USGS NAS

Mariah Boudreau, Kerria Burns, Amanda Jones

12/5/2019

The source is the United States Geological Survey (USGS) Nonindigenous Aquatic Species (NAS) data set. It is a collection of zebra mussel sightings across the United States. Zebra mussels are an incredibly damaging invasive species in the continental US. They have spread drastically over the past several decades, beginning in the Great Lakes. Included is a point map that represents the number of individuals by the color and size of the circle superimposed on a map of the United States. Also included is an Excel spreadsheet including species identification, latitude and longitude, phase of establishment, and date of sighting, among other information.

Figure 1 Zebra mussel sightings represented by size and color of circle – darker and larger circle correspond to a greater number of individuals. Source: USGS NAS The map includes Hydraulic Unit Codes (HUC), a system derived by USGS to delimit subregions within the country.

An inquiry then could be to standardize each HUC by area to find the number of zebra mussels per unit area to determine at which locations the highest density occurs. Similarly, bins of degrees of latitude could be derived (to reduce the step of standardization and bypass the use of HUC altogether, a possible benefit of which would be that measurements of latitude might yield more climatically relevant data rather than political/economic divisions), to again determine the density of zebra mussels per unit (as yet to be defined). This could be combined with climatic information to determine which climate conditions favor the establishment and invasion of zebra mussels to the most severe extent. A time series could be constructed to see if there were certain time periods that showed more movement than others. The type of analysis could be done on a map similar to the one above to see the visual change as well.

Zebra mussels belong to the phylum Mollusca, the group bivalves (unique among them–the only species that attaches to a hard substrate), which are native to the Black, Caspian, and Azov seas (USGS NAS). They were first observed in areas nearby Indiana, Michigan, and Ohio in 1986 (USGS NAS). Zebra mussels likely originated in Europe, establishing a population in Lakes St. Clair and Erie when ships dispersed ballast water (Hebert et al. 1989). The influence of bivalves on ecosystems includes the following: removal of particles from the water column, reducing populations of consumers depending on these particles as food source, increasing populations that use bivalves or their waste products, and making available particles previously used by phytoplankton (Strayer et al. 1999). As a result, their introduction into the United States as a result of commercial trade has had cascading effects on ecosystems, affecting multiple trophic levels and altering the abiotic conditions of the water bodies in which they reside.

The outcome of invasions depends in part on the tolerance of environmental conditions in the area of introduction. Introduced environments with conditions similar to those found in the native range may be more conducive to invasive establishment (Baker and Stebbins, 1965). Additionally, reproductive and physiological characteristics of the invading organism play a role in whether populations become established in the new range. In the case of zebra mussels, free swimming larvae and high fecundity (eggs per female) have been implicated in its proliferation success (Hebert et al. 1989). That zebra mussels are amenable to a wide range of habitats with a flexible reproductive system has aided their spread and establishment in North America where they were introduced via the release of ballast water from ships (Nichols 1996).

Physiological constraints of organisms determine the optimal range in which reproduction can occur, an important factor in the spread and establishment of an invasive species (or any species, for that matter). One such constraint is thermal tolerance, and temperature was shown to affect the timing of gametogenesis (Wacker & Elert 2003). The two experimental groups were raised at different depths to generate the temperature difference, and variation in environmental quality in the surrounding region may explain the differences in egg mass released at the two different depths (Wacker & Elert 2003). In addition, metabolic rates increase with temperature, so that differences in food availability at the two depths could further influence reproductive

investment. Without respective to temperature differences, availability of polyunsaturated fatty acids and food quality in general resulted in changes to reproductive investment (Wacker & Elert 2003). An additional study confirms a threshold of 12 C for the onset of spawning and further identifies two spawning cohorts as the season proceeds from May to August (Borcherding 1991). Gametes are released into the water column over a period of 6 to 8 weeks, where they are fertilized and develop—30,000 to 40,000 eggs may be released by one female, though the actual number may be closer to 1.5 million. Egg release corresponds with temperature, beginning at 12C and peaking at 22C. As a result, juvenile proliferation should track these temperature changes (Hebert et al. 1988). High temperatures alongside low food availability results in reduction in the size of the gonads (Borcherding 1991).

Dreissena tends to colonize structures below 1.2 m. Zebra mussels may spread to larger structures through water intake pipes. Zebra mussels tend to increase water clarity through digestion of suspended sediments—for this reason zebra mussels have been intentionally stocked in lakes outside North America. Food deprived mussels fed indiscriminately on particles of all sizes, but satiated individuals felt only on those in a much smaller size range (MacIsaac 1996).

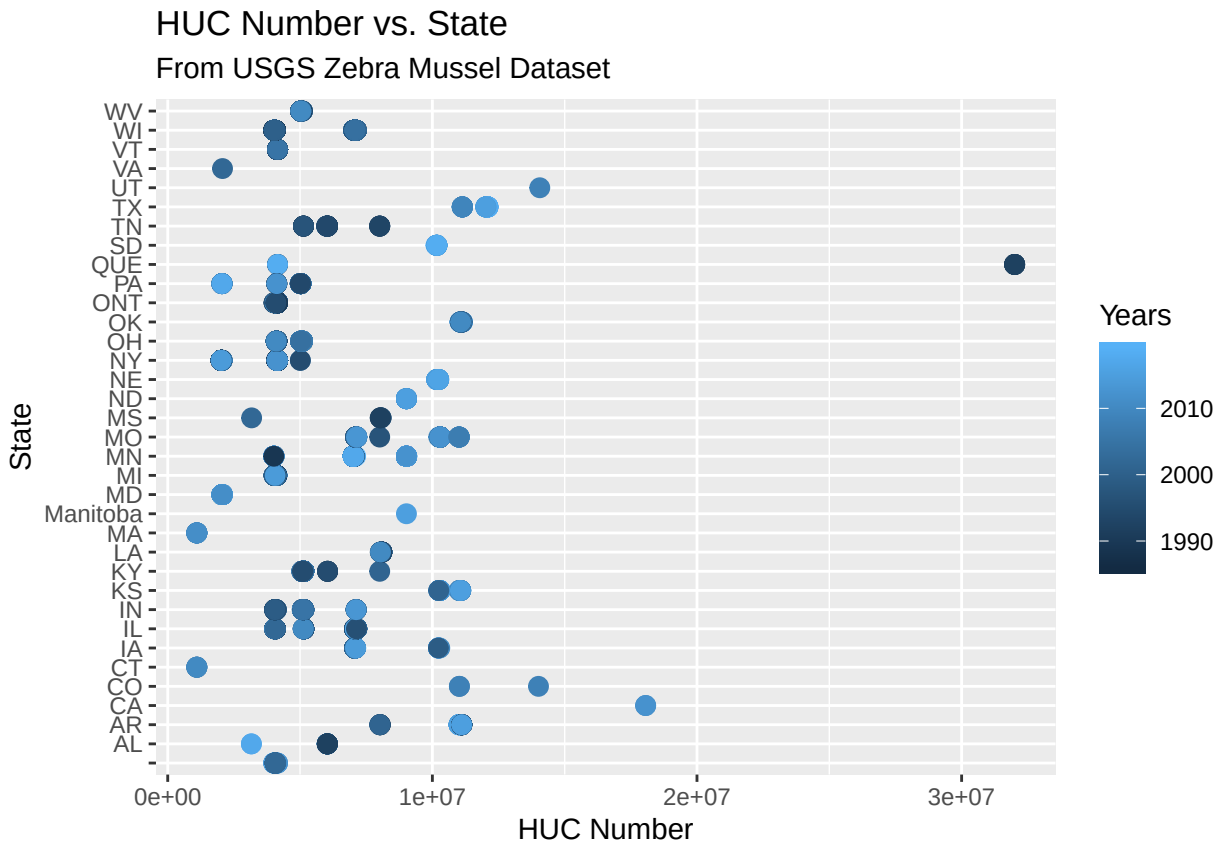
Ambient temperature, seston concentration, and mussel size frequency are three factors that influence Dreissena filtering impact. Maximal filtering rate has been hypothesized to be 5 and 20 C, declining outside of this range. Ingestion rate may also be temperature dependent; however, the results of experiments investigating the effect of temperature are variable. It may still be an important performance factor, and impose a range limitation in southern states. Depending on size, zooplankton may succumb to or evade ingestion by zebra mussels—smaller individuals cannot escape the inflow current and are not rejected, whereas larger individuals may dodge the current or be expelled as a result of irritating the feeding apparatus (MacIsaac 1996).

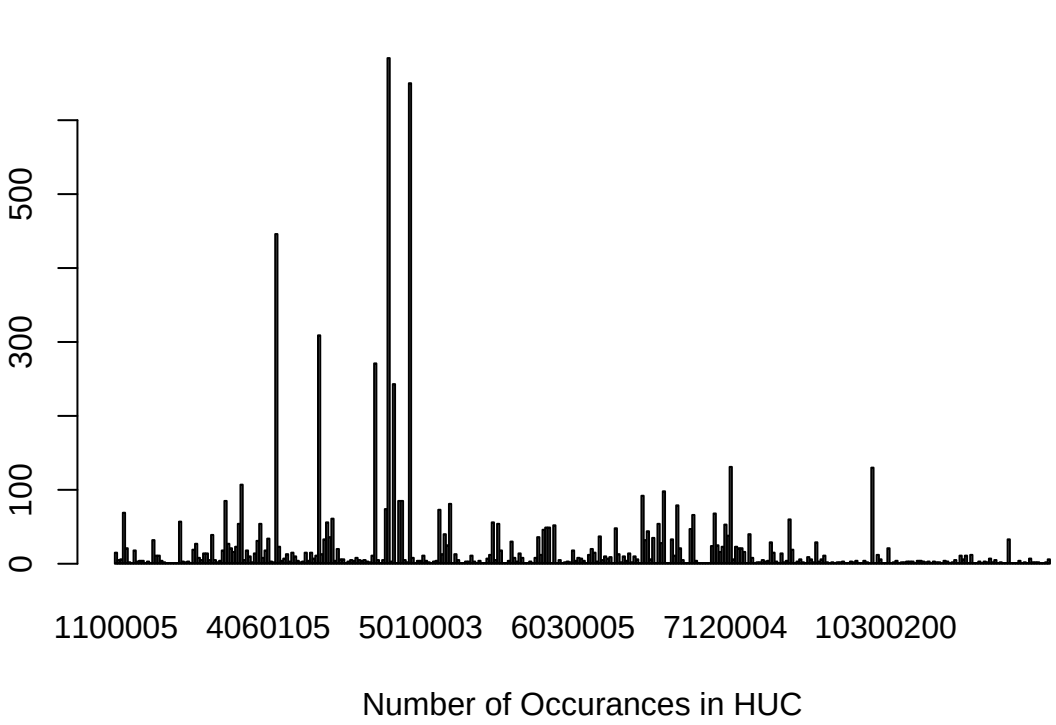
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Initial Visualization

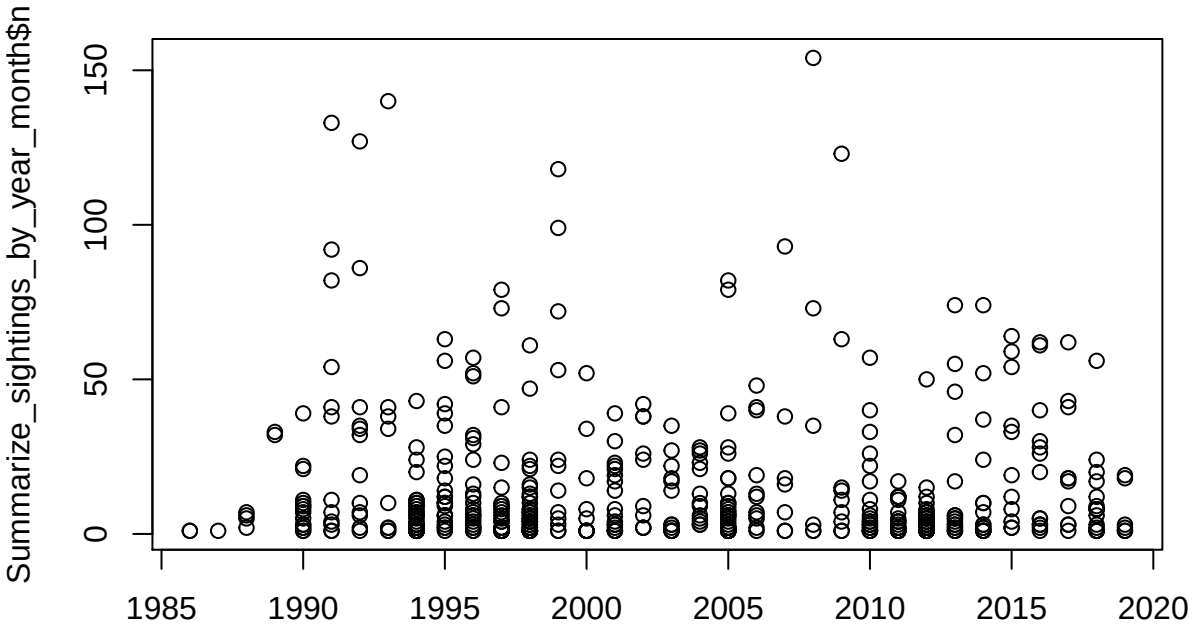
The graph below is one way in which the data set from the USGS zebra mussel data can be visualized. The horizontal axis represents the HUC Number and it is plotted again the state in which that HUC resides. The color of the point exhibits the time when the data was collected. There appears to be an even spread of years throughout the plot given by the various color gradients.

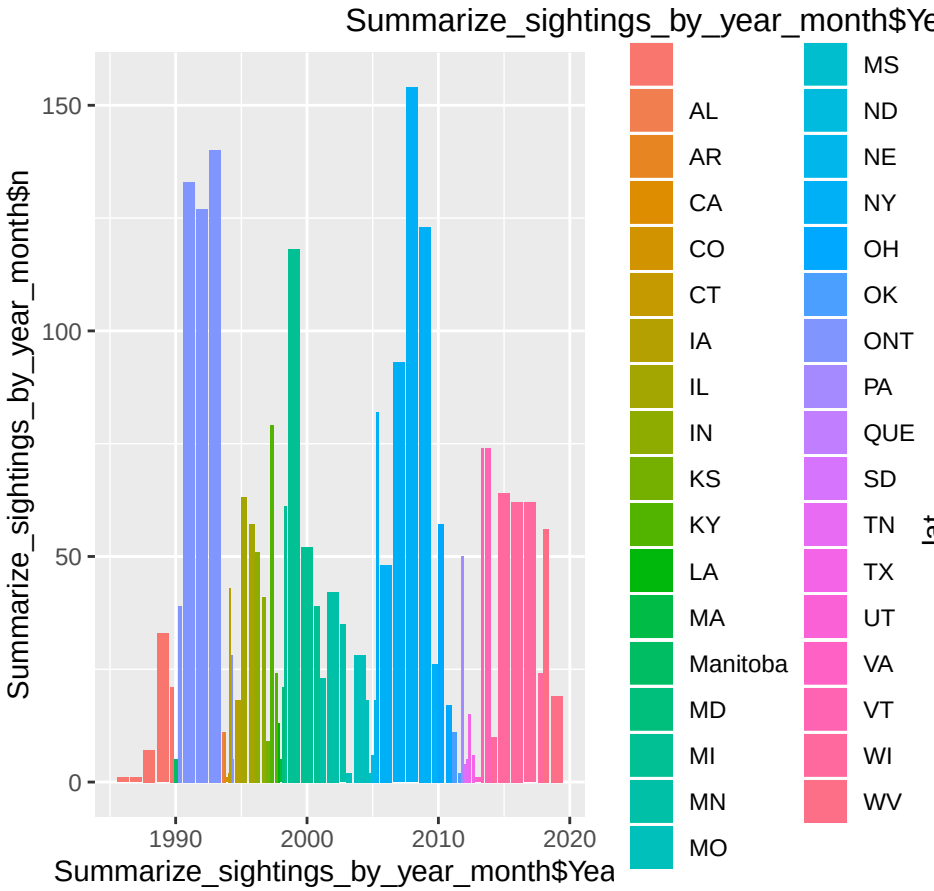
```
## Warning: Removed 576 rows containing missing values (geom_point).
```

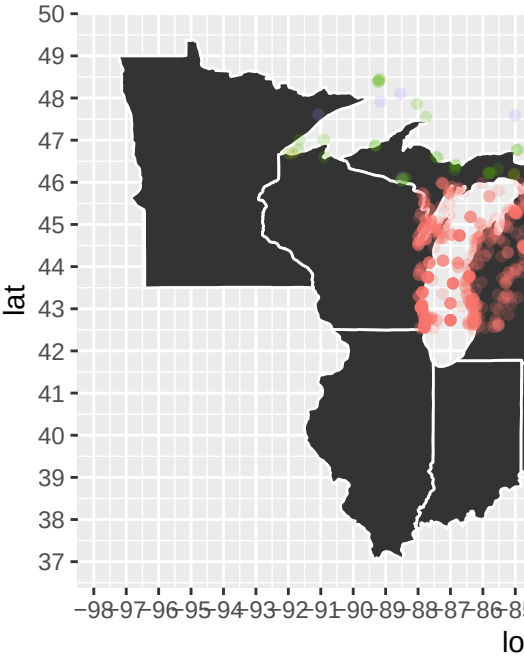

HUC Number vs. State
From USGS Zebra Mussel Dataset
State
WV
WI
VT
VA
UT
TX
TN
SD
QUE
PA
ONT
OK
OH
NY
NE
ND
MS
MO
MN
MI
MD
Manitoba
MA
LA
KY
KS
IN
IL
IA
CT
CO
CA
AR
AL
0e+00
1e+07
2e+07
3e+07
HUC Number
Years
2010
2000
1990


HUC Distribution
0
100
300
500
1100005
4060105
5010003
6030005
7120004
10300200
Number of Occurances in HUC

Maps and graph by state and year

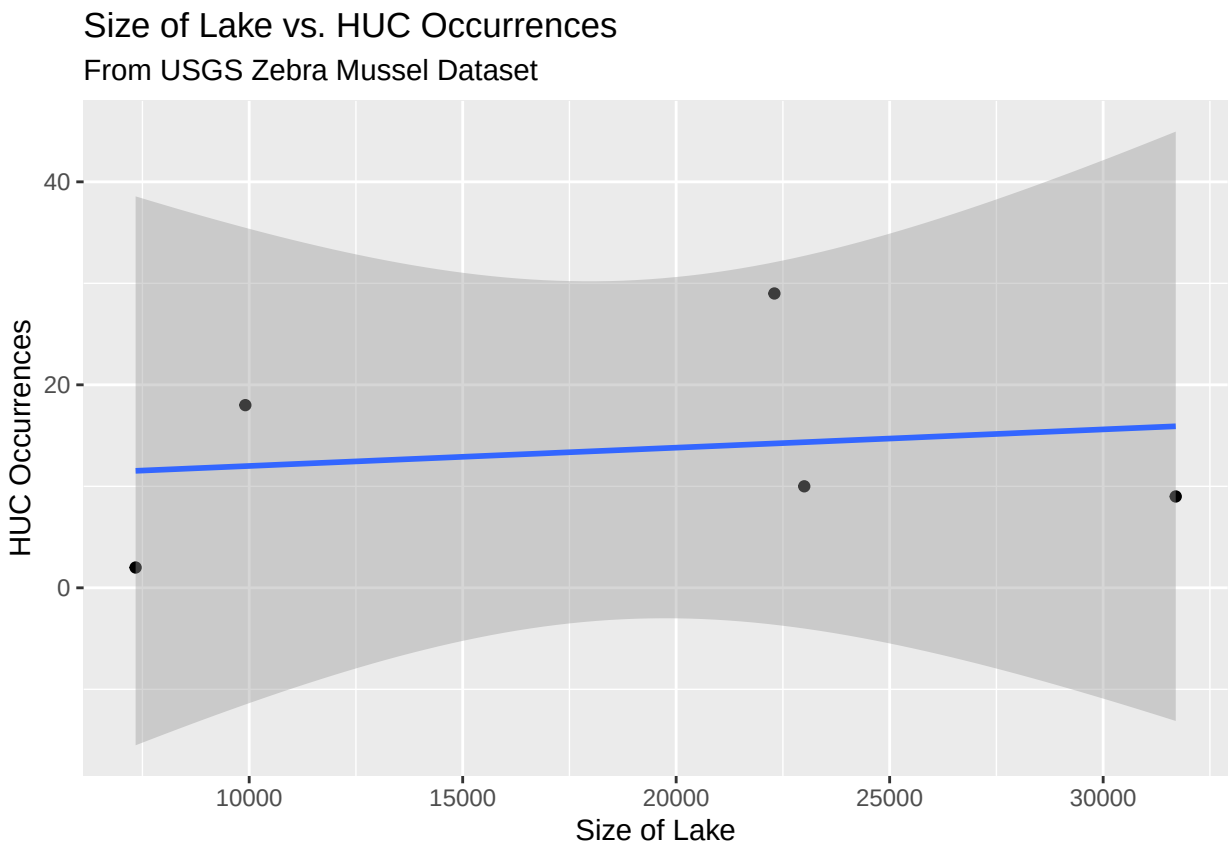






Linear Regression of Size of Lake and HUC Counts

```
##
## Call:
## lm(formula = zm_HUC_occurances ~ size_of_lake, data = sizeDat)
##
## Residuals:
##      1      2      3      4      5
## -6.915 14.778 -4.348  6.011 -9.526
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.020e+01  1.214e+01    0.84    0.462
## size_of_lake 1.802e-04  5.811e-04    0.31    0.777
##
## Residual standard error: 11.72 on 3 degrees of freedom
## Multiple R-squared:  0.03104,    Adjusted R-squared:  -0.2919
## F-statistic: 0.0961 on 1 and 3 DF,  p-value: 0.7768
```



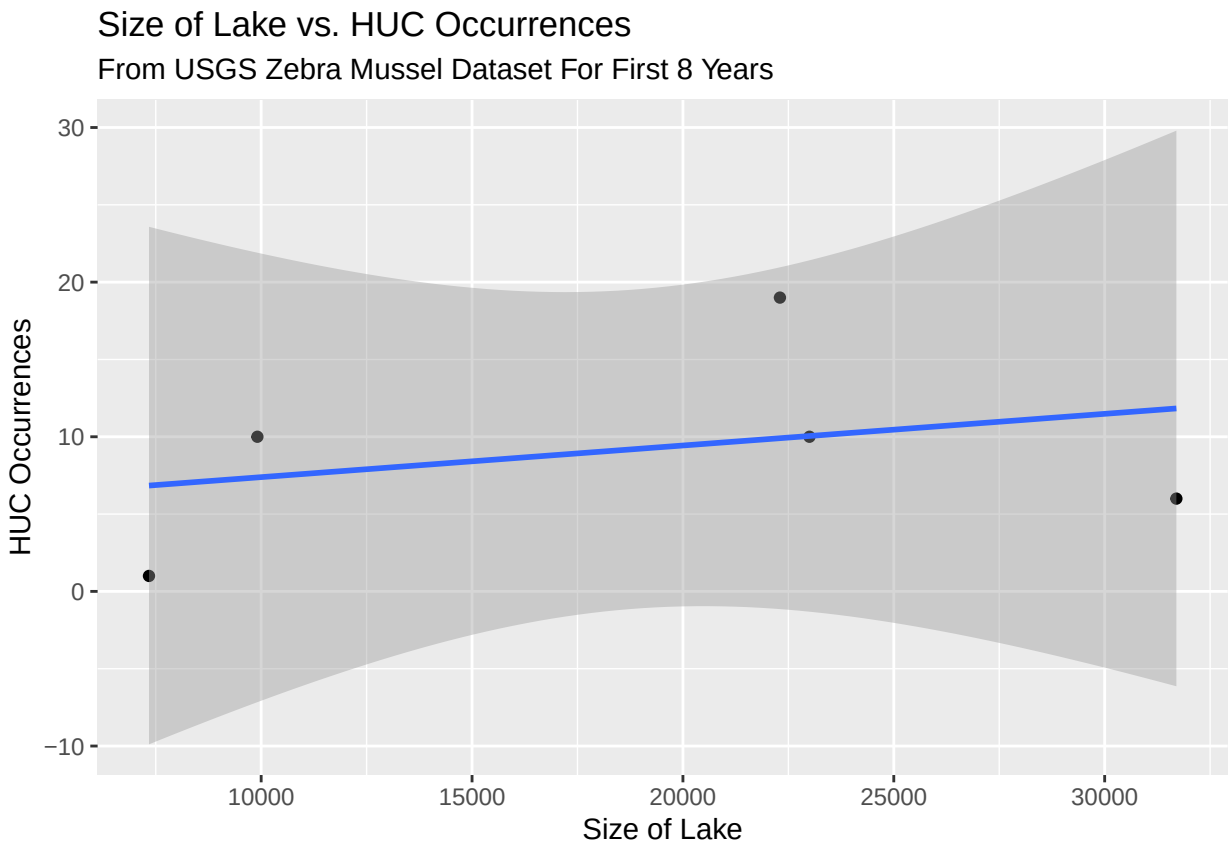
A linear regression analysis was conducted after collecting data on the size of the lake and filtering out the data to only look at the HUC zebra mussel occurrences in the areas of the Great Lakes. The reason for this analysis is to understand if there is a relationship between the size of the lake in water volume, and the number of occurrences on certain HUCs.

Once the analysis was conducted, the p-value associated with the coefficient is not a significant value, and therefore not statistically significant. From the model, its states that a unit increase in the size of the lake leads to a 1.8015439×10^{-4} increase in the HUC Occurrences. Therefore, there is not a significant relationship between the HUC Occurrences, and the size of each Great Lake. The plot also depicts the trend line with the

slope of 1.8015439×10^{-4} and there are confidences bands around the trend line also. This is a small data set with only five data points though, given there are only five Great Lakes.

To conduct further analysis, that data used used the time span of the years 1991-2015 and the data was then split into three different different time period. The time periods are [1991-1999], [2000-2008], and [2009-2015]. Each time period has a linear regression analysis associated with it and the details for each will be described along with a graph of each trend line.

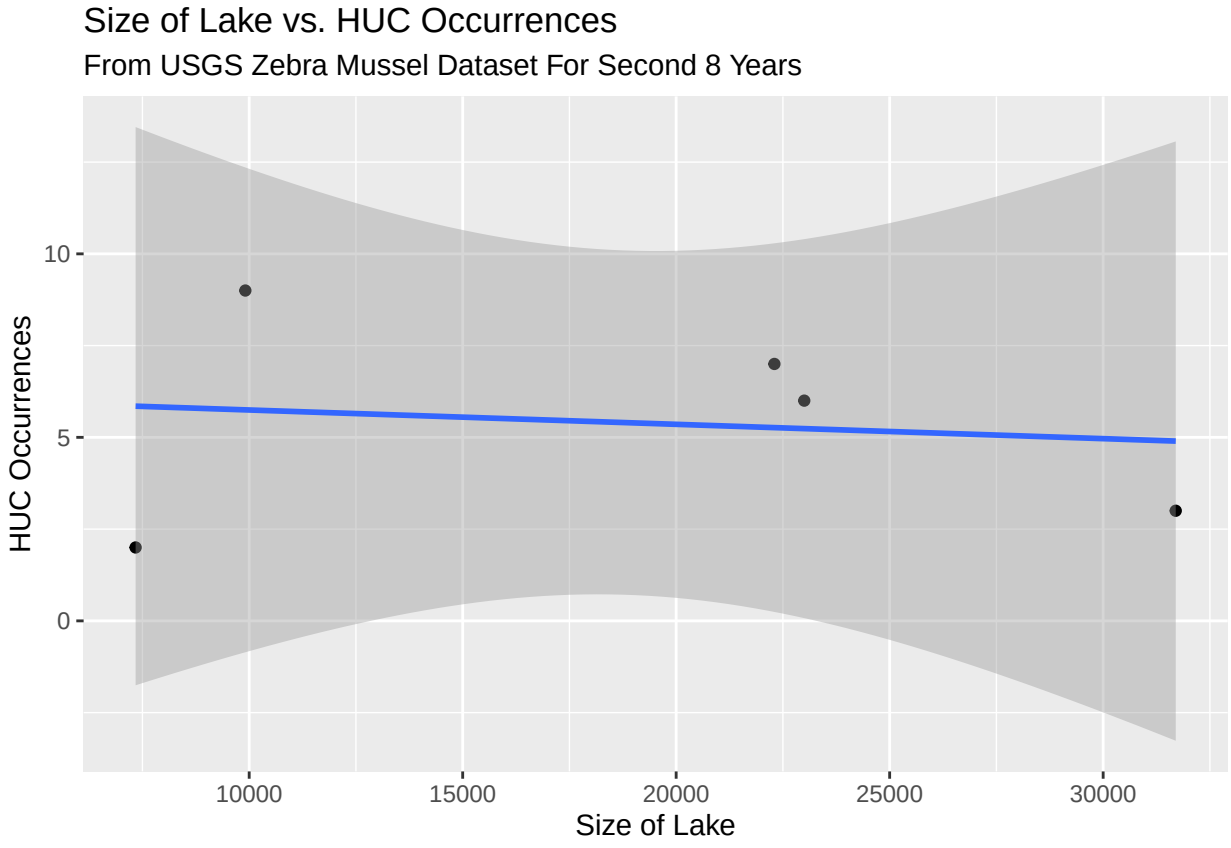
```
##
## Call:
## lm(formula = zm_HUC_occurances_F8 ~ size_of_lake, data = sizeDat)
##
## Residuals:
##       1       2       3       4       5
## -5.83035  9.09380 -0.04949  2.62999 -5.84395
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.3414746  7.5144357   0.711    0.528
## size_of_lake 0.0002047  0.0003596   0.569    0.609
##
## Residual standard error: 7.252 on 3 degrees of freedom
## Multiple R-squared:  0.09748,    Adjusted R-squared:  -0.2034
## F-statistic: 0.324 on 1 and 3 DF,  p-value: 0.609
```



The p-value associated with the coefficient from the data in the range of 1991-1999 is not a significant value, and therefore not statistically significant. From the model, its states that a unit increase in the size of the lake leads to a 2.0469631×10^{-4} increase in the HUC Occurrences between the years of 1991-1999. Therefore,

there is not a significant relationship between the HUC Occurrences, and the size of each Great Lake during that time frame. The plot also depicts the trend line with the slope of 2.0469631×10^{-4} and there are confidences bands around the trend line also.

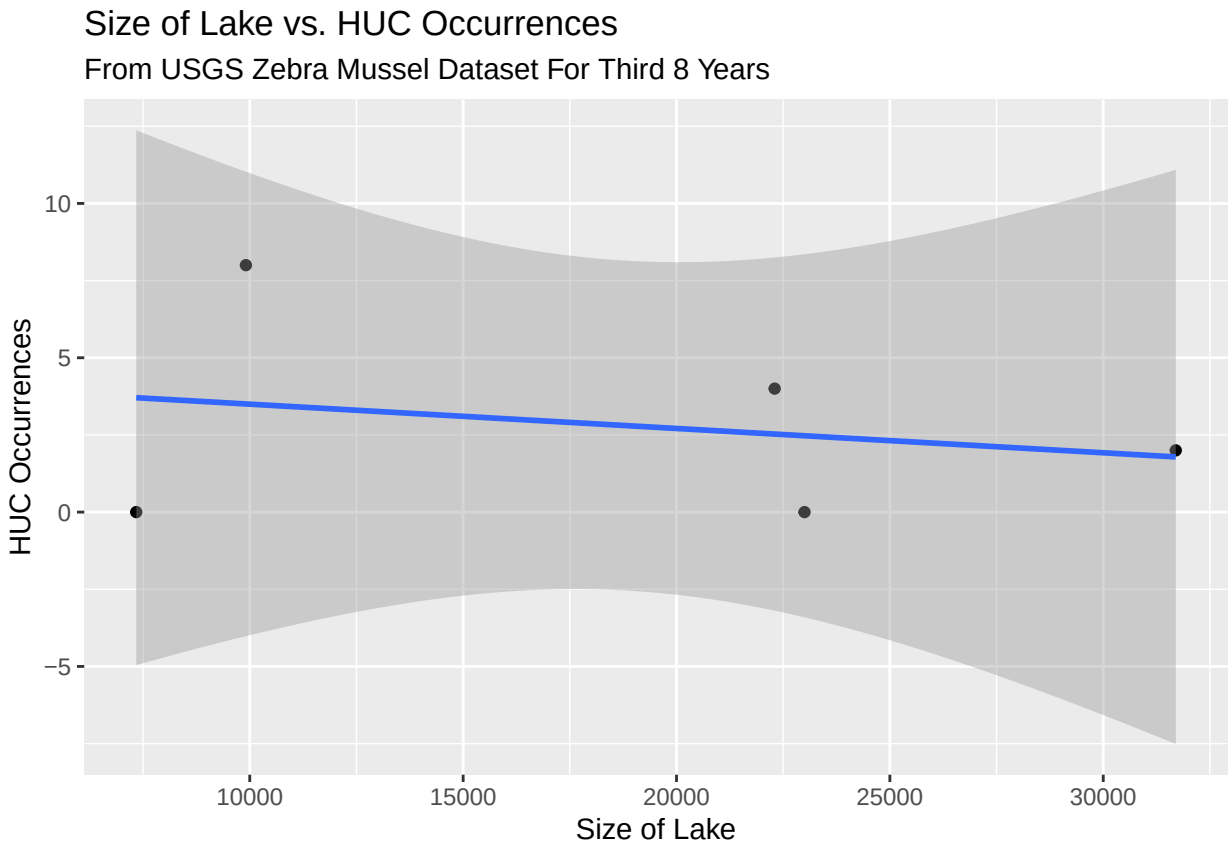
```
##
## Call:
## lm(formula = zm_HUC_occurances_S8 ~ size_of_lake, data = sizeDat)
##
## Residuals:
##       1       2       3       4       5
## -1.8982  1.7347  0.7621  3.2509 -3.8495
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.136e+00  3.415e+00   1.797    0.170
## size_of_lake -3.905e-05  1.634e-04  -0.239    0.827
##
## Residual standard error: 3.295 on 3 degrees of freedom
## Multiple R-squared:  0.01868,    Adjusted R-squared:  -0.3084
## F-statistic: 0.0571 on 1 and 3 DF,  p-value: 0.8265
```



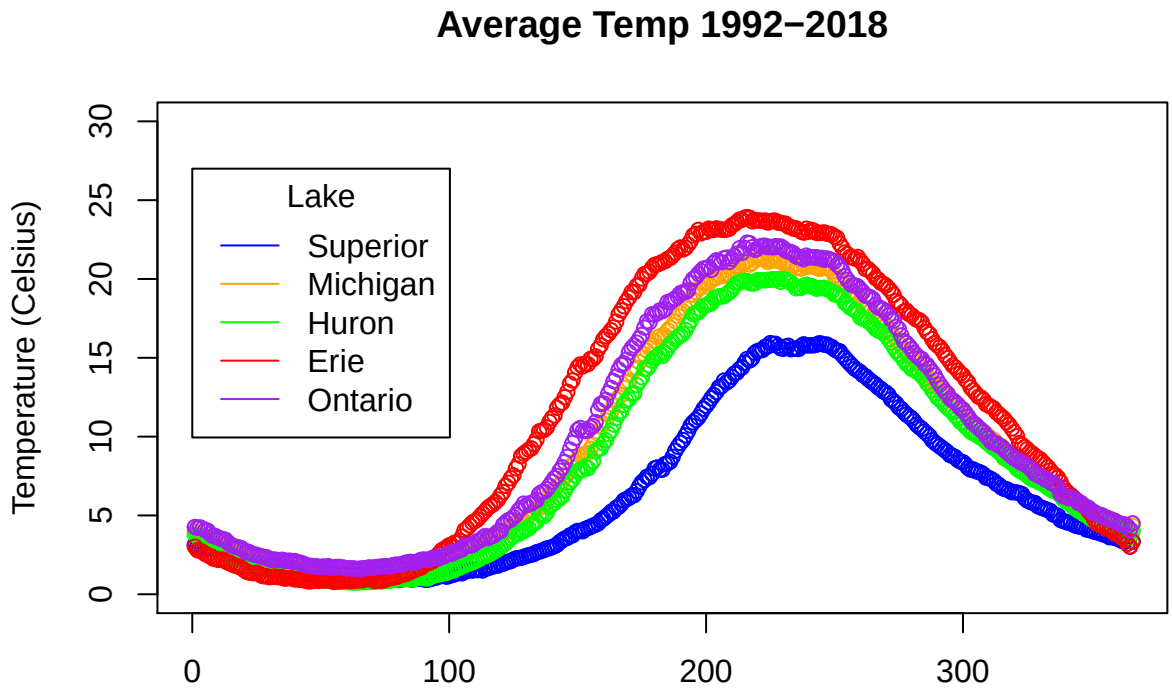
The p-value associated with the coefficient from the data in the range of 2000-2008 is not a significant value, and therefore not statistically significant. From the model, its states that a unit increase in the size of the lake leads to a $-3.9050666 \times 10^{-5}$ increase in the HUC Occurrences between the years of 2000-2008. Therefore, there is not a significant relationship between the HUC Occurrences, and the size of each Great Lake during that time frame. The plot also depicts the trend line with the slope of $-3.9050666 \times 10^{-5}$ and there are confidences bands around the trend line also. This relationship is also negative compared the the first 8 years, which had a positive relationship. Even though neither of the slopes were significant, there could be a change

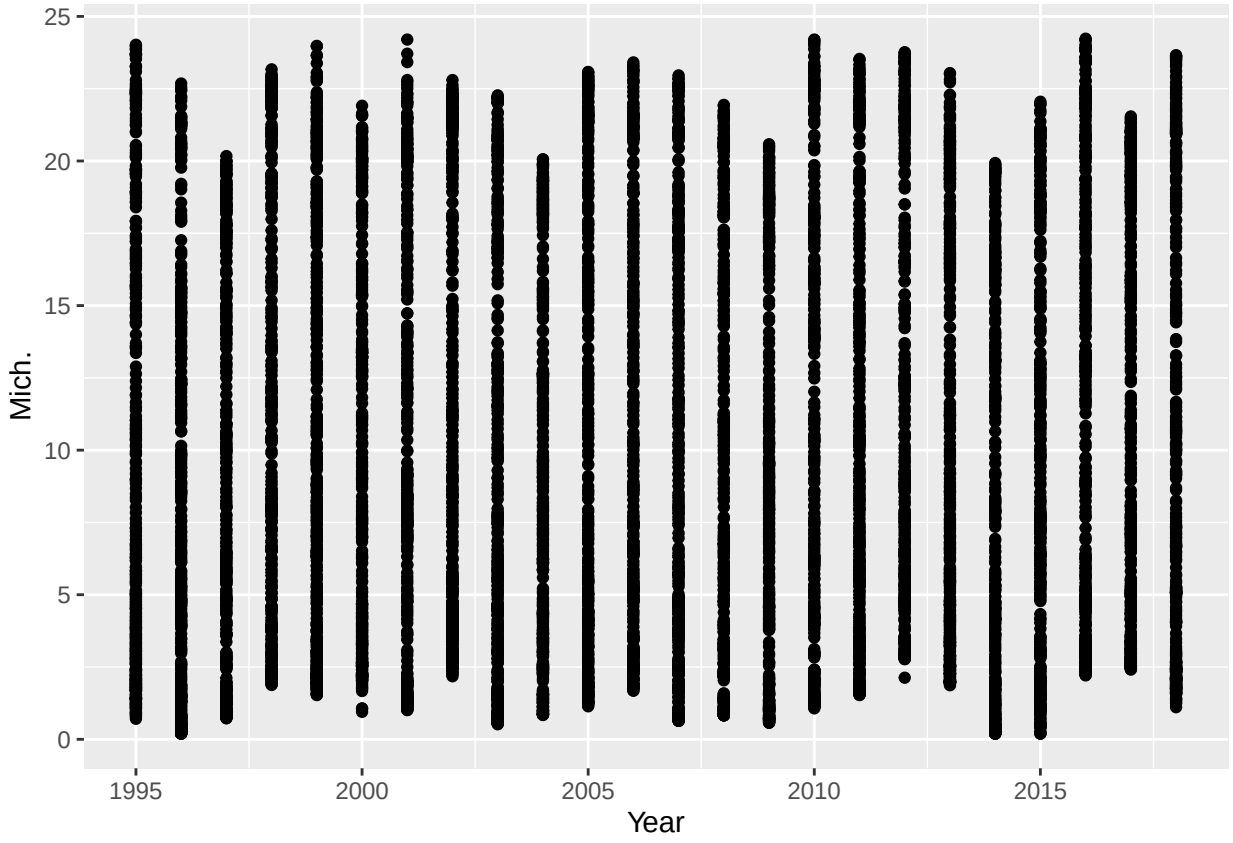
occurring in those time periods that need to be explored.

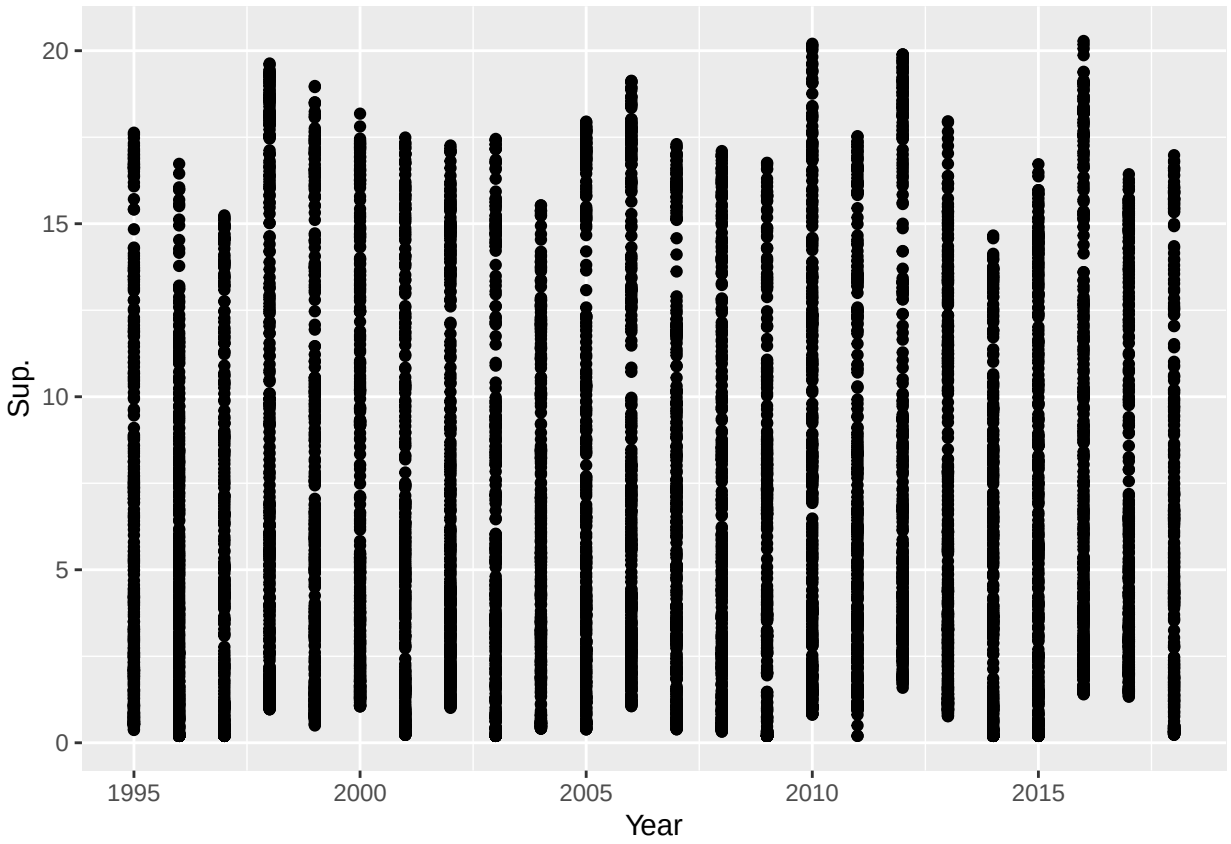
```
##
## Call:
## lm(formula = zm_HUC_occurances_T8 ~ size_of_lake, data = sizeDat)
##
## Residuals:
##       1       2       3       4       5
##  0.2118  1.4717 -2.4732  4.4961 -3.7063
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.284e+00  3.890e+00   1.101    0.351
## size_of_lake -7.874e-05  1.862e-04  -0.423    0.701
##
## Residual standard error: 3.754 on 3 degrees of freedom
## Multiple R-squared:  0.05628,    Adjusted R-squared:  -0.2583
## F-statistic: 0.1789 on 1 and 3 DF,  p-value: 0.7008
```



The p-value associated with the coefficient from the data in the range of 2009-2015 is not a significant value, and therefore not statistically significant. From the model, its states that a unit increase in the size of the lake leads to a -7.87407×10^{-5} increase in the HUC Occurrences between the years of 2009-2015. Therefore, there is not a significant relationship between the HUC Occurrences, and the size of each Great Lake during that time frame. The plot also depicts the trend line with the slope of -7.87407×10^{-5} and there are confidences bands around the trend line also. This coefficient is negative and larger than the previous 8 year period. However, the overall regression analyis shows a positive non significant trend, which insinuates that the first 8 year period has a large affect on the regression analysis since it is the only positive trend out of the three time periods.

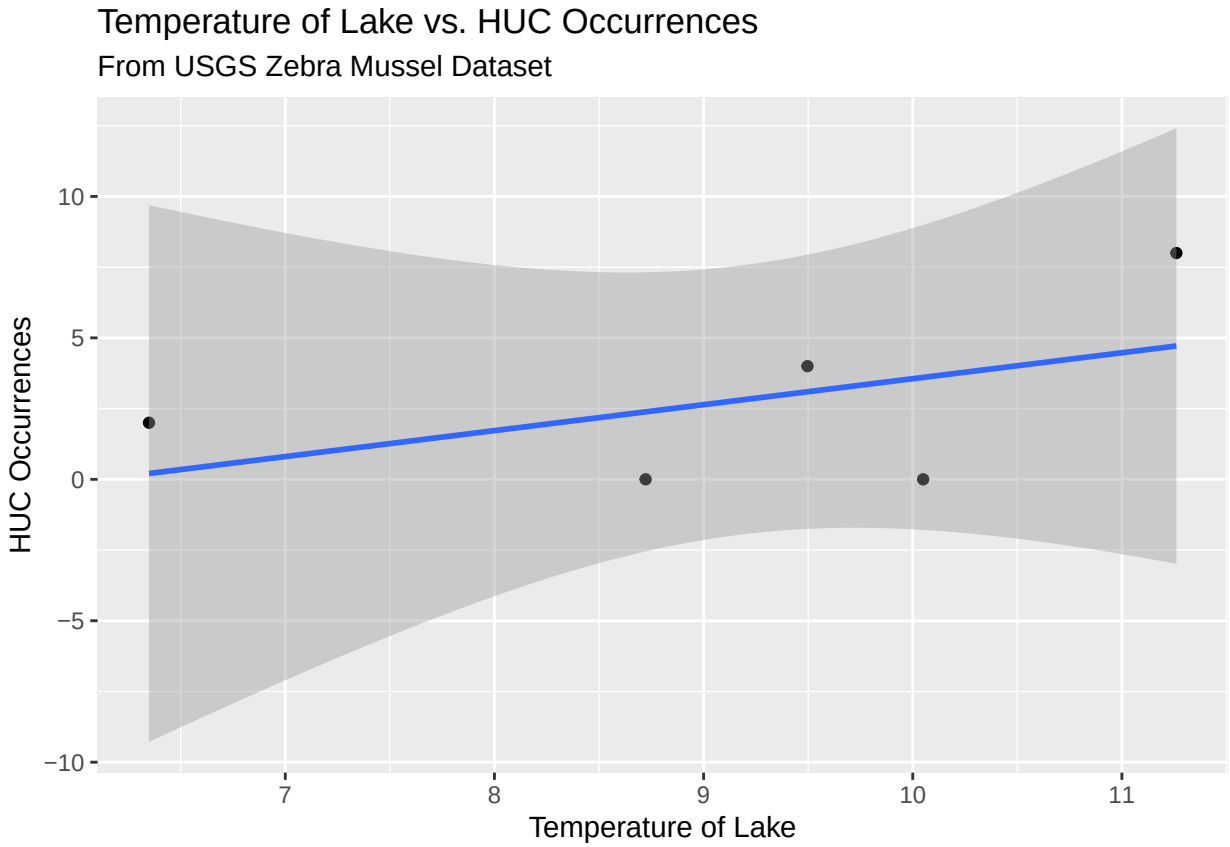

Average Temp 1992−2018
Temperature (Celsius)
0
5
10
15
20
25
30
Lake
Superior
Michigan
Huron
Erie
Ontario
0
100
200
300
Day of the Year


Mich.
25
20
15
10
5
0
1995
2000
2005
2010
2015
Year



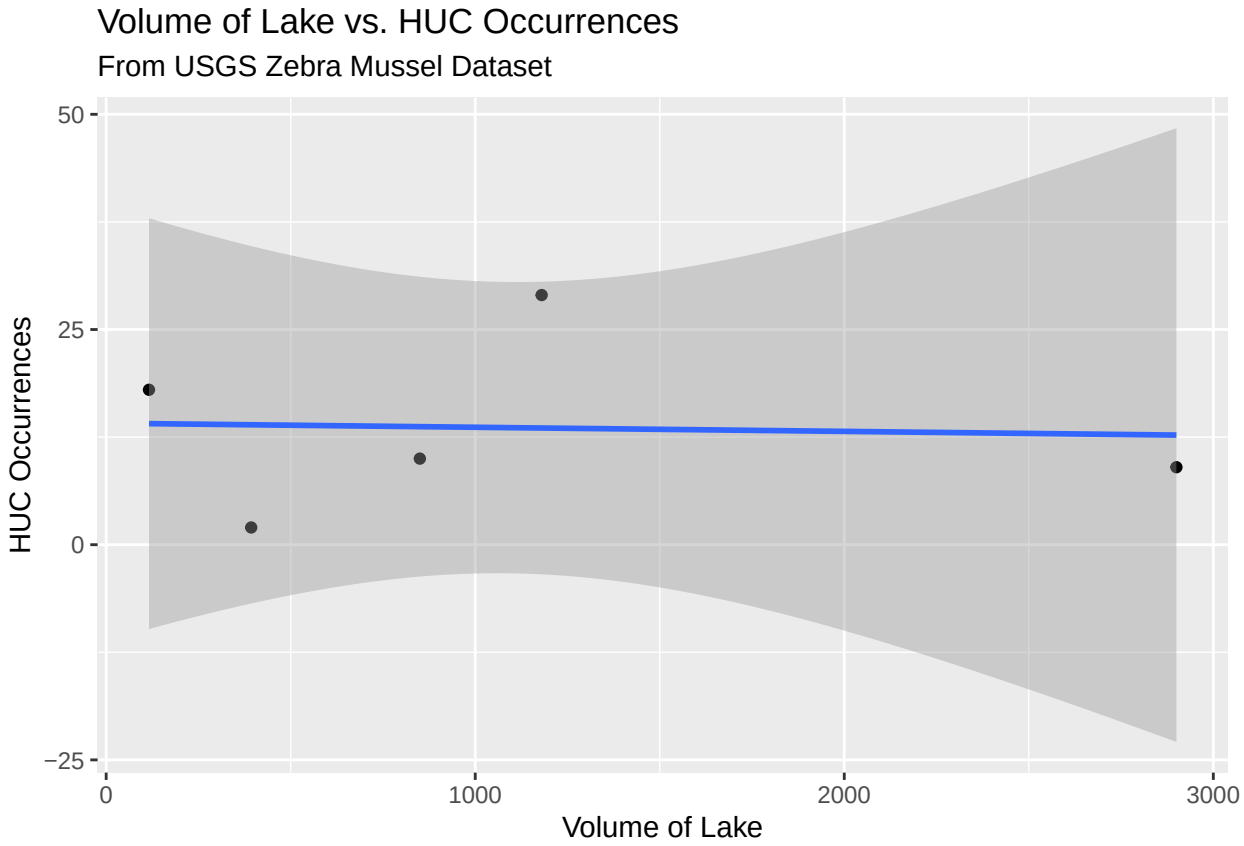
The average temperature of each lake throughout the year shows that Lake Erie get sthe warmest and Lake Superior is the coldest. When specifically looked at, Lake Superior has a high temperature that is about 5 degrees lower than the high temperature of Lake Michigan. There is a lower density of zebra mussels in Lake Superior due to its lower temperatures and harsher conditions.

```
##
## Call:
## lm(formula = zm_HUC_occurances ~ avgTemp, data = TempDat)
##
## Residuals:
##        1        2        3        4        5
##  -0.6765  14.9544  -2.9716   1.5067 -12.8131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.867     29.363    0.03    0.978
## avgTemp        1.388      3.150    0.44    0.689
##
## Residual standard error: 11.54 on 3 degrees of freedom
## Multiple R-squared:  0.06075,    Adjusted R-squared:  -0.2523
## F-statistic: 0.194 on 1 and 3 DF,  p-value: 0.6894
```



Analysis to be done on this, not significant values here.

```
##
## Call:
## lm(formula = zm_HUC_occurances ~ volume_of_lake, data = volumeDat)
##
## Residuals:
##       1       2       3       4       5
##  -3.736  15.444  -3.713   3.937 -11.931
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    14.1184532  7.9560595   1.775    0.174
## volume_of_lake -0.0004766  0.0054405  -0.088    0.936
##
## Residual standard error: 11.89 on 3 degrees of freedom
## Multiple R-squared:  0.002552,   Adjusted R-squared:  -0.3299
## F-statistic: 0.007674 on 1 and 3 DF,  p-value: 0.9357
```



Easton comments

In general, you did a nice job of finding a dataset, learning about the system, and running some analyses. Most of my comments below have to deal with the presentation of the results and organization of the text.

Things to do for final draft:

- In general the figures look nice, but they can be cleaned up with little things like the axes titles
- Reorder the text at the beginning so it reads with a traditional introduction, methods, and results
- add appropriate figure captions
- include formatting to make R markdown output look nicer (e.g. section headers, remove chunk messages)
- Have a short paragraph that summarizes and puts your results in context