Today’s exercise will continue to focus on the `subset()` function, although this time you’ll calculate measures of dispersion like variance, standard deviation, and coefficient of variation. Practice, practice, practice by typing your code each time.

**Part 1: Review of `subset()`**

By now you may be getting comfortable with reading data files (using `read.csv`) and assigning the output to a variable using `<-`. This file is a data frame object (containing a mixture of numeric and character columns). The most common way to specify a particular column within a data frame is using the `$` symbol. For example: `georoc$SIO2` or `georoc$rock.type`. Remember that R is case-sensitive. Use `levels()` to find the names and spellings of values within a text (factor) column within a data frame.

You must specify two things when you are writing a `subset()` function. The first is the desired output columns (or an entire data frame). The second is a logical statement, typically involving `==`, `!`, `<`, or `>`. Subset will give you only the rows of your specified output (either as a data frame or vector) where the logical statement is true. For example, if I wanted a new data frame containing only rocks that are rhyolite, I could write:

```
subset(georoc, georoc$rock.type == "Rhyolite")
```

**Part 2: Subset with multiple criteria**

It is possible to include multiple logical statements within a single subset function, using the logical operators `&` (“AND”) and `|` (“OR”). The “or bar” `|` should be just above the enter key.

For example, if I wanted to find the Co (cobalt) values in andesite rocks that have at least 10% CaO, I could write:

```
subset(georoc$CO, georoc$rock.type == "Andesite" & georoc$CAO > 10)
```

Note that you must write the full column name (data_frame$column) each time.

Here’s how I could find the Co values in rocks that are either basalt or andesite:

```
subset(georoc$CO, georoc$rock.type == "Basalt" | georoc$rock.type == "Andesite")
```

Here are some tips:

1. NEVER combine `&` and `|` in the same subset function! It can be done, but is dangerous. Break it into two steps.
2. Only use `&` if the criteria come from different columns.
3. Only use `|` if the criteria come from the same column.
4. You can always write several steps instead of one complicated subset.
Part 3: Dispersion functions

The R functions to calculate dispersion measures are fairly straightforward. They are:

\texttt{var(x)} \ #calculates the variance

\texttt{sd(x)} \ #calculates the standard deviation

Each function requires a single numeric vector variable as the input. That vector can (and will often) be a column from a data frame or the output from the subset function.

Both of these functions have the built-in option to remove NA values.

\texttt{var(x, na.rm = T)} \ #calculates the variance after removing NA

I have abbreviated TRUE as T (note the single equals sign, not the logical ==).

Exercise

First, read the following data files and store each as a variable (make sure to give each variable an appropriately descriptive name).

http://people.ucsc.edu/~mclapham/earth125/data/chesapeake.csv
http://people.ucsc.edu/~mclapham/earth125/data/georoc.csv
http://people.ucsc.edu/~mclapham/earth125/data/venuscrater.csv

After you have done that, go to the tests & quizzes section of eCommons and work through the practice questions for in-class 3: dispersion.