Variable Report

Cuny MSDA 624 Project 2

Tom Detzel

11/25/2017

```
knitr::opts_chunk$set(echo = TRUE)

# load required packages
suppressMessages(library(easypackages))
suppressMessages(libraries("tidyverse", "nnet", "kernlab", "
```

Get the data

```
# read in the data
df <- read.csv("data/StudentData.csv", header=T, strip.white
# str(df)
```

Get my variables

```
df <- df[, 25:33]
```

Missingness

Missing values aren’t a big problem with these vars.

```
## assumes '0' is an NA
zero_vals <- data.frame(cbind(colSums(df==0)))
colnames(zero_vals) <- "zero_count"
pander(zero_vals)
```

	zero_count
Pressure.Vacuum	0
PH	4
Oxygen.Filler	12
Bowl.Setpoint	2
Pressure.Setpoint	12
Air.Pressurer	0
Alch.Rel	9
Carb.Rel	10
Balling.Lvl	8

Work with complete cases

We lost 40 observations, a de minimus number.

```
# recode '0' as NA
df[df==0] <- NA

# complete cases
df_c <- na.omit(df)
```

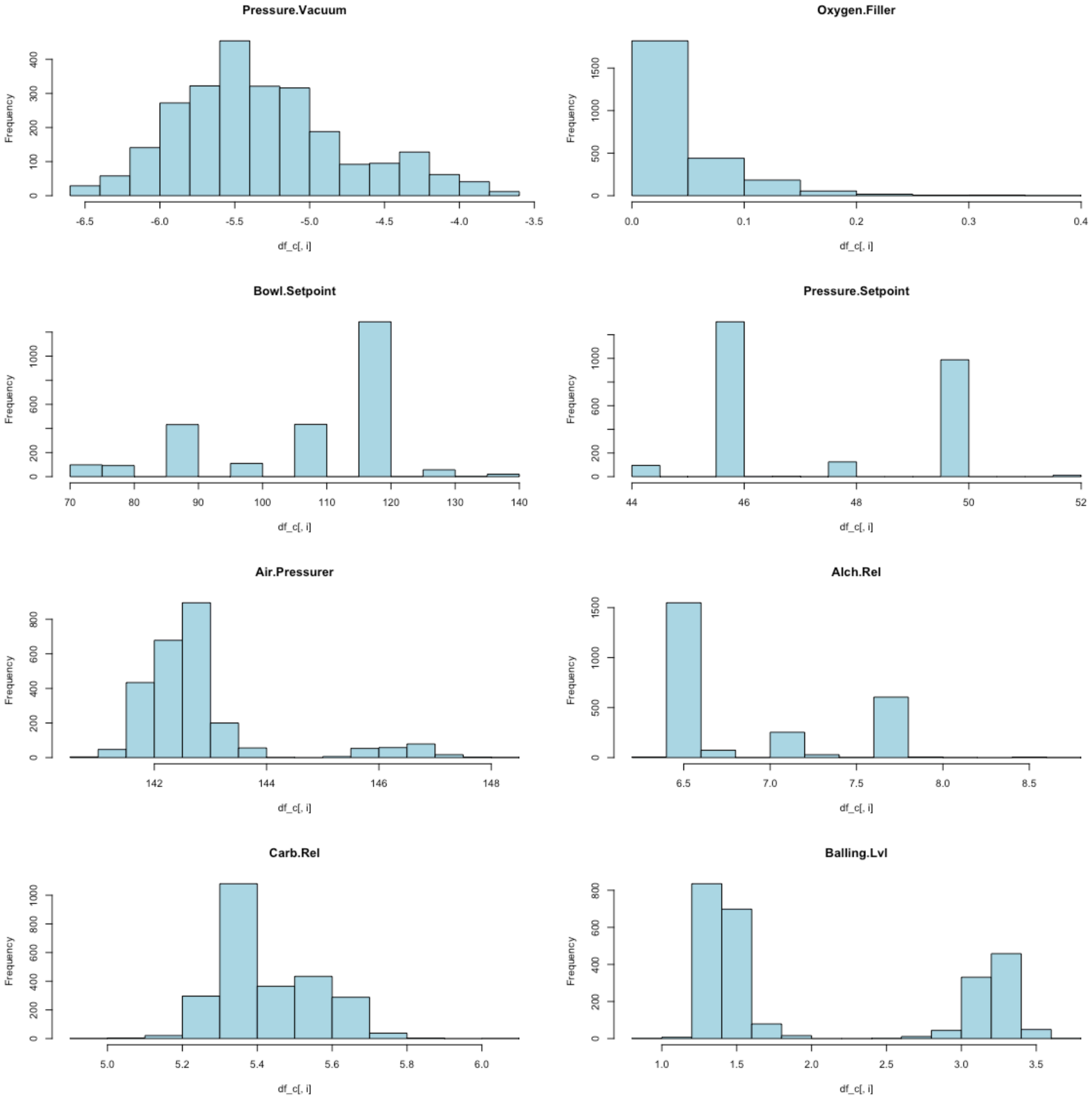
```r
#reorder cols to make PH last
df_c <- df_c[c(1,3:9,2)]
```

Exploratory plots

Histograms and density plots show the bimodal character of several variables: Bowl.Setpoint, Pressure.Setpoint, Alch.Rel, Carb.Rel, and Balling.Lvl.

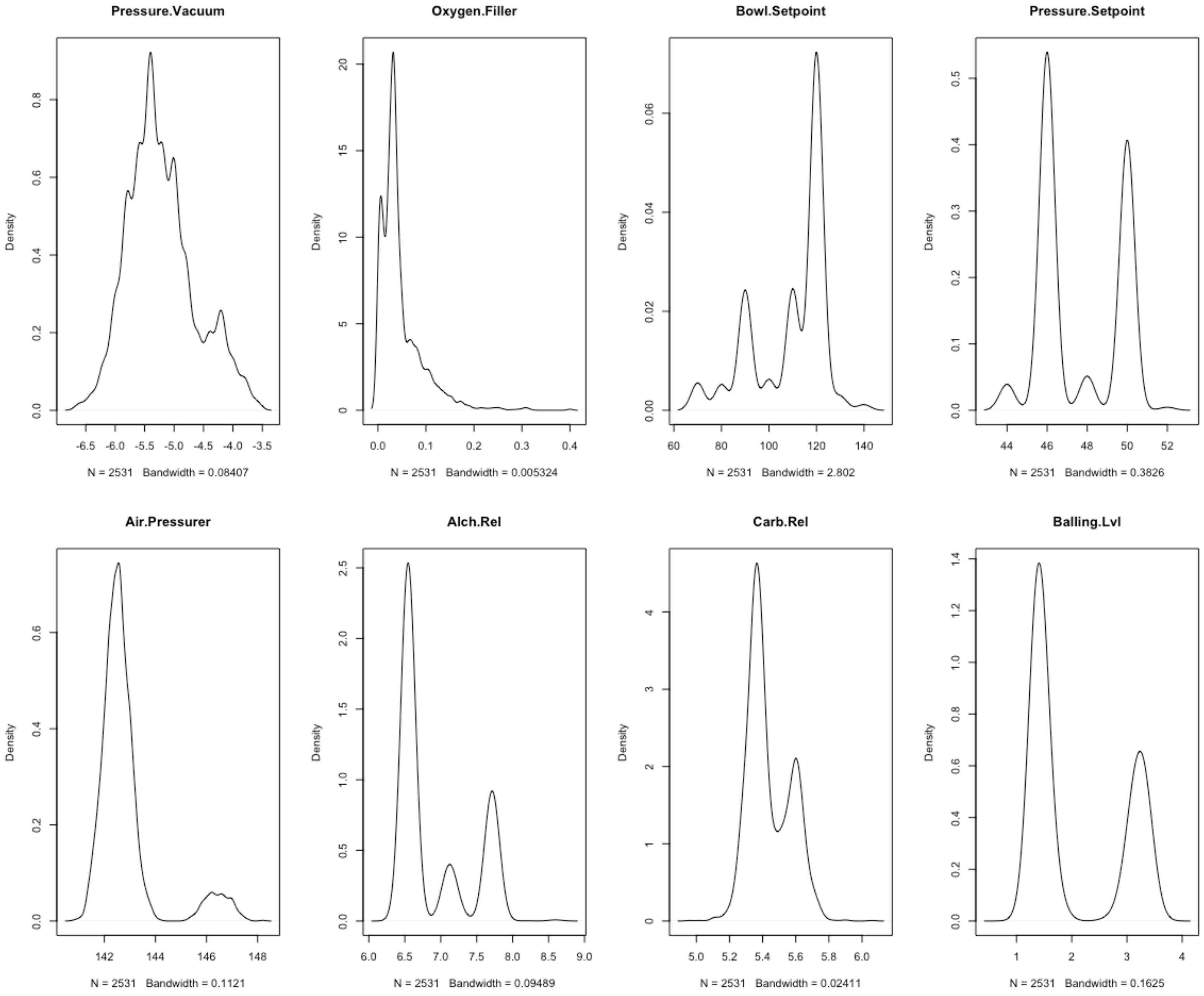
Histograms

```r
# histograms for each variable
par(mfrow=c(4,2))
for(i in c(1:8)) {
  hist(df_c[,i], main=names(df_c)[i], col="lightblue")
}
```



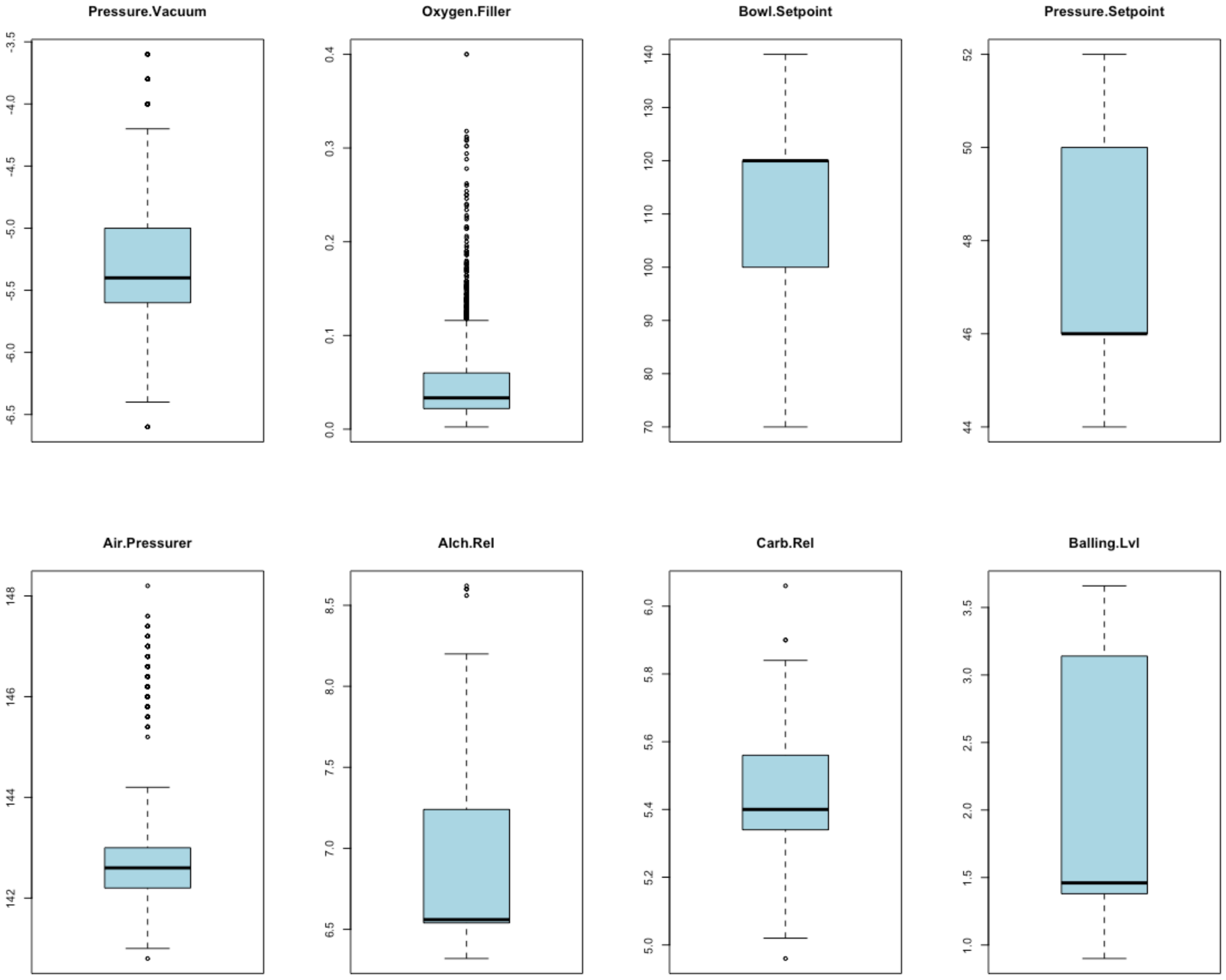
Density plots

```r
# density plot for each var
par(mfrow=c(2,4))
for(i in 1:8) {
  plot(density(df_c[,i]), main=names(df_c)[i])
}
```



Boxplots

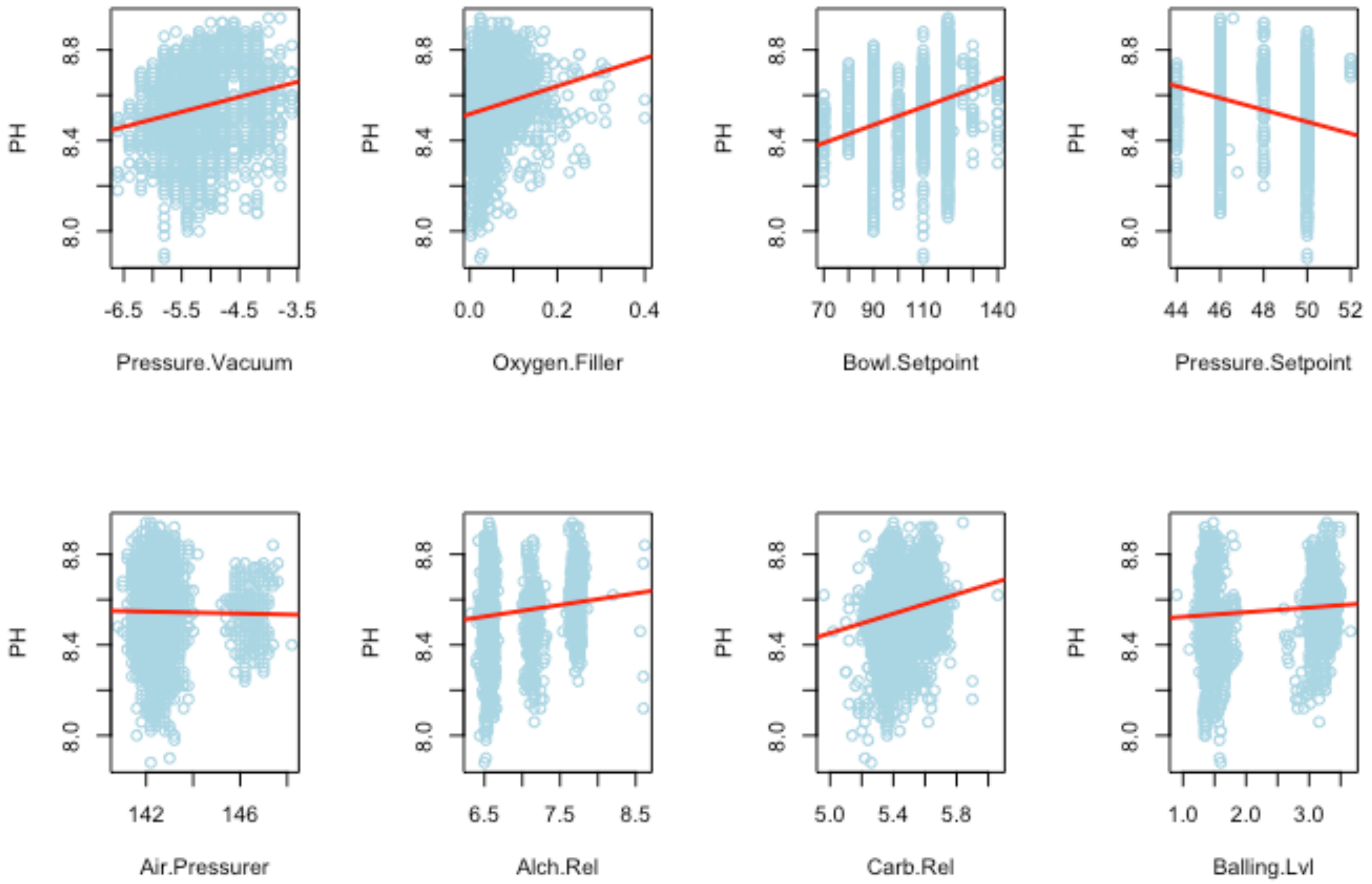
```
# boxplots
par(mfrow=c(2,4))
for(i in 1:8) {
  boxplot(df_c[,i], main=names(df_c)[i], col="lightblue")
}
```



XY plots

Pressure.Vacuum, Oxygen Filler, Bowl.Setpoint and Carb.Rel all have a postive association with PH; Pressure.Setpoint has a modest negative correlation.

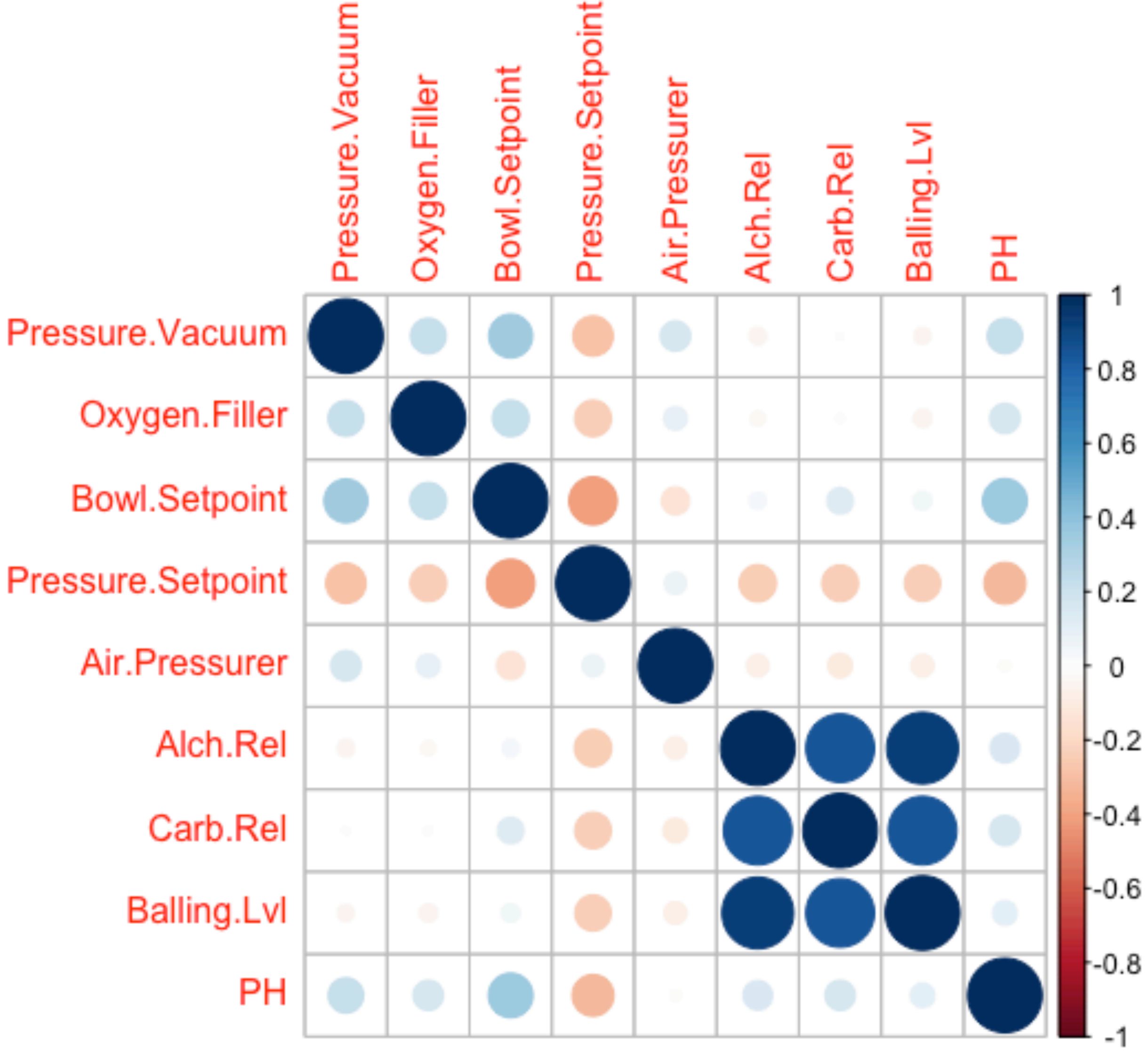
```
par(mfrow=c(2,4))
for(i in 1:8) {
  plot(df_c[,i], df_c$PH, xlab=colnames(df_c)[i], ylab="PH",
  abline(lsfit(df_c[,i], df_c$PH), col="red", lwd=2)
}
```



Correlation plot

Alch.Rel, Carb.Rel and Balling.Lvl are strongly correlated to each other. None of the vars is strongly correlated with PH.

```r
# correlation plot
correlations <- cor(df_c)
corrplot(correlations, method="circle")
```



Variable importance

We'll fit linear, Random Forest and Cubist models to test variable importance. Results suggest that Air.Pressure isn't an important variable and could be eliminated to simplify the model.

```
# fit models

# Run lm using 10-fold cross-validation
trainControl <- trainControl(method="repeatedcv", number=10,
metric <- "RMSE"

# LM
set.seed(100)
fit.lm <- train(PH~., data=df_c, method="lm", metric=metric,
                preProc=c("center", "scale"), trControl=trai

# Cubist
```

```
set.seed(100)
fit.cubist <- train(PH~., data=df_c, method="cubist", metric
                    ppreProc=c("center", "scale"), trControl

# Random Forest
set.seed(100)
fit.rf <- train(PH~., data=df_c, method="rf", metric=metric,
                preProc=c("center", "scale"),
                trControl=trainControl,
                importance=TRUE)
```

Error comparison

```
modelResults <- resamples(list(LM=fit.lm, Cubist=fit.cubist,

summary(modelResults)
```

```
##
## Call:
## summary.resamples(object = modelResults)
##
## Models: LM, Cubist, RF
## Number of resamples: 10
##
## MAE
##           Min. 1st Qu.   Median     Mean 3rd Qu.     Max. NA
## LM     0.11900 0.12220 0.12330 0.12420 0.12710 0.12940
## Cubist 0.06054 0.06673 0.07063 0.07019 0.07363 0.07854
## RF     0.06197 0.06360 0.06973 0.06789 0.07109 0.07371
##
## RMSE
##           Min. 1st Qu.   Median     Mean 3rd Qu.    Max. NA'
## LM     0.14600 0.15290 0.15550 0.15470  0.1573 0.1618
## Cubist 0.08309 0.09536 0.10110 0.09998  0.1051 0.1118
## RF     0.08235 0.09200 0.09832 0.09745  0.1039 0.1082
##
```

```
## Rsquared
##          Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
## LM     0.1261  0.1599 0.1953 0.1889  0.2172 0.2397    0
## Cubist 0.5803  0.6197 0.6515 0.6594  0.7103 0.7507    0
## RF     0.6130  0.6355 0.6687 0.6810  0.7355 0.7664    0
```

Varible importance, Cubist

```
varImp(fit.cubist)
```

```
## cubist variable importance
## 
##                   Overall
## Bowl.Setpoint     100.000
## Balling.Lvl        78.261
## Oxygen.Filler      52.174
## Pressure.Vacuum    50.000
## Alch.Rel           48.913
## Carb.Rel           32.609
## Air.Pressurer       7.609
## Pressure.Setpoint   0.000
```

Variable importance plot, Random Forest

```
varImp(fit.rf)
```

```
## rf variable importance
## 
##                   Overall
## Bowl.Setpoint      100.00
## Pressure.Vacuum     90.33
## Oxygen.Filler       88.08
## Balling.Lvl         87.07
## Alch.Rel            79.93
## Carb.Rel            74.83
## Air.Pressurer       65.54
## Pressure.Setpoint    0.00
```

Variable importance, linear fit

Results show that Pressure.Vaccum, Bowl.Setpoint, Pressure.Setpoint, Alch.Rel and Balling.Lvl all are significant at p <.001. In this model, these predictors explain ony 19 percent of the variation in PH.

```
summary(fit.lm)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58146 -0.09526  0.01523  0.11277  0.32954
##
## Coefficients:
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)        8.545508   0.003075 2779.390  < 2e-16
## Pressure.Vacuum    0.013896   0.003452    4.025 5.86e-05
## Oxygen.Filler      0.007158   0.003270    2.189   0.0287
## Bowl.Setpoint      0.043085   0.003609   11.940  < 2e-16
## Pressure.Setpoint -0.025479   0.003602   -7.074 1.94e-12
## Air.Pressurer      0.003919   0.003218    1.218   0.2234
## Alch.Rel           0.057046   0.008675    6.575 5.87e-11
## Carb.Rel           0.014521   0.006175    2.352   0.0188
## Balling.Lvl       -0.054377   0.008642   -6.292 3.68e-10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.1547 on 2522 degrees of freedo
## Multiple R-squared:  0.1903, Adjusted R-squared:  0.1877
## F-statistic: 74.07 on 8 and 2522 DF,  p-value: < 2.2e-16
```