Statistics for Engineering, 4C3/6C3

Assignment 1

Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 18 January 2013**

Assignment objectives: create suitable data visualizations

Question 1 [10]

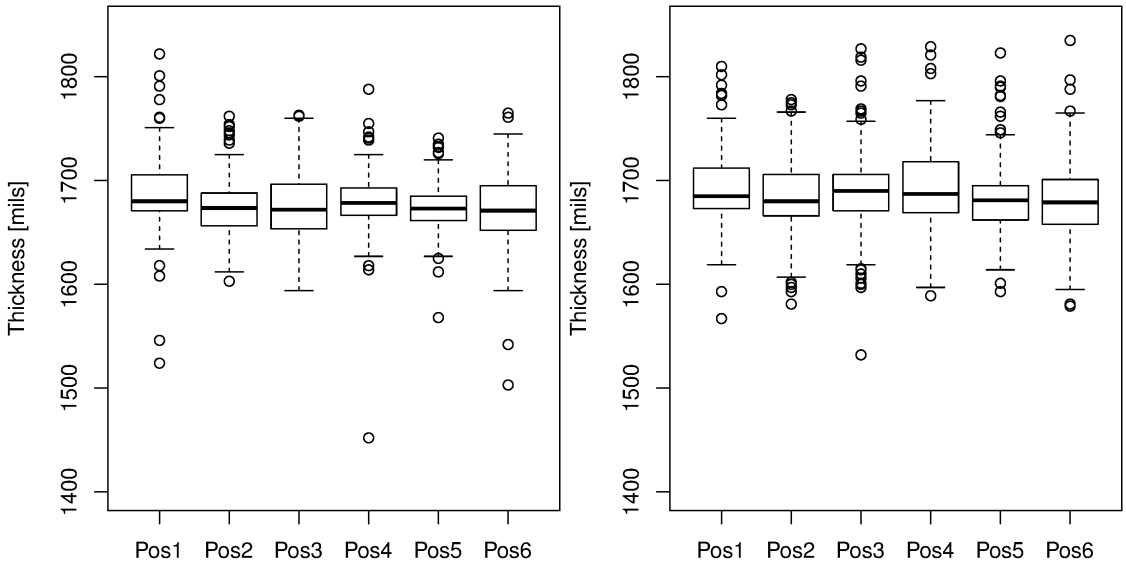
Reproduce the box plot for board thickness that was discussed in class. The board thickness data set is available from the dataset website.

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 4800 to 4900. 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 encountered in your stat prerequisite course. The R code below will generate the following 2 figures:



Left: rows 1 to 100 and *right*: rows 4800 to 4900.

```
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[4800:4900, 2:7]

bitmap('boxplot-for-two-by-six-boards-assign1-2013.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 2 and 4 which have some skew to the higher thicknesses).
- All positions tend to have outliers above and below the median in the second box plot.
- There is on below-average outlier at position 3 in the second set of data. If you look closely in a hardware store, you will often see what it called wane at the edge of a board. This is an example of that, since position 3 (as well as 1, 4 and 6) is at the tip of the board.

Question 2 [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/

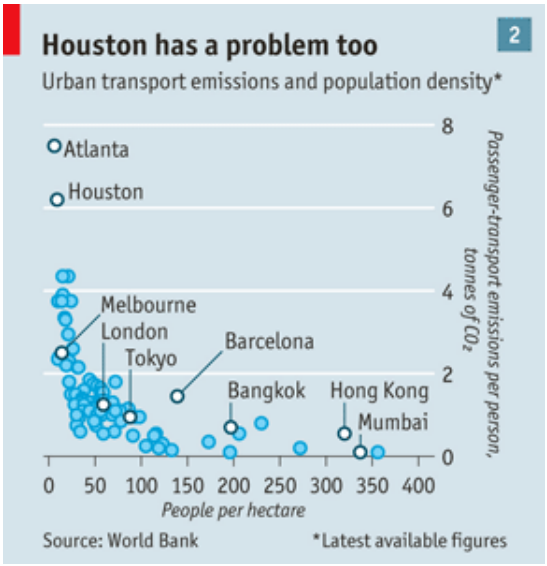
- Histograms are used to show distributions of variables while bar charts are used to compare variables.
- Histograms plot quantitative data with ranges of the data grouped into bins or intervals while bar charts plot categorical data.
- Bars can be reordered in bar charts but not in histograms.
- There 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 3 [5]

In an article published in June 2012, what might this plot's author be asking you to infer?

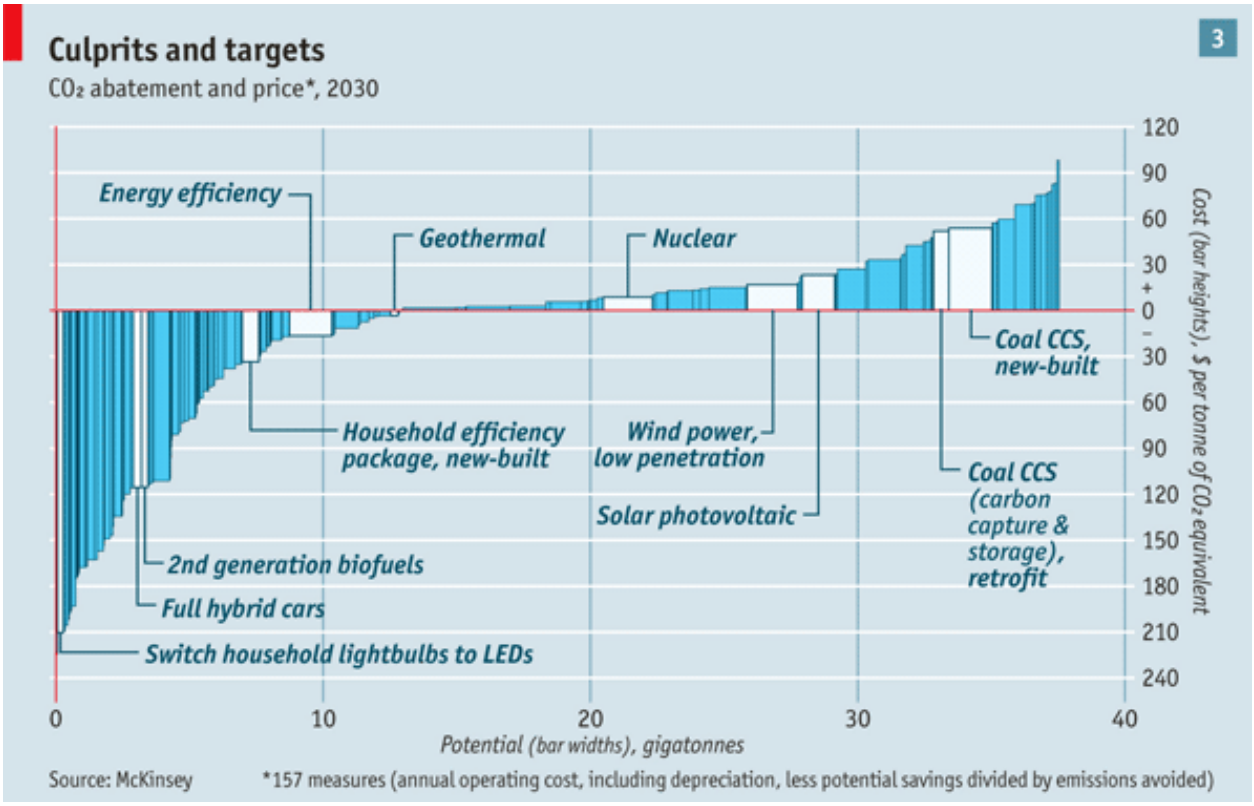
Solution

The plot is asking us to draw a relationship between population density and emissions of carbon dioxide due to passenger transportation. There is not a cause-and-effect relationship, however it is a sensible, and perhaps expected relationship, that comes about due to low *vs* high levels of population density. The relationship is negative and logarithmic: low density areas require their population to use vehicles to get around, while less dense areas can be more efficient (buses, trains and other forms of short-distance urban transport).



The fact that Houston and Atlanta are so different to the other cities with similar population density (e.g. Melbourne) indicates that there is a high level of "driving culture" there and/or low levels of public transport, which would reduce the value on the y-axis.

Question 4 [6C3 students: 8]



Related to the previous question, a cost curve created by a business consultancy, shows different options to cut carbon emission amounts (x-axis).

Describe whether this is an effective visualization. In your answer highlight some of the interesting information you

are able to decode from the visual data.

Solution

I'll be honest and say that I normally find The Economist's data plots very informative and I learn from them. However this one really caused me to scratch my head and took a long time to decode. One of the hallmarks of a good plot is that it *quickly* conveys a difficult data-based concept to the reader.

So after a while I figured out that the y-axis represents the cost and found it interesting that there are many options that have negative cost per tonne of carbon dioxide. From the asterisk at the bottom of the plot, a negative cost can come about due to the savings and low operating cost. For examples LED lighting and hybrid cars, though I still have no idea what the "Energy efficiency" bar refers to.

What I am very unclear about are the bar widths: does the x-axis represent total or incremental gigatonnes of potential savings? For example, does the "Nuclear" option, at around the 21 gigatonne point, save 21 gigatonnes, or is the bar width of about 2 gigatonnes the actual savings? In the latter case, that makes this a really terrible representation of the data and a meaningless x-axis that goes all the way up to 40 gigatonnes.

To be honest, at the end I had more questions than answers, and the original article does not help either. Apparently we are to interpret the bottom corner as options that are cheap and good for the environment. But does that mean the other end is expensive and bad for the environment? I doubt it, because coal CCS, while clearly expensive (the y-axis is large and positive), but that should be good for the environment.

Some of you commented on the (lack) of colour, but bear in mind the magazine only publishes plots in this colour scheme (red, blue and white).

How could this all be improved? My first attempt would be to take the options that are labelled in the plot (there are many bars, but only a few seem to be interesting and labelled), and list these in a short table, each option in its own row. The table's columns would be: name of option, annual operating cost, potential savings, emissions avoided and then the value shown on the y-axis. The user can then quickly scan up and down and recognize the various trade-offs between cost and emissions avoided.

I would be interested to know what your approach would be.

Question 5 [5]

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 these trips are described on the course website.

Plot an interesting visualization from these data. The 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 your written 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 for excellent data graphics 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

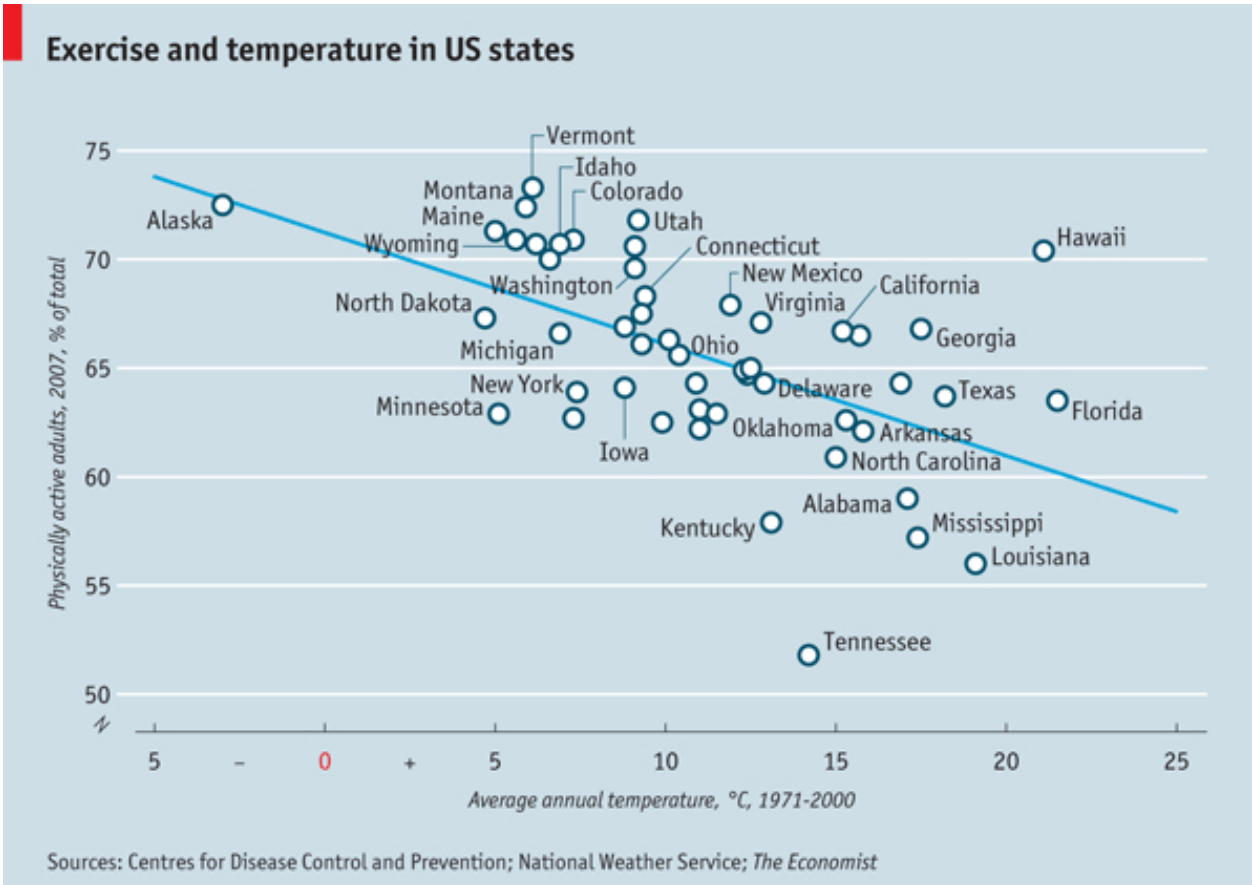
- Idling time (total time minus moving time) when going to or from work (box plot)
- Plotting the maximum speed (or total travel time) against the day of the week (box plot)

Please note: while the Friday travelling times are generally shorter than the other days of the week, this is probably not due to your instructor wanting to rush home (though that is partly true). It is because on Fridays there is consistently lower levels of traffic on all the roads. Other's commented that I speed too much, which is kind-of true since I do like to drive fast, however if you go below 120 on the 407 you really are holding traffic up.

Question 6 [10]

From the 4C3/6C3 final exam, 2012 [8 out of 100 in the exam]

The following figure, taken from The Economist shows the percentage of physically active adults against the average annual temperature, broken down by geographical regions, according to the USA state.



1. Since visualization plots can often stand alone without accompanying text, what is the plot's author asking you to infer from this visualization? **[2]**
2. Is there a causal relationship in the data? Explain your answer. **[2]**
3. The author has shown a linear regression line. Is the intercept term meaningful in this case; please explain. **[2]**
4. Calculate an estimate of the linear model's slope, and give an interpretation for it. **[2]**

Solution

1. To infer that that adults are more physically active in regions with lower average annual temperatures, a statement that might seem counter to usual expectations.

2. No. Causality can never be inferred when the variables cannot be varied independently. In this case, we cannot adjust the average annual temperature, so we cannot imply that this causes those population regions to have lower physical activity. Further, the opposite direction, that average adult physical activity causes the change in average temperature in obviously a bogus statement. We could do an analysis of summer vs winter activity in the same state to see if temperature really does have an effect; or if people who relocate to a colder state become more physically active.
3. "*Yes*": it simply is the level of activity we would expect if there were to exist a region where the *average* annual temperature is 0°C. Intercepts are not meaningful in situations where there are no data close or around the intercept (e.g. when collecting rate data for reactor design); in this case the intercept make sense and can be meaningfully interpreted. Most people answered "*No*" and explained that there were no states where the average temperature is zero, so therefore it is not meaningful. I guess it comes down to semantics and your interpretation of the word "meaningful"; I use the word in the statistical/data analysis interpretation sense here. I will accept either answer for now, but after we've looked at least squares, I would consider *yes* to be the more correct answer.
4. $m \approx \dfrac{70-60}{2.5-22.5} = -0.5$: indicates that for every one degree increase in average annual temperature, we expect the percentage of physically active adults to drop by 0.5%.

Question 7 [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.

Solution

I do recommend you read this, especially if you are of the opinion that pie charts are OK. The article really does present an easy-to-read argument against pie charts that will hopefully convince you otherwise.

END