Group#2 Final

Group 2

4/30/2018

# Introduction

Our project team has obtained a data set containing 2,571 observations with each observation representing a batch of beverages that were produced. Variables included in the data set provide information on the manufacturing process that resulted in the creation of each batch. It is our task to use the 32 predictor variables to develop a model which can be used to predict the PH level in a batch. We have been provided with a training and test set.

The objective is to build multiple models on the training data and then we will then run analysis to determine which model performed best.

## Team Members (alphabetical)

* Brian Kreis
* Bin Lin
* Chris Martin
* Asher Myers

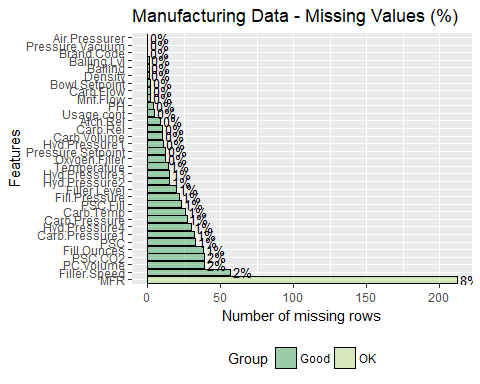
# Dataset

For reproducibility of the results, the data was loaded to and accessed from a Github repository.

# Data Exploration and Statistic Measures

The purpose of the data exploration and statistic measures phase is to understand the data to determine how to process the dataset for modelling.

## Missing and Zero Values



|  |  |
| --- | --- |
|  | x |
| Brand.Code | 0 |
| Carb.Volume | 10 |
| Fill.Ounces | 38 |
| PC.Volume | 39 |
| Carb.Pressure | 27 |
| Carb.Temp | 26 |
| PSC | 33 |
| PSC.Fill | 23 |
| PSC.CO2 | 39 |
| Mnf.Flow | 2 |
| Carb.Pressure1 | 32 |
| Fill.Pressure | 22 |
| Hyd.Pressure1 | 11 |
| Hyd.Pressure2 | 15 |
| Hyd.Pressure3 | 15 |
| Hyd.Pressure4 | 30 |
| Filler.Level | 20 |
| Filler.Speed | 57 |
| Temperature | 14 |
| Usage.cont | 5 |
| Carb.Flow | 2 |
| Density | 1 |
| MFR | 212 |
| Balling | 1 |
| 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 | 1 |

## # A tibble: 5 x 33  
## Brand.Code Carb.Volume Fill.Ounces PC.Volume Carb.Pressure Carb.Temp  
## <fct> <int> <int> <int> <int> <int>  
## 1 "" 0 4 4 1 3  
## 2 A 0 8 7 2 1  
## 3 B 5 16 17 10 15  
## 4 C 1 4 4 2 2  
## 5 D 4 6 7 12 5  
## # ... with 27 more variables: PSC <int>, PSC.Fill <int>, PSC.CO2 <int>,  
## # Mnf.Flow <int>, Carb.Pressure1 <int>, Fill.Pressure <int>,  
## # Hyd.Pressure1 <int>, Hyd.Pressure2 <int>, Hyd.Pressure3 <int>,  
## # Hyd.Pressure4 <int>, Filler.Level <int>, Filler.Speed <int>,  
## # Temperature <int>, Usage.cont <int>, Carb.Flow <int>, Density <int>,  
## # MFR <int>, Balling <int>, Pressure.Vacuum <int>, PH <int>,  
## # Oxygen.Filler <int>, Bowl.Setpoint <int>, Pressure.Setpoint <int>,  
## # Air.Pressurer <int>, Alch.Rel <int>, Carb.Rel <int>, Balling.Lvl <int>

##   
## A B C D   
## 120 293 1239 304 615

We have a number of missing values. Presumably, because we have identified NA values across various brands, both named and unnamed, we would expect that the NA values are not informative. In other words, if information for a particular brand is usually recorded, we would expect failure to record information to be the result of an error and not a typical process change for a particular brand. As our knowledge of the production process itself is limited, we will rely on your subject matter expertise to let us know if this presumption is incorrect.

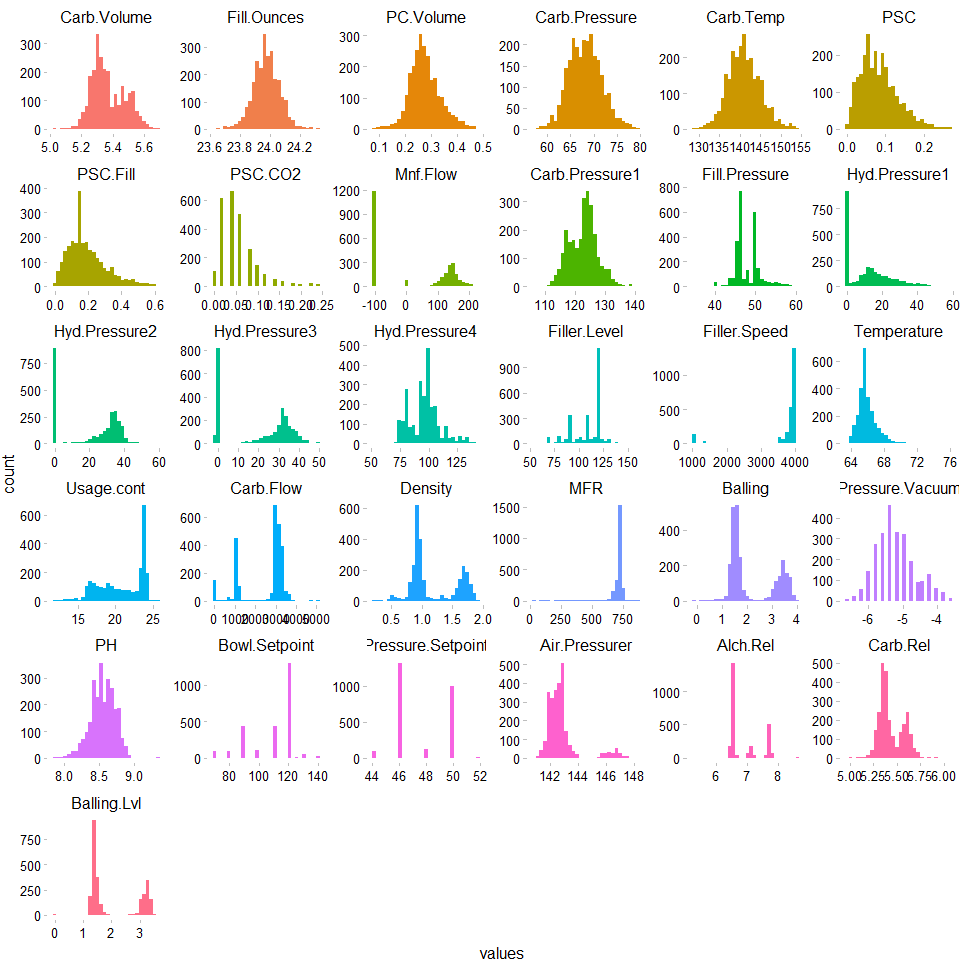
MFR stands out as a variable with a significant number of missing values; however, the percentage of missing values is still low enough where imputation is not unreasonable. We will likely find it necessary to drop the few observations that have an NA value for the response variable.

## Descriptive Statistics and Data Exploration

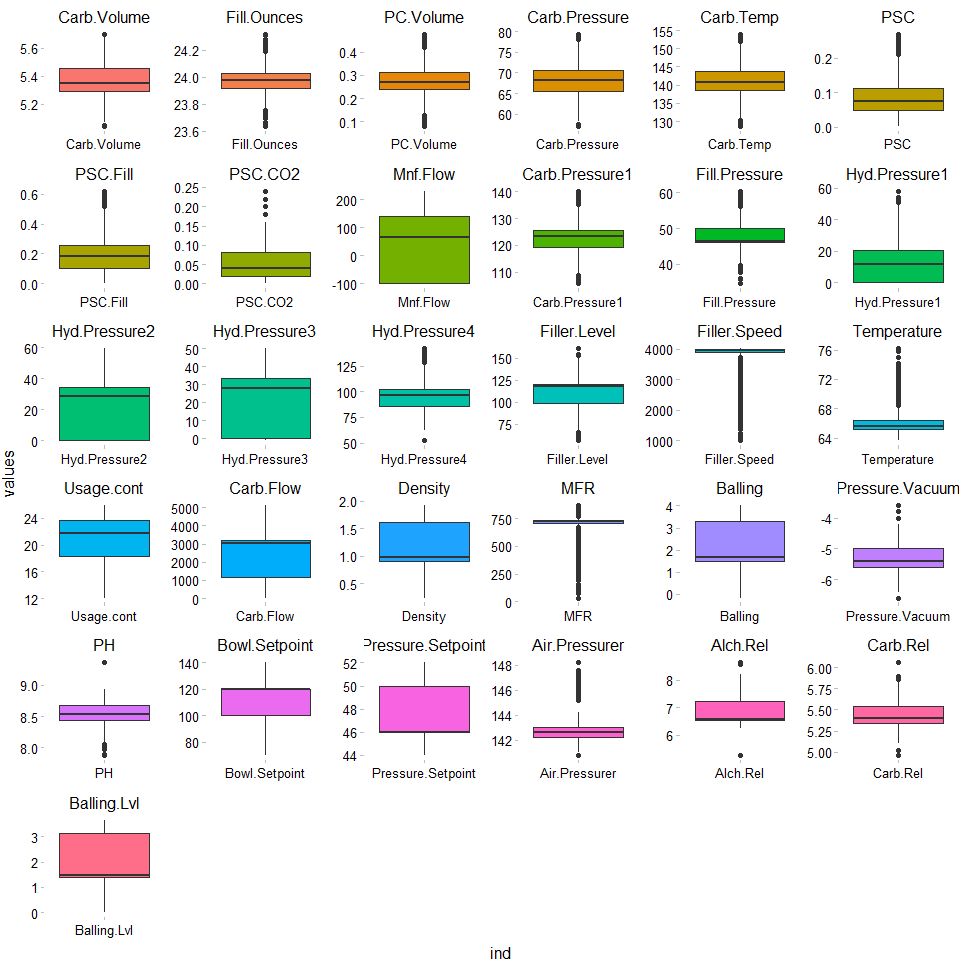
Descriptive statistics was performed for all predictor and response variables to explore the data.

## data1   
##   
## 33 Variables 2571 Observations  
## ---------------------------------------------------------------------------  
## Brand.Code   
## n missing distinct   
## 2571 0 5   
##   
## Value A B C D  
## Frequency 120 293 1239 304 615  
## Proportion 0.047 0.114 0.482 0.118 0.239  
## ---------------------------------------------------------------------------  
## Carb.Volume   
## n missing distinct Info Mean Gmd .05 .10   
## 2561 10 101 1 5.37 0.1198 5.220 5.253   
## .25 .50 .75 .90 .95   
## 5.293 5.347 5.453 5.527 5.553   
##   
## lowest : 5.040000 5.046667 5.066667 5.080000 5.093333  
## highest: 5.646667 5.653333 5.666667 5.673333 5.700000  
## ---------------------------------------------------------------------------  
## Fill.Ounces   
## n missing distinct Info Mean Gmd .05 .10   
## 2533 38 92 0.999 23.97 0.09699 23.83 23.87   
## .25 .50 .75 .90 .95   
## 23.92 23.97 24.03 24.09 24.11   
##   
## lowest : 23.63333 23.65333 23.66667 23.69333 23.70000  
## highest: 24.27333 24.28000 24.30667 24.31333 24.32000  
## ---------------------------------------------------------------------------  
## PC.Volume   
## n missing distinct Info Mean Gmd .05 .10   
## 2532 39 454 1 0.2771 0.06689 0.1910 0.2100   
## .25 .50 .75 .90 .95   
## 0.2392 0.2713 0.3120 0.3579 0.3890   
##   
## lowest : 0.07933333 0.08666667 0.09000000 0.09133333 0.09533333  
## highest: 0.46600000 0.46800000 0.46933333 0.47066667 0.47800000  
## ---------------------------------------------------------------------------  
## Carb.Pressure   
## n missing distinct Info Mean Gmd .05 .10   
## 2544 27 106 1 68.19 3.991 62.8 63.8   
## .25 .50 .75 .90 .95   
## 65.6 68.2 70.6 72.8 74.0   
##   
## lowest : 57.0 57.6 58.2 58.4 58.6, highest: 78.2 78.4 78.6 78.8 79.4  
## ---------------------------------------------------------------------------  
## Carb.Temp   
## n missing distinct Info Mean Gmd .05 .10   
## 2545 26 123 1 141.1 4.522 135.0 136.2   
## .25 .50 .75 .90 .95   
## 138.4 140.8 143.8 146.4 148.2   
##   
## lowest : 128.6 129.0 129.2 129.6 129.8, highest: 153.0 153.4 153.6 153.8 154.0  
## ---------------------------------------------------------------------------  
## PSC   
## n missing distinct Info Mean Gmd .05 .10   
## 2538 33 129 1 0.08457 0.05443 0.018 0.028   
## .25 .50 .75 .90 .95   
## 0.048 0.076 0.112 0.152 0.178   
##   
## lowest : 0.002 0.004 0.006 0.008 0.010, highest: 0.260 0.264 0.266 0.268 0.270  
## ---------------------------------------------------------------------------  
## PSC.Fill   
## n missing distinct Info Mean Gmd .05 .10   
## 2548 23 32 0.997 0.1954 0.1291 0.04 0.06   
## .25 .50 .75 .90 .95   
## 0.10 0.18 0.26 0.36 0.42   
##   
## lowest : 0.00 0.02 0.04 0.06 0.08, highest: 0.54 0.56 0.58 0.60 0.62  
## ---------------------------------------------------------------------------  
## PSC.CO2   
## n missing distinct Info Mean Gmd .05 .10   
## 2532 39 13 0.959 0.05641 0.04319 0.02 0.02   
## .25 .50 .75 .90 .95   
## 0.02 0.04 0.08 0.10 0.14   
##   
## Value 0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14 0.16 0.18  
## Frequency 108 613 661 502 257 144 82 49 39 20  
## Proportion 0.043 0.242 0.261 0.198 0.102 0.057 0.032 0.019 0.015 0.008  
##   
## Value 0.20 0.22 0.24  
## Frequency 18 22 17  
## Proportion 0.007 0.009 0.007  
## ---------------------------------------------------------------------------  
## Mnf.Flow   
## n missing distinct Info Mean Gmd .05 .10   
## 2569 2 487 0.975 24.57 127.3 -100.2 -100.2   
## .25 .50 .75 .90 .95   
## -100.0 65.2 140.8 158.0 175.1   
##   
## lowest : -100.2 -100.0 0.2 2.0 4.8, highest: 210.2 213.4 216.2 223.0 229.4  
## ---------------------------------------------------------------------------  
## Carb.Pressure1   
## n missing distinct Info Mean Gmd .05 .10   
## 2539 32 140 1 122.6 5.32 115.0 116.4   
## .25 .50 .75 .90 .95   
## 119.0 123.2 125.4 128.2 130.4   
##   
## lowest : 105.6 106.4 107.0 107.6 108.8, highest: 138.2 138.4 139.0 139.4 140.2  
## ---------------------------------------------------------------------------  
## Fill.Pressure   
## n missing distinct Info Mean Gmd .05 .10   
## 2549 22 108 0.993 47.92 3.353 44.00 45.56   
## .25 .50 .75 .90 .95   
## 46.00 46.40 50.00 51.20 53.92   
##   
## lowest : 34.6 35.8 36.0 36.2 37.8, highest: 58.8 59.2 59.6 60.0 60.4  
## ---------------------------------------------------------------------------  
## Hyd.Pressure1   
## n missing distinct Info Mean Gmd .05 .10   
## 2560 11 245 0.965 12.44 13.63 0.0 0.0   
## .25 .50 .75 .90 .95   
## 0.0 11.4 20.2 30.2 36.6   
##   
## lowest : -0.8 -0.6 -0.4 -0.2 0.0, highest: 51.8 52.2 52.4 53.8 58.0  
## ---------------------------------------------------------------------------  
## Hyd.Pressure2   
## n missing distinct Info Mean Gmd .05 .10   
## 2556 15 207 0.972 20.96 17.81 0.0 0.0   
## .25 .50 .75 .90 .95   
## 0.0 28.6 34.6 38.0 39.6   
##   
## lowest : 0.0 0.2 0.6 2.0 2.2, highest: 56.8 57.0 57.2 57.8 59.4  
## ---------------------------------------------------------------------------  
## Hyd.Pressure3   
## n missing distinct Info Mean Gmd .05 .10   
## 2556 15 192 0.968 20.46 17.47 0.00 0.00   
## .25 .50 .75 .90 .95   
## 0.00 27.60 33.40 37.60 40.05   
##   
## lowest : -1.2 0.0 1.6 2.6 4.4, highest: 49.2 49.4 49.6 49.8 50.0  
## ---------------------------------------------------------------------------  
## Hyd.Pressure4   
## n missing distinct Info Mean Gmd .05 .10   
## 2541 30 40 0.996 96.29 14.38 76 80   
## .25 .50 .75 .90 .95   
## 86 96 102 114 122   
##   
## lowest : 52 62 64 70 72, highest: 134 136 138 140 142  
## ---------------------------------------------------------------------------  
## Filler.Level   
## n missing distinct Info Mean Gmd .05 .10   
## 2551 20 288 0.999 109.3 16.53 79.8 88.6   
## .25 .50 .75 .90 .95   
## 98.3 118.4 120.0 120.8 122.7   
##   
## lowest : 55.8 59.0 60.8 64.0 64.4, highest: 151.8 152.4 153.2 154.0 161.2  
## ---------------------------------------------------------------------------  
## Filler.Speed   
## n missing distinct Info Mean Gmd .05 .10   
## 2514 57 244 0.999 3687 529.8 1015 3392   
## .25 .50 .75 .90 .95   
## 3888 3982 3998 4010 4014   
##   
## lowest : 998 1002 1004 1006 1008, highest: 4022 4024 4026 4028 4030  
## ---------------------------------------------------------------------------  
## Temperature   
## n missing distinct Info Mean Gmd .05 .10   
## 2557 14 56 0.996 65.97 1.352 64.4 64.6   
## .25 .50 .75 .90 .95   
## 65.2 65.6 66.4 67.4 68.2   
##   
## lowest : 63.6 63.8 64.0 64.2 64.4, highest: 74.2 75.0 75.8 76.0 76.2  
## ---------------------------------------------------------------------------  
## Usage.cont   
## n missing distinct Info Mean Gmd .05 .10   
## 2566 5 481 1 20.99 3.309 16.20 16.68   
## .25 .50 .75 .90 .95   
## 18.36 21.79 23.75 24.02 24.14   
##   
## lowest : 12.08 12.46 12.72 12.80 12.82, highest: 24.38 24.86 25.08 25.78 25.90  
## ---------------------------------------------------------------------------  
## Carb.Flow   
## n missing distinct Info Mean Gmd .05 .10   
## 2569 2 533 1 2468 1080 46 1052   
## .25 .50 .75 .90 .95   
## 1144 3028 3186 3308 3409   
##   
## lowest : 26 28 30 32 34, highest: 3832 3834 3846 4666 5104  
## ---------------------------------------------------------------------------  
## Density   
## n missing distinct Info Mean Gmd .05 .10   
## 2570 1 78 0.998 1.174 0.4044 0.80 0.86   
## .25 .50 .75 .90 .95   
## 0.90 0.98 1.62 1.74 1.78   
##   
## lowest : 0.24 0.30 0.34 0.44 0.46, highest: 1.84 1.86 1.88 1.90 1.92  
## ---------------------------------------------------------------------------  
## MFR   
## n missing distinct Info Mean Gmd .05 .10   
## 2359 212 587 1 704 47.92 644.3 672.0   
## .25 .50 .75 .90 .95   
## 706.3 724.0 731.0 735.8 740.6   
##   
## lowest : 31.4 76.8 95.4 112.6 114.2, highest: 814.6 835.4 840.0 849.4 868.6  
## ---------------------------------------------------------------------------  
## Balling   
## n missing distinct Info Mean Gmd .05 .10   
## 2570 1 217 0.999 2.198 0.9866 1.298 1.350   
## .25 .50 .75 .90 .95   
## 1.496 1.648 3.292 3.588 3.714   
##   
## lowest : -0.170 0.160 0.346 0.448 0.648, highest: 3.914 3.934 3.964 3.984 4.012  
## ---------------------------------------------------------------------------  
## Pressure.Vacuum   
## n missing distinct Info Mean Gmd .05 .10   
## 2571 0 16 0.986 -5.216 0.6333 -6.0 -5.8   
## .25 .50 .75 .90 .95   
## -5.6 -5.4 -5.0 -4.4 -4.2   
##   
## Value -6.6 -6.4 -6.2 -6.0 -5.8 -5.6 -5.4 -5.2 -5.0 -4.8  
## Frequency 9 20 58 143 275 327 460 331 322 192  
## Proportion 0.004 0.008 0.023 0.056 0.107 0.127 0.179 0.129 0.125 0.075  
##   
## Value -4.6 -4.4 -4.2 -4.0 -3.8 -3.6  
## Frequency 92 96 130 63 41 12  
## Proportion 0.036 0.037 0.051 0.025 0.016 0.005  
## ---------------------------------------------------------------------------  
## PH   
## n missing distinct Info Mean Gmd .05 .10   
## 2567 4 52 0.999 8.546 0.1946 8.26 8.32   
## .25 .50 .75 .90 .95   
## 8.44 8.54 8.68 8.76 8.80   
##   
## lowest : 7.88 7.90 7.98 8.00 8.02, highest: 8.88 8.90 8.92 8.94 9.36  
## ---------------------------------------------------------------------------  
## Oxygen.Filler   
## n missing distinct Info Mean Gmd .05 .10   
## 2559 12 338 0.999 0.04684 0.04386 0.00260 0.00476   
## .25 .50 .75 .90 .95   
## 0.02200 0.03340 0.06000 0.10440 0.13600   
##   
## lowest : 0.0024 0.0026 0.0028 0.0034 0.0036, highest: 0.3080 0.3100 0.3120 0.3180 0.4000  
## ---------------------------------------------------------------------------  
## Bowl.Setpoint   
## n missing distinct Info Mean Gmd .05 .10   
## 2569 2 11 0.858 109.3 15.68 80 90   
## .25 .50 .75 .90 .95   
## 100 120 120 120 120   
##   
## Value 70 80 90 100 110 120 122 126 130 134  
## Frequency 99 96 434 112 437 1307 1 10 51 2  
## Proportion 0.039 0.037 0.169 0.044 0.170 0.509 0.000 0.004 0.020 0.001  
##   
## Value 140  
## Frequency 20  
## Proportion 0.008  
## ---------------------------------------------------------------------------  
## Pressure.Setpoint   
## n missing distinct Info Mean Gmd   
## 2559 12 8 0.802 47.62 2.107   
##   
## Value 44.0 46.0 46.4 46.6 46.8 48.0 50.0 52.0  
## Frequency 96 1322 1 1 1 125 1002 11  
## Proportion 0.038 0.517 0.000 0.000 0.000 0.049 0.392 0.004  
## ---------------------------------------------------------------------------  
## Air.Pressurer   
## n missing distinct Info Mean Gmd .05 .10   
## 2571 0 32 0.989 142.8 1.09 141.6 141.8   
## .25 .50 .75 .90 .95   
## 142.2 142.6 143.0 143.6 146.2   
##   
## lowest : 140.8 141.0 141.2 141.4 141.6, highest: 147.0 147.2 147.4 147.6 148.2  
## ---------------------------------------------------------------------------  
## Alch.Rel   
## n missing distinct Info Mean Gmd .05 .10   
## 2562 9 53 0.99 6.897 0.5078 6.50 6.52   
## .25 .50 .75 .90 .95   
## 6.54 6.56 7.24 7.72 7.76   
##   
## lowest : 5.28 6.24 6.32 6.40 6.42, highest: 7.86 8.20 8.56 8.60 8.62  
## ---------------------------------------------------------------------------  
## Carb.Rel   
## n missing distinct Info Mean Gmd .05 .10   
## 2561 10 42 0.996 5.437 0.1433 5.28 5.30   
## .25 .50 .75 .90 .95   
## 5.34 5.40 5.54 5.62 5.66   
##   
## lowest : 4.96 5.02 5.10 5.12 5.14, highest: 5.80 5.84 5.86 5.90 6.06  
## ---------------------------------------------------------------------------  
## Balling.Lvl   
## n missing distinct Info Mean Gmd .05 .10   
## 2570 1 82 0.998 2.05 0.8902 1.32 1.34   
## .25 .50 .75 .90 .95   
## 1.38 1.48 3.14 3.32 3.36   
##   
## lowest : 0.00 0.54 0.90 1.10 1.18, highest: 3.44 3.46 3.48 3.50 3.66  
## ---------------------------------------------------------------------------

We have siginificant skewness in a number of variables, and, depending on our choice of regression technique may require transformation.



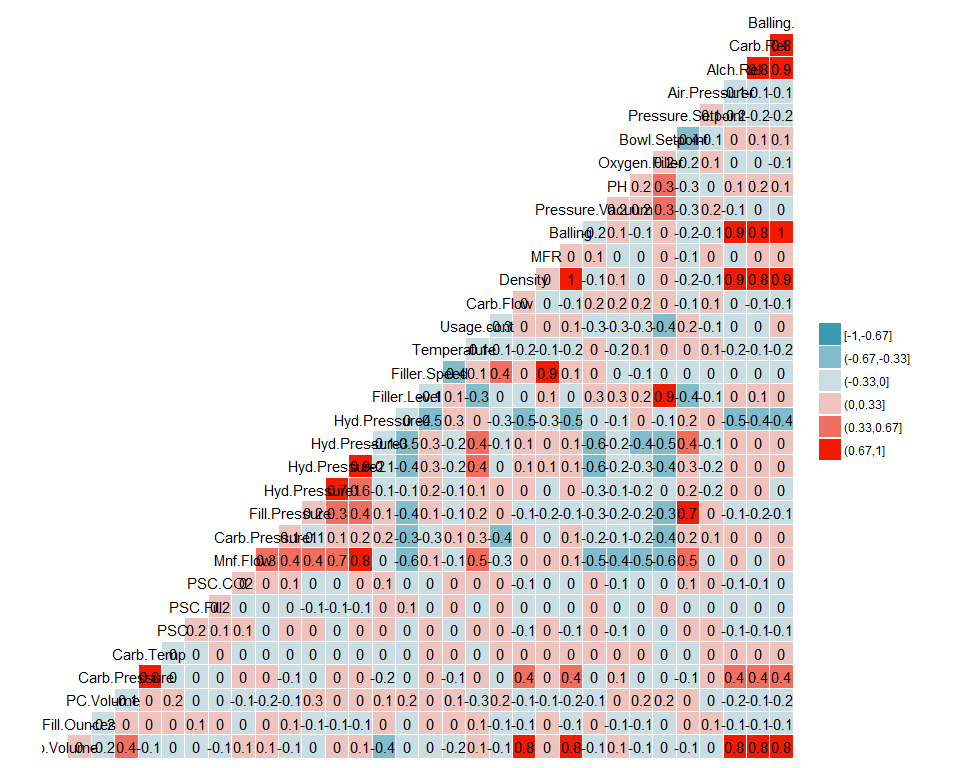
We have a varying mix of distributions for the different data points. For example, - Discrete: The variable Pressure.Setpoint appears to represent a categorical setting ranging from 44 to 52; however, we do not know at first glance if our training set contains all possible values that we might see in the test set or in general practice.  
- Normal Continuous: Carb.Temp appears to be normally distributed. - Multi Modal data: Density appears to have a large number of values around 1 and another large concentration of values around 1.5.



Ther boxplots above demonstrate the skewness of a number of variables as we have discussed.

## Correlation

The tables below represent correlation between response and predictor variables. There is high collinearity among certain variables. For instance in Density and Alch.Rel. This is also true of Alch.Res and Brand D.



Below is a look at the unique values for each variable. As mentioned, it appears that some of these variables represent various non continuous settings.

## Brand.Code Carb.Volume Fill.Ounces PC.Volume   
## 5 102 93 455   
## Carb.Pressure Carb.Temp PSC PSC.Fill   
## 107 124 130 33   
## PSC.CO2 Mnf.Flow Carb.Pressure1 Fill.Pressure   
## 14 488 141 109   
## Hyd.Pressure1 Hyd.Pressure2 Hyd.Pressure3 Hyd.Pressure4   
## 246 208 193 41   
## Filler.Level Filler.Speed Temperature Usage.cont   
## 289 245 57 482   
## Carb.Flow Density MFR Balling   
## 534 79 588 218   
## Pressure.Vacuum PH Oxygen.Filler Bowl.Setpoint   
## 16 53 339 12   
## Pressure.Setpoint Air.Pressurer Alch.Rel Carb.Rel   
## 9 32 54 43   
## Balling.Lvl   
## 83

## Data Manipulation

### Dummy Variables

data1$A <- ifelse(data1$Brand.Code == "A", 1, 0)  
data1$B <- ifelse(data1$Brand.Code == "B", 1, 0)  
data1$C <- ifelse(data1$Brand.Code == "C", 1, 0)  
data1$D <- ifelse(data1$Brand.Code == "D", 1, 0)  
data1 <- data1 %>% select(-Brand.Code)  
  
  
toPred$A <- ifelse(toPred$Brand.Code == "A", 1, 0)  
toPred$B <- ifelse(toPred$Brand.Code == "B", 1, 0)  
toPred$C <- ifelse(toPred$Brand.Code == "C", 1, 0)  
toPred$D <- ifelse(toPred$Brand.Code == "D", 1, 0)  
toPred <- toPred %>% select(-Brand.Code)

### Handling Missing Values

missingPH <- which(is.na(data1$PH))  
data1 <- data1[-missingPH, ]  
  
  
#The following will impute the data. For efficiency the data sets can be loaded from github below.  
# ##For Windows to run in parallel  
# library(parallel)  
# library(doParallel)  
#   
# cluster <- makeCluster(detectCores() - 1) # convention to leave 1 core for OS  
# registerDoParallel(cluster)  
#   
# #impute missing training data  
# set.seed(123)  
# dfImputed <- missForest(data1, parallelize = 'forests')  
# write.csv(dfImputed$ximp, "StudentDataImputedMF")  
#   
# #impute missing test data  
# set.seed(123)  
# predImputed <- missForest(toPred, parallelize = 'forests')  
# write.csv(predImputed$ximp, "PredictImputedMF")  
#   
#   
# #turn off parallel processing  
# stopCluster(cluster)  
# #resume use of the sequential backend  
# registerDoSEQ()  
  
#read imputed train set  
imputed <- read.csv("https://raw.githubusercontent.com/624-Group2/Project-2/master/StudentDataImputedMF")  
imputed <- imputed %>% select(-X)  
sum(is.na(imputed))

## [1] 0

#read imputed test set  
predictImp <- read.csv("https://raw.githubusercontent.com/624-Group2/Project-2/master/PredictImputedMF")  
predictImp <- predictImp %>% select(-X)  
sum(is.na(predictImp))

## [1] 0

# Test & Training Sets

Prior to testing our models on the actual prediction data set, it is prudent to evaluate our models against data where the response is known, so our predictions can be compared.

smp <- floor(0.70 \* nrow(imputed))  
  
  
set.seed(123)  
train\_index <- sample(seq\_len(nrow(imputed)), size = smp, replace = FALSE)  
  
train\_set <- imputed[train\_index, ]  
validation\_set <- imputed[-train\_index, ]

# Linear Regression

Box-Cox Temperature, Filler.Speed, Filler.Level, Hyd.Pressure1, Hyd.Pressure2, Hyd.Pressure3, Mnf.Flow, Balling, MFR, Usage.cont, Carb.Flow, Density, Air.Pressure

## Ordinary Linear Regression

lmTune <- train(PH ~ ., data=imputed, method='lm', preProcess=c('BoxCox', 'center', 'scale'))

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =  
## trainInfo, : There were missing values in resampled performance measures.

lmTune$terms

## PH ~ Carb.Volume + Fill.Ounces + PC.Volume + Carb.Pressure +   
## Carb.Temp + PSC + PSC.Fill + PSC.CO2 + Mnf.Flow + Carb.Pressure1 +   
## Fill.Pressure + Hyd.Pressure1 + Hyd.Pressure2 + Hyd.Pressure3 +   
## Hyd.Pressure4 + Filler.Level + Filler.Speed + Temperature +   
## Usage.cont + Carb.Flow + Density + MFR + Balling + Pressure.Vacuum +   
## Oxygen.Filler + Bowl.Setpoint + Pressure.Setpoint + Air.Pressurer +   
## Alch.Rel + Carb.Rel + Balling.Lvl + A + B + C + D  
## attr(,"variables")  
## list(PH, Carb.Volume, Fill.Ounces, PC.Volume, Carb.Pressure,   
## Carb.Temp, PSC, PSC.Fill, PSC.CO2, Mnf.Flow, Carb.Pressure1,   
## Fill.Pressure, Hyd.Pressure1, Hyd.Pressure2, Hyd.Pressure3,   
## Hyd.Pressure4, Filler.Level, Filler.Speed, Temperature, Usage.cont,   
## Carb.Flow, Density, MFR, Balling, Pressure.Vacuum, Oxygen.Filler,   
## Bowl.Setpoint, Pressure.Setpoint, Air.Pressurer, Alch.Rel,   
## Carb.Rel, Balling.Lvl, A, B, C, D)  
## attr(,"factors")  
## Carb.Volume Fill.Ounces PC.Volume Carb.Pressure  
## PH 0 0 0 0  
## Carb.Volume 1 0 0 0  
## Fill.Ounces 0 1 0 0  
## PC.Volume 0 0 1 0  
## Carb.Pressure 0 0 0 1  
## Carb.Temp 0 0 0 0  
## PSC 0 0 0 0  
## PSC.Fill 0 0 0 0  
## PSC.CO2 0 0 0 0  
## Mnf.Flow 0 0 0 0  
## Carb.Pressure1 0 0 0 0  
## Fill.Pressure 0 0 0 0  
## Hyd.Pressure1 0 0 0 0  
## Hyd.Pressure2 0 0 0 0  
## Hyd.Pressure3 0 0 0 0  
## Hyd.Pressure4 0 0 0 0  
## Filler.Level 0 0 0 0  
## Filler.Speed 0 0 0 0  
## Temperature 0 0 0 0  
## Usage.cont 0 0 0 0  
## Carb.Flow 0 0 0 0  
## Density 0 0 0 0  
## MFR 0 0 0 0  
## Balling 0 0 0 0  
## Pressure.Vacuum 0 0 0 0  
## Oxygen.Filler 0 0 0 0  
## Bowl.Setpoint 0 0 0 0  
## Pressure.Setpoint 0 0 0 0  
## Air.Pressurer 0 0 0 0  
## Alch.Rel 0 0 0 0  
## Carb.Rel 0 0 0 0  
## Balling.Lvl 0 0 0 0  
## A 0 0 0 0  
## B 0 0 0 0  
## C 0 0 0 0  
## D 0 0 0 0  
## Carb.Temp PSC PSC.Fill PSC.CO2 Mnf.Flow Carb.Pressure1  
## PH 0 0 0 0 0 0  
## Carb.Volume 0 0 0 0 0 0  
## Fill.Ounces 0 0 0 0 0 0  
## PC.Volume 0 0 0 0 0 0  
## Carb.Pressure 0 0 0 0 0 0  
## Carb.Temp 1 0 0 0 0 0  
## PSC 0 1 0 0 0 0  
## PSC.Fill 0 0 1 0 0 0  
## PSC.CO2 0 0 0 1 0 0  
## Mnf.Flow 0 0 0 0 1 0  
## Carb.Pressure1 0 0 0 0 0 1  
## Fill.Pressure 0 0 0 0 0 0  
## Hyd.Pressure1 0 0 0 0 0 0  
## Hyd.Pressure2 0 0 0 0 0 0  
## Hyd.Pressure3 0 0 0 0 0 0  
## Hyd.Pressure4 0 0 0 0 0 0  
## Filler.Level 0 0 0 0 0 0  
## Filler.Speed 0 0 0 0 0 0  
## Temperature 0 0 0 0 0 0  
## Usage.cont 0 0 0 0 0 0  
## Carb.Flow 0 0 0 0 0 0  
## Density 0 0 0 0 0 0  
## MFR 0 0 0 0 0 0  
## Balling 0 0 0 0 0 0  
## Pressure.Vacuum 0 0 0 0 0 0  
## Oxygen.Filler 0 0 0 0 0 0  
## Bowl.Setpoint 0 0 0 0 0 0  
## Pressure.Setpoint 0 0 0 0 0 0  
## Air.Pressurer 0 0 0 0 0 0  
## Alch.Rel 0 0 0 0 0 0  
## Carb.Rel 0 0 0 0 0 0  
## Balling.Lvl 0 0 0 0 0 0  
## A 0 0 0 0 0 0  
## B 0 0 0 0 0 0  
## C 0 0 0 0 0 0  
## D 0 0 0 0 0 0  
## Fill.Pressure Hyd.Pressure1 Hyd.Pressure2 Hyd.Pressure3  
## PH 0 0 0 0  
## Carb.Volume 0 0 0 0  
## Fill.Ounces 0 0 0 0  
## PC.Volume 0 0 0 0  
## Carb.Pressure 0 0 0 0  
## Carb.Temp 0 0 0 0  
## PSC 0 0 0 0  
## PSC.Fill 0 0 0 0  
## PSC.CO2 0 0 0 0  
## Mnf.Flow 0 0 0 0  
## Carb.Pressure1 0 0 0 0  
## Fill.Pressure 1 0 0 0  
## Hyd.Pressure1 0 1 0 0  
## Hyd.Pressure2 0 0 1 0  
## Hyd.Pressure3 0 0 0 1  
## Hyd.Pressure4 0 0 0 0  
## Filler.Level 0 0 0 0  
## Filler.Speed 0 0 0 0  
## Temperature 0 0 0 0  
## Usage.cont 0 0 0 0  
## Carb.Flow 0 0 0 0  
## Density 0 0 0 0  
## MFR 0 0 0 0  
## Balling 0 0 0 0  
## Pressure.Vacuum 0 0 0 0  
## Oxygen.Filler 0 0 0 0  
## Bowl.Setpoint 0 0 0 0  
## Pressure.Setpoint 0 0 0 0  
## Air.Pressurer 0 0 0 0  
## Alch.Rel 0 0 0 0  
## Carb.Rel 0 0 0 0  
## Balling.Lvl 0 0 0 0  
## A 0 0 0 0  
## B 0 0 0 0  
## C 0 0 0 0  
## D 0 0 0 0  
## Hyd.Pressure4 Filler.Level Filler.Speed Temperature  
## PH 0 0 0 0  
## Carb.Volume 0 0 0 0  
## Fill.Ounces 0 0 0 0  
## PC.Volume 0 0 0 0  
## Carb.Pressure 0 0 0 0  
## Carb.Temp 0 0 0 0  
## PSC 0 0 0 0  
## PSC.Fill 0 0 0 0  
## PSC.CO2 0 0 0 0  
## Mnf.Flow 0 0 0 0  
## Carb.Pressure1 0 0 0 0  
## Fill.Pressure 0 0 0 0  
## Hyd.Pressure1 0 0 0 0  
## Hyd.Pressure2 0 0 0 0  
## Hyd.Pressure3 0 0 0 0  
## Hyd.Pressure4 1 0 0 0  
## Filler.Level 0 1 0 0  
## Filler.Speed 0 0 1 0  
## Temperature 0 0 0 1  
## Usage.cont 0 0 0 0  
## Carb.Flow 0 0 0 0  
## Density 0 0 0 0  
## MFR 0 0 0 0  
## Balling 0 0 0 0  
## Pressure.Vacuum 0 0 0 0  
## Oxygen.Filler 0 0 0 0  
## Bowl.Setpoint 0 0 0 0  
## Pressure.Setpoint 0 0 0 0  
## Air.Pressurer 0 0 0 0  
## Alch.Rel 0 0 0 0  
## Carb.Rel 0 0 0 0  
## Balling.Lvl 0 0 0 0  
## A 0 0 0 0  
## B 0 0 0 0  
## C 0 0 0 0  
## D 0 0 0 0  
## Usage.cont Carb.Flow Density MFR Balling Pressure.Vacuum  
## PH 0 0 0 0 0 0  
## Carb.Volume 0 0 0 0 0 0  
## Fill.Ounces 0 0 0 0 0 0  
## PC.Volume 0 0 0 0 0 0  
## Carb.Pressure 0 0 0 0 0 0  
## Carb.Temp 0 0 0 0 0 0  
## PSC 0 0 0 0 0 0  
## PSC.Fill 0 0 0 0 0 0  
## PSC.CO2 0 0 0 0 0 0  
## Mnf.Flow 0 0 0 0 0 0  
## Carb.Pressure1 0 0 0 0 0 0  
## Fill.Pressure 0 0 0 0 0 0  
## Hyd.Pressure1 0 0 0 0 0 0  
## Hyd.Pressure2 0 0 0 0 0 0  
## Hyd.Pressure3 0 0 0 0 0 0  
## Hyd.Pressure4 0 0 0 0 0 0  
## Filler.Level 0 0 0 0 0 0  
## Filler.Speed 0 0 0 0 0 0  
## Temperature 0 0 0 0 0 0  
## Usage.cont 1 0 0 0 0 0  
## Carb.Flow 0 1 0 0 0 0  
## Density 0 0 1 0 0 0  
## MFR 0 0 0 1 0 0  
## Balling 0 0 0 0 1 0  
## Pressure.Vacuum 0 0 0 0 0 1  
## Oxygen.Filler 0 0 0 0 0 0  
## Bowl.Setpoint 0 0 0 0 0 0  
## Pressure.Setpoint 0 0 0 0 0 0  
## Air.Pressurer 0 0 0 0 0 0  
## Alch.Rel 0 0 0 0 0 0  
## Carb.Rel 0 0 0 0 0 0  
## Balling.Lvl 0 0 0 0 0 0  
## A 0 0 0 0 0 0  
## B 0 0 0 0 0 0  
## C 0 0 0 0 0 0  
## D 0 0 0 0 0 0  
## Oxygen.Filler Bowl.Setpoint Pressure.Setpoint  
## PH 0 0 0  
## Carb.Volume 0 0 0  
## Fill.Ounces 0 0 0  
## PC.Volume 0 0 0  
## Carb.Pressure 0 0 0  
## Carb.Temp 0 0 0  
## PSC 0 0 0  
## PSC.Fill 0 0 0  
## PSC.CO2 0 0 0  
## Mnf.Flow 0 0 0  
## Carb.Pressure1 0 0 0  
## Fill.Pressure 0 0 0  
## Hyd.Pressure1 0 0 0  
## Hyd.Pressure2 0 0 0  
## Hyd.Pressure3 0 0 0  
## Hyd.Pressure4 0 0 0  
## Filler.Level 0 0 0  
## Filler.Speed 0 0 0  
## Temperature 0 0 0  
## Usage.cont 0 0 0  
## Carb.Flow 0 0 0  
## Density 0 0 0  
## MFR 0 0 0  
## Balling 0 0 0  
## Pressure.Vacuum 0 0 0  
## Oxygen.Filler 1 0 0  
## Bowl.Setpoint 0 1 0  
## Pressure.Setpoint 0 0 1  
## Air.Pressurer 0 0 0  
## Alch.Rel 0 0 0  
## Carb.Rel 0 0 0  
## Balling.Lvl 0 0 0  
## A 0 0 0  
## B 0 0 0  
## C 0 0 0  
## D 0 0 0  
## Air.Pressurer Alch.Rel Carb.Rel Balling.Lvl A B C D  
## PH 0 0 0 0 0 0 0 0  
## Carb.Volume 0 0 0 0 0 0 0 0  
## Fill.Ounces 0 0 0 0 0 0 0 0  
## PC.Volume 0 0 0 0 0 0 0 0  
## Carb.Pressure 0 0 0 0 0 0 0 0  
## Carb.Temp 0 0 0 0 0 0 0 0  
## PSC 0 0 0 0 0 0 0 0  
## PSC.Fill 0 0 0 0 0 0 0 0  
## PSC.CO2 0 0 0 0 0 0 0 0  
## Mnf.Flow 0 0 0 0 0 0 0 0  
## Carb.Pressure1 0 0 0 0 0 0 0 0  
## Fill.Pressure 0 0 0 0 0 0 0 0  
## Hyd.Pressure1 0 0 0 0 0 0 0 0  
## Hyd.Pressure2 0 0 0 0 0 0 0 0  
## Hyd.Pressure3 0 0 0 0 0 0 0 0  
## Hyd.Pressure4 0 0 0 0 0 0 0 0  
## Filler.Level 0 0 0 0 0 0 0 0  
## Filler.Speed 0 0 0 0 0 0 0 0  
## Temperature 0 0 0 0 0 0 0 0  
## Usage.cont 0 0 0 0 0 0 0 0  
## Carb.Flow 0 0 0 0 0 0 0 0  
## Density 0 0 0 0 0 0 0 0  
## MFR 0 0 0 0 0 0 0 0  
## Balling 0 0 0 0 0 0 0 0  
## Pressure.Vacuum 0 0 0 0 0 0 0 0  
## Oxygen.Filler 0 0 0 0 0 0 0 0  
## Bowl.Setpoint 0 0 0 0 0 0 0 0  
## Pressure.Setpoint 0 0 0 0 0 0 0 0  
## Air.Pressurer 1 0 0 0 0 0 0 0  
## Alch.Rel 0 1 0 0 0 0 0 0  
## Carb.Rel 0 0 1 0 0 0 0 0  
## Balling.Lvl 0 0 0 1 0 0 0 0  
## A 0 0 0 0 1 0 0 0  
## B 0 0 0 0 0 1 0 0  
## C 0 0 0 0 0 0 1 0  
## D 0 0 0 0 0 0 0 1  
## attr(,"term.labels")  
## [1] "Carb.Volume" "Fill.Ounces" "PC.Volume"   
## [4] "Carb.Pressure" "Carb.Temp" "PSC"   
## [7] "PSC.Fill" "PSC.CO2" "Mnf.Flow"   
## [10] "Carb.Pressure1" "Fill.Pressure" "Hyd.Pressure1"   
## [13] "Hyd.Pressure2" "Hyd.Pressure3" "Hyd.Pressure4"   
## [16] "Filler.Level" "Filler.Speed" "Temperature"   
## [19] "Usage.cont" "Carb.Flow" "Density"   
## [22] "MFR" "Balling" "Pressure.Vacuum"   
## [25] "Oxygen.Filler" "Bowl.Setpoint" "Pressure.Setpoint"  
## [28] "Air.Pressurer" "Alch.Rel" "Carb.Rel"   
## [31] "Balling.Lvl" "A" "B"   
## [34] "C" "D"   
## attr(,"order")  
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## attr(,"intercept")  
## [1] 1  
## attr(,"response")  
## [1] 1  
## attr(,".Environment")  
## <environment: R\_GlobalEnv>  
## attr(,"predvars")  
## list(PH, Carb.Volume, Fill.Ounces, PC.Volume, Carb.Pressure,   
## Carb.Temp, PSC, PSC.Fill, PSC.CO2, Mnf.Flow, Carb.Pressure1,   
## Fill.Pressure, Hyd.Pressure1, Hyd.Pressure2, Hyd.Pressure3,   
## Hyd.Pressure4, Filler.Level, Filler.Speed, Temperature, Usage.cont,   
## Carb.Flow, Density, MFR, Balling, Pressure.Vacuum, Oxygen.Filler,   
## Bowl.Setpoint, Pressure.Setpoint, Air.Pressurer, Alch.Rel,   
## Carb.Rel, Balling.Lvl, A, B, C, D)  
## attr(,"dataClasses")  
## PH Carb.Volume Fill.Ounces PC.Volume   
## "numeric" "numeric" "numeric" "numeric"   
## Carb.Pressure Carb.Temp PSC PSC.Fill   
## "numeric" "numeric" "numeric" "numeric"   
## PSC.CO2 Mnf.Flow Carb.Pressure1 Fill.Pressure   
## "numeric" "numeric" "numeric" "numeric"   
## Hyd.Pressure1 Hyd.Pressure2 Hyd.Pressure3 Hyd.Pressure4   
## "numeric" "numeric" "numeric" "numeric"   
## Filler.Level Filler.Speed Temperature Usage.cont   
## "numeric" "numeric" "numeric" "numeric"   
## Carb.Flow Density MFR Balling   
## "numeric" "numeric" "numeric" "numeric"   
## Pressure.Vacuum Oxygen.Filler Bowl.Setpoint Pressure.Setpoint   
## "numeric" "numeric" "numeric" "numeric"   
## Air.Pressurer Alch.Rel Carb.Rel Balling.Lvl   
## "numeric" "numeric" "numeric" "numeric"   
## A B C D   
## "numeric" "numeric" "numeric" "numeric"

summary(lmTune)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.52024 -0.07897 0.01068 0.08667 0.46254   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.546e+00 2.596e-03 3291.574 < 2e-16 \*\*\*  
## Carb.Volume -7.780e-03 7.559e-03 -1.029 0.303457   
## Fill.Ounces -5.711e-03 2.795e-03 -2.043 0.041110 \*   
## PC.Volume -7.824e-03 3.273e-03 -2.390 0.016908 \*   
## Carb.Pressure -3.093e-03 1.113e-02 -0.278 0.781077   
## Carb.Temp 6.587e-03 1.014e-02 0.649 0.516175   
## PSC -4.249e-03 2.795e-03 -1.520 0.128511   
## PSC.Fill -4.169e-03 2.730e-03 -1.527 0.126835   
## PSC.CO2 -5.613e-03 2.687e-03 -2.089 0.036781 \*   
## Mnf.Flow -8.457e-02 5.756e-03 -14.694 < 2e-16 \*\*\*  
## Carb.Pressure1 3.232e-02 3.284e-03 9.842 < 2e-16 \*\*\*  
## Fill.Pressure 7.646e-03 3.835e-03 1.994 0.046255 \*   
## Hyd.Pressure1 1.583e-04 4.533e-03 0.035 0.972135   
## Hyd.Pressure2 -2.517e-02 8.696e-03 -2.894 0.003835 \*\*   
## Hyd.Pressure3 5.858e-02 9.247e-03 6.335 2.80e-10 \*\*\*  
## Hyd.Pressure4 -8.456e-05 4.342e-03 -0.019 0.984464   
## Filler.Level -1.670e-02 8.182e-03 -2.040 0.041417 \*   
## Filler.Speed 2.217e-02 1.285e-02 1.726 0.084487 .   
## Temperature -1.922e-02 3.168e-03 -6.066 1.51e-09 \*\*\*  
## Usage.cont -2.288e-02 3.433e-03 -6.664 3.26e-11 \*\*\*  
## Carb.Flow 1.186e-02 3.883e-03 3.054 0.002283 \*\*   
## Density -1.716e-02 8.120e-03 -2.113 0.034700 \*   
## MFR -1.906e-02 1.244e-02 -1.531 0.125777   
## Balling -8.535e-02 1.341e-02 -6.363 2.35e-10 \*\*\*  
## Pressure.Vacuum -1.382e-02 4.096e-03 -3.373 0.000756 \*\*\*  
## Oxygen.Filler -1.319e-02 3.727e-03 -3.539 0.000408 \*\*\*  
## Bowl.Setpoint 5.075e-02 8.479e-03 5.985 2.47e-09 \*\*\*  
## Pressure.Setpoint -1.652e-02 3.940e-03 -4.194 2.84e-05 \*\*\*  
## Air.Pressurer -3.630e-03 2.854e-03 -1.272 0.203565   
## Alch.Rel 2.379e-02 1.072e-02 2.218 0.026614 \*   
## Carb.Rel 4.569e-03 5.861e-03 0.780 0.435690   
## Balling.Lvl 9.416e-02 1.597e-02 5.896 4.23e-09 \*\*\*  
## A -6.084e-03 7.848e-03 -0.775 0.438299   
## B 3.373e-02 6.867e-03 4.912 9.59e-07 \*\*\*  
## C -2.376e-02 4.808e-03 -4.942 8.23e-07 \*\*\*  
## D 1.637e-02 1.192e-02 1.374 0.169541   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1315 on 2531 degrees of freedom  
## Multiple R-squared: 0.4266, Adjusted R-squared: 0.4186   
## F-statistic: 53.79 on 35 and 2531 DF, p-value: < 2.2e-16

## Partial Least Squares

ctrl <- trainControl(method = "cv")  
  
plsTune <- train(PH ~ ., data=imputed, method='pls', preProcess=c('BoxCox', 'center', 'scale'), tuneLength=5, trControl=ctrl)  
summary(plsTune)

## Data: X dimension: 2567 35   
## Y dimension: 2567 1  
## Fit method: oscorespls  
## Number of components considered: 5  
## TRAINING: % variance explained  
## 1 comps 2 comps 3 comps 4 comps 5 comps  
## X 15.34 24.75 39.89 47.86 51.31  
## .outcome 25.30 33.93 36.49 38.06 39.79

## PCA

## Ridge-Regression

ridgeGrid <- data.frame(.lambda=seq(0, 0.1, length=15))  
ridgeTune <- train(PH ~ ., data=imputed, method='ridge', preProcess=c('BoxCox', 'center', 'scale'), tuneGrid=ridgeGrid, trControl=ctrl)  
ridgeTune

## Ridge Regression   
##   
## 2567 samples  
## 35 predictor  
##   
## Pre-processing: Box-Cox transformation (23), centered (35), scaled (35)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 2310, 2309, 2312, 2309, 2311, 2310, ...   
## Resampling results across tuning parameters:  
##   
## lambda RMSE Rsquared MAE   
## 0.000000000 0.1324771 0.4122643 0.1031271  
## 0.007142857 0.1324202 0.4128107 0.1031975  
## 0.014285714 0.1325113 0.4121021 0.1033144  
## 0.021428571 0.1326401 0.4110432 0.1034471  
## 0.028571429 0.1327778 0.4098847 0.1035768  
## 0.035714286 0.1329152 0.4087129 0.1037053  
## 0.042857143 0.1330490 0.4075609 0.1038360  
## 0.050000000 0.1331783 0.4064415 0.1039658  
## 0.057142857 0.1333029 0.4053592 0.1040938  
## 0.064285714 0.1334231 0.4043147 0.1042112  
## 0.071428571 0.1335391 0.4033069 0.1043216  
## 0.078571429 0.1336513 0.4023344 0.1044251  
## 0.085714286 0.1337601 0.4013950 0.1045249  
## 0.092857143 0.1338658 0.4004869 0.1046225  
## 0.100000000 0.1339687 0.3996081 0.1047170  
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final value used for the model was lambda = 0.007142857.

# MARS

library(parallel)  
library(doParallel)

## Warning: package 'doParallel' was built under R version 3.4.4

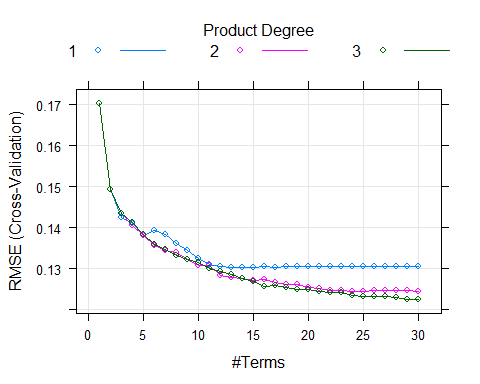
cluster <- makeCluster(detectCores() - 1)   
registerDoParallel(cluster)  
  
  
trainX <- train\_set %>% select(-PH)  
trainY <- train\_set %>% select(PH)  
  
  
  
  
indx <- createFolds(trainY$PH, returnTrain = TRUE)  
ctrl <- trainControl(method = "cv", index = indx)  
  
mars1 <- train(x = trainX, y = trainY$PH,  
 method = "earth",  
 tuneGrid = expand.grid(degree = 1:3, nprune = 1:30),  
 trControl = ctrl)

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =  
## trainInfo, : There were missing values in resampled performance measures.

#turn off parallel processing  
stopCluster(cluster)  
#resume use of the sequential backend  
registerDoSEQ()  
  
mars1

## Multivariate Adaptive Regression Spline   
##   
## 1796 samples  
## 35 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 1618, 1616, 1618, 1617, 1615, 1616, ...   
## Resampling results across tuning parameters:  
##   
## degree nprune RMSE Rsquared MAE   
## 1 1 0.1703757 NaN 0.13756966  
## 1 2 0.1494344 0.2358863 0.11831638  
## 1 3 0.1424579 0.3061415 0.11212236  
## 1 4 0.1411009 0.3196783 0.11145339  
## 1 5 0.1382133 0.3478652 0.10886663  
## 1 6 0.1392922 0.3390082 0.10891503  
## 1 7 0.1383672 0.3484151 0.10697192  
## 1 8 0.1362314 0.3675897 0.10525779  
## 1 9 0.1343316 0.3843240 0.10367747  
## 1 10 0.1325601 0.3999279 0.10240314  
## 1 11 0.1311164 0.4125976 0.10126990  
## 1 12 0.1304793 0.4184978 0.10076868  
## 1 13 0.1303380 0.4200705 0.10066461  
## 1 14 0.1303216 0.4200805 0.10056976  
## 1 15 0.1304014 0.4195646 0.10068380  
## 1 16 0.1304278 0.4193578 0.10074772  
## 1 17 0.1303412 0.4201925 0.10060958  
## 1 18 0.1304505 0.4192190 0.10054860  
## 1 19 0.1304931 0.4188591 0.10060925  
## 1 20 0.1304931 0.4188591 0.10060925  
## 1 21 0.1304931 0.4188591 0.10060925  
## 1 22 0.1304931 0.4188591 0.10060925  
## 1 23 0.1304931 0.4188591 0.10060925  
## 1 24 0.1304931 0.4188591 0.10060925  
## 1 25 0.1304931 0.4188591 0.10060925  
## 1 26 0.1304931 0.4188591 0.10060925  
## 1 27 0.1304931 0.4188591 0.10060925  
## 1 28 0.1304931 0.4188591 0.10060925  
## 1 29 0.1304931 0.4188591 0.10060925  
## 1 30 0.1304931 0.4188591 0.10060925  
## 2 1 0.1703757 NaN 0.13756966  
## 2 2 0.1494344 0.2358863 0.11831638  
## 2 3 0.1435739 0.2933587 0.11336333  
## 2 4 0.1404847 0.3241435 0.11047187  
## 2 5 0.1381057 0.3462844 0.10776950  
## 2 6 0.1357556 0.3677571 0.10551505  
## 2 7 0.1344657 0.3796265 0.10445919  
## 2 8 0.1338555 0.3867209 0.10317081  
## 2 9 0.1321826 0.4002895 0.10139362  
## 2 10 0.1306787 0.4137738 0.09996194  
## 2 11 0.1306991 0.4142029 0.09981815  
## 2 12 0.1284230 0.4348930 0.09851180  
## 2 13 0.1277769 0.4406549 0.09771751  
## 2 14 0.1276478 0.4429696 0.09736621  
## 2 15 0.1272025 0.4483444 0.09721965  
## 2 16 0.1273104 0.4475153 0.09701921  
## 2 17 0.1266020 0.4537457 0.09630248  
## 2 18 0.1262054 0.4580920 0.09581298  
## 2 19 0.1260814 0.4589699 0.09558146  
## 2 20 0.1253128 0.4663889 0.09493973  
## 2 21 0.1251826 0.4682708 0.09453846  
## 2 22 0.1245626 0.4735799 0.09397485  
## 2 23 0.1245672 0.4735820 0.09371483  
## 2 24 0.1244334 0.4747824 0.09351555  
## 2 25 0.1243244 0.4762748 0.09321425  
## 2 26 0.1246585 0.4736861 0.09334017  
## 2 27 0.1246879 0.4732408 0.09340137  
## 2 28 0.1246513 0.4736838 0.09328504  
## 2 29 0.1246423 0.4742276 0.09322185  
## 2 30 0.1244381 0.4760659 0.09280931  
## 3 1 0.1703757 NaN 0.13756966  
## 3 2 0.1494344 0.2358863 0.11831638  
## 3 3 0.1435589 0.2932293 0.11382716  
## 3 4 0.1413107 0.3177965 0.11133241  
## 3 5 0.1382582 0.3469526 0.10790804  
## 3 6 0.1360242 0.3659584 0.10530918  
## 3 7 0.1347189 0.3781417 0.10368184  
## 3 8 0.1331526 0.3922558 0.10225783  
## 3 9 0.1321766 0.4028980 0.10155269  
## 3 10 0.1315789 0.4085377 0.10126362  
## 3 11 0.1299830 0.4229161 0.09972650  
## 3 12 0.1293741 0.4281925 0.09925418  
## 3 13 0.1286980 0.4344938 0.09860808  
## 3 14 0.1277248 0.4430167 0.09776722  
## 3 15 0.1267684 0.4522562 0.09651531  
## 3 16 0.1256195 0.4611607 0.09578949  
## 3 17 0.1259232 0.4589580 0.09606622  
## 3 18 0.1253764 0.4641078 0.09557178  
## 3 19 0.1248459 0.4682651 0.09522896  
## 3 20 0.1248823 0.4711806 0.09517467  
## 3 21 0.1243646 0.4753417 0.09432671  
## 3 22 0.1242037 0.4774226 0.09443660  
## 3 23 0.1240784 0.4784401 0.09431029  
## 3 24 0.1235572 0.4832025 0.09368457  
## 3 25 0.1233248 0.4855206 0.09345023  
## 3 26 0.1231492 0.4873215 0.09306567  
## 3 27 0.1231270 0.4879626 0.09281827  
## 3 28 0.1229792 0.4894506 0.09248285  
## 3 29 0.1225261 0.4930073 0.09221423  
## 3 30 0.1224687 0.4938186 0.09186196  
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final values used for the model were nprune = 30 and degree = 3.

plot(mars1)

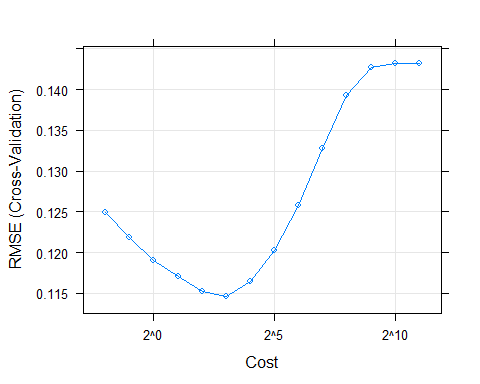


**Support Vector Machines**

cluster <- makeCluster(detectCores() - 1)   
registerDoParallel(cluster)  
  
  
  
  
set.seed(123)  
svmPTune <- train(x = trainX, y = trainY$PH,  
 method = "svmRadial",  
 preProc = c("center", "scale"),  
 tuneLength = 14,  
 trControl = trainControl(method = 'cv'))  
svmPTune

## Support Vector Machines with Radial Basis Function Kernel   
##   
## 1796 samples  
## 35 predictor  
##   
## Pre-processing: centered (35), scaled (35)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 1615, 1617, 1618, 1617, 1616, 1615, ...   
## Resampling results across tuning parameters:  
##   
## C RMSE Rsquared MAE   
## 0.25 0.1249612 0.4720487 0.09395734  
## 0.50 0.1218068 0.4961224 0.09037443  
## 1.00 0.1190658 0.5172826 0.08777865  
## 2.00 0.1170579 0.5324279 0.08613360  
## 4.00 0.1152165 0.5472515 0.08467424  
## 8.00 0.1145571 0.5534973 0.08440365  
## 16.00 0.1165106 0.5432894 0.08634638  
## 32.00 0.1202236 0.5249829 0.08943957  
## 64.00 0.1258161 0.4986690 0.09372242  
## 128.00 0.1327825 0.4660980 0.09866151  
## 256.00 0.1392609 0.4401940 0.10390911  
## 512.00 0.1426860 0.4270693 0.10638408  
## 1024.00 0.1432689 0.4255470 0.10667600  
## 2048.00 0.1432689 0.4255470 0.10667600  
##   
## Tuning parameter 'sigma' was held constant at a value of 0.01861347  
## RMSE was used to select the optimal model using the smallest value.  
## The final values used for the model were sigma = 0.01861347 and C = 8.

plot(svmPTune,   
 scales = list(x = list(log = 2),   
 between = list(x = .5, y = 1)))



#turn off parallel processing  
stopCluster(cluster)  
#resume use of the sequential backend  
registerDoSEQ()  
  
  
vip <- varImp(svmPTune)  
df2 <- data.frame(vip$importance, stringsAsFactors = FALSE)  
vip

## loess r-squared variable importance  
##   
## only 20 most important variables shown (out of 35)  
##   
## Overall  
## Mnf.Flow 100.000  
## Usage.cont 66.766  
## Bowl.Setpoint 54.498  
## Filler.Level 46.784  
## Pressure.Setpoint 43.739  
## Carb.Flow 40.078  
## C 33.508  
## Hyd.Pressure3 25.827  
## Pressure.Vacuum 23.827  
## Hyd.Pressure2 19.220  
## Fill.Pressure 15.895  
## D 15.507  
## MFR 13.527  
## Carb.Rel 10.457  
## Alch.Rel 9.443  
## Oxygen.Filler 9.187  
## Hyd.Pressure4 8.674  
## Temperature 8.193  
## B 7.380  
## A 6.338

**Neural Network**

cluster <- makeCluster(detectCores() - 1)   
registerDoParallel(cluster)  
  
  
nnetGrid <- expand.grid(decay = c(0, 0.01, .1),   
 size = c(1:10),   
 bag = FALSE)  
  
  
set.seed(100)  
nnetTune <- train(x = trainX, y = trainY$PH,  
 method = "avNNet",  
 tuneGrid = nnetGrid,  
 trControl = ctrl,  
 preProc = c("center", "scale"),  
 linout = TRUE,  
 trace = FALSE,  
 maxit = 100,   
 allowParallel = TRUE)  
  
#turn off parallel processing  
stopCluster(cluster)  
#resume use of the sequential backend  
registerDoSEQ()  
  
nnetTune

## Model Averaged Neural Network   
##   
## 1796 samples  
## 35 predictor  
##   
## Pre-processing: centered (35), scaled (35)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 1618, 1616, 1618, 1617, 1615, 1616, ...   
## Resampling results across tuning parameters:  
##   
## decay size RMSE Rsquared MAE   
## 0.00 1 0.1464844 0.2938134 0.11242870  
## 0.00 2 0.1574282 0.3168432 0.10737265  
## 0.00 3 0.1969608 0.2634764 0.10881677  
## 0.00 4 0.1214664 0.4930871 0.09212088  
## 0.00 5 0.1160226 0.5404085 0.08733511  
## 0.00 6 0.1192612 0.5169675 0.08807316  
## 0.00 7 0.1144823 0.5494770 0.08500922  
## 0.00 8 0.1125533 0.5643026 0.08442527  
## 0.00 9 0.1135701 0.5581055 0.08456782  
## 0.00 10 0.1123640 0.5671867 0.08417421  
## 0.01 1 0.1358299 0.3743973 0.10492026  
## 0.01 2 0.1506672 0.3451867 0.10264936  
## 0.01 3 0.1504437 0.3368097 0.10332948  
## 0.01 4 0.1368796 0.4489432 0.09471340  
## 0.01 5 0.1183133 0.5212244 0.08795053  
## 0.01 6 0.1170792 0.5295085 0.08734617  
## 0.01 7 0.1139766 0.5530562 0.08550858  
## 0.01 8 0.1151732 0.5458542 0.08595792  
## 0.01 9 0.1131192 0.5620896 0.08433806  
## 0.01 10 0.1131401 0.5611361 0.08449841  
## 0.10 1 0.1315928 0.4087535 0.10233342  
## 0.10 2 0.1309076 0.4187970 0.10085546  
## 0.10 3 0.1398544 0.3560081 0.10444335  
## 0.10 4 0.1323801 0.4248831 0.09791777  
## 0.10 5 0.1205492 0.5029592 0.09317670  
## 0.10 6 0.1177932 0.5239863 0.08999897  
## 0.10 7 0.1172406 0.5286465 0.08902041  
## 0.10 8 0.1166821 0.5329235 0.08785266  
## 0.10 9 0.1137912 0.5558778 0.08586876  
## 0.10 10 0.1162426 0.5377168 0.08688658  
##   
## Tuning parameter 'bag' was held constant at a value of FALSE  
## RMSE was used to select the optimal model using the smallest value.  
## The final values used for the model were size = 10, decay = 0 and bag  
## = FALSE.

Now lets compare against our test set using the postResample function:

validX <- validation\_set %>% select(-PH)  
validY <- validation\_set %>% select(PH)  
  
  
marsP <- predict(mars1, newdata = validX)  
svmP <- predict(svmPTune, newdata = validX)  
nnetP <- predict(nnetTune, newdata = validX)  
  
  
#MARS  
postResample(pred = marsP, obs = validY$PH)[1:2]

## RMSE Rsquared   
## 0.1238133 0.5112345

#SVM  
postResample(pred = svmP, obs = validY$PH)[1:2]

## RMSE Rsquared   
## 0.1198595 0.5485638

#NNET  
postResample(pred = nnetP, obs = validY$PH)

## RMSE Rsquared MAE   
## 0.11609520 0.57095496 0.08317553

<https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/>