How To Load Data Sets for Analysis

Much of the analysis we do in this course is based on different *data sets*, such as the “POL201” data set, which has information on students in the course, and the “HDEV” data set, which has information on Human Development across countries.

Data sets (or data frames as R calls them) are special types of objects, which contain a number of cases or observations and a number of variables in a matrix format.

Format of the File: In order to run analysis on variables in a data set, we must first load the data set into our R workspace. There are a variety of commands for doing this, depending on the format in which the data file is saved. In this course, all data sets are saved in the comma delimited format and end with “.csv”. The command to load comma delimited files is: *read.csv( ).*

Inside of the parentheses, you must enter the FULL and EXACT location of the file and end with the EXACT name of the file. The whole thing must be place in quotation marks.

Location: Data sets are stored and available in two places in this course:

1. The instructor’s personal website: www.sfu.ca/~sweldon
2. In the “Data Sets” folder on WebCT (lecture site).

Downloading from the Website: For most, this will be the easiest way to download the data sets and load them into your R workspace for this course. For the POL201 data set, for example, this is done with the following command:

```
> POL201 = read.csv(“http://www.sfu.ca/~sweldon/POL201.csv”)
```

First, look at the second part of this command (after the “=” sign). This is the R command (read.csv) with the location and name of the file to be downloaded (“http://www.sfu.ca/~sweldon/POL201.csv”). Take note of the direction of the backslashes “ / ”. Also notice that you need a double backslash at the beginning “ // “.

Second, the “POL201” at the beginning of the “=” sign is the name of the object that we will be creating in our R data set (Note: You can name it anything you want. POL201 is just the convention that I use).

If everything is done correctly, it will look like nothing has happened. That is, it will go directly to the next R prompt.

1. Type “ls()” to see if it is now in your workspace. It should be.
2. Type “names(POL201)” to see the variable names and also make sure it loaded correctly.

Downloading from WebCT: If your security settings prevent you from downloading these types of files, you will need to download the data set from WebCT, save it to your computer, and load it manually from your computer. You will know that this is the case when you type in “names(POL201)” and do not get the names of the variables—though it will still create an object called POL201.

1. Go to the “Data Files” folder on WebCT and click on the appropriate file.

2. Depending on your settings it will either prompt you to save the file or open up automatically into your web browser. You would prefer the former. But, if it opens automatically into your browser, you should then have the option of saving it. Either way, you need to save it onto your computer. If you are using a USB drive regularly, you will want to save it there.

3. You now need to find the EXACT location and NAME of that file. To do that, locate the file, right click on it and choose “properties” (for Mac Users choose “info”). There is quite a bit of info here, but look for “Location”. If it is saved on your Desktop, it should be something like: “C:\Users\Administrator\Desktop”. Whatever it says, this is the exact location and you will need to enter into the R console. Also, double check the spelling and capitalization of the file.

4. Open your R console and at the prompt enter something like:
   POL201 = read.csv(“C://Users/Adminstrator/Desktop/pol201.csv”)

   Notice that I have done 2 things:
   1. I have reversed the backslashes.
   2. I have added a double backslash immediately after “C:”.

   Also notice that I have changed the capitalization of “pol201.csv”. I did this because it was not capitalized when I saved it FOR ME. *You must look at your own file and how you saved it. You also must find your own location.* However, also notice that I have decided to recapitalize it in R by assigning it to POL201.

   Note to Mac Users: You will not need the “C:” at the beginning of the statement. Just start with double back slashes.

   IMPORTANT: Any error here whatsoever (spelling, capitalization, location, backslashes, etc.) will prevent you from loading the data set.

If everything is done correctly, it will look like nothing has happened. That is, it will go directly to the next R prompt.

1. Type “ls()” to see if it is now in your workspace. It should be.
2. Type “names(POL201)” to see the variable names and also make sure it loaded correctly.

Saving Your Workspace:

If you save your workspace, it will save all objects currently in your environment. When you start R back up at a later time, your workspace will automatically be restored. *Thus, if you do everything correctly, you will only have to load a data set once for the entire course.*

Potential Problems:

1. If you downloaded a data set and saved it in a format other than comma delimited (.csv), you will not be able to load it into your R workspace. Make sure you always save it in the comma delimited format.
2. Because of an error in how quotation marks (" ") get transferred from Word to R, you cannot simply copy and paste the commands to download data sets. Either,
    a. Always type in this command directly into the R console, or
    b. Copy and paste the command and go back into it, deleting the quotation marks and simply retyping them.
3. R is case sensitive. Pay attention to this.
4. R requires back slashes " / " , not front slashes " \ " when loading in files.
5. Do not forget that you need both the LOCATION and the NAME of the file to download it. When you right-click and choose properties, it only tells you the location.