PH245_assignment1

#1.

```
#Loading data
setwd("/Users/xiaoyingliu/desktop")
dat1=read.table('dat1.txt',header=F,quote='')
colnames(dat1)=c('word-different',
                        'word-same',
                        'arabic-different',
                        'arabic-same')
head(dat1)
```

```
##   word-different word-same arabic-different arabic-same
## 1            869     860.5            691.0         601
## 2            995     875.0            678.0         659
## 3           1056     930.5            833.0         826
## 4           1126     954.0            888.0         728
## 5           1044     909.0            865.0         839
## 6            925     856.5           1059.5         797
```

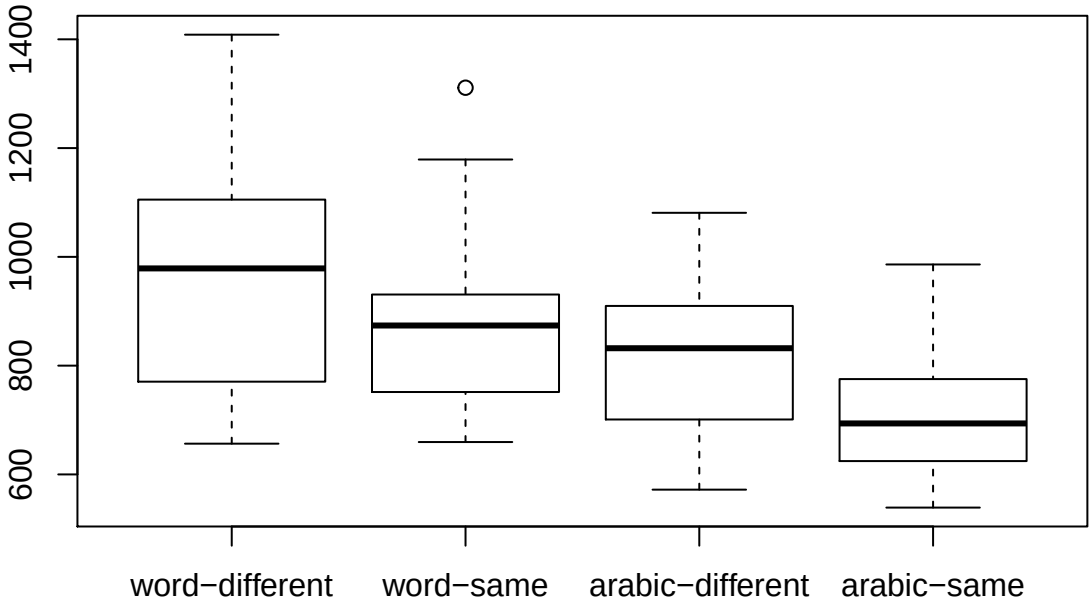
```
#EDA
summary(dat1)
```

```
##  word-different     word-same      arabic-different  arabic-same
##  Min.   : 656.5   Min.   : 659.5   Min.   : 572.0   Min.   :539.0
##  1st Qu.: 772.2   1st Qu.: 752.0   1st Qu.: 706.0   1st Qu.:624.8
##  Median : 978.8   Median : 873.8   Median : 832.0   Median :693.8
##  Mean   : 967.6   Mean   : 875.6   Mean   : 825.3   Mean   :710.9
##  3rd Qu.:1100.9   3rd Qu.: 930.6   3rd Qu.: 907.1   3rd Qu.:770.6
##  Max.   :1408.5   Max.   :1311.0   Max.   :1081.0   Max.   :986.0
```

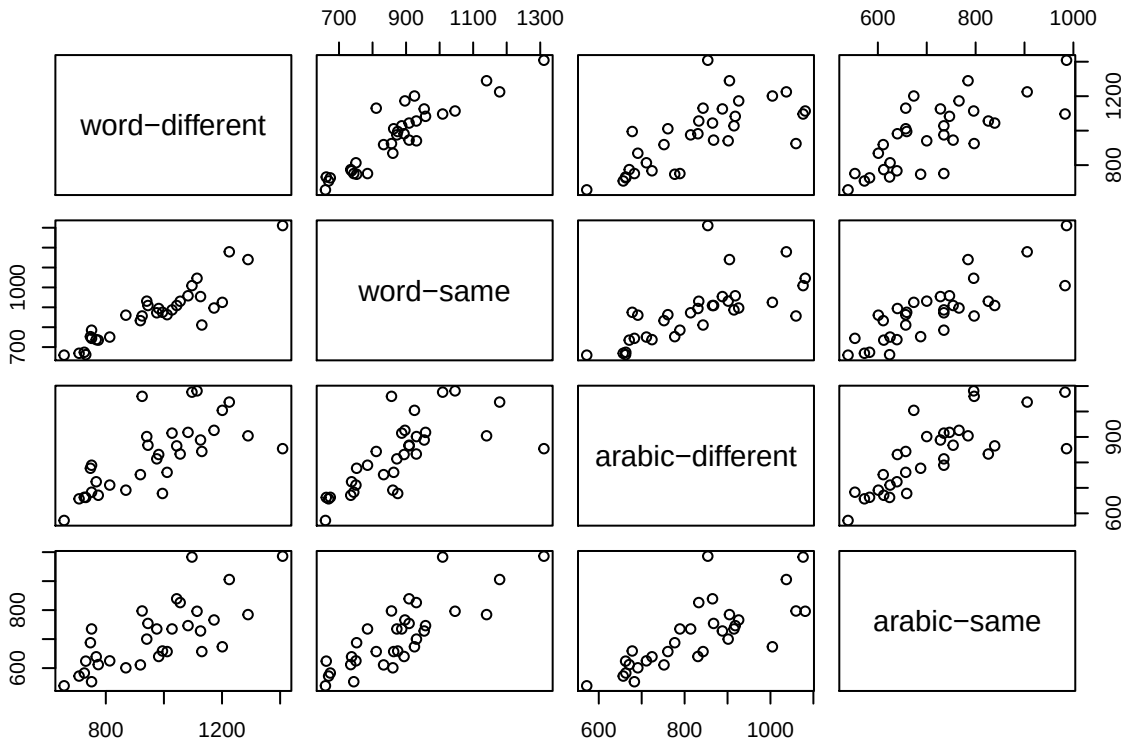
```
nrow(dat1)
```

```
## [1] 32
```

```
boxplot(dat1)
```




```
plot(dat1)
```




```
#comments from EDA
#Variable in the dataset correlates with each other.I will consider there are in total 4 treatments.
#Comparing parity will be considered as level of factors. The response variable is reaction time.
#Null hypothesis: the cogntive processing time of numbers does not depend on how they are presented.
#u1=u2=u3=u4
#Test:repeated measures design
#Reasoning: The treatment is independent to each other and all 4 treatment are testing the median
#of cognitive processing time of numbers.
#Test Statistics:T^2 = n(CXbar)Transpose(CSCTranspose)^-1(CXbar)
```

```r
# Gathering relevant variable data for the test statistic
n = nrow(dat1)
xBar = apply(dat1, 2, mean)
s = cov(dat1)
c = rbind(c(-1, 1, 0, 0),
          c(0, -1, 1, 0),
          c(0, 0, -1, 1)
         )

tsquaredRepeatedMeasures = function(n, xBar, s, c) {
    return( n *
            t( c %*% xBar ) %*%
            solve( c %*% s %*% t(c) ) %*%
            c %*% xBar
          )
}
# Calculating test statistic
observedTestStatistic1= tsquaredRepeatedMeasures(n, xBar, s, c)
print(observedTestStatistic1)
```

```
##           [,1]
## [1,] 153.7275
```

```r
# P-value is tSquared / ( (p)(n-1)/(n-p) ) in the F distribution
# n=nrows, p=degrees of freedom=num variables - 1
observedPValue1 = 1 - pf(q=observedTestStatistic1/ (3*31/29),
                         df1=3,
                         df2=31
                        )
print(observedPValue1)
```

```
##               [,1]
## [1,] 9.433565e-12
```

```r
#Test Statistic Interpretation
#Since the significance level of 0.05, we will reject the null hypothesis that the cognitive processing
#of numbers doesn't depend on numbers are presented. However, the cognitive processing of numbers
#does depend on the way numbers are presented and their parity.
```

#2.

```r
#Loading data
dat2= read.table(file='dat2.txt', header=FALSE, quote='')
colnames(dat2) = c('Fuel',
                       'Repair',
                       'Capital',
                       'EngineType'
                      ) #per mile
head(dat2)
```

```
##    Fuel Repair Capital EngineType
## 1 16.44  12.43   11.23   gasoline
## 2  7.19   2.70    3.92   gasoline
## 3  9.92   1.35    9.75   gasoline
## 4  4.24   5.78    7.78   gasoline
## 5 11.20   5.05   10.67   gasoline
```

```
## 6 14.25   5.78    9.88   gasoline
```

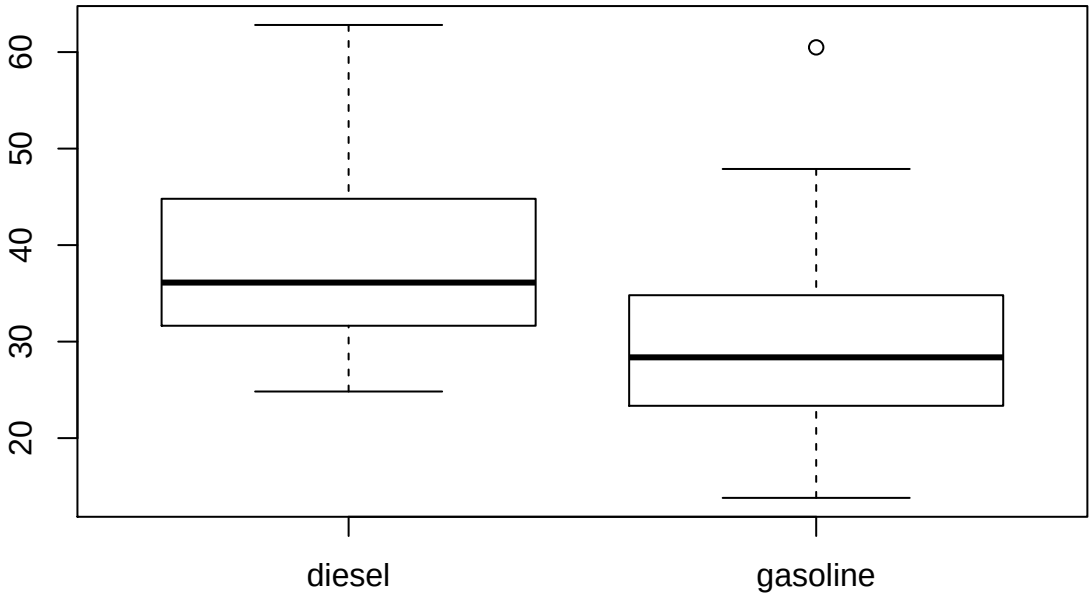
```
# EDA
summary(dat2)
```

```
##       Fuel           Repair          Capital         EngineType
##  Min.   : 4.24   Min.   : 1.350   Min.   : 3.28   diesel  :23
##  1st Qu.: 9.12   1st Qu.: 5.145   1st Qu.: 8.15   gasoline:36
##  Median :10.28   Median : 8.890   Median :11.23
##  Mean   :11.39   Mean   : 9.145   Mean   :12.93
##  3rd Qu.:12.70   3rd Qu.:12.575   3rd Qu.:17.00
##  Max.   :29.11   Max.   :21.520   Max.   :35.18
```

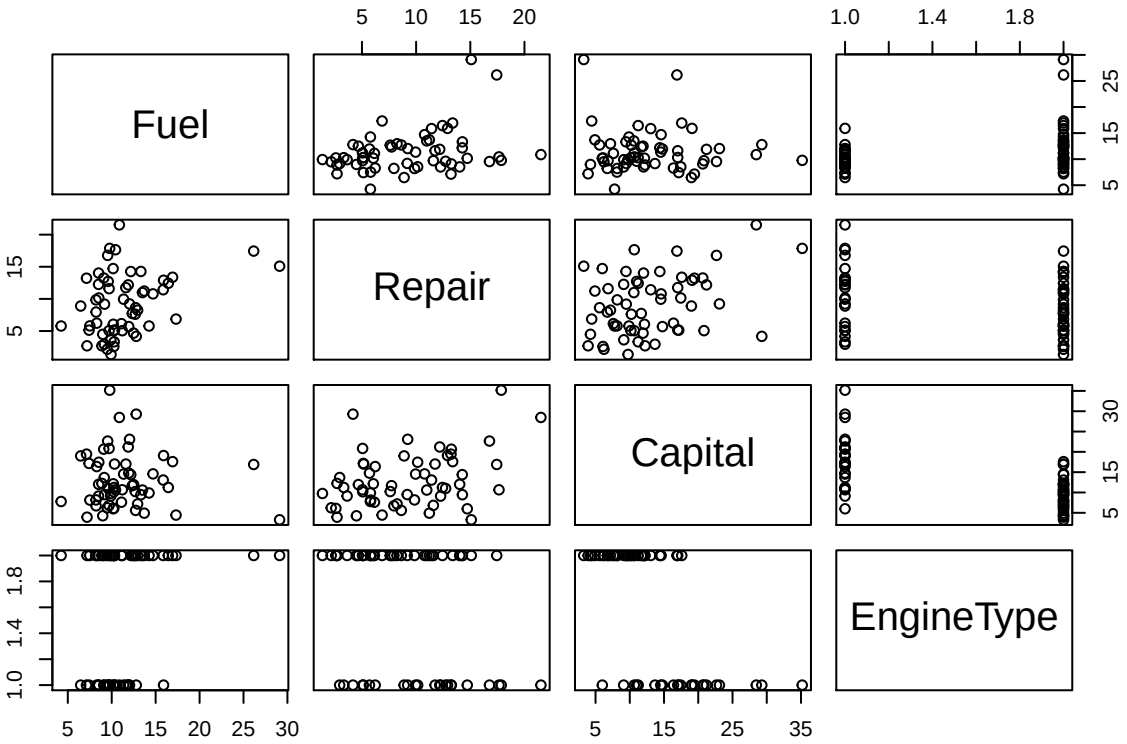
```
nrow(dat2)
```

```
## [1] 59
```

```
boxplot(formula=Fuel+Repair+Capital ~ EngineType, data=dat2)
```




```
plot(dat2)
```




```
#Comments from EDA
#We are testing if the two types of trucks have statistically significantly
#different mean costs from each other.
#Null Hypothesis: u1=u2, u1 is the mean costs of a gasoline truck and u2 is the mean
#costs of a diesel truck. The two types of trucks have the same mean costs per mile
#in regards to the three observed variables.
#Alternative Hypothesis: u1!=u2, where u1 is the mean costs of a gasoline truck
#and u2 is the mean costs of a diesel truck. The two types of trucks do not have
#the same mean costs per mile to in regards to the three observed variables.
#Test: Comparing Mean Vectors from Two Populations.
#Reasoning: We are comparing means from two different populations.
#Test Statistic: (xBar1-xBar2)Transpose * (S(1/n1 + 1/n2))^-1 * (xBar1-xBar2)

gasoline = dat2[dat2$EngineType == 'gasoline',]
diesel = dat2[dat2$EngineType == 'diesel',]
```

```
# preparing variables for tetsing
n1 = nrow(gasoline)
n2 = nrow(diesel)

xBar1 = apply(gasoline[1:3], 2, mean)
xBar2 = apply(diesel[1:3], 2, mean)

s = cov(dat2[1:3])

tsquaredTwoPopMeans = function(n1, n2, xBar1, xBar2, s) {
    return( t(xBar1 - xBar2) %*%
            solve(s * (1/n1 + 1/n2)) %*%
            (xBar1 - xBar2)
          )
}
```

```r
# Calculating test statistic
observedTestStatistic2 = tsquaredTwoPopMeans(n1, n2, xBar1, xBar2, s)
observedTestStatistic2
```

```
##          [,1]
## [1,] 27.36415
```

```r
# P-value is tSquared / ( (n1 + n2 - 2)(p)/(n1+n2-p-1) ) in the F distribution
# n=nrows, p=degrees of freedom=num variables - 1
observedPValue2 = 1 - pf(q=observedTestStatistic2 / ((n1+n2-2)*2/(n1+n2-2-1)),
                         df1=2,
                         df2=n1+n2-1
                        )
observedPValue2
```

```
##              [,1]
## [1,] 1.597538e-05
```

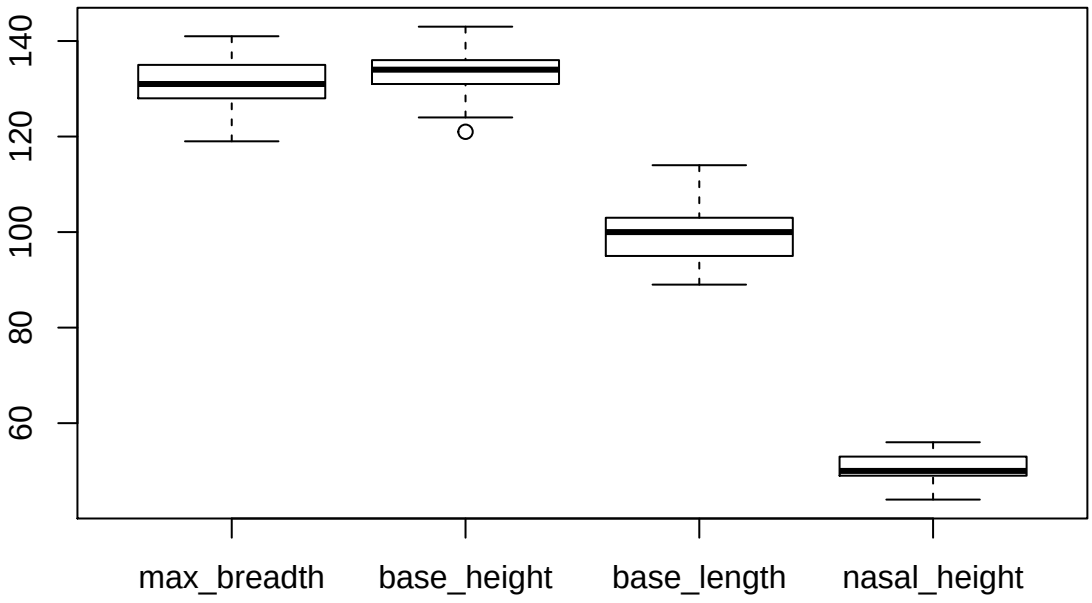
```r
#Test Statistic Interpretation
#With a significance level of .05, we can reject the null hypothesis that
#the two types of trucks (diesel or gasoline) have the same mean costs per mile
#with respect to the three observed variables.
```

#3.

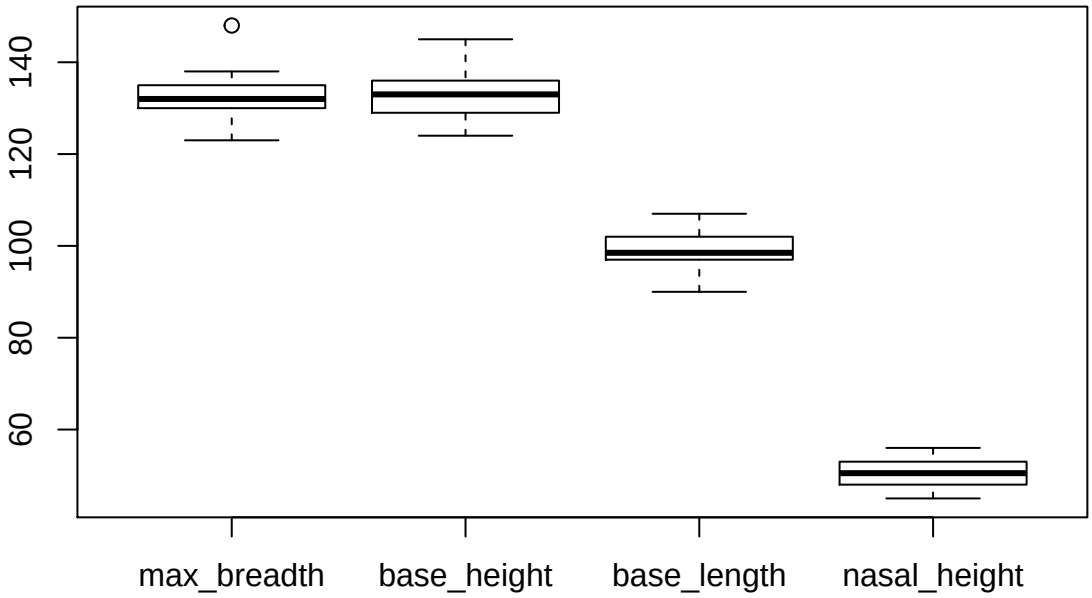
```r
# Loading data
dat3 = read.table(file='dat3.txt', header=FALSE, quote='')
colnames(dat3) = c('max_breadth',
                       'base_height',
                       'base_length',
                       'nasal_height',
                       'time_period'
                      )
head(dat3)
```

```
##   max_breadth base_height base_length nasal_height time_period
## 1         131         138          89           49           1
## 2         125         131          92           48           1
## 3         131         132          99           50           1
## 4         119         132          96           44           1
## 5         136         143         100           54           1
## 6         138         137          89           56           1
```

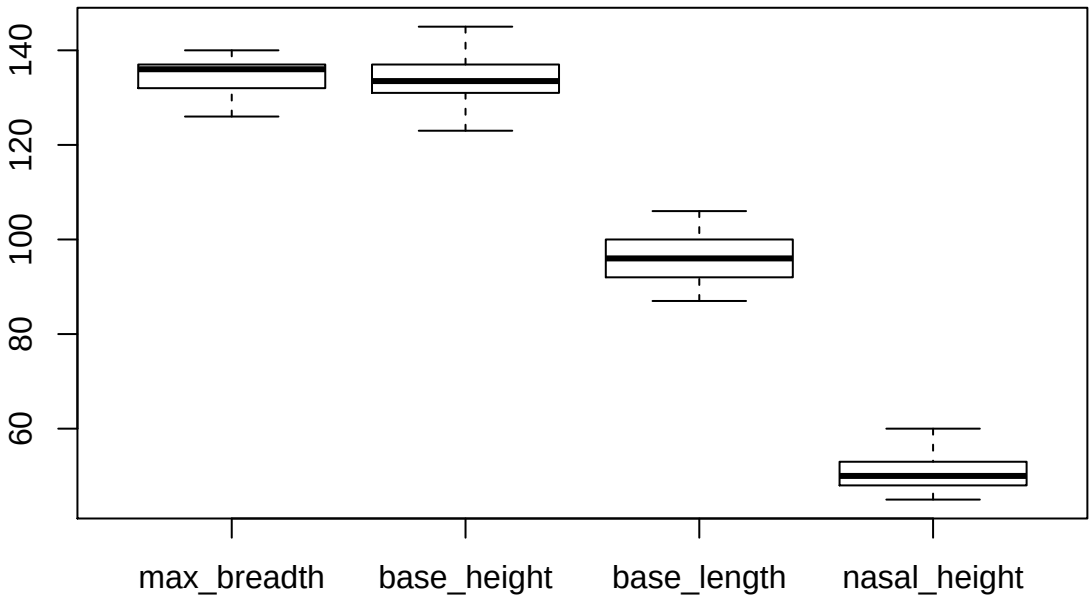
```r
# EDA

period1 = dat3[dat3$time_period == 1,]
period2 = dat3[dat3$time_period == 2,]
period3 = dat3[dat3$time_period == 3,]

boxplot(period1[1:4])
```




```
boxplot(period2[1:4])
```




```
boxplot(period3[1:4])
```




```
#Comments from EDA
#We are testing if humans from resident population over three time periods have varying skull
#sizes which would provide evidence of the resident population interbreeding with immigrant populations
#Null Hypothesis: u1=u2=u3, tha there has been no change in skull size over the
#course of the time periods
#Alternative Hypothesis: At least one ui != uj for any i,j.
#that there has been a change in skull size over the course of the time periods
#Test: One-way MANOVA
#One-way MANOVA is chosen since we have one factor of 3 levels.
#And our variables are dependent to each other(max breadth, base height, base length, nasal height).
```

```
#Statistical Test
timePeriod = as.factor(dat3$time_period)

results = manova(
    cbind(max_breadth, base_height, base_length, nasal_height) ~ timePeriod,
    data=dat3
)

results
```

```
## Call:
##    manova(cbind(max_breadth, base_height, base_length, nasal_height) ~
##     timePeriod, data = dat3)
##
## Terms:
##                 timePeriod Residuals
## resp 1               150.2    1785.4
## resp 2                20.6    1924.3
## resp 3            190.2889 2153.0000
## resp 4              2.0222  840.2000
## Deg. of Freedom          2        87
##
## Residual standard errors: 4.530104 4.703019 4.974648 3.107647
## Estimated effects may be unbalanced
```

```r
summary(results)
```

```
##            Df  Pillai approx F num Df den Df Pr(>F)
## timePeriod  2 0.17221   2.0021      8    170 0.0489 *
## Residuals  87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary.aov(results)
```

```
##  Response max_breadth :
##             Df Sum Sq Mean Sq F value  Pr(>F)
## timePeriod   2  150.2  75.100  3.6595 0.02979 *
## Residuals   87 1785.4  20.522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response base_height :
##             Df Sum Sq Mean Sq F value Pr(>F)
## timePeriod   2   20.6  10.300  0.4657 0.6293
## Residuals   87 1924.3  22.118
##
##  Response base_length :
##             Df  Sum Sq Mean Sq F value  Pr(>F)
## timePeriod   2  190.29  95.144  3.8447 0.02512 *
## Residuals   87 2153.00  24.747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response nasal_height :
##             Df Sum Sq Mean Sq F value Pr(>F)
## timePeriod   2   2.02  1.0111  0.1047 0.9007
## Residuals   87 840.20  9.6575
```

```r
#Test Result Interpretation
#With a significance level of .05, we can reject the null hypothesis
#that no interbreeding occurred.
#From summary results, there was statistically significant variance in two of the
#variables over time.Which is max_breadth and base_length.
```

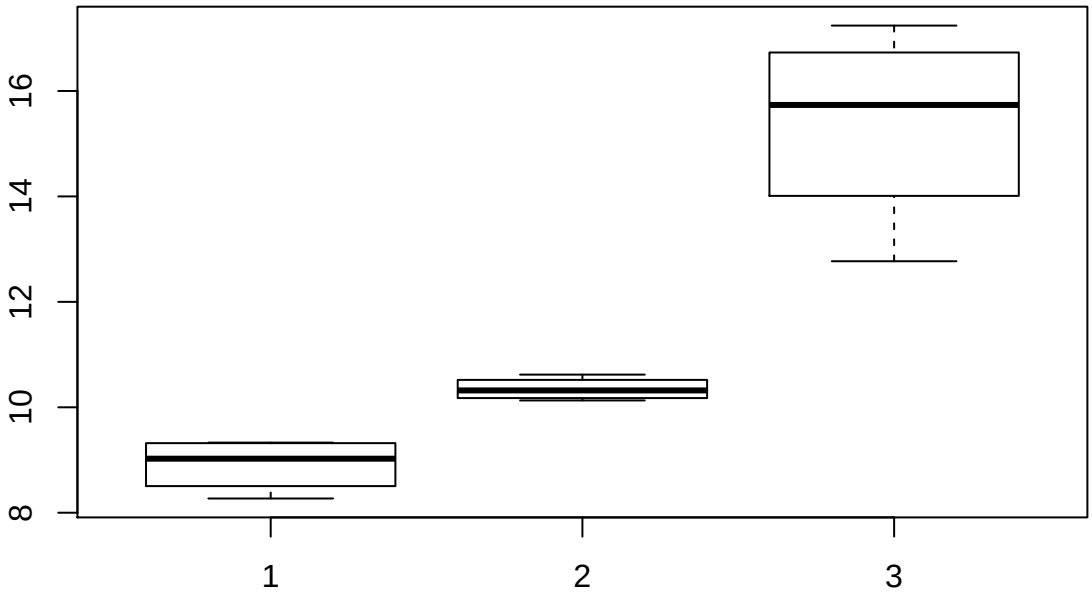
#4.

```r
# Loading data
dat4 = read.table(file='dat4.txt', header=FALSE, quote='')
colnames(dat4) = c('reflectance_green',
                   'reflectance_near_infared',
                   'species',
                   'time_period',
                   'treeID'
                  )
head(dat4)
```

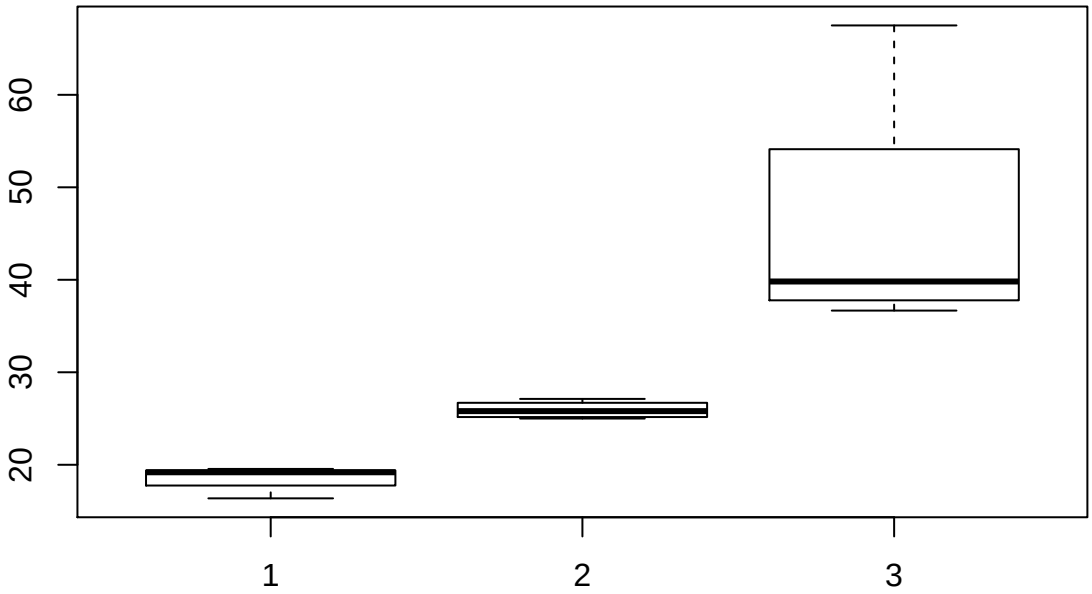
```
##   reflectance_green reflectance_near_infared species time_period treeID
## 1              9.33                   19.14      SS           1      1
## 2              8.74                   19.55      SS           1      2
## 3              9.31                   19.24      SS           1      3
```

```
## 4              8.27                    16.37      SS          1      4
## 5             10.22                    25.00      SS          2      1
## 6             10.13                    25.32      SS          2      2
```

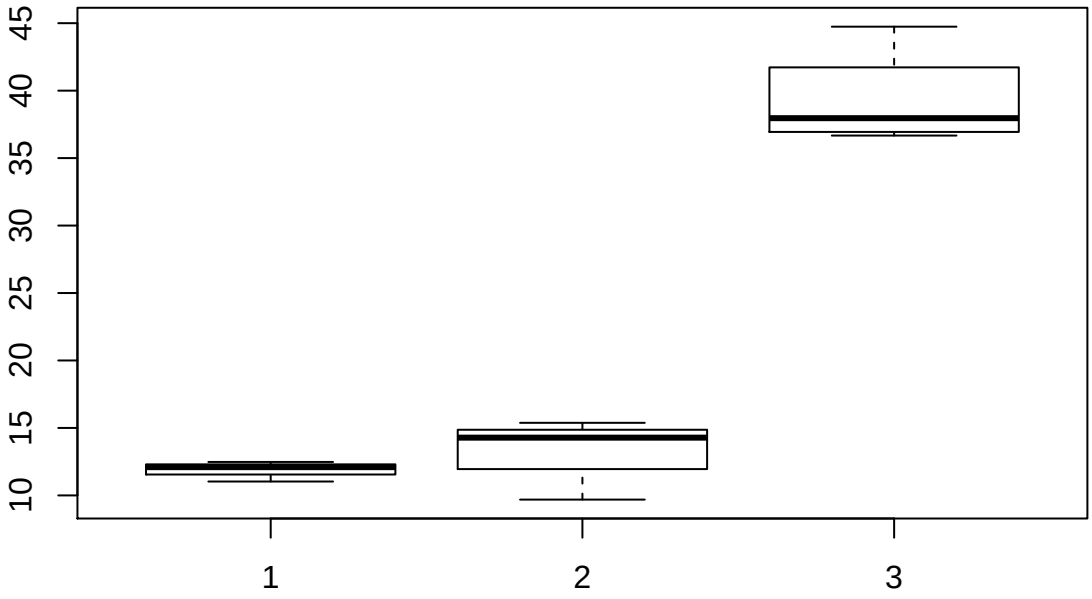
```
#EDA
SS = dat4[dat4$species == 'SS',]
JL = dat4[dat4$species == 'JL',]
LP = dat4[dat4$species == 'LP',]

#EDA of SS
boxplot(reflectance_green ~ time_period, data=SS)
```



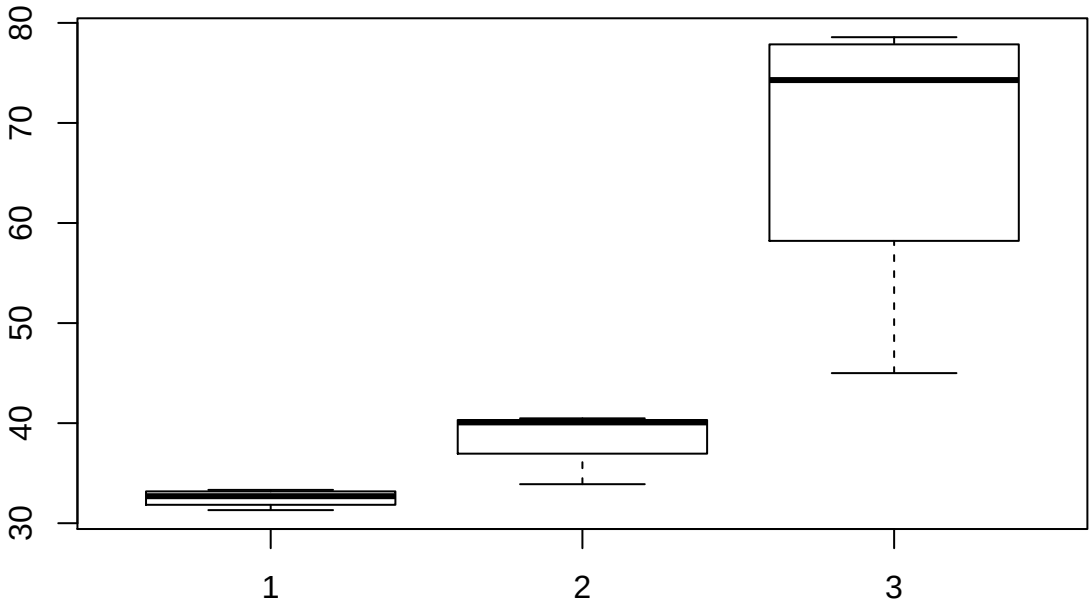

```
boxplot(reflectance_near_infared ~ time_period, data=SS)
```



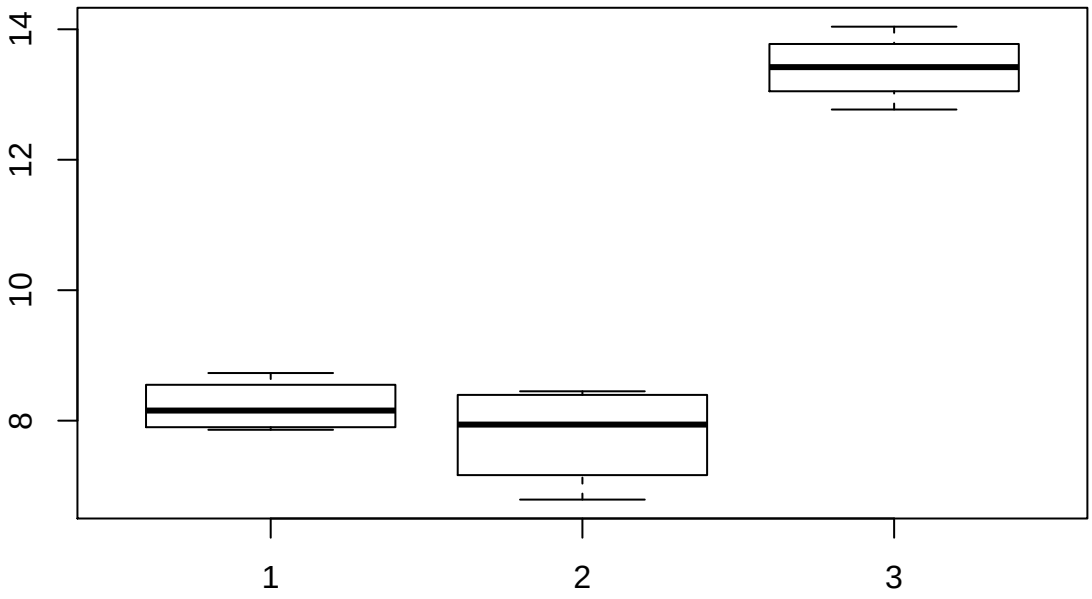

```
#EDA of Species JL
boxplot(reflectance_green ~ time_period, data=JL)
```



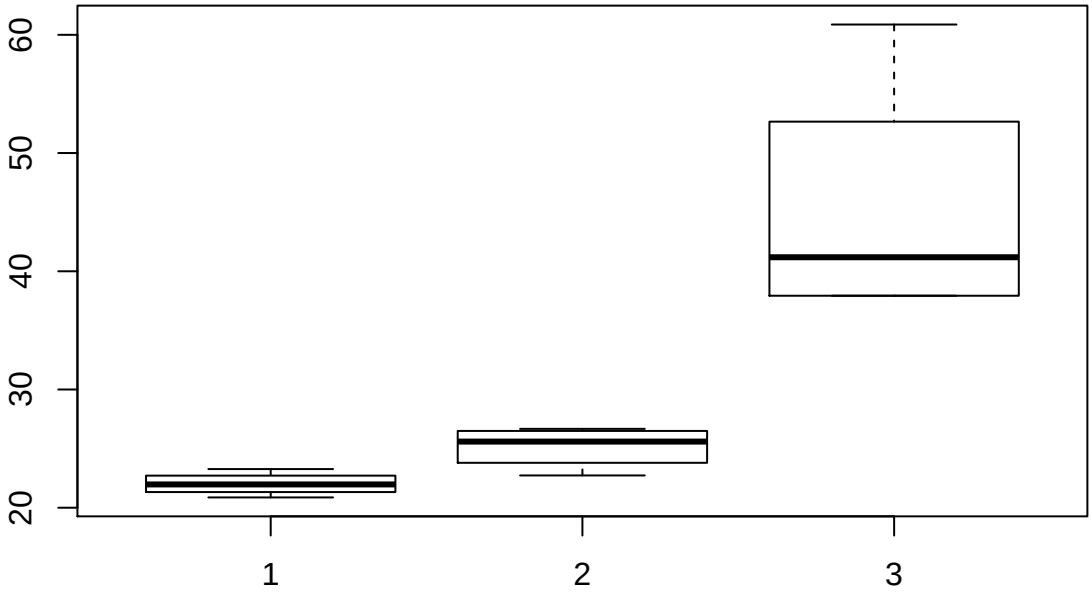

```
boxplot(reflectance_near_infared ~ time_period, data=JL)
```




```
#EDA of Species LP
boxplot(reflectance_green ~ time_period, data=LP)
```




```r
boxplot(reflectance_near_infared ~ time_period, data=LP)
```




```r
###############
#Comments from EDA
#We are testing whether there is a difference between our two dependent variables
#based on our two factors, species and time period. We're also trying to understand
#whether an interaction effect exists between our two independent variables.
#Null Hypothesis: u1=u2=u3, There is no species effect, no time effect, and no interaction effect
#on the green and near-infared reflectance.
#Alternative Hypothesis:There is at least one of: 1) a species effect, 2) a time effect,
#or 3) an interaction effect on the reflectance of the seedlings.
#Test: Two-way MANOVA
#Reasoning: we have two factors (time period, species) with 3 levels each (1, 2, 3; SS, JL, LP)
#and that affects multiple dependent variables (green and near-infared reflectance).
```

```r
#Statistical Test
timePeriod = as.factor(dat4$time_period)
```

```r
species = as.factor(dat4$species)

results = manova(
    cbind(reflectance_green, reflectance_near_infared) ~ timePeriod*species,
    data=dat4
)

results
```

```
## Call:
##    manova(cbind(reflectance_green, reflectance_near_infared) ~ timePeriod *
##     species, data = dat4)
##
## Terms:
##                 timePeriod  species timePeriod:species Residuals
## resp 1            1275.248  965.181            795.808    76.659
## resp 2            5573.806 2026.856            193.549  1769.642
## Deg. of Freedom          2        2                  4        27
##
## Residual standard errors: 1.684997 8.09582
## Estimated effects may be unbalanced
```

```r
summary(results)
```

```
##                    Df  Pillai approx F num Df den Df    Pr(>F)
## timePeriod          2 0.99199  13.2853      4     54 1.330e-07 ***
## species             2 0.96120  12.4915      4     54 2.910e-07 ***
## timePeriod:species  4 0.92116   5.7634      8     54 2.606e-05 ***
## Residuals          27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary.aov(results)
```

```
##  Response reflectance_green :
##                    Df  Sum Sq Mean Sq F value    Pr(>F)
## timePeriod          2 1275.25  637.62 224.578 < 2.2e-16 ***
## species             2  965.18  482.59 169.973 5.027e-16 ***
## timePeriod:species  4  795.81  198.95  70.073 7.341e-14 ***
## Residuals          27   76.66    2.84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response reflectance_near_infared :
##                    Df Sum Sq Mean Sq F value    Pr(>F)
## timePeriod          2 5573.8 2786.90 42.5207 4.537e-09 ***
## species             2 2026.9 1013.43 15.4622 3.348e-05 ***
## timePeriod:species  4  193.5   48.39  0.7383    0.5741
## Residuals          27 1769.6   65.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Test Result Interpretation
#With a significance level of .05, we reject the null hypothesis that there is no species,
#time period, or interaction effect.
#Based on the summary.aov results, there was statistically significant variance
```

```
#in both reflectances for time and species effect. However,
#Only green reflectance have an interation effect(with a significant level ***),
#near_infared reflectance does not have an statistical significant interation effect.
```