GE375: Introduction to Quantitative Environmental Modeling Lab 6: Data Management

***Dataset Management***

**Objective:**

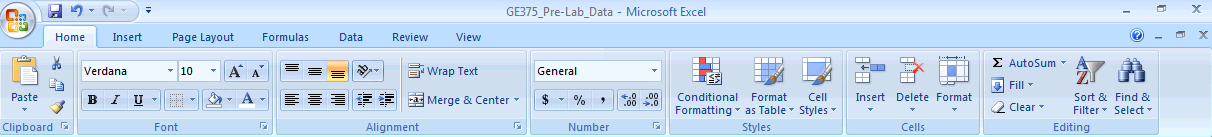
The purpose of this lab exercise is to touch upon some general rules of data management and presentation using survey data collected from a group of BU students.

**Background:**

A few years ago, on the first day of a GE class, students were given a survey that asked them to provide some basic information about themselves. Given only a list of required information, each student provided their answers in slightly different formats (i.e. text, numbers, or a combination of both) and with varying units of measurement or categorical variable names. While no format is correct or incorrect, whenever you are dealing with data information should be presented in the most clear yet concise manner possible. Variables should be clearly labeled and each variable should use a consistent unit of measurement (quantitative values) or set of values (qualitative variables). You also should consider the functionality of the dataset (i.e. how it will be used). In the exercise below we will format a data set two times, the first for the purposes of presentation, and the second for the purpose of archiving data and formatting it so it can be read by other software.

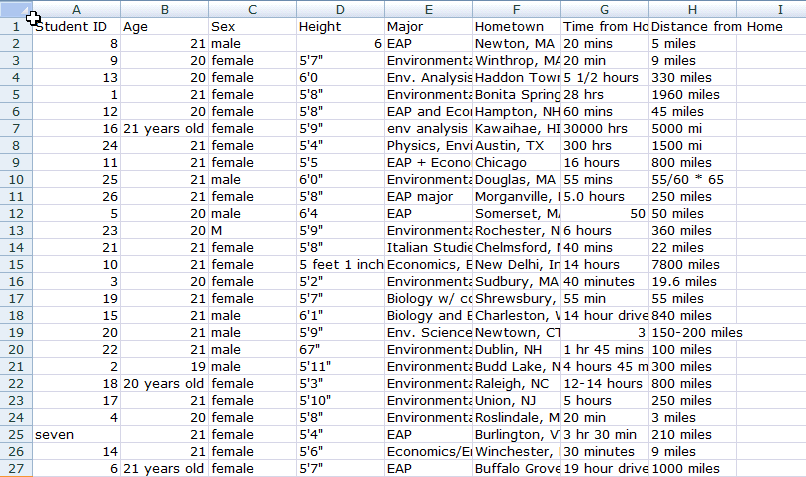
**Basic Excel Functions:**

Below is a screen capture of the Excel 2007 PC toolbar. For this first exercise, the Excel features you will be expected to use can be found under the Home tab, with the exception of the Text to Columns function, which is found under the Data tab.



I. **Raw Data**

The data from the survey was manually entered into Excel exactly as written by each student.



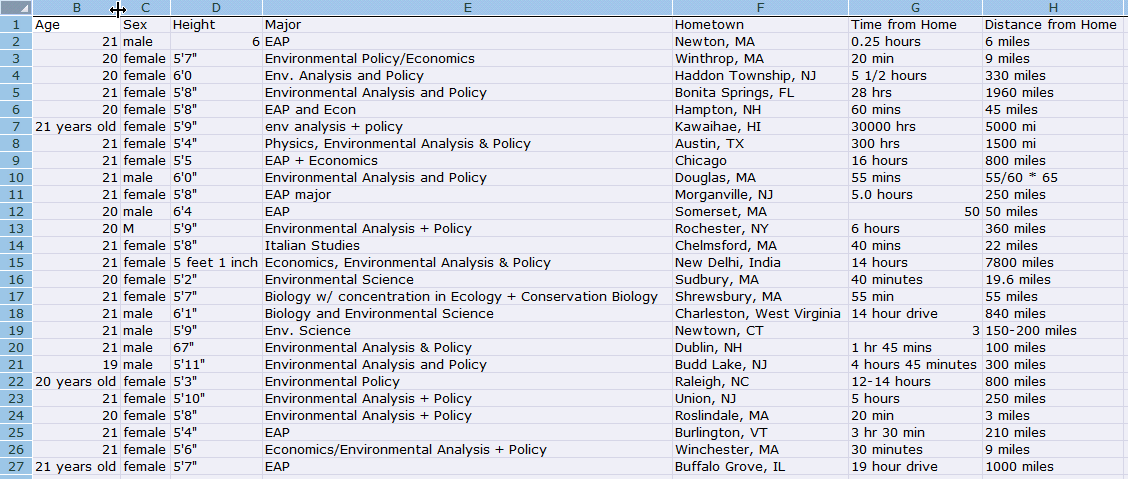
Note: For privacy reasons, student names have been replaced with a Student ID number

*Things to note about data display:* (1) Only text within the default cell width is displayed, unless the cell to the right is empty, and (2) Values in number format are right-justified, while values in text format are left-justified.

**II. Adjust column width to show all data**

*Automatically adjusting the height of a row or width of a column to fit text*: Place your cursor between two row or column headers—you will notice it changes to look like a vertical bar with arrows on either side (see below). Double click to adjust the size of the column to the left. You can also click and drag to manually adjust the size of the row or column.

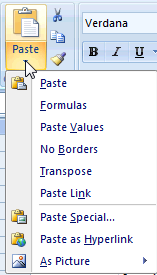
*To adjust all rows and columns at once:* Click the top left corner highlights all rows and all columns, then double click between two row or column headers.



**III. Edit the first four variables: “ID”, “Age”, “Sex” and “Height”**

Remember, you should always use consistent format and units for all entries. Quantitative values, like “ID”, “Age” and “Height” should be entered as numbers using the same unit for all values, with the units shown as part of the variable header.

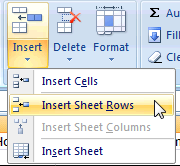
Qualitative values should be entered using simplified, yet easy to understand format. Naming of categorical values should be consistent (i.e. use M/F or male/female, not a combination of both). We will use M/F, as this is common abbreviation and requires fewer characters.

*Updating cell contents:* Update cell values such that all values for a variable have a consistent format and units (height should be converted to inches, so this will require some math). To change the contents of a cell, simply click on the cell you want to update and type in the new information. Remember, if Excel detects text, the cell will be left-justified, while if numbers are detected, the cell will be right-justified.

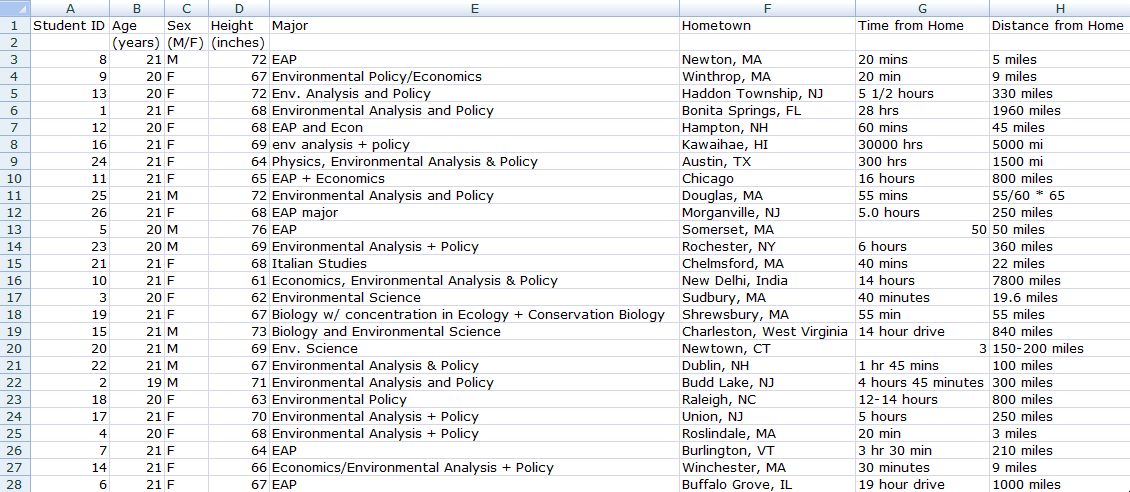
*Copy and Paste:* To copy and paste, use the copy and paste icons or shortcuts (Ctrl+c to copy, Ctrl+v to paste). When pasting, you may select to paste only the value of a cell or the formula contained in the cell, among other options.

*Drag to copy:* To fill multiple cells with the same value or with a sequence, simply click and hold the black box at the lower right corner of a highlighted cell A description...and drag down or across. This trick will be useful for changing ‘male/female’ to ‘M/F’

*Inserting rows:* You will need to add a row for displaying units for each quantitative variable. Go to the Insert pull-down menu and select Insert Sheet Rows to add a row. Note, rows are always added above the selected row.



Once you have finished editing values for the first four variables and formatting the header information, your data should look EXACTLY like this:



**IV. Edit and format “Major” values:**

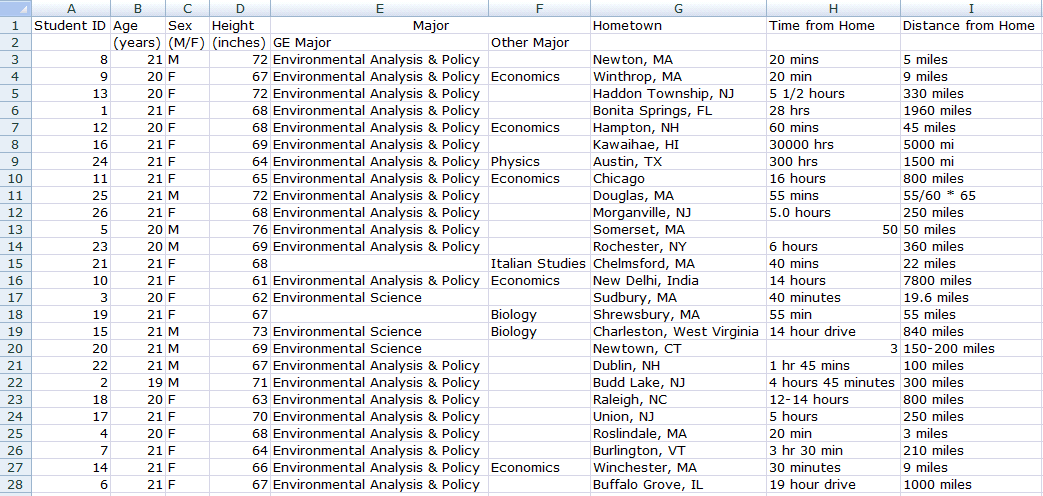
The values for the variable “Major” frequently contain two values due to double-majors. Typically, we will only want one variable value per cell, so we will want to divide majors into each student’s Geography (GE) major and any other majors.

*Insert a new column* in the same way you added a new row, except select Insert Sheet Columns instead. New columns are always created to the left of the highlighted cell/column.

*Edit the header for this variable* to show the variable ‘Major’ with two subdivisions: “GE Major” and “Other Major”. To do this, highlight the cell originally containing ‘Major’ and the cell to its right at the top of the new column, then click A description.... Within the ‘units’ row (row 2), label the original column as “GE Major” and then label the new column “Other Major”.

Finally, *manually update the information for the ‘Major’ variable* placing each person’s major(s) in the appropriate columns and using consistent names for all majors (drag to copy might be helpful here).

Your dataset should now look EXACTLY like this:

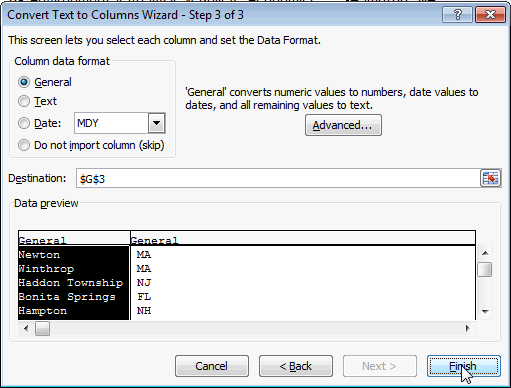
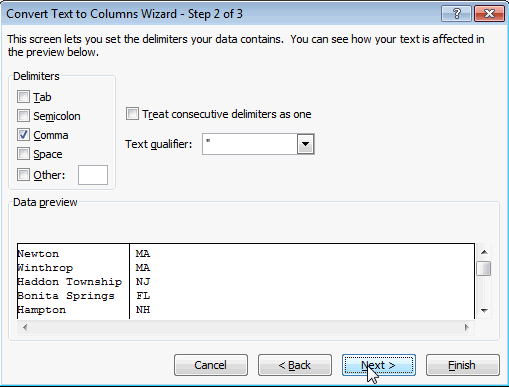
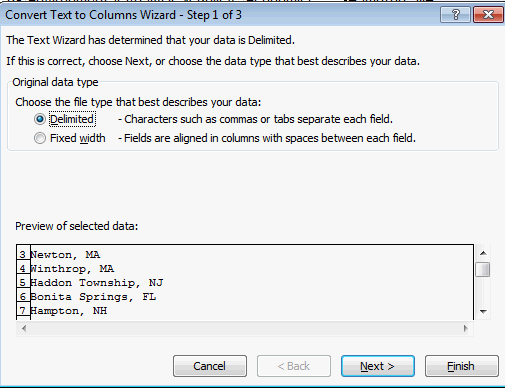


**V. Edit and format “Hometown” values**

Like “Major”, “Hometown” also has two components (City and State/Country), yet unlike “Major”, these values are in a consistent format.

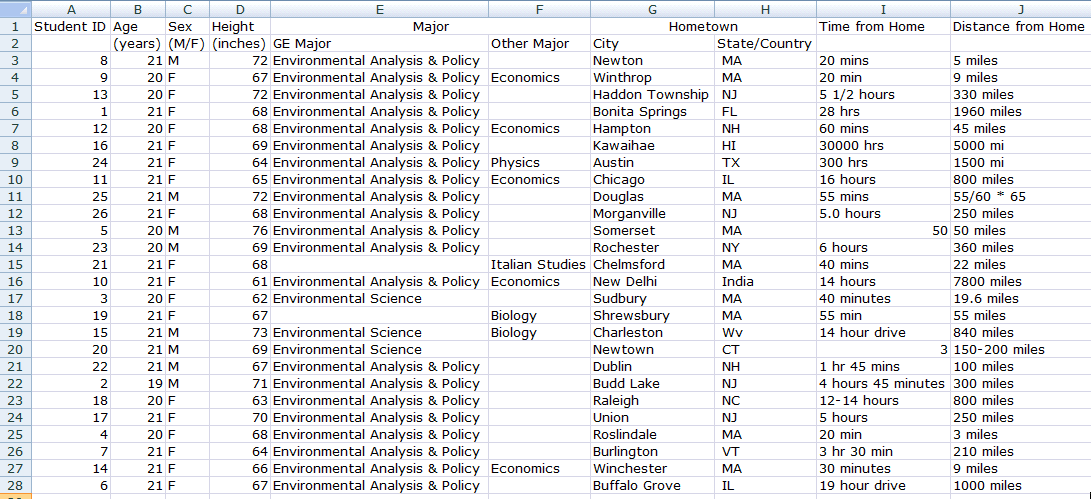
*Insert a new column* to the left of “Time from Home”. This must be done before converting text to columns so that the divided data has somewhere to go.

*Use Data>Text to Columns to convert city, state into two columns:* Highlight all “Hometown” values then click the Text to Columns icon, which can be found under the Data tab (rather than the Home tab we’ve used up until now). Choose ‘delimited’ in the first option window, then check only ‘comma’ in the next. Make sure all values have a consistent format.

A description...

Finally, correct the column header like you did for Major, using “City” and “State/Country” as sub-variable names under “Hometown”.

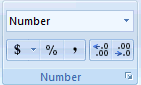
Your dataset should now look EXACTLY like this:

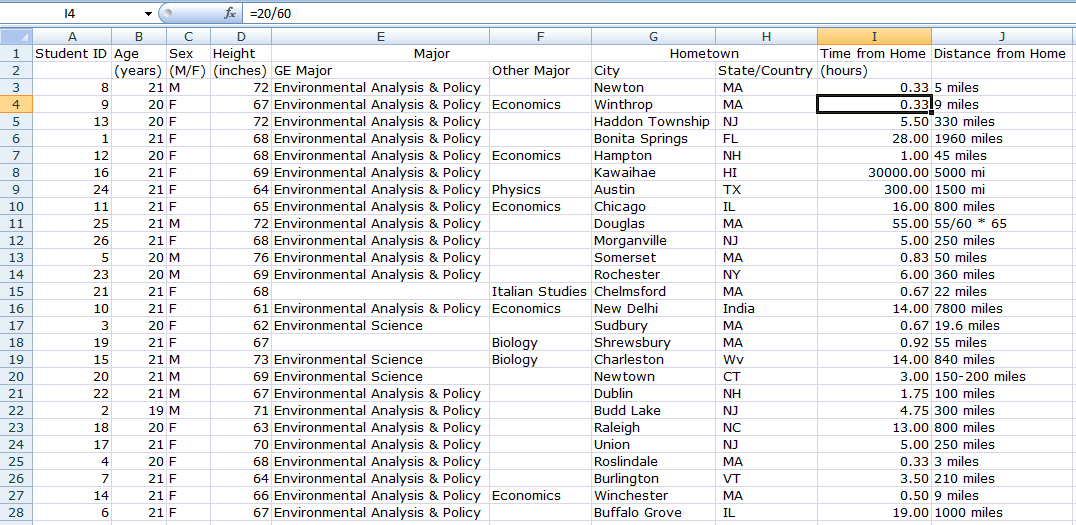


**VI. Edit and format the “Time from Home” values using formulas**

The values for “Time from Home” are a mixture of minutes and hours. For the purposes of this exercise, we want all times in hours.

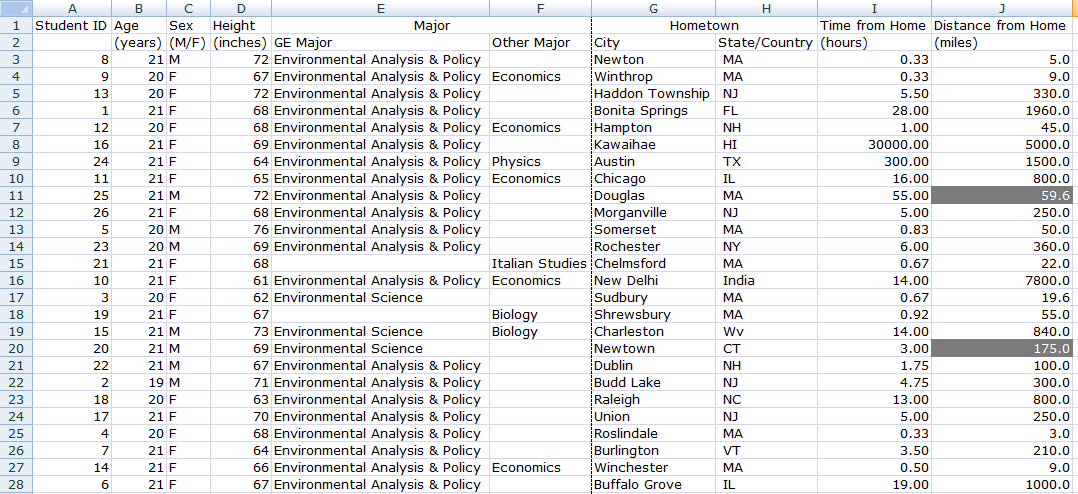
*Change the cell values to minutes using Excel’s built-in calculator:* Simply type “=” followed by an equation, and Excel will display the result in the selected cell. For example, note the highlighted cell shows 0.33 hours, while the formula bar (at the top of the screen) shows “=20/60”, the equation used to convert 20 minutes to hours. Once you have converted all values to hours, make sure to *add the units to the second row of the variable header*.

As a final step, select all values in the “Time from Home” column, and *change the cell format* from General to Number, using two decimal places. Use the Number pull-down menu under the Home tab, then use the icons to the lower right to increase or decrease number of decimal places. 



**VII. Edit and format “Distance from Home” values**

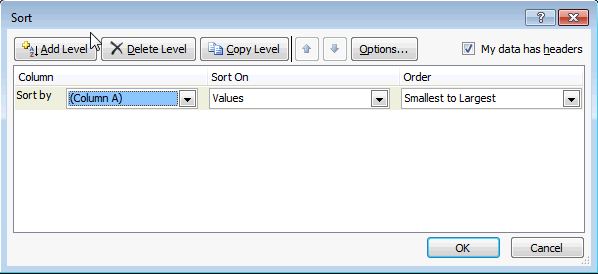
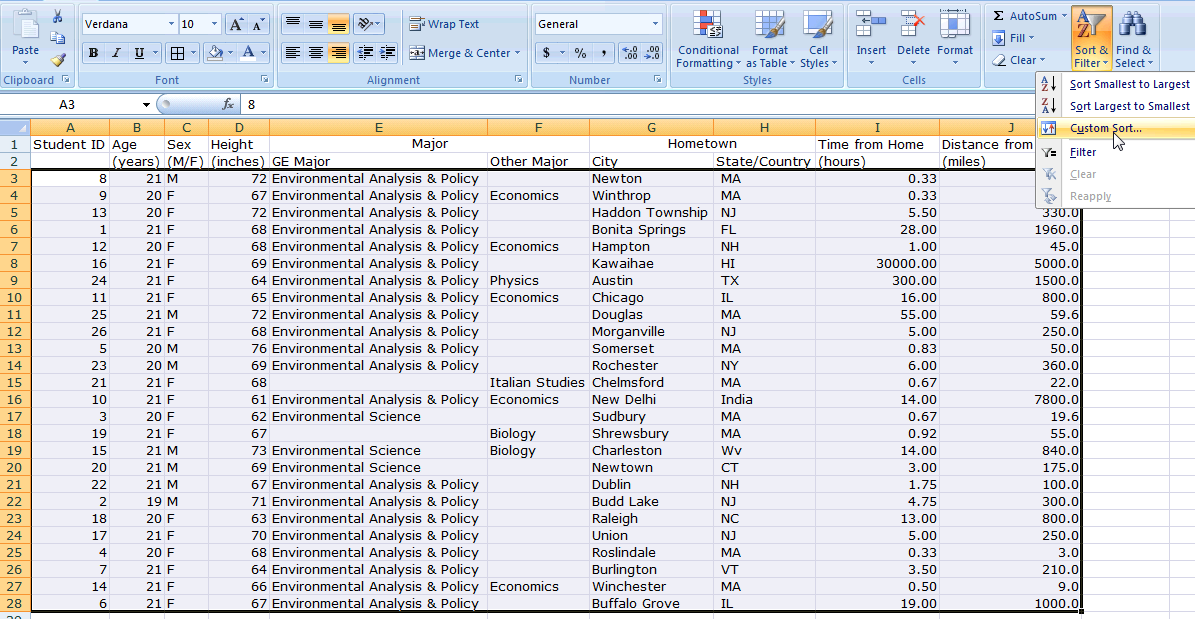
Using the tools and techniques you learned throughout this exercise, edit the “Distance from Home “ values to match the image below. Note, the cells we’ve highlighted with gray backgrounds contain formulas, thus your spreadsheet should also use formulas to calculate these values.



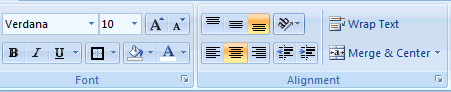
**VIII. Formatting dataset for visual appeal**

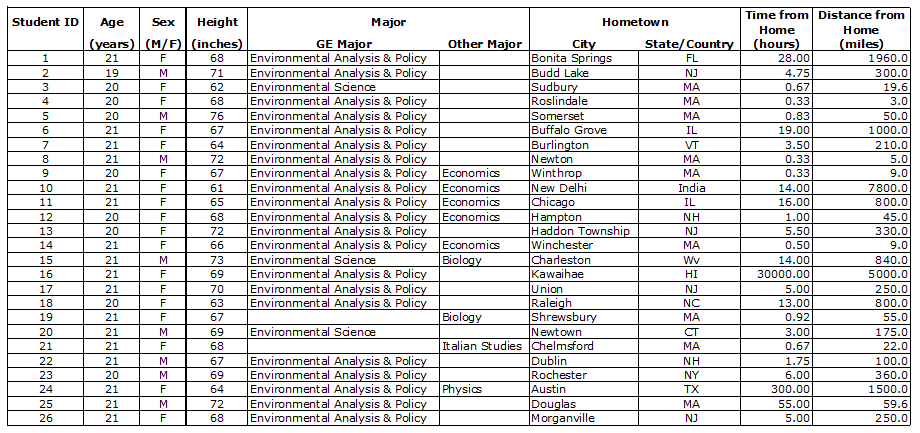
Up until now, all formatting has been performed in an effort to better organize the dataset. Now that we have finished editing, we want to make sure the finished data is easy to read. Note: there are many ways this data could be formatted, but the purpose of this lab is to replicate the formatting shown. If you have not done so already, make sure to **SAVE** your file.

*Organizing data by Student ID:* Use the Sort feature of Excel to reorder the data based on the Student ID. Select all data values (note: do not include variable headers in the selection). Then select ‘Sort & Filter’>’Custom Sort’, and sort by Column A, smallest to largest.



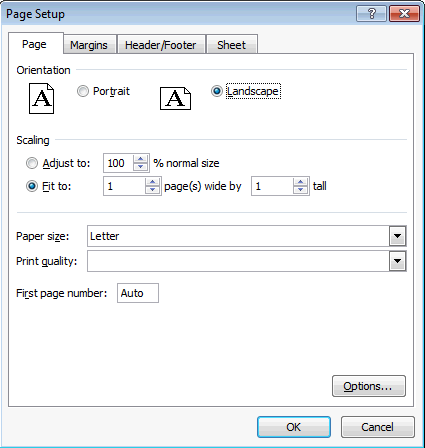
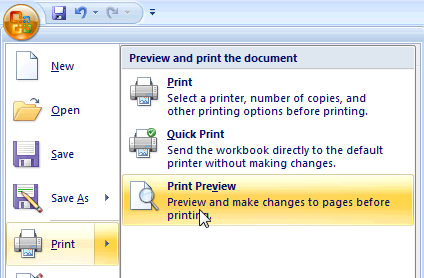
*Cell formatting:* Using icons in the Font and Alignment sections of the Home tab, try to match your dataset to the one shown below (in Print Preview view).





Note: the borders are all identical thickness

Use the *Print Preview* to make sure your dataset matches the example as best as possible. If your entire dataset does not fit on one page, in preview mode, go to ‘Page Setup’ then change the orientation to ‘Landscape’ and select ‘Fit to 1 pages wide by 1 pages tall’.



Note: When viewing your spreadsheet in Print Preview, borders may appear thicker than expected or may not appear at all. These issues will not carry over to the actual printed copy.

**IX. Formatting data for export**

Data input into spreadsheets are commonly export so that they can be read and analyzed using other software outside of Excel, such as the statistical software R. It is also a good practice to archive data in a format that is independent of a particular piece of software or operating systems. Best practices suggest that one should aim to archive data in a format that would still be readable in twenty years. If you think about how much computers have changed since 1994 you will quickly realize that this is a non-trivial challenge. In this section we will format the spreadsheet for export as a “.csv” (comma-separated values) file.

*Make a copy* of your table by selecting everything (Ctl-A on Windows, Command-A on Mac), copying it (Ctl-C), moving to Sheet 2, and then pasting everything (Ctl-V). Double click on the label “Sheet 2” and rename the sheet to “Export”.

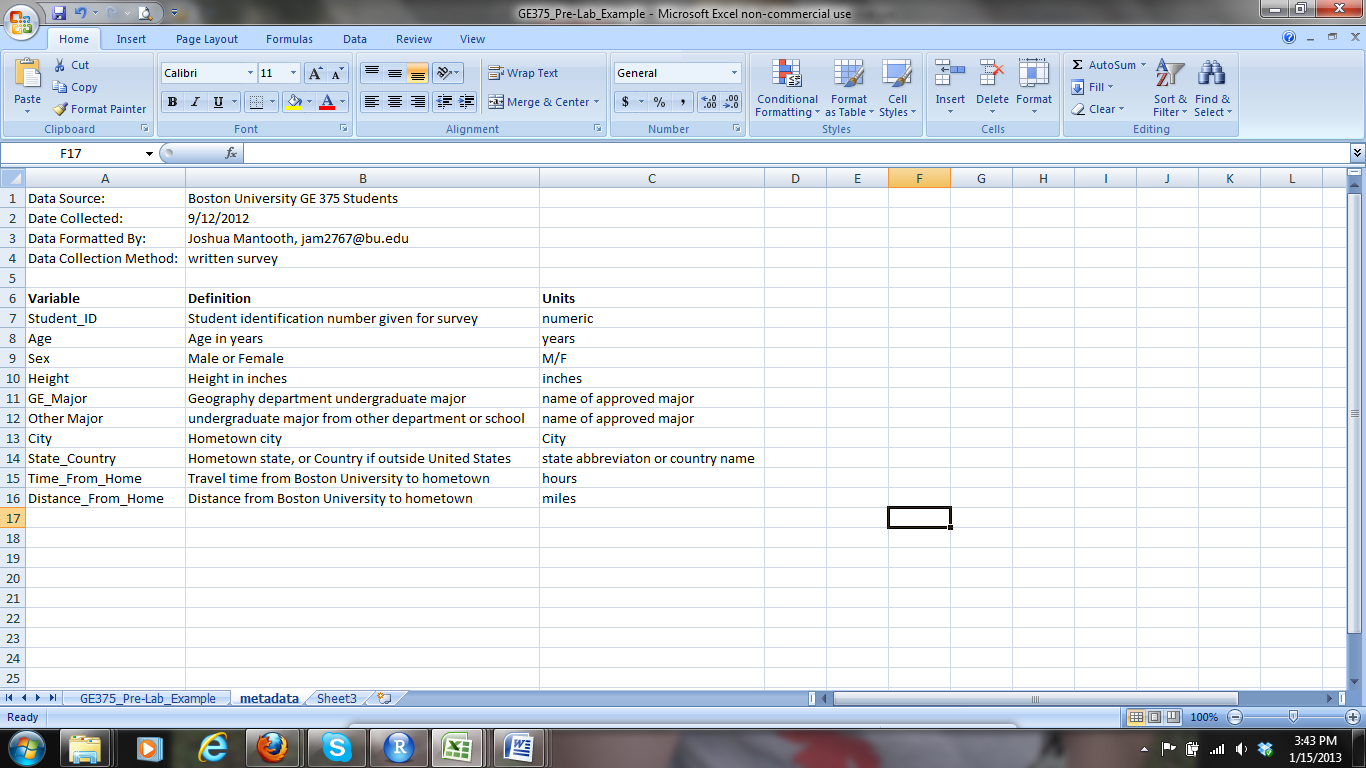
*Header formatting:* First we will need to change our column headers so our .csv file is easier to manipulate. Many programs will only read a single line for the header, so we will move our unit definitions to a new “Metadata” sheet and remove the second row of headers that we previously created. Many programs also don’t like spaces or non-standard characters in header names, so we will also use best practices for renaming the headers. Specifically, we’ll want to replace all spaces between names in the header with an underscore, \_ .

*Metadata:* Metadata can be defined as “data about your data”. More precisely, metadata provides “descriptive information (content, context, quality, structure, and accessibility) about a data product and enables others to search for and use the data product” (definition from [www.dataone.org](http://www.dataone.org)).

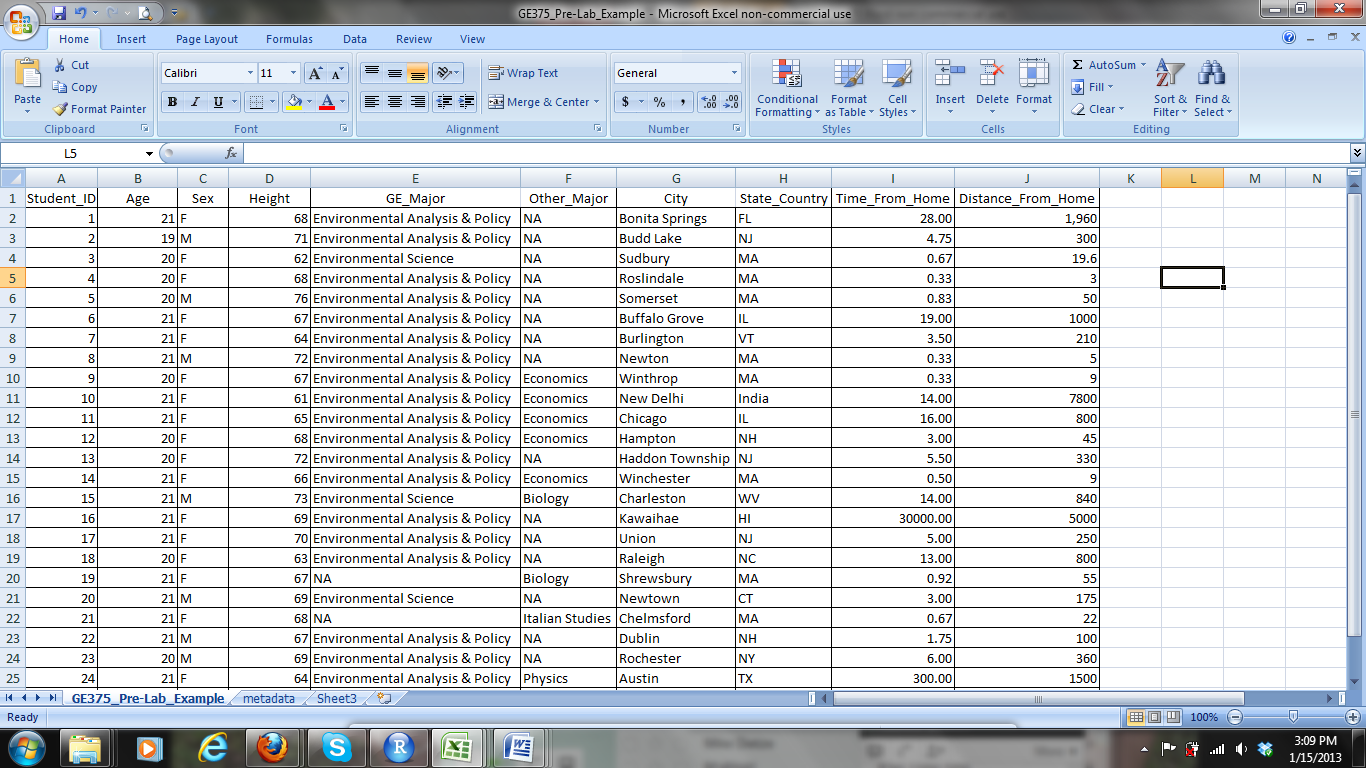
Rename your “Sheet3” as “metadata” and create your metadata table. Some basic guidelines for metadata can be found here: <http://www.dataone.org/best-practices/identify-and-use-relevant-metadata-standards>

You do not have to fill all of the listed categories in, but at a minimum you should include what the data are; who formatted them; when they were collected; and how they were collected.

At a MINIMUM your metadata should contain the following information with the same or similar format:



Your newly formatted header should look like this:



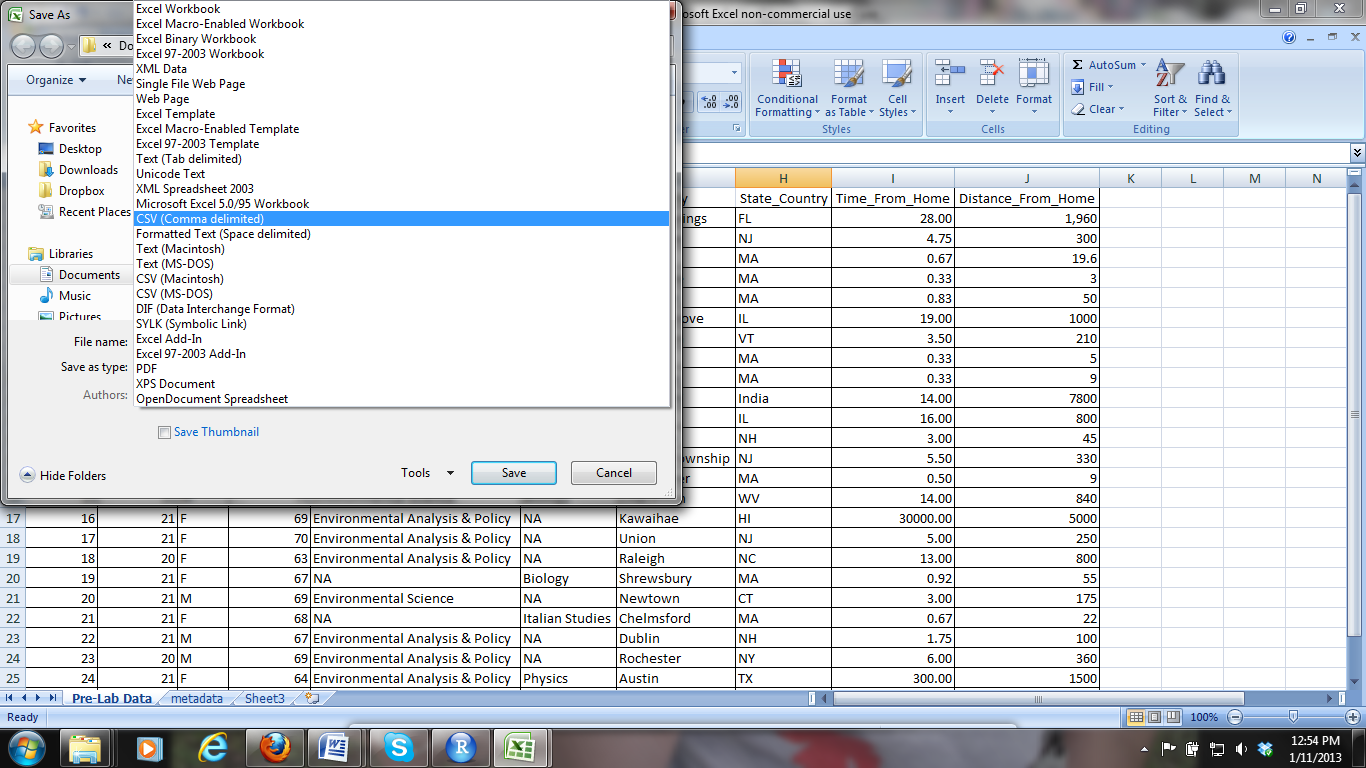
*Formatting text columns*: the columns “GE major” and “other major” both have cells that contain no information. According to best practices these cells should contain an “NA” (a standard abbreviation for ‘not available’). This tells us, and most statistical software, that there is no information in this cell. Replace all empty cells in these columns with “NA”. Replacing blank cells with NA prevents other software from parsing the data incorrectly (i.e. skipping certain columns) or from misinterpreting the data (e.g. assuming that a blank cell equals 0 or NULL).

Your data should now look like this:

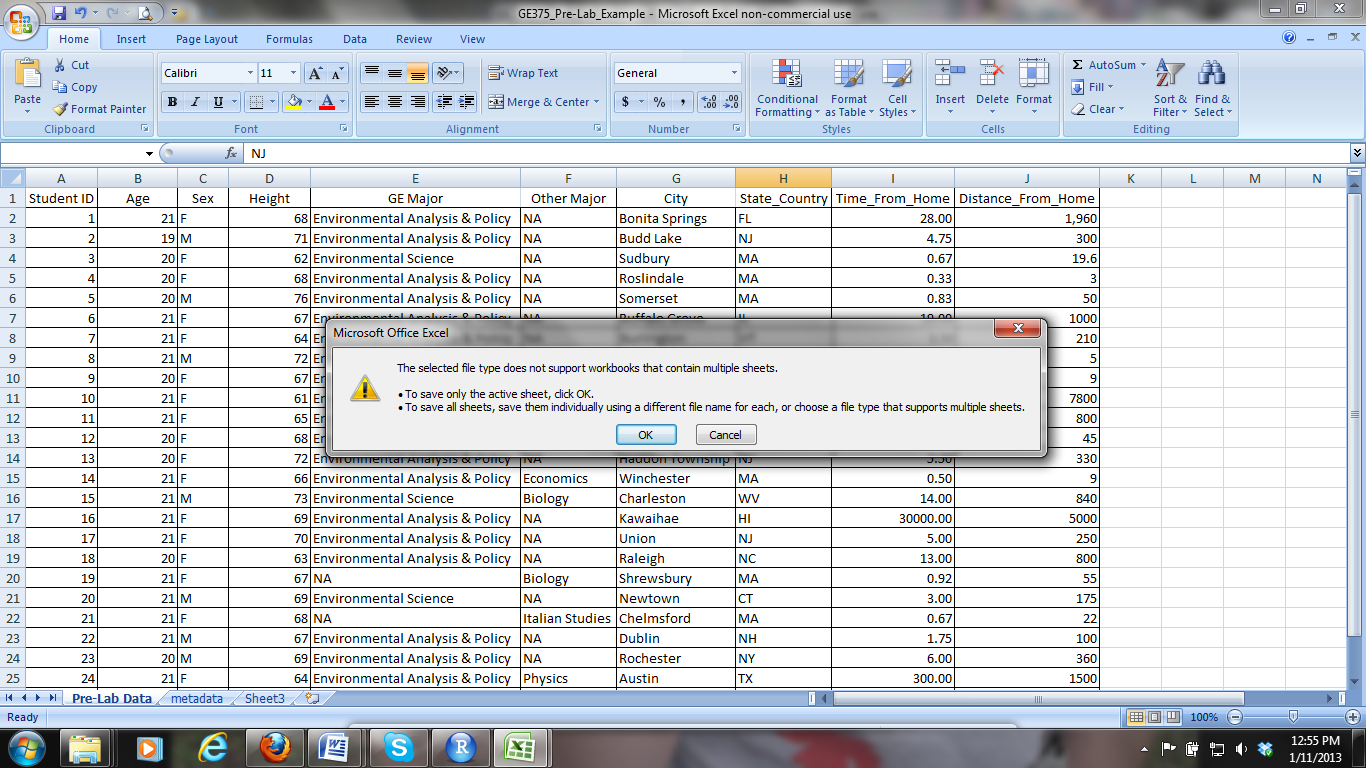


\*Be sure to SAVE the full file before the next step\*

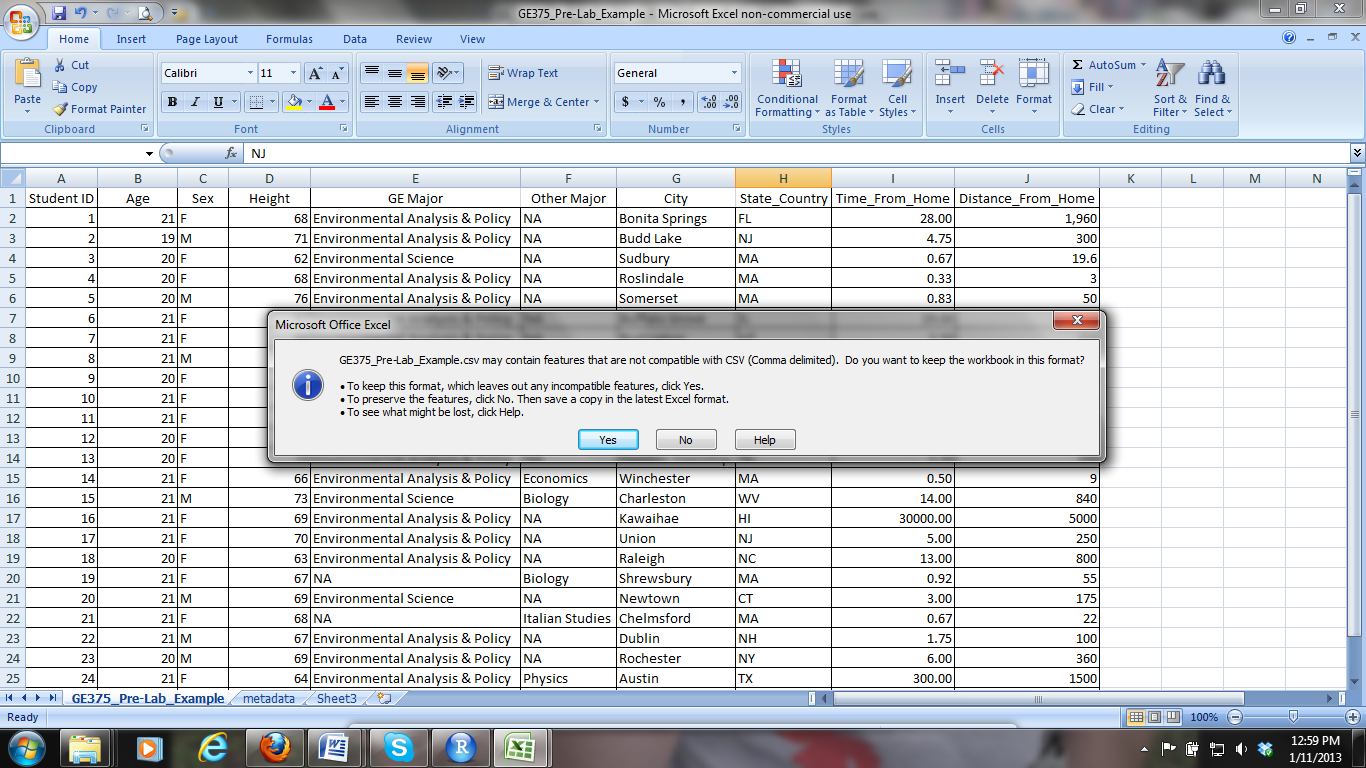
Your spreadsheet should now be ready to be exported/saved as a .csv file. To do this, go to “Save As” then select “Save as type” and choose “CSV (comma delimited)”.



The following warning, informing you that .csv does not support multiple sheets, may appear. This is fine, as we only want to save the main sheet as .csv at this point. Click OK.



Another warning, telling you that you may lose information may then appear. Since we already saved our Excel file before exporting to csv, we will not lose any information. Click OK.



You have now saved your file as a .csv and can now be imported into a number of statistical software packages.

**Turning in the assignment:**

Assignments should be submitted in electronic form. Please include your name and the lab number in your file name (e.g. Lab06\_EmilyFrench.xls) and make sure to submit BOTH the csv and xls files on Blackboard.

Your .xls file should contain three sheets.

* **Sheet 1** should contain your formatted data and should match the reference example
* **Sheet 2** should be your data formatted for export
* **Sheet 3** should contain your metadata

Your csv file should contain your exported data

You will be graded on your attention to detail in editing and formatting the dataset. To receive a perfect score files must use the exact same wording, spacing and borders as the reference example.