Capstone Project: Comparing Neighborhoods in New York and Boston

Introduction:

Imagine you are a realtor who is working with a new company, ABC123, LLC., that just moved from New York to Boston. The company wants to relocate its workforce from New York to Boston. You and your company have been contracted by ABC123, LLC. to find homes/apartments in neighborhoods that make each employee feel at home. How can this be accomplished? This is a relevant issue for many realtors because the post-buy satisfaction of the homeowners is a key metric in measuring the success of a realtor. Moreover, people are becoming more focused on the available amenities in the area around them rather than the house itself when buying housing. People would be more willing to buy a house that isn't perfect for them if there are all the necessary and desired amenities nearby.

Data:

The following data will be used:

- Zip codes (neighborhoods) of Boston and New York City and corresponding latitude and longitude data.
- Foursquare data of venues based on the latitudes and longitudes of those neighborhoods.

The above will allow us to gather data on types of neighborhoods in New York and Boston and segment and cluster them into groups based on their similarities. This will give us the ability to make recommendations on comparable neighborhoods in Boston.

Methodology:

The main idea of this project is to cluster zip codes based on the types of venues most prevalent in those zip codes. I first started by gathering a dataset of zip codes, cities, counties, and latitude and longitudes of those zip codes. This dataset came from USPS. This was then separated into data tables for New York City and Boston. The New York and Boston dataframes were then concatenated into one dataframe with the following columns: zip, city, county, latitude, and longitude. Then I loaded up the Foursquare API with my credentials and pulled the venue data for the previously gathered zip codes. Then I grouped the venues by zip code and got the frequency of the most common venues in each zip code. Then I ran the kmeans clustering algorithm with 10 clusters. I chose 10 clusters because that seemed like a sufficient amount of clusters that would allow us to differentiate the zip codes effectively and build an effective system for the realtor to recommend neighborhoods.

Results:

The algorithm gave us clear clusters which could be used to effectively compare neighborhoods in New York and Boston. An interesting point to note is that there were significantly are more zip codes in New York compared to Boston. As a result, there were a few clusters with only 1 zip codes per city. Listed below is a breakdown of zip codes by cluster.

- Cluster 0: 138 NYC, 9 Boston
- Cluster 1: 0 NYC, 1 Boston
- Cluster 2: 81 NYC, 0 Boston
- Cluster 3: 23 NYC, 6 Boston
- Cluster 4: 33 NYC, 15 Boston
- Cluster 5: 0 NYC, 2 Boston
- Cluster 6: 0 NYC, 1 Boston
- Cluster 7: 1 NYC, 0 Boston
- Cluster 8: 52 NYC, 20 Boston
- Cluster 9: 1 NYC, 5 Boston

Discussion:

As we can see, there are clear Boston zip codes for every New York zip code. This means that we can recommend a bevy of neighborhoods in Boston to the people who are relocating from New York. For example if someone was coming from a New York zipcode in cluster 4, we can clearly recommend any one of the 15 Boston zip codes to those people.

This algorithm has potential for improvement. We could have added more clusters for increased differentiation because, as we can see, clusters 0 and 2 are highly overcrowded and it is very hard to be accurate with such a high number of zip codes in one cluster.

Conclusion:

Overall, the above algorithm can be effectively used by the realtor to recommend neighborhoods to customers.