10-8-19 exploratory analysis

Megan Beaudry

10/8/2019

# Exploratory Analysis

##Prelimnary Loading Load clean data

clean\_water\_quality\_data <- readRDS("~/Desktop/epid8060fall2019/MeganBeaudry-Project\_WQ/data/processed\_data/clean\_water\_quality\_data.rds")

Load libraries

library(ggplot2)  
library(ggthemes)  
library(RColorBrewer)  
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

##Exploratory Analysis - Summary Statistics

summary(clean\_water\_quality\_data$Rainfall)

## Length Class Mode   
## 533 character character

summary(clean\_water\_quality\_data$HF183)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.263 3.330 3.330 3.492 3.632 6.038

summary(clean\_water\_quality\_data$HumM2)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.336 3.847 3.847 3.870 3.847 5.057

summary(clean\_water\_quality\_data$CG01)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.109 3.109 3.109 3.125 3.109 4.871

summary(clean\_water\_quality\_data$LeeSg)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.284 3.284 3.284 3.330 3.284 4.709

summary(clean\_water\_quality\_data$Dog3)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.157 3.157 3.157 3.169 3.157 4.547

summary(clean\_water\_quality\_data$MuBac)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.378 3.555 3.555 3.559 3.555 4.366

summary(clean\_water\_quality\_data$Rum2Bac)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.255 3.457 3.457 3.464 3.457 5.231

summary(clean\_water\_quality\_data$A\_butlzeri\_HSP60)

## Length Class Mode   
## 0 NULL NULL

summary(clean\_water\_quality\_data$Salmonella\_spp\_InvA)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.171 3.171 3.171 3.173 3.171 3.472

summary(clean\_water\_quality\_data$Campylobacter\_spp\_Van\_Dkye)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.487 3.555 3.555 3.571 3.555 4.369

summary(clean\_water\_quality\_data$Enterococcus\_CCE)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.6021 2.0774 2.4297 2.6199 2.8797 52.1365 2

summary(clean\_water\_quality\_data$total\_coliforms)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.255 2.862 3.384 3.067 3.384 3.384 61

summary(clean\_water\_quality\_data$E\_coli)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0000 0.6021 1.2788 1.3119 1.8736 3.3837 3

summary(clean\_water\_quality\_data$Enteroalert)

## Length Class Mode   
## 533 character character

summary(clean\_water\_quality\_data$thermotolerant\_coliforms)

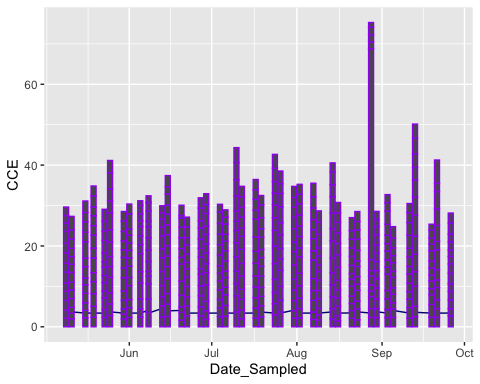
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0000 0.9542 1.3010 1.5138 2.0000 2.9375 1

##Exploratory Analysis - Temporal Trends

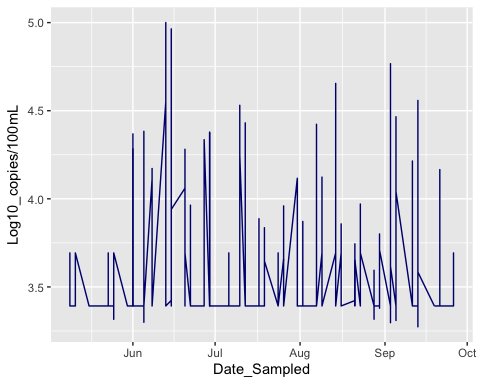
*Temporal Trends All Pond Comparison* Comparison with traditional water quality indicators and Arco. I am really unhappy with this figure. I need to brainstorm a better way to represent data

ggplot(clean\_water\_quality\_data) +   
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_bar(mapping = aes(x = Date\_Sampled, y = Enterococcus\_CCE), color = "purple", stat = "identity") +   
 ylab("CCE")

## Warning: Removed 2 rows containing missing values (position\_stack).

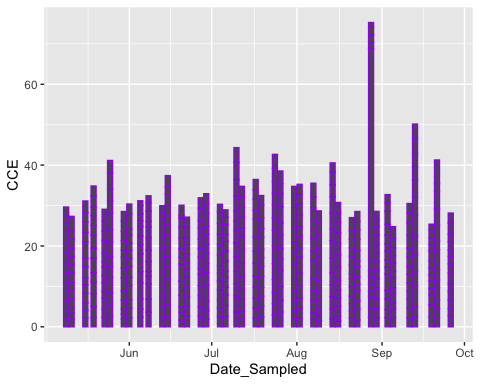


ggplot(clean\_water\_quality\_data) +   
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 ylab(expression("Log10\_copies/100mL"))



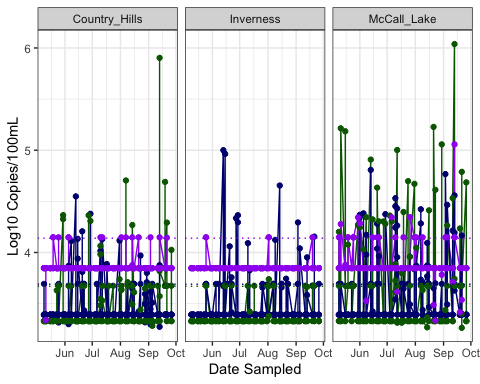
ggplot(clean\_water\_quality\_data) +  
 geom\_bar(mapping = aes(x = Date\_Sampled, y = Enterococcus\_CCE), color = "purple", stat = "identity") +   
 ylab("CCE")

## Warning: Removed 2 rows containing missing values (position\_stack).



The temporal trends by pond of HF183, HumM2 & A. butzleri to see if there was a relationship between source and pathogen

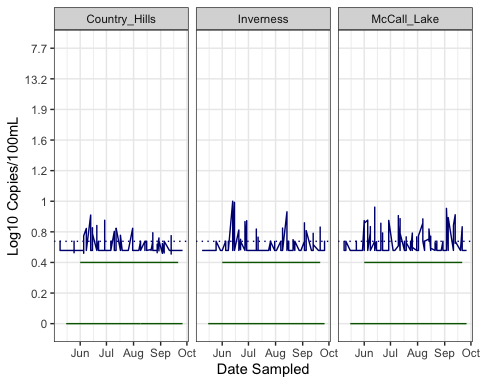
All\_Human\_Arco <- ggplot(clean\_water\_quality\_data) +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = HF183), color = "dark green") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = HumM2), color = "purple") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = HF183), color = "dark green") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = HumM2), color = "purple") +  
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 facet\_wrap(~ Pond) +  
 geom\_hline(yintercept = 3.67, colour = "dark green", linetype = "dotted") +  
 geom\_hline(yintercept = 4.14, colour = "purple", linetype = "dotted") +   
 geom\_hline(yintercept = 3.69, colour = "navy", linetype = "dotted") +  
 theme\_bw()  
All\_Human\_Arco



ggsave(filename = "../../results/All\_Human\_Arco.png",plot = All\_Human\_Arco)

## Saving 5 x 4 in image

All\_rain\_Arco <- ggplot(clean\_water\_quality\_data) +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = Rainfall), color = "dark green") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 facet\_wrap(~ Pond) +  
 geom\_hline(yintercept = 3.69, colour = "navy", linetype = "dotted") +  
 theme\_bw()  
All\_rain\_Arco



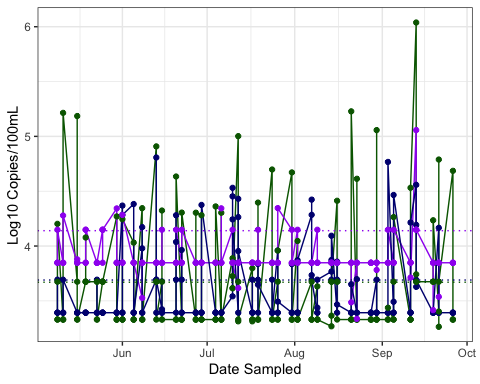
ggsave(filename = "../../results/All\_rain\_Arco.png",plot = All\_rain\_Arco)

## Saving 5 x 4 in image

*Temporal Trends grouped by Pond and Source* Temporal trend of each pond individually with pathogen and source

Part 1: Human and Arco

McCall\_Human\_Arco <- clean\_water\_quality\_data %>%  
 filter(Pond == "McCall\_Lake") %>%  
 ggplot() +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = HF183), color = "dark green") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = HumM2), color = "purple") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_hline(yintercept = 3.67, colour = "dark green", linetype = "dotted") +  
 geom\_hline(yintercept = 4.14, colour = "purple", linetype = "dotted") +   
 geom\_hline(yintercept = 3.69, colour = "navy", linetype = "dotted") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = HF183), color = "dark green") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = HumM2), color = "purple") +  
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 theme\_bw()  
McCall\_Human\_Arco

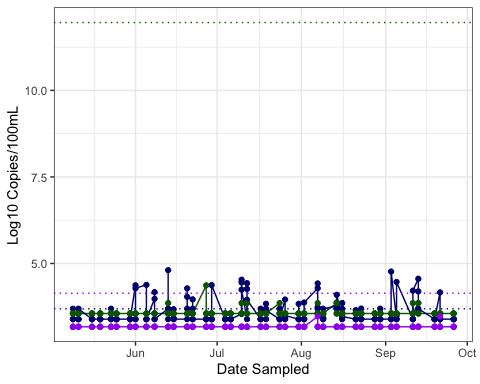


ggsave(filename = "../../results/McCall\_Human\_Arco.png",plot = McCall\_Human\_Arco)

## Saving 5 x 4 in image

All pathogens in McCall

McCall\_Pathogens <- clean\_water\_quality\_data %>%  
 filter(Pond == "McCall\_Lake") %>%  
 ggplot() +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = Campylobacter\_spp\_Van\_Dkye), color = "dark green") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = Salmonella\_spp\_InvA), color = "purple") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_hline(yintercept = 11.96, colour = "dark green", linetype = "dotted") +  
 geom\_hline(yintercept = 4.14, colour = "purple", linetype = "dotted") +   
 geom\_hline(yintercept = 3.69, colour = "navy", linetype = "dotted") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = Campylobacter\_spp\_Van\_Dkye), color = "dark green") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = Salmonella\_spp\_InvA), color = "purple") +  
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 theme\_bw()  
McCall\_Pathogens

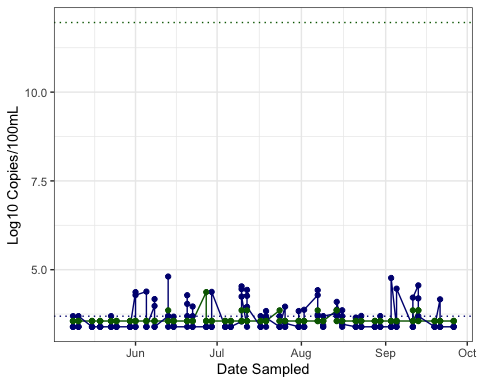


ggsave(filename = "../../results/McCall\_Pathogens.png",plot = McCall\_Pathogens)

## Saving 5 x 4 in image

Took out Salmonella because we didnt find any…However THIS graph looks horrible becasue the LOD for campy is so high compared to the others (i.e., 11 to 3.5 on average). I need to come up with a better to show this if i want to show it. It really is a negative result, as we did not find much Campylobacter spp. However, they are highly related bacteria so some kind of relationship would be nice to show…

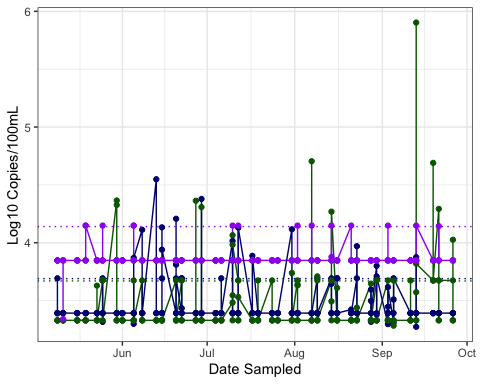
McCall\_Camp\_Arco <- clean\_water\_quality\_data %>%  
 filter(Pond == "McCall\_Lake") %>%  
 ggplot() +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = Campylobacter\_spp\_Van\_Dkye), color = "dark green") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_hline(yintercept = 11.96, colour = "dark green", linetype = "dotted") +  
 geom\_hline(yintercept = 3.69, colour = "navy", linetype = "dotted") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = Campylobacter\_spp\_Van\_Dkye), color = "dark green") +  
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 theme\_bw()  
McCall\_Camp\_Arco



ggsave(filename = "../../results/McCall\_Camp\_arco.png",plot = McCall\_Camp\_Arco)

## Saving 5 x 4 in image

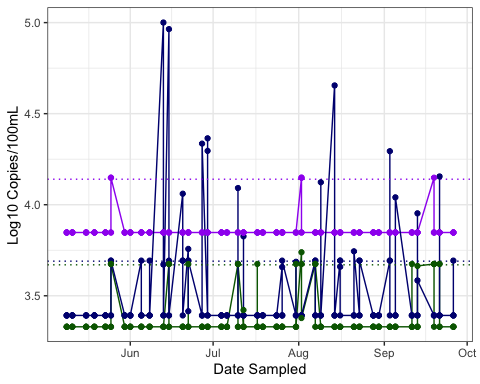
Country\_Hills\_Hum\_Arco<- clean\_water\_quality\_data %>%  
 filter(Pond == "Country\_Hills") %>%  
 ggplot() +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = HF183), color = "dark green") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = HumM2), color = "purple") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_hline(yintercept = 3.67, colour = "dark green", linetype = "dotted") +  
 geom\_hline(yintercept = 4.14, colour = "purple", linetype = "dotted") +   
 geom\_hline(yintercept = 3.69, colour = "navy", linetype = "dotted") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = HF183), color = "dark green") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = HumM2), color = "purple") +  
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 theme\_bw()  
Country\_Hills\_Hum\_Arco



ggsave(filename = "../../results/Country\_Hills\_Hum\_Arco.png",plot = Country\_Hills\_Hum\_Arco)

## Saving 5 x 4 in image

Inverness\_Hum\_Arco <- clean\_water\_quality\_data %>%  
 filter(Pond == "Inverness") %>%  
 ggplot() +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = HF183), color = "dark green") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = HumM2), color = "purple") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_hline(yintercept = 3.67, colour = "dark green", linetype = "dotted") +  
 geom\_hline(yintercept = 4.14, colour = "purple", linetype = "dotted") +   
 geom\_hline(yintercept = 3.69, colour = "navy", linetype = "dotted") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = HF183), color = "dark green") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = HumM2), color = "purple") +  
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 theme\_bw()  
Inverness\_Hum\_Arco

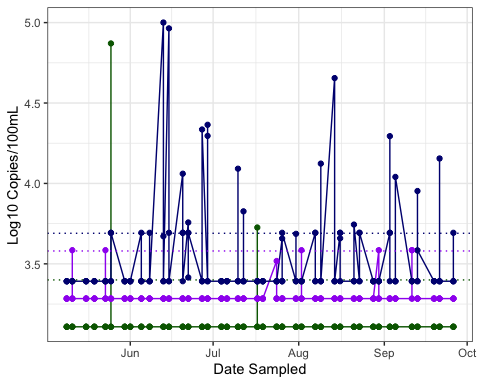


ggsave(filename = "../../results/Inverness\_Hum\_Arco.png",plot = Inverness\_Hum\_Arco)

## Saving 5 x 4 in image

Part 2: Birds and Arco

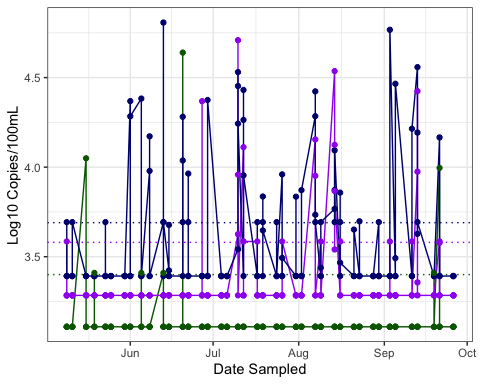
Inverness\_Bird\_Arco <- clean\_water\_quality\_data %>%  
 filter(Pond == "Inverness") %>%  
 ggplot() +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = CG01), color = "dark green") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = LeeSg), color = "purple") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_hline(yintercept = 3.4, colour = "dark green", linetype = "dotted") +  
 geom\_hline(yintercept = 3.58, colour = "purple", linetype = "dotted") +   
 geom\_hline(yintercept = 3.69, colour = "navy", linetype = "dotted") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = CG01), color = "dark green") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = LeeSg), color = "purple") +  
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 theme\_bw()  
Inverness\_Bird\_Arco



ggsave(filename = "../../results/Inverness\_Bird\_Arco.png",plot = Inverness\_Bird\_Arco)

## Saving 5 x 4 in image

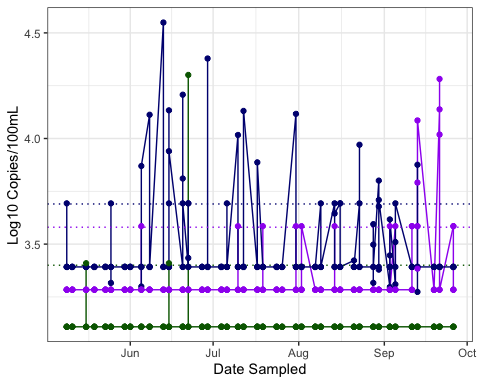
McCall\_Bird\_Arco <- clean\_water\_quality\_data %>%  
 filter(Pond == "McCall\_Lake") %>%  
 ggplot() +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = CG01), color = "dark green") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = LeeSg), color = "purple") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_hline(yintercept = 3.4, colour = "dark green", linetype = "dotted") +  
 geom\_hline(yintercept = 3.58, colour = "purple", linetype = "dotted") +   
 geom\_hline(yintercept = 3.69, colour = "navy", linetype = "dotted") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = CG01), color = "dark green") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = LeeSg), color = "purple") +  
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 theme\_bw()  
McCall\_Bird\_Arco



ggsave(filename = "../../results/McCall\_Bird\_Arco.png",plot = McCall\_Bird\_Arco)

## Saving 5 x 4 in image

Country\_Hills\_Bird\_Arco <- clean\_water\_quality\_data %>%  
 filter(Pond == "Country\_Hills") %>%  
 ggplot() +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = CG01), color = "dark green") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = LeeSg), color = "purple") +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_hline(yintercept = 3.4, colour = "dark green", linetype = "dotted") +  
 geom\_hline(yintercept = 3.58, colour = "purple", linetype = "dotted") +   
 geom\_hline(yintercept = 3.69, colour = "navy", linetype = "dotted") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = A\_butzleri\_HSP60), color = "navy") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = CG01), color = "dark green") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = LeeSg), color = "purple") +  
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 theme\_bw()  
Country\_Hills\_Bird\_Arco

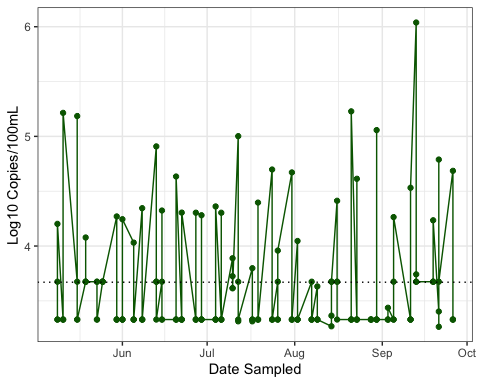


ggsave(filename = "../../results/Country\_Hills\_Bird\_Arco.png", plot = Country\_Hills\_Bird\_Arco)

## Saving 5 x 4 in image

*Temporal Trends One Marker Per Pond*

McCall\_HF183 <- clean\_water\_quality\_data %>%  
 filter(Pond == "McCall\_Lake") %>%  
 ggplot() +  
 geom\_line(mapping = aes(x = Date\_Sampled, y = HF183), color = "dark green") +  
 geom\_point(mapping = aes(x = Date\_Sampled, y = HF183), color = "dark green") +   
 labs (x = "Date Sampled", y = "Log10 Copies/100mL") +  
 geom\_hline(yintercept = 3.67, colour = "black", linetype = "dotted") +  
 theme\_bw() +  
 scale\_color\_discrete(name = "Legend", labels = c("HF183", "LOD95"))  
McCall\_HF183

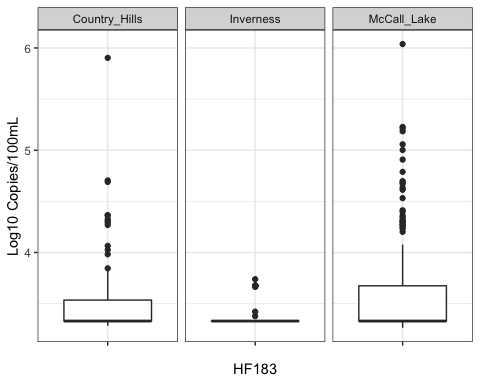


ggsave(filename = "../../results/McCall\_HF183.png", plot = McCall\_HF183)

## Saving 5 x 4 in image

##Exploratory Analysis - Box and Whisker Plots Boxplot of HF183 on all ponds

All\_Box\_HF183 <- ggplot(data = clean\_water\_quality\_data, aes(x = "", y = HF183)) +   
 geom\_boxplot() +  
 facet\_wrap(~ Pond) +  
 theme\_bw() +   
 labs (x = "HF183", y = "Log10 Copies/100mL")  
All\_Box\_HF183

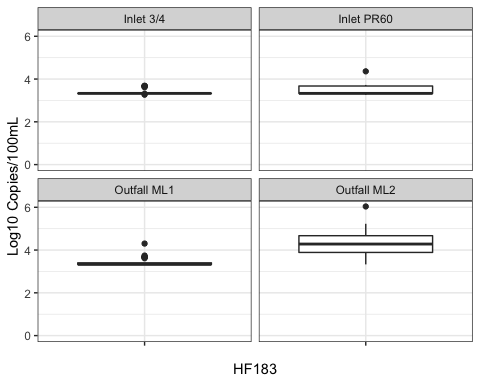


ggsave(filename = "../../results/All\_Box\_HF183.png", plot = All\_Box\_HF183)

## Saving 5 x 4 in image

Since McCall Lake has the most interesting data, lets look at the sampling sites within McCall Lake to see if one site has a higher median (is more contaminated)

McCall\_bysite\_box\_HF183 <- clean\_water\_quality\_data %>%  
 filter(Pond == "McCall\_Lake") %>%  
 ggplot(aes(x = "", y = HF183)) +   
 geom\_boxplot() +  
 coord\_cartesian(ylim = c(0, 6)) +  
 facet\_wrap(~ Sampling\_Site) +  
 theme\_bw() +  
 labs (x = "HF183", y = "Log10 Copies/100mL")  
McCall\_bysite\_box\_HF183



ggsave(filename = "../../results/McCall\_bysite\_box\_HF183.png", plot = McCall\_bysite\_box\_HF183)

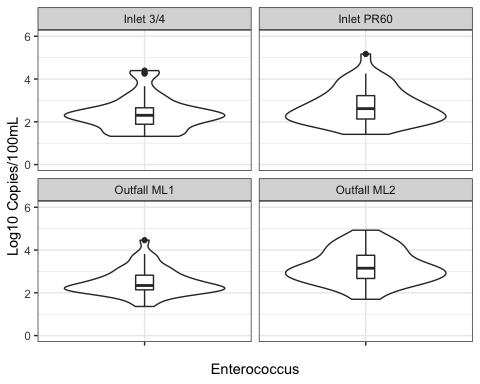
## Saving 5 x 4 in image

##Explatory Analysis - Violin Plots The water quality data measured through molecular methods (i.e., Enterococcus CCE as determined by EPA Method 1611) would be best visualized using a violin plot

McCall\_Enterococcus\_CCE\_Violin <- clean\_water\_quality\_data %>%  
 filter(Pond == "McCall\_Lake") %>%  
 ggplot(aes(x = "", y = Enterococcus\_CCE)) +   
 geom\_violin() +  
 geom\_boxplot(width=0.1) +  
 coord\_cartesian(ylim = c(0, 6)) +  
 facet\_wrap(~ Sampling\_Site) +  
 theme\_bw() +  
 labs (x = "Enterococcus", y = "Log10 Copies/100mL")  
McCall\_Enterococcus\_CCE\_Violin

## Warning: Removed 2 rows containing non-finite values (stat\_ydensity).

## Warning: Removed 2 rows containing non-finite values (stat\_boxplot).



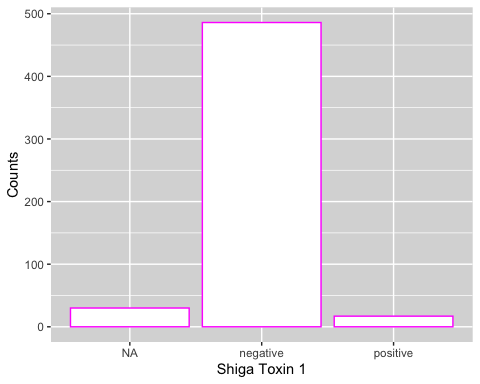
ggsave(filename = "../../results/McCall\_Enterococcus\_CCE\_Violin.png", plot = McCall\_Enterococcus\_CCE\_Violin)

## Saving 5 x 4 in image

## Warning: Removed 2 rows containing non-finite values (stat\_ydensity).  
  
## Warning: Removed 2 rows containing non-finite values (stat\_boxplot).

##Exploratory Analysis - Bar Charts How many samples are positive for shigatoxin 1

Shigatoxin1\_bar <- clean\_water\_quality\_data %>%  
 ggplot() +  
 geom\_bar(aes(x = Shigatoxin\_1), stat = "count", color = "magenta", fill = "white") +  
 labs(x = "Shiga Toxin 1", y = "Counts") +  
 theme(panel.background = element\_rect(fill = "grey85"))  
Shigatoxin1\_bar

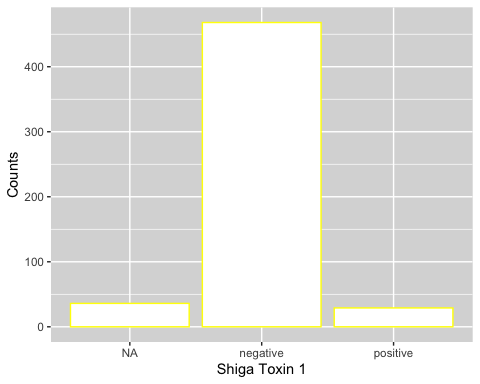


ggsave(filename = "../../results/Shigatoxin1\_bar.png", plot = Shigatoxin1\_bar)

## Saving 5 x 4 in image

How many samples are positive for shigatoxin 2

Shigatoxin2\_bar <- clean\_water\_quality\_data %>%  
 ggplot() +  
 geom\_bar(aes(x = Shigatoxin\_2), stat = "count", color = "yellow", fill = "white") +  
 labs(x = "Shiga Toxin 1", y = "Counts") +  
 theme(panel.background = element\_rect(fill = "grey85"))  
Shigatoxin2\_bar



ggsave(filename = "../../results/Shigatoxin2\_bar.png", plot = Shigatoxin2\_bar)

## Saving 5 x 4 in image