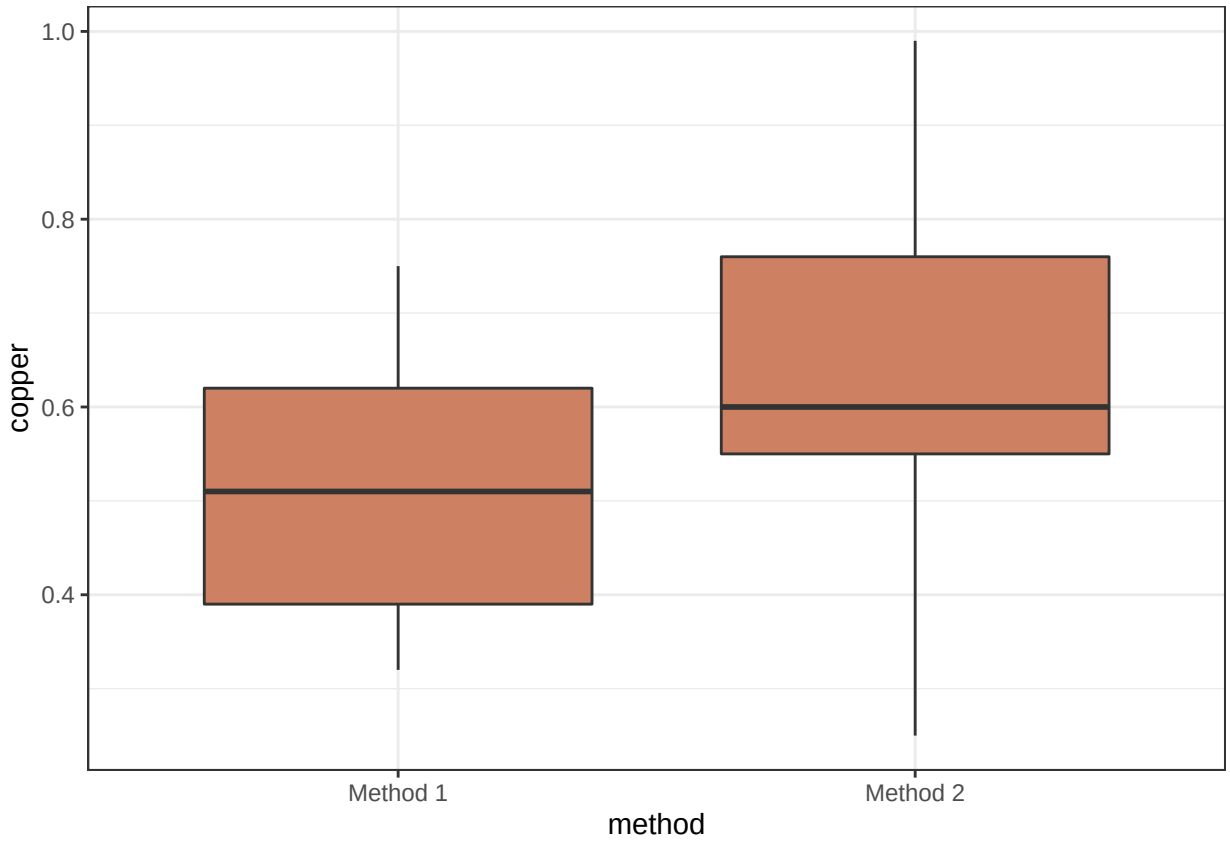
Discussion 2 Solution: Numeric and Graphical Summaries

1. Two methods were studied for the recovery of copper. Thirteen runs were performed using each method, and the fraction of protein recovered was recorded for each run. The results are summarized in the plots below:



   a. Compare the shape, center, and spread of the two methods' fraction of copper recovered.

The data for method 1 is centered at about 0.5 and is roughly symmetrical, the data for method 2 is centered higher, closer to 0.6 and is skewed right. Both have similar spreads with an IQR of about 0.2.

   b. Can you use the above plot to find the mean fraction of copper recovered using Method 1? Explain.

The center line in the box plot represents median, not mean, so it is not possible to read the mean directly off the graph. However, since the data is fairly symmetric, the mean should be quite close to the median, and is probably about 0.5.

   c. Can you use the above plot to find the mean fraction of copper recovered using Method 2? Explain.

For method 2, the data is right skewed, so the mean will be somewhat greater than the median. However, without more information about the data, we cannot be much more precise than saying the mean is somewhat greater than 0.6.

d. After going over the original data again we find out that there was a data entry error. The maximum of Method 1 was accidentially put in as .75, but should have been 1.75. What would this do to the mean of Method 1? What would it do to the median?

Would not change the median, since we are simply making the largest value larger. Would increase the mean.

e. If the mean of the data for Method 2 were 0.63, and another observation was added to the Method 2 group with a value of 0.63, would the standard deviation increase, decrease, or stay the same?

The standard deviation would decrease. The sum of squares would not change, but n would increase, reducing the variance, which would reduce the standard deviation.

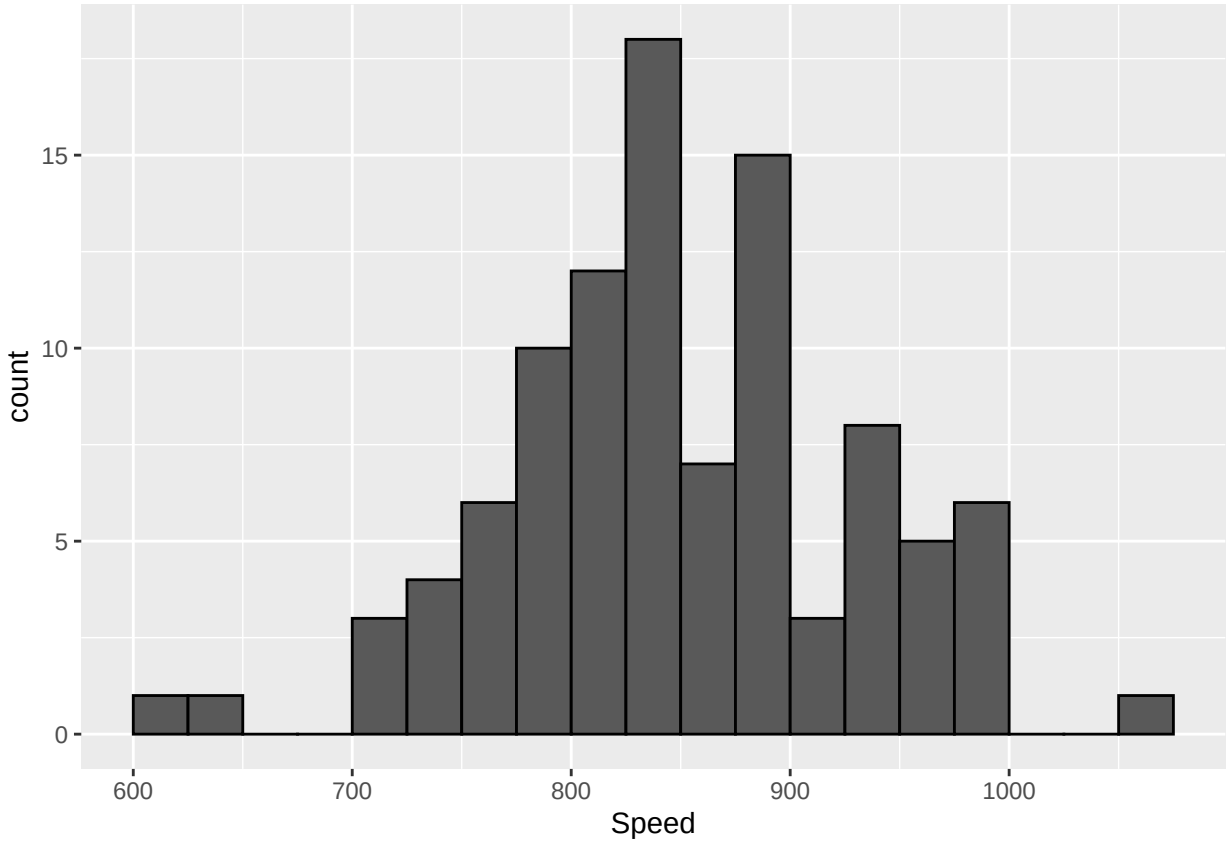
2. The data included in the lightspeed.csv file shows the results of a set of 5 experiments performed by Albert A. Michelson between 1877 and 1882 to measure the speed of light. These were the first in a series of experiments that eventually disproved the existence of "luminiferous aether", the hypothesized medium of propagation for light waves. This research set the stage for the theory of special relativity two decades later.

```
library(tidyverse)

lightspeed <- read_csv("lightspeed.csv")
```

```
## Parsed with column specification:
## cols(
##   Expt = col_double(),
##   Run = col_double(),
##   Speed = col_double()
## )
```

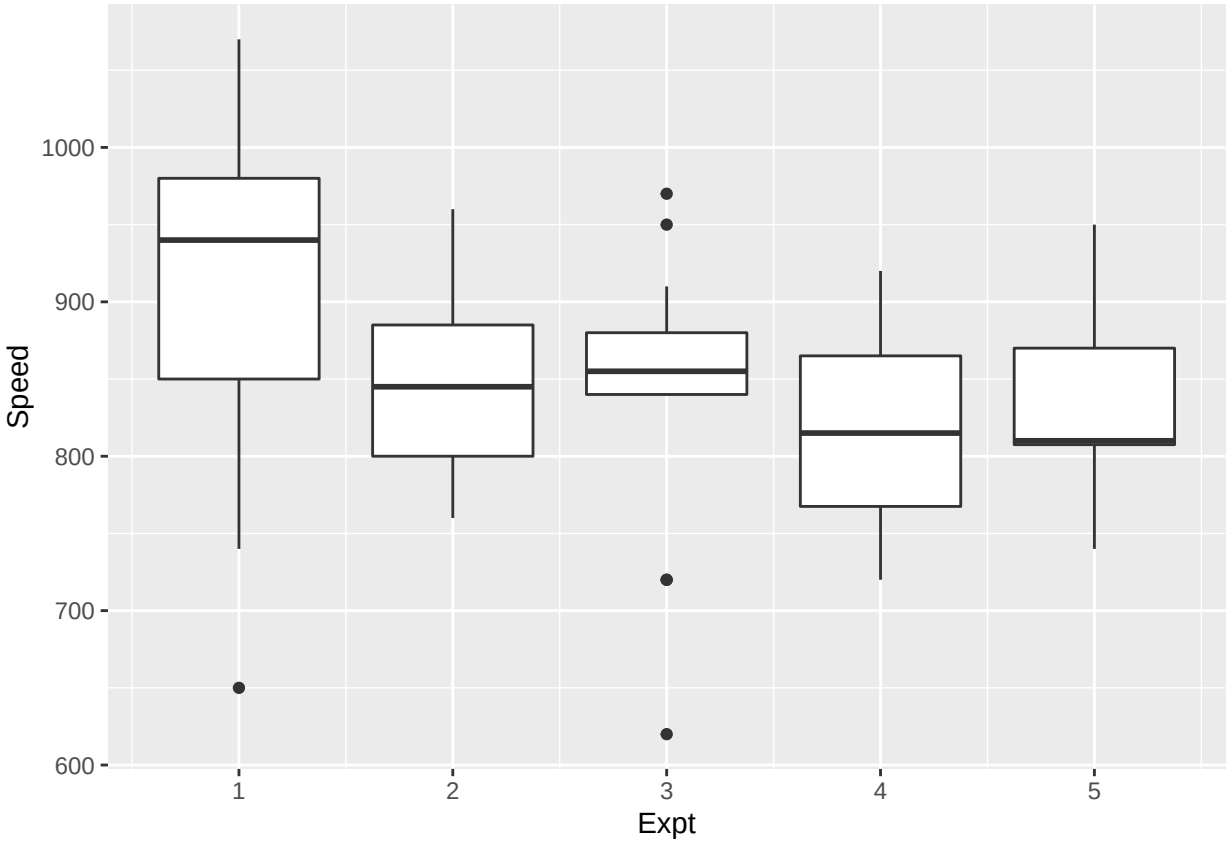
```
ggplot(lightspeed,
       aes(x = Speed)) +
  geom_histogram(binwidth = 25,
                 boundary = 600,    ## this sets one boundary,
                                    ## and creates the rest of the bins
                                    ## using binwidth.
                 color = "black")
```



There are about 16 observations between 750 and 800. If this were a relative frequency histogram, the height of the bar would be 16/100 or 0.16.

c. In R, construct side by side comparative boxplots of the measured speed of light for each of the five experiments. Compare the shape, center, and spread for each of the experiments.

```
ggplot(lightspeed,
       aes(x = Expt, y = Speed, group = Expt)) +  ## fill in the blanks!!!
  geom_boxplot() ## fill in the blanks
```



Experiments 2-5 are centered between 800 and 850, while Experiment 1 is centered closer to 950. Experiments 2,3, and 4 appear to be roughly symmetrical, while Experiment 1 may be skewed left, and Experiment 5 may be skewed right. The spreads of all 5 are roughly comparable, though experiment 3 seems to have the least variation.

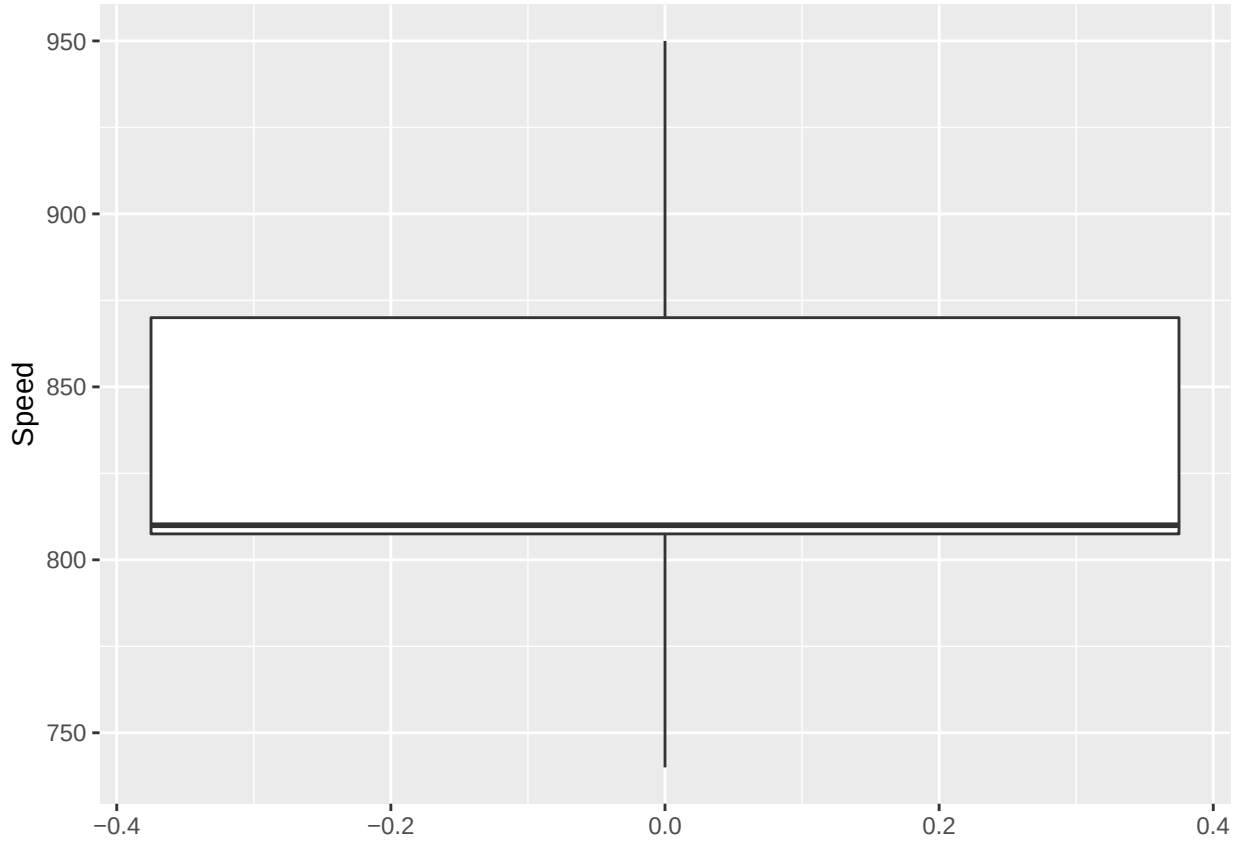
d. Now let's drill down to experiment 3. Using the 1.5 IQR rule, would any of the values in this experiment be considered an outlier?

The IQR is 40, so the outlier cutoffs are 780 and 940. The points at 620, 720, 950, and 970 would be all considered outliers with this criteria.

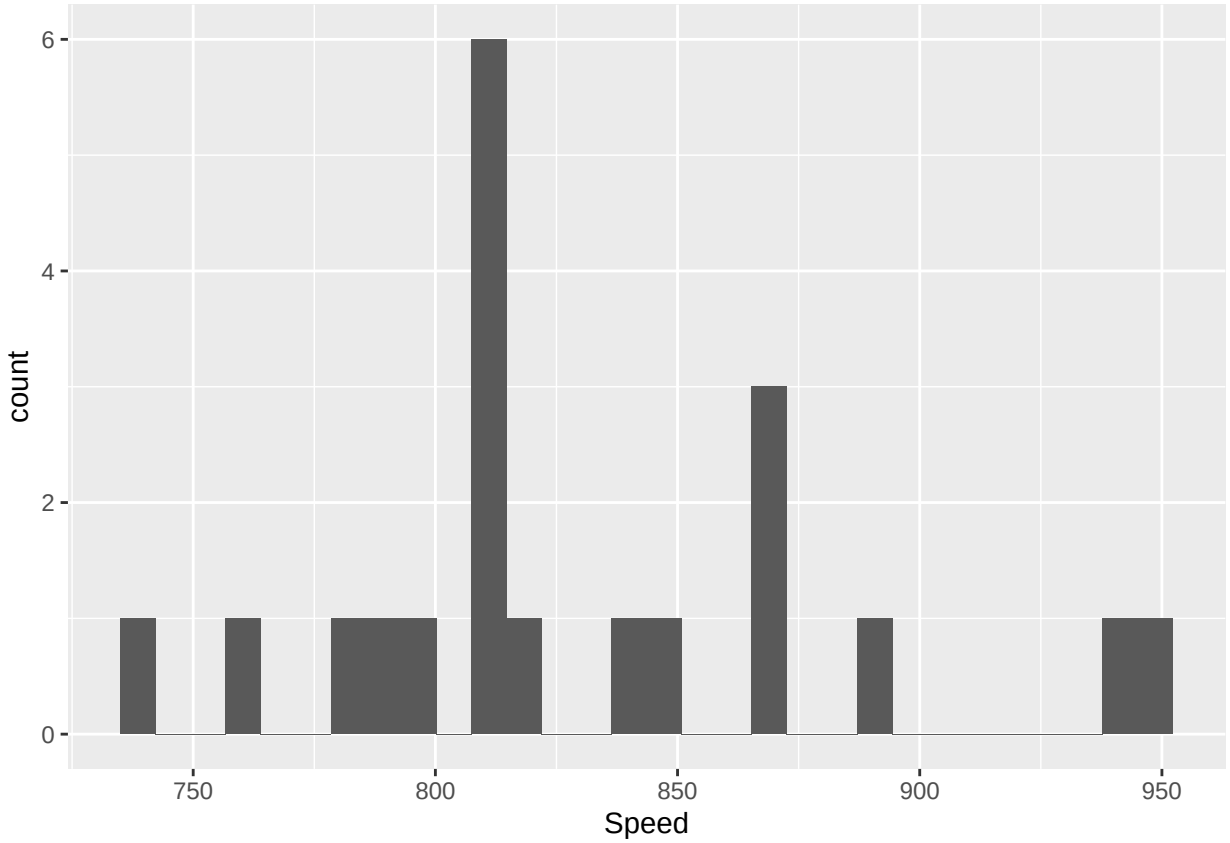
e. Next let's drill down to experiment 5. Construct a histogram and boxplot for the experiment 5 data. Do the different graphical displays seem to say something different about the shape of the data? Explain the apparent contradiction and determine whether the experiment 5 data should be considered skewed or symmetrical.

```
exp5 <- lightspeed %>% filter(Expt==5)

ggplot(data = exp5,
       aes(y = Speed)) +
  geom_boxplot()
```

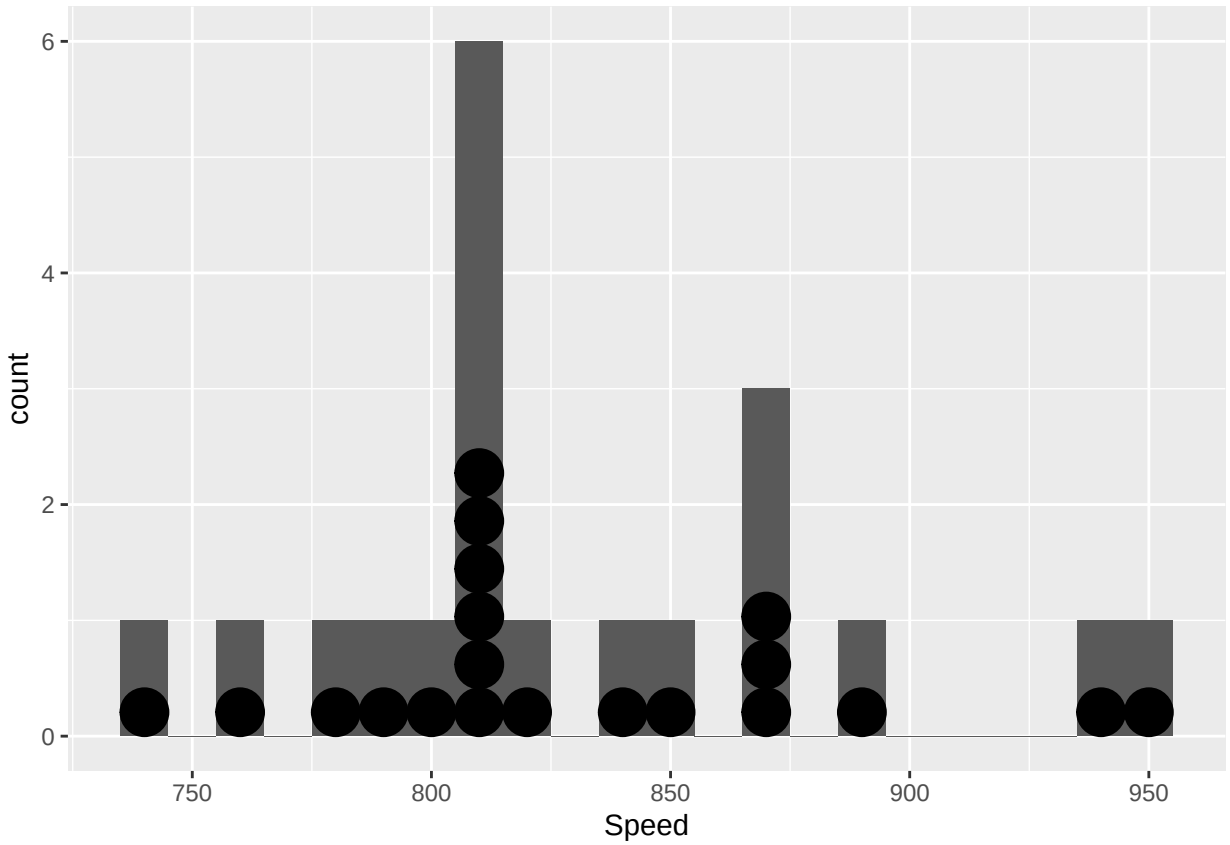



```
ggplot(data = exp5,
       aes(x = Speed)) +
  geom_histogram()
```



This data shows how certain choices of bins can be misleading. The histogram suggests the data is roughly symmetric, while the boxplot shows some right skew. Re-plotting the histogram using ggplot and more bins, and overlaying a dot-plot shows that the data is really closer to right skewed

```
ggplot(data = exp5,
       aes(x = Speed)) +
  geom_histogram(binwidth = 10) +
  geom_dotplot(binwidth = 10)
```


6
4
2
0
count
750
800
850
900
950
Speed