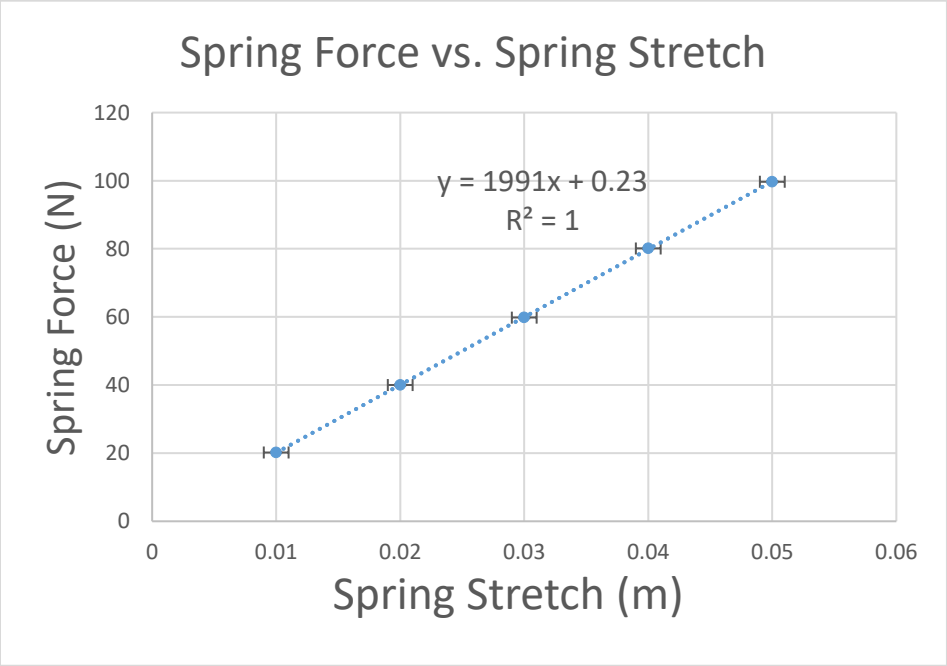
GRAPHING AND TRANSFORMING DATA

USING GRAPHING FOR DATA ANALYSIS

One of the most powerful tools we have for interpreting data is making plots. These plots both allow us to visualize the data and to use analytical tools to calculate experimental values for physical quantities. For example, imagine you have a spring: the more you stretch it, the more force it exerts. You collect the following data:

Spring stretch (m), ±0.001 m.	Spring force (N), ±0.2 N.
0.010	20.2
0.020	40.0
0.030	59.8
0.040	80.1
0.050	99.7

The quickest way to analyze the relationship between the stretch of the spring and the force is to plot the variables against each other. In this experiment, we control the stretch and that stretch determines the force exerted; this means the stretch is the *independent variable* and the force is the *dependent variable*. By convention, when we make a plot we plot the independent variable on the horizontal axis and the dependent variable on the vertical axis, so our plot looks like this:



Vertical error bars are smaller than the data points

Note the following details about the plot above, which you should seek to emulate in your own plots:

- The title has the format [dependent variable] vs. [independent variable].
- Each axis is labeled, with units indicated.
- The scale of the axes is appropriate so trends in the data are obvious.
- The data points have error bars, which are the size of the uncertainty.
- The fact that the vertical error bars are smaller than the data points, and therefore not visible, is noted.
- The equation of the best fit line is displayed.

- The R^2 value is displayed. (The R^2 value indicates how well the linear model fits the data. An R^2 of 1 indicates a perfect fit. In this case, the fit is not quite perfect, but the R^2 is so close to 1 that the computer has rounded it up to 1.)

Since the relationship between the variables does appear to be linear, and the value of the y-intercept is small, we can hypothesize the relationship between the stretch and the force has the mathematical form

$$F = ks$$

Where *F* is the spring force, *s* is the spring stretch, and *k* is a constant (called the spring constant).

When we made our plot and best fit line, the physical variables correspond to the variables in the standard linear equation $y = mx + b$, as shown:

$$\begin{array}{ccccccc} F & = & k & * & s & + & 0 \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ y & = & m & & x & + & b \end{array}$$

Thus, by putting the force on the vertical axis and the stretch on the horizontal axis, the spring constant *k* corresponds to the slope of the best fit line. So, looking at our plot, we see that the experimental best fit line has the equation

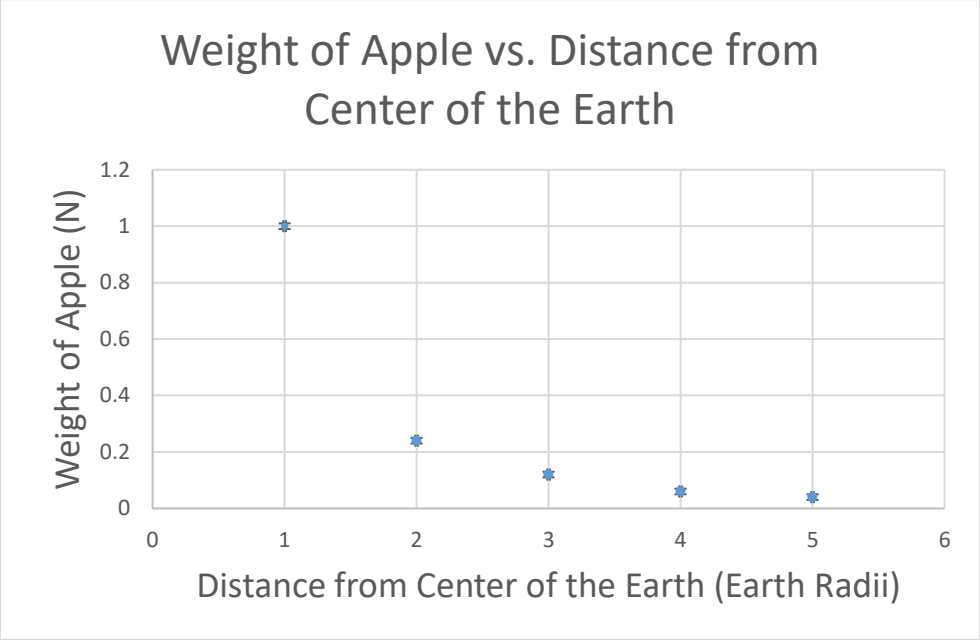
$$\begin{array}{ccccccc} F & = & 1991\left(\frac{N}{m}\right) & * & s & + & 0.23N \\ \nwarrow & & \uparrow & & \uparrow & & \uparrow \\ y & = & m & & x & + & b \end{array}$$

Where we have used the fact that F is on the vertical axis and s is on the horizontal axis. Thus, experimental value for the spring constant *k* is 1991 N/m. (The units are the ratio of the units on the vertical axis (N) to the units on the horizontal axis (m).)
In the above analysis, by using a plot we can see visually and through the R^2 value that our values are linearly correlated. Then, we used the slope of the best fit line to find an experimental value for a physical parameter, *k*.

TRANSFORMING DATA TO ACHIEVE LINEARITY
Now consider a more complicated situation. Say you are in a spacecraft moving away from the earth, and you periodically measure the weight (force due to gravity) of an apple in your spaceship. The data you collect is as follows:

Distance from the center of the Earth (Earth radii) ±0.01 R_E	Weight of the apple (N), ±0.01 N
1.00	1.00
2.00	0.24
3.00	0.12
4.00	0.06
5.00	0.04

If we plot the above data, it's immediately obvious the two variables are not linearly related:



Note that, since the data is clearly nonlinear, the best fit line has been removed.
Looking at the graph, it's clear that the weight decreases with increasing distance from the center of the Earth. But does the decrease follow a mathematical law? If so, which one? There are of course innumerable functions that decrease as their argument increases, but two common ones have the form

$$w = Ae^{-kd} \tag{1}$$

and

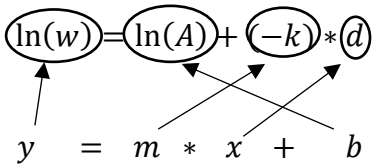
$$w = Bd^{-N} \tag{2}$$

Where *w* represents the weight of the apple, *d* represents its distance from the center of the Earth, and A, *k*, B, and N are all constants. Equation 1 describes what we call an *exponential decay*, while equation 2 describes a *power law*. We can test whether our data fit the above models graphically by transforming the data to achieve linearity.

Let's test the model described by equation 1 above using data transformation. To do, this, take the natural log of both sides of equation 1:

$$\ln(w) = \ln(Ae^{-kd}) = \ln(A) - k * d \tag{3}$$

This equation doesn't immediately look like the equation for a line, but if we assign functions as follows to y and x, it in fact has linear form!



That is, if the data does obey the exponential decay model (equation 1), then if we plot $\ln(w)$ on the vertical axis and d on the horizontal axis, then the points should fall in a straight line with slope equal to $-k$ and y intercept of $\ln(A)$.

In order to make this plot, we must *transform* our weight data by taking its natural logarithm. When we do this, we must also find the uncertainties in the transformed values for $\ln(w)$. These uncertainties can be found using worst case scenario analysis or propagation using calculus, as described in the "Error and Uncertainty" section. Here, we will use the calculus method:

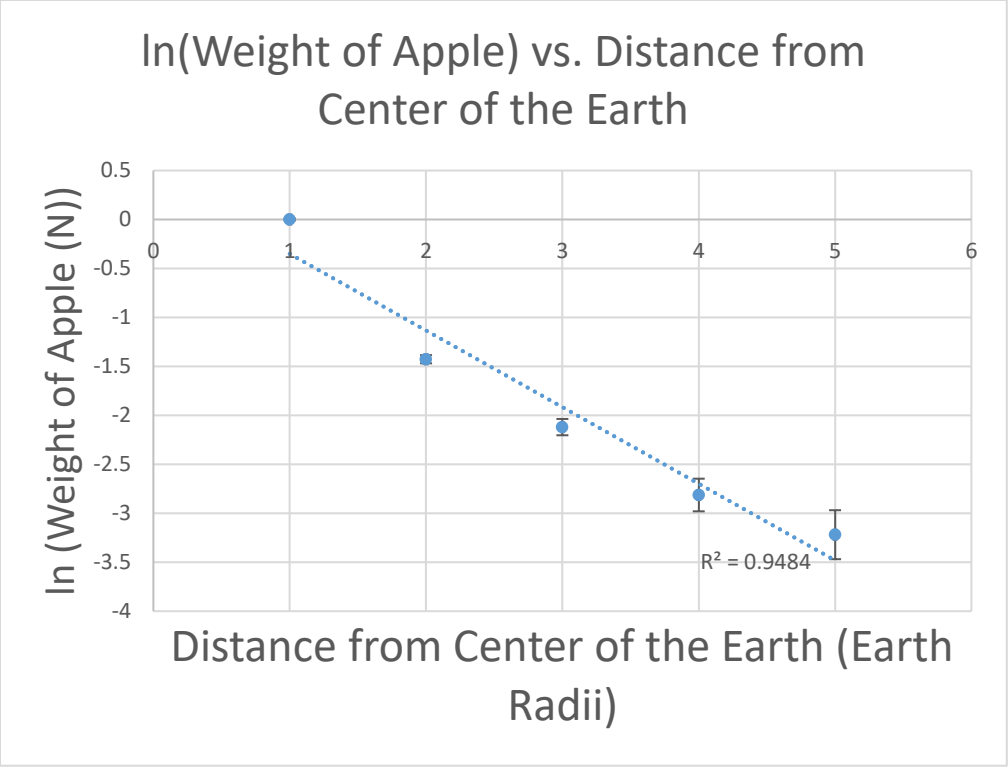
$$\delta\big(ln(w)\big) = \left(\frac{dln(w)}{dw}|_{w_o}\right)\delta w = \frac{1}{w_0}\delta w$$

This yields the following transformed data:

w (N)	$\ln(w)$	Uncertainty in $\ln(w)$ ($\delta\big(ln(w)\big)$)
1	0	0.01
0.24	-1.42712	0.041667
0.12	-2.12026	0.083333
0.06	-2.81341	0.166667
0.04	-3.21888	0.25

Note that there is nothing wrong with the negative values for $\ln(w)$. Also note that even though the uncertainty in w was the same for each data point, the uncertainty in $\ln(w)$ differs from point to point. This is because a small change in the value of a data point with a small w will lead to a large effect on the value of $\ln(w)$.

Plotting $\ln(w)$ against d gives the following plot:



Horizontal error bars are smaller than the data points

It's clear from a glance that the data is still not linear on this plot: there is a marked curvature, as you can tell by looking at the best fit line, which does not fit very well. The relatively low R^2 value confirms this. What this tells us is that the data does not fit the exponential decay model (equation 1).
To test the power law model (equation 2), we again take the natural log of both sides:

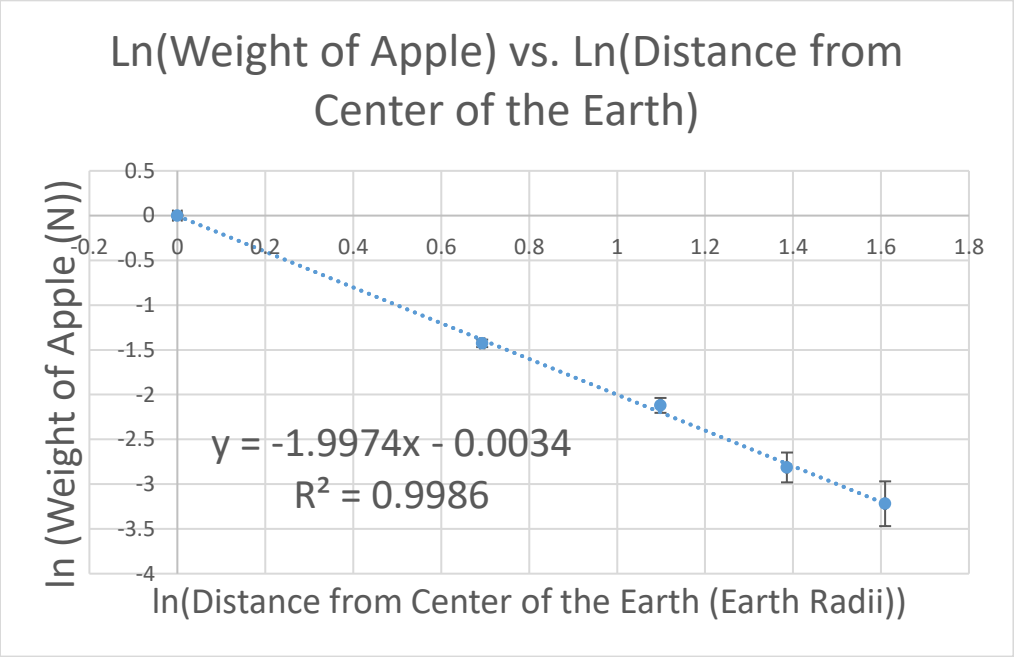
$$\ln(w) = \ln(Bd^{-N}) = \ln(B) - N * ln(d)$$

$$y = m * x + b$$

As you can see, if we plot $\ln(w)$ on the vertical axis against $\ln(d)$, and if the data fits the model, then we should get a straight line with slope N and y intercept $\ln(B)$.
We transform the distance data just as we did with the weight data, getting the following:

d (R_e)	$\ln(d)$	Uncertainty in $\ln(d)$ ($\delta(ln(d))$)
1	0	0.01
2	0.693147	0.005
3	1.098612	0.003333
4	1.386294	0.0025
5	1.609438	0.002

And then we plot $\ln(w)$ vs $\ln(d)$:



Horizontal error bars are smaller than the data points

As you can see, plotting $\ln(w)$ vs. $\ln(d)$ does cause the data to fall on a straight line. The R^2 value close to 1 confirms this. (The fact that the line passes near 0,0 is a coincidence that is not particularly meaningful). This means that power law model (equation 2) is a good fit for the data. Furthermore, the exponent N in equation 2 has an experimental value equal to the slope of the best fit line, -1.9974. (Using a Monte Carlo method, as described in the Error and Uncertainty section, we can estimate the uncertainty of the experimental value for N as ±0.1).

In summary, transforming the data to achieve linearity is a powerful way to check whether two variables are correlated according to a particular mathematical function and to determine experimental values for constants that appear in those functions.