NOx_models_NG

2025-04-02

Libraries and Data Loading

```r
library(here)
```

```
## here() starts at /Users/nicolegutkowski/Desktop/ENV710/ENV710-Group-Project-3B
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(ggplot2)
library(moments)
library(readxl)
library(lubridate)


wetlands <- read.csv(here('Data/Processed/Log_Transformed_Data.csv'))

glimpse(wetlands)
```

```
## Rows: 1,618
## Columns: 20
## $ Date                             <chr> "2015-08-25", "2015-09-15", "2015-09-~
## $ Site                             <chr> "WT3", "WT5", "WT3", "WT4", "WT1", "A~
## $ Temp_C                           <dbl> 23.21, 20.14, 18.13, 18.69, 18.41, 18~
## $ Log_SpCond_mScm                  <dbl> 0.16636154, 0.22713557, 0.09712671, 0~
## $ Log_Cond_mScm                    <dbl> 0.16126815, 0.20863887, 0.08434115, 0~
## $ Log_TDS_mgl                      <dbl> 0.11154137, 0.15357909, 0.06391333, 0~
## $ Log_Sal_ppt                      <dbl> 0.08617770, 0.11332869, 0.04879016, 0~
## $ DO_percent                       <dbl> 30.7, 72.5, 56.6, 62.4, 72.3, 18.5, 7~
## $ DO_mgL                           <dbl> 2.61, 6.56, 5.32, 5.80, 6.80, 1.72, 7~
## $ pH                               <dbl> 6.90, 6.71, 6.59, 6.55, 6.66, 6.71, 6~
## $ Month                            <int> 8, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 1~
## $ Year                             <int> 2015, 2015, 2015, 2015, 2015, 2015, 2~
## $ Season                           <chr> "Summer", "Fall", "Fall", "Fall", "Fa~
## $ Log_Unfiltered_TN_ugL            <dbl> 7.319865, 6.752270, 6.648985, 6.89669~
## $ Log_Filtered_NOx_ugL             <dbl> 5.420535, 4.727388, 3.761200, 5.33271~
## $ Log_Filtered_NHx_ugL             <dbl> 5.817111, 4.859812, 4.934474, 4.82831~
## $ Log_Unfiltered_TP_ugL            <dbl> 4.465908, 4.442651, 4.262680, 4.65396~
## $ Filtered_OP_ugL                  <dbl> 28, 10, 4, 14, 29, 16, 8, 22, 3, 10, ~
## $ Log_TSS_mgL                      <dbl> 3.7135721, 2.5649494, 1.9459101, 2.48~
## $ fecal_coliform_colonies_per100mL <dbl> 65, 5, 90, 30, 0, 5, 20, 45, 30, 60, ~
```

```
colnames(wetlands)
```

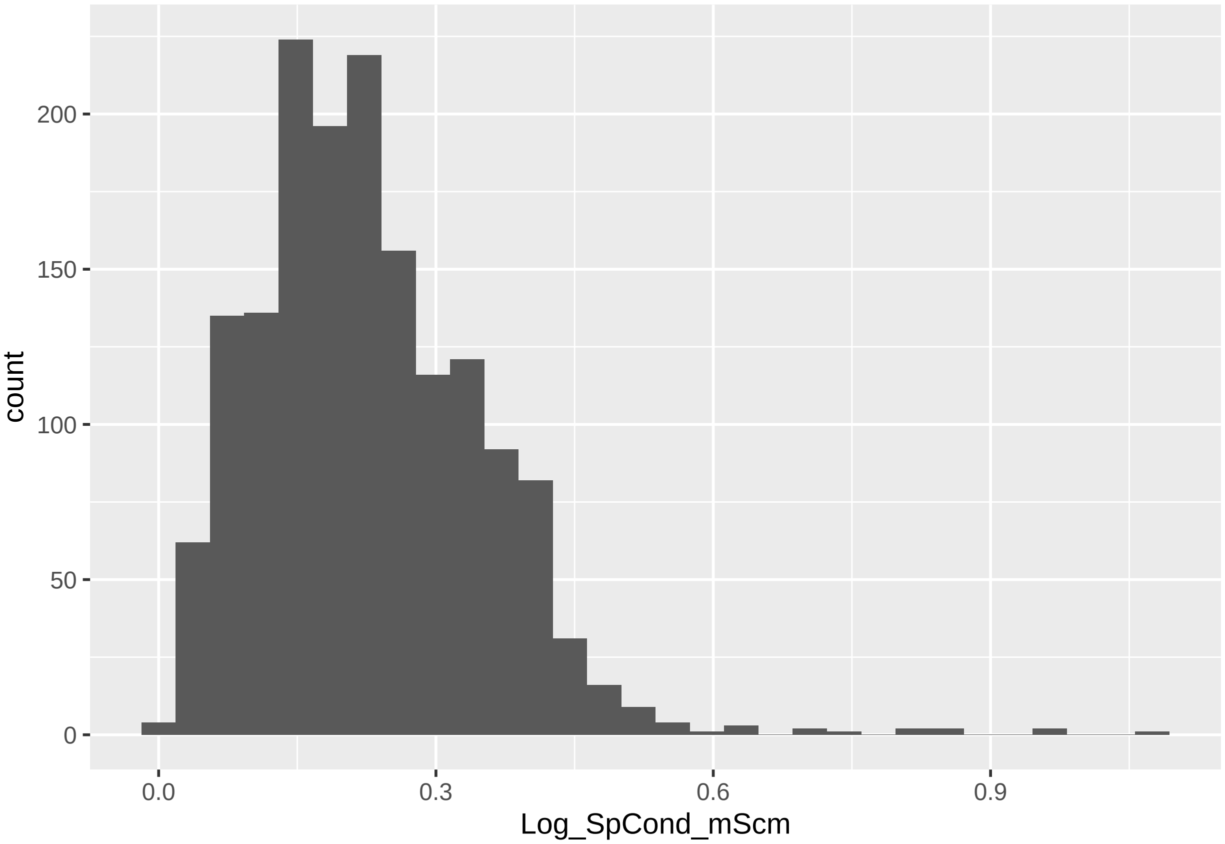
```
##  [1] "Date"                             "Site"                            
##  [3] "Temp_C"                           "Log_SpCond_mScm"                 
##  [5] "Log_Cond_mScm"                    "Log_TDS_mgl"                     
##  [7] "Log_Sal_ppt"                      "DO_percent"                      
##  [9] "DO_mgL"                           "pH"                              
## [11] "Month"                            "Year"                            
## [13] "Season"                           "Log_Unfiltered_TN_ugL"           
## [15] "Log_Filtered_NOx_ugL"             "Log_Filtered_NHx_ugL"            
## [17] "Log_Unfiltered_TP_ugL"            "Filtered_OP_ugL"                 
## [19] "Log_TSS_mgL"                      "fecal_coliform_colonies_per100mL"
```

Data Exploration: Continuous Variables Histograms

```
ggplot(wetlands, aes(x = Log_SpCond_mScm)) +
geom_histogram()
```

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

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_bin()‘).
```



```
skewness(wetlands$Log_SpCond_mScm, na.rm = T)
```

```
## [1] 1.207848
```

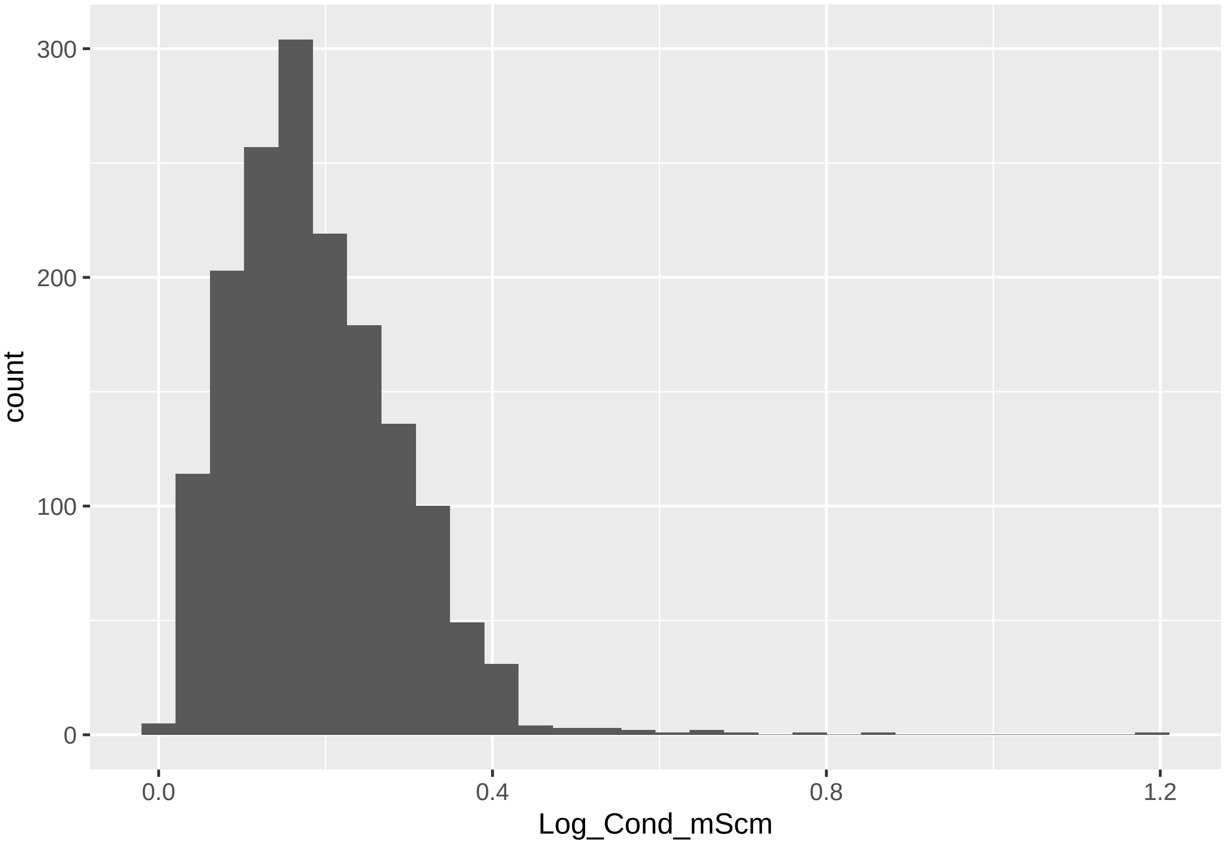
```
kurtosis(wetlands$Log_SpCond_mScm, na.rm = T)
```

```
## [1] 7.072255
```

```
ggplot(wetlands, aes(x = Log_Cond_mScm)) +
geom_histogram()
```

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

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
skewness(wetlands$Log_Cond_mScm, na.rm = T)
```

```
## [1] 1.529835
```

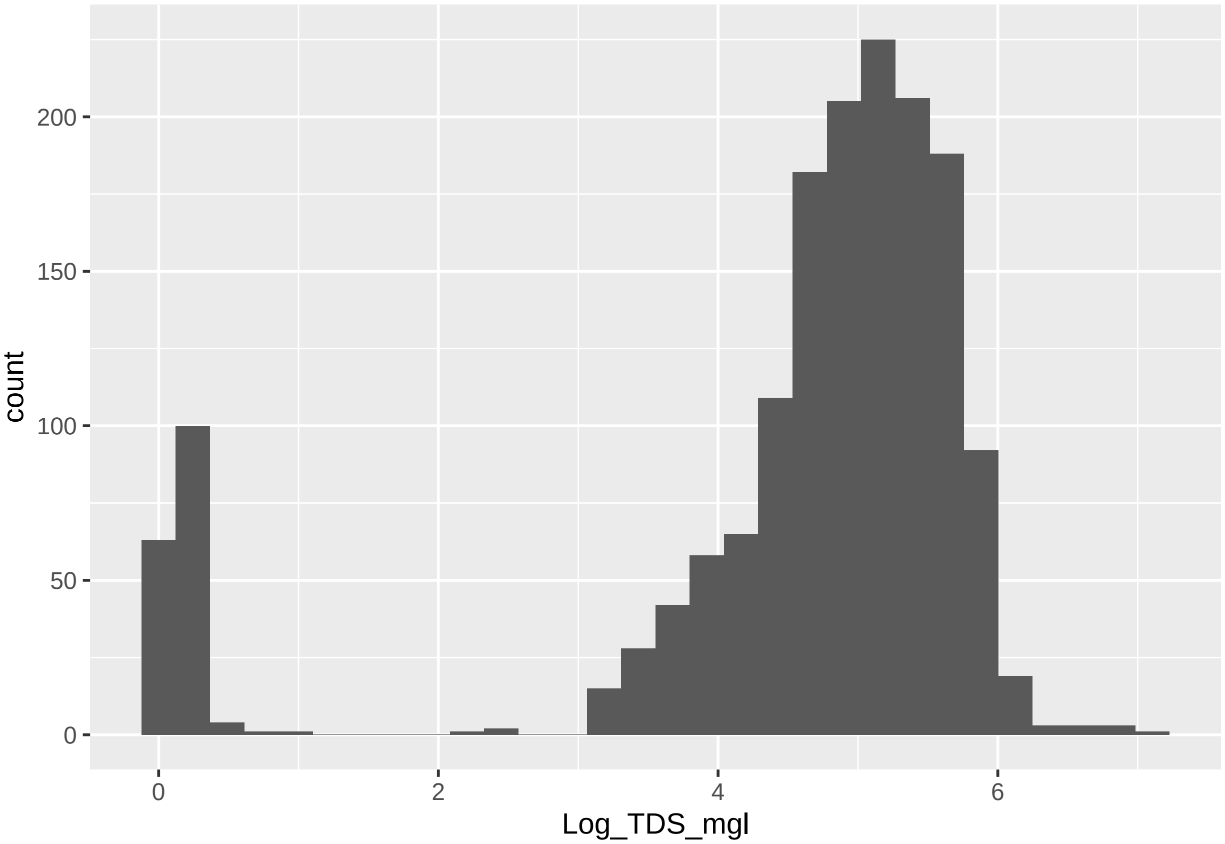
```
kurtosis(wetlands$Log_Cond_mScm, na.rm = T)
```

```
## [1] 10.75839
```

```
ggplot(wetlands, aes(x = Log_TDS_mgl)) +
geom_histogram()
```

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

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



```
skewness(wetlands$Log_TDS_mgl, na.rm = T)
```

```
## [1] -1.922859
```

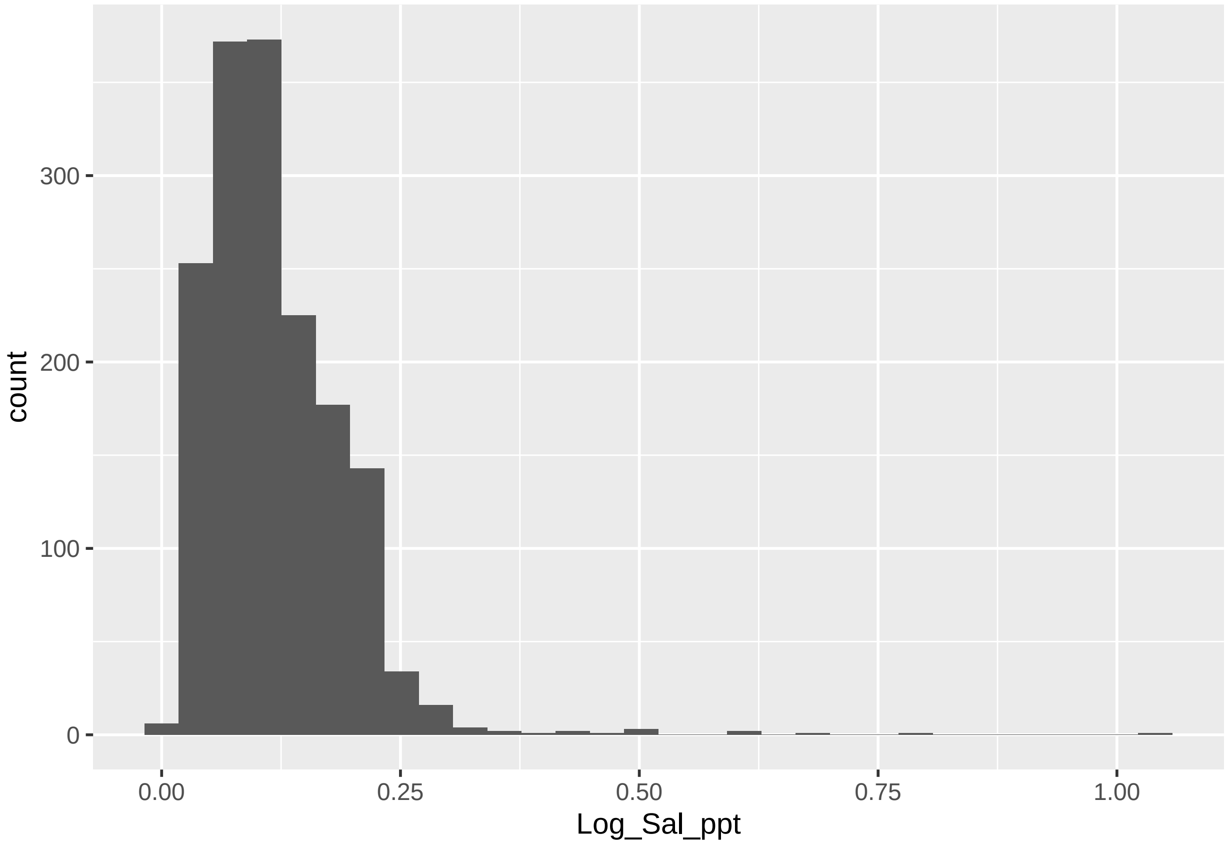
```
kurtosis(wetlands$Log_TDS_mgl, na.rm = T)
```

```
## [1] 5.699103
```

```
ggplot(wetlands, aes(x = Log_Sal_ppt)) +
geom_histogram()
```

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

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_bin()‘).
```



```
skewness(wetlands$Log_Sal_ppt, na.rm = T)
```

```
## [1] 2.898492
```

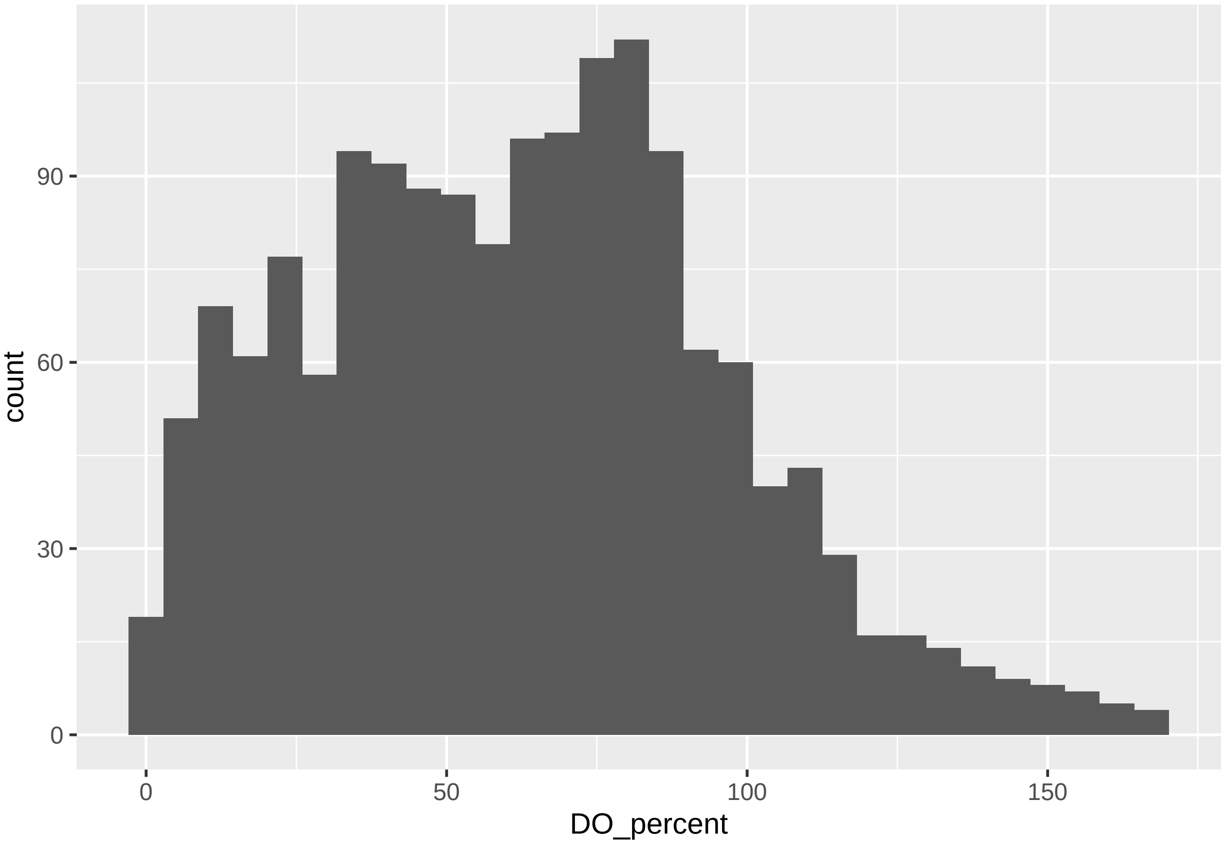
```
kurtosis(wetlands$Log_Sal_ppt, na.rm = T)
```

```
## [1] 25.82352
```

```
ggplot(wetlands, aes(x = DO_percent)) +
geom_histogram()
```

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

```
## Warning: Removed 11 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



```
skewness(wetlands$DO_percent, na.rm = T)
```

```
## [1] 0.348897
```

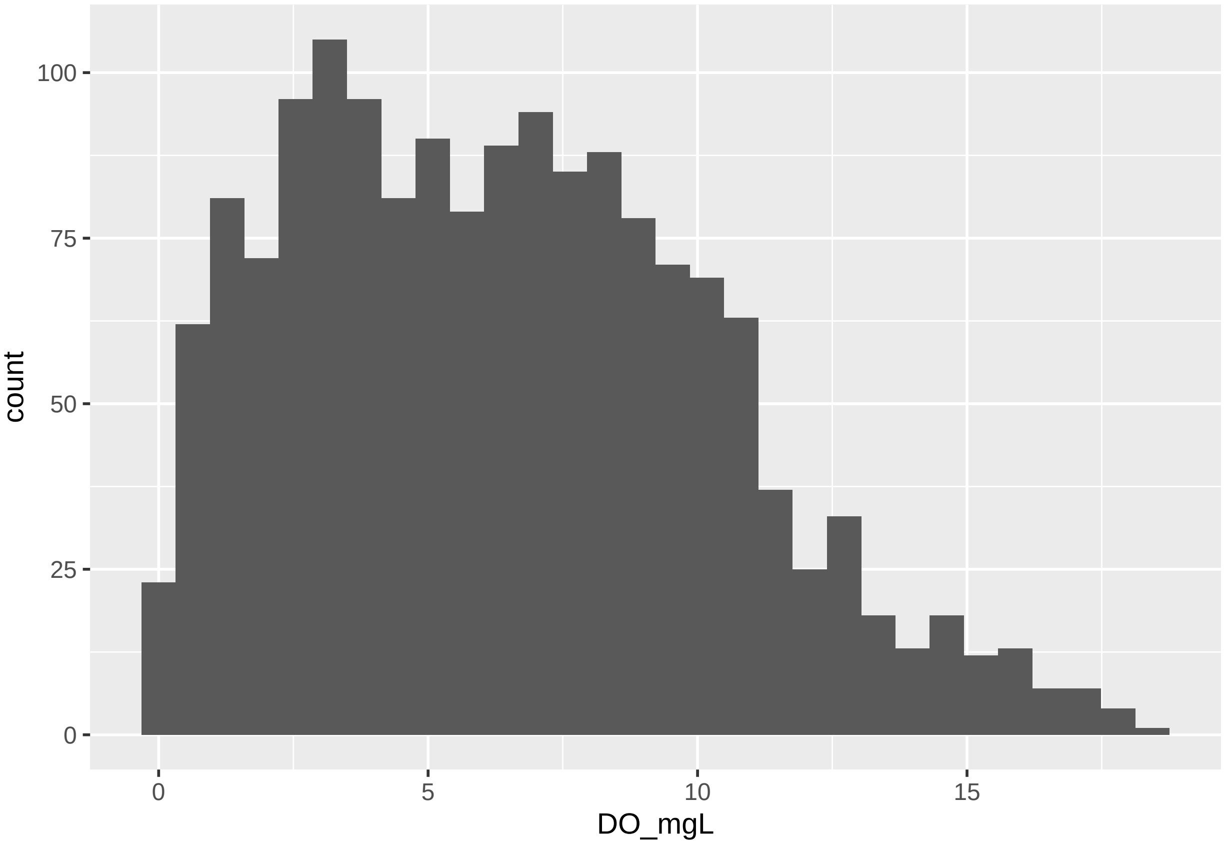
```
kurtosis(wetlands$DO_percent, na.rm = T)
```

```
## [1] 2.748728
```

```
ggplot(wetlands, aes(x = DO_mgL)) +
geom_histogram()
```

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

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



```
skewness(wetlands$DO_mgL, na.rm = T)
```

```
## [1] 0.4772204
```

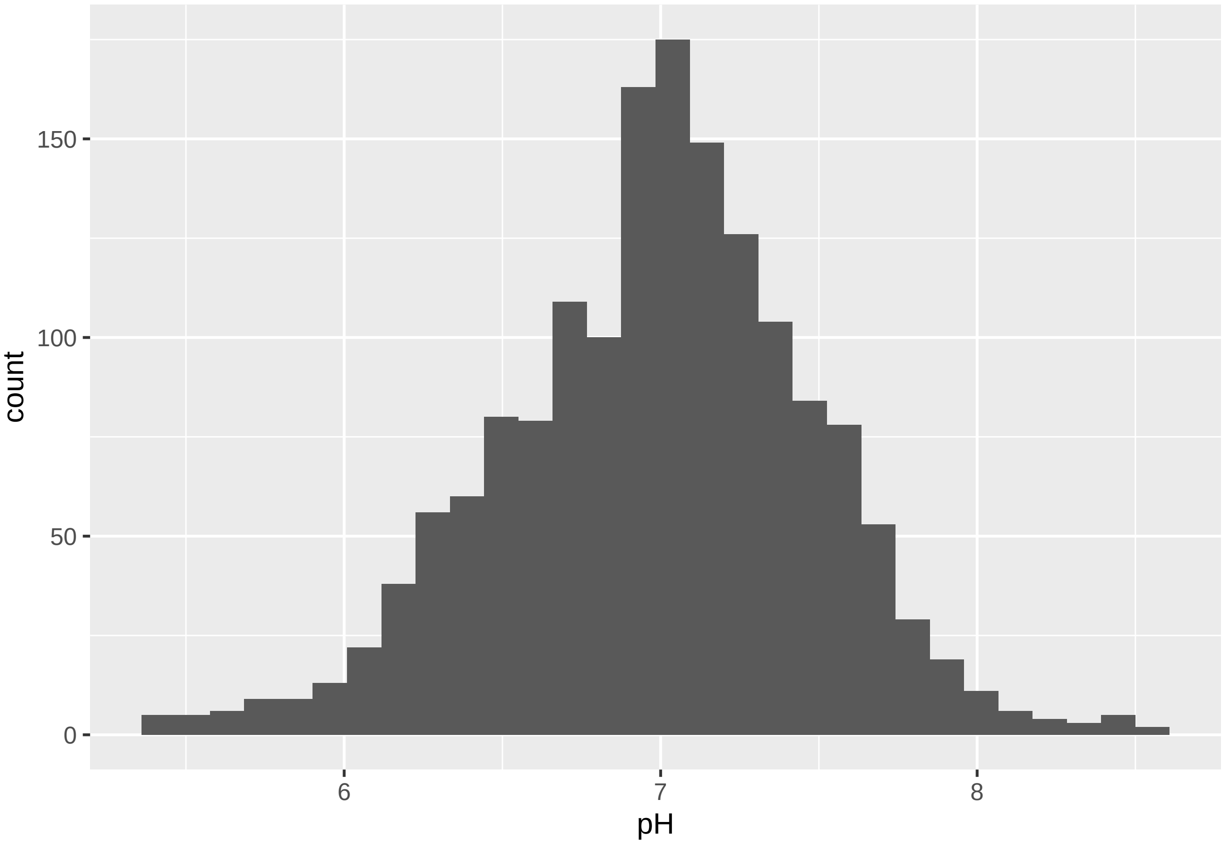
```
kurtosis(wetlands$DO_mgL, na.rm = T)
```

```
## [1] 2.665615
```

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

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

```
## Warning: Removed 16 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
skewness(wetlands$pH, na.rm = T)
```

```
## [1] -0.2046608
```

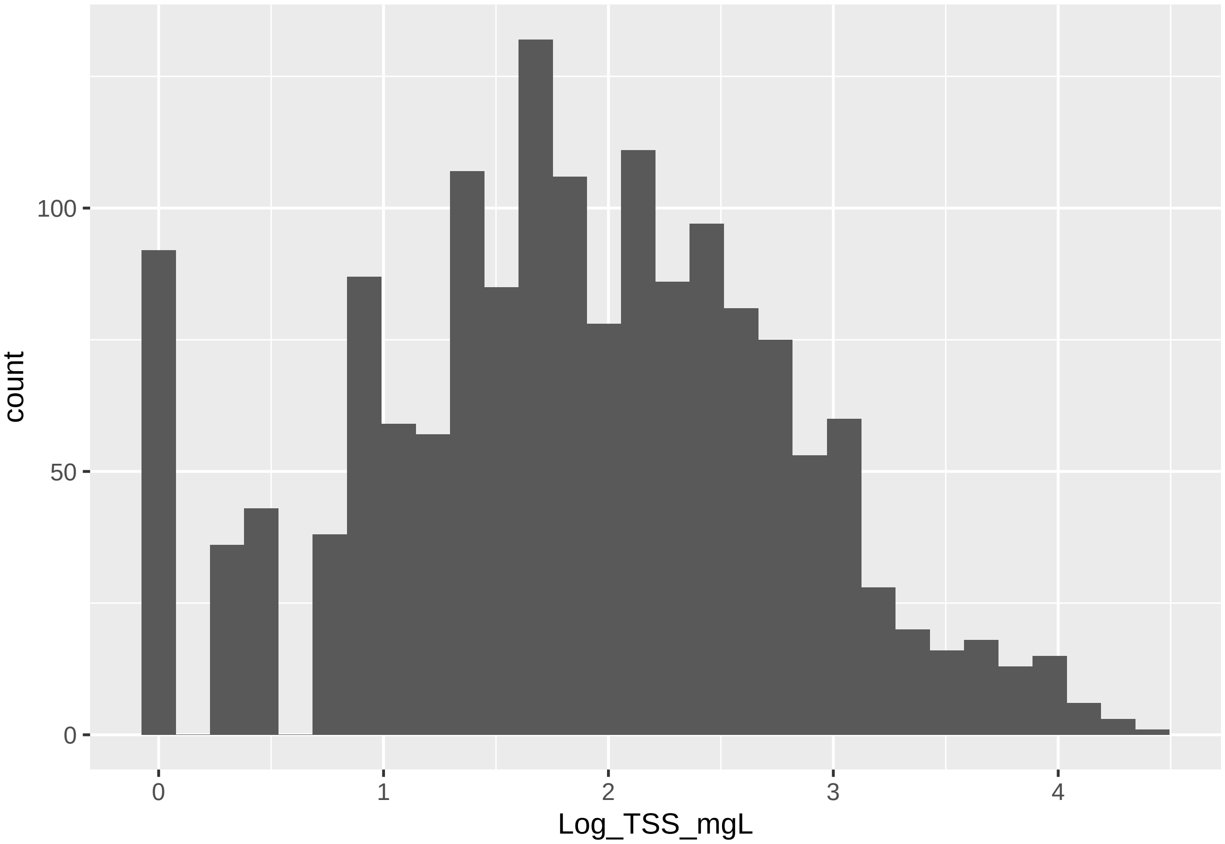
```
kurtosis(wetlands$pH, na.rm = T)
```

```
## [1] 3.325108
```

```
ggplot(wetlands, aes(x = Log_TSS_mgL)) +
geom_histogram()
```

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

```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (‘stat_bin()‘).
```



```
skewness(wetlands$Log_TSS_mgL, na.rm = T)
```

```
## [1] -0.03686939
```

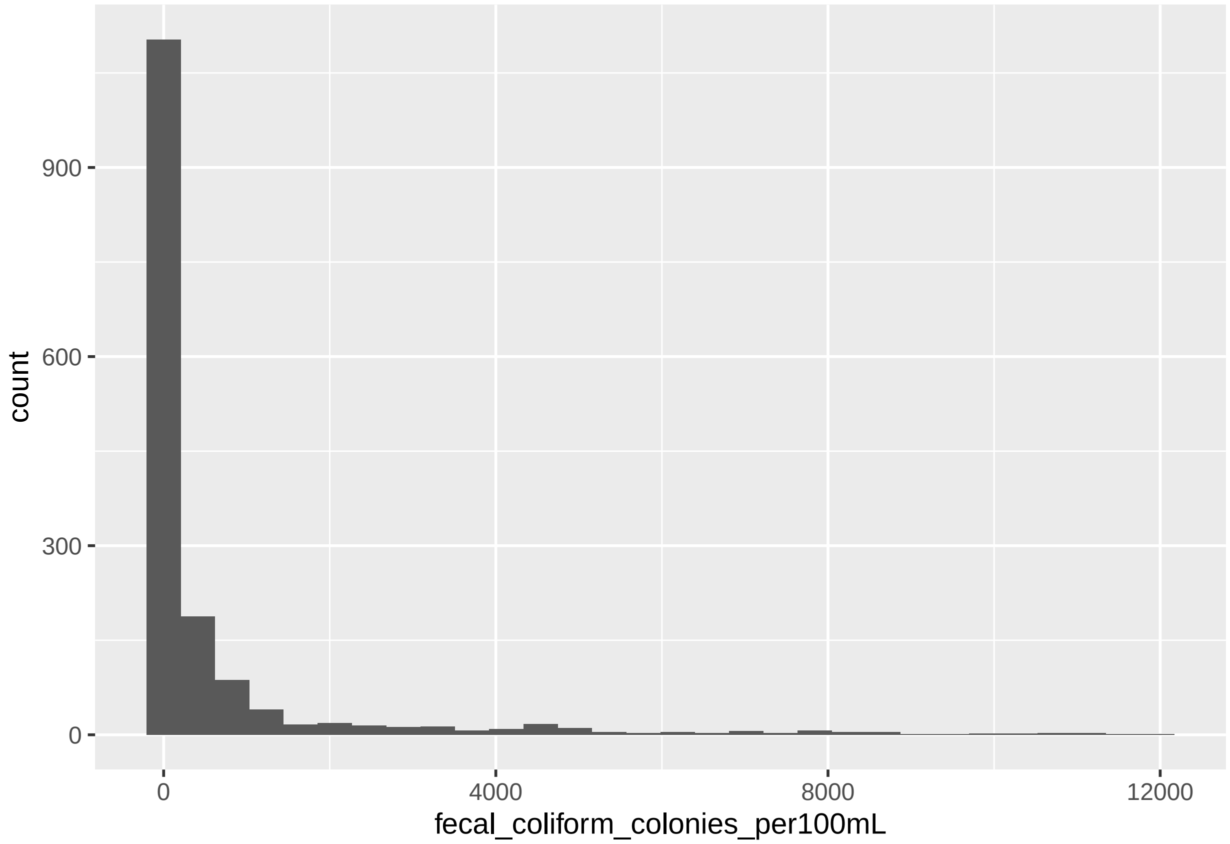
```
kurtosis(wetlands$Log_TSS_mgL, na.rm = T)
```

```
## [1] 2.642163
```

```
ggplot(wetlands, aes(x = fecal_coliform_colonies_per100mL)) +
geom_histogram()
```

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

```
## Warning: Removed 29 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
skewness(wetlands$fecal_coliform_colonies_per100mL, na.rm = T)
```

```
## [1] 3.881634
```

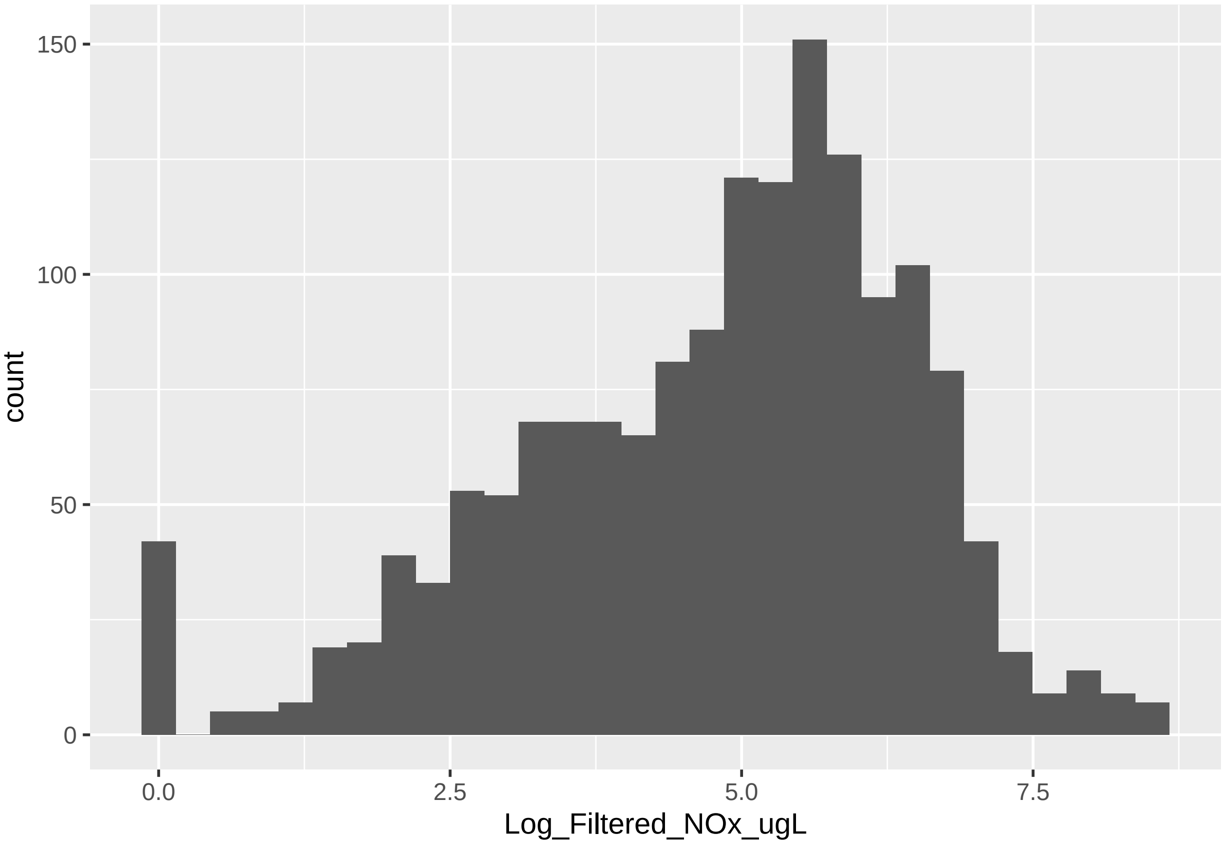
```
kurtosis(wetlands$fecal_coliform_colonies_per100mL, na.rm = T)
```

```
## [1] 19.40906
```

```
ggplot(wetlands, aes(x = Log_Filtered_NOx_ugL)) +
geom_histogram()
```

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

```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
skewness(wetlands$Log_Filtered_NOx_ugL, na.rm = T)
```

```
## [1] -0.644908
```

```
kurtosis(wetlands$Log_Filtered_NOx_ugL, na.rm = T)
```

```
## [1] 3.200968
```

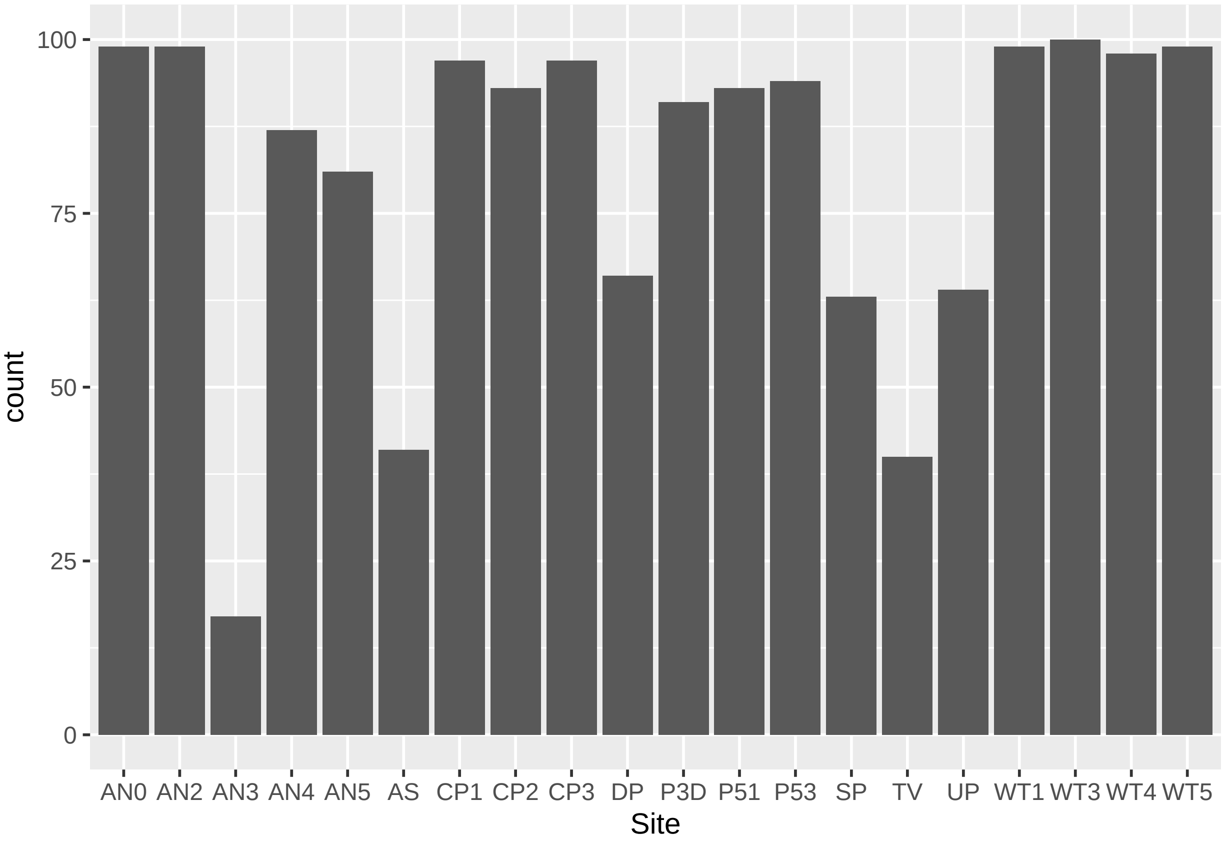
From the histogram plots, and skewness and kurtosis values, the variables that are normally distributed enough for analysis are DO%, DO (mg/L), pH, and log(TSS). Assessments of categorical variables will follow:

Data Exploration: Categorical Variables Bar Charts

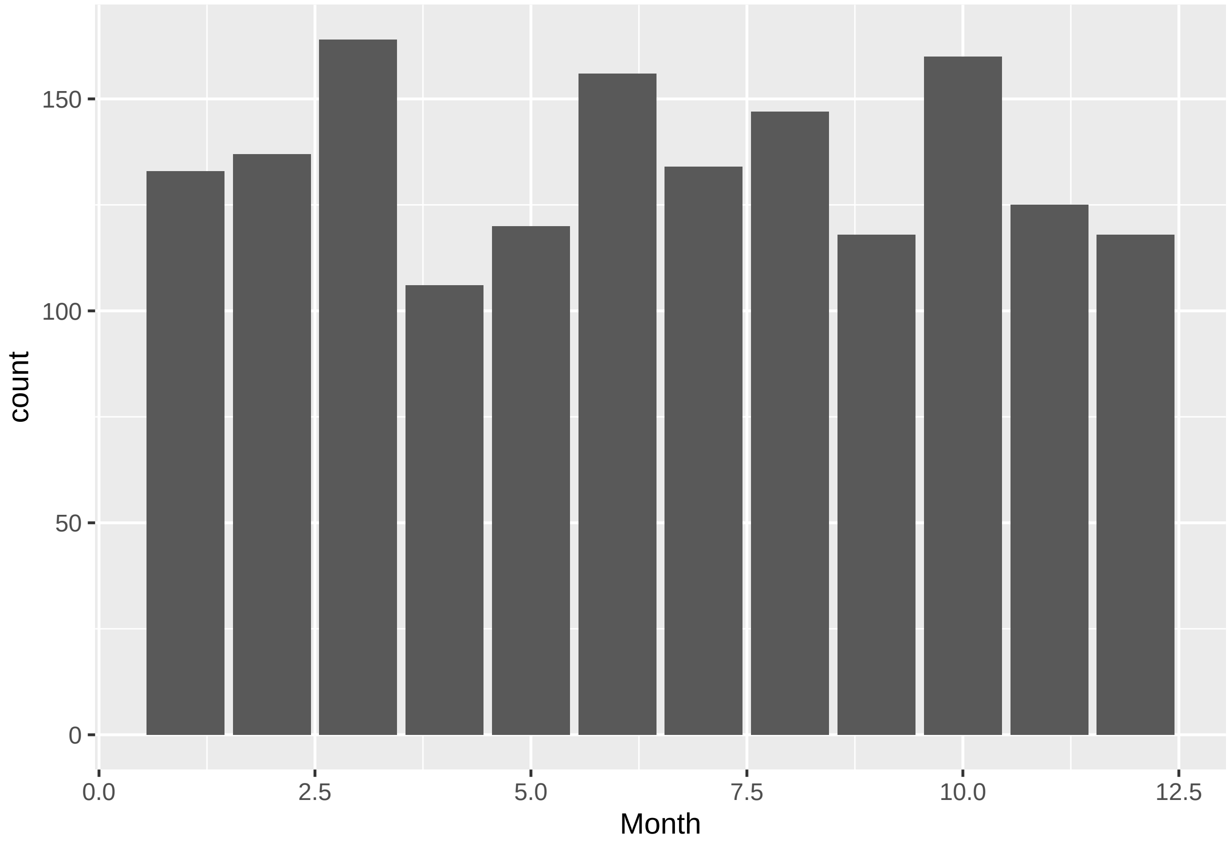
```
unique(wetlands$Site)
```

```
##  [1] "WT3" "WT5" "WT4" "WT1" "AN2" "AN5" "AN0" "P53" "P51" "CP3" "CP2" "AN3"
## [13] "CP1" "AN4" "P3D" "DP"  "UP"  "SP"  "TV"  "AS"
```

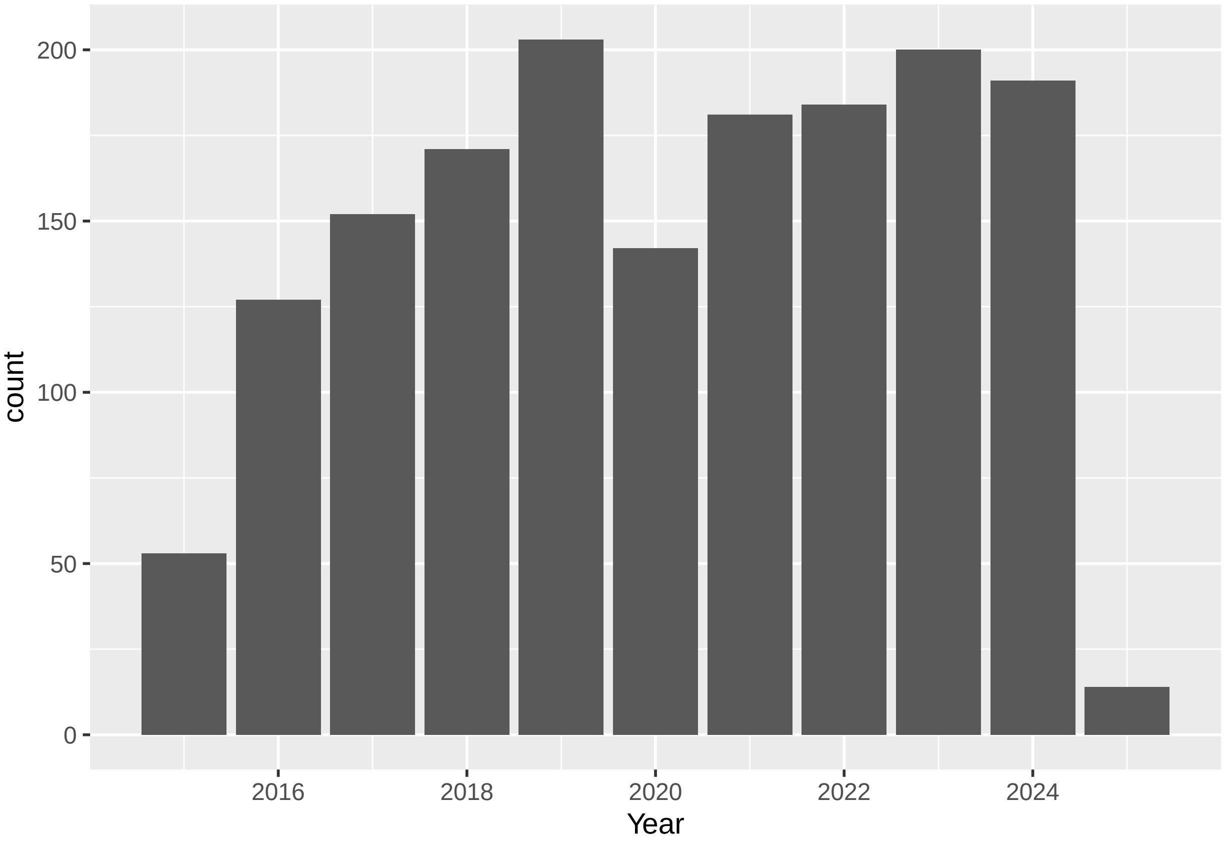
```
ggplot(wetlands, aes(Site))+
  geom_bar()
```



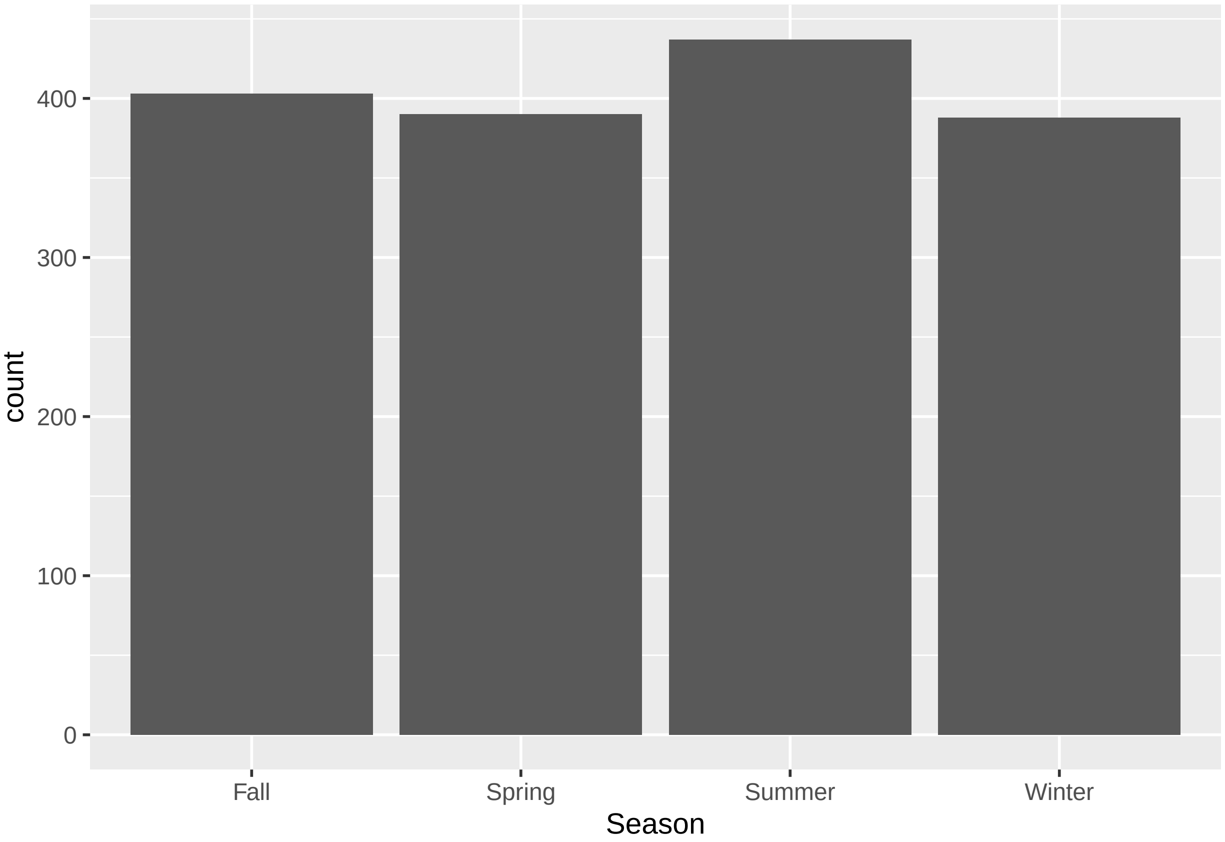
```
ggplot(wetlands, aes(Month))+
  geom_bar()
```



```
ggplot(wetlands, aes(Year))+
  geom_bar()
```



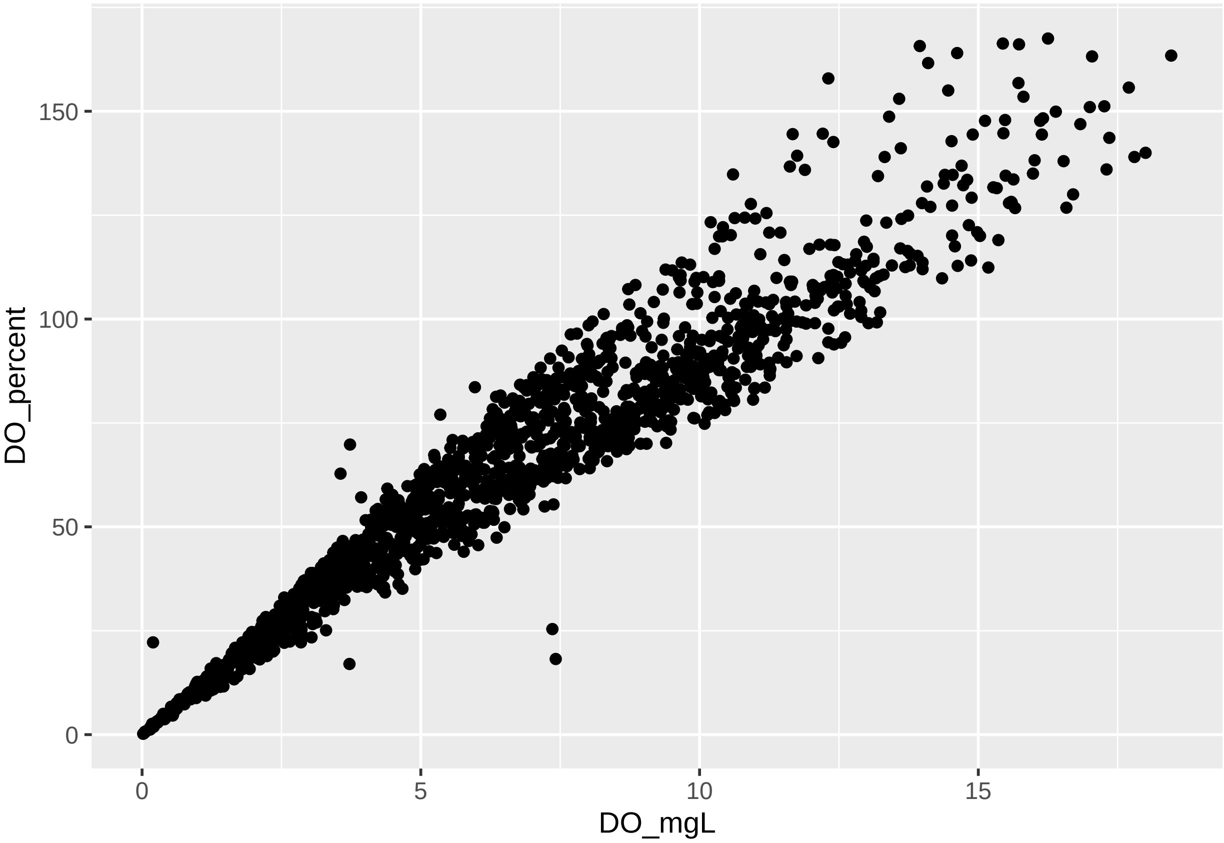
```
ggplot(wetlands, aes(Season))+
  geom_bar()
```



Data Exploration: Scatter and Box Plots

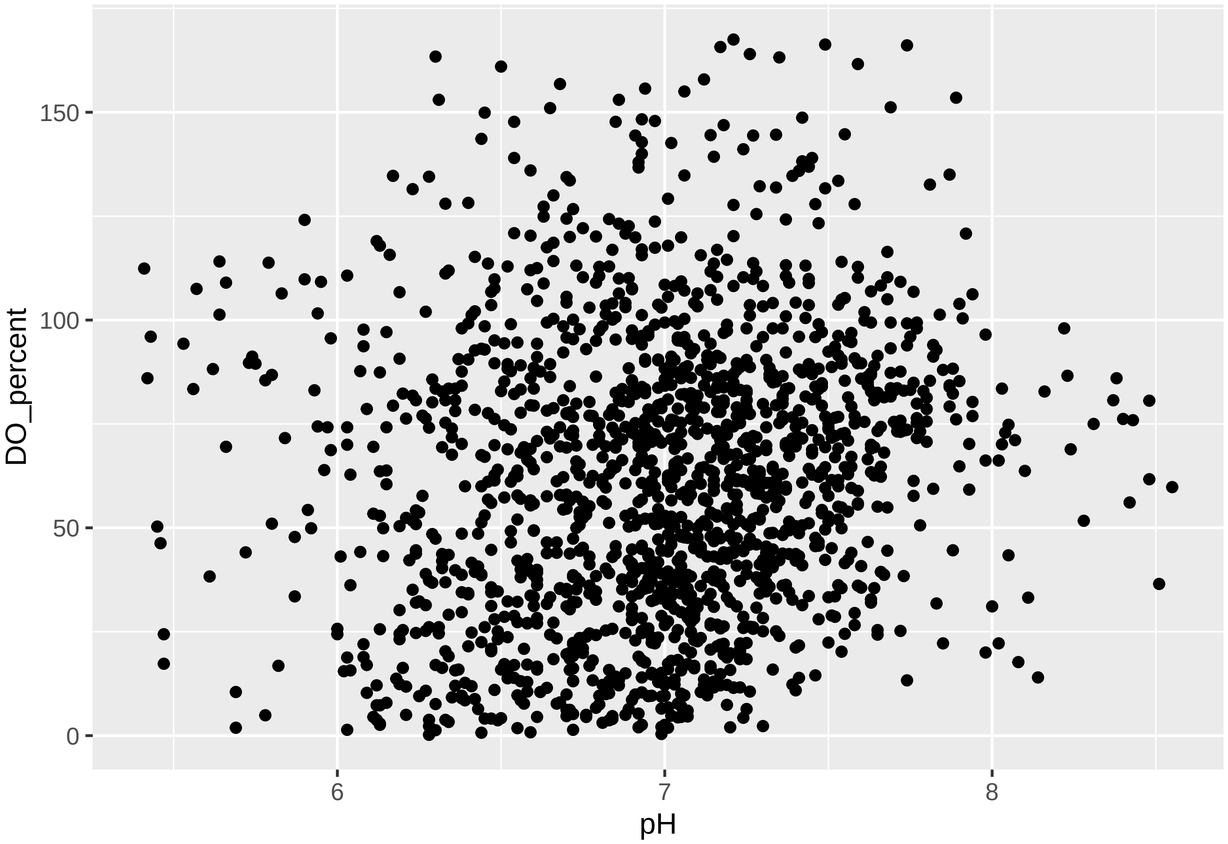
```
#Continuous Variables:
#- DO% / DO mg_L/ pH / LogTSS
#Categorical Variables:
#- Site/ month/ Year/ season

DO_types_plot <- ggplot(wetlands, aes(x = DO_mgL, y = DO_percent)) +
  geom_point()
DO_types_plot
```

```
## Warning: Removed 14 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



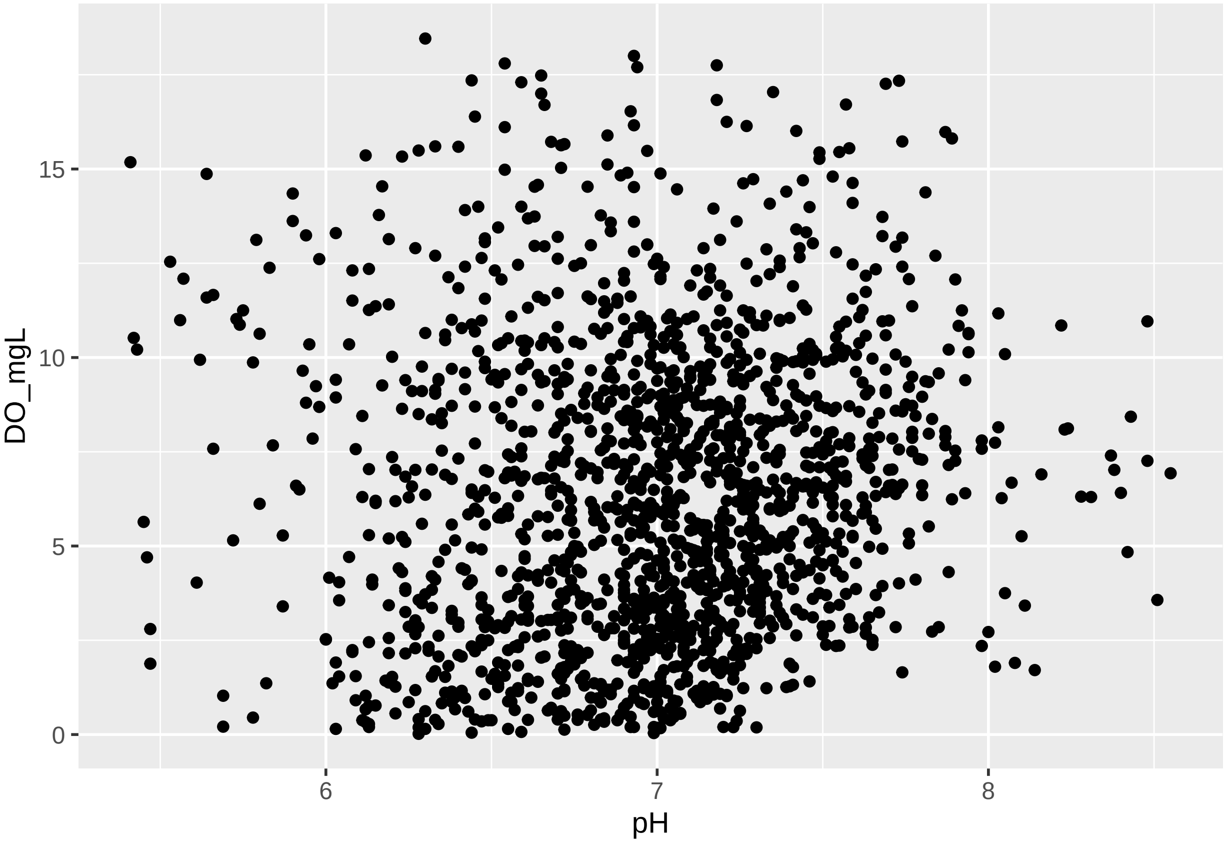
```
#variables extremely correlated (as expected) should not be used in the same model

DO_pH_plot <- ggplot(wetlands, aes(x = pH, y = DO_percent)) +
  geom_point()
DO_pH_plot
```

```
## Warning: Removed 27 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



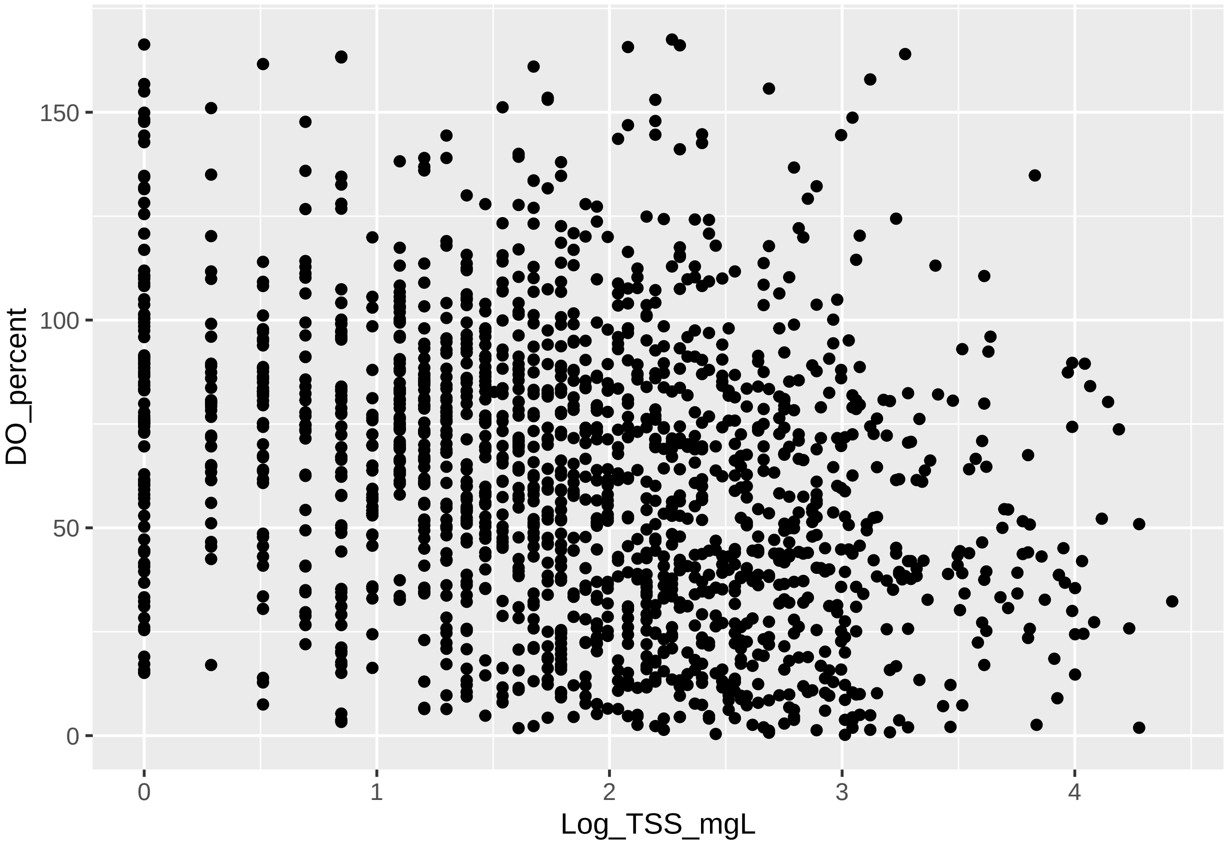
```
DO_pH_plot2 <- ggplot(wetlands, aes(x = pH, y = DO_mgL)) +
  geom_point()
DO_pH_plot2
```

```
## Warning: Removed 24 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



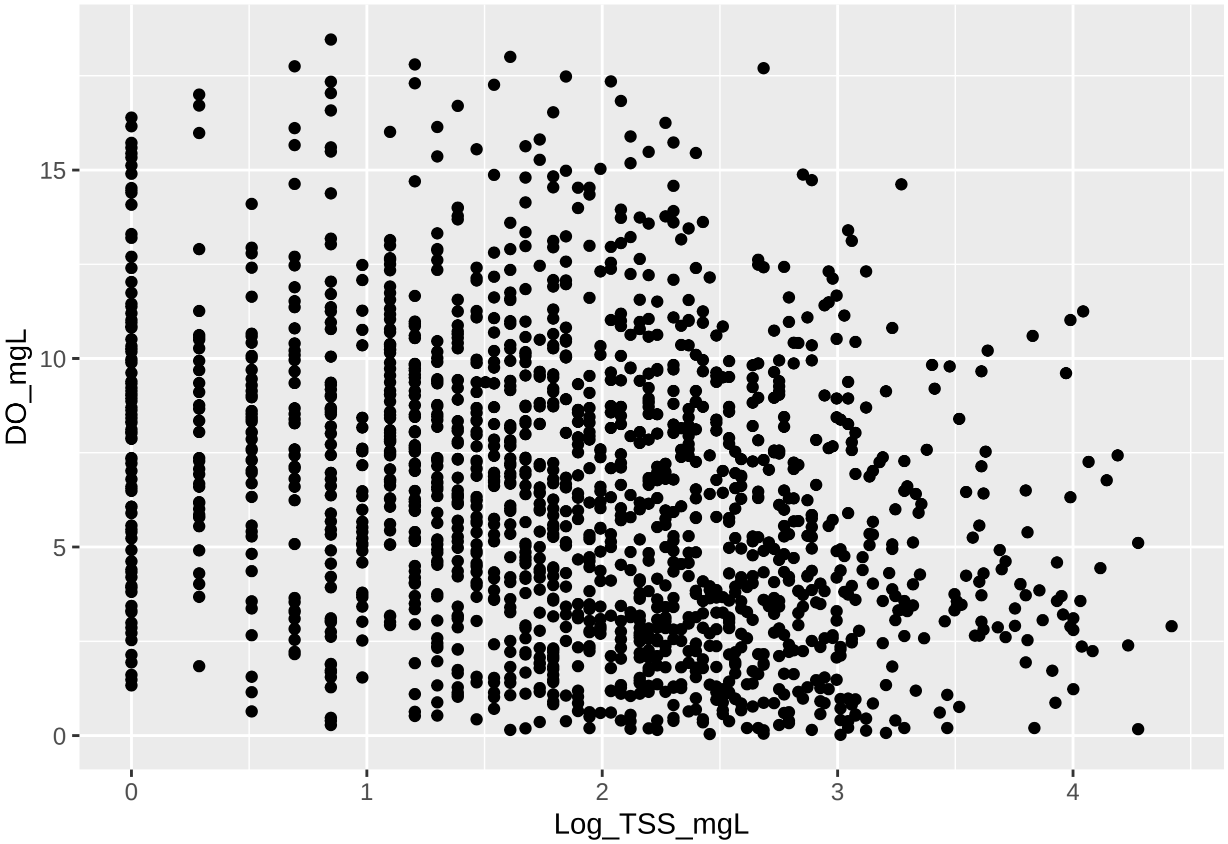
```
#pH does not seem correlated with either DO mgl or DO %

DO_TSS_plot <- ggplot(wetlands, aes(x = Log_TSS_mgL, y = DO_percent)) +
  geom_point()
DO_TSS_plot
```

```
## Warning: Removed 26 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



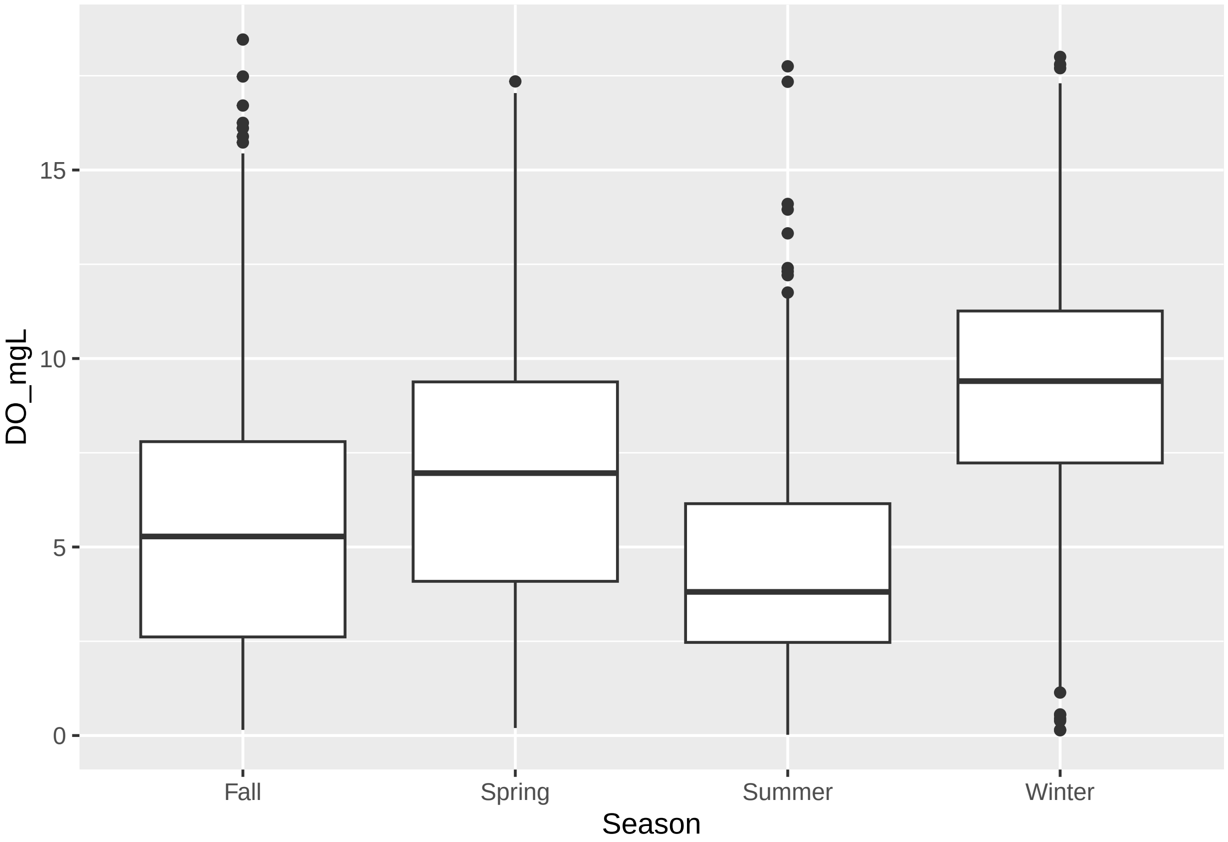
```
DO_TSS_plot2 <- ggplot(wetlands, aes(x = Log_TSS_mgL, y = DO_mgL)) +
  geom_point()
DO_TSS_plot2
```

```
## Warning: Removed 23 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



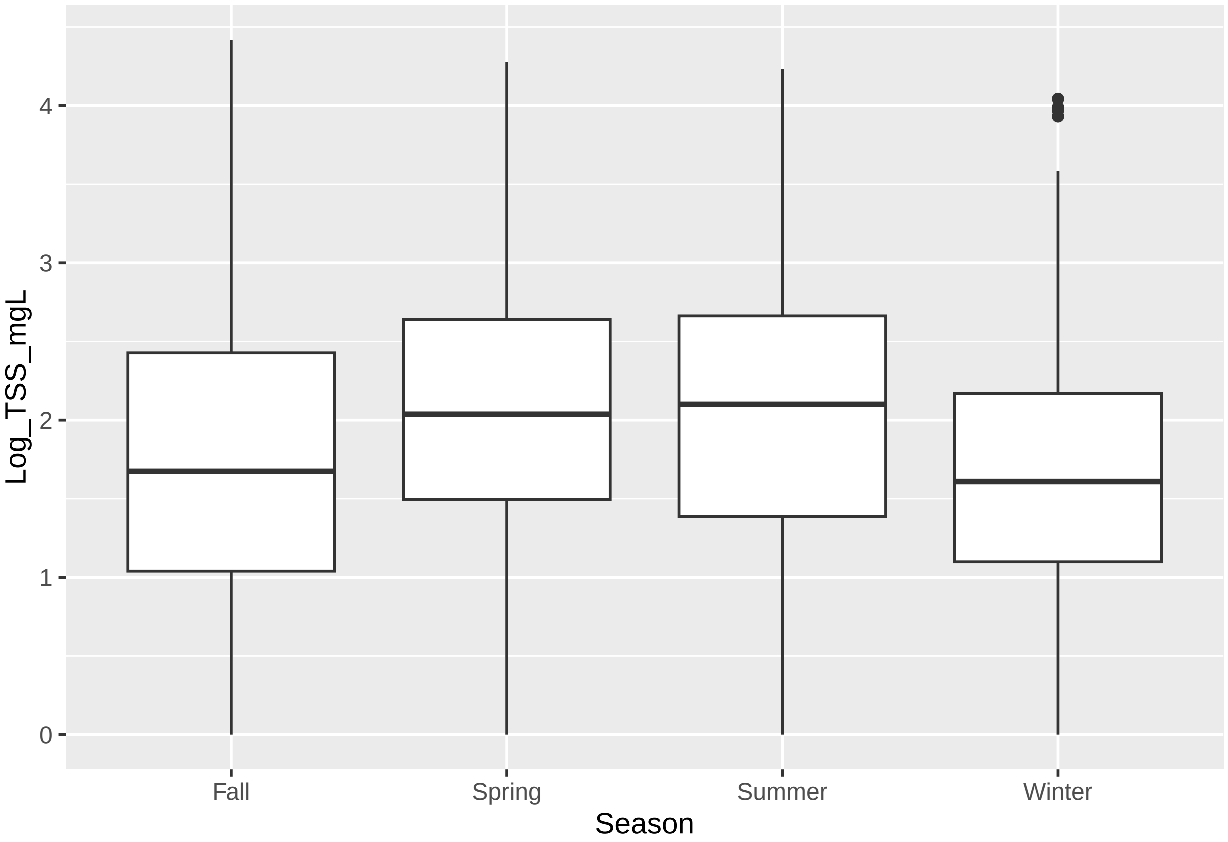
```
#logTSS does not seem correlated with either DO mgl or DO %

DO_season_plot <- ggplot(wetlands, aes(x = Season, y = DO_mgL)) +
  geom_boxplot()
DO_season_plot
```

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



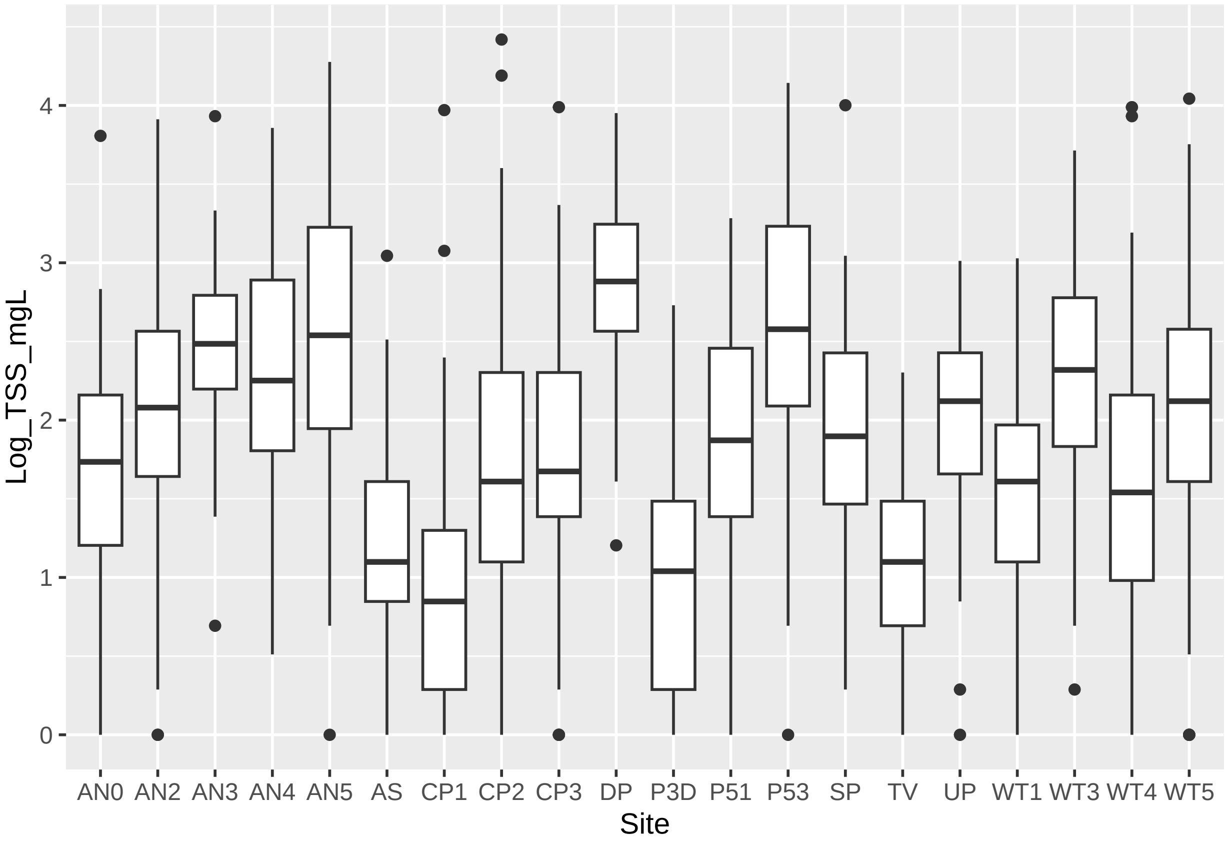
```
#season appears to influence DO mgL but not too much to avoid using together


TSS_season_plot <- ggplot(wetlands, aes(x = Season, y = Log_TSS_mgL)) +
  geom_boxplot()
TSS_season_plot
```

```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



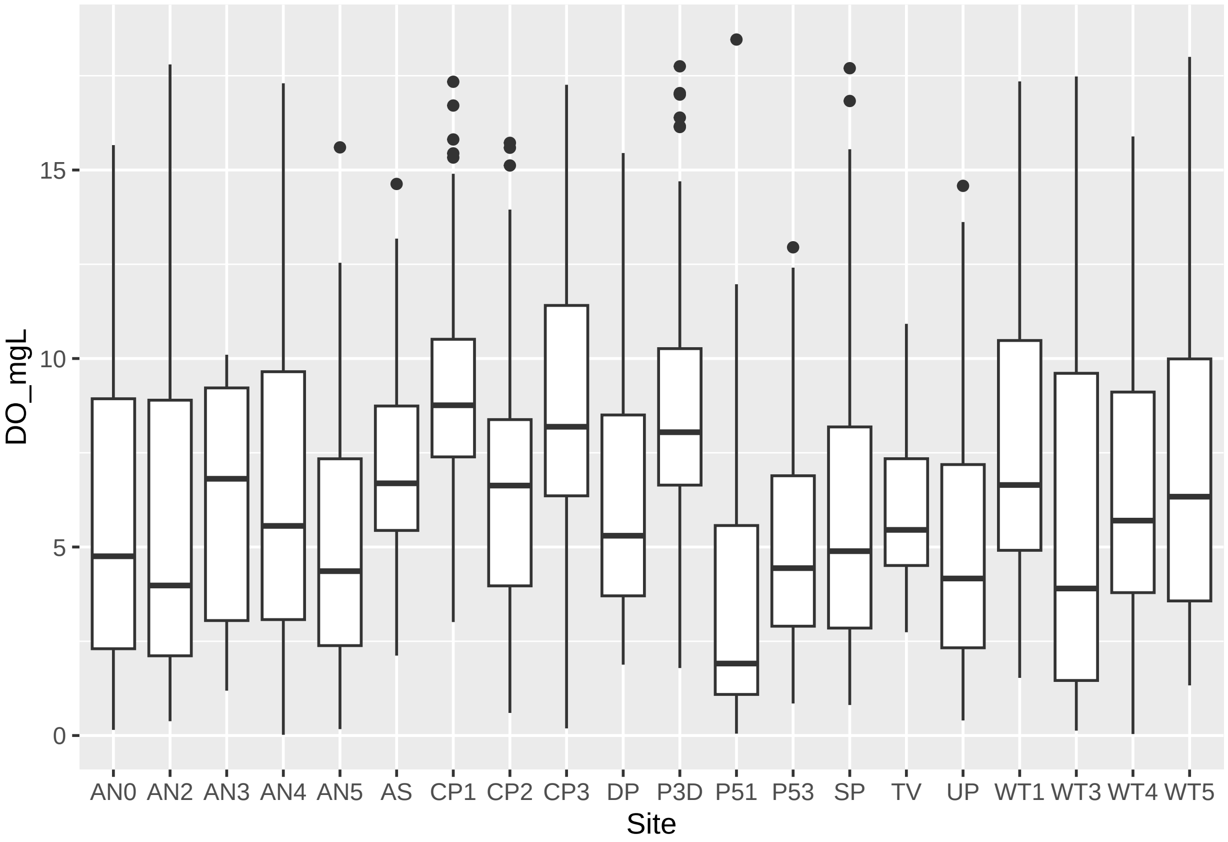
```
#season does not appear to strongly influence TSS

#site based differences?

TSS_site_plot <- ggplot(wetlands, aes(x = Site, y = Log_TSS_mgL)) +
  geom_boxplot()
TSS_site_plot
```

```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



```
#potential influences of site on TSS

O2_site_plot <- ggplot(wetlands, aes(x = Site, y = DO_mgL)) +
  geom_boxplot()
O2_site_plot
```

```
## Warning: Removed 8 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



```
#no signif differences
```

potential further filtering of data ?

removal of sites and years with the lowest number of observations, probably should only use if we are going to do nested linear reg?

```
##filter out first and last years
##filter out AN3/ AS/ SP/TV/ UP

#wetlands2 <- wetlands %>%
 # filter(!(Year  %in% c(2015, 2025))) %>%
 # filter(!(Site %in% c("AN3", "AS", "SP", "TV", "UP", "DP")))

#ggplot(wetlands2, aes(Site))+
 # geom_bar()

#ggplot(wetlands2, aes(Year))+
 # geom_bar()
```

Linear Models

Which Oxygen Measurement to Use

```
#Determining which oxygen measurement would be best to use in the models
wetlands02_perc.lm <- lm(Log_Filtered_NOx_ugL ~ DO_percent,
                       data = wetlands)
summary(wetlands02_perc.lm)
```

```
##
## Call:
## lm(formula = Log_Filtered_NOx_ugL ~ DO_percent, data = wetlands)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6276 -0.9211  0.2515  1.0821  4.7298
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.59903    0.08180   44.00   <2e-16 ***
## DO_percent   0.01853    0.00115   16.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.581 on 1593 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.1403, Adjusted R-squared:  0.1398
## F-statistic:   260 on 1 and 1593 DF,  p-value: < 2.2e-16
```

```
AIC(wetlands02_perc.lm)
```

```
## [1] 5991.301
```

```
#DO % AIC = 5991.301

wetlands02_mgL.lm <-  lm(Log_Filtered_NOx_ugL ~ DO_mgL,
                         data = wetlands)
summary(wetlands02_mgL.lm)
```

```
##
## Call:
## lm(formula = Log_Filtered_NOx_ugL ~ DO_mgL, data = wetlands)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6948 -0.9454  0.2326  1.0749  4.6245
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.70530    0.07675   48.28   <2e-16 ***
## DO_mgL       0.16194    0.01017   15.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.585 on 1596 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.1371, Adjusted R-squared:  0.1366
## F-statistic: 253.6 on 1 and 1596 DF,  p-value: < 2.2e-16
```

```
AIC(wetlandsO2_mgL.lm)
```

```
## [1] 6010.893
```

```
#DO mgL AIC = 6010.893

#The oxygen measurements were fairly similar in predictive power, however the DO% had a slightly more p
```

Notes:

Continuous Variables: - DO% / DO mg_L/ pH / LogTSS Categorical Variables: - Site/ month/ Year/ season

Look at - DO% vs DO mg/L - season v month v year - combinations of season + year/ month + year

Random Effects - site - year - potentially site nested in year