Data Visualisation
Section 1
Introduction

What is Visualisation?
Visualisation is commonly defined as the communicating of information using graphical means or representations. Pictures have been used for millenia as a method of communication of ideas and concepts long before the use of written languages; for example the "Cueva de las Manos" or Cave of Hands in Argentina contains images of prehistoric hands from between 9,000 and 13,000 years ago

Figure 1.1: Image of the hands in the Cueva de las Manos by Mariano

A single picture can contain a welath of information and, in part due to the parallel nature of our visual accuity, can be processed much faster than a comparable page of words or spoken instruction which are processed sequentially. The well known idiom "A picture is worth a thousand words" is proof, if any is needed, that the need to visualise ideas/concepts/instructions/information is nothing new but an idea that has dominated communication for a long time. Indeed, pictures can be used a source of common information transfer between different civilisations/races of people who have no language in common.

Examples of Visualisation in Everyday Life
Before considering, explicitly, the importance of visualisation in today's society, it is an interesting exercise to consider the varied types of data and visualisation typically encountered in everyday life:

- A train or bus map with interconnections and departure/arrival times shown.
- A weather chart showing the expected wind/precipitation/atmospheric pressure.
- A stock market graph showing how shares/exchange rates are varying.
- An exploded view of a piece of machinary showing the various spare parts and their stock numbers.
- A roadsign showing the condition of the road ahead/type of junction/expected dangers/etc.
- A road map to determine the itinerary for a journey.
- A simple formatted number used to represent time; a digital clock.
- A family tree showing the family connections between two or more individuals
- A computer icon on the desktop.

In each of these cases, visualisation has provided an alternative to, or a supplement for, textual or verbal communication. It should be evident, then, that visualisation provides a far richer description of the information encountered than the word-based counterpart. So the following questions are posed:

- Why is this so?
- What situations lend themselves to more effective visualisations?
- What types of information can be readily visualised and what types cannot?
- What are the different methods that can be employed to visualise data, and which ones are best for particular circumstances?

In this module, we shall attempt to answer these questions and consider the tools necessary to effectively visualise data appropriately.

The Importance of Visualisation.

The most obvious reason why visualisation is important to humans stems from the fact that humans have descended from hunter/gatherer species whose sight was one of the key senses for information processing.

The following examples should illustrate the importance of visualisation in decision making and the role of human preferences and training.

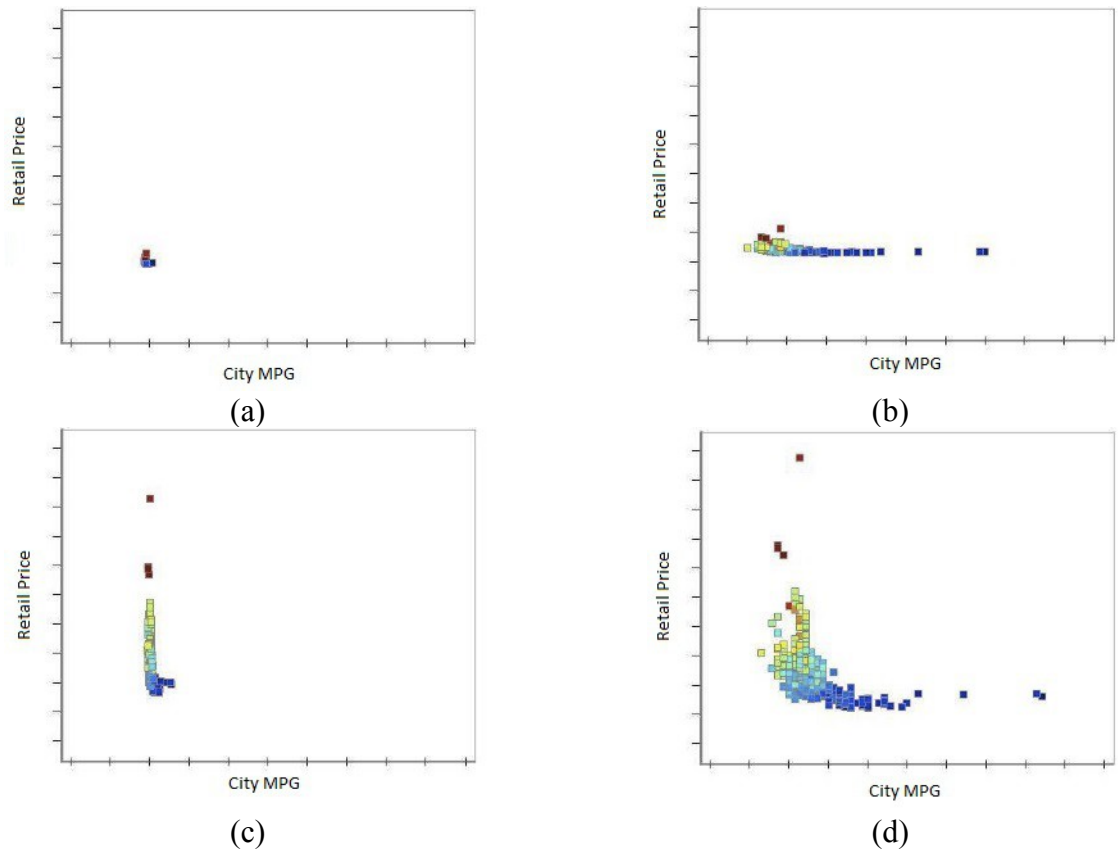

Figure 1.2: Four scatterplots of the same data but scaled differently. (a) both axes have same large scale. (b) large scale in the *y* direction only. (c) large scale in the *x* direction only. (d) both scaled according to data values.

In figure 1.2, four scatterplots of the same data set regarding the retail cost of a vehicle vs its urban fuel economy.

- In the first scatterplot, the scaling chosen was based on a lack of prior knowledge regarding the data set under consideration. The retail price range (*y*-axis) was chosen based on the

range of available vehicles in the market place and the MPG range was chosen based on the difference between the worst and best fuel economies of all available vehicles. The fundamental issue with this approach is whether the data being examined is a representative sample of the entire vehicle market or just a small portion. For example, if the data represented small cars best suited to urban environments, then it would be foolish to consider all other classes of vehicle, from motorcycles to articulated cabs, when scaling for retail prices and fuel economy. Perhaps this approach could be considered as a first iteration but as is evident from figure 1.2(a) the "blob" communicates only one message; scale the axes more appropriately.

- In the second plot, (b), the data was scaled on the complete market vehicle price range (as in (a)) in the *y*-direction but correctly scaled using the actual data set values in the *x*-direction. The improvement is evident horizontally but it is still cluttered vertically.
- In the third plot, (c), the data was scaled correctly in the *y*-direction but using the entire market fuel consumption in the *x*-direction. Not ideal.
- Finally, the fourth plot, (d), shows what happens when both axes are scaled appropriately based on the dataset itself and not on preconceptions. Decisions can now be made regarding the data using this plot.

In figure 1.3 below, the organisational chart (or tree) shows how the reporting structure of the managment in an organisation has been organised. It is evident from the chart that:

- The three VPs report directly to the Chairman
- The Marketing department has the most Consultants.
- The Driver has the most convoluted path back to the Chairman.
- Not all consultants are treated equally; the Sales consultant is one level higher then the others.

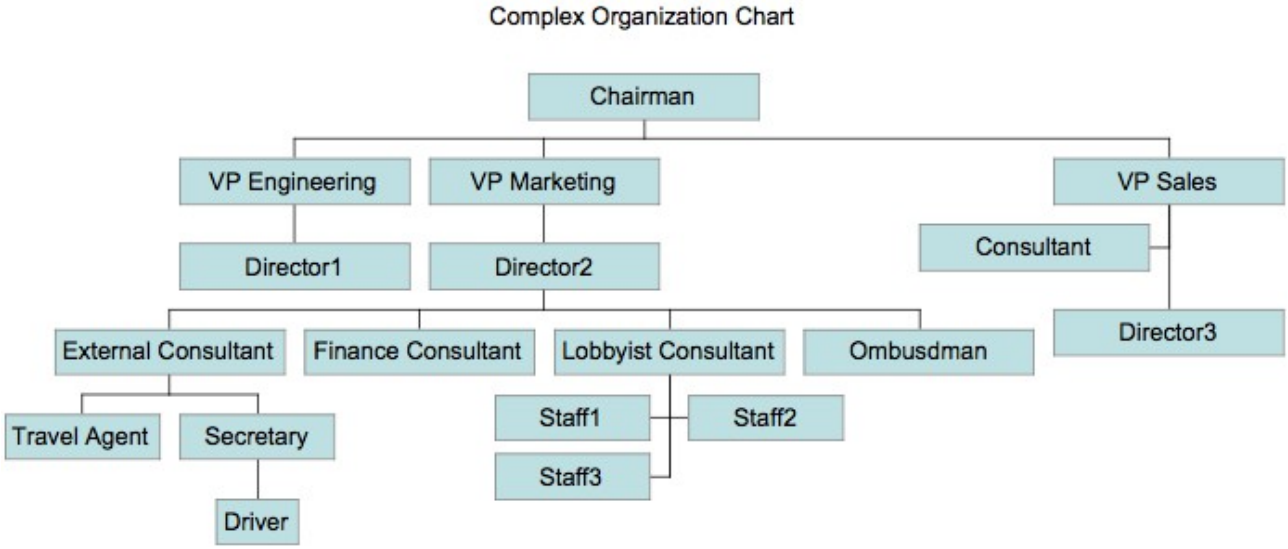

Figure 1.3: A sample organisational chart (tree). The structure is apparent when shown in this form. How difficult would it be to verbally describe this structure succintly and accurately?

The four bullet points above only partially expain the structure of the organisation above. To fully describe it would require a far more detailed and verbose description than is needed when the chart itself is employed.

In today's society, with its unparalleled accumulation of data, the accurate, precise, and concise presentation of this data is more critical than ever. Visualisation has become a cornerstone in the modern exploration, analysis, and dissemination of information. The ever increasing pressure to trawl through massive data sets to find that "nugget" of important/useful data has created a growing

need for tools and techniques to make effective use of this “information overflow”. In virtually every domain, visualisation is becoming an effective tool to assist in the analysis and communication of data/ideas/information.

A Brief History of Visualisation

As the title above suggests, a brief history of visualisation is presented here. Some of the major milestones are examined and their accumulative effect on how data is perceived today discussed.

Early man is credited with the first technique for graphically recording and presenting information. As mentioned at the beginning of this document, there have been cave drawings in existence for many thousands of years. In fact, one of the earliest is the Chauvet-Pont-d'Arc Cave near Vallon-Pont-d'Arc in France. The cave contains hundreds of paintings created approximately 30,000 years ago. While their actual significance has been lost, it is surmised that they were intended to pass on information to future generations.

Figure 1.4: A drawing from the Chauvet-Pont-d'Arc Cave in France

The oldest writing systems used pictures (similar to today's computer icons) to encode words and ideas. Such systems are termed “logograms”. Examples of logograms today would be the € symbol for “Euro”, the @ symbol for “at”, and the ampersand &. The earliest example of a written document is the Kish limestone tablet from Mesopotamia and while mostly pictographic (using logograms) it has the earliest beginnings of the syllabic script called cuneiform. The tablet is stored in the Ashmolean Museum in Oxford.

Figure 1.5: The kish tablet.

Another well known early writing system is from ancient Egypt; i.e. Hieroglyphics. This system contains three major categories: logograms, phonograms and determinatives. The hieroglyphic logogram is a sign/picture that represents the smallest language unit that carries a meaningful

interpretation.
Visualisation methods have developed out of necessity. They aided travel, commerce, religion, and communication. Maps provided information to travellers where planning and/or survival was key. An early example of one such map is the The Tabula Puetingeriana (Peutinger Table, Peutinger Map); an itineraria of the 70,000 miles of Roman road network. The section shown below in Figure 1.6 is from Rome (far left) to the tip of Italy (far right). Africa is at the bottom, the rest of Europe at the top, and the body of water is the Mediterranean Sea. This is an excellent example of a practical map based on the linear routes of Roman roads, but looks nothing like the accurate cartographic projection of the Mediterranean region. However, on closer look, some familiar geographic features can be seen such as the 'boot' of Italy and the island of Sicily on the right.

Figure 1.6: A section of the Tabula Peutingeriana

Such distorted maps arose mainly because relative positions are more important than actual accuracy. In some case, such as the Peutinger Map above, the medium chosen affects the manner of the distortions; the medium being papyrus rolled up to fit in a capsa (tool box) whose width would be restricted but whose length would not. The Peutinger Map is 6.75 meters long but only 0.336 meters wide.
Of course, maps can be used not just to illustrate relative positions but also to convey information about specific locations. Consider John Snow's map of cholera deaths in London in 1854. John Snow (15 March 1813 – 16 June 1858) was an English physician and is considered one of the fathers of modern epidemiology, primarily due to his work in tracing the source of a cholera outbreak in Soho, London, in 1854. His findings highlighted the need for, and ultimately brought about, fundamental changes in the water and waste systems of London, which then led to similar changes in other cities. He was a leader in the adoption of anaesthesia and medical hygiene and is credited with bringing about a significant improvement in general public health around the world.

(a) (b)

Figure 1.7: (a) The complete John Snow Cholera Map. (b) A section around the Broad Street Pump; the source of the infection.

As he was collating deaths in the outbreak, he noticed they were centered around a public water pump on Broad Street. He had the handle of the pump removed and within a short period of time the outbreak stopped; not before over 500 lives were lost. The mechanism he used to collate the data was to take a street map of the area and, for each death that occurred in a house, he drew a bar on the map. Bars were stacked for multiple deaths in a single house. In effect he constructed the first epidemiology map.

Spatial mapping is only one facet of visualisating data. Time series visualisations are also important. Abū al-Rayhān Muhammad ibn Ahmad al-Bīrūnī (who will be referred to from now on as Al-Biruni) completed an extensive astronomical encyclopaedia in the late 10th – early 11th century and one of the most famous illustrations is his depiction of the phases of the moon in orbit shown below in figure 1.8.

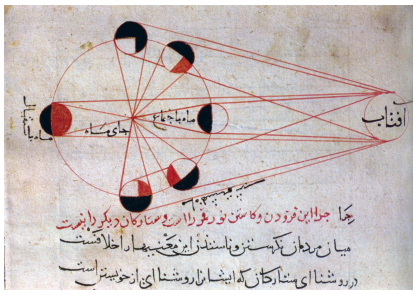

Figure 1.8: Al-Biruni's time series visualisation of the moon phases.

Another excellent illustration of a time series is the Napoleonic March on Moscow by Charles Joseph Minard; a pioneer in the field of information graphics in civil engineering and statistics. He is especially noted for his use of numerical data on geographical maps. The map is a superb example of how, through the clever use of graphics, a wealth of information can be conveyed efficiently and succintly.

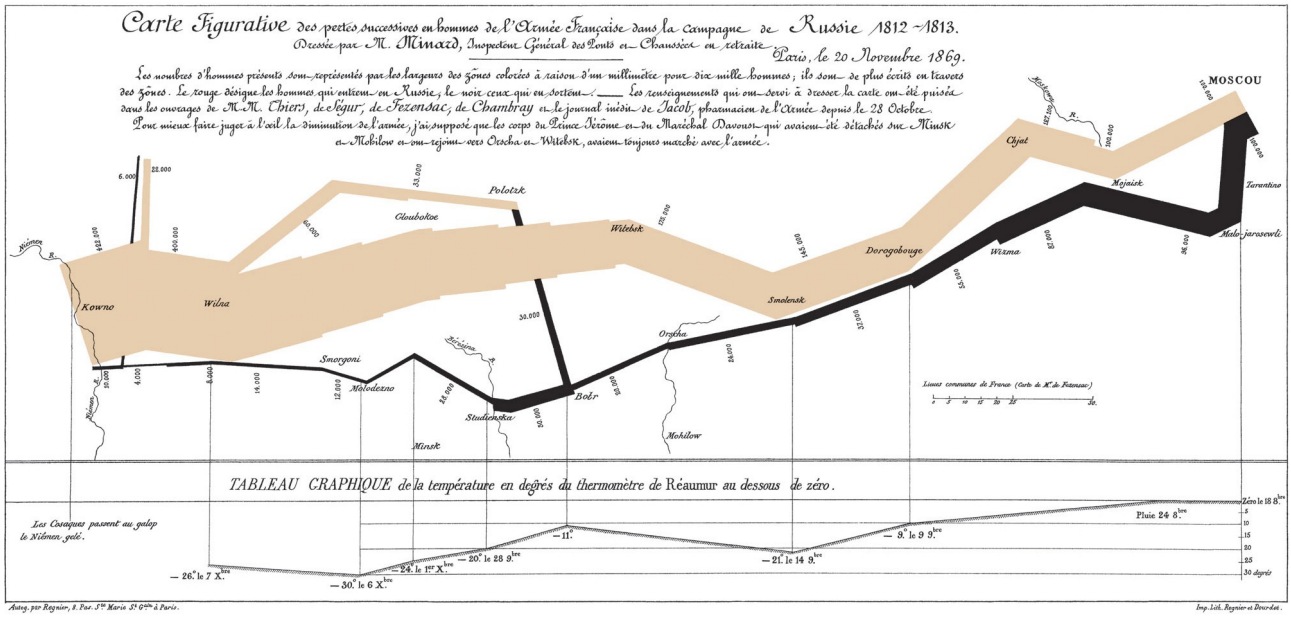

Figure 1.9: Minard's Napoleonic Map of the March on Moscow

In his map, shown above in figure 1.9, Minard uses the horizontal axis to plot distance from the embarkation point on the left to Moscow on the right. The thickness of the line represents the size of the army, initally 422,000 strong, which thins to 100,000 at Moscow. The advance to Moscow is shown in beige and the retreat from Moscow in black. As can be seen from the figure, the final size of the army is 10,000. The temperature along the march is mapped on a separate scale below the primary map but is matched to the locations on the primary.

Minard, through his clever use of graphics, has managed to convey the scope of the catastrophe of this campaign in a simple graph. Methods derived from his work are used today in conveying as much information as possible in a single illustration.
While the representation of data using axes which have physical meaning (time, distance, temperature, density, etc.) is useful, the introduction of abstract mathematical interpretation to the visualisation of data was a significant breakthrough. The most commonly cited examples are those attributable to Florenece Nightingale, the lady with the lamp. She pioneered a method of visualising that bore no resemblence to the geospatial/temporal/physical axes methods discussed previously. She endeavoured to illustrate as clearly as possible the extent of the misery the allied soldiers, in what became known as the Crimean War, were exposed to.

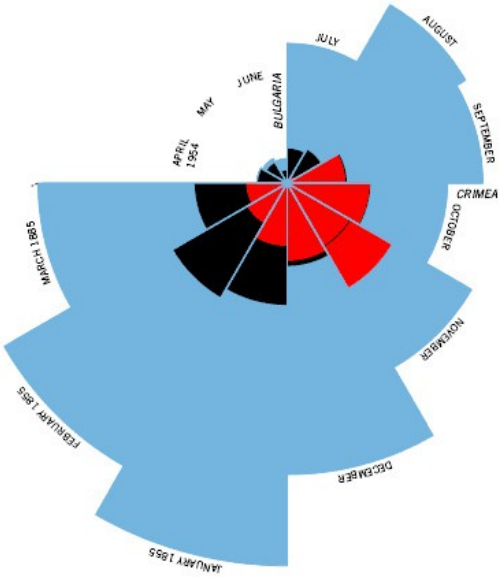

Figure 1.10: Florence Nightingale's abstract method of representing monthly deaths.

She did this by discarding axes completely and by representing a year as a circle subdivided into twelve parts. The size of each monthly slice was based on the number of deaths that occurred in that month. Therefore more real estate was given to higher death rate months than lower ones. This is illustrated above in Figure 1.10. The colours also were given meaning. Blue represents death by disease, red represents deaths from wounds, and black is deaths from all other sources. Thus it was easy for her to report that the greatest number of deaths occurred not from battle but from lack of proper winter clothing during the harsh Russian winter. A fact that contributed, along with William Howard Russell's stark reporting, to the bringing down of the British government of the day.

Visualisation Today

No brief introduction to the area of data visualisation would be complete without considering how visualisation is used in today's society.
One common problem with today's society is that of commuter travel. Up to the Industrial Revolution, travel was the prerogative of the priviliged classes; e.g. the Grand Tour of wealthy English gentry in the eighteenth century. Post Industrial Revolution, the population shifted from local hamlets to large industrial towns to be near the workplace; commuting then was a walk to nearby factories. With the advent of personal transport, such as the automobile, the distance one could travel to get to work increased and so the suburbes were created. Of course, congestion on inadequate roads led to investments in public transport links and buses, railways, light railways, trams, etc. became popular again.
In light of this, the need to accurately display information about travel links in a way that commuters can see what they need to see quickly, with little margin for error, led to the creation of the modern transport map; the first example of which was the London Underground Map (or Tube

Map) created by Harry Beck in 1931. He saw, just as was done in the Peutinger Map, that the relative positions of stations on a line was important and not their actual locations. He distorted distance so that the crowded inner city portion was exploded on the map with the sparse suburbs on the periphery contracted.

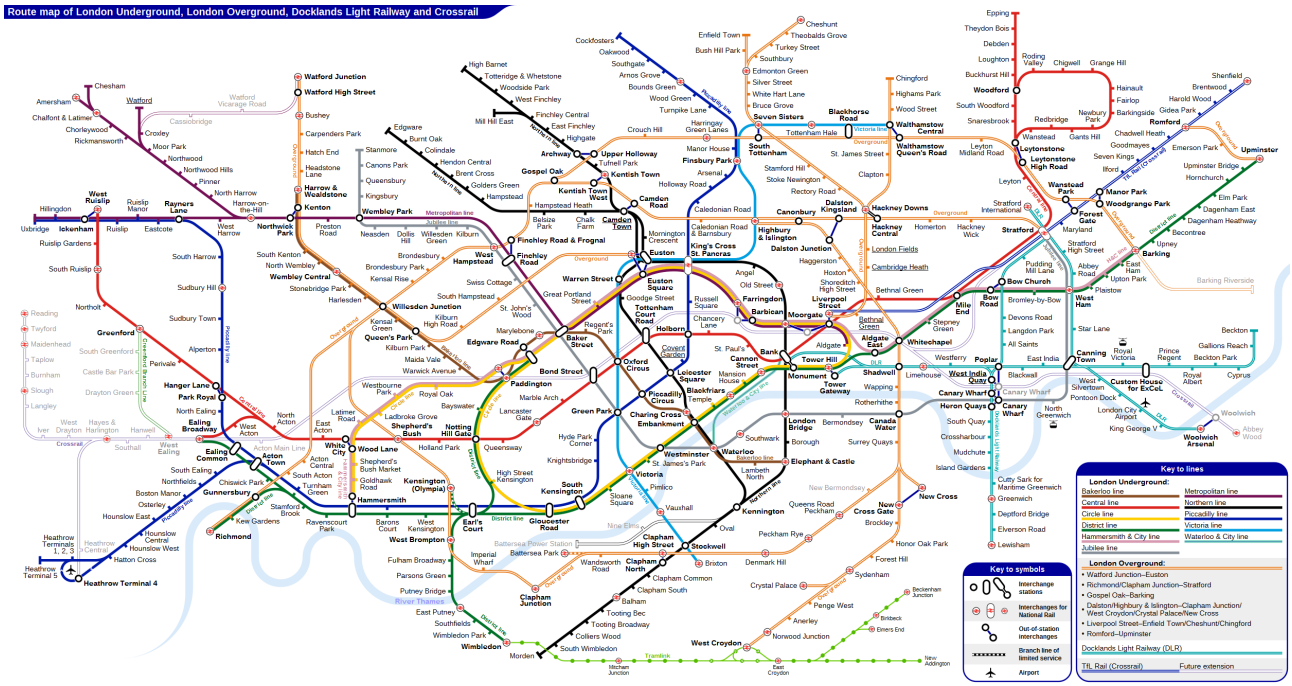

Figure 1.11: The current version of the London Underground Map

When examining distorted views, it is possible for an imprecise interpretation to be made. For example, it is wrong for someone to think that, say, Elephant and Castle on the Northern Line in the map above is closer to Lambeth North than Kennington. However in reality they are not. Such distortions are fine for the map above where interconnection information only is required.

For a street map, accuracy is paramount and distortions are undesirable. Given that maps are 2D and the Earth is a 3D oblate spheroid, distortions are to be expected. However, over a small enough area of the planet's surface, the 2D approximation will be accurate enough for anyone's needs. Consider the street map of Dublin shown below, courtesy of Google Maps.

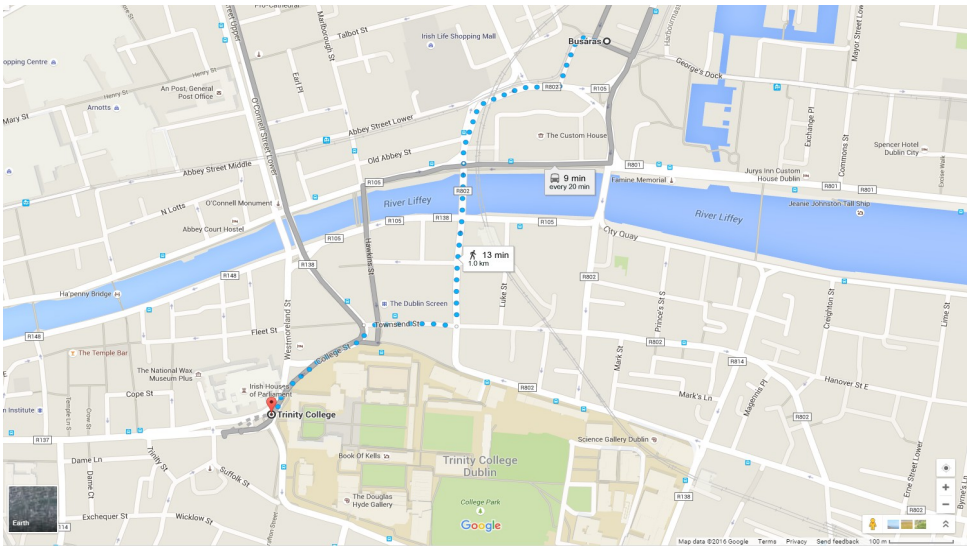

Figure 1.12: Walking directions from Busarus to Trinity College as determined by Google Maps. (Image (c) 2016 Google).

The street map is accurate enough for a pedestrian to use it to walk from Busarus (on Store Street)

to the main entrance to Trinity College (on College Green) following the trail laid out and choosing the correct streets to cross, turns to take and bridges to cross.
It should be apparent that, with maps as with other data sets, it is sometimes appropriate to distort the relative positions of places/objects to display information and at other times it is not.

It is possible for visualisations to provide very precise and specific pieces of information. For example, a digital clock would provide the time at an instant and nothing else. Visualisation is not restricted to graphics but does include numerals and letters which are, in themselves, symbols used to represent an idea and grouped together can be used to convey much more. A famous digital number is that of the US National Debt. The running total is, at 15:08 on 17th January 2016,

$18,901,441,441,271

amounting to $58,545 per citizen in the US.
The figure below is a snapshot from the official website showing just how bad the debt is

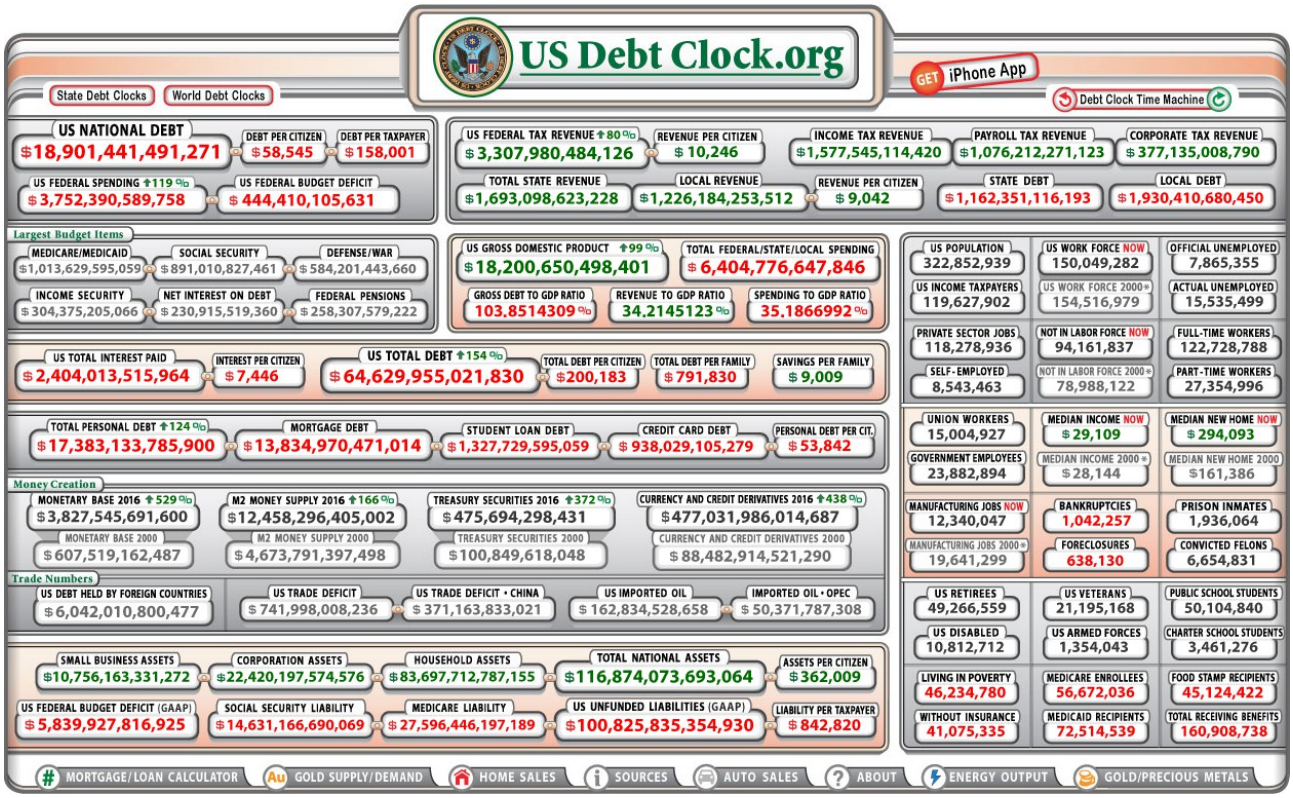

Figure 1.13 The US National Debt (as on 17/01/16).

The ability to provide rich descriptions of data is one of the key strengths of visualisation. In the early seventies, the idea of using graphical methods had been established relatively recently by John Tukey, but there was evidently still a lot of skepticism. In 1973, Francis J. Anscombe published a paper titled, *Graphs in Statistical Analysis*. In it, Anscombe first listed some notions that textbooks were "*indoctrinating*" people with, like the idea that

"numerical calculations are exact, but graphs are rough."

He then presented a table of numbers. It contained four distinct datasets (hence the name *Anscombe's Quartet*), each with statistical properties that were essentially identical:

> The mean of the *x*-values was 9.0. The mean of *y*-values was 7.5. They all had nearly identical variances, correlations, and regression lines (to at least two decimal places).

The table overleaf shows the four distinct data sets used by Anscombe and their scatterplots are shown in Figure 1.14.

Table 1.1 Anscombe's Quartet's Data Sets.

I		II		III		IV	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

The four scatterplots below in Figure 1.14 clearly show that the four data sets bear no resemblance to each other. Yet, their principle statistical measures are identical to two decimal places. Thare are listed in Table 1.2.

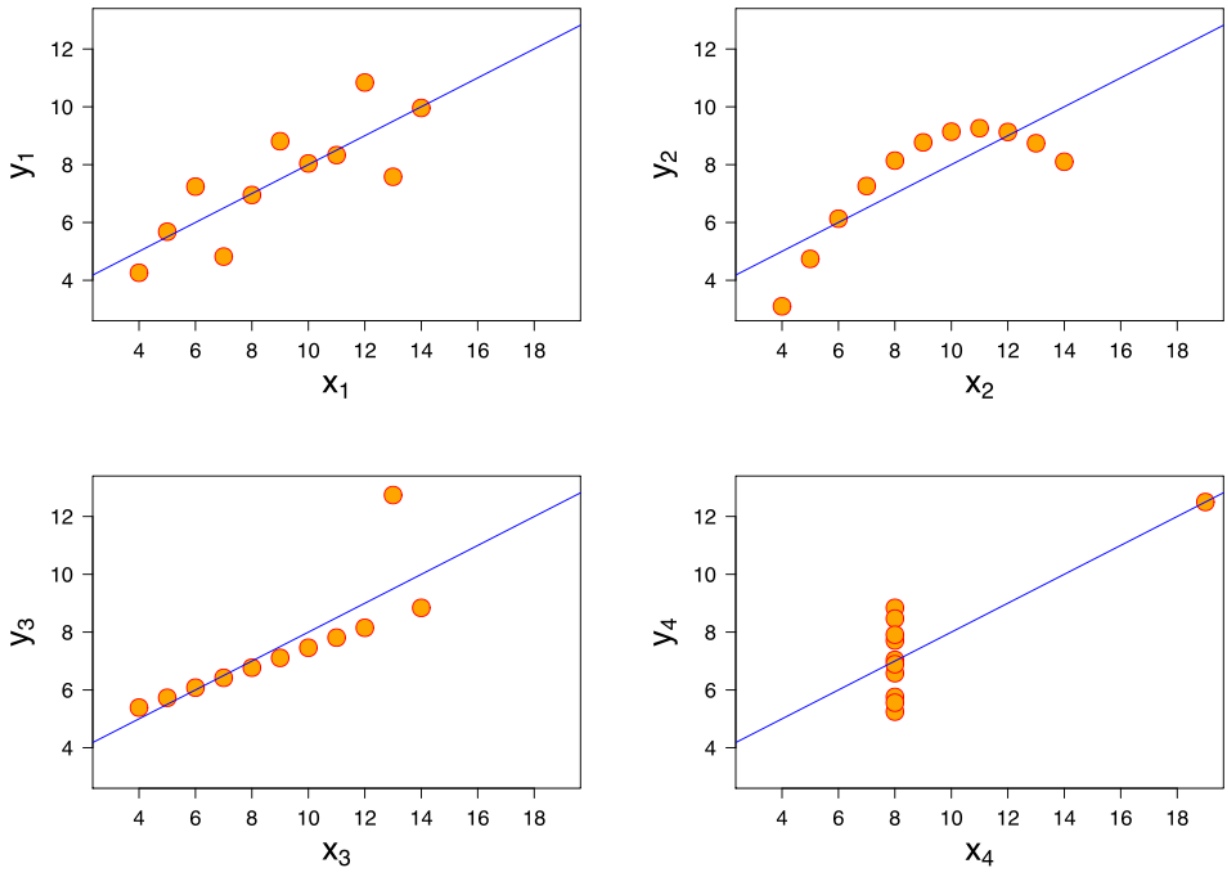

Figure 1.14 Anscombe's Quartet

This example by Anscombe clearly demonstrated that dealing purely with statistical measures is not

enough to ensure uniqueness of the data. In fact, the four-fold degeneracy in the Anscombe Quartet data sets clearly make the case, if one needed to be made, that visualisation should be employed as often as possible to avoid misconceptions and blatant errors.

Table 1.2 Statistics of Anscombe's Quartet.

Property	Value
Mean of x	9
Variance of x	11
Mean of y	7.5
Variance of y	4.1
Correlation	0.82
Linear Regression	$y = 3 + 0.5\,x$

Network representations are becoming increasingly popular as they can represent traffic patterns, computer communications, e-mail exchanges, social media relationships (e.g. friends), and many more. Figure 1.15 shows what could be euphemistically termed “The network of Medicine”

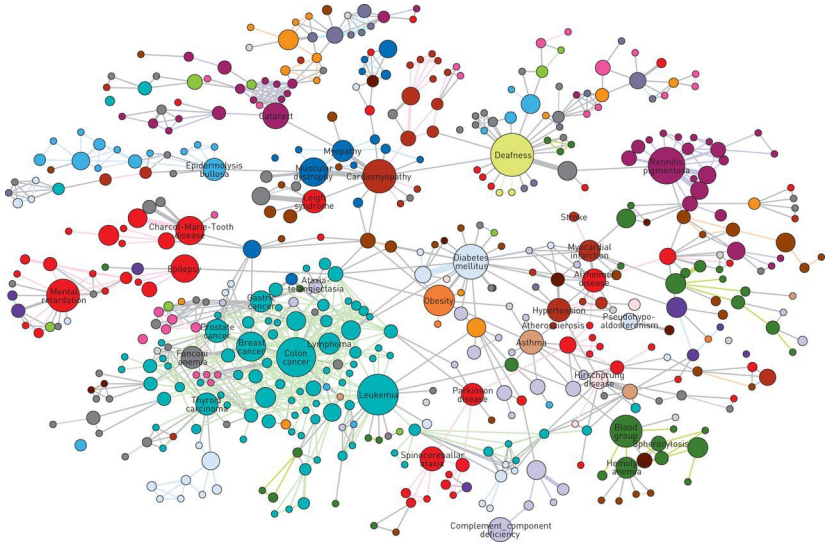

Figure 1.15: Network medicine. From Cancer to Cataracts and everything in between.

This network graph uses colours to separate groups of similarity; i.e. dark green represents blood, blue-green cancer, yellow deafness/hearing, red the heart, purple the eyes/sight, etc. Their relative positions indicate their proximity to other ailments. So it is possible to see interconnections between various groups.

More discussion on some of the topics covered in this section will be done later. In the next section the relationship between Visualisation and other fields is considered.