1. Introduction

1.1 Background

As of the year 2020, the United States housing market is worth an estimated 36.2 trillion dollars. The underlying factors that go into this astronomical amount can be debated, but when it comes to the purchasing of a home, it is the individual that must determine a house's value and decide what metrics that value is based on. As the old allegory 'location, location, location' would indicate, neighborhood characteristics play a major role in the property values of the homes within them. Landing a dream house in the middle of the desert does not leave many options for dinner, regardless of the price paid. Fundamentally, a house is an asset, and to properly value it you cannot overlook its' location. 78

1.2 Problem

Chicago Illinois is one of the most diverse and economically developed cities in the US. The problem this project will focus on is determining the locations in Chicago that a new home buyer will have the highest likelihood of retaining property value based on the characteristics of the surrounding neighborhood.

1.3 Interest

This information would be of interest to new Chicago home buyers, Chicago real estate investment firms, Chicago municipal departments focused in community development and small business owners looking to locate themselves near higher valued neighborhoods.

2. Data

2.1 Data Sources

Chicago is divided into 77 Community Areas. To understand the population characteristics, general information can be pulled from Wikipedia using python's beautifulsoup library and looping through the individual Community Area pages. I will also be pulling data from HousingStudies.org, which utilized the census to build easily accessible tables which can be grouped by Community area. I will also use a complex dataset sourced from the Environmental Data Initiative, which contains a study that analyzed tree canopy in 37 major US cities(including Chicago). Obviously, to understand the surrounding venue characteristics and to complete the picture of the neighborhood, I will be using Foursquare.

2.2 Data Uses

The level of foreclosure activity, mortgage activity, population size, income/poverty levels and tree canopy will be the primary clustering set for the neighborhoods. Each of these elements plays a major role in the shape of the neighborhood and the value within it. Chicago is a notoriously violent city, but I have deliberately left out all demographic data and crime statistics (even though I originally included it) to try to level the obvious bias of the clustering. Once clustered, the venue data will be used to compare community areas that have higher levels of desirable venue types. It will also include finding the highest levels of various venue types in these clusters locations to gain insight in the comparability of neighborhood value and venue type.

2.3 Data Cleaning and Feature Selection

The selection of these specific elements was done by searching the qualities that new home buyers look for in a neighborhood. The though being that the higher the income, tree canopy and mortgage activity, the more attractive the neighborhood is to the general population, and subsequently the properties there are more valuable. The opposite thought being that the higher the foreclosure rate and lower the income and tree canopy, the lower the perceived value of the neighborhood.

The data pulled for foreclosure activity, mortgage activity, population size, and income/poverty did not need to be cleaned as the collection itself was coded to not include anything that would need to be removed. However, the data set to get the tree canopy used geolocation based multipolygons. Utilizing this dataset required the use of the *geopandas library*, which was used to map the polygons to depict percentages of tree cover. After, by utilizing the *shapely* library, I was able to take the geolocation of the Community Areas and check which corresponding polygon it fell into. Storing that polygon's recorded tree canopy percentage in relation to the Community Area as it's own dataset. I was able to then easily join this dataset into the clustering set by Community Area. The foursquare data was acquired using the API and used a 1500 meter radius around the Community Area geolocation point.

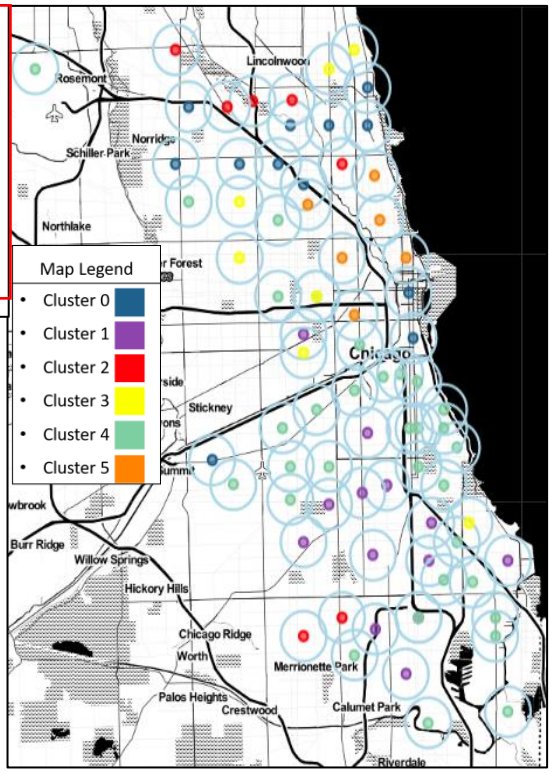
3. Methodology and Data Analysis

3.1 Community Area Clustering

To cluster the Community areas, I first determined the optimal cluster count by utilizing the *yellowbrick* library's KElbowVisualizer. This allows you to quickly see if there is an elbow and to use this point as your cluster count.

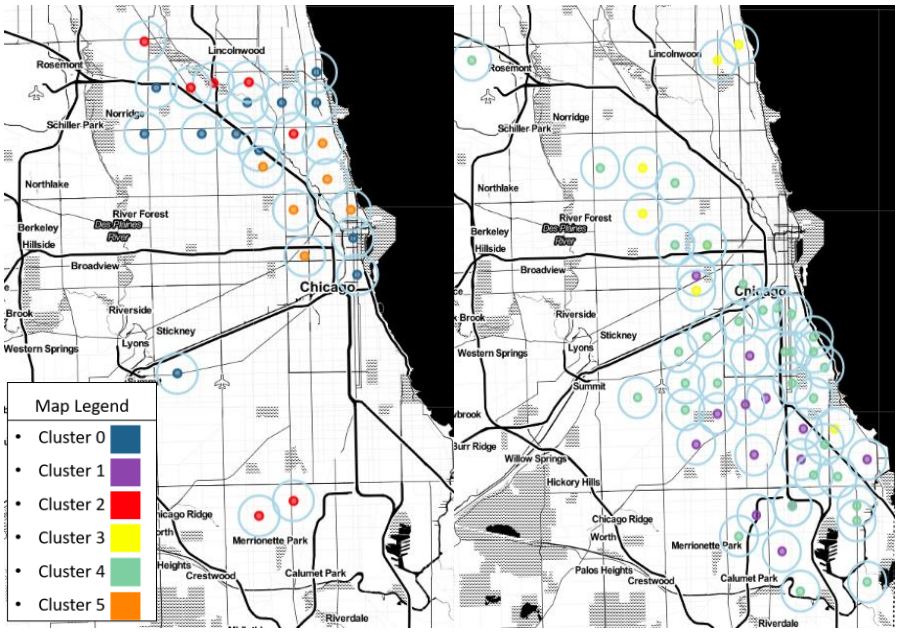
Next was to map the clustered Community Areas and to check the average values of the clustering elements in each cluster.

clus	Population	Poverty	Forclosures	Mortgages	Income	TreeCover	F_to_M
5	80725.333333	9134.070500	184.666667	8336.666667	92215.000000	19.601560	0.022151
2	21655.571429	1480.737857	91.714286	2113.142857	88921.285714	30.561541	0.043402
0	45662.500000	5399.229417	151.500000	3465.000000	68186.765152	17.079920	0.043723
3	69150.714286	17222.681714	364.000000	2599.857143	39794.597403	20.856551	0.140008
4	18645.062500	4076.919969	118.812500	813.843750	39574.750000	17.159290	0.145989
1	34265.307692	9135.112923	477.384615	1470.769231	34070.783217	21.694207	0.324582



The above dataframe shows Foreclosures to Mortgages(F_to_M) and has been sorted by F_to_M in ascending order. This shows clusters 5, 2 and 0 being magnitudes lower in this regard compared to clusters 3,4 and 1. In addition, the income level difference between these two sets of clusters is also substantial. For emphasis, these have been separated with red lines and boxes.

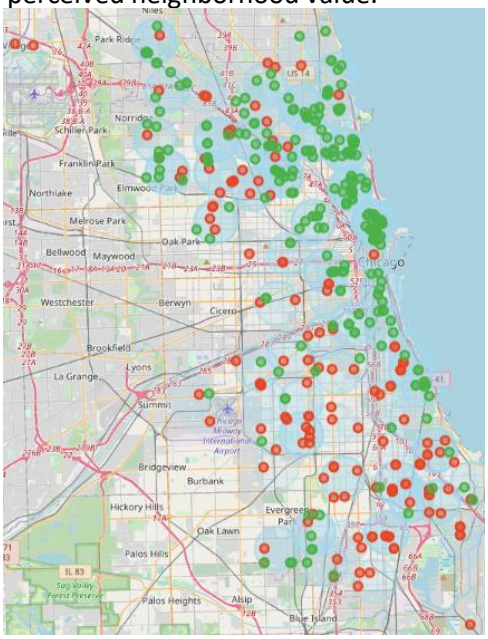
When viewed as separate groups it is much easier to see the separations between the locations of these two groups of clusters. It is not surprising that it should be north and south, as Chicago's south side has a reputation that is notorious worldwide. However the pockets of neighborhoods in the positive grouping that are outside the northern half of the city are interesting, as they show movement to develop new areas of value outside of the established most expensive communities.



3.2 Venues In Community Areas

After getting the venue data from Foursquare, subsets had to be standardized using onehot encoding and grouped by the community area there were closest to. After doing so, it becomes possible to find the most common venue types by neighborhood or cluster, as well as to do analysis on the correlations between venues and the clustering elements. Three correlation tables below were used to help determine the venue groupings to check.

The sort elements have been circled in red for clarity. This shows the various venue types that show the highest correlation to Income, Mortgages, and Foreclosures respectively. The highest correlation values (minimum 0.5) were then used to categorize subsets of the venue data to create a map showing the locations of venues with a positive effect(green) and a negative effect(red) on perceived neighborhood value.



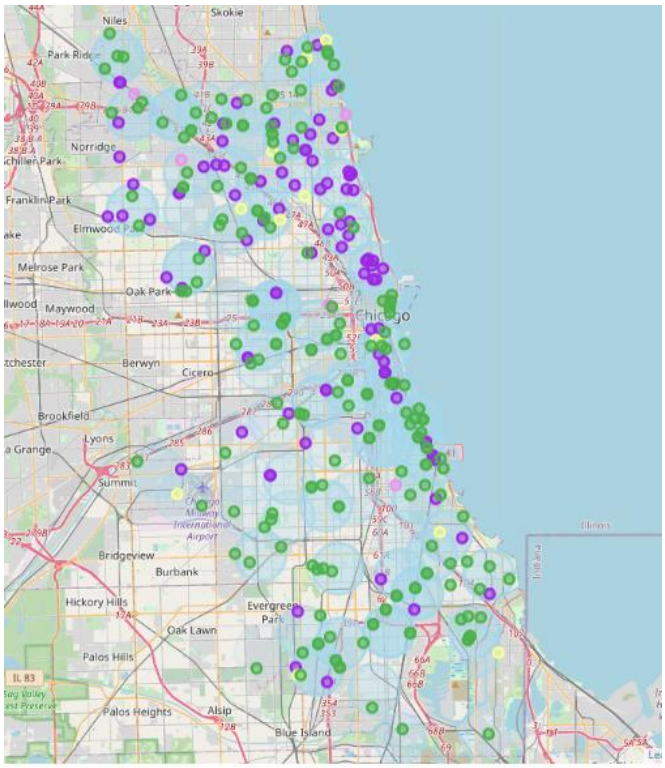
	Poverty	Forclosures	Mortgages	Income
Income	-0.269093	-0.223254	0.670603	1.000000
Italian Restaurant	-0.159296	-0.237101	0.577470	0.685410
Mortgages	0.253647	0.104134	1.000000	0.670603
French Restaurant	-0.070403	-0.136996	0.382434	0.470583
New American Restaurant	0.014177	-0.160386	0.637974	0.464043
Beer Garden	-0.173391	-0.151553	0.154179	0.458813
Mediterranean Restaurant	-0.007979	-0.168370	0.394180	0.453341
Pub	-0.215925	-0.182898	0.214356	0.446854
Salon / Barbershop	0.010946	-0.034896	0.417963	0.419293
Toy / Game Store	-0.018953	-0.022842	0.394279	0.417235

	Poverty	Forclosures	Mortgages	Income
Mortgages	0.253647	0.104134	1.000000	0.670603
Population	0.698121	0.350474	0.813885	0.340984
Income	-0.269093	-0.223254	0.670603	1.000000
New American Restaurant	0.014177	-0.160386	0.637974	0.464043
Cycle Studio	0.066618	0.018520	0.582621	0.285568
Italian Restaurant	-0.159296	-0.237101	0.577470	0.685410
Japanese Restaurant	0.027837	-0.128305	0.570518	0.283520
Brazilian Restaurant	0.147336	-0.036800	0.505907	0.291020
Hostel	0.033778	-0.040824	0.493646	0.340177
Music Venue	0.015317	-0.111570	0.490283	0.364918

	Poverty	Forclosures	Mortgages	Income
Forclosures	0.545242	1.000000	0.104134	-0.223254
Poverty	1.000000	0.545242	0.253647	-0.269093
Discount Store	0.199593	0.537711	-0.338347	-0.466332
Fast Food Restaurant	0.122884	0.468881	-0.334026	-0.432946
Currency Exchange	0.271231	0.460968	-0.047620	-0.217297
Fried Chicken Joint	0.065505	0.379511	-0.236340	-0.295624
Population	0.698121	0.350474	0.813885	0.340984
Fish & Chips Shop	-0.042340	0.315353	-0.077093	-0.076614
Convenience Store	0.026426	0.294519	-0.079979	-0.080689
Pharmacy	0.163515	0.285346	-0.238089	-0.236009

It is easy to see that these match the distribution of the clustering groups mentioned earlier. The green points are much higher in frequency towards the north, while the red is much higher in frequency toward the south. This makes sense considering the items were chosen through their correlation to the same elements that these areas were clustered on.

In addition, a similar map was generated to visualize the location distribution of some of the most desirable public facilities and venues. This included parks, gym/fitness centers, playgrounds, and dog parks. The map to the right shows the location of public parks(green) which are plentiful throughout the city of Chicago. However, fitness centers(purple), dog parks(pink) and playgrounds(yellow) have a very skewed distribution to the same three community clusters highlighted earlier (5, 2 and 0). This only helps to further the conclusion of the analysis.



4. Results

The results show that based on the analysis done, that the community areas contained in clusters 5, 2, and 0 will have higher perceived value from the average consumer. Further, this would indicate that buying a home in one of these three clusters would have a higher likelihood of appreciating in value than would buying in one of the other three clusters. This is based on the communities in clusters 5, 2, and 0 averaging much higher income levels and mortgage activity while being substantially lower in foreclosures relative to the other three clusters. In addition, and most likely due to the previous, the venues located in these three clusters are indicative of higher neighborhood value, and subsequently higher property values.

5. Discussion

I would like to clarify that this conclusion is drawn from a completely objective and data-based standpoint. If you were to look at the demographic make-up of the clusters, the lower perceived value community areas are home to a much higher percentage of minorities and people of color . Demographics were specifically not used, and the goal was to see houses strictly in the capacity of that of an asset. In order to accurately assess risk and valuation for the specific kind of asset that is property, determining the effect that the asset's location has is paramount in assuring the best return on invest. In no way shape or form is this project a comment on, or a valuation of; people, culture or community.

In addition, there appear to be a surprisingly low amount of dog parks and playgrounds in general. This could be because they were labeled as parks, or missed entirely by foursquare, but it would indicate that there is certainly room for more in Chicago if this is an accurate count. Another interesting observation was the correlation of tree cover to higher perceived value in a community area or neighborhood. Any online search for things that add value to a neighborhood will result in something to do with trees. It is interesting to see that this is actually the case, especially when excluding the city center, as it obviously has very little tree cover.

6. Conclusion

To answer the question of where someone newly moving to Chicago should look to purchase property, using the higher perceived value community areas above will increase the likelihood of value retention. These areas posses more wealth per capita, more demand for mortgages, less foreclosures, and contain a much higher degree of desirable venue types. Obviously the closer you get to the city center, the more expensive things become. However, this becomes interesting because the higher perceived value community areas from this analysis span all the way around the city, not just the obvious old money areas in the north. Specifically, Beverly on the lower west side was in the higher valued cluster and still has good access to public transit into the city.