lab 07

Rezaur Rashid

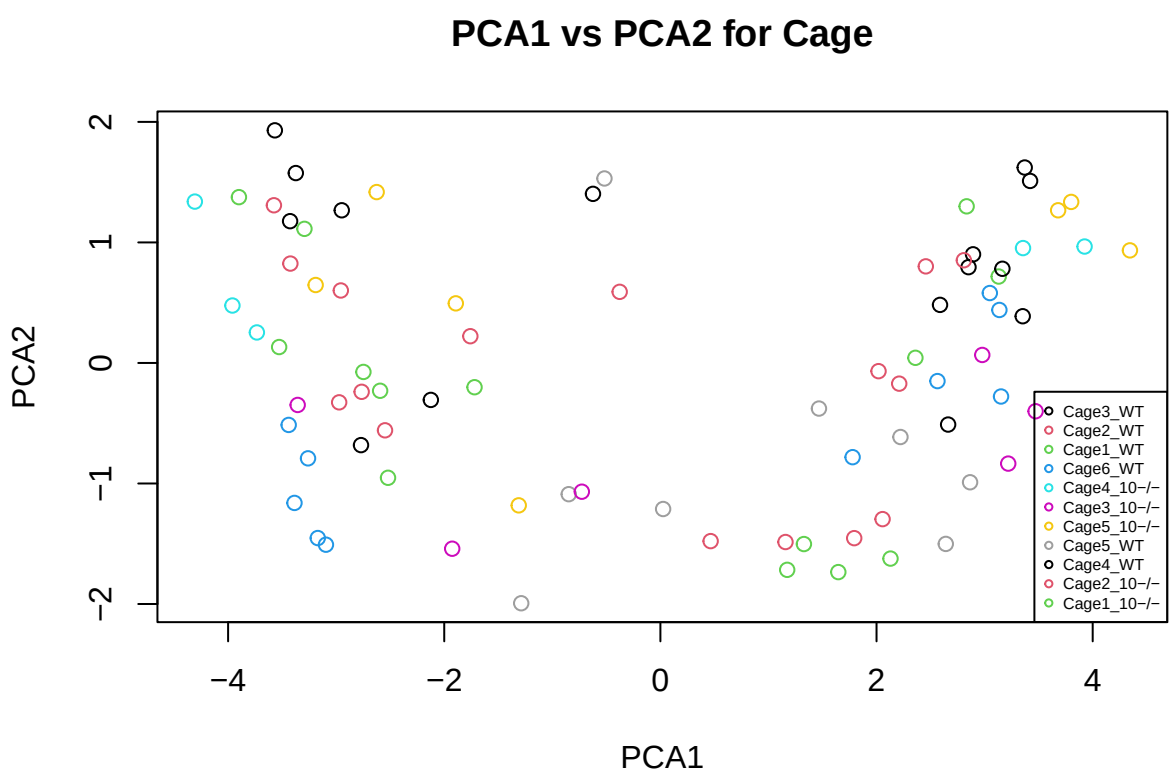
2022-04-28

Part 1: Reading Data

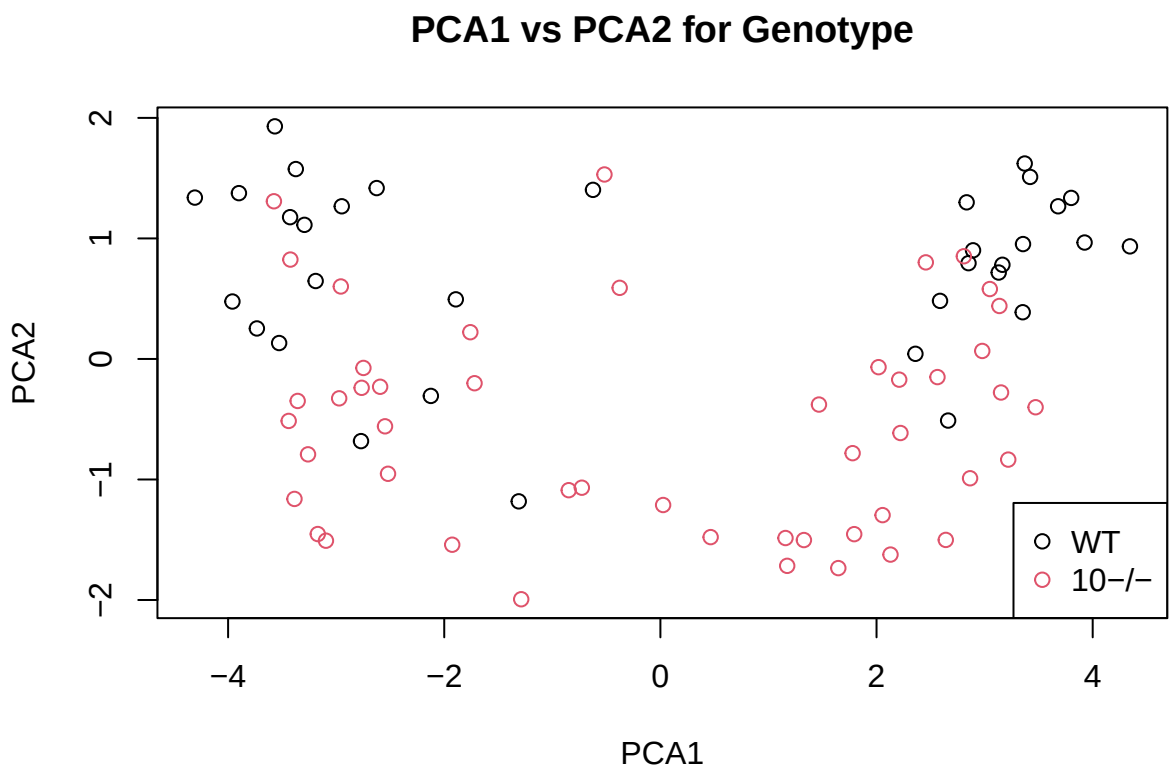
```
## ---------------------- part 1 -------------------------------------

inFileName <- paste("data/lab07data/prePostPhylum.txt", sep ="")

myT <-read.table(inFileName,header=TRUE,sep="\t")
numCols <- ncol(myT)

myColClasses <- c(rep("character",4), rep("numeric", numCols-4))
myT <-read.table(inFileName,header=TRUE,sep="\t",colClasses=myColClasses)
myTData = myT[,5:10]

myPCOA <- princomp(myTData)
PCA1 = myPCOA$scores[,1]
PCA2 = myPCOA$scores[,2]
```

Part 2: PCA analysis and Plotting

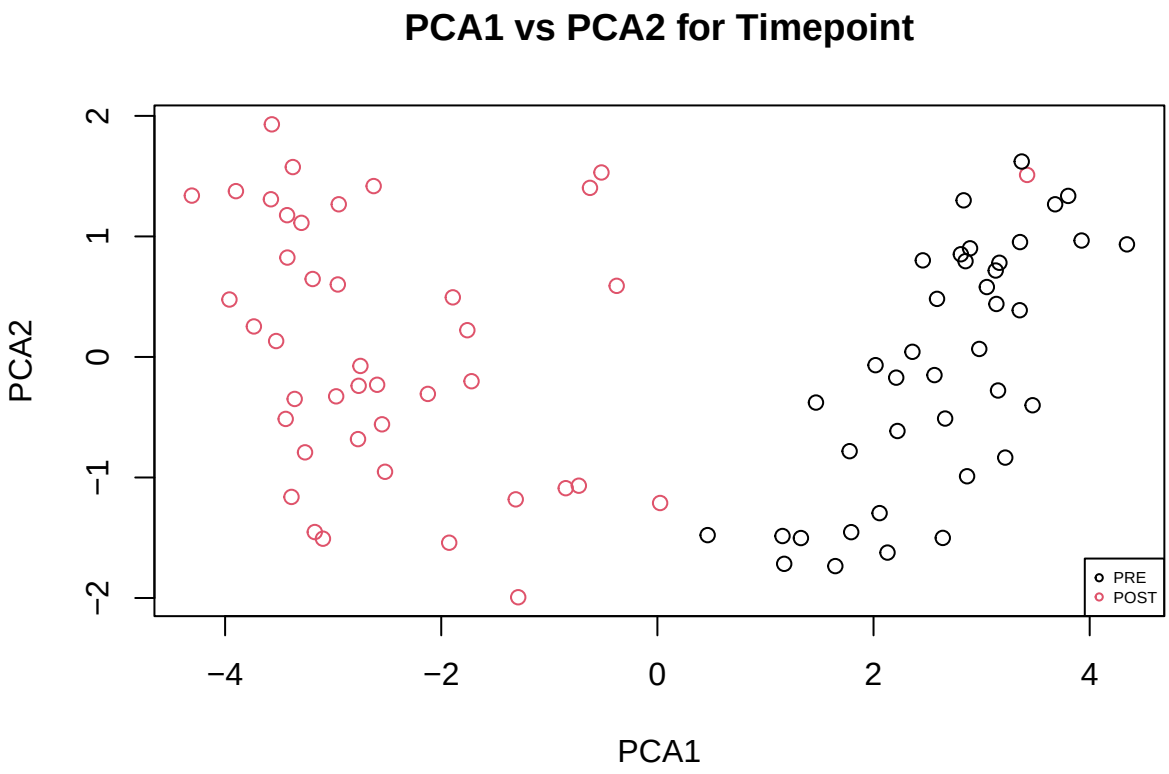
```
## ---------------------- part 2 -------------------------------------

pca_data = data.frame(PCA1, PCA2, myT$cage, myT$genotype, myT$time, stringsAsFactors = TRUE)

## PCA1 vs PCA2 for Cage-------
plot(pca_data$PCA1, pca_data$PCA2, col = pca_data$myT.cage,
     main = 'PCA1 vs PCA2 for Cage', xlab = 'PCA1', ylab = 'PCA2')
legend("bottomright", legend = unique(myT$cage), col = 1:length(unique(myT$cage)),
       inset=c(-0,0), pch=1, cex = 0.5)
```




```
## PCA1 vs PCA2 for Genotype-------
plot(pca_data$PCA1, pca_data$PCA2, col = pca_data$myT.genotype,
     main = 'PCA1 vs PCA2 for Genotype', xlab = 'PCA1', ylab = 'PCA2')
legend("bottomright", legend = unique(myT$genotype), col = 1:length(unique(myT$genotype)),
       inset=c(-0,0), pch=1, cex = 1)
```




```
## PCA1 vs PCA2 for Time-------
plot(pca_data$PCA1, pca_data$PCA2, col = pca_data$myT.time,
     main = 'PCA1 vs PCA2 for Timepoint', xlab = 'PCA1', ylab = 'PCA2')
legend("bottomright", legend = unique(myT$time), col = 1:length(unique(myT$time)),
       inset=c(-0,0), pch=1, cex = 0.5)
```



Part 3:

```
## ---------------------- part 3 -----------------------------------
#cage
cage = factor(pca_data$myT.cage)

myLm = lm(pca_data$PCA1 ~ cage, x=TRUE)
pv_cage_pca1 = anova(myLm)$"Pr(>F)"[1]

myLm2 = lm(pca_data$PCA2 ~ cage, x=TRUE)
pv_cage_pca2 = anova(myLm2)$"Pr(>F)"[1]
pv_cage_pca2
```

```
## [1] 1.629589e-07
```

```
#genotypes
sample1_pca1 = pca_data[pca_data$myT.genotype == 'WT', 'PCA1']
sample2_pca1 = pca_data[pca_data$myT.genotype == '10-/-', 'PCA1']
pv_geno_pca1 = t.test(sample1_pca1, sample2_pca1, var.equal = FALSE)$p.value

sample1_pca2 = pca_data[pca_data$myT.genotype == 'WT', 'PCA2']
sample2_pca2 = pca_data[pca_data$myT.genotype == '10-/-', 'PCA2']
pv_geno_pca2 = t.test(sample1_pca2, sample2_pca2, var.equal = FALSE)$p.value

#timepoints
sample1_pca1 = pca_data[pca_data$myT.time == 'PRE', 'PCA1']
```

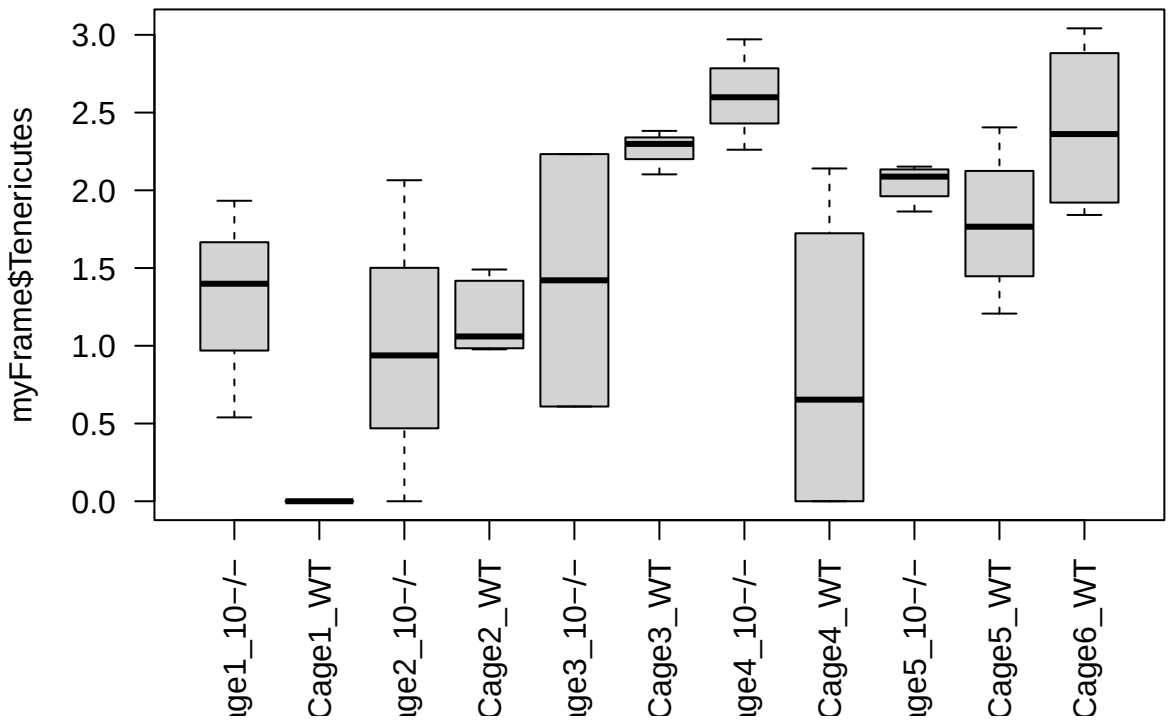
```
sample2_pca1 = pca_data[pca_data$myT.time == 'POST', 'PCA1']
pv_time_pca1 = t.test(sample1_pca1, sample2_pca1, var.equal = FALSE)$p.value

sample1_pca2 = pca_data[pca_data$myT.time == 'PRE', 'PCA2']
sample2_pca2 = pca_data[pca_data$myT.time == 'POST', 'PCA2']
pv_time_pca2 = t.test(sample1_pca2, sample2_pca2, var.equal = FALSE)$p.value

question_3_table = data.frame(
  Category = c('cage', 'genotypes', 'time'),
  PCA1_pValues = c(round(pv_cage_pca1,3), round(pv_geno_pca1,3), round(pv_time_pca1,3)),
  PCA2_pValues = c(round(pv_cage_pca2,3), round(pv_geno_pca2,3), round(pv_time_pca2,3)),
  stringsAsFactors = FALSE)

print(question_3_table)
```

```
##    Category PCA1_pValues PCA2_pValues
## 1      cage        0.992        0.000
## 2 genotypes        0.930        0.000
## 3      time        0.000        0.427
```

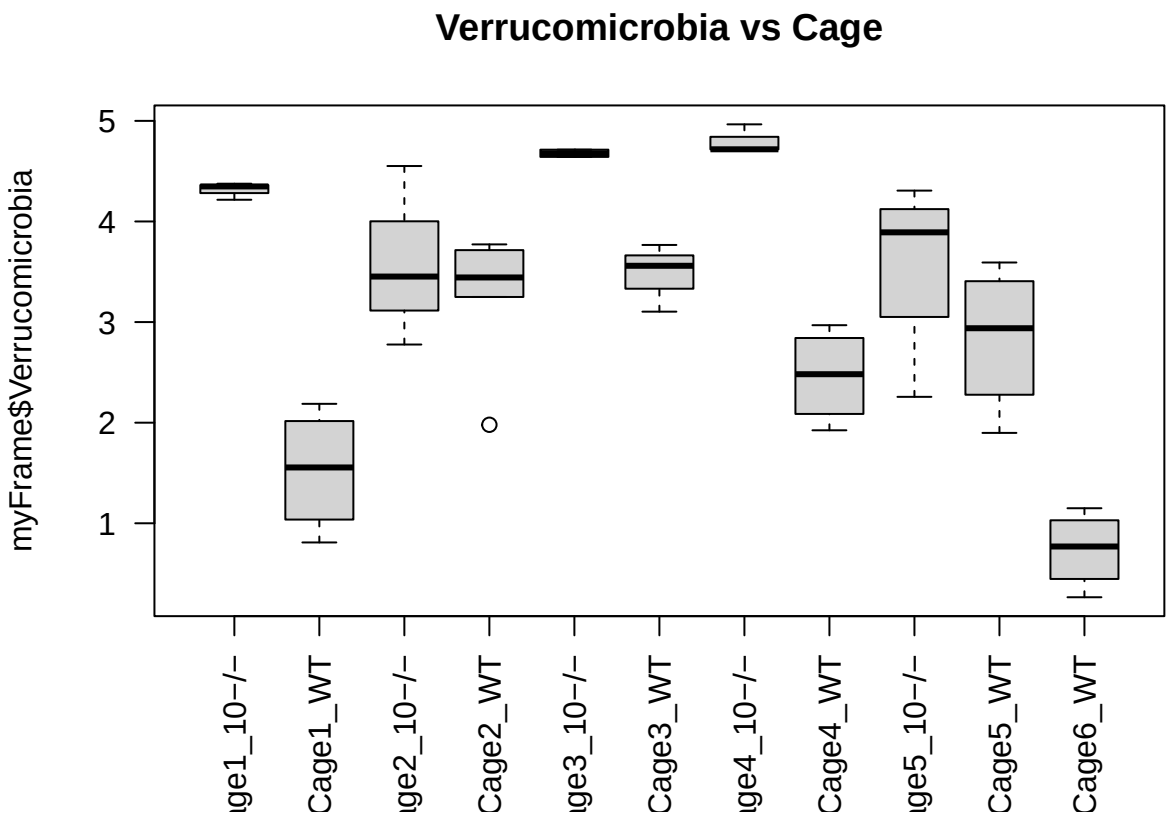
The 'Timepoints' seems to be most associated with the PCA1 axis. We can see it has a small p-values which is less than the significance level. Whereas, the 'Cage' and 'Genotypes' seem to be most associated with the PCA2 axis. Also, we can see that the cage as some effect on the data since they are random for different phyla which can be seen from their p-values from different PCA components.

Part 4:

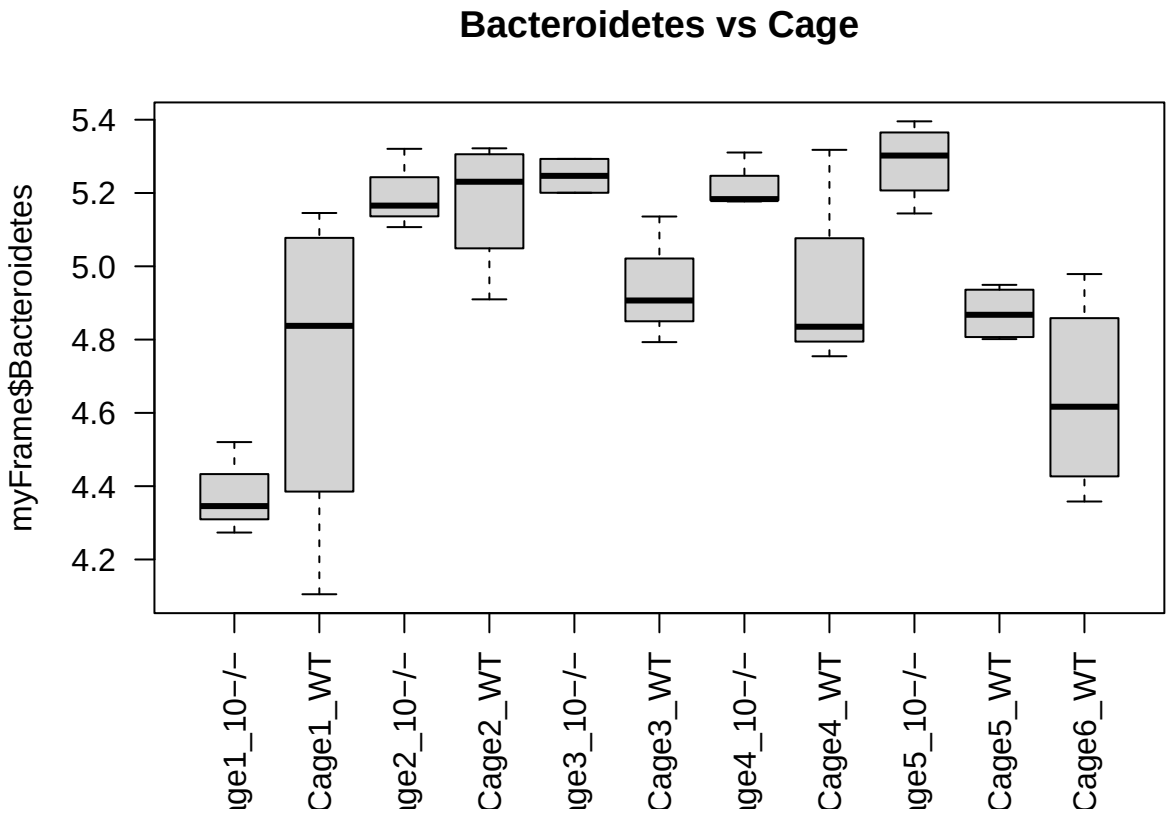
```
## ---------------------- part 4 -----------------------------------
## part 4(a)

post_myT = myT[myT$time == 'POST',]

bug = post_myT[,5:10]
cage = post_myT$cage
genotype = post_myT$genotype
myFrame <- data.frame(bug, cage, genotype, stringsAsFactors = TRUE)

par(mfrow=c(1,1))
boxplot(myFrame$Tenericutes ~ myFrame$cage, main = 'Tenericutes vs Cage', las=2, xlab='')
```



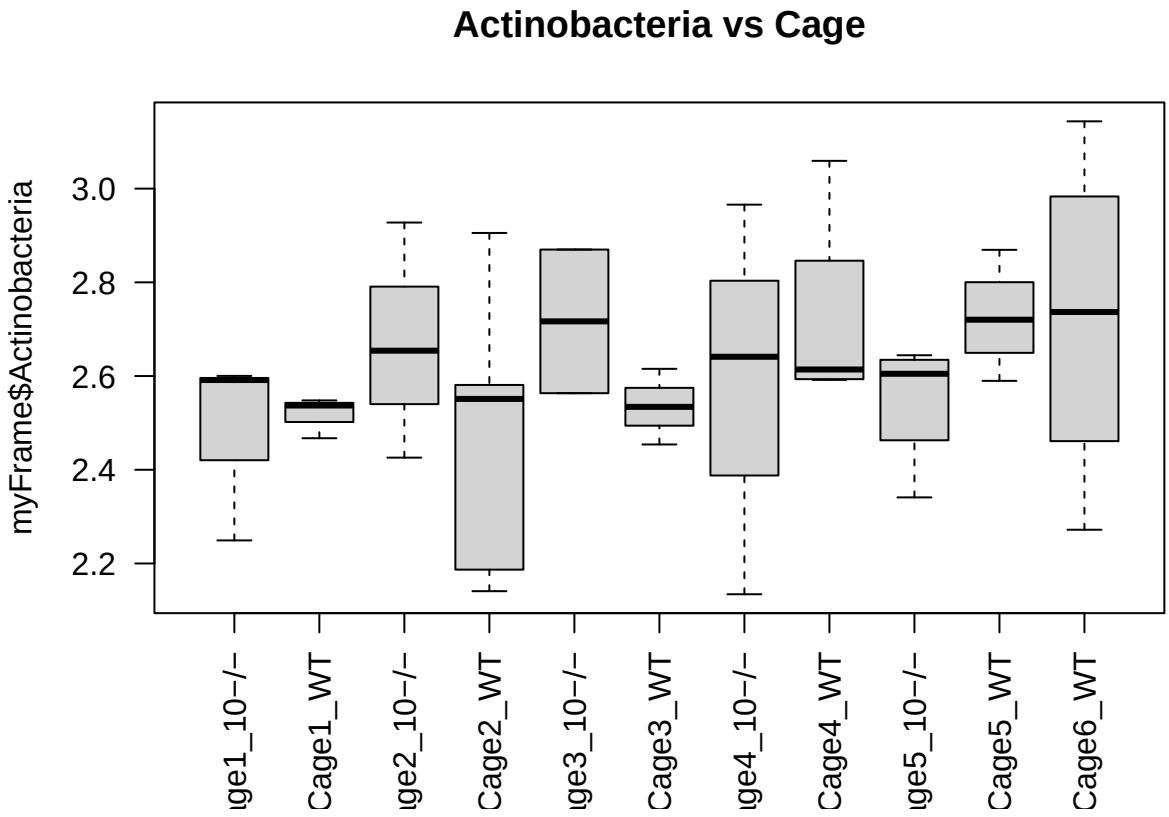

```
boxplot(myFrame$Verrucomicrobia ~ myFrame$cage, main = 'Verrucomicrobia vs Cage', las=2, xlab='')
```



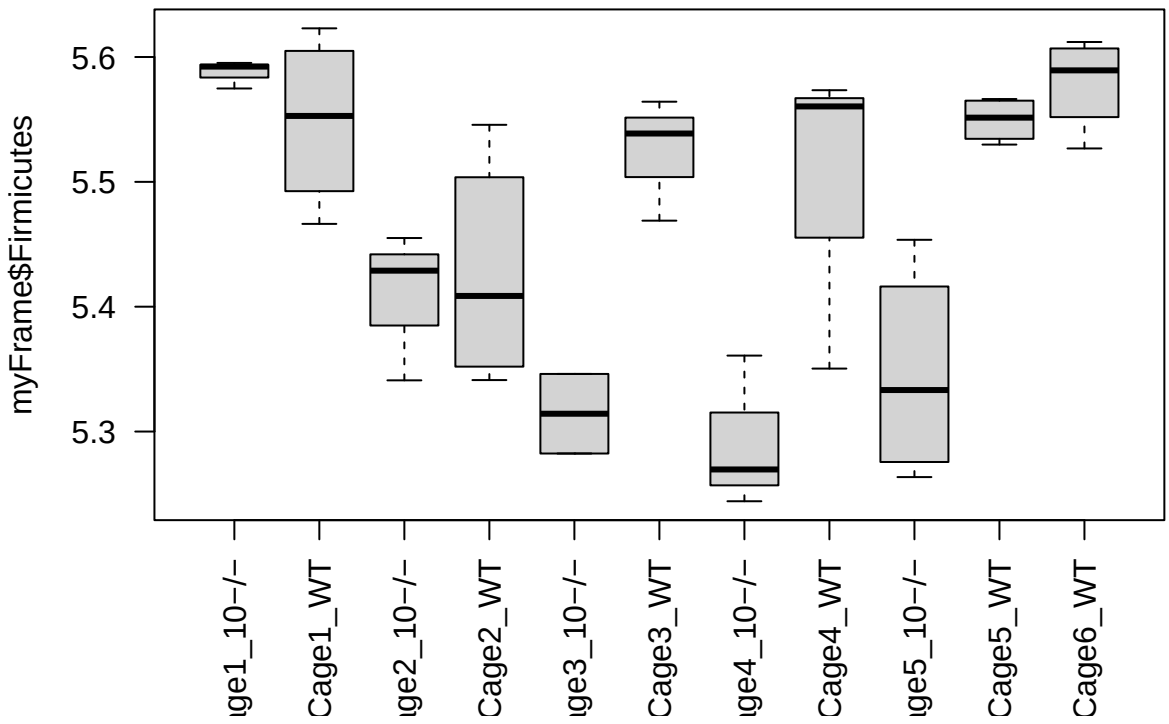

```
boxplot(myFrame$Bacteroidetes ~ myFrame$cage, main = 'Bacteroidetes vs Cage', las=2, xlab='')
```



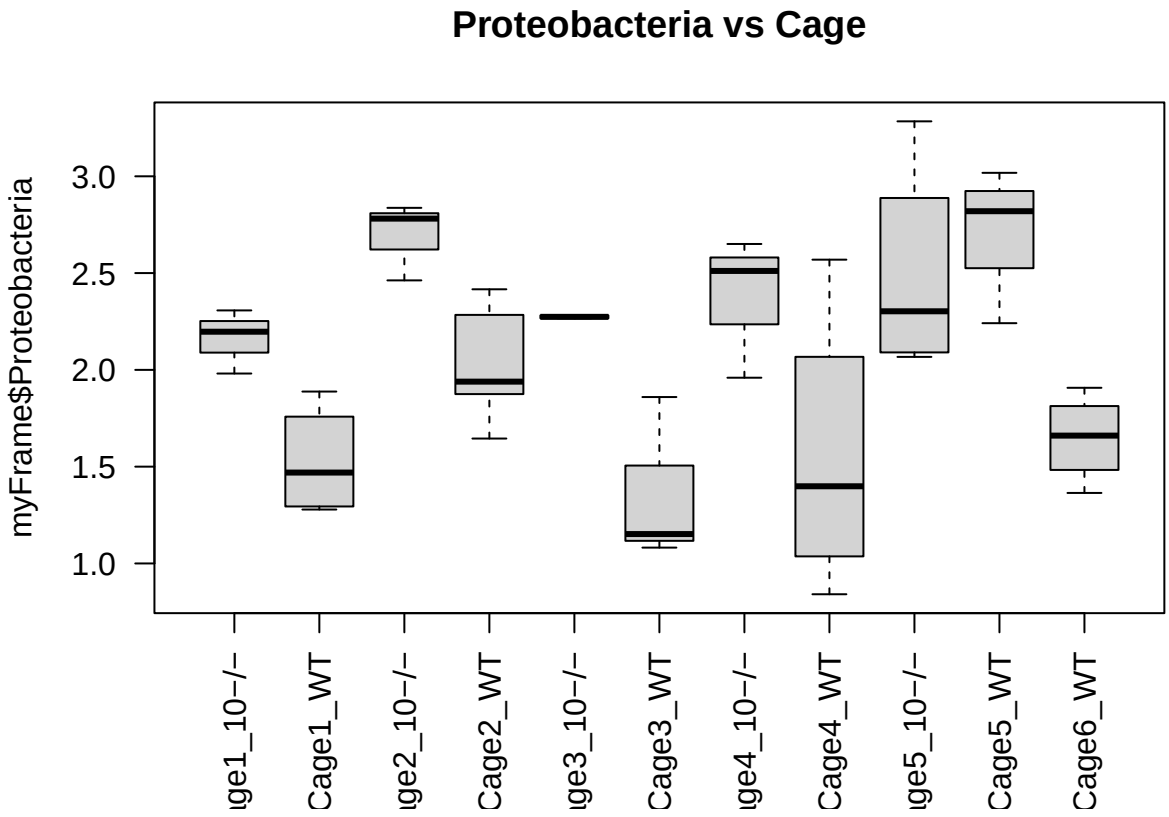

```
boxplot(myFrame$Actinobacteria ~ myFrame$cage, main = 'Actinobacteria vs Cage', las=2, xlab='')
```




```
boxplot(myFrame$Firmicutes ~ myFrame$cage, main = 'Firmicutes vs Cage', las=2, xlab='')
```




```
boxplot(myFrame$Proteobacteria ~ myFrame$cage, main = 'Proteobacteria vs Cage', las=2, xlab='')
```



If we look at the box plots, we notice that there is a cage effect to the phyla. We notice not all the plots have similar relative abundance and also the mean also differs for each phyla in cage categories.

```
library('nlme')

pValuesMixed = vector()
rhoGLS = vector()

for (i in 1:length(bug)) {
  myBug <- bug[,i]
  myData = data.frame(myBug, genotype, cage)
  M.mixed = lme(myBug ~ genotype, method = 'REML', random = ~ 1|cage, data = myData)
  pVal = unclass(summary(M.mixed))$tTable[2,5]
  pValuesMixed[i] = pVal

  M.gls = gls(myBug ~ genotype, method = 'REML', correlation = corCompSymm(form = ~1 | cage), data = myl
  rhoVal = coef(M.gls$modelStruct[1]$corStruct,unconstrained=FALSE)[[1]]
  rhoGLS[i] = rhoVal
}

bugNames = names(bug)

for (i in 1:length(bugNames)) {
  print(paste(bugNames[i],' rho: ', round(rhoGLS[i], 3)))

}
```

```
## [1] "Tenericutes  rho:  0.608"
## [1] "Verrucomicrobia  rho:  0.647"
## [1] "Bacteroidetes  rho:  0.581"
## [1] "Actinobacteria  rho:  -0.04"
## [1] "Firmicutes  rho:  0.564"
## [1] "Proteobacteria  rho:  0.427"
```

```
pValue_adj = p.adjust(pValuesMixed, method = 'BH')

sum(pValue_adj <= 0.1)
```

```
## [1] 3
```

We see that there are 3 types of phyla that are significantly different for genotypes in the Mixed Model at 10% FDR.