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.

```r
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`.

```r
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.

```r
# SWW: surface water withdrawals
# GWW: groundwater withdrawals
# GWD: nonrenewable groundwater depletion
# GWA: groundwater abstractions
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

    ## read each file
      # filename: each file begins with 'rawdata..._v1/'
        # s is either gwd, sw, or gwa and year ranges from 2008 to 2020
        # (specific s and year values are specified by the for loops)
      # each file name is then read into df
        # so we have one data frame with data from all sources and years
    filename = paste0('rawdata/DOI-10-13012-b2idb-4607538_v1/DOI-10-13012-b2idb-4607538_v1/', s, '_', ye
    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)

    df = df %>%
      pivot_longer(cols = !GEOID,
                   names_to = 'src.crop.year',
                   values_to = 'value') %>%
      separate(src.crop.year, into = c('src', 'crop', 'year'), sep = '[.]') %>%
      relocate(GEOID, crop, src, year, value)
    df



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

    ## Each df has values for only one source and year
      # so at the end of each year iteration, we add df to the larger d dataframe
      # after all iterations of source and year, d contains information for all sources and all years
      # (this is why each df has 64460 observations and d has >2.4 million)
    d = rbind(d, df)
  }

  cat('\n') ## start a new line before showing progress for the next source
}
head(df)
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 mean, the change from 2008 to 2020, and the percent change from 2008 to 2020, for each crop and each source (SWW, GWW, GWD).

```
# value = km^3 of water
# mean for each crop and each source = sum of value for all crops / number of crops
  # (for each crop, src)

# total water usage for each crop, source, year, for all census tracts
dd3 = d %>%
  group_by(crop, src, year) %>%
  summarise(value = sum(value))
```

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

```
dd3
```

```
## # A tibble: 780 x 4
## # Groups:   crop, src [60]
##    crop   src   year  value
##    <chr>  <chr> <chr> <dbl>
##  1 barley gwa   2008  1.21
##  2 barley gwa   2009  1.19
##  3 barley gwa   2010  1.11
##  4 barley gwa   2011  1.53
##  5 barley gwa   2012  1.46
##  6 barley gwa   2013  1.27
##  7 barley gwa   2014  0.857
##  8 barley gwa   2015  1.33
##  9 barley gwa   2016  1.12
## 10 barley gwa   2017  1.16
## # i 770 more rows
```

```
# avg annual water usage for each crop and source
dd4 = dd3 %>%
  group_by(crop, src) %>%
  mutate(mean = mean(value)) %>%
  distinct(mean)
dd4
```

```
## # A tibble: 60 x 3
## # Groups:   crop, src [60]
##    crop   src     mean
```

```
##    <chr>  <chr>  <dbl>
##  1 barley gwa   1.19
##  2 barley gwd   0.711
##  3 barley sw    2.20
##  4 corn   gwa   5.65
##  5 corn   gwd   3.52
##  6 corn   sw    5.50
##  7 cotton gwa   2.00
##  8 cotton gwd   1.42
##  9 cotton sw    1.66
## 10 millet gwa   0.0740
## # i 50 more rows
```

```r
# change
dd5 = dd3 %>%
  group_by(crop, src) %>%
  mutate(change = last(value) - first(value)) %>%
  distinct(change)
dd5
```

```
## # A tibble: 60 x 3
## # Groups:   crop, src [60]
##    crop   src    change
##    <chr>  <chr>   <dbl>
##  1 barley gwa    0.0631
##  2 barley gwd   -0.118
##  3 barley sw    -0.508
##  4 corn   gwa    0.617
##  5 corn   gwd   -0.167
##  6 corn   sw    -2.19
##  7 cotton gwa   -0.0846
##  8 cotton gwd    0.154
##  9 cotton sw    -1.05
## 10 millet gwa    0.0241
## # i 50 more rows
```

```r
# percent change
dd6 = dd3 %>%
  group_by(crop,src) %>%
  mutate(percent = (last(value) - first(value)) / first(value)) %>%
  distinct(percent)
dd6
```

```
## # A tibble: 60 x 3
## # Groups:   crop, src [60]
##    crop   src   percent
##    <chr>  <chr>   <dbl>
##  1 barley gwa    0.0521
##  2 barley gwd   -0.174
##  3 barley sw    -0.214
##  4 corn   gwa    0.115
##  5 corn   gwd   -0.0461
##  6 corn   sw    -0.307
```

```
##  7 cotton gwa   -0.0519
##  8 cotton gwd    0.151
##  9 cotton sw    -0.547
## 10 millet gwa    0.272
## # i 50 more rows
```

3. Convert Table 2 to a visualization

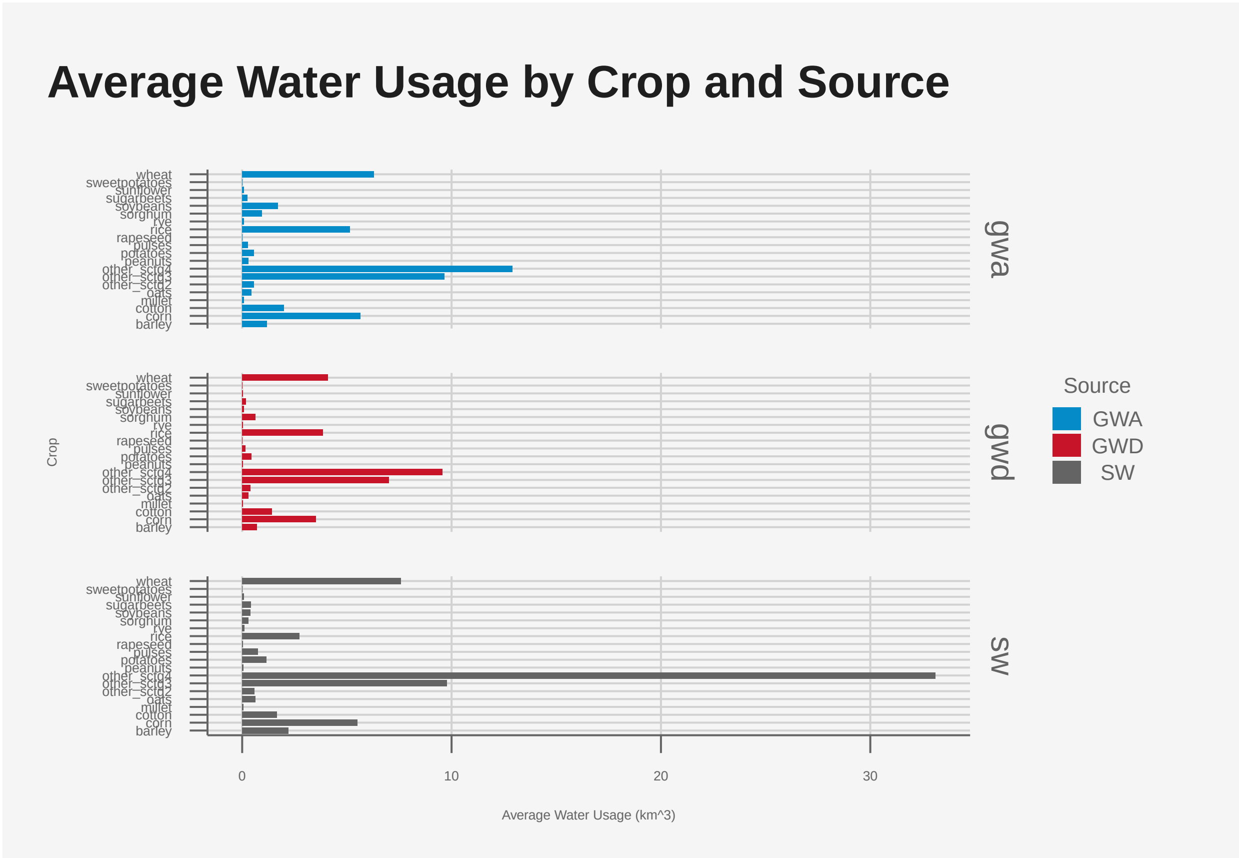
Create a visual representation of the information in Table 2. Create a visualization (or visualizations) that contains mean, change, and percent change in water usage from each crop and source.

```r
# combine mean, change, percent change into 1 dataframe
table2 <- dd3 %>%
  group_by(crop, src) %>%
  mutate(mean = mean(value),
         change = last(value) - first(value),
         percent = (last(value) - first(value)) / first(value)) %>%
  distinct(mean, change, percent)

table2
```

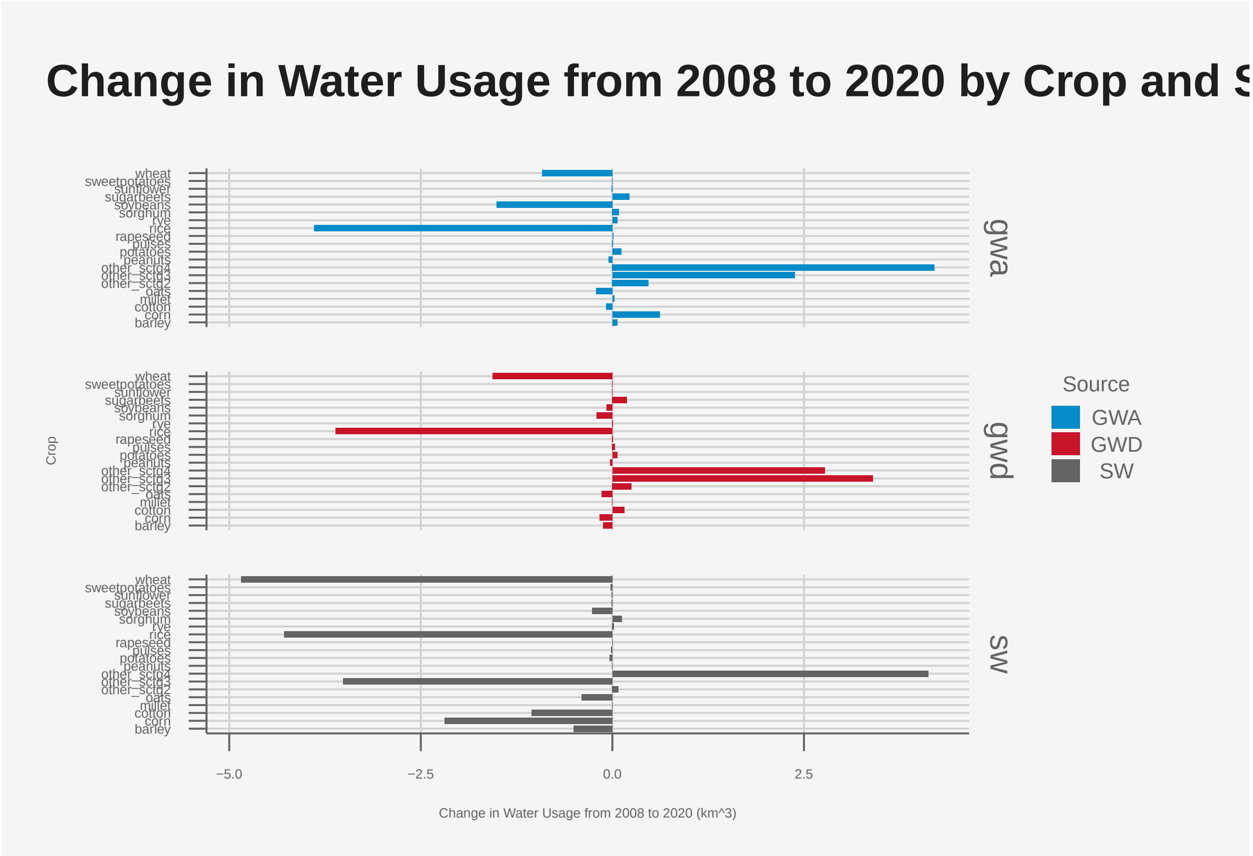
```
## # A tibble: 60 x 5
## # Groups:   crop, src [60]
##    crop   src     mean  change percent
##    <chr>  <chr>  <dbl>   <dbl>   <dbl>
##  1 barley gwa   1.19    0.0631  0.0521
##  2 barley gwd   0.711  -0.118  -0.174
##  3 barley sw    2.20   -0.508  -0.214
##  4 corn   gwa   5.65    0.617   0.115
##  5 corn   gwd   3.52   -0.167  -0.0461
##  6 corn   sw    5.50   -2.19   -0.307
##  7 cotton gwa   2.00   -0.0846 -0.0519
##  8 cotton gwd   1.42    0.154   0.151
##  9 cotton sw    1.66   -1.05   -0.547
## 10 millet gwa   0.0740  0.0241  0.272
## # i 50 more rows
```

```r
# contains mean, change, percent change
# from each crop and source

# Mean
meanvis <- table2 %>% ggplot(aes(y = crop,
                  x = mean,
                  fill = src)) +
  geom_col(width = .8, position = position_dodge()) +
  theme_pub() +
  theme(text = element_text(size = 5),
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8),
        legend.position = "right") +
  labs(title = "Average Water Usage by Crop and Source",
       x = "Average Water Usage (km^3)",
       y = "Crop",
```

```
       fill = "Source") +
   scale_fill_discrete(labels = c("GWA", "GWD", "SW")) +
  facet_grid(src~.)

# Change
changevis <- table2 %>% ggplot(aes(y = crop,
                  x = change,
                  fill = src)) +
  geom_col(width = .8, position = position_dodge()) +
  theme_pub() +
  theme(text = element_text(size = 5),
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8),
        legend.position = "right") +
  labs(title = "Change in Water Usage from 2008 to 2020 by Crop and Source",
       x = "Change in Water Usage from 2008 to 2020 (km^3)",
       y = "Crop",
       fill = "Source") +
   scale_fill_discrete(labels = c("GWA", "GWD", "SW")) +
  facet_grid(src~.)

# Percent Change
percentvis <- table2 %>% ggplot(aes(y = crop,
                  x = percent,
                  fill = src)) +
  geom_col(width = .8, position = position_dodge()) +
  theme_pub() +
  theme(text = element_text(size = 5),
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8),
        legend.position = "right") +
  labs(title = "Percent Change in Water Usage",
       x = "Percent Change in Water Usage",
       y = "Crop",
       fill = "Source") +
   scale_fill_discrete(labels = c("GWA", "GWD", "SW")) +
  facet_grid(src~.)

meanvis
```


Average Water Usage by Crop and Source
wheat
sweetpotatoes
sunflower
sugarbeets
soybeans
sorghum
rye
rice
rapeseed
pulses
potatoes
peanuts
other_scfq4
other_scfq3
other_scfq2
oats
millet
cotton
corn
barley
gwa
gwd
sw
Crop
Source
GWA
GWD
SW
0
10
20
30
Average Water Usage (km^3)

```
changevis
```



Change in Water Usage from 2008 to 2020 by Crop and S
wheat
sweetpotatoes
sunflower
sugarbeets
soybeans
sorghum
rye
rice
rapeseed
pulses
potatoes
peanuts
other_scfg4
other_scfg3
other_scfg2
oats
millet
cotton
corn
barley
gwa
gwd
sw
Crop
Source
GWA
GWD
SW
−5.0
−2.5
0.0
2.5
Change in Water Usage from 2008 to 2020 (km^3)


```
percentvis
```

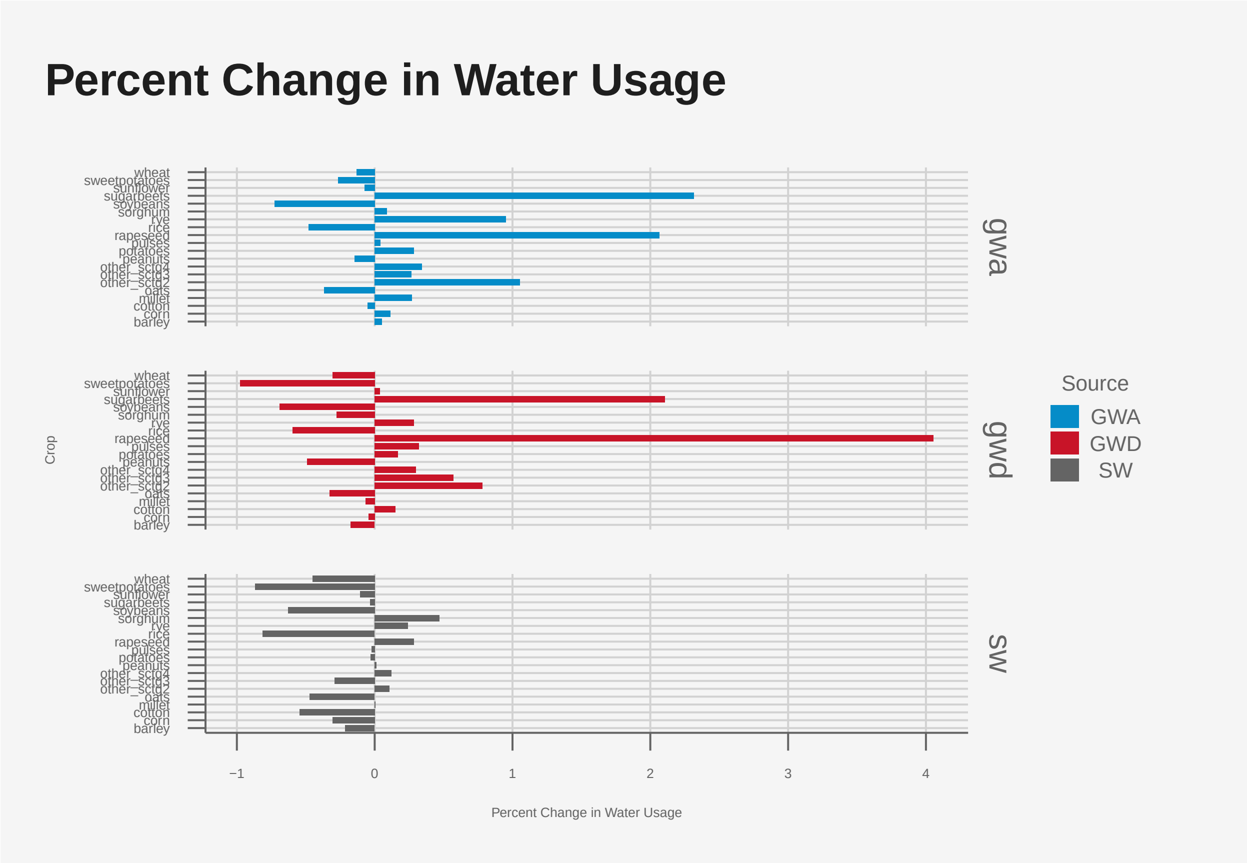


Figure 4

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

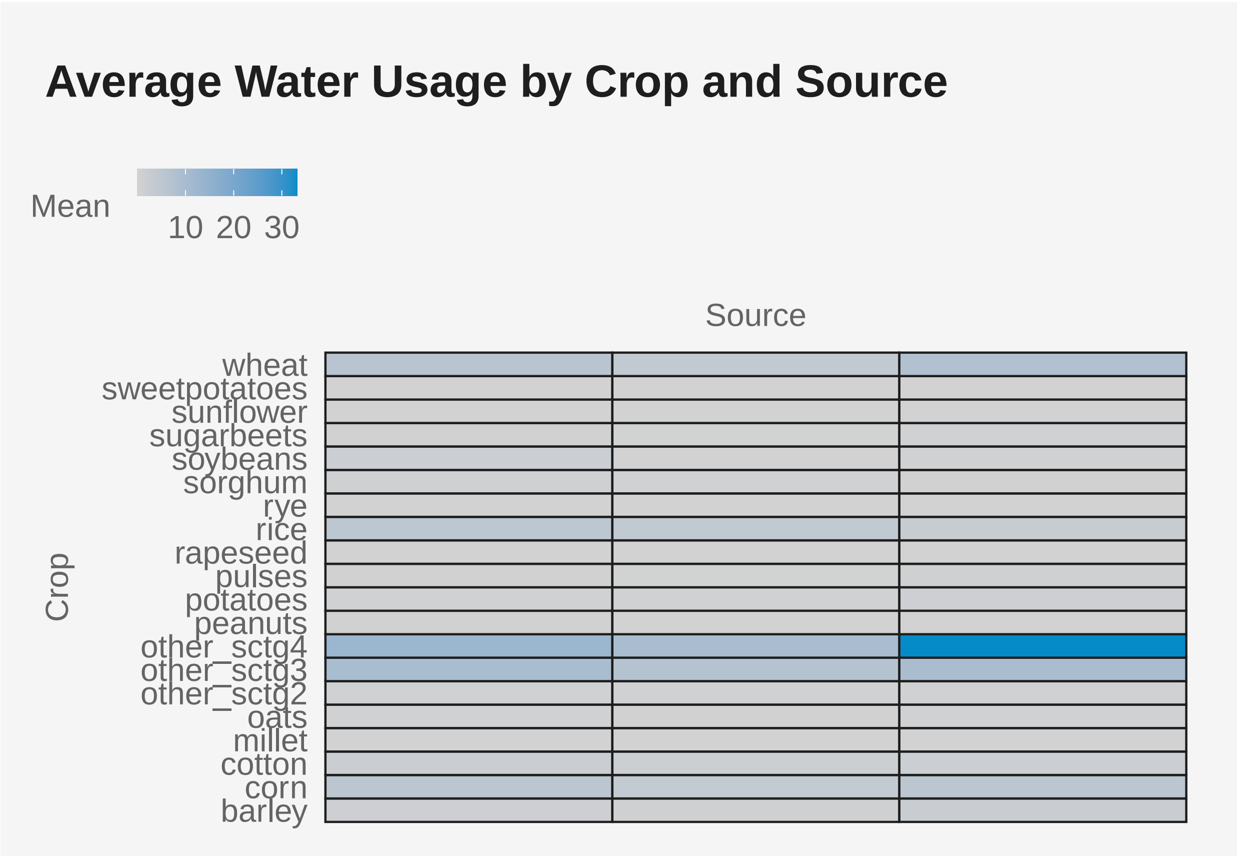
- A. average irrigation water usage by source, colored by crop,
- B. average irrigation water usage by crop, colored by source

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.

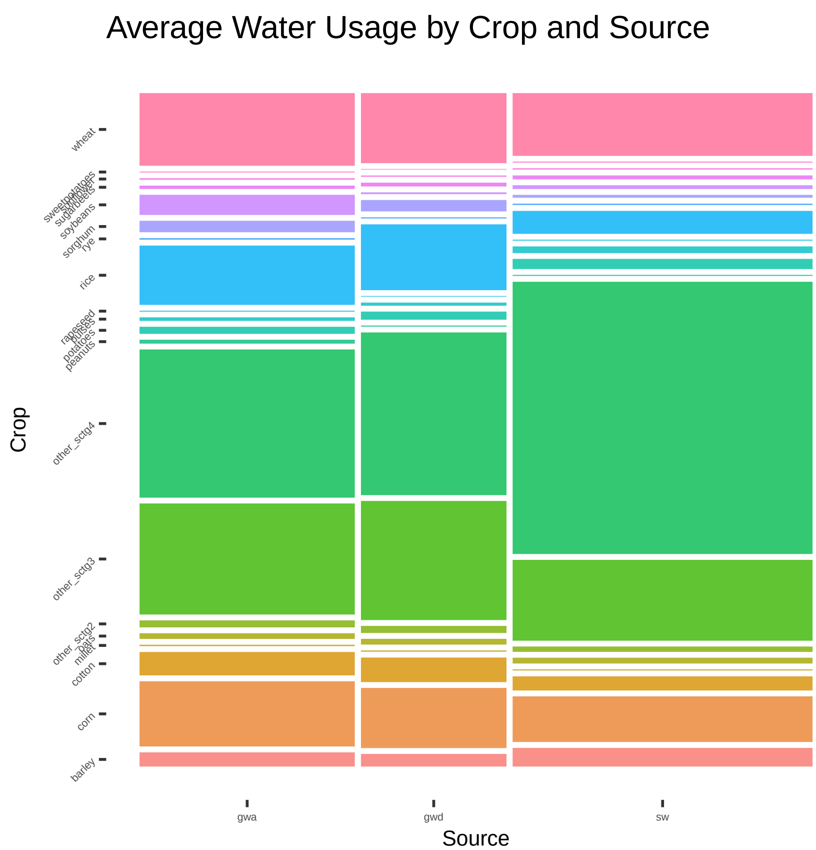
```
g <- table2 %>% ggplot(aes(x = src,
                           y = crop,
                           fill = mean)) +
  geom_tile(linewidth = 0.4,
            show.legend = T,
            color = pubdarkgray) +
  labs(title = "Average Water Usage by Crop and Source",
       x = "Source",
       y = "Crop",
```

```
      fill = 'Mean')

g %>% pub(type = 'grid',
          xbreaks = seq(2, 32, by=2))
```



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

```
library(ggmosaic)

table2 %>%
  ggplot() +
  geom_mosaic(aes(x = product(src), fill = crop, weight = mean), show.legend = F) +
  theme_mosaic() +
  theme(text = element_text(size = 5),
        axis.text.y = element_text(angle = 45, hjust = 1),
        axis.title = element_text(size = 8),
        title = element_text(size = 10)) +
  labs(title = "Average Water Usage by Crop and Source",
       x = "Source",
       y = "Crop")
```



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

Tile plot: It is easy to see the crops and sources where water usage is most heavily concentrated, but the data is concentrated strongly in one particular tile. This makes it more difficult to compare the data across other tiles.

Mosaic plot: The size variation makes it easy to compare water usage across crops and sources. It is difficult, however, to understand water usage for crops that don't use much water, because the mosaic tiles are too small to see and compare them.

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?

Using different color scales makes it easier to see the location and concentration of water use for each crop and source combination. If every color scale ranged from the minimum to maximum value (0.001 to 1.6), it would be more difficult to see the variation in graphs with data concentrated at one end of the spectrum, such as the soy-gwd and other animal feed-gwd plots. But the variation in scale also makes it more difficult to compare data across plots - although cotton and other produce appear to have the same concentration of irrigation from groundwater abstractions, other produce actually uses significantly more water.

The purpose of the figure is to show the location and concentration of irrigation for each source and crop. The figure is not meant to be used to compare one crop or water source to another; it is meant to provide information about each individual crop-source combination. Given this purpose, I would recommend using the same color scale for each plot.

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?

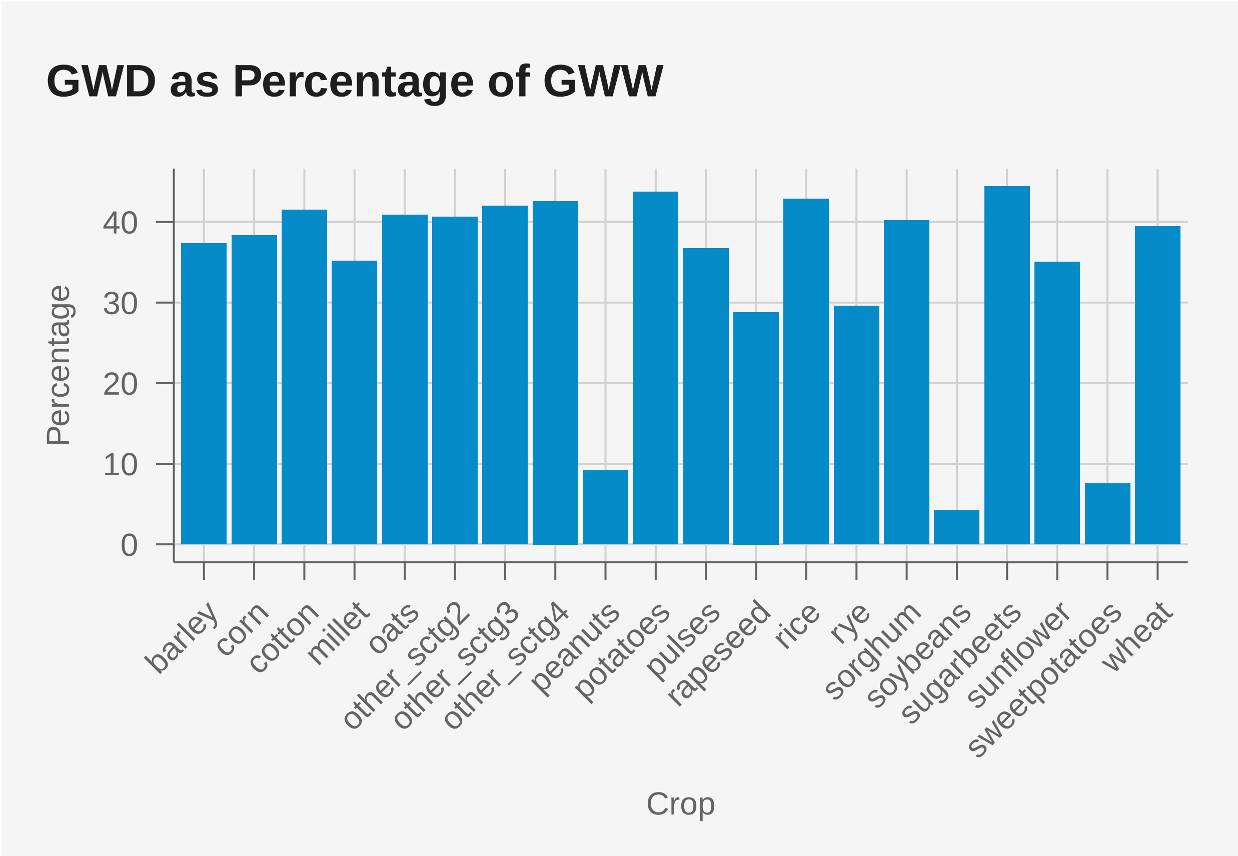
As in Figure 6, using different color scales makes it easier to see the variations in data for each plot, but it also makes it difficult to compare data between plots.

The figure's purpose is to show the difference between estimated and reported withdrawals and compare that difference between surface water and groundwater and between 2010 and 2015. Since the figure is meant to be used for comparison, I would recommend using the same color scale for each plot.

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.

```
t3 <- table2 %>%
  filter(src == "gwa" | src == "gwd") %>%
  group_by(crop) %>%
  mutate(gww = sum(mean)) %>%
  filter(src == "gwd") %>%
    mutate(gwdpercent = mean / gww * 100)

t3 %>% ggplot(aes(x = crop,
           y = gwdpercent,
           fill = src)) +
  geom_col() +
  theme_pub() +
  theme(axis.text.x = element_text(angle = 45,
                                   hjust = 1),
        legend.position = "none") +
  labs(title = "GWD as Percentage of GWW",
       x = "Crop",
       y = "Percentage")
```

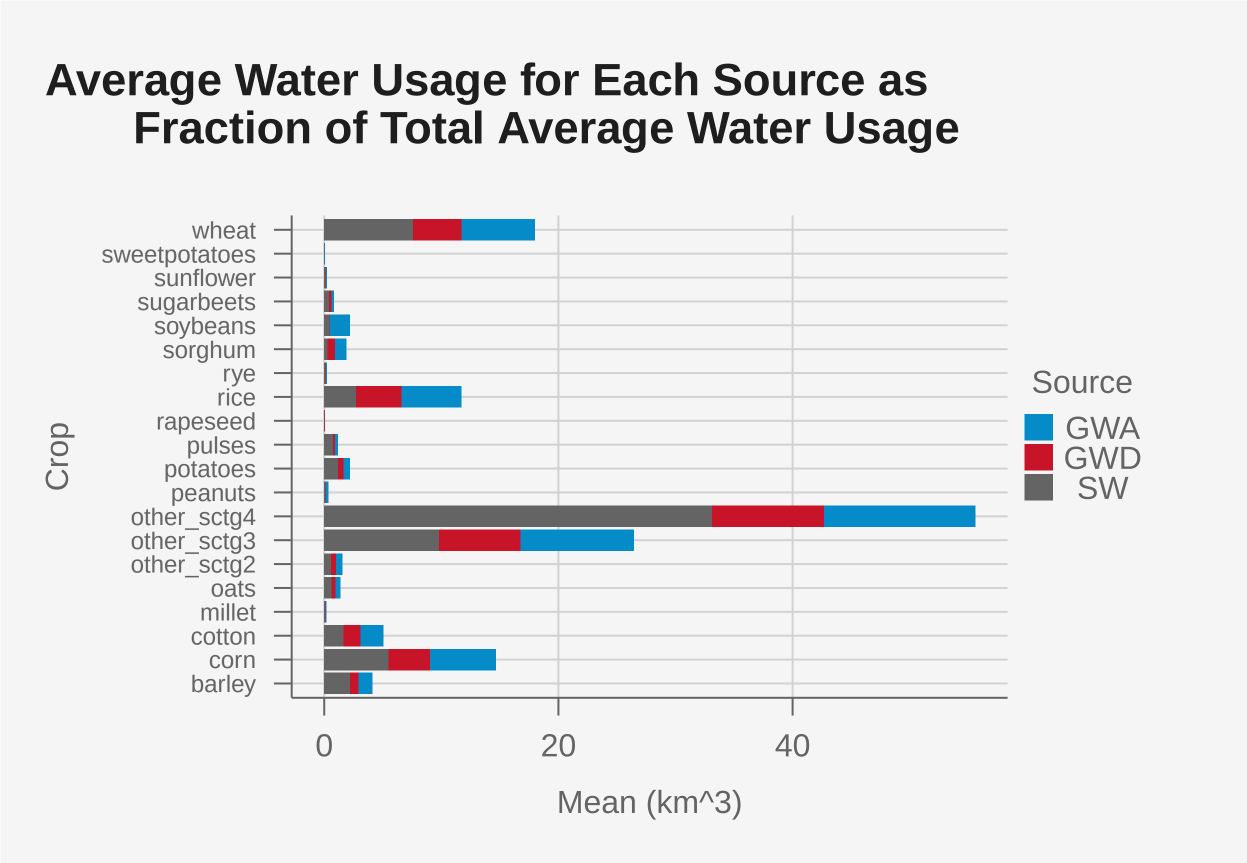


10. Custom visualization

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

For each crop, what is the ratio of average water usage for each water source compared to the total average water usage?

```
table2 %>% ggplot(aes(x = mean,
                      y = crop,
                      fill = src)) +
  geom_col() +
  theme_pub() +
  labs(title = "Average Water Usage for Each Source as
       Fraction of Total Average Water Usage",
       x = "Mean (km^3)",
       y = "Crop",
       fill = "Source") +
  theme(axis.text.y = element_text(size = 9),
        legend.position = "right") +
  scale_fill_discrete(labels = c("GWA", "GWD", "SW"))
```


Average Water Usage for Each Source as Fraction of Total Average Water Usage
wheat
sweetpotatoes
sunflower
sugarbeets
soybeans
sorghum
rye
rice
rapeseed
pulses
potatoes
peanuts
other_sctg4
other_sctg3
other_sctg2
oats
millet
cotton
corn
barley
Crop
0
20
40
Mean (km^3)
Source
GWA
GWD
SW