# *Coursera capstone project: Applied data science*

Saurabh Chaudhari

[Saurabhwvu@gmail.com](mailto:Saurabhwvu@gmail.com)

## 1.Introduction

Mumbai is the capital city of Maharashtra, with a population of around 20 millions. It is second most populated city in India and 7th most populated city in the world. The city of Mumbai consists of two distinct regions: Mumbai City district and Mumbai Suburban district, which form two separate revenue districts of Maharashtra. The western part of the Mumbai Suburban district forms the Western Suburbs and the eastern portion forms the Eastern Suburbs. The suburbs of Chembur, Govandi, Mankhurd and Trombay lie to the south-east of the Eastern Suburbs. These suburbs are generally not considered as part of the Eastern Suburbs and are sometimes referred to as the "Harbour Suburbs".

2.Business Problem

Most of the population in the Mumbai is working population and it need to eat breakfast, lunch, dinner and snacks, while travelling through different modes of transportation such as bus, local trains and taxies. Foods that attract busy people on the go include egg sandwiches, fries, pizza, burgers, microwaveable or cold prepared meals. Beverages such as coffee, tea, wraps, bottled water, soda and juice also sell well.

Thus, the main objective of the project will be to find ideal spots in the mumbai where fast food retail chains can be put up, aiming at the above demographic, thereby helping the owners of the outlets to extract maximum profits out of them.

## 3.Data

The data for this project has been retrieved and processed using through multiple sources.

**3.1 Neighborhood**

The data of the neighborhoods in Mumbai can be extracted out by web scraping using BeautifulSoup library for python. The neighborhood data is scrapped from a Wikipedia webpage. This data is converted into a pandas Data frame. The good thing about it data is that it provides the latitude and longitude of the neighborhoods too and these values are further stored into initial dataframe.

**3.2 Venue data**

From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the FourSquare API, and creating another DataFrame to contain all the venue details along with the respective neighborhoods.

## 4. Methodology

First, I need to get the list of neighborhoods in Mumbai, India. This is possible by extracting the list of neighborhoods from Wikipedia:

**https://en.wikipedia.org/wiki/List\_of\_neighbourhoods\_in\_Mumbai**

I used the web scrapping by utilizing pandas HTML table scraping method as it is easier and more convenient to pull tabular data directly from a web page into the data frame as shown below

#import and use beautifulsoup

from bs4 import BeautifulSoup

soup = BeautifulSoup(data,'html.parser')

#Create lists

area = []

locationinarea = []

lattitude = []

longitude = []

#Append the data

for row in soup.find('table').find\_all('tr'):

cells = row.find\_all('td')

if(len(cells) > 0):

area.append(cells[0].text.rstrip('\n'))

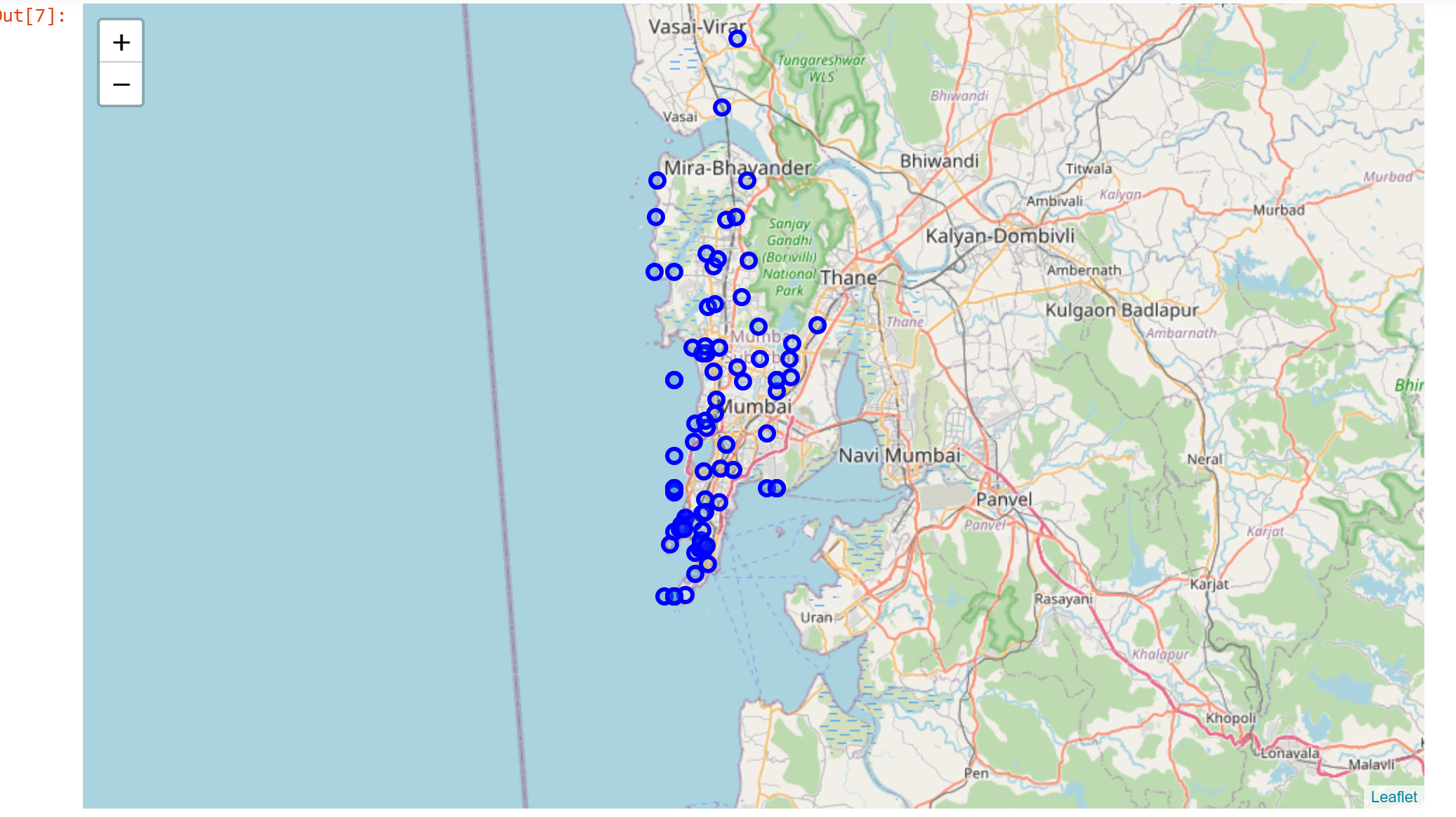
locationinarea.append(cells[1].text.rstrip('\n'))

lattitude.append(cells[2].text.rstrip('\n'))

longitude.append(cells[3].text.rstrip('\n'))

Good thing about this dataset is that it contains Longitude and Lattitude information of these location, which is later used for folium library and Visualization. They are in string format though which was converted into string for further processing.

After gathering these coordinates, I visualize the map of Mumbai using Folium package to verify whether these are correct coordinates as shown below:



Next, I use Foursquare API to pull the list of top 50 venues within 100 meters radius. I have created a Foursquare developer account in the previous week for week 5 assignment that I continued using for this part too.

From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. Then I find out the most common venue visited in each area. Many of these area already has different restaurants and food places as the most visited places. As we are helping to establish a location where the new fast food joints will be profitable, we will eliminate all those area from the consideration

k-means clustering was used from the remaining areas to get the cluster of areas to identify location for setting up new fast food place.K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighborhoods in Mumbai into 5 clusters. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the fast food restaurants.

# set number of clusters

kclusters = 5

Mum\_clustering = Mum\_onehot\_new.drop(["col","Area", "Location in Mumbai"], 1)

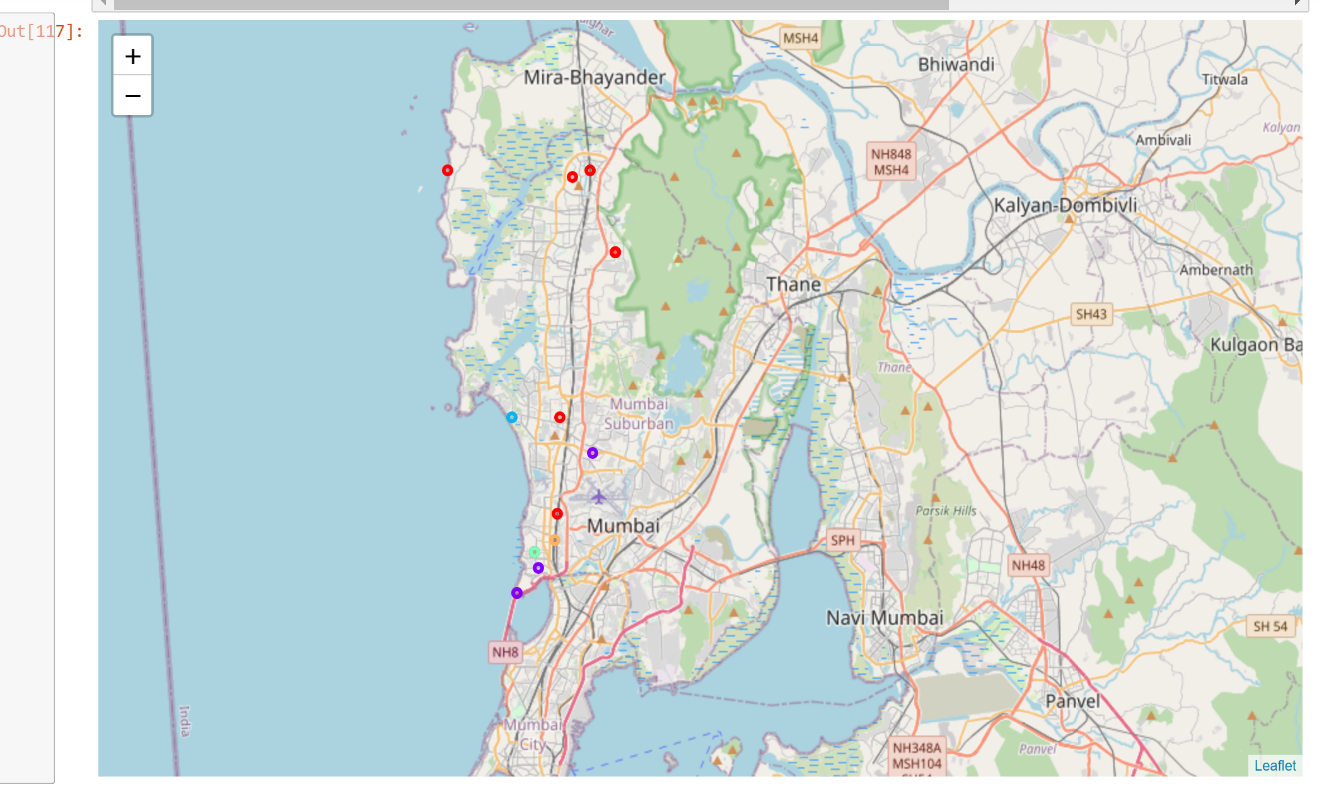
# run k-means clustering

kmeans = KMeans(n\_clusters=kclusters, random\_state=0).fit(Mum\_clustering)

# check cluster labels generated for each row in the dataframe

kmeans.labels\_[0:10]

## 5. Results

The results from k-means clustering show that we can categorize Toronto neighborhoods into 5 clusters based on how many Indian restaurants are in each neighborhood:

The results are visualized in map above with cluster 0 in red, cluster 1 in purple, cluster 2 in blue, cluster 3 in green and cluster 4 in orange. As you can see from the map. Most of the location belong to the cluster 0 in red.

## 6 .Recommendations

Most of the Locations where food place is not a most visited placed are in cluster 0 which is around Khar and Borlivali/Kandivali area and also all these places are generally where people like to visit.

This these will be best location for opening a fast food resturant to have good profit for the business