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.