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 are noted. We view a random sample of the data below.

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 <- obs.sample %>%   
 mutate(Variable = colnames(ph)) %>%  
 dplyr::select(Variable, 1, 2, 3, 4, 5)  
flextable(obs.sample) %>%   
 theme\_vanilla() %>%   
 set\_table\_properties(layout = "autofit")

| **Variable** | **1** | **2** | **3** | **4** | **5** |
| --- | --- | --- | --- | --- | --- |
| ï..Brand.Code | B | A | D | A | B |
| Carb.Volume | 5.346667 | 5.460000 | 5.566667 | 5.486667 | 5.333333 |
| Fill.Ounces | 23.96000 |  | 23.90667 | 23.96667 | 23.90667 |
| PC.Volume | 0.2666667 | 0.1866667 | 0.2773333 | 0.2426667 | 0.2680000 |
| Carb.Pressure | 69.0 | 70.6 | 74.0 | 69.2 | 66.4 |
| Carb.Temp | 143.6 | 141.8 | 143.6 | 140.0 | 139.8 |
| PSC | 0.014 | 0.024 | 0.142 | 0.126 | 0.032 |
| PSC.Fill | 0.08 | 0.28 | 0.10 | 0.08 | 0.20 |
| PSC.CO2 | 0.08 | 0.06 | 0.02 | 0.06 | 0.02 |
| Mnf.Flow | -100.0 | -100.2 | 110.0 | 143.6 | -100.2 |
| Carb.Pressure1 | 121.4 | 119.0 | 124.8 | 124.6 | 121.0 |
| Fill.Pressure | 46.2 | 46.0 | 50.0 | 56.0 | 45.8 |
| Hyd.Pressure1 | 0.0 | 33.4 | 12.6 | 17.8 | 0.0 |
| Hyd.Pressure2 | 0.0 | 38.8 | 27.2 | 33.2 | 0.0 |
| Hyd.Pressure3 | 0.0 | 27.4 | 41.0 | 31.2 | 0.0 |
| Hyd.Pressure4 | 94 | 106 | 88 | 118 | 98 |
| Filler.Level | 120.2 | 120.6 | 79.8 | 70.4 | 119.8 |
| Filler.Speed | 4010 | 3908 | 3808 | 4002 | 3982 |
| Temperature | 65.4 | 65.4 | 64.6 | 66.0 | 67.6 |
| Usage.cont | 19.10 | 23.02 | 16.40 | 24.32 | 19.16 |
| Carb.Flow | 2998 | 2980 | 3288 | 3492 | 2972 |
| Density | 0.92 | 1.70 | 1.68 | 1.58 | 0.88 |
| MFR | 732.8 | 708.2 | 686.8 | 731.6 | 726.4 |
| Balling | 1.498 | 3.440 | 3.390 | 3.142 | 1.398 |
| Pressure.Vacuum | -4.8 | -5.6 | -5.4 | -5.0 | -5.0 |
| PH | 8.80 | 8.60 | 8.66 | 8.30 | 8.72 |
| Oxygen.Filler | 0.024 | 0.062 | 0.049 | 0.045 | 0.056 |
| Bowl.Setpoint | 120 | 120 | 80 | 70 | 120 |
| Pressure.Setpoint | 46 | 46 | 50 | 50 | 46 |
| Air.Pressurer | 142.8 | 141.8 | 147.4 | 141.8 | 141.8 |
| Alch.Rel | 6.52 | 7.06 | 7.76 | 7.10 | 6.54 |
| Carb.Rel | 5.42 | 5.50 | 5.64 | 5.52 | 5.42 |
| Balling.Lvl | 1.42 | 3.22 | 3.06 | 3.12 | 1.40 |

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

| **obs** | **missing** | **median** | **mean** | **sd** | **min** | **max** | **range** | **skew** | **se** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2,571 | 0.00 | 3.0 | 3.4 | 1.1 | 1.0 | 5.0 | 4.0 | 0.0 | 0.0 |
| 2,561 | 0.39 | 5.3 | 5.4 | 0.1 | 5.0 | 5.7 | 0.7 | 0.4 | 0.0 |
| 2,533 | 1.48 | 24.0 | 24.0 | 0.1 | 23.6 | 24.3 | 0.7 | 0.0 | 0.0 |
| 2,532 | 1.52 | 0.3 | 0.3 | 0.1 | 0.1 | 0.5 | 0.4 | 0.3 | 0.0 |
| 2,544 | 1.05 | 68.2 | 68.2 | 3.5 | 57.0 | 79.4 | 22.4 | 0.2 | 0.1 |
| 2,545 | 1.01 | 140.8 | 141.1 | 4.0 | 128.6 | 154.0 | 25.4 | 0.2 | 0.1 |
| 2,538 | 1.28 | 0.1 | 0.1 | 0.0 | 0.0 | 0.3 | 0.3 | 0.8 | 0.0 |
| 2,548 | 0.89 | 0.2 | 0.2 | 0.1 | 0.0 | 0.6 | 0.6 | 0.9 | 0.0 |
| 2,532 | 1.52 | 0.0 | 0.1 | 0.0 | 0.0 | 0.2 | 0.2 | 1.7 | 0.0 |
| 2,569 | 0.08 | 65.2 | 24.6 | 119.5 | -100.2 | 229.4 | 329.6 | 0.0 | 2.4 |
| 2,539 | 1.24 | 123.2 | 122.6 | 4.7 | 105.6 | 140.2 | 34.6 | 0.1 | 0.1 |
| 2,549 | 0.86 | 46.4 | 47.9 | 3.2 | 34.6 | 60.4 | 25.8 | 0.5 | 0.1 |
| 2,560 | 0.43 | 11.4 | 12.4 | 12.4 | -0.8 | 58.0 | 58.8 | 0.8 | 0.2 |
| 2,556 | 0.58 | 28.6 | 21.0 | 16.4 | 0.0 | 59.4 | 59.4 | -0.3 | 0.3 |
| 2,556 | 0.58 | 27.6 | 20.5 | 16.0 | -1.2 | 50.0 | 51.2 | -0.3 | 0.3 |
| 2,541 | 1.17 | 96.0 | 96.3 | 13.1 | 52.0 | 142.0 | 90.0 | 0.5 | 0.3 |
| 2,551 | 0.78 | 118.4 | 109.3 | 15.7 | 55.8 | 161.2 | 105.4 | -0.8 | 0.3 |
| 2,514 | 2.22 | 3,982.0 | 3,687.2 | 770.8 | 998.0 | 4,030.0 | 3,032.0 | -2.9 | 15.4 |
| 2,557 | 0.54 | 65.6 | 66.0 | 1.4 | 63.6 | 76.2 | 12.6 | 2.4 | 0.0 |
| 2,566 | 0.19 | 21.8 | 21.0 | 3.0 | 12.1 | 25.9 | 13.8 | -0.5 | 0.1 |
| 2,569 | 0.08 | 3,028.0 | 2,468.4 | 1,073.7 | 26.0 | 5,104.0 | 5,078.0 | -1.0 | 21.2 |
| 2,570 | 0.04 | 1.0 | 1.2 | 0.4 | 0.2 | 1.9 | 1.7 | 0.5 | 0.0 |
| 2,359 | 8.25 | 724.0 | 704.0 | 73.9 | 31.4 | 868.6 | 837.2 | -5.1 | 1.5 |
| 2,570 | 0.04 | 1.6 | 2.2 | 0.9 | -0.2 | 4.0 | 4.2 | 0.6 | 0.0 |
| 2,571 | 0.00 | -5.4 | -5.2 | 0.6 | -6.6 | -3.6 | 3.0 | 0.5 | 0.0 |
| 2,567 | 0.16 | 8.5 | 8.5 | 0.2 | 7.9 | 9.4 | 1.5 | -0.3 | 0.0 |
| 2,559 | 0.47 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 0.4 | 2.7 | 0.0 |
| 2,569 | 0.08 | 120.0 | 109.3 | 15.3 | 70.0 | 140.0 | 70.0 | -1.0 | 0.3 |
| 2,559 | 0.47 | 46.0 | 47.6 | 2.0 | 44.0 | 52.0 | 8.0 | 0.2 | 0.0 |
| 2,571 | 0.00 | 142.6 | 142.8 | 1.2 | 140.8 | 148.2 | 7.4 | 2.3 | 0.0 |
| 2,562 | 0.35 | 6.6 | 6.9 | 0.5 | 5.3 | 8.6 | 3.3 | 0.9 | 0.0 |
| 2,561 | 0.39 | 5.4 | 5.4 | 0.1 | 5.0 | 6.1 | 1.1 | 0.5 | 0.0 |
| 2,570 | 0.04 | 1.5 | 2.1 | 0.9 | 0.0 | 3.7 | 3.7 | 0.6 | 0.0 |

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