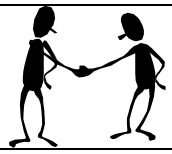
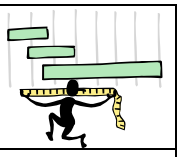
Clean it up!





Plan

A sample of 100 students has been selected from the 2009 CensusAtSchool database.

The students are a mixture of boys and girls and are from all year levels and all regions of New Zealand.

Only the "questions about you" (excluding the ethnicity columns) were selected. Most of the variables are measurement variables. You should be familiar with how these measures were made as your job is going to be to check these measures to ensure they are reliable.

Data

The table below shows part of the data set that you have to explore and "clean".

cleaningdata

	gender	age	country	handed	languages	height	rightfoot	armspan	wrist	neck	popliteal	indexfin...	ringfinger	year	region
1	girl	14	Russia	right	1	149	220	115	15	31	20	5	50	10	Southlan...
2	boy	10	newzeal...	right	1	141	22	138	14	30	41	60	77	6	Auckland...
3	girl	9	newzeal...	right	1	141	18	131	15	129	31	60	17	5	Wellingto...
4	boy	11	newzeal...	right	1	141	22	142	15	29	39	65	65	7	Canterbu...
5	boy	12	newzeal...	right	2	167	27	168	18	34	44	95	105	8	Auckland...
6	girl	14	newzeal...	right	2	175	255	176	17	33	48	81	80	10	Wellingto...
7	girl	13	newzeal...	right	1	162	25	64	15	30	18	80	75	10	Wellingto...
8	girl	14	newzeal...	right	1	158	25	163	18	36	40	97	95	10	Manawa...

You may like to graph the different variables to look for extreme values and to see if these values are suspicious, for example, incorrectly measured or recorded or are legitimate. Alternatively you could scan a printed table of the data, or sort using excel.

For the different variables you should decide what your boundaries are and then explain why values below or above are excluded.

For specific values you should decide whether to impute or exclude. This should be noted. If you decide to impute – explain what you did and why.

Some hints: use other measures of a person to help decide whether a particular measure is correct, for example, look at height and compare with the person's armspan, popliteal length and right foot length.

Reflection

- Why is cleaning data important out in the wider world?
- What were some of the common types of dirty data that was found in the C@S database that you explored?