Predictor Analysis

William Norfolk

11/10/2019

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

library(tidyverse)

## -- Attaching packages ------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v readr 1.3.1  
## v tibble 2.1.3 v purrr 0.3.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v ggplot2 3.2.1 v forcats 0.4.0

## -- Conflicts ---------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(forcats)  
library(ggthemes)  
library(knitr)  
library(naniar)  
library(visdat)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(ggpubr)

## Loading required package: magrittr

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(rpart)  
library(rpart.plot)

#load data. path is relative to project directory  
clean\_data <- readRDS("../Data/processed\_data/data\_cleaned.rds")   
  
glimpse(clean\_data)

## Observations: 22  
## Variables: 13  
## $ sample\_id <fct> P1, P2, H1, H2, VM1, VM2, P1, P2, H1, H2,...  
## $ days\_sig\_rainfall <dbl> 15, 15, 15, 15, 15, 15, 27, 27, 27, 27, 2...  
## $ date <chr> "9.11.19", "9.11.19", "9.11.19", "9.11.19...  
## $ ambient\_air\_high\_c <dbl> 35.61111, 35.61111, 35.61111, 35.61111, 3...  
## $ ambient\_air\_low\_c <dbl> 22.22222, 22.22222, 22.22222, 22.22222, 2...  
## $ water\_temp <dbl> 21.52, 22.93, 22.85, 23.82, 34.50, 26.50,...  
## $ conductivity <dbl> 0.119, 0.077, 0.139, 0.212, 0.049, 0.062,...  
## $ ph <dbl> 7.60, 8.09, 10.46, 8.45, 10.59, 8.90, 7.8...  
## $ final\_ecoli <dbl> 40.0, 425.0, 6450.0, 4300.0, 2.5, 2.0, 19...  
## $ exceeds\_epa <chr> "No", "Yes", "Yes", "Yes", "No", "No", "N...  
## $ confirmed\_salmonella <chr> "N/A", "yes", "N/A", "N/A", "N/A", "no", ...  
## $ xlt4\_salmonella <chr> "no", "yes", "no", "no", "no", "yes", "ye...  
## $ last\_rain <dbl> 15, 15, 15, 15, 15, 15, 27, 27, 27, 27, 2...

clean\_data$days\_sig\_rainfall <- as.factor(as.numeric(clean\_data$days\_sig\_rainfall))

Drop NAs

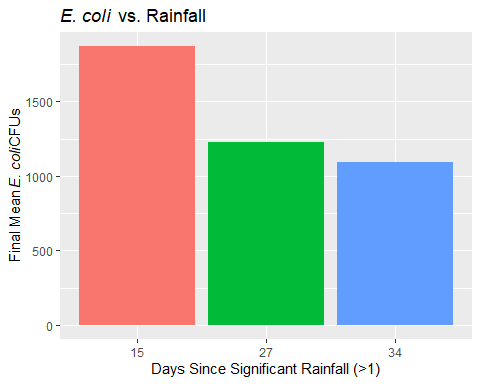
clean\_data <- clean\_data[-c(21,22), ]

Rainfall Plots

aggregate(clean\_data[, 9], list(clean\_data$days\_sig\_rainfall), mean)

## Group.1 final\_ecoli  
## 1 15 1869.917  
## 2 27 1228.643  
## 3 34 1092.500

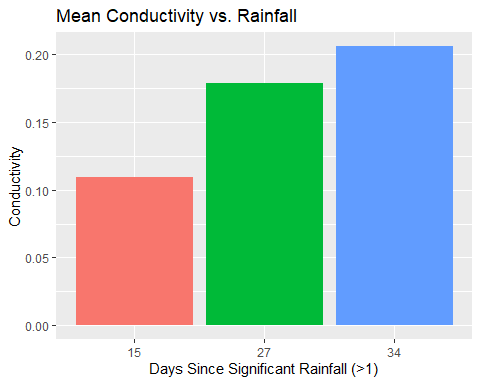
ecoli\_axes <- expression(paste("Final Mean ", italic("E. coli"), "CFUs"))  
title\_here <- expression(paste(italic("E. coli"), " vs. Rainfall"))  
  
rain\_check <- ggplot(clean\_data, aes(x = days\_sig\_rainfall, y = final\_ecoli, fill = days\_sig\_rainfall)) + stat\_summary(fun.y = "mean", geom = "bar") + xlab("Days Since Significant Rainfall (>1)") + ylab(ecoli\_axes) + ggtitle(title\_here) + theme(legend.position = "none")  
  
rain\_check



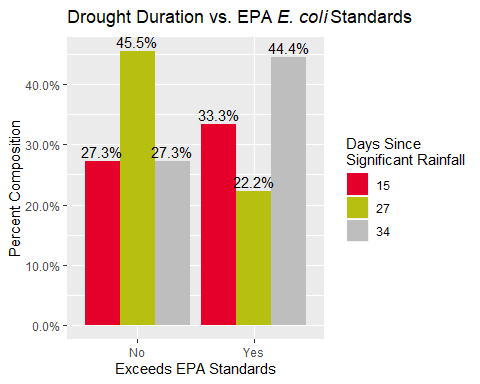
aggregate(clean\_data[, 7], list(clean\_data$days\_sig\_rainfall), mean)

## Group.1 conductivity  
## 1 15 0.1096667  
## 2 27 0.1788571  
## 3 34 0.2061429

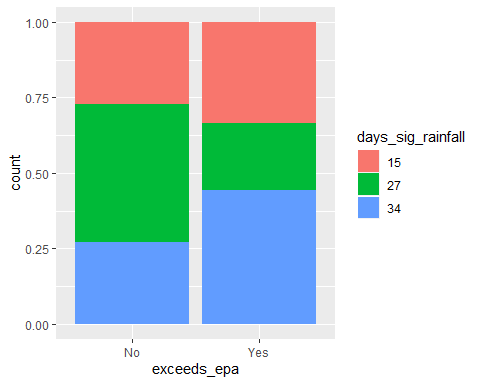
rain\_v\_con <- ggplot(clean\_data, aes(x = days\_sig\_rainfall, y = conductivity, fill = days\_sig\_rainfall)) + stat\_summary(fun.y = "mean", geom = "bar") + xlab("Days Since Significant Rainfall (>1)") + ylab("Conductivity") + ggtitle("Mean Conductivity vs. Rainfall") + theme(legend.position = "none")  
  
rain\_v\_con



side\_bar <- ggplot(clean\_data, aes(x = exceeds\_epa, fill = days\_sig\_rainfall)) + geom\_bar(aes(y = ..count../tapply(..count.., ..x.., sum)[..x..]), position = "dodge") + geom\_text(aes(y = ..count../tapply(..count.., ..x.., sum)[..x..], label = scales::percent(..count../tapply(..count.., ..x.. ,sum)[..x..])), stat = "count", position = position\_dodge(0.9), vjust=-0.3)  
  
title\_1 <- expression(paste("Drought Duration vs. EPA ", italic("E. coli"), "Standards"))  
  
side\_bar + xlab("Exceeds EPA Standards") + ylab("Percent Composition") + scale\_y\_continuous(labels = scales::percent) + ggtitle(title\_1) + theme(legend.position = "right", legend.direction = "vertical") + labs(fill = "Days Since\nSignificant Rainfall") + scale\_fill\_manual(values = c("#E4002B", "#B7BF10", "004E60"))



stack\_bar <- ggplot(clean\_data, aes(x = exceeds\_epa, fill = days\_sig\_rainfall)) + geom\_bar(stat = "count", position = "fill")  
  
stack\_bar



How about a quick ANOVA

mod\_test <- lm(formula = final\_ecoli ~ days\_sig\_rainfall, data = clean\_data)  
  
anova(mod\_test)

## Analysis of Variance Table  
##   
## Response: final\_ecoli  
## Df Sum Sq Mean Sq F value Pr(>F)  
## days\_sig\_rainfall 2 2178189 1089094 0.1947 0.8249  
## Residuals 17 95075985 5592705

anova\_ecoli <- aov(final\_ecoli ~ days\_sig\_rainfall, data = clean\_data)  
  
print(anova\_ecoli)

## Call:  
## aov(formula = final\_ecoli ~ days\_sig\_rainfall, data = clean\_data)  
##   
## Terms:  
## days\_sig\_rainfall Residuals  
## Sum of Squares 2178189 95075985  
## Deg. of Freedom 2 17  
##   
## Residual standard error: 2364.89  
## Estimated effects may be unbalanced

Okay not too great, lets try individual t-test permutations

days\_15 <- subset(clean\_data, days\_sig\_rainfall == "15")  
days\_27 <- subset(clean\_data, days\_sig\_rainfall == "27")  
days\_34 <- subset(clean\_data, days\_sig\_rainfall == "34")  
  
x <- days\_15$final\_ecoli  
y <- days\_27$final\_ecoli  
z <- days\_34$final\_ecoli

library(broom)  
library(purrr)  
  
c <- t.test(x, y)  
  
a <- t.test(x, z)  
  
t <- t.test(y, z)  
  
t\_tab <- map\_df(list(c, a, t), tidy)  
  
kable(t\_tab)

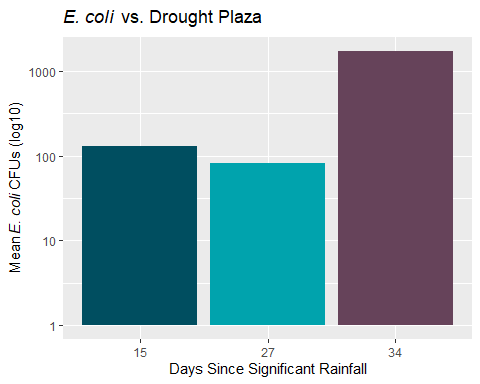
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| estimate | estimate1 | estimate2 | statistic | p.value | parameter | conf.low | conf.high | method | alternative |
| 641.2738 | 1869.917 | 1228.643 | 0.4518381 | 0.6614966 | 9.527792 | -2542.398 | 3824.946 | Welch Two Sample t-test | two.sided |
| 777.4167 | 1869.917 | 1092.500 | 0.5590889 | 0.5895014 | 9.163701 | -2359.582 | 3914.416 | Welch Two Sample t-test | two.sided |
| 136.1429 | 1228.643 | 1092.500 | 0.1181346 | 0.9079225 | 11.955772 | -2375.835 | 2648.121 | Welch Two Sample t-test | two.sided |

Well nothing is significant it looks like unfortunately…

Plaza Over Time

plaza1\_only <- subset(clean\_data, sample\_id == "P1")  
plaza2\_only <- subset(clean\_data, sample\_id == "P2")  
plaza\_total <- rbind(plaza1\_only, plaza2\_only)

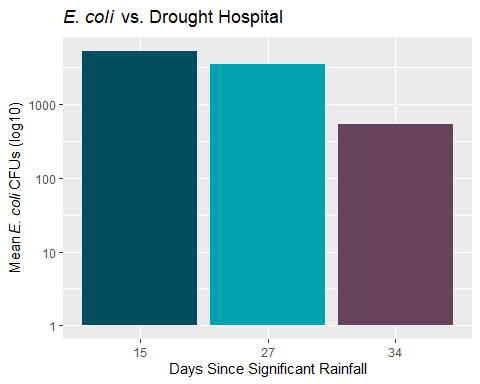
title\_plaza <- expression(paste(italic("E. coli"), " vs. Drought Plaza"))  
mean\_ecoli\_cfu <- expression(paste("Mean ", italic("E. coli "), "CFUs (log10)"))  
  
plaza\_time <- ggplot(plaza\_total, aes(x = days\_sig\_rainfall, y = final\_ecoli, fill = days\_sig\_rainfall)) + stat\_summary(fun.y = "mean", geom = "bar") + theme(legend.position = "none") + xlab("Days Since Significant Rainfall") + ylab(mean\_ecoli\_cfu) + ggtitle(title\_plaza)+ scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#004E60", "#00A3AD", "#66435A"))  
  
plaza\_time



plaza\_arrange <- plaza\_time + expand\_limits(y = c(0:5000)) + ggtitle("Plaza Stream") + xlab("Days Since Rain")

hos1\_only <- subset(clean\_data, sample\_id == "H1")  
hos2\_only <- subset(clean\_data, sample\_id == "H2")  
hos\_total <- rbind(hos1\_only, hos2\_only)

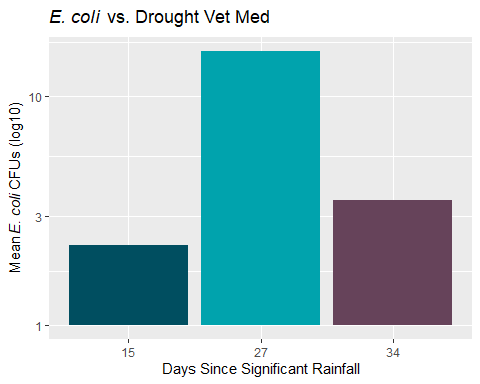
title\_hos <- expression(paste(italic("E. coli"), " vs. Drought Hospital"))  
  
  
hos\_time <- ggplot(hos\_total, aes(x = days\_sig\_rainfall, y = final\_ecoli, fill = days\_sig\_rainfall)) + stat\_summary(fun.y = "mean", geom = "bar") + theme(legend.position = "none") + xlab("Days Since Significant Rainfall") + ylab(mean\_ecoli\_cfu) + ggtitle(title\_hos) + scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#004E60", "#00A3AD", "#66435A"))  
  
hos\_time



hos\_arrange <- hos\_time + expand\_limits(y = c(0:5000)) + ggtitle("Hospital") + xlab("Days Since Rain")

vm1\_only <- subset(clean\_data, sample\_id == "VM1")  
vm2\_only <- subset(clean\_data, sample\_id == "VM2")  
vm\_total <- rbind(vm1\_only, vm2\_only)

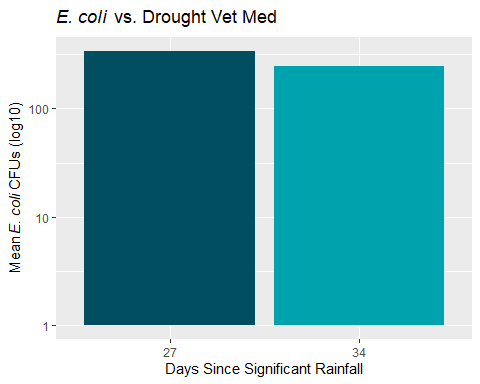
title\_vm <- expression(paste(italic("E. coli"), " vs. Drought Vet Med"))  
  
  
vm\_time <- ggplot(vm\_total, aes(x = days\_sig\_rainfall, y = final\_ecoli, fill = days\_sig\_rainfall)) + stat\_summary(fun.y = "mean", geom = "bar") + theme(legend.position = "none") + xlab("Days Since Significant Rainfall") + ylab(mean\_ecoli\_cfu) + ggtitle(title\_vm)+ scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#004E60", "#00A3AD", "#66435A"))  
  
vm\_time



vm\_arrange <- vm\_time + expand\_limits(y = c(0:5000)) + ggtitle("Vet Med") + xlab("Days Since Rain")

tan\_total <- subset(clean\_data, sample\_id == "T1")

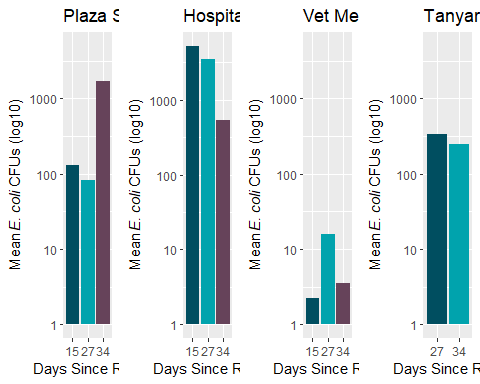
tan\_time <- ggplot(tan\_total, aes(x = days\_sig\_rainfall, y = final\_ecoli, fill = days\_sig\_rainfall)) + stat\_summary(fun.y = "mean", geom = "bar") + theme(legend.position = "none") + xlab("Days Since Significant Rainfall") + ylab(mean\_ecoli\_cfu) + ggtitle(title\_vm)+ scale\_y\_continuous(trans = "log10") + scale\_fill\_manual(values = c("#004E60", "#00A3AD", "#66435A"))  
  
tan\_time



tan\_arrange <- tan\_time + expand\_limits(y = c(0:5000)) + ggtitle("Tanyard") + xlab("Days Since Rain")

ggarrange(plaza\_arrange, hos\_arrange, vm\_arrange, tan\_arrange, nrow = 1, ncol = 4)

## Warning: Transformation introduced infinite values in continuous y-axis  
  
## Warning: Transformation introduced infinite values in continuous y-axis  
  
## Warning: Transformation introduced infinite values in continuous y-axis  
  
## Warning: Transformation introduced infinite values in continuous y-axis



library(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

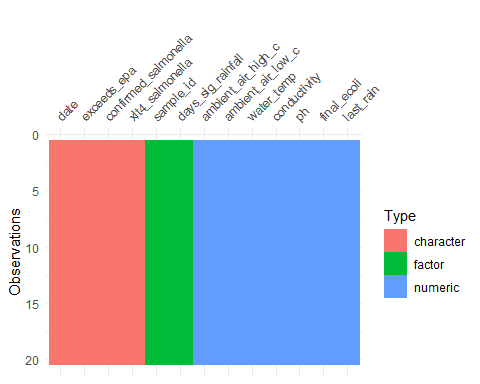
## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

#dual\_line <- plot\_ly(data = clean\_data, x = ~days\_sig\_rainfall, y = ~final\_ecoli, type = "box")  
  
#dual\_line

Tree models are no good unfortunately… likely too few observations?

vis\_dat(clean\_data)



#mod\_data <- clean\_data[!(clean\_data$final\_ecoli == "NA"), ]  
  
#mod\_data <- mod\_data[-c(20,21,22), ]  
  
#vis\_dat(mod\_data)

#```{r} mod\_data <- mod\_data[, c(9,1,2,3,4,5,6,7,8,10,11,12,13)]

#```{r}  
outcomename = "final\_ecoli"

#```{r} set.seed(1111) outcome = “final\_ecoli”

fitControl <- trainControl(method=“repeatedcv”,number=5,repeats=5) Npred <- ncol(mod\_data)-1 resultmat <- data.frame(Variable = names(mod\_data)[-1], Accuracy = rep(0,Npred)) for (n in 2:ncol(mod\_data)) { fit\_single <- caret::train( as.formula(paste(outcomename, “~”,names(mod\_data)[n])) , data = mod\_data, method = “rpart”, trControl = fitControl, na.action = na.pass, tuneLength = 10) resultmat[n-1,2]= max(fit\_singleAccuracy)  
}

print(resultmat)

```