How to clean data using SAS.

Some missing values are coded as 9 (or 99, or 999, etc). Some values are miscoded (i.e. sex = 2 when male = 0 and female = 1). SAS will include a coded missing variable as though it were a real value in all analyses (i.e. a 99 for height will be considered as 99 feet (or meters, or whatever unit is used as missing) and therefore should not be included).

SAS uses a . to denote missing for a numeric variable and a blank for a character variable. So, some data step syntax you will need if you wish to change the value of a variable:

for any numeric variables, e.g. if you wish to change the value of missing (9) to (.), then use the following syntax in the data step:

if *varname* = 9 then *varname* = .;

Thus you would type in:

data *a*; set *what.ever*;
if *varname* = 9 then *varname* = .;

for a character variable, use the following syntax:

if educ = '99' then educ=' ';

If you wish to create a variable which will group the original variables:

if a numeric variable: e.g. ages into young (0-17), medium (18-34) and older (35-55):

```
if 0 < age <= 17 then agecat=' young';
else if 18 <= age <= 34 then agecat='medium' ;
else if 35 <= age <= 55 then agecat='older';
else agecat=' ';
```

if a character variable: e.g. apgar scores:

```
if apgar in('1','2','3') then outcome='poor';
else if apgar in ('4','5','6') then outcome='fair';
etc
```

proc univariate provides descriptive statistics, i.e. mean median mode, stdev, var, max, min, etc for the variables requested. Obviously, **proc univariate** would be meaningless for many discrete variables, or for that matter, the variable **seqn**. The option **plot** (**proc univariate plot**;) produces a stem-and-leaf, a box plot and a normal probability plot. The latter we will learn about later.

To use:

proc univariate plot; var *name of the vars you want* ; run;

will produce a output with all the univariate statistics plus the stem-and-leaf etc, for the variables listed after the **var** command.

proc freq produces one-way to n-way frequency and cross tabulation tables. You are interested in one or two- way frequencies for the variables you have chosen. These will give you the counts of each value of the variables.

To use:

proc freq; tables *name of the vars you want*; run;

will produce a frequency table for all variables listed. If you want a two-way table, say *var1* by *var2* then use:

proc freq; tables *var1*var2*; run;