Exercise: Data Exploration with R

1. Read the table `batting.history.txt` into R. Find the average of PA, AB and H. Create a new data frame consists of the variables Year, Tms, N.Bat, BatAge and the average of PA, AB and H. Export the new data frame into a csv file called `BATTING.csv`.

2. Use data motor trend car road tests (mtcars) from available data frames in R.
   a) Find the factor for the number of cylinders in the car.
   b) Tabulate the percentage for the number of cylinders.
   c) Find the mean, median, first and third quartile as well as the inter-quartile range for the miles per gallon (mpg) of cars observed.
   d) Find the inter-quartile range for the miles per gallon (mpg), displacement (disp), gross horsepower (hp), rear axle ratio (drat), weight (wt) and quarter mile time (qsec) for all the cars.
   e) Tabulate factors for the combination of factors for miles per gallon (mpg) and displacement (disp). Create 4 factors for each variable.

3. Vicente et al. (2006) analysed data from observations of wild boar and red deer reared on a number of estates in Spain. The dataset contains information on tuberculosis (Tb) in both species, and on the parasite Elaphostrongylus cervi, which only infects red deer. In Zuur et al. (2009), Tb was modelled as a function of the continuous explanatory variable, length of the animal, denoted by LengthCT (CT is an abbreviation of cabeza-tronco, which is Spanish for head-body). Tb and Ecervi are shown as a vector of zeros and ones representing absence or presence of Tb and E. cervi larvae. Below, the first seven rows of the spreadsheet containing the deer data are given.

Farm	Month	Year	Sex	LengthClass	LengthCT	Ecervi	Tb
MO	11	00	1	1	75	0	0
MO	07	00	2	1	85	0	0
MO	07	01	2	1	91.6	0	1
MO	NA	NA	2	1	95	NA	NA
LN	09	03	1	1	NA	0	0
SE	09	03	2	1	105.5	0	0
QM	11	02	2	1	106	0	0

   a) Create the data set using functions you have learned so far.
   b) Find the average length of the animals.
   c) Find the number of animals with sex 1 used in the study.
   d) How many animals have tuberculosis in both species?

e) Suppose you decide to square root the length of the animals. Add the results to your data.

f) Find the mean absolute error (MAE) for the lengths of the animals.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \overline{x}|$$