Ellerbee Water Quality

Hank Magan

9/27/2021

1. Introduction

Ellerbee Creek is a tributary of the Neuse River, located in the Piedmont of North Carolina. More specifically, the creek flows through Durham, right by the North Carolina School of Science and Mathematics. The Ellerbee Creek has notably been regarded as the most polluted creek in The Research Triangle region in North Carolina. A variety of data regarding the creek has been compiled into a small dataset, which can be used to examine how certain descriptors regarding the water health are related, as well as why the creek is perceived to be "the most polluted creek in The Triangle" and one of the most polluted in North Carolina. This introduction will examine two measured, commonly found pollutants in detail.

Ammonia (NH_4) is a colorless, pungent gas which happens to be greatly soluble in water. Due to the high solubility of ammonia, it is commonly found in bodies of water such as streams. The compound is found organically, however it is highly concentrated in the discharge a variety of industrial processes, such as fertilization. This discharge often naturally finds itself in bodies of water via runoff, and can also be found in groundwater through the process of infiltration. Luckily for humans, the concentration of ammonia that is present in drinking water has no effect, although the same story is untrue for aquatic life. Ammonia is toxic to fish, and this toxicity increases as both pH and water temperature increase. Due to this, its presence (in concentrated amounts) is linked with poor water quality. Its presence has been recorded in this Ellerbee Creek data and will be looked at in detail shortly.

Additionally, nitrate (NO_3) has been identified to be a harmful molecule in water. Nitrate is a derivative of nitrogen, and therefore occurs organically. With that said, it is most concentrated (and therefore harmful) from man-made runoff, such as waste from human sewage and livestock manure. Additionally, nitrates are used widely in fertilizer, making it even more common as a pollutant. Since livestock is a principle cause of nitrate pollution, agricultural areas are especially vulnerable. When in drinking water, nitrate has to potential to adversely affect humans, which is why the US EDA (United States Environmental Protection Agency) has set the maximum contaminant level at 10mg/L. Perhaps even more worrying is the "algal bloom" caused by nitrate pollution. When nitrogen enters a body of water, it encourages plants to grow. When in high concentrations, algae will bloom at an alarming rate, thus covering the surface of the water in algae. This prevents sunlight from entering the plants in the water, which halts dissolved oxygen production, which kills fish and plants simultaneously.

2. Examine and Clean Data

Having clean data is necessary for performing any form of meaningful data analysis. The term "data cleansing" describes a wide variety of particular processes, all of which work together to transform incomplete, inaccurate, or improperly formatted data into data that is fully complete, making the dataset significantly easier to work with. Essentially, data cleansing prepares data for proper analysis and examination. As such, it is necessary before performing any sort of work on a dataset. To determine if the Ellerbee Creek data is clean, one must examine the structure of the data, along with any other descriptive statistics. R has many commands which are useful for such a task, such as `str`, `summary`, `head`, and `tail`, as demonstrated in the code below.

```
# check descriptive statistics and basic structure
str(ellerbee)
```

```
## '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 ...
```

```
summary(ellerbee)
```

```
##    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     
```

```
head(ellerbee)
```

```
##   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
```

```
tail(ellerbee)
```

```
##    WaterTemp Pressure StrDepth  CFS  DO   pH NO3  Turb Conduct  NH4 saturation
## 15      18.7      755      2.0 3.78 9.4 6.32 1.0 111.0     512 22.7         NA
## 16      18.7      755      2.0 3.78 9.4 6.36 1.0 112.0     507 22.7         NA
## 17      20.6      760      5.5 0.59 8.9 6.95 2.5  17.0     170  1.5         NA
## 18      20.6      760      5.5 0.59 8.5 7.03 2.5  17.0     175  1.8         NA
## 19      22.2      750      2.0   NA  NA 6.82 0.0   0.0     234  0.2         NA
## 20      22.3      750      2.0   NA  NA 6.90 0.0   5.5     230  0.3         NA
```

The first thing to notice is that the dataset is quite small, containing only twenty rows of eleven columns. Moreover, what each column represents can be observed. Each column, as well as its meaning, can be seen below:

1. WaterTemp: in degrees Celsius
2. Pressure: in mmHG
3. StrDepth: stream depth, in feet
4. CFS: cubic flow/sec
5. DO: dissolved oxygen, measured in mg/L
6. pH: acidity
7. NO3: nitrogen content
8. Turb: turbidity, measured in Nephelometric Turbidity Units (NTU)
9. Conduct: conductance
10. NH4: ammonia content
11. saturation

Now that a basic understanding of the data has been reached, it is important to identify and resolve any missing data. This missing data shows up as `NA` in R, and should be replaced before performing analysis as to get accurate results. This process is demonstrated in the code below.

```
# check for missing data
clean <- ifelse(complete.cases(ellerbee) == TRUE, 1, 0)
missing_col <- colnames(ellerbee)[apply(ellerbee, 2, anyNA)]
paste("There are ", dim(ellerbee)[1]-sum(clean),
      " rows with missing data across the following columns:")
```

```
## [1] "There are  18  rows with missing data across the following columns:"
```

```
paste(missing_col)
```

```
## [1] "CFS"        "DO"         "NH4"        "saturation"
```

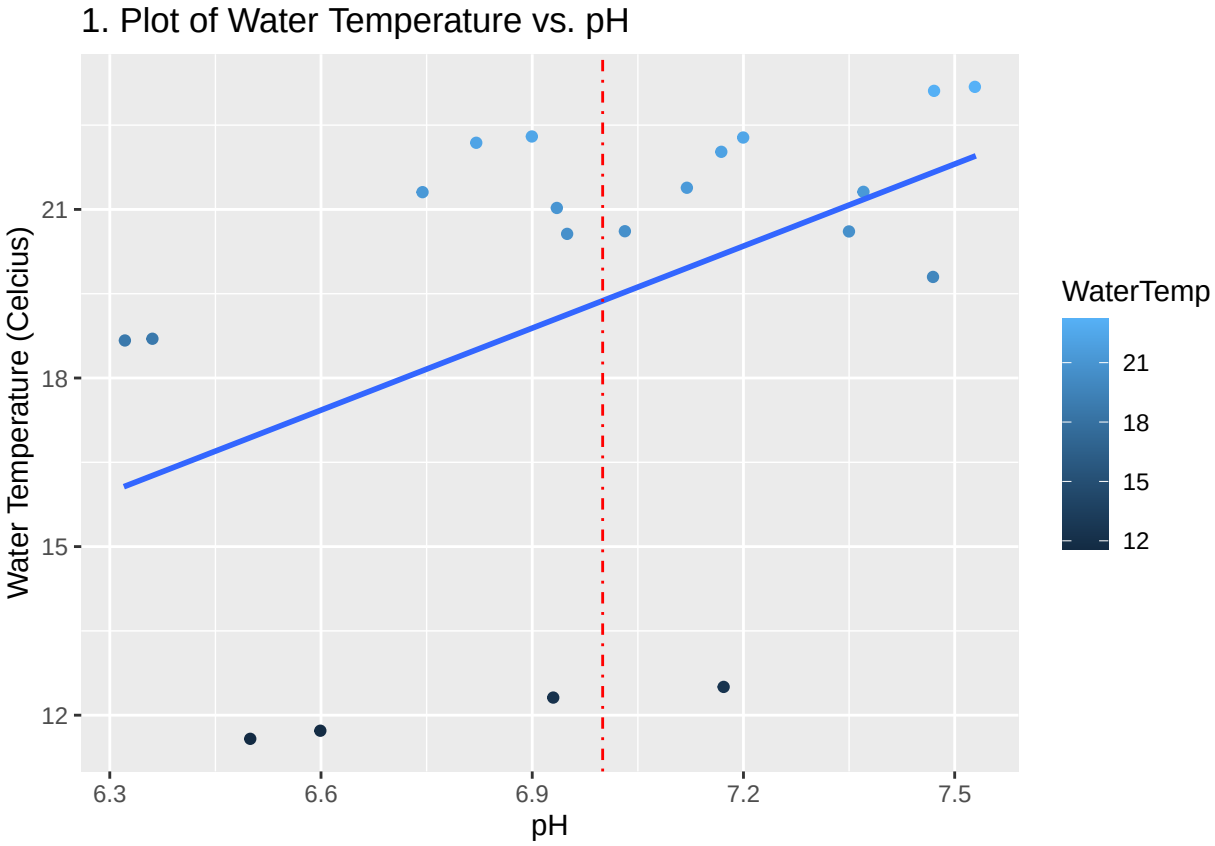
```
# replace NAs with zeroes
ellerbee[is.na(ellerbee)] <- 0
```

Notice that `NA` values were replaced with zeroes. There are many methods of getting rid of `NA` values, such as replacing each with the average of its respective column. It seems that the columns with missing data are not particularly relevant, so replacing with zeroes should be appropriate.

3. Data Modeling

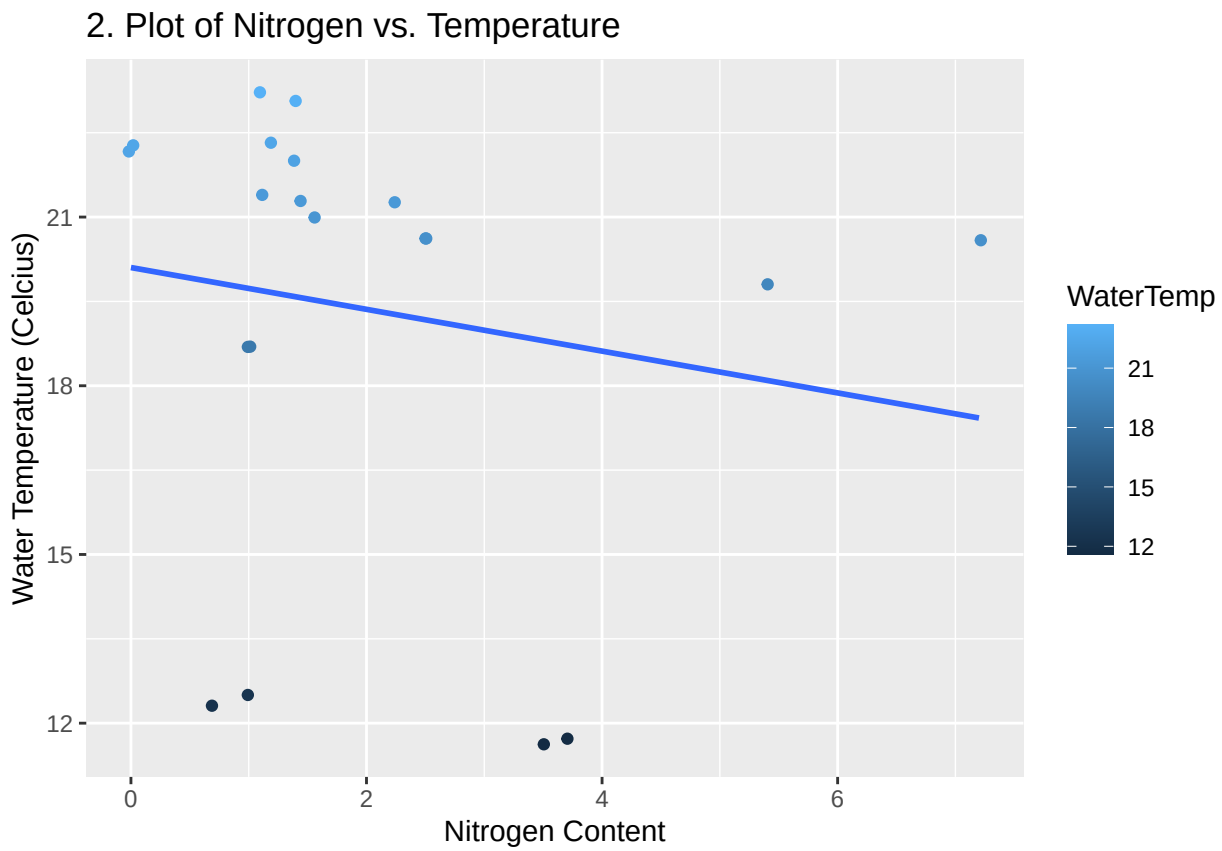
Now that the data has been properly cleaned, analysis may commence. This analysis will be performed primarily via so-called scatter plots. These plots allow correlation to be observed between two variables, which will allow for conclusions to be drawn from the data. It should be noted that the dataset being worked on is very small, and as such significant, uncontested conclusions will not be able to be seen in this report—however, the data can still be examined and correlations may still be observed.

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



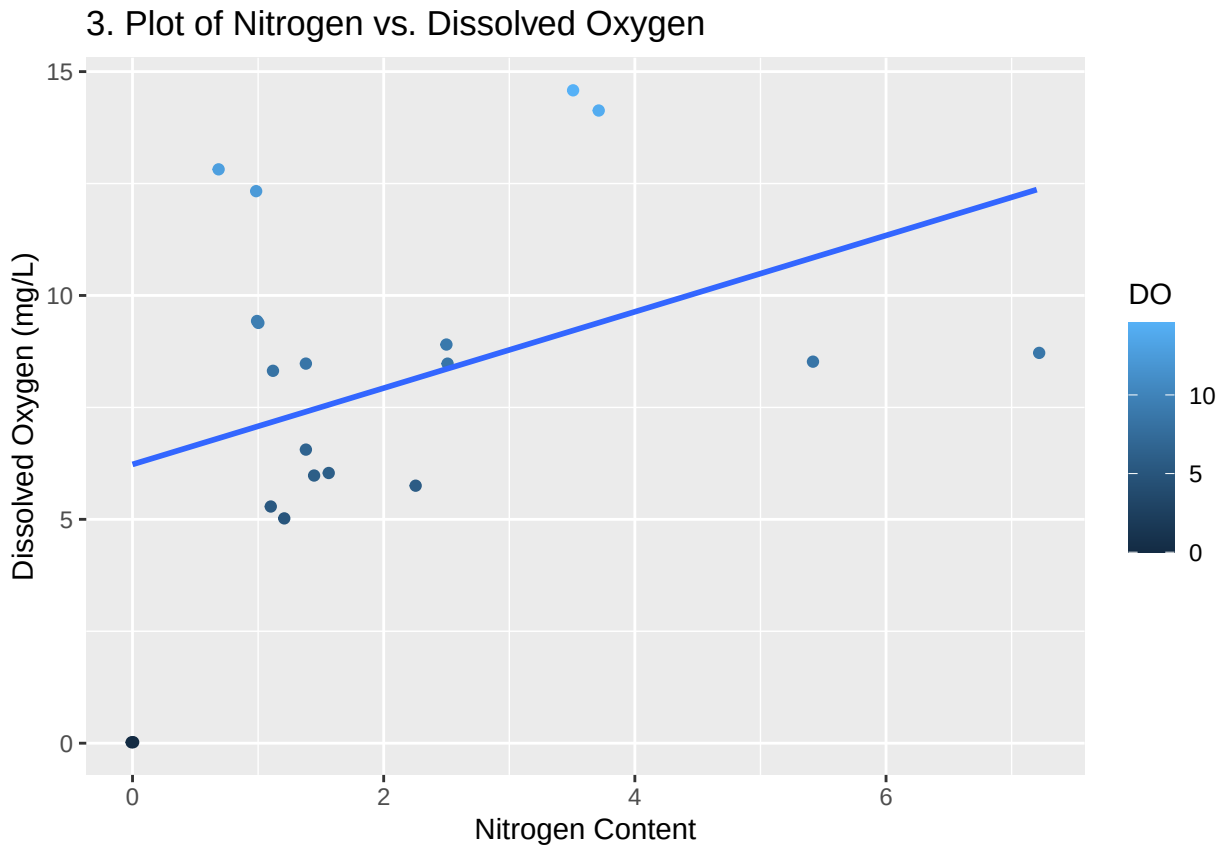
pH is the measure of how acidic or basic water is, and is a crucial indicator of water quality. With aquatic animals (and humans), a certain pH value must be maintained in homeostasis for life, meaning an imbalanced pH can be lethal. This would obviously harm the water quality, however it looks like the pH for Ellerbee Creek is within a reasonable range for life. In fact, many fish can tolerate quite a large pH range, usually hovering around neutral, or a pH level of 7 (marked on the graph by the dotted red line). Interestingly, as the water gets warmer, it also gets more basic (pH increases). Usually the relationship is the opposite.

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



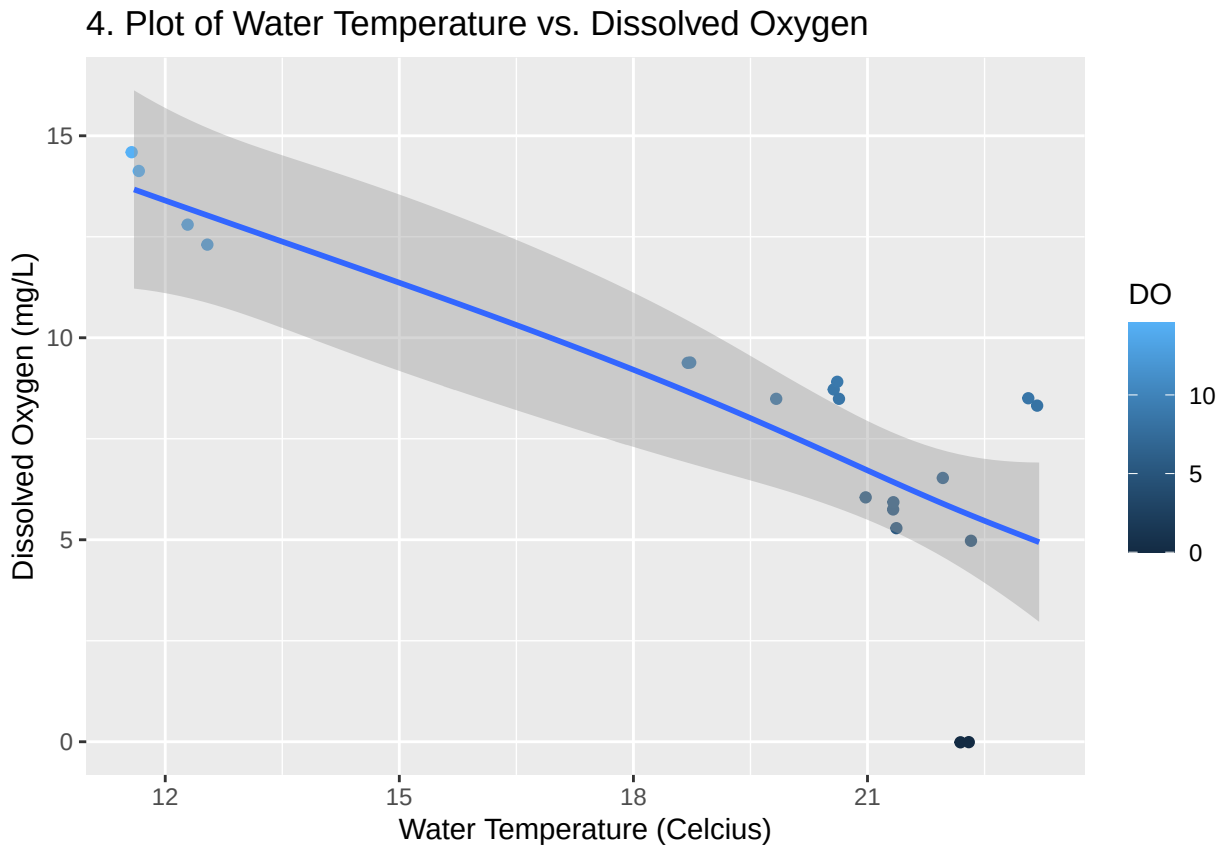
There does not seem to be much correlation between water temperature and nitrate content. The points are wild and all over the place. With that said nitrate can have effects on the water content, despite not affecting temperature. This will be examined in detail in later plots. With that said, observing relationships between uncorrelated variables can still be important. This can eliminate consideration of certain factors when concluding relationships. For example, it was just discovered that pH is positively correlated with temperature, however this relationship us usually inverse. If we were to identify nitrate as factoring into this correlation as a confounding variable, a more significant conclusion could be reached.

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



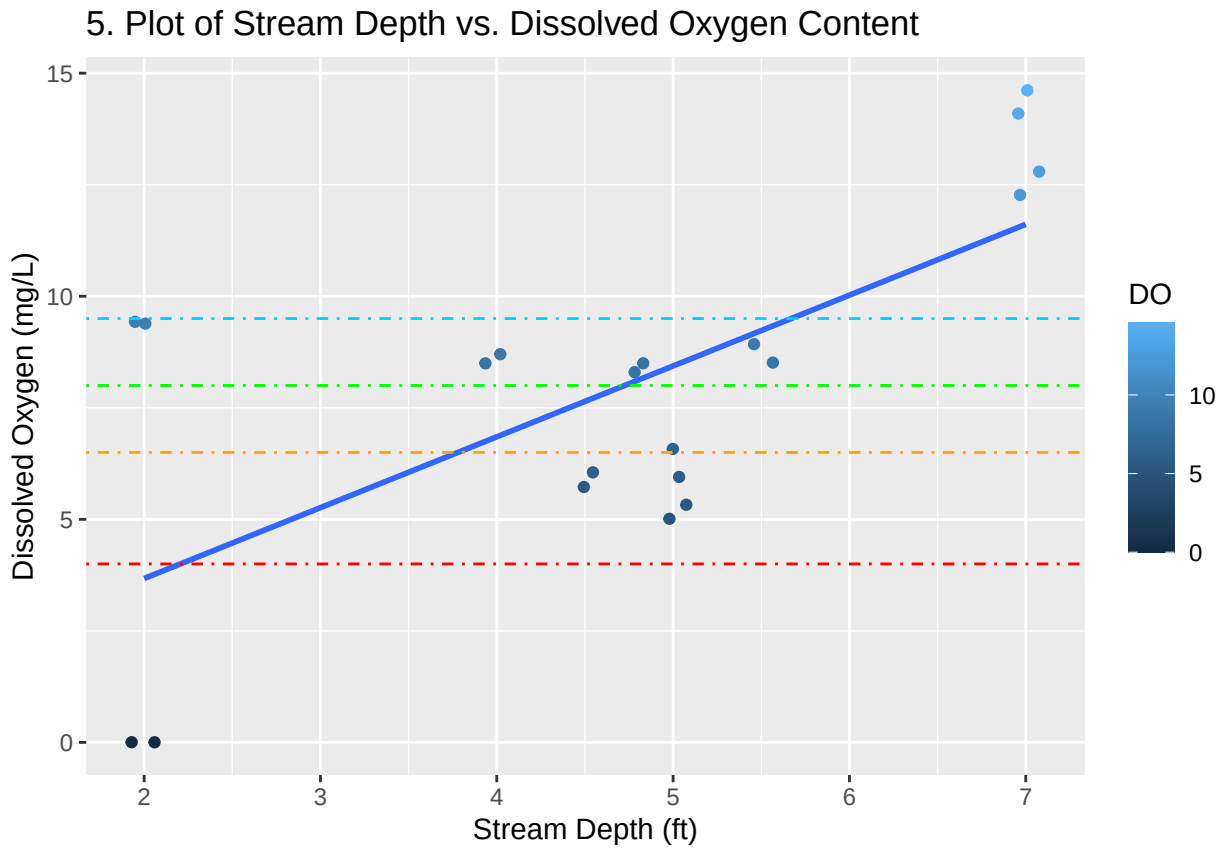
Dissolved oxygen is exactly what it sounds like; oxygen that has been dissolved into water. It is known that oxygen is fundamental to life, and as such, low levels of dissolved oxygen is indicative of poor quality water. In this case, it seems that dissolved oxygen increases as nitrate increases. This is important, as it indicates that the amount of nitrogen present in Ellerbee creek is safe. As mentioned in the introduction, too much nitrogen can cause algal bloom, which would end up depleting oxygen, however the graph does not indicate this. Again, dissolved oxygen is crucial when considering water quality, so it will be examined in detail in the following plots.

```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```



Temperature is also crucial for life, as it often fluctuates quite a bit due to environmental changes. However, as discussed previously, dissolved oxygen content is. This graph shows quite a strong, negative correlation between water temperature and dissolved oxygen content. This is consistent with previous research—cold water can hold more dissolved oxygen than warm water. Due to how quickly the Ellerbee Creek seems to follow this correlation, it seems that it is healthy, at least in this regard.

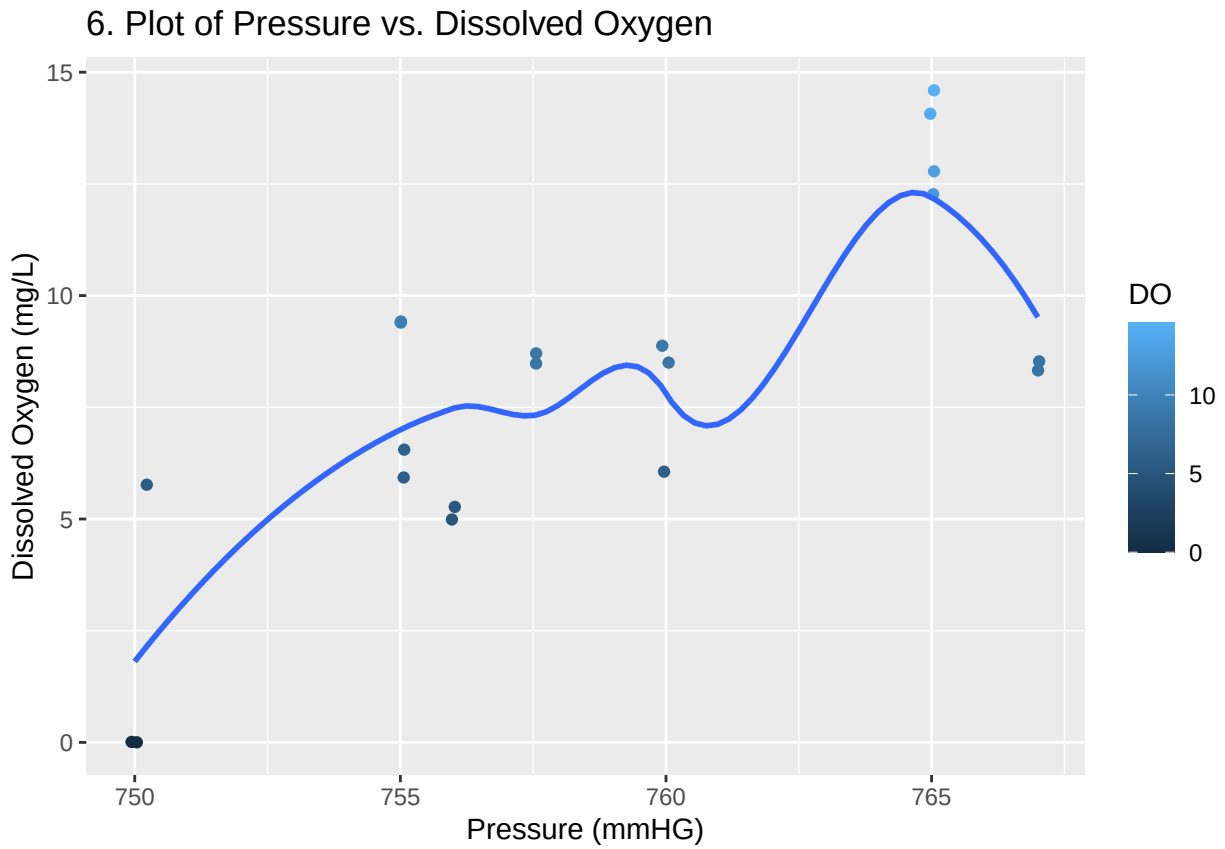
```
## `geom_smooth()` using formula 'y ~ x'
```



This plot further examines dissolved oxygen content. In this graph, a higher stream depth seems to be correlated with a higher content of dissolved oxygen. This result is quite interesting, and there is not quite a discernible reason why this occurs. Of course, with such a small data set, it could be attributed to not having enough data. In fact, in standing bodies of water, this is usually the inverse of what is expected. Regardless, there is little to indicate Ellerbee is particularly unhealthy in terms of dissolved oxygen. Dotted lines have been graphed to demonstrate this, correlating to the dissolved oxygen standards listed below. All of the readings are at least "far", apart from the data artificially manipulated to be zero.

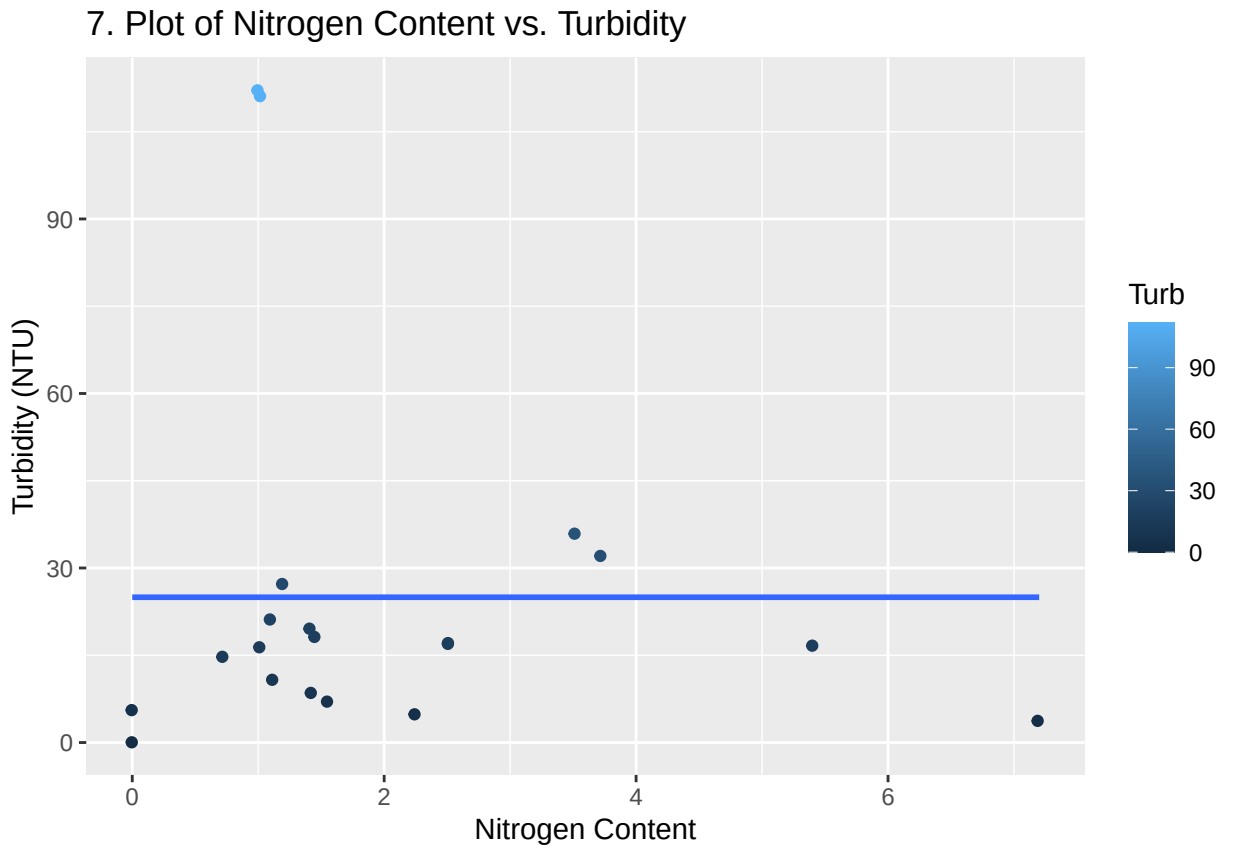
AA. Extraordinary: > 9.5 mg/L
A. Excellent: > 8.0 mg/L
B. Good: > 6.5 mg/L
C. Fair: > 4.0mg/L

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



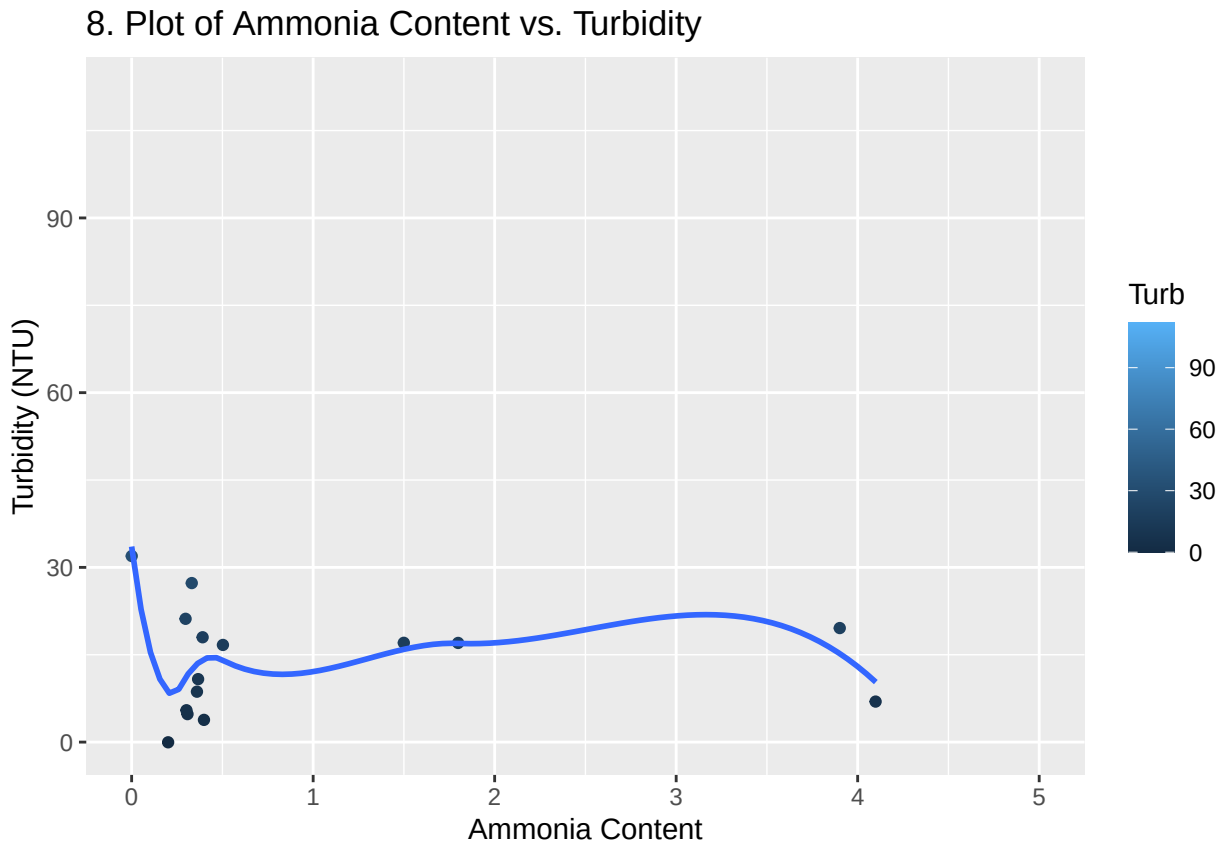
This graphic demonstrates the correlation between pressure, measured in mmHG, and dissolved oxygen content. There is a clear relationship between pressure and dissolved oxygen. This is consistent with the general scientific sentiment. Water under higher pressure can hold more dissolved oxygen. Again, this further demonstrates the healthiness of the creek in terms of dissolved oxygen content, however this is not the only thing that should be considered. The final plots will examine turbidity in the creek, yet another indicator of water quality.

```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```



This graph demonstrates the relationship between nitrogen content and water turbidity. Turbidity is a very important factor in determining water quality. Turbidity is simply the degree to which a substance (in this case, water) is cloudy, or unclear. Obviously, water with low turbidity, such as drinking water, is more likely to be quality. In this case, the linear regression between nitrogen content and turbidity is amazingly nearly a straight line, indicating nearly zero correlation whatsoever between the two variables, which makes sense. Nitrogen is colorless.

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



This graph similarly demonstrates the correlation between ammonia and turbidity. With ammonia, there is a very slight, however inconsistent correlation between its presence and turbidity. When ammonia is presence in water, it has a yellow tint, which could explain the turbidity level. To get a better understanding of what the turbidity value (in NTUs) represents, NTU standards are listen below.

1B. Drinking water: 10
2A. Cold water fishery; all recreation: 10
2B. Cool/warm water fishery, all recreation: 25
2C. Indigenous fish, most recreation: 25

4. References

1. https://www.newsobserver.com/news/local/community/durham-news/article10057493.html
2. https://www.wqa.org/Portals/0/Technical/Technical%20Fact%20Sheets/2014_Ammonia.pdf
3. https://www.wqa.org/Portals/0/Technical/Technical%20Fact%20Sheets/2014_NitrateNitrite.pdf
4. https://www.usgs.gov/special-topic/water-science-school/science/dissolved-oxygen-and-water?qt-science_center_objects=0#qt-science_center_objects