Exercise day 2 with solutions

Introduction to R for Basic Statistics

Exercise B: Reshaping data (Part II)

Consider the data we used for Exercise of day 1.

It is a subset of “follicle” data, collected from patients with cancer that had OTC (ovarian tissue cryopreservation). Follicles were cultured for 8 days and the diameter was collected every 2 days. The aim of the study was to compare the follicles growth over time among different treatment groups.

Question 0 Load data into R :

```
follicle<-read.csv("https://raw.githubusercontent.com/AMeddis/
              IntrotoR-for-Basic-Statistics/main/data_exercise/follicle.csv")

patients<-read.csv("https://raw.githubusercontent.com/AMeddis/
              IntrotoR-for-Basic-Statistics/main/data_exercise/patient.csv")
```

1. Consider Data set *follicle*.
2. Visualize the first lines of the data.
3. Print the summary of the data. Is there any missing values? If yes, how many at Day0 and how many at Day8?

```
head(follicle)
```

```
##   Number Patient     Day0     Day2     Day4     Day6    Day8
## 1      1       1 101.4590 112.1605 135.1980 160.2315 161.500
## 2      2       1  89.8315 141.3770 165.4925       NA      NA
## 3      3       1  90.2835 116.9870 122.6500 127.7305 129.447
## 4      4       1 120.3145 148.8840 166.9970 170.0245 170.740
## 5      5       1  93.0085 112.3135 120.8550 120.9000 120.940
## 6      6       2  83.9085 100.3520 112.6875       NA      NA
```

```
summary(follicle)
```

```
##      Number         Patient          Day0             Day2       
##  Min.   : 1.00   Min.   : 1.0   Min.   : 48.32   Min.   : 59.87  
##  1st Qu.:18.25   1st Qu.: 4.0   1st Qu.: 62.74   1st Qu.: 81.50  
##  Median :35.50   Median : 7.5   Median : 83.91   Median :105.03  
##  Mean   :35.50   Mean   : 7.5   Mean   : 84.70   Mean   :106.85  
##  3rd Qu.:52.75   3rd Qu.:11.0   3rd Qu.: 95.56   3rd Qu.:118.05  
##  Max.   :70.00   Max.   :14.0   Max.   :194.47   Max.   :263.01  
##                                 NA's   :3        NA's   :5       
##       Day4             Day6             Day8       
##  Min.   : 62.97   Min.   : 71.34   Min.   : 72.48  
##  1st Qu.: 96.96   1st Qu.:102.64   1st Qu.:103.78  
##  Median :120.51   Median :126.42   Median :129.45  
##  Mean   :127.01   Mean   :139.46   Mean   :146.24  
##  3rd Qu.:147.24   3rd Qu.:172.66   3rd Qu.:176.93  
##  Max.   :299.65   Max.   :304.94   Max.   :318.88
##  NA's   :11       NA's   :24       NA's   :29
```

4. Use the command *table(dbf$patient)*, interpret the numbers.

```
table(follicle$Patient)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14
##  5  5  5  5  5  5  5  5  5  5  5  5  5  5
```

We have 5 follicles for each patient.

Question 1

1. Calculate mean and standard deviation of the diameter at Day0

```
mean(follicle$Day0, na.rm=TRUE)
```

```
## [1] 84.70384
```

```
sd(follicle$Day0, na.rm=TRUE)
```

```
## [1] 27.6819
```

2. When we encounter into missing, we are often interested in the *complete case analysis* where we exclude patients with missing observations:

   2a. Use the *na.omit* function (excludes all rows that have one missing values) (Run the command *db.CC<-na.omit(NameofDataFrame)**)

   2b. Check the dimension of the new data.frame

   2c. Calculate mean and standard deviation of the diameter at Day0 from db.CC .

   2d. Compare results with the ones in point 1. Did something change? If yes, Why?

```
db.CC<-na.omit(follicle)
dim(db.CC)
```

```
## [1] 41  7
```

```
mean(db.CC$Day0)
```

```
## [1] 91.08129
```

```
sd(db.CC$Day0)
```

```
## [1] 31.47671
```

We can see that mean and standard deviation are different respect to the ones calculated in point 1. This is because the na.omit function is excluding all rows with at least one missing observation. However if one missing was at day 6, this might not be missing at Day0. The mean calculation, with na.rm=TRUE is excluding only the missing at Day0, whereas the complete case consider only follicles that have all observed measurements (for all days)

Question 2 For each follicle the diameter was measured at day 0,2,4,6,8.

1. Are the data in a wide or long format?

Data are in a wide format because we have one row for each follicle and several columns to indicate the measurement at different time points.

2. Convert data from wide to long. **Hint: You can use the function *reshape***

```
db_long<-reshape(db.CC, direction="long",
                 idvar="Number",
                 varying=c("Day0","Day2","Day4","Day6","Day8"),
                 timevar="Day",v.names=c("diameter"),
                 times=c("0","2","4","6","8"))
```

3. How many rows would we expect for each Day? Is it correct? (You can use the command *table(db$Day)*)

```
table(db_long$Day)
```

```
##
##  0  2  4  6  8
## 41 41 41 41 41
```

We have 41 patients with complete data, thus we expect 41 observations for each day

Question 3 Descriptive at baseline (Day0). We would like to create one data set with all characteristics of patients at baseline.

1. Merge the data set obtained in Question 2 and *patient* to add baseline characteristics in the data.frame

```
db_all<-merge(db_long, patient, by="Patient")
```

2. Create a categorical variable for age considering the intervals: (19,30], (30,35], (35,40] (**Hint use the function *cut()***)

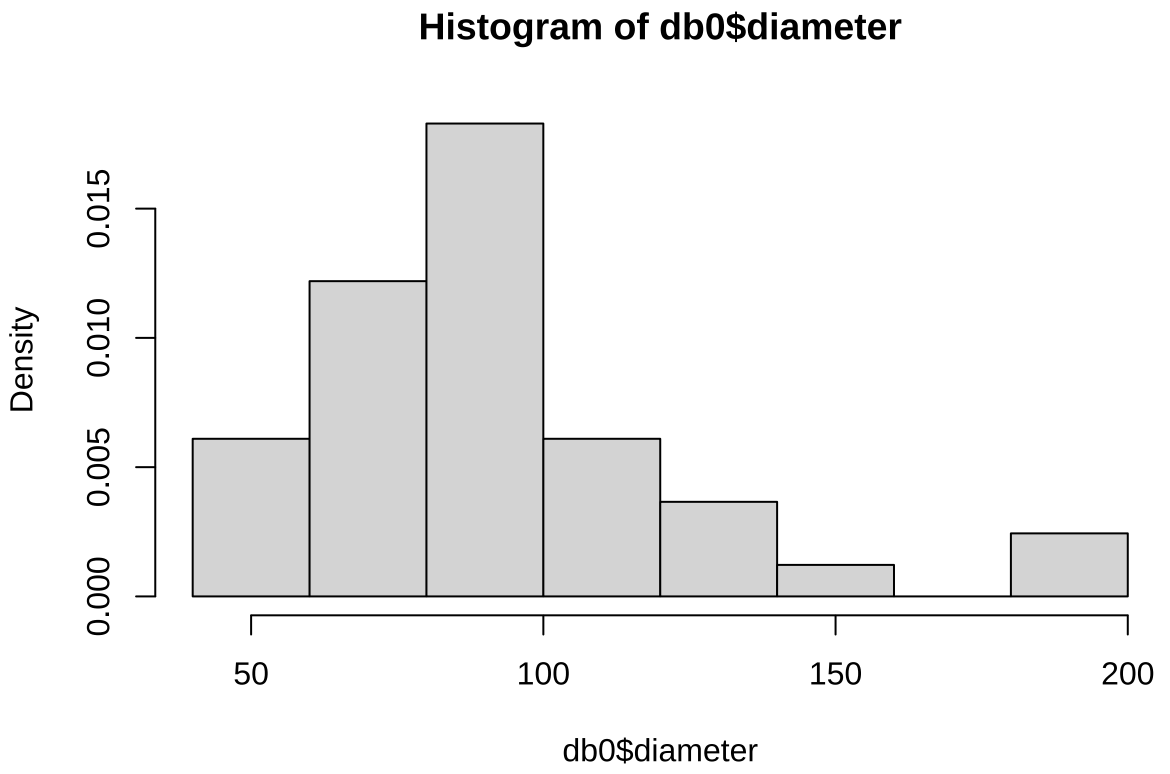
```
db_all$Age_group<-cut(db_all$Age, breaks=c(19,30,35,40))
```

3. Subset from the merged data only observation at baseline (Day 0)

```
db0<-subset(db_all, Day=="0")
```

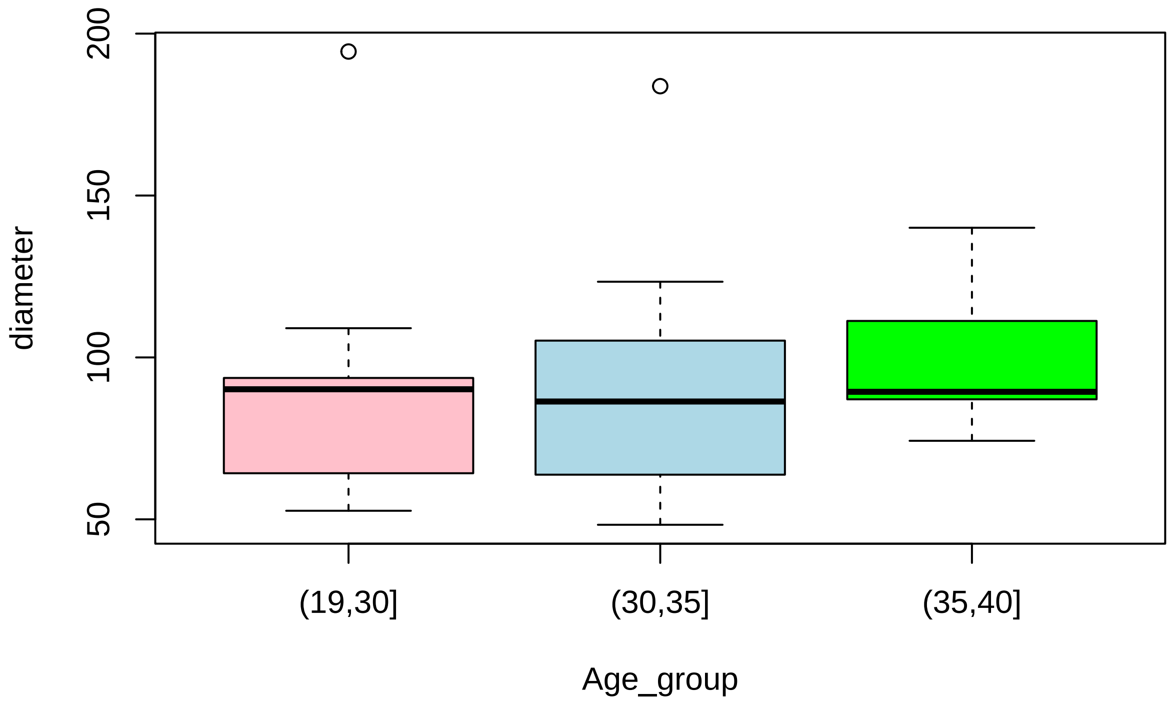
4. Plot the histogram for the density of diameter at Day0.

```
hist(db0$diameter, prob=TRUE)
```



5. Create a Boxplot of diameter at Day 0 by Age category.

```
boxplot(diameter~Age_group, db0, col=c("pink","lightblue","green"))
```



6. Create a table with the counts of follicles per disease and Treatment. Which is the most common disease?

```
table(db0$Disease, db0$Treatment)
```

```
##
##                            FBS hPL HSA
##   Brain_cancer               0   0   1
##   Breast_cancer             26   0   0
##   Chronic_myeloid_leukemia   4   0   0
##   Mb_Hodgkin                 0   5   0
##   Neurological_cancer        3   0   0
##   Sarcoma                    0   0   2
```

Most patients have breast cancer and they have all being treated with FBS

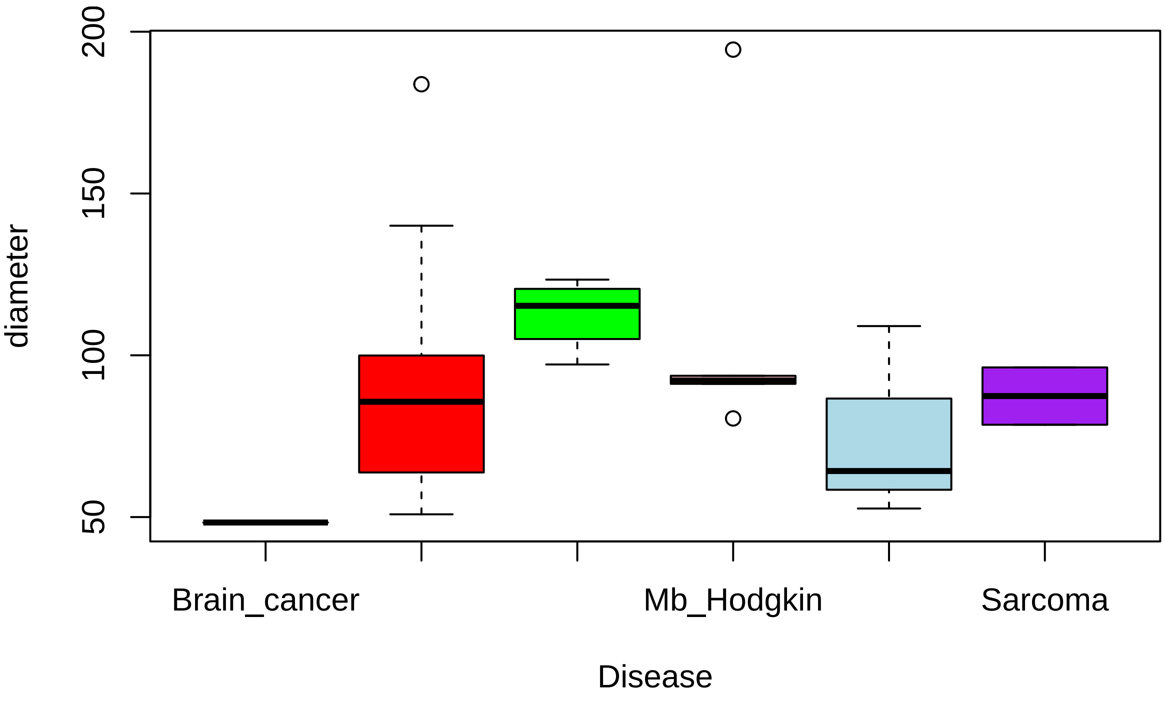
7. Print the mean of diameter at Day0 by Disease (use *tapply()* or *aggregate()*).

   7a. Would you say that the follicle diameter is dependent on the disease? 7b. Show the boxplot of diameter at time 0 by Disease.

```
tapply(db0$diameter,db0$Disease,mean,na.rm=TRUE)
```

```
##             Brain_cancer            Breast_cancer Chronic_myeloid_leukemia
##                 48.32000                 87.77877                112.77750
##               Mb_Hodgkin      Neurological_cancer                  Sarcoma
##                110.39300                 75.30000                 87.39500
```

```
boxplot(diameter~Disease, db0, col=c("gray","red","green","pink","lightblue","purple"))
```



Question 4 : We are interested in the follicle growth over time. We can calculate the diameter difference from time 0 at each time point:

1. Take the subset observations at Day 0. Create a data.frame with only columns *Number* and *diameter*

```
day0<-db0[, c("Number","diameter")]
```

2. Rename the variable of diameter into *diameter0*

```
colnames(day0)<-c("Number","diameter0")
```

3. Merge this data.frame and the long format of your data set (created in Question 3.1) by *Number*.

```
db_join=merge(db_all,day0, by="Number")
```

4. Create a new variable "diam.change" for the difference of diameter at each time point.

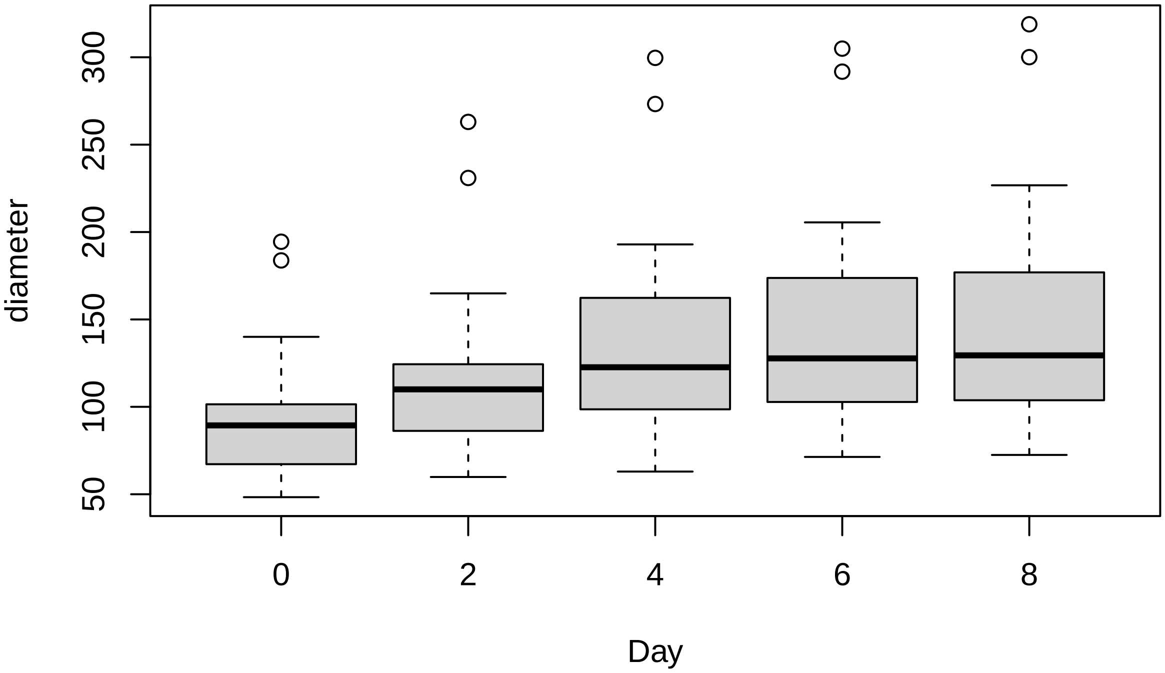
```
db_join$diam.change<-db_join$diameter-db_join$diameter0

head(db_join)
```

```
##   Number Patient Day diameter       Disease Treatment  Age Age_group diameter0
## 1      1       1   0 101.4590 Breast_cancer       FBS 31.6   (30,35]  101.4590
## 2      1       1   4 135.1980 Breast_cancer       FBS 31.6   (30,35]  101.4590
## 3      1       1   2 112.1605 Breast_cancer       FBS 31.6   (30,35]  101.4590
## 4      1       1   6 160.2315 Breast_cancer       FBS 31.6   (30,35]  101.4590
## 5      1       1   8 161.5000 Breast_cancer       FBS 31.6   (30,35]  101.4590
## 6      3       1   4 122.6500 Breast_cancer       FBS 31.6   (30,35]   90.2835
##   diam.change
## 1      0.0000
## 2     33.7390
## 3     10.7015
## 4     58.7725
## 5     60.0410
## 6     32.3665
```

Question 5: Descriptive of diameter change over time

1. Show the Boxplot of diameter by Day. Would you say that the diameter is growing over time?

```
boxplot(diameter~Day , db_join)
```



2. Calculate the median diameter change by Day and Treatment (save the results, you need them for the next step). **use aggregate with formula: diam.change~Day + Treatment**

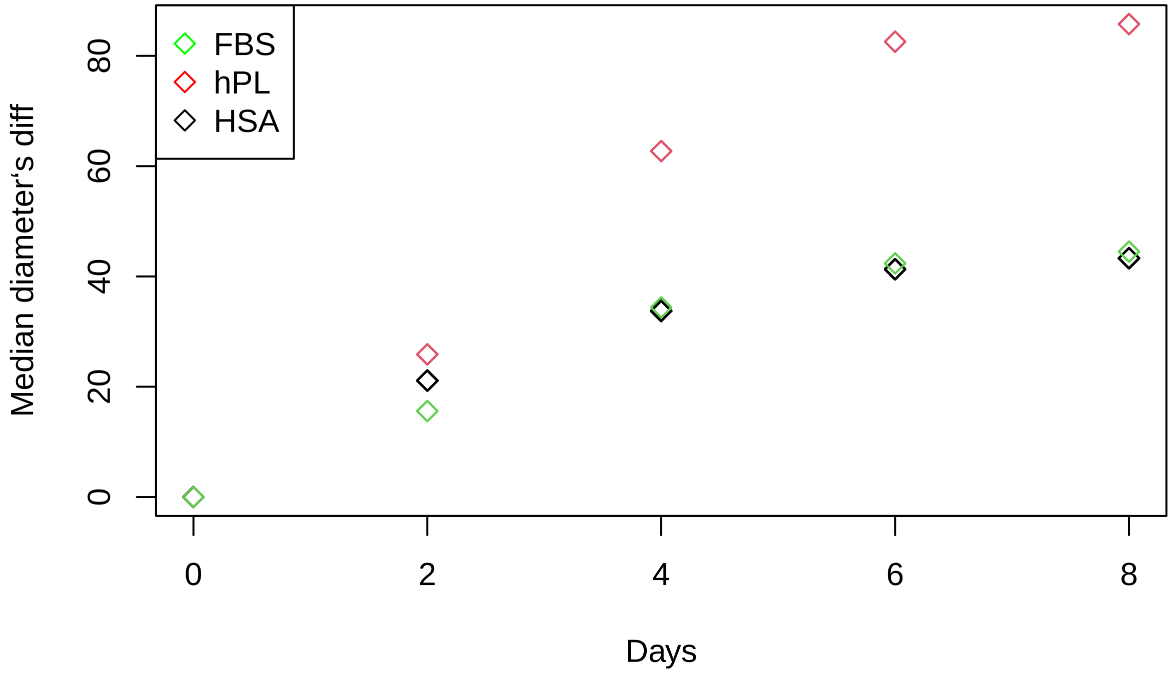
```
db_median<-aggregate(diam.change~ Day + Treatment, db_join, median)
```

3. Plot the diameter growth over time by treatment group:

- 2a. rename the column “diam.change” from the data.frame obtained in the previous point with “median.change”
- 2b. merge the data set with the one obtained in Question 4 (after creating diam.change)
- 2c. create a plot (with points) of the difference from time 0 at varying of days:

```
                    - define the color by Treatment group

                    - precise as name of axis: x= "Days", y="median diameter's diff"

                    - add the legend
```

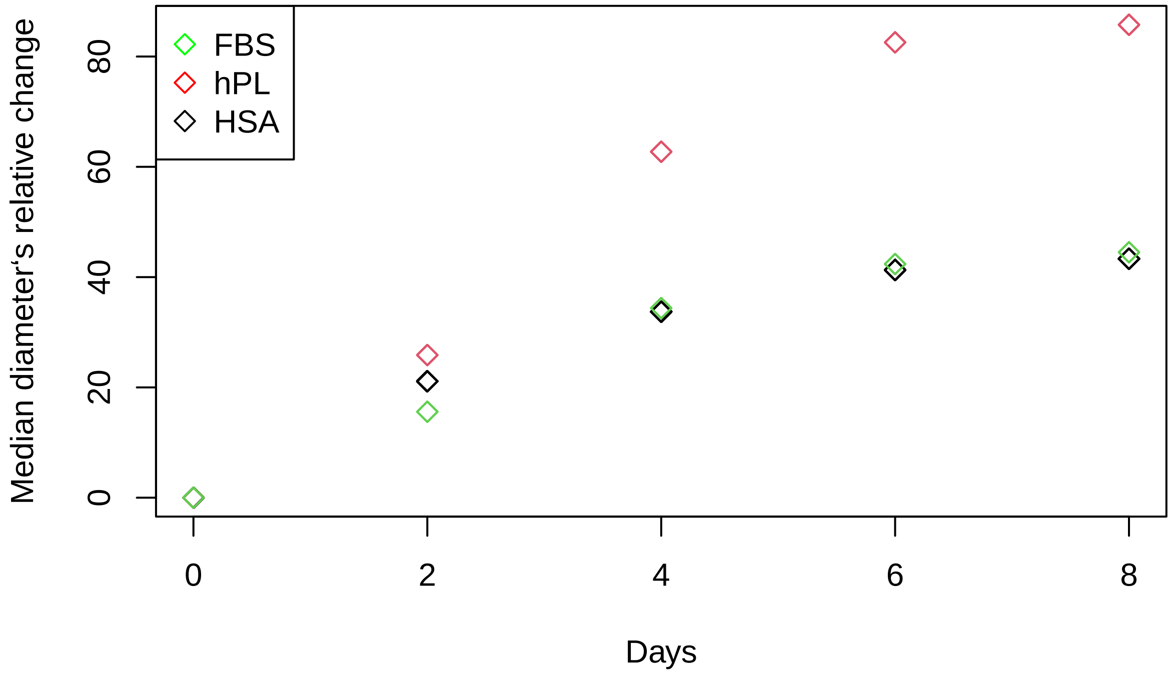
```
colnames(db_median)[3]<-"median.change"
db_join2<-merge(db_join,db_median, by=c("Day","Treatment"))


plot(db_join2$Day, db_join2$median.change,
     col=as.factor(db_join2$Treatment), pch=5,
     xlab="Days", ylab="Median diameter`s diff")
legend("topleft", c("FBS","hPL","HSA"), col=c("green","red","black"), pch=c(5,5,5))
```



4. (Optional) Calculate the relative change for the diameter((diameter - diameter0)/diameter0) and re-create the same plot:

```
db_join$diam.Rchange<-(db_join$diameter-db_join$diameter0)/(db_join$diameter0)
db_medianR<-aggregate(diam.Rchange~ Day + Treatment, db_join, median)

colnames(db_median)[3]<-"median.Rchange"
db_join2<-merge(db_join,db_median, by=c("Day","Treatment"))
```

```r
plot(db_join2$Day, db_join2$median.Rchange,
     col=as.factor(db_join2$Treatment), pch=5,
     xlab="Days", ylab="Median diameter`s relative change")
legend("topleft", c("FBS","hPL","HSA"), col=c("green","red","black"), pch=c(5,5,5))
```


◇ FBS
◇ hPL
◇ HSA
Median diameter's relative change
Days