DATA 624: PREDICTIVE ANALYTICS: Project 2

Melissa Bowman, Frederick Jones, Shoshana Farber, Gabriel Campos

Last edited April 23, 2024

Library

```
library(Amelia)
library(car)
library(caret)
library(corrplot)
library(Cubist)
library(DataExplorer)
library(dplyr)
library(e1071)
library(earth)
library(forcats)
library(forecast)
library(fpp3)
library(gbm)
library(ggplot2)
library(kableExtra)
library(MASS)
library(mice)
library(mlbench)
library(party)
library(randomForest)
library(RANN)
library(RColorBrewer)
library(readxl)
library(rpart)
library(rpart.plot)
library(summarytools)
library(tidyr)
library(VIM)
```

Description

Project #2 (Team) Assignment

This is role playing. I am your new boss. I am in charge of production at ABC Beverage and you are a team of data scientists reporting to me. My leadership has told me that new regulations are requiring us to understand our manufacturing process, the predictive factors and be able to report to them our predictive model of PH.

Please use the historical data set I am providing. Build and report the factors in BOTH a technical and non-technical report. I like to use Word and Excel. Please provide your non-technical report in a business friendly readable document and your predictions in an Excel readable format. The technical report should show clearly the models you tested and how you selected your final approach. Please submit both Rpubs links and .rmd files or other readable formats for technical and non-technical reports. Also submit the excel file showing the prediction of your models for pH.

Data Import

```
train_df <- readxl::read_xlsx('Data/StudentData.xlsx')
test_df <- readxl::read_xlsx('Data/StudentEvaluation.xlsx')
```

StudentData.xlsx is our Training data set. StudentEvaluation.xlsx is our Test data set.

Exporatory Data Analysis

Data Exploration

Initial Exploration

```
glimpse(train_df)
```

```
## Rows: 2,571
## Columns: 33
## $ `Brand Code`        <chr> "B", "A", "B", "A", "A", "A", "A", "B", "B", "B", ~
## $ `Carb Volume`       <dbl> 5.340000, 5.426667, 5.286667, 5.440000, 5.486667, ~
## $ `Fill Ounces`       <dbl> 23.96667, 24.00667, 24.06000, 24.00667, 24.31333, ~
## $ `PC Volume`         <dbl> 0.2633333, 0.2386667, 0.2633333, 0.2933333, 0.1113~
## $ `Carb Pressure`     <dbl> 68.2, 68.4, 70.8, 63.0, 67.2, 66.6, 64.2, 67.6, 64~
## $ `Carb Temp`         <dbl> 141.2, 139.6, 144.8, 132.6, 136.8, 138.4, 136.8, 1~
## $ PSC                 <dbl> 0.104, 0.124, 0.090, NA, 0.026, 0.090, 0.128, 0.15~
## $ `PSC Fill`          <dbl> 0.26, 0.22, 0.34, 0.42, 0.16, 0.24, 0.40, 0.34, 0.~
## $ `PSC CO2`           <dbl> 0.04, 0.04, 0.16, 0.04, 0.12, 0.04, 0.04, 0.04, 0.~
## $ `Mnf Flow`          <dbl> -100, -100, -100, -100, -100, -100, -100, -100, -1~
## $ `Carb Pressure1`    <dbl> 118.8, 121.6, 120.2, 115.2, 118.4, 119.6, 122.2, 1~
## $ `Fill Pressure`     <dbl> 46.0, 46.0, 46.0, 46.4, 45.8, 45.6, 51.8, 46.8, 46~
## $ `Hyd Pressure1`     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ `Hyd Pressure2`     <dbl> NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ `Hyd Pressure3`     <dbl> NA, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ `Hyd Pressure4`     <dbl> 118, 106, 82, 92, 92, 116, 124, 132, 90, 108, 94, ~
## $ `Filler Level`      <dbl> 121.2, 118.6, 120.0, 117.8, 118.6, 120.2, 123.4, 1~
## $ `Filler Speed`      <dbl> 4002, 3986, 4020, 4012, 4010, 4014, NA, 1004, 4014~
## $ Temperature         <dbl> 66.0, 67.6, 67.0, 65.6, 65.6, 66.2, 65.8, 65.2, 65~
## $ `Usage cont`        <dbl> 16.18, 19.90, 17.76, 17.42, 17.68, 23.82, 20.74, 1~
## $ `Carb Flow`         <dbl> 2932, 3144, 2914, 3062, 3054, 2948, 30, 684, 2902,~
## $ Density             <dbl> 0.88, 0.92, 1.58, 1.54, 1.54, 1.52, 0.84, 0.84, 0.~
## $ MFR                 <dbl> 725.0, 726.8, 735.0, 730.6, 722.8, 738.8, NA, NA, ~
## $ Balling             <dbl> 1.398, 1.498, 3.142, 3.042, 3.042, 2.992, 1.298, 1~
## $ `Pressure Vacuum`   <dbl> -4.0, -4.0, -3.8, -4.4, -4.4, -4.4, -4.4, -4.4, -4~
## $ PH                  <dbl> 8.36, 8.26, 8.94, 8.24, 8.26, 8.32, 8.40, 8.38, 8.~
## $ `Oxygen Filler`     <dbl> 0.022, 0.026, 0.024, 0.030, 0.030, 0.024, 0.066, 0~
## $ `Bowl Setpoint`     <dbl> 120, 120, 120, 120, 120, 120, 120, 120, 120, 120, ~
## $ `Pressure Setpoint` <dbl> 46.4, 46.8, 46.6, 46.0, 46.0, 46.0, 46.0, 46.0, 46~
## $ `Air Pressurer`     <dbl> 142.6, 143.0, 142.0, 146.2, 146.2, 146.6, 146.2, 1~
## $ `Alch Rel`          <dbl> 6.58, 6.56, 7.66, 7.14, 7.14, 7.16, 6.54, 6.52, 6.~
## $ `Carb Rel`          <dbl> 5.32, 5.30, 5.84, 5.42, 5.44, 5.44, 5.38, 5.34, 5.~
## $ `Balling Lvl`       <dbl> 1.48, 1.56, 3.28, 3.04, 3.04, 3.02, 1.44, 1.44, 1.~
```

```r
str(train_df)
```

```
## tibble [2,571 x 33] (S3: tbl_df/tbl/data.frame)
##  $ Brand Code       : chr [1:2571] "B" "A" "B" "A" ...
##  $ Carb Volume      : num [1:2571] 5.34 5.43 5.29 5.44 5.49 ...
##  $ Fill Ounces      : num [1:2571] 24 24 24.1 24 24.3 ...
##  $ PC Volume        : num [1:2571] 0.263 0.239 0.263 0.293 0.111 ...
##  $ Carb Pressure    : num [1:2571] 68.2 68.4 70.8 63 67.2 66.6 64.2 67.6 64.2 72 ...
##  $ Carb Temp        : num [1:2571] 141 140 145 133 137 ...
##  $ PSC              : num [1:2571] 0.104 0.124 0.09 NA 0.026 0.09 0.128 0.154 0.132 0.014 ...
##  $ PSC Fill         : num [1:2571] 0.26 0.22 0.34 0.42 0.16 ...
##  $ PSC CO2          : num [1:2571] 0.04 0.04 0.16 0.04 0.12 ...
##  $ Mnf Flow         : num [1:2571] -100 -100 -100 -100 -100 -100 -100 -100 -100 -100 ...
##  $ Carb Pressure1   : num [1:2571] 119 122 120 115 118 ...
##  $ Fill Pressure    : num [1:2571] 46 46 46 46.4 45.8 45.6 51.8 46.8 46 45.2 ...
##  $ Hyd Pressure1    : num [1:2571] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Hyd Pressure2    : num [1:2571] NA NA NA 0 0 0 0 0 0 0 ...
##  $ Hyd Pressure3    : num [1:2571] NA NA NA 0 0 0 0 0 0 0 ...
##  $ Hyd Pressure4    : num [1:2571] 118 106 82 92 92 116 124 132 90 108 ...
##  $ Filler Level     : num [1:2571] 121 119 120 118 119 ...
##  $ Filler Speed     : num [1:2571] 4002 3986 4020 4012 4010 ...
##  $ Temperature      : num [1:2571] 66 67.6 67 65.6 65.6 66.2 65.8 65.2 65.4 66.6 ...
##  $ Usage cont       : num [1:2571] 16.2 19.9 17.8 17.4 17.7 ...
##  $ Carb Flow        : num [1:2571] 2932 3144 2914 3062 3054 ...
##  $ Density          : num [1:2571] 0.88 0.92 1.58 1.54 1.54 1.52 0.84 0.84 0.9 0.9 ...
##  $ MFR              : num [1:2571] 725 727 735 731 723 ...
##  $ Balling          : num [1:2571] 1.4 1.5 3.14 3.04 3.04 ...
##  $ Pressure Vacuum  : num [1:2571] -4 -4 -3.8 -4.4 -4.4 -4.4 -4.4 -4.4 -4.4 -4.4 ...
##  $ PH               : num [1:2571] 8.36 8.26 8.94 8.24 8.26 8.32 8.4 8.38 8.38 8.5 ...
##  $ Oxygen Filler    : num [1:2571] 0.022 0.026 0.024 0.03 0.03 0.024 0.066 0.046 0.064 0.022 ...
##  $ Bowl Setpoint    : num [1:2571] 120 120 120 120 120 120 120 120 120 120 ...
##  $ Pressure Setpoint: num [1:2571] 46.4 46.8 46.6 46 46 46 46 46 46 46 ...
##  $ Air Pressurer    : num [1:2571] 143 143 142 146 146 ...
##  $ Alch Rel         : num [1:2571] 6.58 6.56 7.66 7.14 7.14 7.16 6.54 6.52 6.52 6.54 ...
##  $ Carb Rel         : num [1:2571] 5.32 5.3 5.84 5.42 5.44 5.44 5.38 5.34 5.34 5.34 ...
##  $ Balling Lvl      : num [1:2571] 1.48 1.56 3.28 3.04 3.04 3.02 1.44 1.44 1.44 1.38 ...
```

```r
summary(train_df)
```

```
##   Brand Code         Carb Volume     Fill Ounces      PC Volume
##  Length:2571        Min.   :5.040   Min.   :23.63   Min.   :0.07933
##  Class :character   1st Qu.:5.293   1st Qu.:23.92   1st Qu.:0.23917
##  Mode  :character   Median :5.347   Median :23.97   Median :0.27133
##                        Mean   :5.370   Mean   :23.97   Mean   :0.27712  
##                        3rd Qu.:5.453   3rd Qu.:24.03   3rd Qu.:0.31200  
##                        Max.   :5.700   Max.   :24.32   Max.   :0.47800  
##                        NA's   :10      NA's   :38      NA's   :39       
##  Carb Pressure     Carb Temp          PSC             PSC Fill     
##  Min.   :57.00   Min.   :128.6   Min.   :0.00200   Min.   :0.0000  
##  1st Qu.:65.60   1st Qu.:138.4   1st Qu.:0.04800   1st Qu.:0.1000  
##  Median :68.20   Median :140.8   Median :0.07600   Median :0.1800  
##  Mean   :68.19   Mean   :141.1   Mean   :0.08457   Mean   :0.1954  
##  3rd Qu.:70.60   3rd Qu.:143.8   3rd Qu.:0.11200   3rd Qu.:0.2600  
##  Max.   :79.40   Max.   :154.0   Max.   :0.27000   Max.   :0.6200  
##  NA's   :27      NA's   :26      NA's   :33        NA's   :23      
##     PSC CO2           Mnf Flow       Carb Pressure1  Fill Pressure  
##  Min.   :0.00000   Min.   :-100.20   Min.   :105.6   Min.   :34.60  
##  1st Qu.:0.02000   1st Qu.:-100.00   1st Qu.:119.0   1st Qu.:46.00  
##  Median :0.04000   Median :  65.20   Median :123.2   Median :46.40  
##  Mean   :0.05641   Mean   :  24.57   Mean   :122.6   Mean   :47.92  
##  3rd Qu.:0.08000   3rd Qu.: 140.80   3rd Qu.:125.4   3rd Qu.:50.00  
##  Max.   :0.24000   Max.   : 229.40   Max.   :140.2   Max.   :60.40  
##  NA's   :39        NA's   :2         NA's   :32      NA's   :22     
##  Hyd Pressure1   Hyd Pressure2   Hyd Pressure3   Hyd Pressure4   
##  Min.   :-0.80   Min.   : 0.00   Min.   :-1.20   Min.   : 52.00  
##  1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 86.00  
##  Median :11.40   Median :28.60   Median :27.60   Median : 96.00  
##  Mean   :12.44   Mean   :20.96   Mean   :20.46   Mean   : 96.29  
##  3rd Qu.:20.20   3rd Qu.:34.60   3rd Qu.:33.40   3rd Qu.:102.00  
##  Max.   :58.00   Max.   :59.40   Max.   :50.00   Max.   :142.00  
##  NA's   :11      NA's   :15      NA's   :15      NA's   :30      
##   Filler Level    Filler Speed   Temperature      Usage cont      Carb Flow   
##  Min.   : 55.8   Min.   : 998   Min.   :63.60   Min.   :12.08   Min.   :  26  
##  1st Qu.: 98.3   1st Qu.:3888   1st Qu.:65.20   1st Qu.:18.36   1st Qu.:1144  
##  Median :118.4   Median :3982   Median :65.60   Median :21.79   Median :3028  
##  Mean   :109.3   Mean   :3687   Mean   :65.97   Mean   :20.99   Mean   :2468  
##  3rd Qu.:120.0   3rd Qu.:3998   3rd Qu.:66.40   3rd Qu.:23.75   3rd Qu.:3186  
##  Max.   :161.2   Max.   :4030   Max.   :76.20   Max.   :25.90   Max.   :5104  
##  NA's   :20      NA's   :57     NA's   :14      NA's   :5       NA's   :2     
##     Density           MFR           Balling       Pressure Vacuum 
##  Min.   :0.240   Min.   : 31.4   Min.   :-0.170   Min.   :-6.600  
##  1st Qu.:0.900   1st Qu.:706.3   1st Qu.: 1.496   1st Qu.:-5.600  
##  Median :0.980   Median :724.0   Median : 1.648   Median :-5.400  
##  Mean   :1.174   Mean   :704.0   Mean   : 2.198   Mean   :-5.216  
##  3rd Qu.:1.620   3rd Qu.:731.0   3rd Qu.: 3.292   3rd Qu.:-5.000  
##  Max.   :1.920   Max.   :868.6   Max.   : 4.012   Max.   :-3.600  
##  NA's   :1       NA's   :212     NA's   :1                        
##        PH        Oxygen Filler     Bowl Setpoint   Pressure Setpoint
##  Min.   :7.880   Min.   :0.00240   Min.   : 70.0   Min.   :44.00    
##  1st Qu.:8.440   1st Qu.:0.02200   1st Qu.:100.0   1st Qu.:46.00    
##  Median :8.540   Median :0.03340   Median :120.0   Median :46.00    
##  Mean   :8.546   Mean   :0.04684   Mean   :109.3   Mean   :47.62    
##  3rd Qu.:8.680   3rd Qu.:0.06000   3rd Qu.:120.0   3rd Qu.:50.00    
##  Max.   :9.360   Max.   :0.40000   Max.   :140.0   Max.   :52.00    
##  NA's   :4       NA's   :12        NA's   :2       NA's   :12       
##  Air Pressurer      Alch Rel        Carb Rel      Balling Lvl  
##  Min.   :140.8   Min.   :5.280   Min.   :4.960   Min.   :0.00  
##  1st Qu.:142.2   1st Qu.:6.540   1st Qu.:5.340   1st Qu.:1.38  
##  Median :142.6   Median :6.560   Median :5.400   Median :1.48  
##  Mean   :142.8   Mean   :6.897   Mean   :5.437   Mean   :2.05  
##  3rd Qu.:143.0   3rd Qu.:7.240   3rd Qu.:5.540   3rd Qu.:3.14  
##  Max.   :148.2   Max.   :8.620   Max.   :6.060   Max.   :3.66  
##                  NA's   :9       NA's   :10      NA's   :1
```

```
glimpse(test_df)
```

```
## Rows: 267
## Columns: 33
## $ `Brand Code`        <chr> "D", "A", "B", "B", "B", "B", "A", "B", "A", "D", ~
## $ `Carb Volume`       <dbl> 5.480000, 5.393333, 5.293333, 5.266667, 5.406667, ~
## $ `Fill Ounces`       <dbl> 24.03333, 23.95333, 23.92000, 23.94000, 24.20000, ~
## $ `PC Volume`         <dbl> 0.2700000, 0.2266667, 0.3033333, 0.1860000, 0.1600~
## $ `Carb Pressure`     <dbl> 65.4, 63.2, 66.4, 64.8, 69.4, 73.4, 65.2, 67.4, 66~
## $ `Carb Temp`         <dbl> 134.6, 135.0, 140.4, 139.0, 142.2, 147.2, 134.6, 1~
## $ PSC                 <dbl> 0.236, 0.042, 0.068, 0.004, 0.040, 0.078, 0.088, 0~
## $ `PSC Fill`          <dbl> 0.40, 0.22, 0.10, 0.20, 0.30, 0.22, 0.14, 0.10, 0.~
## $ `PSC CO2`           <dbl> 0.04, 0.08, 0.02, 0.02, 0.06, NA, 0.00, 0.04, 0.04~
## $ `Mnf Flow`          <dbl> -100, -100, -100, -100, -100, -100, -100, -100, -1~
## $ `Carb Pressure1`    <dbl> 116.6, 118.8, 120.2, 124.8, 115.0, 118.6, 117.6, 1~
## $ `Fill Pressure`     <dbl> 46.0, 46.2, 45.8, 40.0, 51.4, 46.4, 46.2, 40.0, 43~
## $ `Hyd Pressure1`     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ `Hyd Pressure2`     <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ `Hyd Pressure3`     <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ `Hyd Pressure4`     <dbl> 96, 112, 98, 132, 94, 94, 108, 108, 110, 106, 98, ~
## $ `Filler Level`      <dbl> 129.4, 120.0, 119.4, 120.2, 116.0, 120.4, 119.6, 1~
## $ `Filler Speed`      <dbl> 3986, 4012, 4010, NA, 4018, 4010, 4010, NA, 4010, ~
## $ Temperature         <dbl> 66.0, 65.6, 65.6, 74.4, 66.4, 66.6, 66.8, NA, 65.8~
## $ `Usage cont`        <dbl> 21.66, 17.60, 24.18, 18.12, 21.32, 18.00, 17.68, 1~
## $ `Carb Flow`         <dbl> 2950, 2916, 3056, 28, 3214, 3064, 3042, 1972, 2502~
## $ Density             <dbl> 0.88, 1.50, 0.90, 0.74, 0.88, 0.84, 1.48, 1.60, 1.~
## $ MFR                 <dbl> 727.6, 735.8, 734.8, NA, 752.0, 732.0, 729.8, NA, ~
## $ Balling             <dbl> 1.398, 2.942, 1.448, 1.056, 1.398, 1.298, 2.894, 3~
## $ `Pressure Vacuum`   <dbl> -3.8, -4.4, -4.2, -4.0, -4.0, -3.8, -4.2, -4.4, -4~
## $ PH                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ `Oxygen Filler`     <dbl> 0.022, 0.030, 0.046, NA, 0.082, 0.064, 0.042, 0.09~
## $ `Bowl Setpoint`     <dbl> 130, 120, 120, 120, 120, 120, 120, 120, 120, 120, ~
## $ `Pressure Setpoint` <dbl> 45.2, 46.0, 46.0, 46.0, 50.0, 46.0, 46.0, 46.0, 46~
## $ `Air Pressurer`     <dbl> 142.6, 147.2, 146.6, 146.4, 145.8, 146.0, 145.0, 1~
## $ `Alch Rel`          <dbl> 6.56, 7.14, 6.52, 6.48, 6.50, 6.50, 7.18, 7.16, 7.~
## $ `Carb Rel`          <dbl> 5.34, 5.58, 5.34, 5.50, 5.38, 5.42, 5.46, 5.42, 5.~
## $ `Balling Lvl`       <dbl> 1.48, 3.04, 1.46, 1.48, 1.46, 1.44, 3.02, 3.00, 3.~
```

```
str(test_df)
```

```
## tibble [267 x 33] (S3: tbl_df/tbl/data.frame)
##  $ Brand Code       : chr [1:267] "D" "A" "B" "B" ...
##  $ Carb Volume      : num [1:267] 5.48 5.39 5.29 5.27 5.41 ...
##  $ Fill Ounces      : num [1:267] 24 24 23.9 23.9 24.2 ...
##  $ PC Volume        : num [1:267] 0.27 0.227 0.303 0.186 0.16 ...
##  $ Carb Pressure    : num [1:267] 65.4 63.2 66.4 64.8 69.4 73.4 65.2 67.4 66.8 72.6 ...
##  $ Carb Temp        : num [1:267] 135 135 140 139 142 ...
##  $ PSC              : num [1:267] 0.236 0.042 0.068 0.004 0.04 0.078 0.088 0.076 0.246 0.146 ...
##  $ PSC Fill         : num [1:267] 0.4 0.22 0.1 0.2 0.3 ...
##  $ PSC CO2          : num [1:267] 0.04 0.08 0.02 0.02 0.06 ...
##  $ Mnf Flow         : num [1:267] -100 -100 -100 -100 -100 -100 -100 -100 -100 -100 ...
##  $ Carb Pressure1   : num [1:267] 117 119 120 125 115 ...
##  $ Fill Pressure    : num [1:267] 46 46.2 45.8 40 51.4 46.4 46.2 40 43.8 40.8 ...
##  $ Hyd Pressure1    : num [1:267] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Hyd Pressure2    : num [1:267] NA 0 0 0 0 0 0 0 0 0 ...
##  $ Hyd Pressure3    : num [1:267] NA 0 0 0 0 0 0 0 0 0 ...
##  $ Hyd Pressure4    : num [1:267] 96 112 98 132 94 94 108 108 110 106 ...
##  $ Filler Level     : num [1:267] 129 120 119 120 116 ...
##  $ Filler Speed     : num [1:267] 3986 4012 4010 NA 4018 ...
##  $ Temperature      : num [1:267] 66 65.6 65.6 74.4 66.4 66.6 66.8 NA 65.8 66 ...
##  $ Usage cont       : num [1:267] 21.7 17.6 24.2 18.1 21.3 ...
##  $ Carb Flow        : num [1:267] 2950 2916 3056 28 3214 ...
##  $ Density          : num [1:267] 0.88 1.5 0.9 0.74 0.88 0.84 1.48 1.6 1.52 1.48 ...
##  $ MFR              : num [1:267] 728 736 735 NA 752 ...
##  $ Balling          : num [1:267] 1.4 2.94 1.45 1.06 1.4 ...
##  $ Pressure Vacuum  : num [1:267] -3.8 -4.4 -4.2 -4 -4 -3.8 -4.2 -4.4 -4.4 -4.2 ...
##  $ PH               : logi [1:267] NA NA NA NA NA NA ...
##  $ Oxygen Filler    : num [1:267] 0.022 0.03 0.046 NA 0.082 0.064 0.042 0.096 0.046 0.096 ...
##  $ Bowl Setpoint    : num [1:267] 130 120 120 120 120 120 120 120 120 120 ...
##  $ Pressure Setpoint: num [1:267] 45.2 46 46 46 50 46 46 46 46 46 ...
##  $ Air Pressurer    : num [1:267] 143 147 147 146 146 ...
##  $ Alch Rel         : num [1:267] 6.56 7.14 6.52 6.48 6.5 6.5 7.18 7.16 7.14 7.78 ...
##  $ Carb Rel         : num [1:267] 5.34 5.58 5.34 5.5 5.38 5.42 5.46 5.42 5.44 5.52 ...
##  $ Balling Lvl      : num [1:267] 1.48 3.04 1.46 1.48 1.46 1.44 3.02 3 3.1 3.12 ...
```

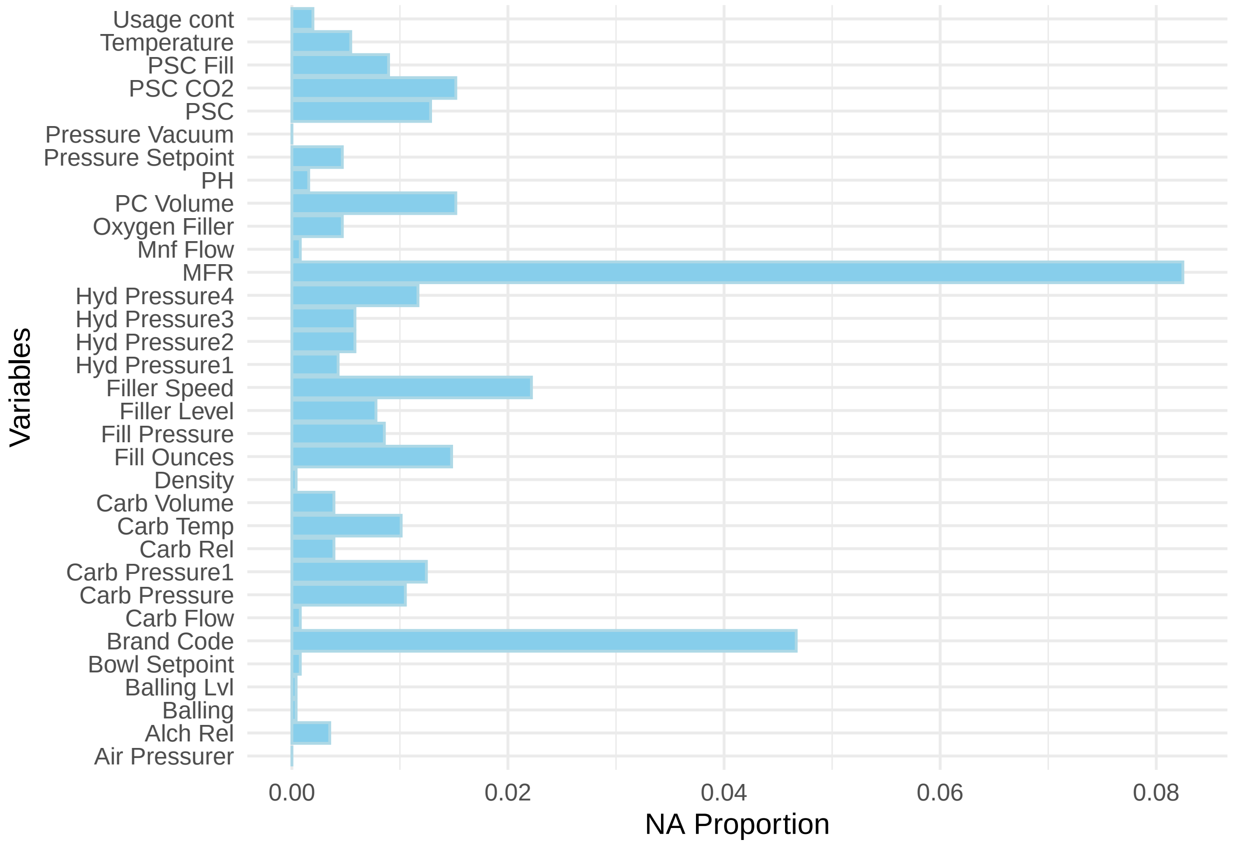
```
summary(test_df)
```

```
##   Brand Code         Carb Volume     Fill Ounces      PC Volume      
##  Length:267         Min.   :5.147   Min.   :23.75   Min.   :0.09867  
##  Class :character   1st Qu.:5.287   1st Qu.:23.92   1st Qu.:0.23333  
##  Mode  :character   Median :5.340   Median :23.97   Median :0.27533  
##                     Mean   :5.369   Mean   :23.97   Mean   :0.27769  
##                     3rd Qu.:5.465   3rd Qu.:24.01   3rd Qu.:0.32200  
##                     Max.   :5.667   Max.   :24.20   Max.   :0.46400  
##                     NA's   :1       NA's   :6       NA's   :4        
##  Carb Pressure     Carb Temp          PSC             PSC Fill     
##  Min.   :60.20   Min.   :130.0   Min.   :0.00400   Min.   :0.0200  
##  1st Qu.:65.30   1st Qu.:138.4   1st Qu.:0.04450   1st Qu.:0.1000  
##  Median :68.00   Median :140.8   Median :0.07600   Median :0.1800  
##  Mean   :68.25   Mean   :141.2   Mean   :0.08545   Mean   :0.1903  
##  3rd Qu.:70.60   3rd Qu.:143.8   3rd Qu.:0.11200   3rd Qu.:0.2600  
##  Max.   :77.60   Max.   :154.0   Max.   :0.24600   Max.   :0.6200  
##                  NA's   :1       NA's   :5         NA's   :3       
##     PSC CO2           Mnf Flow       Carb Pressure1  Fill Pressure  
##  Min.   :0.00000   Min.   :-100.20   Min.   :113.0   Min.   :37.80  
##  1st Qu.:0.02000   1st Qu.:-100.00   1st Qu.:120.2   1st Qu.:46.00  
##  Median :0.04000   Median :   0.20   Median :123.4   Median :47.80  
##  Mean   :0.05107   Mean   :  21.03   Mean   :123.0   Mean   :48.14  
##  3rd Qu.:0.06000   3rd Qu.: 141.30   3rd Qu.:125.5   3rd Qu.:50.20  
##  Max.   :0.24000   Max.   : 220.40   Max.   :136.0   Max.   :60.20  
##  NA's   :5                           NA's   :4      NA's   :2
##  Hyd Pressure1    Hyd Pressure2    Hyd Pressure3    Hyd Pressure4
##  Min.   :-50.00   Min.   :-50.00   Min.   :-50.00   Min.   : 68.00
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.: 90.00
##  Median : 10.40   Median : 26.80   Median : 27.70   Median : 98.00
##  Mean   : 12.01   Mean   : 20.11   Mean   : 19.61   Mean   : 97.84
##  3rd Qu.: 20.40   3rd Qu.: 34.80   3rd Qu.: 33.00   3rd Qu.:104.00
##  Max.   : 50.00   Max.   : 61.40   Max.   : 49.20   Max.   :140.00
##                   NA's   :1        NA's   :1        NA's   :4
##   Filler Level     Filler Speed   Temperature      Usage cont      Carb Flow
##  Min.   : 69.2   Min.   :1006   Min.   :63.80   Min.   :12.90   Min.   :   0
##  1st Qu.:100.6   1st Qu.:3812   1st Qu.:65.40   1st Qu.:18.12   1st Qu.:1083
##  Median :118.6   Median :3978   Median :65.80   Median :21.44   Median :3038
##  Mean   :110.3   Mean   :3581   Mean   :66.23   Mean   :20.90   Mean   :2409
##  3rd Qu.:120.2   3rd Qu.:3996   3rd Qu.:66.60   3rd Qu.:23.74   3rd Qu.:3215
##  Max.   :153.2   Max.   :4020   Max.   :75.40   Max.   :24.60   Max.   :3858
##  NA's   :2       NA's   :10     NA's   :2       NA's   :2
##     Density            MFR            Balling       Pressure Vacuum
##  Min.   :0.060   Min.   : 15.6   Min.   :0.902   Min.   :-6.400
##  1st Qu.:0.920   1st Qu.:707.0   1st Qu.:1.498   1st Qu.:-5.600
##  Median :0.980   Median :724.6   Median :1.648   Median :-5.200
##  Mean   :1.177   Mean   :697.8   Mean   :2.203   Mean   :-5.174
##  3rd Qu.:1.600   3rd Qu.:731.5   3rd Qu.:3.242   3rd Qu.:-4.800
##  Max.   :1.840   Max.   :784.8   Max.   :3.788   Max.   :-3.600
##  NA's   :1       NA's   :31      NA's   :1       NA's   :1
##     PH           Oxygen Filler     Bowl Setpoint   Pressure Setpoint
##  Mode:logical   Min.   :0.00240   Min.   : 70.0   Min.   :44.00
##  NA's:267       1st Qu.:0.01960   1st Qu.:100.0   1st Qu.:46.00
##                 Median :0.03370   Median :120.0   Median :46.00
##                 Mean   :0.04666   Mean   :109.6   Mean   :47.73
##                 3rd Qu.:0.05440   3rd Qu.:120.0   3rd Qu.:50.00
##                 Max.   :0.39800   Max.   :130.0   Max.   :52.00
##                 NA's   :3         NA's   :1       NA's   :2
##  Air Pressurer      Alch Rel        Carb Rel      Balling Lvl
##  Min.   :141.2   Min.   :6.400   Min.   :5.18   Min.   :0.000
##  1st Qu.:142.2   1st Qu.:6.540   1st Qu.:5.34   1st Qu.:1.380
##  Median :142.6   Median :6.580   Median :5.40   Median :1.480
##  Mean   :142.8   Mean   :6.907   Mean   :5.44   Mean   :2.051
##  3rd Qu.:142.8   3rd Qu.:7.180   3rd Qu.:5.56   3rd Qu.:3.080
##  Max.   :147.2   Max.   :7.820   Max.   :5.74   Max.   :3.420
##  NA's   :1       NA's   :3       NA's   :2
```

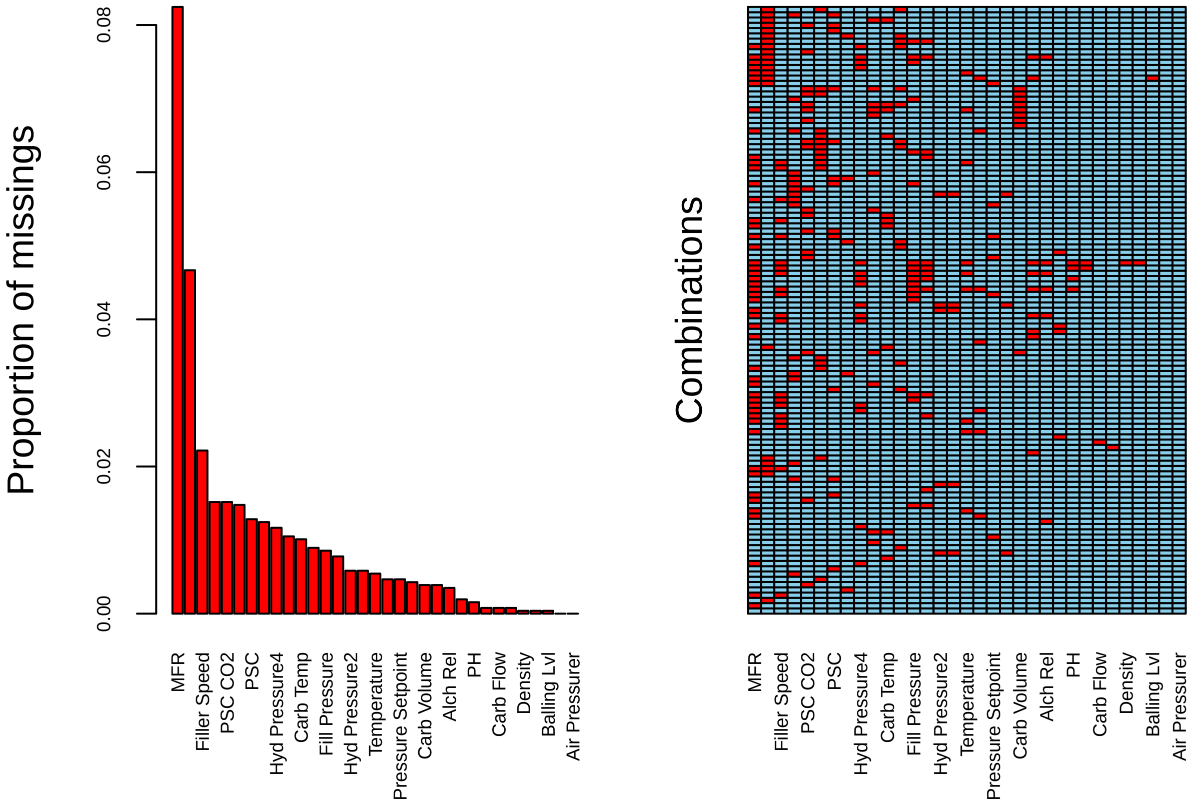
NA Proportions

```r
missing_train_df <- train_df %>%
                summarise(across(everything(), ~mean(is.na(.)))) %>%
                pivot_longer(cols = everything(), names_to = "variable", values_to = "na_proportion")

# Create a bar plot using ggplot2
ggplot(missing_train_df, aes(x = variable, y = na_proportion)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "lightblue") +
  theme_minimal() +
  labs(y = "NA Proportion", x = "Variables") +
  coord_flip()
```



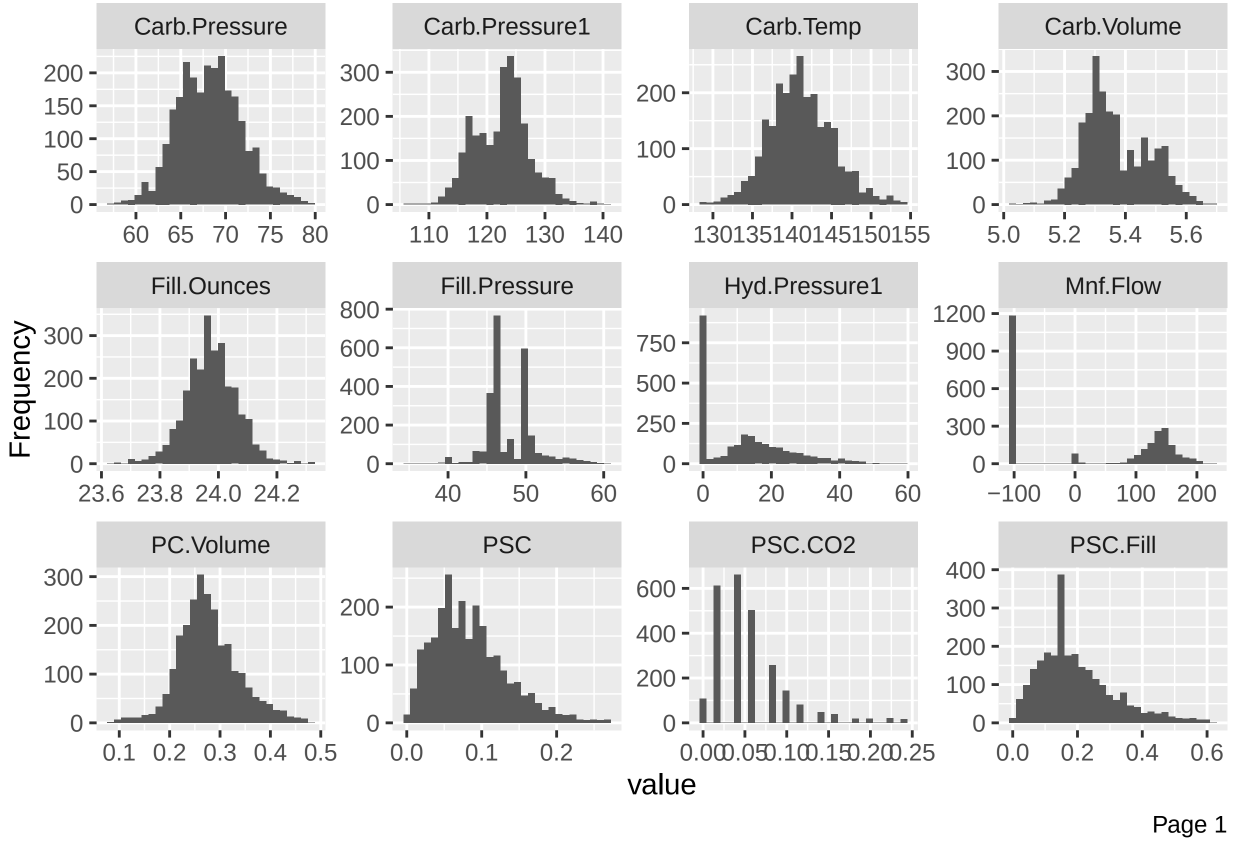
```
VIM::aggr(train_df, numbers=T, sortVars=T, bars = FALSE,
          cex.axis = .6)
```

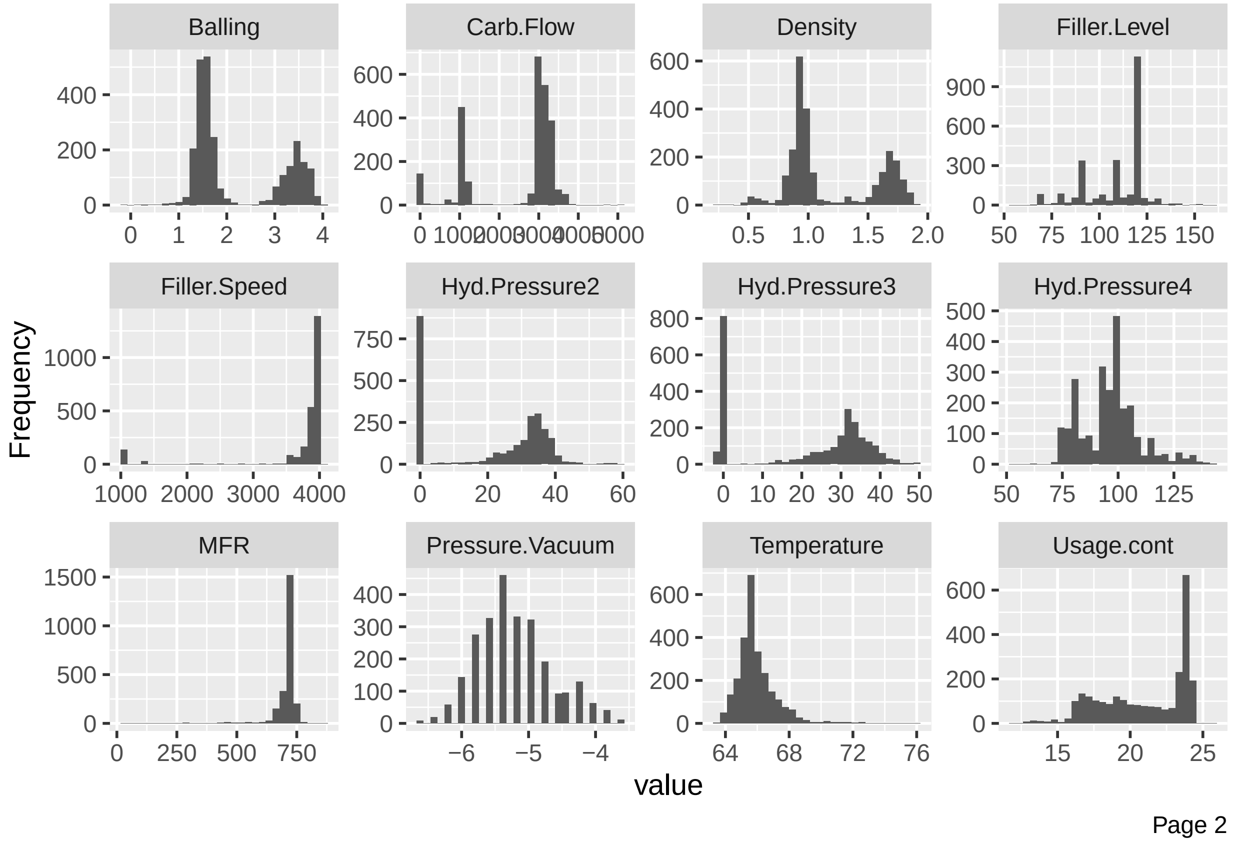


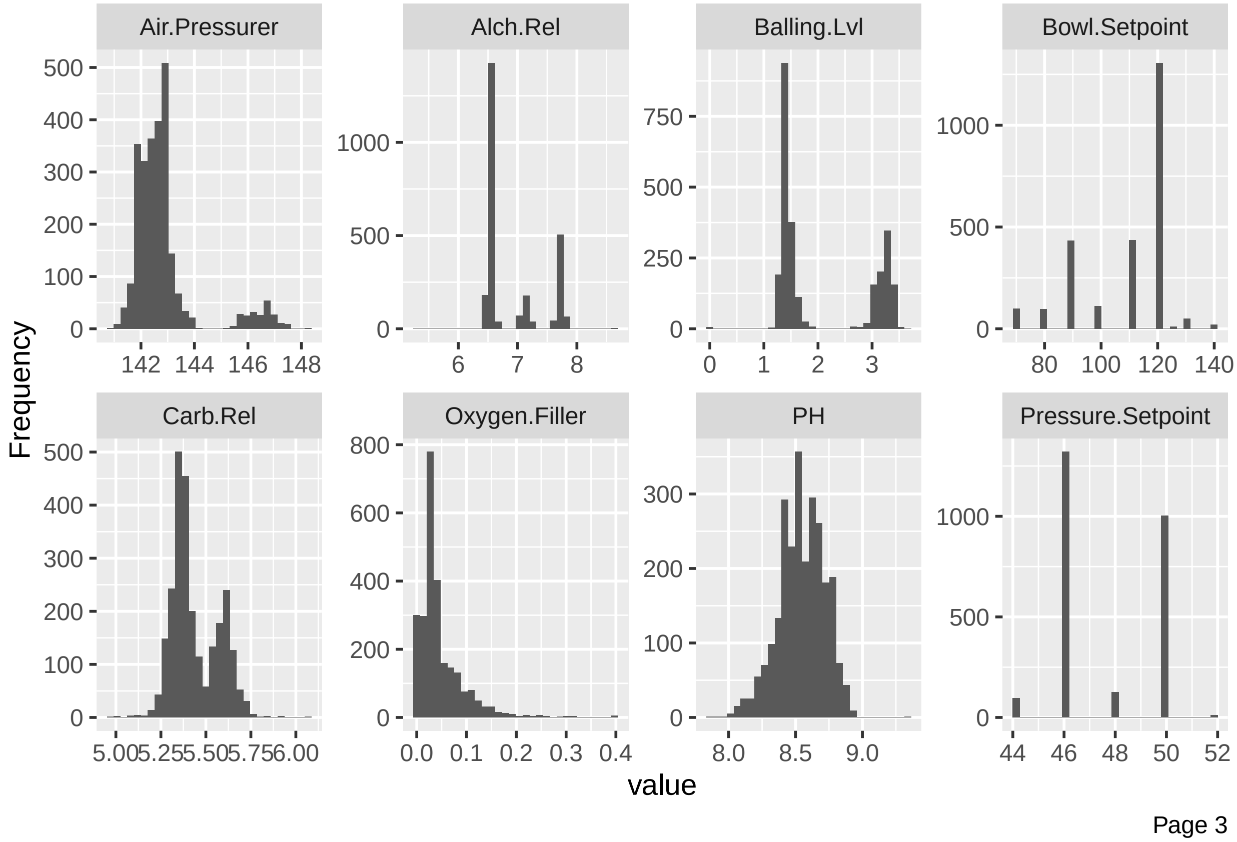
```
##
##  Variables sorted by number of missings:
##           Variable        Count
##                MFR 0.0824581875
##         Brand Code 0.0466744457
##       Filler Speed 0.0221703617
##          PC Volume 0.0151691949
##            PSC CO2 0.0151691949
##        Fill Ounces 0.0147802412
##                PSC 0.0128354726
##     Carb Pressure1 0.0124465189
##      Hyd Pressure4 0.0116686114
##      Carb Pressure 0.0105017503
##          Carb Temp 0.0101127966
##           PSC Fill 0.0089459354
##      Fill Pressure 0.0085569817
##       Filler Level 0.0077790743
##      Hyd Pressure2 0.0058343057
##      Hyd Pressure3 0.0058343057
##        Temperature 0.0054453520
##      Oxygen Filler 0.0046674446
##  Pressure Setpoint 0.0046674446
##      Hyd Pressure1 0.0042784909
##        Carb Volume 0.0038895371
##           Carb Rel 0.0038895371
##           Alch Rel 0.0035005834
##          Usage cont 0.0019447686
##                  PH 0.0015558149
##            Mnf Flow 0.0007779074
##           Carb Flow 0.0007779074
##       Bowl Setpoint 0.0007779074
##             Density 0.0003889537
##             Balling 0.0003889537
##         Balling Lvl 0.0003889537
##     Pressure Vacuum 0.0000000000
##       Air Pressurer 0.0000000000
```

Distribution

```r
DataExplorer::plot_histogram(train_df, nrow = 3L, ncol = 4L)
```




Balling
Carb.Flow
Density
Filler.Level
Filler.Speed
Hyd.Pressure2
Hyd.Pressure3
Hyd.Pressure4
MFR
Pressure.Vacuum
Temperature
Usage.cont
Frequency
value
0
200
400
0 1 2 3 4
600
0 1000 2000 3000 4000 5000
0.5 1.0 1.5 2.0
300
900
50 75 100 125 150
500
1000
1000 2000 3000 4000
250
750
0 20 40 60
800
0 10 20 30 40 50
100
300
50 75 100 125
1500
0 250 500 750
−6 −5 −4
64 68 72 76
15 20 25
Page 2



Initial Findings

- Data consists of 2571 observations with 33 columns
- `Brand Code`:
  - Type character
  - Unordered categorical values
- Predictors:
  - Primarily doubles
  - 4 can be considered integers
  - High range variables:
    1. `Mnf Flow` -100.20 to 220.40
    2. `Hyd Pressure1` -50.00 to 50.00
    3. `Hyd Pressure2` -50.00 to 61.40
    4. `Hyd Pressure3` -50.00 to 49.20
    5. `Hyd Pressure4` 68.00 to 140.00
- About 8% of the values for `MFR` is missing.

- `Brand Code` is missing about 5%
- Filler Speed is missing about 2%
- Remaining Variables have roughly 1% or less missing.

- `Pressure.Vacuum`, `Air.Pressurer` have no NAs

- The Distribution of the variables can be grouped as **left skewed**, **right skewed** and for symmetric we can categorized as **relatively normal**
  - Relatively Normal Distributions:
    - `Carb.Pressure`
    - `Carb.Temp` -`Fill.Ounces`
    - `PC.Volume`
    - `PH`
  - Left-skew Distributions:
    - `Carb.Flow`
    - `Filler.Speed`
    - `Mnf.Flow`
    - `MFR`
    - `Bowl.Setpoint`
    - `Filler.Level`
    - `Hyd.Pressure2`
    - `Hyd.Pressure3` -`Usage.cont`
    - `Carb.Pressure1`
    - `Filler.Speed`
  - Right-skew Distributions:
    - `Pressure.Setpoint`
    - `Fill.Pressure`
    - `Hyd.Pressure1`
    - `Temperature`
    - `Carb.Volume`
    - `PSC`
    - `PSC.CO2`
    - `PSC.Fill`
    - `Balling`
    - `Density`
    - `Hyd.Pressure4`
    - `Air.Pressurer`
    - `Alch.Rel`
    - `Carb.Rel`
    - `Oxygen.Filler`
    - `Balling.Lvl`
    - `Pressure.Vacuum`

```
unique(train_df$`Brand Code`)
```

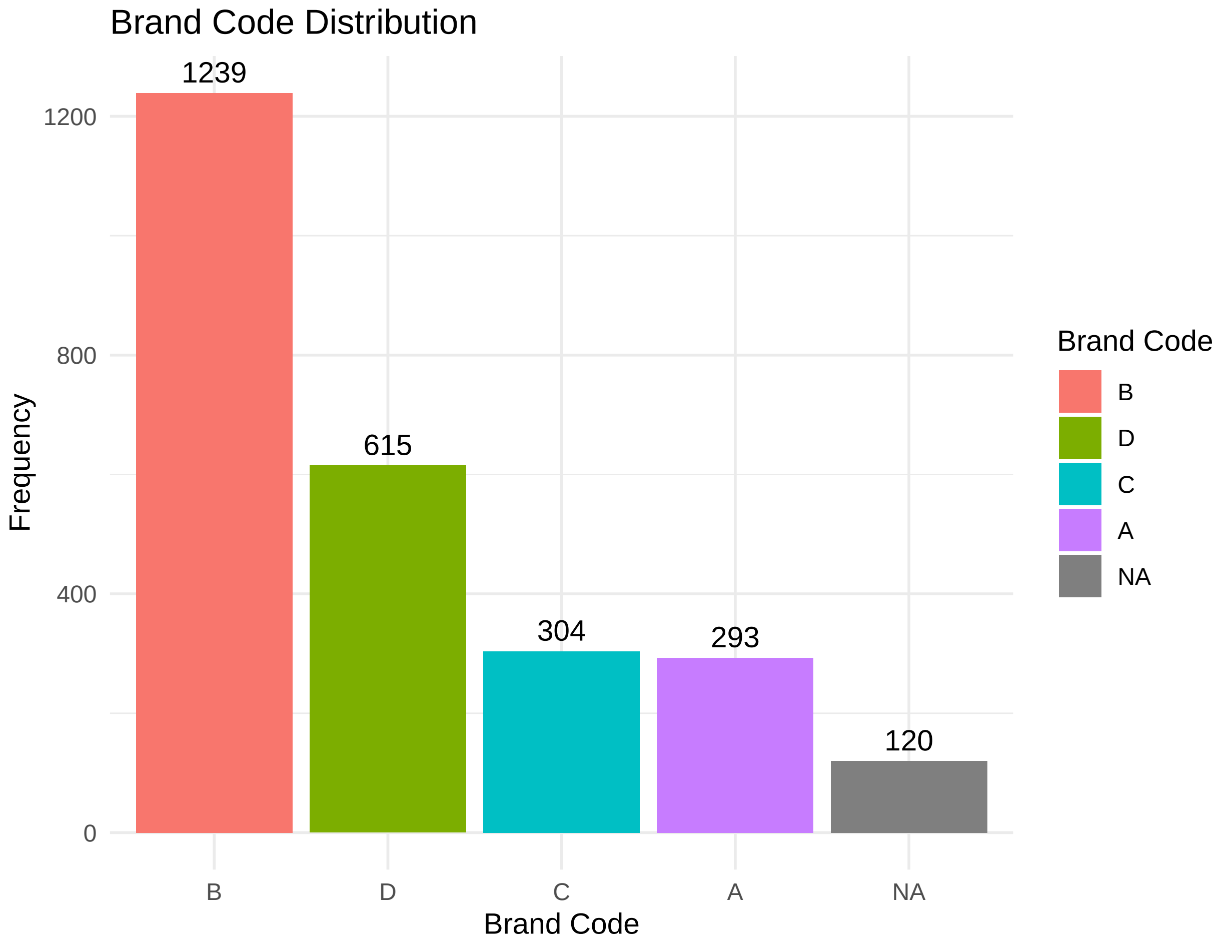
```
## [1] "B" "A" "C" "D" NA
```

Brand Code Distribution

Noting that `Brand Code` has 4 categorical values outside of NA (**A,B,C,D**), further investigation of each values distribution is needed.

```
train_df %>%
  mutate(`Brand Code` = factor(`Brand Code`, levels = names(sort(table(`Brand Code`), decreasing = TRUE
  ggplot(aes(x = `Brand Code`, fill = `Brand Code`)) +
  geom_bar(stat = "count") +
  geom_text(stat = 'count', aes(label = ..count..), vjust = -0.5, color = "black") +
  labs(title = 'Brand Code Distribution', x = 'Brand Code', y = 'Frequency') +
  theme_minimal()
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



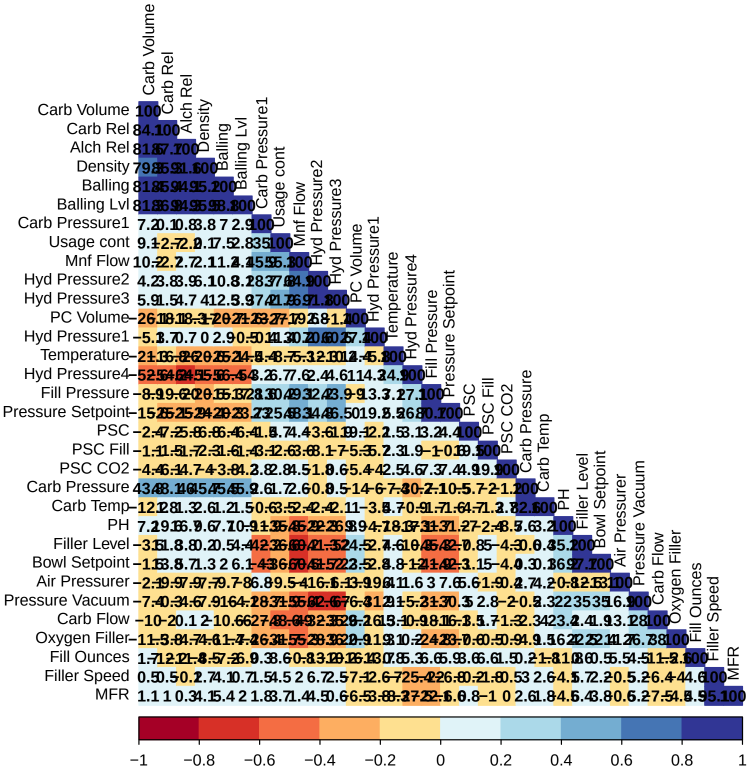
Correlation

General

```
train_numeric_df <- train_df %>%
  dplyr::select(where(is.numeric)) %>%
  na.omit()

# Calculate correlation matrix
train_numeric_cor <- cor(train_numeric_df)
```

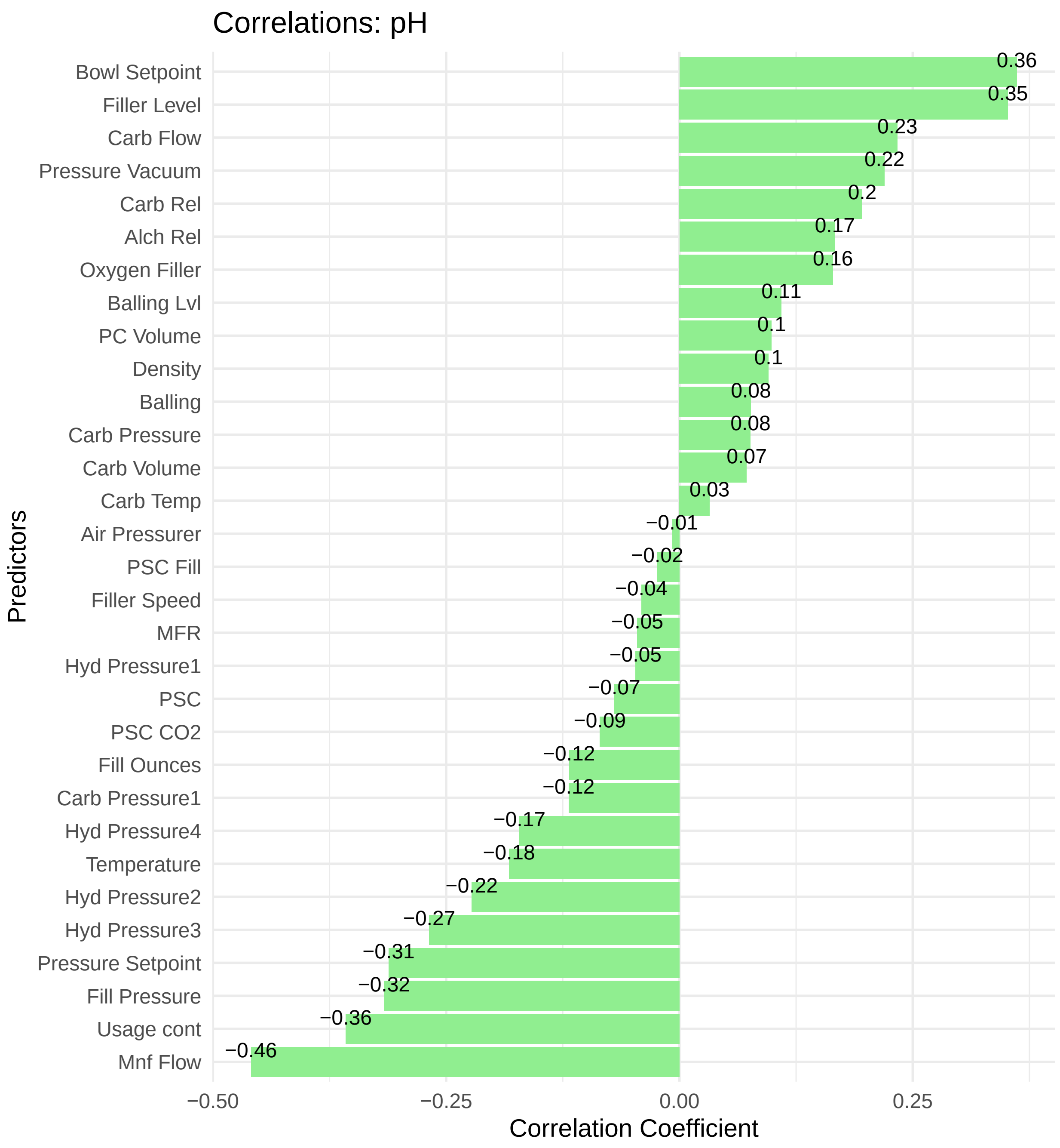
```r
# Generate the correlation plot
corrplot(train_numeric_cor,
         method = "color",
         tl.col = "black",
         col = brewer.pal(n = 10,
                          name = "RdYlBu"),
         type = "lower",
         order = "hclust",
         addCoef.col = "black",
         number.cex = 0.8,
         tl.cex = 0.8,
         cl.cex = 0.8,
         addCoefasPercent = TRUE,
         number.digits = 1)
```


Carb Volume
Carb Rel
Alch Rel
Density
Balling
Balling Lvl
Carb Pressure1
Usage cont
Mnf Flow
Hyd Pressure2
Hyd Pressure3
PC Volume
Hyd Pressure1
Temperature
Hyd Pressure4
Fill Pressure
Pressure Setpoint
PSC
PSC Fill
PSC CO2
Carb Pressure
Carb Temp
PH
Filler Level
Bowl Setpoint
Air Pressurer
Pressure Vacuum
Carb Flow
Oxygen Filler
Fill Ounces
Filler Speed
MFR
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1

PH

With PH being our response variable, assessing PH correlation with other variables is needed.

```r
train_numeric_df %>%
  dplyr::select(-PH) %>%  # Exclude 'PH' from predictors if needed
  cor(train_numeric_df$PH) %>%  # Calculate correlations with 'PH'
  as.data.frame() %>%
  rownames_to_column(var = "Predictor") %>%
  filter(Predictor != "PH") %>%  # Ensure 'PH' is not included as its own predictor
  mutate(Predictor = fct_reorder(factor(Predictor), V1)) %>%  # Reorder factors by correlation for plot
  ggplot(aes(x = Predictor, y = V1, label = round(V1, 2))) +
    geom_col(fill = "lightgreen") +
    geom_text(color = "black", size = 3, vjust = -0.3) +
    coord_flip() +
    labs(title = "Correlations: pH", x = "Predictors", y = "Correlation Coefficient") +
    theme_minimal()
```



Correlation Findings

Multicolliniarity is a concern, based on our plots, considering the number of predictor variables with significant correlation.

Data Cleanup

- Transform `Brand Code` which will be mutated to categorized factors as in **r chunk** `brand_code_dist`.
- Identify unhelpful data:

- Identifying variables with zero variance (`zeroVar`) variables
- Identify near-zero variance (nzv).
- Remove an rows with NAs in our response variable, as it will interfere with analysis in the future.

```
train_df%>%
  dplyr::filter(!is.na(PH))
```

```
## # A tibble: 2,567 x 33
##    `Brand Code` `Carb Volume` `Fill Ounces` `PC Volume` `Carb Pressure`
##    <chr>                <dbl>         <dbl>       <dbl>           <dbl>
##  1 B                     5.34          24.0       0.263            68.2
##  2 A                     5.43          24.0       0.239            68.4
##  3 B                     5.29          24.1       0.263            70.8
##  4 A                     5.44          24.0       0.293            63
##  5 A                     5.49          24.3       0.111            67.2
##  6 A                     5.38          23.9       0.269            66.6
##  7 A                     5.31          23.9       0.268            64.2
##  8 B                     5.32          24.2       0.221            67.6
##  9 B                     5.25          24.0       0.263            64.2
## 10 B                     5.27          24.0       0.231            72
## # i 2,557 more rows
## # i 28 more variables: `Carb Temp` <dbl>, PSC <dbl>, `PSC Fill` <dbl>,
## #   `PSC CO2` <dbl>, `Mnf Flow` <dbl>, `Carb Pressure1` <dbl>,
## #   `Fill Pressure` <dbl>, `Hyd Pressure1` <dbl>, `Hyd Pressure2` <dbl>,
## #   `Hyd Pressure3` <dbl>, `Hyd Pressure4` <dbl>, `Filler Level` <dbl>,
## #   `Filler Speed` <dbl>, Temperature <dbl>, `Usage cont` <dbl>,
## #   `Carb Flow` <dbl>, Density <dbl>, MFR <dbl>, Balling <dbl>, ...
```

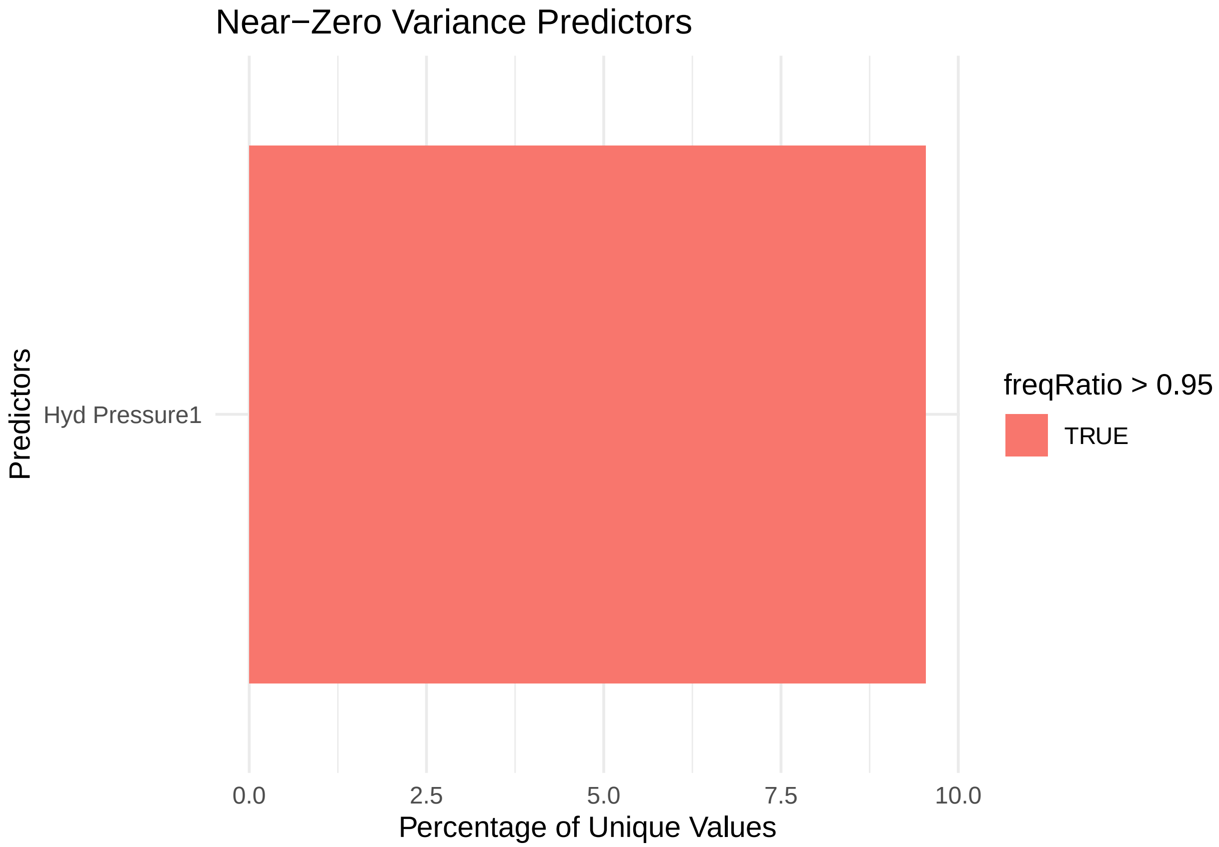
```
train_df<-train_df%>%
  dplyr::filter(!is.na(PH))
```

```
train_df<- train_df %>%
  dplyr::mutate(`Brand Code` = factor(`Brand Code`,
                          levels = c('A','B','C','D','not known'),
                          ordered = FALSE))
```

```
nzv_df <- nearZeroVar(train_df, saveMetrics= TRUE)
nzv_df <- as.data.frame(nzv_df) %>%
  rownames_to_column(var = "Predictor")

nzv_filtered_df <- nzv_df %>%
  filter(nzv == TRUE)

ggplot(nzv_filtered_df, aes(x = Predictor, y = percentUnique, fill = freqRatio > 0.95)) +
  geom_col(position = "dodge") +
  coord_flip() +
  labs(title = "Near-Zero Variance Predictors",
       x = "Predictors",
       y = "Percentage of Unique Values") +
  theme_minimal()
```



```
print(nzv_filtered_df)
```

```
##       Predictor freqRatio percentUnique zeroVar  nzv
## 1 Hyd Pressure1  31.03704      9.544215   FALSE TRUE
```

Modeling

Preliminary Data Processing

Pre-processing Steps:

- Transform the data using as.dataframe() otherwise `preProcess` function from `caret` fails
- Remove separate response variable from predictors
- leverage `caret` package method preProcess to transform data using methods:
  - knnImpute - nearest neighbor to impute missing data
  - nzv = remove near-zero values identified above
  - corr = filters out highly correlated values addressing multicollinearity
  - center = subtracts the mean of the predictor's data (again from the data in x) from the predictor values
  - scale = divides by the standard deviation.
  - BoxCox = normalizes data
- Use the `predict` function to process the list variables created with `preProcess()` to recreate the dataframe.

- Rejoin PH to the dataframe.

```
set.seed(1234)

train_df<- as.data.frame(train_df)

#remove pH from the train data set in order to only transform the predictors
train_preprocess_df <- train_df %>%
  dplyr::select(-c(PH))

preProc_ls <- preProcess(train_preprocess_df, method = c("knnImpute", "nzv", "corr", "center", "scale",

train_preProc_df <- predict(preProc_ls, train_preprocess_df)
train_preProc_df$PH <- train_df$PH
# To verify no NAs produced when recombining
train_preProc_df%>%
  dplyr::filter(is.na(PH))
```

```
##  [1] Brand Code        Carb Volume       Fill Ounces       PC Volume
##  [5] Carb Pressure     Carb Temp         PSC               PSC Fill
##  [9] PSC CO2           Mnf Flow          Carb Pressure1    Fill Pressure
## [13] Hyd Pressure2     Hyd Pressure4     Temperature       Usage cont
## [17] Carb Flow         MFR               Pressure Vacuum   Oxygen Filler
## [21] Bowl Setpoint     Pressure Setpoint Air Pressurer     Alch Rel
## [25] Carb Rel          PH
## <0 rows> (or 0-length row.names)
```

Data Partition

```
training_set_df <- createDataPartition(train_preProc_df$PH, p=0.8, list=FALSE)

train_proc_df <- train_preProc_df[training_set_df,]
eval_proc_df <- train_preProc_df[-training_set_df,]
```

PLS

```
set.seed(222)
y_train <- subset(train_proc_df, select = -c(PH))
y_test <- subset(eval_proc_df, select = -c(PH))
```

```
set.seed(2341)
#generate model
pls_model <- train(y_train, train_proc_df$PH,
                      method='pls',
                      metric='Rsquared',
                      tuneLength=10,
                      trControl=trainControl(method = "cv",  number = 10))
#evaluate model metrics
plsPred <-predict(pls_model, newdata=y_test)
plsReSample <- postResample(pred=plsPred, obs = eval_proc_df$PH)
plsReSample %>% kable() %>% kable_paper()
```

	x
RMSE	0.1296989
Rsquared	0.3892951
MAE	0.1030896