Transforming the Data

In this second part of the two-part project, we will investigate access to safe and affordable drinking water. We will transform the data.

In this project we will use the same dataset to the one we used for understanding the data. The previous data had n 'year' column. This column consists of years from 2000-2020 showing the access levels of various countries over the years.

-- Becoming familiar with the dataset.

we import the data to see how different it differs from what we used in the previous part of the project. We can see that the dataset is now without the income group column, and replaced with the year column.

-- Investigating year representation.

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

\* During which years were this data recorded?

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

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

we need to sort name, and year columns before we proceed.

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. We can determine that data were collected twice per country by checking the number of times the country appeared in the data, and how many distinct years data was recorded.

we will create a new column called year\_diff - year difference. We will 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.

Our condition will be the country name. If the name in the next row is strictly the same as the country name in the next row, it fulfills the condition. When the condition is True, we will calculate the year difference, but when the condition is False we will return nothing.

Because we have set our if statement to return an empty string, when the names are not the same, we can assume that if we find year\_diff = 0, we have duplicate data, because the subtraction of the same year should be 0, so we can infer that year\_diff = 0 equates to duplicates.

1. Range of the Year Difference

The range is the difference between the maximum and minimum values:

Range=MAX−MIN=5−1=4

The average (2.40) lies within this range and gives us a measure of central tendency, representing the typical year difference in your dataset.

2. Average Compared to Minimum

The average year difference (2.40) is significantly larger than the minimum (1). This suggests that the minimum is on the lower end of the distribution, and there may be relatively few instances where the year difference is as low as 1.

3. Average Compared to Maximum

The average year difference (2.40) is less than half of the maximum (5). This indicates that the maximum year difference is a potential outlier or a value that is not typical of the majority of the data.

4. Implications of the Comparison

Since the average is much closer to the minimum (1) than the maximum (5), it suggests that most data points are clustered around smaller year differences, with the larger differences (closer to 5 years) being less frequent or skewing the distribution slightly.

In summary, the average is more representative of the typical year difference than the extreme values (min or max). The max of 5 likely represents a few exceptional cases, while the average shows that most differences are around 2 to 3 years.

-- Investigating the Annual Rates of Change (ARC).

We want to see if access to services is improving or declining across national, urban, and rural areas. The UN uses ARC to see whether the proportion of access to drinking water is declining or improving.

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

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

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:

We create three new columns namely ARC\_n, ARC\_r, ARC\_u in the dataset sheet. These columns represent the annual rates of change for the country's national, rural, and urban populations.

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

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 0. 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 does not necessarily indicate less progress in changing access to basic water services because it takes it into account countries that already have 100% access.

-- Investigating access by area.

In our previous investigation of access by area, we observed that rural populations on average have lower access to basic water services thab the national or urban populations. Now we want to investigate whether countries have made significant efforts to improve this in the years leading up to 2020.

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

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

a. We want to calculate the number of countries per area that have full access and ARC equal to zer0, smaller than zero, and greater than zero.

In the summary sheet, we will find the number of countries that have null ARC values.

We will proceed to calculate the number of countries that have full access accross both years, i,e the number of countries where the access is 100% for both years reported. we have entries that are greater than 100% which is not possible. We will create three new columns called wat\_bas\_n (rounded) wat\_bas\_r (rounded) wat\_bas\_u (rounded) which will contain rounded values of original columns.

Wewill create a new column called ARC\_n\_full in the working sheet that determines IF the country names are the same AND that both wat\_bas\_n (rounded) columns for that country is greater than 99%. Rreturn "full access" if true.

We will create same columns for ARC\_r\_FULL, and ARC\_u\_FULL.

In the summary sheet we will find the number of countries that have full access per population, i.e the number of "full access" occurences in each way of the newly created columns ARC\_n\_FULL, ARC\_r\_FULL, ARC\_u\_FULL.

We will find the number of countries that have their ARC values equal to zero that doesn't already have full access for each population types, national, rural, and urban.

Calculate same for countries that had ARC < 0 and doesnt't already have full access.

And the number of countries where ARC > 0 and doesn't have full access.

We can check that we've considered all countries for the confitions: NULLS, full access, ARC = 0, ARC < 0, ARC > 0.

b. We will calculate the difference between the Annual Rates of Change between rural and urban populations per country. We will create a new column called ARD\_diff in the working 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.

-- Investigating access by region.

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

\* How does ARC compare across different regions?

\* What us the influence of national population size on the ARC?

a. Our original dataset did,t include the region information, so we will have to amend our dataset to investigate access by region.

We will import regions.csv into a new sheet called Regions.

We will add a new column in the working sheet called region and use any LOOKUP function to add the region based on the country name.

In the summary sheet, we will find

- the number of countries per region

- The average ARC on a national level per region.

- The average ARC in rural areas per region

- The average ARC in urban areas per region.

We will visualize access by region to investigate the relatoinship between the national and rural ARCs, aswell as population size and region.

We have analyzed the data, calculated many different statistical measures and visualized the data.

Sub-Saharan Africa, home to the largest population in the world with limited access 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 Ar=frica's water crisis.

Summary.

Originally we had 16 columns. After analysis we have 28 columns including newly created features.

We investigated the year difference by creating year\_diff column.

We investigated the Annual Rates of Change for national (ARC\_n), rural(ARC\_r), and urban (ARC\_u)

We investigated access by area. we created new columns wat\_bas\_n (rounded), wat\_bas\_r (rounded), wat\_bas\_u (rounded), ARC\_n\_FULL, ARC\_r\_FULL, ARC\_u\_FULL, ARC\_diff

We also investigated by region, we added a new column region.

In the summary sheet we have a summary of the dataset year and difference, including median, minimum, maximum.

We have a histogram of the year column.

We have the median, minimum, maximum of each of the three newly created ARC columns based on the rural, urban, and national ARC columns.

The number of countries that has no ARC value, full access, ARC<0, ARC>0 for each of the three newly created ARC columns.

A histogram of the difference in ARC values for rural versus urban areas.

A summary per region which at least includes the number of countries per region, the average ARC on a national level, the average ARC values in a rural and urban areas.

A visualization that represents the national ARC, rural ARC, region, and population size.

The dataset provided didn't represents the entire year range from 2000 to 2020 for at least some countries.

The average year difference across all countries for the dataset is 4.80 years.

Observation of the distribution of the year column shows that the distribution is neither normal, negatively, or positively skewed since there are two distinct peaks.

The avergae Annual Rates of Change of access to basic water services for rual populations across all countries is 0.48%.

16 countries' national population had a 0% ARC, excluding countries that have 100% access.

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.

South Sudan and Morocco had the highest absolute difference between urban and rural Annual Rates of Change.

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.

On average Sub-Saharan region saw the greatest improvement in access to basix water services on a national level (considering the Annual Rates of Change) over the dataset time period.

The average population size per country in Sub-Saharan Africa is smaller than most of other regions. Countries in the Sub-Saharan Africa region observed a greater spread in rural and urban values than other regions.

Countries with larger populations generally observed national ARC values between 0% and 1%