Marina Kate Stephens
Coursera Capstone Project: Battle of the Neighborhoods

Medical Help in Mumbai

Introduction:

Around five million people die every year — almost a third of them in India— due to inadequate healthcare (Times of India). Your organization has raised significant funds to open a free health clinic in Mumbai, the 7th largest city in the world, to help reduce deaths that could be avoided with proper healthcare. Mumbai is a huge city, but you currently only have funding to open one free clinic. Which area of Mumbai most needs this assistance?

Data:

To answer this question, I will use the data from two different sources:

1. The first is geolocation data website Foursquare, which will allow me to see where all the clinics and hospitals are located in Mumbai.
2. The second is 99 Acres, an Indian real estate database website. I will look at property values across the neighborhoods of Mumbai, and then I will assume that average property value and average income are directly related. Found here: https://www.99acres.com/property-rates-and-price-trends-in-mumbai

I will then cluster neighborhoods by average income, number of healthcare venues, and location. I will suggest neighborhoods with lowest income and least healthcare options to be the location for the free health clinic.

Methodology:

Entire report can be found here:
https://github.com/mkatestephens/Coursera_Capstone/blob/master/Capstone%20Project%20Report%20FINAL.ipynb

1. Import libraries and download dependencies.
    a. Import numpy library to handle data in a vectorized manner.
    b. Import pandas library for data analysis in dataframe format.
    c. Import json library in order to work with json files.
    d. Install geopy client for geocoding services.
    e. Import requests library to handle the API requests from Foursquare.
    f. Import matplotlib module for creating data visualizations.
    g. Import KMeans from sklearn in order to use the machine learning known as k-means clustering to split data into like groups.
    h. Install folium to render maps.

2. Import the property value dataset from 99acres and cleanup the dataframe.
    a. I had saved the property data from 99acres as a csv file on github. Use the pandas read function to retrieve the data from the file.
    b. Drop unnecessary columns.
    c. Assign PIN codes (Indian postal codes) to each neighborhood.
    d. Assign latitude and longitude to each neighborhood using the PIN codes and requesting information from OpenStreetMap.
    e. Create dataframe with only needed columns: 'locality name', 'Price Range', 'Latitude', 'Longitude'.
    f. Remove rows with missing values.
    g. Convert price range column to an integer that is an average of the two numbers in the property value range.

3. Create map of Mumbai to visualize neighborhood location data.
    a. Use mean latitude and longitude from the dataframe to find coordinates for Mumbai and use latitude and longitude from the dataframe to create marker for each neighborhood on the map.
    b. Remove outliers and recreate map.

4. Connect to Foursquare and use their data to find information on location of health care facilities in Mumbai
    a. Define URL in order to make data calls from Foursquare.
    b. Use URL to call all medical venues in Mumbai.

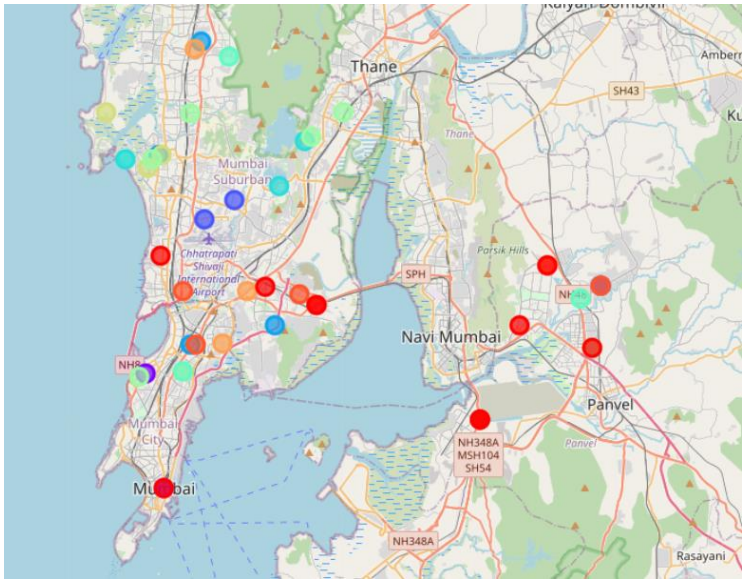
c. Convert json file into a dataframe.
d. Create map of all medical venues in Mumbai using latitude and longitude retrieved from Foursquare.

5. Find all medical venues within 500m of each neighborhood.
    a. Change dataframe to include only the 50 neighborhoods with lowest property value (since we are only concerned about low income neighborhoods).
    b. Use foursquare API to find number of venues for each neighborhood and add that number to the dataframe.

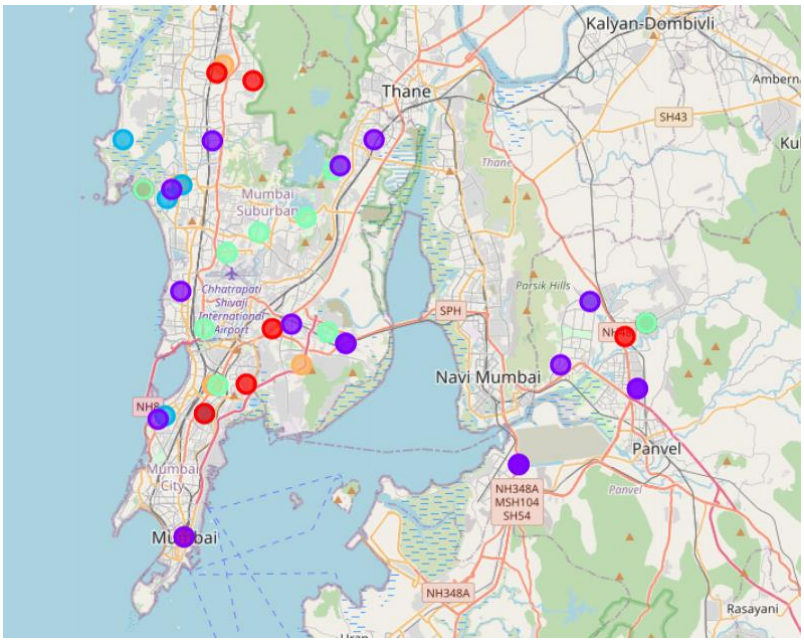
6. Cluster the neighborhood data based on location, property value, and number of nearby medical venues.
    a. First run k-means clustering with 10 groups and map the results.
        i. Add cluster labels for each neighborhood to dataframe
        ii. Map the neighborhoods with color of marker based on cluster group.
    b. The results were unclear, so I repeated the process with only 5 groups.

Results:

Map of first K-means clustering (10 groups):



Map of second K-means clustering (5 groups):



In the second map, you can see that cluster 3 (light green color) is the cluster with the most geographical consistency and (according to the dataframe) a low average number of nearby medical venues. Therefore, I would suggest building the new medical clinic somewhere in the middle of these neighborhoods to best serve the community.

Discussion & Conclusion

Finding the best location to build something in a huge cosmopolitan city like Mumbai is a difficult task. Mumbai is over 230 square miles, has over 500 unique neighborhoods, and has over 24 million citizens. Mumbai has some of the most glamorous neighborhoods and poorest slums in the world. The unique mixture of traditional and modern, rich and poor, local and foreign, makes Mumbai a very unique city with its own unique issues. There is not one single location where those in need are living.

Using the machine learning k-means clustering was only somewhat helpful in determining the best location for our free medical clinic. Visualizing the data on the maps was very useful in understanding where there is a lack of medical care and where the lower income neighborhoods are. The clustering algorithm only returned one group of neighborhoods that were located close together and how low average income. However, the connection was weak. If one is absolutely determined to use machine learning to decide the location for something like a free health clinic,

they might need to look at more factors than just income and location. Or, possibly a more complex machine learning algorithm is needed to find the best spot.