HW3-drafting-viz

AUTHOR
Tom Gibbens-Matsuyama

1. Which option do you plan to pursue?

I am going to pursue the first option, which is the one I planned on doing since the start.

2. Reinstate your question(s). Has this changed at all since HW #1? How so?

From working with the data I have boiled it down to understanding the differences between Sacramento county and Ventura County. I have past, present, and future data in one shape or form for both of these counties. I don't believe I have the best data as there is room for a lot of interpretation, but I can use each one to represent a figure on my visualization. I am interested in how different these two counties are given their geographic differences. A better comparison would be to directly measure between a more northern and more southern county, but these counties had the most data.

3. Explain which variables from your data set(s) you will use to answer your question(s) and how

The variables of interest are as follows:

- `org_id`: An organizational ID that refers to a specific water agency
- `forecast_start_date`: Start date for projected data
- `supplier_name`: Name of the supplier

These variables are present in each dataset. `supplier_name` and `org_id` will be used to identify the water agencies in Ventura and Sacramento counties. These will be paired with `forecast_start_date` to plot the respective values from each dataset onto a time series plot.

The variables of interest within the past dataset (`historical_production`) are:

- `water_produced_or_delivered`: If the water is being produced or delivered
- `water_type`: What type of facility the water is coming from or going to
- `quanitity_acre_feet`: In units acre-feet, the quanitity of water from each observation

These variables, along with the general variables, will be used to identify the differences of water type production/delivery between Sacramento and Ventura counties.

The variables of interest with the present dataset (`water_shortage`) are:

- `state_standard_shortage_level`: A shortage level on a scale of 0 - 6 that tells us the water shortage state of each agency

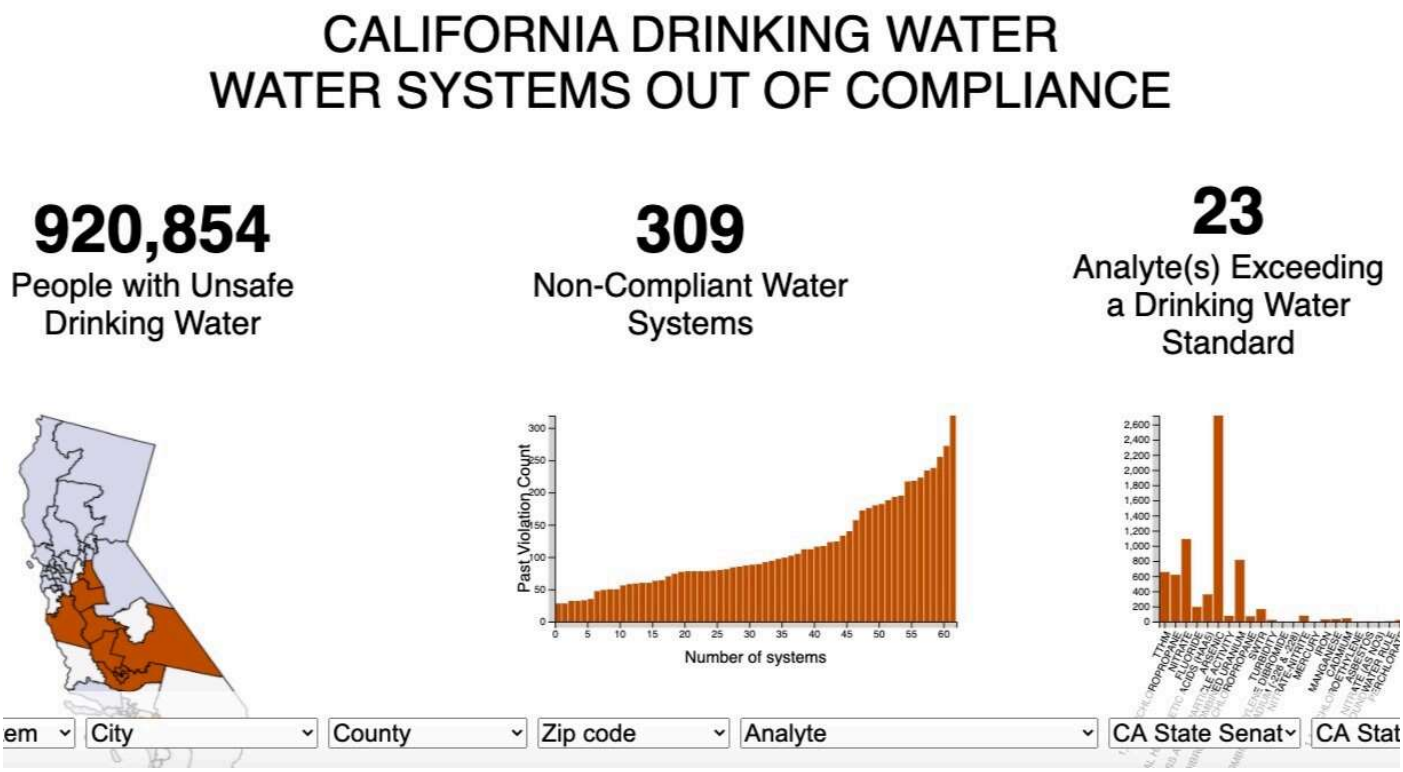
This data is presented every month from 2022 to 2024. It will used to show trends in the counties water levels.

The variables of interest with the projection dataset (`five_year_shortage`) are:

- `water_use_acre_feet`: projected water used for each agency
- `water_supplies_acre_feet`: projected water supply for each agency
- `benefit_supply_augmentation_acre_feet`: projected supply augmentated (bought) for each agency
- `benefit_demand_reduction_acre_feet`: projected demand reduced for each agency

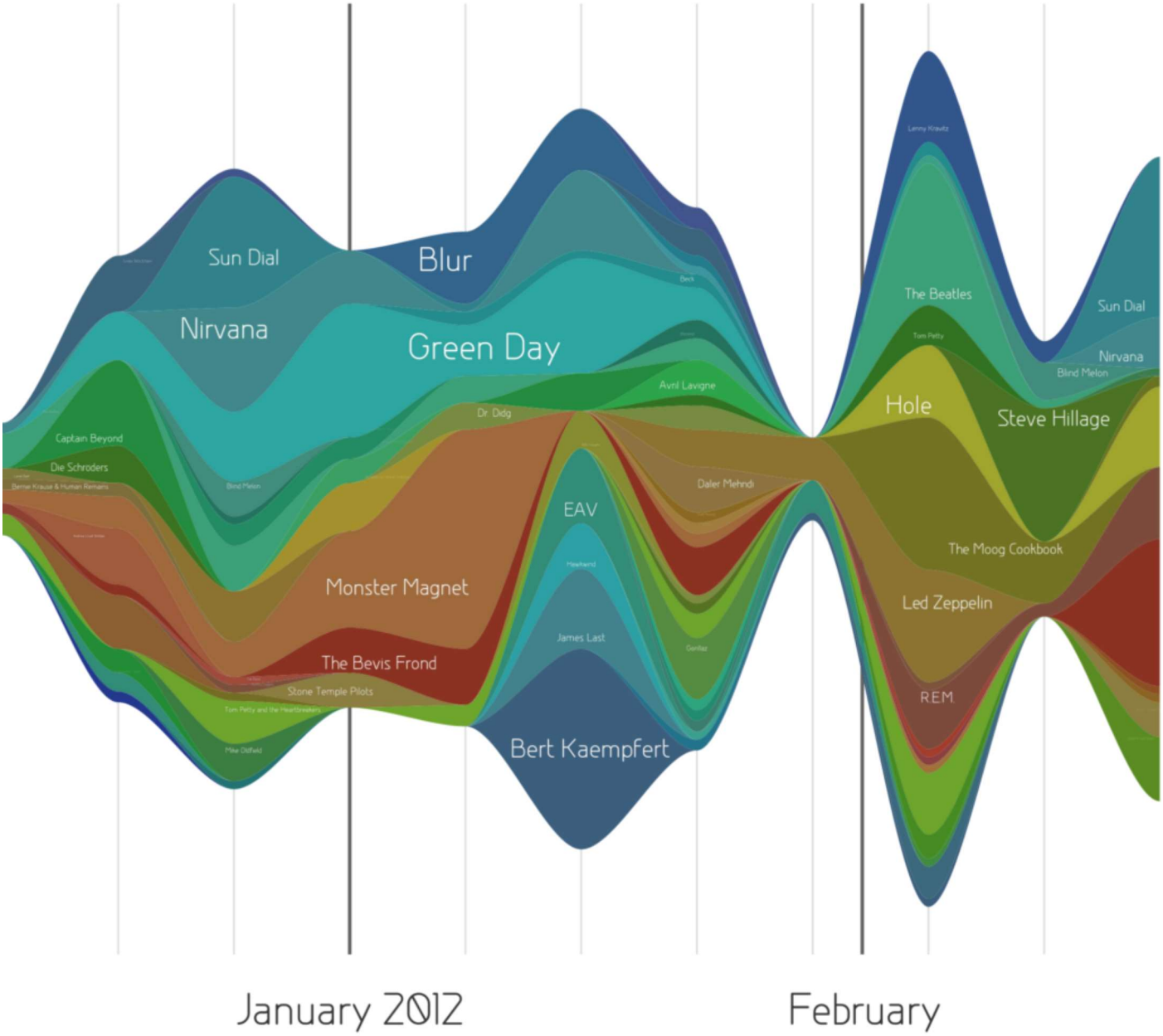
These variables are projections reported from individual water agencies during the 2020 Urban Water Management Plan submission. They are representative of the worst consecutive 5 years on average. Each row is an projected observation for years 2021 - 2025 because this data was submitted in 2020.

4. Find at least two data visualizations



Visualization 1

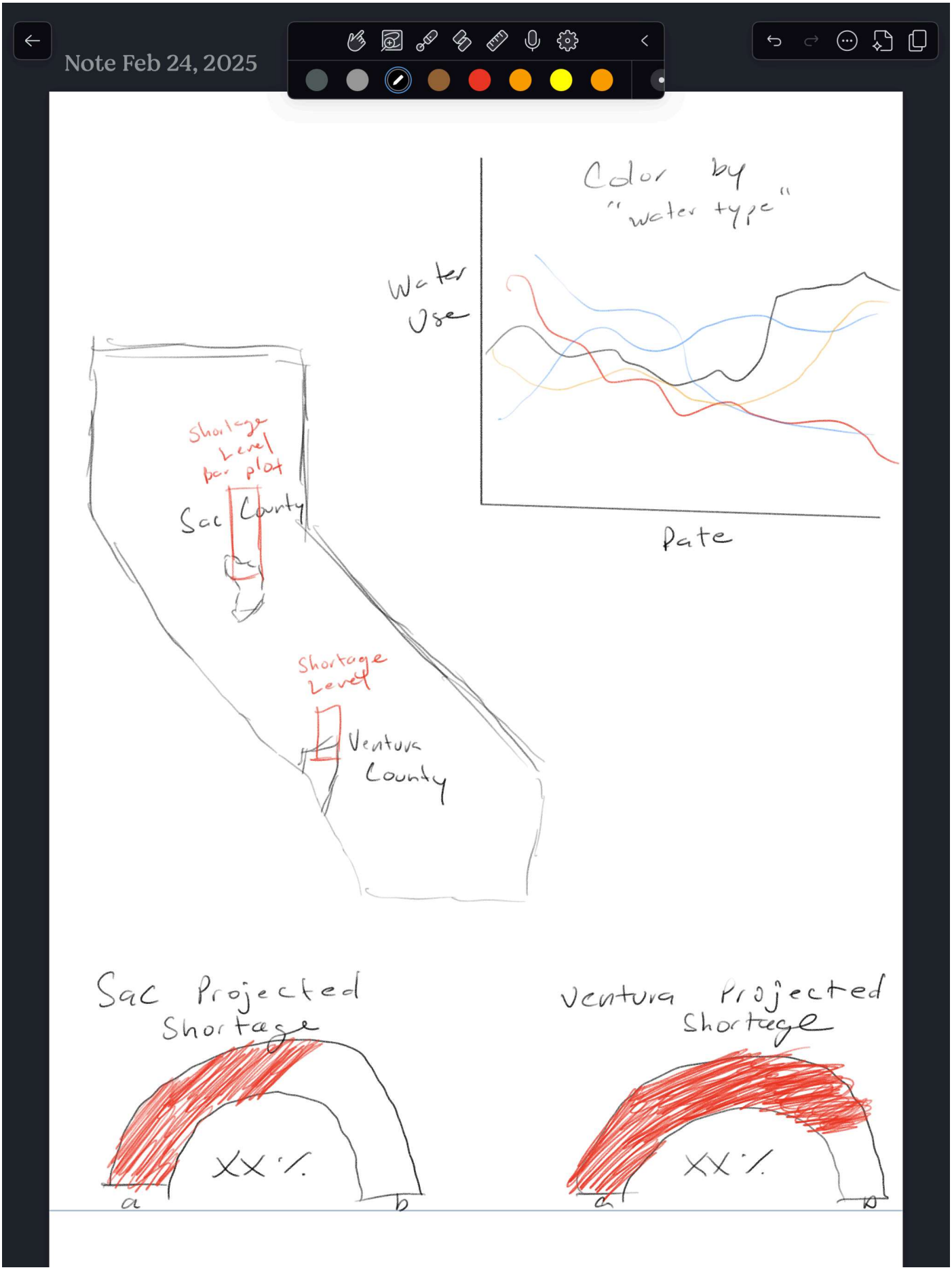
This is a great visualization. It really brings home the point that simpler can be better. It has some easy to read information and it drives the point home that there are many people out there that don't have safe drinking water. I think producing something along this line can be affective. Of course, I want to try and produce something with a little more pizzazz as the examples we have seen from class are great.



Visualization 2

This is an interesting plot for time series data. makes it has a little more

5. Draw anticapated visualizations



Hand Drawn Visualizations

6. Mock up visualizations using code

- have your data plotted (if you're experimenting with a graphic form(s) that was not explicitly covered in class, we understand that this may take some more time to build; you should have as much put together as possible)
- use appropriate strategies to highlight / focus attention on a clear message
- include appropriate text such as titles, captions, axis labels
- experiment with colors and typefaces / fonts
- create a presentable / aesthetically-pleasing theme (e.g. (re)move gridlines / legends as appropriate, adjust font sizes, etc.)

Let's code it

Load libraries

```
# Load libraries
library(tidyverse)
library(here)
library(spnaf)
library(stringr)
library(ggExtra)
library(tmap)
library(sf)
library(janitor)
library(ggridges)
```

Load data

```
# Main data
water_shortage <- read_csv(here("data", "actual_water_shortage_level.csv"))
five_year_shortage <- read_csv(here("data", "five_year_water_shortage_outlook.csv"))
historical_production <- read_csv(here("data", "historical_production_delivery.csv"))
population <- read_csv(here("data", "population_clean.csv"))

# Map boundary data
ca_counties <- st_read(here("data", "ca_counties", "CA_Counties.shp"))
```

```
Reading layer `CA_Counties' from data source
  `C:\MEDS\EDS-240\Gibbens-Matsuyama-eds240-HW4\data\ca_counties\CA_Counties.shp'
  using driver `ESRI Shapefile'
Simple feature collection with 58 features and 19 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: -13857270 ymin: 3832931 xmax: -12705030 ymax: 5162404
Projected CRS: WGS 84 / Pseudo-Mercator
```

```
ca_boundary <- st_read(here("data", "ca_state", "CA_State.shp"))
```

```
Reading layer `CA_State' from data source
  `C:\MEDS\EDS-240\Gibbens-Matsuyama-eds240-HW4\data\ca_state\CA_State.shp'
  using driver `ESRI Shapefile'
Simple feature collection with 1 feature and 18 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: -13857270 ymin: 3832931 xmax: -12705030 ymax: 5162406
Projected CRS: Popular Visualisation CRS / Mercator
```

Plot 1

This plot is the water shortage and surplus for both counties

▼ Show the code

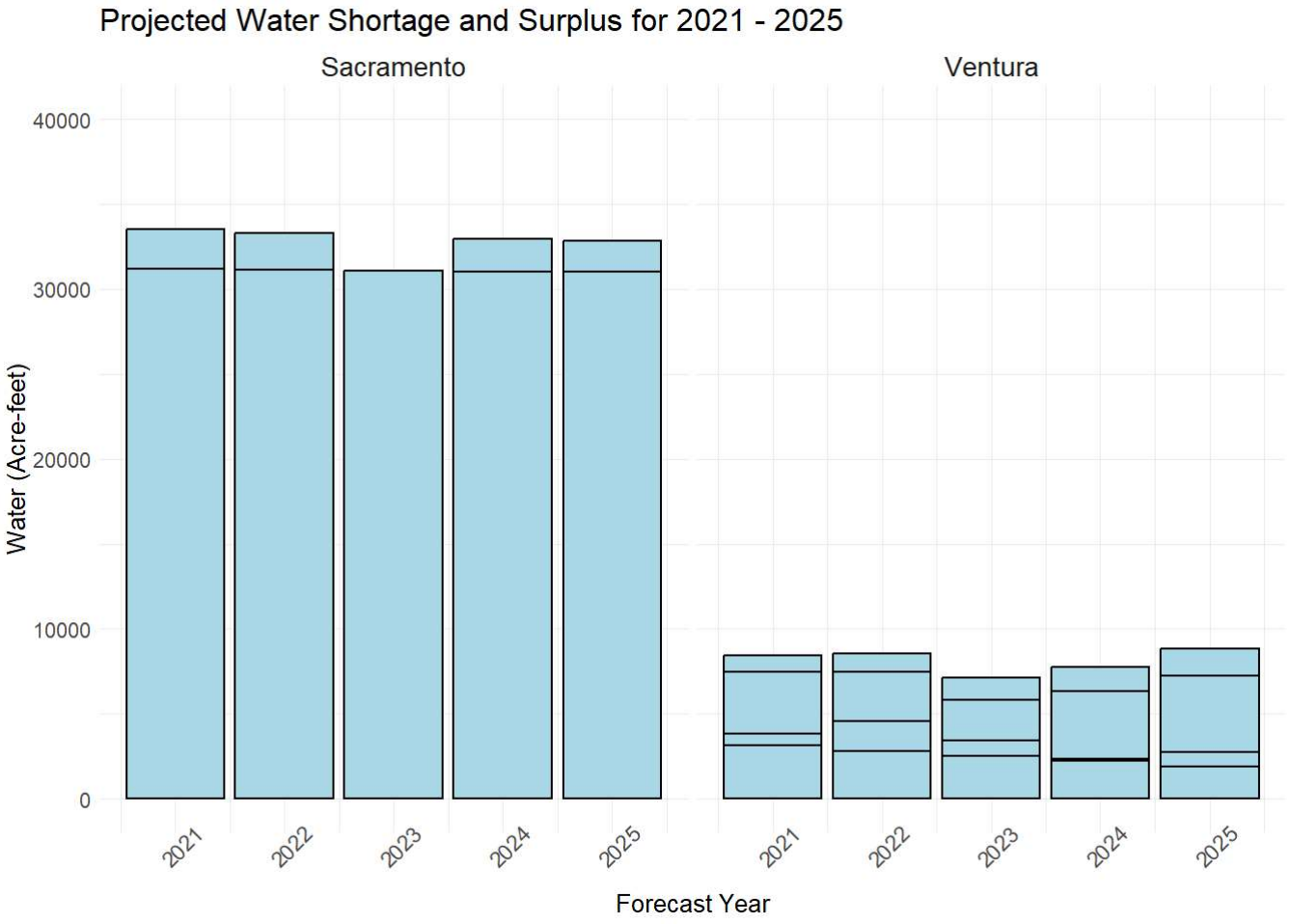
```
# Five year data, filter to Ventura and Sacramento County water agencies
five_filtered <- five_year_shortage %>%
  filter(org_id %in% c(376, 2158, 2629, 2631, 372, 2132, 2140, 2683, 2130))

# Create a Shortage/Surplus column
five_filtered$difference <- five_filtered$water_supplies_acre_feet - five_filtered$water_use_acre_

# Mutate new column for Ventura and Sacramento
five_filtered <- five_filtered %>%
mutate(county = case_when(
    str_detect(supplier_name, fixed("ventura", ignore_case = TRUE)) ~ "Ventura",
    str_detect(supplier_name, fixed("sacramento", ignore_case = TRUE)) ~ "Sacramento",
    TRUE ~ "other"
  ))

# Create ggplot of shortage/surplus for sac and ventura
ggplot(five_filtered, aes(x = forecast_start_date, y = difference,
                          fill = case_when(
                            difference > 0 ~ "lightblue",
                            difference < 0 ~ "firebrick"
                          ),
                          color = "black")) +
  geom_col() +
  facet_wrap(~county) +
    scale_fill_manual(values = c("lightblue" = "lightblue",
                                 "firebrick" = "firebrick")) +
  scale_color_manual(values = c("black" = "black")) +
  scale_y_continuous(limits = c(0, 40000),
                     breaks = seq(-10000, 40000, by = 10000)) +
    labs(x = "Forecast Year",
       y = "Water (Acre-feet)",
       title = "Projected Water Shortage and Surplus for 2021 - 2025") +
       #subtitle = "Projections for randomly selected water agencies in California") +
```

```
  theme_minimal(base_size = 7)+
  theme(
    legend.position = "none",
    plot.title = element_text(size = rel(1.7)),
    plot.subtitle = ggtext::element_textbox(size = rel(1.25),
                                            margin = margin(t = 2, r = 0,
                                                            b = 6, l = 0),
                                            padding = margin(t = 5, r = 0,
                                                             b = 5, l = 0)),
    axis.text = element_text(size = rel(1.2)),
    axis.text.x = element_text(angle = 45),
    axis.title.y = element_text(size = rel(1.4)),
    axis.title.x = element_text(size = rel(1.4)),
    strip.text = element_text(size = rel(1.5))
  )
```



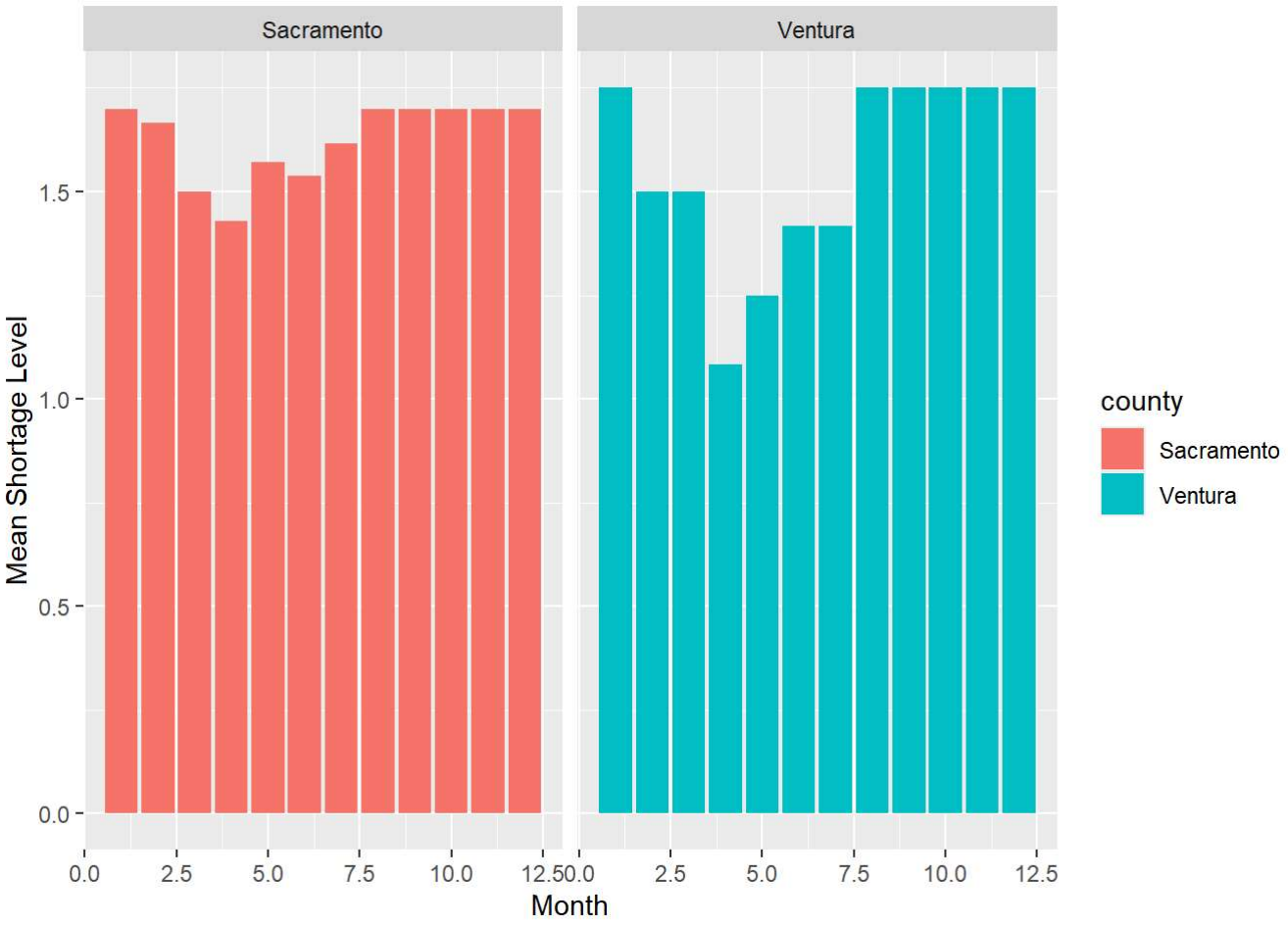
Plot 2

This plot is the water shortage level for both counties

▼ Show the code

```r
# Filter to Ventura and Sacramento County Water agencies
level_filtered <- water_shortage %>%
  filter(org_id %in% c(376, 2158, 2629, 2631, 372, 2132, 2140, 2683, 2130)) %>%
  mutate(county = case_when(
    str_detect(supplier_name, fixed("ventura", ignore_case = TRUE)) ~ "Ventura",
    str_detect(supplier_name, fixed("sacramento", ignore_case = TRUE)) ~ "Sacramento",
    TRUE ~ "other"
  )) %>%
  filter(!is.na(state_standard_shortage_level)) %>%
  mutate(year = year(start_date),
         month = month(start_date)) %>%
  group_by(county, month) %>%
  summarise(mean_level = mean(state_standard_shortage_level))


# Create bar plot
ggplot(level_filtered, aes(x = month, y = mean_level, fill = county)) +
  geom_col() +
  facet_wrap(~county) +
  labs(x = "Month",
       y = "Mean Shortage Level")
```



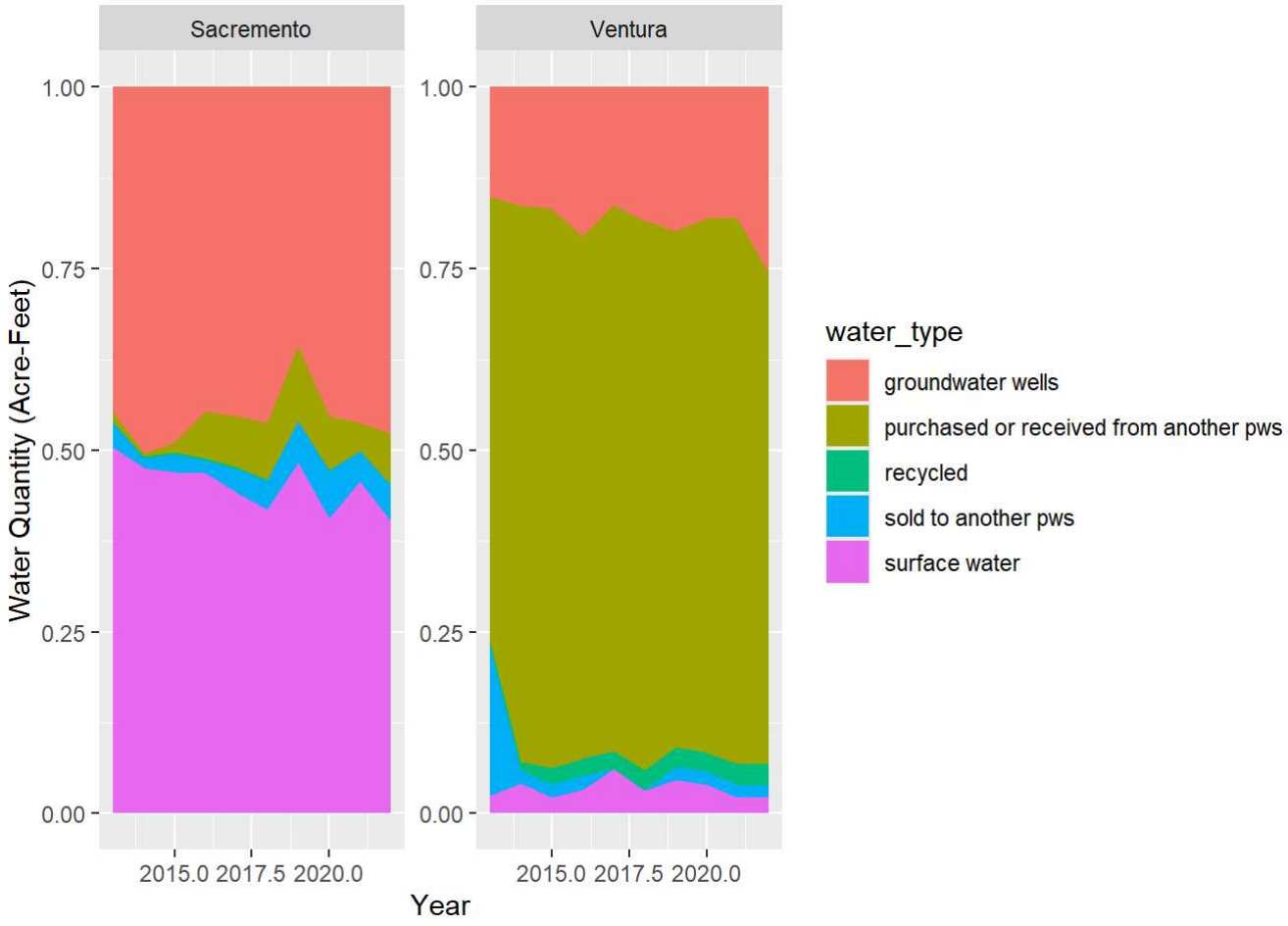
Plot 3: Date and Water Type Line Plot

This plot is the quantity of water types that each county produces.

▼ Show the code

```
# Filter data to
historical_filter2 <- historical_production %>%
  filter(org_id %in% c(376, 2158, 2629, 2631, 372, 2132, 2140, 2683, 2130),
         !is.na(quantity_acre_feet),
         water_produced_or_delivered == "water produced",
         !water_type %in% c("non-potable (total excluded recycled)", "non-potable water sold to a
    mutate(county = case_when(
    org_id %in% c(376, 2158, 2629, 2631) ~ "Ventura",
    org_id %in% c(372, 2132, 2140, 2683, 2130) ~ "Sacremento",
    TRUE ~ "other"
  )) %>%
  group_by(start_date, water_produced_or_delivered, water_type, county) %>%
  summarise(quantity = sum(quantity_acre_feet)) %>%
  ungroup() %>%
  mutate(year = year(start_date)) %>%
  group_by(year, county, water_type) %>%
  summarise(quantity = sum(quantity)) %>%
  group_by(year, county) %>%
  mutate(total_quantity = sum(quantity),
         proportion = quantity / total_quantity) %>%
  ungroup()

# Area plot for proportions
ggplot(historical_filter2, aes(x = year, y = proportion, fill = water_type)) +
  geom_area() +
  facet_wrap(~county, scales = "free") +
  scale_x_continuous(limits = c(2013, 2022)) +
  labs(x = "Year",
       y = "Water Quantity (Acre-Feet)")
```

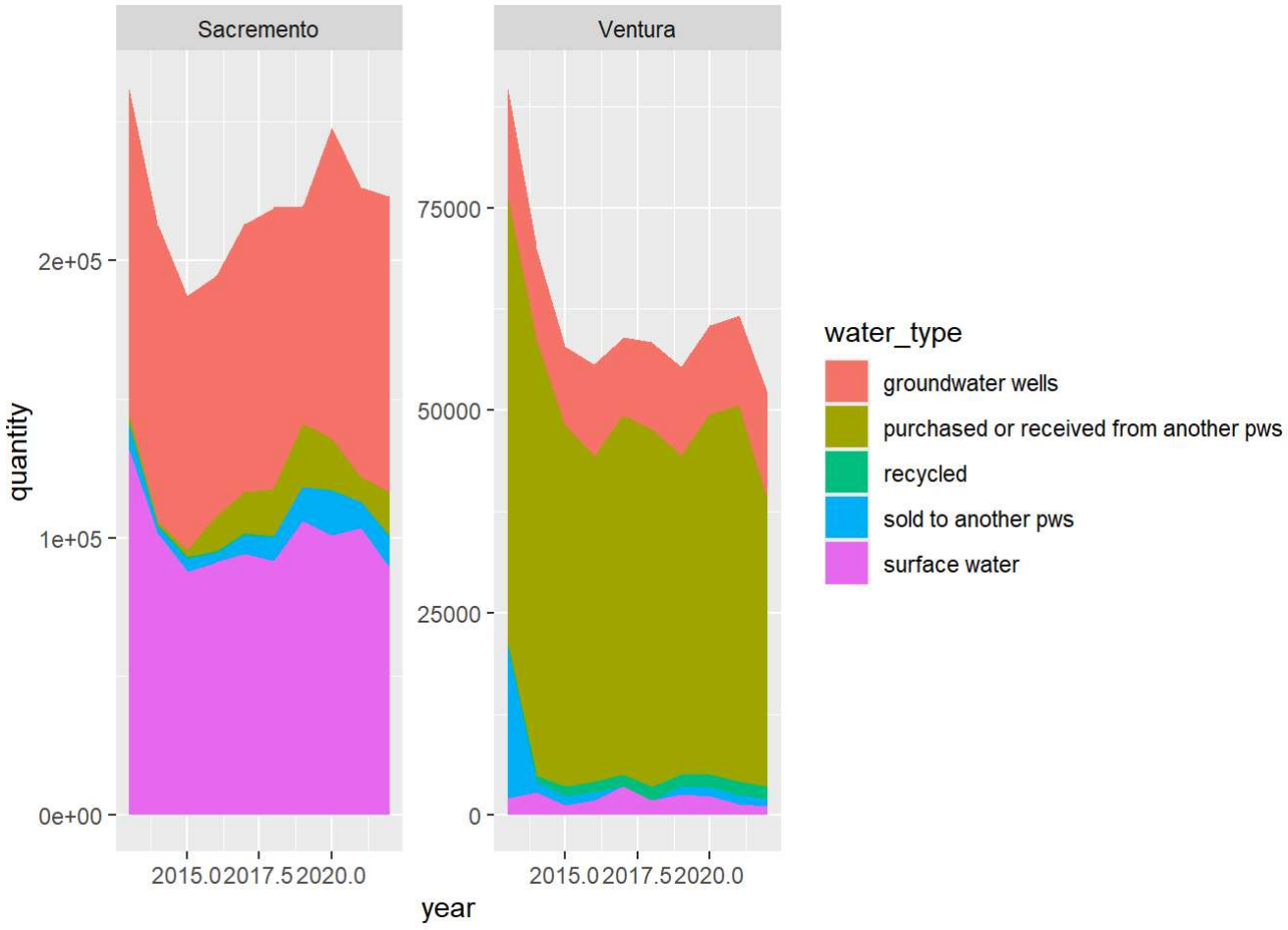


▼ Show the code

```
  theme(
    legend.position = "none"
  )
```

```
List of 1
 $ legend.position: chr "none"
 - attr(*, "class")= chr [1:2] "theme" "gg"
 - attr(*, "complete")= logi FALSE
 - attr(*, "validate")= logi TRUE
```

▶ Show the code



Map

▼ Show the code

```
ca_counties <- ca_counties %>%
  clean_names() %>%
  mutate(color = case_when(
    name == "Ventura" ~ "lightblue",
    name == "Sacramento" ~ "firebrick",
    TRUE ~ "gray50"
  ))

tm_shape(ca_boundary) +
  tm_borders() +
  tm_shape(ca_counties) +
  tm_polygons(fill = "color")
```

7.

a. What challenges did you encounter or anticipate encountering as you continue to build / iterate on your visualizations in R? If you struggled with mocking up any of your three visualizations (from #6, above), describe those challenges here.

I had some trouble with the data wrangling part of this. I changed my question from my previous assignments and had to start over. The wrangling took me some time. Unfortunately, it doesn't look like I did much because my outputs are so ugly. The line graph that I originally drawn on my sketch is not how I pictured it. First, I switch it from a line graph to an area plot. I am not sure as to which one is better as I am not a fan of either. The tough part is that I am comparing two counties so I needed to facet the plot. I need to figure out a way to make this presentable on an infographic. I planned on having the bar plot pop up on the map as a single bar, but I don't think that would do much. Even the one I have now is not very informative. On the sketch, I wanted to have the projected shortage percentage for both on the bottom, but neither has shortage from their projections.

b. What ggplot extension tools / packages do you need to use to build your visualizations? Are there any that we haven't covered in class that you'll be learning how to use for your visualizations?

Everything I've done is from class. I wanted to make a streamgraph from the streamgraph library but I was running into some issues. I scratched it and just used an area plot instead.

c. What feedback do you need from the instructional team and / or your peers to ensure that your intended message is clear? The question I want to answer is which county is at greater risk of drought? From the data presented, I am not sure if I am able to answer this question. Having shortage levels and projections can tell us some information but does it really answer this question? The one take away from the area plot Ventura buys a lot of its water, so maybe I could trace it back to the source.