Graphical & tabular methods of obtaining information from data

BEA140 Quantitative Methods - Module 2

UNIVERSITY *of* TASMANIA

Stem & leaf charts

1.5	2.9	0.8	4.2	3.5	2.1	1.8	4.2	2.6	1.3	3.1	2.0

stem	leaf	frequency	cumulative frequency
0	.8	1	1
1	.3 .5 .8	3	4
2	.0 .1 .6 .9	4	8
3	.1 .5	2	10
4	.2 .2	2	12

Stem & leaf charts allow one to visualise data and are useful to indicate range, concentration and structure of data. As data values are visible, it's possible to identify outliers (observations that deviate markedly from other members of the sample and which really ought to be examined/questioned, though should not be discarded without finding evidence that they are indeed measurement/recording errors rather than genuine data), data rigging and measurement errors (e.g. if too much but not all data ends in '0', one might wonder if the data was rounded inconsistently when recorded).

Frequency distribution tables

15	29	8	42	35	21	18	42	26	13	31	20

Frequency distribution tables group data for easier: analysis, interpretation, and subsequent presentation.

time	(absolute) frequency	class mark	relative frequency	cumulative frequency	cumulative percentage
	f_i	X_i	f_i/n	Σf_i	$(\Sigma f_i/n) * 100$
$0 \leq X < 10$	1	4.5	0.0833	1	8.33%
$10 \leq X < 20$	3	14.5	0.25	4	33.33%
$20 \leq X < 30$	4	24.5	0.3333	8	66.67%
$30 \leq X < 40$	2	34.5	0.1667	10	83.33%
$40 \leq X < 50$	2	44.5	0.1667	12	100.00%

Note: The class mark of a class is a representative value for all values within that class, and is calculated as the mean/average of the lowest and highest possible values within that class.

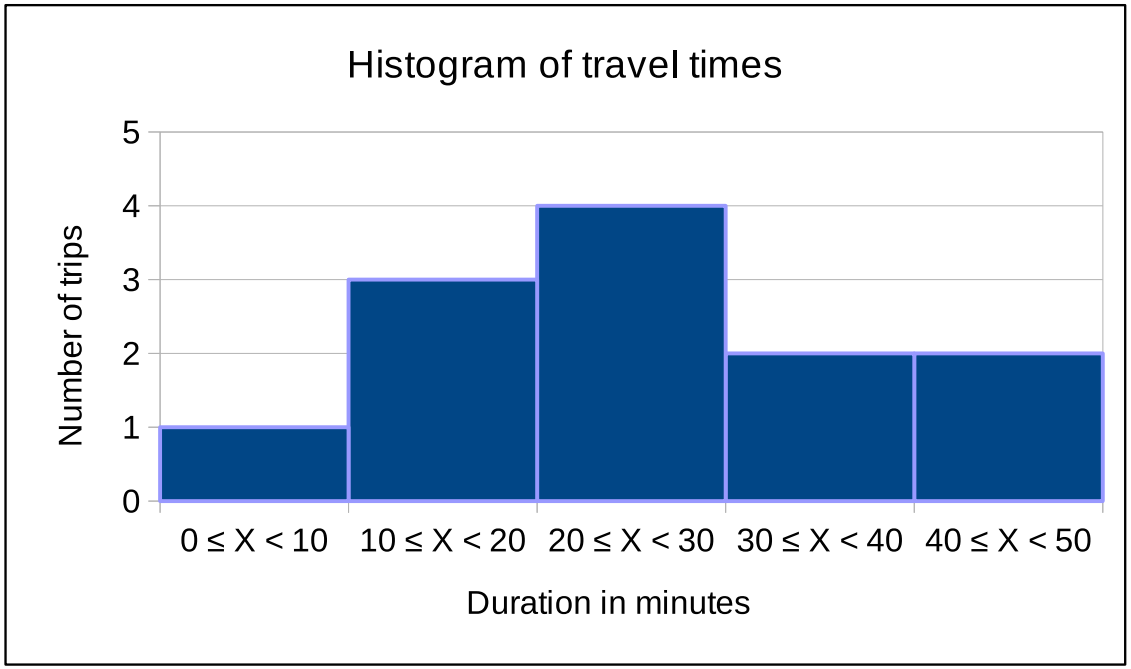
Notes on frequency distribution tables

- Classes are non-overlapping (i.e. are mutually exclusive) and cover all of the data (i.e. are collectively exhaustive);
- Class widths are uniform;
- It is unwise to leave "open classes", as it lessens the information content. E.g. a value in the open class $X > 50$ could be 51 or 251;
- The shaded part (last three columns) of the frequency distribution table on the previous slide is optional. If one intends to construct a relative frequency histogram or polygon based on the table, then one would usually include the relative frequency column. Likewise if you were intending to construct an ogive you would include the cumulative frequency column(s);

Notes on frequency distribution tables

- It is usually wise to use class widths that are 1, 2, 4 or 5 times a power of 10, the choice of which impacts on the number of classes; and
- When building a frequency distribution table there is no objectively best rule about the selection of class widths and the number of classes. If you choose too few classes the reader may not see important aspects of the structure, but if you choose too many the reader may be swamped with detail. Generally the more data you've got, the more classes you should have. Various authors have suggested some rules, including Sturges (recommended number of classes = ceiling(1.44 * ln(n) + 1)). Such rules should be thought of as no more than providing an indication or guideline. They can all provide poor guidance if the data is heavily skewed or contains outliers or the sample size is small (say $n < 30$). Applying Sturges' rule to our data gives recommended number of classes = ceiling(1.44 * ln(12) + 1) = ceiling(6.74) = 5. (by mere coincidence is equal to the number of classes we used).

Histograms



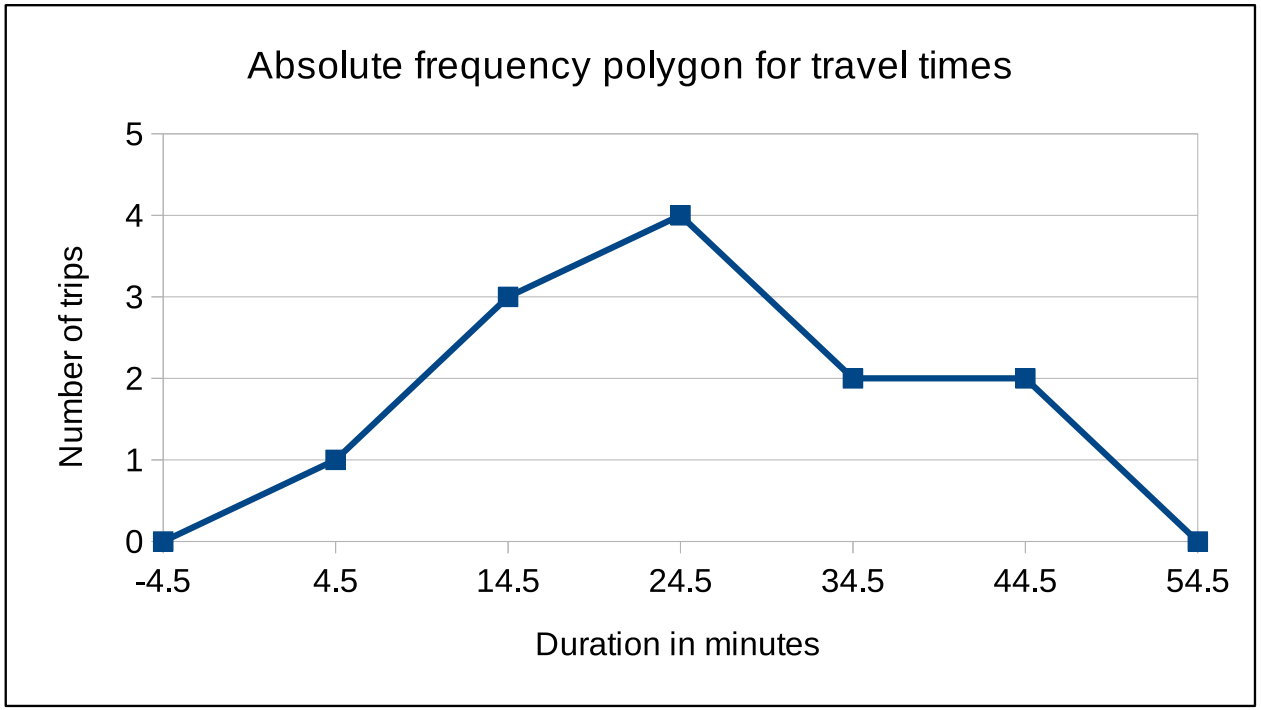
A **histogram** is a graphical presentation of a freq. dist. table as a set of joined equal-width bars, each of height equal to the frequency of that class. Note that histograms are most appropriate for continuous data, in which case the centre of each bar is the class mark.

Note: The horizontal axis of a histogram should show the class limits rather than class labels, however it has become acceptable to show class labels (as above) since it is much easier with software like Excel.

Stem & leaf plots vs. histograms

- Stem & leaf plots retain all information about the data values; whereas
- Histograms allows for considerably more flexibility in bar/class width.

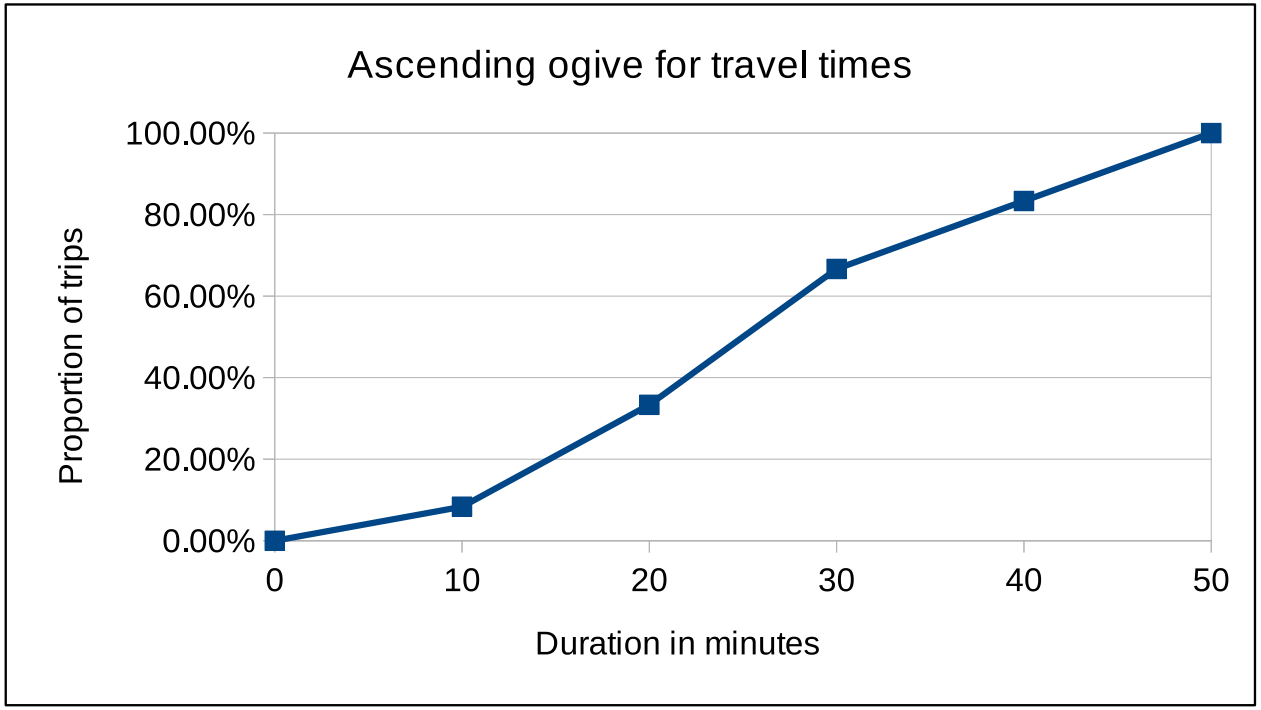
Frequency polygons



A **frequency polygon** has an unsmoothed line joining points on a histogram defined by class mark and frequency, with zero end classes.

Note: Frequency polygons essentially convey they same information as histograms, and may be presented in either absolute or frequency form.

Ascending ogives

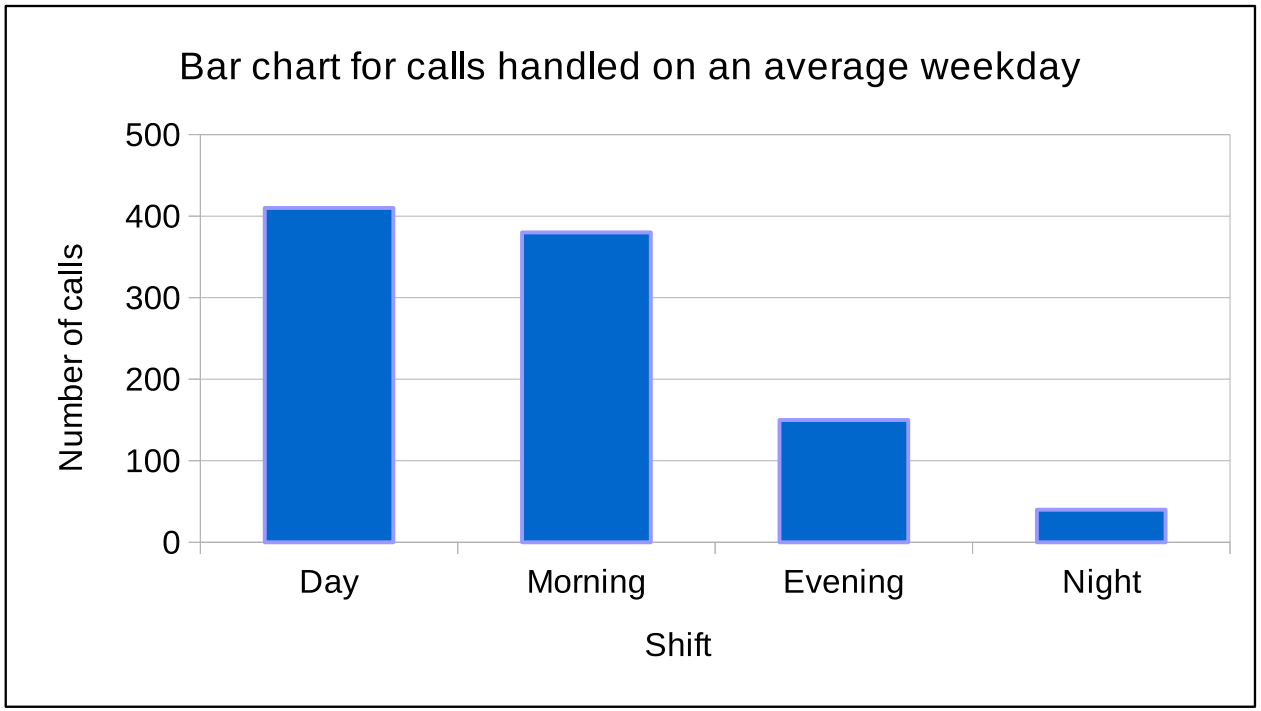


An **ascending ogive** is a graph of cumulative frequency distribution. It is sometimes used to help convey information on distribution tails. For example, to answer questions like:

(i) What proportion of trips took less than 30 minutes?

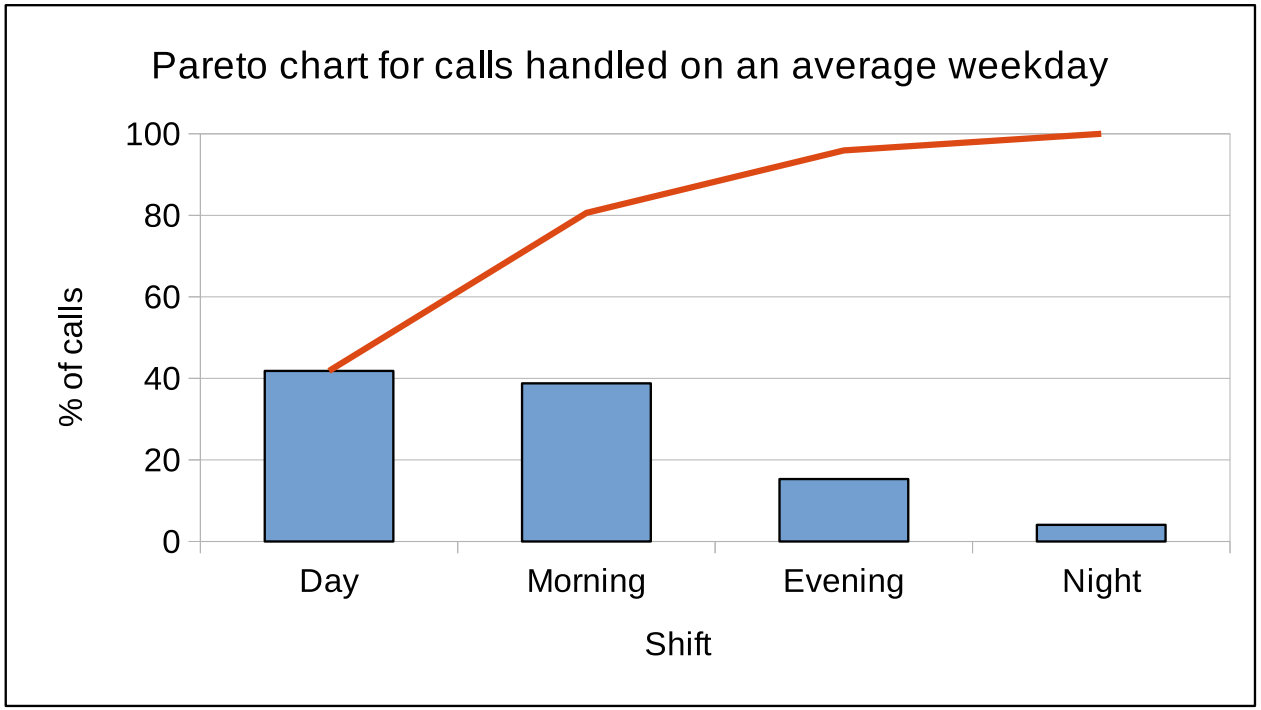
(ii) How long do the fastest 20% of trips take?

Bar charts



Note: Bar charts are appropriate for categorical data.

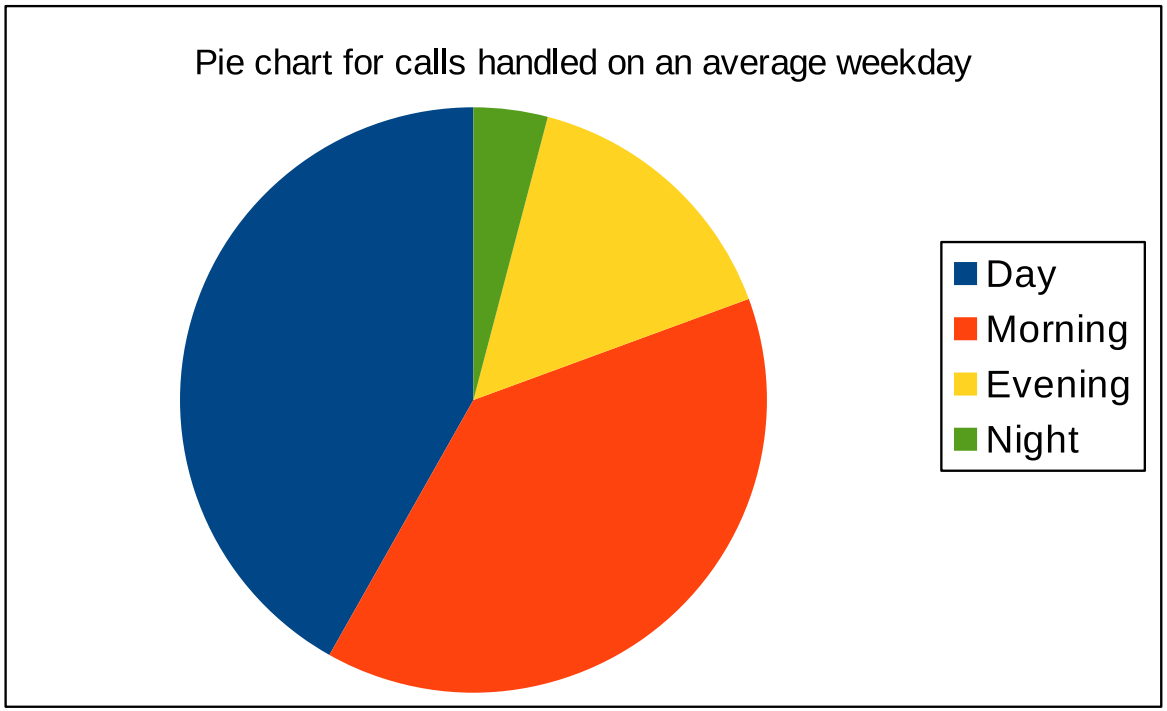
Pareto charts



Pareto charts are a useful variation on bar charts where: categories are ranked from most to least common, left to right across the chart; and a line shows the cumulative level.

Note: Pareto charts are a common tool in quality management, especially for analysing and communicating the magnitude of problems, and for prioritising action.

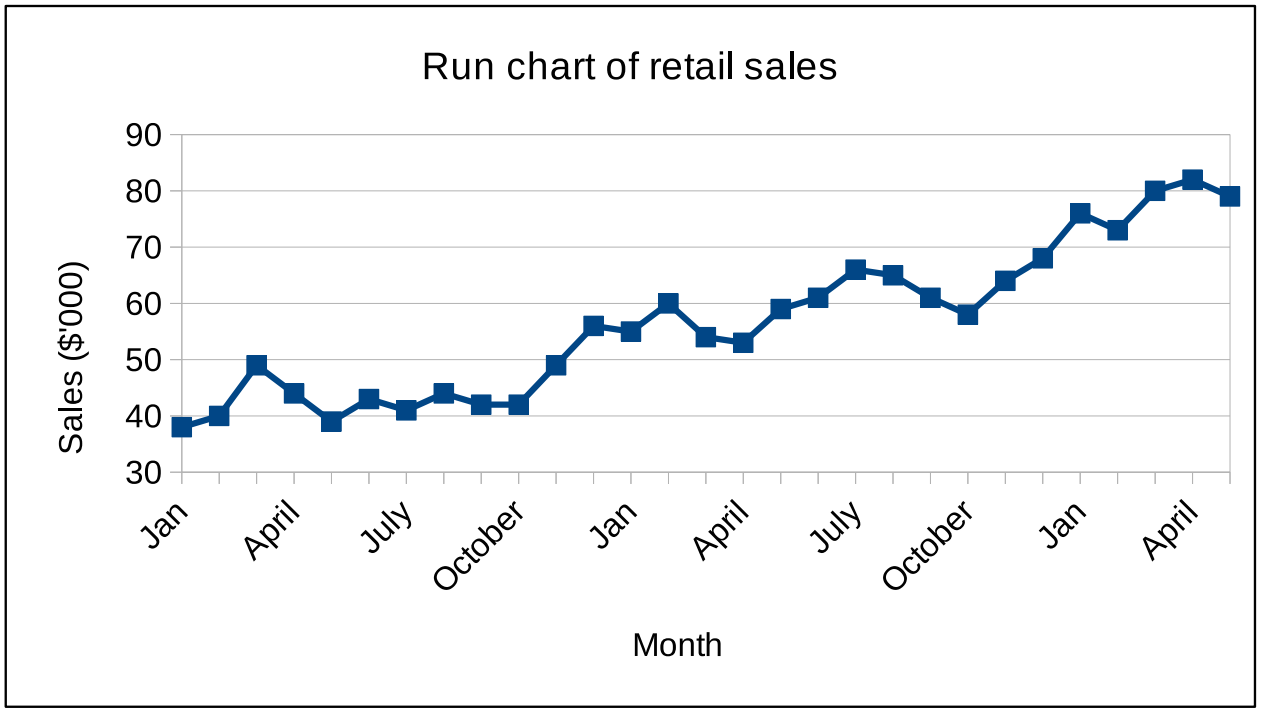
Pie charts



Pie charts show comparisons of proportions from each category.

Note: Pie charts can be good for low-tech audiences.

Run charts



Run charts are a good tool for illustrating one or more (numerical) variable(s) over time (or some other progression, for example the order in which the data was collected).

They can allow identification of trends and periodicity.

What makes a good chart/graph?

- A good chart/graph is clear, uncluttered and well labelled. It should not leave the reader trying to guess what it is about, and should give the reader a 'feel' for the data;
- Good charting/graphing technique requires that there is minimal risk that the reader of a chart/graph needs to make possibly incorrect assumptions. Every chart/graph should have a descriptive heading and labelled axes. In addition it is often useful to identify the data source in a footnote, and if more than one symbol has been used a clear legend should also be included; and
- Charts/graphs are both tools of analysis and tools of communication. That is, they not only allow one to better understand key features of the data, but they also allow one to communicate that to others. As the saying goes - "a picture paints a thousand words".

. . . that's it for now!