Data624 - Project 2

Esteban Aramayo, Coffy Andrews-Guo, LeTicia Cancel, Joseph Connolly, Ian Costello

7/16/2022

Table of Contents

# Project Summary

For this project, we are tasked to predict the “potential of hydrogen” more commonly known as pH in a range of beverages. Two data sets are provided one for modeling and the other to use for predictions. The modeling data set contains 33 different features, including the target variable (PH). The predictor variables are all numeric except for Brand Code, which has a categorical.

## Definition of pH

The pH of a solution is meant to describe its acidity or alkalinity (basic). The pH scale ranges from zero to fourteen, where values closer to zero are extremely acidic and usually quite dangerous if handled improperly. Values closer to fourteen are more basic, that is alkaline, and are **also** quite dangerous if handled improperly. Values in the middle of the scale usually defined as “around seven” are so-called neutral. Pure water, milk, sea water, and human saliva are often examples of neutral substances though can tend slightly either alkaline or acidic.

## Libraries Used

* Standard libraries applied (e.g., tidyverse, psych, ggplot2)
* Portions of the data exploration where assisted by the DataExplorer package
* Of particular use throughout the project is the caret package that provides for flexible pre-processing, modeling options, and tuning controls.

### Code Reproducibility / Repeatability

The reproducibility and repeatability of the analysis is important on building upon findings or providing comparison results. The datasets are structured and the machine learning models (classification and regression), will require a broad approach on code organizing, code reporting, and many other features. Learning best practices for writing reproducible code is an iterative learning process on repeating steps to troubleshoot, modify, and gain insights for accurate results. The set.seed function is set at 1234 to provide exact conditions to reproduce these processes.

## Data Exploration

The variable Brand Code is excluded as it is the only categorical variable which would require different operations (such as creating dummy variables) than the numeric columns. It also would seem that Brand Code would have little impact on pH of specific products.

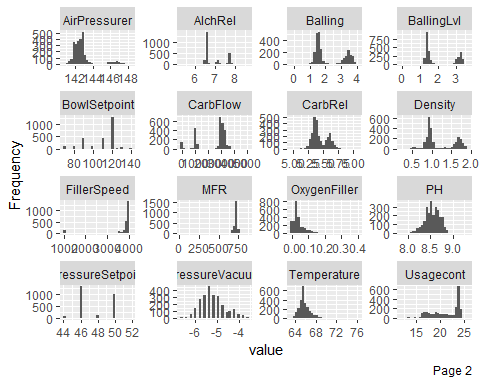
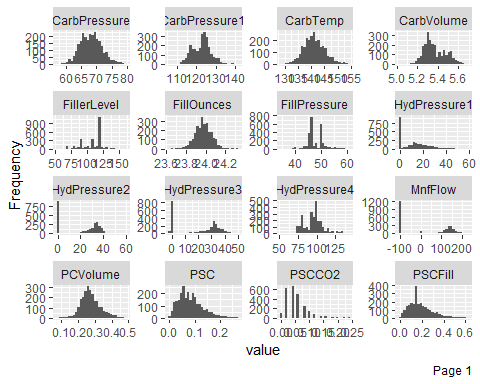
Using describe on the remaining features provides the first glimpse of the shape of these variables. Of note are the skew and kurtosis values, which would at first blush generally indicate (or not) a normal distribution.

#> vars n mean sd median trimmed mad min  
#> CarbVolume 1 2561 5.37 0.11 5.35 5.37 0.11 5.04  
#> FillOunces 2 2533 23.97 0.09 23.97 23.98 0.08 23.63  
#> PCVolume 3 2532 0.28 0.06 0.27 0.27 0.05 0.08  
#> CarbPressure 4 2544 68.19 3.54 68.20 68.12 3.56 57.00  
#> CarbTemp 5 2545 141.09 4.04 140.80 140.99 3.85 128.60  
#> PSC 6 2538 0.08 0.05 0.08 0.08 0.05 0.00  
#> PSCFill 7 2548 0.20 0.12 0.18 0.18 0.12 0.00  
#> PSCCO2 8 2532 0.06 0.04 0.04 0.05 0.03 0.00  
#> MnfFlow 9 2569 24.57 119.48 65.20 21.07 169.02 -100.20  
#> CarbPressure1 10 2539 122.59 4.74 123.20 122.54 4.45 105.60  
#> FillPressure 11 2549 47.92 3.18 46.40 47.71 2.37 34.60  
#> HydPressure1 12 2560 12.44 12.43 11.40 10.84 16.90 -0.80  
#> HydPressure2 13 2556 20.96 16.39 28.60 21.05 13.34 0.00  
#> HydPressure3 14 2556 20.46 15.98 27.60 20.51 13.94 -1.20  
#> HydPressure4 15 2541 96.29 13.12 96.00 95.45 11.86 52.00  
#> FillerLevel 16 2551 109.25 15.70 118.40 111.04 9.19 55.80  
#> FillerSpeed 17 2514 3687.20 770.82 3982.00 3919.99 47.44 998.00  
#> Temperature 18 2557 65.97 1.38 65.60 65.80 0.89 63.60  
#> Usagecont 19 2566 20.99 2.98 21.79 21.25 3.19 12.08  
#> CarbFlow 20 2569 2468.35 1073.70 3028.00 2601.14 326.17 26.00  
#> Density 21 2570 1.17 0.38 0.98 1.15 0.15 0.24  
#> MFR 22 2359 704.05 73.90 724.00 718.16 15.42 31.40  
#> Balling 23 2570 2.20 0.93 1.65 2.13 0.37 -0.17  
#> PressureVacuum 24 2571 -5.22 0.57 -5.40 -5.25 0.59 -6.60  
#> PH 25 2567 8.55 0.17 8.54 8.55 0.18 7.88  
#> OxygenFiller 26 2559 0.05 0.05 0.03 0.04 0.02 0.00  
#> BowlSetpoint 27 2569 109.33 15.30 120.00 111.35 0.00 70.00  
#> PressureSetpoint 28 2559 47.62 2.04 46.00 47.60 0.00 44.00  
#> AirPressurer 29 2571 142.83 1.21 142.60 142.58 0.59 140.80  
#> AlchRel 30 2562 6.90 0.51 6.56 6.84 0.06 5.28  
#> CarbRel 31 2561 5.44 0.13 5.40 5.43 0.12 4.96  
#> BallingLvl 32 2570 2.05 0.87 1.48 1.98 0.21 0.00  
#> max range skew kurtosis se  
#> CarbVolume 5.70 0.66 0.39 -0.47 0.00  
#> FillOunces 24.32 0.69 -0.02 0.86 0.00  
#> PCVolume 0.48 0.40 0.34 0.67 0.00  
#> CarbPressure 79.40 22.40 0.18 -0.01 0.07  
#> CarbTemp 154.00 25.40 0.25 0.24 0.08  
#> PSC 0.27 0.27 0.85 0.65 0.00  
#> PSCFill 0.62 0.62 0.93 0.77 0.00  
#> PSCCO2 0.24 0.24 1.73 3.73 0.00  
#> MnfFlow 229.40 329.60 0.00 -1.87 2.36  
#> CarbPressure1 140.20 34.60 0.05 0.14 0.09  
#> FillPressure 60.40 25.80 0.55 1.41 0.06  
#> HydPressure1 58.00 58.80 0.78 -0.14 0.25  
#> HydPressure2 59.40 59.40 -0.30 -1.56 0.32  
#> HydPressure3 50.00 51.20 -0.32 -1.57 0.32  
#> HydPressure4 142.00 90.00 0.55 0.63 0.26  
#> FillerLevel 161.20 105.40 -0.85 0.05 0.31  
#> FillerSpeed 4030.00 3032.00 -2.87 6.71 15.37  
#> Temperature 76.20 12.60 2.39 10.16 0.03  
#> Usagecont 25.90 13.82 -0.54 -1.02 0.06  
#> CarbFlow 5104.00 5078.00 -0.99 -0.58 21.18  
#> Density 1.92 1.68 0.53 -1.20 0.01  
#> MFR 868.60 837.20 -5.09 30.46 1.52  
#> Balling 4.01 4.18 0.59 -1.39 0.02  
#> PressureVacuum -3.60 3.00 0.53 -0.03 0.01  
#> PH 9.36 1.48 -0.29 0.06 0.00  
#> OxygenFiller 0.40 0.40 2.66 11.09 0.00  
#> BowlSetpoint 140.00 70.00 -0.97 -0.06 0.30  
#> PressureSetpoint 52.00 8.00 0.20 -1.60 0.04  
#> AirPressurer 148.20 7.40 2.25 4.73 0.02  
#> AlchRel 8.62 3.34 0.88 -0.85 0.01  
#> CarbRel 6.06 1.10 0.50 -0.29 0.00  
#> BallingLvl 3.66 3.66 0.59 -1.49 0.02

## Data Visualizations

### Feature Histograms

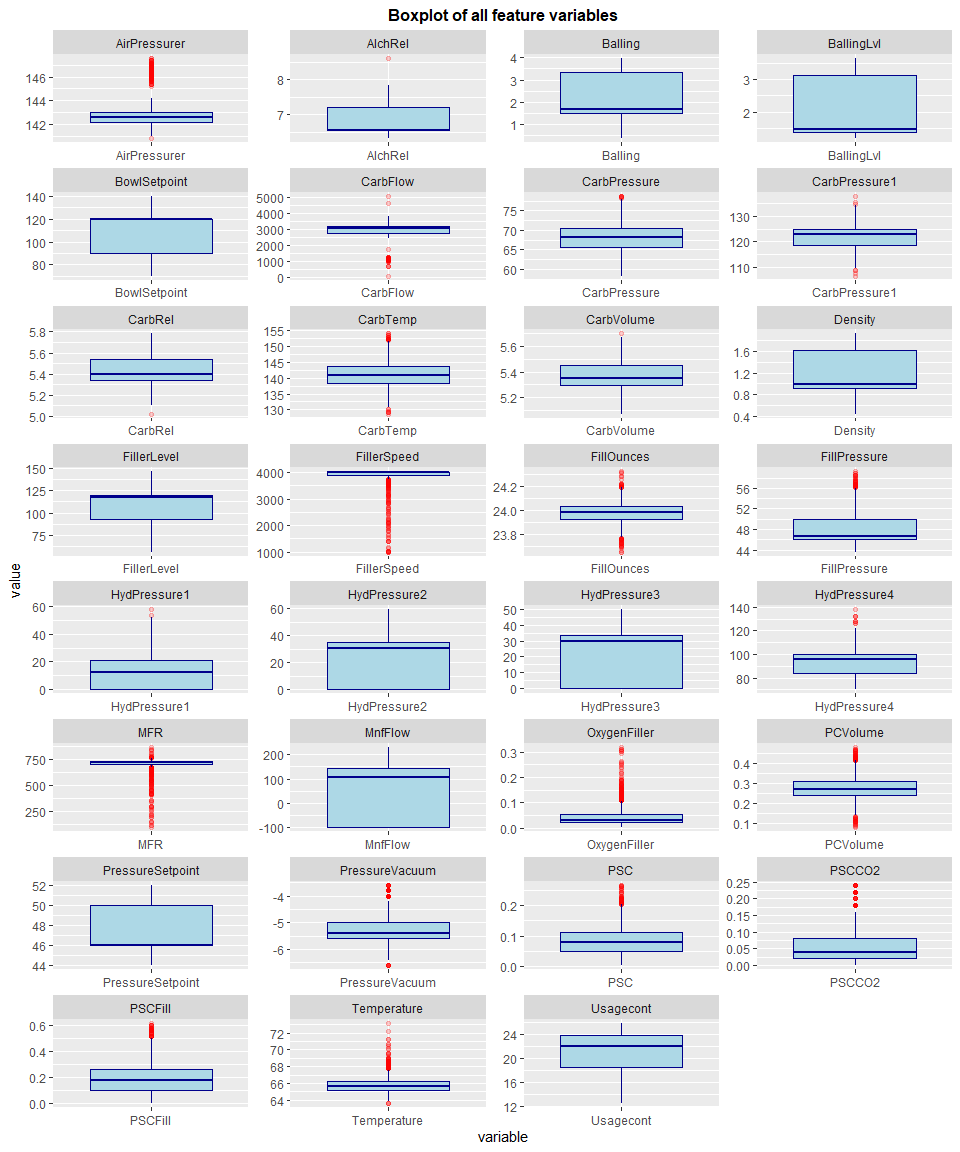
Variable histograms below better illustrate the shape of the feature distributions. The PH distribution appears to be normal, this is a requirement for linear regressions. For the other variables, the distributions are a mix. Some appear quite normal, such as Carb Pressure, Carb Temp, Fill Ounces. Others are bimodel, that is having two “peaks” in the distributions such as Balling Lvl, Density, Carb Rel. The third category of variable are those that may or may not have rather normal distributions but also have significant outliers such as Hyd Pressure, Filler Speed.



### Box Plots and Outliers

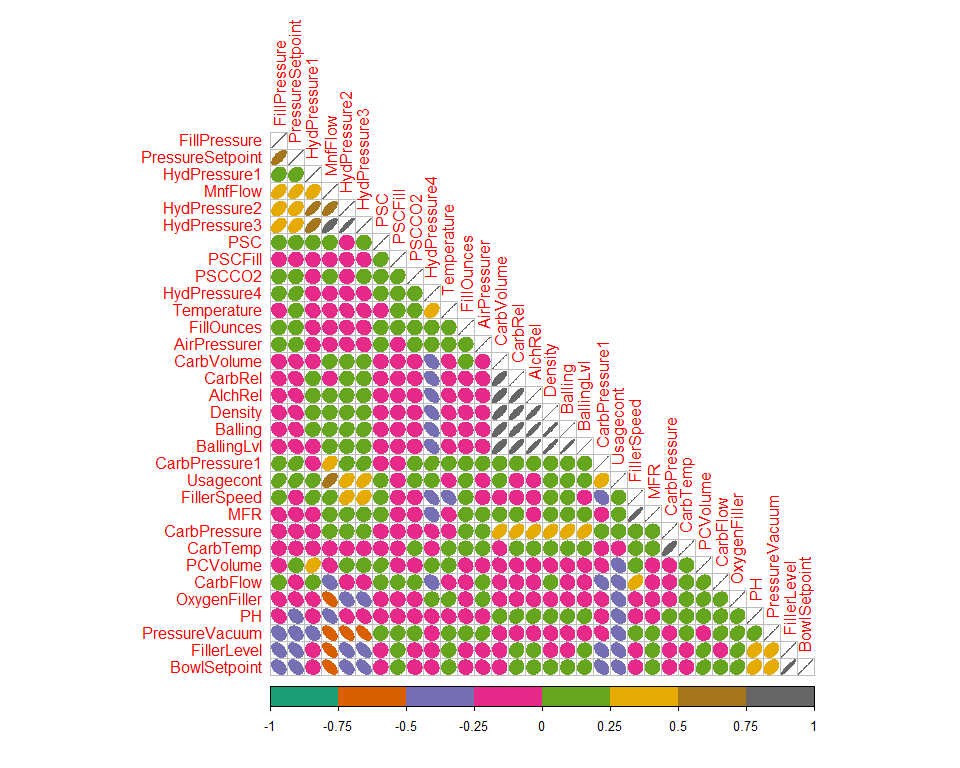
Using box plots will provide more insight into the outliers of these features. Deciding whether or not to smooth, delete, or impute outliers is a critical step in preparing data for a model. Based on these data, it is likely that the test set will also have outliers and so it may be more advantageous to keep them as is to train the model when it encounters them in the test set.

Perhaps it is better to focus on including important features, or features with more normal distributions than treating specific features for their outliers.



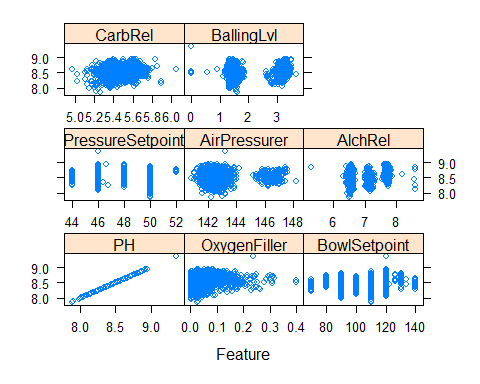
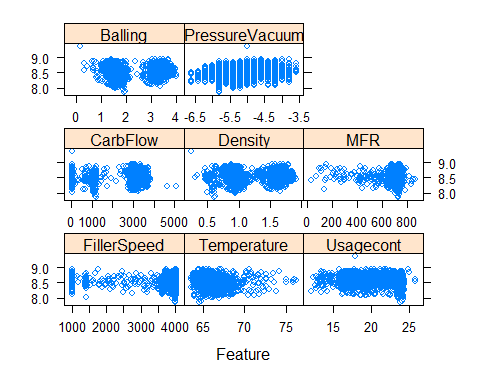
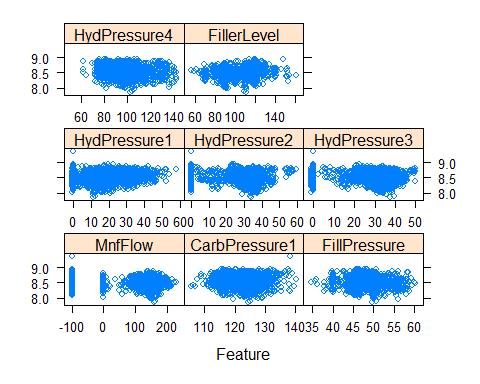
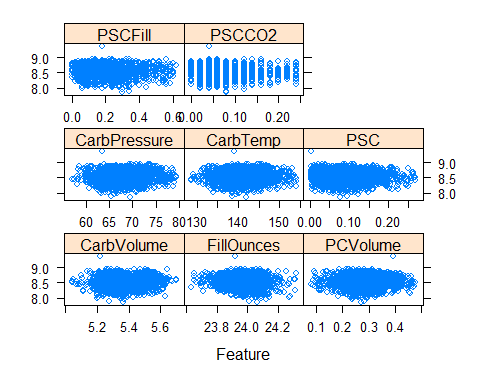
### Correlation

Because of the amount of features in this data set, multicollinearity is an issue. Below, the correlation plot indicates a few features that interact with each other and the target variable. Including all these features will negatively impact the model as they are interrelated. Later the variables with correlation values over 0.75 will be removed.



### Relationship to Target

Another way to analyze and select features is understanding the relationship between the predictor and target variables. From these graphs, one cannot see particularly strong relationships by these graphs alone. Filter Level and Pressure Vacuum are slight exceptions with a hint of a positive relationship. Given the information from these graphs, the bimodal distributed variables should likely be removed.



# Data Preparation

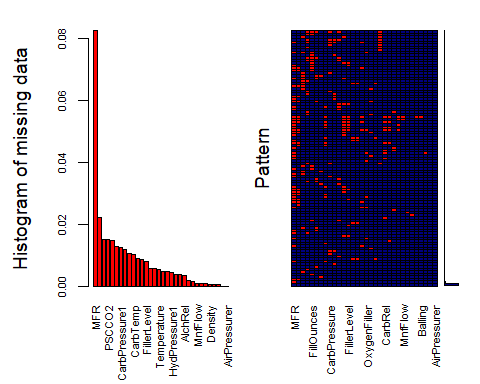
### Near-Zero Variance

Checking for variables with a near-zero variance that will not be valuable for use in the predictions, HydPressure1 can be removed.

#> [1] "HydPressure1"

### Missing Data and Imputation

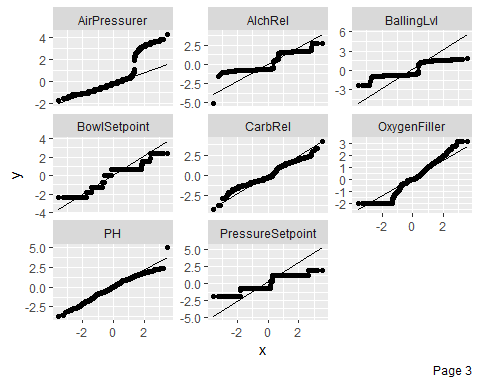
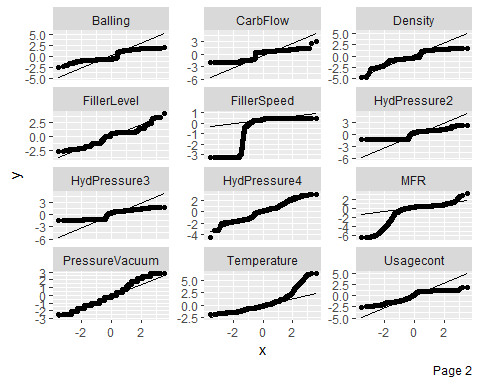
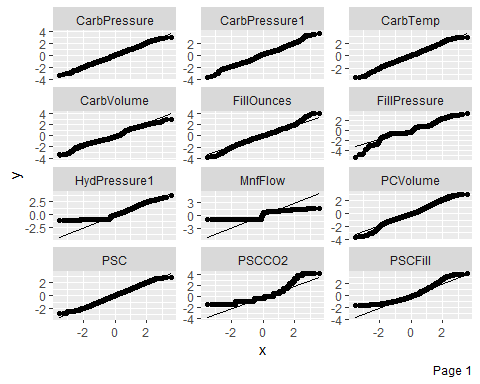
Making good use of the caret package and the preProcess() function, the data sets are scaled and centered, transformed via BoxCox, and missing values imputted using knn imputation process is a single steps and line of code.



#>   
#> Variables sorted by number of missings:   
#> Variable Count  
#> MFR 0.0824581875  
#> FillerSpeed 0.0221703617  
#> PCVolume 0.0151691949  
#> PSCCO2 0.0151691949  
#> FillOunces 0.0147802412  
#> PSC 0.0128354726  
#> CarbPressure1 0.0124465189  
#> HydPressure4 0.0116686114  
#> CarbPressure 0.0105017503  
#> CarbTemp 0.0101127966  
#> PSCFill 0.0089459354  
#> FillPressure 0.0085569817  
#> FillerLevel 0.0077790743  
#> HydPressure2 0.0058343057  
#> HydPressure3 0.0058343057  
#> Temperature 0.0054453520  
#> OxygenFiller 0.0046674446  
#> PressureSetpoint 0.0046674446  
#> HydPressure1 0.0042784909  
#> CarbVolume 0.0038895371  
#> CarbRel 0.0038895371  
#> AlchRel 0.0035005834  
#> Usagecont 0.0019447686  
#> PH 0.0015558149  
#> MnfFlow 0.0007779074  
#> CarbFlow 0.0007779074  
#> BowlSetpoint 0.0007779074  
#> Density 0.0003889537  
#> Balling 0.0003889537  
#> BallingLvl 0.0003889537  
#> PressureVacuum 0.0000000000  
#> AirPressurer 0.0000000000

## Feature Q-Q Plots

The final visualization used to determine features to model are the ‘QQ’ plots. More normal distributions will have plotted values along the diagonal line and are on balance more useful for predictions. There are some variables that have near-horizontal plots that will not provide much insight when used for predictions.



#> CarbVolume FillOunces PCVolume CarbPressure CarbTemp PSC  
#> 1 -0.2591753 -0.09421399 -0.19507352 0.02886039 0.06465558 0.5187053  
#> 2 0.5570433 0.36301548 -0.61278463 0.08537553 -0.33842374 0.8621309  
#> 3 -0.7815371 0.97384081 -0.19507352 0.75095554 0.92792831 0.2581731  
#> 4 0.6791650 0.36301548 0.29745709 -1.50181215 -2.25840822 0.4534529  
#> 5 1.0996025 3.89377242 -3.03688810 -0.25622467 -1.07462875 -1.3491908  
#> 6 0.1224434 -0.55068100 -0.09527442 -0.42931917 -0.64901770 0.2581731  
#> PSCFill PSCCO2 MnfFlow CarbPressure1 FillPressure HydPressure1  
#> 1 0.5487360 -0.3813757 -1.042583 -0.7955317 -0.5774383 -1.000348  
#> 2 0.2091248 -0.3813757 -1.042583 -0.2005518 -0.5774383 -1.000348  
#> 3 1.2279585 2.4068150 -1.042583 -0.4973487 -0.5774383 -1.000348  
#> 4 1.9071809 -0.3813757 -1.042583 -1.5688144 -0.4419470 -1.000348  
#> 5 -0.3002920 1.4774181 -1.042583 -0.8809846 -0.6459377 -1.000348  
#> 6 0.3789304 -0.3813757 -1.042583 -0.6249704 -0.7149475 -1.000348  
#> HydPressure2 HydPressure3 HydPressure4 FillerLevel FillerSpeed Temperature  
#> 1 -1.27918 -1.280596 1.5780323 0.7753167 0.4567709 0.05658733  
#> 2 -1.27918 -1.280596 0.7813743 0.5825000 0.4248112 1.24545800  
#> 3 -1.27918 -1.280596 -1.1256324 0.6858048 0.4928786 0.80960228  
#> 4 -1.27918 -1.280596 -0.2708563 0.5240130 0.4768107 -0.25430483  
#> 5 -1.27918 -1.280596 -0.2708563 0.5825000 0.4727988 -0.25430483  
#> 6 -1.27918 -1.280596 1.4510488 0.7006616 0.4808247 0.20992436  
#> Usagecont CarbFlow Density MFR Balling PressureVacuum  
#> 1 -1.5581836 0.4017171 -0.7454051 0.3158695 -0.8589593 2.133539  
#> 2 -0.4444370 0.6330169 -0.6057750 0.3495872 -0.7515585 2.133539  
#> 3 -1.1131981 0.3823037 1.0929845 0.5042481 1.0141113 2.484420  
#> 4 -1.2124529 0.5429786 1.0124376 0.4210436 0.9067105 1.431776  
#> 5 -1.1367248 0.5342327 1.0124376 0.2747725 0.9067105 1.431776  
#> 6 0.9777057 0.4190036 0.9713761 0.5765085 0.8530101 1.431776  
#> OxygenFiller BowlSetpoint PressureSetpoint AirPressurer AlchRel  
#> 1 -0.37435417 0.7080026 -0.5599319 -0.1862599 -0.6183630  
#> 2 -0.22041456 0.7080026 -0.3491492 0.1533607 -0.6678614  
#> 3 -0.29481549 0.7080026 -0.4538621 -0.7010823 1.5061006  
#> 4 -0.08439545 0.7080026 -0.7762371 2.7707509 0.6032083  
#> 5 -0.08439545 0.7080026 -0.7762371 2.7707509 0.6032083  
#> 6 -0.29481549 0.7080026 -0.7762371 3.0859285 0.6416115  
#> CarbRel BallingLvl  
#> 1 -0.91517751 -0.6549488  
#> 2 -1.08432640 -0.5630274  
#> 3 2.89116829 1.4132824  
#> 4 -0.09733043 1.1375182  
#> 5 0.06085013 1.1375182  
#> 6 0.06085013 1.1145379

# Model Building

## Repeated Cross Validation

Because the test data values are not provided, it is not to our advantage to further split the training data into another test and train set. This is because the new “test” set may contain values that the training model will need when applied later to the actual test data. To side step this problem, the models will deploy a repeated cross validation approach that will split the training data up into chunks and train/test on those chunks.

While this may yield lower model power values in the training evaluation, the reading indicates that models built in this way will have better performance when applied to actual test data.

## Models Tested

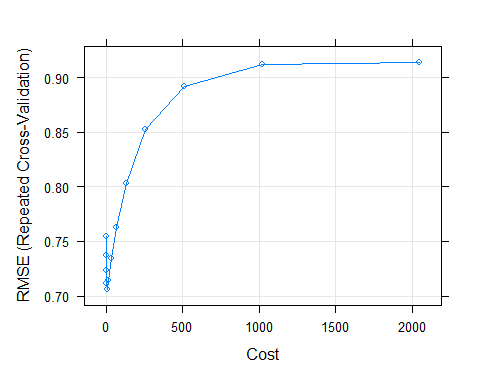
A series of models are tested based on the homework 2 problems. The models range from the simple linear regression to the more complex cubist and neural networks. A table is below that provides the full account of the performance values.

#> Linear Regression   
#>   
#> 2571 samples  
#> 31 predictor  
#>   
#> No pre-processing  
#> Resampling: Cross-Validated (5 fold, repeated 5 times)   
#> Summary of sample sizes: 2057, 2057, 2056, 2056, 2058, 2057, ...   
#> Resampling results:  
#>   
#> RMSE Rsquared MAE   
#> 0.8095471 0.3456735 0.6256341  
#>   
#> Tuning parameter 'intercept' was held constant at a value of TRUE

#> k-Nearest Neighbors   
#>   
#> 2571 samples  
#> 31 predictor  
#>   
#> No pre-processing  
#> Resampling: Cross-Validated (5 fold, repeated 5 times)   
#> Summary of sample sizes: 2056, 2056, 2056, 2058, 2058, 2058, ...   
#> Resampling results across tuning parameters:  
#>   
#> k RMSE Rsquared MAE   
#> 5 0.7342074 0.4675390 0.5395836  
#> 7 0.7360465 0.4639144 0.5455681  
#> 9 0.7381316 0.4622794 0.5497828  
#>   
#> RMSE was used to select the optimal model using the smallest value.  
#> The final value used for the model was k = 5.

#> Multivariate Adaptive Regression Spline   
#>   
#> 2571 samples  
#> 31 predictor  
#>   
#> No pre-processing  
#> Resampling: Cross-Validated (5 fold, repeated 5 times)   
#> Summary of sample sizes: 2056, 2057, 2057, 2057, 2057, 2056, ...   
#> Resampling results across tuning parameters:  
#>   
#> degree nprune RMSE Rsquared MAE   
#> 1 2 0.8816134 0.2240313 0.6926979  
#> 1 12 0.7859032 0.3837325 0.5984355  
#> 1 23 0.7776703 0.3973118 0.5875659  
#> 1 34 0.7776703 0.3973118 0.5875659  
#> 1 45 0.7776703 0.3973118 0.5875659  
#> 1 56 0.7776703 0.3973118 0.5875659  
#> 1 67 0.7776703 0.3973118 0.5875659  
#> 1 78 0.7776703 0.3973118 0.5875659  
#> 1 89 0.7776703 0.3973118 0.5875659  
#> 1 100 0.7776703 0.3973118 0.5875659  
#> 2 2 0.8805504 0.2259635 0.6898880  
#> 2 12 0.7903516 0.3860659 0.5937940  
#> 2 23 0.7615646 0.4321817 0.5618842  
#> 2 34 0.7581974 0.4390154 0.5569192  
#> 2 45 0.7582783 0.4390478 0.5568293  
#> 2 56 0.7582783 0.4390478 0.5568293  
#> 2 67 0.7582783 0.4390478 0.5568293  
#> 2 78 0.7582783 0.4390478 0.5568293  
#> 2 89 0.7582783 0.4390478 0.5568293  
#> 2 100 0.7582783 0.4390478 0.5568293  
#> 3 2 0.8784290 0.2296007 0.6893230  
#> 3 12 0.7853974 0.3922782 0.5915852  
#> 3 23 0.7617679 0.4343172 0.5604322  
#> 3 34 0.7562676 0.4495523 0.5475190  
#> 3 45 0.7499752 0.4560620 0.5441648  
#> 3 56 0.7494493 0.4567320 0.5437624  
#> 3 67 0.7494493 0.4567320 0.5437624  
#> 3 78 0.7494493 0.4567320 0.5437624  
#> 3 89 0.7494493 0.4567320 0.5437624  
#> 3 100 0.7494493 0.4567320 0.5437624  
#>   
#> RMSE was used to select the optimal model using the smallest value.  
#> The final values used for the model were nprune = 56 and degree = 3.

#> Support Vector Machines with Radial Basis Function Kernel   
#>   
#> 2571 samples  
#> 31 predictor  
#>   
#> Pre-processing: centered (31), scaled (31)   
#> Resampling: Cross-Validated (5 fold, repeated 5 times)   
#> Summary of sample sizes: 2056, 2057, 2057, 2056, 2058, 2057, ...   
#> Resampling results across tuning parameters:  
#>   
#> C RMSE Rsquared MAE   
#> 0.25 0.7546737 0.4404707 0.5579292  
#> 0.50 0.7369672 0.4635103 0.5384103  
#> 1.00 0.7237922 0.4806607 0.5249102  
#> 2.00 0.7118922 0.4965450 0.5150742  
#> 4.00 0.7060958 0.5048682 0.5126792  
#> 8.00 0.7062750 0.5075089 0.5148971  
#> 16.00 0.7144952 0.5030256 0.5234681  
#> 32.00 0.7340862 0.4887459 0.5399280  
#> 64.00 0.7627963 0.4680835 0.5625921  
#> 128.00 0.8031439 0.4396459 0.5929255  
#> 256.00 0.8524261 0.4074496 0.6267172  
#> 512.00 0.8920375 0.3822413 0.6526881  
#> 1024.00 0.9116467 0.3712516 0.6642284  
#> 2048.00 0.9137748 0.3700958 0.6654772  
#>   
#> Tuning parameter 'sigma' was held constant at a value of 0.02253823  
#> RMSE was used to select the optimal model using the smallest value.  
#> The final values used for the model were sigma = 0.02253823 and C = 4.



#> CART   
#>   
#> 2571 samples  
#> 31 predictor  
#>   
#> No pre-processing  
#> Resampling: Cross-Validated (5 fold, repeated 5 times)   
#> Summary of sample sizes: 2057, 2057, 2056, 2058, 2056, 2057, ...   
#> Resampling results across tuning parameters:  
#>   
#> maxdepth RMSE Rsquared MAE   
#> 1 0.8815596 0.2235588 0.6928833  
#> 2 0.8590453 0.2629717 0.6787900  
#> 3 0.8453314 0.2862346 0.6640514  
#> 4 0.8352063 0.3033180 0.6523718  
#> 5 0.8263863 0.3180867 0.6452712  
#> 6 0.8208213 0.3275036 0.6371896  
#> 7 0.8122001 0.3417138 0.6285510  
#> 8 0.8051946 0.3530071 0.6212110  
#> 9 0.7953609 0.3687140 0.6109835  
#> 10 0.7714252 0.4070259 0.5885329  
#>   
#> RMSE was used to select the optimal model using the smallest value.  
#> The final value used for the model was maxdepth = 10.

#> Stochastic Gradient Boosting   
#>   
#> 2571 samples  
#> 31 predictor  
#>   
#> No pre-processing  
#> Resampling: Cross-Validated (5 fold, repeated 5 times)   
#> Summary of sample sizes: 2057, 2058, 2055, 2058, 2056, 2056, ...   
#> Resampling results across tuning parameters:  
#>   
#> shrinkage interaction.depth n.trees RMSE Rsquared MAE   
#> 0.01 1 100 0.8899726 0.2700360 0.7066270  
#> 0.01 1 200 0.8509569 0.3188173 0.6717809  
#> 0.01 1 300 0.8290918 0.3400672 0.6524173  
#> 0.01 1 400 0.8166500 0.3528407 0.6413312  
#> 0.01 1 500 0.8078555 0.3634468 0.6336803  
#> 0.01 1 600 0.8012111 0.3719557 0.6277714  
#> 0.01 1 700 0.7959058 0.3790032 0.6230724  
#> 0.01 1 800 0.7913282 0.3850977 0.6188678  
#> 0.01 1 900 0.7874723 0.3903196 0.6153311  
#> 0.01 1 1000 0.7842990 0.3943908 0.6121507  
#> 0.01 3 100 0.8398570 0.3801385 0.6636492  
#> 0.01 3 200 0.7872951 0.4175246 0.6166201  
#> 0.01 3 300 0.7641639 0.4396233 0.5952897  
#> 0.01 3 400 0.7502387 0.4551801 0.5822993  
#> 0.01 3 500 0.7397545 0.4678049 0.5727898  
#> 0.01 3 600 0.7312423 0.4782758 0.5649348  
#> 0.01 3 700 0.7243978 0.4865010 0.5585153  
#> 0.01 3 800 0.7184976 0.4936282 0.5531039  
#> 0.01 3 900 0.7134079 0.4996949 0.5482262  
#> 0.01 3 1000 0.7089644 0.5050447 0.5439714  
#> 0.01 5 100 0.8176666 0.4273365 0.6437017  
#> 0.01 5 200 0.7595306 0.4611974 0.5909233  
#> 0.01 5 300 0.7339567 0.4845402 0.5675446  
#> 0.01 5 400 0.7182233 0.5013362 0.5528975  
#> 0.01 5 500 0.7065684 0.5146117 0.5423610  
#> 0.01 5 600 0.6977041 0.5248051 0.5343063  
#> 0.01 5 700 0.6904061 0.5332033 0.5274790  
#> 0.01 5 800 0.6837082 0.5410789 0.5212847  
#> 0.01 5 900 0.6784275 0.5473066 0.5165437  
#> 0.01 5 1000 0.6736886 0.5528558 0.5121998  
#> 0.01 7 100 0.8025904 0.4570228 0.6305116  
#> 0.01 7 200 0.7399277 0.4917138 0.5737893  
#> 0.01 7 300 0.7126036 0.5152265 0.5484156  
#> 0.01 7 400 0.6960045 0.5322245 0.5332951  
#> 0.01 7 500 0.6841195 0.5449967 0.5225741  
#> 0.01 7 600 0.6748411 0.5552493 0.5139231  
#> 0.01 7 700 0.6672072 0.5637720 0.5067742  
#> 0.01 7 800 0.6607915 0.5709515 0.5008119  
#> 0.01 7 900 0.6557202 0.5766662 0.4958997  
#> 0.01 7 1000 0.6510613 0.5819607 0.4914873  
#> 0.10 1 100 0.7832766 0.3956157 0.6112342  
#> 0.10 1 200 0.7662500 0.4158600 0.5926039  
#> 0.10 1 300 0.7596287 0.4239135 0.5831045  
#> 0.10 1 400 0.7561228 0.4283596 0.5782830  
#> 0.10 1 500 0.7542511 0.4309998 0.5753975  
#> 0.10 1 600 0.7521535 0.4341160 0.5725533  
#> 0.10 1 700 0.7500994 0.4373157 0.5704089  
#> 0.10 1 800 0.7491938 0.4387852 0.5685543  
#> 0.10 1 900 0.7482834 0.4403811 0.5678707  
#> 0.10 1 1000 0.7478052 0.4413001 0.5668609  
#> 0.10 3 100 0.7159315 0.4927580 0.5480372  
#> 0.10 3 200 0.6901638 0.5253012 0.5228086  
#> 0.10 3 300 0.6771489 0.5419945 0.5097235  
#> 0.10 3 400 0.6678278 0.5539767 0.5011545  
#> 0.10 3 500 0.6620257 0.5614090 0.4955108  
#> 0.10 3 600 0.6581568 0.5664021 0.4917668  
#> 0.10 3 700 0.6548186 0.5709062 0.4883200  
#> 0.10 3 800 0.6519778 0.5746642 0.4858068  
#> 0.10 3 900 0.6496532 0.5778580 0.4838754  
#> 0.10 3 1000 0.6476914 0.5805401 0.4816398  
#> 0.10 5 100 0.6835934 0.5361501 0.5196127  
#> 0.10 5 200 0.6598024 0.5652237 0.4967324  
#> 0.10 5 300 0.6500346 0.5772042 0.4864846  
#> 0.10 5 400 0.6436321 0.5854399 0.4800157  
#> 0.10 5 500 0.6387098 0.5917505 0.4752688  
#> 0.10 5 600 0.6359664 0.5954471 0.4727908  
#> 0.10 5 700 0.6341140 0.5980317 0.4706282  
#> 0.10 5 800 0.6329035 0.5997624 0.4692830  
#> 0.10 5 900 0.6316690 0.6014714 0.4680993  
#> 0.10 5 1000 0.6310559 0.6023747 0.4674958  
#> 0.10 7 100 0.6617631 0.5645549 0.5003082  
#> 0.10 7 200 0.6422992 0.5874624 0.4809028  
#> 0.10 7 300 0.6327989 0.5992583 0.4710611  
#> 0.10 7 400 0.6282194 0.6049408 0.4660868  
#> 0.10 7 500 0.6249570 0.6092041 0.4620826  
#> 0.10 7 600 0.6231346 0.6115586 0.4603154  
#> 0.10 7 700 0.6219204 0.6132303 0.4594629  
#> 0.10 7 800 0.6215887 0.6138268 0.4587729  
#> 0.10 7 900 0.6214572 0.6141270 0.4586449  
#> 0.10 7 1000 0.6212159 0.6145606 0.4581797  
#>   
#> Tuning parameter 'n.minobsinnode' was held constant at a value of 8  
#> RMSE was used to select the optimal model using the smallest value.  
#> The final values used for the model were n.trees = 1000, interaction.depth =  
#> 7, shrinkage = 0.1 and n.minobsinnode = 8.

#> Cubist   
#>   
#> 2571 samples  
#> 31 predictor  
#>   
#> No pre-processing  
#> Resampling: Cross-Validated (5 fold, repeated 5 times)   
#> Summary of sample sizes: 2056, 2057, 2057, 2057, 2057, 2057, ...   
#> Resampling results across tuning parameters:  
#>   
#> committees neighbors RMSE Rsquared MAE   
#> 1 0 0.7451550 0.4828956 0.5095084  
#> 1 5 0.7110425 0.5369391 0.4805090  
#> 1 9 0.7151228 0.5269322 0.4836603  
#> 10 0 0.6144533 0.6256302 0.4418918  
#> 10 5 0.5764165 0.6679684 0.4113235  
#> 10 9 0.5807689 0.6628977 0.4140357  
#> 20 0 0.6065476 0.6373009 0.4362906  
#> 20 5 0.5676378 0.6779514 0.4047315  
#> 20 9 0.5722403 0.6731160 0.4075701  
#>   
#> RMSE was used to select the optimal model using the smallest value.  
#> The final values used for the model were committees = 20 and neighbors = 5.

## Model Results

The following table indicates scores of RMSE, R-Squared, and MAE. The RMSE, or the Root Mean Square Error, is the square root of the variance of the residuals, which indicates the absolute fit of the model according to the data. In other words, how close each observed data points are to the model’s predicted values. Lower values of RMSE indicate a better fit, as this is a good measure of how accurately the model predicts the response. R-squared indicates the goodness of fit for the model, and it ranges from 0 to 1. 0 being an indication that the proposed model does not improve prediction over the mean model, while 1 indicates perfect prediction. MAE, or Mean of the Abosolute Errors, is a measure of errors among paired observations within a model. This is another measuse of accuracy, and when a MAE score is generated, the lower score indicates a better performance. For our model selection, we will be consistent with the model output of R, and choose the best performing model according to the RMSE score. [RMSE, R-Squared](https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/#:~:text=As%20the%20square%20root%20of%20a%20variance%2C%20RMSE,variable.%20Lower%20values%20of%20RMSE%20indicate%20better%20fit.) [MAE](https://www.statology.org/mean-absolute-error-in-r/#:~:text=In%20statistics%2C%20the%20mean%20absolute%20error%20%28MAE%29%20is,yi%3A%20The%20observed%20value%20for%20the%20ith%20observation)

Model RMSE Rsquared MAE

Linear Regression 0.8095471 0.3456735 0.6256341 KNN 0.7342074 0.4675390 0.5395836 MARS 0.7494493 0.4567320 0.5437624 SVM 0.7060958 0.5048682 0.5126792 CART 0.7714252 0.4070259 0.5885329 Stoch. Gradient Boosting 0.6212159 0.6145606 0.4581797 *Cubist 0.5722403 0.6731160 0.4075701*

Looking at the table above, the best performing model according to the RMSE is the Cubist Model

# Model Selection

#> # A tibble: 6 x 33  
#> BrandCode CarbVolume FillOunces PCVolume CarbPressure CarbTemp PSC PSCFill  
#> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
#> 1 D 5.48 24.0 0.27 65.4 135. 0.236 0.4   
#> 2 A 5.39 24.0 0.227 63.2 135 0.042 0.22  
#> 3 B 5.29 23.9 0.303 66.4 140. 0.068 0.1   
#> 4 B 5.27 23.9 0.186 64.8 139 0.004 0.2   
#> 5 B 5.41 24.2 0.16 69.4 142. 0.04 0.3   
#> 6 B 5.29 24.1 0.212 73.4 147. 0.078 0.22  
#> # ... with 25 more variables: PSCCO2 <dbl>, MnfFlow <dbl>, CarbPressure1 <dbl>,  
#> # FillPressure <dbl>, HydPressure1 <dbl>, HydPressure2 <dbl>,  
#> # HydPressure3 <dbl>, HydPressure4 <dbl>, FillerLevel <dbl>,  
#> # FillerSpeed <dbl>, Temperature <dbl>, Usagecont <dbl>, CarbFlow <dbl>,  
#> # Density <dbl>, MFR <dbl>, Balling <dbl>, PressureVacuum <dbl>, PH <lgl>,  
#> # OxygenFiller <dbl>, BowlSetpoint <dbl>, PressureSetpoint <dbl>,  
#> # AirPressurer <dbl>, AlchRel <dbl>, CarbRel <dbl>, BallingLvl <dbl>

#> [1] 8.580132 8.481056 8.596498 8.630728 8.509520 8.574694

#> ""