Exploratory Analysis EHSC 8310

William Norfolk

11/3/2019

library(readxl)  
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 0.3.2  
## v tidyr 0.8.3 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

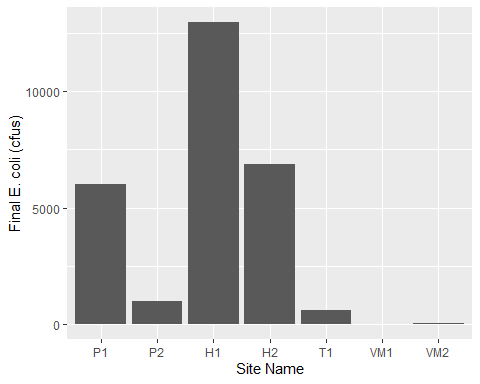
#load data. path is relative to project directory  
clean\_data <- readRDS("../Data/processed\_data/data\_cleaned.rds")

glimpse(clean\_data)

## Observations: 22  
## Variables: 12  
## $ sample\_id <fct> P1, P2, H1, H2, VM1, VM2, P1, P2, H1, H2,...  
## $ last\_rain <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ 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...

site\_vs\_ecoli <- ggplot(clean\_data, aes(x = sample\_id, y = final\_ecoli)) + geom\_bar(stat = "identity")  
  
site\_vs\_ecoli + xlab("Site Name") + ylab("Final E. coli (cfus)")

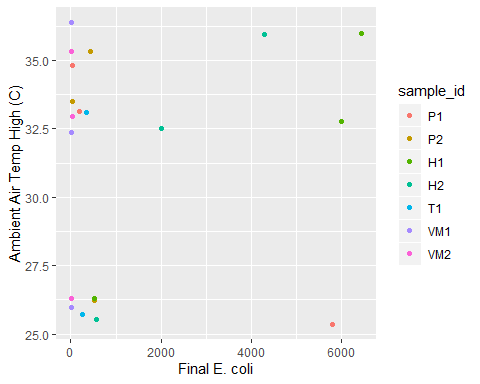
## Warning: Removed 2 rows containing missing values (position\_stack).



Lets look at final E. coli as an oucome.

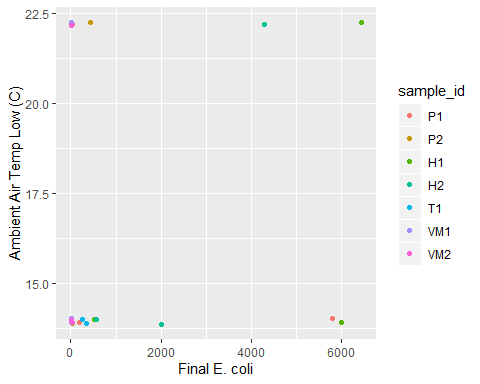
ecoli\_high\_temp\_air <- ggplot(clean\_data, aes(x = final\_ecoli, y = ambient\_air\_high\_c, color = sample\_id)) + geom\_jitter() + xlab("Final E. coli") + ylab("Ambient Air Temp High (C)")  
  
ecoli\_high\_temp\_air

## Warning: Removed 2 rows containing missing values (geom\_point).



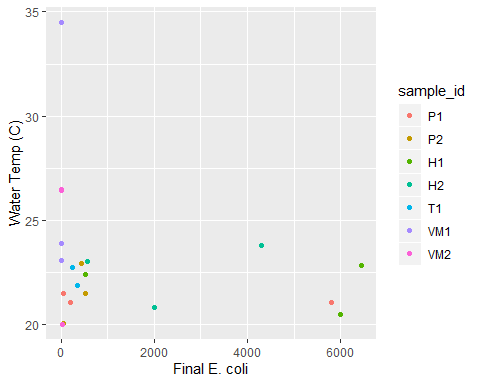
ecoli\_low\_temp\_air <- ggplot(clean\_data, aes(x = final\_ecoli, y = ambient\_air\_low\_c, color = sample\_id)) + geom\_jitter() + xlab("Final E. coli") + ylab("Ambient Air Temp Low (C)")  
  
ecoli\_low\_temp\_air

## Warning: Removed 2 rows containing missing values (geom\_point).



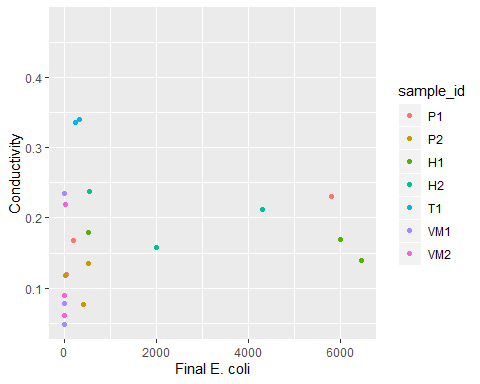
ecoli\_water\_temp <- ggplot(clean\_data, aes(x = final\_ecoli, y = water\_temp, color = sample\_id)) + geom\_jitter() + xlab("Final E. coli") + ylab("Water Temp (C)")  
  
ecoli\_water\_temp

## Warning: Removed 2 rows containing missing values (geom\_point).



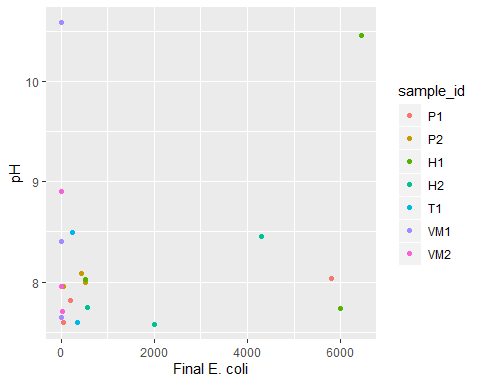
ecoli\_cond <- ggplot(clean\_data, aes(x = final\_ecoli, y = conductivity, color = sample\_id)) + geom\_jitter() + xlab("Final E. coli") + ylab("Conductivity")  
  
ecoli\_cond

## Warning: Removed 2 rows containing missing values (geom\_point).



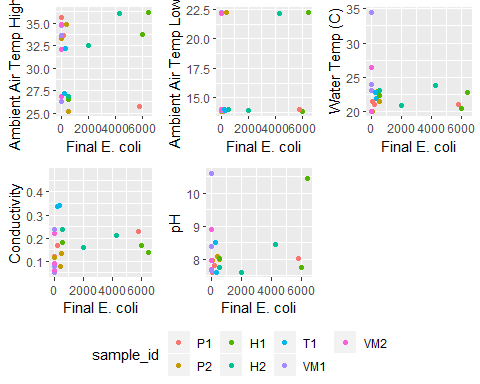
ecoli\_ph <- ggplot(clean\_data, aes(x = final\_ecoli, y = ph, color = sample\_id)) + geom\_jitter() + xlab("Final E. coli") + ylab("pH")  
  
ecoli\_ph

## Warning: Removed 2 rows containing missing values (geom\_point).



ggarrange(ecoli\_high\_temp\_air, ecoli\_low\_temp\_air, ecoli\_water\_temp, ecoli\_cond, ecoli\_ph, ncol = 3, nrow = 2, common.legend = TRUE, legend = "bottom")

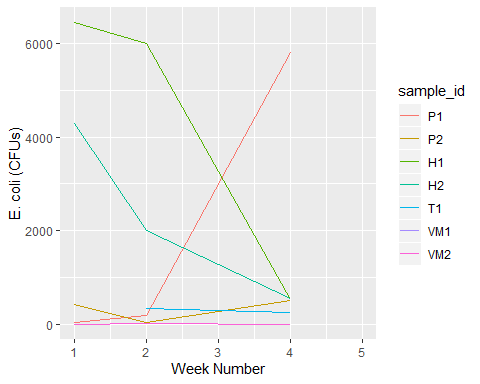
## Warning: Removed 2 rows containing missing values (geom\_point).  
  
## Warning: Removed 2 rows containing missing values (geom\_point).  
  
## Warning: Removed 2 rows containing missing values (geom\_point).  
  
## Warning: Removed 2 rows containing missing values (geom\_point).  
  
## Warning: Removed 2 rows containing missing values (geom\_point).  
  
## Warning: Removed 2 rows containing missing values (geom\_point).



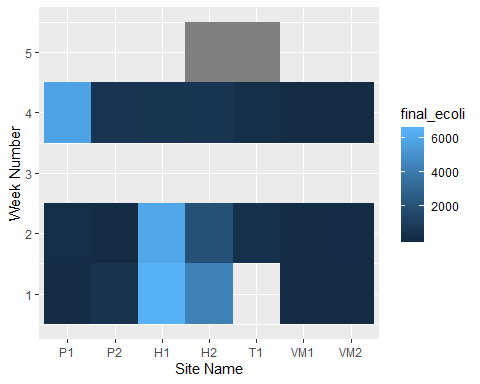
add\_week\_count <- clean\_data %>% dplyr::mutate(week\_count = recode(date,  
 "9.11.19" = 1,  
 "9.23.19" = 2,  
 "9.30.19" = 4,  
 "10.3.19" = 5))

date\_freq <- ggplot(add\_week\_count, aes(x = week\_count, y = final\_ecoli, color = sample\_id)) + geom\_line()  
  
date\_freq + xlab("Week Number") + ylab("E. coli (CFUs)")

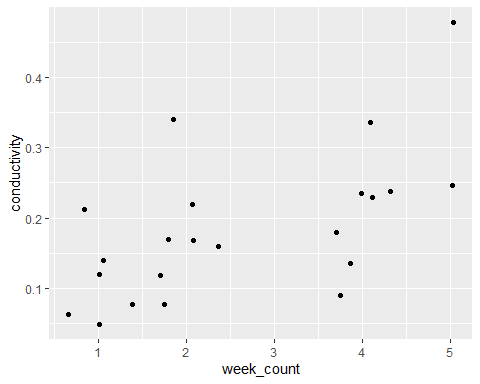
## Warning: Removed 2 rows containing missing values (geom\_path).



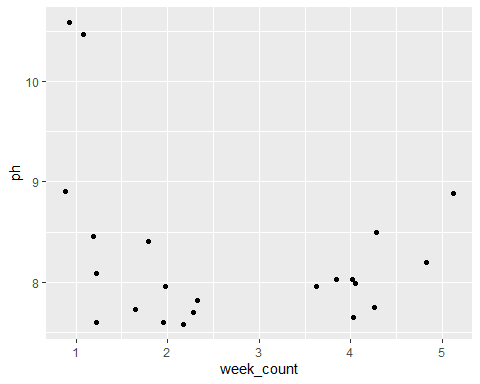
y <- ggplot(add\_week\_count, aes(x = sample\_id, y = week\_count))   
  
y + geom\_tile(aes(fill = final\_ecoli)) +xlab("Site Name") + ylab("Week Number")



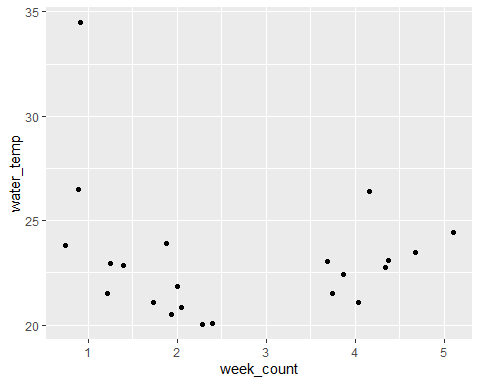
ggplot(add\_week\_count, aes(x = week\_count, y = conductivity)) + geom\_jitter()



ggplot(add\_week\_count, aes(x = week\_count, y = ph)) + geom\_jitter()

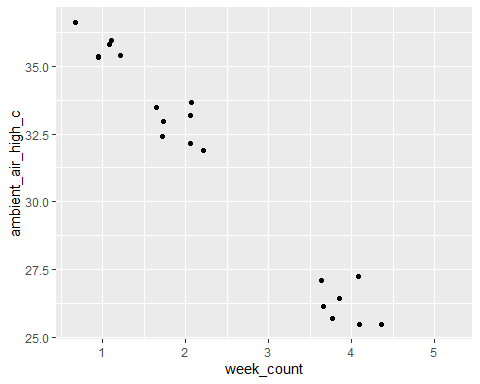


ggplot(add\_week\_count, aes(x = week\_count, y = water\_temp)) + geom\_jitter()



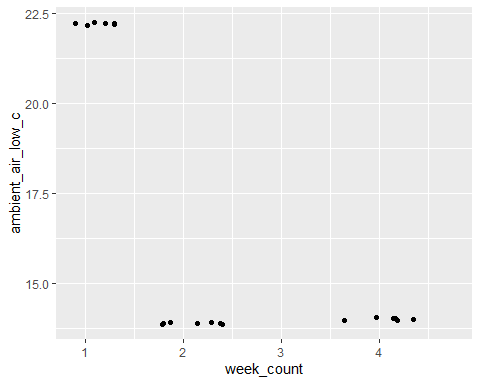
ggplot(add\_week\_count, aes(x = week\_count, y = ambient\_air\_high\_c)) + geom\_jitter()

## Warning: Removed 2 rows containing missing values (geom\_point).



ggplot(add\_week\_count, aes(x = week\_count, y = ambient\_air\_low\_c)) + geom\_jitter()

## Warning: Removed 2 rows containing missing values (geom\_point).



#ggplot(add\_week\_count, aes(x = week\_count, y = last\_rain)) + geom\_jitter()