Spread of COVID-19 in Mumbai

Balasubramanian Viswanathan

September 20, 2020

Table of Contents

Spread of COVID-19 in Mumbai

1 Introduction

1.1 Context

Mumbai, formerly known as Bombay, is the captial of the state of Maharashtra in South Western India. It is known as the financial capital of India. Mumbai is one of the densely populated cities in the World with a population of approximately 12 million as per the 2011 census. Mumbai's population is approxiately 20 million as per the estimate of United Nations in 2018. Mumbai has a tropical climate with the monsoon season spanning between June and September. It is a vibrant city with several places of interest ranging from Beaches to places of worship to famous monuments. It has thousands of restaurants serving a variety of food across the city. Mumbai has an excellent pubilc transportation system with a successful Bus and Suburban railway network covering the width and breadth of the city.

Maharashtra has the highest number of COVID-19 Positive cases among all the states in India as of September 2020 and Mumbai is one of the top cities in India in terms of number of Positive cases.

1.2 Problem

Mumbai is a densely populated, well connected city from within and outside of the state and the country with an agile and active life style and is one of the cities that has seen a consistent growth in COVID-19 cases over the last many months. The objective behind this exercise is to understand how the COVID-19 Positive cases have spread across Mumbai and the relationship the same has with various elements like venues of interest to people, geographical and population related parameters with the goal of arriving at few observations and high level insights that may be of help in fine tuning the actions being taken by the Government in combating this disease.

While there may not be a direct business outcome, this exercise would be of interest to the Government including the Municipal Corporation of Greater Mumbai and will relate to health and socio-economic welfare of the residents.

2 Literature Review

Studied some material available on COVID-19 and it's impact on Mumbai and referred below are few articles and papers in this regard :

1. The COVID-19, Migration and Livelihood in India A Background Paper for Policy Makers International Institute for Population Sciences, Mumbai The COVID-19, Migration and Livelihood in India by R. B. Bhagat, Harihar Sahoo, Sahoo Archana and K Roy

   studies the impact of COVID-19 on migrant workers in large Indian cities including Mumbai.

2. Why Mumbai Was & Is India's Worst Covid Hotspot by Kunal Purohit

   covers how Mumbai has been impacted by COVID-19 and cites some of the challenges faced by few individuals.

3. COVID-19 Awareness Among Healthcare Students and Professionals in Mumbai Metropolitan Region: A Questionnaire-Based Survey by by Pranav D. Modi , Girija Nair, Abhay Uppe, Janhavi Modi, Balaji Tuppekar, Amit S. Gharpure, Deepak Langade

   This study assesses the awareness of COVID-19 disease and related infection control practices among healthcare professionals and students in the Mumbai Metropolitan Region.

4. India coronavirus: 'More than half of Mumbai slum-dwellers had Covid-19'

   A survey found that more than half the residents of slums in three areas in Mumbai tested positive for antibodies to the coronavirus.

   The material referred to above covers COVID-19 in detail and the impact on people and practices being followed. This submission is a complementary work and focuses more on understanding how the COVID-19 Positive cases have spread across Mumbai and the relationship the same has with various elements like venues of interest to people, geographical and population related parameters. We conclude by arriving at few observations and high level insights on clusters created by pooling wards that exhibit similar characteristics.

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

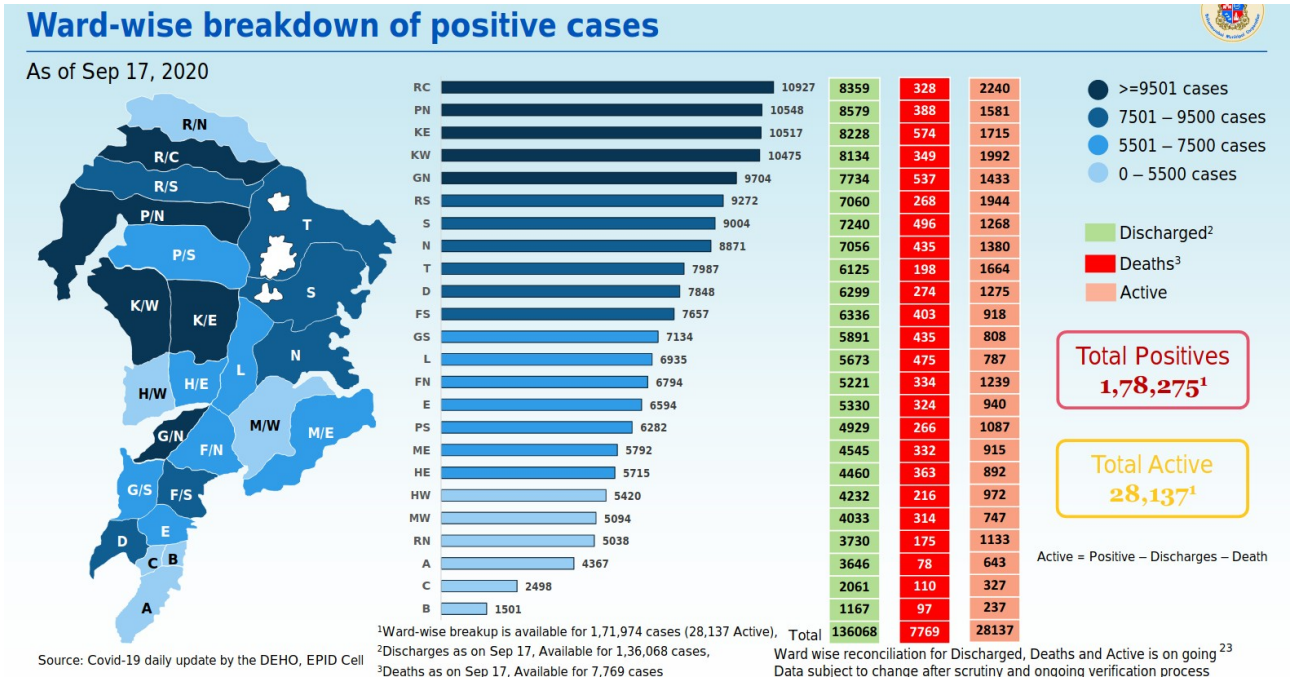
The '**Population'** column strongly correlates with the '**Households**' column as they are directly related. The '**Density per Square Kilometer**'column is derived by using the '**Population**' and '**Land Area (SKM)**' columns. For all the processing needed to arrive at a Cluster of Wards, the '**Households'** 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.

4 Results Section

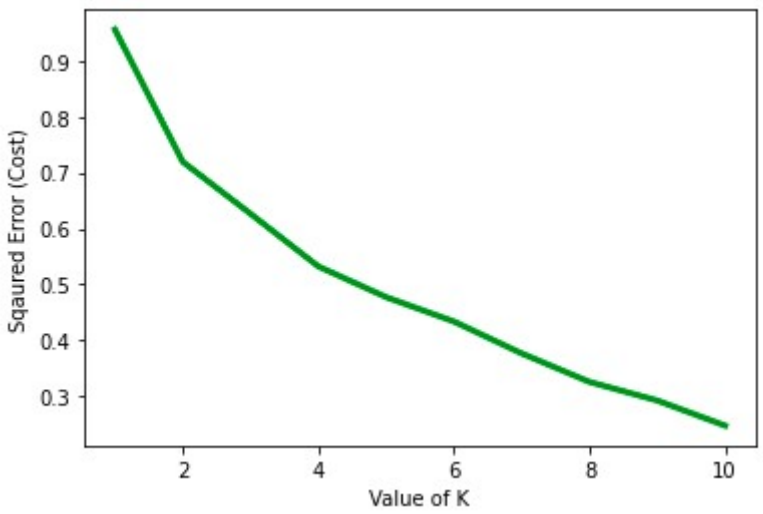
4.1 Location of Wards in Mumbai

The information on the Wards in Mumbai along with the Geographical coordinates are utilized to arrive at a Folium map as shown below :

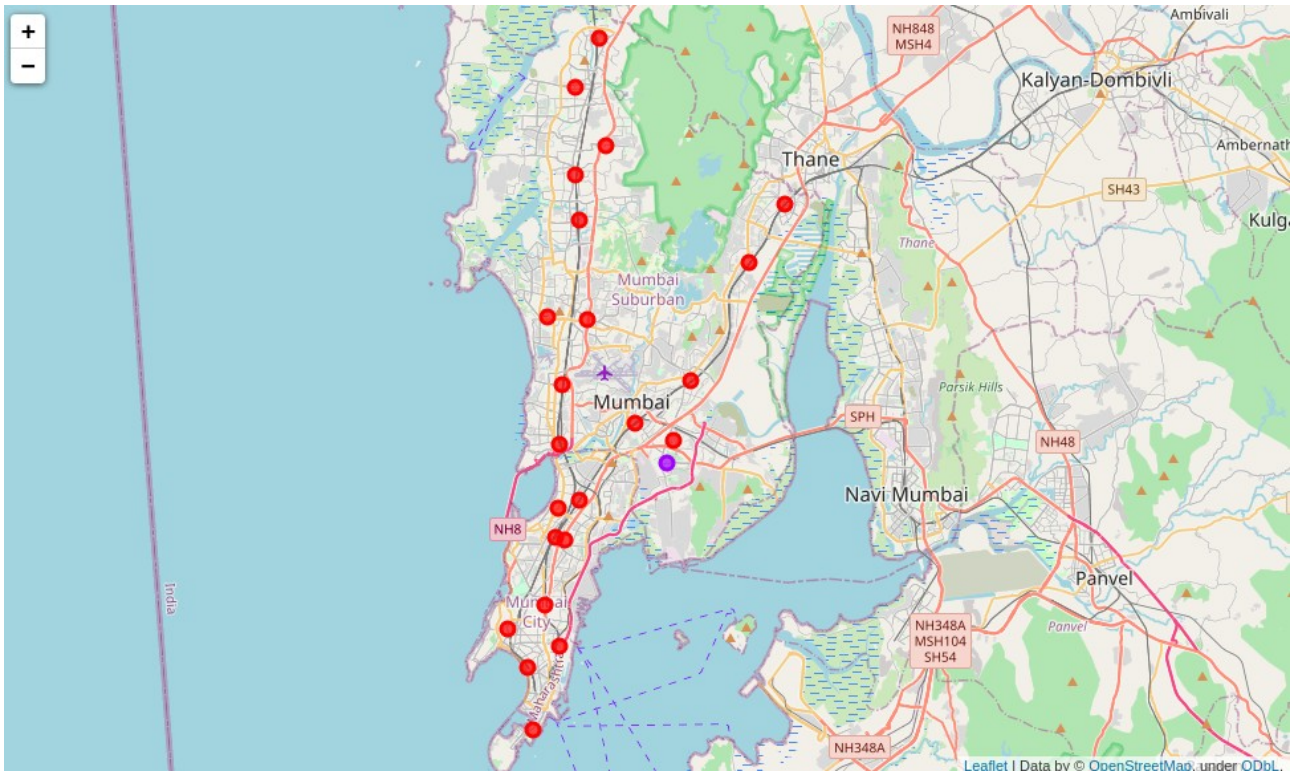
The Wards are distributed primarily in the North to South line with an East to West fork. Mumbai has been organized in this manner by the administrators of the city.

4.2 Clusters of Wards based on Top 10 Venues in Mumbai

As per the Elbow method, the category of venues for all the Mumbai Wards gathered by using the Foursquare APIs are ideally split into two clusters as seen by using the Kmeans algorithm :



Clusters of Mumbai Wards based on similarities of Top 10 venues in each of the Wards are shown below :



From the above graph, 23 of the 24 Wards appear to be similar and hence have been pooled together.

The DataFrame below shows that most of the common venues serve Food and/or beverages in Cluster 0.

			(SKM)					venue	venue	venue	venue	venue	venue	venue	ver
0	A	Colaba	12.5	210847	18.915091	72.825969	0	Indian Restaurant	Café	Hotel	Bakery	Coffee Shop	Scenic Lookout	Cricket Ground	Ita Restau
1	B	Sanhurst Road	2.5	140633	18.956310	72.839838	0	Indian Restaurant	Dessert Shop	Ice Cream Shop	Café	Market	Restaurant	Chinese Restaurant	BBQ J
2	C	Marine Lines	1.8	202922	18.945670	72.823781	0	Indian Restaurant	Café	Fast Food Restaurant	Train Station	Coffee Shop	Juice Bar	Gastropub	Pi Pl
3	D	Grant Road	6.6	382841	18.964447	72.813573	0	Indian Restaurant	Restaurant	Bakery	Bar	Coffee Shop	Snack Place	Chinese Restaurant	P
4	E	Byculla	7.4	440335	18.976622	72.832794	0	Indian Restaurant	Café	Lounge	Chinese Restaurant	Restaurant	Coffee Shop	Hotel	

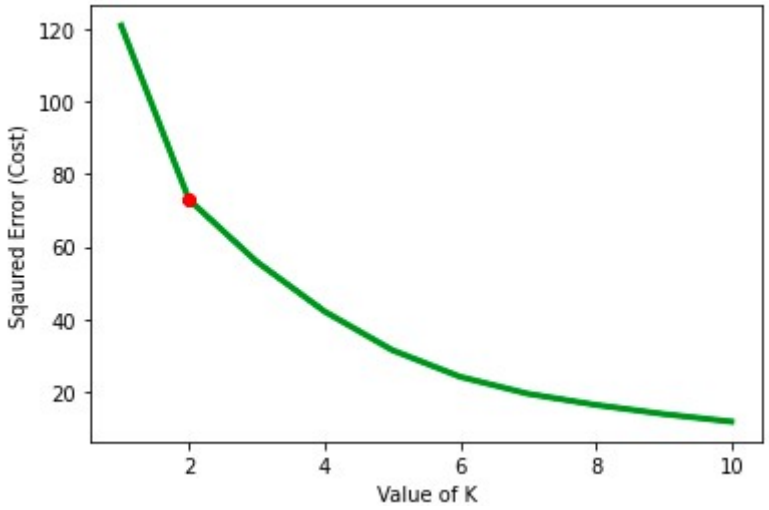
The DataFrame below shows a mixed type of common venues including food, market and others in Cluster 1.

Land Area (SKM)	Population	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
32.5	674850	19.046066	72.895473	1	Chinese Restaurant	Vegetarian / Vegan Restaurant	Asian Restaurant	Food	Women's Store	Flea Market	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant

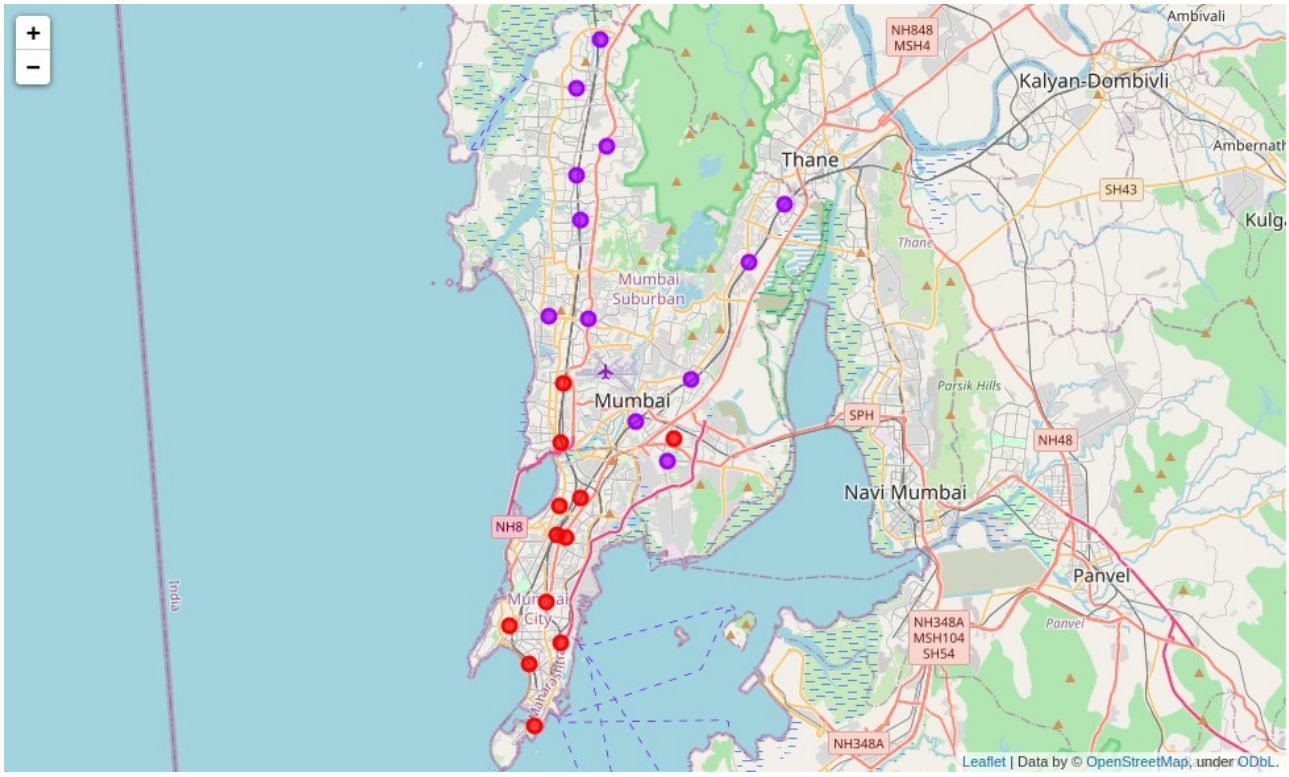
4.3 Clusters of Mumbai Wards based on multiple features

Information on Wards including Land Area, Population, COVID-19 Cases and Top 10 Venues in Mumbai are utilized to cluster Wards that are similar.

As per the Elbow method the Wards of Mumbai are ideally grouped into two clusters :



Here’s a map of Mumbai split into two clusters of Wards based on diverse features as mentioned above :



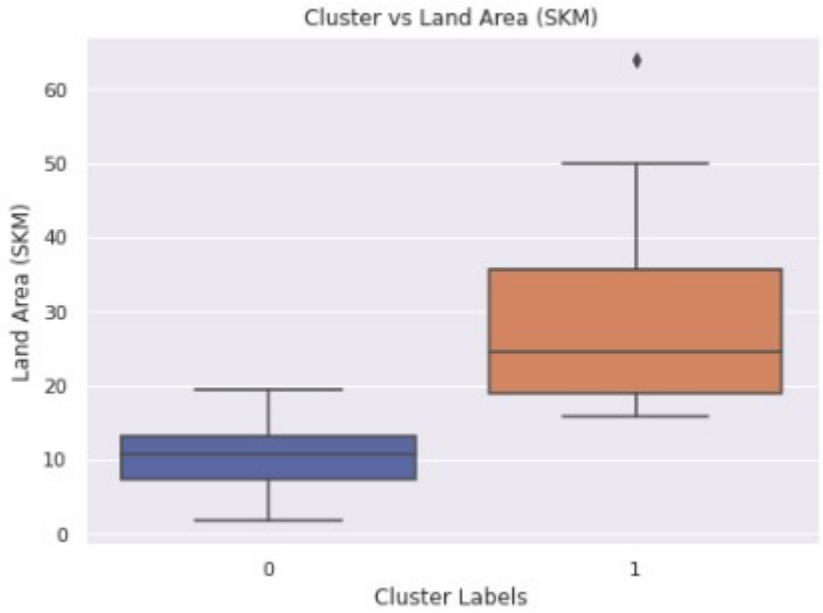
It’s evident from the map that the Wards have been split into two groups in the North and South directions.

4.4 Clusters and their relationship to some of the features of the model

In this subsection we compare Clusters based on different features and also based on select features within a Cluster to unearth the similarities within a Cluster and differences between Clusters.

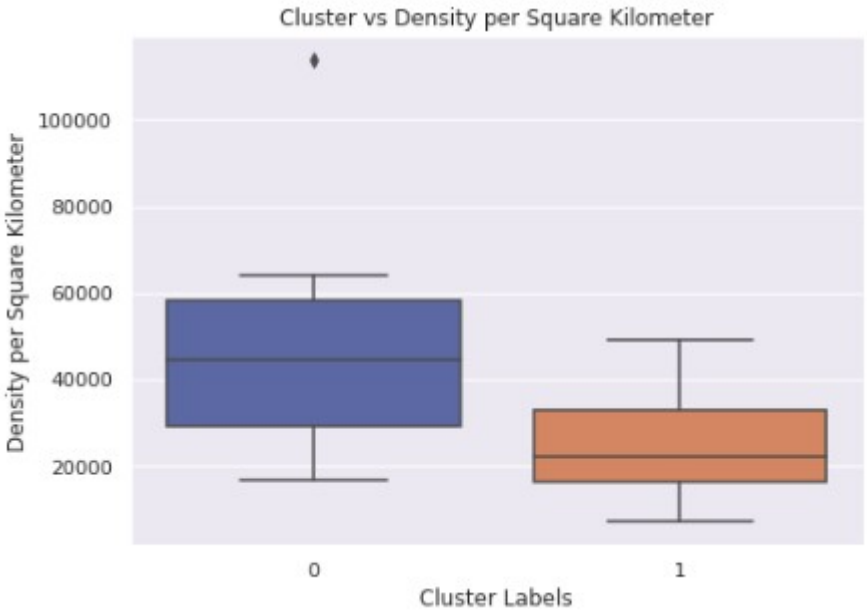
4.4.1 Clusters vs Land Area

The histogram below shows a clear differentiation between the Clusters based on the Land Area of the Wards that make up the same measured in Square Kilometers. Cluster 0 includes almost all the Wards that are smaller in size than Cluster 1. The largest Ward in Cluster 0 is marginally bigger than the smallest Ward in Cluster 1.



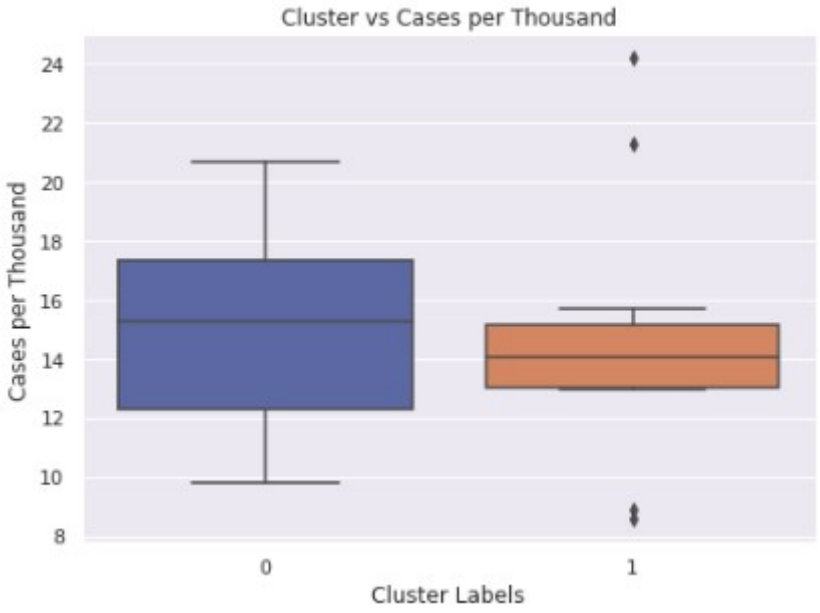
4.4.2 Clusters vs Density per Square Kilometer

The histogram below shows the differentiation between the Clusters based on the Density per Square Kilometer of the Wards. Cluster 0 includes more of the denser Wards of the City and contrast this with the earlier observation of Cluster 0 comprising of smaller Wards of the city.



4.4.3 Clusters vs Cases per Thousand

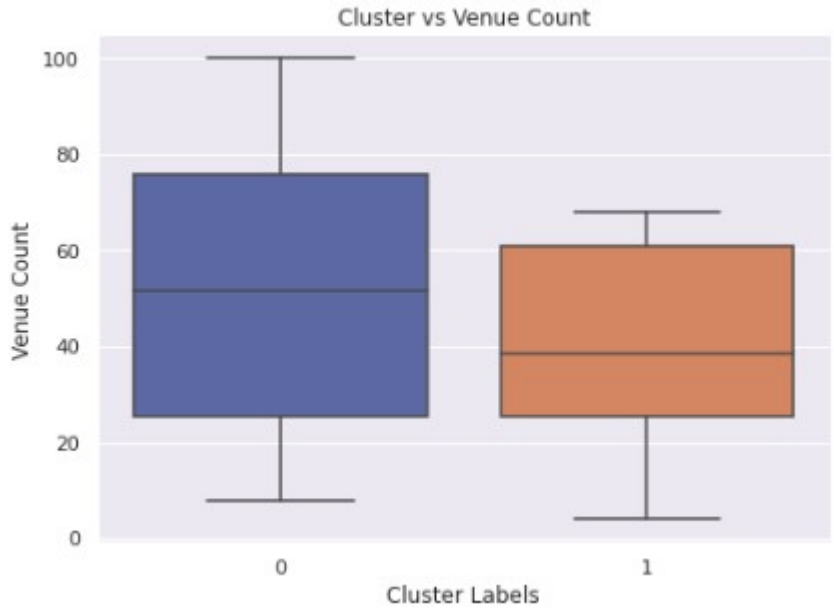
The histogram below shows a clear differentiation between the Clusters based on the number of COVID-19 Positive Cases per Thousand people of the Wards:



Cluster 0 shows a wide variation in the number of Cases per Thousand people across Wards when compared with Cluster 1 while excluding the outliers.

4.4.4 Clusters vs Venue Count

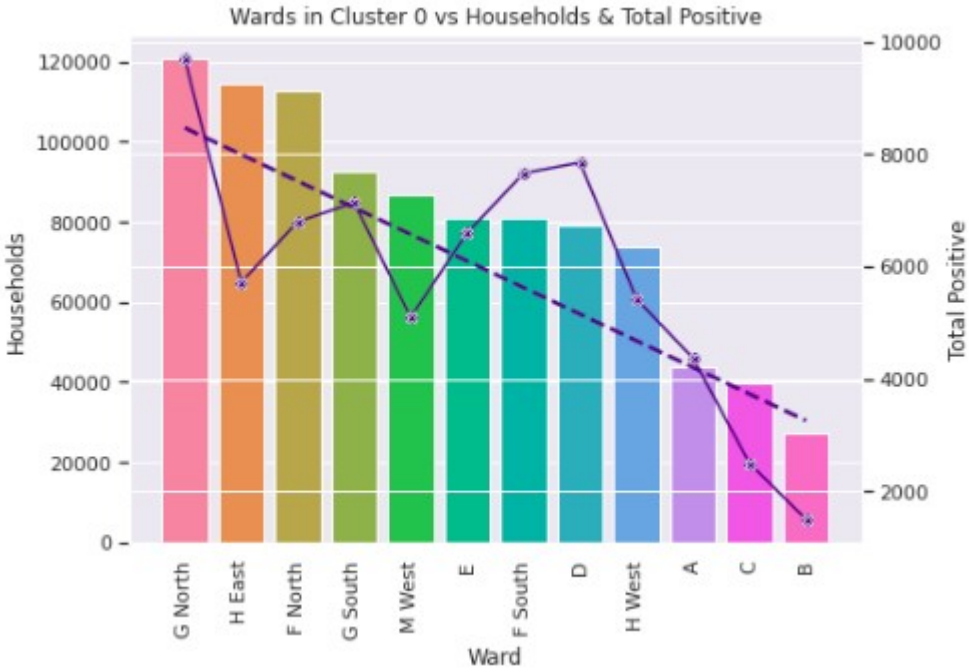
The histogram below shows the differentiation between the Clusters based on the number of common venues in each of the Wards:



Cluster 0 includes more Venues of interest than Cluster 1 even though Cluster 1 on an average includes a much larger area.

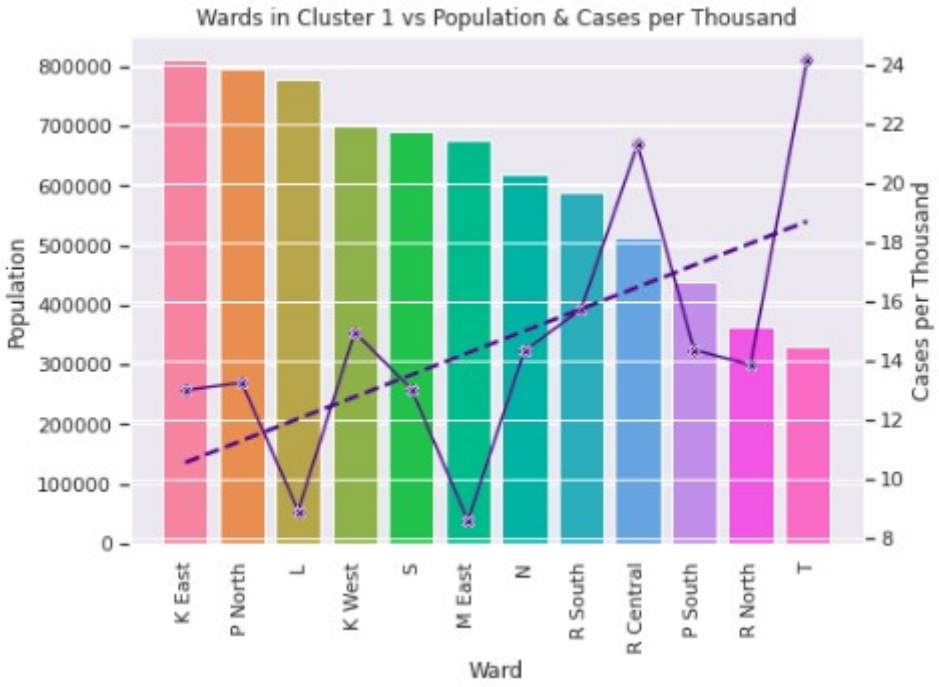
4.4.5 Households vs Total Positive Cases in Cluster 0

The graph below shows a direct relationship between the number of Households and the Total COVID-19 Positive Cases of the Wards in Cluster 0:



4.4.6 Population vs Cases per Thousand in Cluster 1

The graph below shows an inverse relationship between the population and the number of COVID-19 Positive Cases per Thousand people of the Wards in Cluster 1:



5 Discussion Section

Here are some key observations based on the above:

- Clusters are split based on the size of the land areas with smaller & the larger ones forming 0 and 1 respectively.
- Cluster 0 includes all the downtown locations with predominantly higher densities.
- The Cases per Thousand per Ward in Cluster 0 is in the range of ~10 to ~21 i.e. 2X ratio between min and max.
- The Cases per Thousand per Ward in Cluster 1 is in the range ~13 to ~16 i.e. a very limited band except for the outliers.
- There are more venues of interest in Cluster 0, as it includes all the downtown areas, than in Cluster 1.
- For Cluster 0, the Total Positive per Ward appears to relate more directly to size of the Households of the Wards.
- For Cluster 1, the Cases per Thousand per Ward appears to inversely relate to the size of the Population of the Wards.
- Information on the venues and their categories identified using the Foursquare API seem to provide sparse information.
- Venue Data, being sparse, appear to have had very little influence on the final outcome.
- The aggregate population size and the Total Positive Cases of Cluster 0 are less when compared with Cluster 1.
- Average number of Cases per Thousand per Ward is ~15 for Cluster 0 and ~14 for Cluster 1.

6 Conclusion Section

Insights

Here are some thoughts to conclude this exercise:

- The spread of the virus is slightly more in Cluster 0 than Cluster 1.
- In Cluster 0, focus for combating the disease may be more along the lines of the number of Households i.e. Wards with more # of Households are prioritized over those with less.
- In Cluster 1, focus for combating the disease may be based on the size of the Population in an increasing order i.e. priortize Wards with lesser population over those with more.
- A more broader strategy may need to be followed to contain the spread in Cluster 0 as the Cases per Thousand per ward is in a wider range.
- A homogeneous approach may work well in containing the spread in Cluster 1 with outliers handled as appropriate

Future Directions

Here are few areas to be investigated going forward:

- Recompute the Clusters by including Total Deaths per Ward as a new feature
- Look at alternate source of Location Data and arrive at a better understanding of the relationship between Venues and the spread of COVID-19 in Mumbai.

7 Reference and Acknowledgements

Housing Sales Prices & Venues Data Analysis of Istanbul by Sercan Yıldız

Predicting the Improvement of NBA players by Zhenfeng Liu

Capstone Project - The Battle of the Neighborhoods

The Literature section in this document captures more material of interest to gain better understanding of the topic with focus on Mumbai.

Would like to take this opportunity to thank Coursera and IBM for coming up with this well structured specialization on Data Science and for following a hands on approach right through the program that made this exercise an interesting and a nice learning experience as well.