Liyang Chen

2021/5/22

https://github.com/clycly-ai/A3.git

Introduction

Part 1 is to analyze Gene data

```
knitr::opts_chunk$set(echo = TRUE)

## Step 1: Download Gene File and Table making
download.file(url ="https://raw.githubusercontent.com/markziemann/SLE71
2_files/master/assessment_task3/bioinfo_asst3_part1_files/gene_expressi
on.tsv", destfile ="gene_expression.tsv")

## Showing the first six values as an example
read.table(file = 'gene_expression.tsv', sep = '\t', header = TRUE, nrow
s = 6)
```

```
##            GeneID SRR5150592 SRR5150593
## 1 ENSG00000223972          1          0
## 2 ENSG00000227232          0          1
## 3 ENSG00000278267          0          0
## 4 ENSG00000243485          0          0
## 5 ENSG00000284332          0          0
## 6 ENSG00000237613          0          0
```

```
data <- read.table(file = 'gene_expression.tsv', sep = '\t', header = TR
UE, nrows = 6)
data
```

```
##            GeneID SRR5150592 SRR5150593
## 1 ENSG00000223972          1          0
## 2 ENSG00000227232          0          1
## 3 ENSG00000278267          0          0
## 4 ENSG00000243485          0          0
## 5 ENSG00000284332          0          0
## 6 ENSG00000237613          0          0
```

```
## Adding a new column which is the mean of the other columns
### Mean
data_2 <- read.table(file = 'gene_expression.tsv', sep = '\t', header =
TRUE)
x <- apply(data_2[,2:3], 1, mean)

### New Data Table with a new column of Mean
data_3 <- data_2
vec <- c(x)
data_3$Mean <- vec

### New Table to show first six values with Mean
```

```
data_4 <- data_3
head(data_4,6)
```

```
##           GeneID SRR5150592 SRR5150593 Mean
## 1 ENSG00000223972          1          0  0.5
## 2 ENSG00000227232          0          1  0.5
## 3 ENSG00000278267          0          0  0.0
## 4 ENSG00000243485          0          0  0.0
## 5 ENSG00000284332          0          0  0.0
## 6 ENSG00000237613          0          0  0.0
```

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

```
##List the 10 genes with the highest mean expression
### data_3 includes full data with Mean
data_5 <- data_3[with(data_3,order(-Mean)),]
head(data_5,10)
```

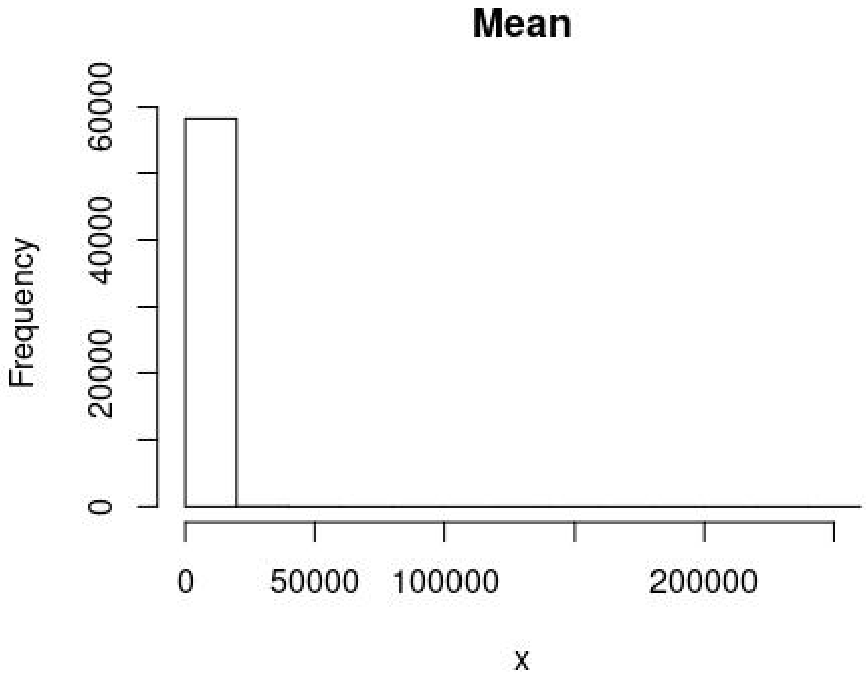
```
##                GeneID SRR5150592 SRR5150593     Mean
## 8683  ENSG00000115414     311857     206347 259102.0
## 58210 ENSG00000210082     145916     163288 154602.0
## 20619 ENSG00000075624     133983     116762 125372.5
## 58234 ENSG00000198886      91596      99943  95769.5
## 42896 ENSG00000137801      95158      74546  84852.0
## 58222 ENSG00000198804      79832      84774  82303.0
## 58238 ENSG00000198786      74570      83589  79079.5
## 25675 ENSG00000196924      88225      66413  77319.0
## 58225 ENSG00000198712      76108      77108  76608.0
## 49030 ENSG00000108821      80342      60127  70234.5
```

```
### this provides the highest 10 Mean values

## Determine the number of genes with a mean <10
### x involves a group mean values
count <- length(which(x < 10))
count
```

```
## [1] 43124
```

```
### 43124 genes that their mean are < 10

## Histogram plot of the mean values
```

```
### x involves a group mean values
hist(x, breaks=10, main="Mean")
```

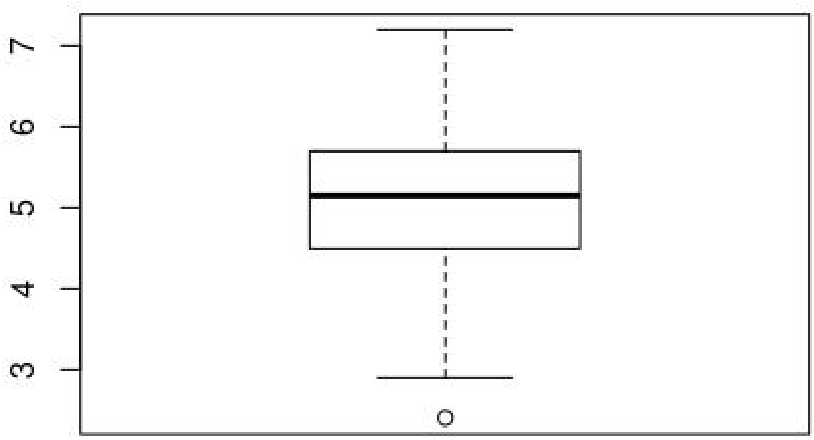


Introduction

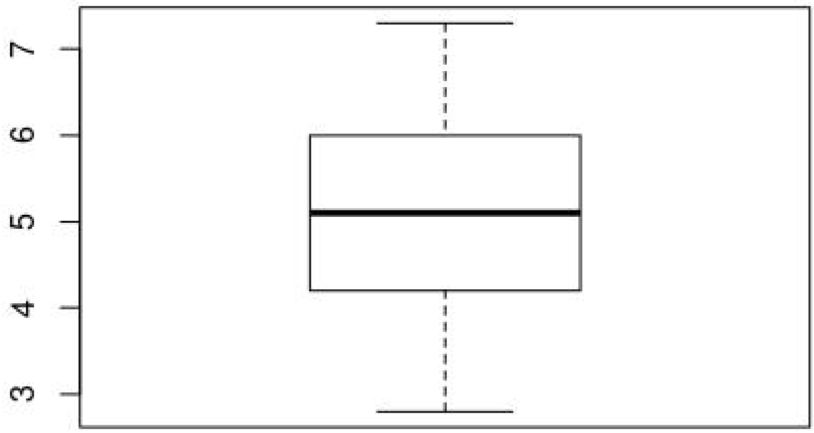
Part 1 Task 2 is to analyze Plant data

```
knitr::opts_chunk$set(echo = TRUE)

## Import csv file and Calculation
### Ensure column names
download.file(url ="https://raw.githubusercontent.com/markziemann/SLE71
2_files/master/assessment_task3/bioinfo_asst3_part1_files/growth_data.c
sv", destfile ="growth_data.csv")

data <- read.table(file = 'growth_data.csv', sep = '\t', header = TRUE,
nrows = 1)
data

##   Site.TreeID.Circumf_2004_cm.Circumf_2009_cm.Circumf_2014_cm.Circum
f_2019_cm
## 1                                       northeast,A003,5.2,10.1,19.9,
38.9

### The bold font at the top of the table is the column name of this gro
up of data

### Mean and Standard Deviation Calculation
### Mean of tree circumference in the northeast: the start(x) and the en
d(y)
### The calculation is conducted in the most basic and least efficient m
ethod
data.frame <- read.table(file = 'growth_data.csv', sep = '\t', header =
TRUE)
c1 <- c(5.2, 4.9, 3.7, 3.8, 3.8, 5.9, 4.4, 5.3, 7.1, 3.8, 5.4, 3.5, 2.4,
 5.9, 6.5, 2.9, 5.0, 7.2, 5.0, 6.5, 5.4, 6.6, 4.7, 6.3, 5.1, 4.8, 3.7, 5.
1, 5.3, 5.4, 5.8, 4.3, 4.5, 6.2, 5.0, 5.5, 4.8, 3.2, 4.8, 5.7, 4.7, 5.3,
 5.5, 5.6, 5.4, 5.7, 5.4, 6.9, 3.9, 5.1)
x <- mean(c1)
x

## [1] 5.078

c2 <- c(38.9, 37.0, 28.1, 18.5, 18.4, 28.4, 50.0, 25.8, 34.2, 28.4, 40.5,
 26.0, 11.8, 67.6, 31.4, 21.8, 24.2, 82.5, 56.5, 49.1, 61.8, 75.7, 53.6,
 30.5, 38.4, 54.6, 41.7, 58.0, 59.9, 61.7, 28.0, 49.0, 51.8, 46.7, 24.1,
 41.5, 23.1, 24.2, 36.0, 43.2, 35.5, 25.5, 26.5, 27.2, 40.9, 27.4, 61.2,
 78.2, 19.0, 38.6)
y <- mean(c2)
y

## [1] 40.052
```

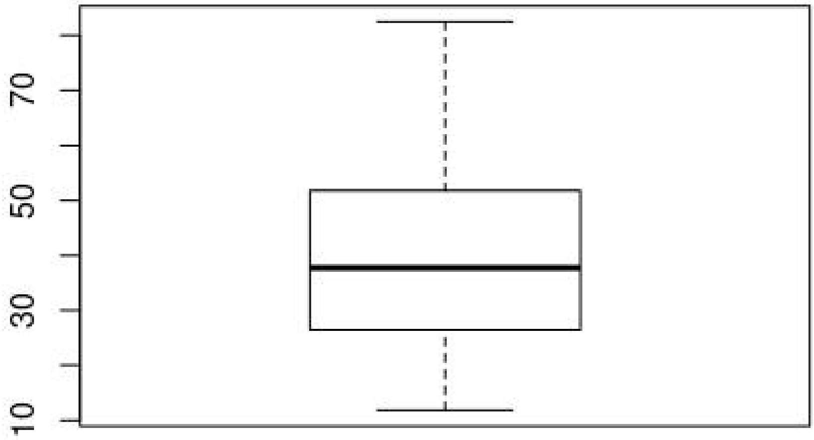
```
### Mean of tree circumference in the southwest: the start(x1) and the e
nd(y1)
c3 <- c(5.3, 5.2, 6.2, 5.1, 3.6, 6.6, 6.6, 5.1, 4.1, 4.4, 3.9, 5.4, 4.7,
 4.2, 3.9, 5.0, 3.9, 6.1, 3.8, 7.3, 5.1, 5.3, 6.4, 2.8, 6.0, 4.1, 6.4, 3.
7, 5.5, 4.7, 4.2, 4.7, 5.4, 3.9, 3.3, 6.1, 4.7, 5.0, 6.0, 5.6, 5.7, 6.2,
 4.5, 6.4, 6.1, 5.0, 5.0, 7.0, 3.2, 5.4)
x1 <- mean(c3)
x1

## [1] 5.076

c4 <- c(88.7, 38.8, 103.9, 58.3, 59.8, 75.5, 110.6, 84.9, 69.3, 33.3, 44.
8, 40.6, 79.3, 47.6, 29.2, 84.2, 64.9, 69.4, 28.5, 54.9, 57.7, 60.8, 72.
6, 21.3, 100.6, 47.0, 73.4, 41.8, 92.4, 53.5, 31.8, 35.7, 40.8, 66.2, 37.
4, 46.2, 78.0, 83.8, 45.3, 94.6, 64.6, 46.8, 33.7, 48.0, 102.0, 38.0, 57.
2, 52.8, 36.8, 61.3)
y1 <- mean(c4)
y1

## [1] 59.772

### Standard Deviation of tree circumference in the northeast: the start
(x2) and the end(y2)
x2 <- sd(c1)
x2

## [1] 1.059127

y2 <- sd(c2)
y2

## [1] 16.90443

### Standard Deviation of tree circumference in the southwest: the start
(x3) and the end(y3)
x3 <- sd(c3)
x3

## [1] 1.060527

y3 <- sd(c4)
y3

## [1] 22.57784

summary(cars)

##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00

## Box plot of tree circumference at the start and end of the study

### Box plot of the start
### Box plot of the NORTHWEST, c1 is data of the start in the northeast
boxplot(c1)
```



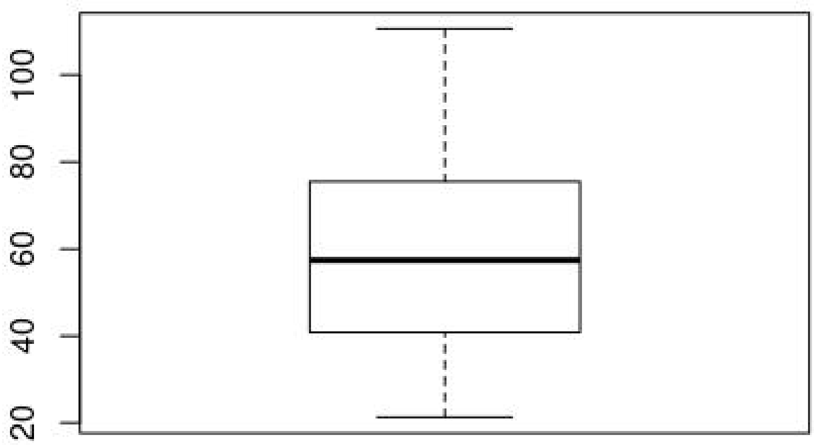
```
### Box plot of the SOUTHWEST, c3 is data of the start in the southwest
boxplot(c3)
```



```
### Box plot of the end
### Box plot of the NORTHWEST, c2 is data of the end in the northeast
boxplot(c2)
```



```
### Box plot of the SOUTHWEST, c4 is data of the end in the southwest
boxplot(c4)
```


20
40
60
80
100