2.1 In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?

In order for a student to really understand the concept that the author wishes you to demonstrate, I will also call the response variable the variable of interest. What is the variable that the researcher really interested in finding more about? This is the response variable.

(a) The amount of time spent studying for a statistics exam and the grade on the exam.

The variable of interest (response variable) is the grade on the exam- the final culmination of studying. We would like to associate the amount one studies (explanatory variable) and the final grade on the exam.

(b) The weight in kilograms and height in centimeters of a person.

Here it is difficult to say which is which. However, I will use current U.S. culture phenomena and say that the response variable (the variable of interest) is weight in kilograms. However, there is an association between height and weight, thus the explanatory variable is the height in centimeters.

(c) Inches of rain in the growing season and the yield of corn in bushels per acre.

I would have to guess that the variable of interest is the yield of corn in bushels per acre; after all, food is important to survival. Obviously, we need water for the crop to grow adequately, thus the measure of rainfall in inches is the explanatory variable.

(d) A student's scores on; the SAT math exam and the SAT verbal exam.

Both tests measure the ability of a student to perform well in a university setting. Thus, there is no clear choice for explanatory or response variable. So we would want to explore the relationship between the two variables.

(e) A family's income and the years of education their eldest child completes.
2.1 A study of reading ability in schoolchildren chose 60 fifth-grade children at random from a school. The researchers had the children's scores on an IQ test and on a test of reading ability. Figure 2.6 plots reading test score (response) against IQ score (explanatory).

(a) Explain why we should expect a positive association between IQ and reading score for children in the same grade. Does the scatterplot show a positive association?

Yes, we would expect that on average children that score a higher IQ score would also demonstrate a higher reading ability. It would not be perfect however, because reading is a learned skill, while an IQ tests intelligence potential.

The scatterplot indeed indicates a positive association, but not a strong one.

(b) A group of four points appear to be outliers. In what way do these children's IQ and reading scores deviate from the overall pattern?

These four values seem to violate the one assumption we are making about our model, which is a constant deviation about the center line.

The way researchers handle such outliers is by finding out what made those values deviate so much from the overall pattern. Often we find that these values belong to a different categorical group.

(c) Ignoring the outliers, is the association between IQ and reading scores roughly linear? Is it very strong? Explain your answers.

Suppose that we omit the outliers then the association is positive but not strong. Why? By strong we mean it adheres closely to a particular pattern, basically close to the blue line that depicts the center of all the normal distributions. But if we say this that means the standard deviation of the normal distributions shown in the “tube” will be small.

The graph shows that the spread about the center line is large which then means a not so strong association.
The main purpose of the study cited in Exercise 2.3 was to ask whether schoolchildren can estimate their own reading ability. The researchers had the children's scores on a test of reading ability. They asked each child to estimate his or her reading level, on a scale from 1 (low) to 5 (high). Figure 2.8 is a scatterplot of the children's estimates (response) against their reading scores (explanatory).

(a) What explains the "stair-step" pattern in the plot?

The variable in the response variable is discreet with only five possible values, \{1, 2, 3, 4, 5\}, but this is associated with a variable that is probably not discreet or if it is discreet, it appears to have a wider possibility of values \{0, 1, 2, \ldots, 100\}. So when we associate the two, we get a lot of repetition in the response variable, creating the stair step look.

(b) Is there an overall positive association between reading score and self-estimate?

The association is positive but extremely weak. How do I know it is positive? If I look at reading ability scores from 70 to 100 it seems to be associated with a self estimate of 5. While a low reading ability score is associated on average with lower scores. The problem is that there is too much overlap between the response variable for a given explanatory value.

(c) There is one, clear outlier. What is this child's self-estimated reading level? Does this appear to over- or underestimate the level as measured by the test?

The value of the reading level is about 12 and the self estimate is 4. This value is an obvious overestimate of the child’s perceived reading ability with respect to his peers. A more common value would have been 3 or less.

Often the percent of an animal species in the wild that survive to breed again is lower following a successful breeding season. This is part of nature's self-regulation, tending to keep population size stable. A study of merlins (small falcons) in northern Sweden observed the number of breeding pairs in an isolated area and the percent of males (banded for identification) who returned the next breeding season. Here are data for nine years:
(a) Why is the response variable the *percent* of males that return rather than, the *number* of males that return?

When one reads the question posed it is clear that the variable of interest is the percent of animals left to breed again.

(b) Make a scatterplot. To emphasize the pattern, also plot the mean response for years with 29 and 38 breeding pairs and draw lines connecting the mean responses for the six values of the explanatory variable.

(c) Describe the pattern. Do the data support the theory that a smaller percent of birds survive following a successful breeding season?

Yes, the pattern is however not linear. A linear model seems to work for a while, but when we look at the case when we had 38 breeding pairs, the percentage of returning males went up slightly from the value associated with 33 breeding pairs.
2.10 Many plants and animals have "biological clocks" that coordinate activities with the time of day. When researchers looked at the length of the biological cycles in the plant *Arabidopsis* by measuring leaf movements, they found that the length of the cycle is not always 24 hours. The researchers suspected that the plants adapt their clocks to their north-south position. Plants don't know geography, but they do respond to light, so the researchers looked at the relationship between the plants' cycle lengths and the length of the day on June 21 at their locations. The data file *ex02-010.dat* has data on cycle length and day length, both in hours, for 146 plants. Plot cycle length as the response variable against day length as the explanatory variable. Does there appear to be a positive association? Is it a strong association? Explain your answers.

There seems to be a positive association, but it is extremely weak to the point I am willing to say there is no association. What do I mean by this, are not the points adhering to a particular pattern (except for the outlier on top and on the right.)? What we would like to see is a strong difference in length of daylight for different values of cycle lengths but this does not seem to be the case. So for a cycle length of 26, what could be an associated length of daylight? By looking at the data it could range between 15 and 19 some hours, which could be said for almost all values of cycle length.

2.15 Table 2.3 (page 120) shows the progress of world record times (in seconds) for the 10,03 meter run up to mid-2004. Concentrate first on the women's world record) Make a scatterplot with year as the explanatory variable. Describe the pattern of improvement over time that your plot displays.