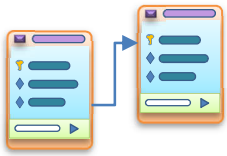
Advanced Data Skills

Checklist — Advanced Data Skills



		Tasks
☐	1. Vocab	1
☐	2. Working with Data	1, 2, 3, 4, 5
☐	3. Database Project: Planning	1, 2, 3, 4, 5, 6, 7, 8, 9
☐	4. Database Project: Building	1, 2, 3
☐	5. Database Project: Populating	1, 2, 3, 4, 5
☐	6. Database Project: Querying	1, 2, 3
☐	7. Database Project: SQL	1, E
☐	8. Database Project: Reporting	1, 2, 3, 4, GF
☐	9. Encoding Recap	1, 2, 3, 4, 5, 6, 7
☐	10. Image Compression	1, 2, 3, 4, 5, E
☐	11. AV Compression	1, 2, 3, E
☐	12. Data Visualisation	1, 2, 3
☐	13. Spreadsheet References	1, 2, 3, 4, 5, 6
☐	14. Spreadsheet Maths	1, 2, 3, 4, 5, E1, E2, E3, E4
☐	15. Spreadsheet Finance	1, 2, 3, 4, 5, 6, 7, 8
☐	16. Spreadsheet Ecology	1, 2
☐	17. Encryption	1, 2, 3, 4, 5

Data plays an ever-increasing role in our lives. It affects our education and work, our social lives and hobbies, our interests and everyday activities. Data is collected when we travel on a train, walk down a city street or visit the supermarket. It's delivered to us when we browse websites or watch videos and listen to music online. All this movement and storage of data must be managed. Much of the data is private, so not only does access to it need to be controlled but the data must be protected during its transfer between computers.

Aim: *To introduce the topic and consider the separation of content, presentation and behaviour.*

This topic will consider how data is compressed and encrypted for safe and efficient transfer between devices and how it is stored, manipulated and retrieved in spreadsheets and databases.

Task 1 – Separating Content, Presentation and Behaviour

One method of streamlining the transfer of all this data is to separate the *content* from the *presentation* and *behaviour*. What do these terms mean? They are explained in the context of a news website below.

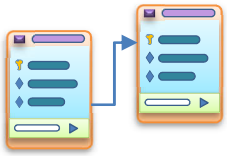
The Content
The actual information. The content is why the publication exists.

The Presentation
The design elements including the layout, the fonts and colours, the icons, logos and the location of links.

The Behaviour
The interactivity that occurs when you click, swipe or run your mouse over objects.

a. Arrange the ideas about creating a website into three separate groups: Content, Presentation and Behaviour.

Working with Data (page 2)



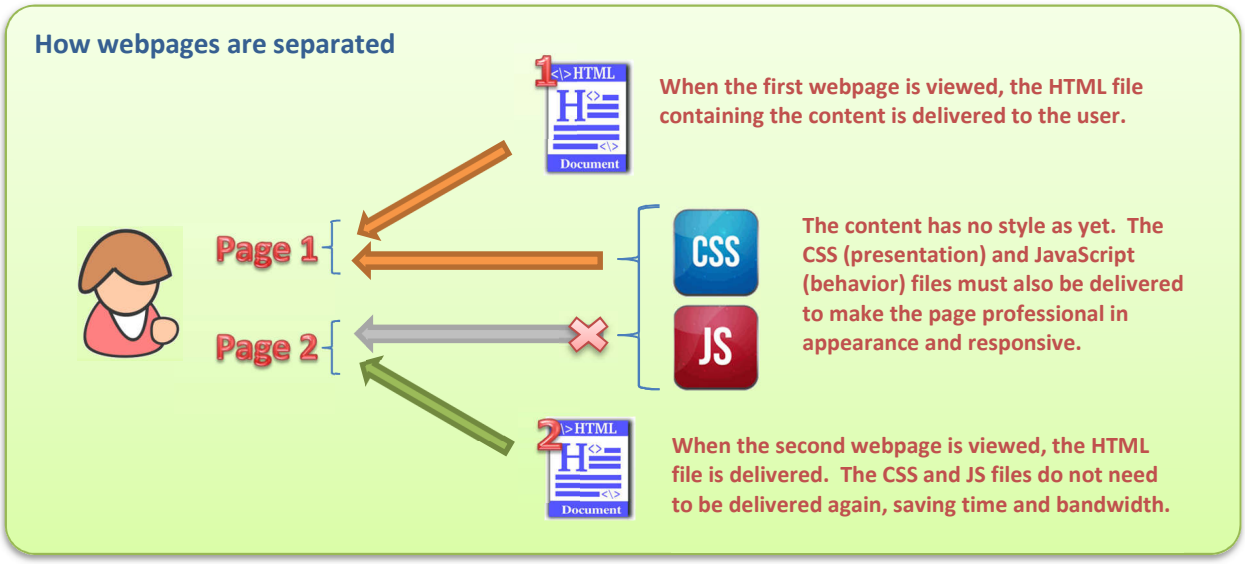
Task 1 – Separating Content, Presentation and Behaviour (cont.)

b. The statements below tell the story of why we try and separate content from presentation and behaviour. Place the statements in the order that makes the most sense. Find further help online if needed.

1. *One benefit was that a well-designed template could be utilised; our subject specialist needed only to fill in the template with their own content. This led to professional looking pages with a consistent feel across the website.*
2. *As things progressed, our content writer/designer also had to grapple with the new interactive website behaviours. This resulted in a lot of poor programming. The situation had to change.*
3. *Another benefit is that the design controls and programming could be placed in separate files which only needed to be transferred once to the website visitor, however many pages they viewed. This reduced the data transfer and improved browsing speeds.*
4. *In the early days of the internet, a single person might design and build a whole website about a subject.*
5. *Over time, designers picked up the design work and techies picked up the programming. Things improved quickly.*
6. *One final benefit is that search engines could get a much better feel for what a webpage was about when all that design and programming code was safely out of the way. Locating information with a search engine became quick and easy.*
7. *As it turned out, this separation of content, presentation and behaviour had huge benefits for a range of reasons.*
8. *This internet pioneer would write the content and create the design. There were a lot of bad looking websites.*
9. *With the templates in use, design changes were easy. A switch of font in one stylesheet could make a difference to all the webpages that picked it up, perhaps numbering in the thousands.*

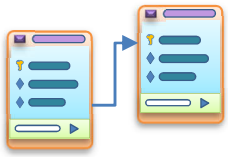
Task 1 Extension

The diagram below shows how a webpage is separated into content, presentation and behavioural elements.



Research some other areas relevant to the separation of content, presentation and behaviour then describe one of them in simple terms. Examples included metadata, XML, tagged PDFs and LaTeX.

3 Database Project: Planning — Advanced Data Skills



Databases are a vital component of the information age and the designing of them is a hugely important skill. You may well have created a small database yourself to hold a list of contacts. The fields were probably laid out in a single table, the data entered and a report printed. Things become more difficult when you want to organise a large quantity of related data, such as information about a whole sports competition, including teams, players and match results. This will be your challenge.

Aim: *To plan the structure of a relational database using primary and foreign keys.*

Task 1 – Flat File Databases

It is possible to hold a great deal of data in a single table like the one shown below. This is a form of *flat file database*.

Sport

ID	First_Name	Last_Name	Address	Team	Team Strip	Game_Date	Start_Time	Home_Team	Away_Team	Home_Goals	Away_Goals
1	Sarah	Layton	15 Wardell Rd	Lansville	Red & White	29/09/2018	2:00 PM	Lansville	Granthorp	3	1
2	Jamie	Ambler	71 Seaview St	Lansville	Red & White	29/09/2018	2:00 PM	Lansville	Granthorp	3	1
3	Sarah	Layton	15 Wardell Rd	Lansville	Red & White	6/10/2018	3:30 PM	St Marks	Lansville	2	2
4	Jamie	Ambler	71 Seaview St	Lansville	Red & White	6/10/2018	3:30 PM	St Marks	Lansville	2	2
5	Sarah	Layton	15 Wardell St	Lansville	Red & White	13/10/2018	2:00 PM	Lansville	North Howe	2	2
6	Jamie	Ambler	71 Seaview St	Lansville	Red & White	13/10/2018	2:00 PM	Lansville	North Howe	2	2

Although this large table could potentially hold all the data about our competition, consider the following questions:

a. How many different ideas are we trying to cram into the table? The address of each player is included. What else is there?

b. How many times has Sarah Layton's name and address been entered into the table? Can you find the mistake?

c. If Jamie Ambler changes address, how many records would need editing to keep the data current?

d. This data involves 2 players, 1 team and 3 matches. How big would the table have to be to include 15 players per side in a 10-team competition and show the results of all the matches? Make some assumptions and come up with a figure.

The fact is that large tables with repeated data are difficult to manage, prone to errors and not very useful when it comes to getting the information out in a report. The secret to all these problems is to organise the data into several related tables.

Task 2 – Relational Databases

A database with related tables is called a *Relational Database Management System* (RDBMS). They are also known more simply as *relational databases*. The three tables below contain all the same information as the single table above. This time, however, we have organised the data into a relational database. Try to answer the questions below:

1. What are the names of the tables?

2. How many times has Sarah Layton's name and address been entered into the database?

3. If Jamie Ambler changed address, how many records would you have to edit to keep the data current?

4. If you wanted to print a list of the results, how would you go about it?

Players

ID	First_Name	Last_Name	Address
1	Sarah	Layton	15 Wardell Rd
2	Jamie	Ambler	71 Seaview St

Teams

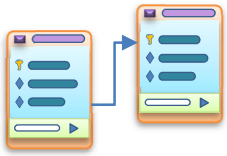
ID	Team	Team Strip
1	Lansville	Red & White

Fixtures

ID	Game_Date	Start_Time	Home_Team	Away_Team	Home_Goals	Away_Goals
1	29/09/2018	2:00 PM	Lansville	Granthorp	3	1
2	6/10/2018	3:30 PM	St Marks	Lansville	2	2
3	13/10/2018	2:00 PM	Lansville	North Howe	2	2

Relational databases make the storing, updating and retrieving of data much more straightforward.

Database Project: Planning (page 2)



Task 3 – Relationships

Relationships dictate how the tables in a relational database are connected. The tables on the right are from a database owned by a company who sell sporting goods. Look through the tables and work out how they are related.

Note: *This is a simplified database. It assumes each customer only purchases one item.*

Use the tables and connections shown to answer the questions below.

Suppliers

Company	Email	Contact	SupplierID
Elliot and Co	simone@thebagco.	Simone Edwards	Supp01
Master Systems	mh@mastersystem	Mark Harvard	Supp02
Grenville Sports	anika@grenvillespc	Anika Lo	Supp03

Products

ProductID	Product	Cost	SupplierID
Prod01	Scoreboard	269.50	Supp02
Prod02	Net	95.00	Supp03
Prod03	Ball	39.95	Supp03
Prod04	Posts	115.75	Supp01

Orders

OrderID	ProductID	CustomerID	Price
Ord01	Prod03	Cust03	69.95
Ord02	Prod01	Cust02	445.00
Ord03	Prod03	Cust01	69.95
Ord04	Prod02	Cust03	145.00

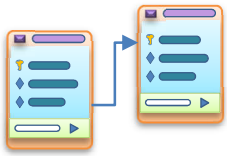
Customers

First_Name	Last_Name	CustomerID	Email
Jo	Summerby	Cust01	josumm1@her
Eric	Chang	Cust02	echang@gone
Jenny	Williams	Cust03	simonwilliams

1. How many orders were taken during this time?
2. Which product does *Elliot and Co* supply?
3. Which *OrderID* relates to Jo Summerby's purchase?
4. How many balls were sold altogether?
5. What did Jenny Williams purchase?
6. Who should the company contact to stock more nets for the store?
7. How much profit does the company make when it sells a net? This is the difference between the amount it pays the supplier and the amount it sells the net for.
8. Eric Chang had a problem with the product he purchased. It was over 12 months since his purchase and the warranty had expired. The company suggested he should contact the supplier directly. Who should he contact?

Task 4 – Database Design Project: The Purpose

Your challenge is to build a database to run a small competition for a sport or hobby (football, netball, gaming, chess or anything else with participants, teams and matches). The teams should all play against each other over the course of the competition. You will need to keep a log of the players, the teams and the matches (similar to the example in Task 1). The first step is to think about the reports that you would like to generate from the database – this is what it is for, after all. We have come up with four ideas below. For each one, suggest the details that might be displayed in the report.



SQL stands for *Structured Query Language*. This is programming language used to retrieve information from, or make changes to, a database. It's not normally necessary to think about the SQL that is being used in Access; creating a query is possible with either the *query wizard* or design tools. However, it is always SQL doing the work behind the scenes. The tools in Access are SQL builders; they help produce the SQL that communicates with the database.

Aim: *To look at basic SELECT SQL queries in a relational database.*

The SQL code for a SELECT query can be broken down into the four parts below. As we are now using more than one table, you must be specific about the tables that each field exists in. This means, for example, using the column name *Players.First_Name* rather than simply *First_Name*.

SELECT	some columns	*e.g. display First_Name from the Players table and Team_Name from Teams, in that order.*
FROM	some tables	*e.g. from the Players and Teams tables.*
WHERE	some criteria is matched	*e.g. only if the player's last name is 'Ambler'.*
ORDER BY	some columns	*e.g. sort by First_Name.*

As we are using more than one table, a join condition must be specified when you select the tables. For a simple join between two tables, the SQL code will look something like the example below:

FROM Teams INNER JOIN Players ON Teams.TeamID = Players.TeamID

Specify the two tables that data is being retrieved from, in this case *Teams* and *Players*.

Specify the join conditions, in this case only retrieve data where the *TeamID* is the same in each table.

Example 1: Find the team of a named player

SELECT Players.First_Name, Players.Last_Name, Teams.Team_Name

FROM Teams INNER JOIN Players ON Teams.TeamID = Players.TeamID

WHERE Players.Last_Name='Ambler'

ORDER BY Players.First_Name

Example 2: Create a team list for a certain team

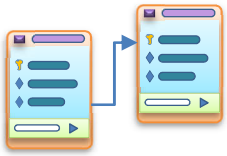
SELECT Teams.Team_Name, Players.First_Name, Players.Last_Name

FROM Teams INNER JOIN Players ON Teams.TeamID = Players.TeamID

WHERE Teams.Team_Name="City"

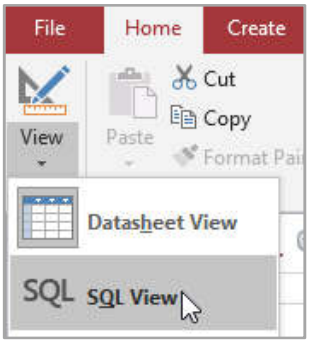
ORDER BY Players.Last_Name

Database Project: SQL (page 2)



Task 1 – Editing SQL

The two examples on the last page were very much like the queries created using the Query Designer in previous activities. We can simply open these two queries and view the SQL generated by Access.



a. In your database, double-click to open the query *Team from Name*.

b. Select '**Home / View / SQL View**'.

```
1 Team from Name
SELECT Players.First_Name, Players.Last_Name, Teams.Team_Name
FROM Teams INNER JOIN Players ON Teams.TeamID = Players.TeamID
WHERE (((Players.Last_Name)='Ambler'));
```

c. Compare Access' SQL code with the example on the last page. You might notice that there are some extra brackets and a semi-colon. Neither of these are required with this simple query; this is Access preparing for something more complicated. We also omitted any sort order in our original query as we only had one 'Ambler'.

d. View the SQL code for the *Players in Team* query.

```
2 Players in Team
SELECT Players.First_Name, Players.Last_Name, Teams.Team_Name
FROM Teams INNER JOIN Players ON Teams.TeamID = Players.TeamID
WHERE (((Teams.Team_Name)="City"));
```

e. Try editing the SQL and re-running the query. You could, for example:

1. Change the order that the columns are displayed by moving the fields listed in line 1. Remember to keep the *table name.column* format, e.g.:

   SELECT Teams.Team_Name, Players.First_Name, Players.Last_Name

2. Add a column, such as the Date of Birth (DOB) of each player in the team e.g.:

   SELECT Teams.Team_Name, Players.First_Name, Players.Last_Name, Players.DOB

3. Add or change the sort order in line 4. Try the code *ASC* (ascending) or *DESC* (descending) for any given column. You may also add further sorting levels, e.g.:

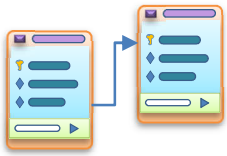
   ORDER BY Players.DOB DESC, Players.First_Name ASC

   Notes: *The code ASC is not required. If omitted, it's assumed you mean 'ascending'.*

   This is a good example of something that is not possible without SQL. You may sort on multiple fields in the query designer, but the sort will be applied to the columns in the order that they appear (left to right). With SQL, you may display the columns in any order and sort the data in any order.

Extension

1. Try typing some SQL queries from scratch and run them to see if they work. Debug any errors.
2. The *Fixture List* query from the last set of tasks is more complicated because there are two join conditions for the *Fixtures* table. Have a look at this and see if you can decipher the JOIN statement. The points total query is more complicated still.



You've most likely experienced the frustration of waiting for images to appear or videos to start playing on your device. The fact is that the data that makes up the media file must be transferred to your device, stored and then processed before use. The bigger the file, the longer this series of events takes to complete. It's in everyone's interest to try and make the file sizes of images, sounds and video as small as possible. The question is, are you willing to sacrifice quality in order to speed things up?

Aim: *To study how image files compress data for easier transfer, storage and*

Task 1 – Lossy and Lossless Compression

Complete the passages below using the words given.

colours	*data*	*storage*	*compact*	*people*	*product*	*lossy*
internet	*retrieved*	*original*	*clarity*	*lossless*	*processing*	*quality*

Files are compressed so that they require less _____ space, less _____ power and can be downloaded from the _____ more quickly. There are two main types of compression: lossy and _____. Lossy compression removes some of the _____ from the file resulting in a lower quality _____. Examples include images with fewer _____ or sound files with less _____. The discarded data cannot be _____ so the output will never be full _____ (although the differences may be undetectable to most _____). Lossless compression keeps all the data so that the output is exactly the same quality as the _____. Clever algorithms are used to _____ the data without losing any of it. The space savings of lossless compression are not as good as they are with _____ compression.

Task 2 – Testing Lossless and Lossy Compression

The image below-right is a rough approximation of a ball (not a very good one but we are only using 81 pixels and 2 colours). We'll attempt to compress this image using both lossless and lossy compression and see what the output would be like.

a. The Original Image

We are working with a 1-bit image. This means that each pixel is either black or white (although we have used grey so that ink isn't wasted by those printing this sheet). In our system, a black pixel is given the value 1 and a white pixel given the value 0.

Complete the bitmap code below for the image. It requires 81 bits.

000 111 000 001

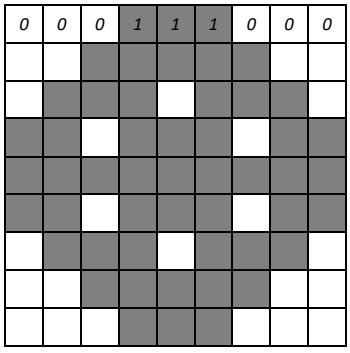
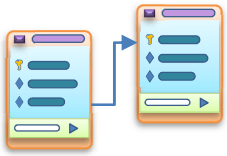


Image Compression (page 2)



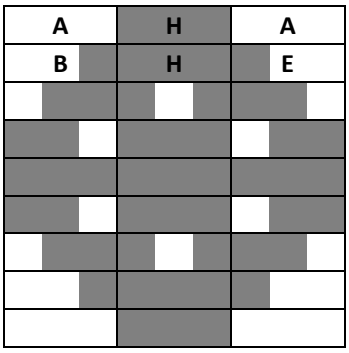
Task 2 – Testing Lossless and Lossy Compression (cont.)

b. Lossless Compression

There are many ways of encoding an image so that no data is lost. Our method will look at each block of 3 pixels and assign it a letter (A, B, C etc.) using the mapping below.

Complete the encoding by placing a letter in each block of 3 pixels.

A		000	E		100
B		001	F		101
C		010	G		110
D		011	H		100



What is the encoding for the image using this method? **A H A B H E** _____

How many characters do we now need to store the image? _____

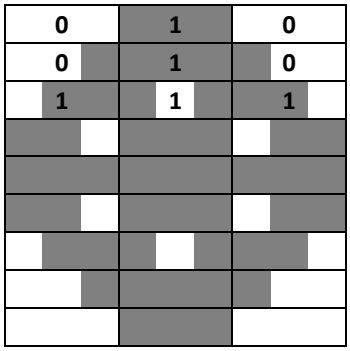
What must every device have before it can make sense of the code? _____

Note: *The more observant amongst you may notice that we haven't actually saved any data in this case; letters are generally encoded as 8-bit Unicode. However, the principle of looking at groups of pixels and assigning them a symbol successfully saves a lot of space when using larger blocks of pixels containing numerous colours.*

c. Lossy Compression

We'll finish with a very crude look at lossy compression. This time, we will take the average of each group of 3 pixels and assign it a single bit value.

- If 2 or 3 of the pixels in a group are black, we will give the group the value 1.
- If 2 or 3 of the pixels in a group are white, we will give the group the value 0.



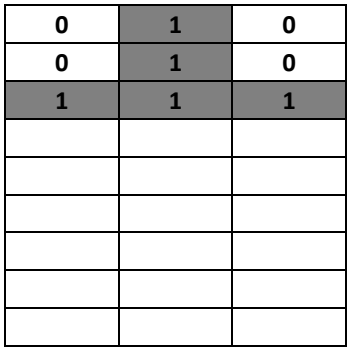
What is the encoding for the image using this method?

0 1 0 0 1 0 _____

How many bits do we need to store the image? _____

When it comes to reconstructing our image, we haven't got much data to play with. Each '0' value will have to be displayed as a group of three white cells and each '1' will be a group of three black cells.

Complete the reconstructed image on the right and describe how it is different from the original.



_____

_____

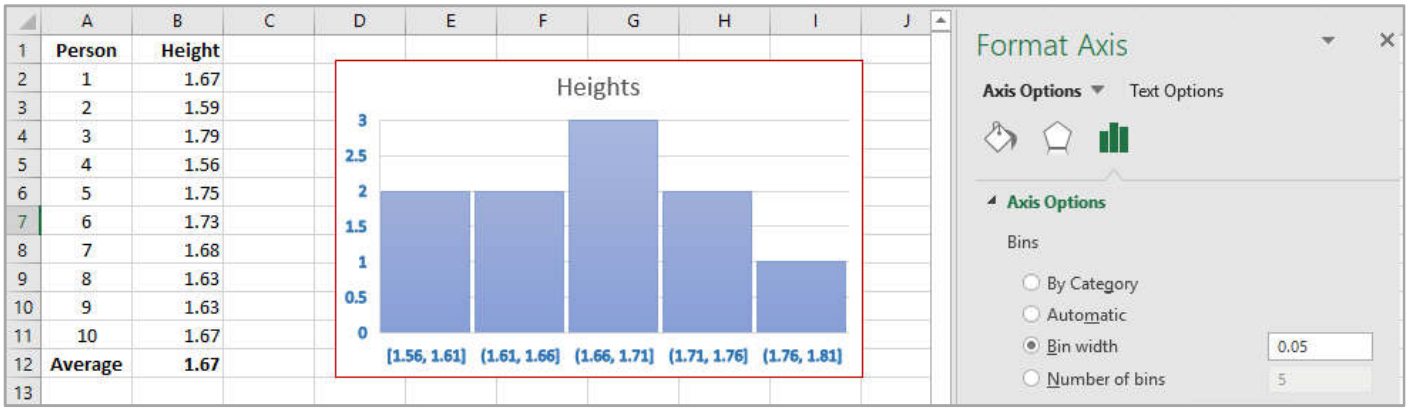
Data visualisation is all about displaying data in a visual way in order to help understand its significance. You will have probably used a spreadsheet to create bar and pie charts. We'll be looking at some other forms of data visualisation.

Aim: *To study a few different methods of visualising data.*

Task 1 – Histograms

Quantitative data might be discrete or continuous. Discrete data generally takes integer values, for example the number of points scored in a match. Continuous data can have any value, for example someone's height (if you accurately measured the heights of 30 people in a group, you'd get 30 different values). Histograms are a way of displaying continuous data on something that looks like a bar chart. It does this by collecting data from set ranges into sections called *bins*.

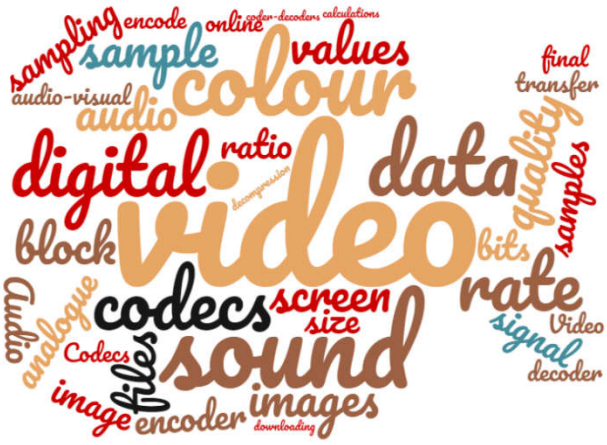
Collect, copy or create some continuous data and produce a histogram in Excel or Google Sheets (the options and results are slightly different in each application). The blue bars in the example below are the *bins*, each of which displays the frequency of occurrences within its range. Format your histogram using the tools available.



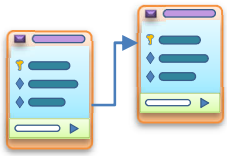
Task 2 – Tag Clouds

A tag cloud is a visualisation of words that are used in a piece of text. Tag clouds often depict keyword metadata on websites. The size of the words usually relates to the number of times they appear in the text. At times, they represent the significance of the word.

Create a tag cloud from a piece of text. There are lots of free tools on the internet to do this (we used www.wordclouds.com).



- You may start by copying and pasting text from a webpage. Wikipedia is a good source.
- You should be able to edit the wordlist once it has been created. Remove any words that are not specifically related to the subject.
- Choose colours and a shape etc.



It can be a lot of fun using spreadsheets to model how populations change over time. We will consider how quickly the rabbit population of Australia increased after they were introduced by a farmer in 1859 and then explore the massive numbers involved in cell division.

Aim: *To build ecological models in a spreadsheet.*

Task 1 – Breeding Like Rabbits

Rabbits breed quickly. They were introduced to Australia in 1859 by a homesick Englishman called Thomas Austin, who released only 24 wild rabbits onto his farmland as game for shooting parties. Within six years, the population had grown to 22 million. Your challenge is to build a spreadsheet that demonstrates how the number of rabbits can increase this quickly.

The first section of this task uses some simple assumptions to model an exploding population of rabbits. The extension task attempts to get these numbers under control so that we arrive at roughly 22 million rabbits after a period of 6 years. The assumptions you can make when developing your initial model are as follows:

- Thomas released 24 rabbits. We'll assume that this is 12 breeding pairs i.e. 12 mature females capable of reproducing along with 12 mature males (although in reality it's not likely to have involved all the males);
- Assume an average of 10 babies per litter - 5 females and 5 males;
- The gestation period for rabbits is about a month. It's a month before the babies are born;
- Female rabbits can get pregnant again almost immediately after giving birth. And they do;
- Assume that rabbits reach maturity at 6 months old. After this time, the offspring themselves can breed;

- For various reasons, some young rabbits will die. Assume that 40% of the rabbits in any litter do not make it through to maturity. This is an infant mortality rate of 40%.

Solution Tips

a. Create a new workbook named 'Spreadsheet Ecology' and a worksheet called 'Rabbits'. Add a top section where you can include assumptions such as the average litter size and mortality rate. Refer to these numbers in your calculations.

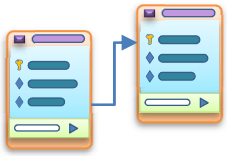
b. Your solution should be iterative. This means that you work through repeated calculations that use the previous month's results.

c. Start with the first month. Consider the number of rabbits born and how many will survive. How many rabbits will there be after 1 month? How many breeding pairs will there be at this point?

d. Fill your calculations down for 6 months. Things get more interesting at this time as the first litter will have reached maturity. Your formulas will have to change slightly. Get ready for the explosion!

	A	B	C
1		Initial breeding pairs	12
2		Average litter size	10
3		Mortality rate	40%
4			
5	**Months**	**Breeding Pairs**	**Litter**
6			
7	1		
8	2		
9	3		
10	4		

e. Fill these calculations down for 6 years. What is the size of the rabbit population at this point? If the cells are formatted with *Number* style then all the digits will be displayed (rather than 1.345E+10 etc.).



Task 1 – Breeding Like Rabbits (cont.)

Task 1 Extension A

After six months, our model arrived at a figure of 35,983,044,969,504 rabbits. This is about 35 trillion (using the more common US version of a trillion). The population of rabbits in Australia six months after their introduction was thought to be around 22 million. Why is our calculated figure so much higher?

Your next task is to develop your spreadsheet to take account of some further assumptions and see if you can arrive at a figure closer to 22 million (which is still an awful lot of rabbits, remember).

Further assumptions:

- Thomas' 24 rabbits might not have included 12 breeding pairs. Assume that only some of these were capable of breeding at the time of release;
- The litter size could have been less than 10 on average. It depends on the breed of rabbit introduced;
- The mortality rate of the juveniles might be higher than 40%;
- In the wild, rabbits generally only breed for 8 or 9 months per year. They don't tend to breed during winter;
- Rabbits don't live forever. You could assume that they live for about 2 years on average in the wild, although this will vary across populations. The fact that some of the original population were killed for sport will also affect this.

Solution Tips

a. Start by copying your last solution into a new worksheet called 'Rabbits Ext'.

b. Add a column which contains a "Y" if the month is in the breeding season and an "N" if it's not. You could use formulas to control how many months are in the breeding season from the top of the spreadsheet.

c. Add a couple of cells, one that displays the final number of rabbits (copied from lower down) and another which tells you how far away from 22 million you are (the error). This makes it easy to see the effects of changing your assumptions.

D	E	F	G	H	I
	Mortality rate	68.47%		After 6 years	21999996
	Months in season	8		Error	4

d. Use different fill colours for ranges of cells that contain different formulas, especially when they change part of the way down a column. This makes editing easier when you are refining your model.

e. Fine-tune your mortality rate until you get an error as close to zero as you can. With a mortality rate of 68.4714121% we were able to arrive at a total of 21,999,996 rabbits, only four away from the target of 22 million.

Task 1 Extension B

1. For a really difficult challenge, add a control that allows you to alter the average lifespan of the rabbits. One step might be to offer a choice between (say) 2 or 3 years, but any more than this is very tough spreadsheet project.

2. The story in Australia was probably a little more complicated than a farmer introducing 24 rabbits to his farm. Find out if there are any further assumptions you can make.