K-means Analysis of Chicagoland Suburbs

A Comparison of Zillow Median Home Values with Foursquare location Data

Prepared by Sean Burke

11/7/2019

Introduction

Background

The Chicagoland area is one of the largest metropolitan areas in the United States. Outside the city limits of Chicago there is a collar of suburbs that many people call home. While the suburbs are very similar in some respects, they also vary in many different ways including home values and nearby business and entertainment venues. This project will explore and compare the suburbs of Chicago in Cook County using median home value data from Zillow and venue location data from Foursquare.

Description of Project

Different suburbs in the Chicagoland area have very different home values and very different amenities available. There are many variables in determining the price of a home, one of those variables would be the available amenities. In this case we will attempt to compare the suburbs of Chicago using Foursquare location data to give an idea of what neighborhoods have similar entertainment and other venues available. This data will then be cross-referenced to median home value prices of those neighborhoods. Analysis of this data should show neighborhoods that might be desirable because they include the same amenities as higher priced neighborhoods. While other variables such as school rankings play a large role in determining home prices in various neighborhoods for many people availble entertainment venues in a neighborhood may be the most important variable to them. By exploring this data one could look for neighborhoods where they can buy a more affordable home while still having access to a similar set of amenities that are available in some of the more expensive suburbs.

Target Audience

The results are useful to anyone involved in real estate transactions in the Chicago suburbs, people looking to research Chicago suburban neighborhoods to buy a home, investors looking for possible up and coming neighborhoods which may be more likely to experience rapid appreciation of property values and those interested in discovering neighborhoods in the Chicago suburbs that are similar to other neighborhoods they are more familiar with.

Neighborhoods are always changing and often business investment in an area preceeds gains in housing prices. If a suburb is becoming popular and businesses are opening new exciting experiences in the area, home prices in that area may begin to increase as people move to that area to be closer to those amenities. By comparing the foursquare location data of different suburbs with median home values it may be possible to find neighborhoods that have had significant business investment but haven't seen a large increase in home values yet. This data could be used to target neighborhoods for investment in housing to meet the demand of new buyers who will want to move to this area to take advantage of the available amenities.

Data

For this project data will be used from Zillow and Foursquare in conjunction with latitude and longitude data from arcgis.

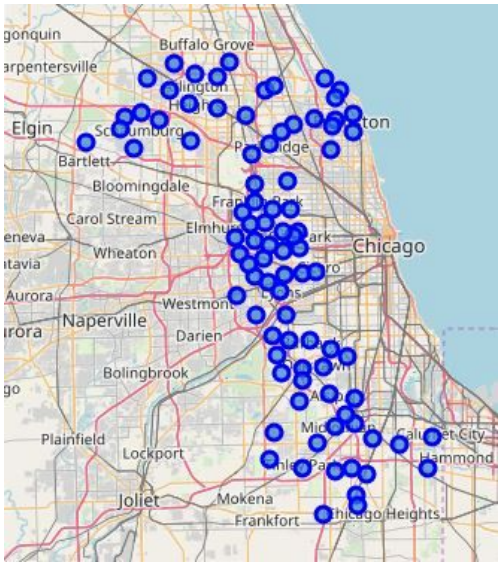
Median Home Value Index

Zillow provides detailed information about real estate in the United States. Zillow uses available information from various sources about properties such as school rankings, previous sale prices, current listing prices, neighborhood crime, and much more to allow Zillow to estimate home values throughout the United States. Additionally they can compile this data by zip code, state or other boundaries to estimate median home values for specific neighborhoods or states.

The Zillow home value index data was combined in a pandas dataframe with latitude and longitude data using arcgis as shown below.

	Zip Code	City	Zillow Home Value Index	Latitude	Longitude
0	60803	Alsip	172200	41.671395	-87.733745
1	60004	Arlington Heights	320200	42.111150	-87.980430
2	60005	Arlington Heights	293200	42.069047	-87.992100
3	60010	Barrington	473900	42.110325	-88.157915
4	60104	Bellwood	159800	41.881870	-87.870935
5	60163	Berkeley	188900	41.886505	-87.908110

The folium library was used to visualize these data as shown below.



Service and Entertainment Venues

Foursquare provides data on entertainment and services venues available in requested locations. Foursquare has a detailed database of businesses throughout the United States with information about these businesses such as location and type of business.

After running k-means clustering several outlier communities were noticed. Cross-checking these outlier suburbs with the venue counts returned by Foursquare showed that these outliers such as Steger, Barrington, Willow Springs and Chicago Heights were all suburbs with low venue counts returned by Foursquare. Below are examples of suburbs with low venue counts that were put into clusters containing only one suburb.

```
In [58]: suburb_merged_5.loc[suburb_merged_5['Cluster Labels'] == 5, suburb_merged_5.columns[[0] + [1] + [2] + list(range(5, suburb_merge
         d_5.shape[1]))]]
```

Out[58]:

	Zip Code	City	Zillow Home Value Index	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
86	60475	Steger	93100	5	Golf Course	Fast Food Restaurant	Auto Garage	Southern / Soul Food Restaurant	Moving Target	Candy Store	Liquor Store	Sausage Shop	Sandwich Place	Sporting Goods Shop

```
In [59]: suburb_merged_5.loc[suburb_merged_5['Cluster Labels'] == 6, suburb_merged_5.columns[[0] + [1] + [2] + list(range(5, suburb_merge
         d_5.shape[1]))]]
```

Out[59]:

	Zip Code	City	Zillow Home Value Index	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
96	60480	Willow Springs	259200	6	Trail	Lake	Coffee Shop	Convenience Store	Nature Preserve	Italian Restaurant	Park	Bar	Brewery	Ice Cream Shop

Personal experience leads me to believe these communities have many more entertainment and service venues than those returned by Foursquare. In order to improve clustering results and interpretation of the data I decided to remove any communities that returned 30 or less venues from Foursquare. Additionally several communities reached the limit of 100 venues returned these data were used in the analysis despite not having access to data for all venues in these locations. Below is a list of zip codes that returned high and low venue counts.

60005	100	60469	39
60805	100	60478	37
60706	100	60465	34
60712	100	60007	34
60707	100	60163	33
60456	100	60558	33
60018	100	60091	33
60304	100	60472	32
60130	100	60422	32
60076	100	60458	32
60415	100	60067	31
60302	100	60464	30
60402	100	60471	27
60202	100	60192	23
60203	100	60476	22
60173	100	60475	18
60453	100	60467	18
60077	100	60439	17
60546	100	60062	16
60525	100	60480	13
60305	100	60411	4
60301	100	60010	4
60201	100		

The foursquare location data was prepared for analysis by k-means clustering by transforming the data into numeric format by one-hot encoding. After removal of low venue count suburbs a total of 88 suburban neighborhoods containing over 6000 venues in over 300 categories were left for analysis.

A dataframe including the top 10 venues in each neighborhood was created to use in the analysis as shown below.

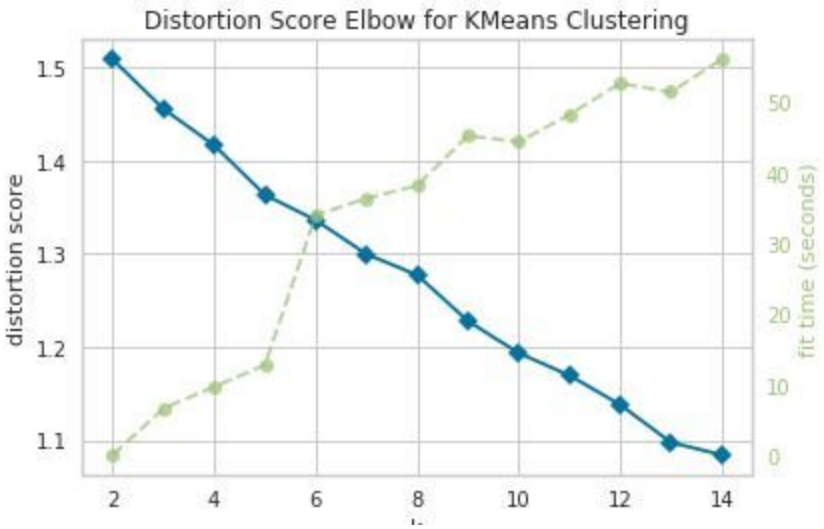
	Zip Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	60004	Park	Grocery Store	Pizza Place	Liquor Store	Bakery	Chinese Restaurant	Clothing Store	Thrift / Vintage Store	Pet Store	Gift Shop
1	60005	Sandwich Place	Italian Restaurant	Sushi Restaurant	Mexican Restaurant	Pizza Place	Bar	Cosmetics Shop	Thai Restaurant	Park	Ice Cream Shop
2	60007	Pizza Place	Mexican Restaurant	Sandwich Place	Bakery	Park	Japanese Restaurant	Shipping Store	Supermarket	Liquor Store	BBQ Joint
3	60008	Sandwich Place	Pizza Place	Fast Food Restaurant	Racetrack	Sports Bar	Coffee Shop	Donut Shop	Skating Rink	Salon / Barbershop	Mexican Restaurant
4	60016	Mexican Restaurant	Donut Shop	Fast Food Restaurant	Sandwich Place	Bar	Bakery	ATM	Breakfast Spot	Café	College Cafeteria

Methodology

Using IBM Watson Studio this project explored the suburbs of Chicago. Data from Zillow and Foursquare were used to create a dataframe that could be used to compare the different suburbs. Foursquare returned over 6,000 venues from over 300 different categories. Using k-means clustering the suburbs were grouped into clusters according to similarity of their Foursquare data. After the clusters were determined the folium library was usued to visualize the clusters on a map of the Chicagoland area. Cluster labels were added to the dataframe so the data could be explored and compared with a focus on differences in median home values for each suburb.

Use of KElbowVisualizer to determine if there is an optimal K value

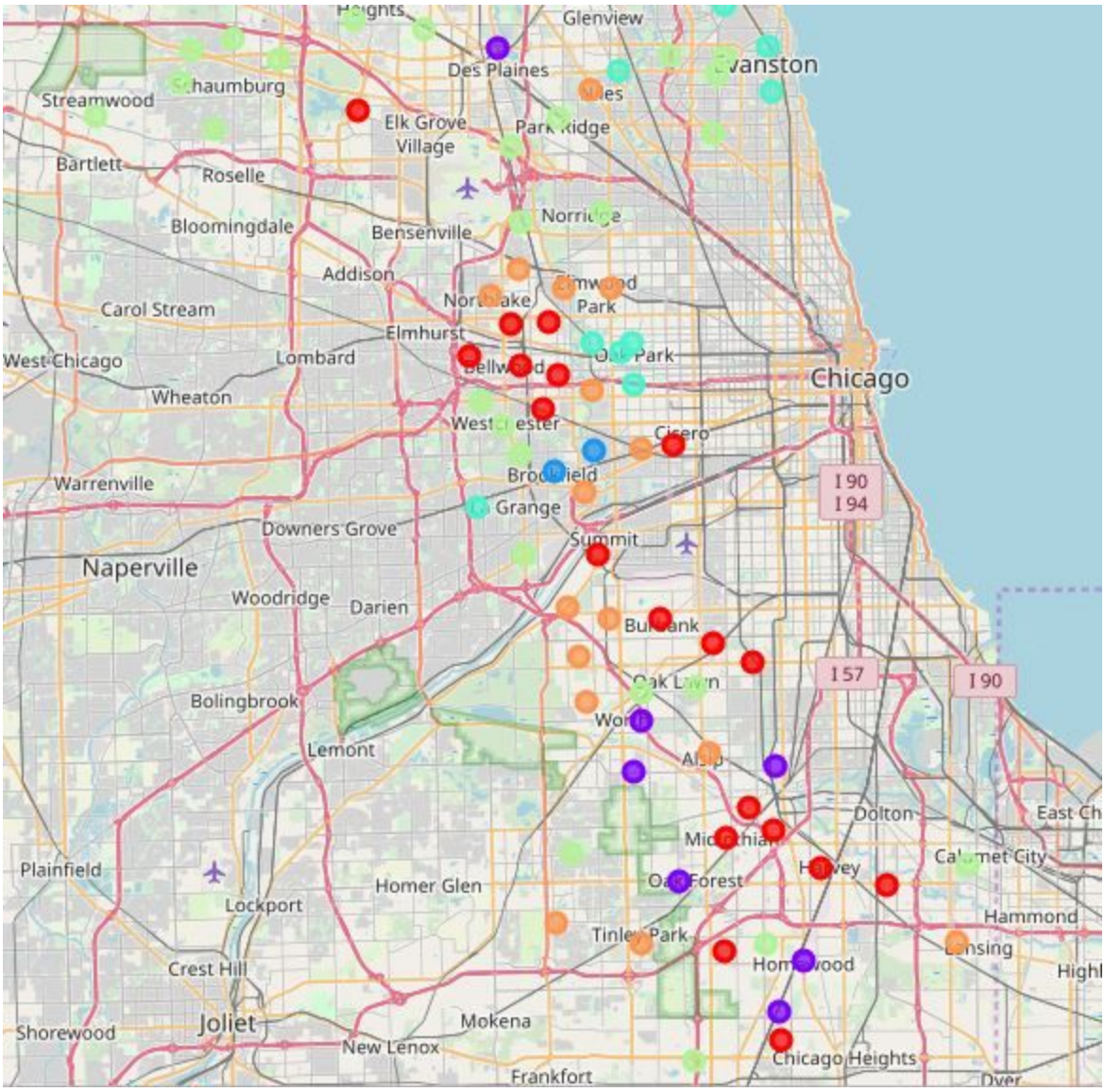
KElbowVisualizer was used to determine if there was an optimal k value for clustering data. Although a k of 6 was recommended the plot was relatively smooth indicating that an optimal k value was not obvious. For this reason several other k values were run to see if further insight could be gained from these models. Analysis of data focused on clusters created using k's equal to 6 and 3.



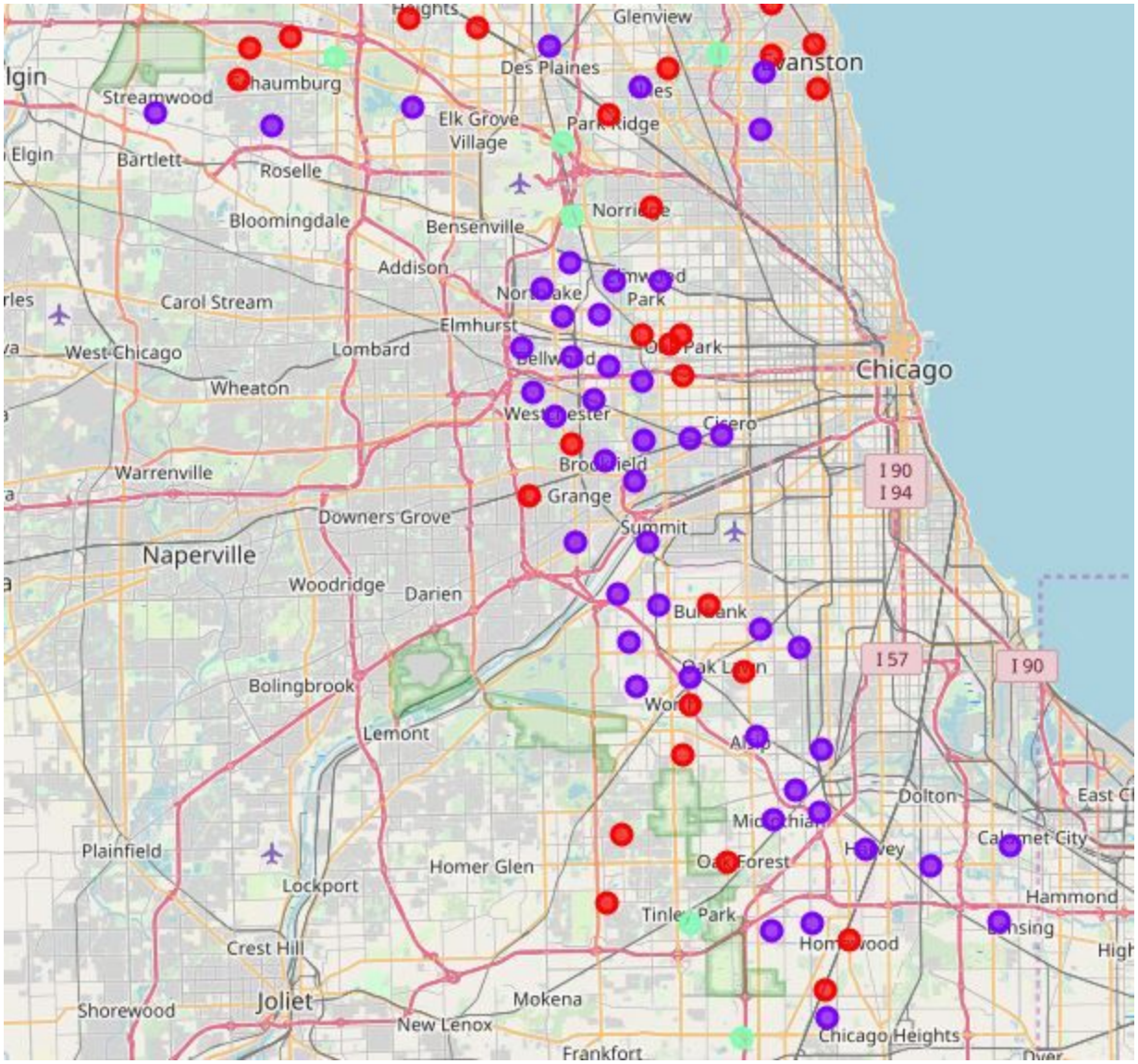
Results

Below are Visualizations and dataframes that explore the Foursquare and Zillow Home Value data. Elbow Visualization of k-means clustering was not conclusive so several k values were run for comparison.

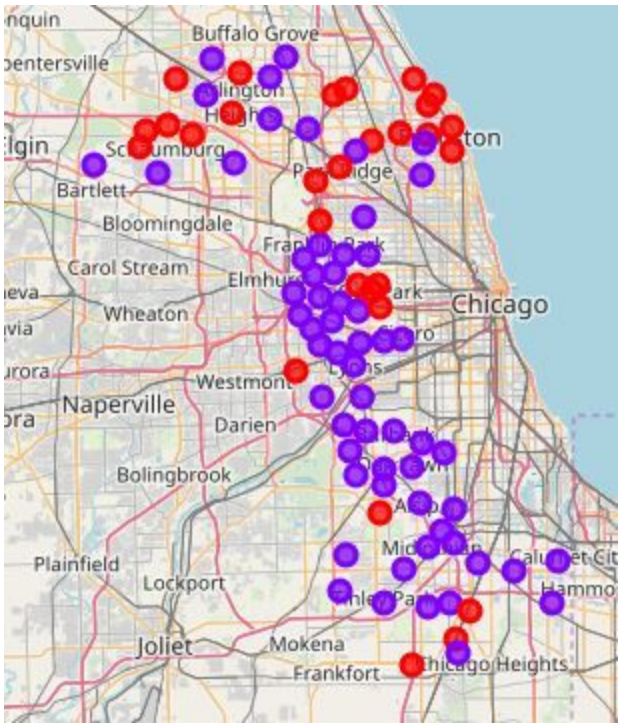
Visualization of 6 Clusters



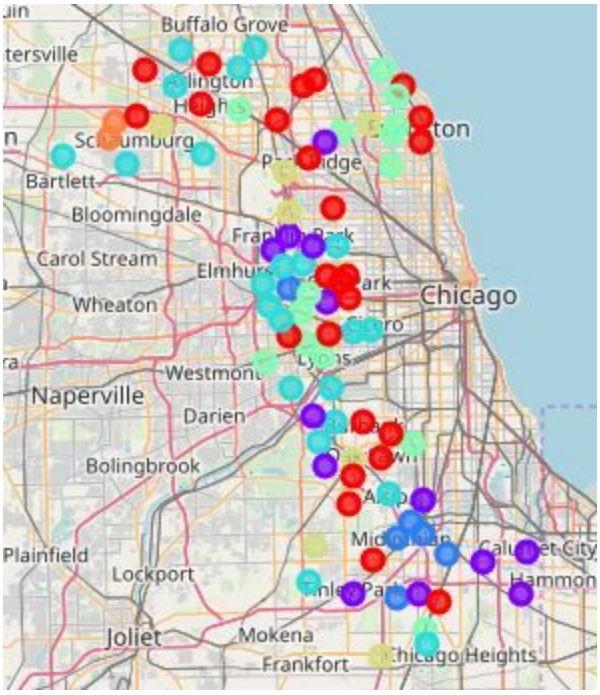
Visualization of 3 Clusters



Visualization of 2 Clusters



Visualization of 7 Clusters



Analysis focused on clustering with k values of 6 and 3. Clustering was performed on Foursquare data, interestingly the clusters returned in most cases grouped suburbs with high median home values into the same cluster despite this information not being used in the k-means analysis. This leads me to believe that high median home value suburbs do have similar entertainment and service venues as returned by Foursquare. The highest median home suburbs are all north shore suburbs. With a k value of 6 the suburbs of Oak Park, Evanston and Morton Grove had the lowest median home values while having similar entertainment and service options compared to the highest median home value suburbs such as Kenilworth and Winnetka as shown below.

	Zip Code	City	Zillow Home Value Index	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
60	60301	Oak Park	180500	3	Italian Restaurant	American Restaurant	Ice Cream Shop	Coffee Shop	Sandwich Place	Park	Breakfast Spot	Grocery Store	Historic Site
23	60202	Evanston	270400	3	Coffee Shop	Pizza Place	Brewery	Sushi Restaurant	Park	Bakery	Thai Restaurant	American Restaurant	Gym
50	60053	Morton Grove	300500	3	Pizza Place	Pharmacy	Donut Shop	Diner	Gym	Spa	Supermarket	Cafeteria	Gas Station
58	60302	Oak Park	343100	3	Park	Historic Site	Italian Restaurant	Mexican Restaurant	Gym	Coffee Shop	Grocery Store	Thai Restaurant	Bar
59	60304	Oak Park	353100	3	Italian Restaurant	Park	Pizza Place	Ice Cream Shop	Coffee Shop	Bar	Fast Food Restaurant	Hot Dog Joint	Breakfast Spot
22	60201	Evanston	417300	3	Coffee Shop	American Restaurant	Bakery	Pizza Place	Bar	Sushi Restaurant	Grocery Store	Café	Gym
73	60305	River Forest	510500	3	Coffee Shop	Italian Restaurant	American Restaurant	Pizza Place	Fast Food Restaurant	Ice Cream Shop	Sandwich Place	Gym	Donut Shop
94	60558	Western Springs	520500	3	Coffee Shop	Grocery Store	American Restaurant	Park	Sports Bar	Nature Preserve	Golf Course	Theater	Gas Station
30	60026	Glenview	552800	3	Pizza Place	Coffee Shop	Sandwich Place	Shopping Mall	Gym	Bank	Sporting Goods Shop	Breakfast Spot	Bar
97	60091	Wilmette	624200	3	Coffee Shop	American Restaurant	Pizza Place	Golf Course	Park	Ice Cream Shop	Diner	Playground	French Restaurant
98	60093	Winnetka	1019100	3	Grocery Store	Park	Train Station	Café	Mexican Restaurant	Coffee Shop	Pizza Place	Bank	Bookstore
40	60043	Kenilworth	1316800	3	Pizza Place	Park	Gym	Pet Store	American Restaurant	Pharmacy	Sushi Restaurant	Supermarket	Beach

With a k value of 3 the list of suburbs similar to the highest median home value suburbs grows and suburbs with even lower median home values are included. The suburbs Homewood, Schaumburg, Worth, Oak Forest and Oak Lawn had the lowest median home values while still being in the same cluster as the highest median home value suburbs Kenilworth and Winnetka. These lower median home value index suburbs are shown in the snippet of the dataframe shown below.

	Zip Code	City	Zillow Home Value Index	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
38	60430	Homewood	155700	0	Bar	Pizza Place	Baseball Field	Ice Cream Shop	Deli / Bodega
81	60195	Schaumburg	159400	0	Indian Restaurant	Grocery Store	Fast Food Restaurant	Korean Restaurant	Pet Store
99	60482	Worth	173000	0	Middle Eastern Restaurant	Bar	Fast Food Restaurant	Sandwich Place	Pizza Place
60	60301	Oak Park	180500	0	Italian Restaurant	American Restaurant	Ice Cream Shop	Coffee Shop	Sandwich Place
56	60452	Oak Forest	189900	0	Bar	Pizza Place	Breakfast Spot	Sports Bar	Ice Cream Shop
57	60453	Oak Lawn	195200	0	American Restaurant	Ice Cream Shop	Bakery	Sandwich Place	Mobile Phone Shop
11	60459	Burbank	196200	0	Pizza Place	Deli / Bodega	Intersection	Gas Station	Burrito Place
79	60194	Schaumburg	199800	0	Indian Restaurant	Gym	Baseball Field	Convenience Store	Pizza Place
26	60422	Flossmoor	214900	0	Park	Hospital	Train Station	Pizza Place	Pharmacy
35	60169	Hoffman Estates	233900	0	Indian Restaurant	Park	Grocery Store	Fast Food Restaurant	Rental Car Location

Discussion

This project aimed to find the suburbs with the lowest median home values that were still similar to the highest median home value suburbs by comparing entertainment and service venues returned by Foursquare using k-means clustering. While there are other variables that potential homeowners and real estate investors may value more, such as public school districts and crime and safety statistics, the information explored in this project still could be of interest to homebuyers.

It would be interesting to include more data such as crime and school information to see how clustering would change, however there are many real estate sites that provide this information. These real estate sites do not include as much information on entertainment and service venues available so by comparing this project with the information available on those sites a homebuyer may be able to make a better decision for their own needs. In particular homebuyers that have minimal interest in available schools or that have a lot of interest in available restaurant and other entertainment options might find this project interesting.

There is a major assumption made in this project that the suburbs with the highest median home values also have the most desirable options for entertainment and services. This may not be the

case and any individual may have very different tastes than the people who live in these high value areas.

The data compiled in this project could be analyzed in other ways by individuals with different needs. For instance the same data could be used to find suburbs that are priced extremely high but that have similar entertainment and service venues as lower priced suburbs. This could indicate that homes in these areas are overpriced or that these areas are prime locations for business investment as the local homeowners may have higher disposable incomes but a lack of entertainment venues where they can spend this income.

Conclusion

The Chicagoland suburbs of Cook County have varied median home values and entertainment and service options. This project aimed to explore the relationship between these variables.

K-means clustering of Foursquare data when compared to Zillow median home values returned some very interesting results in this project. The data provided could help potential homebuyers decide on a neighborhood that they would like to invest in. Further analysis could help business investors determine locations for new entertainment and service venues in areas that are lacking these options but that have home values that suggest residents have available disposable income.

This analysis could be replicated on any number of other locations throughout the United States or throughout the world where this type of data is available.

For k = 6 the suburbs of Oak Park, Evanston and Morton Grove had the lowest median home values in the highest median home value cluster.

For k = 3 the suburbs of Homewood, Schaumburg, Worth, Oak Park, Oak Forest and Oak Lawn had the lowest median home values in the highest median home value cluster.