DATA 624 Predictive Analytics (Project 2)

Beshkia Kvarnstrom, Nikoleta Emanouilidi, Evan McLaughlin, Victor Torres & Vladimir

2024-12-03

INSTRUCTIONS

Hide Assignment Information Instructions 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.

Due on Dec 15, 2024 11:59 PM

Load Packages

The following code below loops through the list of necessary packages and checks to determine if each is installed. If the package is not found it is installed and loaded.

```
## Warning: package 'kableExtra' was built under R version 4.3.3
```

```
## Warning: package 'summarytools' was built under R version 4.3.3
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'randomForest' was built under R version 4.3.3
```

```
## Warning: package 'glmnet' was built under R version 4.3.3
```

```
## Warning: package 'mice' was built under R version 4.3.3
```

DATA EXPLORATION

Load The Data

The script retrieves training (StudentData.xlsx) and test datasets (StudentEvaluation.xlsx) from GitHub, reads them into data frames, and removes the temporary files to maintain a clean workspace.

View The Data

Performs analysis of the Student_Train dataset and display a summary of the structure of the dataset, helping to understand the data’s composition.

```
## There are  2571  observations/cases in the Student Training dataset.
```

```
## There are  33  columns/elements in the Student Training dataset
```

```
## There are  1  Categorical Variables in the Student Training dataset.
```

```
## There are  32  Predictor Variables in the Student Training dataset.
```

Display both a structural overview and a statistical summary of the training dataset.

```
## 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,~
## $ ‘Hyd Pressure2‘     <dbl> NA, NA, NA, 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,~
## $ ‘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.~
```

Brand Code	Carb Volume	Fill Ounces	PC Volume	Carb Pressure	Carb Temp
Length:2571	Min. :5.040	Min. :23.63	Min. :0.07933	Min. :57.00	Min. :128.6
Class :character	1st Qu.:5.293	1st Qu.:23.92	1st Qu.:0.23917	1st Qu.:65.60	1st Qu.:138.4
Mode :character	Median :5.347	Median :23.97	Median :0.27133	Median :68.20	Median :140.8
NA	Mean :5.370	Mean :23.97	Mean :0.27712	Mean :68.19	Mean :141.1
NA	3rd Qu.:5.453	3rd Qu.:24.03	3rd Qu.:0.31200	3rd Qu.:70.60	3rd Qu.:143.8
NA	Max. :5.700	Max. :24.32	Max. :0.47800	Max. :79.40	Max. :154.0
NA	NA’s :10	NA’s :38	NA’s :39	NA’s :27	NA’s :26

Display the first 10 observations/cases in the Student Training dataset.

Brand Code	Carb Volume	Fill Ounces	PC Volume	Carb Pressure	Carb Temp	PSC	PSC Fill
B	5.340000	23.96667	0.2633333	68.2	141.2	0.104	0.26
A	5.426667	24.00667	0.2386667	68.4	139.6	0.124	0.22
B	5.286667	24.06000	0.2633333	70.8	144.8	0.090	0.34
A	5.440000	24.00667	0.2933333	63.0	132.6	NA	0.42
A	5.486667	24.31333	0.1113333	67.2	136.8	0.026	0.16
A	5.380000	23.92667	0.2693333	66.6	138.4	0.090	0.24
A	5.313333	23.88667	0.2680000	64.2	136.8	0.128	0.40
B	5.320000	24.17333	0.2206667	67.6	141.4	0.154	0.34
B	5.246667	23.98000	0.2626667	64.2	140.2	0.132	0.12
B	5.266667	24.00667	0.2313333	72.0	147.4	0.014	0.24

DATA WRANGLING

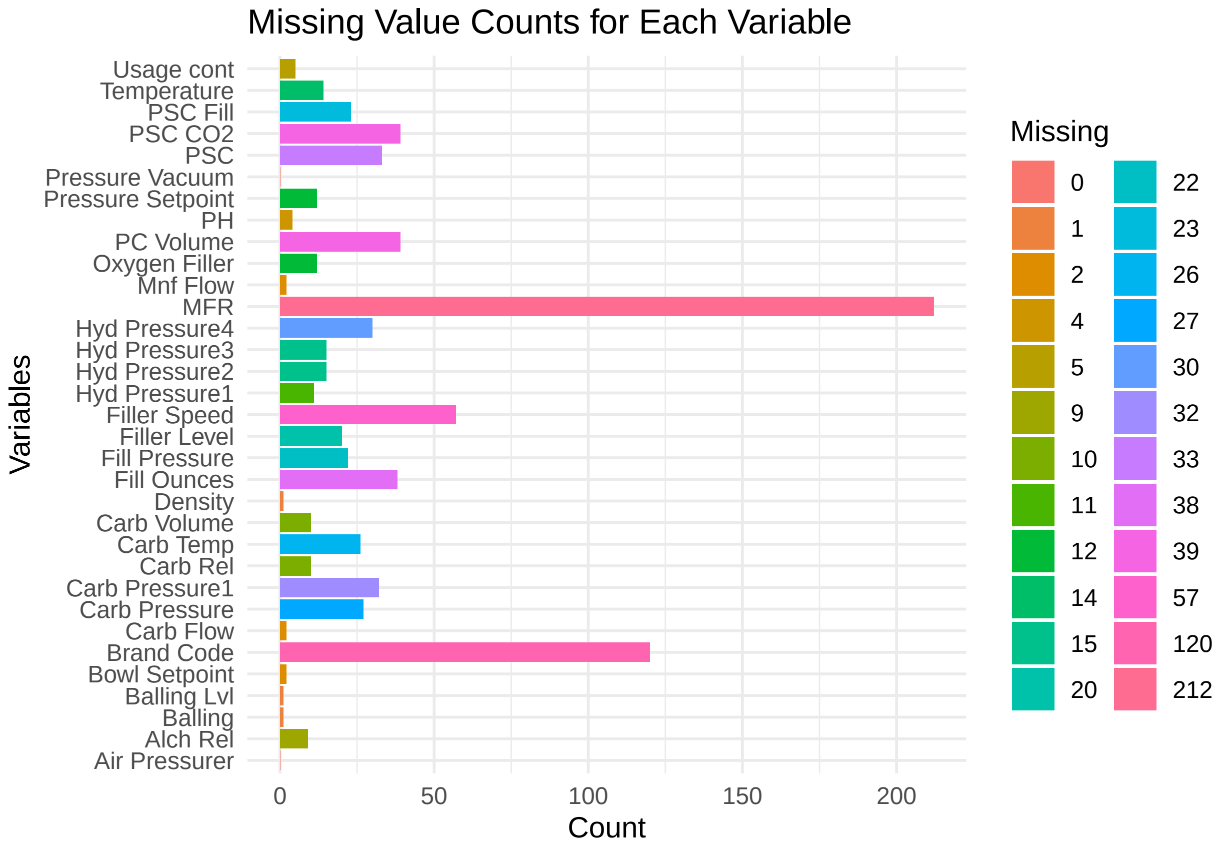
The following functions help clean and preprocess the dataset by handling missing values, outliers, feature scaling, and constant columns.

The following provides a comprehensive analysis of missing data by both displaying a table of missing values per column and generating a bar plot to visualize the distribution of missing values across variables. The table allows for detailed inspection, while the plot gives a quick overview of the missing data across the dataset.

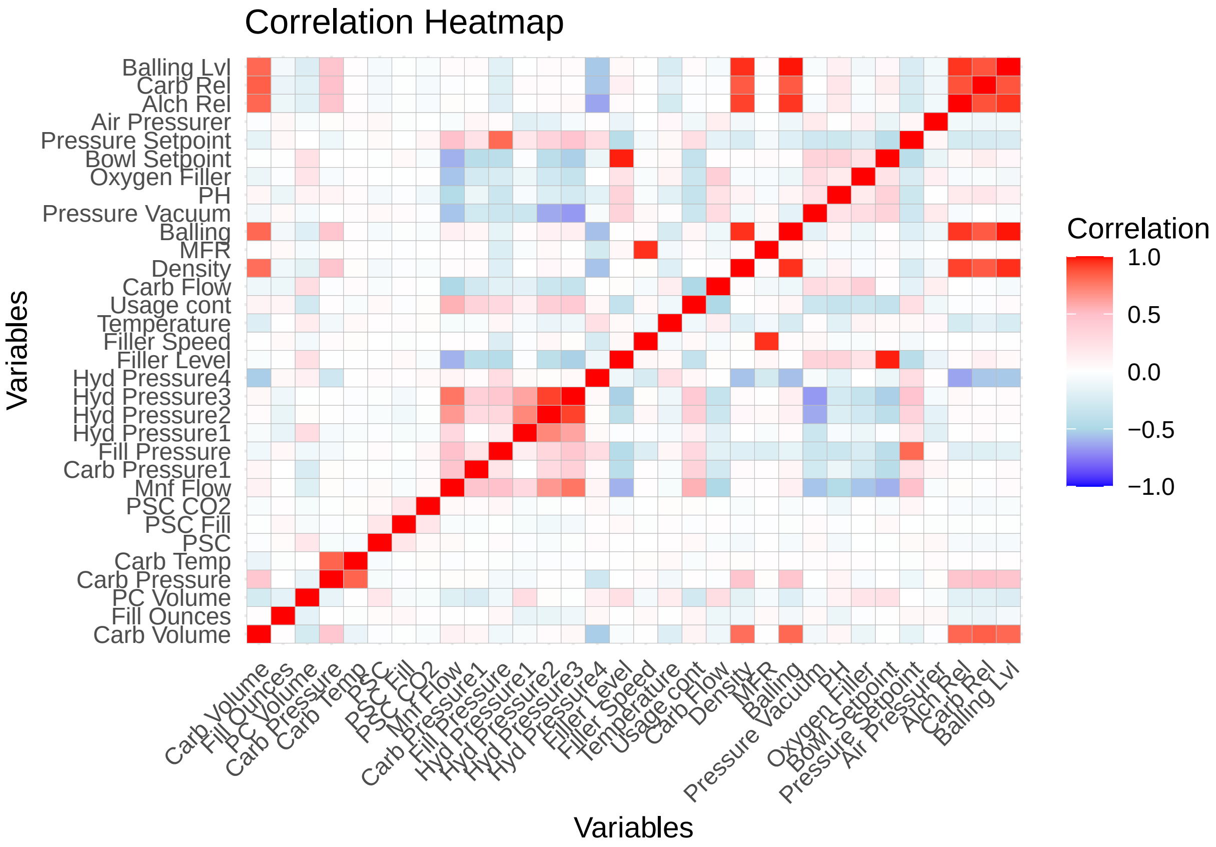
The MFR has the highest percentage of missing data (8.2%), while Density has the least (0%). Many of the variables have relatively low missing data (around 0.1% to 0.5%). While some variables, particularly MFR, Brand Code, Filler Speed, and others, have missing values in the range of 1% to 8%, which will be imputed in the data cleaning process.

```
## Missing values per column:
```

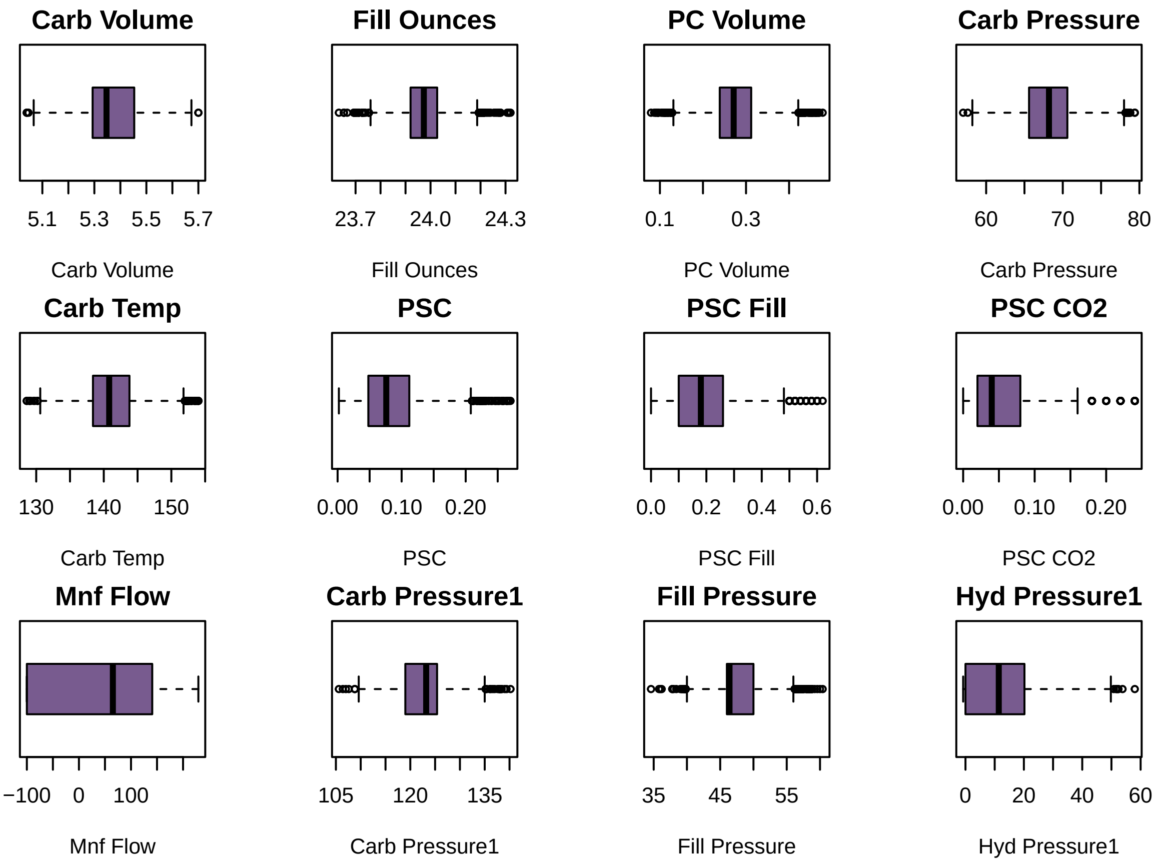
variable	n	percent
MFR	212	8.2%
Brand Code	120	4.7%
Filler Speed	57	2.2%
PC Volume	39	1.5%
PSC CO2	39	1.5%
Fill Ounces	38	1.5%
PSC	33	1.3%
Carb Pressure1	32	1.2%
Hyd Pressure4	30	1.2%
Carb Pressure	27	1.1%
Carb Temp	26	1%
PSC Fill	23	0.9%
Fill Pressure	22	0.9%
Filler Level	20	0.8%
Hyd Pressure2	15	0.6%
Hyd Pressure3	15	0.6%
Temperature	14	0.5%
Oxygen Filler	12	0.5%
Pressure Setpoint	12	0.5%
Hyd Pressure1	11	0.4%
Carb Rel	10	0.4%
Carb Volume	10	0.4%
Alch Rel	9	0.4%
Usage cont	5	0.2%
PH	4	0.2%
Bowl Setpoint	2	0.1%
Carb Flow	2	0.1%
Mnf Flow	2	0.1%
Balling	1	0%
Balling Lvl	1	0%
Density	1	0%

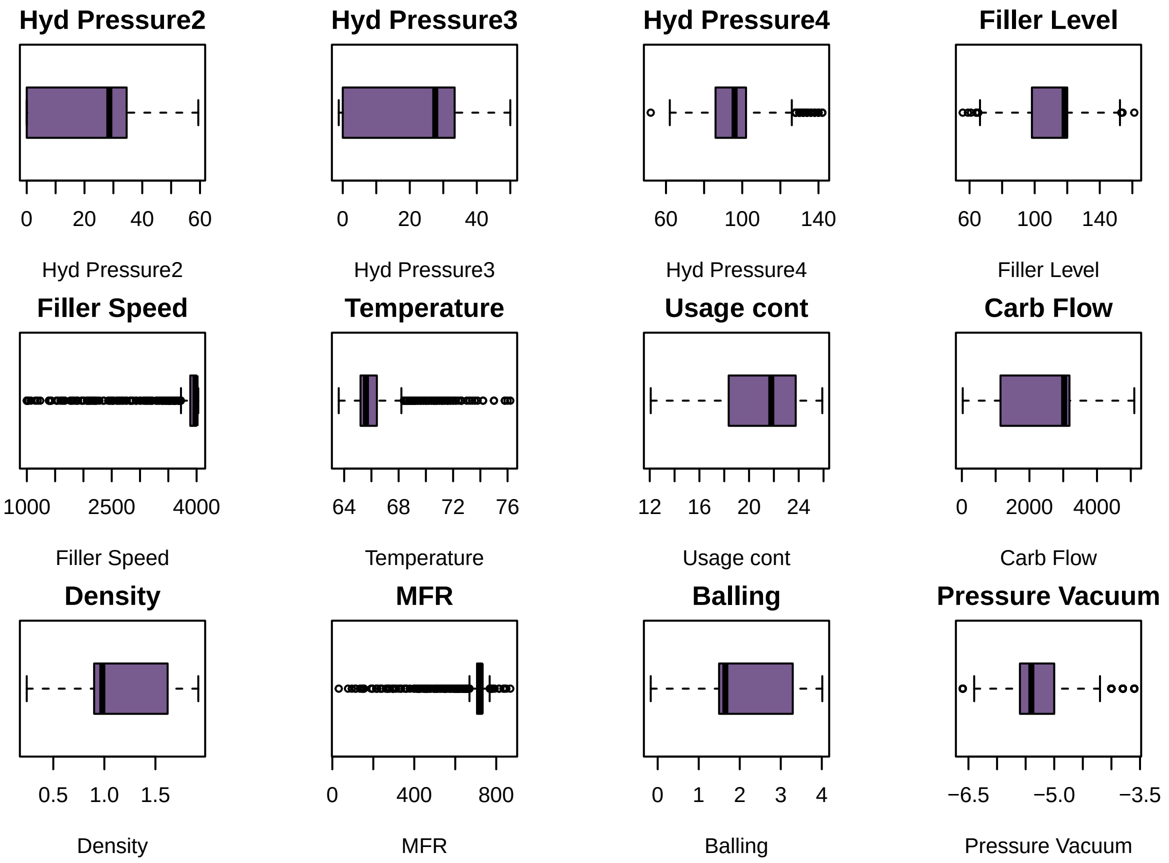


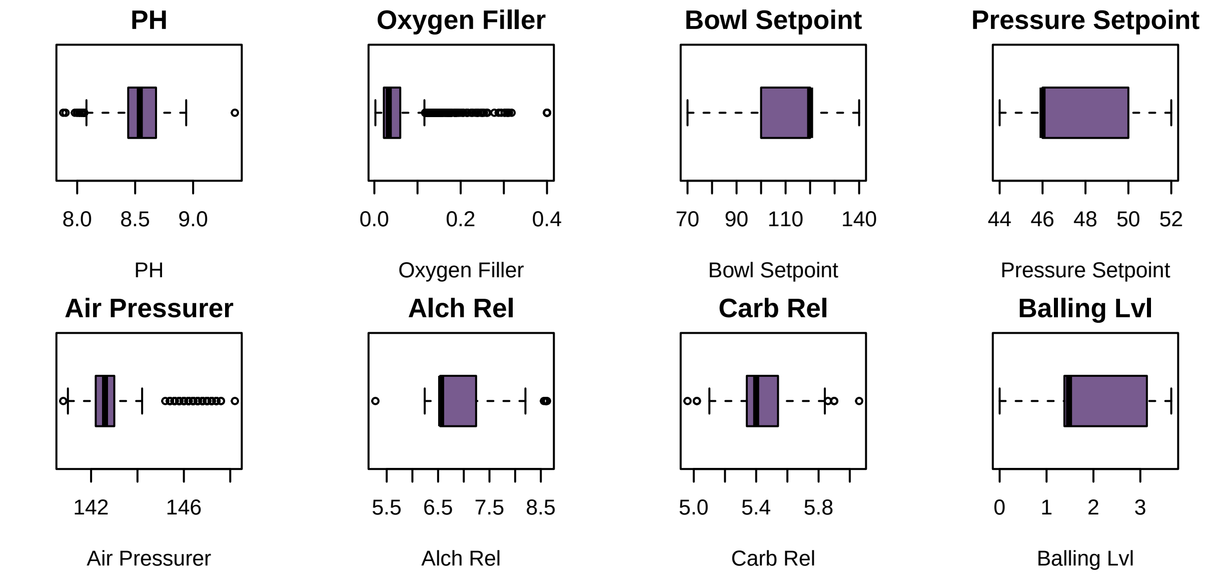
The following calculates and visualizes the correlations between numeric variables in the Student_Train dataset using a heatmap. It highlights the strength and direction of the relationships between variables, with negative correlations shown in blue, positive correlations in red, and zero correlations in gray. The heatmap provides an intuitive way to identify strong or weak correlations in the dataset.

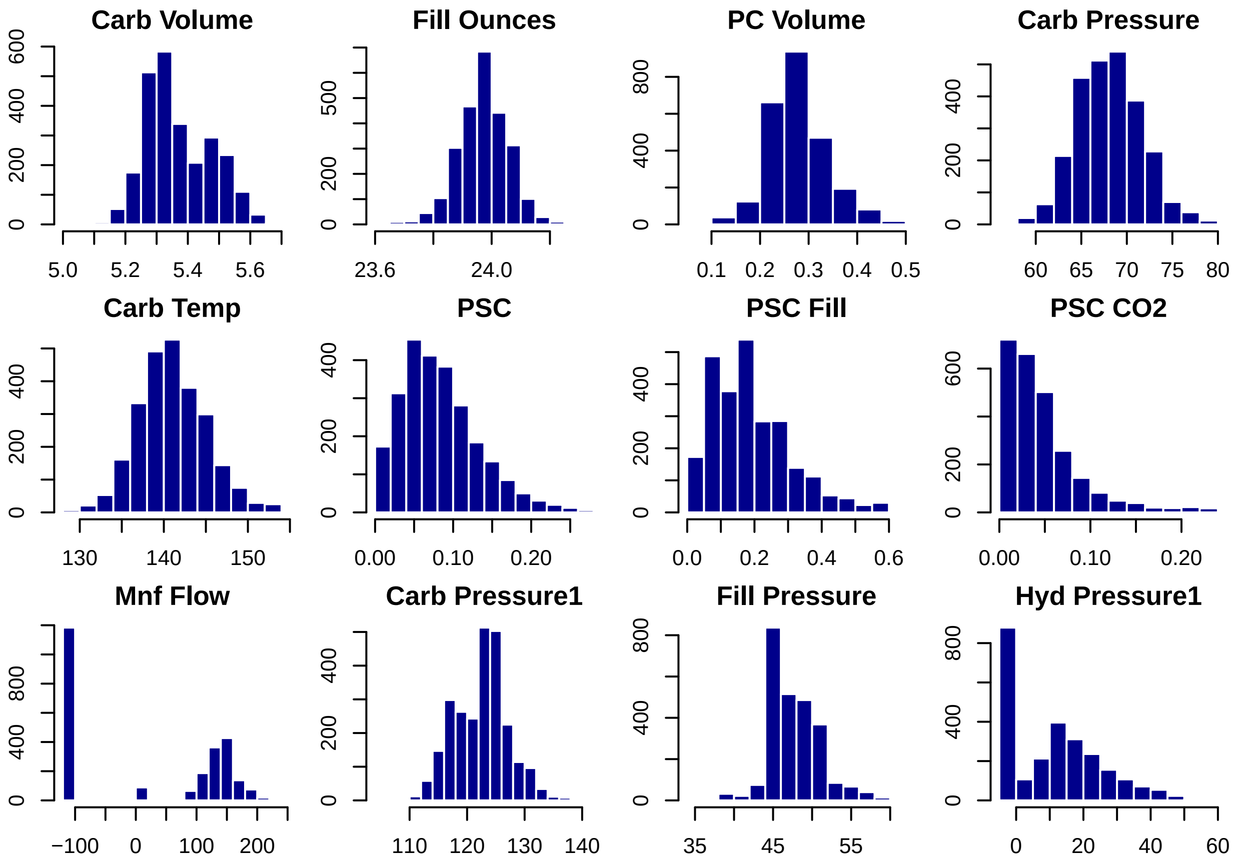


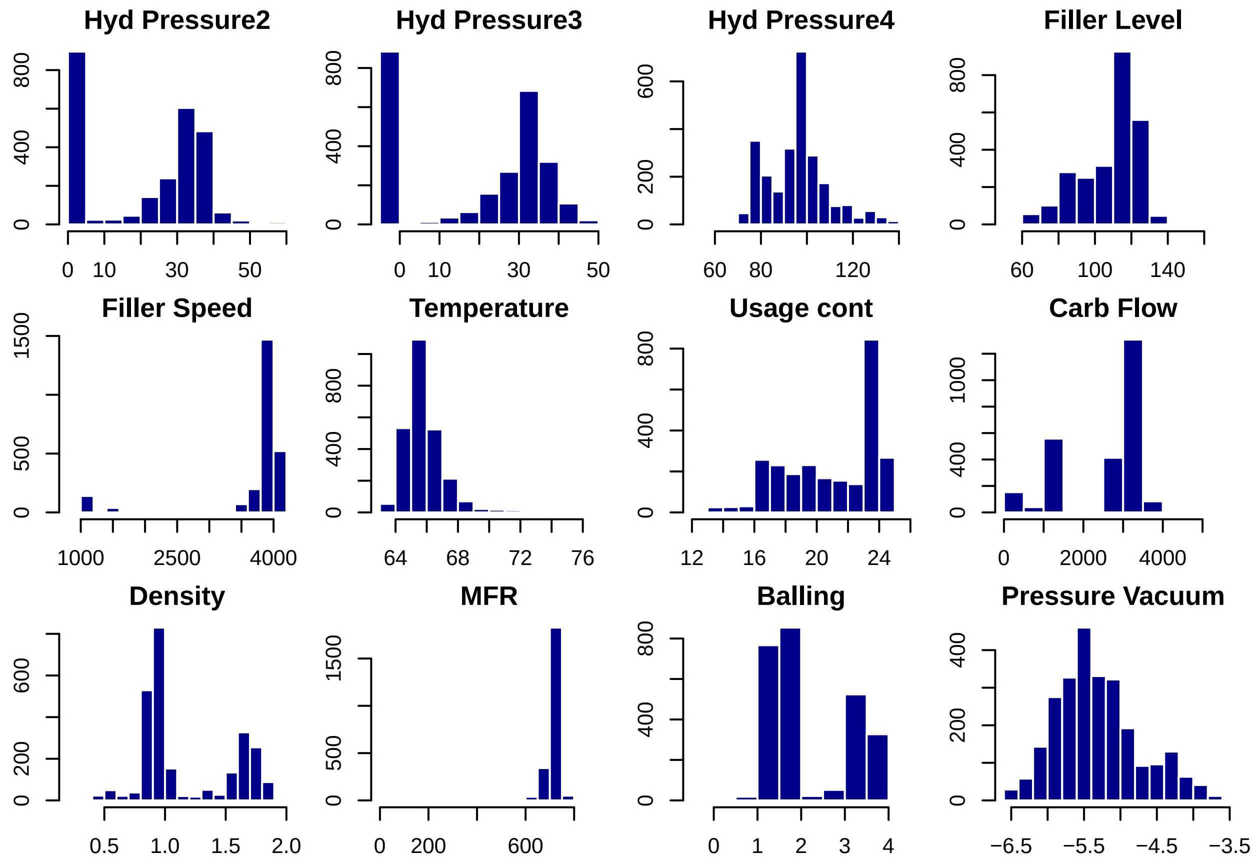
The following boxplots represents the distribution of values for a different variable in the Student_Train dataset (excluding the first column). Each plot is labeled with the variable name, and the boxplots are horizontally oriented with a custom color scheme. This visualization helps to identify the spread, central tendency, and potential outliers for each variable.

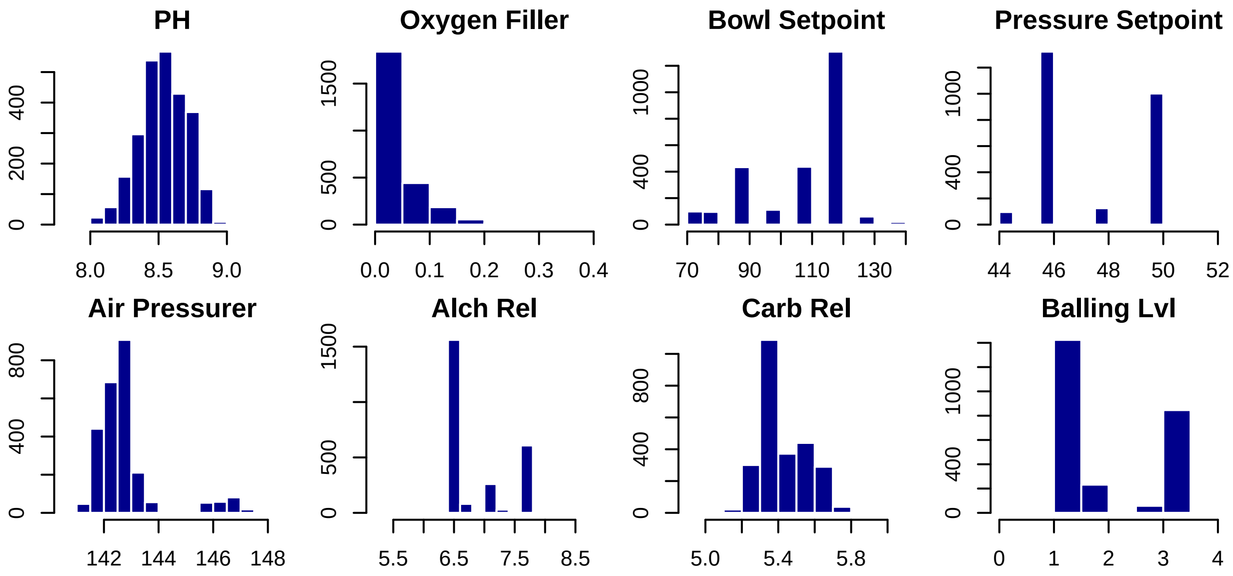

Carb Volume
5.1
5.3
5.5
5.7
Carb Volume
Fill Ounces
23.7
24.0
24.3
Fill Ounces
PC Volume
0.1
0.3
PC Volume
Carb Pressure
60
70
80
Carb Pressure
Carb Temp
130
140
150
Carb Temp
PSC
0.00
0.10
0.20
PSC
PSC Fill
0.0
0.2
0.4
0.6
PSC Fill
PSC CO2
0.00
0.10
0.20
PSC CO2
Mnf Flow
−100
0
100
Mnf Flow
Carb Pressure1
105
120
135
Carb Pressure1
Fill Pressure
35
45
55
Fill Pressure
Hyd Pressure1
0
20
40
60
Hyd Pressure1


Hyd Pressure2
0
20
40
60
Hyd Pressure2
Hyd Pressure3
0
20
40
Hyd Pressure3
Hyd Pressure4
60
100
140
Hyd Pressure4
Filler Level
60
100
140
Filler Level
Filler Speed
1000
2500
4000
Filler Speed
Temperature
64
68
72
76
Temperature
Usage cont
12
16
20
24
Usage cont
Carb Flow
0
2000
4000
Carb Flow
Density
0.5
1.0
1.5
Density
MFR
0
400
800
MFR
Balling
0
1
2
3
4
Balling
Pressure Vacuum
−6.5
−5.0
−3.5
Pressure Vacuum


PH
8.0
8.5
9.0
PH
Oxygen Filler
0.0
0.2
0.4
Oxygen Filler
Bowl Setpoint
70
90
110
140
Bowl Setpoint
Pressure Setpoint
44
46
48
50
52
Pressure Setpoint
Air Pressurer
142
146
Air Pressurer
Alch Rel
5.5
6.5
7.5
8.5
Alch Rel
Carb Rel
5.0
5.4
5.8
Carb Rel
Balling Lvl
0
1
2
3
Balling Lvl


Carb Volume
0
200
400
600
5.0
5.2
5.4
5.6
Fill Ounces
0
200
500
23.6
24.0
PC Volume
0
400
800
0.1
0.2
0.3
0.4
0.5
Carb Pressure
0
200
400
60
65
70
75
80
Carb Temp
0
200
400
130
140
150
PSC
0
200
400
0.00
0.10
0.20
PSC Fill
0
200
400
0.0
0.2
0.4
0.6
PSC CO2
0
200
600
0.00
0.10
0.20
Mnf Flow
0
400
800
−100
0
100
200
Carb Pressure1
0
200
400
110
120
130
140
Fill Pressure
0
400
800
35
45
55
Hyd Pressure1
0
400
800
0
20
40
60


Hyd Pressure2
0
400
800
0 10 30 50
Hyd Pressure3
0
400
800
0 10 30 50
Hyd Pressure4
0
200
600
60 80 120
Filler Level
0
400
800
60 100 140
Filler Speed
0
500
1500
1000 2500 4000
Temperature
0
400
800
64 68 72 76
Usage cont
0
400
800
12 16 20 24
Carb Flow
0
400
1000
0 2000 4000
Density
0
200
600
0.5 1.0 1.5 2.0
MFR
0
500
1500
0 200 600
Balling
0
400
800
0 1 2 3 4
Pressure Vacuum
0
200
400
−6.5 −5.5 −4.5 −3.5



Variables that are highly Right-Skewed are: PSC, PSC Fill, PSC CO2,Hyd Pressure1-3, Oxygen Filler, Air Pressurer. A suggested Transformation would be to apply a logarithmic or a square root transformation to reduce skewness.

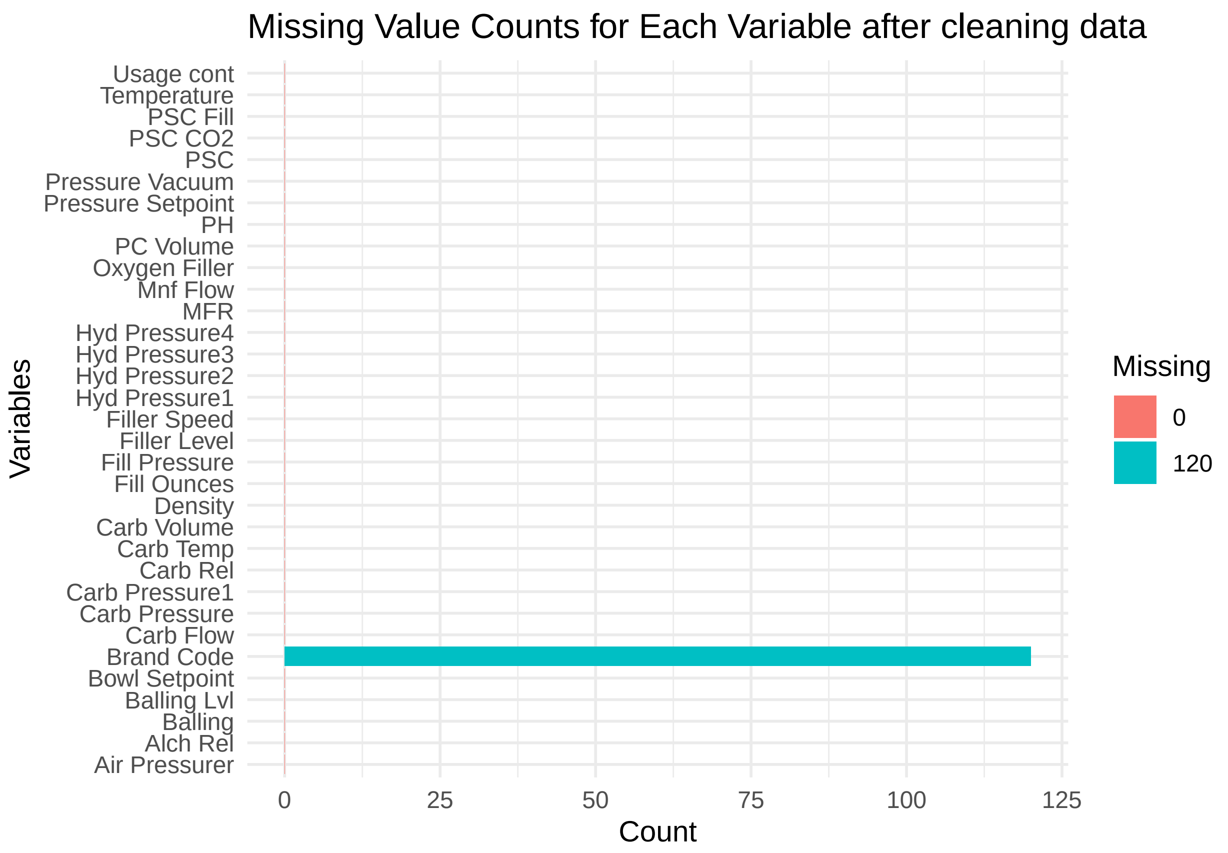
The clean_data function is a comprehensive data preprocessing pipeline that imputes missing values, removes outliers, scales features, and removes constant columns from the dataset. It is applied to the Student_Train dataset to produce a cleaned version, Student_Cleaned, ready for further analysis or modeling. The function uses various techniques like Predictive Mean Matching for imputation, Z-score normalization for scaling, and IQR for outlier detection.

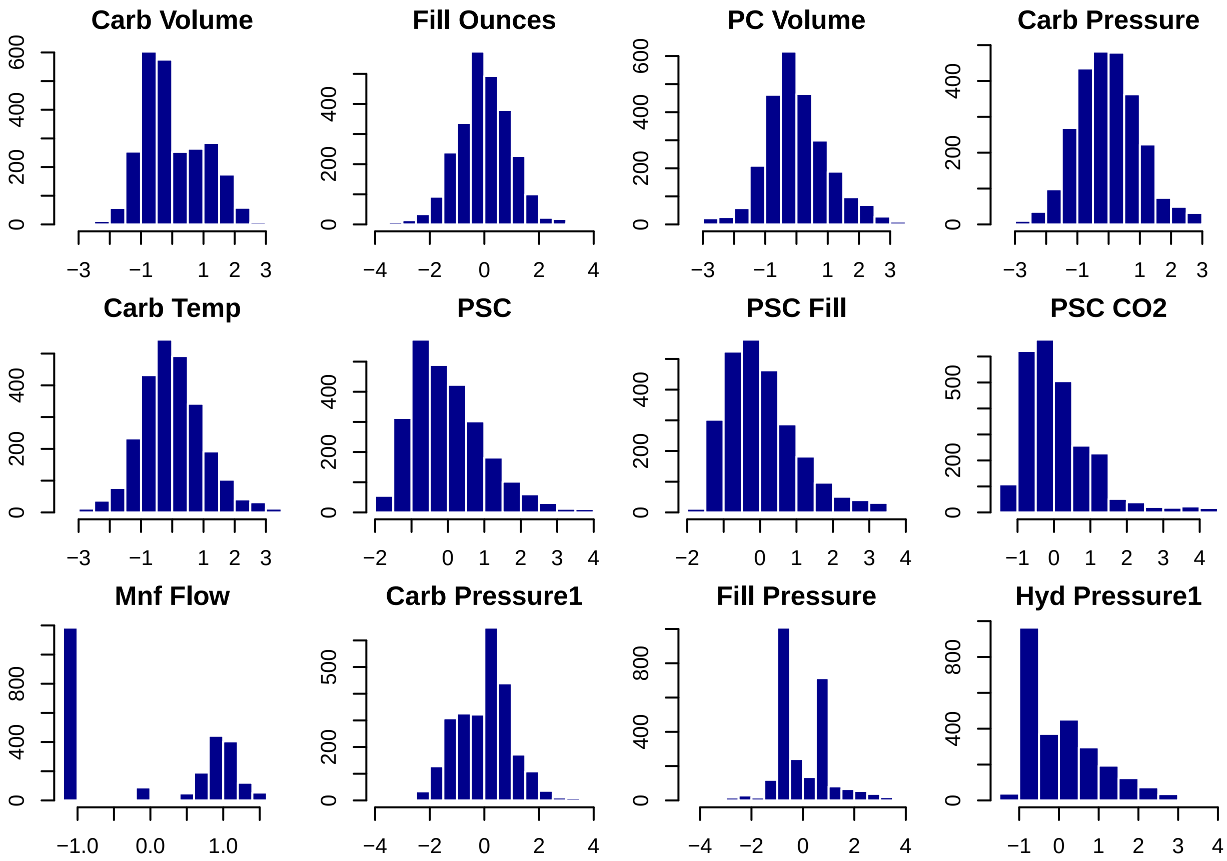
Validation of Cleaning

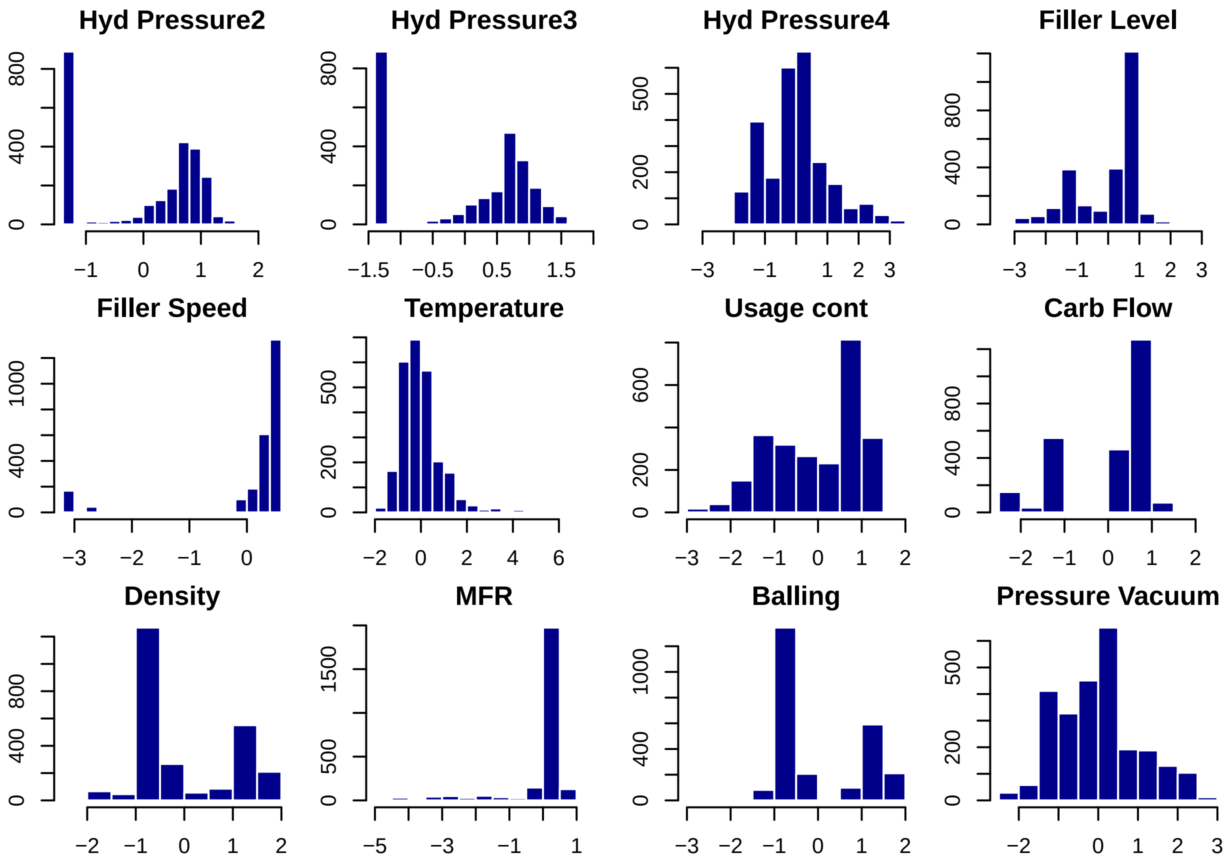
Checks to see if any missing values remain in the cleaned dataset.

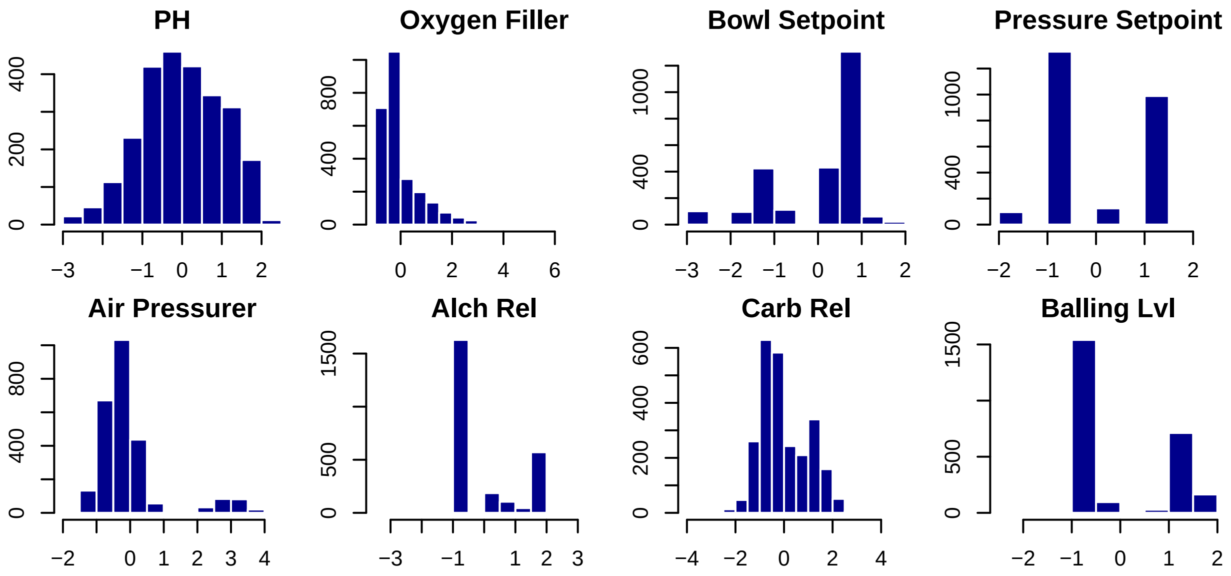
```
## Missing values per column after cleaning data:
```

variable	n	percent
Brand Code	120	4.7%

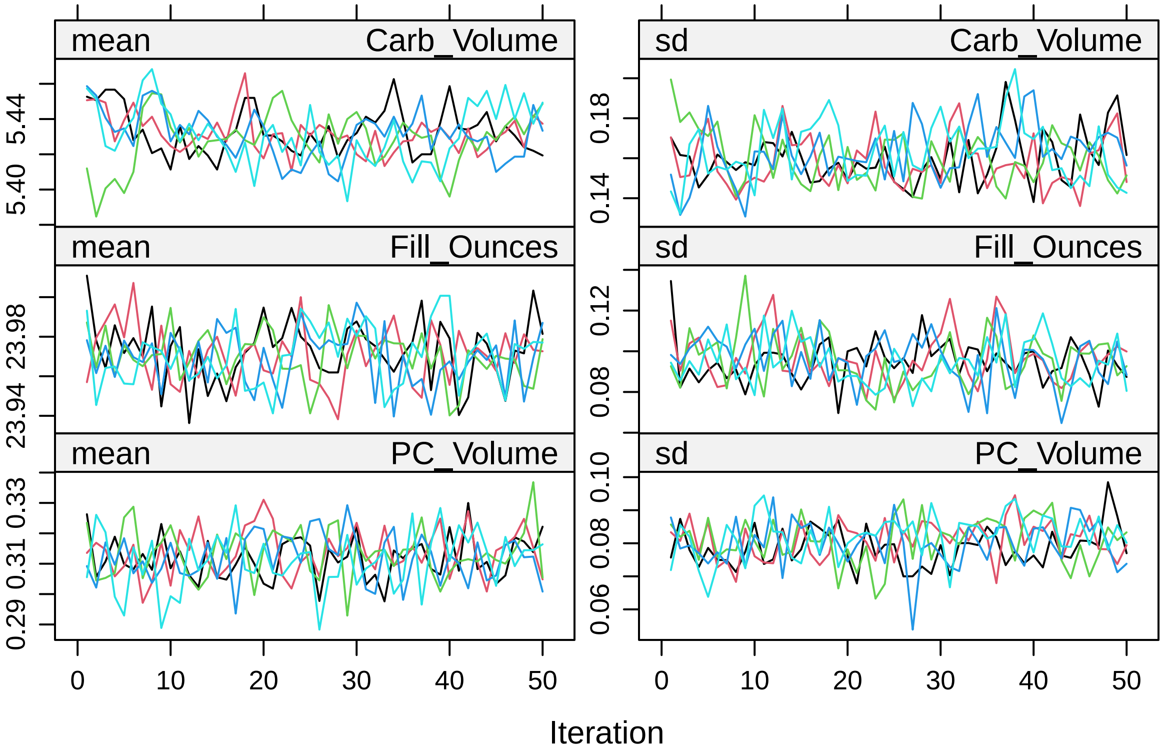

Missing Value Counts for Each Variable after cleaning data
Variables
Usage cont
Temperature
PSC Fill
PSC CO2
PSC
Pressure Vacuum
Pressure Setpoint
PH
PC Volume
Oxygen Filler
Mnf Flow
MFR
Hyd Pressure4
Hyd Pressure3
Hyd Pressure2
Hyd Pressure1
Filler Speed
Filler Level
Fill Pressure
Fill Ounces
Density
Carb Volume
Carb Temp
Carb Rel
Carb Pressure1
Carb Pressure
Carb Flow
Brand Code
Bowl Setpoint
Balling Lvl
Balling
Alch Rel
Air Pressurer
0
25
50
75
100
125
Count
Missing
0
120

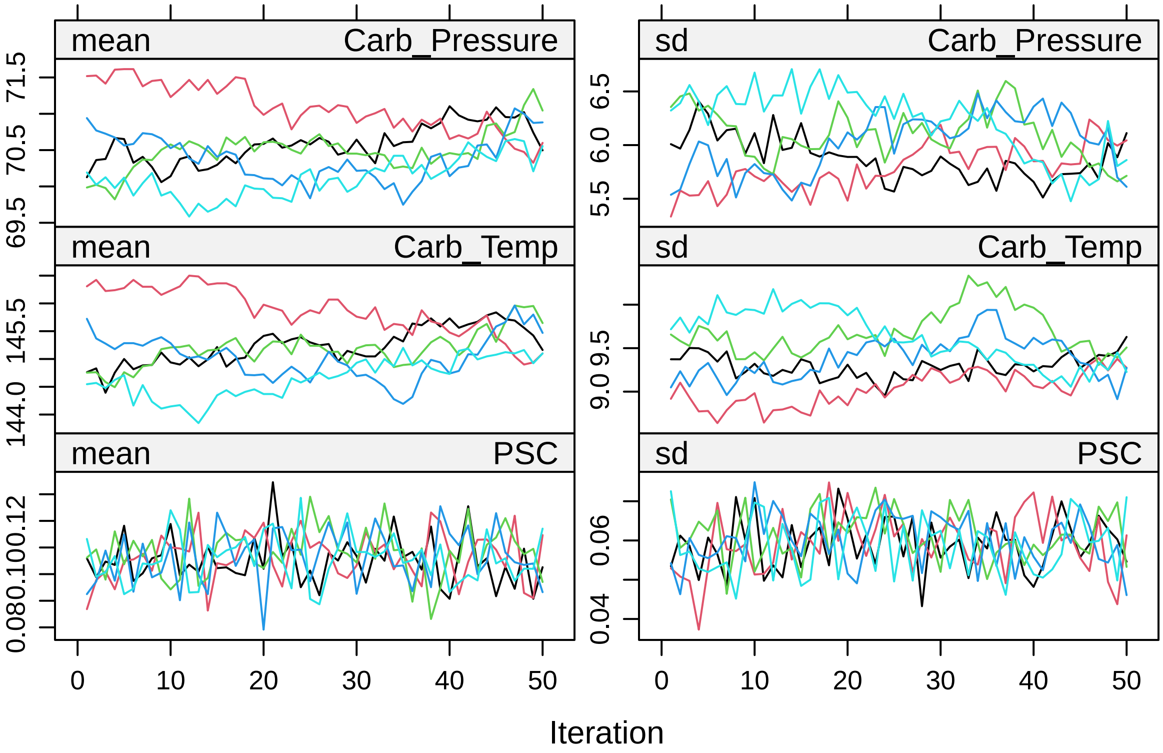

Carb Volume
0
200
400
600
−3
−1
1
2
3
Fill Ounces
0
200
400
−4
−2
0
2
4
PC Volume
0
200
400
600
−3
−1
1
2
3
Carb Pressure
0
200
400
−3
−1
1
2
3
Carb Temp
0
200
400
−3
−1
1
2
3
PSC
0
200
400
−2
0
1
2
3
4
PSC Fill
0
200
400
−2
0
1
2
3
4
PSC CO2
0
200
500
−1
0
1
2
3
4
Mnf Flow
0
400
800
−1.0
0.0
1.0
Carb Pressure1
0
200
500
−4
−2
0
2
4
Fill Pressure
0
400
800
−4
−2
0
2
4
Hyd Pressure1
0
400
800
−1
0
1
2
3
4

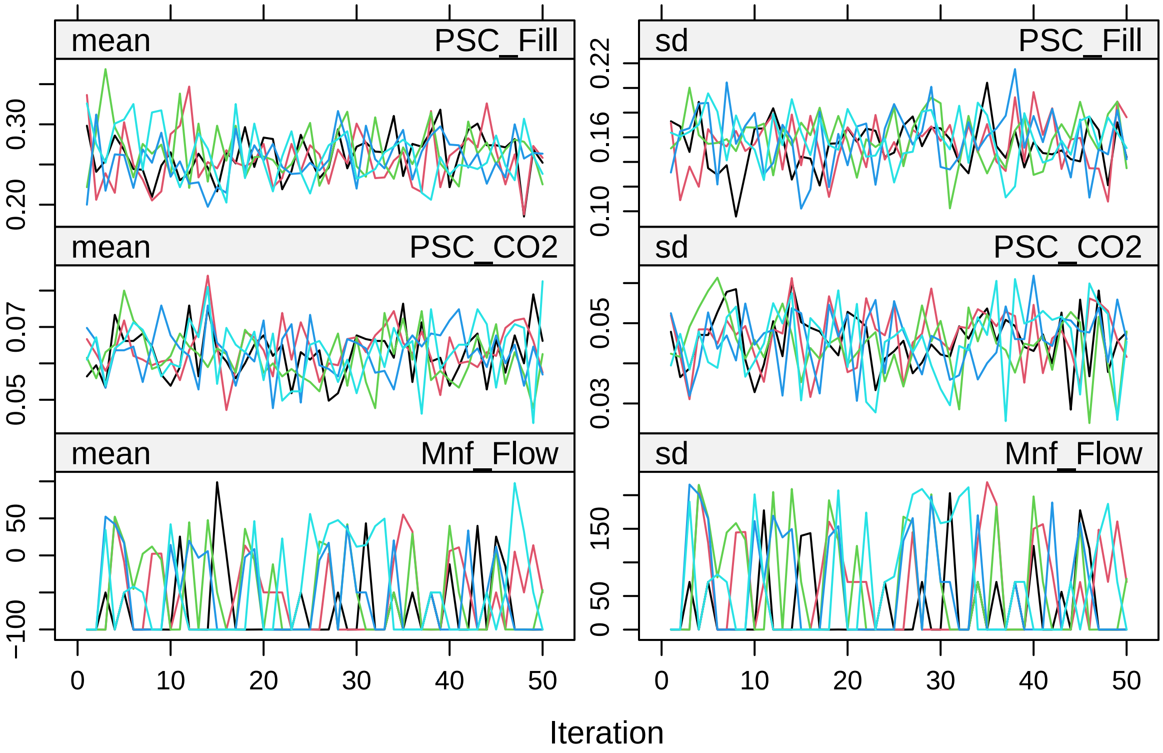

Hyd Pressure2
0
400
800
−1
0
1
2
Hyd Pressure3
0
400
800
−1.5
−0.5
0.5
1.5
Hyd Pressure4
0
200
500
−3
−1
1
2
3
Filler Level
0
400
800
−3
−1
1
2
3
Filler Speed
0
400
1000
−3
−2
−1
0
Temperature
0
200
500
−2
0
2
4
6
Usage cont
0
200
600
−3
−2
−1
0
1
2
Carb Flow
0
400
800
−2
−1
0
1
2
Density
0
400
800
−2
−1
0
1
2
MFR
0
500
1500
−5
−3
−1
1
Balling
0
400
1000
−3
−2
−1
0
1
2
Pressure Vacuum
0
200
500
−2
0
1
2
3



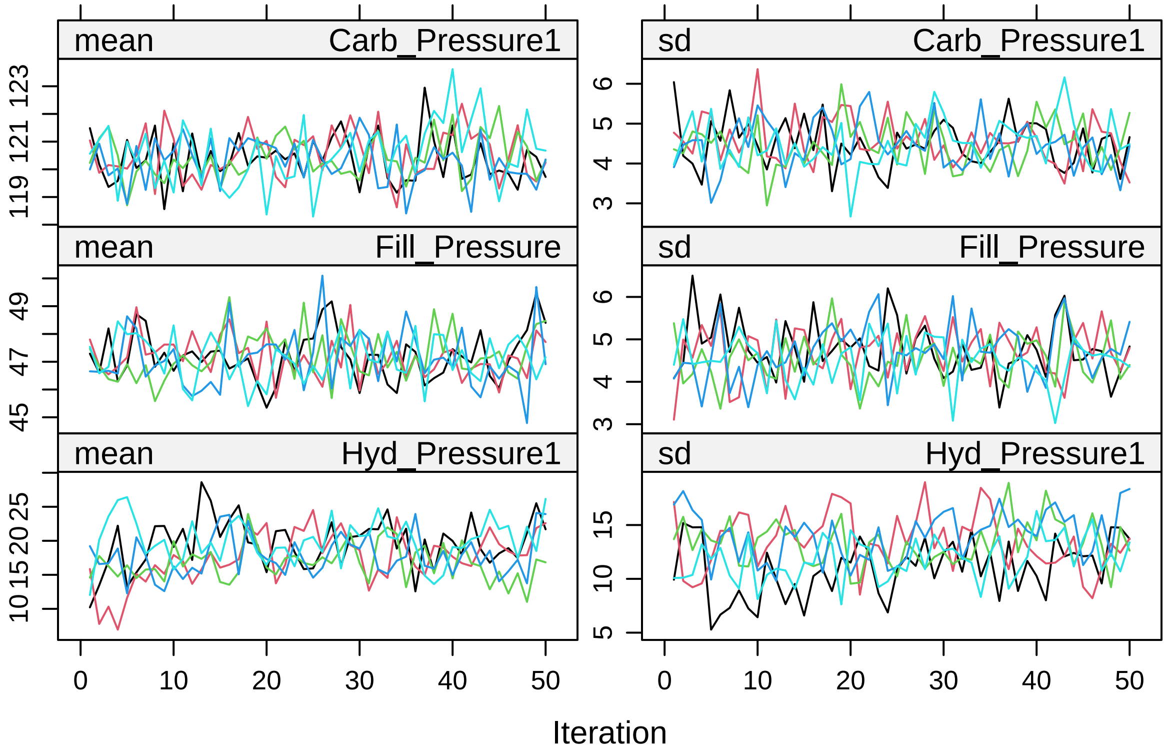
After cleaning the data we need to validate the dataset to check for remaining missing values as well as visualize the distributions to ensure scaling and outlier handling.The imputation process is visualized with a plot, and the density plot for the variable Carb_Volume compares the distribution of imputed and observed values.

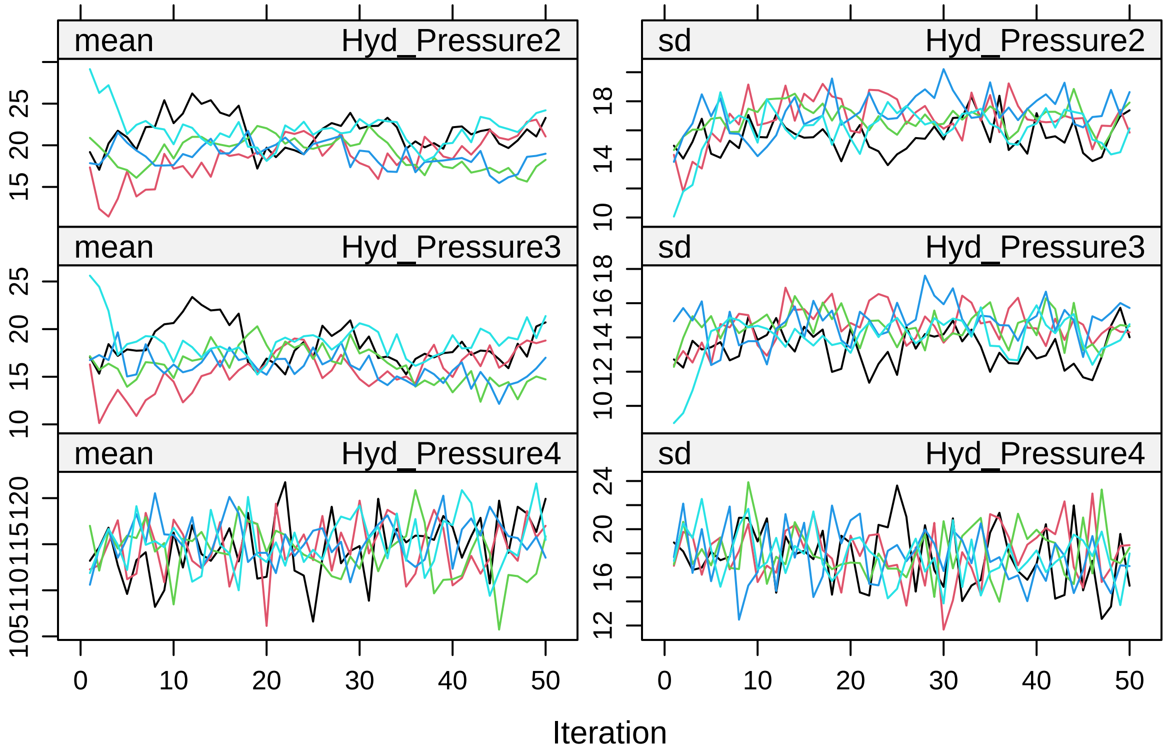

mean
Carb_Volume
sd
Carb_Volume
mean
Fill_Ounces
sd
Fill_Ounces
mean
PC_Volume
sd
PC_Volume
5.40
5.44
0.14
0.18
23.94
23.98
0.08
0.12
0.29
0.31
0.33
0.06
0.08
0.10
0
10
20
30
40
50
Iteration

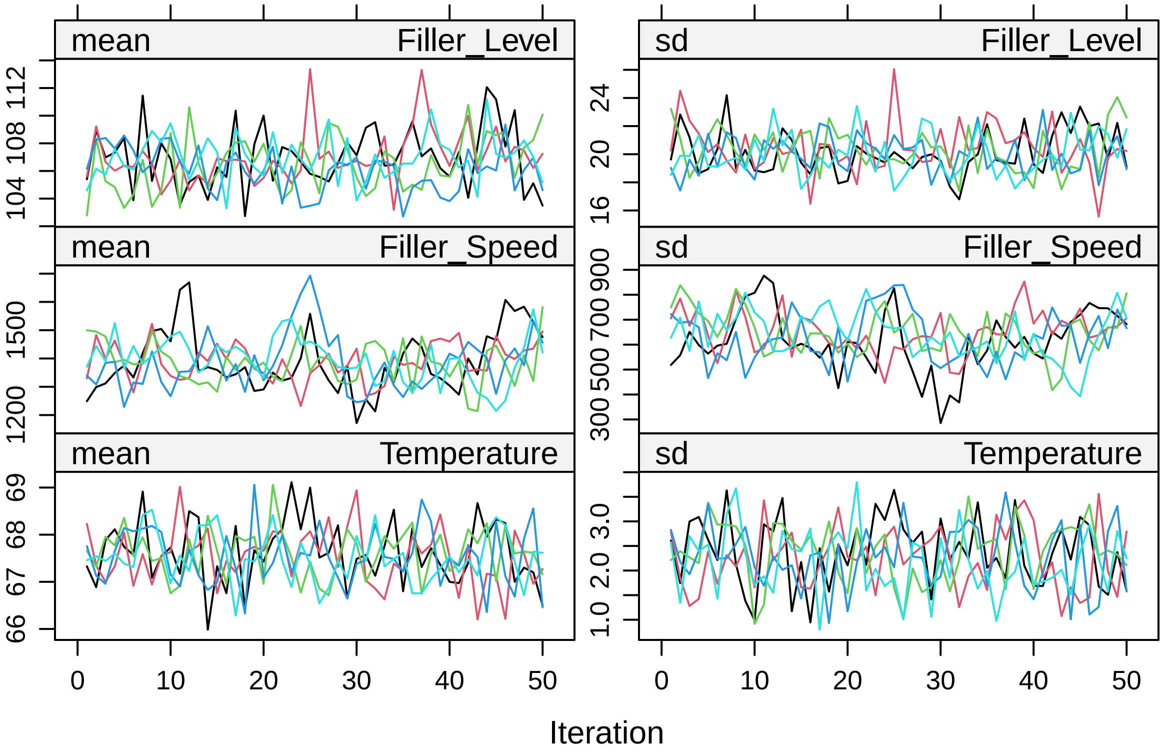

mean
Carb_Pressure
sd
Carb_Pressure
mean
Carb_Temp
sd
Carb_Temp
mean
PSC
sd
PSC
69.5
70.5
71.5
5.5
6.0
6.5
144.0
145.5
9.0
9.5
0.08
0.10
0.12
0.04
0.06
0
10
20
30
40
50
Iteration

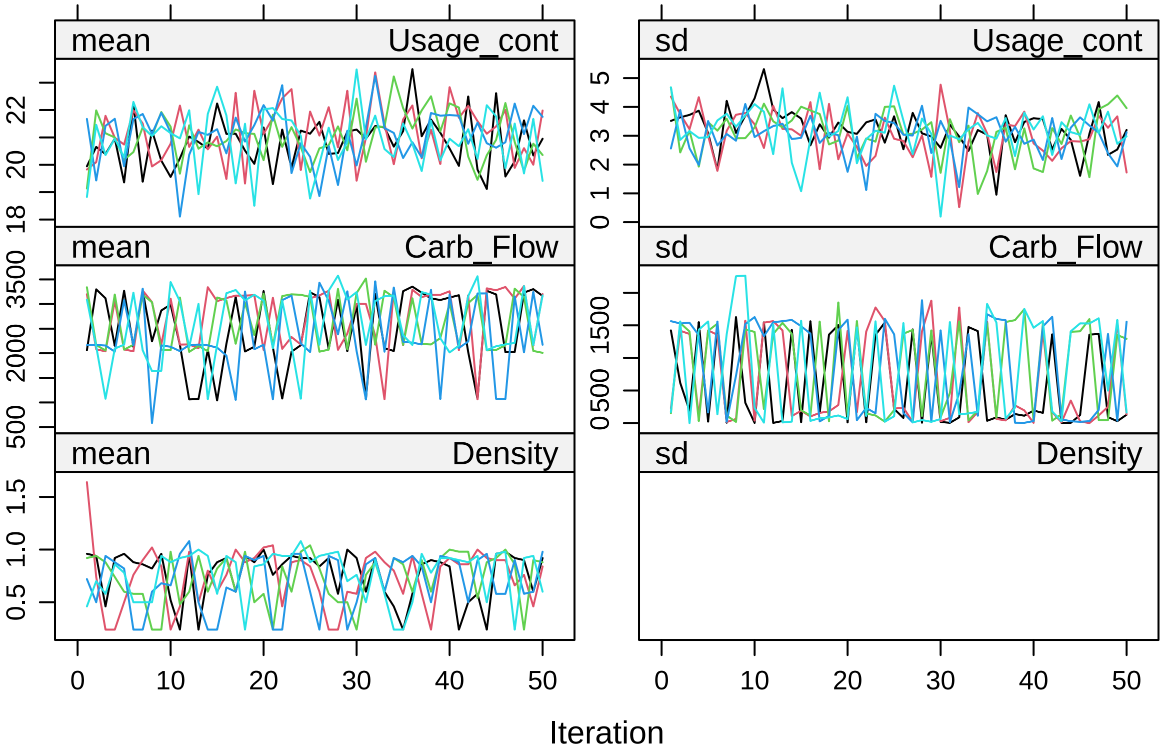


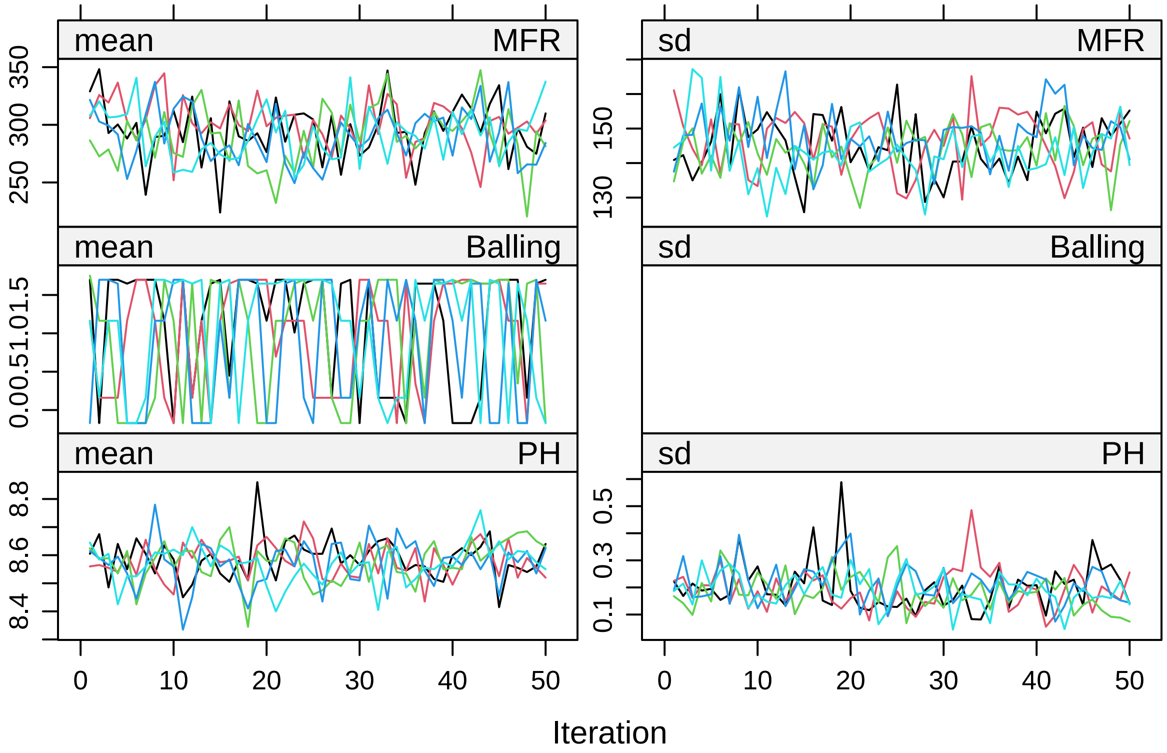
mean
PSC_Fill
sd
PSC_Fill
mean
PSC_CO2
sd
PSC_CO2
mean
Mnf_Flow
sd
Mnf_Flow
0.20
0.30
0.05
0.07
−100
0
50
0.10
0.16
0.22
0.03
0.05
0
50
150
0
10
20
30
40
50
Iteration

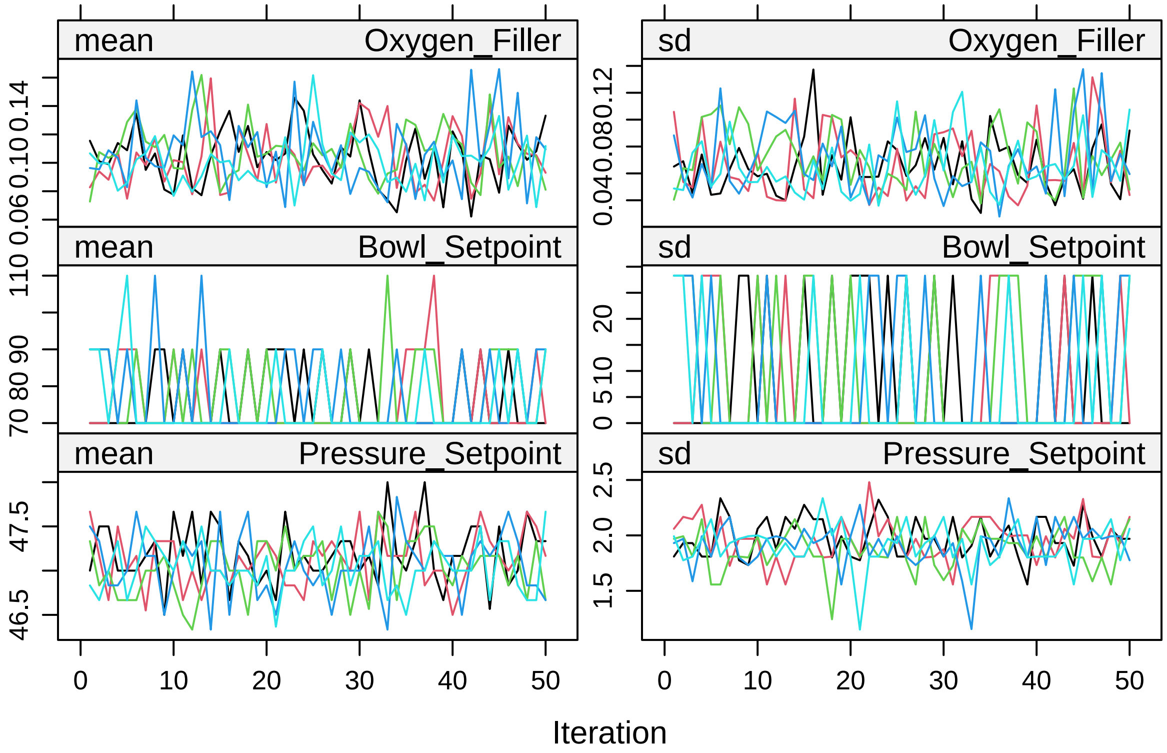

mean
Carb_Pressure1
sd
Carb_Pressure1
mean
Fill_Pressure
sd
Fill_Pressure
mean
Hyd_Pressure1
sd
Hyd_Pressure1
119 121 123
45 47 49
10 15 20 25
3 4 5 6
3 4 5 6
5 10 15
0 10 20 30 40 50
0 10 20 30 40 50
Iteration

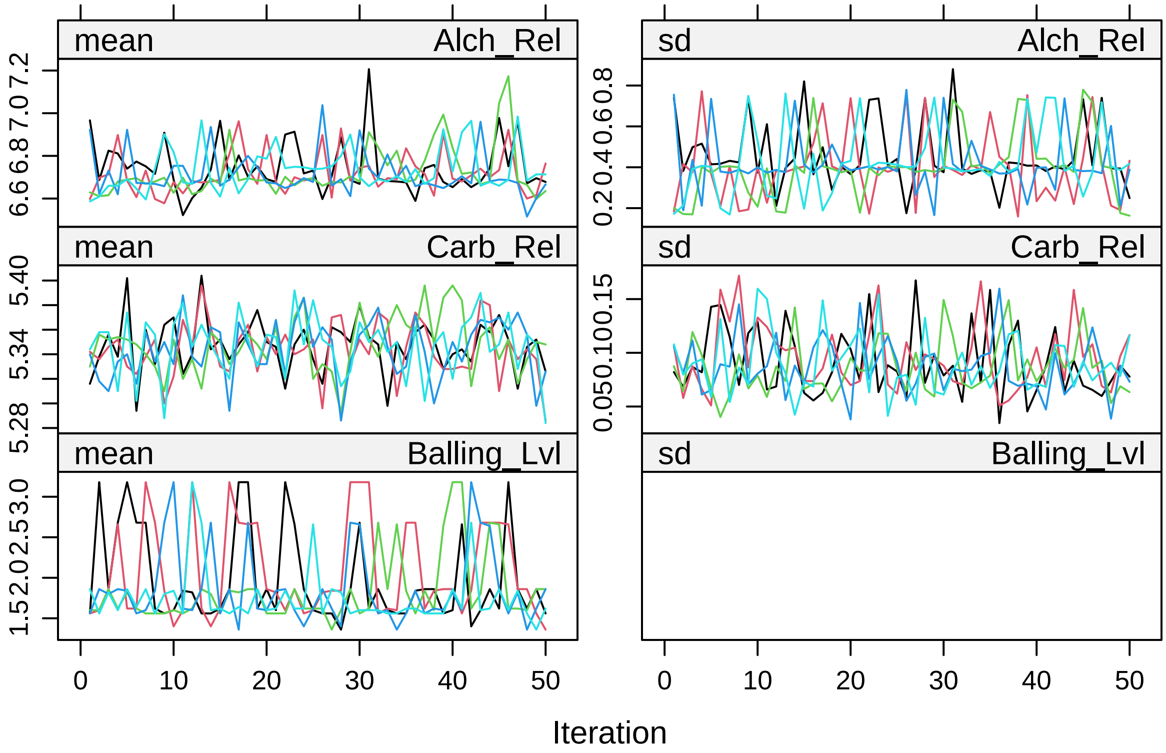

mean
Hyd_Pressure2
sd
Hyd_Pressure2
mean
Hyd_Pressure3
sd
Hyd_Pressure3
mean
Hyd_Pressure4
sd
Hyd_Pressure4
15
20
25
10
14
18
10
15
20
25
10
12
14
16
18
105
110
115
120
12
16
20
24
0
10
20
30
40
50
Iteration

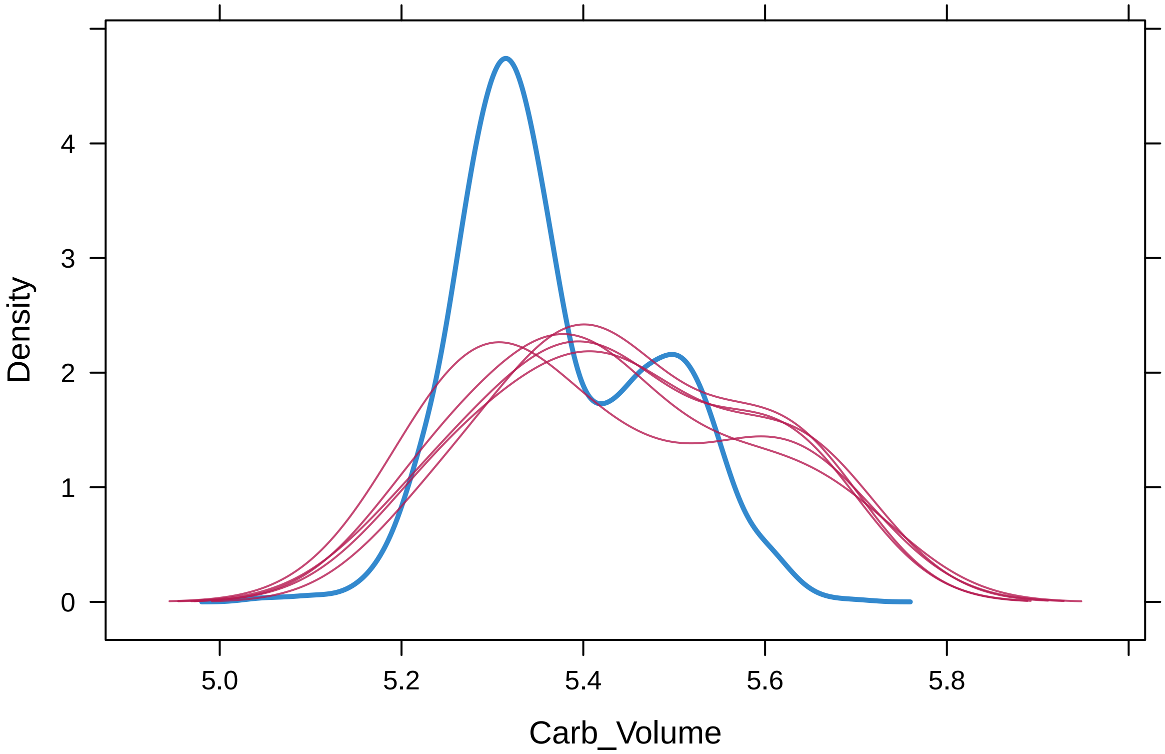

mean
Filler_Level
104
108
112
sd
Filler_Level
16
20
24
mean
Filler_Speed
1200
1500
sd
Filler_Speed
300
500
700
900
mean
Temperature
66
67
68
69
sd
Temperature
1.0
2.0
3.0
0
10
20
30
40
50
0
10
20
30
40
50
Iteration


mean
Usage_cont
sd
Usage_cont
mean
Carb_Flow
sd
Carb_Flow
mean
Density
sd
Density
18
20
22
0
1
2
3
4
5
500
2000
3500
0500
1500
0.5
1.0
1.5
0
10
20
30
40
50
Iteration


mean
MFR
sd
MFR
250
300
350
130
150
mean
Balling
sd
Balling
0.0
0.5
1.0
1.5
mean
PH
sd
PH
8.4
8.6
8.8
0.1
0.3
0.5
0
10
20
30
40
50
0
10
20
30
40
50
Iteration

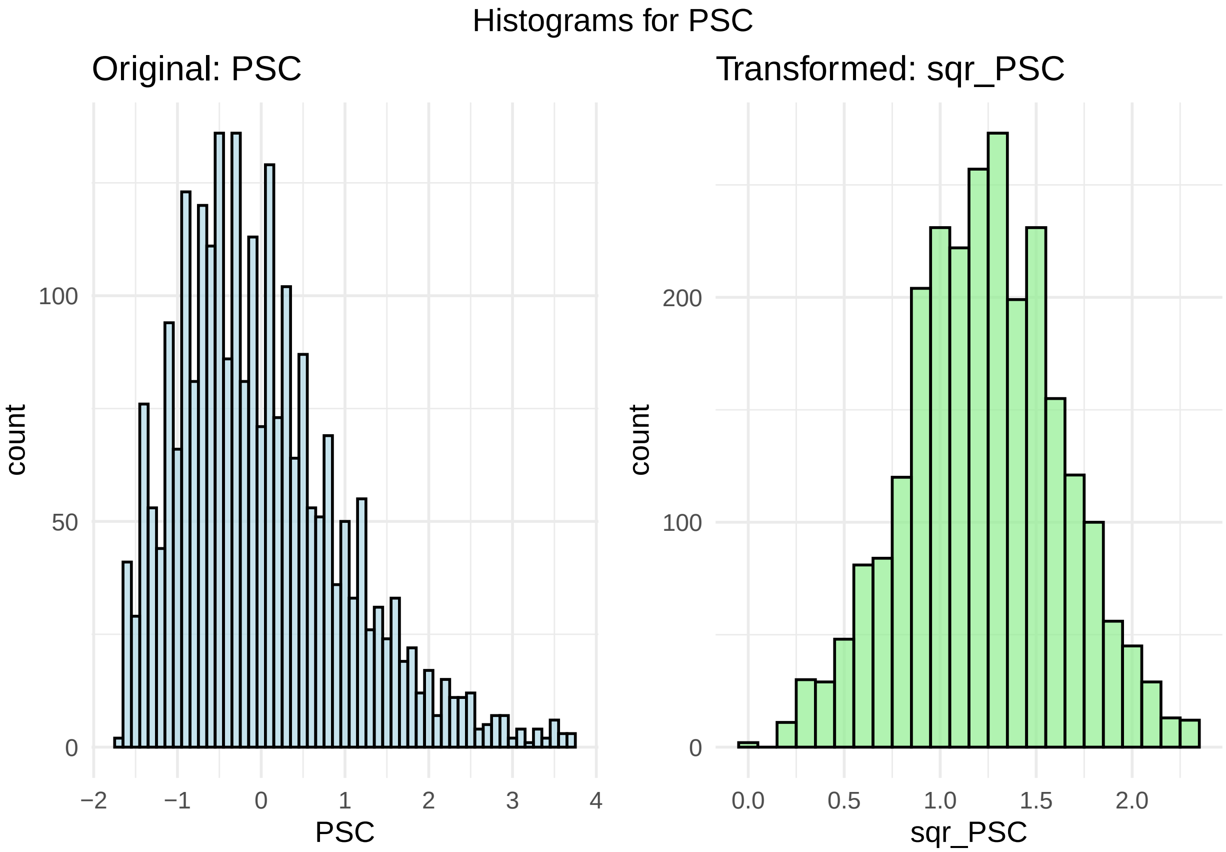

mean
Oxygen_Filler
sd
Oxygen_Filler
mean
Bowl_Setpoint
sd
Bowl_Setpoint
mean
Pressure_Setpoint
sd
Pressure_Setpoint
0.06
0.10
0.14
0.04
0.08
0.12
70
80
90
110
0
5
10
20
46.5
47.5
1.5
2.0
2.5
0
10
20
30
40
50
Iteration

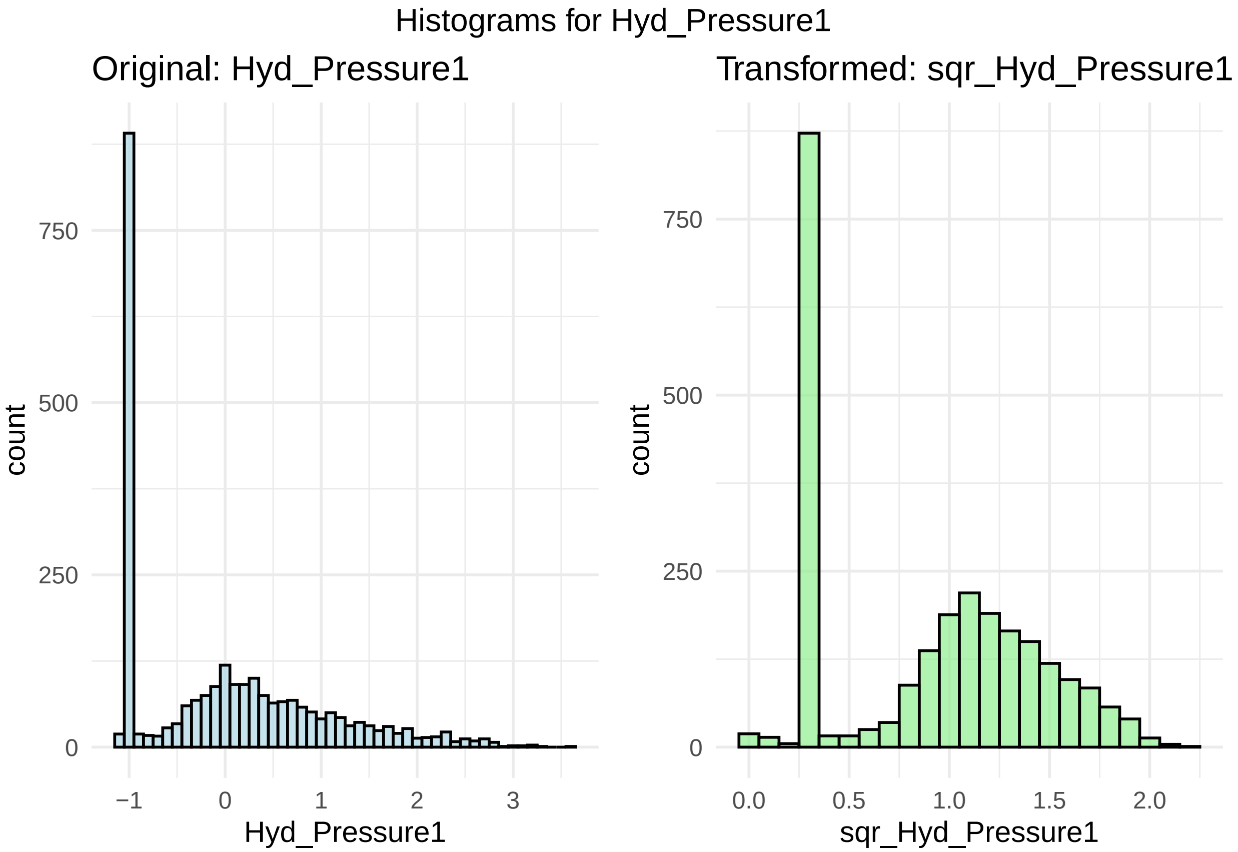

mean
Alch_Rel
6.6
6.8
7.0
7.2
sd
Alch_Rel
0.2
0.4
0.6
0.8
mean
Carb_Rel
5.28
5.34
5.40
sd
Carb_Rel
0.05
0.10
0.15
mean
Balling_Lvl
1.5
2.0
2.5
3.0
sd
Balling_Lvl
0
10
20
30
40
50
Iteration

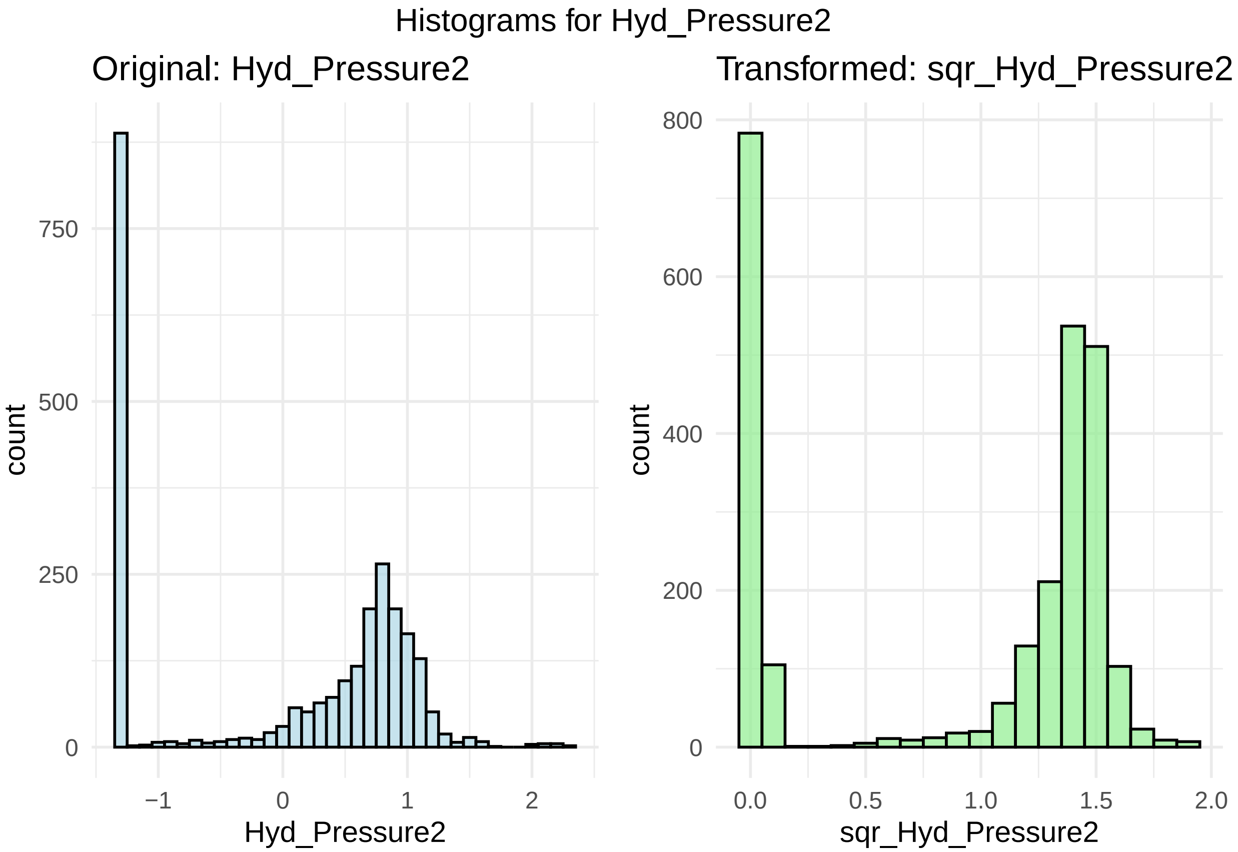


Transformations For the right skewed variables we will use square root transformation

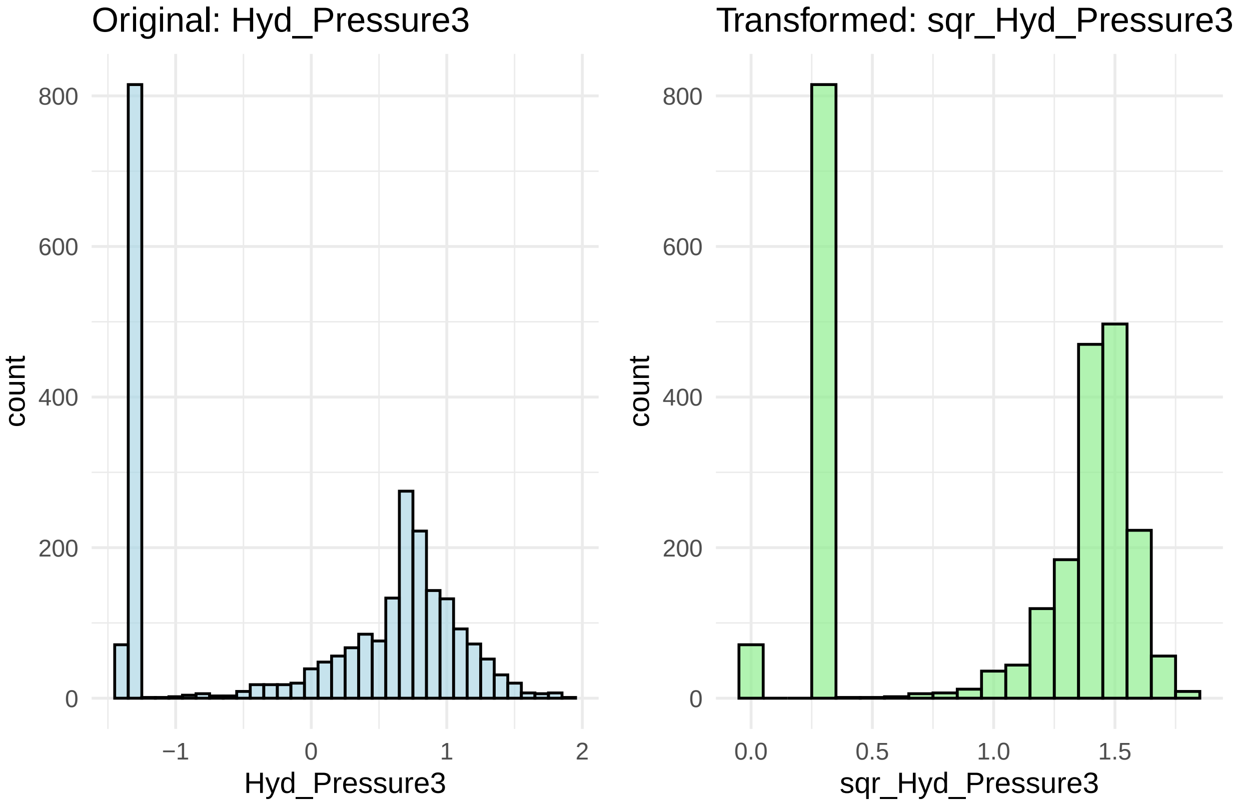
```
## Warning: package 'gridExtra' was built under R version 4.3.3

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:randomForest':
##
##     combine

## The following object is masked from 'package:dplyr':
##
##     combine

## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

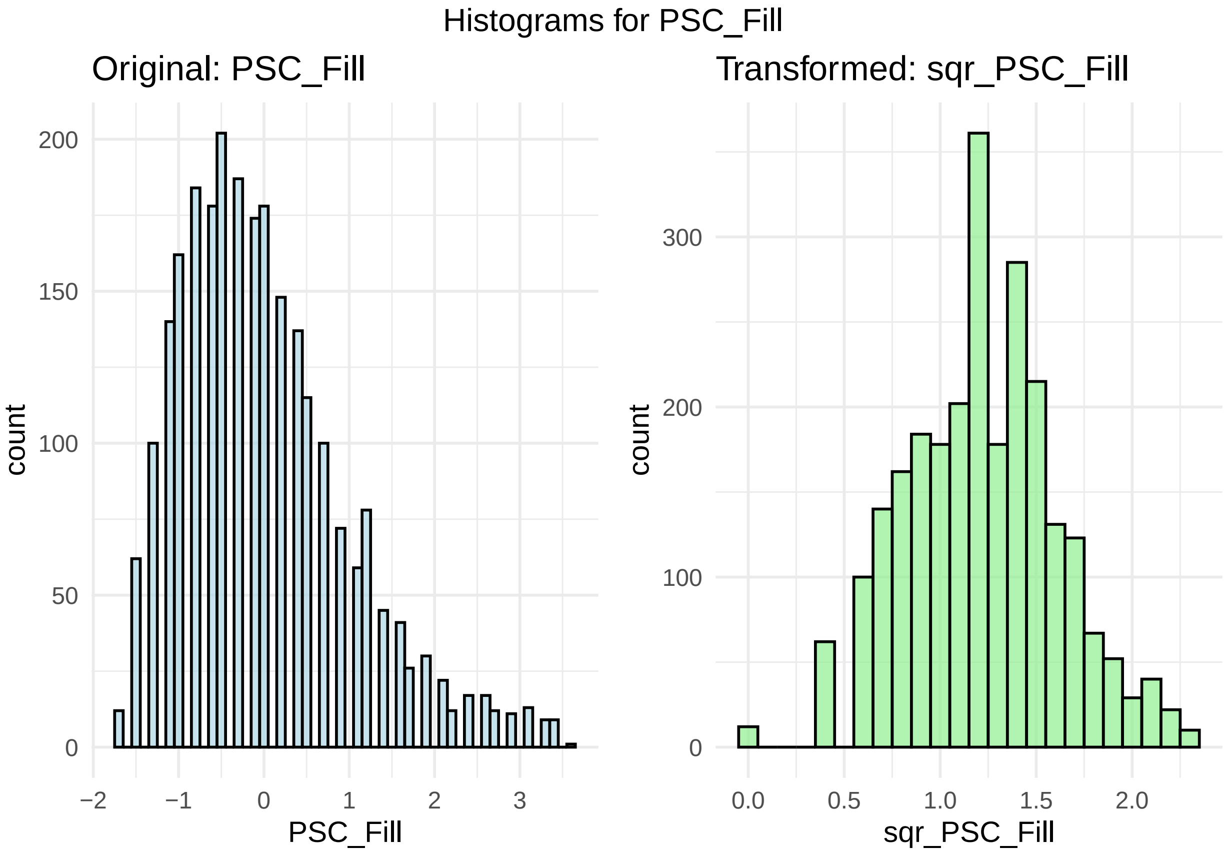

Histograms for PSC
Original: PSC
Transformed: sqr_PSC
count
100
50
0
−2
−1
0
1
2
3
4
PSC
count
200
100
0
0.0
0.5
1.0
1.5
2.0
sqr_PSC

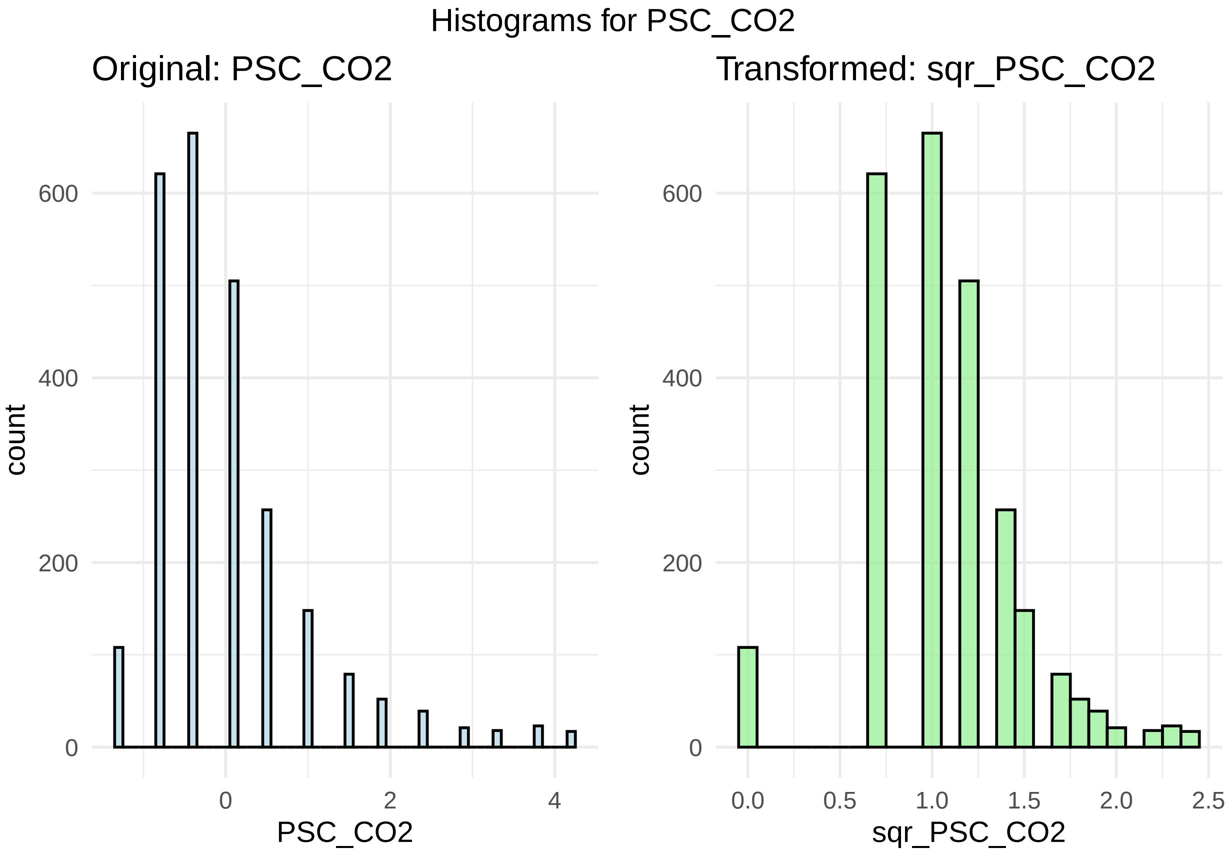

Histograms for Hyd_Pressure1
Original: Hyd_Pressure1
Transformed: sqr_Hyd_Pressure1
count
750
500
250
0
−1
0
1
2
3
Hyd_Pressure1
count
750
500
250
0
0.0
0.5
1.0
1.5
2.0
sqr_Hyd_Pressure1

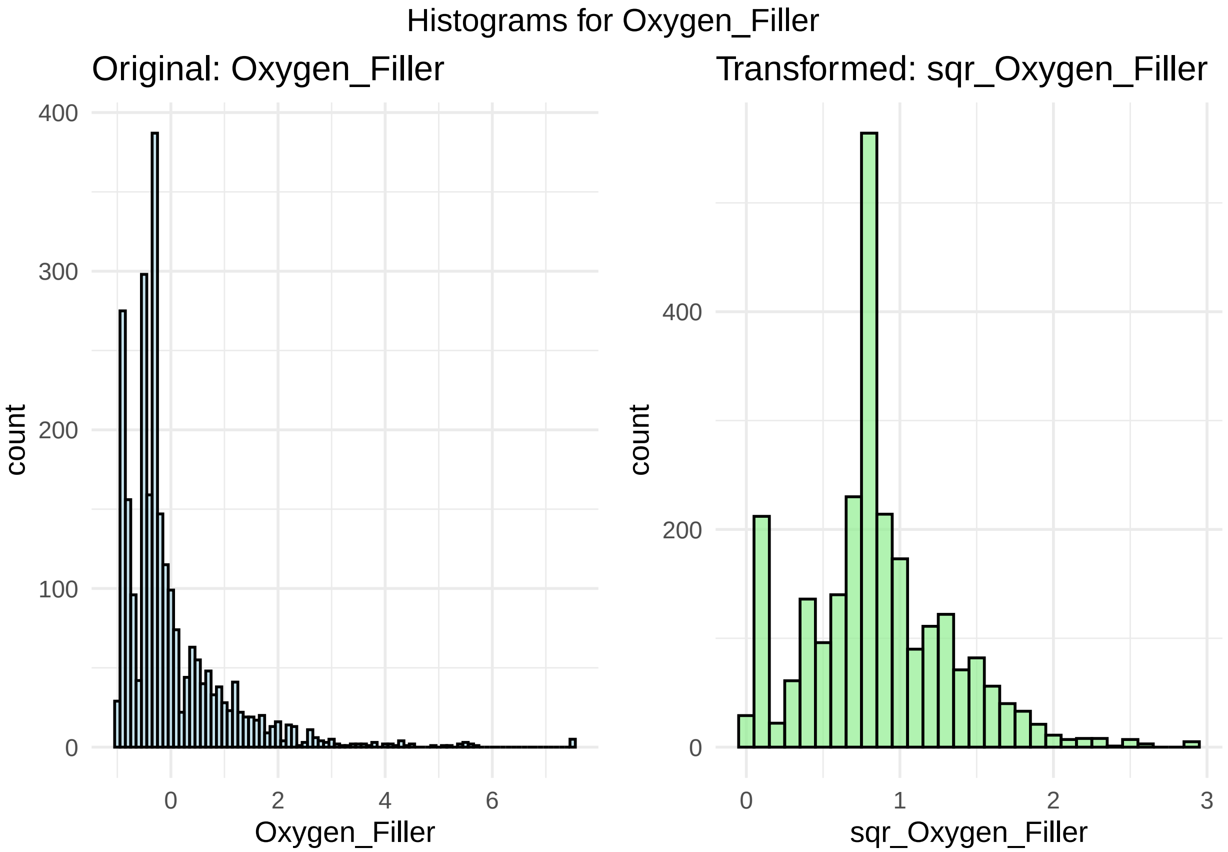

Histograms for Hyd_Pressure2
Original: Hyd_Pressure2
Transformed: sqr_Hyd_Pressure2
count
750
500
250
0
−1
0
1
2
Hyd_Pressure2
count
800
600
400
200
0
0.0
0.5
1.0
1.5
2.0
sqr_Hyd_Pressure2

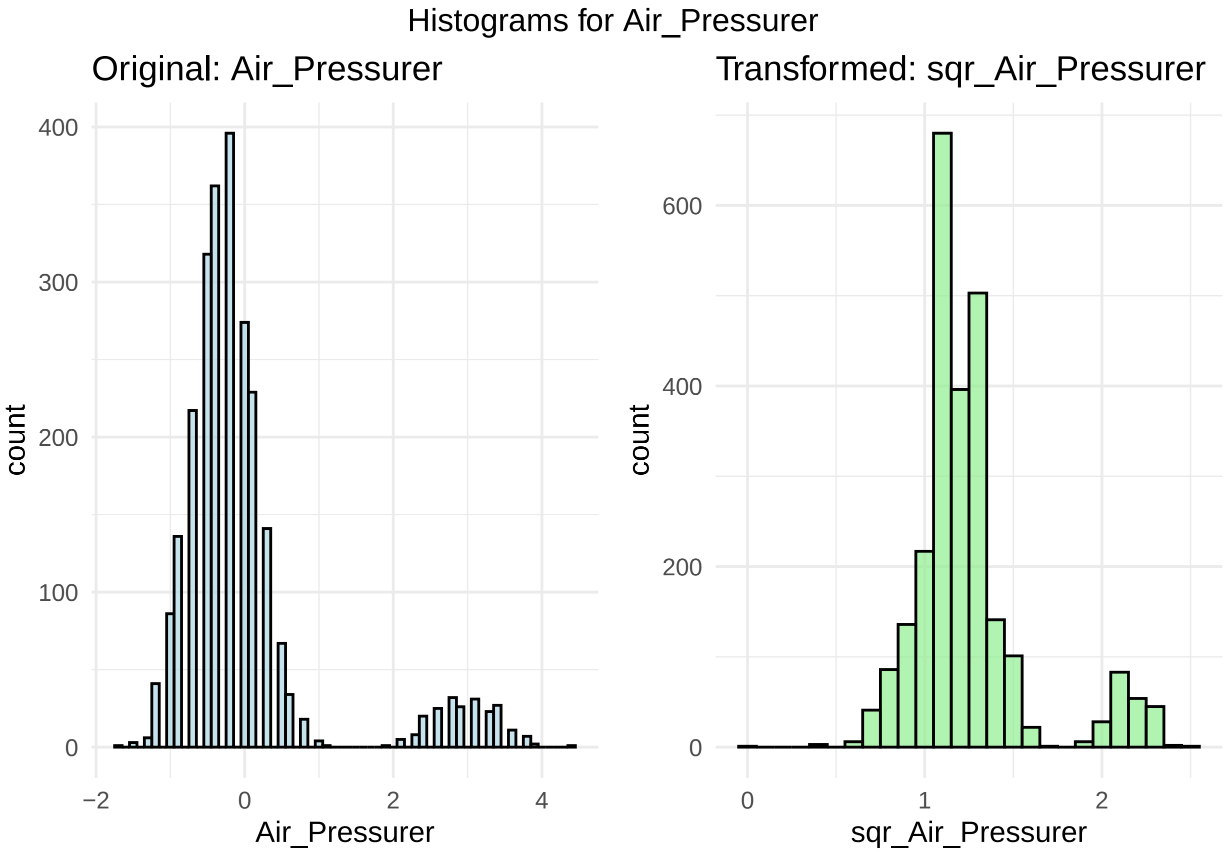
Histograms for Hyd_Pressure3




Histograms for PSC_Fill
Original: PSC_Fill
Transformed: sqr_PSC_Fill
count
200
150
100
50
0
−2
−1
0
1
2
3
PSC_Fill
count
300
200
100
0
0.0
0.5
1.0
1.5
2.0
sqr_PSC_Fill


Histograms for PSC_CO2
Original: PSC_CO2
Transformed: sqr_PSC_CO2
count
0
200
400
600
PSC_CO2
0
2
4
count
0
200
400
600
sqr_PSC_CO2
0.0
0.5
1.0
1.5
2.0
2.5


Histograms for Oxygen_Filler
Original: Oxygen_Filler
Transformed: sqr_Oxygen_Filler
count
400
300
200
100
0
0
2
4
6
Oxygen_Filler
count
400
200
0
0
1
2
3
sqr_Oxygen_Filler



Model Building

Non-Linear Models:

Split the Data 80-20 in training and testing data

KNN

```
##      RMSE  Rsquared       MAE
## 0.8084898 0.3603004 0.6118999
```

SVM

```
##      RMSE  Rsquared       MAE
## 0.7438239 0.4575773 0.5459960
```

MARS

```
## Loading required package: earth
```

```
## Warning: package 'earth' was built under R version 4.3.3
```

```
## Loading required package: Formula
```

```
## Loading required package: plotmo
## Warning: package 'plotmo' was built under R version 4.3.3
```

```
## Loading required package: plotrix
```

```
##      RMSE  Rsquared       MAE 
## 0.7943078 0.3822277 0.5981025
```

Random Forest

```
##      RMSE  Rsquared       MAE 
## 0.6230843 0.6341422 0.4465616
```

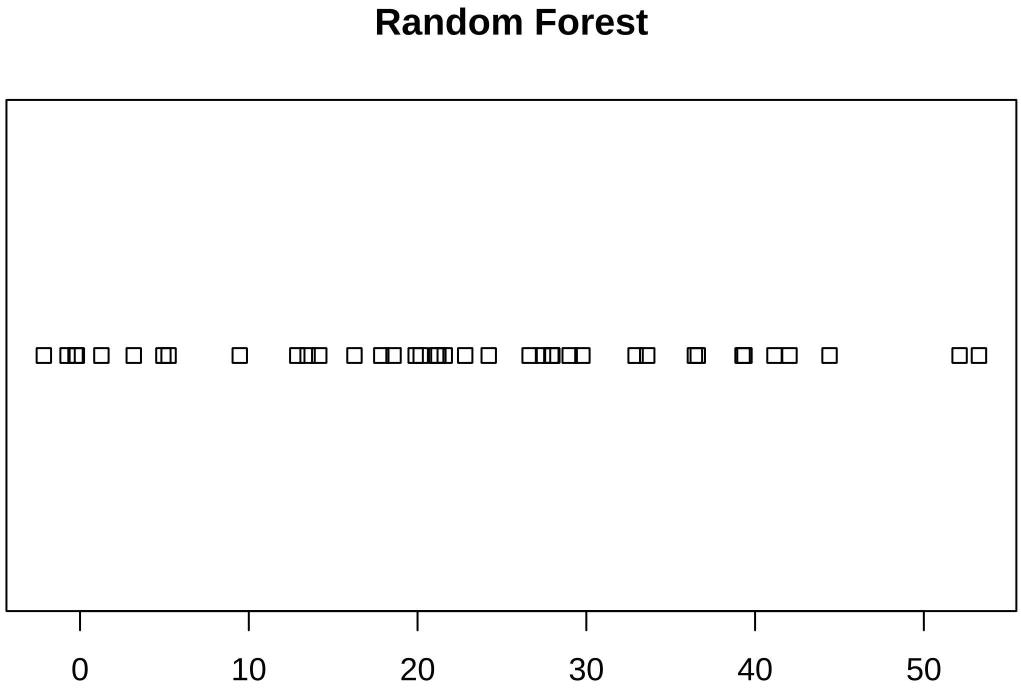
Model Evaluation

```
##                   RMSE  Rsquared       MAE
## KNN          0.8084898 0.3603004 0.6118999
## SVM          0.7438239 0.4575773 0.5459960
## MARS         0.7943078 0.3822277 0.5981025
## RandomForest 0.6230843 0.6341422 0.4465616
```

The non-linear model with the best opimal resampling and test set performance of the four is the Random Forest model. It has the lowest RMSE (a lower RMSE predicts a better performing model) of all the models at: 0.61. It also has the highest Rsquared (a higher Rsquared indicates a better fit) of all the models at 0.64.

Variable Importance

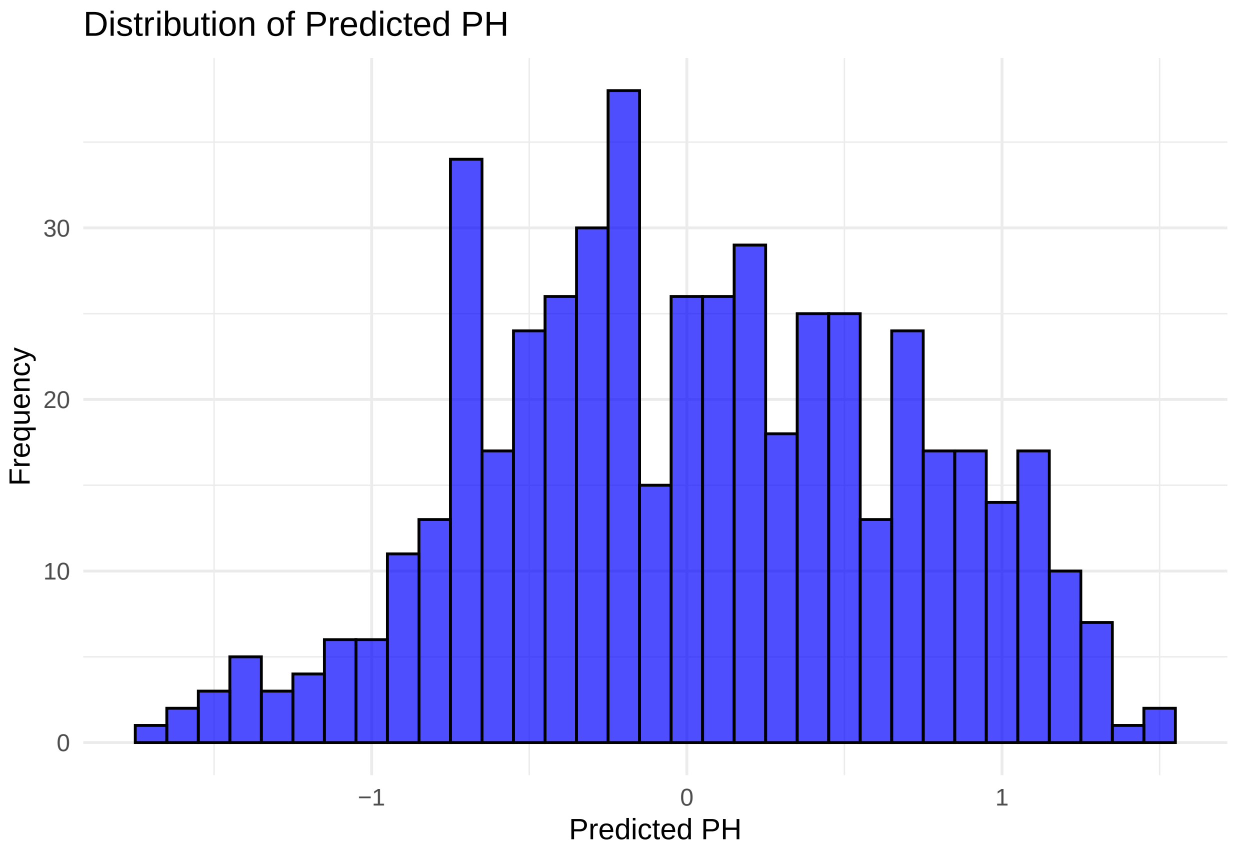
```
##                     Overall
## Carb_Volume      21.6109065
## Fill_Ounces       9.4644973
## PC_Volume        16.2618845
## Carb_Pressure     5.2460040
## Carb_Temp         1.2659327
## PSC               3.1844140
## PSC_Fill         -0.1894354
## PSC_CO2          -0.7388580
## Mnf_Flow         53.2651771
## Carb_Pressure1   29.7683214
## Fill_Pressure    18.5753631
## Hyd_Pressure1    24.2164388
## Hyd_Pressure2    13.4829036
## Hyd_Pressure3    17.8509043
## Hyd_Pressure4    21.2338525
## Filler_Level     22.8215598
## Filler_Speed     29.0142514
## Temperature      52.1176388
## Usage_cont       39.2573111
## Carb_Flow        32.9188973
## Density          27.4348921
## MFR              19.8880467
## Balling          33.6081232
## Pressure_Vacuum  42.0386016
## Oxygen_Filler     36.4329951
## Bowl_Setpoint     26.6399529
## Pressure_Setpoint 14.1736939
## Air_Pressurer     27.9276997
## Alch_Rel          41.1485342
## Carb_Rel          44.4109680
## Balling_Lvl       39.3698205
## sqr_PSC            4.9395307
## sqr_PSC_Fill      -2.1446992
## sqr_PSC_CO2       -0.2671276
## sqr_Hyd_Pressure1 21.1028419
## sqr_Hyd_Pressure2 12.8716246
## sqr_Hyd_Pressure3 20.1869782
## sqr_Oxygen_Filler 36.6042546
## sqr_Air_Pressurer 27.9872572
```



For our optimal Random Forest model, Mnf Flow enjoys the greatest impact (54.42), above others like Carb Pressure1 (27.02) and Carb Volume (21.26), all of which contribute meaningfully to predicting PH. Minimal contributors such as PSC Fill and PSC CO2 can be diesregarded in follow-up analysis, as we should focus optimization strategies on top predictors. We next employed our Random Forest model to form a prediction.

Forecast

```
## # A tibble: 509 x 1
##    Predicted_PH
##           <dbl>
##  1        0.140
##  2        0.208
##  3       -0.0260
##  4       -0.177
##  5       -0.270
##  6       -0.656
##  7       -0.528
##  8        0.0719
##  9       -0.0234
## 10       -0.308
## # i 499 more rows
```



The output would indicate concentration of the predictions around around 0, largely symetric in disttribution and concentration toward the mean, all of which suggests that our model is accurate in predicting PH values within a fairly narrow range.