NEW STUDY IN R

KERAMA JAMES THE DATA ANALYST

2023-04-23

#Another study in R concerning with the dataset cancer  
#First of everything load the library that contains the dataset to enhance the smooth running of the workouts  
library(survival)  
#Secondly load the dataset itself to portray the dataset content  
data("cancer")  
#Type the view command to see the dataset composition  
View(cancer)  
#Get the summary of the dataset to better comprehend what the dataset entails  
summary(cancer)

## inst time status age   
## Min. : 1.00 Min. : 5.0 Min. :1.000 Min. :39.00   
## 1st Qu.: 3.00 1st Qu.: 166.8 1st Qu.:1.000 1st Qu.:56.00   
## Median :11.00 Median : 255.5 Median :2.000 Median :63.00   
## Mean :11.09 Mean : 305.2 Mean :1.724 Mean :62.45   
## 3rd Qu.:16.00 3rd Qu.: 396.5 3rd Qu.:2.000 3rd Qu.:69.00   
## Max. :33.00 Max. :1022.0 Max. :2.000 Max. :82.00   
## NA's :1   
## sex ph.ecog ph.karno pat.karno   
## Min. :1.000 Min. :0.0000 Min. : 50.00 Min. : 30.00   
## 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.: 75.00 1st Qu.: 70.00   
## Median :1.000 Median :1.0000 Median : 80.00 Median : 80.00   
## Mean :1.395 Mean :0.9515 Mean : 81.94 Mean : 79.96   
## 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.: 90.00 3rd Qu.: 90.00   
## Max. :2.000 Max. :3.0000 Max. :100.00 Max. :100.00   
## NA's :1 NA's :1 NA's :3   
## meal.cal wt.loss   
## Min. : 96.0 Min. :-24.000   
## 1st Qu.: 635.0 1st Qu.: 0.000   
## Median : 975.0 Median : 7.000   
## Mean : 928.8 Mean : 9.832   
## 3rd Qu.:1150.0 3rd Qu.: 15.750   
## Max. :2600.0 Max. : 68.000   
## NA's :47 NA's :14

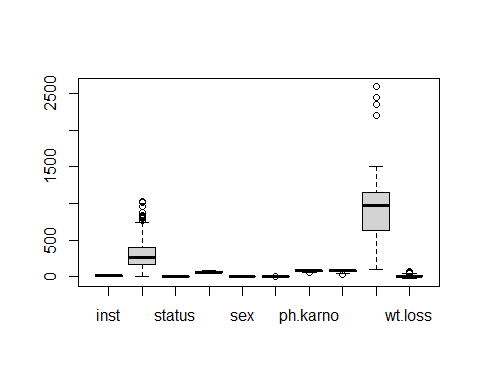
#Get the structure of the dataset to better comprehend what kind of data types the dataset contains  
str(cancer)

## 'data.frame': 228 obs. of 10 variables:  
## $ inst : num 3 3 3 5 1 12 7 11 1 7 ...  
## $ time : num 306 455 1010 210 883 ...  
## $ status : num 2 2 1 2 2 1 2 2 2 2 ...  
## $ age : num 74 68 56 57 60 74 68 71 53 61 ...  
## $ sex : num 1 1 1 1 1 1 2 2 1 1 ...  
## $ ph.ecog : num 1 0 0 1 0 1 2 2 1 2 ...  
## $ ph.karno : num 90 90 90 90 100 50 70 60 70 70 ...  
## $ pat.karno: num 100 90 90 60 90 80 60 80 80 70 ...  
## $ meal.cal : num 1175 1225 NA 1150 NA ...  
## $ wt.loss : num NA 15 15 11 0 0 10 1 16 34 ...

#Get the total number of missing values in the dataset of concern  
sum(is.na(cancer))

## [1] 67

#Workout the outliers concept using the boxplot idea  
boxplot(cancer)  
summary(boxplot(cancer))



## Length Class Mode   
## stats 50 -none- numeric   
## n 10 -none- numeric   
## conf 20 -none- numeric   
## out 28 -none- numeric   
## group 28 -none- numeric   
## names 10 -none- character

#Do away with the missing values by deletion of the NAS by omit concept  
cancer <- na.omit(cancer)  
#Do away with the missing values with the use of the tidyr library and the drop concept  
library(tidyr)

## Warning: package 'tidyr' was built under R version 4.2.3

#Creation of the new dataset from the old dataset that has no missing values  
clean <- drop\_na(cancer)  
View(clean)  
summary(clean)

## inst time status age   
## Min. : 1.00 Min. : 5.0 Min. :1.000 Min. :39.00   
## 1st Qu.: 3.00 1st Qu.: 174.5 1st Qu.:1.000 1st Qu.:57.00   
## Median :11.00 Median : 268.0 Median :2.000 Median :64.00   
## Mean :10.71 Mean : 309.9 Mean :1.719 Mean :62.57   
## 3rd Qu.:15.00 3rd Qu.: 419.5 3rd Qu.:2.000 3rd Qu.:70.00   
## Max. :32.00 Max. :1022.0 Max. :2.000 Max. :82.00   
## sex ph.ecog ph.karno pat.karno   
## Min. :1.000 Min. :0.0000 Min. : 50.00 Min. : 30.00   
## 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.: 70.00 1st Qu.: 70.00   
## Median :1.000 Median :1.0000 Median : 80.00 Median : 80.00   
## Mean :1.383 Mean :0.9581 Mean : 82.04 Mean : 79.58   
## 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.: 90.00 3rd Qu.: 90.00   
## Max. :2.000 Max. :3.0000 Max. :100.00 Max. :100.00   
## meal.cal wt.loss   
## Min. : 96.0 Min. :-24.000   
## 1st Qu.: 619.0 1st Qu.: 0.000   
## Median : 975.0 Median : 7.000   
## Mean : 929.1 Mean : 9.719   
## 3rd Qu.:1162.5 3rd Qu.: 15.000   
## Max. :2600.0 Max. : 68.000

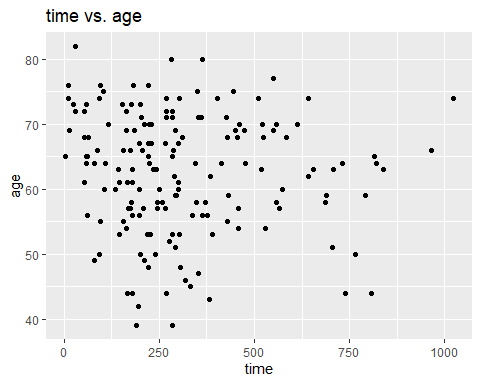
str(clean)

## 'data.frame': 167 obs. of 10 variables:  
## $ inst : num 3 5 12 7 11 1 7 6 12 22 ...  
## $ time : num 455 210 1022 310 361 ...  
## $ status : num 2 2 1 2 2 2 2 2 2 2 ...  
## $ age : num 68 57 74 68 71 53 61 57 57 70 ...  
## $ sex : num 1 1 1 2 2 1 1 1 1 1 ...  
## $ ph.ecog : num 0 1 1 2 2 1 2 1 1 1 ...  
## $ ph.karno : num 90 90 50 70 60 70 70 80 80 90 ...  
## $ pat.karno: num 90 60 80 60 80 80 70 80 70 100 ...  
## $ meal.cal : num 1225 1150 513 384 538 ...  
## $ wt.loss : num 15 11 0 10 1 16 34 27 60 -5 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:61] 1 3 5 12 13 14 16 20 23 25 ...  
## ..- attr(\*, "names")= chr [1:61] "1" "3" "5" "12" ...

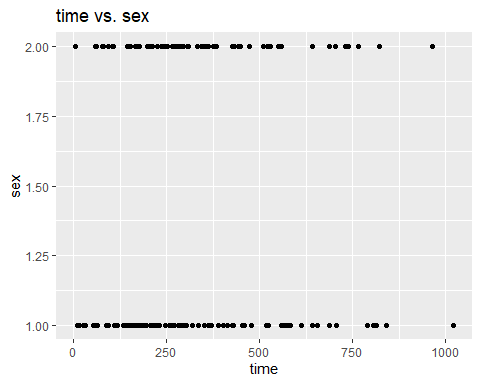
#Data visualization using the scatter plots  
# Load ggplot2 library  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

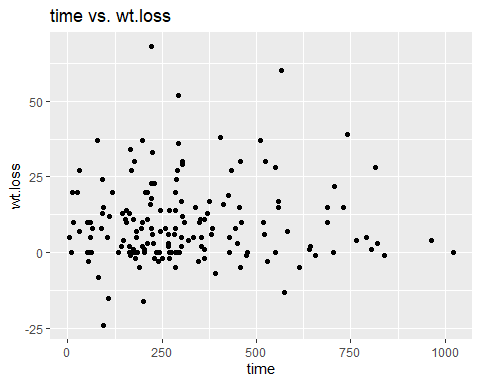
# Scatterplot of time and age  
ggplot(cancer, aes(x = time, y = age)) +   
 geom\_point() +   
 labs(title = "time vs. age", x = "time", y = "age")



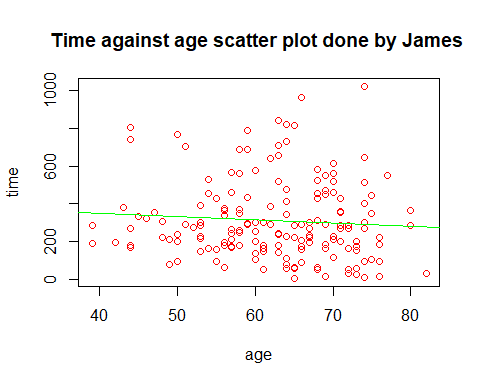
# Scatterplot of time and sex  
ggplot(cancer, aes(x = time, y = sex)) +   
 geom\_point() +   
 labs(title = "time vs. sex", x = "time", y = "sex")



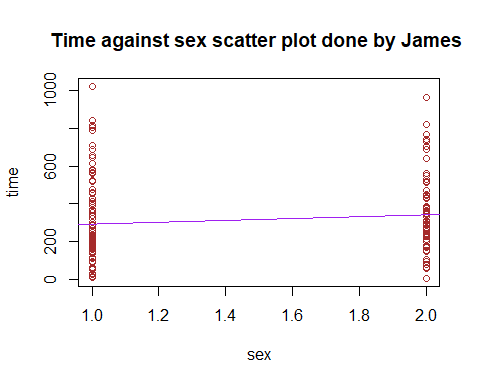
# Scatterplot of time and wt.loss  
ggplot(cancer, aes(x = time, y = wt.loss)) +   
 geom\_point() +   
 labs(title = "time vs. wt.loss", x = "time", y = "wt.loss")



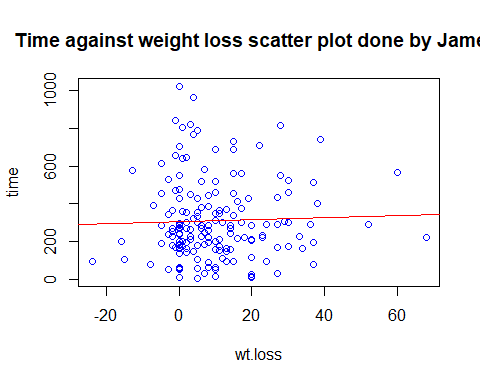
#Data modeling using the linear regression models  
# Fit linear regression models  
model1 <- lm(time ~ age, data = cancer)  
model2 <- lm(time ~ sex, data = cancer)  
model3 <- lm(time ~ wt.loss, data = cancer)  
# Visualize the models  
plot(time ~ age, data = cancer , main = "Time against age scatter plot done by James ", col = "red")  
abline(model1, col = "green")



plot(time ~ sex, data = cancer , main = "Time against sex scatter plot done by James ", col = "brown")  
abline(model2, col = "purple")



plot(time ~ wt.loss, data = cancer , main = "Time against weight loss scatter plot done by James ", col = "blue")  
abline(model3, col = "red")



# Display model summaries  
summary(model1)

##   
## Call:  
## lm(formula = time ~ age, data = cancer)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -300.59 -143.27 -53.88 110.05 732.48   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 421.676 111.598 3.779 0.00022 \*\*\*  
## age -1.786 1.765 -1.012 0.31301   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 209.4 on 165 degrees of freedom  
## Multiple R-squared: 0.006169, Adjusted R-squared: 0.0001456   
## F-statistic: 1.024 on 1 and 165 DF, p-value: 0.313

summary(model2)

##   
## Call:  
## lm(formula = time ~ sex, data = cancer)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -335.17 -132.16 -55.17 101.34 730.85   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 242.12 48.70 4.971 1.65e-06 \*\*\*  
## sex 49.03 33.22 1.476 0.142   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 208.7 on 165 degrees of freedom  
## Multiple R-squared: 0.01303, Adjusted R-squared: 0.007048   
## F-statistic: 2.178 on 1 and 165 DF, p-value: 0.1419

summary(model3)

##   
## Call:  
## lm(formula = time ~ wt.loss, data = cancer)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -303.31 -136.47 -40.42 105.49 717.15   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 304.8488 20.1016 15.16 <2e-16 \*\*\*  
## wt.loss 0.5233 1.2180 0.43 0.668   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 210 on 165 degrees of freedom  
## Multiple R-squared: 0.001117, Adjusted R-squared: -0.004937   
## F-statistic: 0.1846 on 1 and 165 DF, p-value: 0.6681

#Observation of the first and last six rows and columns of the new cancer dataset  
head(clean)

## inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss  
## 2 3 455 2 68 1 0 90 90 1225 15  
## 4 5 210 2 57 1 1 90 60 1150 11  
## 6 12 1022 1 74 1 1 50 80 513 0  
## 7 7 310 2 68 2 2 70 60 384 10  
## 8 11 361 2 71 2 2 60 80 538 1  
## 9 1 218 2 53 1 1 70 80 825 16

tail(clean)

## inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss  
## 221 22 197 1 67 1 1 80 90 1500 2  
## 222 11 203 1 71 2 1 80 90 1025 0  
## 225 13 191 1 39 1 0 90 90 2350 -5  
## 226 32 105 1 75 2 2 60 70 1025 5  
## 227 6 174 1 66 1 1 90 100 1075 1  
## 228 22 177 1 58 2 1 80 90 1060 0

#Observation of the firs and last six rows and columns of the old cancer dataset  
head(cancer)

## inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss  
## 2 3 455 2 68 1 0 90 90 1225 15  
## 4 5 210 2 57 1 1 90 60 1150 11  
## 6 12 1022 1 74 1 1 50 80 513 0  
## 7 7 310 2 68 2 2 70 60 384 10  
## 8 11 361 2 71 2 2 60 80 538 1  
## 9 1 218 2 53 1 1 70 80 825 16

tail(cancer)

## inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss  
## 221 22 197 1 67 1 1 80 90 1500 2  
## 222 11 203 1 71 2 1 80 90 1025 0  
## 225 13 191 1 39 1 0 90 90 2350 -5  
## 226 32 105 1 75 2 2 60 70 1025 5  
## 227 6 174 1 66 1 1 90 100 1075 1  
## 228 22 177 1 58 2 1 80 90 1060 0