Statistics for Engineering, 4C3/6C3

Assignment 1

Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 16 January 2014**

Assignment objectives: creating and interpreting data visualizations

Question 1 [3]

Which types of features can can the human eye easily pick out of a time series plot?

Solution

Features such as sinusoids, spikes, gaps (missing values), upward and downward trends are quickly picked out by the human eye, even in a poorly drawn plot.

Question 2 [4]

Final exam, 2013: Why is the principle of minimizing "data ink" so important in an effective visualization? Give an engineering example of why this important.

Solution

It reduces the time or work to interpret that plot, by eliminating elements that are non-essential to the plot's interpretation. Situations which are time or safety critical are examples, for example in an operator control room, or medical facility (operating room).

Question 3 [10]

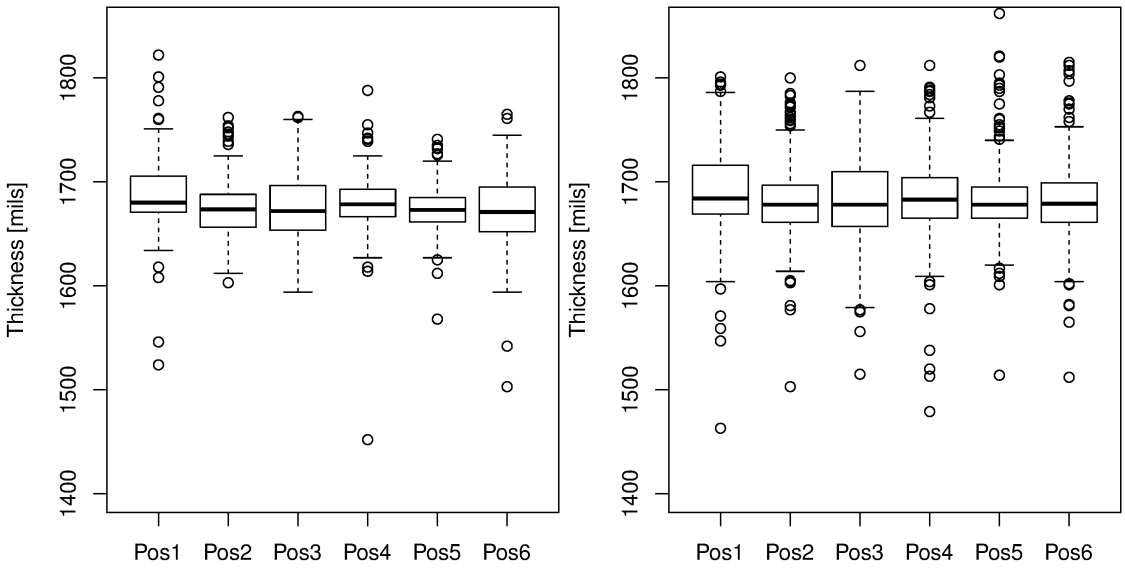
Reproduce the box plot for board thickness that was discussed in class. The board thickness data set is available from the dataset website.

1. Reproduce the figure that was shown in class, using the first 100 rows from the data set. See R code in the course notes.
2. Create a new box plot using rows 3100 to 3300. Interpret any interesting observations from this box plot. Superimpose a target line of 1680 mils.
3. Explain why the thick center line in the box plot is not symmetrical with the outer edges of the box.

This question is to ensure you can install R and use the course dataset site.

Solution

This question was mainly to get you warmed-up to R again, which you may have encountered in your stat prerequisite course; if not, it is a good time to practice some R (or whatever language you prefer to use). The R code below will generate the following 2 figures:



Left: rows 1 to 100 and *right*: rows 3100 to 3300.

```
boards <- read.csv('http://datasets.connectmv.com/file/six-point-board-thickness.csv')
summary(boards)

# Ignore the first date/time column: using only Pos1, Pos2, ... Pos6 columns
first100 <- boards[1:100, 2:7]
later100 <- boards[3100:3300, 2:7]

bitmap('boxplot-for-two-by-six-boards-assign1-2014.png', pointsize=14, res=300,
       type="png256", width=10, height=5)
layout(matrix(c(1,2), 1, 2)) # layout plot in a 1x2 matrix
par(mar=c(2, 4, 0.2, 0.2))   # (bottom, left, top, right) spacing around plot
boxplot(first100, ylab="Thickness [mils]", ylim=c(1400, 1850))
boxplot(later100, ylab="Thickness [mils]", ylim=c(1400, 1850))
dev.off()
```

Some observations noted:

- The second box plot shows the data are more symmetrical for all positions than from the first box plot (except position 1 and 3 which have some skew to the higher thicknesses).
- All positions tend to have outliers above and below the median in the second box plot, far more outliers in fact than in the first box plot.
- There is large outlier at position 5 in the second set of data; perhaps an bump on the edge of the board, in position 5.

The thicker center line is the median; there is no guarantee the median is midway between the first and third quartile, i.e. it is not necessarily halfway along the IQR. The asymmetry of this line indicates (somewhat) how the bulk of the distribution is skewed.

Question 4 [5]

Describe what the main difference(s) between a bar chart and a histogram are.

Solution

The solution is directly from: http://www.forbes.com/sites/naomirobbins/2012/01/04/a-histogram-is-not-a-bar-chart/

- Histograms are used to show distributions of variables while bar charts are used to compare variables.

- Histograms plot quantitative data with ranges of the data grouped into bins or intervals while bar charts plot categorical data.
- Bars can be reordered in bar charts but not in histograms.
- There are no spaces between the bars of a histogram since there are no gaps between the bins. An exception would occur if there were no values in a given bin but in that case the value is zero rather than a space. On the other hand, there are spaces between the variables of a bar chart.
- The bars of bar charts typically have the same width. The widths of the bars in a histogram need not be the same as long as the total area is one hundred percent if percents are used or the total count if counts are used. Therefore, values in bar charts are given by the length of the bar while values in histograms are given by areas.

Question 5 [8]

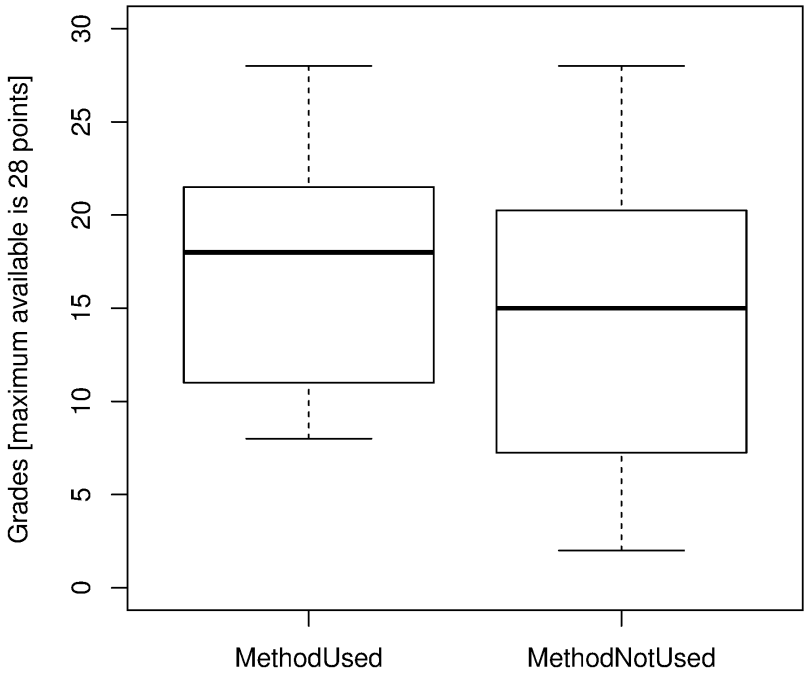
In a question on the final exam in `4M3` there was an open-ended question. The data values are the grades achieved for the answer to that question, broken down by whether the student used a systematic method, or not. No grades were given for using a systematic method; grades were awarded only on answering the question.

A systematic method is any method that assists the student with problem solving (e.g. define the problem, identify knowns/unknowns and assumptions, explore alternatives, plan a strategy, implement the strategy and then check the solution).

Draw two box plots next to each other that compare the two data sets. Also comment on any features you notice in the comparison.

Solution

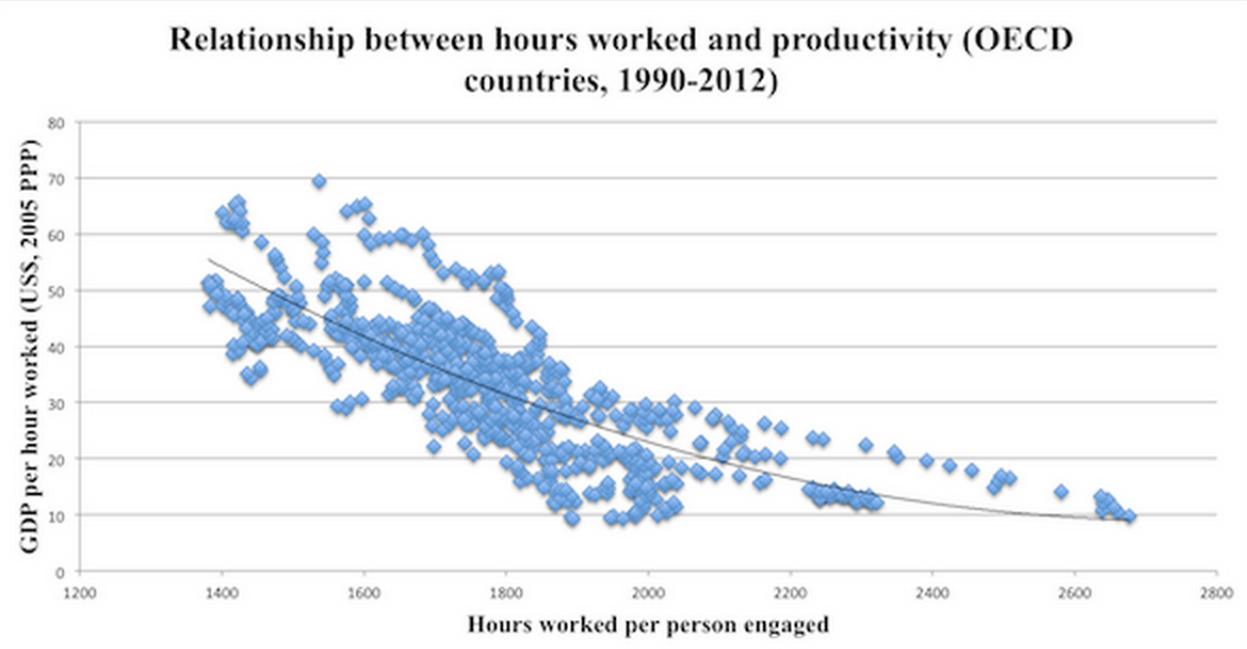
Several points are apparent in the box plot:



- students in either category achieved the highest grade possible
- the spread (interquartile distance) when using the method is smaller
- both box plots show a skew to the lower left tail (compare the median to the first and third quartiles)
- we will use a confidence interval in the next assignment to judge whether this difference is statistical significant or not.

Question 6 [8]

Consider this plot we saw in class (it is also available on-line, with some additional context)



1. What is the plot's author trying to convey with this scatter plot?
2. Do you believe this an effective and complete message (i.e. could you improve it somehow?)
3. Is there a causal mechanism at play between the two variables?
4. How would you confirm or disprove the message the plot's author is making?

Solution

1. The message is likely that longer working hours do not translate into greater earnings (measured with GDP) as might be expected. In fact, the opposite holds: longer working hours are correlated with *lower* earnings (we say: "there's a negative correlation between working hours and earnings"). The axes have been scaled to account for purchasing power.
2. As the original article alludes, there are differences between countries; and given the large number of points on the plot (well over 200) it is safe to assume that there are several points per country, showing the shifts over time. As a result, colour coding, or using different markers to show each country's shift and change over time will provide some additional insight. For example, the line of points stretching from 2200 to 2600 on the x-axis: is that due to one country and in which direction is it moving over time (left or right)?

   Some students rightly pointed out that policy shifts occurred during this period; some countries joined the EU, and that may have lead to a change in the plots. So the picture is by no means complete. However, the picture is almost never complete for any data set.
3. This is a tough one to answer. The data are compelling in their lack of scatter. Usually systems with dubious correlations show a high degree of scatter. As before, colour or shaped codes for each country will give a better idea of cause-effect. I suspect this plot shows a strong correlation simply because there are small clusters for each country that are close together, but the negative trend simply comes from a country-to-country difference.

   As emphasized before in this course, we can only truly tell causality by doing an experiment. Here there are no major ethical obligations, however it is unlikely that you would be able to convince companies to enforce short vs long working hours so you can observe productivity. The time before the change also takes effect is likely very long.

So the answer is yes, maybe there is a causal mechanism here that is plausible (we've often heard that people whose work-life balanced is better are more productive), but we cannot test it explicitly.

4. Also see the prior answer: require experiments over a broad range of employment types and regions, using shorter and longer working hours, and measure the corresponding earnings.

Question 7 [10]

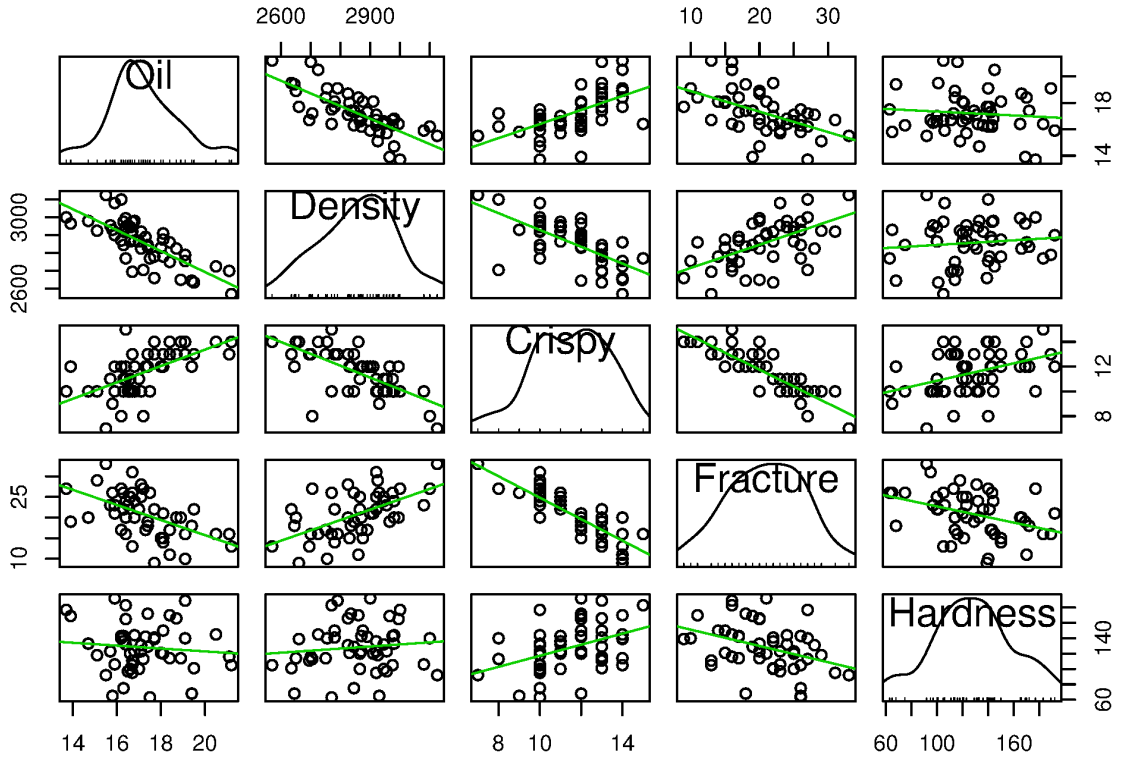
At the start of the class several people indicated they wanted to learn about visualizing more than 3 variables. In class we say a way to visualize at least 5 variables.

Here's another method that you can investigate. Read up about scatterplot matrices, and draw one for the Food texture data set. See the `car` library in R to create an effective scatterplot matrix with the `scatterplotMatrix` function.

Give a couple of bullet-points interpreting the plot.

Solution

A scatterplot matrix can be calculated:




```
food <- read.csv('http://datasets.connectmv.com/file/food-texture.csv')
library(car)

bitmap('scatterplotmatrix-food-data.png', pointsize=14, res=300,  width=7, height=5)

scatterplotMatrix(food[,2:6],       # don't need the non-numeric first column
                  smoother=FALSE)   # hide the smoother and bounds

dev.off()
```

From this plot we see histograms of the 5 univariate distributions on the diagonal plots; the off-diagonal plots are the bivariate correlations between each combination of variable. The trend line (solid light green) shows the linear regression between the two variables. The lower diagonal part of the plot is a 90 degree rotation of the upper diagonal part. Some software packages will just draw either the upper or lower part.

From these plots we quickly gain an insight into the data:

- Most of the 5 variables have a normal-like distribution, except for `Crispy`, but notice the small notches on the middle histogram: they are equally spaced, indicating the variable is not continuous; it is quantized. The `Fracture` variable also displays this quantization.
- There is a strong negative correlation with oiliness and density: oilier pastries are less dense (to be expected).
- There is a positive correlation with oiliness and crispiness: oilier pastries are more crisp (to be expected).
- There is no relationship between the oiliness and hardness of the pastry.
- There is a negative correlation between density and crispiness (based on the prior relationship with `Oil`): less dense pastries (e.g. more air in them) and crispier.
- There is a positive correlation between `Density` and `Fracture`. As described in the dataset file, `Fracture` is the angle by which the pastry can be bent, before it breaks; more dense pastries have a higher fracture angle.
- Similarly, a very strong negative correlation between `Crispy` and `Fracture`, indicating the expected effect that very crispy pastries have a low fracture angle.
- The pastry's hardness seems to be uncorrelated to all the other 4 variables.

Question 8 [0]

Read the short, clearly written article by Stephen Few on the pitfalls of pie charts: Save the pies for dessert, http://www.perceptualedge.com/articles/08-21-07.pdf.

I do recommend you read this. The article presents an easy-to-read argument against pie charts that will hopefully convince you.

Here's a great example from the CRA.

END