Project2\_Group1\_Data624

Angrand, Burke, Deboch, Groysman, Karr

December 1, 2019

## Data 624 Project 2

### Prompt

You are given a simple data set from a beverage manufacturing company. It consists of 2,571 rows/cases of data and 33 columns / variables. **Your goal is to use this data to predict PH (a column in the set).** PH is a measure of acidity/alkalinity, it must conform in a critical range and therefore it is important to understand its influence and predict its values.This is production data. PH is a KPI, Key Performance Indicator. You are also given a scoring set (267 cases). All variables other than the dependent or target. You will use this data to score your model with your best predictions.

**Provided Files**

* Data Dictionary.xlsx *Provides a listing of the columns and their underlying data components*
* StudentData.xlsx *The training dataset for the exercise*
* StudentEvaluation- TO PREDICT.xlsx *The evaluation dataset for the exercise*

**Required Packages**

#Upload library  
library(readxl)  
library(dplyr)  
library(tidyr)  
library(stringr)  
library(ggplot2)  
library(ggcorrplot)  
library(reshape2)

#### 1. Read in Data

* Download from the group’s github repository link to ensure stakeholders can reproduce easily.
* The file is downloaded from git to the user’s default downloads location and read in with the read\_excel function from the readxl module
* Print the dimensions to ensure that the data is consistent with promt specifications
* Train Data: 2,571 observations, 33 predictors
* Eval Data: 267 observations, 33 predictors

**a) Training Dataset - Student Data**

train.loc <- tempfile(fileext = ".xlsx")  
train.dataURL <- "https://raw.githubusercontent.com/mburke65/CUNY\_Data624/master/Project2Folder/ProvidedFiles/StudentData.xlsx"  
download.file(train.dataURL, destfile= train.loc, mode='wb')  
  
  
train.data <- readxl::read\_excel(train.loc,sheet = 1, col\_names =TRUE)  
print(paste("Dimensions of the train.data:", list(dim(train.data))))

## [1] "Dimensions of the train.data: c(2571, 33)"

**b) Evaluation Dataset - StudentEvaluation- TO PREDICT.xlsx**

eval.loc <- tempfile(fileext = ".xlsx")  
eval.dataURL <- "https://raw.githubusercontent.com/mburke65/CUNY\_Data624/master/Project2Folder/ProvidedFiles/StudentEvaluation-%20TO%20PREDICT.xlsx"  
download.file(eval.dataURL, destfile= eval.loc, mode='wb')  
  
eval.data <- read\_excel(eval.loc,sheet = 1, col\_names =TRUE)  
print(paste("Dimensions of the eval.data:", list(dim(eval.data))))

## [1] "Dimensions of the eval.data: c(267, 33)"

#### 2. Exploratory Data Analysis

* 1. Identify the predictors (column names)
  2. For numerical predictors (!= “Brand Code”), explore predictors summary statistics. The summary statistics provide a quick and simple description of the data which helps us have a better understanding of the data. The summary function in R provides the mean, median, number of nulls, min and max.
  3. For categorical predictors (== “Brand Code”), explore the frequency of brands
  4. Graphically highlight nulls
  5. Investigate the correlation between predictors
  6. histogram of all the predictors to better understand the shape and spread of the data

**a. Identify Predictors**

#predictors   
names(train.data)

## [1] "Brand Code" "Carb Volume" "Fill Ounces"   
## [4] "PC Volume" "Carb Pressure" "Carb Temp"   
## [7] "PSC" "PSC Fill" "PSC CO2"   
## [10] "Mnf Flow" "Carb Pressure1" "Fill Pressure"   
## [13] "Hyd Pressure1" "Hyd Pressure2" "Hyd Pressure3"   
## [16] "Hyd Pressure4" "Filler Level" "Filler Speed"   
## [19] "Temperature" "Usage cont" "Carb Flow"   
## [22] "Density" "MFR" "Balling"   
## [25] "Pressure Vacuum" "PH" "Oxygen Filler"   
## [28] "Bowl Setpoint" "Pressure Setpoint" "Air Pressurer"   
## [31] "Alch Rel" "Carb Rel" "Balling Lvl"

**b) Summary Statistics Table**

* Initial Findings: MFR and Brand Code have a significant percentage of null values and the scaling/range of each variable varies

#get the null values  
  
null.values <-as.data.frame(sapply(train.data, function(x) sum(is.na(x))), col.names = "null\_values")%>%  
 tibble::rownames\_to\_column("Predictors")%>%  
 rename(nulls = 2)%>%  
 mutate(Percentage\_Missing = round((nulls/2571)\*100,2))  
  
#get the summary stats, restructure into a readable dataframe. merge on the null.values  
summary.stats <- as.data.frame(summary(train.data))%>%  
 separate(Freq, c("Summary.Stat","Value"), sep = ":")%>%  
 na.omit() %>%  
 mutate(Var2 = as.character(Var2))%>%  
 mutate\_if(is.character, str\_trim)%>%  
 filter(Value !="character " )%>%  
 mutate(Value.Num = factor(Value))%>%  
 select(-c(Var1, Value))%>%  
 rename(Predictors = Var2)%>%  
 spread(Summary.Stat, Value.Num)%>%  
 left\_join(null.values, by = "Predictors")%>%  
 arrange(-nulls)%>%  
 select(-c("1st Qu.","3rd Qu.","Class", "Length" , "NA's" ,'Mode' ))  
  
summary.stats

## Predictors Max. Mean Median Min. nulls  
## 1 MFR 868.6 704.0 724.0 31.4 212  
## 2 Brand Code <NA> <NA> <NA> <NA> 120  
## 3 Filler Speed 4030 3687 3982 998 57  
## 4 PC Volume 0.47800 0.27712 0.27133 0.07933 39  
## 5 PSC CO2 0.24000 0.05641 0.04000 0.00000 39  
## 6 Fill Ounces 24.32 23.97 23.97 23.63 38  
## 7 PSC 0.27000 0.08457 0.07600 0.00200 33  
## 8 Carb Pressure1 140.2 122.6 123.2 105.6 32  
## 9 Hyd Pressure4 142.00 96.29 96.00 52.00 30  
## 10 Carb Pressure 79.40 68.19 68.20 57.00 27  
## 11 Carb Temp 154.0 141.1 140.8 128.6 26  
## 12 PSC Fill 0.6200 0.1954 0.1800 0.0000 23  
## 13 Fill Pressure 60.40 47.92 46.40 34.60 22  
## 14 Filler Level 161.2 109.3 118.4 55.8 20  
## 15 Hyd Pressure2 59.40 20.96 28.60 0.00 15  
## 16 Hyd Pressure3 50.00 20.46 27.60 -1.20 15  
## 17 Temperature 76.20 65.97 65.60 63.60 14  
## 18 Oxygen Filler 0.40000 0.04684 0.03340 0.00240 12  
## 19 Pressure Setpoint 52.00 47.62 46.00 44.00 12  
## 20 Hyd Pressure1 58.00 12.44 11.40 -0.80 11  
## 21 Carb Rel 6.060 5.437 5.400 4.960 10  
## 22 Carb Volume 5.700 5.370 5.347 5.040 10  
## 23 Alch Rel 8.620 6.897 6.560 5.280 9  
## 24 Usage cont 25.90 20.99 21.79 12.08 5  
## 25 PH 9.360 8.546 8.540 7.880 4  
## 26 Bowl Setpoint 140.0 109.3 120.0 70.0 2  
## 27 Carb Flow 5104 2468 3028 26 2  
## 28 Mnf Flow 229.40 24.57 65.20 -100.20 2  
## 29 Balling 4.012 2.198 1.648 -0.170 1  
## 30 Balling Lvl 3.66 2.05 1.48 0.00 1  
## 31 Density 1.920 1.174 0.980 0.240 1  
## 32 Air Pressurer 148.2 142.8 142.6 140.8 0  
## 33 Pressure Vacuum -3.600 -5.216 -5.400 -6.600 0  
## Percentage\_Missing  
## 1 8.25  
## 2 4.67  
## 3 2.22  
## 4 1.52  
## 5 1.52  
## 6 1.48  
## 7 1.28  
## 8 1.24  
## 9 1.17  
## 10 1.05  
## 11 1.01  
## 12 0.89  
## 13 0.86  
## 14 0.78  
## 15 0.58  
## 16 0.58  
## 17 0.54  
## 18 0.47  
## 19 0.47  
## 20 0.43  
## 21 0.39  
## 22 0.39  
## 23 0.35  
## 24 0.19  
## 25 0.16  
## 26 0.08  
## 27 0.08  
## 28 0.08  
## 29 0.04  
## 30 0.04  
## 31 0.04  
## 32 0.00  
## 33 0.00

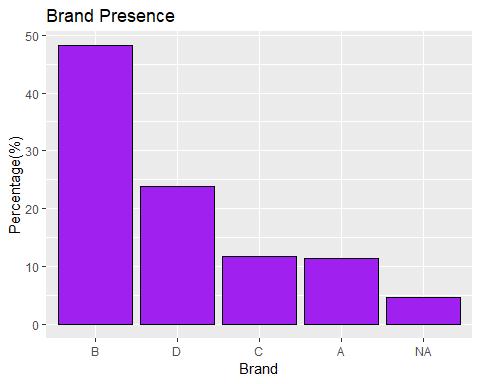
**c) Categorical Frequency - Brand Code**

* Brand Code B is the most popular brand produced by the factory by a significant margin
* Approximately 5% of the Brands dropped identification (null), future null analysis in (d)

brand <- train.data%>%  
 rename(Brand\_Code = 1)%>%  
 group\_by(Brand\_Code)%>% summarise(n = n())%>%  
 arrange(-n)%>%  
 mutate(Presence\_Precentage = round((n/sum(n))\*100,1))  
brand

## # A tibble: 5 x 3  
## Brand\_Code n Presence\_Precentage  
## <chr> <int> <dbl>  
## 1 B 1239 48.2  
## 2 D 615 23.9  
## 3 C 304 11.8  
## 4 A 293 11.4  
## 5 <NA> 120 4.7

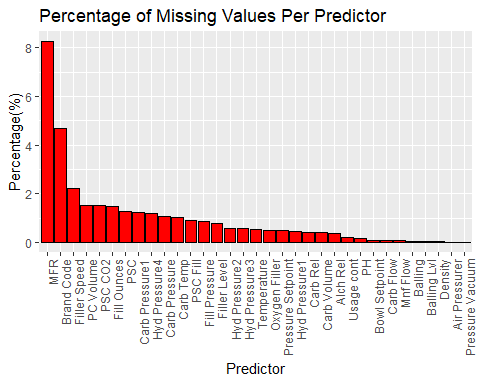
brand%>%  
   
 ggplot(aes(x= reorder(Brand\_Code,-Presence\_Precentage), y = Presence\_Precentage))+  
 geom\_bar(stat = "identity", fill = "purple", color = "black")+   
 ggtitle("Brand Presence")+  
 xlab("Brand")+  
 ylab("Percentage(%)")



**d) Null Display - All Predictors**

* This graph displays the missing values across diverse predictors. In subsequent sections, the group will determine an approach to deal with missing values (i.e. random generation of values, mean replacement or k-neighbor, etc.)

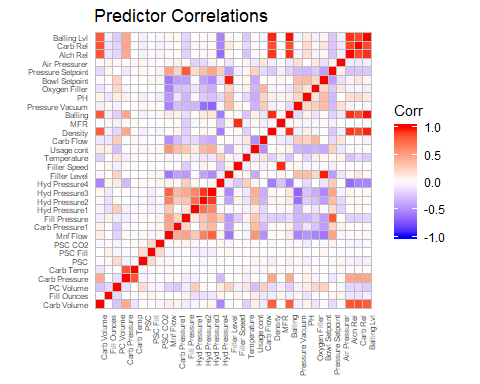
summary.stats%>%  
   
 ggplot(aes(x= reorder(Predictors,-Percentage\_Missing), y = Percentage\_Missing ))+  
 geom\_bar(stat = "identity",fill = "red", color = "black")+   
 theme(axis.text.x=element\_text(angle=90, hjust=1))+   
 ggtitle("Percentage of Missing Values Per Predictor")+  
 xlab("Predictor")+  
 ylab("Percentage(%)")



**e) Data Correlations Display - All Predictors**

* Compute the correlation matrix using the cor() function. Override the “use” parameter, a optional parameter method for computing covariances in the presence of missing values which the PH dataset has.
* Display the correlation matrix ggcorrplot()
* Extract the unique predictor combinations with the highest correlation values (>= .85)
* Balling Lvl is highly correlated to Balling, Density, Alch Rel, Carb Rel
* Carb Rel is highly correlated to Alch Rel, Balling, Density
* Bowl Setpoint is highly correlated to Filler Level
* Balling is highly correlated to Density
* MFR is highly correlated to Filler Speed
* Hyd Pressure3 is highly correlated to Hyd Pressure2
* In subsequent steps, the group will likely remove the highly correlated variables to avoid multicollinearity in our modeling

cor.matrix <- cor(train.data[,-1], use = "na.or.complete")  
ggcorrplot(cor.matrix)+  
 theme(axis.text.x=element\_text(size=rel(.7), angle=90, hjust=1),  
 axis.text.y = element\_text(size=rel(.7), hjust=1))+  
 ggtitle("Predictor Correlations")



reshape2::melt(cor.matrix)%>%  
 rename(Predictor1 = Var1, Predictor2 = Var2, CorrelationValue = value)%>%  
 filter(CorrelationValue != 1)%>%  
 arrange(-CorrelationValue) %>%  
 filter(CorrelationValue >= 0.85)%>%   
 filter (! duplicated(CorrelationValue))%>%  
 arrange(Predictor1, -CorrelationValue)

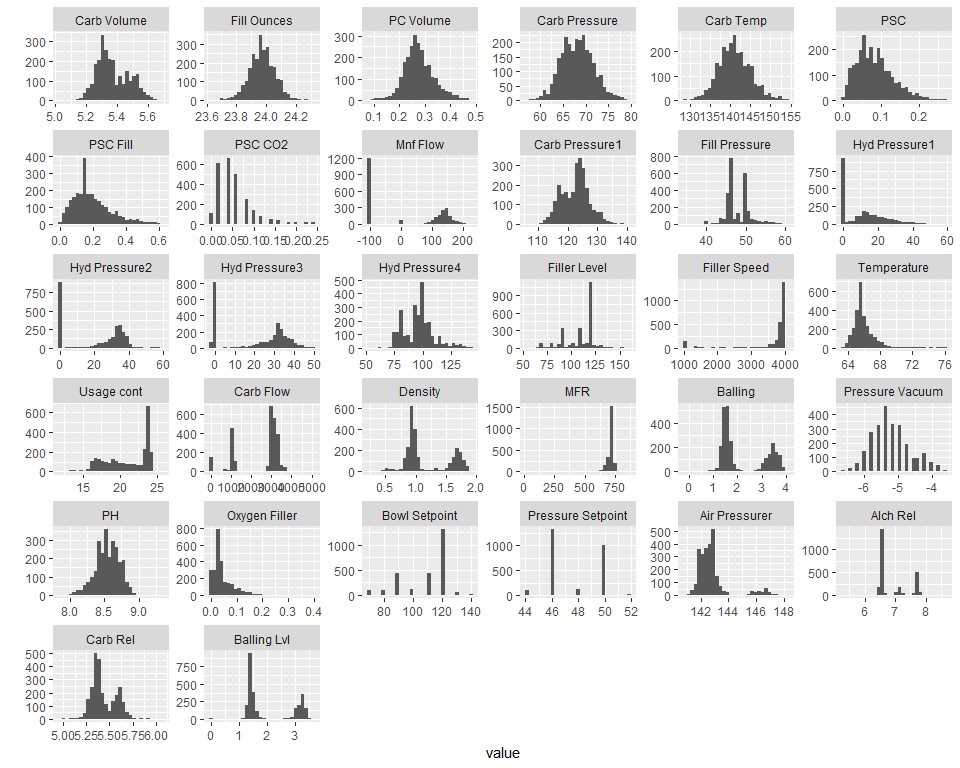
## Predictor1 Predictor2 CorrelationValue  
## 1 Hyd Pressure3 Hyd Pressure2 0.9176010  
## 2 MFR Filler Speed 0.9514224  
## 3 Balling Density 0.9523125  
## 4 Bowl Setpoint Filler Level 0.9773811  
## 5 Alch Rel Balling 0.9412251  
## 6 Alch Rel Density 0.9157798  
## 7 Carb Rel Alch Rel 0.8768039  
## 8 Carb Rel Balling 0.8542582  
## 9 Carb Rel Density 0.8526890  
## 10 Balling Lvl Balling 0.9876727  
## 11 Balling Lvl Density 0.9550900  
## 12 Balling Lvl Alch Rel 0.9429801  
## 13 Balling Lvl Carb Rel 0.8682872

**f.1) Scatter Plots - All Predictors**

* Convert the train.data from wide to long foramt
* Use qplot to generate histograms for each variable
* Initial Findings (NORMAL?): Fill Ounces, PC Volume, Carb Pressure, Carb Temp, Carb Pressure1, PH follow somewhat normal distributions
* Initial Findings (CLUSTER?): Mnf Flow, Hyd Pressure1, Hyd Pressure2, Hyd Pressure3, Filler Speed, and Carb flow values appear more clustered and may need to be examined more closely
* Initial Findings (CATEGORICAL?): Pressure Setpoint, Bowl Setpoint appear to be more categorical as there is very little variety/distribution in the resulting data

melted.train <- melt(train.data[,-1])  
qplot(value, data=melted.train) + facet\_wrap(~variable, scales="free")

## Warning: Removed 724 rows containing non-finite values (stat\_bin).



**f.2) Scatter Plots For Identified Clustered Predictors**

* Use plot() to test the PH levels based on each predictor identified to assess if any values are problematic and should be removed in the modeling step
* Find the counts at each bin to assess if the distribution makes sense

*Function to generate the scatter plot of PH vs. the inputted predictor*

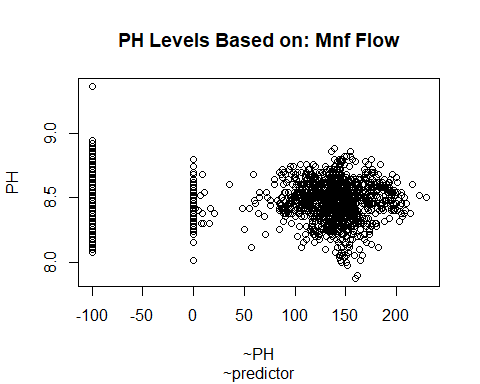
scatter\_analysis <- function(predictor) {  
 subset<- train.data %>%  
 select(c(predictor, "PH"))%>%  
 filter(complete.cases(predictor, "PH"))  
  
 plotgraphic<- plot(subset, aes(PH, predictor))+  
 title(paste0("PH Levels Based on: ", predictor))  
   
 return(plotgraphic)  
}

*Function to generate the counts at each bin*

distribution\_cluster <- function(predictor) {  
 test <- melted.train%>%  
 filter(variable ==predictor )%>%  
 select(c(value))  
   
 histrv<- hist(test$value)  
 breaks <- histrv$breaks  
   
 #loop through the breaks to generate the string range   
 value\_list <- c()  
 i <- 1  
 for (val in breaks){  
 if(which(breaks == val) != 1){  
 value\_list[[i]] <- paste0(breaks[which(breaks == val)-1]," to ",breaks[which(breaks == val)] )  
 i <- i +1  
 }  
 }  
#create a dateframe from the breaks and the counts   
return\_df<- data.frame(breaks= value\_list, counts = histrv$counts)  
#add the percentage at each break point  
return\_df <- return\_df %>%  
 mutate(Percentage\_Break = round((counts/sum(counts))\*100,2))  
return(return\_df)   
}

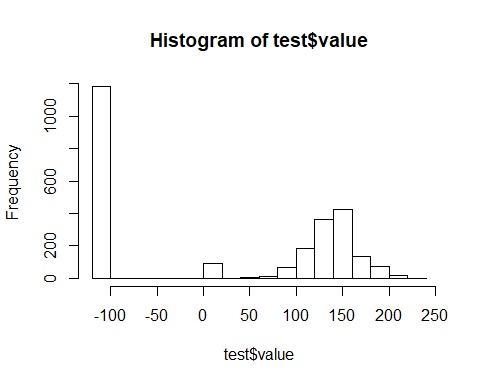
**f.2) Mnf Flow**

scatter\_analysis("Mnf Flow")



## integer(0)

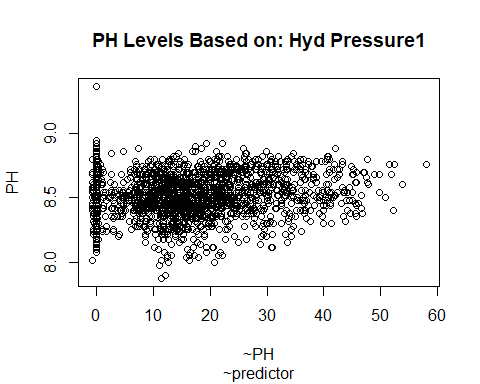
distribution\_cluster("Mnf Flow")



## breaks counts Percentage\_Break  
## 1 -120 to -100 1184 46.09  
## 2 -100 to -80 0 0.00  
## 3 -80 to -60 0 0.00  
## 4 -60 to -40 0 0.00  
## 5 -40 to -20 0 0.00  
## 6 -20 to 0 0 0.00  
## 7 0 to 20 89 3.46  
## 8 20 to 40 2 0.08  
## 9 40 to 60 5 0.19  
## 10 60 to 80 13 0.51  
## 11 80 to 100 65 2.53  
## 12 100 to 120 187 7.28  
## 13 120 to 140 363 14.13  
## 14 140 to 160 427 16.62  
## 15 160 to 180 138 5.37  
## 16 180 to 200 75 2.92  
## 17 200 to 220 19 0.74  
## 18 220 to 240 2 0.08

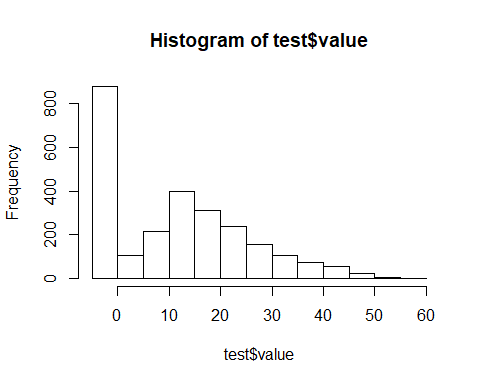
**f.2) Hyd Pressure1**

scatter\_analysis("Hyd Pressure1")



## integer(0)

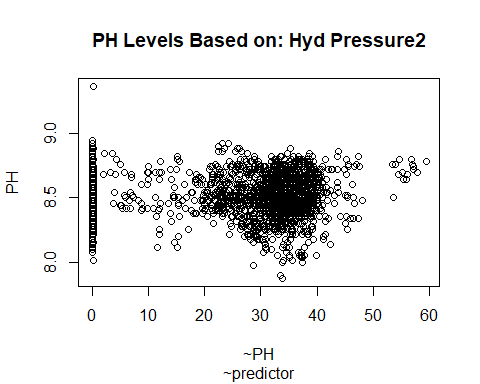
distribution\_cluster("Hyd Pressure1")



## breaks counts Percentage\_Break  
## 1 -5 to 0 879 34.34  
## 2 0 to 5 107 4.18  
## 3 5 to 10 213 8.32  
## 4 10 to 15 396 15.47  
## 5 15 to 20 311 12.15  
## 6 20 to 25 236 9.22  
## 7 25 to 30 156 6.09  
## 8 30 to 35 107 4.18  
## 9 35 to 40 71 2.77  
## 10 40 to 45 54 2.11  
## 11 45 to 50 23 0.90  
## 12 50 to 55 6 0.23  
## 13 55 to 60 1 0.04

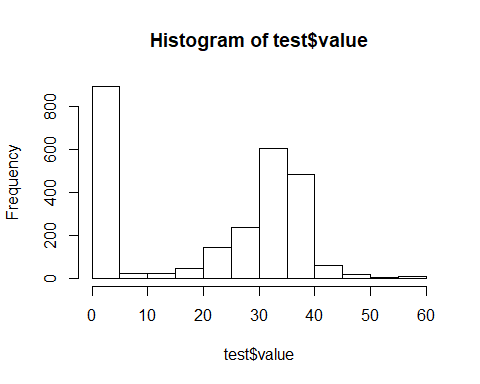
**f.2) Hyd Pressure2**

scatter\_analysis("Hyd Pressure2")



## integer(0)

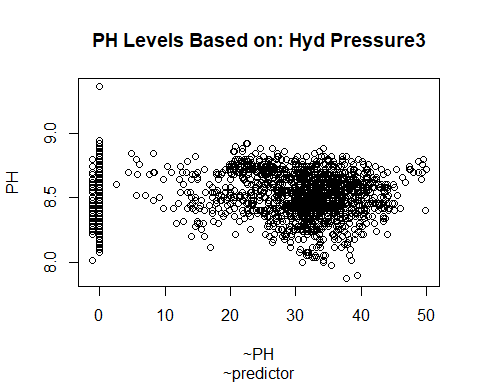
distribution\_cluster("Hyd Pressure2")



## breaks counts Percentage\_Break  
## 1 0 to 5 896 35.05  
## 2 5 to 10 24 0.94  
## 3 10 to 15 25 0.98  
## 4 15 to 20 45 1.76  
## 5 20 to 25 142 5.56  
## 6 25 to 30 239 9.35  
## 7 30 to 35 604 23.63  
## 8 35 to 40 483 18.90  
## 9 40 to 45 62 2.43  
## 10 45 to 50 20 0.78  
## 11 50 to 55 5 0.20  
## 12 55 to 60 11 0.43

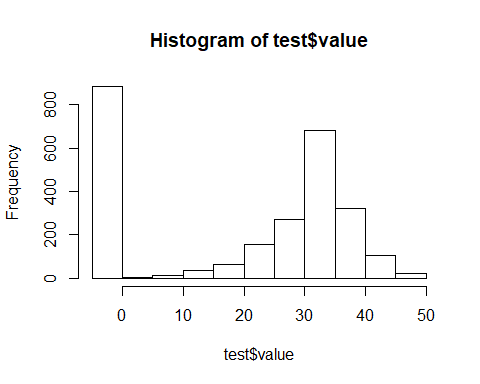
**f.2) Hyd Pressure3**

scatter\_analysis("Hyd Pressure3")



## integer(0)

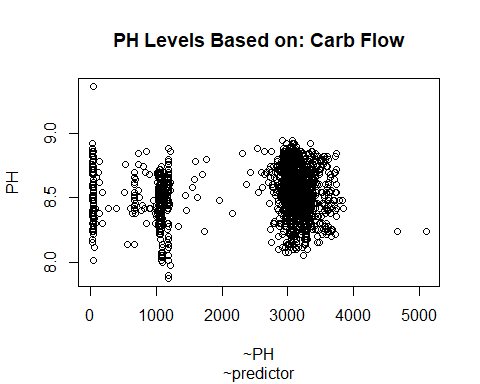
distribution\_cluster("Hyd Pressure3")



## breaks counts Percentage\_Break  
## 1 -5 to 0 884 34.59  
## 2 0 to 5 4 0.16  
## 3 5 to 10 13 0.51  
## 4 10 to 15 35 1.37  
## 5 15 to 20 63 2.46  
## 6 20 to 25 157 6.14  
## 7 25 to 30 269 10.52  
## 8 30 to 35 683 26.72  
## 9 35 to 40 320 12.52  
## 10 40 to 45 107 4.19  
## 11 45 to 50 21 0.82

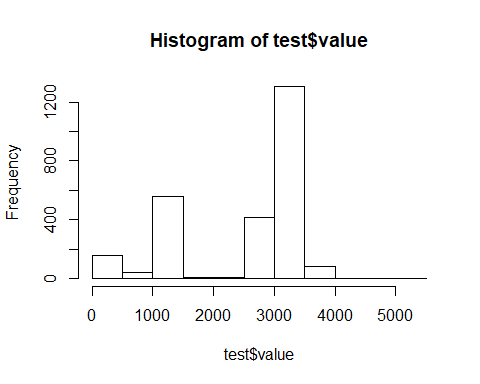
**f.2) Carb Flow**

scatter\_analysis("Carb Flow")



## integer(0)

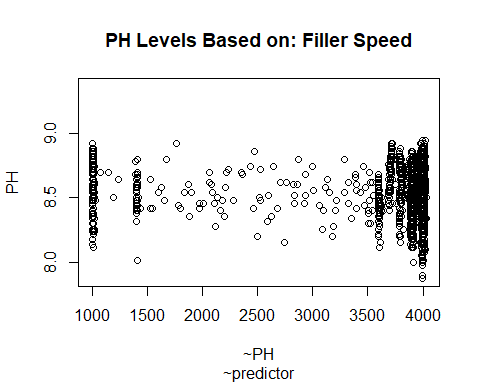
distribution\_cluster("Carb Flow")



## breaks counts Percentage\_Break  
## 1 0 to 500 153 5.96  
## 2 500 to 1000 40 1.56  
## 3 1000 to 1500 558 21.72  
## 4 1500 to 2000 8 0.31  
## 5 2000 to 2500 4 0.16  
## 6 2500 to 3000 413 16.08  
## 7 3000 to 3500 1307 50.88  
## 8 3500 to 4000 84 3.27  
## 9 4000 to 4500 0 0.00  
## 10 4500 to 5000 1 0.04  
## 11 5000 to 5500 1 0.04

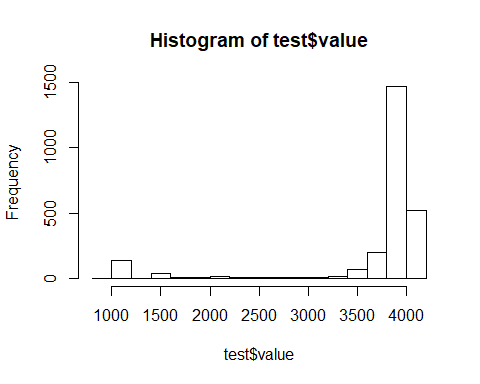
**f.2) Filler Speed**

scatter\_analysis("Filler Speed")



## integer(0)

distribution\_cluster("Filler Speed")



## breaks counts Percentage\_Break  
## 1 800 to 1000 1 0.04  
## 2 1000 to 1200 140 5.57  
## 3 1200 to 1400 3 0.12  
## 4 1400 to 1600 38 1.51  
## 5 1600 to 1800 6 0.24  
## 6 1800 to 2000 6 0.24  
## 7 2000 to 2200 11 0.44  
## 8 2200 to 2400 6 0.24  
## 9 2400 to 2600 8 0.32  
## 10 2600 to 2800 7 0.28  
## 11 2800 to 3000 9 0.36  
## 12 3000 to 3200 9 0.36  
## 13 3200 to 3400 12 0.48  
## 14 3400 to 3600 70 2.78  
## 15 3600 to 3800 198 7.88  
## 16 3800 to 4000 1469 58.43  
## 17 4000 to 4200 521 20.72