DATA 624 Project 2

Jonathan Hernandez

- Training and test data has been loaded and let's read it in (The data have been imputed, removed correlated data etc)

Source: https://github.com/john-grando/data624_hw_group_2/tree/master/Project2

- Preview the training data

```
##        PH        Brand.Code  Carb.Volume     Fill.Ounces
##  Min.   :7.880    : 120     Min.   :5.040   Min.   :23.63
##  1st Qu.:8.440   A: 293     1st Qu.:5.293   1st Qu.:23.92
##  Median :8.540   B:1239     Median :5.347   Median :23.97
##  Mean   :8.546   C: 304     Mean   :5.370   Mean   :23.97
##  3rd Qu.:8.680   D: 615     3rd Qu.:5.453   3rd Qu.:24.03
##  Max.   :9.360              Max.   :5.700   Max.   :24.32
##    PC.Volume       Carb.Pressure     Carb.Temp          PSC
##  Min.   :0.07933   Min.   :57.00   Min.   :128.6   Min.   :0.0020
##  1st Qu.:0.23933   1st Qu.:65.60   1st Qu.:138.4   1st Qu.:0.0480
##  Median :0.27133   Median :68.20   Median :140.8   Median :0.0760
##  Mean   :0.27771   Mean   :68.21   Mean   :141.1   Mean   :0.0847
##  3rd Qu.:0.31267   3rd Qu.:70.60   3rd Qu.:143.8   3rd Qu.:0.1120
##  Max.   :0.47800   Max.   :79.40   Max.   :154.0   Max.   :0.2700
##     PSC.Fill         PSC.CO2           Mnf.Flow       Carb.Pressure1
##  Min.   :0.0000   Min.   :0.00000   Min.   :-100.20   Min.   :105.6
##  1st Qu.:0.1000   1st Qu.:0.02000   1st Qu.:-100.00   1st Qu.:119.0
##  Median :0.1800   Median :0.04000   Median :  64.80   Median :123.2
##  Mean   :0.1964   Mean   :0.05658   Mean   :  24.47   Mean   :122.6
##  3rd Qu.:0.2600   3rd Qu.:0.08000   3rd Qu.: 140.80   3rd Qu.:125.4
##  Max.   :0.6200   Max.   :0.24000   Max.   : 229.40   Max.   :140.2
##  Fill.Pressure   Hyd.Pressure1   Hyd.Pressure2   Hyd.Pressure3
##  Min.   :34.60   Min.   :-0.80   Min.   : 0.00   Min.   :-1.20
##  1st Qu.:46.00   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.00
##  Median :46.40   Median :11.40   Median :28.60   Median :27.60
##  Mean   :47.92   Mean   :12.47   Mean   :20.98   Mean   :20.47
##  3rd Qu.:50.00   3rd Qu.:20.20   3rd Qu.:34.60   3rd Qu.:33.40
##  Max.   :60.40   Max.   :58.00   Max.   :59.40   Max.   :50.00
##  Hyd.Pressure4     Filler.Level    Filler.Speed   Temperature
##  Min.   : 52.00   Min.   : 55.8   Min.   : 998   Min.   :63.60
##  1st Qu.: 86.00   1st Qu.: 98.1   1st Qu.:3825   1st Qu.:65.20
##  Median : 96.00   Median :118.4   Median :3980   Median :65.60
##  Mean   : 96.52   Mean   :109.2   Mean   :3633   Mean   :65.99
##  3rd Qu.:102.00   3rd Qu.:120.0   3rd Qu.:3996   3rd Qu.:66.40
##  Max.   :142.00   Max.   :161.2   Max.   :4030   Max.   :76.20
##    Usage.cont      Carb.Flow       Density         Balling
##  Min.   :12.08   Min.   :  26   Min.   :0.240   Min.   :-0.170
##  1st Qu.:18.36   1st Qu.:1142   1st Qu.:0.900   1st Qu.: 1.496
##  Median :21.80   Median :3028   Median :0.980   Median : 1.648
##  Mean   :21.00   Mean   :2468   Mean   :1.174   Mean   : 2.198
##  3rd Qu.:23.76   3rd Qu.:3187   3rd Qu.:1.620   3rd Qu.: 3.292
```

```
##  Max.   :25.90   Max.   :5104   Max.   :1.920   Max.   : 4.012
##  Pressure.Vacuum  Oxygen.Filler     Bowl.Setpoint   Pressure.Setpoint
##  Min.   :-6.600   Min.   :0.00240   Min.   : 70.0   Min.   :44.00
##  1st Qu.:-5.600   1st Qu.:0.02200   1st Qu.:100.0   1st Qu.:46.00
##  Median :-5.400   Median :0.03340   Median :120.0   Median :46.00
##  Mean   :-5.216   Mean   :0.04711   Mean   :109.3   Mean   :47.61
##  3rd Qu.:-5.000   3rd Qu.:0.06000   3rd Qu.:120.0   3rd Qu.:50.00
##  Max.   :-3.600   Max.   :0.40000   Max.   :140.0   Max.   :52.00
##  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   :6.896   Mean   :5.436   Mean   :2.05
##  3rd Qu.:143.0   3rd Qu.:7.220   3rd Qu.:5.540   3rd Qu.:3.14
##  Max.   :148.2   Max.   :8.620   Max.   :6.060   Max.   :3.66

## 'data.frame':    2571 obs. of  32 variables:
##  $ PH               : num  8.36 8.26 8.94 8.24 8.26 8.32 8.4 8.38 8.38 8.5 ...
##  $ Brand.Code       : Factor w/ 5 levels "","A","B","C",..: 3 2 3 2 2 2 2 3 3 3 ...
##  $ Carb.Volume      : num  5.34 5.43 5.29 5.44 5.49 ...
##  $ Fill.Ounces      : num  24 24 24.1 24 24.3 ...
##  $ PC.Volume        : num  0.263 0.239 0.263 0.293 0.111 ...
##  $ Carb.Pressure    : num  68.2 68.4 70.8 63 67.2 66.6 64.2 67.6 64.2 72 ...
##  $ Carb.Temp        : num  141 140 145 133 137 ...
##  $ PSC              : num  0.104 0.124 0.09 0.162 0.026 0.09 0.128 0.154 0.132 0.014 ...
##  $ PSC.Fill         : num  0.26 0.22 0.34 0.42 0.16 0.24 0.4 0.34 0.12 0.24 ...
##  $ PSC.CO2          : num  0.04 0.04 0.16 0.04 0.12 0.04 0.04 0.04 0.14 0.06 ...
##  $ Mnf.Flow         : num  -100 -100 -100 -100 -100 -100 -100 -100 -100 -100 ...
##  $ Carb.Pressure1   : num  119 122 120 115 118 ...
##  $ Fill.Pressure    : num  46 46 46 46.4 45.8 45.6 51.8 46.8 46 45.2 ...
##  $ Hyd.Pressure1    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hyd.Pressure2    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hyd.Pressure3    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hyd.Pressure4    : int  118 106 82 92 92 116 124 132 90 108 ...
##  $ Filler.Level     : num  121 119 120 118 119 ...
##  $ Filler.Speed     : int  4002 3986 4020 4012 4010 4014 1006 1004 4014 4028 ...
##  $ Temperature      : num  66 67.6 67 65.6 65.6 66.2 65.8 65.2 65.4 66.6 ...
##  $ Usage.cont       : num  16.2 19.9 17.8 17.4 17.7 ...
##  $ Carb.Flow        : int  2932 3144 2914 3062 3054 2948 30 684 2902 3038 ...
##  $ Density          : num  0.88 0.92 1.58 1.54 1.54 1.52 0.84 0.84 0.9 0.9 ...
##  $ Balling          : num  1.4 1.5 3.14 3.04 3.04 ...
##  $ Pressure.Vacuum  : num  -4 -4 -3.8 -4.4 -4.4 -4.4 -4.4 -4.4 -4.4 -4.4 ...
##  $ Oxygen.Filler    : num  0.022 0.026 0.024 0.03 0.03 0.024 0.066 0.046 0.064 0.022 ...
##  $ Bowl.Setpoint    : int  120 120 120 120 120 120 120 120 120 120 ...
##  $ Pressure.Setpoint: num  46.4 46.8 46.6 46 46 46 46 46 46 46 ...
##  $ Air.Pressurer    : num  143 143 142 146 146 ...
##  $ Alch.Rel         : num  6.58 6.56 7.66 7.14 7.14 7.16 6.54 6.52 6.52 6.54 ...
##  $ Carb.Rel         : num  5.32 5.3 5.84 5.42 5.44 5.44 5.38 5.34 5.34 5.34 ...
##  $ Balling.Lvl      : num  1.48 1.56 3.28 3.04 3.04 3.02 1.44 1.44 1.44 1.38 ...

## [1] 2571   32
```

Regression - Forward Selection

- A simple multiple linear regression via forward selection. The regsubsets() function from the leaps package picks the best fit linear regression model and features using 1 feature, 2 features and so on.

```
## [1] "Number of features that make the best model and rmse:  29"
```

```
## [1] "Smallest RMSE:  0.1314"
```

	x
(Intercept)	10.5664267
Brand.CodeA	0.0023316
Brand.CodeB	0.0778317
Brand.CodeC	-0.0677256
Brand.CodeD	0.0564902
Carb.Volume	-0.0686755
Fill.Ounces	-0.0742965
PC.Volume	-0.0991171
Carb.Temp	0.0008175
PSC	-0.0959440
PSC.Fill	-0.0296124
PSC.CO2	-0.1252350
Mnf.Flow	-0.0007094
Carb.Pressure1	0.0070046
Fill.Pressure	0.0022677
Hyd.Pressure2	-0.0009184
Hyd.Pressure3	0.0030794
Filler.Level	-0.0012175
Temperature	-0.0123710
Usage.cont	-0.0071309
Carb.Flow	0.0000122
Density	-0.1252851
Balling	-0.0675429
Pressure.Vacuum	-0.0186663
Oxygen.Filler	-0.3082169
Bowl.Setpoint	0.0034649
Pressure.Setpoint	-0.0087758
Air.Pressurer	-0.0025656
Alch.Rel	0.0726519
Balling.Lvl	0.1032263

- By using the regsubsets() function and looping through all feature number combinations, we were able to find the right linear regression model that minimizes the RMSE between the training data and using the know features and putting it into a lm object, we were able to make predictions of the PH levels in the test dataset.
- Now, let's use elastic net with various parameters α and λ using the train() function

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

```
##       alpha      lambda      RMSE  Rsquared       MAE      RMSESD
## 8 0.4566147 0.001219054 0.1337129 0.3996417 0.1037881 0.005189279
##   RsquaredSD       MAESD
## 8  0.0335675 0.003555362
```

```
## Warning in write.csv(enet_pred_test_PH, "prediction-PH-enet.csv", col.names
## = "PH"): attempt to set 'col.names' ignored
```