Many errors can be committed during data entry. While some are impossible to detect by looking at the raw data, many can be found and corrected. Here are several ways to find and fix, or clean, as many data errors as possible.

Before you analyze your file, you should know three things. First, rarely is a data file 100 percent error free. Because your analyses are only as good as the quality of your data, you should spend some time checking for file accuracy.

Second, data cleaning tasks tend to be easiest using SPSS syntax. The short examples here use SPSS programming code. These fixes can be made using the graphical interface, but the code is not difficult and helps to keep things straight.

Finally, data cleaning requires clear, logical thinking, and should be approached like a logic problem.

**Defining dirty data**
Two sources of detectable errors are data entry and respondent errors. Data entry errors include mistyping responses, entering data out of range or leaving an answer blank when a valid response was included in the questionnaire.

Responsive error includes: failing to accurately follow a skip pattern, writing a response that is difficult to interpret or providing false answers.

**Uncovering dirty data**
Use SPSS Descriptive to run an initial check on your data. Descriptive shows you the minimum and maximum values for each variable in the file. You can immediately see if there are questions that have data entered out of range. For example, if valid response options for a question range from 1 to 5, and the value entered is 6, the response is out of range.

Descriptive also calculate the average value for all cases for each numeric variable. For interval and ratio variables, the average should be interpretable. The mean can provide an initial indication of data accuracy. For example, if your survey was administered to high school seniors, and the average age was 30, you would suspect a data error.

While you can uncover dirty data through routine examinations of output, with Descriptive or Frequencies, you may need to seek out possible problems.

**Seeking data errors in skip patterns**
If you have complicated skip patterns in your instrument, you should write a few SPSS syntax jobs to ensure that respondents followed the skip patterns properly.

For example, suppose your questionnaire has the following structure: a respondent checking 1, “Yes,” to the Head of Household question (HH) is directed to answer the next question, about total household income (TOTINC). Those checking 2, or “No,” are instructed to skip to the question about personal income (PERSINC). To ensure that respondents who answered “No” did not erroneously respond to TOTINC, you create a variable whose values identify the number of cases that correctly and incorrectly followed the skip pattern.

Name the new variable “CLEAN1.” Say that any cases having the value of “0” are clean. What we want to do is create a rule, or logical statement, that gives cases violating the skip pattern a value of “1.” From there, we can identify the caseid of each case with a value of “1,” and go back to the questionnaire and fix the problem.

```
compute clean1=0.
if (hh eq 2 and toinc ge 0) clean1=1.
```

The result of this equation is that anyone who violated the skip pattern, stating they were not the head of household and who answered the total household income question would be assigned a value of “1” for CLEAN.

You would then need to look for cases that have this value, and try to resolve the problem. Ways to identify dirty cases include using the “Go To Case” function in the SPSS Data Editor, or writing a “List” command in the syntax window and asking for the value of the case identification variable.

Looking for illogical responses requires a similar equation. If you want to ensure that PERSINC is always less than TOTINC, write the following command:

```
compute clean2=0.
if (persinc gt totinc)
clean2=1.
```

Again, by running Frequencies on the variable “CLEAN2,” you would see if any cases have a value of 1, identify the caseid, go back to the questionnaire, and see if an entry error has been made. From there you simply change the incorrect value in the SPSS Data Editor.

Sometimes, even after running a series of data checks, you still encounter errors in your analysis. Most importantly, you should know how to identify and correct these errors. While identifying and correcting errors can be clearest through the SPSS command language, the principals are the same if you use the dialog boxes.

Either way, examine your data with a critical eye, and see if your distributions make sense. If they don’t, run a cleaning pass of the data, and see if errors exist. The cleaner the file, the more accurate the analyses. It’s as simple as that.