8.1: Introduction to Bivariate Quantitative Data

Learning Objectives

- Define multivariate data
- Introduce, differentiate, and identify associations and correlations
- Construct and utilize scatter plots to analyze bivariate data
- Distinguish between a linear and a nonlinear relationship
- Explore the differences between causation and correlation

Section 8.1 Excel File: (contains all of the data sets for this section)

Review and Preview

When we observe the world around us, there are a multitude of questions that could be asked about any single object, person, or event. So, when we study a population, it is possible to have many varied interests in the population or each member of the population. Take, for example, the assessment of the general health of an individual by a doctor. When we go to a healthcare provider for an assessment of general health, the doctor considers more than just our height. Multiple factors, like age, sex, height, weight, cholesterol, and glucose, to name a few, are considered. To get an accurate understanding of our general health, multiple variables must be considered together. We collect multivariate data when we are interested in a set of variables from each individual being studied. As this is just an introductory text, we will limit our considerations to bivariate quantitative data, meaning that we only consider analyses with only two quantitative variables of interest.

Bivariate Data: Types of Association and Models

Consider the ages at which married couples gave their wedding vows. For each married couple, we are interested in both the age of the bride and the age of groom. As such, we are considering bivariate data. We sampled 15 different married couples and tabulated the data below.

Table 8.1.1: Age of bride and groom on wedding day

Married Couple	Groom's Age (years)	Bride's Age (years)
1	20	21
2	26	20
3	32	34
4	30	30
5	21	22
6	29	28
7	26	25
8	34	34
9	29	28
10	55	50
11	30	26
12	43	39
13	30	29
14	24	22
15	20	19

Even with just 15 married couples, the data is difficult to digest and summarize. In taking the time to compare the ages of the bride and groom for each married couple, we come to the inclination that it is fairly common for the ages to be somewhat close together. This inclination aligns with the intuition built from previous experience. As such, we expect that there is a relationship between the age of the groom and the age of the bride. When the bride is young, we expect a young groom. When the groom is old, we expect an old bride. Given this expected pattern in the ages, we describe the association as positive. If we consider the ages of the bride and groom as quantitative variables, then as one variable increases, we expect the other variable to increase as well. Equivalently, if one variable decreases, we expect the other variable to decrease. As one variable changes in amount, the other variable changes in the same direction. As such, positive associations between variables are often referred to as direct relationships or positively correlated.

We built an intuition for this relationship with our previous experiences and by comparing the data line-by-line. Bivariate data will not always center around topics so commonplace and with so few observations. We will need to develop methodologies for facilitating such comparisons both visually and analytically. A common way to visualize bivariate quantitative data is by constructing a scatter plot. We are interested in establishing the relationship between the age of the bride and the age of the groom on their wedding day. The first column of the data simply labels the married couples and does not provide any information regarding the desired relationship. We focus on the second and third columns of our tabulated data. We can treat each married couple as a coordinate pair (Groom's Age, Bridge's Age) and then plot the 15 points on a coordinate plane to obtain a visualization of the relationship between the quantitative variables age of groom and age of bride. We have done so below.

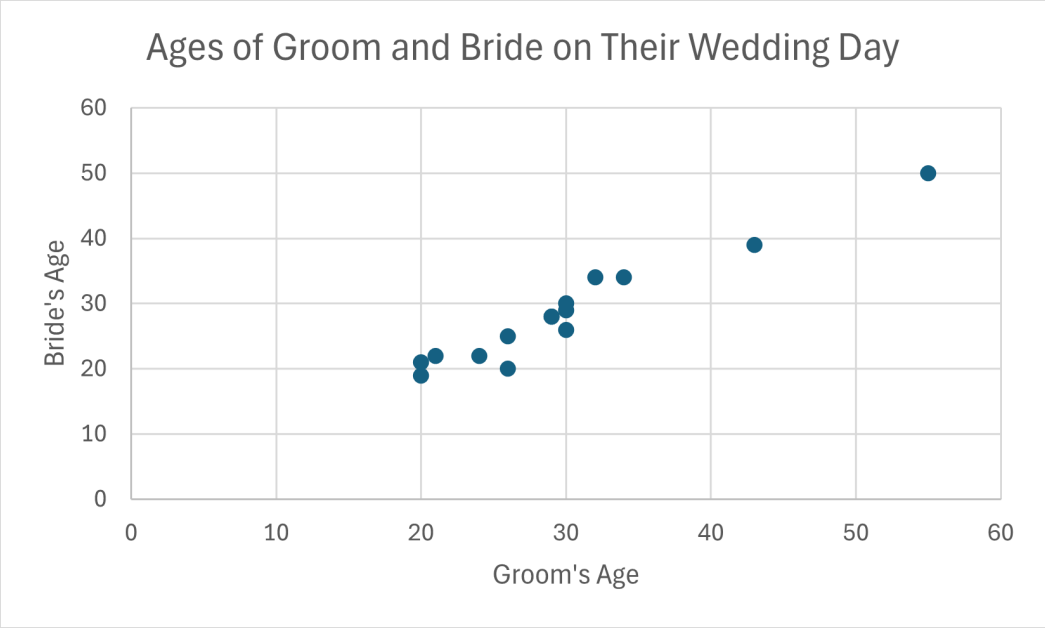

Figure 8.1.1: Scatter plot of groom's age vs bride's age on wedding day

The scatter plot quickly and easily confirms our initial intuitions. There is a relationship between the age of the groom and the age of the bride on their wedding day. As one variable increases, so does the other. Recall that our initial thoughts were that the ages were fairly close together and, as a result, we concluded that as one increased so would the other. We need to take a closer look to see how closely they are related. If the ages of the groom and bride are close together, we would expect that the data points would fall close to the line where the age of the groom equals the age of the bride. We have plotted such a line on the scatter plot below.

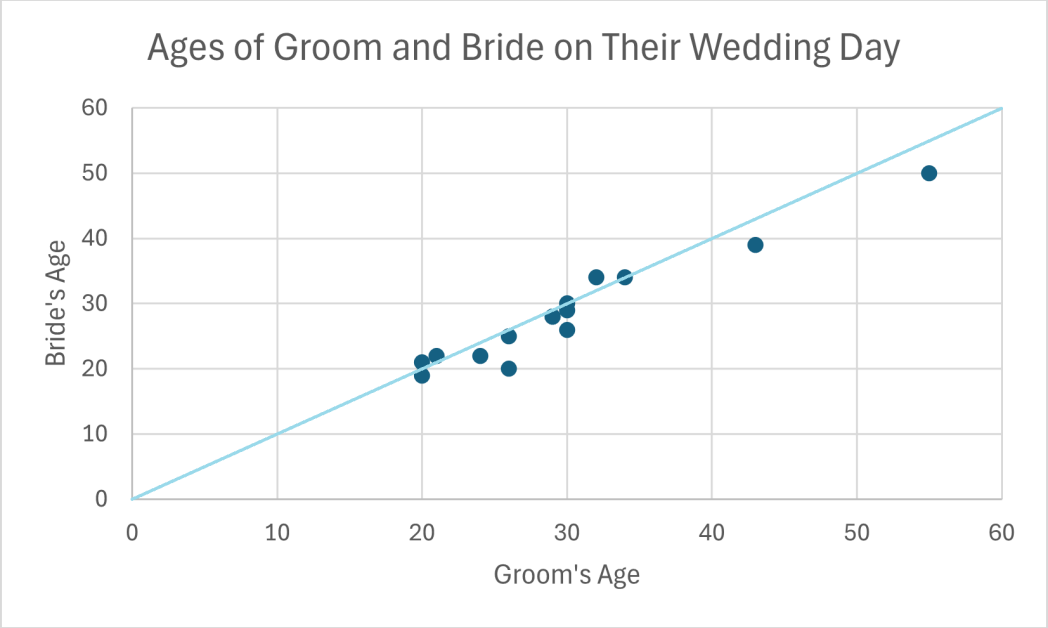

Figure 8.1.2: Scatter plot with the line $y = x$ of groom's age vs bride's age on wedding day

The line fits the data fairly well and gives credence to the idea that the relationship between the ages of the bride and groom can be modeled using a linear function. Recall that a linear function (straight line), is characterized by having a rate of change, a constant slope. The slope is the ratio of the vertical change to the horizontal change (rise over run). Since this is constant, we expect that the change in one variable is proportional to the change in the other variable meaning $\delta y = m\delta x$, where m is the slope of the line, regardless of the particular values of the variables. This seems to be true of the scatter plot at hand, but perhaps there is another line, with a different slope or y-intercept, that represents the data better. Consider the following scatter plot with the additional linear function represented using the blue dotted line.

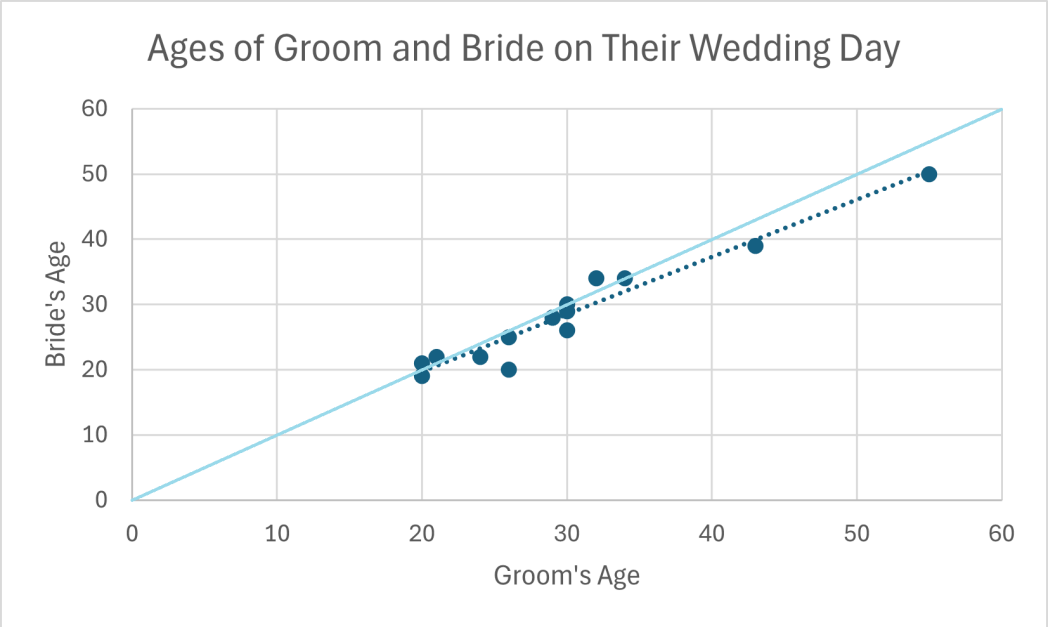

Figure 8.1.3: Scatter plot with two lines of groom's age vs bride's age on wedding day

Both of these lines appear to fit the data fairly well. We will eventually ask the question of how to decide which is better. As of right now, we conclude that when considering married couples, the two quantitative variables age of bride and age of groom appear to have a direct relationship and the relationship is likely to be modeled using a linear function.

Text Exercise 8.1.1

1. When we were considering the bivariate data related to the ages of the bride and groom on their wedding day, we asserted that the two quantitative variables displayed a positive correlation. It is possible to have a negative correlation (also referred to as a negative association or inverse relationship) between two quantitative variables. A negative correlation occurs when, as one variable increases, the other variable decreases, or equivalently: as one variable decreases, the other increases. Remaining within the context of marriage, identify bivariate data that would likely display a negative correlation. Explain your reasoning.

Answer

Given such a broad topic, the answers can vary quite tremendously. To check your solution, ensure that, for each member of a sample, two quantitative measurements are taken. This makes the quantitative data bivariate. Consider the relationship between the values of the two measurements. Will large values of one variable correspond to small values in the other? Does one decrease as the other increases or vice-versa?

A rather simple example relates the number of children in the household with the amount of one-on-one time a husband and wife get to spend with each other. As the number of children in the household increases, the responsibilities of the husband and wife as parents occupy a greater proportion of their time. Most couples realize the need to continually spend time together; so, we would not expect the amount of one-on-one time to diminish to 0, but nonetheless, real and felt decreases in one-on-one time is expected as the family grows in size.

2. Remaining in the context of marriage, give an example of bivariate data in which there is little to no correlation. That would mean if one variable is large, the other may be either small or large, and as one variable is small, the other may be small or large.

Answer

Consider the average height of the couple along with their annual income. In principle, these two variables have nothing to do with each other. While we could imagine reasons why these two quantities may correlate, there are many other factors, such as genetics, which determine height. We could reasonably expect to see short couples with low income, short couples with high income, tall couples with low income, and tall couples with high income.

Just like before, there are innumerably many possible answers to this question. If you picked two quantities where knowing what one is does not inform what the other will be, then your example likely works.

So far, we have described the concept of association within bivariate quantitative data as whether or not there is a relationship between the two variables. If there is an association, we are interested in what generally happens or is expected to happen to the value of one variable as the other variable is changed. If the directions of the changes match, the association is said to be positively correlated. If the directions of the changes are opposite, the association is said to be negatively correlate. It is possible to have relationships between two variables that have positive associations on certain intervals and negative associations on others. In such a case there is an association because there is a relationship, but there is no correlation because the relationship between the variables cannot be simply described as increasing or decreasing. Such relationships, however, would not be modeled well by a single linear function and, therefore, fall out of the scope of this course.

Text Exercise 8.1.2

The following questions center around the bivariate data related to diamonds, the gemstones that point to unwavering and lifelong love, according to popular culture. The quality of a diamond depends on various factors: the cut, color, clarity, and weight. The general shape or cut of the diamond plays an important role, but when the quality of the cut is referenced, the focus is on the proportions of the cut diamond as they relate to reflecting light back through the diamond. The color points to the general hue of the diamond; while, the clarity points to the presence of internal or surface defects on the diamond. The weight of the diamond is typically measured in carats. The prices of various round diamonds with super ideal cuts, flawless clarity, and icy white hue were observed along with their weights from several top national retailers (Brilliant Earth and Blue Nile). The diamonds were then ordered to produce the following table.

Table 8.1.2: Weight (carats) and price ($) of diamonds

Diamond	Weight (carats)	Price (US dollars)
1	0.37	1, 160
2	0.52	2, 430
3	0.63	3, 860
4	0.63	3, 430
5	0.68	3, 900
6	0.7	4, 050
7	0.77	5, 020
8	0.83	6, 830
9	0.95	9, 560
10	1.18	12, 830
11	1.24	13, 610
12	1.3	14, 830
13	1.3	14, 890
14	1.37	15, 990
15	1.43	18, 890
16	1.61	24, 000
17	1.67	23, 870
18	2	39, 300
19	2.02	38, 580
20	2.22	54, 360
21	2.23	63, 820
22	2.34	60, 160
23	2.39	49, 130
24	2.52	65, 600
25	2.56	62, 130
26	3.5	144, 120
27	3.56	192, 710
28	3.88	157, 280
29	4.42	236, 360
30	5.02	322, 260
31	5.06	374, 870
32	5.8	359, 600
33	6.04	398, 190
34	7.13	543, 230

Diamond	Weight (carats)	Price (US dollars)
35	7.32	530, 320

1. What sort of association do you expect with this bivariate data? Explain.

Answer

As diamonds increase in size and weight, the rarity of the diamonds increases. And as rarity increases, the price increases. We expect a positive correlation.

2. Construct a scatter plot of this bivariate data to check your intuition from the previous part of this exercise. Be sure to label the graph and both axes.

Answer

Using the data tabulated in the provide Excel file and under the guidance of the Excel guide, we construct the following scatter plot.

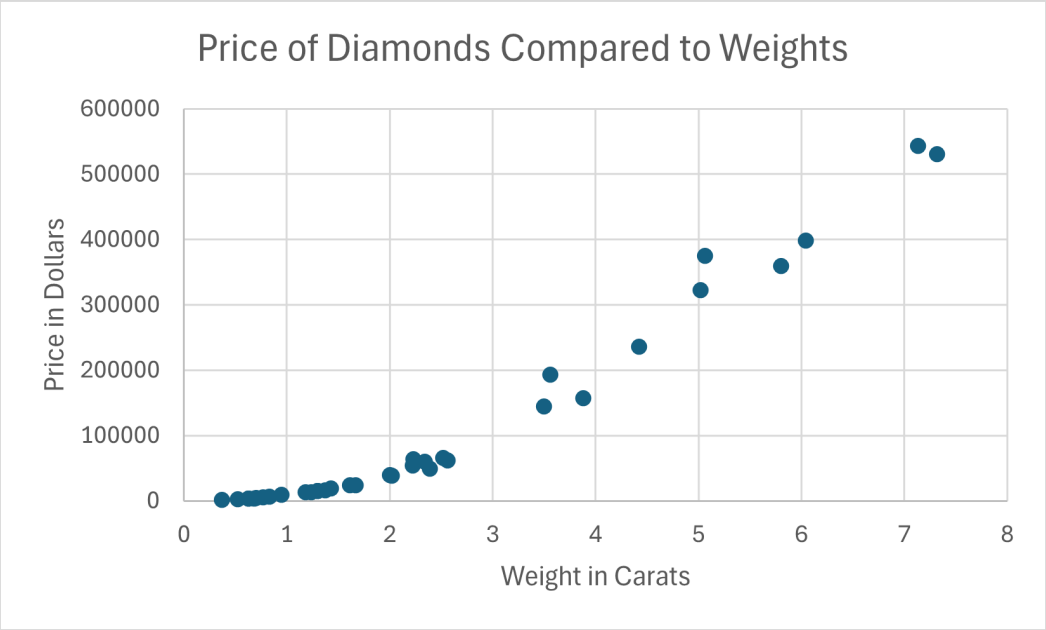

Figure 8.1.4 Scatter plot of weight (carats) vs price ($) of diamonds

As predicted, we see that as the weight of diamond increases, the price of the diamond also increases. We have confirmed that the correlation is positive.

3. If we were trying to model the association of this bivariate data with a function, would a linear function fit the data well?

Answer

In looking at the scatter plot, it appears that the rate at which the price is increasing as the weight increases is not constant. The rate of change when the weight is between 0 and 2 carats is perhaps fairly steady but is different from the rate of change when the weight is between 3 and 8 carats. Consider the slopes of the two lines drawn below.

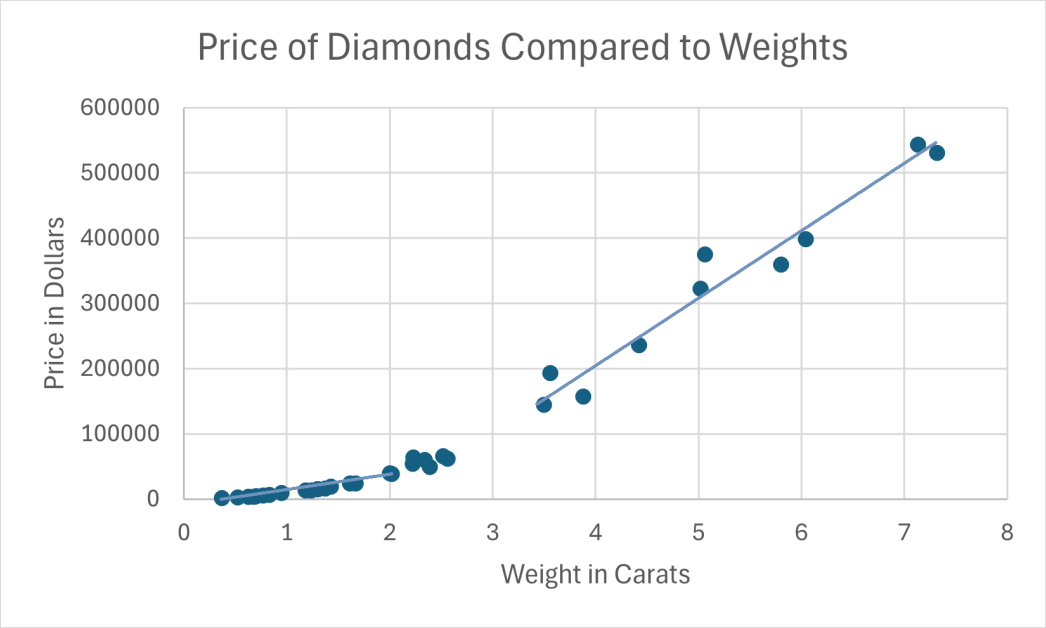

Figure 8.1.5 Scatter plot of weight (carats) vs price ($) of diamonds comparing slope

The slope of the line of the left is positive but significantly less steep than the slope of the line on the right. This leads us to conclude that a linear function is probably not the best model for the bivariate data.

We will limit our discussion in the text to using linear functions, but there are other types of functions that can be used as well. We modeled this same data using a power function to produce the following scatter plot and model below. Interested readers are encouraged seek more advanced statistical texts to address such content.

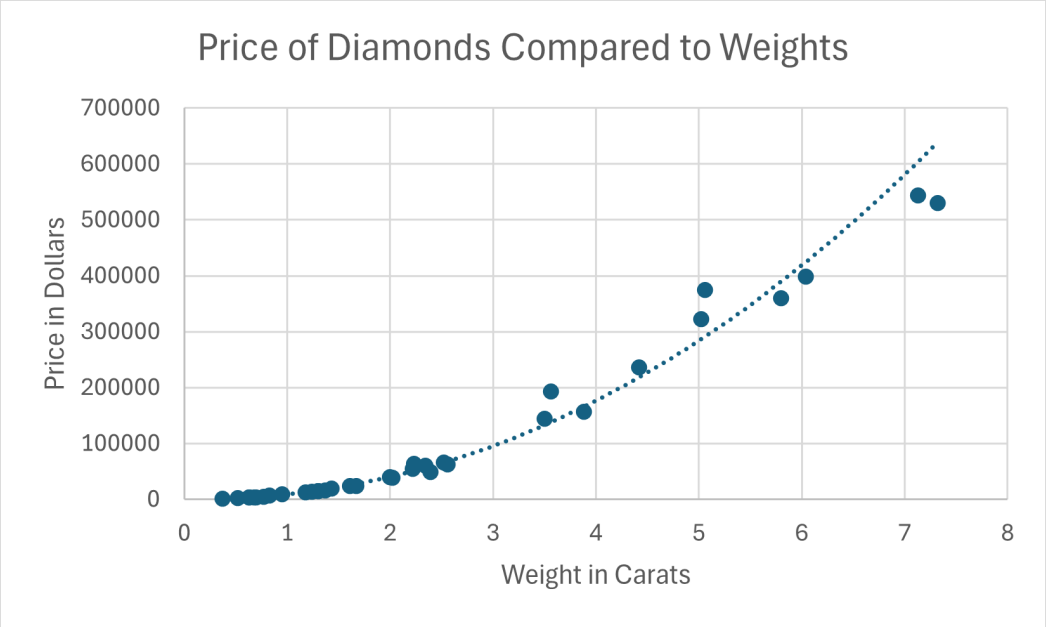

Figure 8.1.6 Scatter plot of weight (carats) vs price ($) of diamonds power function

Bivariate Data: Strength of Association

There are many instances when the relationship is not as clear as we have seen so far; the association is not as strong. Indeed there are cases, where there is no association. Consider the following scatter plot which relays the measurements the age of the bride at

the time of marriage with the number of children the couple has over the course of their marriage from a random sample of married couples.

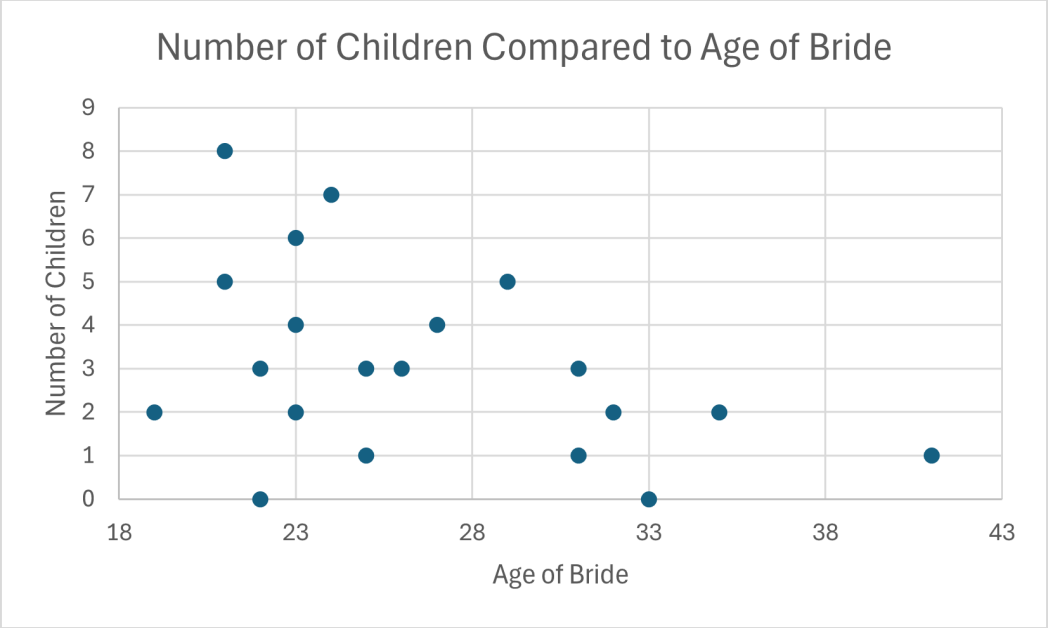

Figure 8.1.7: Scatter plot of age of bride vs number of children

Data in Tabulated Form

Table 8.1.3 Table of age of bride vs number of children

Married Couple	Bride's Age (years)	Number of Children
1	19	2
2	21	8
3	21	5
4	22	0
5	22	3
6	23	6
7	23	4
8	23	2
9	24	7
10	25	1
11	25	3
12	26	3
13	27	4
14	29	5
15	31	3
16	31	1
17	32	2

Married Couple	Bride's Age (years)	Number of Children
18	33	0
19	35	2
20	41	1

The scatter plot shows a general decline in the number of children as the age of the bride increases. This has a quite natural explanation from a basic understanding of human biology. This does not, however, explain the fact that our data looks as it does. Up until now, the coordinate pairs on our scatter plot looked somewhat like closely packed paths through the coordinate plane. Now our scatter plot looks similar to a shaded triangle. There are couples with 0, 1, and 2 children all throughout the presented age range. There are many factors that contribute to the number of children born through a marriage. There is a negative association between the age of the bride and the number of children, but the strength of the association is not as strong as the other examples we have seen due to a number of other factors. The closer the points are clustered to form a tightly packed path, the stronger the relationship; the less densely the data is packed along a path, the weaker the relationship. If there is no path, there is no relationship. We would see a similar sort of loss of strength if in our search for diamond prices and weights we did not first narrow our search to diamonds of the same cut, color, and clarity.

Text Exercise 8.1.3

For each scatter plot, indicate the type and strength of the association between the two variables and if a linear function would model the relationship well.

1.

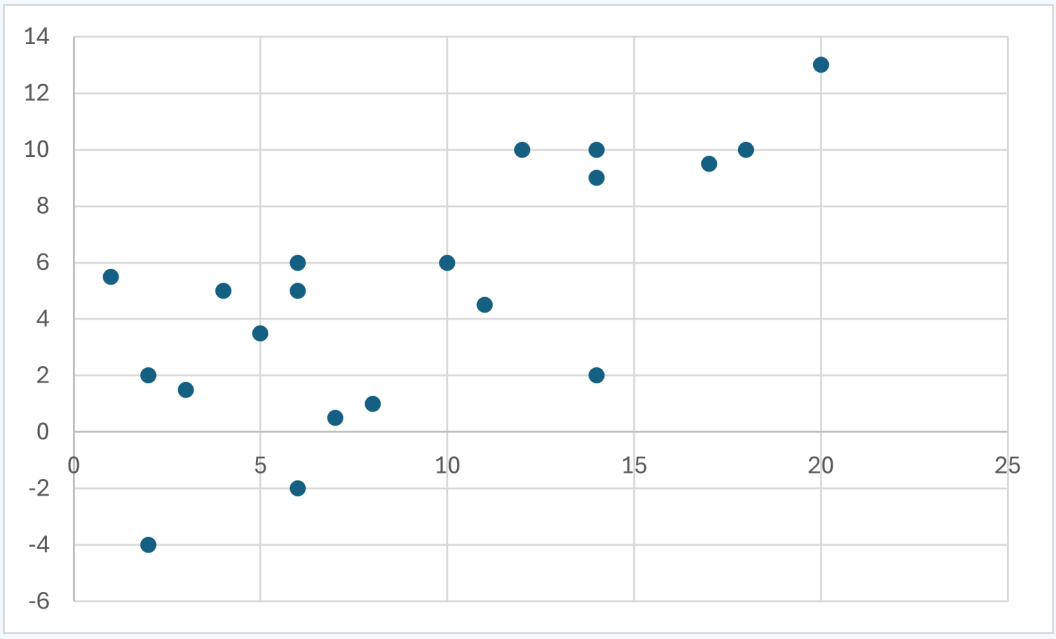

Figure 8.1.8: Scatter plot

Answer

The scatter plot indicates a positive association between the two variables, but the points on the scatter plot are not tightly packed together vertically. This leads us to say that the association is not as strong as some of the other associations seen in this section. It does look like a linear function could be used to model the data.

2.

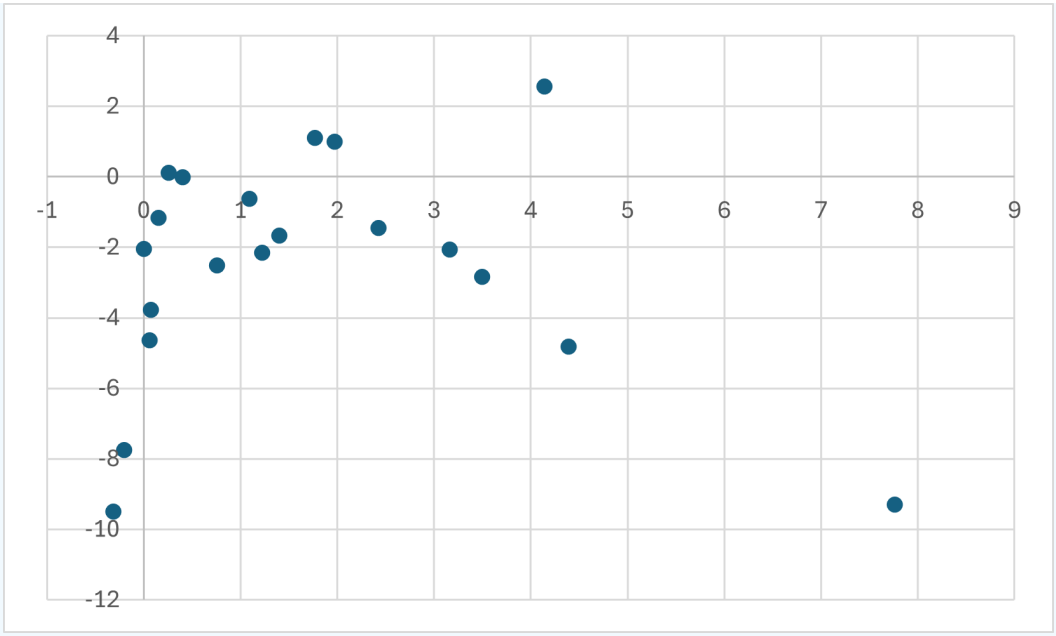

Figure 8.1.9: Scatter plot

Answer

The scatter plot indicates that the variables show a positive association when the horizontal variable is less than about 2 but a negative association after that. Given that the relationship does not appear to be a simply positive or negative association, we would say there is no correlation. The scatter plot does indicate that an association between the variables is present, possibly to a stronger degree than the previous part of this text exercise. Modeling this data would be best be done with a function that is increasing and then decreasing. As such, a linear function would not model the data well.

3.

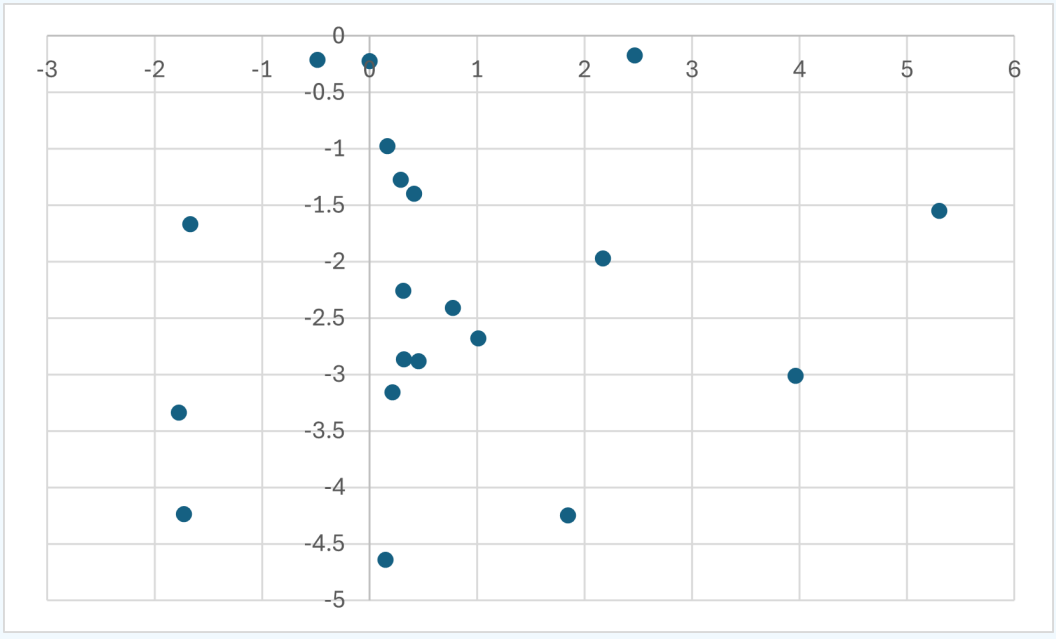

Figure 8.1.10: Scatter plot

Answer

This scatter plot does not seem to admit to any particular association. The data seems scattered fairly randomly over the coordinate plane. It is possible to envision a steep line with negative slope through the origin fitting the data okay, but it

leaves a lot of points farther away from the line, and it seems just as easy to envision a shallow line with positive slope through $(0, -2.75)$ fitting the data okay, but again leaving a lot of points far away from the model. As such, we say that there is likely no association, or at best, a very weak association between the variables. When there is no association or a very weak association, no function will serve as a suitable model.

4.

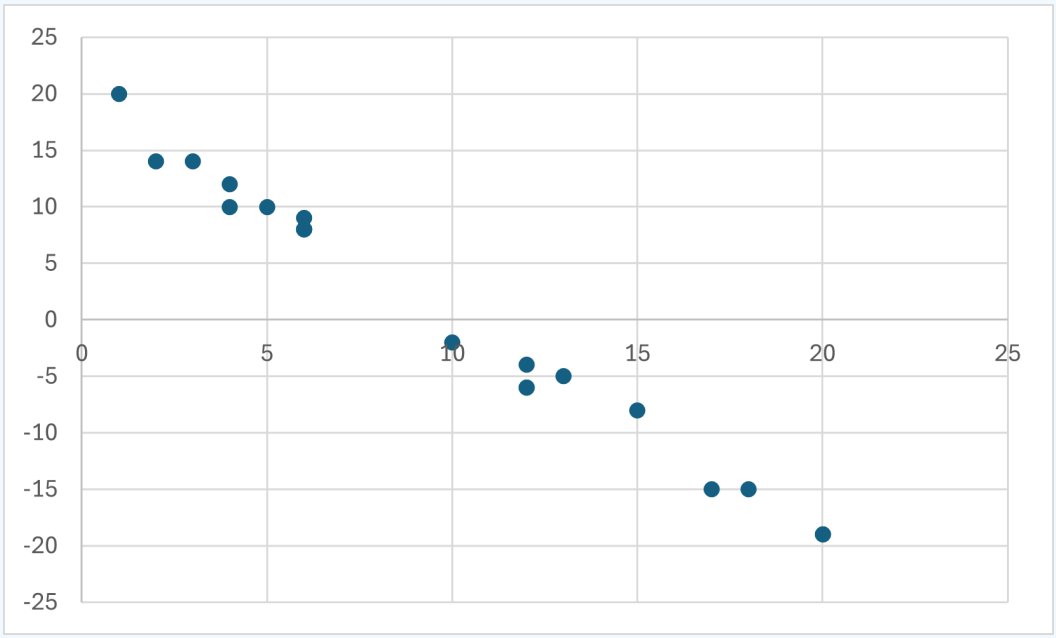

Figure 8.1.11: Scatter plot

Answer

This scatter plot indicates a negative correlation fairly clearly. The association appears to be fairly strong with a linear function being well suited to the model the relationship between the variables.

When assessing associations visually, care must be given. We assess the strength of the association based on how tightly packed the paths are formed by the data. This visual assessment can be greatly distorted by the scale used on the scatter plot. Consider the following scatter plot constructed from the data sets that produced the scatter plots in the first (blue), second (green), and fourth (orange) parts of the previous text exercise.

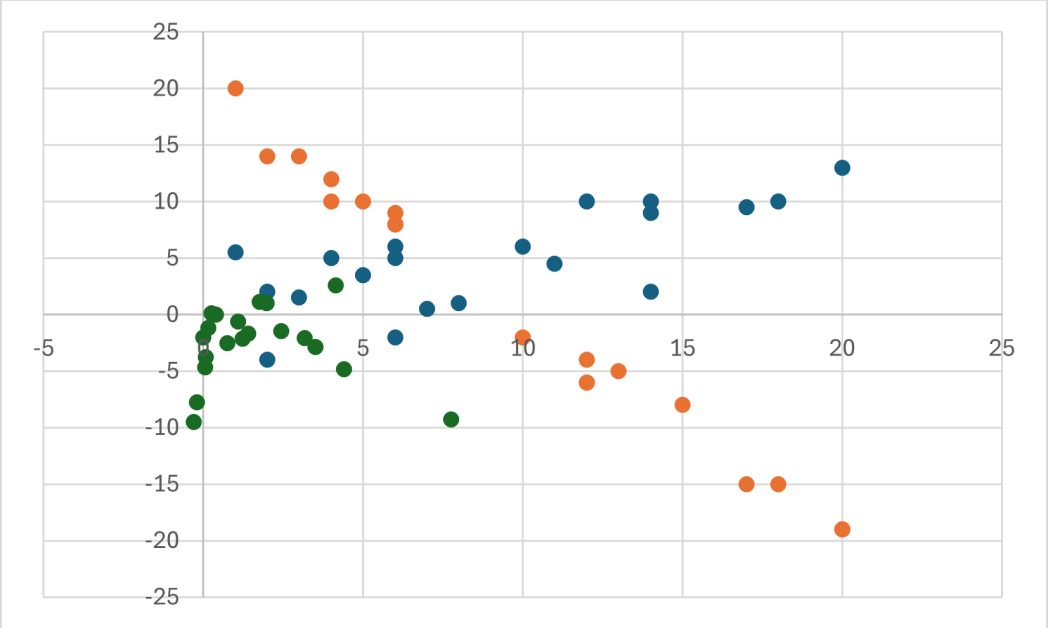

Figure 8.1.12: Comparing scatter plots on same scale

Our understanding of the data from the first and fourth parts of the previous exercise remain the same and in someways are strengthened. The blue data from the first part is positively associated and appears to follow a linear model. The orange data from the fourth part is negatively associated and a linear function still appears to model the data well. Having the two plotted on the same coordinate plane, we can feel confident that the association of the orange data is stronger than that of the blue data because the points form a path that is more densely packed than the blue path. Our confidence in the conclusions relating to the second part of the previous exercise, the green data, may be slightly shaken. At this scale, it seems a little more reasonable to conclude that the green data might be modeled using a linear function with a negative slope. The appearance of a possible linear relationship becomes even more pronounced when the scale is altered again in the following scatter plot.

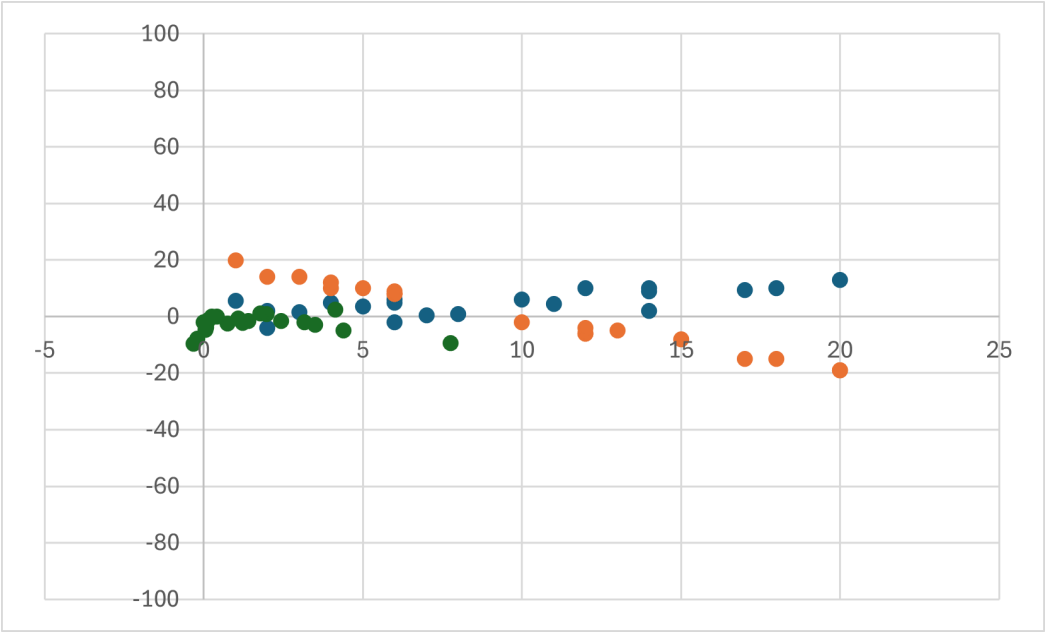

Figure 8.1.13: Comparing scatter plots on same larger scale

Hopefully, we recognize the egregious nature of this last scatter plot. There is no need for such a large scaling of the vertical axis. In the first scatter plot with the three data sets, the scale needed to incorporate all the data present and, therefore, could not be scaled in precisely the same way as the scatter plots were scaled in the original text exercise. Sometimes, there are legitimate

reasons for plotting multiple variables on a single coordinate plane, but when we do so, we must exercise caution. We want to develop analytical methods of assessing the association between variables that does not depend on the scaling of the values either graphically presented or based on the units used in the calculation. We do so later in this chapter.

Correlation and Causation

In this section we address the common adage that "correlation does not imply causation." If quantity A somehow causes or creates quantity B, then one would expect to see an association between the two variables. What the adage is saying, is that the converse is not true; the existence of a correlation could have many explanations other than cause and effect. In reality, the principle applies to more than just correlation; it applies to any association. Perhaps the adage is assembled as an appeal to the allure of alliteration. Setting phrasing aside, the adage warns us that when studying bivariate data, we must carefully discern the conclusions that we draw from the presence of associations.

While the name may be unfamiliar, the concept of causation should be familiar enough; think of it as referencing the ideas of cause and effect. Suppose a pianist begins to play Pachelbel's Canon in D Major at the start of the entrance procession at a wedding. With each press of a key, a hammer strikes a taut string which subsequently reverberates the desired note throughout the space. The pianist, the cause, makes the music, the effect. There is such a possibility when considering quantitative variables as well.

For various reasons, many engaged couples focus on their physiques as part of the preparation for their upcoming wedding. In many cases, this preparation centers around losing weight with a focus on caloric intake. In order to lose weight, one must consistently have a daily caloric deficit. If one wants to gain weight, one must consistently have a daily caloric surplus. Each can be attained through various combinations of food consumption, hydration, and exercise. When in a state of caloric deficiency, the effect is weight loss. We say there is a causal relationship between the quantitative variables caloric deficit and weight loss. When there is a relationship, there is an association. In other words, when there is causation there is association. If we were to collect data on this topic, we would see that the larger the deficit the more weight is lost. We would say that caloric deficit and weight loss are positively correlated. In this case, the positive causal relationship implies a positive correlation.

This is not the typical setup when we are studying the world around us. In general, we do not start with a causal relationship. We start with two variables of interest, collect data, and consider a scatter plot to see if there is an association. Just because there is an association does not mean that the relationship is causal. For example, suppose the happy newly weds are cataloging their gifts so that they may send thank you letters. In doing so, they notice that the people who traveled very far to attend the wedding gave gifts which tended to be more expensive. Their most expensive gifts, in fact, were given by the people who traveled the farthest. Similarly, people who did not travel very far tended to give cheaper gifts. In short, they have noticed an association between two quantitative variables: price of the gift and the distance traveled to the wedding. Is it reasonable to assert that there is a cause and effect relationship? The couple may speculate that when someone puts in a large amount of effort to attend, they are psychologically predisposed to a larger monetary loss and therefore are inclined to spend more than someone who put in little effort to travel. This hypothesis asserts that the travel distance is causing the price of the gift. We wish to emphasize that such a conclusion is premature. While it may be true, there are other explanations of the correlation. Of their loved ones who live far away, perhaps only those with a lot of money were able to attend; those people would be able to afford more expensive gifts. This would suggest that wealth is causing both the willingness to travel long distances and buy expensive gifts. Alternatively, perhaps those who are willing to travel long distances are those who have very close relationships with the couple and thus are willing to spend more money on gifts. Or, it could be a total coincidence; perhaps other weddings did not observe a correlation between these quantities. More work is needed to ascertain which of these explanations, if any, are correct. Concluding any cause and effect based solely on the observation of correlation is erroneous.

In general, if quantity A correlates with quantity B, there are many explanations for that correlation. It could be the case that A causes B. It could be the case that B causes A. It could be the case that a totally separate quantity, C, is causing both A and B; this latter case is referred to as **common response**. Or, as tends to happen with small data sets, the correlation could just be total coincidence: a pattern emerging from random chance and is falsified upon collecting more data. We call these **spurious correlations**. Establishing correlation is comparatively easy, but establishing cause and effect is quite difficult; the latter requires controlled experimentation accompanied with explanatory power. Throughout this chapter, we will explore how to measure correlation, but we will not be making any assertions regarding causation.
