Lesson 10: Bringing it all together from import to graph to result

Adam Zabell

1/3/2018

# From Import to Graph

These lessons were designed to build and reinforce each other. It’s possible to write purely linear code, but **purrr** makes the code easier to read, and remember, and use again. Each csv can be read and examined in isolation, but **rbind** and **dplyr** make it possible to aggregate and discover trends. What follows are the steps to replicate the disocvery of one particular problem in the mock data: excessively good R2 data.

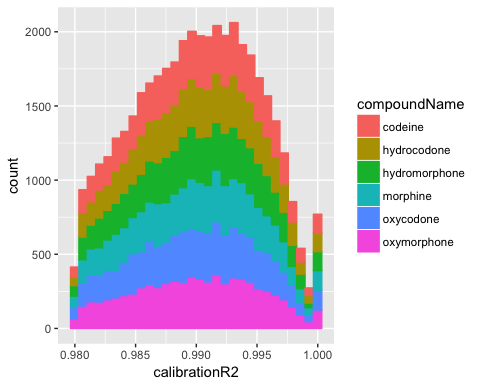
For this lesson, we’re fudging slightly by already knowing what we want to start looking for.

workingDir <- file.path(rootDir,"class\_data")  
oneYearBatches <- list.files(workingDir,pattern="xlsx$") %>%  
 file.path(workingDir,.) %>%  
 map\_dfr(read.xlsx) %>%  
 as\_tibble() %>%  
 type\_convert()

## Parsed with column specification:  
## cols(  
## batchName = col\_character(),  
## instrumentName = col\_character(),  
## compoundName = col\_character(),  
## reviewerName = col\_character(),  
## batchCollectedTimestamp = col\_datetime(format = ""),  
## reviewStartTimestamp = col\_datetime(format = ""),  
## reviewCompleteTimestamp = col\_datetime(format = "")  
## )

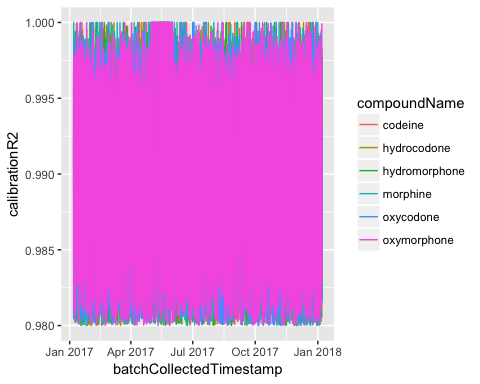
ggplot(oneYearBatches, aes(x=calibrationR2,color=compoundName,fill=compoundName)) +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



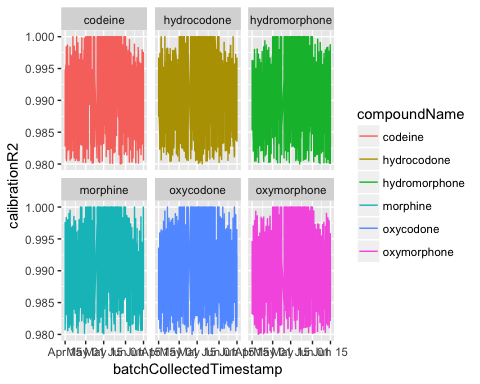
ggplot(oneYearBatches, aes(x=batchCollectedTimestamp,y=calibrationR2,color=compoundName)) +  
 geom\_line()

## Warning in as.POSIXlt.POSIXct(x): unknown timezone 'zone/tz/2017c.1.0/  
## zoneinfo/America/Indiana/Indianapolis'

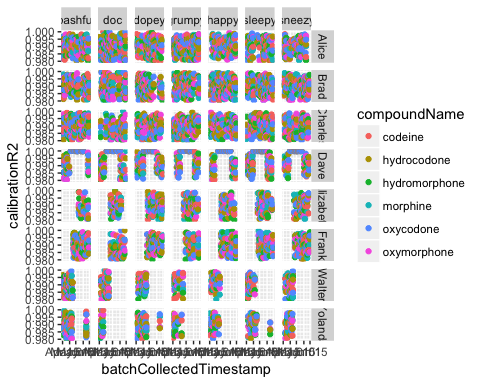


There’s something interesting going on with the R2 values in the month of May, where a large number of them report a value of 1.0 – a perfect fit. Let’s focus on that month, and spread out the data so we can clarify whether it’s all compounds or just oxymorphone (the magenta color on top).

mayPlot <- oneYearBatches %>%   
 filter(batchCollectedTimestamp > ymd("2017-04-15"), batchCollectedTimestamp < ymd("2017-06-15")) %>%  
 ggplot(aes(x=batchCollectedTimestamp,y=calibrationR2,color=compoundName))   
mayPlot +  
 geom\_line() +  
 facet\_wrap(~ compoundName)



mayPlot +  
 geom\_point() +  
 facet\_grid(reviewerName ~ instrumentName)



Whatever is going on, it looks like reviewer ‘Dave’ is the only person it is happening to.

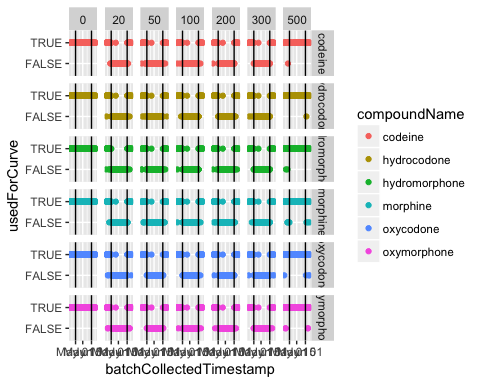
# Graph to Result

Based on the batch-level data, we can see that ‘Dave’ – and apparently only Dave – has perfect R2 values on every batch of data he reviewed throughout the month of May. Digging deeper will require merging information from the batch level with information at the sample (and possibly peak) level.

oneYearSamples <- list.files(workingDir,pattern="csv$") %>%  
 file.path(workingDir,.) %>%  
 map\_dfr(read\_csv)  
davesData <- oneYearSamples %>%  
 left\_join(select(oneYearBatches,-calibrationSlope,-calibrationIntercept)) %>%  
 filter(batchCollectedTimestamp > ymd("2017-04-20"),   
 batchCollectedTimestamp < ymd("2017-06-10"),  
 sampleType == "standard",  
 reviewerName == "Dave")

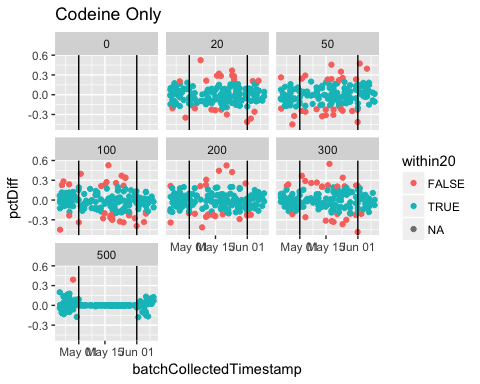
The following plots of davesData provide compelling evidence for what happened: Dave unselected the middle five calibrators in order to draw a straight line and maximize the R2 term.

davesData %>% ggplot(aes(x=batchCollectedTimestamp,y=usedForCurve,color=compoundName)) +  
 geom\_point() +  
 facet\_grid(compoundName ~ expectedConcentration) +  
 geom\_vline(xintercept = as.numeric(as\_datetime(c("2017-05-01","2017-06-01"))),linetype=1,colour="black")



davesData %<>% mutate(pctDiff = (concentration - expectedConcentration) / expectedConcentration,  
 within20 = abs(pctDiff) <= 0.2)  
davesData %>% filter(compoundName == "codeine") %>%  
 ggplot(aes(x=batchCollectedTimestamp,y=pctDiff,color=within20)) +  
 geom\_point() +  
 facet\_wrap(~ expectedConcentration) +  
 ggtitle("Codeine Only") +  
 geom\_vline(xintercept = as.numeric(as\_datetime(c("2017-05-01","2017-06-01"))),linetype=1,colour="black")

## Warning: Removed 164 rows containing missing values (geom\_point).



The second plot shows that calibrators were dropped regardless of whether they would be within 20% of the expected concentration, suggesting that they were dropped for some other reason. The data does not say why ‘Dave’ did this, but there are a couple of good guesses here which revolve around training.

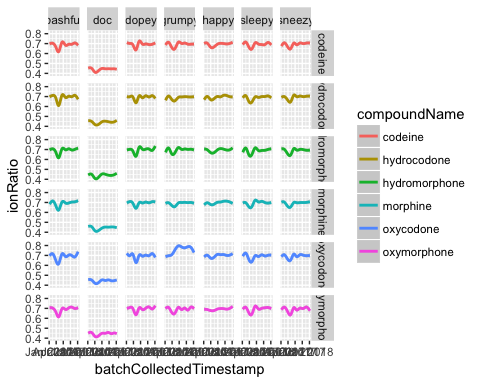
We intentionally included several other issues within the database, which will require aggregation and plotting to discover.

Exercise : Revealing an ion ratio problem  
Ion ratios are particularly sensitive to instrument conditions, because the secondary fragmentation pattern requires a highly tuned spectrometer and the selection of two ions which don’t change much when the mass spec flucturates. We can look for both outlier spikes and stability trends, and separate those out across instruments, compounds, and sample types. There is one obviously outlying instrument, which is worth investigating later. Excluding that instrument, which compound+instrument shows an outlier pattern? Is there a problem here? What additional variables would help besides the ones captured in the sample data.frame?

# 1 #  
oneYearSamples %>%   
 left\_join(oneYearBatches) %>%  
 ggplot(aes(x=batchCollectedTimestamp,y=ionRatio,color=compoundName)) +  
 geom\_smooth() +  
 facet\_grid(compoundName ~ instrumentName)

## Joining, by = c("batchName", "compoundName")

## `geom\_smooth()` using method = 'gam'



# 2 #