Understanding exploratory plots in an introductory data analysis course

Ann-Lee Wang
University of Malaya, Institute of Mathematical Sciences
50603 Kuala Lumpur, Malaysia.
E-mail : j2wang@umcsd.um.edu.my

Exploratory plots are new concepts for students in an introductory data analysis course. The students in this course would have at most taken an introductory statistic course. Initially most of these students would hand in some hilarious and erroneous plots. It seems desirable to set down clearly the dos and don'ts of plotting exploratory plots.

To begin with one teaches the technical details of how to plot each type of exploratory plots. These include the concept of unit of measurement, the scale used, the actual size of the plots, the labelling of the plots and the evaluation of the quartiles (in the case of box plot). As far as possible there should be no changes in scale along an axis in a plot. Adequate explanations need to be given for any change in scale. Plots should not be restricted to a small portion of the available space. All the dots of a dot plot have to be plotted on or above one single line.

When students have understood how to draw the plot, it is time to talk about the appearance of the plot or the aesthetic or artistic aspect of a plot. Exploratory plots are used to display data. These plots should be constructed without any misleading features. In a dot plot, the dots should not be seen dancing or zigzagging horizontally or vertically or even leaning to the left or to the right. The dots are placed on or above the axis line and the ticks marking the values on the axis line are placed below the axis line. The dots or circles should be more or less of the same size. In a stem-and-leaf plot, the leaves have to be written out evenly and not be squeezed together in one stem and widely separated in another stem. This is to avoid creating the impression that one stem is much longer than another stem when it actually is not longer and perhaps may even be much shorter. The leaves have to be ordered from the lowest value to the highest value. The number of lines per stem and the number of stems used need to be appropriate for each particular set of data to give a balanced and elegant look to the stem-and-leaf plot. A horizontal box plot with well-proportioned box and whiskers is easier to interpret than a similar vertical plot.

The next concept to be taught appears to be the most difficult concept for the students to grasp. This is the interpretation of the plot. Here students have to learn to describe the data by stating the pattern and any special features of the data. They have to be able to answer questions such as the followings. What can each plot says about the shape, location and spread of a given set of data? Will the plot gives some idea as to the outlying data points and gaps in the data? Are the data points clustered in several separate groups? Do different plots show different things for the same set of data? How to reconcile the differing conclusions, if any? Are the eyes misleading the brain?

Some students do not seem to be able to see the plots as giving evidence to support one view or another. It is quite true of course that sometime the data points are such that it is difficult to describe the data in a meaningful way. Data set with too few points also makes it difficult for the students to describe the data. In an

introductory course such as this, perhaps each set of data should be at least of size thirty. In this course, data sets of fifty points or less are plotted by hand.

Some students are very uneasy about their inability to describe the data set just by looking at the plots. It may be useful then to emphasis once more the importance of plotting neat plots. Some additional simple arithmetic calculations may also help the students to describe the data.

By plotting a dot plot or a stem-and-leaf plot, the students would have sorted the data set. However not many students realise this. Most students would sort the data before plotting.

There is also another point that need to be stressed. A quick glance at the data set may suggest that a simple linear transformation may be performed before plotting commences. The transformation may be of the form $Y = (X - k)/c$, or $Y = a + bX$ where *k, c, a* and *b* are constants, *X* is the raw data value and *Y* is the transformed value. These transformed values may make it easier to plot the data. Care has to be taken to explain that transformation of the data is to make the task of plotting easier not more difficult. Some students tend to carry out transformations that make it more difficult to plot the data.

RESUME:

Daus ce travaille on considere les problems avec l' instruction des "exploratory plots" daus un cours elementaire de analyse de data.