Coursework

The file `gap.csv` is available on Moodle, and contains the GDP per capita, and the life expectancy for 142 different countries from 1952 to 2007. This data is from gapminder.org.

Load the data into R using the command

```
dat <- read.csv('gap.csv')
```

Start by splitting the data into two data frames, one containing GDP per capita, and the other life expectancy data.

```
gdp <- dat[,3:14]
years <- seq(1952, 2007,5)
colnames(gdp) <- years
rownames(gdp) <- dat[,2]

lifeExp <- dat[,15:26]
colnames(lifeExp) <- years
rownames(lifeExp) <- dat[,2]
```

In this project I would like you to write a short report analysing these data using PCA and MDS. I would like you to include the following:

- Some basic exploratory data analysis plots, showing how GDP and life expectancy have changed over the past 70 years. In particular, you should calculate the average life expectancy and GDP per capita for each continent through time and plot these. But you can also include other plots of your choice.
- Carry out principal component analysis based on your preferred choice of $\boldsymbol{S}$ or $\boldsymbol{R}$. Note that for GDP per capita, it is best to work with the log of the GDP as the values vary over several orders of magnitude between countries.

```
log_gdp <- log(gdp)
```

- Calculate the proportion of variation explained by each of the principal components, and provide a scree plot. Discuss how many principal components you would choose to retain.
- Look at the leading principal components for `log_gdp` and `lifeExp` and provide an interpretation for each component you have chosen to retain.
- Provide scatter plots of combinations of the first three principal component scores, indicating on the plot the names of the countries. Colour the data points by the continent they belong to. Identify and discuss any countries that have interesting characteristics based on your analysis. Can you explain what happened in any of these countries?
- Perform multidimensional scaling using the combined dataset of GDP and life expectancy:

```
combined <- cbind(log(dat[,3:14]), dat[,15:26])
```

Find and plot a 2-dimensional representation of the data. As before, colour each data point by the continent it is on. Discuss the similarity of this plot with your previous plots.