Data Analysis of Ellerbe Creek

Advaith Cheruvu

9/29/2021

Importance of Study

Assessing the water quality of local bodies of water is very important for maintaining safe drinking water, a healthy ecosystem, and preventing environmental damage. This document will focus on the data from Ellerbe Creek which is a tributary of the Neuse River in North Carolina, United States. The Ellerbe creek flows for more than 20 miles through North Durham, eventually draining into the Neuse river. This stream has often considered the most polluted stream in The Triangle and a watershed improvement plan was put in place in 2010. Knowledge about specific water quality indicators would be very helpful in determining relationships between indicators as well as the overall environmental health of the stream. This shows that studying the water quality data of this stream can help local conservation efforts focus on specific indicators and assess the current damage of the stream. The Ellerbe Creek data I will be analyzing has a 11 indicators that were measured.

Overview Water Temperature

Water temperature affects a multitude of other water quality factors. Firstly, a higher water temperature means that there are fewer dissolved gasses. This can reduce the available oxygen for animals as well as carbon dioxide for plants. Also, some animals and plants cannot survive under temperatures that are too hot or cold. These animals are sensitive to temperature changes, showing the importance of maintaining a normal temperature. Also, since water a high heat capacity, it’s temperature doesn’t change very easily, which has what allowed some animals and plants to adapt to the nearly constant water temperature.

Overview of Pressure

Pressure is a result of the weight of the above air. This force often causes stream flow and, subsequently, turbidity to increase. Pressure can make some animals feel unpleasant, causing them to migrate or move if there is long term discomfort. Pressure can be constant or infrequently change based on the environment. This causes the stressors to take form either individually or in combination. Overall, pressure helps maintain a healthy ecosystem.

Stream Depth

Stream depth affect characteristics of water quality because of the effect stream depth has on the temperature. For example, a deeper stream would likely be cooler than a more shallow stream since the sunlight cannot penetrate to the deepest layers of the stream as easily. This helps determine the healthy temperature range for the animals and plants in a stream. Also, streams with a constant depth likely flow slower than streams with a changing depth since there is nothing to slow down the stream.

Stream flow

The flow of a stream is directly related to the amount of water moving from the watershed to a stream channel. This is affected by weather, seasons, vegetation. Faster flowing streams will likely be less affected by pollution than small slower streams. Also, the flowing speed of a stream determines the kinds of organisms that can live in the stream (since some prefer faster and others prefer slower). Lastly, fast moving streams likely have higher levels of dissolved oxygen than slower streams since faster moving streams are better aerated when they are flowing.

Dissolved Oxygen

Dissolved oxygen is important for all aquatic animals to breathe. Dissolved oxygen levels are caused by organic material levels, season changes, temperature, and the flow of the water. Most organisms have a dissolved oxygen tolerance range. For levels below 3 mg/L some sensitive animals usually migrate away. Bodies of water with dissolved oxygen levels below 1 mg/L are hypoxic and typically devoid of life.

pH: Acidity

pH is a measure of how acidic/basic water is, going from zero to fourteen. Water is usually at 7, meaning that it is neutral, with drinking water being anywhere between 6.5 and 8.5. Again, there are organisms that are very sensitive to pH changes. Changes in population of these organisms indicate major changes in the ecosystem.

NO3: Nitrate

Nitrogen is cycled through the ecosystem through the nitrogen cycle. This cycle creates NO3 which can cause damage in high quantities. NO3 is a fertilizer for some species, favoring those species and decreasing the biodiversity of an ecosystem. NO3, in high quantities, also has the effect of acidifying water bodies, causing further ecosystem damage. Denitrifying bacteria naturally cycle nitrogen, so it is important reduce the production of NO3 and allow these bacteria to cycle nitrogen.

Turbidity

Turbidity, measured in units of Nephelometric Turbidity Units (NTU), is the measure of the relative clarity of water. A higher turbidity means that the water is more cloudy and opaque while a lower turbidity means that the water is more transparent and clear. Turbidity is mostly caused by suspended particles in the water. This means that higher particulate matter more greatly affects the light penetration and ecological productivty. Plants at the deepest part of a body of water cannot effectively photosynthesize if the turbidity is too high.

Conductance

Conductance is the measurement of water to pass an electric current. This can be affected by dissolved salts and inorganic chemicals. As salinity increases, conductivity also decreases, but most organic compounds hinder the conductivity of the water. Also, warmer water has a higher conductivity than cooler water. Each body of water tends to have a range of conducitivaty that can be used a baseline for comparison analysis.

NH4: Ammonia

Ammonia is another form of nitrogen that exists in the ecosystem. Unlike NO3, ammonia is not a nutrient, but instead causes damage to aquatic organisms. Ammonia, in high levels can become a toxin to aquatic organisms. This can cause toxic buildup in internal tissues and blood, potentially causing death. Ammonia is greatly affected by pH and temperature. Again, we must allow the nitrogen cycle to filter out ammonia to maintain a constant healthy level.

Saturation

Saturation is the measure of dissolved gasses in a body of water. This is affected by the temperature of the water, since excited water molecules can hold more gasses. Gas saturation levels must remain relatively constant in order to avoid certain risks. If there isn’t enough dissolved oxygen, then there is likely eutrophication in the body of water.

Setting up Workspace and Installing Packages

The code below shows the preparation of the workspace and cleaning up the environment, as well as installing and loading packages.

```
#
# clean up and set up
rm(list=ls())
setwd("/Users/advai/Documents/DSFS")
#
# install and load libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggdendro)
library(ggalt)
```

```
## Registered S3 methods overwritten by 'ggalt':
##   method                  from
##   grid.draw.absoluteGrob  ggplot2
##   grobHeight.absoluteGrob ggplot2
##   grobWidth.absoluteGrob  ggplot2
##   grobX.absoluteGrob      ggplot2
##   grobY.absoluteGrob      ggplot2
```

```
library(plotluck)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
## 
##     combine
```

```
source("myfunctions.R")
#
```

Loading and Looking at Water Quality Data

After reading the water quality data into R studio, I can look at it using a variety of commands. The most important commands are “summary” which reports the quartile data, “str” which reports the type of data in the dataframe, and “names” which reports the names of the columns of data.

```
#
# load wq data
wq <- read.csv(file = "C:\\Users\\advai\\Documents\\DSFS\\ellerbee.csv",header=T)
#
# looking at the data
names(wq)
```

```
##  [1] "WaterTemp"  "Pressure"   "StrDepth"   "CFS"        "DO"        
##  [6] "pH"         "NO3"        "Turb"       "Conduct"    "NH4"       
## [11] "saturation"
```

```
summary(wq)
```

```
##    WaterTemp        Pressure        StrDepth         CFS        
##  Min.   :11.60   Min.   :750.0   Min.   :2.00   Min.   : 0.590  
##  1st Qu.:18.70   1st Qu.:755.0   1st Qu.:4.00   1st Qu.: 4.643  
##  Median :20.80   Median :757.5   Median :4.90   Median : 8.005  
##  Mean   :19.36   Mean   :758.6   Mean   :4.68   Mean   : 9.558  
##  3rd Qu.:22.05   3rd Qu.:765.0   3rd Qu.:5.50   3rd Qu.:13.793  
##  Max.   :23.20   Max.   :767.0   Max.   :7.00   Max.   :21.890  
##                                                 NA's   :6       
##        DO              pH             NO3              Turb       
##  Min.   : 5.00   Min.   :6.320   Min.   :0.000   Min.   :  0.00  
##  1st Qu.: 6.16   1st Qu.:6.801   1st Qu.:1.000   1st Qu.:  8.20  
##  Median : 8.50   Median :6.990   Median :1.400   Median : 16.85  
##  Mean   : 8.81   Mean   :6.997   Mean   :1.998   Mean   : 24.97  
##  3rd Qu.: 9.40   3rd Qu.:7.237   3rd Qu.:2.500   3rd Qu.: 22.74  
##  Max.   :14.60   Max.   :7.530   Max.   :7.200   Max.   :112.00  
##  NA's   :2                                                        
##     Conduct            NH4              saturation   
##  Min.   : 170.0   Min.   :  0.2000   Min.   :0.930  
##  1st Qu.: 182.7   1st Qu.:  0.3375   1st Qu.:1.002  
##  Median : 233.9   Median :  0.4500   Median :1.150  
##  Mean   : 413.7   Mean   : 42.2311   Mean   :1.133  
##  3rd Qu.: 702.0   3rd Qu.:  4.0500   3rd Qu.:1.275  
##  Max.   :1022.0   Max.   :600.0000   Max.   :1.300  
##                   NA's   :2          NA's   :14     
```

```
str(wq)
```

```
## 'data.frame':    20 obs. of  11 variables:
##  $ WaterTemp : num  12.5 12.3 11.7 11.6 19.8 20.6 23.1 23.2 21.4 22.3 ...
##  $ Pressure  : num  765 765 765 765 758 ...
##  $ StrDepth  : num  7 7 7 7 4 4 4.8 4.8 5 5 ...
##  $ CFS       : num  NA NA NA NA 7.46 ...
##  $ DO        : num  12.3 12.8 14.1 14.6 8.5 8.7 8.5 8.3 5.3 5 ...
##  $ pH        : num  7.17 6.93 6.6 6.5 7.47 7.35 7.47 7.53 7.12 7.2 ...
##  $ NO3       : num  1 0.7 3.7 3.5 5.4 7.2 1.4 1.1 1.1 1.2 ...
##  $ Turb      : num  16.3 14.7 32 36 16.7 ...
##  $ Conduct   : num  717 712 702 702 1022 ...
##  $ NH4       : num  100 600 NA NA 0.5 0.4 0.36 0.37 0.3 0.33 ...
##  $ saturation: num  1.1 1.2 1.3 1.3 0.93 0.97 NA NA NA NA ...
```

```
dim(wq)
```

```
## [1] 20 11
```

```
class(wq)
```

```
## [1] "data.frame"
```

```
glimpse(wq)
```

```
## Rows: 20
## Columns: 11
## $ WaterTemp  <dbl> 12.5, 12.3, 11.7, 11.6, 19.8, 20.6, 23.1, 23.2, 21.4, 22.3,~
## $ Pressure   <dbl> 765.0, 765.0, 765.0, 765.0, 757.5, 757.5, 767.0, 767.0, 756~
## $ StrDepth   <dbl> 7.0, 7.0, 7.0, 7.0, 4.0, 4.0, 4.8, 4.8, 5.0, 5.0, 4.5, 4.5,~
## $ CFS        <dbl> NA, NA, NA, NA, 7.46, 7.23, 14.01, 12.04, 14.52, 21.89, 8.0~
## $ DO         <dbl> 12.30, 12.80, 14.10, 14.60, 8.50, 8.70, 8.50, 8.30, 5.30, 5~
## $ pH         <dbl> 7.170, 6.930, 6.600, 6.500, 7.470, 7.350, 7.470, 7.530, 7.1~
## $ NO3        <dbl> 1.00, 0.70, 3.70, 3.50, 5.40, 7.20, 1.40, 1.10, 1.10, 1.20,~
## $ Turb       <dbl> 16.30, 14.70, 32.00, 36.00, 16.70, 3.80, 8.60, 10.80, 21.25~
## $ Conduct    <dbl> 717.00, 712.00, 702.00, 702.00, 1022.00, 1013.00, 195.70, 1~
## $ NH4        <dbl> 100.00, 600.00, NA, NA, 0.50, 0.40, 0.36, 0.37, 0.30, 0.33,~
## $ saturation <dbl> 1.10, 1.20, 1.30, 1.30, 0.93, 0.97, NA, NA, NA, NA, NA, NA,~
```

```
head(wq)
```

```
##   WaterTemp Pressure StrDepth  CFS   DO   pH NO3 Turb Conduct   NH4 saturation
## 1      12.5    765.0        7   NA 12.3 7.17 1.0 16.3     717 100.0       1.10
## 2      12.3    765.0        7   NA 12.8 6.93 0.7 14.7     712 600.0       1.20
## 3      11.7    765.0        7   NA 14.1 6.60 3.7 32.0     702    NA       1.30
## 4      11.6    765.0        7   NA 14.6 6.50 3.5 36.0     702    NA       1.30
## 5      19.8    757.5        4 7.46  8.5 7.47 5.4 16.7    1022   0.5       0.93
## 6      20.6    757.5        4 7.23  8.7 7.35 7.2  3.8    1013   0.4       0.97
```

As you can see, these commands provide useful information about the quality of the data as well as what the data is.

Replacing Column Names

Based on the “names” command from earlier, the column names of the data in this dataset is very ambiguous. By using a “dplyr” command “rename” we can rename the columns so that it is easier to read.

```
#
# rename column
names(wq) <- c("Water_temperature", "Pressure","Stream_depth", "Cubic_flow_sec", "Dissolved_oxyg
en", "pH", "NO3", "Turbidity", "Conductance", "NH4", "Saturation")
```

```
#
# renamed columns
names(wq)
```

```
##  [1] "Water_temperature" "Pressure"          "Stream_depth"
##  [4] "Cubic_flow_sec"    "Dissolved_oxygen"  "pH"
##  [7] "NO3"               "Turbidity"         "Conductance"
## [10] "NH4"               "Saturation"
```

Checking for Missing Data

Most columns have missing data, either as “NA” or data that just doesn’t make sense. I can check for missing data using the commands below which print the rows with missing data. After knowing that I have missing data, I can run the command below which tells me which columns have missing data.

```
#
# Check for missing data
clean <- ifelse(complete.cases(wq) == TRUE, 1, 0)
paste("There are", dim(wq)[1]-sum(clean), "rows with missing data.")
```

```
## [1] "There are 18 rows with missing data."
```

```
#
# Find which columns have missing data
missingDataCol <-colnames(wq)[apply(wq, 2, anyNA)]
paste0("The following columns have missing data: ", missingDataCol)
```

```
## [1] "The following columns have missing data: Cubic_flow_sec"
## [2] "The following columns have missing data: Dissolved_oxygen"
## [3] "The following columns have missing data: NH4"
## [4] "The following columns have missing data: Saturation"
```

Replacing Missing Data

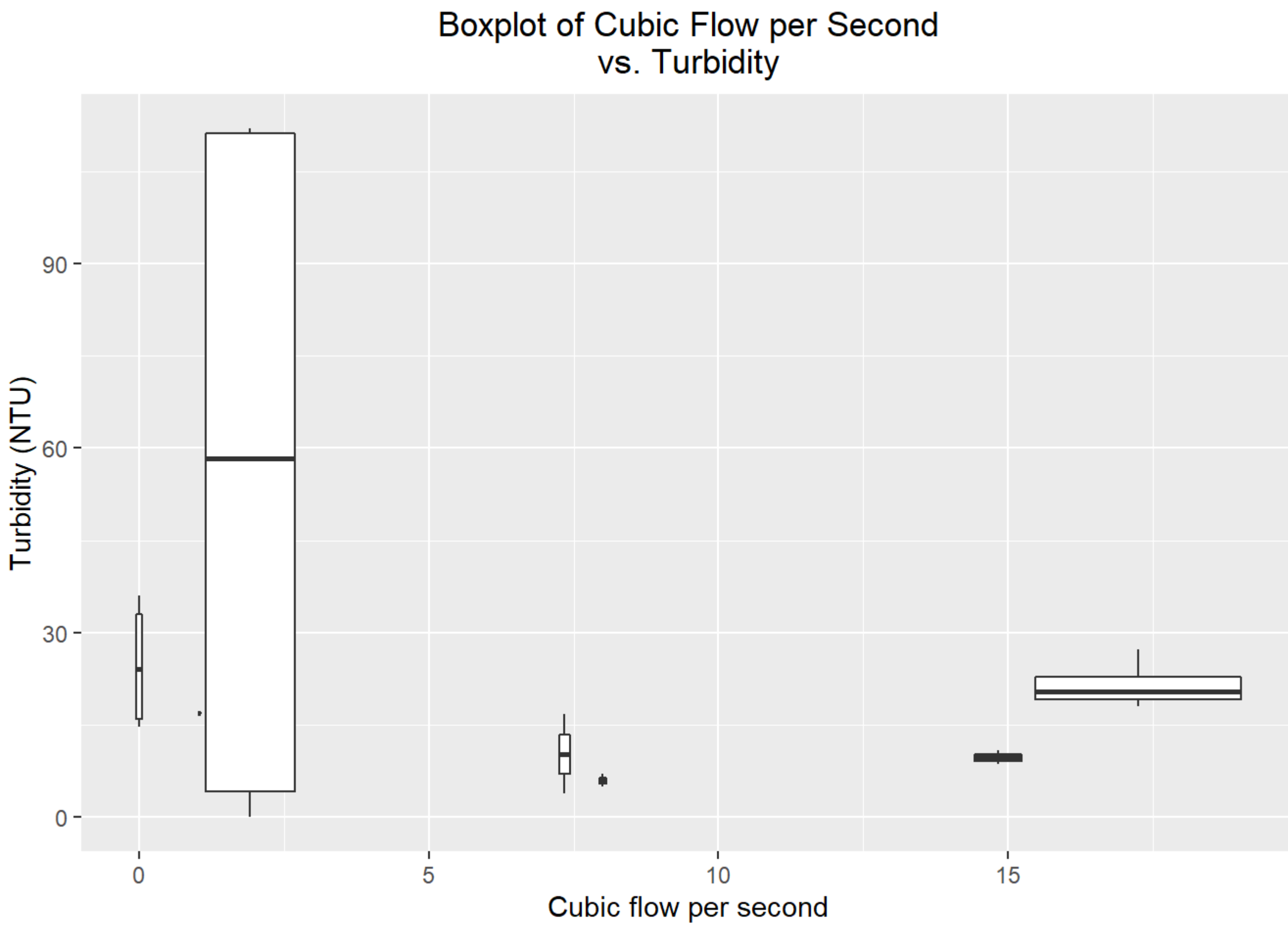
I am going to replace the missing values with 0 since there isn’t enough data to confidently replace the missing data with the mean.

```
# replace columns of missing data with 0
wq[is.na(wq)] <- 0
```

Box Plots

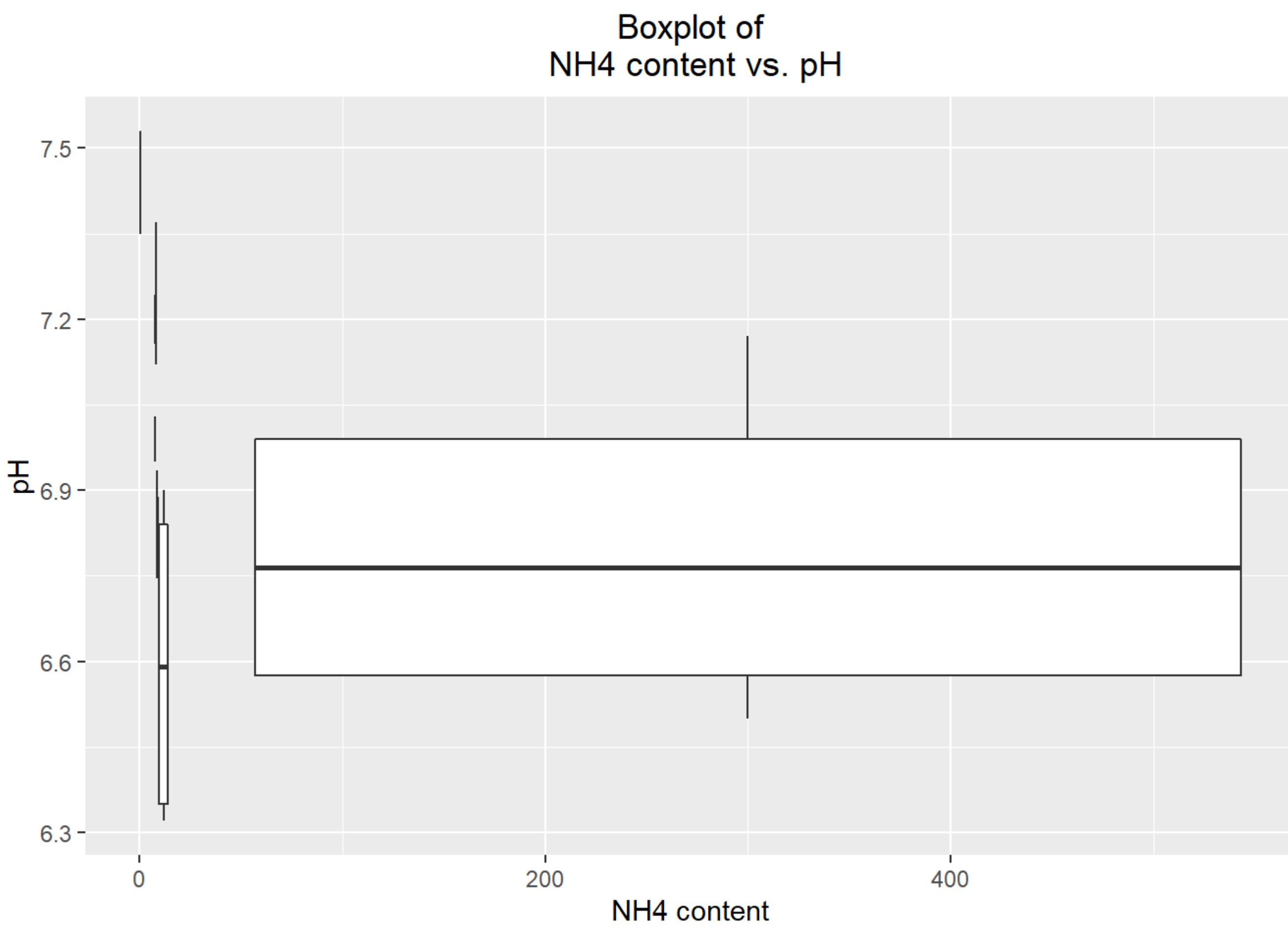
I will be creating 8 graphics showing the relationship between a few variables in this dataset. First, I’ll be creating a box plot to show the relationship between the flow of the stream and the turbidity

```
# cfs vs Turbidity boxplot
bp1 <- ggplot(wq, aes(group = Stream_depth,        # groups the data
                         x=Cubic_flow_sec,           # sets the x
                         y=Turbidity))+              # sets the y
  geom_boxplot() +                                   # tells R to create a boxplot
  xlab("Cubic flow per second") +                    # labels the x axis
  ylab("Turbidity (NTU)") +                                   # labels the y axis
  ggtitle("Boxplot of Cubic Flow per Second\nvs. Turbidity") +      # labels the plot
  theme(plot.title = element_text(hjust = 0.5))      # centers the title

# Prints the plot
bp1
```



This box plot is important to the analysis because this relationship reveals a few patterns. Firstly, the turbidity of the slow moving streams is more varied than the turbidity of the other streams, indicating that there is a great imbalance within the streams at Ellerbe creek. Also, the fact that the fast moving streams generally tend to increase in turbidity shows that the health of the fast moving streams is not in good standing. Conservation efforts should focus on the fast moving streams with some highly turbid slow moving streams. Next, I'll create another box plot showing the relationship between NH4 and pH.

```
# Boxplot of NH4 vs. pH


bp2 <- ggplot(wq, aes(group = Stream_depth,         # groups the data
                     x=NH4,            # sets the x to population
                     y=pH))+              # sets the y to deaths
  geom_boxplot() +                                    # tells R to create a boxplot
  xlab("NH4 content") +                     # labels the x axis
  ylab("pH") +                                    # labels the y axis
  ggtitle("Boxplot of\n NH4 content vs. pH") +      # labels the plot
  theme(plot.title = element_text(hjust = 0.5))

#
# Prints the plot
bp2
```

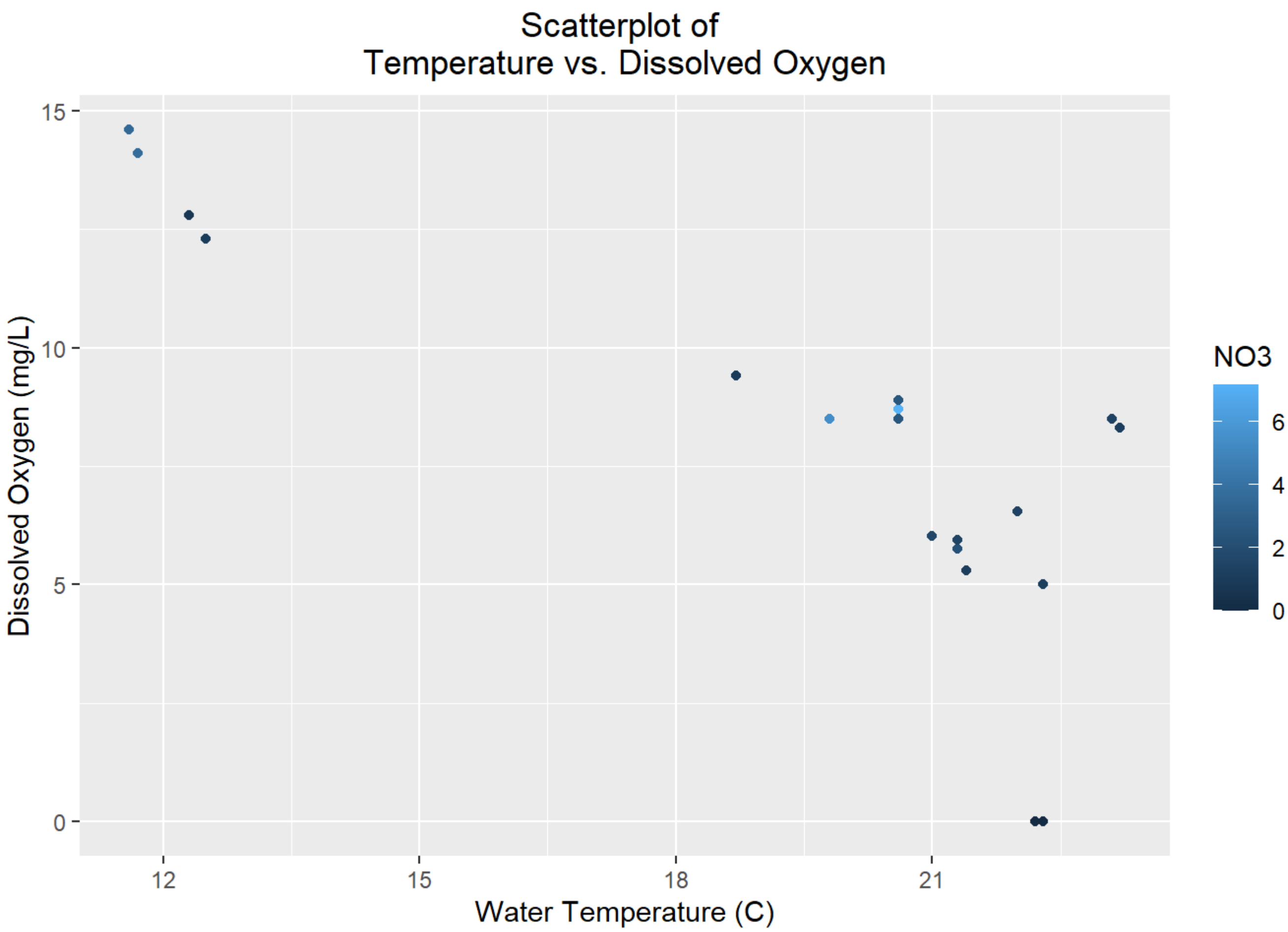


This box plot is important for the analysis because it shows the high range in NH4 content across Ellerbe creek, as well as the varying pH. The pH of the streams both above and below 7, meaning that they are either too basic or acidic. Also, the high range of NH4 content shows that some streams are extremely more polluted than others. This shows that conservation efforts should focus on NH4 content within the streams of Ellerbe Creek.

Scatterplots

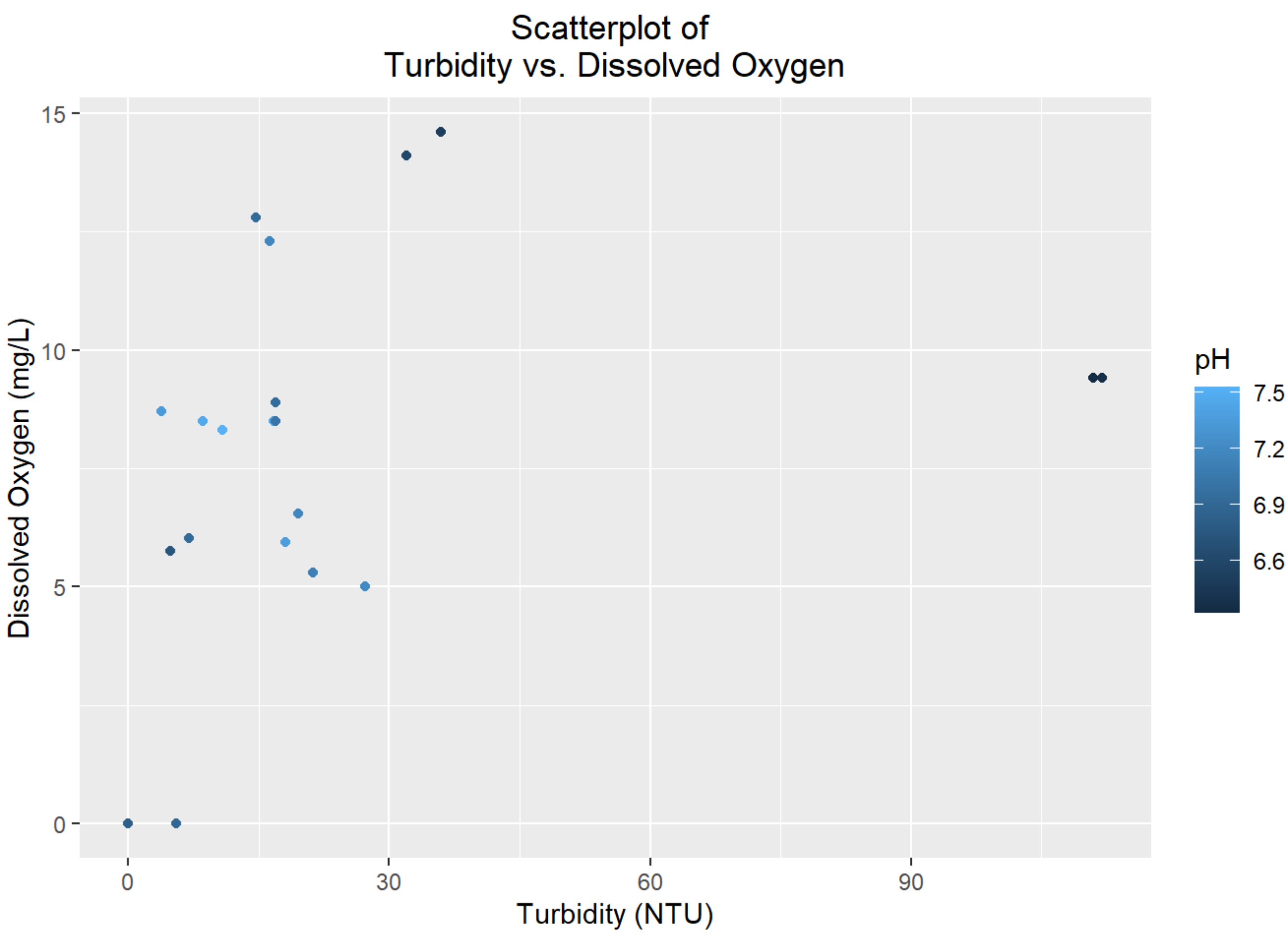
This scatter plot shows the relationship between water temperature and dissolved oxygen.

```
#
# Water Temperature vs. Dissolved Oxygen with NO3

sp1 <-ggplot(wq, aes(
  x=Water_temperature,               # sets the x to adjusted death rate
  y=Dissolved_oxygen,          # sets the y to standard error of the adjusted death rate
  scale_color_gradient(low = "darkblue", high = "deepskyblue")))+ # creates the color gradient
  xlab("Water Temperature (C)") +                      # labels the x axis
  ylab("Dissolved Oxygen (mg/L)") +  # labels the y axis
  ggtitle("Scatterplot of \nTemperature vs. Dissolved Oxygen")+ # labels the plot
  theme(plot.title = element_text(hjust = 0.5))+   # centers the title
  geom_point(aes(color=NO3))                # colors the plot based on NO3

# Prints the plot
sp1
```



This scatter plot is important to the analysis because it shows some key indicators of aquatic environment health. Though some streams have high dissolved oxygen, others are far below. This coupled with the fact that some of these streams have high NO3 indicate some level of eutrophication. Even if the water temperature is high, the water isn't holding as much dissolved oxygen as the cooler streams. This next scatter plot shows the relationship between turbidity and dissolved oxygen.

```
#
# Turbidity vs. Dissolved Oxygen with pH
sp2 <-ggplot(wq, aes(
  x=Turbidity,
  y=Dissolved_oxygen,
  scale_color_gradient(low = "darkblue", high = "deepskyblue")))+
  xlab("Turbidity (NTU)") +
  ylab("Dissolved Oxygen (mg/L)") +
  ggtitle("Scatterplot of \nTurbidity vs. Dissolved Oxygen")+
  theme(plot.title = element_text(hjust = 0.5))+   # centers the title
  geom_point(aes(color=pH))                # colors the plot based on pH

# Prints the plot
sp2
```



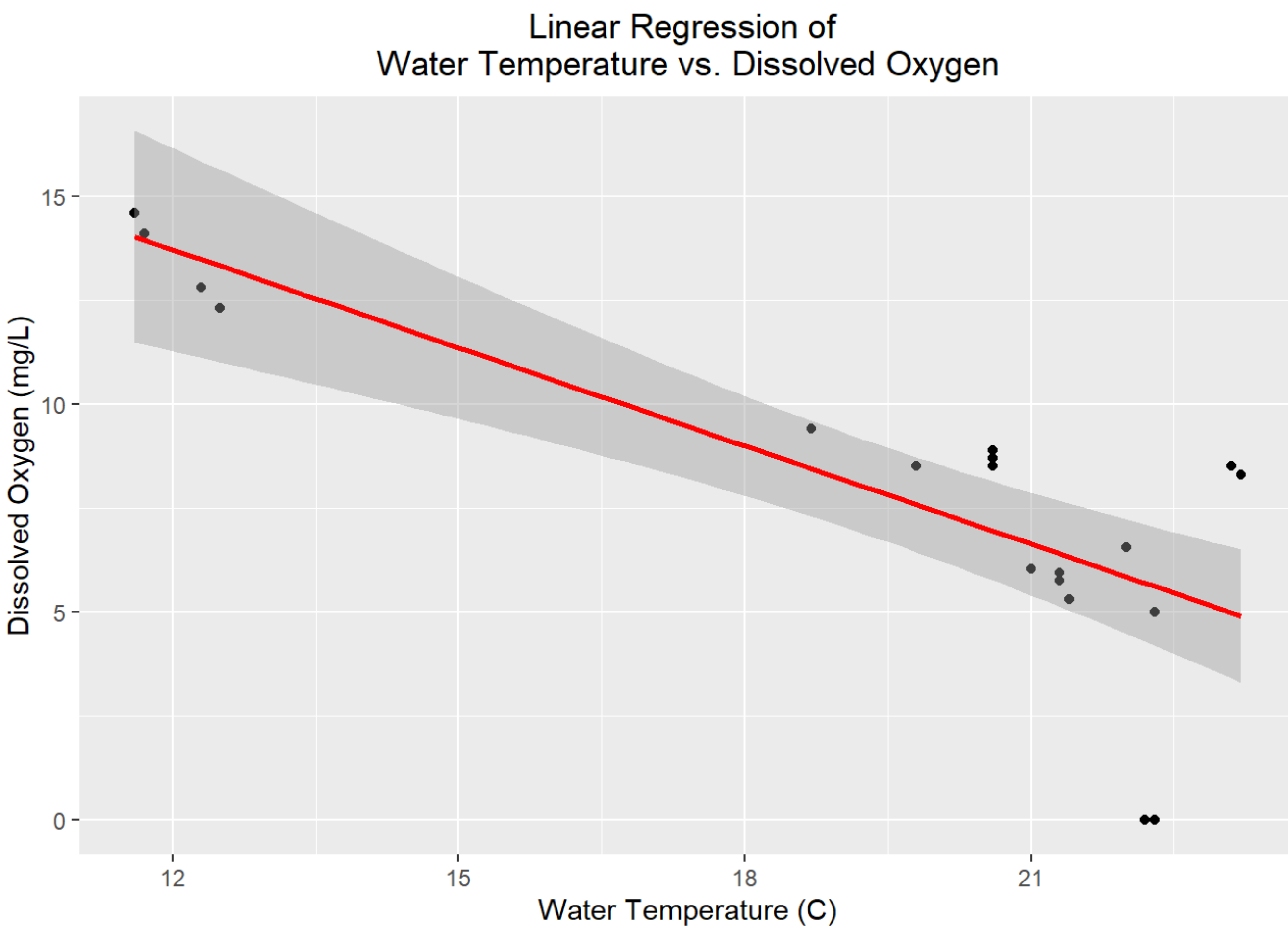
This scatter plot is important to the analysis because shows turbidity and pH in addition to the the dissolved oxygen concentration from the other scatter plot. This plot indicates that most streams are on the borderline between healthy and slightly unhealthy with some streams being extremely turbid. Also, the streams that are more central in terms of dissolved oxygen tend to have a higher pH. Conservation efforts should focus on these few select streams the most and work their way across all the streams in Ellerbe creek.

Linear Regression Plots

This linear regression plot shows the relationship between water temperature and dissolved oxygen.

```
# Water Temperature vs. Dissolved Oxygen

lm1 <- ggplot(wq, aes(x = Water_temperature, y = Dissolved_oxygen)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red") +
  xlab("Water Temperature (C)") +
  ylab("Dissolved Oxygen (mg/L)") +
  ggtitle("Linear Regression of \nWater Temperature vs. Dissolved Oxygen")+
  theme(plot.title = element_text(hjust = 0.5))

# Prints the plot
lm1
```

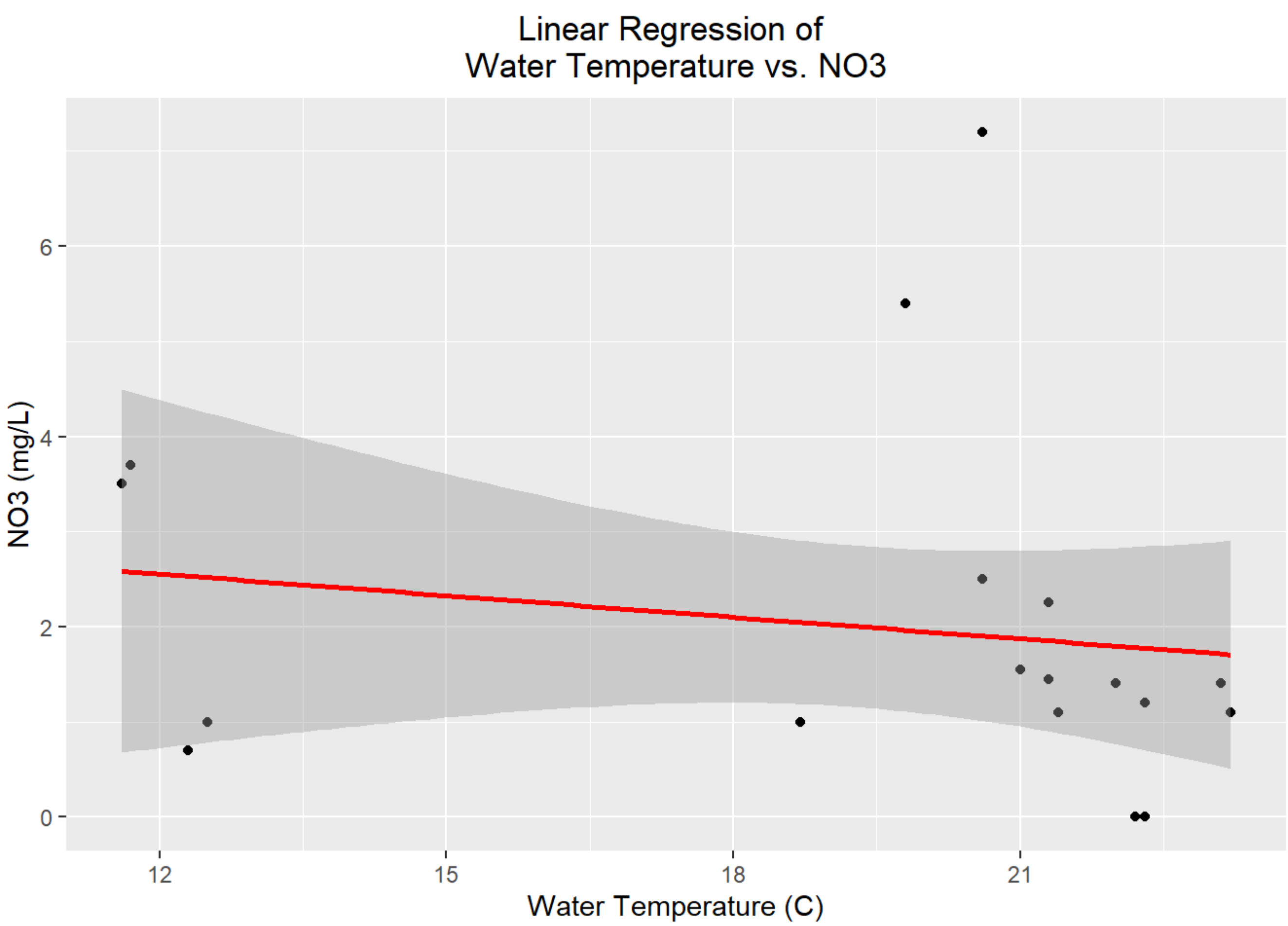
```
## `geom_smooth()` using formula 'y ~ x'
```



This linear regression shows the relationship between dissovled oxygen and water temperature. Prior research indicated that these two indicators should be related, and the Ellerbe creek data supports this, despite the relatively few data points. This graph also shows that many of the streams are warm and have low dissolved oxygen, which could damage the ecosystem. This next linear regression plot shows the relationship between water temperature and NO3

```
# Linear regression of Water Temperature vs. NO3
lm2 <- ggplot(wq, aes(x = Water_temperature, y = NO3)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red") +
  xlab("Water Temperature (C)") +
  ylab("NO3 (mg/L)") +
  ggtitle("Linear Regression of \nWater Temperature vs. NO3")+
  theme(plot.title = element_text(hjust = 0.5))

# Prints the plot
lm2
```

```
## `geom_smooth()` using formula 'y ~ x'
```

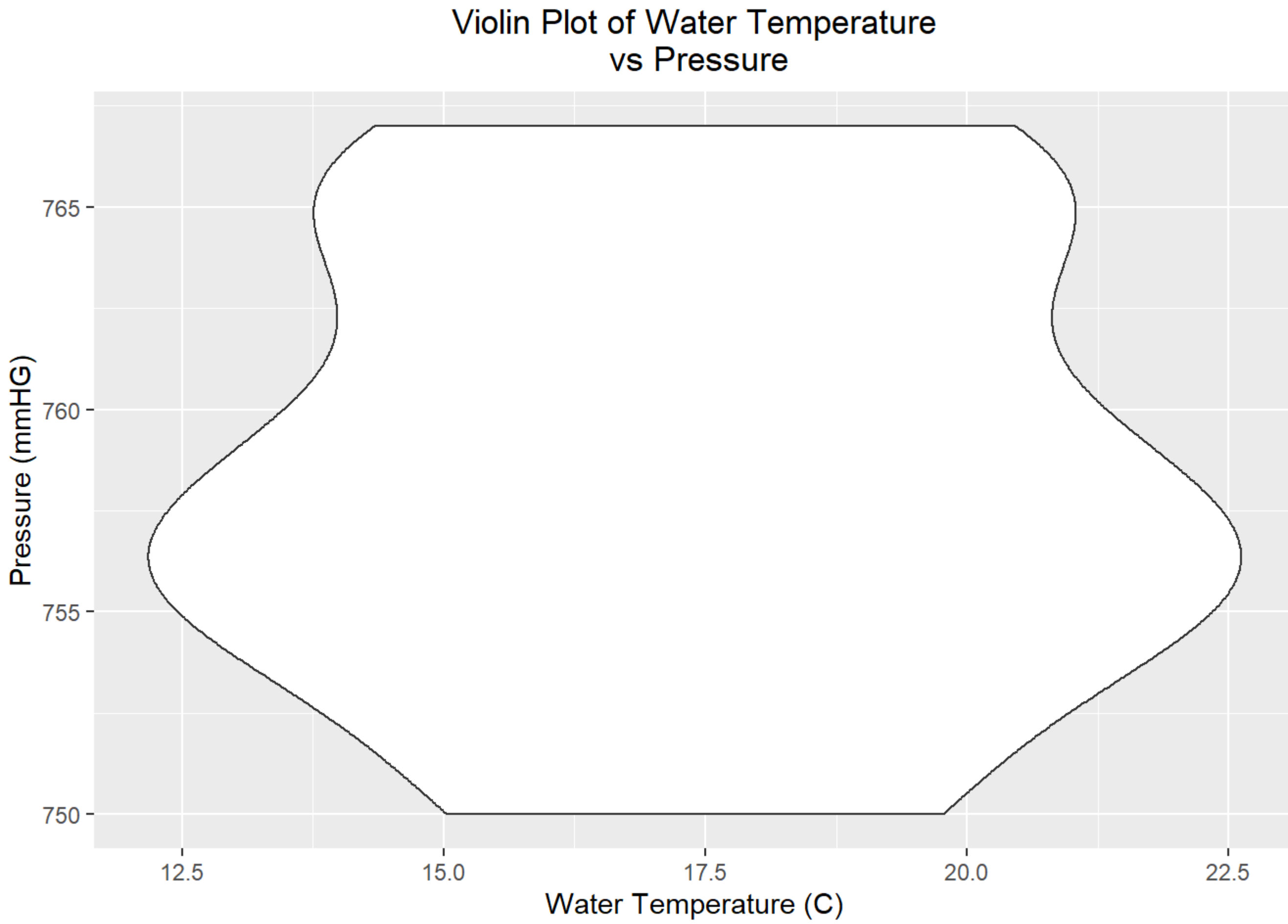


This linear regression plot is important for the analysis because it highlights the lack of relationship between water temperature and NO3. Unlike with dissolved oxygen, NO3 appears to have no relationship with with the water temperature since the data points are too spread out. Conservationists can focus on the few streams with extremely high NO3 and temperature though.

Violin Plots

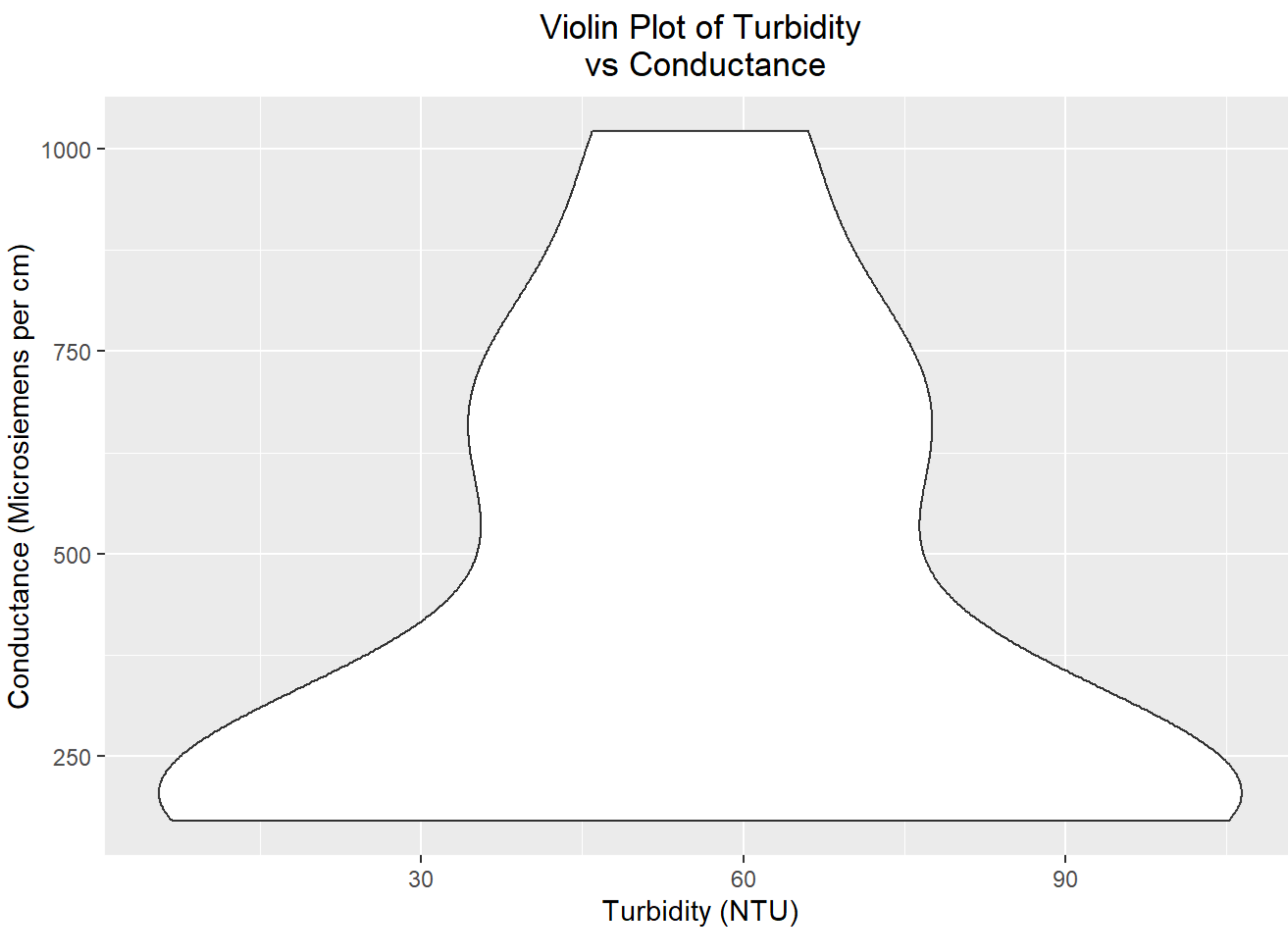
This violin plot shows the relationship between water temperature and pressure.

```
# Water temperature vs pressure
vp1 <- ggplot(wq, aes(x=Water_temperature,                              # sets the x to populati
on
                      y=Pressure)) +                                    # sets the y to deaths
  xlab("Water Temperature (C)") +                                     # labels the x axis
  ylab("Pressure (mmHG)") +                                                    # labels the y axis
  ggtitle("Violin Plot of Water Temperature\n vs Pressure") +           # labels the plot
  theme(plot.title = element_text(hjust = 0.5))+                      # centers the title
  geom_violin()                                                     # tells R to create a boxplot on top
of the violin plot

# Prints the plot
vp1
```



This violin plot is important for the analysis because it shows the relationship between pressure and water temperature. This graph indicates that most pressure values occur at a wide variety of temperatures. These changes in pressure and temperature can harm wildlife and are likely connected to other factors. This next violin plot shows the relationship between turbidity and conductance.

```
# Turbidity vs Conductance
vp2 <- ggplot(wq, aes(x=Turbidity,                            # sets the x to population
                      y=Conductance)) +                            # sets the y to deaths
  xlab("Turbidity (NTU)") +                                 # labels the x axis
  ylab("Conductance (Microsiemens per cm)") +                                        # labe
ls the y axis
  ggtitle("Violin Plot of Turbidity\n vs Conductance") +          # labels the plot
  theme(plot.title = element_text(hjust = 0.5))+              # centers the title
  geom_violin()                                            # tells R to create a boxplot on top
of the violin plot

# Prints the plot
vp2
```




```
#
# EOF (End of File)
```

This violin plot is important for the analysis because it highlights the relationship between turbidity and conductance. The plot indicates that most conductance values are lower with high conductance values becoming more rare in relation to the turbidity. High turbidity that decreases the conductance could indicate that the suspended matter is organic and resists electrical current. This eliminates the possibility of the suspended matter being salts or any other substance that increases electrical current.

Work Cited

“5.1 Stream Flow.” EPA, Environmental Protection Agency, 6 Mar. 2012, archive.epa.gov/water/archive/web/html/vms51.html.

Chase, Valerie. “Nitrate in Ecosystems and Drinking Water.” Carolina.com, Carolina Biological Supply Company, 2021, www.carolina.com/teacher-resources/Interactive/nitrate-in-ecosystems-and-drinking-water/tr10810.tr.

“Ellerbe Creek.” Wikipedia, Wikimedia Foundation, 18 June 2020, en.wikipedia.org/wiki/Ellerbe_Creek.

“Freshwater Multiple Stressor Toolbox.” Pressures - Freshwater Information, 2021, fis.freshwatertools.eu/index.php/infolib/pressures.html.

“Indicators: Conductivity.” EPA, Environmental Protection Agency, 1 June 2021, www.epa.gov/national-aquatic-resource-surveys/indicators-conductivity.

Kreger, Chris. Water Quality, 10 Nov. 2004, www.cotf.edu/ete/modules/waterq3/WQassess4d.html.

“Turbidity and Water.” Turbidity and Water, U.S. Department of the Interior, 2021, www.usgs.gov/special-topic/water-science-school/science/turbidity-and-water?qt-science_center_objects=0#qt-science_center_objects.

Water Quality, waterquality.wordpress.ncsu.edu/.