# Checking Column Requirements for DB

Loading the correct libraries and dataset

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(readxl)  
Class\_Dataset\_for\_digital\_badge <- read\_excel("Class Dataset for digital badge.xlsx")

Checking the Requirements

count\_numeric <- length(select\_if(Class\_Dataset\_for\_digital\_badge,is.numeric))   
count\_cat <- length(select\_if(Class\_Dataset\_for\_digital\_badge,is.character))  
  
  
  
if (count\_numeric < 2 & count\_cat < 2){  
 print("There must be at least two categorical columns and two numerical columns")  
} else if (count\_numeric < 2){  
 print('There needs to be at least two numerical columns')  
}else if (count\_cat < 2){  
 print('There needs to be at least two categorical columns')  
}else {  
 print('The requirements have been satisifed')   
 }

## [1] "The requirements have been satisifed"

Some Statistical Tests

Age\_sum <- summary(Class\_Dataset\_for\_digital\_badge$Age)  
print(Age\_sum)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.00 17.00 18.00 17.85 19.00 20.00

t.test(Class\_Dataset\_for\_digital\_badge$Pre\_Weight,Class\_Dataset\_for\_digital\_badge$Post\_Weight)

##   
## Welch Two Sample t-test  
##   
## data: Class\_Dataset\_for\_digital\_badge$Pre\_Weight and Class\_Dataset\_for\_digital\_badge$Post\_Weight  
## t = -1.1417, df = 98.735, p-value = 0.2563  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.758579 1.013125  
## sample estimates:  
## mean of x mean of y   
## 82.15455 83.52727

t.test(Class\_Dataset\_for\_digital\_badge$Pre\_Weight)

##   
## One Sample t-test  
##   
## data: Class\_Dataset\_for\_digital\_badge$Pre\_Weight  
## t = 116.02, df = 54, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 80.73486 83.57423  
## sample estimates:  
## mean of x   
## 82.15455

summary(lm(data=Class\_Dataset\_for\_digital\_badge, formula = Post\_Weight ~ Gender + Age + Treatment))

##   
## Call:  
## lm(formula = Post\_Weight ~ Gender + Age + Treatment, data = Class\_Dataset\_for\_digital\_badge)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.383 -3.796 -1.127 3.846 16.855   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 73.8415 11.9603 6.174 1.09e-07 \*\*\*  
## GenderM -1.1372 1.8851 -0.603 0.5490   
## Age 0.7021 0.6632 1.059 0.2948   
## TreatmentCont -4.7147 1.8808 -2.507 0.0154 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.914 on 51 degrees of freedom  
## Multiple R-squared: 0.1307, Adjusted R-squared: 0.07956   
## F-statistic: 2.556 on 3 and 51 DF, p-value: 0.06545

BadgeDB2 <- pivot\_longer(data=Class\_Dataset\_for\_digital\_badge, cols = 'Pre\_Weight':'Post\_Weight', values\_to = 'Weight', names\_to = "Pre/Post")  
  
BadgeDB2$PatientID <- NA  
count = 1  
ID = 1  
for (i in seq\_len(nrow(BadgeDB2))){  
 if (count %% 2 == 0) {  
 BadgeDB2$PatientID[i] <-ID  
 ID <- ID +1  
 count <- count -1  
 }  
 else{  
 BadgeDB2$PatientID[i] <-ID  
 count <- count + 1  
 }  
   
}  
  
  
ggplot(data = BadgeDB2[1:20,], mapping = aes( x=PatientID ,y = Weight,   
 color = `Pre/Post`)) +  
 geom\_point() +  
 scale\_y\_continuous(name="Weight in LBs") +  
 labs(title = 'Changes in Weight', x = "Patient ID")

