Coliforms in The Panama Canal Watershed: CapDap Final Report

Edward Habeck IV

12/13/2022

Introduction & Hypotheses

Coliforms are microorganisms commonly spread through feces, often known to contaminate sources of water. The abundances of such organisms are studied from water supplies around the world, as are the optimal conditions for them to thrive in laboratory settings. The data which I'm working with depicts water samples from the Panama Canal Watershed, a system responsible for unifying nearby bodies of water to maintain the hydraulic function of The Panama Canal. Coliform count was gathered from four different sample sites: Silvopastures, Secondary Forests, Mature Forests, and Cattle Pastures between seasons. Additionally, variables such as temperature, pH, and dissolved oxygen were measured. From these variables, I have constructed four hypotheses.

Hypothesis #1

Levels of dissolved oxygen will decrease with temperature increase. Studies have demonstrated that increased temperatures correspond to decreased dissolved oxygen content (Walczynska & Sobczyk, 2017). To analyze this from my data, I'll perform simple linear regression to quantify the relationship between dissolved O2 of samples and recorded temperature.

Hypothesis #2

I predict that abiotic factors such as pH will differ between sample sites. As locations where samples were gathered from are characterized by completely different features (e.g. pastures and forests), the composition of water and/or soil should differ significantly as well. I will test the relationship between pH and sample site along with the relationship between *all* sample sites from their pH values.

Hypothesis #3

Coliform count will differ between sample sites due to differences in environmental composition. If variables such as pH differ significantly between locations, coliform growth should be directly affected. Studies have demonstrated that *E. coli* bacteria optimally survive between 20°C and 40°C (Kumar & Libchaber, 2013). For pH, researchers analyzing samples from Osijek-Baranja, Croatia observed a greater abundance of coliforms under a slightly acidic to neutral water pH (Habuda-Stanić, et al, 2013).

Hypothesis #4

Variables will differ in their abilities to predict coliform abundance, and this can be visualized through amultiple regression model. This test will determine how much of a role specific variables play into coliform count, or if variables outside of those collected better explain this.

Data Analysis

To begin, R's memory is cleared and required packages are installed.

```
rm(list = ls())
library(here)
```

```
## here() starts at /home/ewhabe19/Biostatistics Fall 2022/Capstone-Project
```

```
library(ggfortify)
```

```
## Loading required package: ggplot2
```

```
library(tidyverse)
```

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

```
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.0.1      v forcats 0.5.1
## v purrr   0.3.5
```

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

```
## The here package allows for data set to be imported. ggfortify includes everything needed for produc
```

Now, to import the initial data set containing all variables.

```
Bacteria_Data <- read.csv(here("Data", "AS-Raw-Data-Indicator-bacteria-project.csv"), stringsAsFactors="

Bacteria_Data <- na.omit(Bacteria_Data)

## na.omit() assigns itself to Bacteria_Data to remove any rows containing NA. This is necessary in tid
```

Here, I'll look further into the data to get an idea of basic statistics. This includes, min, max, means, and quartiles.

```
summary(Bacteria_Data)
```

```
##     SampleID              Land_Use       Date     Season
##  SF-AS  : 15   Cattle_Pasture  :82   3/26/19:  6   Dry: 70
##  SP-AS  : 15   Mature_Forest   :37   8/27/19:  6   Wet:145
##  CP-AS  : 14   Secondary_Forest:45   2/18/19:  5
##  MF-AS  : 14   Silvopasture    :51   3/11/19:  5
##  CP-W1  : 11                         4/8/19 :  5
##  CP-W3  : 11                         5/28/19:  5
##  (Other):135                         (Other):183
##                Land_Season   Ecoli_MPN         LOG_Ecoli       Coliform_MPN
##  Cattle_Pasture_Wet  :56   Min.   :    0.5   Min.   :0.1761   Min.   :    0.5
##  Secondary_Forest_Wet:31   1st Qu.:   61.0   1st Qu.:1.7923   1st Qu.: 8212.0
##  Silvopasture_Wet    :30   Median :  168.0   Median :2.2279   Median :17328.0
##  Mature_Forest_Wet   :28   Mean   :  773.6   Mean   :2.0566   Mean   :21998.3
##  Cattle_Pasture_Dry  :26   3rd Qu.:  370.0   3rd Qu.:2.5694   3rd Qu.:31062.0
##  Silvopasture_Dry    :21   Max.   :48400.0   Max.   :4.6849   Max.   :80666.7
##  (Other)             :23
##   LOG_Coliform          pH         Conductivity         TDS
##  Min.   :0.1761   Min.   :3.200   Min.   :0.0200   Min.   : 21.0
```

```
##  1st Qu.:3.9145   1st Qu.:6.300   1st Qu.:0.1000   1st Qu.: 69.0
##  Median :4.2388   Median :6.800   Median :0.1600   Median : 84.0
##  Mean   :4.1059   Mean   :6.573   Mean   :0.1942   Mean   : 92.0
##  3rd Qu.:4.4922   3rd Qu.:7.000   3rd Qu.:0.2400   3rd Qu.:112.5
##  Max.   :4.9067   Max.   :8.600   Max.   :0.7500   Max.   :225.0
##
##   Temperature     DissolvedO2      Turbidity          Hardness
##  Min.   :23.60   Min.   :1.220   Min.   :  0.000   Min.   :  0.7
##  1st Qu.:25.20   1st Qu.:4.855   1st Qu.:  1.545   1st Qu.: 25.0
##  Median :26.00   Median :6.080   Median :  3.370   Median : 25.0
##  Mean   :26.14   Mean   :5.815   Mean   :  7.103   Mean   : 38.9
##  3rd Qu.:26.75   3rd Qu.:6.950   3rd Qu.:  6.500   3rd Qu.: 50.0
##  Max.   :33.40   Max.   :8.130   Max.   :143.700   Max.   :250.0
##
```

There do appear to be rather large ranges for some variables, such as Coliform_MPN. However, I'll get a closer look at relevant variables prior to their particular tests.

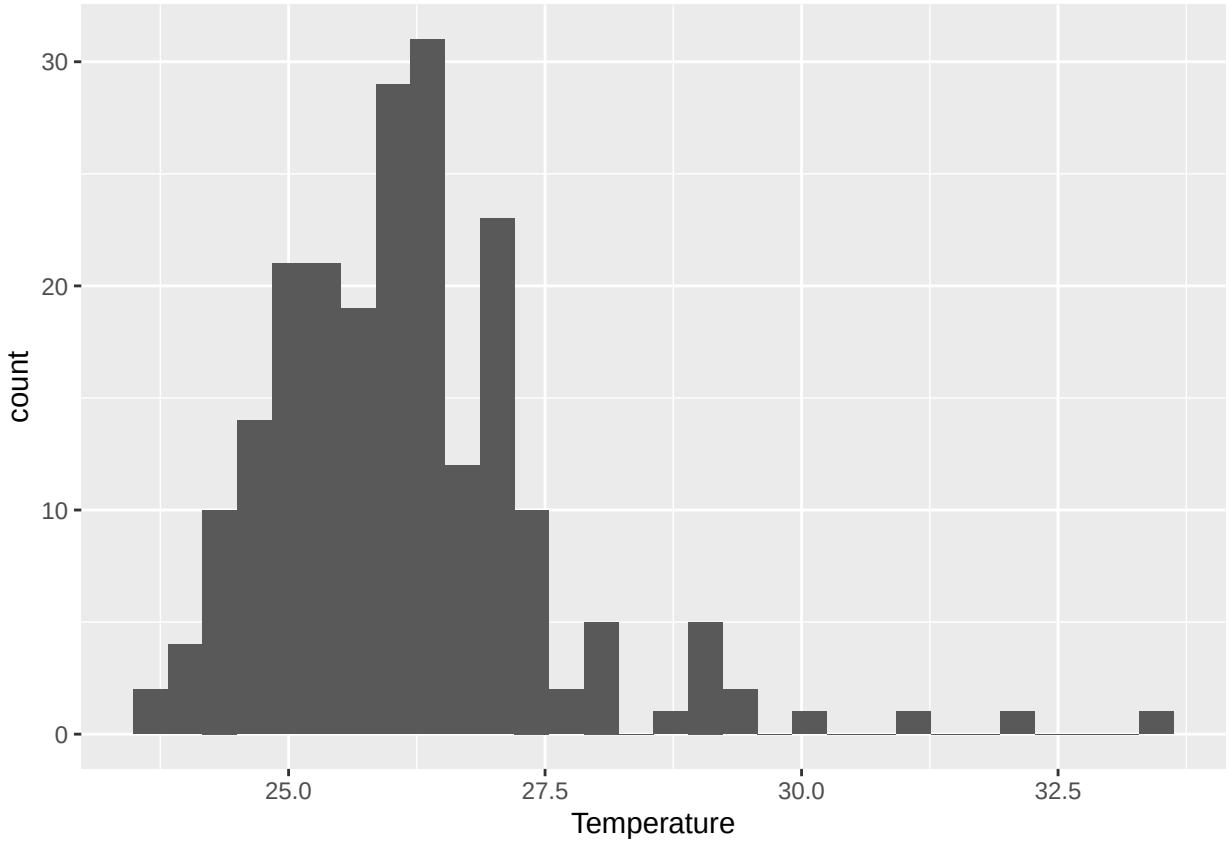
Test #1: Simple Linear Regression to Demonstrate Effects of Temperature on Dissolved O2 Content

```
Bacteria_Data1 <- Bacteria_Data[,13:14]

## Columns 13 and 14 are selected via indexing, as they contain values for temperature and dissolved O2

Bacteria_Data1 <- na.omit(Bacteria_Data1)
```

First, I'll make histograms to better visualize my ranges of dissolved O2 and temperature data.

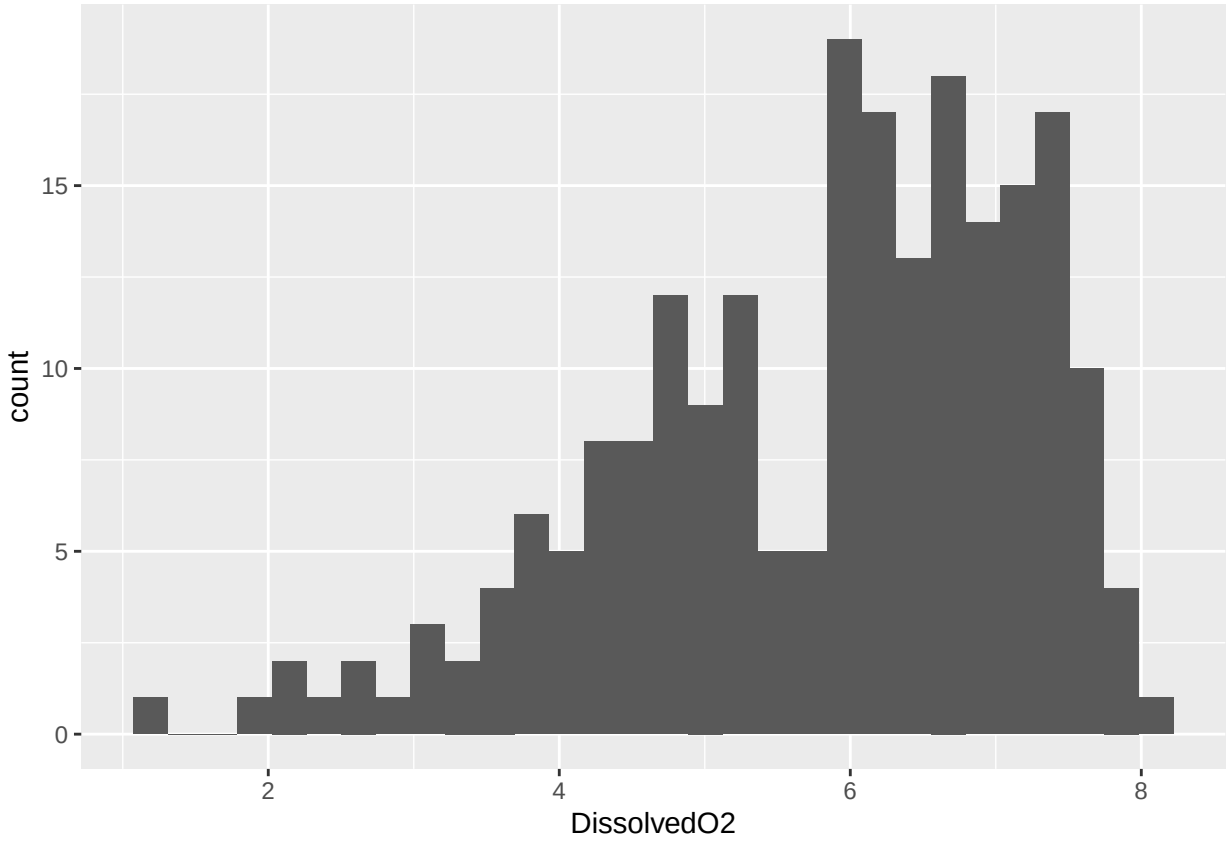
```
ggplot(Bacteria_Data1, aes(x=Temperature)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(Bacteria_Data1, aes(x=DissolvedO2)) +
  geom_histogram()
```

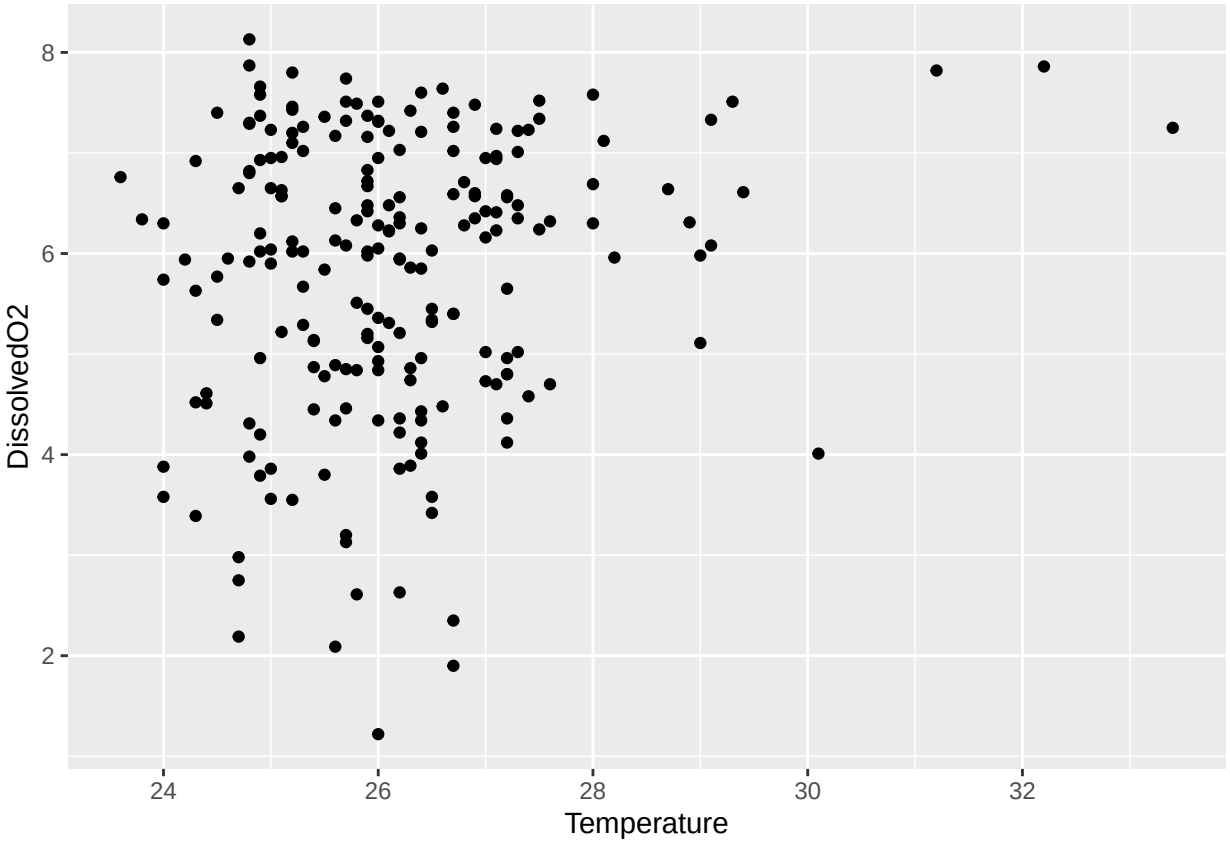
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Dissolved oxygen seems to have an average value of ~6 unspecified units (which I presume to be ppm), while the mean temperature appears to be ~26C. While it looks to me that both variables have a centralized region of points, they aren't evenly distributed along the histogram. In the case of temperature, there are multiple points which trail off one end or the other. Dissolved oxygen has fewer isolated points, but depicts a wide range of values with fewer points towards one side. If there were to be one or two particular points which were very isolated from all other values, then I would be apt to remove them. However, as points seem to trail off from concentrated areas, I'm hesitant to call any outliers.

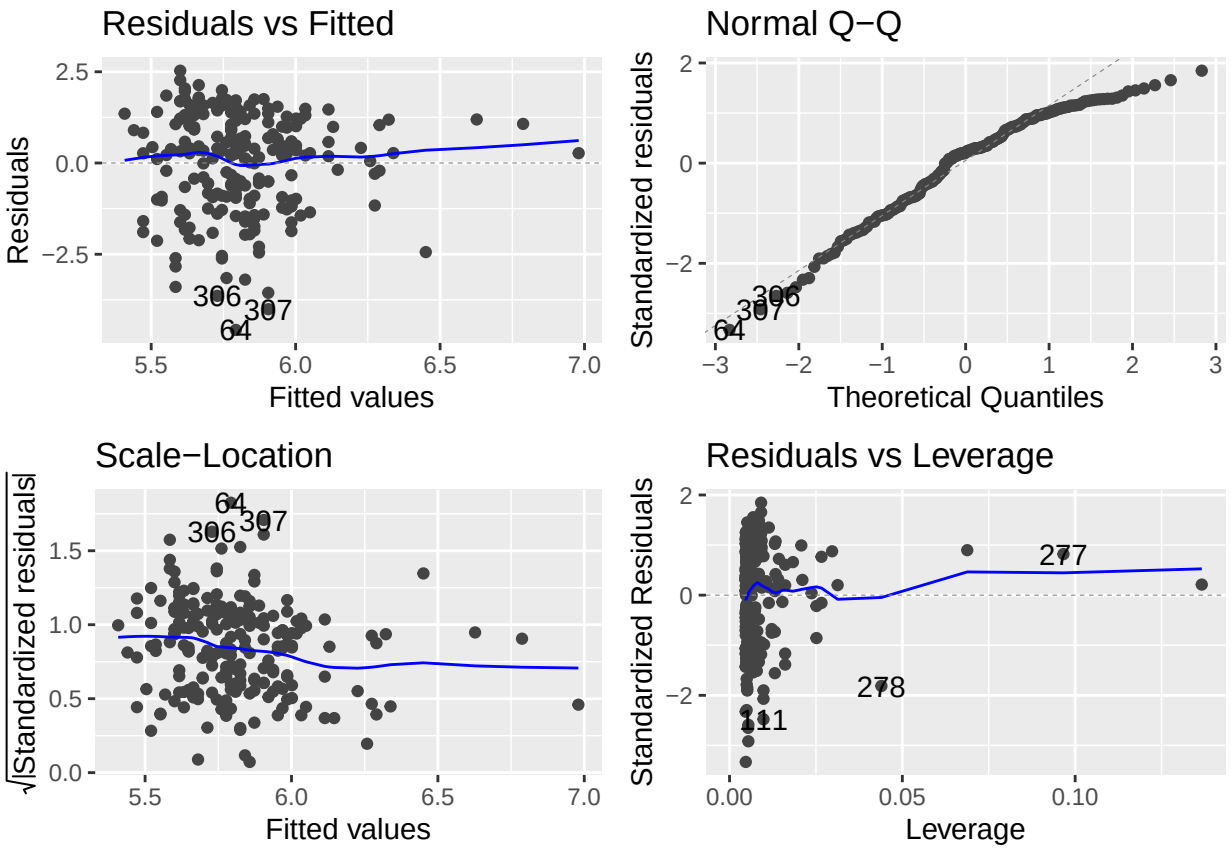
I will now create a basic scatterplot depicting where dissolved O2 values fall along a temperature range. This is an appropriate plot to use, as both temperature and dissolved O2 are forms of continuous data.

```
ggplot(Bacteria_Data1, aes(x=Temperature, y=DissolvedO2)) +
  geom_point()
```



There isn't a noticeably strong connection between variables in this scatterplot. I do still believe that there is at least a weak positive connection based upon the arrangement of points, however. I would estimate the y-intercept of this plot to be ~4.6, and the slope to be ~0.375 (two points on my estimated line being ~0.3 (rise) and ~0.8 (run) away from one another). This imaginary line is most definitely moving in a positive direction, meaning that my hypothesis of decreased dissolved oxygen in higher temperatures may be refuted. This will still require me to perform statistical tests and create a real linear model, though.

Autoplot will determine if this data is adequate for statistical testing, visualized in fitted values (which should be flat lines) and theoretical quantiles (ideally points following a linear trend).



My fitted values are relatively flat with some deviation. Theoretical quartile points generally follow a linear trend, however, there is significant deviation towards its end. At the moment, a good model for working with oddly distributed data cannot be found, so standard testing with a linear model will be conducted.

Now, to perform my test using my standard linear model. As both temperature and dissolved O2 are continuous forms of data, simple linear regression should be used.

```
##
## Call:
## lm(formula = DissolvedO2 ~ Temperature, data = Bacteria_Data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5731 -0.9432  0.2717  1.0690  2.5294
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.62323    1.80274   0.900   0.3689
## Temperature  0.16038    0.06888   2.329   0.0208 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.377 on 213 degrees of freedom
## Multiple R-squared:  0.02482,    Adjusted R-squared:  0.02024
## F-statistic: 5.422 on 1 and 213 DF,  p-value: 0.02082
```

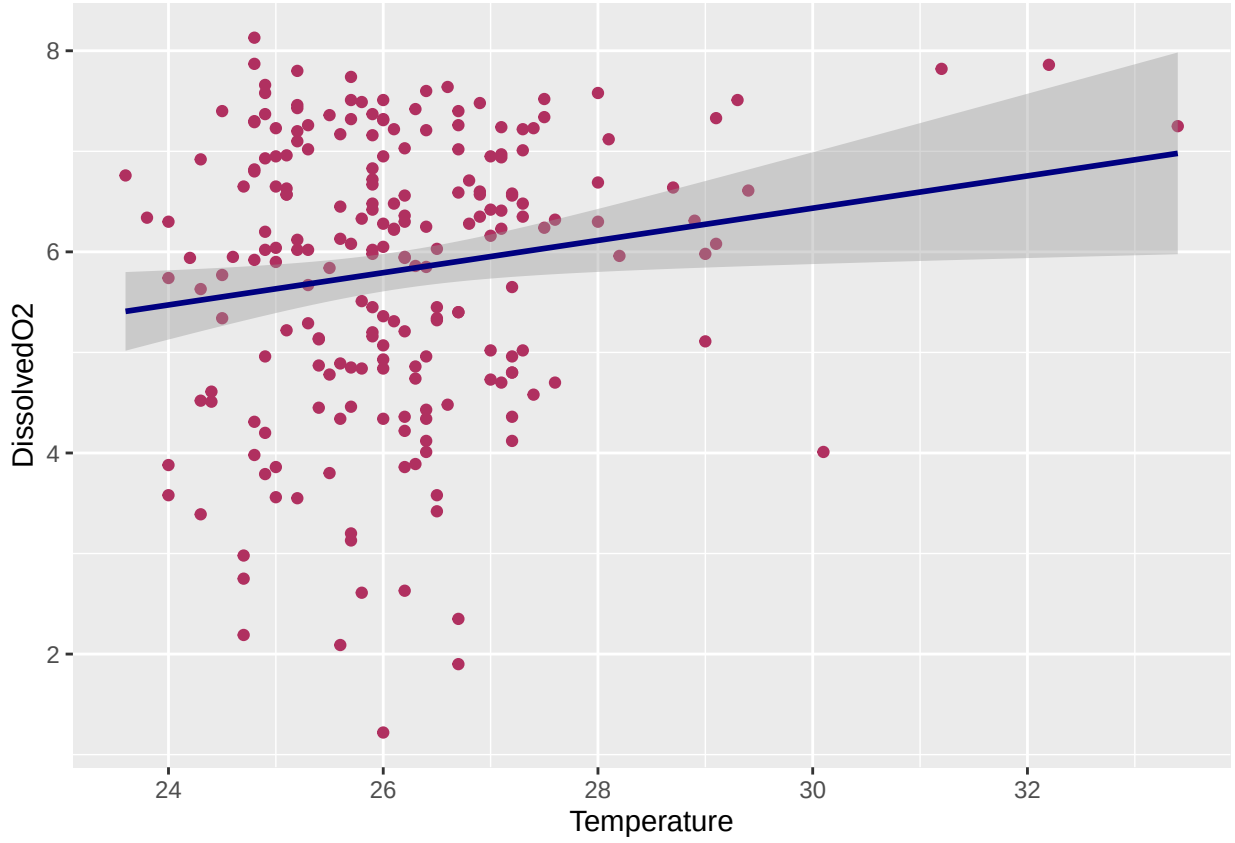
This summary confirms that temperature is a significant predictor of dissolved O2 in samples from the Panama Canal Watershed, its p-value being 0.0208 (less that 0.05). One thing to consider here is the R-squared value,

which is 0.02024. Even though there is a significant relationship between temperature and dissolved O2, this R-squared value tells me that temperature only describes ~2% of dissolved O2 content. This means that the remaining ~98% of dissolved O2 content can be explained through other variables.

Finally, here is a depiction of the linear relationship between points when plotting temperature against dissolved O2.

```
ggplot(Bacteria_Data1, aes(x=Temperature, y=DissolvedO2)) +
  geom_point(colour="maroon") +
  geom_smooth(method = 'lm', colour="navy")
```

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




```
## geom_smooth with 'lm' creates a linear model of the data. This additional feature also depicts a wea
```

This plot refutes my initial hypothesis of dissolved oxygen content decreasing in response to temperature increase. If my hypothesis were to be supported, the linear model would not be moving in a positive direction. Though there *is* technically a significant correlation between water temperature and dissolved oxygen content, my r-squared value demonstrates that temperature can only describe ~2% of dissolved oxygen in samples. Therefore, temperature shouldn't be considered the best indicator of dissolved oxygen in these samples to begin with.

Test #2: One-Way Anova to Determine the Relationship Between Sample Site and pH

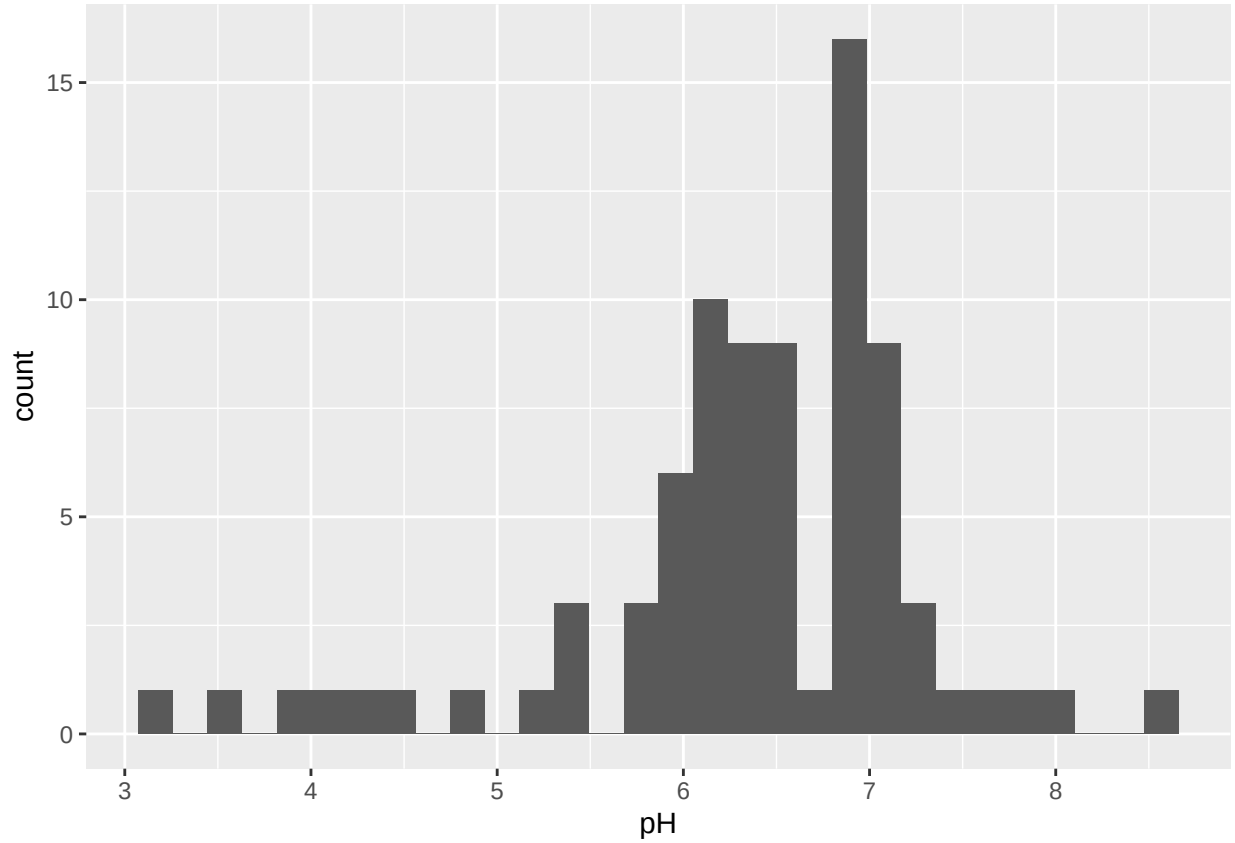
I begin by setting up my new dataset and removing rows containing NA values. I will then create four histograms (one of pH values for each sample site) to make comparisons and predictions before testing.

```
Bacteria_Data2 <- Bacteria_Data[c(2, 10)]

Bacteria_Data2 <- na.omit(Bacteria_Data2)
```

```
Cattle <- Bacteria_Data %>% filter(Land_Use == "Cattle_Pasture")

Mature <- Bacteria_Data %>% filter(Land_Use == "Mature_Forest")

Secondary <- Bacteria_Data %>% filter(Land_Use == "Secondary_Forest")

Silvopasture <- Bacteria_Data %>% filter(Land_Use == "Silvopasture")

## New data sets are created specific to each sample site. This way, plots can be easily constructed fr
```

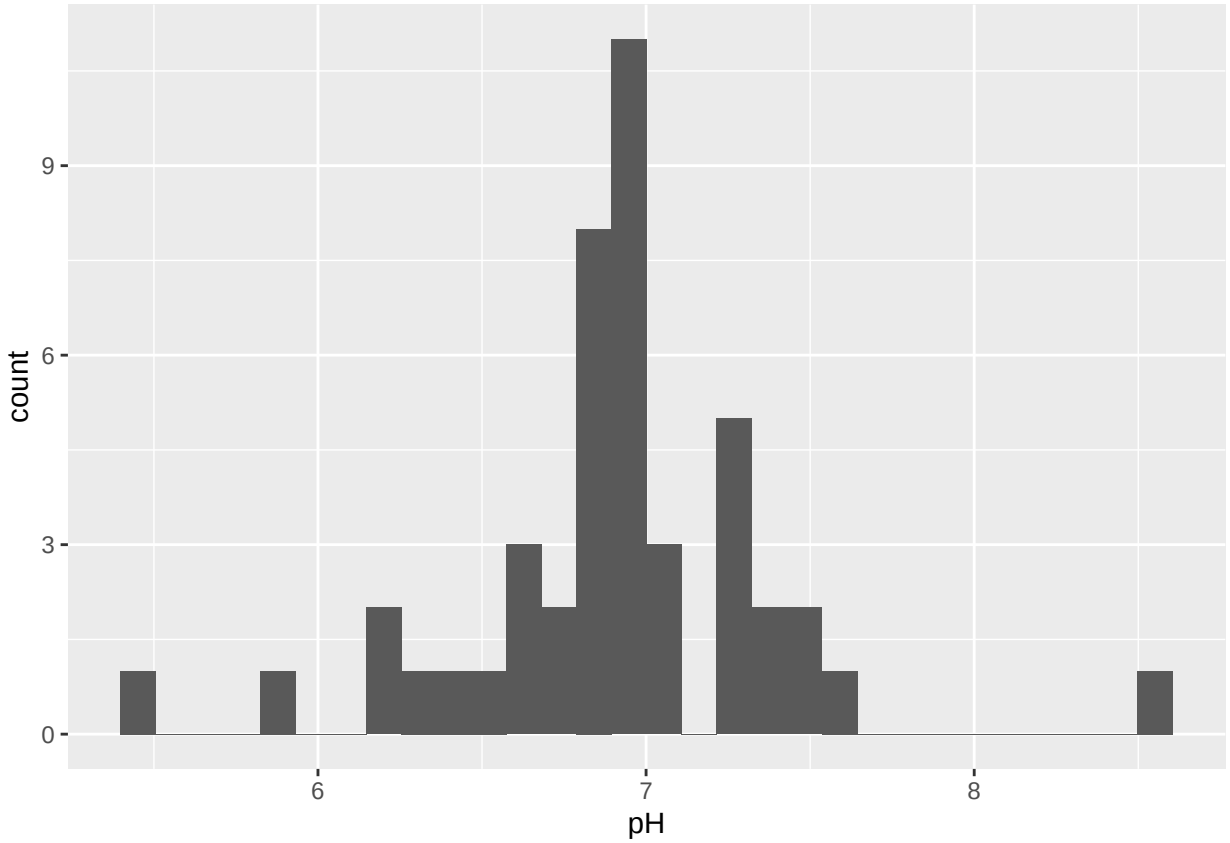
```
ggplot(Cattle, aes(x=pH))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



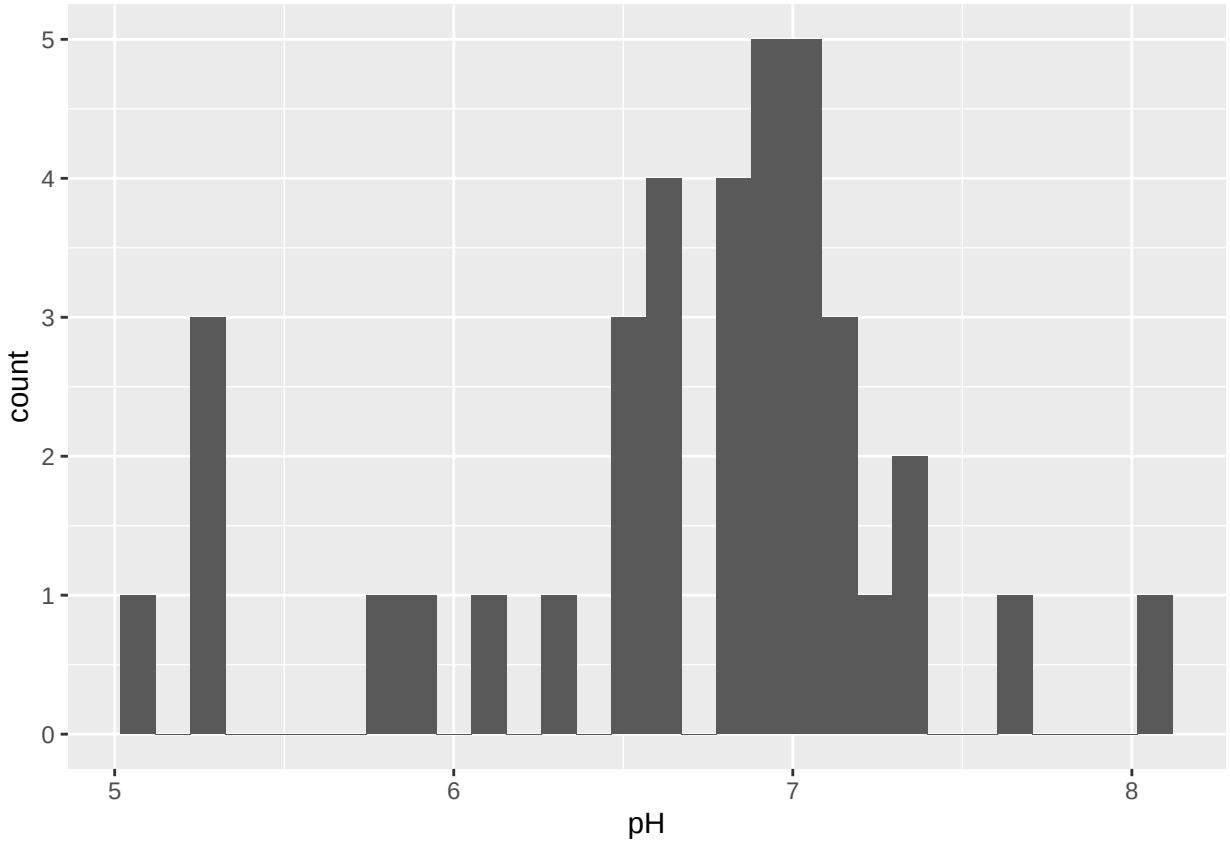

```
ggplot(Secondary, aes(x=pH))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



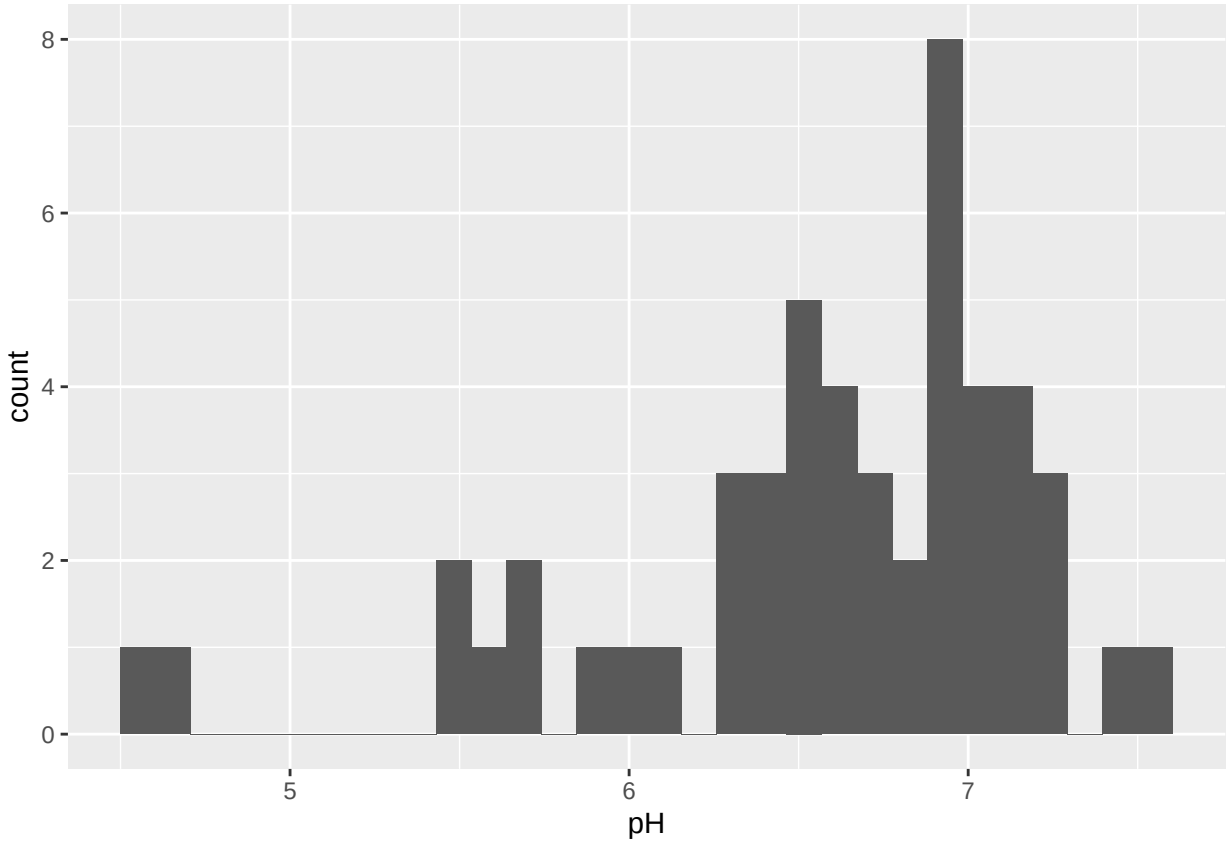

```
ggplot(Mature, aes(x=pH))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(Silvopasture, aes(x=pH))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The approximate means from my visual estimations appear to be as follows:

Cattle Pasture: ~6.5

Secondary Forest: ~6.7

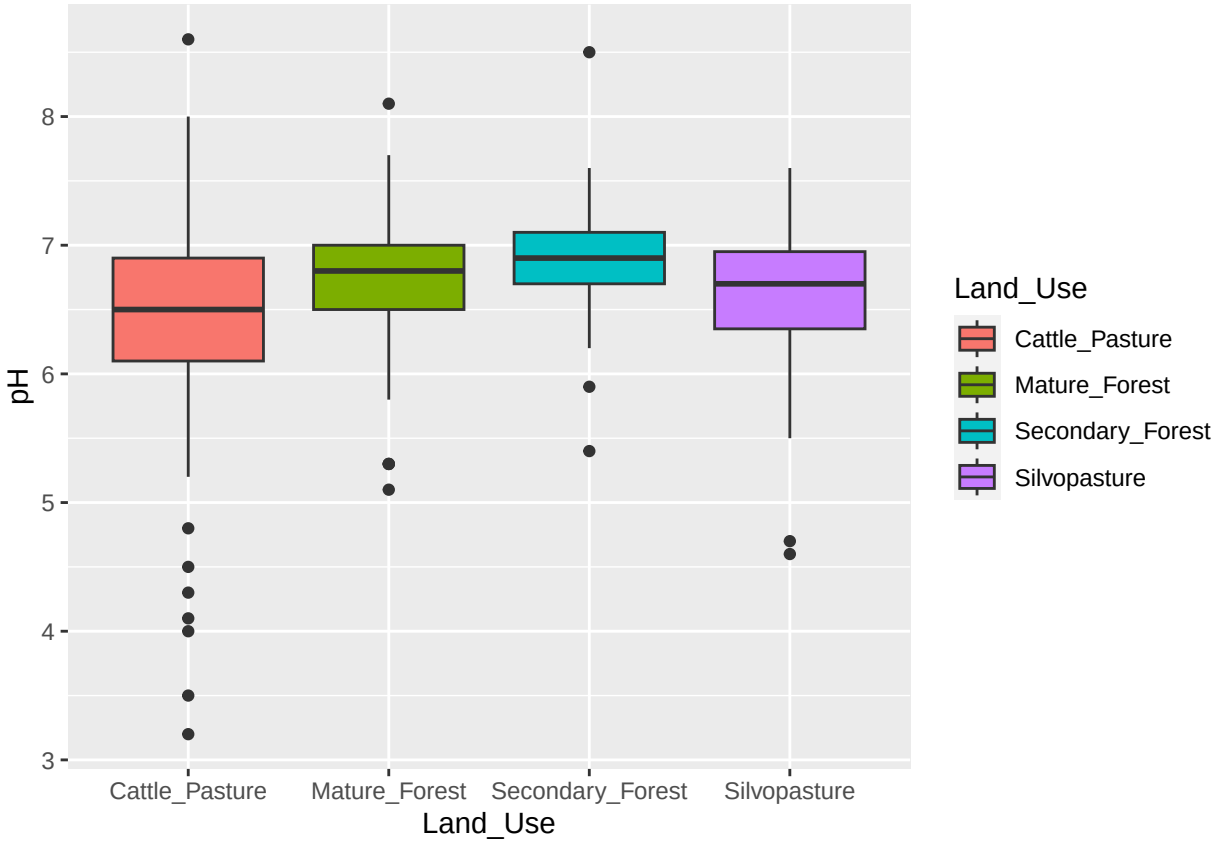
Mature Forest: ~6.8

Silvopasture: ~6.8

There do not appear to be any major differences between these histograms. Though data is distributed differently across sample sites, a large degree of overlap occurs at around 6.7-7.1. From this, I presume there will be at most a weak positive realatiosnhip, but there don't seem to be any significant differences between any sites prior to running tests.

Now to create a figure depicting pH by sample site. A boxplot will be an appropriate model, as pH is a continuous form of data and sample site is categorical.

```
ggplot(Bacteria_Data2, aes(x = Land_Use, y = pH, fill=Land_Use)) +
  geom_boxplot()
```

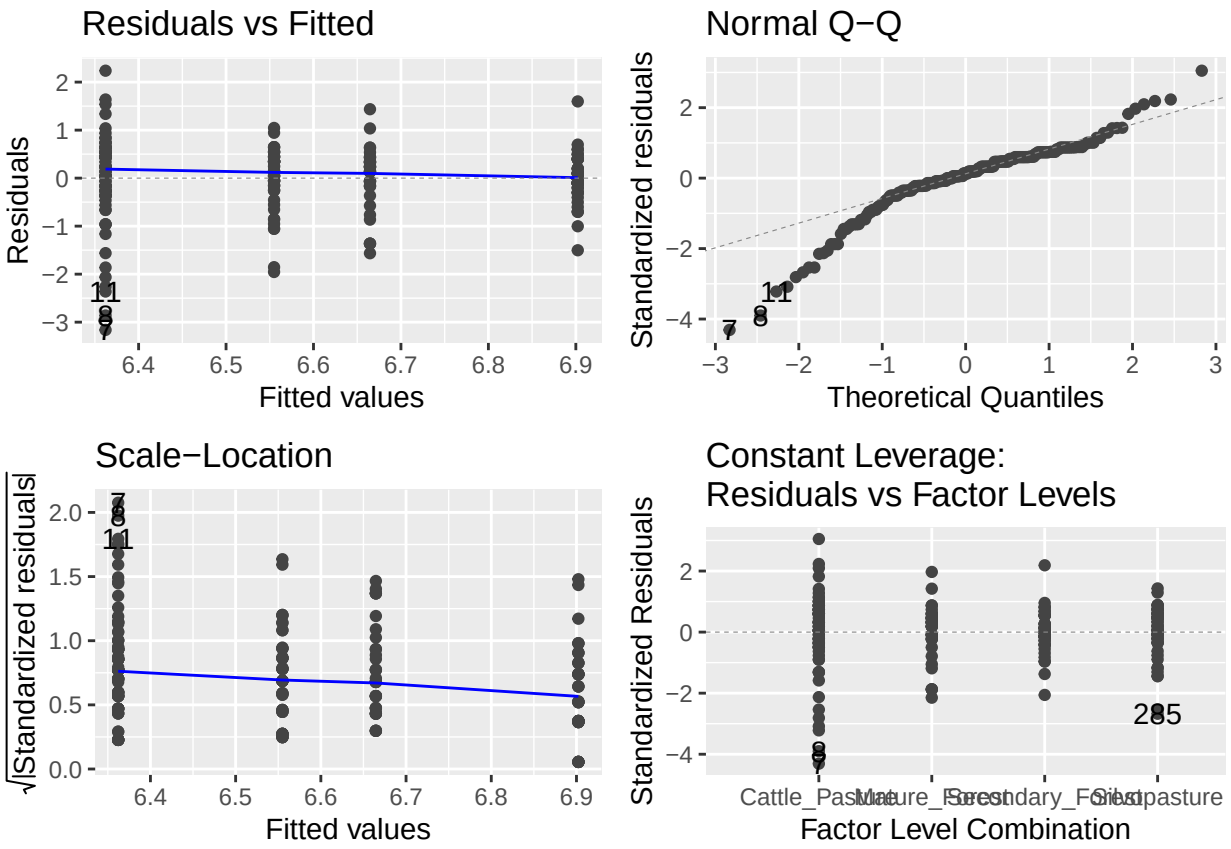



```
## Land_Use will be my x-axis, as it is categorical. Points should be plotted above each site from the
```

As I'd expected, there still does not appear to be much variation between boxes. Again, there is a high degree of overlap in points around 6.7-7.1, though each site has very differently distributed values and ranges.

As the predictor variable is categorical (Land_Use) and the response variable is continuous (pH), a one-way Anova test should be used to determine significance in their relationship.

```
Land_pH <- lm(pH ~ Land_Use, data=Bacteria_Data2)

autoplot(Land_pH)
```



Fitted value lines are relatively flat, but theoretical quantiles deviate from the linear model towards the beginning points. As a good model for this irregular data cannot be found at the moment, I will continue to use a standard linear model for statistical tests.

```
anova(Land_pH)
```

```
## Analysis of Variance Table
##
## Response: pH
##            Df  Sum Sq Mean Sq F value   Pr(>F)
## Land_Use    3   8.849 2.94951  5.4105 0.001322 **
## Residuals 211 115.025 0.54514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to this test, pH is significantly dependent on sample site. However, we can look further into how sample sites individually relate to one another by pH. From my initial boxplot, I predicted that sites would **not** be significantly different from one another. To determine this, a Tukey test will be conducted.

```
library(multcompView)

Bacteria_Data2aov <- aov(pH ~ Land_Use, data=Bacteria_Data2)

## Anova test performed on solely Land_Use. This tells us that coliform count differs significantly bet

summary(Bacteria_Data2aov)
```

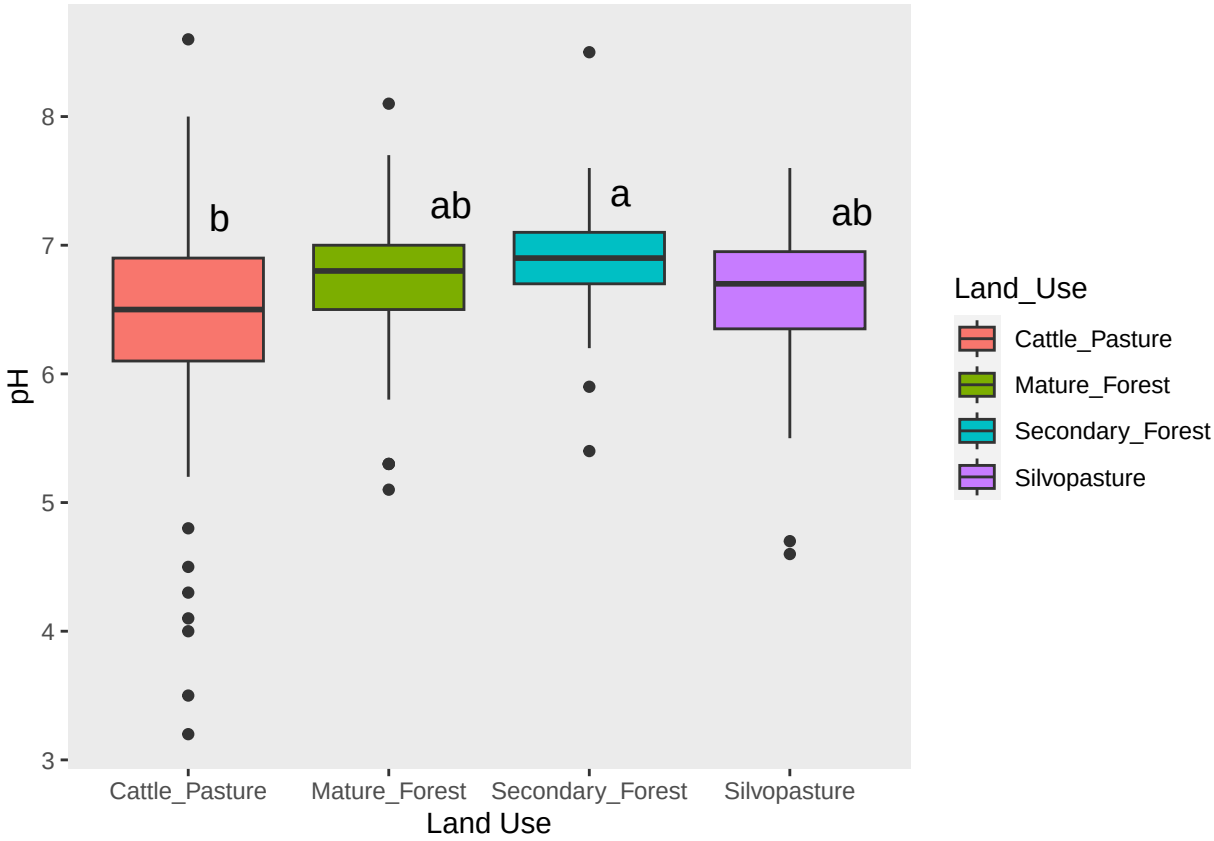
```
##              Df Sum Sq Mean Sq F value  Pr(>F)
```

```
## Land_Use      3   8.85  2.9495   5.411 0.00132 **
## Residuals   211 115.02  0.5451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Bacteria_Data2Tukey <- TukeyHSD(Bacteria_Data2aov, conf.level=.95)

##Tukey test performed--set to a variable.

Bacteria_Data2Tk <- group_by(Bacteria_Data2, Land_Use) %>%
  summarise(mean=mean(pH), quant = quantile(pH, probs = 0.75)) %>%
  arrange(desc(mean))

##Statistics regarding mean and quartiles made into a table, which is then used to create letters on ou

Bacteria_Data2cld <- multcompLetters4(Bacteria_Data2aov, Bacteria_Data2Tukey)
Bacteria_Data2cld <- as.data.frame.list(Bacteria_Data2cld$Land_Use)
Bacteria_Data2Tk$Bacteria_Data2cld <- Bacteria_Data2cld$Letters

## Letters assigned to different levels of Land_Use on table (A, B, AB).

print(Bacteria_Data2cld)
```

```
##                 Letters monospacedLetters LetterMatrix.a LetterMatrix.b
## Secondary_Forest       a                a            TRUE          FALSE
## Mature_Forest         ab               ab            TRUE           TRUE
## Silvopasture          ab               ab            TRUE           TRUE
## Cattle_Pasture         b                b           FALSE           TRUE
```

```
ggplot(Bacteria_Data2, aes(Land_Use, pH)) +
  geom_boxplot(aes(fill=Land_Use)) +
  xlab("Land Use") +
  ylab("pH") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
   geom_text(data = Bacteria_Data2Tk, aes(x = Land_Use, y = quant, label = Bacteria_Data2cld), size = 5
```




```
##Box-plot created and formatted. Letters must be raised and increased in size once set on plot. Label
```

Letters which match indicate no statistically significant difference from one another. Letters which differ (not counting boxes with two letters) *do* significantly differ.

This box-plot demonstrates that Cattle Pasture, Mature Forest, and Silvopasture do not differ significantly from one another. Secondary Forest, Mature Forest, and Silvopasture do not either. This means that the only significant difference in pH of sample sites occurs between Cattle Pasture and Secondary Forest.

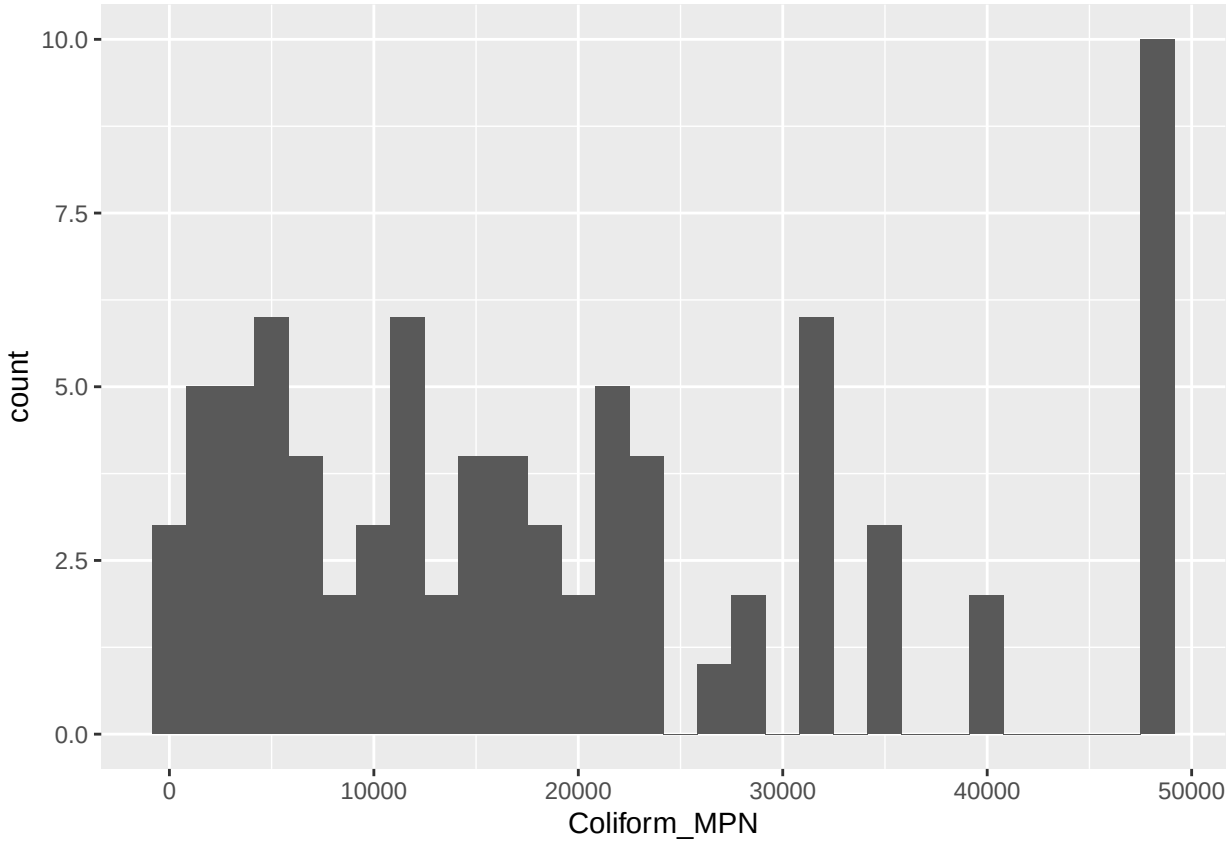
Test #3: One-Way Anova to Determine the Relationship Between Sample Site and Coliform Abundance

```
Bacteria_Data3 <- Bacteria_Data[c(2, 8)]

Bacteria_Data3 <- na.omit(Bacteria_Data3)
```

I will begin by checking how coliform count data is distributed among the four sample sites. A histogram will again be created for each.

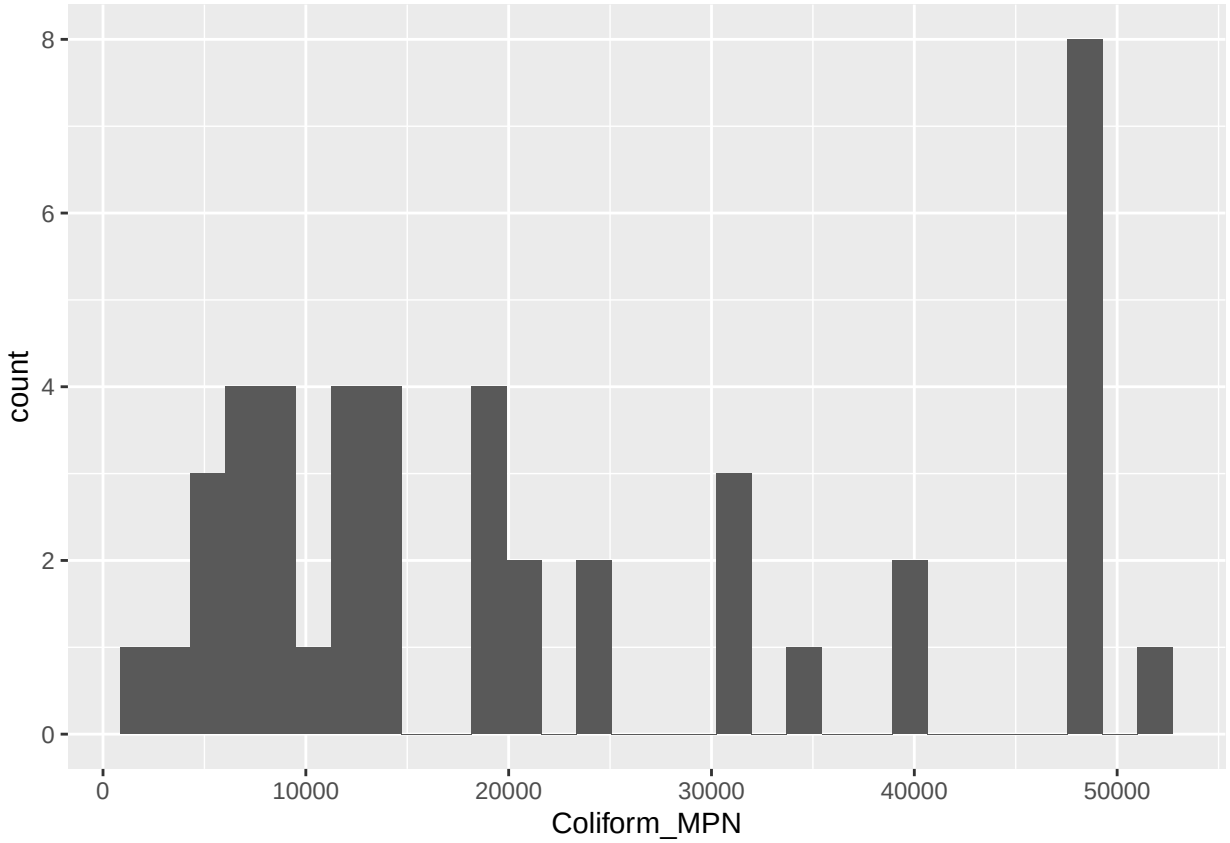
```
ggplot(Cattle, aes(x=Coliform_MPN))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



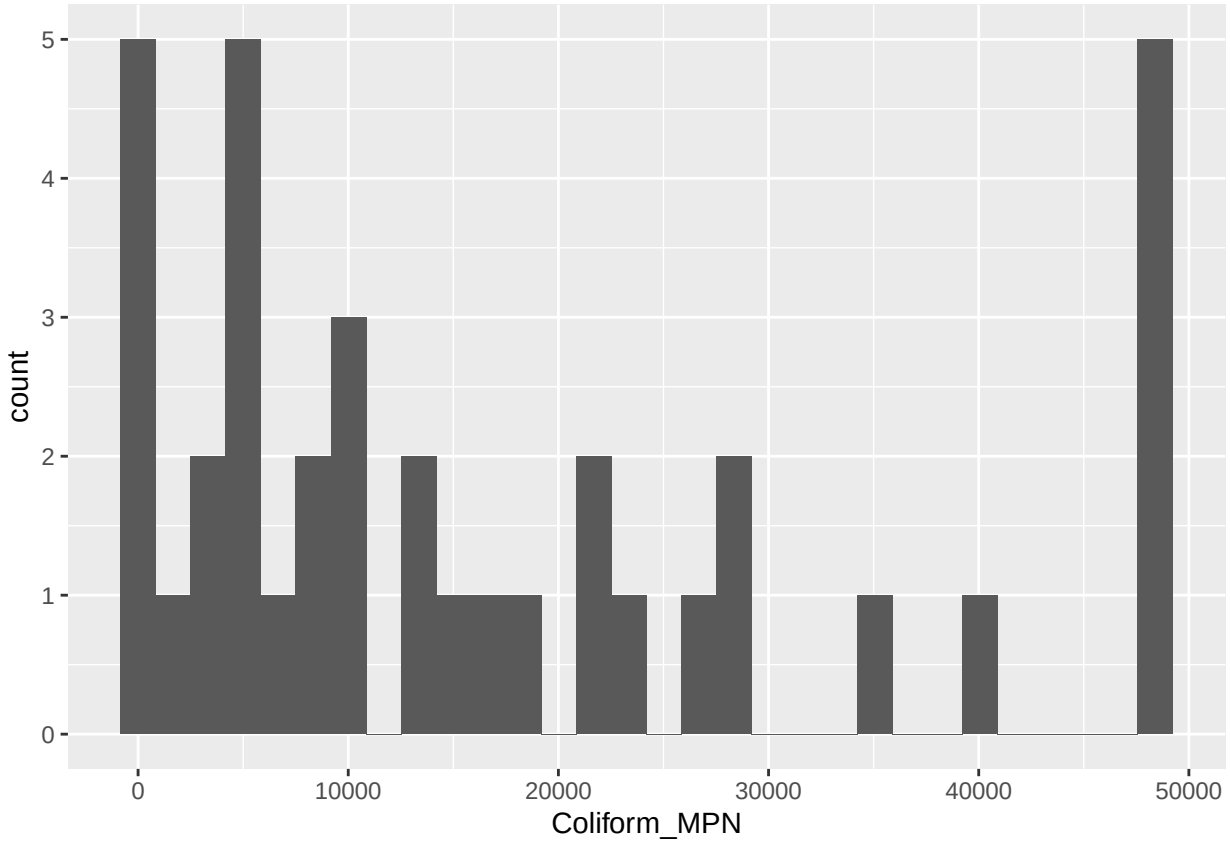

```
ggplot(Secondary, aes(x=Coliform_MPN))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



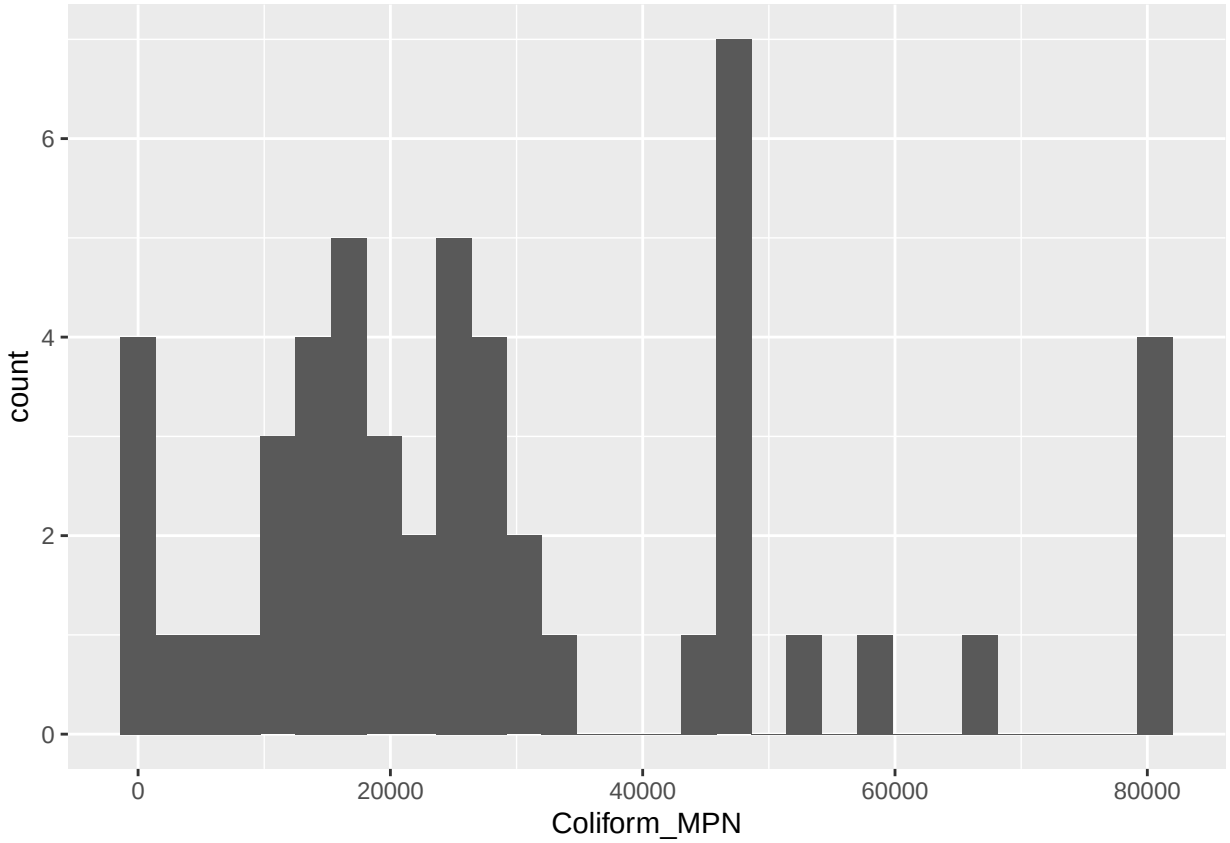

```
ggplot(Mature, aes(x=Coliform_MPN))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(Silvopasture, aes(x=Coliform_MPN))+
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



For coliform count, my estimated means are as follows:

Cattle Pasture: ~20,000 MPN

Secondary Forest: ~20,000 MPN

Mature Forest: ~11,000 MPN

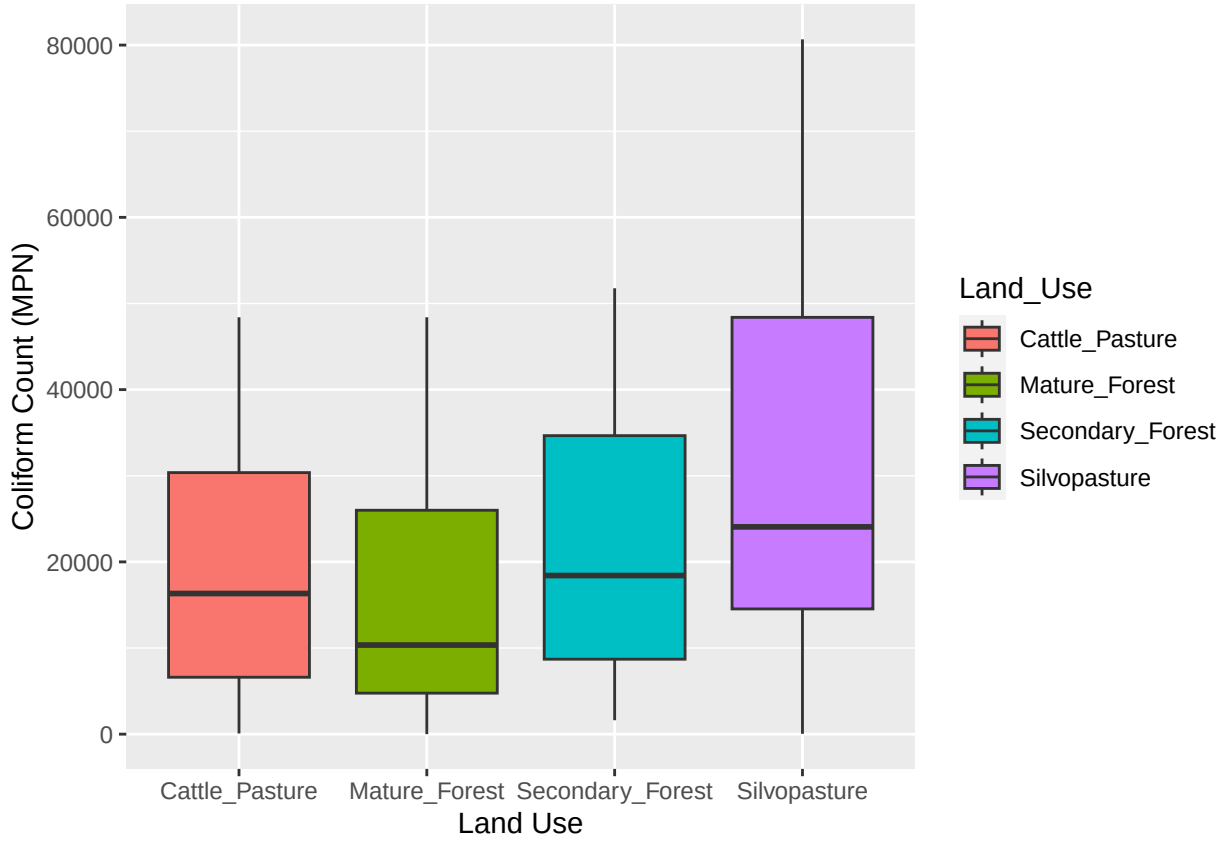
Silvopasture: ~22,000 MPN

There is a far clearer difference between sites for mean coliform count when compared to pH values. The only issue I'm facing in estimating a mean is how spread out points are on my histograms. Though data is concentrated in one area on some plots (e.g. Silvopasture), it appears much more randomly distributed on others (e.g. Cattle Pasture). All sites other than Silvopasture have a large spike in values around ~48,000 MPN, which is quite far from the next greatest concentration of values. If this were to appear differently, my mean estimations would be far lower in sites where these spikes were observed.

From these histograms, I'm predicting that coliform counts of Cattle Pastures, Secondary Forests, and Silvopastures will not significantly differ. However, I assume that Mature Forest **will** significantly differ from other sites.

As this data contains a categorical variable (Land_Use), a boxplot should suffice to view relationships with coliform count.

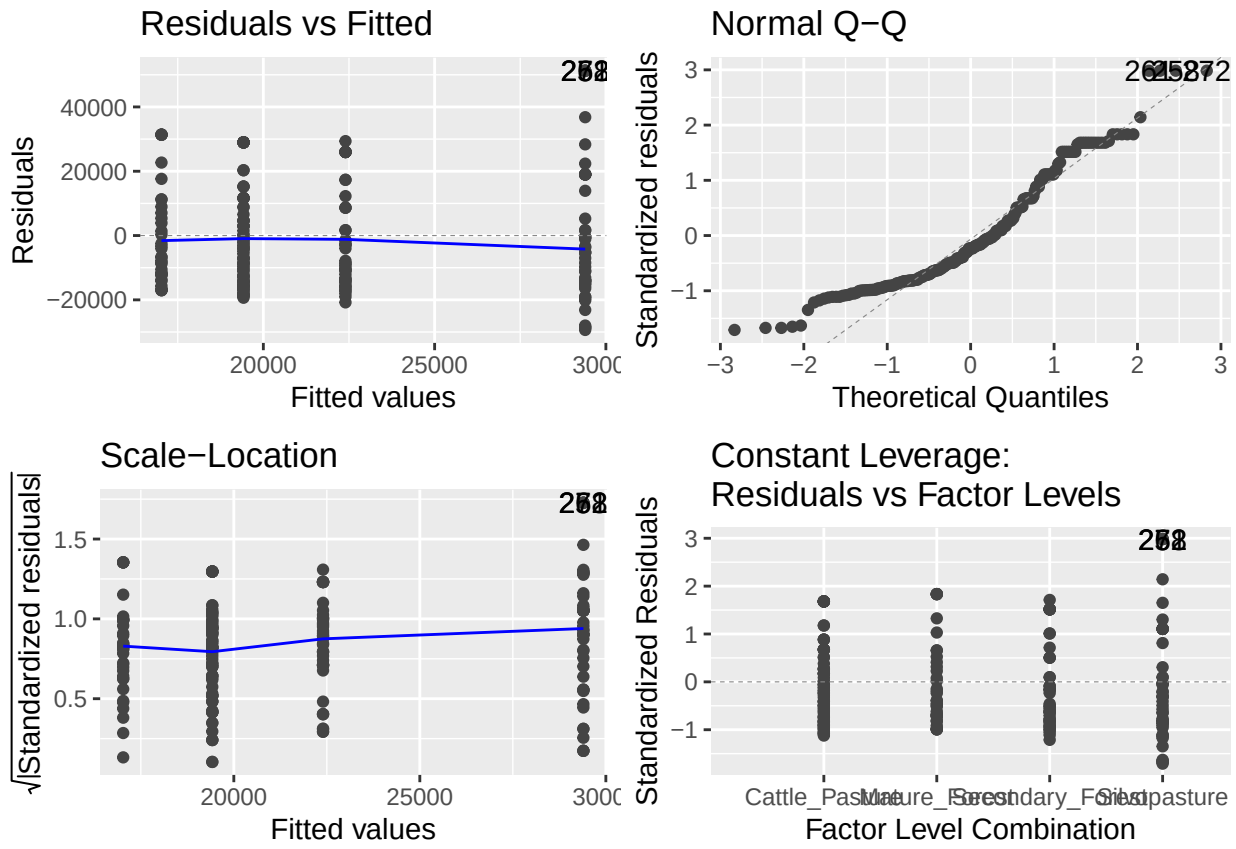
```
ggplot(Bacteria_Data3, aes(x = Land_Use, y = Coliform_MPN, fill=Land_Use)) +
  geom_boxplot() +
  xlab("Land Use") +
  ylab("Coliform Count (MPN)")
```



Data seems much more differentiated here without any major overlap. I predict that there is a significant difference in coliform count between sample sites (particularly Silvopasture) based upon this data visualization.

Now, to set up a linear model for autoplot and one-way Anova testing.

```
Land_Coliform <- lm(Coliform_MPN ~ Land_Use, data=Bacteria_Data3)

autoplot(Land_Coliform)
```



Theoretical quantile points deviate towards the beginning, like pH and dissolved O2. As I had done with other such variables, I will continue with a standard linear model for now.

```
anova(Land_Coliform)
```

```
## Analysis of Variance Table
##
## Response: Coliform_MPN
##            Df     Sum Sq    Mean Sq F value   Pr(>F)
## Land_Use    3 4.2576e+09 1419196073  4.7155 0.003305 **
## Residuals 211 6.3503e+10  300963849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(Land_Coliform)
```

```
##
## Call:
## lm(formula = Coliform_MPN ~ Land_Use, data = Bacteria_Data3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29356 -13819  -4019  11641  51270
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  19421       1916  10.137  < 2e-16 ***
## Land_UseMature_Forest        -2395       3436  -0.697  0.48659
```

```
## Land_UseSecondary_Forest     2977       3218   0.925  0.35602
## Land_UseSilvopasture         9975       3094   3.224  0.00146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17350 on 211 degrees of freedom
## Multiple R-squared:  0.06283,    Adjusted R-squared:  0.04951
## F-statistic: 4.716 on 3 and 211 DF,  p-value: 0.003305
```

The one-way Anova test gives us statistical significance of coliform count from sites in relation to Cattle_Pasture (as this comes first alphabetically). While Mature Forest and Secondary Forest do not differ significantly from Cattle Pasture, Silvopasture does (p-value=0.00146). In order to show relationships between all sites relative to one another, a Tukey test should be performed. From there, I can create a box-plot with letters relative to significance.

```
library(multcompView)

Bacteria_Data3aov <- aov(Coliform_MPN ~ Land_Use, data=Bacteria_Data3)

## Anova test performed on solely Land_Use. This tells us that coliform count differs significantly bet

summary(Bacteria_Data3aov)
```

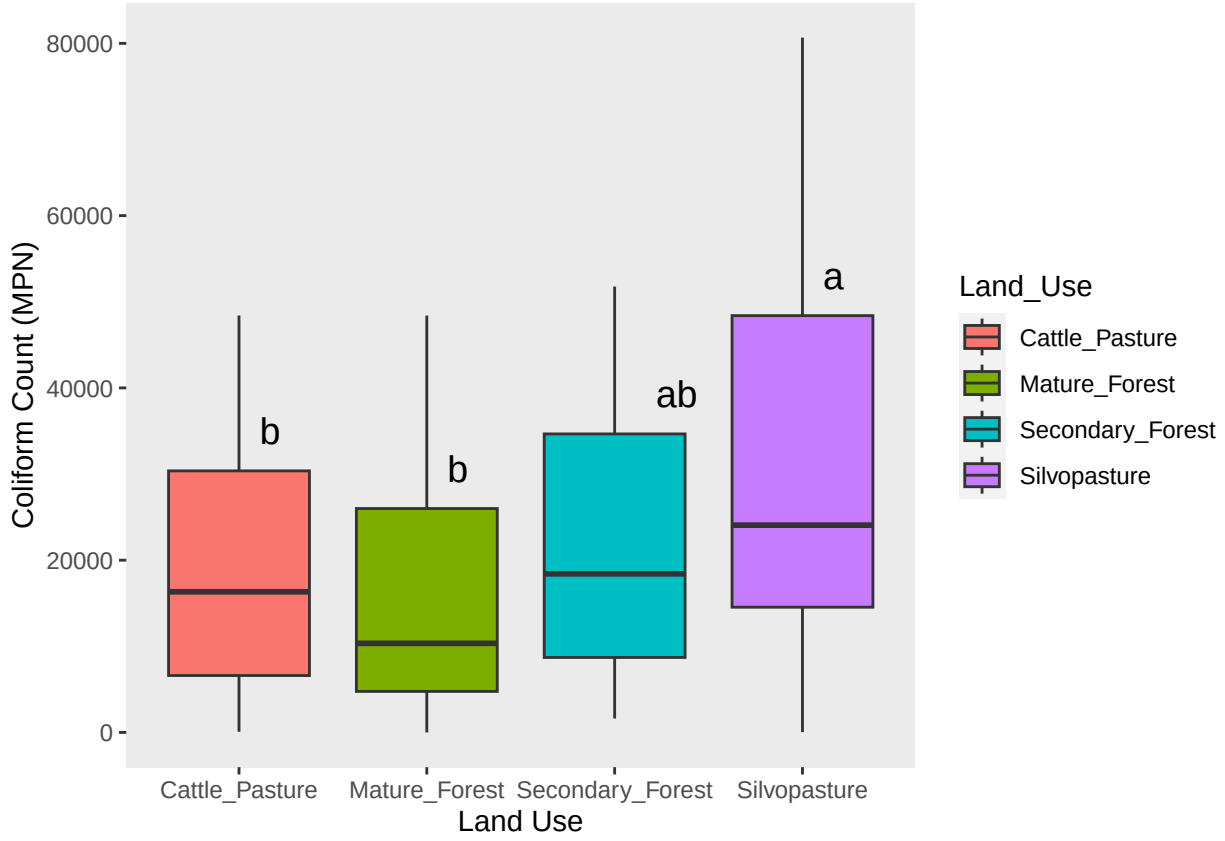
```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Land_Use      3 4.258e+09 1.419e+09   4.716 0.0033 **
## Residuals   211 6.350e+10 3.010e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Bacteria_Data3Tukey <- TukeyHSD(Bacteria_Data3aov, conf.level=.95)

##Tukey test performed--set to a variable.

Bacteria_Data3Tk <- group_by(Bacteria_Data3, Land_Use) %>%
  summarise(mean=mean(Coliform_MPN), quant = quantile(Coliform_MPN, probs = 0.75)) %>%
  arrange(desc(mean))

##Statistics regarding mean and quartiles made into a table, which is then used to create letters on ou

Bacteria_Data3cld <- multcompLetters4(Bacteria_Data3aov, Bacteria_Data3Tukey)
Bacteria_Data3cld <- as.data.frame.list(Bacteria_Data3cld$Land_Use)
Bacteria_Data3Tk$Bacteria_Data3cld <- Bacteria_Data3cld$Letters

## Letters assigned to different levels of Land_Use on table (A, B, AB).

print(Bacteria_Data3cld)
```

```
##                 Letters monospacedLetters LetterMatrix.a LetterMatrix.b
## Silvopasture          a                a            TRUE          FALSE
## Secondary_Forest     ab               ab            TRUE           TRUE
## Cattle_Pasture        b                b           FALSE           TRUE
## Mature_Forest         b                b           FALSE           TRUE
```

```
ggplot(Bacteria_Data3, aes(Land_Use, Coliform_MPN)) +
  geom_boxplot(aes(fill=Land_Use)) +
  xlab("Land Use") +
```

```
  ylab("Coliform Count (MPN)") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
   geom_text(data = Bacteria_Data3Tk, aes(x = Land_Use, y = quant, label = Bacteria_Data3cld), size = 5
```




```
##Box-plot created and formatted. Letters must be raised and increased in size once set on plot. Label
```

From this plot, we can observe several correlations between sites. First, Cattle Pasture, Mature Forest, and Secondary Forest do not significantly differ from one another in terms of coliform count. Additionally, Secondary Forest and Silvopasture do not significantly differ either. However, this chart shows that coliform counts from Cattle Pasture and Mature Forest **do** differ significantly from those of Silvopasture.

Test #4: Multiple Regression Model to Demonstrate the Effects of Several Variables on Coliform Abundance

```
Bacteria_Data4 <- Bacteria_Data[c(2, 8, 10:11, 13:14)]

Bacteria_Data4 <- na.omit(Bacteria_Data4)
```

To begin, a linear model is created by combining multiple x-variables together as a factor of Coliform_MPN. Summary() is then used to demonstrate the properties associated with this model.

```
Bacteria_Data4aov <- aov(Coliform_MPN ~ pH + Temperature + DissolvedO2 +Conductivity + Land_Use, data=Ba

summary(Bacteria_Data4aov)

##              Df    Sum Sq   Mean Sq F value  Pr(>F)
```

```
## pH             1 6.146e+07 6.146e+07   0.202 0.65357
## Temperature    1 7.596e+07 7.596e+07   0.250 0.61784
## DissolvedO2    1 5.564e+06 5.564e+06   0.018 0.89256
## Conductivity   1 2.571e+08 2.571e+08   0.845 0.35903
## Land_Use       3 4.378e+09 1.459e+09   4.797 0.00298 **
## Residuals    207 6.298e+10 3.043e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This multiple-regression model proves that from data included in this test, the only significant variable in predicting coliform count is sample site/Land_Use. This refutes my previous hypothesis that pH or temperature played a role in coliform abundance. The test shows that factors besides pH, temperature, dissolved O2, or conductivity must contribute to the prevalence of coliforms in certain locations. I shouldn't need a figure to depict this information, since all that I need is provided through this multi-variable analysis.

Biological Significance

The first component of my data set I sought to analyze was the correlation between water temperature and dissolved oxygen. I predicted that dissolved oxygen content would decrease in samples taken under higher temperatures. My simple linear regression test proved there to be a significant relationship between the two variables in this data set (p=0.0208). A positive trend was observed when plotting a linear model, refuting my initial hypothesis and demonstrating a slight increase in dissolved O2 in warmer temperatures. However, an r-squared value of 0.0204 indicates that while significant, temperature is a very weak predictor of dissolved oxygen content. It is highly probable that more significant variables are responsible for dissolved oxygen content in these samples.

Next, pH values of samples from all sites were observed. Knowing that coliforms tend to grow within a specific pH range, I was curious to see how pH may differ between sites. Upon comparing histograms of pH and creating a simple boxplot, I ended up with very similar-appearing average values for each site with a high degree of overlap at ~6.5-7. From this, I predicted that there would not be significant variation in pH between sample sites. Using a one-way Anova test, it was determined that sample site *was* indeed a significant indicator of pH (p=0.001322). I was also interested to visualize how **individual** sites differed in relation to one another. Again, I predicted that there would not be a significant relationship between any sites and pH, as there was very little difference between locations when estimating a mean. A Tukey test was performed for this purpose, demonstrating that Silvopastures differed significantly from both Cattle Pastures and Mature Forests in their pH values.

To measure variation in coliform abundance between sites, another one-way Anova test was utilized. I predicted that coliform count would be significantly different between sites due to varying environmental conditions (such as pH, which *was* proven to differ between sample locations). The estimated mean coliform count by location was far more discernible than means of pH values. Ranges of data were differently distributed, and certain means were clearly lower than others. Performing my one-way Anova test supported this hypothesis (p=0.003305), demonstrating that coliform count was dependent on sample location. Another Tukey test was then performed to visualize relationships between individual sites. From my histograms, I predicted that Mature Forests' coliform abundance would differ from those seen in all other locations. However, my test proved that coliform abundance differed significantly between Silvopastures, and both Cattle Pastures and Mature Forests.

In my final test, I looked to determine the degree to which several variables influenced coliform abundance in all samples. For this, I needed to use a multiple regression model. It was demonstrated that of sample site, pH, temperature, dissolved O2, and water conductivity, sample site was the only significant predictor of recorded coliform count (p-value=0.00298). This also refuted my initial hypothesis that factors such as temperature or pH would impact coliform abundance. My conclusion is that variables aside from those tested are responsible for different coliform counts between sites.

Challenges Faced

This project required me to greatly branch out from material we learned in class. I needed to do independent research regarding both statistical tests and formatting of plots. First, I went back to get a clearer understanding of test determination. With this, I became more able to select statistical tests based upon the data types associated with them. Another challenge I faced was learning how to carry out new tests we haven't covered in class. It took many tries to get down Tukey and aov, but I eventually understood them upon consulting various resources. This helped me to feel more confident in interpreting code which others have produced, something which is important in further navigating R.