Tutorial 09 - Variables, functions, data structures, and input/output

When to use what data structure in R

To some extent this is based on personal preference or what a particular function in R requires as its input.

There are, however, some clear cases where one data structure is the right one for the job:

- a list is useful for information that is linked in someway (e.g. from the same analysis or observation), bt differs a lot in its "shape"
- a dataframe is probably most often used because it holds tabular data with different data types and is the default data structure used when data is read in from a file

Factors in R

Factors are a data mode that we haven't talked much about.

These are sometimes useful when conducting statistical tests in R, but otherwise are kind of a pain to work with.

When reading data in from a file or creating a dataframe, it is often useful to specify the argument 'stringsAsFactors=FALSE' so that character data remains as a character data mode rather than being converted to factors (the default in R).

Challenge - Starting with R

- Create a vector containing 5 names
  - access the 3rd name using square brackets
  - access the 1st and 2nd name using square brackets
- Create a list containing your vector of 5 names from above and a 2x2 matrix containing the numbers 1 to 4
  - access the 4th and 5th name from the vector/list in your list/dictionary
  - access the number in the 1st row and 2nd column of the matrix
- Load the `wages.csv` file as a dataframe
  - access the 15th row of that dataframe with square brackets
  - find the minimum wage in the entire dataframe
  - find the gender and education level of the individual that earns the highest wage in the entire data frame