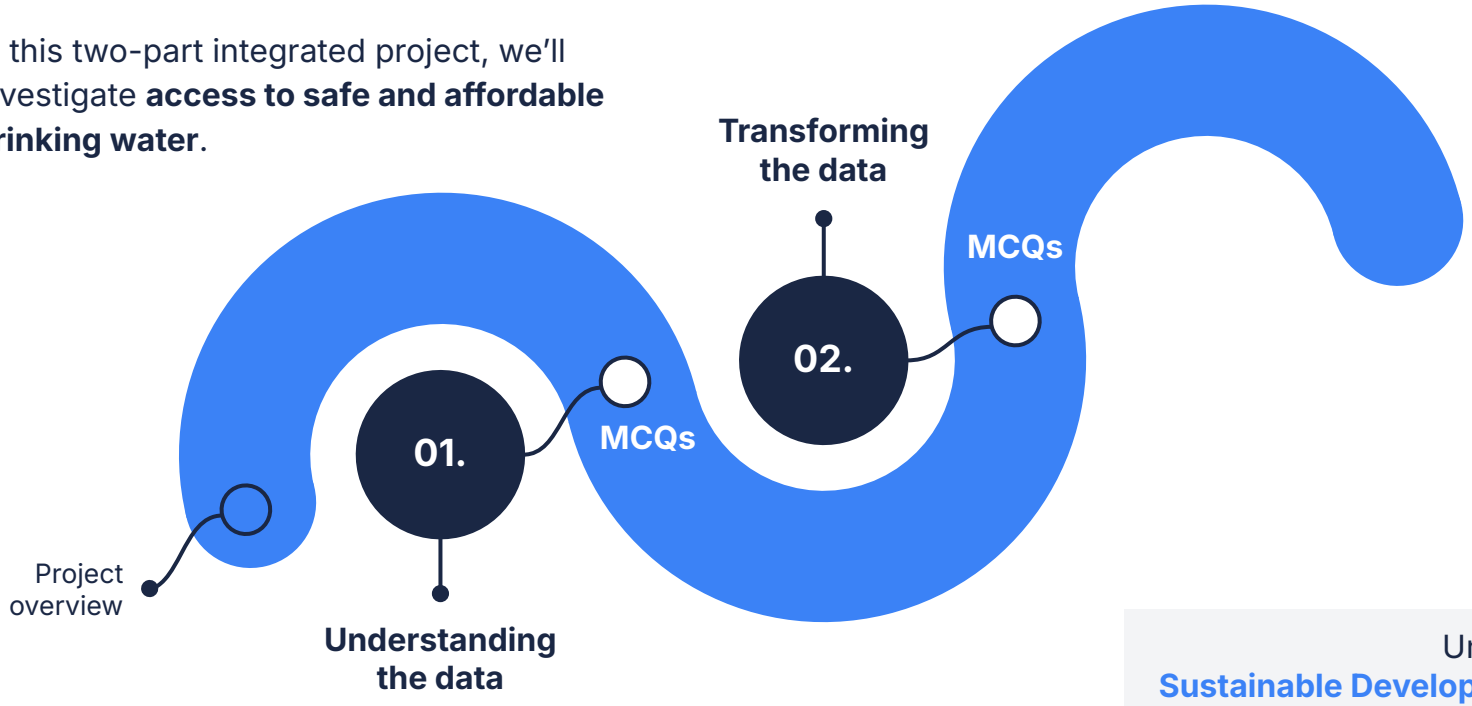
EXPLORE AI ACADEMY

Integrated project: Access to drinking water

Understanding the data

Integrated project overview

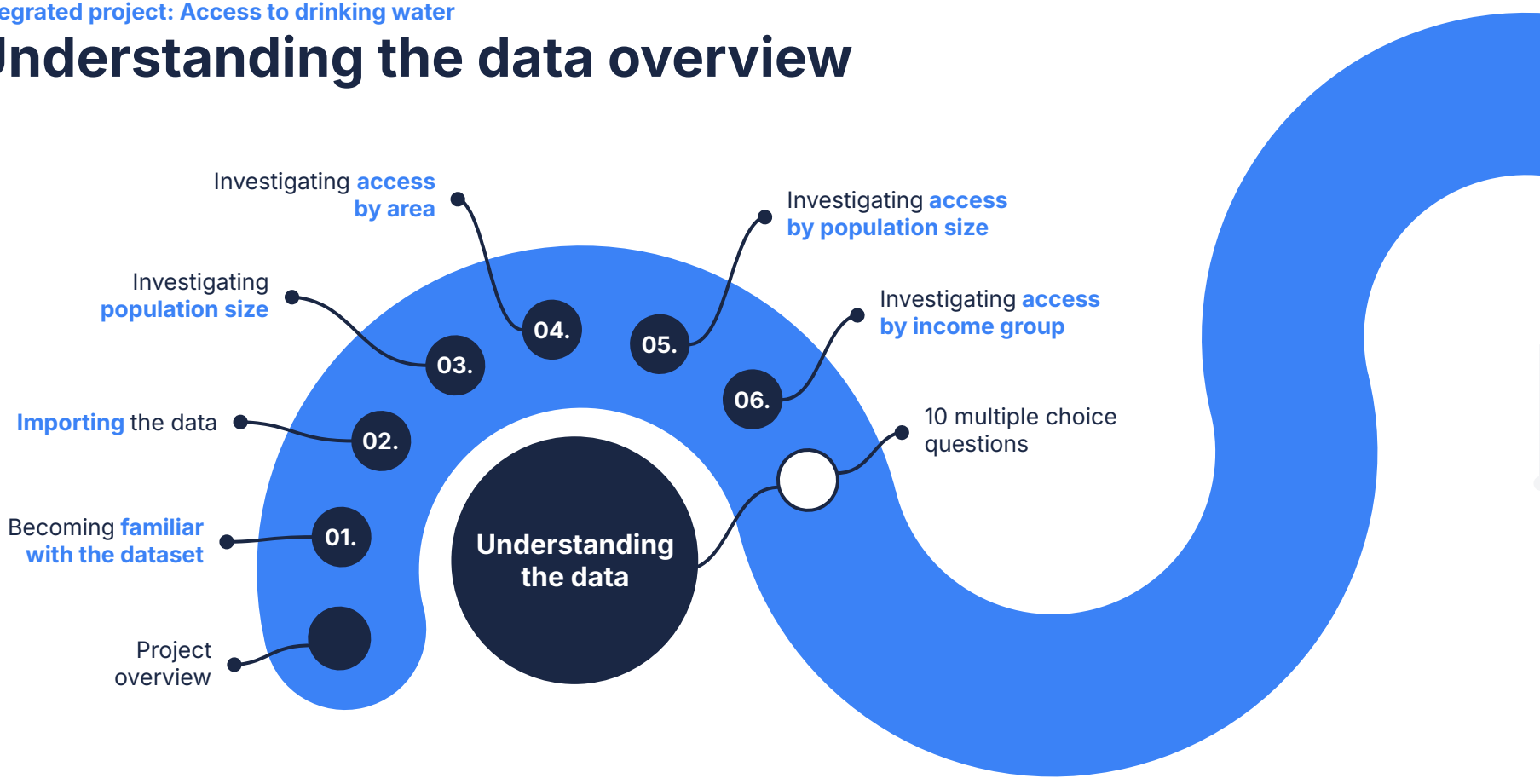
In this two-part integrated project, we'll investigate **access to safe and affordable drinking water**.



United Nations
Sustainable Development Goal 6
Clean water and sanitation

Integrated project: Access to drinking water

Understanding the data overview

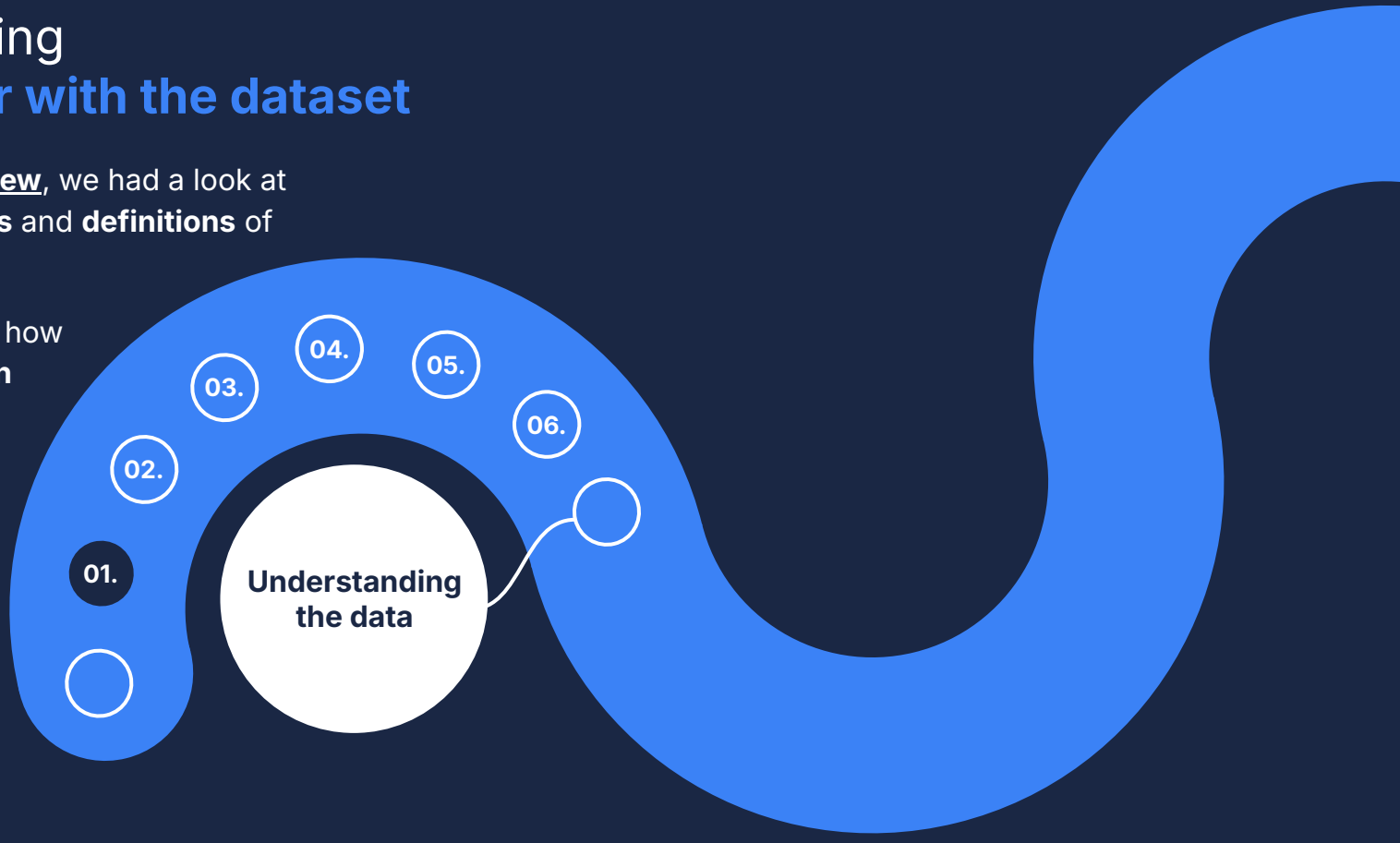


01.

Becoming **familiar with the dataset**

In the **project overview**, we had a look at the different **features** and **definitions** of this dataset.

Now we want to see how this is **represented in the data**.



Becoming familiar with the dataset

> We'll take a look at the WHO/UNICEF JMP (Joint Monitoring Programme for Water Supply, Sanitation, and Hygiene) **Estimates on the use of water** dataset for **2020.**

Estimates on the use of water (2020)

name

The country or area name.

income_group

The country's classification according to income group.

pop_n

The national population size estimate in thousands.

pop_u

The urban population share estimate in percentage points (%).

wat_bas_n

The estimated **national** share of people with at least **basic** service (%)*.

wat_lim_n

The estimated **national** share of people with **limited** service (%).

wat_unimp_n

The estimated **national** share of people with **unimproved** service (%).

wat_sur_n

The estimated **national** share of people with **surface** service (%).

*Although the JMP usually defines service by five levels, the dataset only includes four levels, with at least basic including both safely managed and basic services.

Becoming familiar with the dataset

`wat_bas_r`

The estimated **rural** share of people with at least **basic** service (%).

`wat_lim_r`

The estimated **rural** share of people with **limited** service (%).

`wat_unimp_r`

The estimated **rural** share of people with **unimproved** service (%).

`wat_sur_r`

The estimated **rural** share of people with **surface** service (%).

`wat_bas_u`

The estimated **urban** share of people with at least **basic** service (%).

`wat_lim_u`

The estimated **urban** share of people with **limited** service (%).

`wat_unimp_u`

The estimated **urban** share of people with **unimproved** service (%).

`wat_sur_u`

The estimated **urban** share of people with **surface** service (%).

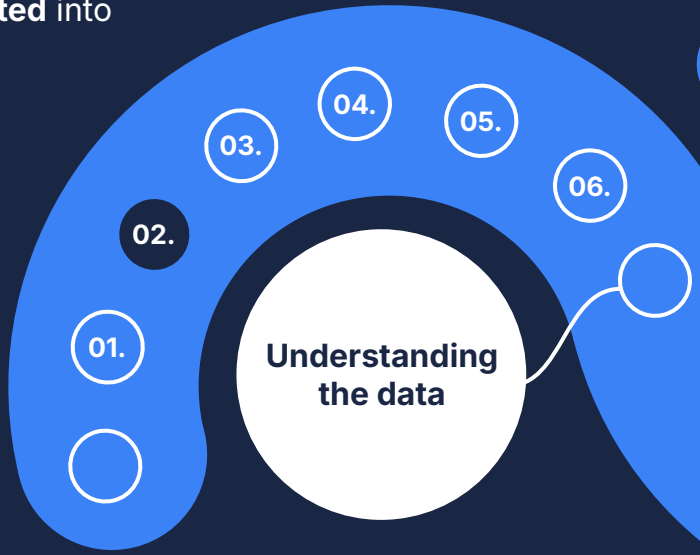
We have a total of 16 features (or columns) in our dataset, 12 of which are service-level percentage shares.

Which data types do we expect for the different features in this dataset?

02.

Importing the data

In order to investigate the data, we need to **import our data**, ensuring that it is **properly separated** into columns and rows.



A. Make sure that we have access to a **Google account**.

B. Create a **new blank spreadsheet** with Google Sheets.

C. **Download** the Estimates on the use of water (2020) dataset as a CSV.

D. **Import** the file into a blank spreadsheet.

Importing the data

When importing data, we can either assume that values are comma-separated (as per the file extension) or have Google Sheets automatically detect the separator type.

A. We see that even if we set the separator type option as "Detect automatically", the column headers are **not separated** because **semicolons** were used as separators, rather than commas as in the rest of the dataset.

```
name;year;pop_n;pop_u;wat_bas_n;wat_lim_n;wat_unimp
_n;wat_sur_n;wat_bas_r;wat_lim_r;wat_unimp_r;wat_su
r_r;wat_bas_u;wat_lim_u;wat_unimp_u;wat_sur_u
Croatia,2020,4105.268066,57.55299759,,,,,,,,100,0,
0,0
Croatia,2015,4232.874023,56.15500259,,,,,,,,100,0,
0,0
```

B. We now also need to **check for other instances** of semicolon separators to ensure that values aren't misinterpreted later.

name;year;pop_n;prop_u;wat_bas_n;wat_lim_n;wat_unimp_n;wat_sur_n;wat_bas_r;wat_lim_r;wat_unimp_r;wat_sur_r;wat_bas_u;wat_lim_u;wat_unimp_u;wat_sur_u				
Croatia	2020	4105.268066	57.55299759	...
Croatia	2015	4232.874023	56.15500259	...

Importing the data

C. Use `Data > Split text to columns` in the toolbar to **split the column headers** into the appropriate columns.

We know that we have **16 features**, so we can check that each column A to P has a column name, i.e. we have 16 columns.

D. Use `CTRL + F (Command + F)` to **search for any other instances** of semicolons and use the same method as in the previous step to split the text into two columns.

We find five occurrences in different rows across the sheet.

C. How can we find and focus only on the **affected rows**?

We know we have 16 features, so we need 16 cells in each row. We also see that missing values are represented as NAN in our dataset. So we can **count** the number of values (both numbers and text) and apply a **filter**.

Importing the data

E. Isolate and fix the incorrectly imported cells.

01. Let's create a new feature called **`value_cnt`** that counts the number of cells in a row that has a value. It should count regardless of whether the value is text or number, so we use the **`COUNTA()`** function.

What would happen if we used the function **`COUNT()`** rather than **`COUNTA()`**?

02. Add a filter to this new feature using either the **`Create a filter`** option in the toolbar or by right-clicking on the column header. Unselect "16", the value we expect, from the filter to only observe the rows that imported incorrectly.

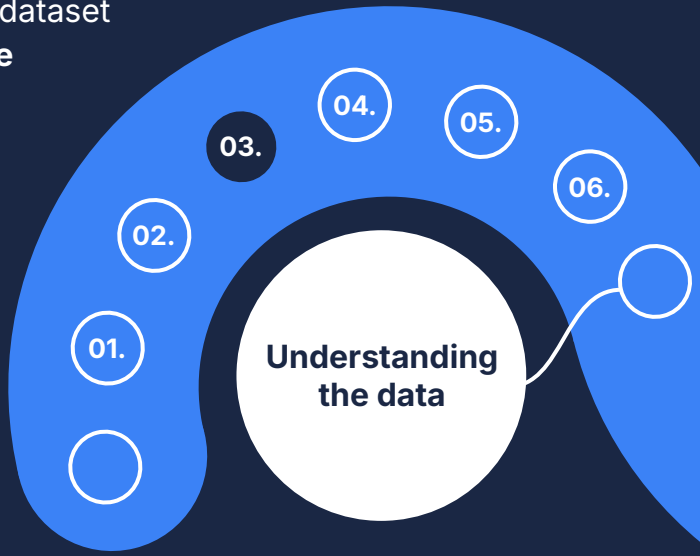
03. Fix the five rows by moving the cells after the cell with the semicolon occurrence to the correct cells, then use **`Data > Split text to columns`**. Remove the filter and check that all values in **`value_cnt`** are now equal to 16.

`Data > Split text to columns` will overwrite adjacent cells when it is applied to a single cell, however, if it is applied to an entire column, the adjacent column will automatically be shifted.

03. Investigating population size

We want to **summarise the national population** size to better understand how the dataset **represents the entire world population**.

In **2020**, the world population was estimated to be **7.821 billion**, **55%** of which lived in **urban** areas*.



A. How do the world population estimates **compare** to the provided dataset populations?

B. How does the urban population share **compare** to the rural population?

*World Cities Report 2020: The Value of Sustainable Urbanization

Investigating population size

A. Create a summary to compare the dataset population to the estimated world population.

01. Let's create a new sheet called **`Global 2020 report`**.

02. In this new sheet, determine the total national population size using the feature **`pop_n`**. Add the estimated world population (7.821 billion) to this sheet as well.

Remember, **`pop_n`** is in 1000s while the estimated world population is provided in billions. In order for these values to be comparable, we would need to ensure they are in the same unit.

03. We also want to compare the world urban population to the dataset urban population. We need the total number of people living in urban areas from our dataset, which we can either compare as a number or percentage to the world urban population value. Let's create a new feature in our dataset sheet called **`pop_u_val`**, which is the number of people living in urban areas per country (row). We will use our features **`pop_n`** and **`pop_u`** to determine this new feature.

Remember, **`pop_u`** is in percentage number. There is a difference between percentage represented by a ratio versus a number, so we need to divide our percentage by 100 before we multiply.

Investigating population size

04. In our **`Global 2020 report`** sheet, let's determine the total urban population from our dataset using the newly created **`pop_u_val`** feature.

Why couldn't we calculate the total urban population from the original urban population feature, **`pop_u`** ?

05. We know that the estimated world urban population in 2020 was 55% of the total population. Let's estimate how many people that would be from the world population of 7.821 billion in our **`Global 2020 report`** sheet.

06. We know that the estimated urban share worldwide is 55%. Let's calculate the urban share in the **`Global 2020 report`** sheet using the total national population (**`pop_n`**) and the total urban population (**`pop_u_val`**).

We see that our dataset's national and urban values are in the same value order as the global estimates. But how can we **represent our findings in a more quantitative way**?

Investigating population size

07. Let's calculate the percentage difference to determine the difference between the number of people included in our dataset and the estimated number of people in the world in 2020. Let's also calculate the percentage difference of our urban population totals and percentages.

Percentage difference refers to the difference between the relative magnitude of two values, expressed as a percentage of the average of those values. We calculate it as:

$$\texttt{Percentage difference} = \frac{|\Delta V|}{\left[\frac{\sum V}{2}\right]} \times 100 = \frac{|V_2 - V_1|}{\left[\frac{(V_2+V_1)}{2}\right]} \times 100$$

Where V_2 and V_1 are the two values being compared.

The pipe symbols in the above equation represent the absolute value which is always a positive number or a zero. We can use the **`ABS()`** function in Google Sheets in the following way:
`ABS(value_2 - value_1)`

We note that the percentage difference for the number of people and percentage share of people living in urban areas differ, although they represent the same intrinsic value. Why is that?

Investigating population size

B. Create a visualisation to compare the share of the national population living in urban versus rural areas.

01. Let's create a line chart on our **`Global 2020 report`**. Our independent variable (the cause) on the x-axis will be the national population size (**`pop_n`**). Our dependent variables (the effect(s)) on the y-axis are the urban and rural population share percentages.

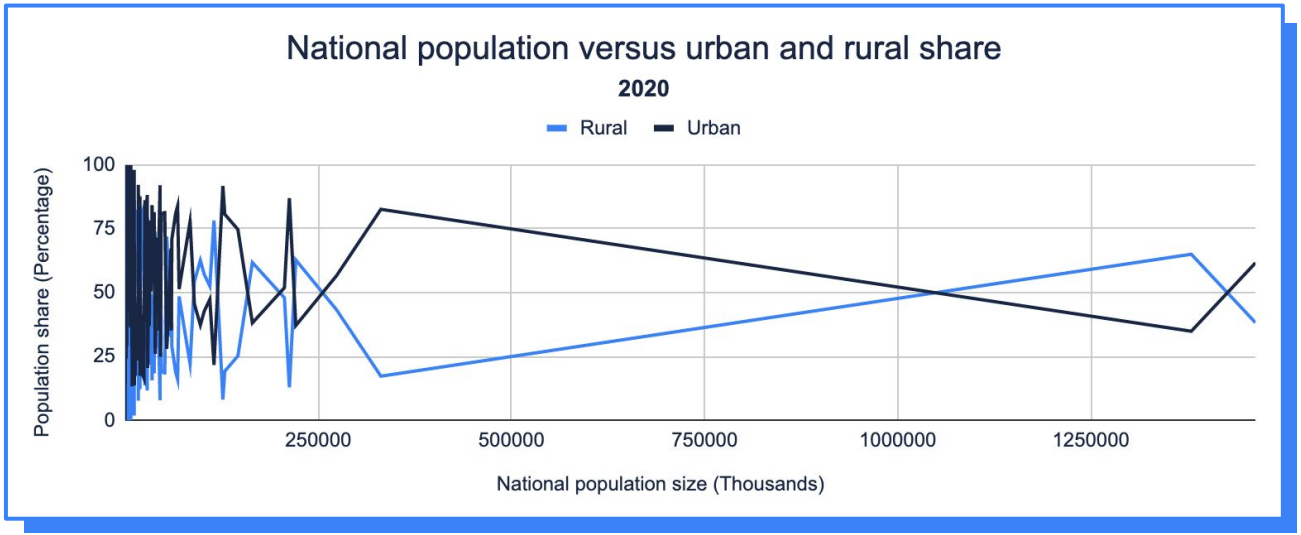
02. We already have the urban share in percentage in the **`pop_u`** column. We don't have the rural share but we know we can assume that we only have two groups, urban and rural. In other words, the sum of the urban and rural shares should be equal to 100% for each country. Create a new feature called **`pop_r`** that is the rural share of the population in percentage using the features **`pop_u`** and **`pop_n`**.

03. Add **`pop_n`** as the x-axis and **`pop_u`** and the newly created **`pop_r`** features as the **`Series`** columns. Add appropriate chart and axis titles.

Which insights can we gain from this data visualisation?

Investigating population size

04. We find that our data visualisation is not very comprehensible, due to some of our **pop_n** values being much larger than our other values, in other words, **outliers**.



In which ways can we change our visualisation to be more comprehensible without changing the meaning of our data or insights?

Investigating population size

05. To deal with the outliers in our dataset and thereby increase the readability of our data visualisation we have multiple options, each with its own advantages and disadvantages:

   a. Delete the outliers from our dataset.
   Advantage: We don't have to change anything in our data visualisation as it would update automatically.
   Disadvantage: We lose information about some of the countries included in the dataset in this visualisation and any future analysis we might need to do.

   b. We edit our visualisation by adding a maximum cut-off value to our x-axis.
   Advantage: Our visualisation is more comprehensible and we don't lose information in our dataset for any future analysis we might need to do.
   Disadvantage: Our visualisation doesn't represent the entire dataset, only a subset of it.

   c. We change the unit of our x-axis.
   Advantage: Our visualisation is much more comprehensible and we don't lose information in our dataset for any future analysis we might need to do.
   Disadvantage: Our x-axis doesn't represent the true difference between population sizes.

Investigating population size

06. Let's consider option c, changing the unit of our x-axis. Currently, our x-axis (**`pop_n`**) is represented in thousands, i.e. if the **`pop_n = 53771.30078`** (Kenya's population size), then the actual population size is **`pop_n`** multiplied by 1000, which equates to approximately 53,771,300 people, or 53.77 million people. Let's create a new feature in our original dataset sheet called **`pop_n (m)`** which is the national population size **rounded up to the nearest million**.

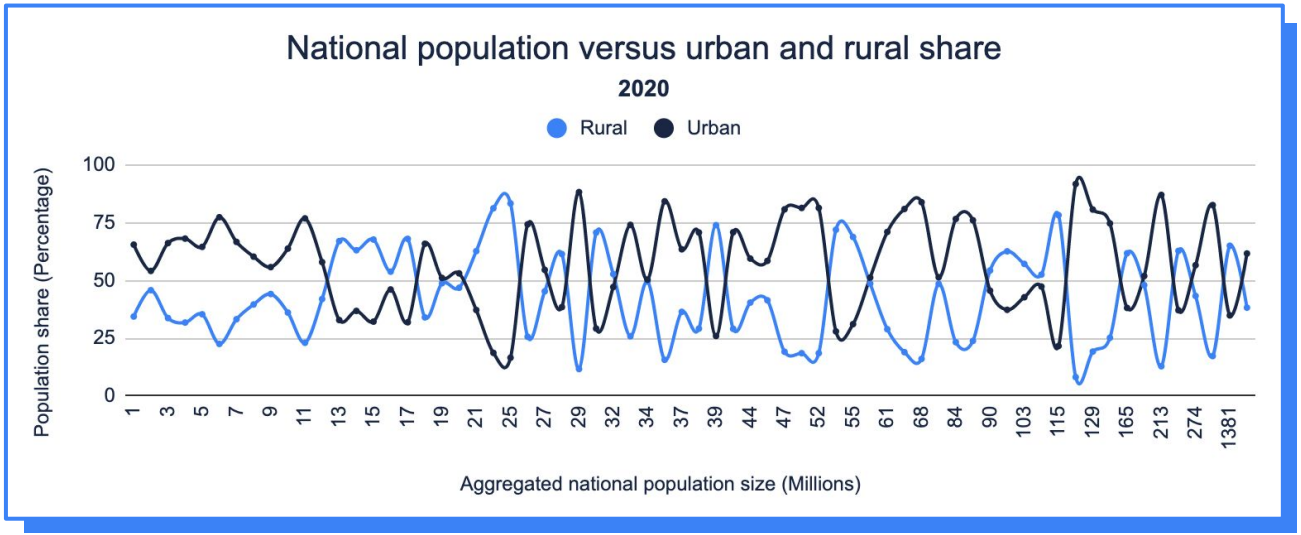
What would happen if we used the **`ROUND()`** rather than the **`ROUNDUP()`** function in Google Sheets?

07. Change the x-axis column in the previously created line chart to the newly created **`pop_n (m)`** column, selecting the **`Aggregate`** option in the chart editor under the x-axis options, and **`Average`** as the way to aggregate for both series we added previously. Remember to update the x-axis title to appropriately represent the new feature we are using.

Why are we aggregating and how is it different to aggregating previously with the x-axis in thousands? Why are we using the average function to aggregate rather than sum, count, or any other function?

Investigating population size

08. We decided to use smooth lines for our line chart and to add a point representation of our aggregated data point.



Considering that the sum of the urban and rural shares should be equal to 100% for each country, is there any other type of chart that might have been more appropriate to use in this case?

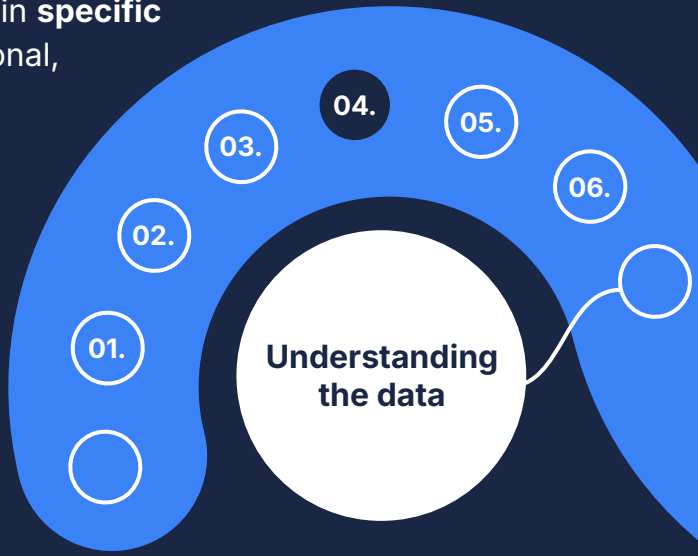
04. Investigating access by area

We want to investigate **what access to water** at the different service levels looks like for people in **specific types of areas** (national, urban, and rural).

We'll use the **measures of central tendency** and **spread**.

A. What is the **tendency** and **spread** of the different water access features?

B. How do these **measures** of the water access compare across different types of areas?



Investigating access by area

A. Calculate the measures of central tendency and spread of the four national service levels.

01. In our previously created **`Global 2020 report`** sheet, let's determine the maximum of each of the four national water access columns, i.e. `wat_bas_n`, `wat_lim_n`, `wat_unimp_n`, and `wat_sur_n`.

02. We see that our maximum for `wat_bas_n` exceeds 100%, which is not possible since 100% access means that every person in that country has access to the service. Let's go back to our dataset sheet and create a new feature called `wat_bas_n (rounded)`. We only round up to the decimal to which our value exceeds 100% because we are not rounding any of our other three service columns. We can use a filter on our new column to check that we now only have values below 100%.

Erroneous data entries are often due to the different ways in which data are collected. In this case, having a percentage value that exceeds 100% is probably due to the value being calculated from various different metrics and/or other values.

Why do we calculate a new feature to round rather than using the **`Decrease decimal places`** option in the toolbar?

Investigating access by area

If we only wanted to round the values that exceed 100% down to 100%, how would we be able to use a conditional statement to do that?

03. On the **`Global 2020 report`** sheet where we calculated the maximum of each of the water levels, we change the reference to **`wat_bas_n`** on the function to the newly created feature, **`wat_bas_n (rounded)`**.

04. We see that we get a #VALUE! error for our maximum function. Looking at the typical values in our **`wat_bas_n (rounded)`** column using a filter, we see that we have #VALUE! errors on our rounding function. Let's use the filter and change the #VALUE! entries in this column to a text **`NAN`** value. Now our maximum function should work in the **`Global 2020 report`** sheet.

How could we have used the previous conditional statement to account for any error values?

05. Since we've already determined the maximum values, we know that we don't have any values above approximately 37% for **`wat_lim_n`**, **`wat_unimp_n`**, and **`wat_sur_n`**. Let's also determine the minimum values for all four national service levels in the **`Global 2020 report`** sheet.

Investigating access by area

06. For a normal distribution, we expect that the mean, median, and mode would be equal. Let's calculate these three features for each of the four service level columns namely, `wat_bas_n`, `wat_lim_n`, `wat_unimp_n`, and `wat_sur_n`.

07. We see that the mean, median, and mode are not equal for any of the four features. Let's also calculate the interquartile range (IQR) and standard deviation for these four features to gain some insight into how the data are distributed.

Remember, the interquartile range cannot be calculated with a single function in Google Sheets. We need to use functions to determine the first and third quartiles in order to calculate the IQR.

08. For `wat_bas_n`, we see that the mean is relatively high and the interquartile range is about a tenth of the range, which indicates that our data are concentrated around a point closer to 100% than 0%. This means that the majority of people represented in the data have access to at least basic water services on a national level.

How can we visualise our measures of central tendency and spread for the different water access features for each of the areas, national, urban, and rural in order to easily compare the results?

Investigating access by area

B. Visualise the five-number summary of the four access features across the three different types of areas.

09. Let's create a box and whisker diagram (called a candlestick chart in Google Sheets) in the **`Global 2020 report`** sheet from our measures of central tendency and spread summary for all 12 features namely, `wat_bas_n`, `wat_lim_n`, `wat_unimp_n`, `wat_sur_n`, `wat_bas_r`, `wat_lim_r`, `wat_unimp_r`, `wat_sur_r`, `wat_bas_u`, `wat_lim_u`, `wat_unimp_u`, and `wat_sur_u`.

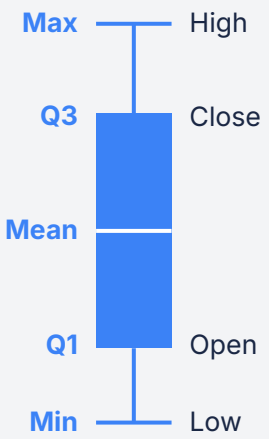
Each of the 12 features will represent one instance of a box and whisker plot, each of which will be represented on the x-axis. We'll interpret each of the candlestick chart values in the following way:

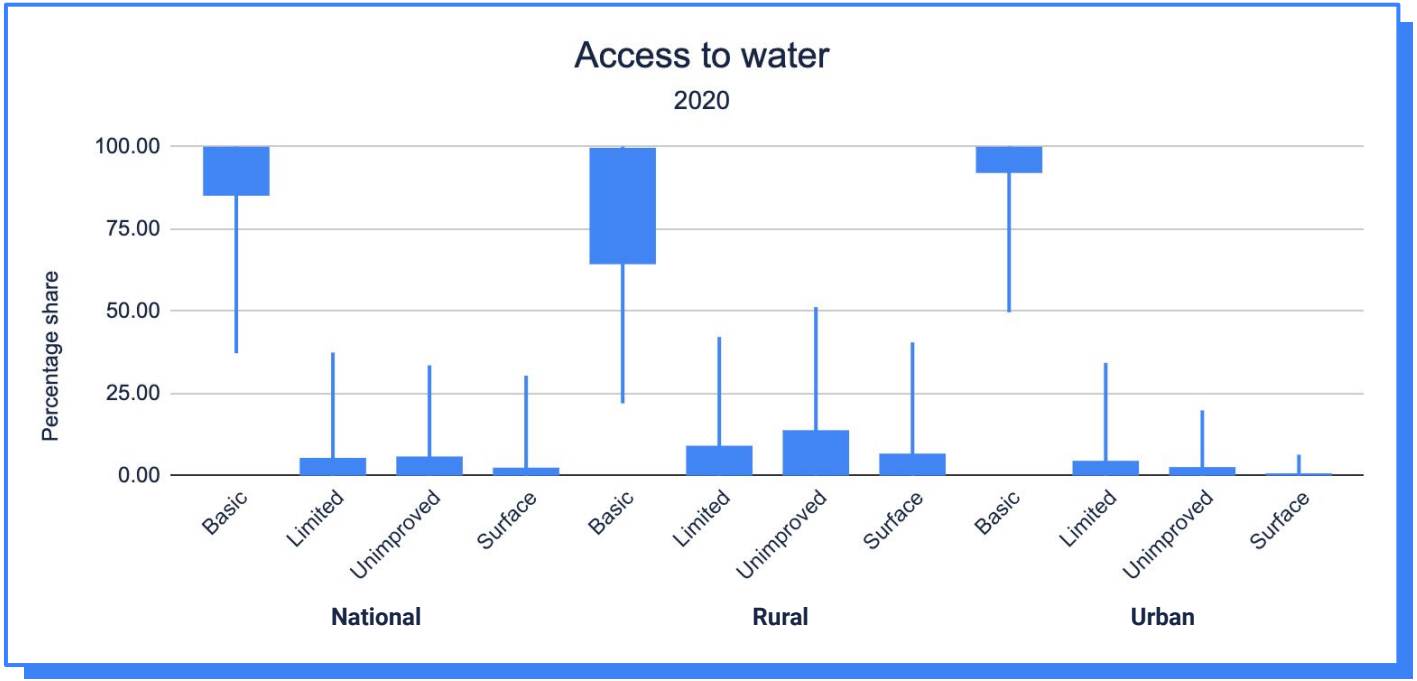
a. Low and high are our **minimums** and **maximums**, represented by the starts and ends of the whisker lines.
b. Open and close are our **first** and **third quartiles** (Q1 and Q3) which represent the starts and ends of the boxes.

Remember, the entire box and whisker represents the range of the feature, while the box represents the interquartile range. The mean is located in the middle of the box.

Investigating access by area

Box and whisker interpretation



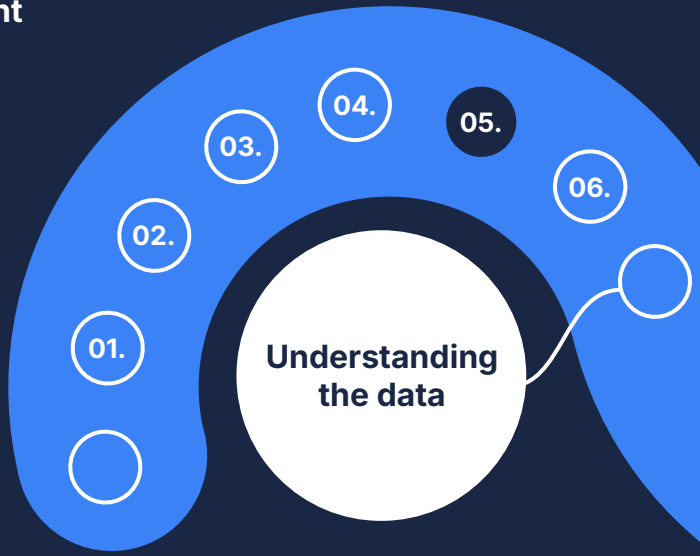


Based on the box and whisker plot, what can we say about the distributions of the different features?

05.

Investigating
access by population size

We want to investigate what **access to water** at the different service levels looks like for **different population sizes**.



A. What does the **national access** to water look like based on national **population** size?

B. What does the **urban access** to water look like based on urban **population** size?

C. What does the **rural** access look like?

Investigating access by population size

A. Visualise the **national access** to water on all four levels based on the **national population** size.

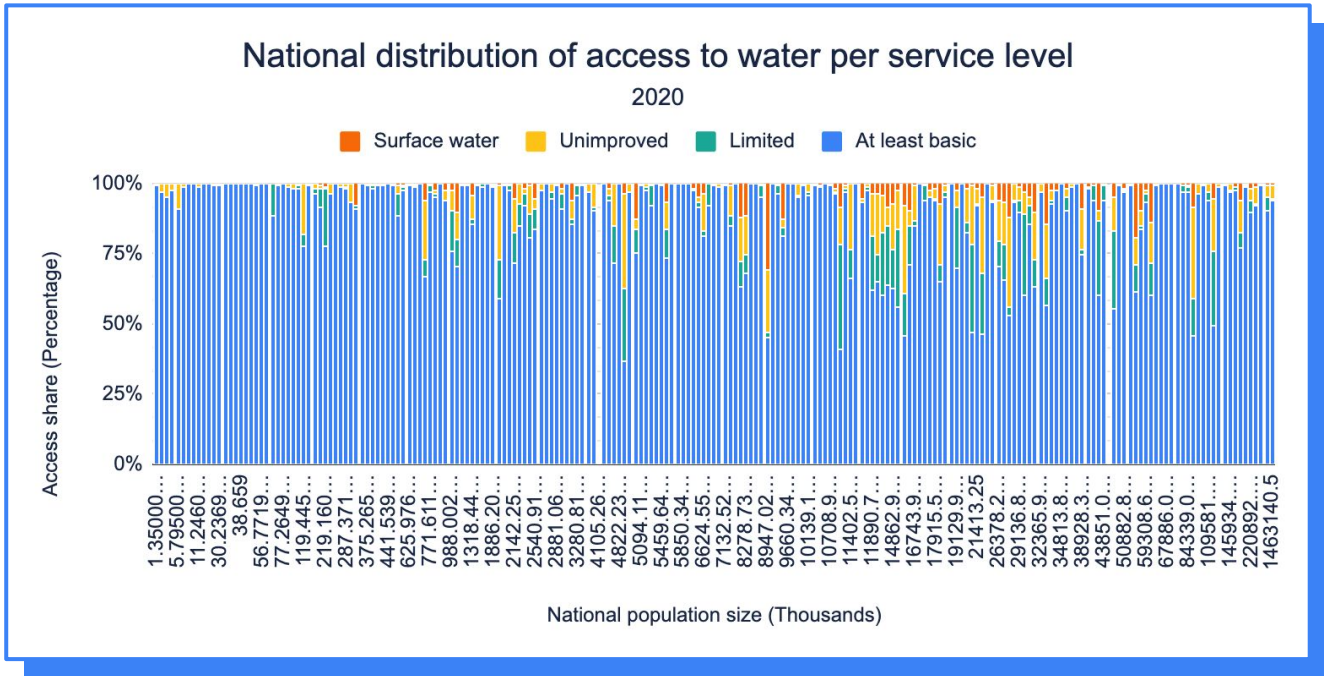
01. In our previously created **`Global 2020 report`** sheet, let's create a bar chart. Since we want to investigate how population size affects access levels, population size (**`pop_n`**) will be our independent variable (the cause), and our four access features (`wat_bas_n`, `wat_lim_n`, `wat_unimp_n`, and `wat_sur_n`) will be the dependent variables (the effects).

02. Because our access levels are percentages and we know that an individual or household can only be at one of those levels, we'll change our bar chart to a **`100% stacked column chart`**.

03. As in our investigation on population size previously, we observe that using the original population sizes in thousands is relatively unintuitive because the bars are too narrow and we have too many of them.

Remember, each bar now represents a single country (row) in our dataset, although not every country's population size is a label due to space constraints.

Investigating access by population size



How would our visualisation change if we rather used **`pop_n (m)`** which is the national population size **rounded up to the nearest million**?

Investigating access by population size

B. Visualise the **urban access** to water on all four levels based on the **urban population**.

01. In our previously created **`Global 2020 report`** sheet, let's create another **`100% stacked column chart`**, similar to what we created in A, but only for urban areas. Let's use urban population share (**`pop_u`** in percentage) as our independent variable.

02. In order to avoid a messy bar chart, we are going to create a new feature called **`pop_u (rounded)`**, which is the urban population share (**`pop_u`**) rounded to the nearest whole number. Use this new feature as the x-axis in the **`100% stacked column chart`**, and set the aggregations to **`Average`**.

03. We notice that our x-axis isn't arranged from zero to a hundred, as expected. Let's order our dataset based on **`pop_u (rounded)`**, before we consider any insights.

Consider how our national bar chart now has an unordered x-axis. Why is that, and how can we change our dataset or visualisations so that all of our bar charts are always in the correct order?

Investigating access by population size

C. Visualise the **rural access** to water on all four levels based on the **rural population**.

01. In our previously created **`Global 2020 report`** sheet, let's create another **`100% stacked column chart`**, similar to what we created in B, but only for rural areas. Let's use rural population share as our independent variable.

02. We are again going to create a new feature called **`pop_r (rounded)`**, which is the rural population share (**`pop_r`**) rounded to the nearest whole number. Use this new feature as the x-axis in the **`100% stacked column chart`**, and set the aggregations to **`Average`**.

Comparing the 100% stacked column charts for national, urban, and rural, what can we say about access? Is it related to population size or area type share?

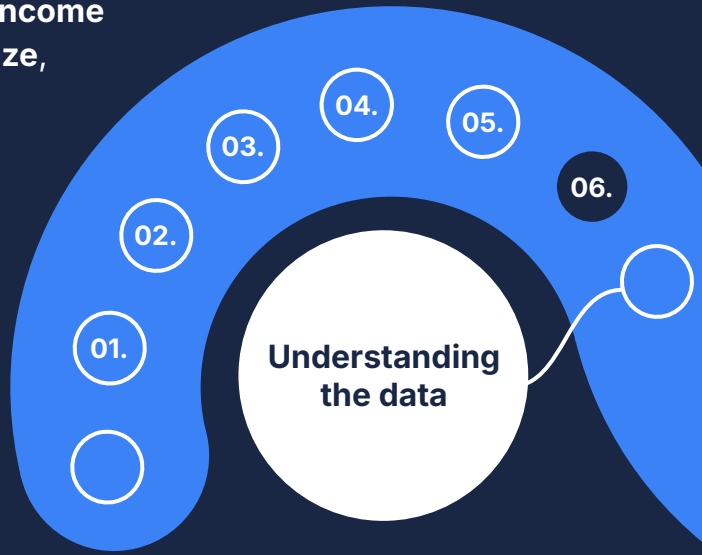
06.

Investigating
access by income group

We want to investigate the **relationship between** GNI (gross national income) or **income group**, **population size**, **urbanisation**, and **national water access**.

Economies are classified into **four income groups** based on gross national income (GNI) per capita.

A. What is the effect of **national population size** and **urbanisation** on **GNI** and **water access**?



Investigating access by income group

A. Summarise the dataset to group and investigate by the four income groups.

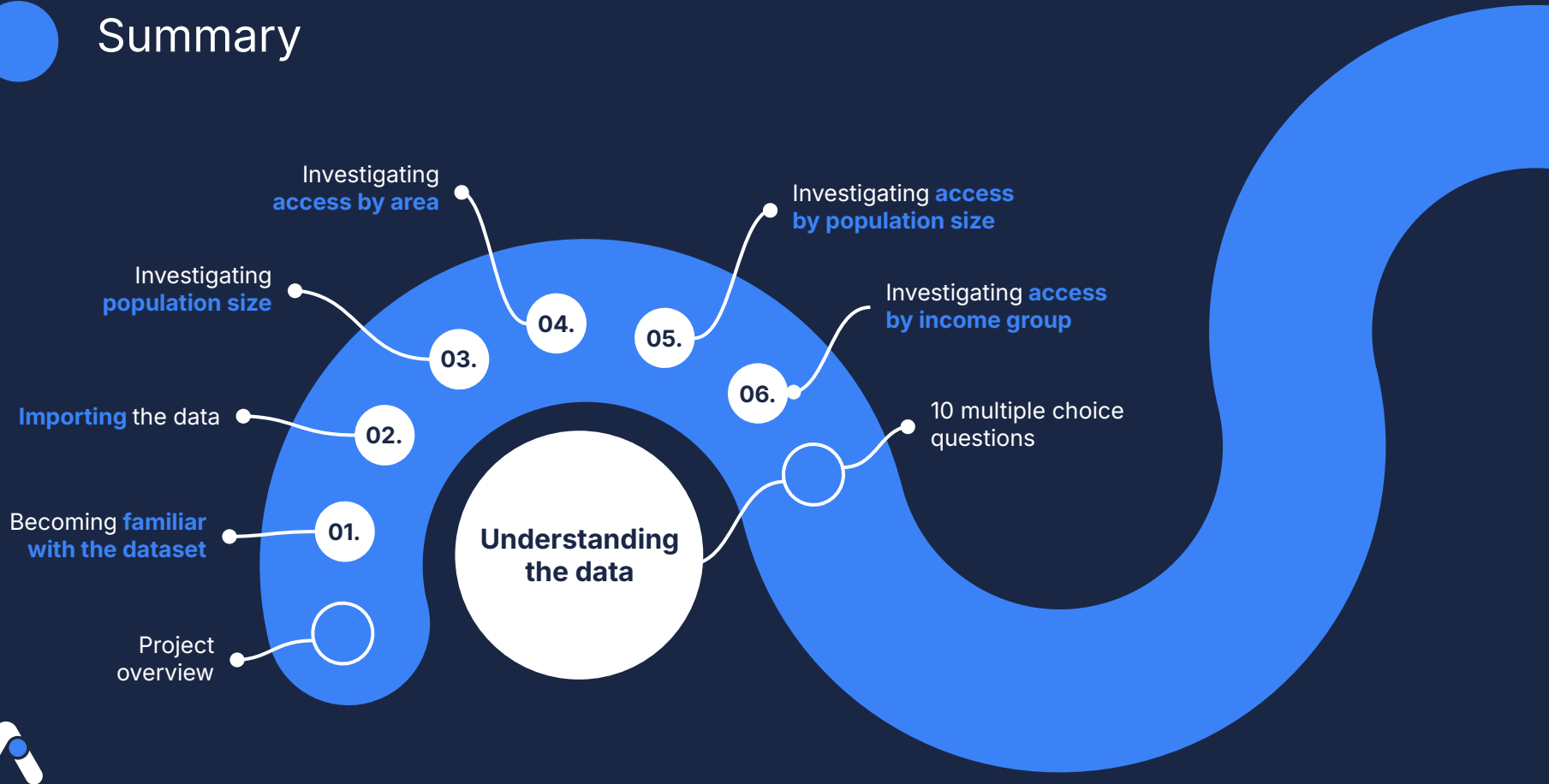
01. In our previously created **`Global 2020 report`** sheet, let's create a pivot table. Because we want to group by income group, we will set our rows to the income group feature, **`income_group`**.

02. Our values are the sum of the population size (**`pop_n`**), the average urban share (**`pop_u`**), and the average national share of basic (`wat_bas_n`), limited (`wat_lim_n`), unimproved (`wat_unimp_n`), and surface (`wat_sur_n`) access.

03. Let's also visualise our summary data. In order to better investigate the link between income group and the other features, it would be useful if we could sort our x-axis on income groups more appropriately. Let's convert our text column **`income_group`** in the dataset sheet to numbers, where NAN is 0, Low income is 1, Lower middle income is 2, Upper middle income is 3, and High income is 4.

You can use any visualisation and any method to convert the text income groups to numbers, including creating new columns or finding and replacing values.

Summary


Project overview
Becoming familiar with the dataset
01.
Importing the data
02.
Investigating population size
03.
Investigating access by area
04.
Understanding the data
Investigating access by population size
05.
Investigating access by income group
06.
10 multiple choice questions

Integrated project: Access to drinking water

Estimates on the use of water (2020)

You should have at least the following in the **imported dataset sheet:**

01. Becoming **familiar with the dataset**
    - Original 16 features
02. **Importing** the data
03. Investigating **population size**
    - New: `value_cnt`
04. Investigating **access by area**
    - New: `pop_u_val`, `pop_r`, `pop_n (m)`
05. Investigating **access by population size**
    - New: `wat_bas_n (rounded)`
06. Investigating **access by income group**
    - New: `pop_u (rounded)`, `pop_r (rounded)`

23 features

Including the original and newly created features

Integrated project: Access to drinking water

Global 2020 report

You should have at least the following in the **newly created sheet:**

03. Investigating **population size**

- A summary of the dataset population size and estimated world population, which includes urban percentage share and the percentage difference between all of the features.
- A line chart of the national population versus the urban and rural population shares.

04. Investigating **access by area**

- The maximum, minimum, mean, mode, median, first and third quartiles, the interquartile range, and the standard deviation for each of the 12 water access features.
- A box and whisker plot for all 12 water access features.

05. Investigating **access by population size**

- Three 100% stacked column charts, one each for national, rural, and urban population size or percentage versus the four different service levels.

Global 2020 report

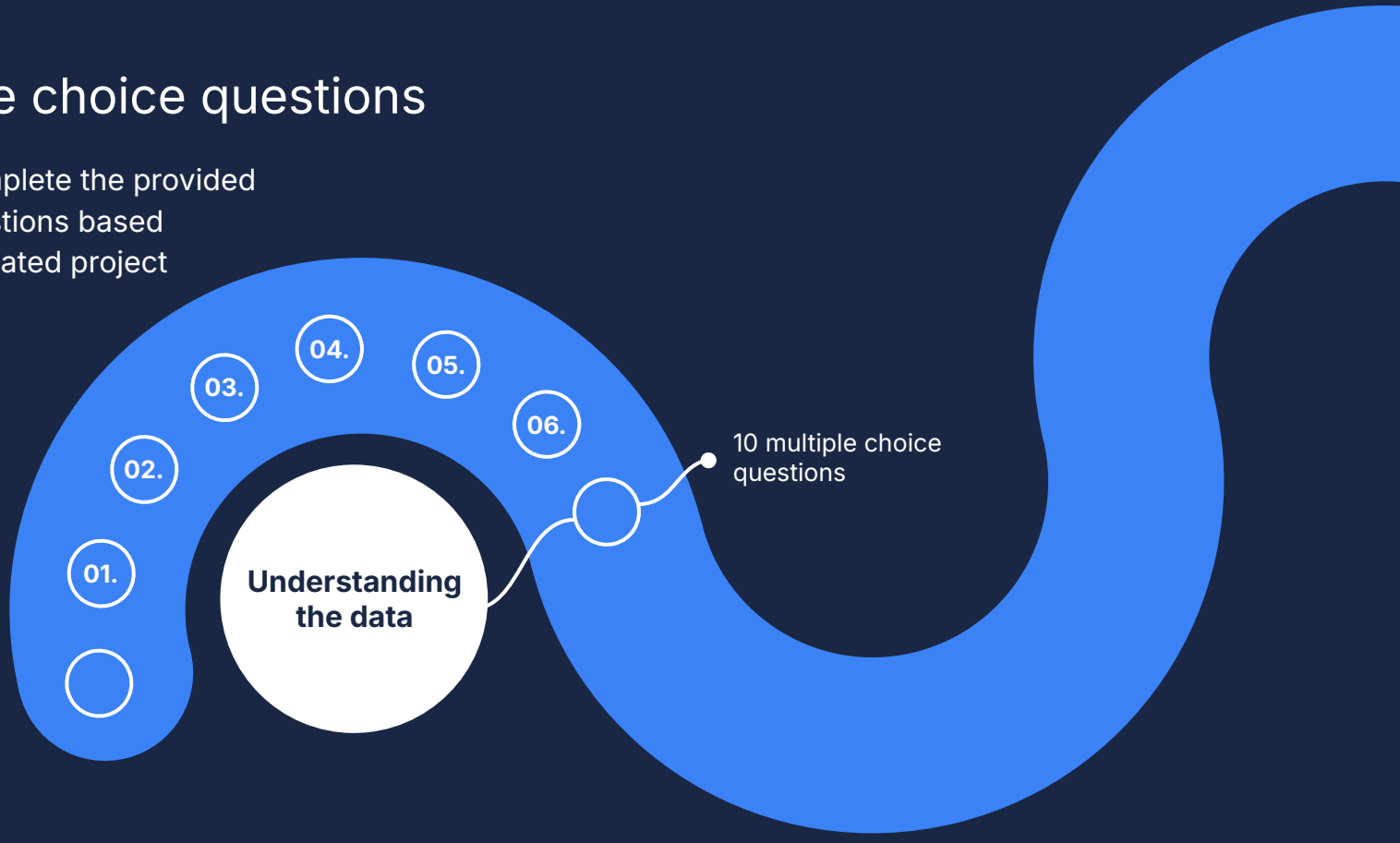
06. Investigating **access by income group**

- A pivot table for income group versus the sum of the national population and the averages for the urban population, basic, limited, unimproved, and surface access shares.
- A visualisation (of choice) for income group versus the different average shares in the created pivot table.

You **will need to create** the additional features, summaries, and visualisations as per the project instructions in order to **complete the compulsory assessment** (MCQs).

MCQs
Multiple choice questions

You will need to complete the provided multiple choice questions based on the various integrated project sections.



Understanding the data MCQs

01. What is the percentage difference between the dataset and the estimated world urban population size (the total number of people living in urban areas) for 2020?

02. True or false? Based on the national population versus urban and rural share visualisation, where the x-axis is the aggregated national population size and the y-axis is the population share in percentage, the average population shares across all population sizes are approximately 50% because the chart lines fluctuate equally above and below the 50% line.

03. True or false? The distribution of the national basic service feature (wat_bas_n) is more similar to the distribution of the urban basic feature (wat_bas_u) than it is to the national limited feature (wat_lim_n).

04. What is the interquartile range of the rural surface service feature, wat_sur_r?

Understanding the data MCQs

05. Which of the following statements are true based on the created 100% stacked column chart for urban population share versus access to the various water levels?

- Countries with greater urban population shares are more likely to provide basic water service than countries with smaller urban population shares.
- Countries with smaller urban population shares are more likely to provide basic water service than countries with greater urban population shares.
- Basic, limited, unimproved, and surface-level urban access do not increase or decrease based on the share of urban population, and we can therefore not estimate any type of relationship between water access and urbanisation.
- Basic, limited, unimproved, and surface-level urban access increase and decrease based on the share of urban population, and we can estimate that the relationship between water access and urbanisation is constant.

06. True or false? Based on the created 100% stacked column chart for rural population share versus access to the various water levels, there are countries with approximately 100% access to the basic service level across all rural population shares (0% to 100% share of rural population).

07. Based on the created pivot table, what is the national average percentage of access to limited services (wat_lim_n) for low-income countries?

Understanding the data MCQs

08. Which of the following statements are true for population sizes and shares across the different income groups according to the dataset?

- More people included in this dataset live in low-income countries than in any of the other types of economies.
- High-income countries are on average more urbanised than low, lower-middle, and upper-middle-income countries.
- More people included in this dataset live in lower-middle-income countries than in any of the other types of economies.
- On average, the greater the GNI the more urbanised.

09. Which of the following options has the highest national percentage for each of the service levels for the different income groups according to the pivot table?

- Basic access in high-income countries; Limited access in low-income countries; Unimproved access in low-income countries; Surface access in low-income countries
- Basic access in unidentified economies (NAN); Limited access in lower-middle-income countries; Unimproved access in lower-middle-income countries; Surface access in lower-middle-income countries
- Basic access in unidentified economies (NAN); Limited access in low-income countries; Unimproved access in low-income countries; Surface access in low-income countries
- Basic access in high-income countries; Limited access in lower-middle-income countries; Unimproved access in lower-middle-income countries; Surface access in lower-middle-income countries

Understanding the data MCQs

10. True or false? Visualising the pivot table values for national access versus income group indicates that as urbanisation increases, so does the share of the population with basic water access, and as GNI increases, limited, unimproved, and surface water access decreases.