Data

# Data Description

The problem was described in the Scenario-Business problem section as city clustering under the scope of similar climatological weather conditions. The data used for solving this problem were searched and retrieved from various climatology sites such as weatherbase.com. Despite the fact that there are huge amounts of data regarding climatology reports in cities and countries, an excel file which contains a summary of climatological data per city or per country, was never been created. Thus, *I was forced to create my dataset from scratch*. The excel file consists of 108 rows and 14 columns. From 87 countries 108 cities were selected: 87 capitals and 21 other major/known city. Only two rows contain some NaN values, but I decided to keep them just to exclude them later on, at the data preprocessing section.

The columns include the following not climate information: Country, City, Latitude and Longitude. This will help us create folium dots on a map that represent all cities reported in the excel file.

The rest of the columns contain climatological variables such as Average Temperature, Average High Temperature, Average Low Temperature, Average Precipitation, Average Number of Days With Precipitation, Average Length of Day, Average Number of Days Above 30-32 Degrees C, Average Number of Days Below 0 C, Average Relative Humidity and finally Average Wind Speed. These will be our variables which will help us find similarities among the aforementioned 108 cities by feeding them to our clustering algorithm.

The goal is to create a colored dot world map, where each dot represents a city and each color similar climate characteristics.