Coursera Capstone

IBM Data Science Capstone Project

Rohan Datta

June 2020

Introduction:

With the pandemic all over the world it has been proven that doctors and medical infrastructure are critical for our survival and fighting against this global issue. Thus it is important to ensure there is medical reach to all parts of the world. I live in India and the government has made huge efforts to ensure medical attention reaches all kinds of people from all walks of life. I would like to contribute to this effort and make this available to all people equally. Thus, I believe finding the location for a hospital to maximize reach is a great problem statement for a neighborhood analysis. Thus I chose a big city of India with many places that are semi urban or rural, Mumbai. It is the commercial capital of the country yet expands a lot and is often shorthanded as the population in that city is a lot, around 1.84 crore people.

Business Problem:

The business problem is to analyze and select the best locations in the city of Mumbai, India to open a new hospital. Using data analytics and the data science methodology, this project aims to provide solutions to the specific question: Where should a hospital be opened in the city of Mumbai, India?

Target Audience:

This project will be useful for healthcare workers and the general people alike as it aims to bring more healthcare services to a larger number of people. It will help people stay safe and secure from a global pandemic and rest assured that they have the necessary services close by if an emergency occurs and there is a need.

Data:

We will need the following data to solve this problem:

- List of neighbourhoods in Mumbai to define the scope and granularity of this project.

- Latitude and Longitudinal values of all this neighbourhoods to plot the maps and get further data.
- Venue data from FourSquare related to hospitals to perform clustering and decipher the best area for hospital location.

Sources of the data and where and how to extract them are as follows:

- This Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai#Mumbai_neighbourhood_coordintes) contains a list of neighbourhoods in Mumbai, with a total of 87 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.
- After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the hospital category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.