Statistics for Engineers

© Kevin Dunn, 2015
kevin.dunn@mcmaster.ca
http://learnche.mcmaster.ca/

Univariate Data Analysis
Copyright, sharing, and attribution notice

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 Unported License. To view a copy of this license, please visit http://creativecommons.org/licenses/by-sa/4.0/

This license allows you:

- **to share** - to copy, distribute and transmit the work, including print it
- **to adapt** - but you must distribute the new result under the same or similar license to this one
- **commercialize** - you *are allowed* to use this work for commercial purposes
- **attribution** - but you must attribute the work as follows:
  - “Portions of this work are the copyright of Kevin Dunn”, *or*
  - “This work is the copyright of Kevin Dunn”

(when used without modification)
We appreciate:

▶ if you let us know about **any errors** in the slides
▶ **any suggestions to improve the notes**

All of the above can be done by writing to

kevin.dunn@mcmaster.ca

or anonymous messages can be sent to Kevin Dunn at

http://learnche.mcmaster.ca/feedback-questions

If reporting errors/updates, please quote the current revision number:

Please note that all material is provided “as-is” and no liability will be accepted for your usage of the material.
Topics covered in this section on data visualization

Data visualization
- Use of colour
- Tables
- Scatter plots
- Time-series plots
- Bar plots
- Box plots
Some ways you can impress your colleagues and manager in the future

- **Co-worker**: Here are the yields from a batch system for the last 3 years (1256 data points), can you help me:
  - understand more about the time-trends in the past 3 year?
  - efficiently summarize the yield from all batches run in 2010?

- **Manager**: effectively summarize the (a) number and (b) types of defects on 17 aluminum grades for the past 12 months

- **Yourself**: 24 different variables being measured vs time (5 readings per minute, over 300 minutes) for each batch we produce; how can we visualize these 36,000 data points?
  - see next slides
Batch systems generate large quantities of extremely valuable data.
Batch systems generate large quantities of extremely valuable data.

Data from a single batch:

- Collector tank level
- Agitator Speed
- Dryer Temp SP
- Dryer Temp

Data from many batches:

- Collector tank level
- Jacket Temp SP
- Dryer Temp
Recommended references for this section: Data visualization

Why bother learning about this topic: it’s too easy!

This class might seem too easy; too obvious. It is!

- The human eye and brain are excellent at pattern recognition, sorting through signal and noise.
- We can easily cope with bad plots; but good plots save time and show a clearer, more honest picture.
- Cliches: “Let the data speak for themselves”, “Plot the data”

**Strong suggestion:** find a bad plot (journal publications, an old lab report that you have written); upload it to the forums and criticize the plot. Why is it bad?
We need good plots to make decisions quickly, correctly, and confidently.
Time-series plots show a univariate piece of information in 2 dimensions

- (usually) have the horizontal $x$-axis show time or sequence order
- the other axis: the data values

- Our eyes can deal with high data density, sinusoids, spikes, patterns, can separate noise from signal, and recognize outliers.
An example of a bad time-series plot; what problems can you identify?
Notice how the plot’s “message” is entirely different now

A first attempt at fixing the prior visualization
Poor plots; default settings in plotting software create cluttered plots

The plot has been slightly improved here:
Use separate, parallel axes rather to compare plots (and minimal data ink)

These non-default settings can take a long time to set (10 minutes for this example)
Sparklines are a type of time-series plot

- except, we omit the horizontal and vertical axes (they are implicit)
- Read more about them from http://yint.org/sparklines

- Useful for financial trends
- Built into Excel 2010
- Great for iPods, cell phones, tablet computers
  - because they are of high density and small size
- Our eye can detect 250 dots (points) per linear inch and 650 points per square inch.
Sparklines are used on various websites now to show high-density graphics.

<table>
<thead>
<tr>
<th>Company name</th>
<th>Price</th>
<th>Change</th>
<th>Chg %</th>
<th>d</th>
<th>m</th>
<th>y</th>
<th>Mkt Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>XFN iShares S&amp;P TSX C...</td>
<td>28.77</td>
<td>+0.04</td>
<td>0.14%</td>
<td></td>
<td></td>
<td></td>
<td>861.90M</td>
</tr>
<tr>
<td>HXU Horiz. BetaPro S&amp;...</td>
<td>23.23</td>
<td>+0.14</td>
<td>0.61%</td>
<td></td>
<td></td>
<td></td>
<td>65.81M</td>
</tr>
<tr>
<td>HXD Horizons BetaPro S...</td>
<td>6.39</td>
<td>-0.04</td>
<td>-0.62%</td>
<td></td>
<td></td>
<td></td>
<td>68.54M</td>
</tr>
<tr>
<td>HFD Horiz. Beta. S&amp;P/...</td>
<td>3.58</td>
<td>+0.01</td>
<td>0.28%</td>
<td></td>
<td></td>
<td></td>
<td>9.39M</td>
</tr>
<tr>
<td>HEU Hor. Beta. S&amp;P/TS...</td>
<td>5.97</td>
<td>-0.05</td>
<td>-0.83%</td>
<td></td>
<td></td>
<td></td>
<td>15.95M</td>
</tr>
<tr>
<td>HFU Horizons BetaPro ...</td>
<td>19.96</td>
<td>-0.04</td>
<td>-0.20%</td>
<td></td>
<td></td>
<td></td>
<td>20.40M</td>
</tr>
<tr>
<td>XEG iShares S&amp;P TSX C...</td>
<td>17.15</td>
<td>+0.03</td>
<td>0.18%</td>
<td></td>
<td></td>
<td></td>
<td>642.00M</td>
</tr>
<tr>
<td>HGU Ho. Bta. S&amp;P/TSX ...</td>
<td>7.30</td>
<td>-0.09</td>
<td>-1.22%</td>
<td></td>
<td></td>
<td></td>
<td>103.74M</td>
</tr>
<tr>
<td>HQD Ho. Beta. S&amp;P/TSX...</td>
<td>20.07</td>
<td>+0.26</td>
<td>1.31%</td>
<td></td>
<td></td>
<td></td>
<td>24.03M</td>
</tr>
<tr>
<td>XMA iShares S&amp;P TSX C...</td>
<td>12.69</td>
<td>+0.10</td>
<td>0.79%</td>
<td></td>
<td></td>
<td></td>
<td>143.53M</td>
</tr>
<tr>
<td>XTR iShares Diversifi...</td>
<td>12.00</td>
<td>+0.03</td>
<td>0.25%</td>
<td></td>
<td></td>
<td></td>
<td>702.64M</td>
</tr>
</tbody>
</table>

Screenshot from Google Finance 08 January 2014

Notice how you clearly detect correlations in stock prices (stocks that move together).

Screenshot from Drupal website to track software bugs.
Example of sparklines in everyday use

[Wikipedia: File:12leadECG.jpg]
Keep the x-axis spacing constant on time-series plots: helps interpretation

- Use another plot (e.g. below the original) to zoom in on details
Provide an honest message to your viewer

Adjust for inflation when plotting money values against time.

Screenshot from Google Finance.

Example of car sales: http://www.duke.edu/~rnau/411infla.htm
Show a reasonable amount of historical data for context

1. Got to buy some of this stock!

2. But, here is some more context

3. And, even further context

4. To finish: all available data
Important learning points from time-series plots

- avoid using colour as your message
- use honest scaling on your $x$-axis and $y$-axis
- the human eye (and brain) can deal with vast quantities of data: exploit it
Bar plots are univariate plots, on a 2-D axis.

Use a bar plot when you have many categories, and the literal interpretation does not depend on category order.
Very different messages come across, even though the data are identical. There’s no direct message here. This message is more clear: the reader can quickly see their greatest expenses.
You should not use a bar plot to show time-series data

Rather use a time-series plot, which is much less wasteful and shows the trends more clearly.
Bar plots can be wasteful as each data point is repeated several times:

1. left edge (line) of each bar
2. right edge (line) of each bar
3. the height of the colour in the bar
4. the number’s position (up and down along the y-axis)
5. the top edge of each bar, just below the number
6. the number itself
A general plotting principle: “Maximize thee data ink ratio”, within reason

Maximize data ink ratio

\[
\text{Maximize data ink ratio} = \frac{\text{total ink for data}}{\text{total ink for graphics}}
\]

\[
= 1 - \text{proportion of ink that can be erased without loss of data information}
\]

For example, rather use a table for a handful of data points:

<table>
<thead>
<tr>
<th>Province</th>
<th>Profit ($ '000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontario</td>
<td>562</td>
</tr>
<tr>
<td>Manitoba</td>
<td>423</td>
</tr>
<tr>
<td>Quebec</td>
<td>231</td>
</tr>
<tr>
<td>Nova Scotia</td>
<td>181</td>
</tr>
<tr>
<td>Alberta</td>
<td>90</td>
</tr>
<tr>
<td>British Columbia</td>
<td>82</td>
</tr>
</tbody>
</table>
Don’t use cross-hatching, textures, or unusual shading in the plots: it creates visual vibrations.
Worst bar plot ever?

Actual example from a “production report” board at a company.
Bar plots often benefit from a horizontal presentation, especially if
- there is some ordering to the categories
- the labels do not fit side-by-side

You can place the labels inside the bars

You should usually start the non-category axis at zero.
Unnecessary plot embellishments are not required

- Avoid unnecessary “extras” to enliven the plot
- “If the statistics are boring, then you’ve got the wrong numbers” [Tufte]

But living in a rented semi-detached home with three college students means she’s eager to find her own space. She was also careful enough to save $40,000 for a down payment during her university years by running a College Pro Painters franchise.

Buyers today can get a variable-rate mortgage at prime or 2.25 per cent, and in many cases cheaper after discounting.

But even at the prime rate, it would cost only $1,347 to carry a $400,000 home with an amortization of 35 years and a 5 per cent down payment. By comparison, an average two-bedroom condo in the Toronto area costs $1,487 per month to rent.

That’s a compelling reason for home ownership.

[Toronto Star, 2010]
Consider a vector of temperature data

For example, this is the weather in Hamilton, Ontario, on 8 January 2015:

**Current Conditions**

- **Temperature**: -11°C
- **Condition**: Light Snowshower
- **Pressure**: 102.0 kPa
- **Tendency**: Falling
- **Visibility**: 19 km
- **Temperature**: -10.6°C
- **Dewpoint**: -15.8°C
- **Humidity**: 86%
- **Wind**: SW 43 gust 54 km/h
- **Wind Chill**: -22
Outliers will be defined later: they are unusual data points that are far away from the “bulk” of the data.
Box plots: compared to a pure normal distribution

[Wikipedia has some really great illustrations to explain statistical concepts, such as this plot]
Case study: lumber cutting

The log is completely scanned by lasers, and a few seconds later a computer determines lumber cuts that will maximize the economic value per log.

The log is rotated and guided into rigid saw blades to achieve the predicted result.
Case study: lumber cutting

After cutting, the thickness is measured at 6 locations; target = 1680 mils

Actual thickness of a 2x6 is = 1500 mils; a little extra is added to compensate for the lumber drying out
Box plots are very effective for comparing similar variables (in the same units of measurement)
Box plots: some alternatives you might see in practice

There is no agreed on definition:

- can use the mean instead of the median
- outliers shown as dots, where an outlier is most commonly defined as any point 1.5 IQR distance units above and below the median.
- use the 2nd percentile (instead of median \(-1.5 \cdot IQR\))
- use the 98th percentile (instead of median \(+1.5 \cdot IQR\))
- add the density histogram onto the box plot: *violin plot*
  - Now we can see some of the distortion at positions 1 and 3 (next slide)
Variations on a theme: the violin plot as an alternative to the box plot
Scatter plots help understand the relationship between two variables

- It is a two dimensional plot of two variables (vectors).
- Each marker is the intersection of the values from the data vectors.

**Intention of a scatter plot**

Asks the viewer to draw a causal relationship between the two variables
A scatter plot showing a cause-and-effect relationship

And in many cases, that causal relationship actually exists.
Scatter plots

However, not all scatter plots show causal phenomenon.

Student 2013, Hawra: “Although scatter plots may imply a cause and effect relationship exists, it is not a ‘tool’ to test the existence of a possible relationship.”
Three variables: so 3 scatter plot combinations could have been drawn. Which are correlations, and which are actually cause-and-effect?

We will answer the question in the Experiments section.
Scatter plots: is there cause and effect here?

http://yint.org/working-hours explains the relationship and the data source.
Scatter plots can be greatly improved from the software defaults by:

- making each axis as tight as possible
- avoiding heavy grid lines
- use the least amount of ink
- not distorting the axes
Note the tight axes, low amount of data ink: scatter plots are efficient.
There is an unfounded fear that others won’t understand your scatter plot

- Plant control room: seldom see scatter plots.
- Japanese newspapers frequently use scatterplots
- He shows 12 year olds can interpret such plots.

**Key point**
The producers of charts must assume their audience is capable of interpreting them.

Rather, assume that if you can understand the plot, so will your audience.
Take a diversion to watch this YouTube video http://yint.org/rosling-video
Hans Rosling has used these to great effect to illustrate issues related to International Health

Variables shown in the figure:

1. \(x\)-axis: income per person
2. \(y\)-axis: children per woman
3. marker area: population
4. colour: life-expectancy [20 to 80]
5. time-based animation: on the GapMinder website you can “play” the graph over time

The [http://gapminder.org](http://gapminder.org) site allows you to select many interesting variables on the axes.

Watch a video explanation of these data: [http://yint.org/rosling-video](http://yint.org/rosling-video)
3D glasses used to visualize process data in 6-dimensions?

It would be possible with 3D glasses.
Scatter plots lose density information: recover it with some jitter

R code:

plot(education, vocabulary)  
plot(jitter(education), jitter(vocabulary))
Recover distribution and spread information with box plots

Used the scatterplot(...) function from the library(car) in R to create these.
Consider adding histograms when you are exploring the data to learn about the system.
Investigate this plot in your own time

- Why did the author use a time-series plot to show correlation?
- Would the plot be more informative as a 2D-scatter plot?
- Redraw a rough version of this plot as a scatter plot instead.
- What if you were to repeat this analysis for multiple regions/countries/cities. How would you show (visualize) the correlations effectively?
- Read the full article for details: http://yint.org/lead-and-crime

Sources: Rick Nevin, USGS, DOJ
Some data visualization blogs worth reading

FlowingData | Visualization and Statistics
flowingdata.com
$5.2 million in extra cab tips, found in public data ... software packages, which come ready-made with commands for statistical analysis and data visualization.
Learning - 19 Maps That Will Blow Your ... - Projects - Chart-topping songs

Visually Blog: Visual Content Marketing Blog
blog.visual.ly
Discover the latest news and insights in this visual content marketing blog from the ...
These Seven Cardinal Sins are sure to miscommunicate your data and are ...

information aesthetics - Data Visualization & Information ...
informatieaesthetica.com
Visualizing Publicly Available US Government Data Onl ». 19 September .... Recent entries, Recommended blogs, Recommended articles (more »). Visualizing ...

Information Is Beautiful
www.informationisbeautiful.net
Dedicated to distilling the world’s data, information and knowledge into beautiful, interesting and, above all, useful visualizations, infographics and diagrams.

What is your favorite data visualization blog? - Cross Validated
stats.stackexchange.com/.../what-is-your-favorite-data-visualization-blog
What is the best blog on data visualization? I'm making this question a community wiki since it is highly subjective. Please limit each answer to one link.

GE Data Visualization
visualization.geblogs.com
At GE, we believe data visualization is a powerful way to simplify complexity. We are committed to creating visualizations that advance the conversation about ...
Tables

Tables are for **comparative** data analysis on **categorical objects**.

<table>
<thead>
<tr>
<th></th>
<th>Bank loan monthly payments</th>
<th>Monthly lease payment</th>
<th>Minimum downpayment for lease</th>
<th>Total interest paid over 48 months</th>
<th>Monthly insurance payment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ford Fusion</strong></td>
<td>552</td>
<td>395</td>
<td>0</td>
<td>2,529</td>
<td>180</td>
</tr>
<tr>
<td><strong>Honda Civic</strong></td>
<td>538</td>
<td>424</td>
<td>0</td>
<td>2,466</td>
<td>236</td>
</tr>
<tr>
<td><strong>Mazda 3</strong></td>
<td>506</td>
<td>478</td>
<td>1,000</td>
<td>2,318</td>
<td>251</td>
</tr>
<tr>
<td><strong>Toyota Yaris</strong></td>
<td>435</td>
<td>490</td>
<td>1,000</td>
<td>1,992</td>
<td>198</td>
</tr>
<tr>
<td><strong>VW Golf</strong></td>
<td>596</td>
<td>550</td>
<td>2,500</td>
<td>2,730</td>
<td>244</td>
</tr>
</tbody>
</table>

- **categorical objects**: the cars
- Note the rows are in *default* alphabetical order.
- We can make the table “tell a story” if we reorder the rows by some other variable.
  - e.g. monthly insurance payment
Tables

- Compare defect types (columns) for different product grades (rows)
- Categorical variables appear in the **rows** and **columns** here

<table>
<thead>
<tr>
<th>Total defects</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A4636</td>
<td>131</td>
<td>37</td>
<td>21</td>
<td>28</td>
<td>45</td>
</tr>
<tr>
<td>A2524</td>
<td>86</td>
<td>20</td>
<td>24</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>A3713</td>
<td>75</td>
<td>17</td>
<td>13</td>
<td>18</td>
<td>27</td>
</tr>
<tr>
<td>A4452</td>
<td>73</td>
<td>5</td>
<td>33</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>A4088</td>
<td>72</td>
<td>14</td>
<td>16</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>A2103</td>
<td>68</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>A2156</td>
<td>68</td>
<td>16</td>
<td>13</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>A3681</td>
<td>66</td>
<td>12</td>
<td>16</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>A1366</td>
<td>50</td>
<td>11</td>
<td>15</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>A2610</td>
<td>39</td>
<td>5</td>
<td>7</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>728</td>
<td>151</td>
<td>171</td>
<td>162</td>
<td>7</td>
</tr>
</tbody>
</table>

- Which defects cost us the most money?
Tables

- Defect frequency
  - If 1850 lots of grade A4636 (first row): defect A rate $= 1/50$
  - If 250 lots of grade A2610 (last row): defect A rate $= 1/50$
  - Redraw table on production rate basis

- If comparing defects over different grades: go down the table (show fraction within the column)

- If comparing defects within grade: go across table (show fraction with the row)
  - Could weight each column by cost of defect
Tables

Common pitfalls: using pie charts when tables will do

I cannot explain the pitfalls of pie charts as well as Stephen Few does: **Save the pies for dessert** (please read)
Tables vs pie charts: plenty of bad examples

Composition of total expenses, 2010-11

- **Children's and Social Services Sector**: 11% ($13.9 B)
- **Justice Sector**: 3% ($4.4 B)
- **Postsecondary and Training Sector**: 6% ($8.1 B)
- **Other programs**: 17% ($22.0 B)
- **Education Sector**: 17% ($21.4 B)
- **Health Sector**: 37% ($46.1 B)
- **Interest on Debt**: 8% ($10.0 B)

*Excludes Teachers' Pension Plan.

Note: numbers may not add to 100 due to rounding

CARRIE COCKBURN/THE GLOBE AND MAIL  SOURCE: ONTARIO MINISTRY OF FINANCE

---

**Equipment**
- Equipment (88%)
- Site Preparation (12%)
2. Arbitrarily ordering of the rows

<table>
<thead>
<tr>
<th>Model</th>
<th>Bank loan monthly payments</th>
<th>Monthly lease payment</th>
<th>Minimum downpayment for lease</th>
<th>Total interest paid over 48 months</th>
<th>Monthly insurance payment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ford Fusion</td>
<td>552</td>
<td>395</td>
<td>0</td>
<td>2,529</td>
<td>180</td>
</tr>
<tr>
<td>Honda Civic</td>
<td>538</td>
<td>424</td>
<td>0</td>
<td>2,466</td>
<td>236</td>
</tr>
<tr>
<td>Mazda 3</td>
<td>506</td>
<td>478</td>
<td>1,000</td>
<td>2,318</td>
<td>251</td>
</tr>
<tr>
<td>Toyota Yaris</td>
<td>435</td>
<td>490</td>
<td>1,000</td>
<td>1,992</td>
<td>198</td>
</tr>
<tr>
<td>VW Golf</td>
<td>596</td>
<td>550</td>
<td>2,500</td>
<td>2,730</td>
<td>244</td>
</tr>
</tbody>
</table>
3. using excessive grid lines

<table>
<thead>
<tr>
<th>Total defects</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A4636</td>
<td>131</td>
<td>37</td>
<td>21</td>
<td>28</td>
<td>45</td>
</tr>
<tr>
<td>A2524</td>
<td>86</td>
<td>20</td>
<td>24</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>A3713</td>
<td>75</td>
<td>17</td>
<td>13</td>
<td>18</td>
<td>27</td>
</tr>
<tr>
<td>A4452</td>
<td>73</td>
<td>5</td>
<td>33</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>A4088</td>
<td>72</td>
<td>14</td>
<td>16</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>A2103</td>
<td>68</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>A2156</td>
<td>68</td>
<td>16</td>
<td>13</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>A3681</td>
<td>66</td>
<td>12</td>
<td>16</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>A1366</td>
<td>50</td>
<td>11</td>
<td>15</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>A2610</td>
<td>39</td>
<td>5</td>
<td>7</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>728</td>
<td>151</td>
<td>171</td>
<td>162</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total defects</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A4636</td>
<td>131</td>
<td>37</td>
<td>21</td>
<td>28</td>
<td>45</td>
</tr>
<tr>
<td>A2524</td>
<td>86</td>
<td>20</td>
<td>24</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>A3713</td>
<td>75</td>
<td>17</td>
<td>13</td>
<td>18</td>
<td>27</td>
</tr>
<tr>
<td>A4452</td>
<td>73</td>
<td>5</td>
<td>33</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>A4088</td>
<td>72</td>
<td>14</td>
<td>16</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>A2103</td>
<td>68</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>A2156</td>
<td>68</td>
<td>16</td>
<td>13</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>A3681</td>
<td>66</td>
<td>12</td>
<td>16</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>A1366</td>
<td>50</td>
<td>11</td>
<td>15</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>A2610</td>
<td>39</td>
<td>5</td>
<td>7</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>728</td>
<td>151</td>
<td>171</td>
<td>162</td>
<td>7</td>
</tr>
</tbody>
</table>
## Tables

**Interesting example: comparing two treatments**

<table>
<thead>
<tr>
<th></th>
<th>Corrosion resistance</th>
<th>Surface roughness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>K135</td>
<td>0.3</td>
<td>0.22</td>
</tr>
<tr>
<td>K136</td>
<td>0.45</td>
<td>0.39</td>
</tr>
<tr>
<td>P271</td>
<td>0.22</td>
<td>0.24</td>
</tr>
<tr>
<td>P275</td>
<td>0.4</td>
<td>0.44</td>
</tr>
<tr>
<td>S561</td>
<td>0.56</td>
<td>0.36</td>
</tr>
<tr>
<td>S567</td>
<td>0.76</td>
<td>0.51</td>
</tr>
</tbody>
</table>

- Coating A or B are applied to different products
- K-series, P-series, S-series
- How does the coating affect corrosion and surface roughness?
Sankey diagrams

Canada's Energy Flow 2007

Unit: Petajoules (PJ)

I highly recommend reading Tufte’s 4 books: contain remarkable examples of how to bring data to life.
Colour

- Colour is effective, but:
  - readers could be colour-blind,
  - document read from a gray-scale print out

- There is **no standard colour progression** (blues, greens, yellows, orange, red).
- Safest colour progression is gray-scale axis: from black to white
  - satisfies colour-blind readers
  - looks good in printed form
General summary

No general advice that applies in every instance. Useful tips nevertheless:

- To understand causality, you must show causality: use bivariate scatter plots (sometimes line plots also work well)
- Plots and text go together: a plot = paragraph of text
  - add labels to plots for outliers and interesting points
  - add equations
  - add small summary tables
- Avoid codes: “A = grade TK133”, “B = grade RT231”
Avoid unnecessary “extras” to enliven the plot

“If the statistics are boring, then you’ve got the wrong numbers”.

But living in a rented semi-detached home with three college students means she’s eager to find her own space. She was also careful enough to save $40,000 for a down payment during her university years by running a College Pro Painters franchise.

Buyers today can get a variable-rate mortgage at prime or 2.25 per cent, and in many cases cheaper after discounting.

But even at the prime rate, it would cost only $1,347 to carry a $400,000 home with an amortization of 35 years and a 5 per cent down payment. By comparison, an average two-bedroom condo in the Toronto area costs $1,487 per month to rent.

That's a compelling reason for home ownership.
General summary

- Adjust for inflation if plot involves money and time
- Maximize the data-ink ratio $= \frac{\text{ink for data}}{\text{total ink for graphics}}$.
  1. eliminate non-data ink
  2. erase redundant data-ink.
- Maximize data density: 250 data points per linear inch, and 625 data points per square inch.
Good plotting is not difficult. It just takes time and thought.