Problem Set 3

[Glacial Lakes]. The data at this link contain labels of glacial lakes the Hindu Kush Himalaya, created during an ecological survey in 2015 by the International Centre for Integrated Mountain Development.

a. How many lakes are in this dataset? What are the latitude / longitude coordinates of the largest lakes in each Sub-basin?

3624 lakes are in this dataset.

```
library(tidyverse)
library(sf)
library(tmap)

data <- read_sf('data/GL_3basins_2015.dbf')
nrow(data)
```

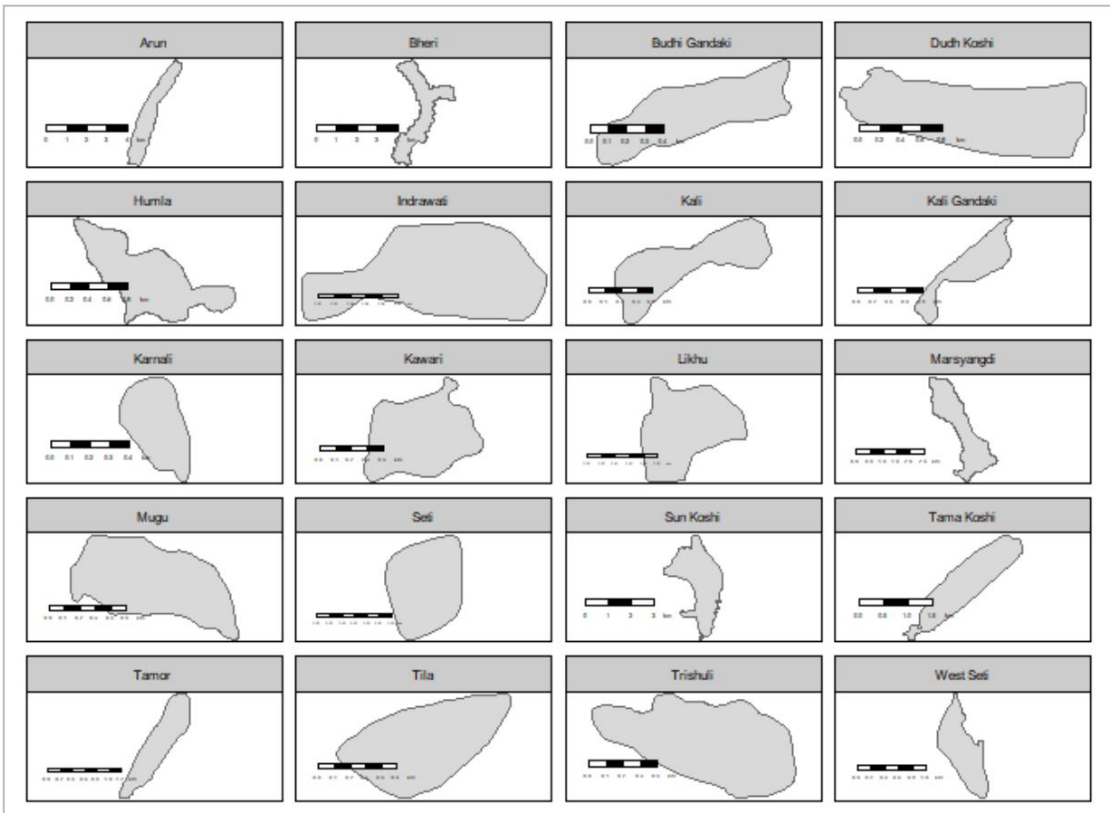
	Latitude		Longitude
	<dbl>		<dbl>
1	28.4	1	86.3
2	28.3	2	85.8
3	27.9	3	86.9
4	27.9	4	87.9
5	27.9	5	86.4
6	28.0	6	85.7
7	27.7	7	86.5
8	29.2	8	82.9
9	29.4	9	82.4
10	29.6	10	81.6
11	29.8	11	82.4
12	29.8	12	81.5
13	29.9	13	81.6
14	30.1	14	81.8
15	30.6	15	80.2
16	28.4	16	84.1
17	28.5	17	85.5
18	28.6	18	84.6
19	28.7	19	83.9
20	29.2	20	83.7

```
largest <- data %>%
  group_by(Sub_Basin) %>%
  top_n(1,Area)

largest['Latitude']
largest['Longitude']
```

b. Plot the polygons associated with each of the lakes identified in step (a). Hint: To make the

polygons legible, use wrap_plots in the patchwork package to combine separate plots for each lake.

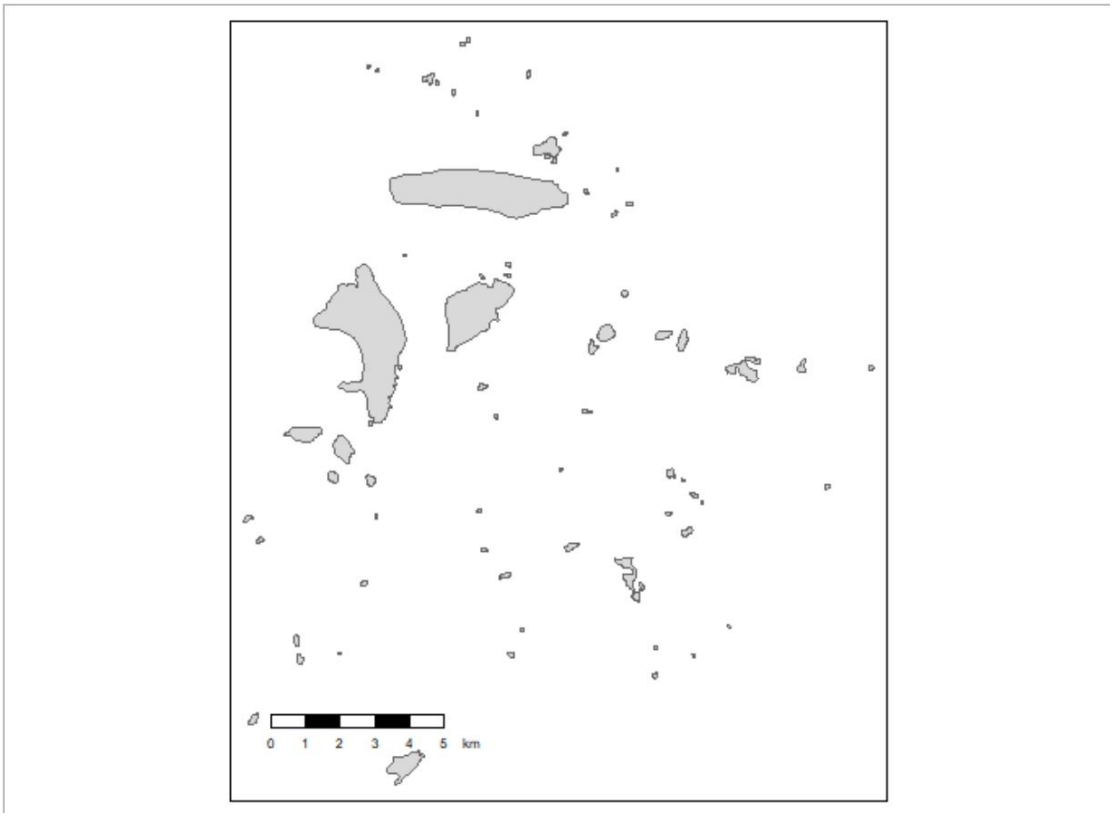



```
tm_shape(largest) +
   tm_polygons(title='lake',size = 0.1) +
   tm_layout(legend.outside = TRUE)+
   tm_facets("Sub_Basin") +
   tm_scale_bar(position = c("left", "bottom"))
```

b. Visualize the lakes with latitude between 28.2 and 28.4 and with longitude between 85.8 and 86. Optionally, add a basemap associated with each lake.

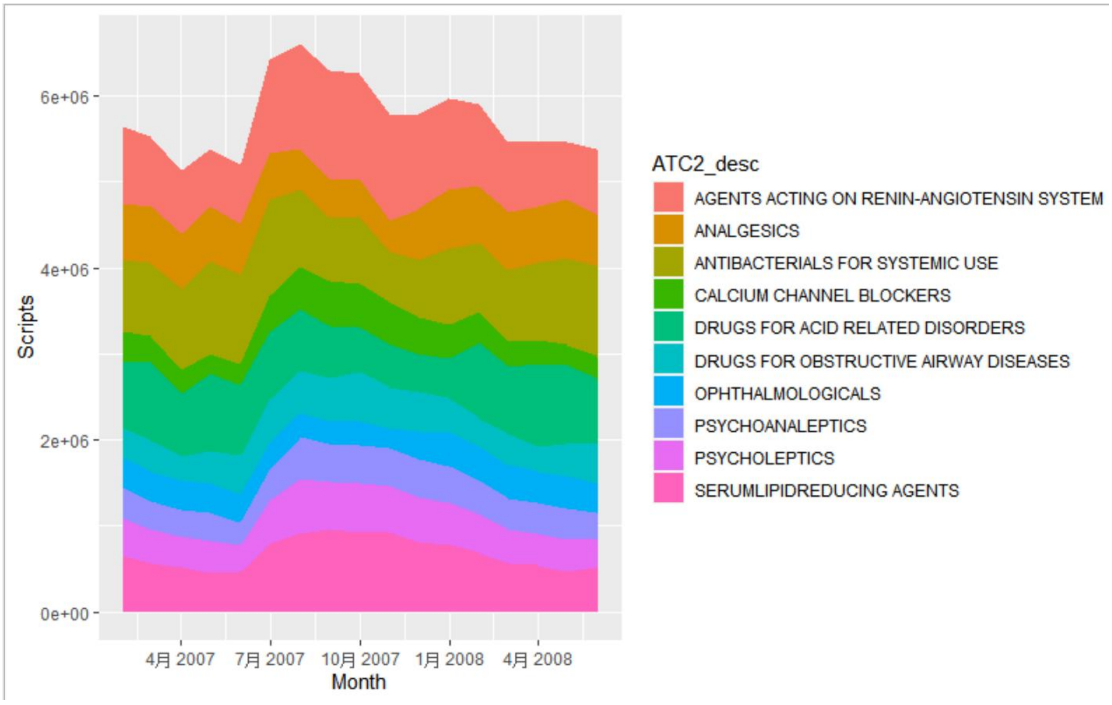
```
criteria <- data %>%
   filter(Latitude > 28.2 & Latitude < 28.4 & Longitude > 85.8 & Longitude < 86)

tm_shape(criteria) +
   tm_polygons(title='lake',size = 0.1)+
   tm_layout(legend.outside = TRUE)+
   tm_scale_bar(position = c("left", "bottom"))
```



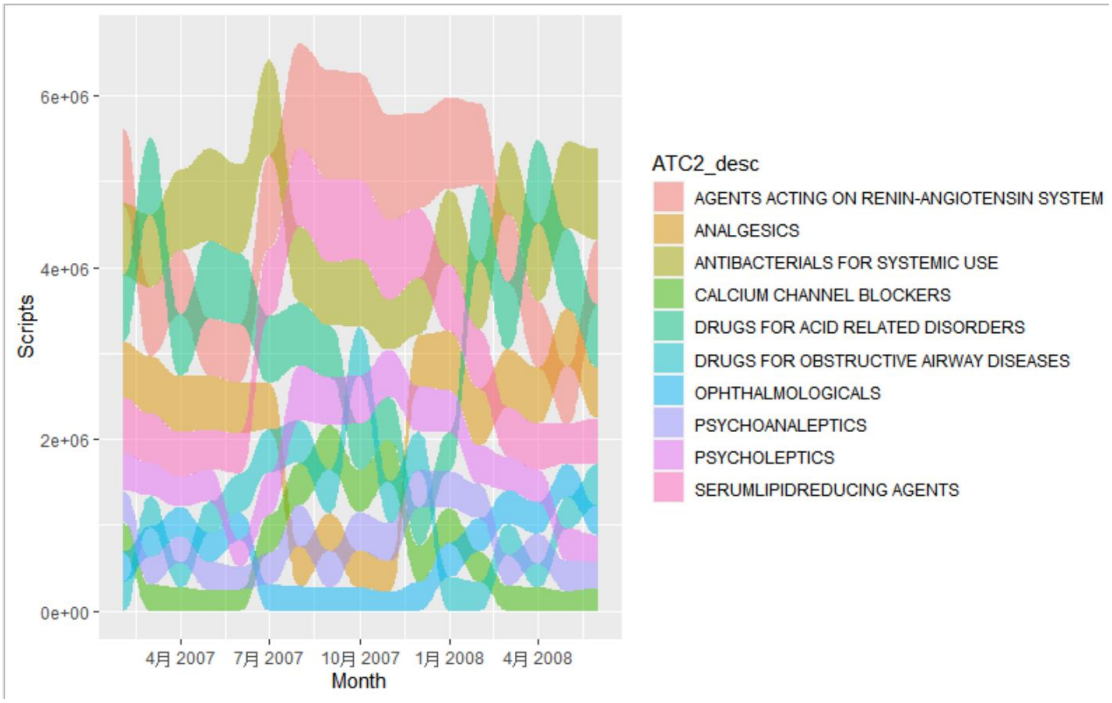
[Australian Pharmaceuticals II] The code below takes the full PBS dataset from the previous problem and filters down to the 10 most commonly prescribed pharmaceutical types. This problem will ask you to implement and compare two approaches to visualizing this dataset.

a. Implement a stacked area visualization of these data.



```
ggplot(pbs) +
  geom_area(aes(Month, Scripts,fill = ATC2_desc))
```

b. Implement an alluvial visualization of these data.



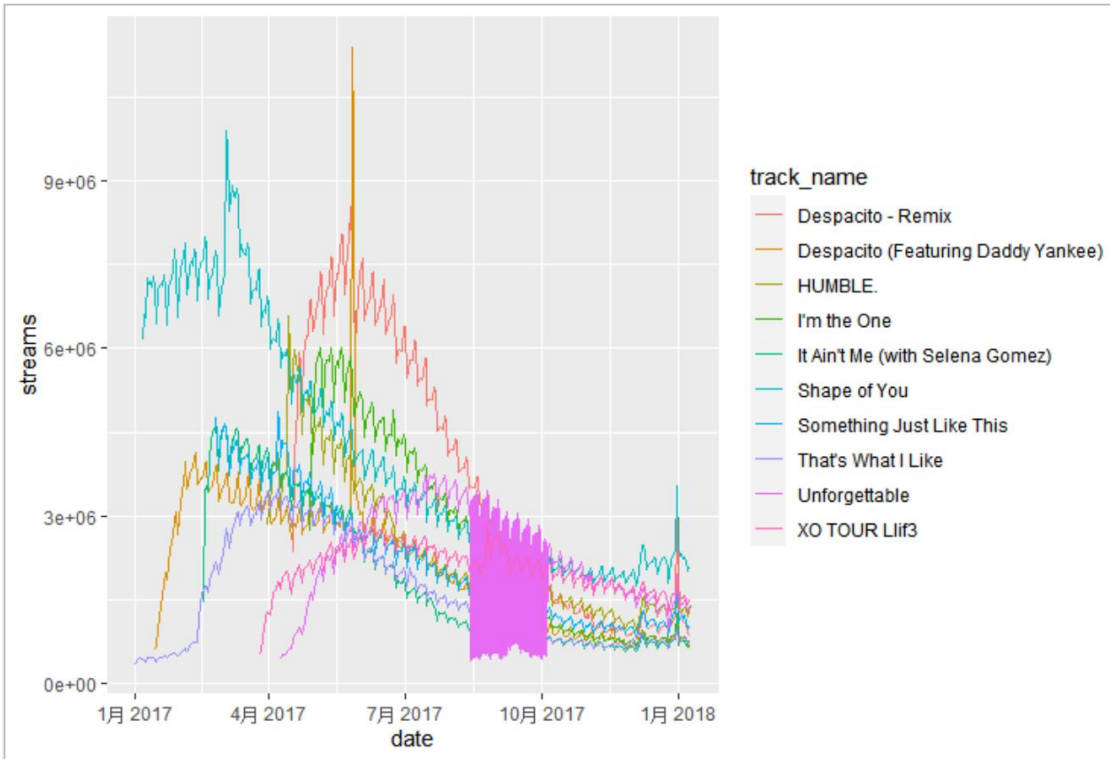

```
ggplot(pbs) +
  geom_alluvium(aes(Month, Scripts,fill = ATC2_desc,alluvium = ATC2_desc),decreasing = FALSE)
```

c. Compare and contrast the strengths and weaknesses of these two visualization strategies. Which user queries are easier to answer using one approach vs. the other?

```
Stacked area graphs make it easier to tell how the totals at any timepoint breakdown across groups. While alluvial plot explicitly rank the groups, because at each time point, the groups are sorted from the largest to smallest contribution.
```

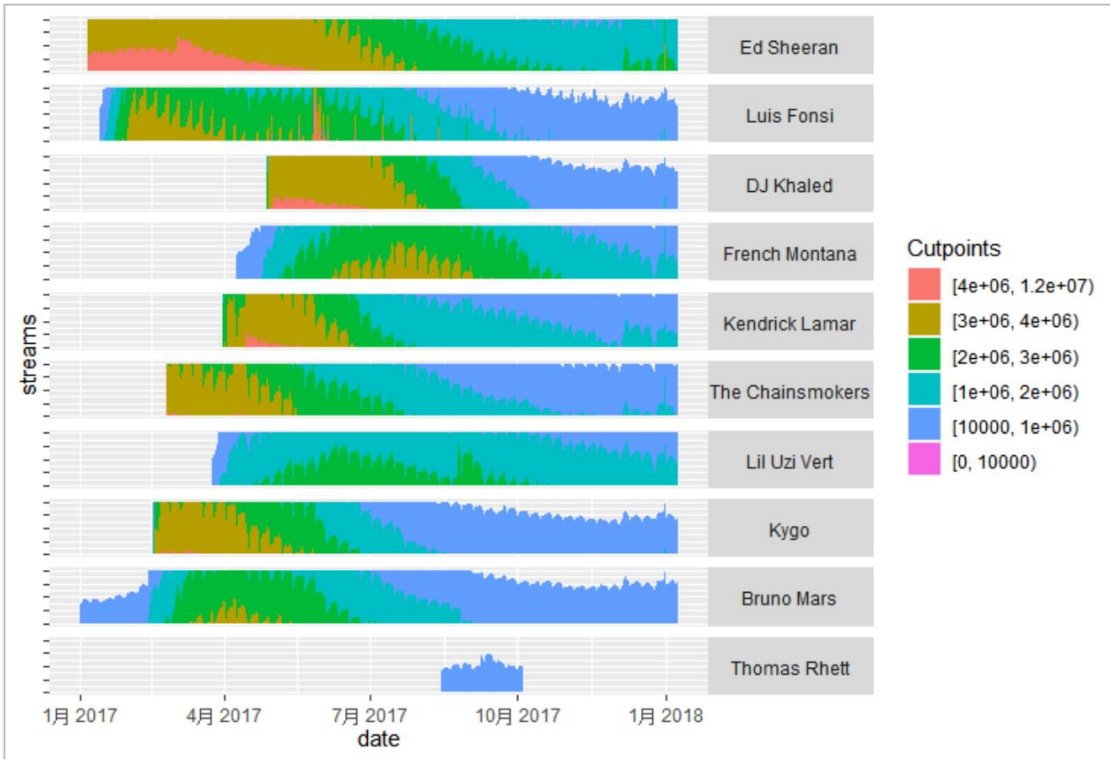
[Spotify Time Series II] The code below provides the number of Spotify streams for the 40 tracks with the highest stream count in the Spotify 100 dataset for 2017. This problem will ask you to explore a few different strategies that can be used to visualize this time series collection.
For a better view of the graph, I only chose top 10 of the data.

a. Design and implement a line-based visualization of these data.




```
ggplot(spotify) +
   geom_line(aes(date,streams,col = track_name)) +
   theme(legend.position = 'right')
```

b. Design and implement a horizon plot visualization of these data.




```
max_stream <- max(spotify$streams)
cutpoints <- c(0, 1000000, 2000000, 3000000, 4000000, 12000000)
ggplot(spotify) +
   geom_horizon(aes(date, streams, fill = ..Cutpoints..), origin=10000, horizonscale = cutpoints)+
   facet_grid(reorder(artist,-streams) ~ .)+
   theme(strip.text.y = element_text(angle=0), axis.text.y = element_blank())
```

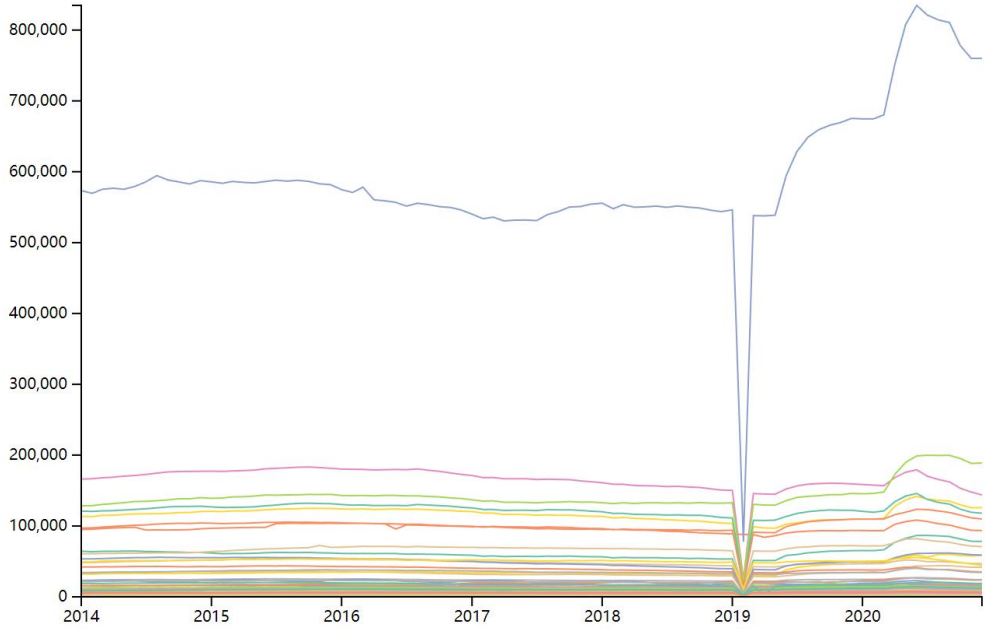
c. Building from the static views from (a - b), propose, but do not implement, a visualization of this dataset that makes use of dynamic queries. What would be the structure of interaction, and how would the display update when the user provides a cue? Explain how you would use one type of D3 selection or mouse event to implement this hypothetical interactivity

```
For both (a) and (b), it's hard to tell the accurate figure at a specific timepoint. I'd like to implement an interaction like this: when use's mouse is hovering on the graph, the accurate date and number of streams will show up.

In D3, I would use .property function and change the sub dataset accordingly and then render the plot.
```

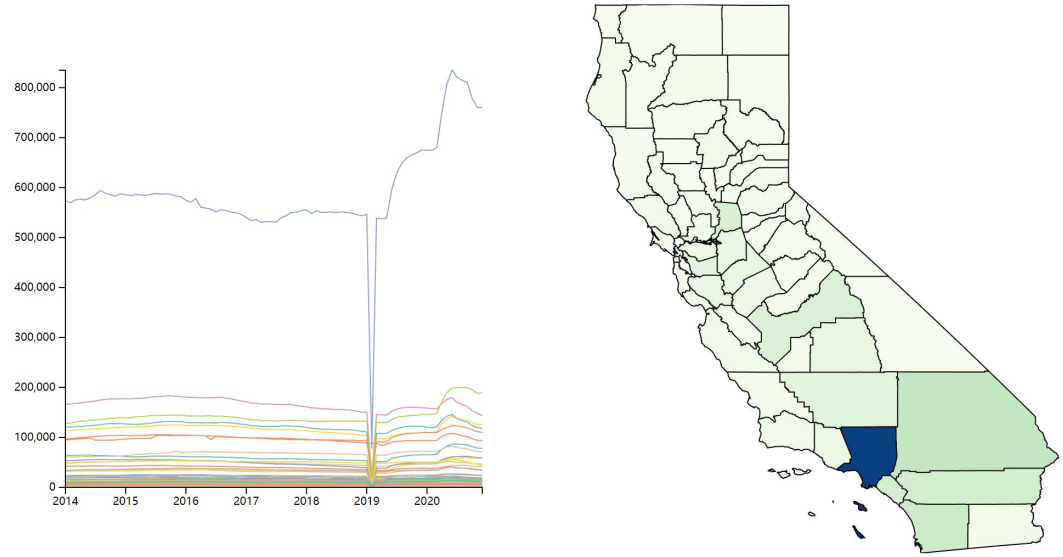
[CalFresh Enrollment II] In this problem, we will develop an interactively linked spatial and temporal visualization of enrollment data from CalFresh, a nutritional assistance program in California. We will use D3 in this problem.

a. Using a line generator, create a static line plot that visualizes change in enrollment for every county in California.




```
See a.html and a.js
```

b. On the same page as part (a), create a choropleth map of California, shaded in by the average enrollment for that county over the full time window of the dataset.




```

See b.html and b.js
```

C. Propose one interactive, graphical query the links the combined spatial + temporal view from (a - b). For example, you may consider highlighting time series when a county is hovered, highlighting counties when specific series are brushed, or updating choropleth colors when a subsetted time range is selected.

```

I wanna propose a design that, when the mouse is hovering over the geographical part, the corresponding temporal data will be highlighted. And vice versa, when the mouse is hovering over the temporal graph, the correspongding part in the geographical data will be highlighted.
```

d. Implement your proposal from part (c).

```

See d.html and d.js
```

Temporal and Geospatial Commands

A. geom_stream

```

It adds a layer to the ggplot() function. The display for temporal data looks like a stacked area chart, but oriented around a central line.
A specific example is, when given a time series dataset for population of each different,
```

geom_stream function can show the change of population of different countries by time, and makes it easier to tell how the totals at any timepoint breakdown across groups.
```

B. tm_lines

```
It adds on a layer for tm_shape() function for geospatial vector data. Basically, this function use spatial data to draw spatial lines.
A specific example is, to draw the road network, we can add tm_lines() on tm_shape().
```

C. rast

```
This function was mainly used for raster data. It creates a SpatRaster. We can add more layers on it.
A specific example is, we can read in satellite raster dataset with rast() function, and then add layer with functions like tm_raster() on tm_shape().
```

D. geom_horizon

```
It adds a layer to the ggplot() function. The display for temporal data are collapsed on top of one another. In this way, even though there may be a large range in the realized y-values, the amount of space needed to display them all is relatively small.

A specific example is, when given a time series dataset for tourist population of different attraction sites, even if the volume at some time point is several times higher than other time points, geom_horizon function can show the change across attraction sites by time in a relatively small space.
```

[Visualization for Everyday Decisions] We routinely encounter spatial and temporal data in everyday life, from the from the dates on a concert calendar to the layout of buttons in an elevator. This problem asks you to critically reflect on the way these data are represented.

a. Describe one type of spatial or temporal dataset (loosely defined) that you encounter in a nonacademic context. What visual encodings are used? Does it implement any interactivity?

```
```
```

https://www.nbcnews.com/health/article/covid-hospitalization-hot-spots-across-u-s-five-charts-n1285965

This page shows Covid hospitalizations by state. The first graph shows Covid hospitalizations per 100,000 residents, and the second graph shows Change in hospitalizations over the last two weeks. The third graph shows share of hospitals reporting critical staffing issues and the fourth graph shows percentage of hospital beds in use.

The interactivity part is, when you hover your mouse on a specific state, it shows more specific Covid hospitalization data for this state.
```

b. What questions does the visualization help you answer? How easily can you arrive at an accurate answer?

```
I can tell how capable/overwhelmed a state is for Covid and how severe the Covid issue is by the four graphs provided.

Colors were used to represent the data, but if you hover your mouse on specific state, accurate data will show up. So it's pretty easy to arrive at an accurate answer.
```

c. In the spirit of the re-imagined parking sign project, propose an alternative representation of these data (or an alternative way of interacting with the same representation). Why did you propose the design that you did? What advantages do you think it has?

```
I feel 4 graphs make the article long and dull to read. So I would rather combine them into one graph, with a select input on the top for 4 type of options. The interactivity part is good, so I will keep them as what they were.

My design would be simpler and cleaner to read. And when users are encouraged to try different options, they will tend to think more towards the problem.
```
```