Investigating Access to Drinking Water

Project Summary/Overview

A two-part integrated project on investigating access to safe and affordable drinking water (SDG 6: Clean water & Sanitation) using google sheets. The first part of the project covers understanding the features in our dataset and the second involves transforming the features in our data.

Tools Used

- Google Sheets/Microsoft Excel

Datasets (CSV files)

- **Project 1**
  WHO/UNICEF JMP (Joint Monitoring Programme for Water Supply, Sanitation, and Hygiene) Estimates on the use of water dataset for 2020.
    - *This dataset has a total of 16 features (or columns), 12 of which are service-level percentage shares.*

- **Project 2**
  WHO/UNICEF JMP (Joint Monitoring Programme for Water Supply, Sanitation, and Hygiene) Estimates on the use of water dataset for 2000 - 2020.
    - *This dataset has a total of 16 features (or columns), 12 of which are service-level percentage shares.*

NB: *Refer to data dictionary for more in-depth descriptions of what each feature or column represents*

Project 1: Understanding the Data

The first part of the project covers understanding the features in our dataset. This involved:

- Importing and Data cleaning
    - Importing the CSV file into google sheets and ensuring that all column in the data are properly delimited(seperated by a comma) using the ***split text to column*** option under the **Data tab**
    - Creation of a new column "**value_count"** to count the number of features in a row, ensuring each row had 16 features.

- Investigating population size
    - Creation of a new worksheet "**Global 2020 report** ". This sheet contains:
        - An aggregation for the total national population size (pop_n) and the value for the estimated world population.
        - Values for the urban percentage share. This involves creating a new feature "**pop_u_val**" and the percentage difference between all of the features.

Pop_u_val *is the number of people living in urban areas per each country. It is calculated from our features pop_n and pop_u (%).*

- A line chart of the national population versus the urban and rural population shares. Involved the creation of two new features in our dataset ("**pop_r"** , "**pop_n(m)**".

Pop_r *is the percentage share of the population living in rural areas, calculated using the pop_u and pop_n features.*

pop_n(m) *is the national population size rounded up to the nearest million.*

- 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 in our "**Global 2020 report** ".
  - A box and whisker plot for all 12 water access features.

- 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 access to water (Basic, Limited, Unimproved, and Surface). This involved creation of two new features ("**pop_u (rounded)**", "**pop_r (rounded)**") to make our visualisation neater.

pop_u (rounded) *and* ***pop_r (rounded)*** *are the urban population share (pop_u) and the rural population share (pop_r) rounded to the nearest whole number.*

- Investigating access by income group
  - Creation of 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 the income group versus the different average shares in the created pivot table.

*All visualisation are stored in a seperate worksheet "****visualisations"***

This part of the project enabled us answers questions such as:

1. What the national average percentage of access to limited services (wat_lim_n) for low-income countries is
2. if the share of the national population with basic water access increases as urbanisation increases and if limited, unimproved, and surface water access to water decreases as GNI increases.

Project 2: Transforming the Data

- Importing and Data cleaning
    - Importing the CSV file into google sheets and ensuring that all column in the data are properly delimited(seperated by a comma) using the split text to column option under the Data tab
    - Creation of a new column “value_count” to count the number of features in a row, ensuring each row had 16 features.

  *The “**income_group”** feature in this dataset has been removed and replaced with a “**year”** feature that represents the years over which our data was collected.*

- Investigating Year Representation
    - To observe which years are represented for which countries, we sort entire dataset by just the **“name**” and “**year**” features (use Data > Sort range > Advanced range sort options)
    - Calculate the average difference in years for data entries per country in a new column “**y_diff**’ Using if statements to only subtract two years if the country in the **name** feature is the same for the new feature y_diff else return an empty string.

  A y_diff value = 0 indicates the presence of duplicate rows in our dataset which need to be removed.
    - Creation of a new worksheet “**summary**” in which we:
        - Calculated the average, minimum, and maximum year difference (y_diff)
        - Created a histogram of the year column (used the minimum and maximum values on the horizontal axis).

- Investigating ARC
    - Create three new columns “**ARC_n”**, “**ARC_r”**, and “**ARC_u”** to represent the annual rate of change for the country’s national (n), rural (r), and urban (u) populations.

        - ARC_n (national) is calculated using wat_bas_n, ARC_r (rural) using wat_bas_r and ARC_u (urban) using wat_bas_u. This helps determine whether the proportion of access to drinking water is declining or increasing.

        - Use an iferror function to replace #VALUE! errors with the string “null”

  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 a variable at different points in time and dividing the result by the number of years that separate the two values

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

- Investigating Access by area

- Created three new features in our original dataset sheet "**wat_bas_n (rounded)"**, **"wat_bas_r (rounded)**", and "**wat_bas_u (rounded)**". Which are the original access to basic water services columns (wat_bas_n, wat_bas_r, wat_bas_u) rounded to zero decimal places.
- Created additonal three new features "**ARC_n_full, "ARC_n_full", "ARC_n_full"** in our original data sheet to indicate whether a country has full access for its national, rural and urban populations.
    - This was done by using IF statements and AND logic to return "full access". For example for **ARC_n_Full**, IF the country names were the same AND both wat_bas_n (rounded) features for that country are > 99% for both years we output "full access". (Similar calculations were done for **ARC_u_Full and ARC_r_Full)**

- In our summary sheets we calculated the number of countries that have no ARC value, full access, ARC < 0, and ARC > 0 for each of the three newly created ARC columns.
- Plotted a histogram of the difference in ARC values for rural versus urban areas.

- Created a new feature "ARC_diff" in the dataset to calculate the difference between the rural ARC (ARC_r) and urban ARC (ARC_u).
    - Visualised the newly created ARC_diff feature in a histogram.

- Investigsating access by region
    - Imported the Regions.csv into a new worksheet.
    - Created a new feature "**region**" in our original dataset and used a LOOKUP function to add the region based on the country name.
    - Calculated the number of countries per region, the average ARC on a national level, and the average ARC values in rural and urban areas in the summary sheets.
    - Created a visualisation that showcased the national ARC against the rural ARC, as well as the region and national population size.

This part of the project enabled us answers questions such as:

1. What the average year difference across all countries in our dataset is.
2. What the average annual rates of change (ARC) of access to basic water services for rural and urban populations across all countries is.
3. Identify which region(s) saw the greatest improvement in access to basic water services on a national level between 2000-2020.