**2. The Type of Data to be Collected**

**2.1 Using the data contained on Wikipedia:**

I obtained a list of [the most visited museums in the world](https://en.wikipedia.org/wiki/List_of_most_visited_museums), as it contained the essential data attributes to build my dataset. The key data elements included :

* Individual Museums Names
* Country of Origin
* Visitors per year

Using [BeautifulSoup](https://en.wikipedia.org/wiki/Beautiful_Soup_(HTML_parser)) (a python package to scrape website data), I scraped the Wikipedia webpage to extract the necessary data/tags and convert the raw data into a [pandas Dataframe](https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.DataFrame.html) for additional analysis.

**2.2 Obtaining the location coordinates**

The next step entailed obtaining the geographic coordinates (latitude and longitude) for each museum.

Using the Google API, I created a request to obtain the geo-coordinates for each museum. For e.g., the request for the first museum, which included the latitude and longitude coordinates of the Louvre in Paris were obtained as follows:

* https://maps.googleapis.com/maps/api/geocode/json?address=Louvre +Paris.
* The API request specifies the output in [JSON format](https://en.wikipedia.org/wiki/JSON) (open-standard file format that uses human-readable text to transmit data)

The detail steps for the scrapping process can be [obtained here.](https://github.com/romalanv/Capstone-Project-Setting-Up-A-New-Resturant/blob/master/Data_Files/North%20America_%20Top%20twenty%20museums%20.ipynb)

The layout of the final combined dataset of the geo-coordinates and museums scraped above can be accessed via the [following link.](https://github.com/romalanv/Capstone-Project-Setting-Up-A-New-Resturant/blob/master/Data_Files/Top%20Museums%20Lat%20Long%20.ipynb)

The combination of the museums and the geo-coordinates allowed me to develop a meaningful model.