PH 245 Homework 1

Kunal Mishra

Problem 1

In [1]:
```
# Loading data
pb1_data = read.table(file='Data-HW1-Cognition.dat', header=FALSE, quote='')
colnames(pb1_data) = c('word-different',
                        'word-same',
                        'arabic-different',
                        'arabic-same'
                       )
head(pb1_data)
```

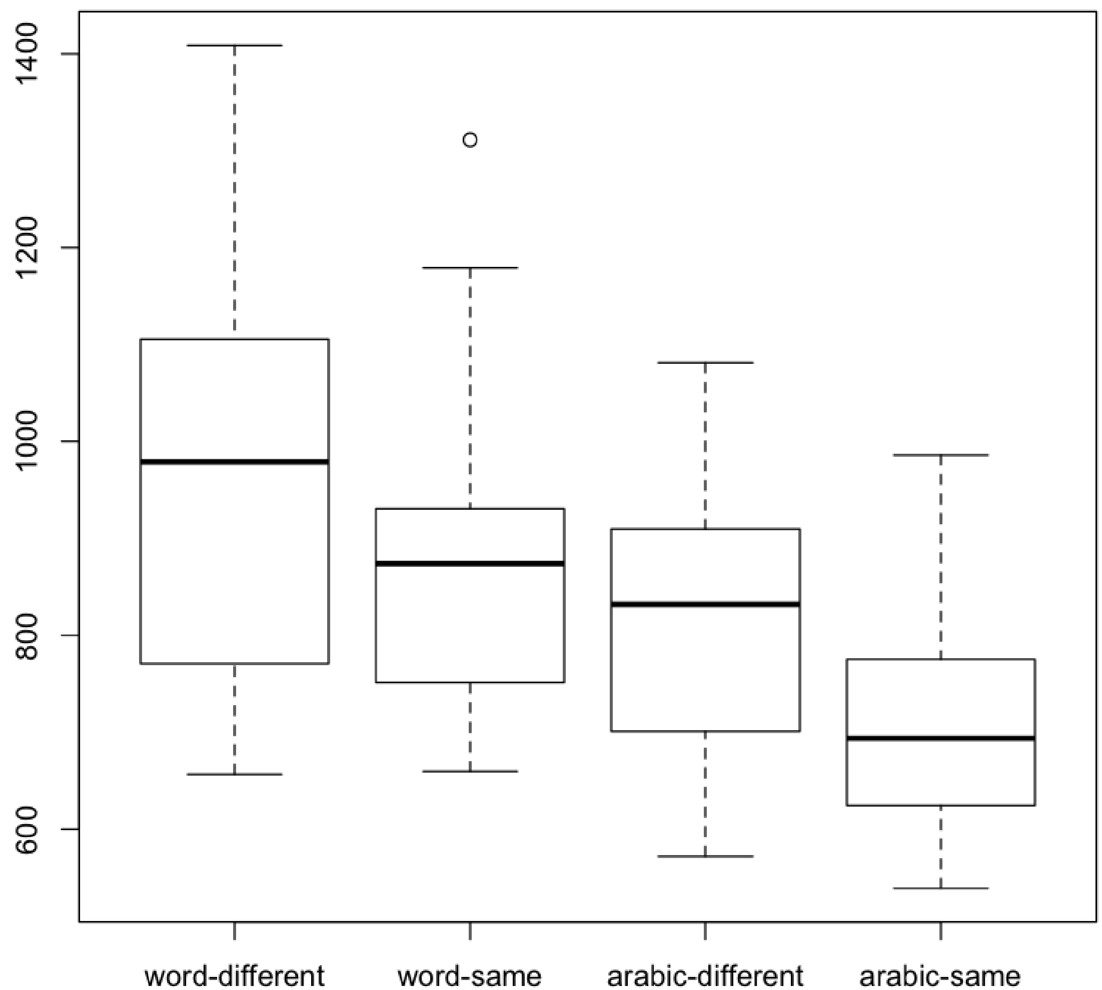
word-different	word-same	arabic-different	arabic-same
869	860.5	691.0	601
995	875.0	678.0	659
1056	930.5	833.0	826
1126	954.0	888.0	728
1044	909.0	865.0	839
925	856.5	1059.5	797

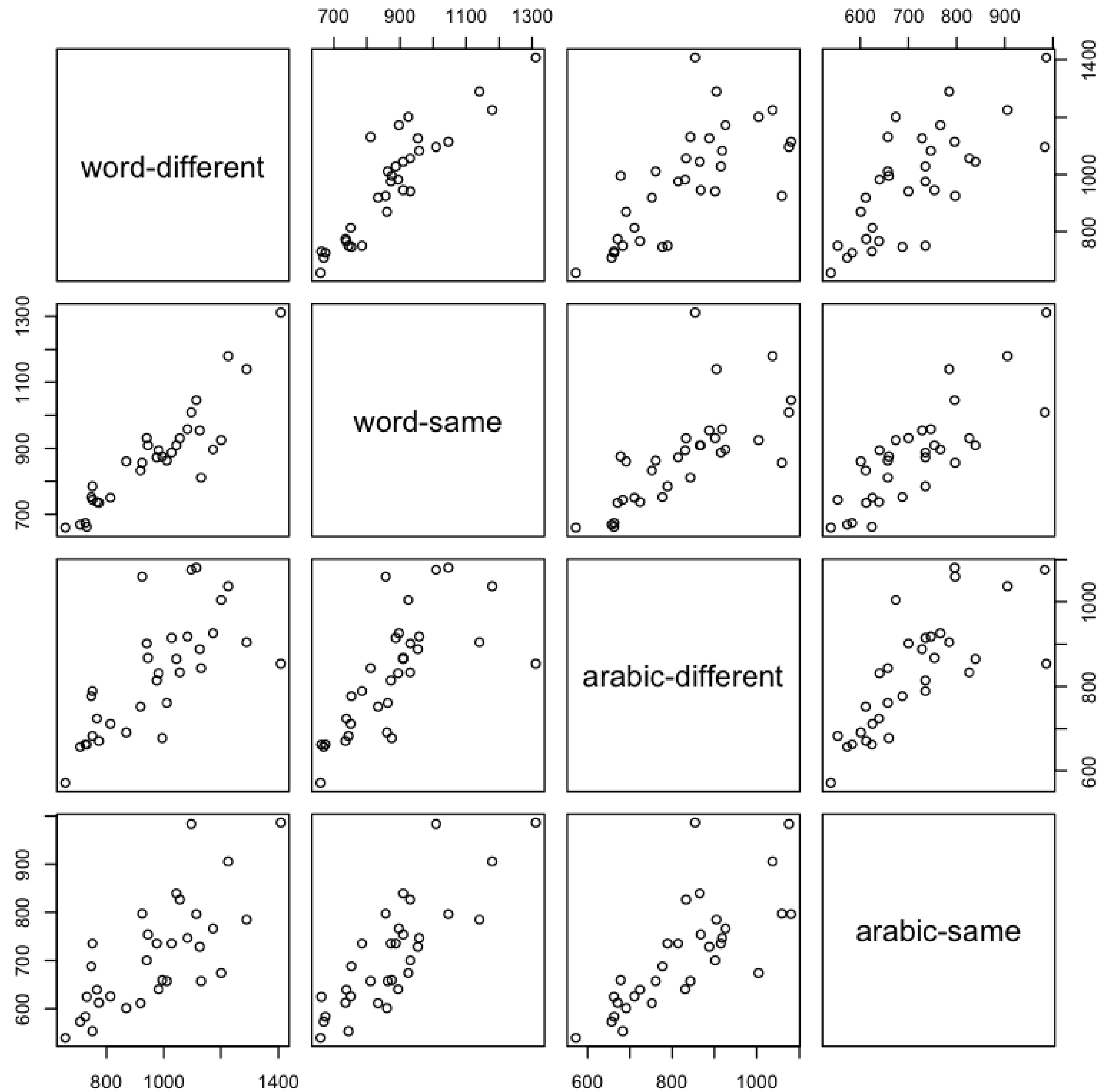
In [2]:
```
# Exploratory Data Analysis
summary(pb1_data)
nrow(pb1_data)
```

```
 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
```

32

In [3]:
```
# Exploratory Data Analysis
boxplot(pb1_data); plot(pb1_data)
```





General Comments

1. It seems like every measured variable in the dataset correlates with every other variable
2. *Each* subject was treated with the all 4 treatments, so this study design has me leaning toward an intra-subject repeated measures design. One issue I have with doing this is that I'm treating the 4 measured variables as 4 seperate treatments whereas it is more intuitive to think about it as 2 treatments (word-format and Arabic-digit-format) and comparing parity (same and different) as factors or levels of factors.

    - Treatment 1: Word-Same
    - Treatment 2: Word-Different
    - Treatment 3: Arabic-digit-Same
    - Treatment 4: Arabic-digit-Different
    - Further resources on factors: [http://stattrek.com/statistics/dictionary.aspx?definition=Factor (http://stattrek.com/statistics/dictionary.aspx?definition=Factor)](http://stattrek.com/statistics/dictionary.aspx?definition=Factor)
    - Response variable: Reaction time

3. *Null Hypothesis*: Cu = 0; u1 = u2 = u3 = u4; The congitive processing of numbers **doesn't** depend on the way numbers are presented or their parity.
4. *Alternative Hypothesis*: Cu != 0; At least one ui != uj for some i, j in set(1, 2, 3, 4); The cognitive processing of numbers **does** depend on the way numbers are presented and their parity
5. Test: Repeated Measures design
6. Test Statistic: T^2 = n(CXbar)Transpose(CSCTranspose)^-1(CXbar)

In [4]:
```
# Gathering relevant variable data for the test statistic
n = nrow(pb1_data)

xBar = apply(pb1_data, 2, mean)

s = cov(pb1_data)

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
          )
}
```

In [5]:
```
# Calculating test statistic and p-value
observedPb1TestStatistic = tsquaredRepeatedMeasures(n, xBar, s, c)
print(paste('Test Statistic', observedPb1TestStatistic))

# P-value is tSquared / ( (p)(n-1)/(n-p) ) in the F distribution
# n=nrows, p=degrees of freedom=num variables - 1
observedPb1PValue = 1 - pf(q=observedPb1TestStatistic / (3*31/29),
                        df1=3,
                        df2=31
                       )
print(paste('P-Value:', observedPb1PValue))
```

```
[1] "Test Statistic 153.727505641501"
[1] "P-Value: 9.43356504023996e-12"
```

Test Statistic Interpretation

With a significance level of .05, our p-value indicates that we can reject the null hypothesis that the cognitive processing of numbers doesn't depend on doesn't depend on the way numbers are presented or their parity. Rather, we have evidence that at least one ui != uj for some i, j in set(1, 2, 3, 4) and that the cognitive processing of numbers does depend on the way numbers are presented and their parity.

In particular, I suspect based on this evidence and our initial EDA with the boxplots, that within our two factors word format < arabic-digit-format and different < same in terms of ease of comprehension.

Problem 2

In [6]:
```
# Loading data
pb2_data = read.table(file='Data-HW1-Transportation.dat', header=FALSE, quot
colnames(pb2_data) = c('Fuel',
                       'Repair',
                       'Capital',
                       'EngineType'
                      ) #All cost of transport per mile
head(pb2_data)
```

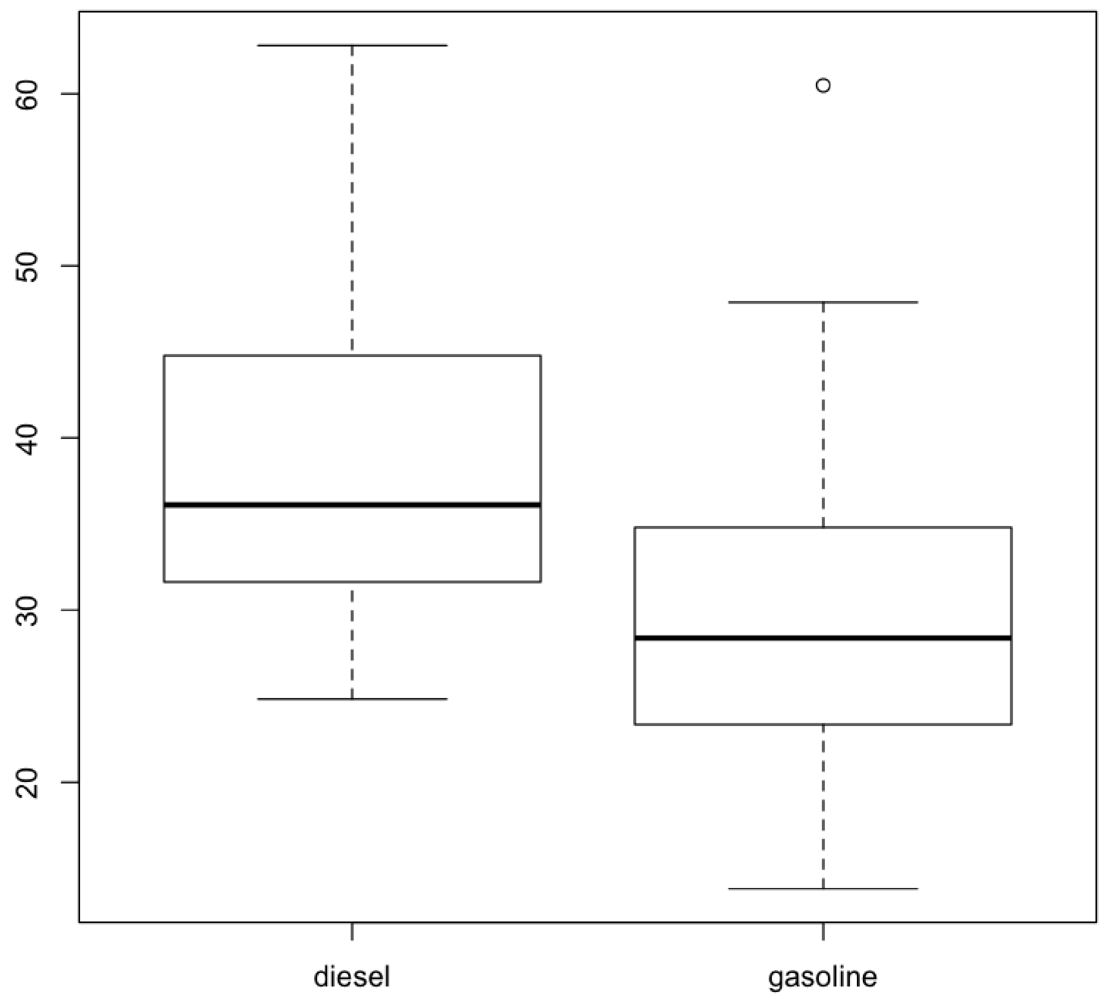
Fuel	Repair	Capital	EngineType
16.44	12.43	11.23	gasoline
7.19	2.70	3.92	gasoline
9.92	1.35	9.75	gasoline
4.24	5.78	7.78	gasoline
11.20	5.05	10.67	gasoline
14.25	5.78	9.88	gasoline

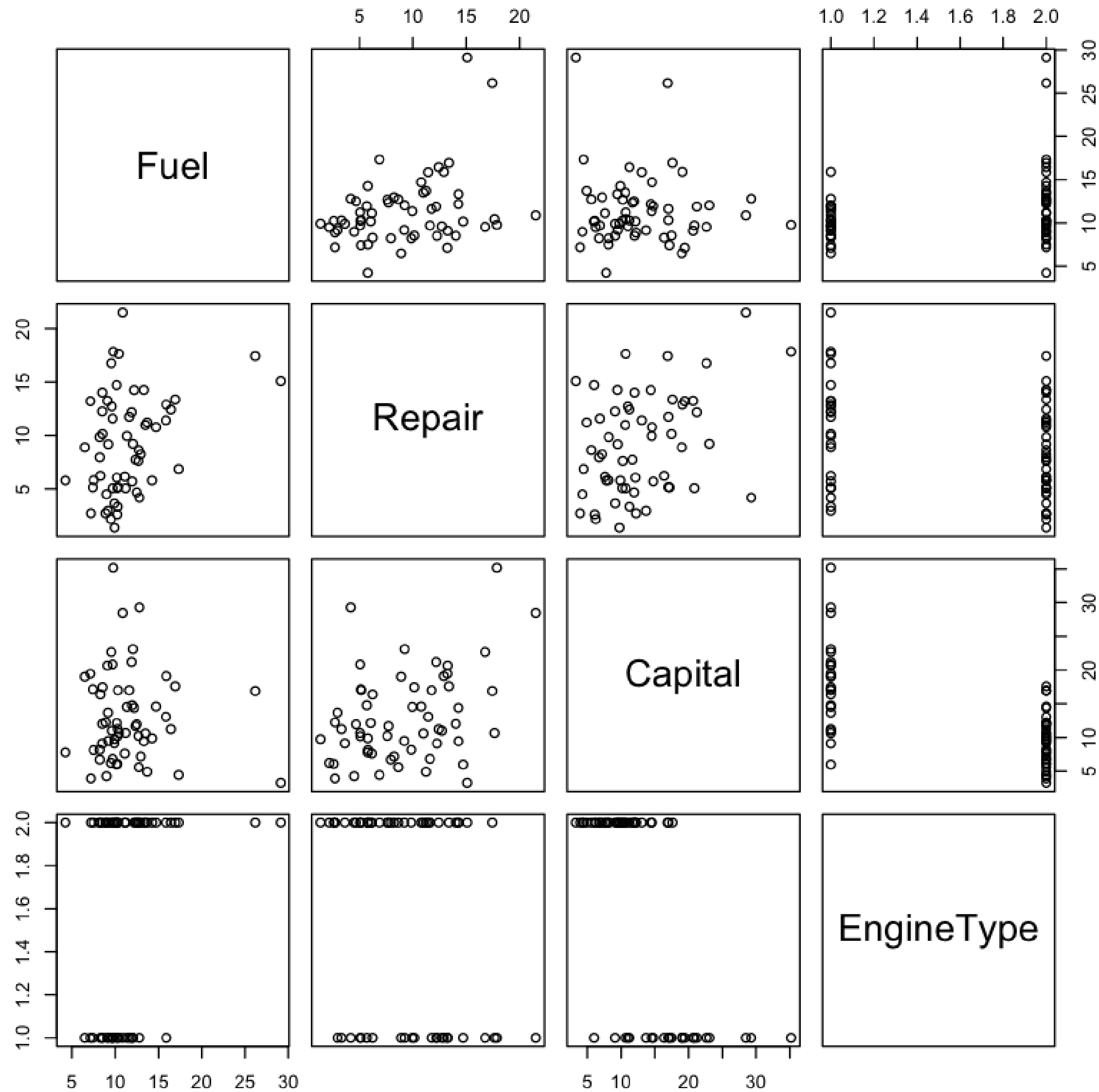
In [7]:
```
# EDA
summary(pb2_data)
nrow(pb2_data)
```

```
      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
```

59

In [8]:
```
# EDA
boxplot(formula=Fuel+Repair+Capital ~ EngineType, data=pb2_data)
plot(pb2_data)
```





General Comments

1. The question we're examining is if the two types of trucks have statistically significantly different mean costs from each other. Intuitively, we're delving into whether the variance in cost of our observed samples is due to pure chance or whether there is a systematic difference in cost between the two types of trucks.
2. *Null Hypothesis*: u1-u2=0, where u1 is the mean vector of costs of a gasoline truck and u2 is the mean vector of costs of a diesel truck. The two types of trucks (diesel or gasoline) have the same mean costs per mile to operate with respect to the three observed variables.
3. *Alternative Hypothesis*: u1-u2!=0, where u1 is the mean vector of costs of a gasoline truck and u2 is the mean vector of costs of a diesel truck. The two types of trucks (diesel or gasoline) do not have the same mean costs per mile to operate with respect to the three observed variables.
4. Test: Comparing Mean Vectors from Two Populations
5. Test Statistic: (xBar1-xBar2)Transpose * (S(1/n1 + 1/n2))^-1 * (xBar1-xBar2)

In [9]:
```
# Filtering dataset
gasoline = pb2_data[pb2_data$EngineType == 'gasoline',]
diesel = pb2_data[pb2_data$EngineType == 'diesel',]
```

In [10]:
```
# Gathering relevant variable data for the test statistic

n1 = nrow(gasoline)
n2 = nrow(diesel)

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

s = cov(pb2_data[1:3])

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

In [11]:
```
# Calculating test statistic and p-value
observedPb2TestStatistic = tsquaredTwoPopMeans(n1, n2, xBar1, xBar2, s)
print(paste('Test Statistic', observedPb2TestStatistic))

# P-value is tSquared / ( (n1 + n2 - 2)(p)/(n1+n2-p-1) ) in the F distribut
# n=nrows, p=degrees of freedom=num variables - 1
observedPB2PValue = 1 - pf(q=observedPb2TestStatistic / ((n1+n2-2)*2/(n1+n2-
                         df1=2,
                         df2=n1+n2-1
                        )
print(paste('P-Value:', observedPB2PValue))
```

```
[1] "Test Statistic 27.3641495407546"
[1] "P-Value: 1.59753787296602e-05"
```

Test Statistic Interpretation

With a significance level of .05, our p-value indicates that we can reject the null hypothesis that the two types of trucks (diesel or gasoline) have the same mean costs per mile to operate with respect to the three observed variables.

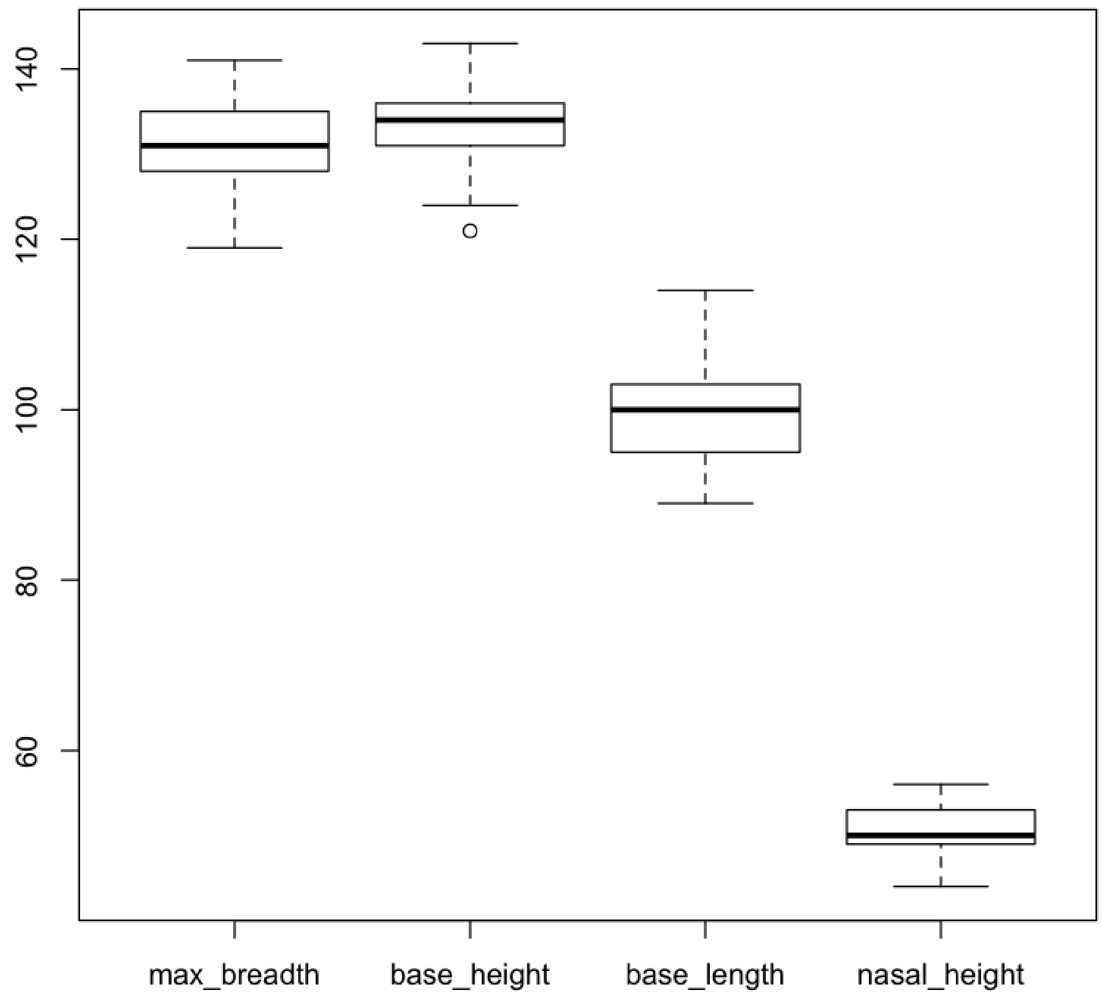
I suspect, based on this evidence and our initial EDA with the boxplots, that diesel-engine trucks are more expensive to operate than gasoline-engine trucks on a per mile basis.

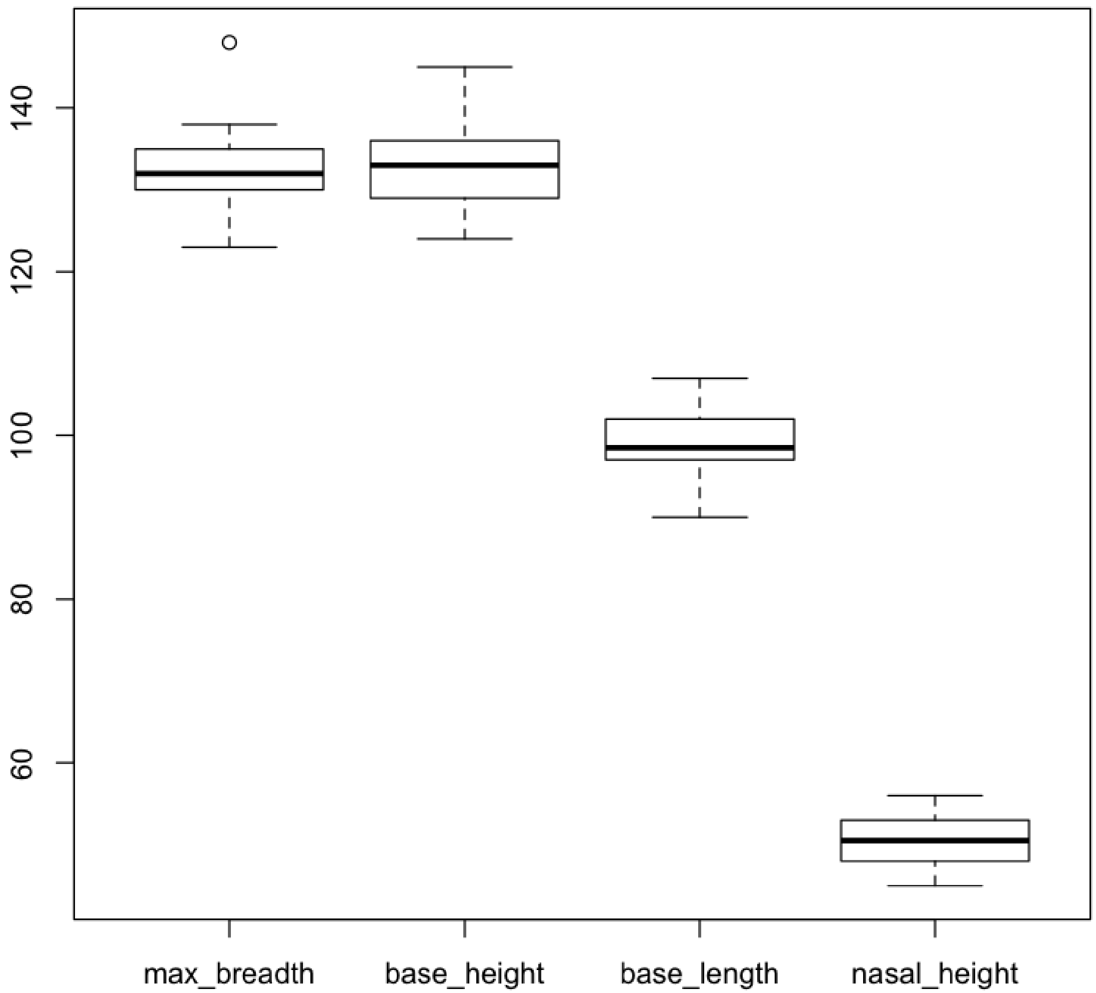
Problem 3

In [12]:
```
# Loading data
pb3_data = read.table(file='Data-HW1-Skull.dat', header=FALSE, quote='')
colnames(pb3_data) = c('max_breadth',
                       'base_height',
                       'base_length',
                       'nasal_height',
                       'time_period'
                      )
head(pb3_data)
```

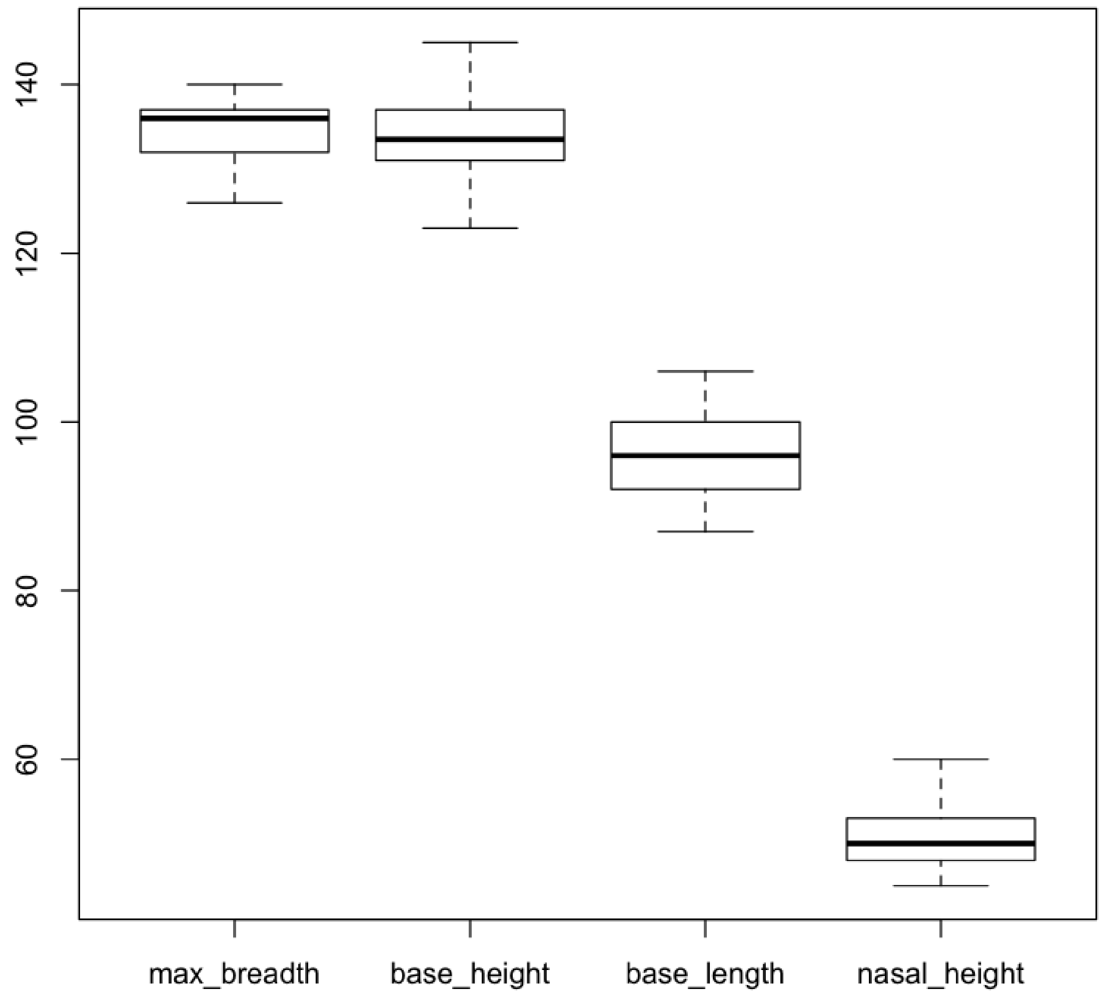
max_breadth	base_height	base_length	nasal_height	time_period
131	138	89	49	1
125	131	92	48	1
131	132	99	50	1
119	132	96	44	1
136	143	100	54	1
138	137	89	56	1

In [13]:
```
# EDA

period1 = pb3_data[pb3_data$time_period == 1,]
period2 = pb3_data[pb3_data$time_period == 2,]
period3 = pb3_data[pb3_data$time_period == 3,]

boxplot(period1[1:4])
boxplot(period2[1:4])
boxplot(period3[1:4])
```

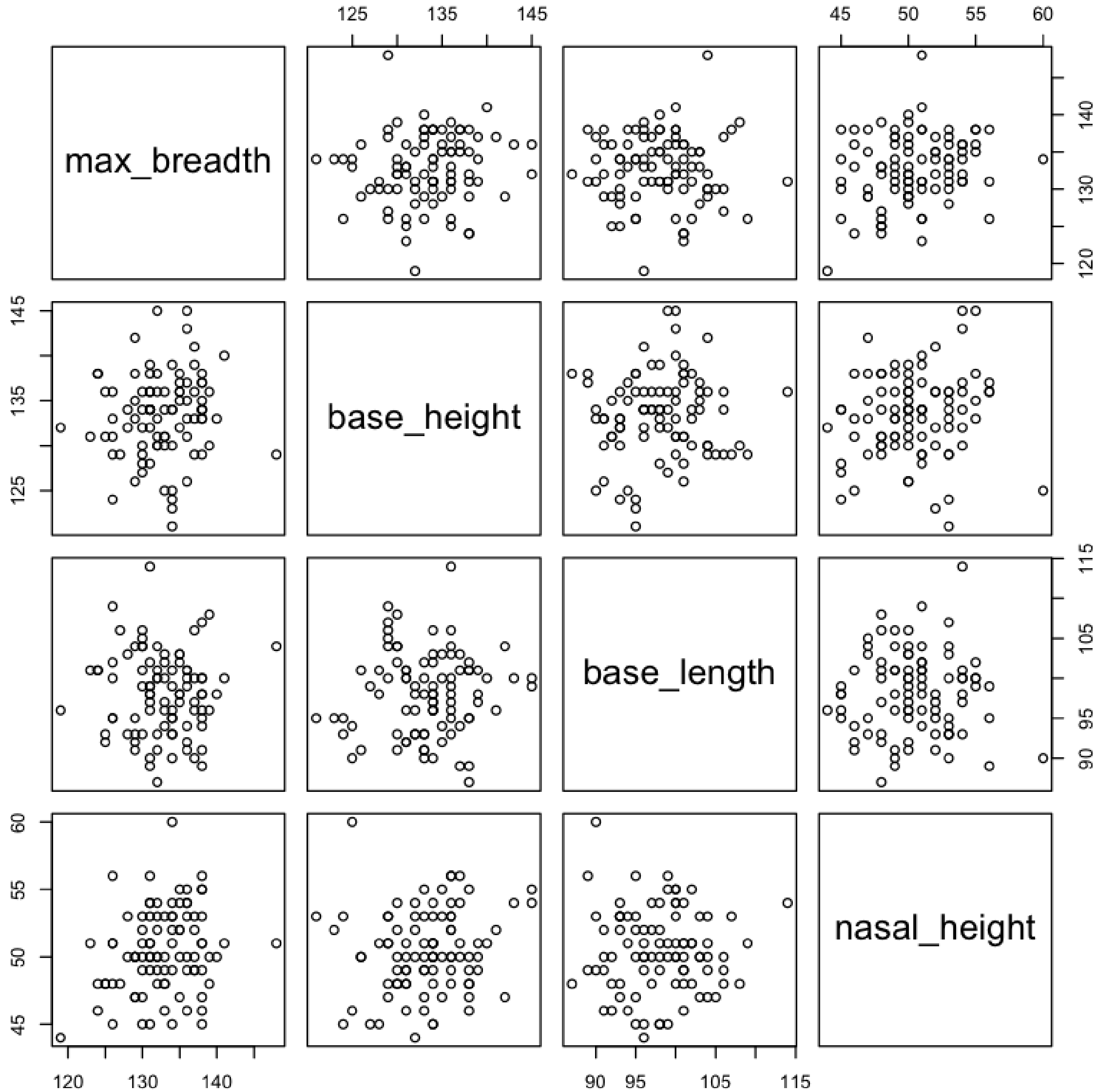




140
120
100
80
60
max_breadth
base_height
base_length
nasal_height


140
120
100
80
60
max_breadth
base_height
base_length
nasal_height

```
In [14]: plot(pb3_data[1:4])
```



General Comments

1. The question we're examining is 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.
2. *Null Hypothesis*: u1=u2=u3, where each u is a mean vector consisting of the 4 measurements for that time period. There has been no change in skull size over the course of the time periods
3. *Alternative Hypothesis*: At least one ui != uj for some i, j in set(1, 2, 3). There has been a change in skull size over the course of the time periods
4. Test: One-way MANOVA
5. Reasoning: It makes sense to go with this test because we have only one factor (time period) with 3 levels (1, 2, 3) and that affects multiple dependent variables (max breadth, base height, base length, nasal height), which is why this is the multivariate case and not the univariate.
6. Further resources for One-way Manova

- https://statistics.laerd.com/spss-tutorials/one-way-manova-using-spss-statistics.php (https://statistics.laerd.com/spss-tutorials/one-way-manova-using-spss-statistics.php)
- http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance#compute-manova-in-r (http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance#compute-manova-in-r)

7. There doesn't seem to be a particularly strong correlation among the variables

In [15]:

```
# Running Statistical Test
timePeriod = as.factor(pb3_data$time_period)

results = manova(
    cbind(max_breadth, base_height, base_length, nasal_height) ~ timePeriod,
    data=pb3_data
)

results
```

```
Call:
   manova(cbind(max_breadth, base_height, base_length, nasal_height) ~
    timePeriod, data = pb3_data)

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
```

In [16]:

```
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
```

In [17]:
```
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
```

Test Result Interpretation

With a significance level of .05, our p-value of .0489 indicates that we can reject the null hypothesis that no interbreeding occurred.

Based on the summary results, there was statistically significant variance in two of the variables over time (max_breadth and base_length)

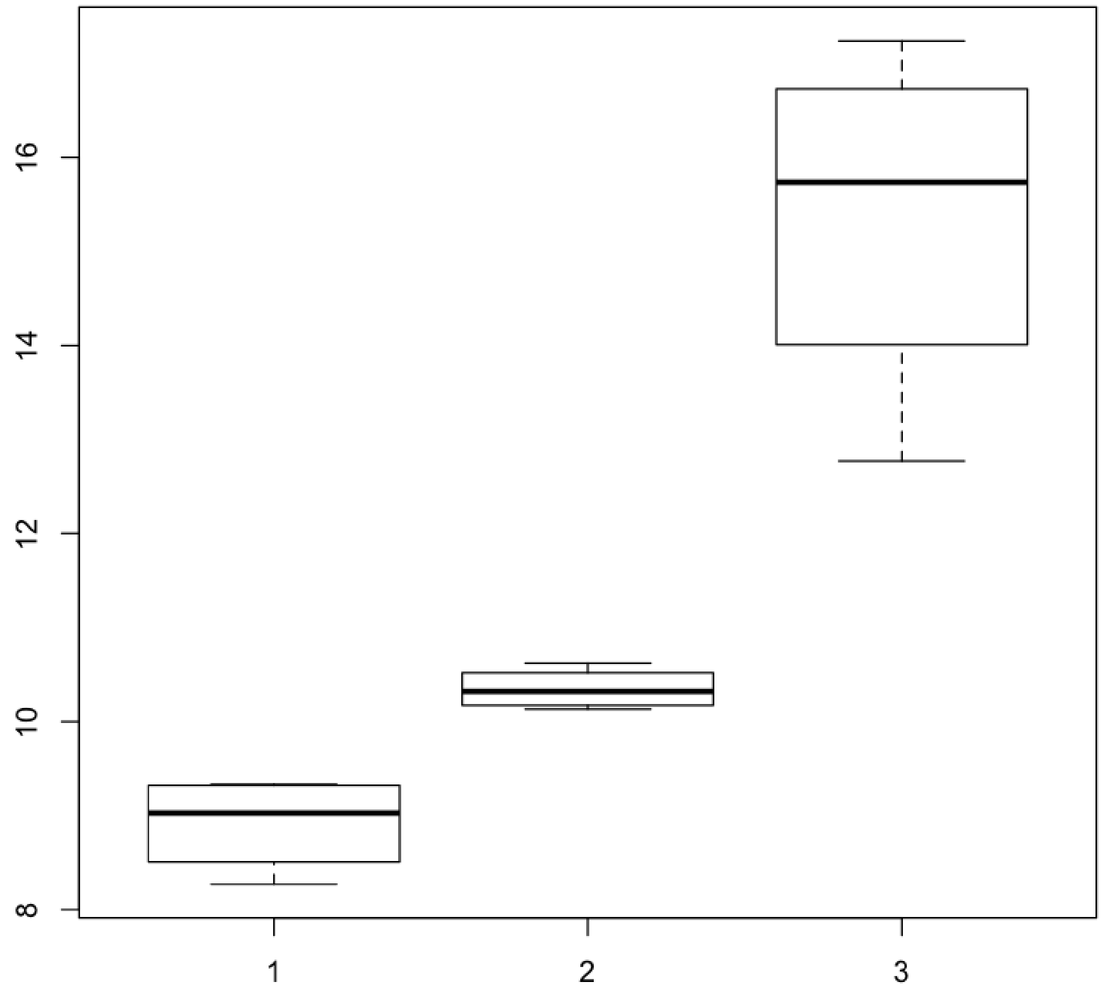
Problem 4

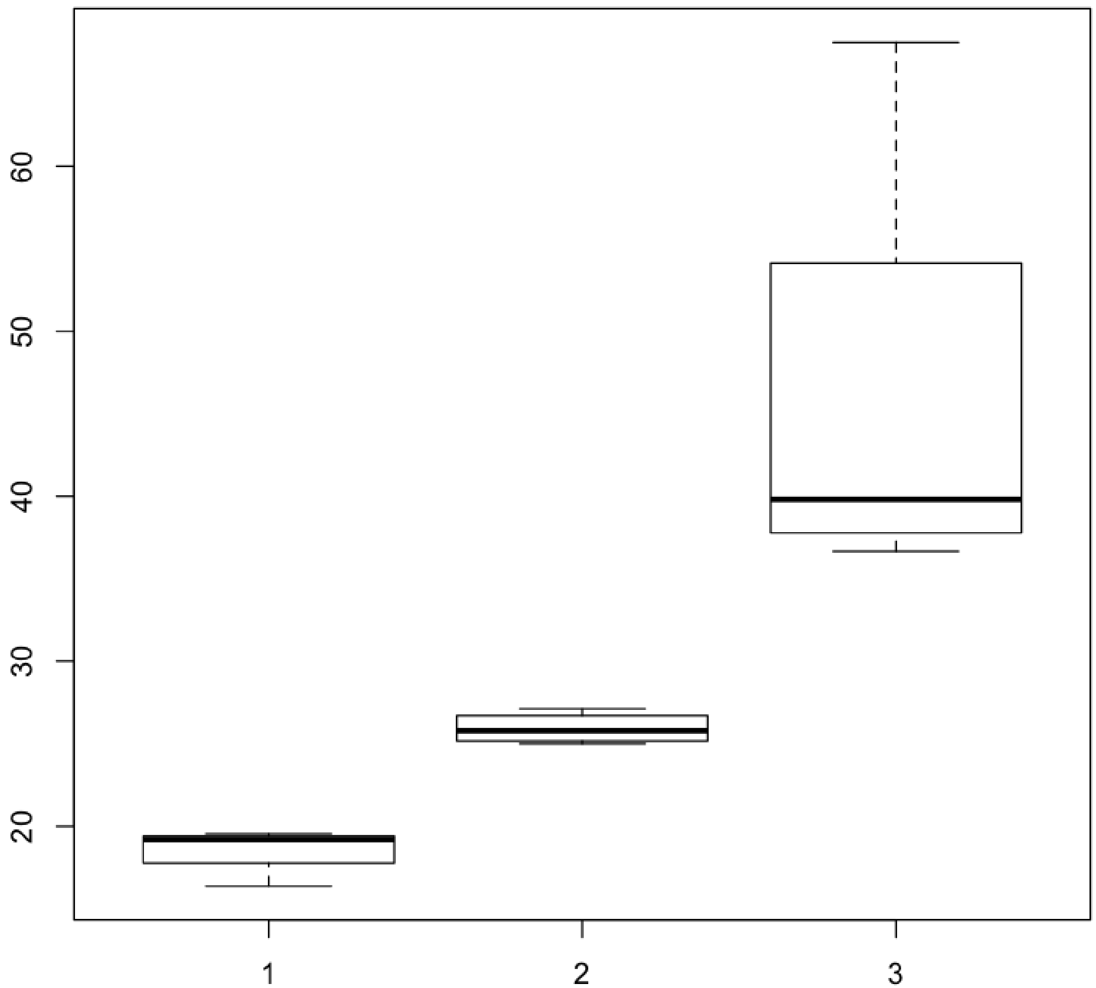
In [18]:
```
# Loading data
pb4_data = read.table(file='Data-HW1-Sensing.dat', header=FALSE, quote='')
colnames(pb4_data) = c('reflectance_green',
                       'reflectance_near_infared',
                       'species',
                       'time_period',
                       'treeID' #Unique to each species
                      )
head(pb4_data)
```

reflectance_green	reflectance_near_infared	species	time_period	treeID
9.33	19.14	SS	1	1
8.74	19.55	SS	1	2
9.31	19.24	SS	1	3
8.27	16.37	SS	1	4
10.22	25.00	SS	2	1
10.13	25.32	SS	2	2

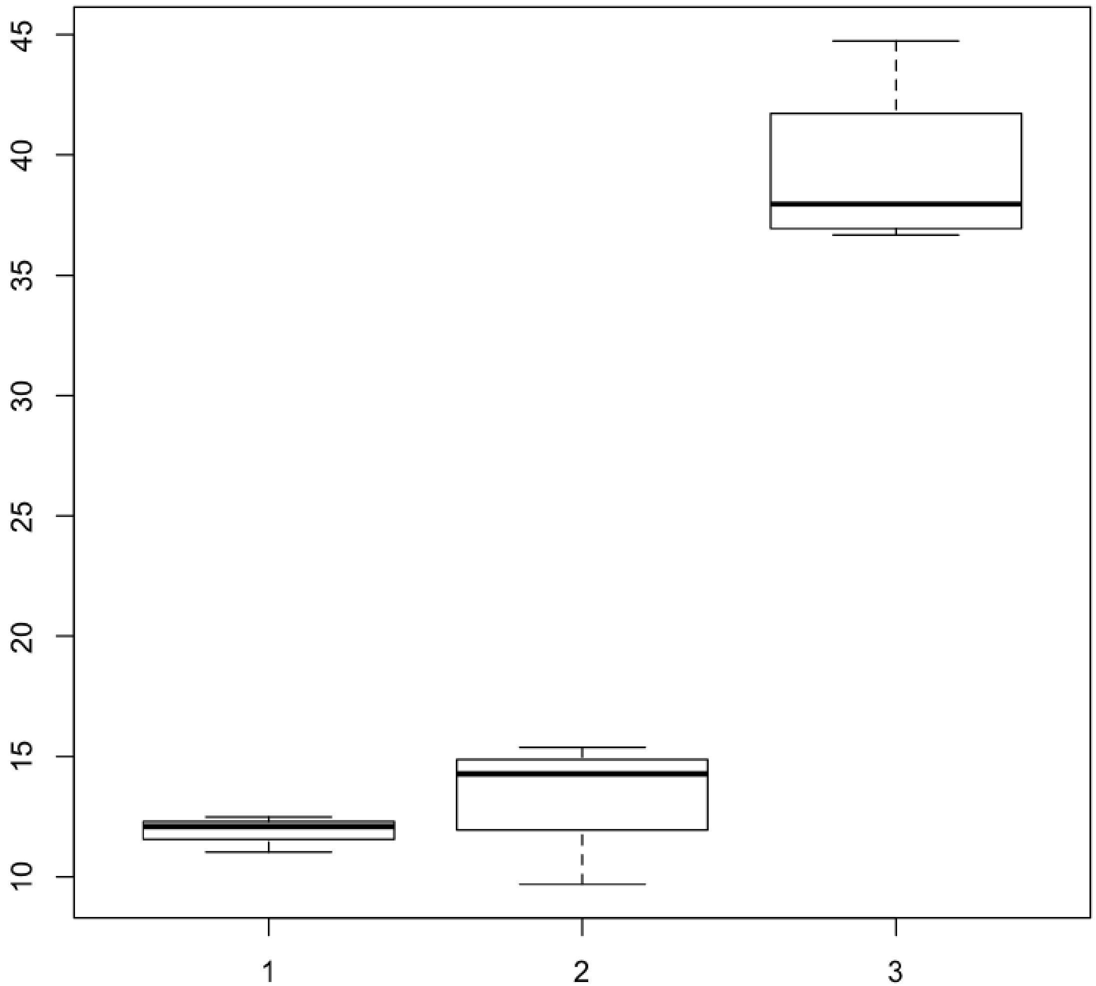
In [19]:
```
#Partitioning dataset into distinct species for EDA
SS = pb4_data[pb4_data$species == 'SS',]
JL = pb4_data[pb4_data$species == 'JL',]
LP = pb4_data[pb4_data$species == 'LP',]
```

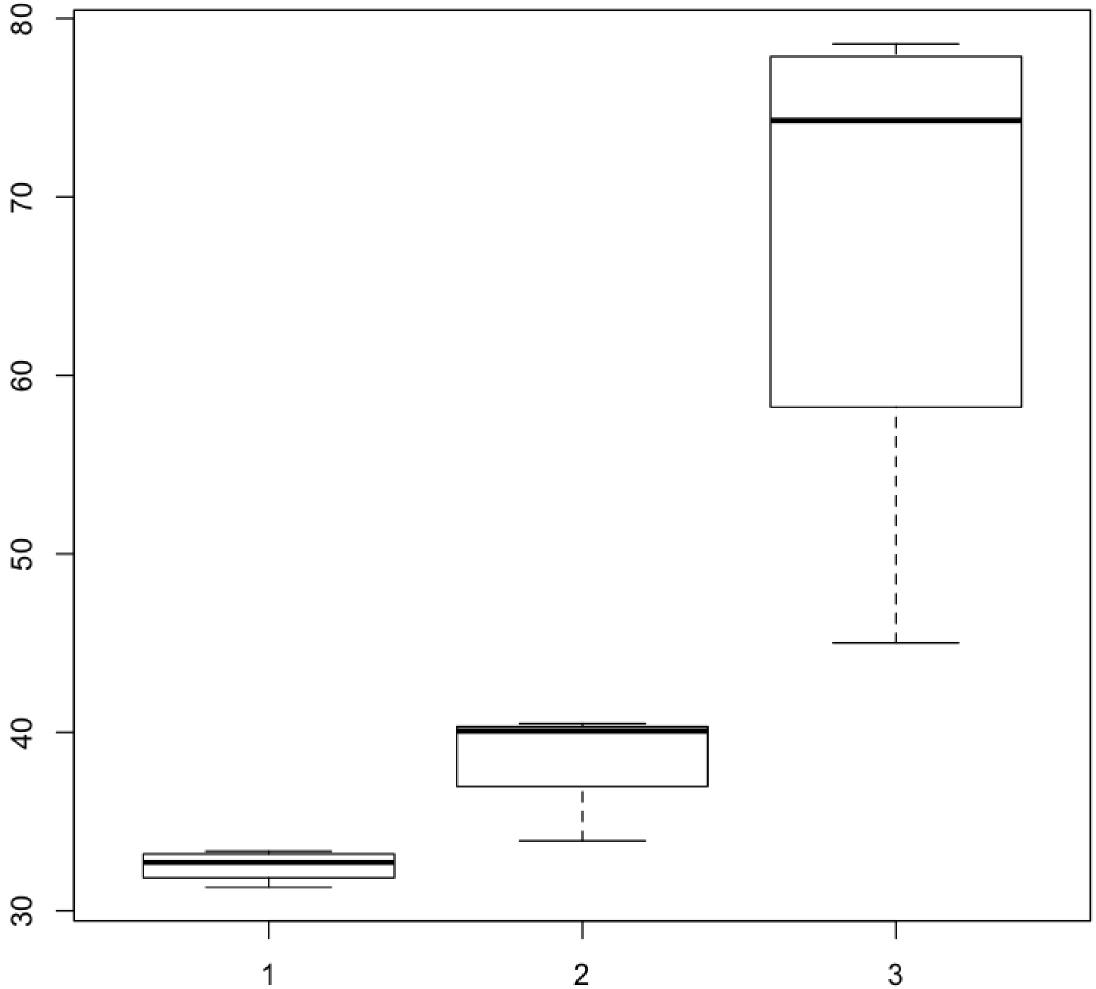
In [20]:
```
#EDA of Species SS
boxplot(reflectance_green ~ time_period, data=SS)
boxplot(reflectance_near_infared ~ time_period, data=SS)
```



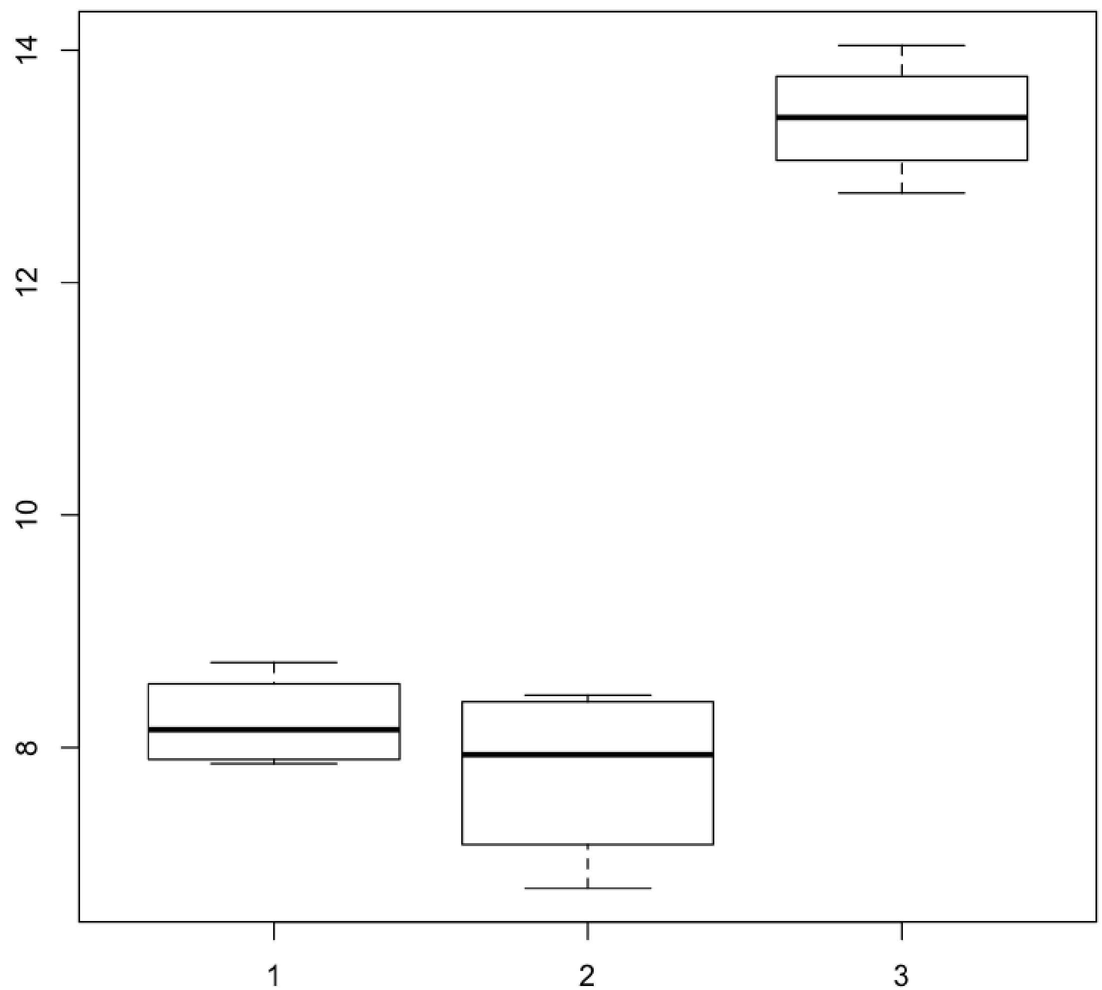

60
50
40
30
20
1
2
3

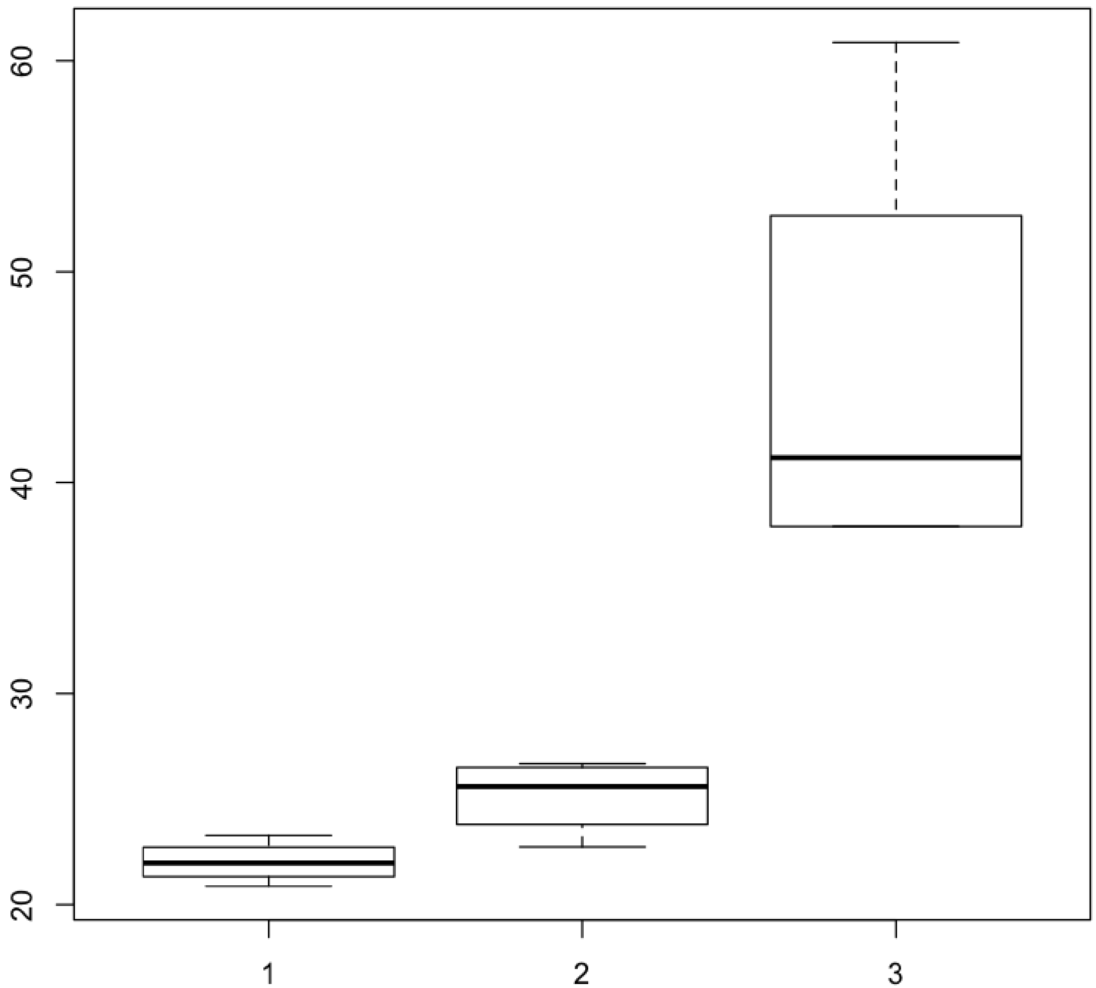
In [21]:
```
#EDA of Species JL
boxplot(reflectance_green ~ time_period, data=JL)
boxplot(reflectance_near_infared ~ time_period, data=JL)
```




80
70
60
50
40
30
1
2
3

In [22]:
```
#EDA of Species LP
boxplot(reflectance_green ~ time_period, data=LP)
boxplot(reflectance_near_infared ~ time_period, data=LP)
```





General Comments

1. The question we're examining is whether there is a difference between our two dependent variables (green reflectance and near-infared reflectance) 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 (factors).
2. *Null Hypothesis*: u1=u2=u3, where each u is a matrix composed of three mean vectors, each consisting of the 2 reflectance measurements for a time period while each matrix corresponds to a species. There is no species effect, no time effect, and no interaction effect on the green and near-infared reflectance of the seedlings.
3. *Alternative Hypothesis*: At least one ui != uj for some i, j in set(1, 2, 3). 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.
4. Test: Two-way MANOVA
5. Reasoning: It makes sense to go with this test because 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). The presence of more than one dependent variable in our analysis explains why we're choosing MANOVA over the univariate case.

6. Further resources for Two-way Manova
   - Understanding two way MANOVA: https://statistics.laerd.com/spss-tutorials/two-way-manova-using-spss-statistics.php (https://statistics.laerd.com/spss-tutorials/two-way-manova-using-spss-statistics.php)
   - Using MANOVA in R: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance#compute-manova-in-r (http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance#compute-manova-in-r)
   - Looking at the Interaction Effect: https://www.r-bloggers.com/r-tutorial-series-two-way-anova-with-interactions-and-simple-main-effects/ (https://www.r-bloggers.com/r-tutorial-series-two-way-anova-with-interactions-and-simple-main-effects/)
7. From EDA, it appears infared reflectance is higher than green reflectance for corresponding time periods across all species.
8. From EDA, it seems like there would be a species effect for both reflectances as across species the boxplots indicates fairly different values for all of them
9. From EDA, Reflectance steadily increases for both reflectances in seedlings as our time period increases across all species.

In [23]:

```
# Running Statistical Test
timePeriod = as.factor(pb4_data$time_period)
species = as.factor(pb4_data$species)

results = manova(
    cbind(reflectance_green, reflectance_near_infared) ~ timePeriod*species,
    data=pb4_data
)

results
```

```
Call:
   manova(cbind(reflectance_green, reflectance_near_infared) ~ timePeriod
*
    species, data = pb4_data)

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
```

In [24]:
```
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
```

In [25]:
```
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
```

Test Result Interpretation

With a significance level of .05, our p-values are much smaller and indicate that we can 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 due to a time and species effect. However, our evidence suggest that the an interaction effect was only applicable to green reflectance, and there was no evidence of an interaction for near_infared reflectance.

In [ ]: