The Olympic Games have always mixed pure sporting spectacle with national competition. During the Cold War the USA and the Soviet Union competed fiercely to win the most medals in each games. On a somewhat milder level, in Britain we often compare our medal count with that of Australia, one of our traditional sporting rivals. If you were in the UK during the summers of 2012 and 2016 you cannot have missed the excitement caused by the UK's success relative to previous years.

This competition is usually expressed in terms of the number of medals won by each country's athletes ((Figure 1 top panel). However, many interested watchers, especially those from smaller countries, have pointed out that the medal table is hardly a fair reflection of a country's sporting prowess. Some countries have a strong tradition of sporting excellence, but are simply too small to make an impact in terms of total medals. These commentators would rather look at the per capita medal count (Figure 1 bottom panel).

Looking at the per capita map above though, we see that large areas of the world are still very under-represented. Specifically, poorer countries do not win many medals per head of population. There are many reasons for this, including a lack of investment in sport and facilities, and fewer individuals who are wealthy enough to devote their life to training. As such, it has been suggested that we should compensate for wealth when measuring a country's Olympic performance.

In this practical you will investigate how the number of medals a country wins can be predicted from national population and GDP, and how consistent these relationships are, using the glm function in R.

Please check and follow the instructions given in Practicals 1 on documenting your work. You need to return your report using Minerva by the deadline. Use tables and figures to support your reasoning. The evaluation focuses on the quality of your report. Please respect academic integrity statement and avoid plagiarism.

Begin by downloading the data file medal_pop_gdp_data_statlearn.csv from Minerva.

This data file contains the following information for 71 countries (those that won at least one gold medal in each of the last three games):

- Country name (as recognised by the IOC)
- Population
- GDP (in billions of US dollars)
- Medals won in Beijing 2008, London 2012 and Rio 2016

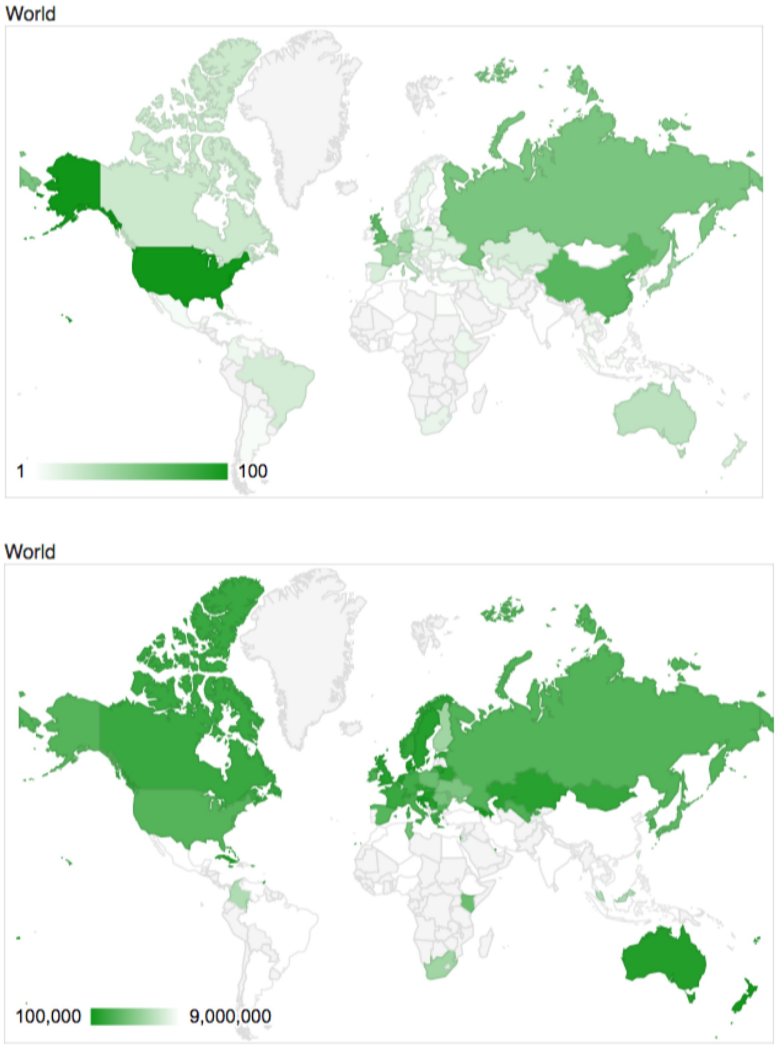

Figure 1: Total medals per country (top) and medals per capita (bottom) in the Rio 2016 Olympic Games (credit: http://www.medalspercapita.com/)

Tasks

1. Use Population and GDP as inputs and medal count in 2012 Olympics as outputs for the linear regression model. Assess the model predictions using the same inputs against the medal count in 2016.
2. Repeat the task 1 for log-transformed inputs. Which model performs better. Justify your reasoning. Discuss potential benefits and reasons for using the transformation.
3. Use K-means algorithm (kmeans function in R) to cluster the log-transformed inputs. Repeat the task 1 by training a linear regression model for each cluster (partitioning of the inputs) using the corresponding outputs. Combine the predictions suitably. Validate the optimal number of clusters. Justify your choice. Discuss potential benefits of the approach.
4. Derive and compute the probability that a country wins at least one medal given the estimated model parameters. Use the UK as the country and the model from task 2.

5. Perform model selection for the different models from tasks 1 to 3, respectively, and report your results. Which model would you choose to accurately predict the medal count? Justify your reasoning. Explain the model and comment on your findings.