Project 2 in R

DATA 624-01 Group 3

Z. Palmore, K. Popkin, K. Potter, C. Nan, J. Ramalingam

7/8/2021

Table of Contents

# Summary

# Data Exploration

Data is loaded and all packages used throughout the report are noted. We view a random sample of the data below. It shows the randomly selected observations as the first, second, third, fourth, and fifth for each variable.

library(utils)  
library(psych)  
library(tidyverse)  
library(corrplot)  
library(MASS)  
library(caret)  
library(flextable)  
ph <- read.csv(  
 "https://raw.githubusercontent.com/palmorezm/msds/main/624/Projects/Project2/StudentData%20-%20TO%20MODEL.csv")  
obs.sample <- as.data.frame(t(head(sample\_n(ph[1:33], 5), 5)))  
obs.sample <- as.data.frame(lapply(obs.sample[2:33,],   
function(x) round(as.numeric(as.character(x)),1)))  
obs.sample <- rbind(ph$ï..Brand.Code, obs.sample)  
obs.sample <- obs.sample %>%   
 mutate(Variable = colnames(ph)) %>%  
 dplyr::select(Variable, X1, X2, X3, X4, X5)  
flextable(obs.sample) %>%   
 theme\_vanilla() %>%   
 set\_table\_properties(layout = "autofit")

| **Variable** | **X1** | **X2** | **X3** | **X4** | **X5** |
| --- | --- | --- | --- | --- | --- |
| ï..Brand.Code | B | A | B | A | A |
| Carb.Volume | 5.3 | 5.5 | 5.3 | 5.4 | 5.3 |
| Fill.Ounces | 24.1 |  | 24 | 24.1 | 23.9 |
| PC.Volume | 0.3 | 0.4 | 0.3 | 0.2 | 0.2 |
| Carb.Pressure | 66 | 65.6 | 64.2 | 71.8 | 68 |
| Carb.Temp | 139.4 | 135.8 | 138 | 143.8 | 142 |
| PSC | 0.1 | 0.1 | 0 | 0.1 | 0.1 |
| PSC.Fill | 0.1 |  | 0.1 | 0.1 | 0.1 |
| PSC.CO2 | 0.1 | 0 | 0 | 0 | 0 |
| Mnf.Flow | -100 | -100.2 | 137 | 120.4 | 164.6 |
| Carb.Pressure1 | 130.2 | 132.8 | 126.6 | 124 | 124.2 |
| Fill.Pressure | 44.2 | 44.4 | 50 | 50.2 | 50.2 |
| Hyd.Pressure1 | 0 | 0 | 9.8 | 7.2 | 13 |
| Hyd.Pressure2 | 0 | 0 | 36.8 | 30.4 | 46.4 |
| Hyd.Pressure3 | 0 | 0 | 33.6 | 25.8 | 40.6 |
| Hyd.Pressure4 | 130 | 108 | 114 | 94 | 98 |
| Filler.Level | 100.2 | 125 | 90.2 | 112 | 91 |
| Filler.Speed | 1008 | 4010 | 3994 | 3990 | 3908 |
| Temperature | 69.8 | 65.6 | 68 | 66.8 | 65 |
| Usage.cont | 22 | 17 | 23.4 | 24 | 23.7 |
| Carb.Flow | 30 | 2706 | 3298 | 1068 | 1078 |
| Density | 0.7 | 1.7 | 0.9 | 1 | 1 |
| MFR |  | 736.4 | 728.2 | 731.4 | 697.8 |
| Balling | 1 | 3.3 | 1.4 | 1.6 | 1.6 |
| Pressure.Vacuum | -4 | -4.8 | -5 | -5.4 | -5.6 |
| PH | 8.5 | 8.6 | 8.6 | 8.5 | 8.3 |
| Oxygen.Filler |  | 0.1 | 0 | 0 | 0 |
| Bowl.Setpoint | 100 | 126 | 90 | 110 | 90 |
| Pressure.Setpoint | 50 | 46 | 50 | 50 | 50 |
| Air.Pressurer | 147 | 142.2 | 142.6 | 143.2 | 142.4 |
| Alch.Rel | 6.5 | 7 | 6.6 | 6.5 | 6.5 |
| Carb.Rel | 5.4 | 6.1 | 5.3 | 5.4 | 5.3 |
| Balling.Lvl | 1.5 | 3.2 | 1.4 | 1.4 | 1.4 |

We note a few interesting facts before exploring further. We notice that data is missing. These appear as empty spots on the table. It also appears that due to the conversion of spreadsheet formats from excel to csv when hosting the data, our first column variable name is legible but needs the “i..” symbol removed. None of the other 33 variables need to be renamed. Our target variable, PH, appears to be on a scale between 1 and 10, as it should be, while the others vary between roughly -100 (Mnf.Flow) and greater than 4000 (Filler.Speed). We take a closer look at these variables with some descriptive statistics.

ph.desc <- ph %>%   
 describe() %>%  
 mutate("pm" = round(((2571 - n)/2571)\*100, 2),   
 obs = n,   
 med = median) %>%   
 round(digits = 1) %>%   
 mutate(var = colnames(ph)) %>%  
 dplyr::select(var, obs, pm, mean, med,  
 sd, min, max, skew, se)  
flextable::flextable(ph.desc) %>%   
 theme\_vanilla() %>%   
 set\_table\_properties(layout = "autofit")

| **var** | **obs** | **pm** | **mean** | **med** | **sd** | **min** | **max** | **skew** | **se** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ï..Brand.Code | 2,571 | 0.0 | 3.4 | 3.0 | 1.1 | 1.0 | 5.0 | 0.0 | 0.0 |
| Carb.Volume | 2,561 | 0.4 | 5.4 | 5.3 | 0.1 | 5.0 | 5.7 | 0.4 | 0.0 |
| Fill.Ounces | 2,533 | 1.5 | 24.0 | 24.0 | 0.1 | 23.6 | 24.3 | 0.0 | 0.0 |
| PC.Volume | 2,532 | 1.5 | 0.3 | 0.3 | 0.1 | 0.1 | 0.5 | 0.3 | 0.0 |
| Carb.Pressure | 2,544 | 1.0 | 68.2 | 68.2 | 3.5 | 57.0 | 79.4 | 0.2 | 0.1 |
| Carb.Temp | 2,545 | 1.0 | 141.1 | 140.8 | 4.0 | 128.6 | 154.0 | 0.2 | 0.1 |
| PSC | 2,538 | 1.3 | 0.1 | 0.1 | 0.0 | 0.0 | 0.3 | 0.8 | 0.0 |
| PSC.Fill | 2,548 | 0.9 | 0.2 | 0.2 | 0.1 | 0.0 | 0.6 | 0.9 | 0.0 |
| PSC.CO2 | 2,532 | 1.5 | 0.1 | 0.0 | 0.0 | 0.0 | 0.2 | 1.7 | 0.0 |
| Mnf.Flow | 2,569 | 0.1 | 24.6 | 65.2 | 119.5 | -100.2 | 229.4 | 0.0 | 2.4 |
| Carb.Pressure1 | 2,539 | 1.2 | 122.6 | 123.2 | 4.7 | 105.6 | 140.2 | 0.1 | 0.1 |
| Fill.Pressure | 2,549 | 0.9 | 47.9 | 46.4 | 3.2 | 34.6 | 60.4 | 0.5 | 0.1 |
| Hyd.Pressure1 | 2,560 | 0.4 | 12.4 | 11.4 | 12.4 | -0.8 | 58.0 | 0.8 | 0.2 |
| Hyd.Pressure2 | 2,556 | 0.6 | 21.0 | 28.6 | 16.4 | 0.0 | 59.4 | -0.3 | 0.3 |
| Hyd.Pressure3 | 2,556 | 0.6 | 20.5 | 27.6 | 16.0 | -1.2 | 50.0 | -0.3 | 0.3 |
| Hyd.Pressure4 | 2,541 | 1.2 | 96.3 | 96.0 | 13.1 | 52.0 | 142.0 | 0.5 | 0.3 |
| Filler.Level | 2,551 | 0.8 | 109.3 | 118.4 | 15.7 | 55.8 | 161.2 | -0.8 | 0.3 |
| Filler.Speed | 2,514 | 2.2 | 3,687.2 | 3,982.0 | 770.8 | 998.0 | 4,030.0 | -2.9 | 15.4 |
| Temperature | 2,557 | 0.5 | 66.0 | 65.6 | 1.4 | 63.6 | 76.2 | 2.4 | 0.0 |
| Usage.cont | 2,566 | 0.2 | 21.0 | 21.8 | 3.0 | 12.1 | 25.9 | -0.5 | 0.1 |
| Carb.Flow | 2,569 | 0.1 | 2,468.4 | 3,028.0 | 1,073.7 | 26.0 | 5,104.0 | -1.0 | 21.2 |
| Density | 2,570 | 0.0 | 1.2 | 1.0 | 0.4 | 0.2 | 1.9 | 0.5 | 0.0 |
| MFR | 2,359 | 8.2 | 704.0 | 724.0 | 73.9 | 31.4 | 868.6 | -5.1 | 1.5 |
| Balling | 2,570 | 0.0 | 2.2 | 1.6 | 0.9 | -0.2 | 4.0 | 0.6 | 0.0 |
| Pressure.Vacuum | 2,571 | 0.0 | -5.2 | -5.4 | 0.6 | -6.6 | -3.6 | 0.5 | 0.0 |
| PH | 2,567 | 0.2 | 8.5 | 8.5 | 0.2 | 7.9 | 9.4 | -0.3 | 0.0 |
| Oxygen.Filler | 2,559 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 2.7 | 0.0 |
| Bowl.Setpoint | 2,569 | 0.1 | 109.3 | 120.0 | 15.3 | 70.0 | 140.0 | -1.0 | 0.3 |
| Pressure.Setpoint | 2,559 | 0.5 | 47.6 | 46.0 | 2.0 | 44.0 | 52.0 | 0.2 | 0.0 |
| Air.Pressurer | 2,571 | 0.0 | 142.8 | 142.6 | 1.2 | 140.8 | 148.2 | 2.3 | 0.0 |
| Alch.Rel | 2,562 | 0.3 | 6.9 | 6.6 | 0.5 | 5.3 | 8.6 | 0.9 | 0.0 |
| Carb.Rel | 2,561 | 0.4 | 5.4 | 5.4 | 0.1 | 5.0 | 6.1 | 0.5 | 0.0 |
| Balling.Lvl | 2,570 | 0.0 | 2.1 | 1.5 | 0.9 | 0.0 | 3.7 | 0.6 | 0.0 |

In the header the abbreviations stand for observations (obs), percent missing (missing), median, mean, standard deviation, minimum, maximum, range, skew, and standard error. These are calculated for each variable

plot(ph$PH)  
numerics <- ph %>%   
 dplyr::select(is.numeric)   
res <- cor(numerics, use = "complete.obs")  
round(res, 2)  
corrplot::corrplot(res, type = "upper", order = "hclust",   
 tl.col = "black", tl.srt = 45)  
numerics[is.na(numerics)] <- median(numerics, na.rm=TRUE)  
ph %>%   
 select(PH, Oxygen.Filler) %>%   
 mutate(ValueIndex = 1:length(ph$PH)) %>%   
 ggplot() +   
 geom\_point(aes(ValueIndex, PH, color = Oxygen.Filler))  
ggplot(ph, aes(PH)) + geom\_histogram()   
ggplot(ph, aes(PH, PH)) + geom\_violin()

ggplot() + geom\_boxplot(aes(ph$PH))

# Data Preparation

# replaces NA with median (given a removal of missing values in calculation)  
for (i in colnames(ph)) {  
 ph[[i]][is.na(ph[[i]])] <- median(ph[[i]], na.rm=TRUE)  
}  
sum(is.na(ph))

# select numeric variables  
ph.numerics <- ph %>%   
 dplyr::select(is.numeric)  
# remove outliers based on IQR  
for (i in colnames(ph.numerics)) {  
 iqr <- IQR(ph.numerics[[i]])  
 q <- quantile(ph.numerics[[i]], probs = c(0.25, 0.75), na.rm = FALSE)  
 qupper <- q[2]+1.5\*iqr  
 qlower <- q[1]+1.5\*iqr  
 outlier\_free <- subset(ph.numerics, ph.numerics[[i]] > (q[1] - 1.5\*iqr) & ph.numerics[[i]] < (q[2]+1.5\*iqr) )  
}  
ph.numerics <- outlier\_free  
# join outlier free numerics with categorical   
Brand.Code <- ph$ï..Brand.Code  
df <- cbind(Brand.Code, ph.numerics)  
sum(is.na(df$Brand.Code))

# Model Building

model.pm <- lm(PH~.,df)  
pm <- stepAIC(model.pm, trace = F, direction = "both")  
p <- summary(pm)$call  
pm <- lm(p[2], df)  
summary(pm)

model.ks <- lm(PH~., df)  
summary(model.ks)

data <- df %>%   
 dplyr::select(Brand.Code,   
 Carb.Volume,  
 Fill.Ounces,  
 PC.Volume,  
 Carb.Temp,  
 PSC,   
 PSC.Fill,   
 PSC.CO2,  
 PH,  
 Mnf.Flow,   
 Carb.Pressure1,   
 Fill.Pressure,   
 Hyd.Pressure2,   
 Hyd.Pressure3,   
 Filler.Level,  
 Temperature,   
 Usage.cont,  
 Carb.Flow,  
 Density,  
 Balling,  
 Pressure.Vacuum,   
 Oxygen.Filler,   
 Bowl.Setpoint,   
 Pressure.Setpoint,   
 Alch.Rel,  
 Balling.Lvl)

# Split 70-30 training test  
set.seed(1102)  
index <- createDataPartition(data$PH, p = .7,   
 list = FALSE, times = 1)  
train <- data[index,]  
test <- data[-index,]

# Model Selection

# Predictions

# Conclusion