# Background

I work at a tourist office that provides travel packages to people who want to visit Greece for the summer. Lately we’ve experience higher demand for people wanting to visit the Greek islands and have with tasked with the problem of designing custom packages to offer. These packages would include multiple islands per package that the client can visit during their stay in Greece

Greece has a great number of islands cited as between 166 and 227. However, for the purposes of this analysis the focus is on the major 40 islands that are the main touristic destinations and can be accessed via boat and/or plane. Smaller islands would require special arrangements for travel and are therefore not in scope of the analysis.

The problem we are facing is to try and identify groups of islands that are either very touristic and therefore appeal to a younger demographic looking for fun and partying or less touristic and appeal to an older demographic looking for relaxation.

The question we need to answer is how can we find groups of islands that are similar in terms of their popularity with the tourists

# Data Preparation

In order to find the answer to the question above we first need to define our universe (i.e. which islands to select). A list of the islands that we can use can be found online on the Wikipedia page [[https://simple.wikipedia.org/wiki/List\_of\_Greek\_islands](https://simple.wikipedia.org/wiki/List_of_Greek_islands')]. This is a list of the 40 islands that we consider for their touristic popularity.

Next we will be looking at the Foursquare database to identify the venues on these islands. We will be doing that based on the island name for two reasons:

* The latitude and longitude are not readily available and most importantly
* The islands are of different sizes and in close proximity to one another therefore defining a radius around a central location would yield incorrect results. Often times the coordinates of an island are those of it’s main town (‘Chora’) and therefore are located near the port and not in the centre of the islands

We will set the limit to a very high value in order to include all the venues on the island rather than only the top x. We will review the request responses for any inconsistencies or bad data and cleanse the dataset. Finally we will extract the features we think are needed for the clustering of the islands and perform the ML algorithm to group them together.

The results of the data preparation and analysis can be found on:

<https://github.com/zdetor54/ibm-data-science-capstone/blob/master/GreekTourism_FinalCapstoneProject-Part1.ipynb>