Final_LC_JB

Luke Clement, JuJuan Brown

2022-05-04

Load Packages

```
#install.packages("tidyverse")
#install.packages("rvest")
#install.packages("naniar")
#install.packages("readxl")
library("readxl")
library("naniar")
library("tidyverse")
library("rvest")
library("maps")
library("stringr")
library("glue")
```

Proposal and Introduction

Title: The Impacting Factors on Water's Survival

Purpose:

98.28% of all water on Earth is in liquid form with the rest being in glaciers, ice, snow, and other small forms. Of that, 97.5% of that is salt water, which without desalination is unusable for human consumption, which means that we can safely use only a small fraction of the planet's liquid water. Our goal is to bring into perspective how small this fraction of water is and to see if it is for us as a society to invest in improved technologies for desalination of saltwater and purification of groundwater.

To accomplish this, we are going to look at the water usage for each country in the world. From here, we will look into the progression of the water usage over time. From there, our goal is to see how long our current supply of usable water would last if we continued following the same pattern of water usage.

Other environmental issues like pollution and greenhouse gas emissions have often been assumed to be linked to population and economic growth. To that end, we wanted to investigate whether there was a connection between a countries economy and their water usage.

Data:

1. https://www.wri.org/data/aqueduct-projected-water-stress-country-rankings

This is the World Resource Institute and it is where we will pull out data for water stress. Water stress is the situation where there is not enough water to provide the needs of people and the environment in a local area. This cite includes this data on global scale which will allow us to see which countries are struggling with water currently. This data, fortunately, was already in the form of an Excel doc. This table also includes predictions for water stress in countries around the world up until 2050.

2. https://www.worldometers.info/water/

We did need to scrape this data to use it. This data has 179 countries and instead of measuring water stress, it simply gives the amount of water that a country is using annually. This data was also very nice since it

included daily water use per capita and the population of these countries at the time. The only downside to this data set was that it was gather in 2009 so it is a little old. We do have other data sets that are more recent so we can try to make projections based on this data and we can see how accurate those projections were by measuring it up against our other more modern data sets.

3. https://www.imf.org/en/Publications/WEO/weo-database/2021/October/download-entire-database #For this dataset, download "By Countries"

This is the data that we will use to measure a country's economic health. We will pull GDP, GDP per capita, and measure the population for more recent years. This is especially nice since the data goes to as recent as 2020. It also provides predictions for certain variables through 2025.

Variables**:**

Country - We will be comparing different countries.

PPPPC - Gross domestic product based on purchasing power parity (PPP) per capita GDP. This will be used to measure the economic status of a country on a per person basis. (GDP per capita)

PPPGDP - Gross domestic product based on purchasing-power-parity (PPP) valuation of the country. This will be used to measure the total economic output of a country without accounting for population size. (GDP)

Water Stress - "Water stress measures total annual water withdrawals (municipal, industrial, and agricultural) expressed as a percentage of the total annual available blue water. Higher values indicate more competition." 5 indicates severe water stress, while 0 indicates no water stress.

Yearly Water Used - The total amount of water used by a country. This will be used to measure the annual amount of water consumed by country so it can be compared to other variables to see if there is a correlation.

Population - The population of each country. This will be compared with water usage.

Project:

There is approximately 326 Quintilian gallons of water on the planet. Of that massive amount, 98.28% of all water on Earth is in liquid form with the rest being in glaciers, ice, snow, and other small forms. That sounds great until you realize that 97.5% of that liquid is salt water, which without desalination is unusable for human consumption, which means that if don't make changes, it wont be long until our current amount of consumable water wont be enough. Our goal is to bring into perspective how small this fraction of water is and to show how important it is for us as a society to invest in improve technologies for desalination of saltwater and purification of groundwater.

We are hoping to deliver an analysis on the factors that determine water consumption and to estimate how much longer water supplies will last based on select factors. We want to bring perspective how much water we have and to show how important it is for us as a society to invest in improve technologies for desalination of saltwater and purification of groundwater. We are hoping to have a graph showing how much longer before we run out of blue water based on a statistical method for calculating future water consumption. We are hoping to create bar graphs/scatter plots comparing effects of certain factors on water usage. We are also hoping to to show what would happen if starting now, we stopped wasting water.

Gathering Data:

```
#We had to code things separately, so to unite this into one document, we needed to double define somet

#Scraping data
(webpage <- read_html("https://www.worldometers.info/water/"))
```

```
## {html_document}
## <html lang="en">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
## [2] <body>\n<script> (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[ ...
```

```r
raw_data<-webpage %>%
  html_nodes("td:nth-child(2) a , td:nth-child(1) a") %>%
  html_text()
#Cleaning the data by removing commas
annual_water<-raw_data %>%
  str_subset("[:digit:]") %>%
  str_remove_all("\\,") %>%
  as.numeric()
#Turning it into a tibble
water_data<-tibble(
  countries = raw_data %>%
  str_subset("[:alpha:]"),
  water_usage=annual_water
)

#The two data sets I downloaded
money_data <-  read_excel("Money2.xlsx")
water_stress<-
read_excel("aqueduct-water-stress-country-rankings-data-set.xlsx", sheet = "2020 BAU")
#"C:/Users/jajua/Downloads/water_stress.xlsx"

# Creating Backup Versions of our important data
water_data_default <- water_data
money_data_default <- money_data
water_stress_data_default <- water_stress

#Duplicated Variable Names
water <-read_excel("aqueduct-water-stress-country-rankings-data-set.xlsx", sheet = "2020 BAU")
  #read_xlsx("aqueduct-water-stress-country-rankings-data-set.xlsx",sheet = 2)

water_usage_data<-water_data %>% rename("water_data" = "water_usage")
water_scarcity <- water %>%
  rename("countries" = "Name") %>%
  mutate(countries = case_when(countries == "Republic of Serbia" ~ "Serbia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "United States of America" ~ "United States", TRUE ~ countri
  mutate(countries = case_when(countries == "United Republic of Tanzania" ~ "Tanzania", TRUE ~ countrie
  mutate(countries = case_when(countries == "Republic of Serbia" ~ "Serbia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Republic of the Congo" ~ "Republic of Congo", TRUE ~ counti

water_usage_data <- water_usage_data %>%
  mutate(countries = case_when(countries == "United States of America" ~ "United States", TRUE ~ countri
  mutate(countries = case_when(countries == "Czech Republic (Czechia)" ~ "Czech Republic", TRUE ~ counti
  mutate(countries = case_when(countries == "Tanzania" ~ "United Republic of Tanzania", TRUE ~ countrie
  mutate(countries = case_when(countries == "Republic of Serbia" ~ "Serbia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Congo" ~ "Republic of Congo", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Guinea-Bissau" ~ "Guinea Bissau", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "DR Congo" ~ "Democratic Republic of the Congo", TRUE ~ cou
  mutate(countries = case_when(countries == "Republic of the Congo" ~ "Republic of Congo", TRUE ~ counti
  mutate(countries = case_when(countries == "Saint Lucia" ~ "St. Lucia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Saint Kitts & Nevis" ~ "St. Kitts and Nevis", TRUE ~ counti
  mutate(countries = case_when(countries == "United Republic of Tanzania" ~ "Tanzania", TRUE ~ countrie
money_data_default <-read_excel("Money2.xlsx")
  # read_xlsx("Money2.xlsx")
```

```r
  # read_excel("C:/Users/jajua/Downloads/money_data.xlsx")
money_data <-
  money_data_default %>%
  filter(`WEO Subject Code` %in% c("PPPPC", "PPPGDP", "LP", "NGDPD")) %>%
  select(-c(`WEO Country Code`, ISO, `Country/Series-specific Notes`, `Estimates Start After`, `Subject
  rename("WEO" = "WEO Subject Code")

population_data <-
  money_data %>%
  filter(WEO == "LP") %>%
  select(Country,`2020`) %>%
  rename("countries" = "Country") %>%
  rename("Population" = `2020`) %>%
  mutate(countries = case_when(countries == "Brunei Darussalam" ~ "Brunei", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Guinea-Bissau" ~ "Guinea Bissau", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "The Gambia" ~ "Gambia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Islamic Republic of Iran" ~ "Iran", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Korea" ~ "South Korea", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Kyrgyz Republic" ~ "Kyrgyzstan", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Slovak Republic" ~ "Slovakia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "St. Vincent and the Grenadines" ~ "St. Vincent & Grenadines
  mutate(countries = case_when(countries == "North Macedonia" ~ "Macedonia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Taiwan Province of China" ~ "Taiwan", TRUE ~ countries)) %
  mutate(countries = case_when(countries == "Lao P.D.R." ~ "Laos", TRUE ~ countries))

PPPGDP_data <-
  money_data %>%
  filter(WEO == "PPPGDP") %>%
  select(Country,`2020`) %>%
  rename("countries" = "Country") %>%
  rename("PPPGDP" = '2020') %>%
  mutate(countries = case_when(countries == "Brunei Darussalam" ~ "Brunei", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Guinea-Bissau" ~ "Guinea Bissau", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "The Gambia" ~ "Gambia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Islamic Republic of Iran" ~ "Iran", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Korea" ~ "South Korea", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Kyrgyz Republic" ~ "Kyrgyzstan", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Slovak Republic" ~ "Slovakia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "St. Vincent and the Grenadines" ~ "St. Vincent & Grenadines

PPPPC_data <- money_data %>%
  filter(WEO == "PPPPC") %>%
  select(Country,`2020`) %>%
  rename("countries" = "Country") %>%
  rename("PPPPC" = '2020') %>%
  mutate(countries = case_when(countries == "Brunei Darussalam" ~ "Brunei", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Guinea-Bissau" ~ "Guinea Bissau", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "The Gambia" ~ "Gambia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Islamic Republic of Iran" ~ "Iran", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Korea" ~ "South Korea", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Kyrgyz Republic" ~ "Kyrgyzstan", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Slovak Republic" ~ "Slovakia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "St. Vincent and the Grenadines" ~ "St. Vincent & Grenadines
```

```
  mutate(countries = case_when(countries == "North Macedonia" ~ "Macedonia", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Lao P.D.R." ~ "Laos", TRUE ~ countries))


#Luke's Definition of water_data
water_data <- full_join(water_scarcity,water_usage_data)
```

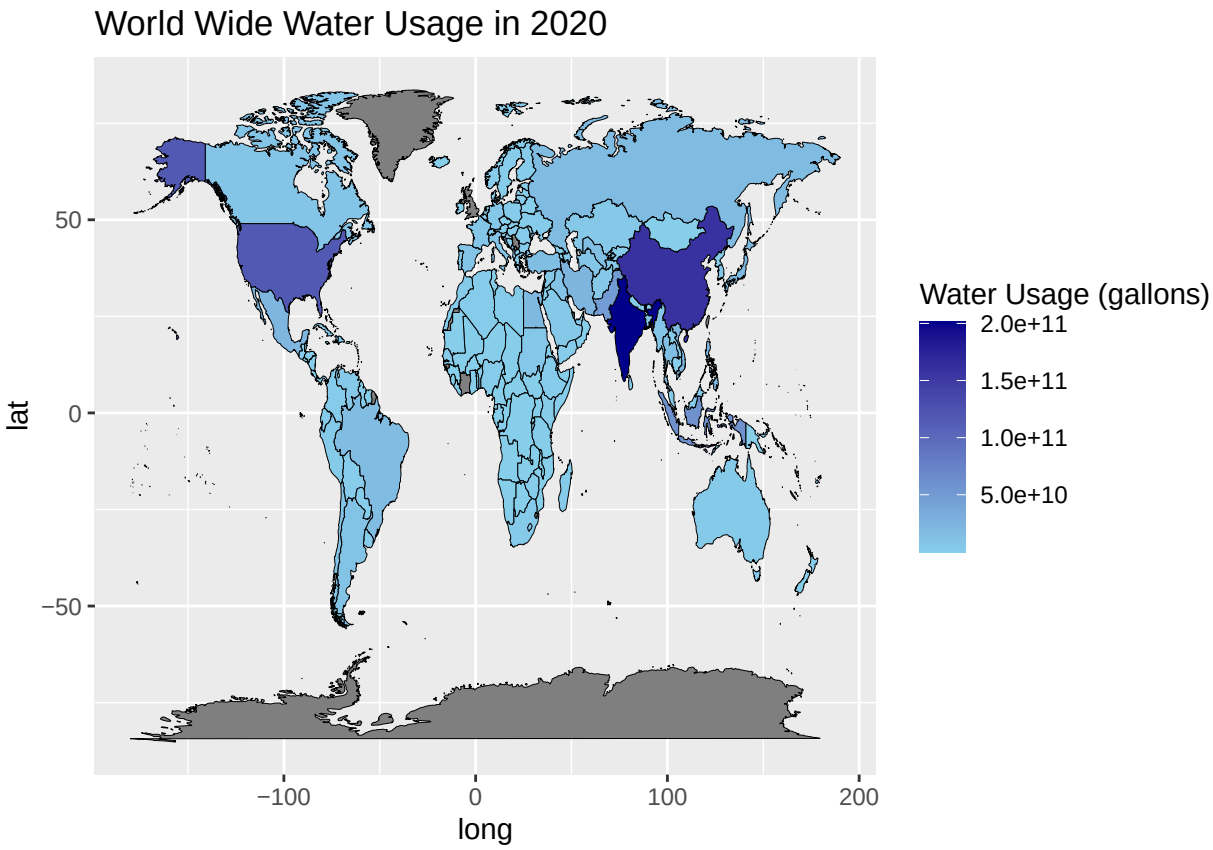
```
## Joining, by = "countries"
```

Looking into our data:

The first way we chose to help show how the remaining 81.5 Quintilian gallons of usable water was distributed around the Earth was with a choropleth map of each country's water usage. Our economic data also happened to have good values for populations, so we needed to combined the economic data set with the water data set. Since our data came from different sources, the naming of certain countries was not consistent between the two and the biggest challenge with creating this map was to find a way of giving all the maps consistent naming of all the countries. Once we did this, creating the map was fairly straightforward.

```
#Filtering the desired Economic Data
money_data <- money_data_default %>%
  filter(`WEO Subject Code` %in% c("PPPPC", "PPPGDP", "LP", "NGDPD")) %>%
  select(-c(`WEO Country Code`, ISO, `Country/Series-specific Notes`,
            `Estimates Start After`, `Subject Notes`)) %>%
  rename("region"= "Country")
```

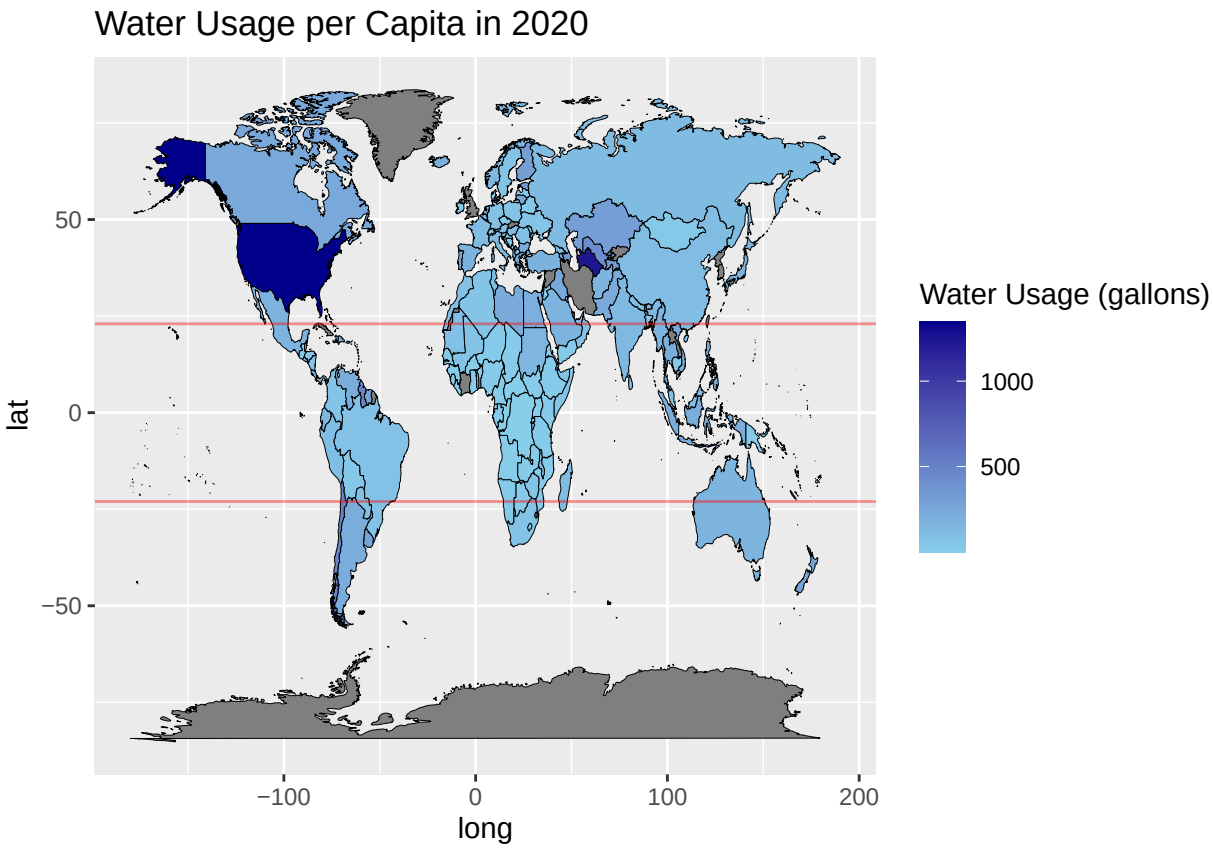
```
world<-map_data("world")


water_data <- water_data_default %>%
  mutate(countries = case_when(countries == "United States" ~ "USA",
                               TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Czech Republic (Czechia)" ~
                                 "Czech Republic", TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Serbia" ~ "Republic of Serbia",
                               TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "DR Congo" ~
                                 "Democratic Republic of the Congo",
                               TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Guinea-Bissau" ~ "Guinea Bissau",
                               TRUE ~ countries)) %>%
  mutate(countries = case_when(countries == "Congo" ~ "Republic of Congo",
                               TRUE ~ countries)) %>%
  mutate(water_usage= water_usage*0.2641720524) %>%
  rename("region"= "countries") %>%
  full_join(world, water_data, by= "region")

ggplot()+
  geom_map(data= water_data, map = world,
           aes(long, lat, map_id = region, fill = water_usage),
           color = "black", size = 0.1)+
  scale_fill_continuous(low = "skyblue", high = "darkblue",
                        name= "Water Usage (gallons)")+
  labs(title = "World Wide Water Usage in 2020")
```



We expected the U.S do be higher than most other countries, but were very shocked to find that countries like India had a higher overall water usage than the United States. From here, we wanted to explore why countries in what appears to be the Middle East have such high water usage. The first way we did this was to compare the worldwide water use in each country per capita.

```
#Finding water usage per capita

money_data <-money_data %>%
  filter(`WEO Subject Code`!="NGDPD") %>%
  filter(`WEO Subject Code`!="PPPGDP") %>%
  filter(`WEO Subject Code`!="PPPPC")

pop_data<- full_join(water_data, money_data, by= "region") %>%
  mutate(region = case_when(region == "United States" ~ "USA",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "Czech Republic (Czechia)" ~
                              "Czech Republic",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "Serbia" ~ "Republic of Serbia",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "DR Congo" ~
        "Democratic Republic of the Congo", TRUE ~ region)) %>%
  mutate(region = case_when(region == "Guinea-Bissau" ~ "Guinea Bissau",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "Congo" ~ "Republic of Congo",
                            TRUE ~ region)) %>%
  mutate(per_capita_use = (water_usage/(`2020`*1000000)))
```

```
#I have to manually enter certain countries because the name fix didnt work
pop_data$per_capita_use[c(78720:84472)]<-(1346.364)

ggplot()+
  geom_map(data= pop_data, map = world, aes(long, lat, map_id = region,
                fill = per_capita_use), color = "black", size = 0.1)+
  scale_fill_continuous(low = "skyblue", high = "darkblue",
                        name= "Water Usage (gallons)")+
  geom_hline(yintercept = c(-23, 23), color= "red", alpha = .4)+
  labs(title = "Water Usage per Capita in 2020")
```



Originally, we thought temperature was the reason why these areas high such an unexpectedly high overall water usage. To account for this, we included horizontal lines to draw the viewers focus to the torrid zone. The torrid zone is consistently the hottest area on on Earth, and it is the area between these two red lines. From this we can see that the per capita usage from the Earth's hottest countries is not noticeably higher than the per capita water use in other parts of the world, so we can infer that a country's temperature does not play a massive role in its overall water usage.

This graph made it clear that the U.S had an abnormally high per capita usage, which was to be expected, but we couldnt figure out what wouldve made the overall water usage of the Middle East so high if individuals weren't using more water than usual. From here, we wanted to see if there were connections to water usage that were outside of the control of the people like international trade. To bring this into perspective, we began seeing looking into connections between a country's economic health and its water usage.

Figuring Out What Variables Impact Water Usage:

We looked at three variables to see if they correlated with water usage and water stress/scarcity. These three

variables include Gross Domestic Product (GDP), GDP per capita/per person, and population size for every country. Note water stress and water scarcity are the same thing just named differently so that our code could be merged

We first looked at gdp per capita vs water usage. Does gdp per capita affect water usage?

We did this by taking gdp per capita from the money data set and water usage from the water worldometer dataset. First we looked at the covariance of each deciding it was best to look at it by taking the log of each variable. We created intervals to place countries into based on their gdp per capita. We then took the average water usage of these intervals and graphed the variables on a bar graph. A scatter plot was created by graphing each country by their gdp per capita and their corresponding water usage. We took the log of each variable to help them fit on the scatter plot. A line of best fit was also applied.

```
#water_data <- full_join(water_scarcity,water_usage_data) %>% drop_na()

# Builds data and interval
  ppppc_water <- full_join(PPPPC_data,water_usage_data) %>% drop_na() %>%
                 mutate(interval = PPPPC)
```

```
## Joining, by = "countries"
```

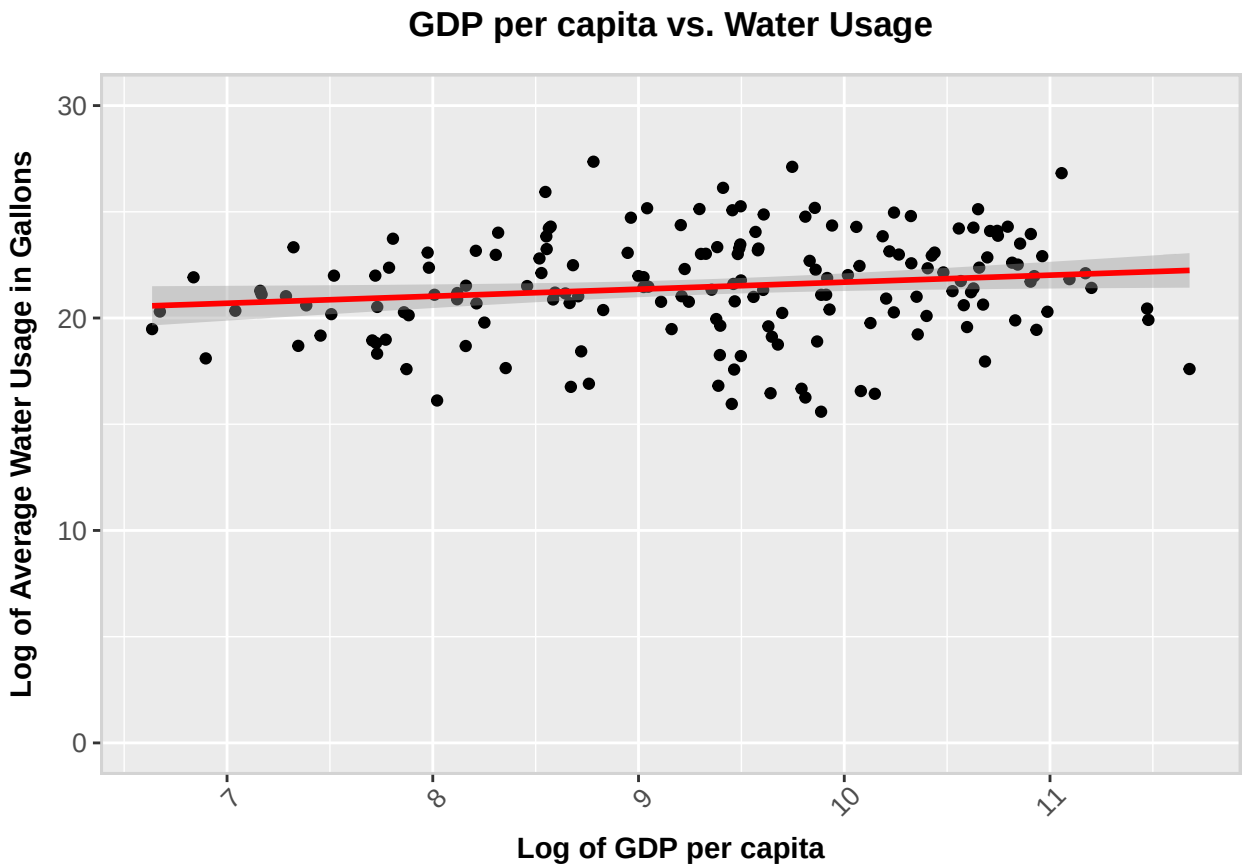
```
  ppppc_water_interval <- c(0,2500,10000,20000,50000,120000)

# Covariance
  ppppc_water %>%
  mutate(PPPPC = log(PPPPC)) %>%
  mutate(water_data = log(water_data)) %>%
  select(PPPPC, water_data) %>%
  cor() # No correlation between gdp per capita and water usage
```

```
##                PPPPC water_data
## PPPPC      1.0000000  0.1579074
## water_data 0.1579074  1.0000000
```

```
# Grouping by interval
  i <- 2
  while(i <= length(ppppc_water_interval))
  {
    ppppc_water <- ppppc_water %>%
    mutate(interval = case_when(( ppppc_water_interval[i-1] <= PPPPC & PPPPC <  ppppc_water_interval[i]
    i <- i + 1
  }

# Builds gdp per capita vs water usage scatter plot
ppppc_water %>%
    ggplot(aes(x = log(PPPPC),y = log(water_data))) + geom_point(color = "black") +
    stat_smooth(method = "lm", col = "red") +
    scale_y_continuous(limits = c(0,30)) +
  theme(axis.text.x = element_text(angle=45, hjust=0.8),
        panel.border = element_rect(color = "light gray",
                                    fill = NA,
                                    size = 1)) +
    theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
  labs(x = "Log of GDP per capita",
```

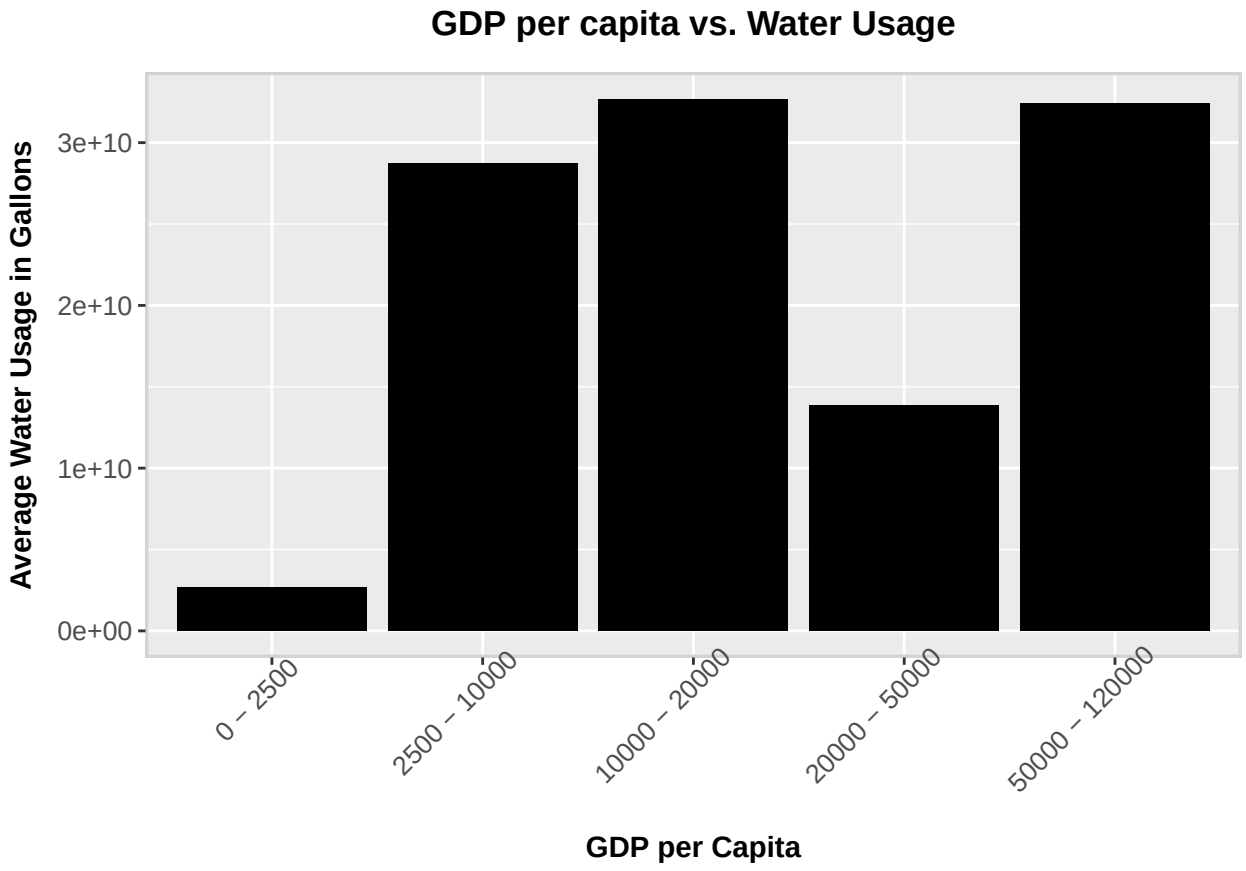
```
      y = "Log of Average Water Usage in Gallons",
      title = "GDP per capita vs. Water Usage") +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Builds gdp per capita vs water usage bar graph
 ppppc_water %>%
    group_by(interval) %>%
    summarise(ppppc_water_mean = mean(water_data)) %>%
    ggplot(aes(x = as.factor(interval),y = ppppc_water_mean)) + geom_bar(stat = "identity", fill = "Bla
  theme(axis.text.x = element_text(angle=45, hjust=0.8),
        panel.border = element_rect(color = "light gray",
                                    fill = NA,
                                    size = 1)) +
    theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
  scale_x_discrete(labels = c("0 - 2500",
                              "2500 - 10000",
                              "10000 - 20000",
                              "20000 - 50000",
```

```
                                "50000 - 120000")) +
  labs(x = "GDP per Capita",
       y = "Average Water Usage in Gallons",
       title = "GDP per capita vs. Water Usage") +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```



The covariance is close to 0 for this comparison showing that there is a very small correlation between gdp per capita and water usage. The graph also does not appear to be consistently increasing or increasing in any meaningful way. The scatter plot with the line of best fit it almost horizontal showing very little correlation. Therefore, gdp per capita is not strongly correlated with water usage.

Next we looked at gdp vs water usage. Does gdp affect water usage?

We did this by taking gdp from the money data set and water usage from the water worldometer dataset. First we looked at the covariance of each deciding it was best to look at it by taking the log of each variable. We created intervals to place countries into based on their gdp. We then took the average water usage of these intervals and graphed them on a bar graph. We took the log of water usage in this case since the original graph had an exponential curve and the smaller bars would otherwise be drowned out. A scatter plot was created by graphing each country by their gdp and their corresponding water usage. We took the log of each variable to help them fit on the graph. A line of best fit was also applied.

```
# Builds data set
  pppgdp_water <- full_join(PPPGDP_data,water_usage_data, by = "countries")  %>% drop_na() %>%
       mutate(interval = PPPGDP)
```

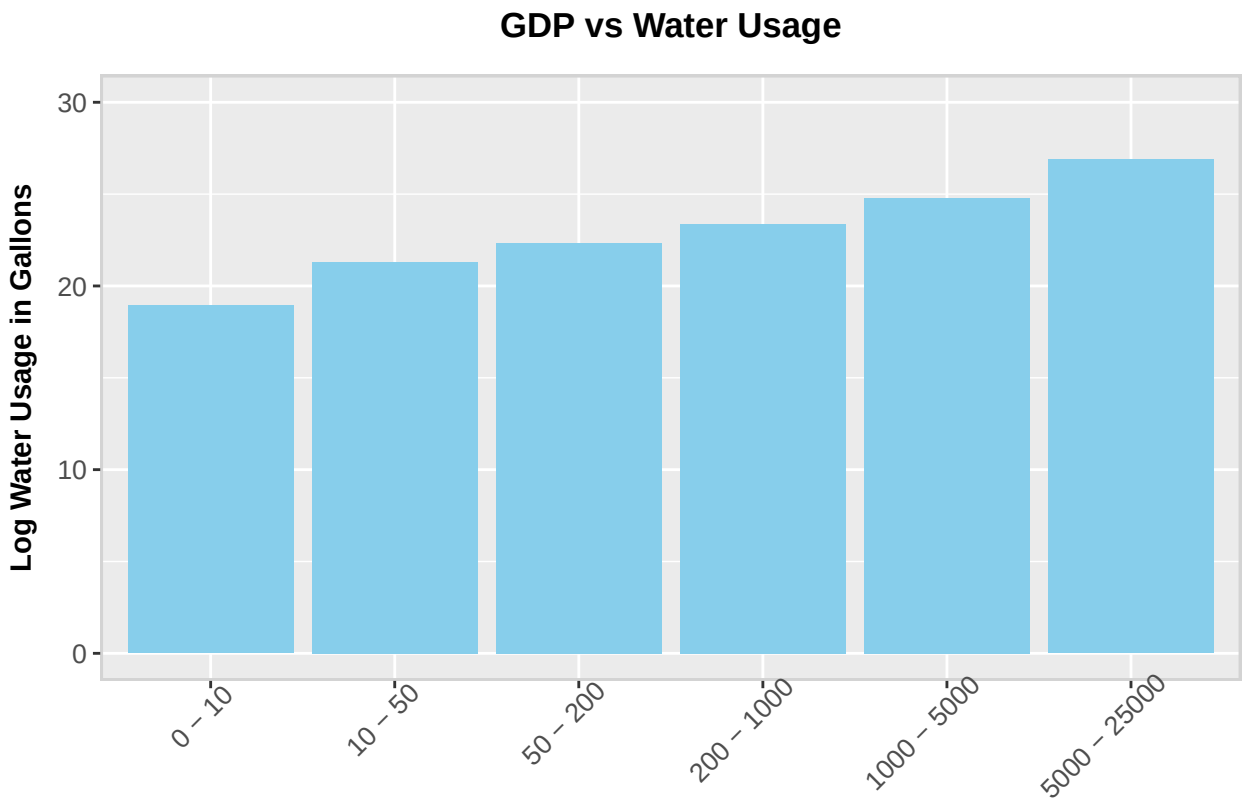
```
  intervals_pppgdp <- c(0,10,50,200,1000,5000,25000)

# Covariance
pppgdp_water %>%
  mutate(water_data = log(water_data)) %>%
  mutate(PPPGDP = log(PPPGDP)) %>%
  select(-countries,-interval) %>%
  cor()
```

```
##                PPPGDP water_data
## PPPGDP     1.0000000  0.8577835
## water_data 0.8577835  1.0000000
```

```
pppgdp_water %>%
  select(-countries,-interval) %>%
  cor() # Strong correlation between GDP and water usage
```

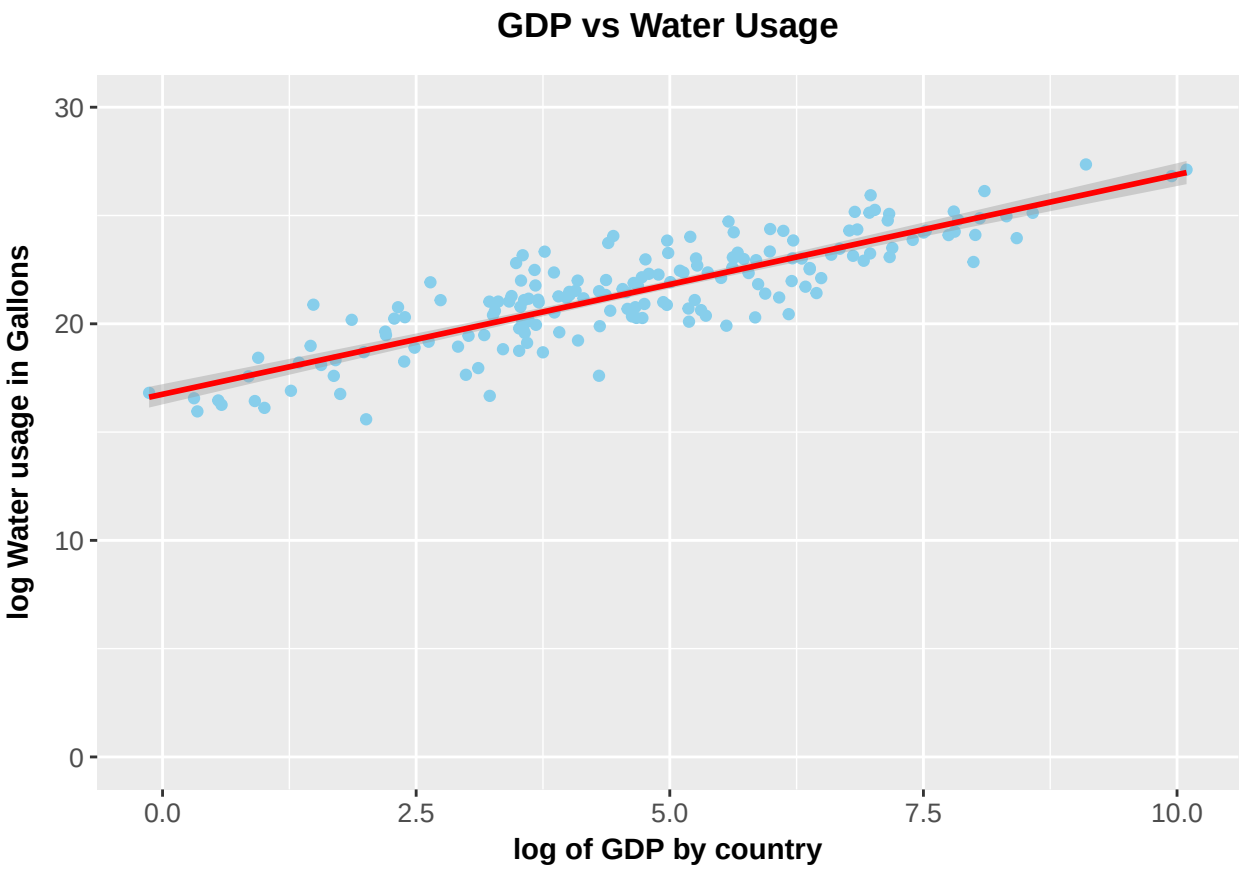
```
##                PPPGDP water_data
## PPPGDP     1.0000000  0.8293665
## water_data 0.8293665  1.0000000
```

```
# Creates interval
  i <- 2
  while(i <= length(intervals_pppgdp))
  {
    pppgdp_water <- pppgdp_water %>%
      mutate(interval = case_when((intervals_pppgdp[i-1] <= PPPGDP & PPPGDP < intervals_pppgdp[i]) ~ in
    i <- i + 1
  }

# Builds gdp vs water usage bar graph
 pppgdp_water %>%
    group_by(interval) %>%
    summarise(pppgdp_water_mean = mean(water_data)) %>%
    ggplot(aes(x = as.factor(interval),y = log(pppgdp_water_mean))) + geom_bar(stat = "identity",fill =
    theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
   scale_y_continuous(limits = c(0,30)) +
   labs(x = "GDP by Country in Millions of Dollars",
       y = "Log Water Usage in Gallons",
       title = "GDP vs Water Usage") +
   scale_x_discrete(labels = c("0 - 10",
                               "10 - 50",
                               "50 - 200",
                               "200 - 1000",
                               "1000 - 5000",
                               "5000 - 25000")) +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```




```
# Builds gdp vs water usage scatter plot
 pppgdp_water %>%
  ggplot(aes(x = log(PPPGDP),y = log(water_data))) + geom_point(color = "sky Blue") +
  stat_smooth(method = "lm", col = "red") +
  labs(x = "log of GDP by country",
       y = "log Water usage in Gallons",
       title = "GDP vs Water Usage") +
  scale_y_continuous(limits = c(0,30))  +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The covariance is close to 1 for this comparison showing that there is a very strong correlation between gdp and water water usage. The graph appears to be consistently increasing. As gdp increases, so does water usage. The scatter plot with the line of best fit is very closely aligned showing that there is a strong correlation. Therefore, gdp is strongly correlated with water usage.

Next we looked at population vs water usage. Does population affect water usage?

We did this by taking the population data from the money data set and water usage from the water worldometer dataset. First we looked at the covariance of each deciding it was best to look at it by taking the log of each variable. We created intervals to place countries into based on their population. We then took the average water usage of these intervals and graphed them on a bar graph. We took the log of water usage in this case since the original graph had an exponential curve and the smaller bars would otherwise be drowned out. A scatter plot was created by graphing each country by their population and their corresponding water usage. We took the log of each variable to help them fit on the graph. A line of best fit was also applied.

```
# Combines data set
pop_water <- full_join(population_data,water_usage_data) %>% drop_na() %>%
             mutate(water_capita = water_data/(Population*1000000)) %>%
             mutate(Population = (Population*1000000)) %>%
             mutate(interval = Population)
```

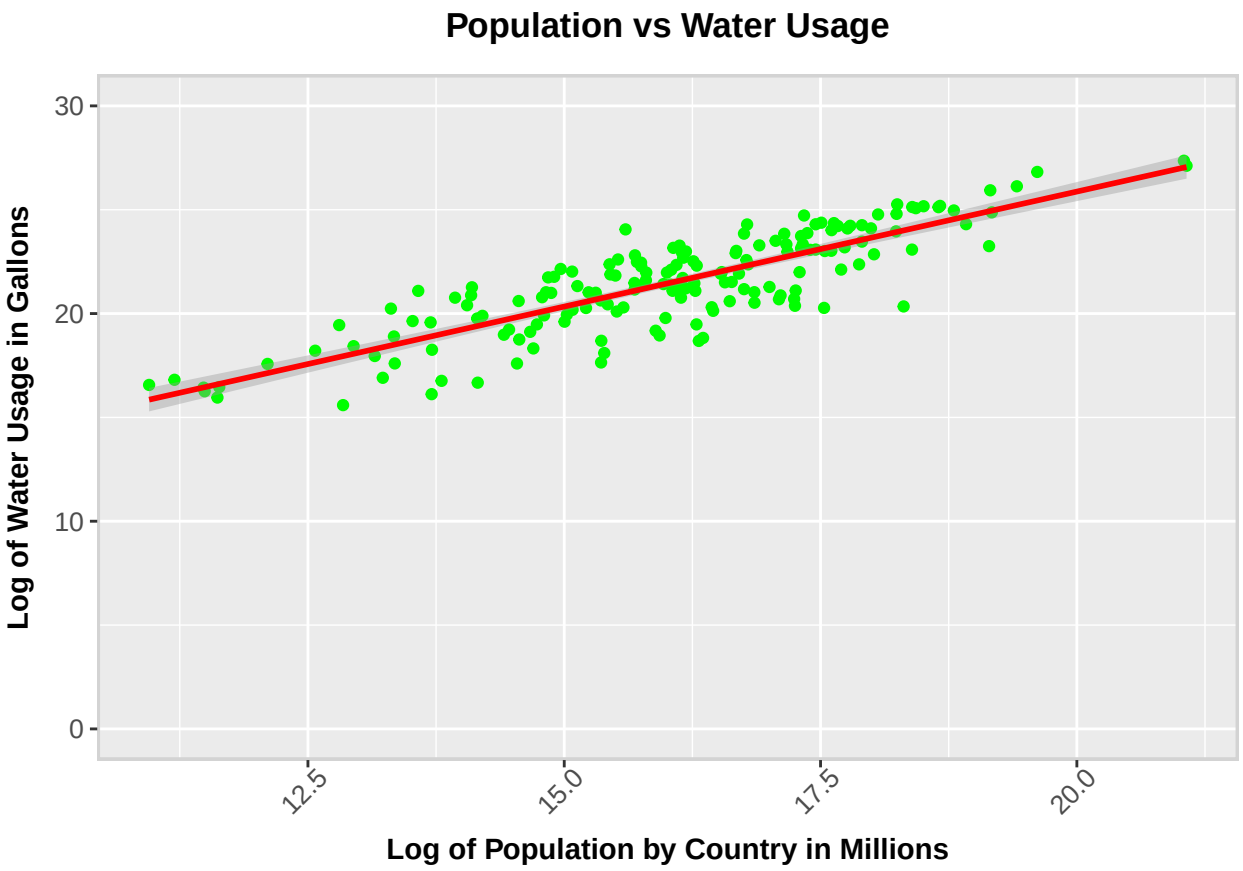
```
## Joining, by = "countries"
```

```
# Covariance
pop_water %>%
  mutate(water_data = log(water_data)) %>%
  mutate(Population = log(Population)) %>%
  select(-countries,-water_capita,-interval) %>%
  cor()
```

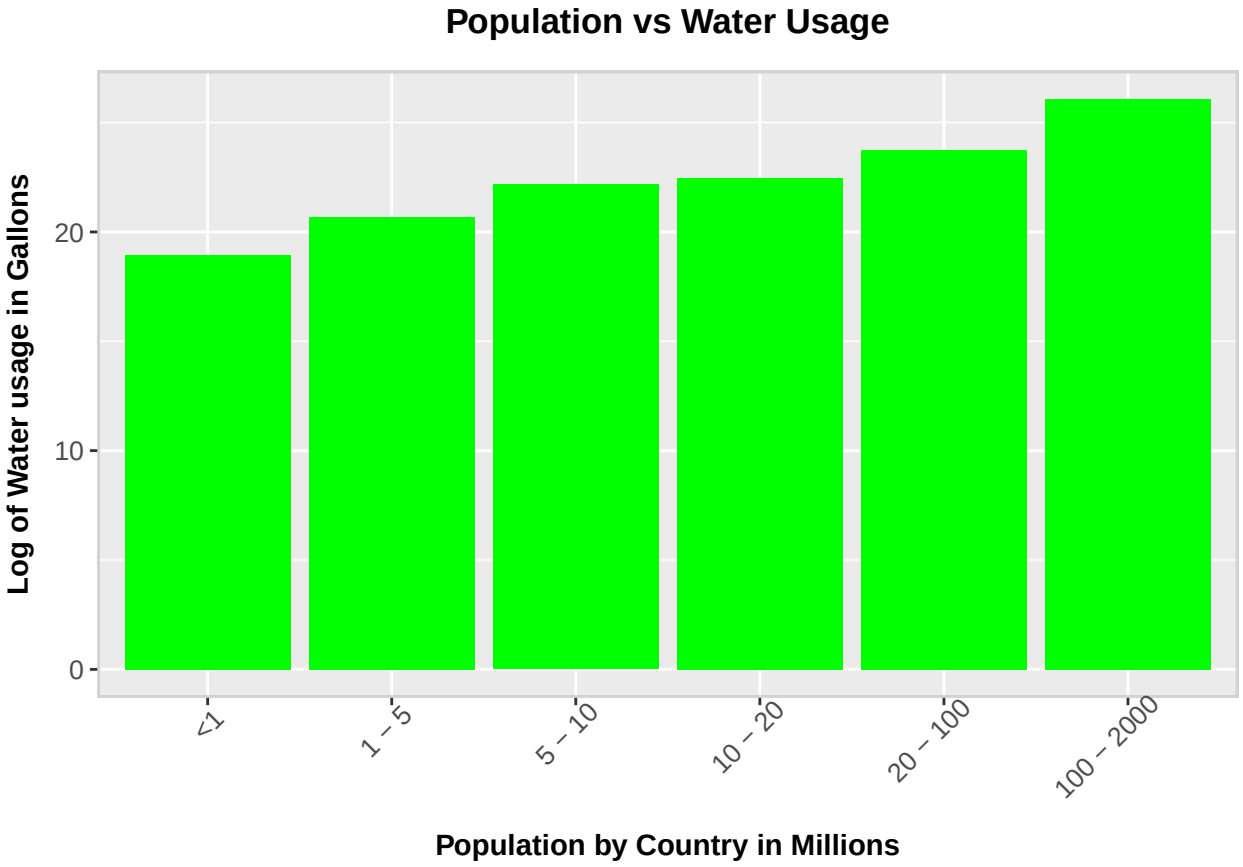
```
##            Population water_data
## Population  1.0000000  0.8479351
## water_data  0.8479351  1.0000000
```

```r
# Builds interval
pop_water_interval <- c(0,1000000,5000000,10000000,20000000,100000000,2000000000)

i <- 2
  while(i <= length(pop_water_interval))
  {
    pop_water <- pop_water %>%
    mutate(interval = case_when((pop_water_interval[i-1] <= Population & Population < pop_water_interval
    i <- i + 1
  }

# Builds population vs water scarcity scatter plot
pop_water %>%
  ggplot(aes(x = log(Population), log(water_data))) + geom_point(color = "Green") +
  stat_smooth(method = "lm", col = "red") +
  scale_y_continuous(limits = c(0,30))  +
  theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
  labs(x = "Log of Population by Country in Millions",
       y = "Log of Water Usage in Gallons",
       title = "Population vs Water Usage") +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Builds population vs water usage bar graph
 pop_water %>%
    group_by(interval) %>%
    summarise(mean_water = mean(water_data)) %>%
    ggplot(aes(x = as.factor(interval/1000000),y =log(mean_water))) + geom_bar(stat = "identity",fill =
    theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
  scale_x_discrete(labels = c("<1",
                              "1 - 5",
                              "5 - 10",
                              "10 - 20",
                              "20 - 100",
                              "100 - 2000")) +
  labs(x = "Population by Country in Millions",
       y = "Log of Water usage in Gallons",
       title = "Population vs Water Usage") +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```



The covariance is close to 1 for this comparison showing that there is a very strong correlation between population and water usage. The graph appears to be consistently increasing. As population increases, so does water usage. The scatter plot with the line of best fit is very closely aligned showing that there is a strong correlation. Therefore, population size is strongly correlated with water usage.

Next we looked at population vs water stress/scarcity. Does population affect water stress?

We did this by taking the population data from the money data set and water stress from the water stress dataset. First we looked at the covariance of each deciding it was best to look at it by taking the log of population and leaving water stress alone. We created intervals to place countries into based on their population. We then took the average water stress of these intervals and graphed the variables on a bar graph. The bar graph does not take the log of population, but either way the correlation is low. A scatter plot was created by graphing each country by their population and their corresponding water stress level. We took the log of population to help the dots fit on the scatter plot. A line of best fit was also applied.

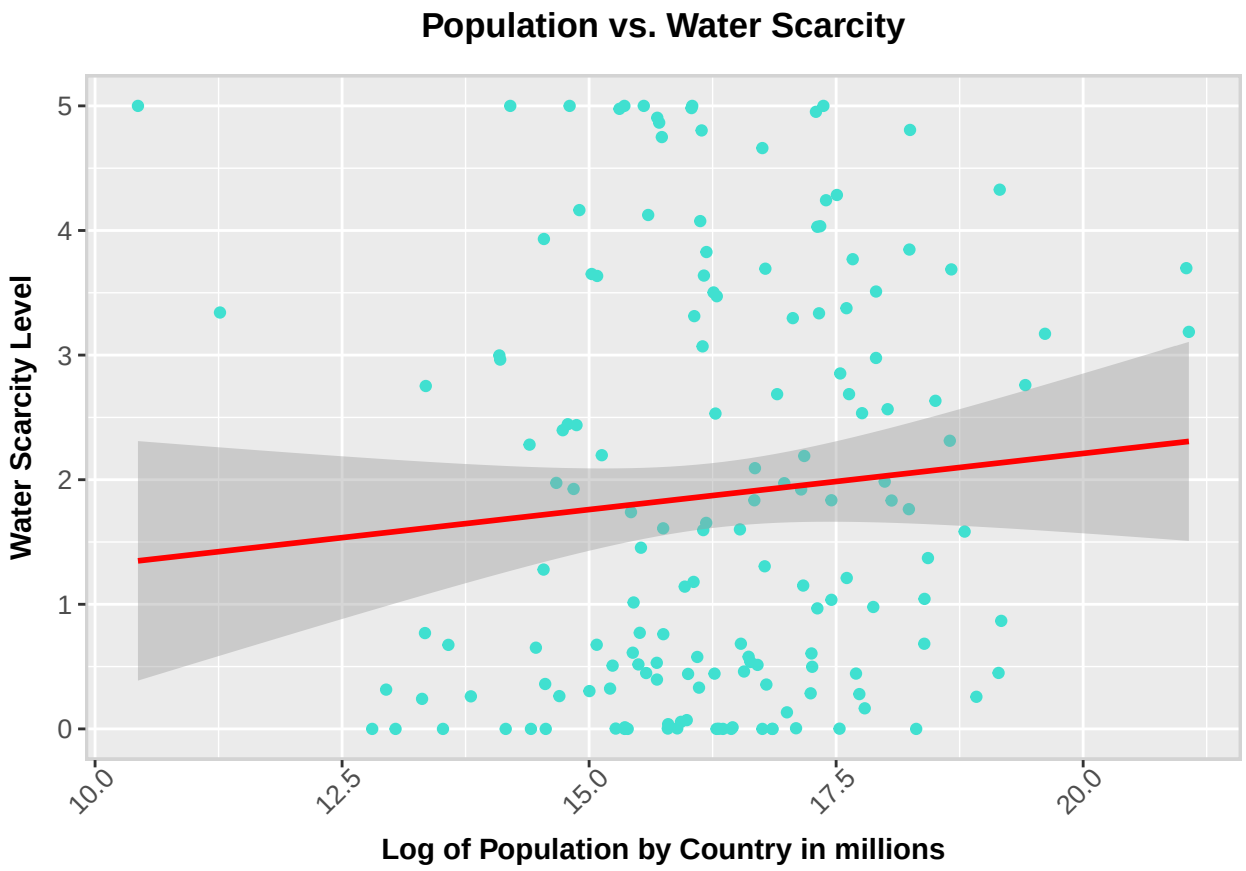
```
# Builds data set
pop_scarcity <- full_join(population_data,water_scarcity, by = "countries") %>% drop_na %>%
  mutate(Population = (Population*1000000)) %>%
  mutate(interval = Population)

# Covariance
pop_scarcity %>%
  mutate(Population = log(Population)) %>%
  select(Population,`All Sectors`) %>%
  cor()
```

```
##             Population All Sectors
## Population  1.00000000  0.08956672
```

```
## All Sectors 0.08956672  1.00000000
```

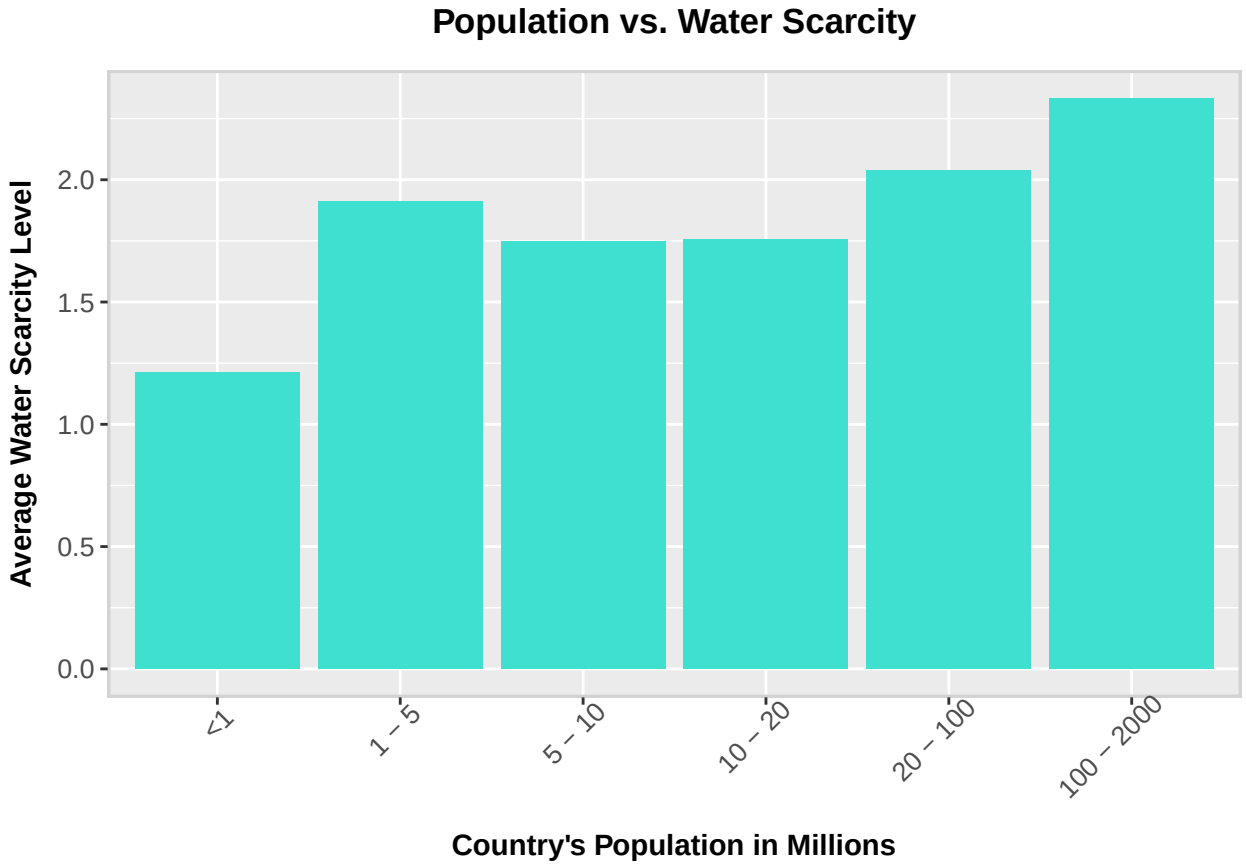
```
# Builds interval
pop_scarcity_interval <- c(0,1000000,5000000,10000000,20000000,100000000,2000000000)
i <- 2
  while(i <= length(pop_scarcity_interval))
  {
    pop_scarcity <- pop_scarcity %>%
    mutate(interval = case_when((pop_scarcity_interval[i-1] <= Population & Population < pop_scarcity_i
    i <- i + 1
  }

# Builds population vs scarcity scatter plot
pop_scarcity %>%
  ggplot(aes(x = log(Population), `All Sectors`)) + geom_point(color = "turquoise") +
  stat_smooth(method = "lm", col = "Red") +
  theme(axis.text.x = element_text(angle=45, hjust=0.8),
        panel.border = element_rect(color = "light gray",
                                    fill = NA,
                                    size = 1)) +
    theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
  labs(x = "Log of Population by Country in millions",
       y = "Water Scarcity Level",
       title = "Population vs. Water Scarcity") +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Builds population vs water scarcity bar graph
pop_scarcity %>%
  group_by(interval) %>%
  summarise(mean_sector = mean(`All Sectors`)) %>%
  ggplot(aes(as.factor(interval/1000000),mean_sector)) + geom_bar(stat = "identity",fill = "turquoise")+
  theme(axis.text.x = element_text(angle=45, hjust=0.8),
        panel.border = element_rect(color = "light gray",
                                    fill = NA,
                                    size = 1)) +
    theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
  scale_x_discrete(labels = c("<1",
                              "1 - 5",
                              "5 - 10",
                              "10 - 20",
                              "20 - 100",
                              "100 - 2000")) +
  labs(x = "Country's Population in Millions",
       y = "Average Water Scarcity Level",
       title = "Population vs. Water Scarcity") +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
```

```
        plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```



The covariance is close to 0 for this comparison showing that there is a very weak correlation between population and water scarcity. The graph appears to be slightly increasing, but it's not a strong connection. As population increases, water scarcity somewhat increases. The scatter plot with the line of best fit is not closely aligned showing that there is not a strong correlation. Therefore, population size is not strongly correlated with water usage. There is a chance this may be do to an error in our analysis or the fact a non-linear relationship exists but according to our data they are not strongly correlated.

Next we looked at gdp per capita vs water stress/scarcity. Does gdp per capita affect water stress?

We did this by taking the gdp per capita data from the money data set and water stress from the water stress dataset. First we looked at the covariance of each deciding it was best to look at it by taking the log of gdp per capita and leaving water stress alone. We created intervals to place countries into based on their gdp per capita. We then took the average water stress of these intervals and graphed the variables on a bar graph. The bar graph does not take the log of gdp per capita, but either way the correlation is moderate. A scatter plot was created by graphing each country by their gdp per capita and their corresponding water stress level. We took the log of gdp per capita to help the dots fit on the scatter plot. A line of best fit was also applied.

```
ppppc_scarcity <- full_join(PPPPC_data,water_scarcity, by = "countries") %>% drop_na %>%
  mutate(interval = PPPPC)

# Covariance
ppppc_scarcity %>%
  select(PPPPC,`All Sectors`) %>%
  cor()
```

```
##                PPPPC All Sectors
```

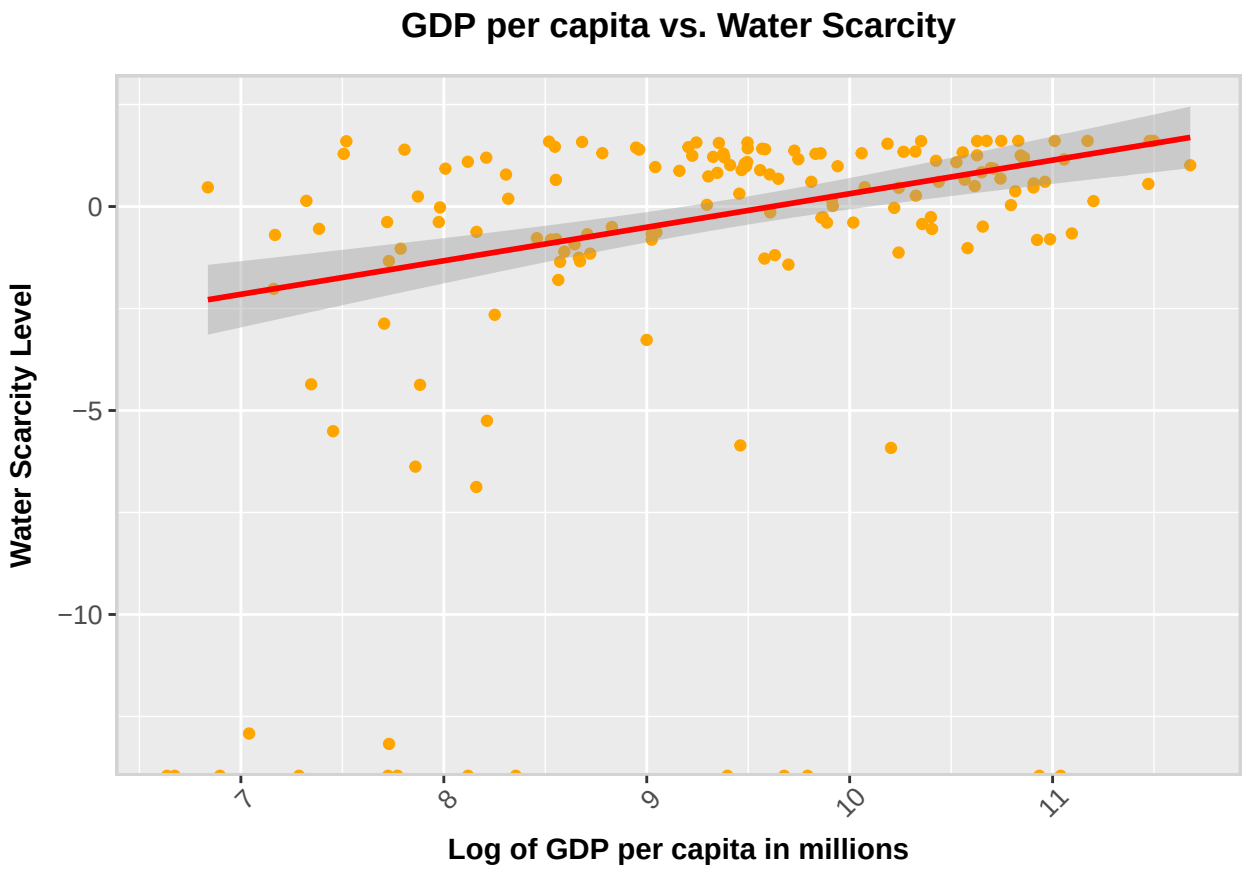
```
## PPPPC       1.0000000   0.2812989
## All Sectors 0.2812989   1.0000000
```

```
ppppc_scarcity %>%
  mutate(PPPPC = log(PPPPC)) %>%
  select(PPPPC,`All Sectors`) %>%
  cor()
```

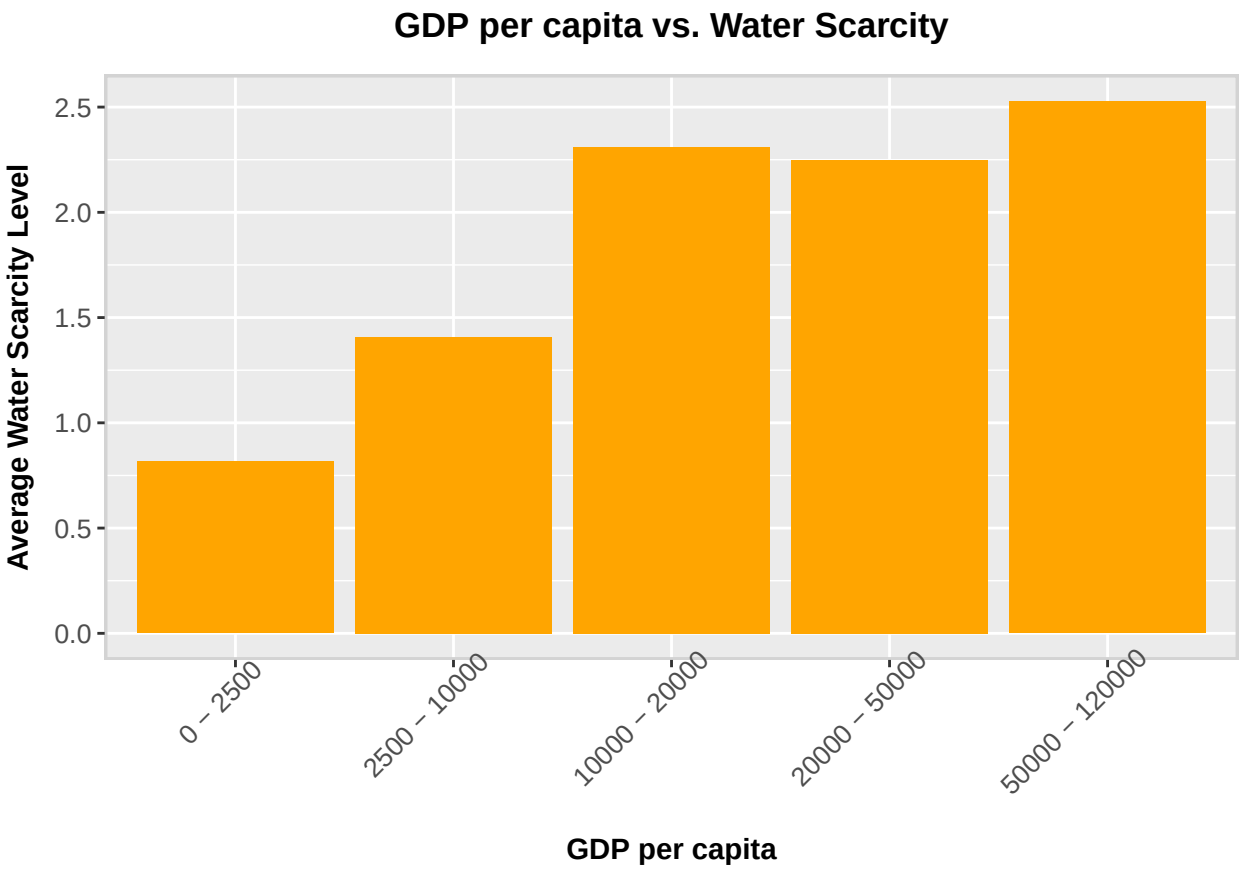
```
##                 PPPPC All Sectors
## PPPPC       1.0000000   0.3547221
## All Sectors 0.3547221   1.0000000
```

```
# Builds interval
ppppc_scarcity_interval <- c(0,2500,10000,20000,50000,120000)

i <- 2
  while(i <= length(ppppc_scarcity_interval))
  {
    ppppc_scarcity <- ppppc_scarcity %>%
    mutate(interval = case_when((ppppc_scarcity_interval[i-1] <= PPPPC & PPPPC < ppppc_scarcity_interval
    i <- i + 1
  }

# Builds GDP per capita vs water scarcity scatter plot
 ppppc_scarcity %>%
  ggplot(aes(x = log(PPPPC), log(`All Sectors`))) + geom_point(color = "orange") +
  stat_smooth(method = "lm", col = "red") +
  theme(axis.text.x = element_text(angle=45, hjust=0.8),
        panel.border = element_rect(color = "light gray",
                                    fill = NA,
                                    size = 1)) +
    theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
  labs(x = "Log of GDP per capita in millions",
       y = "Water Scarcity Level",
       title = "GDP per capita vs. Water Scarcity") +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Builds GDP per capita vs scarcity bar graph
ppppc_scarcity %>%
  group_by(interval) %>%
  summarise(mean_sector = mean(`All Sectors`)) %>%
  ggplot(aes(as.factor(interval),mean_sector)) + geom_bar(stat = "identity", fill = "orange") +
  theme(axis.text.x = element_text(angle=45, hjust=0.8),
        panel.border = element_rect(color = "light gray",
                                    fill = NA,
                                    size = 1)) +
    theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
  scale_x_discrete(labels = c("0 - 2500",
                              "2500 - 10000",
                              "10000 - 20000",
                              "20000 - 50000",
                              "50000 - 120000")) +
  labs(x = "GDP per capita",
       y = "Average Water Scarcity Level",
       title = "GDP per capita vs. Water Scarcity") +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```



The covariance is closer to 0 than 1, but it's there is still a bit of a correlation. So this comparison shows that there is a slight correlation between gdp per capita and water scarcity. The graph appears to be increasing, but not at a consistent rate. As gdp per capita increases, so does water scarcity but not consistently. The scatter plot with the line of best fit is not very closely aligned implying the correlation is not as strong as the graph would make it seem. Therefore, gdp per capita is somewhat correlated with water scarcity.

Next we looked at gdp vs water stress/scarcity. Does gdp affect water stress?

We did this by taking the gdp data from the money data set and water stress from the water stress dataset. First we looked at the covariance of each deciding it was best to look at it by taking the log of gdp while leaving water stress alone. We created intervals to place countries into based on their gdp. We then took the average water stress of these intervals and graphed the variables on a bar graph. The bar graph does not take the log of gdp, but either way the correlation is moderate. A scatter plot was created by graphing each country by their gdp and their corresponding water stress level. We took the log of gdp to help the dots fit on the scatter plot. A line of best fit was also applied.

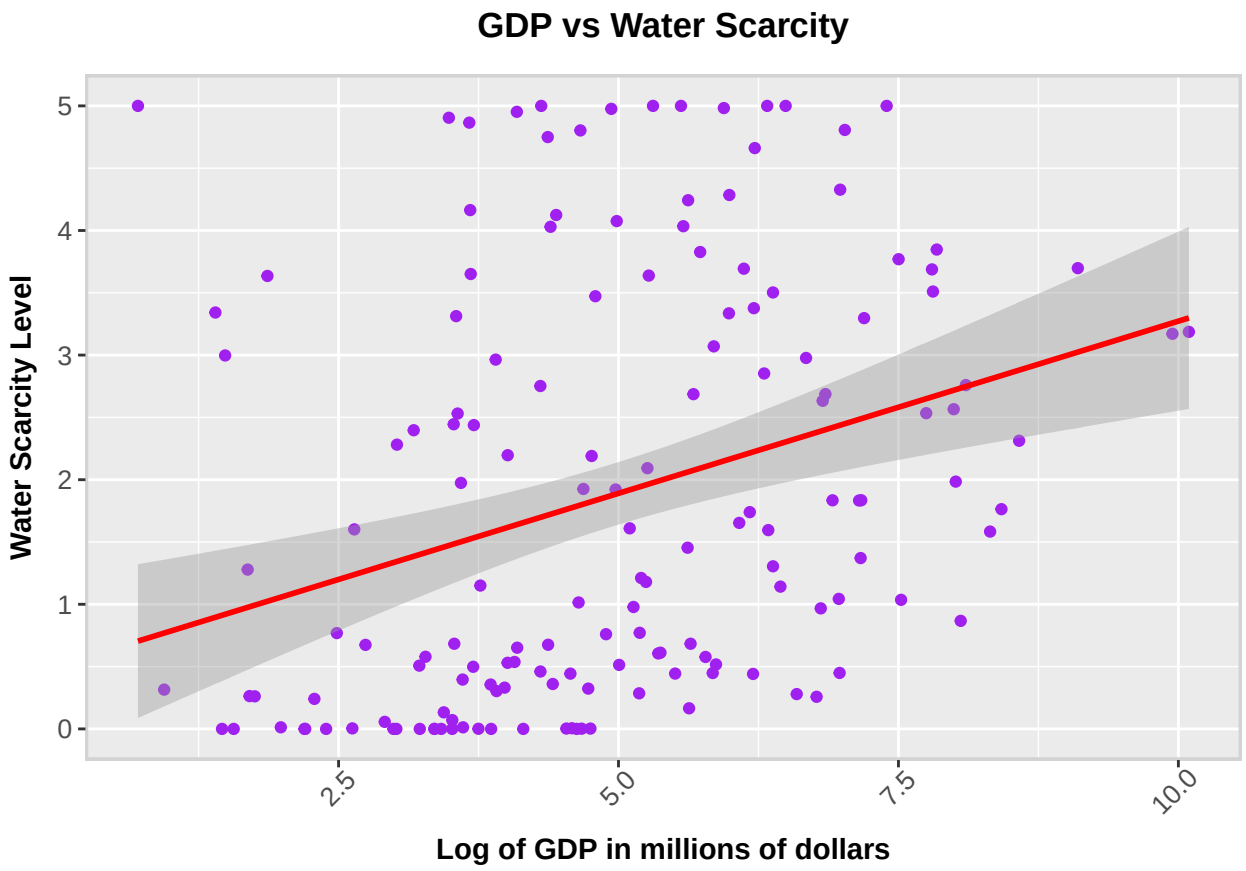
```
# Builds data set
pppgdp_scarcity <- full_join(PPPGDP_data,water_scarcity, by = "countries") %>% drop_na %>%
  mutate(interval = PPPGDP)

# Covariance
pppgdp_scarcity %>%
  mutate(PPPGDP = log(PPPGDP)) %>%
  select(PPPGDP,`All Sectors`) %>%
  cor()
```

```
##                PPPGDP All Sectors
## PPPGDP      1.0000000   0.3136735
```

```
## All Sectors 0.3136735   1.0000000
```

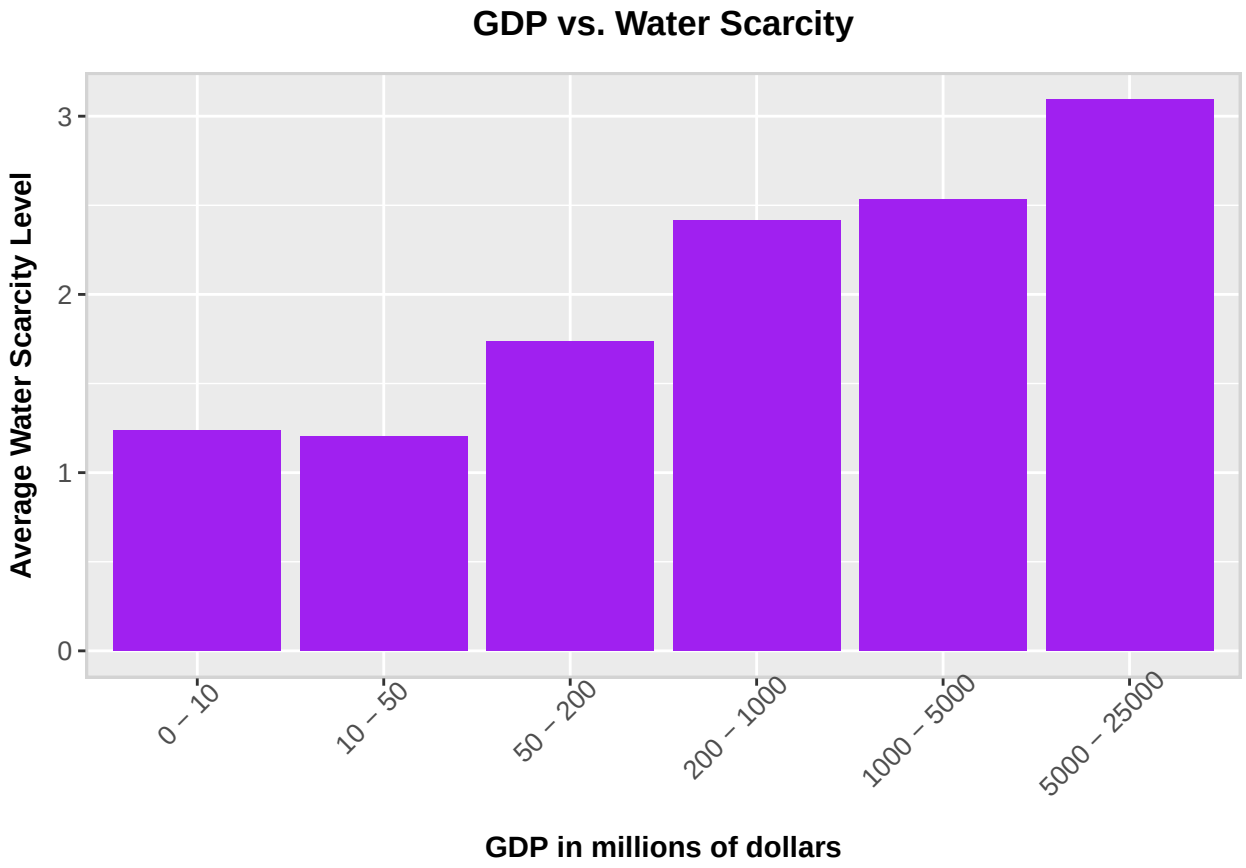
```
# Builds interval
pppgdp_scarcity_interval <-c(0,10,50,200,1000,5000,25000)
i <- 2
  while(i <= length(pppgdp_scarcity_interval))
  {
    pppgdp_scarcity <- pppgdp_scarcity %>%
    mutate(interval = case_when((pppgdp_scarcity_interval[i-1] <= PPPGDP & PPPGDP < pppgdp_scarcity_int
    i <- i + 1
  }

# Builds GDP vs water scarcity scatter plot
 pppgdp_scarcity %>%
  ggplot(aes(x = log(PPPGDP), `All Sectors`)) + geom_point(color = "purple") +
  stat_smooth(method = "lm", col = "red") +
  theme(axis.text.x = element_text(angle=45, hjust=0.8),
        panel.border = element_rect(color = "light gray",
                                    fill = NA,
                                    size = 1)) +
    theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
  labs(x = "Log of GDP in millions of dollars",
       y = "Water Scarcity Level",
       title = "GDP vs Water Scarcity") +
  theme( axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
         plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold")) # Higher population incr
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Builds GDP vs water scarcity bar graph
 pppgdp_scarcity %>%
  group_by(interval) %>%
  summarise(mean_sector = mean(`All Sectors`)) %>%
  ggplot(aes(as.factor(interval),mean_sector)) + geom_bar(stat = "identity", fill = "purple") +
  theme(axis.text.x = element_text(angle=45, hjust=0.8),
        panel.border = element_rect(color = "light gray",
                                    fill = NA,
                                    size = 1)) +
    theme(axis.text.x = element_text(angle=45, hjust=0.8),
          panel.border = element_rect(color = "light gray",
                                      fill = NA,
                                      size = 1)) +
  scale_x_discrete(labels = c("0 - 10",
                              "10 - 50",
                              "50 - 200",
                              "200 - 1000",
                              "1000 - 5000",
                              "5000 - 25000")) +
  labs(x = "GDP in millions of dollars",
       y = "Average Water Scarcity Level",
       title = "GDP vs. Water Scarcity") +
  theme(axis.text=element_text(size=10),
         axis.title=element_text(size=11,vjust = 0.5,face="bold"),
         axis.title.y=element_text(vjust = 3),
         axis.title.x=element_text(vjust = 0.5),
```

```
        plot.title=element_text(size=13,hjust = 0.5, vjust = 3, face="bold"))
```



The covariance is closer to 0 than 1, but it's there is still a bit of a correlation. So this comparison shows that there is a slight correlation between gdp and water scarcity. The graph appears to be increasing, but only at a slightly consistent rate. As gdp per capita increases, so does water scarcity. The scatter plot with the line of best fit is not very closely aligned implying the correlation is not as strong as the graph would make it seem. Therefore, gdp per capita is somewhat correlated with water scarcity. Maybe more then gdp per capita.

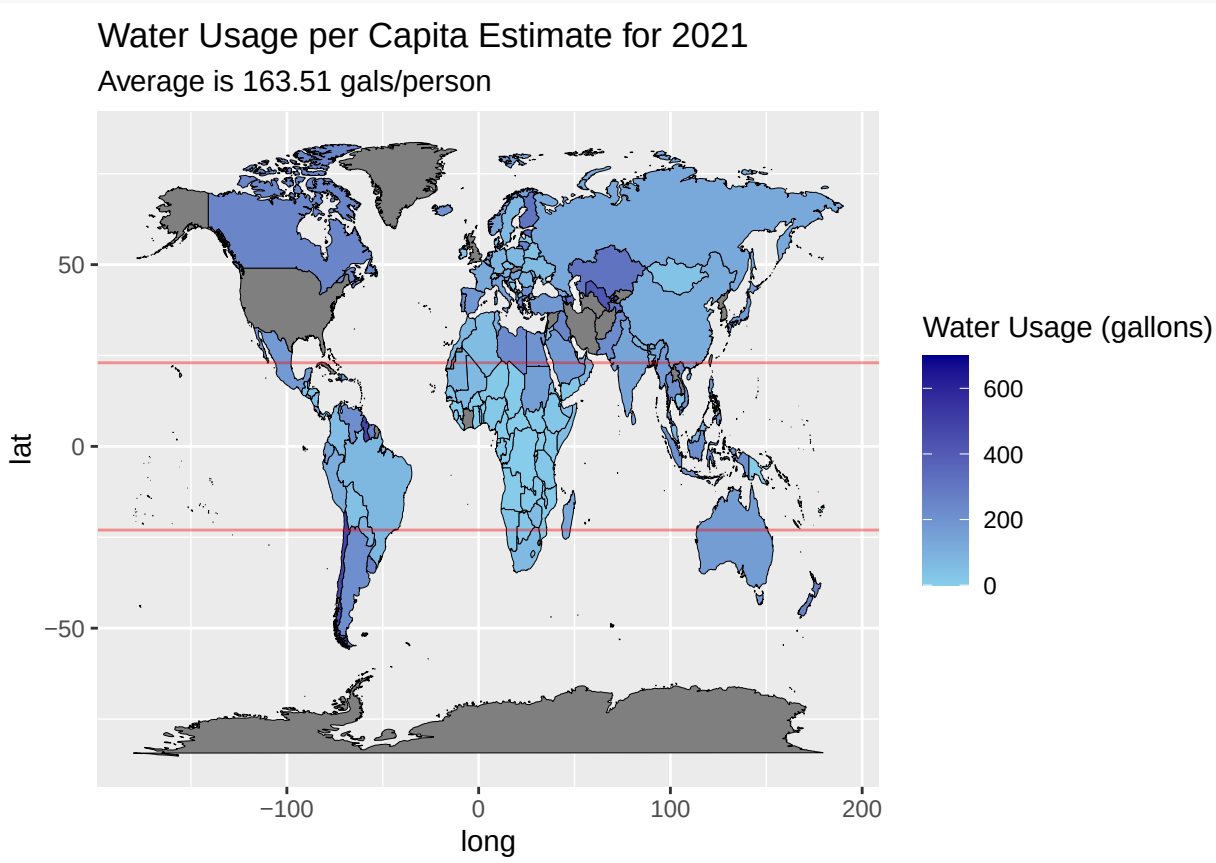
What if We Stopped Wasting Water?

From this point, it was clear that there were factors outside of the control of individuals that resulted in increased water usage. Our next question for the project was if the global usage of water were to stay the same, how long would it take for our situation to improve.

The first challenge with this was that in our data set that contained the populations from 1980 to 2020 and predictions from 2020 to 2026, didnt estimate the population of every countries and those that they skipped had an "n/a" put in their place. This created problems because for the math needed to do the predictions, NAs were being read as strings which returned errors instead of an NA output. I originally used the function replace_with_na_all(~.x=="n/a") thinking that this would solve the issue, but for some reason, this code ran indefinitely. Instead of replace just those strings, I decided to replace all the strings in the data set with NAs and then simple reassign the countries afterward. Though this was a bit more tedious, it did work. From here, we added a new column to data set for each here we wanted an estimate for.

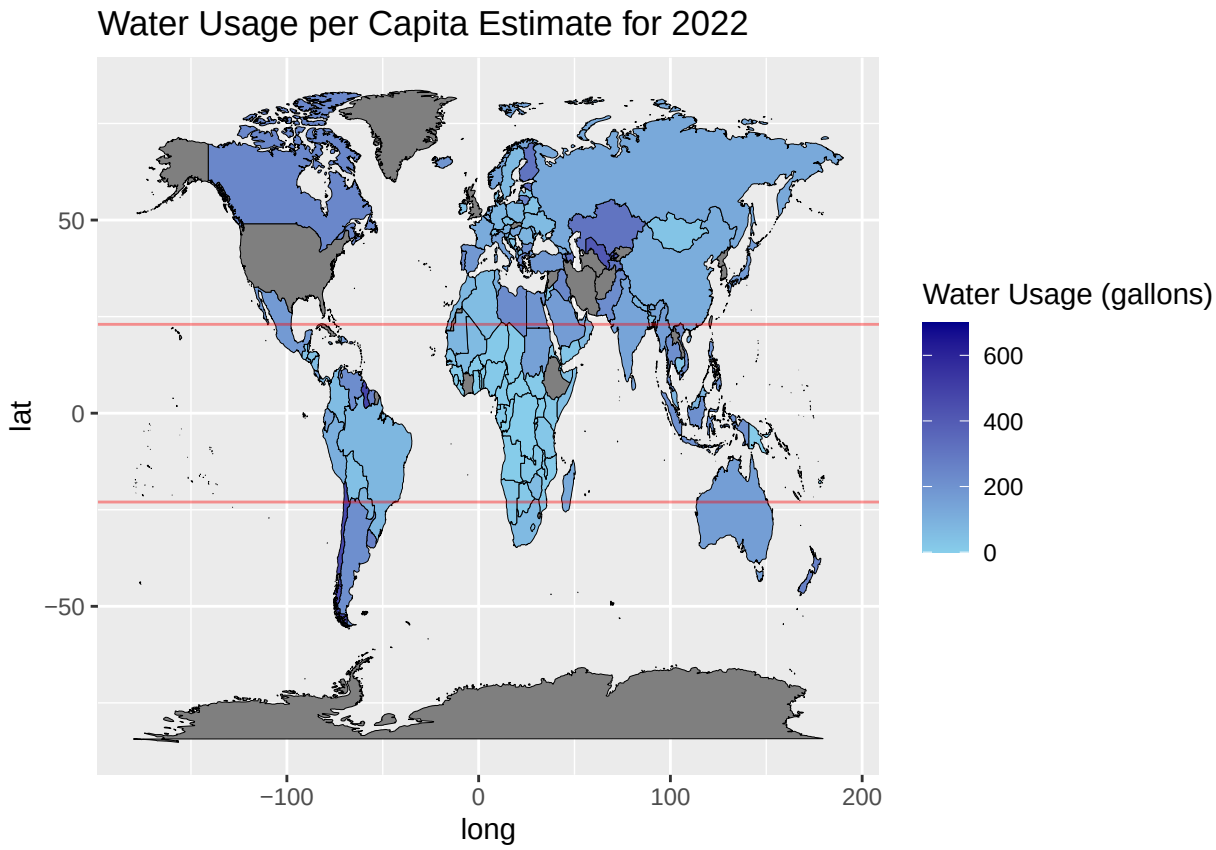
```
#Making Prediction Graphs

test <- data.frame(lapply(pop_data, function(x) as.numeric(as.character(x)))) %>%
  mutate(est_1980=(water_usage/(X1980*1000000))) %>%
  mutate(est_1990=(water_usage/(X1990*1000000))) %>%
```

```
  mutate(est_2000=(water_usage/(X2000*1000000))) %>%
  mutate(est_2010=(water_usage/(X2010*1000000))) %>%
  mutate(est_2021=(water_usage/(X2021*1000000))) %>%
  mutate(est_2022=(water_usage/(X2022*1000000))) %>%
  mutate(est_2023=(water_usage/(X2023*1000000))) %>%
  mutate(est_2024=(water_usage/(X2024*1000000))) %>%
  mutate(est_2025=(water_usage/(X2025*1000000))) %>%
  mutate(est_2026=(water_usage/(X2026*1000000))) %>%
  mutate(region=pop_data$region)

#2021 Estimate
avg_2021<-round(mean(test$est_2021, na.rm = T), 2)
ggplot()+
  geom_map(data= test, map = world,
aes(long, lat, map_id = region, fill = est_2021), color = "black", size = 0.1)+
  scale_fill_continuous(low = "skyblue", high = "darkblue",
                        name= "Water Usage (gallons)",
                        limit=c(0,700))+
  geom_hline(yintercept = c(-23, 23), color= "red", alpha = .4)+
  labs(title = "Water Usage per Capita Estimate for 2021", subtitle = glue("Average is {avg_2021} gals/
```

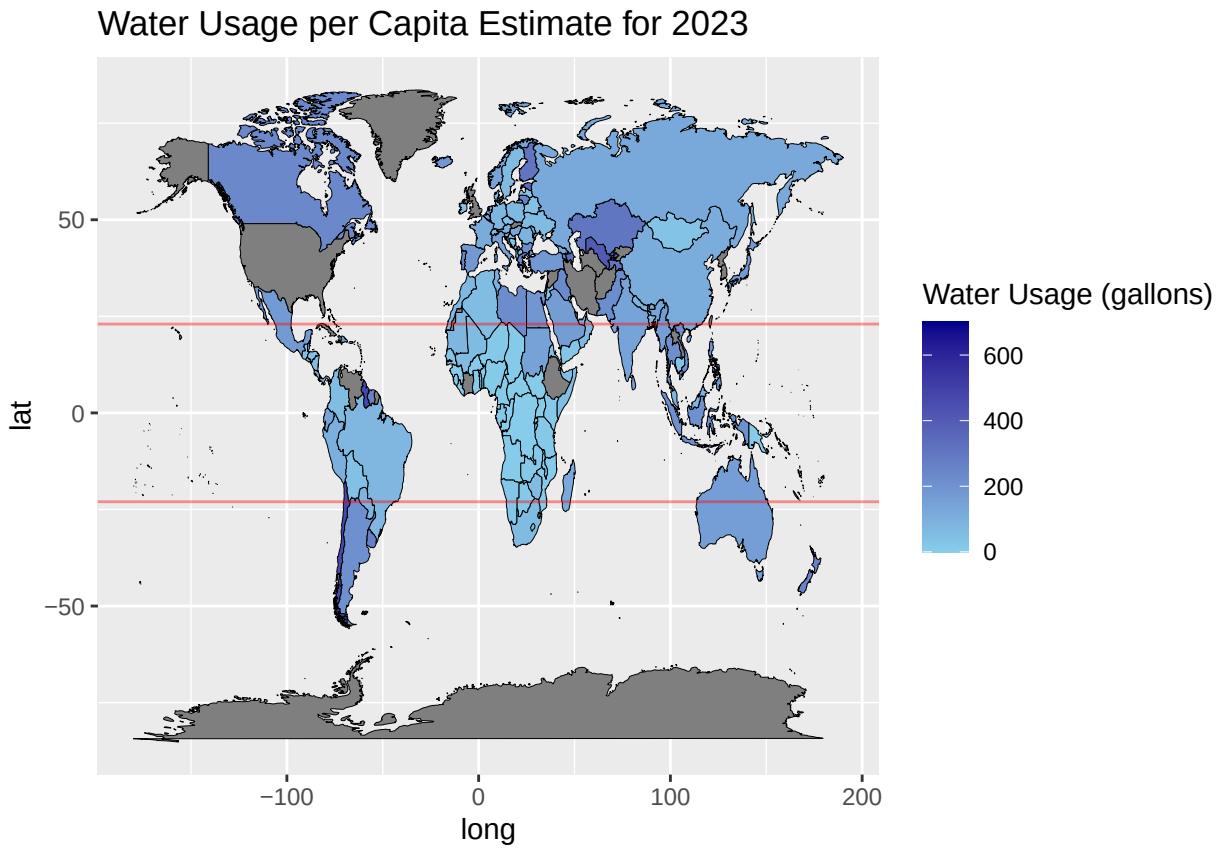



```
#2022 Estimate
ggplot()+
  geom_map(data= test, map = world,
aes(long, lat, map_id = region, fill = est_2022), color = "black", size = 0.1)+
  scale_fill_continuous(low = "skyblue", high = "darkblue",
                        name= "Water Usage (gallons)",
```

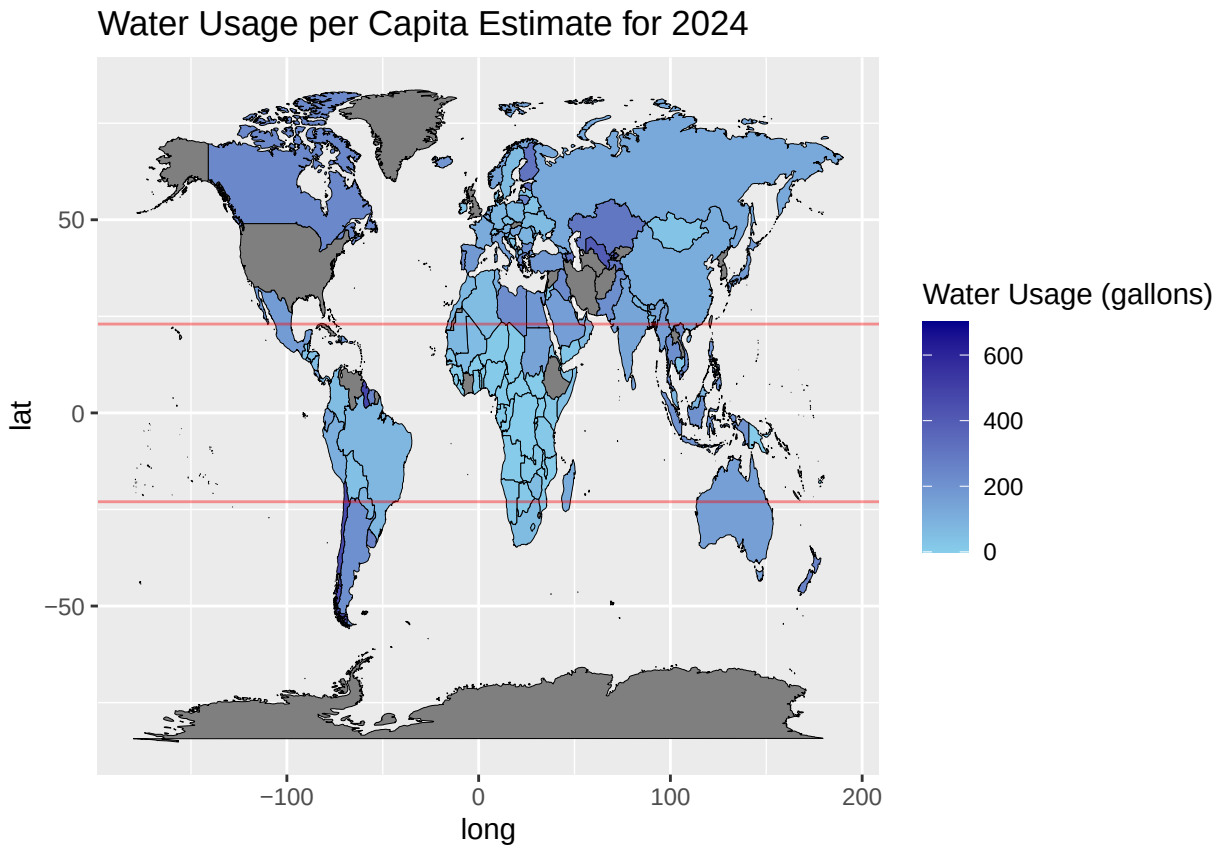
```
                        limit=c(0,700))+
  geom_hline(yintercept = c(-23, 23), color= "red", alpha = .4)+
  labs(title = "Water Usage per Capita Estimate for 2022")
```



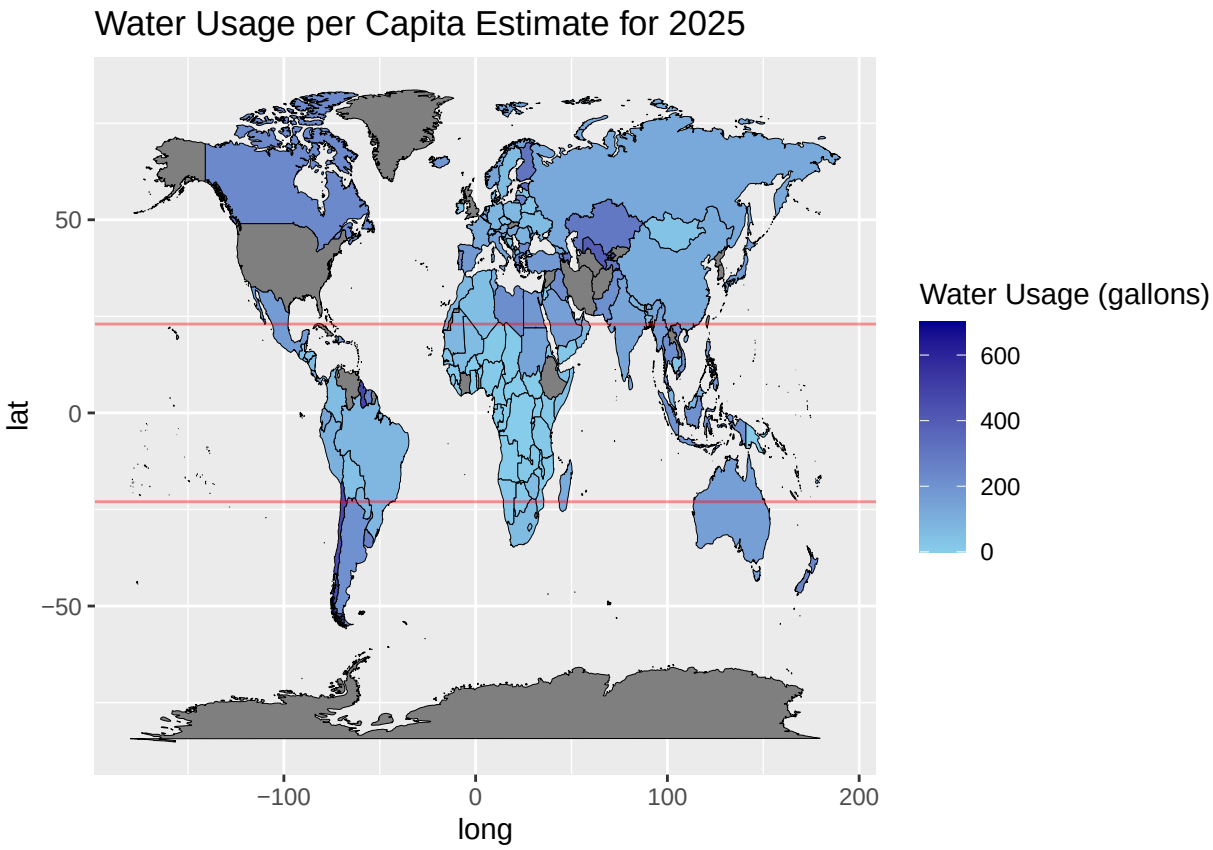

```
#2023 Estimate
ggplot()+
  geom_map(data= test, map = world,
aes(long, lat, map_id = region, fill = est_2023), color = "black", size = 0.1)+
  scale_fill_continuous(low = "skyblue", high = "darkblue",
                        name= "Water Usage (gallons)",
                        limit=c(0,700))+
  geom_hline(yintercept = c(-23, 23), color= "red", alpha = .4)+
  labs(title = "Water Usage per Capita Estimate for 2023")
```



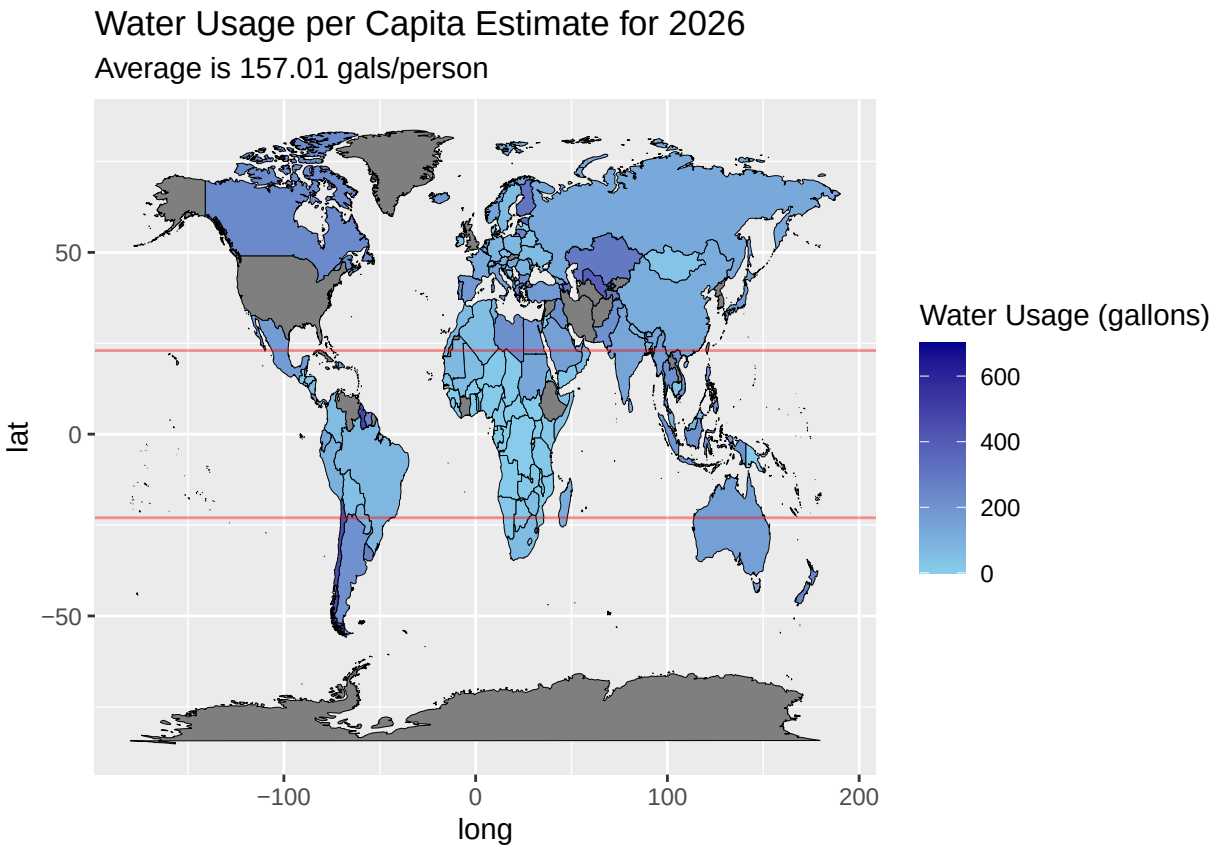

```
#2024 Estimate
ggplot()+
  geom_map(data= test, map = world,
aes(long, lat, map_id = region, fill = est_2024), color = "black", size = 0.1)+
  scale_fill_continuous(low = "skyblue", high = "darkblue",
                        name= "Water Usage (gallons)",
                        limit=c(0,700))+
  geom_hline(yintercept = c(-23, 23), color= "red", alpha = .4)+
  labs(title = "Water Usage per Capita Estimate for 2024")
```




```
#2025 Estimate
ggplot()+
  geom_map(data= test, map = world,
aes(long, lat, map_id = region, fill = est_2025), color = "black", size = 0.1)+
  scale_fill_continuous(low = "skyblue", high = "darkblue",
                        name= "Water Usage (gallons)",
                        limit=c(0,700))+
  geom_hline(yintercept = c(-23, 23), color= "red", alpha = .4)+
  labs(title = "Water Usage per Capita Estimate for 2025")
```




```
#2026 Estimate
avg_2026<-round(mean(test$est_2026, na.rm = T), 2)
ggplot()+
  geom_map(data= test, map = world,
aes(long, lat, map_id = region, fill = est_2026), color = "black", size = 0.1)+
  scale_fill_continuous(low = "skyblue", high = "darkblue",
                        name= "Water Usage (gallons)",
                        limit=c(0,700))+
  geom_hline(yintercept = c(-23, 23), color= "red", alpha = .4)+
  labs(title = "Water Usage per Capita Estimate for 2026", subtitle = glue("Average is {avg_2026} gals/
```



It is very clear that if the global usage of water could stay consistent for the next 4 years, we would be able to bring our average global water usage per capita to just over 157 gallons per person. As a point of reference, several studies we found showed that a person could comfortably live off of 101.5 gallons a day. This 101.5 includes things like being able to drink a gallon of water, take two ten minute shows and a bath, and several other things all in one day.

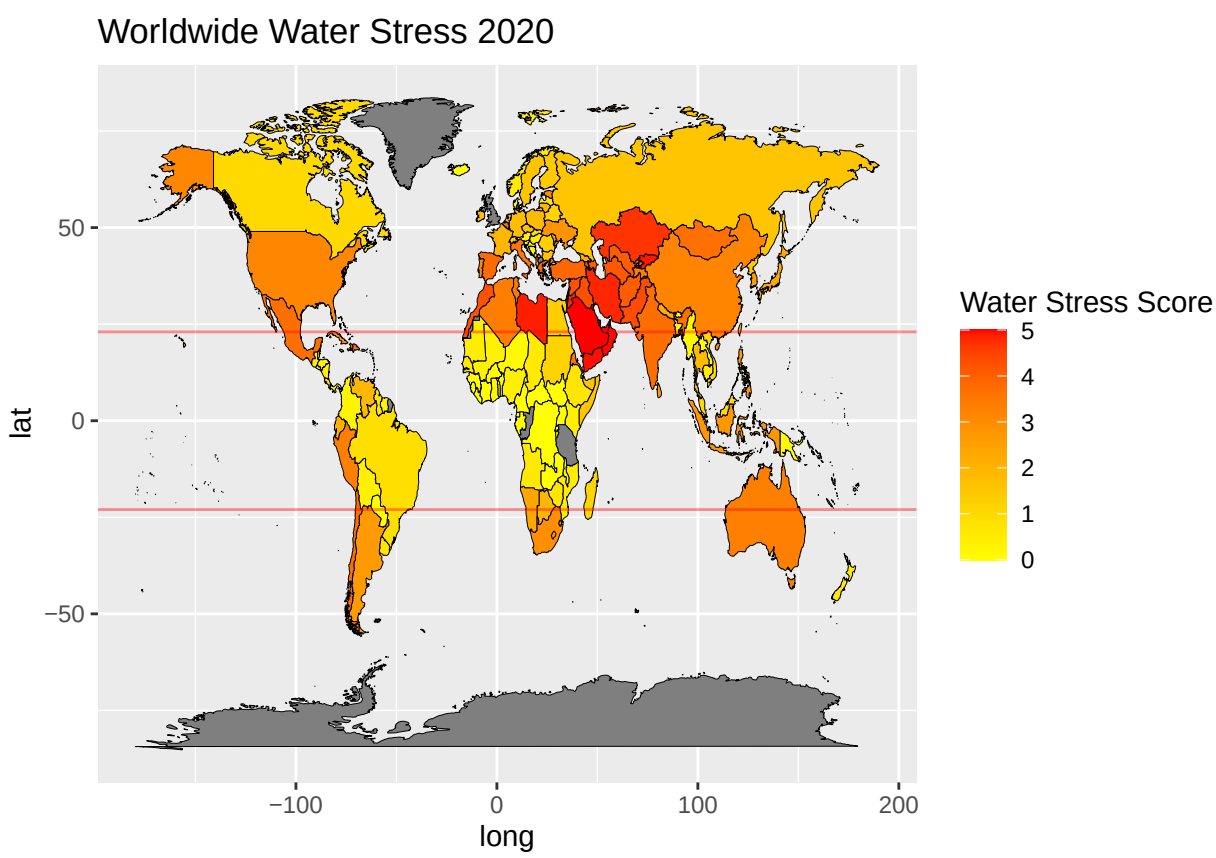
Water Stress:

In discovering this, we realized that the sum of freshwater used, according to our data, accounted for less than 1% of the Earth's total freshwater reserves, which seemed abnormal. This would suggest that is no need for concern for our global water usage at all. This conclusion seemed to contradict the majority of professional opinions which made us consider using a different measurement for our relationship with water instead of only usage. Water stress is commonly accepted to occur when "the demand for water exceeds the available amount during a certain period or when poor quality restricts its use". We wanted to see how much stress on water the world's usage habits create.

The water stress score scale gives a country a score from 0 to 5 based on the percentage of its total annual water withdrawals compared to the total annual amount of fresh water that country can access. Freshwater "withdrawal" refers to water that is either temporary or permanently altered after it is used, meaning water that is considered "withdrawn" is not considered "fresh" after it completes the water cycle because the water itself has become chemically altered and extra treatment is required to return it to its natural state. An example of this is water that is used for fracking. Even if that water is boiled and condenses, it would not be safe for humans because it has undergone chemical alterations that aren't safe.

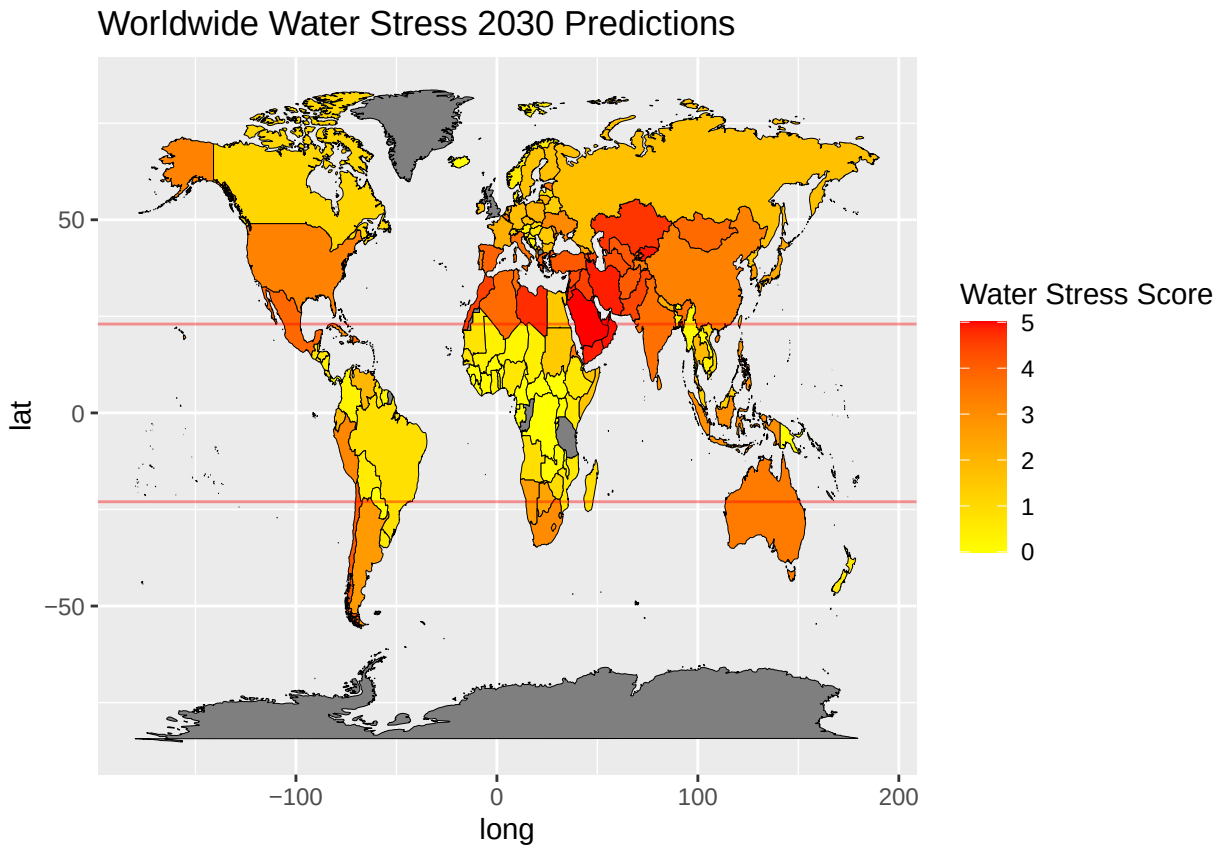
If withdraws account for less than 10%, they receive something from 0-1, 10%-20% receive something from 1-2, 20%-40% receives something from 2-3, 40%-80% receive something from 3-4, 80%+ receive something from 4-5. This graph makes it very clear that by 2040, easily over half the world will be lacking fresh water and that for most of them, somewhere between 20% and 40% of the water they are using will require additional treatment before it is beneficial to humans again.

```r
#Showing how close we are to using the Earth's total amount of freshwater

water_stress <- water_stress_data_default %>%
  rename(region=Name) %>%
  full_join(world, water_stress, by= "region") %>%
  mutate(region = case_when(region == "United States of America" ~
                              "USA", TRUE ~ region)) %>%
  mutate(region = case_when(region == "Czech Republic (Czechia)" ~
                              "Czech Republic", TRUE ~ region)) %>%
  mutate(region = case_when(region == "Serbia" ~ "Republic of Serbia",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "DR Congo" ~
                      "Democratic Republic of the Congo", TRUE ~ region)) %>%
  mutate(region = case_when(region == "Guinea-Bissau" ~
                              "Guinea Bissau", TRUE ~ region)) %>%
  mutate(region = case_when(region == "Congo" ~
                              "Republic of Congo", TRUE ~ region))

ggplot()+
  geom_map(data= water_stress, map = world,
  aes(long, lat, map_id = region, fill = water_stress$`All Sectors`),
  color = "black", size = 0.1)+
  scale_fill_continuous(low = "yellow", high = "Red",
                        name= "Water Stress Score")+
  geom_hline(yintercept = c(-23, 23), color= "red", alpha = .4)+
  labs(title = "Worldwide Water Stress 2020")
```

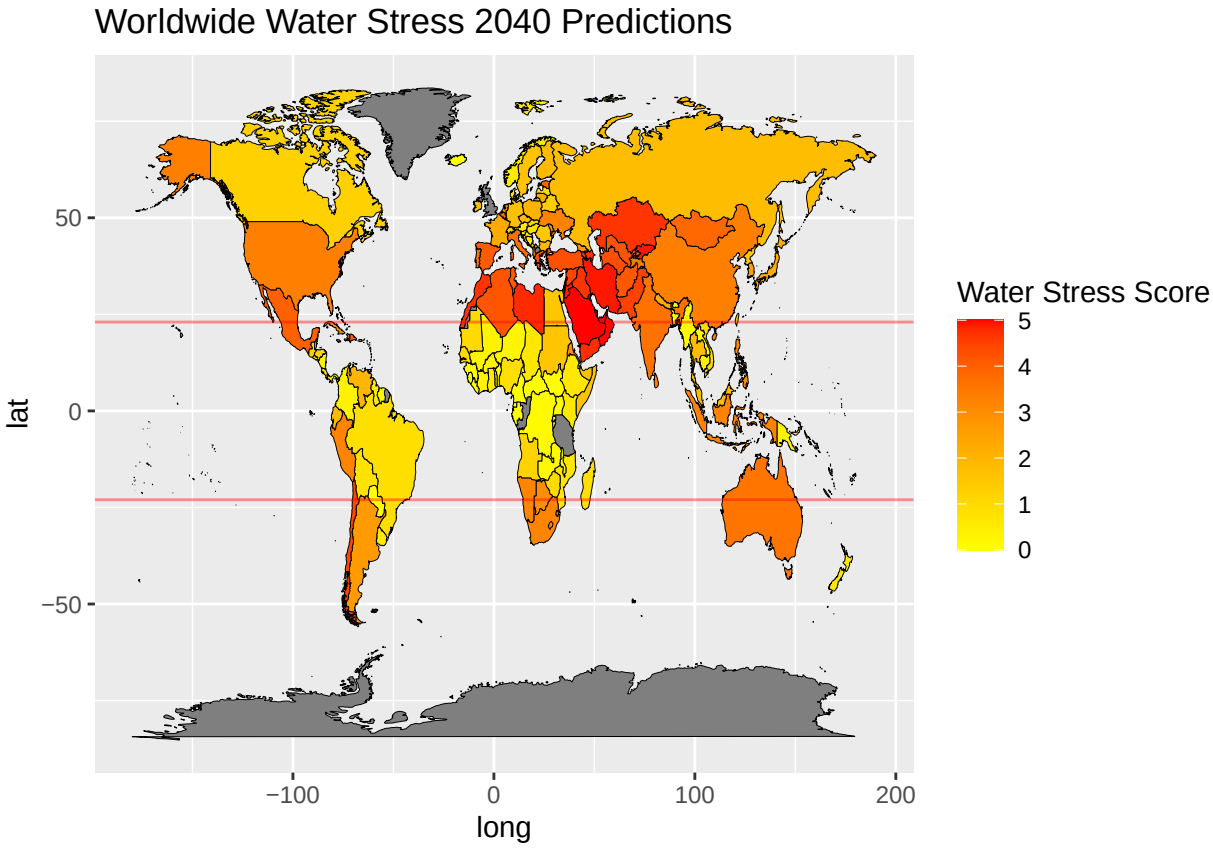


This seemed much more like what we were expecting. The first world countries and Middle East have the highest water stress and those that are economically and technologically underdeveloped had the smaller Water Stress Scores. One thing that was a bit shocking to us what the Central Africa seems to experience some of the least severe water stress on Earth which felt weird. Just like before, we wanted to project what things would look like for the future if we do not change our current water treatment habits.

```
#Water Stress Predictions
bau_2030 <- read_excel("aqueduct-water-stress-country-rankings-data-set.xlsx",
    sheet = "2030 BAU") %>%
  rename(region=Name) %>%
  full_join(world, bau_2030, by= "region") %>%
  mutate(region = case_when(region == "United States of America" ~ "USA",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "Czech Republic (Czechia)" ~
                              "Czech Republic", TRUE ~ region)) %>%
  mutate(region = case_when(region == "Serbia" ~ "Republic of Serbia",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "DR Congo" ~
                              "Democratic Republic of the Congo",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "Guinea-Bissau" ~ "Guinea Bissau",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "Congo" ~ "Republic of Congo",
                            TRUE ~ region))

ggplot()+
  geom_map(data= bau_2030, map = world,
  aes(long, lat, map_id = region, fill = bau_2030$`All Sectors`),
  color = "black", size = 0.1)+
  scale_fill_continuous(low = "yellow", high = "Red",
                        name= "Water Stress Score")+
  geom_hline(yintercept = c(-23, 23), color= "red", alpha = .4)+
  labs(title = "Worldwide Water Stress 2030 Predictions")
```




```
bau_2040 <- read_excel("aqueduct-water-stress-country-rankings-data-set.xlsx",
    sheet = "2040 BAU") %>%
  rename(region=Name) %>%
  full_join(world, bau_2040, by= "region") %>%
  mutate(region = case_when(region == "United States of America" ~
                              "USA", TRUE ~ region)) %>%
  mutate(region = case_when(region == "Czech Republic (Czechia)" ~
                              "Czech Republic", TRUE ~ region)) %>%
  mutate(region = case_when(region == "Serbia" ~
                              "Republic of Serbia", TRUE ~ region)) %>%
  mutate(region = case_when(region == "DR Congo" ~
                              "Democratic Republic of the Congo",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "Guinea-Bissau" ~ "Guinea Bissau",
                            TRUE ~ region)) %>%
  mutate(region = case_when(region == "Congo" ~ "Republic of Congo",
                            TRUE ~ region))

ggplot()+
  geom_map(data= bau_2040, map = world,
  aes(long, lat, map_id = region, fill = bau_2040$`All Sectors`),
  color = "black", size = 0.1)+
  scale_fill_continuous(low = "yellow", high = "Red",
                        name= "Water Stress Score")+
  geom_hline(yintercept = c(-23, 23), color= "red", alpha = .4)+
  labs(title = "Worldwide Water Stress 2040 Predictions")
```



Now with using water stress as our measurement for the world's relationship with freshwater, we still see that there is very little change in the amount of water stress that exists if no major changes are made to how much we use water, but this graph does make the severity of the situation a bit more clear.

Limitations

We only had one water usage data set instead of one with multiple years. More analyses could be done if we had information from year to year. We were not able to perform as many predictive analyses for the future concerning water data and stress as we were hoping. We could have done this if we had more time and research. Our data primarily examines a linear relationship. It is possible that strong correlations can be found using non-linear forms of analysis. Differences in how we analysed our data versus how someone else would go about it could also reveal different conclusions such as not looking at the data using a log function or always using a function. Sometimes we used it and other times we felt it was best not to.

Conclusions

The two biggest factors contributing to increased water usage are increases in population size and total economic output. Water stress was more difficult to correlate, but GDP and GDP per capita had some correlation with water stress. Therefore the primary factor in our research for increasing water usage and scarcity is increases in total economic output. As economic and industrial demands increase, so does the demand for water.

Concerning future water availability, the data does do not spell certain doom for all life on Earth, but it does show us that if we aren't careful, by 2040 we could be in a place where water is scarce and nearly half the world will not be able to meet water demands. Additionally in our geographic analysis, it is important to note that temperature is not a major factor affecting water scarcity.

Solutions

Since it would be unreasonable to halt economic progress, solutions should come in the form of minimizing

water usage or increasing the available supply of water. One example would be regulating the amount of water that can be used in industry if and where possible. Another is making appliances that require water more efficient such as clothes and dish washers. Ways to increase water availability is by purifying ocean water or by building fresh water lakes. While these solutions might be possible in theory they are incredibly expensive to put into practice.

Future Work

With more time and data on year to year water usage changes, we could create a predictive model for how much water will be used in the future. We only looked at three major factors, but there are many more factors that could be examined such as country development index, climate, etc