Chapter 1

Visualisation

1.1 Introduction

Visualisation is a relatively new term which describes the process of representing information or ideas by diagrams or graphs. There can be a number of reasons why such a representation might be produced. One common use of graphical representations is to communicate information. This relies on a careful choice of a display which highlights the facts known to be present in a given set of information so that others can be made aware of them. A second common use of graphics is to discover unknown features which might be hidden in a set of information. Such "exploratory" use of graphics is a relatively recent innovation; made possible by the development and widespread availability of computer graphics. A third related use of graphics is to try to gain a deeper insight into some familiar situation.

Although these three uses of pictorial representations are quite different, they use many of the same tools, because revealing information to ourselves is not all that different from revealing it to others.

Pictorial representations are attractive because our visual processing systems are very highly developed and we can process very large amounts of information when it is presented to us in a visual form. That said, it is important to recognise that our visual systems have not been designed to handle modern-day problems. Discovering important structure in the corporate spreadsheet is a very different problem from those encountered in the kind of hunting and gathering which has gotten our species by for most of its existence.

In the next three sections we'll show examples of the use of graphics for both presentation and exploratory purposes.

1.2 Graphics for Communication

In June of 1812, Napoleon began his fatal Russian campaign, a landmark in the history of the destructive potential of warfare. Virtually all of continental Europe was under his control, and the invasion of Russia was an attempt to force Tsar Alexander I to submit once again to the terms of a treaty that Napoleon had imposed upon him four years earlier. Having gathered nearly half a million soldiers, from France as well as all of the vassal states of Europe, Napoleon entered Russia at the head of the largest army ever seen. The Russians, under Marshal Kutuzov, could not realistically hope

to defeat him in a direct confrontation. Instead, they begin a defensive campaign of strategic retreat, devastating the land as they fell back and harassing the flanks of the French. As the summer wore on, Napoleon's massive supply lines were stretched ever thinner, and his force began to decline. By September, without having engaged in a single pitched battle, the French Army had been reduced by more than two thirds from fatigue, hunger, desertion, and raids by Russian forces.

Nonetheless, it was clear that unless the Russians engaged the French Army in a major battle, Moscow would be Napoleon's in a matter of weeks. The Tsar insisted upon an engagement, and on September 7, with winter closing in and the French army only 110 km from the city, the two armies met at Borodino Field. By the end of the day, 108,000 men had died, but neither side had gained a decisive victory. Kutuzov realized that any further defence of the city would be senseless, and he withdrew his forces, prompting the citizens of Moscow to began a massive and panicked exodus. When Napoleon's army arrived on September 14, they found a city depopulated and bereft of supplies, a meagre comfort in the face of the oncoming winter. To make matters much, much worse, fires broke out in the city that night, and by the next day the French were lacking shelter as well.

After waiting in vain for Tsar Alexander to offer to negotiate, Napoleon ordered his troops to begin the march home. Because the route south was blocked by Kutuzov's forces (and the French were in no shape for a battle) the retreat retraced the long, devastated route of the invasion. Having waited until mid-October to depart, the exhausted French army soon found itself in the midst of winter; in fact, in the midst of an unusually early and especially cold winter. Temperatures soon dropped to well below freezing, cossacks attacked stragglers and isolated units, food was almost non-existent, and the march was eight hundred kilometres. Ten thousand men survived.

Nearly 50 years later, in 1861, the French engineer Charles Joseph Minard (1781-1870) created a map showing a graphical representation of the campaign. A reproduction of the map as it appears in Tufte [10] is given as figure 1.1. The map was described by E. J. Marey [7] as "seeming to defy the pen of the historian by its brutal eloquence". It is generally recognised as one of the best graphical displays every created.

Minard's map is remarkable because it condenses a large amount of information into a very simple form. It shows how the size of the army changed over time and how the temperature varied during the retreat. In addition, it is possible to see the effect of individual battles; the disastrous crossing of the Berezina river is particularly apparent. Historical accounts give long and vivid descriptions of the campaign, but none is able to reduce it to such a brutal and clear summary.

Graphs, such as the Minard map, communicate information in a very effective way. A large amount is known about Napoleon's campaign; many books been written about it, filled with minute detail. By reading these books it is possible to gain a deep understanding of the 1812 campaign, but few people bother to do so because it takes too much time and effort. On the other hand, a quick look at Minard's map, together with a short explanation of the context, will give most people a a feeling for the fate of "La Grande Armée".

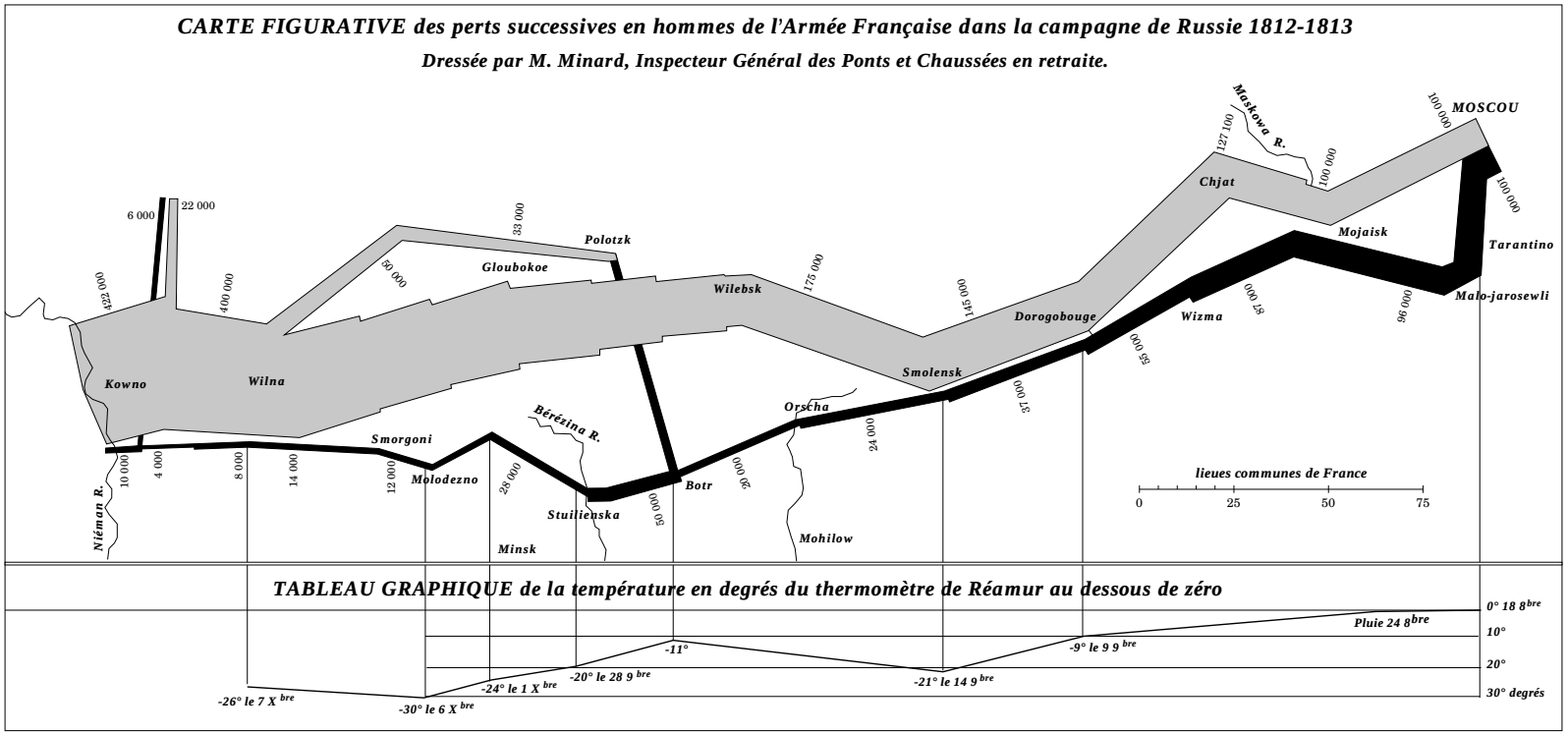


Figure 1.1: Charles Joseph Minard's map of Napoleon's 1812 campaign.

1.3 Graphics for Discovery

Forms of lotto are played worldwide and lots of people have theories about how to make money at the game. In this section we'll examine some information about a particular lotto game, and see whether it might be possible to play it profitably.

The game we'll look at is the daily "pick it" lottery run by the state of New Jersey in the USA. In the game, each player selects a three digit number between 000 and 999. A winning number is selected by picking three digits at random by picking a one of 10 balls randomly from each of three containers which hold the numbers $0, \ldots, 9$. All players who hold the winning numbers split the prize money for the game, so the size of the prize depends on the number of players who choose the winning numbers.

The state makes the results of the games public so it is possible to study the game and see if there is a good strategy for playing it. Here are the winning-number/payoff results for 254 consecutive games of "Pick-It" lotto.

(810, $190.0), (156, $120.5), (140, $285.5), (542, $184.0), (507, $384.5),
(972, $324.5), (431, $114.0), (981, $506.5), (865, $290.0), (499, $869.5),
(020, $668.5), (123, $83.0), (356, $188.0), (015, $449.0), (011, $289.5),
(160, $212.0), (507, $466.0), (779, $548.5), (286, $260.0), (268, $300.5),
(698, $556.5), (640, $371.5), (136, $112.5), (854, $254.5), (069, $368.0),
(199, $510.0), (413, $102.0), (192, $206.5), (602, $261.5), (987, $361.0),
(112, $167.5), (245, $187.0), (174, $146.5), (913, $205.0), (828, $348.5),
(539, $283.5), (434, $447.0), (357, $102.5), (178, $219.0), (198, $292.5),
(406, $343.0), (079, $332.5), (034, $532.5), (089, $445.5), (257, $127.0),
(662, $557.5), (524, $203.5), (809, $373.5), (527, $142.0), (257, $230.5),
(008, $482.5), (446, $512.5), (440, $330.0), (781, $273.0), (615, $171.0),
(231, $178.0), (580, $463.5), (987, $476.0), (391, $290.0), (267, $176.0),
(808, $195.0), (258, $159.5), (479, $296.0), (516, $177.5), (964, $406.0),
(742, $182.0), (537, $164.5), (275, $137.0), (112, $191.0), (230, $298.0),
(310, $110.0), (335, $353.0), (238, $192.5), (294, $308.5), (854, $287.0),
(309, $203.5), (026, $377.5), (960, $211.5), (200, $342.0), (604, $259.0),
(841, $231.0), (659, $348.0), (735, $159.0), (105, $130.5), (254, $176.0),
(117, $128.5), (751, $159.0), (781, $290.0), (937, $335.0), (020, $514.0),
(348, $191.0), (653, $304.5), (410, $167.0), (468, $257.0), (077, $640.0),
(921, $142.0), (314, $146.0), (683, $356.0), (000, $96.0), (963, $295.0),
(122, $237.0), (018, $312.5), (827, $215.0), (661, $442.5), (918, $127.0),
(110, $127.0), (767, $756.0), (761, $228.5), (305, $132.0), (485, $256.0),
(008, $374.5), (808, $262.5), (648, $286.5), (508, $264.0), (684, $380.5),
(879, $357.5), (067, $478.5), (282, $511.5), (928, $218.0), (733, $353.0),
(518, $162.5), (441, $184.0), (661, $548.0), (219, $166.5), (310, $147.5),
(771, $240.0), (906, $386.0), (235, $130.5), (396, $287.5), (223, $230.0),
(695, $480.5), (499, $247.5), (042, $380.0), (230, $238.5), (623, $237.5),
(300, $214.5), (380, $394.5), (646, $416.5), (553, $392.5), (182, $244.5),
(158, $202.0), (744, $371.5), (894, $553.0), (689, $293.5), (978, $295.0),
(314, $178.0), (337, $334.5), (226, $226.0), (106, $194.0), (299, $388.5),
(947, $353.0), (896, $404.0), (863, $348.0), (239, $163.5), (180, $216.5),
(764, $283.0), (849, $388.5), (087, $567.5), (975, $250.5), (092, $478.0),
(701, $267.5), (402, $326.5), (001, $369.0), (884, $512.5), (750, $341.0),
(236, $188.5), (395, $386.0), (999, $239.0), (744, $480.5), (714, $105.0),
(253, $227.0), (711, $130.5), (863, $384.5), (496, $294.5), (214, $154.0),

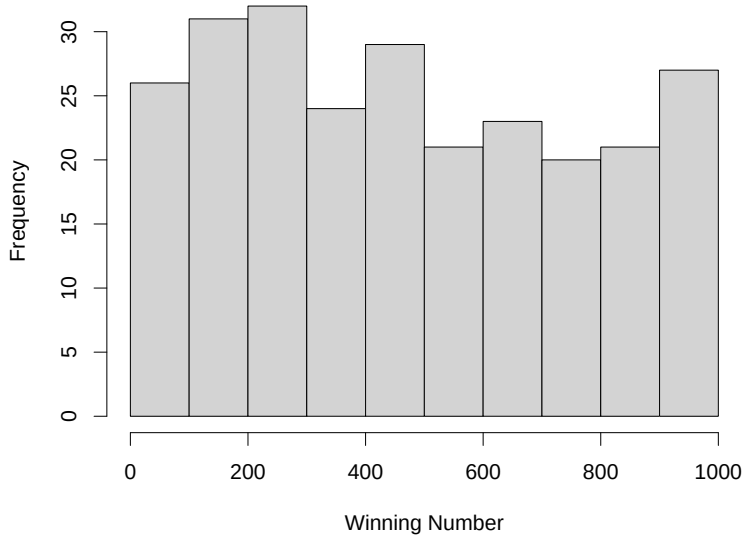


Figure 1.2: The distribution of the winning numbers for the "Pick-It" lottery.

(430, $324.0), (107, $116.0), (781, $229.0), (954, $301.5), (941, $334.0),
(416, $143.5), (243, $212.0), (480, $448.0), (111, $126.5), (047, $417.5),
(691, $276.5), (616, $303.0), (253, $211.0), (477, $373.0), (011, $209.5),
(114, $207.5), (133, $195.0), (293, $317.0), (812, $170.5), (197, $230.0),
(358, $143.0), (007, $361.0), (996, $452.0), (842, $260.5), (255, $308.5),
(374, $206.0), (693, $256.5), (383, $291.0), (099, $421.5), (474, $295.5),
(333, $119.5), (467, $268.5), (515, $221.0), (357, $151.5), (694, $314.5),
(919, $313.5), (424, $323.5), (274, $204.0), (913, $241.0), (919, $637.0),
(245, $214.0), (964, $348.0), (472, $191.5), (935, $384.0), (434, $220.0),
(170, $285.5), (300, $335.0), (476, $251.5), (528, $131.5), (403, $328.0),
(677, $392.0), (559, $509.0), (187, $235.5), (652, $249.5), (319, $129.5),
(582, $303.0), (541, $201.5), (016, $365.0), (981, $346.5), (158, $210.5),
(945, $334.0), (072, $376.5), (167, $215.5), (077, $312.0), (185, $239.5),
(209, $221.0), (893, $388.0), (346, $154.5), (515, $268.5), (555, $127.0),
(858, $537.5), (434, $427.5), (541, $272.0), (411, $197.0), (109, $167.5),
(761, $292.0), (767, $170.0), (597, $486.5), (479, $262.0)

All the available information about the lottery results is here, but it is not presented in an especially useful form. Some interesting features can be seen by reading the numbers (can you tell what they are?), but they are not obvious and must be searched for carefully.

Since we are much better at processing visual information than numeric information let's try a few visual displays to see what they reveal. We'll start by using a histogram to look at the distribution of the winning numbers (see figure 1.2).

The histogram shows a slight peak at about 200, but this can be explained by random variability. If the numbers are chosen randomly, we would expect the winning number to fall into each of the histogram cells with probability $1/10$. The expected count in each cell would thus be $254 \times 1/10 = 25.4$ with a standard deviation of

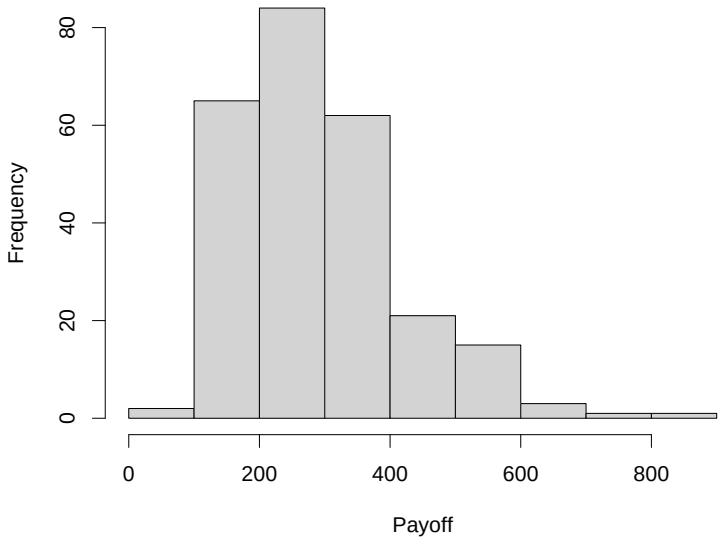


Figure 1.3: The distribution of the winning payoffs for the "Pick-It" lottery.

$\sqrt{254 \times 1/10 \times 9/10} = 4.78$. This looks to be about right for the plot.

What about the winning payoff? This can also be investigated by producing a histogram of the payoffs as shown in figure 1.3. This plot looks very different from the random distribution shown in figure 1.2. Payoffs are most frequently between \$100 and \$400, but there some which are much larger and a few which are pitifully small.

The knowledge that there are sometimes very large payoffs is only useful to us if there is some way that we can choose ticket numbers which are likely to lead to these large payoffs. To investigate, this we need to determine whether there is a relationship between the winning ticket number and the payoff amount.

One way to check for a relationship is to make a plot of the payoff amount against the winning ticket number. Figure 1.4 shows such a scatterplot. The plot has a very interesting feature; the group of payoffs at the extreme left appear to be quite a bit higher than the rest of payoffs. This group consists of those tickets with a number whose leading digit was zero!

This suggests a different way of looking at this data set; partition the values into groups by the leading digit of the ticket number and compare the values of the payoff by group. One way to do this is to produce a set of parallel box-and-whisker plots; as in figure 1.5.

The plot shows that the highest payoffs generally occur when the leading digit is zero and that a payoff when the leading digit is a 1, 2, or 3, is generally less than when the leading digit is 4, 5, 6, 7, 8 or 9. It also shows that there are a number of extreme points, at both the high and low ends of the payoff scale.

The ticket numbers corresponding to the extreme payoffs can be determined by labelling the extreme points in the scatterplot with their ticket number. Figure 1.6 shows the extreme points labelled in this way. They all show repeated digits.

On the basis of the plots we have made, we can now begin to formulate some theories which will help explain the structure of the lottery observations and help us if we play the game.

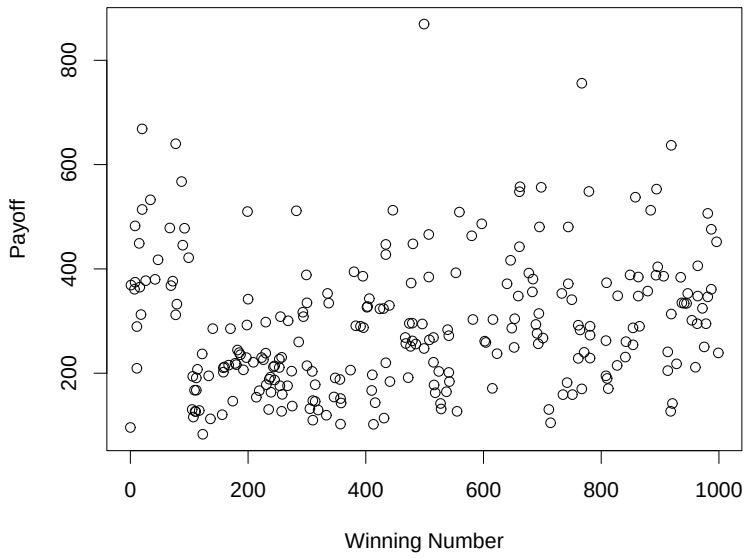


Figure 1.4: Winning payoff plotted against winning number. Numbers with leading digit zero generally seem to produce higher payoffs

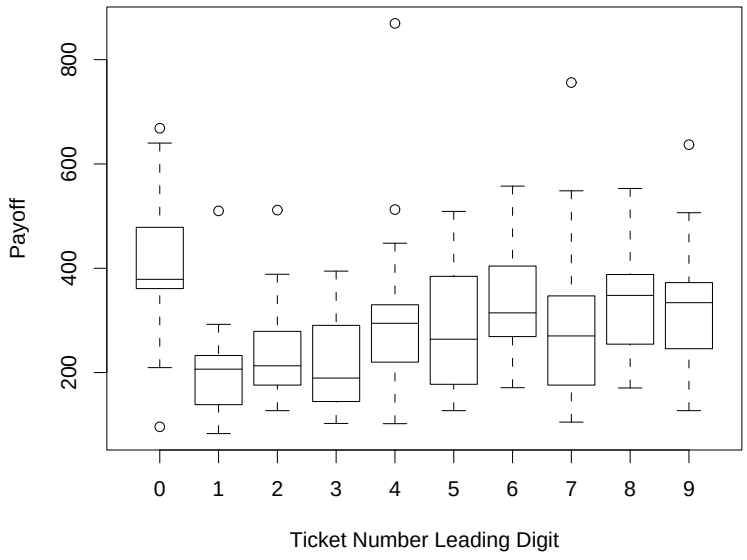


Figure 1.5: Box-plots comparing the winning payoffs when the tickets are grouped by the leading digit of their number.

- Players tend not to pick numbers whose leading digit is zero. This means that fewer people select these numbers and so the payoff is higher when they are the winning numbers.

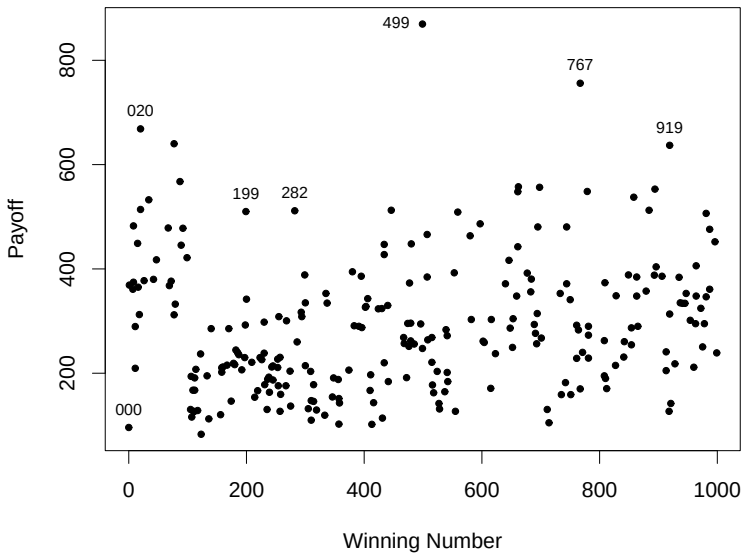


Figure 1.6: A scatterplot wit the extreme points labelled by their ticket number.

- People tend not to pick numbers with repeated digits; possibly because of an intuitive belief that such numbers are less likely. This results in fewer winners and so such numbers tend to yield higher payoffs.

There is also some evidence that players tend to pick numbers whose first digit is small (but not zero). This is the reason that the pay-outs whose leading digits are 1, 2, and 3 tend to produce lower payoffs.

Finally, a look back at the original data values shows that some special numbers are best avoided because of their popularity. The numbers 000, 123, 333 and 555 all had very low payoffs. To get a high payoff it would be best to avoid this kind of number.

Despite the fact that we have some promising leads on how to maximise our returns in the "Pick-It" lottery, we still need to consider whether the game is worth playing. To make a rational decision about this, we must weigh how much we expect to win against how much we pay for a ticket. In this game we pay a dollar to play and we expect, on average, to win somewhat less than a dollar. The rational decision is clearly not to play, but gambling decisions are rarely made rationally.

1.4 Graphics for Insight

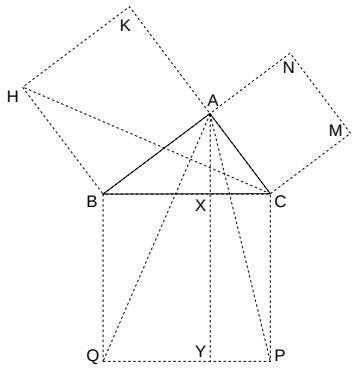
Pythagoras' theorem is one of the oldest and best known mathematical results. For thousands of years the standard Western proof of this result has been that presented by the Greek Philospher/Mathematician Euclid. A version of this proof, taken from C. V. Durrell's *Elementary Geometry* [5], is reproduced below.

THEOREM (Pythagoras' Theorem)

In any right-angled triangle, the square of the hypotenuse is equal to the sum of the squares on the sides containing the right angle.

Given ∠BAC is the right angle.

To prove the square on BC = the square on the BA + the square on AC.



Let ABHK, ACMN, BCPQ be the squares on AB, AC, BC.

Join CH, AQ. Through A, draw AXY parallel to BQ, cutting BC, QP at X, Y.

Since ∠BAC and ∠BAK are right angles, KA and AC are in the same straight line.

Again, ∠HBA = 90° = ∠QBC.

Add to each ∠ABC, ∴ ∠HBC = ∠ABQ.

In the △s HBC, ABQ,

HB = AB, sides of square,

CB = QB, sides of square,

∠ HBC = ∠ ABQ, proved.

∴△HBC = △ABQ (2 sides, inc. angle).

Now △HBC and square HA are on the same base HB and between the same parallels HB, KAC;

∴ △HBC = $\frac{1}{2}$ square HA.

Also, △ABQ and rectangle BQYX are on the same base BQ and between the same parallels BQ, AXY.

∴ △ABQ = $\frac{1}{2}$ rect. BQYX.

∴ square HA = rect. BQYX.

Similarly, by joining AP, BM, it can be shown that square MA = rect. CPYX;

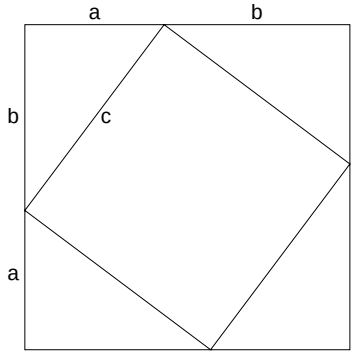
∴ square HA + square MA = rect. BQYX + rect. CPYX

= square BP. Q.E.D.

In fact, there are many proofs of Pythagoras' theorem and Euclid's proof is one of the more difficult of them. It is, admittedly, a direct proof, but it requires a good deal of explanation. There are some 63 encoded references from the proof to the associated diagram, and a good deal of thinking is required before the idea underlying the proof becomes clear (I'm not sure that it ever did become clear to me in high-school).

By contrast, I'll present two simpler proofs below, to show just how simple the result is when looked at the right way. Both "proofs" consist of diagrams and a small amount of additional calculation.

The first proof is based on a diagram which consists of four copies of a right-angled triangle arranged in the following figure.



The figure is enclosed by a square whose sides have length $a+b$, and thus whose area is $(a+b)^2$. This "large square" clearly consists of the four triangles, each of area $ab/2$, together with an "inner square" whose sides have length c and whose area is thus c^2. Equating areas, we see that

$$(a+b)^2 = 4(ab/2) + c^2.$$

Expanding and simplifying, this becomes

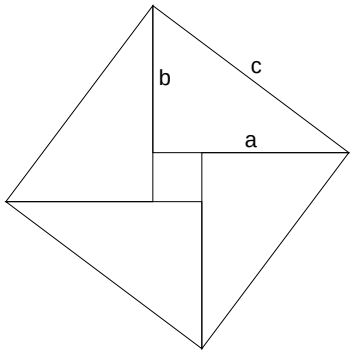
$$a^2 + b^2 + 2ab = 2ab + c^2,$$

and, cancellation of the $2ab$ term yields

$$a^2 + b^2 = c^2.$$

(You should note that we have glossed over some details here. For example, how do we know that the inner figure is a square?)

A second simple proof of Pythagoras' theorem is based on a similar diagram. Again, four copies of the triangle are arranged around a central square.



Simple algebra again reveals that $a^2 + b^2 = c^2$.

This second proof is very old. The diagram can be found in the classical Chinese mathematics book *Zhou bi suan jing* (ca -600 to $+300$), which predates Euclid and is contemporary with Pythagoras. (It is possible that the diagram was not in the original text, but was added by its primary commentator Zhao Shuang in the third century C.E.)

In fact the result is even older. It was known know to the Babylonians a thousand years before Pythagoras, and it possible that they also used diagrams of the form above to prove it. Despite this, Pythagoras is generally credited with giving the first proof of the result.

The lesson of this section is that diagrams and graphs can be a major aid in theoretical as well as practical studies. A good choice of diagram can often simplify proofs and provide insight into such problems.