As discussed in class, the Pearson correlation coefficient misses non-linear relationships and is also sensitive to outliers — the Spearman correlation can sometimes find correlations that Pearson misses. It is defined as the Pearson correlation of the rank order of the data. That means it also varies from -1 (perfectly anti-correlated) to $+1$ (perfectly correlated), with 0 meaning uncorrelated.

If the data has $x = [.6, .4, .2, .1, .5]$ then the ranks are $r = [5, 3, 2, 1, 4]$. For data $y = [403, 54, 7, 2, 148]$, the ranks $s = [5, 3, 2, 1, 4]$ are the same†, so the Spearman correlation is 1, whereas the Pearson is less than one. Both functions are available in scipy.stats (as pearsonr() and spearmanr()).

Defined as the Pearson correlation for the ranks, the Spearman correlation is written

$$\rho = \frac{\mathrm{Cov}[r,s]}{\sigma[r]\sigma[s]} \ , \tag{1}$$

where $\mathrm{Cov}[r,s] = E[(r-E[r])(s-E[s])]$ (generalizing the $\mathrm{Var}[x] = E[(x-E[x])^2]$, with $\mathrm{Cov}[x,x] = \mathrm{Var}[x]$). The formula for the Spearman correlation coefficient is given at http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient in terms of the difference $d_i = r_i - s_i$ between ranks, in this easily calculable form:

$$\rho = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)} \ . \tag{2}$$

It is straightforward to verify that (1) reduces to (2):

First note that the ranks r_i and s_i for n data points always run through the integers from 1 to n, in some orders. Thus

$$E[s] = E[r] = \frac{1}{n}\sum_i i = \frac{1}{n}\frac{n(n+1)}{2} = \frac{1}{2}\frac{(n+1)}{2} \ ,$$

$$E[s^2] = E[r^2] = \frac{1}{n}\sum_i i^2 = \frac{1}{n}\frac{1}{6}n(n+1)(2n+1) = \frac{1}{6}(n+1)(2n+1) \ ,$$

$$\text{and} \qquad \mathrm{Var}[s] = \mathrm{Var}[r] = E[r^2] - (E[r])^2 = \frac{1}{6}(n+1)(2n+1) - \frac{1}{4}(n+1)^2 = \frac{1}{12}(n^2-1) \ .$$

Next write the covariance in the form $\mathrm{Cov}[r,s] = E[rs] - E[r]E[s]$ (generalizing $\mathrm{Var}[x] = E[x^2] - (E[x])^2$, and derived in the same way). Then use $E[(r-s)^2] = E[r^2] - 2E[rs] + E[s^2]$ to write $E[rs] = E[r^2] - \frac{1}{2}E[(r-s)^2]$, together with $\sigma[r] = \sigma[s] = \sqrt{\mathrm{Var}[r]}$, to give:

$$\rho = \frac{\mathrm{Cov}[r,s]}{\sigma[r]\sigma[s]} = \frac{E[rs]-(E[r])^2}{\mathrm{Var}[r]} = \frac{\mathrm{Var}[r] - \frac{1}{2}E[(r-s)^2]}{\mathrm{Var}[r]} = 1 - \frac{\frac{1}{2}\frac{1}{n}\sum_{i=1}^n (r_i-s_i)^2}{\frac{1}{12}(n^2-1)} = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2-1)} \ ,$$

in agreement with (2).

† Actually the second was generated from the first by taking the integer part of $\exp(10x)$