Statistics for Engineering, 4C3/6C3

Assignment 1

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 12 January 2012**

Assignment objectives: create suitable data visualizations

Question 1 [2]

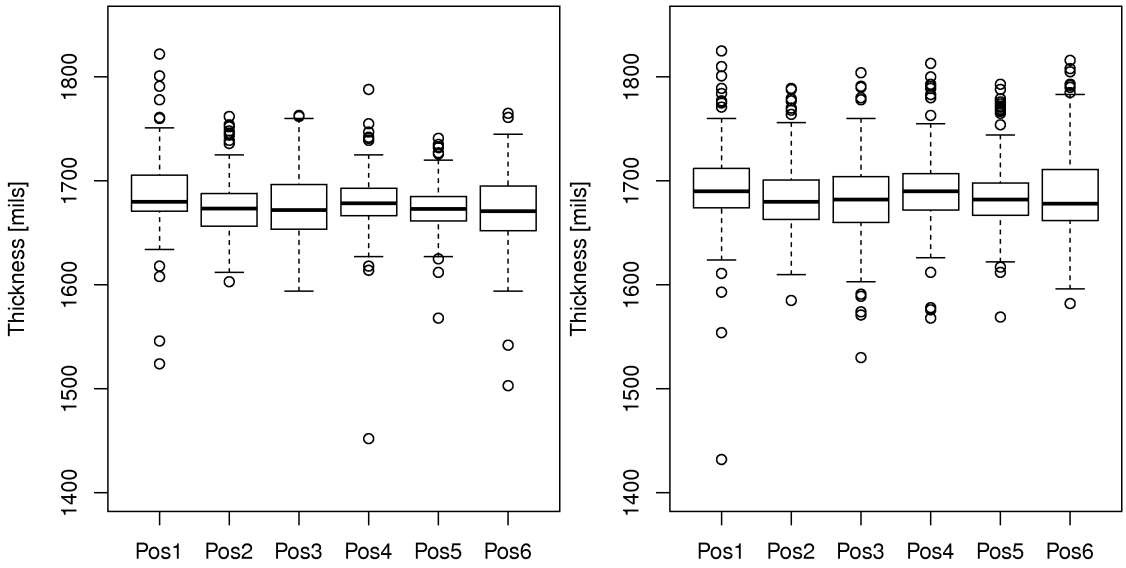
Reproduce the box plot for board thickness that was discussed in class. The board thickness data set is available from the dataset website.

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 4500 to 4600. Interpret any interesting observations from this box plot.

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

Solution

The R code below will generate the following 2 figures:




```
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[4500:4600, 2:7]

bitmap('boxplot-for-two-by-six-boards.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 positions 1 through 5 than from the first box plot.
- However position 6 is less symmetrical (positive skew) than from before, since the median is lower down.
- All positions tend to outliers above the median in the second box plot.
- One below-average outlier appears quite strongly at position 1 in the second set of data.

Question 2 [4]

The instructor uses an app to track his GPS coordinates as he drives to work and back to Hamilton each day. The app collects the location and elevation data every 5 meters, or every 2 seconds, roughly 4000 data points per trip. Data for about 200 trips are described on the course website.

Plot any two interesting visualizations from these data.

Each visualization should be accompanied only by (a) the question you are trying to ask the data set, (b) the plot that you draw, and (c) a single short sentence that summarizes the answer to your question. In other words, the visualization should answer the question, not the text.

Questions should be interesting (i.e. not something like "what is the average trip duration"), but more challenging.

Please feel free to use R, Excel, MATLAB, Python, or any other tool to answer this question, but ensure your plots obey the general guidelines covered in this section of the course.

Solution

There are many potentially interesting questions and plots to use here, illustrating that the visualization should be made relevant to the objective. Some questions you could consider are related to the following plots:

- Fraction of the time moving when on the 407 toll highway or not (box plot)
- Travel times when leaving before or after 08:00 (box plot)
- Average moving speed vs day of the week (box plot)
- Average moving speed when going to or from work (box plot)
- Fuel economy plotted against average speed
- Idling time (total time minus moving time) wen going to or from work (box plot)
- Plotting the maximum speed (or total travel time) against the day of the week (box plot)

END