Data

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

1. List of neighborhoods in Manhattan, NYC. This defines the scope of this project, which is confined to the Manhattan, NYC.
2. Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and to get the venue data.
3. Venue data, particularly data related to restaurant. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_Manhattan_neighborhoods) contains a list of neighborhoods in Manhattan, NYC, with a total of 25 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful-soup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use **Foursquare API** to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. 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 section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.
