3.1: Types of Visualizations

Suppose we want to study the composition of 1.69-oz (47.9-g) packages of plain M&Ms. We obtain 30 bags of M&Ms (ten from each of three stores) and remove the M&Ms from each bag one-by-one, recording the number of blue, brown, green, orange, red, and yellow M&Ms. We also record the number of yellow M&Ms in the first five candies drawn from each bag, and record the actual net weight of the M&Ms in each bag. Table 3.1.1 summarizes the data collected on these samples. The bag id identifies the order in which the bags were opened and analyzed.

Table 3.1.1. Analysis of Plain M&Ms in 47.9 g Bags.

bag	store	blue	brown	green	orange	red	yellow	yellow_first_five	net_weight
1	CVS	3	18	1	5	7	23	2	49.287
2	CVS	3	14	9	7	8	15	0	48.870
3	Target	4	14	5	10	10	16	1	51.250
4	Kroger	3	13	5	4	15	16	0	48.692
5	Kroger	3	16	5	7	8	18	1	48.777
6	Kroger	2	12	6	10	17	7	1	46.405
7	CVS	13	11	2	8	6	17	1	49.693
8	CVS	13	12	7	10	7	8	2	49.391
9	Kroger	6	17	5	4	8	16	1	48.196
10	Kroger	8	13	2	5	10	17	1	47.326
11	Target	9	20	1	4	12	13	3	50.974
12	Target	11	12	0	8	4	23	0	50.081
13	CVS	3	15	4	6	14	13	2	47.841
14	Kroger	4	17	5	6	14	10	2	48.377
15	Kroger	9	13	3	8	14	8	0	47.004
16	CVS	8	15	1	10	9	15	1	50.037
17	CVS	10	11	5	10	7	13	2	48.599
18	Kroger	1	17	6	7	11	14	1	48.625
19	Target	7	17	2	8	4	18	1	48.395
20	Kroger	9	13	1	8	7	22	1	51.730
21	Target	7	17	0	15	4	15	3	50.405
22	CVS	12	14	4	11	9	5	2	47.305
23	Target	9	19	0	5	12	12	0	49.477
24	Target	5	13	3	4	15	16	0	48.027
25	CVS	7	13	0	4	15	16	2	48.212
26	Target	6	15	1	13	10	14	1	51.682
27	CVS	5	17	6	4	8	19	1	50.802
28	Kroger	1	21	6	5	10	14	0	49.055
29	Target	4	12	6	5	13	14	2	46.577
30	Target	15	8	9	6	10	8	1	48.317

Having collected our data, we next examine it for possible problems, such as missing values (Did we forget to record the number of brown M&Ms in any of our samples?), for errors introduced when we recorded the data (Is the decimal point recorded incorrectly for any of the net weights?), or for unusual results (Is it really the case that this bag has only yellow M&M?). We also examine our data to identify interesting observations that we may wish to explore (It appears that most net weights are greater than the net weight listed on the individual packages. Why might this be? Is the difference significant?) When our data set is small we usually can identify possible problems and interesting observations without much difficulty; however, for a large data set, this becomes a challenge. Instead of trying to examine individual values, we can look at our results visually. While it may be difficult to find a single, odd data point when we have to individually review 1000 samples, it often jumps out when we look at the data using one or more of the approaches we will explore in this chapter.

Dot Plots

A dot plot displays data for one variable, with each sample's value plotted on the *x*-axis. The individual points are organized along the *y*-axis with the first sample at the bottom and the last sample at the top. Figure 3.1.1 shows a dot plot for the number of brown M&Ms in the 30 bags of M&Ms from Table 3.1.1. The distribution of

points appears random as there is no correlation between the sample id and the number of brown M&Ms. We would be surprised if we discovered that the points were arranged from the lower-left to the upper-right as this implies that the order in which we open the bags determines whether they have many or a few brown M&Ms.

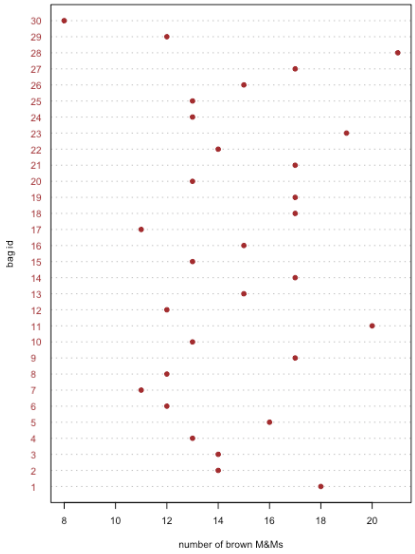

Figure 3.1.1: Dot plot for the brown M&Ms in each of the 30 bags included in Table 3.1.1.

Stripcharts

A dot plot provides a quick way to give us confidence that our data are free from unusual patterns, but at the cost of space because we use the *y*-axis to include the sample id as a variable. A stripchart uses the same *x*-axis as a dot plot, but does not use the *y*-axis to distinguish between samples. Because all samples with the same number of brown M&Ms will appear in the same place—making it impossible to distinguish them from each other—we stack the points vertically to spread them out, as shown in Figure 3.1.2.

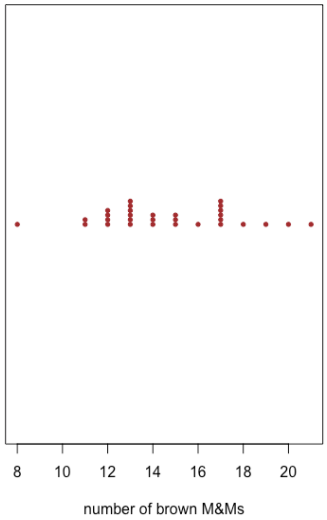

Figure 3.1.2: Stripchart for the brown M&Ms in each of the 30 bags included in Table 3.1.1.

Both the dot plot in Figure 3.1.1 and the stripchart in Figure 3.1.2 suggest that there is a smaller density of points at the lower limit and the upper limit of our results. We see, for example, that there is just one bag each with 8, 16, 18, 19, 20, and 21 brown M&Ms, but there are six bags each with 13 and 17 brown M&Ms.

Because a stripchart does not use the *y*-axis to provide meaningful categorical information, we can easily display several stripcharts at once. Figure 3.1.3 shows this for the data in Table 3.1.1. Instead of stacking the individual points, we jitter them by applying a small, random offset to each point. Among the things we learn from this stripchart are that only brown and yellow M&Ms have counts of greater than 20 and that only blue and green M&Ms have counts of three or fewer M&Ms.

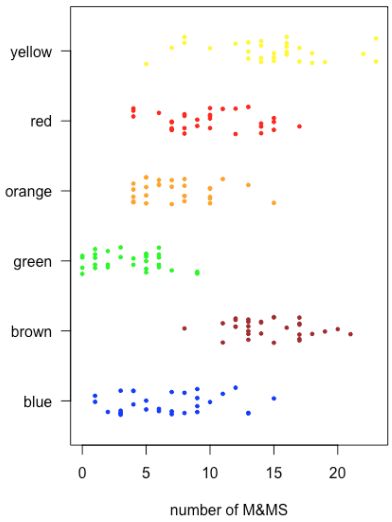

Figure 3.1.3: Stripcharts for each color of M&Ms in each of the 30 bags included in Table 3.1.1.

Box and Whisker Plots

The stripchart in Figure 3.1.3 is easy for us to examine because the number of samples, 30 bags, and the number of M&Ms per bag is sufficiently small that we can see the individual points. As the density of points becomes greater, a stripchart becomes less useful. A box and whisker plot provides a similar view but focuses on the data in terms of the range of values that encompass the middle 50% of the data.

Figure 3.1.4 shows the box and whisker plot for brown M&Ms using the data in Table 3.1.1. The 30 individual samples are superimposed as a stripchart. The central box divides the *x*-axis into three regions: bags with fewer than 13 brown M&Ms (seven samples), bags with between 13 and 17 brown M&Ms (19 samples), and bags with more than 17 brown M&Ms (four samples). The box's limits are set so that it includes at least the middle 50% of our data. In this case, the box contains 19 of the 30 samples (63%) of the bags, because moving either end of the box toward the middle results in a box that includes less than 50% of the samples. The difference between the box's upper limit (19) and its lower limit (13) is called the interquartile range (IQR). The thick line in the box is the median, or middle value (more on this and the IQR in the next chapter). The dashed lines at either end of the box are called whiskers, and they extend to the largest or the smallest result that is within $\pm 1.5 \times \text{IQR}$ of the box's right or left edge, respectively.

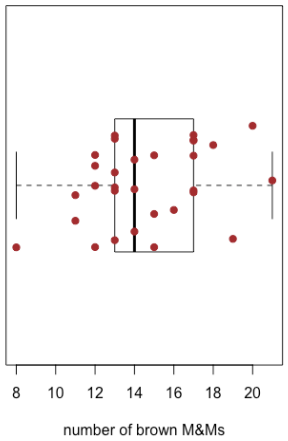

Figure 3.1.4: Box-and-whisker plot for the brown M&Ms in each of the 30 bags included in Table 3.1.1 showing individual samples as a jittered stripchart.

Because a box and whisker plot does not use the *y*-axis to provide meaningful categorical information, we can easily display several plots in the same frame. Figure 3.1.5 shows this for the data in Table 3.1.1. Note that when a value falls outside of a whisker, as is the case here for yellow M&Ms, it is flagged by displaying it as an open circle.

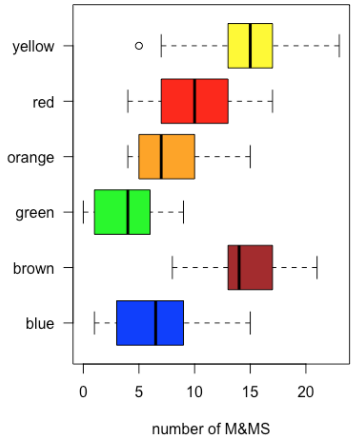

Figure 3.1.5: Box-and-whisker plots for each of the 30 bags included in Table 3.1.1 organized by color.

One use of a box and whisker plot is to examine the distribution of the individual samples, particularly with respect to symmetry. With the exception of the single sample that falls outside of the whiskers, the distribution of yellow M&Ms appears symmetrical: the median is near the center of the box and the whiskers extend equally in both directions. The distribution of the orange M&Ms is asymmetrical: half of the samples have 4–7 M&Ms (just four possible outcomes) and half have 7–15 M&Ms (nine possible outcomes), suggesting that the distribution is skewed toward higher numbers of orange M&Ms (see Chapter 5 for more information about the distribution of samples).

Figure 3.1.6 shows box-and-whisker plots for yellow M&Ms grouped according to the store where the bags of M&Ms were purchased. Although the box and whisker plots are quite different in terms of the relative sizes of the boxes and the relative length of the whiskers, the dot plots suggest that the distribution of the underlying data is relatively similar in that most bags contain 12–18 yellow M&Ms and just a few bags deviate from these limits. These observations are reassuring because we do not expect the choice of store to affect the composition of bags of M&Ms. If we saw evidence that the choice of store affected our results, then we would look more closely at the bags themselves for evidence of a poorly controlled variable, such as type (Did we accidentally purchase bags of peanut butter M&Ms from one store?) or the product's lot number (Did the manufacturer change the composition of colors between lots?).

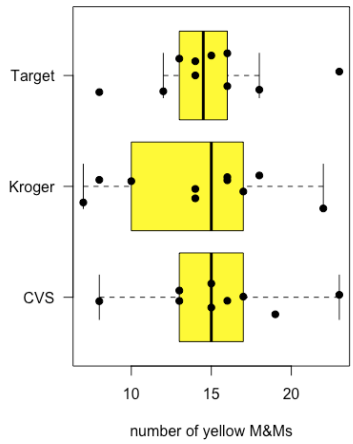

Figure 3.1.6: Box-and-whisker plots for yellow M&Ms for each of the 30 bag in Table 3.1.1 organized by the store where the bags were purchased.

Bar Plots

Although a dot plot, a stripchart and a box-and-whisker plot provide some qualitative evidence of how a variable's values are distributed—we will have more to say about the distribution of data in Chapter 5—they are less useful when we need a more quantitative picture of the distribution. For this we can use a bar plot that displays a count of each discrete outcome. Figure 3.1.7 shows bar plots for orange and for yellow M&Ms using the data in Table 3.1.1.

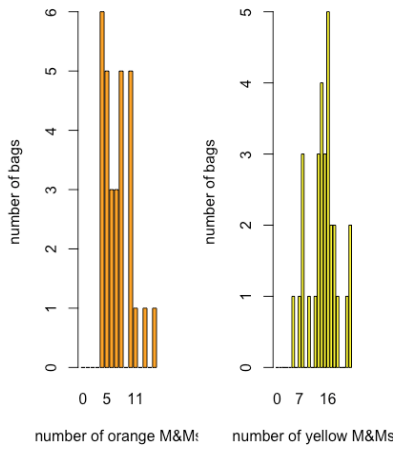

Figure 3.1.7: Bar plots for orange M&Ms and yellow M&Ms using the data in Table 3.1.1.

Here we see that the most common number of orange M&Ms per bag is four, which is also the smallest number of orange M&Ms per bag, and that there is a general decrease in the number of bags as the number of orange M&M per bag increases. For the yellow M&Ms, the most common number of M&Ms per bag is 16, which falls near the middle of the range of yellow M&Ms.

Histograms

A bar plot is a useful way to look at the distribution of discrete results, such as the counts of orange or yellow M&Ms, but it is not useful for continuous data where each result is unique. A histogram, in which we display the number of results that fall within a sequence of equally spaced bins, provides a view that is similar to that of a bar plot but that works with continuous data. Figure 3.1.8, for example, shows a histogram for the net weights of the 30 bags of M&Ms in Table 3.1.1. Individual values are shown by the vertical hash marks at the bottom of the histogram.

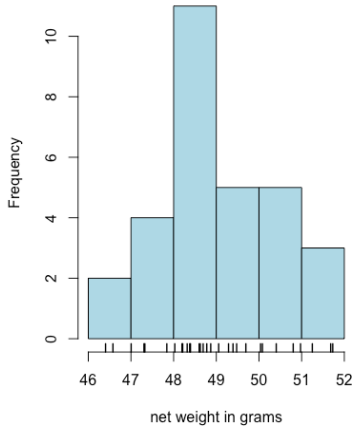

Figure 3.1.8: Histogram of net weights for the data in Table 3.1.1. There are, for example, four bags of M&Ms with net weights between 47 g and 48 g.