Water consumption worldwide – comparison of average per capita water usage across different countries.

Name: Batyrali Bazar

Date: 5/23/2025

1. Introduction to data

1) Description:

Data is taken from Kagle website, it refers to 'Water Use Statistics - Worldometer'. Additionally, some data from the World Bank website.

Variables accessible:

Country	Qualitative
Yearly Water Used (billions of liters)	Quantitative
Daily Water Used Per Capita (liters)	Quantitative
Population	Quantitative
GDP per Capita($)	Quantitative
GDP Current($)	Quantitative
Urban Population(%)	Quantitative
Industry Water Withdrawals(%)	Quantitative
Water Usage Efficiency($/Liter/Day)	Quantitative
Industrial Economic Intensity($)	Quantitative

There is no time series data, only cross-sectional. Because data from one point in time.

2) Analysis goals:

Purpose of the project:
To analyze if available data regarding water can show us meaningful insights, to understand if it is suitable for precise analysis.

Research questions:
Which factor affects water the most? Additionally:

- Is there any factors?
- Which factors affect water consumption per capita?(main question of the topic)

Can we calculate water usage efficiency?
Does higher water consumption level lead to better usage efficiency?

Which countries use water the most?
In average, how much water do we use worldwide?
Is there a difference in water usage between different groups of countries?

Hypotheses:
There is no statistically significant difference in variances among the groups.
There is no difference in daily water usage in different groups by GDP.
There is no difference in water usage efficiency between different levels of urbanization.

3) Data limitations:

I must mention that the main data set used for analysis contains data from various ranges of years (around 1996-2024). It adds limitation to data reliability and comparability. However, since there is no other options available and in accordance with project guidance, analysis was made on available dataset. Also agricultural and climate data were not used, mainly because in terms of country they are too much averaged, and didn't show anything at all.

2. Data Preparation and Cleaning

1) Packages used:

General	**pandas, numpy**
For linear regression	sklearn
For statistical tests	scipy
For visualization	matplotlib, seaborn, plotly.express

2) Loading and manipulating the main data set:

Main data set was loaded form .csv file and was examined it to understand structure and content. It covers water consumption and social indicators worldwide.

Two columns(total water usage, water usage per capita) data types were changed from object(string) to float. Additionally, 'Population' column data type was changed from object(string) to int.

In the column 'Country' entry 'Russia' was changed to 'Russian Federation' for easier data sets merge later.

Column 'Yearly Water Used' data was converted from thousands of liters to billions of liters to improve readability and analysis.

3) Adding additional data as a columns:

4 additional data sets were loaded and used to extract necessary columns. Additionally renamed for proper merger. Made an inner merge, key = Country.

Additional data loaded:

GDP per capita

GDP total

Urbanization %

Water used for industry

4) Data cleaning:

Data was cleaned after uploading all the necessary data sets, was decided to drop null and NA values. Cleaning left 153 non-null values to use.

5) Creating new columns from existing data:

Two new columns have been created:

'Water Usage Efficiency($/Liter/Day)' – by dividing GPD per capita by daily water use. Obviously, shows water usage efficiency.

'Industrial Economic Intensity($)' – by multiplying freshwater withdrawals for industry and GDP per capita. It stands for $ produced per unit of industrial water withdrawals.

6) Setting up data for ANOVA test:

Countries were divided into 3 groups by percentage of population in urban areas.
This is a setup to use in a ANOVA test later.

3. Descriptive Statistics

Descriptive Statistics were provided for entire data set columns in notebook. I have decided to show the key points for each column in the report.

1) Yearly Water Used (billions of liters):

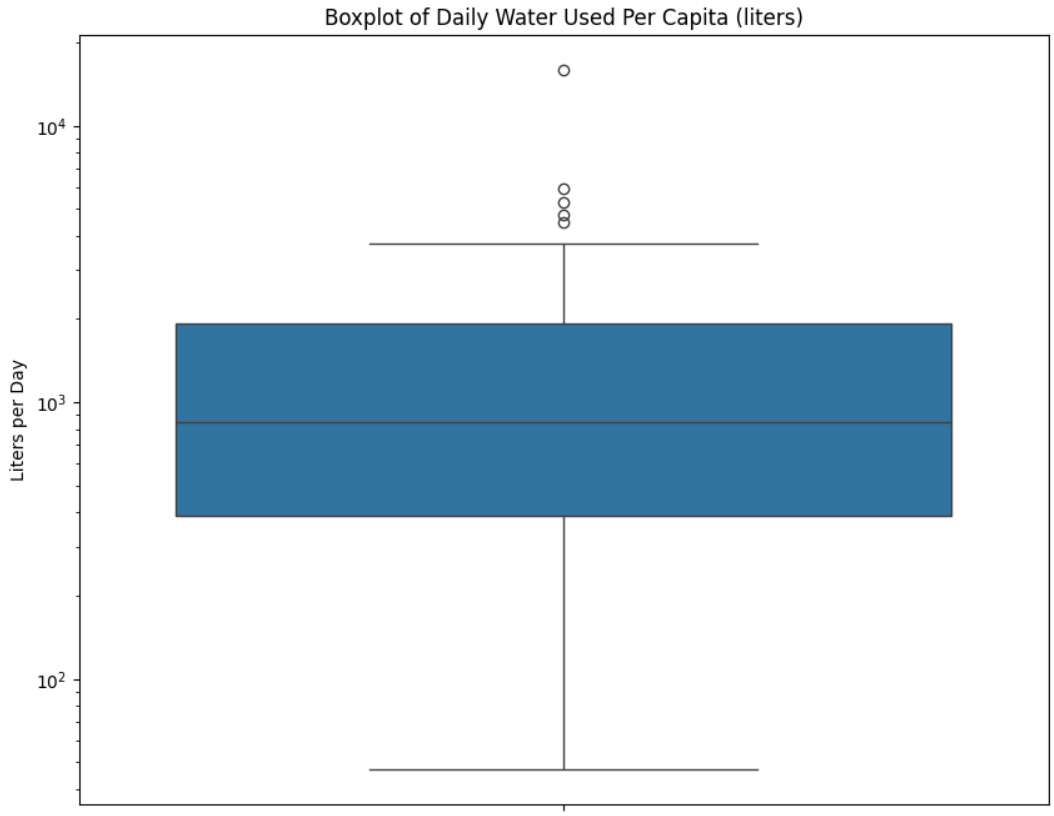
Mean ~ 23 400 billion liters.
Median ~ 2100 billion liters.
It has very high variability, standard deviation ~ 88 400 with a huge range ~ 760 000.

Overall need to mention that big consumers make average seem higher than it is for others.

2) Daily Water Used Per Capita (liters):

Mean ~ 1363 liters.
Median ~ 850 liters.
Again, high range ~15 800, suggesting significant differences between countries.

I have decided to show boxplot only for (main topic) Daily water usage per capita. It is generally skewed towards lower values, as indicated by the position of the median closer to Q1 than Q3. The majority of daily water usage per person falls between approximately 500 and 1000 liters. There are, however, several individuals or data points with significantly higher water consumption



3) Population:

Mean ~ 41 million.
Median ~ 8 million.
Range here stands for ~ 1.39 billion, it shows that I have both small and large countries in my data set.

4) GDP per Capita ($):

Mean ~ 18 500 $.
Median ~ 6800 $.
High range ~ 225 800 again shows inclusion of both high and low level groups.

It has a high variability, which may show that it affects water consumption. (we will check it later)

5) GDP Current ($):

Very large scale: mean ~ 620 billion$ and range ~ 26 trillion$ range. That has been mostly affected by countries like USA ad China.

6) Urban Population (%) and Industry Water Withdrawals (%):

They have moderate means and medians. Ranges (urbanization ~86%, industry water ~92%). Obviously, it is unrealistic to have 100% on ranges here, but still values are quite high.

7) Water Usage Efficiency ($/Liter/Day):

Mean ~ 28.7$.
Median ~ 9.1$.
Range ~ 604.
It shows that some countries have significantly higher efficiency, and from other side some significantly less.

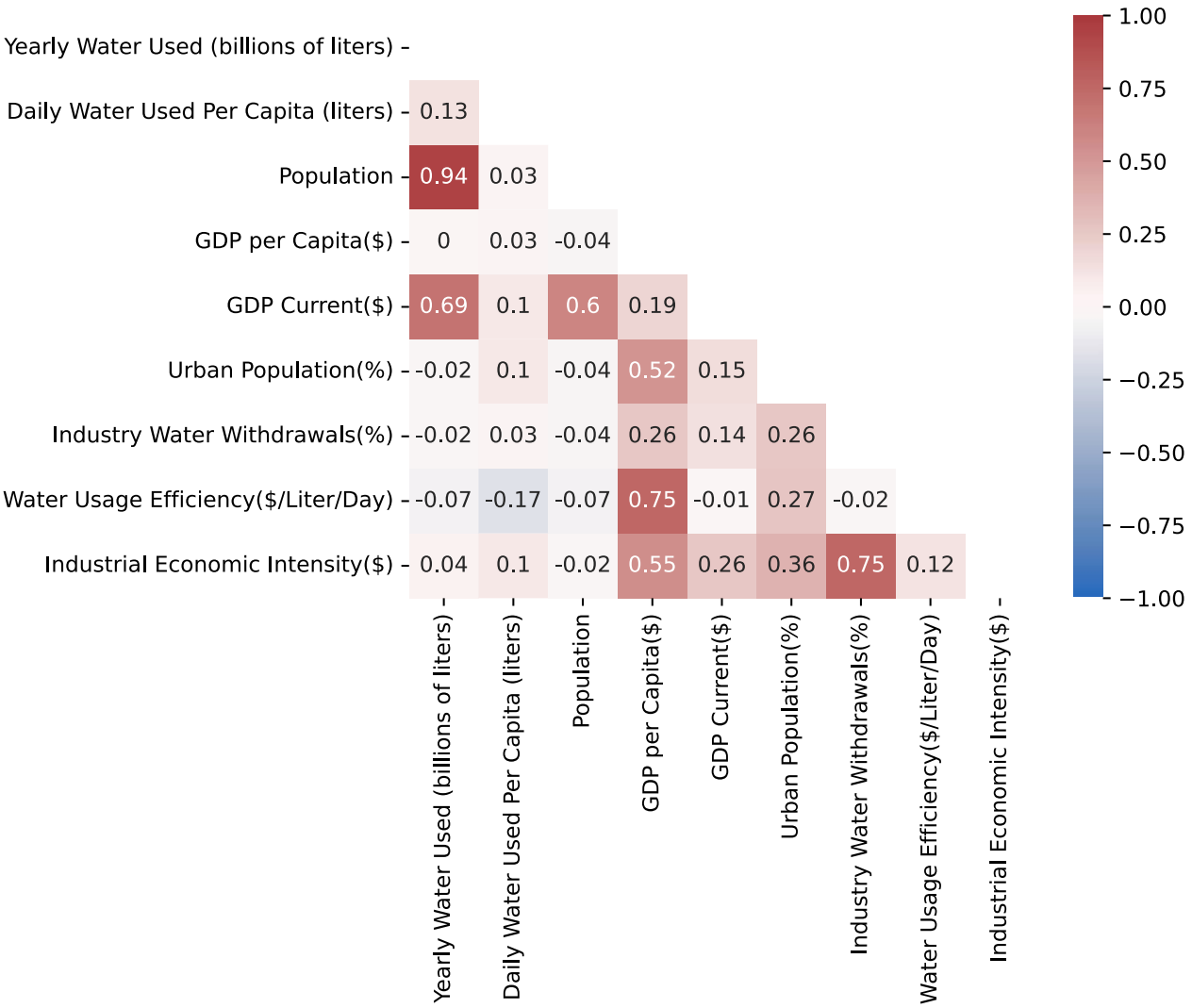
8) Industrial Economic Intensity ($):

Large variability (mean ~511 000$, std dev ~1 million$).
Affected by stronger industrialist countries.

9) Conclusion:

The descriptive statistical analysis showed that water consumption and economic indicators showed significant differences between countries. The yearly water consumption data showed major outliers because it spanned from 2 billion to more than 760,000 billion liters. The wide range of water use per person indicates different patterns of consumption among the population. The dataset includes countries with populations between 10 million and 1 billion people and GDP per capita ranging from $2000 to more than $200,000.

4. Correlation Analysis



Pearson Correlation matrix heatmap:

1) Objective:

To find factors that affect/correlate with water. For that two groups(with columns related to water) were considered.

2) 0.40 < r < 0.70 Moderate Correlation:

I. Between 'GDP per Capita' and 'Industrial Economic Intensity' (0.55): We can assume that high GDP per capita does not show significant impact on how effective is industry uses water withdrawals. In simple words if country is richer, it doesn't mean that it produce more $. Theoretically it also affected by counties like Ireland, which has artificially high GDP per capita.

II. Between 'GDP total(current$)' and 'Yearly water used(0.69): Countries with higher GDP more likely consume more water yearly. Value r here is

'nearly strong', so if between two, total GDP's impact on water is rather significant than not.

3) 0.70<r<0.90 Strong Correlation:

I. Between 'Industrial Economic Intensity' and 'Industry Water Withdrawals'(0.75): We can see that number of water pumped has higher impact on $ generated than GDP per capita. And summing up with 'Industrial Economic Intensity', first assumption is: availability of resource is more important than wealthiness, in case of industrial output.

II. Between 'GDP per capita' and 'Water Usage Efficiency'(0.75): Taking into consideration what I mentioned before, there are several countries with very high GDP but small population. Still, I assume that as countries become richer, they manage water resources better (due to infrastructure or technologies). Difference with 'Industrial Economic Intensity' is that it is limited by water used per capita.

4) r>=0.90 Very Strong Correlation:

Between 'Population' and 'Yearly water used'(0.94): One and only, expected very high correlation. Simple enough, more people = more water usage.

5) Conclusion:

As a conclusion, first of all, have to mention that I am surprised that pure water consumption is affected by so few list variables. Especially 'Daily Water Used per capita', literally 0 correlation with anything. Since main topic is per capita water usage, I have to conclude that water, like other natural phenomena, is more likely to be nearly unaffected by external influences (if we compare it in world scales of course). But still, we got some meaningful insight(population's affect).

5. Linear Regression

1) Objective:

Linear regression was used to analyze if it is possible to predict yearly water usage by population, water usage efficiency by GDP per capita, Daily water consumption per capita by GDP per capita. Main purpose of these linear regressions, is to check how predictable water is.

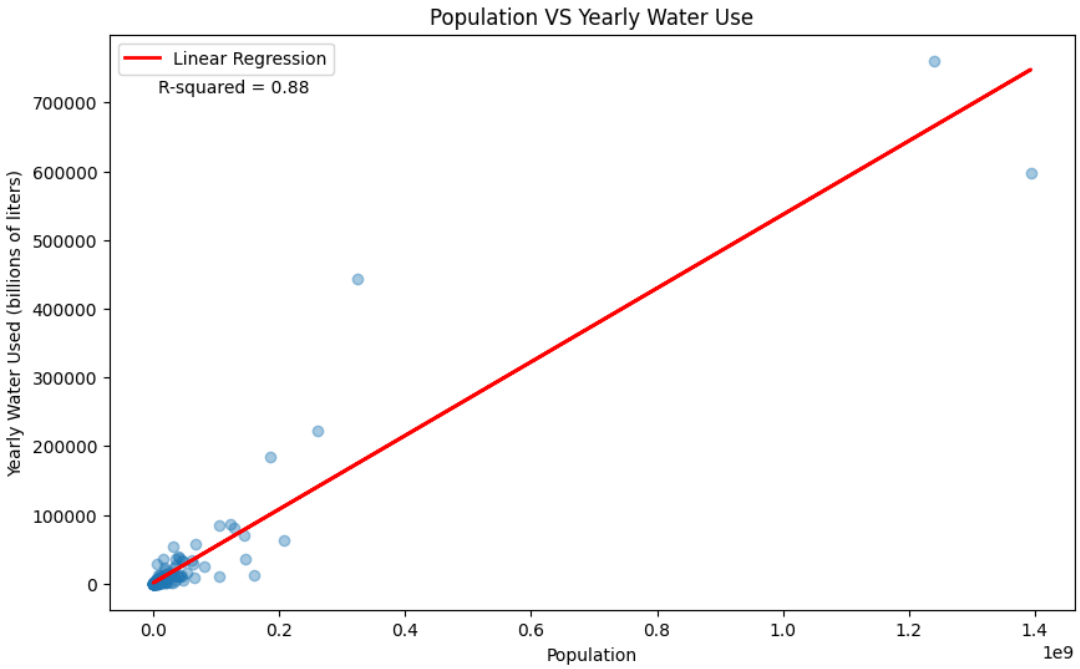
2) Variables Definition

Dependent variables	Independent Variables
'Yearly Water Used (billions of liters)'	'Population'
'Water Usage Efficiency($/Liter/Day)'	'GDP per Capita($)'
'Daily Water Used Per Capita (liters)'	'GDP per Capita($)'

This predictors were selected basically because of environmental science's logic and hypothesis.

3) Linear Regression Models:

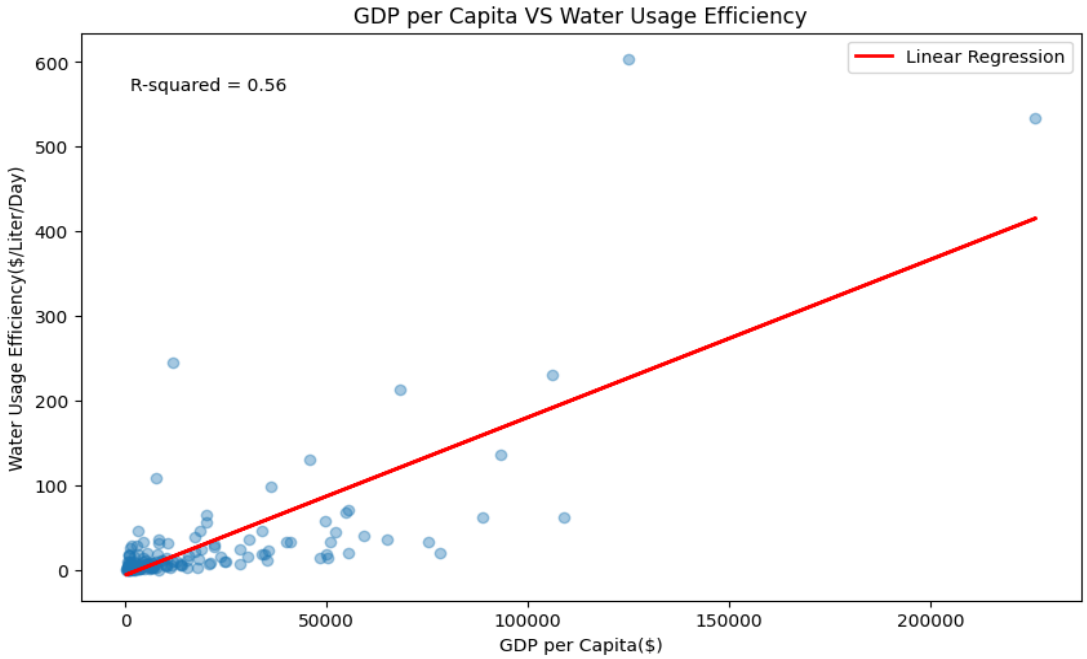
I. Prediction of 'Yearly Water Used (billions of liters)' by 'Population'



Intercept: [1020.96069246]
Coefficient: [0.00053597]

As you can see R-squared value here is 0.88, it means that 88% of the variability in Y is explained by the model. According to Ecology/Environmental science (r2>0.70) it is considered very strong connection. We can assume that it is possible to predict yearly water usage by population.

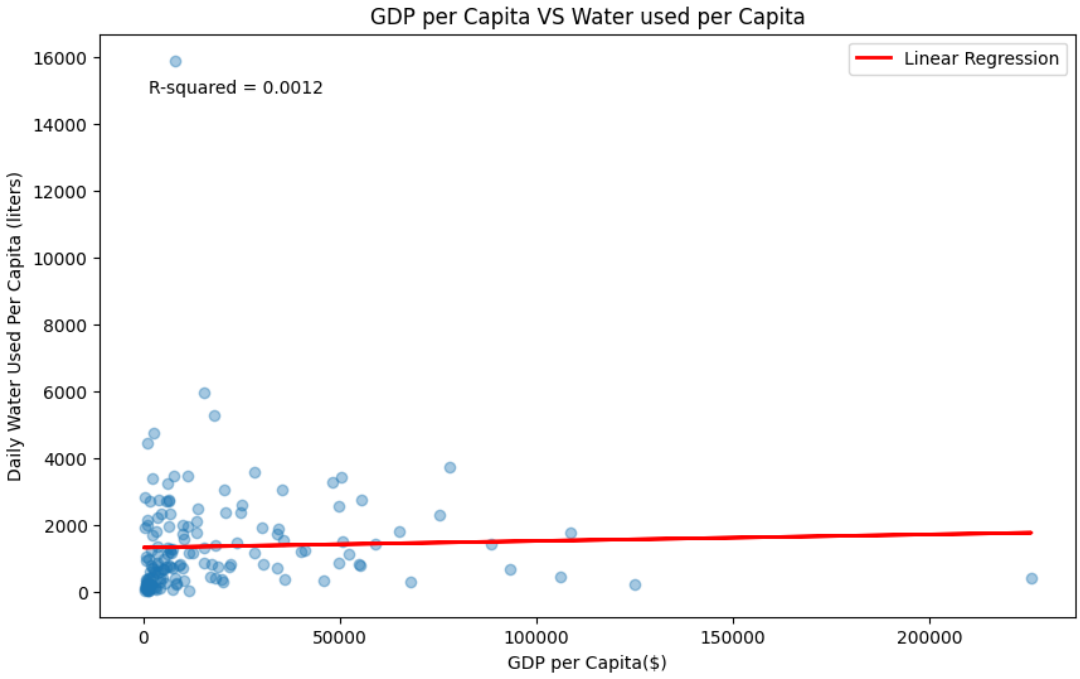
II. Prediction of 'Water Usage Efficiency($/Liter/Day)' by 'GDP per Capita($)'



Intercept: [-5.87308194]
Coefficient: [0.00186376]

R-squared = 0.56, 56% of variability Y explained. According to Ecology/Environmental science (r2>0.50) it is considered strong connection. According to this model, I can assume that higher GDP per capita is associated with higher water usage efficiency, even without R-squared being too high.

III. Prediction of 'Daily Water Used Per Capita (liters)' by 'GDP per Capita($)'



Intercept: [1326.89459221]
Coefficient: [0.0019546]

R – squared = 0.0012, which is extremely low. GDP per capita is very ineffective to predict daily water usage per capita. This example showcases most of the attempts to predict daily water usage. It is important to mention, I don't have any variable that will predict it well.

4) Conclusion

Linear regression models gave more confidence to say that yearly water usage is highly affected by population and that water usage efficiency is affected by GDP per capita (even if not too high).

Now we have already better picture and can assume that **there are factors that affect water overall**, **but none of them affect daily water usage per capita** - which is main topic.

6. Statistical Tests

1) Assumption checks: Normality, Homoscedasticity, Independence.

Normality test was not conducted, because sample size is 153, Central Limit theorem applies. Also, independence test was not conducted, because there are no time series data. To test homoscedasticity, Levene's test was used to prepare variables for ANOVA test:

Null Hypothesis: there is no statistically significant difference in variances among the groups
Alternative Hypothesis: there is statistically significant difference in variances among the groups

Results:
Levene's Test Statistic: 2.3017
Levene's Test P-value: 0.1036

Since **P-value: 0.1 > 0.05**, failed to reject null hypothesis. There is no statistically significant difference in variances among the groups prepared for Anova test.

2) Welch's T-Test

I have decided to check water consumption per capita between two groups, with GDP per capita lower than mean, and higher . For that, was decided to use Welch's T-test instead of Student's, mainly because in real world data it is safer to assume that variances are unequal.

> Null Hypothesis: there is no difference in daily water usage in different groups by GDP
> Alternative Hypothesis: there is significant difference in daily water usage in different groups by GDP

Results:

Welch's T-value: -1.06

P-value: 0.29

P-value: 0.29>0.05, fail to reject null hypothesis, no statistically significant difference in per capita water consumption between the two groups with different gpd per capita.

3) Anova test

Lastly, I have decided to check if urbanization level affects water usage efficiency.

> Null Hypothesis: there is no difference in water usage efficiency between different levels of urbanization
> Alternative Hypothesis: there is significant difference in water usage efficiency between different levels of urbanization

Results:

Anova f-value: 3.44

P-value: 0.03

We can see that **P-value: 0.03<0.05**, we reject null hypothesis, so there is statistically significant difference in water usage efficiency between different levels of urbanization.

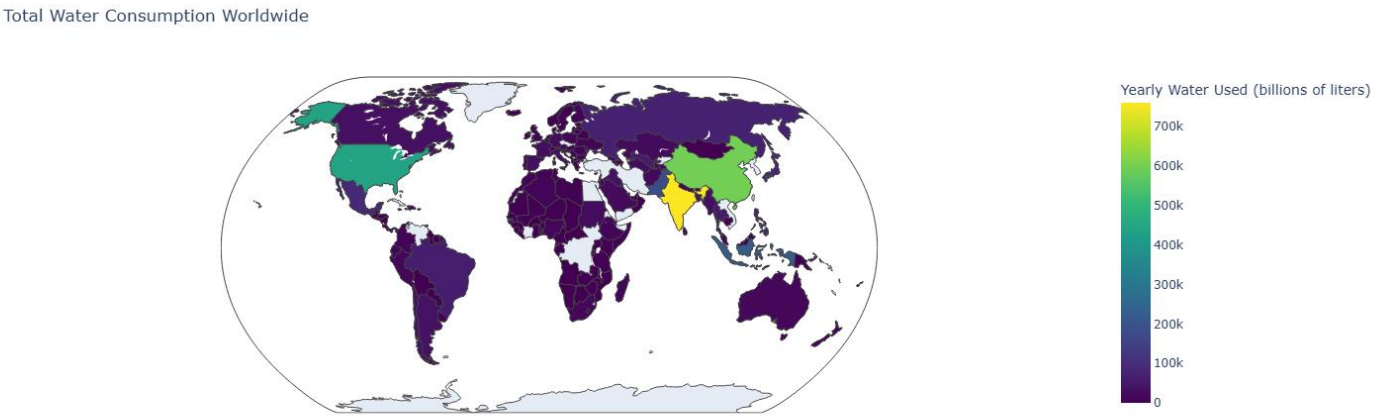
4) Conclusion

We can see that some kind of pattern is being traced. **There are factors that affect water** somehow, but still, **nothing affects daily water usage per capita**.

7. Data Visualization

1) World Maps

I. Figure 1: World Map of Water Consumption

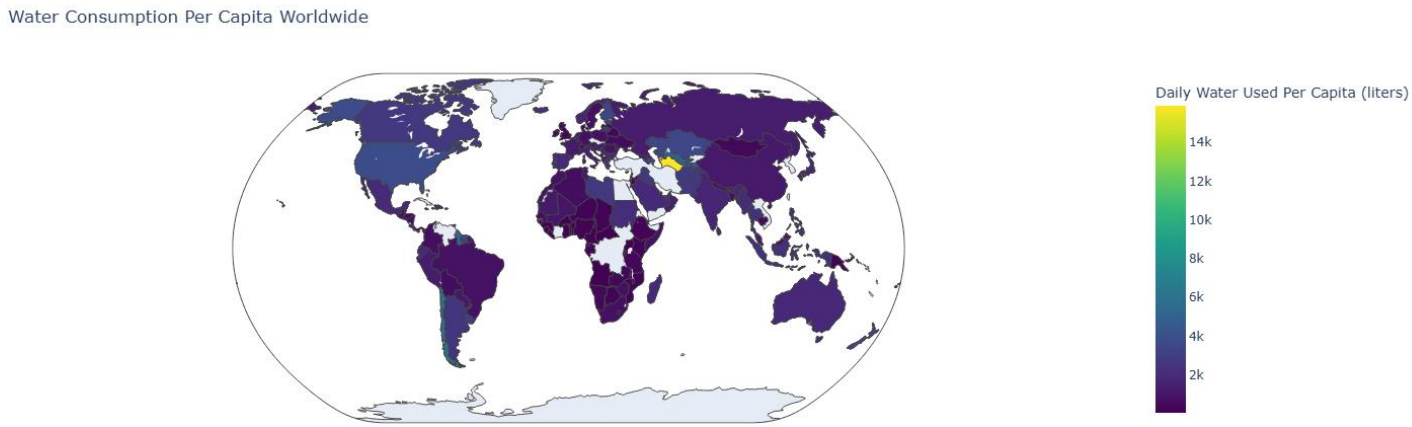


*It is interactive in python notebook

Choropleth map was used to show total water consumption for countries. Lighter color represents higher consumption. Highest value: India – 761000 billions of liters.

There is a trend, countries with higher population tend to have higher water consumption.

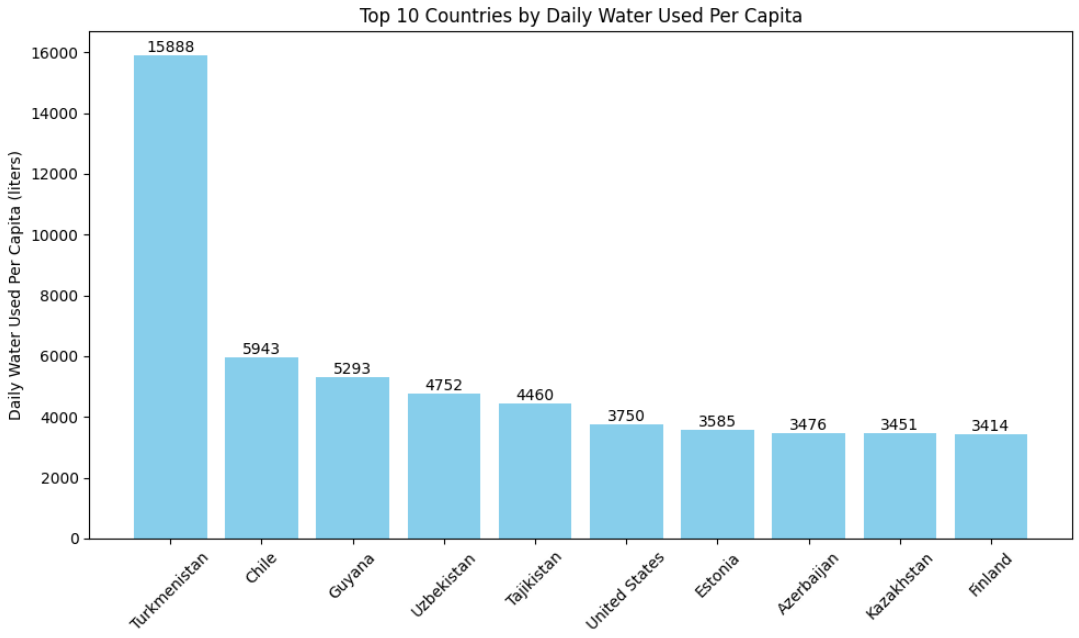
II. Figure 2: World Map of Per Capita Water Consumption



Here is water consumption per capita on choropleth map. Lighter color represents higher consumption. We can see the contrast, only one highly outstanding country here: Turkmenistan – 15.888 liters a day. Factors here unclear, mainly because of limited access to comprehensive data.

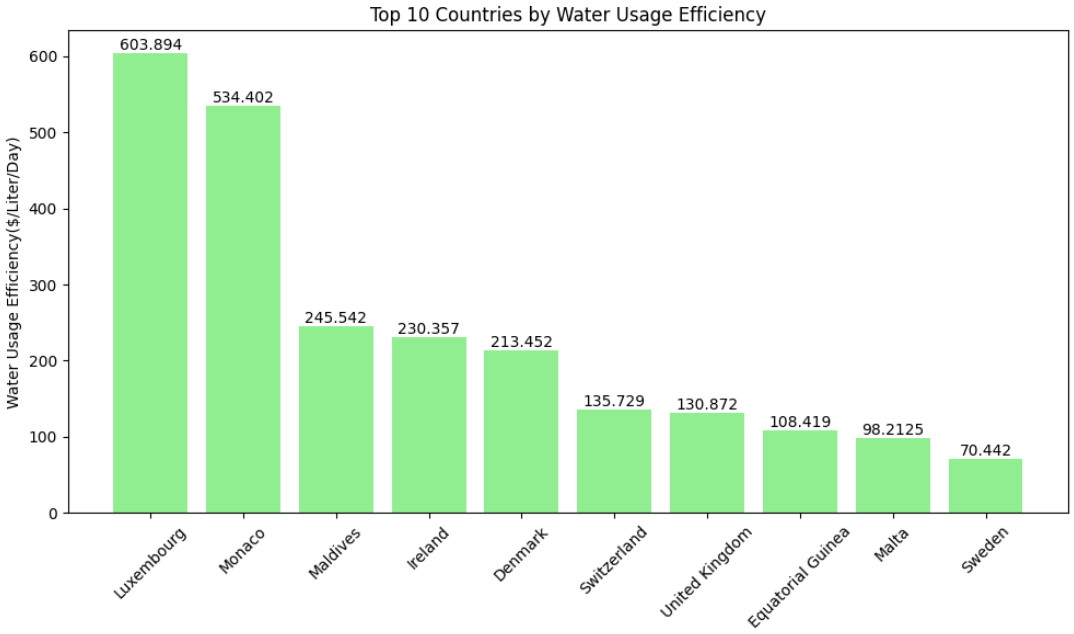
2) Bar Charts

I. Figure 3: Bar Chart of Daily Water Consumption per Capita



On this bar chart we can see that Turkmenistan is the leader, I assume because of high water usage in agriculture (there is no dataset specifically about agriculture water use, so it is difficult to prove statistically). One of the questions for this dataset was *'Does higher water consumption level lead to better usage efficiency?'*, now we got first part of the answer.

II. Figure 4: Bar Chart of Water Usage Efficiency



Differences are clearly outstanding. None of the countries from top 10 water per capita consumption are in top 10 by water usage efficiency. Country with highest water usage efficiency is Luxemburg, followed by Monaco. We can assume that **GDP per capita affects water usage efficiency more significantly**.

Conclusion

Project goal was to analyze worldwide water consumption, especially water consumption per capita. Economic and demographic factors were used for comparison.

Answering our questions from the beginning, I can conclude that **there are factors that affect water consumption**. This evidence is supported by a significant correlation between population and total yearly water consumption, ANOVA test on water usage efficiency across urbanization levels, linear regression predictions of yearly water usage and efficiency. Additionally, visualizations lend further support to the conclusions.

In summary:

- Factor that **affected water the most is Population**, due to strongest correlation with total water usage.
- Water usage efficiency was calculated. It is assumed that higher water consumption does not lead to higher efficiency.
- Countries that use water the most: Countries like Turkmenistan, Chile, and Guyana lead in per capita water usage, while populous countries like China and India have high total usage.
- The average daily per capita water use is approximately 1,363 liters, but with wide variation between countries.
- ANOVA showed significant differences in water use efficiency between urbanization groups, but no difference in per capita use by GDP levels.

Regarding to the main question, in this dataset **none of the factors affect daily water consumption per capita.** Correlation, linear regression and Welch's T-test didn't show any statistically significant relationships. Potential reasons for that:

- Inconsistency of the data set, including data from too wide range.
- Water consumption per capita is influenced by complex factors. Cultural, geographical and infrastructural factors are difficult to capture.
- Lack of proper agricultural data available.
- Fact that data on temperature gets averaged in the scope of countries.

These results highlight the need for more comprehensive and consistent data, incorporating a broader range of variables, to fully understand the drivers of daily per capita water use.

Sources:

Data Visualization and Presentation method class: W6-W9

https://www.kaggle.com/datasets/shuvokumarbasak4004/global-water-usage-statistics

https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS

https://data.worldbank.org/indicator/ER.H2O.FWIN.ZS