Table 1 SAS Macro Checklist

After running the Table 1 macro on your data, perform the following checks:

1. Look at the "N=" value. Is this the number of observations you were expecting?

Continuous Variables:

1. Are the mean and median values reasonable?
2. Are the mean and median values similar? If not, you may need to think about transforming this variable or creating a categorical version. If the mean is much larger than the median, the variable may be right-skewed. This may suggest that a predictor variable will not meet the linearity assumption if modeled as an untransformed continuous predictor in regression models (you should check this in any fitted models), in which case consider logarithmic transformation or modeling as a categorical predictor (e.g., quartiles). For an outcome variable, skewness may suggest that modeling it as an untransformed numeric outcome may not be the most meaningful scale, for example, if a 1 point difference is more important at low values of the outcome than it is at high values of the outcome. Logarithmic transformation may again be worth considering; this is appropriate if, for example, a 50% difference is equally important at both low and high values of the outcome.
3. For variables that can only take on non-negative values, is the SD more than half as large as the mean? This also indicates right-skewness (see #2).
4. Are the min and max values (that are shown with the median) consistent with the possible values for this variable? If not there is probably a data problem that should be investigated and systematically corrected.
5. Are the min or max more than 3 SD's from the mean? If so, these may be overly-influential outliers that need to be considered when presenting summaries and performing analyses. The median is often a better summary than mean when outliers are present. Logarithmic transformation often helps with positive outliers. DFBeta's can be calculated for regression models to assess the impact each observation has on the regression estimates. Sensitivity analyses can also be performed by deleting outliers and re-running models to ensure that results are not substantially affected.
6. If dates are shown, consider the min and max dates. Are they reasonable? Date values are particularly prone to corruption when moving from one source to another, e.g. from an EXCEL spreadsheet to SAS, so it is important to check that they are reasonable.

Categorical variables:

1. Are the categories labeled correctly? Look for misspelled words. Often misspellings occur in data, e.g., Mlae for Male, or data have been recorded in different ways, e.g. males recorded as “Male” and “M”. In both examples the computer acts as if those as two different categories. In addition, some statistical software is affected by capitalization. For example, R considers “Male” and “male” as different categories.
2. Do some variables have a large number of missing values? If so, was this expected or is there a problem with the source data?
3. Do some categories need to be combined due to small counts? Especially consider categories with counts less than 5 as candidates for combining.