CSC3031: Additional Worksheet

TJ McKinley (t.mckinley@exeter.ac.uk)

You should keep an R script file (or preferably an R Markdown file) for all your work. Your script file should be neat, readable and commented throughout (but be succinct). Since we are working with data sets you are encouraged to use the `tidyverse` philosophy for consistency throughout. It would also be good practice to put your answers together with your code as an R Markdown document to produce a final PDF or HTML to act as a future reference.

1. Download the `titanic.csv` data file from the ELE page. Data collected from the Titanic disaster describe the survival status of individual passengers on the Titanic (it does not contain information for the crew). The variables recorded are:

   - `pclass`: Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd);
   - `survival`: (0 = No; 1 = Yes);
   - `name`;
   - `gender`: (Male; Female);
   - `age`: in years;
   - `fare`: Passenger Fare (in Pre-1970 British Pounds).

   You are interested in factors affecting the survival of individuals during the disaster. Firstly, read the data into R.

   (a) Use a suitable R command (or commands) to return the number of rows and columns of the dataset.

   (b) For each variable in the data set, identify whether the variable is:

   - explanatory or response;
   - numeric or categorical.
   - If it should be categorical, is it recognised as a 'factor' in R? If not, then turn it into a 'factor'. Ensure that it has the correct levels in the order defined by the data dictionary.
   - Remove any variables that do not make sense as either the response, or as an explanatory variable.

   (c) For each categorical variable, produce a bar chart of counts in each category. For each numeric variable, produce a box-and-whisker plot and a kernel density plot. Discuss which you think is more informative for each variable and why.

   (d) Assuming the response variable is `survived`, produce a series of suitable plots exploring the relationship between the response and each explanatory variable in turn.

   (e) Write a short passage summarising your thoughts about the observed patterns in the data, focussing on the potential relationships between the response and each explanatory variable.

   (f) Install the package `GGally`. This provides some additional functions for producing more complex plots, but built on `ggplot2`. Explore the website http://ggobi.github.io/ggally/, in particular the `ggpairs` function, and try to produce an informative multivariate summary plot of these data.