Analysis of Water Quality in Ellerbe Creek

Samuel Dummer

9/28/2021

Introduction

Water pollution is a problem in many cities since there are many industries that release toxic chemicals into water that causes many chemicals to go in the water. There is also the problem of runoff which is one of the biggest causes of nitrification in water bodies. In this paper, we will specifically be analyzing the water quality of Ellerbe Creek which runs through Durham.

In this data set, there are 11 major variables that were used to study the creek. Firstly, they took the temperature of the water. This is important because the temperature of the water can affect many chemicals especially the quantity of dissolved oxygen. This data generally ranged from about 11 Celsius to 23 Celsius.

Next, there was pressure which also has a correlation between the levels of chemicals in the creek. This data ranged from 750 to 767 mmHG.

Moving on, the next variable taken into account is stream depth. This is important in understanding the velocity of the flow and any weather or other factors that may have caused this. This variable ranged from 2 to 7 feet.

CFS a.k.a. cubic flow/sec describes the velocity of the flow and the amount of water being moved. This usually has a lot of correlation. This data ranged from about .59 to 21.9 cubic flow/sec.

DO is very important in terms of water quality. DO is important for most aquatic wildlife since it is what allows fish to breath underwater. This variable ranged from 5 to 14.6 mg/L.

The next variable being discussed is pH. pH is also very important in terms of water quality. Many animals cannot live in environments that have a different pHs. The pH rangeds from 6.3 tp 7.6

Next, NO3 was tested which is important to figure out whether or not the body of water has a capability of hosting an algal bloom which would kill many animals in the body of water. This data ranged from 0 to 7.2.

Additionally, turbulence was taken into account. A higher turbulence may mean that the water stream is moving faster or something is disturbing it. This value ranges from 0 to 112 NTU,

Conductance was also measured in the creek which can help to identify how many solutes are in the water, since distilled water is not conductive the solutes in water cause it to be conductive. These values range from 170 to 1022.

The levels of ammonia were also tested. Ammonia is also an important chemical in water since ammonia is toxic to fish, and in high levels it can kill the fish. Ammonia also has the possibility to change the pH of the water. This variable ranged from .2 to 600.

Lastly, the saturation of the water which most likely describes how saturated the chemicals were in the water. This value ranged from .93 to 1.3.

Setting Up Directories, Cleaning Environment, and More

Before we can create and graphs, we need to set up the working directory and clean the environment. This helps to clear up the the environment from the previous time the environment was used. Next we also need to load in any libraries that we believe will be used in our code. In this specific script, I loaded in the "tidyverse", "ggplot2", "gridExtra" and "stringr" libraries. These libraries help to give us many different/more efficient functions to use. Lastly, we also loaded in a user function in case we needed to use some functions from there.

```
rm(list=ls())
setwd("C:/Users/isabe/Desktop/RFLoder")
source("myfunctions.R")
library(tidyverse)
library(stringr)
library(ggplot2)
library(gridExtra)
```

Reading in Data

Once everything is set up, we are now able to load in the data set that we will be using. For this we used the read_csv file to read in our file names "ellerbee.csv".

```
pollution <- read.csv("ellerbee.csv")
```

Quickly Cleaning the Data

After loading in the data, we realized that there were a lot of missing values. To fix this problem we used an ifelse statement that would check if the value was an NA values then replace it with the mean of the variable if it was missing. We did this for the CFS, DO, saturation, and NH4 variables.

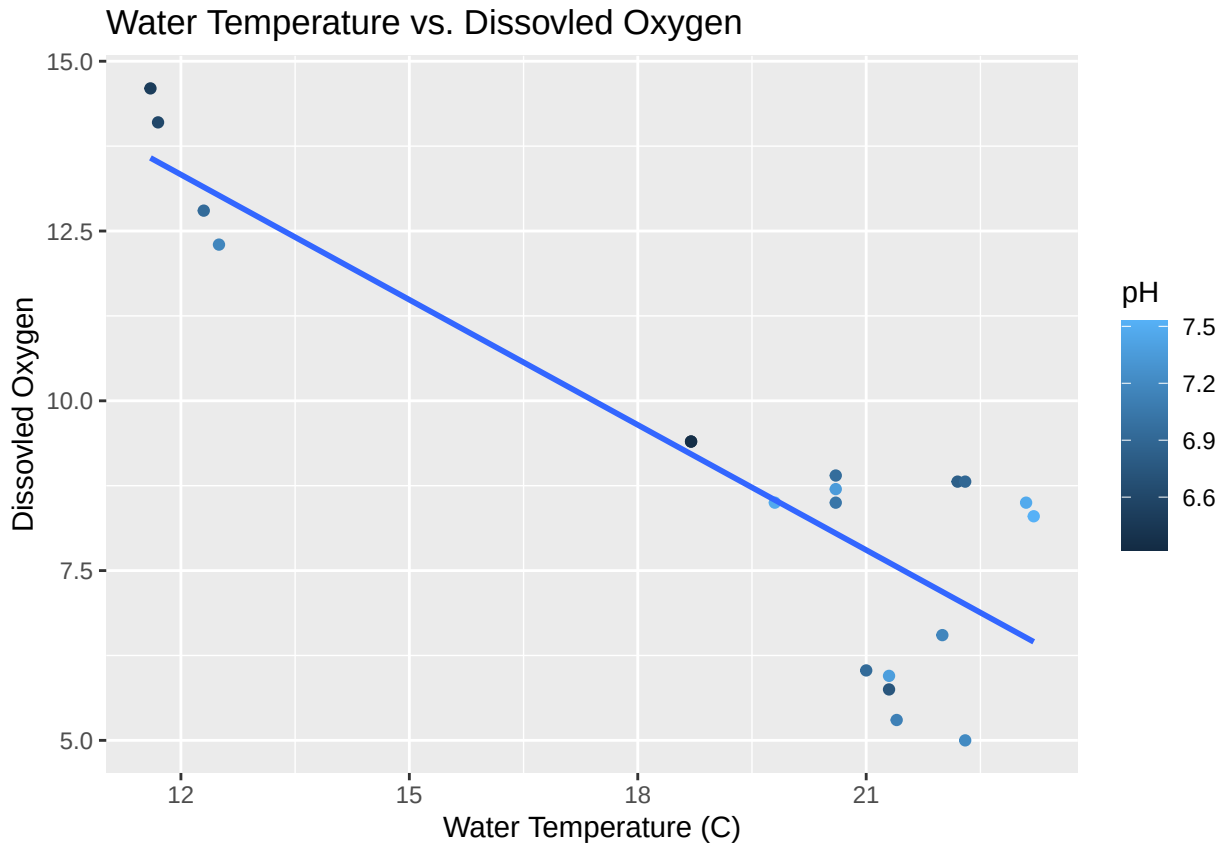
```
pollution$CFS <- ifelse(is.na(pollution$CFS), (mean(pollution$CFS, na.rm=TRUE)), pollution$CFS)
pollution$DO <- ifelse(is.na(pollution$DO), (mean(pollution$DO, na.rm=TRUE)), pollution$DO)
pollution$saturation <- ifelse(is.na(pollution$saturation), (mean(pollution$saturation, na.rm=TRUE)), p
pollution$NH4 <- ifelse(is.na(pollution$NH4), (mean(pollution$NH4, na.rm=TRUE)), pollution$NH4)
head(pollution)
```

```
##   WaterTemp Pressure StrDepth      CFS   DO   pH NO3 Turb Conduct       NH4
## 1      12.5    765.0        7 9.557857 12.3 7.17 1.0 16.3     717 100.00000
## 2      12.3    765.0        7 9.557857 12.8 6.93 0.7 14.7     712 600.00000
## 3      11.7    765.0        7 9.557857 14.1 6.60 3.7 32.0     702  42.23111
## 4      11.6    765.0        7 9.557857 14.6 6.50 3.5 36.0     702  42.23111
## 5      19.8    757.5        4 7.460000  8.5 7.47 5.4 16.7    1022   0.50000
## 6      20.6    757.5        4 7.230000  8.7 7.35 7.2  3.8    1013   0.40000
##   saturation
## 1       1.10
## 2       1.20
## 3       1.30
## 4       1.30
## 5       0.93
## 6       0.97
```

Plotting Eight Separate Graphs

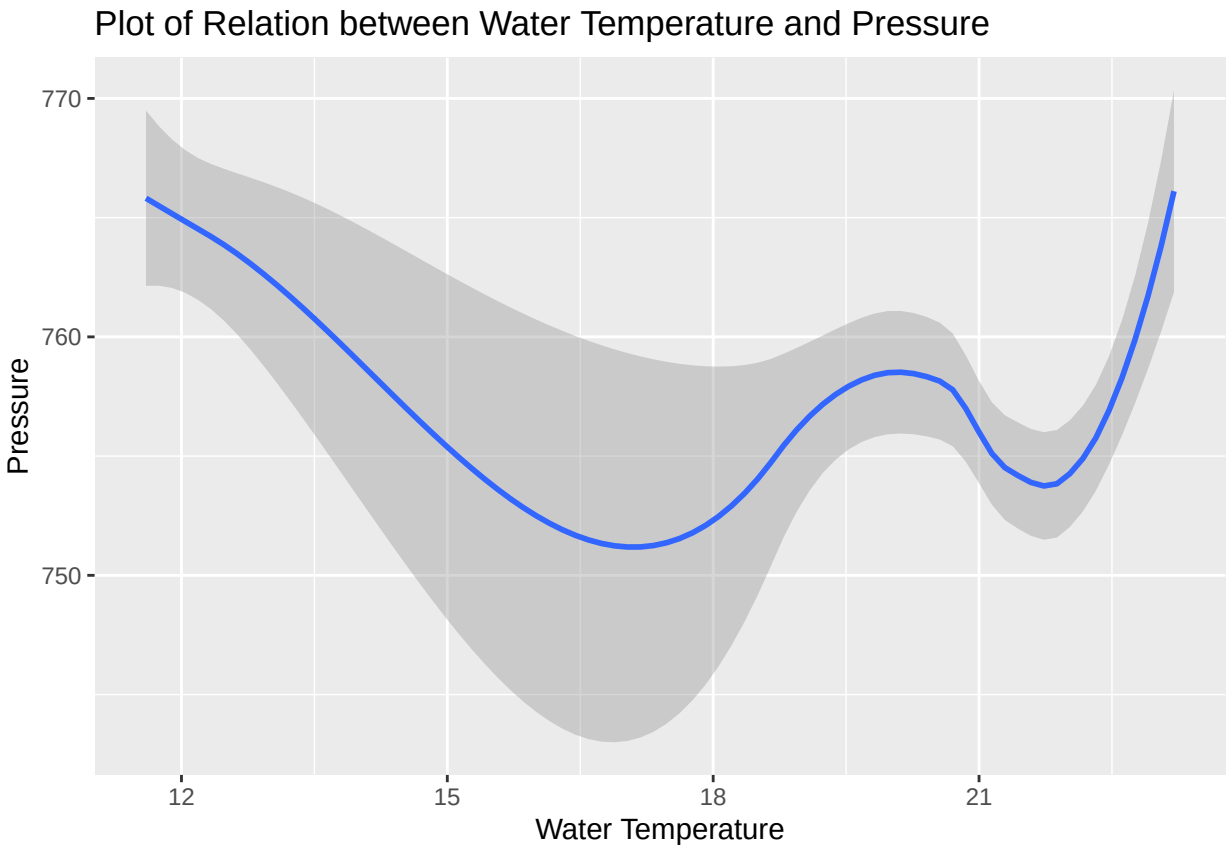
1. Scatter plot of Water Temperature vs. Dissolved Oxygen This scatter plot helps to show the relation between Dissolved Oxygen(DO), Water Temperature, and pH. The results of this graph help to solidify the fact that the higher temperature water is the lower levels of dissolved oxygen. In this scatter plot, we can can observe the negative correlation between DO and Water Temperature, additionally, pH seems to be higher the higher temperature the water is, but this isn't completely proven. There are also some outliers, but they align with the linear regression which strengthens the correlation.

```
first <- ggplot(pollution, aes(WaterTemp, DO, color = pH)) +
  geom_point() +
  geom_smooth(method = lm, se = F) +
  labs(title = "Water Temperature vs. Dissovled Oxygen", x = "Water Temperature (C)", y = "Dissovled Oxy
first
```



2. Plot of Relation between Water Temperature and Pressure Next we wanted to know if there was any correlation between Water Temperature and Pressure. We decided to graph a linear regression using the "loess" method. We also decided to keep the standard error. When looking at the graph, it seems as if there is little correlation between pressure and water temperature, but it does look like at a certain temperature the pressure is at its lowest. This might have some other correlation other than water temperature, though.

```
second <-  ggplot(pollution, aes(WaterTemp, Pressure)) +
  geom_smooth(method = "loess") +
  labs(title = "Plot of Relation between Water Temperature and Pressure", x = "Water Temperature")
second
```



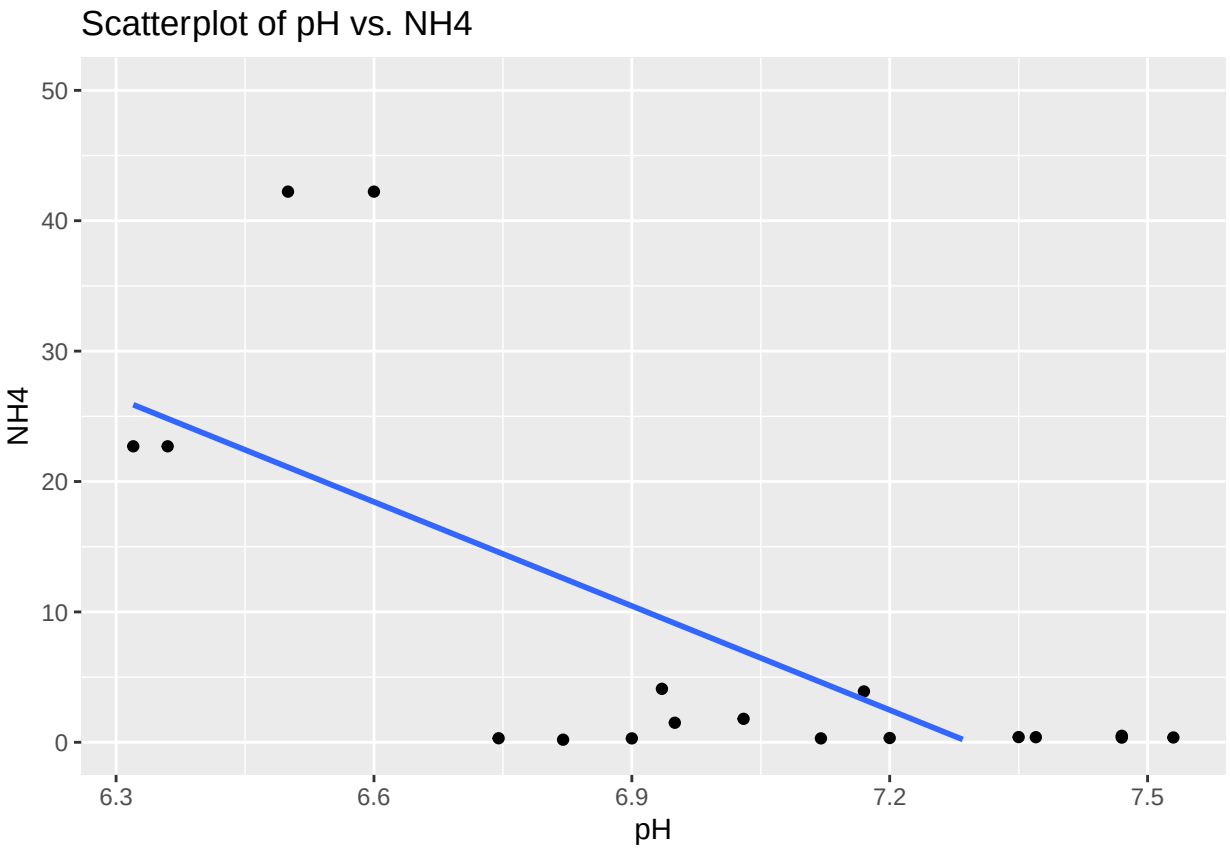
3. Scatter Plot of pH vs NH4 Since ammonia is very basic, we decided to see if there was a correlation of the pH of the water and the level of NH4. To do so we graphed a scatter plot with a simple linear regression through it. We also made the y limit 0 to 50 since there seemed to be some outliers. The graph overall showed is that there was some negative correlation but nothing significant. This negative correlation makes some sense becuase that means the higher pH the less NH4 and the lower pH the more which is what we believed would happen.

```
third <- ggplot(pollution, aes(pH, NH4)) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Scatterplot of pH vs. NH4") +
  ylim(0, 50)
third
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

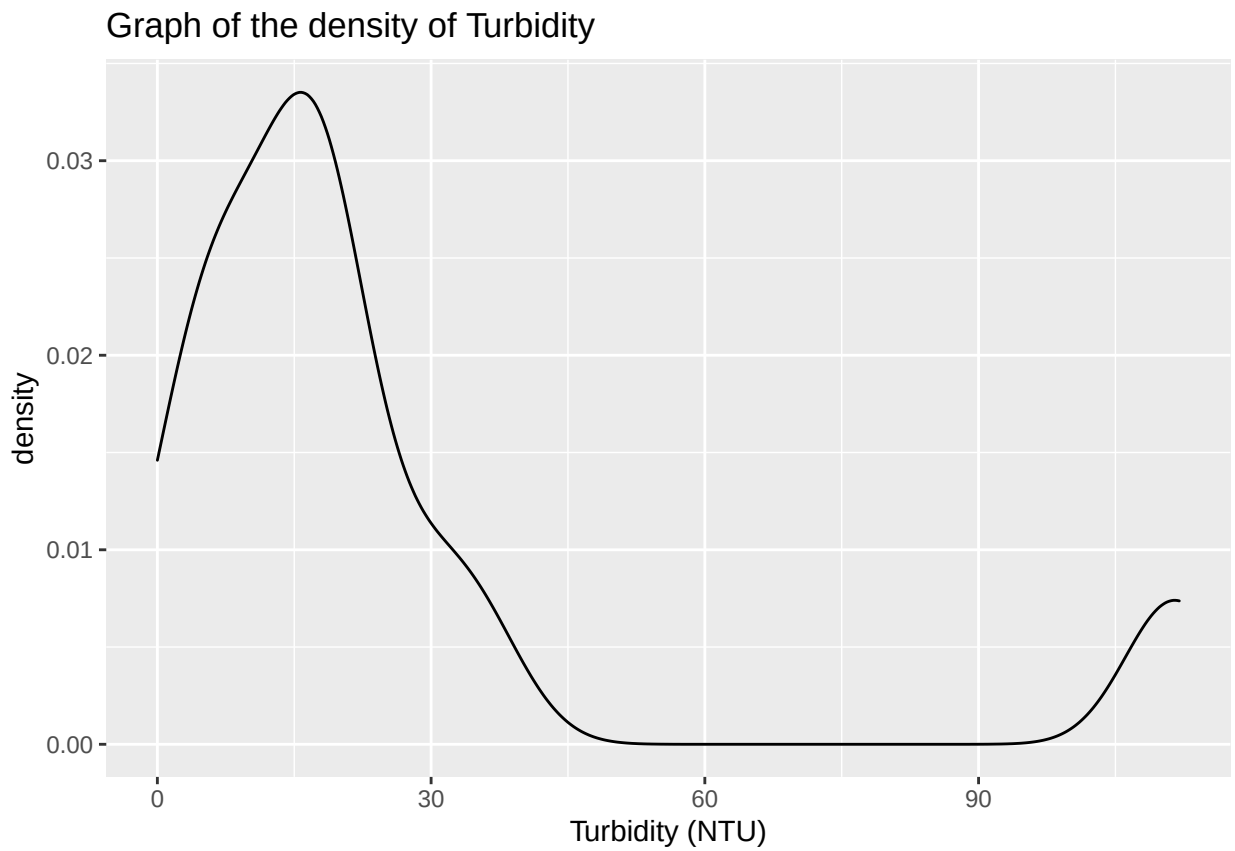
```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## Warning: Removed 16 rows containing missing values (geom_smooth).
```



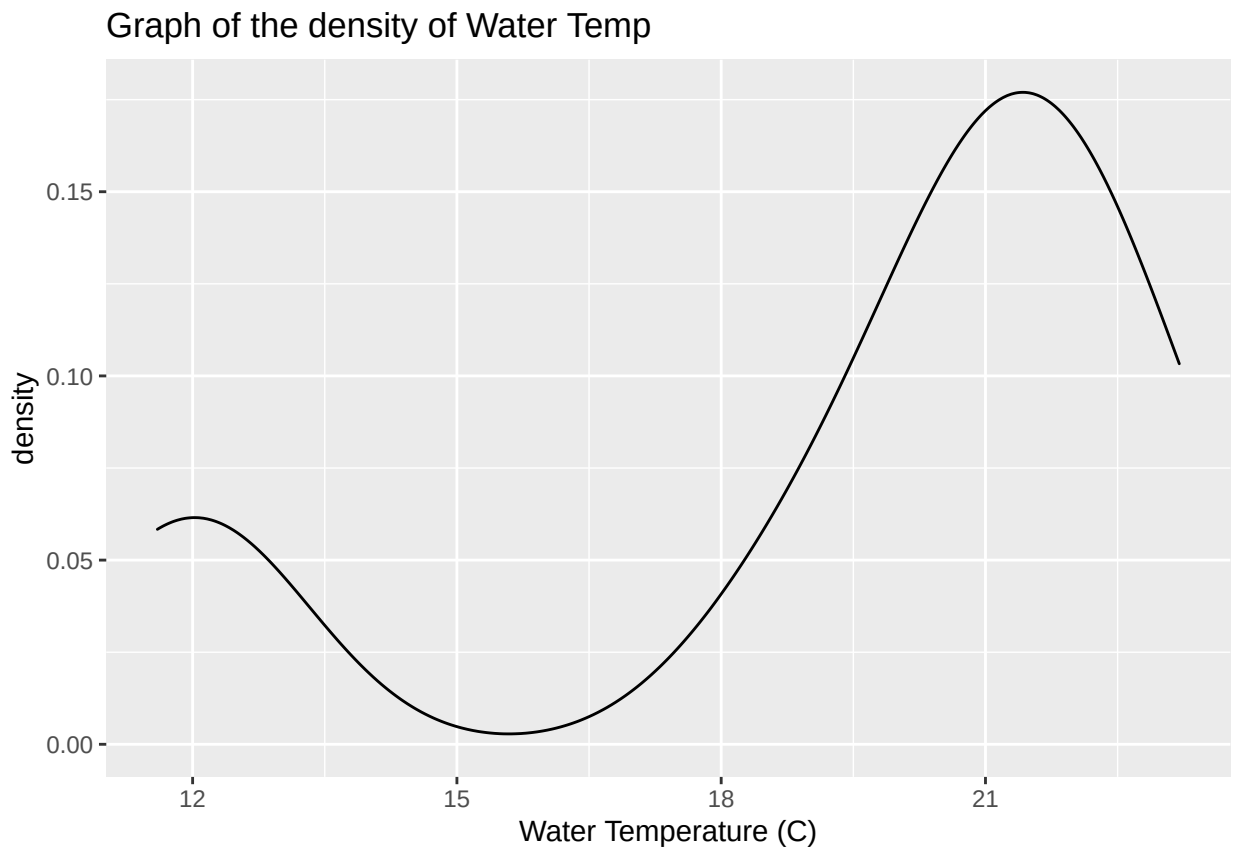
4. Graph of the Density of Turbidity Next we wanted to see the general distribution of Turbidity. We used a density plot to do so because overall this would show a better distribution than a histogram since we don't have many observations. While graphing the density plot, we noticed that the distribution was fairly normal, except for a random outlier at 100 NTU which was intersting. This may be a false reading or an even that had happened that caused the turbidity to spike.

```
fourth <- ggplot(pollution, aes(x = Turb)) +
  geom_density(kernel = "gaussian") +
  labs(title = "Graph of the density of Turbidity", x = "Turbidity (NTU)")
fourth
```



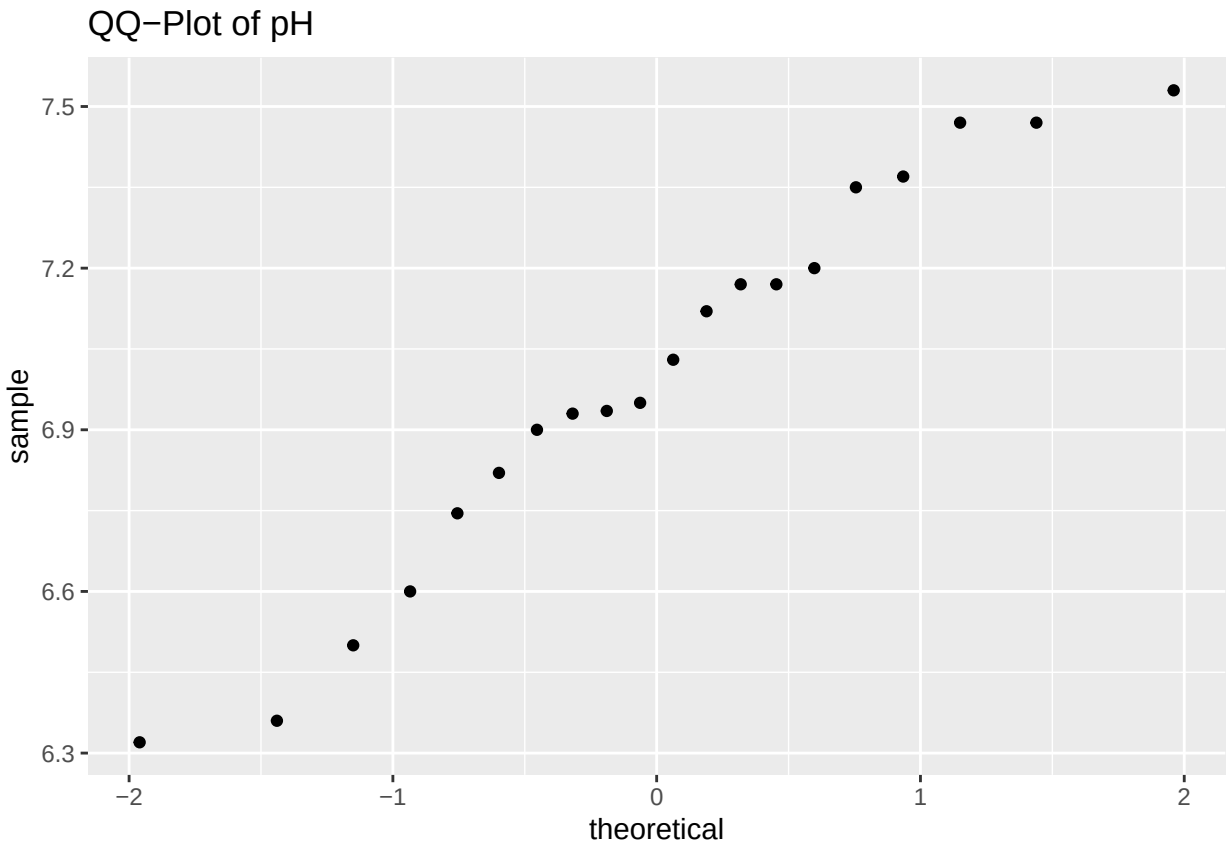
5. Graph of the density of Water Temp Next, since I wanted to see is this outlier was also consistent with other data, I graphed a density plot of Water Temperature. Again, this is a better way to show the distribution of the data since there are so little values. Similar to the last plot, there seemed to be an outlier at 12 which could correlate with the outlier in turbidity. The rest of the data seems to follow a normal distribution as expected.

```
fifth <- ggplot(pollution, aes(x = WaterTemp)) +
  geom_density(kernel = "gaussian") +
  labs(title = "Graph of the density of Water Temp", x = "Water Temperature (C)")
fifth
```



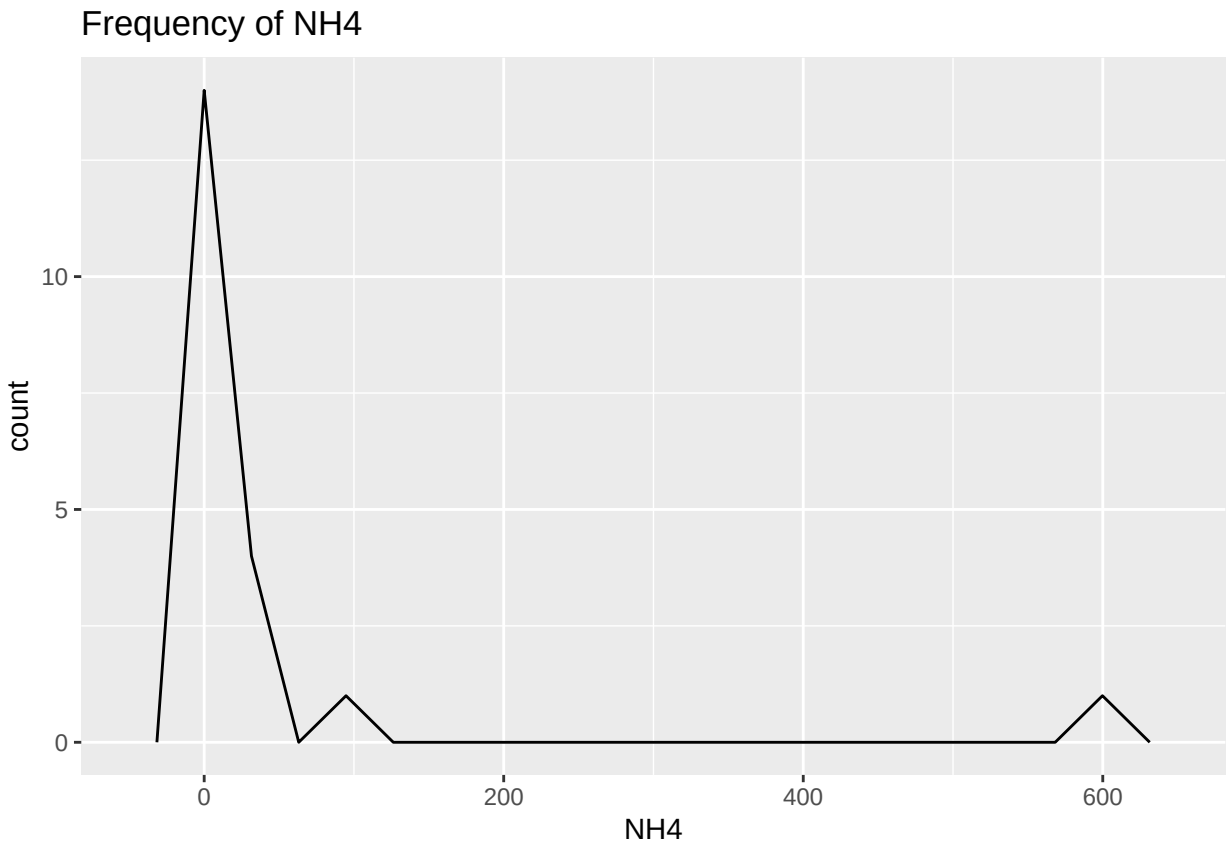
6. QQ-Plot of pH Another way that we can check the distribution of a variable is through QQ-Plots. We decided that it was important to check the distribution of pH. The qq-plot of pH seems to be generally linear which is a good sign of normal distribution. There is some varience meaning it isn't perfect, but overall the data of pH seems to be normally distributed. This means that there is a chance that the outliers viewed in the passed two plots might just be some sor of mistake or misreading.

```
sixth <-  ggplot(pollution) +
  geom_qq(aes(sample = pH)) +
  labs(title = "QQ-Plot of pH")
sixth
```



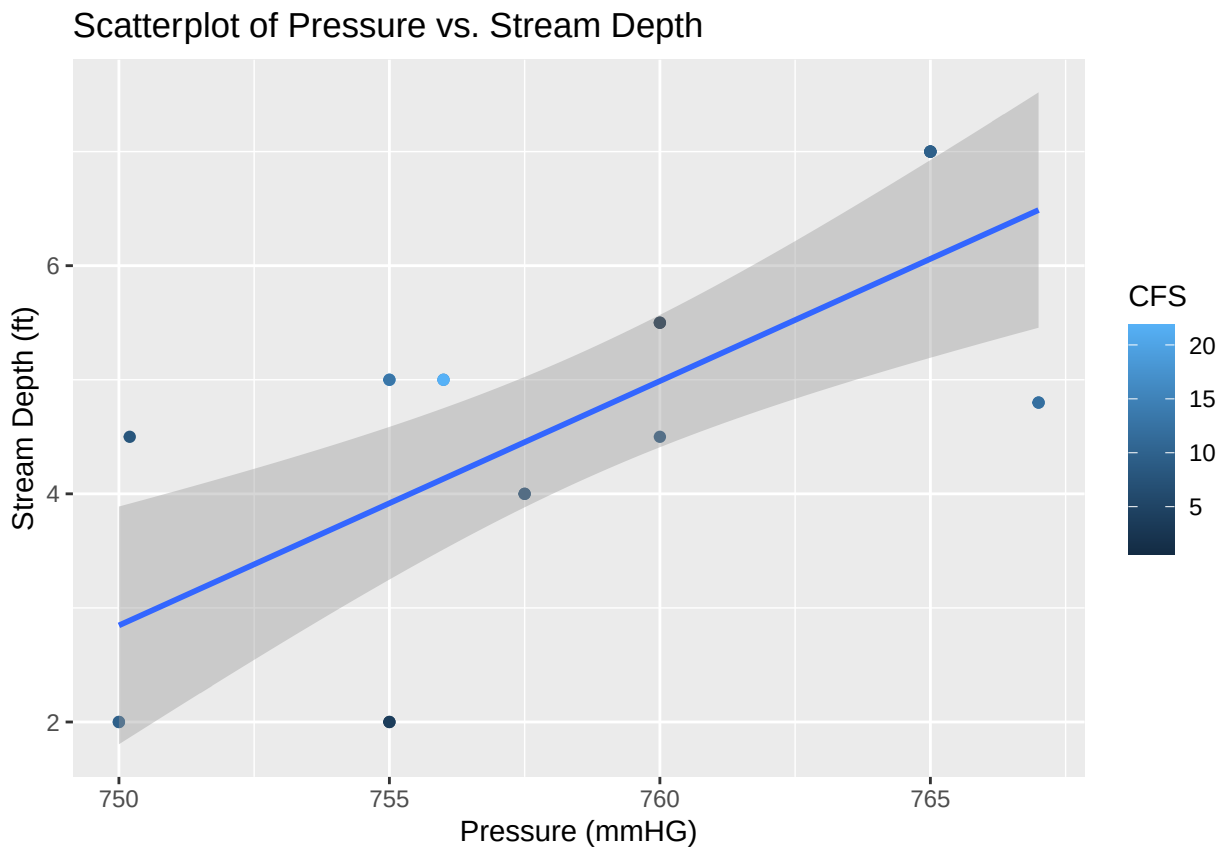
7. Graph of the Frequency This time we decided on a frequency graph that is very similar to a histogram, but the frequency is graphed by lines. We decided that it was important to test our hypothesis about the outliers one last time. This time we graphed the frequency of NH4. While observing this graph we noticed again that there was an outlier which furthermore provides proof that some event may have happened to cause this leap in values. This event may not have effected pH which is why we see no change.

```
seventh <- ggplot(pollution, aes(NH4)) +
  geom_freqpoly(bins = 20) +
  labs(title = "Frequency of NH4")
seventh
```



8. Scatter plot of Pressure vs. Stream Depth After diving into these mysterious outliers, we decided that it would be interesting to see of there was any correlation between pressure, stream depth, and cubic flow. To find this correlation we created a similar graph to the first one we created. We plotted a scatter plot while coloring the points based on cubic flow. We also decided to keep the standard error. The graph helps to show that there is a positive correlation between the pressure and the depth, but no real correlation with cubic flow. The correlation was not very strong, but it was visible.

```
eighth <- ggplot(pollution, aes(Pressure, StrDepth, color = CFS)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Scatterplot of Pressure vs. Stream Depth", y = "Stream Depth (ft)", x = "Pressure (mmHG)
eighth
```


Scatterplot of Pressure vs. Stream Depth
Stream Depth (ft)
6
4
2
750
755
760
765
Pressure (mmHG)
CFS
20
15
10
5