Unassessed Coursework

This coursework is **unassessed** but is useful preparation for the assessed coursework. You should submit your solution at the lecture on Tuesday 23rd February.

The file UScancer.csv contains the number of cigarettes sold per person (CIG) in 43 US states in 1960 together with the number of deaths per one hundred thousand population from various forms of cancer: bladder (BLAD), lung (LUNG), kidney (KID) and leukemia (LEUK).

You have been asked by the manager of a statistical unit within the US government to conduct an analysis of these data and present your findings in a statistical report. The file Unassessed_Template.pdf contains a partially completed report to get you started. Your report should:

1. include a histogram of each type of cancer and comment on the shape of each distribution.
2. include a boxplot for each type of cancer (preferably on the same graph) and comment on any differences in their prevalence.
3. estimate the mean and standard deviation for each of the variables and comment on any interesting features of the data such as outliers (if any).
4. investigate if there is a relationship between the number of cigarettes sold and the number of deaths from the various forms of cancer.

Your report should be a maximum of 750 words. Please state the number of words used at the bottom of your report. Reports which exceed this limit will be penalised.

Hints and tips:

1. You are strongly advised to read section 3 of the lecture notes which describes the requirements of a statistical report.
2. You are only expected to use the statistical techniques and terminology described in sections 1 and 2 of the lecture notes.
3. The data and template report are available on the module moodle page. You can read the data into `R` using the following:
   (a) Right-click on the file (e.g. UScancer.csv) on the moodle page and select Save Link As or Save Target As.

(b) Save the file to your directory (e.g. Z drive on university computers).

(c) In R, go to File and Change Dir and select the directory where you saved the data file.

(d) To read the data into R, use the command

```
data<-read.csv("UScancer.csv")
```

(e) If you want to use commands such as hist(CIG), you will first need to attach the column names to the data using

```
attach(data)
```

DO NOT open the data file in EXCEL - it will screw up the formatting.