2. Data

2.1 Data Requirement and Collection

We needed the data about population of Mumbai, pollution data, and neighborhoods data. We collected the data from different website. This is common data. We could collect from Wikipedia and normally search in google. We would also use foursquare location data to segment the neighborhoods. Therefore we used following data.

2.1.1 Population Growth of Mumbai

Here, a questions arise in mind that why is population growth data is important and what is the connection between a hospital situated and population growth?

It's very important to know about population growth to start a new business in market. Population growth is big factor to impact on any types of business and opening a hospital is one of them. Suppose a city has low population and there are a number of hospital are opened there, what will happen, there is less chance to run a new business there also it will have to face most challenging environment. Now any one can think deeply that if population increases, more traffic occurs in that area, resultant pollution rate is also increased that effect on health of people. So it was also important to know about pollution rate of the neighborhoods of Mumbai city.

2.1.2 Pollution Rate Dataset

We collected the dataset from different sources and combine them together in a single dataset. We extracted the pollution data of each neighborhoods having the pollutants PM2.5, PM10, NO2, CO, NH3, and OZONE.

Again the same question why this data is important?

Pollution rate of the city help to know about health status of people in neighborhoods of the city. Lots of people are suffering from many types of diseases like heart diseases, cancer, breathing problem due to pollution and many more diseases and everyone wants to go to near hospital for treatment. So, we

have to find such as location where no any hospital or less hospital in neighborhood. So our next step is to collect the neighborhoods dataset.

2.1.3 Neighborhoods data

We extracted the data from Wikipedia and scrap this data by pandas library.

2.1.4 Foursquare Location Data

Last data we will use foursquare location data. It will help to extract the venues within the range. In this case, we will use foursquare credential information like client id, client secret, and version. Also we take a limit of venues in a particular range.

3. Data Cleaning and Feature Extraction

1. Our first data of Population data is in csv file. It has 86 rows and 4 columns. Each row explains year, population, growth rate and growth of the population in that year. We will use all columns of the dataset and see the variation of each year.
2. Second data of pollution data is also in csv file. It has 114 rows and 5 columns. The first column represents location in which neighborhood exists.
   Second column is pollutants that are PM 2.5, PM10, NO2, NH3, SO2, OZONE and CO. Each pollutant has average values in microgram per meter cube. Last two columns represent minimum and maximum value of pollutants.
3. Third dataset is in url form. And we will extract from Wikipedia with the help of pandas library. It has 93 rows and 4 columns.
   First column represents Neighborhood of Mumbai city. Second column is Location detail where neighborhood exists and last two columns are latitude and longitude. These are coordinates where neighborhood is located.
   First we will use geocoder python library and find the latitude and longitude of Mumbai city location.
4. Last dataset is used from foursquare location data where we will segment and cluster the neighborhood of Mumbai city. We will use our foursquare credential to create a url and extract the location data from it. We will use k

means clustering to find the best k value so that we may find the best clusters in our map.

4. Methodology

4.1 Analytic Approach

4.1.1 Population Data

First we will explore the population data by using Plotly Library. Plotly is a python library that makes the graph more attractive. We can zoom in and zoom out in graph that is created by Plotly. We can see each value of the year on each points on the graph. We will see the growth rate of population that how increase the data over each year.

4.1.2 Pollution Data

Second step we will import pollution data. We will also use Plotly Library to plot the bar graph and decide to choose best location that is free from pollution or good or satisfactory in condition. In this dataset we will plot each location where neighborhood exists and see the AQI (stands for air quality index) decide which location is good for opening hospital.

4.1.3 Neighborhoods Data

Last we will focus on neighborhood clustering. For analytical approach we will use clustering technique of k means machine learning algorithm and segment of neighborhoods. K means clustering technique is the best way for clustering the data points of similar characteristics. We plotted a graph of cluster to choose the best value of k and decided how to apply it to segment the neighborhood of Mumbai city. We decided the best cluster by the help of distance between the points within the cluster that is inertia. Inertia means that how far the points within the cluster.