# K12 data tools vignette: where are the lobsters?

Nicholas Horton (nhorton@amherst.edu), Danny Pimentel, and Michelle Wilkerson

August 29, 2022

## Contents

**Overview** **1**

**Motivating data** **1**

**CODAP** **2**

Load document or data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

Explore data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

**R** **7**

**Google Sheets** **11**

**Tableau** **13**

## Overview

To explore how the selection of data analysis tools can fundamentally shape what is highlighted in an investigation, we conducted a comparative analysis using several popular tools from multiple tool genres: R (scripting), CODAP (visual/pedagogical), Google Sheets (spreadsheet), and Tableau (visual/commercial).

The associated commissioned paper and supplementary materials can be found at https://nicholasjhorton.github.io/K12-Data-Tools.

## Motivating data

Using each tool, we loaded the same CSV (comma separated value) dataset focused on the recent historical migration patterns of the American Lobster (*Homarus americanus*) off the coast of Northern East Coast of the United States and Southern East Coast of Canada. These data are part of a much larger repository of data about aquatic species maintained by the OceanAdapt project (https://oceanadapt.rutgers.edu). They track position including the mean latitude, longitude, and depth of lobsters for different regions/seasons between the years of 1970 and 2020. The dataset is structured such that each row includes one observation record including `year`, `region`, `latitude`, `longitude`, and `depth`.

It is well established that American lobster populations have been moving north in recent years; examining this northward movement was the focus of our investigation. We selected this examination for the vignettes because it reflects a popular 'alternative' data type (geospatial data); it is deeply embedded in a specific context that connects to other disciplines (ecosystems); and there is a clear and consequential pattern to observe that emerges from relatively messy, multivariate data.

The source files for each of the tools explored here can be found at https://nicholasjhorton.github.io/K12-Data-Tools.

See https://codap.concord.org/app/static/dg/en/cert/index.html#shared=https%3A%2F%2Fcfm-shared.concord.org%2FkgGQg8IJkNr3c9825mnM%2Ffile.json for the CODAP (Common Online Data Analysis Platform) worksheet this filtered dataset came from.

See https://www.nationalacademies.org/our-work/foundations-of-data-science-for-students-in-grades-k-12-a-workshop for more information about the NASEM Foundations of Data Science for Students in Grades K-12 workshop.

### Executive summary of the motivating dataset

- the points indicate the centroid of the catch of lobsters for a particular year
- there is a separate point for each of the main regions/seasons
- there is a shift to the north
- this might indicate impacts of climate change

### Other notes about the analyses

- multivariate displays are helpful in communicating this result
- mapping is also important for this example, providing context in terms of geographic location
- certain data tools allow easy entry to the problem while others allow advanced visualizaton and analysis
- data science features many things, among them a variety of data types (e.g., maps, text, networks)

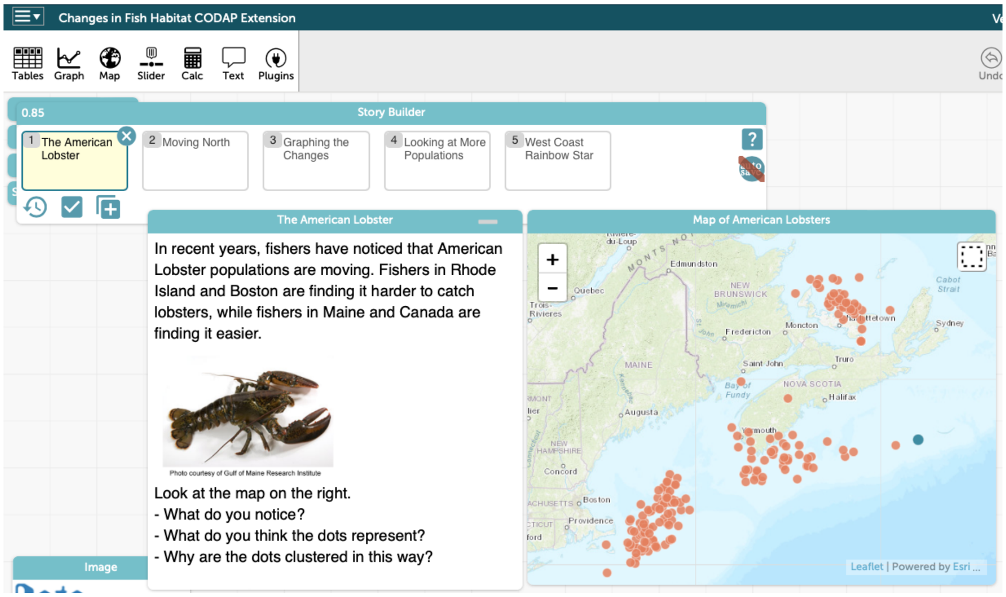
# CODAP

The recommended way to explore the CODAP (Common Online Data Analysis Platform) vignette is to open up the CODAP link (here) and start to explore.

The following figures provide highlights of what students would see as they work through this activity.

## Load document or data

Figure 1 displays the starting screen of the CODAP lobster activity.



In recent years, fishers have noticed that American Lobster populations are moving. Fishers in Rhode Island and Boston are finding it harder to catch lobsters, while fishers in Maine and Canada are finding it easier.

Look at the map on the right.
- What do you notice?
- What do you think the dots represent?
- Why are the dots clustered in this way?

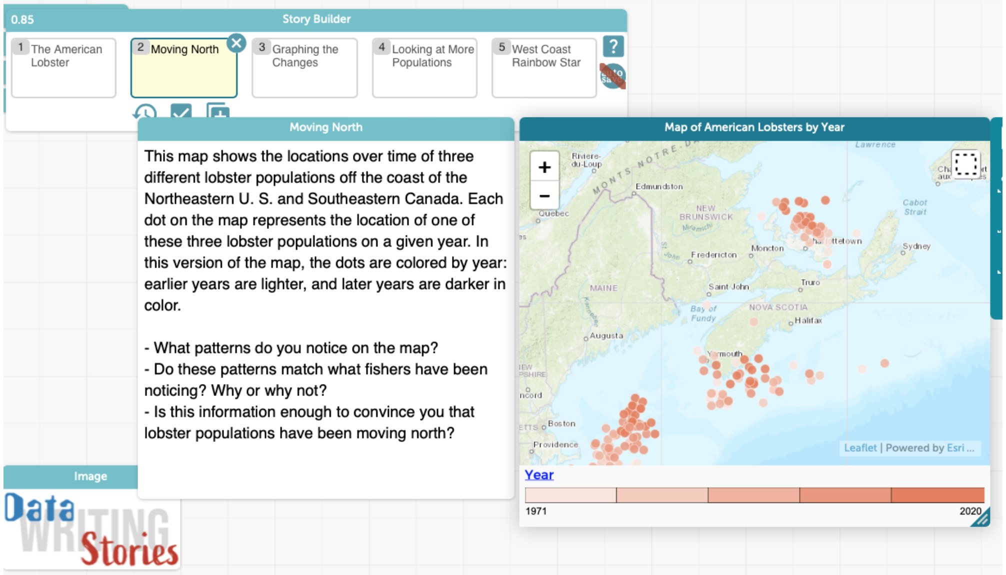
Figure 1: The opening screen from the CODAP activity which introduces the activity.

It includes five “moments” of a Writing Data Stories (WDS) module (https://concord.org/newsletter/2020-spring/writing-data-stories) to engage students in exploring the lobster data, and then comparing it to data collected about other aquatic species. The module features additional background information (and prompts for students), and starts with a map of the Northeast United States and Atlantic Canada. The fish data could also be freshly loaded by dragging and dropping the CSV file into a new blank CODAP document. Because the data includes clearly labeled latitude and longitude, opening a map will automatically display these observations.

## Explore data

### Explore map

Figure 2 steps through the next box in WDS Lobster Story module. Here the unit of observation is defined and the year of measurement is included as a variable used to shade the colors of the points.



This map shows the locations over time of three different lobster populations off the coast of the Northeastern U. S. and Southeastern Canada. Each dot on the map represents the location of one of these three lobster populations on a given year. In this version of the map, the dots are colored by year: earlier years are lighter, and later years are darker in color.

- What patterns do you notice on the map?
- Do these patterns match what fishers have been noticing? Why or why not?
- Is this information enough to convince you that lobster populations have been moving north?

Figure 2: Displaying the next step in the story module. Here the year of measurement is used to color points on the map.

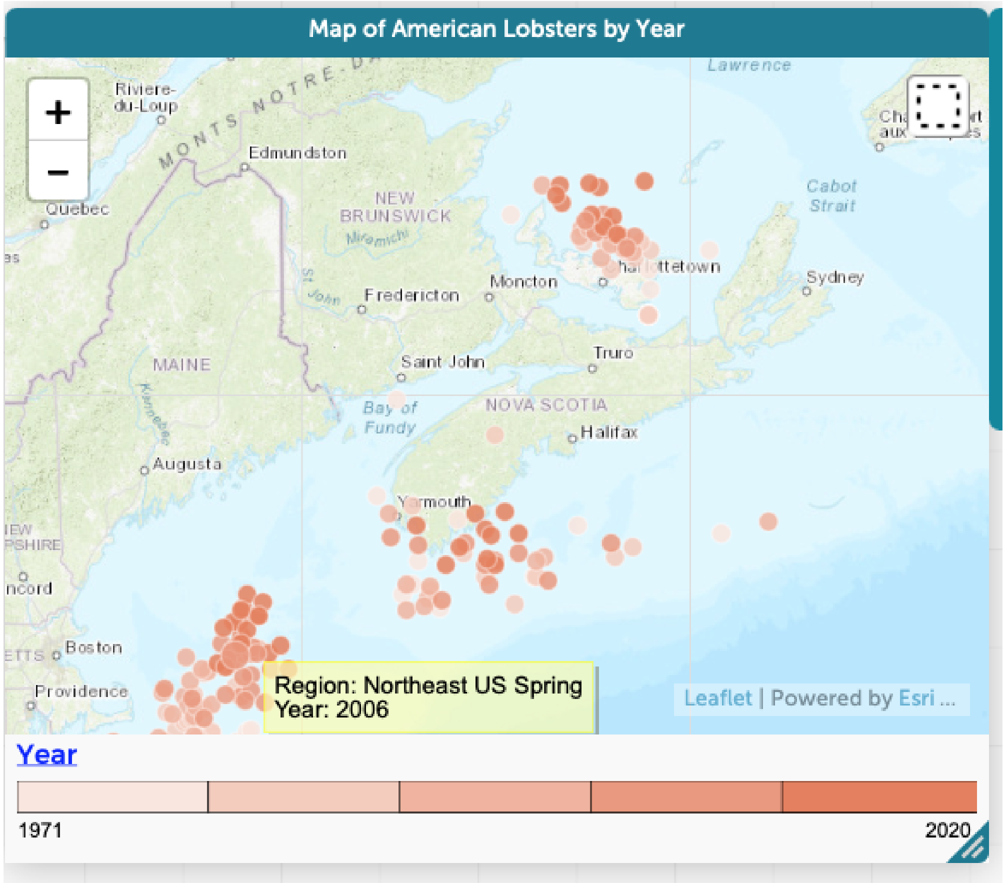
Figure 3 displays a close-up of the map of lobster locations where points are colored based on year of measurement. Data values for a selected point (2006 Northeast US Spring) are displayed: users can move their mouse to highlight any point of interest.

### Scatterplot of latitude

Figure 4 displays the map along with a linked scatterplot of latitude versus year. The plot can be generated on-the-fly by creating a new graph and dragging the desired attributes (Year and Latitude) to the relevant axes. Users can also click on the axis and select the desired attribute from a menu. When a selected point is clicked, it is highlighted on both visual displays. The graph provides some indication of a positive slope, potentially related to climate change.

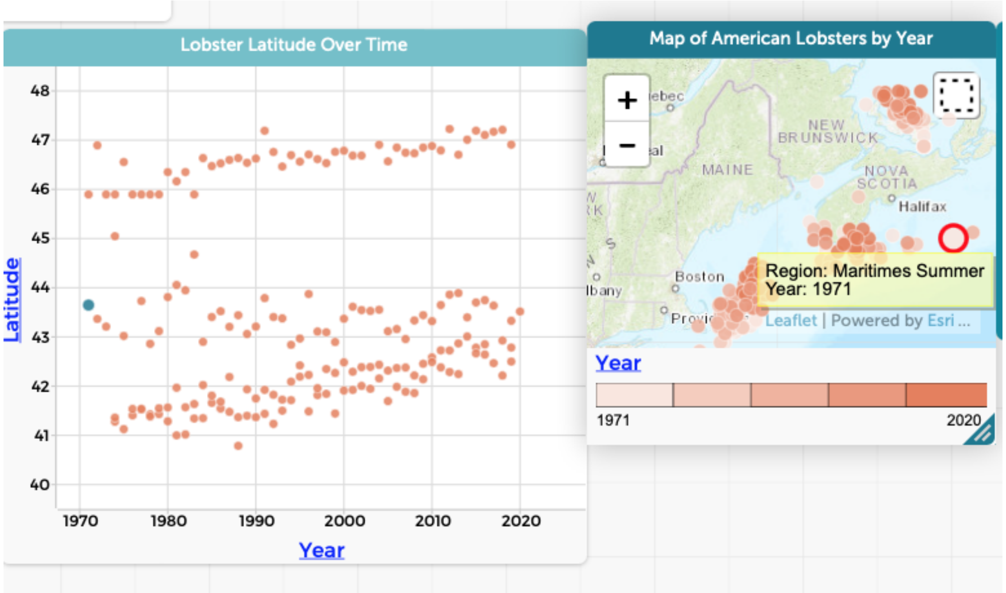
Figure 5 is a similar display, where a regression line is added to help describe the pattern. The regression line is generated using a menu attached to the graph window. Clearly the regression line is not accounting for the three separate populations and may be misleading.

Figure 6 brings a third variable (region) into the scatterplot by dragging the attribute onto the body of the graph, which helps the analyst distinguish the separate populations.



Region: Northeast US Spring
Year: 2006

Figure 3: CODAP allows users to highlight points from the map display (and display attributes of that point).



Region: Maritimes Summer
Year: 1971

Figure 4: Multiple linked displays can be created, with the capability of selecting points that will be highlighted in both displays.

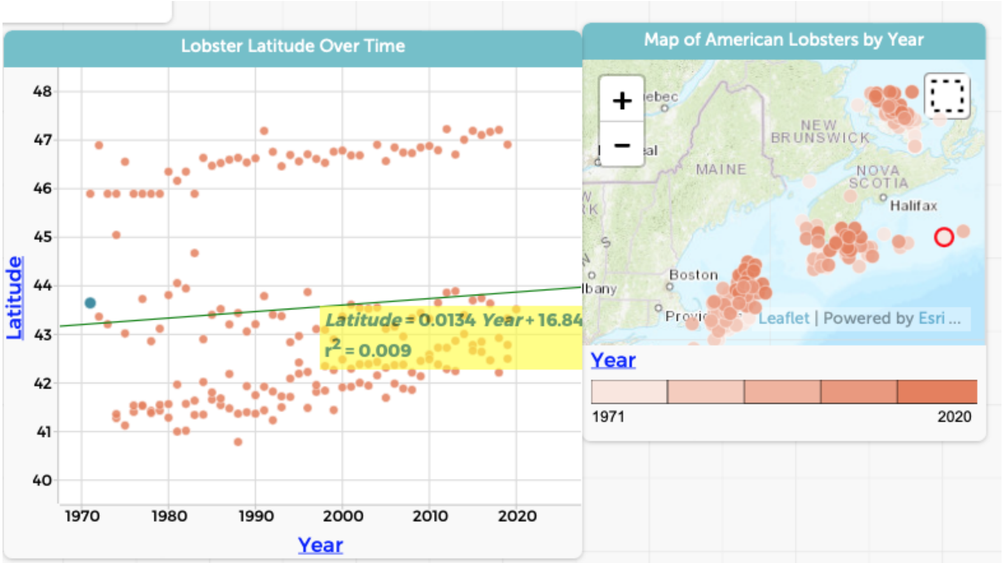


Figure 5: A similar display can include annotations, in this case a linear regression line

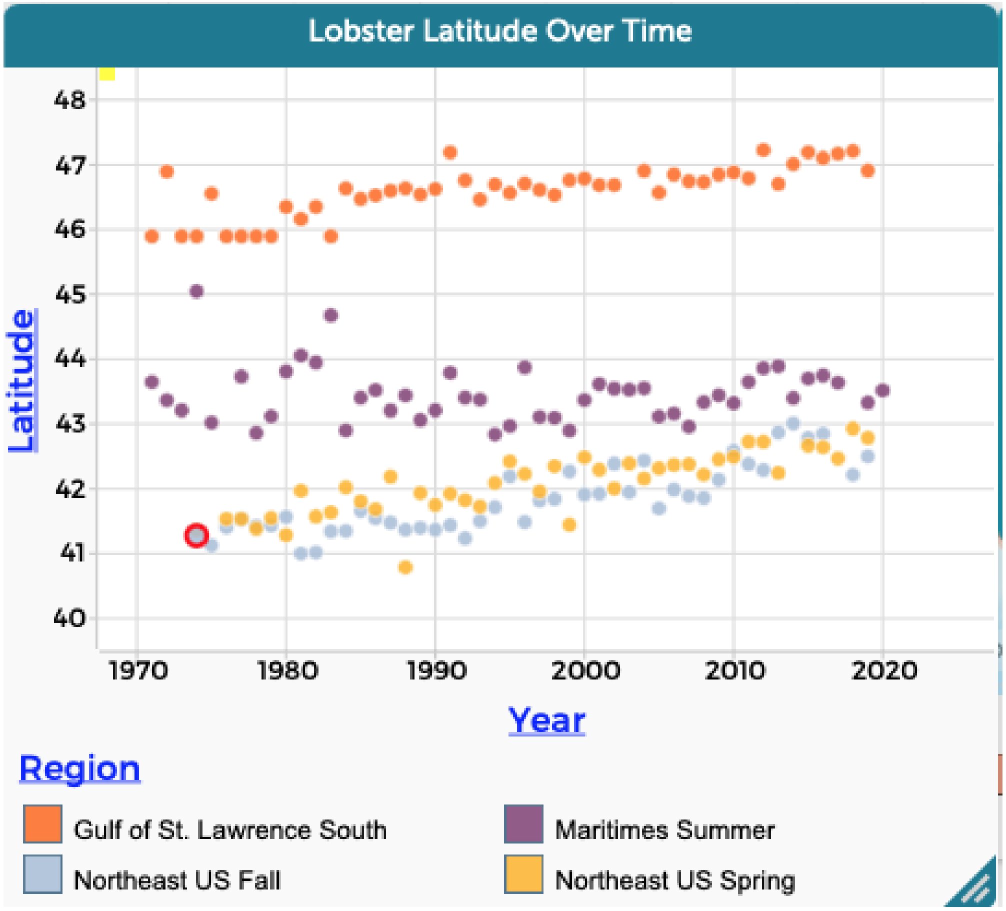
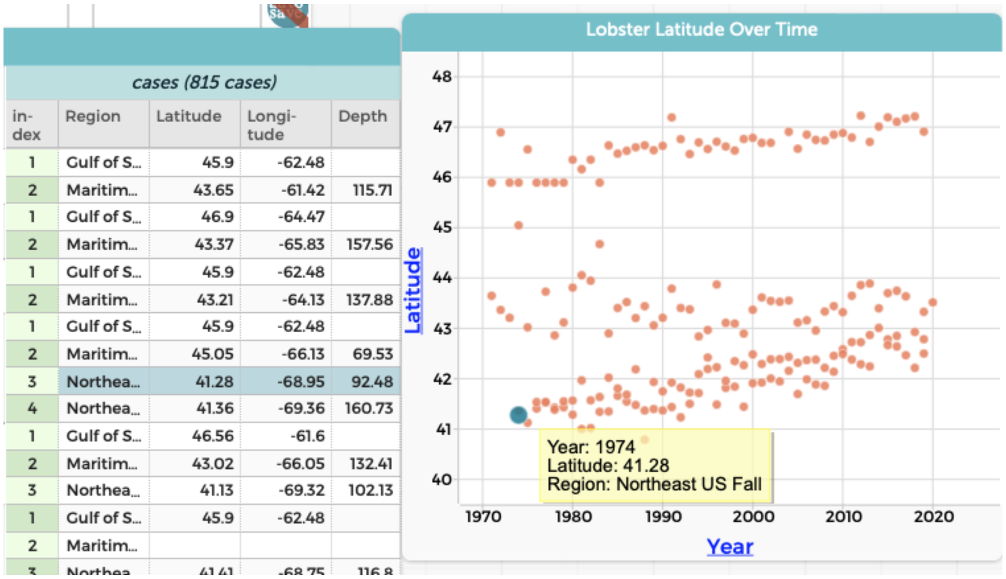


Figure 6: Scatterplots can display three (or more) variables. In this case, the region of the population is used to color points.

Figure 7 demonstrates how a case table can also be displayed. Here the selected point (Northeast 1974) shows in both linked displays.



*cases (815 cases)*

| in-dex | Region | Latitude | Longi-tude | Depth |
|---|---|---|---|---|
| 1 | Gulf of S... | 45.9 | -62.48 | |
| 2 | Maritim... | 43.65 | -61.42 | 115.71 |
| 1 | Gulf of S... | 46.9 | -64.47 | |
| 2 | Maritim... | 43.37 | -65.83 | 157.56 |
| 1 | Gulf of S... | 45.9 | -62.48 | |
| 2 | Maritim... | 43.21 | -64.13 | 137.88 |
| 1 | Gulf of S... | 45.9 | -62.48 | |
| 2 | Maritim... | 45.05 | -66.13 | 69.53 |
| 3 | Northea... | 41.28 | -68.95 | 92.48 |
| 4 | Northea... | 41.36 | -69.36 | 160.73 |
| 1 | Gulf of S... | 46.56 | -61.6 | |
| 2 | Maritim... | 43.02 | -66.05 | 132.41 |
| 3 | Northea... | 41.13 | -69.32 | 102.13 |
| 1 | Gulf of S... | 45.9 | -62.48 | |
| 2 | Maritim... | | | |
| 3 | Northea... | 41.41 | -68.75 | 116.8 |

Figure 7: CODAP can link case tables and graphical displays as a way to highlight and explore individual points.

More information about CODAP can be found at https://codap.concord.org.

# R

R is an open-source scripting environment for statistical computing and graphics that has been widely adopted (https://www.r-project.org). RStudio is an integrated development environment (IDE) that was designed to facilitate use of R for experts and new users (https://www.rstudio.com/products/rstudio). The `mosaic` and `ggformula` packages were designed to provide a simplified interface to modeling and visualization in R (https://doi.org/10.32614/RJ-2017-024).

We use these packages to explore the lobster data. The analyses are scaffolded using an R Markdown file (https://rmarkdown.rstudio.com, see also https://escholarship.org/uc/item/90b2f5xh) which is a text-based format to undertake reproducible analyses.

### Ingest data

We begin by reading in the data. The base pipe operator (`|>`) allows the output of one function to be provided as the input to the next. This piping helps clarify the flow of operations (read the CSV, clean up the names, including only the lobsters, and clump the Northeast US catch into a single group).

```
lobsters <- readr::read_csv("fishdata.csv") |>
  janitor::clean_names() |>
  filter(common_name == "American lobster") |>
  mutate(region_3 = if_else(
    region %in% c("Northeast US Fall", "Northeast US Spring"),
    "Northeast US",
    region)
  )
glimpse(lobsters)
```

```
## Rows: 187
## Columns: 8
## $ common_name <chr> "American lobster", "American lobster", "American lobster"~
## $ species     <chr> "Homarus americanus", "Homarus americanus", "Homarus ameri~
## $ year        <dbl> 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980~
## $ region      <chr> "Gulf of St. Lawrence South", "Gulf of St. Lawrence South"~
## $ latitude    <dbl> 45.90, 46.90, 45.90, 45.90, 46.56, 45.90, 45.90, 45.90, 45~
## $ longitude   <dbl> -62.48, -64.47, -62.48, -62.48, -61.60, -62.48, -62.48, -6~
## $ depth       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ region_3    <chr> "Gulf of St. Lawrence South", "Gulf of St. Lawrence South"~
```

We see that there are 187 observations (each one a geographical location for each region for a given year).

### Exploration

### Summary statistics and distributions

A number of univariate summaries can be calculated.

```
tally(~ region_3, margins = TRUE, data = lobsters)
```

```
## region_3
## Gulf of St. Lawrence South           Maritimes Summer
##                         48                         50
##               Northeast US                      Total
##                         89                        187
```

```
tally(~ region_3, format = "percent", data = lobsters)
```

```
## region_3
## Gulf of St. Lawrence South           Maritimes Summer
##                      25.67                      26.74
##               Northeast US
##                      47.59
```

```
tally(~ year, margins = TRUE, data = lobsters)
```

```
## year
##  1970  1971  1972  1973  1974  1975  1976  1977  1978  1979  1980  1981  1982
##     1     2     2     2     4     3     4     4     4     4     4     4     4
##  1983  1984  1985  1986  1987  1988  1989  1990  1991  1992  1993  1994  1995
##     4     4     4     4     4     4     4     4     4     4     4     4     4
##  1996  1997  1998  1999  2000  2001  2002  2003  2004  2005  2006  2007  2008
##     4     4     4     4     4     4     4     3     4     4     4     4     4
##  2009  2010  2011  2012  2013  2014  2015  2016  2017  2018  2019  2020 Total
##     4     4     4     4     4     3     4     4     3     3     4     1   187
```

```
favstats(~ latitude, data = lobsters) # needed for bottom and top
```

```
##    min    Q1 median   Q3   max  mean    sd   n missing
##  40.79 41.95  42.96 45.9 47.23 43.55 1.972 185       2
```

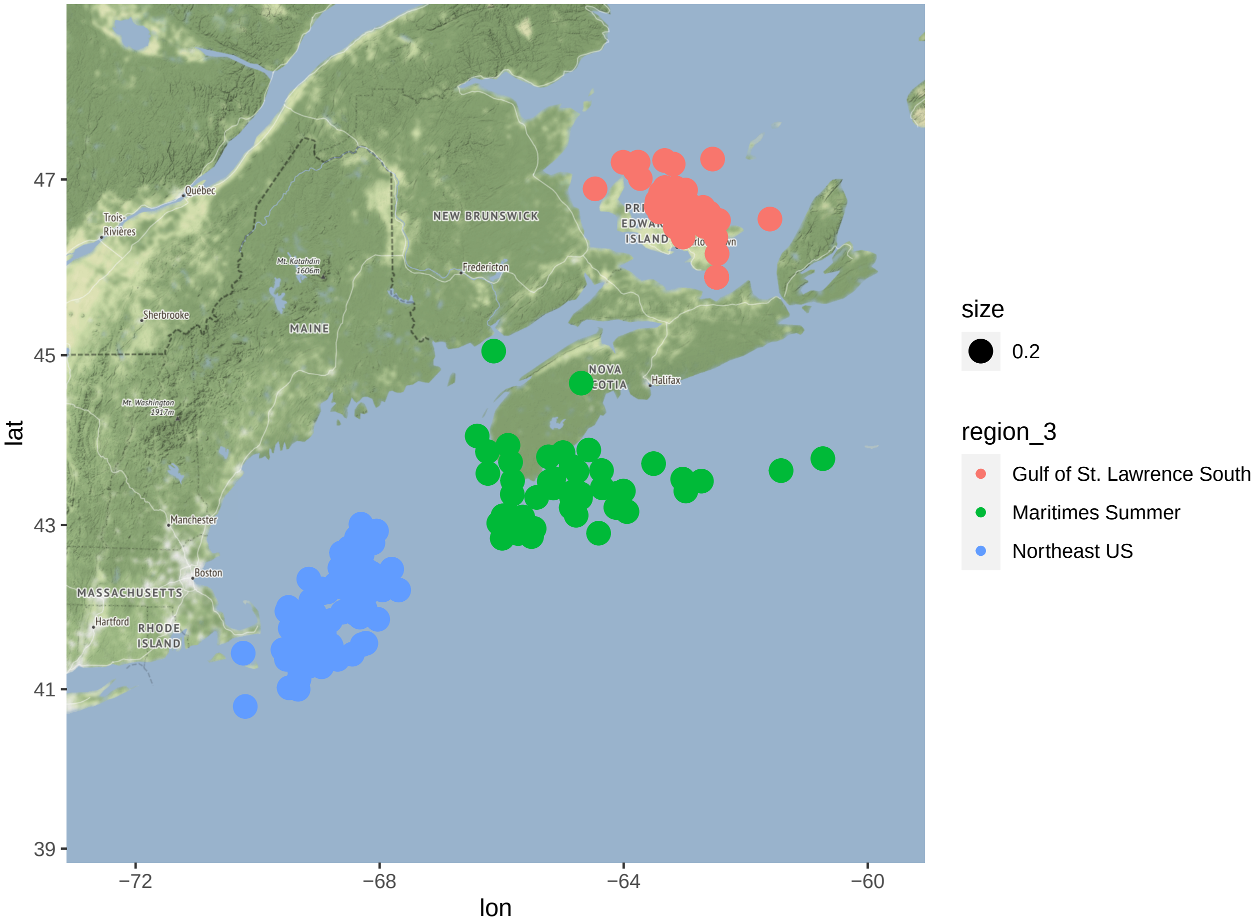
```
favstats(~ longitude, data = lobsters) # needed for left and right
```

```
##     min     Q1 median     Q3    max   mean    sd   n missing
##  -70.24 -68.72 -66.13 -63.45 -60.74 -66.25 2.632 185       2
```

A straightforward way to identify what map region to display is by calculating summary statistics for latitude and longitude.

### Simple map

```
myLocation <- c(
  left = -71.0,
  bottom = 40.78,
  right = -60.70,
  top = 47.23
)
myMap <- get_map(
  location = myLocation,
  source = "stamen",
  maptype = "watercolor",
  crop=FALSE
)
ggmap(myMap) +
  geom_point(
    aes(x = longitude, y = latitude, color = region_3, size = 0.2),
    data = lobsters
  )
```

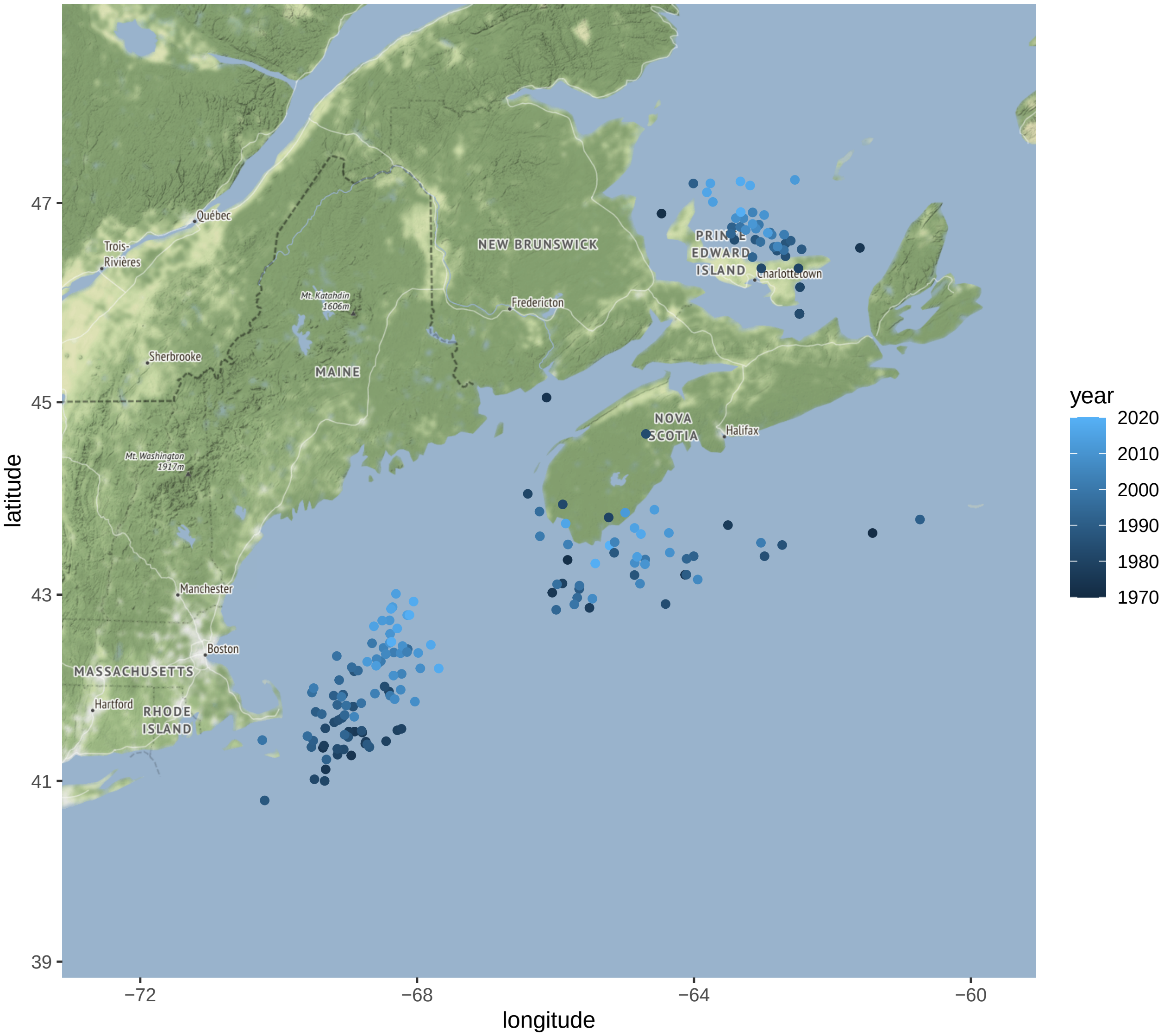


Routines in the `ggmap` package (https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf) are used to generate the maps.

### More sophisticated map

A more sophisticated map colors the points (depending on year) and cleans up the axis labels. We see the trend of northward movement when comparing the years.

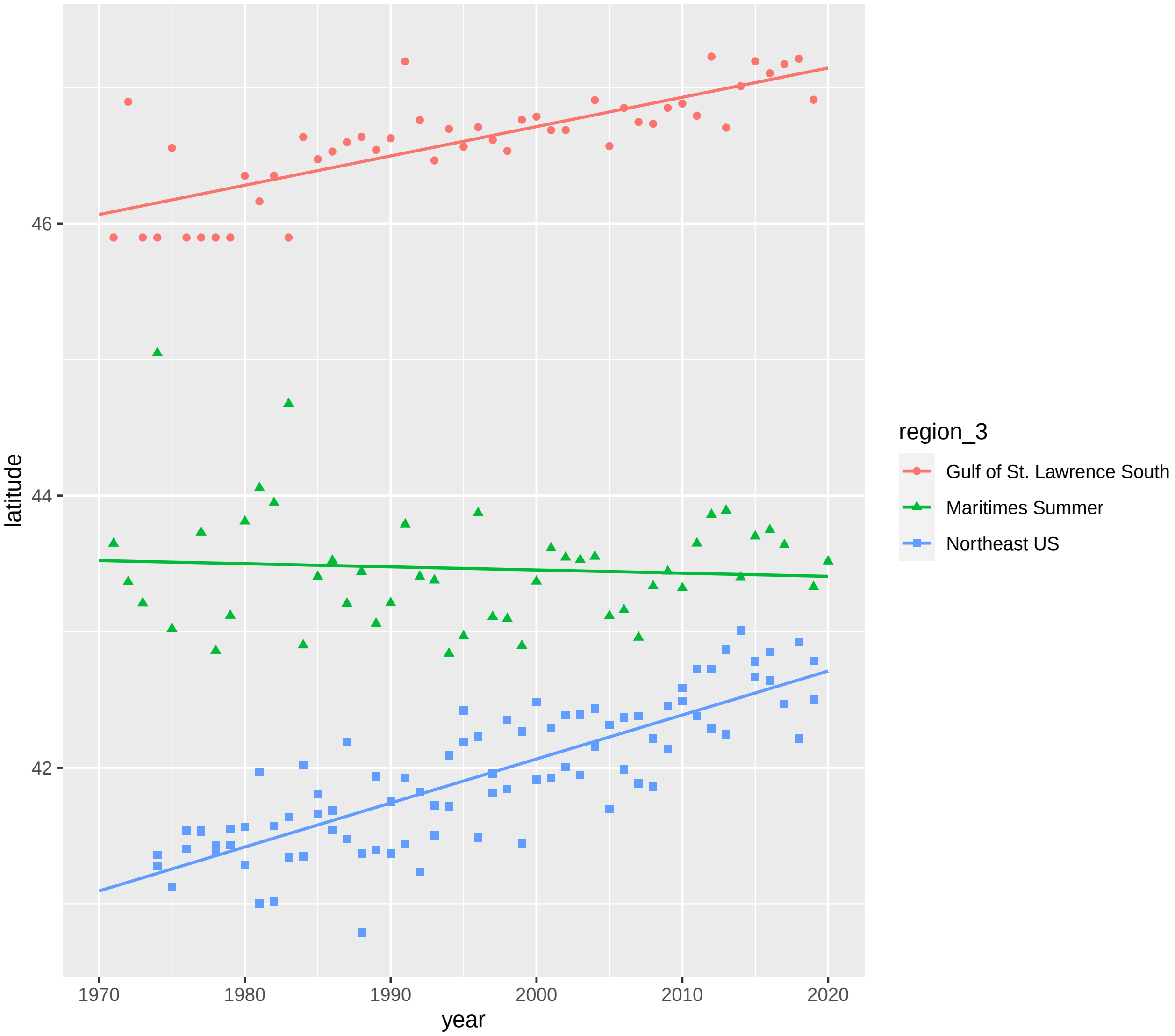
```r
ggmap(myMap) +
  geom_point(
    aes(x = longitude, y = latitude, color = year),
    data = lobsters
  ) +
  labs(x = "longitude", y = "latitude")
```



### Scatterplot of latitude

A scatterplot is an alternative representation of these data. Here we generate a scatterplot where the points are colored by the three level region variable.

```
gf_point(
  latitude ~ year,
  color = ~ region_3,
  shape = ~ region_3,
  data = lobsters
) |>
  gf_lm()
```



Both of these displays are static, however dynamic graphical displays are available within R (see for example https://mdsr-book.github.io/mdsr2e/ch-vizIII.html).

It is possible to further create a map that distinguishes regions by hue and years by saturation using similar commands; though we chose to end the vignette here.

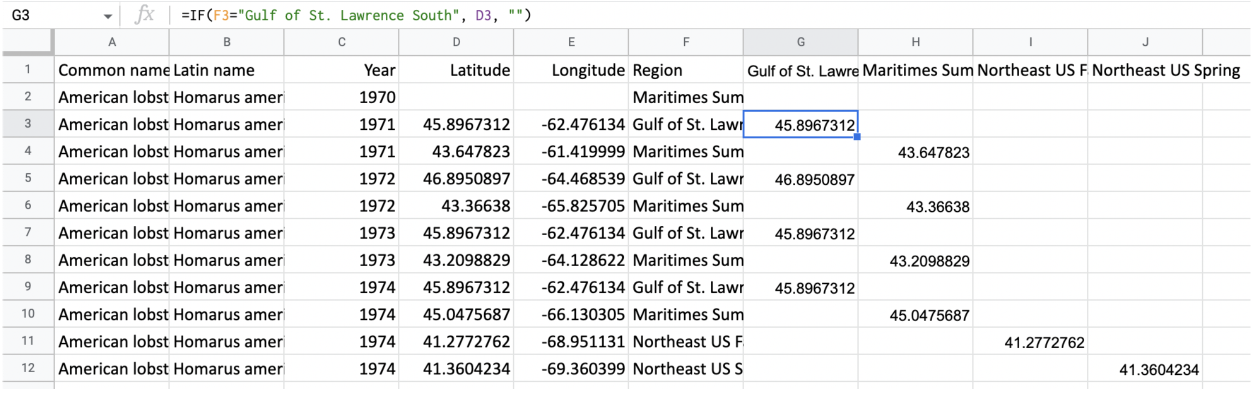
## Google Sheets

The data are also available within Google Sheets: https://docs.google.com/spreadsheets/d/14MhVPLtXXB20vMecX0UGQpFW0jvNtkgx_51nDsU6ncA/edit#gid=0

To create a similar plot of fish latitude over time in Google Sheets, the data needed additional transformations. Observations of each of the four regional types (Gulf of St. Lawrence, Maritimes Summer, Northeast US Fall, and Northeast US Spring) needed to be recoded as distinct series in order to be visually distinguished from one another in a scatterplot.

### Reformat data

To create these separate series, we used the formula =IF([value in region column]="Desired Series Region", [value in latitude column],""). This created four new columns; for each observation the appropriate column contained the observed latitude (see 8).



```
G3 ▾ fx =IF(F3="Gulf of St. Lawrence South", D3, "")
```

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Common name | Latin name | Year | Latitude | Longitude | Region | Gulf of St. Lawre | Maritimes Sum | Northeast US F | Northeast US Spring |
| 2 | American lobst | Homarus amer | 1970 | | | Maritimes Sum | | | | |
| 3 | American lobst | Homarus amer | 1971 | 45.8967312 | -62.476134 | Gulf of St. Lawr | 45.8967312 | | | |
| 4 | American lobst | Homarus amer | 1971 | 43.647823 | -61.419999 | Maritimes Sum | | 43.647823 | | |
| 5 | American lobst | Homarus amer | 1972 | 46.8950897 | -64.468539 | Gulf of St. Lawr | 46.8950897 | | | |
| 6 | American lobst | Homarus amer | 1972 | 43.36638 | -65.825705 | Maritimes Sum | | 43.36638 | | |
| 7 | American lobst | Homarus amer | 1973 | 45.8967312 | -62.476134 | Gulf of St. Lawr | 45.8967312 | | | |
| 8 | American lobst | Homarus amer | 1973 | 43.2098829 | -64.128622 | Maritimes Sum | | 43.2098829 | | |
| 9 | American lobst | Homarus amer | 1974 | 45.8967312 | -62.476134 | Gulf of St. Lawr | 45.8967312 | | | |
| 10 | American lobst | Homarus amer | 1974 | 45.0475687 | -66.130305 | Maritimes Sum | | 45.0475687 | | |
| 11 | American lobst | Homarus amer | 1974 | 41.2772762 | -68.951131 | Northeast US F | | | 41.2772762 | |
| 12 | American lobst | Homarus amer | 1974 | 41.3604234 | -69.360399 | Northeast US S | | | | 41.3604234 |

Figure 8: Formula and example of recoded region series. This formatting is required to color observations by region on the scatterplot.

### Generate graphical display

A scatter plot can then be generated using five columns of the table (Year plus the four new constructed columns). If the option to simply create a chart is selected with these columns, Google Sheets chooses a bar chart style the yields an inappropriate visualization; the user must know that a scatter plot will generate the desired outcome. The axes of the plot must also be manually adjusted by entering a minimum value using a format option that is initially hidden from the user.

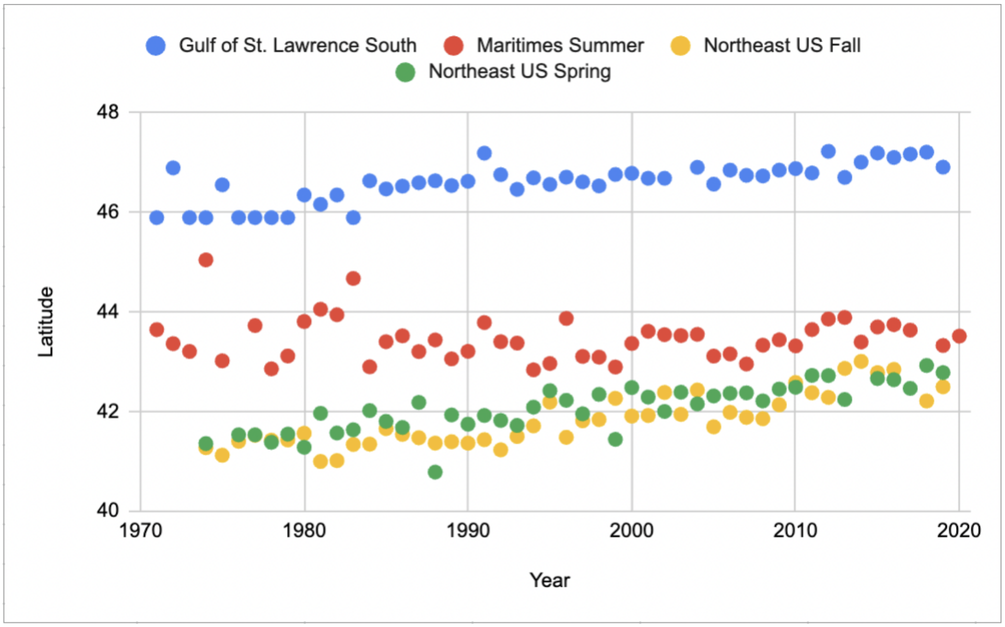


Figure 9: Google Sheets scatter plot of recoded data.

It was not straightforward to generate the map within Google Sheets.

# Tableau

The data within Tableau are accessible using the following file: https://nicholasjhorton.github.io/K12-Data-Tools/static/fishdata.twb

Figure 10 displays a scatter plot of data, colored by region. Scatterplots are generated by dragging and dropping available attributes directly onto a plot space or associated fields in the interface. Additional, prominently featured interface options are available to color and otherwise format the data display.

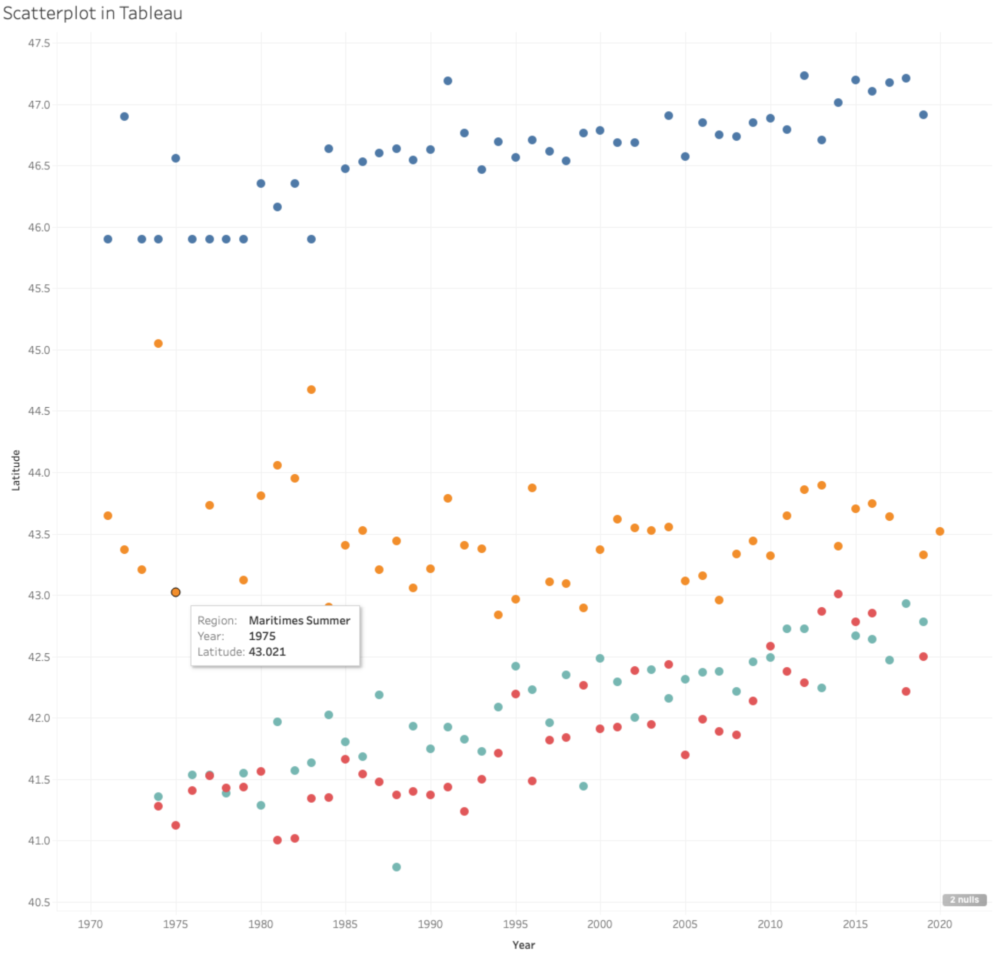
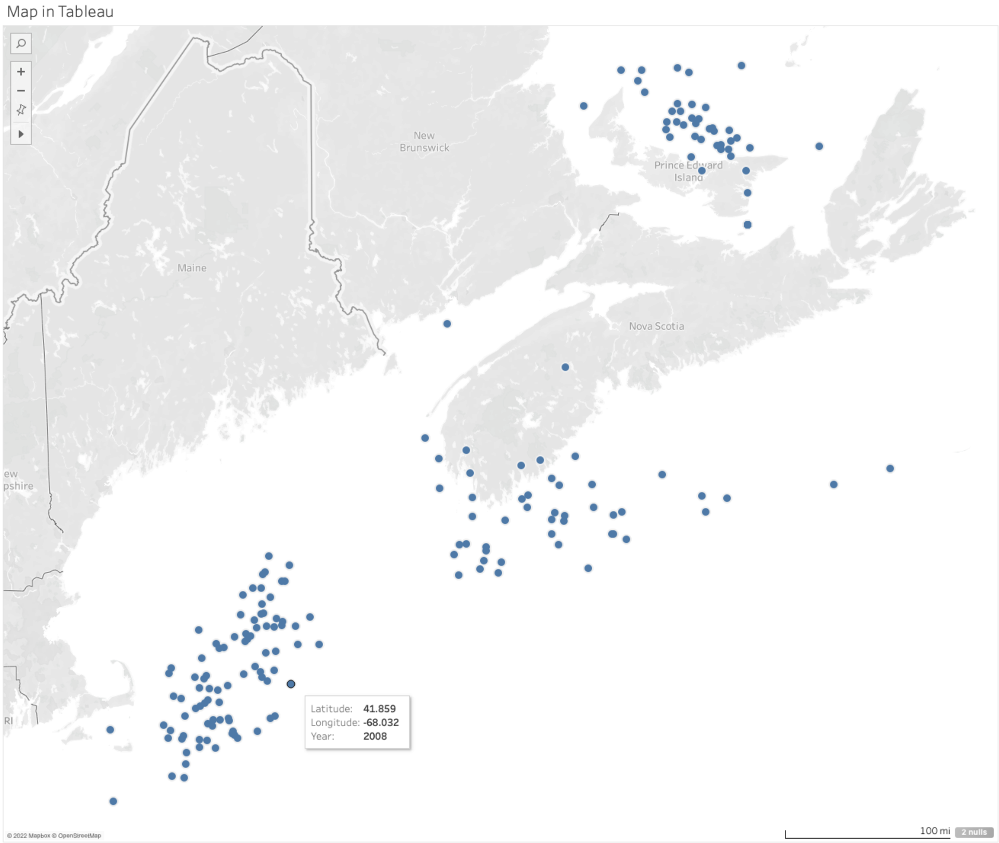


Figure 10: Tableau scatter plot of data.

Figure 11 displays a map of the locations. For both displays, a point can be selected to highlight information about that observation.

It was not straightforward to color the points on the map based on year.



Map in Tableau

Latitude: 41.859
Longitude: -68.032
Year: 2008

Figure 11: Tableau map of data.