Estimates on the Use of Water Project

## Part 1: Understanding Data.

This project is based on the WHO/UNICEF JMP (Joint Monitoring Programme for Water Supply, Sanitation, and Hygiene) estimates on the use of water dataset for 2020.

In this project we will analyze to understand the data in the dataset provided. We will use this understanding to uncover insights about the access of water across the countries in the world. We will understand which countries made progress in national, rural, and urban populations.

## **Familiarizing with the data.**

This is the most important part of any analysis project. You have to become familiar with the dataset. This is the first step that aims you to understand the data, and allows you to know the relationships between various columns to be able to make informed decisions.

The Dataset is estimates on the use of water (2020).

There are a total of 213 rows.

* name - name of the country or area name
* income\_group - The country's clarification 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 at least basic service (%).
* wat\_sur\_n - The estimated national share of people with surface service (%).
* 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 improved service (%).
* wat\_sur\_u - The estimated urban share of people with surface service (%).

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

### **Importing Data**

The data is in csv format. The table headers were separated by semi-colons. During importing certain (5) rows were separated by semi-colons therefore did not split to a different column. Using Text-to-split function to split the columns.

### **Investigating Population Size**

We want to summarize the national population size to better understand how the dataset represents the entire world population. In 2020 the estimated world population was estimated to be 7.821 billion, 55% of which lived in urban areas.

In this part we will find:

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

We will create a summary sheet called Global 2020 Report to compare the dataset population to the estimated world population.

In this new sheet we determine the total national population size using the column pop\_n. The pop\_n column is represented in thousands, so we divide the column by 1000000 to have a figure in the same format (Billions).

We also want to compare the world’s urban population to the dataset’s urban population. We need the total number of people living in urban areas from our dataset, which we can compare as a number or a percentage to the world urban population value.

We create a new column in the working sheet called pop\_u\_val, which is the number of living in urban areas per country. We will use pop\_n, and pop\_u (represented in percentage) to determine this new column.

We know that the estimated world urban population in 2020 was 55% of the total population, which is 4.3 billion people living in urban areas.

We find the urban share of the dataset using the pop\_n and pop\_u columns. We can see that we have relatively same values for the dataset (4.283), and world (4.302)

We will 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 (44%).

We will also find the percentage difference of the urban population totals and percentages in the dataset (44%).

*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.*

*Percentage difference refers to the difference between the relative magnitude of two values, expressed as a percentage of the average of those values.*

* How the urban population share compares to the rural population?

We will create a visualization to compare the share of the national population living in urban versus rural areas. We will create a line chart with national population ***(pop\_n)*** as the *independent variable - cause*. Our *dependent variable - effects* - on the y-axis will be the rural, and urban populations.

*We find that our data is not very comprehensible, due to some of our pop\_n values being much larger than our other values, outliers. If we change the chart to a scatter plot you can clearly see the outliers.*

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

* **Delete the outliers from our dataset.**

1. Advantage: We donʼt have to change anything in our data visualization as it would update automatically.
2. Disadvantage: We lose information about some of the countries included in the dataset in this visualization and any future analysis we might need to do.

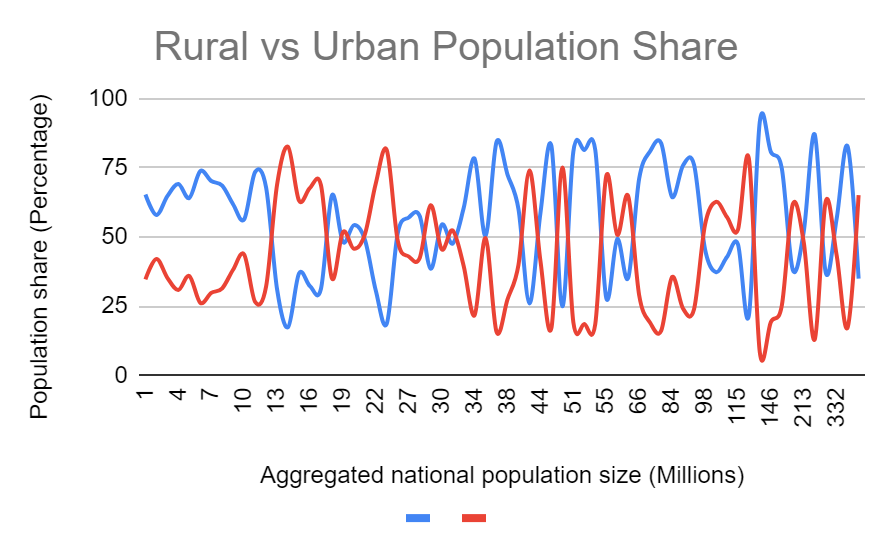
* **We edit our visualization by adding a maximum cut-off value to our x-axis.**

1. Advantage: Our visualization is more comprehensible and we don't lose information in our dataset for any future analysis we might need to do.
2. Disadvantage: Our visualization doesn't represent the entire dataset, only a subset of it.

* **We change the unit of our x-axis.**

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

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.



### **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, rural, and urban***).

We will use measures of central tendency and spread in this section.

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

In the Global 2020 Report sheet, we will find the maximum of each of the four national water access columns, **wat\_bas\_n, wat\_lim\_n, wat\_unimp\_n, wat\_sur\_n.**

We found that the maximum for national access to basic water service (wat\_bas\_n) exceeds 100%, which is not possible. Because the highest is 100% and we cannot have past that. We will fix that by adding a new column **(wat\_bas\_n rounded)**, and we will round the values that exceeds 100% and then report the maximum value on the report sheet.

For a normal distribution, we expect that the mean, median, and mode would be equal. We will find the three features for each of the four services.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | **Maximum** | **Minimum** | **Mean** | **Median** | **Mode** |
| national | Basic | 100 | 37.20240205 | 89.86478244 | 97.3481397 | 100 |
| Limited | 37.42696287 | 0 | 3.865265345 | 0.4707829181 | 0 |
| Unimproved | 33.53911377 | 0 | 4.418010654 | 0.8536474704 | 0 |
| Surface | 30.36979308 | 0 | 1.915488584 | 0 | 0 |
| rural | Basic | 100 | 21.98279234 | 81.33591803 | 90.73161683 | 100 |
| Limited | 42.16438068 | 0 | 5.836562548 | 1.812330835 | 0 |
| Unimproved | 51.21598167 | 0 | 8.732261484 | 3.229749713 | 0 |
| Surface | 40.51813242 | 0 | 4.224039638 | 0.2196897375 | 0 |
| urban | Basic | 100 | 49.66166495 | 94.68983975 | 98.10662849 | 100 |
| Limited | 34.27978009 | 0 | 3.281838654 | 0.5 | 0 |
| Unimproved | 19.83580681 | 0 | 1.71878521 | 0.3498740029 | 0 |
| Surface | 6.362160532 | 0 | 0.3131148443 | 0 | 0 |

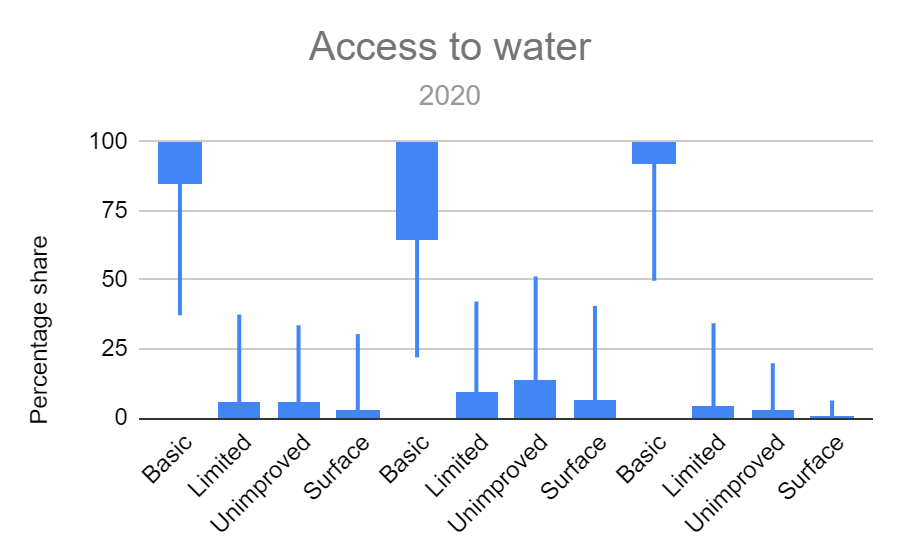
We see that the mean, median, and mode are not equal for any of the four columns, we will also find the Interquartile range and standard deviation for these four columns to gain some more insights into how the data are distributed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **QR1** | **QR3** | **IQR** | **SD** |
| national | Basic | 85.64331096 | 99.88752524 | 14.24421428 | 15.08742684 |
| Limited | 0 | 4.851219157 | 4.851219157 | 6.960846633 |
| Unimproved | 0.03390622265 | 5.279309922 | 5.245403699 | 7.174477036 |
| Surface | 0 | 1.878301993 | 1.878301993 | 3.923973987 |
| rural | Basic | 64.8254211 | 99.11924249 | 34.29382139 | 21.50663741 |
| Limited | 0 | 8.553175126 | 8.553175126 | 8.704765846 |
| Unimproved | 0.1584109169 | 13.2663393 | 13.10792838 | 35.1380243 |
| Surface | 0 | 6.157471481 | 6.157471481 | 11.82849434 |
| urban | Basic | 92.56287752 | 99.94941803 | 7.38654051 | 8.03886161 |
| Limited | 0 | 3.966347184 | 3.966347184 | 5.63506705 |
| Unimproved | 0 | 2.059001302 | 2.059001302 | 3.23457032 |
| Surface | 0 | 0.1597056849 | 0.1597056849 | 0.8696316426 |

For wat\_bas\_n **(Basic)**, we see that the mean is relatively high and the interquartile range is about 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 do these measures of the water access compare across different types of areas?

We will visualize the five-number summary of the four access features across the three different types of areas. We will create a box and whisker diagram in the report sheet from our measure of central tendency and spread summary for all 12 columns.

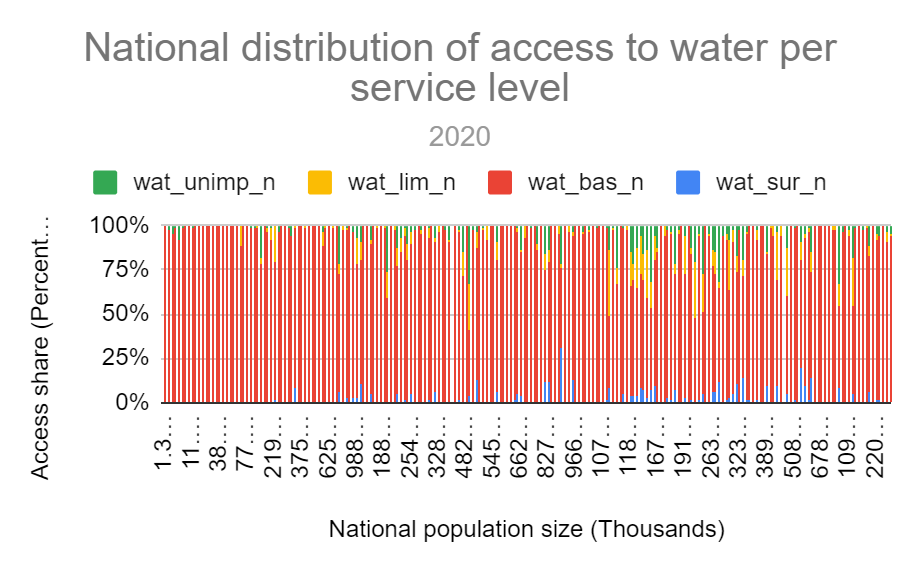


### **Investigating Access by Population Size.**

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

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

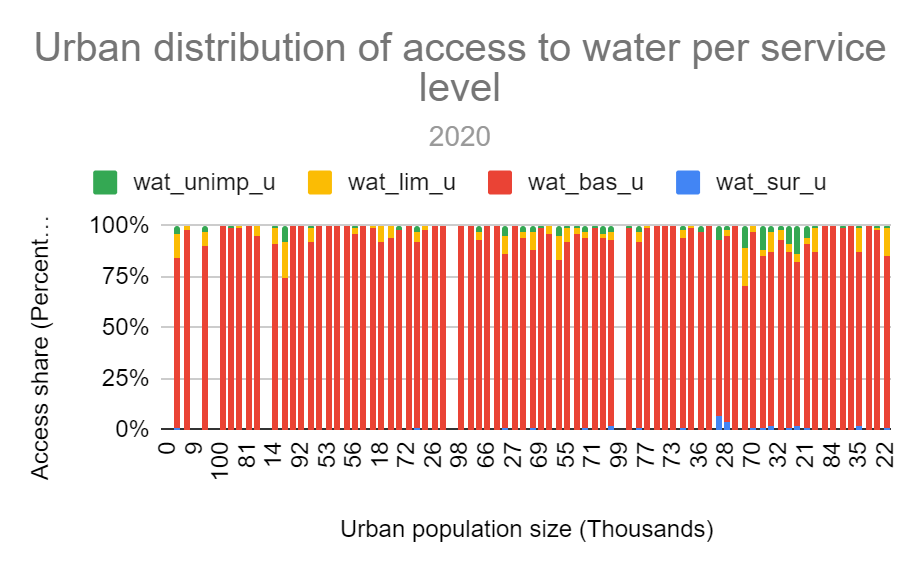
We will visualize the national access to water on all four levels based on the national population size.

we want to investigate how population size affects access levels, population size **(pop\_n)** will be our independent variable (cause), and our four access columns **(wat\_bas\_n, wat\_lim\_n, wat\_unimp\_n, wat\_sur\_n)** will be the dependent variables (effects).

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

We will visualize the urban access to water on all four levels based on the urban population.

We will create a similar visualization like the previous one we did for national, but this time for urban population.



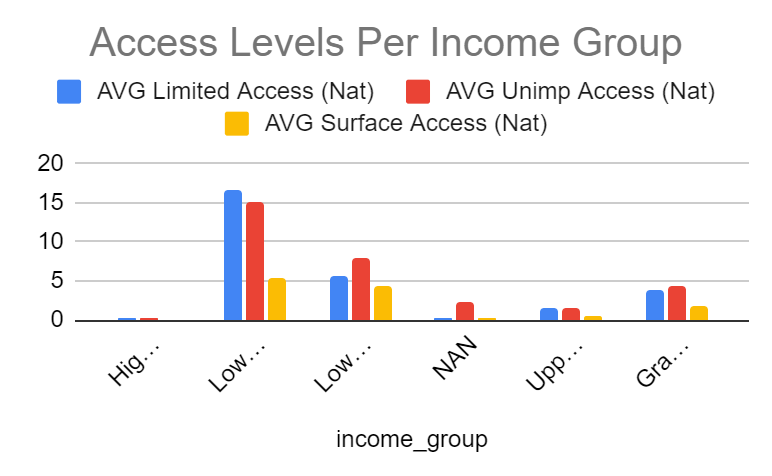
### **Investigating access by income group**

We want to investigate the relationship between Gross National Income or income group, population size, urbanization, and national water access.

Economies are classified into four income groups based on gross national income per capita.

* What is the effect of national population size and urbanization on GNI and water access?

In our report sheet we will create a pivot table. Because we want to group by income group column.



Summary

The percentage difference between the dataset and the estimated world urban population size (the total number of people living in urban areas) for 2020 is 1.70%. The total urban population size (the total number of people living in urban areas, NOT the percentages) for the dataset needs to determine by first calculating the number of people in urban areas per country based on the population size per country (pop\_n) and the urban share (pop\_u), and then the sum of these urban totals for the entire dataset.

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).

Considering the measures of central tendency and spread, we can see that the maximum for both wat\_bas\_n and wat\_bas\_u is 100%. We also see that the minimums, means, medians, first and third quartiles as well as the interquartile range are more similar between wat\_bas\_n and wat\_bas\_u than wat\_lim\_n. Considering the box and whisker plot, we clearly see that wat\_bas\_n and wat\_bas\_u sit at the top of our plot, while wat\_lim\_n sits at the bottom. If we had to imagine what the distributions would look like based on the measures and chart, we know that wat\_bas\_n and wat\_bas\_u will exhibit negatively skewed distributions because the outliers occur to the left of the number line, while wat\_lim\_n exhibits a positively skewed distribution because outliers occur to the right of the number line.

Countries with greater urban population shares are more likely to provide basic water service than countries with smaller urban population shares. When the sort is correctly applied to the variable used on the x-axis (pop\_u (rounded)), we see that toward the right of the number line, limited, unimproved, and surface access decreases, and basic increases. When the sort is applied to the rural population share, we see the opposite, with basic at a maximum on the left of the number line (but still at the highest values) because the number line is in descending order. When the dataset is sorted on the national population size, there is no order to the number line and seemingly no relationship between access and urbanisation.

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). When the sort is correctly applied to the variable used on the x-axis (pop\_r (rounded)), we see ~100% basic access at a rural population share of 2%, 5 – 8%, 12%, 16%, 20%, 28%, 40%, 46 – 47%, 49%, 59%, and 100%, although there seems to be some trend where, on average, basic access decreases with rural population share increase.

Based on the created pivot table, the national average percentage of access to limited services (wat\_lim\_n) for low-income countries is 16.55%

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.

Basic access in high-income countries; Limited access in low-income countries; Unimproved access in low-income countries; Surface access in low-income countries

Visualizing the pivot table values for national access versus income group indicates that as urbanization increases, so does the share of the population with basic water access, and as GNI increases, limited, unimproved, and surface water access decreases.

There is a clear relationship between urbanization (average of urban share) and water access shares. As urbanization increases, so does the average of basic access (average of basic share) increase and limited, unimproved, and surface access decrease. Also, as GNI increases (or income group increases from low to high), basic share increases, and limited, unimproved, and surface shares decrease.

We had 16 original columns from the beginning. We created new column to total 23 columns at the end of the analysis.

* We investigated population size and added a new column, **value\_cnt**
* We investigated access by area and added new columns **pop\_u, pop\_r, pop\_n (m)**
* We investigated access by population size and added new columns **wat\_bas\_n (rounded)**
* We investigated access by income group and added new column **pop\_u (rounded), pop\_r (rounded).**

In the summary sheet.

* A summary of the dataset population size and estimated world population, which includes urban percentage difference between all of the features.
* A line chart of the national population versus the urban and rural population shares.
* The maximum, minimum, mean, mide, 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.
* Three 100% stacked column charts, one each for national, rural, and urban population size or percentage versus the four service levels.
* A pivot table for income group versus the sum of the national and the averages for the urban population, basic, limited, unimproved, and surface shares.
* A visualization for income group versus the different average shares in the created pivot table.