Homework 3

This homework requires you use skills from all previous lectures, although the focus is on data structures and common/routine functions.

1. Load the Camden Boroughs dataset, treating "Not applicable" and "Unknown", "999", "", and "NA" as missing data.
2. Create a new variable, `frl`, that is a numeric version of `Percentage.Claiming.Free.School.Meals` (*Note: Google is your friend here*).
3. Create a notched boxplot of `frl` by `Town`. Add a horizontal blue dashed line to the plot displaying the grand mean, and gray horizontal lines above and below the grand mean to display the 95% confidence interval around the mean. Note that to calculate the standard error around the mean you will have to use a custom function (one does not exist in base R). The basic function for computing the standard error of the mean is `se <- function(x) sqrt(var(x)/length(x))`. That is, the standard error of the mean is the square root of variance divided by the number of observations. However, you will need to modify this function to account for missing data. After modifying the function, you should be able to use it to calculate the standard error of any generic vector, `x`, with `se(x)`.
4. Briefly interpret the plot (hint, run `par(las = 3, cex.axis = 0.75)` prior to producing the plot to get all the towns to display on the x-axis). Discuss both measures of central tendency and spread. Use R Markdown in-text code to refer to something from the data (e.g., a mean, number of observations in a group, etc.).
5. Transform the data frame into a list of data frames separated by `Statutory.Low.Age`. The list should be of length 19. Produce a boxplot for `frl` by `Local.Authority.Name` for only schools in which the `Statutory.Low.Age` is 3. Use the list to produce the same plot for schools in which the `Statutory.Low.Age` is 11.
6. Fit a multiple regression model for schools in which the Statutory Low Age is 3, with `Number.Of.Boys` and `Local.Authority.Name` are modeled as predictors of `frl`. Compute predictor-residual plots for the model (*hint: use a package*).
7. Plot the density of `frl`. Overlay a plot of the likelihood, had the data been generated by a normal distribution with a mean and standard deviation equal to the sample mean and standard deviation. Do the data appear to have been generated by such a distribution? Why or why not?