Chapter 2: Looking at Multivariate Data

- Multivariate data could be presented in tables, but graphical presentations are more effective at displaying patterns.
- We can see the patterns in one variable at a time using univariate graphics like histograms, stemplots, and boxplots.
- For multivariate data, graphics that allow us to look at several variables at once are more useful.
- The simplest such plot is the *scatterplot*.

Scatterplots and Beyond

- A simple 2-dimensional scatterplot is useful for visualizing the relationship between two variables.
- With $q > 2$ variables, we will need other plots (more later).
- For bivariate data, the scatterplot can be enhanced with:
  1. rug plots to show marginal distributions
  2. histograms to show marginal distributions
  3. labeling points with abbreviations to identify interesting observations
  4. a regression line or curve overlain on the plot

The Convex Hull of a Data Set

- When examining a data set, we sometimes seek *robust* measures which ignore outliers.
- An example with one-variable data: The *trimmed mean*
- *Trimming*: Highest and lowest observations are trimmed off – the resulting data set may be more representative of the population.
- With multivariate data, we define the convex hull of the data set as the points on the edge of the smallest (convex) perimeter surrounding the data (see example)
- Trimming off these "outlying" observations on the convex hull may display the relationship between the variables better.
- Robust summary measurements (such as the correlation) can be obtained from this trimmed data set.
- Finding the convex hull can be done with the `chull` function in `R`.

The Chi-plot

- The *chi-plot* is a graphical method for judging whether two variables are independent, based on sample data.
- It is based on whether larger values of one variable tend to occur with larger (or smaller) values of the other variable (see pp. 24-25 for exact details).
- This plot can be done with the `chiplot` function from the text's website.
- When the two variables are truly independent, the points in the chi-plot will fall within a central region.
- If the points fall outside this center region, this indicates the two variables may be related.

The Bivariate Boxplot

- The *bivariate boxplot* is an two-dimensional analogue of the familiar boxplot.
- Instead of "boxes," it displays two ellipses centered at the same point (the bivariate center).
- The smaller ellipse contains the "central 50%" of the data and the wider ellipse contains all the data except the "outliers."
- Two regression lines are displayed (the regression of Y on X and that of X on Y).
- The intersection of the regression lines is the bivariate center of the data.
- If the angle between the lines is highly acute, this indicates a strong correlation between the two variables.
- If the angle is almost a right angle, the correlation between the two variables is weak.
- This plot can be done with the `bivbox` function from the text's website.
- Options exist to use robust measures of center, spread, and association when constructing the bivariate boxplot.

Bivariate Density Estimation

- Two-dimensional density estimates (based on sample data) can give a good picture of a bivariate distribution.
- This estimate can be obtained with the `bivden` function from the text's web-site, and it can be plotted in `R` with the `persp` function (nice 3-D picture) or the `contour` function (on top of a scatterplot).

Plots to Display $q > 2$ Variables

- A *bubble plot* can show three variables on a regular 2-D plot.
- Two variables are plotted as usual on a scatterplot.
- Circular bubbles are drawn around each point, with the size of the bubble representing the value of a third variable for that observation.
- The bubbles can be added using the `symbols` functions in `R`.
- By adding colors, a fourth variable can be incorporated into the scatterplot.
- This can be done with the `ggplot` function in the `ggplot2` package in `R`.
- The `ggplotly` function in the `plotly` package is a useful tool for making ggplot graphics interactive (see examples).

Plots to Display $q > 2$ Variables

- A *scatterplot matrix* is a $q \times q$ array of individual 2-D scatterplots.
- This shows the relationship between *all possible pairs* of variables, but not any 3-way associations, for example.
- This can be done in `R` with the `pairs` function.
- The `ggpairs` function in the `GGally` package allows us to see pairwise relationships between variables for data sets that contain both continuous **and** categorical variables.

Three-Dimensional Scatterplots

- 3-D scatterplots can be drawn in `R` (using the `cloud` function in the `lattice` package, for example, but they are often not easy to interpret visually.
- Using "drop lines" may make such plots more interpretable.
- The `scatterplot3d` function in the `scatterplot3d` package is another way to make 3-D scatterplots.
- The `plot3d` function in the `rgl` package makes 3-D scatterplots that can be rotated interactively and labeled more usefully.

Conditioning Plots

- These plots show scatterplots of two variables, conditional on the value of a third variable.
- These are separate scatterplots for different values of the third variable.
- These are especially useful if the third variable is a categorical (grouping) variable.
- This can be done in `R` with the `coplot` function or with the `xyplot` function in the `lattice` package.

Star Plots

- In a *star plot*, the magnitudes of q variables can be represented graphically for each observation.
- There are q points on the star, and the length of each point represents the value for that observation.
- It may be useful to scale or standardize the variables first.
- This can be done in `R` with the `stars` function.
- Alternatively, the stars could be placed on a 2-D scatterplot, and (like with a bubble plot), stars having $q - 2$ points could be placed over each observation.

Chernoff Faces

- With *Chernoff Faces*, each multivariate observation is displayed with a face.
- We represent q variables using q specific characteristics of the face (eye size, mouth angle, nose size, etc.).
- For example, a wide mouth might correspond to a large value for the fifth variable, say.
- Can be done in `R` with the `faces` function, in the `TeachingDemos` package.

Radar Plots

- *Radar plots* are similar to star plots, but generally look a bit more modern and allow for more customization.
- Radar plots for multiple observations could be placed on top of each other, but if there are more than two or three observations, it's usually better to plot them separately.
- Interpretability can be improved if variables that are similar in some sense are placed next to each other in the data frame used to construct the radar plot.

Combining Information on a Continuous Variable and a Categorical Variable

- *Side-by-side boxplots* are a traditional way to show graphically information about how a continuous variable it related to a categorical variable.
- Basically, separate boxplots show the distribution of the continuous variable at each level of the categorical variable.
- *Pirate plots* are similar, but more advanced: They show more information about the continuous variable than boxplots do (including a density estimate, mean, and interval estimate of the mean).
- Pirate plots can be implemented with the `pirateplot` function in the `yarrr` package in `R`.

Animation Plots

- Static plots are like still photographs: an unchanging snapshot of the information in a data set.
- When the information in a data set changes over time (or over the values of some other variables), it is possible to illustrate the changing picture with an *animation plot* or *dynamic plot*.
- These appear like videos to our eye, since they are sequences of static plots that are frames that flip quickly as the values of the time variable change.
- By default, they are presented as gifs.
- Dynamic plots can be implemented using the `gganimate` package in `R`.

Profile Plots

- Profile plots are a simple way to display several multivariate observations graphically.
- They show connected line segments where the height at each joining point is the value of a variable for that observation.
- Another form of profile plot uses a series of bars to represent the variable values for each multivariate observation.
- These work best when there are relatively few observations (and variables) so that the plot looks “cleaner”.
- Plotting the different observations in different colors (or line types) helps distinguish the observations.
- Profile plots for the same data set can appear different visually when the variables are reordered (a possible weakness).

Andrews Plots

- These convert each data vector into a Fourier series, and the resulting curves are plotted together.
- For a data vector $(x_1, x_2, x_3, x_4, x_5, \ldots)$, the corresponding Fourier curve would be

$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \cdots$$

- The curves are plotted on the domain $t \in (-\pi, \pi)$.
- Again, the order of the variables does affect the appearance of the plot.
- Advantage: The distances between curves in an Andrews plot reflect correctly the pairwise distances between observations.
- Disadvantage: The roles of the individual variables are not as apparent in Andrews plots as in the other types of plots.