**COURSERA CAPSTONE**

**OPENING A NEW GYM IN MUMBAI, INDIA**

**INTRODUCTION**

**Mumbai**  is the [capital city](https://en.wikipedia.org/wiki/Capital_city) of the [Indian](https://en.wikipedia.org/wiki/India) [state](https://en.wikipedia.org/wiki/States_and_union_territories_of_India) of [Maharashtra](https://en.wikipedia.org/wiki/Maharashtra). According to [United Nations](https://en.wikipedia.org/wiki/United_Nations), as of 2018, Mumbai was the second most populous city in India after [Delhi](https://en.wikipedia.org/wiki/Delhi) and the seventh most populous city in the world with a population of almost 20 million. As per Indian government population census of 2011, Mumbai was the [most populous city](https://en.wikipedia.org/wiki/List_of_cities_in_India_by_population) in India with an estimated [city proper](https://en.wikipedia.org/wiki/City_proper) population of 12.5 million living under [Municipal Corporation of Greater Mumbai](https://en.wikipedia.org/wiki/Municipal_Corporation_of_Greater_Mumbai). Mumbai is the centre of the [Mumbai Metropolitan Region](https://en.wikipedia.org/wiki/Mumbai_Metropolitan_Region), the sixth most populous metropolitan area in the world with a population of over 23 million. Mumbai lies on the [Konkan](https://en.wikipedia.org/wiki/Konkan) coast on the west coast of India and has a deep [natural harbour](https://en.wikipedia.org/wiki/Natural_harbour). In 2008, Mumbai was named an [alpha world city](https://en.wikipedia.org/wiki/Alpha_world_city).

**BUSINESS PROBLEM**

The objective of this capstone project is to analyse and select the best locations in the city of Mumbai , India to open a new Gym. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question i.e; In the city of Mumbai, India if a business man is looking to open a new Gym , where would you recommend to open it?

**Target Audience of this project**

This project is particularly useful to Businessmans , property dealers and fitness freaks . This project is timely as the city is currently suffering from oversupply of Gyms.

**DATA**

**To solve the problem , we need the following data:**

* List of neighbourhoods in Mumbai.
* Latitude and Longitudes coordnates of those neighbourhoods.
* Venue data, particularly the data related to Gyms.

**Sources of data and methods to extract them**

This Wikipedia page(<https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai>) contains a list of neighbourhoods in Mumbai with a total of 93 neighbourhoods. We will convert this into a panas dataframe.

After this we will use FourSquare API to get the venue data for those neighbourhoods. This project will make use of many data science skills from data prepreocessing to , working with API(Foursquare), data cleaning , data wrangling , to machine learning and map visualization(Folium).

**METHODOLOGY**

Firstly , we need to get the list of neigbourhoods in the city of Mumbai. Fortunately the list is available in the Wikipedia page(<https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai>). We will convert this into pandas dataframe and luckily it also has the geographical coordinates. After populating the data into a panda dataframe and then visualize the neighbourhoods in a map using Folium package.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 metres. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and foursquare secret key. Foursquare will return the venue data in JSON format and we will extract the venue name , venue category and venues geographical coordinates. Then , we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of each occurrence of each venue category.

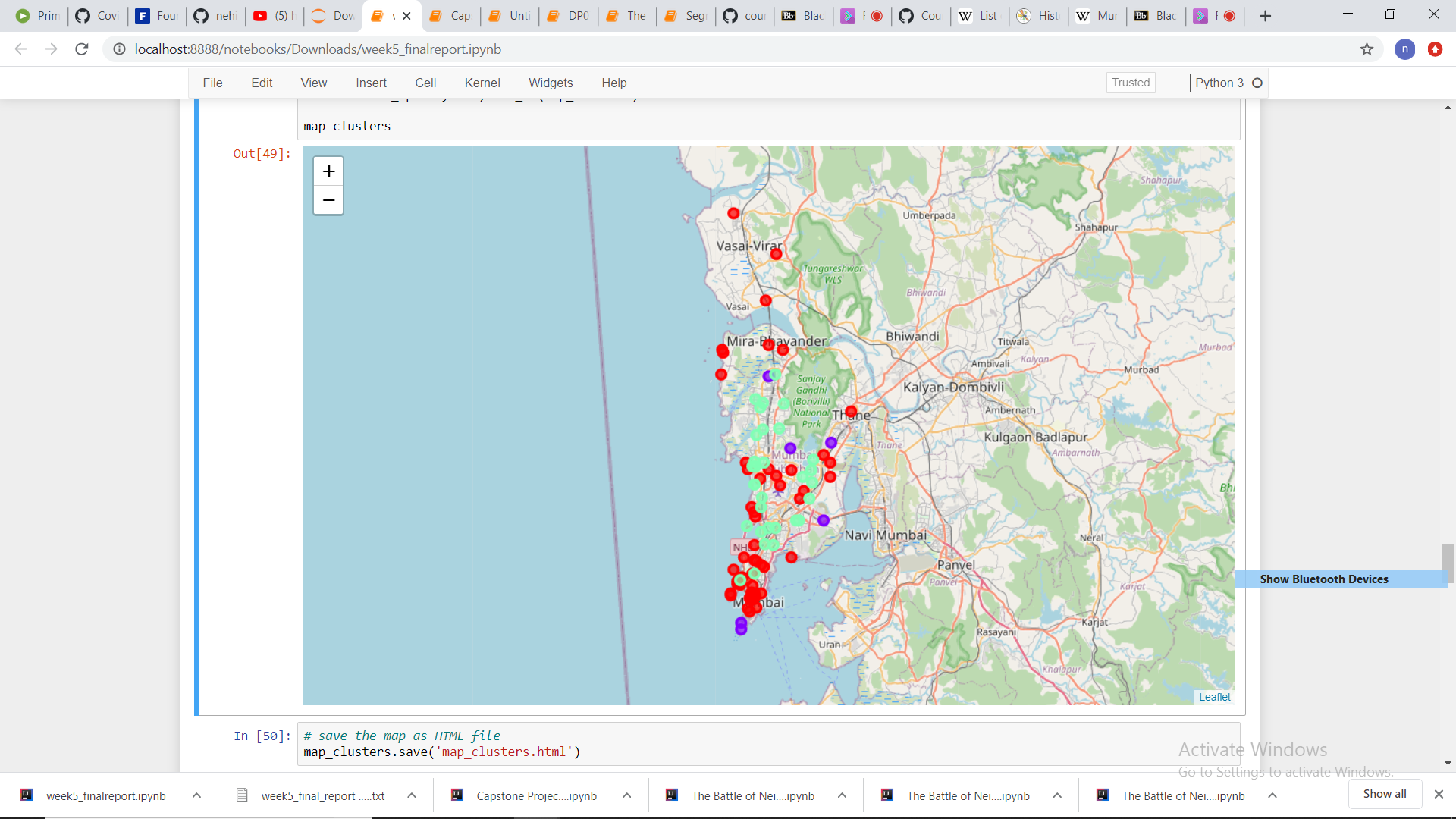
Lastly, we will perform clustering on the data by using k-means clustering . K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. We will cluster the neighbourhoods into 3 clusters based on their frequency for “Gym”. The results will allow us to identify which neighbourhoods have higher concentration of gyms while which neighbourhoods have fewer number of gyms.

**RESULTS**

The results from the k-means clustering shows that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence if “Gym”:

* Cluster 0: Areas with high concentration of Gyms
* Cluster 1: Areas with low number to no existence of Gyms
* Cluster 2: Areas with moderate number of shopping malls

The results of the clustering are visualized in the map below



**DISCUSSION**

As observations noted from the map in the Results section , most of the gyms are concentrated in the lower area of the Mumbai city , with the highest number in cluster 0 and moderate number in cluster 2. On the other hand, cluster 1 has very low number to no gyms in the areas. This represents a great opportunity and high potential areas to open new gyms as there is very little to no competition from existing gyms. Therefore, this project recommends property developers and businessmans to capitalize on these findings to open new gyms in neighbourhoods in cluster 1 with little to no competition .

**Limitations and Suggestions for Future Research**

In this project, we only consider one factor i.e. frequency of occurrence of gyms , there are other factors such as population and income of residents that could influence the location decision of a new gym. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. This project made use of the the free Sandbox Tier Account of Foursquare API that come with limitations as the number of API calls and the results returned. Future research could make use of paid account to byepass these limitations and obtain more results.

**CONCLUSION**

In this project, we have gone through the process of identifying the business problem , specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. businessmans and investos regarding the best locations to open a new gym. To answer the business question that was raised in the introduction session , the answer proposed by this project is: The Areas in cluster 1 are the most preferred locations to open a new gym.