Analysis of Best Neighborhoods to Move Into in Massachusetts

Nirushan Udayakumar

Introduction

It is very common for people to undergo relocation for work. It is a very stressful process, and yet despite that, since 1999 between 2.8 and 4.5 million people relocate for work every year in the US. In fact, as of 2019, 62% of workers in the US said they would relocate for work.

Relocation is as stressful as it is because there are so many factors to consider, and the research required can take days. It is such an intensive process that you would not want to relocate more than once simply because you did not perform your research properly the first time. I took on this capstone to tackle the problem of performing research into where one should relocate. This report is targeting those who would consider relocating for work, but are not looking forward to performing the research required. This tool could also be valuable to employers, to offer as a starting point for new hires who require relocation.

Some of the factors that need to be considered are the cost of housing, safety of the neighborhood, and nearby venues and amenities. This capstone will aggregate these factors for each neighborhood in a given state in the US (ie. Massachusetts), identify nearby venues based on what is important to you (such as food venues or gyms), and then cluster similar neighborhoods.

This tool should allow its users to evaluate neighborhoods based on their own merits and no other external factors, such as distance to job, as people are more likely to switch jobs than they are to relocate; it is normal to switch jobs within a metropolitan area.

This report will walk you through the data that was used, the analysis performed on the data, the results of the analysis, and will then make recommendations and conclusions of the data.

Data

As mentioned previously, some of the key pieces of information required for relocation are the cost of housing for the neighborhood, the safety of the neighborhood, and nearby venues. There are other important factors as well, but they may vary a lot between people, so they will not be covered in this report.

There are four main sets of data that will be utilized in this project:

- Neighborhood location data
    - URL: https://www.zipcodestogo.com/Massachusetts/
- House pricing data
    - URL: https://www.rentdata.org/states/massachusetts/2020
- Safety data
    - URL: https://www.safewise.com/blog/safest-cities-massachusetts/
- Foursquare venue data

The neighborhood location data contains the zip codes of each neighborhood in the given state; some neighborhoods span multiple zip codes. The housing price data set has the average rental price for each of a 0 to 4 bedroom apartments in each neighborhood. The safety data set includes the safety ranking, as well as the rates of property and violent crimes, of each of the 139 safest neighborhoods in the state. Finally, Foursquare is used to extract information of nearby venues for each neighborhood.

Several steps of data cleaning were performed to each data set. In brief, the following processes were performed:

- Neighborhood location data
    - The table of zip codes and neighborhood names was imported and only those two columns were kept
- House pricing data
    - The table of rental pricing was imported. Logically, the prices of each type of apartment should be relative, so we will arbitrarily decide to keep the 1-bedroom column and drop the others
    - The words “town” and “city” were deleted from each neighborhood name, to allow for merging of dataframes later
    - “Metro” in each neighborhood was split into its own column, and then wasused to drop neighborhoods that are not considered a metro
- Safety data
    - The top 20 safest neighborhoods are displayed as infographics, so the data was scraped using BeautifulSoup - specifically focusing on the name of the neighborhood, their safety ranking, and the violent crime (VC) & property crime (PC) rates
    - The rest of the neighborhoods were imported via the table and only relevant columns were kept
- Foursquare venue data
    - Foursquare was used to gather the number of a particular type of venue within a given distance of each neighborhood, based off user preferences
    - For the sake of this capstone, I have chosen to look at restaurants and gyms, as those are important to me
    - The tool will count the number of venues in the area for each category of venues and then compile the data into a dataframe

    - If there exists a neighborhood with either no restaurants or no gyms, the neighborhood was removed

The different data sets and dataframes, once cleaned, were then merged together to only capture neighborhoods that were found amongst all the data sets.

The geographical US state (ie. Massachusetts) and the venues of choice will be indicated in the beginning as global variables, so analyzing a different state or different venues is as easy as changing these variables.

The neighborhoods will be clustered and displayed on a map. This will be discussed in further detail in a later section.

Methodology

Exploratory Data Analysis

It has been decided that a clustering model will be used to analyze the neighborhoods - this is because we are dealing with a large dataset where we expect to see similarities and differences between neighborhoods and would like to group them by their similarities, such that users can best determine what neighborhoods would best suit their needs.

Prior to applying the clustering model, however, a little bit of exploratory data analysis was performed to see if there were any clear correlations between some of the variables in question; knowing this may give insights into the performance of the clustering model. For instance, if there is a strong correlation between two variables, it can be quite reasonably hypothesized that the clusters will differ greatly between those variables.

When relocating, one of the biggest factors people immediately consider is cost of housing. As a result of this, I chose to perform some rudimentary correlation analysis between cost of housing and other variables. The various zip codes were grouped by neighborhood such that there would be fewer entries plotted for visual analysis purposes. The results are seen below.

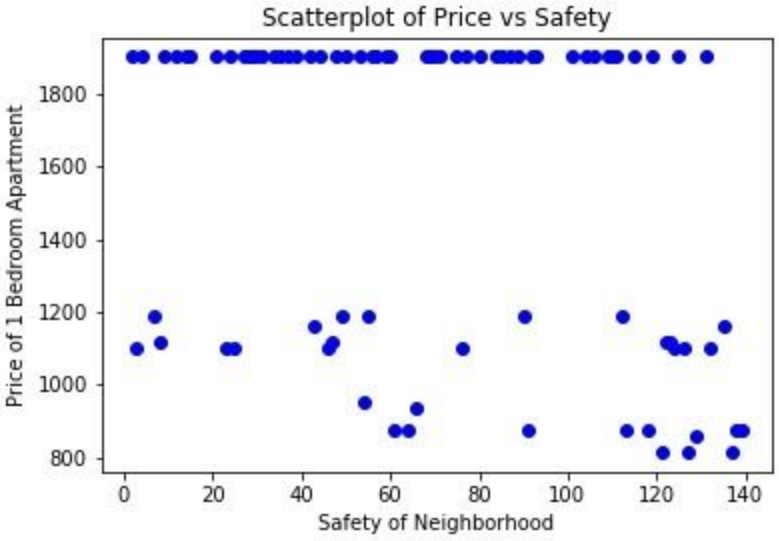

Figure 1 - Plot of the price of a 1 bedroom apartment vs the safety of the neighborhood

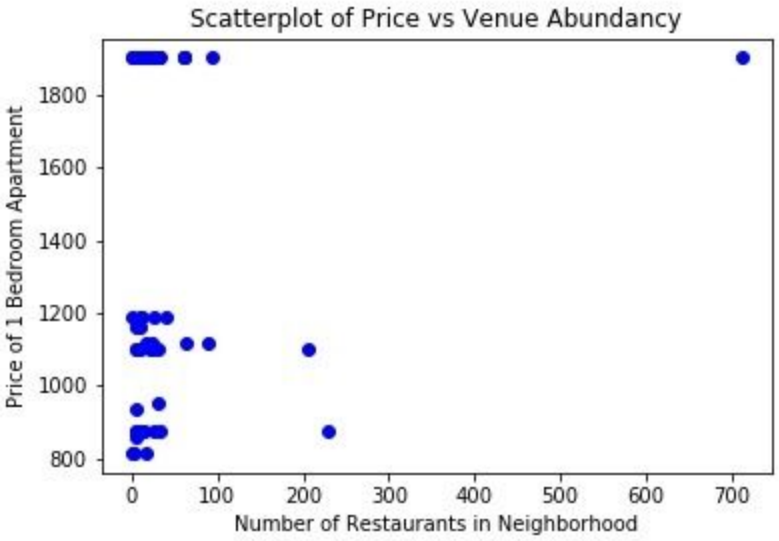

Figure 2 - Plot of the price of a 1 bedroom apartment vs the number of restaurants in the neighborhood

Figure 1 attempts to see the correlation between price and safety, with the hypothesis that safer neighborhoods have a higher cost of housing. Figure 2 attempts to see the correlation between price and abundance of venues/restaurants, with the hypothesis that neighborhoods with more restaurants will be in areas with more expensive housing.

As can be seen in both scatter plots, there is no clear correlation between the variables. In figure 2, there is a point that may appear to be an outlier in the top right corner, and this plot would give indication that the data point should be removed, however since this plot shows data points of grouped neighborhoods, this point will not be removed as it as accumulation of data points.

The fact that there is no clear trend or correlation between variables strengthens the fact that an unsupervised machine learning model should be used to properly analyze the data for similarities.

Machine Learning Methods

K means clustering was used as the ML method for this set of data. There are several reasons for this.

- There is no clear correlation between variables, so an unsupervised method would be ideal
- We are looking at similarities and differences between neighborhoods, so clustering is preferred
- Clustering is based off similarities in input variables and not off geographical location

The inputs used for the model for each zip code were: cost of housing, overall safety ranking, PC rate, VC rate, number of restaurants, and number of gyms. Six clusters were chosen as the k value for the algorithm, and then the clusters were analyzed. The table below shows the inputs for the zip codes, displaying only the first few rows.

	Zip Code	Neighborhood	1 BR	Safety Ranking	Violent Crime (per 1,000)	Property Crime (per 1,000)	Number of Restaurants	Number of Gyms
0	1001	Agawam	875	113	3.6	14.8	12	1
1	1002	Amherst	875	64	3	4.3	9	6
2	1003	Amherst	875	64	3	4.3	9	4
3	1004	Amherst	875	64	3	4.3	9	7
4	1040	Holyoke	875	139	9.7	40.5	14	2
5	1060	Northampton	875	118	4	18.9	2	2
6	1061	Northampton	875	118	4	18.9	15	3
7	1063	Northampton	875	118	4	18.9	15	4

Table 1 - A snippet of a table showing the inputs for each zip code and their associated values

Results

The following table will show snippets of the top, middle, and bottom of a table that displays all the zip codes that were used, their associated inputs for the model, and their cluster numbers produced as a result of the k means clustering algorithm.

	Zip Code	Neighborhood	Cluster Number	1 BR	Safety Ranking	Violent Crime (per 1,000)	Property Crime (per 1,000)	Number of Restaurants	Number of Gyms
0	1001	Agawam	1	875	113	3.6	14.8	12	1
1	1002	Amherst	1	875	64	3	4.3	9	6
2	1003	Amherst	1	875	64	3	4.3	9	4
3	1004	Amherst	1	875	64	3	4.3	9	7
4	1040	Holyoke	1	875	139	9.7	40.5	14	2
5	1060	Northampton	1	875	118	4	18.9	2	2
6	1061	Northampton	1	875	118	4	18.9	15	3
7	1063	Northampton	1	875	118	4	18.9	15	4
.									
.									
.									
49	1757	Milford	5	937	66	2.1	8	5	17
50	1760	Natick	4	1900	75	1.5	11.4	10	2
51	1806	Woburn	4	1900	53	1.2	9.4	3	2
52	1815	Woburn	4	1900	53	1.2	9.4	15	5
53	1888	Woburn	4	1900	53	1.2	9.4	14	5
54	1803	Burlington	4	1900	84	1.5	13.7	30	7
55	1805	Burlington	4	1900	84	1.5	13.7	30	4
56	1810	Andover	2	1117	8	0.2	4.8	30	6
57	1899	Andover	2	1117	8	0.2	4.8	4	1
58	5501	Andover	3	1117	8	0.2	4.8	30	12
59	1824	Chelmsford	2	1188	55	0.9	10.8	11	4
60	1826	Dracut	2	1188	49	1.3	8.7	1	4
.									
.									
.									

150	2301	Brockton	2	1160	135	9	21.6	4	2
151	2302	Brockton	2	1160	135	9	21.6	6	1
152	2324	Bridgewater	2	1160	43	2.1	4.2	4	1
153	2331	Duxbury	0	1900	24	0.4	5.7	5	2
154	2340	Hanover	0	1900	31	0.1	8.3	5	1
155	2359	Pembroke	0	1900	27	0.8	4.7	2	5
156	2361	Plymouth	0	1900	101	3.8	8.6	2	1
157	2362	Plymouth	0	1900	101	3.8	8.6	5	4
158	2368	Randolph	0	1900	106	3.1	13.2	6	2

Table 2 - Snippet of a table showing the inputs for each zip code, their associated values, and their cluster numbers produced from the k means algorithm

Six clusters (from 0 to 5) were produced by the algorithm, grouping similar neighborhoods together. The discussion section will provide insights into the results and make recommendations. To do so, table 3 was produced, which displays the mean value of each input for each of the clusters. These can be compared to each other between clusters and to the mean of all the neighborhoods.

	Cluster Number	1 BR	Safety Ranking	Violent Crime (per 1,000)	Property Crime (per 1,000)	Number of Restaurants	Number of Gyms
mean	0	1900.000	94.985	3.788	14.206	20.646	15.415
mean	1	868.667	119.250	6.887	21.054	14.333	3.667
mean	2	1140.100	85.100	3.655	11.540	14.750	3.650
mean	3	1117.000	8.000	0.200	4.800	30.000	12.000
mean	4	1900.000	73.643	2.482	10.139	11.786	5.893
mean	5	1065.381	104.238	4.914	16.195	15.190	6.667
mean, full table	1.786	1533.585	94.321	4.135	14.392	16.730	9.308

Table 3 - A summary of each cluster, showing the mean values for each input for all the clusters

Discussion

Table 3 was used and analyzed to produce table 4, which shows the percent variation of each variable for each cluster, relative to the average of the entire data set.

	Cluster Number	1 BR	Safety Ranking	Violent Crime (per 1,000)	Property Crime (per 1,000)	Number of Restaurants	Number of Gyms
mean	0	23.89%	0.70%	-8.40%	-1.29%	23.41%	65.61%
mean	1	-43.36%	26.43%	66.56%	46.29%	-14.32%	-60.61%
mean	2	-25.66%	-9.78%	-11.61%	-19.82%	-11.83%	-60.79%
mean	3	-27.16%	-91.52%	-95.16%	-66.65%	79.32%	28.92%
mean	4	23.89%	-21.92%	-39.98%	-29.55%	-29.55%	-36.69%
mean	5	-30.53%	10.51%	18.84%	12.53%	-9.20%	-28.38%
mean, full table	1.786	1533.585	94.321	4.135	14.392	16.730	9.308

Table 4 - The percent variation of each input for each cluster when compared to the mean of the entire data set. Darker colors in the gradient indicate a more favorable result.

These two tables can be used to gain insights and draw conclusions about each of the clusters.

- Neighborhoods in cluster 0 tend to be more expensive in terms of housing costs, are average in terms of safety, and have plenty of venues (restaurants and gyms).
- Neighborhoods in cluster 1 are by far the cheapest, have a very poor safety rating, have a slightly lower than average number of restaurants, and very few gyms.
- Those in cluster 2 are cheaper than average, slightly safer than average, have a slightly lower than average number of restaurants, and a very low number of gyms.
- Neighborhoods in cluster 3 are cheaper than average, extremely safe, and have plenty of venues; however, there is only 1 zip code that falls within this cluster.
- Those in cluster 4 are more expensive than average, are a little lower than average in terms of safety, and have fewer than average number of venues.
- Finally, neighborhoods in cluster 5 are cheaper than average, slightly safer than the average (though they have a high property crime rate), have a slightly lower than average number of restaurants, and fewer than average number of gyms.

This information is summarized in the form of pros/cons in table 5.

Cluster Number	Pros	Cons
0	- Average safety - Plenty of venues	- More expensive than average
1	- Cheapest neighborhood	- Poor safety rating - Slightly lower than average # of restaurants - Very few gyms
2	- Cheaper than average - Slightly safer than average	- Slightly lower than average # of restaurants - Very few gyms
3	- Cheaper than average - Extremely safe - Plenty of venues	- Only 1 zip code falls within this cluster
4	- No radical characteristics (none are far higher or lower than the average)	- More expensive than average - Slightly lower than average safety ranking - Fewer than average # of venues
5	- Cheaper than average - Slightly safer than average	- High property crime rate - Slightly lower than average # of restaurants - Fewer than average # of gyms

Table 5 - Table summarizing the pros and cons of each cluster

As price is usually the strongest driving factor for people, I will attempt to make recommendations based on this assumption. As cluster 3 is a bit of an anomaly, I will ignore it.

- For those who are price insensitive (can and are willing to spend more on housing), look at either clusters 0 or 4, where cluster 1 is preferred for those who value having a number of venues nearby
- For those who are extremely price sensitive (cannot and/or are not willing to spend very much on housing), I would recommend neighborhoods in cluster 1, though it comes with the trade-off of safety
- For those who are looking for sensible options in terms of pricing, look at clusters 2, 3 or 5. Cluster 5 would be preferred over cluster 2 if having access to slightly fewer gyms is important, but it comes with higher property crimes so I would recommend a home security system
    - Cluster 3 is your most well-rounded cluster - none of the neighborhoods excel in any category, but they also do not have any major issues

Conclusion

This report addresses the issue of relocation, where the primary audience is those who must relocate for work. This report performs research into the neighborhoods of a US state to cluster similar neighborhoods based on factors such as housing costs, safety, and access to venues. The report focuses on Massachusetts, but the state in question can change quite easily in the code by the global variable defined at the top.

The data was pulled from a number of sources, cleaned, analyzed, and then a k-means clustering algorithm was performed to cluster the neighborhoods. Six clusters were formed, if this report were to be redone, five clusters would be more appropriate, as one of the clusters only contained a single neighborhood and could likely have been grouped within one of the other clusters.

Aside from changing the number of clusters, future work can also include performing the same analysis for other US states to which people often relocate for work, such as New York and California.