BIRN
BALKAN INVESTIGATIVE REPORTING NETWORK | ALBANIA

GETTING STARTED IN DATA JOURNALISM

By Crina Boroş and Lawrence Marzouk

Published by:
Balkan Investigative Reporting Network in Albania
Blv. Gjergj Fishta, Nr. 5/22, Tirana, Albania
http://birn.eu.com/

Authors:
Lawrence Marzouk
Crina Boros

Contributors:
Besar Likmeta
Ivan Angelovski

English Editor:
Marcus Tanner

Graphic Design:
Jurgena Tahiri

This publication has been financially supported by the National Endowment for Democracy (NED) and Rockefeller Brothers Fund (RBF). The views expressed here do not necessarily reflect those of NED and RBF.

BIRN
BALKAN INVESTIGATIVE REPORTING NETWORK | ALBANIA

	A	B	C
1	GETTING STARTED IN DATA JOURNALISM		
2		by	
3	Crina Boroş	and	Lawrence Marzouk

About the Authors

Lawrence Marzouk

Lawrence Marzouk is an editor for Balkan Investigative Reporting Network (BIRN).

He leads cross-border teams of journalists, deploying huge volumes of Freedom of Information requests, scraping and using old-fashioned reporting tricks to delve into high-level corruption in the Balkans and beyond.

He started his career in 2003 with regional newspapers in London and later in Brighton. He helped his first paper to national awards with his editing of the coverage of July 2005 London bombings and picked up nominations and prizes for his investigations at The Argus, Brighton.

In 2009, he moved to Kosovo to work for Balkan Investigative Reporting Network, where he edited its English-language newspaper Prishtina Insight and launched a new investigative journalism portal, Gazeta Jeta ne Kosove.

In 2010 and 2011, he won best anti-corruption investigation of the year at awards organised by the UN Development Programme in Kosovo.

Lawrence is the author of the Follow the Paper Trail, a guide to document-based investigative journalism in Kosovo and Exposing the Truth, a manual for journalists in Albania.

Since 2013, he has focussed on building and training cross-border teams of journalists for BIRN. Their stories have prompted criminal proceedings and caused political tremors, besides scooping national and international awards.

He has also been experimenting with new ways to present stories and working on making documents and data more accessible.

Crina Boroş is an investigative journalist with a focus on public accountability and policy transparency. She reports on conflicts of interest, vulnerable groups, problematic policies and use of public funds.

Her first articles were published in Romania at the age of 17, for which she won two awards for Best Feature and Best Interview. After wrapping up her Master's degree in London, she became foreign correspondent for a national news agency, covering the economic crisis, labour rights and current affairs.

In 2011, she started contributing to the Centre for Investigative Journalism (CIJ) in London, where she was mentored to become a data journalist and trainer. From there, she launched her UK career by freelancing for the BBC, Exaro News, OffshoreAlert and the Black Sea. She has produced front-page and cross-border investigations about corrupt authorities, tax-dodgers, abuse of minorities and war victims.

She has led large-scale survey-driven reporting projects about women's rights for the Thomson Reuters Foundation; has picked up a Best Investigation prize with ICIJ's Swiss Leaks team for revealing murky cash sheltered by banking secrecy; has sued the European Parliament over its lack of transparency for MEPs' general expenses; was nominated for a Paul Foot Award for her British fishing-rights reporting for Greenpeace - Unearthed; and has helped openDemocracy's initiative to investigate press freedom.

Crina is currently reporting with Investigate Europe and trains data journalism internationally.

crinaboros.tumblr.com

Crina Boroş

Contents Page

8

CHAPTER I - Introduction

10

CHAPTER II - What data journalism is and why you should bother

10 Definition
11 Case studies
11 Albania
13 Making a Killing
15 BIRN Macedonia: The True Cost of Skopje 2014
15 Winner of the Data Journalism Awards for Best Investigation: The Unfounded

17

CHAPTER III - Finding credible sources of data - browsing with Google

17 Google – the basics
22 Other useful domains

24

CHAPTER IV - Finding credible sources of data - beyond Google

24 Open data sources in Albania
26 International sources (shtim)
29 Freedom of Information (FOI)
30 FOI law in Albania
30 How to access information from institutions across the world
32 Scraping - Pentagon arms investigation
35 Big data leaks

39

CHAPTER V - Interviewing data

39 Finding stories in structured information - Maths and spreadsheet tools for journalists
48 Essential formulas
53 Useful numbers
63 Summarising with charts
72 Curating and summarising data
83 Selective counting
85 Digging deeper with pivot tables

104

CHAPTER VI - Preparing your data for analysis

103 Importing data
- 104 CSVs
- 106 TXTs
- 110 Importing data from web
- 113 Formatting data

114 Data cleaning quick tricks
- 114 Find & replace
- 114 Trimming
- 116 Concatenate
- 118 Text-to-columns
- 123 Splitting dates
- 129 Spotting problems

131 Recap

132

CHAPTER VII - How to visualize

135

CHAPTER VIII - Further reading

135 Sources on Data Journalism
135 Data journalist lexicon
136 Future Learning
137 Useful websites

CHAPTER I

Introduction

The world is awash with data. Some are of immediate use to journalists – a spreadsheet of politicians' earnings for example, or data showing crime rates in a capital city. Much, however, is not easily accessible or of obvious interest at first glance.

Who would have imagined that a Hollywood film would one day emerge from the pages of Boston's archdiocesan directories explaining the changing positions of clergymen? We are, of course, referring to Spotlight, the movie which told the tale of the Boston Globe's Pulitzer-winning, pioneering U.S. investigation into paedophilia in the Catholic Church. It led to global reform in part thanks to the discovery of patterns in dusty tomes that were then transposed to a spreadsheet.

This manual, although universally accessible, was written with Albanian journalists in mind. Its aim is to help reporters understand the power of harnessing data and deliver impactful story ideas, some of which, we hope, will hold power to account, expose corruption and wrongdoing, or be transposed to the silver screen.

We will start by showing you how to find sources of data through smart boolean searches, useful online libraries and crowdsourcing.

Open source resources available to journalists in Albania are growing, with a wealth of local, regional and national databases, so making good use of data is an increasingly efficient and powerful way of reporting.

We will also show in this manual how to clean and then interview data to obtain precise answers to your questions, revealing trends, anomalies or intriguing statistics.

You will learn about key tools to present your data in different formats: sometimes this will make it easier to spot the story, but it will also allow you to give your audience the information in easily digestible formats, such as charts and maps.

Over the past decade, data journalism has become a buzzword in media circles, grabbing the attention of hardcore notepad-wielding hacks, popping up on university syllabuses across the world, and delighting the many in our industry who revel in digital novelty.

What is important to remember is that data journalism is not a fad or distinct from everyday reporting. It is just a different way of approaching a story – by investigating how systematic a problem is – and a set of skills that complements traditional methods. Journalists in the U.S. have been doing it since the ‘60s, where computer-assisted reporting became a subset of skills investigative journalists employed.

This manual aims to set you on a path to adopting “data journalism” if you choose, but it also seeks to demonstrate that you don’t need to be a “data journalist” to use these incredibly useful skills in your everyday work.

CHAPTER II

What data journalism is and why you should bother

Definition

It is easier to define data journalism by what it is not. It is not social science, although we use polling, statistics and other related methods in reporting. It is not mathematics, although we need to know how to compute a trend and do primary calculations. It is not pretty graphs and sleek, interactive maps, although we often visualise for analysis and illustration. It is not hardcore coding, although we use code to analyse, scrape, or make tables talk to each other and to us. It is not hacking, we don't do that.

Data journalism comes from traditional investigative reporting.

The authors of this manual do not pretend this definition is exhaustive. But a traditional data reporter is an investigative or deep-digging journalist with high numeracy and solid digital skills.

We see data journalism as the craft of crunching numbers or analysing text or images or metadata to find patterns or outliers that could form the basis of a new story, or give us colour and clues, which the same journalist should be able to report out and develop into a story.

Whether the story is about Members of Parliament's expenses; how many times a politician uses the words "jobs" or "employment" in their electoral campaign speeches; the change in a country's forest surface over the past 50 years under the influence of questionable policies; gaps in air pollution data; or whether we compare how bias is affected by race, how race influences poverty and how poverty, in turn, influences performance. These are all legitimate examples of data journalism.

Sometimes data itself can be a story. This was the case with Wikileaks, Offshore Leaks, Luxleaks, Swiss Leaks, Panama Papers and all the other leaks where the revelation of information itself has been celebrated as a novelty item, as well as rocking the political, diplomatic, financial and media establishment.

But data will never tell you why something happens, or how. It may hint at the answer to these questions or what you should turn your attention to.

"What is data?" you may legitimately ask. It is an open question. But one convention still stands: if someone wishes to do data journalism, they need to gain a package of digital skills that allows them to:

- obtain or harvest information in a manner in which it can be (easily) analysed by a spreadsheet software, a coding language, an app etc.;
- clean and curate information into a format that makes room for manipulation or digital analysis;
- analyse the information with an eye for what a story is;
- visualise it for analysis;
- write compellingly using clear, easy to understand language.

Why should I bother learning the skills of a data journalist?

Investigative reporting is sometimes about taking what appears to be a one-off event and revealing and explaining that it is part of a trend or a system. Analysing data is a powerful way to do this.

In possession of a list of facts, such as a database, you're going beyond hearsay, an anecdote or circumstantial evidence, although reporting often starts this way. You can investigate whether a problem is widespread or an odd occurrence.

Plus, data allow you to follow money, map connections and scan for patterns of behaviour in a company, industry or government. It can also reveal the human impact of a policy,or illustrate consumer behaviour, to name a few favourite uses.

Case studies

a. Albania

BIRN Albania frequently uses data, both domestic and internationally sourced, as the foundation to build its investigations.

Here are some striking recent examples:

Data exposed lobbying spending:

BIRN Albania has used data from the US's Foreign Agents Registration Unit, FARA, and Albania's Central Election Commission (CEC) to investigate party political finances in a number of investigations. One of these stories, entitled "The LSI paid more than $300,000 for Lobbying in the US and declared a third to the CEC," revealed how Albania's former junior government partner, the Socialist Movement for Integration, LSI, splashed out on US lobbyists but did not declare its expenses properly in its annual report to the Central Electoral Commission. FARA "Supplemental Statement' forms, were used to calculated the total amount the LSI spent in 2016. That was then compared to the annual financial statement for 2016 that the LSI filed with the Central Electoral Commission. The LSI - under the lobbying/consultancy section of its financial statement - declared around 11 million lek ($85,000) in expenses for 2016. The investigation also detailed what benefits the payments made to US lobby firms like Global Security and Innovative Strategies (GSIS) and the McKeon Group secured for the party leader Ilir Meta. This included arranging meetings with senators to a ticket for the inauguration of US President Donald Trump. Meta was later elected Albania's President with the support of the ruling Socialist Party. His party is now in opposition headed by his wife Monika Kryemadhi.

In November 2017, BIRN Albania revealed that the Democratic Party had hidden hundreds of thousands of dollars in payments to its US lobbyist, Stonington Strategies, by routing payments based on a shell company in Edinburgh, Scotland. The secret payments were never declared to Albania's Central Electoral Commission. Following publication of the story, the Tirana prosecutor's office launched an investigation against the Democratic Party.

Business records reveal Albania's questionable Public Private Partnerships

BIRN Albania has used business registry records to shed light on Albania's public private partnerships, PPPs, particularly those in the health sector but also other PPP bids. Some of the stories looked specifically at companies that had been awarded concessionary agreements worth tens of millions of euros, their shareholders and investments, while others took a broader view on the profits for PPPs. Here are three examples.

- "The golden business of concessions secured record profits in the 2015" This analysed financial statements published at the business registry, the National Business Centre, of ten concessionary companies, such as Tirana Airport Partners, which runs Albania's only international airport, or ALEAT, which prints passports, revealing that these companies are the most profitable in Albania, raking in 85 million euro in profits in 2015 alone. The investigation also revealed that these companies often owe their profit margins to their monopoly status in the market.
- "The suspicious company that was awarded the 86 million dialysis concession This used a combination of business registration records, tender documents and on-the-ground reporting, to track the owner of Evita Sh.p.k – which won the lucrative PPP for the dialysis service – to a run-down family owned bar in the suburbs of the city of Berat. She was the cousin of the executive director of the Amerikan Hospital in Tirana, which would than take a minority stake in the

project, and had no previous experience in the healthcare field.

- "Olsi Rama's partner behind the VAT concession agreement bid" This used business registration records in Albania and Hong Kong to reveal that the company behind a concession agreement bid worth more than 100 million euros for the VAT collections service, MCN Sh.p.k, was owned by a business partner of Olsi Rama – the brother of Prime Minister Edi Rama. Following the revelation the contract was not awarded.

b. Making a Killing

International investigations that penetrate a secretive world can also be rooted in data.

When BIRN set out to the look at the flow of weapons from Eastern Europe to the Middle East, the task ahead was complex: how to infiltrate the shadowy world to reveal arms deals that the neither the countries nor the weapons brokers wanted placed in the spotlight?

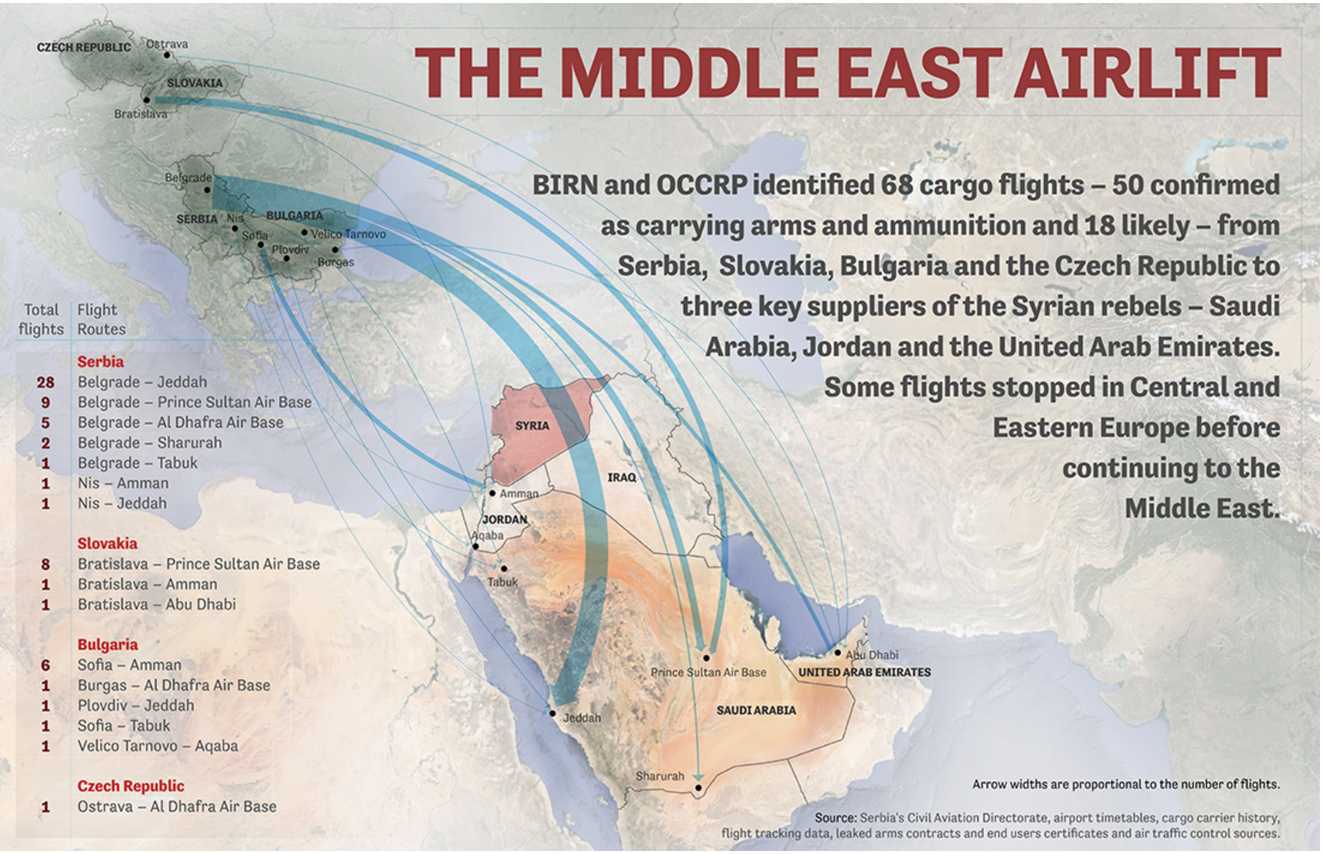
Based on a few news report and discussions with experts, reporters had a hunch that weapons from the across the region had ended up on the Syrian battlefield. But they needed to go beyond the singular to prove a worldwide trend so they turned to data.

The approach reporters hit on was first to identify open data sources that could provide proof of the weapons transfers. This was later divided into three strands: flight tracking, arms export reporting and social media analysis.

These were pinpointed as key to the story as they would provide proof that weapons were being flown to the Middle East; would demonstrate an unusual spike in the arms trade between Eastern Europe and, principally, Saudi Arabia, since the start of the Syrian war; and show that weapons from Central and Eastern Europe were in use on the battlefield.

This data analysis formed the backbone of the story which, alongside on and off-the-record interviews and a slew of leaked documents, revealed a 1.2billion euro arms pipeline to Syria [http://www.balkaninsight.com/en/article/making-a-killing-the-1-2-billion-euros-arms-pipeline-to-middle-east-07-26-2016]

Data from a flight tracking website (flightradar24.com), airport timetables and plane-spotting forums was entered into a spreadsheet of hundreds of flights.



THE MIDDLE EAST AIRLIFT

BIRN and OCCRP identified 68 cargo flights – 50 confirmed as carrying arms and ammunition and 18 likely – from Serbia, Slovakia, Bulgaria and the Czech Republic to three key suppliers of the Syrian rebels – Saudi Arabia, Jordan and the United Arab Emirates. Some flights stopped in Central and Eastern Europe before continuing to the Middle East.

Total flights	Flight Routes
	Serbia
28	Belgrade – Jeddah
9	Belgrade – Prince Sultan Air Base
5	Belgrade – Al Dhafra Air Base
2	Belgrade – Sharurah
1	Belgrade – Tabuk
1	Nis – Amman
1	Nis – Jeddah
	Slovakia
8	Bratislava – Prince Sultan Air Base
1	Bratislava – Amman
1	Bratislava – Abu Dhabi
	Bulgaria
6	Sofia – Amman
1	Burgas – Al Dhafra Air Base
1	Plovdiv – Jeddah
1	Sofia – Tabuk
1	Velico Tarnovo – Aqaba
	Czech Republic
1	Ostrava – Al Dhafra Air Base

Arrow widths are proportional to the number of flights.

Source: Serbia's Civil Aviation Directorate, airport timetables, cargo carrier history, flight tracking data, leaked arms contracts and end users certificates and air traffic control sources.

We tracked more than 50 weapons flights from Central and Eastern Europe to the Middle East.
Click here for a copy of our final spreadsheet of flight data.

- Vast amounts of data made public from sources such as UN trade data [https://comtrade.un.org/], EU arms export reports
- [https://eeas.europa.eu/headquarters/headquarters-homepage_en]
- and national reporting [https://www.sipri.org/databases/national-reports]
  were also transposed into a second spreadsheet which plotted arms trade trends.

A third spreadsheet was formed of data collected from social media accounts connected to rebel groups. Reporters pinpointed videos and photos of weapons from Central Europe, making notes of the dates, the group using them and details identifying the weapons or ammunition, such as lot numbers. With that information, reporters were able to trace the weapons back to factories in Central and Eastern Europe.

Data from these three sources all provided a compelling case that vast quantities of weaponry were being shipped to Saudi Arabia before being diverted to Syria. But reporters wanted to go further and found documents.

Equipped with an impressive body of evidence, they found it relatively easy to convince officials to speak on and off the record and to secure key internal documents

c. BIRN Macedonia: The True Cost of Skopje 2014

BIRN Macedonia has produced a series of innovative and award-winning databases drawn from analysis of public records.

One of the most striking was Skopje 2014 Uncovered [skopje2014.prizma.birn.eu.com/en], a project that included a database of buildings, new facades, sculptures, monuments, fountains and other structures that formed components of Skopje's vast, publicly funded makeover.

The eight-month investigation drew on data procured through the Access to Public Information Act, the official web page of the Public Procurement Bureau, the "Skopje 2014" audit and a joint report by the government, the Skopje municipality of Centar and the Ministry of Culture, presented after the 2013 local elections.

It revealed how from the initially announced price tag of 80 million euros, designed to give the Macedonian capital a new/old look, the cost had risen to 684 million euros. It also revealed the identities of the construction companies that had cashed in on the project.

d. Winner of the Data Journalism Awards for Best Investigation: The Unfounded

Like many stories, this award-winning investigation started with a snippet found in a small study from an Ottawa law professor into something called police service "unfounded" rates.

When police believe an allegation is baseless and that no crime has occurred it is recorded as "unfounded", a code Statistics Canada stopped recording in 2002.

The Globe decided that in order to get an accurate national picture, it needed to collect data from the roughly 175 police forces in Canada, using Freedom of Information requests.

Our probe revealed that from 2010-2014, one in five people who reported sexual assaults to the police had their case dismissed as “unfounded” and that sex-assault complaints are nearly twice as likely to be designated “unfounded” as physical assault allegations.

You can find more technical information on this story and other shortlisted piece from 2017 through this page:
www.globaleditorsnetwork.org/programmes/data-journalism-awards/dja-2017-shortlist

It is a great source of tips and inspiration.

CHAPTER III

Finding credible sources of data - browsing with Google

The world is awash with data in manifold form - locked into books, online and even in people's heads.

The first job of a journalist looking to use data for an investigative piece is to pinpoint what is, or could be made, available.

The second job is sourcing it – perhaps by crowdsourcing, clever online searches, a freedom of information request or poring over a reference manual at the local library.

In the next two sections, we will help you find credible sources of data using Google tricks as well as exploring other avenues to track down the information you need to make great journalism.

Google – the basics

While it's important to understand that Google indexes only about 20% of the entire web, that 20 per cent is often the most interesting and definitely the most accessible.

So, learning how to tease out great results from this internet search behemoth is an essential skill for any reporters.

Here are some tips to narrow down your searches:

Tip 1:

Imagine the search result before you type

Before you start banging search terms into Google, take a deep breath, think, and imagine what your ideal result would look like, like research guru Henk van Ess advises.

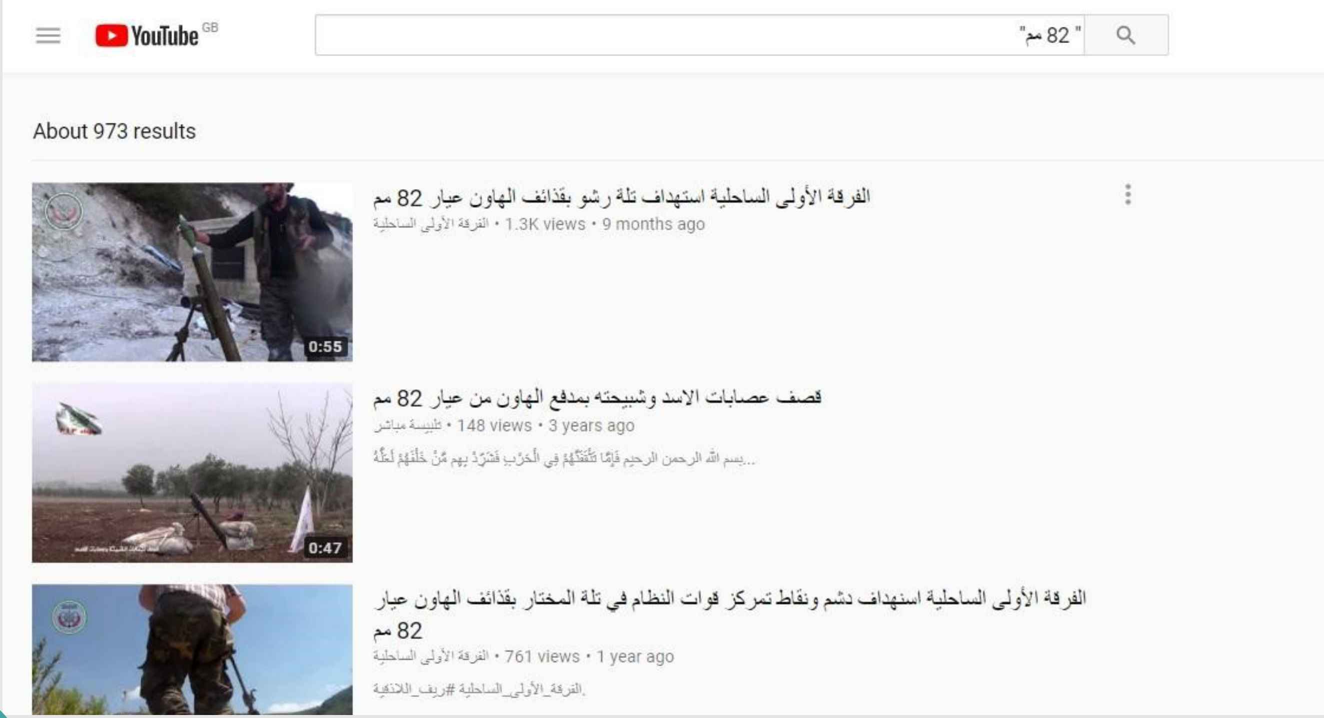
This is important because Google might be smart and can compensate for many of idiosyncrasies of the human brain, but it works much better with clear and careful instructions.

So, if you are looking for a spreadsheet containing crime statistics for Albania, typing "crime statistics for Albania" into Google might not be your best option. Why? Because that exact phrase may not appear on a spreadsheet, meaning that a Google search will miss it.

The first thing to consider is that any text in such a spreadsheet is likely to be written in Albanian. Second, it might not have a handy, summarizing title. Try, for example, searching for two different types of official crime categories written in Albanian which would likely be listed. You might need to spend a bit of time doing some background research at this point to get your search right, but it is worth it.

If you're looking to collect personal testimonies from online media, you need to remember that people writing about themselves tend to use "I" or "me" or "my" - a search term which isn't immediately obviously, but can help you hit the jackpot.

When we were trying to pinpoint Eastern European arms in Syria through social media, we learnt how militia groups referred to various items. Searches for " " 82 مم " mm in Arabic, a calibre of mortar shells mostly produced in Eastern Europe – were far more successful in pinpointing useful posts and videos than searches for "Serbian weapons in Syria".



" 82 مم"

About 973 results

الفرقة الأولى الساحلية استهداف تلة رشو بقذائف الهاون عيار 82 مم
الفرقة الأولى الساحلية • 1.3K views • 9 months ago

قصف عصابات الاسد وشبيحته بمدفع الهاون من عيار 82 مم
طليعة مباشر • 148 views • 3 years ago
بسم الله الرحمن الرحيم فَإِمَّا تَثْقَفَنَّهُمْ فِي الْحَرْبِ فَشَرِّدْ بِهِم مَّنْ خَلْفَهُمْ لَعَلَّهُ...

الفرقة الأولى الساحلية استهداف دشم ونقاط تمركز قوات النظام في تلة المختار بقذائف الهاون عيار 82 مم
الفرقة الأولى الساحلية • 761 views • 1 year ago
الفرقة_الأولى_الساحلية #ريف_اللاذقية.

Tip 2:

Quotation marks

Never forget to use quotation marks around your search term. If, for example, you are searching for references to Edi Rama, and do not include the quote marks ("Edi Rama"), Google will return pages with references to "Edi" and "Rama", but not necessarily the exact phrase "Edi Rama".

Remember that names are often misspelled in official and media reports. ry variants, and also consider whether the name has been transliterated from another script such as Greek, Russian or Arabic. There are a multitude of ways to spell Mohamad, or Aleppo, for example. Some people also anglicise their name, so "Ioannis" becomes "John". Others take nicknames or abbreviations.

Official documents often contain middle names and full official titles, so it's worth adding the ever-useful asterix "*" to your search term, indicating that there may be words in between.

So, if you are looking to do some research on the local administration in Brighton, and type "Brighton council", as it is often referred to colloquially, into Google you will miss out on most useful results. This is because the actual name is "Brighton and Hove City Council." Had you included the versatile "*", and searched for "Brighton * council", Google would also have also found you results for the official title as well as "Brighton Borough Council"; "Brighton City Council" and many more.

Tip 3:

Finding trusted sources

Google is the world's biggest library, but is largely unregulated. This means that swathes of information will need to be discarded because it's simply not credible.

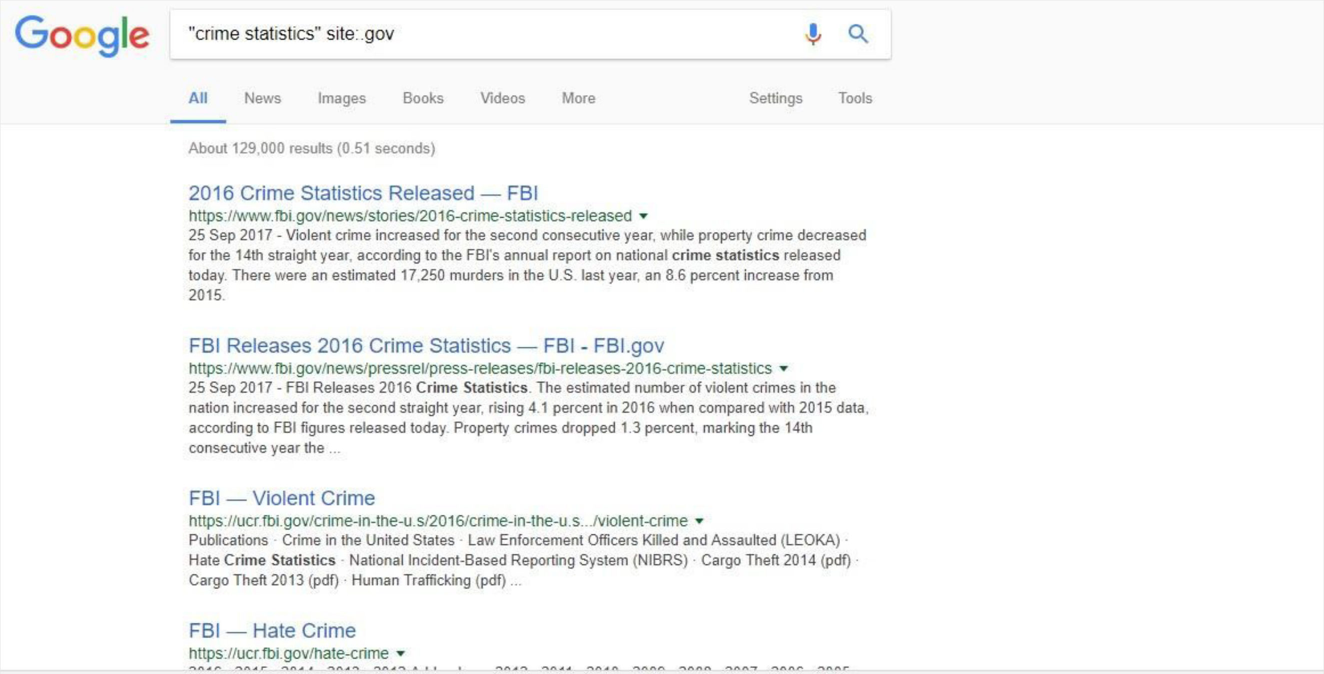
Luckily, there is a way of narrowing your searches to more reputable sources.

Site:

Governments are often excellent sources of information, and Google allows you to narrow your results to those found on certain webpages.

Most countries in the world have “gov” in the url of their official websites. The US has “.gov”, the UK has “gov.uk”, Albania “gov.al”, and Kosovo “rks-gov.net”. Exceptions exist within countries and internationally, of course: the Prime Minister' s office in Albania is kryeministria.al. German official websites end with just the country prefix of “.de”. Many parliaments have their own website: parlament.al for example in Albania and parliament.co.uk in the UK.

By typing your search term, followed by site:gov in Google, you will be searching for any web pages with your search term within websites with domain names that end in “.gov”. In simple terms, if “crime statistics” site:.gov is typed in a Google browser it would return pages that mention the term “crime statistics” on most US government webpages.



"crime statistics" site:.gov

About 129,000 results (0.51 seconds)

2016 Crime Statistics Released — FBI
https://www.fbi.gov/news/stories/2016-crime-statistics-released ▾
25 Sep 2017 - Violent crime increased for the second consecutive year, while property crime decreased for the 14th straight year, according to the FBI's annual report on national **crime statistics** released today. There were an estimated 17,250 murders in the U.S. last year, an 8.6 percent increase from 2015.

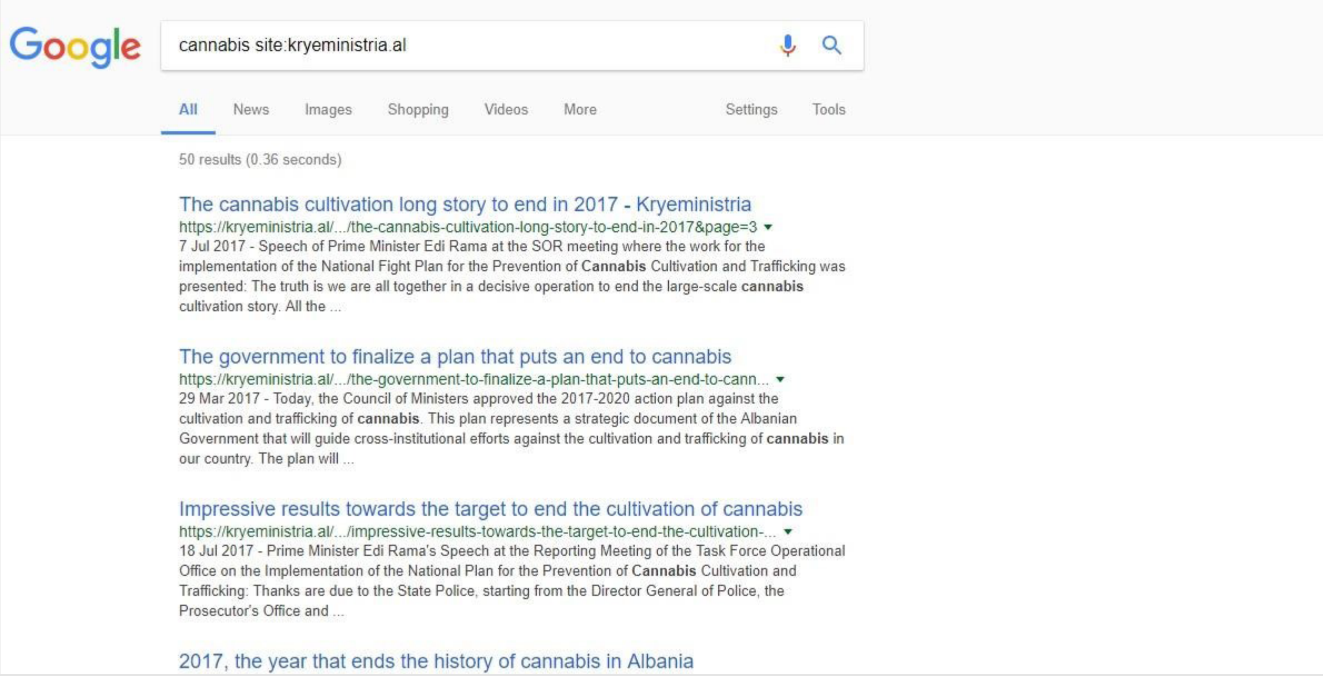
FBI Releases 2016 Crime Statistics — FBI - FBI.gov
https://www.fbi.gov/news/pressrel/press-releases/fbi-releases-2016-crime-statistics ▾
25 Sep 2017 - FBI Releases 2016 **Crime Statistics**. The estimated number of violent crimes in the nation increased for the second straight year, rising 4.1 percent in 2016 when compared with 2015 data, according to FBI figures released today. Property crimes dropped 1.3 percent, marking the 14th consecutive year the ...

FBI — Violent Crime
https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.../violent-crime ▾
Publications · Crime in the United States · Law Enforcement Officers Killed and Assaulted (LEOKA) · Hate **Crime Statistics** · National Incident-Based Reporting System (NIBRS) · Cargo Theft 2014 (pdf) · Cargo Theft 2013 (pdf) · Human Trafficking (pdf) ...

FBI — Hate Crime
https://ucr.fbi.gov/hate-crime ▾

If you want to search on a particular website, add the full url after “site:” For example:

- find any mention of cannabis on the Albanian Prime Minister’s website.



50 results (0.36 seconds)

The cannabis cultivation long story to end in 2017 - Kryeministria
https://kryeministria.al/.../the-cannabis-cultivation-long-story-to-end-in-2017&page=3
7 Jul 2017 - Speech of Prime Minister Edi Rama at the SOR meeting where the work for the implementation of the National Fight Plan for the Prevention of **Cannabis** Cultivation and Trafficking was presented: The truth is we are all together in a decisive operation to end the large-scale **cannabis** cultivation story. All the ...

The government to finalize a plan that puts an end to cannabis
https://kryeministria.al/.../the-government-to-finalize-a-plan-that-puts-an-end-to-cann...
29 Mar 2017 - Today, the Council of Ministers approved the 2017-2020 action plan against the cultivation and trafficking of **cannabis**. This plan represents a strategic document of the Albanian Government that will guide cross-institutional efforts against the cultivation and trafficking of **cannabis** in our country. The plan will ...

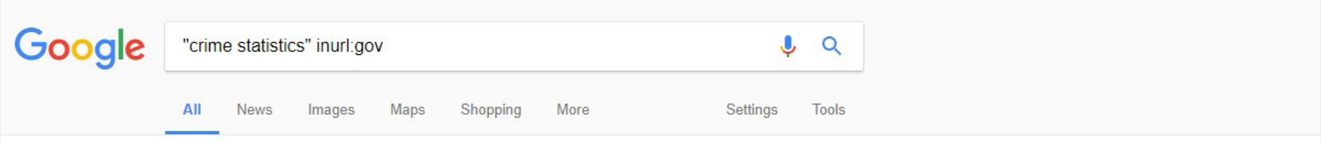
Impressive results towards the target to end the cultivation of cannabis
https://kryeministria.al/.../impressive-results-towards-the-target-to-end-the-cultivation-...
18 Jul 2017 - Prime Minister Edi Rama's Speech at the Reporting Meeting of the Task Force Operational Office on the Implementation of the National Plan for the Prevention of **Cannabis** Cultivation and Trafficking: Thanks are due to the State Police, starting from the Director General of Police, the Prosecutor's Office and ...

2017, the year that ends the history of cannabis in Albania

Another option is to add the versatile asterisk, called “wild card”, to the “site:” search. So, for example, “crime statistics” site:gov.* will browse through websites that end with gov.al or with gov.uk, etc.

⊖ inurl:

Typing a search term followed by inurl:gov will browse through web pages where “gov” is found in the whole url, therefore adding websites structured such as gov.al in Albania.



You will find that the first results include US, UK and Australian official websites.

Other useful domains

While “.gov” is the commonest and most helpful domain, many public bodies do not use it. You should, therefore, do some research beforehand to learn how useful websites structure their url.

For example, police departments in the UK end with police.uk, while in the United States – there are a variety of domains, from “.net” to “.gov”

The United Nations, World Bank and the Organisation for Security and Co-operation in Europe, OSCE, all use “.org”, which is common among not-for-profits and think-tanks.

Here is a list of useful domains to try:

:state. – used by some US state governments.
:parliament.* will search through dozens of parliament websites across the world.
:org useful for some international organisations, NGOs and think-tanks which are often good sources of data.
:.mil US military websites.
:.mil.* other countries’ military websites.
:.int popular domain with organisations such as the World Health Organisation, Interpol and NATO.

Filetype

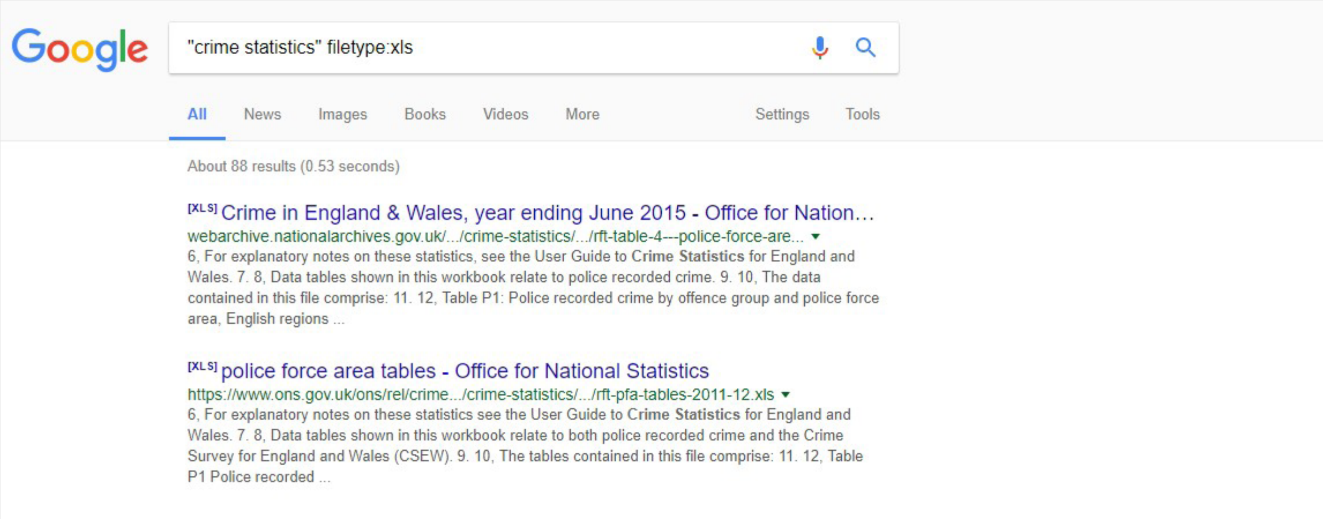
Narrowing down your search based on the type of document using "filetype:" is incredibly helpful when seeking official documents and tables.

Organisations and governments upload often upload important documents in the form of PDFs, Word documents, spreadsheets or presentations.

The filetype: search function allows you to hone in on these documents.

If you type your search term followed by Filetype:PDF and inurl:gov, Google will search for all pages with your search term in a PDF where the domain name includes gov.

The most frequently used file types are the following: pdf; xls (spreadsheets); ppt (presentations); doc (Word documents); xlsx - a Microsoft Excel open format spreadsheet; csv (Comma-Separated Values).



About 88 results (0.53 seconds)

[XLS] Crime in England & Wales, year ending June 2015 - Office for Nation...
webarchive.nationalarchives.gov.uk/.../crime-statistics/.../rft-table-4---police-force-are...
6, For explanatory notes on these statistics, see the User Guide to Crime Statistics for England and Wales. 7. 8, Data tables shown in this workbook relate to police recorded crime. 9. 10, The data contained in this file comprise: 11. 12, Table P1: Police recorded crime by offence group and police force area, English regions ...

[XLS] police force area tables - Office for National Statistics
https://www.ons.gov.uk/ons/rel/crime.../crime-statistics/.../rft-pfa-tables-2011-12.xls
6, For explanatory notes on these statistics see the User Guide to Crime Statistics for England and Wales. 7. 8, Data tables shown in this workbook relate to both police recorded crime and the Crime Survey for England and Wales (CSEW). 9. 10, The tables contained in this file comprise: 11. 12, Table P1 Police recorded ...

CHAPTER IV

Finding credible sources of data - beyond Google

While Google is a highly efficient way of quickly retrieving data, remember that it can at best search 20 per cent of the web (There are other search engines, such as DuckDuckGo, that may produce different results or protect your identity better).

Many of the most useful sources of data are walled off from prying eyes of search engines, often requiring you to search directly from the webpage. For key public institutions' websites which are indexed by Google (many are indexed in Albania), it is useful to browse them rather than pinning all your hopes on Google, or on one search engine.

Here are some useful sources of data, both open and closed, to fuel your stories.

Open data sources in Albania

Albania has a host of useful websites providing useful information and statistics in a variety of formats.

These range from **financa.gov.al**, which provides treasury expenditures data, including each single expenditure carried out by public agencies in the country, to one of the most comprehensive business registry in the world.

Here are 15 most useful websites, but remember that key information may also be found on other government websites.

qkb.gov.al The main database for business registration in Albania contains business registry, financial statements of biggest businesses (LLC and corporate) as well licenses that such businesses have. As such, qkb.gov.al is an unrivalled source for journalists. It contains names of owners, (companies or individuals), names of company board members, administrators and addresses. Journalists use this database for many purposes, from understanding the value or the return of a company to identifying persons related to a company or industry as potential sources of information.

financa.gov.al The official webpage of the Ministry of Finance. This ministry produces several reports that help journalists to understand the economic reality of the country. Here you can find revenues and expenditures on an aggregated level, but also information on imports of sensitive goods like tobacco or petroleum. The most important part of this website is the treasury expenditures data. These provide data on each single expenditure that a public agency in the country has carried out. Whether it is a municipality bill or a central government PR bill, it has to go through the treasury and will leave a trace on this database. Searching can be time-consuming,but rewarding. Since 2016, this website also publishes quarterly financial reports of state-owned or partially state-owned corporations.

instat.gov.al The website of the Albanian Statistical Agency, the main statistical body, contains statistical-level data and scores of reports on various issues. Journalists can use it to discover context about social issues, such as employment data or the level of education among the workforce.

instasgis.gov.al The platform for data visualization of Instat. It contains rich data on several levels, (national, local or grid) on scores of indicators, including employment, levels of urbanization etc.

dogana.gov.al The official page of the Customs Service in Albania contains a section of statistical reports on imports, exports and tax revenues of the government.

qbz.gov.al The Official Gazette of the Republic of Albania contains almost all government decisions, laws, rules and regulations.

amf.gov.al The page of the Financial Supervision Authority, which monitors and controls the financial sector, except for banks. It contains several reports, including data on securities and lists of companies and individual agents engaged in insurance.

ere.gov.al The website of the Energy Supervisory Board, which monitors everything to do with energy, especially electricity and gas. It is the place to go to find out what is going on in the energy sector, including companies operating or building power plants and as well traders.

parlament.al Parliament's website. It contains laws, or amendments of laws, and the explanatory notes of the proposed legal changes. It also contains scores of reports from several institutions and minutes of parliamentary debates. It is helpful, when a journalist is investigating an issue to search there and find out whether a parliamentary debate has been held on the issue previously.

cec.gov.al The website of the Central Election Commission. Here you can find the names of candidates (going back to 2000), or their electoral standings. When trying to learn about someone's background, searching here may help to see whether they have ever been engaged in politics. The political system in Albania functions in such a way that tens of thousands of people find themselves in the lists of candidacy of various parties.

app.gov.al The website of Albania Procurement Agency, a centralized agency through which almost all public procurements go. Journalists use it to understand who is gaining most from public money and as well, and find possible links between various state entities and certain businesses.

kpp.gov.al The appeals body for public procurements. When one of the parties disagrees with a public procurement decision, it appeals to the KPP, which then takes a decision and publishes it. Here you might find the names of people dissatisfied about certain procurement decisions who could prove useful sources in the future. .

planifikimi.gov.al The central authority for building permits in Albania. Construction is a politically-sensitive sector, so this might be worth searching for newsworthy decisions, or for the name of persons involved in one of the most lucrative and allegedly politically controlled markets in the country.

caa.gov.al The webpage of Albania's Competition Authority has information on ongoing investigations and the rulings of its Competition Commission.

gjykata.gov.al The main case file database of the judicial system in Albania includes First Instance Courts, Appeals Court and Administrative Courts.

gjykatatirana.gov.al The website of Tirana First Instance Court, the largest in Albania, has an electronic archive of cases going back to 2002. However, it's difficult to search since the court decided in early 2017 to anonymize the verdicts.

gjykataelarte.gov.al The website of the Albanian Supreme Court, has a database of civil and criminal case verdicts.

gjk.gov.al The website of the Constitutional Court has a database of verdicts.

International sources

A wealth of data is available from international organisations, civic groups and others, often in easy-to-use formats.

Here are a few essentials worth checking out:

World Bank Data: data.worldbank.org/
Hundreds of indicators – from access-to-electricity rates to short-term debt – are available from across the world. World Bank data is based on government-issued information. Consider using this as leverage with your own government or statistical agency if they appear unwilling to issue similar data.

Eurostat: ec.europa.eu/eurostat
Eurostat offers a dizzying array of statistics, from the usual trade data to the more obscure, such as minimum wages across the region or cargo tonnage transported between European airports.

World Health Organisation: who.int/gho/database/en/
From road traffic injuries to reported cases of cholera, this is the place to look for health-related data.

United Nations data: data.un.org/
The UN offers an extraordinary range of statistics, covering its many different organisations that collect data on tourism, refugees and development, to name but three.

OECD data: data.oecd.org/
The Organisation for Economic Cooperation and Development has data in more than 20 different fields for its 35 member countries.

USAID: usaid.gov/data
Data collected as part of USAID, the US government's development arm, has been made machine-readable. Grab your spreadsheet!

UN Comtrade: comtrade.un.org/
This easy-to-use international trade database allows to track movements of goods between countries. More information is available later in this section.

Engima: enigma.io
Enigma is a private database providing U.S. open data content in an easily searchable way. It is particularly useful if you are tracking U.S. public procurement records and transport information, such as bills of lading.

Open Corporates: opencorporates.org
A vast and ever growing trove of company data from 100 million firms. Extra search facilities are available for journalists. It is worth checking out 2016-launched opengazettes.com, which includes information from official gazettes.

Interpol databases: https://www.interpol.int/INTERPOL-expertise/Databases
A database of crimes monitored by the INTERPOL.

Investigative Dashboard: https://data.occrp.org/ and https://investigativedashboard.org/
A very useful free resource for company-related and corporate information.

Example: Using Comtrade data in investigations

The UN's Comtrade database formed a central part of one of BIRN's recent arms investigation.
Following a spate of uncomfortable headlines in 2013 about Croatia's role in supplying weapons for Syrian rebels, the authorities in Zagreb drastically reduced the amount of information that they make public.

In particular, Croatia stopped publishing data in its yearly arms export reports on the destination of any of its weapons or ammunition.

Undeterred, reporters from BIRN looked for an alternative source of information.

The first breakthrough came when reporters discovered that Croatia had continued to supply this information to the EU in its yearly arms export report. The downside to this was a time-lag of two years, so that data from 2014 was being published in 2016.

However, more up-to-date data emerged from an unlikely source – trade statistics provided to the UN.

This information is produced by Croatia's statistical agency and supplied to the EU (Eurostat) and the UN.

BIRN used the UN database Comtrade as it is simpler than Eurostat. We searched for exports by monetary value and weight from Croatia to Middle Eastern countries under the commodity code "93" for arms.

Under each parent code, there are sub-codes with four and six digits. So, "93" is for all arms, "9306" represents ammunition, and 930621 is "shotgun cartridges". You can explore these easily within the Comtrade website.

We discovered that the weight of the export was also provided under the four or six-digit code, as well as the monetary value, unlike for the two digit codes.

It is possible to search yearly, but also month by month, even in the current year meaning that up-to-date figures are often available.

Other commodities include tobacco products with the code “24” and drinks (including spirits) with “22”.

Freedom of Information (FOI)

Freedom of Information requests are a powerful means to obtaining key data for your stories.

Vast amounts of valuable information are generated by public bodies, but by no means are all of them automatically made public.

Here are some initial tips to approaching an FOI request.

a) Know what you are looking for. It pays to do background research and find out what data are collected, where the documents are kept and what they are called. All of this vastly increases your chances of getting the necessary information quickly and also means you can think around problems. For example, a request for information on the number of unpaid parking tickets issued to foreign number plates in the city of Brighton and Hove was turned down on the grounds that the council did not collect information specifically on foreign number plates. However, when asked to provide data on the number plates which differed from the standard UK format, the council obliged.

b) Most FOI laws are clear: even when an exemption applies to part of a requested document, the whole document should be released with the relevant parts redacted. Make sure that you point this out if an official attempts to deny you access.

c) A public interest test applies to most exemptions. Don’t be fobbed off. Try to prove this “overriding public interest”.

d) Appeal – FOI requests are denied for a variety of reasons, some genuine, others self-serving. Unless there are strong reasons not to, appeal a decision denying your request. We have been successful in overturning decisions in countries ranging from the United Kingdom to Serbia. Even if your appeal fails, remember that using all the legal resources available to you will make it easier for you and your colleague to obtain information in the future, as institutions become accustomed to behaving more openly.

a) FOI law in Albania

A Freedom of Information law has been on the Albanian statute books since 1999 and was updated in September 2014, providing greater access for the public to official documents as well as concrete penalties for public officials who refused to make information available.

While FOI journalism is far from commonplace in Albania today, it is also not unusual and it is possible to obtain interesting and important data through this method.

Three categories of institutions are subject to the FOI law in Albania:

- all administrative institutions, including the council of ministers, ministries, municipalities, government agencies, public advisory council and boards, as well as the armed forces, the police and the Republican Guard
- commercial companies where the state controls the majority of the shares, or companies that provide a public service. One example of such companies is the state-owned oil producer, Albpetrol, or the military export-import company, MEICO.
- Every person or company that through a law or sublegal act has been granted authority to perform a public service. This may include companies that perform public functions in the fieds of education, telecommunication or health. Although such companies are not necessarily recipients of public funds, the law judged that they have a direct impact on the public, which makes them subject to public requests for transparency.
- FOI requests can be submitted by an individual, including foreign citizens, or by organisations.

b) How to access information from institutions across the world

One of the most important skills that a journalist needs is the ability to find ways around problems.

If you can't get the information you need from your first port of call, you need to think of clever ways of collecting data from other sources.

That approach is key to Freedom of Information requests: if your request is turned down in one country, you should immediately think about what information you can request from other states or international organisations, such as the EU or the World Bank.

Most Freedom of Information laws do not preclude foreigners from applying for information. Even if they do, there are simple ways to get around this problem, such as asking a friend to apply for you.

In recent years, we have secured FOI requests related to investigations in Central and Easter Europe including from the US, UK, EU and the Netherlands.

United Kingdom

The UK offers an efficient and comprehensive FOI system. Citizens of Albania can submit FOI requests to UK public bodies, including those operating in the Balkans, such as the embassies, the Department for International Development or the British Council.

Almost every organisation has an FOI officer and contact details are usually available on the relevant website.

Documents available include internal correspondence, reports, statistics, etc.

To follow what other people are searching for, or submit an application yourself, you could use the website whatdotheyknow.com. It's an excellent website for tracking down interesting datasets which have already been put into the public domain. Before you use the website, remember that they tend to publish the requests and any interim answers on their website immediately. You may have to ask them to refrain from publication before the final answer.

United States

Anyone can apply for documents in the United States, where the FOI process is often slow but transparent.

European Union

Public access to documents is guaranteed at the European Parliament, the Council and Commission, and you don't need to be a citizen or resident of an EU member state to apply. You can use **www.asktheeu.org** to make the process easier, although this makes the request public on the group's website, unless you specifically ask the group to keep it secret.

The World Bank

The World Bank has its own Access to Documents policy, which means you can apply for previously unreleased reports and data. However, the WB is not subject to FOI legislations.

Other countries

Dozens of countries, ranging from Germany to Belize, also have FOI laws. For a full list, visit **http://en.wikipedia.org/wiki/Freedom_of_information_legislation**.

In Europe you can ask for the help of Access Info Europe (**www.access-info.org**) which works to promote the use of FOI laws by journalists.They have a useful website called Legal Leaks, which includes a handy handbook full of tips (**www.legalleaks.info/toolkit.html**).

Scraping - Pentagon arms investigation

Huge amounts of data are publicly available but poorly organised. One way to unleash the power of the data is to scrape it from its original source and repackage it in a spreadsheet.

Here's an example of how. "The Pentagon's $2.2 Billion Soviet Arms Pipeline Flooding Syria" was almost entirely based on public information that the United States government makes available on several websites like USA Spending or Federal Procurement Data System. But the only way to get to the interesting data we needed was to scrape.

Basically, scraping is downloading data from a web page or an online database and storing it locally, so you can play with it and organize it in a way that helps you work on your story.

Data are all around the web, but often organised in a way that is readable to humans, but not to machines. That's "unstructured" data.

Your goal is to make it structured, to end up with a clean set of data stored in a table with only the information you need for the story.

However, it's a long road to get to there.

Scraping is not simple. There are a lot of tools, browser add-ons and apps that offer one-click solutions but those often don't work on larger sets of data and can't handle multiple pages or, if they do, they are expensive.

Good scraping requires some coding skills. Even if you use free and open source tools like Scrapy or BeautifulSoup you will need to understand programming to follow online tutorials for scraping. The good news is that there are a lot of free online tutorials to help you get started.

Those tutorials will teach you how to extract information from a website and export it into a table that you can then manipulate.

Chrome extensions such as Scraper and Data Miner, for example, are excellent tools for simple scraping jobs.

A more complex and more expensive solution are websites such as import.io which enables you to make scrapers for multiple pages.

If you do have some coding skills, several websites, like USAspending.gov for example, offer an application programming interface, API, that you can use to dig your own data.

If not, there are readily available solutions like enigma.io or data.world that already have done half a job for you; they have scraped public databases so you can use it to dig your own stories.

That's what we did for the Pentagon story. We had spotted that for each public payment to a contractor, the origin of the goods was declared. This meant that it was possible to see where the US military was buying weapons from in Central and Eastern Europe.

Unfortunately, there was no easy way to search through the US federal procurement data system by "country of origin", so we needed a Plan B.

We decided to turn to enigma.io, a website that offers a comprehensive repository of public data. They have already scraped complete USAspending.gov, so we didn't have to, and could concentrate on filtering what was relevant to our story.

That sounds easy, but it was far from that. The dataset of US government contracts lists every dollar contracted by every US agency, so, with all the filters applied, the first set of data we downloaded was well over 100MB in one .csv file. CSV stands for Comma Separated Values, and it's just one big text file containing columns separated by commas and rows in separate paragraphs. You import that in a program that handles tables, like MS Excel, Libreoffice Calc or Google Sheets.

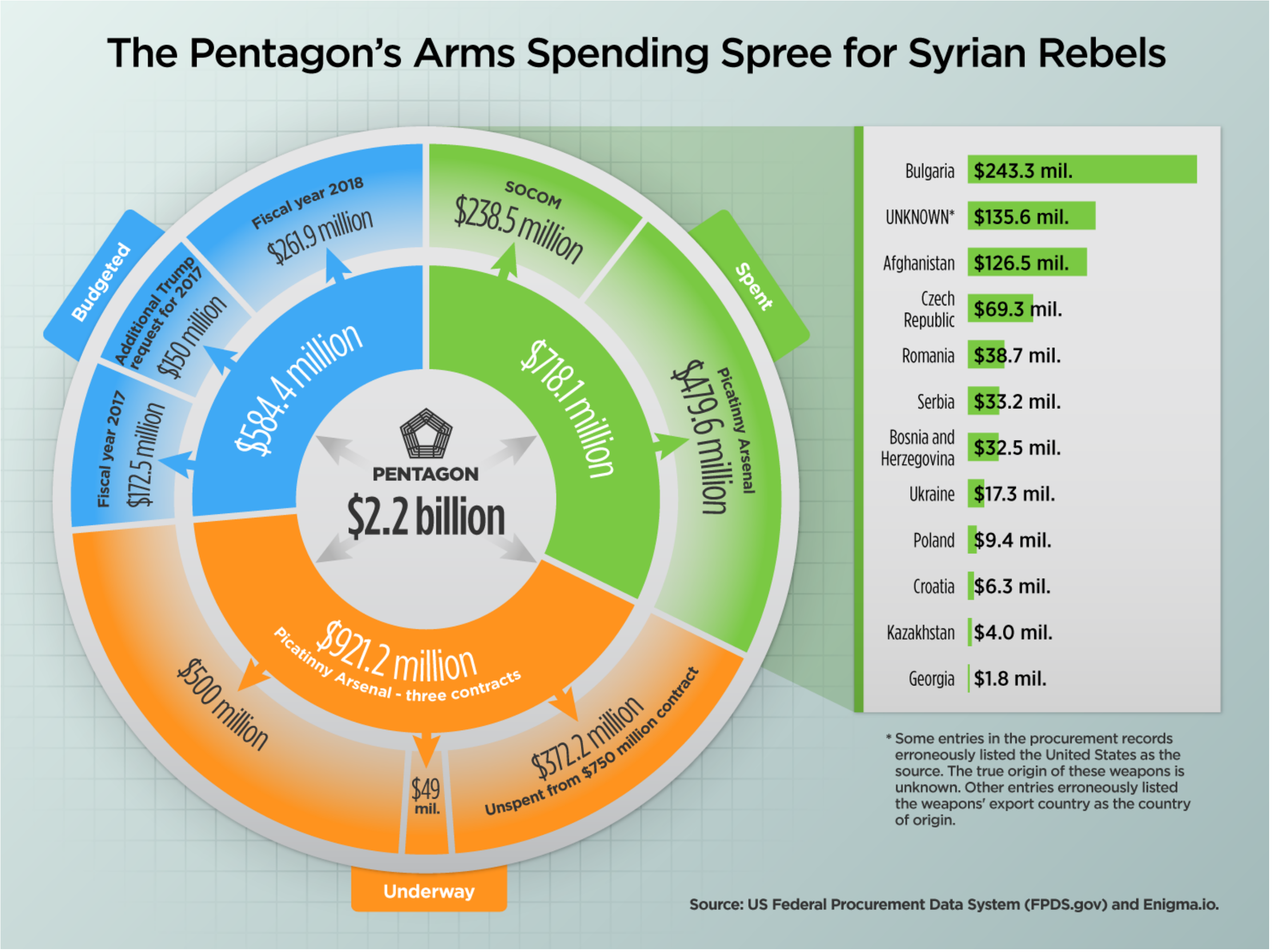
As we were facing this type of data for the first time, we had to become acquainted with it pretty fast.

The data contained all the weapons and ammunition procured by Pentagon from 2011 to 2016. But to figure out what we needed, we had to understand the procurement processes at the US Department of Defense and decode the technical lexicon.

Without doing background research, the data would have been meaningless.

Since we were all new to this dataset, we had to go back and forth several times before we were satisfied with the gathered information.

We focused our attention only on so-called "non-standard" weapons procured from Central and Eastern European countries and intended, or likely to be intended, for Syria. At the same time, because the clerks filing this information were sometimes careless, we needed to go through all other data and dig out valuable contracts hidden behind incomprehensible codes.


The Pentagon's Arms Spending Spree for Syrian Rebels

* Some entries in the procurement records erroneously listed the United States as the source. The true origin of these weapons is unknown. Other entries erroneously listed the weapons' export country as the country of origin.

Source: US Federal Procurement Data System (FPDS.gov) and Enigma.io.

This table was the end product of our number-crunching, revealing from where the US was sourcing its weapons.

After removing all the data that was not relevant to our work, we reduced it to less than 1MB, which was much easier to work with. That shows why it's important to know exactly what you're looking for, what your story is about and what data you need to tell it. Click here for a copy of our final dataset.

It shows why it's important to ask the right question. Only then can you expect data to give you answers.

However, to trust the gathered information, we had to verify it.

Luckily, US procurement data are stored in multiple databases.

To verify every contract we were interested in, we used the US online procurement records database in the US federal procurement data system. Likewise, we checked the most problematic contracts with officials at the Pentagon.

Thanks to all of that, we were able to catch the Pentagon's attempt to hide some embarrassing data. We found out that, after we started to work on the story and after asking the Pentagon inconvenient questions, the US Department of Defense staff went into FPDS database and changed several mentions of Syria. As we had stored the originals, we were able to expose this attempt to rewrite history.

Big data leaks

Large data dumps are increasingly common, so journalists need to be able to extract the information needed to produce a story.

So, what do you do when you get a leak of 300GB of messy data?

Handling that amount of data is not easy and you need to know exactly what you want in order to manage it properly.

There are a few ways such data leaks come to you. Either a source gives them to you directly, as in the case of the Snowden leaks or the Panama Papers, or a hacker uploads it on the web as a set of files you need to download, as in the case of Hacking Team or the Silk Way Airlines files, which we used for our Pentagon investigations - http://www.balkaninsight.com/en/article/the-pentagon-s-2-2-billion-soviet-arms-pipeline-flooding-syria-09-12-2017.

Or, maybe you just got a huge pile of paper documents you need to digitize in order to handle it properly, as in Yanukovych Leaks.

In any case, you need to act fast, especially if it's something of immediate public interest, or your competition is on the case.

Your job is to quickly assess the material, organize it, recognize the significance of each document, verify it and start publishing.

But the first thing is to sort it all out and get to know the data.

It can be overwhelming, so you need to get on top of it as soon as possible.

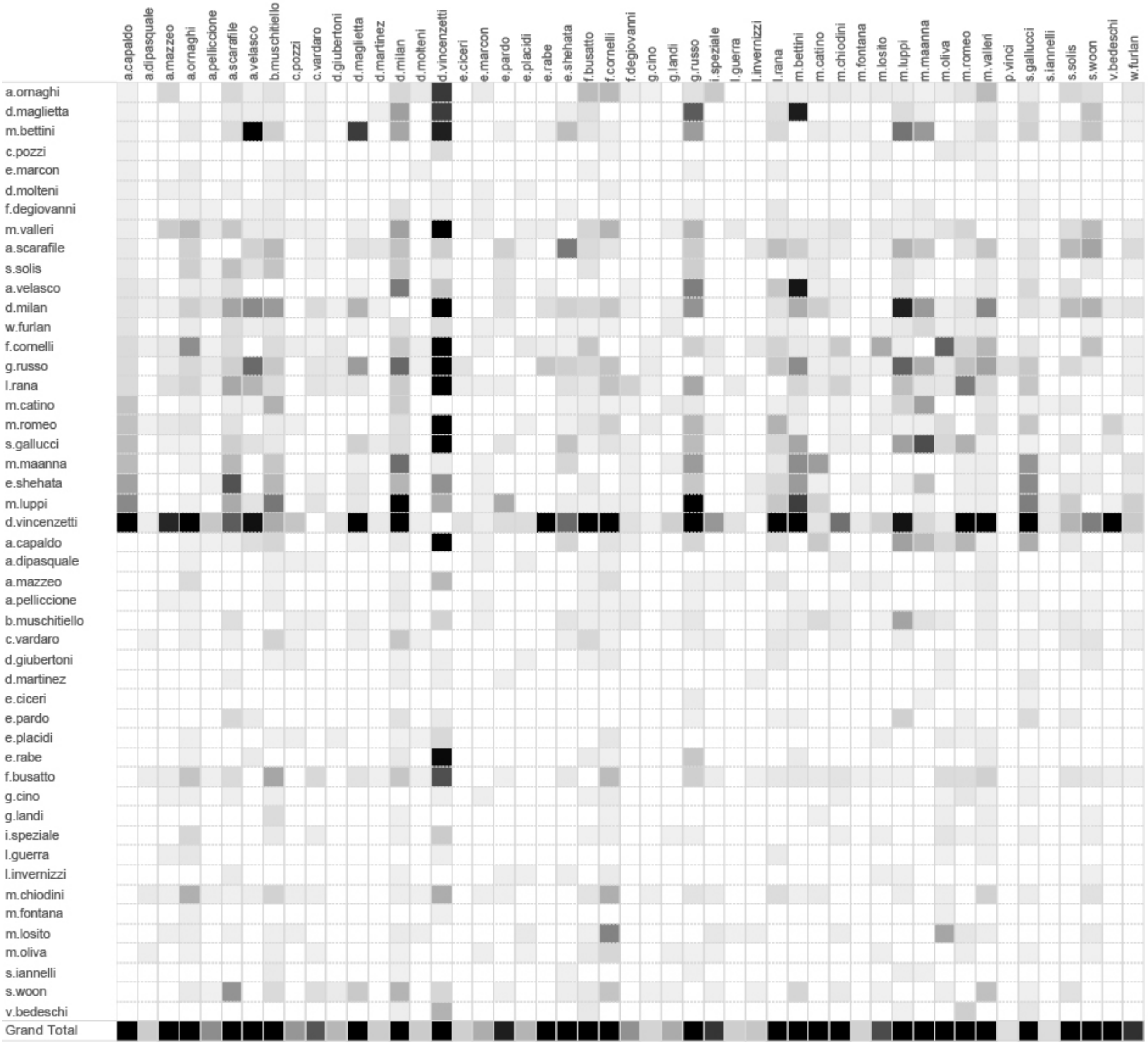
- What did you actually receive?
- What does the source say and how does it compare to what's in front of you?
- Does it look legit? Who can verify it?
- What's the general content?
- Can you spot some leads?
- Is there anything you need to act on immediately?

When you get to grips with what' s actually in your hands, you work out what potential stories you might be able to extract from the data, then begin organising in such a way that you can prove or disprove your bias or story ideas.

A great example of this method is the work carried out by the people from The Share Foundation on the Hacking Team files.

Hacking Team, HT, is an Italian company specializing in software tools which can be used by cyber snooping. Reporters without Borders labelled them as one of the corporate enemies of internet. Hacking Team insists it acts responsibly.

In July 2015, an unknown hacker publicly provided the links to 400GB of internal files, invoices, email inboxes and source code of their software.



This is the heat map produced by the Share Foundation showing interactions between Hacking Team staff

While everyone was focusing on HT's contracts, their tools and countries involved, people from Share Foundation decided to do a story about the inner workings of Hacking Team, based on metadata from thousands of emails provided in the leak.

They managed to reconstruct the exact behaviour of HT's CEO David Vincenzetti (Mister D.), what he was getting up, when he was traveling abroad, when were crises occurred at HT and how he was behaving.

Investigating metadata is one of the main focuses of Share Foundation, so they formulated the main approach in their investigation accordingly: "We wanted to question a somewhat heretical argument that bulk metadata contains sensitive information about the private life of internet users."

So, how did they do it?

First they obtained the data provided by the hacker. Then they went through it, figuring out all possible stories, deciding on the one that was in their focus - the metadata investigation.

They realised that they could extract headers from hundreds of thousands emails from HT mailbox accounts, and that's when the story started to form.

When you know what you want, it's just a matter of choosing the right tools to get it.

Email headers contain enormous amount of data. Most of it is useless for a focused investigation, and you reporters need to make some sense of it.

For this type of investigation, the required data were: the subject of each email, the date and time, the email addresses of people involved and their names.

Since mailbox files had the extension .pst, that meant that those were files of Microsoft Outlook, the ubiquitous email software from Microsoft. However, Outlook was not of use in this investigation as its only job with emails is to open them and show the content.

They needed a tool to search through those files and extract and sort only the information they needed.

The tool they used was Outlook Export, a freeware program from Code Two that has several freeware programs for handling Microsoft files.

Using that, reporters were able to choose the metadata that were important for their investigation and export them into a separate table.

Now, when they had structured data in a table, it was just a matter of translating that data into a story.

But how do you translate hundreds of thousands of data rows into a story? How do you spot the important trends?

There are several ways, but one of the best is to look at the data from a different perspective.

Using open source visualization tools, such as the powerful but easy to use Gephi, reporters were able to graphically represent the extracted data in several ways.

Each representation gave them a new insight - who was communicating with whom inside the organization and outside, when would they went on holiday, when would they wake up, who was working from which part of the world, who were the main partners, etc.

That's how they were able to identify the email address of the CEO as the central point in inner communications of the people in HT. This information defined the rest of the story.

By focusing on the CEO, reporters could answer the question they were asking: can you reconstruct a person' s behaviour based solely on metadata of their emails?
The answer was, yes.

By investigating metadata with only open source and freeware tools that are available to everyone, reporters were able to gain deep insight into the personal habits and activities of a CEO of a "hacking" company, and prove their premise that metadata contain sensitive information about the private lives of internet users that can be used to monitor the person' s behaviour "on a much deeper level than traditional surveillance", as the reporters said in the conclusion.

CHAPTER V

Interviewing data

FINDING STORIES IN STRUCTURED INFORMATION - MATHS AND SPREADSHEET TOOLS FOR JOURNALISTS

Welcome to the world of rows and columns! Fancy database managers sometimes use the term “tabular data” to refer to something less threatening-sounding: a table. Don’t let technical lingo intimidate you.

In fact, don’t let maths do that, either. We’ll cover Computer-Assisted Reporting (CAR) essential skills step by step, so that you’ll approach the spreadsheet like a magical place that can squeeze answers from a digital source.

Because that is what tables really are: sources you should be able to find and talk to. Unlike humans, they can’t leave the room. But you need to speak their language and understand results.

The very basics - know the environment

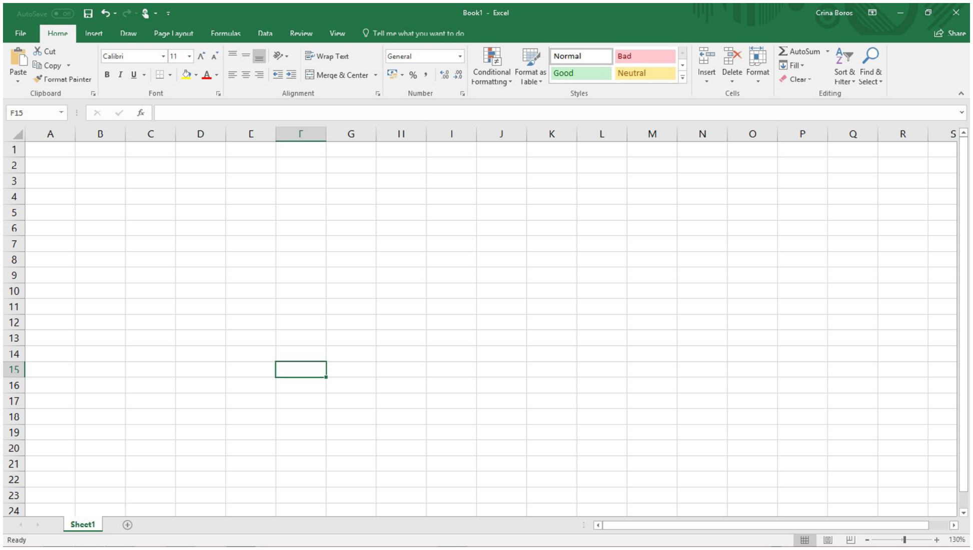


Table 1:
an Excel spreadsheet

From the variety of spreadsheets, we found Excel to be the most powerful, mainly due to its statistical power. Some (but not all) of the exercises below can be replicated using free and/open software. We recommend that you compare and contrast it with spreadsheets, simply because Microsoft is neither free, nor open.

We will make our first steps in MS Office - Excel. (Libre Office and Google spreadsheets are free alternatives.) Let's begin.

Open your spreadsheet. At the bottom of our table, you will see the name of the sheet: "Sheet1" by default. If you double-click it, you can rename it. We recommend that you always do.

There's a plus sign, a star or a spreadsheet icon next to the sheet's name. If you click on it, you will create another sheet - "Sheet 2". Try to keep the number of sheets small because spreadsheets start to slow down and to even malfunction as they become bigger.

An Excel spreadsheet will hold a particular number of rows and columns. Different Excel versions have different maximum capacities. Check your version.

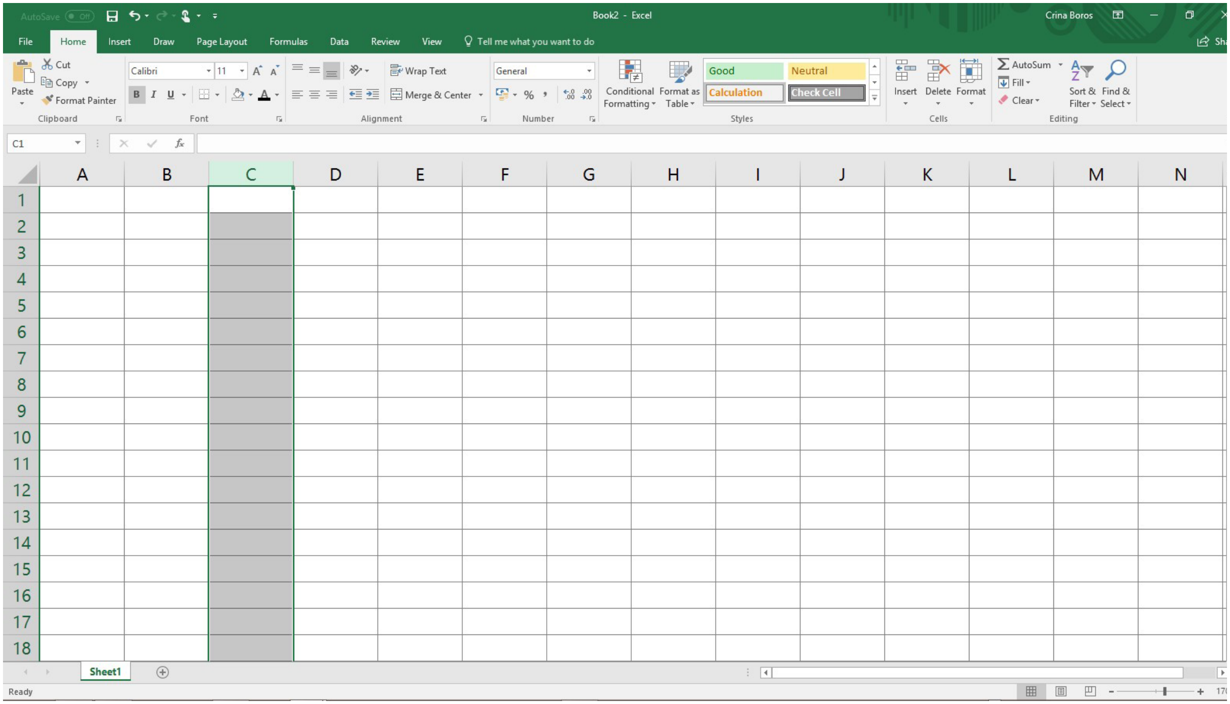


Table 2

Columns (selection in Table 2) are also known as variables.

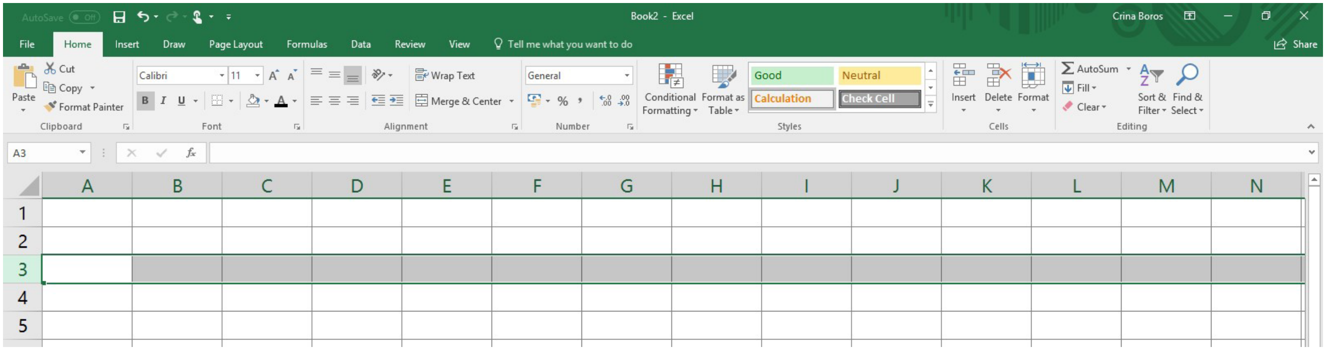


Table 3

Rows are also called records. Make a note of the number of rows you have in any dataset and keep an eye on it as you interrogate your data.

Always ask your data source to specify the particulars of your data: number of rows, number of columns and MB (GB for other formats) in order to anticipate whether your software can handle it.

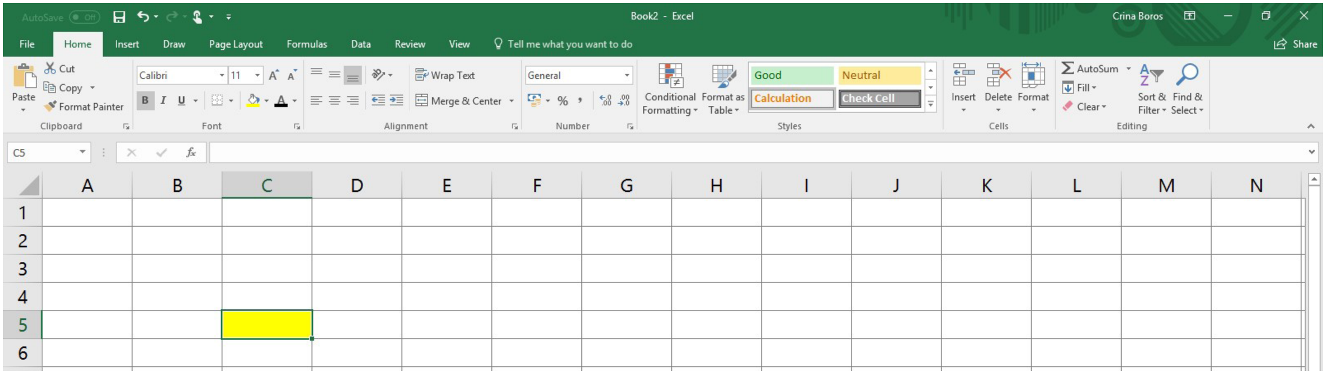


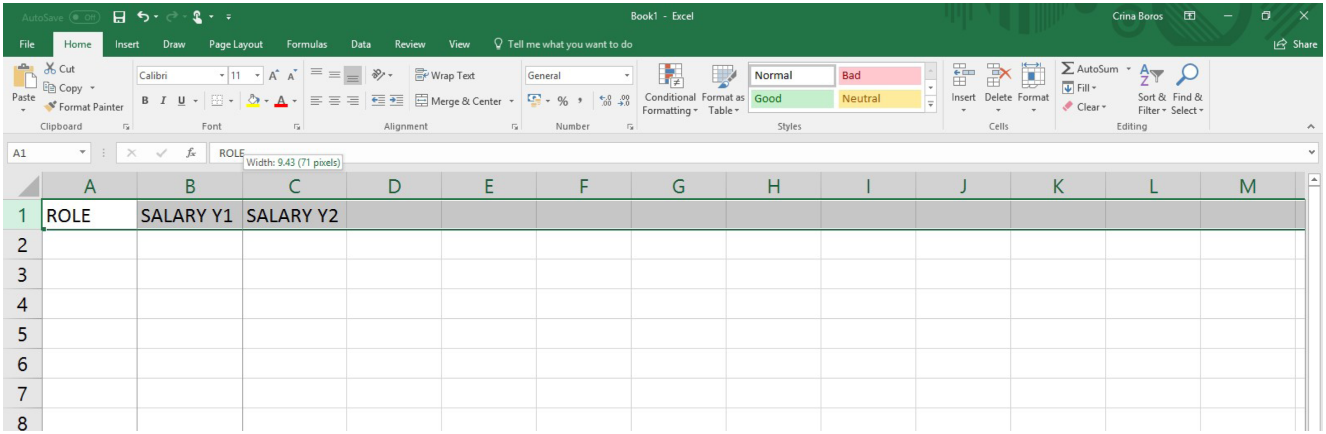
Table 4

Each sheet is made of cells with unique address: A1, B2, C3 etc.

On top of your sheet, you will see several tabs. These might differ, based on your software version. You should see File, Home, Insert, Data, Formulas, View, Review, Page Layout. There might be others. You will often switch between these, depending on what analysis you engage in.

Excel versions for Apple machines came to maturity in 2016. If you're working on a Mac and use older Excel versions, think about upgrading. This book might frustrate those on older Office versions for Mac. PCs were already in a strong position with Excel 2007, but an important jump in analysis power was achieved in 2010. Make sure you're all set with a decent version of the programme before going further into this chapter.

Let's enter some data in row 1, otherwise referred to as the TOP ROW.



ROLE

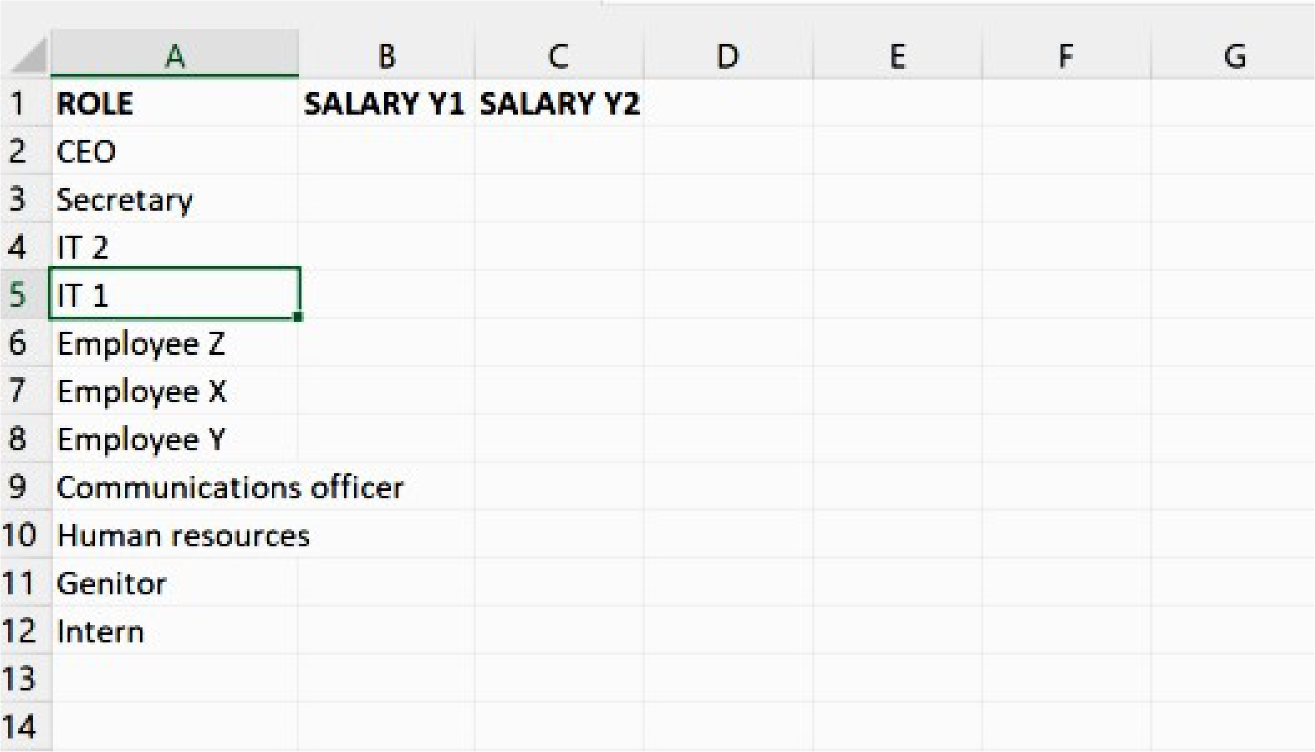
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ROLE	SALARY Y1	SALARY Y2										
2													
3													
4													
5													
6													
7													
8													

Table 5

We will enter and analyse income data. Write up three column headers (the names of columns) as in Table 5.

Tables are often made up of tens of thousands of records. It is important to never lose sight of what information you're actually looking at. One way to do that is to freeze the top row in place - this way, you can scroll all the way to the bottom of your spreadsheet but still see your column headers.

Place your cursor on the number of the row (1) > click on View > Freeze panes > Freeze top row. Similarly, you can also freeze the first column.



	A	B	C	D	E	F	G
1	**ROLE**	**SALARY Y1**	**SALARY Y2**				
2	CEO						
3	Secretary						
4	IT 2						
5	IT 1						
6	Employee Z						
7	Employee X						
8	Employee Y						
9	Communications officer						
10	Human resources						
11	Genitor						
12	Intern						
13							
14							

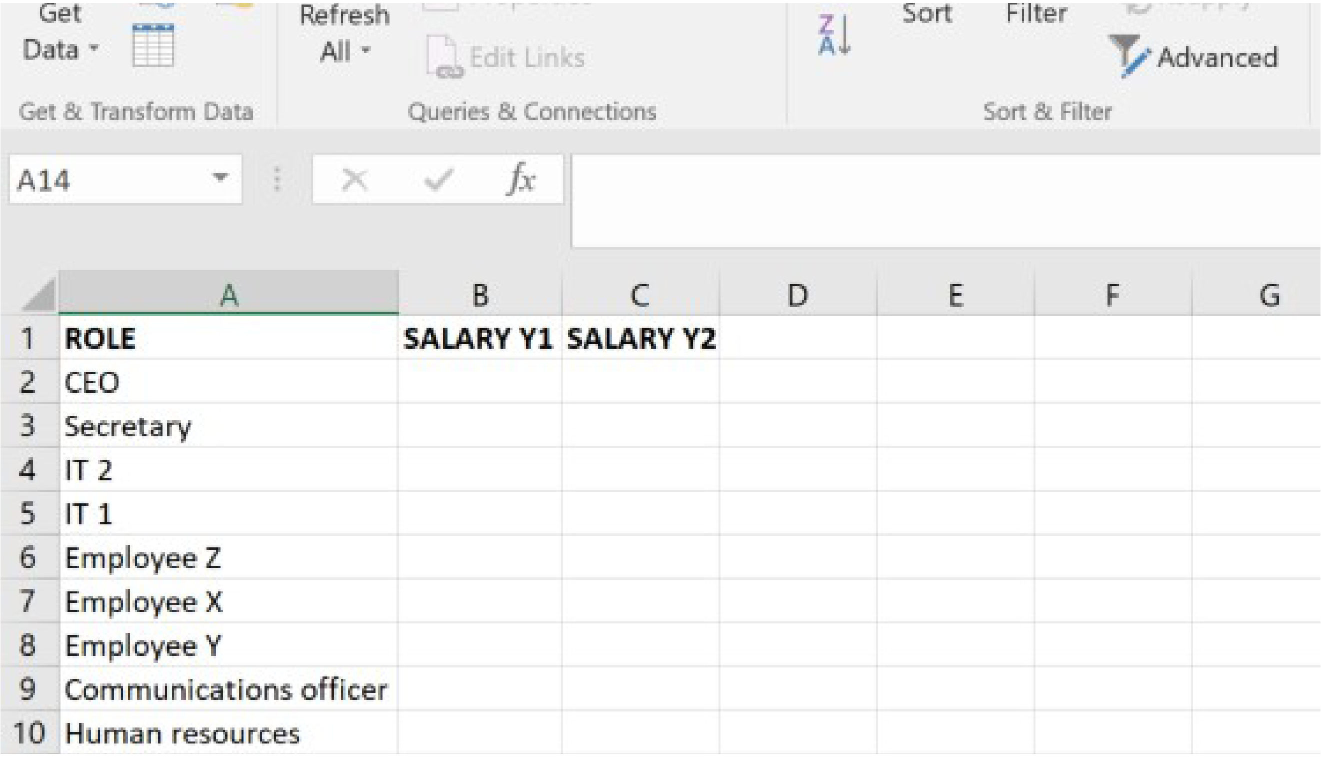
Table 6

Type in job descriptions, like in Table 6.

Some of the text will bleed into the next column. To fix this, click on the border line between the letters of two columns (say, A and B, above row 1); then click and drag to set your preferred column width. It will look neat, like in Table 7.

Or, instead of clicking and dragging, you can simply double-click the border-line between two columns, right at the top.

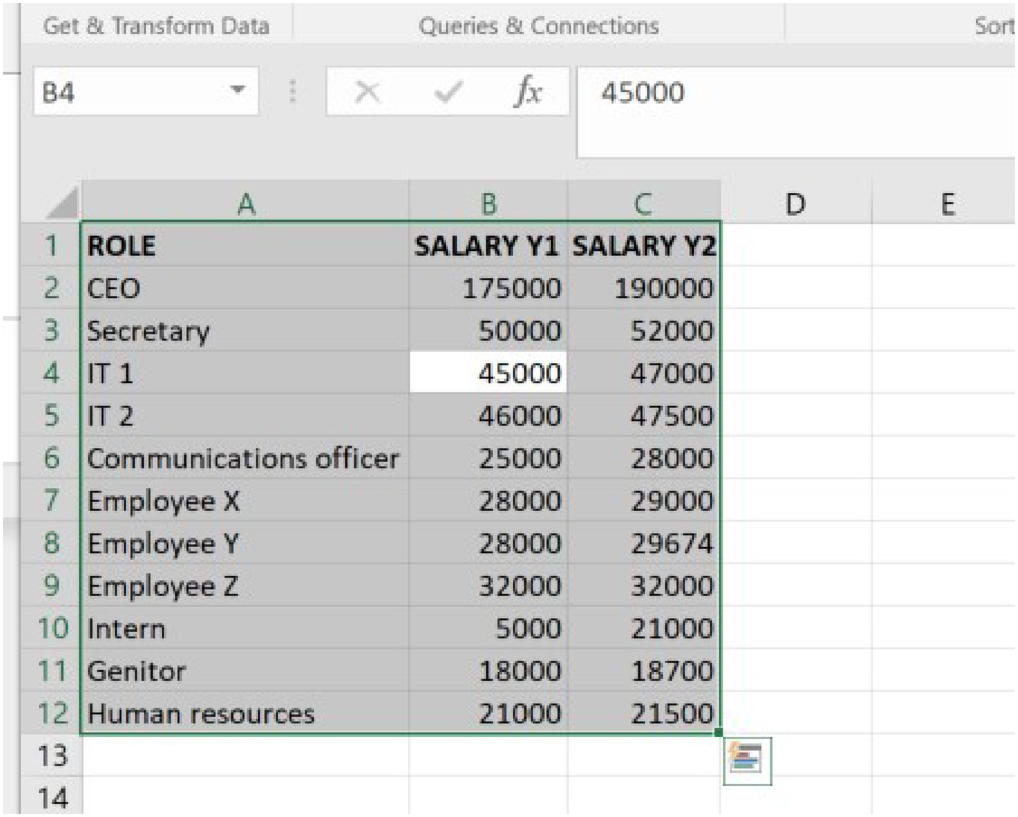
Be patient with double-clicking if you feel clumsy the first time, and watch the cursor changing shapes and colours (black or white) as you will perform various double-clicking tasks in a spreadsheet.



	A	B	C	D	E	F	G
1	**ROLE**	**SALARY Y1**	**SALARY Y2**				
2	CEO						
3	Secretary						
4	IT 2						
5	IT 1						
6	Employee Z						
7	Employee X						
8	Employee Y						
9	Communications officer						
10	Human resources						

Table 7
You can download a version of this table by clicking here.

Data entry and formatting

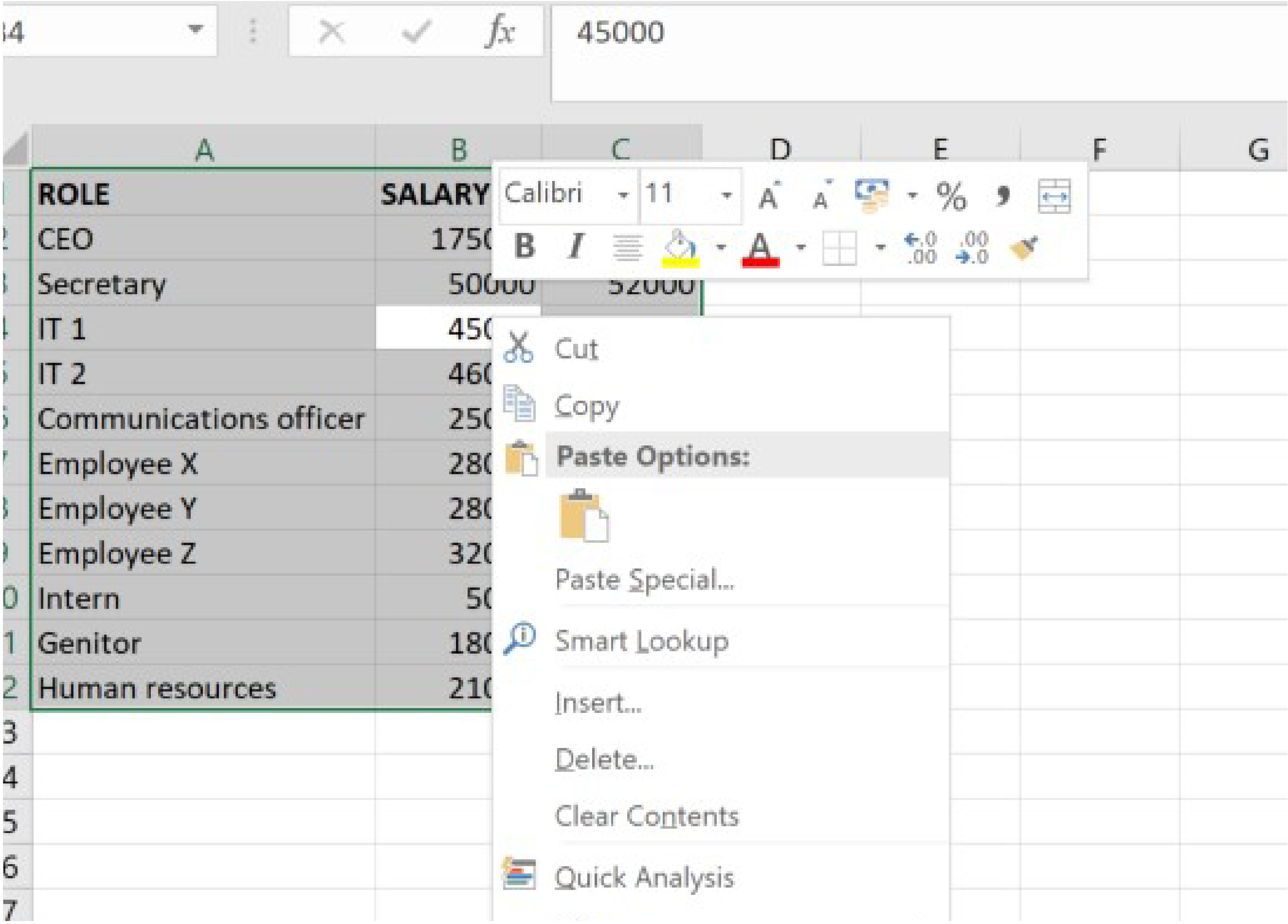


B4 | 45000

ROLE	SALARY Y1	SALARY Y2
CEO	175000	190000
Secretary	50000	52000
IT 1	45000	47000
IT 2	46000	47500
Communications officer	25000	28000
Employee X	28000	29000
Employee Y	28000	29674
Employee Z	32000	32000
Intern	5000	21000
Genitor	18000	18700
Human resources	21000	21500

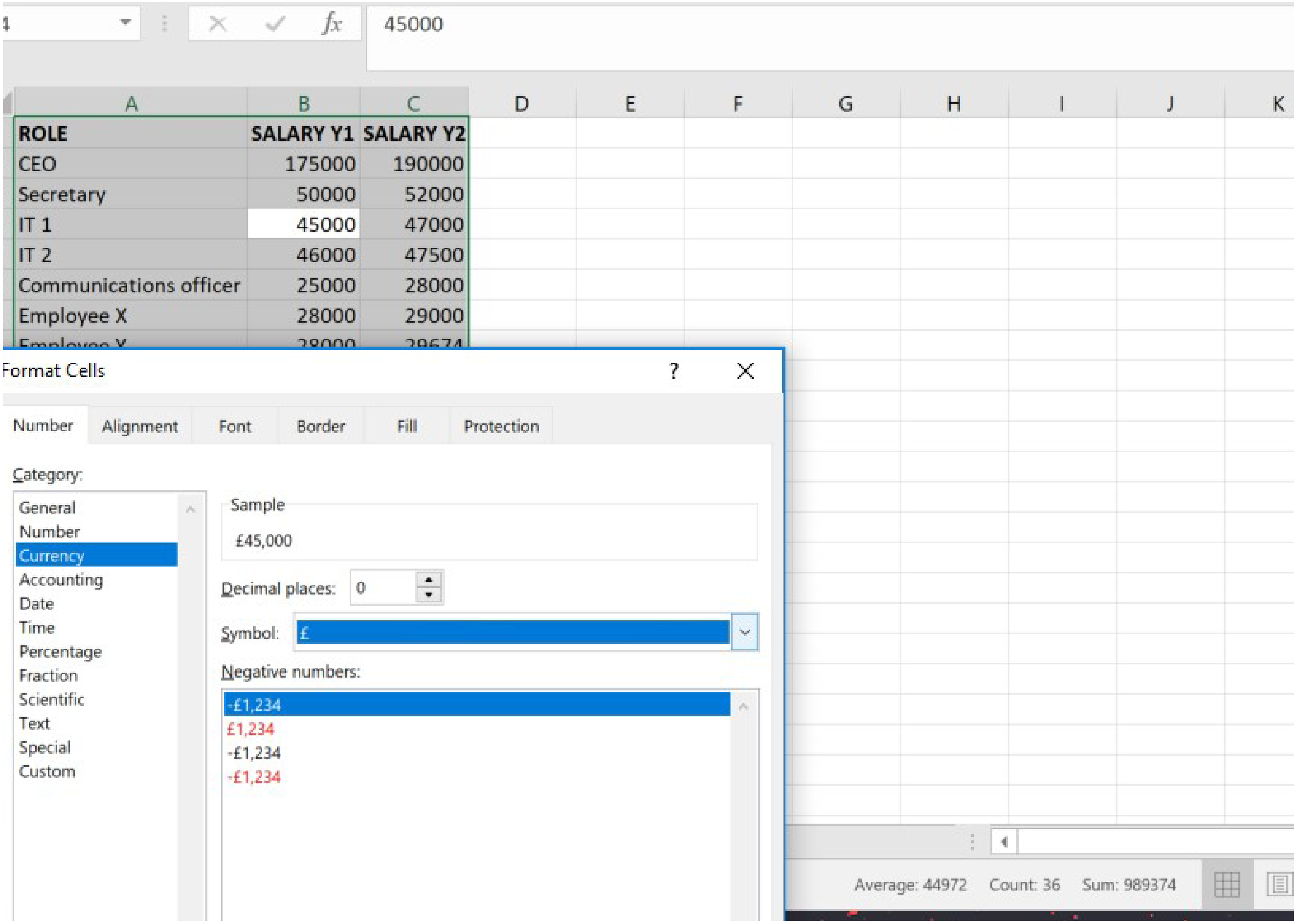
Table 8

When you enter numbers, do not use the keyboard to format decimals and currencies – this may lead to errors when running functions. Instead,
> click anywhere in the data
> CTRL+A to select it (MAC users, you may have to press Command+A)
> follow with Right click
> format cells (like in Table 9).



ROLE	SALARY	
CEO	175	
Secretary	50000	52000
IT 1	450	
IT 2	460	
Communications officer	250	
Employee X	280	
Employee Y	280	
Employee Z	320	
Intern	50	
Genitor	180	
Human resources	210	

Table 9



45000

ROLE	SALARY Y1	SALARY Y2
CEO	175000	190000
Secretary	50000	52000
IT 1	45000	47000
IT 2	46000	47500
Communications officer	25000	28000
Employee X	28000	29000
Employee Y	28000	29674

Format Cells

Category: Currency

Sample

£45,000

Decimal places: 0

Symbol: £

Negative numbers:

-£1,234
£1,234
-£1,234
-£1,234

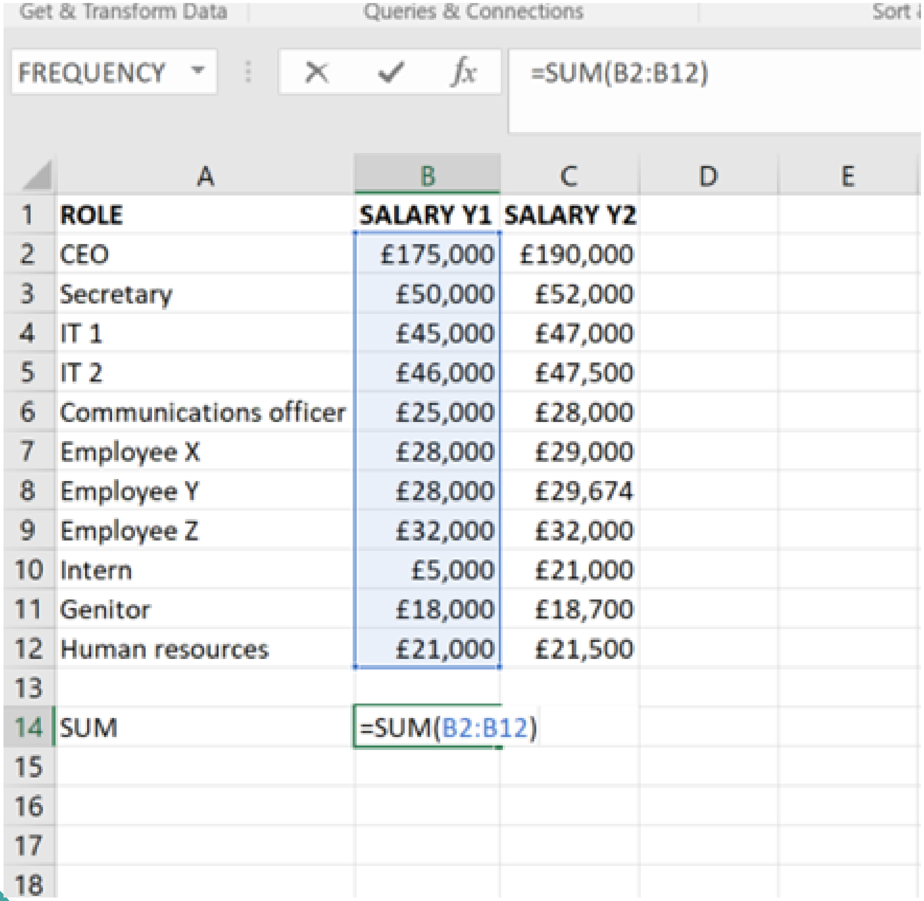
Average: 44972 Count: 36 Sum: 989374

Table 10

A window with several tabs opens up. The first tab is “Number”. The third option down is “Currency” – select it. Here you can set the currency symbol from the drop-down menu, as well as assign a number of decimals (Table 10). Under the first option – Numbers – you may tick the thousand separator, so that the number is reader-friendly.

Essential formulas

Knowing how to compute your numbers is a must. Calculate your sums, averages, frequencies, differences between largest, average and and smallest, or percentages and percentage changes with the following formulas. These numbers are called “descriptive statistics”. They describe a situation, a person, a comparison.



=SUM(B2:B12)

	A	B	C	D	E
1	ROLE	SALARY Y1	SALARY Y2		
2	CEO	£175,000	£190,000		
3	Secretary	£50,000	£52,000		
4	IT 1	£45,000	£47,000		
5	IT 2	£46,000	£47,500		
6	Communications officer	£25,000	£28,000		
7	Employee X	£28,000	£29,000		
8	Employee Y	£28,000	£29,674		
9	Employee Z	£32,000	£32,000		
10	Intern	£5,000	£21,000		
11	Genitor	£18,000	£18,700		
12	Human resources	£21,000	£21,500		
13					
14	SUM	=SUM(B2:B12)			
15					
16					
17					
18					

Table 11

>SUM

Before you run any maths on your table, make sure you leave a blank row between your table and future computations (Table 11). You want your original data to keep their integrity and not get mixed up with your analysis.

We also recommend that you always work on a copy of your original table and never on the raw information.

To calculate the total yearly income budget, you first need to add up all the salaries from year one. To add up numbers, you will use a function called SUM.

In EXCEL, functions start with = (equal sign).

Most functions will require parenthesis / round brackets.

You can add up by listing all the numbers.

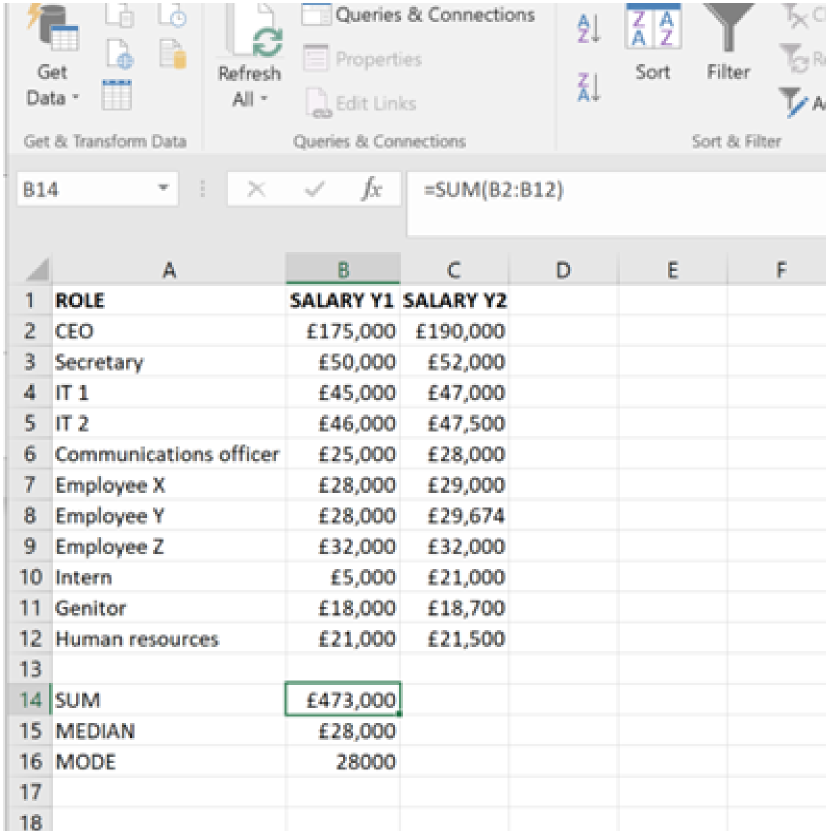
=SUM(No1, No2, ...No12)

This method is not helpful if you have many numbers, but it’s useful when we have to add a few scattered ones.

In our case, we want to add up every salary down one column. It’s quicker to **define a range** by telling Excel with which number to start and which to finish with. We do this by using **:** (colon). The formula becomes:

=SUM(B2:B12)

Write the formula in B14 and hit “Enter” (Table 11). Summing up all incomes received in Year 1 tells us a company’s or institution’s yearly income budget.



=SUM(B2:B12)

	A	B	C	D	E	F
1	ROLE	SALARY Y1	SALARY Y2			
2	CEO	£175,000	£190,000			
3	Secretary	£50,000	£52,000			
4	IT 1	£45,000	£47,000			
5	IT 2	£46,000	£47,500			
6	Communications officer	£25,000	£28,000			
7	Employee X	£28,000	£29,000			
8	Employee Y	£28,000	£29,674			
9	Employee Z	£32,000	£32,000			
10	Intern	£5,000	£21,000			
11	Genitor	£18,000	£18,700			
12	Human resources	£21,000	£21,500			
13						
14	SUM	£473,000				
15	MEDIAN	£28,000				
16	MODE	28000				
17						
18						

Table 12

>WHAT'S AVERAGE?

One of the most commonly reported numbers is the average. More precisely, the mean. There are several types of averages - numbers relied on to indicate central tendencies, or to describe what's 'typical' in a context. The mean is the one everyone knows and feels comfortable to use.

The mean is calculated by summing all numbers from a group, and then dividing the result by the count of numbers of the same group.

Like a sum, a mean describes a situation. It also answers a question. One way to decide what formula to use is to articulate your question you first.

In this context, our question is: given the total income budget of an organisation, if everyone was paid the same, what would that common salary be?

=AVERAGE(B2:B12)

When it comes to money, however, the mean is sometimes misleading. If one of the numbers is very high or very low compared to the others (outliers), the mean will provide a misleading impression of the typical salary.

Outliers are values which stand out from most values on the list.

You have two options here. Some choose to take out the highest and/or lowest values, or the top and bottom 1% and to calculate the mean again. The result is called a TRIMMED MEAN.

But a better option would be to calculate **another average value**, called MEDIAN:

=MEDIAN(B2:B12)

A median value splits a list in two equal halves: half of the numbers are above the median value and half are below.

To see the split, you should sort your records. (We cover sorting in the next chapter).

Calculating the median when dealing with money values is preferable, simply because it eliminates the problems that outliers can create by distorting (skewing) an average. Reporting skewed results as genuine deceives the audience.

A median gets rid of the distortion created by outliers and is, therefore, a more accurate central value. It helps create a better understanding of the "typical" house price or salary, as Sarah Cohen explains in "Numbers in the Newsroom".

However, a median can only be computed if a list of values is complete: all the salaries from a ministry; all the property prices from an area; the pay for each football player in a league; the expenses of all members of a parliament, etc.

So which average do you report and when?

If the values of mean and the median are close, report the mean because this is a number everyone understands and no further explanation is needed of what the average is. If there is a significant difference between the two, report the median, but only if you have a full list of the values you're analysing. If you don't, which is often the case, compute a **trimmed mean**.

If you have to report a trimmed mean, consider reporting the average value before and after trimming, to give a sense of the distortion that the outliers cause. This can make your reporting more interesting and nuanced.

Sometimes, neither the average nor the median can describe the 'typical' situation or value. For example, a quick look at Romanian wages would reveal that at least one in three of the working population earns the national minimum wage or slightly below.

Reporting for Investigate Europe, one of the authors started with this number, called MODE, to reveal what has led to a dramatic impact on the country's remuneration policy and mass economic migration (see How Romania Sold Out Its Workers for IMF and EU Cash - The Black Sea; and How Romania became an EU workers' Rights Guinea Pig - EUobserver).

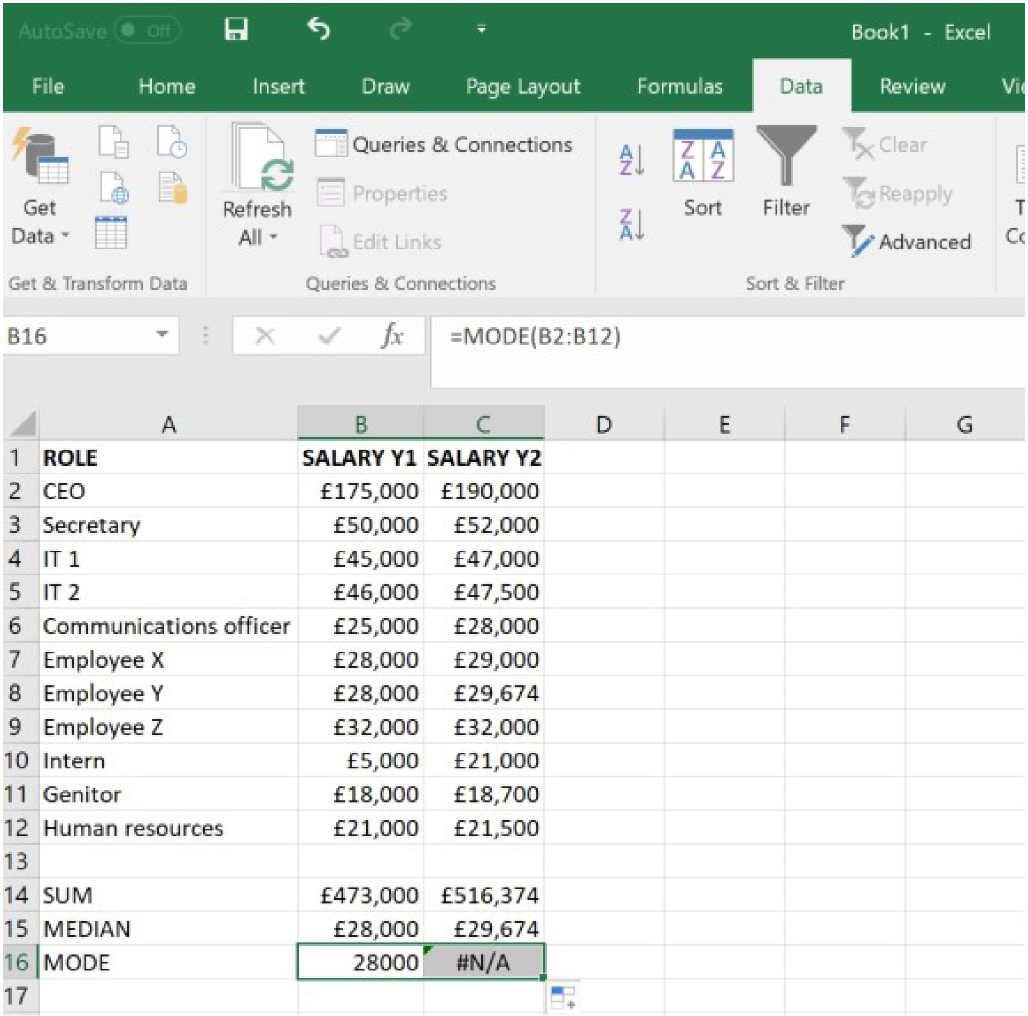
Modes calculate common values. The analysts behind polls or elections are ideal for modal values, which represent the most common result. The formula is:

=MODE(B2:B12)

In our case, the modal value for salaries in year 1 is identical to the median.

To replicate the maths across the second column of numbers, click the lower right corner of the cell containing the formula, hold and drag across to the second column.

The #N/A result – not applicable – in the case of Mode for the salaries in Year 2 column means that there are no two identical salaries.



B16 =MODE(B2:B12)

	A	B	C
1	ROLE	SALARY Y1	SALARY Y2
2	CEO	£175,000	£190,000
3	Secretary	£50,000	£52,000
4	IT 1	£45,000	£47,000
5	IT 2	£46,000	£47,500
6	Communications officer	£25,000	£28,000
7	Employee X	£28,000	£29,000
8	Employee Y	£28,000	£29,674
9	Employee Z	£32,000	£32,000
10	Intern	£5,000	£21,000
11	Genitor	£18,000	£18,700
12	Human resources	£21,000	£21,500
13			
14	SUM	£473,000	£516,374
15	MEDIAN	£28,000	£29,674
16	MODE	28000	#N/A
17			

Table 13

Sum, mean, median, mode are called “descriptive statistics”: they are values that we compute in order to describe a situation, a scene, a population, a person.

To determine which number to report (and when), calculate all of the above whenever possible. Report all of them, if your space allows it. But make sure to include all the values in a methodology document if the space allocated for the story doesn' t.

And when you analyse news yourself, note that one way for a report to misinform is to hold back some of these values.

Tip

If your record contains thousands of salaries, it may not be practical to search for identical ones. Instead, you may wish to create salary ranges that suit a context.

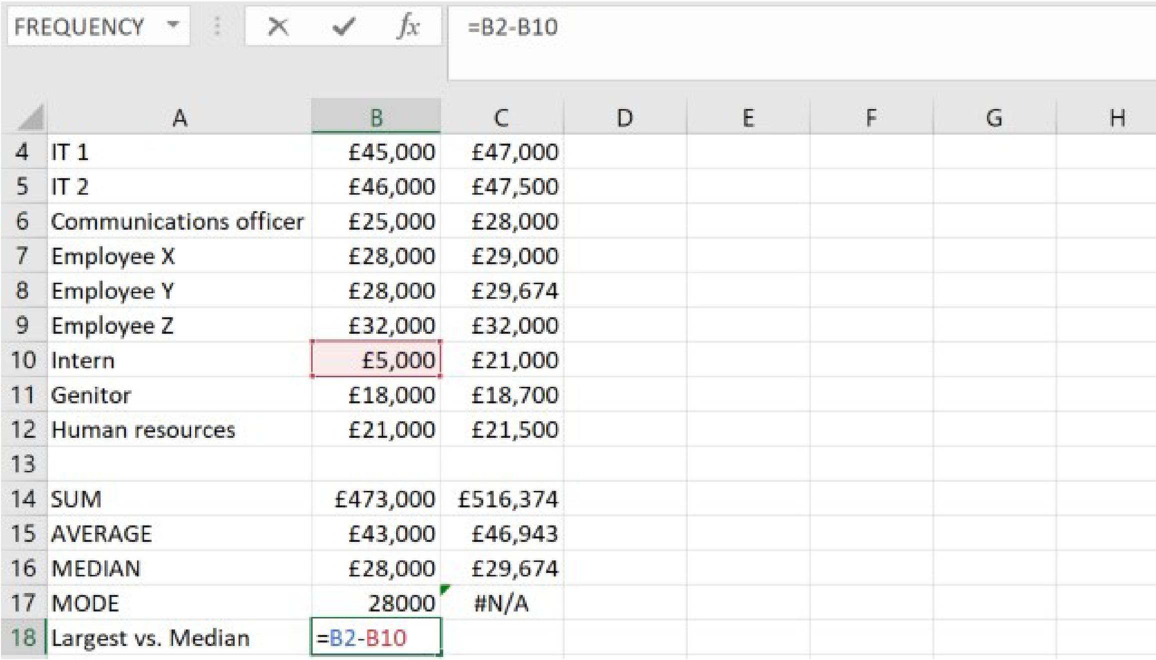
USEFUL NUMBERS

Which difference?

A lot of our work is comparing and contrasting, meaning we look at differences and similarities, disparities and trends. There are a few numbers to help paint a picture.

>SUBTRACTION AND PERCENTAGES

A useful function for finding discrepancies is **subtraction**. Use it to find the difference between the highest number and the median (=B2-B10) (Table 14) or the difference between the highest and lowest, between the median and mean, between the median and the lowest number, etc. This will give you an idea of what's going on in a group you're studying.



FREQUENCY × ✓ fx =B2-B10

	A	B	C	D	E	F	G	H
4	IT 1	£45,000	£47,000					
5	IT 2	£46,000	£47,500					
6	Communications officer	£25,000	£28,000					
7	Employee X	£28,000	£29,000					
8	Employee Y	£28,000	£29,674					
9	Employee Z	£32,000	£32,000					
10	Intern	£5,000	£21,000					
11	Genitor	£18,000	£18,700					
12	Human resources	£21,000	£21,500					
13								
14	SUM	£473,000	£516,374					
15	AVERAGE	£43,000	£46,943					
16	MEDIAN	£28,000	£29,674					
17	MODE	28000	#N/A					
18	Largest vs. Median	=B2-B10						

Table 14

The beauty of numbers is that you can compare them. What is the purchasing-power of one country compared to another? What is the difference in GDP per capita across Europe? How have salaries evolved year-on-year for the same role?

We must, however, ensure that we are comparing apples to apples and pears to pears.

For example, we could compare the minimum wages across various countries, but that would not be meaningful unless we also take into account the types and sizes of the economies we're analysing. Comparing GDP per capita would be fairer, although you can expect that number to be skewed as well. For example, in the UK, statistics quote the national average wage and the London average wage separately because including London, where wages are much higher, would distort the picture for the rest of the country.. Check out UK's income and tax borough by borough.

Let's return to our table. Say you want to compare the change in salaries year on year. Subtract the two and look at the raw difference. This is a good start. But it's not enough for an apples to apples comparison.

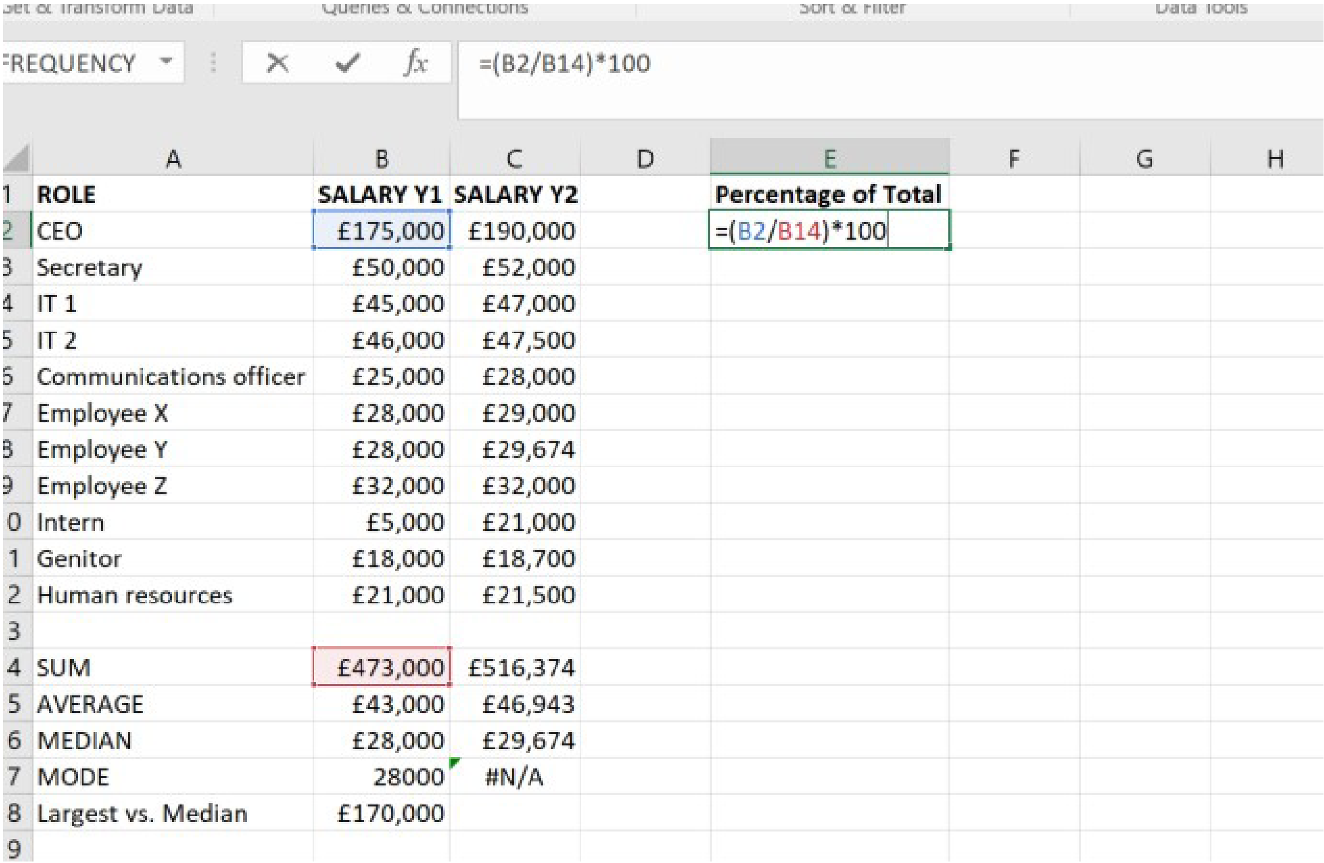
Percentage changes are often more meaningful.

Let's look down one column and calculate ***the percentage each salary represents from the total*** yearly income budget (Table 15) for YEAR 1.

The formula is:

=(THE SALARY/THE TOTAL)*100

OR =(B2/B14)*100

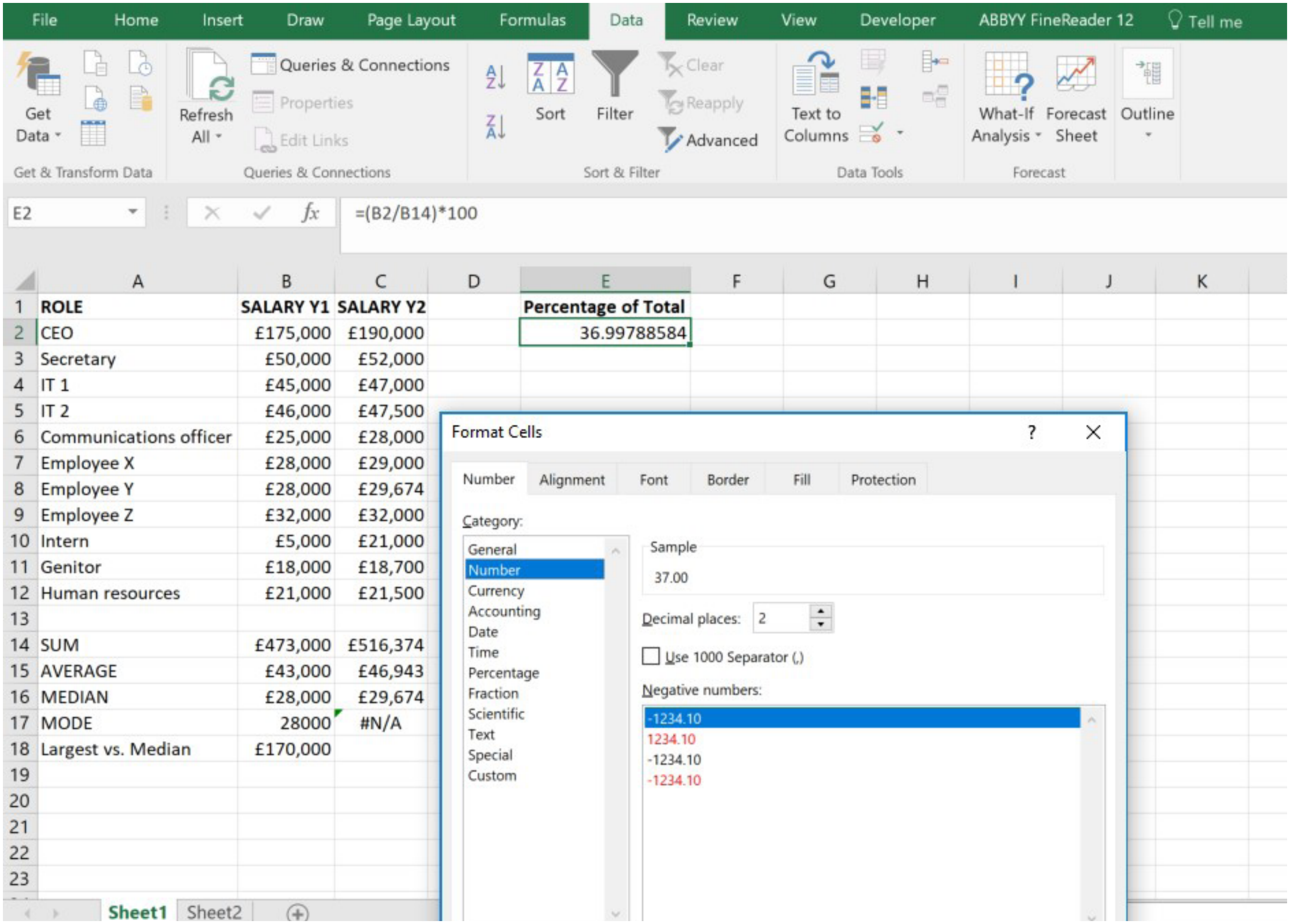


=(B2/B14)*100

	A	B	C	D	E	F	G	H
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total			
2	CEO	£175,000	£190,000		=(B2/B14)*100			
3	Secretary	£50,000	£52,000					
4	IT 1	£45,000	£47,000					
5	IT 2	£46,000	£47,500					
6	Communications officer	£25,000	£28,000					
7	Employee X	£28,000	£29,000					
8	Employee Y	£28,000	£29,674					
9	Employee Z	£32,000	£32,000					
0	Intern	£5,000	£21,000					
1	Genitor	£18,000	£18,700					
2	Human resources	£21,000	£21,500					
3								
4	SUM	£473,000	£516,374					
5	AVERAGE	£43,000	£46,943					
6	MEDIAN	£28,000	£29,674					
7	MODE	28000	#N/A					
8	Largest vs. Median	£170,000						
9								

Table 15

The result is a number with many decimals. To reduce the decimals, or to round up the number:
>right-click in the cell address with your value
> Format cells
> select the NUMBER tab from the window
> Number. In this window, adjust the decimals and tick the 1000 separator (Table 16).



E2 =(B2/B14)*100

	A	B	C	D	E
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total
2	CEO	£175,000	£190,000		36.99788584
3	Secretary	£50,000	£52,000		
4	IT 1	£45,000	£47,000		
5	IT 2	£46,000	£47,500		
6	Communications officer	£25,000	£28,000		
7	Employee X	£28,000	£29,000		
8	Employee Y	£28,000	£29,674		
9	Employee Z	£32,000	£32,000		
10	Intern	£5,000	£21,000		
11	Genitor	£18,000	£18,700		
12	Human resources	£21,000	£21,500		
13					
14	SUM	£473,000	£516,374		
15	AVERAGE	£43,000	£46,943		
16	MEDIAN	£28,000	£29,674		
17	MODE	28000	#N/A		
18	Largest vs. Median	£170,000			

Sample
37.00

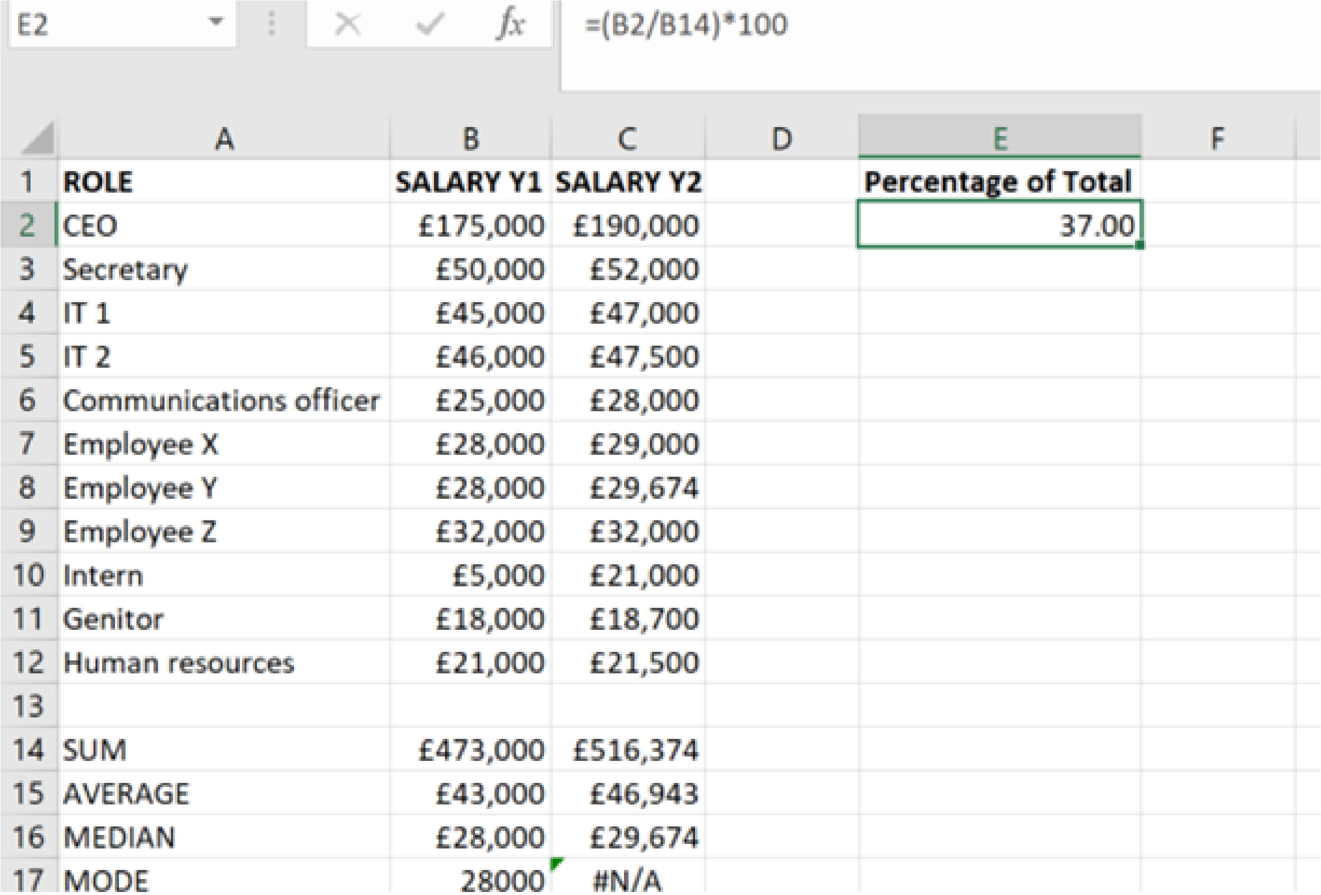
Decimal places: 2

□ Use 1000 Separator (,)

Negative numbers:
-1234.10
1234.10
-1234.10
-1234.10

Table 16

You can copy the formula by dragging the lower-right corner down the column (Table 17). Or double click the lower-right corner of the cell.



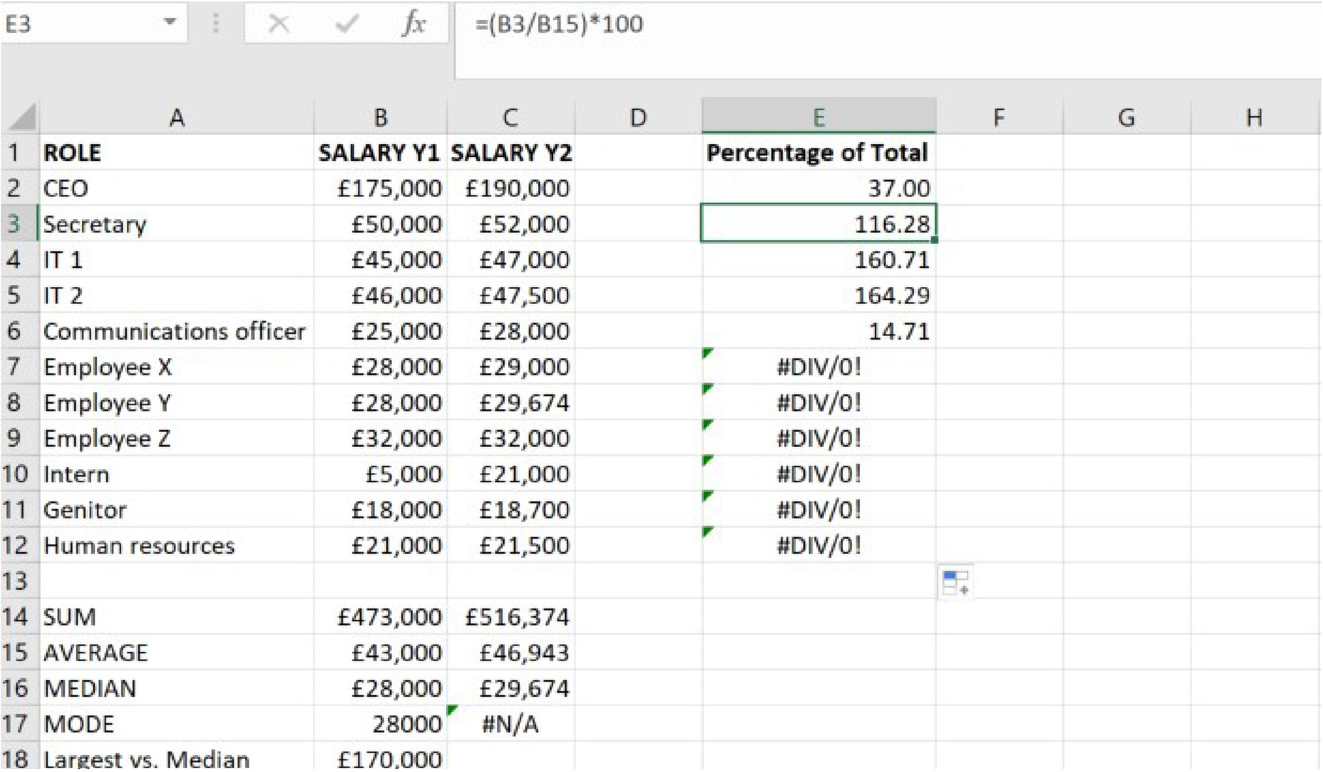
E2 =(B2/B14)*100

	A	B	C	D	E	F
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total	
2	CEO	£175,000	£190,000		37.00	
3	Secretary	£50,000	£52,000			
4	IT 1	£45,000	£47,000			
5	IT 2	£46,000	£47,500			
6	Communications officer	£25,000	£28,000			
7	Employee X	£28,000	£29,000			
8	Employee Y	£28,000	£29,674			
9	Employee Z	£32,000	£32,000			
10	Intern	£5,000	£21,000			
11	Genitor	£18,000	£18,700			
12	Human resources	£21,000	£21,500			
13						
14	SUM	£473,000	£516,374			
15	AVERAGE	£43,000	£46,943			
16	MEDIAN	£28,000	£29,674			
17	MODE	28000	#N/A			

Table 17

What we see is numbers that don't look quite right. Unless we tweak our formula, we will run into an error (Table 18). The error **#DIV/0!** is Excel telling you: "You're dividing a number by 0, silly!"

By the way, don't get demoralised when you see an error message. There are countless forums, YouTube videos and help out there helping you to fix a problem. And, once you learn how, you'll never forget. Onwards!



E3 =(B3/B15)*100

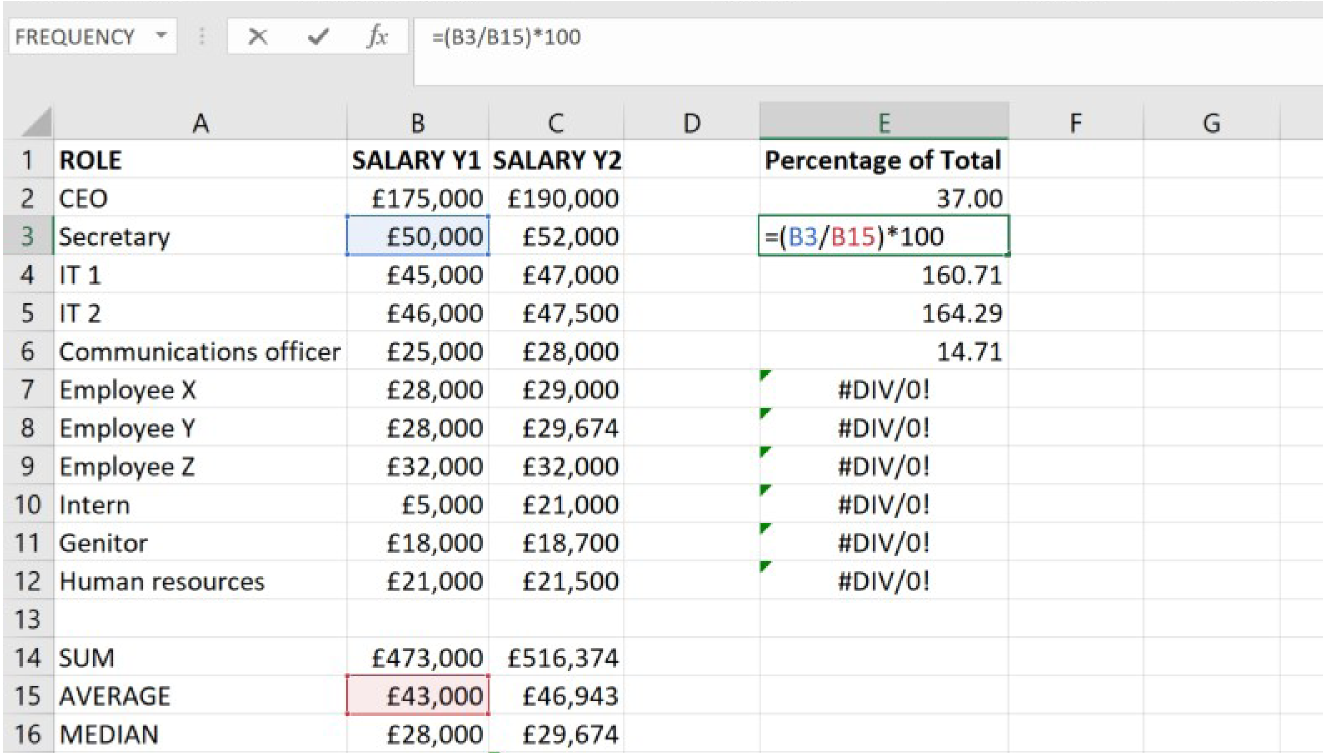
	A	B	C	D	E	F	G	H
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total			
2	CEO	£175,000	£190,000		37.00			
3	Secretary	£50,000	£52,000		116.28			
4	IT 1	£45,000	£47,000		160.71			
5	IT 2	£46,000	£47,500		164.29			
6	Communications officer	£25,000	£28,000		14.71			
7	Employee X	£28,000	£29,000		#DIV/0!			
8	Employee Y	£28,000	£29,674		#DIV/0!			
9	Employee Z	£32,000	£32,000		#DIV/0!			
10	Intern	£5,000	£21,000		#DIV/0!			
11	Genitor	£18,000	£18,700		#DIV/0!			
12	Human resources	£21,000	£21,500		#DIV/0!			
13								
14	SUM	£473,000	£516,374					
15	AVERAGE	£43,000	£46,943					
16	MEDIAN	£28,000	£29,674					
17	MODE	28000	#N/A					
18	Largest vs. Median	£170,000						

Table 18

Click on the second percentage in E3. This number also does not look right.

Above the columns, in the **formula bar (fx)**, you see the fully unpicked formula behind a value or an error. You can always check your formulas against the Fx bar by clicking into any cell. You can also easily manipulate formulas by typing directly in the formula bar. Use it to your advantage.

It looks like we have not divided the salary by the total, but by the next number down, which in our case is the median (Table 19). This is a good example to remember: when in doubt, check the maths behind a number in the Formula Bar.



=(B3/B15)*100

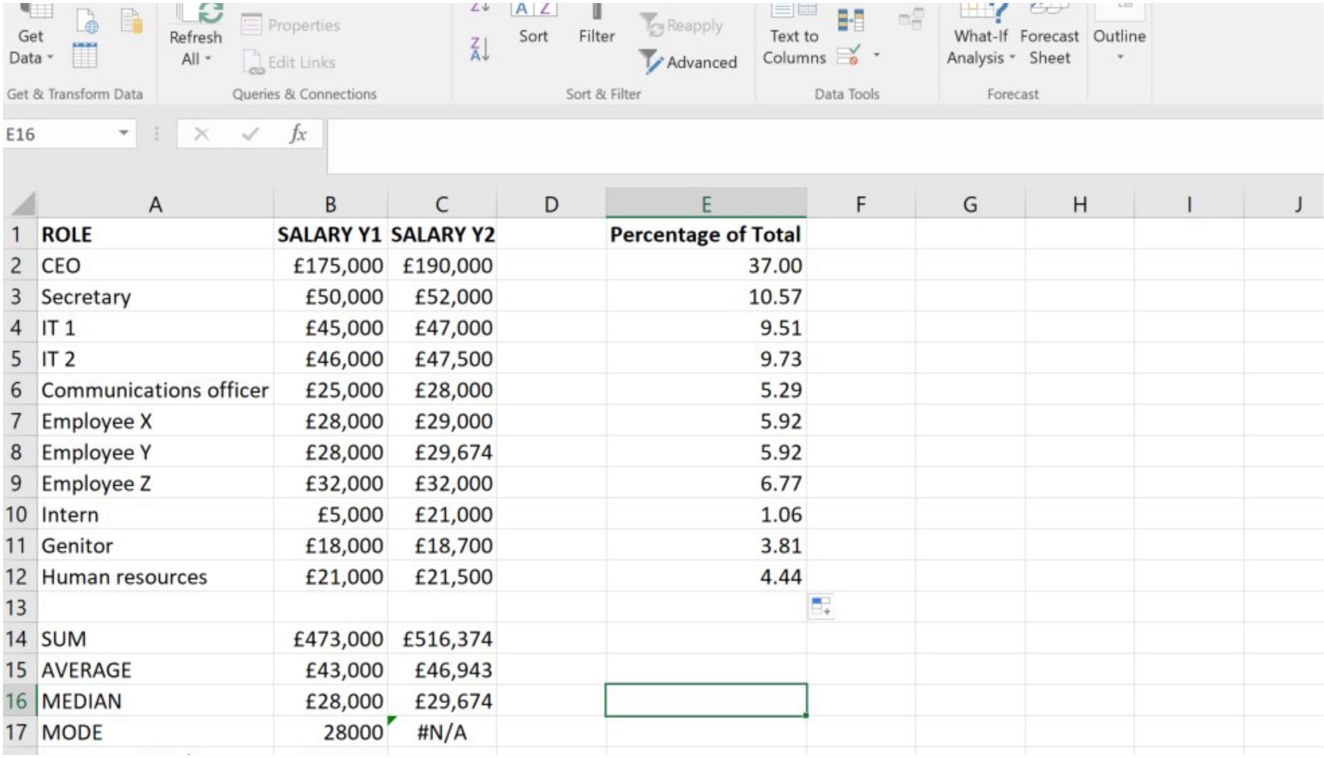
	A	B	C	D	E	F	G
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total		
2	CEO	£175,000	£190,000		37.00		
3	Secretary	£50,000	£52,000		=(B3/B15)*100		
4	IT 1	£45,000	£47,000		160.71		
5	IT 2	£46,000	£47,500		164.29		
6	Communications officer	£25,000	£28,000		14.71		
7	Employee X	£28,000	£29,000		#DIV/0!		
8	Employee Y	£28,000	£29,674		#DIV/0!		
9	Employee Z	£32,000	£32,000		#DIV/0!		
10	Intern	£5,000	£21,000		#DIV/0!		
11	Genitor	£18,000	£18,700		#DIV/0!		
12	Human resources	£21,000	£21,500		#DIV/0!		
13							
14	SUM	£473,000	£516,374				
15	AVERAGE	£43,000	£46,943				
16	MEDIAN	£28,000	£29,674				

Table 19

To fix the problem and get meaningful values, what we want to do is **anchor the total in place** and have only the salaries shifting from one row to another.

The anchor is the dollar sign **$**.
Click back in E2, where our first percentage is. Your current formula is
=(B2/B14)*100
The anchored formula will be **=(B2/$B$14)*100**

Run your double-click down the column again and your percentage of total values will be correct (Table 20). It looks like the CEO has got more than a third of the total income budget.



	A	B	C	D	E	F	G	H	I	J
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total					
2	CEO	£175,000	£190,000		37.00					
3	Secretary	£50,000	£52,000		10.57					
4	IT 1	£45,000	£47,000		9.51					
5	IT 2	£46,000	£47,500		9.73					
6	Communications officer	£25,000	£28,000		5.29					
7	Employee X	£28,000	£29,000		5.92					
8	Employee Y	£28,000	£29,674		5.92					
9	Employee Z	£32,000	£32,000		6.77					
10	Intern	£5,000	£21,000		1.06					
11	Genitor	£18,000	£18,700		3.81					
12	Human resources	£21,000	£21,500		4.44					
13										
14	SUM	£473,000	£516,374							
15	AVERAGE	£43,000	£46,943							
16	MEDIAN	£28,000	£29,674							
17	MODE	28000	#N/A							

Table 20

Percentage change over years

So far, we have looked down one column. Now look across columns and compare the changes in salaries from the first to the second year.

First, the **PERCENT CHANGE FORMULA**:

$$((\text{NEW no.} - \text{OLD no.})/\text{OLD no.})$$
$$*100$$

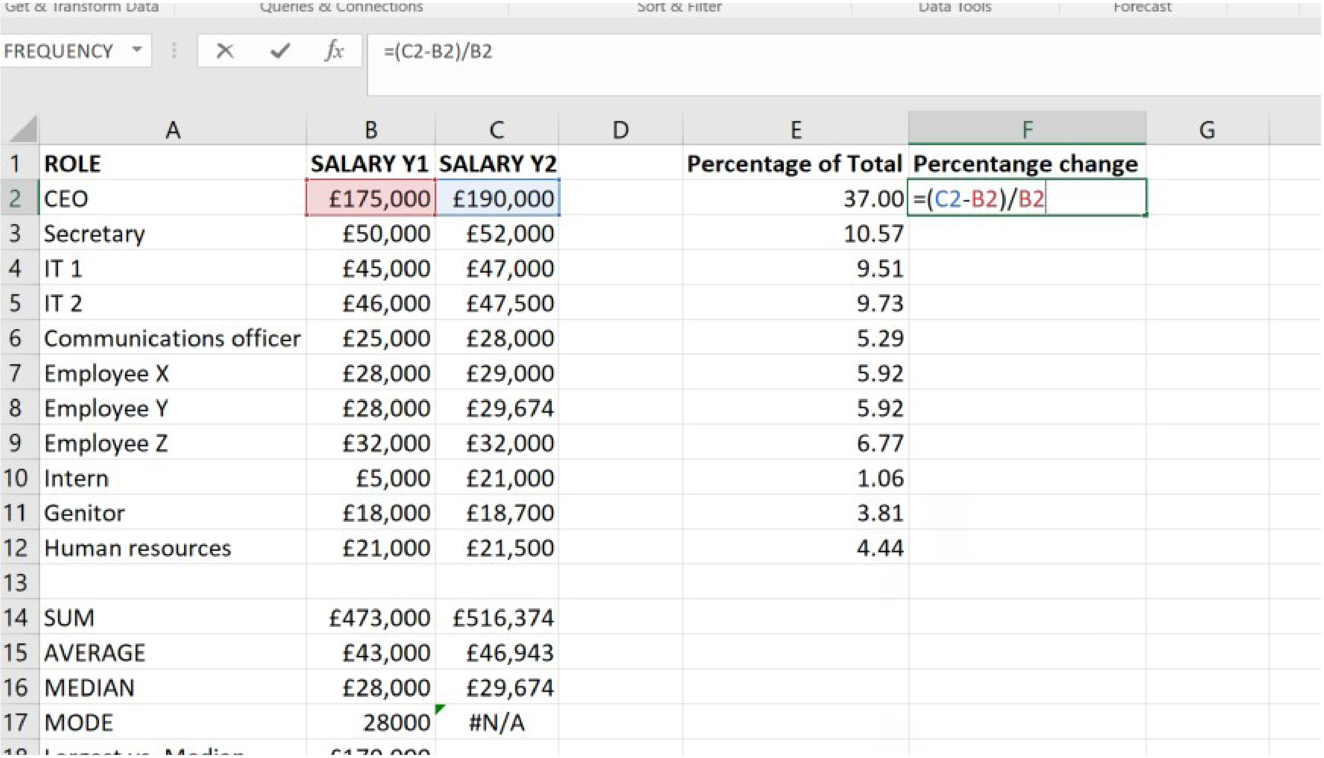
Or, in short,

(NOO)%

(Table 21)

After running the equation (NEW no. – OLD no.)/OLD no., you will get a result that looks more like a jumble of numbers. You can't work with that, yet. To transform it into a user-friendly percentage, including adding the percentage symbol to the final result, go on the HOME tab and, in the middle of the ribbon, you will find the % symbol. Select your long number and hit %. Hooray, you've got a healthy-looking percentage!

If you choose to multiply *100 rather than using the percentage sign (because you might like to do it all in one go), you won't see the % (percentage) symbol.



=(C2-B2)/B2

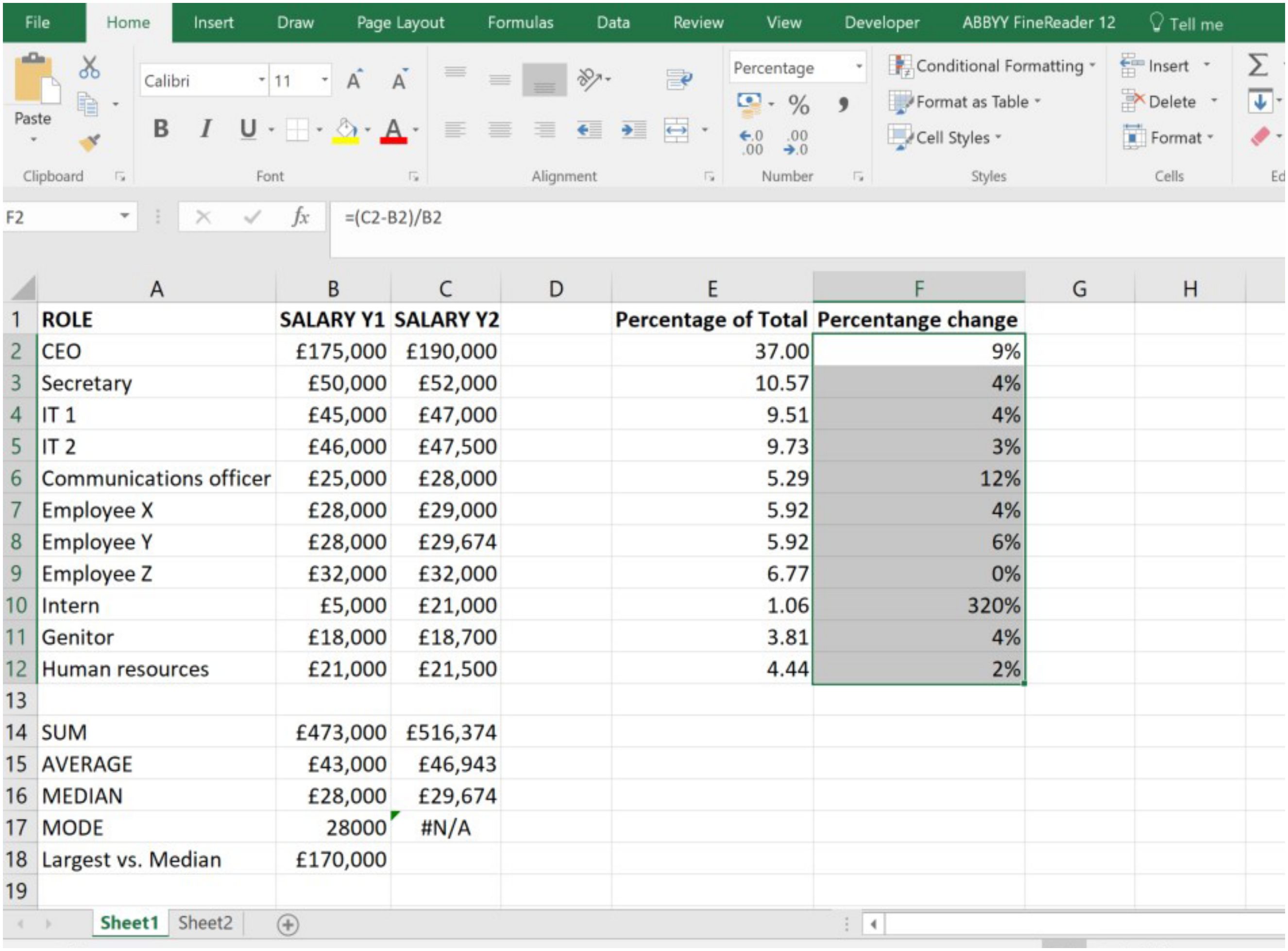
	A	B	C	D	E	F	G
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total	Percentange change	
2	CEO	£175,000	£190,000		37.00	=(C2-B2)/B2	
3	Secretary	£50,000	£52,000		10.57		
4	IT 1	£45,000	£47,000		9.51		
5	IT 2	£46,000	£47,500		9.73		
6	Communications officer	£25,000	£28,000		5.29		
7	Employee X	£28,000	£29,000		5.92		
8	Employee Y	£28,000	£29,674		5.92		
9	Employee Z	£32,000	£32,000		6.77		
10	Intern	£5,000	£21,000		1.06		
11	Genitor	£18,000	£18,700		3.81		
12	Human resources	£21,000	£21,500		4.44		
13							
14	SUM	£473,000	£516,374				
15	AVERAGE	£43,000	£46,943				
16	MEDIAN	£28,000	£29,674				
17	MODE	28000	#N/A				
18	Largest vs. Median	£170,000					

Table 21

The New Number is the latest value/salary. The Oldest Number is the first value/salary.
This translates into:

=(C2-B2)/B2

Next, click on the HOME tab. In the middle of the Home ribbon, find the % sign. Click on it. You now have 9%. Drag down the column or double-click the result to reveal the percentage changes for all the salaries (Table 22).



F2 =(C2-B2)/B2

	A	B	C	D	E	F	G	H
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total	Percentange change		
2	CEO	£175,000	£190,000		37.00	9%		
3	Secretary	£50,000	£52,000		10.57	4%		
4	IT 1	£45,000	£47,000		9.51	4%		
5	IT 2	£46,000	£47,500		9.73	3%		
6	Communications officer	£25,000	£28,000		5.29	12%		
7	Employee X	£28,000	£29,000		5.92	4%		
8	Employee Y	£28,000	£29,674		5.92	6%		
9	Employee Z	£32,000	£32,000		6.77	0%		
10	Intern	£5,000	£21,000		1.06	320%		
11	Genitor	£18,000	£18,700		3.81	4%		
12	Human resources	£21,000	£21,500		4.44	2%		
13								
14	SUM	£473,000	£516,374					
15	AVERAGE	£43,000	£46,943					
16	MEDIAN	£28,000	£29,674					
17	MODE	28000	#N/A					
18	Largest vs. Median	£170,000						
19								

Table 22

SUMMARISING WITH CHARTS

Data analysis does not only comprise maths or functions. A graph can sometimes do a great job to summarise a situation that your audience can understand at a glance, without reading copious amounts of text. And let's face it, editors love a bit of visualised data.

Bar charts and pie charts

Excel provides a few graphs, charts and illustration options that help us understand a trend, an existent relation between two variables, or the importance of a number. The most popular ones are pie charts, bar charts, scatter plots with trend lines, all found on the INSERT ribbon. Newer versions of Excel offer mapping tools and pivot charts.

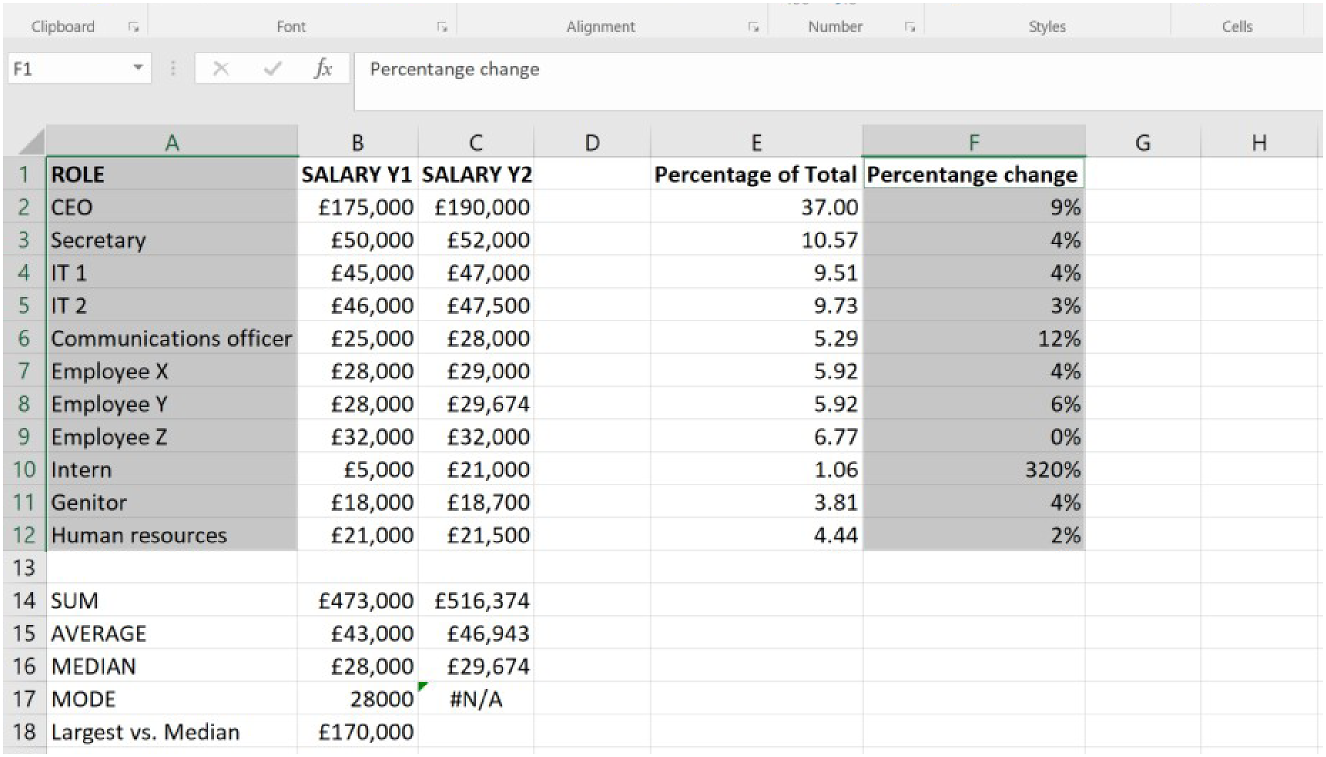
Data visualisation is an art in its own rights. This manual does not teach the art of visualising data, but rather a quick and easy spreadsheet visualisation of results for analysis purposes.

It is nonetheless important to note that for the purpose of analysis, you should use the simplest versions of any chart or graph, that is to say a 2D version, with no shadows or sophisticated design. This is because 3D and shadows often manipulate the perception of an image, making one slice of a pie chart seem bigger or different from its actual size. Steer clear of pretty graphics that manipulate the reader. but challenge their use when you see them.

Use design with care and always make it secondary to and supportive of what you're actually presenting and reporting: new information to an audience. And remember: the design is not the goal, the story is. Without the story and the reporting efforts behind it, there would be no graph. No data design can ever substitute for, or exist in the absence of, good data reporting.

Let's use some of the numbers we've been computing on a chart.
First, we need to select two columns that are not direct neighbours: A (employees' roles) and F (salary percentage changes):
>Click in A1.
>Hold SHIFT and click on A12.
>Release.
>Hold CTRL (or Command for MACs) and click in F1.
>Release.
>Hold SHIFT and click F12.

The selection looks like the one in Table 23.



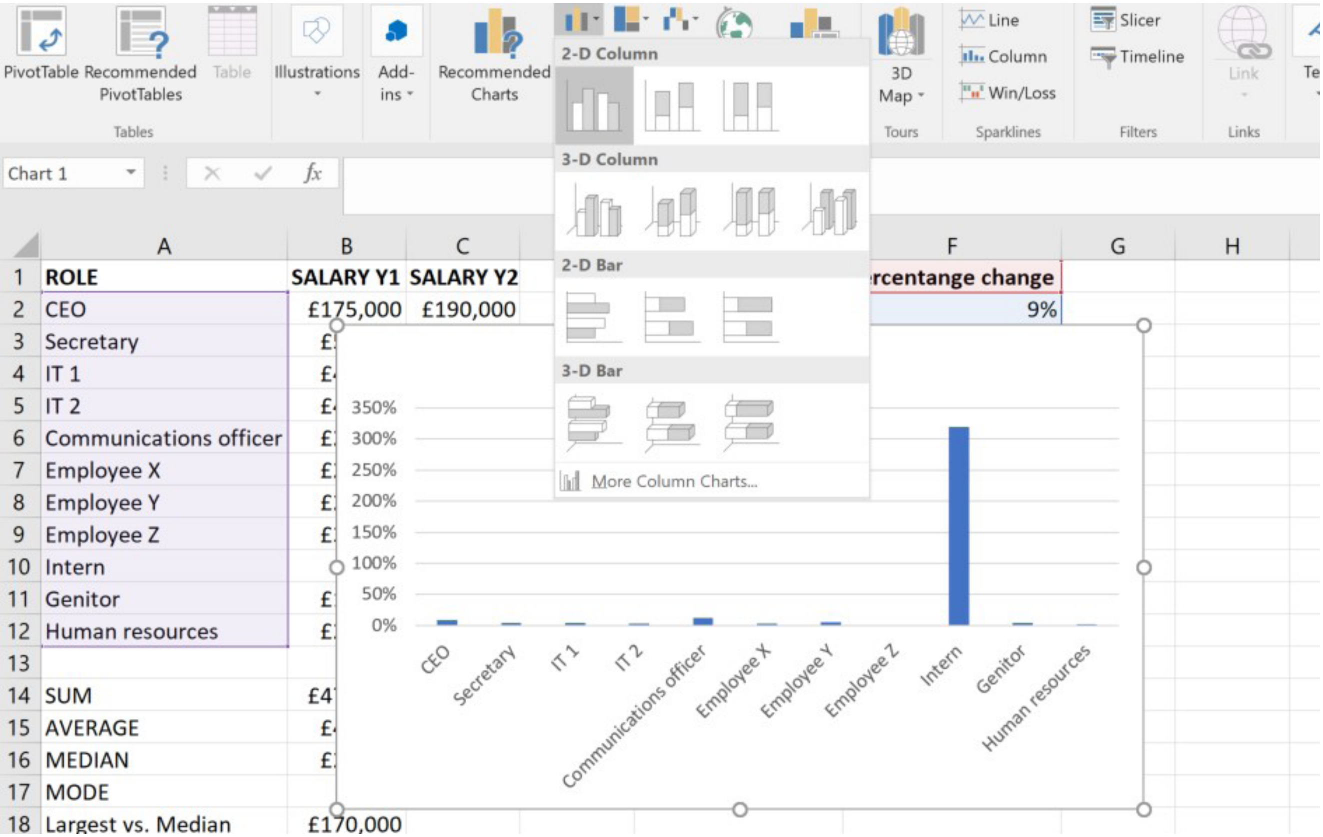
Percentange change

	A	B	C	D	E	F	G	H
1	**ROLE**	**SALARY Y1**	**SALARY Y2**		**Percentage of Total**	**Percentange change**		
2	CEO	£175,000	£190,000		37.00	9%		
3	Secretary	£50,000	£52,000		10.57	4%		
4	IT 1	£45,000	£47,000		9.51	4%		
5	IT 2	£46,000	£47,500		9.73	3%		
6	Communications officer	£25,000	£28,000		5.29	12%		
7	Employee X	£28,000	£29,000		5.92	4%		
8	Employee Y	£28,000	£29,674		5.92	6%		
9	Employee Z	£32,000	£32,000		6.77	0%		
10	Intern	£5,000	£21,000		1.06	320%		
11	Genitor	£18,000	£18,700		3.81	4%		
12	Human resources	£21,000	£21,500		4.44	2%		
13								
14	SUM	£473,000	£516,374					
15	AVERAGE	£43,000	£46,943					
16	MEDIAN	£28,000	£29,674					
17	MODE	28000	#N/A					
18	Largest vs. Median	£170,000						

Table 23

Time introduce a chart!

Click on the INSERT tab.
In the middle of its ribbon, you will find the CHARTS area.
Select the Bar charts > 2-D > Clustered columns to chart the percentage increases (Table 24).
At a glance, you can see the winner: one of the interns has had a massive increase in salary, percentage wise.



	A	B	C	F	G	H
1	ROLE	SALARY Y1	SALARY Y2	rcentange change		
2	CEO	£175,000	£190,000	9%		
3	Secretary	£				
4	IT 1	£				
5	IT 2	£				
6	Communications officer	£				
7	Employee X	£				
8	Employee Y	£				
9	Employee Z	£				
10	Intern					
11	Genitor	£				
12	Human resources	£				
13						
14	SUM	£4				
15	AVERAGE	£				
16	MEDIAN	£				
17	MODE					
18	Largest vs. Median	£170,000				

Table 24

If you click on any of the bars in the chart, you will select them all. Right-click to see a formatting window. Select ADD DATA LABELS (Table 25) to have the values showing on top of the bars (Table 26).

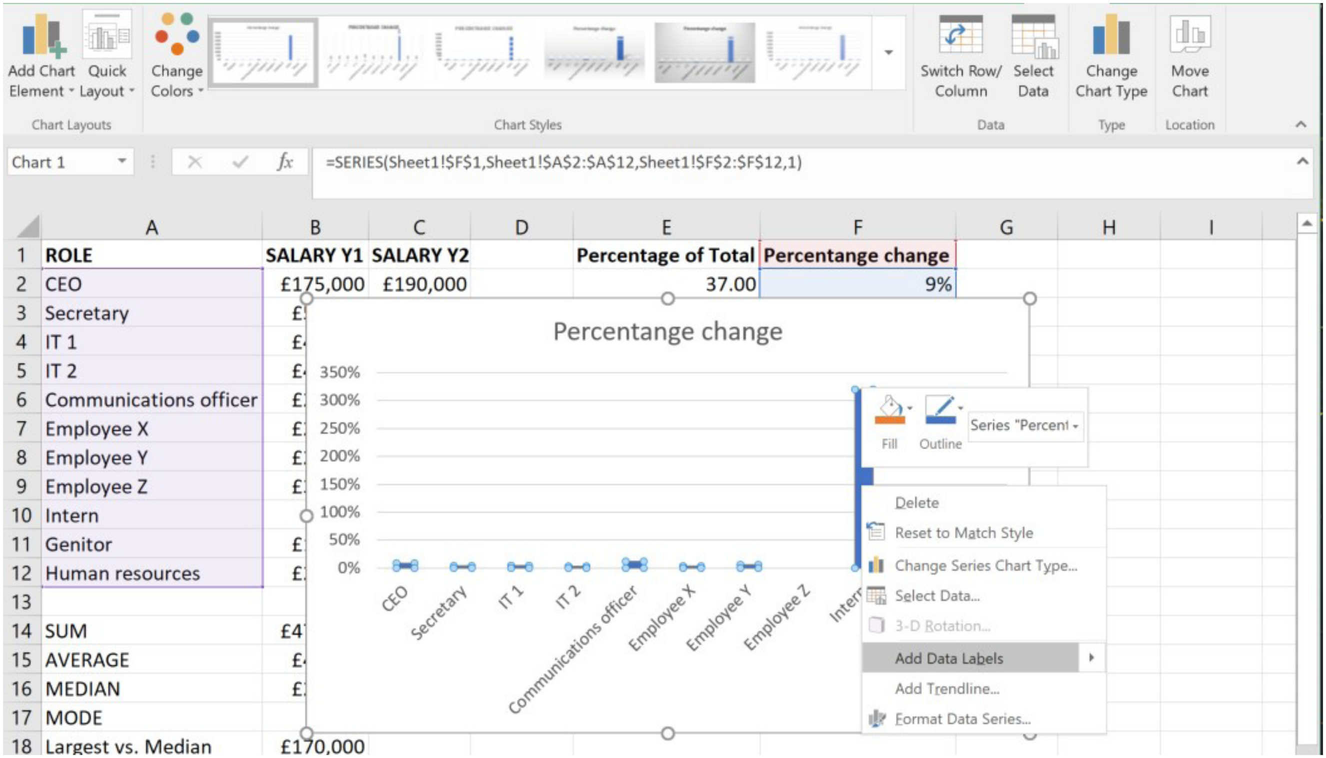


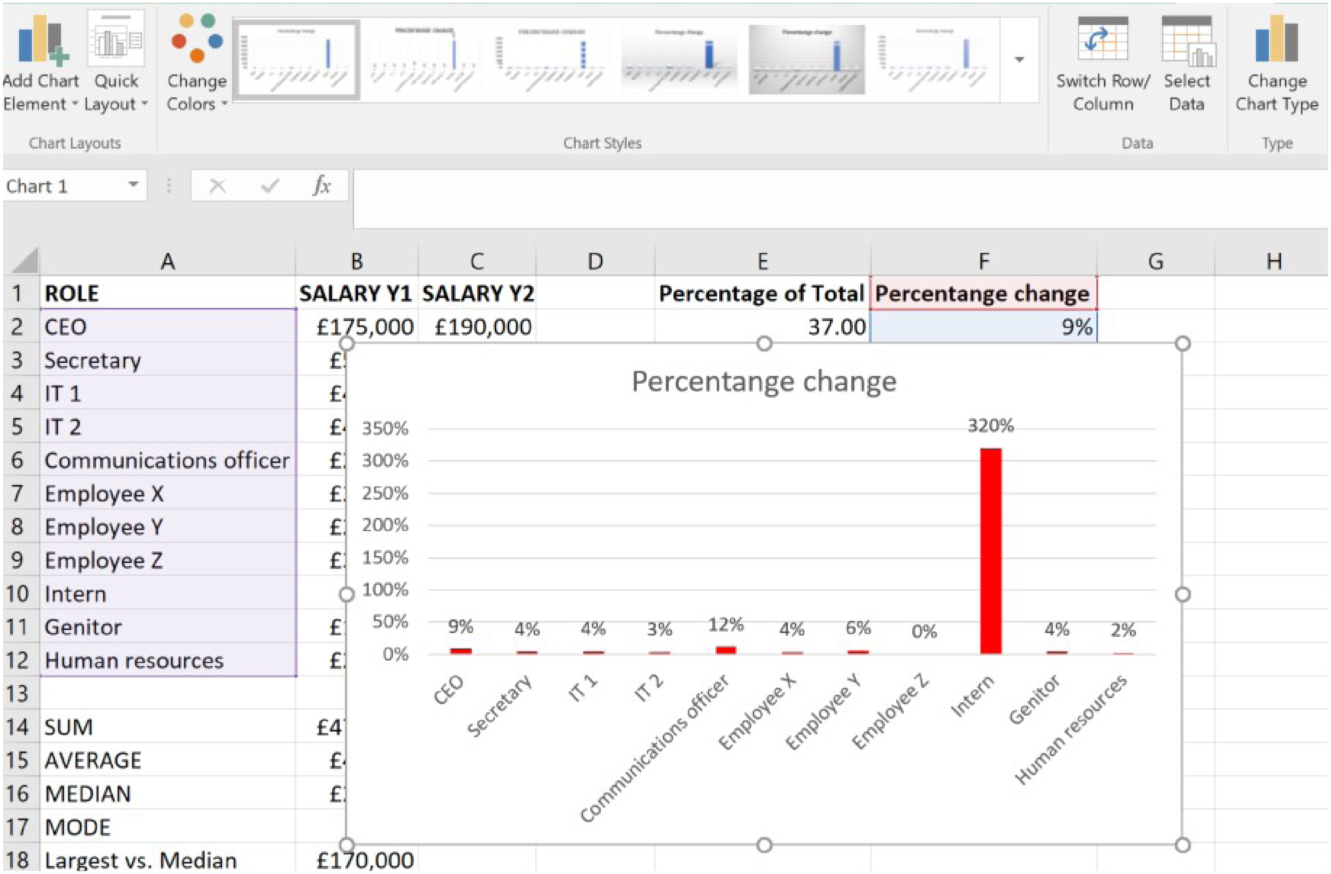
Chart 1

=SERIES(Sheet1!$F$1,Sheet1!$A$2:$A$12,Sheet1!$F$2:$F$12,1)

	A	B	C	D	E	F
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total	Percentange change
2	CEO	£175,000	£190,000		37.00	9%
3	Secretary	£				
4	IT 1	£				
5	IT 2	£				
6	Communications officer	£				
7	Employee X	£				
8	Employee Y	£				
9	Employee Z	£				
10	Intern					
11	Genitor	£				
12	Human resources	£				
13						
14	SUM	£4				
15	AVERAGE	£				
16	MEDIAN	£				
17	MODE					
18	Largest vs. Median	£170,000				

Table 25

Change the colour by clicking on the FILL bucket and select another colour. You can also adjust one bar at a time. And you can also use the settings here to adjust the space between columns, as well as adding frames around the bars or around the chart itself.

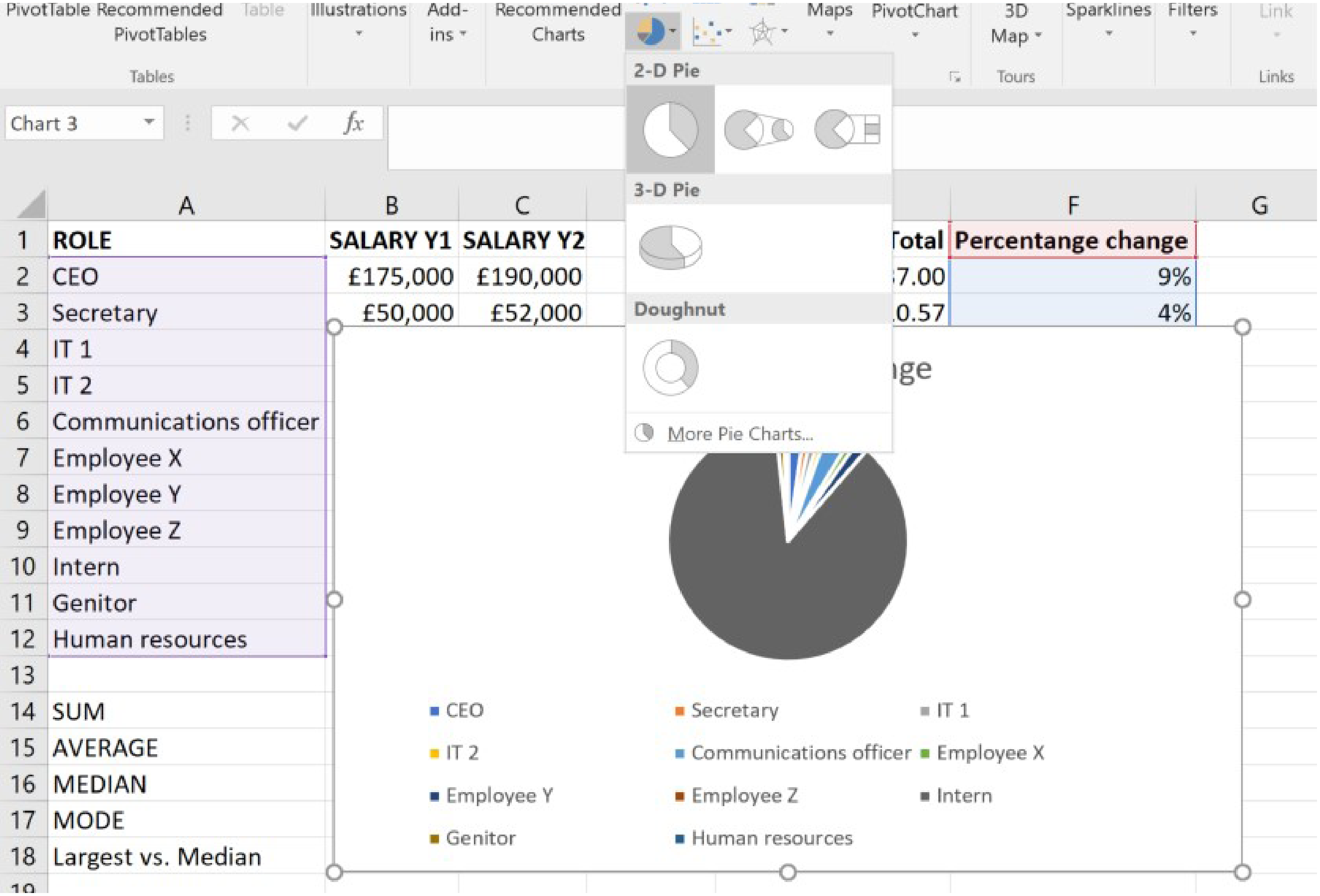


	A	B	C	D	E	F	G	H
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total	Percentange change		
2	CEO	£175,000	£190,000		37.00	9%		
3	Secretary	£						
4	IT 1	£						
5	IT 2	£						
6	Communications officer	£						
7	Employee X	£						
8	Employee Y	£						
9	Employee Z	£						
10	Intern							
11	Genitor	£						
12	Human resources	£						
13								
14	SUM	£4						
15	AVERAGE	£						
16	MEDIAN	£						
17	MODE							
18	Largest vs. Median	£170,000						

Table 26

If you decide that vertical lines are not the best way to see your numbers, you could decide to slice a pie chart.

Make the same selection, and on the INSERT ribbon, Charts area, select INSERT PIE OR DOUGHNUT CHART > 2-D Pie (Table 27). Data labels, colour changes and frames can be added. You can also give your chart a title of your choice.



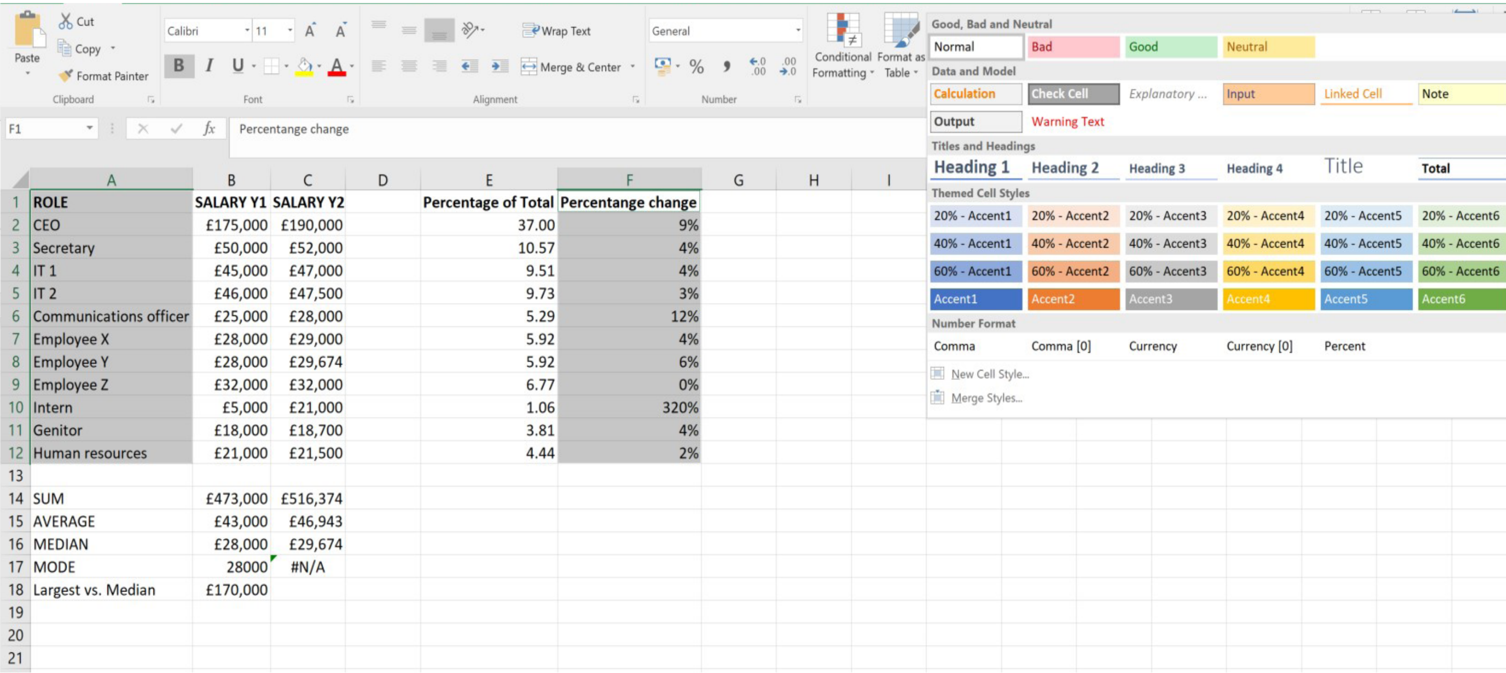
	A	B	C		F	G
1	ROLE	SALARY Y1	SALARY Y2	Total	Percentange change	
2	CEO	£175,000	£190,000	7.00	9%	
3	Secretary	£50,000	£52,000	0.57	4%	
4	IT 1					
5	IT 2					
6	Communications officer					
7	Employee X					
8	Employee Y					
9	Employee Z					
10	Intern					
11	Genitor					
12	Human resources					
13						
14	SUM					
15	AVERAGE					
16	MEDIAN					
17	MODE					
18	Largest vs. Median					

Table 27

When you chart values, try several options and decide which one works best case by case.
To delete a chart, click anywhere in the white space of the graph, then press DELETE on your keyboard.

Tip

Before you start computing numbers, you may wish to make the spreadsheet look more inviting. You can use the HOME ribbon STYLES and CELLS area to play around with different formatting options, colours and cell sizes (Table 28).



F1 Percentange change

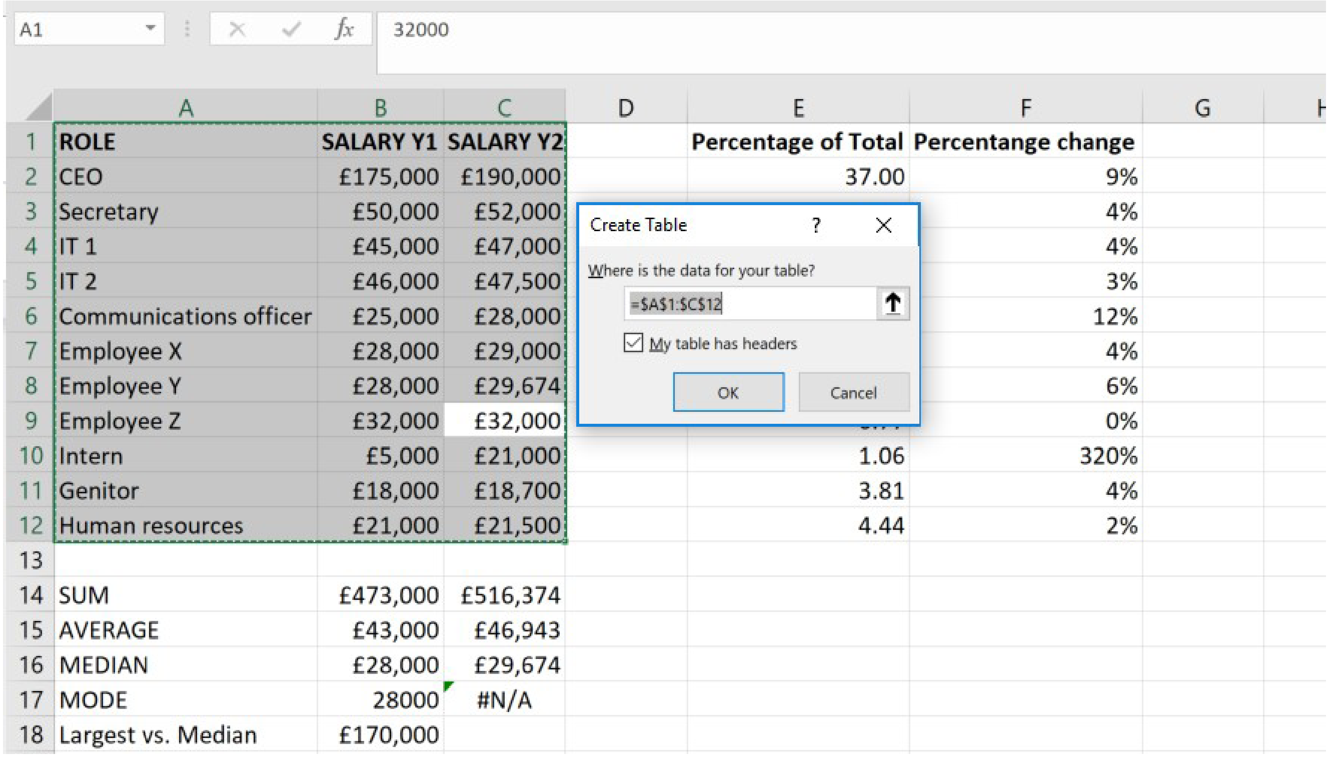
	A	B	C	D	E	F	G	H	I
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total	Percentange change			
2	CEO	£175,000	£190,000		37.00	9%			
3	Secretary	£50,000	£52,000		10.57	4%			
4	IT 1	£45,000	£47,000		9.51	4%			
5	IT 2	£46,000	£47,500		9.73	3%			
6	Communications officer	£25,000	£28,000		5.29	12%			
7	Employee X	£28,000	£29,000		5.92	4%			
8	Employee Y	£28,000	£29,674		5.92	6%			
9	Employee Z	£32,000	£32,000		6.77	0%			
10	Intern	£5,000	£21,000		1.06	320%			
11	Genitor	£18,000	£18,700		3.81	4%			
12	Human resources	£21,000	£21,500		4.44	2%			
13									
14	SUM	£473,000	£516,374						
15	AVERAGE	£43,000	£46,943						
16	MEDIAN	£28,000	£29,674						
17	MODE	28000	#N/A						
18	Largest vs. Median	£170,000							
19									
20									
21									

Table 28

Tip

There may be cases when you have several tables on one sheet. Each table has to be formatted into official tables. EXCEL has to be told that a range is TABLE 1, another range is TABLE 2, etc. This is particularly important when you're looking to join matching records in two or more tables to create a more comprehensive dataset.

To transform a dataset into a table, select the table > press CTRL+T (Table 29A).



A1 | 32000

	A	B	C	D	E	F
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total	Percentange change
2	CEO	£175,000	£190,000		37.00	9%
3	Secretary	£50,000	£52,000			4%
4	IT 1	£45,000	£47,000			4%
5	IT 2	£46,000	£47,500			3%
6	Communications officer	£25,000	£28,000			12%
7	Employee X	£28,000	£29,000			4%
8	Employee Y	£28,000	£29,674			6%
9	Employee Z	£32,000	£32,000		[illegible]	0%
10	Intern	£5,000	£21,000		1.06	320%
11	Genitor	£18,000	£18,700		3.81	4%
12	Human resources	£21,000	£21,500		4.44	2%
13						
14	SUM	£473,000	£516,374			
15	AVERAGE	£43,000	£46,943			
16	MEDIAN	£28,000	£29,674			
17	MODE	28000	#N/A			
18	Largest vs. Median	£170,000				

Create Table

Where is the data for your table?

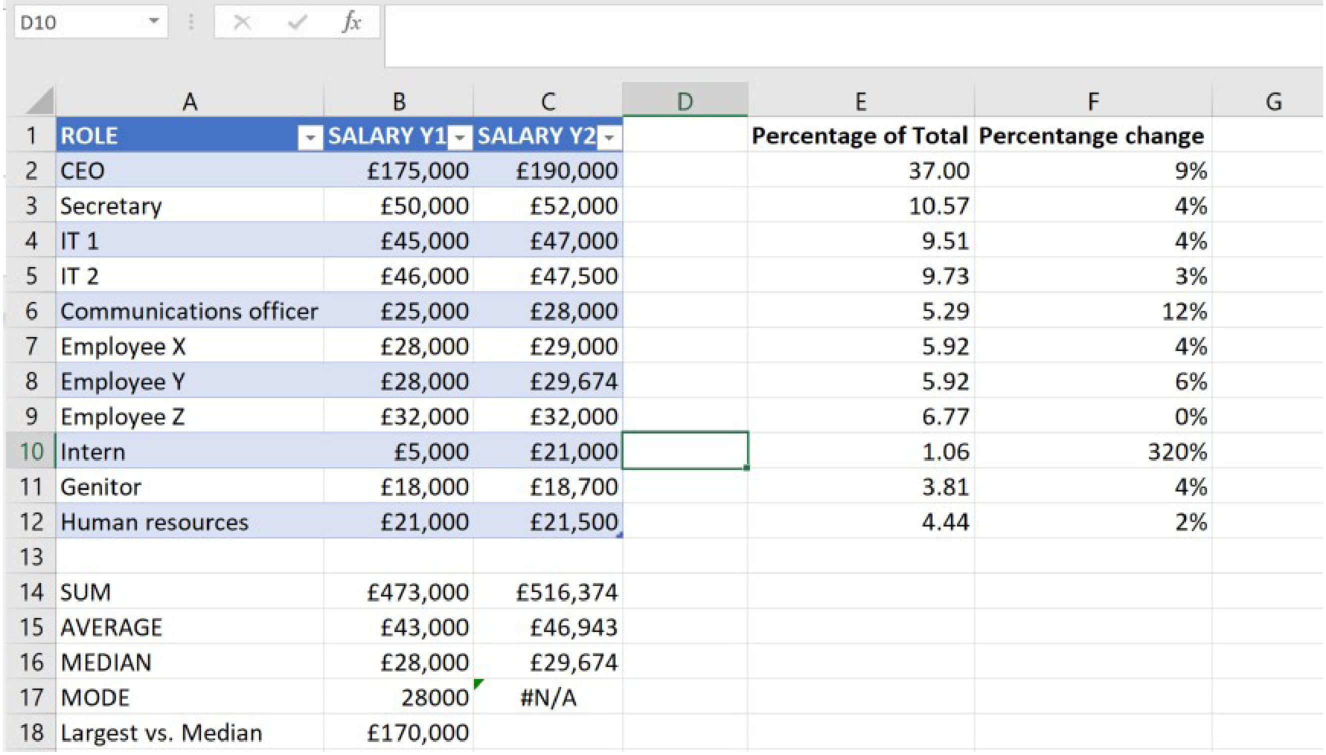
=$A$1:$C$12

☑ My table has headers

OK | Cancel

Table 29A

To exert more control over how a table will look like, use the STYLES section:
> select your data (CTRL+A)
> click on the HOME tab
> select Styles
> format as tables (Table 29 A). A window displaying the data range, anchored, comes up. >You can edit the data range if you'd like to include more rows and columns. When you're ready, click OK. The table will appear formatted (table 29B).



D10

	A	B	C	D	E	F	G
1	ROLE	SALARY Y1	SALARY Y2		Percentage of Total	Percentange change	
2	CEO	£175,000	£190,000		37.00	9%	
3	Secretary	£50,000	£52,000		10.57	4%	
4	IT 1	£45,000	£47,000		9.51	4%	
5	IT 2	£46,000	£47,500		9.73	3%	
6	Communications officer	£25,000	£28,000		5.29	12%	
7	Employee X	£28,000	£29,000		5.92	4%	
8	Employee Y	£28,000	£29,674		5.92	6%	
9	Employee Z	£32,000	£32,000		6.77	0%	
10	Intern	£5,000	£21,000		1.06	320%	
11	Genitor	£18,000	£18,700		3.81	4%	
12	Human resources	£21,000	£21,500		4.44	2%	
13							
14	SUM	£473,000	£516,374				
15	AVERAGE	£43,000	£46,943				
16	MEDIAN	£28,000	£29,674				
17	MODE	28000	#N/A				
18	Largest vs. Median	£170,000					

Table 29B

CURATING AND SUMMARISING DATA

There are a few actions that a spreadsheet can take to curate the data as you prefer. You may want your data listed in a certain way; edited to focus on particular companies / projects / stakeholders; or, maybe you're interested in records that meet certain conditions. Spreadsheets offer tools that help with an overview, or zoom in on details or aggregate thousands of records in split seconds.

For this section, we'll start with WHO's Life Expectancy dataset.

Sorting & Filtering

Think of sorting as a tool to organise your data.
Filtering is for getting a close up picture of a detail or to curate a dataset.
The sort and filter tools are located on the DATA ribbon in the SORT & FILTER area. These tools can also be used to clean data or to ask questions.

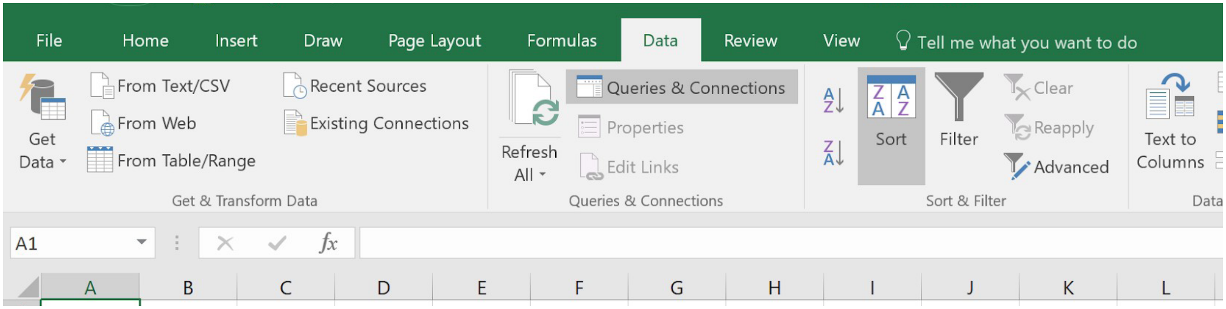


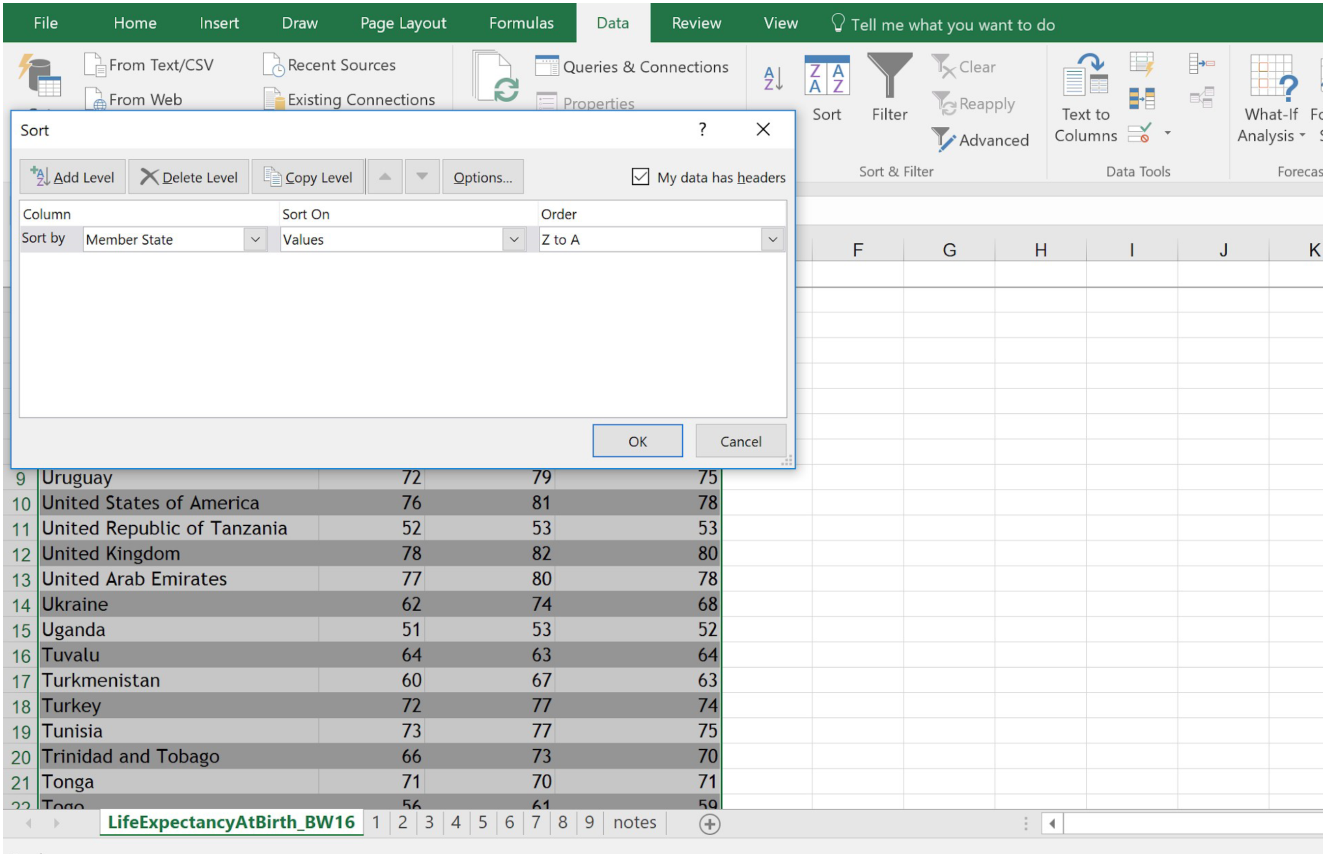
Figure 1

A-Z or Z-A buttons perform quick sorts: alphabetically or reverse alphabetically, ascending or descending. If you click in a column, you can sort the whole table based on that column.

WARNING: Older versions of Excel only sorted that one column, but not the rest of the records. **This has made grown people cry!** Make sure you always select the whole table (with CTRL+A or COMMAND+A) before hitting the sort button.

For a more flexible sort, use the sort wizard (button). This will open a dialogue window for a custom sort, where you can add one or several conditions.

Download Life_expectancy_ed_WHS2008 for the next exercise by clicking here.



Sort

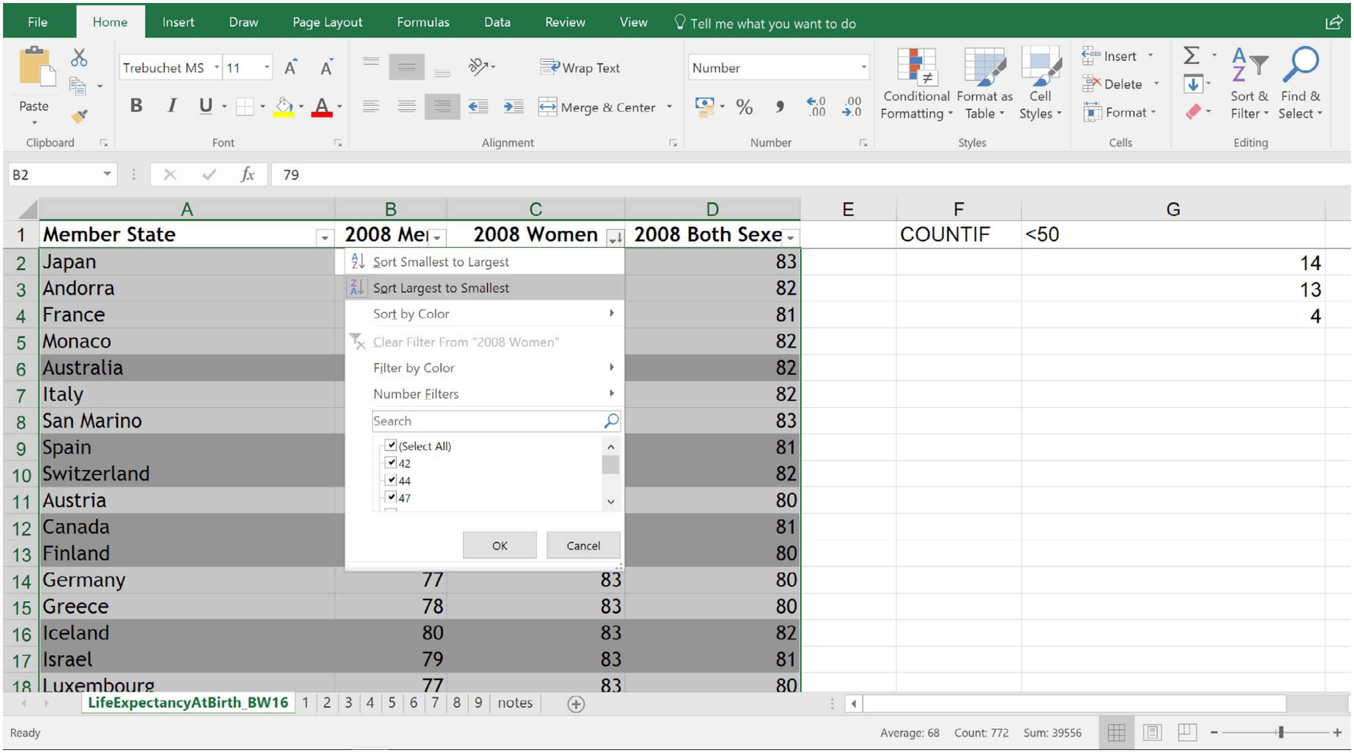
☑ My data has headers

Column		Sort On	Order
Sort by	Member State	Values	Z to A

9	Uruguay	72	79	75
10	United States of America	76	81	78
11	United Republic of Tanzania	52	53	53
12	United Kingdom	78	82	80
13	United Arab Emirates	77	80	78
14	Ukraine	62	74	68
15	Uganda	51	53	52
16	Tuvalu	64	63	64
17	Turkmenistan	60	67	63
18	Turkey	72	77	74
19	Tunisia	73	77	75
20	Trinidad and Tobago	66	73	70
21	Tonga	71	70	71
22	Togo	56	61	59

Table 30

Sort longest to shortest life expectancy for women - **in which part of the world do women live the longest?**
> Click CTRL+A to select your table.
>Select the DATA tab and click on the FILTER symbol. This will add a small drop-down button to each one of your table's columns.
>Click on cell C1, where the section header is.
>Select "Largest to Smallest" and press "Enter" (Table 31).
The results will show that women lived on average to be 86 in Japan in 2008, the highest longevity for females at that time (Table 32).

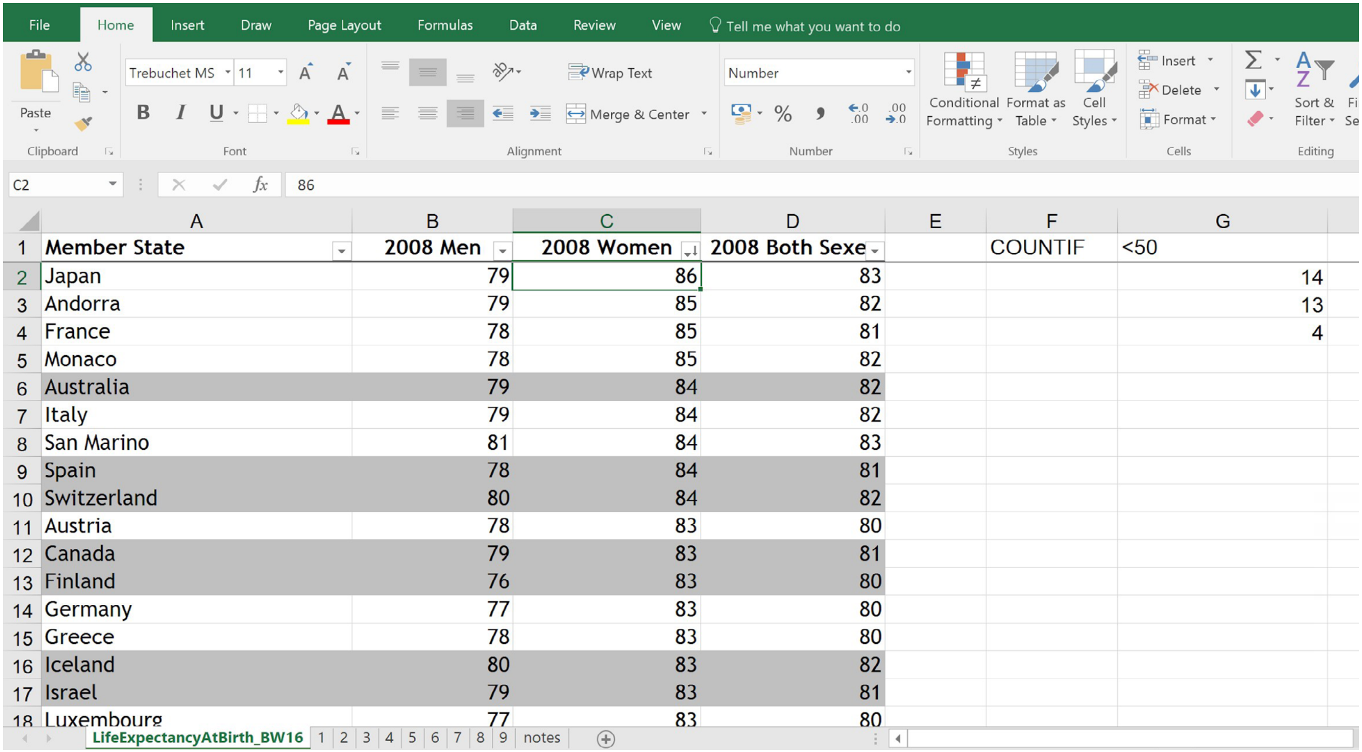


B2 | 79

	A	B	C	D	E	F	G
1	Member State	2008 Mei	2008 Women	2008 Both Sexe		COUNTIF	<50
2	Japan			83			14
3	Andorra			82			13
4	France			81			4
5	Monaco			82			
6	Australia			82			
7	Italy			82			
8	San Marino			83			
9	Spain			81			
10	Switzerland			82			
11	Austria			80			
12	Canada			81			
13	Finland			80			
14	Germany	77	83	80			
15	Greece	78	83	80			
16	Iceland	80	83	82			
17	Israel	79	83	81			
18	Luxembourg	77	83	80			

Average: 68 Count: 772 Sum: 39556

Table 31



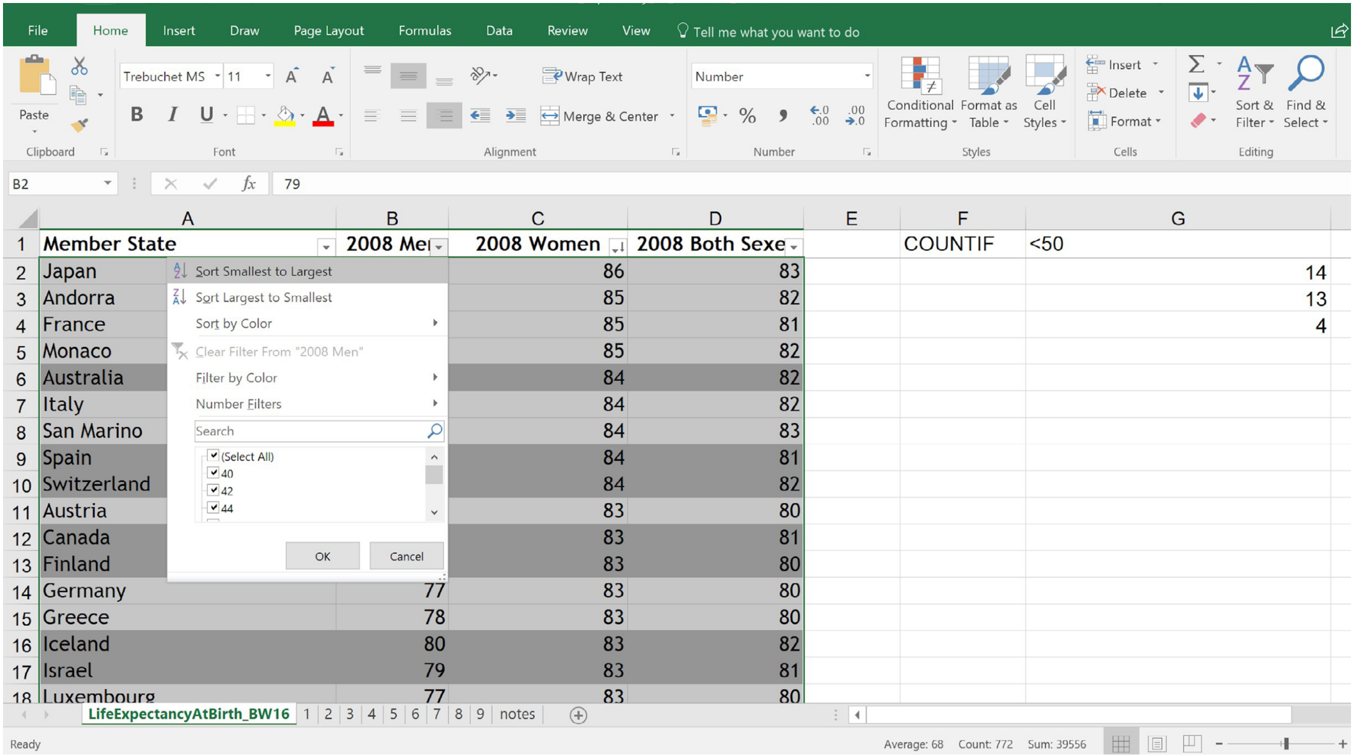
C2 | 86

	A	B	C	D	E	F	G
1	Member State	2008 Men	2008 Women	2008 Both Sexe		COUNTIF	<50
2	Japan	79	86	83			14
3	Andorra	79	85	82			13
4	France	78	85	81			4
5	Monaco	78	85	82			
6	Australia	79	84	82			
7	Italy	79	84	82			
8	San Marino	81	84	83			
9	Spain	78	84	81			
10	Switzerland	80	84	82			
11	Austria	78	83	80			
12	Canada	79	83	81			
13	Finland	76	83	80			
14	Germany	77	83	80			
15	Greece	78	83	80			
16	Iceland	80	83	82			
17	Israel	79	83	81			
18	Luxembourg	77	83	80			

Table 32

Switch to life expectancy for men. **In which part of the world do men have the shortest lives?**
>Click on the filter button in column header B1 and select “Sort Largest to Smallest” (Table 33). Result: Afghanistan - where men live to be 40 on average (Table 34).

How close are you from this age? It’s worthwhile testing the impact of a result / answer on yourself when searching for a story or colour.



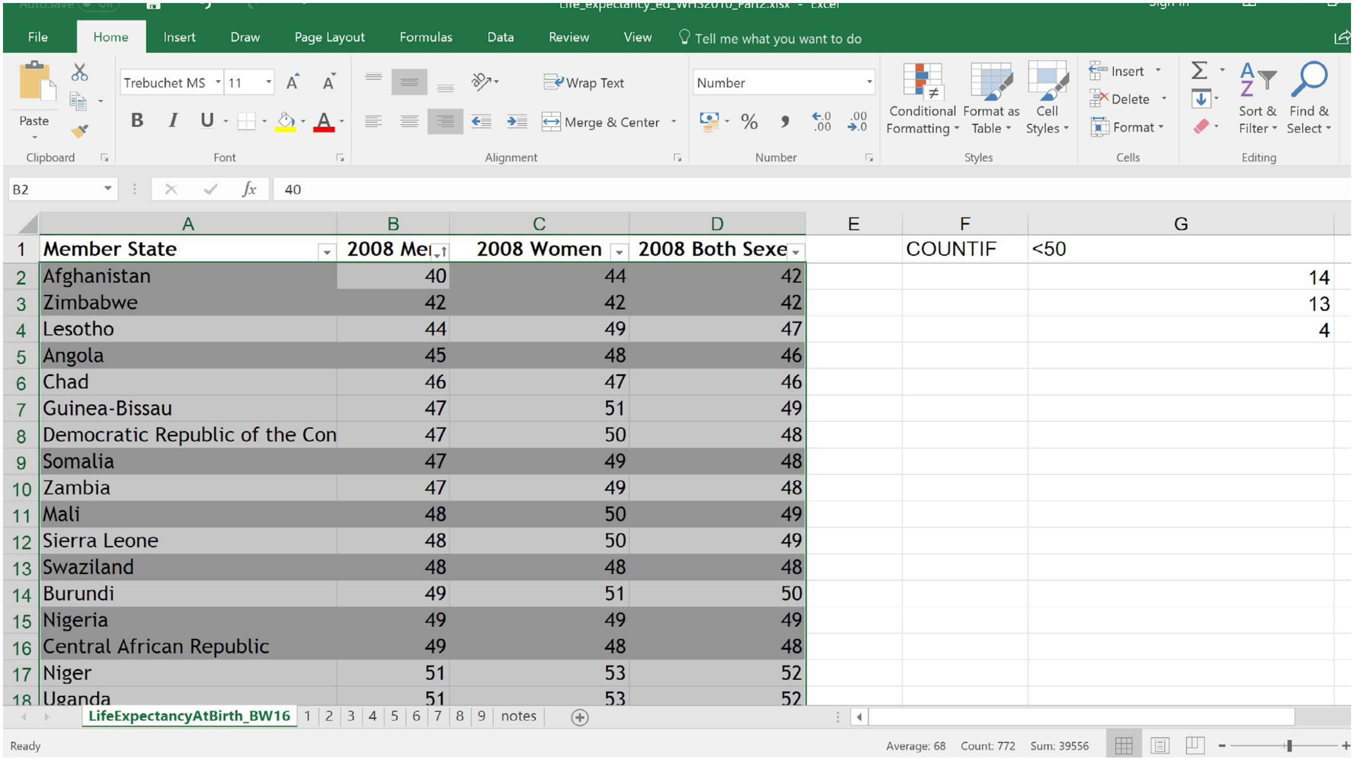
B2 | fx | 79

☑ (Select All)
☑ 40
☑ 42
☑ 44

	A	B	C	D	E	F	G
1	Member State	2008 Me	2008 Women	2008 Both Sexe		COUNTIF	<50
2	Japan		86	83			14
3	Andorra		85	82			13
4	France		85	81			4
5	Monaco		85	82			
6	Australia		84	82			
7	Italy		84	82			
8	San Marino		84	83			
9	Spain		84	81			
10	Switzerland		84	82			
11	Austria		83	80			
12	Canada		83	81			
13	Finland		83	80			
14	Germany	77	83	80			
15	Greece	78	83	80			
16	Iceland	80	83	82			
17	Israel	79	83	81			
18	Luxembourg	77	83	80			

Average: 68 Count: 772 Sum: 39556

Table 33



B2 | 40

	A	B	C	D	E	F	G
1	Member State	2008 Me	2008 Women	2008 Both Sexe		COUNTIF	<50
2	Afghanistan	40	44	42			14
3	Zimbabwe	42	42	42			13
4	Lesotho	44	49	47			4
5	Angola	45	48	46			
6	Chad	46	47	46			
7	Guinea-Bissau	47	51	49			
8	Democratic Republic of the Con	47	50	48			
9	Somalia	47	49	48			
10	Zambia	47	49	48			
11	Mali	48	50	49			
12	Sierra Leone	48	50	49			
13	Swaziland	48	48	48			
14	Burundi	49	51	50			
15	Nigeria	49	49	49			
16	Central African Republic	49	48	48			
17	Niger	51	53	52			
18	Uganda	51	53	52			

LifeExpectancyAtBirth_BW16 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | notes

Average: 68 Count: 772 Sum: 39556

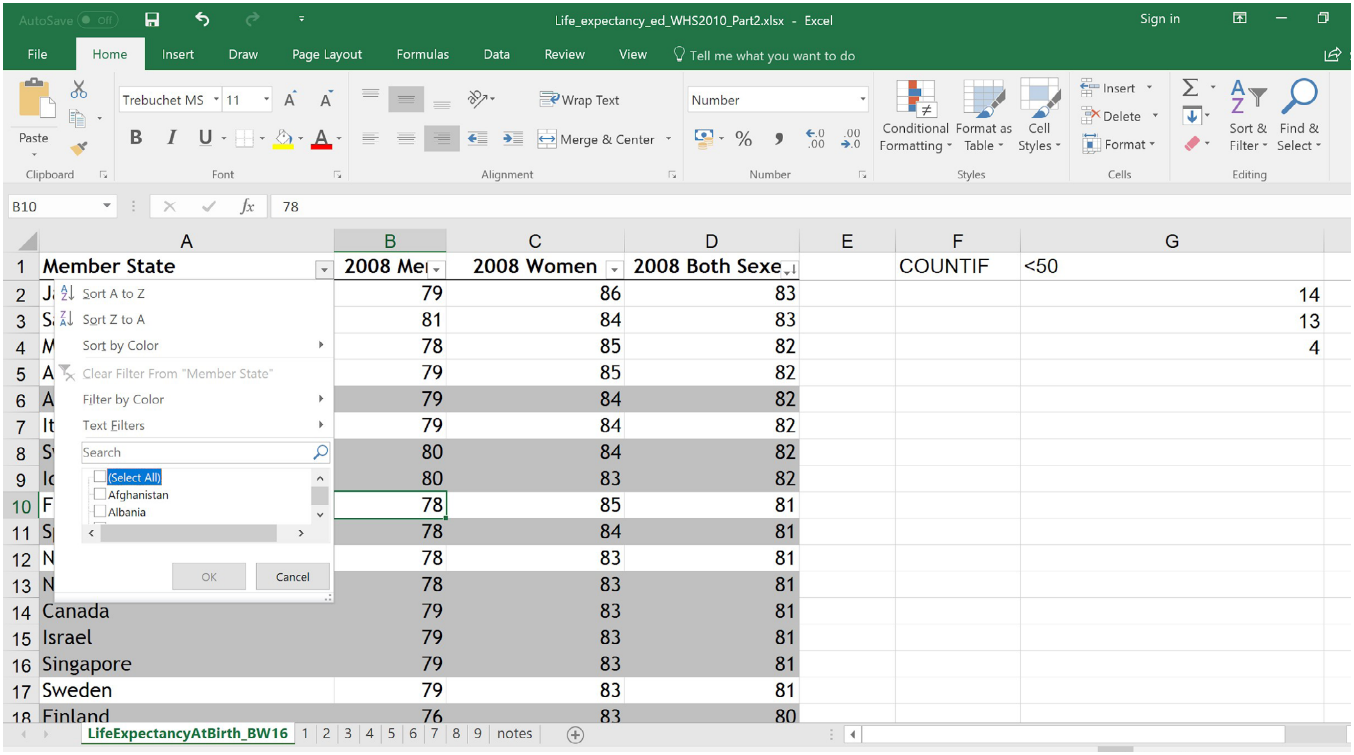
Table 34

The same result can be obtained by clicking anywhere in the column of interest and use one of the sorting shortcut buttons on the data ribbon. A to Z means an alphabetic or ascending sort (depending on the type of data), while Z to A means reverse-alphabetic or descending sort.

Next, run the sort on the average life expectancy for both sexes. **Where do we all live the longest lives?** *(Answer: Japan, 83.)*

Note that ***data or their analysis don't tell you why something happens***. At its best, you may see correlations if you later get into statistics. At this stage, you might be tempted to speculate on the reasons why we live longer lives in one country, but not in others. And while this makes for a good conversation with colleagues, you will need to talk to experts to find the actual reason before reporting.

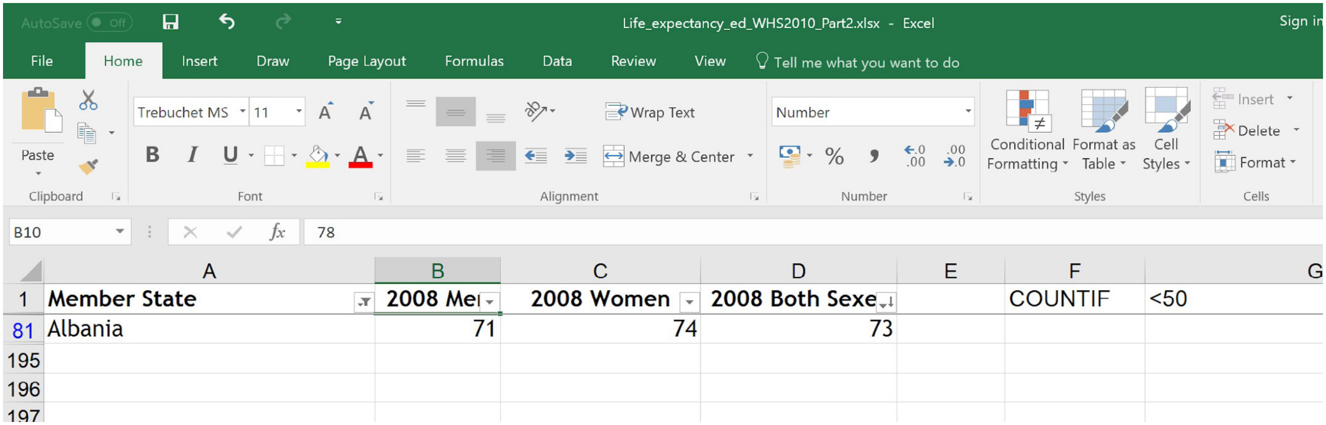
At this stage, you may be interested in your country. This means you will need to filter out all the other ones. To do that, click on A1, to open the dropdown menu for Member States. Notice that all countries are selected. Deselect them by removing the tick from “Select all” (Table 35).



	A	B	C	D	E	F	G
1	Member State	2008 Me	2008 Women	2008 Both Sexe		COUNTIF	<50
2	J	79	86	83			14
3	S	81	84	83			13
4	M	78	85	82			4
5	A	79	85	82			
6	A	79	84	82			
7	It	79	84	82			
8	S	80	84	82			
9	Ic	80	83	82			
10	F	78	85	81			
11	S	78	84	81			
12	N	78	83	81			
13	N	78	83	81			
14	Canada	79	83	81			
15	Israel	79	83	81			
16	Singapore	79	83	81			
17	Sweden	79	83	81			
18	Finland	76	83	80			

Table 35

To bring just your country of interest, type "Albania" in the search box and click OK (Table 36).

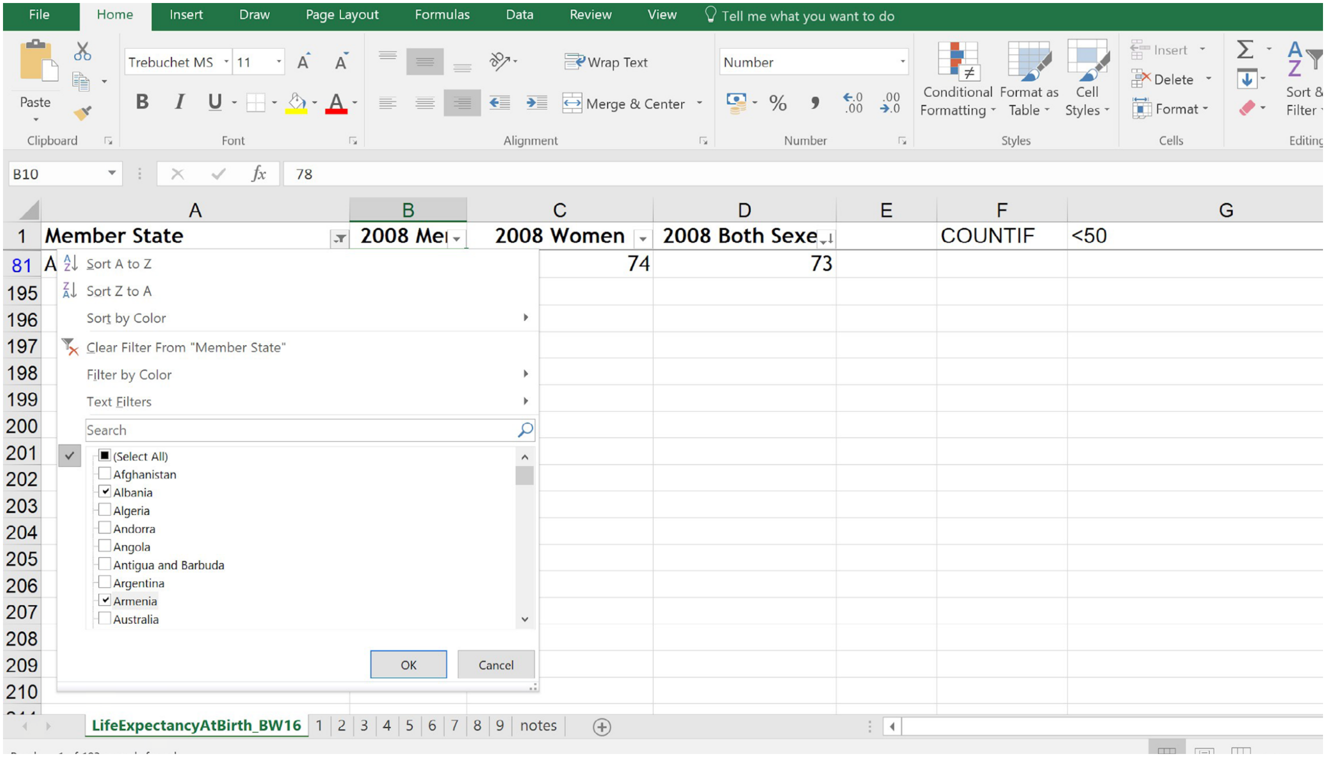


	A	B	C	D	E	F	G
1	Member State	2008 Me	2008 Women	2008 Both Sexe		COUNTIF	<50
81	Albania	71	74	73			
195							
196							
197							

Table 36

Now that you have found out the 2008 life expectancy in Albania, let's compare it with other countries: neighbours, best and worst countries, countries with similar GDP and socio-economic past, etc.

In the same drop-down menu, select several items by scrolling and ticking your chosen boxes (Table 37). To make the process more user-friendly, click and drag the bottom-right corner of the menu to enlarge it.



B10: 78

	A	B	C	D	E	F	G
1	Member State	2008 Me	2008 Women	2008 Both Sexe		COUNTIF	<50
81	A		74	73			

■ (Select All)
□ Afghanistan
☑ Albania
□ Algeria
□ Andorra
□ Angola
□ Antigua and Barbuda
□ Argentina
☑ Armenia
□ Australia

Table 37

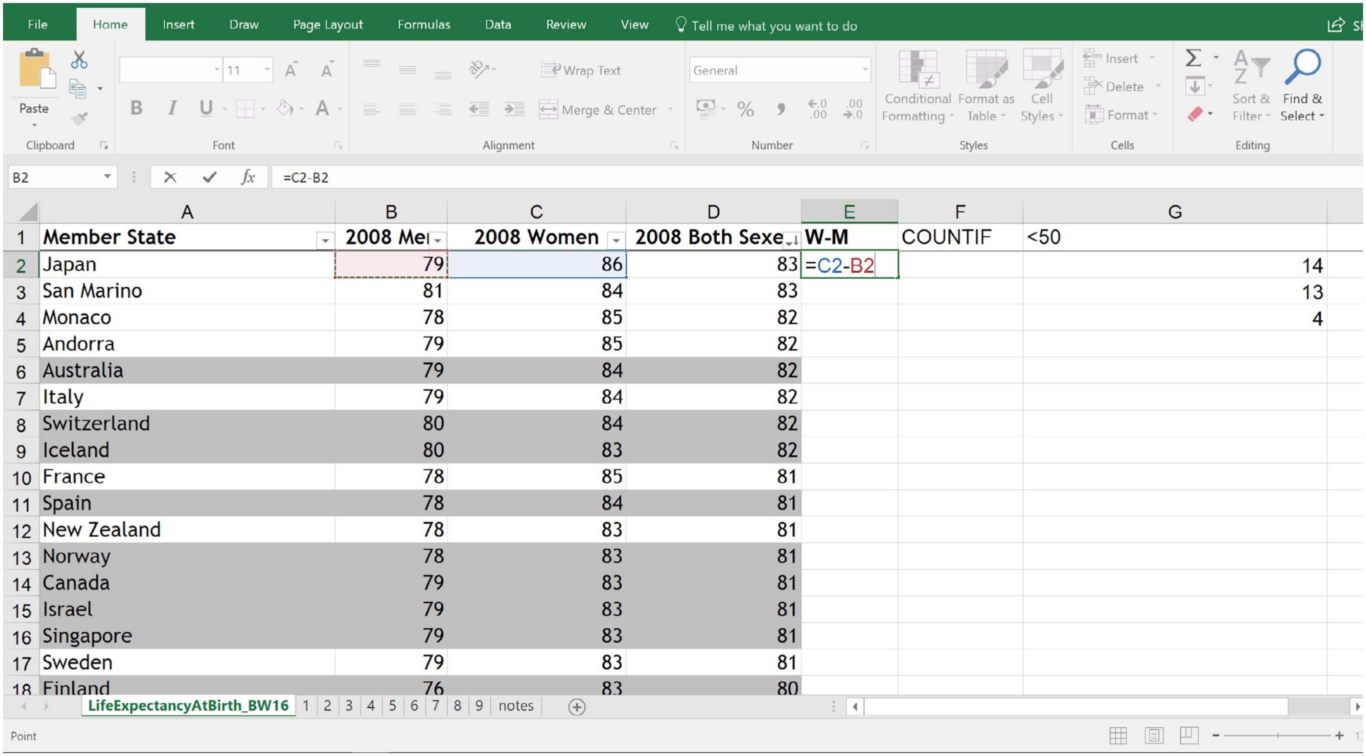
To bring all other records back, go inside the menu of the column you've filtered on, tick SELECT ALL > OK.

What else can you do with this data?
Use curiosity and instincts generously, no matter what data you have to work with.

As a next step, **create your own column**, to calculate **the difference in life expectancy between the two sexes**. Women live, on average, longer than men in most countries, so your column might be called W-M (Women-Men). Type this column header in E1.

In E2, write the formula once and hit enter:
=C2-B2.

The minus sign is the dash on your keyboard. No function word! (Table 38)

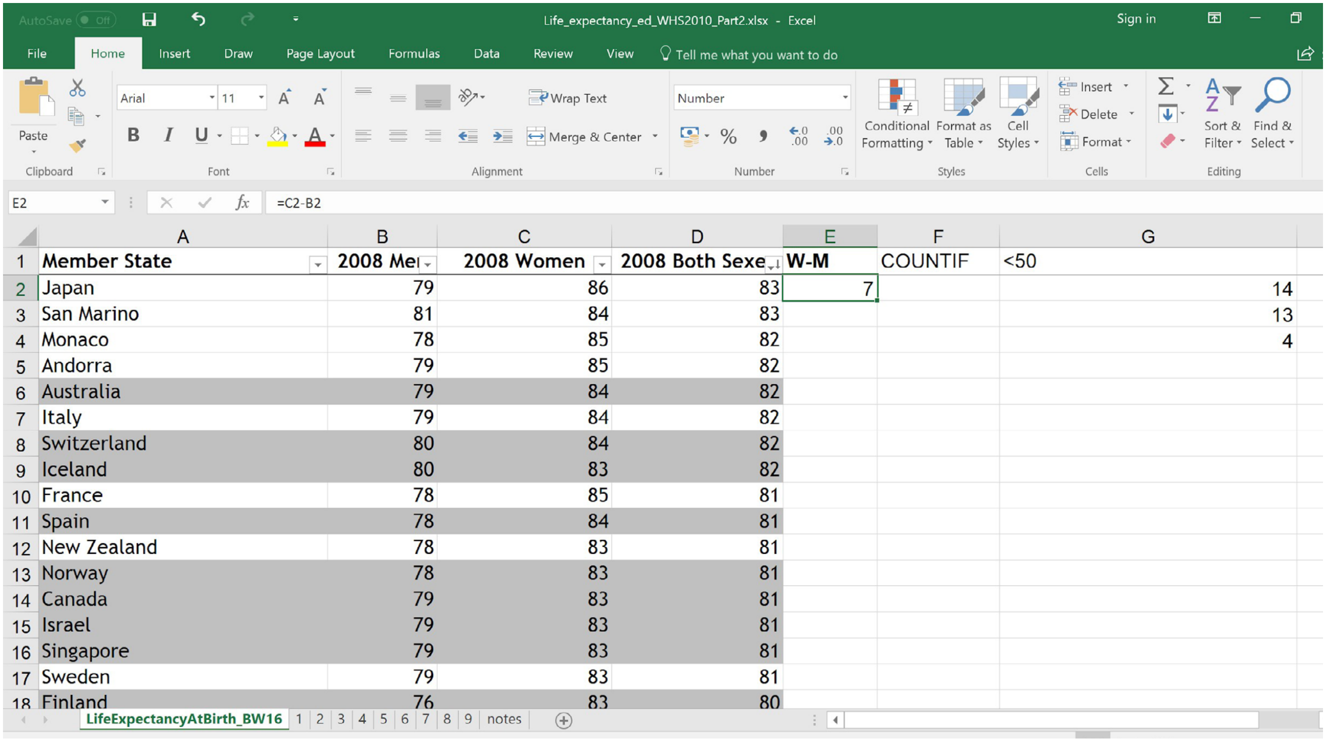


B2 | =C2-B2

	A	B	C	D	E	F	G
1	Member State	2008 Me	2008 Women	2008 Both Sexe	W-M	COUNTIF	<50
2	Japan	79	86	83	=C2-B2		14
3	San Marino	81	84	83			13
4	Monaco	78	85	82			4
5	Andorra	79	85	82			
6	Australia	79	84	82			
7	Italy	79	84	82			
8	Switzerland	80	84	82			
9	Iceland	80	83	82			
10	France	78	85	81			
11	Spain	78	84	81			
12	New Zealand	78	83	81			
13	Norway	78	83	81			
14	Canada	79	83	81			
15	Israel	79	83	81			
16	Singapore	79	83	81			
17	Sweden	79	83	81			
18	Finland	76	83	80			

Table 38

We won’t retype the formula for each row –life is too short! Instead, select E2 – this is a cell that has both a number and the maths behind it. Either click-hold its lower-right corner and drag down to where your data ends; or double-click the selected cell’s lower right corner – where you can see a minuscule square – and the rest of the rows will automatically be populated with the results you’re after. To successfully double-click, place your cursor – now a white, fat cross – on the cell’s lower right corner. When it becomes a slim, black cross, double-click.



E2 =C2-B2

	A	B	C	D	E	F	G
1	Member State	2008 Me	2008 Women	2008 Both Sexe	W-M	COUNTIF	<50
2	Japan	79	86	83	7		14
3	San Marino	81	84	83			13
4	Monaco	78	85	82			4
5	Andorra	79	85	82			
6	Australia	79	84	82			
7	Italy	79	84	82			
8	Switzerland	80	84	82			
9	Iceland	80	83	82			
10	France	78	85	81			
11	Spain	78	84	81			
12	New Zealand	78	83	81			
13	Norway	78	83	81			
14	Canada	79	83	81			
15	Israel	79	83	81			
16	Singapore	79	83	81			
17	Sweden	79	83	81			
18	Finland	76	83	80			

Table 39

Sort your data in multiple ways, to get more colour for your reporting!

Tip: When engaging with the "why" and "how", think about what other datasets you could correlate it with: years of war? Level of education? Affordable decent food? GDP?

SELECTIVE COUNTING

>COUNTIF

Sometimes you want to count very quickly only those records with certain parameters. The quick selective count is done with the help of a command called **COUNTIF**.

Let's find out in how many countries life expectancy is lower than 50 years of age.

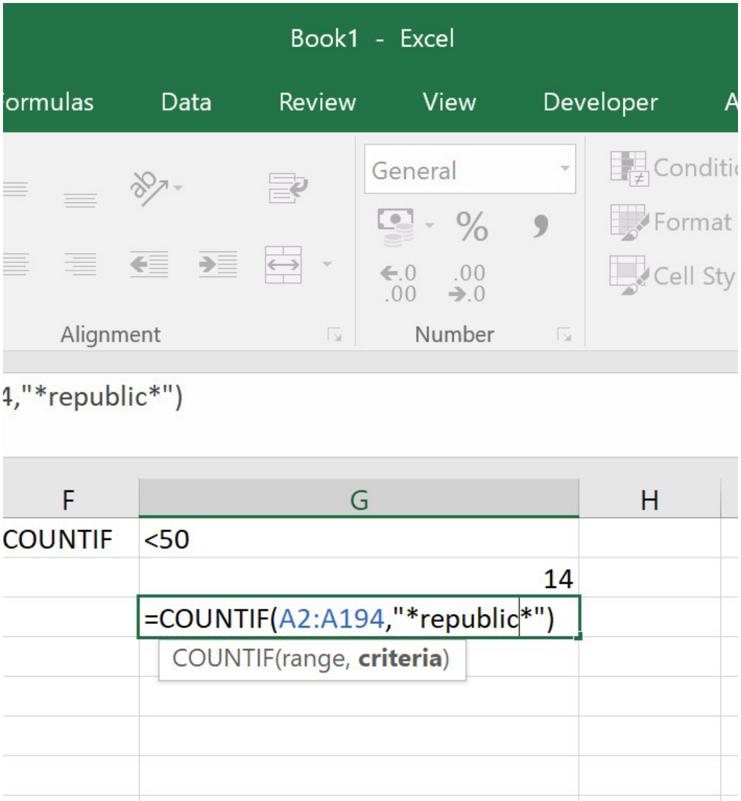
=COUNTIF(D2:D194,"<50")

Result:14

We can also find out how many countries contain the word "Republic" in their name:

=COUNTIF(A2:A194,"*republic*")

Result:13



4,"*republic*")

F	G	H
COUNTIF	<50	
	14	
	=COUNTIF(A2:A194,"*republic*")	

COUNTIF(range, **criteria**)

Figure 2

The asterisk / star is a WILD CARD that is just as useful when analysing data as it is for online searches. In the context of a spreadsheet, it allows us to run imprecise commands based on a condition.

For example, the formula in Figure 2 can be read like this:

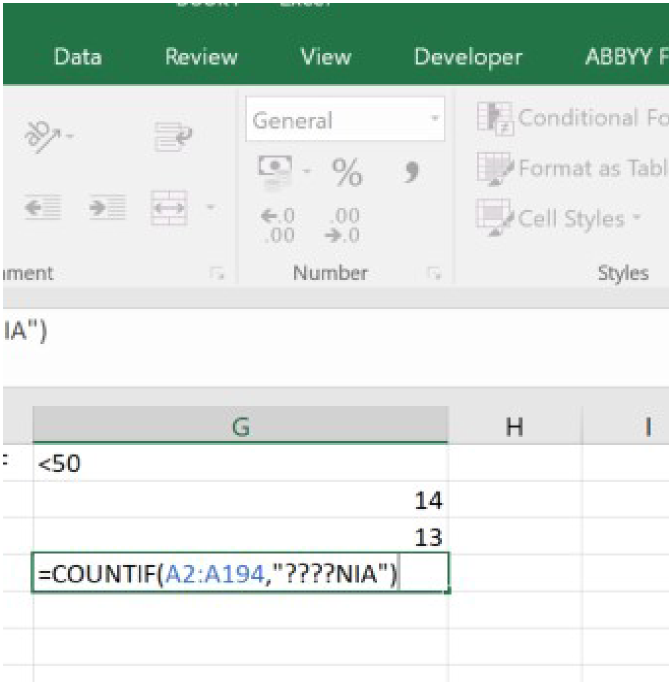
> In the list ranging from A2 to A194, count all the countries that start with anything, end in anything and contain the word "republic" anywhere in their title. This means that the count is tailored to catch any country that features the word "Republic" anywhere in its name.

To add to its level of flexibility, we can also tell Excel to include the number of letters in a condition. We will use "?" to represent a letter.

For example, let's count all those names that end in "NIA" and have seven letters:

=COUNTIF(A2:A194,"????NIA")

RESULT: 4



IA")

G
<50
14
13
=COUNTIF(A2:A194,"????NIA")

Figure 3

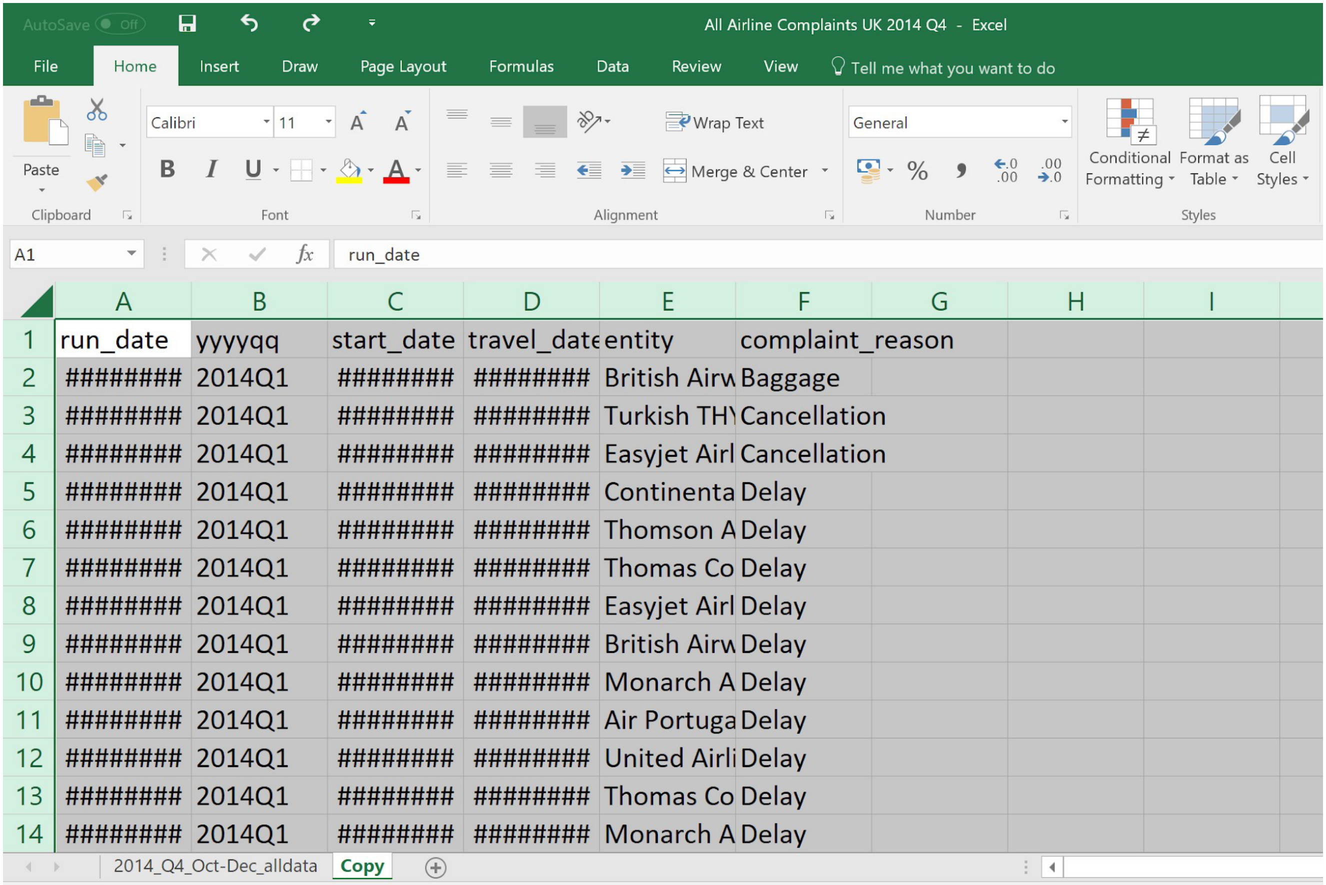
You can get creative with COUNTIF. For a comprehensive list of what the formula has to offer, visit Microsoft's dedicated site: https://support.office.com/en-gb/article/COUNTIF-function-e0de10c6-f885-4e71-abb4-1f464816df34.

DIGGING DEEPER WITH PIVOT TABLES

The Pivot Table is a most useful tool that allows for deeper data mining and interrogation. It is a very popular data interviewing tool. It can aggregate, filter, custom-sort and compute, all in one go! Download the file "All Airline Complaints UK 2014 Q4" by clicking here.

Let's open a dataset containing complaints filed with airlines by customers flying from the UK anywhere in the world (Airline Complaints dataset).

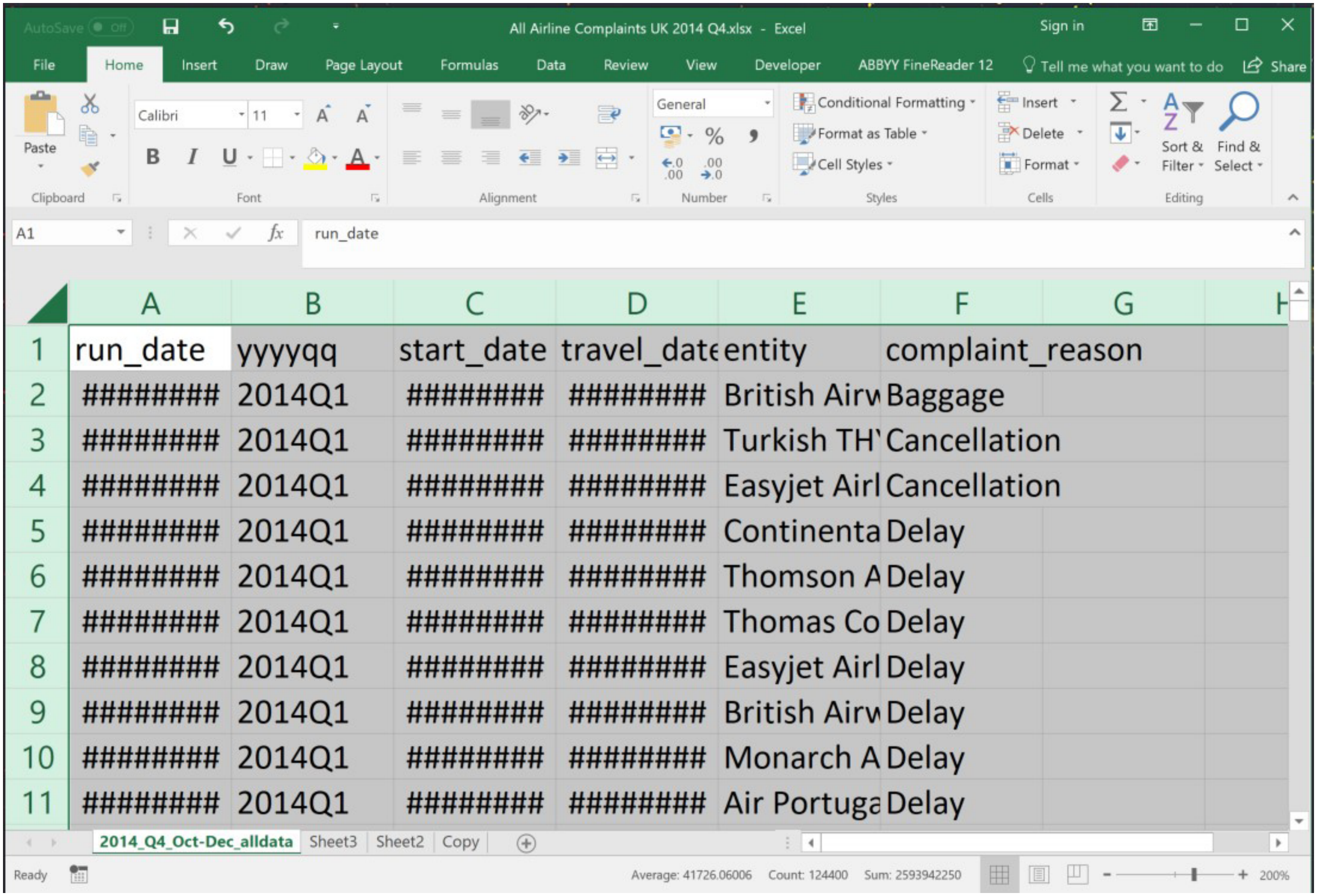
When data in a column take up more space than initially available, the information displayed might look like a row of hashtags (#) (Table 40). This is not an error or dirty data. To see the information, simply expand the column width.



	A	B	C	D	E	F	G	H	I
1	run_date	yyyyqq	start_date	travel_date	entity	complaint_reason			
2	########	2014Q1	########	########	British Airw	Baggage			
3	########	2014Q1	########	########	Turkish THY	Cancellation			
4	########	2014Q1	########	########	Easyjet Airl	Cancellation			
5	########	2014Q1	########	########	Continenta	Delay			
6	########	2014Q1	########	########	Thomson A	Delay			
7	########	2014Q1	########	########	Thomas Co	Delay			
8	########	2014Q1	########	########	Easyjet Airl	Delay			
9	########	2014Q1	########	########	British Airw	Delay			
10	########	2014Q1	########	########	Monarch A	Delay			
11	########	2014Q1	########	########	Air Portuga	Delay			
12	########	2014Q1	########	########	United Airli	Delay			
13	########	2014Q1	########	########	Thomas Co	Delay			
14	########	2014Q1	########	########	Monarch A	Delay			

Table 40

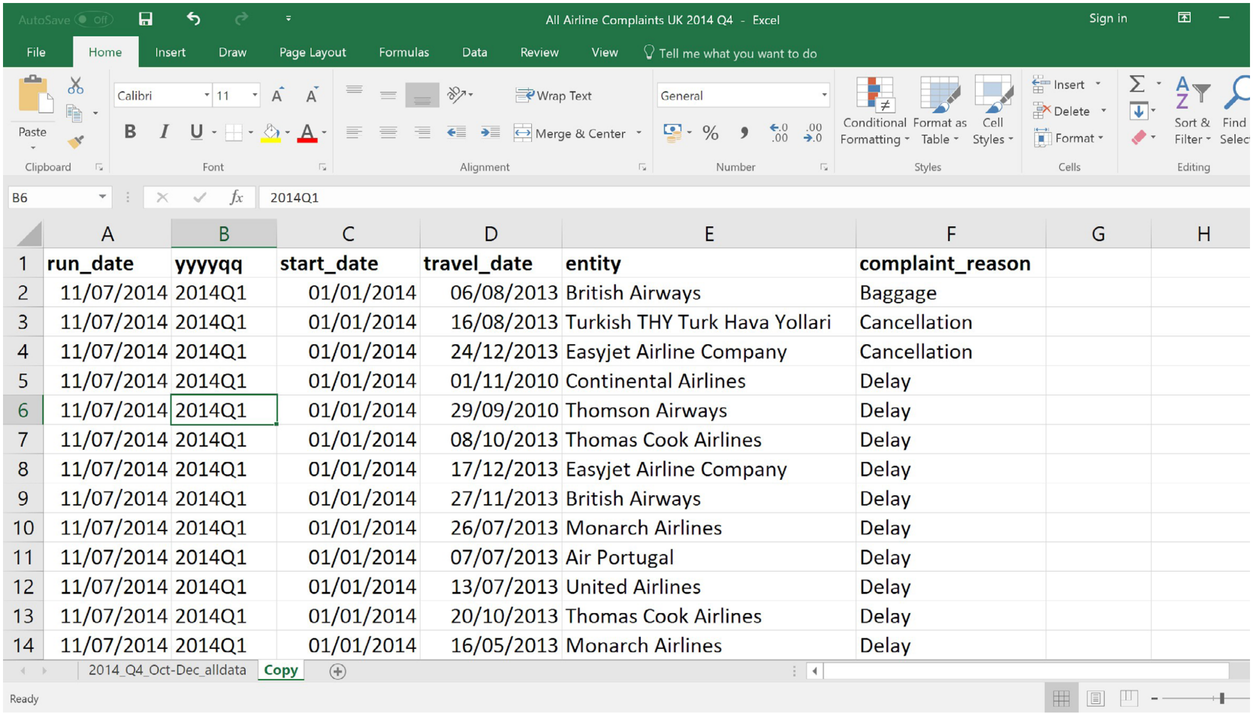
At the bottom of the page, next to the sheet name, you will find an icon (a plus in a circle or a star on a spreadsheet). Click on it to create a new, blank sheet. Select all the data on your original sheet by clicking on the top-left rectangle.



	A	B	C	D	E	F	G
1	run_date	yyyyqq	start_date	travel_date	entity	complaint_reason	
2	########	2014Q1	########	########	British Airw	Baggage	
3	########	2014Q1	########	########	Turkish TH'	Cancellation	
4	########	2014Q1	########	########	Easyjet Airl	Cancellation	
5	########	2014Q1	########	########	Continenta	Delay	
6	########	2014Q1	########	########	Thomson A	Delay	
7	########	2014Q1	########	########	Thomas Co	Delay	
8	########	2014Q1	########	########	Easyjet Airl	Delay	
9	########	2014Q1	########	########	British Airw	Delay	
10	########	2014Q1	########	########	Monarch A	Delay	
11	########	2014Q1	########	########	Air Portuga	Delay	

Table 41

Like in a word document, CTRL+C / Command +C to copy the data. In the new sheet, preferably in A1, follow with CTRL+V / Command +V. Enlarge columns by dragging the border lines between two column letters left or right. Work on the copy, not on the original.



	A	B	C	D	E	F
1	run_date	yyyyqq	start_date	travel_date	entity	complaint_reason
2	11/07/2014	2014Q1	01/01/2014	06/08/2013	British Airways	Baggage
3	11/07/2014	2014Q1	01/01/2014	16/08/2013	Turkish THY Turk Hava Yollari	Cancellation
4	11/07/2014	2014Q1	01/01/2014	24/12/2013	Easyjet Airline Company	Cancellation
5	11/07/2014	2014Q1	01/01/2014	01/11/2010	Continental Airlines	Delay
6	11/07/2014	2014Q1	01/01/2014	29/09/2010	Thomson Airways	Delay
7	11/07/2014	2014Q1	01/01/2014	08/10/2013	Thomas Cook Airlines	Delay
8	11/07/2014	2014Q1	01/01/2014	17/12/2013	Easyjet Airline Company	Delay
9	11/07/2014	2014Q1	01/01/2014	27/11/2013	British Airways	Delay
10	11/07/2014	2014Q1	01/01/2014	26/07/2013	Monarch Airlines	Delay
11	11/07/2014	2014Q1	01/01/2014	07/07/2013	Air Portugal	Delay
12	11/07/2014	2014Q1	01/01/2014	13/07/2013	United Airlines	Delay
13	11/07/2014	2014Q1	01/01/2014	20/10/2013	Thomas Cook Airlines	Delay
14	11/07/2014	2014Q1	01/01/2014	16/05/2013	Monarch Airlines	Delay

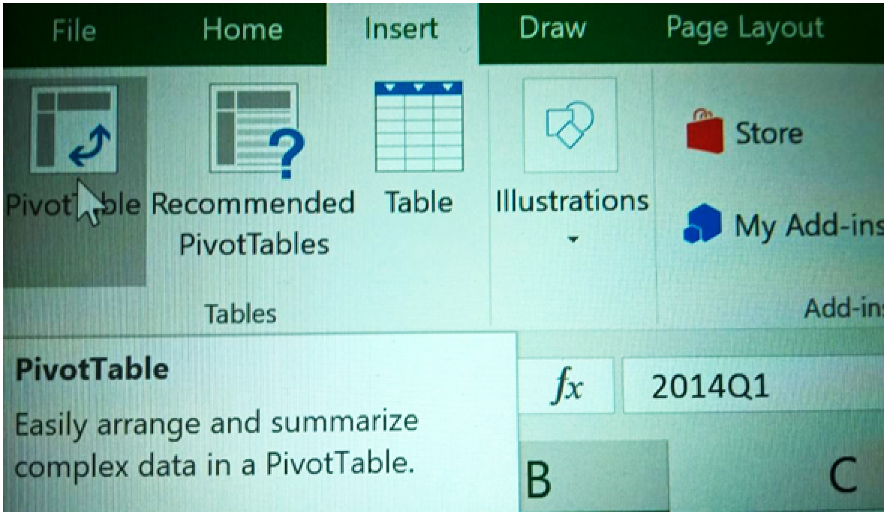
Table 42: all text fits in cells and is visible

We have thousands of records openly available from UK's Civil Aviation Authority (CAA). You'll need to become familiar with the content of each column in order to ask smart questions. Use filters to eyeball the information in each column, before proceeding to your digital interview.

This source can tell you about the number of complaints for each airline, what people complain about, when the flight took place and what kind of complaints have been filed.

To create a PIVOT TABLE:

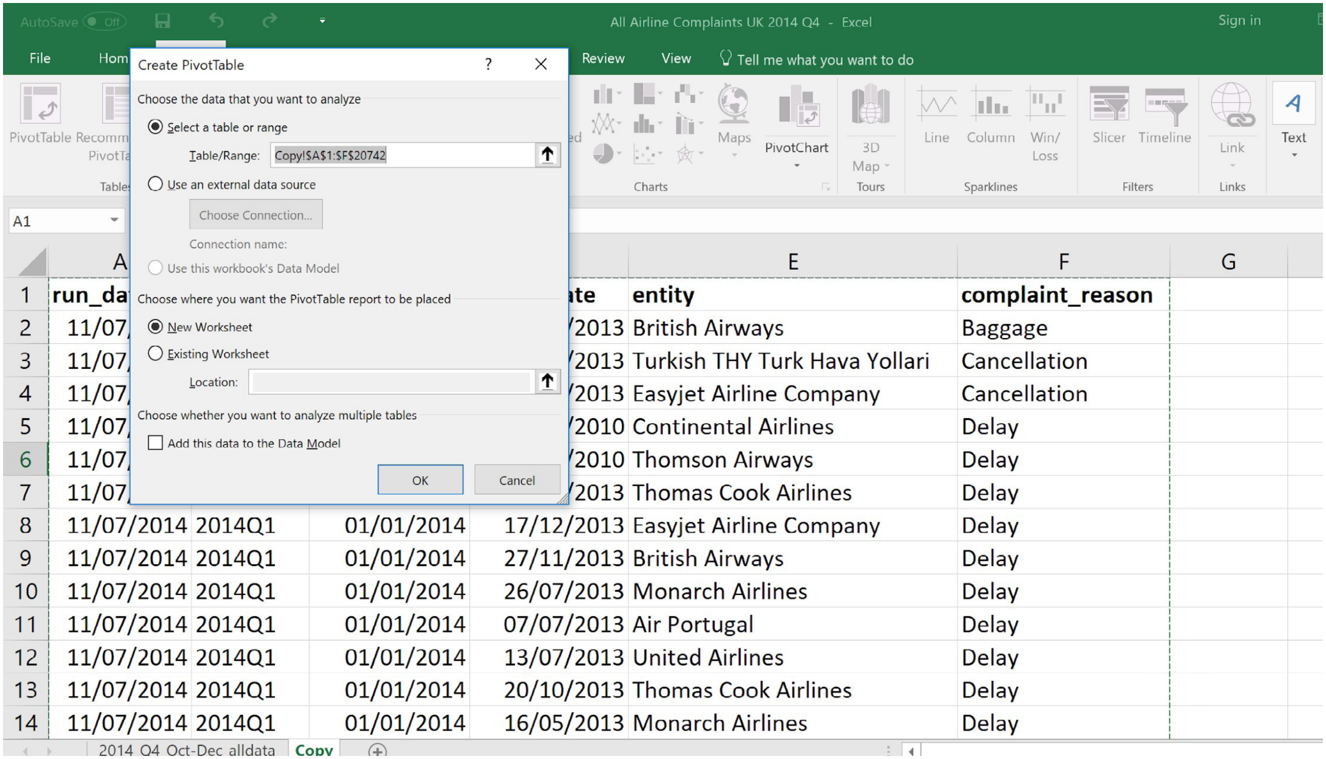
- Select all your data (CTRL+A / COMMAND+A)
- Click on the INSERT tab
- Select Pivot Table by clicking on the button (in the older Excel versions for MACs, this will be under DATA)
- Ensure the PivotTable on a New Worksheet is selected
- Click OK.



PivotTable

Easily arrange and summarize complex data in a PivotTable.

Figure 4: Insert pivot table 1/3



Create PivotTable

Choose the data that you want to analyze

● Select a table or range

Table/Range: Copy!$A$1:$F$20742

○ Use an external data source

Choose Connection...

Connection name:

○ Use this workbook's Data Model

Choose where you want the PivotTable report to be placed

● New Worksheet

○ Existing Worksheet

Location:

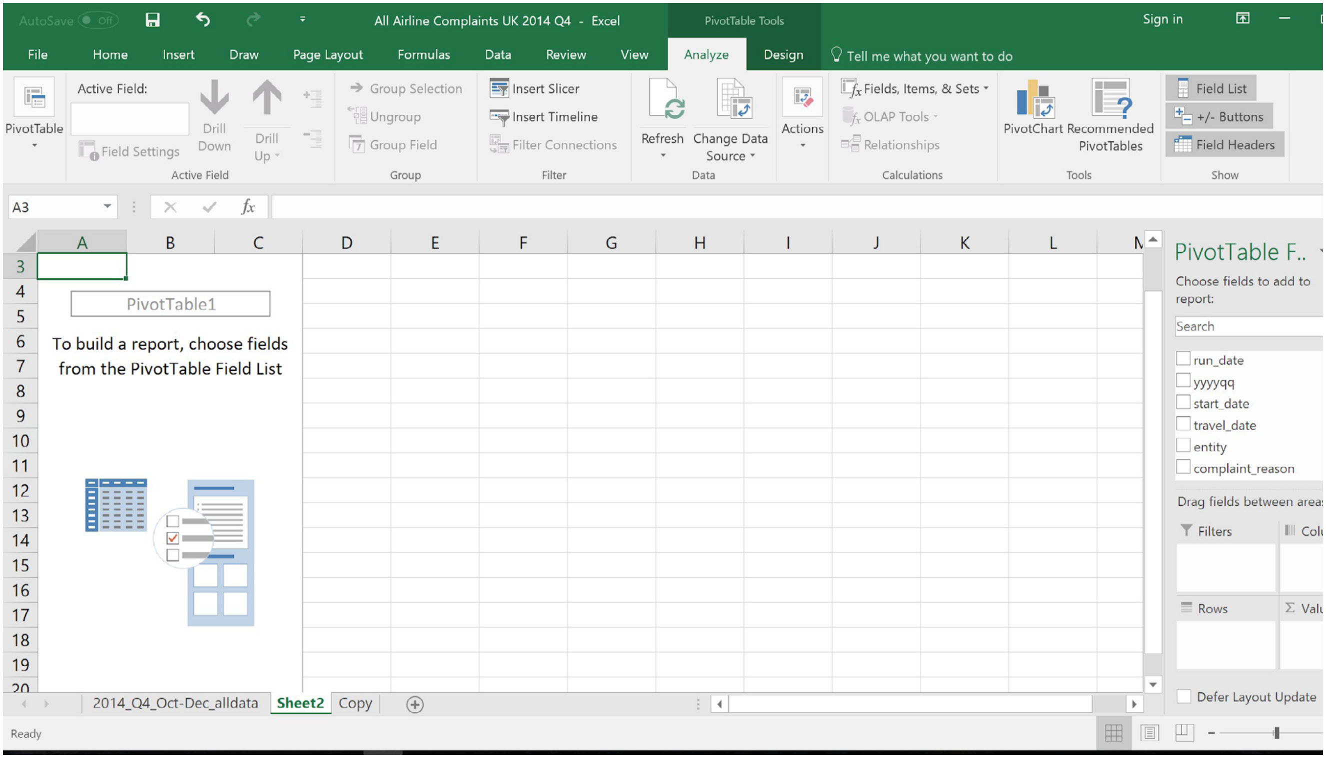
Choose whether you want to analyze multiple tables

□ Add this data to the Data Model

OK | Cancel

	A				E	F	G
1	run_da			te	entity	complaint_reason	
2	11/07,			/2013	British Airways	Baggage	
3	11/07,			/2013	Turkish THY Turk Hava Yollari	Cancellation	
4	11/07,			/2013	Easyjet Airline Company	Cancellation	
5	11/07,			/2010	Continental Airlines	Delay	
6	11/07,			/2010	Thomson Airways	Delay	
7	11/07,			/2013	Thomas Cook Airlines	Delay	
8	11/07/2014	2014Q1	01/01/2014	17/12/2013	Easyjet Airline Company	Delay	
9	11/07/2014	2014Q1	01/01/2014	27/11/2013	British Airways	Delay	
10	11/07/2014	2014Q1	01/01/2014	26/07/2013	Monarch Airlines	Delay	
11	11/07/2014	2014Q1	01/01/2014	07/07/2013	Air Portugal	Delay	
12	11/07/2014	2014Q1	01/01/2014	13/07/2013	United Airlines	Delay	
13	11/07/2014	2014Q1	01/01/2014	20/10/2013	Thomas Cook Airlines	Delay	
14	11/07/2014	2014Q1	01/01/2014	16/05/2013	Monarch Airlines	Delay	

Table 43: Pivot Table 2/3



To build a report, choose fields from the PivotTable Field List

Table 44: Pivot Table space 3/3

The data are now hidden from sight (but still available in your copy). A list with the column headers can be seen on the right. There are also four rectangles in the right-side working space: one to create filters; one to display the information as rows, one to display the data as columns and one to do the maths and counting. You can click and drag a column header from the listing above into one or more rectangles to display a column's content.

To get rid of the selection, simply untick the column names in the list.

If your working space disappears in the process of working in the pivot table, click in the pivot table area (spreadsheet, left) to bring it all back.

Any data source should supply a **Data Dictionary** or metadata or record layout alongside any database released.

This document should contain an explanation of the terms and codes, the size of the data, the types of data, the number of columns and rows, how the data were collected, or if something was redacted out of the data for any reason.

Most of the time, you would have to do the chasing for this document. And more often than not, it will be incomplete.

Get to know your source by listing the content of each column. To do so, click and drag a column header under the rows section in the Pivot Table working space (Table 44) one by one.

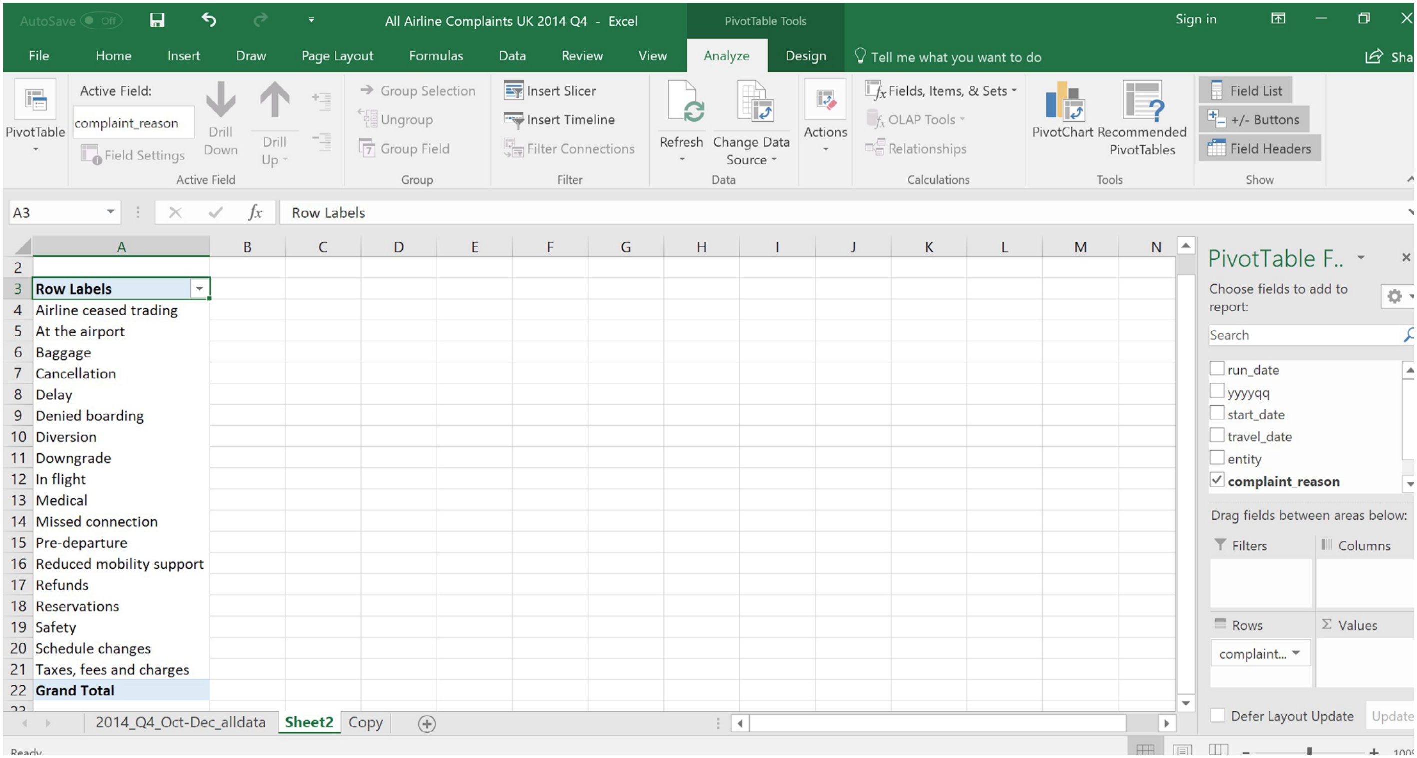
Make notes of anything that's unclear, mark duplicates, dirty data or vague information. Data are usually dirty or incomplete. This means that you will have to do more reporting and cleaning before you can run a data analysis undisturbed.

Tip: Go back to the data source and ask them about the gaps in the data. Do not leave significant blanks unresolved – sometimes they are the story!

Ask:

- how are the data collected? Is there an online form or a paper form?
- how are the data filed and when?
- who is responsible for the database's integrity?
- are state agencies allowed to simply not report data when they wish to hide something? Are there any reviews / audits / penalties in place?
- how often are the data filed? When is the next round due? Who checks the data when they are filed?
- how is the record keeping working vs. how is it supposed to work?
- where is the data dictionary? does it contain an explanation of all the terms? What do all the terms and codes mean?
- do you need to speak with a scientist / expert / inside source to understand how anchored in reality the data are? Or what they leave out?
- You may need to file press or Freedom of Information request to fill the gaps in the data.
- Always report missing data, because it skews the results. But do not use this as an excuse not to do your best to investigate and fill in the gaps.

In the words of the late legendary David Donald, Data Editor, professor and CAR trainer extraordinaire, before you do anything, KNOW THY DATA!



	A
2	
3	**Row Labels**
4	Airline ceased trading
5	At the airport
6	Baggage
7	Cancellation
8	Delay
9	Denied boarding
10	Diversion
11	Downgrade
12	In flight
13	Medical
14	Missed connection
15	Pre-departure
16	Reduced mobility support
17	Refunds
18	Reservations
19	Safety
20	Schedule changes
21	Taxes, fees and charges
22	**Grand Total**

Table 44

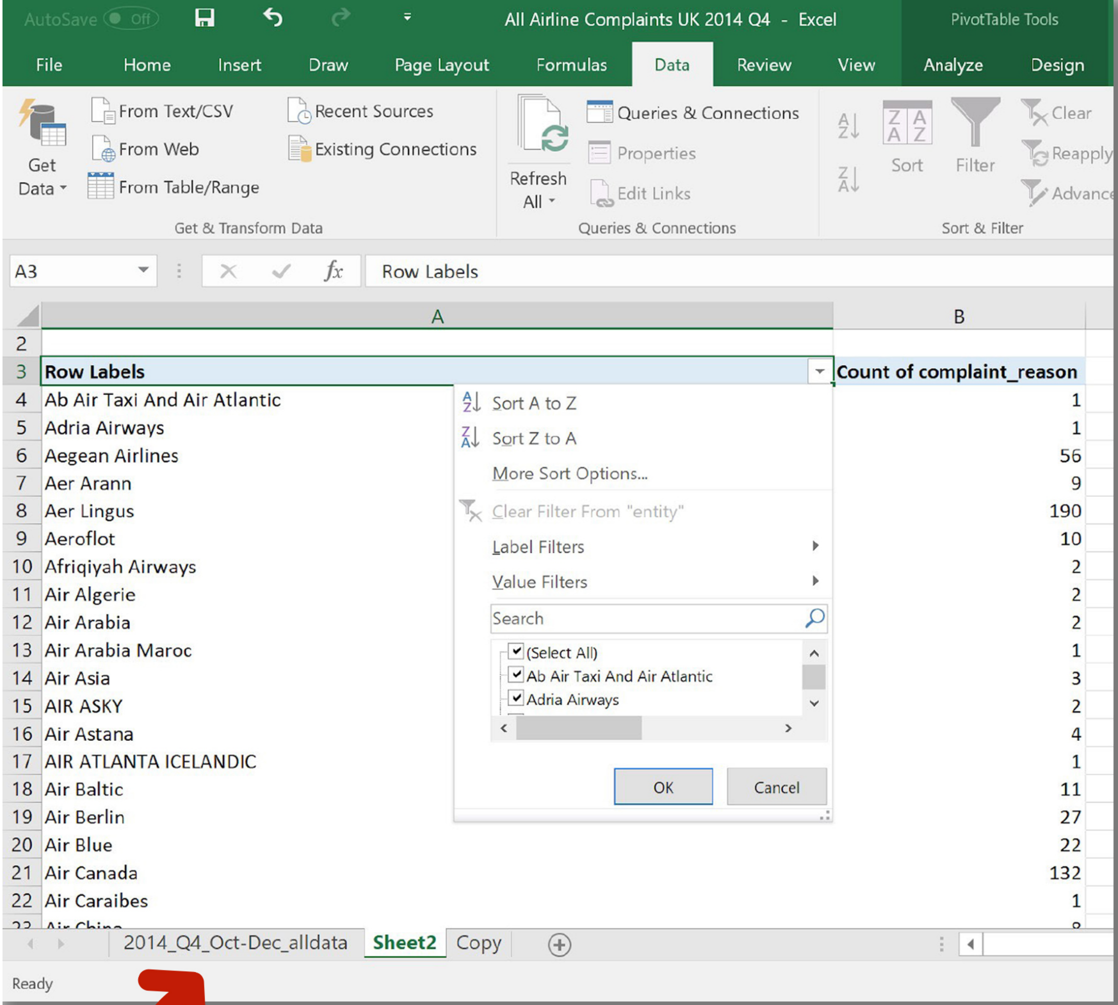
Here are a few questions you could ask this source:
Q: Which airline received the most complaints?

To answer, you would need to have ENTITIES in rows and complaint reason in values, where they will be counted.

To list the findings, from largest to smallest count of complaints, click in the new table > Row Labels drop-down > More Sort Options > arrange the data Descending > by Count of complaint reason (Tables 46 and 47). This gives you the listing of the complaint counts per airline.

When you report, you must compare apples to apples. Raw numbers are not good enough for this. You will need to find out the number of flights for the period covered by the data, the number of passengers and calculate the annual complaint per passenger.

You should also take into account that the number of complaints may not be all about quality of service but may also reflect cultural or practical differences between countries or airlines.



Row Labels	Count of complaint_reason
Ab Air Taxi And Air Atlantic	1
Adria Airways	1
Aegean Airlines	56
Aer Arann	9
Aer Lingus	190
Aeroflot	10
Afriqiyah Airways	2
Air Algerie	2
Air Arabia	2
Air Arabia Maroc	1
Air Asia	3
AIR ASKY	2
Air Astana	4
AIR ATLANTA ICELANDIC	1
Air Baltic	11
Air Berlin	27
Air Blue	22
Air Canada	132
Air Caraibes	1

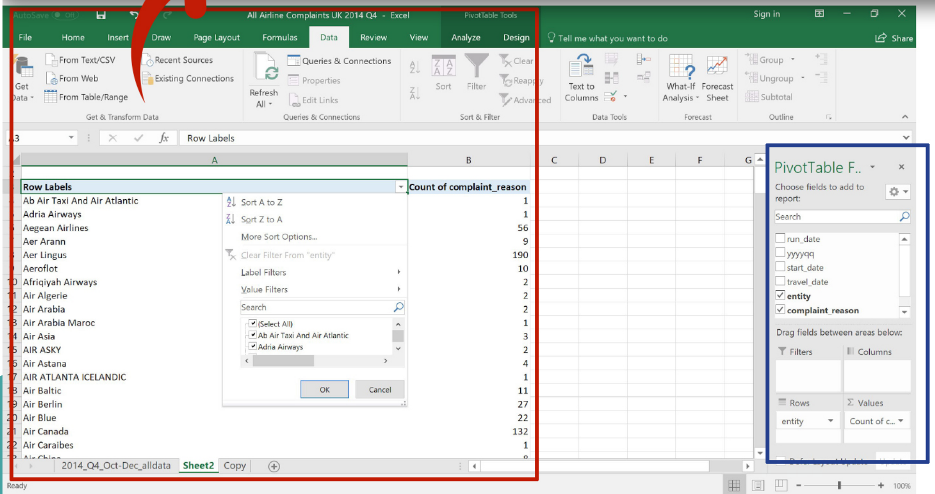
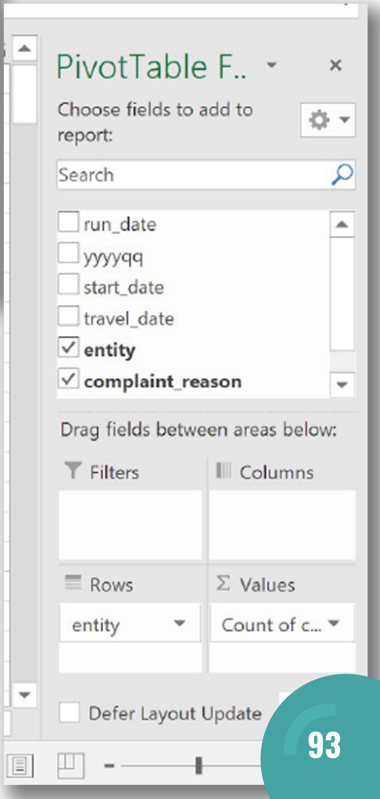
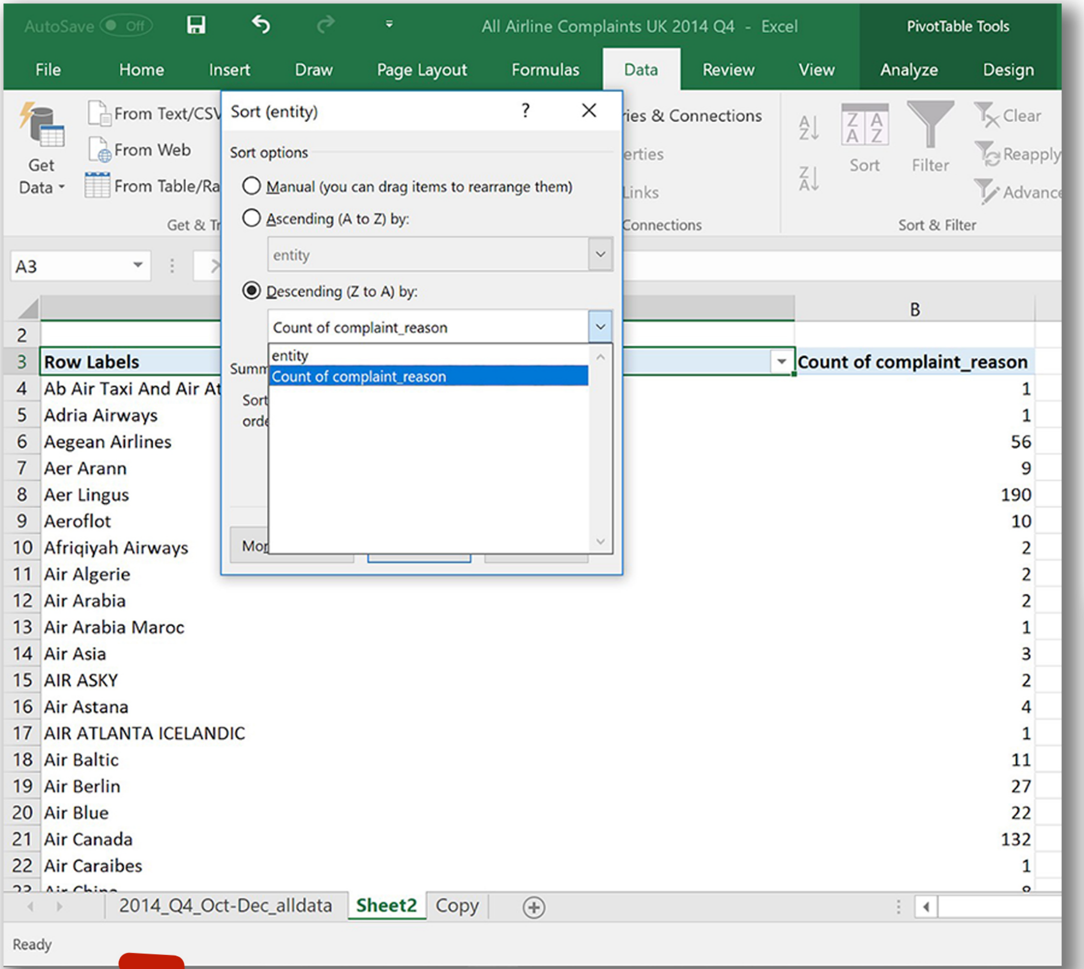


Table 45:
More sort options





Sort (entity)

Sort options

○ Manual (you can drag items to rearrange them)

○ Ascending (A to Z) by:

entity

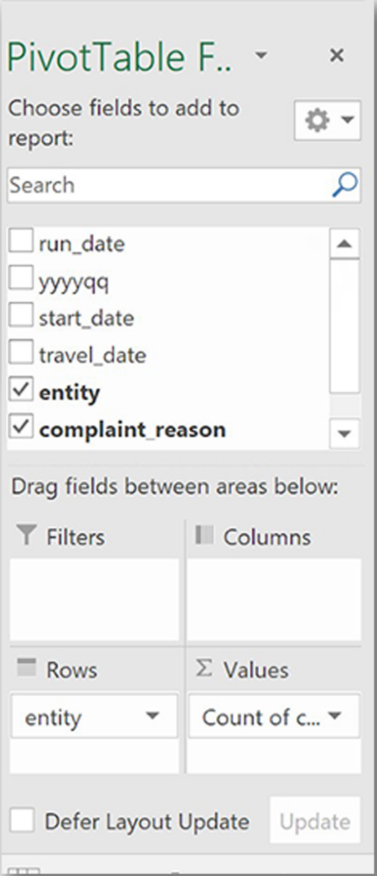
● Descending (Z to A) by:

Count of complaint_reason

- entity
- Count of complaint_reason

	A	B
2		
3	Row Labels	Count of complaint_reason
4	Ab Air Taxi And Air At	1
5	Adria Airways	1
6	Aegean Airlines	56
7	Aer Arann	9
8	Aer Lingus	190
9	Aeroflot	10
10	Afriqiyah Airways	2
11	Air Algerie	2
12	Air Arabia	2
13	Air Arabia Maroc	1
14	Air Asia	3
15	AIR ASKY	2
16	Air Astana	4
17	AIR ATLANTA ICELANDIC	1
18	Air Baltic	11
19	Air Berlin	27
20	Air Blue	22
21	Air Canada	132
22	Air Caraibes	1

Table 46:
count of complaint reason



PivotTable F..

Choose fields to add to report:

Search

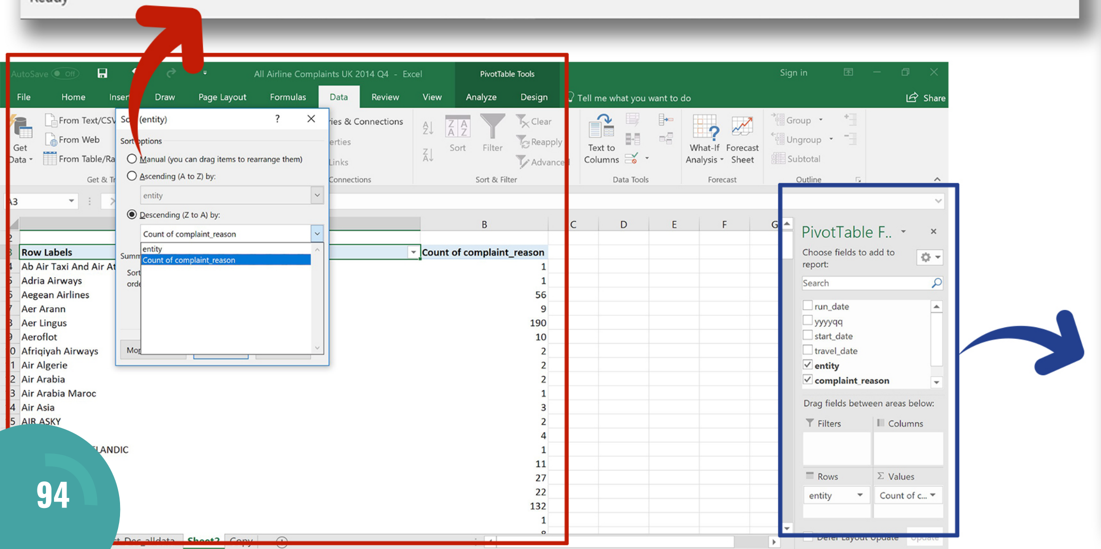
- □ run_date
- □ yyyyqq
- □ start_date
- □ travel_date
- ☑ entity
- ☑ complaint_reason

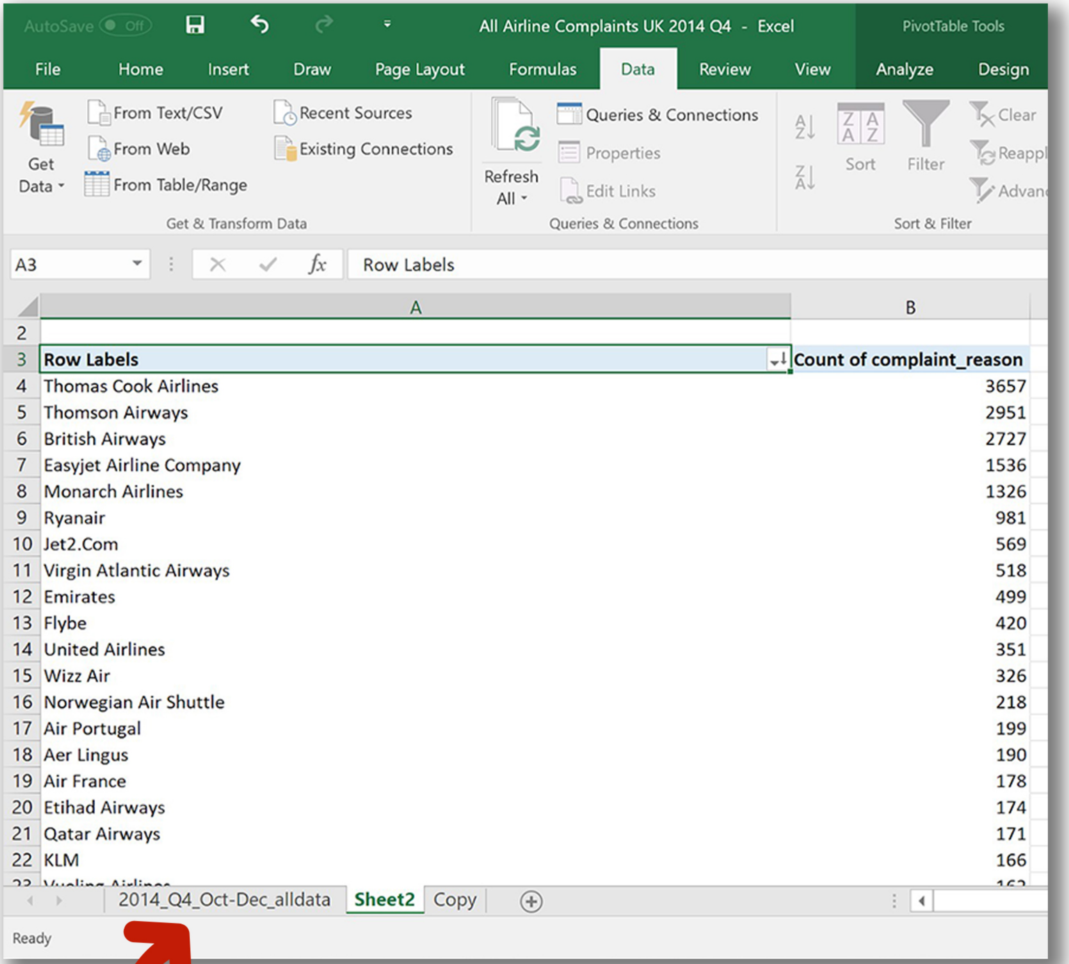
Drag fields between areas below:

Filters	Columns

Rows	Values
entity	Count of c...

□ Defer Layout Update

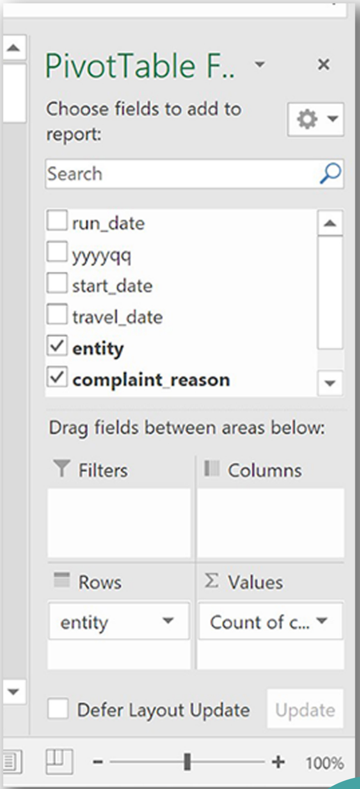




A3 Row Labels

	Row Labels	Count of complaint_reason
4	Thomas Cook Airlines	3657
5	Thomson Airways	2951
6	British Airways	2727
7	Easyjet Airline Company	1536
8	Monarch Airlines	1326
9	Ryanair	981
10	Jet2.Com	569
11	Virgin Atlantic Airways	518
12	Emirates	499
13	Flybe	420
14	United Airlines	351
15	Wizz Air	326
16	Norwegian Air Shuttle	218
17	Air Portugal	199
18	Aer Lingus	190
19	Air France	178
20	Etihad Airways	174
21	Qatar Airways	171
22	KLM	166

Table 47: first pivot


PivotTable F..

Choose fields to add to report:

Search

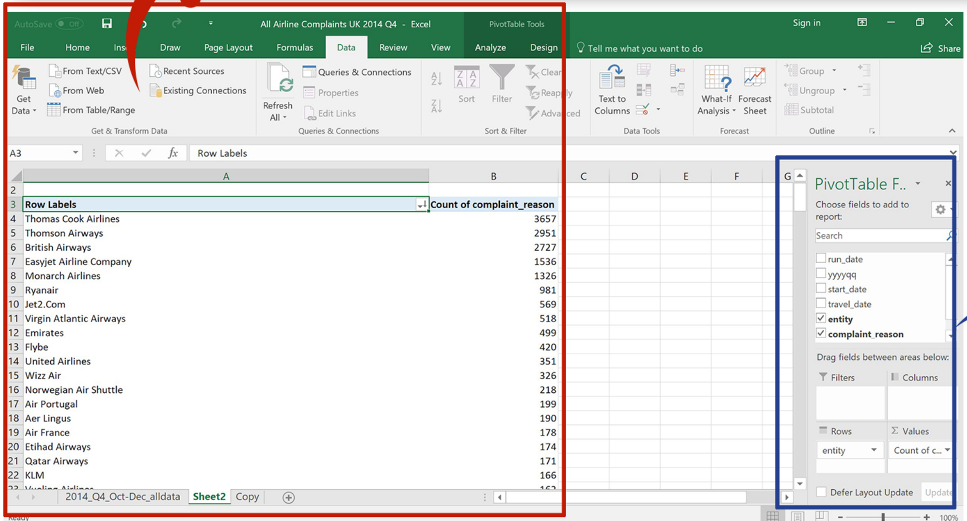
- ☐ run_date
- ☐ yyyyqq
- ☐ start_date
- ☐ travel_date
- ☑ **entity**
- ☑ **complaint_reason**

Drag fields between areas below:

Filters	Columns

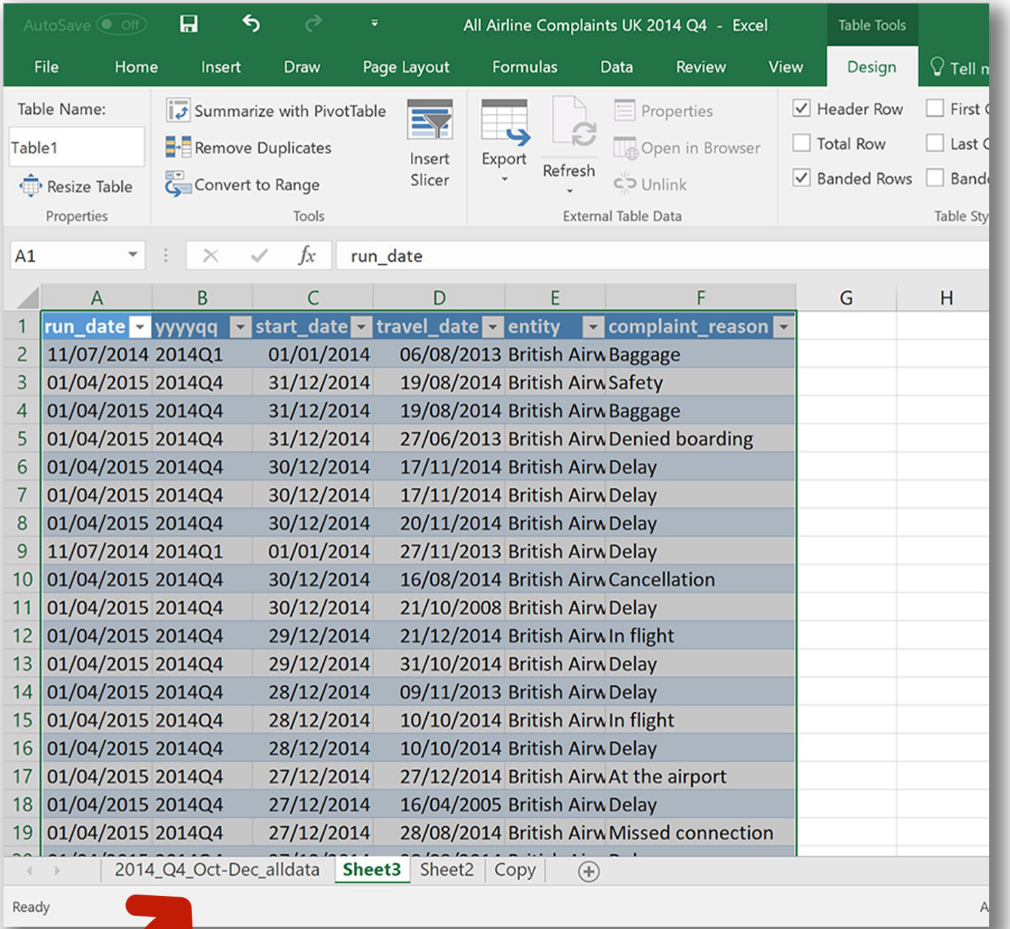
Rows	Values
entity	Count of c...

☐ Defer Layout Update



A3 Row Labels

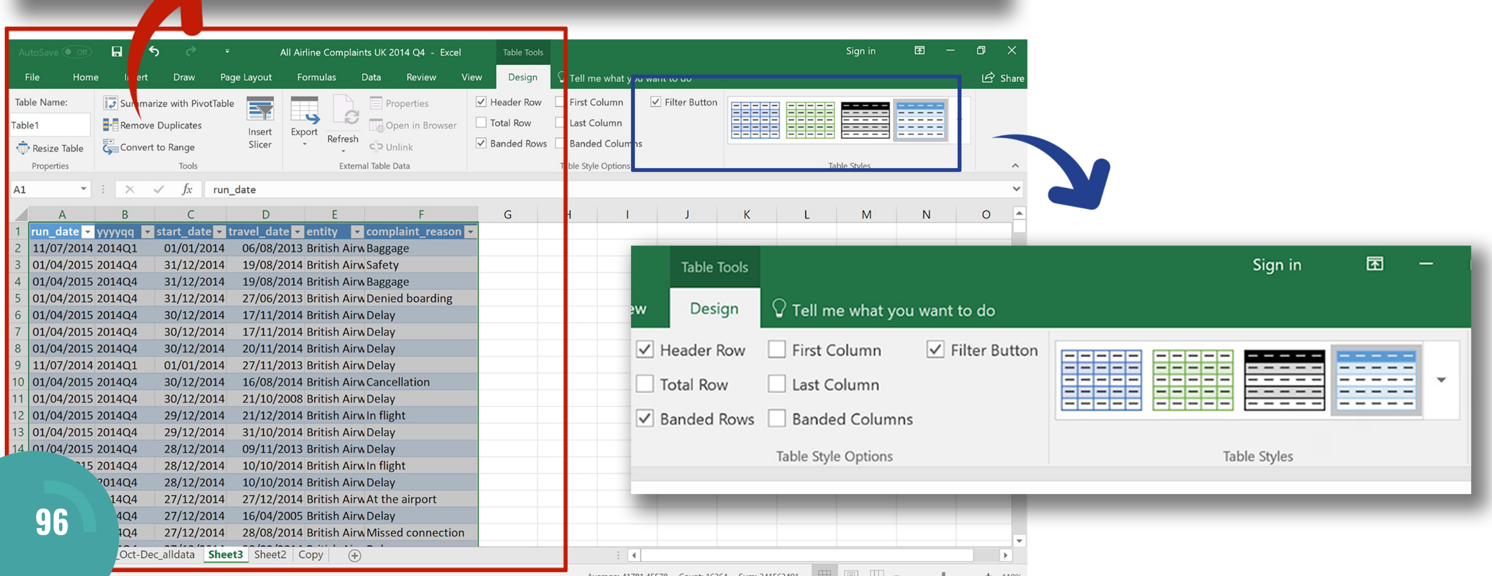
	Row Labels	Count of complaint_reason
4	Thomas Cook Airlines	3657
5	Thomson Airways	2951
6	British Airways	2727
7	Easyjet Airline Company	1536
8	Monarch Airlines	1326
9	Ryanair	981
10	Jet2.Com	569
11	Virgin Atlantic Airways	518
12	Emirates	499
13	Flybe	420
14	United Airlines	351
15	Wizz Air	326
16	Norwegian Air Shuttle	218
17	Air Portugal	199
18	Aer Lingus	190
19	Air France	178
20	Etihad Airways	174
21	Qatar Airways	171
22	KLM	166



	A	B	C	D	E	F
1	run_date	yyyyqq	start_date	travel_date	entity	complaint_reason
2	11/07/2014	2014Q1	01/01/2014	06/08/2013	British Airw	Baggage
3	01/04/2015	2014Q4	31/12/2014	19/08/2014	British Airw	Safety
4	01/04/2015	2014Q4	31/12/2014	19/08/2014	British Airw	Baggage
5	01/04/2015	2014Q4	31/12/2014	27/06/2013	British Airw	Denied boarding
6	01/04/2015	2014Q4	30/12/2014	17/11/2014	British Airw	Delay
7	01/04/2015	2014Q4	30/12/2014	17/11/2014	British Airw	Delay
8	01/04/2015	2014Q4	30/12/2014	20/11/2014	British Airw	Delay
9	11/07/2014	2014Q1	01/01/2014	27/11/2013	British Airw	Delay
10	01/04/2015	2014Q4	30/12/2014	16/08/2014	British Airw	Cancellation
11	01/04/2015	2014Q4	30/12/2014	21/10/2008	British Airw	Delay
12	01/04/2015	2014Q4	29/12/2014	21/12/2014	British Airw	In flight
13	01/04/2015	2014Q4	29/12/2014	31/10/2014	British Airw	Delay
14	01/04/2015	2014Q4	28/12/2014	09/11/2013	British Airw	Delay
15	01/04/2015	2014Q4	28/12/2014	10/10/2014	British Airw	In flight
16	01/04/2015	2014Q4	28/12/2014	10/10/2014	British Airw	Delay
17	01/04/2015	2014Q4	27/12/2014	27/12/2014	British Airw	At the airport
18	01/04/2015	2014Q4	27/12/2014	16/04/2005	British Airw	Delay
19	01/04/2015	2014Q4	27/12/2014	28/08/2014	British Airw	Missed connection

Now that you have your list, if you're curious about one airline in particular – e.g. British Airways – double-click on the number next to the name. This will open a sheet with all the complaints received by the airline in question (Table 48).

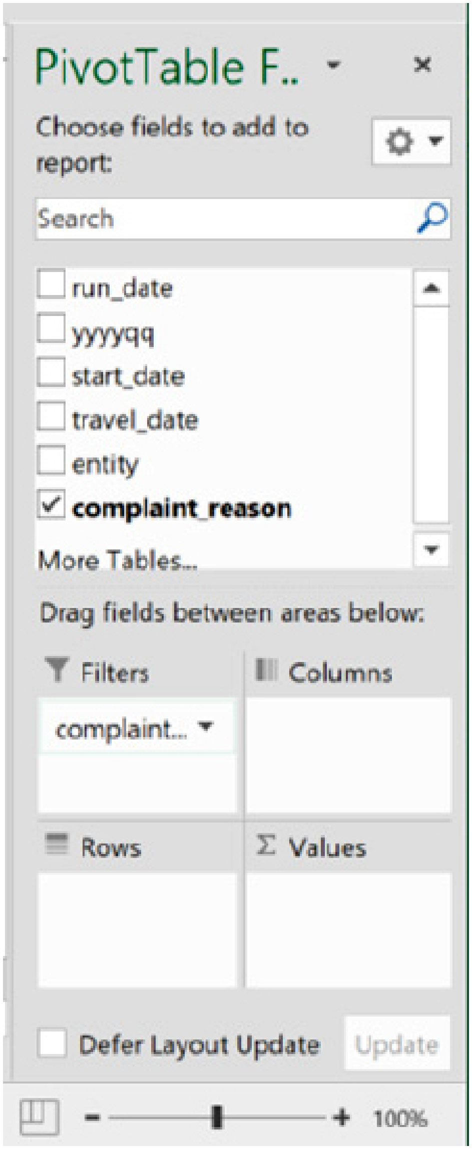
Table 48



☑ Header Row
☐ Total Row
☑ Banded Rows
☐ First Column
☐ Last Column
☐ Banded Columns
☑ Filter Button

Q: How many complaints were there about safety, who got them and when?
When you begin answering a question, start small.

Create a filter with the complaint reason



PivotTable F..

Choose fields to add to report:

Search

□ run_date
□ yyyyqq
□ start_date
□ travel_date
□ entity
☑ **complaint_reason**

More Tables...

Drag fields between areas below:

Filters: complaint...

Columns

Rows

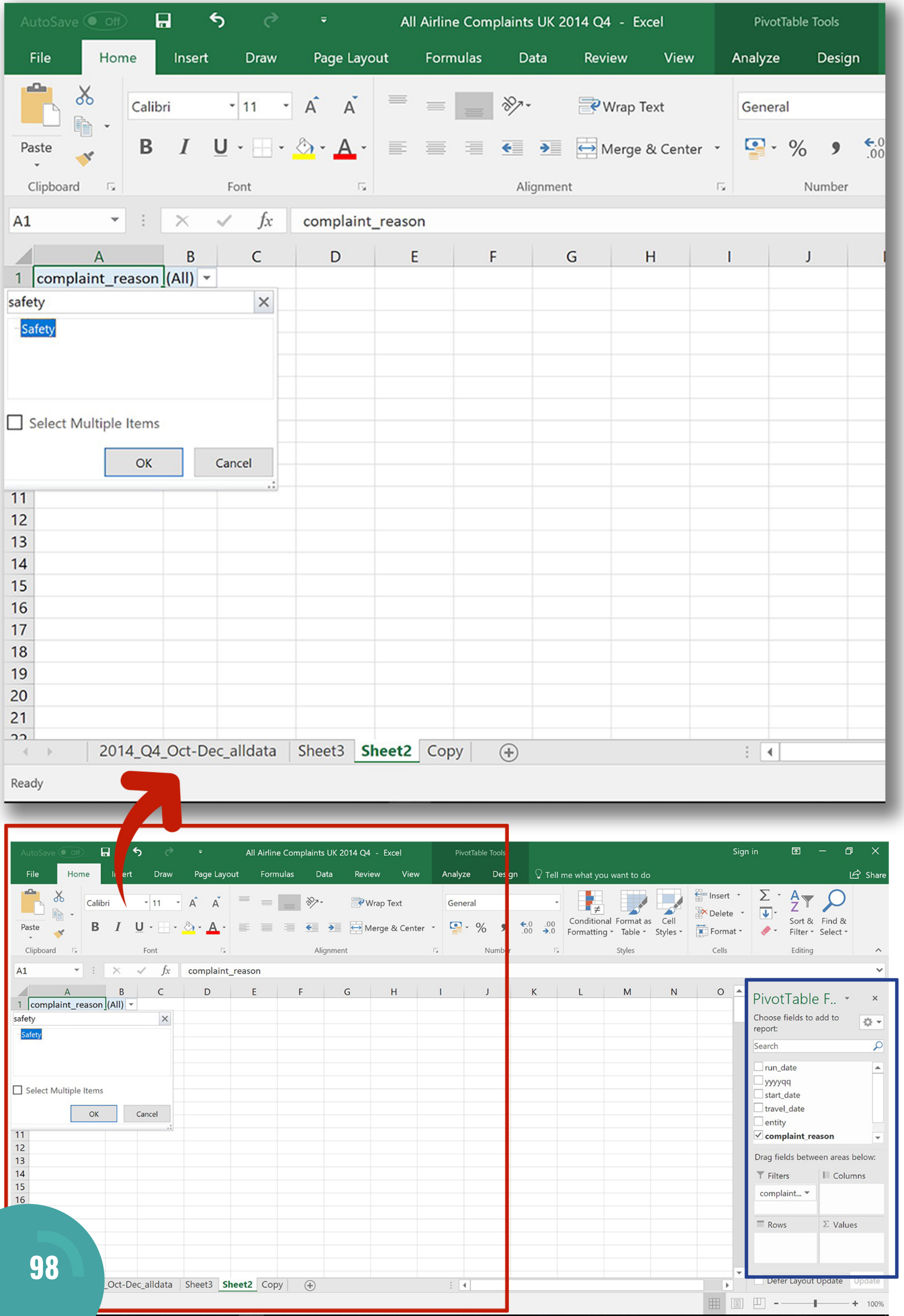
Values

□ Defer Layout Update

Figure 5

On the sheet, open the drop-down button near "complaint reason" and select just "safety". Any results that will appear will be related to this "complaint reason".

Table 49



All Airline Complaints UK 2014 Q4 - Excel

A1 fx complaint_reason

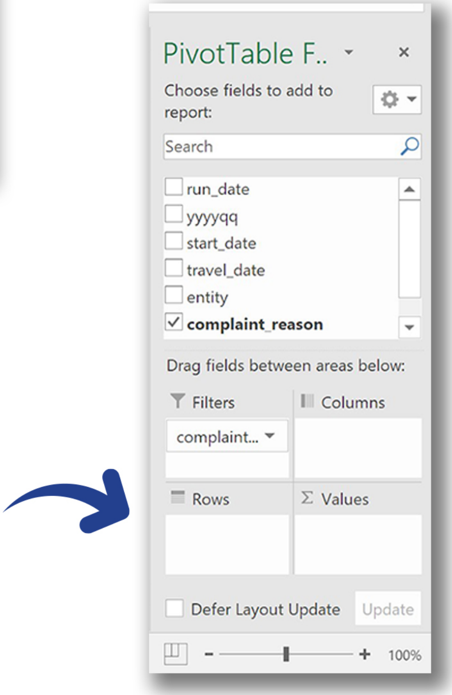
1 complaint_reason (All)

safety

Safety

□ Select Multiple Items

OK Cancel



PivotTable F..

Choose fields to add to report:

Search

□ run_date
□ yyyyqq
□ start_date
□ travel_date
□ entity
☑ complaint_reason

Drag fields between areas below:

Filters: complaint...

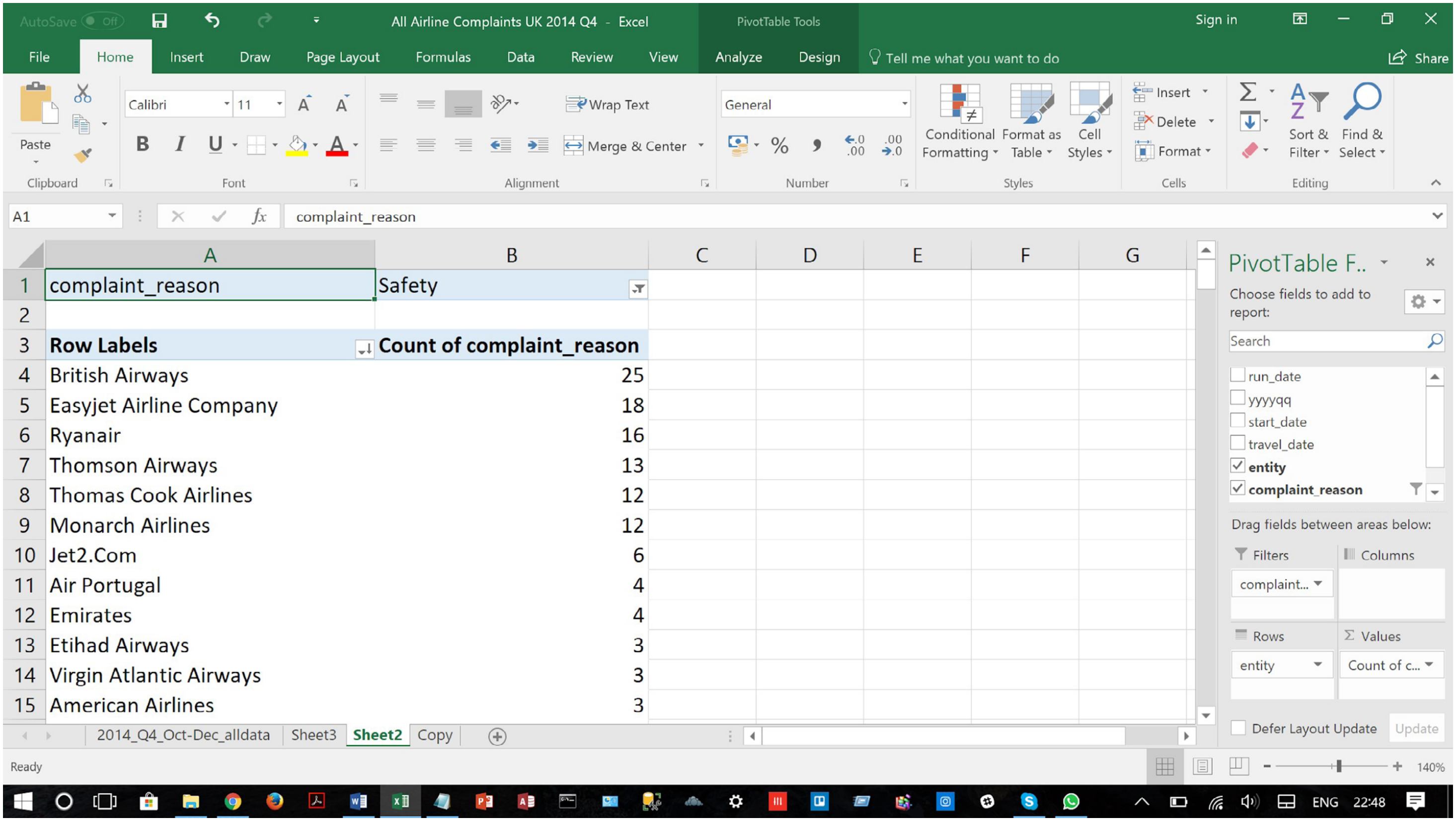
Columns

Rows

Values

□ Defer Layout Update

Add “entity” in rows and the “complaint reason” in values for the count (Table 50).
On the sheet, under “row labels”, select again More Sort Options > Descending > Count of complaint reason. The result is as in table 50.



All Airline Complaints UK 2014 Q4 - Excel

A1 complaint_reason

	A	B
1	complaint_reason	Safety
2		
3	Row Labels	Count of complaint_reason
4	British Airways	25
5	Easyjet Airline Company	18
6	Ryanair	16
7	Thomson Airways	13
8	Thomas Cook Airlines	12
9	Monarch Airlines	12
10	Jet2.Com	6
11	Air Portugal	4
12	Emirates	4
13	Etihad Airways	3
14	Virgin Atlantic Airways	3
15	American Airlines	3

PivotTable Fields

Choose fields to add to report:

- □ run_date
- □ yyyyqq
- □ start_date
- □ travel_date
- ☑ entity
- ☑ complaint_reason

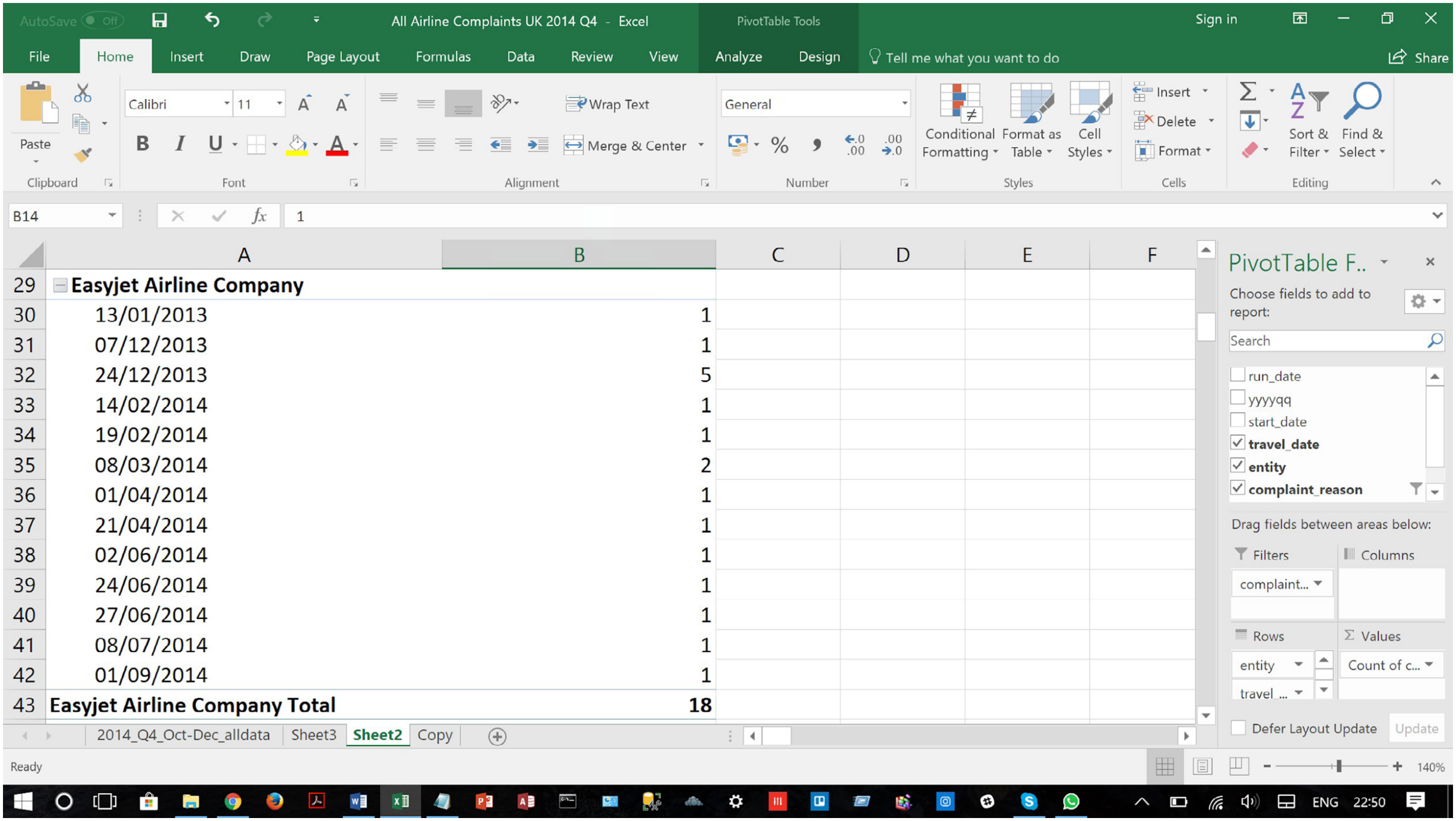
Drag fields between areas below:

Filters: complaint...
Columns:
Rows: entity
Values: Count of c...

□ Defer Layout Update

Table 50

Now you can play around with the variables to squeeze more insight. If you add the “travel date”, we can curate the data to show safety complaints count per airline with a breakdown per date (Table 51).



B14 | *fx* 1

	A	B
29	Easyjet Airline Company	
30	13/01/2013	1
31	07/12/2013	1
32	24/12/2013	5
33	14/02/2014	1
34	19/02/2014	1
35	08/03/2014	2
36	01/04/2014	1
37	21/04/2014	1
38	02/06/2014	1
39	24/06/2014	1
40	27/06/2014	1
41	08/07/2014	1
42	01/09/2014	1
43	Easyjet Airline Company Total	18

PivotTable Fields

Choose fields to add to report:

- □ run_date
- □ yyyyqq
- □ start_date
- ☑ travel_date
- ☑ entity
- ☑ complaint_reason

Drag fields between areas below:

Filters: complaint...

Columns:

Rows: entity, travel ...

Values: Count of c...

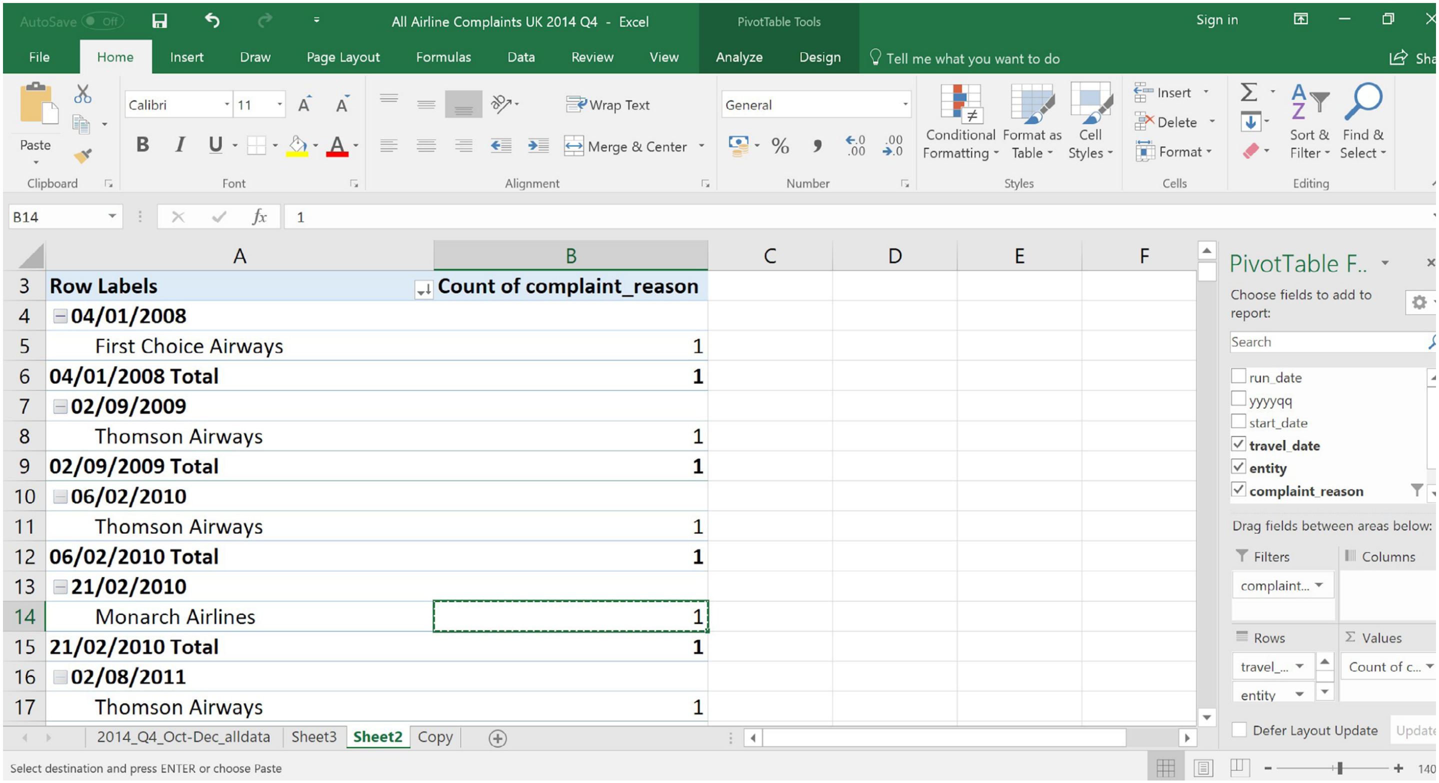
□ Defer Layout Update

Table 51

Another example: in the query you have already created, you can display the count of safety-related complaints per date, with a breakdown per airline by simply reversing the order of the columns in row labels (Table 52). This is particularly interesting if you're looking at complaints around an event, such as an ash cloud, a bombing incident, travel around Christmas and New Year, etc.

Tips:

- Try alternating the row vs. column display, to find a formatting that works for your analysis.
- When considering issues such as safety, small numbers may be meaningful. Do not disregard them just because they don't seem impressive at a first glance. It's all about context.



B14 | 1

	A	B
3	**Row Labels**	**Count of complaint_reason**
4	**⊟04/01/2008**	
5	First Choice Airways	1
6	**04/01/2008 Total**	**1**
7	**⊟02/09/2009**	
8	Thomson Airways	1
9	**02/09/2009 Total**	**1**
10	**⊟06/02/2010**	
11	Thomson Airways	1
12	**06/02/2010 Total**	**1**
13	**⊟21/02/2010**	
14	Monarch Airlines	1
15	**21/02/2010 Total**	**1**
16	**⊟02/08/2011**	
17	Thomson Airways	1

PivotTable F..

Choose fields to add to report:

Search

☐ run_date
☐ yyyyqq
☐ start_date
☑ **travel_date**
☑ **entity**
☑ **complaint_reason**

Drag fields between areas below:

Filters	Columns
complaint...	

Rows	Values
travel_... entity	Count of c...

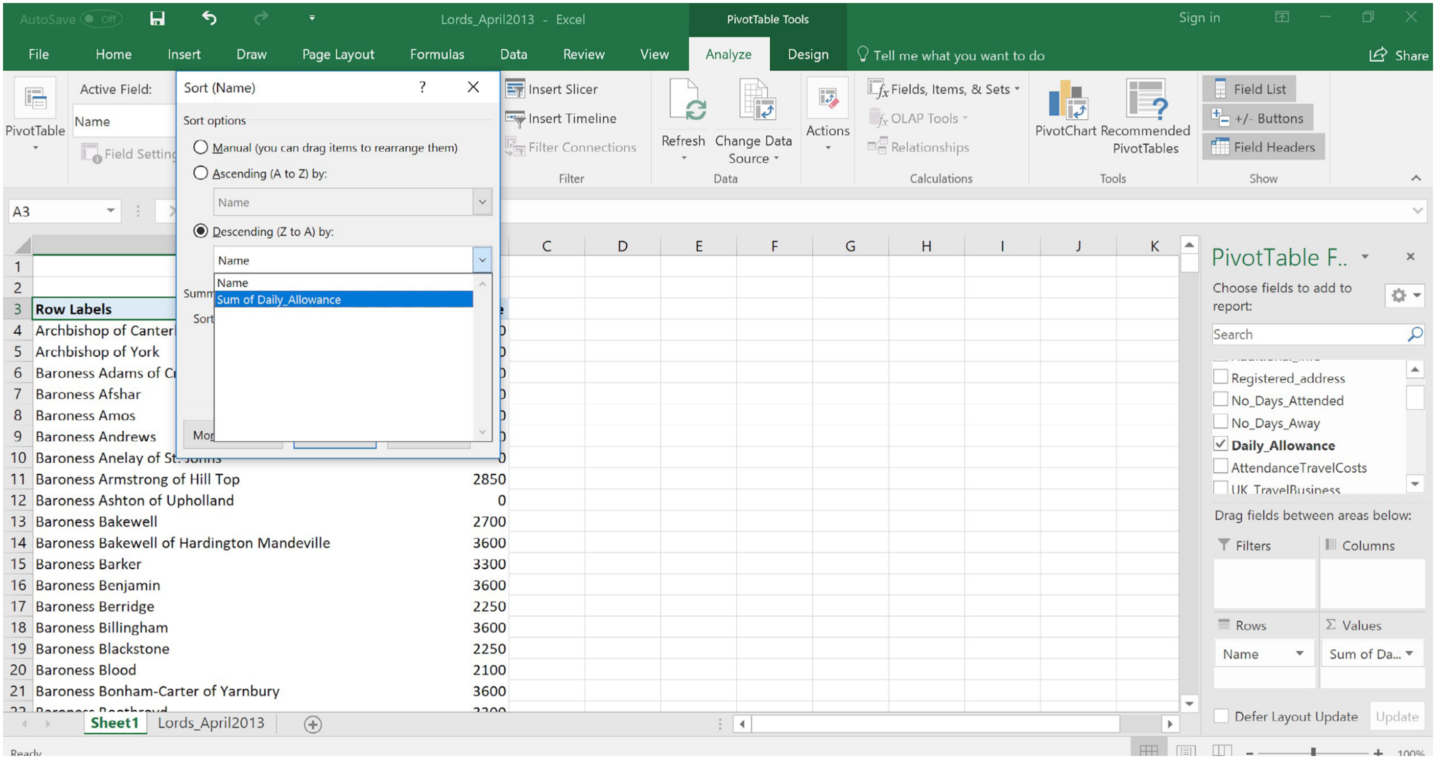
☐ Defer Layout Update Update

Table 52

Practice:

Download the British Lords' expenses dataset. Insert a pivot table and look at who claimed the most bucks for their work. By dropping the "Daily allowance" in values and the "name" in rows, you can sum the total allowance. In Row labels, select More Sort Options > Descending > Sum of daily Allowance (Table 53).

Tip: If the VALUES box counts rather than sums your numbers click on the drop-down button next to the function, select Value Field Settings and SUM from the list of descriptive statistics available.



Sort (Name)

Sort options

○ Manual (you can drag items to rearrange them)

○ Ascending (A to Z) by: Name

● Descending (Z to A) by: Name

Name

Sum of Daily_Allowance

	Row Labels	
4	Archbishop of Canter	0
5	Archbishop of York	0
6	Baroness Adams of Cr	0
7	Baroness Afshar	0
8	Baroness Amos	0
9	Baroness Andrews	0
10	Baroness Anelay of St. Johns	0
11	Baroness Armstrong of Hill Top	2850
12	Baroness Ashton of Upholland	0
13	Baroness Bakewell	2700
14	Baroness Bakewell of Hardington Mandeville	3600
15	Baroness Barker	3300
16	Baroness Benjamin	3600
17	Baroness Berridge	2250
18	Baroness Billingham	3600
19	Baroness Blackstone	2250
20	Baroness Blood	2100
21	Baroness Bonham-Carter of Yarnbury	3600

PivotTable Fields

Choose fields to add to report:

□ Registered_address
□ No_Days_Attended
□ No_Days_Away
☑ **Daily_Allowance**
□ AttendanceTravelCosts
□ UK_TravelBusiness

Drag fields between areas below:

Filters | Columns

Rows: Name | Values: Sum of Da...

□ Defer Layout Update | Update

Table 53

IMPORTANT: You have run the numbers and you have found interesting comparisons. But be careful: when it comes to data analysis, a key question is **"compared to what?"**. Even if it is our job to describe how bad an airline's service is, it would be fair to compare **apples to apples**. Line companies are different to low-costs with regards to size, number (and sometimes type) of passenger, capacity, costs, services, etc. Use averages and percentages to create a level playing-field and to make them all comparable.

This means that you have to ask the source for more information. Your story will rarely rely on only what's already publicly available. It's all good to make a start, but even beat reporting might require an added comparison field. Do some data engineering yourself and ask for more or scrape your own columns from reports, websites, etc.

a) Recap: CAR capital commandments

- You should never assume that you know anything about the data. Always ask for the metadata / record layout / data dictionary.
- A data explainer / metadata should be provided immediately, for free, and should not make the scope of yet another Freedom of Information request.
- You should never assume that the data are clean. Check, check and then check again.
- Study transparency and data protection laws.
- Find out how the data were collected, the frequency of collection, the frequency of release, release dates and platforms.
- When you see an error message, do not panic. Read it and work through the knots. Don't let setbacks or error messages stop your CAR.
- Data are rarely pure and never complete. Find out what's missing from yours.
- Ask yourself: what is fair? Avoid apples-to-pears comparisons.
- If you don't use it, you lose it.

CHAPTER VI

Preparing your data for analysis

Importing and cleaning data will always be the most boring and laborious of all stages from your CAR journey. But we cannot ignore these important steps.

These stages, however, are very informative and they allow you to get to know your digital source. It' s like reading up on that ex-government official before scheduling an interview.

IMPORTING DATA

Excel can open several file types. Popular Excel file types are .xls, .xlsx and .xlm, among many supported by Excel. Some of them do not simply open, but have to be invited inside Excel's rows and columns. When a file does not display well at a click of a button, we will have to perform an import manoeuvre.

Among Excel's non-dedicated file types are .CSV (comma-separated values) and .txt (text files). Both CSV and TXT are also associated with a flat file type.

Importing data from an XML (eXtensible Markup Language), CSV or TXT files is not always straightforward.

CSVs

Some CSV type files are not formatted to display a user-friendly spreadsheet: the text is often in one column and bleeding over the others. When this is the case (click here to download the spreadsheet Hunting_Accidents.txt)

- Open a blank spreadsheet
- Click on the DATA Tab
- From "Get and Transform Data" section, select "From Text/CSV"

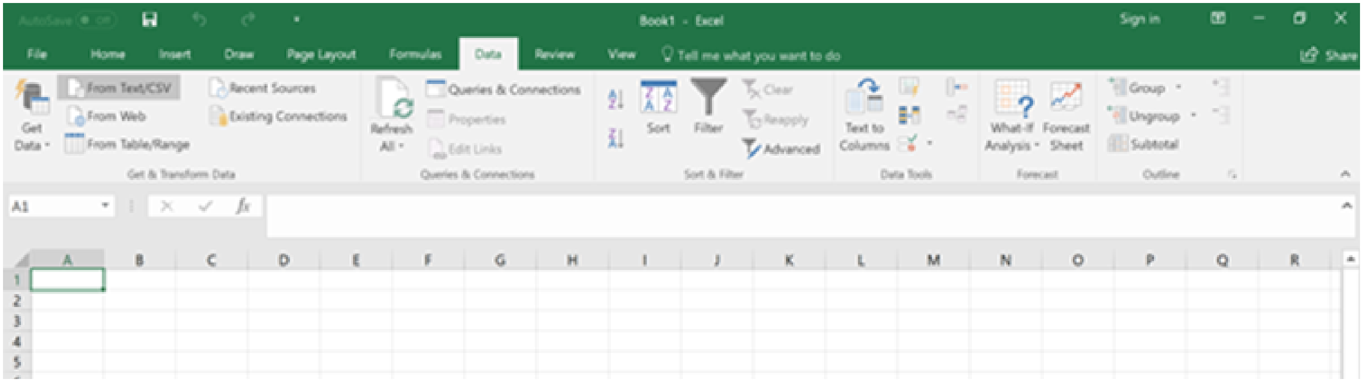
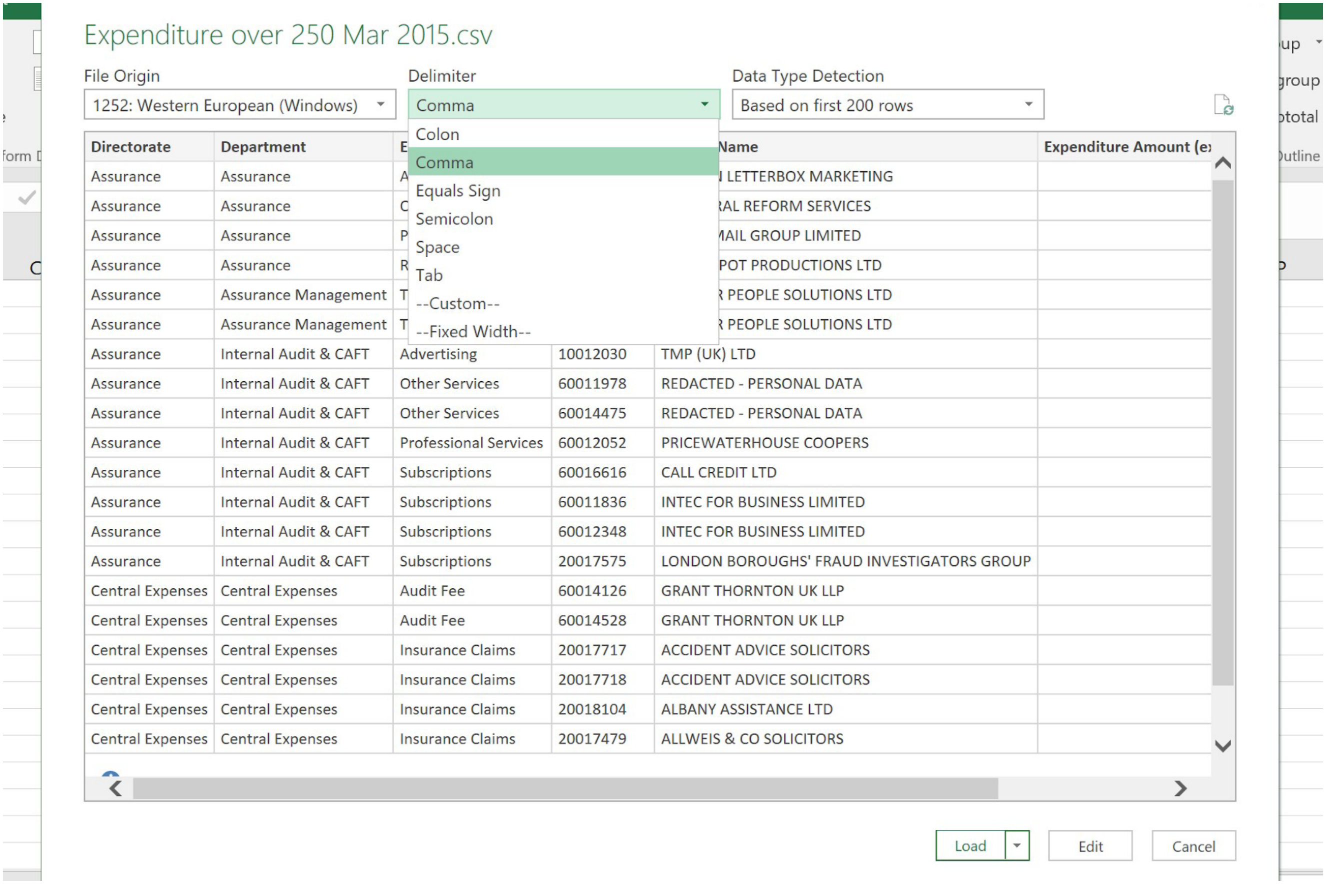


Table 54

- An import window comes up
- Select “All Files” and choose your CSV
- Click on “Import”
- The window that follows is a preview of how your data will look like once imported. Under “Delimiter”, choose “Comma” for a CSV type file, then click “Load”.



Expenditure over 250 Mar 2015.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

Data Type Detection: Based on first 200 rows

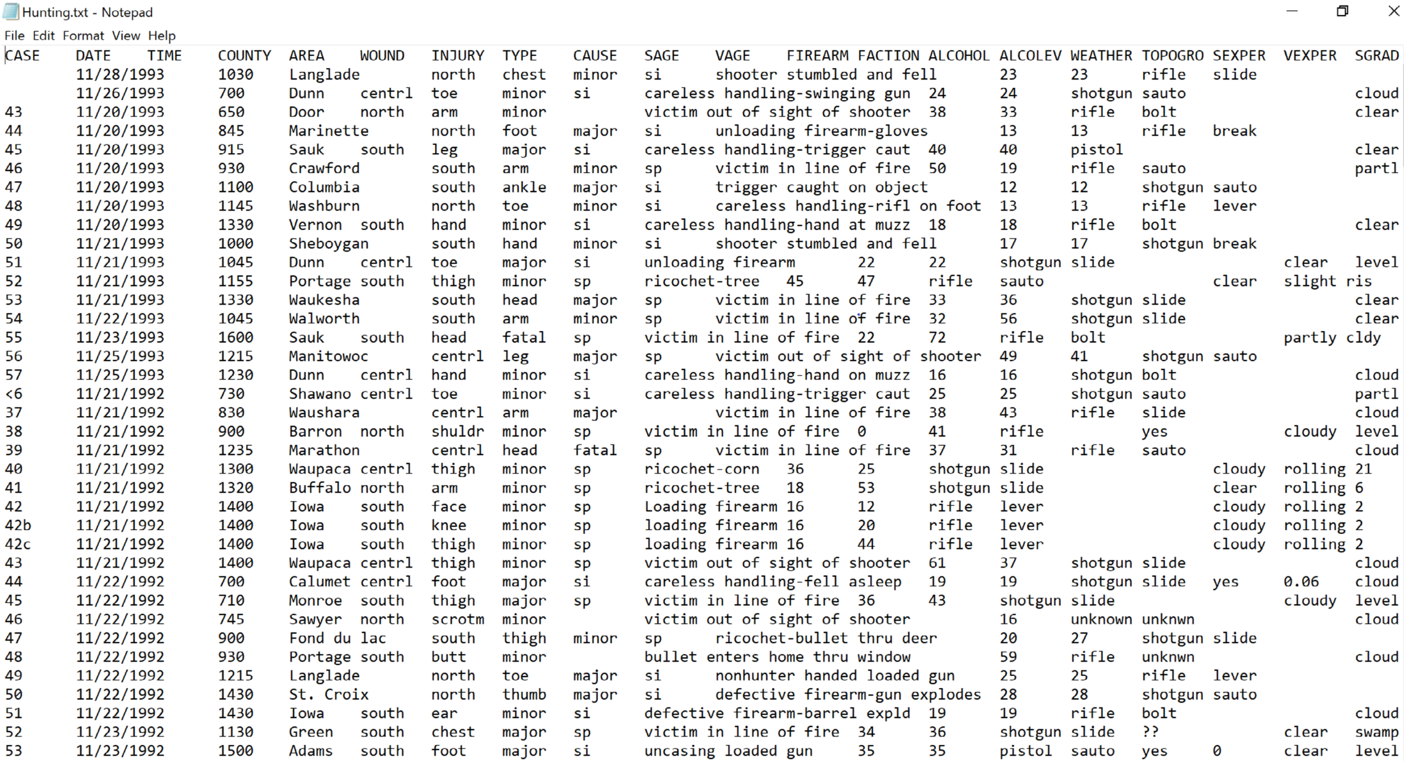
Directorate	Department	E		Name	Expenditure Amount (e
Assurance	Assurance	A		LETTERBOX MARKETING	
Assurance	Assurance	C		RAL REFORM SERVICES	
Assurance	Assurance	P		AIL GROUP LIMITED	
Assurance	Assurance	R		POT PRODUCTIONS LTD	
Assurance	Assurance Management	T		PEOPLE SOLUTIONS LTD	
Assurance	Assurance Management	T		PEOPLE SOLUTIONS LTD	
Assurance	Internal Audit & CAFT	Advertising	10012030	TMP (UK) LTD	
Assurance	Internal Audit & CAFT	Other Services	60011978	REDACTED - PERSONAL DATA	
Assurance	Internal Audit & CAFT	Other Services	60014475	REDACTED - PERSONAL DATA	
Assurance	Internal Audit & CAFT	Professional Services	60012052	PRICEWATERHOUSE COOPERS	
Assurance	Internal Audit & CAFT	Subscriptions	60016616	CALL CREDIT LTD	
Assurance	Internal Audit & CAFT	Subscriptions	60011836	INTEC FOR BUSINESS LIMITED	
Assurance	Internal Audit & CAFT	Subscriptions	60012348	INTEC FOR BUSINESS LIMITED	
Assurance	Internal Audit & CAFT	Subscriptions	20017575	LONDON BOROUGHS' FRAUD INVESTIGATORS GROUP	
Central Expenses	Central Expenses	Audit Fee	60014126	GRANT THORNTON UK LLP	
Central Expenses	Central Expenses	Audit Fee	60014528	GRANT THORNTON UK LLP	
Central Expenses	Central Expenses	Insurance Claims	20017717	ACCIDENT ADVICE SOLICITORS	
Central Expenses	Central Expenses	Insurance Claims	20017718	ACCIDENT ADVICE SOLICITORS	
Central Expenses	Central Expenses	Insurance Claims	20018104	ALBANY ASSISTANCE LTD	
Central Expenses	Central Expenses	Insurance Claims	20017479	ALLWEIS & CO SOLICITORS	

Table 55a
You should now be the proud owner of a formatted CSV table.

TXTs

When you have to import a .txt flat file, open the file with NotePad before importing and observe how the columns are delimited. Are they separated by tab, space or semicolon? This will help you assign the correct delimiter in the importing process.

Once you have ascertained that, follow the same path described for importing a CSV file type. In the case of .TXT files, it is particularly important to open them before importing. A simple NotePad software would be enough for a viewing that allows you to learn whether the columns are separated by a space, a tab, a colon etc.



```
CASE    DATE    TIME    COUNTY  AREA    WOUND   INJURY  TYPE    CAUSE   SAGE    VAGE    FIREARM FACTION ALCOHOL ALCOLEV WEATHER TOPOGRO SEXPER  VEXPER  SGRAD
        11/28/1993      1030    Langlade        north   chest   minor   si      shooter stumbled and fell       23      23      rifle   slide
        11/26/1993      700     Dunn    centrl  toe     minor   si      careless handling-swinging gun  24      24      shotgun sauto                   cloud
43      11/20/1993      650     Door    north   arm     minor           victim out of sight of shooter  38      33      rifle   bolt                    clear
44      11/20/1993      845     Marinette       north   foot    major   si      unloading firearm-gloves        13      13      rifle   break
45      11/20/1993      915     Sauk    south   leg     major   si      careless handling-trigger caut  40      40      pistol                          clear
46      11/20/1993      930     Crawford        south   arm     minor   sp      victim in line of fire  50      19      rifle   sauto                   partl
47      11/20/1993      1100    Columbia        south   ankle   major   si      trigger caught on object        12      12      shotgun sauto
48      11/20/1993      1145    Washburn        north   toe     minor   si      careless handling-rifl on foot  13      13      rifle   lever
49      11/20/1993      1330    Vernon  south   hand    minor   si      careless handling-hand at muzz  18      18      rifle   bolt                    clear
50      11/21/1993      1000    Sheboygan       south   hand    minor   si      shooter stumbled and fell       17      17      shotgun break
51      11/21/1993      1045    Dunn    centrl  toe     major   si      unloading firearm       22      22      shotgun slide                   clear   level
52      11/21/1993      1155    Portage south   thigh   minor   sp      ricochet-tree   45      47      rifle   sauto                   clear   slight ris
53      11/21/1993      1330    Waukesha        south   head    major   sp      victim in line of fire  33      36      shotgun slide                   clear
54      11/22/1993      1045    Walworth        south   arm     minor   sp      victim in line of fire  32      56      shotgun slide                   clear
55      11/23/1993      1600    Sauk    south   head    fatal   sp      victim in line of fire  22      72      rifle   bolt                    partly cldy
56      11/25/1993      1215    Manitowoc       centrl  leg     major   sp      victim out of sight of shooter  49      41      shotgun sauto
57      11/25/1993      1230    Dunn    centrl  hand    minor   si      careless handling-hand on muzz  16      16      shotgun bolt                    cloud
<6      11/21/1992      730     Shawano centrl  toe     minor   si      careless handling-trigger caut  25      25      shotgun sauto                   partl
37      11/21/1992      830     Waushara        centrl  arm     major           victim in line of fire  38      43      rifle   slide                   cloud
38      11/21/1992      900     Barron  north   shuldr  minor   sp      victim in line of fire  0       41      rifle           yes             cloudy  level
39      11/21/1992      1235    Marathon        centrl  head    fatal   sp      victim in line of fire  37      31      rifle   sauto                   cloud
40      11/21/1992      1300    Waupaca centrl  thigh   minor   sp      ricochet-corn   36      25      shotgun slide                   cloudy  rolling 21
41      11/21/1992      1320    Buffalo north   arm     minor   sp      ricochet-tree   18      53      shotgun slide                   clear   rolling 6
42      11/21/1992      1400    Iowa    south   face    minor   sp      Loading firearm 16      12      rifle   lever                   cloudy  rolling 2
42b     11/21/1992      1400    Iowa    south   knee    minor   sp      loading firearm 16      20      rifle   lever                   cloudy  rolling 2
42c     11/21/1992      1400    Iowa    south   thigh   minor   sp      loading firearm 16      44      rifle   lever                   cloudy  rolling 2
43      11/21/1992      1400    Waupaca centrl  thigh   minor   sp      victim out of sight of shooter  61      37      shotgun slide                   cloud
44      11/22/1992      700     Calumet centrl  foot    major   si      careless handling-fell asleep   19      19      shotgun slide   yes     0.06    cloud
45      11/22/1992      710     Monroe  south   thigh   major   sp      victim in line of fire  36      43      shotgun slide                   cloudy  level
46      11/22/1992      745     Sawyer  north   scrotm  minor           victim out of sight of shooter          16      unknown unknwn                  cloud
47      11/22/1992      900     Fond du lac     south   thigh   minor   sp      ricochet-bullet thru deer       20      27      shotgun slide
48      11/22/1992      930     Portage south   butt    minor           bullet enters home thru window          59      rifle   unknwn                  cloud
49      11/22/1992      1215    Langlade        north   toe     major   si      nonhunter handed loaded gun     25      25      rifle   lever
50      11/22/1992      1430    St. Croix       north   thumb   major   si      defective firearm-gun explodes  28      28      shotgun sauto
51      11/22/1992      1430    Iowa    south   ear     minor   si      defective firearm-barrel expld  19      19      rifle   bolt                    cloud
52      11/23/1992      1130    Green   south   chest   major   sp      victim in line of fire  34      36      shotgun slide   ??              clear   swamp
53      11/23/1992      1500    Adams   south   foot    major   si      uncasing loaded gun     35      35      pistol  sauto   yes     0       clear   level
```

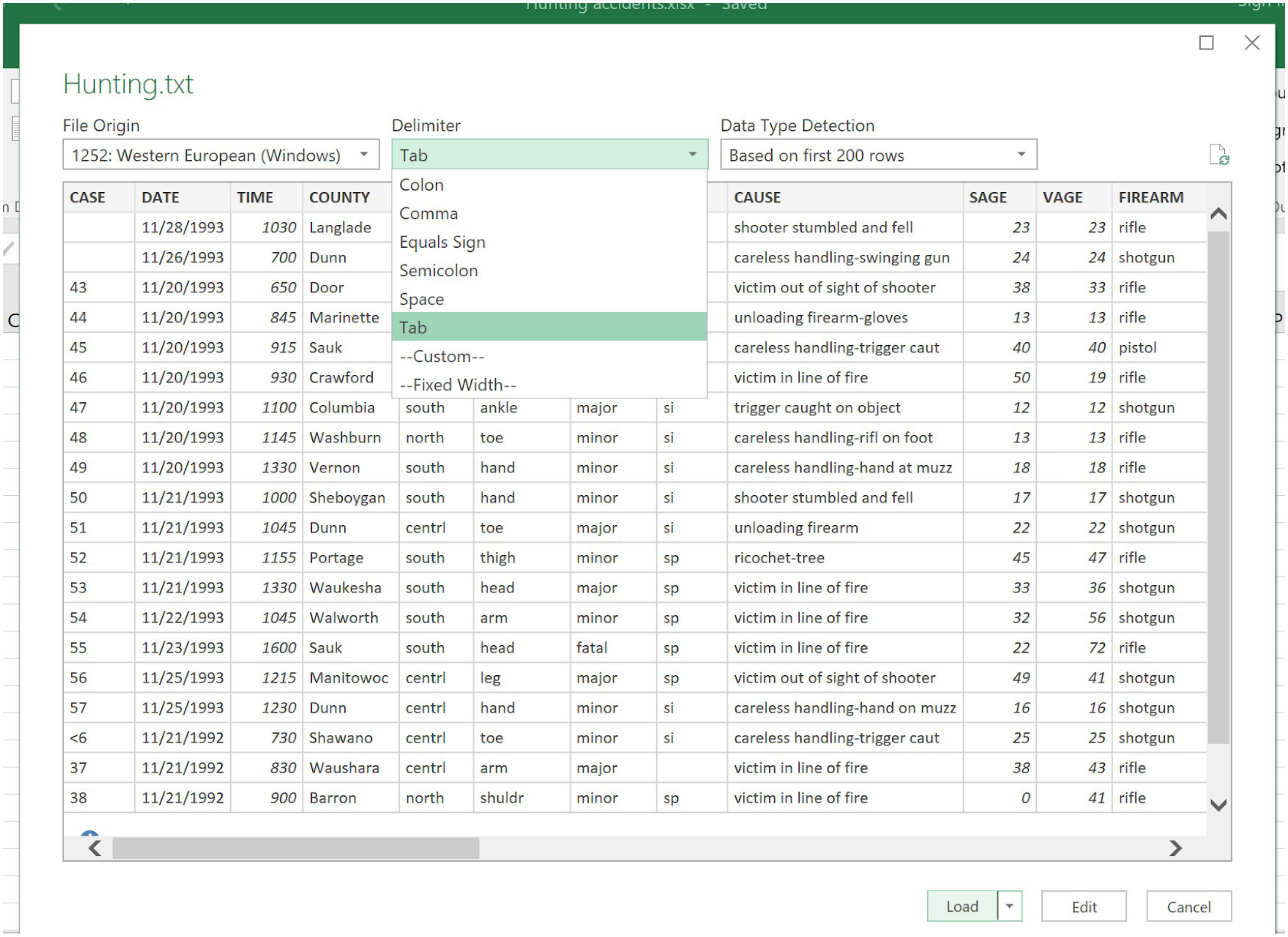
Table 55b (Hunting accidents dataset)

The columns in the US dataset that contain records of accidents that have taken place during hunting are separated by a tab. We will remember this during import.

Open a blank spreadsheet

- Click on the DATA Tab
- From “Get and Transform Data” section, select “From Text/CSV”
- Browse for the dataset, select it and click “Import”
- The import dialog continues. Click on the “Delimiter drop-down button menu and select “Tab”.

(Note: if the data is delimited by a symbol not featured in the shortlist, select “-Custom-” and enter the correct symbol.)



Hunting.txt

File Origin: 1252: Western European (Windows)

Delimiter: Tab

Data Type Detection: Based on first 200 rows

CASE	DATE	TIME	COUNTY					CAUSE	SAGE	VAGE	FIREARM
	11/28/1993	1030	Langlade					shooter stumbled and fell	23	23	rifle
	11/26/1993	700	Dunn					careless handling-swinging gun	24	24	shotgun
43	11/20/1993	650	Door					victim out of sight of shooter	38	33	rifle
44	11/20/1993	845	Marinette					unloading firearm-gloves	13	13	rifle
45	11/20/1993	915	Sauk					careless handling-trigger caut	40	40	pistol
46	11/20/1993	930	Crawford					victim in line of fire	50	19	rifle
47	11/20/1993	1100	Columbia	south	ankle	major	si	trigger caught on object	12	12	shotgun
48	11/20/1993	1145	Washburn	north	toe	minor	si	careless handling-rifl on foot	13	13	rifle
49	11/20/1993	1330	Vernon	south	hand	minor	si	careless handling-hand at muzz	18	18	rifle
50	11/21/1993	1000	Sheboygan	south	hand	minor	si	shooter stumbled and fell	17	17	shotgun
51	11/21/1993	1045	Dunn	centrl	toe	major	si	unloading firearm	22	22	shotgun
52	11/21/1993	1155	Portage	south	thigh	minor	sp	ricochet-tree	45	47	rifle
53	11/21/1993	1330	Waukesha	south	head	major	sp	victim in line of fire	33	36	shotgun
54	11/22/1993	1045	Walworth	south	arm	minor	sp	victim in line of fire	32	56	shotgun
55	11/23/1993	1600	Sauk	south	head	fatal	sp	victim in line of fire	22	72	rifle
56	11/25/1993	1215	Manitowoc	centrl	leg	major	sp	victim out of sight of shooter	49	41	shotgun
57	11/25/1993	1230	Dunn	centrl	hand	minor	si	careless handling-hand on muzz	16	16	shotgun
<6	11/21/1992	730	Shawano	centrl	toe	minor	si	careless handling-trigger caut	25	25	shotgun
37	11/21/1992	830	Waushara	centrl	arm	major		victim in line of fire	38	43	rifle
38	11/21/1992	900	Barron	north	shuldr	minor	sp	victim in line of fire	0	41	rifle

Table 56

- The assessment of the information you see in the dialog preview is based on the first 200 rows of data. If the information is entered differently from row 201 onwards, the wizard will not pick this up. But this problem is rare.
- Click on LOAD and wait for the spreadsheet to deliver a formatted table, where the filters have already been turned on. The right-hand side bar also tells us how many rows we have imported -255 in this case. Close this side-bar when you're ready.

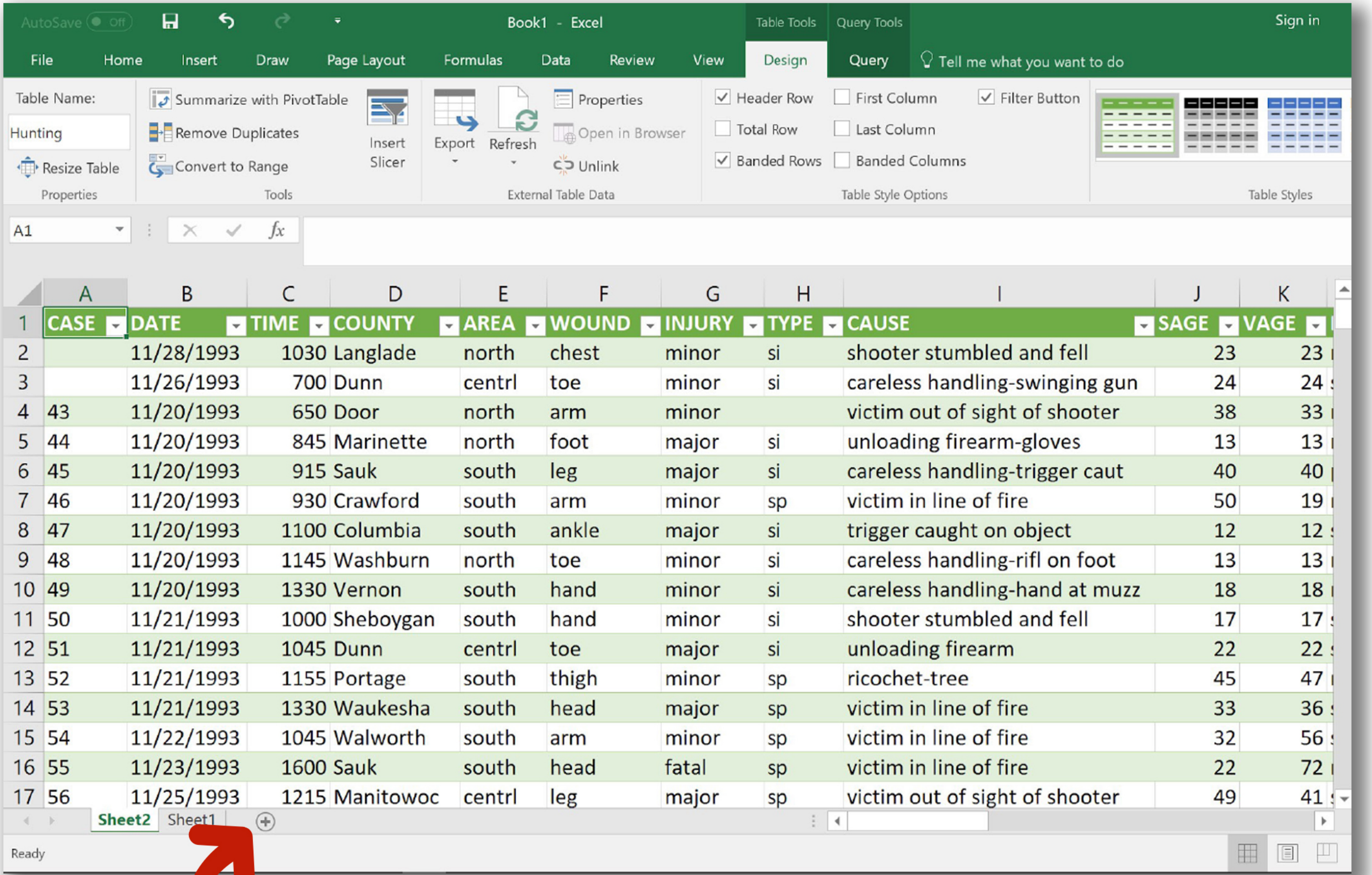


Table Name: Hunting

☑ Header Row ☐ First Column ☑ Filter Button
☐ Total Row ☐ Last Column
☑ Banded Rows ☐ Banded Columns

	A	B	C	D	E	F	G	H	I	J	K
1	CASE	DATE	TIME	COUNTY	AREA	WOUND	INJURY	TYPE	CAUSE	SAGE	VAGE
2		11/28/1993	1030	Langlade	north	chest	minor	si	shooter stumbled and fell	23	23
3		11/26/1993	700	Dunn	centrl	toe	minor	si	careless handling-swinging gun	24	24
4	43	11/20/1993	650	Door	north	arm	minor		victim out of sight of shooter	38	33
5	44	11/20/1993	845	Marinette	north	foot	major	si	unloading firearm-gloves	13	13
6	45	11/20/1993	915	Sauk	south	leg	major	si	careless handling-trigger caut	40	40
7	46	11/20/1993	930	Crawford	south	arm	minor	sp	victim in line of fire	50	19
8	47	11/20/1993	1100	Columbia	south	ankle	major	si	trigger caught on object	12	12
9	48	11/20/1993	1145	Washburn	north	toe	minor	si	careless handling-rifl on foot	13	13
10	49	11/20/1993	1330	Vernon	south	hand	minor	si	careless handling-hand at muzz	18	18
11	50	11/21/1993	1000	Sheboygan	south	hand	minor	si	shooter stumbled and fell	17	17
12	51	11/21/1993	1045	Dunn	centrl	toe	major	si	unloading firearm	22	22
13	52	11/21/1993	1155	Portage	south	thigh	minor	sp	ricochet-tree	45	47
14	53	11/21/1993	1330	Waukesha	south	head	major	sp	victim in line of fire	33	36
15	54	11/22/1993	1045	Walworth	south	arm	minor	sp	victim in line of fire	32	56
16	55	11/23/1993	1600	Sauk	south	head	fatal	sp	victim in line of fire	22	72
17	56	11/25/1993	1215	Manitowoc	centrl	leg	major	sp	victim out of sight of shooter	49	41

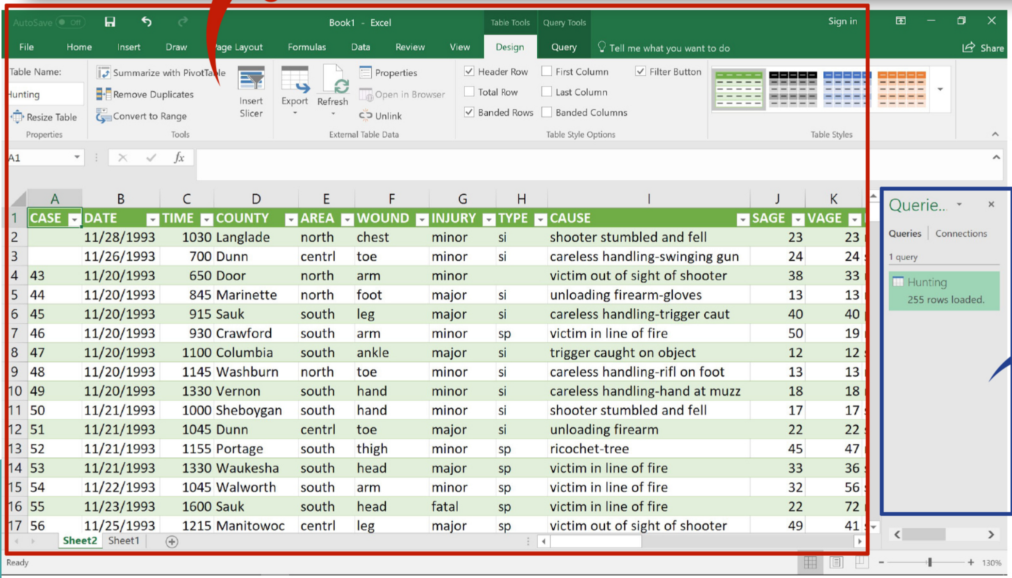


Table Name: Hunting

☑ Header Row ☐ First Column ☑ Filter Button
☐ Total Row ☐ Last Column
☑ Banded Rows ☐ Banded Columns

	A	B	C	D	E	F	G	H	I	J	K
1	CASE	DATE	TIME	COUNTY	AREA	WOUND	INJURY	TYPE	CAUSE	SAGE	VAGE
2		11/28/1993	1030	Langlade	north	chest	minor	si	shooter stumbled and fell	23	23
3		11/26/1993	700	Dunn	centrl	toe	minor	si	careless handling-swinging gun	24	24
4	43	11/20/1993	650	Door	north	arm	minor		victim out of sight of shooter	38	33
5	44	11/20/1993	845	Marinette	north	foot	major	si	unloading firearm-gloves	13	13
6	45	11/20/1993	915	Sauk	south	leg	major	si	careless handling-trigger caut	40	40
7	46	11/20/1993	930	Crawford	south	arm	minor	sp	victim in line of fire	50	19
8	47	11/20/1993	1100	Columbia	south	ankle	major	si	trigger caught on object	12	12
9	48	11/20/1993	1145	Washburn	north	toe	minor	si	careless handling-rifl on foot	13	13
10	49	11/20/1993	1330	Vernon	south	hand	minor	si	careless handling-hand at muzz	18	18
11	50	11/21/1993	1000	Sheboygan	south	hand	minor	si	shooter stumbled and fell	17	17
12	51	11/21/1993	1045	Dunn	centrl	toe	major	si	unloading firearm	22	22
13	52	11/21/1993	1155	Portage	south	thigh	minor	sp	ricochet-tree	45	47
14	53	11/21/1993	1330	Waukesha	south	head	major	sp	victim in line of fire	33	36
15	54	11/22/1993	1045	Walworth	south	arm	minor	sp	victim in line of fire	32	56
16	55	11/23/1993	1600	Sauk	south	head	fatal	sp	victim in line of fire	22	72
17	56	11/25/1993	1215	Manitowoc	centrl	leg	major	sp	victim out of sight of shooter	49	41

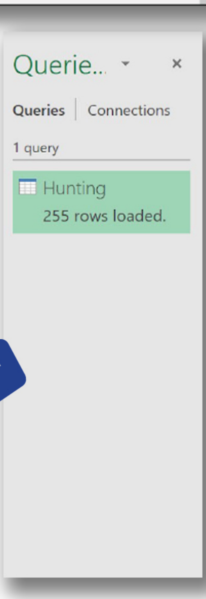
Querie..

Queries | Connections

1 query

Hunting

255 rows loaded.



Querie..

Queries | Connections

1 query

Hunting

255 rows loaded.

Table 57

Importing data from web

In the same manner, use the “From web” command to import data from a website that does not offer the option to download the dataset as any type of file.

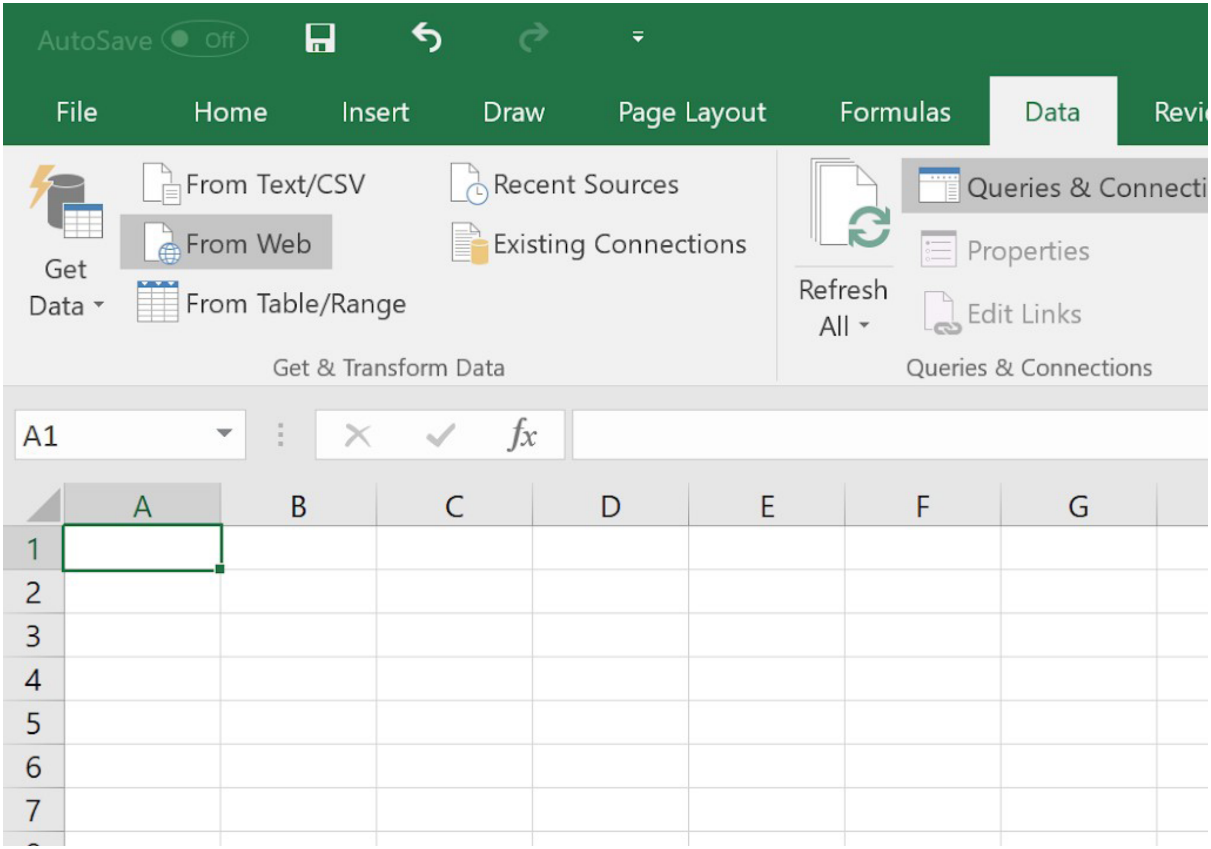
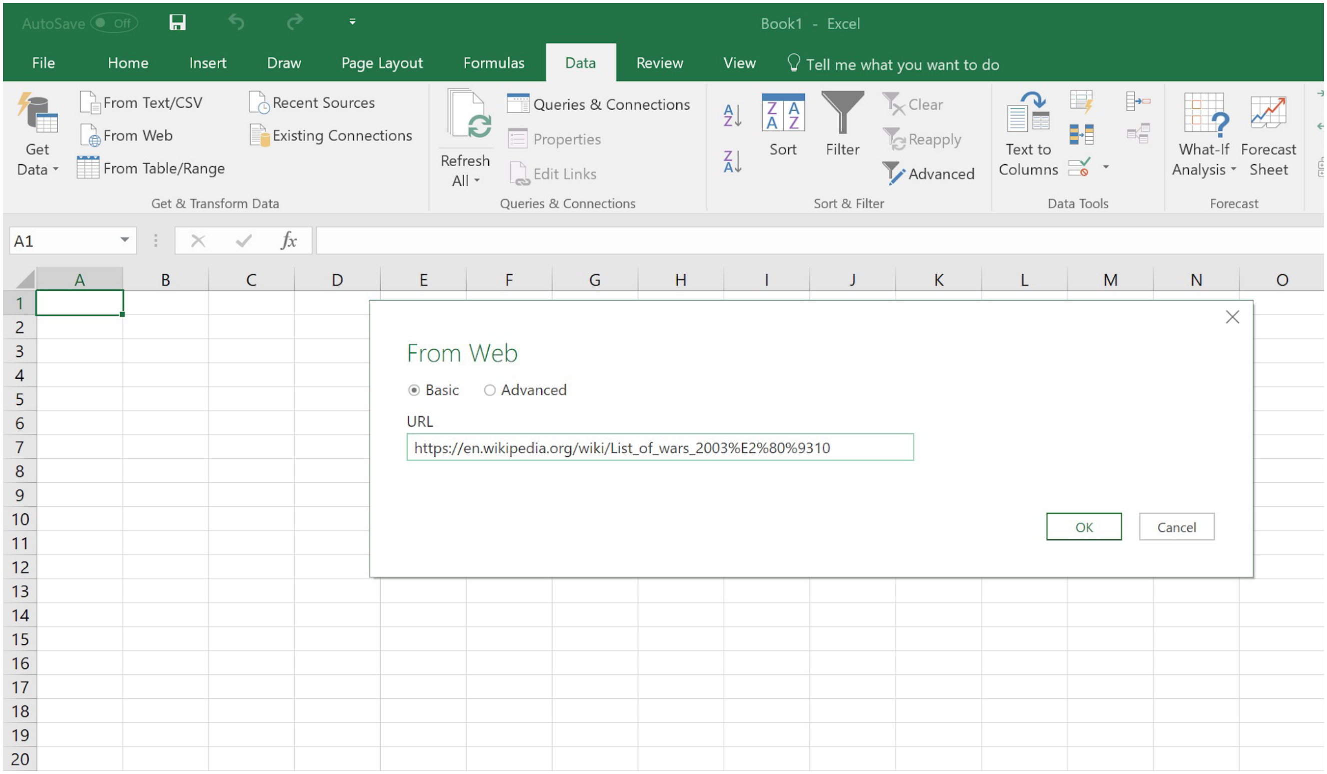


Figure 6

Open a blank Excel sheet > select the DATA tab > click on "From web"
In the import wizard, copy+paste or type the link you're interested in. Then click OK.



From Web

● Basic ○ Advanced

URL

https://en.wikipedia.org/wiki/List_of_wars_2003%E2%80%9310

Table 58

In the Navigator window, select the table on the left-hand side. On your right, the table from the Wikipedia page is displayed. When you see it, click LOAD (Table 59).

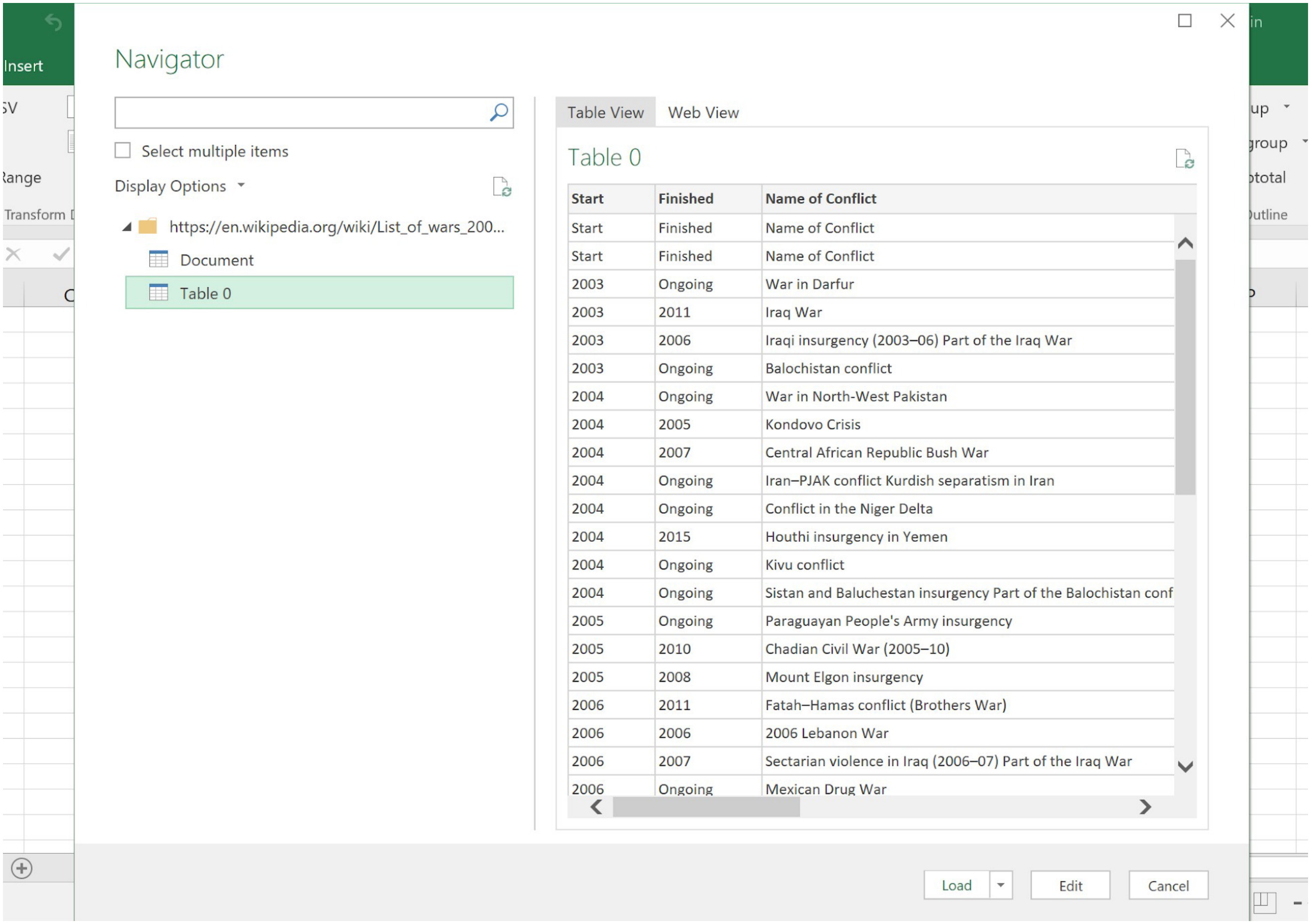


Table 0

Start	Finished	Name of Conflict
Start	Finished	Name of Conflict
Start	Finished	Name of Conflict
2003	Ongoing	War in Darfur
2003	2011	Iraq War
2003	2006	Iraqi insurgency (2003–06) Part of the Iraq War
2003	Ongoing	Balochistan conflict
2004	Ongoing	War in North-West Pakistan
2004	2005	Kondovo Crisis
2004	2007	Central African Republic Bush War
2004	Ongoing	Iran–PJAK conflict Kurdish separatism in Iran
2004	Ongoing	Conflict in the Niger Delta
2004	2015	Houthi insurgency in Yemen
2004	Ongoing	Kivu conflict
2004	Ongoing	Sistan and Baluchestan insurgency Part of the Balochistan conf
2005	Ongoing	Paraguayan People's Army insurgency
2005	2010	Chadian Civil War (2005–10)
2005	2008	Mount Elgon insurgency
2006	2011	Fatah–Hamas conflict (Brothers War)
2006	2006	2006 Lebanon War
2006	2007	Sectarian violence in Iraq (2006–07) Part of the Iraq War
2006	Ongoing	Mexican Drug War

Table 59

From the website, only the table will be extracted. It will appear formatted (Table 60). Make sure you verify that Excel has extracted a complete table of the data listed on the website. Then you're good to go.

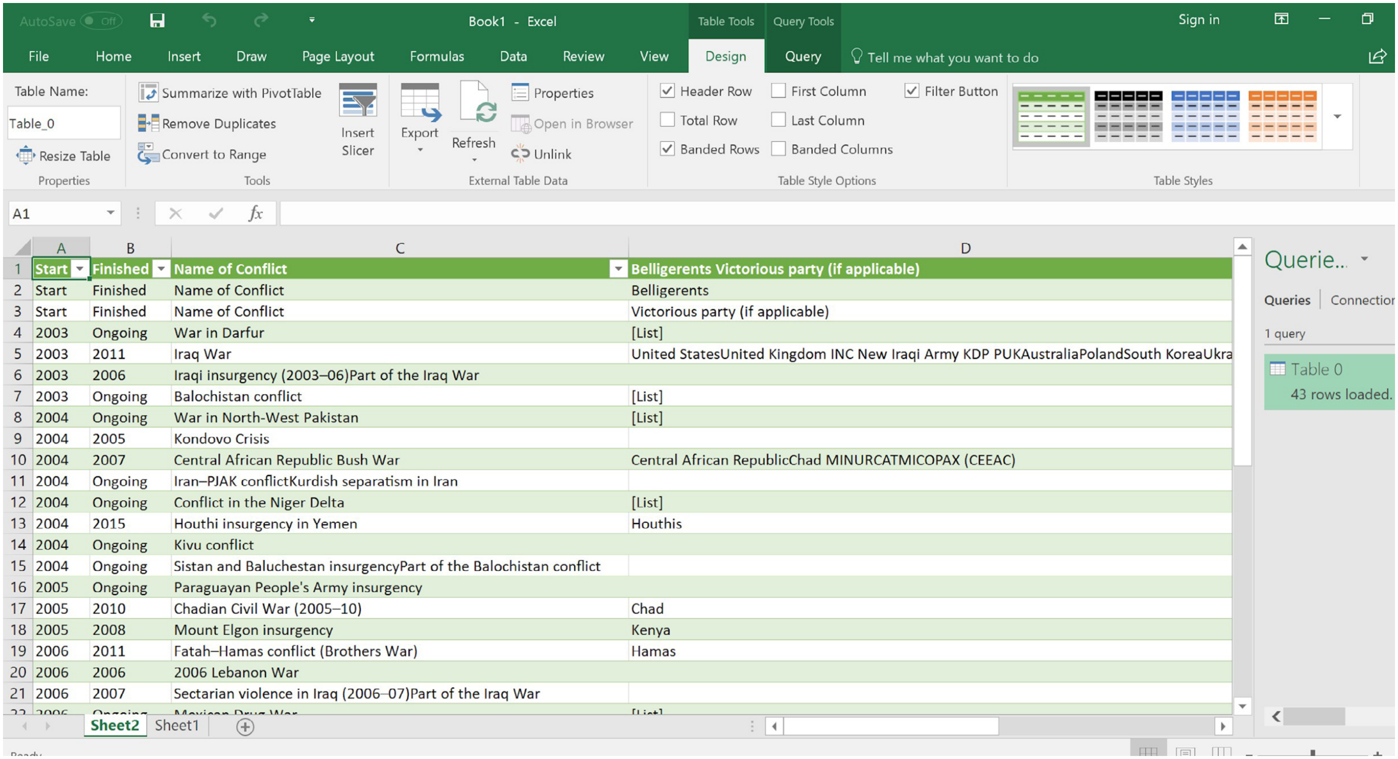


Table Name: Table_0

☑ Header Row ☐ First Column ☑ Filter Button
☐ Total Row ☐ Last Column
☑ Banded Rows ☐ Banded Columns

	A	B	C	D
1	Start	Finished	Name of Conflict	Belligerents Victorious party (if applicable)
2	Start	Finished	Name of Conflict	Belligerents
3	Start	Finished	Name of Conflict	Victorious party (if applicable)
4	2003	Ongoing	War in Darfur	[List]
5	2003	2011	Iraq War	United StatesUnited Kingdom INC New Iraqi Army KDP PUKAustraliaPolandSouth KoreaUkra
6	2003	2006	Iraqi insurgency (2003–06)Part of the Iraq War	
7	2003	Ongoing	Balochistan conflict	[List]
8	2004	Ongoing	War in North-West Pakistan	[List]
9	2004	2005	Kondovo Crisis	
10	2004	2007	Central African Republic Bush War	Central African RepublicChad MINURCATMICOPAX (CEEAC)
11	2004	Ongoing	Iran–PJAK conflictKurdish separatism in Iran	
12	2004	Ongoing	Conflict in the Niger Delta	[List]
13	2004	2015	Houthi insurgency in Yemen	Houthis
14	2004	Ongoing	Kivu conflict	
15	2004	Ongoing	Sistan and Baluchestan insurgencyPart of the Balochistan conflict	
16	2005	Ongoing	Paraguayan People's Army insurgency	
17	2005	2010	Chadian Civil War (2005–10)	Chad
18	2005	2008	Mount Elgon insurgency	Kenya
19	2006	2011	Fatah–Hamas conflict (Brothers War)	Hamas
20	2006	2006	2006 Lebanon War	
21	2006	2007	Sectarian violence in Iraq (2006–07)Part of the Iraq War	
22	2006	Ongoing	Mexican Drug War	[List]

Table 0
43 rows loaded.

Table 60

FORMATTING DATA

In everyday spreadsheet prep work, data may need to be formatted in order for your analysis to avoid errors and imprecision.

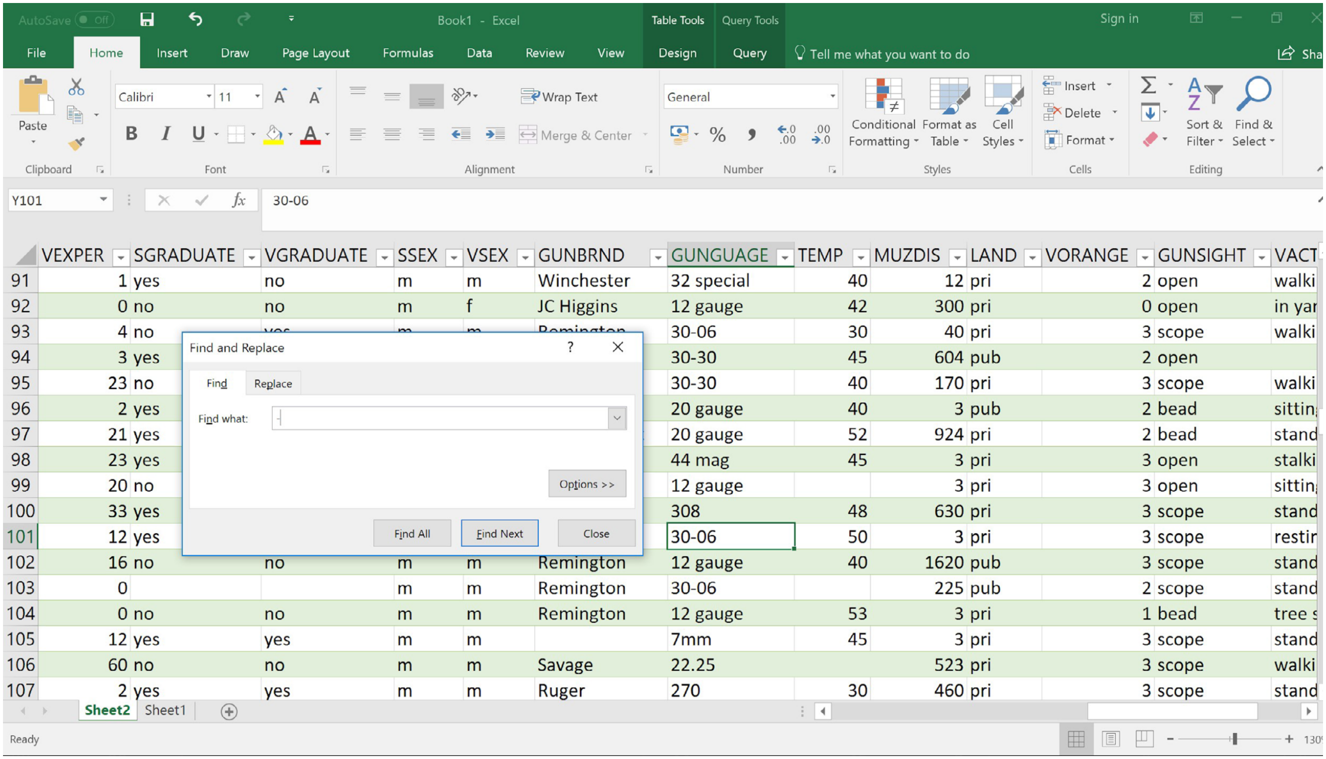
A frequent problem is when numbers (float) are entered as text (string / text) or general information (General or VarChar). Other frequent errors occur when splitting text or dates across several columns. Extra spaces between words are precision-unfriendly too, and need to be deleted.

Whether you must dismiss digital dirt or parse (curate) information in a more suitable format, there are a few steps you can take to prepare the data before querying it.

DATA CLEANING QUICK TRICKS

- ## FIND & REPLACE

The Find / Find and Replace tool can be used to locate text, numbers, symbols etc.; or to remove stray punctuation marks or spaces. Like in a Word document, click CTRL+F (PCs) / COMMAND+F (MACs) and type a word or punctuation mark in the search box. Excel is a Microsoft programme, so certain operators work like in Word. You may wish to type a term's substitute in the Replace box, or leave the Replace box empty to get rid of data.



Y101 | 30-06

Find and Replace

Find | Replace

Find what: -

Options >>

Find All | Find Next | Close

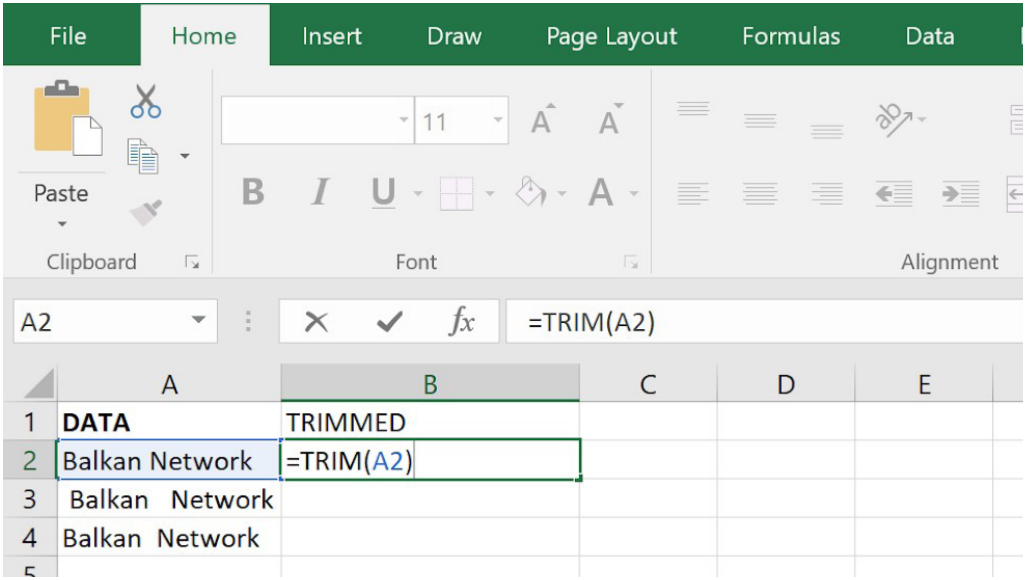
	VEXPER	SGRADUATE	VGRADUATE	SSEX	VSEX	GUNBRND	GUNGUAGE	TEMP	MUZDIS	LAND	VORANGE	GUNSIGHT	VACT
91	1	yes	no	m	m	Winchester	32 special	40	12	pri	2	open	walki
92	0	no	no	m	f	JC Higgins	12 gauge	42	300	pri	0	open	in yar
93	4	no	yes	m	m	Remington	30-06	30	40	pri	3	scope	walki
94	3	yes					30-30	45	604	pub	2	open	
95	23	no					30-30	40	170	pri	3	scope	walki
96	2	yes					20 gauge	40	3	pub	2	bead	sittin
97	21	yes					20 gauge	52	924	pri	2	bead	stand
98	23	yes					44 mag	45	3	pri	3	open	stalki
99	20	no					12 gauge		3	pri	3	open	sittin
100	33	yes					308	48	630	pri	3	scope	stand
101	12	yes					30-06	50	3	pri	3	scope	restir
102	16	no	no	m	m	Remington	12 gauge	40	1620	pub	3	scope	stand
103	0			m	m	Remington	30-06		225	pub	2	scope	stand
104	0	no	no	m	m	Remington	12 gauge	53	3	pri	1	bead	tree s
105	12	yes	yes	m	m		7mm	45	3	pri	3	scope	stand
106	60	no	no	m	m	Savage	22.25		523	pri	3	scope	walki
107	2	yes	yes	m	m	Ruger	270	30	460	pri	3	scope	stand

Table 61

- ## TRIMMING

Often, data come with rogue extra spaces that constitute digital dirt (Table 62). To remove extra spaces, often difficult to spot, run a function called TRIM.

=TRIM(Cell address)

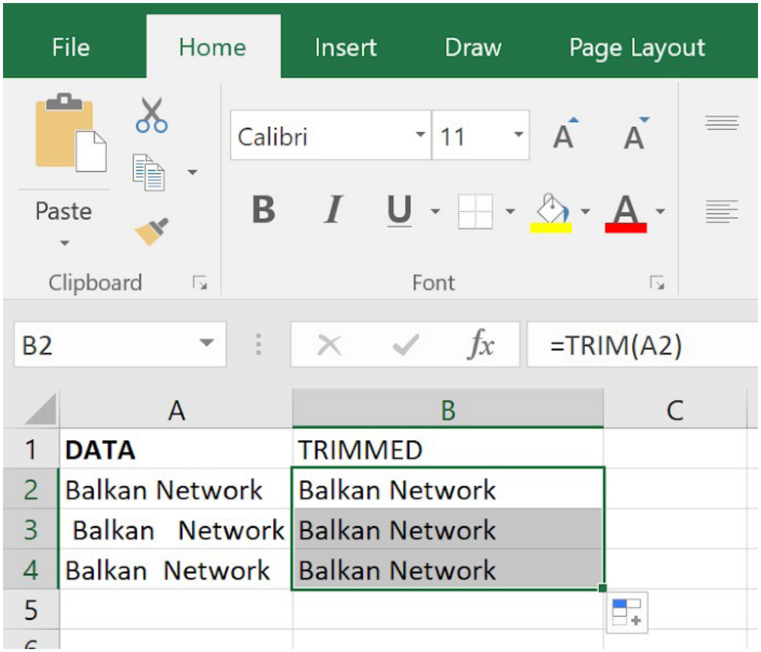


A2 =TRIM(A2)

	A	B	C	D	E
1	**DATA**	TRIMMED			
2	Balkan Network	=TRIM(A2)			
3	Balkan Network				
4	Balkan Network				

Table 62

Write the formula once and double-click down the column to clean all cells (Table 63). You’re not ready just yet.



B2 =TRIM(A2)

	A	B	C
1	**DATA**	TRIMMED	
2	Balkan Network	Balkan Network	
3	Balkan Network	Balkan Network	
4	Balkan Network	Balkan Network	
5			

Table 63

Your freshly-trimmed column B is dependent on the existence of column A. If you delete column A, column B would lose its data. You should avoid function-results dependencies. Select the newly-trimmed column, copy it (Copy + C or Command + C), right-click in B1 > paste as values. At first, nothing appears to be changed. But if you click in B2, the TRIM function will no longer show up in the formula bar (Fx). The column is now ready for analysis.

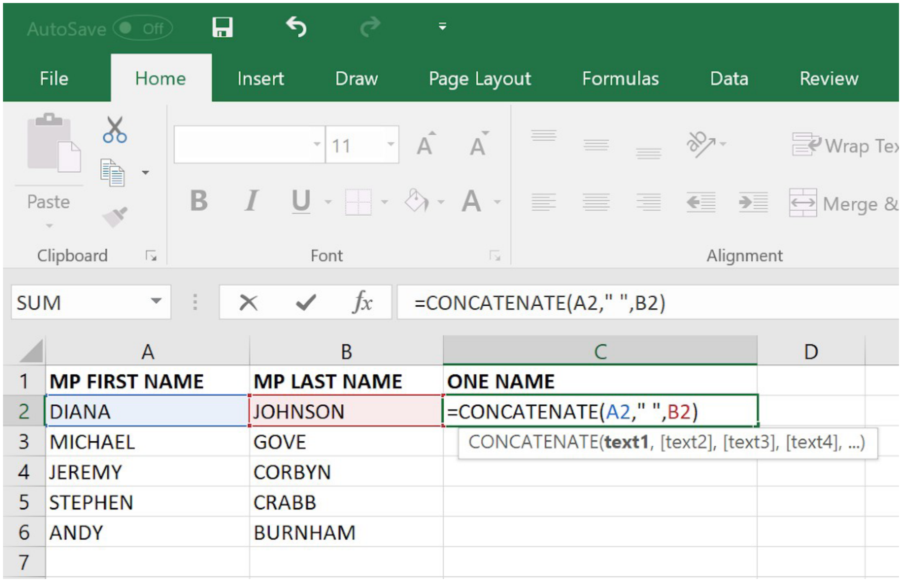
Tip: Keep both the original and the clean column, so you can always have a point of reference and check yourself against the original data.

- CONCATENATE

At times, you may need to bring bits of text together, and they will happen to be sitting in different columns. Say you have a list of names. The first name is in one column and the last name is in another column. In order to bring the two together, we need to CONCATENATE.

=CONCATENATE(cell address," ",cell address)

In Excel's case, a space does not equal nothing. Rather, it is information and a spreadsheet is aware of its presence or absence. That is why there must be a space between quotation marks in between the two cell addresses. If we omit this, Excel will stick the two names together without any space between them.

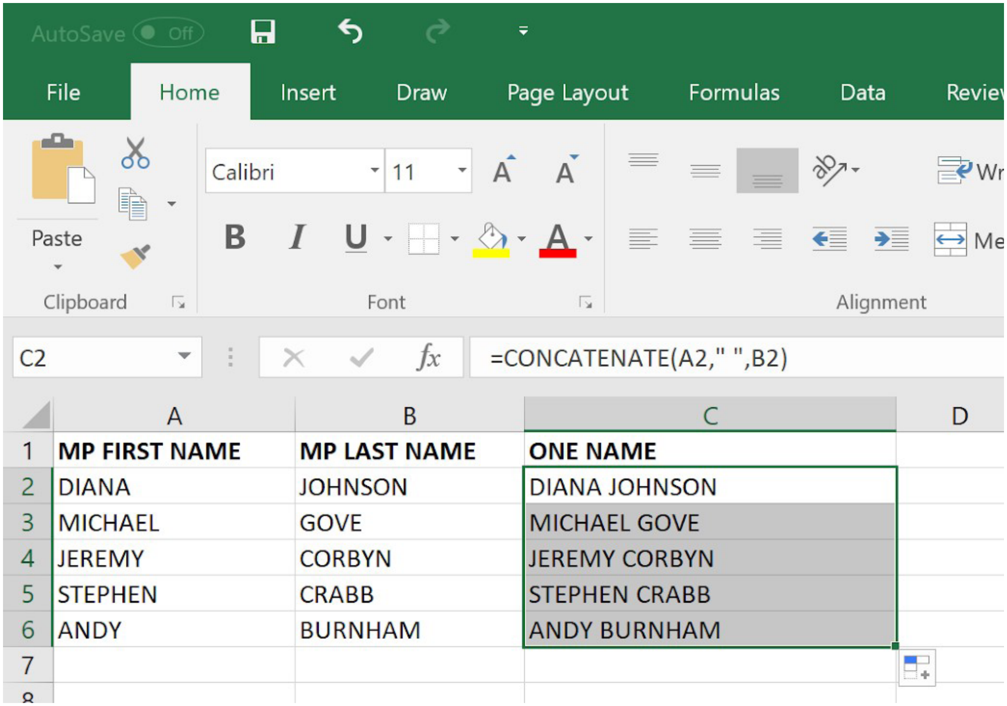


=CONCATENATE(A2," ",B2)

	A	B	C	D
1	MP FIRST NAME	MP LAST NAME	ONE NAME	
2	DIANA	JOHNSON	=CONCATENATE(A2," ",B2)	
3	MICHAEL	GOVE	CONCATENATE(**text1**, [text2], [text3], [text4], ...)	
4	JEREMY	CORBYN		
5	STEPHEN	CRABB		
6	ANDY	BURNHAM		
7				

Table 64

Once concatenated, repeat Copy>Paste as Values explained above, to ready the data for analysis.



C2 =CONCATENATE(A2," ",B2)

	A	B	C	D
1	MP FIRST NAME	MP LAST NAME	ONE NAME	
2	DIANA	JOHNSON	DIANA JOHNSON	
3	MICHAEL	GOVE	MICHAEL GOVE	
4	JEREMY	CORBYN	JEREMY CORBYN	
5	STEPHEN	CRABB	STEPHEN CRABB	
6	ANDY	BURNHAM	ANDY BURNHAM	
7				

Table 65

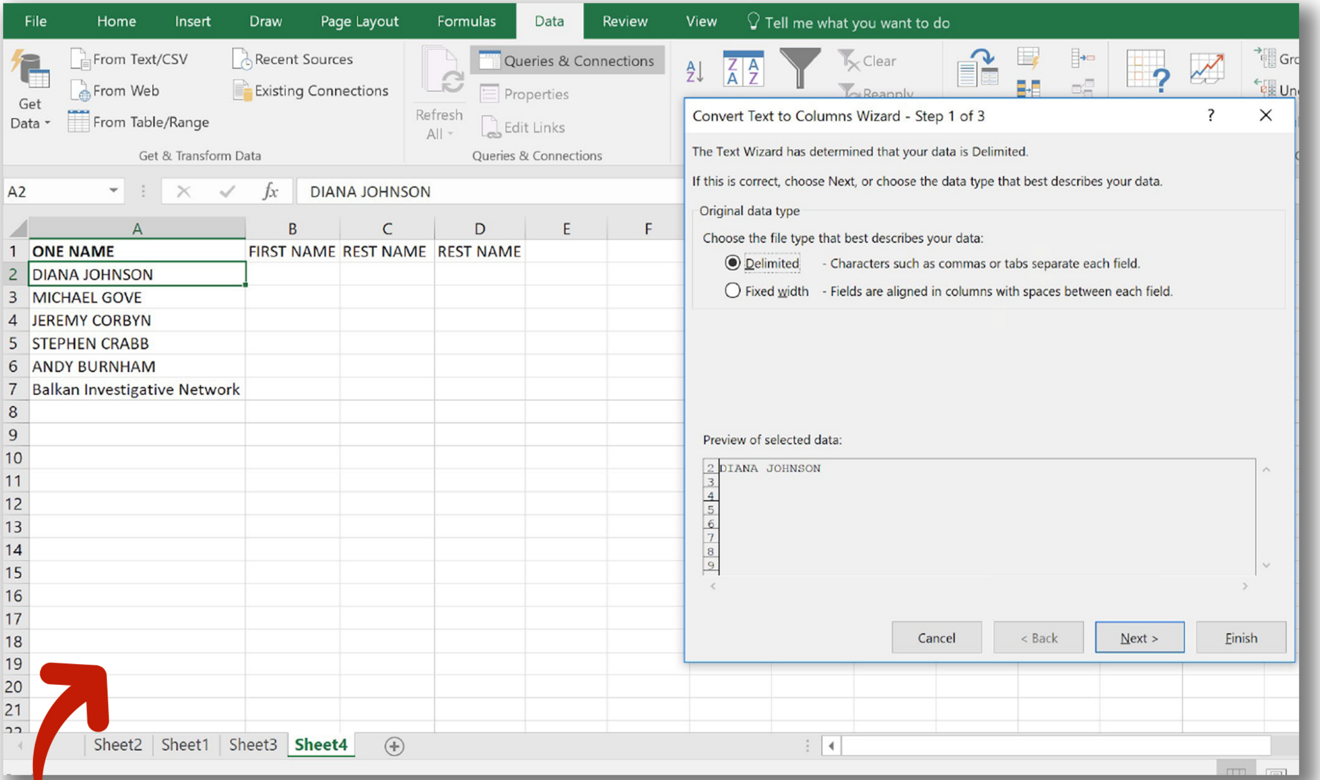
Tip: When joining names, some of them might be longer than two words. Take the longest name as your base, and write the first formula as such before double-clicking down the column. This way, you will correctly concatenate every record.

- TEXT-TO-COLUMNS

To split a name over several columns, use the TEXT-TO-COLUMNS command from the DATA ribbon:

>Click on the cell address you wish to split

>Data > Text to columns



A2 fx DIANA JOHNSON

	A	B	C	D
1	ONE NAME	FIRST NAME	REST NAME	REST NAME
2	DIANA JOHNSON			
3	MICHAEL GOVE			
4	JEREMY CORBYN			
5	STEPHEN CRABB			
6	ANDY BURNHAM			
7	Balkan Investigative Network			

Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

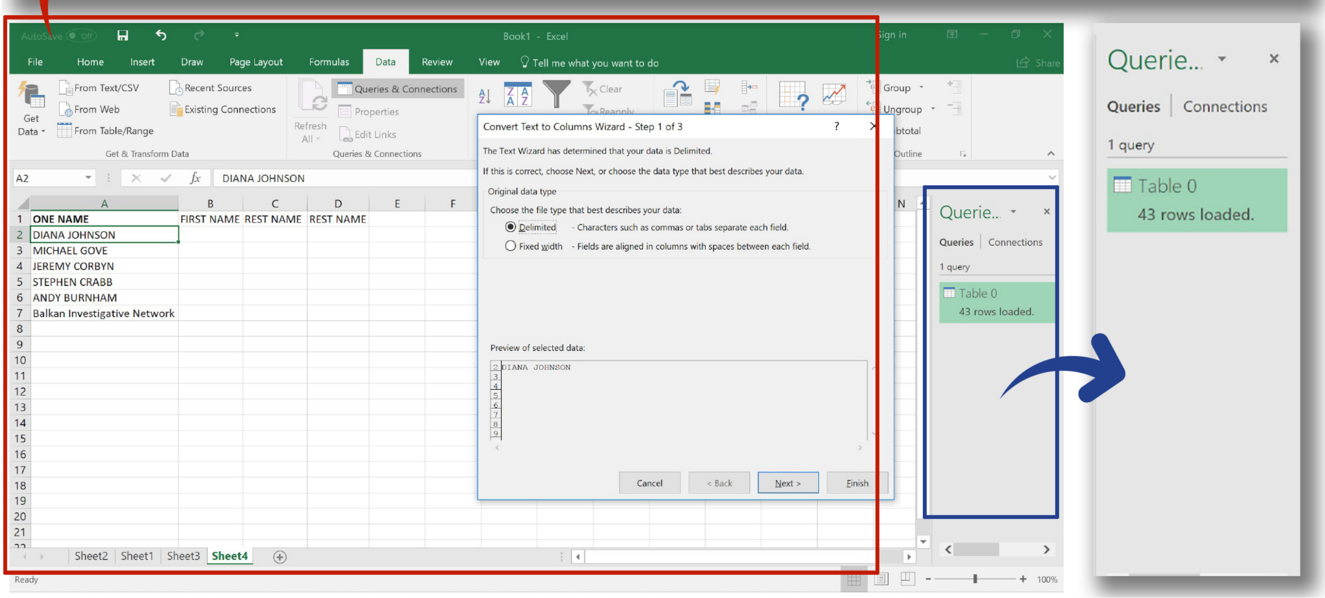
Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.

○ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

```
2 DIANA JOHNSON
3
4
5
6
7
8
9
```



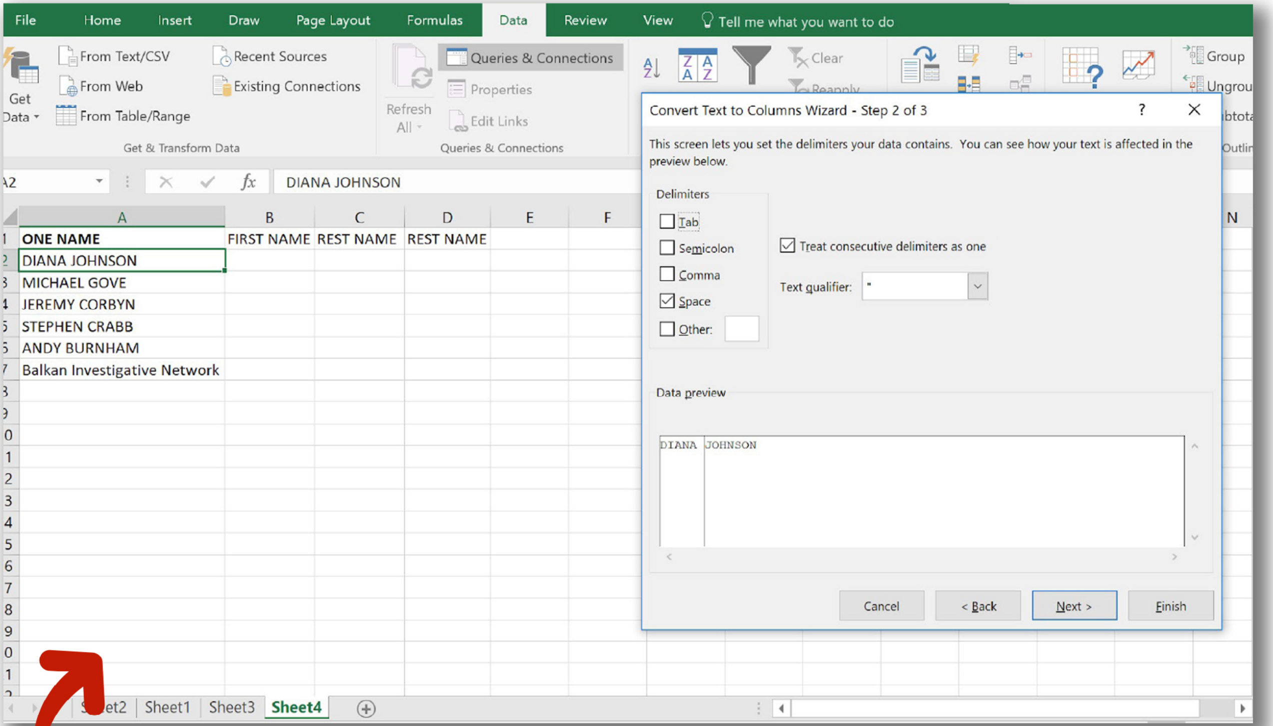
Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Table 66

> Choose the delimiter: in this case, it’s space, but it can also be tab, semicolon, comma, or any other punctuation mark or symbol;



A2 fx DIANA JOHNSON

A	B	C	D
ONE NAME	FIRST NAME	REST NAME	REST NAME
DIANA JOHNSON			
MICHAEL GOVE			
JEREMY CORBYN			
STEPHEN CRABB			
ANDY BURNHAM			
Balkan Investigative Network			

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

□ Tab
□ Semicolon
□ Comma
☑ Space
□ Other:

☑ Treat consecutive delimiters as one

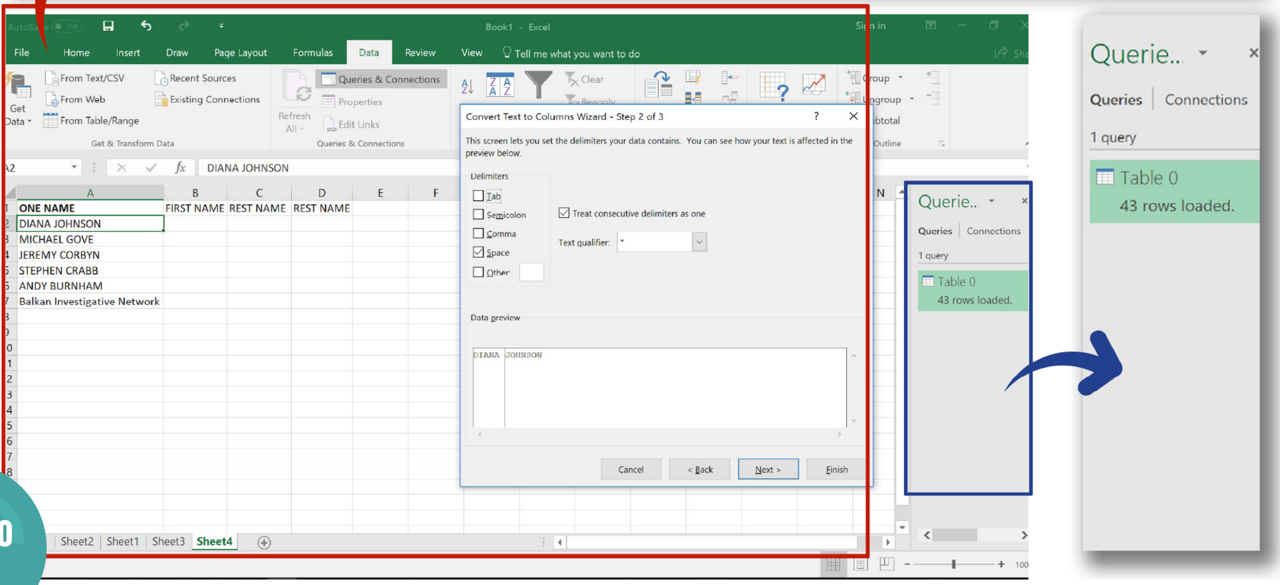
Text qualifier: "

Data preview

```
DIANA JOHNSON
```

Cancel | < Back | Next > | Finish

Table 67



A2 fx DIANA JOHNSON

ONE NAME | FIRST NAME | REST NAME | REST NAME
DIANA JOHNSON
MICHAEL GOVE
JEREMY CORBYN
STEPHEN CRABB
ANDY BURNHAM
Balkan Investigative Network

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

□ Tab
□ Semicolon
□ Comma
☑ Space
□ Other:

☑ Treat consecutive delimiters as one

Text qualifier: "

Data preview

DIANA JOHNSON

Cancel | < Back | Next > | Finish

Queries & Connections

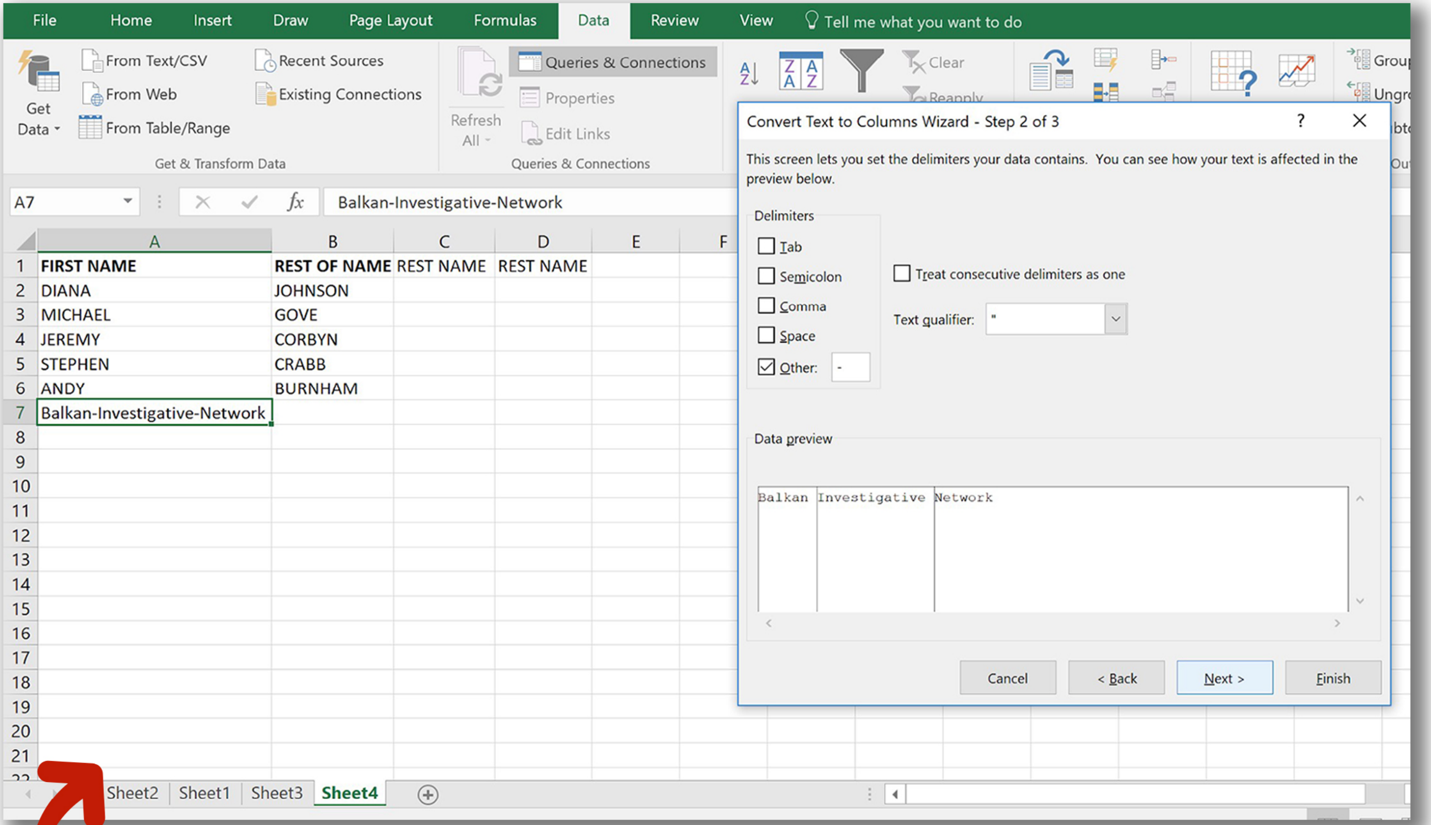
Queries | Connections

1 query

Table 0
43 rows loaded.

> In the next window, select the type of information in each column. “General” lets Excel decide what type of information it’s dealing with. The “Date” option allows you to format the date [for example, when you want to change the European format Day Month Year (DMY) to the US format – Month Day Year (MDY)]. >Finish

Get hands-on and establish the delimiter yourself. In Table 68, BIRN is a 3-word name, where words are separated by a dash “-“. In such cases, select OTHER as the delimiter and enter the dash from your keyboard into the box. The text will appear split over three columns, correctly.



A7 — Balkan-Investigative-Network

	A	B	C	D
1	FIRST NAME	REST OF NAME	REST NAME	REST NAME
2	DIANA	JOHNSON		
3	MICHAEL	GOVE		
4	JEREMY	CORBYN		
5	STEPHEN	CRABB		
6	ANDY	BURNHAM		
7	Balkan-Investigative-Network			

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

□ Tab

□ Semicolon

□ Comma

□ Space

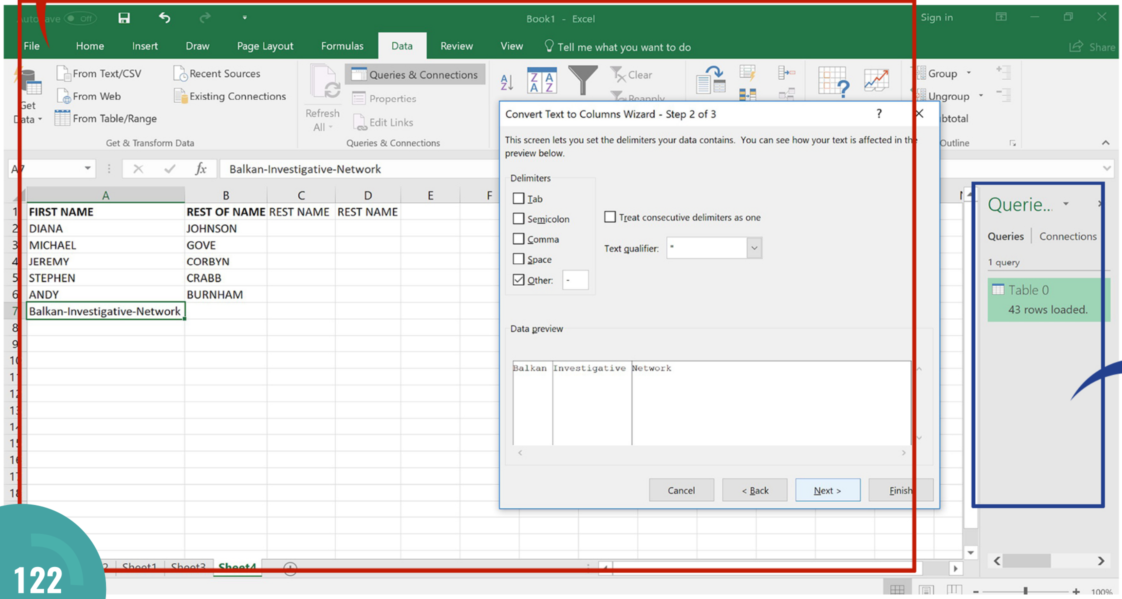
☑ Other: -

□ Treat consecutive delimiters as one

Text qualifier: "

Data preview

```
Balkan Investigative Network
```


A7 — Balkan-Investigative-Network

	A	B	C	D
1	FIRST NAME	REST OF NAME	REST NAME	REST NAME
2	DIANA	JOHNSON		
3	MICHAEL	GOVE		
4	JEREMY	CORBYN		
5	STEPHEN	CRABB		
6	ANDY	BURNHAM		
7	Balkan-Investigative-Network			

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

□ Tab

□ Semicolon

□ Comma

□ Space

☑ Other: -

□ Treat consecutive delimiters as one

Text qualifier: "

Data preview

Balkan Investigative Network

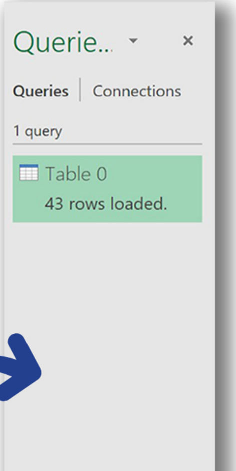
Queries...

Queries | Connections

1 query

Table 0

43 rows loaded.


Querie...

Queries | Connections

1 query

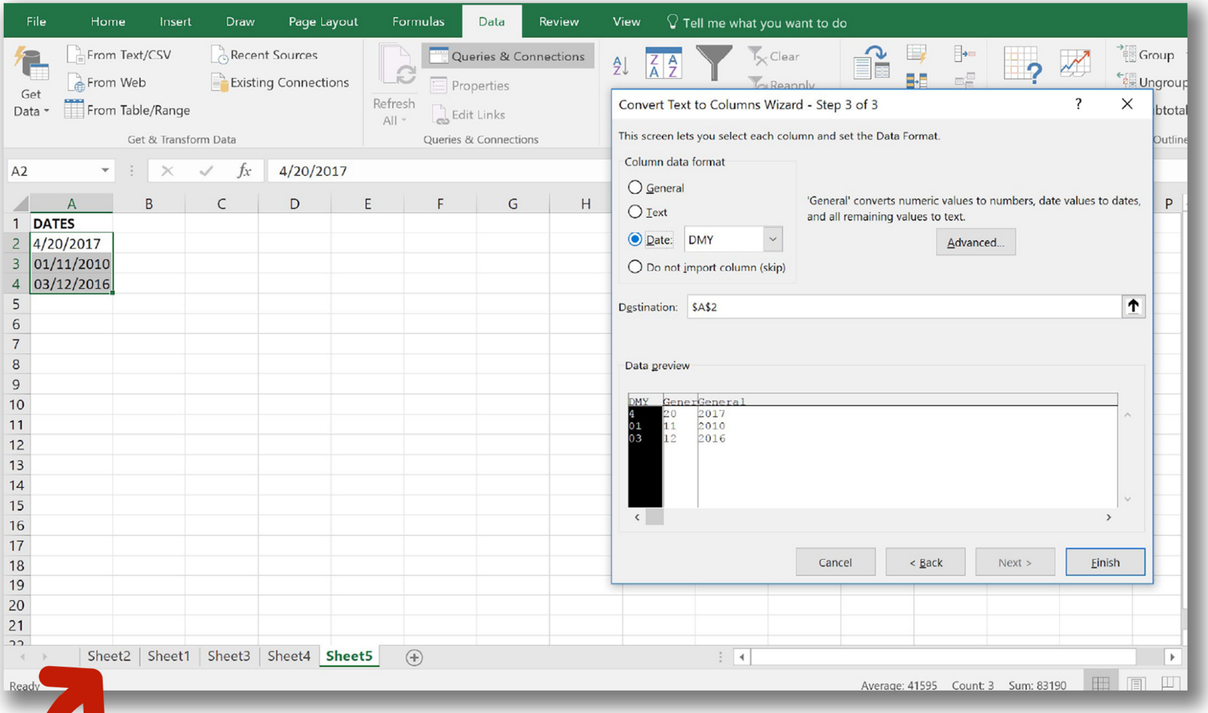
Table 0

43 rows loaded.

Table 68

- ## SPLITTING DATES

Splitting dates with TEXT-TO-COLUMNS is counter-intuitive. Try the logical path - Cell > Data>Text to columns > Select Delimiter > Select DATE as data type



A2 fx 4/20/2017

A
DATES
4/20/2017
01/11/2010
03/12/2016

Convert Text to Columns Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

○ General

○ Text

● Date: DMY

○ Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Advanced...

Destination: $A$2

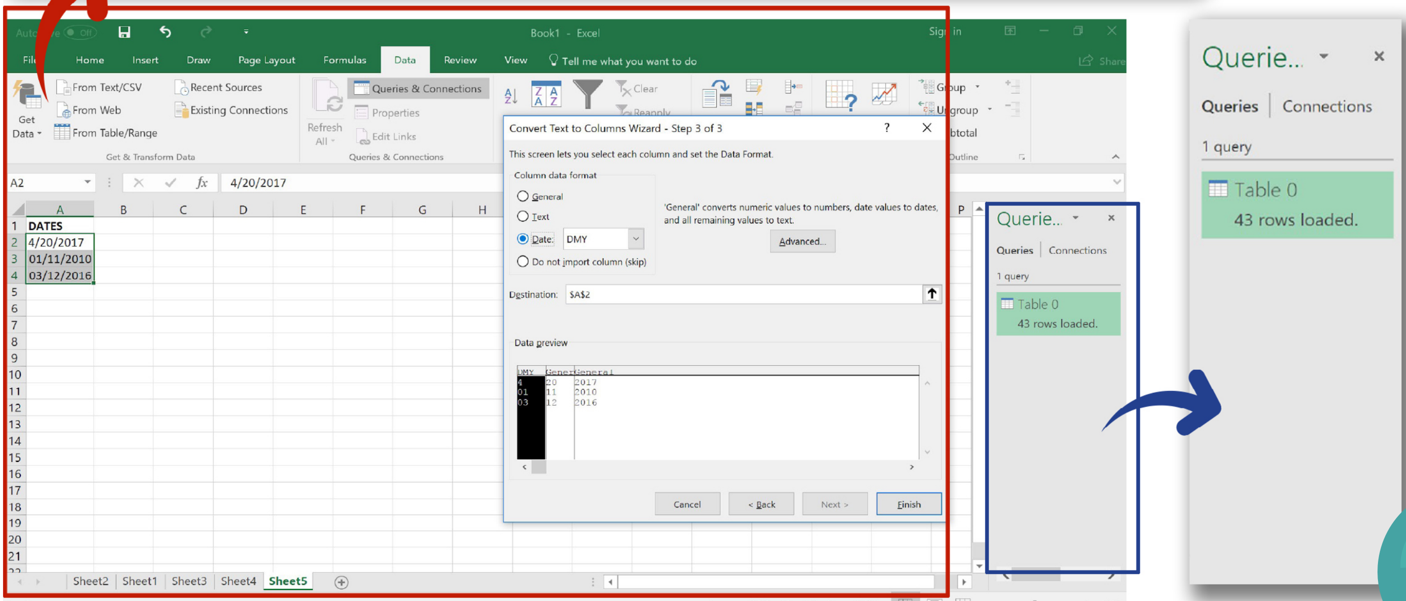
Data preview

```
DMY GenerGeneral
4   20   2017
01  11   2010
03  12   2016
```

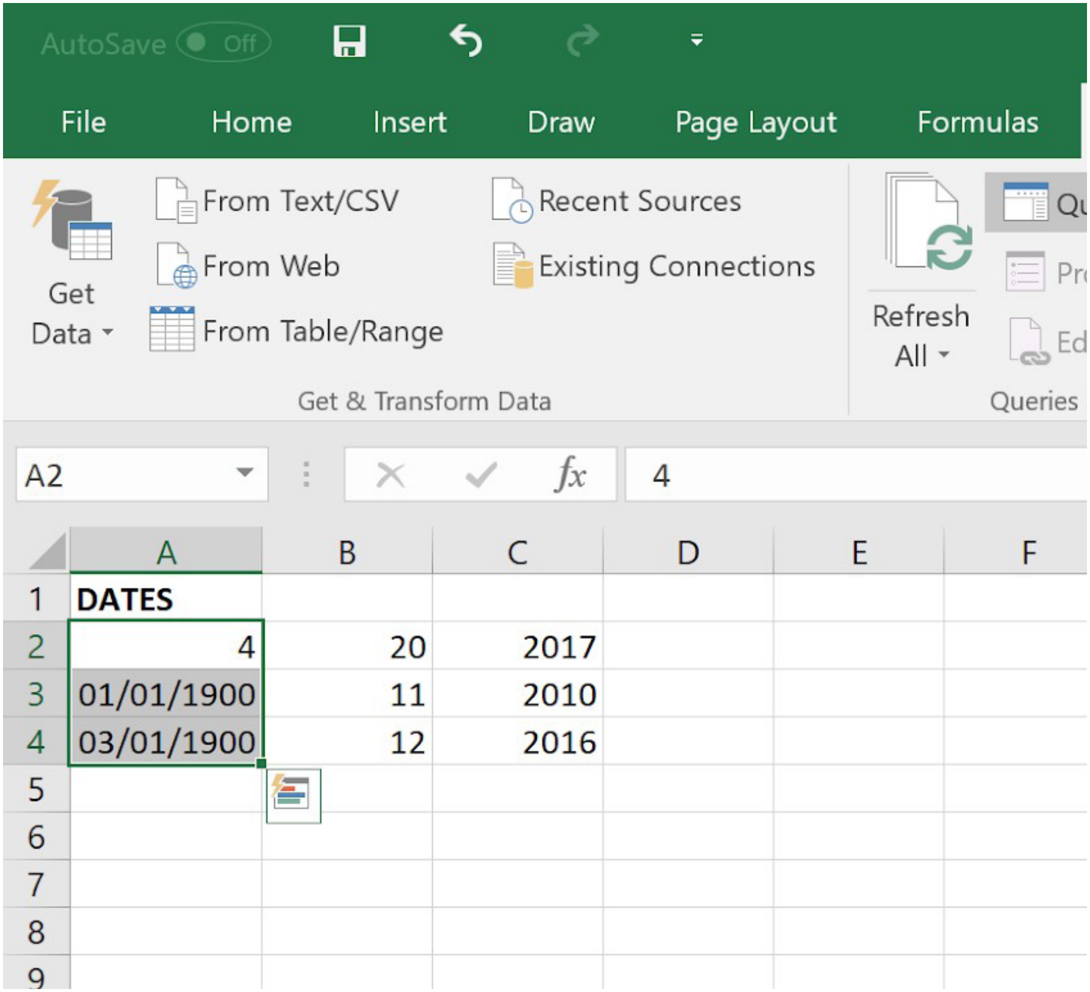
Cancel | < Back | Next > | Finish

Average: 41595 Count: 3 Sum: 83190

Table 69



> Finish. This will be the result:



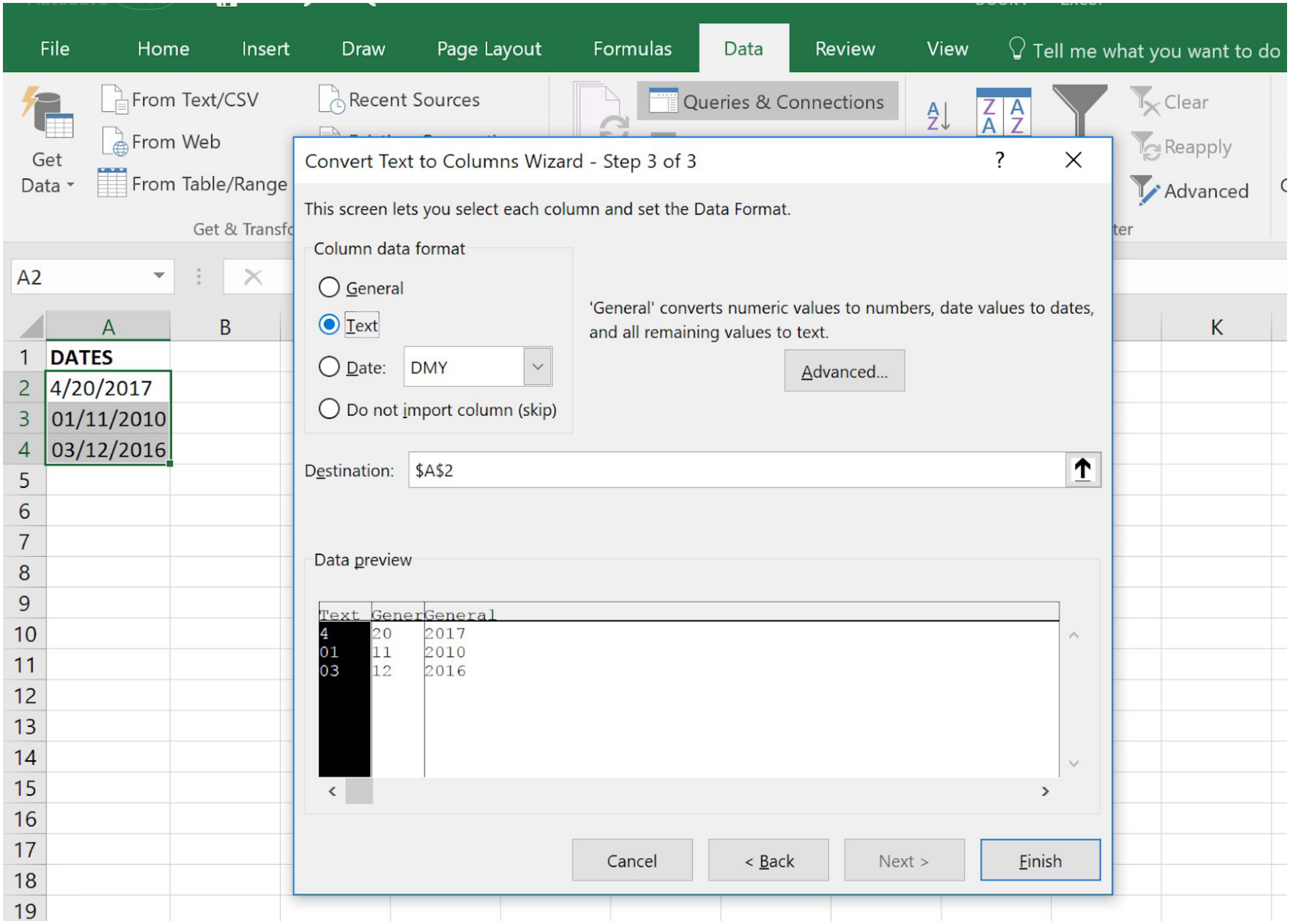
A2 | 4

	A	B	C	D	E	F
1	**DATES**					
2	4	20	2017			
3	01/01/1900	11	2010			
4	03/01/1900	12	2016			
5						
6						
7						
8						
9						

Table 70

Excel is messy when splitting dates. It auto-correct dates, taking them back to 1900s. In order for the dates to be split correctly, we need to manipulate the programme.

At the stage where you assign a type to the data, select text instead of date for each of the three date sections.



	A
1	DATES
2	4/20/2017
3	01/11/2010
4	03/12/2016

Convert Text to Columns Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

○ General

● Text

○ Date: DMY

○ Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

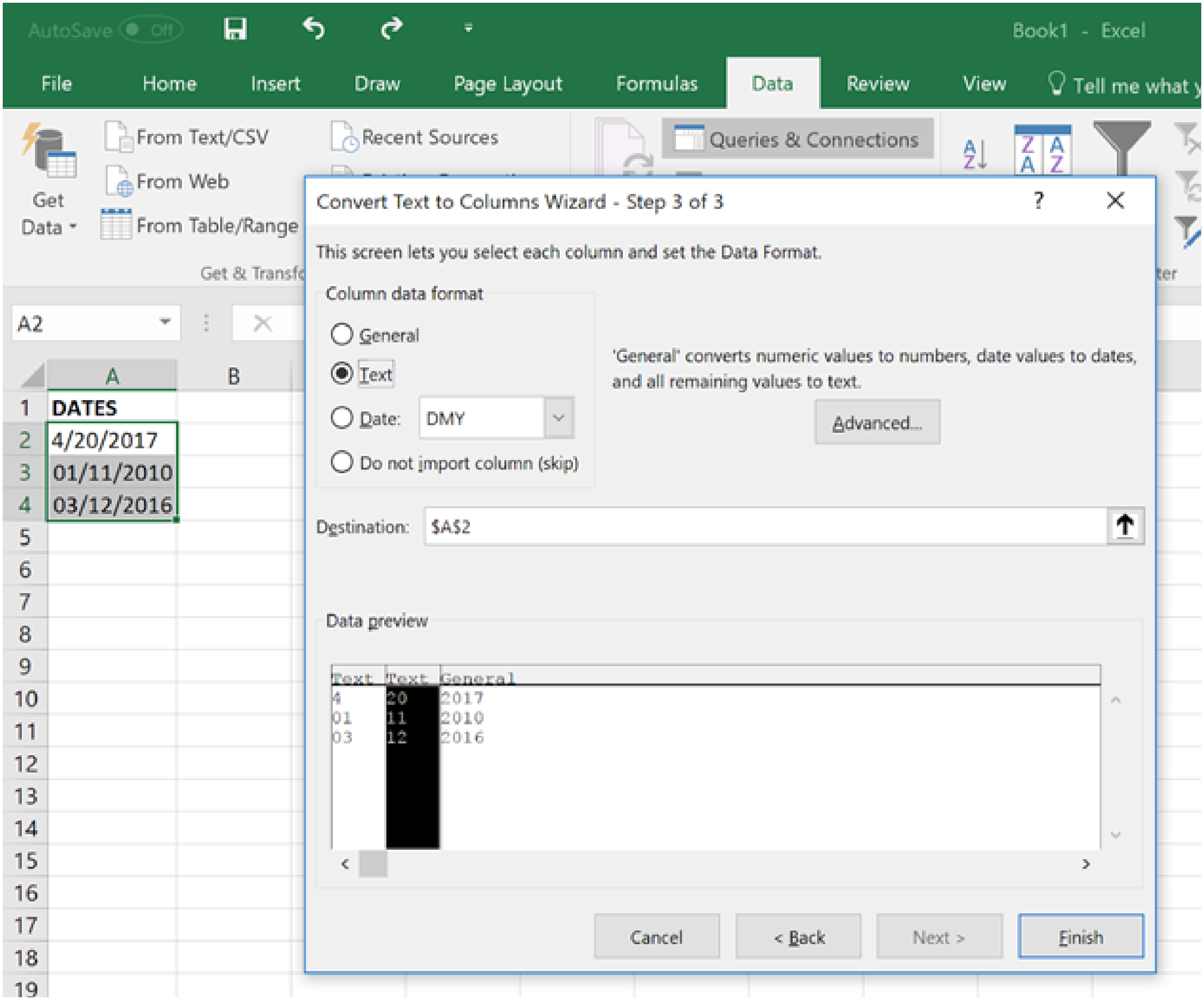
Advanced...

Destination: $A$2

Data preview

```
Text GenerGeneral
4    20   2017
01   11   2010
03   12   2016
```

Table 71



A2

	A	B
1	DATES	
2	4/20/2017	
3	01/11/2010	
4	03/12/2016	

Convert Text to Columns Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

○ General

● Text

○ Date: DMY

○ Do not import column (skip)

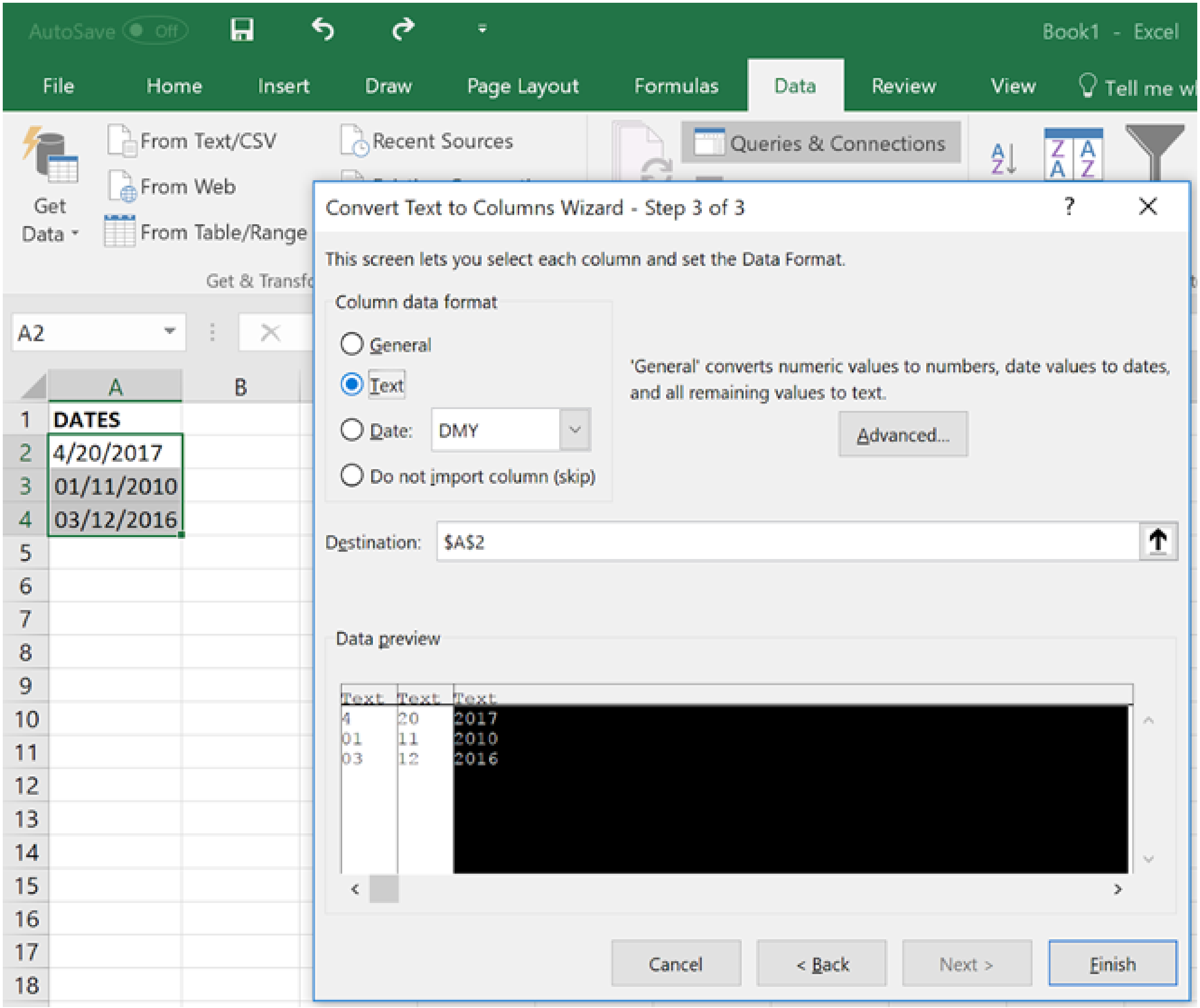
'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Destination: $A$2

Data preview

```
Text Text General
4    20   2017
01   11   2010
03   12   2016
```

Table 72



	A	B
1	DATES	
2	4/20/2017	
3	01/11/2010	
4	03/12/2016	

Convert Text to Columns Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

○ General

● Text

○ Date: DMY

○ Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

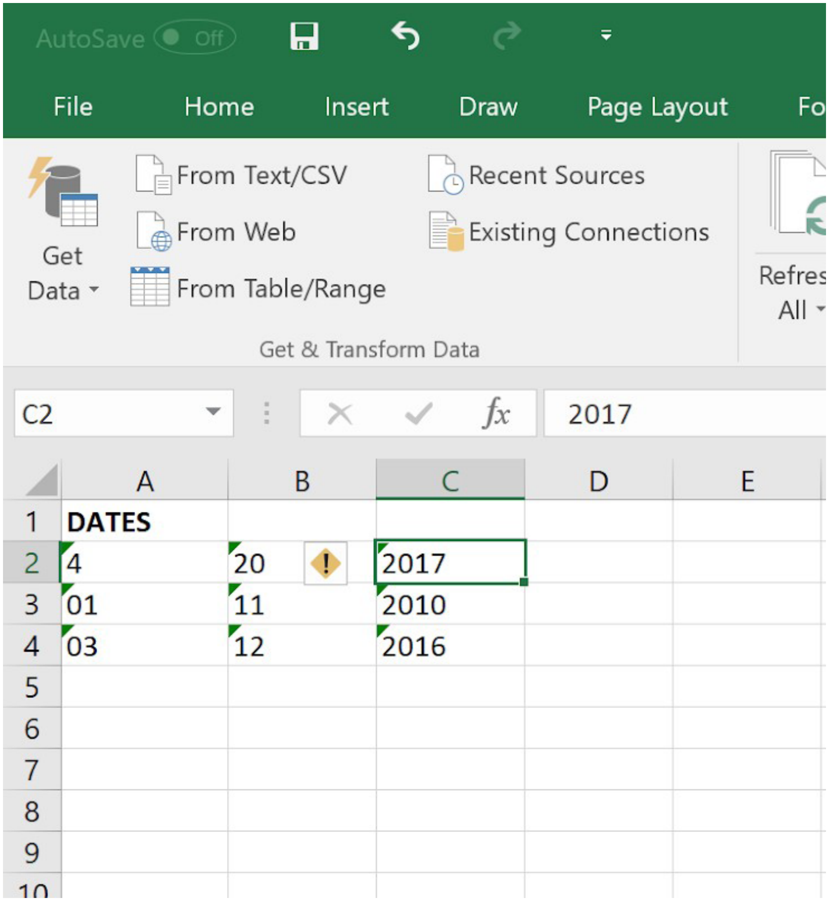
Destination: $A$2

Data preview

```
Text Text Text
4    20   2017
01   11   2010
03   12   2016
```

Table 73

This trick will result in an unproblematic date-to-columns split:



C2: 2017

	A	B	C	D	E
1	DATES				
2	4	20	2017		
3	01	11	2010		
4	03	12	2016		
5					
6					
7					
8					
9					
10					

Table 74

Numbers will appear formatted as text. Excel embeds a comment behind every single number, signaled by green corners, to tell you just this: “You cannot do maths on these numbers, they are stored as text!”.

To transform numbers stored as text into values, select your test-numbers, click on the yellow diamond containing an exclamation mark and select “Convert to number” (Table 75.).

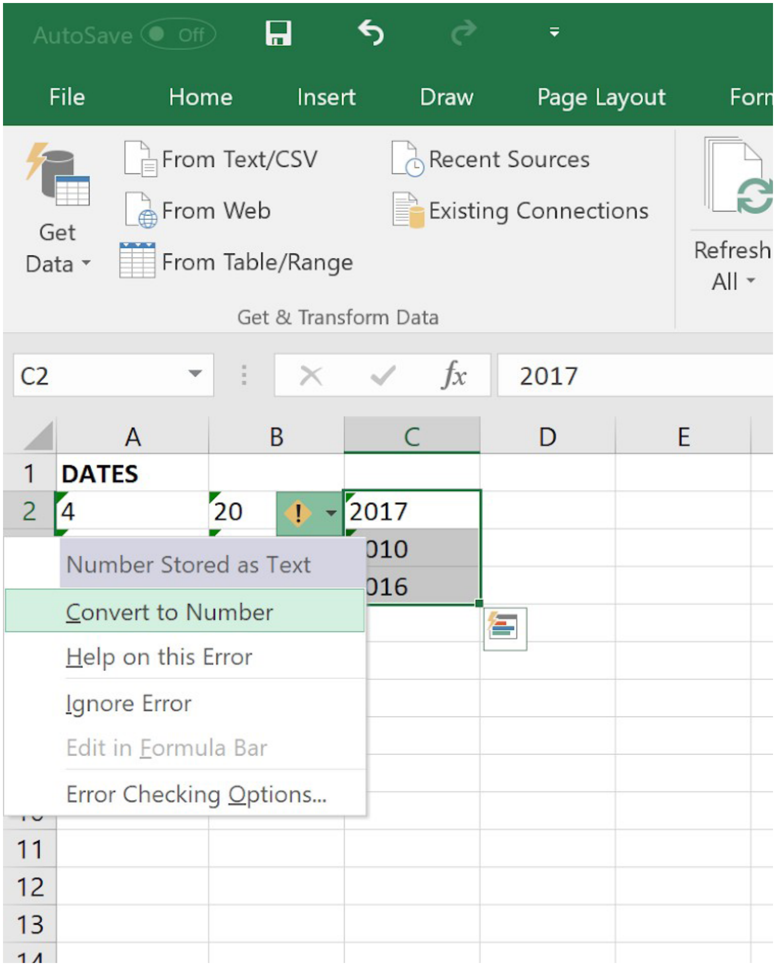


Table 75

- SPOTTING PROBLEMS

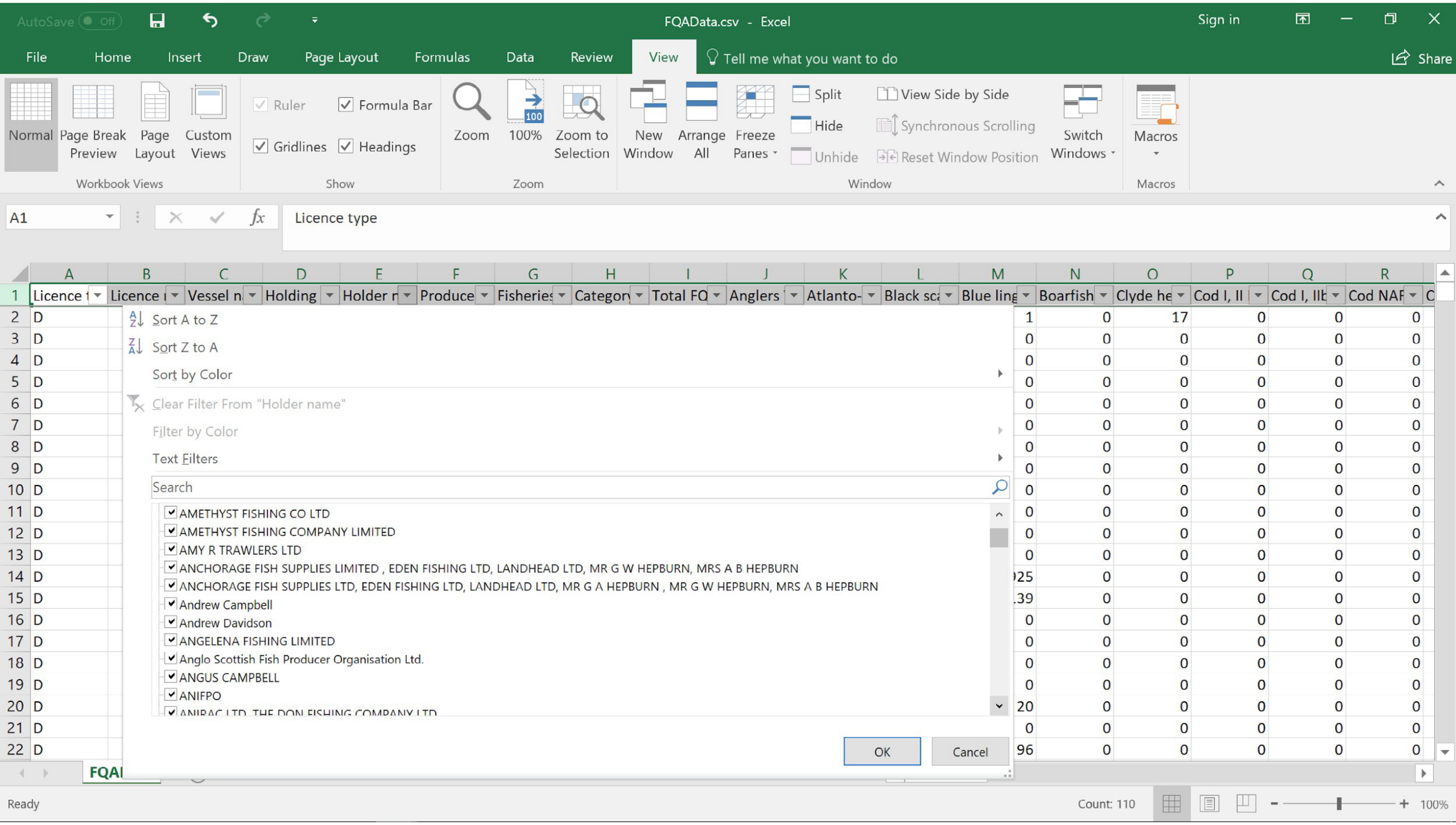
Data journalism is not for the lazy. It may speed up certain operations, but it requires a lot of pre-analysis labour. Just as an athlete needs to train for a marathon, as part of your data undertaking, do not overlook eyeballing: nothing beats it when it comes to getting to know your digital source and spotting errors.

When your data are big, this step can seem exhausting. But you shouldn't skip this step in your rush to helpers, like OpenRefine. Whether before of after automatised cleaning, DO:

- Use the SORT function to order each column containing text alphabetically or descending, depending on the data.

- Turn on the filters for each column and inspect the data in the filter window. Take notes of the issues you find
- Identify misspelt data, weird-looking dates or numbers that don't make sense.

You can use this to verify whether a name of a person or a company, for example, is entered several times in different ways.



A1 | Licence type

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Licence t	Licence	Vessel n	Holding	Holder n	Produce	Fisheries	Category	Total FQ	Anglers	Atlanto-	Black sca	Blue ling	Boarfish	Clyde he	Cod I, II	Cod I, IIb	Cod NAF
2	D												1	0	17	0	0	0
3	D												0	0	0	0	0	0
4	D												0	0	0	0	0	0
5	D												0	0	0	0	0	0
6	D												0	0	0	0	0	0
7	D												0	0	0	0	0	0
8	D												0	0	0	0	0	0
9	D												0	0	0	0	0	0
10	D												0	0	0	0	0	0
11	D												0	0	0	0	0	0
12	D												0	0	0	0	0	0
13	D												0	0	0	0	0	0
14	D												)25	0	0	0	0	0
15	D												.39	0	0	0	0	0
16	D												0	0	0	0	0	0
17	D												0	0	0	0	0	0
18	D												0	0	0	0	0	0
19	D												0	0	0	0	0	0
20	D												20	0	0	0	0	0
21	D												0	0	0	0	0	0
22	D												96	0	0	0	0	0

☑ AMETHYST FISHING CO LTD
☑ AMETHYST FISHING COMPANY LIMITED
☑ AMY R TRAWLERS LTD
☑ ANCHORAGE FISH SUPPLIES LIMITED , EDEN FISHING LTD, LANDHEAD LTD, MR G W HEPBURN, MRS A B HEPBURN
☑ ANCHORAGE FISH SUPPLIES LTD, EDEN FISHING LTD, LANDHEAD LTD, MR G A HEPBURN , MR G W HEPBURN, MRS A B HEPBURN
☑ Andrew Campbell
☑ Andrew Davidson
☑ ANGELENA FISHING LIMITED
☑ Anglo Scottish Fish Producer Organisation Ltd.
☑ ANGUS CAMPBELL
☑ ANIFPO
☑ ANIRAC LTD, THE DON FISHING COMPANY LTD

Table 76: Fish Quota Allocation data from the UK (DEFRA)

When you become confident in Excel, learn to use a more advanced tool that will help you clean duplicates is Open Refine - http://openrefine.org/. Although not infallible, it will allow you to clean large datasets faster. Pair it with eyeballing and patience and you will soon have a healthy, clean looking newborn dataset, ready for interrogation.

RECAP:

- When you start on your data journey, never work on the original dataset. You must always make a few copies and work from one of these. You don’t want to change the original, make a mistake and not be able to go back.
-
- Re-run your analysis several times. Best practice is to ensure you get the same results three times in a row. This is particularly important when you are the only data reporter in your team. If you have a colleague who can verify your analysis, they will also need to start from scratch, from an unedited copy. They less they know about the data, the better they will be at spotting things you’ve missed.
-
- Keep a log as you progress. Note the data source, the changes you make during the cleaning; the number of records before and after the data cleaning; mention the errors you’ve corrected, the columns you’ve added, the analysis carried out and its results. This will make it easy for you or your team to retrace your steps, for your editor to understand what you’ve done and for the lawyers (yours or theirs) to verify your findings.
-
- Field-test your results: do your findings make sense in the real world? Take a sample through field-reporting to see if the reality matches the data.
-
- Keep in mind that you may have to work with a reporter (at least!), maybe a developer, an illustrator, a communications specialist, a lawyer and several editors. Make sure you know the data and can explain your findings in your sleep!
-
- Most importantly, never assume you know anything about the data. N.E.V.E.R. Ask for explanations, fill in the gaps and understand what it can or cannot show.
-
- Check, check and check again. And don' t be afraid to verify your analysis against the source right before publication, or even before. It’s an elegant way of finding out about the problems in the data, the problems in the system reporting the data, and to give the source a right to comment.

CHAPTER VII

How to visualise

Making sense of large sets of data is demanding. When you have thousands of rows across a number of tables, or you're dealing with hundreds of events, dates, persons, companies, organisations and all of their relations, you can easily get lost and lose focus. Enter data visualisation.

Data visualisation tools are often used towards the end of a project as an afterthought to help your audience better understand the subject.

But these tools can be extremely useful from the very start of your project for you, the reporter, to properly understand the data you're dealing with in the first place.

Mapping out people and companies to better understand their relations, who's connected to whom, making timelines to get a deeper insight into the course of events, the cause and effect, drawing graphs to get on top of how numbers are changing will help you to figure out what's really going on, and ultimately, help you do better stories.

There are a number of free and open source programs to help you with that.

If you're working, for example, on a story about arms trade between certain countries, you'll have thousands of rows of information on who sold what, who's the buyer, who was the intermediary and what the price was. You need a tool that will immediately analyse those data and pinpoint the most important relations in your dataset. If you have all the names of all the companies involved, you could immediately spot who earned the most.

Gephi, for example, offers a powerful set of tools for graphs and networks analysis. It's data driven, meaning that you don't actually draw the visualisation yourself. It's drawn by the data you provide. The program makes the visualisation for you.

You can do the similar with Tableau Public, if you use Windows or Mac, or Google Fusion Tables, which is maybe more user friendly, but also more limited in comparison to Gephi.

In programs like Maltego Casefile, or Visual Investigative Scenarios (VIS) you can manually draw connections and relations between people, companies, documents, real estates, etc.

Casefile is intended with investigators in mind and you can install it on your computer. The Organized Crime and Corruption Reporting Project created VIS with investigative journalists in mind; and you have to use it online, after you sign up for it.

If your story requires geographical maps, two of the most used tools are Qgis and ArcGIS. Both have their flaws and advantages, but the biggest advantage of Qgis is that it's free and open source, while use of ArcGIS can be quite expensive.

If you want to try more advanced tools, you should look at the R-Project, a software environment for statistical computing and graphics, as well as D3, a JavaScript library for visualizing data.

Whatever tool you use, the most important thing is to be on top of the data. Always check and recheck every conclusion you draw from a certain data visualisation. Getting a new view of complex information can be very helpful, but can lead to misleading conclusions. For example, geographic profile maps that do not take into account population density often end up just reflecting where most people live.

You need to be aware of this problem when using data visualisation in presenting your story to the public.

The purpose of data visualisation is not to beautify the article. Its goal is to help your readers or viewers understand the story and grasp the significance and the scale of what you're saying.

You can use a number of different types of charts, diagrams, graphs and maps. What you use will depend on the type of data you want to present, their relations, the range and the change.

For each type of data and each type of relationship or concept, you have many available types of visualisations: do you need a map, or a timeline or a way to show some kind of hierarchy or range or movement? You can use this The Data Visualisation Catalogue [https://datavizcatalogue.com/] to help you decide the type of visualisation that best suits your story, but also to explore all the possibilities with data visualisation.

For example, stories with a strong chronological narrative are best presented with a timeline.
With tools like TimeLineCurator and Timeline.js you can easily translate events into a linear form with text, images, even videos, so your readers can figure out the exact chain of what happened.

One of the best ways to present data is with infographics. In one image, you can present a number of information, charts, timelines, maps, etc. There are a number of online tools to play around with, but if you want to do it properly, you'll have to learn to use advanced graphic tools such as Adobe graphic suite or open source variants like Inkscape, or hire a designer.

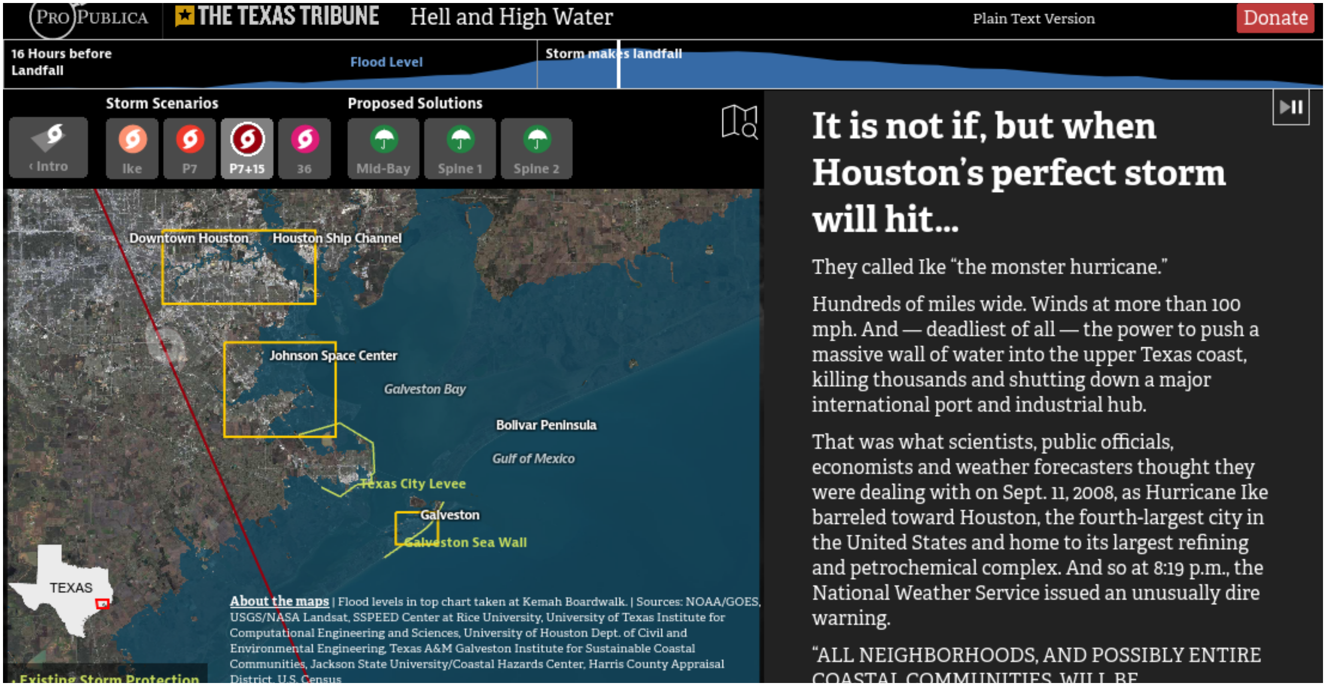
A general rule of thumb is to keep it short and simple, keep it focused and always have in mind the reader and what the reader needs to take in from your graphics.

The infographic is there to inform the reader or viewer in one glimpse. The reader needs to know immediately what the infographic is about and where to look for data that will support what your story said.

However, if you want really compelling graphics at the cutting edge of what today's technology can provide, consider getting a coder and the designer on board.

A good coder can bring your data to life and make it a dynamic. Your job will be to provide the data and to overview its accuracy, but cooperation with good designer and a good coder will provide you with tools to make a powerful impact.

Just look at the Texas Tribune and ProPublica's Hell And High Water or Brazil's Crime Data or Washington Post's article on US foreign aid budget.



About the maps | Flood levels in top chart taken at Kemah Boardwalk. | Sources: NOAA/GOES, USGS/NASA Landsat, SSPEED Center at Rice University, University of Texas Institute for Computational Engineering and Sciences, University of Houston Dept. of Civil and Environmental Engineering, Texas A&M Galveston Institute for Sustainable Coastal Communities, Jackson State University/Coastal Hazards Center, Harris County Appraisal District, U.S. Census

It is not if, but when Houston's perfect storm will hit...

They called Ike "the monster hurricane."

Hundreds of miles wide. Winds at more than 100 mph. And — deadliest of all — the power to push a massive wall of water into the upper Texas coast, killing thousands and shutting down a major international port and industrial hub.

That was what scientists, public officials, economists and weather forecasters thought they were dealing with on Sept. 11, 2008, as Hurricane Ike barreled toward Houston, the fourth-largest city in the United States and home to its largest refining and petrochemical complex. And so at 8:19 p.m., the National Weather Service issued an unusually dire warning.

"ALL NEIGHBORHOODS, AND POSSIBLY ENTIRE COASTAL COMMUNITIES WILL BE

The Texas Tribune and ProPublica's Hell And High Water

All these stories were produced by teams of reporters, programmers and designers, so, for big projects, consider adding new members to your team.

CHAPTER VIII

Further reading

Sources on Data Journalism

Precision Journalism: a reporter's introduction to Social Science Methods (4th edition), by Philip Meyer

Numbers in the newsroom, by Sarah Cohen
The Data Journalism Handbook, edited by Liliana Bounegru, Lucy Chambers and Jonathan Gray

Data Journalism or Computer Assisted Reporting, by Elena Egawhary and Cynthia O'Murchu **Statistics for Journalists**, by Connie St Louis

Data journalist toolkit - a list of the useful programs/aps for data journalists
The spreadsheet: choose Excel 2016 or newer. No other spreadsheet programme
Scrapers: Python
Statistical analysis: R
Intelligent text editors: Beautiful Soup or Sublime Text, among many

DB Browser for SQLite (for more advanced users)
Google Chrome browser: Scraper app for harvesting data and Fusion Tables for mapping

Data journalist lexicon - a dictionary of the key words

Data: information

Tabular data: information in rows and columns or that can be analysed using a rows-and-columns analysis tool.

Data Protection: policy and measure in place to protect sensitive information from being unjustifiably placed in the public domain. For a more localised description, check your country's legal texts' own definition.

Personal information – names, mobile phones, email addresses that contain names; date of birth and any item of information that is specific to and may help identify a person. Holding personal information of or on other people may require a special licence and extra security measures set in place to protect the information.

Data comptroller licence: the permit that allows anyone in the United Kingdom to legally hold personal data of other people. The license is released after filing an online request, filling in a security survey and paying a tax.

Variable: information that varies. In a spreadsheet context, a column.
Field: row
Record: row
Data entry: typing or copying information in a database
Flat file: .txt or .csv file
CSV: a file type in which the information arranged in columns is separated and delimited by a comma. It usually comprises only one sheet.
Txt: text file type; usually opens with NotePad
Xml: eXtensible Markup Language. It is “information wrapped in tags” (https://www.w3schools.com/xml/xml_whatis.asp)
String: text
Float: numbers
Data dictionary: metadata
Metadata: information about the data: how big is the dataset, how many records, how many rows, the explanation of the type of information and what different codes mean
Parsing data: to divide; to edit types of data into a preferred shape, form and type
Record layout: metadata

Future Learning

If you’re a working professional rather than a student, build your learning around your reporting needs. The spreadsheet is a good place to start from, not only because, as reporters, we have to deal with maths, but because you can apply it to your reporting straight away.

You should bring yourself to the level where you can analyse data, scrape websites and build small data visualisations for analysis (rather than for design).

Know your tools. You can make fast progress by learning one spreadsheet function a day. Here is a list with categories of functions - http://www.excel-easy.com/functions.html.

Start small: If you're covering crime, start with crime rates. Then add reports, sources and demographic data. Begin with the most basic data available and aim to get the most detailed granularity collected by the institution. This means getting down to the bare bones of the data collected. Often, you will get a hint of what is available from a department's fill-in forms, whether digital or in paper. Where there's a form, there is an archive, which is digitised to a greater or lesser extent, depending on the department's will and resources.

Ask how the data are collected to figure out to what level you can harvest.

Ask how the data are collected, so that you know all the details stored on the subject.

Use blank forms – collect them or photograph them everywhere you see them because every field on paper is a row or probably a column in a database.

Learn data protection, security and transparency laws. They will help you source information that nobody else has. This can be particularly difficult when you start out, especially if you're the first reporter asking for it. But so much for the better if you can fight a transparency corner.

Work with coders. Their skills can open new perspectives on what data are and what you can get from what is already in the public domain.

Learn Excel well and a coding language. Make them part of your everyday reporting.

Useful websites

For Albanian web sources please refer to Chapter 4 of this manual

International websites:

Land registry/cadastral info

- Croatia: http://www.katastar.hr/dgu/
- Serbia: http://katastar.rgz.gov.rs/KnWebPublic/
- Macedonia: http://www.katastar.gov.mk/prebmk_address/searchadd.aspx
- Bosnia: http://www.katastar.ba/pregled

- Montenegro: http://www.nekretnine.co.me/me/Katastarski_podaci.asp

Considerably more information will be available directly from the registry and some registries also provide extra services online for a fee.

Court documents

- www.bailii.org

The UK and Irish High Court records can be searched here and can prove extremely helpful given the number of international business disputes resolved in London.

- www.pacer.gov

Signing up to Pacer takes a little time, but is worth it as it provides you with access to the vast majority of US court records.

- https://courtconnect.courts.delaware.gov

Delaware, a popular US jurisdiction for offshore firms, provides its superior court records here.

- http://iapps.courts.state.ny.us/iscroll/

New York superior court records available here.
Washington DC: https://www.dccourts.gov/cco/maincase.jsf
Washington DC superior court records available here.

- http://www.eccourts.org/

The Eastern Caribbean Supreme Court covers many of the most common tax havens, including British Virgin Islands, and uploads judgements to its website which are searchable

- www.offshorealert.com/

Offshore Alert also provides paid-for access to court documents in Caymans, Bahamas and other tricky jurisdictions here http://www.offshorealert.com/

Freedom of Information

- www.asktheeu.org

Submit a Freedom of Information request to the EU through this website.

- whatdotheyknow.org

Submit a Freedom of Information to the UK institutions with this website.

- www.legalleaks.info

Legalleaks provides a guidance to submitting Freedom of Information requests across Europe.

Domain name information

- domaintools.com

One of the best known and comprehensive services to find current and historical information on websites, including the name of the registrant and contact information.

- whoisology.com

A free tool to look up domain name information.

- sameid.net/

Test links between websites using the same Analytics account (and some other code). Use is limited for non-paying customers.

- archive.org

Often called the Wayback Machine. Use this to find archived web pages and documents which have been removed from the internet.

Facebook

- researchclinic.net/graph.html

Internet research specialist Paul Myers's excellent guide to using Facebook Graph, Facebook's powerful search function (don't forget to set the language to US English).

- inteltechniques.com/OSINT/facebook.htm

A very useful search function which allows you to search Facebook Graph without using complicated commands.

- Graph.tips

Another Graph tool which provides a means of easily searching Facebook.

Social media searches

- www.echosec.net

EchoSEC is excellent at finding geo-located posts from Twitter and Flickr. A premium version offers more options.

Pinpoint geolocated YouTube videos.

- yomapic.com

Allows you to search Instagram and the Russian Facebook equivalent VK by location, although only in week-long slots. It does, however, allow you to find the geo-locations of all posts from a particular individual, meaning that you can trace all of their movements.

Useful databases

- app.enigma.io/

A treasure trove of a database with very useful documents on US imports (bills of lading), US company registration information and a host of other data.

- documentcloud.org/public/search/Kosovo

A site frequently used by journalists to archive documents with a handy search function.

- Birnsource.com

Balkan Investigative Reporting Network's own database of documents. It is due to include scraped public registers from the Balkans.

- World Bank Data: https://data.worldbank.org/

The WB collects a wide range of data from government agencies across the globe. The data is openly available for download.

- Eurostat: ec.europa.eu/eurostat

The European Union's statistical service. The data is openly available to download.

- World Health Organisation data: www.who.int/gho/database/en/

World health data and statistics openly available.

- United Nations data: data.un.org/

From energy to environment, from commodities to gender and demographics, there is data waiting to be processed.

- OECD data: https://data.oecd.org/

A broad catalog of datasets available - https://data.oecd.org/searchresults/?r=+f/type/indicators

- US Aid: https://www.usaid.gov/data

How and where does the U.S. spend it's financial aid funds? Find out here.

Tip:

Other countries - foreign, trade, defence, home etc. departments - produce data about your own country. Think tanks and some companies do, too. Go cross-border with your reporting and use their resources.

Lobbying

- http://www.fara.gov/

Find out which US lobbying firms are employed by foreign governments and who they have been lobbying.

- www.opensecrets.org/www.opensecrets.org/

A useful website for tracking corporate lobbying in the US.

- http://ec.europa.eu/transparencyregister/public/homePage.do

Find out who is lobbying who at the EU.

Tracking transport

- www.flightradar24.com

One of the best websites for tracking planes based on a range of parameters.

- marinetraffic.com

Track ships as they sail and investigate the history of a particular vessels.

- equasis.org

A free database of ship ownership.

Weapons databases

- www.sipri.org/databases

Stockholm International Peace Research Institute provides a very handy searchable database of weapons transfers.

- nisat.prio.org/trade-database/

The Norwegian Initiative on Small Arms Transfers, as its name suggests, is geared towards the small arms trade. A very handy tool for those investigating the arms trade.

- thearmstradetreaty.org

Official yearly country filings on arms imports and exports

Further CAR study

Short list of free online DDJ / CAR courses
MOOCs: Massive Open Online Courses - http://mooc.org/

Short list of favourite support forums
Investigative Reporters and Editors: www.ire.org – join the mailing list for free and live email forum support concerning any data, investigative journalism, ethical and editorial questions.

Data Journalism handbook

- https://github.com/DatajournalismBIRN/BIRN-Data-Journalism-Handbook

GETTING STARTED IN DATA JOURNALISM

Tirana, 2018