MATH4068: Coursework, Spring 2023

- This coursework is ASSESSED and is worth 20% of the total module mark for **MATH4068. This coursework is different to the coursework for MATH3030.**

- **Deadline**: Coursework should be submitted via the coursework submission area on the Moodle page by **3pm Monday 8th May (8/5/23)**.

- Do not spend more time on this project than it merits – it is only worth 20% of the module mark.

- **What to include:** your coursework is to write a report showing you data analysis of the given dataset. Below details the dataset you will be using, some code to get you started and some of the methods you should be using. Throughout you should provide interpretation of your results where relevant.

- **Format**: The format expected is the reports generated from using R Markdown in R Studio, where text, relevant codes and plots are all included in a clear format. Therefore, it is **strongly advised you use R Markdown in R Studio** to create your reports and then submit a single pdf or html file that has been produced by R Markdown in R Studio. **Do not submit raw markdown or R code** - raw code (i.e. with no output, plots, analysis etc) will receive a mark of 0. You may submit a word document however you will need to make sure it is similar in style to an Rmarkdown report e.g., include relevant code within your report and so I strongly advise against using word.

- **Report length**: There is no page limit, however your report should not be too long. You should aim to convey the important details in a way which is easy to follow, but not excessively long. Think about your reader, and try to help them quickly understand the key points. Avoid repetition and long print-outs of uninteresting numerical output.

- Please post any questions about the coursework on the Piazza discussion boards. This will ensure that all students receive the same level of support. Please be careful not to ask anything on the discussion boards that reveals any part of your solution to other students.

- The live sessions will be a good place to ask questions you may have about the coursework. I will not be meeting students 1-1 to discuss the coursework outside of these times.

- **Plagiarism and Academic Misconduct** For all assessed coursework it is important that you submit your own work. By submitting your coursework, you are agreeing to the universities policy on academic misconduct. Some information about plagiarism is given on the Moodle webpage.

- **Grading** The coursework will be marked out of 20:
  - 10 marks for technical content, use of R, and appropriate methods
  - 10 marks for presentation and interpretation of results.

  See attatched for details the dataset you will be using, some code to get you started and some of the methods you should be using

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 commands

```
gap.raw <- read.csv('gap.csv')
gap <- gap.raw
gap[,3:14]<- log(gap.raw[,3:14])
```

Note that for GDP per capita, it is best to work with log(GDP), as the code above does, when doing statistical analysis, as the values vary over several orders of magnitude between countries. For ease of plotting, it may be useful to split the data into two data frames, one containing GDP per capita, and the other life expectancy data.

```
gdp <- exp(gap[,3:14])
years <- seq(1952, 2007,5)
colnames(gdp) <- years
rownames(gdp) <- gap[,2]
lifeExp <- gap[,15:26]
colnames(lifeExp) <- years
rownames(lifeExp) <- gap[,2]
```

In this project, you will analyse this data using the methods we have looked at during the module.

- Begin by creating some basic exploratory data analysis plots, showing how GDP and life expectancy have changed over the past 70 years.

Principal component analysis

- Carry out principal component analysis on the log(GDP) data and on the life-expectancy data using your preferred choice of ***S*** or ***R***.

- 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 in each case.

- Look at the leading principal components for the log(GDP) and the life expectancy data, 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?

Multidimensional scaling

- Perform multidimensional scaling using the combined dataset of log(GDP) and life expectancy, i.e., using

  ```
  gap[,3:26]
  ```

  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.

Hypothesis test

- Consider the log(GDP) and life expectancy of each country in the year 2007. Conduct a multivariate hypothesis test to test whether there was a statistically significant difference between the mean log(GDP) and life expectancy of Asian and European countries in the year 2007. Were the continents more similar in the year 1952?

Linear discriminant analysis

We will now look at whether linear discriminant analysis can be used to successfully separate the continents.

- Use linear discriminant analysis to train a classifier to predict the continent of each country using the log(GDP) and life expectancy from 1952-2007. Test the accuracy of your model by randomly splitting the data into test and training sets, and calculating the predictive accuracy on the test set.

- Give a plot of the 2d projection of the data onto the first two eigenvectors found by Fisher's discriminant analysis approach. Discuss the difference between this plot and the plot you found using PCA.

Clustering

- Apply k-means clustering to the data. Give a plot of the final clusters you find, and discuss how you chose the number of clusters.
- Apply agglomerative hierarchical clustering. Try a variety of methods and give one or two carefully selected plots that you feel represent the most successful clustering. Note that as well as changing the measure of the distance between clusters (i.e., complete/single/average linkage etc) you may also want to consider scaling the data before computing the distance matrix, i.e., using

```
gap.scaled <- gap
gap.scaled[,3:26] <- scale(gap[,3:26])
```

- Discuss the similarity of the clusters you find using hierarchical clustering with the clusters you found using k-mean clustering, and the whether the countries naturally cluster by continent or not.

Linear regression:

- Use a linear regression approach to predict the 2007 life expectancy from the GDP values. Explain your choice of regression method and assess its accuracy.