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(pls)
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.~
```

```
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 ...
```

```
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)
```

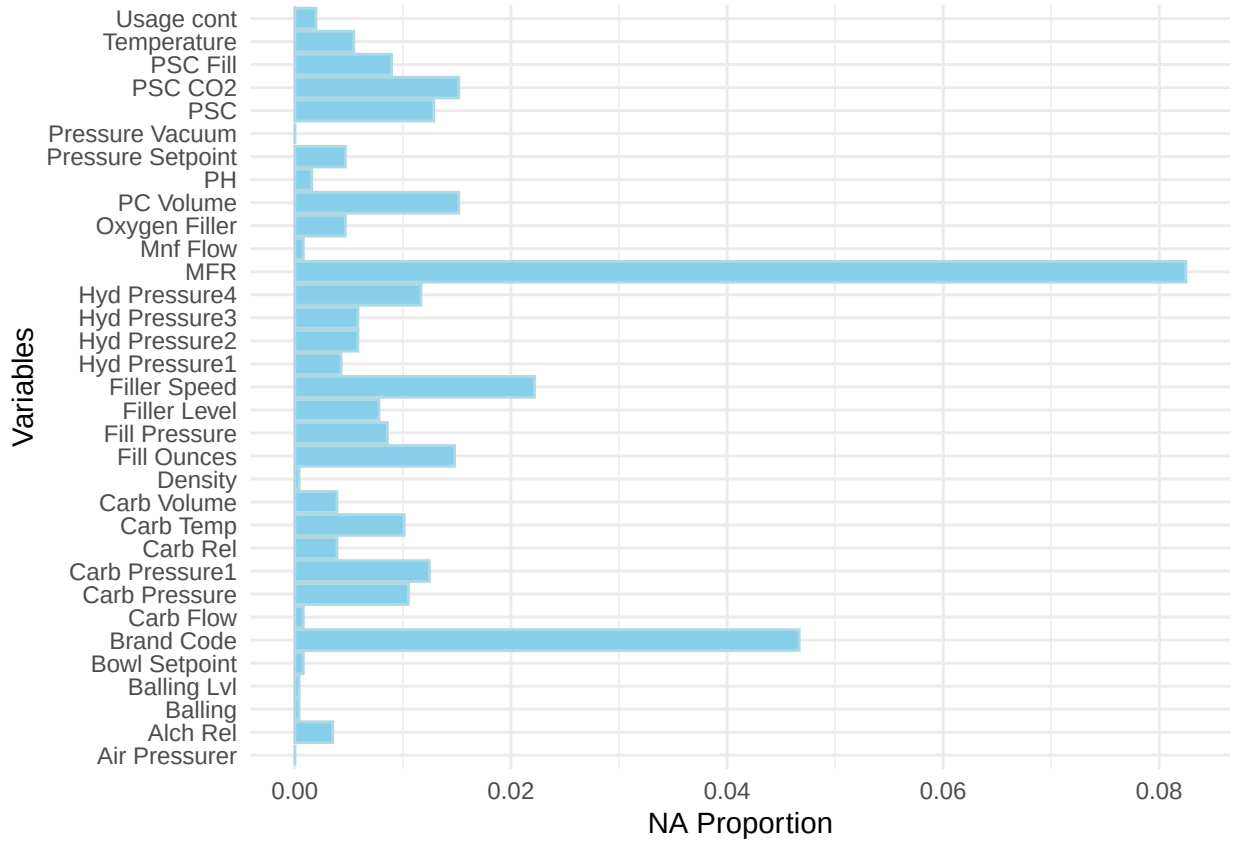
```
##   Brand Code          Carb Volume     Fill Ounces      PC Volume
##  Length:267         Min.   :5.147   Min.   :23.75   Min.   :0.09867
##  Class :character   1st Qu.:5.287   1st Qu.:23.92   1st Qu.:0.23333
##  Mode  :character   Median :5.340   Median :23.97   Median :0.27533
##                     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
```

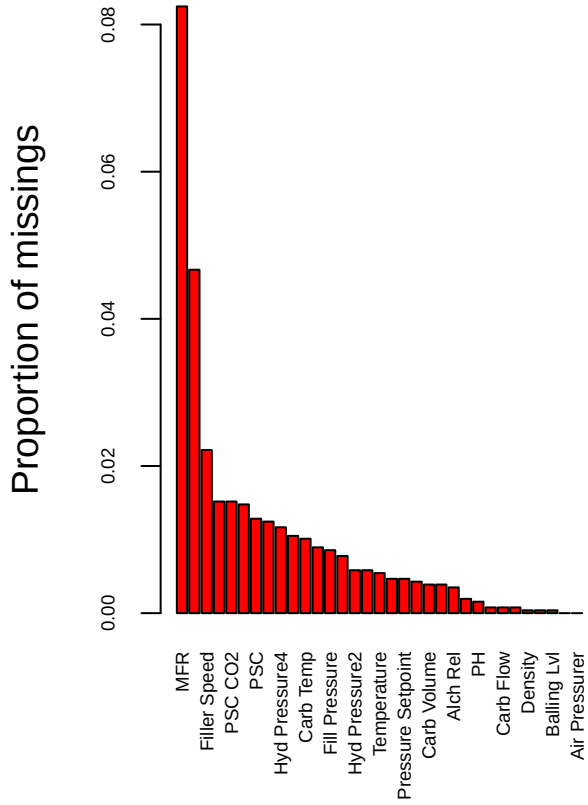
NA Proportions

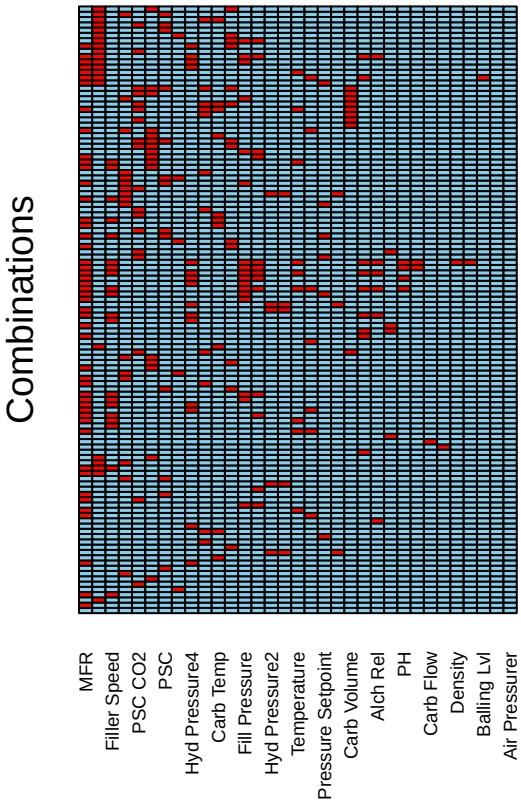
```
missing_train_df <- train_df %>%
                summarise(across(everything(), ~mean(is.na(.)))) %>%
                pivot_longer(cols = everything(), names_to = "variable", values_to = "na_proportion")

# 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)
```



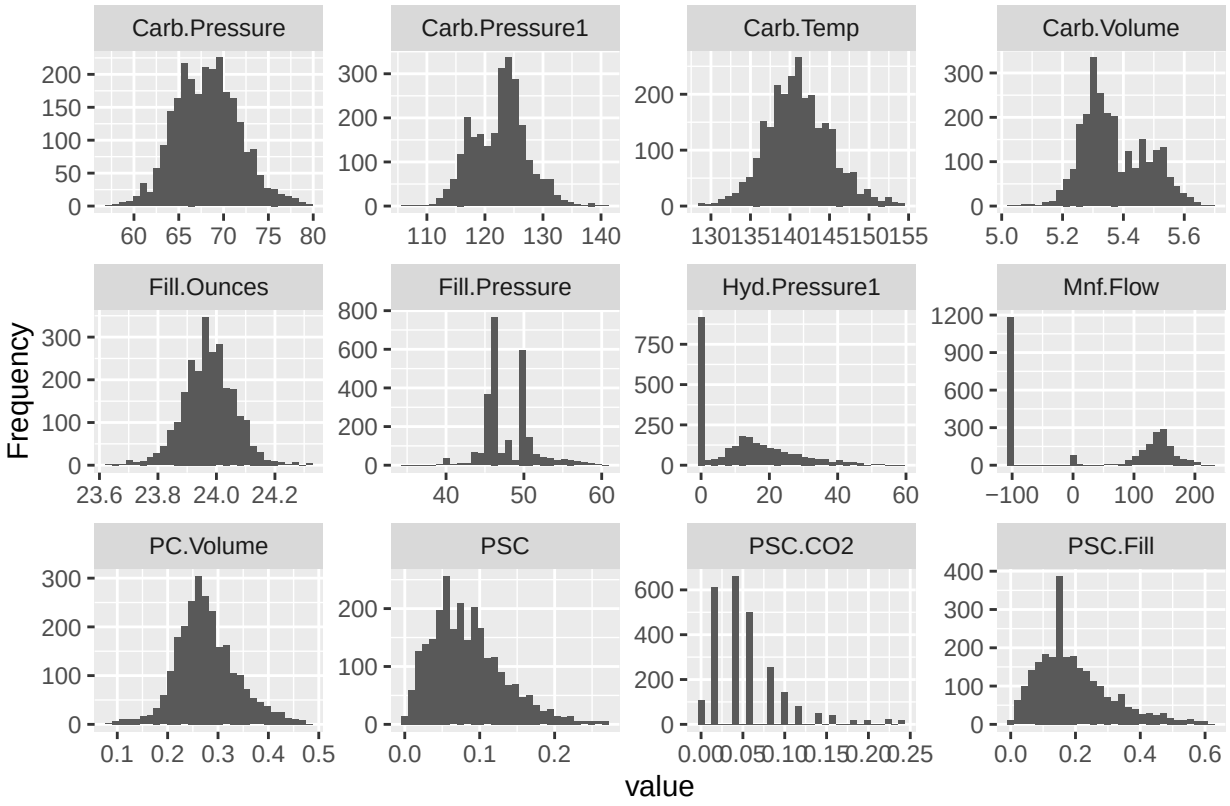



```
##
##  Variables sorted by number of missings:
##            Variable        Count
##                 MFR 0.0824581875
##          Brand Code 0.0466744457
##        Filler Speed 0.0221703617
##           PC Volume 0.0151691949
##             PSC CO2 0.0151691949
##         Fill Ounces 0.0147802412
##                 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
```

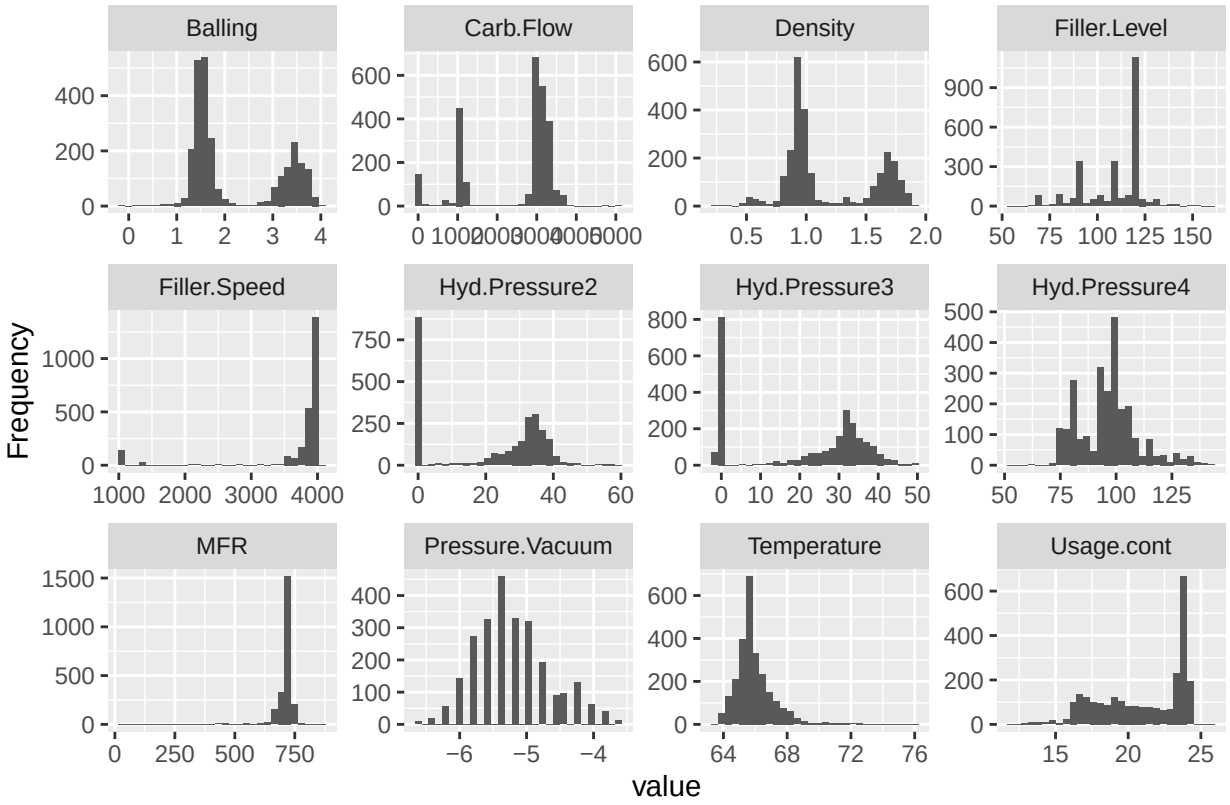
```
##          Usage cont 0.0019447686
##                  PH 0.0015558149
##           Mnf Flow 0.0007779074
##          Carb Flow 0.0007779074
##      Bowl Setpoint 0.0007779074
##             Density 0.0003889537
##             Balling 0.0003889537
##         Balling Lvl 0.0003889537
##    Pressure Vacuum 0.0000000000
##      Air Pressurer 0.0000000000
```

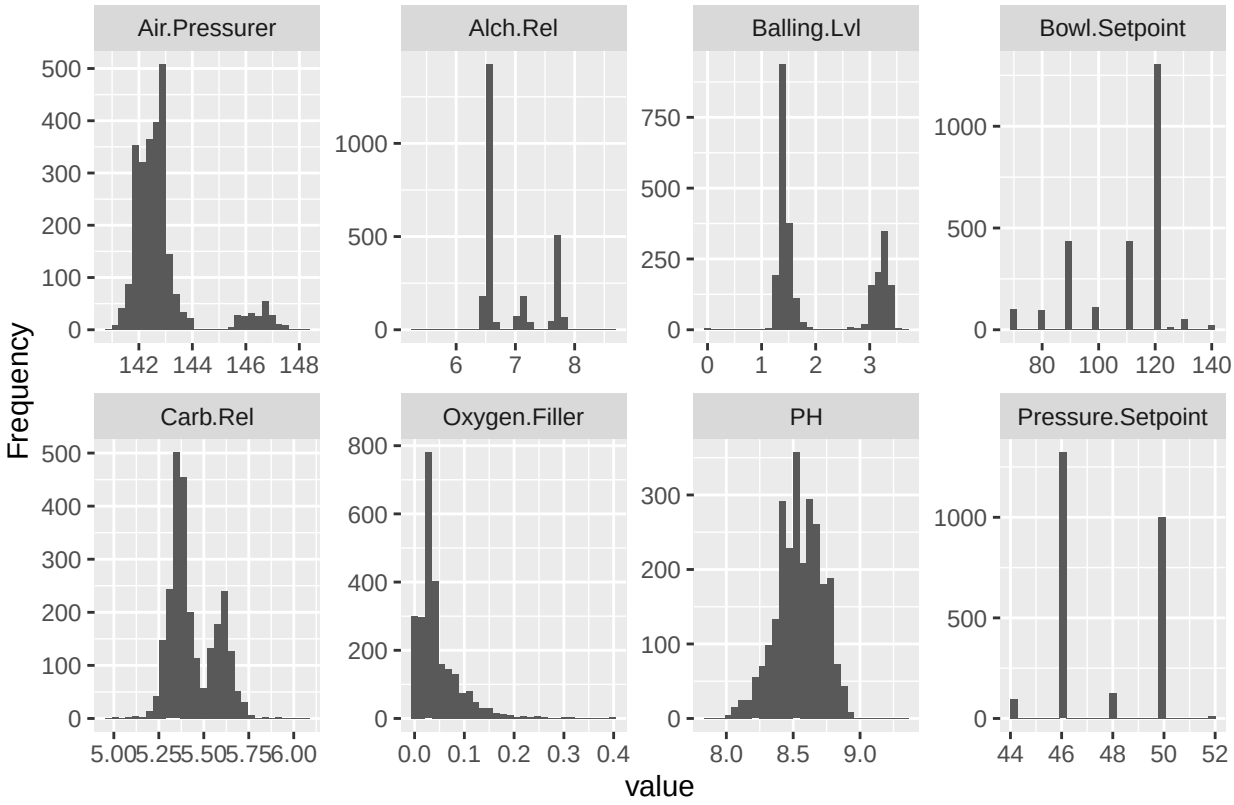
Distribution

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



Page 1


Balling
Carb.Flow
Density
Filler.Level
Filler.Speed
Hyd.Pressure2
Hyd.Pressure3
Hyd.Pressure4
MFR
Pressure.Vacuum
Temperature
Usage.cont
Frequency
value



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:
    i. `Mnf Flow` -100.20 to 220.40
    ii. `Hyd Pressure1` -50.00 to 50.00
    iii. `Hyd Pressure2` -50.00 to 61.40
    iv. `Hyd Pressure3` -50.00 to 49.20
    v. `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
```

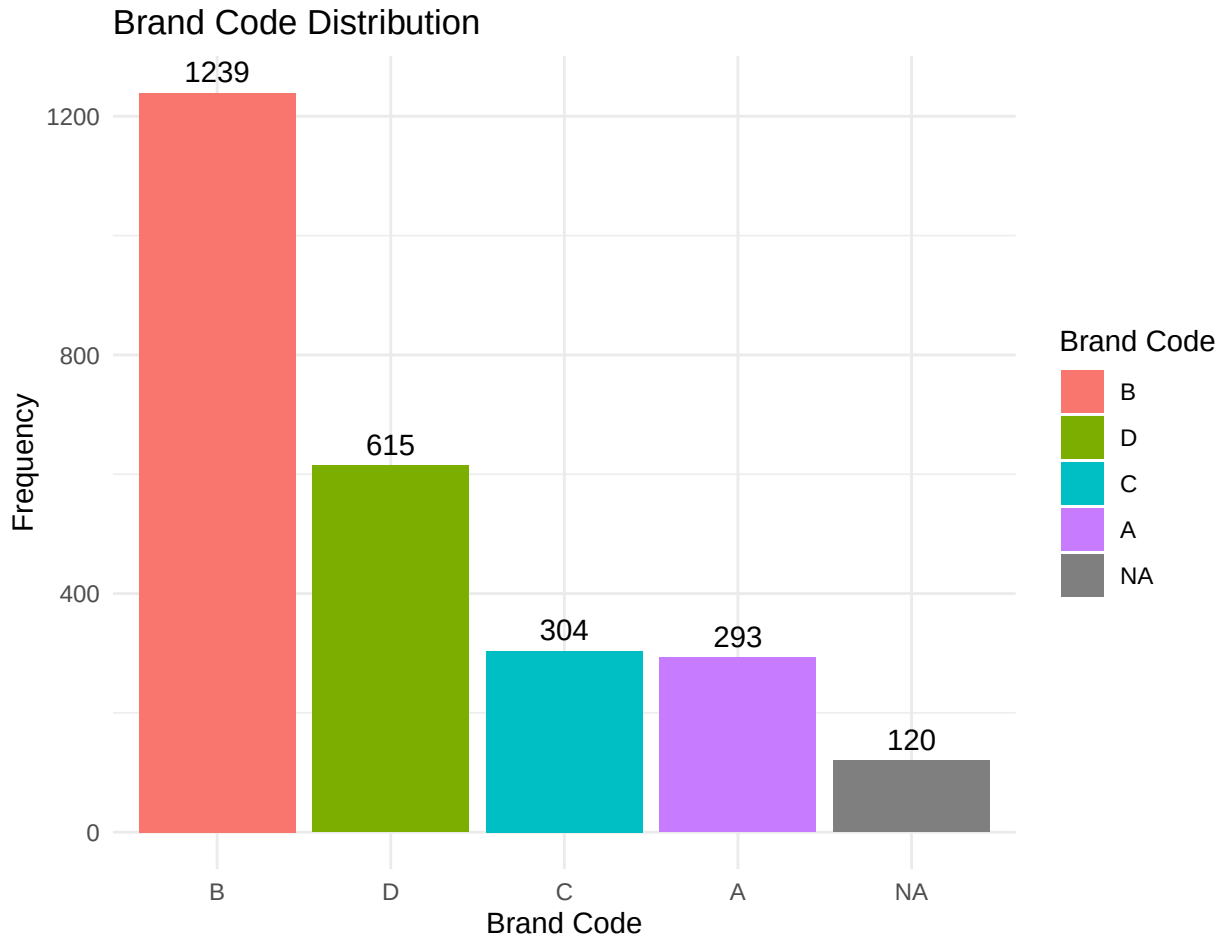
Brand Code Distribution

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

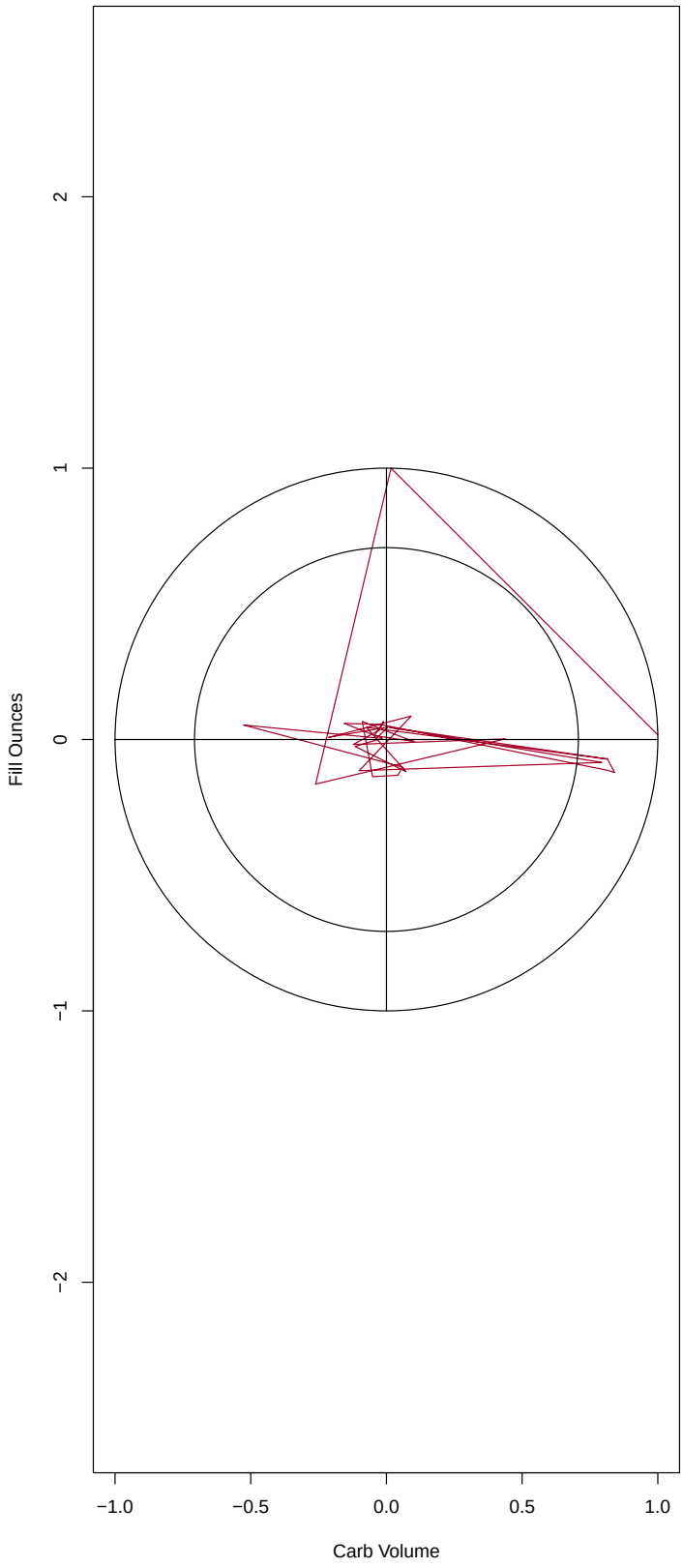
```
train_numeric_df <- train_df %>%
  dplyr::select(where(is.numeric)) %>%
  na.omit()

# 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)
```

```
## Warning in plot.window(...): "method" is not a graphical parameter

## Warning in plot.window(...): "tl.col" is not a graphical parameter

## Warning in plot.window(...): "order" is not a graphical parameter

## Warning in plot.window(...): "addCoef.col" is not a graphical parameter

## Warning in plot.window(...): "number.cex" is not a graphical parameter

## Warning in plot.window(...): "tl.cex" is not a graphical parameter

## Warning in plot.window(...): "cl.cex" is not a graphical parameter

## Warning in plot.window(...): "addCoefasPercent" is not a graphical parameter

## Warning in plot.window(...): "number.digits" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): plot type 'lower' will be truncated to first
## character

## Warning in plot.xy(xy, type, ...): "method" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "tl.col" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "order" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "addCoef.col" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "number.cex" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "tl.cex" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "cl.cex" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "addCoefasPercent" is not a graphical
## parameter

## Warning in plot.xy(xy, type, ...): "number.digits" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "method" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.col" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "order" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "addCoef.col" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "number.cex" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.cex" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "cl.cex" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "addCoefasPercent"
## is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "number.digits" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "method" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.col" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "order" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "addCoef.col" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "number.cex" is
## not a graphical parameter
```

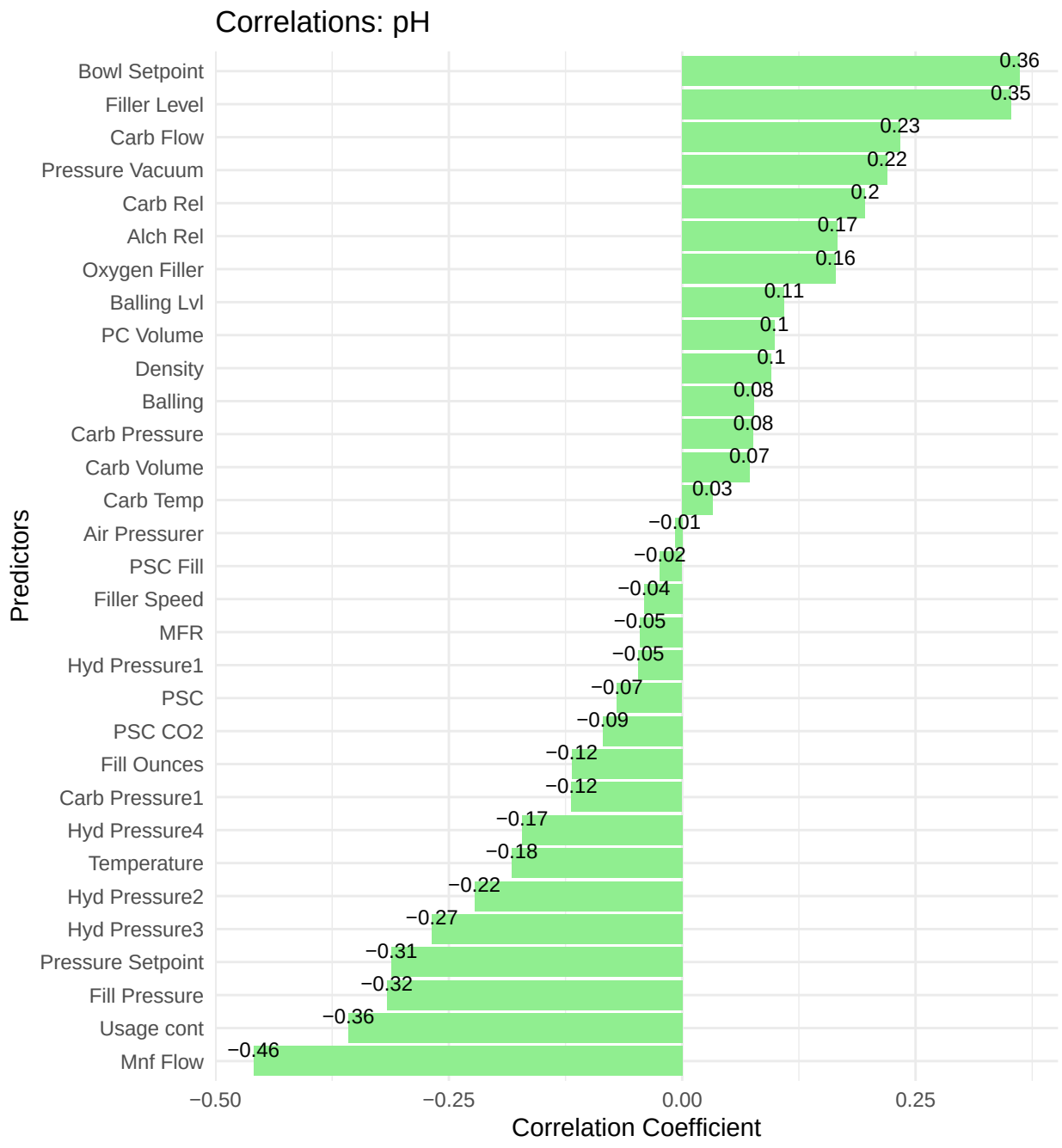
```
## Warning in axis(side = side, at = at, labels = labels, ...): "tl.cex" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "cl.cex" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "addCoefasPercent"
## is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "number.digits" is
## not a graphical parameter

## Warning in box(...): "method" is not a graphical parameter

## Warning in box(...): "tl.col" is not a graphical parameter

## Warning in box(...): "order" is not a graphical parameter

## Warning in box(...): "addCoef.col" is not a graphical parameter

## Warning in box(...): "number.cex" is not a graphical parameter

## Warning in box(...): "tl.cex" is not a graphical parameter

## Warning in box(...): "cl.cex" is not a graphical parameter

## Warning in box(...): "addCoefasPercent" is not a graphical parameter

## Warning in box(...): "number.digits" is not a graphical parameter

## Warning in title(...): "method" is not a graphical parameter

## Warning in title(...): "tl.col" is not a graphical parameter

## Warning in title(...): "order" is not a graphical parameter

## Warning in title(...): "addCoef.col" is not a graphical parameter

## Warning in title(...): "number.cex" is not a graphical parameter

## Warning in title(...): "tl.cex" is not a graphical parameter

## Warning in title(...): "cl.cex" is not a graphical parameter

## Warning in title(...): "addCoefasPercent" is not a graphical parameter

## Warning in title(...): "number.digits" is not a graphical parameter
```


Fill Ounces
2
1
0
−1
−2
−1.0
−0.5
0.0
0.5
1.0
Carb Volume

PH

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

```
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.

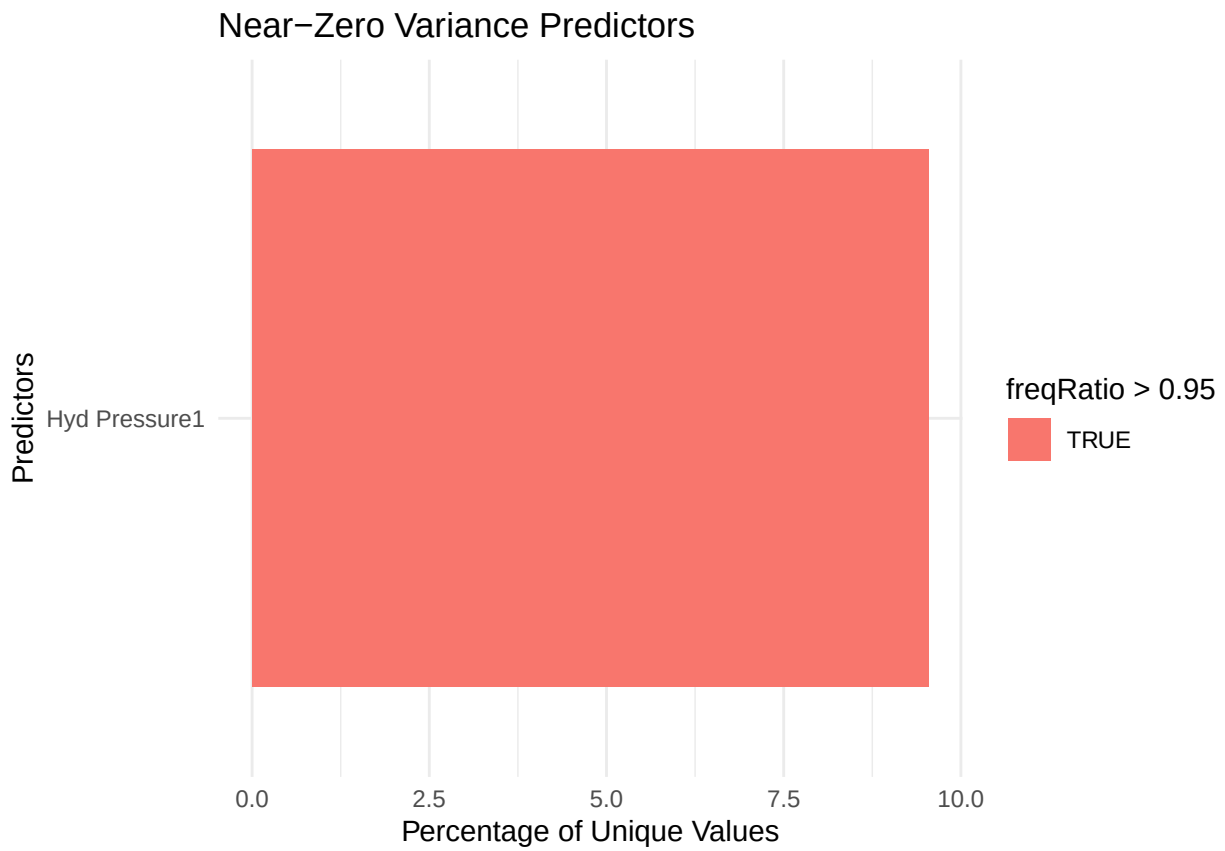
```
train_df%>%
  dplyr::filter(!is.na(PH))
```

```
## # 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

```
train_proc_pls_df<-train_proc_df
eval_proc_pls_df<-eval_proc_df
```

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

```
set.seed(2341)
#generate model
pls_model <- train(y_train, train_proc_pls_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_pls_df$PH)
```

```
plsReSample %>% kable() %>% kable_paper()
```

	x
RMSE	0.1296989
Rsquared	0.3892951
MAE	0.1030896

plsr

```
head(train_proc_df)
```

```
##   Brand Code Carb Volume Fill Ounces  PC Volume Carb Pressure   Carb Temp
## 1          B  -0.2604332 -0.09614345 -0.1976056    0.02866702  0.06388654
## 2          A   0.5557054  0.36141542 -0.6161602    0.08517463 -0.33925947
## 3          B  -0.7827437  0.97268084 -0.1976056    0.75066595  0.92887640
## 4          A   0.6778151  0.36141542  0.2959196   -1.50180156 -2.25347122
## 5          A   1.0982114  3.89471605 -3.0451586   -0.25638005 -1.07441886
## 8          B  -0.4544596  2.27612380 -0.9298223   -0.14185547  0.11344703
##          PSC   PSC Fill    PSC CO2  Mnf Flow Carb Pressure1 Fill Pressure
## 1  0.5175131  0.5492983 -0.3817538 -1.042891     -0.7954714    -0.5774383
## 2  0.8613677  0.2097078 -0.3817538 -1.042891     -0.1997239    -0.5774383
## 3  0.2566553  1.2284792  2.4047915 -1.042891     -0.4970772    -0.5774383
## 4  0.3467262  1.9076602 -0.3817538 -1.042891     -1.5676630    -0.4419470
## 5 -1.3527171 -0.2996779  1.4759431 -1.042891     -0.8809202    -0.6459377
## 8  1.3287015  1.2284792 -0.3817538 -1.042891      0.3497949    -0.3084269
##   Hyd Pressure2 Hyd Pressure4 Temperature Usage cont  Carb Flow        MFR
## 1     -1.281777     1.5550020  0.05864358 -1.5592991  0.3995372  0.3158695
## 2     -1.281777     0.7913786  1.25049580 -0.4450060  0.6313808  0.3495872
## 3     -1.281777    -1.1395149  0.81354700 -1.1140953  0.3800781  0.5042481
## 4     -1.281777    -0.2556205 -0.25302826 -1.2133988  0.5411308  0.4210436
## 5     -1.281777    -0.2556205 -0.25302826 -1.1376335  0.5323643  0.2747725
## 8     -1.281777     2.3274766 -0.57045397 -0.7482698 -1.6730192 -1.2370726
##   Pressure Vacuum Oxygen Filler Bowl Setpoint Pressure Setpoint Air Pressurer
## 1        2.132323   -0.37262498     0.7073527        -0.5589443    -0.1861009
## 2        2.132323   -0.21784950     0.7073527        -0.3481354     0.1532988
## 3        2.482975   -0.29265441     0.7073527        -0.4528613    -0.7005884
## 4        1.431020   -0.08109182     0.7073527        -0.7752764     2.7689863
## 5        1.431020   -0.08109182     0.7073527        -0.7752764     2.7689863
## 8        1.431020    0.35150632     0.7073527        -0.7752764     2.9267953
##     Alch Rel    Carb Rel   PH
## 1 -0.6193417 -0.91493855 8.36
## 2 -0.6688482 -1.08402872 8.26
## 3  1.5054728  2.89008568 8.94
```

```
## 4  0.6024313 -0.09737543 8.24
## 5  0.6024313  0.06075021 8.26
## 8 -0.7692315 -0.74774471 8.38
```

```r
head(eval_proc_df)
```

```
##    Brand Code Carb Volume Fill Ounces   PC Volume Carb Pressure  Carb Temp
## 6           A   0.1211481  -0.5529393 -0.09760493    -0.4294515 -0.6495941
## 7           A  -0.5196224  -1.0089721 -0.11976904    -1.1376901 -1.0744189
## 9           B  -1.1849736   0.0562914 -0.20875871    -1.1376901 -0.1866775
## 10          B  -0.9827131   0.3614154 -0.74306071     1.0749986  1.5196367
## 11          B  -0.4544596  -0.6289978 -0.27585702    -0.5457008 -0.3904998
## 17          C  -1.3895605  -3.2023743  0.66934490    -1.2582988 -0.7019985
##            PSC    PSC Fill     PSC CO2  Mnf Flow Carb Pressure1 Fill Pressure
## 6   0.25665527  0.37950304 -0.38175376 -1.042891    -0.62483205    -0.7149475
## 7   0.92672091  1.73786497 -0.38175376 -1.042891    -0.07260156     1.2111487
## 9   0.99106069 -0.63926840  1.94036726 -1.042891    -0.36951356    -0.5774383
## 10 -1.85057354  0.37950304  0.08267045 -1.042891    -0.58222579    -0.8545226
## 11  0.01638641 -0.12988268 -0.38175376 -1.042891    -0.62483205    -0.3749434
## 17  0.86136772  0.03991256  0.08267045 -1.042891     0.18108302     0.3311381
##    Hyd Pressure2 Hyd Pressure4 Temperature Usage cont  Carb Flow        MFR
## 6      -1.281777    1.43492395  0.21236515  0.9778345  0.4168643  0.5765085
## 7      -1.281777    1.89992556 -0.09648185 -0.1615909 -2.0346046  0.4469659
## 9      -1.281777   -0.42210038 -0.41101308 -0.9219638  0.3671254  0.6070455
## 10     -1.281777    0.92624675  0.51566407 -2.2196711  0.5148522 -0.2803259
## 11     -1.281777   -0.09378017 -0.73136889 -0.7230313  0.5345553  0.3533388
## 17     -1.281777   -0.25562047  2.09034745 -1.3277691  0.6934231 -5.3371304
##    Pressure Vacuum Oxygen Filler Bowl Setpoint Pressure Setpoint Air Pressurer
## 6          1.431020   -0.29265441     0.7073527        -0.7752764      3.083959
## 7          1.431020    0.74681198     0.7073527        -0.7752764      2.768986
## 9          1.431020    0.71199354     0.7073527        -0.7752764      3.551612
## 10         1.431020   -0.37262498     0.7073527        -0.7752764      2.768986
## 11         1.431020   -0.08109182     0.7073527        -0.7752764      2.768986
## 17         1.781671    1.87990602     0.7073527        -0.7752764      3.083959
##      Alch Rel    Carb Rel   PH
## 6   0.6408410  0.06075021 8.32
## 7  -0.7188096 -0.41893354 8.40
## 9  -0.7692315 -0.74774471 8.38
## 10 -0.7188096 -0.74774471 8.50
## 11 -0.7692315 -0.74774471 8.34
## 17 -0.7692315 -0.74774471 8.58
```

missing values

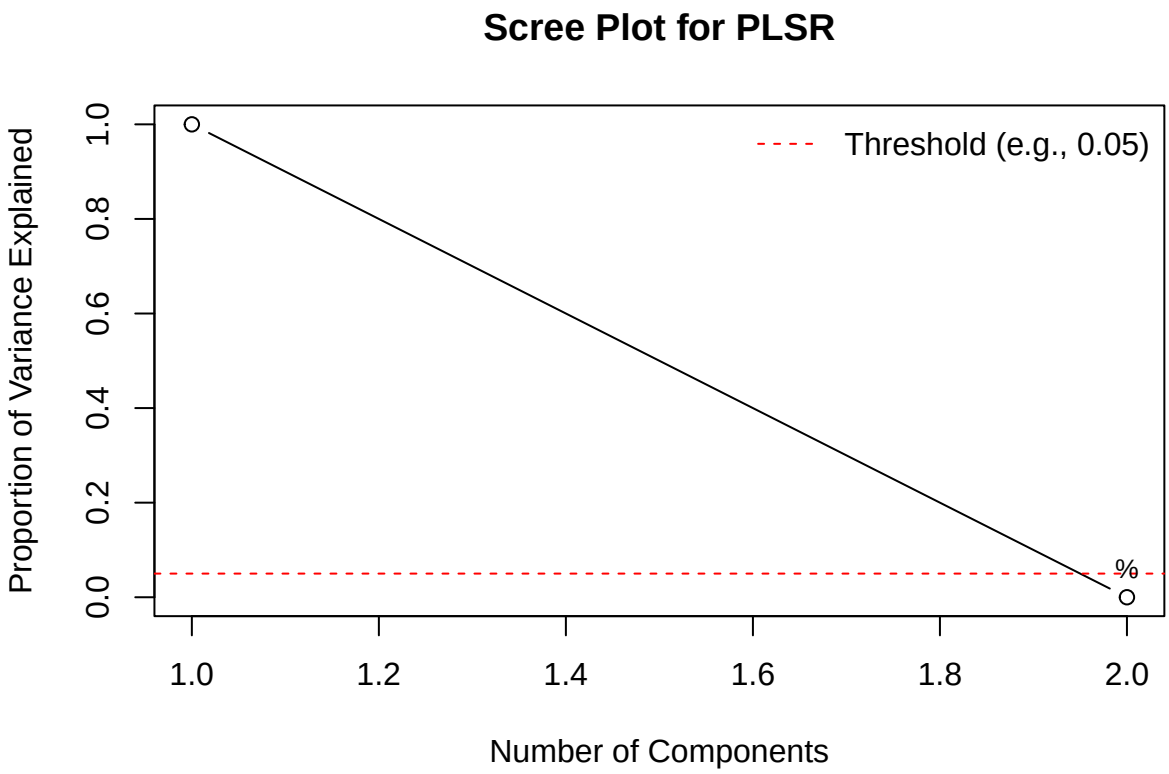
```r
eval_proc_df <- na.omit(eval_proc_df)
train_proc_df <- na.omit(train_proc_df)
```

```r
# Install and load the necessary packages
library(pls)

# Assuming train_proc_df contains your training data
```

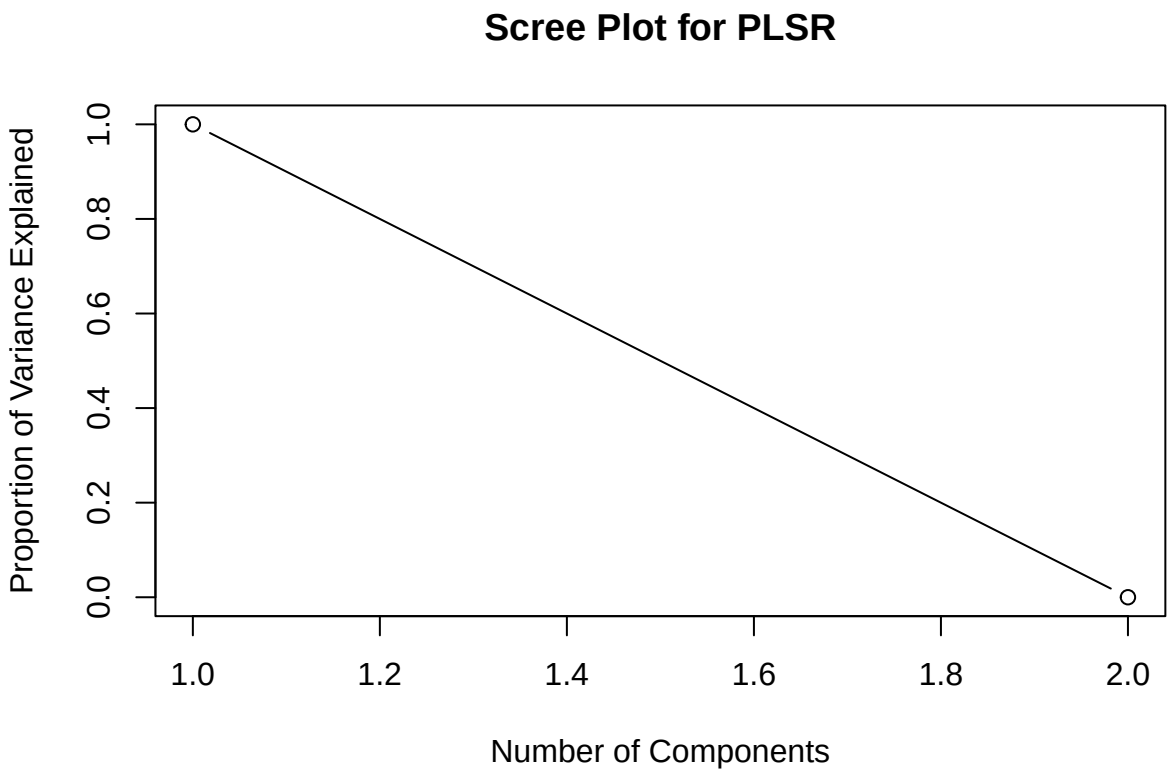
```r
# Fit PLSR model
plsr_model <- plsr(PH ~ ., data = train_proc_df, ncomp = 10)  # Set a reasonable maximum number of comp

# Extract the proportion of variance explained by each component
variance_explained <- summary(plsr_model)$val$prop
```

```
## Data:    X dimension: 1964 28
##  Y dimension: 1964 1
## Fit method: kernelpls
## Number of components considered: 10
## TRAINING: % variance explained
##     1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X      16.8    20.71    26.60     38.6    43.78    48.32    50.88    54.61
## PH     23.7    34.81    37.69     38.6    39.72    40.11    40.33    40.41
##     9 comps  10 comps
## X     59.11     61.99
## PH    40.45     40.49
```

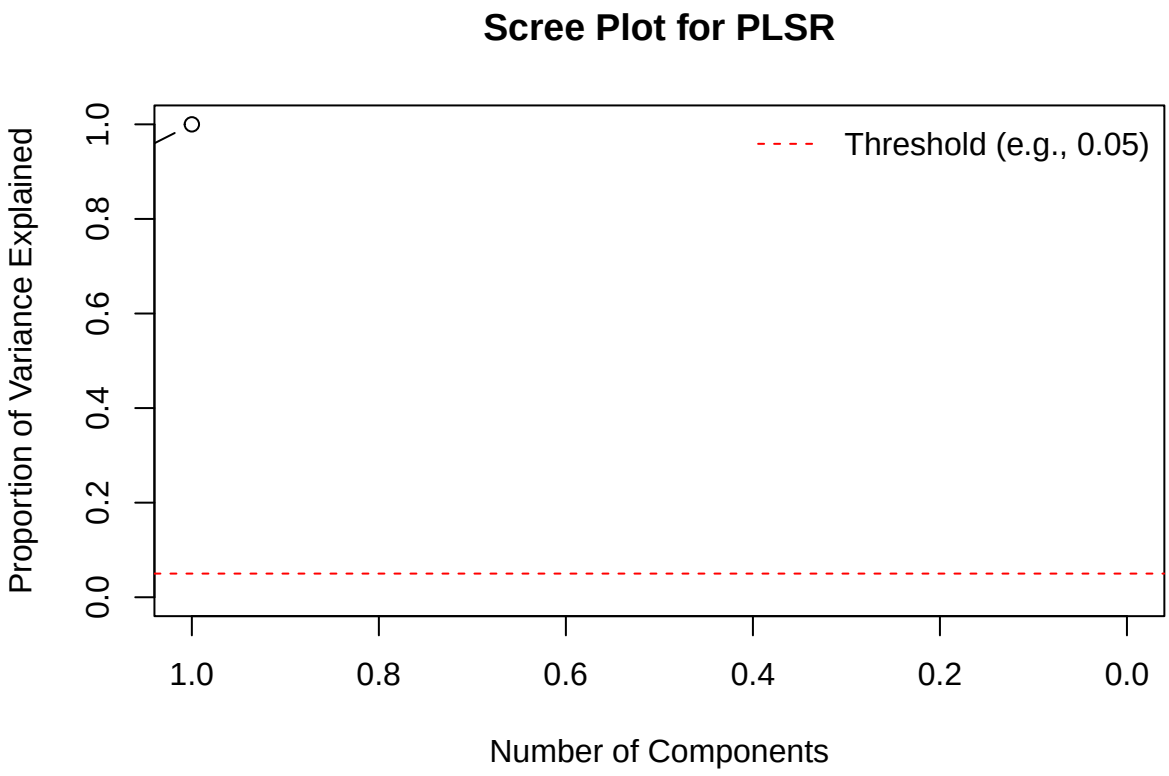
```r
# Create a scree plot
plot(1:length(variance_explained), variance_explained, type = "b",
     xlab = "Number of Components", ylab = "Proportion of Variance Explained",
     main = "Scree Plot for PLSR")

# Add a horizontal line at 0.05 for reference (adjust as needed)
abline(h = 0.05, col = "red", lty = 2)

# Add text indicating the percentage of variance explained by each component
text(1:length(variance_explained), variance_explained,
     labels = paste0(round(variance_explained * 100, 2), "%"),
     pos = 3, cex = 0.8)

# Add a legend
legend("topright", legend = "Threshold (e.g., 0.05)", lty = 2, col = "red", bty = "n")
```




```r
# Create a scree plot for the first few components
plot(1:length(variance_explained), variance_explained, type = "b",
     xlab = "Number of Components", ylab = "Proportion of Variance Explained",
     main = "Scree Plot for PLSR")
```



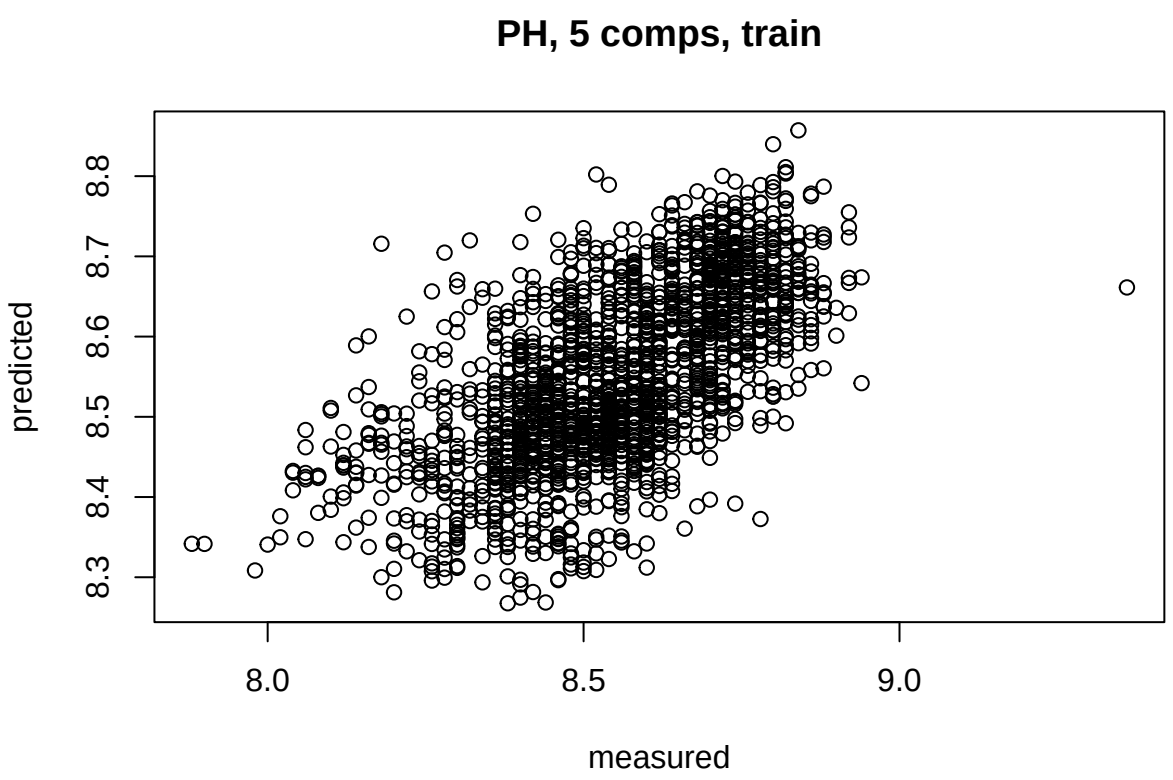

```r
# Zoom in on the first few components (adjust xlim as needed)
xlim <- c(1, min(10, length(variance_explained)))  # Adjust the maximum number of components if needed
plot(1:length(variance_explained), variance_explained, type = "b",
     xlab = "Number of Components", ylab = "Proportion of Variance Explained",
     main = "Scree Plot for PLSR", xlim = xlim)

# Add a horizontal line at 0.05 for reference (adjust as needed)
abline(h = 0.05, col = "red", lty = 2)

# Add text indicating the percentage of variance explained by each component
text(1:length(variance_explained), variance_explained,
     labels = paste0(round(variance_explained * 100, 2), "%"),
     pos = 3, cex = 0.8)

# Add a legend
legend("topright", legend = "Threshold (e.g., 0.05)", lty = 2, col = "red", bty = "n")
```




```r
# Fit PLSR model
plsr_model <- plsr(PH ~ ., data = train_proc_df, ncomp = 5)  # Specify the number of components (e.g., 

summary(plsr_model)
```

```
## Data:    X dimension: 1964 28
##  Y dimension: 1964 1
## Fit method: kernelpls
## Number of components considered: 5
## TRAINING: % variance explained
##    1 comps  2 comps  3 comps  4 comps  5 comps
## X    16.8    20.71    26.60     38.6    43.78
## PH   23.7    34.81    37.69     38.6    39.72
```

```r
# Predict PH values for evaluation/test set
predictions <- predict(plsr_model, newdata = eval_proc_df)

plot(plsr_model)
```



model summary explaination

The summary provided gives information about the Partial Least Squares Regression (PLSR) model that was fitted to your data. Here's an explanation of the key elements:

1. **Data Dimensions**:
   - X dimension: 1964 rows and 28 columns
   - Y dimension: 1964 rows and 1 column
   - This indicates that your dataset has 1964 observations (rows) and 28 predictor variables (X) along with 1 response variable (Y).
2. **Fit Method**:
   - Kernel PLS (Partial Least Squares) was used as the fitting method for the model. Kernel PLS is a variant of PLS that can handle non-linear relationships between predictors and the response variable.
3. **Number of Components Considered**:
   - The model considered up to 5 components in the analysis. Components represent the latent variables extracted by PLS that explain the maximum covariance between the predictor variables (X) and the response variable (Y).
4. **Training: % Variance Explained**:
   - For each number of components (from 1 to 5), the percentage of variance explained by the model in both the predictor variables (X) and the response variable (PH) is provided.
   - For example, with 5 components, the model explains 43.78% of the variance in the predictor variables (X) and 39.72% of the variance in the response variable (PH).

Overall, the summary provides insights into how well the PLSR model captures the variance in the data and how many components are needed to explain a significant portion of the variance. Higher percentages indicate that the model captures more variance in the data, suggesting better predictive performance.

model evaluation

```
sum(is.na(predictions))
```

```
## [1] 0
```

```
# Calculate Mean Squared Error (MSE)
on <- predictions - eval_proc_df$PH
on <- on^2
mse <- mean(on)

mse# Calculate R-squared (R²)
```

```
## [1] 0.01795639
```

```
actual <- eval_proc_df$PH
ss_total <- sum((actual - mean(actual))^2)
ss_residual <- sum((actual - predictions)^2)
r_squared <- 1 - (ss_residual / ss_total)

# Print MSE and R²
cat("Mean Squared Error (MSE):", mse, "\n")
```

```
## Mean Squared Error (MSE): 0.01795639
```

```
cat("R-squared (R²):", r_squared, "\n")
```

```
## R-squared (R²): -2.271855
```

explaination

The Mean Squared Error (MSE) and R-squared (R^2) are two common metrics used to evaluate the performance of regression models like Partial Least Squares Regression (PLSR).

1. **Mean Squared Error (MSE)**:
    - The MSE measures the average squared difference between the predicted values and the actual values.
    - A lower MSE indicates that the model's predictions are closer to the actual values on average.
    - In your case, the MSE value of 0.01795639 suggests that, on average, the squared difference between the predicted PH values and the actual PH values is approximately 0.018.
2. **R-squared (R^2)**:

- The R-squared (R^2) value represents the proportion of variance in the dependent variable (PH) that is explained by the independent variables (predictors) in the model.
- R^2 ranges from 0 to 1, where 1 indicates a perfect fit (the model explains all the variance), and 0 indicates that the model does not explain any of the variance.
- However, R^2 can also be negative, which typically occurs when the model performs worse than a horizontal line (a model that simply predicts the mean of the dependent variable for all observations).
- In your case, the negative R^2 value of -2.271855 suggests that the model performs worse than a horizontal line, indicating poor predictive performance. This could be due to various reasons such as overfitting, multicollinearity among predictors, or the model not capturing the underlying relationships in the data adequately.

Overall, based on these values, it seems that the PLSR model is not performing well in explaining the variance in the dependent variable (PH) and making accurate predictions. Further investigation and potentially model refinement or feature engineering may be necessary to improve the model's performance.

Regression Tree model

```
# Fit regression tree model to training data
tree_model <- rpart(PH ~ ., data = train_proc_df)

# Make predictions on evaluation/test data
tree_predictions <- predict(tree_model, newdata = eval_proc_df)

# Calculate Mean Squared Error (MSE)
tree_mse <- mean((tree_predictions - eval_proc_df$PH)^2)

# Calculate R-squared (R²)
tree_actual <- eval_proc_df$PH
tree_ss_total <- sum((tree_actual - mean(tree_actual))^2)
tree_ss_residual <- sum((tree_actual - tree_predictions)^2)
tree_r_squared <- 1 - (tree_ss_residual / tree_ss_total)

# Print MSE and R²
cat("Regression Tree Model:\n")
```

```
## Regression Tree Model:
```

```
cat("Mean Squared Error (MSE):", tree_mse, "\n")
```

```
## Mean Squared Error (MSE): 0.01510054
```

```
cat("R-squared (R²):", tree_r_squared, "\n")
```

```
## R-squared (R²): 0.4497026
```

```
summary(tree_model)
```

```
## Call:
## rpart(formula = PH ~ ., data = train_proc_df)
##   n= 1964
##
##            CP nsplit rel error    xerror       xstd
## 1  0.21379182      0 1.0000000 1.0021968 0.03351630
## 2  0.07313917      1 0.7862082 0.7885809 0.03107633
## 3  0.03838786      2 0.7130690 0.7172996 0.02871537
## 4  0.03035244      3 0.6746812 0.6796095 0.02727335
## 5  0.01997908      4 0.6443287 0.6498360 0.02704588
## 6  0.01457737      6 0.6043706 0.6325127 0.02657459
## 7  0.01431694      7 0.5897932 0.6218431 0.02609776
## 8  0.01414008      8 0.5754763 0.6155020 0.02601272
## 9  0.01398458     10 0.5471961 0.6044168 0.02548074
## 10 0.01237477     11 0.5332115 0.5924817 0.02517028
## 11 0.01232729     12 0.5208367 0.5806313 0.02514690
## 12 0.01038548     13 0.5085095 0.5633837 0.02398801
## 13 0.01000000     14 0.4981240 0.5572620 0.02413688
##
## Variable importance
##          Mnf Flow     Bowl Setpoint     Oxygen Filler        Usage cont
##                16                13                11                10
##     Hyd Pressure2        Brand Code Pressure Setpoint          Alch Rel
##                10                 9                 8                 5
##          Carb Rel       Carb Volume     Air Pressurer     Hyd Pressure4
##                 4                 3                 2                 2
##               MFR     Fill Pressure       Temperature   Pressure Vacuum
##                 2                 2                 1                 1
##         Carb Flow         PC Volume
##                 1                 1
##
## Node number 1: 1964 observations,    complexity param=0.2137918
##   mean=8.548126, MSE=0.03003946
##   left son=2 (1065 obs) right son=3 (899 obs)
##   Primary splits:
##       Mnf Flow          < -0.6236492  to the right, improve=0.2137918, (0 missing)
##       Usage cont        < 0.7937777   to the right, improve=0.1724400, (0 missing)
##       Bowl Setpoint     < 0.3392728   to the left,  improve=0.1601176, (0 missing)
##       Pressure Setpoint < 0.6056506   to the right, improve=0.1327233, (0 missing)
##       Brand Code        splits as  LRLR-, improve=0.0989594, (0 missing)
##   Surrogate splits:
##       Bowl Setpoint     < 0.3392728   to the left,  agree=0.904, adj=0.790, (0 split)
##       Hyd Pressure2     < -1.275671   to the right, agree=0.836, adj=0.642, (0 split)
##       Usage cont        < 0.8092994   to the right, agree=0.812, adj=0.590, (0 split)
##       Oxygen Filler     < 0.3944159   to the left,  agree=0.804, adj=0.573, (0 split)
##       Pressure Setpoint < 1.059325    to the right, agree=0.778, adj=0.515, (0 split)
##
## Node number 2: 1065 observations,    complexity param=0.03838786
##   mean=8.474498, MSE=0.02087151
##   left son=4 (794 obs) right son=5 (271 obs)
##   Primary splits:
##       Alch Rel      < 1.325189    to the left,  improve=0.10188820, (0 missing)
##       Brand Code    splits as  LLLR-, improve=0.09532440, (0 missing)
##       Carb Rel      < 0.8999047   to the left,  improve=0.06655158, (0 missing)
```

```
##       Usage cont    < 0.7782826   to the right, improve=0.06651554, (0 missing)
##       Hyd Pressure4 < -0.6396263  to the right, improve=0.05679734, (0 missing)
##   Surrogate splits:
##       Brand Code    splits as  LLLR-, agree=0.990, adj=0.959, (0 split)
##       Hyd Pressure4 < -0.6396263  to the right, agree=0.934, adj=0.742, (0 split)
##       Carb Rel      < 0.7510094   to the left,  agree=0.915, adj=0.668, (0 split)
##       Carb Volume   < 0.9492772   to the left,  agree=0.907, adj=0.635, (0 split)
##       Temperature   < -0.975735   to the right, agree=0.788, adj=0.166, (0 split)
##
## Node number 3: 899 observations,    complexity param=0.07313917
##   mean=8.63535, MSE=0.02687004
##   left son=6 (117 obs) right son=7 (782 obs)
##   Primary splits:
##       Brand Code    splits as  RRLR-, improve=0.17863050, (0 missing)
##       Carb Rel      < -0.6650818  to the left,  improve=0.08273612, (0 missing)
##       Temperature   < 0.451824    to the right, improve=0.07131096, (0 missing)
##       Mnf Flow      < -1.043728   to the right, improve=0.06616111, (0 missing)
##       Air Pressurer < 1.562371    to the right, improve=0.05834929, (0 missing)
##   Surrogate splits:
##       Carb Rel    < -0.7693944  to the left,  agree=0.882, adj=0.094, (0 split)
##       Carb Volume < -1.42392    to the left,  agree=0.879, adj=0.068, (0 split)
##       PSC         < -2.596811   to the left,  agree=0.872, adj=0.017, (0 split)
##       PC Volume   < 2.670291    to the right, agree=0.871, adj=0.009, (0 split)
##       PSC CO2     < 4.030276    to the right, agree=0.871, adj=0.009, (0 split)
##
## Node number 4: 794 observations,    complexity param=0.01997908
##   mean=8.447557, MSE=0.02056154
##   left son=8 (561 obs) right son=9 (233 obs)
##   Primary splits:
##       Usage cont    < 0.7435282   to the right, improve=0.07188052, (0 missing)
##       Brand Code    splits as  RRLR-, improve=0.05910629, (0 missing)
##       Air Pressurer < 0.4058977   to the left,  improve=0.05772227, (0 missing)
##       Carb Rel      < -0.6650818  to the right, improve=0.03766888, (0 missing)
##       Carb Volume   < -0.1642649  to the right, improve=0.03593818, (0 missing)
##   Surrogate splits:
##       PC Volume      < 0.5794366   to the left,  agree=0.782, adj=0.258, (0 split)
##       Air Pressurer  < 0.9065399   to the left,  agree=0.780, adj=0.249, (0 split)
##       Carb Pressure1 < -0.475806   to the right, agree=0.775, adj=0.232, (0 split)
##       Mnf Flow       < 0.8190117   to the right, agree=0.764, adj=0.197, (0 split)
##       Bowl Setpoint  < 0.3392728   to the left,  agree=0.757, adj=0.172, (0 split)
##
## Node number 5: 271 observations
##   mean=8.553432, MSE=0.01342254
##
## Node number 6: 117 observations,    complexity param=0.01457737
##   mean=8.456239, MSE=0.0403209
##   left son=12 (94 obs) right son=13 (23 obs)
##   Primary splits:
##       MFR           < -0.143788   to the right, improve=0.18230420, (0 missing)
##       Hyd Pressure2 < -1.275671   to the left,  improve=0.17414250, (0 missing)
##       Oxygen Filler < 1.175577    to the left,  improve=0.14308680, (0 missing)
##       Alch Rel      < -0.6792033  to the right, improve=0.11688260, (0 missing)
##       PC Volume     < 1.306121    to the left,  improve=0.09001194, (0 missing)
##   Surrogate splits:
```

```
##       Oxygen Filler  < 1.519984    to the left,  agree=0.855, adj=0.261, (0 split)
##       Fill Pressure  < -0.2754094  to the left,  agree=0.838, adj=0.174, (0 split)
##       Alch Rel       < -0.4496261  to the left,  agree=0.838, adj=0.174, (0 split)
##       Carb Pressure1 < 1.271921    to the left,  agree=0.829, adj=0.130, (0 split)
##       Carb Flow      < -1.807311   to the right, agree=0.829, adj=0.130, (0 split)
##
## Node number 7: 782 observations,    complexity param=0.03035244
##   mean=8.662148, MSE=0.01933963
##   left son=14 (87 obs) right son=15 (695 obs)
##   Primary splits:
##       Air Pressurer   < 1.562371    to the right, improve=0.11840560, (0 missing)
##       Mnf Flow        < -1.043728   to the right, improve=0.09512144, (0 missing)
##       Pressure Vacuum < 1.255694    to the right, improve=0.07265005, (0 missing)
##       Carb Pressure1  < -0.1361419  to the left,  improve=0.05547662, (0 missing)
##       Bowl Setpoint   < 0.3392728   to the left,  improve=0.04635601, (0 missing)
##   Surrogate splits:
##       PC Volume     < -2.588699   to the left,  agree=0.893, adj=0.034, (0 split)
##       Oxygen Filler < 2.809448    to the right, agree=0.890, adj=0.011, (0 split)
##       Bowl Setpoint < 0.3392728   to the left,  agree=0.890, adj=0.011, (0 split)
##       Alch Rel      < -0.8457995  to the left,  agree=0.890, adj=0.011, (0 split)
##
## Node number 8: 561 observations,    complexity param=0.01414008
##   mean=8.422781, MSE=0.01972703
##   left son=16 (259 obs) right son=17 (302 obs)
##   Primary splits:
##       Bowl Setpoint     < -1.005019   to the left,  improve=0.06742954, (0 missing)
##       Carb Flow         < -1.370701   to the right, improve=0.05103851, (0 missing)
##       Pressure Setpoint < -1.361094   to the right, improve=0.04767844, (0 missing)
##       Air Pressurer     < 0.4058977   to the left,  improve=0.03801666, (0 missing)
##       Brand Code        splits as  RRLR-, improve=0.03625682, (0 missing)
##   Surrogate splits:
##       Oxygen Filler < -0.1205227  to the right, agree=0.847, adj=0.668, (0 split)
##       Carb Flow     < 0.6457554   to the right, agree=0.809, adj=0.587, (0 split)
##       MFR           < 0.1577768   to the left,  agree=0.704, adj=0.359, (0 split)
##       Mnf Flow      < 0.9738216   to the left,  agree=0.674, adj=0.293, (0 split)
##       Alch Rel      < -0.6938289  to the right, agree=0.629, adj=0.197, (0 split)
##
## Node number 9: 233 observations,    complexity param=0.01997908
##   mean=8.50721, MSE=0.01753428
##   left son=18 (25 obs) right son=19 (208 obs)
##   Primary splits:
##       Fill Pressure < -0.6804426  to the left,  improve=0.2897868, (0 missing)
##       Temperature   < 0.2885325   to the right, improve=0.2639579, (0 missing)
##       MFR           < -0.3088805  to the left,  improve=0.1754761, (0 missing)
##       Carb Flow     < 0.6568252   to the left,  improve=0.1455480, (0 missing)
##       Carb Rel      < -0.6650818  to the right, improve=0.1227717, (0 missing)
##   Surrogate splits:
##       Carb Volume   < 1.479062    to the right, agree=0.901, adj=0.08, (0 split)
##       PSC CO2       < 4.030276    to the right, agree=0.901, adj=0.08, (0 split)
##       Usage cont    < -2.380988   to the left,  agree=0.901, adj=0.08, (0 split)
##       Hyd Pressure4 < -1.958041   to the left,  agree=0.897, adj=0.04, (0 split)
##       Carb Rel      < 1.408008    to the right, agree=0.897, adj=0.04, (0 split)
##
## Node number 12: 94 observations,    complexity param=0.01232729
```

```
##   mean=8.41383, MSE=0.02845129
##   left son=24 (65 obs) right son=25 (29 obs)
##   Primary splits:
##       Alch Rel        < -0.6938289  to the right, improve=0.27193900, (0 missing)
##       Oxygen Filler   < 0.7463949   to the left,  improve=0.16591830, (0 missing)
##       PC Volume       < 0.1821734   to the right, improve=0.15253260, (0 missing)
##       Pressure Vacuum < -0.1469131  to the right, improve=0.08770621, (0 missing)
##       Temperature     < 0.5904693   to the right, improve=0.07315877, (0 missing)
##   Surrogate splits:
##       PC Volume   < -0.5989576  to the right, agree=0.734, adj=0.138, (0 split)
##       Fill Ounces < 2.00727     to the left,  agree=0.723, adj=0.103, (0 split)
##       Carb Flow   < 0.6778836   to the left,  agree=0.723, adj=0.103, (0 split)
##       Carb Volume < -2.709161   to the right, agree=0.713, adj=0.069, (0 split)
##       PSC Fill    < -1.233552   to the right, agree=0.713, adj=0.069, (0 split)
##
## Node number 13: 23 observations
##   mean=8.629565, MSE=0.05143894
##
## Node number 14: 87 observations
##   mean=8.526897, MSE=0.01628347
##
## Node number 15: 695 observations,    complexity param=0.01431694
##   mean=8.679079, MSE=0.01714563
##   left son=30 (60 obs) right son=31 (635 obs)
##   Primary splits:
##       Oxygen Filler   < 1.658362    to the right, improve=0.07088360, (0 missing)
##       Pressure Vacuum < -0.4975648  to the left,  improve=0.06363751, (0 missing)
##       Mnf Flow        < -1.043728   to the right, improve=0.05199326, (0 missing)
##       Carb Pressure1  < -0.1361419  to the left,  improve=0.04694552, (0 missing)
##       Brand Code      splits as  LR-R-, improve=0.04028876, (0 missing)
##   Surrogate splits:
##       Pressure Vacuum < -1.54952    to the left,  agree=0.915, adj=0.017, (0 split)
##
## Node number 16: 259 observations,    complexity param=0.01414008
##   mean=8.383398, MSE=0.01841934
##   left son=32 (122 obs) right son=33 (137 obs)
##   Primary splits:
##       Carb Rel      < -0.6650818  to the right, improve=0.1933139, (0 missing)
##       Oxygen Filler < -0.225209   to the left,  improve=0.1845186, (0 missing)
##       Air Pressurer < -0.1010727  to the left,  improve=0.1311222, (0 missing)
##       Mnf Flow      < 0.9529013   to the right, improve=0.1009128, (0 missing)
##       Bowl Setpoint < -1.581144   to the right, improve=0.0962006, (0 missing)
##   Surrogate splits:
##       Carb Volume < -0.03683637 to the right, agree=0.761, adj=0.492, (0 split)
##       Brand Code  splits as  LRLR-, agree=0.737, adj=0.443, (0 split)
##       Alch Rel    < -0.5948131  to the right, agree=0.718, adj=0.402, (0 split)
##       Temperature < -0.01891914 to the right, agree=0.645, adj=0.246, (0 split)
##       MFR         < 0.08770308  to the right, agree=0.637, adj=0.230, (0 split)
##
## Node number 17: 302 observations,    complexity param=0.01398458
##   mean=8.456556, MSE=0.01837754
##   left son=34 (86 obs) right son=35 (216 obs)
##   Primary splits:
##       Brand Code        splits as  RRLR-, improve=0.14865820, (0 missing)
```

```
##       Carb Rel          < -1.691452   to the left,  improve=0.09754918, (0 missing)
##       Carb Flow         < -1.295901   to the right, improve=0.07524696, (0 missing)
##       Air Pressurer     < 0.4058977   to the left,  improve=0.06347581, (0 missing)
##       Pressure Setpoint < 0.1925251   to the right, improve=0.05556516, (0 missing)
##   Surrogate splits:
##       Alch Rel        < -0.7946755  to the left,  agree=0.825, adj=0.384, (0 split)
##       Pressure Vacuum < -1.54952    to the left,  agree=0.755, adj=0.140, (0 split)
##       Hyd Pressure4   < 0.7961838   to the right, agree=0.752, adj=0.128, (0 split)
##       Hyd Pressure2   < 0.6170547   to the left,  agree=0.745, adj=0.105, (0 split)
##       Oxygen Filler   < -0.06504574 to the right, agree=0.745, adj=0.105, (0 split)
##
## Node number 18: 25 observations
##   mean=8.3016, MSE=0.00604544
##
## Node number 19: 208 observations
##   mean=8.531923, MSE=0.01322322
##
## Node number 24: 65 observations
##   mean=8.355077, MSE=0.02903422
##
## Node number 25: 29 observations
##   mean=8.545517, MSE=0.002066112
##
## Node number 30: 60 observations
##   mean=8.565667, MSE=0.02396122
##
## Node number 31: 635 observations,    complexity param=0.01038548
##   mean=8.689795, MSE=0.01517145
##   left son=62 (84 obs) right son=63 (551 obs)
##   Primary splits:
##       Pressure Vacuum < -0.4975648  to the left,  improve=0.06360031, (0 missing)
##       Mnf Flow        < -1.043728   to the right, improve=0.04983939, (0 missing)
##       Carb Volume     < 0.01072681  to the right, improve=0.04391202, (0 missing)
##       Carb Pressure1  < -0.1361419  to the left,  improve=0.04325966, (0 missing)
##       Carb Rel        < -0.01831261 to the right, improve=0.04296771, (0 missing)
##   Surrogate splits:
##       Bowl Setpoint < 0.3392728   to the left,  agree=0.883, adj=0.119, (0 split)
##       Hyd Pressure2 < 1.081078    to the right, agree=0.874, adj=0.048, (0 split)
##       Fill Pressure < 3.073191    to the right, agree=0.871, adj=0.024, (0 split)
##       Usage cont    < -2.244187   to the left,  agree=0.869, adj=0.012, (0 split)
##
## Node number 32: 122 observations
##   mean=8.320164, MSE=0.01960981
##
## Node number 33: 137 observations
##   mean=8.439708, MSE=0.01062765
##
## Node number 34: 86 observations,    complexity param=0.01237477
##   mean=8.373721, MSE=0.02780708
##   left son=68 (13 obs) right son=69 (73 obs)
##   Primary splits:
##       Carb Rel    < -1.515491   to the left,  improve=0.3052931, (0 missing)
##       Mnf Flow    < 1.129468    to the right, improve=0.2848475, (0 missing)
##       Temperature < -0.4907335  to the left,  improve=0.2696386, (0 missing)
```

```
##       MFR         < 0.3737364   to the right, improve=0.2015336, (0 missing)
##       Usage cont  < 1.105252    to the right, improve=0.1798218, (0 missing)
##   Surrogate splits:
##       Temperature < -0.6509114  to the left,  agree=0.907, adj=0.385, (0 split)
##       Carb Volume < -1.287007   to the left,  agree=0.895, adj=0.308, (0 split)
##       Mnf Flow    < 1.13951     to the right, agree=0.872, adj=0.154, (0 split)
##       MFR         < 0.4777202   to the right, agree=0.872, adj=0.154, (0 split)
##
## Node number 35: 216 observations
##   mean=8.489537, MSE=0.01080349
##
## Node number 62: 84 observations
##   mean=8.610238, MSE=0.009128515
##
## Node number 63: 551 observations
##   mean=8.701924, MSE=0.01498069
##
## Node number 68: 13 observations
##   mean=8.155385, MSE=0.02154793
##
## Node number 69: 73 observations
##   mean=8.412603, MSE=0.01892062
```

1. **Mean Squared Error (MSE)**: This metric represents the average squared difference between the actual and predicted values of the target variable (PH, in this case). A lower MSE indicates better model performance in terms of prediction accuracy. In your case, the MSE of 0.0151 suggests that, on average, the squared difference between the actual and predicted PH values is relatively low, indicating a reasonably good fit of the regression tree model to the data.

2. **R-squared (R^2)**: This metric measures the proportion of the variance in the target variable that is explained by the independent variables in the model. An R-squared value closer to 1 indicates that a larger proportion of the variance in the target variable is explained by the model, suggesting a better fit. Your R-squared value of 0.4497 indicates that the regression tree model explains approximately 45% of the variance in the PH values. While this value is moderate, it suggests that there is still room for improvement in capturing the variability of the target variable.

Overall, based on these metrics, the regression tree model appears to provide a reasonably good fit to the data, with a relatively low MSE and a moderate level of explained variance (R-squared). However, further analysis and possibly model refinement may be beneficial to improve predictive accuracy and capture more of the variability in the PH values. 1. **Call**: It shows the call that was used to fit the regression tree model, indicating the formula and the dataset.

2. **Complexity Parameter (CP)**: The complexity parameter is used to control the size of the tree. A larger CP results in a smaller tree, which helps prevent overfitting. The CP values in each node represent the cost complexity of that node. As the tree grows, the CP increases.

3. **Variable Importance**: This section shows the importance of each predictor variable in the model. It indicates how much each variable contributes to the decision-making process in the tree.

4. **Node Summary**: Each node in the tree is summarized, showing the number of observations, the mean value of the response variable (PH), and the mean squared error (MSE) associated with that node.

5. **Primary Splits**: These are the variables and values used to split the data at each node. The "improve" value indicates how much the split improves the model's performance.

6. **Surrogate Splits**: Surrogate splits are alternative splits used when the primary split is missing. These splits provide backup options for making decisions.

7. **Mean and MSE for Terminal Nodes**: The terminal nodes (leaf nodes) represent the final segments of the tree where predictions are made. For each terminal node, the mean value of the response variable (PH) and the mean squared error (MSE) are provided.

This summary helps interpret the structure of the regression tree model, showing how the data is split based on different predictor variables and providing insights into the predictive performance of the model at different segments.

model Rsquared values

1. **Mean Squared Error (MSE)**: This metric represents the average squared difference between the actual and predicted values of the target variable (PH, in this case). A lower MSE indicates better model performance in terms of prediction accuracy. In your case, the MSE of 0.0151 suggests that, on average, the squared difference between the actual and predicted PH values is relatively low, indicating a reasonably good fit of the regression tree model to the data.

2. **R-squared (R^2)**: This metric measures the proportion of the variance in the target variable that is explained by the independent variables in the model. An R-squared value closer to 1 indicates that a larger proportion of the variance in the target variable is explained by the model, suggesting a better fit. Your R-squared value of 0.4497 indicates that the regression tree model explains approximately 45% of the variance in the PH values. While this value is moderate, it suggests that there is still room for improvement in capturing the variability of the target variable.

Overall, based on these metrics, the regression tree model appears to provide a reasonably good fit to the data, with a relatively low MSE and a moderate level of explained variance (R-squared). However, further analysis and possibly model refinement may be beneficial to improve predictive accuracy and capture more of the variability in the PH values.

```
# Plot the regression tree model
rpart.plot(tree_model, main = "Regression Tree Model")
```

Regression Tree Model