Assignment 2. ST661 2018
Catherine Hurley
Due on Wednesday November 7 6pm.

You should complete this assignment using Rmarkdown. Place the printed html file in in the box labelled ST661/ST663 in the ground floor of Logic house (under stairs).
Also upload the .Rmd file to Moodle.

1. Download the `cdc1.Rdata` file from moodle and load with

```
load("yourfolder/cdc1.Rdata")
```

   which gives a dataset called `cdc1`, a subset of the data used in class.

   (a) Change `exerany`, `smoke100` and `hlthplan` to be factors with suitable levels.

   (b) Calculate the proportion of men that have health very good or better. Do the same for women. Who has the better health?

   (c) Use `subset` to extract the smokers. For the smokers, calculate the proportion of men that have health very good or better. Do the same for women.

   (d) Repeat (c) for non smokers.

   (e) Based on your calculations of (c) and (d), compare the health of men and women, and the health of smokers and non smokers.

2. The definition of sample skewness of a set of data $x_1, x_2, \ldots, x_n$ is

$$\frac{m_3}{m_2^{3/2}}$$

   where

$$m_3 = \sum_1^n (x_i - \bar{x})^3/n$$

   and

$$m_2 = \sum_1^n (x_i - \bar{x})^2/n$$

   (a) Write function called `skew` to calculate the skewness of a numeric vector.

   (b) Test your function on `mtcars$wt` .

   (c) Test it also on a set of 50 numbers generated from the standard normal distribution.

   (d) Using `replicate`, replicate the calculation of (c) 1000 times. Draw a histogram of the results and calculate the mean skewness value.

   (e) Repeat (c) and (d), this time using a set 50 numbers generated from the exponential distribution with parameter 1.

   (f) Compare the results of (d) and (e).

   (g) Write code which calculates the skewness for every numeric variable in cdc1. Your code should construct a named vector of skewness values, with NA's for non numeric variables. Your code should not give any warnings.

3. (a) Make a plot of weight-wtdesire versus age for the cdc1 data. Use pch=20 and a colour vector which assigns colours as follows:

weight > wtdesire and age > 30	red
weight > wtdesire and age <= 30	blue
weight <= wtdesire and age > 30	orange
weight <= wtdesire and age <= 30	cyan

   Hint: use ifelse.

   (b) Recall in a previous lecture we used boxplot.stats to find indices of outliers. Write a function using boxplot.stats called boxplot.out that given a numeric vector returns "low" if there are low outliers, "high" if there are high outliers, "both" if there are high and low outliers, and "none" if there are no outliers.

(c) For the cdc1 data, make boxplots showing weight-wtdesire versus genhlth.

(d) Using boxplot.out and tapply, construct a vector indicating the type of outlier (high, low, both or none) in each boxplot.

(e) Construct a color vector that is "red" for boxplots that have high and low outliers, "blue" for just high outliers, "green" for just low outliers, and "yellow" otherwise. Use this vector to colour the boxplots.

4. Download kobe.csv from Moodle and load it in with

```
load("kobe.csv")
```

In the data frame kobe, every row records a shot taken by Kobe Bryant. If he hit the shot (made a basket), a hit, H, is recorded in the column named basket, otherwise a miss, M, is recorded.

(a) Write a while loop to calculate how many throws are required to reach 3 hoops.

(b) Write another while loop to calculate how many throws are required to reach 3 more hoops hoops.

Remaining questions are optional.

(c) Construct a vector containing the number of throws required to get 3 hoops.

(d) Construct a vector containing the number of throws required to get 3 hoops. Do not use for, if, while. Instead use cumsum, diff.

(e) Define the length of a shooting streak to be the number of consecutive baskets made until a miss occurs. For example, in Game 1 Kobe had the following sequence of hits and misses from his nine shot attempts in the first quarter:

`H M | M | H H M | M | M | M`

Within the nine shot attempts, there are six streaks, which are separated by a `|` above. Their lengths are one, zero, two, zero, zero, zero (in order of occurrence). Write a function called `streaklen` that calculates the streak lengths in Kobe's baskets. Tabulate the results and draw a barplot.