Assignment\_3

Ajay Kanubhai Patel

2022-10-21

#Loading the package

#Load packages to convert file in PDF.  
  
if(!require(tinytex)){install.packages("tinytex")}

## Loading required package: tinytex

#This sets the working directory

This section is for the basic set up. It will clear all the plots, the console and the workspace. It also sets the overall format for numbers.

if(!is.null(dev.list())) dev.off()

## null device   
## 1

cat("\014")

rm(list=ls())  
options(scipen=9)

#To read Excel file in R data frame.

if(!require(readxl)){install.packages("readxl")}

## Loading required package: readxl

library("readxl")

#This sets the working directory

knitr::opts\_chunk$set(echo = TRUE)  
knitr::opts\_knit$set(root.dir = 'D:/Final Assignment/DATA/Assignment3')

This section is for the basic set up. It will clear all the plots, the console and the workspace. It also sets the overall format for numbers.

#To read Excel file in R data frame.

if(!require(readxl)){install.packages("readxl")}  
library("readxl")

if(!require(pastecs)){install.packages("pastecs")}

## Loading required package: pastecs

library("pastecs")  
  
if(!require(lattice)){install.packages("lattice")}

## Loading required package: lattice

library("lattice")  
  
if(!require(vcd)){install.packages("vcd")}

## Loading required package: vcd

## Loading required package: grid

library("vcd")  
  
if(!require(HSAUR)){install.packages("HSAUR")}

## Loading required package: HSAUR

## Loading required package: tools

library("HSAUR")  
  
if(!require(rmarkdown)){install.packages("rmarkdown")}

## Loading required package: rmarkdown

library("rmarkdown")  
  
  
if(!require(ggplot2)){install.packages("ggplot2")}

## Loading required package: ggplot2

library("ggplot2")

getwd()

## [1] "D:/Final Assignment/DATA/Assignment3"

Assignment03\_AP <- read\_excel("PROG8430\_Assign03\_22F.xlsx")  
  
  
Assignment03\_AP <- as.data.frame(Assignment03\_AP )

Initial Transformation

1. Rename all variables with your initials appended.

# to change all column name by appending my initials (Ajay Patel = AP) and separate it by “\_”. # head(data,n) displays 1st n rows present in our excel file. head(Assignment03\_BDSA,08)= shows first 8 rows. If number is not provided by default is shows 1st 6 rows

colnames(Assignment03\_AP) <- paste(colnames(Assignment03\_AP), "AP", sep = "\_")   
  
  
head(Assignment03\_AP)

## ID\_AP gender\_AP HR\_AP BP\_AP Wgt1\_AP Wgt2\_AP Exercise\_AP Hgt\_AP Smoke\_AP  
## 1 1 female Norm Norm 118.6 121.5 158 67.8 N  
## 2 2 female Norm Norm 143.1 146.6 152 65.9 N  
## 3 3 female Norm High 105.3 107.3 205 69.3 N  
## 4 4 female Norm Norm 119.5 120.9 151 65.4 N  
## 5 5 female Norm High 130.9 132.1 178 65.8 N  
## 6 6 female Norm High 90.0 91.4 204 63.1 N  
## Drink\_AP Group\_AP WBC\_AP Income\_AP  
## 1 Y Control 5193 125000  
## 2 Y Control 5705 NA  
## 3 N Test 7680 NA  
## 4 Y Control 7342 NA  
## 5 N Control 7714 NA  
## 6 Y Control 3851 NA

1. Transform character variables to factor variables.

# first we analyze structure of our data and then perform required operation.

str(Assignment03\_AP)

## 'data.frame': 500 obs. of 13 variables:  
## $ ID\_AP : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ gender\_AP : chr "female" "female" "female" "female" ...  
## $ HR\_AP : chr "Norm" "Norm" "Norm" "Norm" ...  
## $ BP\_AP : chr "Norm" "Norm" "High" "Norm" ...  
## $ Wgt1\_AP : num 119 143 105 120 131 ...  
## $ Wgt2\_AP : num 122 147 107 121 132 ...  
## $ Exercise\_AP: num 158 152 205 151 178 204 91 160 246 153 ...  
## $ Hgt\_AP : num 67.8 65.9 69.3 65.4 65.8 63.1 67.9 61.7 69.2 72.2 ...  
## $ Smoke\_AP : chr "N" "N" "N" "N" ...  
## $ Drink\_AP : chr "Y" "Y" "N" "Y" ...  
## $ Group\_AP : chr "Control" "Control" "Test" "Control" ...  
## $ WBC\_AP : num 5193 5705 7680 7342 7714 ...  
## $ Income\_AP : num 125000 NA NA NA NA NA NA NA NA NA ...

# to transform chr variables into factor variables.

# There are total 6 varibles in character form.

Assignment03\_AP$gender\_AP <- as.factor(Assignment03\_AP$gender\_AP)  
Assignment03\_AP$HR\_AP <- as.factor(Assignment03\_AP$HR\_AP)  
Assignment03\_AP$BP\_AP <- as.factor(Assignment03\_AP$BP\_AP)  
Assignment03\_AP$Smoke\_AP <- as.factor(Assignment03\_AP$Smoke\_AP)  
Assignment03\_AP$Drink\_AP <- as.factor(Assignment03\_AP$Drink\_AP)  
Assignment03\_AP$Group\_AP <- as.factor(Assignment03\_AP$Group\_AP)  
str(Assignment03\_AP)

## 'data.frame': 500 obs. of 13 variables:  
## $ ID\_AP : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ gender\_AP : Factor w/ 1 level "female": 1 1 1 1 1 1 1 1 1 1 ...  
## $ HR\_AP : Factor w/ 3 levels "High","Low","Norm": 3 3 3 3 3 3 3 3 3 3 ...  
## $ BP\_AP : Factor w/ 3 levels "High","Low","Norm": 3 3 1 3 1 1 3 3 3 3 ...  
## $ Wgt1\_AP : num 119 143 105 120 131 ...  
## $ Wgt2\_AP : num 122 147 107 121 132 ...  
## $ Exercise\_AP: num 158 152 205 151 178 204 91 160 246 153 ...  
## $ Hgt\_AP : num 67.8 65.9 69.3 65.4 65.8 63.1 67.9 61.7 69.2 72.2 ...  
## $ Smoke\_AP : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Drink\_AP : Factor w/ 2 levels "N","Y": 2 2 1 2 1 2 1 2 2 2 ...  
## $ Group\_AP : Factor w/ 2 levels "Control","Test": 1 1 2 1 1 1 2 1 1 1 ...  
## $ WBC\_AP : num 5193 5705 7680 7342 7714 ...  
## $ Income\_AP : num 125000 NA NA NA NA NA NA NA NA NA ...

Q1 (2).

Reduce Dimensionality

1. Apply the Missing Value Filter to remove appropriate columns of data.

# To identify missing values we apply summary()func and observe NA’s for

#each column.

#There are 492 records (almost 98.4%) in Income\_AP which are NA(Not Available).

#To drop Income\_AP column Assignment03\_AP[-c(13)].

summary(Assignment03\_AP)

## ID\_AP gender\_AP HR\_AP BP\_AP Wgt1\_AP   
## Min. : 1.0 female:500 High: 77 High: 94 Min. : 49.2   
## 1st Qu.:125.8 Low : 18 Low : 34 1st Qu.:114.3   
## Median :250.5 Norm:405 Norm:372 Median :131.7   
## Mean :250.5 Mean :131.0   
## 3rd Qu.:375.2 3rd Qu.:149.6   
## Max. :500.0 Max. :199.3   
##   
## Wgt2\_AP Exercise\_AP Hgt\_AP Smoke\_AP Drink\_AP  
## Min. : 50.2 Min. : 67.0 Min. :59.80 N:427 N:140   
## 1st Qu.:115.2 1st Qu.:147.0 1st Qu.:65.00 Y: 73 Y:360   
## Median :133.7 Median :175.5 Median :67.05   
## Mean :132.9 Mean :176.9 Mean :67.01   
## 3rd Qu.:151.6 3rd Qu.:207.0 3rd Qu.:69.00   
## Max. :201.5 Max. :297.0 Max. :78.90   
##   
## Group\_AP WBC\_AP Income\_AP   
## Control:250 Min. : 3851 Min. : 2000   
## Test :250 1st Qu.: 6195 1st Qu.: 7000   
## Median : 6902 Median : 53500   
## Mean : 6896 Mean : 66625   
## 3rd Qu.: 7532 3rd Qu.:128000   
## Max. :10652 Max. :148000   
## NA's :492

Assignment03\_AP <- Assignment03\_AP[-c(13)]  
  
head(Assignment03\_AP)

## ID\_AP gender\_AP HR\_AP BP\_AP Wgt1\_AP Wgt2\_AP Exercise\_AP Hgt\_AP Smoke\_AP  
## 1 1 female Norm Norm 118.6 121.5 158 67.8 N  
## 2 2 female Norm Norm 143.1 146.6 152 65.9 N  
## 3 3 female Norm High 105.3 107.3 205 69.3 N  
## 4 4 female Norm Norm 119.5 120.9 151 65.4 N  
## 5 5 female Norm High 130.9 132.1 178 65.8 N  
## 6 6 female Norm High 90.0 91.4 204 63.1 N  
## Drink\_AP Group\_AP WBC\_AP  
## 1 Y Control 5193  
## 2 Y Control 5705  
## 3 N Test 7680  
## 4 Y Control 7342  
## 5 N Control 7714  
## 6 Y Control 3851

1. Apply the Low Variance Filter to remove appropriate columns of data.

#only for numerical columns.

stat.desc(Assignment03\_AP) #Consider coef of var for the low variance,

## ID\_AP gender\_AP HR\_AP BP\_AP Wgt1\_AP Wgt2\_AP  
## nbr.val 500.0000000 NA NA NA 500.0000000 500.000000  
## nbr.null 0.0000000 NA NA NA 0.0000000 0.000000  
## nbr.na 0.0000000 NA NA NA 0.0000000 0.000000  
## min 1.0000000 NA NA NA 49.2000000 50.200000  
## max 500.0000000 NA NA NA 199.3000000 201.500000  
## range 499.0000000 NA NA NA 150.1000000 151.300000  
## sum 125250.0000000 NA NA NA 65478.0000000 66452.900000  
## median 250.5000000 NA NA NA 131.6500000 133.700000  
## mean 250.5000000 NA NA NA 130.9560000 132.905800  
## SE.mean 6.4614240 NA NA NA 1.1942880 1.198161  
## CI.mean.0.95 12.6949496 NA NA NA 2.3464527 2.354061  
## var 20875.0000000 NA NA NA 713.1618677 717.794415  
## std.dev 144.4818328 NA NA NA 26.7050907 26.791686  
## coef.var 0.5767738 NA NA NA 0.2039241 0.201584  
## Exercise\_AP Hgt\_AP Smoke\_AP Drink\_AP Group\_AP  
## nbr.val 500.0000000 500.00000000 NA NA NA  
## nbr.null 0.0000000 0.00000000 NA NA NA  
## nbr.na 0.0000000 0.00000000 NA NA NA  
## min 67.0000000 59.80000000 NA NA NA  
## max 297.0000000 78.90000000 NA NA NA  
## range 230.0000000 19.10000000 NA NA NA  
## sum 88440.0000000 33503.40000000 NA NA NA  
## median 175.5000000 67.05000000 NA NA NA  
## mean 176.8800000 67.00680000 NA NA NA  
## SE.mean 1.8917697 0.13089697 NA NA NA  
## CI.mean.0.95 3.7168156 0.25717712 NA NA NA  
## var 1789.3963928 8.56700778 NA NA NA  
## std.dev 42.3012576 2.92694513 NA NA NA  
## coef.var 0.2391523 0.04368131 NA NA NA  
## WBC\_AP  
## nbr.val 500.0000000  
## nbr.null 0.0000000  
## nbr.na 0.0000000  
## min 3851.0000000  
## max 10652.0000000  
## range 6801.0000000  
## sum 3447776.0000000  
## median 6902.0000000  
## mean 6895.5520000  
## SE.mean 46.1807869  
## CI.mean.0.95 90.7327494  
## var 1066332.5403768  
## std.dev 1032.6337881  
## coef.var 0.1497536

summary(Assignment03\_AP)

## ID\_AP gender\_AP HR\_AP BP\_AP Wgt1\_AP   
## Min. : 1.0 female:500 High: 77 High: 94 Min. : 49.2   
## 1st Qu.:125.8 Low : 18 Low : 34 1st Qu.:114.3   
## Median :250.5 Norm:405 Norm:372 Median :131.7   
## Mean :250.5 Mean :131.0   
## 3rd Qu.:375.2 3rd Qu.:149.6   
## Max. :500.0 Max. :199.3   
## Wgt2\_AP Exercise\_AP Hgt\_AP Smoke\_AP Drink\_AP  
## Min. : 50.2 Min. : 67.0 Min. :59.80 N:427 N:140   
## 1st Qu.:115.2 1st Qu.:147.0 1st Qu.:65.00 Y: 73 Y:360   
## Median :133.7 Median :175.5 Median :67.05   
## Mean :132.9 Mean :176.9 Mean :67.01   
## 3rd Qu.:151.6 3rd Qu.:207.0 3rd Qu.:69.00   
## Max. :201.5 Max. :297.0 Max. :78.90   
## Group\_AP WBC\_AP   
## Control:250 Min. : 3851   
## Test :250 1st Qu.: 6195   
## Median : 6902   
## Mean : 6896   
## 3rd Qu.: 7532   
## Max. :10652

# we can observe that Hgt\_AP has low variance.  
  
table(Assignment03\_AP$Hgt\_AP) # It displays how many records has same value

##   
## 59.8 59.9 60.3 60.6 60.7 60.8 60.9 61.2 61.5 61.6 61.7 61.8 61.9 62.1 62.2 62.4   
## 1 1 1 1 1 2 3 2 1 2 2 1 4 2 2 1   
## 62.5 62.6 62.7 62.9 63 63.1 63.2 63.3 63.4 63.5 63.6 63.7 63.8 63.9 64 64.1   
## 1 1 2 4 4 3 6 3 4 3 4 2 3 9 2 5   
## 64.2 64.3 64.4 64.5 64.6 64.7 64.8 65 65.1 65.2 65.3 65.4 65.5 65.6 65.7 65.8   
## 5 4 7 8 7 7 3 3 6 9 8 1 4 8 5 11   
## 65.9 66 66.1 66.2 66.3 66.4 66.5 66.6 66.7 66.8 66.9 67 67.1 67.2 67.3 67.4   
## 6 7 7 8 4 6 5 5 5 10 3 5 9 8 4 6   
## 67.5 67.6 67.7 67.8 67.9 68 68.1 68.2 68.3 68.4 68.5 68.6 68.7 68.8 68.9 69   
## 10 8 7 7 9 3 6 2 5 4 9 5 5 8 7 5   
## 69.1 69.2 69.3 69.4 69.5 69.6 69.7 69.8 69.9 70 70.1 70.2 70.3 70.4 70.5 70.6   
## 4 11 5 7 12 2 4 4 5 4 2 2 4 6 1 1   
## 70.7 70.8 70.9 71 71.1 71.2 71.3 71.4 71.5 71.6 71.7 71.8 71.9 72.1 72.2 72.3   
## 3 3 3 2 1 3 2 2 2 2 2 2 1 2 2 2   
## 72.4 72.5 72.6 72.8 72.9 73 74.1 74.7 75.6 77.5 78.4 78.9   
## 3 1 1 1 1 1 1 2 1 1 1 1

# For example, 9 records has 67.1 value in Hgt\_AP column.  
  
Assignment03\_AP <- Assignment03\_AP[-c(8)] #removes Hgt\_AP column.  
  
head(Assignment03\_AP)

## ID\_AP gender\_AP HR\_AP BP\_AP Wgt1\_AP Wgt2\_AP Exercise\_AP Smoke\_AP Drink\_AP  
## 1 1 female Norm Norm 118.6 121.5 158 N Y  
## 2 2 female Norm Norm 143.1 146.6 152 N Y  
## 3 3 female Norm High 105.3 107.3 205 N N  
## 4 4 female Norm Norm 119.5 120.9 151 N Y  
## 5 5 female Norm High 130.9 132.1 178 N N  
## 6 6 female Norm High 90.0 91.4 204 N Y  
## Group\_AP WBC\_AP  
## 1 Control 5193  
## 2 Control 5705  
## 3 Test 7680  
## 4 Control 7342  
## 5 Control 7714  
## 6 Control 3851

1. Apply the High Correlation Filter to remove appropriate columns of data.

#High correlation between two variables means they have similar trends and are #likely to carry similar information.

#No correlation available between numerical and nominal columns.

#pearson, spearman,kendall methos can be used to measure the degree of #association between two variables.

# can only check for numerical and we have 4 column with numeric data so

#n(n-1)/2 (4\*3/2 = 6) combination should be checked.

# I have checked by three methods just for knowledge.

#by spearman method  
#Speaman is non-parametric and therefore makes no normalacy assumption  
  
cor(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$Wgt2\_AP,method = "spearman")

## [1] 0.9990139

cor(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$Exercise\_AP,method = "spearman")

## [1] -0.1848344

cor(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$WBC\_AP,method = "spearman")

## [1] -0.002327948

cor(Assignment03\_AP$Wgt2\_AP,Assignment03\_AP$Exercise\_AP,method = "spearman")

## [1] -0.1808802

cor(Assignment03\_AP$Wgt2\_AP,Assignment03\_AP$WBC\_AP,method = "spearman")

## [1] -0.001722567

cor(Assignment03\_AP$Exercise\_AP,Assignment03\_AP$WBC\_AP,method = "spearman")

## [1] 0.08459301

#by pearson method  
#assumes normalacy  
  
cor(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$Wgt2\_AP,method = "pearson")

## [1] 0.9993236

cor(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$Exercise\_AP,method = "pearson")

## [1] -0.2027378

cor(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$WBC\_AP,method = "pearson")

## [1] 0.00158806

cor(Assignment03\_AP$Wgt2\_AP,Assignment03\_AP$Exercise\_AP,method = "pearson")

## [1] -0.1998686

cor(Assignment03\_AP$Wgt2\_AP,Assignment03\_AP$WBC\_AP,method = "pearson")

## [1] 0.001131027

cor(Assignment03\_AP$Exercise\_AP,Assignment03\_AP$WBC\_AP,method = "pearson")

## [1] 0.07299489

# by kendall method  
#Kendall rank correlation (non-parametric) is an alternative to Pearson's   
#correlation (parametric)  
  
cor(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$Wgt2\_AP,method = "kendall")

## [1] 0.9767528

cor(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$Exercise\_AP,method = "kendall")

## [1] -0.1257533

cor(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$WBC\_AP,method = "kendall")

## [1] -0.000344912

cor(Assignment03\_AP$Wgt2\_AP,Assignment03\_AP$Exercise\_AP,method = "kendall")

## [1] -0.1230769

cor(Assignment03\_AP$Wgt2\_AP,Assignment03\_AP$WBC\_AP,method = "kendall")

## [1] 0.0003930487

cor(Assignment03\_AP$Exercise\_AP,Assignment03\_AP$WBC\_AP,method = "kendall")

## [1] 0.05623635

#Wgt1\_AP and Wgt2\_AP are highly correlated to each other so going to drop # Wgt2\_AP column

Assignment03\_AP <- Assignment03\_AP[-c(6)]  
head(Assignment03\_AP)

## ID\_AP gender\_AP HR\_AP BP\_AP Wgt1\_AP Exercise\_AP Smoke\_AP Drink\_AP Group\_AP  
## 1 1 female Norm Norm 118.6 158 N Y Control  
## 2 2 female Norm Norm 143.1 152 N Y Control  
## 3 3 female Norm High 105.3 205 N N Test  
## 4 4 female Norm Norm 119.5 151 N Y Control  
## 5 5 female Norm High 130.9 178 N N Control  
## 6 6 female Norm High 90.0 204 N Y Control  
## WBC\_AP  
## 1 5193  
## 2 5705  
## 3 7680  
## 4 7342  
## 5 7714  
## 6 3851

1. Drop any variables that do not contribute any useful analytical information at all.

Answer: Here, I am going to drop gender\_AP columns as it contains only Females so it is not quite useful for analytic purpose.

Assignment03\_AP <- Assignment03\_AP[-c(2)]  
head(Assignment03\_AP)

## ID\_AP HR\_AP BP\_AP Wgt1\_AP Exercise\_AP Smoke\_AP Drink\_AP Group\_AP WBC\_AP  
## 1 1 Norm Norm 118.6 158 N Y Control 5193  
## 2 2 Norm Norm 143.1 152 N Y Control 5705  
## 3 3 Norm High 105.3 205 N N Test 7680  
## 4 4 Norm Norm 119.5 151 N Y Control 7342  
## 5 5 Norm High 130.9 178 N N Control 7714  
## 6 6 Norm High 90.0 204 N Y Control 3851

Q1 (3).

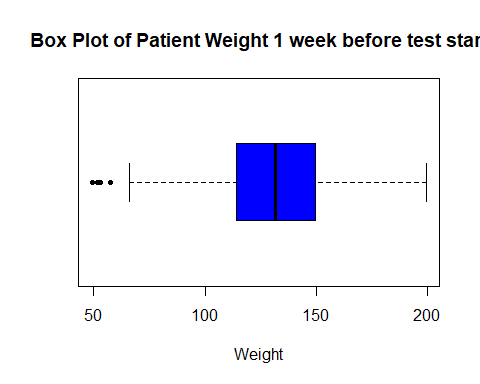
Outliers

1. Create boxplots of all relevant variables (i.e. numeric, non-binary) to determine outliers.

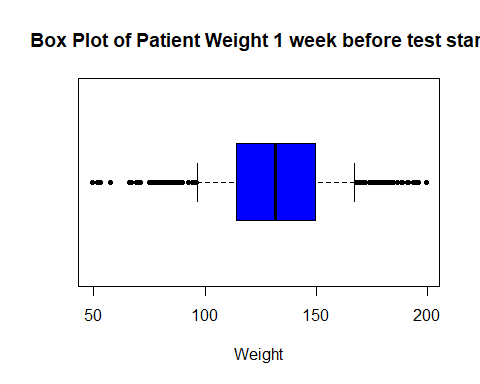
str(Assignment03\_AP)

## 'data.frame': 500 obs. of 9 variables:  
## $ ID\_AP : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ HR\_AP : Factor w/ 3 levels "High","Low","Norm": 3 3 3 3 3 3 3 3 3 3 ...  
## $ BP\_AP : Factor w/ 3 levels "High","Low","Norm": 3 3 1 3 1 1 3 3 3 3 ...  
## $ Wgt1\_AP : num 119 143 105 120 131 ...  
## $ Exercise\_AP: num 158 152 205 151 178 204 91 160 246 153 ...  
## $ Smoke\_AP : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Drink\_AP : Factor w/ 2 levels "N","Y": 2 2 1 2 1 2 1 2 2 2 ...  
## $ Group\_AP : Factor w/ 2 levels "Control","Test": 1 1 2 1 1 1 2 1 1 1 ...  
## $ WBC\_AP : num 5193 5705 7680 7342 7714 ...

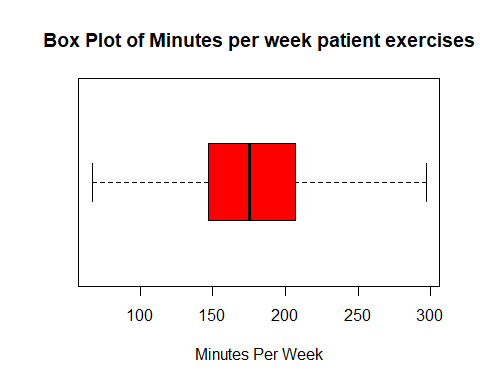
boxplot(Assignment03\_AP$Wgt1\_AP,  
 main="Box Plot of Patient Weight 1 week before test start",  
 xlab="Weight",  
 col="blue", horizontal=TRUE, pch=20)



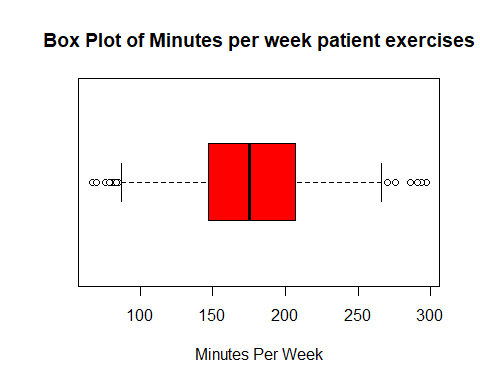
# let's shrink the graph and observe  
  
boxplot(Assignment03\_AP$Wgt1\_AP,  
 main="Box Plot of Patient Weight 1 week before test start",  
 xlab="Weight",  
 col="blue", horizontal=TRUE, pch=20, range = 0.5)



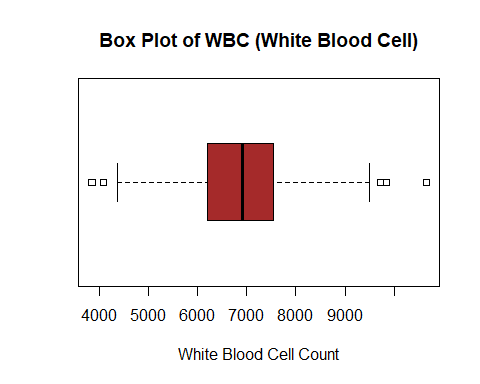
boxplot(Assignment03\_AP$Exercise\_AP,   
 main="Box Plot of Minutes per week patient exercises",  
 xlab="Minutes Per Week",  
 col="red", horizontal=TRUE, pch=21)



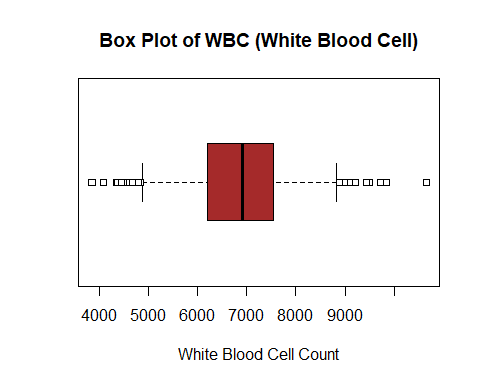
# let's shrink the graph and observe  
  
boxplot(Assignment03\_AP$Exercise\_AP,   
 main="Box Plot of Minutes per week patient exercises",  
 xlab="Minutes Per Week",  
 col="red", horizontal=TRUE, pch=21, range = 1)



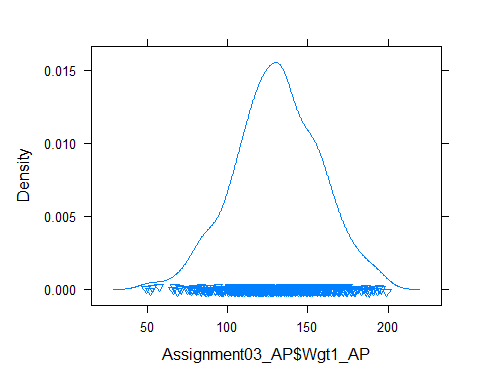
boxplot(Assignment03\_AP$WBC\_AP,   
 main="Box Plot of WBC (White Blood Cell)",  
 xlab="White Blood Cell Count",  
 col="brown",horizontal=TRUE, pch=22)



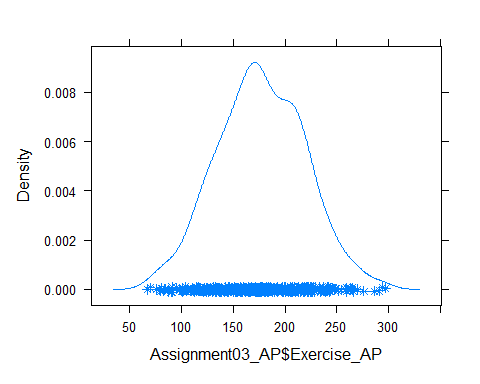
# let's shrink the graph and observe  
  
boxplot(Assignment03\_AP$WBC\_AP,   
 main="Box Plot of WBC (White Blood Cell)",  
 xlab="White Blood Cell Count",  
 col="brown",horizontal=TRUE, pch=22, range = 1)



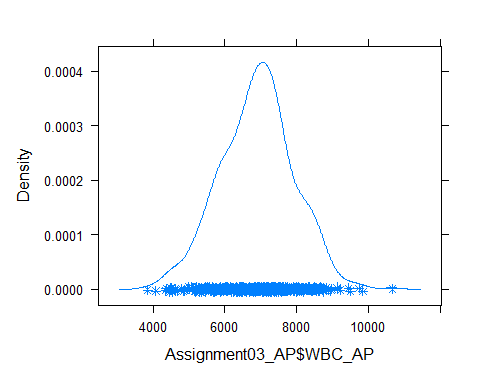
densityplot( ~ Assignment03\_AP$Wgt1\_AP, pch=6)



densityplot( ~ Assignment03\_AP$Exercise\_AP, pch=8)

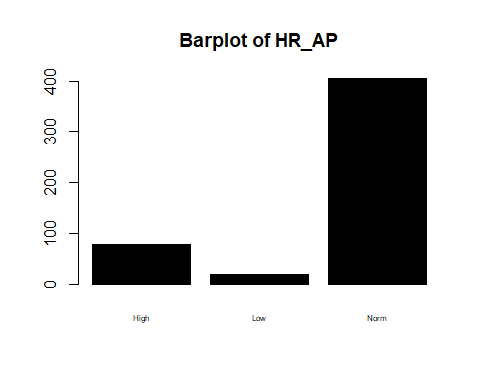


densityplot( ~ Assignment03\_AP$WBC\_AP, pch=8)

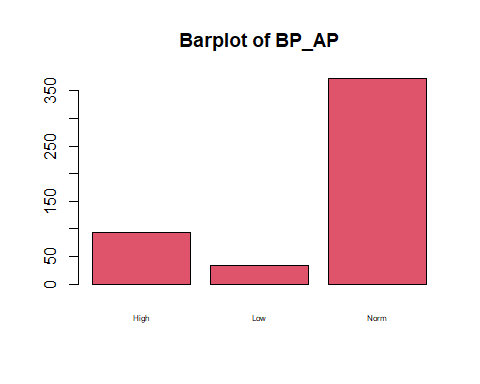


# For factor variables.(for non-binary variables only)

Assignment03\_AP$HR\_AP <- as.factor(Assignment03\_AP$HR\_AP)  
Assignment03\_AP$BP\_AP <- as.factor(Assignment03\_AP$BP\_AP)  
  
  
barplot(table(Assignment03\_AP$HR\_AP), cex.names=0.5, main="Barplot of HR\_AP",   
 col = 1)



barplot(table(Assignment03\_AP$BP\_AP), cex.names=0.5, main="Barplot of BP\_AP",   
 col = 2)

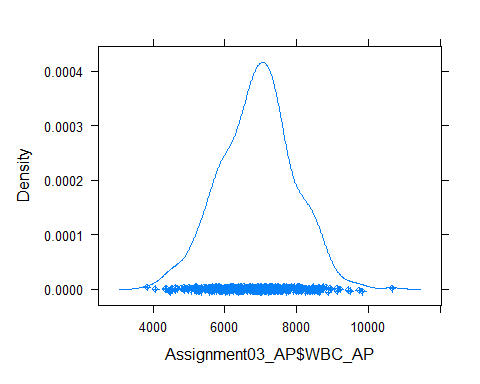


1. Comment on any outliers you see and deal with them appropriately.

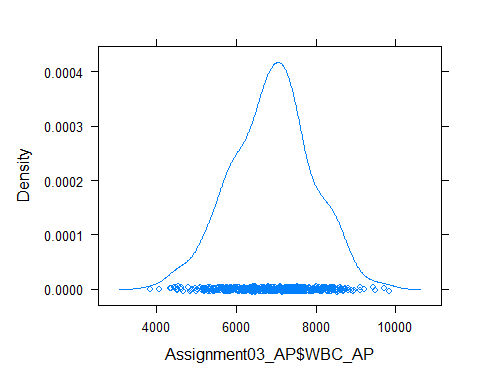
Conclusion:1. wgt1\_AP looks good. 2. Exercise\_AP looks good. 3. WBC\_AP has one outlier. low HR\_AP and low BP\_AP categories as comparatively small values than that of normal category still they are useful.

#To remove a outlier of WBC\_AP at its max value.

densityplot( ~ Assignment03\_AP$WBC\_AP, pch=10)



nr <- which(Assignment03\_AP$WBC\_AP == max(Assignment03\_AP$WBC\_AP))   
#above code is to detect Row number with max value   
Assignment03\_AP <- Assignment03\_AP[-c(nr),]  
densityplot( ~ Assignment03\_AP$WBC\_AP, pch=21)

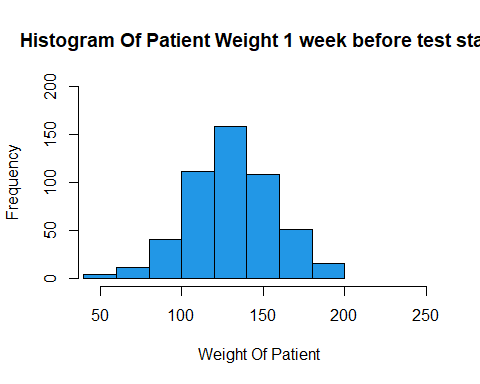


Q2.

Organizing Data

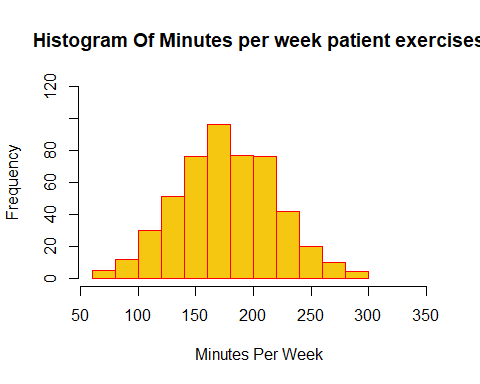
1. Scatter Plots
2. Create a histogram for one of the Weight variables.

hist(Assignment03\_AP$Wgt1\_AP,  
 main = "Histogram Of Patient Weight 1 week before test start",  
 xlab = "Weight Of Patient",  
 xlim = c(45,250),  
 ylim = c(0,200),  
 col = 4, border = "black")



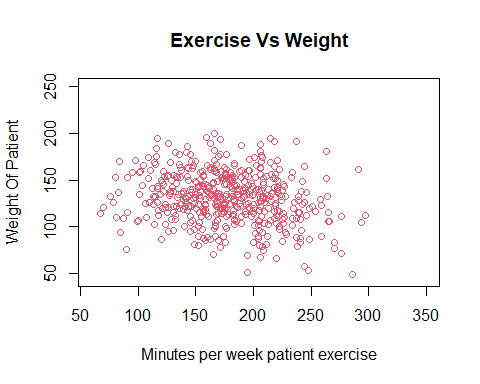
1. Create a histogram for Exercise.

hist(Assignment03\_AP$Exercise\_AP,  
 main = "Histogram Of Minutes per week patient exercises",  
 xlab = "Minutes Per Week",  
 xlim = c(60,350),  
 ylim = c(0,120),  
 col = 7, border="red")



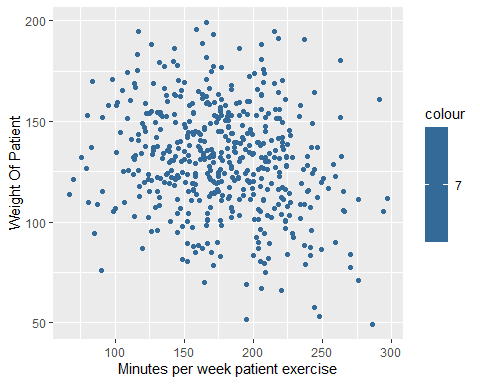
1. Create a scatter plot showing the relationship between Exercise and Weight. (note: Exercise should be on the x-axis, Weight should be the y-axis)

plot(Assignment03\_AP$Wgt1\_AP~Assignment03\_AP$Exercise\_AP,  
 main = "Exercise Vs Weight",  
 xlab = "Minutes per week patient exercise",  
 ylab = "Weight Of Patient",  
 xlim = c(60,350),   
 ylim = c(45,250),  
 col = 2)

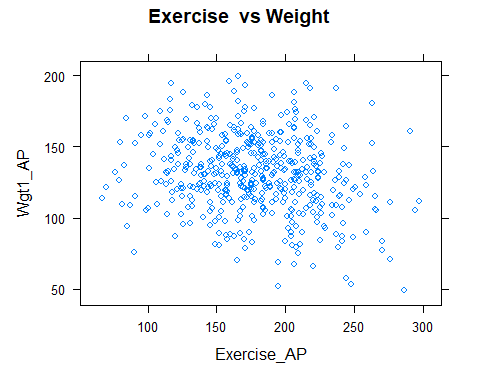


1. What conclusions, if any, can you draw from the chart? From the above chart (scatter plot), we can see that there is not any linear relation between these two variables.
2. Calculate a correlation coefficient between these two variables (Wgt1\_AP & Exercise\_AP) What conclusion you draw from it?

ggplot(data = Assignment03\_AP, aes(x = Exercise\_AP,   
 y = Wgt1\_AP,   
 color = 7)) +  
 geom\_point()+ labs(y = "Weight Of Patient",  
x = "Minutes per week patient exercise")



xyplot(Wgt1\_AP~Exercise\_AP, data=Assignment03\_AP, main="Exercise vs Weight",   
 colors=TRUE)



cor(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$Exercise\_AP)

## [1] -0.1993163

cor.test(Assignment03\_AP$Wgt1\_AP,Assignment03\_AP$Exercise\_AP,  
 method="spearman")

## Warning in cor.test.default(Assignment03\_AP$Wgt1\_AP,  
## Assignment03\_AP$Exercise\_AP, : Cannot compute exact p-value with ties

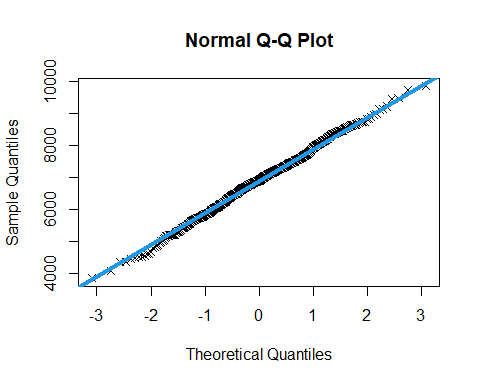
##   
## Spearman's rank correlation rho  
##   
## data: Assignment03\_AP$Wgt1\_AP and Assignment03\_AP$Exercise\_AP  
## S = 24468852, p-value = 0.00004502  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## -0.1815849

Conclusion: There is no linear relationship as our 0.00 ≤ |𝑟| < 0.25.

Q3 Inference

1. Normality
2. Create a QQ Normal plot of White Blood Cell counts.

qqnorm(Assignment03\_AP$WBC\_AP, pch=4, frame=TRUE)  
qqline(Assignment03\_AP$WBC\_AP, col=4, lwd=4)



1. Conduct a statistical test for normality on White Blood Cell counts.

shapiro.test(Assignment03\_AP$WBC\_AP) #Shapiro-Wilk Normality Test

##   
## Shapiro-Wilk normality test  
##   
## data: Assignment03\_AP$WBC\_AP  
## W = 0.99738, p-value = 0.6203

1. Is White Blood Cell count normally distributed? What led you to this conclusion?

* Yes, White Blood Cell Count is normally distributed.
* 1.as p-value = 0.6 which is GREATER THAN 0.05 (P>0.05) that means we cannot reject Null hypothesis which state that variable is normally distributed.
* 2.From QQ Normal Plot and qq line all the points are on and close to line which indicates normality of variable.(from Q3 a)

Statistically Significant Differences

1. Compare White Blood Cell counts between the treatment and control group using a suitable hypothesis test.

# to perform F test to see whether variance are equal or not.  
  
Ftest\_AP <- var.test(WBC\_AP~Group\_AP, data = Assignment03\_AP)  
Ftest\_AP

##   
## F test to compare two variances  
##   
## data: WBC\_AP by Group\_AP  
## F = 1.0601, num df = 248, denom df = 249, p-value = 0.6456  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.8262879 1.3602413  
## sample estimates:  
## ratio of variances   
## 1.060132

# from F-test p > 0.05 that means no significant difference in variances or  
# variance of variables are almost same.  
  
# I have described reasons behind choosing this test below.  
  
Ttest\_AP <- t.test(WBC\_AP~Group\_AP, data = Assignment03\_AP, var.equal=TRUE)  
Ttest\_AP

##   
## Two Sample t-test  
##   
## data: WBC\_AP by Group\_AP  
## t = -2.4461, df = 497, p-value = 0.01479  
## alternative hypothesis: true difference in means between group Control and group Test is not equal to 0  
## 95 percent confidence interval:  
## -400.7539 -43.7367  
## sample estimates:  
## mean in group Control mean in group Test   
## 6776.679 6998.924

1. Explain why you chose the test you did.
   1. Data is Independent
   2. Data is normally distributed (From Shapiro-Wilks Test)
   3. Variance is unknown, but equal (From F-Test)

* A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.
* Reference:Bevans, R. (2022, July 9). An introduction to t-tests. Scribbr. Retrieved October 20, 2022, from #<https://www.scribbr.com/statistics/t-test/>

1. Do you have strong evidence that White Blood Cell counts are different between the treatment and control groups?

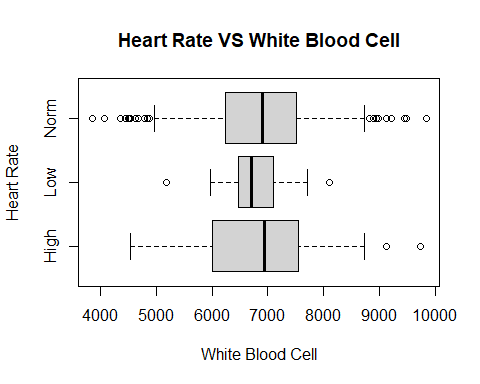
* Yes, From the T-Test, I can say that White Blood Cell Counts are different between the treatment and control groups as p-value is less than 0.05 which provide evidence against Null Hypothesis.

1. Multiple Statistical Differences
2. Determine if White Blood Cell count varies by Heart Rate Level using ANOVA (statistical) and a sequence of boxplots (graphical).

summary(aov(WBC\_AP~HR\_AP, data = Assignment03\_AP))

## Df Sum Sq Mean Sq F value Pr(>F)  
## HR\_AP 2 303459 151729 0.145 0.865  
## Residuals 496 517657299 1043664

boxplot(WBC\_AP~HR\_AP, data = Assignment03\_AP,  
 main = "Heart Rate VS White Blood Cell",  
 ylab = "Heart Rate",  
 xlab = "White Blood Cell", horizontal = TRUE,  
 range = 1)



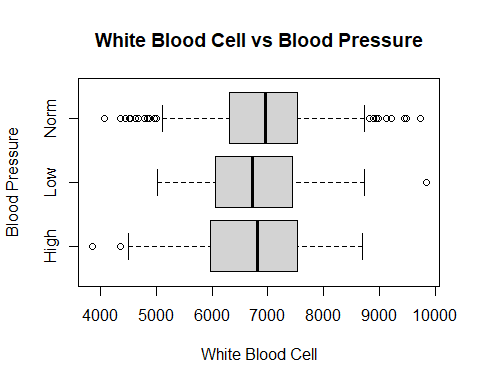
Conclusion: No, White Blood Cell count do not vary by Heart Rate Level as p-value is greater than 0.05 so mean of all the groups is almost same.

1. Determine if White Blood Cell count varies by Blood Pressure Level using ANOVA and a sequence of boxplots.

summary(aov(WBC\_AP~BP\_AP, data = Assignment03\_AP))

## Df Sum Sq Mean Sq F value Pr(>F)  
## BP\_AP 2 2231837 1115918 1.073 0.343  
## Residuals 496 515728921 1039776

boxplot(WBC\_AP~BP\_AP, data = Assignment03\_AP,  
 main = "White Blood Cell vs Blood Pressure",  
 ylab = "Blood Pressure",  
 xlab = "White Blood Cell", horizontal = TRUE,  
 range = 1)



Conclusion: No, White Blood Cell count do not vary by Blood Pressure Level as we can see p-value derived from ANOVA Test is greater than 0.05 which indicates that means of all BP categories are probably same. Moreover, Boxplot also indicates same result.