Pset 1 - Water usage

425/625

Spring 2024

Introduction

Water scarcity is a major issue in many parts of the world. According to the United Nations, "About two billion people worldwide don't have access to safe drinking water today (SDG Report 2022), and roughly half of the world's population is experiencing severe water scarcity for at least part of the year (IPCC). These numbers are expected to increase, exacerbated by climate change and population growth (WMO)."

In this problem set, we will investigate water usage estimates by crop in the United States. The `.csv` for this data set comes from here (by checking Select All and clicking Get Custom Zip) and the associated academic journal article is here. See this thread on X for a summary.

Read the academic article to familiarize yourself with the basics of the water usage data. You don't need to know how these water usage levels were estimated, so you can skip over those parts. We are going to focus on visualizing the water levels using the estimates that they generated.

Data preparation

The `.zip` file `rawdata/DOI-10-13012-b2idb-4607538_v1.zip` contains one `.csv` file per source (SWW, GWW, GWD) per year from 2008 to 2020. There are also a couple of `.txt` files in the folder. We can use `unzip` with `list = TRUE` to see what's in the `.zip` file.

```
unzip(zipfile = 'rawdata/DOI-10-13012-b2idb-4607538_v1.zip',
      list = TRUE) ## this lists the filename, but does not unzip the file
```

```
##                                          Name  Length                Date
## 1        DOI-10-13012-b2idb-4607538_v1/readme.txt    1053 2023-10-29 14:08:00
## 2      DOI-10-13012-b2idb-4607538_v1/gwa_2008.csv 2274812 2023-10-29 14:08:00
## 3      DOI-10-13012-b2idb-4607538_v1/gwa_2009.csv 2274812 2023-10-29 14:08:00
## 4      DOI-10-13012-b2idb-4607538_v1/gwa_2010.csv 2200859 2023-10-29 14:08:00
## 5      DOI-10-13012-b2idb-4607538_v1/gwa_2011.csv 2274812 2023-10-29 14:08:00
## 6      DOI-10-13012-b2idb-4607538_v1/gwa_2012.csv 2274812 2023-10-29 14:08:00
## 7      DOI-10-13012-b2idb-4607538_v1/gwa_2013.csv 2274812 2023-10-29 14:08:00
## 8      DOI-10-13012-b2idb-4607538_v1/gwa_2014.csv 2274812 2023-10-29 14:08:00
## 9      DOI-10-13012-b2idb-4607538_v1/gwa_2015.csv 2200859 2023-10-29 14:08:00
## 10     DOI-10-13012-b2idb-4607538_v1/gwa_2016.csv 2275517 2023-10-29 14:08:00
## 11     DOI-10-13012-b2idb-4607538_v1/gwa_2017.csv 2275517 2023-10-29 14:08:00
## 12     DOI-10-13012-b2idb-4607538_v1/gwa_2018.csv 2275517 2023-10-29 14:08:00
## 13     DOI-10-13012-b2idb-4607538_v1/gwa_2019.csv 2275517 2023-10-29 14:08:00
## 14     DOI-10-13012-b2idb-4607538_v1/gwa_2020.csv 2275517 2023-10-29 14:08:00
## 15     DOI-10-13012-b2idb-4607538_v1/gwd_2008.csv  211884 2023-10-29 14:08:00
## 16     DOI-10-13012-b2idb-4607538_v1/gwd_2009.csv  208249 2023-10-29 14:08:00
## 17     DOI-10-13012-b2idb-4607538_v1/gwd_2010.csv  214546 2023-10-29 14:08:00
## 18     DOI-10-13012-b2idb-4607538_v1/gwd_2011.csv  213608 2023-10-29 14:08:00
## 19     DOI-10-13012-b2idb-4607538_v1/gwd_2012.csv  210157 2023-10-29 14:08:00
```

```
## 20     DOI-10-13012-b2idb-4607538_v1/gwd_2013.csv  207564 2023-10-29 14:08:00
## 21     DOI-10-13012-b2idb-4607538_v1/gwd_2014.csv  209619 2023-10-29 14:08:00
## 22     DOI-10-13012-b2idb-4607538_v1/gwd_2015.csv  208683 2023-10-29 14:08:00
## 23     DOI-10-13012-b2idb-4607538_v1/gwd_2016.csv  206644 2023-10-29 14:08:00
## 24     DOI-10-13012-b2idb-4607538_v1/gwd_2017.csv  206188 2023-10-29 14:08:00
## 25     DOI-10-13012-b2idb-4607538_v1/gwd_2018.csv  206429 2023-10-29 14:08:00
## 26     DOI-10-13012-b2idb-4607538_v1/gwd_2019.csv  208246 2023-10-29 14:08:00
## 27     DOI-10-13012-b2idb-4607538_v1/gwd_2020.csv  208252 2023-10-29 14:08:00
## 28      DOI-10-13012-b2idb-4607538_v1/sw_2008.csv 2274792 2023-10-29 14:08:00
## 29      DOI-10-13012-b2idb-4607538_v1/sw_2009.csv 2274792 2023-10-29 14:08:00
## 30      DOI-10-13012-b2idb-4607538_v1/sw_2010.csv 2200839 2023-10-29 14:08:00
## 31      DOI-10-13012-b2idb-4607538_v1/sw_2011.csv 2274792 2023-10-29 14:08:00
## 32      DOI-10-13012-b2idb-4607538_v1/sw_2012.csv 2274792 2023-10-29 14:08:00
## 33      DOI-10-13012-b2idb-4607538_v1/sw_2013.csv 2274792 2023-10-29 14:08:00
## 34      DOI-10-13012-b2idb-4607538_v1/sw_2014.csv 2274792 2023-10-29 14:08:00
## 35      DOI-10-13012-b2idb-4607538_v1/sw_2015.csv 2200839 2023-10-29 14:08:00
## 36      DOI-10-13012-b2idb-4607538_v1/sw_2016.csv 2275497 2023-10-29 14:08:00
## 37      DOI-10-13012-b2idb-4607538_v1/sw_2017.csv 2275497 2023-10-29 14:08:00
## 38      DOI-10-13012-b2idb-4607538_v1/sw_2018.csv 2275497 2023-10-29 14:08:00
## 39      DOI-10-13012-b2idb-4607538_v1/sw_2019.csv 2275497 2023-10-29 14:08:00
## 40      DOI-10-13012-b2idb-4607538_v1/sw_2020.csv 2275497 2023-10-29 14:08:00
## 41 DOI-10-13012-b2idb-4607538_v1/dataset_info.txt    3894 2023-10-29 14:08:00
```

Before summarizing/visualizing this data, we'll want to join these data sets. We could certainly unzip the file manually. We can also do this in R using `unzip`.

```
unzip(zipfile = 'rawdata/DOI-10-13012-b2idb-4607538_v1.zip',
      junkpaths = TRUE,
      exdir = 'rawdata') ## gets rid of paths, keeps only filenames
```

1. Join data First, let's create a data set with all years/crops together in one data frame. Below is some code to help you get started. Add comments to each place there is `##` to explain what the chunk of code is doing. Then add code to the `Tranforming data` Section to transform the data into a data frame with 5 columns: `GEOID`, `crop`, `source`, `year`, and `value` (indicating km^3 of water).

Note that `eval = F` at the start of the chunk will prevent this chunk from evaluating when you knit the document. You can temporarily remove it if you'd like, but you'll want to add it back before knitting the document so that knitting takes less time.

```
sources = c('gwd', 'sw', 'gwa')
years = 2008:2020
d = NULL

for(s in sources){
  cat(s, '') ## show progress

  for(year in years){
    cat(year, '') ## show progress

    ##
    filename = paste0('rawdata/', s, '_', year, '.csv')
    df = read.csv(filename)
    head(df)
```

```
    ## Tranform data ########################################
    ## Use `pivot_longer`, `separate`, and/or other functions to transform this
    ## data frame into a data frame with 5 columns:
    ## GEOID, crop, source, year, and value (indicating km^3 of water)
    ##
    ## You can use the code below to check your work.
    ##
    df <- df %>%
      pivot_longer(cols = -"GEOID", names_to = "crop", values_to = "value") %>%
      separate(col = crop, into = c("src", "crop","year"), sep = "\\.")

    ## end of transforming data ############################

    ##
    d = rbind(d, df)
  }

  cat('\n') ## start a new line before showing progress for the next source
}
head(d)
tail(d)
```

Data exploration and summaries

Let's load the data we'll use for the rest of the assignment. This is the data set created in #1, so if you were unable to finish #1, you can still do the rest of the assignment.

```
d = readRDS('data/water.usage.rds')
head(d)
```

```
## # A tibble: 6 x 5
##   GEOID crop        src   year  value
##   <int> <chr>       <chr> <chr> <dbl>
## 1  1001 barley      gwd   2008      0
## 2  1001 corn        gwd   2008      0
## 3  1001 cotton      gwd   2008      0
## 4  1001 millet      gwd   2008      0
## 5  1001 oats        gwd   2008      0
## 6  1001 other_sctg2 gwd   2008      0
```

2. Summaries of data Find the **annual** mean, the change from 2008 to 2020, and the percent change from 2008 to 2020, for each crop and each source (SWW, GWW, GWD).

```
dd1 = d %>%
  group_by(crop, src, year) %>%
  summarize(value = sum(value)) %>%
  group_by(crop, src) %>%
  mutate(mean = mean(value))
```

```
## `summarise()` has grouped output by 'crop', 'src'. You can override using the
## `.groups` argument.
```

```
dd1
```

```
## # A tibble: 780 x 5
## # Groups:   crop, src [60]
```

```
##    crop   src   year  value  mean
##    <chr>  <chr> <chr> <dbl> <dbl>
##  1 barley gwa   2008  1.21   1.19
##  2 barley gwa   2009  1.19   1.19
##  3 barley gwa   2010  1.11   1.19
##  4 barley gwa   2011  1.53   1.19
##  5 barley gwa   2012  1.46   1.19
##  6 barley gwa   2013  1.27   1.19
##  7 barley gwa   2014  0.857  1.19
##  8 barley gwa   2015  1.33   1.19
##  9 barley gwa   2016  1.12   1.19
## 10 barley gwa   2017  1.16   1.19
## # i 770 more rows
```

```
dd2 <- dd1 %>%
  pivot_wider(names_from = year, values_from = value) %>%
  mutate(change = `2020` - `2008`,
         percent_change = (change/`2008`)*100) %>%
  select(-`2009`, -`2010`, -`2011`, -`2012`, -`2013`, -`2014`, -`2015`, -`2016`, -`2017`, -`2018`, -`20

dd2
```

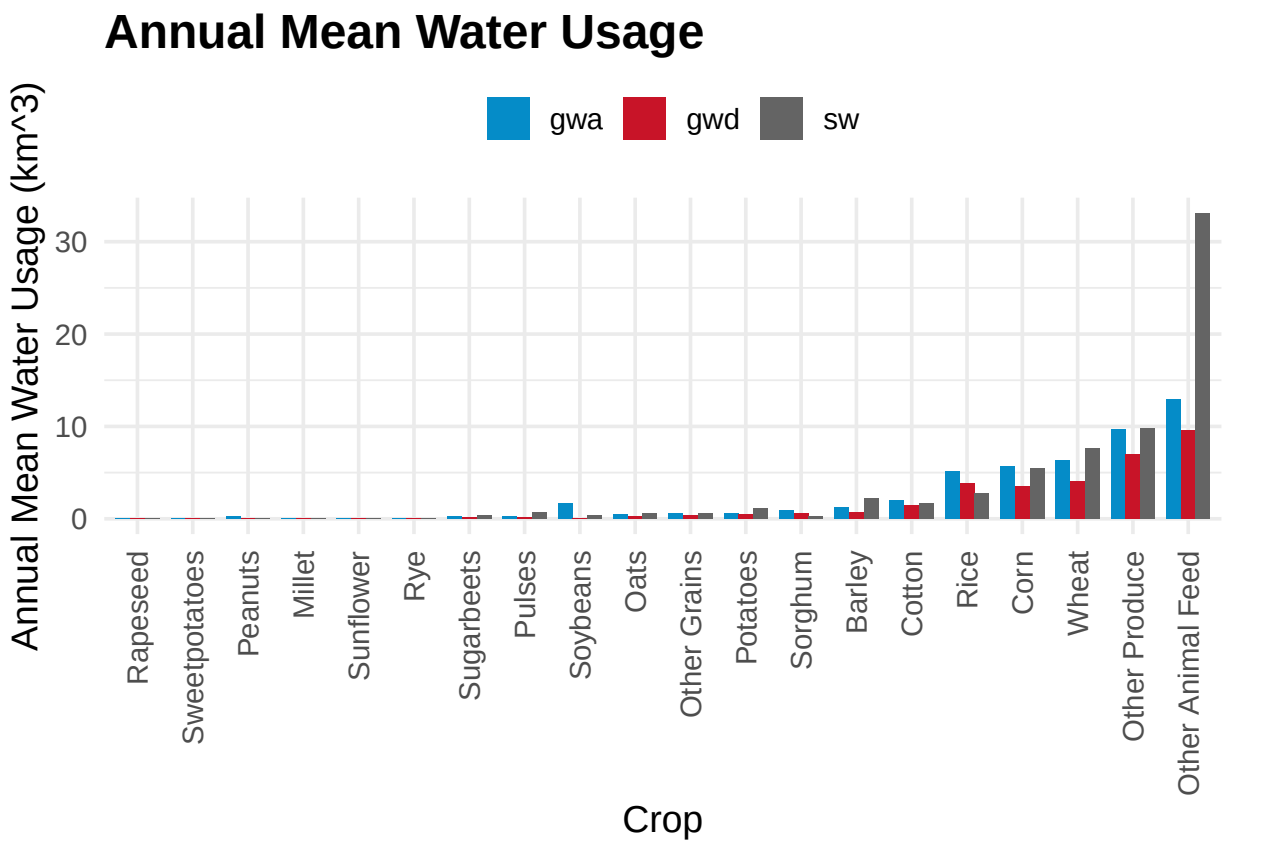
```
## # A tibble: 60 x 5
## # Groups:   crop, src [60]
##    crop   src     mean  change percent_change
##    <chr>  <chr>  <dbl>   <dbl>          <dbl>
##  1 barley gwa   1.19    0.0631          5.21
##  2 barley gwd   0.711  -0.118          -17.4
##  3 barley sw    2.20   -0.508          -21.4
##  4 corn   gwa   5.65    0.617           11.5
##  5 corn   gwd   3.52   -0.167           -4.61
##  6 corn   sw    5.50   -2.19           -30.7
##  7 cotton gwa   2.00   -0.0846          -5.19
##  8 cotton gwd   1.42    0.154           15.1
##  9 cotton sw    1.66   -1.05           -54.7
## 10 millet gwa   0.0740  0.0241          27.2
## # i 50 more rows
```

3. Convert Table 2 to a visualization

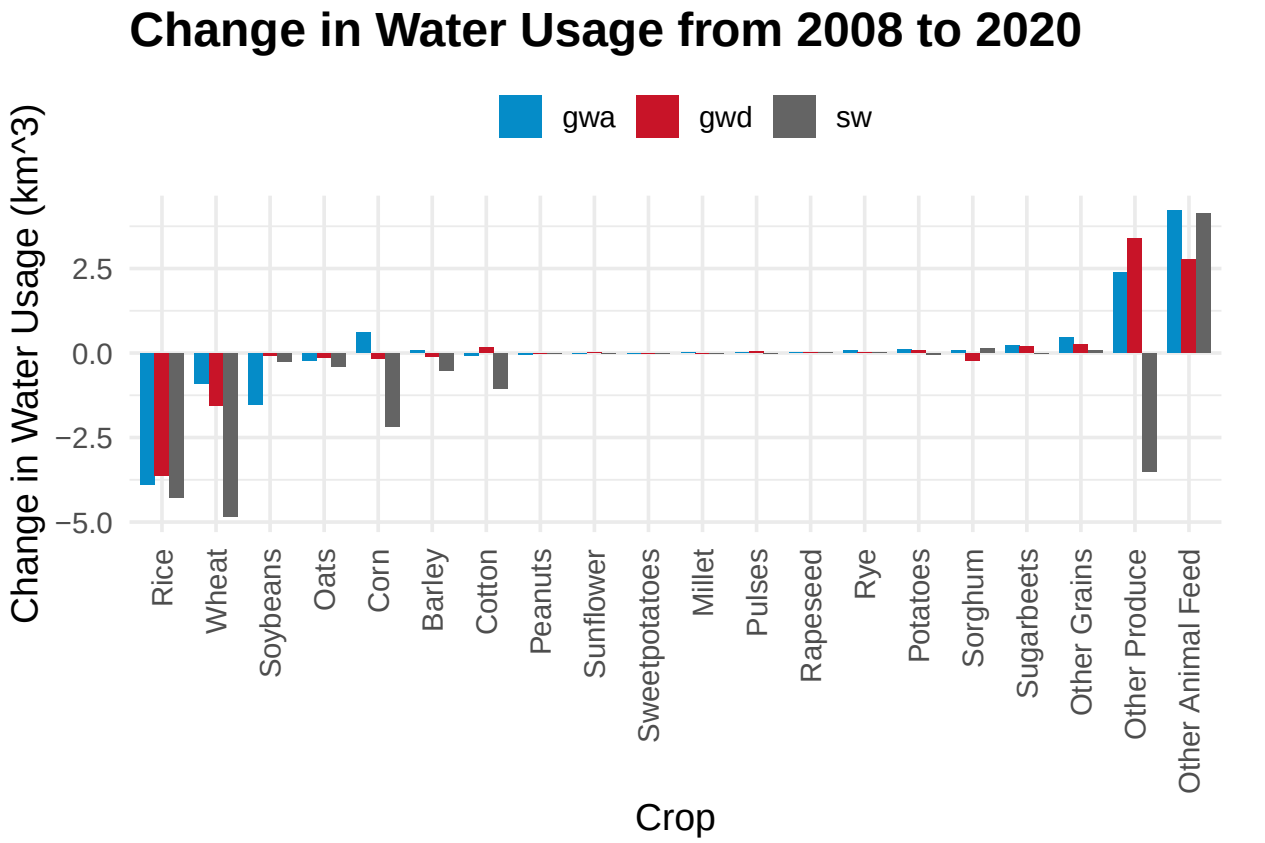
Create a visual representation of the information in Table 2. Create a visualization (or visualizations) that contains mean, change, and percent change in water usage from each crop and source.

```
# Function to create a plot with an enhanced appearance
create_plot <- function(data, y_value, y_label, title) {
  data %>%
    ggplot(aes(x = fct_reorder(str_to_title(crop), get(y_value)), y = get(y_value), fill = src)) +
    theme_pub() + # Use the pubtheme package
    geom_col(position = "dodge", width = 0.8) + # Adjust width of the bars
    labs(title = title, x = "Crop", y = y_label) +
    theme_minimal(base_size = 14) + # Use a minimal theme with a base font size
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
          legend.position = "top", # Move legend to the bottom
          legend.title = element_blank(), # Remove the legend title
          plot.title = element_text(face = "bold", size = 18), # Bold and larger plot title
```

```
          axis.title = element_text(size = 14)) # Larger axis titles
}

dd2 <- dd2 %>%
  mutate(crop = case_when(
    crop == "other_sctg2" ~ "Other Grains",
    crop == "other_sctg3" ~ "Other Produce",
    crop == "other_sctg4" ~ "Other Animal Feed",
    TRUE ~ as.character(crop) # Keep other crop names as they are
  ))

# Plot for Mean Water Usage
mean_plot <- create_plot(dd2, "mean", "Annual Mean Water Usage (km^3)", "Annual Mean Water Usage")

# Plot for Change in Water Usage
change_plot <- create_plot(dd2, "change", "Change in Water Usage (km^3)", "Change in Water Usage from 2

# Plot for Percent Change in Water Usage
percent_change_plot <- create_plot(dd2, "percent_change", "Percent Change in Water Usage (%)", "Percent

# Print the plots
print(mean_plot)
```




```
print(change_plot)
```




```
print(percent_change_plot)
```

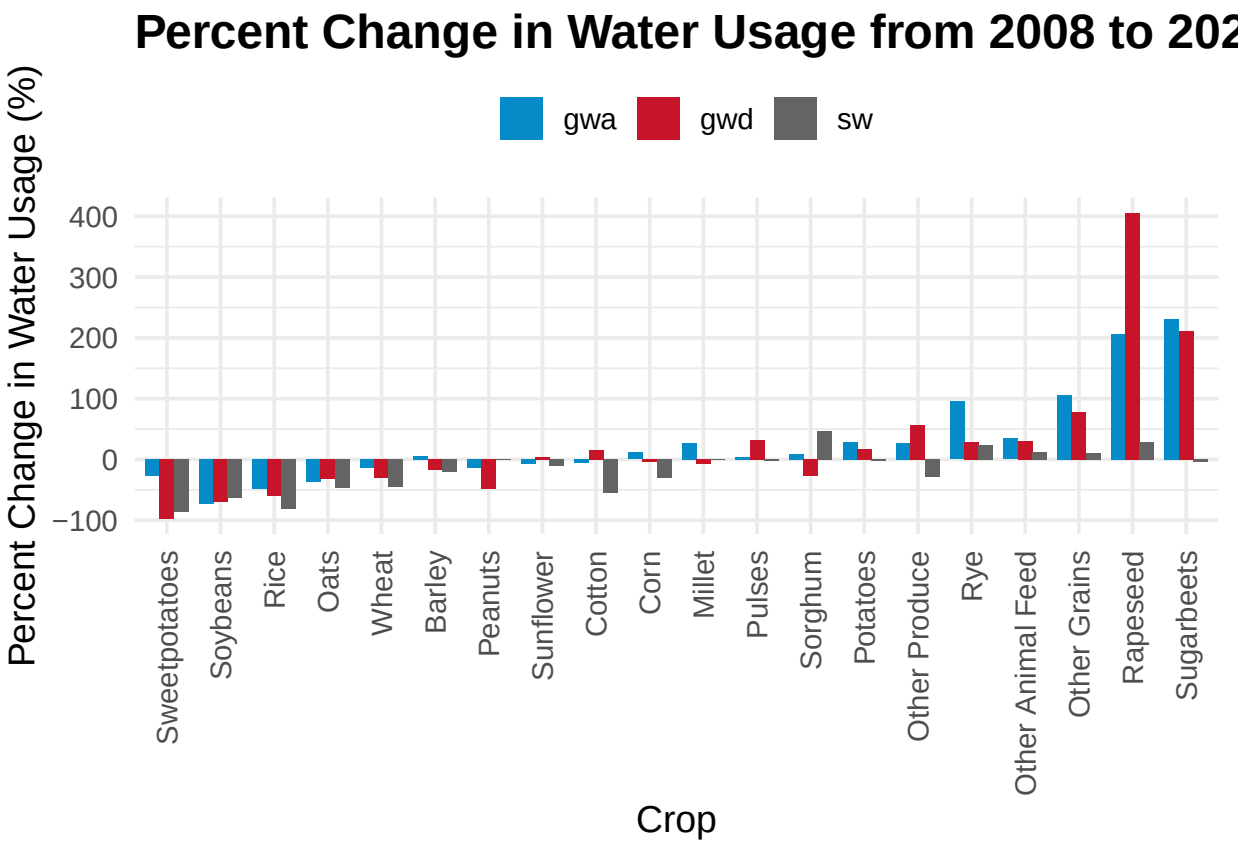


Figure 4

Figure 4 shows the average water usage by crop and source.

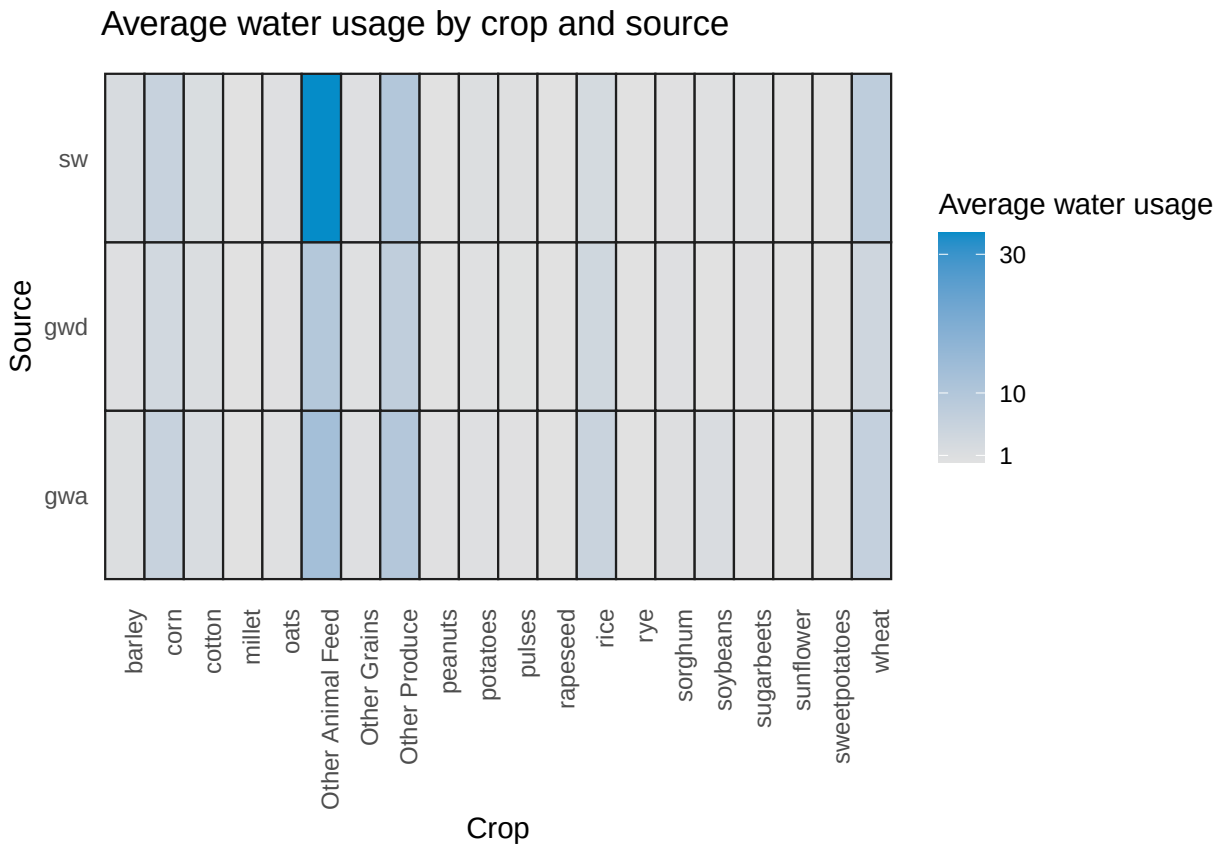
- A. average irrigation water usage by source, colored by crop,
- B. average irrigation water usage by crop, colored by source

Two other options for visualizing a numeric variable broken down by two different categorical variable would be a tile plot/grid plot (e.g. https://github.com/bmacGTPM/pubtheme?tab=readme-ov-file#grid-plot) and a mosiac plot (https://haleyjeppson.github.io/ggmosaic/).

4. Create a tile plot/grid plot of the data in Figure 4.

```
grid_plot = ggplot(dd2,
           aes(x    = crop,
               y    = src,
               fill = mean)) +
  geom_tile(linewidth   = 0.4,
            show.legend = T,
            color       = pubdarkgray) +
  scale_fill_gradient(low      = pubgradgray,
                      high     = pubblue,
                      na.value = pubmediumgray, ## same color as below
                      oob      = squish,
                      breaks   = c(1, 10, 30)) +

  labs(title    = 'Average water usage by crop and source',
```

```
       fill = 'Average water usage',
       x = 'Crop',
       y = 'Source') +
  theme(axis.text.x = element_text(angle = 90, hjust=1),
        axis.ticks.x=element_blank(), axis.ticks.y=element_blank(),
        panel.background = element_blank())


grid_plot
```

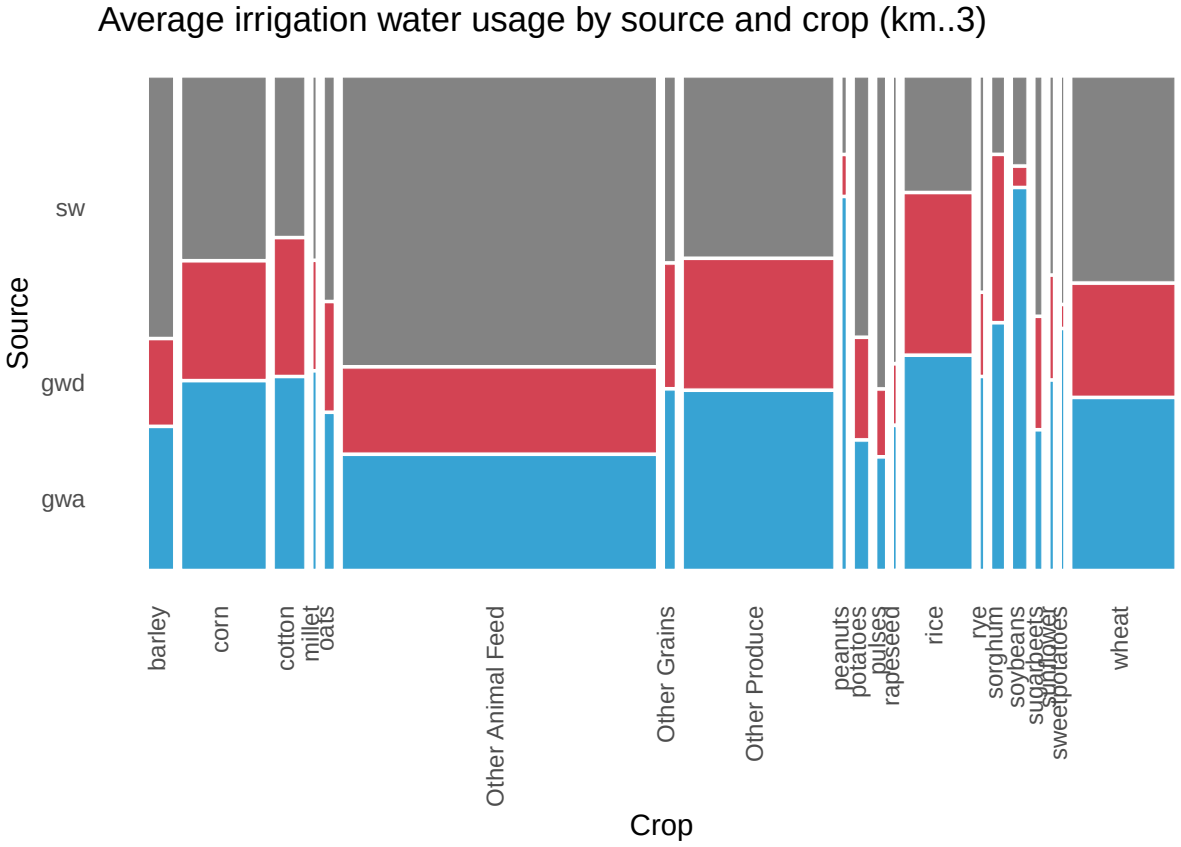


5. Create a mosiac plot of the data in Figure 4.

```
library(ggmosaic)

mosaic_plot = dd2 %>% arrange(desc(mean)) %>%
     ggplot() +
   geom_mosaic(aes(x = product(src,crop), fill=src, weight = mean), show.legend = FALSE) +
    labs(title = "Average irrigation water usage by source and crop (km^3)",
    x = 'Crop', y = "Source") +
    theme(axis.text.x = element_text(angle = 90, hjust=1, vjust=0.25),
    axis.ticks.x=element_blank(), axis.ticks.y=element_blank(),
    panel.background = element_blank())


mosaic_plot
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <cb>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <86>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <cb>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <86>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <cb>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <86>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <cb>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <86>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <cb>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <86>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <cb>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <86>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <cb>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <86>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <cb>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
```

```
## conversion failure on 'Average irrigation water usage by source and crop
## (km^3)' in 'mbcsToSbcs': dot substituted for <86>
```



6. What are the benefits (other than it fits on one plot) and drawbacks of these two plots?

The grid plot does well on showing the average water usage by crop and source, and it emphasizes on the crop with high average on water usage. But it is difficult to compare the water usage across crops. Especially, the scales of the gradient color is hard to balance given that some crops of water usage like 'Other Animal Feed' has a much greater mean than the rest. The audience may not be able to see the differences in water usage across crops with little water usage.

The mosaic plot is similar to the grid plot as it emphasizes the emphasizes on the crop with great average water usage. It is easier to compare the water usage across crops, but it is difficult to compare the water usage across sources. The audience may not be able to see the differences in water usage across sources with little water usage as well.

7. Figure 6

Figure 6 uses a different color scale for each plot. Discuss the benefits and drawbacks of this choice. What was the main purposes of this figure? Given the main purpose, would you recommend using the same color scale, or different color scales, for each plot?

The main purpose of this figure is to demonstrate the details of water usage in the States in 2020, including the distributions of water uasge of each crop and source and the differences between each crop and source. The different color scales for each plot is beneficial because it is easy to tell the distribution of water usage for each crop and source, i.e., in which region of America consumes the most amount of water. The drawback of this choice is that it is difficult to compare the water usage across crops and sources. I would recommend

using different color scales for each plot as the purpose of this figure is to show the distribution of water usage for each crop and source, not to compare the water usage across crops and sources.

8. Figure 8

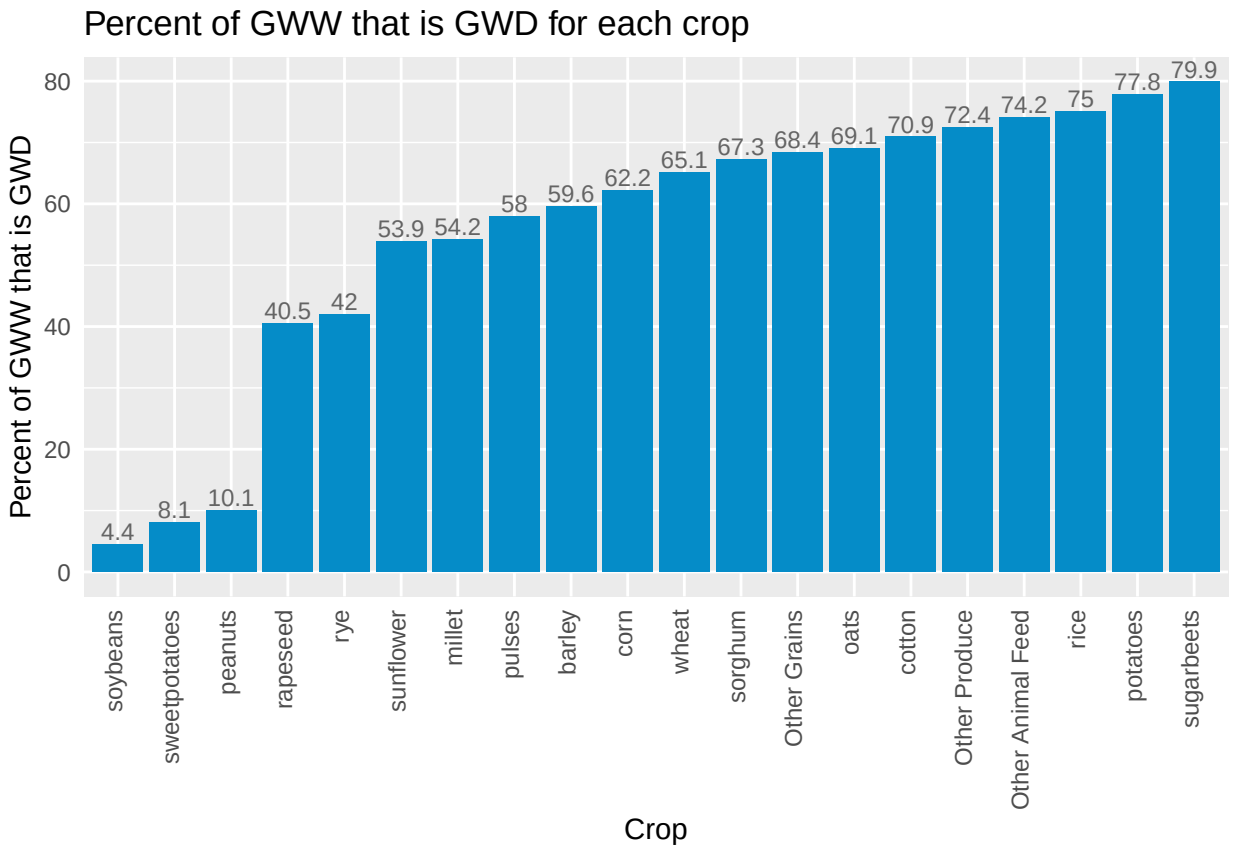
Figure 8 also uses a different color scale for each plot. Discuss the benefits and drawbacks of this choice. What was the main purposes of this figure? Given the main purpose, would you recommend using the same color scale, or different color scales, for each plot?

The main purpose of figure 8 is to show the difference of PCR estimates of water usage and the USGS reported water usage. The good thing of using a different scale for each plot is that it is easy to tell where in the States does the estimates of PCR and the USGS reported values differ the most. The drawback of this choice is that it is difficult to compare the PCR estimates of water usage and the USGS reported water usage across different years. Given that in the original paper, the authors emphasize on the biased estimation of PCR in western areas and try to argue that the PCR model has potential issues, I would say using different scales is better for this purpose. This is because we don't need to show that the error of PCR estimates actually decreases over time, but we need to show that the PCR estimates are biased in the western areas.

9. Breakdown of GWW

The paper notes in Section 3.1 that $GWW = GWW_{sustainable} + GWW_{unsustainable}$, and that $GWD = GWW_{unsustainable}$. Create a visualization showing the percent of GWW that is GWD for each crop. Use the mean values for water usage.

```
dd2 %>%
  filter(src == "gwa" | src == "gwd") %>%
  select(crop, src, mean) %>%
  pivot_wider(names_from = src, values_from = mean) %>%
  mutate(percent_of_GWD = gwd/gwa * 100) %>%
  ggplot(aes(x = reorder(crop,percent_of_GWD), y = percent_of_GWD, label = round(percent_of_GWD,1))) +
  geom_col() +
  geom_text(vjust = -0.25, size = 3) +
  labs(title = "Percent of GWW that is GWD for each crop",
       x = "Crop",
       y = "Percent of GWW that is GWD") +
  theme(axis.text.x = element_text(angle = 90, hjust=1, vjust=0.25),
        axis.ticks.x=element_blank(), axis.ticks.y=element_blank())
```

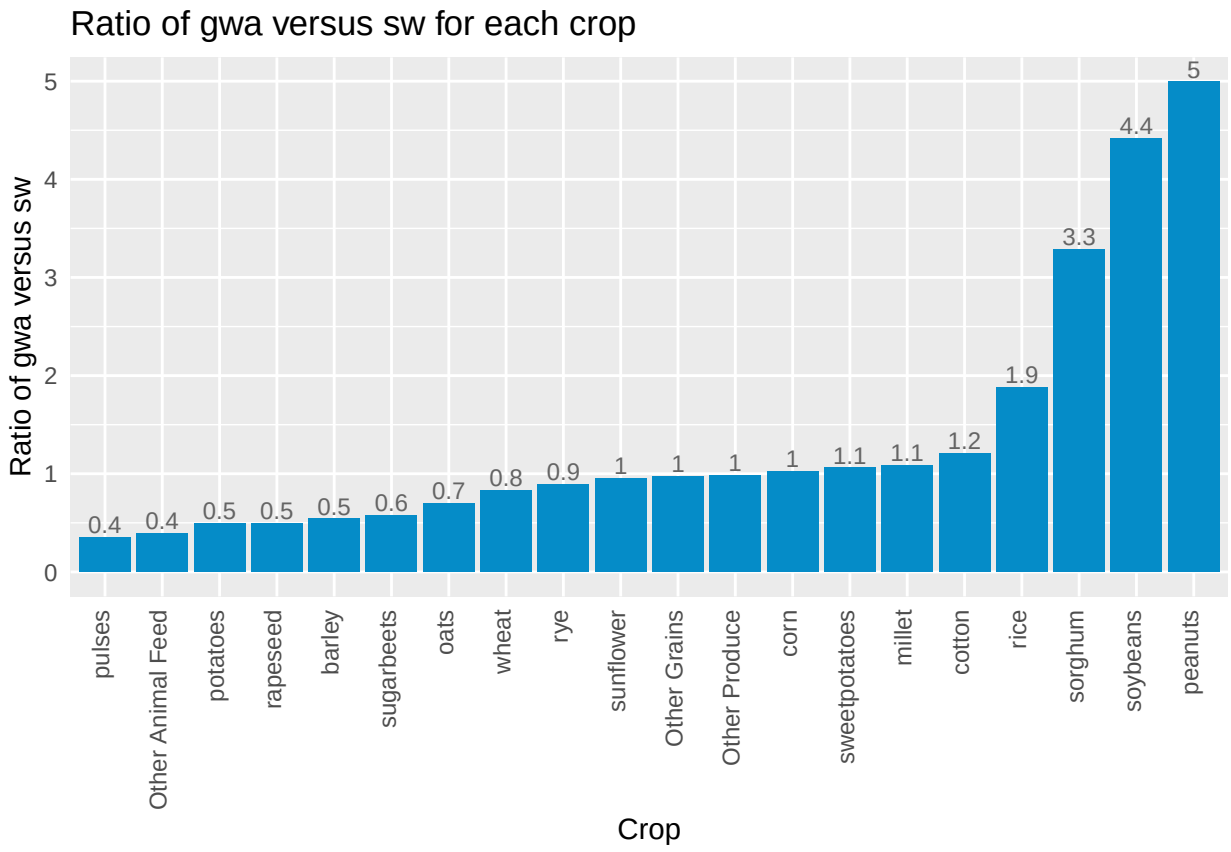


10. Custom visualization

What is another question you have about this data? Create a visualization that attempt to answer your question.

Question: What is the ratio of gwa versus sw for each crop using the mean value?

```
dd2 %>%
  filter(src == "gwa" | src == "sw") %>%
  select(crop, src, mean) %>%
  pivot_wider(names_from = src, values_from = mean) %>%
  mutate(ratio = gwa/sw) %>%
  ggplot(aes(x = reorder(crop,ratio), y = ratio, label = round(ratio,1))) +
  geom_col() +
  geom_text(vjust = -0.25, size = 3) +
  labs(title = "Ratio of gwa versus sw for each crop",
       x = "Crop",
       y = "Ratio of gwa versus sw") +
  theme(axis.text.x = element_text(angle = 90, hjust=1, vjust=0.25),
        axis.ticks.x=element_blank(), axis.ticks.y=element_blank())
```


Ratio of gwa versus sw for each crop
Ratio of gwa versus sw
0
1
2
3
4
5
0.4
0.4
0.5
0.5
0.5
0.6
0.7
0.8
0.9
1
1
1
1
1.1
1.1
1.2
1.9
3.3
4.4
5
pulses
Other Animal Feed
potatoes
rapeseed
barley
sugarbeets
oats
wheat
rye
sunflower
Other Grains
Other Produce
corn
sweetpotatoes
millet
cotton
rice
sorghum
soybeans
peanuts
Crop