Cleaning Dirty Data

Data integrity is a growing issue for many companies who, by law, have to store and protect accurate personal data. The accuracy of these data is compromised by human errors when the data are entered and errors that occur when data are transmitted from one computer to another. The consequences of these errors can involve huge costs for companies.

Can you think of reasons why errors in personal data would cost companies a lot of money?

The *CensusAtSchool* online questionnaire 2008/2009 www.censusatschool.org.uk asked children their height, foot length, length of index and ring fingers, hair colour, the soap location they would like to live in plus other questions. Even though these data are anonymous and are solely used to create teaching resources, it can be very time consuming to correct errors.

The spread sheet 'Cleaning Dirty Data.xls', (hard copy on page 2), contains **a few** of the errors found in the *CensusAtSchool* 2008/2009 database. Your task is to clean these data.

Task A

Highlight the data you would remove. Explain why you would remove these data.

- Remember these data were entered in 2008/2009.
- Do the ages and dates of birth make sense?
- Have any children got feet bigger than their height?
- Are some children too tall for their age?

Task B

Suggest changes you would make to the data to make it more accurate. Learners were asked 'What is your natural hair colour?' and 'Which soap location would you prefer to live in?' Each question had drop down response menus with choices 'dark brown', 'light brown', 'black', 'blonde', 'red' and 'other' for hair colour and 'Coronation Street', 'Albert Square', 'Summer Bay', 'Ramsey Street', 'Emmerdale', 'other' and 'none' for soap locations.

- Can you fit the 'Other Hair Colour' entries into the 'Hair Colour' categories, e.g. could auburn be considered light brown?
- Can you do a similar thing for the 'Other Soap Locations'?

Dirty Spread Sheet

Gender	Age	Date of Birth	Height (cm)	Foot Length (cm)	Index Finger (cm)	Ring Finger	Hair Colour	Other Hair Colour	Soap Location	Other Soap Location
M	12	08/01/1996	157	23	6.7	6.5	Black		Other	home and away
M	12	14/07/1997	146	22	5.5	7	Blonde		Other	neighbours
F	11	24/04/1997	1.40	19	6	6.30	Dark Brown		Ramsey Street	
M	11	26/01/1998	200	500	50		Dark Brown		Summer Bay	
F	13	20/12/1995	165	14	5	3	Light Brown		Other	home and away
F	12	23/02/1996	149	22	6.2	5.2	Dark Brown		Other	eastenders
F	5	01/01/1894	209	35	11	12	Other	blue	None	
F	13	12/11/1995	152	21	7.5	6.5	Other	auburn	None	
F	13	30/01/1996	150	21	8.5	7.5	Other	auburn	Albert Square	
F	12	12/02/1996	150	22	7	5.5	Other	blonde and brown	Coronation Street	
M	11	28/04/1997	127	13	10	4	Other	dark blonde	None	
F	13	19/10/1995	165	19	7.5	6.5	Other	dark blonde	Albert Square	
M	14	13/05/1991	145	22	6.1	7	Red		Albert Square	
F	15	29/11/1993	162	22	8	7	Other	ginger	Other	
F	16	14/06/1995	164	23	6	7	Other	mousey brown	Emerdale	
M	14	05/06/1994	180	26	7	9	Other	normal brown	None	
F	12	25/03/1996	158	26	7.5	7	Other	strawberry blonde	Coronation Street	
M	16	05/09/1994	178	29	11	8.5	Other	very very dark brown	None	
M	14	17/01/1995	176	27	8	7	Other	dark blonde	Other	Naighbours
F	5	25/12/1997	1.1	1		2	Other	pink		