# Spread of COVID-19 in Mumbai

Balasubramanian Viswanathan

September 20, 2020

# Spread of COVID-19 in Mumbai

## 1 Introduction

...

## 2 Literature Review

...

## 3 Methodology Section

### 3.1 Source of Data

The following are the data needed for analyzing the distribution of Positive Cases of COVID-19 in Mumbai.

#### 3.1.1 Population

We require the population of Mumbai and ideally split at a Ward level. As ‘Ward’ is an administrative unit within the city, observations, insights and actions can be based or targeted at more manageable sized areas and the population of the city. Here is a list of websites that have information of interest in this regard :

1. Information on Mumbai Ubran population is available at: https://www.censusindia.gov.in/pca/SearchDetails.aspx?Id=623672
2. Information on Mumbai Suburban population is availlable at: https://www.censusindia.gov.in/pca/SearchDetails.aspx?Id=623607
3. Mumbai City map inclusive of Wards is available at: https://portal.mcgm.gov.in/irj/portal/anonymous/qlmumbaimap
4. Mumbai Wards & Districts: Population & Density by Sector 2001 is available at: http://www.demographia.com/db-mumbaidistr91.htm

While the information on www.demographia.com appears to be dated, the #s on this page tally reasonably well with the data from the Census India site with an advantage of providing a bunch of data including Households, Population, Land Area at a Ward level all consolidated in one table. Data from www.demographia.com are used for all further work as part of this exercise.

#### 3.1.2 Location Data i.e. Venues of Interest

The Foursquare API has been used to identify venues of interest from a given geographical location i.e. latitude and longitude by querying https://api.foursquare.com

#### 3.1.3 Latitude and Longitude

We convert Ward and Area level information to latitude and longitude by using the Nominatim library that leverages OpenStreetMap for geographical information.

### 3.1.4 COVID-19 Data

Municipal Corporation of Greater Mumbai, publishes a COVID-19 dashboard and the same is available at:
http://stopcoronavirus.mcgm.gov.in/assets/docs/Dashboard.pdf

This is a graphically rich multi page document and does not lend itself to be scraped with any of the available PDF reader libraries. Hence, Ward level COVID-19 data is manually captured from this Dashboard onto a spreadsheet which in turn is being read into the program for further processing.

## 3.2 Data Analysis

As mentioned above, data taken from www.demographia.com include all the necessary information about the Wards in Mumbai. Some of the initial rows are dropped and the names of the columns are extracted from table and used to rename the columns of the DataFrame used for processing. '**Density per Square Mile**' column is redundant as the same information is available as '**Density per Square Kilometer**'. The top and bottom rows of the DataFrame are shown below :

| | Ward | Area | Land Area (SKM) | Households | Population | Density per Square Kilometer |
|---|---|---|---|---|---|---|
| **0** | A | Colaba | 12.5 | 43661 | 210847 | 16868 |
| **1** | B | Sanhurst Road | 2.5 | 27225 | 140633 | 56936 |

| | Ward | Area | Land Area (SKM) | Households | Population | Density per Square Kilometer |
|---|---|---|---|---|---|---|
| **22** | S | Bhandup | 64.0 | 148731 | 691227 | 10800 |
| **23** | T | Mulund | 45.4 | 73540 | 330195 | 7270 |

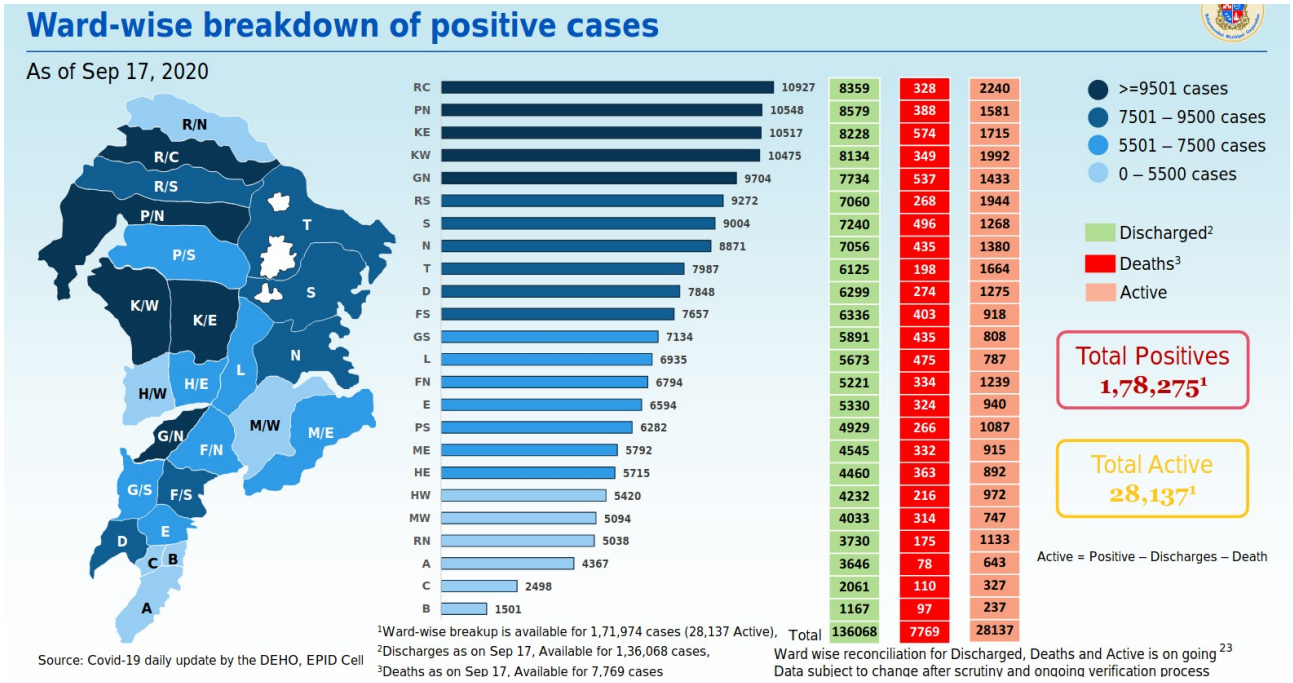
The '**Population'** column strongly correlates with the '**Households**' column as they are directly related. The '**Density per Square Kilometer**'column is derived by using the '**Population**' and '**Land Area (SKM)**' columns. For all the processing needed to arrive at a Cluster of Wards, the '**Households'** and the '**Density per Square Kilometer**' columns are not utilized.

The '**Ward**' column is critical to this exercise and is used as the primary key for a variety of operations. We use the '**Area**' column as an alternate field to do a Ward to Geographical coordinate lookup as some of the Ward names do not have entries on the OpenStreeMap and hence is used as as a proxy for the '**Ward**' selectively. Wards without a valid Geographical coordinates are dropped. The '**Land Area (SKM)**', '**Population**', '**Latitude**', '**Longitude**' are used for creating clusters while running the data by the Kmeans algorithm. The '**Latitude**' and '**Longitude**' are also utilized while generating the Folium based maps to visualize Mumbai and it's Wards.

The DataFrame (top 5 rows) with the Geograhical coordinates is shown below :

| | Ward | Area | Land Area (SKM) | Population | Latitude | Longitude |
|---|---|---|---|---|---|---|
| **0** | A | Colaba | 12.5 | 210847 | 18.915091 | 72.825969 |
| **1** | B | Sanhurst Road | 2.5 | 140633 | 18.956310 | 72.839838 |
| **2** | C | Marine Lines | 1.8 | 202922 | 18.945670 | 72.823781 |
| **3** | D | Grant Road | 6.6 | 382841 | 18.964447 | 72.813573 |
| **4** | E | Byculla | 7.4 | 440335 | 18.976622 | 72.832794 |

As highlighted earlier, data on the Total Positive cases in Mumbai are manually entered in a spreadsheet from a Dashboard published by the Municipal Corporation of Greater Mumbai.




The names of the Wards are updated to ensure their full form as opposed to abbreviated forms are utilized. For this exercise as the focus is on the number of Positive cases, data related to the deaths are ignored by dropping the '**Total Deaths**' column from the Data Frame. The DataFrame with information on the Total Positive cases on a per Ward basis is shown below:

| | Ward | Total Positive |
|---|---|---|
| 0 | R Central | 10927 |
| 1 | P North | 10548 |
| 2 | K East | 10517 |
| 3 | K West | 10475 |
| 4 | G North | 9704 |

Data from both the above DataFrames have been merged all the categorical information including Ward and Area names have been removed in the DataFrame shown below :

| | Land Area (SKM) | Population | Latitude | Longitude | Total Positive |
|---|---|---|---|---|---|
| 0 | 12.5 | 210847 | 18.915091 | 72.825969 | 4367 |
| 1 | 2.5 | 140633 | 18.956310 | 72.839838 | 1501 |
| 2 | 1.8 | 202922 | 18.945670 | 72.823781 | 2498 |

This is necessary as only numerical fields are needed to utilize some of the machine learning algorithms to process the data.

## 3.3 Approach

The overall approach is similar to one that was followed for the exercise on Segmenting and Clustering Neighborhoods in New York City completed earlier. Data related to Mumbai like the Wards of the city, the Area (central location within the Ward), Land Area in Square Kilo Meter, Population are gathered to start with. Latitude and Longitude of the Ward are looked up for the Wards by using the Nominatim library. These are utilized to visualize the Wards by using a Folium map to gain an understanding of where the Wards are located within the city.

Foursquare APIs are utilized to gather information on the venues of interest nearby all the Wards by limiting the radius of the search as a function of the Land Area of each of the Wards. We ensure no venue is counted more than once by deleting duplicate (read redundant) venues that may be returned by the Foursquare API. All the category of venues gathered from the Foursquare APIs are group with the Wards of Mumbai and split into clusters by using the Kmeans algorithm based on top 10 common venues in each of the Wards. These Clusters are visualized by using a Folium map to understand how the wards are clustered based on Top 10 common venues in each of the Wards.

Total Positive cases of COVID-19 reported in Mumbai is merged with Ward and related data along with the Top common venues in each of the Wards. Mumbai Ward details along with Total Positive case data are normalized using the Standard score method to ensure differences in the range of each of these features do not impact the model while training with these data.

| | Afghan Restaurant | Airport | Airport Lounge | Airport Service | American Restaurant | Aquarium | Arcade | Art Gallery | Asian Restaurant | Athletics & Sports | ... | Vegetarian / Vegan Restaurant | Water Park | Wine Bar | Wine Shop | Women's Store | ( |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.0 | 0.0 | 0.0 | 0.0 | 0.000 | 0.0 | 0.000 | 0.015873 | 0.015873 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | -0.51 |
| **1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.025 | 0.0 | 0.025 | 0.000000 | 0.000000 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | -1.19 |

2 rows × 175 columns

| merican staurant | Aquarium | Arcade | Art Gallery | Asian Restaurant | Athletics & Sports | ... | Vegetarian / Vegan Restaurant | Water Park | Wine Bar | Wine Shop | Women's Store | Land Area (SKM) | Population | Latitude | Longitude | Total Positive |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.0 | 0.0 | 0.017544 | 0.0 | ... | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 2.980588 | 1.039031 | 0.745209 | 2.240199 | 0.74746 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.052632 | 0.0 | ... | 0.052632 | 0.0 | 0.0 | 0.0 | 0.0 | 1.717263 | -0.912616 | 1.053795 | 2.751295 | 0.33397 |

The resultant dataset is run by the Kmeans algorithm to arrive at clusters of Wards which are similar. The number of clusters are determined by using the elbow method. The resultant output is visualized with a Folium map.