

# Battle of Neighborhoods – Provide Choices for Residential Homes based on School and child care needs

Anasuya Devi Kompella

IBM Data Science Capstone Project



# Background



- Finding right school is a big challenge for parents and in case of public schools, there are a few more parameters to be considered.
- In USA, kids can only attend public schools that are assigned to their residence. So, if parents prefer a particular school, they need to reside in the geographical boundary assigned to the school.
- Hence choice of schools for kids influences the choice of houses the parents are likely to rent or buy.
- If parents are looking to buy a house, they would like to make sure that all levels of associated schools are good. And if they have child care needs like daycare, pre-school, having good facilities near by may also be factor that influences their decision



# Problem Definition

- There are existing residential property search websites like HAR.com, Zillow.com etc. Search on these sites can be done in following manner:
  - School search - Given an address, zipcode or city, lists out the schools (with filters for ratings and other scores) and then for each school, there is a link to search for homes in the school's boundary.
  - Home search - While searching for homes, under details of each home, associated school details are given with options to click on each school and get its ratings score etc.
- However if the requirement is to find out a list of neighborhoods in a city (or list of cities) that are best suited as per the family's educational and child care needs, the above mentioned sites do not cater to the same.
  - For eg: 1. Which are the neighborhoods that are best for families with teenage kids with good ratings of high school ?
  - 2. Which neighborhoods are best for families with young children who need daycares/pre-schools now and good elementary school going forward?



# Data Sources

- List of Neighborhoods and their Geographical Coordinates
  - This list of neighborhoods was compiled offline from multiple sources as a csv file. It consists of Neighborhoods from a cluster of cities in Greater Houston.
  - Latitude and Longitude for each neighborhood is obtained
- List of schools and day cares within a given radius of each neighborhood.
- Ratings of each of the schools obtained from a centralized source
  - School rating – Rating assigned based on accountability scores of schools
  - Parent Rating – Computed based on the ratings given by parents



# Sample Data

```
data = pd.read_csv(neighborhoodsFile)
data.head()
```

Out[3]:

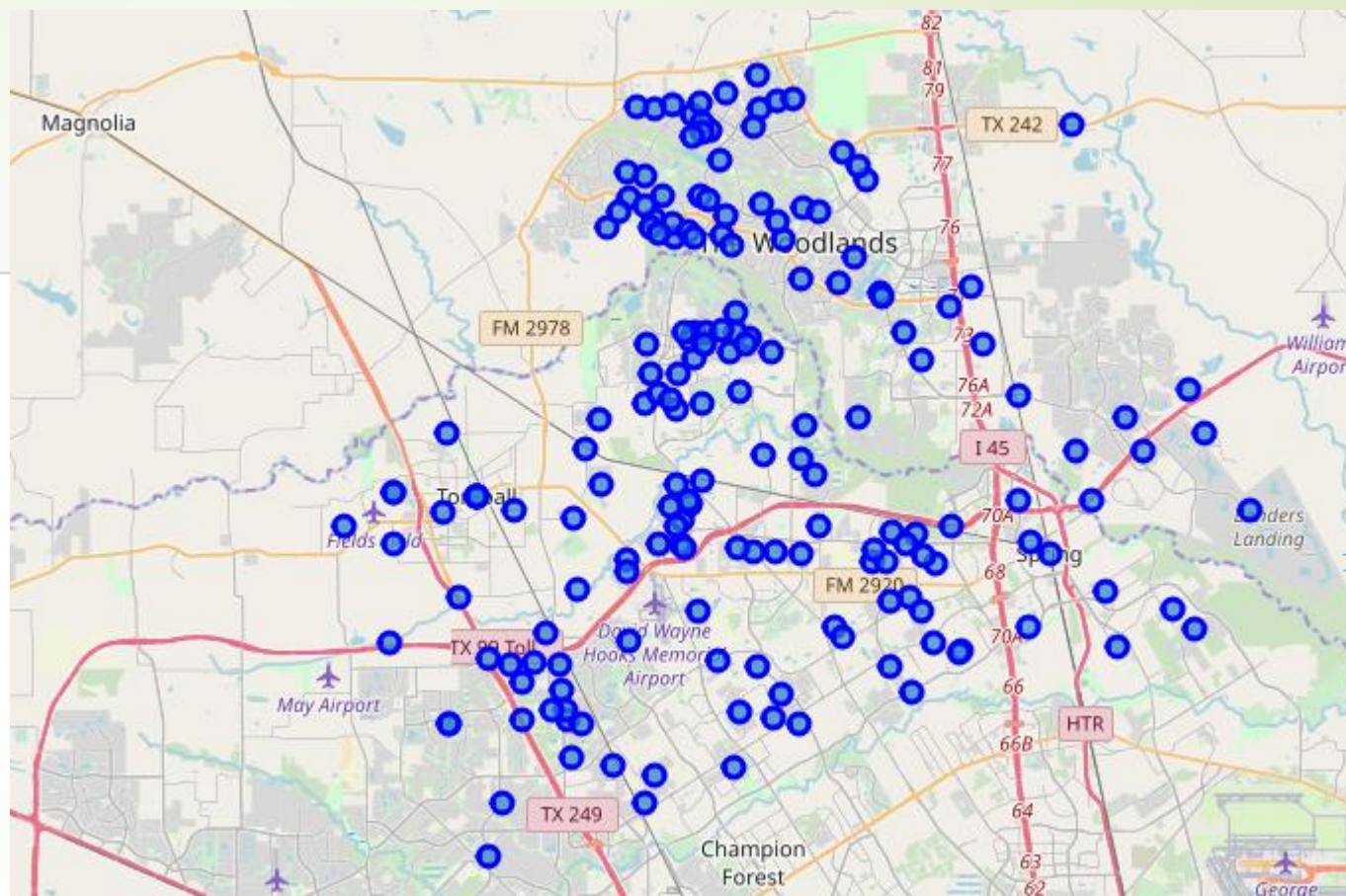
	Neighborhood	City
0	Acacia Park	Spring, TX
1	Albury Trails Estates	The Woodlands, TX
2	Alden Bridge	Spring, TX
3	Alden Bridge Hollylaurel	Spring, TX
4	Alden Trace	Spring, TX

Base list of neighborhoods

Out[5]:

	Neighborhood	City	Resolved Name	Latitude	Longitude
0	Acacia Park	Spring, TX	(Acacia Park, Alden Bridge, The Woodlands, Mon...	30.218422	-95.529759
1	Albury Trails Estates	The Woodlands, TX	(Albury Trails Estates, Harris County, Texas, ...	30.069389	-95.580927
2	Alden Bridge	Spring, TX	(Alden Bridge, The Woodlands, Montgomery Count...	30.213204	-95.517991
3	Alden Bridge Hollylaurel	Spring, TX	(Hollylaurel, Alden Bridge, The Woodlands, Mon...	30.223754	-95.518385
4	Alden Trace	Spring, TX	(Alden Trace, Alden Bridge, The Woodlands, Mon...	30.208315	-95.519652

Neighborhoods resolved to Geographical coordinates



# Sample Data (contd)

Nearby School Data from  
Foursquare Location API

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	School	School Latitude	School Longitude	School Category
0	Acacia Park	30.218422	-95.529759	Woodlands Civic Ballet	30.207763	-95.527541	School
1	Acacia Park	30.218422	-95.529759	Bush Elementary School	30.211384	-95.517843	School
2	Acacia Park	30.218422	-95.529759	Barbara Pierce Bush Elementary	30.211636	-95.517427	Elementary School
3	Acacia Park	30.218422	-95.529759	Kumon Math and Reading Center of The Woodlands...	30.208512	-95.529289	School
4	Acacia Park	30.218422	-95.529759	Legacy Preparatory Christian Academy	30.225233	-95.545745	High School

	School	gsRating	parentRating	city	state
0	Bush Elementary School	3	3	Houston	TX
1	Galatas Elementary	10	4	Spring	TX
2	Powell Elementary	9	3	Spring	TX
3	Tough Elementary School	10	4	The Woodlands	TX
4	The Woodlands Christian High School	NaN	5	The Woodlands	TX

School rating data from  
[greatSchools.org](https://greatschools.org)

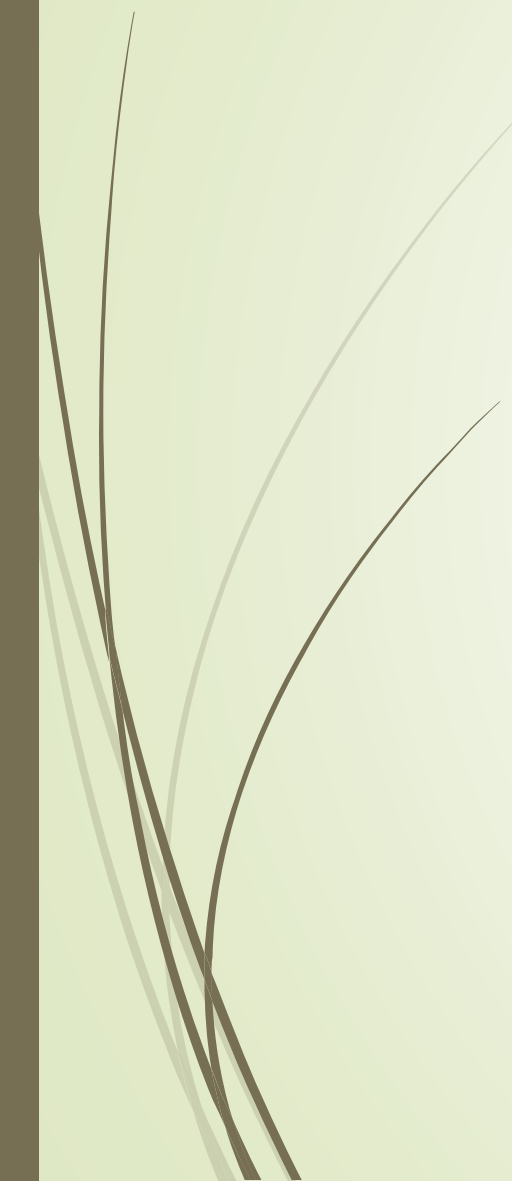


# Methodology Overview

- The goal is to provide supporting information for a family deciding on a neighborhood depending on their educational and child care needs. It is achieved by clustering a given set of neighborhoods as per existence of day cares, pre-schools and the ratings of schools nearby. It involves the following:
  - **Data Gathering:**
    - Getting LatLong Coordinates for pre-defined set of neighborhoods
    - Getting nearby schools for these neighborhoods using foursquare API
    - Obtaining ratings of the schools from [greatschools.org](http://greatschools.org)
  - **Data Preparation:**
    - Massaging the data for clustering
  - **Clustering:**
    - Clustering the data using KMeans Algorithm
    - Analyzing resulting clusters



# Scope/ Assumptions

- While there could be many parameters for deciding suitability of a house, only the parameters given below are considered to be in scope of this project:
    - Rating of elementary school, Middle School and High School within given radius
    - For private schools, there is no great school rating, only parent rating.
    - Existence of Day care or Pre-school facilities within given radius
  - A list of neighborhoods in Greater Houston comprising from a few cities are taken as input.
- 



# Methodology

## ■ Data Gathering:

### 1. Geographical Data:

- Latitude and Longitude of Neighborhoods obtained using GeoLocator API
- Get list of nearby schools for each of neighborhood using Foursquare Location API. Search function is used with category as 'School' since we are only interested in Schools and not all venues.
- Compile all the data into a dataframe.

### 2. Location based Venue Data:

- Analysis of the school data returned by Foursquare API shows that the schools are of multiple varieties like public or private schools, swim schools, driving schools, etc. Since we are only interested in educational institutes and child care centers, only the schools that are in the categories of Elementary school, Middle School, High School and Day Care/ Pre-school are filtered out.
- List of Unique Schools is obtained from the filtered list above.

### 3. School Ratings:

- Ratings of Elementary, Middle and High Schools are obtained from greatSchools.org.
- While retrieving data from greatSchools.org, sometimes multiple schools matching the textual search are returned and not just the school with given name. The school name is matched using text similarity measure to avoid issues of getting a wrong school data.
- Also schools with same name in multiple cities may be returned from greatschools, in such a case, expected city is also matched. These measures are taken to ensure accuracy of the school data.

# Methodology (contd)

## ► Data Preparation:

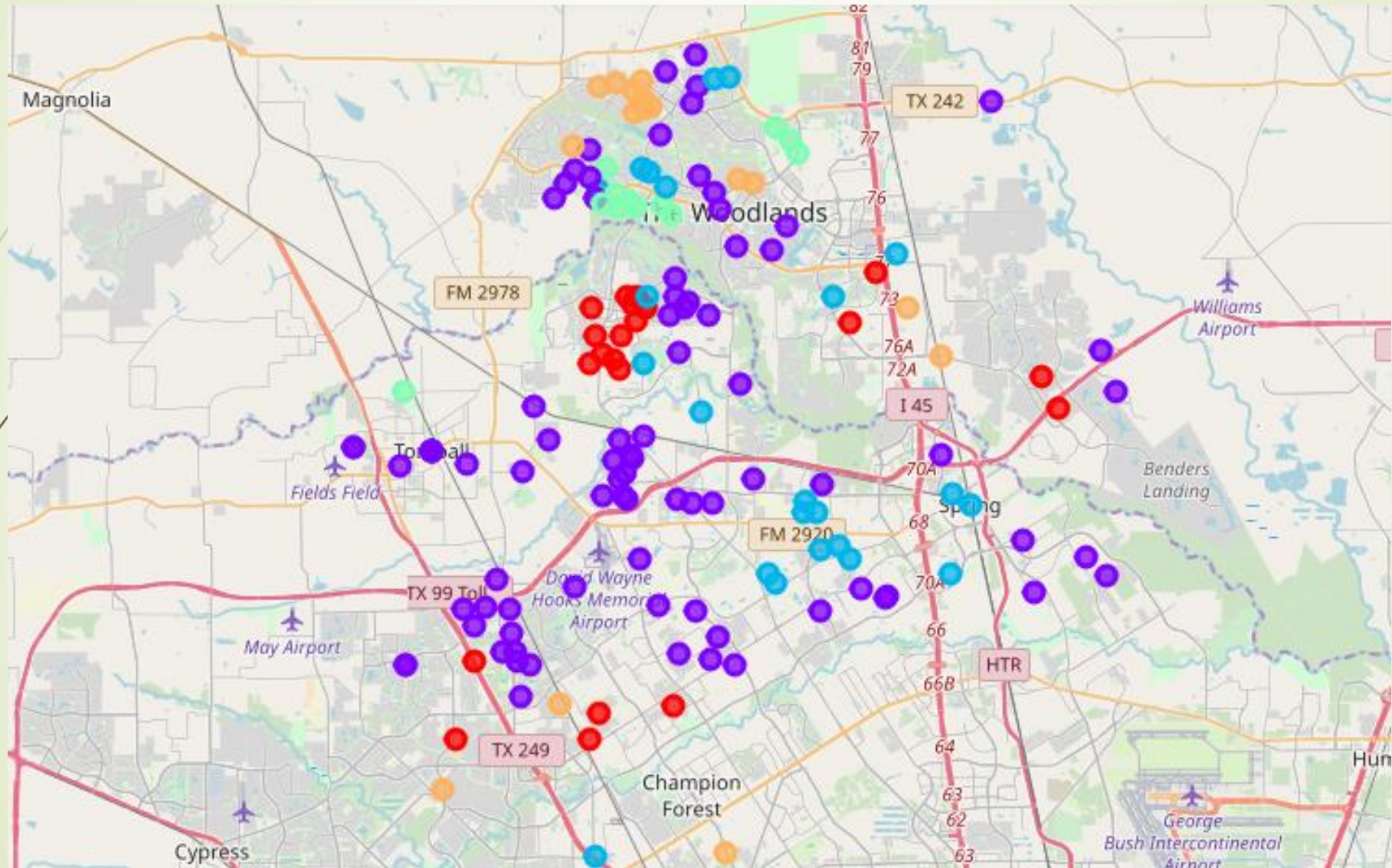
- Pre-school and day care existence for each of the neighborhood is obtained using onehot encoding
- Dataframe functions are used to massage and normalize data. This involves grouping based on neighborhood and collating elementary, middle and high school ratings from greatschools and parents as separate columns along with 0/1 for existence of pre-school and day care for each neighborhood.
- Data is now ready for clustering

Neighborhood	gsElementaryRating	parentElementaryRating	gsMiddleRating	parentMiddleRating	gsHighschoolRating	parentHighschoolRating	Daycare	Preschool
Acacia Park	0.3	0.6	0.0	0.0	0.0	0.0	0.0	0.0
Alden Bridge	0.3	0.6	0.0	0.0	0.0	0.0	0.0	0.0
Alden Bridge Hollylaurel	0.3	0.6	0.0	0.0	0.0	0.0	0.0	0.0
Alden Trace	1.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0
Altwood	1.0	0.8	0.0	0.0	0.0	1.0	1.0	0.0

## ► Clustering:

- Kmeans algorithm is used for clustering data into 5 clusters.
- Each cluster contents are given along with analysis for cluster components

# Results – Map View







# Analysis of Clusters

## ➤ **Cluster0: Neighborhoods near to :**

- Elementary Schools with good ratings
- Middle schools
- Childcare facilities like pre-schools

## ➤ **Cluster1: Neighborhoods near to :**

- Elementary Schools with average to good ratings
- Few Middle school assignments
- Almost no Childcare facilities

## ➤ **Cluster2: Neighborhoods near to :**

- Average to good rated Elementary Schools,
- Choice of public and private high schools
- Childcare facilities like day care-





# Analysis of Clusters

- **Cluster3: Neighborhoods near to :**

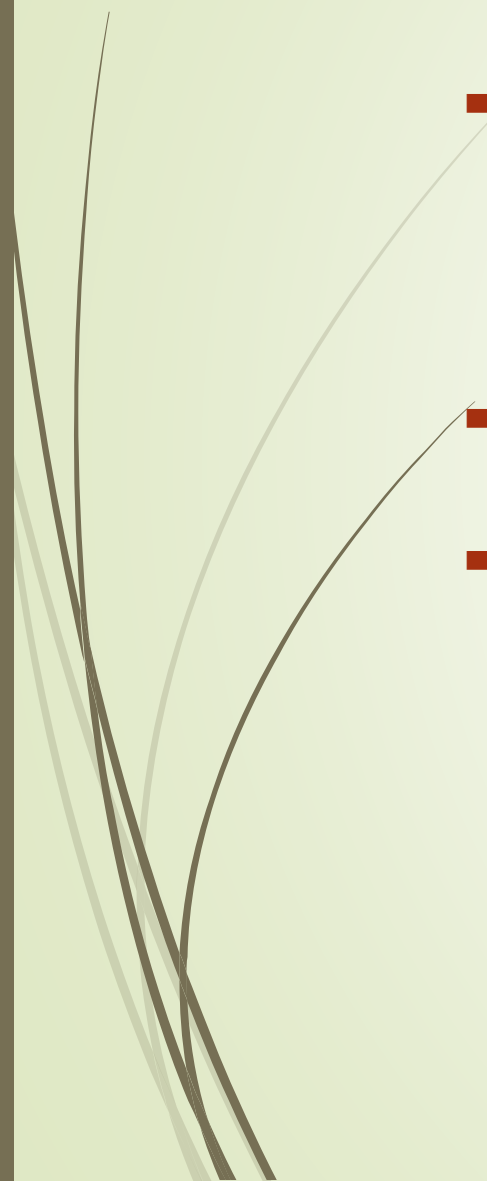
- Mostly private high schools with good parent rankings, and
- Childcare facilities like day care

- **Cluster4: Neighborhoods near to :**

- Elementary Schools with good ratings
- Childcare facilities like pre-schools



# Discussion

- ▶ Since there is a requirement to use Foursquare Location API as part of this course, the schools are taken from given radius (distance from the neighborhood) rather than assignment. The idea is to demonstrate the usage of data and capability of clustering algorithms to provide decision support information.
  - ▶ Same concept can be extended to identifying the assigned schools to make it lot more useful. This project is to be taken as a starting point.
  - ▶ Foursquare API results need to be analyzed and filtered appropriately to obtain only relevant information. The methods used in this project are purely based on my own assumptions and the problem definition.
- 

# Conclusion

Following conclusions may be drawn from the characteristics of 5 different clusters to help decide which neighborhoods to explore for a family home based on the criteria being satisfied by each cluster.

<b>Neighborhoods in Cluster#</b>	<b>Suitability for Which kind of Families</b>
0	<ul style="list-style-type: none"><li>• With Young kids starting with elementary school and going up to middle school, and</li><li>• With childcare needs</li></ul>
1	<ul style="list-style-type: none"><li>• With young kids going to/in elementary school, and</li><li>• With no childcare needs</li></ul>
2	<ul style="list-style-type: none"><li>• With teenage kids going to high school exploring both public and private school options, and</li><li>• With young kids needing child care</li></ul>
3	<ul style="list-style-type: none"><li>• With teenage kids and preference to private schools, and</li><li>• With young kids with child care needs</li></ul>
4	<ul style="list-style-type: none"><li>• With young kids going to elementary schools</li><li>• With childcare needs</li></ul>