DATA 624 Spring 2019: Project-2

Ahmed Sajjad, Harpreet Shoker, Jagruti Solao, Chad Smith, Todd Weigel

April 29, 2019

Loading the data set

```
project_df <-read.csv("StudentData.csv",sep=",",header=TRUE,stringsAsFactors=FALSE)
dim(project_df)
```

```
## [1] 2571   33
```

From the above result dataset has 33 variables and 2571 observations.so we have 32 predictor variables and 1 target variable. Lets look at summary of data

```
summary(project_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
##  Min.   : 55.8   Min.   : 998   Min.   :63.60   Min.   :12.08
##  1st Qu.: 98.3   1st Qu.:3888   1st Qu.:65.20   1st Qu.:18.36
##  Median :118.4   Median :3982   Median :65.60   Median :21.79
##  Mean   :109.3   Mean   :3687   Mean   :65.97   Mean   :20.99
```

```
##  3rd Qu.:120.0   3rd Qu.:3998   3rd Qu.:66.40   3rd Qu.:23.75
##  Max.   :161.2   Max.   :4030   Max.   :76.20   Max.   :25.90
##  NA's   :20      NA's   :57     NA's   :14      NA's   :5
##    Carb.Flow       Density           MFR           Balling
##  Min.   :  26   Min.   :0.240   Min.   : 31.4   Min.   :-0.170
##  1st Qu.:1144   1st Qu.:0.900   1st Qu.:706.3   1st Qu.: 1.496
##  Median :3028   Median :0.980   Median :724.0   Median : 1.648
##  Mean   :2468   Mean   :1.174   Mean   :704.0   Mean   : 2.198
##  3rd Qu.:3186   3rd Qu.:1.620   3rd Qu.:731.0   3rd Qu.: 3.292
##  Max.   :5104   Max.   :1.920   Max.   :868.6   Max.   : 4.012
##  NA's   :2      NA's   :1       NA's   :212     NA's   :1
##  Pressure.Vacuum        PH        Oxygen.Filler     Bowl.Setpoint
##  Min.   :-6.600   Min.   :7.880   Min.   :0.00240   Min.   : 70.0
##  1st Qu.:-5.600   1st Qu.:8.440   1st Qu.:0.02200   1st Qu.:100.0
##  Median :-5.400   Median :8.540   Median :0.03340   Median :120.0
##  Mean   :-5.216   Mean   :8.546   Mean   :0.04684   Mean   :109.3
##  3rd Qu.:-5.000   3rd Qu.:8.680   3rd Qu.:0.06000   3rd Qu.:120.0
##  Max.   :-3.600   Max.   :9.360   Max.   :0.40000   Max.   :140.0
##                   NA's   :4       NA's   :12        NA's   :2
##  Pressure.Setpoint Air.Pressurer      Alch.Rel        Carb.Rel
##  Min.   :44.00     Min.   :140.8   Min.   :5.280   Min.   :4.960
##  1st Qu.:46.00     1st Qu.:142.2   1st Qu.:6.540   1st Qu.:5.340
##  Median :46.00     Median :142.6   Median :6.560   Median :5.400
##  Mean   :47.62     Mean   :142.8   Mean   :6.897   Mean   :5.437
##  3rd Qu.:50.00     3rd Qu.:143.0   3rd Qu.:7.240   3rd Qu.:5.540
##  Max.   :52.00     Max.   :148.2   Max.   :8.620   Max.   :6.060
##  NA's   :12                        NA's   :9       NA's   :10
##   Balling.Lvl
##  Min.   :0.00
##  1st Qu.:1.38
##  Median :1.48
##  Mean   :2.05
##  3rd Qu.:3.14
##  Max.   :3.66
##  NA's   :1
```

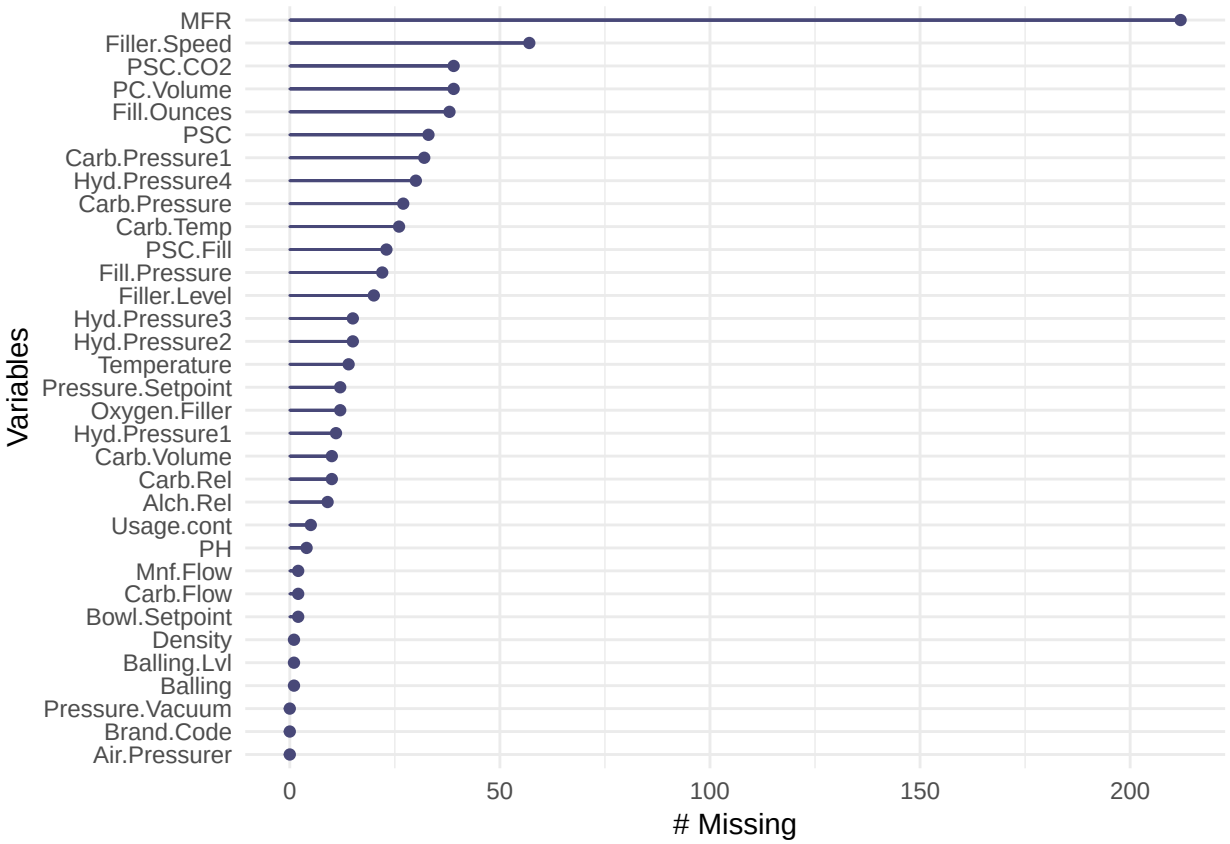
```
#str(project_df)
```

From the summary results - The variable Brandcode is a character type .All the other variables are numeric and integers. We can see null values in all variables except Air.Pressure and Pressure.Vacuum.We can have a look through plot

```
project_df[, c(16, 18, 21, 28)] <- sapply(project_df[, c(16, 18, 21, 28)], as.numeric) # converting int
```

Lets look at the missing values in the dataset

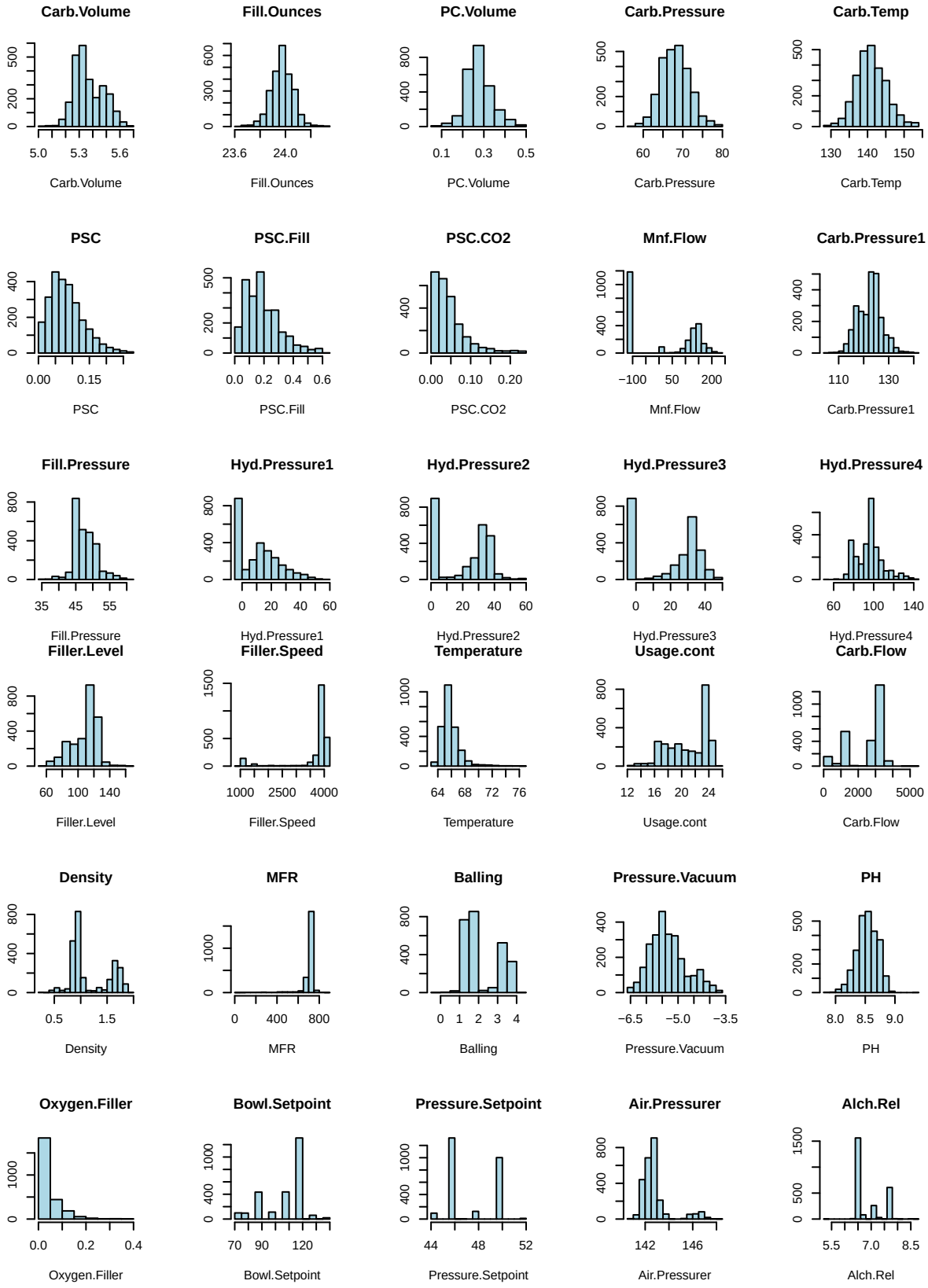
```
library(naniar)
gg_miss_var(project_df)
```

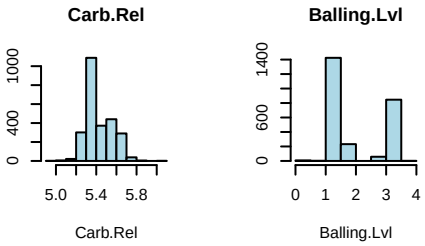


The above plot shows that MFR has most missing values.

To have a better look to see distribution of the variables lets plot histogram of all variables

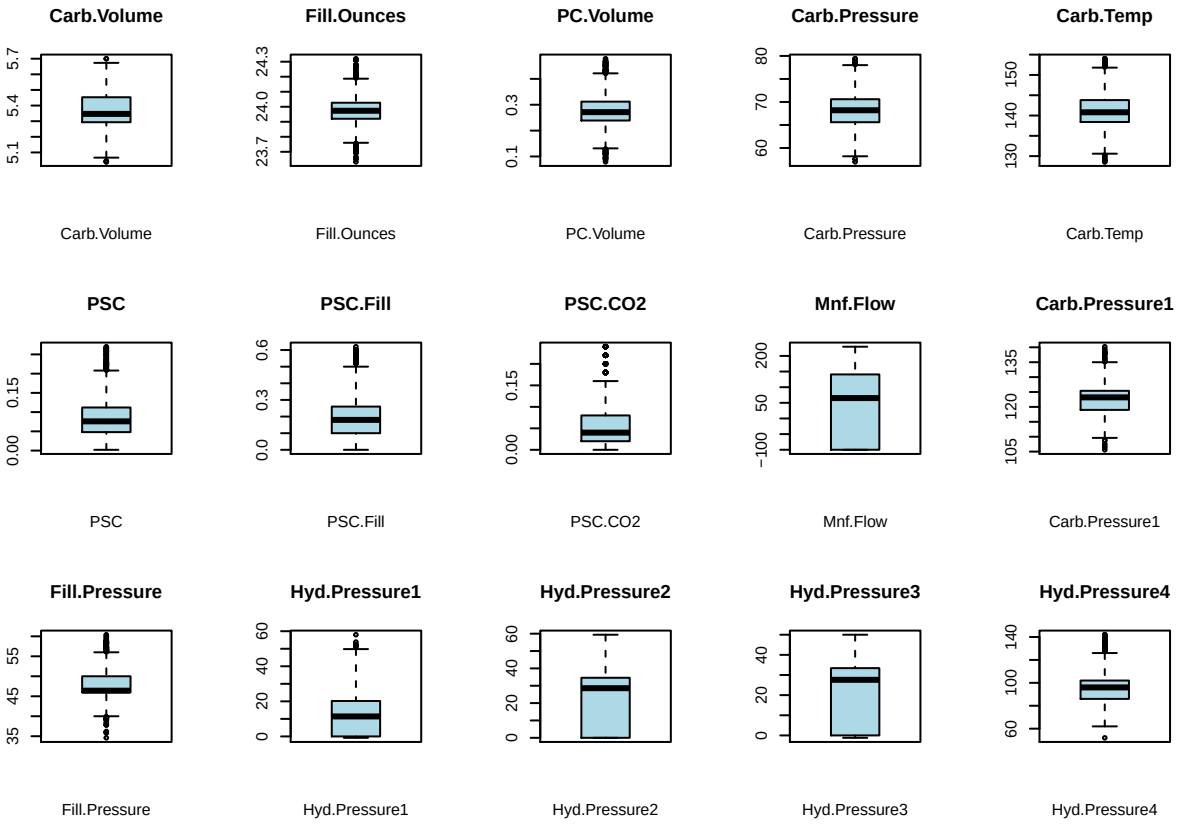
```
par(mfrow = c(3,5), cex = .5)
project_df <- project_df[2:ncol(project_df)]
for(i in colnames(project_df)){
hist(project_df[,i], xlab = names(project_df[i]),
  main = names(project_df[i]), col="light blue", ylab="")
}
```

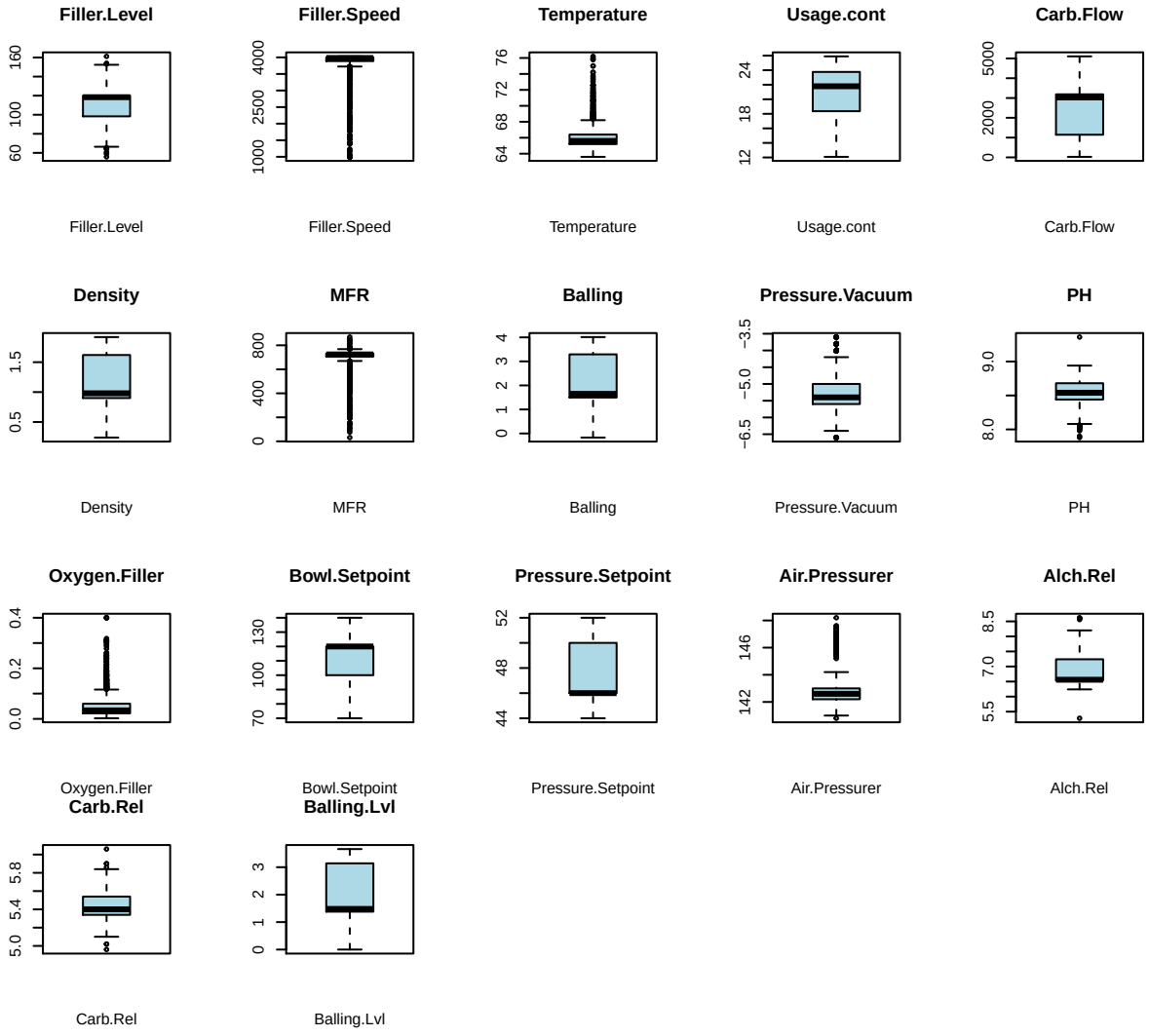

Carb.Volume
Fill.Ounces
PC.Volume
Carb.Pressure
Carb.Temp
PSC
PSC.Fill
PSC.CO2
Mnf.Flow
Carb.Pressure1
Fill.Pressure
Hyd.Pressure1
Hyd.Pressure2
Hyd.Pressure3
Hyd.Pressure4
Filler.Level
Filler.Speed
Temperature
Usage.cont
Carb.Flow
Density
MFR
Balling
Pressure.Vacuum
PH
Oxygen.Filler
Bowl.Setpoint
Pressure.Setpoint
Air.Pressurer
Alch.Rel



From the above some of the variables shows nearly normal distributions.Some of variables shows strong skewness that means the presence of outliers. Also some variables have many near to zero values. Some varibles need transformations here. (need to update all these) Plotting boxplots of all variables will give better understanding with outliers and provide a view to look for what approach should be considered to fix the outliers.

```
par(mfrow = c(3,5), cex = .5)
for(i in colnames(project_df)){
boxplot(project_df[,i], xlab = names(project_df[i]),
  main = names(project_df[i]), col="light blue", ylab="")
}
```





(update here with method to fix outliers)

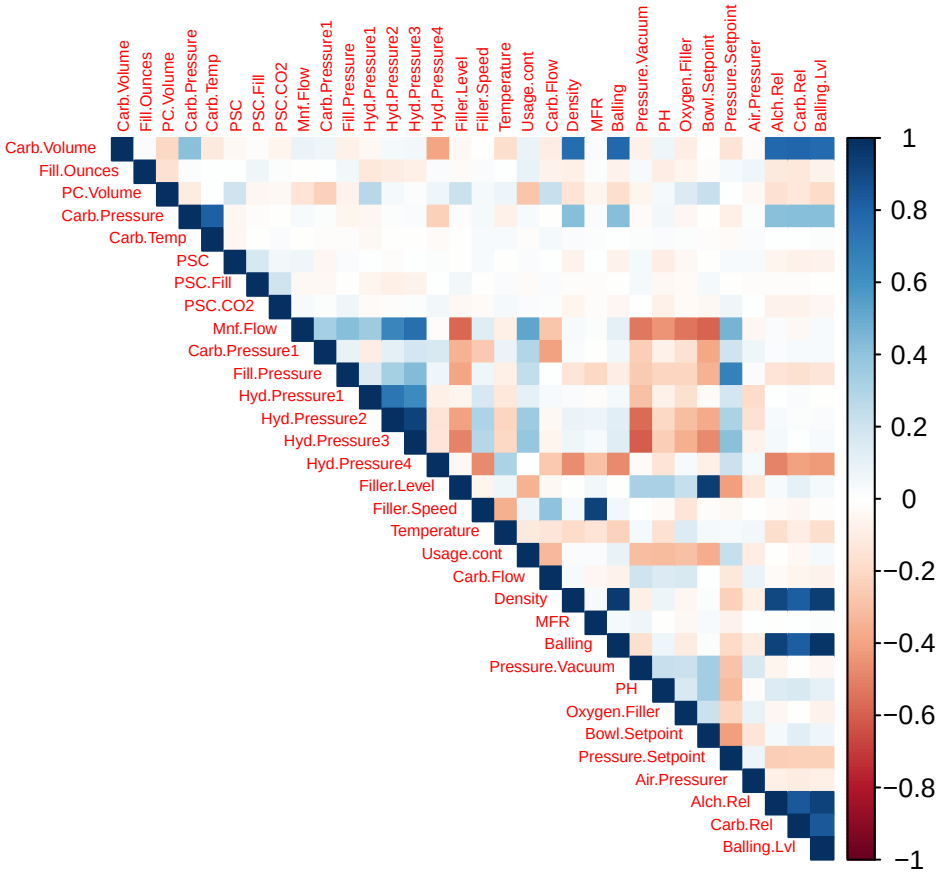
CORRELATIONS BETWEEN VARIABLES :-

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
plotcorr =cor(project_df,use="pairwise.complete.obs", method = "pearson")

corrplot(plotcorr, method = "color",type = "upper", order = "original", number.cex = .7,tl.pos = "td",t
```




```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
hc = findCorrelation(plotcorr, cutoff=0.75)
hc
```

```
##  [1] 23  9 14 32 30 21 31 16 17  5
```

We have found out the variables that are very highly correlated with each other.we can remove these variables (need to discuss with everyone here) Let us now look at the correlation between the target (pH) variable and the predictors.

```
library(corrr)
project_df %>% correlate() %>% focus(PH)
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
## # A tibble: 31 x 2
##    rowname             PH
##    <chr>            <dbl>
##  1 Carb.Volume     0.0635
##  2 Fill.Ounces    -0.0951
##  3 PC.Volume       0.0458
##  4 Carb.Pressure   0.0596
```

```
##  5 Carb.Temp       0.0289
##  6 PSC            -0.0900
##  7 PSC.Fill       -0.0363
##  8 PSC.CO2        -0.0756
##  9 Mnf.Flow       -0.447 
## 10 Carb.Pressure1 -0.0793
## # ... with 21 more rows
```