Water Quality Case Study

Andrew Wang

10/12/2020

Introduction

In this report, we will be analyzing a set of water quality data retrieved from Ellerbee Creek, in Durham North Carolina. First, we'll go over some relevant information regarding water quality and its indicators. After that, we will dive into the graphical analysis conducted by the ggplot2 package of R, and discuss conclusions and the water quality of Ellerbee Creek.

Water quality is a term used to describe the condition of a body of water, pertaining to its physical and chemical attributes. Poor water quality is a risk for aquatic ecosystems, and dangerous for people to drink. Water quality is measuring through a variety of variables, called indicators. Below, we explain the various indicators we are using in the dataset and what they mean.

Water temperature The temperature of water is an important indicator in water quality because all organisms have optimal temperature ranges. Furthermore, the temperature of the water influences other indicators, like dissolved oxygen content. The rate of chemical reactions increases with higher water temperature. However, colder water in the world ocean usually has more nutrients, such as dissolved oxygen. The temperature range of water is very wide, and it is hard to put even a general range on it.

Pressure Pressure refers to the atmospheric pressure of the water, which is measured in mmHg, or millimeters of mercury. Atmospheric pressure influences various aspects of water chemistry, such as boiling point and dissolved oxygen content. One standard "atomosphere" of pressure is 760 mmHg, and it is the pressuer at sea level.

Stream depth Stream depth is very simple, it is just how deep a stream is. However, it can influence many factors of a body of water, such as temperature and dissolved oxygen. A stream which is not very deep will have more uniform indicators, while a stream which is deep will see more variation in its water quality indicators.

Cubic flow (per second) The cubic flow of a stream is a property of moving water that has a impact on many other indicators, like temperature and dissolved oxygen. The flow of a stream is measured by the volume of water that passes over a certain point in a given time. Flow is a combination of water volume and velocity, upon which bodies of water with large cubic flow can disperse pollutants easier, while the opposite is true for those with a low cubic flow. The flow of a stream has an impact on the amount of erosion, as a quickly moving stream will erode more sediment. The average streamflow for a body of moving water is 136 cubic feet per second (CFS).

Dissolved oxygen Dissolved oxygen is the amount of oxygen that has been dissolved into the water. It is an indicator of how healthy a water system is. Each type of aquatic organism requires a different level of dissolved oxygen. Furthermore, temperature and dissolved oxygen are closely linked, and they shared an inverse relationship. Dissolved oxygen is measured in milligrams per L (mg/L) or as a percentage. Generally, a dissolved oxygen of above 8 mg/L is considered sufficent for most aquatic life.

pH The pH value of a body of water is a measure of how acidic or basic it is. The pH scale functions from 0 to 14, values closer to 0 being more acidic and values closer to 14 being more basic. The pH of a water indicates if it is changing chemically. Its a measure of the balance between the position Hydrogen ions and the negative Hydroxide ions. Most fish can tolerate pH values between 5 and 9, but the best waters usually have pH values between 6.5 and 8.2. The normal pH range for water is 6.5 - 8.5.

Nitrogen content Nitrogen is an essential nutrient which is required by all organisms in the creation of amino acids. In order for aquatic plants to take in nitrogen, it must be in the form of Ammonia (covered later), Nitrates, or Nitrites. While nitrogen is an essential compound for growth, too much nitrogen is bad because it will lead to eutrophication and then degradation of water quality. Generally, levels between 0.1 and 0.2 parts per million of nitrogen in a water body is considered good.

Turbidity Turbidity is a measure of water quality which tells the degree to which water loses its transparency due to the particles in it. Generally, higher turbidity is associated with lower quality water, as it means the water has more particles in it, and is not pure water. Turbidity is measuring in Nephelometric Turbidity Units, and rivers generally have an average turbidity of 10 NTU.

Conductance Conductance is the ability of water to carry electricity. The conductivity of water also indicates the presence of other conductive chemicals, such as salt or heavy metals. Things like salinity and temperature have an impact on the conductivity of the water, as both factors increase water's conductivity. Typical drinking water has a range of 200 to 800 µS/cm, which is microSiemens per centimeter. Siemens are the unit of electrical conductivity in the International System of Units.

Ammonia Ammonia, (NH3) is a chemical compound commonly used in industrial cleaners and plant fertilizers. It is extremely soluble in water, and depending on the dosage, considered toxic to many organisms. Accumulation of Ammonia over time in animals is toxic, and can lead to death. Environmental limits from the US government put Ammonia levels between 0.25 and 32.5 parts per million(ppm).

Saturation Saturation refers to the oxygen saturation, or the dissolved oxygen content, of a body of water. It is calculated as a percentage of dissolved Oxygen relative to the maximum saturation at a given temperature and depth. However, it is important to note that the saturation percentage of a body of water can go above 100%, as the air and water will not be in constant equilibrium. Photosynthesis is a common cause for the saturation percentage to go over 100%, as is the case in the dataset.

Data setup

In this section, we are going over various functions and methods for cleaning data in R, which will allow us to better perform analysis later on it. First, we clean up and set up the R environment with the working directory. Keep in mind that your directory structure is likely different from mine, and adjust accordingly.

```
# clean up and setup
rm(list=ls()) # clean up any old stuff in R
setwd("C:/Users/hyper/OneDrive/Desktop/Desktop Folders/Programming/R/Assignments/Week
 6") # go to this folder
```

After this is done, we import the packages we're using. Keep in mind that you will likely need to install these packages first before using them.

```
#library import
library(tidyverse)
library(ggcorrplot)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Finally, we can import our .csv file, and assign it to a variable. I run some basic observations on the dataset as a starting point.

```
#input csv file
water <- read.csv("ellerbee.csv")
view(water)
summary(water)
```

```
##    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
```

Data cleanup

Next, we need to clean the data, by replacing all the missing (NA) values with the means of their column. We do that below.

```
#fix columns CFS, DO, NH4, and saturation
water$CFS <- ifelse(is.na(water$CFS),mean(water$CFS,na.rm=TRUE),water$CFS)
water$DO <- ifelse(is.na(water$DO),mean(water$DO,na.rm=TRUE),water$DO)
water$NH4 <- ifelse(is.na(water$NH4),mean(water$NH4,na.rm=TRUE),water$NH4)
water$saturation <- ifelse(is.na(water$saturation),mean(water$saturation,na.rm=TRUE),wa
ter$saturation)
```

I want to make sure that my data does not have any missing data points, so I use the following debug point to make sure.

```
print(sum(is.na(water))) #debug point
```

```
## [1] 0
```

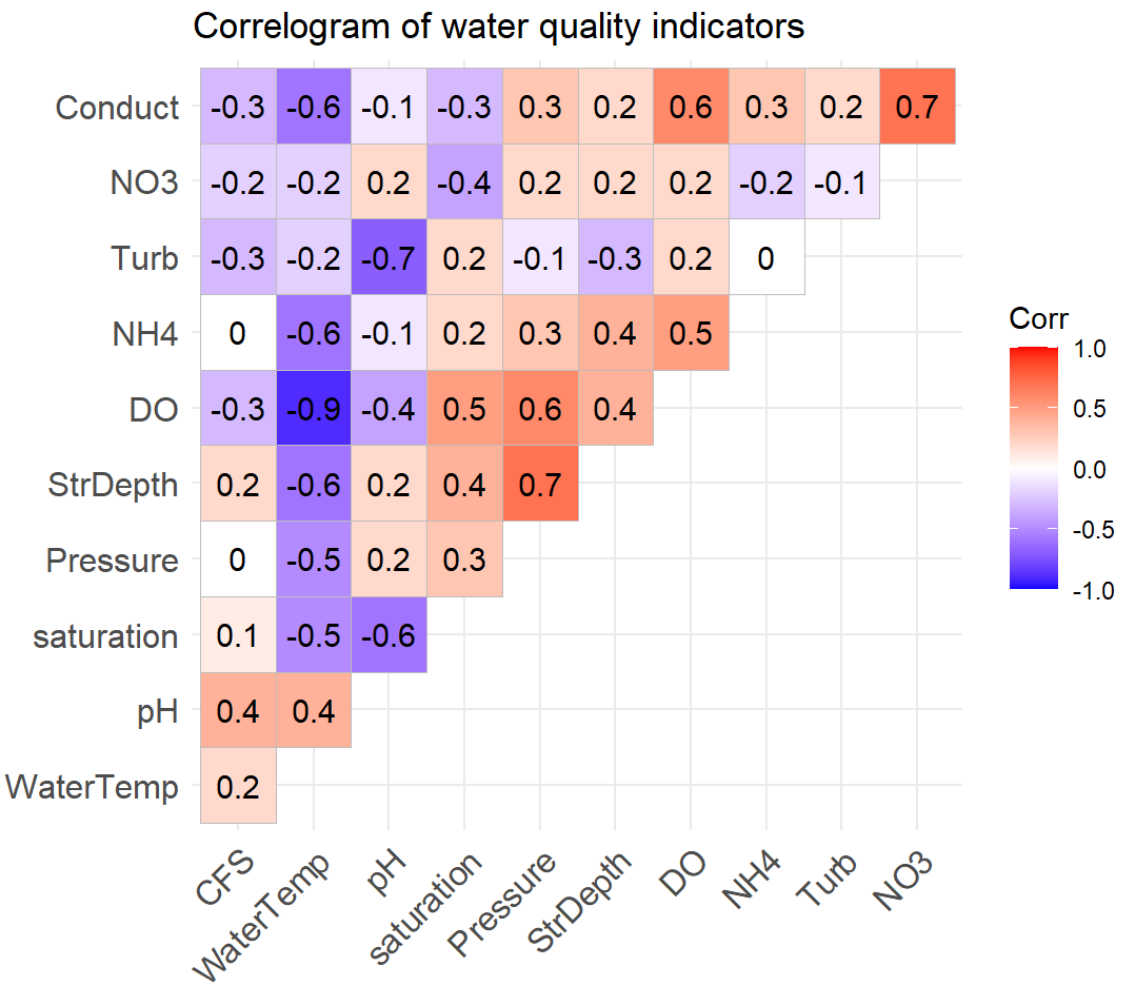
Now that all the cleanup and setup is done, we can start on the graphical analysis of the data using ggplot2 (a graphing package contained in tidyverse). Below, I explain 8 graphs I have used for my data analysis.

Graphical analysis

Correlogram of variables

I wanted to first get a correlogram of my water quality indicators so that I could have a general overview of how variables are related in my dataset. I felt like this was an important first step in analysis so that I could target specific relationships to further analyze, and it would give me a colorful representation of the relationships in the data.

```
# correlogram for all indicating variables
waterCor <- round(cor(water), 1)
corPlot <- ggcorrplot(waterCor,
           hc.order = TRUE,
           type = "upper",
           lab = TRUE,
           title="Correlogram of water quality indicators")
corPlot
```



Based of the results from this graph, we can immediately notice two patterns. There is one streak of dark blue in the water temperature column. Looking closer, this shows that there is a strong inverse relationship between water temperature and other indicators like NH4, dissolved oxygen, and stream depth. Furthermore, we notice a blob of darker red near the center of the plot, suggesting a strong relationship between variables like pressure, stream depth, and dissolved oxygen.

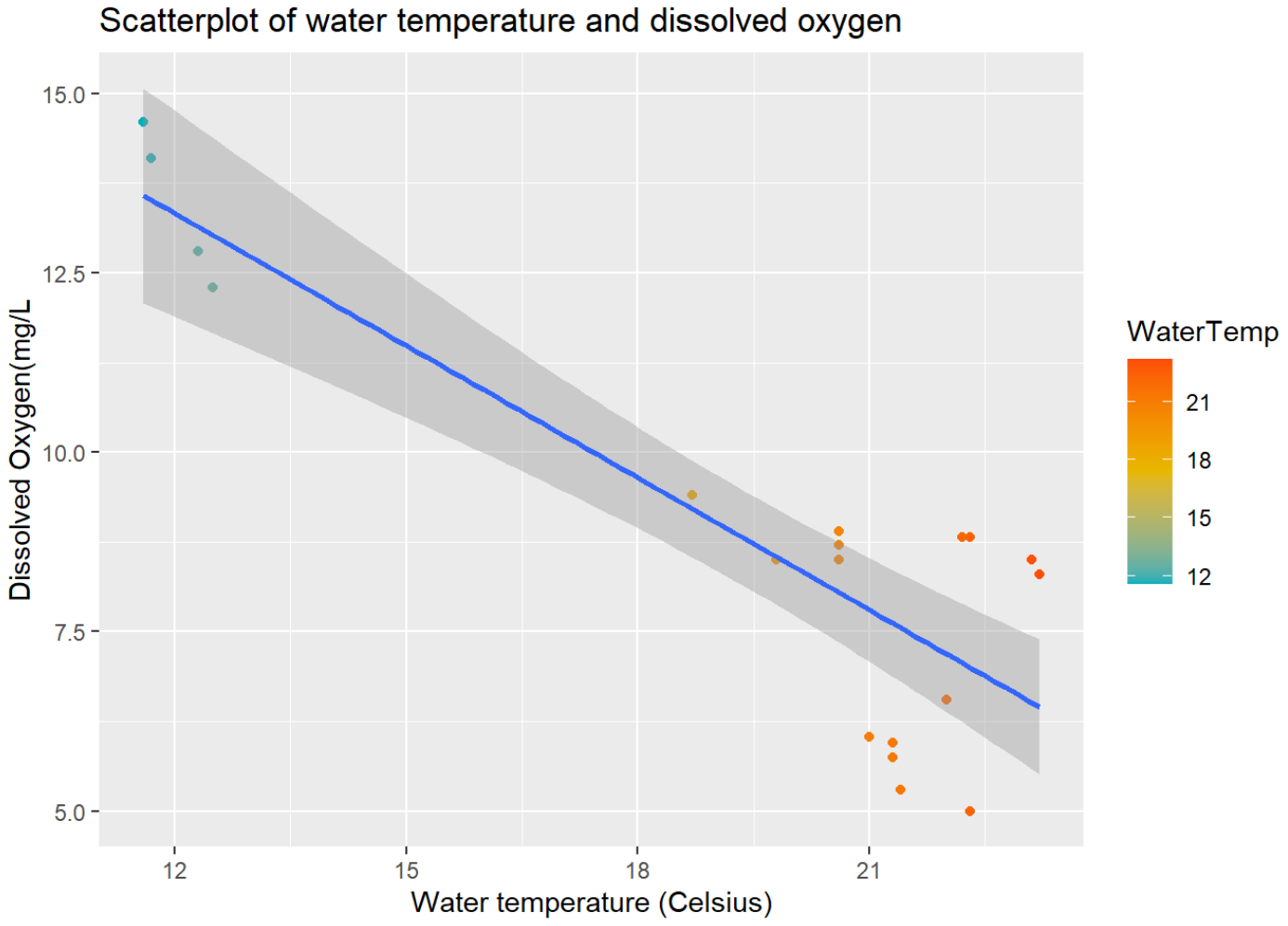
In the next four plots, I dive into what I deem are the most interesting relationships between indicators, which are primarily those with high correlation values.

Scatterplot between dissolved oxygen and water temperature

Firstly, I wanted to observe a graphical representation of the relationsihp between dissolved oxygen and water temperature. It has been proven already that there exists and inverse relationship between water temperature and dissolved oxygen content, and it would appear that this dataset is no exception. On the following graph, I have color-coded the water temperature for aesthetic purposes, and imposed a line of best fit through a linear model.

```
#scatterplot between DO and watertemp
DOWT <- ggplot(water, aes (x = WaterTemp, y = DO)) +
  geom_point(aes(color = WaterTemp)) +
  geom_smooth(method = "lm") +
  labs(x = "Water temperature (Celsius)", y = "Dissolved Oxygen(mg/L", title = "Scatter
plot of water temperature and dissolved oxygen") +
  scale_color_gradientn(colors = c("#00AFBB", "#E7B800", "#FC4E07"))
DOWT
```

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



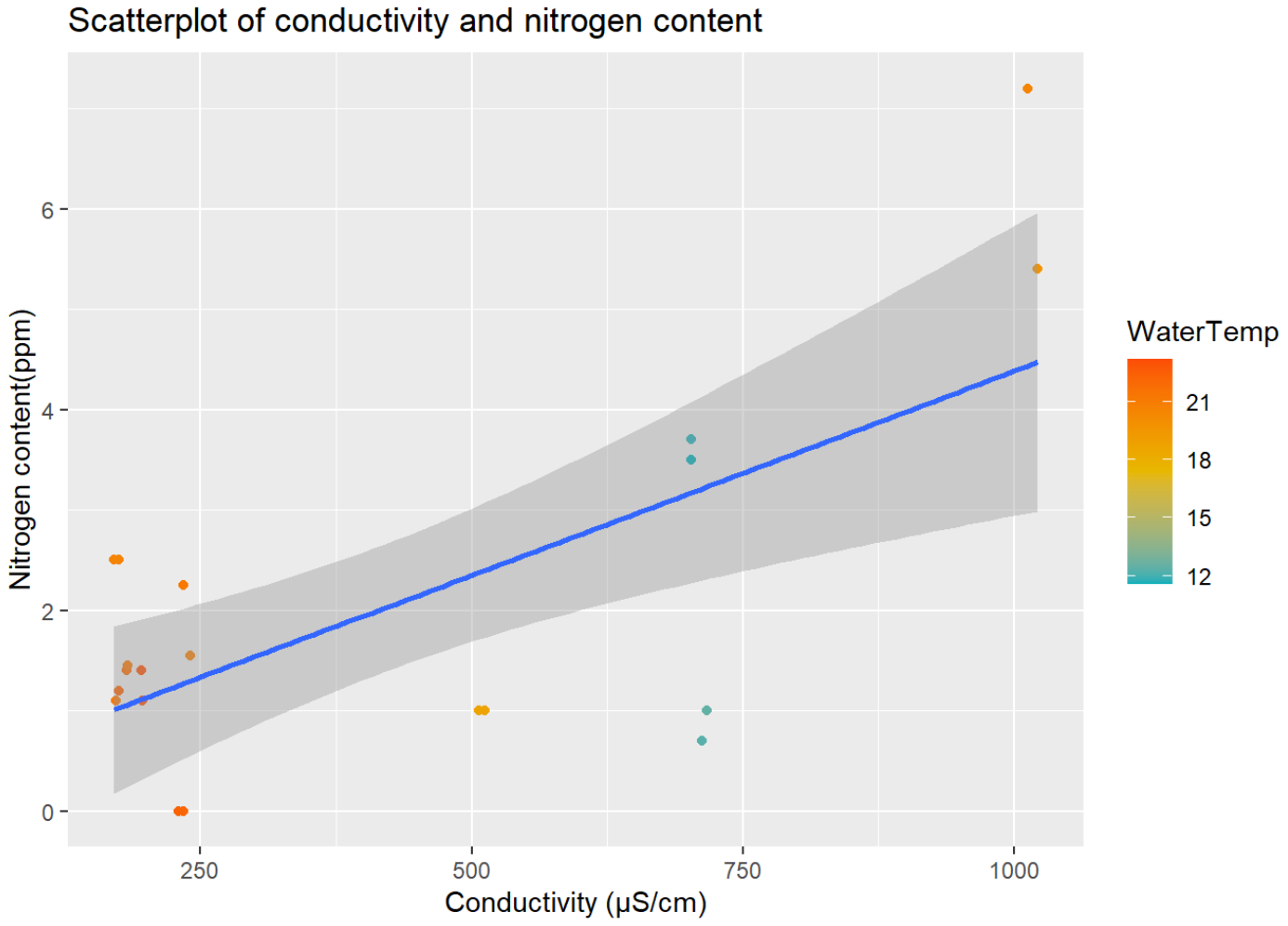
Based off this plot, we can recognize that there indeed exists a strong inverse relationship between dissolved oxygen and water temperature. Furthermore, the width of the confidence interval surrounding the line of best fit shows us that this is a very closely related set of data.

Scatterplot between conductivity and nitrogen content

Next, I noticed that conductivity and nitrogen content shared a high positive correlation, which suggests that nitrogen plays a large role in the ability of a body of water to conduct electricity. I have done nearly everything the same in the following graph as I did for the first one, although I kept the color-coding still as water temperature to see how water temperature would play into the relationship.

```
#scatterplot between NO3 and conductivity
NO3C <- ggplot(water, aes (x = Conduct, y = NO3)) +
  geom_point(aes(color = WaterTemp)) +
  geom_smooth(method = "lm") +
  labs(x = "Conductivity (µS/cm)", y = "Nitrogen content(ppm)", title = "Scatterplot of
conductivity and nitrogen content") +
  scale_color_gradientn(colors = c("#00AFBB", "#E7B800", "#FC4E07"))
NO3C
```

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



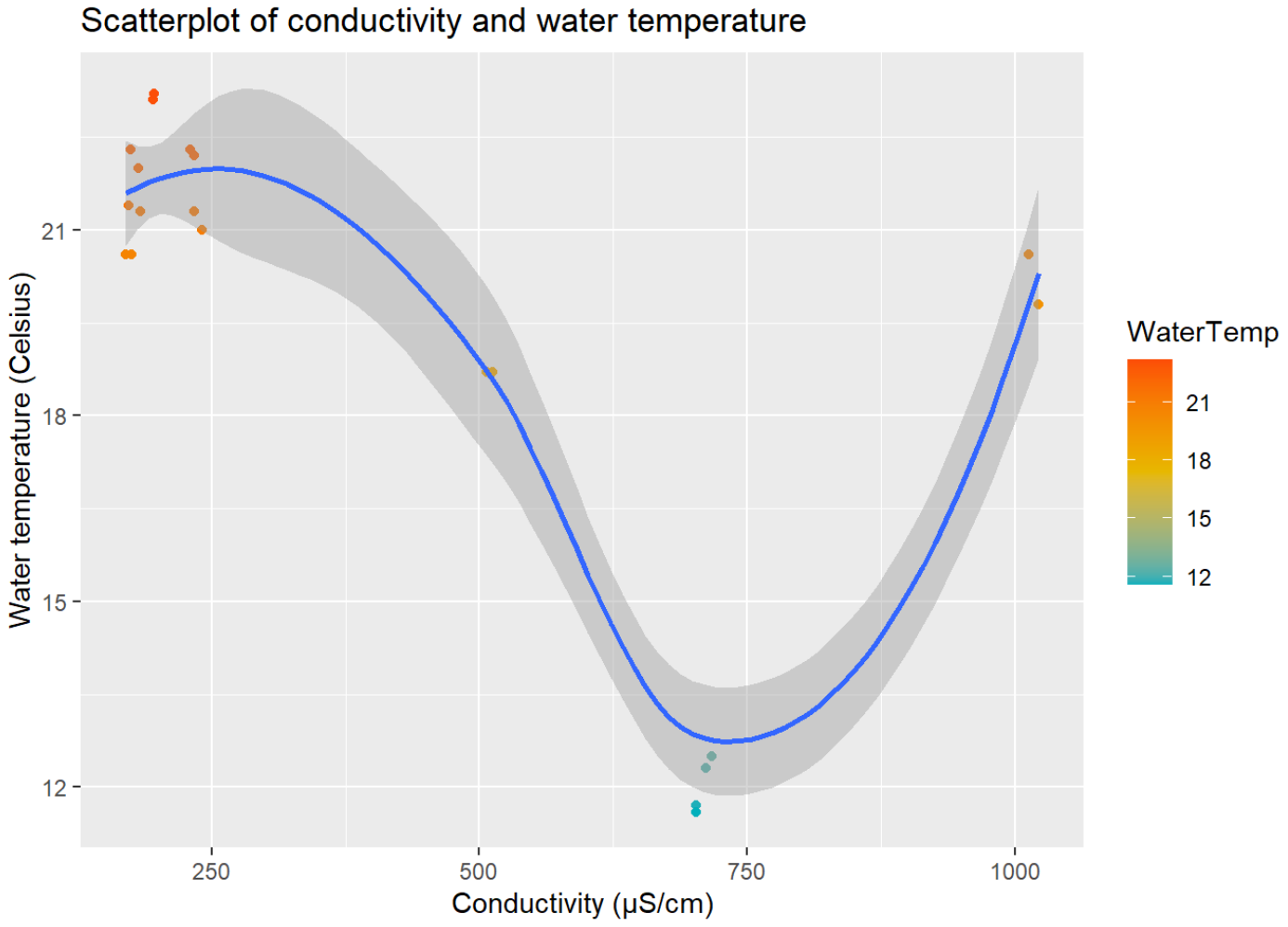
Based off this plot, we can obviously notice the correlation between conductivity and nitrogen content. This shows me that an increased amount of nitrogen in a body of water results in higher conductivity for that body of water, meaning nitrogen is a good conductor. We can see that the relationship between conductivity and nitrogen content is weaker than that of water temperature and dissolved oxygen content, as the confidence interval is larger. Finally, by color-coding by water, we can also notice that there is a column of cold data, around when the conductivity is 750 µS/cm. This odd coloration may be an outlier or a fluke, but it is worth looking into, which we do in our next plot.

Scatterplot of conductivity and water temperature

From our findings in the previous plot, I decided that I wanted to evaluaate the relationship between water conductivity and water temperature. Looking back at the correlogram, we can see that they have a correlation of -0.6, suggesting an above average inverse relationship. I decieded to switch from linear modeling to a loess model, because I believe that the relationship is more parabolic than linear.

```
#scatterplot between conductivity and water temperature
WTC <- ggplot(water, aes (x = Conduct, y = WaterTemp)) +
  geom_point(aes(color = WaterTemp)) +
  geom_smooth(method = "loess") +
  labs(x = "Conductivity (µS/cm)", y = "Water temperature (Celsius)", title = "Scatterp
lot of conductivity and water temperature") +
  scale_color_gradientn(colors = c("#00AFBB", "#E7B800", "#FC4E07"))
WTC
```

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



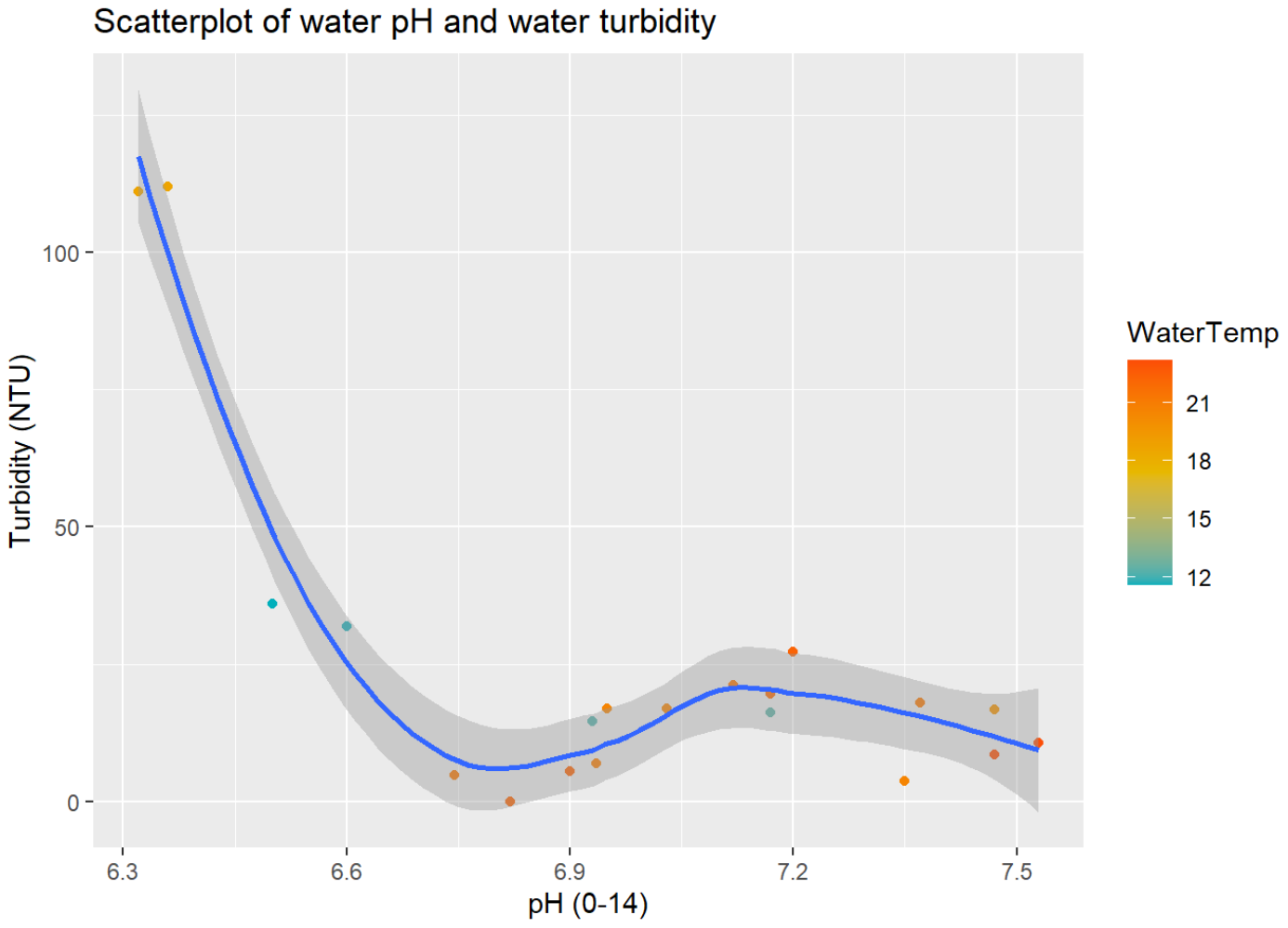
It seems that the water temperature's relationship with conductivity is somewhat parabolic, as the water temperature dips around 750 µS/cm, before going back up. However, it is hard to tell if this is intentional or a fluke of the data measurement, as we don't have nearly enough data points to reach a reasonable conclusion. Based off what we have, I would say that cold water will generally have a conductivity around 750 µS/cm.

Scatterplot of pH and turbidity

Next, I wanted to look at a grpahical relationship between the water pH and its turbidity, as I saw it had a strong inverse correlation (-0.7). In the following graph, I still use the loess model, as using a linear model wasn't accurate enough. Furthermore, I still kept the color-coding for water temperature, to see if there were any interesting occurrences to pick up.

```
#scatterplot between pH and turbidity
PHT <- ggplot(water, aes (x = pH, y = Turb)) +
  geom_point(aes(color = WaterTemp)) +
  geom_smooth(method = "loess") +
  labs(x = "pH (0-14)", y = "Turbidity (NTU)", title = "Scatterplot of water pH and wat
er turbidity") +
  scale_color_gradientn(colors = c("#00AFBB", "#E7B800", "#FC4E07"))
PHT
```

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



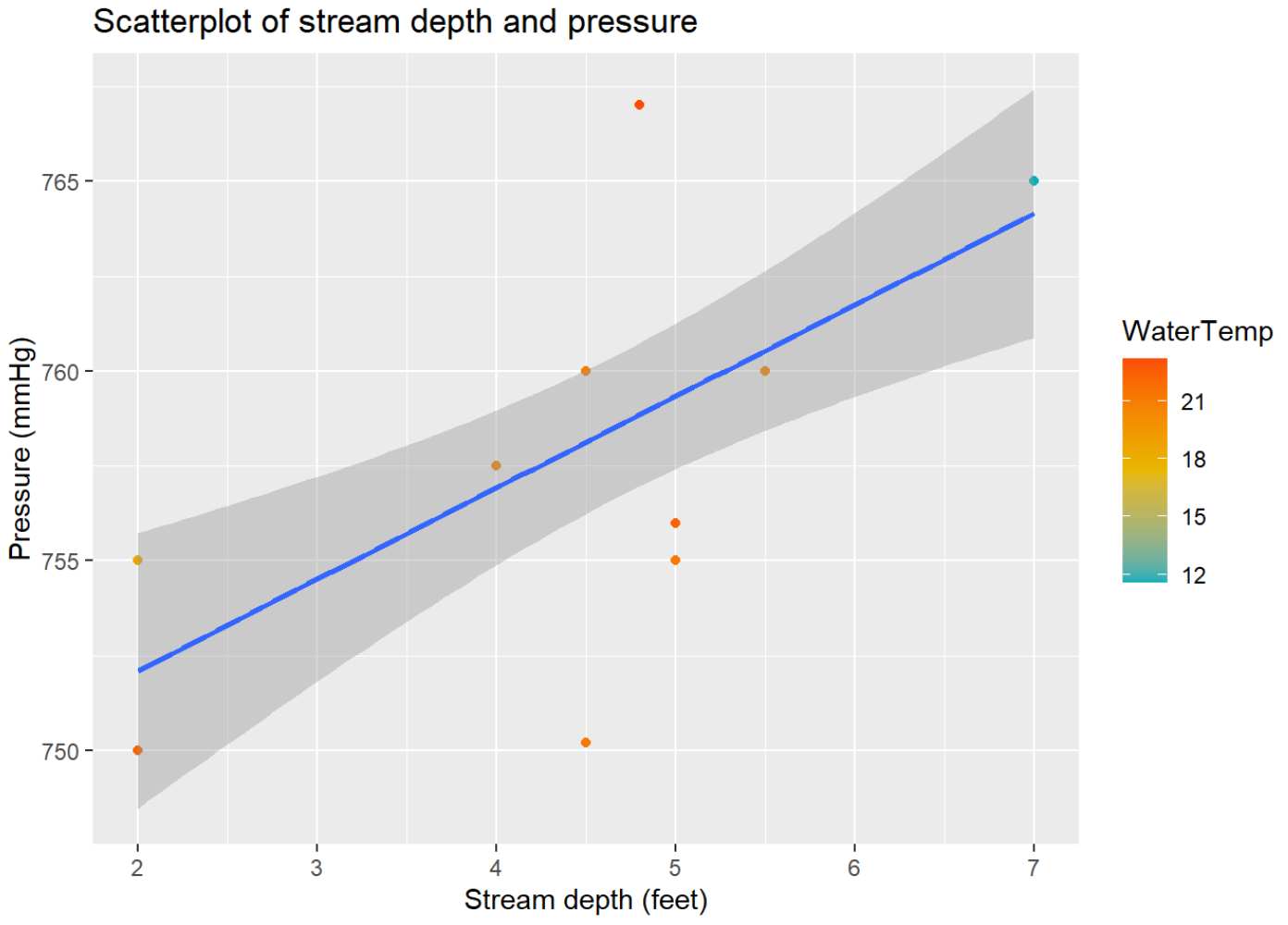
I found this scatterplot particularly interesting because of its closer relationship in low pH values but then variation at higher pH values. It seems like the turbidity increased as the pH decreased, but then once past 6.7 pH there was more fluctuation and uncertainty. I can say with a fair degree of confidence that

turbidity will go up as pH goes down, but I don't want to make any generalizations about the relationship between the pH and turbidity around 7 pH because of the fluctuation.

**Scatterplot of stream depth and pressure In this scatterplot, I decided to look at the relationship between stream depth and pressure (atmospheric pressure), which has a positive correlation of 0.7.

```
#scatterplot between pressure and strdepth
PSD <- ggplot(water, aes (x = StrDepth, y = Pressure)) +
  geom_point(aes(color = WaterTemp)) +
  geom_smooth(method = "lm") +
  labs(x = "Stream depth (feet)", y = "Pressure (mmHg)", title = "Scatterplot of stream depth and pressure") +
  scale_color_gradientn(colors = c("#00AFBB", "#E7B800", "#FC4E07"))
PSD
```

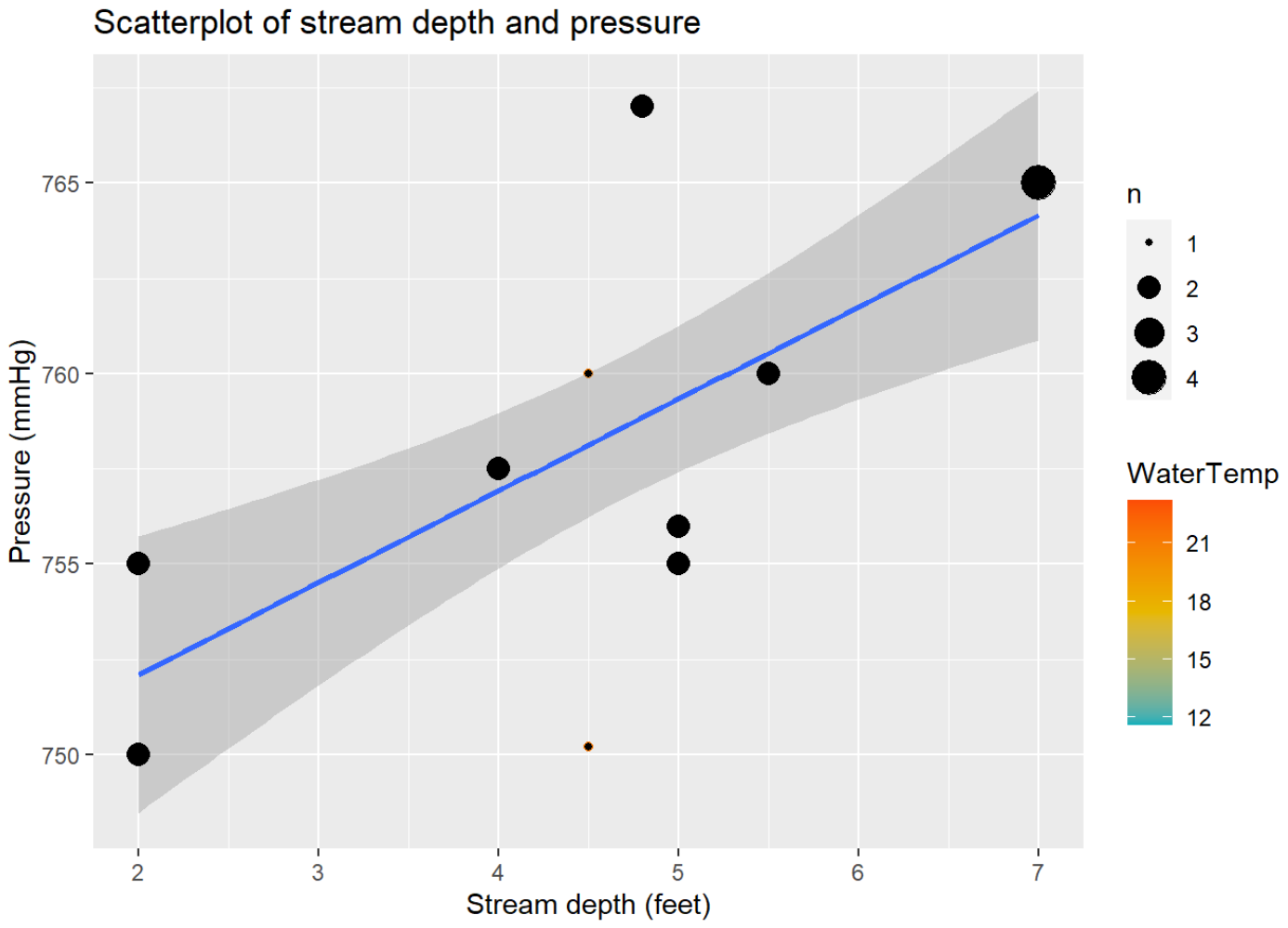
```
## `geom_smooth()` using formula 'y ~ x'
```



My initial thoughts where that this graph was not very well correlated, as it seemed there wasn't discernible pattern to the data points. However, it occurred to me that there were probably point stack on top of each other, so I added the following to the plot.

```
# add size reference
PSD <- PSD + geom_count()
PSD
```

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

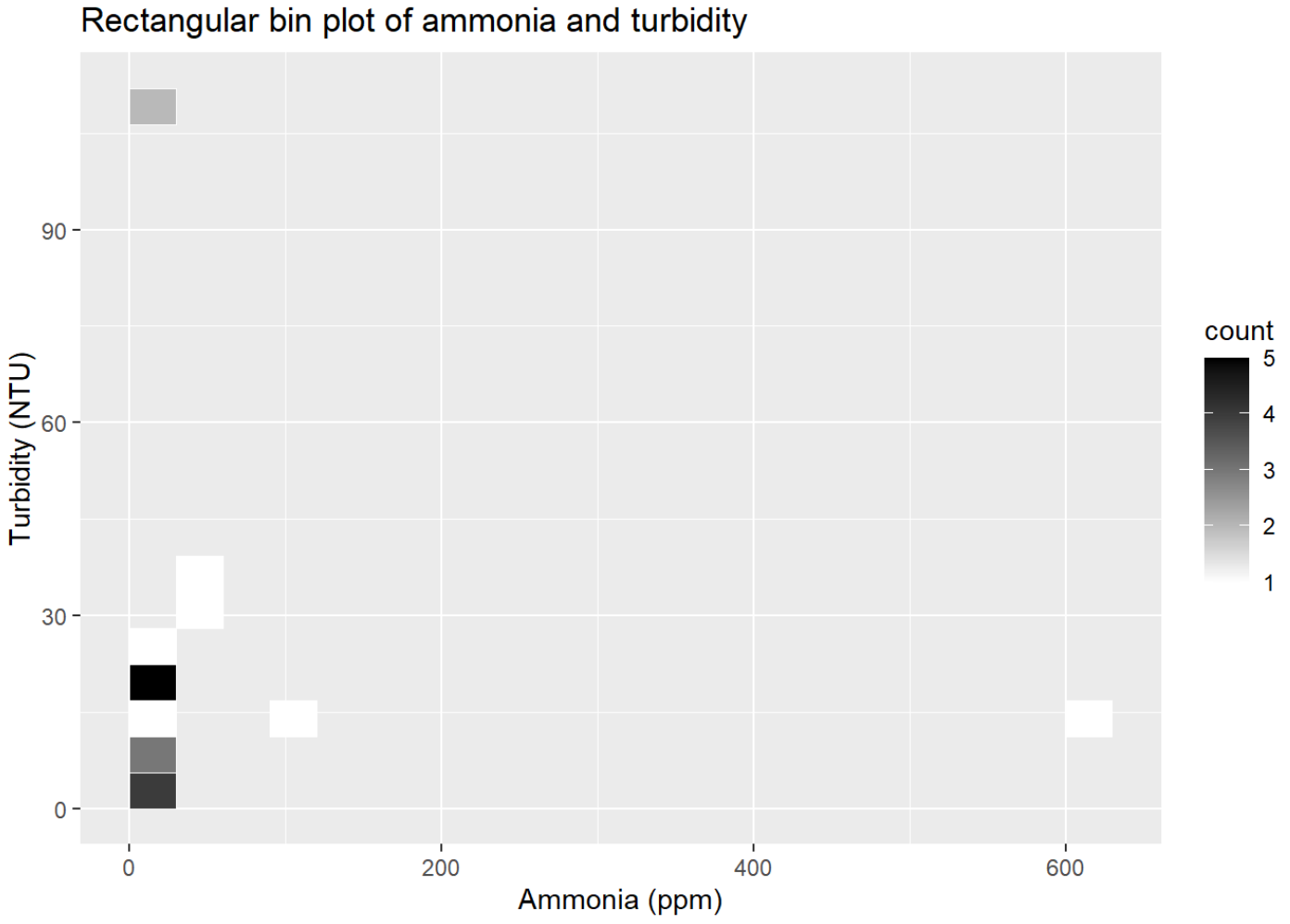


After accounting for the number of occurences, it makes a lot more sense why the relationship has a high correlation value. Although the data points have a lot of spread,the points on the line have a much higher occurrence than the points off the line, which would've soldified the linear model. It seems that there is a mostly linear relationship between stream depth and pressure, although there is some descrepancy around the 4-5 feet range of stream depth, where there are pressure fluctuations.

Rectangular bin plot of ammonia by turbidity

For this next plot I wanted to look at a pair of indicators with a very low correlation value, which one would expect to be higher. I decided to look at the relationship between ammonia and turbidity, as I had expected them to be at least somewhat related, but to my surprise they had a correlation of 0. I thought they would be related by ammonia is used in plant growth, and therefore a highly amount of ammonia would led to higher numbers of plants. This in turn would lead to increased turbidity, as there are more plants in the water. I decided to color-code the graph in black and white for better contrast value, as the default color scheme was hard to discern.

```
# heat map of NH4  by turbidity
NH4T <- ggplot(water, aes(NH4, Turb)) +
  geom_bin2d(bins = 20, color ="white") +
  scale_fill_gradient(low =  "#ffffff", high = "#000000") +
  labs(x = "Ammonia (ppm)", y = "Turbidity (NTU)", title = "Rectangular bin plot of amm
onia and turbidity")
NH4T
```

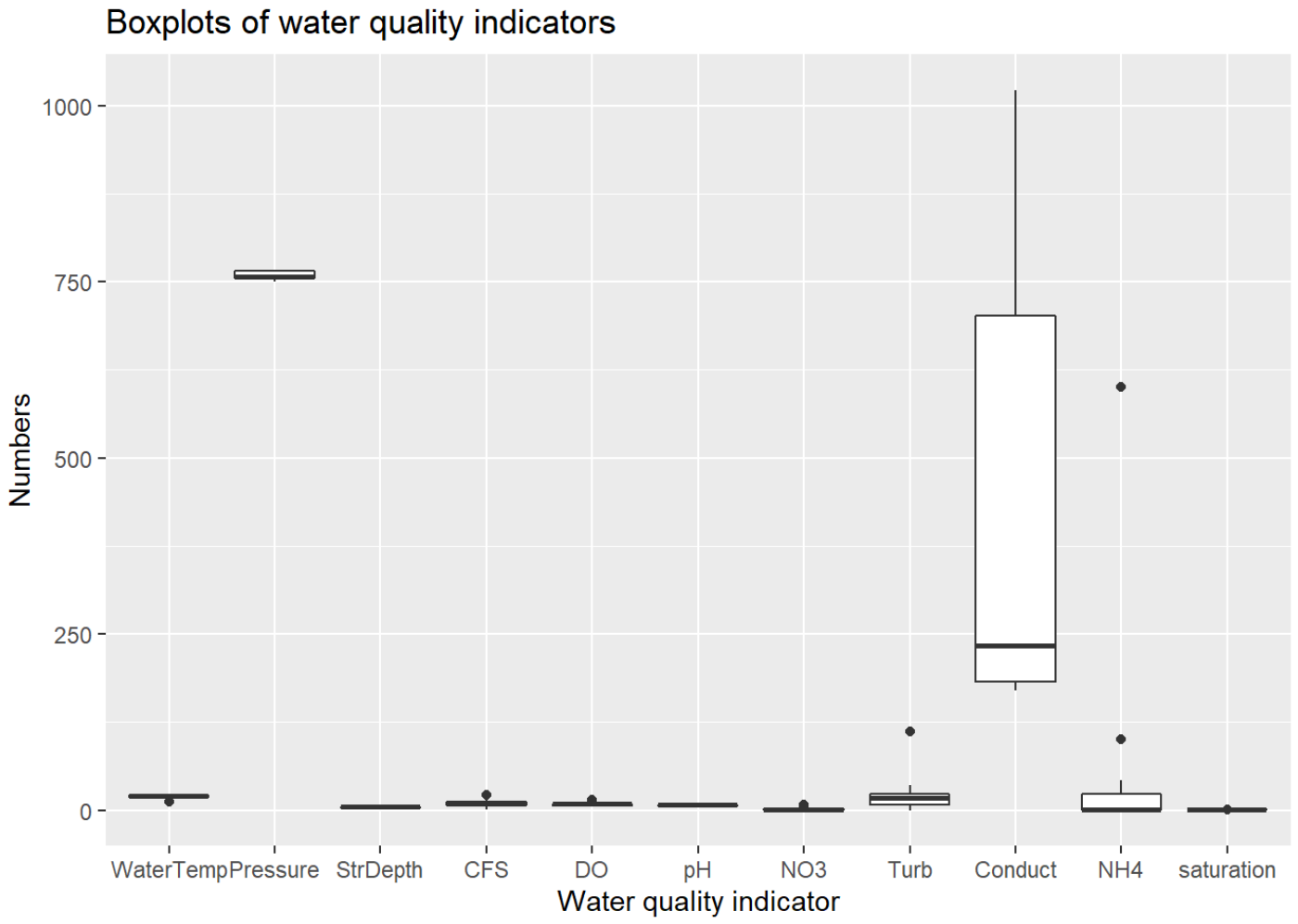


Looking at this graph, the first thing I noticed was the concentration of darker colors in the bottom left corner. This shows me that there is a high frequency of these numbers, which should've led to a stronger correlation. However, there are two extreme outliers on both axis of the graph, which would've ruined any correlation the values had before. While I still believe that there is some relationship between ammonia content and turbidity, this dataset does not reflect that, even though the existance of extreme outliers may have skewed the correlation calculations.

Array of boxplots

For my final plot, I wanted to get another overview of my indicator columns, so I decided to make an array of boxplots. The purpose of this plot was essentially to give an overview of all the data that we looked at, so we could make decisions regarding its variance.

```
boxplots <- ggplot(stack(water), aes(x = ind, y = values)) +
  geom_boxplot() +
  labs(x = "Water quality indicator", y = "Numbers", title = "Boxplots of water quality
indicators")
boxplots
```



The biggest takeaway I got from this graph was the large variation present in the conductivity of water. The other water quality indicators were very localized and compact, not representing much variation.

Concluding statements

From this graphical analysis of water quality data at Ellerbee Creek, I have two primary takeaways:

1. The water temperature and dissolved oxygen are the two best predictors for other water quality indicators. Their correlation values with other indicators were consistently high.
2. Nitrogen content and cubic flow per second are the two worst predictors for other water quality indicators. Their correlation values with other indicators were consistently low.

Other analysis about specific relationships between variables can be found under the plots which discuss them.