The following guidelines have been written to produce data sets (e.g., using Excel) conducive to import in both **R** and **SAS** and, thus, are more restrictive than specifications needed for only R or only SAS.

Here are some general guidelines on formatting data to be conducive to statistical analysis (and data-import)

- The first row and only the first row should have the headers (column names / variable names)
- Variable names should meet certain criteria
  - Should start with a letter
  - Should have no spaces or symbols other than underscore and period (there should only be letters, numbers, and underscore _). Do *not* use "/" or "%" in a column heading or variable name. Ideally period (.) and underscore (_) should also be avoided in Excel/CSV data files, as these can create problems with SAS (.) or LaTeX (_). Camel-case can be used to help differentiate words (e.g., instead of "baseline weight" or "baseline_weight" you could use "BaselineWeight").
  - Check for spaces before or after a word in a cell – these can cause problems and should be removed.
  - Should be unique (there should be only *one* column with that heading)
  - The cell (that is the heading of the first column) should not be "ID" (this creates problems with Excel and opening *.csv files). Something like SubjectNum, SubjID, or PatientID is fine. http://www.alunr.com/excel-csv-import-returns-an-sylk.../
- There should be no formatting (e.g., subscript).
- **Ideally the file should be in a comma-delimited (*.csv)** or tab-delimited (*.txt) format (instead of *xls or *.xlsx).
  - Anything that would be lost when converting from .xls to .csv or .txt (e.g., colors, font formatting) should not be used, so this is actually *helpful* to remove anything that might be problematic in importing data.
  - If you do send an Excel (*.xls or *.xlsx) file you can have multiple tabs (multiple worksheets, this is preferable to sending multiple *files*, but each worksheet should have a name helpful for identifying the content.)
- The data in any one column should be of a consistent type; i.e., characters and numbers should not be mixed for any variable that needs to be analyzed as a number. So for example, if Blood Pressure was not taken for a given individual, that cell should just be left blank, comments should not be inserted (e.g., "Not Done" or "N/A"). If a column uses dates, the same date format should be used for all cells in that column (e.g., do not use both 8/24/2016 and 8/24/16)
  - If the data in the column is (appears) mixed, for example, ID's that are a mix of numbers (e.g., some are "1", "2", "3", while others are characters or alphanumeric (e.g., "A1", "B1", "B2"), then it is important to have the characters appear in the first few lines.
- I cannot analyze information like <10 for a variable that is numeric. I can't do anything with that. It can be entered as 10 or left as blank (missing). It cannot be analyzed as "less than 10".
- Variable values are case-sensitive. For example, "Control", "control", and "CONTROL" would be interpreted as 3 different values instead of one (so it is also important to avoid typos, e.g., "control"). For this reason, it is advisable to use short/simple values and avoid long words/phrases.
- Check that there are no stray spaces before or after a value. For example, if your variable is Gender and values include: `"male" " male" and "male "` ; these will be read as three *different* values instead of one.
- If the file format is Excel, ideally only a single worksheet will be used (again, think of how this would convert to a *.csv or *.txt file). However, as comments (e.g., "Gender (1=male, 0-female)") are not appropriate for column headings, one option is to have the data in one worksheet (that could be converted to *.csv or *.txt without loss of information) and have a second worksheet labeled "DataDictionary", "Key", or "Metadata" where comments like that (e.g., Gender 1=male, 0=female) can be stored for reference (again, it would be best to avoid formatting or special characters and stick to plain text so this information can be included in a report if needed).
- Avoid the following as variable values (i.e., do not have cells with the following)
  - quotation marks, e.g., do not have a height as 5'4" or put quotes around strings such as "Jane Doe" for a name versus Jane Doe.
  - Having commas in a cell. For example, having names as Doe, Jane will likely cause problems

- Suggestions for **Indicator Variables** (a.k.a. **Dummy Variables**). Using number codes for levels of a categorical variable can help avoid many of the problems with words. For example, using 1 and 2 for gender instead of "male" and "female" or even "m" and "f" avoids the problem of mistakes (e.g. sometimes using "m" and sometimes "M"). The problem, of course, is remembering whether 1 is male or female. I avoid these problems, I try to name the variable (for binary categories) in a way that the answer is, effectively "yes" or "no". It is standard for "1" to mean "yes" (or "TRUE") and "0" to mean "no" (or "FALSE"). So instead of having a variable called "Gender", if the variable is called "female", then females would be "1" (i.e., "yes, this subject is female") and males would be "0" (i.e., "is this subject female? No/False"). To be even more clear, instead of naming the variable "female", you may want to name it, for example, "IsFemale".

- **Dates** and **Dates and Times**. Dates and times can be particularly problematic for import and analysis. It is vital that they be in a *consistent* and *standard* format. I highly recommend using a format date option in Excel (but *saving* the file as .csv). The following formats worked for import into SAS and R (see example below). These are obviously NOT a comprehensive list. If you have dates or date-times automatically generated by a program, please contact me early in the process so I can test the format to see if they are compatible with import or if you will need to convert them before sending to me.
    - for Date:
        - MM/DD/YYYY (e.g., 10/28/1992) and
        - YYYY-MM-DD (e.g., 1992-10-28)
        - unformatted numbers: YYYYMMDD (e.g., 19921028) or MMDDYYYY (e.g., 10281992) as long as position is consistent (e.g., Day is always the 3rd and 4th digit) – for example, it would *not* work to have October 4, 2016 be written as 1042016, it would need to be 10042016. These will be read in as a number (not as a "date") but I can easily create a date in R or SAS.
        - It also works to have separate columns for Month, Day, and Year (all as numbers)
    - for Date-Times: MM/DD/YYYY HH:MM (e.g., 10/28/1992 19:23) – notice the hour is in military (24-hour) time.

mydate1	mydate2	mydatetime1
10/12/1974	1974-10-12	10/12/1974 1:00
6/13/2016	2016-6-13	6/13/2016 2:34
10/4/2015	2015-10-4	10/4/2015 23:23
1/3/2012	2012-1-3	1/3/2012 8:58
1/4/2011	2011-1-4	1/4/11 14:00

    - It is vital to be *consistent*. Whatever format you use, stick with it. Do not mix formats within a column!

<u>Guidelines for data file names</u>

- **The name should be informative** to help me identify the project and possibly version. Having your name (the clients name) the project (if I have multiple projects with you) and the date (e.g., in format YYYYMMDD) is helpful. For example: SmithSurvivalAnalysisData20170328.csv. Keep in mind I get many data files from many clients, so files named "Data for Shannon" or "For Statistical Analysis" are not helpful on my end. Also, keep in mind you may send me greater than one file over time (e.g., when data is added) so having a date in the file name is helpful in keeping track. For the purposes of keeping a data-analysis audit trail, it is best that I (the statistician) not be the one to change a filename (or any data, variable names within a data file).
- The name should have no spaces
- The name should have no symbols other than underscore (_), though ideally no symbols at all.
- It can be helpful to give the name in "Camel Case" (for readability in lieu of spaces or underscore), for example: SmithSurvivalAnalysisData20170328.csv

Suggestions for Longitudinal Data

Formatting data when the same variable is taken over multiple time points. There are two common ways of doing this ("wide" and "long").

The long method, has one column for each variable and multiple rows for each subject (one row for each time for each subject). Here is a simple example of the "long" format.

SubjectNum	DateYYYYMMDD	WeightPounds	BPsystolic	BPdiastolic
1	20160103	145	120	90
1	20160204	142	115	89
2	20151114	193	140	110
2	20160307	174	132	98
3	20150630	220	145	109
3	20150828	214	140	104
3	20160405	198	128	90

If you also have variables that are only collected once (things that would not change over the course of the study), you can keep that information in a separate database (data file), with one line per subject instead of copying over many lines in the longitudinal data set. The key is to have an identifier that matches that in the longitudinal data set (ideally with the same variable name) – it is easy for me to put these two data sets together in R or SAS.

SubjectNum	Female	DOB	TreatmentAssigned
1	1	19711214	1
2	0	19780403	1
3	0	29730522	2