MAT220 Lecture 2: visualizing Data

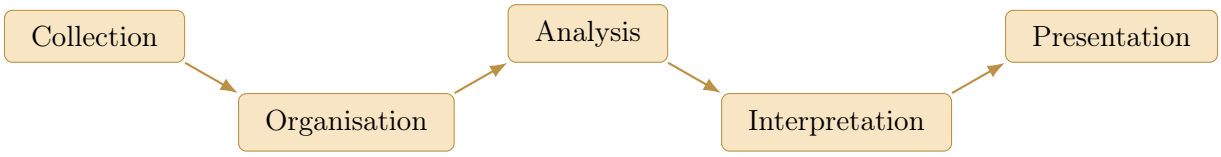
Contents

In Lecture 1, we introduced the basic language of statistics: data, variables, populations, samples and the workflow of a statistical study. In this lecture, we will start our discussion of descriptive statistics by focusing on the visual presentation of data. At this stage in the course, we have not yet introduced any formulas. Instead, today we will study the idea of *communication*: how to turn a data set into a visual summary that others can quickly scan and understand. In keeping with the main point of the previous lecture, different data sets may have different structures, and therefore the "correct" presentation will depend on this structure.

---

1 Data Visualization as Presentation

Last lecture we saw that, on a basic level, a statistical study can be broken down into five key components:



We emphasized that this workflow is more like a *chain* of dependencies: if any of the boxes is flawed in its own way, then that is like "breaking the chain of validity", and the results of the study will become unreliable. For instance, if we collect inaccurate or biased data in the first place, then it does not matter how good our statistical analysis is, the result will be as unreliable as our initial data. Or, if you collect accurate data in an unbiased way, but you apply the wrong formulas to it, then the overall study will be dubious.

Later on in the course, we will spend a lot of time discussing the analysis and interpretation stages of a statistical study. This is an extremely important element of statistical practice: not only do we need to know which formulas to correctly apply, but we also need to understand *why* these formulas are correct and *what they imply*. However, in today's lecture we will assume that all of these deeper considerations are already established, and we will jump straight to the final step of the process: presentation.

Definition

Data visualization is the presentation of data using visual objects such as tables, graphs, charts, diagrams and maps. The goal is to make important structure in the data easier to see.

Of course, statistical studies have the potential to be extremely insightful. However, our insight is only ever as good as our *representation* of it: if data is presented poorly, then any underlying lessons or narratives may be impossible to identify. In other words, accurate data is useless unless we can clearly see the patterns inside of it.

1.1 Presentation is more art than science

Generally speaking, there is no perfect way to visualize a data set. As we will see, different data sets admit many different representations. Therefore, we as the presenters of the data have to make

informed choices about which elements of our data we want to emphasize. Different presentations will have different effects on the audience, and therefore they will communicate slightly different messages. For this reason, data presentation is more of an art form than a strict science: there is no "formula" that we can apply to create the best presentation of data, and instead we need to use a mix of practical judgement and aesthetic taste. As we will see, often the right mix will depend on our data, our context, and our audience.

Important point

Data presentation is more art than science. A graph is not just a neutral container for information, instead it carries the choices made by the presenter along the way, and these choices will affect what the audience notices.

Of course, this does not mean that *anything* goes. As we will see, graphs can be genuinely bad, ugly, misleading or outright dishonest. Throughout today's lecture, we will slowly identify some key guidelines for good data presentation. Along the way, we will survey some basic representations of data. We will discuss:

1. quantitative data,
2. categorical data,
3. time-dependent data, and
4. paired data.

Finally, we will finish with a case study designed for this lecture, so that you can see several ways in which the same data can be presented.

2 Representing Quantitative Data

Recall that a data set is called "quantitative" whenever it consists of numerical measurements or counts for which arithmetic is meaningful. Because numerical data has a mathematical structure (order, arithmetic, and so on), it is often useful to keep this extra structure in mind when representing the data visually.

Generally, it is not particularly useful to present large amounts of quantitative data directly. To illustrate this, consider the following table of data, which lists the shoe sizes of 42 people, measured in centimetres:

Shoe sizes of 42 people					
25	27	28	24	26	26
26	24	30	25	25	24
26	24	27	24	24	23
27	24	28	26	25	23
24	24	25	24	24	24
25	24	25	23	26	25
27	25	24	27	27	25

The table above simply displays the raw data. Although this presentation is very open and honest,

it is not particularly insightful because we cannot easily see any useful patterns in the data. For instance, just from the numbers alone, we cannot quickly answer potentially useful questions such as:

- Which shoe size is most common?
- How are the shoe sizes roughly distributed in the group?
- Are there any outliers?

The answers to these questions are contained within the data *somewhere*, but they are hidden. It would be much better if these hidden structures were uncovered through a different representation.

2.1 Frequency tables and histograms

When working with quantitative data, it is often useful to group the data into intervals of the same size. These intervals are often called *classes*, or more commonly *bins*. From this grouping, we can then create a frequency table, pictured below. For clarity, we redraw the same raw shoe-size data below next to its grouped frequency table.

Shoe sizes of 42 people

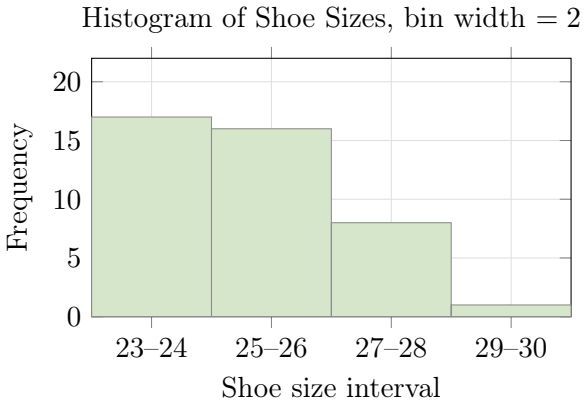
25	27	28	24	26	26
26	24	30	25	25	24
26	24	27	24	24	23
27	24	28	26	25	23
24	24	25	24	24	24
25	24	25	23	26	25
27	25	24	27	27	25

→

Class Interval	Frequency
23–24	17
25–26	16
27–28	8
29–30	1

Here we have chosen to collect the 42 data points into bins of size 2, which are simply intervals of length 2 that run across the data. As you can see, the frequency table is already an improvement, since we can immediately see that lots of the data is clustered around the 23 to 26 range.

We will now use the information in the frequency table to plot a graph: the x-axis displays the intervals, and the y-axis displays the counts. Diagrams of this type are known as *histograms*.



Definition: Histogram

A **histogram** is a graph used to represent quantitative data by grouping values into intervals called bins. The height of each bar represents how many data values fall into that bin.

Histograms display the same basic information that can be found in the frequency table. However, the information is much more *accessible*, because we can quickly see how the frequencies relate to each other.

2.2 Relative frequency distributions

A frequency tells us how many observations fall into a given bin. Sometimes, however, we care more about proportions than the raw frequencies themselves. Fortunately, these proportions can be calculated with a simple procedure.

Definition: Relative Frequency

The **relative frequency** of a class is

$$\text{Relative Frequency} = \frac{\text{Bin Frequency}}{\text{Total Count}}.$$

It tells us the proportion of the data set that lies inside that bin.

For example, using our shoe-size data, the first class has frequency 17 and there are 42 people in total. So, its relative frequency is:

$$\frac{17}{42} \approx 0.405 = 40.5\%.$$

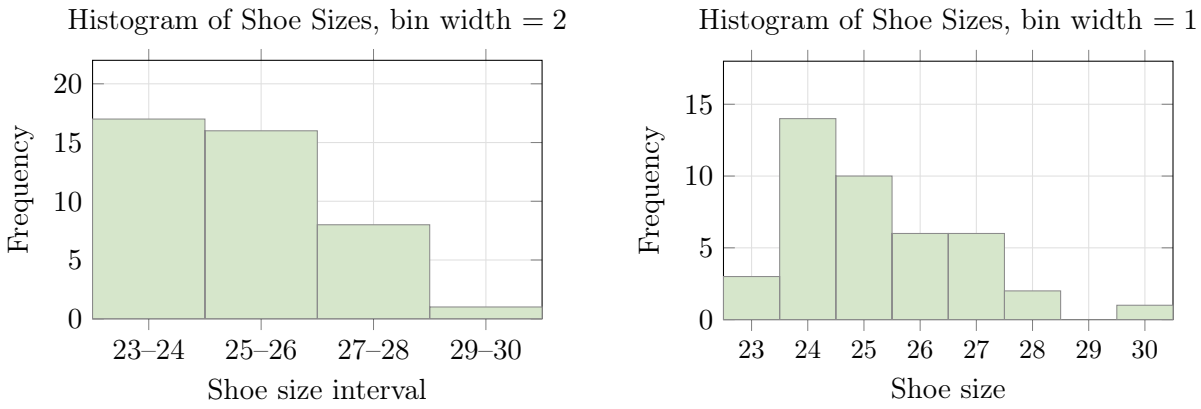
The table below is what is known as a *relative frequency table*.

Class Interval	Frequency	Relative Frequency
23–24	17	$17/42 \approx 40.5\%$
25–26	16	$16/42 \approx 38.1\%$
27–28	8	$8/42 \approx 19.0\%$
29–30	1	$1/42 \approx 2.4\%$

Relative frequency is useful when comparing data sets of different sizes. For example, if one class has 42 students and another class has 120 students, raw counts may not be directly comparable. Percentages often make the comparison clearer.

2.3 Abuse of histograms: changing bin width

Histograms are useful, but they are also very easy to manipulate. The main way to do this is by altering the *bin width*, since this may cause different shapes to occur as data shuffles around into different bins. For example, our shoe-size data can also be counted into bins of width 1, and this results in a very different-looking histogram:



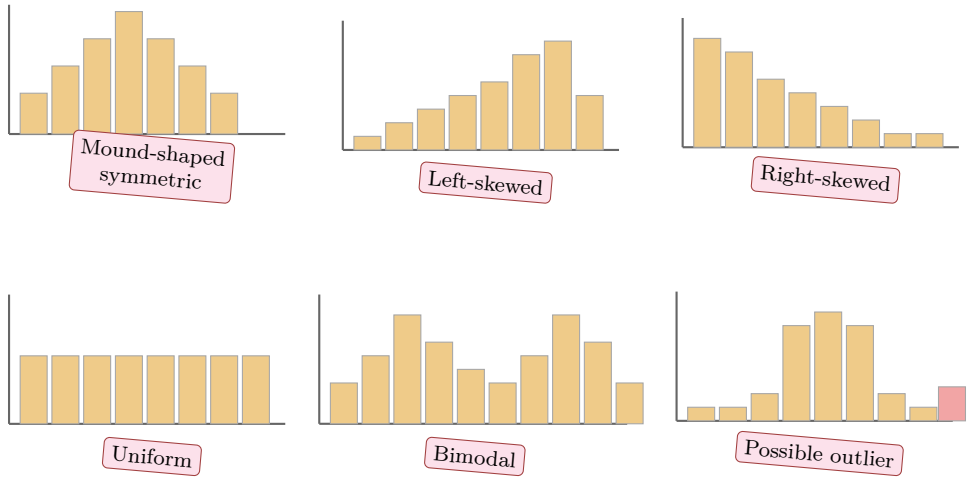
As you can see, the histogram on the right-hand side provides a higher resolution description of the data. Here we can see the distribution of each individual value. Notice how some of this data was obscured using a bin width of 2: in the case of the shoe sizes in the 23–24 bin, the first histogram merely tells us that there are 17 entries. However, the histogram on the right communicates that there are *many* more people with a shoe size 24 than there are of size 23. This extra information is entirely lost when we expand the bin size to width 2.

Important warning

Changing the bin width can change the visual story of a histogram. A responsible presenter should choose bins that reveal the structure of the data, rather than bins that exaggerate or hide a preferred conclusion.

2.4 Distribution shapes

Usefully, the overall shape of a histogram tells us something about the general way that the data is distributed. Some common histogram shapes are drawn below.



We will now describe these important shapes one-by-one.

2.4.1 Mound-shaped symmetric distributions

A mound-shaped symmetric distribution is one in which the histogram rises toward the middle and then falls away in a roughly balanced way. The two sides do not have to be perfectly identical, although in lots of cases they may well be. The main idea is that the centre is the most populated part of the distribution, and the data is spread out in both directions from that centre.

Mound-shaped symmetric distribution

A **mound-shaped symmetric distribution** is a distribution whose histogram has a central mound and whose left and right sides are approximately balanced.

Many naturally varying measurements, such as adult height, are often roughly mound-shaped, but not all randomly sampled data has this shape.

2.4.2 Skewed distributions

A skewed distribution is one where one tail is longer than the other, i.e. there is an inherent asymmetry. Confusingly, we name the skewness after the direction in which there is a *longer* tail, i.e. if the longer tail is on the left of the distribution, then we call that graph "left-skewed".

Skewed distributions

A distribution is **skewed** when one side has a longer tail than the other.

- **Right-skewed:** the longer tail is on the right.
- **Left-skewed:** the longer tail is on the left.

The long tail often contains unusually large or unusually small observations. For example, income data is often right-skewed because most people earn ordinary amounts, while a small number of people earn extremely large amounts.

2.4.3 Uniform distributions

A uniform distribution is one in which every class has the same frequency, causing the histogram to look flat or rectangular.

Uniform distribution

A **uniform distribution** is a distribution where the classes have approximately equal frequencies.

Exact uniform distributions are uncommon in observational real-life data, but they are theoretically important, especially later during our discussions of probability theory. Uniform distributions represent a situation in which no interval is particularly special compared to any other.

2.4.4 Bimodal distributions

A bimodal distribution has two noticeable peaks. This can often indicate that the data set is mixing two different populations, or that there are two distinct "explanatory forces" acting on the population.

The idea is simple: the graph looks like two copies of the symmetric mound-shaped distribution. An individual peak may suggest a subgroup, mechanism, or repeated pattern in the data. In the case of a bimodal distribution there are *two* mounds, which can mean that there are *two* different properties or forces acting on the data.

Bimodal distribution

A **bimodal distribution** is a distribution with two main peaks, usually separated by at least one lower-frequency class.

For example, we could measure pedestrian traffic passing through a popular urban train station like Shinjuku. The (relative) frequency distribution will display two clear mounds: one centred around 8 a.m. and another centred around 6 p.m. This is because there are two peaks in traffic as all of the commuters go to and from work, respectively.

2.5 What histograms tell us

Histograms and relative-frequency histograms can show us how quantitative data is roughly distributed. By looking at histograms, we can often conclude:

- whether the data distribution is symmetric, skewed, uniform or bimodal,
- whether there are possible outliers,
- which data intervals contain the most data, and
- how spread out the data is.

3 Representing Categorical Data

Recall that categorical data consists of labels, names, groups or categories. Categorical data does not necessarily have any inherent mathematical structure that we can use to our advantage. For example, if our categories were blood types, then there is no sense in which we can "add" type A-negative and type O-positive together in some meaningful way. Since there is so little mathematical structure attached to categorical data, we are quite restricted in what we can do: the main operations we can perform are to count frequencies and to compare proportions. These two operations lead to two common representations of categorical data:

- **bar charts**, which represent frequencies or percentages, and
- **pie charts**, which represent proportions of a whole.

We will now discuss these separately.

3.1 Bar charts

A bar chart represents individual categories using separate bars in a graph, similar to a histogram. The length or height of each bar represents a value, usually a frequency or percentage.

Definition: Bar Chart

A **bar chart** is a graph for categorical data in which each category is represented by a separate bar. The height or length of the bar represents the frequency, percentage or value of that category.

Good bar charts follow three basic rules.

1. Bars can be vertical or horizontal.
2. Bars should have uniform width and uniform spacing.
3. The same measurement scale should be used for every bar.

The gaps between bars are important. A histogram represents intervals on a numerical scale. A bar chart represents distinct category labels, so the bars are usually visually separated.

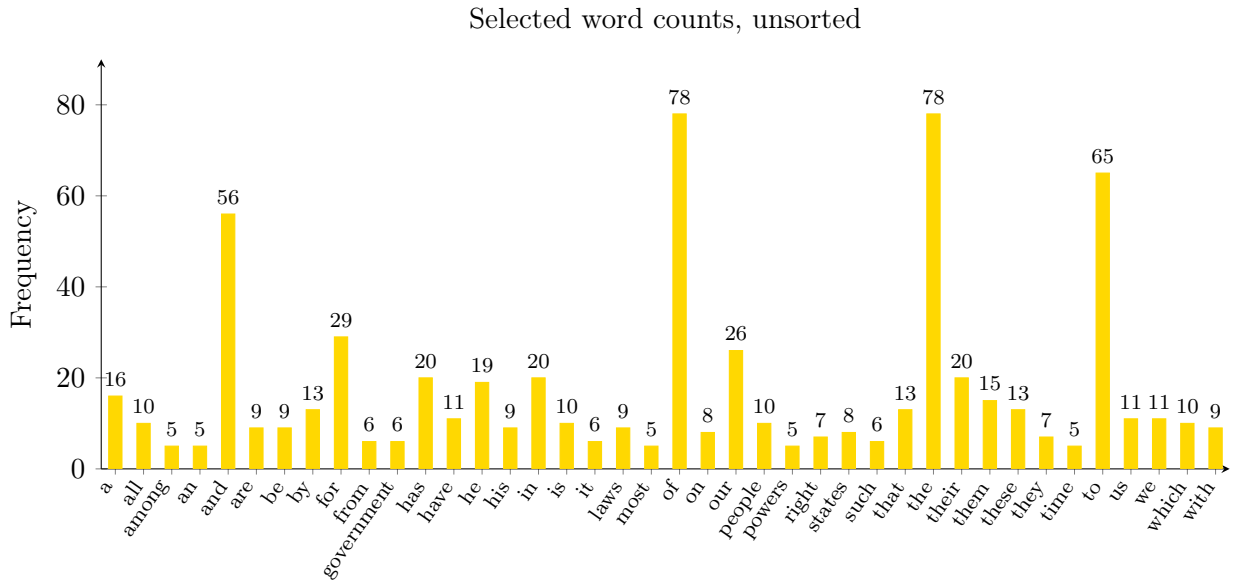
A Pareto chart is a bar chart whose bars are arranged in decreasing order. The most common categories are placed on the left, and the least common categories are placed on the right.

Definition: Pareto Chart

A **Pareto chart** is a bar chart in which the bars are arranged from largest to smallest. It is useful when we want to quickly identify the most frequent categories.

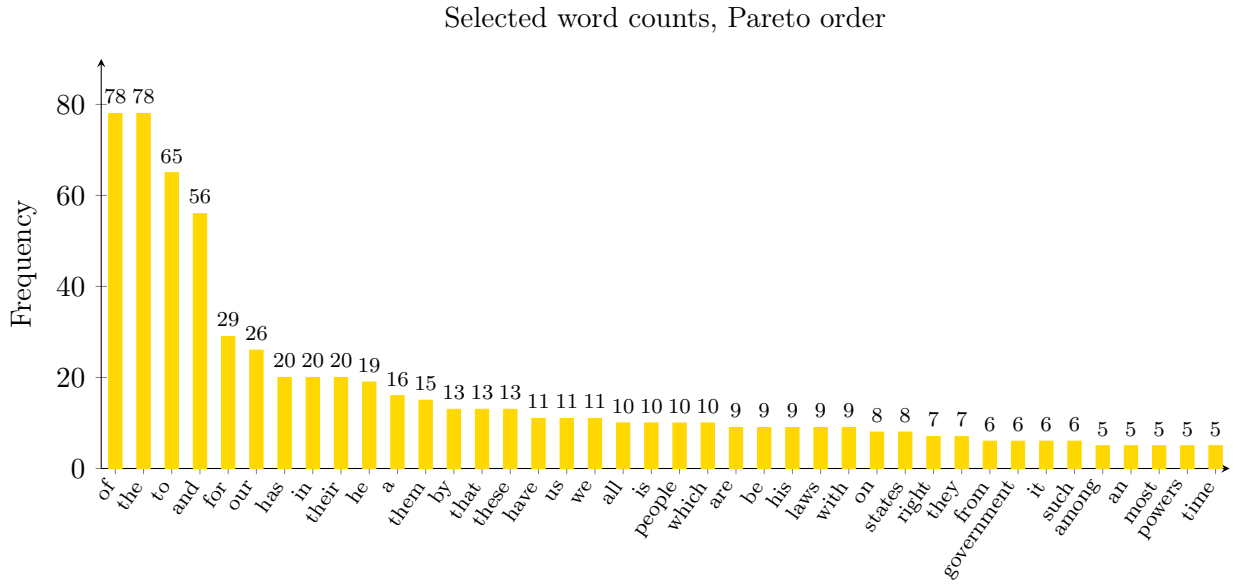
3.2 Example: The Declaration of Independence

Suppose we count frequently occurring words in the Declaration of Independence. Using this particular word count, the main body contains 1333 words, so a bar chart that contains every word frequency would be quite cluttered. Instead, we will plot all words occurring at least 5 times as a bar chart:



As you can see, the graph is readable, but still quite crowded. Because there are so many individual

bars, it becomes hard to quickly answer questions like "which words are among the most frequent?" This can be improved by using a Pareto chart instead:



The Pareto chart contains the same counts, but the ordering makes the dominant categories much easier to identify. In other words, the ordering and harmony of in the Pareto chart is easier on the eye.

3.3 Pie charts

A pie chart represents proportions as slices of a circle. To make a pie chart, we first calculate *proportions* by counting how often each category occurs, then dividing by the total count.

Definition: Proportion

A **proportion** is a part-to-whole comparison:

$$\text{Proportion} = \frac{\text{Frequency}}{\text{Total Count}}.$$

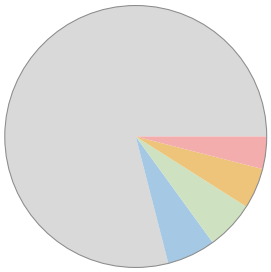
As an example, we use the data from the previous section. Observe that the word "the" occurs 78 times in the Declaration of Independence, out of a total of 1333 words. Therefore, the proportion of the word "the" is:

$$\frac{78}{1333} \approx 0.0585 = 5.85\% \approx 6\%,$$

that is, "the" appears approximately 6% of the time.

Definition: Pie Chart

A **pie chart** is a graph in which categories are represented as wedges of a circle. The size of each wedge represents the proportion of the total belonging to that category.



A readable pie chart has only a few substantial slices.

Important warning

Pie charts work best when there are only a few wedges. If a pie chart contains many tiny slices, the audience may not be able to view or compare the categories accurately.

Exercise

Suppose a class of 40 students is asked for their favourite convenience store. The results are:

7-Eleven : 18, FamilyMart : 12, Lawson : 8, Other : 2.

(a) What proportion of students chose 7-Eleven?
(b) Would a pie chart be reasonable for this data set?
(c) Would a Pareto chart be reasonable for this data set?

Solution

(a) The proportion is $18/40 = 0.45 = 45\%$.
(b) Yes. There are only four categories, so a pie chart would still be readable.
(c) Yes. A Pareto chart would also be reasonable, especially if we want to show the most popular store quickly.

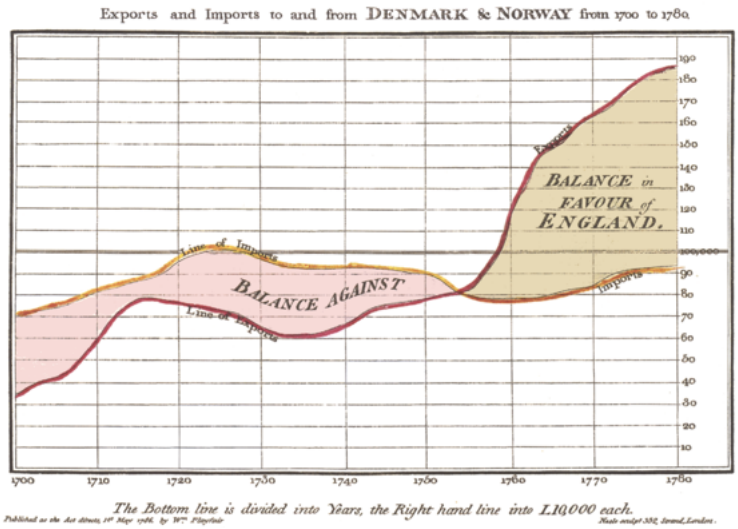
4 Representing Time-Dependent Data

Sometimes we collect data by repeatedly observing the same object. This allows us to track changes in a property over time. For example, we might measure our body weight every week, track the growing height of a sunflower every day, record the number of hours studied each evening, or follow the popularity of a word across many decades. This kind of data is called *time-dependent data* or *time-series data*.

Definition: Time-Series Graph

A **time-series graph** is a graph in which time-dependent data is plotted in chronological order. The horizontal axis usually represents time, and the vertical axis represents the quantity being tracked.

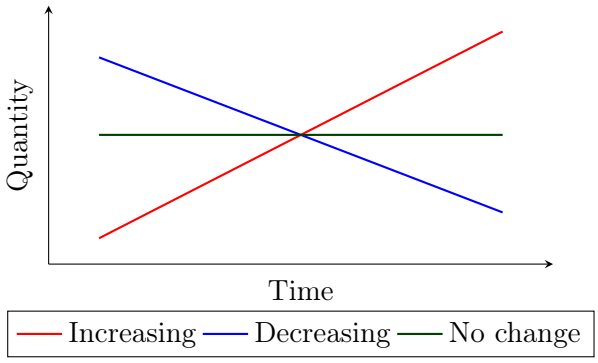
One of the earliest known examples of a statistical time-series graph was produced by William Playfair in the late eighteenth century:



As you can see, a time-series graph is extremely intuitive: we can quite clearly see the change in a quantity over time. We will now briefly review several useful types of change.

4.1 Linear change

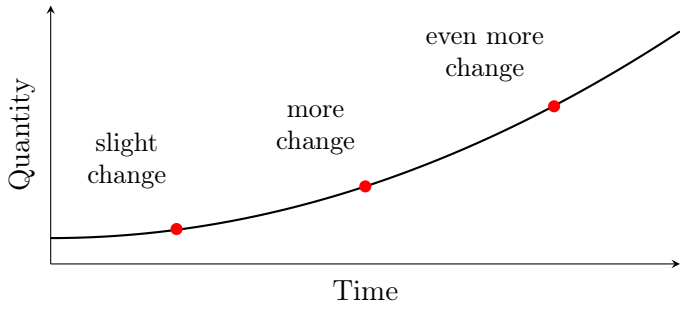
The simplest kind of change over time is *linear*. This means that the data set traces out a straight line when plotted on a time-series graph. Geometrically, the rate of change is represented by the *gradient* of the line:



The increasing line shows a quantity that grows over time. The decreasing line shows a quantity that falls over time. The flat line shows no change. In each case, the main visual idea is the same: the shape of the line tells us the direction and speed of change.

4.2 Non-linear change

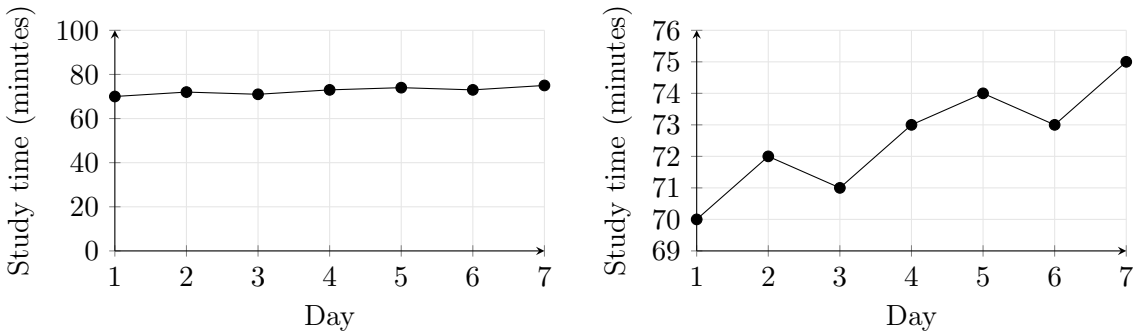
In most cases, change does not happen at a constant rate. Instead, the change itself can *speed up* or *slow down* depending on the process being measured. Consider the following:



In the graph above, we have an example of *acceleration*, since the rate of change itself is increasing with time.

4.3 Scale and narrative

Time-series graphs are especially sensitive to scale: changing the vertical axis can make a small change look dramatic, or make a large change look normal. To illustrate this, consider the two graphs below:



The two graphs above are both built from the same underlying data set. However, by adjusting the scale of the vertical axis we can make the data appear to be very different. The choice of scale subtly encodes *different narratives*: the first graph suggests that there is little change in the data, whereas the second graph suggests that the data is quite volatile. This observation motivates the following.

Important warning

Different scales can create different narratives. When reading a time-series graph, always check the axes before interpreting the story.

This is one of the most common ways that data visualizations mislead people. A graph may be technically correct, but still visually misleading if the scale is chosen to exaggerate the effect.

5 Representing Paired Data

Sometimes the data we collect naturally comes in pairs. This usually happens when we study the same object and collect different properties at the same time, or when we want to compare two properties of each observation. Examples would include:

- Left and right foot length for the same person.
- Travel time and distance travelled for students on their way to campus.
- The height and weight of people living in Tokyo.

Definition: Paired Data

Paired data occurs when each observation consists of two related measurements. We often study paired data to investigate whether there is a relationship between the two variables.

5.1 Example: commute time and distance

Consider the following data set of 12 students.

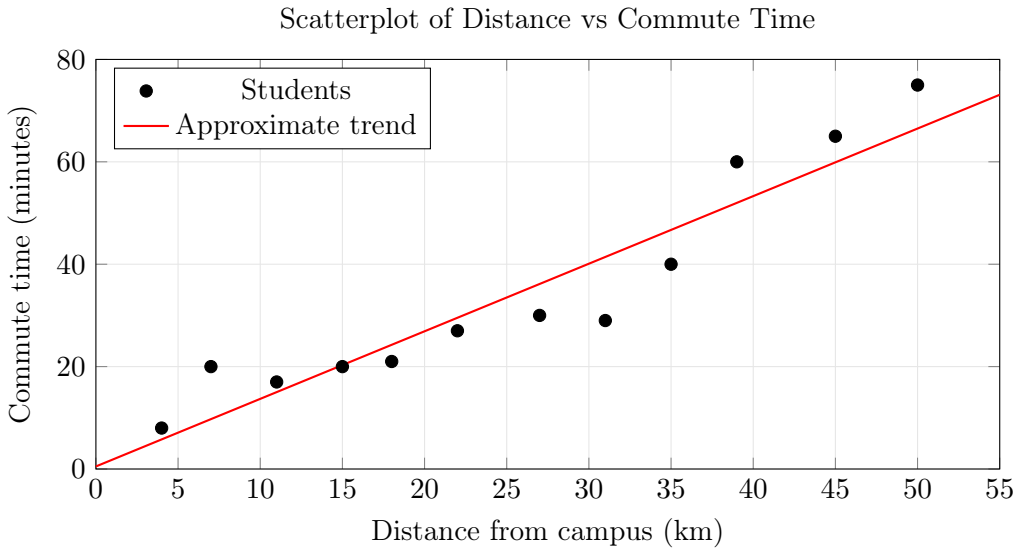
Student	Distance from campus (km)	Commute time (minutes)
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75

This data set is naturally paired because each row contains two linked pieces of information about the same individual student. We can split this pair into two variables: the first variable is distance from campus, and the second variable is commute time. A *scatterplot* is the natural graph for this situation.

Definition: Scatterplot

A **scatterplot** is a graph for paired quantitative data. Each observation is represented by a point, where the horizontal coordinate gives one variable and the vertical coordinate gives the other variable.

To create the scatterplot, we plot one variable on the x-axis and the other on the y-axis:

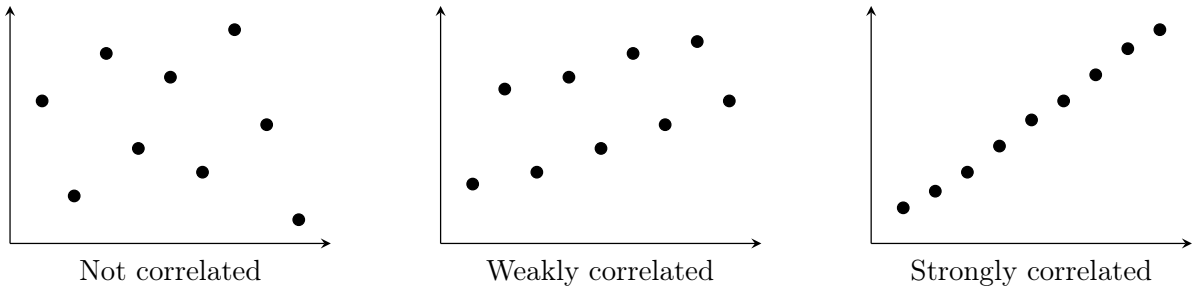


5.2 Correlation

The scatterplot drawn above shows a clear positive relationship between the two variables, meaning that students who live farther from campus tend to have longer commute times. Of course this is not a perfect relationship, since commute times generally depend on more than just distance alone. However, there is still a clear, general pattern between the two variables. When paired data shows a general pattern, we say that the variables are *correlated*.

Definition: Correlation

Correlation means that two variables tend to vary together. A scatterplot can show whether paired data is not correlated, weakly correlated or strongly correlated.



Later in the course, we will study some methods for how to describe correlation numerically. But, for now, it is enough simply to recognise the basic idea:

- If the points form no clear pattern, the variables may be uncorrelated.
- If the points roughly move in one direction but with a lot of scatter, the variables may be weakly correlated.
- If the points lie close to a line, the variables may be strongly linearly correlated. If they lie close to a curve, the variables may be strongly associated.

A negative correlation would show a downward trend instead.

In the commute-time example, we drew an approximate line through the data. The act of finding a best-fitting line is called **linear regression**. We will study this more formally later in the course.

Linear regression

Linear regression is a method for finding a line that summarises the relationship between two quantitative variables. In this lecture, we only use it visually. The formal method will come later.

5.3 Correlation is not causation

As we discussed in Lecture 1, correlation does not automatically mean causation. A scatterplot can suggest that two variables are related, but it does not by itself prove that one variable causes the other.

Example. Suppose distance from campus and commute time are strongly correlated. It is reasonable to think that living farther away often contributes to a longer commute. However, the exact commute time may also depend on train lines, transfer stations, bicycle access, traffic, walking speed and whether the student lives near a convenient station.

A good scatterplot may *suggest* a causal relationship, but the interpretation still requires context.

6 Case Study: Konbinis in Japan

Suppose that we have a company that produces anpan. They want to sell their products in convenience stores (konbinis) somewhere in Japan, and they are trying to do some market research to properly inform their decision. They task a poor intern, Jenny, to do some research about the distribution of konbinis nationwide, and then to present the results. Simply put, Jenny's task is to:

- figure out how konbinis are distributed within Japan, and then
- present the data to her bosses.

So, Jenny uses a data set of convenience stores listed by prefecture:

Prefecture	Stores
Hokkaido	3003
Saga	360
Yamanashi	459
Fukui	337
Tokyo	7405
Ehime	583
Miyagi	1161
Gifu	854
Ibaraki	1432
Hiroshima	1205
Fukushima	904
Chiba	2701

Prefecture	Stores
Aomori	610
Kagawa	406
Aichi	3642
Okayama	797
Tochigi	921
Yamaguchi	565
Toyama	485
Kagoshima	665
Akita	449
Yamagata	447
Gunma	900
Kyoto	1077

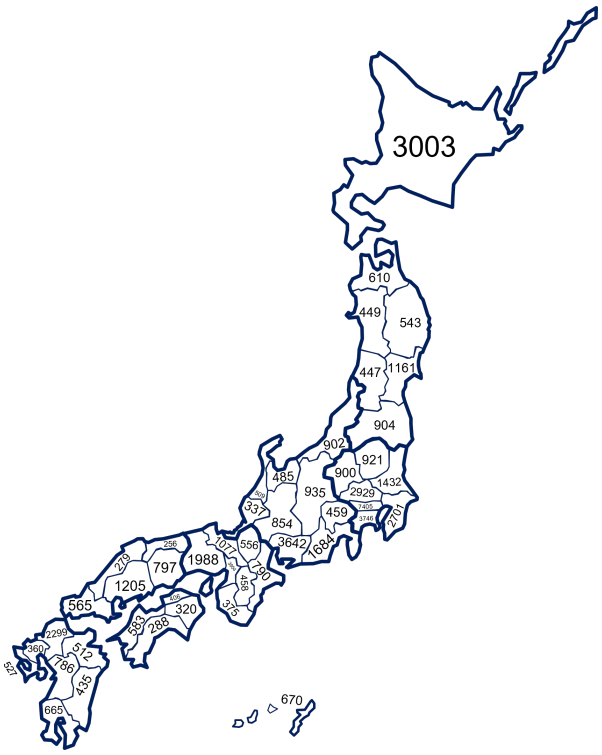
Prefecture	Stores
Shizuoka	1684
Kochi	288
Tottori	256
Shimane	279
Okinawa	670
Niigata	902
Nagano	935
Miyazaki	435
Oita	512
Wakayama	375
Kumamoto	786
Kanagawa	3746

Prefecture	Stores
Ishikawa	509
Nagasaki	527
Iwate	543
Saitama	2929
Fukuoka	2299
Shiga	556
Osaka	3954
Hyogo	1988
Mie	790
Nara	458
Tokushima	320

We will now explore some ways in which to present this data visually.

6.1 Map representations

The data written above is precise, but it is not very easy to interpret. As you can see, it contains all of the relevant information, however, it is very difficult to make sense of any underlying patterns in the data. Even if we were to plot this honestly on a map of Japan, the results would be uninformative:

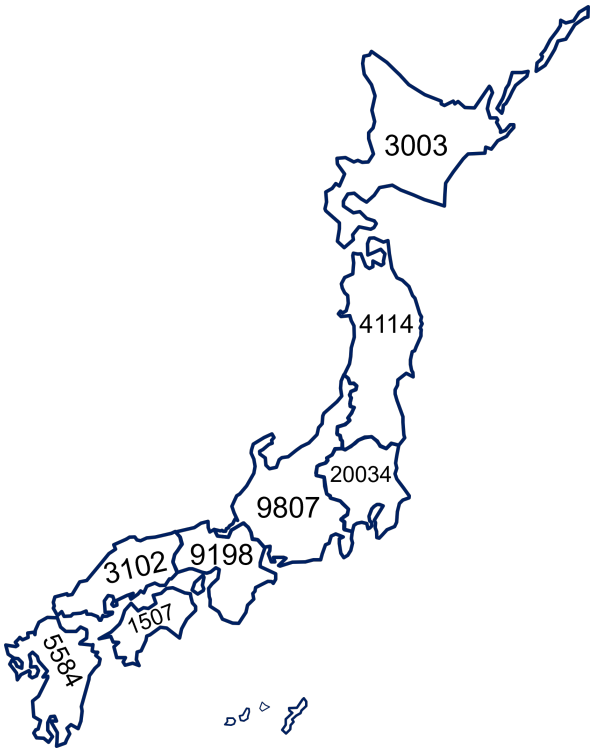


As you can see, this map is not a good way to represent this data. Although it has hints of a good idea, there are three main issues:

1. the numbers are too small to read,
2. it is still difficult to see any sort of pattern in the data,
3. even if we *could* see all of the numbers, their exact values may not actually be relevant.

One approach to solving this might be to reduce the resolution of the map by aggregating the data into regions instead of prefectures. Firstly, we may rewrite the counts into the main regions of Japan:

Region	Convenience Stores
Hokkaido	3003
Tohoku	4114
Kanto	20034
Chubu	9807
Kansai	9198
Chugoku	3102
Shikoku	1597
Kyushu	5584
Okinawa	670



As you can see, this data tells a clearer story: Kanto has the largest number of convenience stores by far, followed by Chubu and Kansai. However, we have lost detail in the process of making the data easier to read. We will now resolve this.

6.2 Coloured representations

Given the context of this presentation, it is most likely that Jenny doesn't actually need to present the exact numerical data of her study. In other words, the raw data may display *too much* information to be genuinely useful. We can suppress this information whilst maintaining its patterns by encoding quantity via a colour gradient. In the table below, the higher counts are coloured further towards blue, and the lower counts are coloured further towards red, with white being in the middle.

Prefecture	Stores
Hokkaido	3003
Saga	360
Yamanashi	459
Fukui	337
Tokyo	7405
Ehime	583
Miyagi	1161
Gifu	854
Ibaraki	1432
Hiroshima	1205
Fukushima	904
Chiba	2701

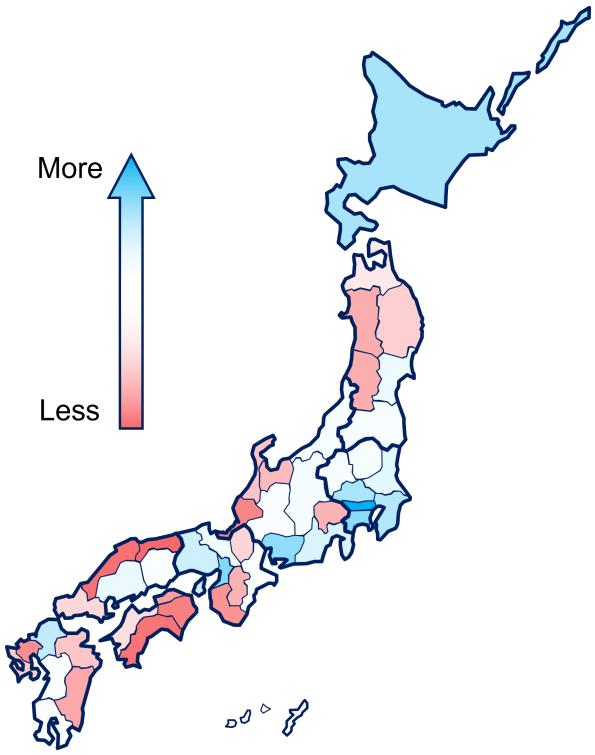
Prefecture	Stores
Aomori	610
Kagawa	406
Aichi	3642
Okayama	797
Tochigi	921
Yamaguchi	565
Toyama	485
Kagoshima	665
Akita	449
Yamagata	447
Gunma	900
Kyoto	1077

Prefecture	Stores
Shizuoka	1684
Kochi	288
Tottori	256
Shimane	279
Okinawa	670
Niigata	902
Nagano	935
Miyazaki	435
Oita	512
Wakayama	375
Kumamoto	786
Kanagawa	3746

Prefecture	Stores
Ishikawa	509
Nagasaki	527
Iwate	543
Saitama	2929
Fukuoka	2299
Shiga	556
Osaka	3954
Hyogo	1988
Mie	790
Nara	458
Tokushima	320

We can take this data and simply shade each prefecture with its corresponding colour. Of course,

this is a loss of information compared to simply presenting the raw counts. However, the payoff is greater: we can immediately see the geographic pattern in the data.



As you can see, the patterns in the data are more detailed than the regional map plotted previously. Here, we see that there are clear hotspots of concentration around urban areas: Tokyo, Nagoya, Osaka, and to some extent Sendai, Hiroshima and Fukuoka. Conversely, outside of populated areas we see a large drop-off in konbini frequency: Shimane, Tottori, Shikoku and northern Tohoku all present a lower-than-average count. Aside from these two general observations, we also notice that Hokkaido is a clear outlier.

Although the above is a better representation than the previous graphs, there is still a sense in which it can be improved: within the context of Jenny and her company, it is likely that the CEOs already know that urban areas have more convenience stores than the national average. Therefore, the study doesn't quite display anything new. However, this does not mean that there is nothing else hiding away inside.

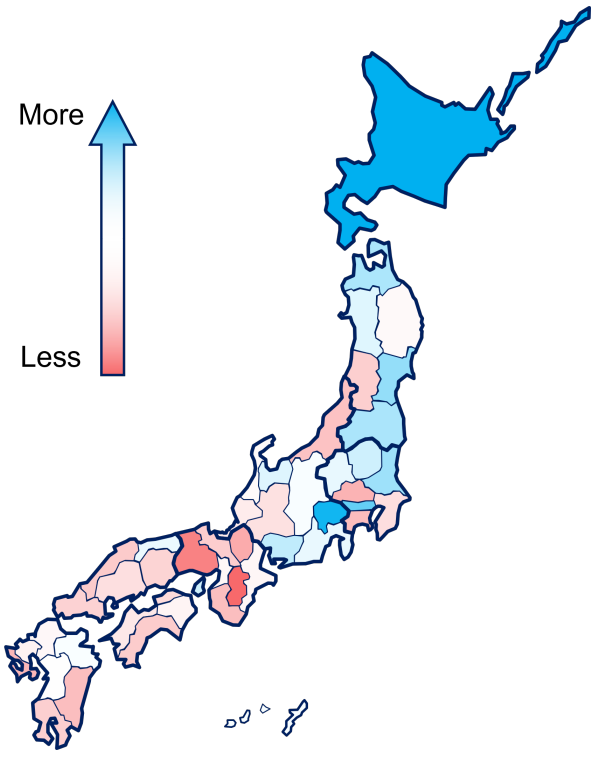
To illustrate a different pattern in this data, we will now divide the store count by the population, then multiply by 100,000, effectively creating a "konbini per 100,000 people" measure. In the table below (again labelled with our colour gradient) we display this extra information.

Prefecture	Stores	Per 100k
Hokkaido	3003	57.43
Yamanashi	459	56.64
Tokyo	7405	52.65
Miyagi	1161	50.40
Ibaraki	1432	49.92
Fukushima	904	49.29
Aomori	610	49.24
Aichi	3642	48.26
Tochigi	921	47.62
Toyama	485	46.83
Akita	449	46.77
Gunma	900	46.38
Shizuoka	1684	46.33
Tottori	256	46.22
Okinawa	670	45.63
Nagano	935	45.62

Prefecture	Stores	Per 100k
Oita	512	45.53
Kumamoto	786	45.19
Ishikawa	509	44.91
Iwate	543	44.83
Fukuoka	2299	44.74
Osaka	3954	44.72
Mie	790	44.60
Tokushima	320	44.46
Saga	360	44.33
Fukui	337	43.91
Ehime	583	43.65
Gifu	854	43.14
Hiroshima	1205	43.01
Chiba	2701	42.96
Kagawa	406	42.69
Okayama	797	42.18

Prefecture	Stores	Per 100k
Yamaguchi	565	42.07
Kagoshima	665	41.85
Yamagata	447	41.83
Kyoto	1077	41.75
Kochi	288	41.62
Shimane	279	41.54
Niigata	902	40.96
Miyazaki	435	40.65
Wakayama	375	40.63
Kanagawa	3746	40.54
Nagasaki	527	40.13
Saitama	2929	39.87
Shiga	556	39.31
Hyogo	1988	36.35
Nara	458	34.56

This table tells a very different story from the raw total counts: some prefectures (especially Yamanashi and those in northern Tohoku) suddenly switch from red to blue, meaning that they are suddenly over-represented. Conversely, some of the urban areas from before now pass into the red as they are scaled by population. Tokyo, however, stays highly represented in the data, and Hokkaido is still an outlier. What is perhaps more interesting is the pattern displayed after plotting onto our map of Japan. Compare the graph below to that of the raw count.



We see that there is a clear East versus West split in terms of konbini per capita: the east of Japan typically scores higher, and the west of Japan typically scores lower. In the context of Jenny's

presentation, she may therefore suggest investigating the east of Japan further, depending on whether the company values store density, total store count, or something else.

6.3 Some observations on data visualization

Finally, we finish with some important observations regarding data visualization.

Important point

Data visualization is usually an intentional *loss* of information in the pursuit of insight. Aside from honesty and accuracy, you should be guided by the principle that your presentation should be useful and interesting, ideally showing the audience something they might not otherwise see.

Presentation of data is always context dependent. As we have seen throughout this lecture, there can always be ways to warp narratives based on choices made by the presenter. Whenever you observe data in the future, it is a good idea to ask yourself the following questions.

1. **Data type:** What structure does the data have: quantitative, categorical, time-dependent, paired, or perhaps some combination?
2. **Purpose:** What question is the graph trying to answer?
3. **Scale:** Are the axes and scales honest, and easy to read?
4. **Grouping:** If the data is grouped, are the bins or categories fair and reasonable?
5. **Loss of information:** What details could have been hidden or simplified?
6. **Conflicts of interest:** Who created this data, and why might they be presenting it in this way?
7. **Narrative:** What conclusion does the visualization encourage the audience to make?
8. **Alternative views:** Would another representation tell a different (but equally valid) story?