1.14 Using your own data

There are no short-cuts when it comes to data preparation. Regardless of what statistical software package you use, if your data does not follow certain rules, analysing it will produce strange results.

Here are some tips to follow when preparing your own datasets. But don't worry, since this course is not about preparing data, we have made several datasets available that we will use throughout the course (more about these on the next page).

We cannot help you to load your own data into iNZight, but you are welcome to use the discussion pages to seek help from others on the course.

Some general data preparation rules

1. You may wish to eliminate data which is:
    - **Incomplete** e.g. critical pages of a questionnaire are missing.
    - **Not collected according to instructions** e.g. the data was collected after the cut-off date or the respondent was not properly qualified.
    - **Not of interest** e.g. there is no variation (all values identical).
2. **Make corrections** to illegible, incomplete, inconsistent or ambiguous answers e.g. impossible dates of birth, membership of non-existent categories.
3. All numeric data should have the **same units**. E.g. All minutes, not seconds and minutes.
4. **Manipulate** the data where it requires weighting or scale transformations.
5. **Assign codes** to answers to assist with ordering or other data manipulations.
6. **Reformat** the data (e.g. names, scales, ordering) so as to make it accessible.
7. Give each variable a **unique name** (preferably with letters and numbers only).

8. Allow for **annotations** in the data e.g. total rows, documentation or comments.
9. Verify that **blanks** are actually blank and not spaces.
10. **Remove units** from data, e.g. 5 years/5 yrs/5yr should all be just 5 and $10 is just 10.

iNZight data preparation rules

Follow these rules when you use the iNZight data visualisation software:

1. The **first row** of cells must contain the names of the variables.
2. **Variable names should be unique** and cannot include spaces. They are case sensitive.
3. Everything in the column **beneath a variable name** is treated as data recorded on that variable.
4. Any **punctuation**, **control characters** and **alphabetical characters** will cause the data to be treated as a categorical variable e.g.
    - **Spaces** and **control characters** will turn your numeric variables (numerical data) into categorical (a category).
    - **Money**: do not use money formatting or currency signs. $10,000CR should be 10000 without commas and $1,000,000 should be 1000000.
    - **Dates**: If you enter "4/3/2014" it will be interpreted as a categorical variable. Think about how you want to use the date. For example you can split it into three variables – year month and day, or concatenate e.g. 4 Mar 2014 becomes 20140304, which is numeric and able to be ordered.
5. If a row starts with # then it is a **comment row** and is not imported. You may wish to use this for documentation or to exclude suspect data from your file.