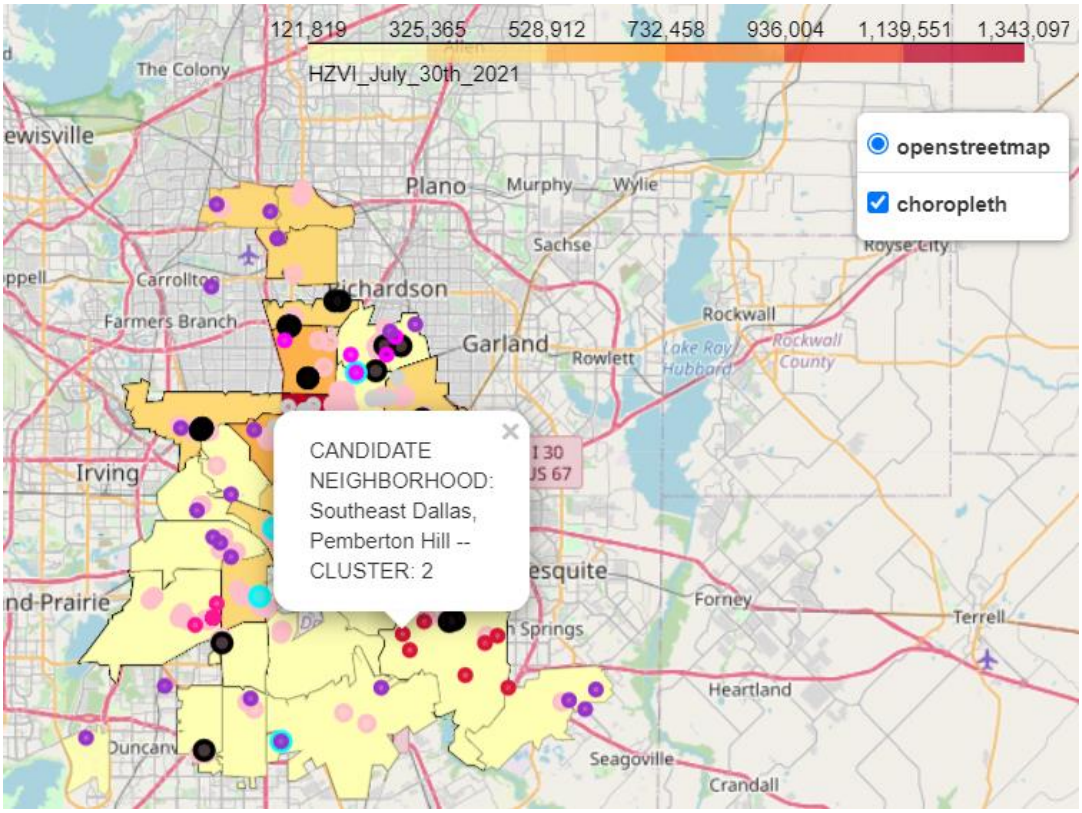
6. Southeast Dallas, Riverway Estates/Bruton Terrace

7. Southeast Dallas, Cedar Run

The remaining less attractive candidate neighborhoods, albeit still competitive may be found on the map above. Their colors can be identified from the map as 'dimgray', 'royalblue', 'steelblue'

3. HDBSCAN Clustering

Although K-means has the highest speed with increasing volume of data, for our small dataset we may use the slower DBSCAN and its variant HDBSCAN to account for its (K-means) shortcoming of not assigning a fixed number of clusters. It is also beneficial only when we know that the data has spherical cluster shapes, which our data may not.



The above breakdown of candidate neighborhoods in each cluster also indicates that HDBSCAN has done a better job in clustering them than DBSCAN.

Based on above criteria, cheaper property prices and proximity to booming food neighborhoods, potential attractive neighborhoods could be:

1. Downtown Dallas, Uptown

2. West Dallas, Trinity Grove

3. East Dallas, Lower Greenville

4. South Central Dallas, Skyline Heights

5. East Dallas, Baylor/Meadows

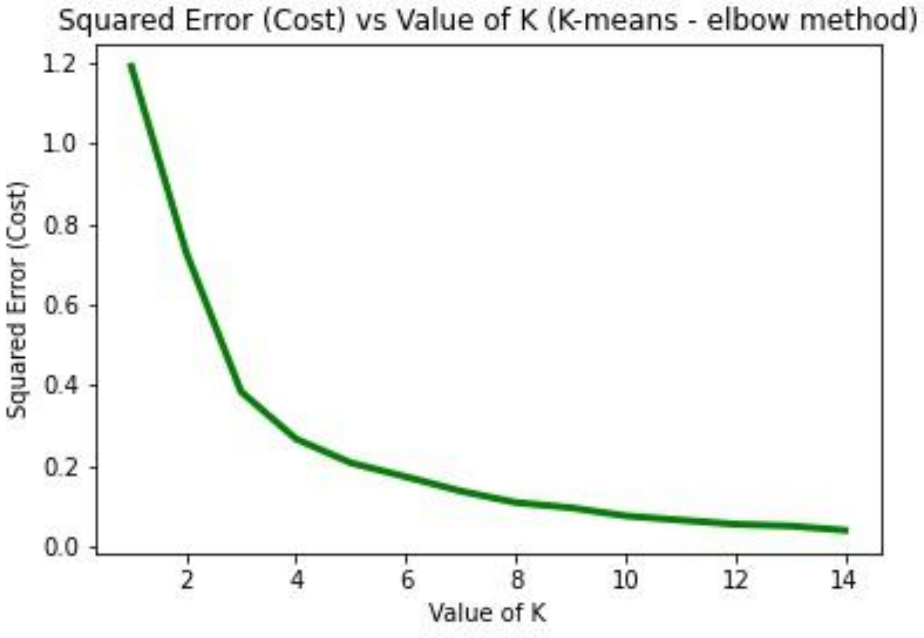
6. Southeast Dallas, Riverway Estates/Bruton Terrace

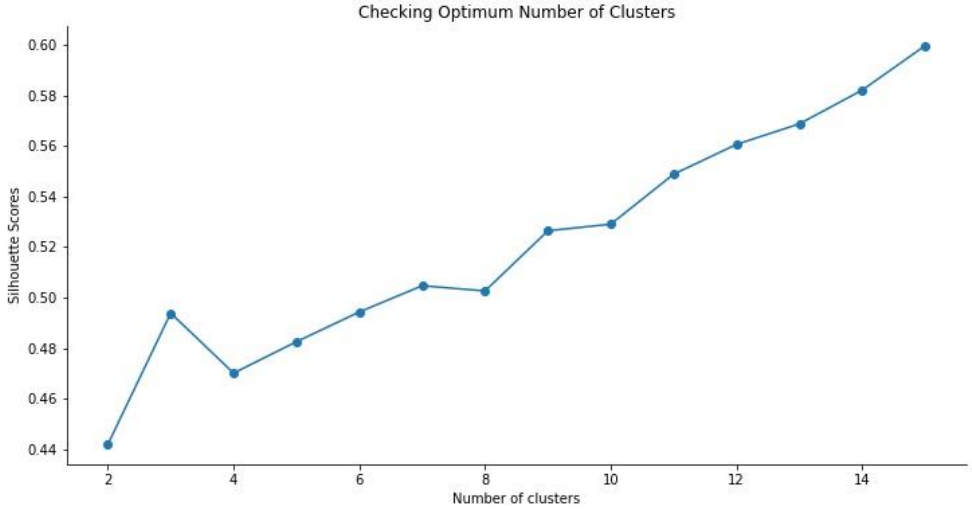
7. Southeast Dallas, Cedar Run

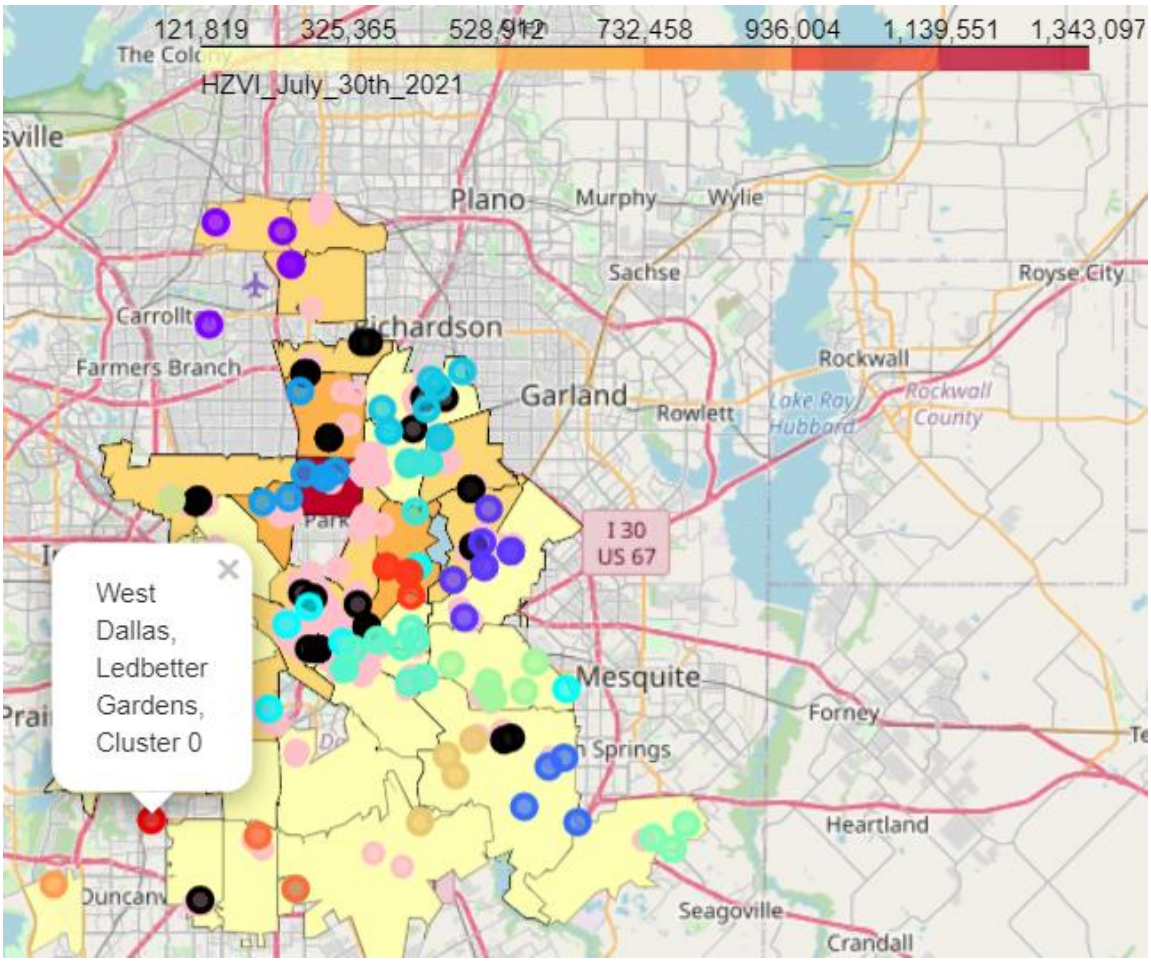
The remaining less attractive candidate neighborhoods, albeit still competitive may be found on the map above. Their colors can be identified from the map as 'darkorchid', 'deeppink', 'crimson', 'darksalmon', 'gainsboro', 'lightgrey'

4. K-Means Clustering

Using the Elbow method (naïve) and Silhouette Score, the optimal clusters K = 15 is obtained as below:


Squared Error (Cost) vs Value of K (K-means - elbow method)
Squared Error (Cost)
Value of K
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
12
14


Checking Optimum Number of Clusters
Silhouette Scores
Number of clusters
0.60
0.58
0.56
0.54
0.52
0.50
0.48
0.46
0.44
2
4
6
8
10
12
14



The above breakdown of candidate neighborhoods in each cluster from the map indicates that K-Means has done at least as good a job in clustering as compared to HDBSCAN.

Based on above criteria, cheaper property prices and proximity to booming food neighborhoods, potential attractive neighborhoods could be:

1. Downtown Dallas, Uptown

2. West Dallas, Trinity Grove

3. East Dallas, Lower Greenville

4. South Central Dallas, Skyline Heights

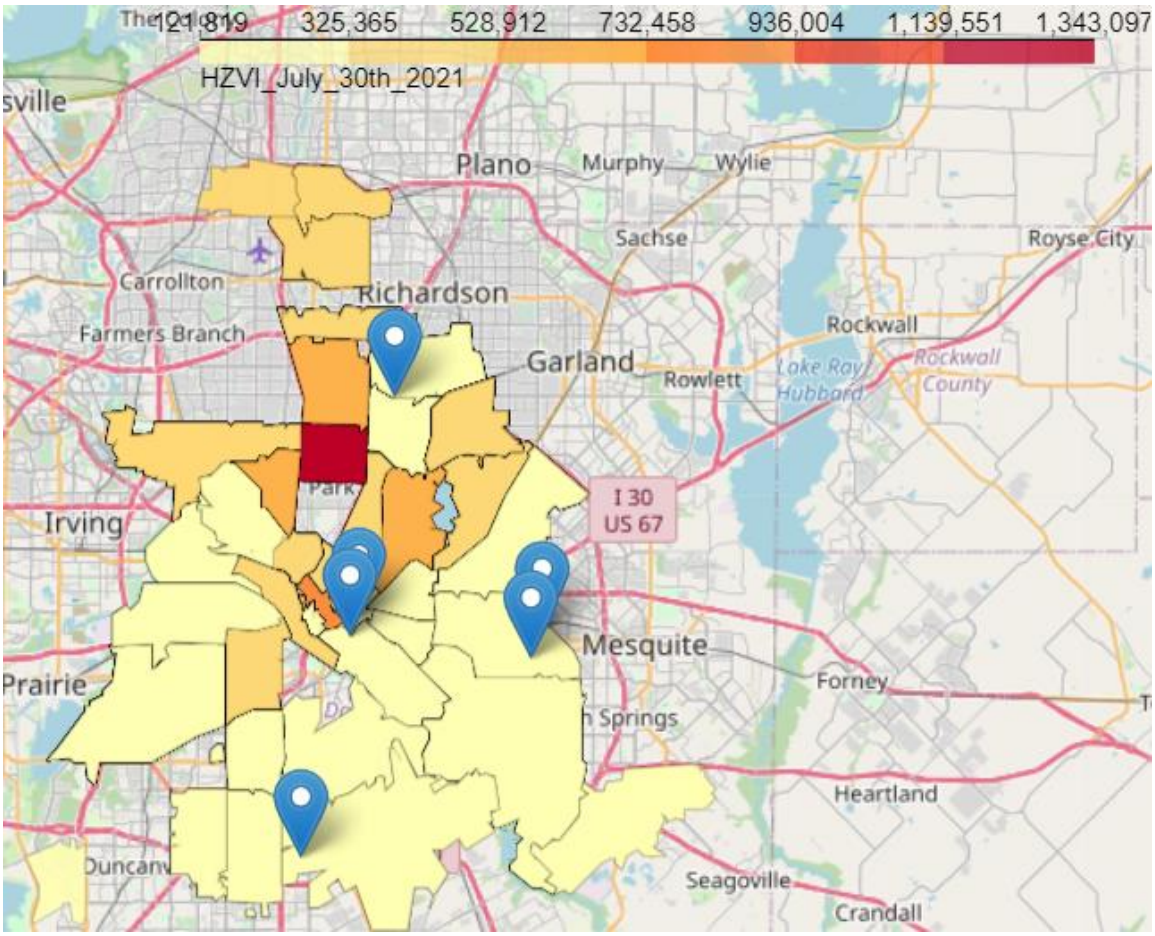
5. East Dallas, Baylor/Meadows

6. Southeast Dallas, Riverway Estates/Bruton Terrace

7. Southeast Dallas, Cedar Run

The remaining less attractive candidate neighborhoods, albeit still competitive may be found on the map above. Their colors can be identified from the map as 'Rajah', 'Blue jeans', 'Dark turquoise', 'Orange (Crayola)', 'Blue (RYB)', 'Flax', 'Aquamarine', 'Yellow-green (Crayola)', 'Blue (Crayola)', 'Turquoise', 'Violet (color wheel)', 'Red (RYB)', 'Red', 'Light green'

Finally, we display the best 7 neighborhoods optimal for opening an Indian, Asian, or Chinese restaurant:



5. Results and discussion

From our preliminary exploration, we found that only 3.7% of all of 1000+ restaurants (bars were excluded in this search) in Dallas are Indian, Chinese, or Asian. This leaves much scope to open a restaurant of these types in a bustling city with a diverse population.

Our analysis shows that even though there is a larger number of restaurants in the center of Dallas as opposed to other parts, some of the final recommended neighborhood centers are within the constraints of being 300m away from any Indian, Asian, or Chinese restaurant, and 500m away from any other restaurant.

Neighborhoods like 'Deep Ellum', 'Bishop Arts', 'Greenville', 'Trinity Groves', 'Knox-Henderson', 'Oak Lawn', 'Design District', 'Downtown', 'Lakewood' offer a combination of popularity among tourists, strong socio-economic dynamics, thriving entertainment and restaurant scene. Therefore, it would be wiser to open a restaurant in a neighborhood nearer to these neighborhoods, dubbed as ’Good Food Neighborhoods’.

Next, three clustering algorithms were applied: DBSCAN, HDBSCAN and K-means to cluster the resulting neighborhoods. The K-Means and HDBSCAN algorithms clustered the neighborhoods more appropriately than the DBSCAN, given the parameters they were assigned. The parameter K clusters for K-Means were chosen based on the Elbow method (naïve method) and Silhouette score and K was chosen to be 14. For DBSCAN (and its variant, HDBSCAN), heuristics were applied obtained from literature. Since, there were 2 dimensions, the MinPts parameter was set to 4. For the other parameter, epsilon (ε), the crook of the elbow on the plot of ‘(Sorted) average distance between each candidate neighborhood and its 4 nearest neighbors’ (Y) v/s ‘(Sorted) neighborhood number’ (X) was assigned as ε.

Finally, the candidate neighborhoods were further filtered by viewing the maps of the 3 clustering algorithms used, and 7 neighborhoods were chosen based on the above constraints and a further constraint. This was the property price in that zip code. Properties may be leased for short term projects (or lack of purchasing power, or for legal reasons

imposed by the government), or bought for long term projects. The neighborhoods which belonged to an inexpensive area were recommended.

The 7 neighborhoods were:

1. Downtown Dallas, Uptown

2. West Dallas, Trinity Grove

3. East Dallas, Lower Greenville

4. South Central Dallas, Skyline Heights

5. East Dallas, Baylor/Meadows

6. Southeast Dallas, Riverway Estates/Bruton Terrace

7. Southeast Dallas, Cedar Run

6. Conclusion

Purpose of this project was to identify Dallas neighborhoods close to center with low number of restaurants (particularly Indian, Asian, or Chinese) to aid stakeholders in narrowing down the search for optimal location for a new Italian restaurant.

This, of course, does not imply that these neighborhood centers are in fact optimal locations for a new restaurant! Purpose of this analysis was to only provide info on areas close to booming restaurant neighborhoods but not crowded with existing restaurants (particularly Indian, Asian, or Chinese) - it is entirely possible that there is a very good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Recommended neighborhood centers should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition and appropriate property prices, but also other factors considered, and all other relevant conditions met.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like levels of real estate availability, noise / proximity to major roads, social and economic dynamics of every neighborhood etc.