# CS910 Exercise Sheet 2: Trying out tools

The exercises on this sheet ask you to accomplish a number of tasks. It's up to you to choose tools that will help you achieve them — either from those covered in lectures, or others that you are familiar with. Note: the marks are only indicative for the relative weights of the questions.

The results should not exceed 4 sides of paper – this should be more than enough to express your findings; any more and you are including a lot of unnecessary detail.

**Warning:** some of the questions require you to figure out how to do things in your tool of choice that have not been discussed in detail in lectures. You'll need to do some searching on the web and trial and error to find the commands that will produce the output you want.

## Manipulating Data

In this section, we will use the "automobile" data set, which contains information about the characteristics of cars from the 1980s.

This is available from: `http://archive.ics.uci.edu/ml/datasets/Automobile`

1. How many vehicle models are produced by manufacturers beginning with 'm' (e.g. mazda)? Describe how you produced your answer. [10]

2. How many different (unique) combinations of {fuel type, aspiration, number of doors, body style, drive wheels, engine location} are there present in the data (counting examples where there is an unknown value as distinct)? For example, there is a (gas, std, two, hatchback, rwd, front) combination in the data, so this counts 1; there is no (diesel, std, two, wagon, fwd, rear) combination in the data, so this is not counted. [15]

   What is the answer if we remove examples where there is some unknown value within this set of attributes?

3. What is the median and average price of four door vehicles? Describe briefly the steps you followed to achieve this. [15]

## Plotting Data

In this section, we will use the "abalone" data, which contains information about physical measurements of these sea creatures.

This is available from: `http://archive.ics.uci.edu/ml/datasets/Abalone`

4. Create a scatter plot that shows the distribution of height (on the x-axis) against length (on the y-axis). Ensure the axes are labeled clearly. Include a line of best fit, and give the equation. Describe the outliers in the data. [20]

5. Which pairs of numeric variables out of {length, diameter, height, whole weight, shucked weight, viscera weight, shell weight} have a correlation coefficient of more than 0.95? Outline briefly the steps you followed to answer this question. [20]

   **Hint:** Find a built-in function in your tool of choice that will compute the pearson product moment correlation coefficient for you. Can you find a way to apply it to all pairs of variables in turn?

6. Generate a single plot showing the (empirical) cumulative distribution function for the number of rings for each of the three sexes (male, female, infant). Outline briefly the steps you followed to do this. [20]

   **Hint:** In R, the "subset" and "lines" functions may be useful. For spreadsheets or gnuplot, you may have to manipulate the data first before plotting.