**What's a factor and why would you use it?**

In this chapter you dive into the wonderful world of **factors**.

The term factor refers to a statistical data type used to store categorical variables. The difference between a categorical variable and a continuous variable is that a categorical variable can belong to a **limited number of categories**. A continuous variable, on the other hand, can correspond to an infinite number of values.

It is important that R knows whether it is dealing with a continuous or a categorical variable, as the statistical models you will develop in the future treat both types differently. (You will see later why this is the case.)

A good example of a categorical variable is the variable 'Gender'. A human individual can either be "Male" or "Female", making abstraction of inter-sexes. So here "Male" and "Female" are, in a simplified sense, the two values of the categorical variable "Gender", and every observation can be assigned to either the value "Male" of "Female".

# What's a factor and why would you use it? (2)

To create factors in R, you make use of the function [factor()](http://www.rdocumentation.org/packages/base/functions/factor). First thing that you have to do is create a vector that contains all the observations that belong to a limited number of categories. For example, gender\_vectorcontains the sex of 5 different individuals:

gender\_vector <- c("Male","Female","Female","Male","Male")

It is clear that there are two categories, or in R-terms **'factor levels'**, at work here: "Male" and "Female".

The function [factor()](http://www.rdocumentation.org/packages/base/functions/factor) will encode the vector as a factor:

factor\_gender\_vector <- factor(gender\_vector)

# What's a factor and why would you use it? (3)

There are two types of categorical variables: a **nominal categorical variable**and an **ordinal categorical variable**.

A nominal variable is a categorical variable without an implied order. This means that it is impossible to say that 'one is worth more than the other'. For example, think of the categorical variable animals\_vector with the categories "Elephant", "Giraffe", "Donkey" and "Horse". Here, it is impossible to say that one stands above or below the other. (Note that some of you might disagree ;-) ).

In contrast, ordinal variables do have a natural ordering. Consider for example the categorical variable temperature\_vector with the categories: "Low", "Medium" and "High". Here it is obvious that "Medium" stands above "Low", and "High" stands above "Medium".