Data

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

1. **List of neighbourhoods in Mumbai:** defines the scope of this project which is confined to the city of Mumbai, the financial capital of the country India in Asia.
2. **Latitude and longitude coordinates:** of those neighbourhoods which is required in order to plot the map and also to get the venue data.
3. **Venue data:** specifically, the data related to the shopping malls. We will use this data to perform clustering on the neighborhood.

Sources of data and Methods of Extraction:

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Mumbai) contains a list of neighbourhoods in Mumbai, with a total of 42 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and **beautiful soup**[1] **package**.

Then we will get the geographical coordinates of the neighborhood using Python **Geocoder package**[2] which will give us the latitude and longitude coordinates of the neighbourhoods

Foursquare[3] **API:** to get the venue data for those neighbourhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward.

This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

In the next section, we will present the Methodology where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique(s) that was used in the project.