Local Government Training Institute (LGTI) Short Course Data Curriculum

Data Fundamentals Course

Module 4: Verifying, Cleaning and Analysing Data

Disclaimer: This work is developed by School of Data with funding from The World Bank Tanzania Data programme. For more information visit https://www.schoolofdata.org. We have adopted some of this content from the SudanData Learning Content, developed by Eva Constantaras.

Module 4: Verifying, Cleaning and Analysing Data

STUDENT WORKBOOK

In this module we will work through introductions to three stages of the data pipeline.

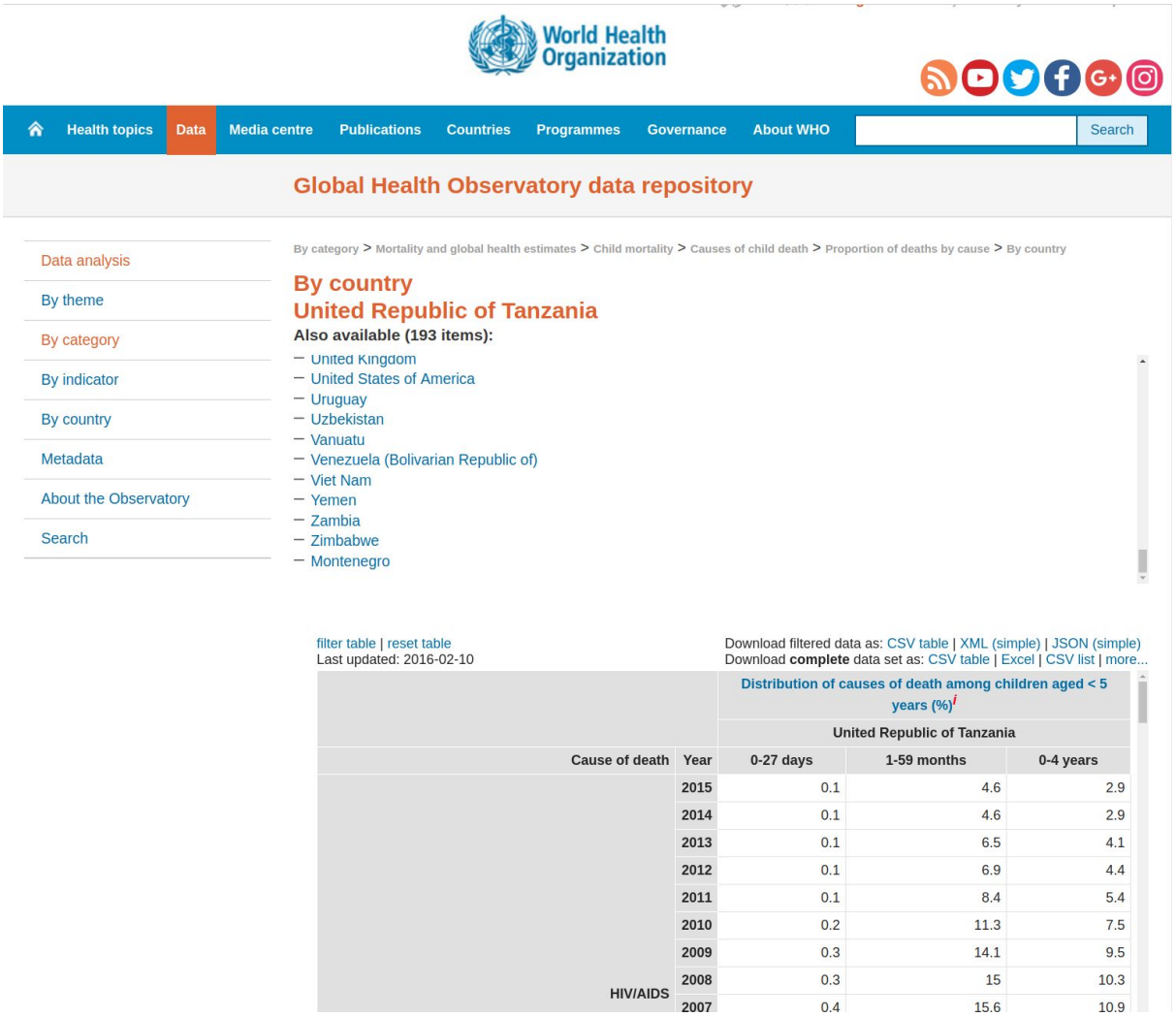
1. Data Verification:
2. Data Cleaning:

3. Data Analysis : Key to producing solid data driven analysis is the ability to evaluate data quality and apply basic statistical principles for accurate interpretation. This module will introduce basic concepts of data organization and cleaning as well as questions to help us evaluate the source of the data. Next we will look at basic calculations that can transform numbers into ratios, comparisons and rounded figures that audiences can more easily understand. Next, we will cover essential statistics to ensure data is interpreted correctly and that we recognize data manipulation. We will also go over one of the most powerful tools that a spreadsheet has to offer, the pivot table.

Lesson 1: Verifying Data

Data spreadsheets often hold a wealth of information in a very compact format. Before analyzing data, it is important to understand what the data is measuring and what all descriptions, labels and other contextual information means to ensure a correct interpretation of the data.

Often sector-specific data is produced by specialized professionals who use a lot of jargon and abbreviations to save space in data files. By doing a bit of research and following best practices when it comes to citing the source of the data, it is much easier to understand the data in context. Many datasets come with a codebook or a glossary that explain all the data labels and measurements.



Take a look at this dataset, which documents the causes of death of children under the age of five in the region:
http://apps.who.int/gho/data/view.main.ghe3002015-TZA?lang=en
Select options to see under five child deaths by country, then go to **Download complete data set as** as a **CSV list**.

This is a commonly used dataset to help evaluate the health of children in a country. When looking at a raw dataset like this it is important to thoroughly review all the information before starting your analysis. This section walks through some questions to ask about the data before using it.

Sources of Data	Understanding the Indicators	Units of Measure
Data Questions		
What organization produced this data? Where did the organization source the data from or are they the original source? Can I find an explanation of the data? Is there a link to data on the Spreadsheet? How old is the data?	What do the indicators mean? Can I look up definitions of indicators I don't understand What are the differences between the age categories? What is the difference between a rate or a Percentage? Is this data available using other measures from another source? What indicators are not included in this data that would provide more context?	What do the numbers mean? What is the unit of measure? What is the difference between a rate or a Percentage? Is this data available using other measures from another source?
Public Service Questions		
Does the data come from a trusted source? Is the data current enough	What would the public want to know about this data?	Does the unit of measure accurately put the data into context?

to be relevant? Can I find more information about the data source?	Do the indicators answer the questions I want to ask? What other information would explain the data	Does the unit of measure help the public gauge the risk? What text do I need to explain the units to my audience?

Responsible data hinges on the author verifying the validity of the data before reporting or presenting on it. Without being an expert in data science, there are a list of questions that can help casual data users identify signs of suspicious or untrustworthy data. Using this list to evaluate data each time will minimise mistakes. Being aware of common data reporting mistakes by others will also ensure responsible data use.

Essential Questions to Ask a Dataset

Below are a set of questions that can be used to interrogate your data:

- Where do these numbers come from?
    - What institution published this data?
    - Does this institution have a record for reliable data collection?
    - Is the report available on their website?
- Who recorded them?
    - Did the institution gather the data itself or did it outsource to another company?
    - What training did the employees undergo?
- How?
    - Was data collected by going to the primary source or was it gathered from a report?
    - Were these the results of a survey in which some people were documented or a census in which almost everyone is documented?
- For what purpose was this data collected?

    - Was this data collected to report to a funder to show that targets have been met?
    - Was the data collected by an outside auditor?
- How do we know it is complete?
    - Can we interview the data collectors?
    - Is there an explanation of the limitations of the data?
- What are the demographics?
    - Who was data collected about and who was left out?
    - Were rural and urban areas represented? Men and women? Abled and disabled?
- Is this the right way to quantify this issue?
    - What exactly is the data measuring and does your story match?
- Who is not included in these figures?
    - Was any group left out because of difficult of access? (such as disabled, people living in areas of violence)
- Who is going to look bad or lose money as a result of these numbers?
    - Was the study commissioned by an organization that is trying to prove that their projects are effective?
    - Was the study commissioned by an outspoken critic on the topic?
- Is the data consistent from day to day, or when collected by different people?
    - If collected over years, was the data collected by the same group following the same methodology?
- What arbitrary choices had to be made to generate the data?
    - How were decisions on things such as sample size made?
- Is the data consistent with other sources? Who has already analyzed it?
    - Are there other data sets on the same topic and to the results align?
- Does it have known flaws? Are there multiple versions?
    - Does the methodology explain potential errors in the data? Are there different copies in different places?

Lesson 2: Cleaning Data

Organizing Data

Before beginning data analysis to help answer our hypothesis and questions, we have to be able to understand the information we have. Data is organized under a specific set of standardized rules to make it easier for us to see our data. In our work, we will mostly be working with data tables in spreadsheets, not databases, but many of the same organizational principles apply.

A data table is a spreadsheet that's organized in such a way that the human eye can understand. To reach a conclusion, you can analyze the data table as a whole instead of going row by row. A database is a dataset organized into columns with each discrete data record in a different row. This organization system allows a computer to analyze the data and recognize similarities, allowing you to reach general conclusions about the data. Each column head is labeled by the category of data it contains and each row is a separate record. Each column indicates the type of data in that row, whether it be names, ages, gender, organization, etc.

Data Standardization

When working with a database, information can come from different sources, have incomplete fields, come in different structures, and include errors such a double entries or misspellings. This complicates the analysis process and although we recognize the mistakes, the computer does not know how to handle them.

Data standardization or cleaning is the process of cleaning the data and is an important step for data work. One of the first steps in cleaning data is to ensure that all the column headers are correct and complete and that the data type in each row matches the column header.

Many data cleaning processes allow us to clean the entire database with the same set of tools. If the database has addresses, dates, ages, measures, the first step is to decide a standard way to enter these fields into the database.

For example, this is a non-standardized date column

Date
12 February 2017
12/2/2017
2/12/2017
12/2/17
12/feb/2017

There is not a correct format as long as all of the dates are formatted the same and the computer understands the date format. It's important to choose a format that is most convenient for the whole database. In this case we decided on DD/MM/YY. So now our data after being cleaned looks like this:

Date
12/02/2017
12/02/2017
12/02/2017
12/02/2017
12/02/2017

Each date record is now in the identical format: DD/MM/YYYY

How to Standardize Data

The basic principle is to ensure that all the data is entered in the same format, often in all capital letters, without any extra spaces. For example, here is a data set:

Non-Standardised Data:

Name	Date of Birth	Address	Salary
Titus Jambo	16 April 1975	41 Julius Nyerere St. plot 23	TSh 2500000
Mary Yambesi	31/5/2000	6 Nkrumah st plot 12	4,000,000 Tanzanian Shillings

Standardised Data

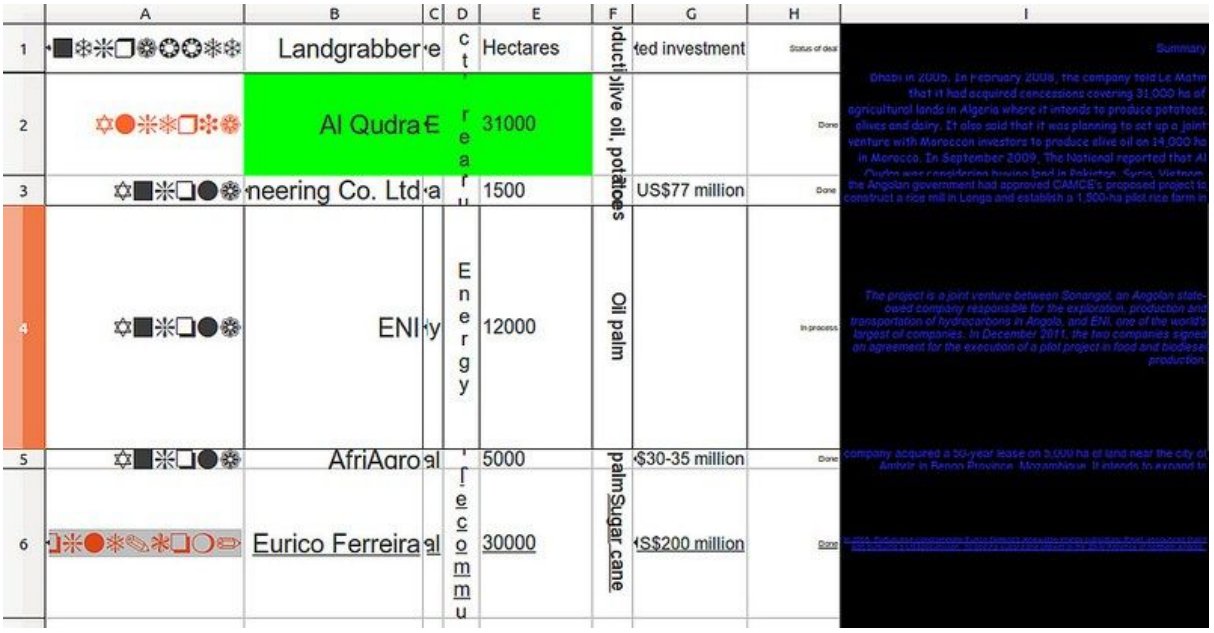
NAME	DATE OF BIRTH	ADDRESS	SALARY (TSh)
TITUS JAMBO	16/04/1975	41 JULIUS NYERERE STREET PLOT 23	2500000
MARY YAMBESI	31/05/2000	6 NKRUMAH STREET PLOT 12	4000000

There are spreadsheet functions and special programs specialized in data cleaning. Excel filters, search and replace, trim and other functions are sufficient for basic data cleaning. For more advanced cleaning, OpenRefine has features that can handle even the messiest data set. We will review these tools in the labs.

Data Formatting

In spreadsheets, formatting and data don't go together. If you have ever received data from someone, you have probably seen something that is incomprehensible as the below. Creators of spreadsheets spend a lot of time formatting data in ways that they find useful to them. They change the text size, colour in the cells, introduce borders and lines to give emphasis to the things they are interested in. However, when is data is shared, the formatting is often not useful and maybe not even understandable to the receiver of that data. Formatting may get in the way of the receiver to use the data.

Below is an example of the consequences of spreadsheet formatting:



In order to learn how to present your raw data to others so that they are able to understand and use it, we will first have to learn how to mess up your data. For these exercises, we will be using a spreadsheet, published on the Tanzanian Open Data Portal on "Dropout in Secondary Schools - Number of Dropouts in Secondary School by Sex and Region, 2015." You can download the data from the data portal here:

http://opendata.go.tz/dataset/idadi-ya-wanafunzi-walioacha-au-kuachishwa-shule-sekondari-kwa-mkoa/resource/1ba82df6-fb76-4dfd-8d73-b6c38897bb96

In his excellent essay “The Art of Spreadsheets” John Raffensperger lists 37 ways that you can hide data in a spreadsheet. Here are 10 of them:

- Do not share the file. This is the most common way of hiding information, and the most effective. We discussed this in module two when we learned about open data.
- Hide the sheet. In order to hide a sheet in a spreadsheet, you will need at least two sheets. Then, by clicking ‘Format’, ‘Sheet’, ‘Hide’, your sheet will be hidden from the person that you share that data with.
- Similarly, a row can be hidden by clicking ‘Format’, ‘Row’, ‘Hide’.
- A column can be hidden by clicking ‘Format’, ‘Column’, ‘Hide’.
- You can hide the cell and protect the sheet by clicking ‘Format’, ‘Cells’, ‘Protection’, ‘Hidden’, then ‘Tools’, ‘Protection’. This shows a display, but hides the formula: =if(1, “Peace!”, “Attack at dawn.”).
- Make the column too narrow: Format, Column, Width, 0.
- For formulas that are likely to be zero, use Tools, Options, View, and clear the Zero values box. For example: =IF(1, 0, “Attack at dawn.”).
- Use a formula that returns a blank: =IF(1, “ ”, “Attack at dawn.”).
- Create a complicated formula that displays the information, but format it as text (with Format, Cells, Number, Text, or just start the cell with a single quotation mark), so the formula is displayed rather than the output.
- Format the font with Wingdings: Format, Cells, Font, Wingdings. This displays unintelligible characters.

Using John Raffensperger’s list as inspiration, your task is to mess up the Tanzanian school Data data as much as possible. Marks will be awarded for:

- making the presentation just bad enough that someone using the data might be tempted to think they can still use it!
- the use of colour and font effects in ways that really offend the eye
- ingenuity in hiding bits of data in plain sight.

When you’re finished making a mess, consider how you would undo it and persuade others not to treat their data in this way.

The following reading will help you better understand formatting.

1. The help documentation for the most common spreadsheet tools outlines the different ways that you can change the way that a spreadsheet displays data. Depending on which spreadsheet you're using, visit these sites to refresh your memory: Libre Office – Formatting, Excel – Apply Cell Style, Google Docs – Cell Style
2. "The Art of Spreadsheets" by John Raffensperger has a list of 37 ways to hide information in spreadsheets.
3. Spreadsheet legend Chandoo looks at how to Boss proof your spreadsheets.
4. The nuclear option: how to remove all formatting from a spreadsheet (LibreOffice)

Whitespace in Spreadsheets

	A
1	Data
2	Your data
3	Your data
4	Your data
5	Your Data

Look at the data above, one error inserted you can clearly spot, but the others are far harder to spot with the naked eye:

- Added extra white spaces at the ends of entries.
- Tabs that are inserted at the ends of lines
- Line breaks and 'carriage returns', which you insert by pressing enter (or Ctrl-Enter).

They're called "non-printable" characters, and aren't displayed all the time in spreadsheets. But you will still feel their sinister presence as they seriously affect data analysis. **This is because spreadsheets treat these sorts of characters as real data.** Ignoring the column label, in the data above you can see four terms that are essentially the same. The spreadsheet, however, sees four different, distinct pieces of data. If you were trying to count how many times "Your Data" was mentioned, a spreadsheet would only show a single entry.

We therefore need to set traps to catch these non-printable characters. We can do the same in our spreadsheets. By the end of this section you should have:

- some knowledge about how non-printable characters cause errors in data
- tried out different functions and features of the spreadsheet that will remove them

Download and open this sample spreadsheet on your computer. In column A is the data from above, with different sorts of non-printable characters. In columns B-E are four easy methods of removing non-printable characters from your data using:

- the TRIM function (in column B)
- the CLEAN function (in column C)
- the TRIM and CLEAN functions .... together! (in column D)
- the "Paste Special" feature (in column E)

There are two easy ways to remove **whitespace** and **newlines** from a spreadsheet. Both are equally as effective.

Use the Find and Replace feature

Both whitespace and newlines can be "seen" by the spreadsheet.

1. Open the find/replace tool (Shortcut: Ctrl-H or Command-Shift-H).
2. Select "Regular expressions". This feature enables the spreadsheet to search for patterns, and not just specific characters.
3. In the input area for Find type **[:space:]** and click "Find". This is a regular expression that searches only for spaces that are at the end of the text in a cell.
4. Running this search will show you the cells in this worksheet that have one or more trailing whitespaces.
5. To remove the trailing whitespace that have been found, click "Find All". Make sure the input area for 'Replace with' is empty. Then click on "Replace All". Perform this operation until the spreadsheet tells you, "The search key was not found".

Use the TRIM and CLEAN functions

Trailing whitespace and newlines are common enough problems in spreadsheets that there are two specialised functions – clean and trim – that can be used to remove them. This is a little more detailed, so follow the steps carefully:

1. In your spreadsheet, the GRAIN dataset should be in 'Sheet1'. Create a new worksheet for your spreadsheet, called Sheet2.
2. In cell A1 of the new worksheet you have just created enter the following formula: **=CLEAN(TRIM(Sheet1.A1))** and press enter. This will take the content of cell A1 from Sheet1, which is your original data, and reproduce it in Sheet2 without any invisible character, new lines or trailing whitespace.
3. Find out the full data range of Sheet1: It will be A1 to I417. In Sheet2, select cell A1 and then copy it (Shortcut: Ctrl+C). In the same sheet select the range A1 to I417 and paste the formula into it (Shortcut: Ctrl+V). The complete dataset from Sheet1 will be reproduced in Sheet2, without any the problematic invisible characters.
4. To work further on this data, you will have to now remove the formulas you used to clean it. This can be done with the **Paste Special** operation. In Sheet2, select the complete dataset and copy it. Put the cursor in Cell A1, and then go to **Edit → Paste Special**. This enables you to choose what attributes of the cell you want to paste: we want to paste everything except the formulas.
5. Double click on any cell, and you will see that it just contains data, not a formula. If you like, run through the steps outlined in Problem 1 to make the text 'wrap' in cells, and adjust the column widths.

Fixing Numbers that Aren't Numbers

Three is not a number. Nor is a million. At least not when they are typed in as text in a cell in a spreadsheet. Your spreadsheet is your awkward and pedantic friend that needs everything to be precise, defined and consistent. If you don't do this, your spreadsheet will become confused. This is not dissimilar to what we discussed earlier around the difference between human and machine readability. For humans, 3 Million and 3000000

can be easily understood as the same thing but for your spreadsheet, these are two distinct entries. Here's how you and your spreadsheet might see things differently:

With spreadsheets, there are three basic data formats. Or in other words, three ways you can enter data into a spreadsheet cell:

- as a number, like 100.
- as text, like one hundred.
- as a formula, like =SUM(99,1), which creates a "calculated value".

To avoid creating confusion in your spreadsheet, follow two basic rules of thumb:

- Be consistent and **don't mix them up** in a column of data.
- Let the spreadsheet know **what type of data** is in each column.

If you ignore these rules, your spreadsheet will have difficulty doing all the useful things for which we entered the data in the first place, like adding, subtracting, counting, sorting, filtering and so on.

Download a copy of the GRAIN database of land grabs onto your computer. Load the spreadsheet and consider the following:

1. What sort of data type do you think the data in each column should be? Find out what type of data the spreadsheet thinks it is by right clicking on the column heading, selecting Format Cells and looking at the "Numbers tab". What do you see? Try choosing different options from the Category and Format lists and see what happens to the data.
2. Where there are some numbers in a column, can you add them up and see what happens? Use your common sense: does the sum look too big or too small? Does it produce an error? This may tell you there's something wrong with the data
3. Where there is text or numbers, try sorting in ascending order, and see what happens. Does it behave as you would expect?
4. Use Auto-Filter to display the distinct values in a column so you can see what sort of data is in the column. Does anything stand out to you as strange, or inconsistent? Can you see mixes of different data types?

After probing the data in these four different ways, what is your impression about how easy the data is to analyse using the spreadsheet? How could you improve the data?

Lesson 3: Analysing Data

Introduction to The Math you Need to Start

Math seems to be a scary thing for many people. If you tend to get scared by thinking about numbers and what to do with them, this lesson is for you. We will calm your fears and show you how much you can do – with counting, adding, and dividing numbers.

Common problems we address

When working with data, we run into a few common problems we will address with the methods we learn here:

- What is “normal”?
- What is different/special?
- How do I compare two different entities (e.g. countries)?.

What am I dealing with?

When working with multiple values in data, one of the key points of information you can get is how the data is distributed. This helps you figure out what you are dealing with.

Range

The first thing you want to look at is the **range** of your data: from where to where does your data stretch? Does it start with small numbers? Large numbers? Does it run from negative to positive? All this is essential information that will help you to deal with your data.

Looking at the range will also help you to find errors in your data. Let's say you are comparing body heights and that you've asked people to enter their size in centimetres. Say you find that your data's values range from 127 to 622. There is clearly a mistake: 6m tall people are very rare, and the likelihood for you to have caught one is relatively small. You should go back to your data and check it.

What do you have to do to find your range? Simply go through your data and find the **minimum** and **maximum** value—the lowest and the highest, respectively. In a spreadsheet, you can do this with the formulas `=MIN` and `=MAX`.

Let's say you have the following data (let's take body sizes):

163.1 162.2 210.5 201.0 188.7 182.6 153.0 173.5 146.6 148.0

Question: What is the range of your dataset?

Hint: find the lowest number (minimum) and the highest number (maximum).

Answer: Our range is from: 146.6 to 210.5

How many do we have?

The next important question we can ask ourselves is: how many things do we have? e.g. how many people did we survey? How many countries do we know things of? etc. This might seem simple, but for statistics, this is very important.

How do we get it? Simply count! In spreadsheets, this would be the formula `=COUNT` or `=COUNTA`.

In the data below: How many data points do we have?

163.1 162.2 210.5 201.0 188.7 182.6 153.0 173.5 146.6 148.0

10! Easy, isn't it? Simply by counting and looking at our range, we already have valuable information! Let's look at more.

Distribution

Next you want to look at is how the data is distributed. This is commonly done with a plot called a **histogram**. A histogram simply counts how often each value appears and proceeds from there.

So how do we do this? This is commonly done by **binning** data. What does this mean? Basically, we create **bins**, which are the ranges of numbers we care about.

Let's do this for the data we used above. Our data ranges from 146.6 to 210.5—let's create reasonable bins. Let's say we use 140-160, 160-180, 180-200, and 200-210. Then we go on and—surprise—count. How many values do we have between 140 and 160? How many between 160-180? And so on.

Result:

Bin	Number
140-160	3
160-180	3
180-200	2
200-220	2

Fantastic- let's draw this as a simple graphic:

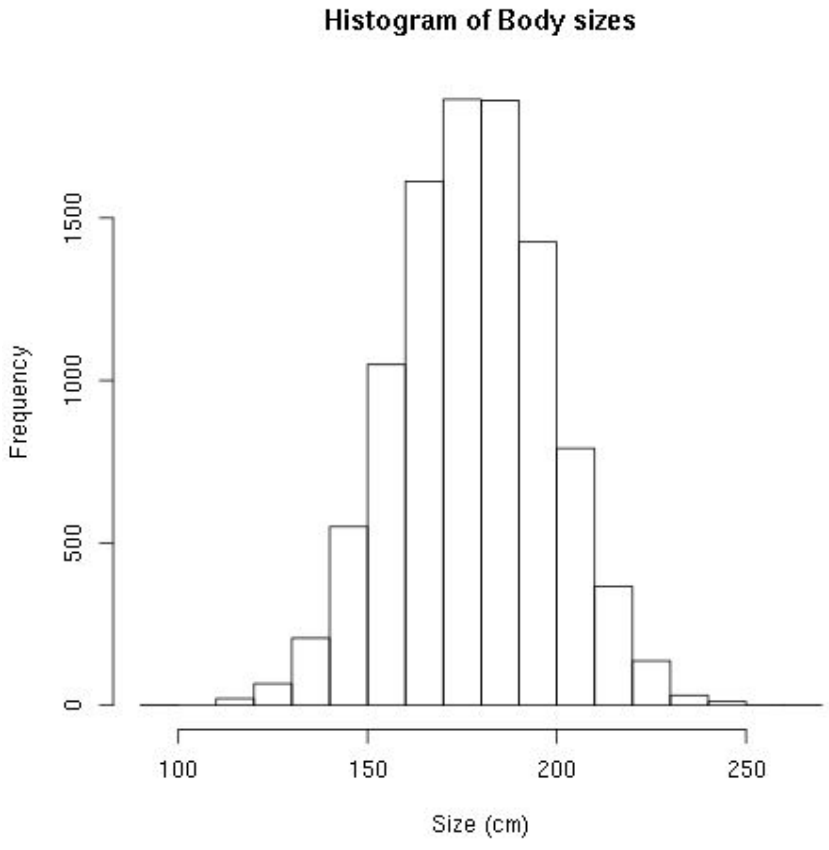
```
1    140 - 160:    ***
2    160 - 180:    ***
3    180 - 200:    **
```

4	200 - 220: **

This is all there is to histograms, which show us how our data is distributed.

While this doesn't tell us much if we have only ten values, this is super useful if we have more.

Things we want to look at: how many peaks do we see? Is there a single or multiple peaks? (Multiple peaks can tell us something about different groups present.) Is this a ***normal distribution*** where there is a clear peak and the sides are equally distributed (as below)?



Or is the distribution "skewed" (i.e. there is a peak on the left and a long tail to the right)?

The distribution tells you what kind of further descriptors are practical to use.

What is Normal?

One of the next questions often is: what is “normal”? What we mean by this is: how can I tell whether something is worth looking at? (We’ll figure out what is special further below.)

There are different ways to find this out.

Mean

The mean (or “average”) is the most commonly used way of looking at what is “normal”. We know it from newspaper articles claiming that the “average income” is increasing or decreasing and so on. But how do we calculate it? Quite simple: we sum up all the numbers we have and divide the result by the number of numbers we have. For example, the mean of 1, 2, 3, 4 = (1+2+3+4)/4 = 10/4 = 2.5.

Can you calculate the average of the heights we used before?

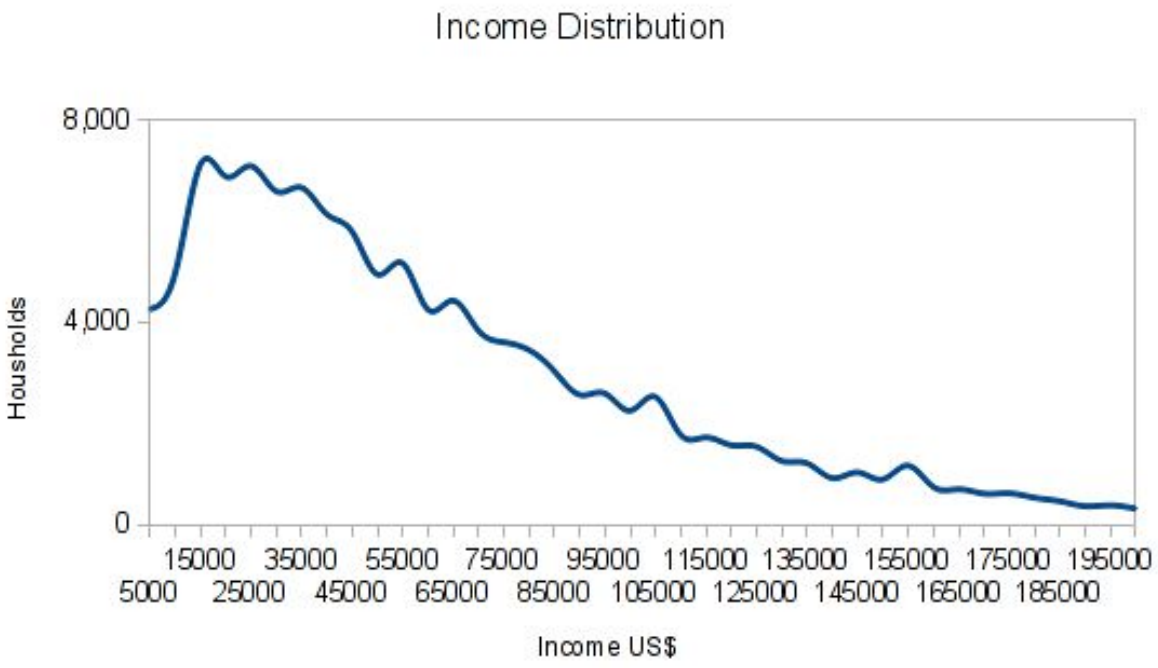
163.1 162.2 210.5 201.0 188.7 182.6 153.0 173.5 146.6 148.0

Answer: the mean is 172.92.

The mean is a great tool if your data is normally distributed. In that case, it tells you quite a bit about where the maximum of the distribution is and thus what you would perceive as normal.

Median

Let’s look at a different example. If we look at income distribution in countries, the distribution is not normal. It rather looks something like:



Now, if you look at the mean income, that might be quite a number. But if you earn less than the mean, you could still earn more than half of the population simply because the majority of the population earns so little. The median tells us this. To calculate the median, we simply sort the data we have and pick the value right in the middle.

Let's calculate the median of the following data:

162.0 159.1 169.9 191.3 195.9 139.8 186.0

First we'll sort the data (ascending or descending does not matter)

139.8 159.1 162.0 169.9 186.0 191.3 195.9

Then we'll pick the value right in the middle: 169.9 – this is our median!

What if we have an even number of values? We simply take the mean of the two values right in the middle.

Can you calculate the median of:

163.1 162.2 210.5 201.0 188.7 182.6 153.0 173.5 146.6 148.0 ?

Result: 168.3, the average of 163.1 and 173.5

Mode

Sometimes neither the mean or the median really tell us what we want to know. Let's look at a survey where we asked people how many siblings they have. They responded:

0, 1, 1, 1, 1, 2, 2, 2, 3, 5

well have a mean of 1.8 siblings, a median of 1.5 siblings. But what we really want to know is: How many siblings do the *most* people say they have. So we start counting.

1	0 - 1
2	1 - 4
3	2 - 3
4	3 - 1
5	5 - 1

We see 1 is the most frequent answer. This is the mode.

What do we do if the data is *not* discrete? We create ***bins*** as we did above, then count.

Sometimes you will not end up with a clear winner. There might be two different values that get the same number of counts. In case they are clearly separate, we call this a bi-modal distribution (or multi-modal in case it's more than two).

How big is the variation in the data?

The next thing we want to know is how big the variation is in our data. Two measures come in handy: the standard deviation and the median absolute deviation. The standard deviation comes with the mean and is very frequently used. The median absolute

deviation is less well known and would be the best to use if you're using the median already.

Standard Deviation

So let's look at the first thing, the standard deviation. It tells us how much, on average, data points are off the mean. We calculate it by summing up the square of the differences of the values and the mean, then dividing that sum by the number of measurements minus one, then taking the square root of that—did you even pay attention?

More importantly: If we do have a normal distribution – 68.27 percent of data points will fall within one standard deviation from the mean and 95.45 percent within 2 standard deviations from the mean. So it gives us a good idea where most of our data is. Hard to remember- this illustration shows it pretty clearly:

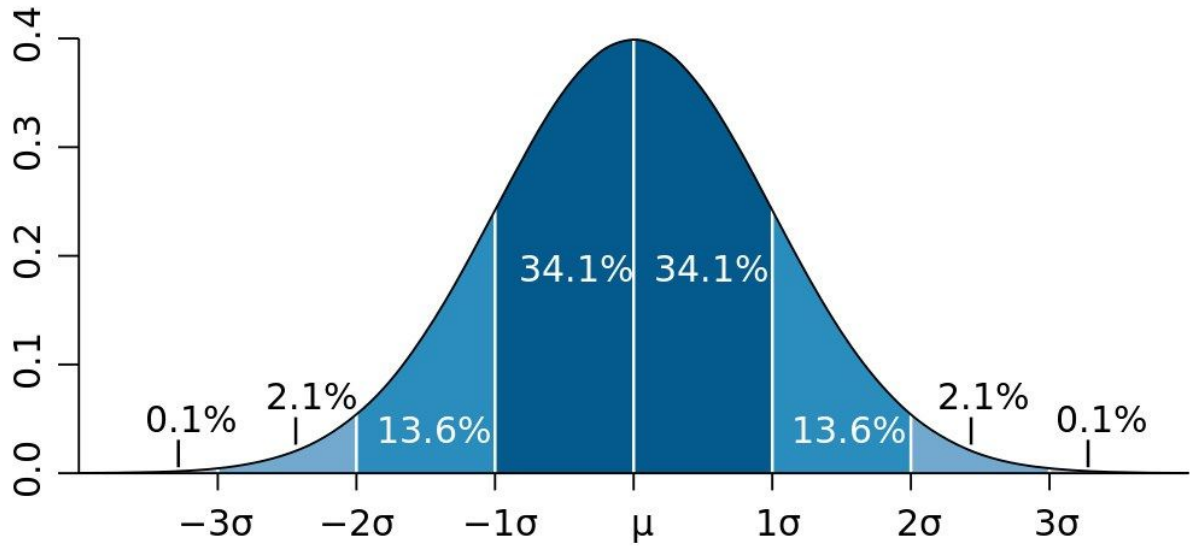


Image CC-by Wikipedia Member Mwtoews

Sounds complicated; let's do it.

Let's take our data above: 1, 2, 3, 4.

We already know the mean is 2.5.

Let's calculate the standard deviation:

Value	Difference to mean	Squared Difference
1	-1.5	2.25
2	-0.5	0.25
3	0.5	0.25
4	1.5	2.25

So we'll sum up the squared differences: that's 5.

We'll divide 5 by our number of data points minus 1.

5/(4-1)

That's 5/3.

And now we'll take the square root of that.

And we arrive at 1.291. – this means that 68.27% of measures will fall in this distance from the mean (assuming we do have a normal distribution).

Sounds complicated! It is a little. But remember, it's just adding stuff up, multiplying, and dividing. No big magic here. Luckily, if you need it, spreadsheets have a formula for this: `=STDEV`.

Median Absolute Deviation

As we said, the standard deviation works well when you can use the mean, since it's based on the mean. But what about when you use the median? Use the median absolute deviation. It works very similarly, but it's easier: you calculate your median and the

absolute difference between each value and the median. Then you calculate the median of the differences.

For example: our data is

1, 2, 3, 4, 5

The median is 3.

The differences are: 2 1 0 1 2 – sorted, 0 1 1 2 2.

The median absolute deviation is: 1.

Sounds feasible, right?

Normalizing Data – aka. comparing things

Once we have an idea of what we're talking about, let's figure out how we compare things.

Let's say we want to compare two countries that are quite different in several aspects. If we want to compare their GDPs, for example, we can do so, but it will not tell us anything useful. If, for example, one country is very big and another is very small, the bigger will have the higher GDP. Does that mean it's more productive? No. We'll have to compare them on equal footing. Usually this is done using something that tells us how big a country is; in countries, that's often population.

To compare productivity, we can divide the GDP by the population. This is called normalization. Now we can compare the GDP per capita. This is so commonly done that you will probably have heard of this indicator.

Another way of normalizing values is to use percentages. For example, if you want to compare what countries spend on health, it is great to normalize this by the GDP (e.g. to say this country spends a relatively great amount on health, whereas this other country doesn't). Or think of

elections: we commonly encounter percentages there (e.g. we calculate how many people voted for party A and divide this by the number of valid votes).

Finding out what's special – Z-Scores

Z or standard scores are a good way to figure out what is special. Let's say you have election results and want to find places that are interesting to report on. One thing you can do is figure out where a party has done exceptionally well. Z-Scores are great for this.

The Z-Score of a measurement is calculated as

(x – mean)/standard deviation.

The Z-Score gives you a value's distance in standard deviations from the mean. You can now set arbitrary limits and say: places where the votes with a Z-score higher than 2 are interesting because this means that extraordinarily many people voted for a specific party. A Z-score below -2 is also interesting because exceptionally few people voted. You can make this a little finer by looking at Z-scores for counties or regions (often there are regional differences and so on). Remember – 95.45 percent of measures fall into 2 standard deviations from the mean (if you have normally distributed data) this means: there is less than a 5 percent change for the Z score to be higher (or lower) than 2. 5% makes you pretty stand out. Not enough: 3 Standard deviations give you less than 1% of chance (99.73% of points are within 3 standard deviations from the mean in a normal distribution).

Overall, the Z-Score is a handy addition to your toolbox for figuring out what data values are different.

In this lesson, we talked about the basic math you need to understand quantitative data. We talked about how to figure out what you are dealing with (this is commonly called descriptive statistics). We went

on to figure out how we could compare apples to oranges (if we wanted to), and finally we looked at Z-Scores to be able to figure out what is special about something.

Simplifying Data

Let's explore basic data related to child health in Tanzania. Child health is a complex issue that often varies widely by geography, income, and other factors. Comparing basic data from a variety of sources can help develop a more complete picture of the major challenges facing child health in the country.

Simplifying Percentages

When people read a series of percentages, they have a hard time connecting with the subject of the data. So always try to simplify a percentage to a fraction or a population rate so that the audience can imagine how many people are affected by an issue. A common denominator, or the biggest number that fits evenly into both the percentage and the total (100%) can help simplify the numbers.

Understanding percentages

To understand how to convert percentages, take a look at these:

33% = $\frac{33}{100}$ = 3/10 (divide top and bottom by 3) = $\frac{1}{3}$

75% = $\frac{75}{100}$ (divide top and bottom by 25) = $\frac{3}{4}$

Example

Here are a few examples to further illustrate this point, a fraction or a population rate can help simplify the following statements about maternal health, which use a percentage.

Percentage	**Fraction**	**Population rate**
34.2% of pregnant women are anaemic.	One third of pregnant women are anaemic	One in three pregnant women anaemic.

74.1% of pregnant women receive prenatal care.	Three quarters of pregnant women receive prenatal care.	Three out of four pregnant women receive prenatal care.

Exercise: Simplifying Percentages

Simplify the following statements that contain percentages by using fractions or population rates:

- In Tanzania, 90% of newborns were protected against tetanus in 2016
- In Tanzania, 63.7% of births were attended by skilled health staff in 2016
- In Tanzania, health made up 15.9% of the national budget in 2012

Calculating Rates

Rates are also useful to simplify statements that quote percentages. For instance, to calculate what part of a population is affected by a condition, divide the total number of people by the number of people affected – the resulting number is your rate.

Example

> **Here is a statement:** In 2015, "13.2% of 1-59 month old baby deaths were caused by injury"

Let's convert the percentage to a rate:

- You understand that 13.2% = $\frac{13.2}{100}$
- We want the numerator to be 1 so let's divide the numerator and denominator by 13.2, which gives us $\frac{1}{7.57}$
- Based on this calculation we can say that 1 in 7.46 child deaths are caused by injury
- Since 7.57 is not an even number, multiply by 10, which gives $\frac{10}{75.7}$
- Now round it off to get $\frac{10}{75}$

"10 in every 75 children who die, die of injury"

Comparing Numbers

Now let's try comparing two statements that use percentages, and re-write them in a clear and comprehensible manner for readers.

For example, let's simplify these two statements:

- 49.7% of births are registered in rural areas
- 84.5% of births are registered in urban areas

To do this convert the percentages in the two statements to fractions:

$$49.7\% = \frac{50}{100} = \frac{5}{10} = \frac{1}{2}$$

$$84.\%5 = \frac{84.5}{100} = \frac{17}{20}$$

So now we can say:

- One out of two births are registered in rural areas
- 17 out of 20 births are registered in urban areas.

But this can still be further simplified by using the same denominator in both statements:

- 10 out of 20 births are registered in rural areas
- 17 out of 20 births are registered in urban areas

Rounding Off Numbers

Large, complex numbers can cause your audience to stop paying attention to your story. Use rounded, easy to understand numbers can ensure people understand the magnitude of the number without getting lost in all the digits:

Examples	**Rounded Off**
Tanzania's population is 55.57 million	Nearly 60 million people live in Tanzania
Dodoma's population is 2.084 million	Nearly two million people live in Dodoma
Tanzania's fertility rate is 5.08	Women in Tanzania have on average between five and six children.

Comparing Numbers: Exercise

Consider these statements and answer the following questions.

Statements

- XX.X% receiving diarrhea treatment (% of children under 5 receiving oral rehydration and continued feeding), 2010
- XX.X% receiving diarrhea treatment (% of children under 5 who received ORS packet), 2010
- XX.X% receiving diarrhea treatment (% of children under 5 receiving oral rehydration and continued feeding), 2015
- XX% receiving diarrhea treatment (% of children under 5 who received ORS packet), 2015

Questions

- Do you understand the indicators?
- Simplify the numbers to make comparisons:
    - Compare treatment in 2010
    - Compare treatment in 2015
    - Compare 2010 with 2015
- What other data do you need to tell the story of child diarrhea?

Spreadsheet Formulae

This tutorial uses Google spreadsheets to analyse data. Other spreadsheet programs work in a similar way – play around and see how they differ. There is sample data for this tutorial here: http://bit.ly/2CntL6g

A quick introduction to common spreadsheet symbols

Now that you have a sense of how spreadsheet formula work, here's a quick introduction to some of the most common formula symbols that you are likely to come across.

The symbols

These are all 'basic maths functions' – the kind of things you would find on a simple calculator.

=

Tells your spreadsheet that you are writing a formula. This is the first thing that should go in your formula cell. (NOTE: A spreadsheet assumes that *everything* that begins with an '=' is a formula... so be careful how you use it!)

+

Add

-

Subtract

*

Multiply (this would be ‘x’ on a calculator)

/

Divide (this would be ‘:’ on a calculator)

Tip: Get your symbols in the right order

It is worth remembering that basic maths rules about the order of functions apply. For example, the formula =3+5*2 will give you 13, NOT 16. If you want to change the order of function you’ll need parentheses: Formulas inside parentheses will be evaluated before any other formula. If you want the formula above to result in 16 you’ll need to type: =(3+5)*2

Have a go at using these formula in the ‘play sheet’ of your spreadsheet until you feel comfortable with them. You should find that they work pretty much as you would expect them to.

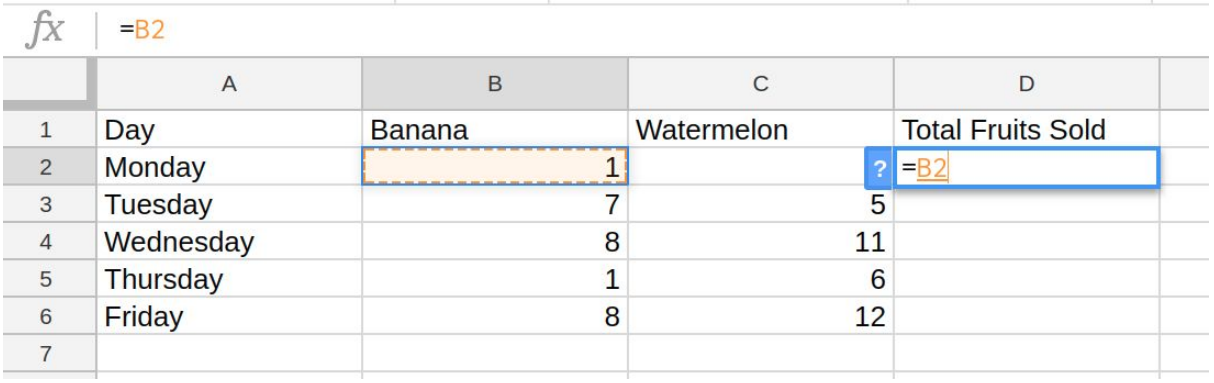
What if you wanted to add more numbers? You could always add them manually using + or you could use SUM a formula to sum up all the values in the given range. Let’s try to calculate how many bananas, watermelons and total fruit we sold during the week: Go to cell B7 and type =SUM(B2:B6) this will add the numbers of bananas.

Walkthrough: Using spreadsheets to add up values.

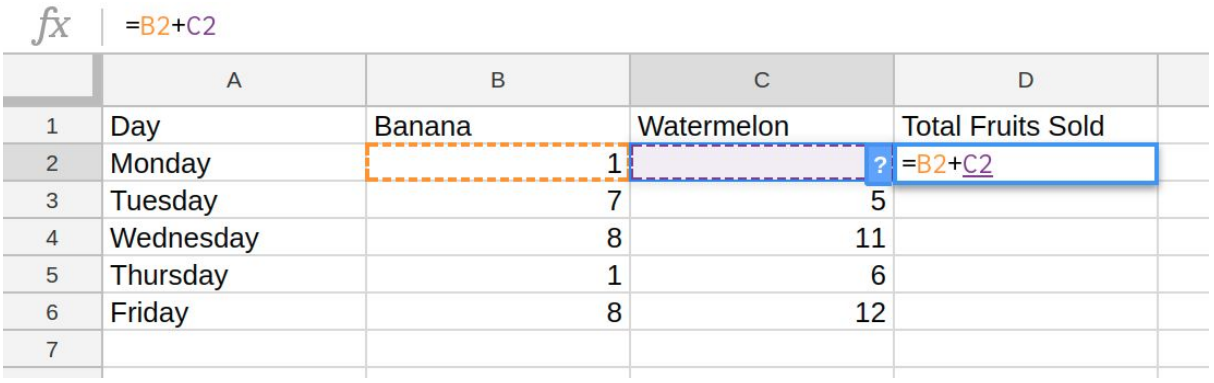
Let’s calculate the total of fruits sold.

1. Get the example data and create a copy.
2. To start, move to the first row.
3. Each formula in a spreadsheet starts with =

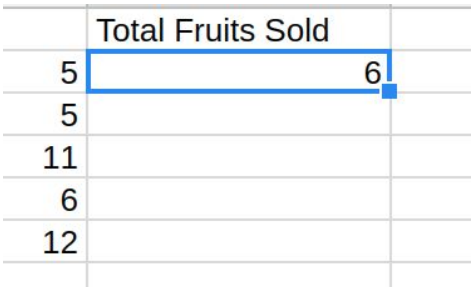
4. Enter = and select the first cell you want to add. Notice how the cell reference appears in the formula?



5. now type + and select the second cell you want to add



6. Press Enter or tab .
7. The formula disappears and is replaced by the value.

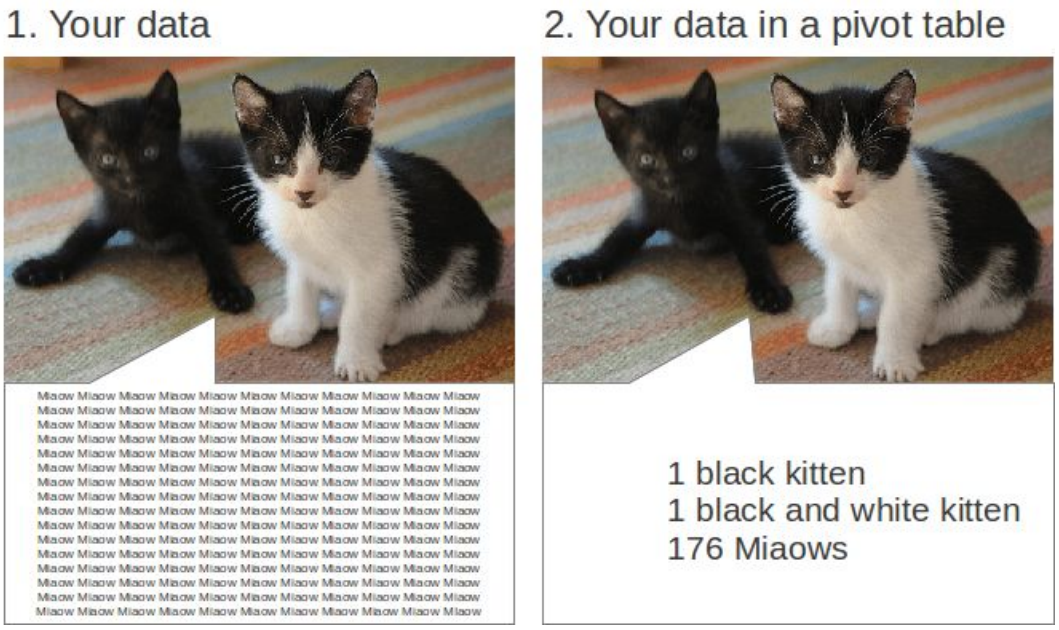


8. Try changing the number in one of the original cells (apples or plums) you should see the value in total update automatically.
9. You can type each formula individually, but it also possible to cut and paste or drag formulas across a range of cells.
10. Copy the formula you have just written (using ctrl + c) and paste it into the cell below (using ctrl + v), you will get the sum of the two numbers on the row below.
11. Alternatively click on the lower right corner of the cell (the blue square), and drag the formula down to the bottom of the column. Watch the 'total' column update. Feels like magic!

Task: Create a formula to calculate the total amount of bananas and watermelons sold during the week.

An Introduction to Pivot Tables

Now that you have learned some basic formulas and statistical concepts, it's time to introduce you to a powerful descriptive analysis tool for spreadsheets called pivot tables. In a nutshell, this is what they do:



"More Kittens", courtesy Hitchster. Some rights reserved: CC-BY 2.0. Adapted by Tactical Tech.

Pivot tables summarise complete datasets without you having to write formulae, create new columns, or arrange your data in any particular way. They enable you to combine data in ways that reveal the relationships that exist in the data, and show it to you in a new light. They don't change your data, they have a stack of useful functions built in and using them effectively can cut down on a lot of repetitive tasks, saving you time.

A combination of a number of spreadsheet functions – like sort and filter, and some formulae – work in a similar way to pivot tables in that they enable you to rearrange and pull out small bits of data more easily. Pivot tables do these things for the complete dataset, and present it to you so you can see it all.

Getting Started with Pivot Tables:

Take a look at the sample dataset about the selection of snacks on the author's kitchen shelves. It has data about their main ingredient, quantity, and the risk they pose to the author's health:

	A	B	C	D
1	**Snack**	**Ingredient**	**Quantity**	**Risk to Health**
2	Biscuit	Chocolate	10	High
3	Cake	Chocolate	3	High
4	Sandwich	Cheese	4	Medium
5	Biscuit	Jam	20	Low
6	Biscuit	Cream	20	Medium
7	Sandwich	Cucumber	3	Low
8	Cake	Orange	5	Low
9	Cake	Cream	2	High
10				

Start by building a pivot table using the data from the sample sheets:

- Select all the data. You can do this by selecting cell A1 and dragging the mouse to cell D9, or holding down Ctrl-A (Cmd-A on Apple Mac computers).
- With the data now selected choose Data → Pivot Table. This will create a new pivot table sheet.
- To the right, you will see a pivot table editor. In your original data, the columns are ‘snacks’, ‘ingredients’, ‘quantities’, ‘risk to health’. **Let’s pivot them, that is, turn a column into a row**.
- In your pivot table editor, click ‘add’ in the row section and then click snack. ’

	A	B
1	Filter	
2		
3	Snack	
4	Biscuit	
5	Cake	
6	Sandwich	
7	**Total Result**	
8		

So, what has happened to the data?

In the original data, “Biscuit” is mentioned 3 times: the Pivot table shows it only once. “Sandwich” is mentioned 2 times: the Pivot table shows it only once. And so on. The Pivot table has grouped and summarised the data in the Snack column of your raw dataset. It answers the question of what different types of snack are included in the data.

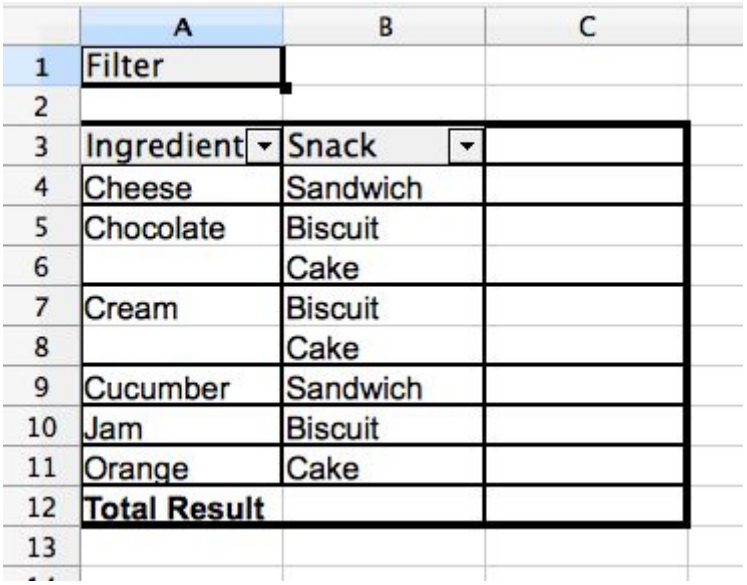
Pivot tables can be created with more than one Row Field. Using the sample dataset, let’s choose another row of data to add:

1. Go back to your pivot table editor, this time, click ‘add’ and then ‘Ingredient’ in the row field area. The data produced by the pivot table will now look different:

	A	B	C
1	Filter		
2			
3	Snack	Ingredient	
4	Biscuit	Chocolate	
5		Cream	
6		Jam	
7	Cake	Chocolate	
8		Cream	
9		Orange	
10	Sandwich	Cheese	
11		Cucumber	
12	**Total Result**		
13			
14			

What’s happened this time? In the same way as before, the pivot table has also grouped and summarised the data about ‘Ingredients’. The great thing about this is that it has grouped the data about ingredients to show them for each type of snack. We can turn this around to give another view, from the perspective of the ingredient, not the treat.

To do this, go back to your pivot table editor. Drag the 'ingredients' above 'snacks' in the row field. This will re-create the pivot table with the new layout. This is how the data in it will look:



	A	B	C
1	Filter		
2			
3	Ingredient	Snack	
4	Cheese	Sandwich	
5	Chocolate	Biscuit	
6		Cake	
7	Cream	Biscuit	
8		Cake	
9	Cucumber	Sandwich	
10	Jam	Biscuit	
11	Orange	Cake	
12	**Total Result**		
13			

In this pivot table the groups of values are arranged in a different way. Rather than showing the ingredients that go into each snack, this shows the types of snack that contain a particular ingredient. Now, let's try it out on a larger dataset where we can see the value of a pivot table more dramatically.

For this exercise, we are going to use a GRAIN dataset that can be downloaded from from the Datahub.

Before you get started, it's important to spend a bit of time familiarising yourself with this dataset. A good (but more time consuming) way of doing this is to work through the the steps that we did above in the cleaning section of the module.

If you don't have time right now, the basics of this dataset are as below:

- the dataset has been made by GRAIN, a research and advocacy organisation which works to support biodiversity and sustainable, community-controlled food systems.

- each row of the dataset contains details about the sale of a huge amount of agricultural land in a country, often in the global south.
- the columns contain data about the names of investors and the countries where they are based, the country where the land deal has been carried out, the size of the land deal, and the amount of money invested to purchase the land, and whether the deal went ahead.

To create a pivot table in the GRAIN dataset the steps are the same:

1. Select the complete dataset (from cell A1 to I417). Remember that if you don't select data, it won't be included in the pivot table.
2. From the top menu, select **Data → Pivot Table.** This will create a new sheet with your pivot table.

The GRAIN dataset has nine columns of data. We will be add different combinations of fields into the **Row Fields** part of the pivot table to answer specific questions.

We'll walk through one of the questions to get you started: **"In which countries has land been acquired?"**

1. The data you need to answer this is in column A, labelled 'Landgrabbed'.
2. Select the complete dataset. Go to Data → Pivot Table. This will create a pivot table.
3. In your pivot table editor, go to the row field area and click 'add'. Select 'landgrabbed'
4. The data in this pivot table will be as below, a list of countries:

We can now build on this list to increase our understanding of what is in the dataset. For example, in the pivot table editor, return to the Row Field area and click 'add' and select 'Landgrabber'. Now we can answer the question: "Which companies have acquired land in which countries?"

Here's the first few rows of data that you'll get in the pivot table:

	A	B	C
1	Filter		
2			
3	Landgrabbed	Landgrabber	
4	Algeria	Al Qudra	
5	Angola	AfriAgro	
6		CAMC Engineering	Co. Ltd
7		ENI	
8		Eurico Ferreira	
9		Lonrho	
10		Quifel Natural Resources	
11	Argentina	Adecoagro	
12		AgroGeneration	
13		Al-Khorayef Group	

For extra points, try reversing the order of the tiles and creating a pivot table from that layout. It will show you the same data but arranged around the investor (the 'Landgrabber') rather than the country where land has been acquired. Here's a bit of the data you'll get from that layout:

Now you're pretty much an expert, here are a few more questions that you can answer by adding in data to the Row Fields of a pivot table. Have a go at these:

1. In which countries are investors based (their base)?
2. In which countries are investors based, and where did they acquire land?
3. Which investors are working in which sectors?
4. Which investors are working in which sectors, and how did they use the land they purchased? Tip: data on how acquired land was used is in the column called 'Production'.
5. Which companies work in which sectors, broken down by base country?
6. What are the names of investors that have made similar sized land acquisitions, and in which countries did they make those acquisitions?
7. What were similar sized land acquisitions used for, and in which country, and what is the status of the deal

Make a pivot table even more useful by adding ‘values’ or ‘data fields’

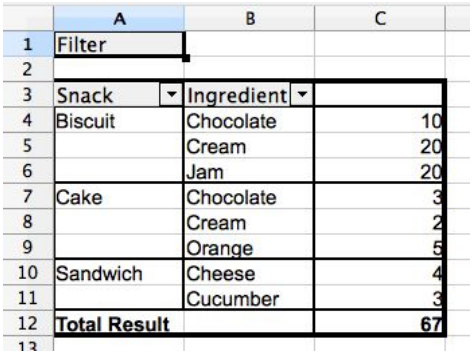
In Section 1 we tried out building sorted and grouped lists that can use your data to answer questions. But what else can a pivot table do? In this section we’ll look at how the ‘values” part of the pivot table works.

To get started, build a pivot table of the different types of snacks again, as outlined in Section 1 above. This time however, we’ll add in a “value” that will calculate how many of each type of snack there. To do this, in your pivot table editor click ‘add’ in the value field area and select quantity.

So, what’s happened?

The pivot table has grouped and summarised the data on the types of snacks, which you put into a Row Field. The data on the quantity of snacks – which you put in the value field – has been added up to create a total for each type of snack. Let’s add in another Row Field, just as we did in Section 1, and see what it tells us:

1. Go back to your pivot table editor and click ‘add’ in the row field area to select ‘ingredient’
2. The data shown will change again. This time, the types of snack are sub-grouped by the sort of ingredient, along with the quantities:



	A	B	C
1	Filter		
2			
3	Snack	Ingredient	
4	Biscuit	Chocolate	10
5		Cream	20
6		Jam	20
7	Cake	Chocolate	3
8		Cream	2
9		Orange	5
10	Sandwich	Cheese	4
11		Cucumber	3
12	**Total Result**		**67**

We can apply the same steps to the GRAIN dataset on landgrabbing to create more useful summary views of the data. For example, let's find out how much land was reported as being acquired in each country:

You will do this by first clicking 'add' in the row field area and selecting 'landgrabbed' and then clicking 'add' in the value field area and selecting 'hectares'. The effect is the same as with the example above in the short task. The data in the Row Field is summarised and grouped to show a list of countries, without duplicates. The data in the value field has been added up to give a total figure for each country. Here are some sample rows of what this pivot table will produce:

As before, we can continue to ask questions of the data by adding in different value fields. The data above shows the amount of land acquired in each country. Add in 'Status of deal' as a row field to refine this picture even further and show which deals are done, in process, proposed and so on.

Using your knowledge of choosing Row Fields, and now adding Value Fields, try creating pivot tables which show the following:

- a little profile for each investor, showing the countries where they have acquired land, and the size of the land area they have acquired e.g. a pivot table that shows Adecoagro reportedly made deals in Argentina for 242000 ha, Brazil for 165000 ha and Uruguay for 8600 ha.
- The total amount that that each investor has invested to acquire land e.g. this pivot table should show that Saxonian Estates reportedly made investments totalling USD 7.7 million.
- The amount of land that has been acquired, organised by investment sector e.g. this pivot table will show that 160,000 ha have been acquired by investors that work in the telecommunications sector.
- The amount of investment made, organised by the size of the land acquired, showing the country where the land was acquired e.g. the pivot table you make here should be able to quickly show us that land deals of 6000 ha were made in Australia for USD 335 million, in Russia for USD 39 million and in Nigeria where there is no record of the amount invested.

- Bonus features: change which aspects of data are shown

The fields that you add to pivot tables have two useful features you should know about. We'll provide a walkthrough below, but here's an overview:

The data that we have positioned in the Value Field of the pivot table is often just added up – that is, where there are multiple values they are added together to show the "sum" (this is the default setting). However, the pivot table can show this data differently by:

- picking out the highest (the "max") or lowest (the "min") values from a list.
- giving a total of the number of values (the "count").
- calculating the data as a percentage or running total

As with the Row Fields, you can have more than one value field in a pivot table. This means you can display different aspects of the same data next to each other.

Here's an example pivot table layout that demonstrates both these features.

To get there, build your pivot table as usual but this time, add 'hectares' into the value field four different times. Under 'summarise by', select 'sum' for the first, 'count' for the second, 'min' for the third and 'max' for the fourth.

This pivot table will show four pieces of data for each country where land has been acquired: the number (or 'count') of deals where the amount of land is recorded, the largest acquisition ('max'), the smallest acquisition ('min') and the total amount of land ('sum').

Adding columns to pivot tables

In the previous sections, we looked at how to add row fields and data fields to your pivot tables. We also looked at how to sort and filter data in pivot tables, and how to

adapt the display of data to pick out the largest and smallest values in a list. In this section, we'll add the final basic component: Column Fields.

After building nearly 30 pivot tables in this course, we're sure you're now getting the hang of this. The next step is to choose the data that can be a Column Field in your pivot table.

Take as a starting point the pivot table you made about snacks in Section 2 (this should have the two rows 'snacks' and 'ingredients' and the value 'quantity'). Now, return to the pivot table editor and click the 'add' button in the Column Fields area, select 'risk to health'. Your new pivot table will look like this:

	A	B	C	D	E	F
1	Filter					
2						
3	Sum - Quantity		Risk to He			
4	Snack	Ingredient	High	Low	Medium	Total Result
5	Biscuit	Chocolate	10			10
6		Cream			20	20
7		Jam		20		20
8	Cake	Chocolate	3			3
9		Cream	2			2
10		Orange		5		5
11	Sandwich	Cheese			4	4
12		Cucumber		3		3
13	**Total Result**		**15**	**28**	**24**	**67**
14						

The effect of adding the Column Field is to further sub-group the data. The version that includes columns enables you to see at a glance which the high risk snacks are, what they are made of, and how many of them there are. Better avoid chocolate biscuits and cream cake!

Now, lets return to the GRAIN dataset, we can see how adding this final dimension affects how the data is shown. Create a basic pivot table which shows how much land ('Hectares') has been acquired in each country ('Landgrabbed'). This time include the 'Status of deal' field in the Column Fields area of the pivot table layout editor:

The effect should be quite predictable for you by now. The pivot table will give an overview of the total amounts of land acquired for each country, broken down by the status of the deal. The 'Status of deal' field is a fairly convenient field to add to the Column Fields area. When summarised by the pivot table it has only five distinct categories. This means it fits easily into the screen area! Something like 'Production', which has over 100 categories, would not be as easy to view.

Have a go at changing the layout of the pivot table whilst keeping 'Status of deal' as a column:

- Replace the tile in the Row Fields with 'Landgrabber' (ie. the investor) and change the tile in Value Fields to 'Projected Investment' (ie. the amount paid for land). This shows how much money investors have tied up in done deals, deals that are signed, proposed and so on.
- Replace the Row Fields with 'Sector' and the Value Fields with a count of the number of investors. We covered how to do this in Section 2's

Adding charts to pivot tables

You can chart data that is produced from a pivot table. Having both a summary of the data, and a chart is a way of further exploring and coming to an understanding of the data you have. Using the GRAIN data, here's a simple example of how it works.

Once again, create a basic pivot table which shows the amount of land purchased in each country: drag 'Landgrabbed' into the Row Fields and 'Hectares' into the Value Fields.

First, sort the data so the largest land deal is at the top of the list:

1. In the row field, under 'order' select ascending and 'sort by' SUM of Hectare
2. Second, add a chart: Select cells A1 through B67.

3. In the top menu, go to Insert → Chart. We want to great a column chart with Landgrabbed on the x-axis.
4. Third, refine the chart to show only the 10 countries where the most land has been acquired. By hiding rows in the pivot table, we can change what data is shown in the chart. Select rows 12 to 67. In the top menu. The chart will change to the below, which is far easier to grasp.