236523 - Introduction to Bioinformatics - Spring 21 - HW1

Roy Bernea - 316315829 ; Alon Hacohen - 311587653

# Qusetion 1

Following is the code and answers for the questions under the “Question 1” Section - as instructed in the assignment:

### Q1.1

x1<-runif(6, min = 1, max = 20)  
x2<-runif(6, min = 1, max = 20)  
x3<-runif(6, min = 1, max = 20)  
x4<-runif(6, min = 1, max = 20)  
mat1<-rbind(x1,x2,x3,x4)  
print(mat1)

## [,1] [,2] [,3] [,4] [,5] [,6]  
## x1 10.195603 4.198767 14.01098 6.201380 19.702665 4.140229  
## x2 1.405454 11.656965 13.60391 2.164546 11.627533 10.287406  
## x3 2.082238 5.464033 11.53788 11.056389 4.324356 16.819614  
## x4 18.434759 18.436876 11.04295 2.278413 1.209627 4.469715

### Q1.2

sum\_of\_rows<-apply(mat1,1,sum)  
print(sum\_of\_rows)

## x1 x2 x3 x4   
## 58.44963 50.74581 51.28451 55.87234

### Q1.3

tentative.normalization <- function(mat,npar){  
 max\_vector\_row <-apply(mat,1,max)  
 max\_vector\_col <-apply(mat,2,max)  
 if(npar==1){  
 return(mat/max\_vector\_row)  
 }  
 if(npar==2){  
 return(t(t(mat)/max\_vector\_col))  
 }  
 return(mat)  
}

### Q1.4

mat1.norm.rows<-tentative.normalization(mat1,1)  
mat1.norm.columns<-tentative.normalization(mat1,2)  
  
print(mat1.norm.rows)

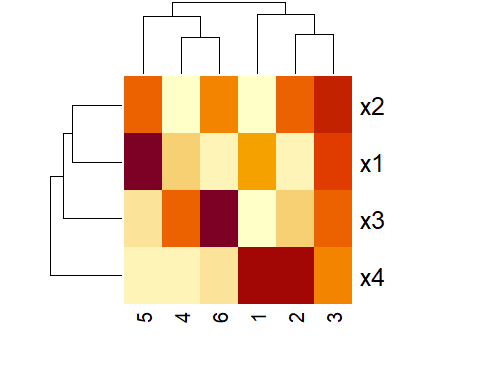
## [,1] [,2] [,3] [,4] [,5] [,6]  
## x1 0.5174733 0.2131065 0.7111213 0.3147483 1.00000000 0.2101355  
## x2 0.1033125 0.8568837 1.0000000 0.1591121 0.85472013 0.7562097  
## x3 0.1237982 0.3248608 0.6859775 0.6573509 0.25710198 1.0000000  
## x4 0.9998852 1.0000000 0.5989599 0.1235791 0.06560912 0.2424334

print(mat1.norm.columns)

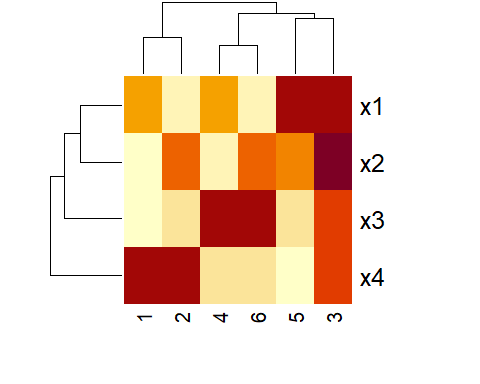
## [,1] [,2] [,3] [,4] [,5] [,6]  
## x1 0.55306409 0.2277374 1.0000000 0.5608866 1.00000000 0.2461548  
## x2 0.07623935 0.6322636 0.9709458 0.1957733 0.59015025 0.6116315  
## x3 0.11295175 0.2963644 0.8234880 1.0000000 0.21948077 1.0000000  
## x4 1.00000000 1.0000000 0.7881637 0.2060720 0.06139409 0.2657442

### Q1.5

heatmap(mat1.norm.rows)



heatmap(mat1.norm.columns)



# Question 2

Following is the code and answers for the questions under the “Question 2” Section - as instructed in the assignment:

### Q2.1

data1.file <- file.path("./data1.csv")  
data2.file <- file.path("./data2.csv")  
data1 <- read.csv(data1.file)  
data2 <- read.csv(data2.file)

### Q2.2

data1\_type <- class(data1)  
print(data1\_type)

## [1] "data.frame"

The type of data1 is **data.frame**. As you may see in our code, the function used for this is the class() function, which allows us to figure out the data structure of an object.

### Q2.3

library(dplyr)  
  
data2\_wo\_NA <- filter(na.omit(data2))  
control <- filter(data2\_wo\_NA, Treatment == "Placebo")  
treatment <- filter(data2\_wo\_NA, Treatment == "HF123")  
print(control)

## Treatment Blood\_Sugar\_Levels  
## 1 Placebo 21.51  
## 2 Placebo 28.14  
## 3 Placebo 23.45  
## 4 Placebo 23.68  
## 5 Placebo 19.79  
## 6 Placebo 28.40  
## 7 Placebo 20.98  
## 8 Placebo 22.51  
## 9 Placebo 20.10  
## 10 Placebo 26.91  
## 11 Placebo 26.25

print(treatment)

## Treatment Blood\_Sugar\_Levels  
## 1 HF123 25.71  
## 2 HF123 26.37  
## 3 HF123 22.80  
## 4 HF123 25.34  
## 5 HF123 24.97  
## 6 HF123 28.14  
## 7 HF123 29.58  
## 8 HF123 30.92  
## 9 HF123 34.02  
## 10 HF123 21.90  
## 11 HF123 31.53  
## 12 HF123 20.73

* We chose to already remove the “NA” rows from the data, as they do not provide any additional information for our uses.

### Q2.4

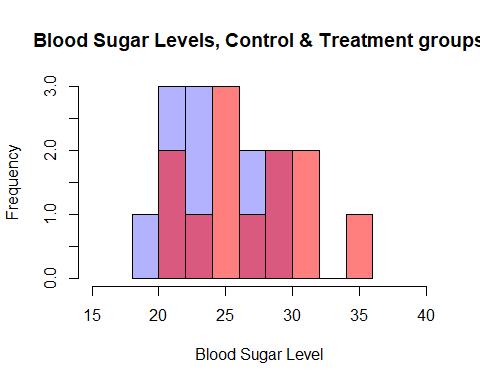
control\_vec\_BSL <- apply(control["Blood\_Sugar\_Levels"],1,as.double)  
treatment\_vec\_BSL <- apply(treatment["Blood\_Sugar\_Levels"],1,as.double)  
mean\_control <- mean(control\_vec\_BSL)  
mean\_treatment <- mean(treatment\_vec\_BSL)  
mean\_difference <- abs(mean\_control-mean\_treatment)  
  
print(mean\_difference)

## [1] 3.041439

As you may notice, the Mean Difference is **3.041439**. This mean difference is **not significant** - this variable is not of any statistical value regarding the context of the question, and does not really help us determine any conclusions about the data.

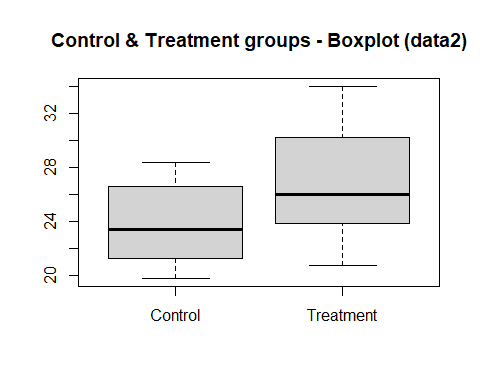
### Q.2.5

hist(control\_vec\_BSL, col=rgb(0,0,1,0.3), main="Blood Sugar Levels, Control & Treatment groups", xlab="Blood Sugar Level", xlim=c(15,40))  
hist(treatment\_vec\_BSL, col=rgb(1,0,0,0.5), add=T)

 \* Please notice that the *blue* represents the **control** group, and the *red* represents the **treatment** group.

### Q.2.6

boxplot(control\_vec\_BSL, treatment\_vec\_BSL, names=c("Control","Treatment"), main="Control & Treatment groups - Boxplot (data2)")



### Q2.7

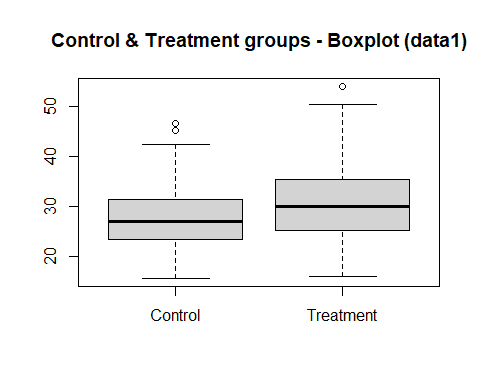
t.test(control\_vec\_BSL, treatment\_vec\_BSL)

##   
## Welch Two Sample t-test  
##   
## data: control\_vec\_BSL and treatment\_vec\_BSL  
## t = -2.0001, df = 20.458, p-value = 0.05894  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -6.2088384 0.1259596  
## sample estimates:  
## mean of x mean of y   
## 23.79273 26.83417

The p-value calculated in the t-test is **0.05894**. This p-value is *larger than 0.05*, which means that we can not contradict the **Null Hypothesis** based on it. In simpler words, it means that we can not state that the difference between the *control* group and the *treatment* group is significant (at least not based on *data2*), therefore we are unsure if the treatment has any real effect on the patients.

### Q2.8

data1\_wo\_NA <- filter(na.omit(data1))  
control1 <- filter(data1\_wo\_NA, Treatment == "Placebo")  
treatment1 <- filter(data1\_wo\_NA, Treatment == "HF123")  
control\_vec\_BSL\_1 <- apply(control1["Blood\_Sugar\_Levels"],1,as.double)  
treatment\_vec\_BSL\_1 <- apply(treatment1["Blood\_Sugar\_Levels"],1,as.double)  
  
boxplot(control\_vec\_BSL\_1, treatment\_vec\_BSL\_1, names=c("Control","Treatment"), main="Control & Treatment groups - Boxplot (data1)")



t.test(control\_vec\_BSL\_1, treatment\_vec\_BSL\_1)

##   
## Welch Two Sample t-test  
##   
## data: control\_vec\_BSL\_1 and treatment\_vec\_BSL\_1  
## t = -7.179, df = 727.79, p-value = 1.737e-12  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.913876 -2.232920  
## sample estimates:  
## mean of x mean of y   
## 27.39611 30.46951

This time, based on data1, the p-value calculated in the t-test is **1.737x10-12)**. This p-value is *smaller than 0.05*, which means that we **can** indeed contradict the **Null Hypothesis** based on it. In simpler words, it means that this time, based on data 1, we **can** state that the difference between the *control* group and the *treatment* group **is significant**, therefore it seems that the treatment does have a real effect on the patients.

### Q2.9

When comparing between the t-tests of data1 vs t-test of data2, we trust the result of the first test more (meaning the **t-test performed on data1**). The reason for that is that data1 includes 846 records, while data2 includes only 24 records. This means we rely on data from a lot more patients when performing the t-test on data1, which allows the distribution to probably be closer to the one in real-life. This also affects the p-value of the t-test, as seen in previous questions. Another way to see that data1 is more diverse than data 2 (which will lead to better results) is looking at the *Gender* column - examining the records of data1 and data2 shows us that the majority of the patients in data2 (over 90%) are females, while the patients in data1 divide into almost even groups of males and females.

# Question 3

Following is the code for generating the GRanges objects - as instructed in the assignment:

### Q3.1

CLIP\_data <- read.table(file.path("./CLIP\_data.BED"),header=TRUE)  
library("GenomicRanges")  
gr <- GRanges(CLIP\_data)  
gr

## GRanges object with 44 ranges and 2 metadata columns:  
## seqnames ranges strand | name score  
## <Rle> <IRanges> <Rle> | <character> <numeric>  
## [1] chr1 567412-567473 + | G10.1 0.653213  
## [2] chr1 10231530-10231560 + | G100.1 -0.124939  
## [3] chr1 161433585-161433627 + | G1000.1 1.237894  
## [4] chr16 69782164-69782223 - | G10009.1 0.778151  
## [5] chr16 69789483-69789507 - | G10010.1 0.698970  
## ... ... ... ... . ... ...  
## [40] chr1 161510012-161510135 + | G1010.1 0.361096  
## [41] chr16 87417585-87417708 - | G10100.1 1.531156  
## [42] chr16 87866215-87866241 - | G10103.1 0.301030  
## [43] chr16 87903029-87903088 - | G10104.1 0.522879  
## [44] chr16 87915229-87915289 - | G10105.1 0.845098  
## -------  
## seqinfo: 2 sequences from an unspecified genome; no seqlengths

### Q3.2

gr\_flank <- flank(gr, width=200, start=FALSE, both=FALSE)  
gr\_flank

## GRanges object with 44 ranges and 2 metadata columns:  
## seqnames ranges strand | name score  
## <Rle> <IRanges> <Rle> | <character> <numeric>  
## [1] chr1 567474-567673 + | G10.1 0.653213  
## [2] chr1 10231561-10231760 + | G100.1 -0.124939  
## [3] chr1 161433628-161433827 + | G1000.1 1.237894  
## [4] chr16 69781964-69782163 - | G10009.1 0.778151  
## [5] chr16 69789283-69789482 - | G10010.1 0.698970  
## ... ... ... ... . ... ...  
## [40] chr1 161510136-161510335 + | G1010.1 0.361096  
## [41] chr16 87417385-87417584 - | G10100.1 1.531156  
## [42] chr16 87866015-87866214 - | G10103.1 0.301030  
## [43] chr16 87902829-87903028 - | G10104.1 0.522879  
## [44] chr16 87915029-87915228 - | G10105.1 0.845098  
## -------  
## seqinfo: 2 sequences from an unspecified genome; no seqlengths