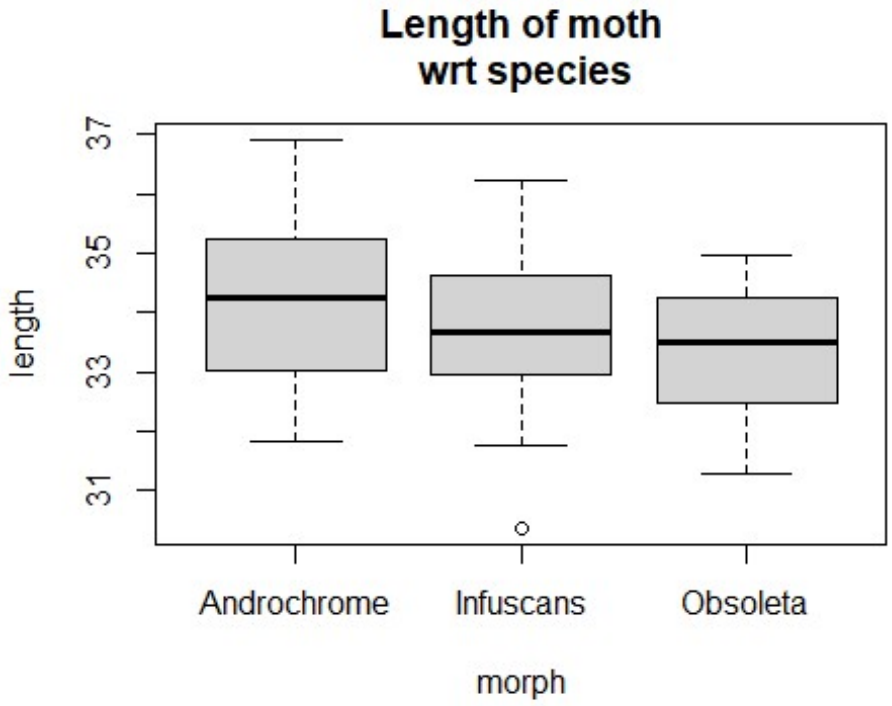
Question 4 - BIOS14 (HT2021)

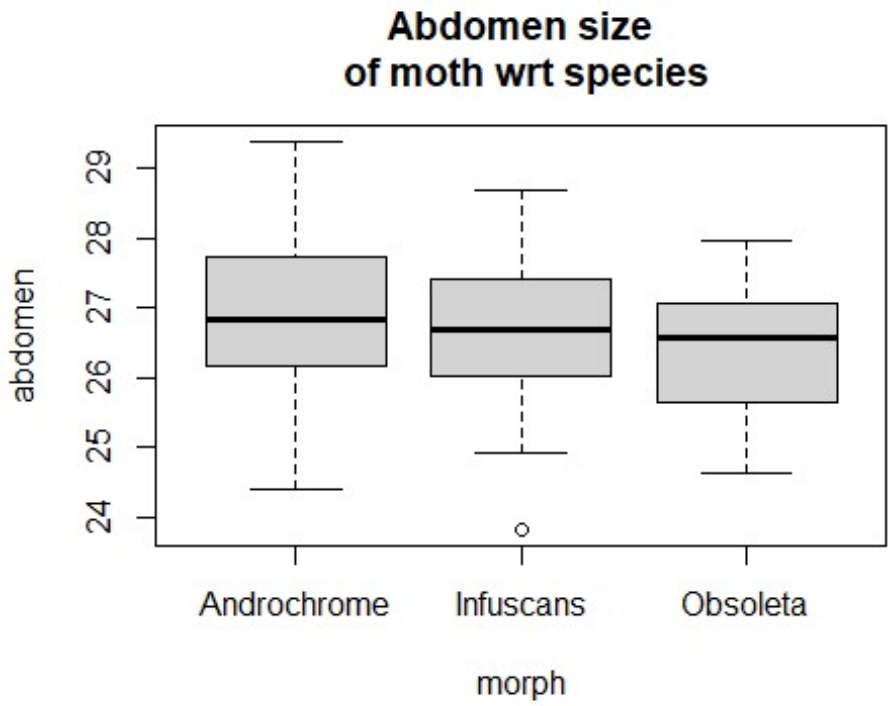
Read and explore dataset

```
#Question 4

rm(list=ls())

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Warning: package 'survival' was built under R version 4.1.2

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

library(lmtest)

## Warning: package 'lmtest' was built under R version 4.1.2

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

library(car)

## Loading required package: carData

library(MVN)

## Warning: package 'MVN' was built under R version 4.1.2

#read and explore data
morph <- read.csv('morphs.csv')

morph$morph <- factor(morph$morph)
```

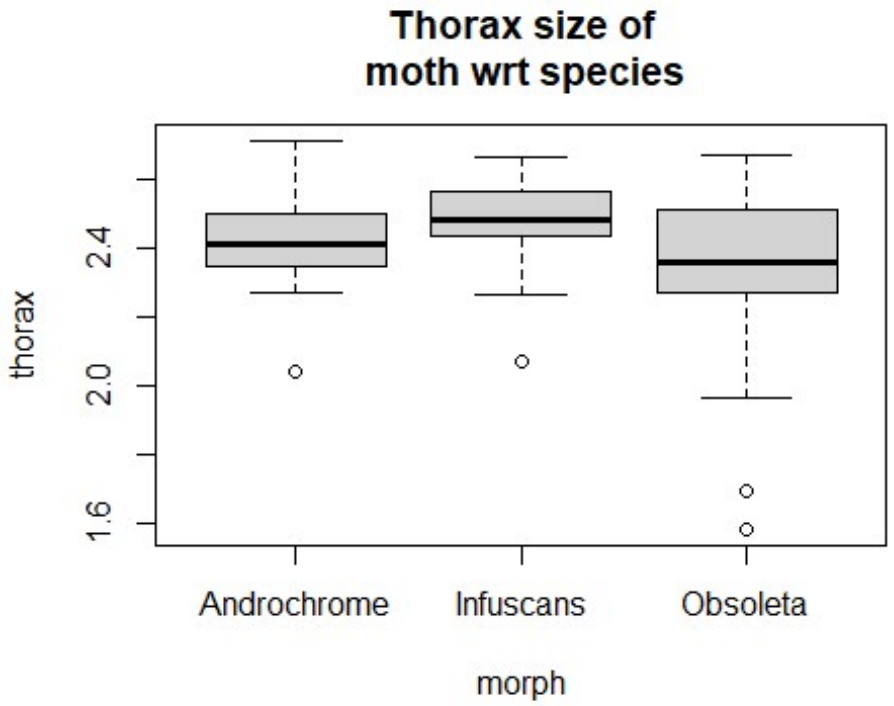
```
plot(length~morph, data=morph, main='Length of moth\n wrt species')
```



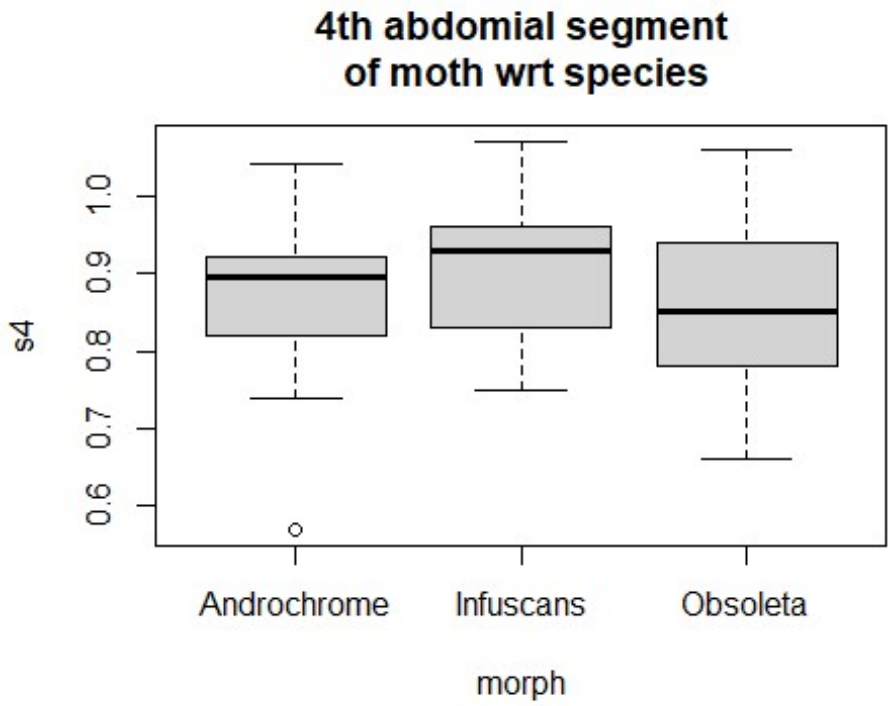

```
plot(abdomen~morph, data=morph, main='Abdomen size\n of moth wrt species')
```



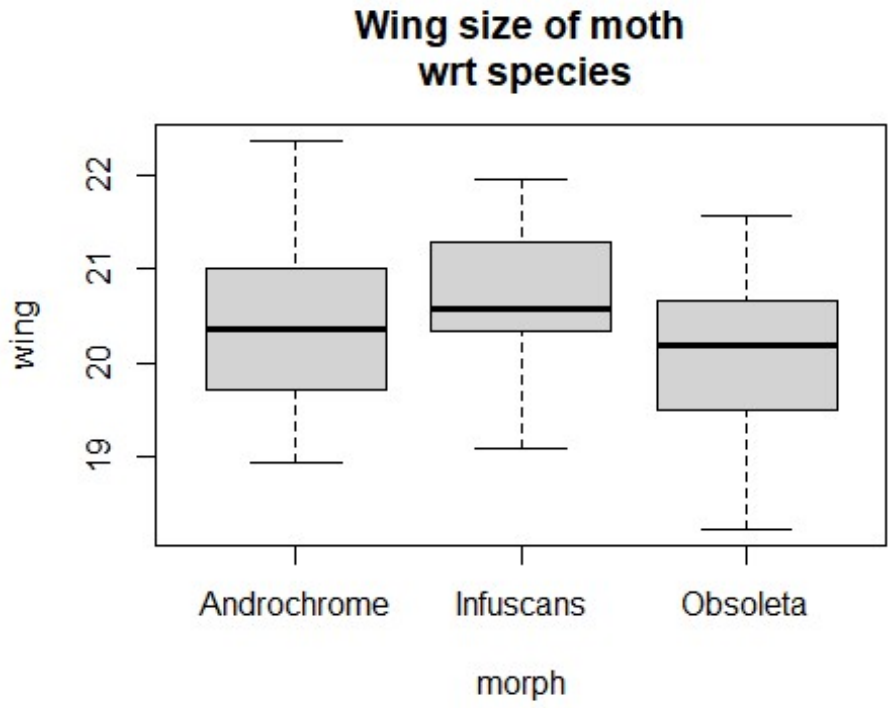

```
plot(thorax~morph, data=morph, main='Thorax size of\n moth wrt species')
```




```
plot(s4~morph, data=morph, main='4th abdomial segment\n of moth wrt species')
```




```
plot(wing~morph, data=morph, main='Wing size of moth\n wrt species')
```



Part (a)

The data set contains five morphological data that are measures of fly length, abdomen width, thorax size and, width of the 4th abdominal segment, and wing size.

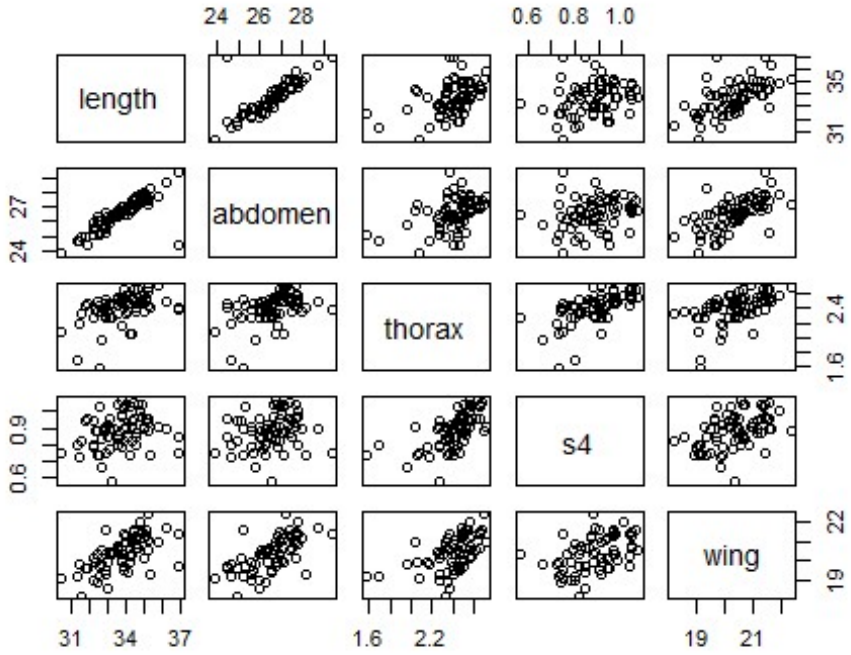
Here, I have used **Principal Component Analysis** to reduce the five traits into a single variable. This variable is the principal component score, which is provided to each observation in the data set by taking into account weight age of each measure to provide a singular value that is representative of each measure/variable and their correlation with one another.

The code below was used to perform the analysis

```
#variable reduction using PCA

#check assumptions of PCA

#assumption 1: sufficient sample size
#5*number of i/p variables = 5*5 = 25 < 68 (number of samples)
#true

#assumption 2: input variables are correlated
rcorr(as.matrix(morph[,2:6]))
```

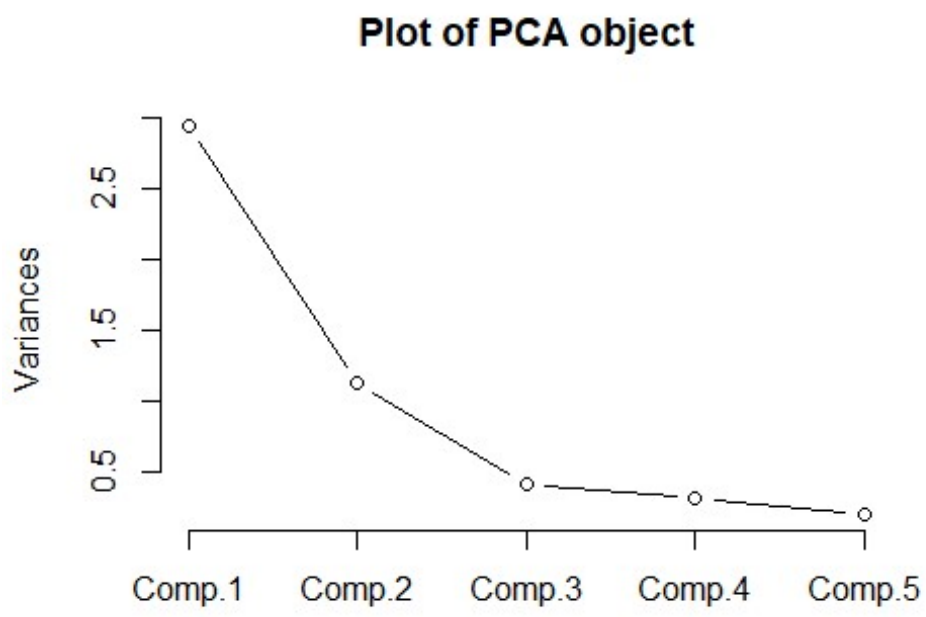
```
##         length abdomen thorax   s4 wing
## length    1.00    0.81   0.47 0.25 0.62
## abdomen   0.81    1.00   0.45 0.23 0.64
## thorax    0.47    0.45   1.00 0.60 0.52
## s4        0.25    0.23   0.60 1.00 0.45
## wing      0.62    0.64   0.52 0.45 1.00
##
## n= 68
##
##
## P
##         length abdomen thorax s4     wing
## length         0.0000  0.0000 0.0370 0.0000
## abdomen 0.0000         0.0001 0.0587 0.0000
## thorax  0.0000 0.0001         0.0000 0.0000
## s4      0.0370 0.0587  0.0000        0.0001
## wing    0.0000 0.0000  0.0000 0.0001
```

```
#all variables correlated with each other - relationship is also
statistically significant

#assumption 3: linearity of relationship between variables
pairs(morph[,2:6])
```




```
#no patterns observed with plot data between any variables :
#all have linear relationship with one another

#assumption 4: no outliers - removing all outliers
#outliers - fall > 3SD away from mean
outlier <- NULL
for(i in 2:6){
  outlier1 <- morph[,i]>=mean(morph[,i])+3*sd(morph[,i]) |
morph[,i]<=mean(morph[,i])-3*sd(morph[,i])
  outlier2 <- which(outlier1==TRUE)
  outlier <- c(outlier, outlier2)
  outlier <- unique(outlier)
}
outlier

## [1] 57 61

#remove outliers
morph_mod <- morph[-outlier,]
str(morph_mod)

## 'data.frame':    66 obs. of  6 variables:
##  $ morph  : Factor w/ 3 levels "Androchrome",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ length : num  32.4 35 36.9 34.9 34 ...
##  $ abdomen: num  26.1 27.7 29.4 27.8 26.8 ...
##  $ thorax : num  2.5 2.57 2.37 2.49 2.59 2.41 2.71 2.27 2.41 2.35 ...
```

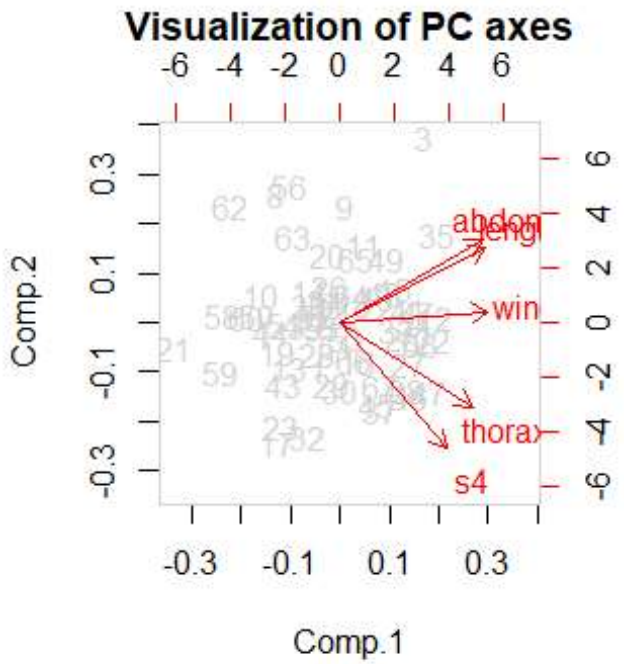
```
##  $ s4     : num  0.88 0.91 0.75 0.97 1.04 0.84 0.91 0.57 0.74 0.74 ...
##  $ wing   : num  20.3 20.1 21.4 21.4 21.3 ...

#2 samples removed

#construct PCA

morph.pca <- princomp(morph_mod[,2:6], cor=TRUE)

#get results
summary(morph.pca)

## Importance of components:
##                           Comp.1    Comp.2     Comp.3    Comp.4     Comp.5
## Standard deviation     1.7152662 1.0631414 0.64210960 0.5621486 0.44640416
## Proportion of Variance 0.5884276 0.2260539 0.08246095 0.0632022 0.03985533
## Cumulative Proportion  0.5884276 0.8144815 0.89694246 0.9601447 1.00000000

plot(morph.pca, type='lines', main='Plot of PCA object')
```




```
#first two components important- variances > 1

loadings(morph.pca)

##
## Loadings:
##         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
```

```
## length   0.479  0.398  0.297  0.202  0.695
## abdomen  0.470  0.443  0.153  0.232 -0.711
## thorax   0.436 -0.447  0.493 -0.601
## s4       0.351 -0.665 -0.122  0.648
## wing     0.486        -0.794 -0.352
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

```
#visualize
biplot(morph.pca, col=c('lightgrey', 'red'), main="Visualization of PC axes")
```




```
#adding PC scores to the existing data frame
morph_mod$PC1 <- morph.pca$scores[,1]
morph_mod$PC2 <- morph.pca$scores[,2]

#Here, PC2 does not include wing size in the scoring variable.
#Hence, we can consider only PC1 scores as the new variable that measures
overall size of the damselfly.
```

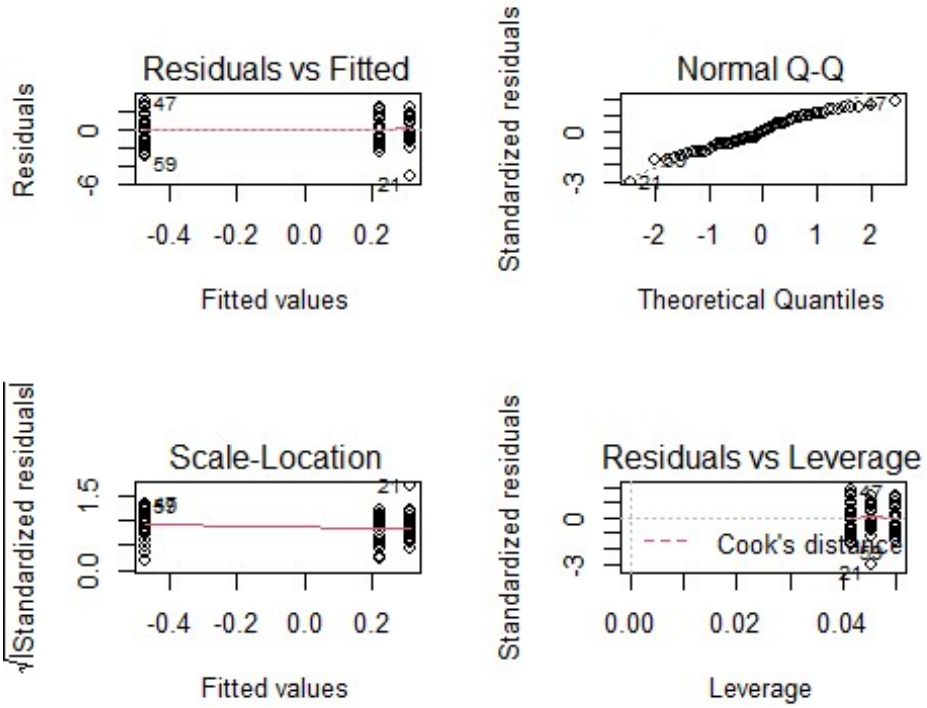
Part (b)

To test if female morphs differ in overall size, we use the new variable we obtained from the previous analysis. Here, I have performed **one-way ANOVA** for both PC1 and PC2.

However, only PC1 is significant for our analysis and hence, I have illustrated the results of only PC1 here.

Since we are doing an ANOVA analysis where we compare groups based on a categorical variable and not a linear model, I have eliminated the intercept value and forced comparison of each of the groups with zero as a reference value.

```
#PC1
fit.pc1 <- lm(PC1~morph-1, data=morph_mod)

#assumption 1: independent observations
#assume that data collected is independent

#assumption 2: Normality of residuals
par(mfrow=c(2,2))
plot(fit.pc1)
```




```
par(mfrow=c(1,1))
#residuals normality tested

#assumption 3: Homogeneity of variances
leveneTest(fit.pc1)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.2567 0.7744
##       63
```

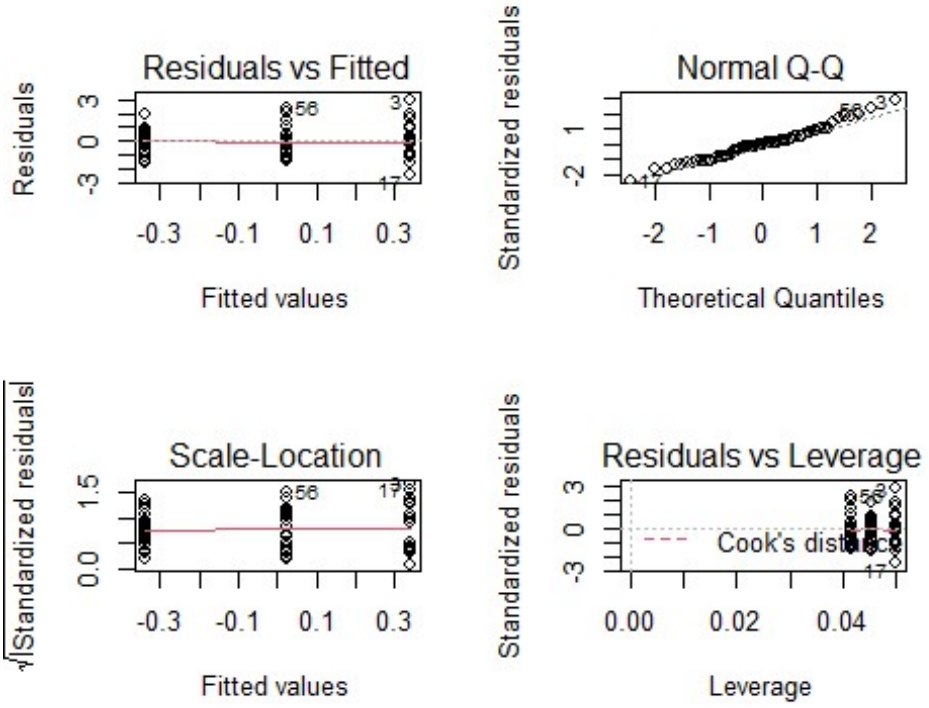
```r
#variances are homogeneous - test is not  significant

anova(fit.pc1)
```

```
## Analysis of Variance Table
##
## Response: PC1
##           Df  Sum Sq Mean Sq F value Pr(>F)
## morph      3   8.401  2.8002  0.9496 0.4222
## Residuals 63 185.780  2.9489
```

```r
summary(fit.pc1)
```

```
##
## Call:
## lm(formula = PC1 ~ morph - 1, data = morph_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0203 -1.1545 -0.0676  1.4606  2.9735
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## morphAndrochrome   0.2211     0.3840   0.576    0.567
## morphInfuscans     0.3113     0.3661   0.850    0.398
## morphObsoleta     -0.4696     0.3505  -1.340    0.185
##
## Residual standard error: 1.717 on 63 degrees of freedom
## Multiple R-squared:  0.04326,    Adjusted R-squared:  -0.002297
## F-statistic: 0.9496 on 3 and 63 DF,  p-value: 0.4222
```

```r
#PC2
fit.pc2 <- lm(PC2~morph-1, data=morph_mod)

#assumption 1: independent observations
#assume that data collected is independent

#assumption 2: Normality of residuals
par(mfrow=c(2,2))
plot(fit.pc2)
```




```
par(mfrow=c(1,1))
#residuals normality tested

#assumption 3: Homogeneity of variances
leveneTest(fit.pc2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.5795 0.5632
##       63
```

```
#variances are homogeneous - test is not  significant

anova(fit.pc2)
```

```
## Analysis of Variance Table
##
## Response: PC2
##           Df Sum Sq Mean Sq F value Pr(>F)
## morph      3  4.828  1.6094  1.4533 0.2358
## Residuals 63 69.770  1.1075
```

```
summary(fit.pc2)
```

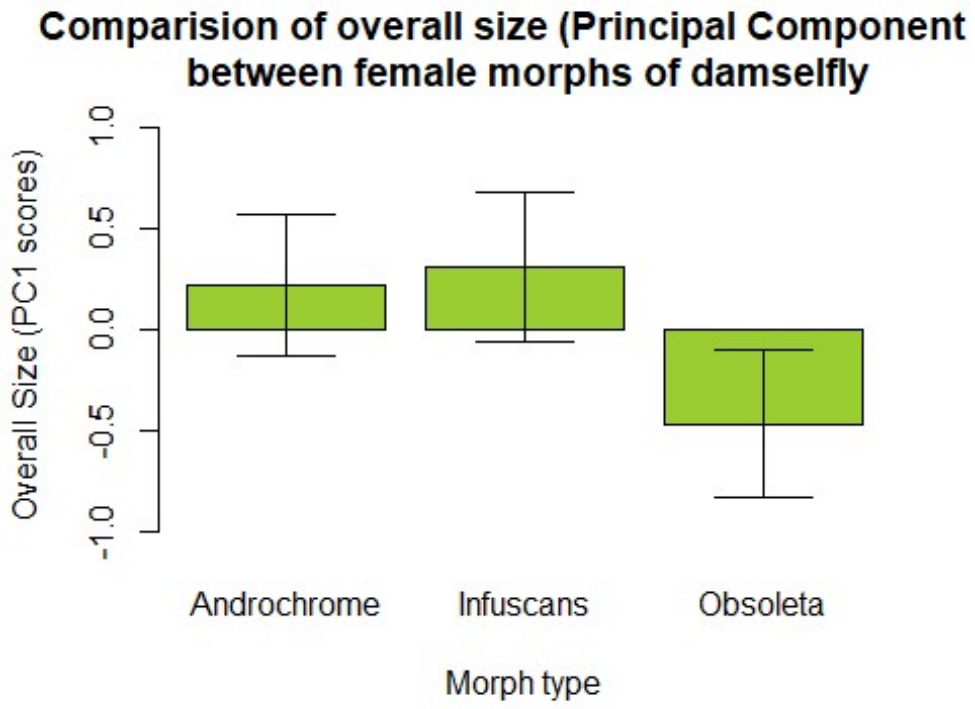
```
##
## Call:
## lm(formula = PC2 ~ morph - 1, data = morph_mod)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.43232 -0.80542 -0.00488  0.51869  2.89955
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## morphAndrochrome  0.33963    0.23531   1.443    0.154
## morphInfuscans   -0.33742    0.22436  -1.504    0.138
## morphObsoleta     0.02628    0.21481   0.122    0.903
##
## Residual standard error: 1.052 on 63 degrees of freedom
## Multiple R-squared:  0.06472,    Adjusted R-squared:  0.02019
## F-statistic: 1.453 on 3 and 63 DF,  p-value: 0.2358
```

From the analysis, we can see that there is no significant difference in overall size between the three morphs; implies that the prey they prefer is likely similar.

Below is the illustration of results for PC1

```
#illustrate results - for PC1 which is representative of the overall size
std.err <- function(x) sd(x, na.rm=TRUE)/sqrt(sum(!is.na(x)))

mean.pc1 <- aggregate(list(mean=morph_mod$PC1),
by=list(morph=morph_mod$morph), mean, na.rm=TRUE)
se.pc1 <- aggregate(list(se=morph_mod$PC1), by=list(morph=morph_mod$morph),
std.err)

bp1.dat <- cbind(mean.pc1, se.pc1)

bp1 <- barplot(mean~morph, data=bp1.dat,
               main='Comparision of overall size (Principal Component 1)\n
between female morphs of damselfly',
               xlab='Morph type',
               ylab='Overall Size (PC1 scores)',
               names=c('Androchrome', 'Infuscans', 'Obsoleta'),
               col='yellowgreen',
               ylim=c(min(bp1.dat$mean)-0.6, max(bp1.dat$mean)+0.7))
arrows(bp1, bp1.dat$mean+bp1.dat$se, bp1, bp1.dat$mean-bp1.dat$se, code=3,
angle=90)
```



Part (c)

The one-way ANOVA analysis in part (b) was performed under the important assumption that the observations are all independent.

The design of this study could be improved by collecting data across years from different study sites. This would improve the randomness of the samples, reduce any dependency within observations and hence give unbiased results.