AguerosR\_Assignment2.doc

Robert Agueros

5/20/2021

Q1: Load the data

df <- read.csv("https://stats.idre.ucla.edu/stat/data/rdm/patient\_pt1\_dm.csv")

Q2: Display the summary statistics of the data

summary(df)

## hospital hospid docid dis\_date   
## Length:120 Min. :1.000 Length:120 Length:120   
## Class :character 1st Qu.:1.000 Class :character Class :character   
## Mode :character Median :1.000 Mode :character Mode :character   
## Mean :1.483   
## 3rd Qu.:2.000   
## Max. :2.000   
## sex age test1 test2   
## Length:120 Min. : 34.19 Min. :-99.000 Min. :-99.000   
## Class :character 1st Qu.: 47.75 1st Qu.: 1.560 1st Qu.: 2.249   
## Mode :character Median : 51.83 Median : 3.107 Median : 4.162   
## Mean : 53.59 Mean : -1.989 Mean : -1.226   
## 3rd Qu.: 55.01 3rd Qu.: 5.707 3rd Qu.: 6.166   
## Max. :357.89 Max. : 12.416 Max. : 17.228   
## pain tumorsize co2 wound   
## Min. :1.000 Min. : 49.10 Min. :-98.0000 Min. :2.000   
## 1st Qu.:4.000 1st Qu.: 62.18 1st Qu.: 1.4954 1st Qu.:5.000   
## Median :5.000 Median : 68.23 Median : 1.5859 Median :6.000   
## Mean :5.325 Mean : 70.08 Mean : -0.8901 Mean :5.592   
## 3rd Qu.:6.000 3rd Qu.: 77.48 3rd Qu.: 1.6860 3rd Qu.:7.000   
## Max. :9.000 Max. :109.01 Max. : 1.9424 Max. :9.000   
## mobility ntumors remission lungcapacity   
## Min. :2.000 Min. :0.00 Min. :0.00 Min. :-99.0000   
## 1st Qu.:5.000 1st Qu.:1.00 1st Qu.:0.00 1st Qu.: 0.5141   
## Median :6.000 Median :2.00 Median :0.00 Median : 0.7457   
## Mean :6.033 Mean :3.15 Mean :0.35 Mean :-18.3280   
## 3rd Qu.:7.000 3rd Qu.:5.00 3rd Qu.:1.00 3rd Qu.: 0.8721   
## Max. :9.000 Max. :9.00 Max. :1.00 Max. : 0.9982   
## married familyhx smokinghx cancerstage   
## Min. :0.0000 Length:120 Length:120 Length:120   
## 1st Qu.:0.0000 Class :character Class :character Class :character   
## Median :1.0000 Mode :character Mode :character Mode :character   
## Mean :0.6333   
## 3rd Qu.:1.0000   
## Max. :1.0000   
## lengthofstay wbc rbc bmi   
## Min. :3.000 Length:120 Min. :4.360 Min. :18.45   
## 1st Qu.:4.000 Class :character 1st Qu.:4.825 1st Qu.:24.51   
## Median :5.000 Mode :character Median :4.978 Median :27.82   
## Mean :5.308 Mean :4.970 Mean :29.38   
## 3rd Qu.:6.000 3rd Qu.:5.150 3rd Qu.:34.31   
## Max. :8.000 Max. :5.535 Max. :58.00

Q3: Convert the data type of wbc to numeric.

df$wbc <- as.numeric(df$wbc)

## Warning: NAs introduced by coercion

is.numeric(df$wbc)

## [1] TRUE

str(df)

## 'data.frame': 120 obs. of 24 variables:  
## $ hospital : chr "UCLA" "UCLA" "UCLA" "UCLA" ...  
## $ hospid : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ docid : chr "1-1" "1-1" "1-1" "1-1" ...  
## $ dis\_date : chr "6-Sep-09" "7-Jan-11" "4-Sep-10" "25-Jun-10" ...  
## $ sex : chr "male" "female" "male" "male" ...  
## $ age : num 65 53.9 41.4 46.8 51.9 ...  
## $ test1 : num 3.7 2.63 -99 3.89 1.42 ...  
## $ test2 : num 8.086 0.803 2.126 1.349 2.195 ...  
## $ pain : int 4 2 3 3 4 3 4 5 4 5 ...  
## $ tumorsize : num 68 64.7 86.4 53.4 51.7 ...  
## $ co2 : num 1.53 1.68 1.45 1.57 1.42 ...  
## $ wound : int 4 3 3 4 5 4 4 5 5 5 ...  
## $ mobility : int 2 2 2 2 2 2 3 9 9 9 ...  
## $ ntumors : int 0 0 0 0 0 0 2 0 9 0 ...  
## $ remission : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ lungcapacity: num 0.801 0.326 0.848 0.886 -99 ...  
## $ married : int 0 0 0 0 1 1 1 1 0 1 ...  
## $ familyhx : chr "no" "no" "no" "no" ...  
## $ smokinghx : chr "former" "former" "former" "never" ...  
## $ cancerstage : chr "II" "II" "I" "II" ...  
## $ lengthofstay: int 6 6 5 6 5 4 6 4 7 4 ...  
## $ wbc : num 6088 6700 7163 6443 6801 ...  
## $ rbc : num 4.87 4.68 5.27 4.98 5.2 ...  
## $ bmi : num 24.1 29.4 21.6 29.8 27.1 ...

Q4: Display a frequency table for each of the following categorical variables: sex, smokinghx, and familyhx. Do you find any invalid values? Invalid values: dfsmokinghx is -99, and df$familyhx is -99.

table(df$sex)

##   
## 12.2 female male   
## 1 74 45

table(df$smokinghx)

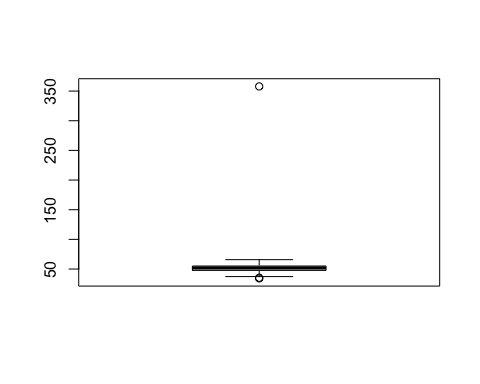
##   
## -99 current former never   
## 6 26 22 66

table(df$familyhx)

##   
## -99 no yes   
## 6 97 17

Q5: Display a boxplot for each of the following numerical variables: age, test1, test2, co2, and lungcapacity. Do you find any invalid values? Yes. Age: 357.89001, Test 1: -99.00000, Test2: -99.00000, Co2: -98, lungcapacity: -99 and -98.

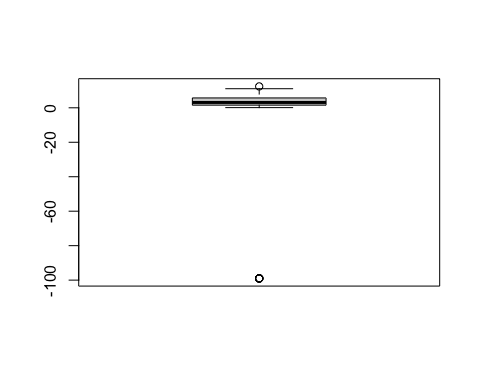
boxplot(df$age)



boxplot.stats(df$age)

## $stats  
## [1] 37.25225 47.70946 51.83233 55.05831 65.80417  
##   
## $n  
## [1] 120  
##   
## $conf  
## [1] 50.77238 52.89228  
##   
## $out  
## [1] 35.31930 357.89001 34.19229

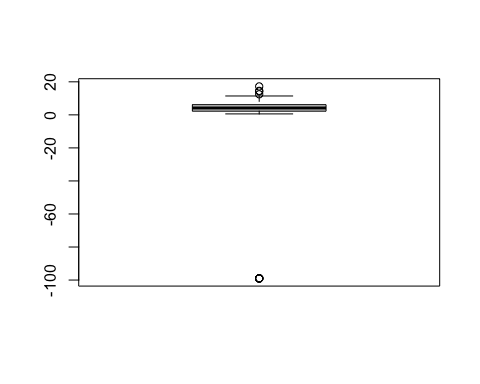
boxplot(df$test1)



boxplot.stats(df$test1)

## $stats  
## [1] 0.1048958 1.5348145 3.1065208 5.7227144 11.0714440  
##   
## $n  
## [1] 120  
##   
## $conf  
## [1] 2.502485 3.710557  
##   
## $out  
## [1] -99.00000 -99.00000 -99.00000 -99.00000 -99.00000 -99.00000 -99.00000  
## [8] 12.41639

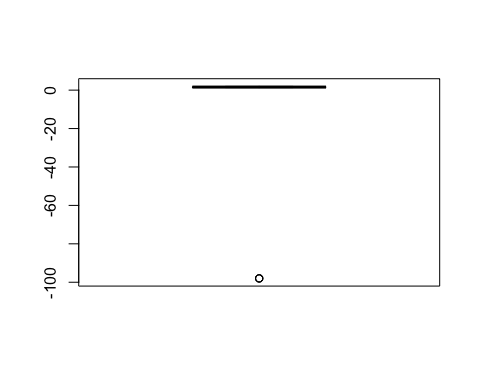
boxplot(df$test2)



boxplot.stats(df$test2)

## $stats  
## [1] 0.5807591 2.2311410 4.1620321 6.1669598 11.4315750  
##   
## $n  
## [1] 120  
##   
## $conf  
## [1] 3.594355 4.729709  
##   
## $out  
## [1] -99.00000 17.22758 -99.00000 -99.00000 12.44937 14.53654 -99.00000  
## [8] -99.00000 -99.00000 14.22523 -99.00000

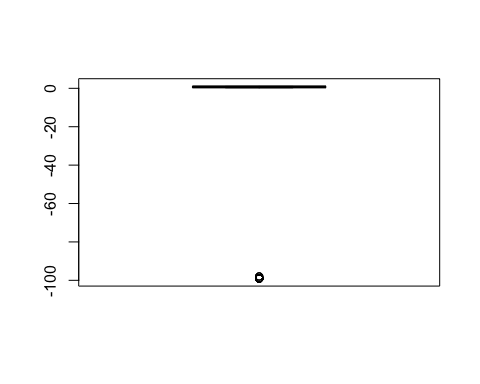
boxplot(df$co2)



boxplot.stats(df$co2)

## $stats  
## [1] 1.327440 1.495279 1.585878 1.689664 1.942401  
##   
## $n  
## [1] 120  
##   
## $conf  
## [1] 1.557841 1.613914  
##   
## $out  
## [1] -98 -98 -98

boxplot(df$lungcapacity)



boxplot.stats(df$lungcapacity)

## $stats  
## [1] 0.2949074 0.5056119 0.7456918 0.8769102 0.9982018  
##   
## $n  
## [1] 120  
##   
## $conf  
## [1] 0.6921381 0.7992455  
##   
## $out  
## [1] -99 -99 -98 -99 -99 -99 -99 -99 -99 -99 -98 -99 -99 -99 -99 -99 -99 -99 -99  
## [20] -98 -99 -99 -99

Q6: Recode all invalid values found in previous two steps as missing values (NA).

df$sex[df$sex==12.2] <- NA  
df$smokinghx[df$smokinghx== -99] <- NA  
df$familyhx[df$familyhx == -99] <- NA  
df$age[df$age == 357.8901] <- NA  
df$test1[df$test1 == -99.00000] <- NA  
df$test2[df$test2 == -99.00000] <- NA  
df$co2[df$co2 == -98] <- NA  
df$lungcapacity[df$lungcapacity == -99] <- NA  
df$lungcapacity[df$lungcapacity == -98] <- NA

Q7: Count the number of missing values in the dataset.

sum(is.na(df))

## [1] 55

Q8: Use its mean to replace missing values for each numeric variable, and use it most frequent value to replace missing values for each categorical variable.

df$age[is.na(df$age)] <- mean(df$age, na.rm = TRUE)  
df$test1[is.na(df$test1)] <- mean(df$test1, na.rm = TRUE)  
df$test2[is.na(df$test2)] <- mean(df$test2, na.rm = TRUE)  
df$co2[is.na(df$co2)] <- mean(df$co2, na.rm = TRUE)  
df$lungcapacity[is.na(df$lungcapacity)] <- mean(df$lungcapacity, na.rm = TRUE)  
df$wbc[is.na(df$wbc)] <- mean(df$wbc, na.rm = TRUE)  
df$sex[is.na(df$sex)] <- "female"  
df$familyhx[is.na(df$familyhx)] <- "no"  
df$smokinghx[is.na(df$smokinghx)] <- "never"

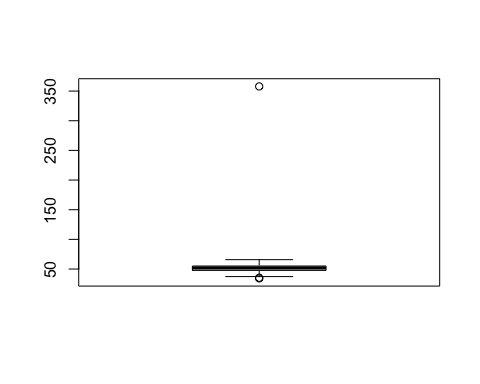
Q9: Count the number of missing values again to ensures all of them have been replaced.

sum(is.na(df))

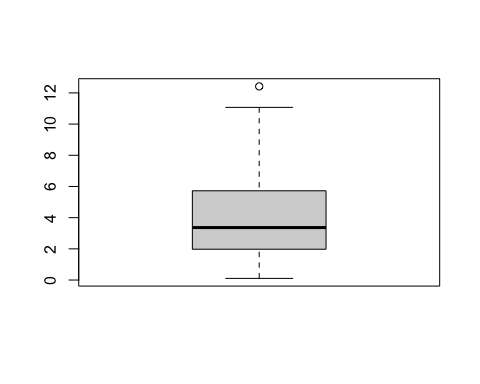
## [1] 0

Q10: Redo the boxplots in step 5 to see what they look like after invalid values have been replaced.

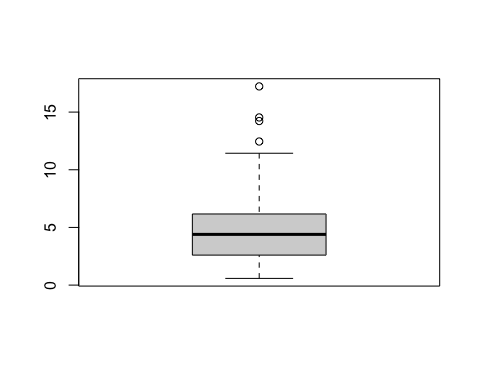
boxplot(df$age)



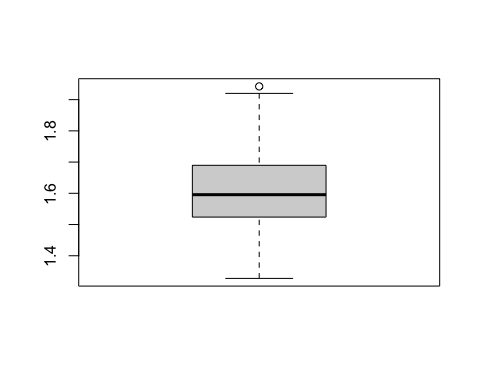
boxplot(df$test1)



boxplot(df$test2)



boxplot(df$co2)



boxplot(df$lungcapacity)

