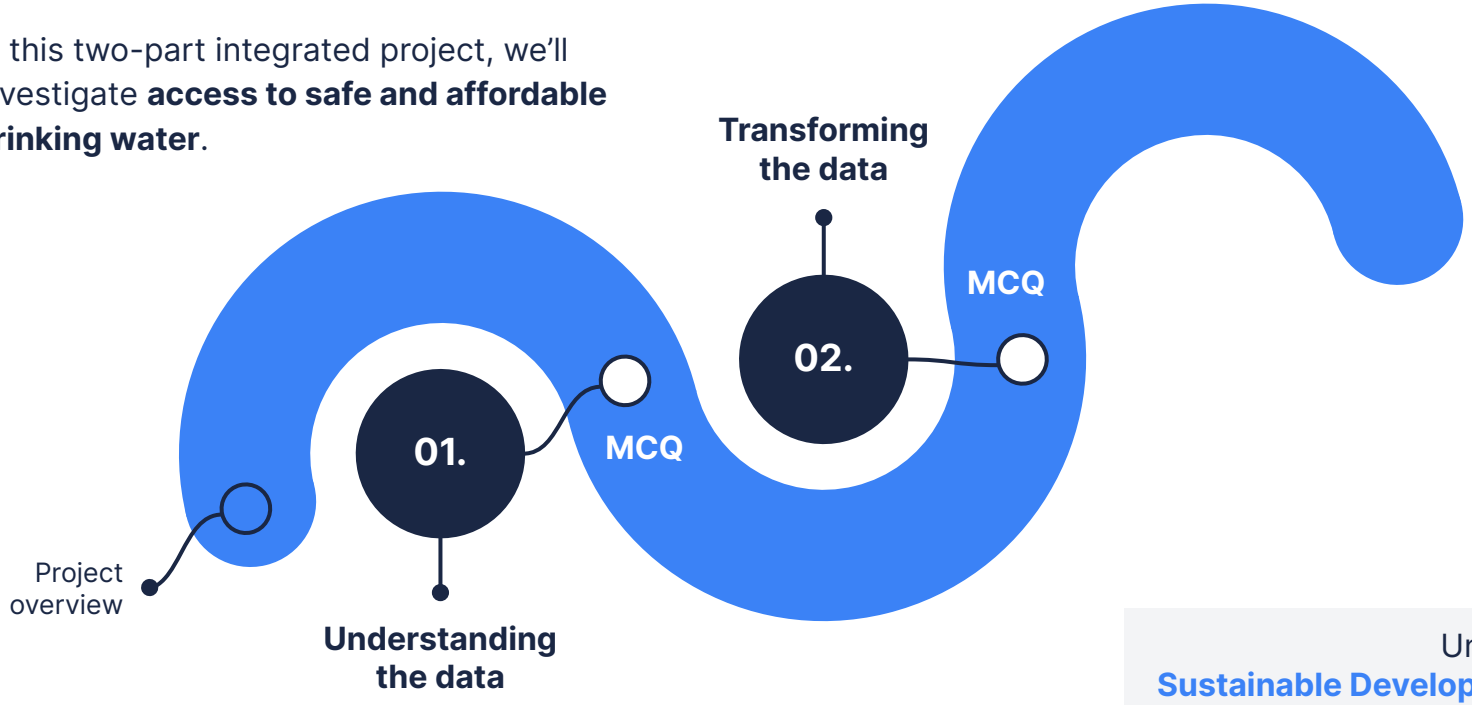
EXPLORE AI ACADEMY

Integrated project: Access to drinking water

Transforming the data

Integrated project overview

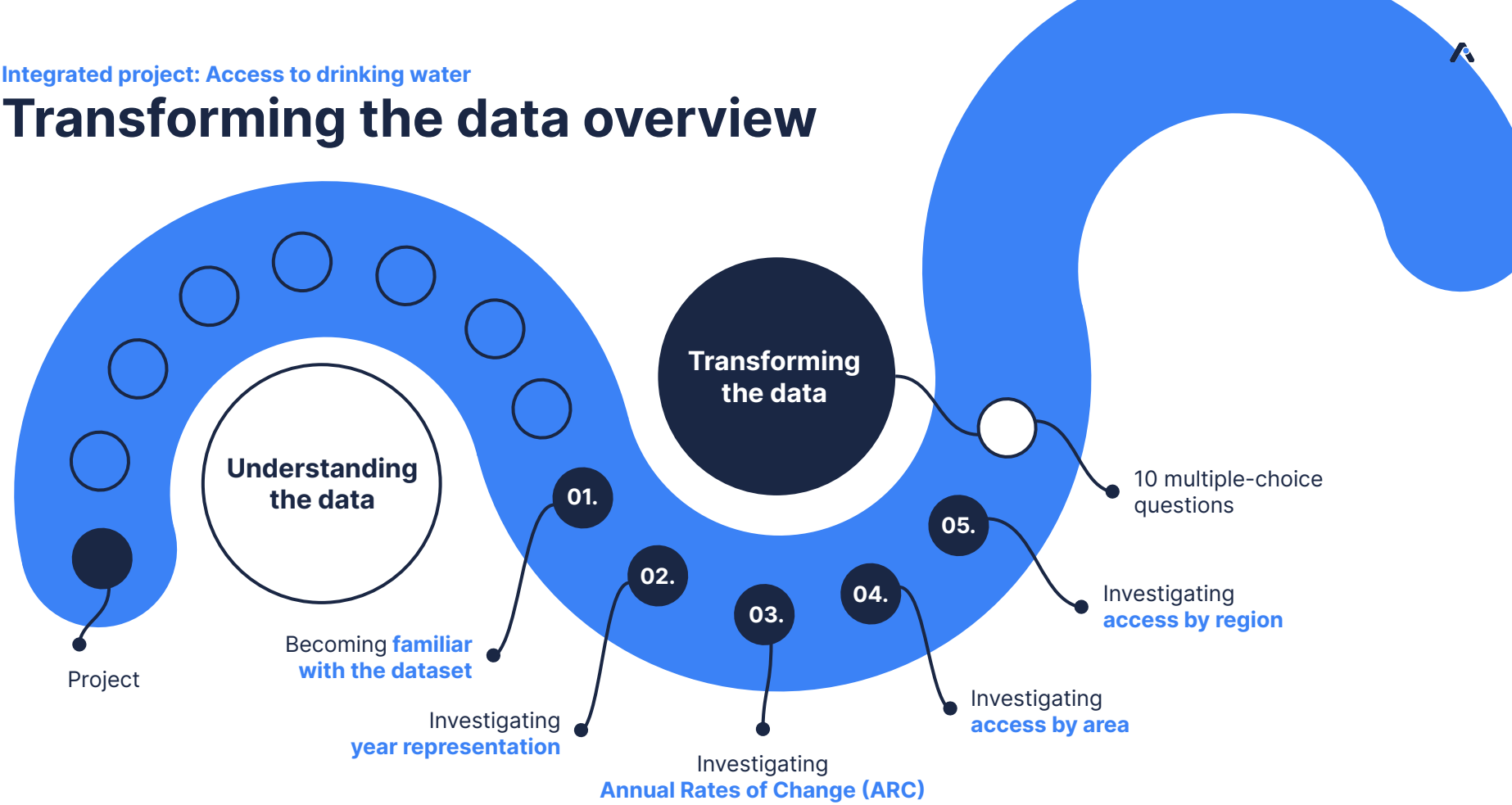
In this two-part integrated project, we'll investigate **access to safe and affordable drinking water**.



United Nations
Sustainable Development Goal 6
Clean water and sanitation

Integrated project: Access to drinking water

Transforming the data overview



01. Becoming **familiar with the dataset**

How does the **imported dataset differ** from the Estimates on the use of water (2020) dataset?

A.

Transforming the data

01.
02.
03.
04.
05.

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

Now we want to see how these features are **represented in the dataset**.

Becoming familiar with the dataset

> We'll take a look at the same WHO/UNICEF JMP **Estimates on the use of water** dataset, but now it ranges from **2000** to **2020.**

A. We **import the data** (`Estimates on the use of water (2000-2020).csv`) to see how it differs from what we used in the previous part of the project.

Considering the column names, we observed that the **`income_group`** feature has been removed and a **`year`** feature has been added.

Considering that the dataset title includes "2000-2020", how can we confirm that our dataset does represent this time period?

How do we know if the **difference between years per country** are similar?

02.

Investigating **year representation**

A. During which years were this data recorded?

B. What is the **average** number of **years** between data collections per country?

Transforming the data

01.

02.

03.

04.

05.

In order to **understand what our dataset represents**, we need to investigate when and how often data were collected.

Becoming familiar with the dataset

A. To observe which **years** are represented for which countries, we **sort** both by **`name`** and **`year`**.

01. If we sort without considering the row containing column names, they are sorted with the other rows and it might not end up where we expect it to be. To avoid this, right click on `row 1 > View more row actions > Freeze up to row 1`.

02. If we were to sort only by **`name`** or **`year`** using right click > `Sort Sheet A to Z` (or Z to A), we won't be sure whether we are considering both parameters. We will rather use the `Data > Sort range > Advanced range sort options`. Because we are now sorting on range and not the entire sheet, we need to make sure that we select our entire dataset. Add two sorts: `Sort by Column A` (name) then by `Column B` (year), both `A → Z`.

At first glance, it seems like the data were not recorded for every year in each country, so when we want to look at the change over time, we will have to account for that. How can we determine whether the data were only collected twice per country?

Becoming familiar with the dataset

B. Calculate the average difference in years for data entries per country.

01. In the dataset sheet, **create a new column** (feature) called **`y_diff`** (year difference).

From the first few rows in the sorted sheet, we assume we only have two entries per country, so theoretically, we can subtract the second **`year`** from the first **`year`** per country (**`name`**), i.e. **`y_diff = year(n+1) - year(n)`**, where n is the row number.

02. Use an **if statement** to only subtract two years if the country in **`name`** is the same for the new feature **`y_diff`**. If they are not the same, return an empty string. We are assuming that per country, the text will be exactly the same.

Note, the comparison operator for strictly equal in Google Sheets is a single equal sign ("**`=`**") rather than double as we use in pseudocode. We can also use the **`EQ()`** function to check if the two strings are the same.

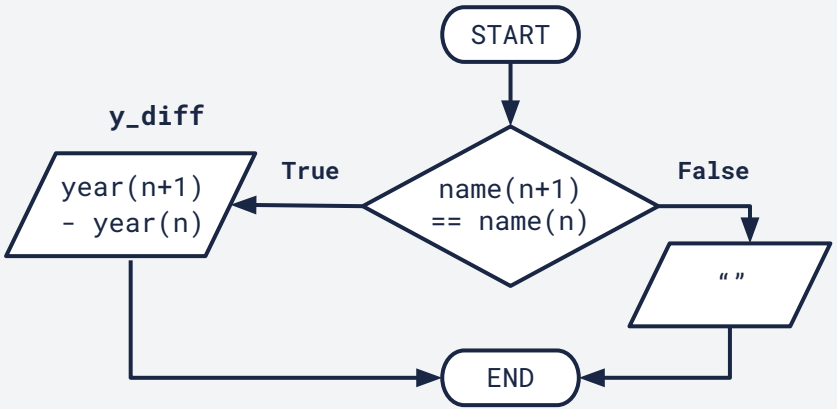
What can we expect to happen to our **`y_diff`** statement if we incorrectly assumed that the text will be exactly the same per country?

Becoming familiar with the dataset

What does this if statement look like?

Our condition will be the country name in the next row (**`name(n+1)`**) is strictly the same as the country name in the current row (**`name(n)`**). When the condition is **`True`**, we will calculate the year difference, but when the condition is **`False`**, we will return nothing.

Flowchart



Pseudocode

```
START

If name(n+1) == name(n) then
- y_diff = year(n+1) - year(n)
Else
- y_diff = “”
End if

END
```

Becoming familiar with the dataset

03. Because we've set our if statement to return an empty string when the names are not the same, we can assume that if we find `y_diff = 0`, we have duplicate rows. **Find and remove the erroneous duplicate** values from the dataset and fix the broken `y_diff` formula for the adjacent rows.

Why do we infer that `y_diff = 0` equates to duplicates based on our conditional statement?

04. In the newly created summary sheet, calculate the **average year difference across all countries**, rounded to two decimal places. Calculate the **minimum** and **maximum year difference** in the same sheet.

How does the average year difference compare to the minimum and maximum year difference? How does the average year difference compare to the minimum and maximum years represented in the dataset?

05. In the newly created sheet, **create a histogram** of the `year` column. Note the minimum and maximum values on the horizontal axis.

What can we say about the distribution of `year` in this dataset?

03.

Investigating
Annual Rates of Change (ARC)

What is the **ARC** for the **national**, **rural**, and **urban** areas per country? **A.**

What is the **average** of the **different ARCs** for all countries? **B.**

Transforming the data

01.

02.

03.

04.

05.

We want to see if access to **services is improving or declining** across national, urban, and rural areas.

The United Nations (UN) uses **Annual Rates of Change (ARC)** to see whether the proportion of access to drinking water is declining or improving.

Investigating Annual Rates of Change (ARC)

The United Nations (UN) uses Annual Rates of Change (ARC) to see whether the proportion of access to drinking water is declining or increasing. The **Annual Rates of Change (ARC)** is a statistical measure used to express the average yearly change rate of a variable over a certain period of time.

It's calculated by taking the difference between the end and start values of the dataset and dividing the result by the number of years that separate the two values:

$$ARC_x = \frac{P_{x,y2} - P_{x,y1}}{Y_2 - Y_1}$$

Where ARC_x is the annual rate of change for the indicator x, $P_{x,y1}$ and $P_{x,y2}$ is the estimate for indicator x in reference to year 1 (Y_1) and year 2 (Y_2) in percentage.

In **Google Sheets**, we need to do a row calculation between two years for the same country to calculate the ARC. Based on our column names, our ARC equation is*:

```
ARC_x = (wat_bas_x(n+1) - wat_bas_x(n))/(year(n+1) - year(n))
```

Where **_x** represents the different areas; **ARC_n** (national) we calculate using **wat_bas_n**, **ARC_r** (rural) using **wat_bas_r** and **ARC_u** (urban) using **wat_bas_u**. Remember, our **wat_bas_x** values are already in percentages.

*Since we ordered by year in ascending order, we need to subtract the greater year value from the smaller value.

Investigating ARC

A. Calculate ARC per country, i.e. only calculate the ARC between two years of the same country name.

01. **Create three new columns** namely **`ARC_n`**, **`ARC_r`**, and **`ARC_u`** in the dataset sheet. These columns represent the Annual Rates of Change (ARC) for the country's national (n), rural (r), and urban (u) populations.

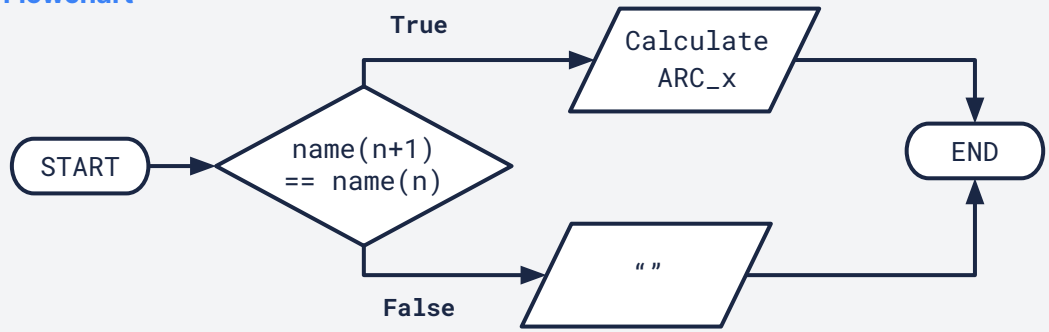
02. **Calculate the ARC** for each of the new features when the two country names are the same. As with calculating **`y_diff`**, we only want to calculate the ARC between two years when it's the same country. In other words, we will only calculate ARCs if the country name is the same for the two years.

Considering that we have the same condition on **`year`** and **`name`** for **`ARC_n`**, **`ARC_r`**, and **`ARC_u`** as **`y_diff`**, what would the flowchart and pseudocode look like and how would we change the Google Sheets formula?

Since we are calculating the same value based on different columns, we can use both absolute and relative references in the formula to ensure that the correct features are being used for the condition and calculation of the different ARC columns.

Investigating Annual Rates of Change (ARC)

Flowchart



Pseudocode

```
START

If name(n+1) == name(n) then
- ARC_x = (wat_bas_x(n+1)-wat_bas_x(n))/(year(n+1)-year(n))
Else
- ARC_x = “”
End if

END
```

Our condition will be: the country name in the next row (**name(n+1)**) is strictly the same as the country name in the current row (**name(n)**). When the condition is **True**, we will calculate the **ARC_x**, but when the condition is **False**, we return nothing.

Investigating Annual Rates of Change (ARC)

Let's consider what the formula will look like for **ARC_n** and what we expect as the output.

	A	B	...	E	...	R		
1	name	year	...	wat_bas_n	...	ARC_n	name(n+1) == name(n)	ARC_x
2	Afghanistan	2015		61.33978081		=IF(A3=A2, (E3-E2)/(B3-B2),)	A3=A2 → True	=(E3-E2)/(B3-B2)
3	Afghanistan	2020		75.09141325		=IF(A4=A3, (E4-E3)/(B4-B3),)	A4=A3 → False	
4	Albania	2015		93.39432534		=IF(A5=A4, (E5-E4)/(B5-B4),)	A5=A4 → True	=(E5-E4)/(B5-B4)
5	Albania	2020		95.06803883		=IF(A6=A5, (E6-E5)/(B6-B5),)	A6=A5 → False	
6	Algeria	...		...		...	...	...

Why are we not using absolute references on rows?

	A	B	...	E	...	R
1	name	year	...	wat_bas_n	...	ARC_n
2	Afghanistan	2015		61.33978081		=IF($A3=$A2, (E3-E2)/($B3-$B2),)
3	Afghanistan	2020		75.09141325		=IF($A4=$A3, (E4-E3)/($B4-$B3),)
4	Albania	2015		93.39432534		=IF($A5=$A4, (E5-E4)/($B5-$B4),)
5	Albania	2020		95.06803883		=IF($A6=$A5, (E6-E5)/($B6-$B5),)
6	Algeria	...		...		...

Investigating Annual Rates of Change (ARC)

Why do we observe #VALUE! errors in all three new features (**ARC_n**, **ARC_r**, and **ARC_u**)? How can we change our formula to ensure that we don't observe these errors?

03. We see that the error is observed in rows where one or both of our **wat_bas_x** rows are "null". Add the **IFERROR** function to the IF statements in the features **ARC_n**, **ARC_r**, and **ARC_u**, so that if the output **ARC_x** value is an error, the value is replaced with "null".

Would we have observed the #VALUE! errors if our dataset treated missing values as blanks rather than the string "null"?

B. Calculate the **average**, **minimum**, and **maximum** for each of the ARC values for access to **basic** service level for each of the **three population groups**.

01. In the summary sheet, calculate the average, minimum, and maximum of **ARC_n**, **ARC_r**, and **ARC_u**.

Because we divide by the difference in years, the ARC indicates the yearly change in access in percentage points.

Investigating Annual Rates of Change (ARC)

We observe that the rural ARC value is higher than that of the urban and national population. What does this tell us about rural versus urban and national access?

It's easy to jump to the conclusion that the average rural access has increased more significantly per year than urban access because the average ARC value is higher, however, we have not considered observations where access is already 100%.

When access to basic water service per country is reported as 100% for both years, the ARC values are zero. If we calculate the average ARC value over all countries and a relatively large proportion of those values are equal to zero, our average would be lower.

In other words, a lower ARC average (as calculated before) does not necessarily indicate less progress in changing access to basic water services because it takes into account countries that already have 100% access.

Considering that our water access features haven't been rounded, how can we determine the number of countries that have full access to basic water services?

04.

Investigating **access by area**

What does the **change** in access to **basic** water look like for **different areas**? **A.**

How does the **ARC differ** between **rural** and **urban** populations? **B.**

Transforming the data

01. 02. 03. 04. 05.

In our previous investigation of access by area, we observed that **rural populations on average have lower access** to basic water services than the national or urban populations.

Now, we want to investigate whether countries have made **significant enough effort to improve** this in the years leading up to 2020.

Investigating access by area

A. Calculate the number of countries per area that have full access and Annual Rates of Change equal to zero, smaller than zero, and greater than zero.

01. In the summary sheet, calculate the **number of countries that have missing ARC values**, i.e. the number of "null" occurrences in each of the columns **`ARC_n`**, **`ARC_r`**, and **`ARC_u`**.

02. In the summary sheet, also calculate the number of countries that have full access across both years, i.e. the number of countries where access is 100% for both years reported.

As in the previous part of the project, we have access to basic water service entries that are greater than 100%, which is not possible.

a. Create three new columns in the original dataset sheet called **`wat_bas_n (rounded)`**, **`wat_bas_r (rounded)`**, and **`wat_bas_u (rounded)`** that is the original access to basic water services columns (**`wat_bas_n`**, **`wat_bas_r`**, **`wat_bas_u`**) **rounded** to zero decimal places.

Considering that we are now rounding, for example, 99.6% to 100%, what does a 100% access actually mean?

Investigating access by area

b. Create a new column called **ARC_n_full** in the original data sheet that determines **IF** the country names are the same **AND** that both **wat_bas_n (rounded)** features for that country are > 99% for both years. Return “full access” if it is true.

We’ve already used an if statement to determine whether the cells we are referring to are for the same country in the previous section. You can use the following pseudocode to calculate **ARC_n_full**:

Pseudocode

```
START

If name(n) == name(n+1) AND wat_bas_n (rounded)(n) == 100 AND wat_bas_n (rounded)(n+1) == 100 then
- ARC_n_full = “full access”
End if

END
```

c. Create two more columns for **ARC_r_full** and **ARC_u_full** that similarly calculate whether a country has full access for its rural and urban populations.

Investigating access by area

d. In the summary sheet, calculate the number of countries that have full access per population, i.e. the number of "full access" occurrences in each of the newly created columns **`ARC_n_full`**, **`ARC_r_full`**, and **`ARC_u_full`**.

03. Calculate the **number of countries that have ARC values equal to zero that doesn't already have full access** for each of the population types: national, rural, and urban.

04. Calculate the **number of countries where ARC < 0 and doesn't have full access** for each of the population types: national, rural, and urban.

05. Calculate the **number of countries where ARC > 0 and doesn't have full access** for each of the population types.

We can check for **full access** by using the **not equal operator** (**<>**) in our `COUNTIFS()` formula on the "full access" string in the **`ARC_n_full`**, **`ARC_r_full`**, and **`ARC_u_full`** columns.

We can check that we've considered all countries for the conditions: no value, full access, ARC = 0, ARC < 0, and ARC > 0, by summing the number of countries for each and comparing it to the total number of countries.

Investigating access by area

B. Calculate the difference between the Annual Rates of Change between rural and urban populations per country.

01. Create a new feature called **`ARC_diff`** in the dataset sheet that calculates the difference between the rural ARC **(`ARC_r`)** and urban ARC **(`ARC_u`)** for every second row since these rows are empty.

Remember, **`ARC_r`** and **`ARC_u`** are already in percentage points, so the difference between them will also be in percentage points.

02. We again observe #VALUE! errors. Change the difference formula to account for this error.

If we were to calculate the percentage difference (the absolute difference divided by the average of the two) between **`ARC_r`** and **`ARC_u`**, what additional error would we have, and how could we change the difference formula to account for it?

03. Create a histogram of the newly created **`ARC_diff`** feature.

05.

Investigating **access by region**

How does **ARC** compare across **different regions**? **A.**

What is the influence of **national population size** on the **ARC**? **B.**

Transforming the data

01. 02. 03. 04. 05.

The UN often uses **classification by region** as a way to group various countries and investigate a region's progress in the SDGs.

We want to investigate whether more or less **progress has been made in increasing access to basic water services** in specific regions across the world.

Investigating access by region

A. Our original dataset didn't include region information, so we'll have to amend our dataset to investigate access by region.

01. Import the **`Regions.csv`** into a new sheet.
02. Add a new column to the original dataset called **`region`** and use any **LOOKUP function** to add the **region** based on the country name.
03. In the summary sheet, use any preferred method(s) to **calculate**:
    a. The **number of countries per region**.
    b. The average Annual Rates of Change on a **national** level per region.
    c. The average Annual Rates of Change in **rural** areas per region.
    d. The average Annual Rates of Change in **urban** areas per region.

Which built-in Google Sheet functions or methods allow us to quickly group by and summarise data?

Investigating access by region

B. Visualise access by region to investigate the relationship between the national and rural Annual Rates of Change, as well as population size and region.

01. **Create a visualisation** that represents the **national ARC** versus the **rural ARC**, as well as the **region** and **national population size**.

Do we observe any patterns in this visualisation related to region or a specific relationship between the national and rural Annual Rates of Change?

Integrated project: Access to drinking water

Weaving ARCs into a story

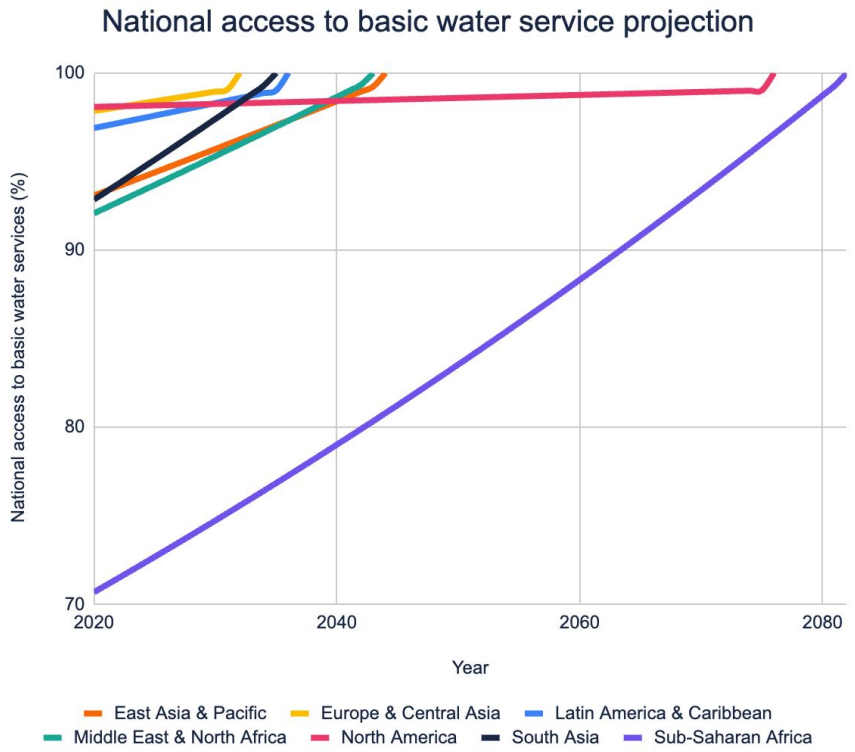
We've done an analysis of the data, calculating many different statistical measures and visualising the data.

But why does this matter and how is it actionable?

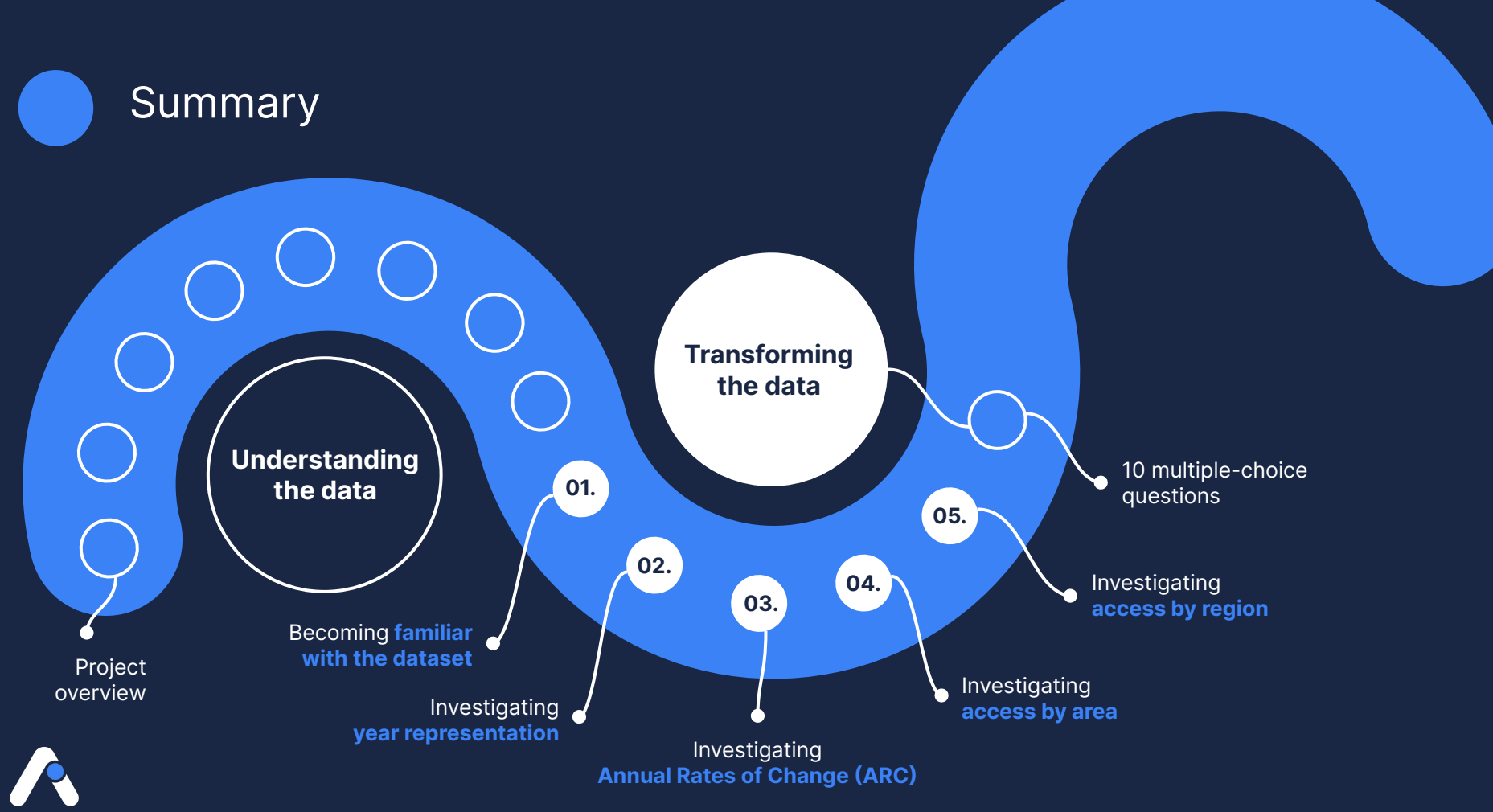
Sub-Saharan Africa, home to the largest population in the world with limited access to water, is at the centre of our findings. Despite some progress in recent years, Sub-Saharan Africa will only have full access to water by approximately 2080 if the current rate of change doesn't improve.

These findings enable us to weave a powerful narrative about Africa's water crisis.

Without significant intervention, millions of Africans will face a scarcity of clean water for the next ~60 years.



Summary


Understanding the data
Transforming the data
Project overview
Becoming familiar with the dataset
01.
Investigating year representation
02.
Investigating Annual Rates of Change (ARC)
03.
Investigating access by area
04.
Investigating access by region
05.
10 multiple-choice questions

Integrated project: Access to drinking water

Estimates on the use of water (2000-2020)

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

01. Becoming **familiar with the dataset** — Original 16 features

02. Investigating **year representation** — New: **`y_diff`**

03. Investigating **Annual Rates of Change (ARC)** — New: **`ARC_n`**, **`ARC_r`**, **`ARC_u`**

04. Investigating **access by area** — New: **`wat_bas_n (rounded)`**, **`wat_bas_r (rounded)`**, **`wat_bas_u (rounded)`**, **`ARC_n_full`**, **`ARC_r_full`**, **`ARC_u_full`**, **`ARC_diff`**

05. Investigating **access by region** — New: **`region`**

28 features

Including the original and newly created features

Global 2000-2020 report

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

02. Investigating **year representation**

- A summary of the dataset year and year difference, including the median, minimum, and maximum.
- A histogram of the year column.

03. Investigating **Annual Rates of Change (ARC)**

- The median, minimum, and maximum of each of the three newly created ARC columns based on the national, rural, and urban change in access.

04. Investigating **access by area**

- The number of countries that has no ARC value, full access, ARC < 0, and ARC > 0 for each of the three newly created ARC columns.
- A histogram of the difference in ARC values for rural versus urban areas.

Integrated project: Access to drinking water

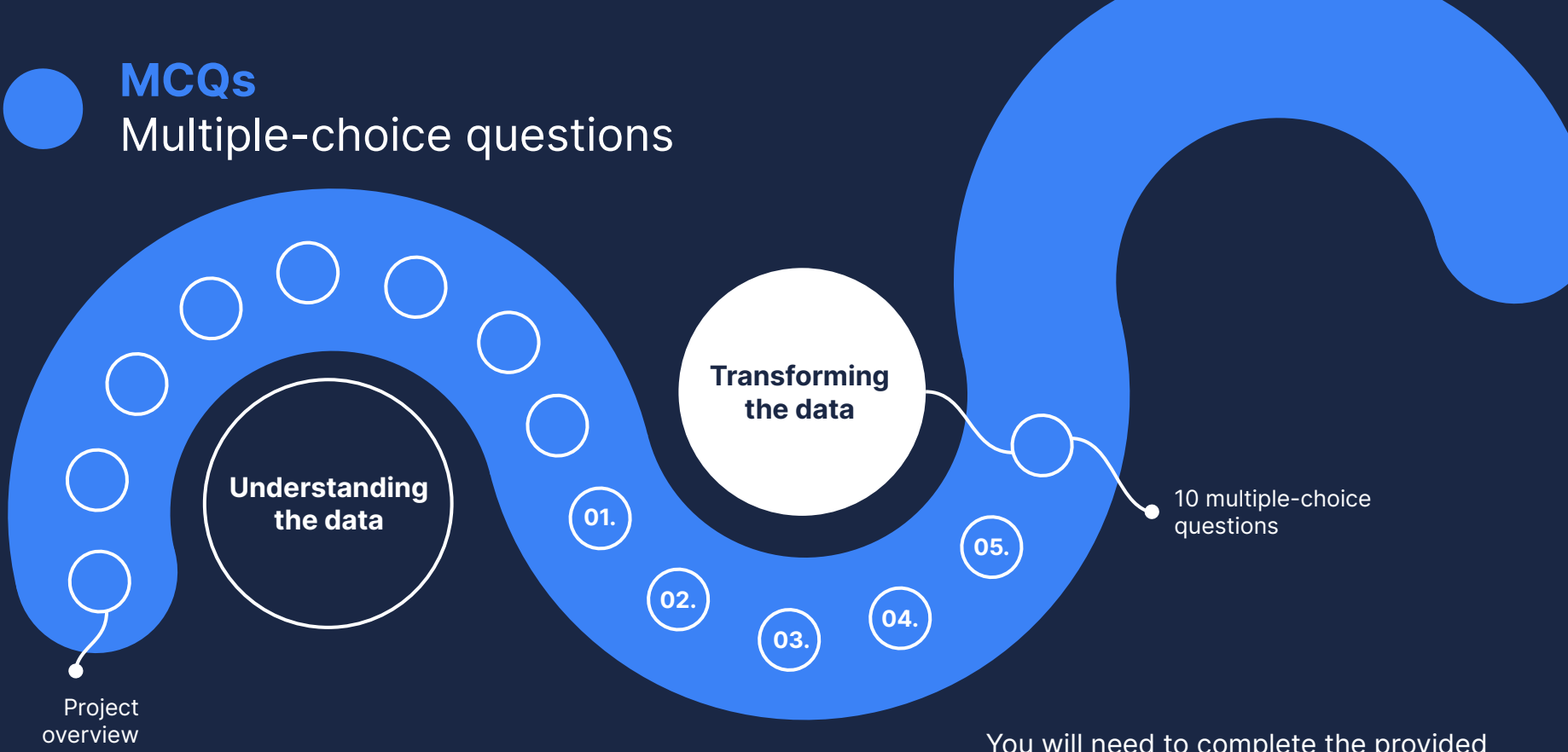
Global 2000-2020 report

05. Investigating **access by region**

- A summary per region which at least includes the number of countries per region, the average ARC on a national level, and the average ARC values in rural and urban areas.
- A visualisation (of choice) that represents the national ARC, rural ARC, region, and population size.

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.

Transforming the data MCQs

01. True or false? The dataset provided represents the entire year range from 2000 to 2020 for at least some countries.

02. What is the average year difference across all countries for the dataset?

03. Which of the following observations best represents the distribution of the year column?

- Nothing can be said about the distribution because we have too few data points.
- The distribution is neither normal, negatively, nor positively skewed since there are two distinct peaks.
- The distribution is positively skewed since the peak is to the left of the number line.
- The distribution is negatively skewed since the peak is to the left of the number line.

04. What is the average Annual Rates of Change (ARC) of access to basic water services for rural populations (ARC_r) across all countries?

Transforming the data MCQs

05. How many countries' national populations had a 0% Annual Rates of Change, excluding countries that have 100% access, across the time period?

06. Considering that a negative Annual Rates of Change (ARC) indicates a decrease in access to water from one year to the next, and a zero ARC indicates no change, which of the following statements is most true based on the data?

- Although access to basic water services on a national level remained unchanged for more countries than in rural and urban areas, rural areas had the greatest decrease in access.
- Although access to basic water services increased for more countries in rural areas than on a national level, more countries had full access to basic water services in urban areas.
- Most countries had a similar change in access to basic water services across all types of population areas.
- Although access to basic water services on a national level increased for more countries, more countries had a decrease in access in urban than rural areas.

07. Which two countries had the highest absolute difference between urban and rural Annual Rates of Change?

Transforming the data MCQs

08. Which of the following statements is most true about the distribution of the difference in ARC values between rural and urban areas?

- More countries had higher Annual Rates of Change in urban areas than in rural areas since a greater number of difference values falls to the one side of the number line.
- The Annual Rates of Change were similar across urban and rural areas for most countries since the peak is close to the middle of the number line.
- More countries had higher Annual Rates of Change in rural areas than in urban areas since a greater number of difference values falls to the one side of the number line.
- We cannot say anything about the distribution because some of the histogram bins are empty.

09. On average, which region saw the greatest improvement in access to basic water services on a national level (considering the Annual Rates of Change) over the dataset time period?

Understanding the data MCQs

10. Based on the visualisation investigating the relationship between the rural and national Annual Rates of Change, national population size, and region, which of the following statements are true?

- Only countries in the Sub-Saharan Africa region observed a decrease in basic water access on a national level and in rural areas.
- The average population size per country in Sub-Saharan Africa is smaller than that of most other regions.
- Countries in the Sub-Saharan Africa region observed a greater spread in rural and national ARC values than other regions.
- Countries with larger populations generally observed national ARC values between 0% and 1%.