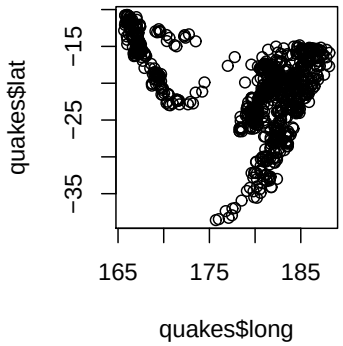
verify this conclusion. This ends the discussion of graphical tools that can be applied to univariate data. The natural extension is to consider graphical tools for multivariate data.

20.2 Multivariate data

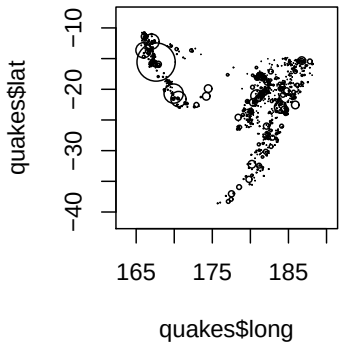
A multivariate data set consists of data values that come in pairs, triples, etc. A simple example of a bivariate data set is the collection of the heights of n couples in which the x_i value corresponds to the husband's height and y_i corresponds to the wife's height, for $i = 1, 2, \ldots, n$. A scatterplot is an effective graphical tool for an initial assessment of a bivariate data set. Returning to the earthquake data set contained in the built-in data set `quakes`, a scatterplot of the longitude of the epicenter of the quake on the horizontal axis versus the latitude of the epicenter of the quake can be generated with the `plot` function:

```
> plot(quakes$long, quakes$lat)          # longitude vs. latitude of epicenters
```



This plot reveals a pattern associated with the locations of the $n = 1000$ earthquakes. The axis labels default to the names of the two arguments to `plot`. It does not, however, incorporate the magnitude of the earthquake. The magnitudes of the earthquakes are stored in `quakes$mag`. The `symbols` function can be used to generate a scatterplot that also incorporates the magnitude of the earthquakes with

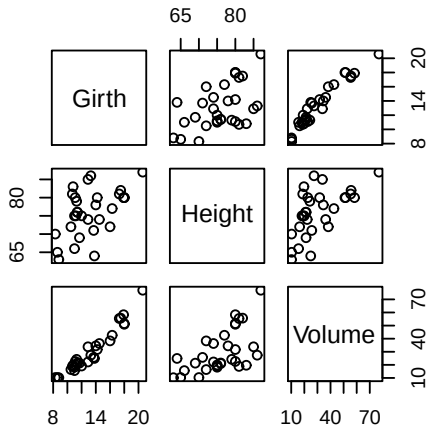
```
> symbols(quakes$long, quakes$lat, circles = 10 ^ quakes$mag)   # include mag.
```



Even though there were more earthquakes that occurred in the eastern cluster, the western cluster contains the strongest earthquakes.

Data in three dimensions (trivariate data) is typically more difficult to visualize using graphics than bivariate data. Consider the built-in data set named `trees`, which consists of the `Girth` (the diameter, in inches, measured 4 ft., 6 in. above the ground), `Height` (in feet), and `Volume` (in cubic feet) of $n = 31$ felled black cherry trees. The `pairs` function draws two-dimensional scatterplots of all possible pairs of the data values:

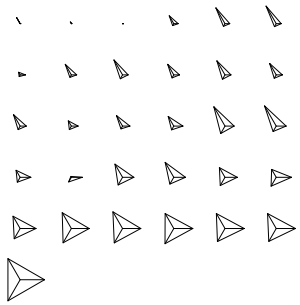
```
> pairs(trees)                          # pairwise Girth, Height, & Volume
```



One observation that is immediately apparent from the scatterplots is that the two variables with the highest positive correlation are `Girth` and `Volume`. Since `Girth` is also an easier measure to obtain than the `Height` of a standing black cherry tree, it is considered the better predictor of the `Volume` of a standing black cherry tree.

A second way of viewing trivariate data (or, more generally multivariate data) is with the `stars` function. For the `trees` data set, a call to `stars` is

```
> stars(trees)                          # star plot for tree data
```



which produces a star for each of the $n = 31$ trees that contains the three variables (girth, height, and volume) by row. A slight variant of a star diagram uses the `draw.segments` argument to modify the look of each data point. For this data set, they look a bit more like a birds-eye view of the trees.