K12 Data Tools Vignette (R)

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

August 28, 2022

Contents

Overview

To explore how the selection of data analysis tools can fundamentally shape what is highlighed in an investigation, we conducted a comparative analysis using free, popular tools from three distinct genres: R (scripting), CODAP (visual), and Tableau (see supplementary materials at https://nicholasjhorton.github.io/K12-Data-Tools). 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 thr ee distinct groups of lobsters between the years of 1970 and 2010. 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 (geographic data); it is deeply embedded in a specific context that connects to other disciplines; and there is a clear pattern to observe amidst relatively messy, multivariate data.

This file demonstrates how to carry out these analyses using R (https://www.r-project.org).

The source files 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 file this dataset came from.

Executive summary of 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
- there is a shift to the north
- this might indicate impacts of climate change

Other notes about the vignettes

- mapping is important for this example
- data science features many things, among them a variety of data types (e.g., maps, text, networks)
- multivariate thinking is important (moving beyond univariate and bivariate visualizations)

R

Ingest data

```
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"~
```

Exploration

Summary statistics and distributions

```
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
```

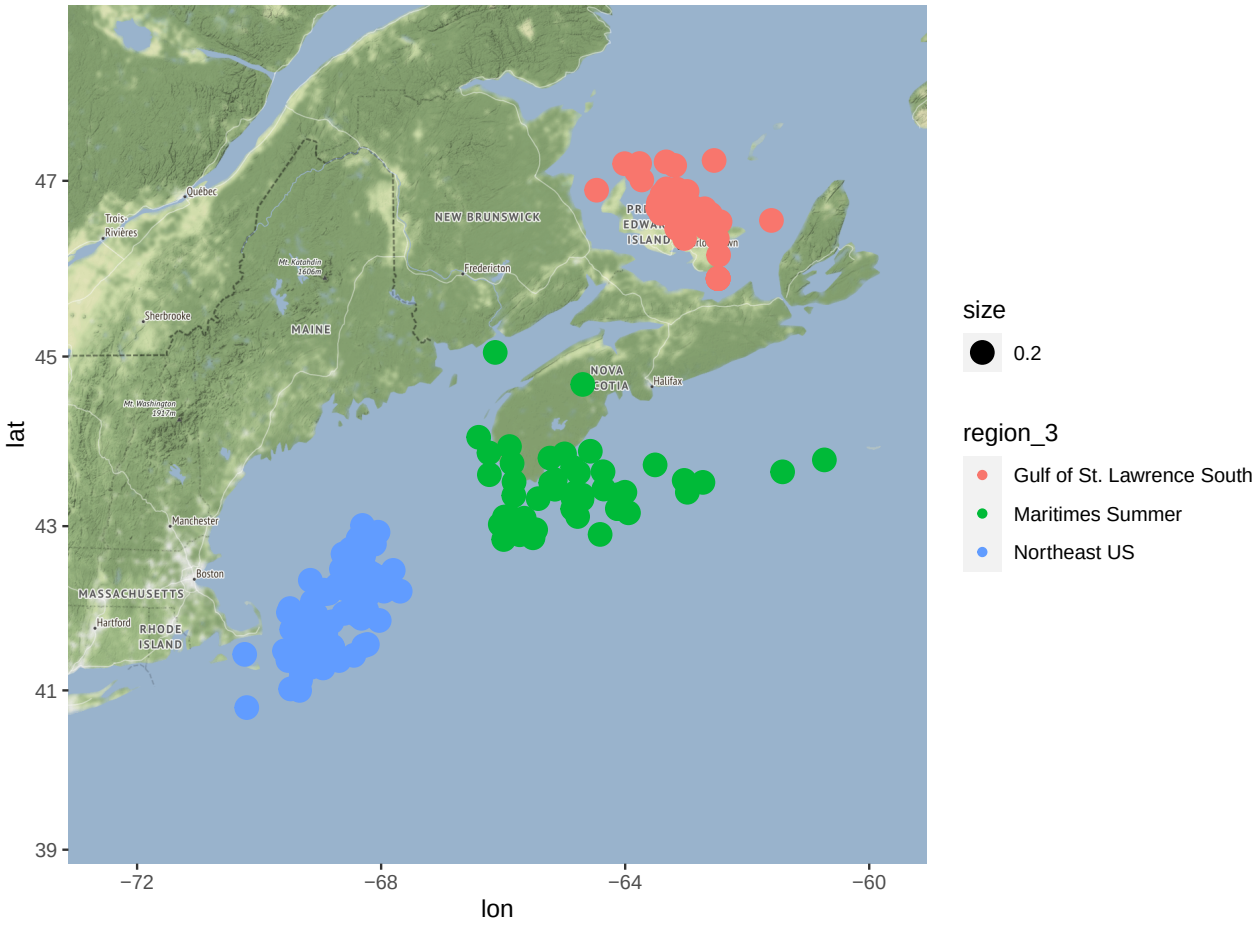
```
##    min    Q1 median   Q3   max  mean    sd   n missing
##  40.79 41.95  42.96 45.9 47.23 43.55 1.972 185       2
```

```
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
```

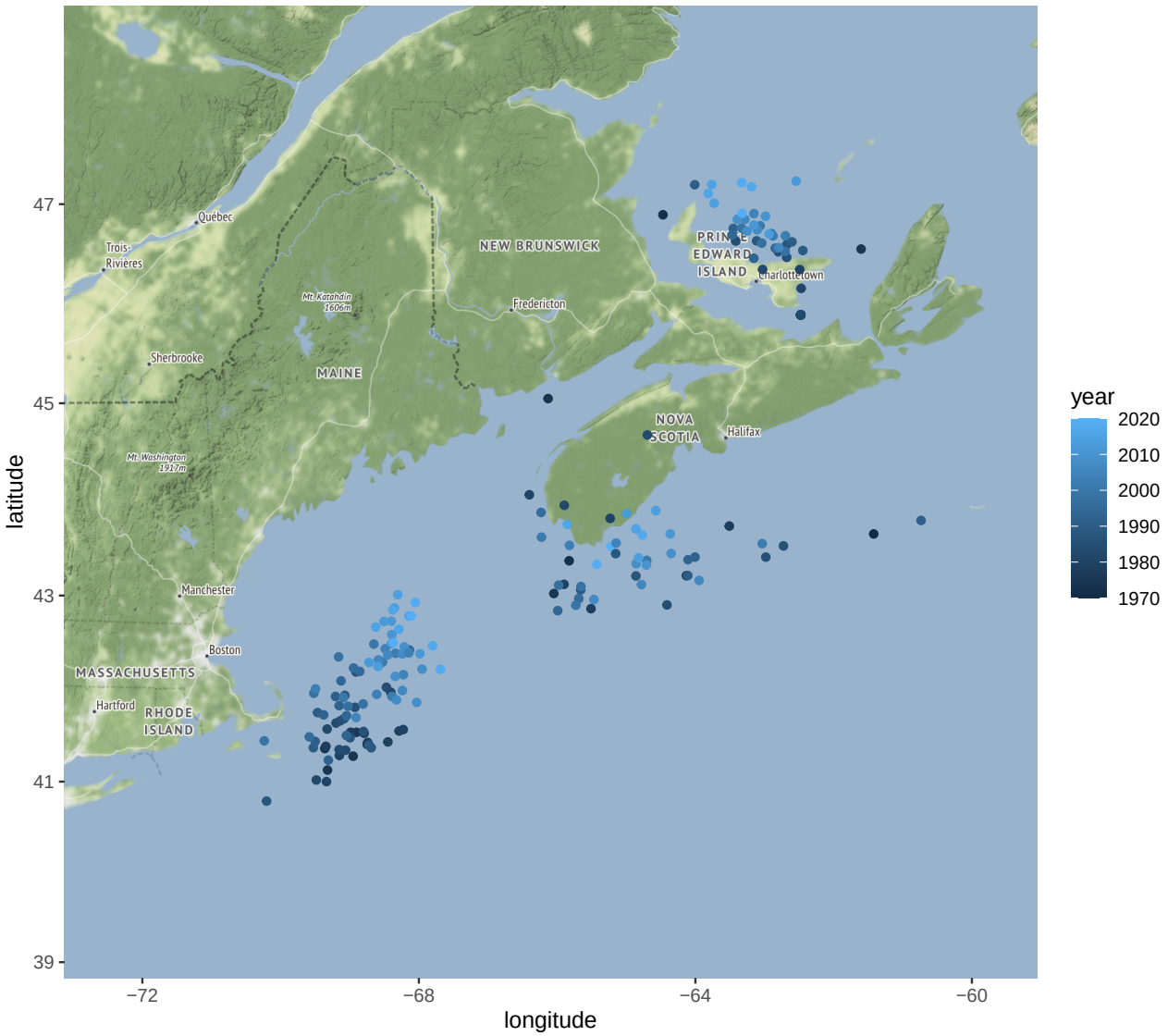
Simple map

```
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
  )
```



More sophisticated map

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



Scatterplot of latitude

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


latitude
46
44
42
1970
1980
1990
2000
2010
2020
year
region_3
Gulf of St. Lawrence South
Maritimes Summer
Northeast US