Data Management Culminating Assignment: Correlation Between World Population and Global Water Use

2022-06-16

Arshia Barootkoob Dezfooli

MDM4U

Ms. Sardellitti

Table of Contents(pages)

Introduction

Water use and the amount of water that we consume everyday has become an interesting and crucial topic for years. Simultaneously, world population and its growth has made some of the people worried about the future of our planet, Earth. Basically, the more people we have on our planet, the more water is expected to be consumed everyday. That's what makes people worried. According to the World Wildlife Fund (Canada's largest international conservation organization), 1.1 billion people are dealing with water shortage in the world(Jay, 2020). These concerns together, provide enough reasons for investigation. Therefore, I decided to technically check if there is really a correlation between world population and the amount of global water consumption.

Thesis Question

Is there a correlation between world population and global water use?

Supporting Arguments

1. Every human needs an average of 3.2 liters of water to drink every day which means that they would need a minimum amount of 1168 liters of water per year. This means adding one person to the population of the world would add nearly 1200 liters of water to the amount of global water consumption (per year).
2. Some people have kidney problems which results in having a higher need of drinkable water throughout their life. An example of these problems is nephropathy.
3. Due to accidents, there is a chance that water valves are left open. The wasted amount of water is considered a consumed amount of water. This probability can get higher as more people exist. Data shows humans waste 30 gallons of water everyday(Gov, 2020).

Hypothesis

I assume, there will be a strong, positive correlation between world population and global water consumption. As the world population increases, global water consumption will also increase.

Data Analysis

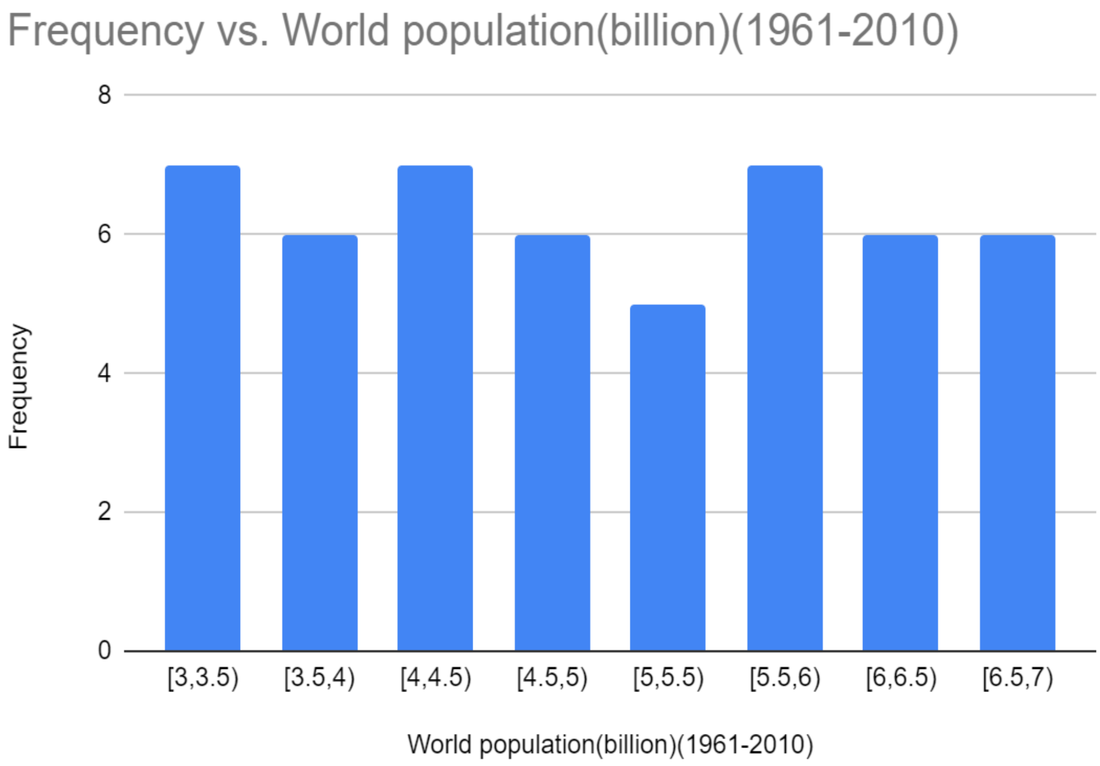
First set of data (World population) can be found here:

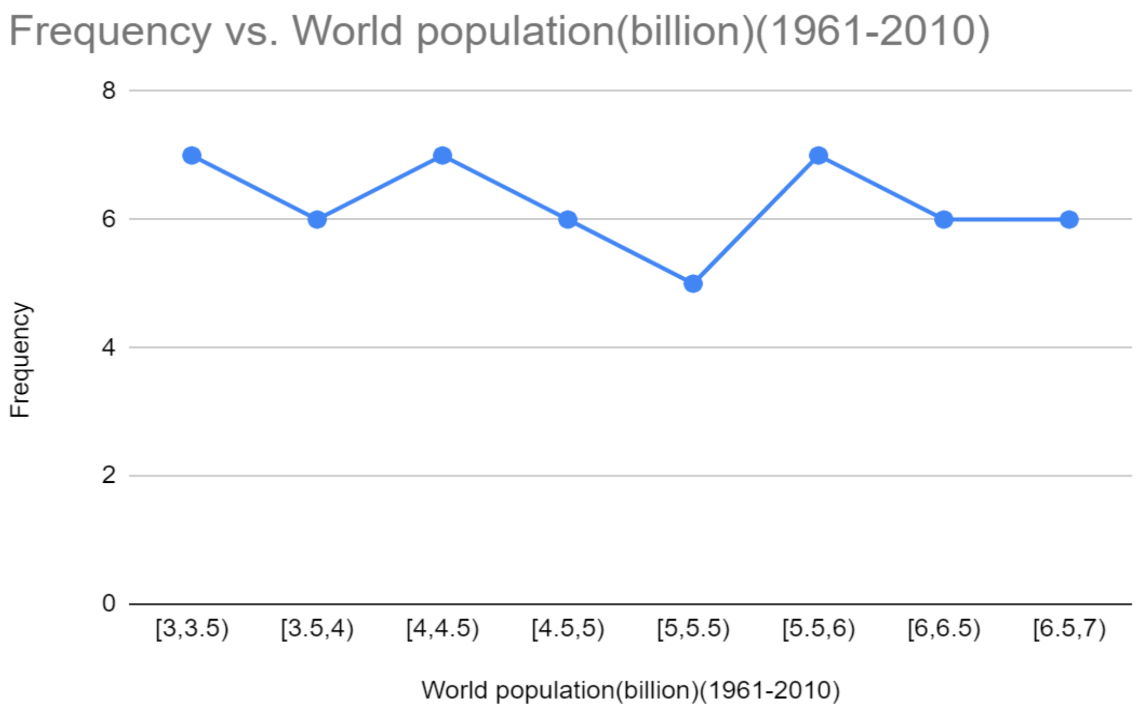
[World Population Growth - Our World in Data]()

Second set of data (global water use) can be found here:

[Water Use and Stress - Our World in Data]()

One Variable Data (World population):





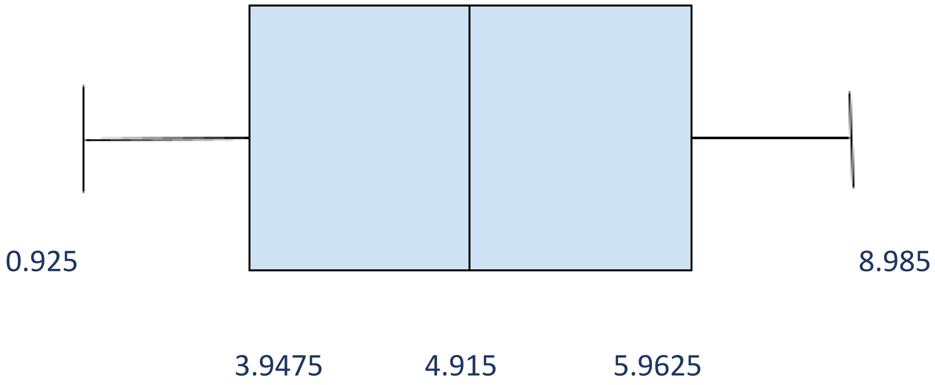
Outlier Calculations

IQR = Q3-Q1	Q3+1.5(IQR)
= 5.9625-3.9475	= 5.9625+1.5(2.015)
= 2.015	= 8.985 (max)

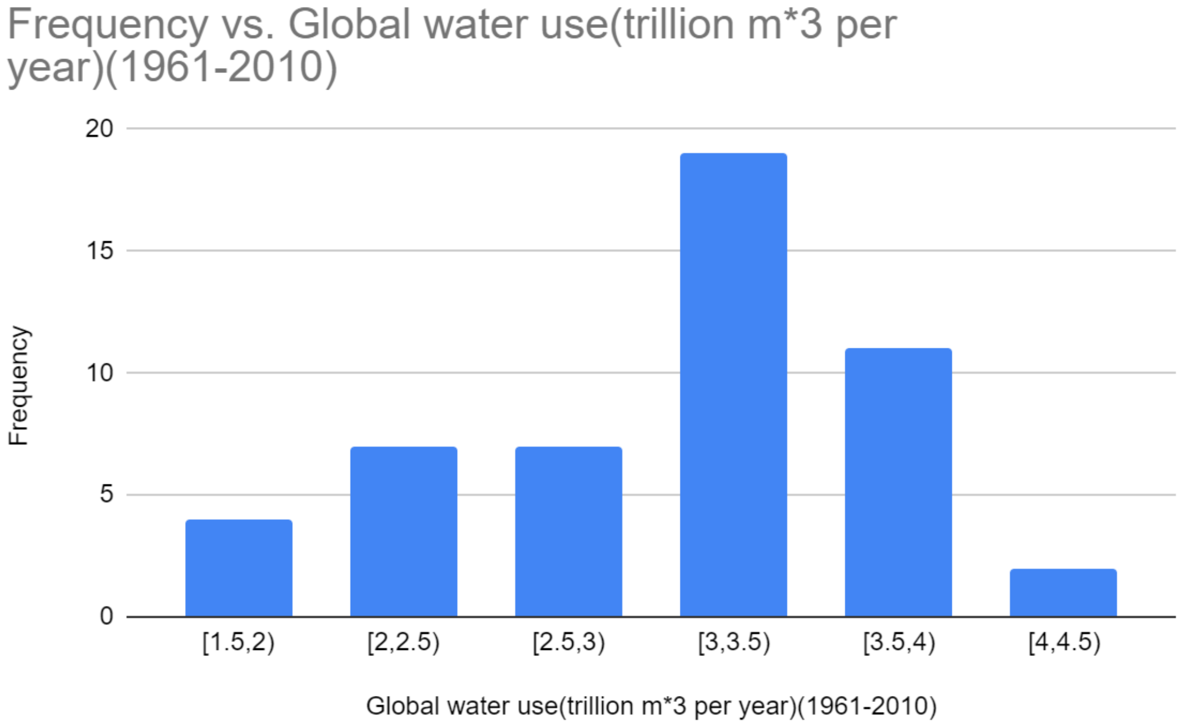
Q1-1.5(IQR)
=3.9475-1.5(2.015)
= 0.925 (min)

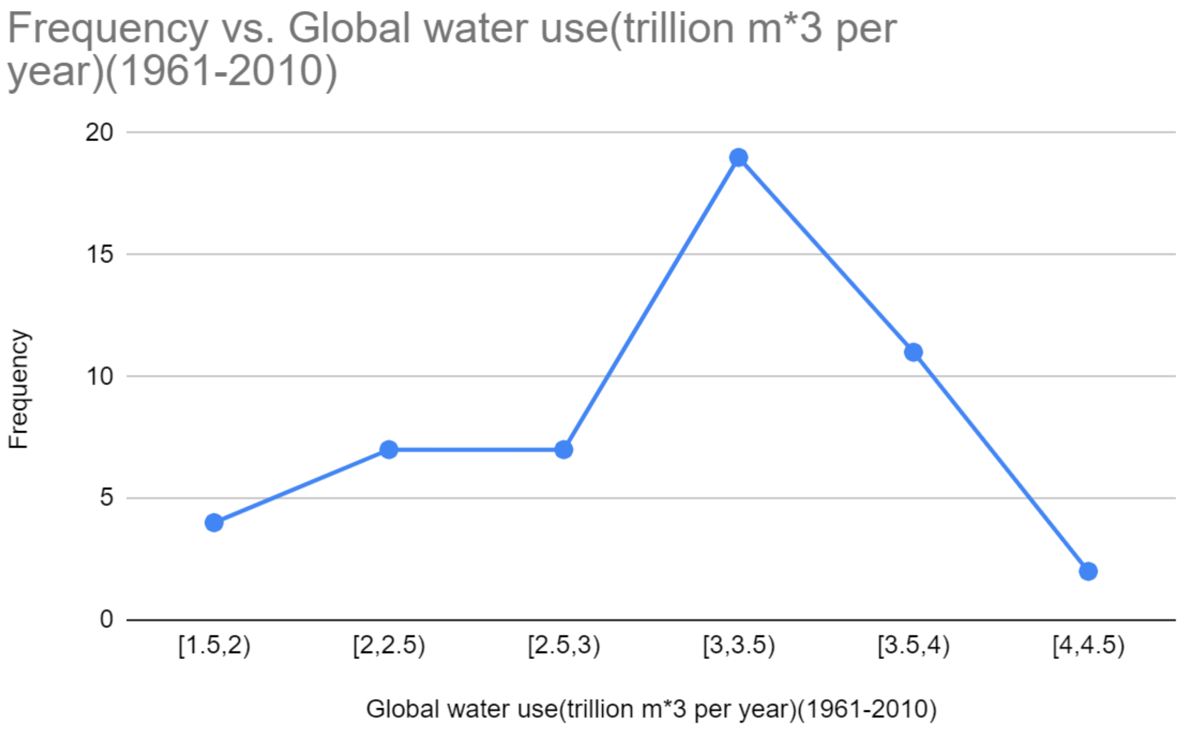
Data for the first variable is between our min and max values.
Therefore, there are no outliers.

Box and Whisker plot



One Variable Data (Global Water Use):





Outlier Calculations

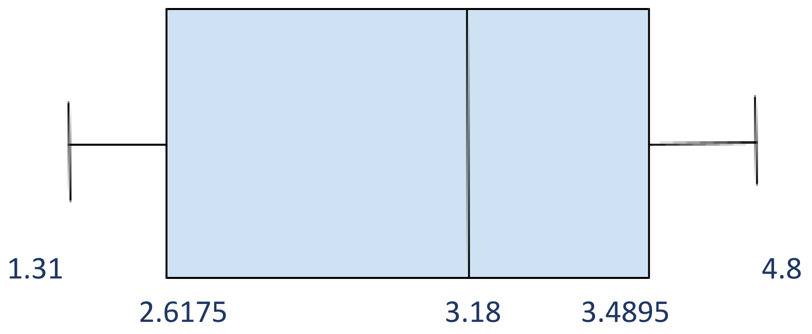
IQR= Q3-Q1
= 3.4895-2.6175
= 0.872

Q3+1.5(IQR)
= 3.4895+1.5(0.872)
= 4.8 (max)

Q1-1.5(IQR)
= 2.6175-1.5(0.872)
= 1.31(min)

Data for the second variable is between our min and max values.
Therefore, there are no outliers.

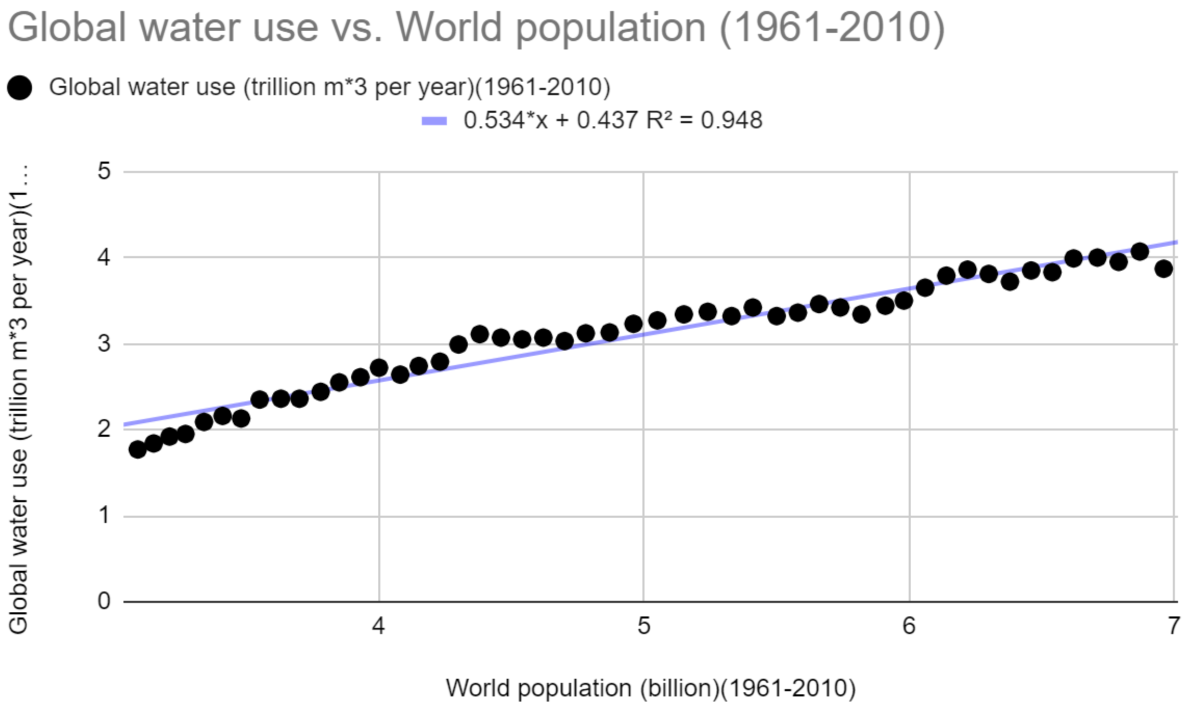
Box and Whisker plot



Measures of Central Tendency

	World population	Global water use
Mean	**4.958**	**3.083**
Median	**4.915**	**3.18**
Mode	**none**	**2.36**
Q1	**3.9475**	**2.6175**
Q3	**5.9625**	**3.49**

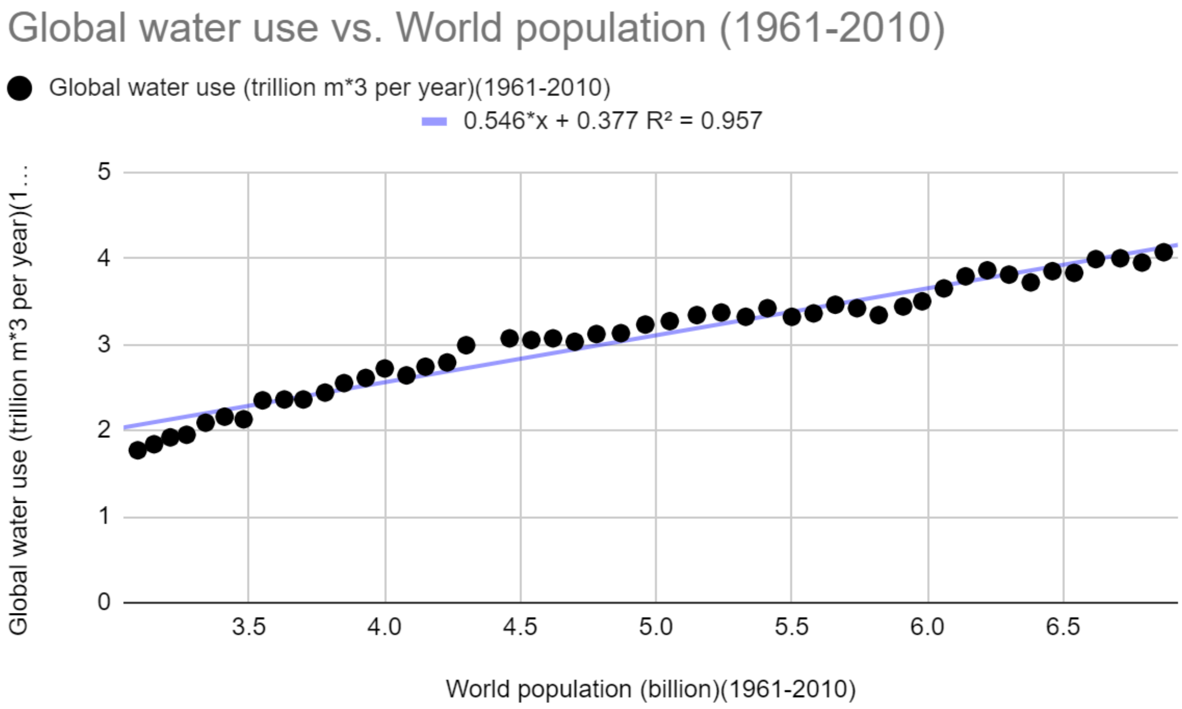
Two Variable Data



Equation of the line of best fit: y= 0.534x + 0.437

The coefficient of determination: (R^2)= 0.948
The correlation coefficient: (R)= 0.974
Apparent outliers: (4.38,3.11) & (6.96,3.87)

Two Variable Data(with removing the apparent outliers)

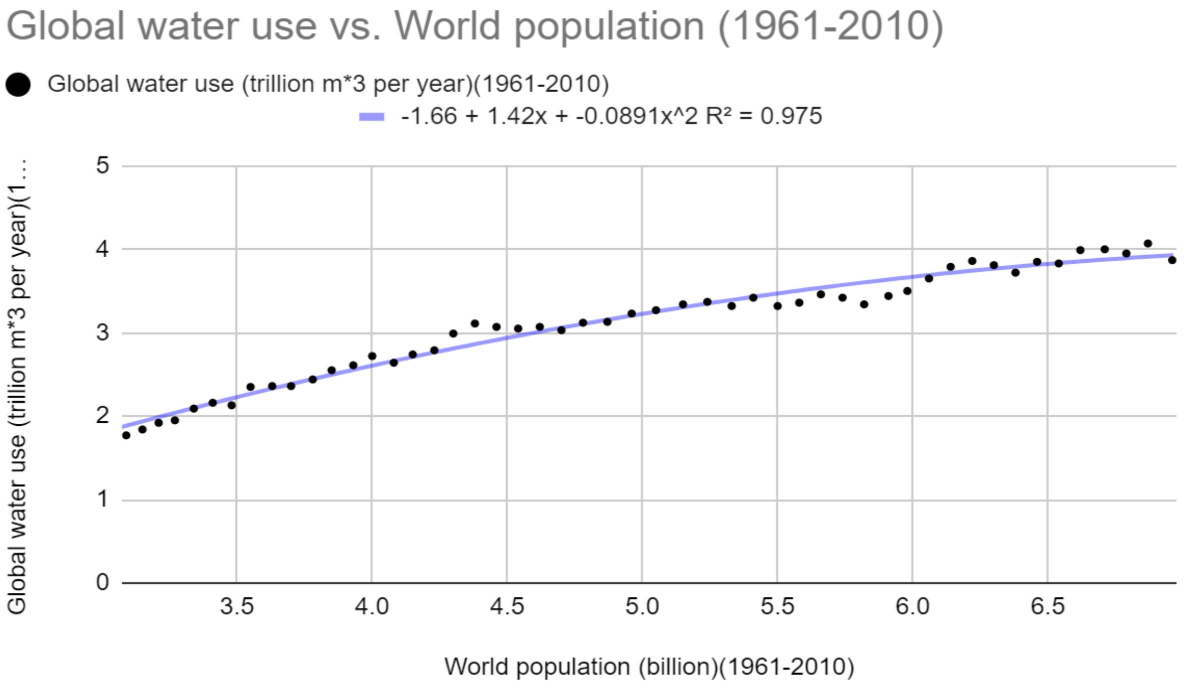


Removing the apparent outliers did not cause a significant change in the correlation.

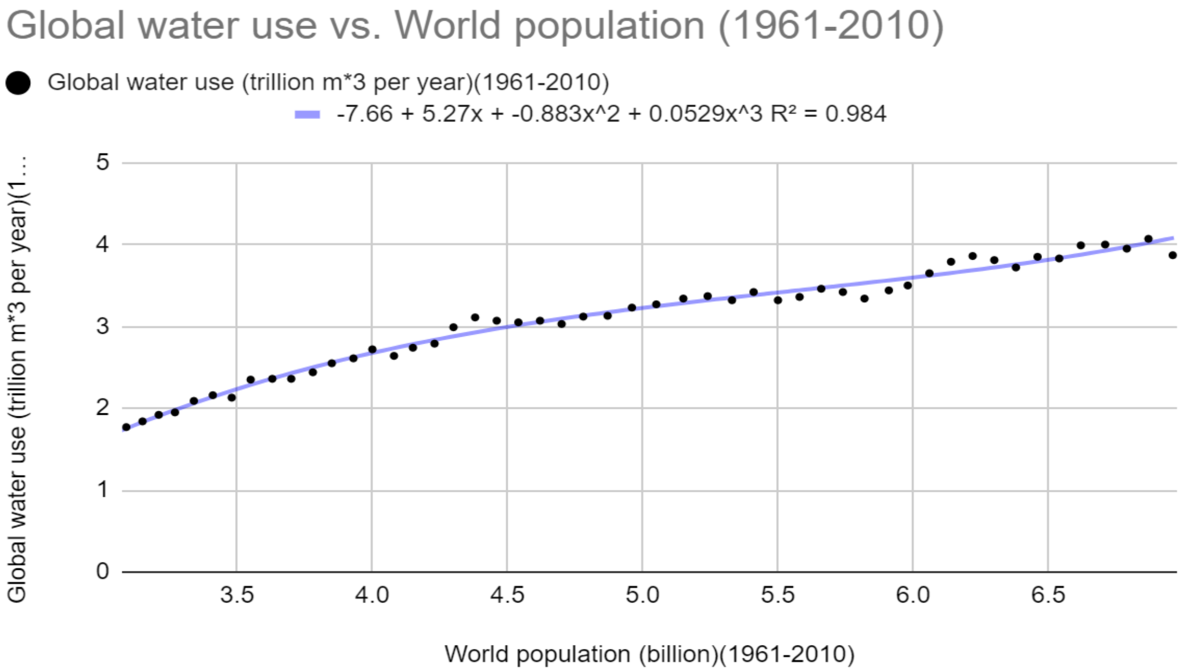
The R^2 value increased by 0.009, so technically the correlation became stronger, but only by a tiny amount. (New correlation coefficient is 0.978)

Residual value of the first outlier= 0.334 (approximately)
Residual value of the second outlier= -0.28 (approximately)

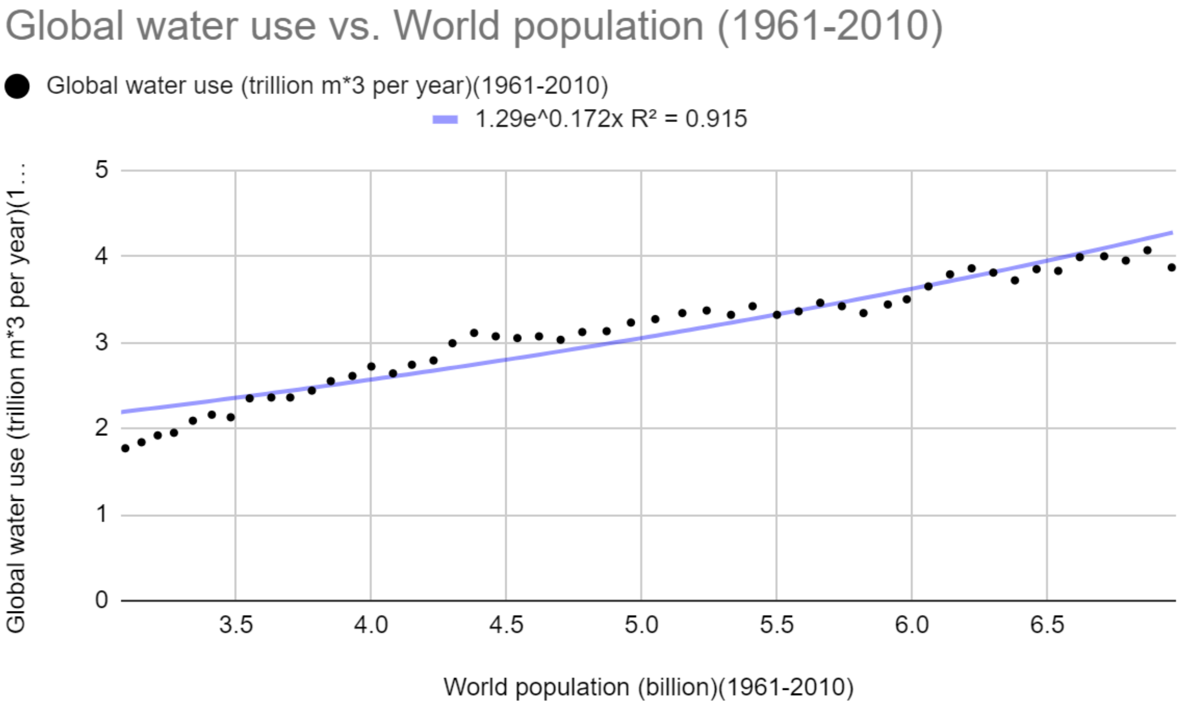
Non-Linear Regression (quadratic)



Cubic:



Exponential:



Model	Equation of the Line/Curve of Best Fit	Coefficient of Determination (r²)
Quadratic	y= -0.0891x^2+1.42x-1.66	0.975
Cubic	y= 0.0529x^3-0.883x^2+5.27x-7.66	0.984
Exponential	y= 1.29e^0.172x	0.915

Data Discussion

One Variable – World Population

The interval with the lowest frequency is [5,5.5). Among my data, from 1961 to 2010, the highest world population relates to 2010, with 6.96 billion people and the lowest world population relates to 1961, with 3.09 billion people. By looking at the frequency polygon it is apparent that the data is not distributed unimodally and there is a kind of a fluctuation. Also, we can see that the highest frequency happens three times throughout the frequency polygon(and the line chart). The mean is 4.958, and the median is 4.915. There is not a significant difference between the mean and the median. In this case the mean is greater than the median by 0.043. This situation can indicate that there is some consistency in the data. In other words, the mean and median are not too directed into one side of the data, whether it's the first data point's side or the last. There's no mode which means that no specific population happened twice between 1961 and 2010. The Box and Whisker Plot for these data does not have any outliers, and the median doesn't seem to shift towards Q1 or Q3.

One Variable - Global Water Use

The interval with the highest frequency is [3,3.5). The highest global water use is related to 2009, with 4.07 trillion m^3 per year and the lowest global water use relates to 1961 with 1.77 trillion m^3 per year. By looking at the frequency polygon, it is evident that the data is distributed unimodally, or in one peak. Also, we can see that the lowest frequency happens in the last interval. The mean is 3.083, and the median is 3.18. There is approximately a difference of 0.01 between the mean and the median. The mode is 2.36 which appears 2 times out of 50 values. Therefore, the mode does not have a big impact on the mean, median or the consistency of the data. The Box and Whisker plot for these data does not have any outliers, and the median is shifted towards Q3.

Two variables

The data has a strong, direct correlation. This is because R > 2/3, as R = 0.974. The slope of the line of best fit is 0.534. This means for every 1 billion increase in the world population, there is a 0.534 trillion m^3 increase in global water use per year. The y-intercept of the line of best fit is 0.437. Technically, this means that if there were 0 people in the world, there would still be 0.437 trillion m^3 global water use per year. This is actually referring to the amount of water that is either wasted or not actually used by humans at all.

Linear or Non-Linear Regression?

In general, a positive linear regression seems to be a good model as we expect the global water use to increase while observing an increase in the world population (as it was stated in the supporting arguments). However, this type of correlation doesn't have to be shown with a straight line. The growth of population might imply the growth of global water use, but it doesn't necessarily have to be that consistent. That's why a wide cubic model seems to be a better representation of my data. Aside from the visual and general explanations, mathematically, the more the coefficient of determination, the better representation of the model(data) its line/curve of best fit can be. In this case, the Cubic model is the most appropriate regression model for my data since it has the highest coefficient of determination(r^2).

In other words, the order in which the models can be set from best(most appropriate) to worst(least appropriate) based on their value of r^2 to represent the data is:

1. Cubic
2. Quadratic
3. Linear (r^2=0.948)
4. Exponential

Causal Relationship

The causal relationship of the data is a cause & effect relationship. This is because by looking at the graphs, scatter plot and the whole data, we can get the general understanding that the more the population grows, the higher the amount of global water use gets. In other words, as the world population increases, the amount of global water consumption increases as well(as mentioned in the hypothesis). Also, mathematically, our R value for the data was 0.974 which indicates a strong direct correlation which implies a cause and effect relationship. (More justification is stated after Extraneous Variables inside the Important note).

Extraneous Variables that might have affected my dependent variable(Global water consumption)

1.Wasting water. Each year, an incredible amount of water is wasted. As mentioned in the third supporting argument, every human wastes 30 gallons of water every day. This can have a negative impact on the amount of global water use in the sense of reality. In other words, what we are seeing as the amount of water used by people every year is potentially not the real quantity of water.

2.Allocating the human water consumption sources to other types of usage(like agriculture). Well, we know that every year, people use a specific amount of water. We know that this consumed quantity of water is recorded as a fact every year. But what if the wrong resource of consumption is recorded? For instance, if an agricultural resource of water use is recorded as a part of my dependent variable, the amount of global water use (for that specific year of wrong record) is wrong.

3.Unknown sources of water consumption. Aside from the fact that there are potentially some sources of water use that were not even included in the data at the time, sometimes, data is recorded and it’s not necessarily wrong, but it’s not essentially right either. In fact, sometimes, we don’t know what the data is and we don’t know where it really comes from. In this case, that unknown data can be that unknown source of water consumption. In order to make it easier to visualize, imagine a water valve which is used for different purposes. It is used for daily usage, agricultural and environmental purposes. We don’t know when, where and how water is exactly used through this valve. We have the data, but basically, it's unknown data. (Regarding the independent variable, there might have been some data exclusions as well, meaning that the real world population is potentially more than the data submitted)

Important note: Although these extraneous variables might have affected the global water use, they are not special variables. On a larger scale, they don’t happen at specific times. In other words, each year, there might have been a chance that any of these variables would have affected the data. With the consideration of this idea, my extraneous variables can’t interfere with the fact that the correlation between my independent variable(world population) and dependent variable(global water use) is a cause and effect relationship.

Conclusion

All in all, there is a strong positive correlation between the world population and global water use. This means my hypothesis was correct, as I predicted a strong positive correlation between these two variables. I got the data from OurWorldinData site, and had no problem in collecting it except for the existence of so many graphs which made it a little hard to find my intended data. Both data sets have no outliers, so there are no values that are skewing the results. These data do have a cause and effect relationship, as based on the data and two-variable data graphs, increase in world population directly causes the global water use to increase.(Considering that some external factors might have affected both variables). After doing this assignment, it would be interesting to research how the GDP(Gross Domestic Product) of different countries would impact the amount of their specific water uses, the combination of which would be considered as the global water use. This is because, I think it would be interesting to see if there is a correlation between the wealth of a country and its amount of water consumption. Would a lower GDP indicate a financial flop and therefore a decrease in the amount of water production and therefore a decrease in the amount of water consumption? I think this would be a compelling follow up.

References

Roser, Max, et al. “World Population Growth.” *Our World in Data*, 9 May 2013,

https://ourworldindata.org/world-population-growth.

Ritchie, Hannah, and Max Roser. “Water Use and Stress.” *Our World in Data*, 20 Nov. 2017,

https://ourworldindata.org/water-use-stress.

CA, WWF. “World Wildlife Fund Canada.” *WWF.CA*, 8 June 2022,

https://wwf.ca/.

WaterFilterMarket, Jay. “Water Shortages: Causes, Effects, and Solutions.” *The Water Filter*, 15 Jan. 2020,

https://thewaterfiltermarket.com/water-shortages/.

CA, WWF. “Water Scarcity.” *WWF*, World Wildlife Fund, 2022,

https://www.worldwildlife.org/threats/water-scarcity.

Gunnars, Kris. “How Much Water Should You Drink per Day?” *Healthline*, Healthline Media, 5 Nov. 2020,

https://www.healthline.com/nutrition/how-much-water-should-you-drink-per-day#:~:text=How%20much%20water%20you%20need%20depends%20on%20a,15.5%20cups%20%283.7%20liters%29%20a%20day%20for%20men.

Gov, Doh. “Stop Water Waste the Average Person Unknowingly Wastes.” *Stop Water Waste*, Apr. 2020,

https://doh.wa.gov/sites/default/files/legacy/Documents/Pubs/331-450.pdf.

Appendix 1 – Raw Data

World population (billion)(1961-2010)	Global water use (trillion m*3 per year)(1961-2010)
3.09	1.77
3.15	1.84
3.21	1.92
3.27	1.95
3.34	2.09
3.41	2.16
3.48	2.13
3.55	2.35
3.63	2.36
3.7	2.36
3.78	2.44
3.85	2.55
3.93	2.61
4	2.72
4.08	2.64
4.15	2.74
4.23	2.79
4.3	2.99
4.38	3.11
4.46	3.07
4.54	3.05
4.62	3.07
4.7	3.03
4.78	3.12
4.87	3.13
4.96	3.23
5.05	3.27

5.15	3.34
5.24	3.37
5.33	3.32
5.41	3.42
5.5	3.32
5.58	3.36
5.66	3.46
5.74	3.42
5.82	3.34
5.91	3.44
5.98	3.5
6.06	3.65
6.14	3.79
6.22	3.86
6.3	3.81
6.38	3.72
6.46	3.85
6.54	3.83
6.62	3.99
6.71	4
6.79	3.95
6.87	4.07
6.96	3.87

Appendix 2 – Grouped Data

World population(billion)(1961-2010)	Frequency
[3,3.5)	7
[3.5,4)	6
[4,4.5)	7
[4.5,5)	6
[5,5.5)	5
[5.5,6)	7
[6,6.5)	6
[6.5,7)	6

Global water use(trillion m*3 per year)(1961-2010)	Frequency
[1.5,2)	4
[2,2.5)	7
[2.5,3)	7
[3,3.5)	19
[3.5,4)	11
[4,4.5)	2