pH Prediction Exploration and Transformation

Heather Geiger

Libraries

```
library(ggplot2)
library(tidyr)
library(dplyr)
library(gridExtra)
library(VennDiagram)
library(corrplot)
```

Exploratory analysis

Read in data.

Read in CSV files, which are just the CSV versions of the original Excel sheets.

```
training <- read.csv("pH_prediction_training_data.csv",header=TRUE,stringsAsFactors=F
ALSE)
test <- read.csv("pH_prediction_test_data.csv",header=TRUE,stringsAsFactors=FALSE)
```

Set aside target (PH) from training into a separate variable. Remove entirely from test.

Then, combine training and test into one data frame.

```
training_target <- training$PH

training <- training[,setdiff(colnames(training),"PH")]
test <- test[,colnames(training)]

alldata <- data.frame(rbind(training,test),
    Data = rep(c("Training","Test"),times=c(nrow(training),nrow(test))),
    stringsAsFactors=FALSE)
```

Within variables

Unique values per variable

One useful quick thing to check can be the number and type of unique values per variable.

If for example there are only 12 unique values, and those values are 1-12, then you know that the variable may represent counts rather than a continuous numeric range.

```
unique_per_var <- apply(alldata,2,function(x)length(unique(x)))
```

```
## [1] "List of number of unique values per variable:"
```

```
unique_per_var[order(unique_per_var)]
```

```
##              Data       Brand.Code Pressure.Setpoint     Bowl.Setpoint 
##                 2                5               10                12 
##          PSC.CO2   Pressure.Vacuum         PSC.Fill     Air.Pressurer 
##                14               17               33                35 
##     Hyd.Pressure4         Carb.Rel         Alch.Rel      Temperature 
##                42               43               54                60 
##          Density      Balling.Lvl      Fill.Ounces      Carb.Volume 
##                80               84               94               105 
##     Carb.Pressure     Fill.Pressure        Carb.Temp              PSC 
##               107              112              125               132 
##    Carb.Pressure1     Hyd.Pressure3     Hyd.Pressure2          Balling 
##               141              199              218               229 
##     Hyd.Pressure1      Filler.Speed      Filler.Level     Oxygen.Filler 
##               251              260              297               351 
##         PC.Volume       Usage.cont          Mnf.Flow         Carb.Flow 
##               467              490              510               558 
##               MFR 
##               620
```

```
## [1] "Count of unique values per variable for select variables:"
```

```
vars_with_relatively_few_unique <- colnames(alldata)[order(unique_per_var)[2:6]]

apply(alldata[,vars_with_relatively_few_unique],2,function(x)table(x,useNA="ifany"))
```

```
## $Brand.Code
## x
##         A    B    C    D
##  128  328 1368  335  679
##
## $Pressure.Setpoint
## x
## 44.0 45.2 46.0 46.4 46.6 46.8 48.0 50.0 52.0 <NA>
##  105    1 1450    1    1    1  143 1108   14   14
##
## $Bowl.Setpoint
## x
##   70   80   90  100  110  120  122  126  130  134  140 <NA>
##  108  111  469  124  486 1446    1   10   58    2   20    3
##
## $PSC.CO2
## x
## 0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 <NA>
##  128  677  736  542  283  158   95   54   41   20   20   22   18   44
##
## $Pressure.Vacuum
## x
## -3.6 -3.8 -4.0 -4.2 -4.4 -4.6 -4.8 -5.0 -5.2 -5.4 -5.6 -5.8 -6.0 -6.2 -6.4
##   14   45   71  143  114  101  210  353  364  510  357  308  151   65   22
## -6.6 <NA>
##    9    1
```

Looks like brand is blank for a significant number of observations.

In Pressure.Setpoint, the vast majority of values are multiples of 2. So the ones that are between multiples (45.2, 46.4, 46.6, and 46.8) are unusual.

Similarly in Bowl.Setpoint, 122, 126, and 134 are unusual in being not multiples of 10.

Zeros per variable

Sometimes variables will have a distribution where there are many zeros, but then the nonzero part of the distribution looks like a relatively standard continuous numeric variable.

Let’s see if this is the case for any of the variables here.

```
num_zeros_per_var <- rep(0,times=ncol(alldata))

for(i in 1:ncol(alldata))
{
    num_zeros_per_var[i] <- length(which(is.na(alldata[,i]) == FALSE & alldata[,i] ==
0))
}

names(num_zeros_per_var) <- colnames(alldata)

num_zeros_per_var[num_zeros_per_var > 0]
```

```
##      PSC.Fill       PSC.CO2 Hyd.Pressure1 Hyd.Pressure2 Hyd.Pressure3
##            12           128           928           854           898
##     Carb.Flow   Balling.Lvl
##             1            10
```

We find the Hyd.Pressure variables have a very large number of observations equal to 0.

Repeated values per variable

Any notable often-repeated nonzero values?

```
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
```

```
nonzero_mode_per_var <- data.frame(Variable = colnames(alldata),
    Num.unique = unique_per_var,
    Most.common.nonzero.value = rep(NA,times=ncol(alldata)),
    Num.obs = rep(NA,times=ncol(alldata)),
    stringsAsFactors=FALSE,check.names=FALSE,
    row.names=colnames(alldata))

for(i in 1:ncol(alldata))
{
    mymode <- Mode(alldata[alldata[,i] != 0,i])
    nonzero_mode_per_var$Most.common.nonzero.value[i] <- mymode
    nonzero_mode_per_var$Num.obs[i] <- length(which(alldata[,i] == mymode))
}
```

```
nonzero_mode_per_var <- nonzero_mode_per_var[nonzero_mode_per_var$Num.unique >= 30,]
nonzero_mode_per_var <- nonzero_mode_per_var[which(is.na(nonzero_mode_per_var$Most.co
mmon.nonzero.value) == FALSE),]

nonzero_mode_per_var[order(nonzero_mode_per_var$Num.obs,decreasing=TRUE),2:4]
```

```
##                Num.unique Most.common.nonzero.value Num.obs
## Mnf.Flow              510                       -100     671
## Fill.Pressure         112                         46     452
## Air.Pressurer          35                      142.6     436
## Alch.Rel               54                       6.52     421
## Temperature            60                       65.4     299
## Carb.Rel               43                       5.36     288
## Hyd.Pressure4          42                         98     270
## Density                80                       0.94     254
## Balling               229                      1.548     239
## PSC.Fill               33                       0.14     224
## Oxygen.Filler         351                     0.0026     223
## Balling.Lvl            84                       1.36     223
## Filler.Level          297                        120     193
## Filler.Speed          260                       4010     185
## Hyd.Pressure2         218                        0.2     127
## Fill.Ounces            94                      23.98     117
## Carb.Volume           105                5.306666667     102
## Hyd.Pressure3         199                       -1.2      82
## Carb.Pressure1        141                      124.2      74
## Carb.Temp             125                        140      70
## Carb.Pressure         107                       68.2      68
## PSC                   132                      0.056      65
## Carb.Flow             558                         44      56
## Usage.cont            490                      23.88      46
## Hyd.Pressure1         251                       12.4      28
```

Get table for top most common values for a few of these variables.

```
for(var in c("Mnf.Flow","Hyd.Pressure2","Hyd.Pressure3"))
{
    print(var)
    freq_per_var <- data.frame(table(alldata[,var]))
    print(freq_per_var[order(freq_per_var$Freq,decreasing=TRUE)[1:5],])
}
```

```
## [1] "Mnf.Flow"
##       Var1 Freq
## 2     -100  671
## 1   -100.2  636
## 3      0.2   92
## 234  133.4   10
## 242    135   10
## [1] "Hyd.Pressure2"
##     Var1 Freq
## 2      0  854
## 3    0.2  127
## 134 33.4   40
## 139 34.4   38
## 142   35   36
## [1] "Hyd.Pressure3"
##     Var1 Freq
## 3      0  898
## 2   -1.2   82
## 117 31.8   53
## 116 31.6   39
## 119 32.2   39
```

We find that Mnf.Flow most common values are -100.2/-100 (together over half of observations), then 0.2 (92 observations).

In addition to the zeros, Hyd.Pressure 2 and 3 each have another repeated value (0.2 for Hyd.Pressure2, -1.2 for Hyd.Pressure3).

Missing values

How many missing values do we find per variable?

```
missing_per_var <- rep(0,times=ncol(alldata))

for(i in 1:ncol(alldata))
{
    missing_per_var[i] <- length(which(is.na(alldata[,i]) == TRUE))
}

names(missing_per_var) <- colnames(alldata)

missing_per_var[order(missing_per_var)]
```

```
##        Brand.Code              Data   Pressure.Vacuum     Air.Pressurer
##                 0                 0                 1                 1
##       Balling.Lvl         Mnf.Flow        Carb.Flow           Density
##                 1                 2                 2                 2
##           Balling     Bowl.Setpoint       Usage.cont       Carb.Volume
##                 2                 3                 7                11
##     Hyd.Pressure1         Alch.Rel         Carb.Rel Pressure.Setpoint
##                11                12                12                14
##     Oxygen.Filler     Hyd.Pressure2     Hyd.Pressure3       Temperature
##                15                16                16                16
##      Filler.Level     Fill.Pressure         PSC.Fill     Carb.Pressure
##                22                24                26                27
##         Carb.Temp     Hyd.Pressure4    Carb.Pressure1               PSC
##                27                34                36                38
##         PC.Volume       Fill.Ounces          PSC.CO2      Filler.Speed
##                43                44                44                67
##               MFR
##               243
```

We find most variables are missing in at least a few observations. Including, brand code should maybe have some as well, once we convert the blanks to NA.

How many missing variables do we tend to find in a given observation?

```
missing_per_obs <- apply(alldata,1,function(x)length(which(is.na(x) == TRUE)))

table(missing_per_obs)
```

```
## missing_per_obs
##    0    1    2    3    4    5    6    7    8   11   14
## 2335  311  129   39   13    3    3    1    2    1    1
```

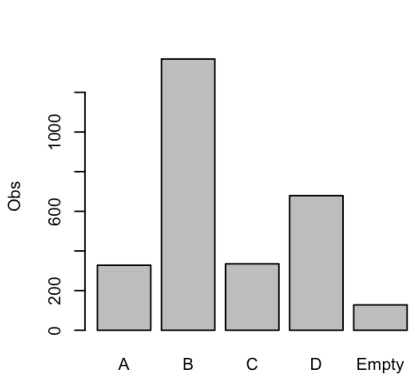
Most observations are only missing data for at most one or two variables, which is good.

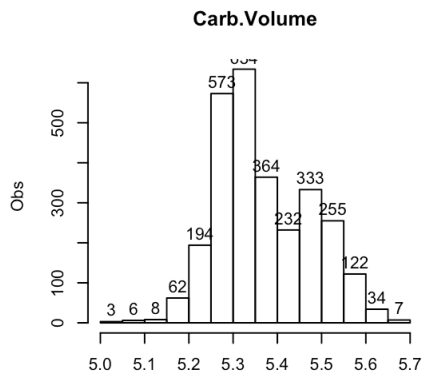
Overall distribution of values

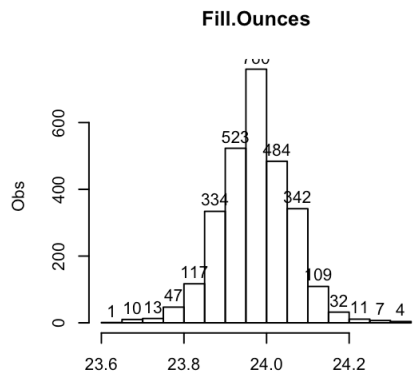
Let’s make a simple histogram (or barplot if categorical) per variable.

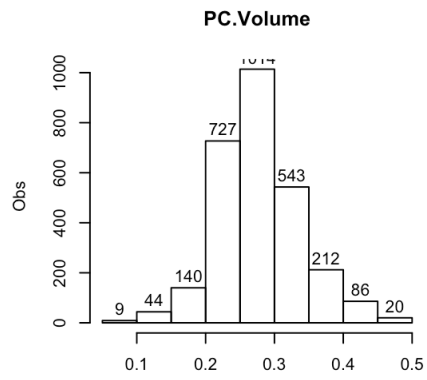
```
alldata$Brand.Code[alldata$Brand.Code == ""] <- "Empty"
```

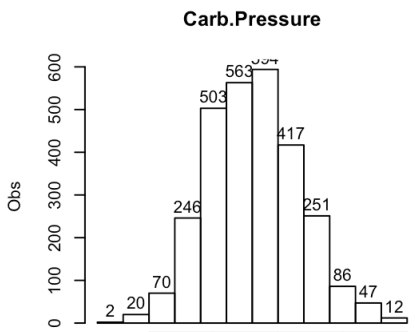
```
par(mfrow=c(8,4))

barplot(table(alldata$Brand.Code),ylab="Obs")

for(var in setdiff(colnames(alldata),c("Data","Brand.Code")))
{
    hist(alldata[,var],ylab="Obs",xlab="",main=var,labels=TRUE)
}
```

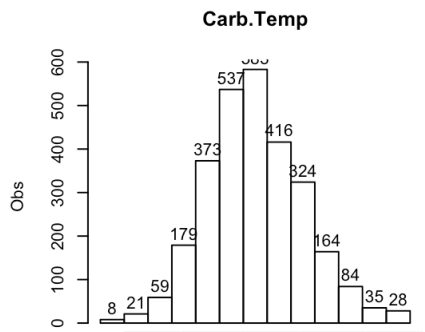


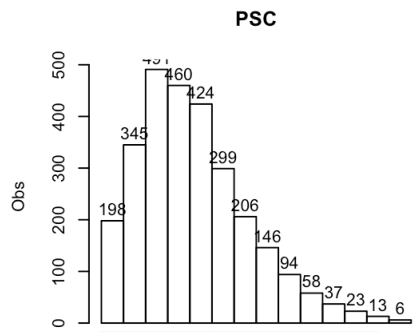


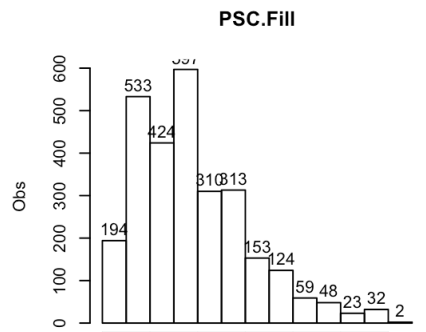


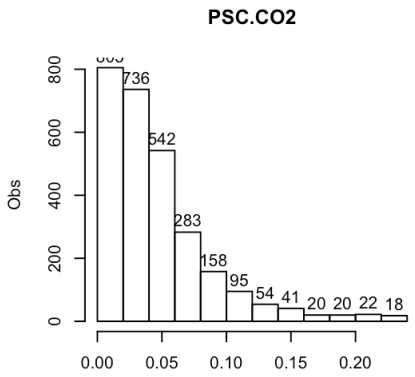


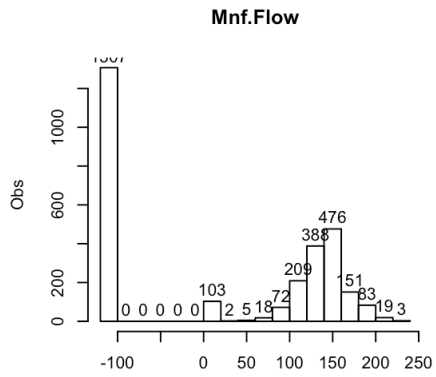


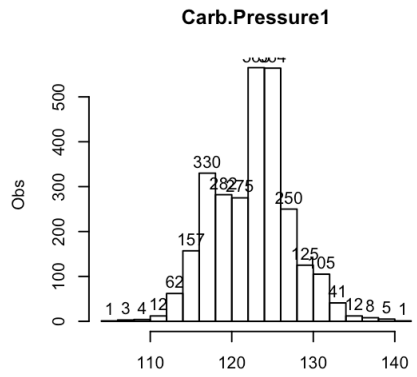


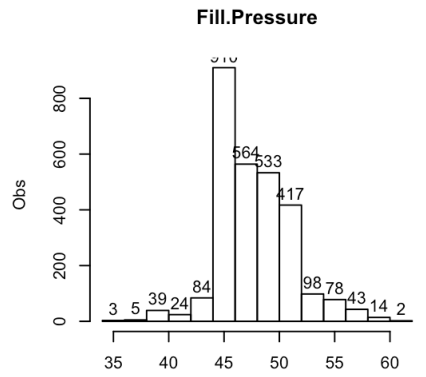


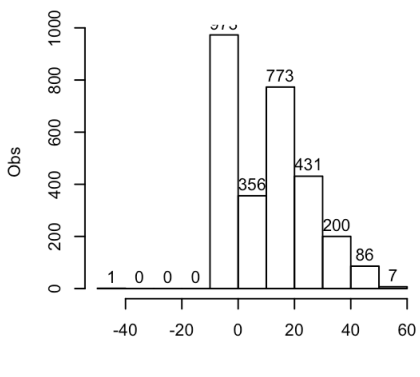


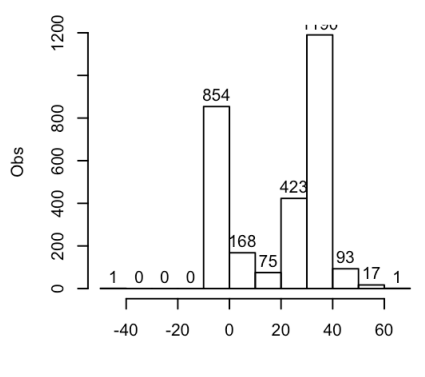


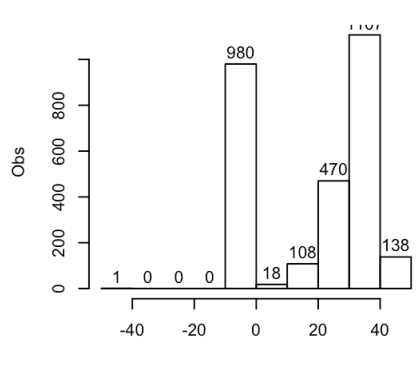


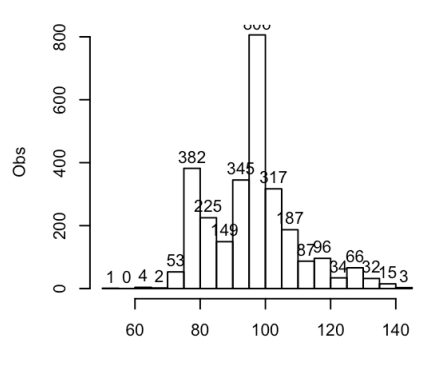


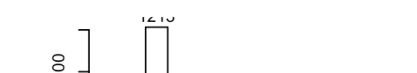


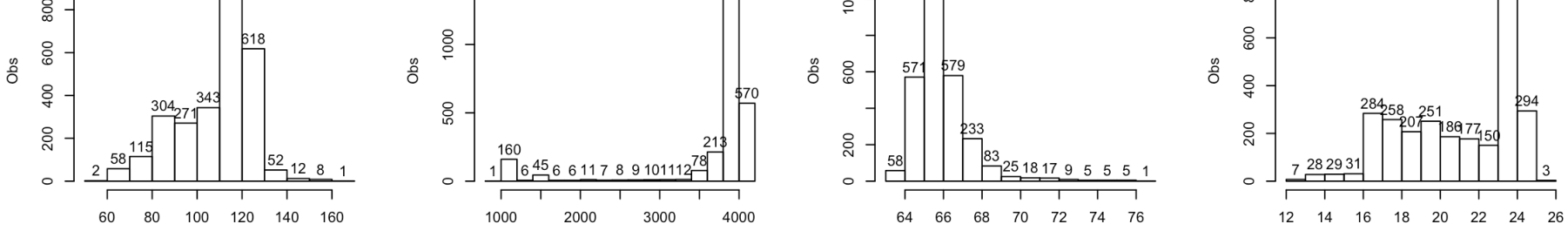


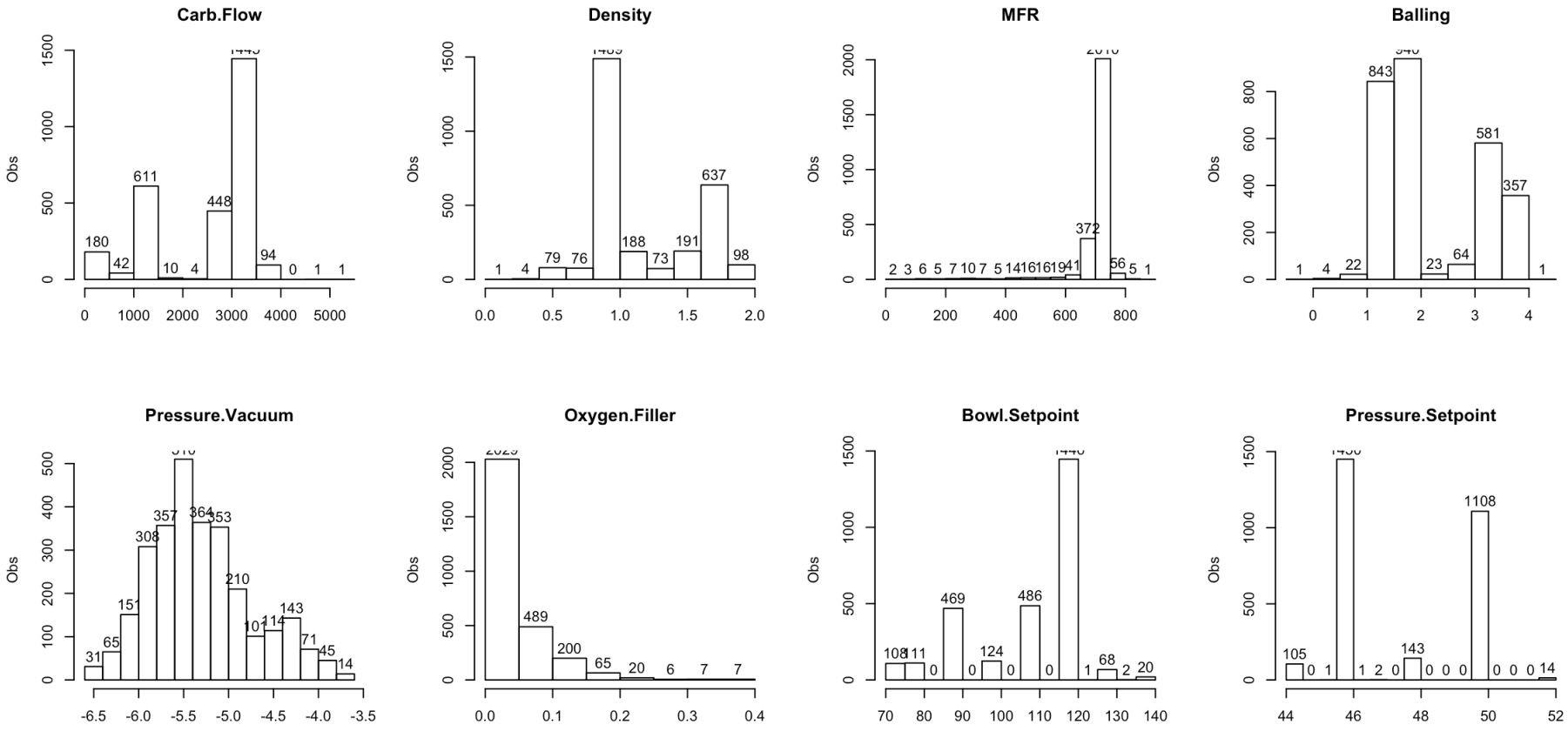


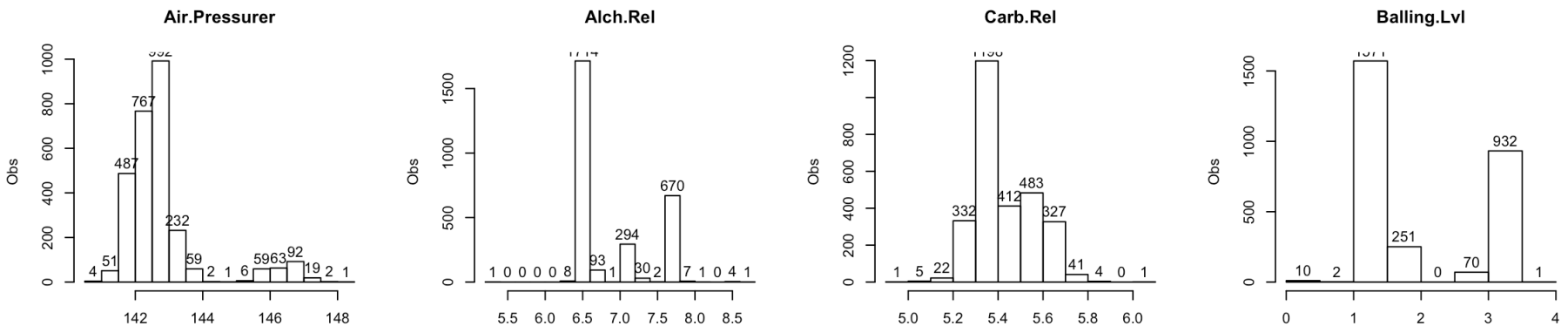








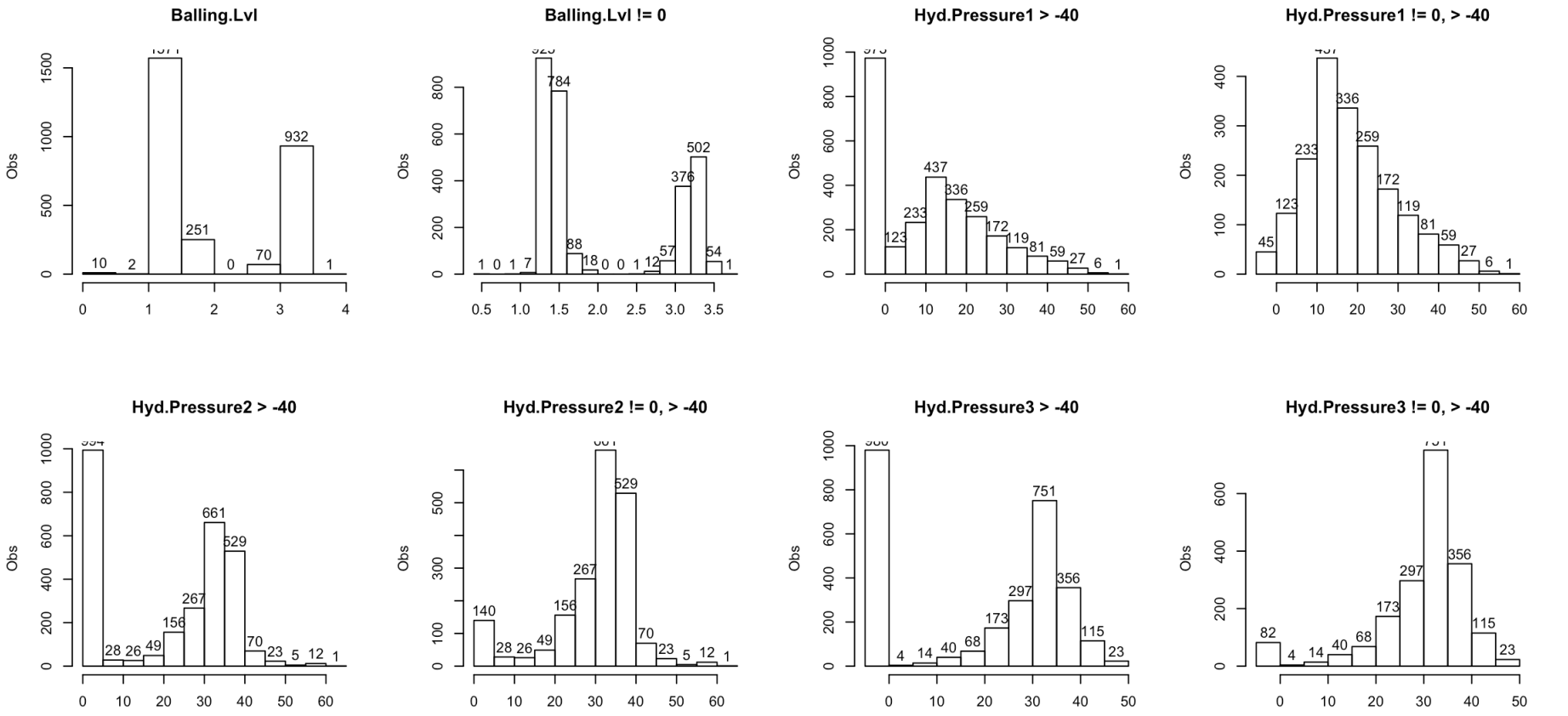




For a few of the variables that had zeros, let's also plot original histogram side-by-side with histogram of nonzero values.

Also remove the one very low outlier in Hyd.Pressure.

```
par(mfrow=c(2,4))

var = "Balling.Lvl"

hist(alldata[,var],ylab="Obs",xlab="",main=var,labels=TRUE)
hist(alldata[alldata[,var] != 0,var],ylab="Obs",xlab="",main=paste0(var," != 0"),labe
ls=TRUE)

for(var in c("Hyd.Pressure1","Hyd.Pressure2","Hyd.Pressure3"))
{
    hist(alldata[alldata[,var] > -40,var],ylab="Obs",xlab="",main=paste0(var," > -40"
),labels=TRUE)
    hist(alldata[alldata[,var] != 0 & alldata[,var] > -40,var],ylab="Obs",xlab="",mai
n=paste0(var," != 0, > -40"),labels=TRUE)
}
```



We see variables with various patterns including bimodal or multimodal, varying degrees of skew, and a few particular values being especially common.

Looks like for Hyd.Pressure 1,2, and 3, most values are much larger than zero if they are not exactly equal to 0.

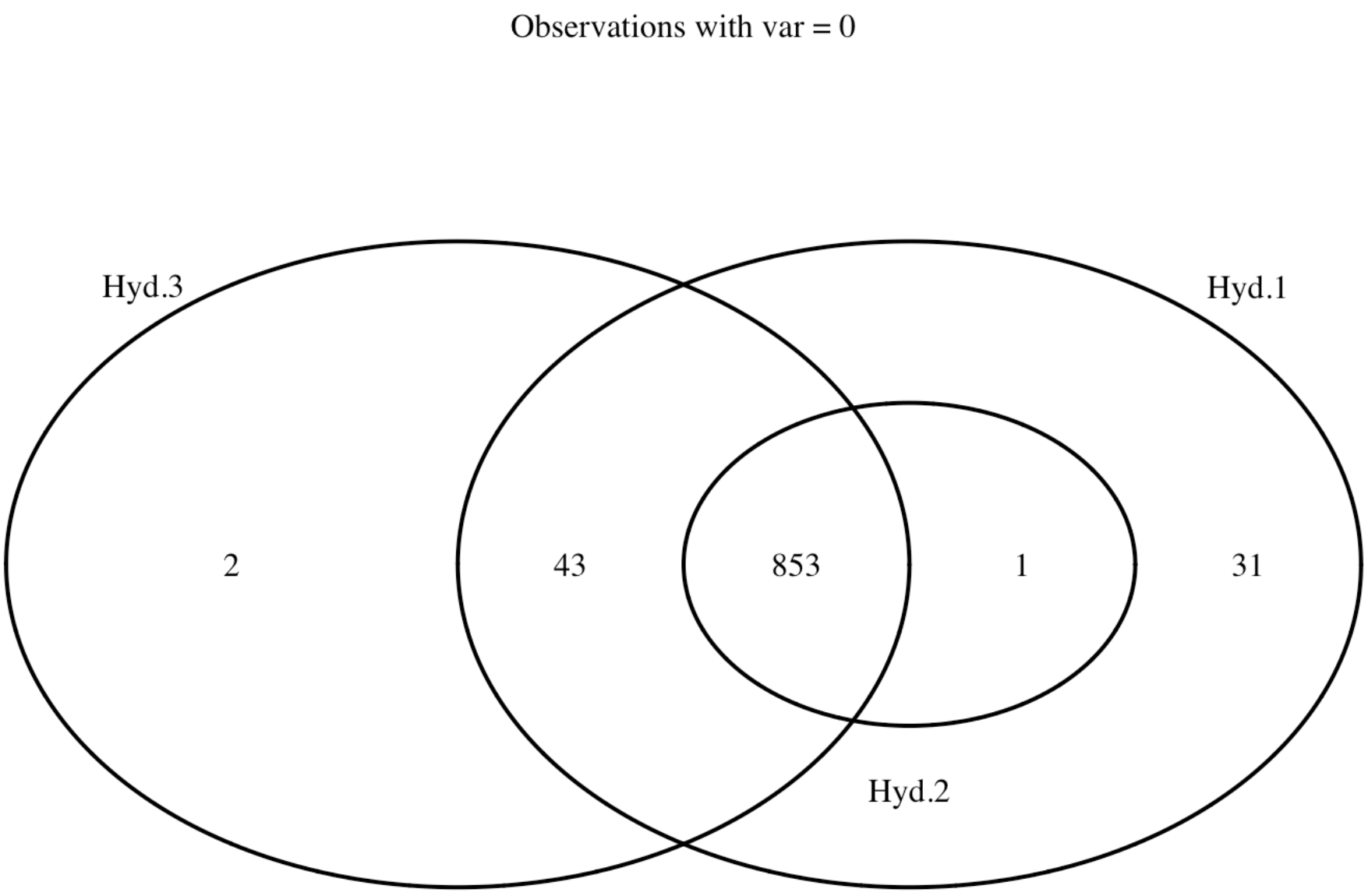
Between variables

Co-occurence of zeros?

One quick obvious thing to check is the correlation between the different Hyd.Pressure variables.

First, do the zeros often co-occur?

```
zero_hyd1 <- which(alldata[,"Hyd.Pressure1"] == 0 & is.na(alldata[,"Hyd.Pressure1"])
== FALSE)
zero_hyd2 <- which(alldata[,"Hyd.Pressure2"] == 0 & is.na(alldata[,"Hyd.Pressure2"])
== FALSE)
zero_hyd3 <- which(alldata[,"Hyd.Pressure3"] == 0 & is.na(alldata[,"Hyd.Pressure3"])
== FALSE)

list_for_venn <- list(Hyd.1 = zero_hyd1,Hyd.2 = zero_hyd2,Hyd.3 = zero_hyd3)

object_for_venn <- venn.diagram(list_for_venn,main="Observations with var = 0",filena
me=NULL)

grid.draw(object_for_venn)
```



Yes, it seems like they definitely do!

One more thing - let's just look a bit more in detail at the observations where Hyd.Pressure2 = 0.2 and/or Hyd.Pressure3 = -1.2.

Have a feeling at least some of these should be 0.

```
length(which(alldata$Hyd.Pressure2 == 0.2))
```

```
## [1] 127
```

```
length(which(alldata$Hyd.Pressure2 == 0.2 & alldata$Hyd.Pressure1 == 0 & alldata$Hyd.
Pressure3 == 0))
```

```
## [1] 43
```

```
length(which(alldata$Hyd.Pressure3 == -1.2))
```

```
## [1] 82
```

```
length(which(alldata$Hyd.Pressure3 == -1.2 & alldata$Hyd.Pressure1 == 0))
```

```
## [1] 26
```

Also, let's print the observations where:

- Hyd.Pressure 1 and 2 are 0, but 3 is not (1 observation)
- Hyd.Pressure 3 is 0, but 1 is not (2 observations)
- Hyd.Pressure 1 is 0, but 3 is not, and 3 is also not -1.2 (32 - 26 = 6 observations)

```
myindices <- which(alldata$Hyd.Pressure1 == 0 & alldata$Hyd.Pressure2 == 0 & alldata$
Hyd.Pressure3 != 0)

alldata[myindices,paste0("Hyd.Pressure",1:3)]
```

```
##      Hyd.Pressure1 Hyd.Pressure2 Hyd.Pressure3
## 1719             0             0           1.6
```

```
myindices <- which(alldata$Hyd.Pressure3 == 0 & alldata$Hyd.Pressure1 != 0)

alldata[myindices,paste0("Hyd.Pressure",1:3)]
```

```
##      Hyd.Pressure1 Hyd.Pressure2 Hyd.Pressure3
## 1215           0.2           0.2             0
## 1376           0.2           0.2             0
```

```
zero_hyd1 <- which(alldata$Hyd.Pressure1 == 0)
zero_or_neg1.2_hyd3 <- which(alldata$Hyd.Pressure3 == 0 | alldata$Hyd.Pressure3 == -1
.2)

myindices <- setdiff(zero_hyd1,zero_or_neg1.2_hyd3)

alldata[myindices,paste0("Hyd.Pressure",1:3)]
```

```
##      Hyd.Pressure1 Hyd.Pressure2 Hyd.Pressure3
## 1                0            NA            NA
## 2                0            NA            NA
## 3                0            NA            NA
## 1719             0           0.0           1.6
## 1962             0          17.2          23.0
## 2572             0            NA            NA
```

Looks like we may want to convert 0.2 to 0 for Hyd.Pressure2 when both Hyd.Pressure 1 and 3 are 0.

Same idea for Hyd.Pressure3. Convert -1.2 to 0 when Hyd.Pressure 1 is 0.

Finally, convert 0.2 to 0 for Hyd.Pressure 1 and 2 when 3 = 0.

Oh, and convert NA to 0 when 2/3 are NA and 1 is 0.

Not 100% sure these conversions are correct, so let's also set aside the original data.

```
alldata_original <- alldata

alldata[which(alldata$Hyd.Pressure2 == 0.2 & alldata$Hyd.Pressure1 == 0 & alldata$Hyd
.Pressure3 == 0),"Hyd.Pressure2"] <- 0
alldata[which(alldata$Hyd.Pressure3 == -1.2 & alldata$Hyd.Pressure1 == 0),"Hyd.Pressu
re3"] <- 0

myindices <- which(alldata$Hyd.Pressure1 == 0.2 & alldata$Hyd.Pressure2 == 0.2 & alld
ata$Hyd.Pressure3 == 0)

alldata[myindices,"Hyd.Pressure1"] <- 0
alldata[myindices,"Hyd.Pressure2"] <- 0

myindices <- which(alldata$Hyd.Pressure1 == 0 & is.na(alldata$Hyd.Pressure2) == TRUE
& is.na(alldata$Hyd.Pressure3) == TRUE)

alldata[myindices,"Hyd.Pressure2"] <- 0
alldata[myindices,"Hyd.Pressure3"] <- 0

myindices <- which(alldata$Hyd.Pressure1 == 0 & alldata$Hyd.Pressure3 == 0 & alldata$
Hyd.Pressure2 == 0.2)

alldata[myindices,"Hyd.Pressure2"] <- 0
```

Now, just curious, what do the other two variables look like when Hyd.Pressure2 is 0.2 or Hyd.Pressure3 is -1.2?

```
myindices1 <- which(alldata$Hyd.Pressure2 == 0.2)

myindices2 <- which(alldata$Hyd.Pressure3 == -1.2)

length(myindices1)
```

```
## [1] 56
```

```
length(myindices2)
```

```
## [1] 56
```

```
length(intersect(myindices1,myindices2))
```

```
## [1] 56
```

```
hyd1_when_hyd2plus3_repeated_values <- as.numeric(as.vector(alldata[myindices1,"Hyd.P
ressure1"]))


table(hyd1_when_hyd2plus3_repeated_values[order(hyd1_when_hyd2plus3_repeated_values)]
)
```

```
##
##   -1 -0.8 -0.6 -0.4 -0.2  0.2  0.4
##    1   22   16    5    1    9    2
```

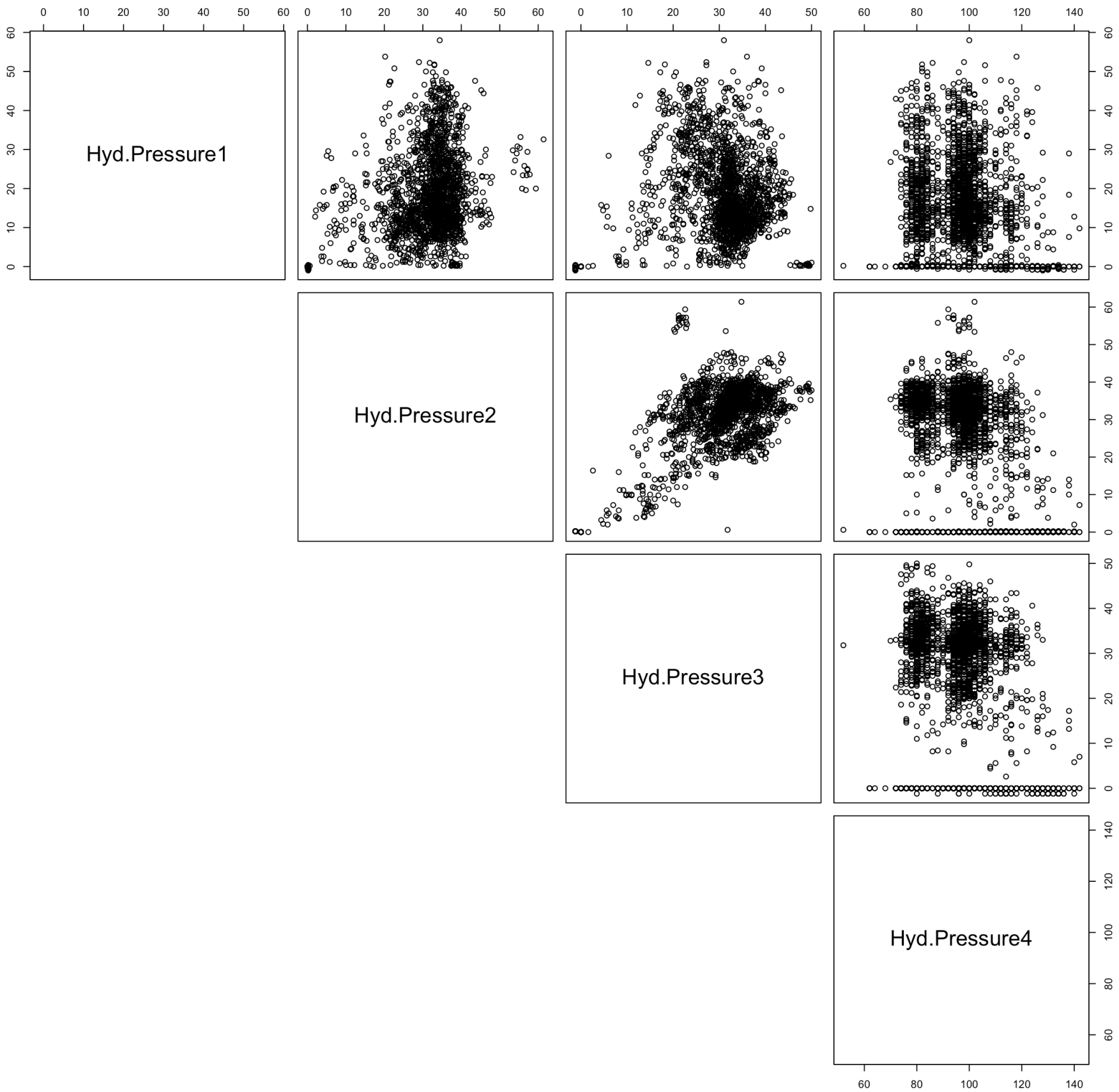
After some transformations in other rows to 0, looks like Hyd.Pressure2 = 0.2 and Hyd.Pressure3 = -1.2 always co-occur.

Then Hyd.Pressure1 is within a range of different low values.

Correlations between certain specific variables

Next, make scatterplots looking at correlation including nonzero values for Hyd.Pressure 1-4.

```
pairs(alldata[alldata[,"Hyd.Pressure1"] > -40,paste0("Hyd.Pressure",1:4)],lower.panel
=NULL,main="Minus 1/2/3 < -40 observation")
```



Other than the common zeros, the Hyd.Pressure variables are not quite as correlated as I expected.

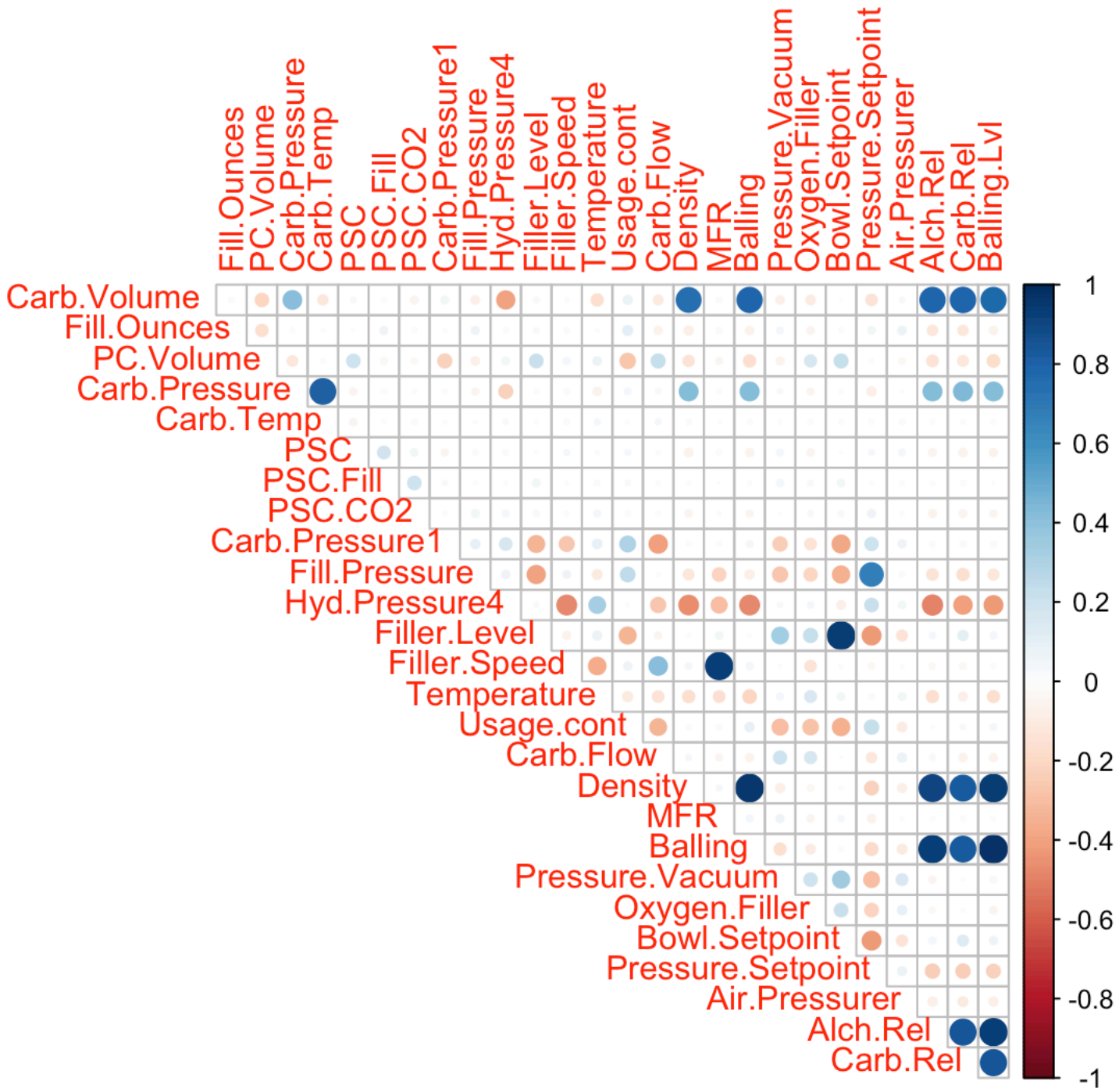
Some correlation between 2 and 3, but it is not that strong other than when both are very low.

Correlations between all predictors

Now, time to take an unbiased look and see which variables are most correlated.

For now, exclude Hyd.Pressure1-3 and Mnf.Flow from this since they have very oft-repeated values. Also exclude the categorical variables.

```
alldata_for_correlations <- alldata[,setdiff(colnames(alldata),c("Data","Brand.Code",
"Hyd.Pressure1","Hyd.Pressure2","Hyd.Pressure3","Mnf.Flow"))]

corrplot(cor(alldata_for_correlations,use="pairwise.complete.obs"),type="upper",diag=
FALSE)
```

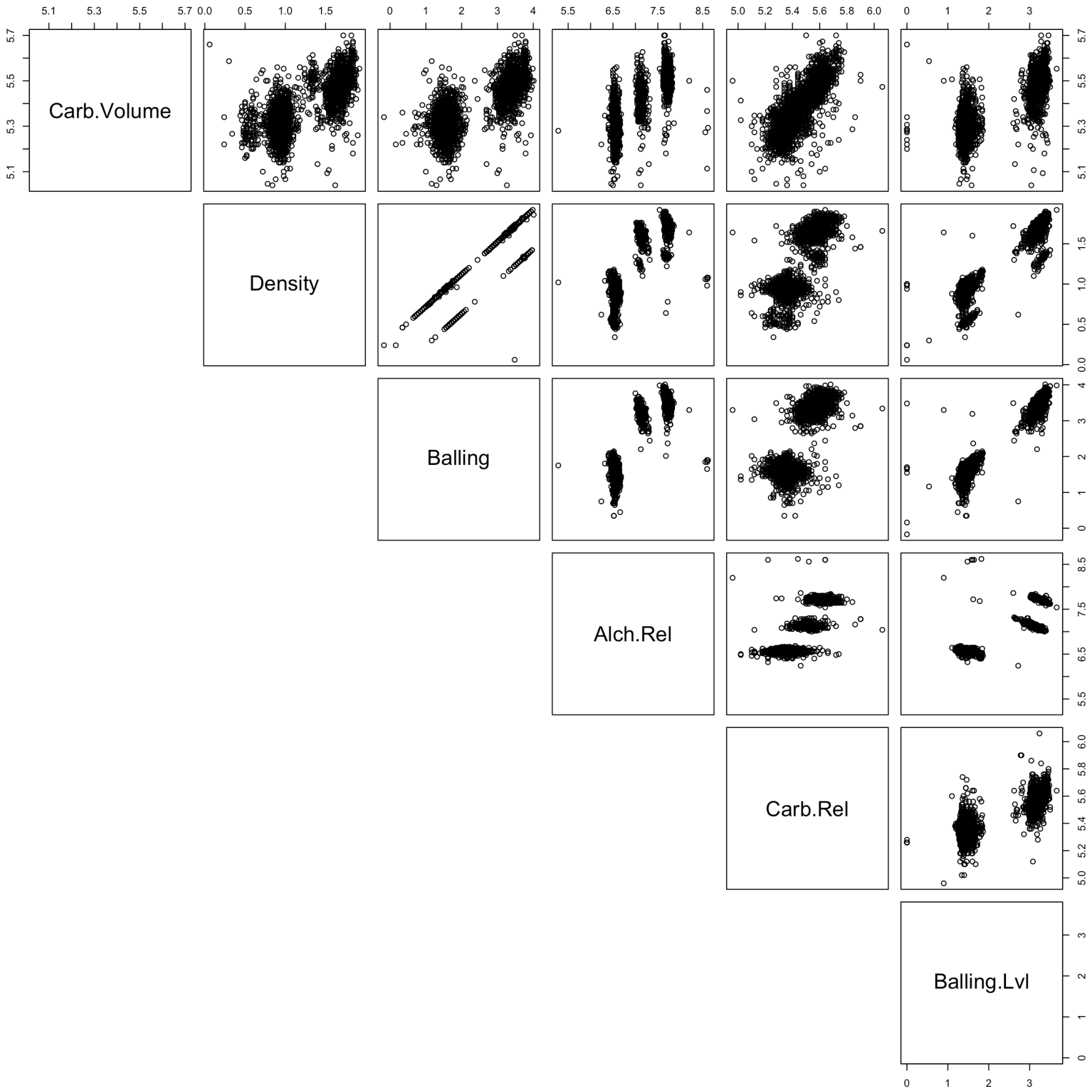


We find the following very strong correlations:

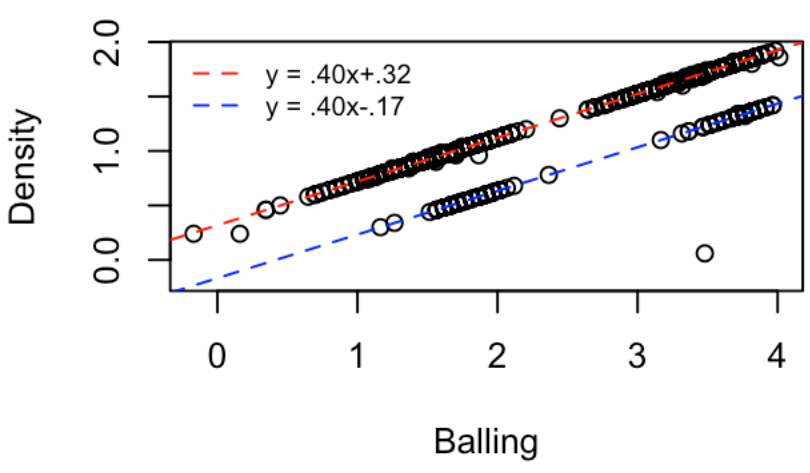
- Carb.Volume with Density, Balling, Alch.Rel, Carb.Rel, and Balling.Lvl (and correlations within these as well)
- Carb.Pressure with Carb.Temp
- Filler.Level with Bowl.Setpoint
- Filler.Speed with MFR

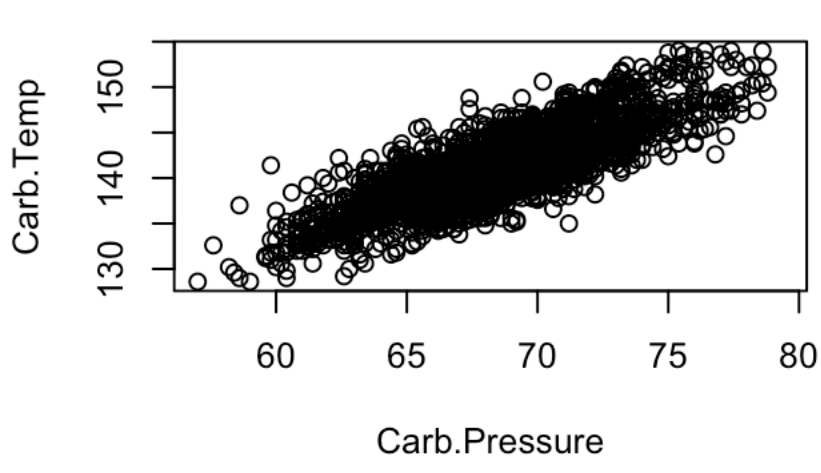
Let’s make some scatterplots!

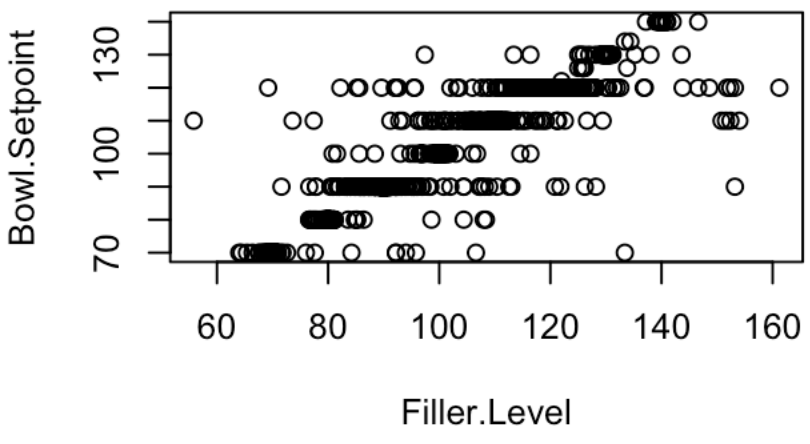
```
pairs(alldata[,c("Carb.Volume","Density","Balling","Alch.Rel","Carb.Rel","Balling.Lvl
")],lower.panel=NULL)
```

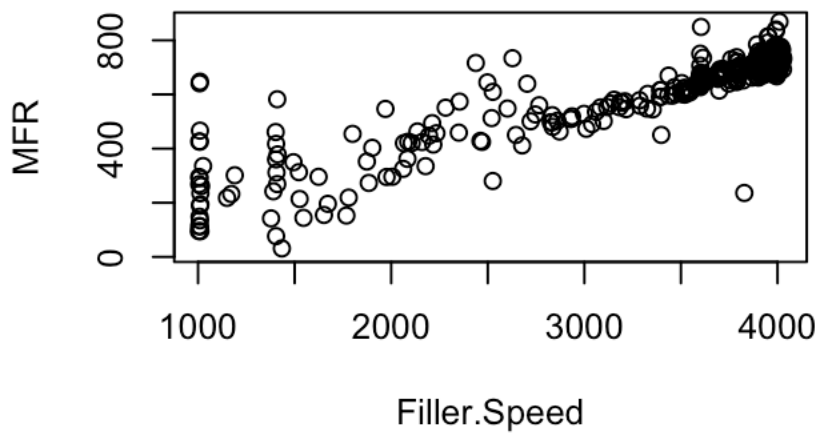


```
par(mfrow=c(2,2))

plot(alldata$Balling,alldata$Density,xlab="Balling",ylab="Density",
    ylim=c(-.20,max(alldata$Density,na.rm=TRUE)))

model1_indices <- which(alldata$Density > (0.4*alldata$Balling))
model2_indices <- which(alldata$Density <= (0.4*alldata$Balling))
model2_indices <- setdiff(model2_indices,which(alldata$Density == min(alldata$Density
,na.rm=TRUE)))

abline(lm(Density ~ Balling,data=alldata[model1_indices,]),lty=2,col="red")
abline(lm(Density ~ Balling,data=alldata[model2_indices,]),lty=2,col="blue")

legend("topleft",
    legend=c("y = .40x+.32","y = .40x-.17"),
    col=c("red","blue"),lty=2,bty="n",cex=0.75)

plot(alldata$Carb.Pressure,alldata$Carb.Temp,xlab="Carb.Pressure",ylab="Carb.Temp")
plot(alldata$Filler.Level,alldata$Bowl.Setpoint,xlab="Filler.Level",ylab="Bowl.Setpoi
nt")
plot(alldata$Filler.Speed,alldata$MFR,xlab="Filler.Speed",ylab="MFR")
```







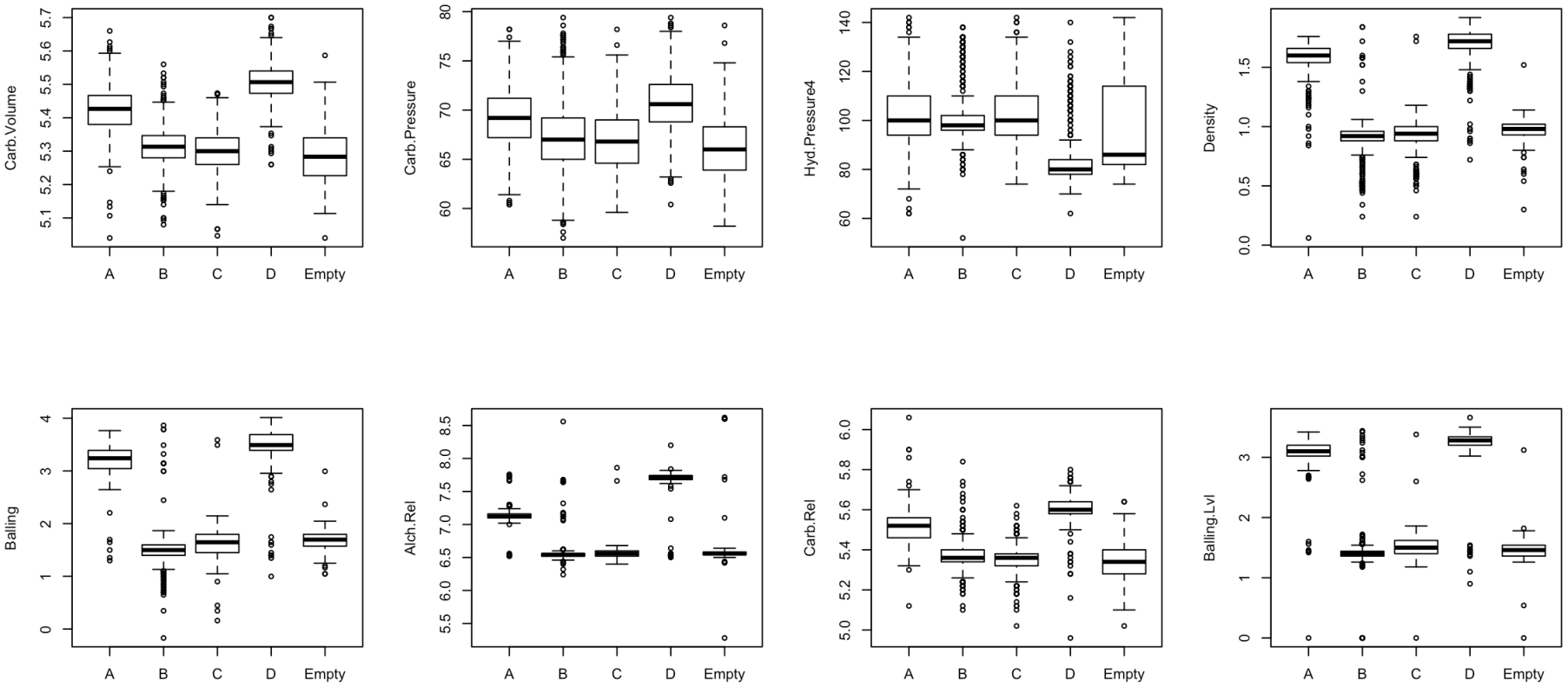


Looks like density and balling are actually almost perfectly correlated. Except, some points are modelled by a different relationship than others.

For the remaining variables, the correlation coefficient generally makes sense with the scatterplot, but not always as expected. Sometimes, there is just a correlation when points from one modal peak of variable1 have higher values for variable2 than points from the other (lower) modal peak of variable1.

Now, let's look at brand code vs. other variables.

```
brand_code <- alldata$Brand.Code

brand_code_dat <- data.frame(A = ifelse(alldata$Brand.Code == "A",1,0),
    B = ifelse(alldata$Brand.Code == "B",1,0),
    C = ifelse(alldata$Brand.Code == "C",1,0),
    D = ifelse(alldata$Brand.Code == "D",1,0),
    stringsAsFactors=FALSE)

correlations_with_brand <- cor(brand_code_dat,alldata_for_correlations,use="pairwise.
complete.obs")

correlations_with_brand_max <- apply(correlations_with_brand,2,function(x)max(abs(x))
)

vars_corr_with_brand <- colnames(correlations_with_brand)[correlations_with_brand_max
> 0.4]

par(mfrow=c(2,4))

for(var in vars_corr_with_brand)
{
    boxplot(alldata[,var] ~ alldata$Brand.Code,ylab=var)
}
```

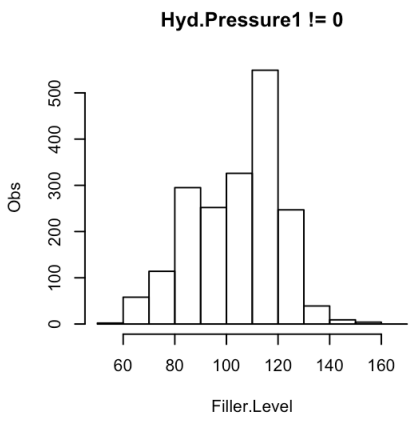


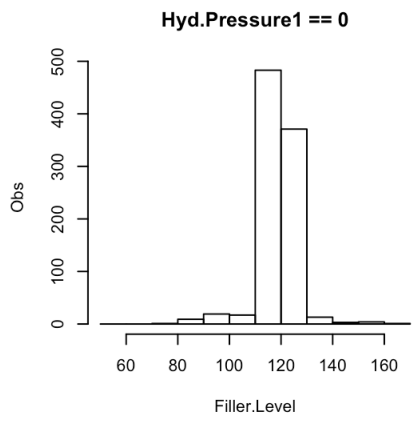
We find the following patterns.

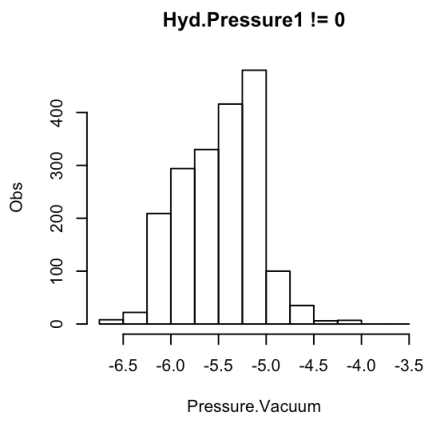
- Higher Carb.Volume, Carb.Pressure (effect not as strong), Density, Balling, Alch.Rel, Carb.Rel, and Balling.Lvl in brands A and D.
- Lower Hyd.Pressure4 in brand D.

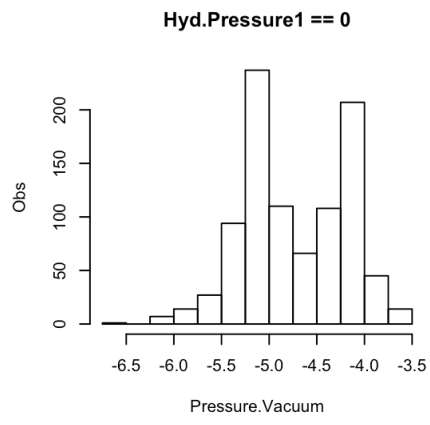
For Hyd.Pressure 1-3, let's see if any patterns for where Hyd.Pressure1 is 0 vs. not.

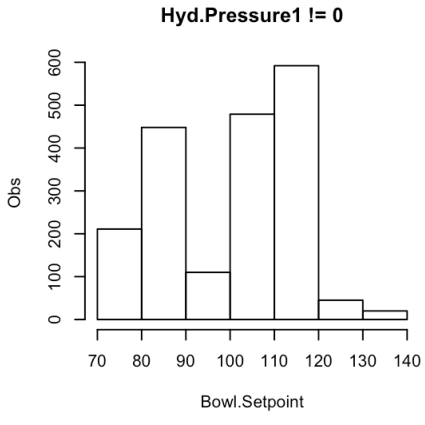
```
correlations_with_hyd_zero <- cor(ifelse(alldata$Hyd.Pressure1 == 0,1,0),alldata_for_
correlations,use="pairwise.complete.obs")

vars_corr_with_hyd_zero <- colnames(correlations_with_brand)[abs(correlations_with_hy
d_zero) >= 0.4]

zero_indices <- which(alldata$Hyd.Pressure1 == 0)
nonzero_indices <- setdiff(1:nrow(alldata),zero_indices)

par(mfrow=c(2,4))

for(var in vars_corr_with_hyd_zero)
{
    if(var == "Filler.Level"){mybreaks=seq(from=50,to=170,by=10)}
    if(var == "Pressure.Vacuum"){mybreaks=seq(from=-6.75,to=-3.5,by=0.25)}
    if(var == "Bowl.Setpoint"){mybreaks=seq(from=70,to=140,by=10)}
    if(var == "Pressure.Setpoint"){mybreaks=seq(from=42,to=54,by=2)}
    hist(alldata[nonzero_indices,var],xlab=var,ylab="Obs",main="Hyd.Pressure1 != 0",b
reaks=mybreaks)
    hist(alldata[zero_indices,var],xlab=var,ylab="Obs",main="Hyd.Pressure1 == 0",brea
ks=mybreaks)
    if(var == "Pressure.Setpoint")
    {
        print(var)
        print(table(alldata[nonzero_indices,var]))
        print(table(alldata[zero_indices,var]))
    }
}
```

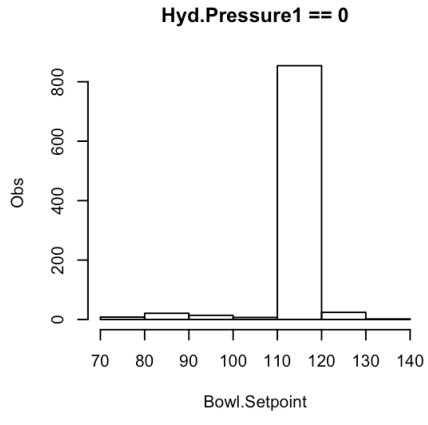


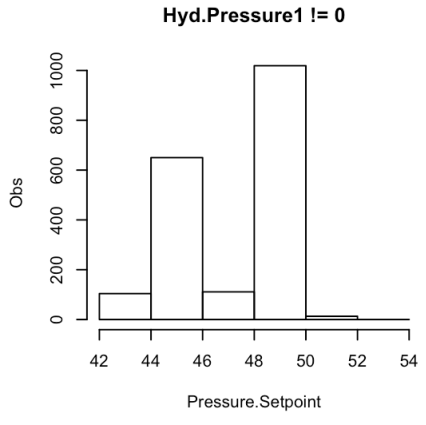


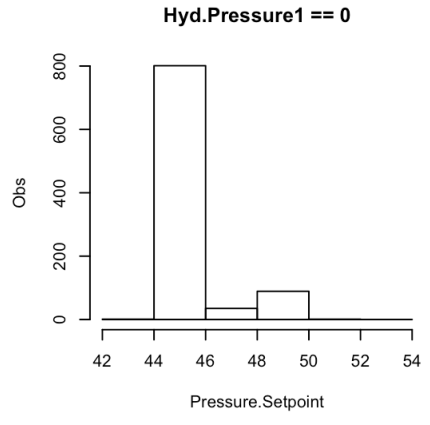












```
## [1] "Pressure.Setpoint"
##
##   44   46   48   50   52
##  104  650  111 1019   13
##
##   44 45.2   46 46.4 46.6 46.8   48   50   52
##    1    1  800    1    1    1   32   89    1
```

We find Filler.Level and Pressure.Vacuum are both higher when Hyd.Pressure1 = 0.

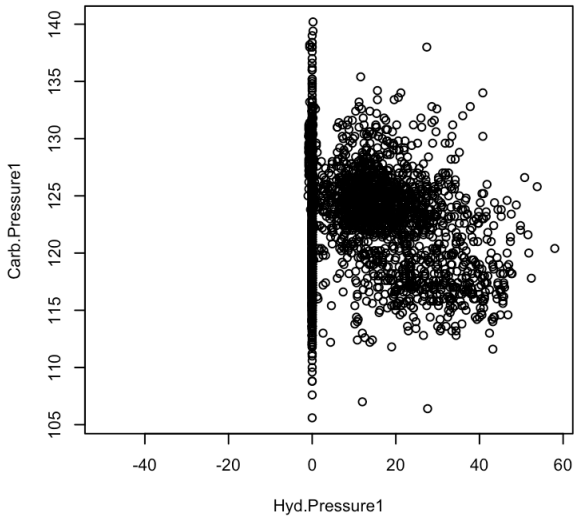
For Bowl.Setpoint and Pressure.Setpoint, we find that the values tend to be less on the extremes and more toward more common values when Hyd.Pressure1 = 0.

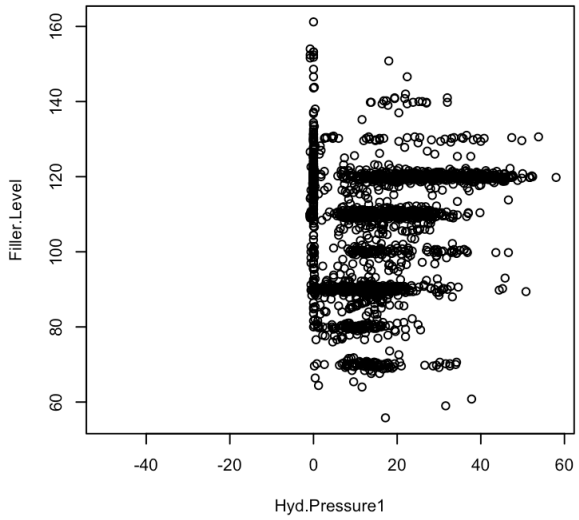
Another question - are Hyd.Pressure 1/2/3 correlated with other variables when nonzero?

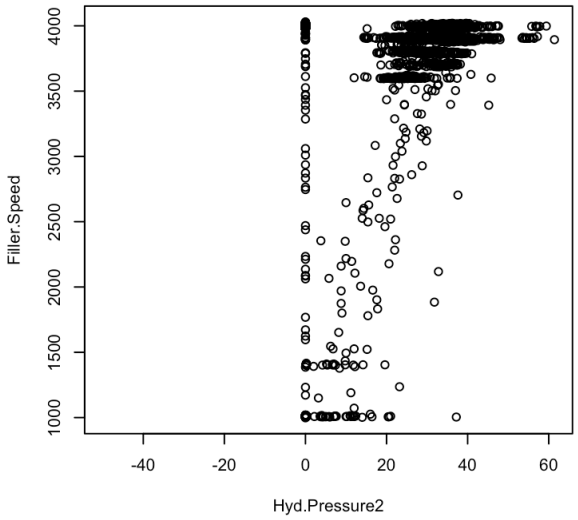
```
alldata_for_correlations <- alldata[nonzero_indices,setdiff(colnames(alldata),c("Data
","Brand.Code","Mnf.Flow"))]

correlations_with_hyd_nonzero <- cor(alldata_for_correlations,use="pairwise.complete.
obs")

correlations_with_hyd_nonzero <- correlations_with_hyd_nonzero[paste0("Hyd.Pressure",
1:3),setdiff(colnames(correlations_with_hyd_nonzero),paste0("Hyd.Pressure",1:3))]

max_correlations_with_hyd_nonzero <- apply(correlations_with_hyd_nonzero,2,function(x
)max(abs(x)))

correlations_with_hyd_nonzero[,max_correlations_with_hyd_nonzero >= 0.4]
```

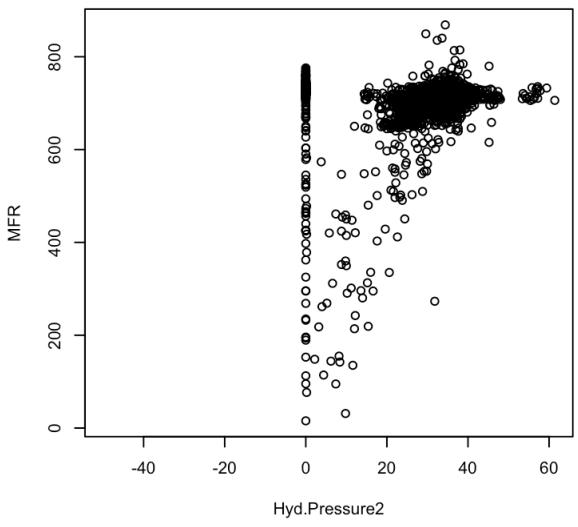
```
##                Carb.Pressure1 Filler.Level Filler.Speed       MFR
## Hyd.Pressure1    -0.44120052   0.41402994    0.2479046 0.0958028
## Hyd.Pressure2    -0.14105977  -0.02229064    0.7591481 0.5826434
## Hyd.Pressure3     0.07446566  -0.27719985    0.7000897 0.4667870
##                Bowl.Setpoint
## Hyd.Pressure1    0.45180545
## Hyd.Pressure2    0.03587979
## Hyd.Pressure3   -0.20903950
```

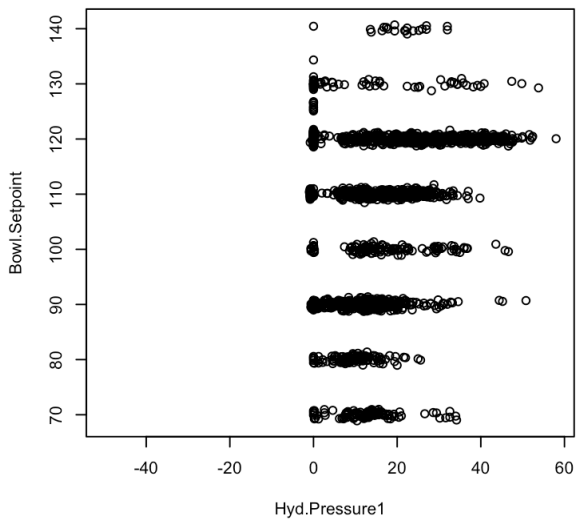
```
par(mfrow=c(2,3))

plot(alldata$Hyd.Pressure1,alldata$Carb.Pressure1,xlab="Hyd.Pressure1",ylab="Carb.Pressure1")
plot(alldata$Hyd.Pressure1,alldata$Filler.Level,xlab="Hyd.Pressure1",ylab="Filler.Level")
plot(alldata$Hyd.Pressure2,alldata$Filler.Speed,xlab="Hyd.Pressure2",ylab="Filler.Speed")
plot(alldata$Hyd.Pressure2,alldata$MFR,xlab="Hyd.Pressure2",ylab="MFR")
plot(alldata$Hyd.Pressure1,alldata$Bowl.Setpoint + rnorm(nrow(alldata),mean=0,sd=0.5),xlab="Hyd.Pressure1",ylab="Bowl.Setpoint")
```





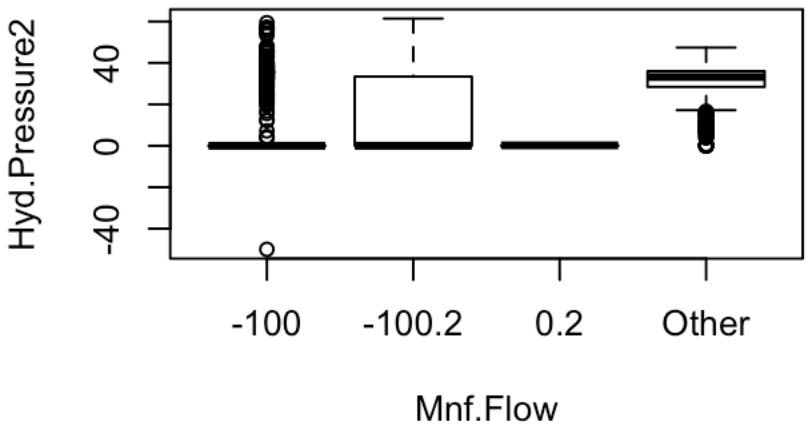


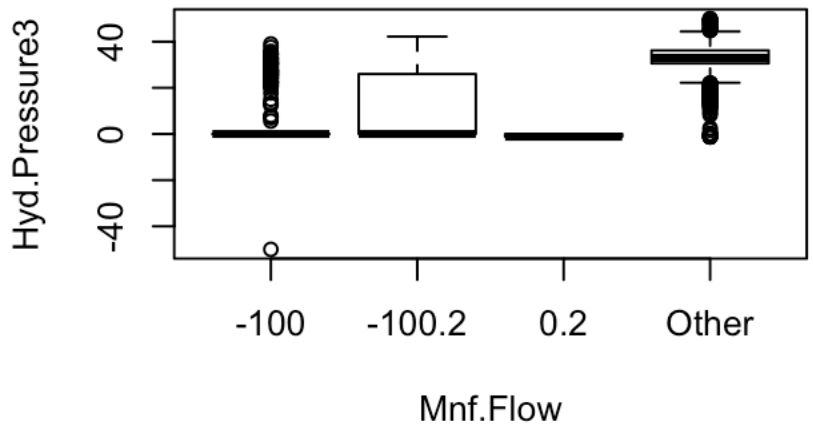


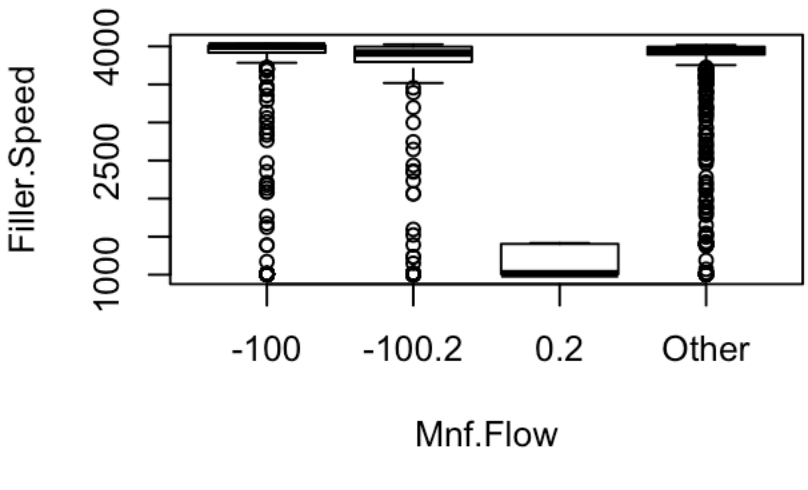


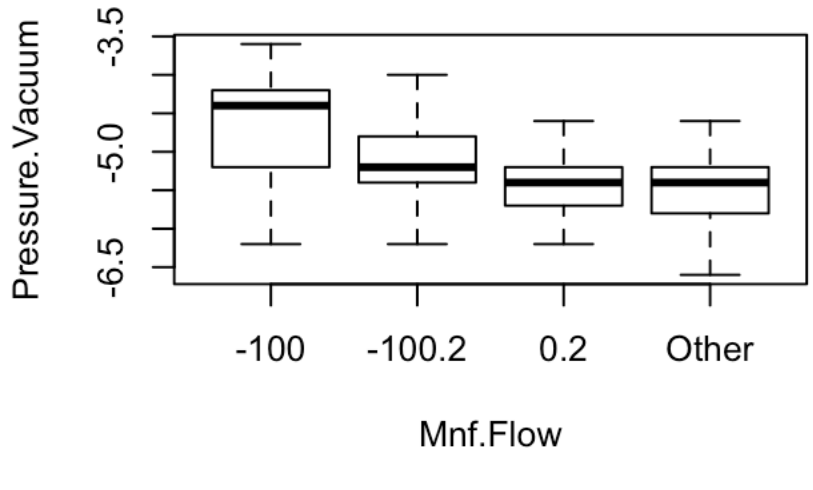
Finally, look at Mnf.Flow at the repeated values vs. others.

```
mnf_flow <- alldata$Mnf.Flow

mnf_flow <- plyr::mapvalues(mnf_flow,
    from=c(-100.2,-100,0.2,setdiff(unique(mnf_flow),c(-100.2,-100,0.2))),
    to=c("-100.2","-100","0.2",rep("Other",times=length(unique(mnf_flow)) - 3)))

mnf_flow_dat <- data.frame(Neg.100point2 = ifelse(mnf_flow == "-100.2",1,0),
    Neg.100 = ifelse(mnf_flow == "-100",1,0),
    Point2 = ifelse(mnf_flow == "0.2",1,0),
    stringsAsFactors=FALSE)

alldata_for_correlations <- alldata[,setdiff(colnames(alldata),c("Data","Brand.Code",
"Mnf.Flow"))]

correlations_with_mnf_flow <- cor(mnf_flow_dat,alldata_for_correlations,use="pairwise
.complete.obs")

correlations_with_mnf_flow_max <- apply(correlations_with_mnf_flow,2,function(x)max(a
bs(x)))

vars_corr_with_mnf_flow <- colnames(correlations_with_mnf_flow)[correlations_with_mnf
_flow_max >= 0.4]

par(mfrow=c(2,2))

for(var in vars_corr_with_mnf_flow)
{
    boxplot(alldata[,var] ~ factor(mnf_flow),ylab=var,xlab="Mnf.Flow")
}
```

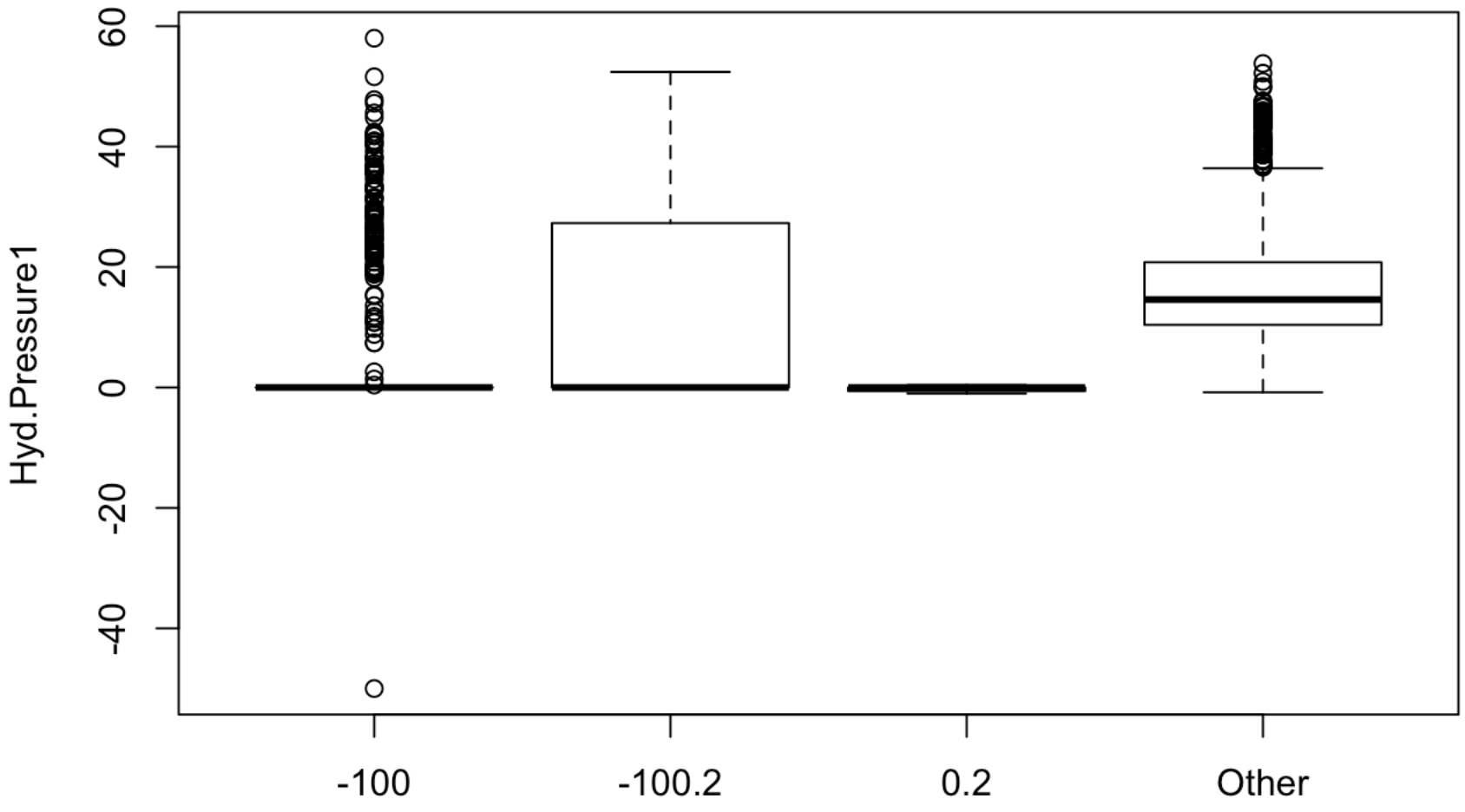









```
par(mfrow=c(1,1))

boxplot(alldata[,"Hyd.Pressure1"] ~ factor(mnf_flow),ylab="Hyd.Pressure1",xlab="Mnf.F
low")
```



Looks like the zeros in Hyd.Pressure 1/2/3 frequently match up to Mnf.Flow <= 0.2.

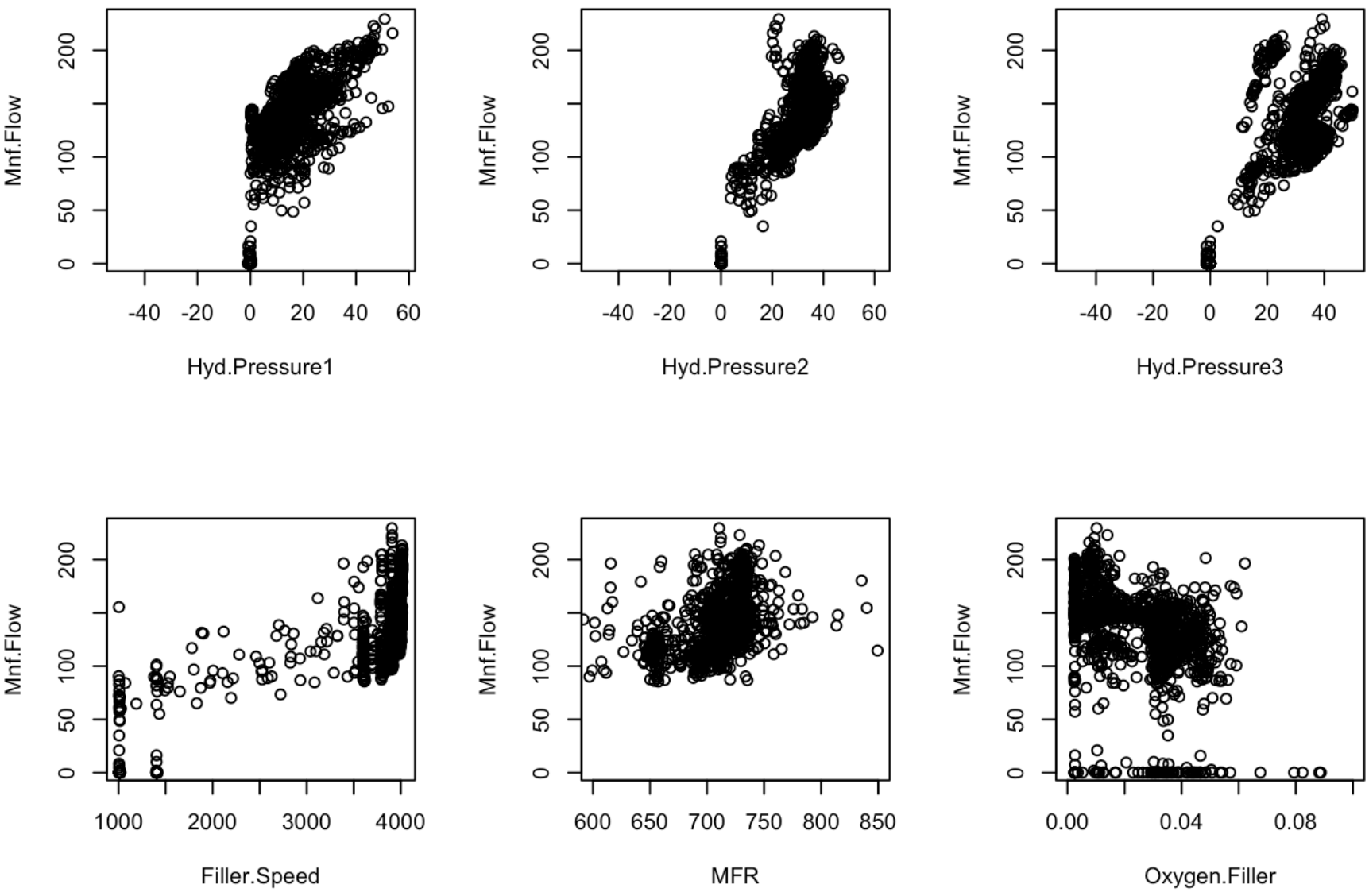
Mnf.Flow 0.2 is associated with much lower Filler.Speed.

The lower repeated Mnf.Flow values are associated with higher Pressure.Vacuum.

```
non_repeated_indices <- which(rowSums(mnf_flow_dat) == 0)

alldata_for_correlations <- alldata[non_repeated_indices,setdiff(colnames(alldata),c(
"Data","Brand.Code","Mnf.Flow"))]

correlations_with_mnf_flow <- cor(alldata$Mnf.Flow[non_repeated_indices],alldata_for_
correlations,use="pairwise.complete.obs")

vars_corr_with_mnf_flow <- colnames(alldata_for_correlations)[abs(correlations_with_m
nf_flow) >= 0.4]

vars_corr_with_mnf_flow
```

```
## [1] "Hyd.Pressure1" "Hyd.Pressure2" "Hyd.Pressure3" "Filler.Speed"
## [5] "MFR"           "Oxygen.Filler"
```

```
par(mfrow=c(2,3))

for(var in vars_corr_with_mnf_flow)
{
    if(var != "MFR" & var != "Oxygen.Filler")
    {
        plot(alldata[,var],alldata$Mnf.Flow,xlab=var,ylab="Mnf.Flow",ylim=c(min(allda
ta$Mnf.Flow[alldata$Mnf.Flow > 0.2],na.rm=TRUE),max(alldata$Mnf.Flow,na.rm=TRUE)))
    }
    if(var == "MFR")
    {
        plot(alldata[,var],alldata$Mnf.Flow,xlab=var,ylab="Mnf.Flow",ylim=c(min(allda
ta$Mnf.Flow[alldata$Mnf.Flow > 0.2],na.rm=TRUE),max(alldata$Mnf.Flow,na.rm=TRUE)),
        xlim=c(600,850))
    }
    if(var == "Oxygen.Filler")
    {
        plot(alldata[,var],alldata$Mnf.Flow,xlab=var,ylab="Mnf.Flow",ylim=c(min(allda
ta$Mnf.Flow[alldata$Mnf.Flow > 0.2],na.rm=TRUE),max(alldata$Mnf.Flow,na.rm=TRUE)),
        xlim=c(0,0.1))
    }
}
```



Looks like we see some of the same variables that were correlated with Mnf.Flow as a factor, along with a few new ones.

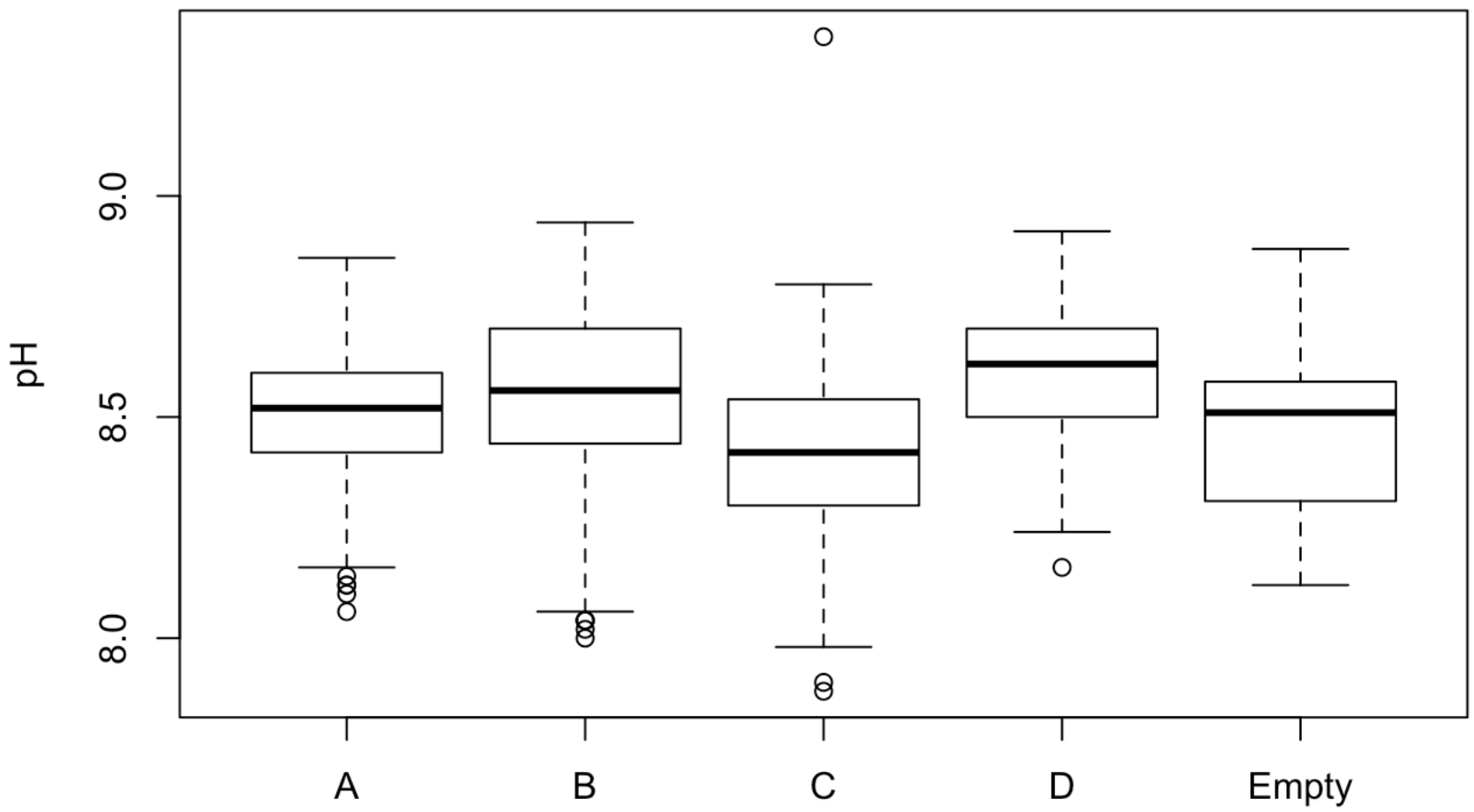
Variables vs. target

Separate out training data.

```
training <- alldata[alldata$Data == "Training",]
```

Start with a simple boxplot for brand code.

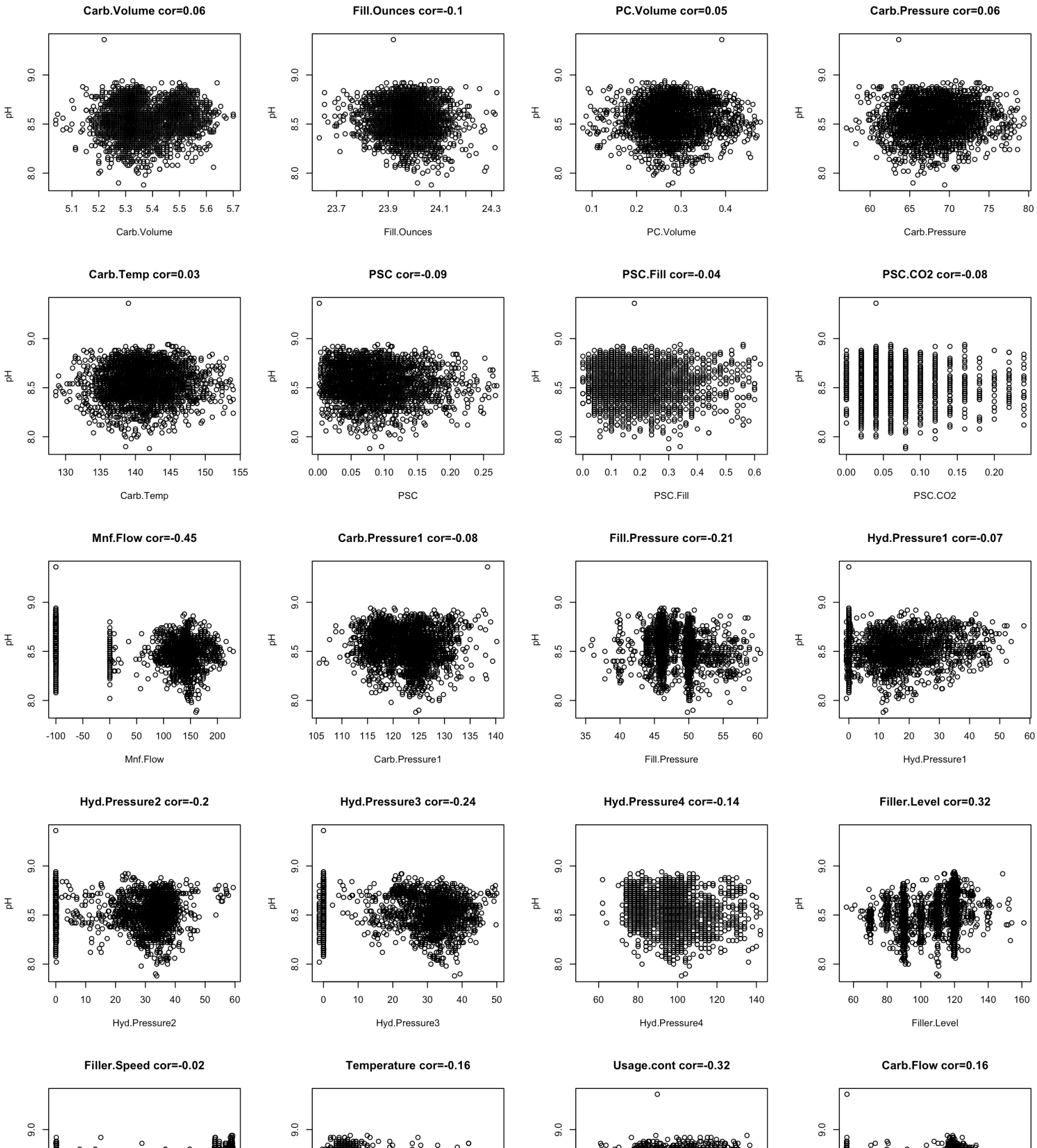
```
boxplot(training_target ~ training$Brand.Code,ylab="pH")
```

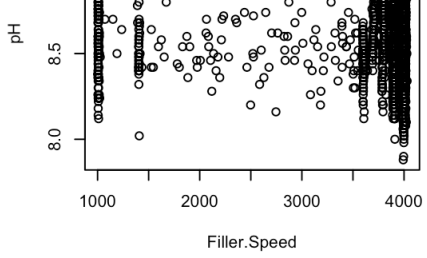


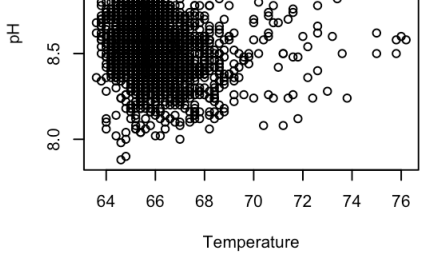
Looks like brand C may have a somewhat lower pH.

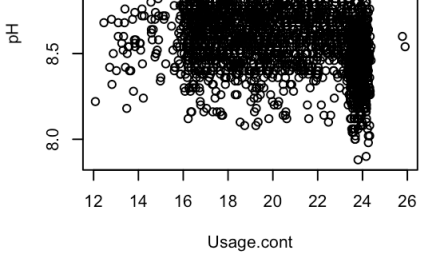
Plot scatterplots for remaining variables.

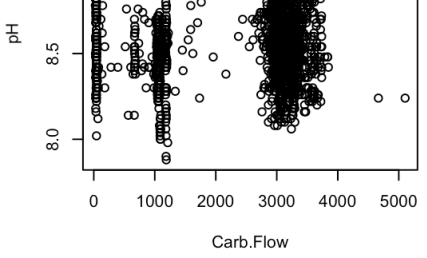
```
par(mfrow=c(8,4))

for(var in setdiff(colnames(training),c("Data","Brand.Code")))
{
    plot(training[,var],training_target,xlab=var,ylab="pH",main=paste0(var," cor=",ro
und(cor(training[,var],training_target,use="pairwise.complete.obs"),digits=2)))
}
```



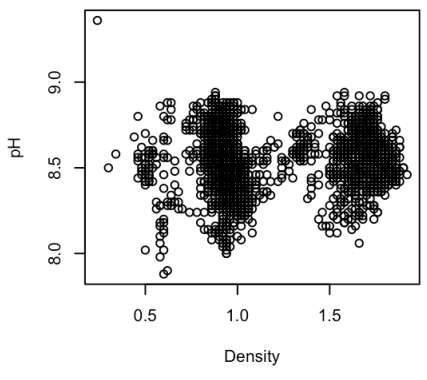




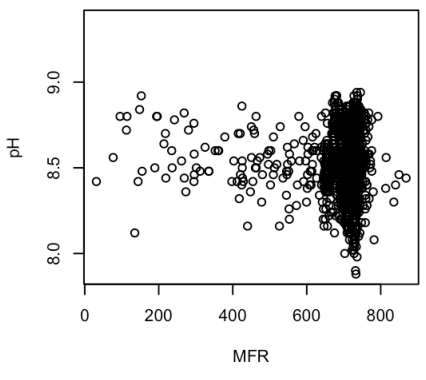




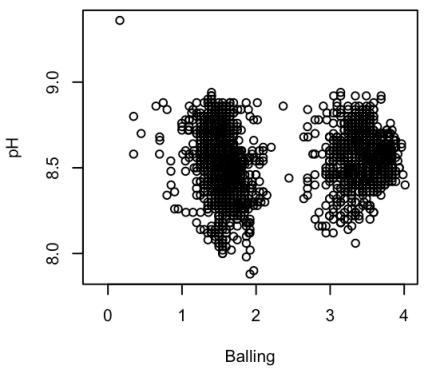
Density cor=0.08



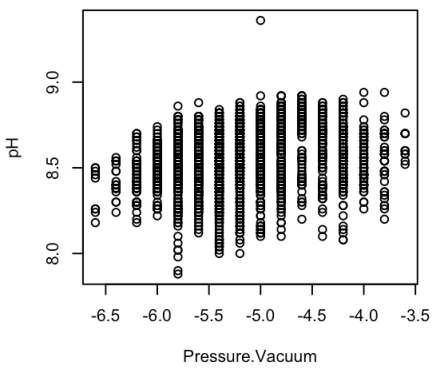
MFR cor=-0.01



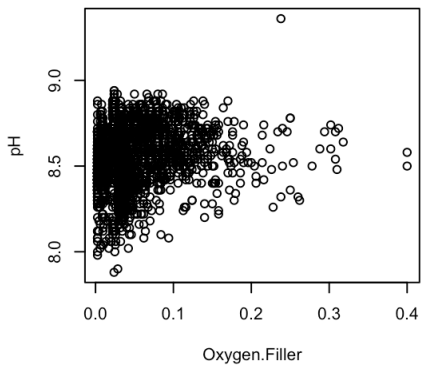
Balling cor=0.07



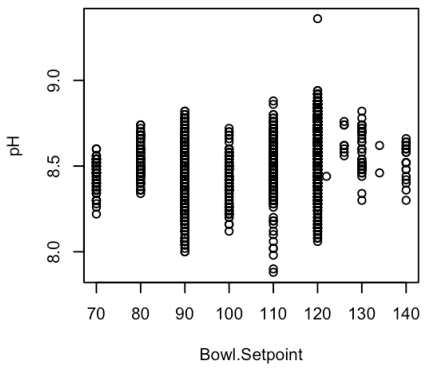
Pressure.Vacuum cor=0.22



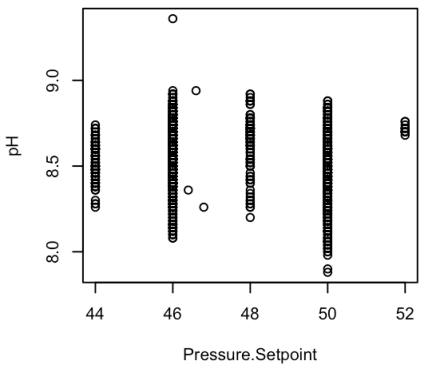
Oxygen.Filler cor=0.17



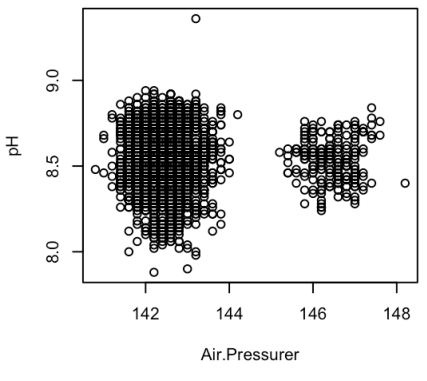
Bowl.Setpoint cor=0.35



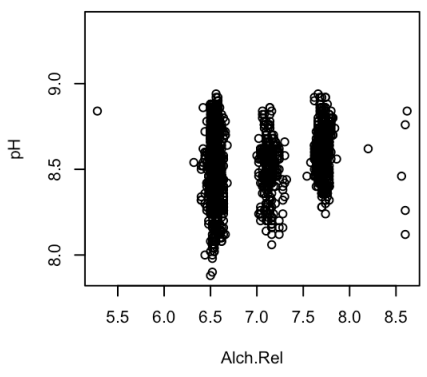
Pressure.Setpoint cor=-0.31



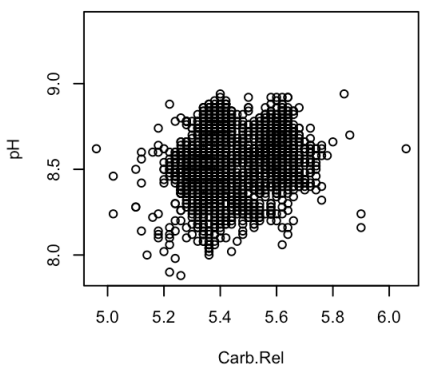
Air.Pressurer cor=-0.01



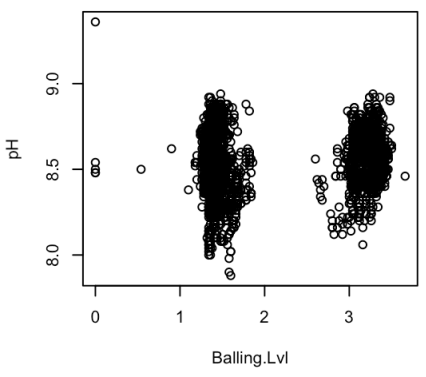
Alch.Rel cor=0.15



Carb.Rel cor=0.16



Balling.Lvl cor=0.1



Looks like there are a fair number of variables that will probably not end up in the final model, as they are not really associated with the target (pH) at all.

There also seem like there may be some non-linear associations. For example, Usage.cont generally has constant pH throughout its range, except some observations with Usage.cont > 23 or so have much lower pH than the other points.