**Comparison of Massachusetts Northshore**

**Communities for Settlement**

Lawrence E. Field

Coursera/IBM Data Science Professional Cert Capstone Project

Table of Contents

[I. Abstract 3](#_Toc2544795)

[II. Introduction 3](#_Toc2544796)

[III. Data 3](#_Toc2544797)

[IV. Methodology 3](#_Toc2544798)

[V. Results 4](#_Toc2544799)

[VI. Discussion 4](#_Toc2544800)

[VII. Conclusion 5](#_Toc2544801)

[VIII. Recommendations 5](#_Toc2544802)

# Abstract

The communities of the area of Massachusetts self-identified as “the Northshore” were selected for the sake of studying them to identify options and homeowner and renter criteria for settlement in this area. The communities of Saugus, Lynn, Nahant, Swampscott, Marblehead, Salem, Peabody, Danvers and Beverly were included in the review. From the standpoint of venues that support the communities and provide diversion to residents, the communities were found to be largely homogenous. The only discriminator was isolation created by geography of the coastline. Other discriminators, such as property taxes, schools and property prices are necessary to create distinctions. The effort necessary to extract this data is beyond the scope of this project.

# Introduction

The Northshore of Massachusetts is a residential region for many people who work in Boston, Massachusetts. In addition, the “Technology Belt” along Route 128 and Techno-pharma in Cambridge are attracting more people every year. Many of these people prefer to live along the coast and engage in the New England seaside way of life; fishing, sailing and diving.

The purpose of this project is to discover discriminators that can be used by people new to the area when they shop for a new residence. The goal of this first phase effort is to review venues for residents’ support, such as laundries and grocery stores and entertainment venues.

Maps from the Jupyter Notebook are very large and produce and extremely large Word file that does not act well on my computer. For that reason, I have removed figures and, instead, reference the reader back to the Notebook.

# Data

Zip codes were used as primary parameters for organization. This was because communities can have more than one zip code. Files were obtained from the USPS. One was a zip code list for the particular communities. The other was a file that listed the latitude and longitudes for zip codes.

Google Maps was used to obtain the names of the various names in the communities.

The Foursquare API was the source of venue data.

# Methodology

Both of the zip code files were downloaded from web sites. The file containing the latitude and longitude data was found to contain errors: several different zip codes shared the same latitude and longitude in the file. Review of the dataframe head suggested the problem and scan of the map create by folium confirmed it in two communities, Lynn and Danvers. The problem was rectified by identifying major landmarks in the effected zip code on the folium map and using nominatim to find their latitude and longitude. The zip code dataframe was updated with the new zip codes. A new folium map was created and the correction to the zip code data confirmed.

The City of Lynn, being the largest of the communities, was subjected to a preliminary study. Lynn has 4 zip codes. Lynn showed substantial homogeneity for 3 zip codes. The fourth zip code was completely different, having little in the way of venues at all. A formal clustering analysis was neither possible nor necessary.

The whole of the Northshore region was subjected to a clustering analysis. A total of 7 clusters seems to provide the information that clustering will provide. At this level of ‘granularity’, a core homogenous region is clearly defined and other regions are distinctive.

# Results

The Jupyter Notebook contains the map that results from repairing the zip code errors. The City of Boston is visible on the bottom center of the map. The Northshore communities are represented by a total of 13 zip codes. Except for the corrected zip codes, the latitudes and longitudes of the zip codes are those of the town post offices, not of the town centers. This was not considered a problem because post offices tend to be located in the commercial centers of communities.

A subsequent map in the Notebook shows the map of Lynn after the zip code latitude and longitude errors had been corrected. For the Wyoma neighborhood, Union Hospital was chosen as a central landmark. Essex Street served for East Lynn and the North Shore Community College Lynn Campus was used for Central Lynn. The location of zip code 01905 was not changed even though Barry Park was more central to West Lynn. It is the post office location and leaving the coordinates there would be consistent with the treatment of other towns.

Kmeans clustering was started with 4 clusters. This produced a large central cluster and a few single zip codes as clusters due to geographical isolation. The number of clusters was increased to 7. An inner larger cluster seemed to stabilize, surrounded by ‘clusters’ of individual zip codes. The map of the 7 kmeans clusters is shown on Figure 3.

# Discussion

The final map in the Notebook shows the 7 clusters. Cluster 1 is central to the region. The types of services and the geographical distances to services and entertainment are very similar. This cluster begins in Saugus and runs through much of Lynn, Salem and Marblehead. The other ‘clusters’ are single zip codes. They are characterized, for clusters 0, 2 and 4, by isolation by coastal geography and coastal venues, such as marinas and seafood restaurants.

Clusters 3, 5 and 6 are all inland and geographically distinct. Even though venues are similar, the geographical separation renders them separate clusters. They are similar to the coastal clusters but, have sets of venues different from the coastal clusters.

# Conclusion

When considering venues for residential service convenience and entertainment, the potential resident can choose from 3 general groups of geographies, even though 7 clusters were identified. They are the coastline communities, the central exurban communities and the more distant inland suburban communities.

The central cluster is homogenous for venues and distance to venues. Much more information is necessary to create further distinctions for this region. Discriminators are parameters such as property taxes, rental rates, school performance parameters and population density.

# Recommendations

This analysis is too shallow to be of value to the new home buyer or renter. More parameters should be considered. A next step would be the proposal of a list of possible parameters, then a search for parameter availability and reliability.

The ages of the buyer/renter’s household may be pertinent to the analysis. Since the Northshore has a strong Council On Aging program in all communities, the number and ages of dependents and the buyer/renter’s ages and future plans may be pertinent to the analysis. The analyst should review how this issue might be considered and evaluate approaches and data availability.