# Lesson

## Data Citizenship

In this lesson you will learn what data citizenship means and how you can identify and make use of it. You will:

1. Explain the use of data in society
2. Explain data literacy concepts
3. Interpret complex data

## Explaining the use of data in society

We will be able to:

1. Explain the technological, economic and societal reasons for the growth of data
2. Explain how data is used and misused by individuals, organisations and society
3. Explain types of bias and its impact on individuals and society
4. Explain types and sources of large datasets and the philosophy of open data
5. Explain the rights and responsibilities of data subjects and data owners

## Explain the technological, economic and societal reasons for the growth of data

Do you recognise the **use of data in business**? Why do businesses want your phone number, email address and social media details to name but a few?

Any business with a website, a social media presence, and accepts electronic payments of any kind is collecting data about customers, user habits, web traffic, demographics, and more. All that data is filled with potential if you can learn to get at it.

[Data helps businesses grow](https://www.grow.com/blog/data-important-business). Check out this link later.

1. Increase revenue (money made) by diversifying products.
2. Gain some customer trust as you target to them individually.
3. Identify areas where revenue (money reduced) is leaking from the business.
4. Identify specific customers that help increase revenue to then be targeted.
5. Monitor how many customers register, use the business and also return or leave.
6. Small businesses that have access to big data – shows trends can be more effective to change to the new trends quicker than big businesses.
7. **YOU ANSWERS**

The data needs to be accurate but also interpreted with no bias to ensure a suitable ‘probability’ of a positive result. If there are bias the data may be interpreted incorrectly and provide an incorrect result.

A picture containing person, indoor

Description automatically generatedWe are fed data based on a bias view all the time. For example a news feed telling you the ‘story of what is happening now and what has happened?’ The two tend to be completely different or inaccurate. News reports want to be first – but not as correct or accurate as the truth. WOKE, FAKE WOKE, lets just call it lies and truth. [Denzel Washington on ‘news’ 2016](https://www.youtube.com/watch?v=U3pV_Mw4mrM&ab_channel=Hame). Watch and listen to this solid set of words from an actor.

A person holding an object

Description automatically generatedNow listen to Tom MacDonald about [Brainwashing](https://www.youtube.com/watch?v=zCBNwGHPZ2M&ab_channel=TomMacDonald), he might not be your go to but here he is trying to show how the data could be used and controlled in the wrong way.

Okay I have displayed my anti-data stance. I know I am fed data that suits others’ agendas and I notice it some of the time. I watch BBC, SKY, CNN, RT, ALJAZERA, EURONEWS channels to see the different ways that ‘news’ is reported and find it odd that more people do not.

Every new channel has its own bias – I like **Euro news** because it reports are generally accurate in my mind (but maybe I am wrong?!?) but that’s just my mindset on the data that the news reports provide. You can easily get by in life by watching one new channel hoping that it is accurate data. The huge problem I see is that there are many news outlets including the Internet feeds, Tic-Tok and even darker and deeper Internet rabbit holes which make what you believe as true to be **inaccurate**. Data is used so many times to confuse us by so many outlets that the truth is hard to identify. Eventually you either agree with an outlet or similar outlets and live a happily informed life of their view or disagree and become more confused by what you think is true! Let’s not even talk about the data that has been **redacted** over centuries because it was inconvenient to admit what did happen when you’re trying to show a nice shiny history to the world.

The **Mau Mau** uprising - or I like to call it the Kenyan’s don’t want us robbing their country dry – but we did!

Timeline

Description automatically generated with medium confidence

No history is true. The **winners write the history**. In this case the government you and I live under lied and destroyed lots of the evidence of murders, rapes, tortures of more than 1 million people. You could say ‘hey its in the past mate’ but it was the 1950’s after World War II. That means we learned from the Germans about concentration camps and made them? I doubt it – it looks like its just a different version of the same thing that appears again and again throughout history – look at the Yugoslav Conflict. There were mass graves, concentration camps, rapes, murders and death.

So what am I trying to say? Data is manipulated to produce results. The British Government in the 1950’s manipulated the data and therefore the ‘story’ to tell the UK masses that Kenya was a friend after the revolt – we went in and ‘saved them’. Sounds like a play book for any 1st world country when they want resources and control of smaller countries – hey we live in a world where we disable countries from developing, unless they follow their masters orders.

In Kenya “God save the Queen and Country” and this was during QE2’s reign but she had just got the Crown. Maybe she didn’t realise?

What could be **legitimate or illegitimate reasons?**

We have seen that Kenya for example was a colonial take over – obviously there were resources that the UK wanted, or it was a strategic location – I have not read anymore on it. The issue could be that the UK government has their own legitimacy – perhaps they seen colonial expansion as a positive in the UK’s perspective. It’s lost in history or buried deep enough to cause woke and fake woke.

So historic data could easily be some made up story from some government, business and even individual(s). Do not worry much about it as no one really does!

**YOUR HISTORICAL EXAMPLES**

Do you recognise the use of data and the limitations in business?

A good way to identify data is locate outlets for a supermarket. I have made use of Google Maps to find Tesco Extra locations in Scotland. The initial search included Tesco Express, but I removed them. Unfortunately the map only indicated Tesco Extra outlets in the central belt of Scotland (that’s annoying).

Map

Description automatically generatedI searched using **tesco extra scotland -express**. This is data being useful to a point but since I could not get all of the outlets in Scotland that means there is a disadvantage because I want to get them all. A limitation of the data based on Google Maps search criteria.

I then did another search in Google which identified a website:

[**https://www.tesco.com/store-locator/directory**](https://www.tesco.com/store-locator/directory)

**It was listed A-C etc, which again was very annoying – why can’t I just see them all?**

**I then expanded all of the areas which had a little down arrow head so I could see them all – but this data is formatted in a way that Tesco want’s me to see – a subset of all the data.**

**So at the moment Tesco are intentionally or non-intentionally not letting me see all of the data in one go. I now note that Dumfries is shown as an area. Why? A decision made by a designer or coder. Again I see all of the data for Dumfries including the Express store!**

**Graphical user interface, text

Description automatically generated**

**Back at the start! I still cannot find a definitive list of stores that are not Express in Scotland.**

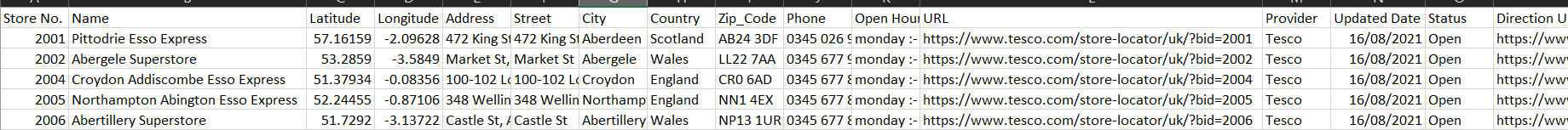
I went back to my initial search and tried:

Map

Description automatically generated**tesco list of all stores csv**

This use of the **csv** means I am looking for a document formatted as csv which is a **comma separated values** file. It brought me to here: <https://www.scrapehero.com/location-reports/Tesco-UK/> There are lots of them.

Further down it gives me a breakdown and offers the download of a **data set** which is the data about the stores. This data set will cost **50 bucks**. So even after my variations of Google searches I still don’t have all of the data for free. I did get a sample of less than 10 locations so I can see the format of the data.



This shows that data even about Tesco stores is worth money. Data is worth money.

Below is an example of data that has been stolen and will be used for **illegitimate** purposes. In this case to access Nord VPN for free using another person’s login details (6000+)



Chart

Description automatically generated

So who would use this non-redacted? Those who want a VPN for free? Those who want to hide behind a VPN but not be easily traced? Criminals, sexual offenders, spy’s? Just folk?

Text

Description automatically generatedI have had my data stolen from 3 sites over the last few years. I do not care as each site is not used by me anymore. This means that my email and old passwords will be in some form of data that people may or may not try to access or sell – data is money BUT we all give it away for free BECAUSE the apps, programs, sign ups etc say ‘we want your data is that okay?’ They should pay us.

**YOUR EXAMPLES OF DATA BEING STOLEN OR RANDSOMED?**

## Activity 1 – Who owns you

Use the **powned** website or another and check all of your email addresses to see when you have been owned or not.

## Activity 2 – Who has stolen data in the news

Identify a group that has stolen data in the news recently and tell me what you found out about it.

## Activity 3 – Horror story

Tell me about an online horror story where data has been ransomed recently to force a blackmail pay out for a large business.

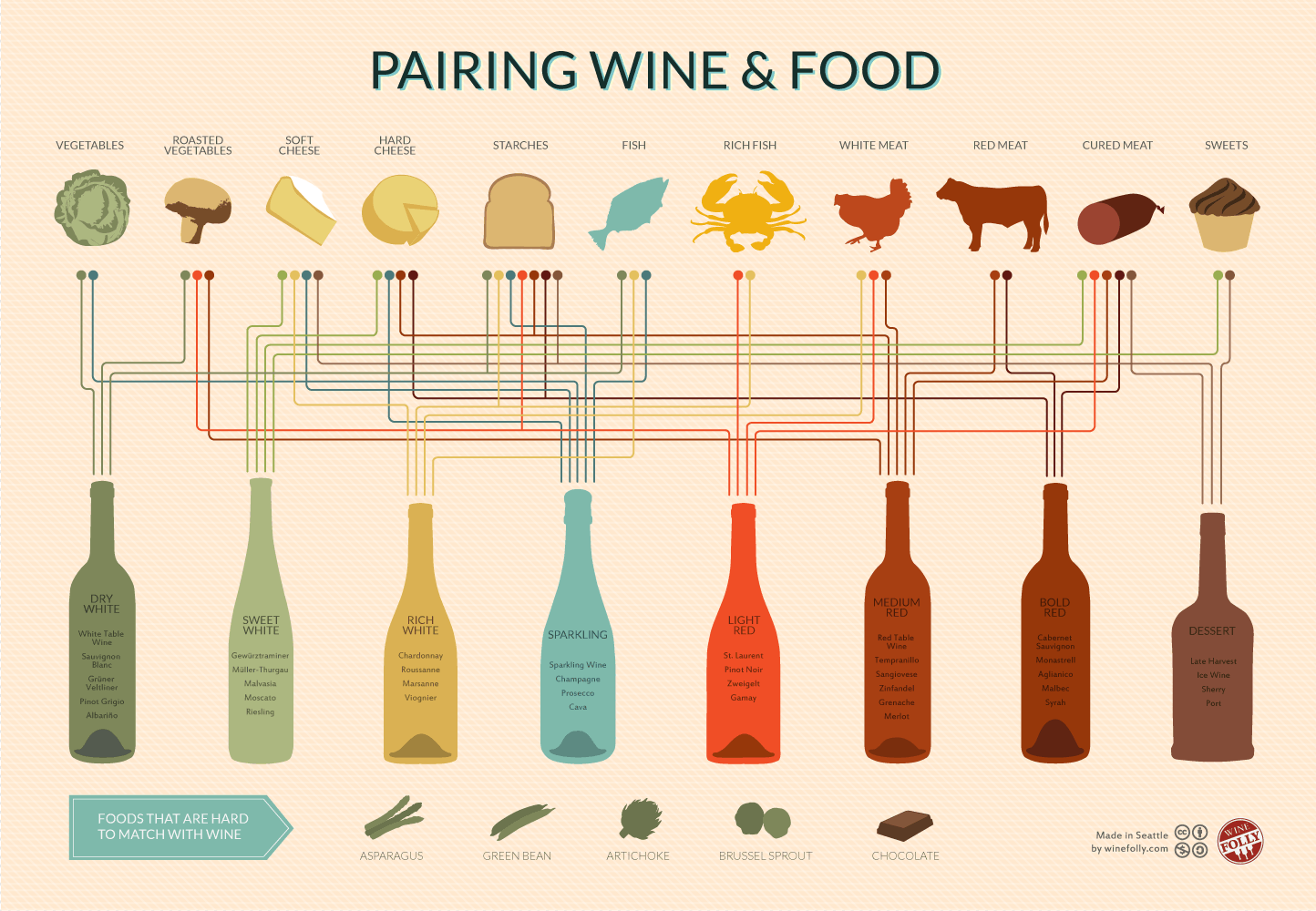
## Activity 4 – Your belief about the twin towers

Research the New York twin towers if you have not already and provide a 300-word breakdown of what you believe happened there as the planes hit and towers came down. This may be hard for some if a family member of friend died there. Try to be subjective on the news information and other outlets to provide your own short review using the data available. You can of course include links to relevant websites, videos etc.

Upload the [activity 1-4 evidence here](https://www.dglearnnet.com/moodle2/mod/assign/view.php?id=99762) on LearnNet.

## Data Visualisation

The concept of data visualisation means taking raw data e.g. csv data and other forms of data and producing the raw data as a visual version to help make others understand the data better.

This means that it might involve things like points, bars, lines, plots, bars etc. This might also mean some form of spider diagram or similar to show how the data is linked together. This also might involve a video explaining what the data means. The example to the right is a visual way of pairing food and wine together but comes from data such as a csv file. It’s all about how the data visualiser creates the visualisation.

Visualizations can include the following features :

* **Indicators**: These show the hierarchy and organization of a collection of data on a given subject. They highlight the most important information.
* **Simplicity**: The information is clear. “A picture is worth a thousand words”. The reader understands the information at hand immediately.
* **Brevity**: The message is short and clear, and no unnecessary information is visible.
* **Originality**: types of data, that seem unrelated at first glance, are collected and displayed in a way that offers readers a new perspective on the subject.
* **Visualization**: #7 is a good example of this, e.g. it means Number 7.
* **Colour**: to draw the reader’s attention to the most important pieces of information, clear and easy-to-understand colour palettes are used.
* **Aesthetics**: The graphics are lively, well-designed and pleasant to look at.

## Data Visualisation Reversed Activity

<https://www.toucantoco.com/hs-fs/hubfs/pairing-wine-and-food.png?width=1200&name=pairing-wine-and-food.png>

Using the link above and make the XLSX or CSV that might be used to then later on create the food and wine visualisation.

**Hint. I started from left to right looking at vegetables and then roast vegetables:**

Table

Description automatically generated

Upload the [Data Visualisation Reversed work here](https://www.dglearnnet.com/moodle2/mod/assign/view.php?id=100589).

## Data factors

The data that is sourced to be analysed and investigated have a number of factors that can affect the data and therefore the analysis and investigation. These are:

* Data Accuracy
* Data Timeliness
* Data Completeness
* Data Volume
* Data Variety
* Data Velocity
* Data Veracity
* Data Bias
* Data Value

## Data Accuracy

If the data that has been sourced is accurate then the investigation and analysis of the data will also be accurate. Of course if the data has inaccuracy this will skew the resulting analysis and investigation. The amount of inaccuracy will affect the analysis greatly if there is a lot of inaccurate data. Imagine 80% of the data was inaccurate and only 20% was accurate – you can see how the analysis and investigation of the data can make a huge difference to the result. Imagine you looked at a data set which showed data in ranges of millions e.g. so many people voted for Labour, SNP and Conservative – it provides a broad analysis of the voting practice of Scotland but loses the accurate for areas if the country and different types of individual people.

## Data Timeliness

If the data that has been sourced is gained just at the right time the result of an analysis and investigation should formulate quickly. However if the data is not gained at the right time but instead later on – the analysis and investigate will be hampered by this loss of time, providing a result later than expected. Imagine your data set was from 1990 so compared to now that data set could be very inaccurate as a representation of now.

## Data Completeness

If the data that has been sourced is complete i.e. what was wanted in its entirety then the analysis and investigation can begin confident in the data being complete. If the data is not complete, this can affect the result because the missing data may be important and alter the result. Imagine some of the data set has missing data – not helpful.

## Data Volume

If the data that has been sourced is a large volume of data it may take longer to analyse and investigate to produce a result. Large or even big data can also be looked at as providing a better result because there is so much of it, allowing for great analysis and investigation.

## Data Variety

Data that is sourced can come in a variety of different formats. Data can be structured (organised) and also unstructured (not organised). Data can be in the form of tables in spreadsheets, data in tables in databases, data that is just text and unstructured, emails, videos, audio. A huge variety of data in many different formats. The analysis and investigation of the data may be hampered by too much variety.

## Data Velocity

If the data that has been sourced arrives at different rates when used ‘live’. The velocity or speed of data arrival can be very quick indeed and the analysis and investigation of the data has to be done very quickly to keep up with the velocity.

## Data Veracity

If the data that has been sourced can be trusted (veracity). So how truthful the data is. There are lots of fake news and of course fake data that supports the fake news. The trustworthiness of the data also infers that you need to trust the data source provider as well. Social media is a great example of data veracity. Can you trust the source of the data when its trending i.e. is it just a huge lie that people choose to follow and trend? Data that is sourced from a trusted provider such as the BBC Weather, you can assume that the veracity is very good and trusted. Data that is sourced from a bias provider such as a far right group could easily mean the data has been manipulated and tarnished to provide a skewed data source that when analysed and investigated can provide the wrong results that work for the far right groups agenda instead.

## Data Bias

If the data that is sourced is bias in some way by the provider you will of course never gain accurate results from the analysis and investigation. Imagine the data that has been provided is only English and taken from English speaking countries only, how could that be a good representation of data that could easily be found in other countries and regions that do not store their data as English. Imagine the data set is about food preparation for curries that has been sourced from the UK and excludes India and other regional sources that could provide a more authentic and accurate data set.

## Data Value

The data that is sourced has value. Sometimes very little or none, but other times a high value. This value may be monetary (dosh) or valued in other ways such as precious to the analyst or investigator. When people trace their family tree there is usually a monetary value which they are willing to pay in order to gain the value of the data about their family tree which is priceless to them.

**Data Analysis and Investigation Activity START**

It is now time to take the ideas about data and analyse and investigate a data set.

**Currency for UK, USA, EURO, CHINA , RUSSIA FROM 2004**

https://trends.google.com/trends/explore?date=all&q=%2Fm%2F01nv4h,%2Fm%2F09nqf,%2Fm%2F02l6h,%2Fm%2F0hn4\_,%2Fm%2F025smh5

Graphical user interface, application

Description automatically generated

1. Which currency has been searched about most of the time?
2. Is the data timely? Explain why?
3. Is the data volume manageable as a dataset?
4. Does the data set have any value? Explain why?
5. Which currency has been searched the least over this vast period of years?

Upload your evidence here.

## Data Visualisation

This is where a data set is turned into a visual aid to help the user of the data understand it better and make judgements using the visualisations.

We are going to make use of the currency data set from Google trends (see above).

1. Download the dataset.
2. Ensure that the dataset is saved as an xlsx file.
3. Rename the tab to data. Just double click or right click RENAME to change it.
4. Cleans the data by removing any rows or columns that are pointless. How? Select rows that are not needed and right click DELETE. Same for columns.
5. Save the data columns as a table. How? Select all of the columns of data and then select AS TABLE. Pick a colour and click OKAY.
6. Create a new tab called visuals. Just click the + circle and give it a name.

## Adding in a slicer might not work

Select the entire table now and click **Insert** 🡪 **Slicer**. The window that appears will offer headings. Just tick each heading and the slicer will appear. Copy the slicers and place them on the visuals tab.

## Adding a line chart visualisation

Click on the visuals somewhere and click **Insert 🡪 Line or Area Chart**. Nothing is in the chart so click on the chart and locate **Select Data**. Click on the data tab and highlight the entire table and click OK. The chart will appear. You can customise the chart in various way including colours, fonts, wording etc. Now with a slicer and see the visuals update automatically.

**Visualisation Activity**

Using the information and example try and create a Visualisation for the following data.

Day,Quantity

1,12

2,22

3,32

4,42

5,32

6,22

7,32

8,52

9,62

10,52

11,53

12,62

13,63

14,72

15,42

16,40

17,32

18,40

19,52

20,49

21,25

22,35

23,45

24,35

25,40

26,45

27,60

28,65

29,55

30,50

31,57

Make use of a **chart** and a **slicer**.

Upload the visualisation here please.

## Counting raw data

## no jokes due to the sensitive content

In this example we will count the raw data from a CSV I will provide about births in Scotland hospitals.

1. Open the CSV and save it as an xlsx.
2. Do some data cleansing and ensure the only columns left are:

A picture containing graphical user interface

Description automatically generated

Note I renamed \_id.

1. Turn the data into a table. For example: A screenshot of a computer

   Description automatically generated with medium confidence
2. Examine the UniqueID column and tell me how many rows of data exist in this dataset.
3. Examine the FinancialYear column and tell me the start year and end year of the dataset rows.
4. Examine the Outcome column and tell me the different unique values used in that column’s data.

I am sure you have all noticed that there are no personal details about the babies and also the hospitals are referenced by codes. This ensures that the data would be difficult to be identified as breaking data protection act rules and regulations. No personal data has been identified.

Table

Description automatically generated

1. Time to rename the tab of the sheet. I want to call my sheet **data**. Just a matter of double-clicking the name and changing it.
2. Now add a new tab and name it analysis. To do that click on the circle with the plus. Then change the name.
3. While on the analysis tab, select cell B2 and type in the following. Table

   Description automatically generatedResize the column width so you can see all the text in the single cell with no overlaps.
4. Select cell C2 and type in =count(now watch me select data then I type ) and press the ENTER key. Graphical user interface, application, table

   Description automatically generatedThe count has counted all of the cells in that column which contain data so I automatically get to know how many unique IDs are there, therefore how many **babies**.
5. Graphical user interface, text, application

   Description automatically generatedNow I want to know how many baby outcomes were **live**. I will type in some text in the analysis tab and use a new function. I typed in =countif(now watch me select data then I type,“Live”) and press the ENTER key. Graphical user interface, application

   Description automatically generated There is an obvious difference between the two numbers. Now you cannot assume at this point although it may look as if all the rest did not survive.
6. Make more text and use the same function =countif but instead of “Live”, type in “Still”. See if you spot anything. Graphical user interface, application, table

   Description automatically generated Not quite 1308. At this point we need to examine the data to work out what else might be there instead. Perhaps the **domain expert** has already explained from the Maternity ward.
7. Let’s go back to the data tab and use the drop down filter on that column to see what else is there. Graphical user interface

   Description automatically generated I have checked for ‘Blanks’ and found none so there is no missing data in that column. I then checked ‘Unknown’ and found lots of entried stating ‘Unknown’. Reset the filter back to the original by selecting all the tick boxes. Did anyone see a quicker way to do the same?
8. Using the ‘Unknown’ status count how many are there and ensure you see the answer in the analysis tab. Graphical user interface, text, application

   Description automatically generatedSo we do not know what unknown means in this context and making a guess is a mistake – do not assume facts, find facts. You could of course guess that unknown might mean the babies died through complications at birth like being breach or ambilical cord issues, but again its guessing which is not worth while considering. Someone could try and play on this unknown and say “Were they murdered by evil nurses and doctors?”. Some people would data the data, make it’s outcome become bias to fit their own agendas. Don’t become one of **them**.
9. Now we will make use of the function called sum which will be used to add up numbers. The numbers we will add up are 6643, 1244 and 64 which should total up to 7951. I type =sum(now watch me select data then I type) and press the ENTER key. Graphical user interface, text, application, email

   Description automatically generatedAs you can visually see it is a match now so we have truly accounted for all of the babies in Scotland in this dataset.
10. Now I am interested in working out the obvious question. What is the percentages related to the three values. See what I have came up with as the sample. Graphical user interface, application

    Description automatically generatedThis initially works out what 1 baby is worth as a percentage of them all. I got E2. From that I used E2 to work out the 3 percentages. I then sumed or added all of the 3 percentages to come up with 100 which proves that I did the calculation correctly. Many people have difficulty working out percentages including **myself**. Lets examine E2 and see what I did.  This means take 100 and divide it by whats in C2. Now I worked out E3.  which means take whats in C3 and multiply it by E2. So in English it means take how many babes were Live and multiply that by how much 1 baby is worth as a percentage. Get your spreadsheet to work out all of the rest.

We could work out lots more using the data. Right now we have barely scratched the surface. We could ask the hard questions such as:

1. Which hospitals have more Live babies compared to others.
2. Which hospitals have more Still babies compared to others.
3. Which hospitals have more Unknown babies compared to others.
4. Based on the questions a to c is there any weird or anomylous data i.e. is any certain hospital suffering more Still compared to the rest.
5. What is the average per hospital for each of a-c.
6. Reverse the questions a-c and look for least instead of more.
7. Look again for data that stands out or seems weird or anomylous.

This is the analysis that you can use to identify problem areas as well as positive areas. You could also break it all down based on yearly performance per hospital and look to see if a hospital is displayin a negative performance compared to others or a positive compared to others.

Maybe you come up with not spikes or troughs in the data and find that all hospitals are operating in similar ways based on the live, still and unknown options. All of this is valid analysis.

Later on you could use this data analysis to then visualise your data using graphs and charts which would be a true reflection of the data with no bias.

1. Create a new sheet tab called **visual**.
2. Produce a **2D column** chart that takes the numerical values for live, still and unknown to show the difference. See my example. A picture containing chart

   Description automatically generatedI have altered the colours used and also edited the title of the chart.
3. Now make a pie chart that uses the 3 percentage values to show the data in a different way. See my example. Chart, pie chart

   Description automatically generated
4. Now [upload your completed spreadsheet xlsx file to here](https://www.dglearnnet.com/moodle2/mod/assign/view.php?id=108927).

**Individual Data Cleaning, Analysis and Visualisation Activity (Cancer)**

1. Locate a **dataset** from a suitable **open data** source. This dataset will contain data about the age of cancer deaths in Scotland.
2. Obtain a copy of the dataset and open it in Excel. Save the dataset as an xlsx file and call the file cancerscotland.
3. Perform data cleansing, by removing columns of data that are not useful or needed. You need to focus on the ages at which people died, the year they died, the gender and also what type of cancer was blamed.
4. Turn the data columns into a table.
5. Work out how many people under 25 have died of cancer?
6. Work out how many people under 25 have died of cancer in 2020?
7. Create a column chart to show 5 and 6.
8. Work out how many people over 59 have died of cancer?
9. Work out how many people over 59 have died of cancer in 2020?
10. Create a column chart to show 8 and 9.
11. Work out how many people under 25 have died of cancer in 2020 that are female?
12. Work out how many people over 59 have died of cancer in 2020 that are female?
13. Create a column chart to show 11 and 12.
14. Work out how many people under 25 have died of cancer of the bladder in 2020 that are female?
15. Work out how many people over 59 have died of cancer of the bladder in 2020 that are female?
16. Create a pie chart to show 14 and 15.
17. Upload the [completed spreadsheet here](https://www.dglearnnet.com/moodle2/mod/assign/view.php?id=108927).

## Data Types start here

Data comes in a variety of types. To start with there are 2 main types of data:

Icon

Description automatically generated

With quantitive it is easier to understand and analyse. You can count, sum, total, average, find minimums, find maximums and generally do lots of arithmetic and mathematics on this type of data. You can ask questions like ‘How many’, ‘How much’ etc.

With qualitative it is more difficult to understand and analyse. The answers tend to be open-ended and **not** measurable very easily. The data can contain opinions instead of facts. You can ask questions like ‘What to you think’, ‘How did it affect you’ etc.

Within these two main data types, there are spcific data types that come down to the actual type of data is stored in a dataset.

|  |  |
| --- | --- |
| **Integer** | Whole numbers only e.g. 7, 3, 9, 101, -7 |
| **Floating Point** | Decimal numbers only e.g. 7.3, 9.101, -7.3 |
| **Character** | A single letter or symbol on the keyboard e.g. “A”, “a”, “@” |
| **String** | A collection of characters e.g. “Welcome”, “This was my idea” |
| **Boolean** | This can be one of two values e.g. true, false |
| **Array** | A structure that can hold multiple pieces of data of the same type |
| **List** | A structure that can hold multiple pieces of data of different types |
| **Date** | A single date using yy/mm/dd e.g. 16/03/22 or other suitable format |
| **Time** | A single time using hh:mm:ss e.g. 20:19:09 or other suitable format |

**21 Questions Activity**

**Data growth**

Data is growing and being used. For example in **technology** to help analyse medical results from trials including Covid 19 to ensure vaccines are safe, to analyse internet network traffic to allow Google to manage their data centres better. Come up with one more reason for technology.

For the **economy** to help the government analyse financial information that could lead to recession or savings for the country. To help the NHS analyse the cost of operations to streamline their services better. To allow banks to analyse stock markets to get better results in trades. Come up with one more reason for economy.

For **society** to help social media platforms analyse what is trending or popular to then cater more to it and alert social media users to these trends. To allow advertisers to analyse social media chatter and trends to aim adverts at the correct group of people. To allow governments to analyse social media chatter and trends to aim politics at the correct group of people. Come up with one more reason for society.

**Data misuse**

Organisations sometimes misuse personal data of groups and individuals. An Example is Dorset Council put three children and their parents under surveillance, both at home and during their daily movements, simply to check whether they lived in a particular school catchment area. This was to prove that the parents were trying to obtain a place for their children at a popular local school when they lived a distance away. Another example is when personal data is used to produce a profile that puts individuals into groups with no proof that they should belong to these groups. This can lead to discrimination from all areas – race, finance, sexual orientation, age etc. Come up with one example of your own.

**Data Ownership**

In organisations, personal data is gathered, stored, managed and analysed. It is this temporary ownership of the data that must be regulated to ensure individuals personal data is not misused. The data owner is usually a responsible person within an organisation who has enough managerial power to ensure the data is not misused.

Responsibilities of the data owner include:

Data is lawfully gathered, accurate, secure, kept for only as long as needed, not given to other organisations unless permission has been gained to do so. Examples of your personal data include what.

**Data value**

Data that organisations hold about people, events, the weather etc has value. Some of that value is short term and others long term. For example in a weather-related organisation the data provided from satellites, weather balloons, weather stations can be turned into value to accurately predict weather patterns. Organisations related to climate change can use climate change data to provide predictive models of future climate change. Organisations involved in stock markets can use data from previous times to attempt to predict a rise or fall in stocks. Come up with three examples of your own.

**Data analysis**

When you analyse data you examine datasets looking for trends and occurrences that may help you come up with a reasonable for the trends and also allow you to come up with a predictive but informed decision or action.

Graphical user interface, text, application, email

Description automatically generated

This simple example extracted from Google Trends indicates that the ‘Im a celebrity – get me out of here’ show peaked in popularity around 12 months ago and now has the start of interest again. Why is this? Also why did the peak drop suddenly for the majority of a year? This is data analysis in action.

**Impact of data analysis**

There can be positive and negative impacts from working with data. Sometimes the data can gathered with bias, allowing for future analysis to be skewed. The data can also be incomplete which leads to incomplete analysis. The data itself could be analysed from a bias perspective meaning the analysis is aimed towards a pre-determined answer or action. Explain what data inaccuracy means in your own words and how this could impact on the analysis.

Explain what data timeliness means in your own words and how this could impact on the analysis.

Looking at the cancer dataset is this data qualitive or quantitive?

Looking at the cancer dataset columns that show ages, what data type is held in them?

Looking at the cancer dataset columns that show sex, what data type is held in them?

Complete the student survey here but before you click submit look at the different questions and tell me what data types are being used? For example either qualitive or quantative, string, integer, character etc.

Make use of the scotlandcrimearea.csv file which is located here.

1. Save as xlsx
2. Convert to table
3. Rename tab to data
4. Add analysis tab
5. Add visuals tab
6. Work out the total of crime for Scotland by sum function
7. Work out the most crime ridden area by sort filter
8. Work out the least crime ridden area by sort filter
9. Work out how many areas there are using counta function
10. Work out the average crime for Scotland by arithmetic
11. Make a column chart showing all areas and crime number.
12. Make a pie chart showing the most crime ridden area and number next to the least crime ridden area.
13. Make a column chart showing the most crime ridden area and number next to the least crime ridden area and also include the average crime number.
14. Make a column chart showing the most crime ridden area and number next to the least crime ridden area and also include the average crime number and then include Dumfries and Galloway.
15. Based on the chart in question 29, what can you tell me in words about Dumfries and Galloway using the data. This will involving you some thought to deduce correlations about the data.
16. If you wanted to live in a different area but not D&G that had just a little less crime which place would that be based on the data?

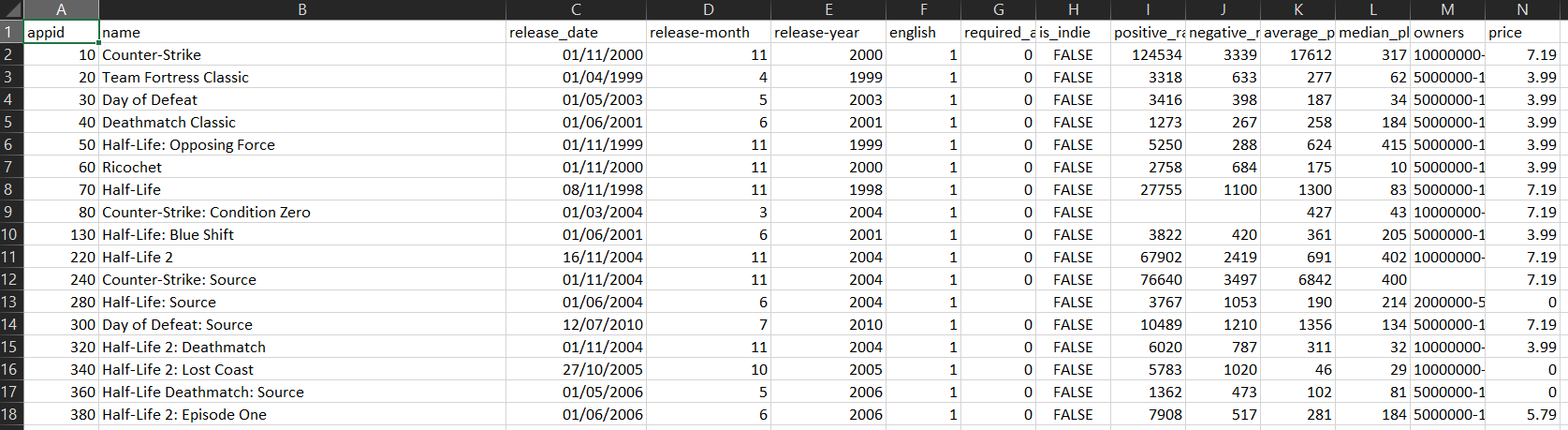
Upload all of [your 31 answers to here](https://www.dglearnnet.com/moodle2/mod/assign/view.php?id=109137).

A picture containing indoor, orange, close, cosmetic

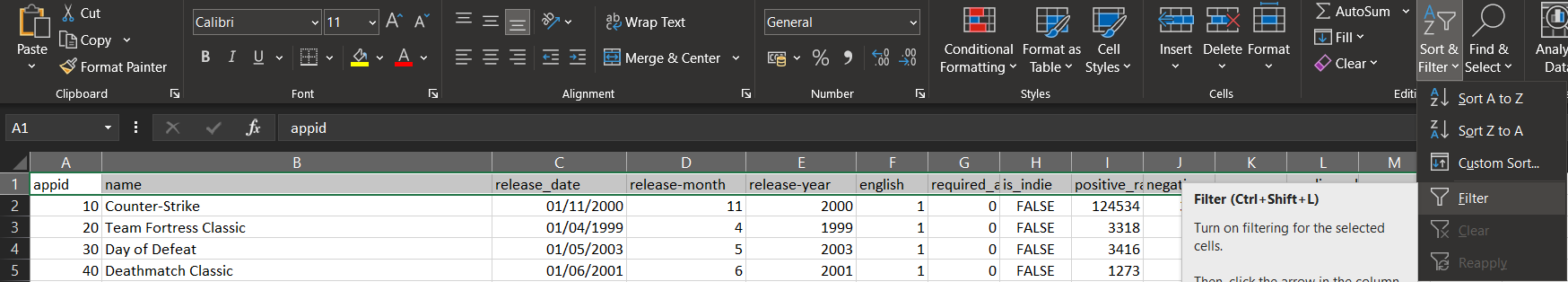
Description automatically generated

IGNORE ALL THE REST! THIS IS A LIVE DOCUMENT AND IS BEING EDITED AS TIME GOES ON.

the games data series provided with the assessment. If we were to try and count the number of years games were release i.e. year 2018, 207 etc it would be a very big and extensive job unless we make use of some smart thinking and a function called **countif**. Below is a small screenshot of the dataset and as you can see there is a release-year column that indicates the years games were released. They are very confusing as its all in different orders.



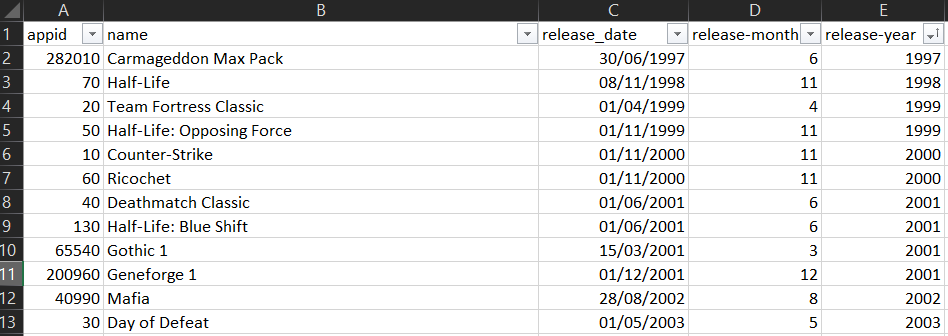
The first thing I would do is take the data set and add in something called a filter. To do this select the headings and then turn on a filter.

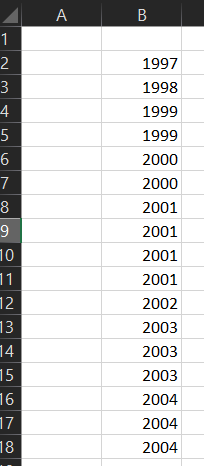


We can now see a filter.

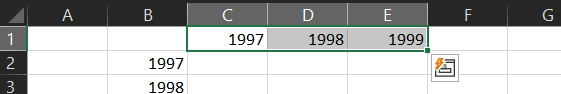


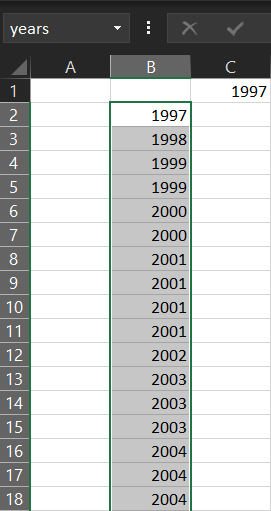
From there I would select the release-year and click on sort smallest to largest and this will show the games in year order starting at 1997 and ending in 2019.



We do not need the names of the games just how many there are release each year so I am going to select only the release-year data which goes down quite far. I selected it all. Copied it and created a new tab and pasted all of the data.

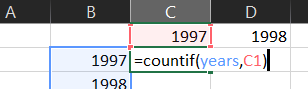
This will be used to work out how many games were release each year by using the **countif** function.

My best idea at this point is to make a set of columns that start at 1997 and ends in 2019. To do this you type in a few of them and then select and drag them to auto-fill the other years.

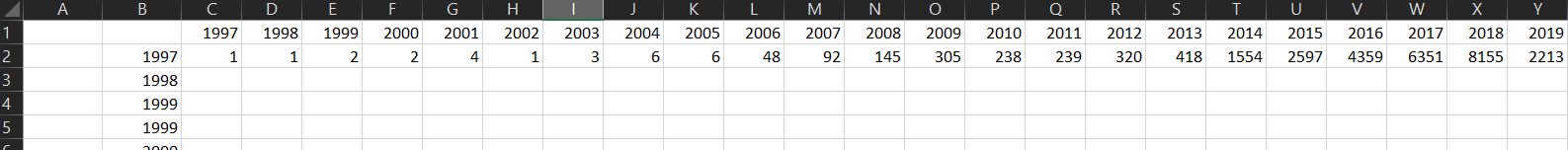


I will now select all of the years in column B and on the left box above type in years so that area I selected is now called years to make it easy to refer to later.

Under 1997 column I will type in =countif(years,=C1)

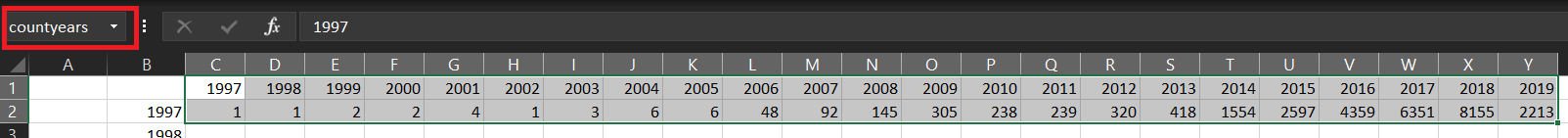


This will now say 1 because out of the years area only 1 is listed for 1997. Now grab the formula and drag it right all the way to 2019.



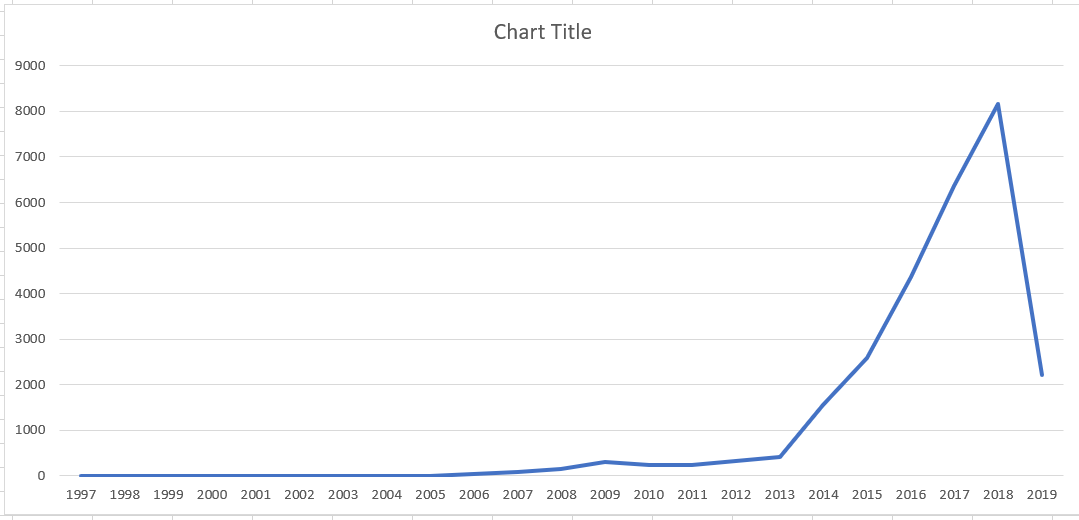
We now have an effective way to count how many games were released based on the year. Wow in 2018 there were 8155 games released.

The next thing I would do is highlight all of the years and counts and name that area countyears.

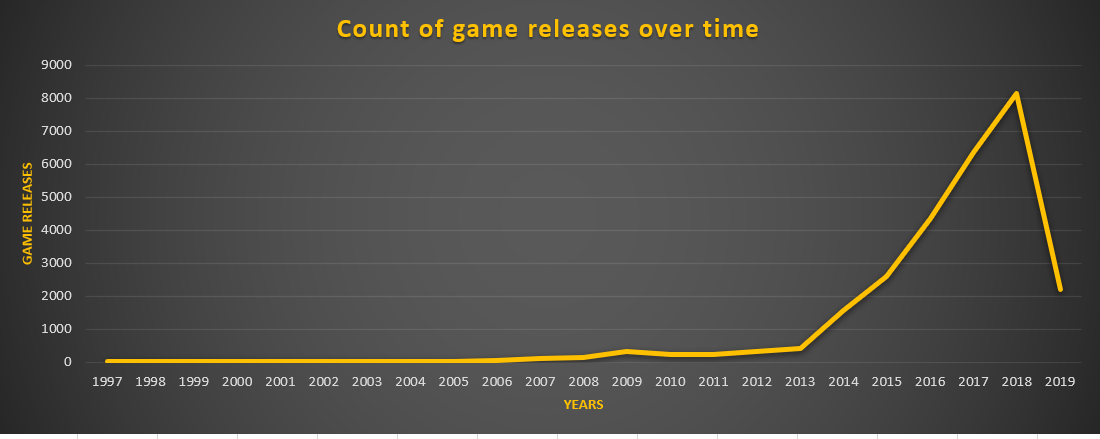


I am going to add another tab and call that visuals.

I then added a line graph and used that area called countyears. I now have.



Next is to make it look pretty and I came up with.



Now I can use this to analyse the number of games released over time and see that the number of games started to increase around 2013 to 2018.

## Identify the difference

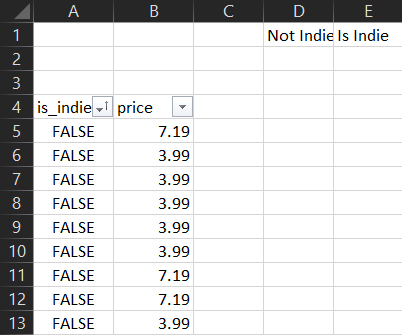
In this example I will use the game data set to see the difference between indie and non-indie games using the price as an average.

First I need the original games data set and then create a new tab called **indie and non-indie difference**. I want to show that using a simple bar chart I can see the average price of games for indie and also non-indie.

I am aiming towards something like this:

In order to build this chart it is really easy compared to others.

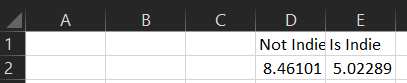
The first thing I would do is copy the is\_indie and price columns from the dataset into a new tab. I would also insert a few rows above the data to move it down a bit to leave me some space to work with.



This gives me all of the data and some space above. I then add in the filter and then sort the data of is\_indie by false so all of the is\_indie false appears at the top i.e. NOT INDIE. I also added in some more headings around D1 so I can work out the average game price for NOT INDIE and also IS INDIE.

With all of the FALSE and Prices selected I called that area nonindie. I then went to the cell below Not Indie and typed in =average(nonindie) which works out the average game price for all games that are NOT INDIE.

I did the same for the TRUE and Prices to work out the average for IS INDIE which provided:



Now just looking at this it is easy to see that Indie games sell for a lesser price than non-indie.

From this small amount of data I can easily produce a simple bar chart that gives a nice visual version.



Now from this visualisation I can answer questions about the Indie data.

## Typical information such as price

I used the same data set again to work on what is the prices between tha games that can show the typical prices and of course trends in prices with all the games.

I made a new tab and copied all of the prices into it. I realise that the price will identify via a nice visualisation how the prices change from high to low. This is what I am aiming for:

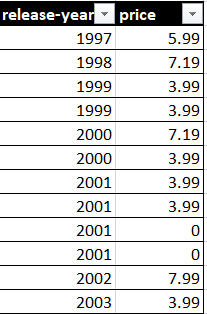
To achieve this type of graph I selected all of the prices and selected another column chart. I could of course select another but this one made sense. Now if I analyse what I see I can see a trend.

## Average prices per year

This is the only real complex idea. I know I have the game dataset and I would like to see how to identify the average prices of games per year. I am aiming towards something like:

In this partial example I started with 1997 and stopped at 2010 but of course this is a partial example, and you will probably need something that uses the entire dataset and will of course provide a more detailed and realistic visual.

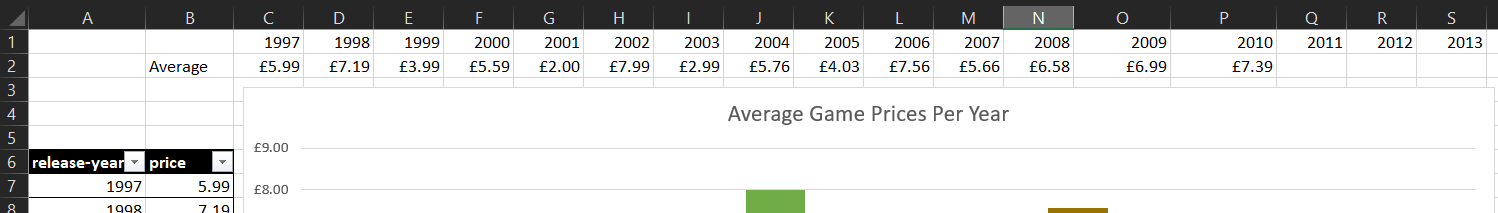
In this example I make a new tab and copied the release-year and price data so I can work out the average using the year and price.



I added in the filter and sorted from lowest to highest on release year so I can see the 1997, 1998, 1999, 2000, 2001 etc at the top.

Now it is going to become complicated slightly.

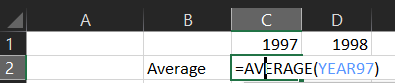
I make all of the years from 1997 to 2010 along the top.



Now I need to select all of the 1997 prices. I called them year97.

I did this for every year up until 2010 because this is a demo. You of course would work all the way to 2019.

Now that all of the year groups have been created I can now go back to the first year of 1997 and work on the averages.



Here I have used the year97 to work out the average – just press enter and see. I did this up to 2010 as it’s a demo.



Now using the years and average prices I can create the expected visual.

[Task 6 – go go go!](https://www.dglearnnet.com/moodle2/mod/assign/view.php?id=102380)

## Preparation for SOLAR test

Today you will cover materials that will help prepare you for the final test which will be a solar test, online, in college and closed book which lasts 60 minutes maximum.

There are 6 questions that you must answer by typing and each question is worth 2 marks in total = 12 marks for the test. You must get 50% of the test correct to pass, which means 6 marks to pass.

The test will take place in the afternoon in class on Monday 29th at 2pm. You will have some time in the morning to prepare. Susan will be coming into the class around 9.40am as well.