eda-final.R

niko

2020-06-09

```r
rm( list = ls() )

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------- ti
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

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

```r
library(fields)
```

```
## Loading required package: spam
```

```
## Loading required package: dotCall64
```

```
## Loading required package: grid
```

```
## Spam version 2.5-1 (2019-12-12) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.
```

```
##
## Attaching package: 'spam'
```

```
## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve
```

```
## Loading required package: maps
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
## See https://github.com/NCAR/Fields for
##  an extensive vignette, other supplements and source code
```

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

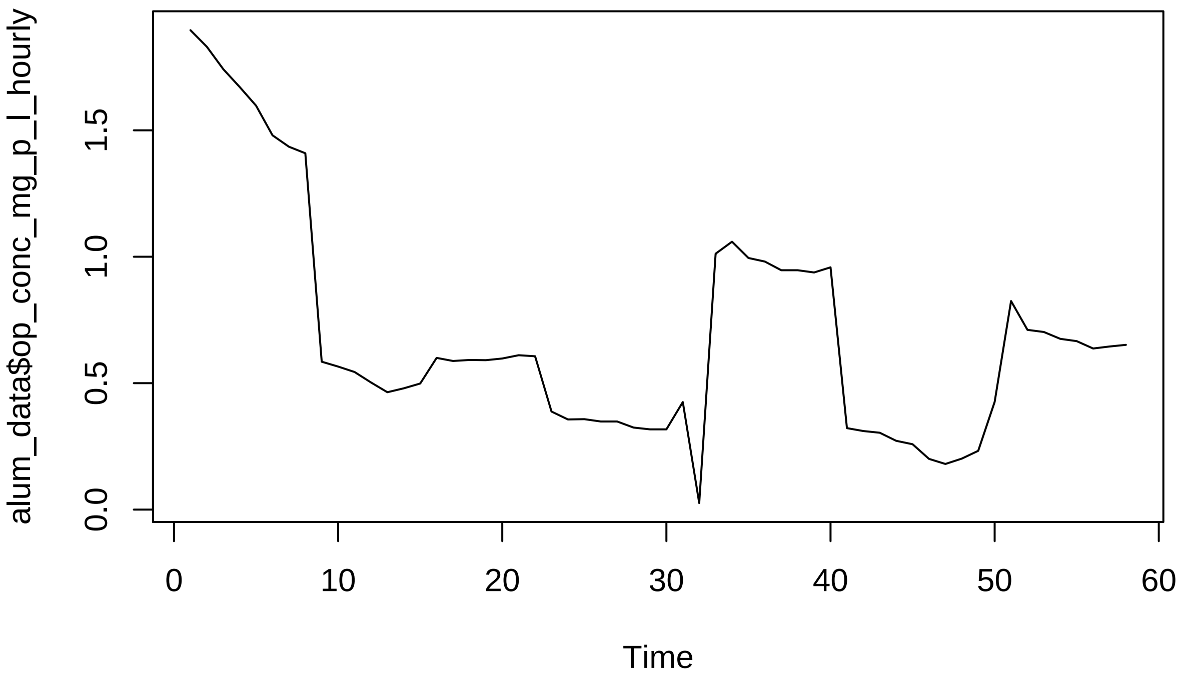
```r
library(latex2exp)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

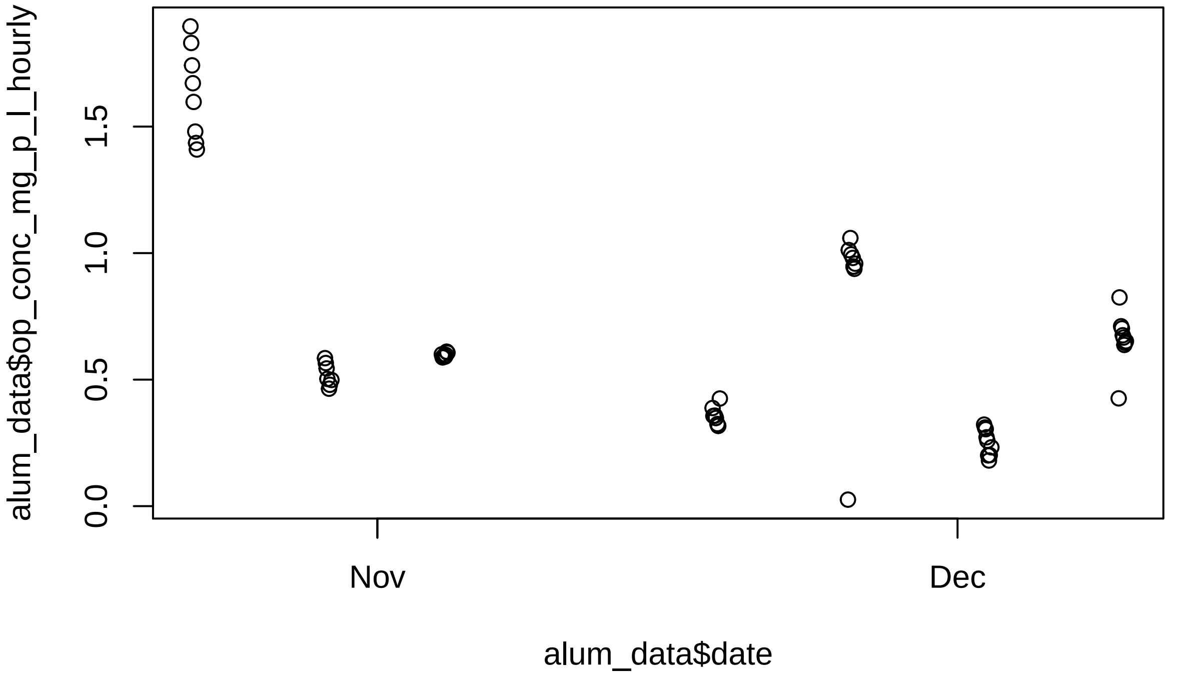
```r
# make sure src is the current working directory

load("../data/final-data.rda")
load("../data/boulderMoWater.rda")

# display the number of observations for each type of coagulant

table(final_data$coagulant)
```

```
##
##          Alum Ferric   None
##      0     58     45      4
```

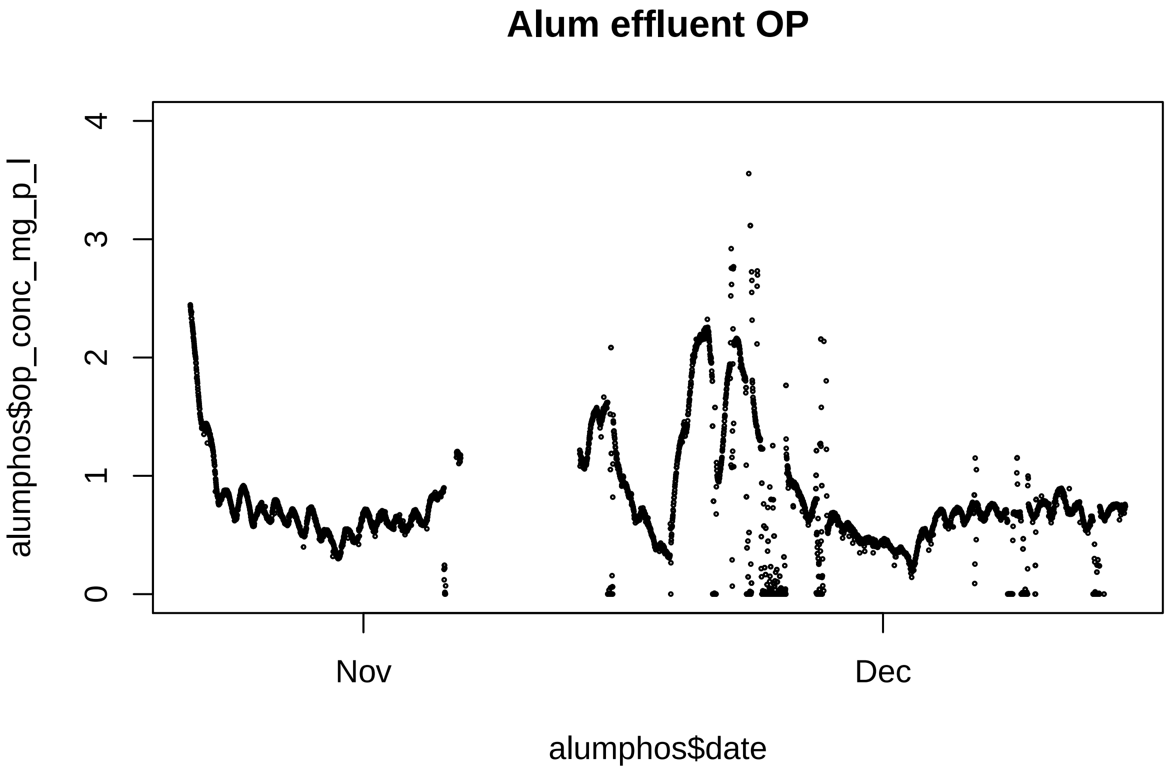
```r
# splitting the data by coagulant

ferr_data <- final_data %>% filter(coagulant == "Ferric")
alum_data <- final_data %>% filter(coagulant == "Alum")
none_data <- final_data %>% filter(coagulant == "None")

# average was found by hour maybe do a moving average instead?

ts.plot(alum_data$op_conc_mg_p_l_hourly)
```



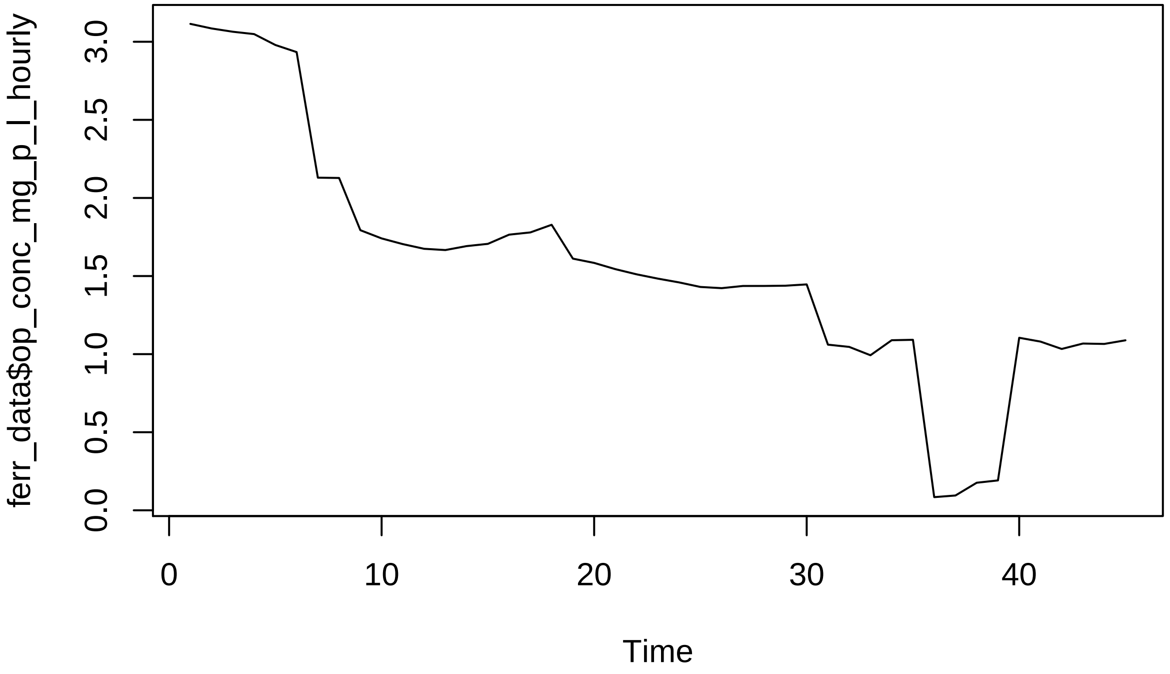
```r
plot(alum_data$date, alum_data$op_conc_mg_p_l_hourly)
```



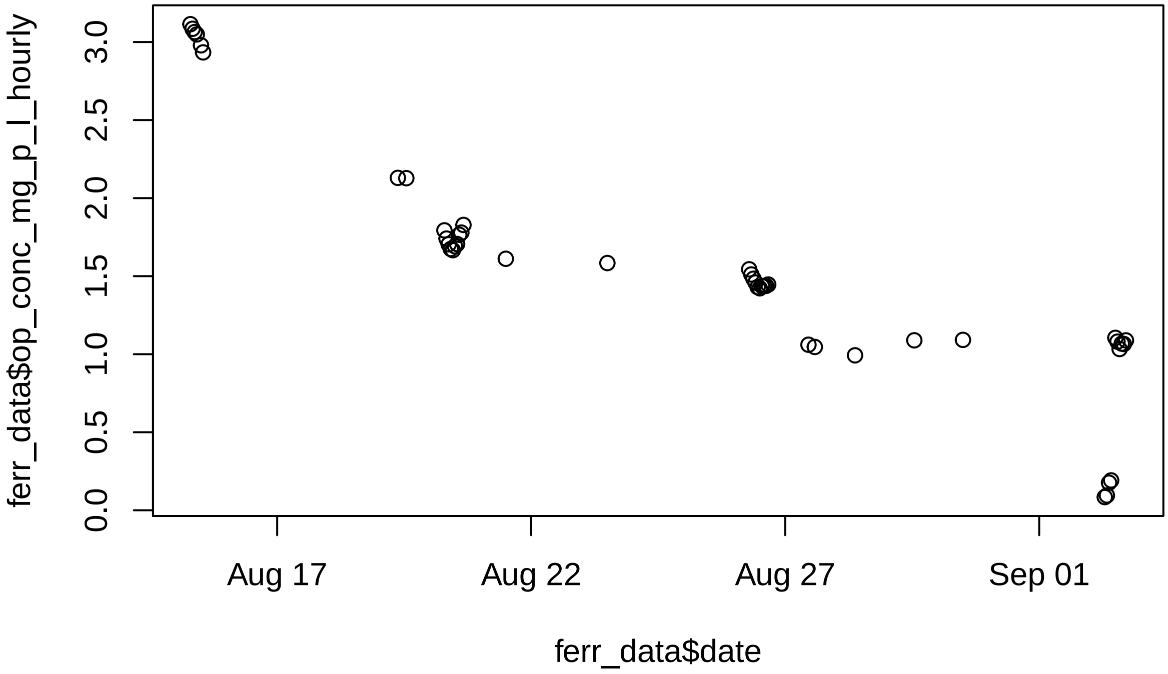
```r
# gap in november looking at this variable at original time scale no average

alumphos <- phosfax_10m %>%
  filter(date >= ymd("2019-10-22")) %>%
  filter(date <= ymd("2019-12-15"))
plot(alumphos$date, alumphos$op_conc_mg_p_l,
     cex = .25,
     main = "Alum effluent OP",
     ylim = c(0,4))
```



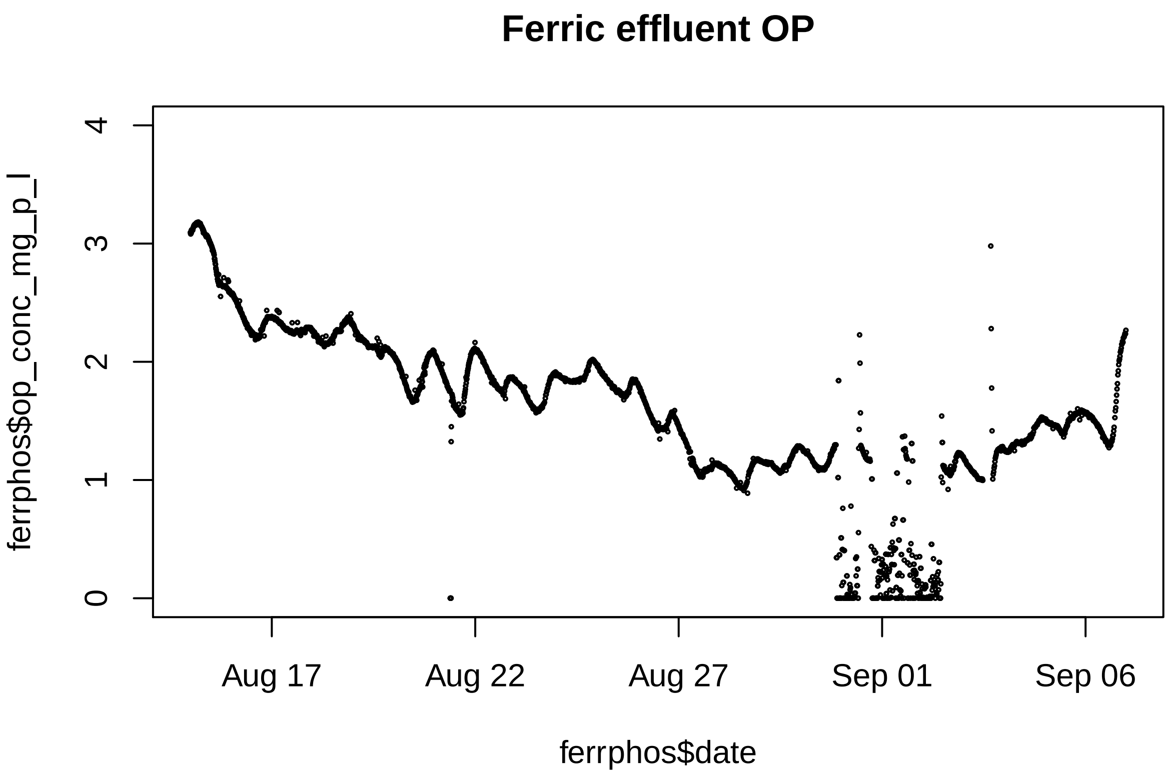
```
# comparing this to above we see 11/18 - 11/25 our merged data doesnt exist. a lot of outliers in this

# what occured between 08/28 - 09/02

ts.plot(ferr_data$op_conc_mg_p_l_hourly)
```



```
plot(ferr_data$date, ferr_data$op_conc_mg_p_l_hourly)
```



```
# under origina time scale we see the same issue occuring

ferrphos <- phosfax_10m %>%
  filter(date >= ymd("2019-08-15")) %>%
  filter(date <= ymd("2019-09-07"))
plot(ferrphos$date, ferrphos$op_conc_mg_p_l,
     cex = .25,
     main = "Ferric effluent OP",
     ylim = c(0,4))
```



```
# numerous values of 0.0004; is this the default error value

# variables that have NA's...dealing with them

final_data$influent_mgd_hourly_avg[4:18]
```

```
##  [1] 16.56255 15.93624 15.29458 13.38298 15.61549  7.59330       NA       NA
##  [9]       NA       NA       NA 15.51021 14.68490 14.35954 13.81169
```

```
final_data$daft_sub_gpm[6:17]
```

```
##  [1] 281.916359 280.675926 282.239102 281.148568         NA         NA
##  [7]         NA         NA         NA   7.339252  67.427778 285.178500
```

```
final_data$daft_sub_gal[8:21]
```

```
##  [1] 282.2887 282.1906       NA       NA       NA       NA       NA       NA
##  [9]       NA       NA       NA 267.3607 281.2609 279.9732
```

```
final_data$abi_mgd[1:12]
```

```
##  [1]        NA        NA        NA        NA        NA        NA 10052.215
##  [8] 11510.382  6216.176  6173.352  6173.352  6173.352
```

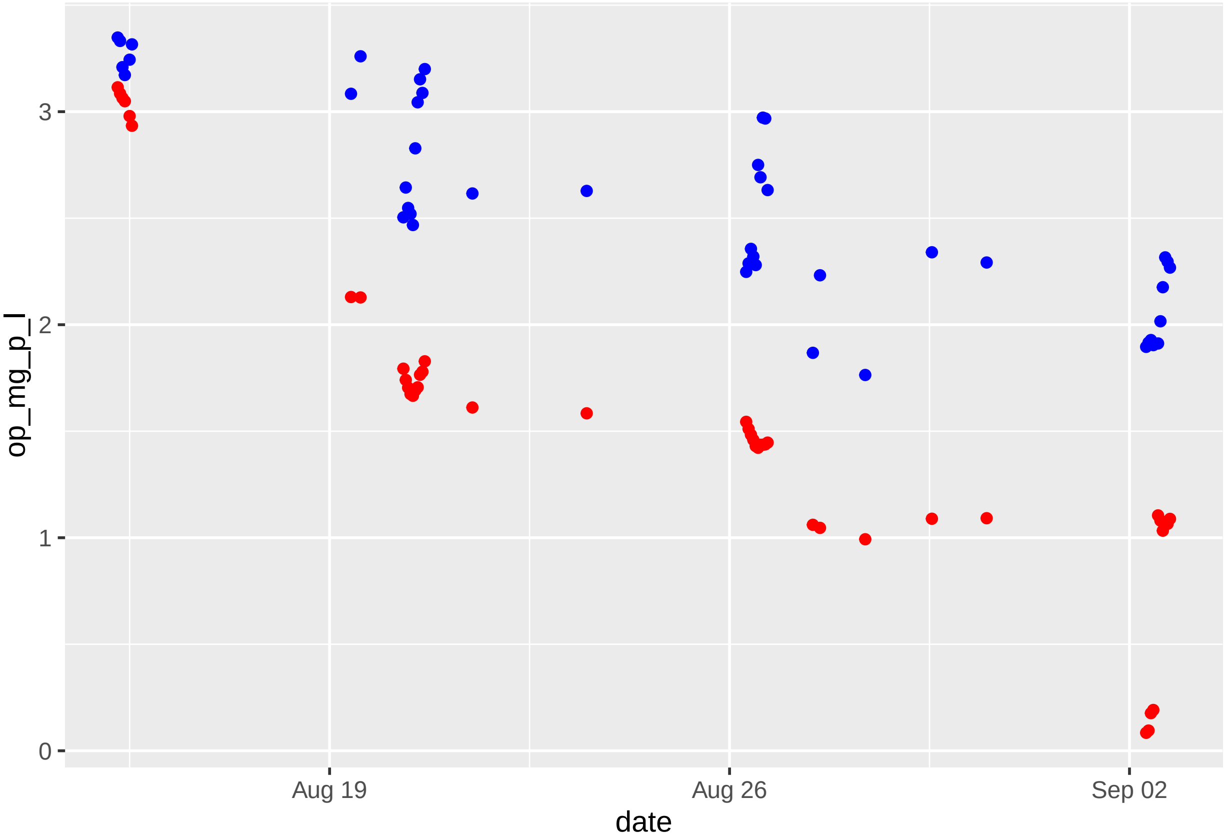
```
final_data$ras_gpm[1:12]
```

```
##  [1]       NA       NA       NA       NA       NA       NA 6531.393 7469.967
##  [9] 5640.932 5655.000 5655.000 5655.000
```

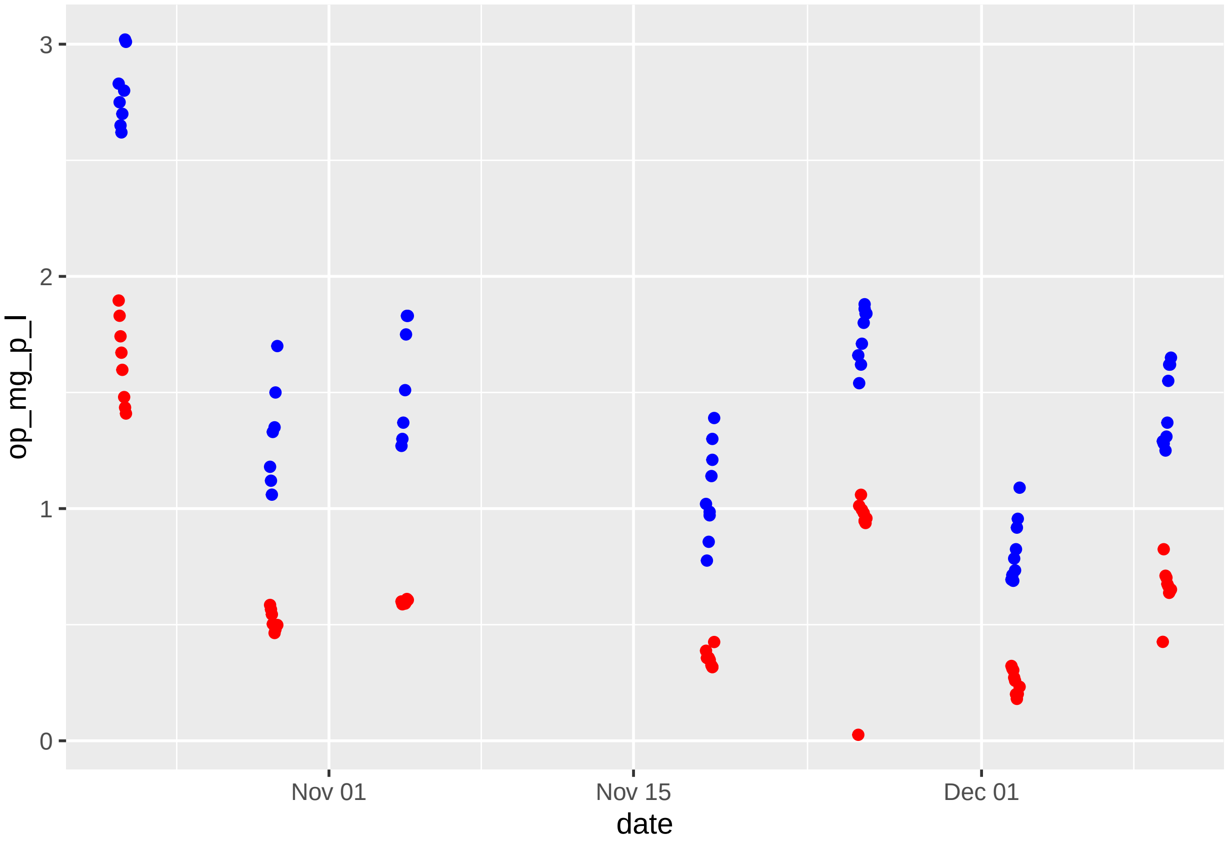
```
final_data$mlws_flow_gpm[7:18]
```

```
##  [1] 294.45824 294.68749 295.53401        NA        NA        NA        NA
##  [8]        NA  12.80994  67.42778 295.33417 295.47741
```

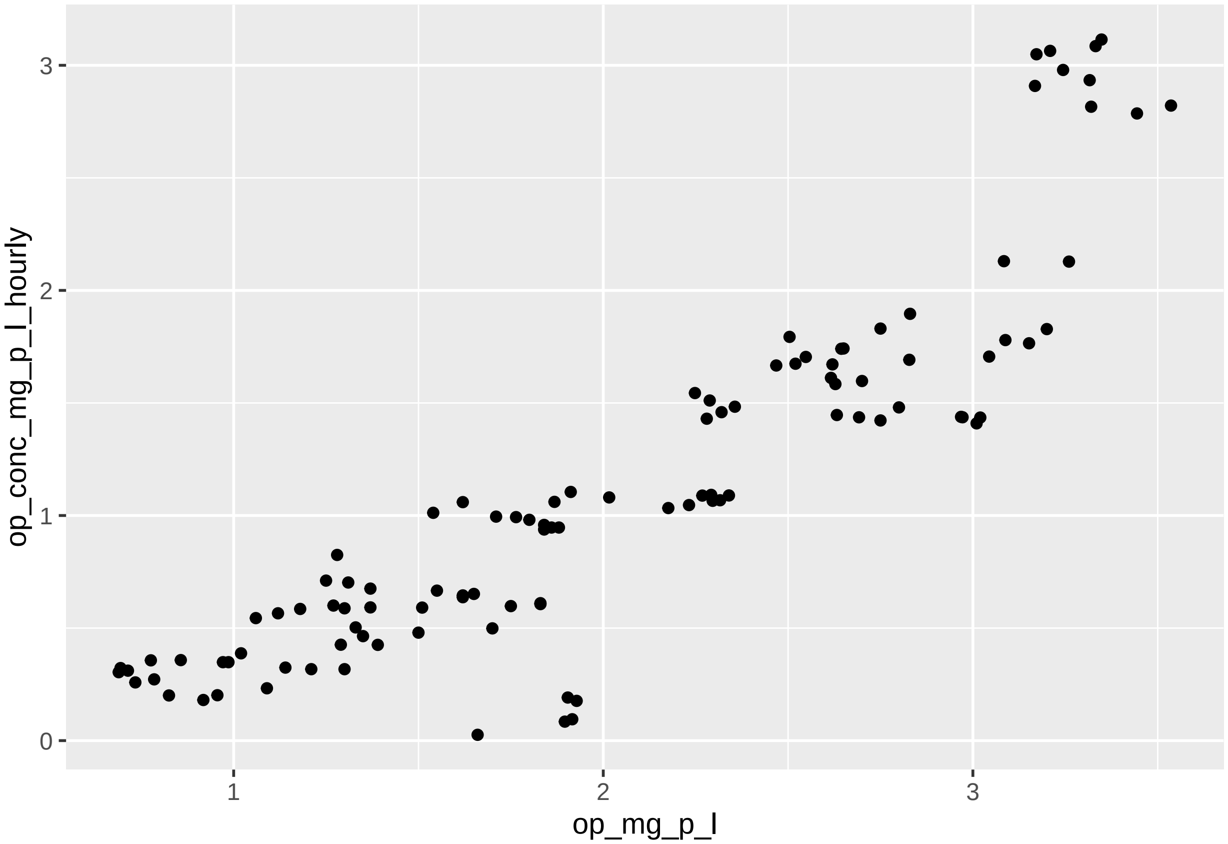
```
# remove? fill with average, moving average?

ferr_data %>% ggplot() +
  geom_point(aes(date, op_mg_p_l), col = "blue") +
  geom_point(aes(date, op_conc_mg_p_l_hourly), col = "red")
```



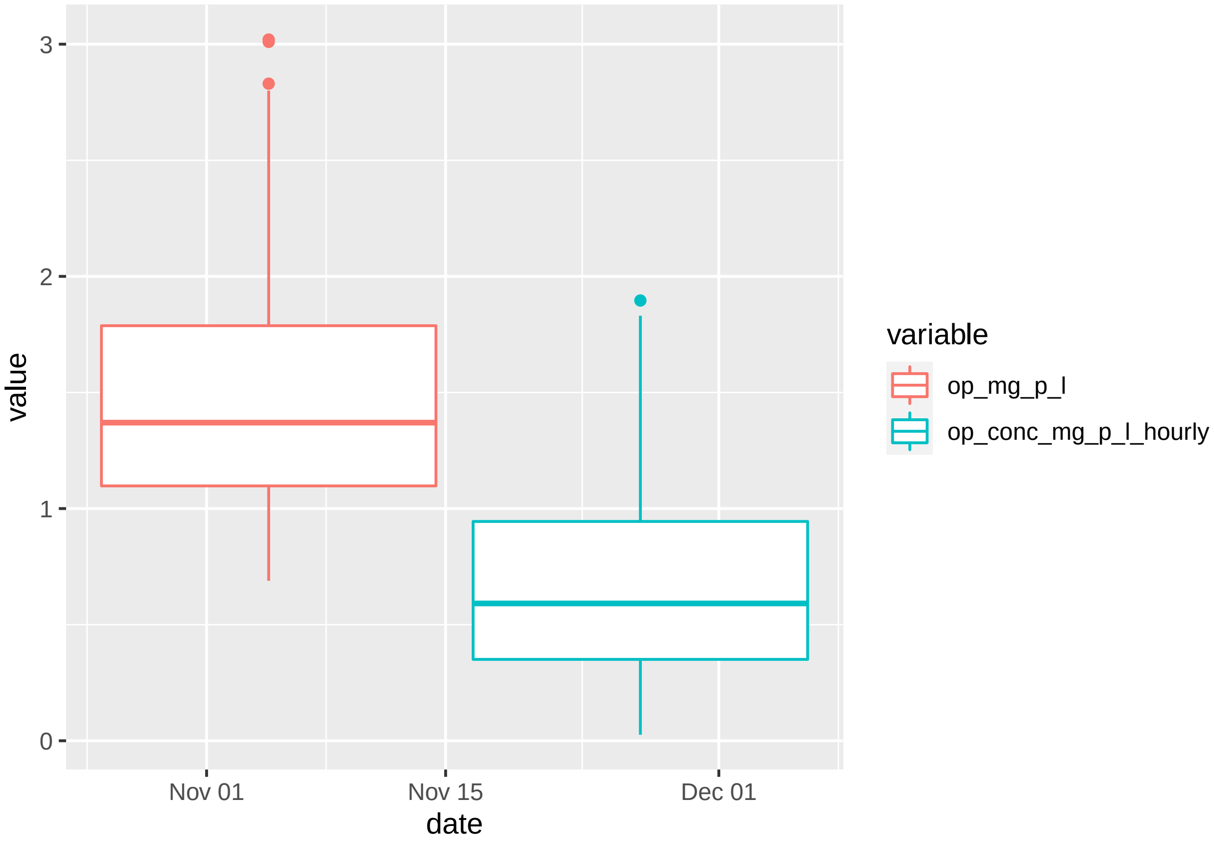
```
alum_data %>% ggplot() +
  geom_point(aes(date, op_mg_p_l), col = "blue") +
  geom_point(aes(date, op_conc_mg_p_l_hourly), col = "red")
```



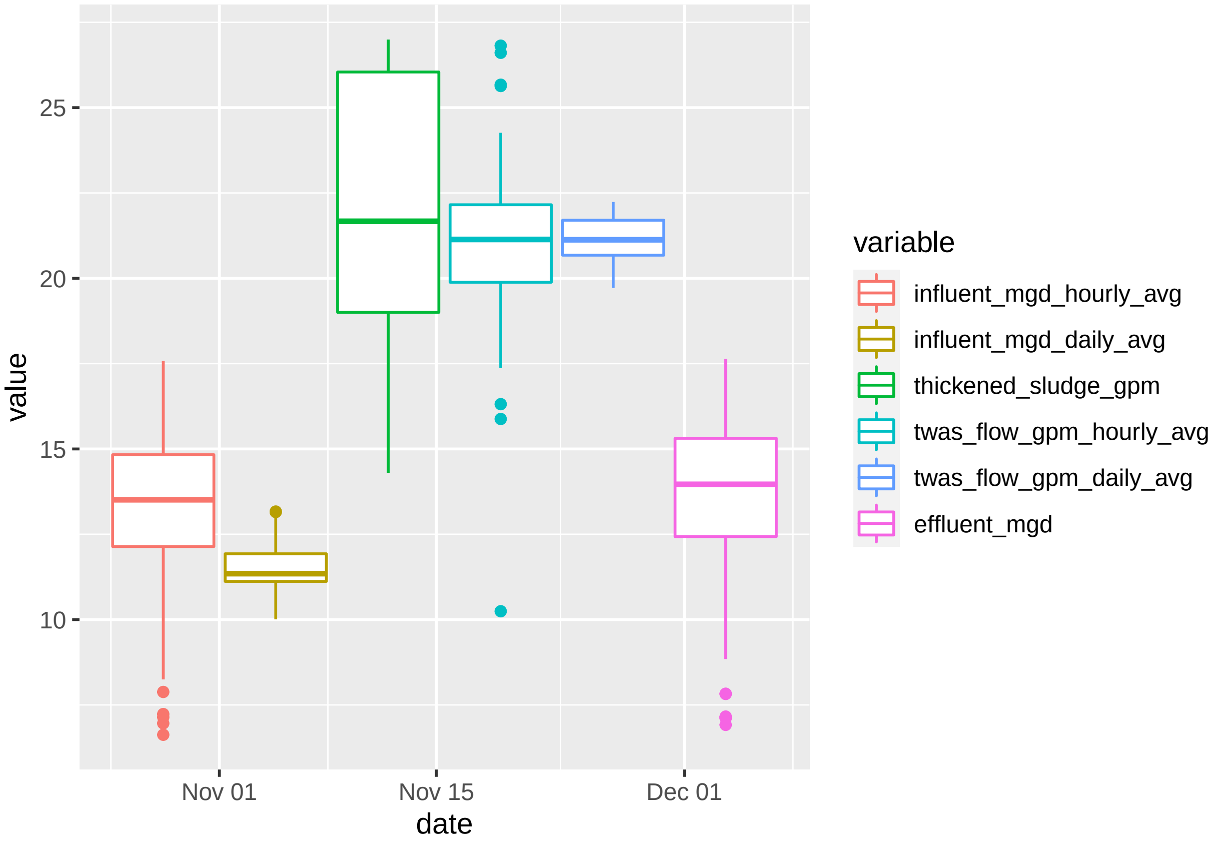
```
# shows linear relationship shifted from effluent being less due to dosing
final_data %>% ggplot() +
  geom_point(aes(op_mg_p_l, op_conc_mg_p_l_hourly))
```



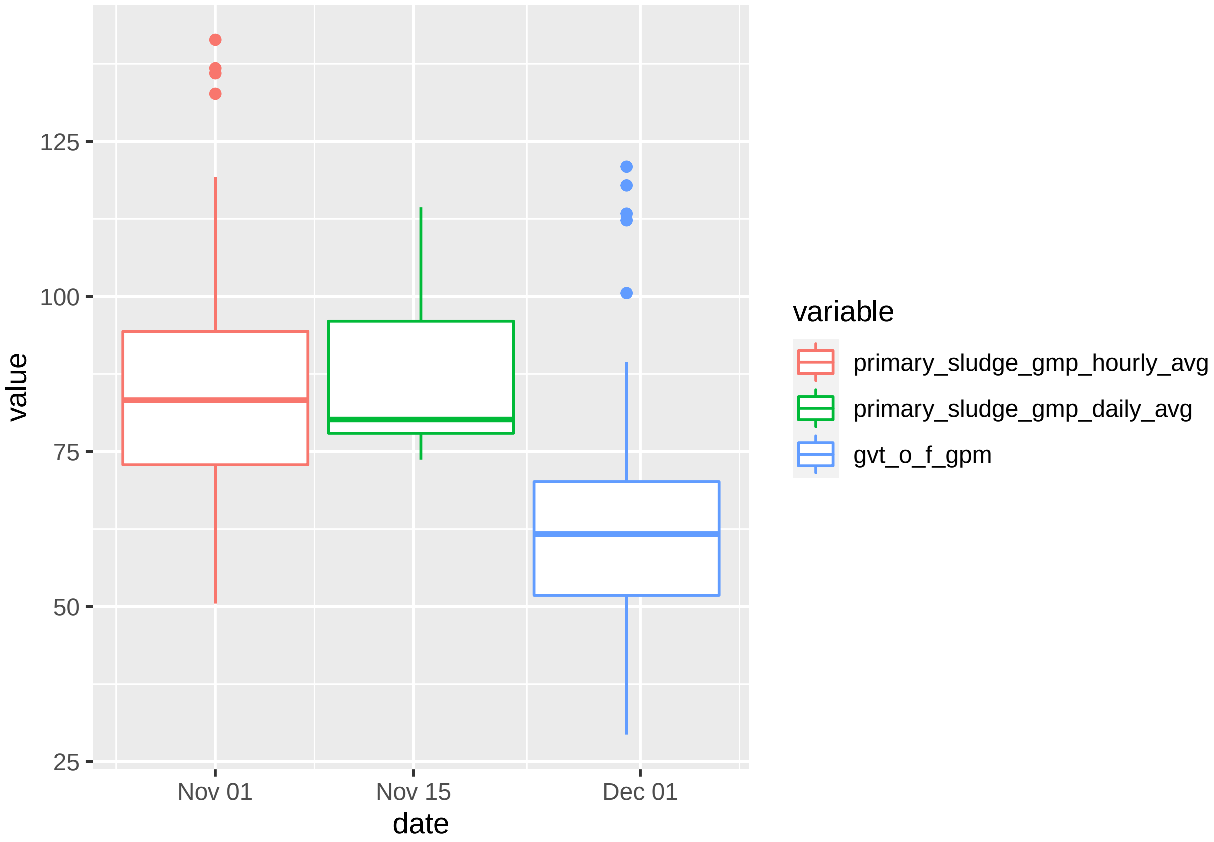
```
alum_melt1 <- melt(alum_data[,c(1,3:4)],id.vars='date', measure.vars=colnames(alum_data[,c(1,3:4)])[-1])
alum_melt2 <- melt(alum_data[,c(1,5:6,10,19:20,22)],id.vars='date', measure.vars=colnames(alum_data[,c(1
alum_melt3 <- melt(alum_data[,c(1,7:8,12)],id.vars='date', measure.vars=colnames(alum_data[,c(1,7:8,12)]
alum_melt4 <- melt(alum_data[,c(1,14:15)],id.vars='date', measure.vars=colnames(alum_data[,c(1,14:15)])
alum_melt5 <- melt(alum_data[,c(1,18)],id.vars='date', measure.vars=colnames(alum_data[,c(1,18)])[-1])
alum_melt6 <- melt(alum_data[,c(1,21)],id.vars='date', measure.vars=colnames(alum_data[,c(1,21)])[-1])
alum_melt7 <- melt(alum_data[,c(1,9,11,13)],id.vars='date', measure.vars=colnames(alum_data[,c(1,9,11,13
alum_melt8 <- melt(alum_data[,c(1,16:17)],id.vars='date', measure.vars=colnames(alum_data[,c(1,16:17)])

#

ggplot(alum_melt1) +
  geom_boxplot(aes(x=date, y=value, color=variable))
```



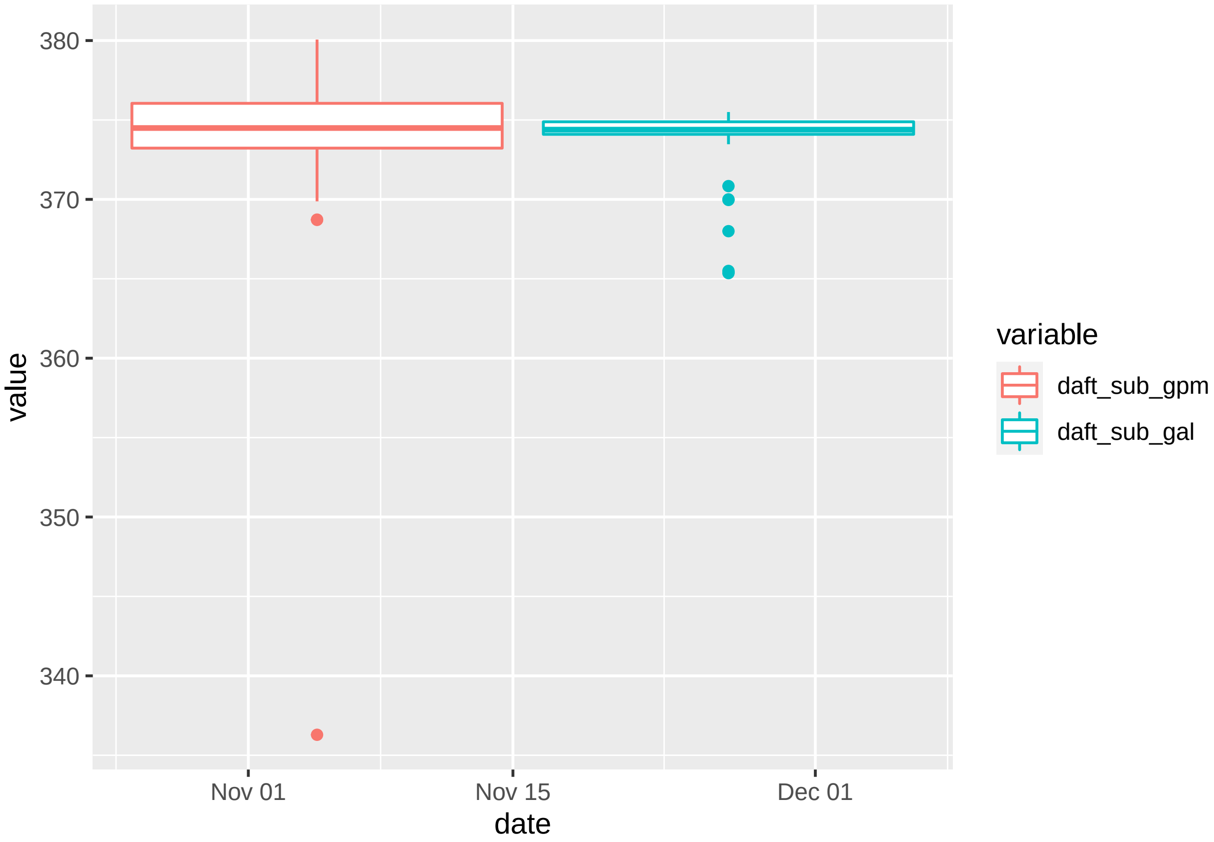
```
#

ggplot(alum_melt2) +
  geom_boxplot(aes(x=date, y=value, color=variable))
```



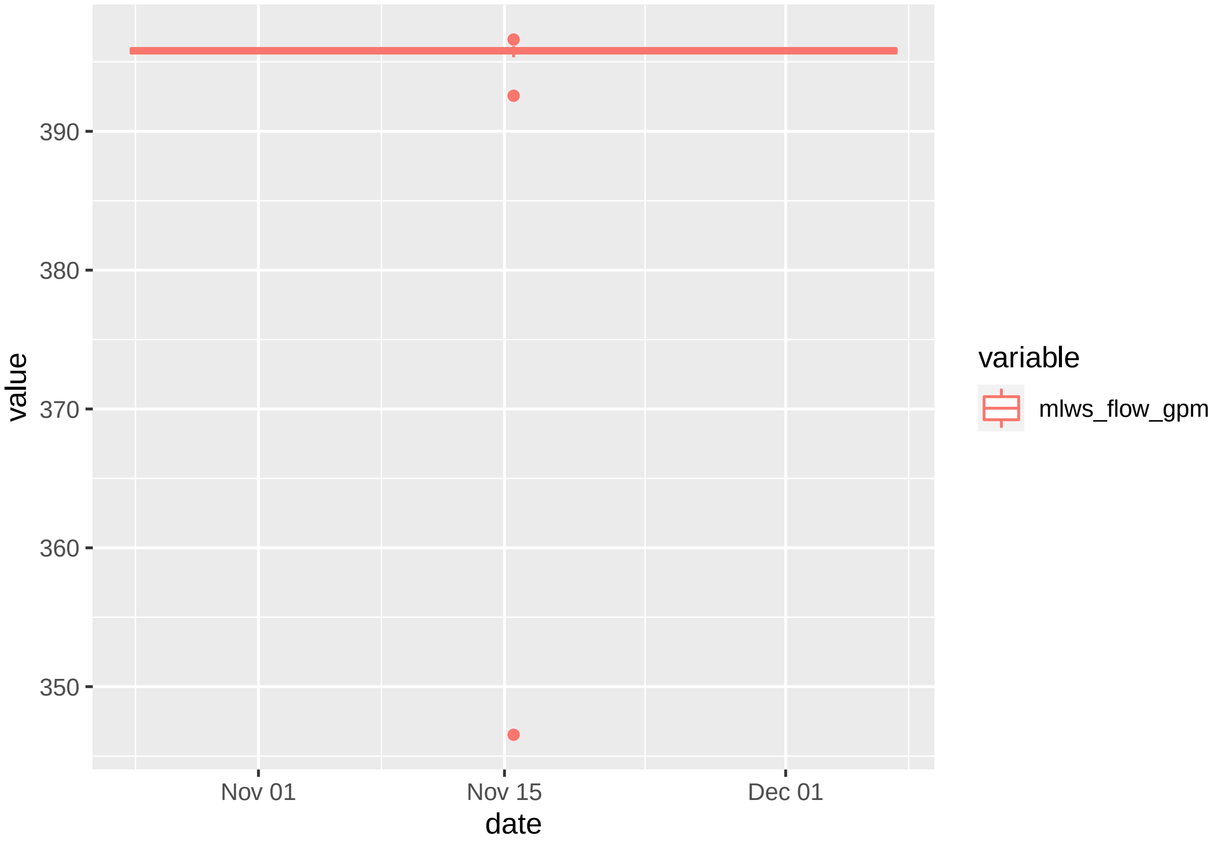
```
#

ggplot(alum_melt3) +
  geom_boxplot(aes(x=date, y=value, color=variable))
```



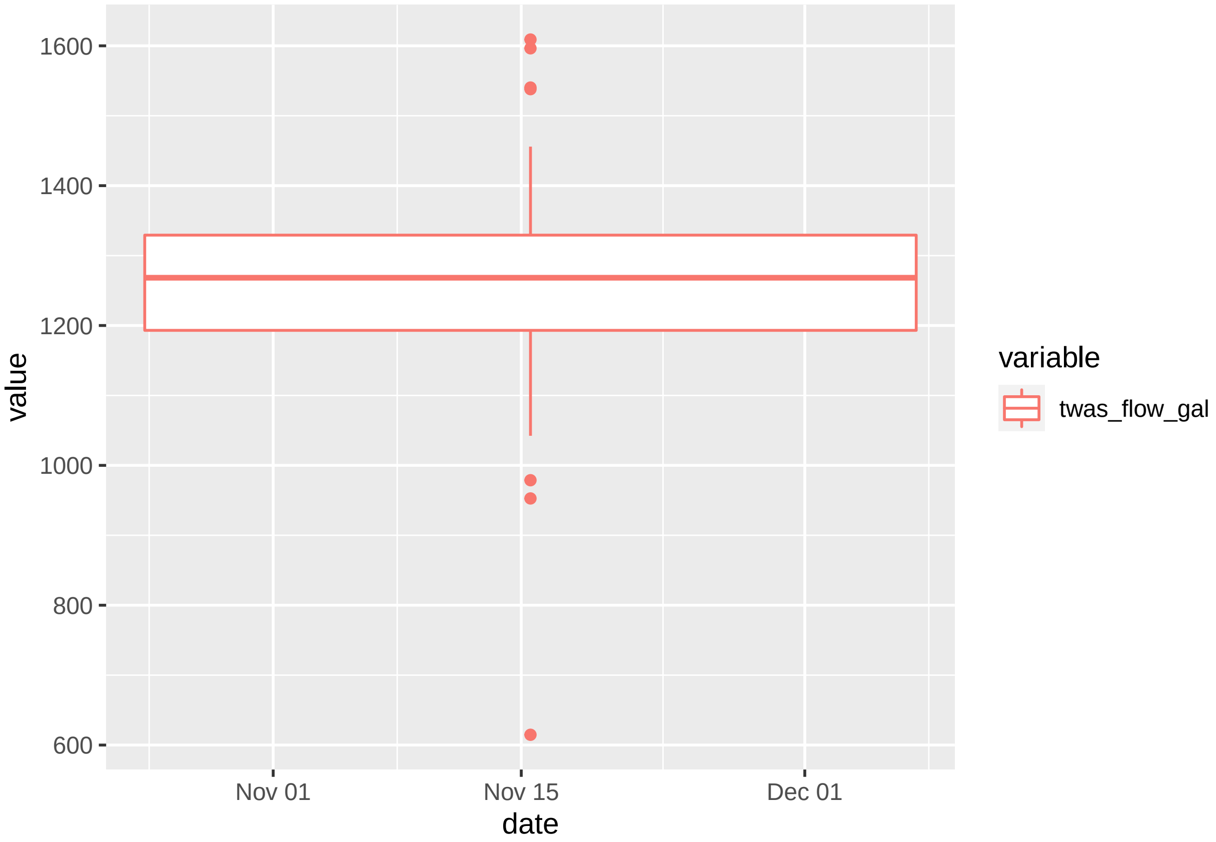
```
#

ggplot(alum_melt4) +
  geom_boxplot(aes(x=date, y=value, color=variable))
```



```
# outlier in melt 5. determine how to approach and then replot

ggplot(alum_melt5) +
  geom_boxplot(aes(x=date, y=value, color=variable))
```

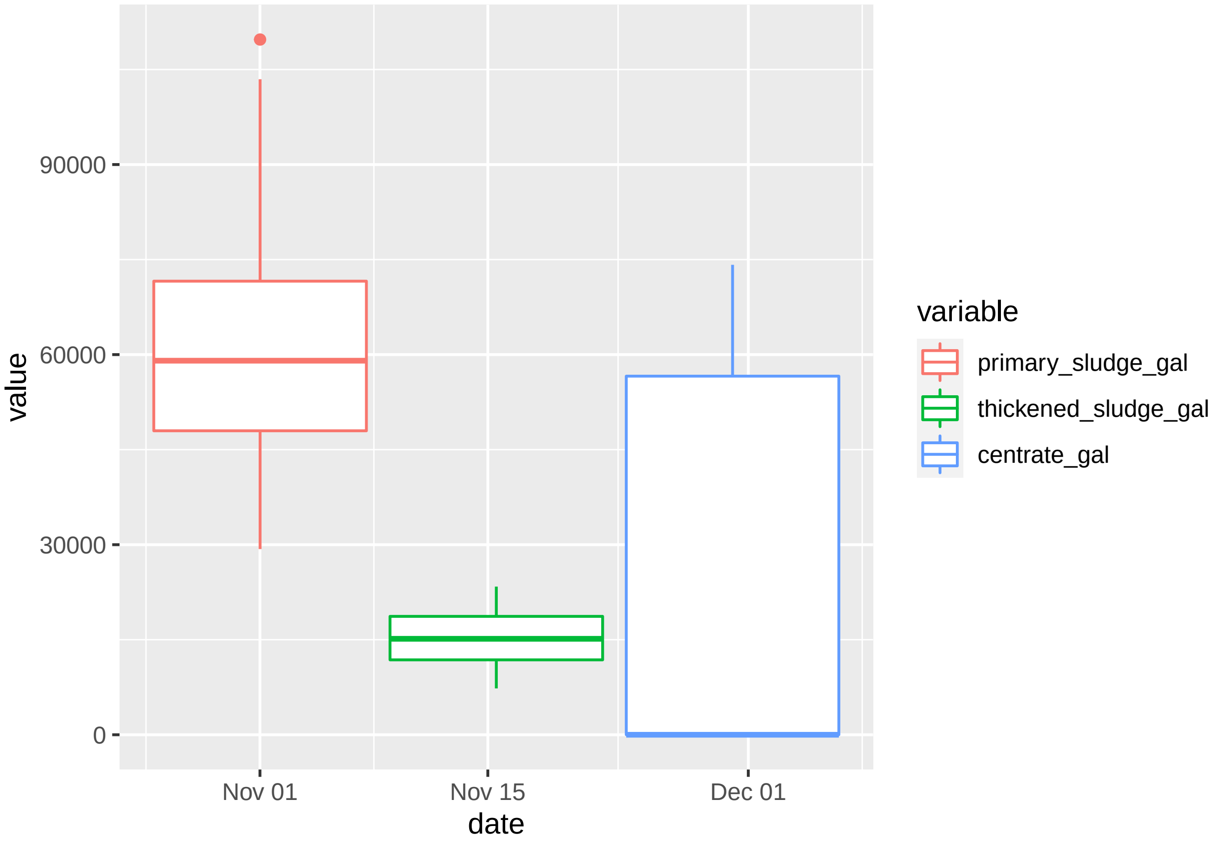


```
#

ggplot(alum_melt6) +
  geom_boxplot(aes(x=date, y=value, color=variable))
```

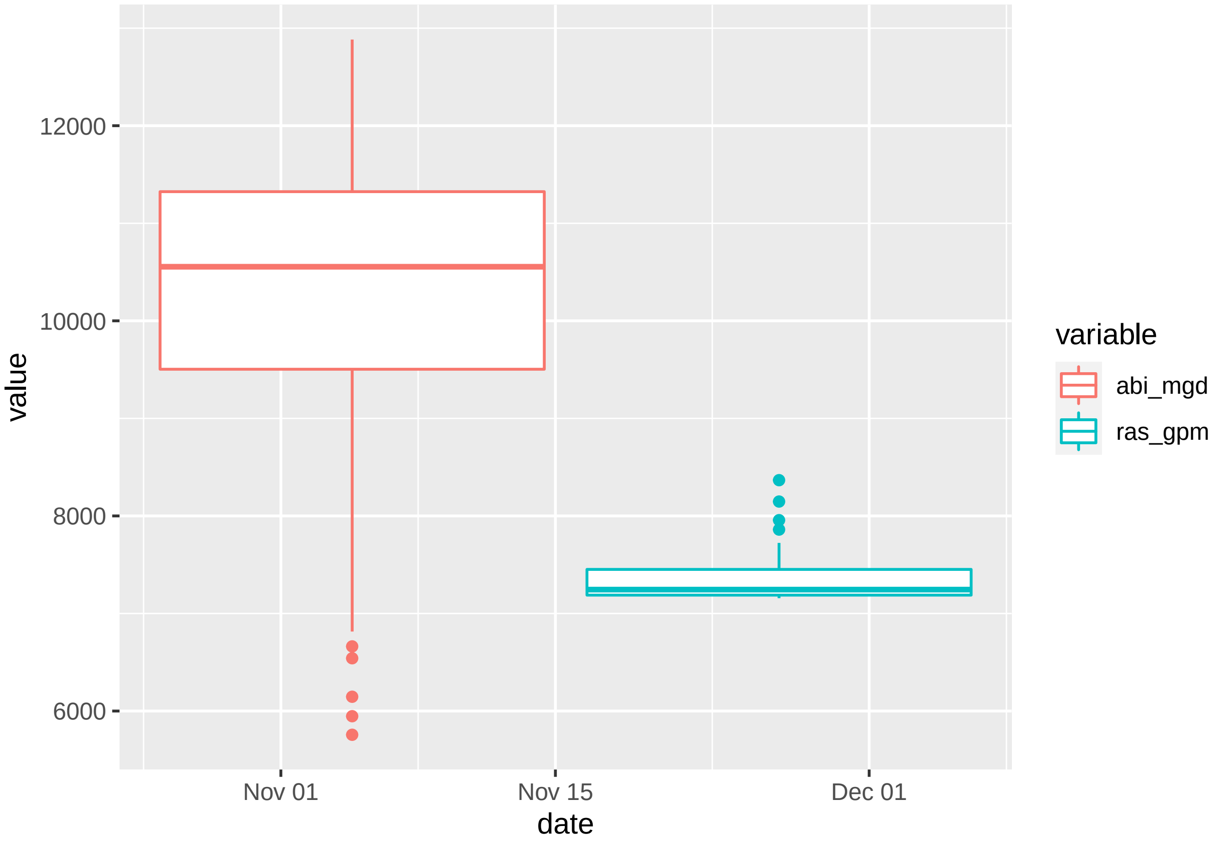


```r
# consider removing centrate_gal

ggplot(alum_melt7) +
  geom_boxplot(aes(x=date, y=value, color=variable))
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



```
ggplot(alum_melt8) +
  geom_boxplot(aes(x=date, y=value, color=variable))
```


12000
10000
8000
6000
value
Nov 01
Nov 15
Dec 01
date
variable
abi_mgd
ras_gpm