The Data-Free Graph

How to streamline your analysis

Donald J. Wheeler

Why bother to plot your data? A simple shortcut is available that will allow you to do your analysis without the data getting in the way. How do you accomplish this breakthrough? Read on.

This marvelous advance in analysis is known as the Data-Free Graph. As usual we begin with a collection of data as in Figure 1.

Roll No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Count	33	37	24	35	22	23	25	23	32	34	33	37	26	36	35	23	26
Roll No.	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Count	22	33	36	38	22	21	23	35	26	35	24	33	21	35	27	26	25
Roll No.	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	
Count	24	23	34	35	34	21	23	34	30	22	25	35	36	37	34	25	

Figure 1: Number of Blemishes per Roll.

Now this is way too much detail. We cannot readily absorb this much information So we simply extract a few summary statistics from our data. First on this list might be the average or a median. A median value for the number of blemishes per roll is 28.5.

Next we need to characterize how the data are spread out above and below the median, so we find the minimum and maximum values for the data. For the numberof blemishes per roll these values are 21 and 38 respectively. To further refine our summary of how the data spread out we also find the quartiles for the data. The quartiles for the number of blemishes per roll are 23 and 35.

Now we assemble these five summary numbers into our graph. We start with a scale that covers the range from the minimum to the maximum values. Next we mark our five summary values on this scale. We draw a rectangle that spreads from the lower quartile to the upper quartile. We divide this rectangle into two by marking the median value on the rectangle. Finally we add a faint, thin line from the upper end of the rectangle to the maximum, and another faint, thin line from the lower end of the rectangle to the minimum. The result is Figure 2.

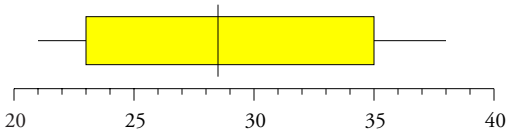

Figure 2: The Data-Free Graph for Blemishes per Roll.

The Data-Free Graph in Figure 2 suggests a broad mound centered near the median value with short tails.

To further illustrate this technique we shall use two more examples. Figure 3 shows the Hot-Metal Transit Times in minutes as well as their Data-Free Graph. The minimum is 20, the lower quartile is 45, the median is 50, the upper quartile is 65, and the maximum is 180.

40	45	125	100	40	40	100	65	55	40	125	65	40	45	95
105	45	110	40	50	120	45	65	105	35	70	55	25	50	55
50	40	40	45	55	50	45	125	55	100	40	70	40	40	110
55	50	30	50	105	45	45	55	50	25	65	60	60	55	70
55	45	100	60	45	145	45	50	65	180	60	45	35	35	55
55	55	50	120	35	45	35	45	55	50	70	45	75	60	45
60	40	60	40	50	60	65	95	65	60	50	25	25	100	50
60	45	35	40	30	180	50	30	30	30	65	130	80	20	45
65	65	45	40	50	25	120	30	115	50	85	40	35	40	40
55	50	25	75	55	50									

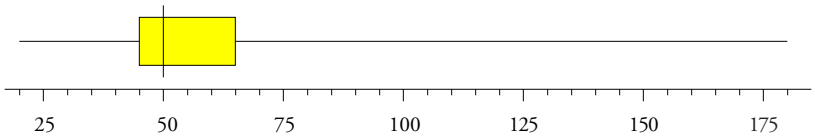

Figure 3 Hot-Metal Transit Times and Their Data-Free Graph

Figure 3 shows an uncomplicated picture of a tight, lop-sided mound and an elongated upper tail.

The Creel Yield Data of Figure 4 has a minimum of 3415, a lower quartile of 3452.5, a median of 3514, an upper quartile of 3524.5, and a maximum of 3542. Figure 4 shows these data along with their Data-Free Graph.

Date	Yields				
8/1	3534	3542	3532	3537	3532
8/2	3533	3524	3524	3525	3527
8/3	3531	3526	3529	3524	3527
8/4	3525	3522	3521	—	—
8/5	3521	3521	3521	3521	3515
8/6	3498	3498	3506	3513	3536
8/7	3526	3529	3524	3525	3520
8/8	3517	3517	3519	3516	3517
8/9	3453	3445	3451	3445	3452
8/10	3445	3449	3454	3447	3446
8/11	3440	3423	3416	3419	3415
8/12	3458	3457	3457	3452	3446
8/13	3448	3451	3453	3453	3455
8/14	3475	3475	3474	3486	3490

Figure 4 The Creel Yield Data and Their Data-Free Graph

Figure 4 shows an uncomplicated picture of a broad, lop-sided mound and an elongated lower tail.

So, rather than wrestling with dozens or hundreds of data, the Data-Free Graph replaces all of the tedious plotting with five simple and easy to obtain statistics. By plotting these five statistics in a graph we combine the simplicity of descriptive statistics with the power of the graph and manage to completely sidestep the original data.

This sidestepping of the original data is a good thing because one of the first principles of good graphics is that equal amounts of data should be represented by equal-sized areas in the graph. But since the Data-Free Graph is made up of descriptive statistics and therefore contains no real data we get to represent the outer 50% of the data with a couple of lines while the inner 50% is represented by our boxes.

If we were to use the four quartiles defined by each of these Data-Free Graphs and represented each quartile by rectangles with equal areas we would end up with the crazy pseudo-histograms shown below.

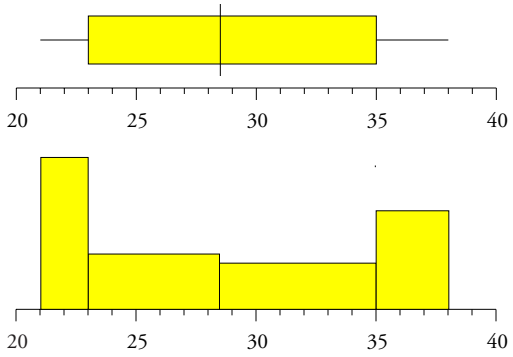

Figure 5 Data-Free Graph and Pseudo-Histogram for the Number of Blemishes per Roll

In Figure 5 we see how the Data-Free Graph hides the concentration of data at each end of the Blemishes per Roll data. The two graphs in Figure 5 deliver distinctly different impressions. When we use the Data-Free Graph we do not have to wonder what these concentrations of data in the tails might mean. We can simply proceed with our analysis without having to stop and explain the messy details.

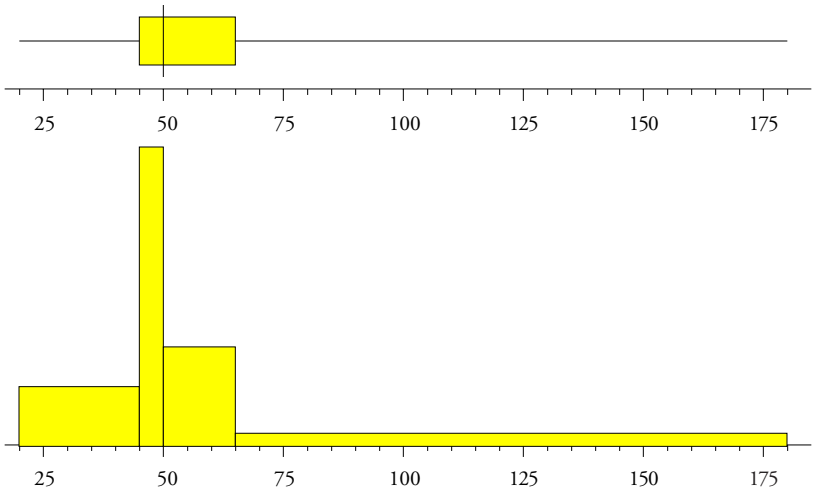

Figure 6 Data-Free Graph and Pseudo-Histogram for the Hot Metal Transit Times

In Figure 6 we see how the Data-Free Graph gives the same weight to the spike between 45 and 50 minutes as it does to the cluster between 50 and 65 minutes. It also gives the same weight to the elongated upper tail as it does to the shorter lower tail.

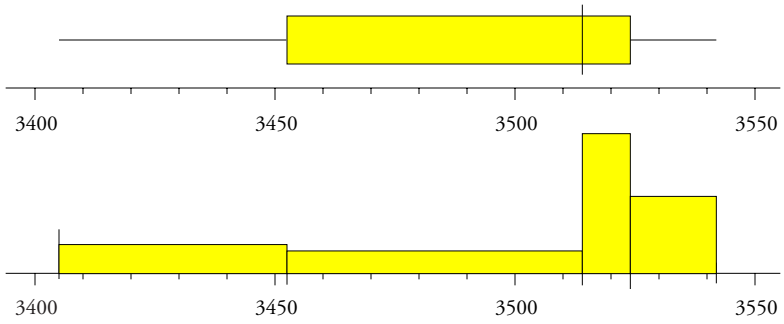

Figure 7 Data-Free Graph and Pseudo-Histogram for the Creel Yield Data

In Figure 7 we see how the Data-Free Graph gives equal emphasis to the second and third quartiles when they represent completely different concentrations of the data. It also gives more weight to the second quartile than it does to the lower tail even though the lower tail is more concentrated. How much simpler everything becomes when we get rid of the data and just use the descriptive statistics!

But what happens if we look at the actual histograms? The data for the number of blemishes per roll from Figure 1 have the histogram shown in Figure 8.

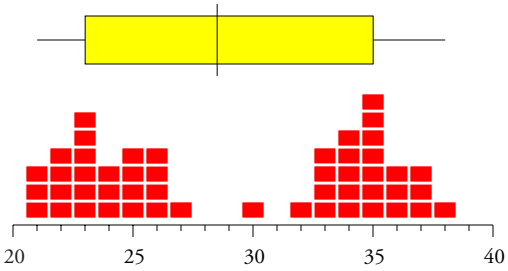

Figure 8 Data-Free Graph and Histogram for the Blemishes per Roll Data

The Blemishes per Roll data have a bimodal histogram. This happened because the rolls were two different sizes. While the blemish rate was reasonably constant, the differences in the areas of the different sized rolls resulted in two different sets of counts. Fortunately, the Data-Free Graph saved us from having to be concerned about this aspect of the data.

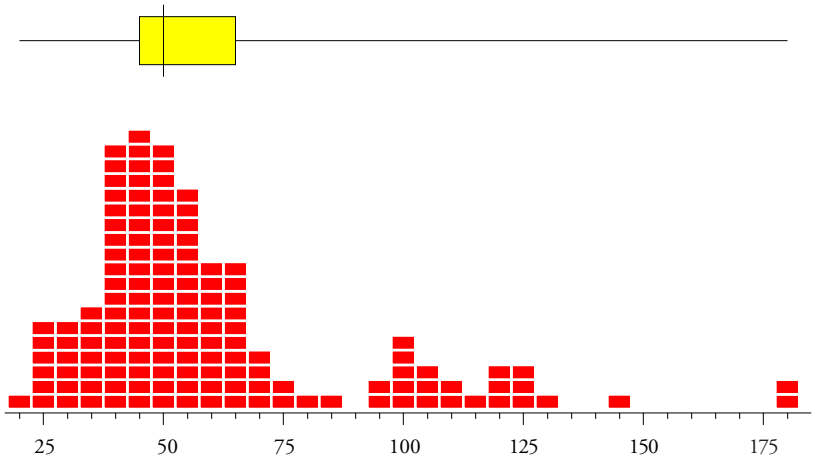

Figure 9 Data-Free Graph and Histogram for the Hot Metal Transit Times

The transit time data have the histogram shown in Figure 9. Here we see two, or possibly even three different mounds within the data. These details get in the way of our statistical analysis, even if they reveal the heart of the problem. Here the railroad crew was moving the hot metal car out from the blast furnace, unhooking, and going off to perform other tasks before returning, sometimes two or three hours later, to deliver the hot metal to the steel furnace. Once more the Data-Free Graph rescues us from having to deal with the realities behind our data!

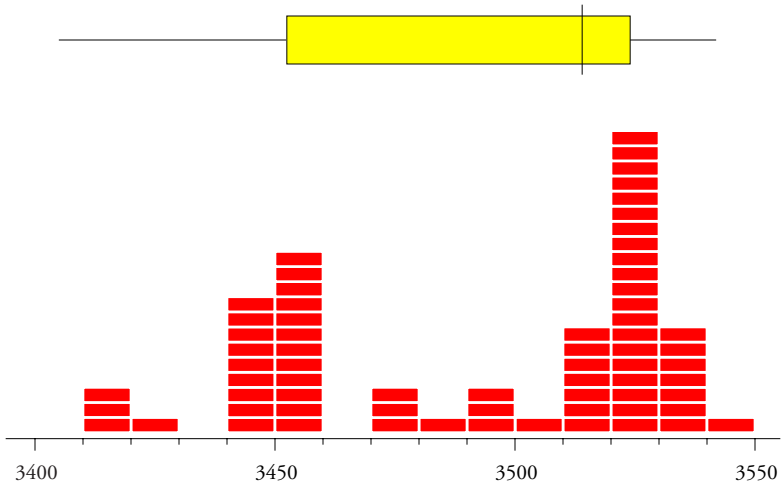

Figure 10 Data-Free Graph and Histogram for the Creel Yield Data

The Creel Yield Data have the histogram shown in Figure 10. These data show up to five distinct mounds. These data come from at least five different production lines. While they are all making the same product, and while the products are all within the specifications, the variation in this material makes it too expensive for the customer to use this product on a high-volume application. However, if we use the Data-Free Graph we are not bothered with these messy details.

Fortunately, most software packages already contain this fool-proof scheme for hiding the tedious information contained within your data. They call the Data-Free Graph a "box and whiskers plot." Regardless of the name used, this plot is guaranteed to replace all of your data with just five simple descriptive statistics whether they are appropriate or not. The Data-Free Graph represents a major breakthrough in statistical analysis because it allows you to paint a very nice and tranquil picture even when your data are messy, nonhomogeneous, and full of interesting signals. After all, you should never let the data get in the way of painting a good picture.

So use the Data-Free Graph and you, too, can be an April fool.