Understand your data and extract the insights that matter

A step-by-step guide to analysing and visualising data for international development

Akvo.

Design, Capture, Understand, Act

#withAkvo

Contents

Introduction 04

1 | What is a data journey and how will it boost the impact of your programme? 06

2 | How to design your data analysis for effective data use 09

3 | Quality data for quality decision making: How to clean your data in three key steps 12

4 | Predict the future, understand the past: The four types of data analysis 18

5 | How to visualise your data in a clear and compelling way 24

Conclusion 32

About Akvo 33

Credits 33

Why this eBook?

This eBook guides international development professionals through every step in the understand phase of the data journey,[1] resulting in valuable insights for data informed decision making. This eBook will help you to:

- Capture your data with analysis in mind
- Prepare, clean and transform your data
- Conduct analyses for a variety of data purposes
- Apply data science concepts to your data
- Generate insights for decision making

Following the data journey

The understand phase comes after the design and capture phases of the data journey. In the data journey philosophy, a thorough design phase is central to the success of any programme. The design phase is conducted at the beginning of a programme but revisited and revised throughout to ensure accuracy. Before you enter into the capture phase of your data journey, the design phase helps you to gain a comprehensive overview of what data is needed as well as the opportunities, challenges and objectives or your programme. In the capture phase, you can set your team up for smooth and reliable data collection at scale and monitor data quality as it comes in. For a deep dive into the design and capture phases of the data journey, download our eBooks:

- Design data-driven programmes that deliver results effectively. [2]
- Capture reliable data in the international development sector.[3]

Making sense of your data

It's easy to think that once you've captured all your data, you can jump right in and start doing cool visualisations and progressions. The truth is, no matter where your data comes from, you will always need to clean it. This is especially true for complex

1 https://akvo.org/our-approach/

2 https://datajourney.akvo.org/ebook-design-data-driven-development-programmes-that-deliver-results-effectively

3 https://datajourney.akvo.org/ebook-capture-reliable-data-in-the-international-development-sector

Design

Gain clarity on the context of your programme, the problem you are trying to solve, the data you need, and the roles and responsibilities of each partner.

Capture

Capture reliable and high quality data from the start. Monitor your data collection to ensure accuracy and track progress.

Understand

Extract the insights that matter. Clean, analyse and visualise your data and turn it into valuable information and knowledge.

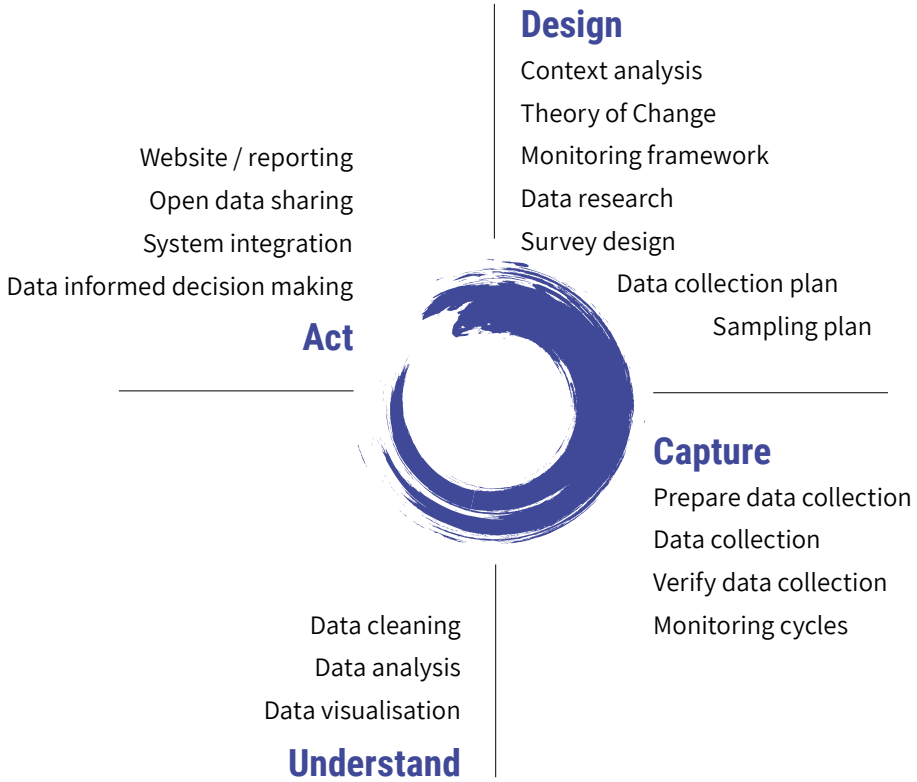
Act

Share insights with the relevant people, generate dialogue, encourage decision making and continuously improve your work.

development programmes, in which surveys have been conducted in multiple countries. Once your data is clean, you need to start asking questions to bring context and value to your data - the more you enrich your data through analysis, the more insightful and valuable your data will be. This is the essence of the understand phase - making sense of your data, translating it into information, and using that information to generate insights for decision making.

Making sense of data collected in multiple programmes in different countries and turning it into unified, digestible and insightful reports can be a real challenge. This eBook will take you through every step of the understand phase, from cleaning your data to visualising it in clear and compelling ways. This way, you can extract the insights that matter and turn your data into valuable information and knowledge.

1 | What is a data journey and how will it boost the impact of your programme?



The data journey methodology consists of four phases: design, capture, understand and act. They form the starting point for organisations to ensure data is used to contribute to lasting and inclusive impact. These phases aren't always consecutive or prescriptive, there may be some overlap, and it may be necessary to go back to a previous phase due to findings at a later stage.

What does each data journey phase consist of?

Design

Gain clarity on the context of your programme, the problem you are trying to solve, the results you are trying to achieve, the partners you'll work with, the data you need to monitor progress, and the roles and responsibilities of each partner. The design phase should enable you to define your data needs and prepare for a smooth data capture process.

Questions to consider include:

- What is the context you are operating in? Who is involved and what is their role?
- Which impact do you want to achieve and which outcomes will contribute to it?
- Which data will you need for which purpose, which data already exists and which do you still need to collect?
- What does the optimal survey design look like to ensure success?
- Which sampling plan fits best and is most cost-effective?

Capture

Collect relevant, high quality data from the start. Implement your data collection plan and track progress. Questions to consider include:

- Are the tools and skills and logistical plan in place to commence data collection?
- How can you verify and ensure the quality of your data on the go?
- How best do you organise monitoring cycles of repeated data collection?

Understand

Clean, analyse and visualise your data and turn it into valuable information. Extract the insights that you can act upon. In the understand phase of your programme, you can generate information which can be interpreted to extract insights.

- What data sources are you planning to combine? Is your data clean and ready for analysis?
- How can you extract insights from your data?
- How will you visualise the data and ensure effective data storytelling?[1]

Act

Share insights with the relevant people, generate dialogue, encourage decision making and continuously improve your work. In the act phase, you'll share your data to influence change.

- How will you share with the key audiences?
- Which systems does the data need to be stored in?
- How can you amplify your insights and create lasting impact?

For international development professionals, following this data journey will ensure smooth and successful implementation of your programme, allowing you to focus on capturing data that matters. In this eBook, we're focusing on the understand phase of the data journey.

1 https://akvo.org/blog/five-tips-for-effective-data-storytelling/

Round out your data journey knowledge

Download the eBook now

Visit our knowledge library

2 | How to design your data analysis for effective data use

Let's say you're building your own house. You start by designing it - the style, the materials, the structure. By the time you get the actual building, you have a pretty good idea of how it's going to look, how much time it's going to take, and what it's going to cost you.

The same goes for a data collection project. Before you go out into the field, it's important to reflect closely on what it is you want to achieve. Before you get to the actual analysis of your data, you should have a clear idea of what type of analysis you'll be applying, why, and what information you expect to get out of it. Without this prep work, you'll simply be collecting bits of information that don't necessarily provide the answers you need to achieve your project's goals. By following the data journey methodology, you will have already covered this in the design phase.[1] However, it's good practice to revisit these steps throughout your project to ensure you're on track. Below, we've summarised the four key steps to analysing your data before you've captured it.

Identify your key questions

Asking the right questions in a clear and concise way can be the most challenging task in any data analysis project. The questions you ask should first of all stem from the problem, impact, and outcomes you have already identified in the design phase of your programme. Let's use an example. As more households make the transition to improved water sources, lower income households that

1 The design phase is covered in depth in our eBook: Design data-driven programmes that deliver results effectively

aren't able to make the transition face increased risk of contamination, particularly from faecal matter. Studies show that this increased risk is accompanied by poor sanitation levels. This project therefore wants to know the status of sanitation in X regions of [country] in order to improve WASH for all citizens.

Thus, a data collection project is commissioned with the following objectives:

- Discover the relationship between income and water and sanitation levels of households per division and district.
- Predict which households are most likely to be vulnerable to water related diseases.

Based on these objectives, questions should be defined. A common pitfall here is that the questions asked are not clearly defined or translated into data analytics concepts. This requires thinking about the specific type of information that is required to answer the question and therefore to achieve the objective. When translating questions into data analytics concepts, the type of objective will most often dictate the type of data, preparation and modelling techniques that will be considered. For example, a prediction goal has very different implications than a descriptive goal, and inferring correlation requires fewer restrictions than inferring causal relations. All of these concepts will be described in detail in the following chapters.

Conduct desk research to see whether data exists

Once you've clarified the questions that you need to ask and the data that you need to answer them, the next step is to see whether secondary data is available. Secondary data is existing data from external sources, as opposed to primary data, which is data that you collect yourself by conducting a survey. In practice, it's rare for an existing dataset to deliver exactly what you are looking for, but secondary data can be useful for a number of reasons. Besides saving resources on collecting data that already exists, it can be used to:

- Triangulate data: This is a method of data validation whereby you cross verify your findings using data from other sources.
- Add layers to your data: For example, you can add population data to your data on water points in order to add context to your data.
- Generate ideas about the design of your survey: Secondary data can be used to inspire and inform your own survey design.
- Set a baseline for your data collection: For example, if water point or WASH data has already been collected for a particular region, you can use that as a baseline for your own data collection.

Open data portals are a fantastic resource for those looking to find secondary data for their programmes. At Open Data Inception,[2] you can find over 2,600 open data portals worldwide.
One disadvantage of secondary data is that it isn't always clear how the data was collected. If the sample strategy isn't clearly documented, for example, then you can't know whether the data is representative. For a more thorough guide on how to conduct data research, check out this blog: How to conduct data research in four steps.[3]

Decide on the output of the data (data types)

The next step is to consider the data types you're going to work with and how you plan on deploying them in your project. The most common distinction is between structured and unstructured data. Most development projects use structured data, which can be subdivided into categorical and numerical data (figure 01). The type of data you work with will determine the type of data analytics you apply.

2 https://opendatainception.io/
3 https://datajourney.akvo.org/blog/how-to-conduct-data-research-for-your-programme-in-four-steps

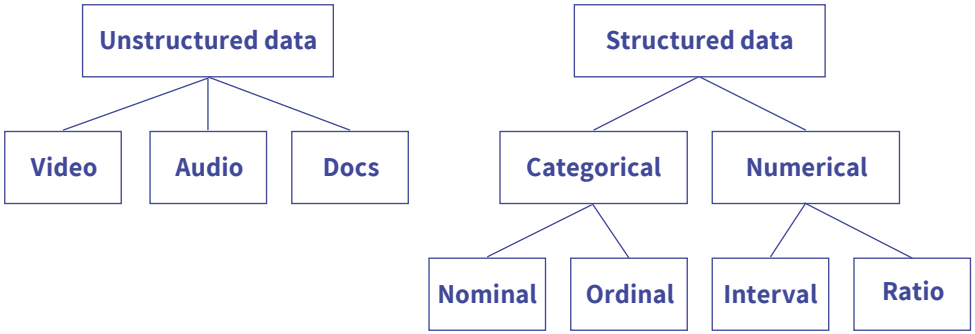


Figure 01: Data types

Categorical data: categorical data describes categories or groups.

- Nominal data: A nominal scale describes a variable with categories that do not have a natural order or ranking. Examples of nominal variables include gender, zip code, eye colour, race and political party.
- Ordinal data: An ordinal scale is one where the order matters, but not the difference between values. Examples of ordinal variables include socioeconomic status (low income, middle income, high income), education level (primary school, high school, BS, MS, PhD), income level (less than 50K, 50K-100K, over 100K) or distance to the water source (<50, 100-200, 200+).

Numerical data: numerical data represents numbers.

- Interval data: An interval scale is one where there is order and the difference between the two values is meaningful. Examples of interval variables include temperature, IQ, pH or any other water quality parameter.
- Ratio data: A ratio variable has all the properties of an interval variable, but has a true zero. In other words, there can be no negative numerical value in ratio data. Examples of ratio variables include age, weight, height, income earned in a week, years of education, and number of children.

Decide which type of analytics you want to apply

There are numerous ways to make sense out of data. The method you choose will depend on the questions you're asking. When these questions are more about explaining what and why things have happened, descriptive and diagnostic analytics will come in handy. If the questions relate more to what could possibly happen in the future, predictive and prescriptive analytics are more appropriate. These types of data analytics are further elaborated upon in chapter four.

Figure 02: A breakdown of the different types of data analytics

3 | Quality data for quality decision making: How to clean your data in three key steps

"Even a good cook will fail with bad ingredients."
— Unknown

However impressive you are in the kitchen, it's unlikely that you'll be able to pull off a fantastic meal with old or expired ingredients. The same goes for data - your analysis will only be as good as the data you use. If the data that goes in is *garbage*, the result of your analysis will also be *garbage* - this is sometimes referred to as GIGO (garbage in, garbage out). Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.[1]

Cleaning your data

There are three key steps to data cleaning. In this chapter, we'll walk you through the three steps, and then we'll show you what happens when you analyse a dataset before you've cleaned it and after you've cleaned it. We'll wrap up with some tips on how to minimise data cleaning.

Safely store your data

Before making any changes, make sure the original raw data is stored safely with a good backup strategy in place. For example, save the original dataset both locally and in a secure online storage.

Tidy up your data

Tidy datasets are easy to manipulate, model and visualise, and have a specific structure:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

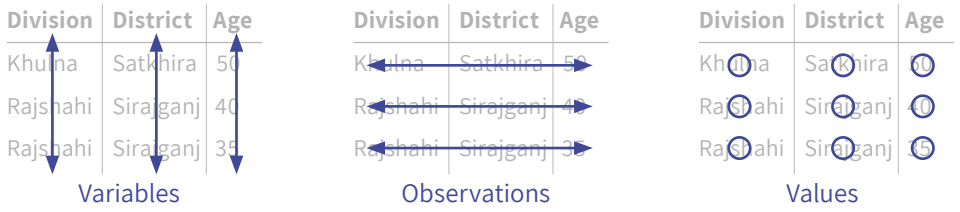


Figure 03: An example of tidy data

1 Han, J., Kamber, M. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001. ISBN 1-55860-489-8.

From here, we can start checking other things:

- Are all your columns are in the right data format (e.g. numbers as numerical and not as text)?
- Does every column have a unique, correct and simple header? For example: “Age” instead of “What is your age?”
- Have you double checked the character encoding? When you import data from CSV for example, the character encoding can be wrong which can result in special characters not being displayed correctly, e.g. "FÃ ©dÃ ©ration" instead of "Fédération". UTF-8 encoding[2] does well with most languages.

Clean your data

At this stage, you can start checking for inconsistencies and inaccuracies in your data using the table below as a guide. This table shows the key elements to data quality - accuracy, consistency, completeness, timeliness, uniqueness, and validity. Each of these elements should be checked for each and every variable as well as a combination of variables.

Accuracy

How close is the value of the data to the true value? In other words, how accurately does the value of the data describe the object or event being described?

There are two types of accuracy.

- **Syntactic accuracy:** In this case, the value might be correct, but it doesn’t belong to the correct domain of the variable. For example: A negative value for duration or age or a percentage higher than 100.
- **Semantic accuracy:** In this case, the value is in the correct domain, but it is not correct. For example: The attribute gender is given the value “female” for a person called John Smith.

Consistency

Do all the values of one variable represent the same definition? For example: Distance recorded in the same unit throughout the dataset.

Completeness

How complete is the dataset with respect to variable values and/or records?

- **Variable values:** Are there values missing for certain variables?
- **Records:** Is the dataset complete for the analysis at hand? For example, you set out to survey 1,000 households but only have 900 completed.

The occurrence of missing values can have different causes.

People might have refused or forgotten to answer a question in a questionnaire, or a variable might not be applicable to a certain object. For instance, the variable “pregnant” with the two possible values, yes and no, does not make sense for men. Of course, one could always enter the value no for the attribute pregnant. But this might lead to a grouping of men with not-pregnant women.

Timeliness

How timely is the data? For example: Data has to be collected within a defined time period.

Uniqueness

Is there any duplicate data?

Checking the uniqueness in a dataset consists of identifying and correcting duplicated rows (observations). For example: A water point that has been mapped twice.

Validity

Does the data conform to the defined rules? For example: The project ID should always be three characters between A-Z. A value such as *abdd* is therefore invalid.

2 https://en.wikipedia.org/wiki/UTF-8

Handling outliers

Once you've checked your data against the data quality elements, you should also check for outliers. An outlier is a data point that differs significantly from other observations.

Household	Income
1	$ 12,000
2	$ 50,000
3	$ 11,000
4	$ 120,000
5	$ 14,000
6	$ 10,000

Figure 04: Example of an outlier

In figure 04, you can see an example of an outlier. While most households have an income of between $10,000 and $50,000, household four has an income of $120,000.

There are different ways to go about solving this:

- Check with the data collector. Did they really earn this amount or was it a typo?

If this is not possible, you'll have to make the decision yourself:

- If you are sure it is a typo, modify the value to $12,000.
- Leave the value (there might be a household that actually earns this) and maybe use mode instead of average to avoid getting a skewed view of the data.
- Replace the value by the mean of the variable.
- Remove the observation based on a statistical threshold, for example when a value is two standard deviations from the mean.

It can be time consuming to check for outliers manually, especially when there are many observations. A simple trick to speed this process up is to visualise the the data with the use of a box plot.

You can also use a scatter plot to identify outliers, as visualised below: You can also use a scatter plot to identify outliers, as visualised below:

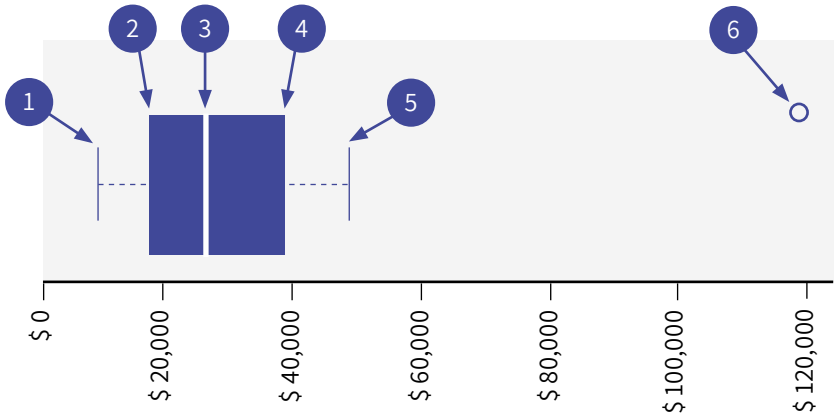


Figure 05: A box plot

1. The minimum (the smallest number in the data set)
2. First quartile (Q1)
3. The median
4. Third quartile (Q3)
5. The maximum
6. An outlier (this value lies outside of the maximum)

You can also use a scatter plot to identify outliers, as visualised below: You can also use a scatter plot to identify outliers, as visualised in figure 06 below:

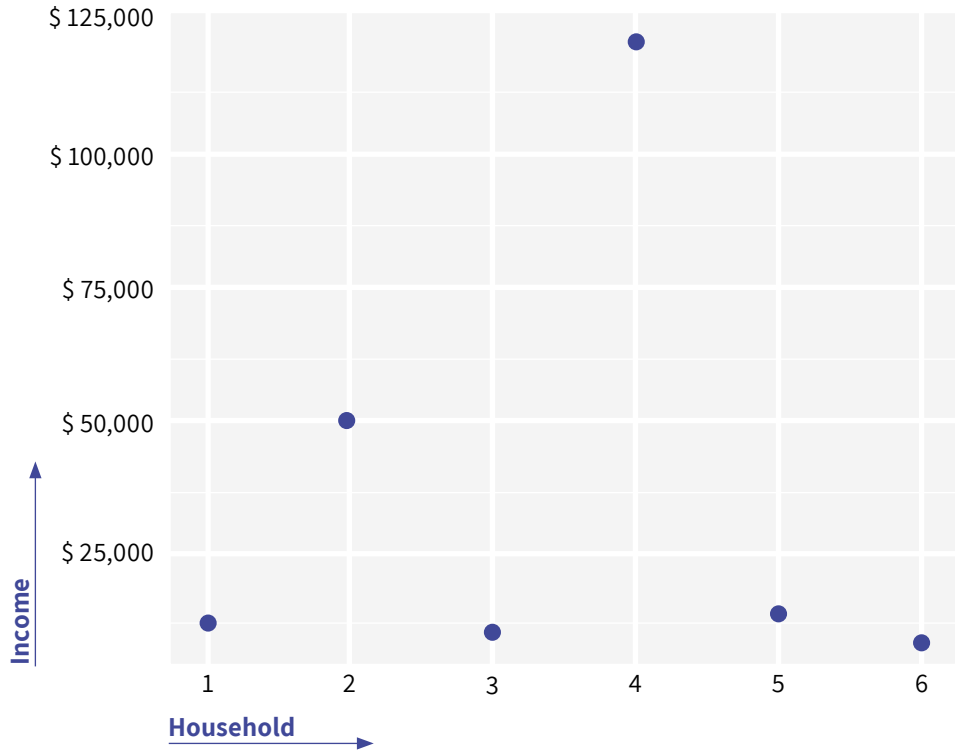


Figure 06: A scatter plot

Data analysis before and after cleaning

Let's see what happens when you analyse data before you've cleaned it. To show how best to approach data cleaning, analysis and visualisation, we are using an example dataset throughout this eBook which you can view here: Example WaSH dataset.[1] The dataset contains information about WaSH facilities in households in Bangladesh, including data on the household, the water source, the toilet facilities, hand washing facilities, and geolocation. See table 01 on page 17 for an overview of the variables present in the data set.

Using this dataset, we'll perform the same analysis before and after cleaning. We want to know what the average cost of a toilet is per type. Based on data that has not yet been cleaned, we can see that an improved toilet costs on average $19,275.

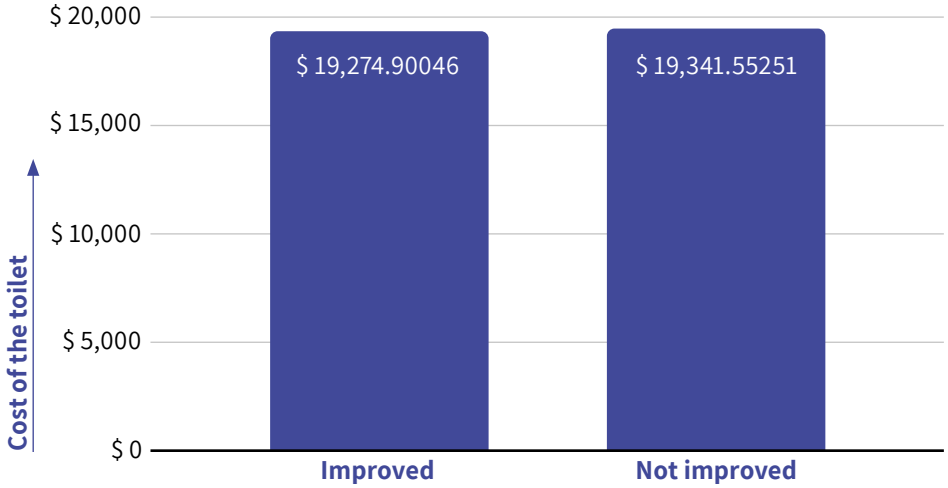


Figure 07: Analysis done with unclean data

1 https://docs.google.com/spreadsheets/d/108H0H7NKqPEbPrQgdLyChHAjeUxSfzA4xiwkHF4Q2nU/edit?usp=sharing

During the data cleaning process, we started by assessing the data quality indicators as described at the beginning of this chapter. In this project, the working group concluded that it is not possible to have a toilet that costs more than $50,000. We decided to investigate further by looking at other indicators. Most mobile data collection tools enable enumerators to add additional data during collection. For example, the geolocation (was it captured at the planned location?), the survey duration (did the interview take a realistic amount of time for the number of questions?) and, for example, a picture (do we see the expected subject?). In our example, the expected survey duration was between 30 and 45 minutes and our outlier was collected in just 11 minutes. Knowing that both the cost and time value were outside our limits, we decided to take the value out.

After cleaning the data and removing the outlier, we performed the same analysis. This time around, we see that an improved toilet costs on average $15,713 (figure 08). This is more than a 20% difference with the uncleaned value. Using unclean data can therefore lead to bad decision making.

Figure 08: Analysis done with clean data

Minimising data cleaning

Data cleaning is time consuming. You can significantly minimise data cleaning by designing your survey with data quality in mind. One way to do this is by limiting the possible answers to a certain question without influencing the actual response. For example, when trying to discover the age of an interviewee, it's advisable to choose a numeric question so that every answer is numeric (e.g. 6, 10, 60). You could leave it at this, but the data collector could mistakenly tap another zero when entering 60. Most data collection tools allow us to further limit the responses. For example, for a household interview you could say the age should be >18 and <116. It should be above 18 as you only want to interview adults and below 116 as this is the oldest living person in the world. For a deep dive into solid survey design, check out this blog: How to design your survey for smooth and reliable data collection.[1]

Additional tips:

- If applicable, design a set of rules that can be checked during data collection to prevent mismatches of answers (If the answer to Q1 is A, then the answer to Q2 cannot be B).
- Train the enumerators thoroughly to optimise data quality during collection.
- Monitor incoming data during data collection and try to correct it as soon as possible, preferably while you still have access to the data collector and the interviewee.
- Keep a change log where you document all data cleaning actions.

1 https://datajourney.akvo.org/blog/how-to-design-your-survey-for-smooth-and-reliable-data-collection

Table 01: Overview of variables in example WaSH data set

Household	Water source	Toilet facilities	Hand washing facilities	Additional
Division	Current main water source	Presence of toilet in the household	Frequency of hand washing	Organisation
District	Water source status	Type of toilet in the household	Hand washing product	Latitude
Gender of the respondent	Water source this year	Status of toilet use	Presence of hand soap	Longitude
Age of the respondent	Water source ownership	Type of toilet used this year	Hand washing facility in the household	Elevation
Number of household members	Water source installation	Toilet owner		
Number of female household members	Water source manager	Cost of the toilet		
Number of children in the household	Distance to water source	Support of the toilet cost		
Highest education level in the household	Length of the trip to the water source			
Main income source				
Income				
Income category				

4 | Predict the future, understand the past: The four types of data analysis

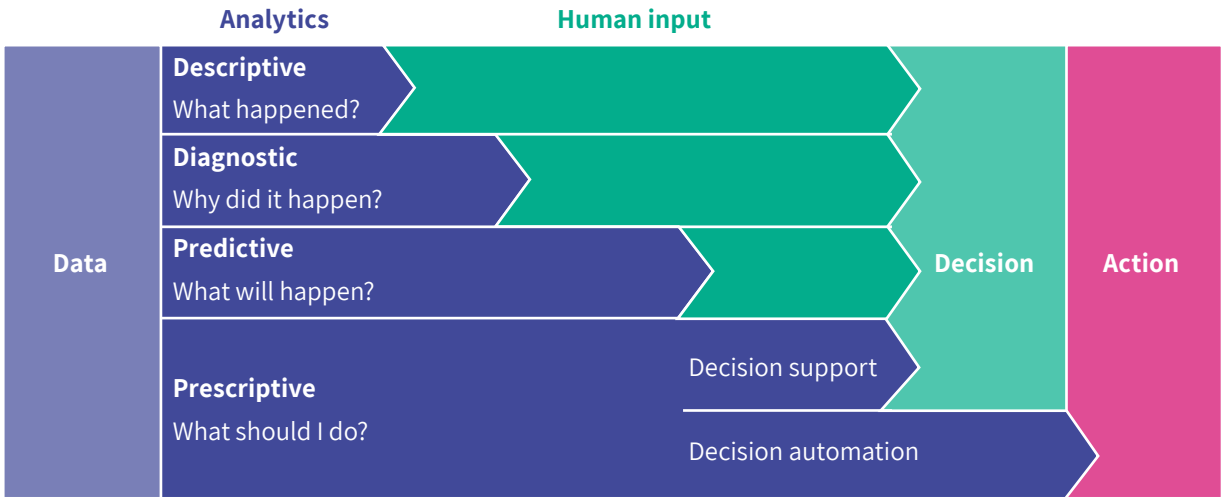


Figure 09: An overview of the four key types of analytics

In this chapter we will discuss four types of data analysis:

1. Descriptive
2. Diagnostic
3. Predictive
4. Prescriptive

Descriptive analytics - What happened?

Descriptive analytics is the first step in data analysis. The goal of descriptive analytics is to find out what happened? For example, what was the average revenue for the month of January? Or how many children between the ages of two and ten attend school? It's the first layer of information that you can get from the data you've collected, either with or without adding data from other sources.

If you're an analyst, you might not know in detail, or at all, what happened that led to the dataset you need to analyse. If this is the case, it's important to seek out context on the sample you're looking

at and the population it was drawn from. For a more detailed guide on the importance of sampling strategies, check out this blog: How to choose your sampling strategy to guarantee relevant results.[1]

If you're a project officer or programme manager, you may already have a good idea of the information your data contains. You may have been involved in the data collection, perhaps even the sampling, and you might have knowledge of what went well and what didn't. Questions that are important to keep in mind when looking at descriptive statistics are:
Did I reach my required sample size?
Were there any local conditions that might have influenced my data? For example, weather conditions that forced you to use other participants than intended.

Using the example dataset[2] introduced in chapter three, we'll give some examples of how to conduct a descriptive analysis.

Tables and bar charts

The easiest and quickest way to look into your data is by using (frequency) tables and bar charts. This only goes for nominal and sometimes ordinal data, as described in chapter two. Making use of the pivot table function in excel, for example, allows you to depict a lot of information. Start with variables that describe your sample, such as gender (figure 10). This will give you an idea of the characteristics of the respondents. A deviation in these characteristics can give you insight into the variables you set out to measure and your potential findings.

1 https://datajourney.akvo.org/blog/how-to-choose-your-sampling-strategy-to-guarantee-relevant-results

2 https://docs.google.com/spreadsheets/d/108H0H7NKqPEbPrQgdLyChHAjeUxSfzA4xiwkHF4Q2nU/edit?usp=sharing

	Female	Male
Count of gender of the respondent	389	261

Figure 10: A table showing the number of female and male participants

You may already have an idea of how certain variables will be represented in your sample. In this case, making a bar chart will offer a lot of insight. Figure 11 shows the education level of the participants, which could be seen as an ordinal variable. By observing this variable in a bar chart instead of a table, you immediately get a sense of the average education levels. In later analyses this can be translated to, for example, how many of them can read or write.

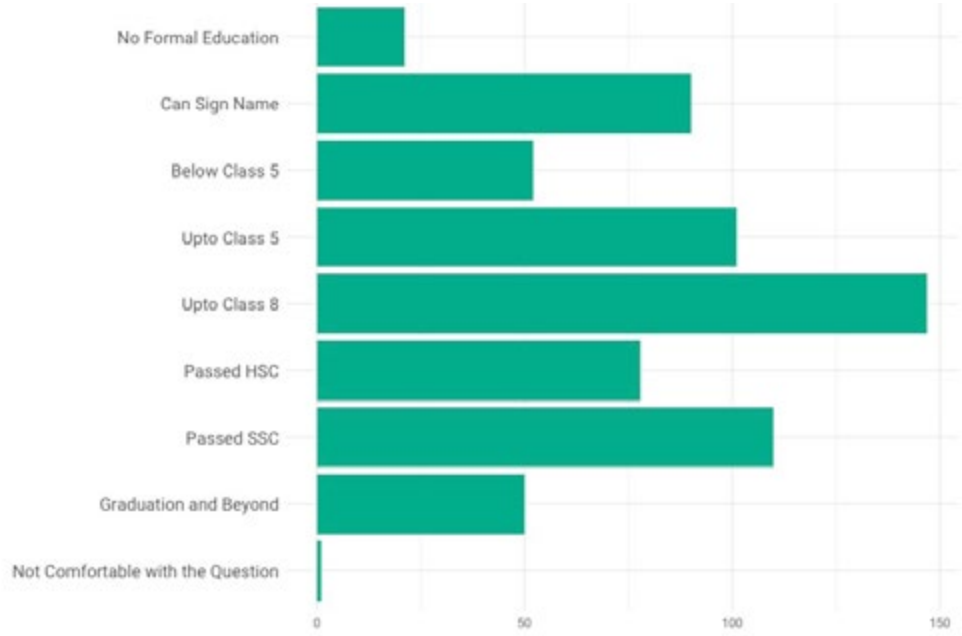


Figure 11: Education levels

After looking at sample characteristics, you can start making overviews of the variables of interest. Figure 12 shows an overview of the types of water sources represented in the data.

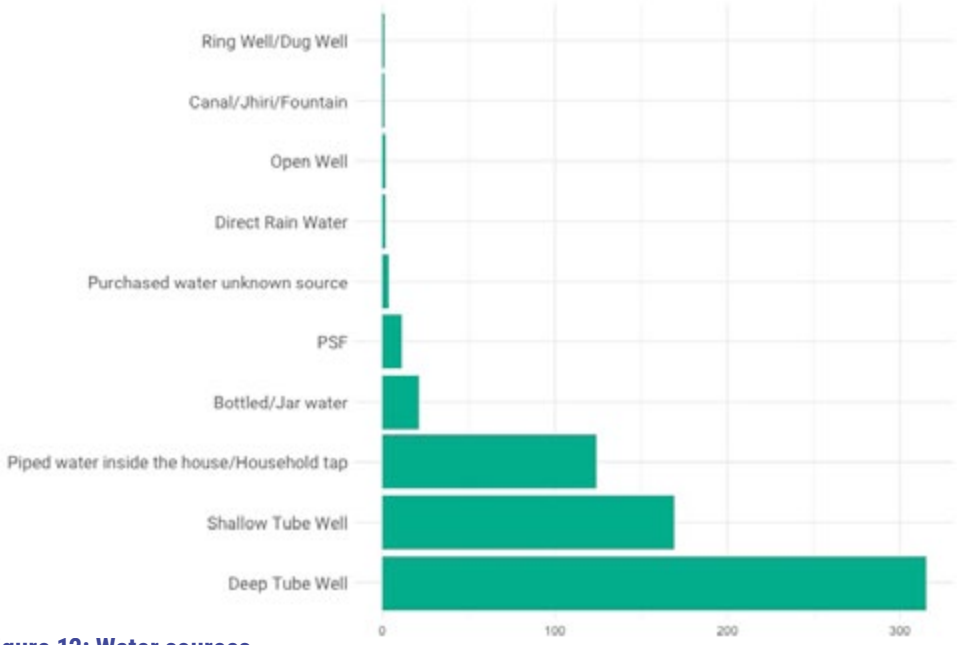


Figure 12: Water sources

The distinction between different water sources alone is not enough to indicate the quality of the water facilities. Other variables impact the quality of the source, such as the walking distance from the household to the source, whether or not you have to wait in line to get water, and whether animals can use the source as well. Figure 13 shows the combination of the type of water source, here categorised as “improved” or “unimproved,” combined with the length of the trip to the water source. Combining these variables already gives more information about the quality of the water facilities. The World Health Organisation (WHO) introduced standards to classify a water source as improved or unimproved. See this introduction to drinking water[3] for more information about this distinction.

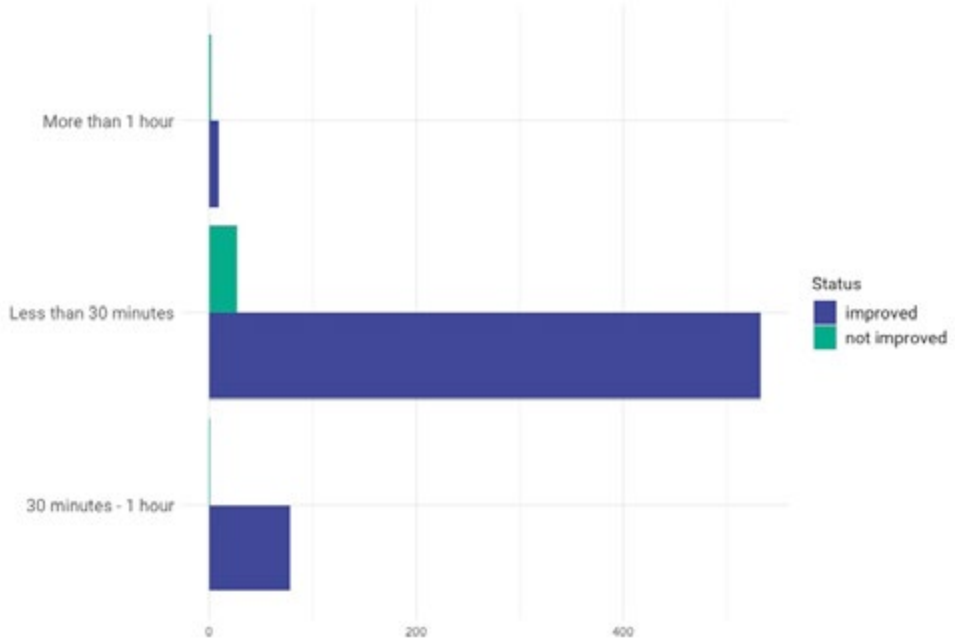


Figure 13: Quality indication of water source - (walking) distance to the water source

3 https://resources.cawst.org/manual/cfd38f83/drinking-water-quality-testing-manual

Histograms

A more advanced way of looking at your ratio or interval variables can be achieved by making a histogram. A histogram visualises the distribution of data over a continuous interval and can be used to see how your data is deviated. For example, if we look at the number of household members, we expect the sample to follow a normal distribution. This basically means you have observations centered around a mean with equal deviations to both sides. Take a look at this lecture by Andy Field[4] for information on the assumption of normality and other forms of bias. Figure ten shows the distribution of the number of household members. For more on distributions, see chapter five of this eBook.

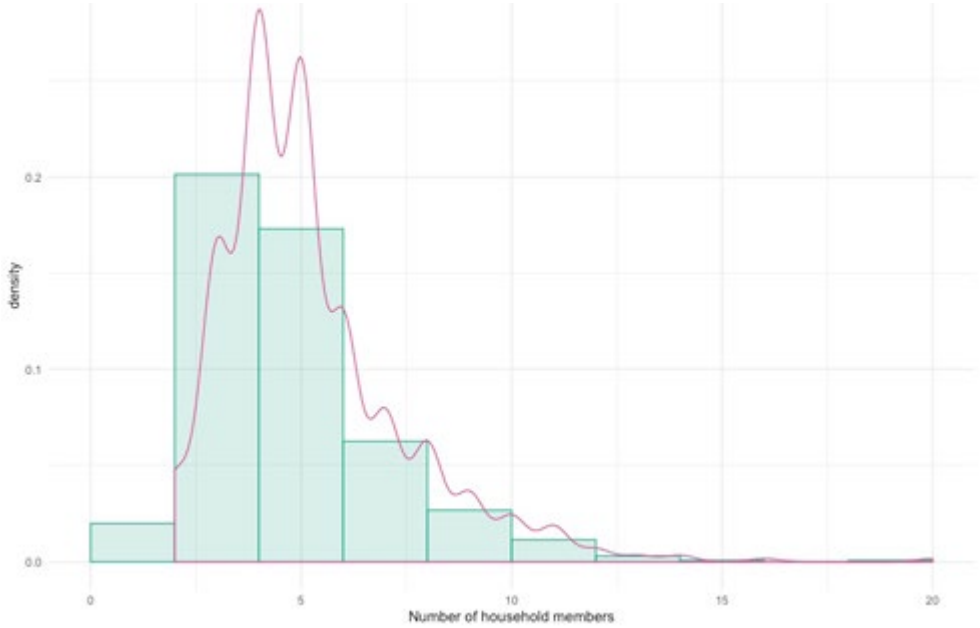


Figure 14: Number of household members

When looking at the distribution (figure ten), you can see that the bars are skewed to the right. This implies that there are a few large households in the set compared to the average. Given that the sample is representative, these data points, and therefore the households, could be outliers. If your sample is not representative, it could mean that you have sampled among the smaller households, and that in reality the mean is higher. If you are in the business of hypothesis testing, this distinction is important.

Diagnostic analytics - Why did this happen?

When you have an overview of what is in your data and what your sample looks like, you might want to know why certain things are happening. Maybe you found that in one district, far less children attend school than in the other districts. Could there be something in your data that shows other ways in which this district is different? With diagnostic analytics, we can go one step deeper and ask the question: Why did this happen? Again, we will use the example dataset[5] introduced in chapter three to give you some examples.

To see why certain behaviour is observed, we can look at a combination of variables. For example, which district has the most unimproved water sources? Are primarily women reporting long walking distances? Are water sources more often broken when they are not looked after by the municipality? Figure 15 shows the ownership of water sources compared to whether the well is functional or not. You might wonder whether the municipalities take good care of the wells or whether the interviewees have enough experience to take care of the well themselves. In this case, you can see that the private wells are a lot more often functional than those installed by the municipalities.

4 https://www.youtube.com/watch?v=GKmvz6SbBoo

5 https://docs.google.com/spreadsheets/d/108H0H7NKqPEbPrQgdLyChHAjeUxSfzA4xiwkHF4Q2nU/edit?usp=sharing

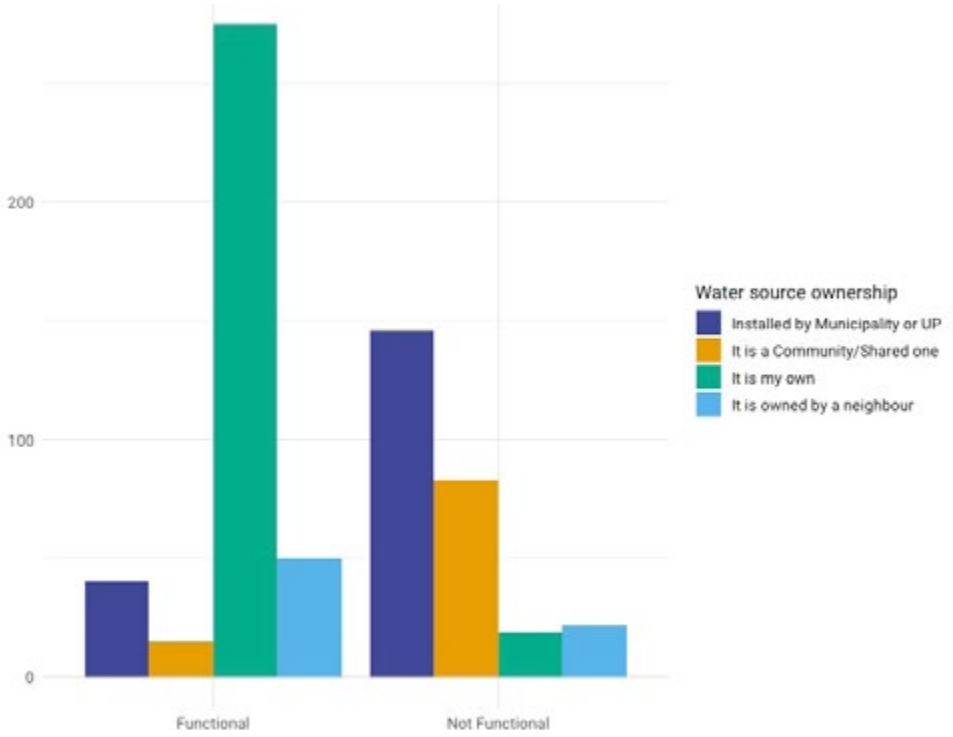


Figure 15: Water source ownership and functionality of the water source

The questions you ask, and the variables you combine, can be based on previous research, by hunches or observations during data collection, or they could be explorative, meaning you *found* them in the data. When you are working with sampled data, the distinction between how you get to these findings is very important. With the example in figure 15, you might have had a reason to investigate water source ownership. For example, qualitative research showed you that the local municipality is understaffed. When you formulate a research question and an accompanying hypothesis in advance, and you have a representative sample, you can test whether the hypothesis can be confirmed using *statistical inference*. Statistical inference refers to the use of statistics to draw conclusions about some unknown aspect of a population based on a random sample from that population. When you can indeed confirm your hypothesis, you can assume that the behaviour you find is something that goes for the entire population. In our case, you could use this to advocate for more staff in the municipality.

If your findings are explorative, you cannot make this assumption. Figure 16 shows the different water sources split according to the gender of the respondent. You can see that women seem to report a lot more *shallow tube wells*. If you never had any reason to expect this difference, you have to be aware of the fact that this could be a chance finding. As we saw in the descriptive analytics part, there were a lot more female respondents, which could account for this difference.

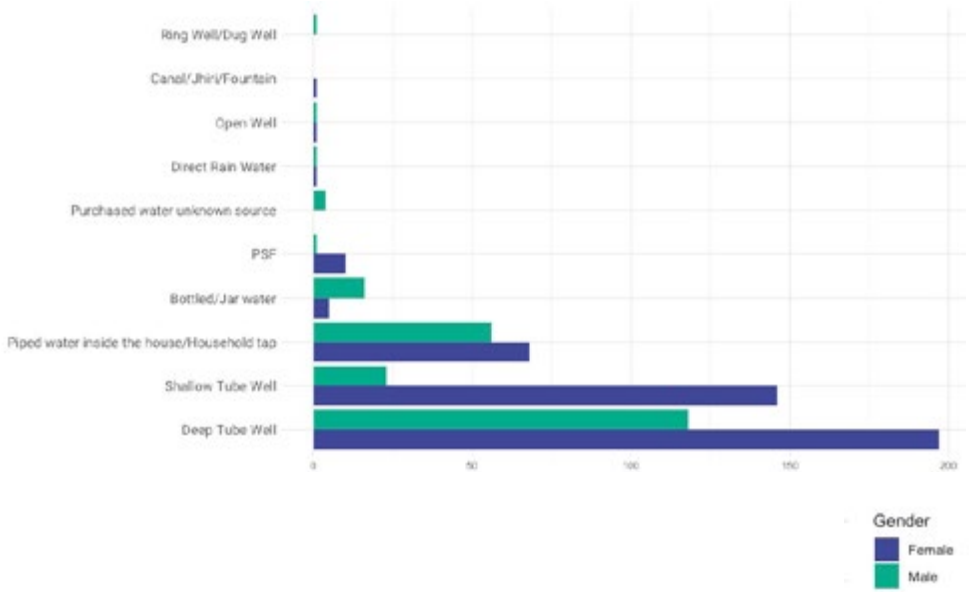


Figure 16: Water sources and gender of the respondent

Predictive analytics - What is likely to happen?

By the time you know why something happened, we might go as far as predicting what is likely to happen next, given our knowledge of previous events. Predictive analytics tries to answer the question: *What is likely to happen?* By using what we learned with descriptive and diagnostic analytics, we can use predictive analytics to look at clusters, tendencies or maybe exceptions that allow us to make a certain prediction.

Let's say that during the diagnostic analysis, you found that shallow tube wells built on a clay surface are often broken when older than five years. This information can be used to investigate the remainder of your population and make an estimate of which wells might need repairing without actually knowing they are broken.

These findings are often referred to as *patterns* in the data. There are generally two ways of looking at these patterns: supervised (e.g. regression) or unsupervised (e.g. clustering).

There are more advanced techniques in predictive analytics that allow you to search for patterns in your data and confirm them at the same time. Let's say, for example, you collected data about water source wells. You know the type, when they were built, who is maintaining them, if they are broken or not, and you used additional data to determine the kind of soil the wells are built on. By feeding this data into a predictive model, you can make a model that allows you to predict when a new well, one that wasn't part of your initial data, might break down. One of these predictive models is logistic regression. Again, see how statistics professor Andy Field[6] explains this concept.

Prescriptive analytics - What should be done?

Now that you have an idea of what is likely to happen, you might want to know what the best course of action is. Prescriptive analytics tries to answer the question: *What should be done?* or *what can we do to make ... happen?* Prescriptive analytics is mostly used in large companies that are looking for advice on, for example, their inventory or supply chain. It goes one step further than descriptive and predictive analytics by recommending possible outcomes. Essentially, you can predict multiple futures and allow companies to assess a number of possible outcomes based upon their actions.

6 https://www.youtube.com/watch?v=37983YYQnWU

5 | How to visualise your data in a clear and compelling way

When the data has been cleaned and analysed, the next step is to share the outcomes with your audience in an intuitive and digestible format. This can be done through data visualisation. Data visualisation is the graphic representation of data in an infographic, map, dashboard, chart, table or graph. Creating an effective data visualisation requires domain expertise, data expertise, and communication/design skills. We therefore encourage interdisciplinary cooperation when creating your visualisations.

By following these next steps, you'll be able to communicate complex datasets in a clear and compelling way to a variety of stakeholders.

Understand the context

Before you get started, it is important to determine the context in which your visualisation is being created. You can do this by asking the who, what, why and how questions. Why are you doing this project (objective)? Who needs to see the data visualisations in order to take action (target audience)? What do you want to communicate (message)? How will you share the message with your target audience (medium)?

It is also important to determine the function of the visualisation in advance. In most data visualisation projects, a distinction can be made between explanatory and exploratory data visualisations. If your aim is to communicate a specific insight, message, or story to your audience (explanatory), you'll use a different type of visualisation than if you want to give your audience the option to become familiar with the data themselves (exploratory).

Choose an appropriate visual display

When the context is clear, it's time to find the right visual display. There are multiple types of data visualisation, and the type you choose depends on the number of variables you want to display and the relationships between those variables. In general, data visualisations can be grouped according to these five categories:

Compare categories

2. Show change over time
3. Display a part of a whole
4. Present relationships between variables
5. Explore geospatial data

Comparing categories

This type of visualisation can be used to facilitate a comparison between categorical or quantitative values. The most commonly used comparison charts are bar charts and histograms.

Bar chart

This is the most common chart type and is generally used when you want to show a comparison between different values of a variable. A bar chart can be displayed both horizontally and vertically, as long as the numerical axis starts with zero in order to avoid inaccuracy.

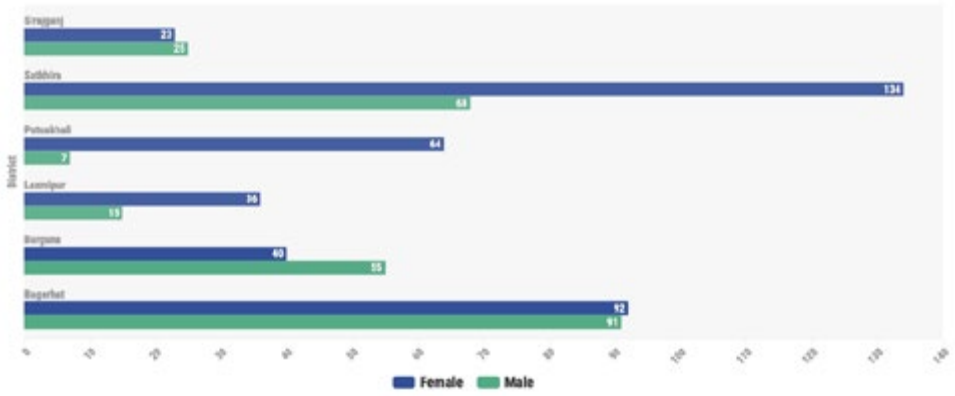


Figure 17: A bar chart showing gender per district

Histogram

A histogram shows how data is spread out compared to one central value. Histograms help to get a general overview of specific values in a dataset and to display extreme or unusual values.

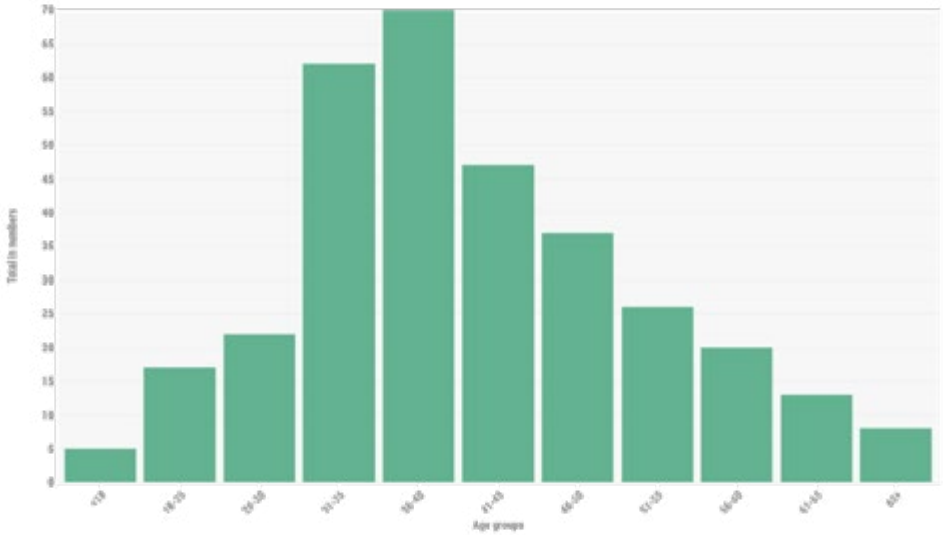


Figure 18: A histogram showing age distribution

You can also use radial charts, dot plots and sankey diagrams to compare categories.

Change over time

The most common time series charts are line charts and area charts.

Line charts

Line charts are used to compare continuous joint values (x axis) to quantitative values (y axis). If you want to display trends over a period of time, line charts are useful. Unlike bar charts, the y axis does not need to start at zero because line charts display a relative pattern.

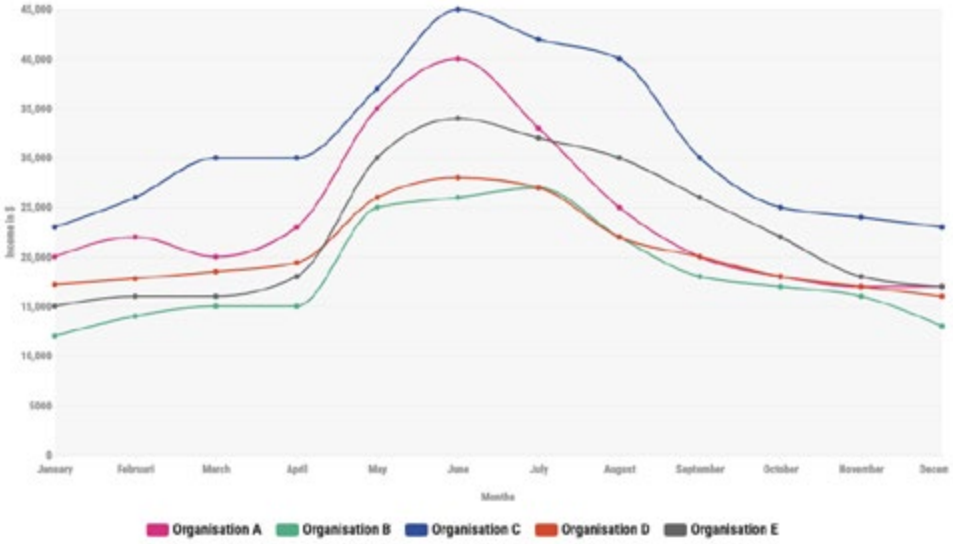


Figure 19: A line chart showing average monthly income over a period of time

Area charts

Area charts are line charts, but the area below the line is filled with a certain colour or texture. Unlike line charts, the y axis should start at zero for accurate interpretation

.

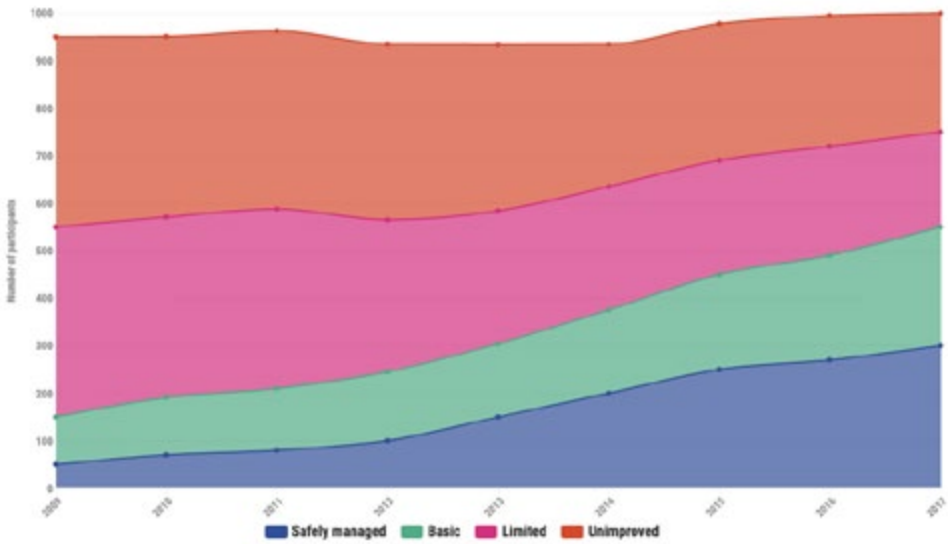


Figure 20: An area chart showing the service level of drinking water access over time

Stacked area charts, candlestick charts and Gantt charts can also be used to display change over time.

Part of a whole

Part of a whole refers to visualisations in which you want to show how something is divided up. The most commonly used visualisations to show part of a whole are pie charts and stacked bar charts.

Pie chart

Pie charts are ideal to get a quick idea of the proportional distribution of the data. Unfortunately, pie charts are also among the most frequently misused charts. They should only be used when:

- All values add up to 100%
- The division between values is clearly visible
- A maximum of four categories is used to divide the values.

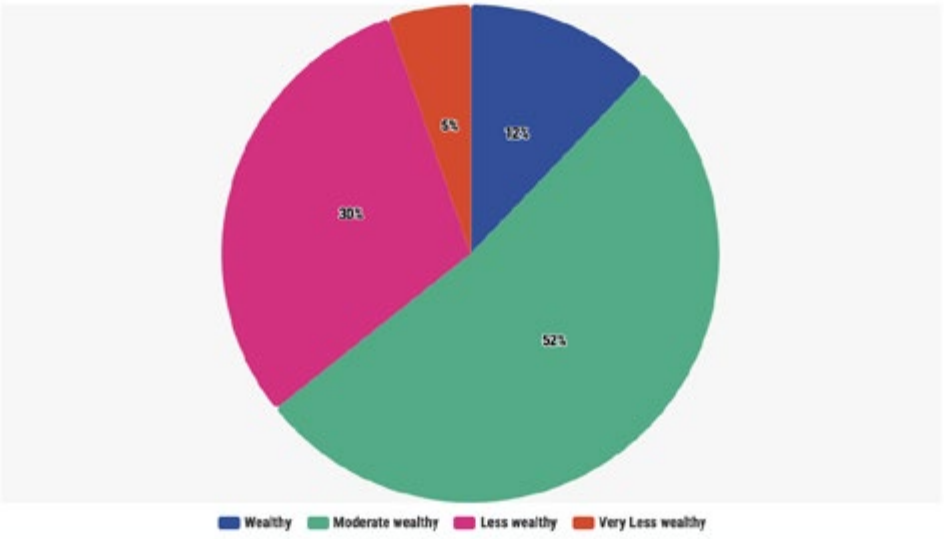


Figure 21: A pie chart showing income categories

Stacked bar chart

Stacked bar charts are used to display how a larger category is divided into smaller categories in one overview; it shows the relationship of each part to the whole. Like pie charts, stacked bar charts should be used carefully and the division between values should be clearly visible.

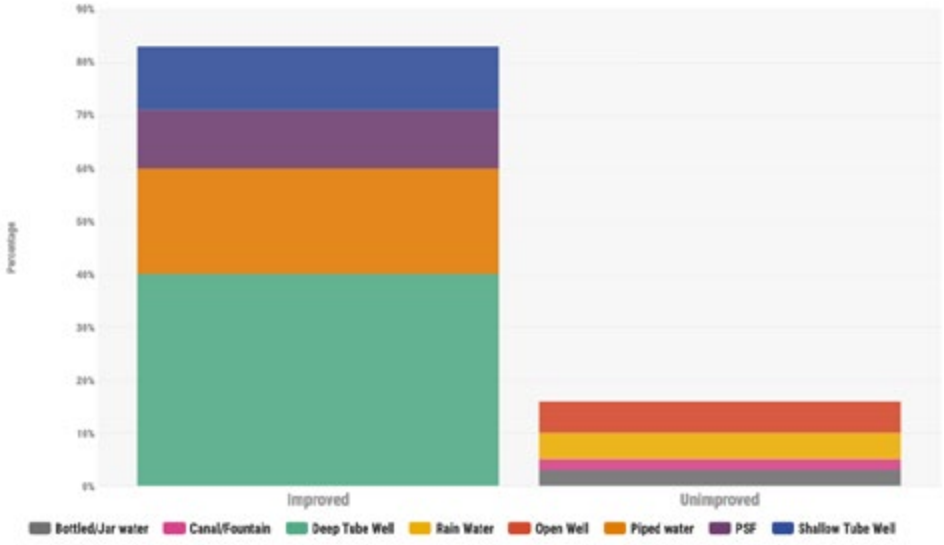


Figure 22: A stacked bar chart showing improved and unimproved water sources by type

Other commonly used charts to display a part of a whole are tree maps and donut charts.

Relationships between variables

Relationship charts can be used to display correlations between two or more variables.

Scatter plot

A scatter plot is a combination of two variables plotted on the x and y axis to reveal patterns, correlations or outliers. They are particularly useful during the data cleaning and data analytics phase to get familiar with the data.

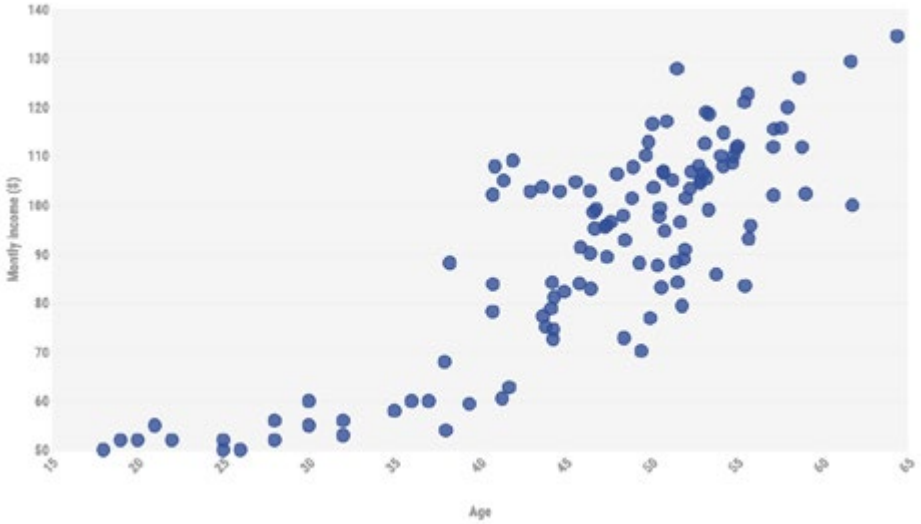


Figure 23: A scatter plot showing the relationship between income and age

Bubble chart

A bubble chart can be used to add a further dimension to the data. Using colour for example, you can distinguish between categories or represent an additional data variable.

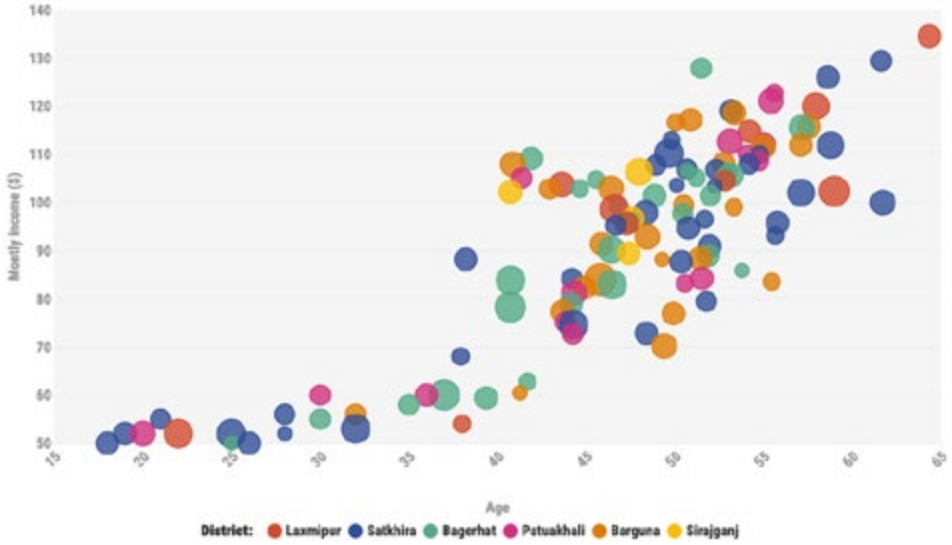


Figure 24: A bubble chart showing the relationship between income and age according to district

Geospatial

When a dataset contains geographic information, the data can also be displayed on a map to highlight the locations. The most common type of geospatial data visualisation is point map and choropleth map.

Point map

Point maps are used to detect spatial patterns in data over a geographic area. By placing equally sized points on a map, point maps can reveal patterns when the points cluster on the map.

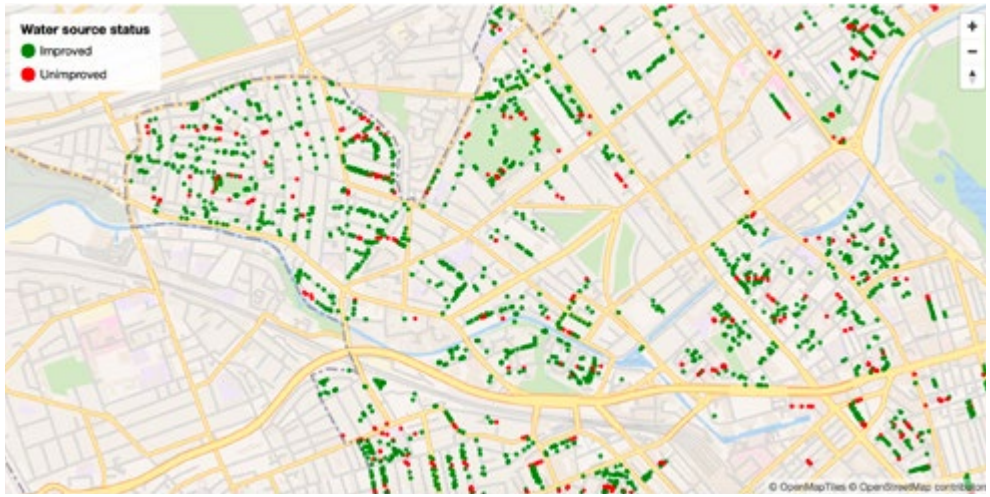


Figure 24: A point map showing water source status per household

Choropleth map

Choropleth maps display divided geographic areas or regions that are coloured, shaded or patterned in relation to the different values of a dataset. This provides a way to visualise values over a geographic area, which can show variations or patterns across a map.

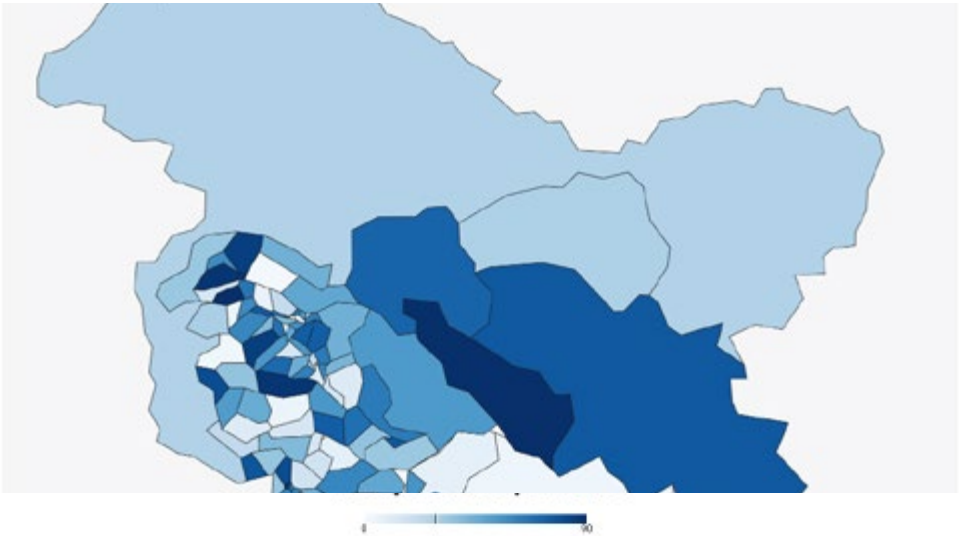


Figure 25: A choropleth map showing water point access per district

Design your data visualisation story

When you've decided how to visualise your data, it's time to go beyond concept and towards design.

There are hundreds of ways to ultimately present your data. If your project is more challenging than the examples used in the first half of this chapter, it's advisable to first sketch how your data visualisation should look. Think, for example, of setting up infographics or interactive data stories.

The next step is to select a tool to make the data visualisation. There are numerous data visualisation tools out there, and each have their pros and cons. This goes beyond budget - be sure to consider functionalities, scale and the skill required to use the tool effectively. For example, some programming languages are extremely suitable for creating customised interactive data visualisations, but also require the right skill. There are also many generic tools that offer a solution for when the technical programming or design skills are missing.

After selecting the right tool, the focus is on actually creating the data visualisations. The final design depends as much on personal taste and the target group as it does on what you want to achieve. In general, these tips are useful when designing data visualisations:

1. **Declutter:** Remove every label, piece of text, drop shadow or other addition that isn't strictly necessary to your visualisation.
2. **Focus:** Annotate and highlight relevant contextual details in the data visualisation to provide additional focus.
3. **Order data intuitively:** Make a logical hierarchy in your data visualisation to improve readability.
4. **Make the right use of colours:** Colours can be used to guide the viewer's eye and draw attention to particular values. Using a good colour palette can clarify your message.
5. **Tell your story:** Good data and a good visualisation alone are not enough to convey a message. A strategic and focused narrative to guide the viewer towards key facts is just as important. For more on this, read this blog on five tips for effective data storytelling.[1]

Share your data visualisation

The final step is to share the data. How you share the data will depend on the questions we asked at the beginning of this chapter: Why are you doing this project (objective)? Who needs to see the data visualisations in order to take action (target audience)? What do you want to communicate (message)? How will you share the message with your target audience (medium)? Data can be shared in the following ways:

Reports

In the development sector, it often happens that data visualisations are published in reports in which the results of a project are shared. These are printed on paper, and allow the publisher to explain findings in more detail.

1 https://datajourney.akvo.org/blog/five-tips-for-effective-data-storytelling

Figure 26: (left) An Akvo Lumen dashboard showing all farmers involved in the Mars Food data collection exercise in Cambodia during 2018

Figure 27: Infographic showing the water and sanitation sustainable development goals

Figure 28: A screenshot of an interactive chart and a narrative from Ghana's WASH data portal.

Dashboards

A dashboard is a collection of charts that offers an overview of the most important indicators of a project. An advantage of dashboards is that they can display real time data, meaning quick insights can be gained from an ongoing project or business progress. (Figure 26)

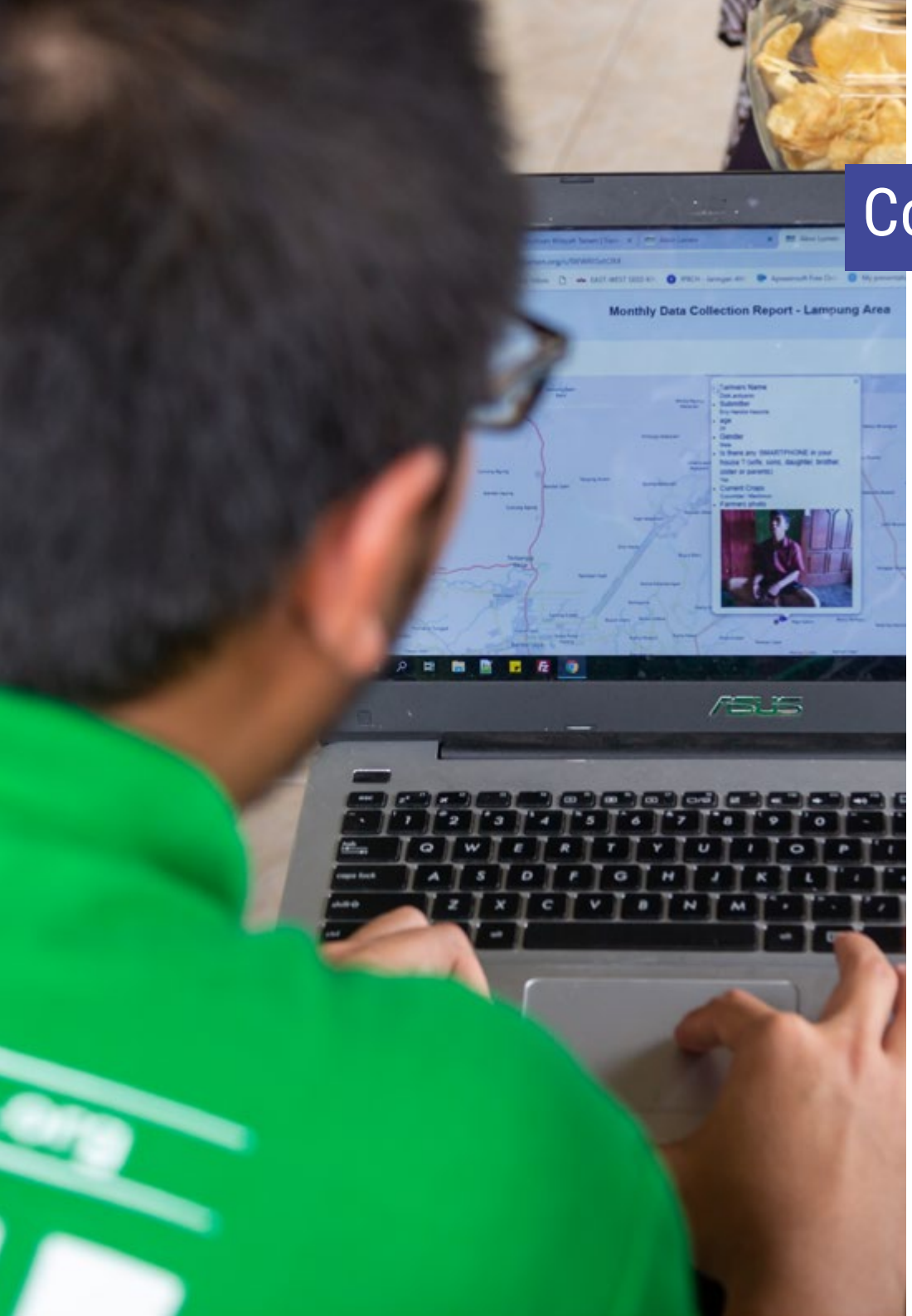
Infographics

An infographic provides an informative representation of different objects with a combination of text and image. Infographics are useful for transferring information, data and knowledge to an external audience. (Figure 27)

Websites

A website is an ideal medium to combine data visualisations with a narrative. The more interactive a website, the more user engagement in exploring the data. (Figure 28)

Whatever medium you choose, it's important to guide the viewer through the narrative and tell them a story. A good data visualisation ends with a call to action to direct or instruct the viewer to do something with the data presented. Ultimately, you want the audience to make decisions which contribute to your programme's intended impact.



Conclusion

The essence of the understand phase is exploration. By exploring your data, you can discover hidden patterns, extract valuable insights, and translate those insights into information and knowledge. Regardless of your field of study, the goal of an understand phase is to inform decision making. There are numerous ways to go about data analysis and visualisation. The key is in choosing the right approach for your programme's objectives and available resources. Conducting a thorough design and capture phase is essential in this process. By taking the time to achieve clarity and capture reliable data, you'll be equipped with the information you need to generate valuable information and knowledge to support decision making.

By following the steps in this eBook, you'll produce credible and compelling evidence for effective decision making. The better you've designed your analysis before capturing the data, cleaned your data according to the key quality indicators, and analysed and visualised it according to the specific context of your programme, the easier it will be for you to inform and support decision making in your programme. Ultimately, this will boost the impact of your development work. Good luck!

About Akvo

We believe in equal access to public services, reliable infrastructure and a safer environment for everyone. We are convinced that this will happen faster if governments and non-governmental organisations become more effective, accountable and collaborative.

Since 2008, we've worked with over 20 governments and 200 organisations in more than 70 countries to improve the way they implement development projects and make decisions using data. We call them partners.

With our combination of tools, services, local expertise and sector knowledge, our partners improve the management of water, sanitation and agriculture, with a strong commitment to accelerating the progress of the sustainable development goals.

With our unique approach to development, we help our partners design their projects so that they can capture and understand reliable data which they can act upon.

Visit us at www.akvo.org to learn more.

Credits

Project manager
Georgia Walker

Authors
Carmen Wolvius, Geert Soet, Lars Heemskerk

Co-Authors
Ilyasse Kabore, Irene Westra, Jana Gombitova

Editor
Georgia Walker

Art direction / graphic design
Linda Leunissen

Photograph and image credits
P04 & P07 WWF by Stefan Kraus (RGB Collective), P31 figure 27 Mars Food Akvo Lumen dashboard,[1] P31 figure 28 Infographic UN,[2] P31 figure 29 CWSA Water Atlas,[3] P32 SmartSeeds by Stefan Kraus (RGB Collective).

1 https://marsfood.akvolumen.org/s/Ea_KbD6MqUY

2 https://www.3blmedia.com/News/Unilever-Releases-Infographics-Demonstrate-Interlinkages-Between-Water-and-Sanitation

3 https://cwsawateratlas.org/data/brong-ahafo-region/service-authority-data


Looking to unlock the power
of data for sustainable change?
Embark on a data journey
#withAkvo and amplify the impact
of your development work
Akvo.
Get in touch now