THE BATTLE OF NEIGHBORHOODS REPORT

(Version 2.0)

By,

Janaranjani Kannan

1 Table of Contents

1 Introduction

STEM is an acronym for the fields of Science, Technology, Engineering, and Mathematics, commonly used among educators, policymakers, and government officials. The demand for STEM students increases day by day. To fill academic gaps, parent seeks tutoring centers to improve their student's grades, raise college exam test scores and gain academic confidence. In this report, we are going to see how an entrepreneur who is in STEM tutoring industry uses Data Science to find the perfect neighborhood to establish his premises.

1.1 Business Problem

A successful entrepreneur who owns multiple tutoring centers in North East US, is looking to expand in South. To start with, he wants to open a tutoring center in Dallas, Texas hence approached the Data Science team to help him locate the best neighborhood where he can establish a tutoring center and run it successfully.

1.2 Introduction about the city

With an estimated population of 1,345,076 and still growing, Dallas is the ninth most-populous city in the U.S. Dominant sectors of its diverse economy include defense, financial services, information technology, telecommunications, and transportation. Dallas is home to 9 Fortune 500 companies within the city limits. Over 41 colleges and universities are located within its metropolitan area, which is the most of any metropolitan area in Texas. The city has a population from a myriad of ethnic and religious backgrounds.

1.3 Target Audience

The learning center franchises such as Kumon, Mathnasium, Huntington etc or any tutoring center owner can use the data analysis discussed in this report to find the best neighborhood to establish their premises.

With the information on hand, lets move on to Data collection.

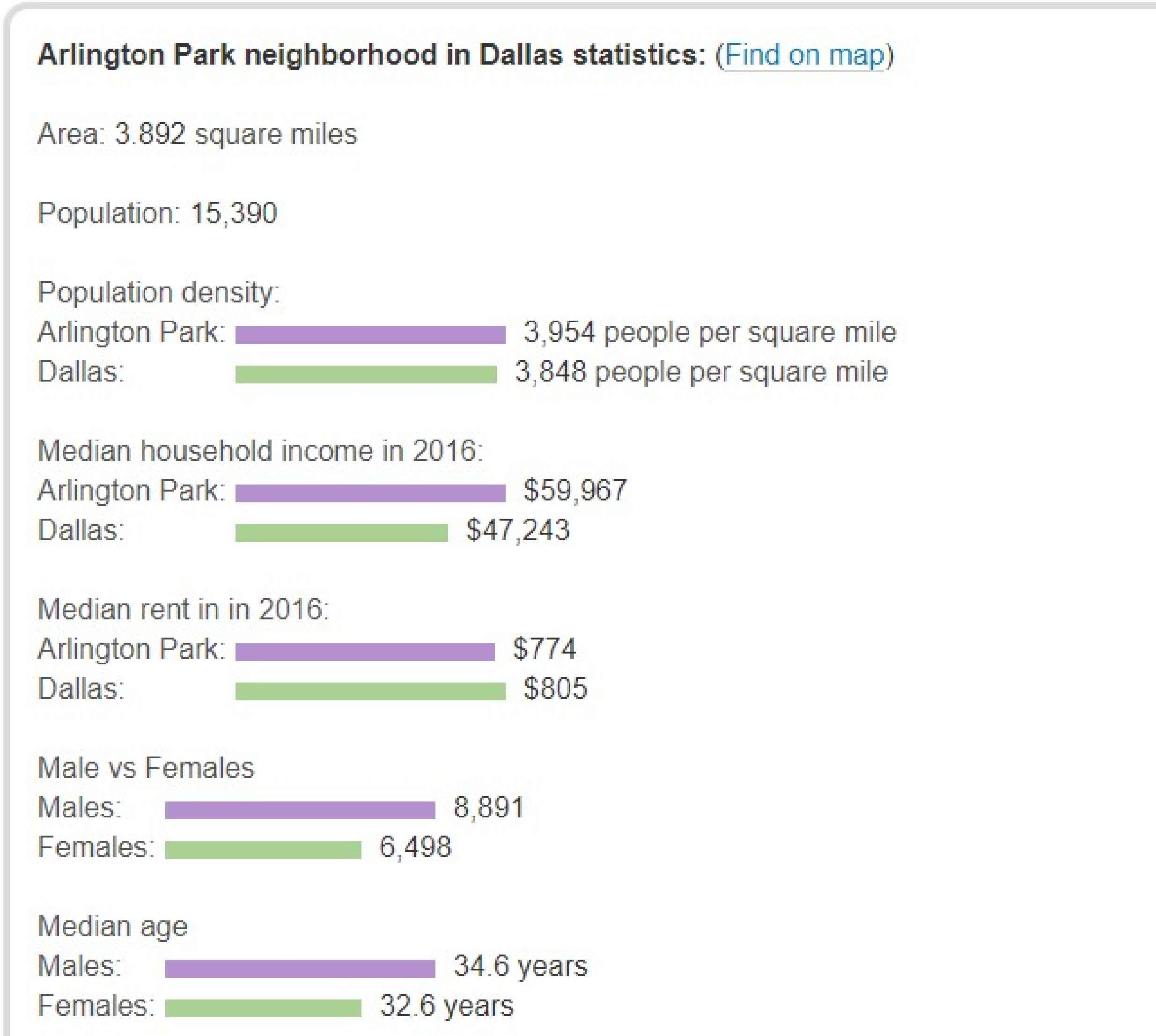
2 Data acquisition and cleaning

The entrepreneur is looking for middle aged, medium to high income families and populous neighborhood who will be interested in using tutoring centers to help their children academically. We are tasked with finding a neighborhood that satisfies entrepreneur's criteria.

2.1 Data Source

i) The portal **'www.city-data.com'** has detailed, informative profiles for every city in the United States. We are looking for Dallas's neighborhoods lists, population in each neighborhood, Male/Female age and Household income which can be collected from this portal.
ii) **Four-Square API** will allow us to collect details about venues around each neighborhood so that we can single out one neighborhood which doesn't have any tutoring center near it so that the success rate of the entrepreneur's tutoring center will be high.

Raw data: Sample data from **‘www.city-data.com’.**



2.2 Data Collection Method

1. Data related to Dallas can be scraped from the portal: ‘www.city-data.com’.
2. Use ‘geopy’ module to extract latitude and longitude of each neighborhood.
3. FourSquare API provides venue details for any mentioned radius based on the extracted latitude and longitude of each neighborhood.

2.3 Example Dataset

	City	Neighborhood Name	Population	Male Avg Age	Female Avg Age	Median Household Income	Latitude	Longitude
0	Dallas	Arlington Park	15390.0	34.6	32.6	59967.0	32.817605	-96.857609
1	Dallas	Belmont	5085.0	35.5	32.6	86718.0	32.813733	-96.782253
2	Dallas	Bent Tree	31951.0	42.0	42.3	100680.0	32.973411	-96.826306

2.4 Feature Selection

We need two different sets of data required for the data analysis. From 'www.city-data.com' portal, we need the following features:

1. City
2. Neighborhood Name
3. Population per neighborhood
4. Median Household Income per neighborhood
5. Male average age per neighborhood
6. Female average age per neighborhood

Additional ones from the module "Geopy" and Four Square API:

7. We can retrieve Latitude and Longitude based on neighborhood address using "Geopy" module.
8. Four Square API provides location data. Using the endpoint: "explore", we can obtain top 100 venues that are within a radius of 2000 meters

2.5 Data cleaning

Let's clean our dataset as follows

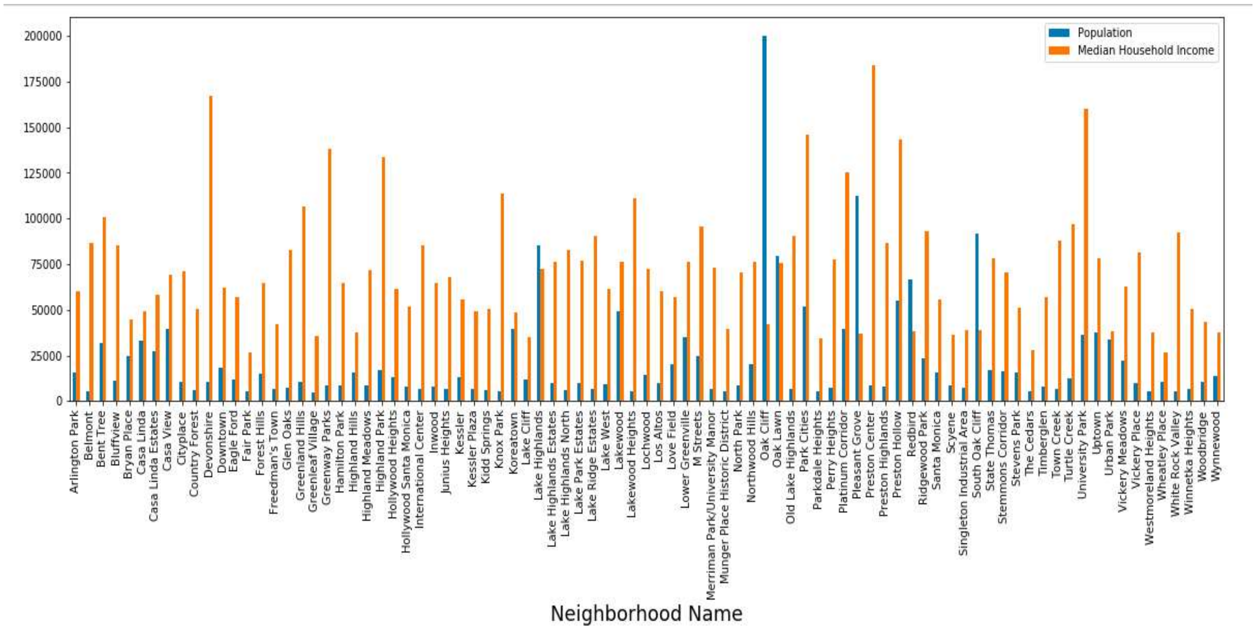
1. Remove duplicate entries.
2. Drop rows if any of its cells have NaN instead of real data.
3. Remove cumulative population number entries such as North Dallas, East Dallas, Far North West Dallas etc since each of these locations are further divided into many neighborhoods which are already part of this dataset.
4. Consider neighborhoods if its population is more than 5000 since 'Population is an important feature in our decision making.

3 Methodology

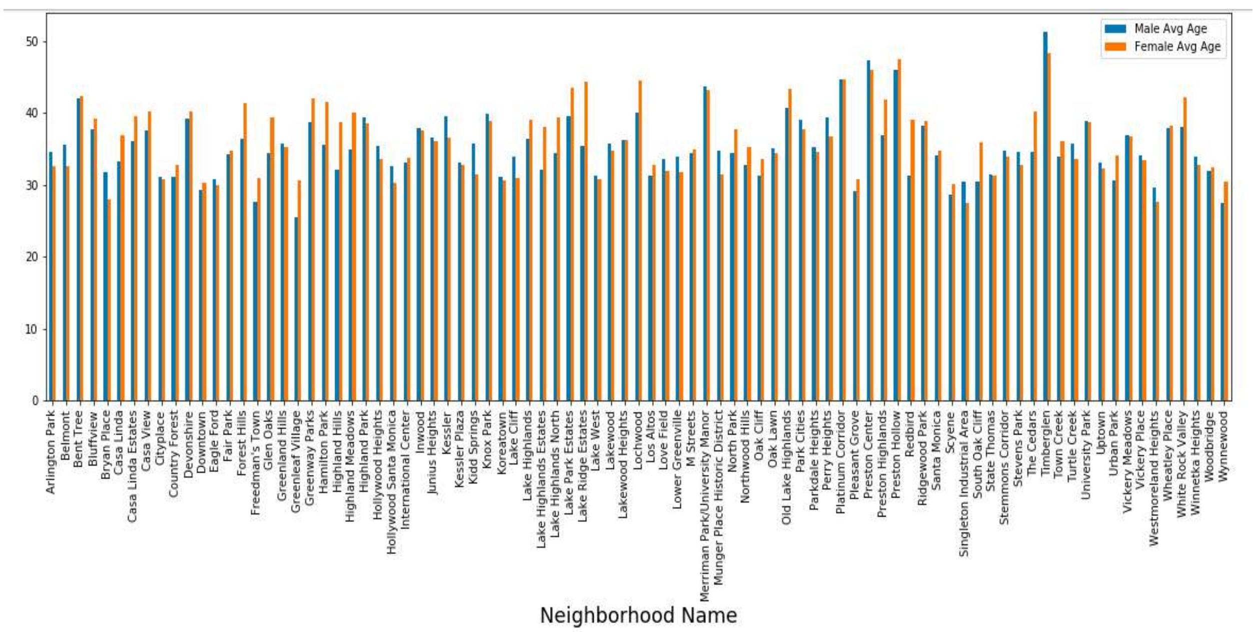
3.1 Exploratory Data Analysis

Let's visualize the initial dataset before we proceed further with our analysis. We are going to create two bar charts to compare 1. Population and Median Household Income and 2. Male and Female average age for each neighborhood.

Below bar chart depicts the comparison between Population and Median Household Income for each neighborhood.

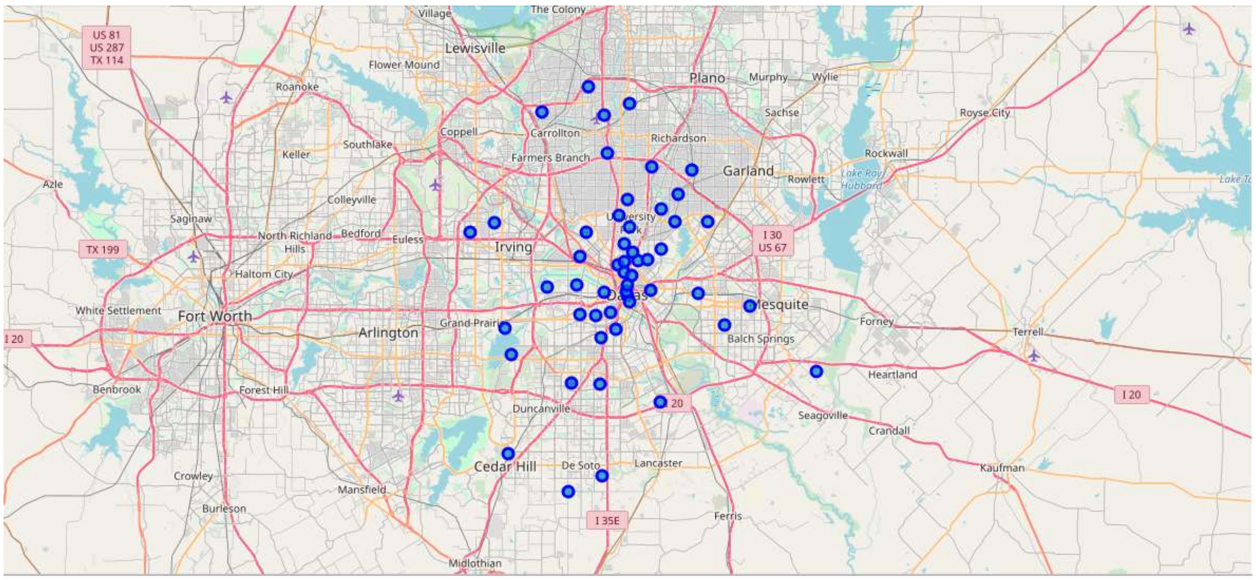


Below bar chart depicts the comparison between Male and Female average age for each neighborhood.



Our goal of this data analysis is to find a situatable neighborhood which has middle aged, medium to high income families and a populous neighborhood. Per the above two pictures, on an average most of the neighborhood consists of middle-aged male/female but the first picture provides us an interesting fact that most populous neighborhood doesn't meet our income criteria. Hence, we need to find some balance between these 4 features in our final dataset to find the perfect neighborhood for our client.

Below is the initial map of Texas with our filtered dataset version of Dallas neighborhoods superimposed on top.



3.2 Four Square API

Four Square API provides location data based on latitude and longitude of each neighborhood. Let us use Four Square endpoint "explore" to retrieve top 100 venues that are within a radius of 2000 meters for all neighborhoods in Dallas. Our dataframe will look like the below one.

	City	Neighborhood Name	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dallas	Arlington Park	32.818	-96.858	Sushi Time	32.821937	-96.856740	Sushi Restaurant
1	Dallas	Arlington Park	32.818	-96.858	New Fine Arts Alternatives	32.821748	-96.856172	Bookstore
2	Dallas	Arlington Park	32.818	-96.858	Hampton Inn & Suites	32.811370	-96.858243	Hotel
3	Dallas	Arlington Park	32.818	-96.858	Jimmy John's	32.821892	-96.855460	Sandwich Place
4	Dallas	Arlington Park	32.818	-96.858	Smokey's John's Bar-B-Que	32.821695	-96.854113	BBQ Joint

Since Venue and Venue Category in the dataset are categorical, we need to convert it into numerical data. We can do that using 'One hot encoding' method which is a process by which categorical variables are converted into a form that could be provided to Machine Learning algorithms to do a better job in prediction. The output of 'One hot encoding' will look like the below picture.

	City	Neighborhood Name	ATM	Accessories Store	African Restaurant	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Water Park	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	Zoo	Zoo Exhibit
0	Dallas	Arlington Park	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	Dallas	Arlington Park	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	Dallas	Arlington Park	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	Dallas	Arlington Park	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	Dallas	Arlington Park	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 301 columns

3.3 Top 10 Venue selection

After changing categorical feature set into a numerical one using 'One hot encoding', we need to select top 10 venues for each neighborhood. We are required to group rows by neighborhood and by taking the mean of the frequency of occurrence of each venue category for that neighborhood.

Dataset will look like the below one:

	Neighborhood Name	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	Arlington Park	Hotel	American Restaurant	Sandwich Place	BBQ Joint	Burger Joint	Mexican Restaurant	Fast Food Restaurant	Rental Car Location	Convenience Store	Gas Station
1	Belmont	American Restaurant	Bar	Pizza Place	Coffee Shop	Mexican Restaurant	Restaurant	Taco Place	Thai Restaurant	New American Restaurant	Grocery Store
2	Bent Tree	Rental Car Location	Pizza Place	Italian Restaurant	Burger Joint	Park	Hotel	Golf Course	Steakhouse	Gas Station	Sushi Restaurant
3	Bluffview	Korean Restaurant	Sushi Restaurant	Sandwich Place	Coffee Shop	Bakery	Fast Food Restaurant	Bubble Tea Shop	Ice Cream Shop	Mexican Restaurant	Pizza Place
4	Bryan Place	Clothing Store	Burger Joint	Cosmetics Shop	Mexican Restaurant	Discount Store	Coffee Shop	Furniture / Home Store	Department Store	Sushi Restaurant	Supplement Shop

3.4 Predictive Modeling (Machine Learning algorithm selection)

The K-means machine learning algorithm is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data.

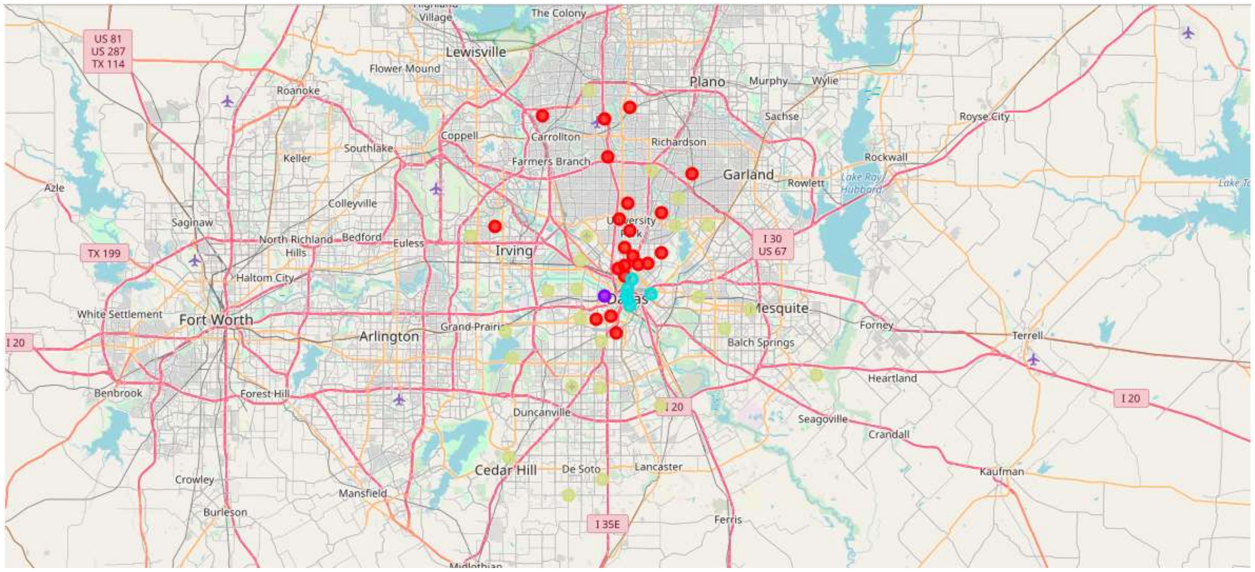
Some real-world applications of k-means:

- Customer segmentation
- Understand what the visitors of a website are trying to accomplish
- Pattern recognition
- Machine learning
- Data compression

We are going to use k-Means for neighborhood segmentation, and we are going to cluster the neighborhoods into 4 clusters.

4 Result

Let us use folium to visualize the dataset(see below picture) and its respective clusters. Data visualization will help us to examine the clusters and categorize them based on the neighborhood characteristics(in our case it's venues).



5 Discussion

As we have segmented the neighborhoods into 4 clusters, we ended up having 4 different types of neighborhood. They are listed below per cluster numbers.

1. **Ethnic** - Immigrants from a particular ethnicity, young couples, budget-conscious singles.
2. **Urban Pioneer** - Near downtown and inner-ring suburbs.
3. **Urban Core** - Downtown, the heart of major metros.
4. **Cul-de-sacs & Kids (Bedroom)** - Middle-aged soccer moms and dads whose lives revolve around their children.

Let us view all the 4 clusters to analyze the data further.

Cluster 1

Cluster 1 - Neighborhood type : Ethnic

```
texas_merged.loc[texas_merged['Cluster Labels'] == 0, texas_merged.columns[list(range(0,6))+ list(range(9, texas_merged.shape[1]))]].sort_values(by='Population',ascend
```

	City	Neighborhood Name	Population	Male Avg Age	Female Avg Age	Median Household Income	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	Oak Cliff	200297	31.3	33.5	41991.0	Mexican Restaurant	Zoo Exhibit	Restaurant	Fried Chicken Joint	Brewery	Taco Place	Italian Restaurant	Gastropub	Convenience Store	Pizza Place
1	Dallas	Oak Lawn	79350	35.1	34.4	75484.0	Gay Bar	Bar	Mexican Restaurant	Seafood Restaurant	Coffee Shop	Hotel	Nail Salon	Italian Restaurant	Salon / Barbershop	Burger Joint
2	Dallas	Preston Hollow	54703	45.9	47.4	143411.0	Bakery	Italian Restaurant	Pizza Place	Ice Cream Shop	Shipping Store	Seafood Restaurant	Sandwich Place	Mexican Restaurant	Spa	Café
3	Dallas	Lakewood	49005	35.8	34.8	76122.0	Mexican Restaurant	Cosmetics Shop	Burger Joint	Bar	Discount Store	Pizza Place	Coffee Shop	Sushi Restaurant	Pet Store	Park
4	Dallas	Uptown	37818	33.0	32.3	78443.0	American Restaurant	Cocktail Bar	Hotel	Sushi Restaurant	Japanese Restaurant	Burger Joint	Coffee Shop	Seafood Restaurant	Steakhouse	Mexican Restaurant
5	Dallas	University Park	35995	38.8	38.7	160096.0	Bakery	Coffee Shop	American Restaurant	Seafood Restaurant	Ice Cream Shop	Mediterranean Restaurant	New American Restaurant	Cupcake Shop	Sandwich Place	Gym / Fitness Center
6	Dallas	Lower Greenville	34956	33.9	31.8	76117.0	Mexican Restaurant	Bar	Pizza Place	New American Restaurant	Grocery Store	Coffee Shop	Thai Restaurant	Burger Joint	Taco Place	Vietnamese Restaurant

Cluster 2

Cluster 2 - Neighborhood type : Urban Pioneer

```
#texas_merged.loc[texas_merged['Cluster Labels'] == 1, texas_merged.columns[list(range(0,6))+ list(range(9, texas_merged.shape[1]))]].sort_values(['Population'],ascend
texas_merged.loc[texas_merged['Cluster Labels'] == 1].sort_values(['Population'],ascending=False).reset_index(drop=True).head(10)
```

	City	Neighborhood Name	Population	Male Avg Age	Female Avg Age	Median Household Income	Latitude	Longitude	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	I Com V
0	Dallas	Singleton Industrial Area	7240	30.4	27.5	38608.0	32.779	-96.826	1.0	Italian Restaurant	Brewery	Event Space	Taco Place	Plaza	Mexican Restaurant	Scenic Lookout	Hotel	Asian Restaurant	

Cluster 3

Cluster 3 - Neighborhood type : Urban Core (Downtown)

```
texas_merged.loc[texas_merged['Cluster Labels'] == 2, texas_merged.columns[list(range(0,6))+ list(range(9, texas_merged.shape[1]))]].sort_values(by='Population',ascend
```

	City	Neighborhood Name	Population	Male Avg Age	Female Avg Age	Median Household Income	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	Casa View	39253	37.6	40.2	69417.0	Hotel	Coffee Shop	Steakhouse	Plaza	Bar	Cocktail Bar	American Restaurant	Park	History Museum	Café
1	Dallas	Downtown	18395	29.2	30.2	61939.0	Hotel	Coffee Shop	Park	Cocktail Bar	American Restaurant	Mexican Restaurant	Steakhouse	Plaza	Gym	Movie Theater
2	Dallas	Wheatley Place	10619	37.8	38.2	26649.0	American Restaurant	Burger Joint	Steakhouse	Japanese Restaurant	Seafood Restaurant	Mexican Restaurant	Cocktail Bar	Coffee Shop	Hotel	New American Restaurant
3	Dallas	Greenway Parks	8275	38.7	42.0	137987.0	Pizza Place	Discount Store	Coffee Shop	Fast Food Restaurant	New American Restaurant	Bar	American Restaurant	Sandwich Place	Hotel	Brewery
4	Dallas	International Center	6772	33.0	33.7	85388.0	Hotel	American Restaurant	Bar	Coffee Shop	Steakhouse	Cocktail Bar	New American Restaurant	Seafood Restaurant	Japanese Restaurant	Mexican Restaurant
5	Dallas	Country Forest	6180	31.1	32.8	50420.0	Tapas Restaurant	Coffee Shop	Restaurant	Hotel	Bar	Nightclub	Spanish Restaurant	Plaza	Scenic Lookout	Café
6	Dallas	Fair Park	5291	34.3	34.7	26811.0	Bar	American Restaurant	Dive Bar	Burger Joint	Coffee Shop	Cocktail Bar	Art Gallery	Pizza Place	Rock Club	Nightclub

Cluster 4

Cluster 4 - Neighborhood type : Cul-de-sacs & Kids (Bedroom)

```
texas_merged.loc[texas_merged['Cluster Labels'] == 3, texas_merged.columns[list(range(0,6))+ list(range(9, texas_merged.shape[1]))]].sort_values(by='Population',ascend
```

	City	Neighborhood Name	Population	Male Avg Age	Female Avg Age	Median Household Income	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	Pleasant Grove	112549	29.1	30.7	36852.0	Fast Food Restaurant	Mexican Restaurant	Pizza Place	Convenience Store	Fried Chicken Joint	Thrift / Vintage Store	Grocery Store	Discount Store	Burger Joint	Gas Station
1	Dallas	Lake Highlands	85169	36.3	39.0	72459.0	Pizza Place	Fast Food Restaurant	Sandwich Place	Convenience Store	Pharmacy	Coffee Shop	Mexican Restaurant	Grocery Store	Bank	Video Store
2	Dallas	Redbird	66535	31.3	39.1	37968.0	Fast Food Restaurant	Discount Store	Convenience Store	Pizza Place	Fried Chicken Joint	Department Store	Wings Joint	Southern / Soul Food Restaurant	Sandwich Place	Big Box Store
3	Dallas	Urban Park	33552	30.6	34.0	38557.0	Fried Chicken Joint	Breakfast Spot	Pharmacy	Discount Store	Chinese Restaurant	Convenience Store	Fast Food Restaurant	Home Service	Gas Station	Mexican Restaurant
4	Dallas	Bryan Place	24660	31.7	28.0	44804.0	Clothing Store	Burger Joint	Cosmetics Shop	Mexican Restaurant	Discount Store	Coffee Shop	Furniture / Home Store	Department Store	Sushi Restaurant	Supplement Shop
5	Dallas	Ridgewood Park	23467	38.2	38.9	92875.0	Fast Food Restaurant	Breakfast Spot	Hardware Store	Sandwich Place	Bank	Liquor Store	Gas Station	Supplement Shop	Garden Center	Big Box Store
6	Dallas	Love Field	20048	33.6	32.0	57009.0	Rental Car Location	Mexican Restaurant	Airport Service	Coffee Shop	Fast Food Restaurant	Convenience Store	Hotel	Sandwich Place	Burger Joint	Shoe Store

Per our business objective, we need to select a neighborhood based on the following features:

1. Middle aged Male/Female
2. Medium to High Income Families
3. Populated neighborhood
4. No other tutoring/learning center within a radius of 2000 meters

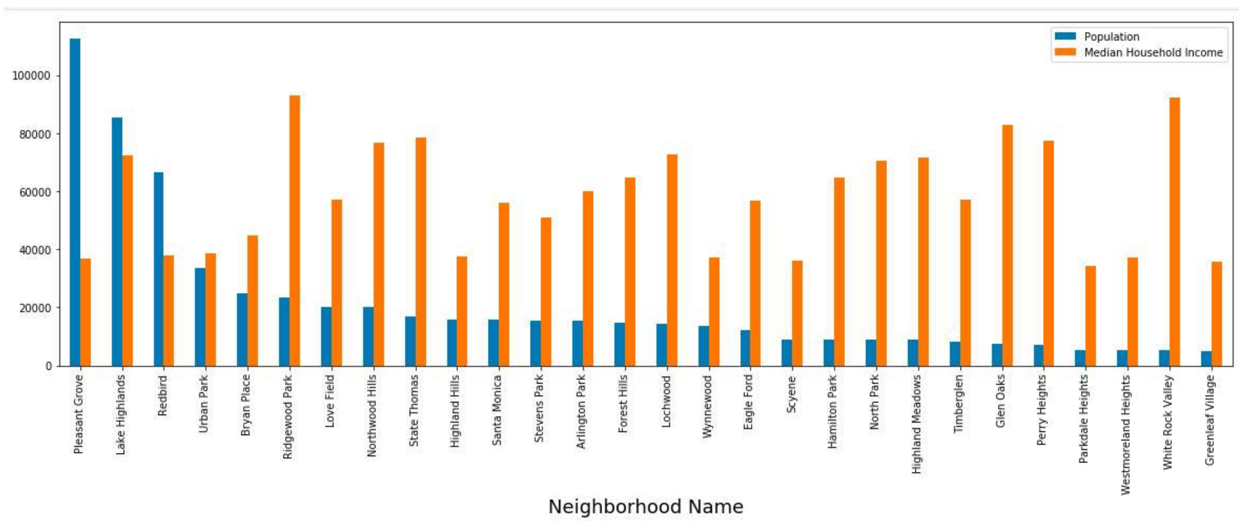
Venue details within a neighborhood allows us to define the neighborhood type. We are looking for neighborhood types like:

1. '*New Urban*' which is populated and near a business hub other than the city's main downtown.
2. *'Cul-de-sacs & Kids (Bedroom)'* which is populated with middle-aged soccer moms and dads whose lives revolve around their children.

Examining the clusters, we can see ***'Cul-de-sacs & Kids'*** neighborhood type is one among them which satisfies our business criteria. Hence we can narrow down to **Lake Highlands** neighborhood (Cluster 4) based on our client's criteria (i.e, middle-aged, medium to high income families and populous neighborhood) instead of 'Pleasant Grove' since the Income feature is not satisfactory.

5.1 Data visualization

As we can see here that the neighborhood: 'Pleasant Grove' is highly populated compared to Lake Highlands however its median household income falls between low-medium income earners hence Lake Highlands is the perfect neighborhood for our client.



6 Conclusion

Based on the data analysis (per the combination of the features such as *Population, Median Household Income, Male/Female average age and Venues list*) and KMeans machine learning algorithm's clusters(neighborhood types), we can conclude that **Lake Highlands** is the perfect neighborhood to start the first tutoring center in Dallas by the client.